A Data Science Big Mechanism for DARPA

Story

Slides

Slide 1 Big Mechanism
Slide 2 Big Data, 1854
Slide 3 Big Data, 2013
Slide 4 The Problem
Slide 5 The Solution: Read, Assemble, Explain
Slide 6 Read, Assemble, Explain for Which Systems?
Slide 7 Complicated Systems
Slide 8 Knowledge is Fragmented and Voluminous
Slide 9 Knowledge is Inconsistent
Slide 10 Technology Development Tasks 1
Slide 11 The Reading Task: Technology Challenges
Slide 12 The Reading Task: The Language
Slide 13 The Reading Task: Previous Results
Slide 14 Technology Development Tasks 2
Slide 15 The Assembly Task – Technology Challenges
Slide 16 The Assembly Task – Ad Hoc Representation
Slide 17 The Assembly Task – Previous Results
Slide 18 Technology Development Tasks 3
Slide 19 The Explanation Task – Technology Challenges
Slide 20 The Explanation Task – Previous Results
Slide 21 Technology Development Tasks 4
Slide 22 The Integration Task – Technology Challenges
Slide 23 Conclusion

Spotfire Dashboard
Research Notes
Meetup
DARPA Memex
Conference on Semantics in Healthcare and Life Sciences
Introduction

Topic Areas

Submissions

Keynote Speakers

Ronald Collette, CISSP
Hilmar Lapp
Deborah McGuinness, Ph.D
Prof. Dr. B. (Barend) Mons

Tech Talks

High Availability and Graph Mining using bigdata® - an Open-Source Graph Database
Successfully Navigating Diagnosis And Treatment In The Age Of Targeted Cancer Therapy
Making Sense of Big Data in Pharma
Using Supercomputer Architecture for Practical Semantic Web Research in the Life Sciences

Presenters

Scott Bahlavooni
Evan Bolton
Aftab Iqbal
Andrew M. Jenkinson
Bernadette Hyland
Hyeongsik Kim
David King
David H. Mason
Chimezie Ogbuji
Tom Plasterer
Eric Prud'Hommeaux
Arash Shaban-Nejad
Robert Stanley

Poster Presentations

Poster 01 A Semantic Computing Platform to Enable Translational and Clinical Omics
Poster 02 Optimization of RDF Data for Graph-based Applications
Poster 03 A Phenome-Guided Drug Repositioning through a Latent Variable Model
Poster 04 Cross-Linking DOI Author URIs Using Research Networking Systems
Poster 05 Simplified Web 3.0 Query Portal for Biomedical Research
Poster 06 A Minimal Governance Layer for the Web of Linked Data
Poster 07 BAO Reveal: Assay Analysis Using the BioAssay Ontology and Open PHACTS
Poster 08 A High Level API for Fast Development of High Performance Graphic Analytics on GPUs
Poster 09 A Nanopublication Framework for Systems Biology and Drug Repurposing
Poster 10 An Interoperable Framework for Biomedical Image Retrieval and Knowledge Discovery
PART I: OVERVIEW

PART II: FULL TEXT OF ANNOUNCEMENT

I. FUNDING OPPORTUNITY DESCRIPTION
   A. Introduction
   B. Program Description/Scope
   C. Technical Areas
      1. Reading
      2. Assembly
      3. Explanation
   D. Program Structure
   E. Schedule/Milestones
   F. Performance Requirements
   G. Government-furnished Property/Equipment/Information
   H. Intellectual Property
   I. Additional Guidance for Proposers

II. AWARD INFORMATION
   A. Awards
   B. Fundamental Research

III. ELIGIBILITY INFORMATION
   A. Eligible Applicants
      1. Federally Funded Research and Development Centers (FFRDCs) and Government Entities
      2. Foreign Participation
   B. Procurement Integrity, Standards of Conduct, Ethical Considerations and Organizational Conflicts of Interest (OCIs)
   C. Cost Sharing/Matching
   D. Other Eligibility Requirements
      Ability to Receive Awards in Multiple Technical Areas - Conflicts of Interest

IV. APPLICATION AND SUBMISSION INFORMATION
   A. Address to Request Application Package
   B. Content and Form of Application Submission
      1. Proposals
      2. Proprietary and Classified Information
   C. Submission Dates and Times
      1. Proposals
D. Funding Restrictions
E. Other Submission Requirements
   1. Unclassified Submission Instructions
   2. Classified Submission Instructions

V. APPLICATION REVIEW INFORMATION
   A. Evaluation Criteria
   B. Review and Selection Process

VI. AWARD ADMINISTRATION INFORMATION
   A. Selection Notices
   B. Administrative and National Policy Requirements
      1. Intellectual Property
      2. Human Subjects Research (HSR)
      3. Animal Use
      4. Export Control
      5. Electronic and Information Technology
      6. Employment Eligibility Verification
      7. System for Award Management (SAM) Registration and Universal Identifier Requirements
      8. Reporting Executive Compensation and First-Tier Subcontract Awards
      9. Updates of Information Regarding Responsibility Matters
      10. Representation by Corporations Regarding Unpaid Delinquent Tax Liability or a Felony Conviction under Any Federal Law
      11. Cost Accounting Standards (CAS) Notices and Certification
      12. Controlled Unclassified Information (CUI) on Non-DoD Information Systems
      13. Safeguarding of Unclassified Controlled Technical Information (This applies only to FAR-based awards)
   C. Reporting
      1. Technical and Financial Reports
      2. Representations and Certifications
      3. Wide Area Work Flow (WAWF)
      4. i-Edison

VII. AGENCY CONTACTS

VIII. OTHER INFORMATION
   A. Frequently Asked Questions (FAQs)
   B. Submission Checklist

Footnotes
1
2
3
A Data Science Big Mechanism for DARPA

DARPA wants to help the DoD get to the essence of cause and effect for cancer from reading the medical literature. The Federal Big Data Working Group Meetup has also been doing that with Semantic Medline - YarcData and Euretos. See the video for Cancer Immunotheraphy (21 minutes) which Science magazine called the biggest breakthrough in 2013 at the end of 2013 and which Dr. Tom Rindflesch (the inventor of Semantic Medline) identified from Semantic Medline as a very important breakthrough in early 2013!

Data Science provides a Data Mining process for doing that which can be applied to DARPA content.

The essence of DARPA’s Big Mechanism is:

- Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.
- Assemble causal fragments into more complete pathways; discover and resolve inconsistencies.
- Explain phenomena in signaling pathways.
- Answer questions, including “reaching down to data,” when it is available.
- Integrate reading, assembly and explanation in a non-pipeline architecture that provides flexible control.
Reading, Assembly and Explanation shouldn’t be pipelined but should use each other opportunistically. Need flexible, non-pipelined control, plus significant software integration.

An intriguing part of the DARPA solicitation is: Prior to the program kick-off, DARPA will develop and provide a "starter kit" to the performers that will contain a first draft of a representation language for biological processes, and means to access the many curated ontologies and databases that are relevant to biological processes. This “starter kit” will be provided to the performers as Government-furnished Information.

The bottom line is:

- It's a big problem
- Many domains
- Distinct communities to coordinate
  - statistical NLP
  - knowledge-based NLP
  - systems biology
  - mathematical biology
  - ontology, databases
  - data mining
  - representation and reasoning
- Potentially a new way to do science

Data Science is all about “reaching down to data” with data mining and the Federal Big Data Working Group Meetup requires Data Science Teams to answer the following questions in their presentations:

- How was the data collected?
- Where is the data stored?
- What are the results?

In fact the new paradigm in scientific research and publishing is being called nanopublications or data publishing which is to put the data first and the paper second so results can be verified and built upon by integration with other data sources. Specifically, Dr. Barend Mons' biomedical data science team is working on a new way (via Open Nanopublications) now in various other editable interfaces.

The DARPA Big Mechanism mashup below is a lot of little mechanisms that need to be integrated into a knowledge base.

Knowledge representation expert John Sowa said recently on Historical Perspectives On Problems of Knowledge Sharing

- Abstract
  - Documents and libraries written and read by humans have been highly successful for the past three millennia. But differences in languages have been a major obstacle for knowledge sharing among humans. Computer systems are much faster than humans, but much less flexible, tolerant, and knowledgeable. This talk will investigate how we can simplify and enhance knowledge sharing among computers and humans.

- Old Fashioned Knowledge Acquisition
Knowledge engineers acquire knowledge from a SME and translate it to some knowledge representation language.

But a KE is also a highly trained professional.

Hiring a SME to train a KE doubles the cost.

**A Migration Path to the Future**

- Enable subject-matter experts to review, update, and extend their knowledge bases with little or no assistance from IT specialists.
- Provide tools that support collaboration, review, and testing by people with different levels and kinds of expertise.

**Knowledge Discovery**

- We need tools that can play the role of Socrates (he was the midwife to his listeners).
- They should help us discover the implicit knowledge and use it to process the huge volumes of digital data.

Professor Michel Dumontier (Stanford BMIR) said recently:

**Abstract**

- I will discuss the current challenge of working with "schema-lite" LOD, various strategies to make sense of this data (mappings/community standards), and how we formalize specific parts (in OWL) so they are fit for purpose (deductive reasoning in conjunction with data mining).

**Main Points:**

- Bio2RDF converts bio-data in RDF format and ensures URI integrity by conferring with its registry of datasets
- Despite all the data, it's still hard to find answers to questions
- Because there are many ways to represent the same data and each dataset represents it differently
- Bio2RDF and SIO powered SPARQL 1.1 federated query: Find chemicals (from CTD) and proteins (from SGD) that participate in the same process (from GOA)
- Knowledge Discovery through Data Integration and Enrichment Analysis
- Tactical Formalization + Automated Reasoning Offers Compelling Value for Certain Problems

The recent Conference on Semantics in Healthcare and Life Sciences (2014) can be "data mined" for similar work.

Some observations:

- Semantic Medline reverses the order by extracting the predications, visualizing their relationships, and discovering answers.
- Nanopublications use Semantic Medline as a baseline to extract "cardinal predications" and compare them to an integrated database of predications. See Open Nanopublications
- An eScience data paper is moving from a narrative based communication into a data-based convincing argument to bridge the gap between publications and data repositories.
- Data Science follows the CRISP-DM (The CRoss-Industry Standard Process for Data Mining) by putting the business problem to be solved and the available data first to guide the process for ultimately deploying a data-driven application that serves the business well.
The later seems to be the essence of the DARPA Big Mechanism Initiative. The recent Conference on Semantics in Healthcare and Life Sciences (2014) shows the difficulty of collecting and storing the data in these disciplines to begin to do data science. So much time and effort is spent in data preparation before data results can be discovered and published in reusable and verifiable form.

So just like for the Euretos-BRAIN, we performed these steps:

- Followed the Data Mining Process (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment)
- Create a Knowledge Base (this includes lots of things like a digital object architecture, metadata, linked data, and ontology) (MindTouch)
- Put the Knowledge Base in a spreadsheet in both relational and graph database formats (Excel)
- Perform analytics and visualizations on the spreadsheet data (Sprofire)
- Develop and Knowledge Model application if needed (Be Informed)

The spreadsheet with the Knowledge Base (MindTouch) and the DRAPA Open Catalog for Software and Publications (Web table to Excel) is attached below and analyzed and visualized in Spotfire.

MORE TO FOLLOW

Slides

Slide 1 Big Mechanism
Slide 4 The Problem

We need to understand the mechanisms of big, complicated systems.

But our knowledge of these mechanisms is increasingly fragmented, voluminous and inconsistent.

Slide 5 The Solution: Read, Assemble, Explain

We need to understand the mechanisms of big, complicated systems.

But our knowledge of these mechanisms is increasingly fragmented, voluminous and inconsistent.

The Solution: Make computers read documents and data, and assemble the fragments they contain into Big Mechanisms that explain causes and effects within systems.
Slide 6 Read, Assemble, Explain for Which Systems?

Slide 7 Complicated Systems

http://www.nature.com/ng/journal/v43...c_id=NG-201112

Cancer biology is a complicated system.

Different authors estimate that on average 7–15 somatic mutations are required for a normal human cell to undergo malignant transformation. The proteins and genes in the figure are all implicated in colorectal cancer.

2. Methyl transferase 
3. Methylation of DNA
4. Protein coded by CTNNB1

Source: Nature 510(7502) 2013
http://www.nature.com/pathobiology/living-workshop/YF_tale2013.html

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch
Slide 8 Knowledge is Fragmented and Voluminous

Cancer biology knowledge is fragmented and voluminous:

PubMed contains 23,000,000 abstracts and grows at roughly 500,000 abstracts per year (source: PubMed)

http://www.biomedcentral.com/1752-0509/6/29

Slide 9 Knowledge is Inconsistent

Cancer biology knowledge is inconsistent:

“...observed remarkably poor agreement (consistently less than 10%) among different databases regarding pathway components.”

“What constitutes a ‘canonical’ pathway is database specific. This inconsistency ... may reflect underlying biology, in that signal transduction events are often context-dependent, or it may reflect the absence of a controlled vocabulary.”

“This raises a significant problem for mechanistic modeling, since ... it is not clear which genes/proteins to include for modeling or experimental measurement.”

– Kirouser et al. BMC Systems Biology 2012, 6:29
http://www.biomedcentral.com/1752-0509/6/29
Slide 10 Technology Development Tasks 1

Technology Development Tasks

- **Read** papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.
- **Assemble** causal fragments into more complete pathways; discover and resolve inconsistencies.
- **Explain** phenomena in signaling pathways. Answer questions, including “reaching down to data,” when it is available.
- **Integrate** reading, assembly and explanation in a non-pipeline architecture that provides flexible control.

Slide 11 The Reading Task: Technology Challenges

The Reading Task: Technology Challenges

Current reading technology extracts semantically shallow assertions; this program intends to “go deep” to extract causality, kinetics, abstraction.

- **Involvement**: "β-catenin is a critical component of Wnt-mediated transcriptional activation."
- **Causal Promotion**: "ARF6 activation promotes the intracellular accumulation of β-catenin."
- **Kinetic**: "L-cells treated with the GSK3β inhibitor LiCl (50 nM) or the proteasome/caspase inhibitor MG132 (25 μM) showed a marked increase in β-catenin fluorescence within 30 – 60 min."
- **Modular**: "...via a mechanism that involves the endocytosis of growth factor receptors and robust activation of extracellular signal-regulated kinases."

Slide 12 The Reading Task: The Language

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC125566/#
The Reading Task: The Language

HBP1 is a repressor of the cyclin D1 gene and inhibits the Wnt signaling pathway. The inhibition of Wnt signaling and growth requires a common domain of HBP1. The apparent mechanism is an inhibition of TCF/LEF DNA binding through physical interaction with HBP1.


Updated: Sat, 19 Sep 2015 00:07:55 GMT

Slide 13 The Reading Task: Previous Results

Extracting Regulatory Gene
Expression Networks from PubMed

Saric, Jensen, Ouzounova, Rojas & Bork EMBL, Heidelberg, Germany

Extracted a regulatory network of 441 pairwise relations from 59,694 PubMed abstracts about Brews’ Yeast.

Semantic accuracy of 83 – 90% for different roles (e.g., mx_prom, contain)

Extremely small problem!

Semantically shallow (no kinetics, modules, etc.); “… we decided against trying to extract how the regulation take place.”
Slide 14 Technology Development Tasks 2

Technology Development Tasks

- **Read** papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.

- **Assemble** causal fragments into more complete pathways; discover and resolve inconsistencies.

- **Explain** phenomena in signaling pathways. Answer questions, including “reaching down to data,” when it is available.

- **Integrate** reading, assembly and explanation in a non-pipeline architecture that provides flexible control.

Slide 15 The Assembly Task – Technology Challenges

The Assembly Task – Technology Challenges

Which entities “match up”? Are fragments semantically consistent? In which formal language is Big Mechanism to be represented?

**Fragment 1**

**Fragment 2**

- HBPI is a repressor of the cyclin D1 gene and inhibits the first signaling pathway.
- HBPI regulates the DMT1 promoter through sequence-specific binding (of the type TCAAT/CTATA) and the activity of HBPI itself is regulated through acetylation at any of 5 sites in the protein.
Slide 16 The Assembly Task – Ad Hoc Representation

http://www.nature.com/nbt/journal/v2.../nbt.1558.html

Slide 17 The Assembly Task – Previous Results

http://www.nature.com/nbt/journal/v2.../nbt.1558.html
Slide 18 Technology Development Tasks

- Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.
- Assemble causal fragments into more complete pathways; discover and resolve inconsistencies.
- Explain phenomena in signaling pathways. Answer questions, including “reaching down to data,” when it is available.
- Integrate reading, assembly and explanation in a non-pipeline architecture that provides flexible control.

Slide 19 The Explanation Task – Technology Challenges

Diverse reasoning methods – probabilistic, deductive, abductive, kinetic simulation, qualitative simulation – and data mining methods, all contribute to explanations; which are best, when, and what’s missing?

Examples:
- Can a contradiction in a causal model be resolved by a kinetic model?
- What are likely consequences of enabling or inhibiting a protein?
- Use data to evaluate pathway models
- Use data to create or modify pathway models

Slide 20 The Explanation Task – Previous Results

The Explanation Task – Previous Results

Detecting and Removing Inconsistencies between Experimental Data and Signaling Network Topologies Using Integer Linear Programming on Interaction Graphs.

Melas, Samaga, Alexopoulos, and Klamt. [We] predict the possible qualitative changes (up, down, no effect) of the activation levels of the nodes for a given stimulus. We detect and remove inconsistencies between measurements and predicted behavior. [We] detect interactions ... and provide suggestions for new interactions that, if included, would significantly improve the goodness of fit.


Slide 21 Technology Development Tasks

Technology Development Tasks

Read papers in cancer biology and extract causal fragments of signaling pathways, represented at all relevant semantic levels.

Assemble causal fragments into more complete pathways, discover and resolve inconsistencies.

Explain phenomena in signaling pathways. Answer questions, including “reaching down to data,” when it is available.

Integrate reading, assembly and exploration in a non-pipeline architecture that provides flexible control.
Slide 22 The Integration Task – Technology Challenges

The Integration Task – Technology Challenges

Reading, Assembly and Explanation shouldn’t be pipelined but should use each other opportunistically. Need flexible, non-pipelined control, plus significant software integration.

Slide 23 Conclusion

Conclusion

It’s a big problem

BIG DATA

Distinct communities to coordinate

statistical NLP
knowledge-based NLP
ontologies, databases
representation and reasoning
mathematical biology

BIG MECHANISM

Many domains

Potentially a new way to do science

Spotfire Dashboard

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App

Media, iframe, embed and object tags are not supported inside of a PDF.

Research Notes

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA

Updated: Sat, 19 Sep 2015 00:07:55 GMT

Powered by mindtouch
Meetup

I keep getting compliments for the Meetup and want to share them with you like: “Thanks again for a wonderful gathering of deep thinkers at the NIH-NSF Big Data event -- that was terrific. Great line up of speakers.”

My count showed that we had nearly 50 attend out of our 130 some members and some of them are asking me what they/we could do next with this information to write concept papers, do pilots, build reference implementations, etc. They include George Mason graduate students, entrepreneurs, data scientists, and Dr. Phil Bourne who asked us to work with Peter Lyster to set up a presentation/demo for him.

So I am asking you all for suggestions. Barend kept referring to Semantic Medline in his presentation and Tom and his staff will continue to develop that with the YarcData experts. We better understand what Barend is talking about from his terms: eScience approaches to ‘in silico’ knowledge discovery; eScience integrator; nanopublications as a substrate for in silico knowledge discovery; Concept Web Alliance, with ‘nanopublications’ as its first brainchild; Nanopublications are currently implemented in the semantic project of the Innovative Medicines Initiative (IMI) called Open PHACTS; and the Helicopter and Digger concepts. See his CSHALS 2014 presentation below.

We are also aware of an upcoming conference: The Semantics for e-science in an intelligent Big Data context, focused on this as well. See below.

So the question I and others have is how do we produce “Nanopublications and data papers”? Are their exemplary examples, standards, etc. or do we need to start experimenting with doing that in a sandbox like a wiki that supports HTML 5, APIs, semantic, etc.

The NIST Data Science Symposium this week had a poster on what looks like “nanopublications with semantics” and I submitted it as a suggested use case for our Meetup and others to work on as part of the new NIST Data Science Program. My poster was on the Semantic Medline-YarcData application and semantic data science mining work we have done with open government data.

Bottom line: How could we best help move the NIH-NSF Big Biomedical Data Initiative along and support the Data Fairport Unconference recommendations?

Semantics Based Biomedical Knowledge Search, Integration and Discovery

http://www.iscb.org/cshals2014-progr...4-keynote-mons

Abstract: Barend will talk about the role of semantic technologies, (under)standards and the nanopublication ecosystem in particular. He will challenge several established views in the field of the semantic web, scholarly communication, intellectual networking, science metrics, peer review and ‘data publishing’ with an emphasis on the barriers to break down in order to allow effective data exposure, sharing and integration in the Big Data era. The context of his talk will be the need for eScience approaches to ‘in silico’ knowledge discovery.

Biography: Barend Mons holds a chair in Biosemantics at the LUMC and is one of the scientific directors of NBIC. In addition he acts as a Life Sciences 'eScience integrator' in the Netherlands eScience centre. Currently, he coordinates the creation of the Data Integration and Stewardship Centre (DISC-ELIXIR) and in that capacity he is also the scientific representative of The Netherlands in the interim board of the ELIXIR ESFRI project.
Barend Mons is a molecular biologist by training and received his PhD on genetic differentiation of malaria parasites from Leiden University (1986). In 2000 he focused on the development of semantic technologies to manage big data and he founded the Biosemantics group. His research is currently focused on nanopublications as a substrate for in silico knowledge discovery. Barend is also one of the founders of the Concept Web Alliance, with 'nanopublications' as its first brainchild. Nanopublications are currently implemented in the semantic project of the Innovative Medicines Initiative (IMI) called Open PHACTS.

The Semantics for e-science in an intelligent Big Data context
http://sepublica.mywikipaper.org/drupal/
Scientific data, semantics for scientific data, e-scholarly, e-science, interoperability in scientific data, linked science, big data in science.

DARPA Memex

Source: https://www.fbo.gov/index?s=opportun...=core&_cview=0
Added: Feb 04, 2014 2:36 pm

The Defense Advanced Research Projects Agency (DARPA) is soliciting proposals for innovative research to maintain technological superiority in the area of content indexing and web search on the Internet. Proposed research should investigate approaches that enable revolutionary advances in science, devices, or systems. Specifically excluded is research that primarily results in evolutionary improvements to the existing state of practice.

The Memex program envisions a new paradigm, where one can quickly and thoroughly organize a subset of the Internet relevant to one’s interests. Memex will address the inherent shortcomings of centralized search by developing technology for domain-specific indexing of web content and domain-specific search capabilities. Memex will develop technology to enable discovery, organization, and presentation of domain relevant content. The new search paradigm will provide fast, flexible, and efficient access to domain-specific content as well as search interfaces that offer valuable insight into a domain that previously remained unexplored.

For further details see attached "DARPA-BAA-14-21" PDF document.

Conference on Semantics in Healthcare and Life Sciences

Source: http://www.iscb.org/cshals2014

Introduction

CSHALS is the premier annual event focused on the practical application of Semantic Web and other semantic technologies to problems in the Life Sciences, including pharmaceutical industry and related areas, such as hospitals/healthcare institutions and academic research labs.

CSHALS has for 7 years been the ISCB reference conference for cutting-edge practitioners of knowledge engineering and forward-thinking strategists in information and knowledge management in the life sciences. Topics covered by
CSHALS 2014 span the continuum between semantic standards development, data analytics, knowledge representation, information and knowledge management, knowledge engineering, and machine learning and reasoning.

The conference is accepting full paper and poster submissions in the topic areas listed below. **New for 2014**, following peer review, accepted papers will be published online in a volume of the CEUR Workshop Proceedings series [http://ceur-ws.org](http://ceur-ws.org).

### Topic Areas

- **Linked Data** - generation, representation, and management of data using the W3C’s Resource Description Framework (RDF) and Linked Data principles.
- **Data Modeling** - Ontologies, Taxonomies - practical application of knowledge engineering principles and techniques to data modeling.
- **Text Analysis, NLP, Question Answering** - use of semantic technologies for processing natural language.
- **Cloud, Parallel, and Distributed Computing** - use of semantic technologies for representation and management of distributed computational resources and workflows.
- **Healthcare eScience** - Translational Medicine, Pharmacogenomics, personal health records (PHR) - patient-centric use of semantic web technologies to handle the integration of clinical and biomolecular data.
- **Business Intelligence, Machine Learning, and Analytics** - semantic applications for data analysis, inference, and decision-making.
- **Human Computer Interfaces** - development of interfaces or design criteria that involve humans in the process of knowledge creation.
- **Clinical Applications** - Use of semantic web technology or architectures, including the orchestration of webApp ecosystems, as part of medical care delivery.
- **Molecular Applications** - Deployments of semantic web technology in biomolecular and biochemical domains.
- **Applications to Emerging Health Disciplines** - Use of the semantic web’s integrative and distributed nature to support emergence of novel applications in health care.

### Submissions

- **Research papers**: up to 12 pages that report on original research, and / or novel methodology that have led to significant outcomes with potential for lasting impact. Papers should demonstrate theoretical soundness of the research and/or strong evaluation.
- **Application papers**: up to 6 pages that report on original research with successful application demonstrated by an improvement over existing methods. Papers may include systems level approaches using established methodology in a novel application domain.
- **Presentation Abstracts**: Abstracts of up to 500 words reporting on ongoing projects that have met key project milestones but are not ready for fully documented disclosure. Submissions may also describe funded research proposals.
- **Posters**: Abstracts of 500 words reporting on early research results focusing on the design and application of semantic technologies in life science scenarios. Priority will be given to submissions offering a functional demo that can be shown at the conference.

Papers are to be prepared using [Springer's LNCS templates](http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA) and submitted via [EasyChair](http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA).  

- Spinger Author Guidelines ([pdf](http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA))
Keynote Speakers

Ronald Collette, CISSP

Updated Jan 28, 2014

Chief Information Officer
Foundation Medicine
Cambridge, MA, USA

Presentation Date and Time:
Thursday, February 27 - 9:00 am

Presentation Title:
The Carpenter to the Screw: Well, I Have a Hammer. You Must Be a Nail

Abstract:
Big Data is the buzz word du jour for the media. But this phrase only scratches the surface of the problem-space that the Industry is faced with today. The issues of Fast Data extreme cardinality, disparate data formats, the demand for data mining, are but a few of the challenges that call for new approaches. A host of technologies from RDF to NLP hold tremendous promise in addressing many of these issues. During the presentation Ron will discuss the evolution, adoption, and state of computer science within Industry from the perspective of an applied engineer; and address some of the limiting factors that are hindering the adoption of Semantic Technologies.

Biography:
From Punch Cards to Big Data, Mr. Collette has spent the better part of 3 decades working in the field of technology. During that time he has worked with in industries that encompass the various sectors aerospace, engineering, construction, finance, cyber security, life sciences, commercial food, and various entities within the federal government. This variety of experience, coupled with the innate need constantly push the envelope, has garnered him a reputation of solving difficult technical issues with practical solutions that blend mature and cutting-edge technologies.

Ron Collette is a regular speaker at a number of security and IT related events such as International Standards Organization (ISO) conference, SecureWorld Expo, ISMS, and InfoSeCon. He is also a regular
columnist and research analyst for Computer Economics and has co-authored two books on information security. *The CISO Handbook: A Practical Guide to Securing Your Company* and the companion publication *CISO Soft Skills: Securing Organizations Impaired by Employee Politics, Apathy and Intolerant Perspectives*. Both of these books are utilized as course material for a number of post-graduate programs on security leadership.

### Hilmar Lapp

*Updated March 03, 2014*

Assistant Director for Informatics  
National Evolutionary Synthesis Center (NESCen)  
Durham, North Carolina, USA

**Presentation Date and Time:**  
Friday, February 28 - 9:00 am

**Presentation Title:**  
*Semantics of and for the Diversity of Life: Opportunities and Perils of Trying to Reason on the Frontier*

[Click here for presentation slides](#)

**Abstract:**  
The stunning diversity of life has long been a target of research, resulting in billions of specimens and hundreds of millions of pages of text in the world’s natural history museum collections and libraries. A wealth of literature documents in meticulous the detail the features shared between contemporary and extinct species, their evolutionary history and their ecological properties. Enabling the full power of computation for this constantly growing, yet woefully incomplete body of knowledge faces enormous challenges, but also offers huge opportunities. These include discovery, such as candidate gene hypothesis generation, knowledge integration, such as using the tree of life as a map, and prediction, such as forecasting the effects of environmental change on biodiversity patterns and human health. I will present examples of both obstacles and opportunities for semantics-based technologies to enable computing on our knowledge of the diversity and evolution of life on earth.

**Biography:**  
Hilmar Lapp (@hlapp, [http://orcid.org/0000-0001-9107-0714](http://orcid.org/0000-0001-9107-0714)) is the Assistant Director for Informatics at the National Evolutionary Synthesis Center.
Center (NESCen). His research interests are in reusable and interoperable software and data, large-scale data integration, and in building open sustainable eScience infrastructure. A biologist by training, he has also more than two decades worth of experience developing software, databases, and other informatics resources. He is co-PI of the NSF-funded Phenoscape project (http://phenoscape.org) on ontological annotation of evolutionary phenotype observations, co-convener of the Biodiversity Information Standards (TDWG) Interest Group on Phylogenetics Standards, senior personnel in the Dryad digital data repository project (http://datadryad.org), and a participant in the Data Integration and Semantics Working Group of DataONE (http://dataone.org).

Deborah McGuinness, Ph.D

*Updated March 02, 2014*

Professor, Computer and Cognitive Science  
Founding Director, Rensselaer Polytechnic Institute’s Web Science Research Center  
Rensselaer Polytechnic Institute (RPI)  
Troy, NY, USA

**Presentation Date and Time:**  
Thursday, February 27 - 1:00 pm

**Presentation Title:**  
*Towards Semantic Health Assistants*

- Click here for PDF of presentation. [PDF](#)

**Abstract:** In this presentation we will present a vision of semantically-enabled health assistants. We will present three examples of semantics in practice in the areas of drug repurposing, mobile health advisors, and population health. We will highlight where semantic technologies are adding value and also present some remaining challenges.

**Biography:** Deborah McGuinness is the Tetherless World Senior Constellation Chair and Professor of Computer and Cognitive Science. She is also the founding director of the Web Science Research Center at Rensselaer Polytechnic Institute. Deborah has been recognized with awards for leadership in Semantic Web research and in bridging Artificial Intelligence (AI) and eScience, significant contributions to
deployed AI applications, and extensive service to the AI community. Deborah is a leading authority on the semantic web and has been working in knowledge representation and reasoning environments for over 25 years. Deborah's primary research thrusts include work on explanation, trust, ontologies, eScience, open data, and semantically-enabled schema and data integration. Prior to joining RPI, Deborah was the acting director of the Knowledge Systems, Artificial Intelligence Laboratory and Senior Research Scientist in the Computer Science Department of Stanford University, and previous to that she was at AT&T Bell Laboratories.

Deborah is also widely known for her leading role in the development of the W3C Recommended Web Ontology Language (OWL), her work on earlier description logic languages and environments, including the CLASSIC knowledge representation system, and more recently, her work on provenance languages and environments, including InferenceWeb, PML, and PROV. She has built and deployed numerous ontology environments and ontology-enhanced applications, including some that have been in continuous use for over a decade at AT&T and Lucent, and two that have won deployment awards for variation reduction on plant floors and interdisciplinary virtual observatories. Recent application thrusts include health informatics and smart environmental monitoring. Some current and recent projects can be explored at [http://tw.rpi.edu/web/person/Deborah_L_McGuinness/Projects](http://tw.rpi.edu/web/person/Deborah_L_McGuinness/Projects). She has published over 200 peer-reviewed papers and has authored granted patents in knowledge based systems, ontology environments, configuration, and search technology.

Deborah also consults with clients wishing to plan, develop, deploy, and maintain semantic web and/or AI applications. Some areas of recent work include: data science, next generation health advisors, ontology design and evolution environments, semantically-enabled virtual observatories, semantic integration of scientific data, context-aware mobile applications, search, eCommerce, configuration, and supply chain management. She is a frequent technology advisory board member, currently with Qualcomm, OrbMedia, SocialWire, and Sandpiper Software. She also advised Applied Semantics, Guru Worldwide, and Cerebra prior to their acquisitions. Deborah has also worked as an expert witness in a number of cases, and has deposition and trial experience. Deborah received her Bachelors degree in Math and Computer Science from Duke University, her Masters degree in Computer Science from University of California at Berkeley, and her Ph.D. in Computer Science from Rutgers University.

Prof. Dr. B. (Barend) Mons

*Updated Jan 28, 2014*
Leiden University Medical Center (LUMC) and Netherlands Bioinformatics Center (NBIC), The Netherlands.

Presentation Date and Time:
Friday, February 28 - 1:00 pm

Presentation Title:
Semantics Based Biomedical Knowledge Search, Integration and Discovery

Abstract: Barend will talk about the role of semantic technologies, (under)standards and the nanopublication ecosystem in particular. He will challenge several established views in the field of the semantic web, scholarly communication, intellectual networking, science metrics, peer review and 'data publishing' with an emphasis on the barriers to break down in order to allow effective data exposure, sharing and integration in the Big Data era. The context of his talk will be the need for eScience approaches to 'in silico' knowledge discovery.

Biography: Barend Mons holds a chair in Biosemantics at the LUMC and is one of the scientific directors of NBIC. In addition he acts as a Life Sciences 'eScience integrator' in the Netherlands eScience centre. Currently, he coordinates the creation of the Data Integration and Stewardship Centre (DISC-ELIXIR) and in that capacity he is also the scientific representative of The Netherlands in the interim board of the ELIXIR ESFRI project.

Barend Mons is a molecular biologist by training and received his PhD on genetic differentiation of malaria parasites from Leiden University (1986). In 2000 he focused on the development of semantic technologies to manage big data and he founded the Biosemantics group. His research is currently focused on nanopublications as a substrate for in silico knowledge discovery. Barend is also one of the founders of the Concept Web Alliance, with 'nanopublications' as its first brainchild. Nanopublications are currently implemented in the semantic project of the Innovative Medicines Initiative (IMI) called Open PHACTS.

Tech Talks

Source: http://www.iscb.org/cshals2014-progr...014-tech-talks

Updated Feb 27, 2014
Tech Talks showcase products and services of relevance to the CSHALS audience. Each Tech Talk is 20 minutes (15 minutes presentation and up to 5 minutes for Q&A) in length and designed to allow organizations to create awareness of new technologies, services, etc., in an informational presentation format.

High Availability and Graph Mining using bigdata® - an Open-Source Graph Database

- Click here for PDF of presentation. PDF

Presenter: Bryan Thompson, SYSTAP LLC., Greensboro, NC United States

Abstract: bigdata(R) is a high performance, scalable, open source graph database platform supporting the RDF data model and edge attributes. I will provide a brief overview of the bigdata platform, summarize some of its key differentiators, including the High-Availability enterprise deployment model, an API for writing graph mining algorithms against RDF data, and efficient representation and query of edge attributes, discuss approaches for combining bigdata clusters with map/reduce processing, and provide a glimpse of new features in our roadmap, including accelerated graph processing on GPUs at 3-billion edges per second.

SYSTAP, LLC leads the development of the bigdata open source platform and offers consulting services related scalable information architectures and services and support for the bigdata platform. Bigdata is available under both open source and commercial licenses.

- http://www.systap.com
- http://www.bigdata.com/blog
- http://sourceforge.net/projects/bigdata/

Biography: Bryan Thompson (SYSTAP, LLC) is the Chief Scientist and co-Founder of SYSTAP, LLC. He is the lead architect for bigdata®, an open source graph database used by Fortune 500 companies including EMC (SYSTAP provides the graph engine for the topology server used in their host and storage management solutions) and Autodesk (SYSTAP provides their cloud solution for graph search). He is the principle investigator for a DARPA research team investigating GPU-accelerated distributed architectures for graph databases and graph mining. He has over 30 years experience related to cloud computing; graph databases; the semantic web; web architecture; relational, object, and RDF database architectures; knowledge management and collaboration; artificial intelligence and connectionist models; natural language processing; metrics, scalability studies, benchmarks and performance tuning; decision support systems.
Successfully Navigating Diagnosis And Treatment In The Age Of Targeted Cancer Therapy

Presenter: Helena Deus from Foundation Medicine, Cambridge, United States

Abstract: Initially, cancer was understood in simple terms as a disease caused by uncontrolled division of abnormal cells which formed tumors that could "metastasize" to invade other tissues and organs in the body. We knew that certain factors played a role in the cause of cancer like environment, lifestyle, genetic and carcinogens. But we didn't have a clear idea on how cancer was caused or what drove it.

The traditional paradigm for solid tumor treatment was:
1. To visually examine a tumor biopsy under a microscope to determine if cancer is present, the type of cancer and how advanced or what stage the cancer is in
2. Traditional cancer treatments include Surgery, Radiation and Traditional Chemotherapy
3. These are still the mainstay of cancer treatment and while they are highly effective in some patients, they are less effective in others and can cause significant side effects

Today, we know that cancer is a disease of the genome. We understand that cancer is a result of the carcinogens, and other factors causing changes, or alterations to our DNA. And we know that increase exposure to these factors increases the chance of gene alterations.

In this talk, I will explain the paradigm and complexities of Targeted Cancer Therapy from an IT and knowledge management perspective and how we are addressing those at Foundation Medicine.

Biography: Dr. Deus is a Senior Scientist at Foundation Medicine. Since 2005 she has worked with oncologists and scientists in finding and developing the best strategies to improve the quality of bioinformatics and translational research solutions. Her main area of expertise is in applying RDF/SPARQL-based technologies to solving difficult data problems in biological domains. With her work on RDFizing the Cancer Genome Atlas and linking it to pubmed, her team won her the 2013 Linked Data Cup (at I-Semantics) and the Semantic Web Challenge - Big Data Track (at ISWC2013).

Making Sense of Big Data in Pharma

Presenter: Andreas Matern, Vice President, Disruptive Innovation, Thomson Reuters Life Sciences, MA, United States

Abstract: Big Data is quickly becoming an overused, and poorly understood, term in technology. This talk will focus on
Big Data for the life sciences: are -omics data the only 'big data'? What's a practical working definition for Big Data in the Life Sciences and does it differ from other areas where data is analyzed at scale? What role does visualization have in Big Data? How do we resolve gaps in life sciences data? How can we spot trends utilizing aggregated information from disparate data sources, and can we effectively ask questions and monitor the ever growing amount of structured and unstructured content that we have access to?

Biography: Andreas Matern is the Vice President, Disruptive Innovation for Thomson Reuters Life Sciences. He and his team devote their efforts to the intersection of content and technology - finding new ways to explore how end users can, without assistance from experts, explore information and find the answers to the questions they are most interested in. Andreas' background is in molecular genetics, but he has spent the entirety of his professional career involved in bioinformatics, where he has worked for over a decade. He has worked at LION Bioscience, Teranode and InforSense before coming to work at Thomson Reuters.

Using Supercomputer Architecture for Practical Semantic Web Research in the Life Sciences

Presenter: Matt Gianni, Cambridge, MA, United States

Abstract: YarcData delivers on the promise of big data analytics and semantic technologies by enabling research organizations to discover unknown relationships and hidden patterns in their structured and unstructured data. Within Life Sciences and Healthcare, YarcData can provide solutions to advance research in critical disease areas, improve the ability to rescue drug candidates, and find new targets for existing drugs.

We will present a use case by which a major cancer research organization is taking a systems biology approach to understanding the development of cancers. By rapidly bringing together data from Medline, TCGA, Pfam, UniProt, and other sources, the organization could model the growth of cancer cells at the molecular level.

All of this research was augmented into one body of interrelated, queryable knowledge, from which researchers could fast-track testing using in-silico hypothesis verification. This resulted in a clear visualization of the connections and associations within the data to help identify promising candidates for new therapies. The solution also overcame typical technical challenges, such as dealing with dynamic datasets, limited performance, and scalability issues as seen with relational data warehouses.

YarcData’s innovative, real-time approach to cancer research discovery allows researchers to identify new patterns in the data without any upfront modeling or advance knowledge of the data relationships. When compared to traditional database strategies and investigative laboratory experiments, the time to discovery was significantly reduced, saving months or years of research with a higher probability of success.

Biography: Matt Gianni is a Life Science Solution Architect at YarcData. Matt is responsible for representing
YarcData’s solutions from a technical and scientific perspective within the Life Science markets. Matt joins us from IDBS where he focused on delivering data management and software solutions in Life Science, with a particular focus on translational research. Over the past 15 years, Matt focused on accelerating drug discovery using computational technologies and has held key technical roles with Elsevier, Symyx, MDL and Exelixis prior to joining YarcData.

Presenters

Source: http://www.iscb.org/cshals2014-progr...014-presenters

Scott Bahlavooni

Towards Semantic Interoperability of the CDISC Foundational Standards

Click here for PDF of presentation PDF BahlavooniCSHALS2014.pdf

Scott Bahlavooni, Biogen Idec, United States
Geoff Low, Medidata Solutions, United Kingdom
Frederik Malfait, IMOS Consulting, Switzerland

The Clinical Data Interchange Standards Consortium (CDISC) mission is "to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare." The CDISC Foundational Standards support the clinical research life cycle from protocol development through analysis and reporting to regulatory agencies. Unfortunately, the majority of the Foundational Standards utilized by pharmaceutical, biologic, and device development companies are published in formats that are not machine-processable (e.g., Microsoft Excel, PDF, etc.). The PhUSE Computational Science Symposium, Semantic Technology (ST) project in consultation with CDISC has been working on representing the CDISC Foundational Standards in RDF based on an ISO 11179-type metamodel. Draft RDF representations are available for five of the CDISC Foundational Standards - CDASH, SDTM, SEND, ADaM, and controlled terminology - and have been published on GitHub. In Q1, 2014, these representations will undergo CDISC public review to facilitate their adoption as a CDISC standard and a standard representation of the CDISC Foundational Standards. Additional activities are ongoing to create RDF representations of conformance checks, the CDISC Protocol Representation Model (PRM), and the Study Design Model (SDS-XML). Further activities are planned around analysis metadata, linking to EHR, and representing clinical trial data in RDF. The presentation will provide an overview of the different models, their relationships to one another, and how they can be used to manage data standards in an ISO 11179-type metadata registry. It explains how to represent CDISC Foundational Standards in a machine readable format that enables full semantic interoperability. The presentation will also highlight how this work can facilitate a planned RDF export format for the CDISC SHARE environment, a metadata registry currently being developed by CDISC to develop and manage CDISC standards in a consistent way.

Evan Bolton

PubChemRDF: Towards a semantic description of PubChem

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA

Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch
PubChem is a community driven chemical biology resource containing information about the biological activities of small molecules located at the National Center for Biotechnology Information (NCBI). With over 250 contributors, PubChem is a sizeable resource with over 125 million chemical substance descriptions, 48 million unique small molecules, and 220 million biological activity results. PubChem integrates this information with other NCBI internal resources (such as PubMed, Gene, Taxonomy, BioSystems, OMIM, and MeSH) but also external resources (such as KEGG, DrugBank, and patent documents).

Semantic Web technologies are emerging as an increasingly important approach to distribute and integrate scientific data. These technologies include the Resource Description Framework (RDF). RDF is a family of World Wide Web Consortium (W3C) specifications used as a general method for concept description. The RDF data model can encode semantic descriptions in so-called triples (subject-predicate-object). For example, in the phrase “atorvastatin may treat hypercholesterolemia,” the subject is “atorvastatin,” the predicate is “may treat,” and the object is “hypercholesterolemia.”

The PubChemRDF project provides a great wealth of PubChem information in RDF format. One of the aims is to help researchers work with PubChem data on local computing resources using semantic web technologies. Another aim is to harness ontological frameworks to help facilitate PubChem data sharing, analysis, and integration with resources external to NCBI and across scientific domains.

This talk will give an overview of the recently released PubChemRDF project scope and some examples of its use.

Aftab Iqbal

GenomeSnip: Fragmenting the Genomic Wheel to Augment Discovery in Cancer Research

Awarded Best Paper

In recognition of outstanding work in the science of semantic technologies applied in health and life sciences, the CSHALS Organizing Committee is pleased to announce awards for Outstanding Papers and Posters at CSHALS2014. The most important research results presented at the conference, as determined by the Program Committee, will be recognized in the following way:

- A cash prize of $250 will awarded by the International Society of Computation Biology for the most outstanding student poster.
- Software licences for Sentient Knowledge Explorer PRO will be awarded by conference sponsors IO Informatics to authors of the most outstanding paper.

Click here for PDF of presentation PDF

Maulik R. Kamdar, National University of Ireland, Ireland

Aftab Iqbal, National University of Ireland, Ireland
Cancer genomics researchers have greatly benefitted from high-throughput technologies for the characterization of genomic alterations in patients. These voluminous genomics datasets when supplemented with the appropriate computational tools have led towards the identification of 'oncogenes' and cancer pathways. However, if a researcher wishes to exploit the datasets in conjunction with this extracted knowledge his cognitive abilities need to be augmented through advanced visualizations. In this paper, we present GenomeSnip, a visual analytics platform, which facilitates the intuitive exploration of the human genome and displays the relationships between different genomic features. Knowledge, pertaining to the hierarchical categorization of the human genome, oncogenes and abstract, co-occurring relations, has been retrieved from multiple data sources and transformed a priori. We demonstrate how cancer experts could use this platform to interactively isolate genes or relations of interest and perform a comparative analysis on the 20.4 billion triples Linked Cancer Genome Atlas (TCGA) datasets.

Andrew M. Jenkinson
The RDF Platform of the European Bioinformatics Institute

The breadth of diversity of data in the life sciences that is available to support research is a valuable asset. Integrating these complex and disparate data does however present a challenge. The Resource Description Framework (RDF) technology stack offers a mechanism for storing, integrating and querying across such data in a flexible and semantically accurate manner, and is increasingly being used in this domain. However the technology is still maturing and has a steep learning curve. As a major provider of bioinformatics data and services, the European Bioinformatics Institute
Institute (EBI) is committed to making data readily accessible to the community in ways that meet existing demand. In order to support further adoption of RDF for molecular biology, the EBI RDF platform (https://www.ebi.ac.uk/rdf) has therefore been developed. In addition to coordinating RDF activities across the institute, the platform provides a new entry point to querying and exploring integrated resources available at the EBI.

Bernadette Hyland

Improving Scientific Information Sharing by Fostering Reuse of Presentation Material

Bernadette Hyland, 3 Round Stones Inc., United States
David Wood, 3 Round Stones Inc., United States
Luke Ruth, 3 Round Stones Inc., United States
James Leigh, 3 Round Stones Inc., United States

Most scientific developments are recorded in published papers and communicated via presentations. Scientific findings are presented within organizations, at conferences, via Webinars and other fora. Yet after delivery to an audience, important information is often left to wither on hard drives, document management systems and even the Web. Accessing data underlying scientific findings has been the Achilles Heel of researchers due to closed and proprietary systems. Because there is no comprehensive ecosystem for scientific findings, important research and discovery is repeatedly performed and communicated over and over. In an ideal world, published papers and presentations would be the start of an ongoing dialogue with peers and colleagues. By definition this dialogue spans geographical boundaries and therefore must involve a global network, international data exchange standards, a universal address scheme and provision for open annotations. Security and privacy are key concerns in discussing and sharing scientific research. 3 Round Stones has created a collaboration system that is platform-agnostic relying on the architecture and standards of the Web using the Open Source Callimachus Linked Data Management system. The collaboration system facilitates the full ecosystem for sharing and annotating scientific findings. It includes modern social network capabilities (e.g., bio, links to social media sites, contact information), Open Annotation support for associating distinct pieces of information, and metrics on content usage. Papers and presentations can be annotated and iterated using a familiar review cycle involving authors, editors and peer reviewers. Proper attribution for content is handled automatically. The 3 Round Stones collaboration system has the potential to foster discovery, access and reuse of scientific findings.

Hyeongsik Kim

Scalable Ontological Query Processing over Semantically Integrated Life Science Datasets using MapReduce

Hyeongsik Kim, North Carolina State University, United States
Kemafor Anyanwu, North Carolina State University, United States

While addressing the challenges of join-intensive Semantic Web workloads has been a key research focus, processing disjunctive queries has not. Disjunctive queries arise frequently in scenarios when querying heterogeneous datasets that have been integrated using rich and complex ontologies such as in biological data warehouses, UniProt, Bio2RDF, and...
Chem2Bio2RDF, etc. Here, the same or similar concepts may be modeled in different ways in the different datasets. Therefore, several disjunctions are often required in queries to describe the different related expressions relevant for a query, either included explicitly or generated as part of an inferencing-based query expansion process. Often, the width (#branches) of such disjunctive queries can be large making them expensive to evaluate. They pose particularly significant challenges when cloud platforms based on MapReduce such as Apache Hive and Pig are used to scale up processing, translating to long execution workflows with large amount of I/O, sorting and network traffic costs. This paper presents an algebraic interpretation of such queries that produces query rewritings that are more amenable to efficient processing on MapReduce platforms than traditional relational query rewritings.

David King

Evolution from Data-analytics to Data-applications: How Modular Software Environments Built on Semantic Linked-Data Enable a New Generation of Collaborative Life-Science Applications

David King, Exaptive Inc., United States
Robert McBurney, The Accelerated Cure Project, United States

This talk proposes a new approach to solving one of the paramount challenges of gaining valuable insight from complex data environments like those found in the life sciences – the integration of not only distributed and disparate data, but also the integration of diverse, specialized, and often siloed scientists and their analytical tools. Traditional tightly-coupled architectures, relying on data-warehousing and monolithic software, are ill-equipped to handle a diverse and dynamic landscape of data, users and tools. Fragmented data cannot be leveraged by equally fragmented software. A successful approach must dynamically link scientists and their data and facilitate their cross-disciplinary collaboration. This talk will explore both the technical details for implementing architectures that accomplish this though linked-data and linked-visualization, and will show how this approach is being leveraged by The Accelerated Cure Project and the Orion Bionetworks Alliance in their efforts to lower the barriers to collaborative Multiple Sclerosis research. The combination of Object-Oriented and Flow-Based programming with Semantic Data standards and recent HTML5 Web advances creates the potential for a new breed of life-science analytics applications. Applications built from this style of combinatorial architecture achieve a loosely-coupled modularity that allows them to leverage the wide variety of components present in complex data systems instead of being hindered by it. By treating datasets more modularly, these applications give liquidity to otherwise frozen and siloed data. By treating algorithms and visualizations as data-agnostic modules, these applications lower the barriers to creative experimentation, a key component to hypothesis generation and insight. By leveraging recent advances in HTML5 Web capabilities, these techniques can launch in web-based, highly cohesive exploratory interfaces that can be rapidly deployed, evolved, and redeployed to a distributed and dynamic workforce of collaborating scientists, analysts, and subject matter experts. The technical details of this talk will be balanced by two real-world use-cases: how it is being applied by The Accelerated Cure Project as a next-generation interface to their unique biosample repository, and how it is being applied by the Orion Bionetworks Alliance as a means to link and visualize data from multiple large-scale clinical studies. A goal of the talk will be to strike a balance between the macro and micro views of modular data-application development to provide attendees with two sets of useful takeaways. At the macro level, the audience will be exposed to a generic architecture for combinatorial interactive interface development that can inform some of their future design decisions. At the micro level, the audience will be introduced to some specific linked-data and visualization techniques - exploratory interfaces that are both thought-provoking and inspiring for the attendees’ future work.
David H. Mason

*Multi-Domain Collaboration for Web-Based Literature Browsing and Curation*

David H. Mason, Concordia University, Canada
Marie-Jean Meurs, Concordia University, Canada
Erin McDonnell, Concordia University, Canada
Ingo Morgenstern, Concordia University, Canada
Carol Nyaga, Concordia University, Canada
Vahé Chahinian, Concordia University, Canada
Greg Butler, Concordia University, Canada
Adrian Tsang, Concordia University, Canada

We present Proxiris, a web-based tool developed at the Centre for Structural and Functional Genomics, Concordia University. Proxiris is an Open Source, easily extensible annotation system that supports teams of researchers in literature curation on the Web via a browser proxy. The most important Proxiris features are iterative annotation refinement using stored documents, Web scraping, strong search capabilities, and a team approach that includes specialized software agents. Proxiris is designed in a modular way, using a collection of Free and Open Source components best suited to each task.

Chimezie Ogbuji

*Lattices for Representing and Analyzing Organogenesis*

Chimezie Ogbuji, Metacognition LLC, United States
Rong Xu, Case Western University, United States

A systems-based understanding of the molecular processes and interactions that drive the development of the human heart, and other organs, is an important component of improving the treatment of common congenital defects. In this work, we present an application of Formal Concept Analysis (FCA) on molecular interaction networks obtained from the integration of two central biomedical ontologies (the Gene Ontology and Foundational Model of Anatomy) and a subset of a cross-species anatomy ontology. We compare the formal concept lattice produced by our method against a cardiac developmental (CD) protein database to verify the biological significance of the structure of the resulting lattice. Our method provides a unique and unexplored computational framework for understanding the molecular processes underlying human anatomy development.

Tom Plasterer

*The Open PHACTS Project: Progress and Future Sustainability*
Tom Plasterer, AstraZeneca, United States
Lee Harland, Connected Discovery, United Kingdom
Bryn Williams-Jones, Connected Discovery, United Kingdom

The Open PHACTS project, an Innovative Medicines Initiative funded effort, has delivered a platform designed to reduce barriers to drug discovery and accelerate collaboration within and outside of biopharma. It has done so by building a community of life science and computational researchers around a use-case driven model, taking advantage of existing community efforts while pushing life science content providers towards the use and provisioning of linked data. The project is pursuing an ENSO (explore new scientific opportunities) call to expand the available content set and look at new distribution models. It is also looking to complete the project in the fall of 2014 and to turn these successes into a self-sustaining foundation, the Open PHACTS Foundation (OPF). We’ll provide a travelogue of this journey along with a look into some of the major contributions, from the Open PHACTS API to the exemplar applications built on top of this platform. From there a snapshot of the impact of Open PHACTS at AstraZeneca will be discussed, from the vantage point of characterising internal and external biochemistry assays using the BioAssay Ontology and Open PHACTS assay content.

Eric Prud’Hommeaux

Eric Prud’Hommeaux, W3C, United States
Charles Mead, W3C, France
Sajjad Hussain, INSERM U1142, France

In December 2012, the FDA solicited advice on how to address major challenges in cross-study data integration. A significant part of the feedback they received was to leverage Semantic Web technologies to capture embedded context for machine processing. They combined this with an already existing effort to standardize a set of around 60 "Therapeutic Areas" (TAs) -- attributes commonly captured for submissions for a particular disease.

In the fall of 2013, FDA asked IBM to provide ontologies for 12 TAs. Initial work on a Renal Transplantation TA has provided a shared ontologies for all TAs grounded in BRIDG, as well as set of emerging practices for efficiently extracting expertise from Subject Matter Experts (SMEs). The resulting test-driven development approach could establish practice for the production of future FDA TAs and, ideally, study design for the long tail of trials not covered by current TAs.

Further, the development of such a corpus of ontologies and the trivial expression of submission data as RDF can effectively create a standard for a clinical study metadata and data repository, subsuming both CDISC SHARE and ODM-based repositories.

Arash Shaban-Nejad

A Semantic-driven Surveillance Model to Enhance Population Health Decision-making
The Population Health Record (PopHR) is a web-based software infrastructure that retrieves and integrates heterogeneous data from multiple sources (administrative, clinical, and survey) in near-real time and supports intelligent analysis and visualization of these data to create a coherent portrait of population health through a variety of indicators. Focused on improving population health decision-making, PopHR addresses common flaws of existing web portals with similar functionality, such as the lack of structure in presenting available indicators, insufficient transparency of computational algorithms and underlying data sources, overly complicated user interface, and poor support for linking different indicators together to draw conclusions about population health. PopHR presents population health indicators in a meaningful way, generates results that are contextualized by public health knowledge, and interacts with the user through a simple and intuitive natural language interface.

Robert Stanley

Safety Information Evaluation and Visual Exploration ("SIEVE")

AstraZeneca ("AZ") Patient Safety Science wanted to improve retrieval of clinical trial data and biometric assessments across studies. Traditionally, evaluation of clinical trials data across studies required manual intervention to deliver desired datasets. A proposal titled Safety Information Evaluation and Visual Exploration ("SIEVE") was sponsored by Patient Safety Science. This took the form of collaboration between AZ and IO Informatics ("IO"). AZ provided the project environment, data resources, subject matter expertise ("SME") and business expertise. IO provided semantic software, data modeling and integration services including solutions architecture, knowledge engineering and software engineering. The project goal was to improve search and retrieve of clinical trials data. SIEVE was to provide a web-based environment suitable for cross-study analysis. The environment was to align across biomarkers, statistics and bioinformatics groups. Over-arching goals included decision-making for biomarker qualification, trial design, concomitant medication analysis and translational medicine. The team analyzed approximately 42,000 trials records, identified by unique subjectIDs. IO’s Knowledge Explorer software was used by IO’s knowledge engineers in collaboration with AZ’s SMEs to explore the content of these records as linked RDF networks. Reference metadata such as studyID, subjectID,
rowID, gender and DoB was central to assuring valid integration. Because almost all docs had both, subjectID and studyID, concatenation of these items as an individual identifier allowed connections that bridged multiple documents for data traversal. 36,000 records contained valid data, each including a unique trial, patient, and at least one row of laboratory data. IO created a semantic data model or “application ontology” to meet SIEVE requirements. The resulting data model and instances were harmonized by application of SPARQL-based rules and inference and were aligned with AZ standards. Data was integrated under this ontology, loaded into a semantic database and connected to IO’s “Web Query” software. The result is a web-based User Interface accessible to end users for cross-study searching, reporting, charting and sub-querying. Methods include “Quick Search” options, shared searches and query building containing nesting, inclusion / exclusion, ranges, etc. Advanced Queries are presented as filters for user entry to search subjects (“views” or “facets”) including Clinical Assays, Therapy Areas, Adverse Events and Subject Demographics. Reports include exporting, charting, hyperlink mapping and results-list based searches. Results include reduced time to evaluate data from clinical trials and to facilitate forward looking decisions relevant to portfolios. Alternatives are less efficient. Trial data could previously be evaluated within a study; however there was no method to evaluate trials data across studies without manual intervention. Semantic technologies applied for data description, manipulation and linking provided mission-critical value. This was particularly apparent for integration and harmonization, in light of differences discovered across resources. IO’s Knowledge Explorer applied data visualization and manipulation, application of inference and SPARQL-based rules to RDF creation. This resulted in efficient data modeling, transformation, harmonization and integration and helped assure a successful project.

Poster Presentations

Source: http://www.iscb.org/cshals2014-progr...-presentations

Poster 01 A Semantic Computing Platform to Enable Translational and Clinical Omics

- Click here for PDF of poster PDF

Jonathan Hirsch, Syapse, United States
Andro Hsu, Syapse, United States
Tony Loeser, Syapse, United States

To bring omics data from benchtop to point of care, labs and clinics must be able to handle three types of data with very different properties and requirements. The first, biomedical knowledge, is highly complex, continually evolving, and comprises millions of concepts and relationships. The second, medical data including clinical health and outcomes records, is temporal, unstructured, and hard to access. The third, omics data such as whole-genome sequence, is structured but voluminous. Attempts to bridge the three have had limited success. No single data architecture allows efficient querying of these types of data. The lack of scalable infrastructure that can integrate complex biomedical knowledge, temporal medical data, and omics data is a barrier to widespread use of omics data in clinical decisionmaking. Syapse has built a data platform that enables capture, modeling, and query of all three data types, along with applications and programming interfaces for seamless integration with lab and clinical workflows. Using a proprietary, semantic layer based on Resource Description Framework (RDF) and related standards, the Syapse platform enables assembly of a dynamic knowledgebase from biomedical ontologies such as SNOMED CT and OMIM as well as customers’ internally curated knowledge. Similarly, HL7-formatted medical data can be imported and
represented as RDF objects. Lastly, Syapse enables federated queries that associate RDF-represented knowledge with omics data while retaining the benefits of scalable storage and indexing techniques. Biologists and clinicians can access the platform through a rich web application layer that enables role-specific customization at any point in the clinical omics workflow. We will describe how biologists and clinicians use Syapse as the infrastructure of an omics learning healthcare system. Clinical R&D performs data mining queries, e.g. selecting patients who share disease and treatment characteristics to identify associations between omics profiles and clinical outcomes. Organizations update clinical evidence such as variant interpretation or pharmacogenetic associations in a knowledgebase that triggers alerts in affected patient records. At point of care, clinical decision support interfaces present internal or external treatment guidelines based on patients’ omics profiles. The Syapse platform is a cloud-based solution that allows labs and clinics to deliver translational and clinical omics with minimal IT resources.

**Poster 02 Optimization of RDF Data for Graph-based Applications**

Ryota Yamanaka, The University of Tokyo, Japan  
Tazro Ohta, Database Center for Life Science, Japan  
Hiroyuki Aburatani, The University of Tokyo, Japan

Various omics datasets are publicly available in RDF via their data repositories or endpoints and this makes it easier to obtain integratable datasets from different data sources. Meanwhile, when we use parts of linked data in our own applications for data analyses or visualization, the data format does not have to be RDF but can be processed into appropriate formats according to usage. In fact, most web applications handle table format data rather than RDF.

Application developers can think of two reasons to convert semantic networks into other data models rather than keeping original RDF in backend databases. One reason is the difficulty of understanding complex RDF schema and writing SPARQL queries, and another reason is that the data model described in RDF is not always optimized for search performance. Consequently, we need practical methods to convert RDF into the data model optimized for each application in order to build an efficient database using parts of linked data.

The simplest method of optimizing RDF data for most applications is to convert it into table format data and store it in relational databases. In this method, however, we need to consider not only table definition but also de-normalization and indices to reduce the cost of table-join operations. As a result, we are focusing on graph databases instead. In graph databases, their data models can naturally describe semantic networks and enable network search operations such as traversal as well as in triplestores.

Although the data models in graph databases are similar in structure to RDF-based semantic networks, they are different in some aspects. For example, in the graph database management system we used in this project, Neo4j, relationships can hold properties, while edges in RDF-based semantic networks do not have properties. We are therefore researching how to fit RDF data to effectively use graph database features for better search performance and more efficient application development.

We have developed tools to convert RDF data (as well as table format data) and loaded sample data into graph database. Currently, we are developing demo applications to search and visualize graph data such as pathway networks. These resources are available at [http://sem4j.org/](http://sem4j.org/)
**Poster 03 A Phenome-Guided Drug Repositioning through a Latent Variable Model**

Halil Bisgin, National Center for Toxicological Research, United States  
Zhichao Liu, National Center for Toxicological Research, United States  
Hong Fang, National Center for Toxicological Research, United States  
Reagan Kelly, National Center for Toxicological Research, United States  
Xiaowei Xu, University of Arkansas at Little Rock, United States  
Weida Tong, National Center for Toxicological Research, United States

The phenome has been widely studied to find the overlaps with genome, which in turn, has identified correlations for diseases. Besides its explanatory power for causalities, the phenome has been also explored for drug repositioning, which is a process of identifying new uses of existing drugs. However, most of current phenome-based approaches limit the search space for candidate drugs with the most similar drug. For a comprehensive analysis of the phenome, we assumed that all phenotypes (indications and side effects) were generated with a probabilistic distribution that can provide the likelihood of new therapeutic indications for a given drug. We employed Latent Dirichlet Allocation (LDA), which introduces latent variables (topics) that were assumed to govern phenome distribution. Hence, we developed our model on the phenome information in Side Effect Resource (SIDER). We first examined the recovery potential of LDA by perturbating the drug-phenotype matrix for each of the 11,183 drug-indication pairs. Those indications were assumed to be unknown and tried to be recovered based on the remaining drug-phenotype pairs. We were able to recover known indications masked during the model development phase with a 70% success rate on a portion of drug-indication pairs (5516 out of 11183) that have probabilities greater than the random chance (p>0.005). After obtaining such a decision criterion that considers both probability and rank, we applied the model on the whole phenome to suggest alternative indications. We were able to retrieve FDA-approved indications of 6 drugs whose indications were not listed in SIDER. For 907 drugs that are present with their indication information, our model suggested at least one alternative treatment option for further investigations. Several of the suggested new uses can be supported with information from the scientific literature. The results demonstrated that the phenome can be further analyzed by a generative model, which can discover the probabilistic associations between drugs and therapeutic uses. In this regard, LDA stands as a promising tool to explore new uses of existing drugs by narrowing down the search space.

**Poster 04 Cross-Linking DOI Author URIs Using Research Networking Systems**

- Click here for PDF of poster [PDF](#)

Nick Benik, Harvard Medical School, United States  
Timothy Lebo, Rensselaer Polytechnic Institute, United States  
Griffin Weber, Harvard Medical School, United States

A proof-of-concept application was created to automatically cross-link publications that were written by the same person by harvesting linked open data from institution-based research networking systems. This is important because it (1) helps people identify related articles when exploring the biomedical literature, (2) gives scientists appropriate credit for the work they have done, and (3) makes it easier to find experts in a subject area. Our presentation will include a demo of an interactive network visualization that allows exploration of these newly created links.
Poster 05 Simplified Web 3.0 Query Portal for Biomedical Research

Justin Lancaster Hydrojoule, LLC, United States

We will report on progress to develop a simpler, generalized web interface to provide a broader group of biomedical researchers useful access to one or more RDF triplestores for biomedical information (Bio2RDF, etc.). The approach "wraps" a SPARQL and/or Virtuoso-type query structure as a tool called by the more user-friendly query interface. The build employs a lightweight client, using JS, JS frameworks and JSON for exchange, with one or more servers providing backend routings and intermediary, user-specific storage (node.js and a specialized graph-store DB). This is a first step on a path toward a longer-term prototype development to create hypothesis generation related to the research query and related to the first rounds of data returned from the triplestore.

Poster 06 A Minimal Governance Layer for the Web of Linked Data

- Click here for PDF of poster PDF

Alexander Grüneberg, University of Alabama at Birmingham, United States
Jonas Almeida, University of Alabama at Birmingham, United States

The absence of governance engines capable of governing semantic web constructs by responding solely to embedded RDF (Resource Description Framework) descriptions of the governance rules remains one of the major obstacles to a web of linked data that can traverse domains with modulated privacy. This is particularly true when the governance is as fluid as the data models described by the RDF assertions. In part based on our previous experience developing and using S3DB ([http://en.wikipedia.org/wiki/S3DB](http://en.wikipedia.org/wiki/S3DB)), we have realized that the most scalable solutions will place minimalist requirements on the governance mechanism and will maximize the use of existing dereferencing conventions. Accordingly, we propose a generalized rule-based authorization system decoupled from services exposed over HTTP.

The authorization rules are written using JSON-LD (JavaScript Object Notation for Linked Data), apply to resources identified by URIs, and can be dynamically inserted and removed. The solution found uses a simple inheritance model in which resources inherit rules from their parent resource. This procedure seeks to advance the recently proposed Linked Data Platform ([http://www.w3.org/TR/ldp/](http://www.w3.org/TR/ldp/)) by weaving it with widely supported web authentication standards. The management of sensitive patient derived molecular pathology data is used as an illustrative case study where the proposed solution is validated. The mediating governance engine is made available as an open source NodeJS module at [https://github.com/ibl/Bouncer](https://github.com/ibl/Bouncer).

Poster 07 BAO Reveal: Assay Analysis Using the BioAssay Ontology and Open PHACTS

Simon Rakov, AstraZeneca, United States
Linda Zander Balderud, AstraZeneca, Sweden
Tom Plasterer, AstraZeneca, United States

Biochemical assay data is complex and highly variable. It contains attributes and terms that are not applicable to all assays, and uses controlled vocabularies inconsistently. Sometimes relevant data is entered in the wrong attribute, or
appears in a different document altogether. The complexity and inconsistency of assay data creates challenges for those who wish to query assay data for a given set of attributes, such as the technology chosen to conduct a given assay.

The BioAssay Ontology (BAO) is an ontology developed by the University of Miami and extended by the Open PHACTS consortium, itself part of the European Union’s Innovative Medicine Initiative (IMI). The purpose of the BAO is to standardize how we represent assay data. It incorporates many standard public ontologies, importing sections of the NCBI taxonomy, Uniprot, the Unit Ontology, the Ontology of Biomedical Investigation and the Gene Ontology, among others. More than 900 PubChem assays have been annotated according to BAO.

We have converted over 400 AstraZeneca primary HTS assays into BAO format in order to evaluate whether this common model can improve project success analyses based on assay technologies, help us to better understand the impact of technology artifacts such as frequent hitters, and improve our ability to employ data mining methodologies against assay data. We have created static visualizations that combine our internal data with the annotated PubChem assays. Most recently, this project has created a dynamic interface, "BAO Reveal," for querying and visualizing BAO data.

Our frequent hitter analysis methodology has found twice as many frequently-hitting assays when assay data is structured using BAO than with previous methods that did not have the granularity of BAO. This has suggested improvements to the data capture process from these assays. The dynamic faceting features and linked biochemical information in BAO Reveal provide researchers ways to investigate the underlying causes of broad assay patterns. This will allow us to focus assay development efforts on the most promising approaches.

BAO Reveal facilitates identification of screening technologies used for similar targets and helps analyze the robustness of a specific assay technology for a biological target. It can identify screening data to confirm assay reproducibility, and also assist frequent hitter analysis. As a linked data application built on Open PHACTS methodologies and other semantic web standards, BAO Reveal is well positioned for exploitation in multiple directions by multiple communities.

**Poster 08 A High Level API for Fast Development of High Performance Graphic Analytics on GPUs**

- Click here for PDF of poster [PDF](#)

Zhisong Fu, SYSTAP LLC., United States

High performance graph analytics are critical for a long list of application domains, ranging from social networks, information systems, security, biology, healthcare and life sciences. In recent years, the rapid advancement of many-core processors, in particular graphical processing units (GPUs), has sparked a broad interest in developing high performance graph analytics on these architectures. However, the single instruction multiple thread (SIMT) architecture used in GPUs places particular constraints on both the design and implementation of graph analytics algorithms and data structures, making the development of such programs difficult and time-consuming.

We present an open source library (MPGraph) that provides a high level abstraction which makes it easy to develop...
high performance graph analytics on massively parallel hardware. This abstraction is based on the Gather-Apply-Scatter (GAS) model as used in GraphLab. To deliver high performance computation and efficiently utilize the high memory bandwidth of GPUs, the underlying CUDA kernels use multiple sophisticated strategies, such as vertex-degree-dependent dynamic parallelism granularity and frontier compaction. Our experiments show that for many graph analytics algorithms, an implementation, with our abstraction, is up to two order of magnitude faster than parallel CPU implementations on up 24 CPU cores and has performance comparable to a state-of-the-art manually optimized GPU implementation. In addition, with our abstraction, new algorithms can be implemented in a few hours that fully exploit the data-level parallelism of the GPU and offer throughput of up to 3 billion traversed edges per second on a single GPU. We will explain the concepts behind the high-level abstraction and provide a starting point for people who want to write high throughput analytics.

MPGraph is now in its second release. Future work will extend the platform to multi-GPU workstations and GPU compute clusters.

**Poster 09 A Nanopublication Framework for Systems Biology and Drug Repurposing**

Jim McCusker, Rensselaer Polytechnic Institute, United States

Kusum Solanki, Rensselaer Polytechnic Institute, United States

Cynthia Chang, Rensselaer Polytechnic Institute, United States

Michel Dumontier, Stanford University, United States

Jonathan Dordick, Rensselaer Polytechnic Institute, United States

Deborah McGuinness, Rensselaer Polytechnic Institute, United States

Systems biology studies interactions between proteins, genes, drugs, and other molecular entities. A number of databases have been developed that serve as a patchwork across the landscape of systems biology, focusing on different experimental methods, many species, and a wide diversity of inclusion criteria. Systems biology has been used in the past to generate hypotheses for drug effects, but has become fragmented under the large number of disparate and disconnected databases. In our efforts to create a systematic approach to discovering new uses for existing drugs, we have developed Repurposing Drugs with Semantics (ReDrugS). Our ReDrugS framework can accept data from nearly any database that contains biological or chemical entity interactions. We represent this information as sets of nanopublications, fine-grained assertions that are tied to descriptions of their attribution and supporting provenance. These nanopublications are required to have descriptions of the experimental methods used to justify their assertions. By inferring the probability of truth from those experimental methods, we are able to create consensus assertions, along with a combined probability. Those consensus assertions can be searched for via a set of Semantic Automated Discovery and Integration (SADI) web services, which are used to drive a demonstration web interface. We then show how associations between exemplar drugs and cancer-driving genes can be explored and discovered. Future work will incorporate protein/disease associations, perform hypothesis generation on indirect drug targets, and test the resulting hypotheses using high throughput drug screening.

**Poster 10 An Interoperable Framework for Biomedical Image Retrieval and Knowledge Discovery**

Syed Ahmad Chan Bukhari, University of New Brunswick Canada, Canada
Biomedical images have an irrefutably central role in life science discoveries. Ongoing challenges associated with knowledge management and utility operations unique to biomedical image data are only recently gaining recognition. Making biomedical image content explicit is essential with regards to medical decision making such as diagnosis, treatment plans, follow-up, data management, data reuse for biomedical research and the assessment of care delivery. In our previous work, we have developed Yale Image finder (YIF), which is a novel BioMedical image search engine that indexes around two million biomedical image data, along with associated meta-data. While YIF is considered to be a veritable source of easily accessible biomedical images, there are still a number of usability and interoperability challenges that have yet to be addressed, including provenance and cross platform accessibility.

To overcome these issues and to accelerate the adoption of YIF for next generation biomedical applications, we have developed a publicly accessible Biomedical Image API with multiple modalities. The core API is powered by a dedicated semantic architecture that exposes Yale Image Finder (YIF) content as linked data, permitting integration with related information resources and consumption by linked data-aware data services. We have established a protocol to transform image data according to linked open data recommendations, and exposed it through a SPARQL endpoint and linked data explorer.

To facilitate the ad-hoc integration of image data with other online data resources, we built semantic web services, such that it is compatible with the SADI semantic web service framework. The utility of the combined infrastructure is illustrated with a number of compelling use cases and further extended through the incorporation of Domeo, a well known tool for open annotation. Domeo facilitates enhanced search over the images using annotations provided through crowdsourcing. In the current configuration, our triplestore holds more than thirty-five million triples and can be accessed and operated through syntactic or semantic solutions. Core features of the framework, namely: data reusability, system interoperability, semantic image search, automatic update and dedicated semantic infrastructure make system a state of the art resource for image data discovery and retrieval.

A demo can be accessed at: http://cbakerlab.unbsj.ca:8080/icyrus/

**Poster 11 HYDRA8: A Graphical Query Interface for SADI Semantic Web Services**

Christopher J. O. Baker, CEO

IPSNP Computing Inc.

Saint John, NB, Canada

HYDRA is a high-performance query engine operating on networks of SADI services representing various distributed resources. Here we present an intuitive end user-oriented querying and data browsing tool is designed construct SPARQL queries for issue to the HYDRA back end. The current prototype permits users to (i) enter key phrases and generate suitable query graphs corresponding to the key phrases; (ii) select between the suggested graphs; (iii) extend the automatically suggested query graphs manually, adding new relations and entities according to the current semantic
data schema (predicate map); (iv) run a query with Hydra and see the results in a tabular form. A demo will be available at the conference.

**Tutorial**


*Updated March 03, 2014*

Those attending the **Callimachus Tutorial on Wednesday, February 26th**, have three ways to maximize their experience. The tutorial will be interactive and projected to the room so everyone can follow along.

1. Those wishing to explore the code at their own pace are welcome to install Callimachus Open Source from [http://callimachusproject.org](http://callimachusproject.org).
2. Tutorial attendees are also being given the opportunity to get access to their very own copy of Callimachus Enterprise running on Rackspace Cloud.
3. Anyone who wishes to write Callimachus applications using Callimachus Enterprise during the conference should send email to cshals@3roundstones.com and a cloud instance will be established for them.

We look forward to seeing you at the tutorial!

**Tutorial – Wednesday, February 26**

1:00 pm – 5:00 pm

*Please note the tutorial is includes as part of the conference registration fee. Delegates wishing to attend should select the tutorial during registration.*

**Enterprise and Scientific Data Sharing Using Callimachus**

[Click here for presentation slides](http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA)

Callimachus Enterprise is an innovative solution to build and host Web applications combining enterprise and public data on the cloud. Applications built on Callimachus have specific benefits for scientific data sharing.

This tutorial will introduce common issues with existing enterprise data sharing and show how Callimachus can help. Callimachus’ particular benefits to scientific enterprises will be highlighted, such as on-the-fly data integration from multiple sources, and Callimachus applications aimed at scientists such as presentation material reuse, tracking and attribution.

Attendees will be led through the creation of simple Callimachus applications to include:

- Data integration from multiple sources
- Data re-use
- Development of visualizations
- Leveraging data in the Web, including big data sets.

Callimachus Enterprise is based upon The Callimachus Project, an Open Source Software project which limits vendor lock in and encourages community involvement. Callimachus implements Web standards to ensure it plays well with others.

**Participants should have some understanding of relational databases and enterprise in general would be helpful. Knowledge of semantics is not necessary as the tutorial will cover the very basics.**

**Presenter:**
Dr. David Wood is CTO of enterprise software vendor 3 Round Stones Inc. He has contributed to the evolution of the World Wide Web since 1999, especially in the formation of standards and technologies for the Semantic Web. He has architected key aspects of the Web to include the Persistent Uniform Resource Locator (PURL) service and several Semantic Web databases and frameworks including the Callimachus Project. He has represented international organizations in the evolution of Internet standards at the International Standards Organization (ISO), the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C) and is currently chair of the W3C RDF Working Group. David is the author of *Programming Internet Email* (O'Reilly, 1999) and *Linked Data: Structured data on the Web* (Manning, 2013) and editor of *Linking Enterprise Data* (Springer, 2010) and *Linking Government Data* (Springer, 2011).

**Tweets**

Source: [https://twitter.com/search?q=%23cshals&src=hash](https://twitter.com/search?q=%23cshals&src=hash)

Results for #cshals
Top / All
1 new result

1.

[Brent Gendleman @brentgendelemen](https://twitter.com/brentgendelemen)  Mar 7
Jim McCusker @jpmccu presents nanopublication framework research for systems biology & drug repurposing #CSHALS http://ow.ly/um9oV ...

2.

5AM Solutions 5@amsolutions Mar 5

Competitive advantages of applying semantic technologies in biomedical research - highlights from #CSHALS - http://hub.am/1icsSkm

3. Followed by jahendler and 1 other

Jim McCusker @jpmccu Mar 5

My report from this year's #CSHALS http://info.5amsolutions.com/blog/bid/157515/Highlights-from-the-Conference-on-Semantics-in-Healthcare-and-Life-Sciences-CSHALS ...

4. Followed by Ontology Summit and 5 others

Kerstin Forsberg @kerfors Mar 4

More interesting presentations from Conference on Semantics in Healthcare and Life Sciences #cshals now published: http://www.iscb.org/cshals2014-program/cshals2014-full-agenda ...

5.

Kerstin Forsberg @kerfors Mar 4


6. Top Tweet
7.  

Jim McCusker @jpmccu  Feb 27

Our (@dlmcguinness, @micheldumontier, and others) poster just got Honorable Mention for Best Poster at #CSHALS! pic.twitter.com/4vRh5niaQE

8.  

Jim McCusker @jpmccu  Feb 27

At #CSHALS Frederick Malfait: RDF allows for incremental data development, because the model IS the implementation.
Joanne Luciano  @JoanneLuciano  Feb 27

#CSHALS  Hoffmann-LaRoche Case Study Overview Calculated ROI – will save 60M in next 3 years using Semantic Technologies

Kerstin Forsberg  @kerfors  Feb 27

The FDA/PhUSE Semantic Technology project http://www.phusewiki.org/wiki/index.php?title=Semantic_Technology ... on #CSHALS program later today http://www.iscb.org/cshals2014-program/cshals2014-full-agenda ... #cdisc

YarcData  @YarcData  Feb 26


Bernadette Hyland  @BernHyland  Feb 26

Great to see about 25 people attending Callimachus talk 1/4 are new to Linked Data the rest are building apps #CSHALS pic.twitter.com/hD8t3yVnSZ
PART I: OVERVIEW

- **Federal Agency Name**: Defense Advanced Research Projects Agency (DARPA), Information Innovation Office (I2O)
- **Funding Opportunity Title**: Big Mechanism
- **Announcement Type**: Initial Announcement
- **Funding Opportunity Number**: DARPA-BAA-14-14

**Catalog of Federal Domestic Assistance Numbers (CFDA)**: 12.910 Research and Technology Development

**Dates**
- Posting Date: January 30, 2014
- Proposal Due Date: March 18, 2014, 12:00 noon (EST)
PART II: FULL TEXT OF ANNOUNCEMENT

I. FUNDING OPPORTUNITY DESCRIPTION

DARPA is soliciting innovative research proposals in the area of reading research papers and abstracts to construct and reason over explanatory, causal models of complicated systems. Proposed research should investigate innovative approaches that enable revolutionary advances in science, devices, or systems. Specifically excluded is research that primarily results in evolutionary improvements to the existing state of practice.

This broad agency announcement (BAA) is being issued, and any resultant selection will be made, using procedures under Federal Acquisition Regulation (FAR) 35.016. Any negotiations and/or awards will use procedures under FAR 15.4 (or 32 CFR 22 for grants and cooperative agreements). Proposals received as a result of this BAA shall be evaluated in accordance with evaluation criteria specified herein through a scientific review process.

DARPA BAAs are posted on the Federal Business Opportunities (FBO) website (http://www.fbo.gov/) and the Grants.gov website (http://www.grants.gov/).

The following information is for those wishing to respond to the BAA.

A. Introduction

Some of the systems that matter most to the DoD are very complicated. Ecosystems, brains, economic and social systems have many parts and processes, but they are studied piecewise, and their literatures and data are fragmented,
distributed, and inconsistent \[1,2\]. It is difficult to build complete, explanatory models of complicated systems, so effects in these systems that are brought about by many interacting factors are poorly understood.

Big Mechanisms are causal, explanatory models of complicated systems in which interactions have important causal effects. The collection of Big Data is increasingly automated, but the creation of Big Mechanisms remains a human endeavor made increasingly difficult by the fragmentation and distribution of knowledge. To the extent that we can automate the construction of Big Mechanisms, we can change how science is done.

The Big Mechanism program will develop technology to read research abstracts and papers to extract fragments of causal mechanisms, assemble these fragments into more complete causal models, and reason over these models to produce explanations. The domain of the program will be cancer biology with an emphasis on signaling pathways.

### B. Program Description/Scope

Although the domain of the Big Mechanisms program is cancer biology, and systems biology and signaling networks are referred to throughout this BAA, the goal of the program is to develop technologies for a new kind of science in which research is integrated more or less immediately – automatically or semi-automatically – into causal, explanatory models of unprecedented completeness and consistency. Cancer pathways are just one example of causal, explanatory models.

Here is one way to factor the technologies in the Big Mechanism program:

1) Read abstracts and papers to extract fragments of causal mechanisms;
2) Assemble fragments into more complete Big Mechanisms;
3) Explain and reason with Big Mechanisms.

As the program evolves, its technologies will probably factor in other ways, so this framework is not intended to be prescriptive or permanent. It does, however, give us a convenient organization for discussing some technical issues in the program below.

The Big Mechanism program will require new research and the integration of several research areas, particularly statistical and knowledge-based Natural Language Processing (NLP); curation and ontology; systems biology and mathematical biology; representation and reasoning; and quite possibly other areas such as visualization, simulation, and statistical foundations of very large causal networks. Machine reading researchers will need to develop deeper semantics to represent the causal and often kinetic models described in research papers. Deductive inference and qualitative simulation will probably not be sufficient to model the complicated dynamics of signaling pathways and will need to be augmented or replaced by probabilistic and quantitative models. Classification and prediction will continue to be important, but causal explanation is primary. On the principle that we should use the knowledge we have, extant databases and ontologies will provide top-down constraints on reading, assembly of big mechanisms and explanation.
C. Technical Areas

As noted, some technologies for the Big Mechanism program can be grouped roughly into three areas: Reading, Assembling and Explaining. Before describing these areas in more detail, we need to say a little about program-wide resources and activities.

Prior to the program kick-off, DARPA will develop a "starter kit" that will contain a first draft of a representation language for biological processes, and means to access the many curated ontologies and databases that are relevant to biological processes. As the program goes forward, its representation language and curated resources will evolve. Whether this is managed by a single performer or by committees formed from several performers will depend on the proposals DARPA receives. However, most proposers should plan to put some effort into program-wide resources, and some might make the development of representations and the curation of resources central to their proposals.

DARPA intends to have at least one integrated demonstration system that reads, assembles and explains big mechanisms. How this work will be done depends on the proposals that DARPA receives. It is not necessary to propose to build an integrated demonstration system. Indeed, a proposal to merely integrate and demonstrate technology, without a well-motivated, innovative vision of how to achieve the goals of the Big Mechanism program, is unlikely to be successful.

Those who propose to build an integrated system should explain how they will provide flexible control of its components. Although one can imagine a pipeline that begins with Reading, then passes fragments to Assembly, which passes big mechanisms to Explanation, one can also see how Assembly could help Reading, or Explanation could constrain what gets assembled. One can envision a query-driven system that actively looks for papers to extend a big mechanism to answer a user's question; or a system that engages a human user when it detects a contradiction in the literature; or a system that searches for therapeutic targets and assembles only pertinent fragments of pathways; or a Reading system that directs Assembly to try several alternative interpretations of a paragraph and return the one that's most consistent with the current big mechanism; and so on. Rigid, pipelined control jeopardizes the Big Mechanism program by limiting how its technologies interact with each other, with exogenous resources such as ontologies and databases, and with users.

With this preamble, we turn now to the three technical areas: Reading, Assembly and Explanation.

1. Reading

In the literature on signaling pathways, research papers usually describe small fragments of big mechanisms. The job of Reading is to extract these fragments. Said differently, the relevant semantics of papers and abstracts will be the causal mechanisms they describe.

Reading researchers should assume that the purpose of reading is to augment or extend known signaling pathways, rather than build pathways de novo. This is to avoid self-inflicted research (the tendency to make research problems harder than they need to be) and to encourage top-down reasoning in Reading. Performers should exploit extant ontologies, databases and other readily available sources of information about relevant signaling pathways.
As with all natural language processing, Reading is bedeviled by ambiguity [5]. The mapping of named entities to biological entities is many-to-many. Context matters, but is often missing; for example, the organism in which a pathway is studied might be mentioned once at the beginning of a document and ignored thereafter. Although the target semantics involves processes, these can be described at different levels of detail and precision. For example, “β-catenin is a critical component of Wnt-mediated transcriptional activation” tells us only that β-catenin is involved in a process; whereas, “ARF6 activation promotes the intracellular accumulation of β-catenin” tells us that ARF6 promotes a process; and “L-cells treated with the GSK3 β inhibitor LiCl (50 mM) . . . showed a marked increase in β-catenin fluorescence within 30 – 60 min” describes the kinetics of a process. Processes also can be described as modular abstractions, as in “. . . the endocytosis of growth factor receptors and robust activation of extracellular signal-regulated kinase”. It might be possible to extract causal skeletons of complicated processes (i.e., the entities and how they causally influence each other) by reading abstracts, but it seems likely that extracting the kinetics of processes will require reading full papers. It is unclear whether this program will be able to provide useful explanations of processes if it doesn’t extract the kinetics of these processes.

The prior knowledge requirement for Reading in this domain is large but not impossibly so (and it probably includes relatively little commonsense knowledge). Much prior knowledge is already encoded in ontologies. Complicated processes decompose into many interacting instances of a few, simpler processes. Putting it crudely, the texts have many semantically unique nouns, but few semantically unique verbs and adjectives.

Signaling pathways are complicated processes, so it will be necessary to develop or adopt representations of processes as the targets for Reading. Texts go in, representations of processes come out. The Big Mechanism program will provide a “starter kit” that includes a first draft of a representation of processes, but other representations might be proposed and will be evaluated according to how they support Reading, Assembly and Explanation. Ideally, one product of the program will be a probabilistic representation of processes that is readable by humans and machines, and is compatible with extant databases and ontologies, and is easily and informatively visualized, and supports Assembly and Explanation, in addition to serving as a target for Reading.

2. Assembly

Given a prior big mechanism and many fragments, the job of Assembly is to extend the mechanism as thoroughly as is warranted by the fragments and other knowledge in databases and ontologies. Assembly also involves finding semantic inconsistencies among fragments, and finding “holes” or parts of big mechanisms for which no causal fragments are known. Consider the following excerpts:

Excerpt 1: HBP1 is a repressor of the cyclin D1 gene and inhibits the Wnt signaling pathway. The inhibition of Wnt signaling and growth requires a common domain of HBP1. The apparent mechanism is an inhibition of TCF/LEF DNA binding through physical interaction with HBP1.[3]

Excerpt 2: HBP1 represses the DNMT1 promoter through sequence-specific binding (of the type TTCATTCATTCA) and the activity of HBP1 itself is regulated through acetylation at any of 5 sites in the protein. Mutation of any acetylation sites abrogates the HBP1 repression activity. The HBP1-mediated repression of the DNMT1 gene then decreases overall DNA methylation. On the p16 gene, HBP1 expression leads to a similar DNA hypomethylation, but HBP1 instead binds to putative HBP1 activation element (of the type GGTTAGG) to give activation.[4]
Which words in these excerpts refer to the same biological entities? Do the excerpts describe the same process at different levels of abstraction, or different processes? If they describe the same process, are the descriptions semantically consistent or does one contradict the other? Do these excerpts extend the big mechanism we already have, or are they redundant? In this example, neither excerpt describes kinetics, but when kinetic parameters are available, are they consistent?

Clearly it will be difficult to answer any of these questions if we lack a formal representation language for Big Mechanisms, a requirement that Assembly shares with Reading.

To avoid self-inflicted research, Assembly should take advantage of what is already known about signaling pathways. Most abstracts and papers describe variations or details or anomalous behaviors of well-known pathways (and the annual evaluations for this program will focus on pathways about which we already know a lot) so Assembly should focus on integrating new fragments into existing big mechanisms and on detecting inconsistencies and holes.

Note that Assembly need not care whether it gets fragments from Reading or from other sources. It should be able to integrate fragments produced by Reading, but it isn't limited to those fragments.

3. Explanation

The word "Explanation" was selected, along with "Big Mechanism," to emphasize causal reasoning about complicated processes, and to counteract the current preoccupation with classification and clustering. That said, the Explanation technical area encompasses many kinds of inference, exemplified by the following questions:

- If there are contradictions between fragments of a causal mechanism (e.g., one fragment says X up-regulates Y, another says the opposite), can they be resolved by a kinetic mechanism?
- What are likely consequences of increasing or decreasing levels of a protein?
- Are there alternative pathways by which a gene can be expressed?
- Which other pathways intersect with this one?
- If there's a hole in our knowledge, can knowledge about orthologs plug it?
- Can we find a small modification to a mechanism that would explain anomalous results?
- Can we compress a Big Mechanism into a smaller one that has the same (or nearly the same) explanatory power?

In addition to these general sorts of questions about signaling pathways, suitably qualified researchers might propose to tackle specific kinds of inference about tumorigenesis and other processes in cancer biology.

Just as Assembly is not limited to fragments produced by Reading, Explanation is not limited to big mechanisms produced by Assembly. As more data are published in machine-readable forms, and as high-throughput sources produce more and more data automatically, we can anticipate extending the Explanation task to incorporate two kinds of data-intensive reasoning:

- Inducing big mechanisms (or extensions of big mechanisms) directly from data, by causal induction algorithms or something similar;
- Checking the predictions of big mechanisms against available data.

Proposers to the Explanation technical area are encouraged to explore additional extensions to this task.

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by MindTouch
D. Program Structure

The Big Mechanism program is designed to last 42 months and consists of three phases. No down-selections are anticipated. However, to continue from one phase to the next participants must produce technology that supports the goals of the program.

Phase 1 is anticipated to be 18 months long. The goals for this phase include:

• Development of a formal representation language for biological processes;
• Extraction of fragments of known signaling networks from a relatively small and carefully selected corpus of texts and encoding of these fragments in a formal language;
• Preliminary work on extracting information about the kinetics of fragments;
• Integration of fragments into models of known signaling networks, detecting inconsistencies and anomalies;
• Some ability to reason about the known networks, including qualitative predictions of the effects of manipulations;
• Coordinated work by one or more performers to make extant ontological and database resources available to all performers;
• Coordinated work on making some technologies interoperable, so that by the end of Phase 1 there is at least one system that performs Reading, Assembly and Explanation; and,
• Design and perhaps implement a flexible control architecture that provides top-down constraints on Reading, Assembly and Explanation.

In Phase 2, which is anticipated to be 12 months long, the goals include (at a minimum):

• Extraction of fragments of at least one signaling network that is not well-understood;
• Extension of this network, detecting inconsistencies and anomalies;
• Publication of this network; and
• Continued work on the extraction of kinetics and integration of kinetics into explanations of signaling networks.

In Phase 3, which is anticipated to be 12 months long, the program goals include:

• Extraction of qualitative and kinetic relations between signaling pathway entities from very large numbers of abstracts and papers, representing many pathways;
• Publication of the most complete and consistent pathways that can be supported by the literature; and
• Demonstration of useful explanatory capabilities; for example, suggesting targets for therapy, or resolving inconsistencies in the publication record, or prospectively suggesting a hypothesis. Demonstration of tight integration of Reading, Assembly, Explanation and available ontological and database resources.

E. Schedule/Milestones

A notional schedule is shown below, include a speculative evaluation schedule. Proposers are encouraged to propose alternative tasks for each phase, as suited to the work being proposed.
In addition to the elements shown above, performers will be expected to attend all DARPA meetings related to the program. At a minimum there will be annual PI meetings, most likely held at DARPA’s facility in Arlington, Virginia. It is likely that there will also be two weeks of onsite participation in Arlington, Virginia, where performers will be expected to send staff members or senior graduate students and/or post-docs to participate in community-wide development efforts. The program manager will also make annual site visits when possible.

F. Performance Requirements

The deliverables for all performers will be assessed at the program’s phase boundaries and include:

- Fundamental research and publications in areas that support the goals of the Big Mechanism program;
- Participation in program-wide activities and support of program-wide resources, including but not limited to participation in integrated demonstrations;
- Running local evaluations associated with claims in the performer’s proposal;
- Participating in program-wide evaluations; and
- Participating in intensive development sessions to be held in Arlington, VA.

In addition, some performers might deliver software, integrated demonstrations, curated resources, representation specifications and/or representation languages, control architectures and so on.

G. Government-furnished Property/Equipment/Information

Prior to the program kick-off, DARPA will develop and provide a “starter kit” to the performers that will contain a first draft of a representation language for biological processes, and means to access the many curated ontologies and databases that are relevant to biological processes. This “starter kit” will be provided to the performers as Government-furnished Information.

H. Intellectual Property

The program will emphasize creating and leveraging open source technology and architecture. Intellectual property rights asserted by proposers are strongly encouraged to be aligned with open source regimes. See Section VI.B.1 for more details on intellectual property.
A key goal of the program is to establish open, standards-based, multi-source, plug-and-play algorithms that allow for interoperability and integration. This includes the ability to easily add, remove, substitute, and modify software and hardware components. This will facilitate rapid innovation by providing a base for future users or developers of program technologies and deliverables. Therefore, it is desired that all noncommercial software (including source code), software documentation, hardware designs and documentation, and technical data generated by the program be provided as deliverables to the Government, with a minimum of Government Purpose Rights (GPR). See Section VI.B.1 for more details on intellectual property.

I. Additional Guidance for Proposers

Proposers are encouraged to bridge between Reading, Assembly and Explanation, either as producers or consumers of technologies. For example, Reading might benefit from top-down context provided by Assembly, or Assembly might become more focused by requirements from Explanation.

Proposers can work in one or all technical areas, or they can ignore these areas altogether and propose another factorization of Big Mechanism technologies, such as a focus on data and curation or some other area that is not reflected in this BAA.

Proposers can submit more than one proposal, but should keep in mind that one goal of the Big Mechanism program is to integrate Reading, Assembly and Explanation.

Big Mechanism research and development has not been tried before, so this is not the time to prescribe how it is done. Though it might frustrate proposers, it is best to leave some aspects of the program under-specified, including:

- The formal framework in which Big Mechanisms are represented (other than what will be provided in the starter kit);
- Which performers or technical areas are responsible for developing this framework;
- Whether Big Mechanism functionality runs offline or online in a user-driven system;
- The extent to which the program incorporates and reasons with published data (in addition to abstracts and articles that discuss data);
- Whether or to what extent Explanation will focus on potential targets for therapy;
- Whether any element of the program is fully automatic or relies to some extent on human assistance; and
- How selected performers should self-organize into teams.

Proposers are encouraged to think creatively about computers assembling big mechanisms that explain extremely complicated systems given a variety of source material, including textual and data sources.

Proposers should envision Big Mechanism systems for signaling pathways, but are also encouraged to think about other complicated systems and dynamics, such as climate change, economic development, or the brain.

Proposers should carefully scope what they can contribute to a Big Mechanism program and should not feel obliged to propose complete, integrated Big Mechanism systems.

However, proposers should be clear about anticipated dependencies and synergies between what they propose and other technologies.
II. AWARD INFORMATION

A. Awards

Multiple awards are anticipated. The level of funding for individual awards made under this solicitation has not been predetermined and will depend on the quality of the proposals received and the availability of funds. Awards will be made to proposers whose proposals are determined to be the most advantageous and provide the best value to the Government, all factors considered, including the potential contributions of the proposed work, overall funding strategy, and availability of funding. See Section V.B. for further information.

The Government reserves the right to:

- select for negotiation all, some, one, or none of the proposals received in response to this solicitation;
- make awards without discussions with proposers;
- conduct discussions with proposers if it is later determined to be necessary;
- segregate portions of resulting awards into pre-priced options;
- accept proposals in their entirety or to select only portions of proposals for award;
- fund proposals in increments with options for continued work at the end of one or more phases;
- request additional documentation once the award instrument has been determined (e.g., representations and certifications); and
- remove proposers from award consideration should the parties fail to reach agreement on award terms within a reasonable time or the proposer fails to provide requested additional information in a timely manner.

Proposals selected for award negotiation may result in a procurement contract, grant, cooperative agreement, or other transaction (OT) depending upon the nature of the work proposed, the required degree of interaction between parties, and other factors. In all cases, the Government contracting officer shall have sole discretion to select award instrument type and to negotiate all instrument terms and conditions with selectees. Proposers are advised that, if they propose grants or cooperative agreements, the Government contracting officer may select other award instruments, as appropriate. Publication or other restrictions will be applied, as necessary, if DARPA determines that the research resulting from the proposed effort will present a high likelihood of disclosing performance characteristics of military systems or manufacturing technologies that are unique and critical to defense. Any award resulting from such a determination will include a requirement for DARPA permission before publishing any information or results on the program. For more information on publication restrictions, see Section II.B.

B. Fundamental Research

It is Department of Defense (DoD) policy that the publication of products of fundamental research will remain unrestricted to the maximum extent possible. National Security Decision Directive (NSDD) 189 established the national policy for controlling the flow of scientific, technical, and engineering information produced in federally funded fundamental research at colleges, universities, and laboratories. NSDD 189 defines fundamental research as follows:

‘Fundamental research’ means basic and applied research in science and engineering, the results of which ordinarily are published and shared broadly within the scientific community, as distinguished from proprietary
research and from industrial development, design, production, and product utilization, the results of which ordinarily are restricted for proprietary or national security reasons.

As of the date of publication of this BAA, the Government expects that program goals as described herein may be met by proposers intending to perform fundamental research. The Government does not anticipate applying publication restrictions of any kind to individual awards for fundamental research that may result from this BAA. Notwithstanding this statement of expectation, the Government is not prohibited from considering and selecting research proposals that, while perhaps not qualifying as fundamental research under the foregoing definition, still meet the BAA criteria for submissions. If proposals are selected for award that offer other than a fundamental research solution, the Government will either work with the proposer to modify the proposed statement of work to bring the research back into line with fundamental research or else the proposer will agree to restrictions in order to receive an award.

Proposers should indicate in their proposal whether they believe the scope of the proposed research is fundamental. For certain research projects, it may be possible that although the research to be performed by the prime proposer is non-fundamental, a subcontractor’s tasks may be considered fundamental research. In those cases, it is the prime proposer’s responsibility to explain in their proposal why its subcontractor’s effort is fundamental research. While proposers should clearly explain the intended results of their research, DARPA shall have sole discretion to determine whether the project is considered fundamental research. Awards for non-fundamental research will include the following statement or similar provision:

There shall be no dissemination or publication, except within and between the contractor and any subcontractors, of information developed under this contract or contained in the reports to be furnished pursuant to this contract without prior written approval of DARPA’s Public Release Center (DARPA/PRC). All technical reports will be given proper review by appropriate authority to determine which Distribution Statement is to be applied prior to the initial distribution of these reports by the contractor. With regard to subcontractor proposals for Contracted Fundamental Research, papers resulting from unclassified contracted fundamental research are exempt from prepublication controls and this review requirement, pursuant to DoD Instruction 5230.27 dated October 6, 1987.

When submitting material for written approval for open publication, the contractor/awardee must submit a request for public release to the PRC and include the following information: 1) Document Information: title, author, short plain-language description of technology discussed in the material (approx. 30 words), number of pages (or minutes of video) and type (e.g., briefing, report, abstract, article, or paper); 2) Event Information: type (e.g., conference, principal investigator meeting, article or paper), date, desired date for DARPA’s approval; 3) DARPA Sponsor: DARPA Program Manager, DARPA office, and contract number; and 4) Contractor/Awardee’s Information: POC name, e-mail address and phone number. Allow four weeks for processing; due dates under four weeks require a justification.

Unusual electronic file formats may require additional processing time. Requests may be sent either to prc@darpa.mil or 675 North Randolph Street, Arlington VA 22203-2114, telephone (571) 218-4235. See http://www.darpa.mil/NewsEvents/Publ...se_Center.aspx for further information about DARPA’s public release process.

III. ELIGIBILITY INFORMATION
A. Eligible Applicants

All responsible sources capable of satisfying the Government's needs may submit a proposal that shall be considered by DARPA.

1. Federally Funded Research and Development Centers (FFRDCs) and Government Entities

FFRDCs and Government entities (e.g., Government/National laboratories, military educational institutions, etc.) are subject to applicable direct competition limitations and cannot propose to this solicitation in any capacity unless the following conditions are met.

- FFRDCs must clearly demonstrate that the proposed work is not otherwise available from the private sector and must provide a letter on official letterhead from their sponsoring organization citing the specific authority establishing the FFRDC’s eligibility to propose to Government solicitations and compete with industry, and compliance with the terms and conditions in the associated FFRDC sponsor agreement. This information is required for FFRDCs proposing as either prime contractors or subcontractors.
- Government entities must clearly demonstrate that the proposed work is not otherwise available from the private sector and provide documentation citing the specific statutory authority (and contractual authority, if relevant) establishing their eligibility to propose to Government solicitations.

At the present time, DARPA does not consider 15 USC § 3710a to be sufficient legal authority to show eligibility. For some entities, 10 USC § 2539b may be the appropriate statutory starting point; however, specific supporting regulatory guidance, together with evidence of agency approval, will still be required to fully establish eligibility.

DARPA will consider eligibility submissions on a case-by-case basis; however, the burden to prove eligibility for all team members rests solely with the proposer.

2. Foreign Participation

Non-U.S. organizations and/or individuals may participate to the extent that such participants comply with any necessary nondisclosure agreements, security regulations, export control laws, and other governing statutes applicable under the circumstances.

B. Procurement Integrity, Standards of Conduct, Ethical Considerations and Organizational Conflicts of Interest (OCIs)

Current Federal employees are prohibited from participating in particular matters involving conflicting financial, employment, and representational interests (18 USC §§ 203, 205, and 208). Prior to the start of proposal evaluation, the Government will assess potential COIs and will promptly notify the proposer if any appear to exist. The Government assessment does NOT affect, offset, or mitigate the proposer’s responsibility to give full notice and planned mitigation for all potential organizational conflicts, as discussed below.

In accordance with FAR 9.5 and without prior approval or a waiver from the DARPA Director, a contractor cannot simultaneously provide scientific, engineering, and technical assistance (SETA) or similar support and be a technical...
performer. As part of the proposal submission, all members of a proposed team (prime proposers, proposed subcontractors and consultants) must affirm whether they (individuals and organizations) are providing SETA or similar support to any DARPA technical office(s) through an active contract or subcontract. Affirmations must state which office(s) the proposer and/or proposed subcontractor/consultant supports and must provide prime contract number(s). All facts relevant to the existence or potential existence of OCIs must be disclosed. The disclosure shall include a description of the action the proposer has taken or proposes to take to avoid, neutralize, or mitigate such conflict. If, in the sole opinion of the Government after full consideration of the circumstances, a proposal fails to fully disclose potential conflicts of interest and/or any identified conflict situation cannot be effectively mitigated, the proposal will be rejected without technical evaluation and withdrawn from further consideration for award.

If a prospective proposer believes a conflict of interest exists or may exist (whether organizational or otherwise) or has a question as to what constitutes a conflict, a summary of the potential conflict should be sent to BigMechanism@darpa.mil before preparing a proposal and mitigation plan.

C. Cost Sharing/Matching

Cost sharing is not required; however, it will be carefully considered where there is an applicable statutory condition relating to the selected funding instrument (e.g., OTs under the authority of 10 USC § 2371).

D. Other Eligibility Requirements

Ability to Receive Awards in Multiple Technical Areas - Conflicts of Interest

A proposer may submit more than one proposal as a prime contractor. Each proposal may cover any number of Technical Areas. If a proposal is submitted for more than one Technical Area, the decision as to which Technical Area(s), if any, to consider for award is at the discretion of the Government. Proposers may receive awards for multiple Technical Areas. There are no conflicts between Technical Areas.

IV. APPLICATION AND SUBMISSION INFORMATION

A. Address to Request Application Package

This document contains all information required to submit a response to this solicitation. No additional forms, kits, or other materials are needed except as referenced herein. No request for proposal (RFP) or additional solicitation regarding this opportunity will be issued, nor is additional information available except as provided at the Federal Business Opportunities or Grants.gov websites (http://www.fbo.gov or http://www.grants.gov/) or referenced herein.

B. Content and Form of Application Submission
1. Proposals

Proposals consist of Volume 1: Technical and Management Proposal (including mandatory Appendix A and optional Appendix B) and Volume 2: Cost Proposal.

Proposers are encouraged to submit concise but descriptive proposals. The Government will not consider pages in excess of the page count limitation, as described herein. Proposals with fewer than the maximum number of pages will not be penalized. Information incorporated into Volume 2: Cost Proposal that is not related to cost will not be considered.

All pages shall be formatted for printing on 8-1/2 by 11-inch paper with a font size not smaller than 12 point. Font sizes of 8 or 10 point may be used for figures, tables, and charts. Document files must be in .pdf, .odx, .doc, .docx, .xls, or .xlsx formats. Submissions must be written in English.

Proposals not meeting the format prescribed herein may not be reviewed.

a. Volume 1: Technical and Management Proposal

The maximum page count for Volume 1 is 14 pages if the proposal covers one of the three technical areas. The page count for Volume 1 can be increased to 16 pages if the proposal covers two of the technical areas and up to 18 pages if the proposal covers all three of the technical areas. The page count can be 20 pages if the proposal covers all three technical areas in an integrated system, or 20 pages if the proposal offers a significantly different factoring of the technical areas than Reading, Assembly and Explanation. These page limits include all figures, tables and charts but not the cover sheet, table of contents or appendices. A submission letter is optional and is not included in the page count. Appendix A does not count against the page limit and is mandatory. While Appendix B does not count against the overall page limit and is optional, it cannot exceed 15 pages.

Additional information not explicitly called for here must not be submitted with the proposal, but may be included as links in the bibliography in Appendix B. Such materials will be considered for the reviewers’ convenience only and not evaluated as part of the proposal.

Volume 1 must include the following components:

i. Cover Sheet: Include the following information.

- Label: “Proposal: Volume 1”
- BAA number (DARPA-BAA-14-14)
- Technical Area(s)
- Proposal title
- Lead organization (prime contractor) name
- Type of business, selected from the following categories: Large Business, Small Disadvantaged Business, Other Small Business, HBCU, MI, Other Educational, or Other Nonprofit
- Technical point of contact (POC) including name, mailing address, telephone, and email
- Administrative POC including name, mailing address, telephone number, and email address
- Award instrument requested: procurement contract (specify type), grant, cooperative agreement or OT.
- Place(s) and period(s) of performance
- Other team member (subcontractors and consultants) information (for each, include Technical POC name, organization, type of business, mailing address, telephone number, and email address)
- Proposal validity period (minimum 120 days)
- Data Universal Numbering System (DUNS) number
- Taxpayer identification number
- Commercial and Government Entity (CAGE) code
- Proposer’s reference number (if any)

ii. Table of Contents

iii. Executive Summary: Provide a two page synopsis of the proposed project. The first page should include answers to the following questions:

- What is the proposed work attempting to accomplish or do?
- How is it done today, and what are the limitations? (Include a list of references (URL or bibliographic) for NOT MORE THAN 5 papers which have informed the development of your proposed approach.)
- Who or what will be affected and what will be the impact if the work is successful?
- How much will it cost, and how long will it take?

The first page of the executive summary should also include a description of the key technical challenges, a concise review of the technologies proposed to overcome these challenges and achieve the project’s goal, and a clear statement of the novelty and uniqueness of the proposed work.

On the second page of the executive summary, performers should provide an editable one-page graphic or illustration of the proposed approach, including expected inputs, technical approach, and outputs.

iv. Goals and Impact: Describe what the proposed team is trying to achieve and the difference it will make (qualitatively and quantitatively) if successful. Describe the innovative aspects of the project in the context of existing capabilities and approaches, clearly delineating the uniqueness and benefits of this project in the context of the state of the art, alternative approaches, and other projects from the past and present. Describe how the proposed project is revolutionary and how it significantly rises above the current state of the art.

Describe the deliverables associated with the proposed project and any plans to commercialize the technology, transition it to a customer, or further the work. Discuss the mitigation of any issues related to sustainment of the technology over its entire lifecycle, assuming the technology transition plan is successful.

v. Technical Plan: Outline and address technical challenges inherent in the approach and possible solutions for overcoming potential problems. Demonstrate a deep understanding of the technical challenges and present a credible (even if risky) plan to achieve the project’s goal. Discuss mitigation of technical risk. Provide appropriate measurable milestones (quantitative if possible) at intermediate stages of the project to demonstrate progress, and a plan for achieving the milestones.
vi. Evaluation Plan: State concisely the research and technology claims and describe how they will be evaluated.

vii. Management Plan: Provide a summary of expertise of the proposed team, including any subcontractors/consultants and key personnel who will be executing the work. Resumes count against the proposal page limit so proposers may wish to include them as links in Appendix B below. Identify a principal investigator (PI) for the project. Provide a clear description of the team’s organization including an organization chart that includes, as applicable, the relationship of team members; unique capabilities of team members; task responsibilities of team members; teaming strategy among the team members; and key personnel with the amount of effort to be expended by each person during the project. Provide a detailed plan for coordination including explicit guidelines for interaction among collaborators/subcontractors of the proposed project. Include risk management approaches. Describe any formal teaming agreements that are required to execute this project.

viii. Capabilities: Describe organizational experience in relevant subject area(s), existing intellectual property, specialized facilities, and any Government-furnished materials or information. Discuss any work in closely related research areas and previous accomplishments.

ix. Statement of Work (SOW): The SOW must provide a detailed task breakdown, citing specific tasks and their connection to the interim milestones and metrics, as applicable. Each year of the project should be separately defined. The SOW must not include proprietary information. For each defined task/subtask, provide:
   - A general description of the objective.
   - A detailed description of the approach to be taken to accomplish each defined task/subtask.
   - Identification of the primary organization responsible for task execution (prime contractor, subcontractor(s), consultant(s), by name).
   - A measurable milestone, i.e., a deliverable, demonstration, or other event/activity that marks task completion.
   - A definition of all deliverables (e.g., data, reports, software) to be provided to the Government in support of the proposed tasks/subtasks.

x. Schedule and Milestones: Provide a detailed schedule showing tasks (task name, duration, work breakdown structure element as applicable, performing organization), milestones, and the interrelationships among tasks. The task structure must be consistent with that in the SOW. Measurable milestones should be clearly articulated and defined in time relative to the start of the project.

xi. Cost Summary: Provide the cost summary as described in Section IV.B.1.b.ii.

xii. Appendix A: This section is mandatory and must include all of the following components. If a particular subsection is not applicable, state “NONE.”

   (1). Team Member Identification: Provide a list of all team members including the prime, subcontractor(s), and consultant(s), as applicable. Identify specifically whether any are a non-US organization or individual, FFRDC and/or Government entity. Use the following format for this list:
<table>
<thead>
<tr>
<th>Individual Name</th>
<th>Role (Prime, Subcontractor or Consultant)</th>
<th>Organization</th>
<th>Non-US? Org.</th>
<th>Non-US? Ind.</th>
<th>FFRDC or Govt?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(2). **Government or FFRDC Team Member Proof of Eligibility to Propose**: If none of the team member organizations (prime or subcontractor) are a Government entity or FFRDC, state “NONE.”

If any of the team member organizations are a Government entity or FFRDC, provide documentation (per Section III.A.1) citing the specific authority that establishes the applicable team member’s eligibility to propose to Government solicitations to include: 1) statutory authority; 2) contractual authority; 3) supporting regulatory guidance; and 4) evidence of agency approval for applicable team member participation.

(3). **Government or FFRDC Team Member Statement of Unique Capability**: If none of the team member organizations (prime or subcontractor) are a Government entity or FFRDC, state “NONE.”

If any of the team member organizations are a Government entity or FFRDC, provide a statement (per Section III.A.1) that demonstrates the work to be performed by the Government entity or FFRDC team member is not otherwise available from the private sector.

(4). **Organizational Conflict of Interest Affirmations and Disclosure**: If none of the proposed team members is currently providing SETA or similar support as described in Section III.B, state “NONE.”

If any of the proposed team members (individual or organization) is currently performing SETA or similar support, furnish the following information:

<table>
<thead>
<tr>
<th>Prime Contract Number</th>
<th>DARPA Technical Office supported</th>
<th>A description of the action the proposer has taken or proposes to take to avoid, neutralize, or mitigate the conflict</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(5). Intellectual Property (IP): If no IP restrictions are intended, state “NONE.” The Government will assume unlimited rights to all IP not explicitly identified as restricted in the proposal.

For all technical data or computer software that will be furnished to the Government with other than unlimited rights, provide (per Section VI.B.1) a list describing all proprietary claims to results, prototypes, deliverables or systems supporting and/or necessary for the use of the research, results, prototypes and/or deliverables. Provide documentation proving ownership or possession of appropriate licensing rights to all patented inventions (or inventions for which a patent application has been filed) to be used for the proposed project. The following format should be used for these lists:

<table>
<thead>
<tr>
<th>Technical Data and/ or Computer Software To be Furnished With Restrictions</th>
<th>Summary of Intended Use in the Conduct of the Research</th>
<th>Basis for Assertion</th>
<th>Asserted Rights Category</th>
<th>Name of Person Asserting Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(List)</td>
<td>(Narrative)</td>
<td>(List)</td>
<td>(List)</td>
<td>(List)</td>
</tr>
<tr>
<td>(List)</td>
<td>(Narrative)</td>
<td>(List)</td>
<td>(List)</td>
<td>(List)</td>
</tr>
</tbody>
</table>

(6). Human Subjects Research (HSR): If HSR is not a factor in the proposal, state “NONE.” If the proposed work will involve human subjects, provide evidence of or a plan for review by an institutional review board (IRB). For further information on this subject, see Section VI.B.2.
(7). Animal Use: If animal use is not a factor in the proposal, state “NONE.”
If the proposed research will involve animal use, provide a brief description of the plan for Institutional Animal Care and Use Committee (IACUC) review and approval. For further information on this subject, see Section VI.B.3.

(8). Representations Regarding Unpaid Delinquent Tax Liability or a Felony Conviction under Any Federal Law: Per Section VI.B.10, complete the following statements.

(a) The proposer represents that it is [ ] is not [ ] a corporation that has any unpaid Federal tax liability that has been assessed, for which all judicial and administrative remedies have been exhausted or have lapsed, and that is not being paid in a timely manner pursuant to an agreement with the authority responsible for collecting the tax liability.

(b) The proposer represents that it is [ ] is not [ ] a corporation that was convicted of a felony criminal violation under Federal law within the preceding 24 months.

(9). Cost Accounting Standards (CAS) Notices and Certification: Per Section VI.B.11, any proposer who submits a proposal which, if accepted, will result in a CAS-compliant contract, must include a Disclosure Statement as required by 48 CFR 9903.202. The disclosure forms may be found at http://www.whitehouse.gov/omb/procurement_casb. If this section is not applicable, state "NONE."

(10). Subcontractor Plan: Pursuant to Section 8(d) of the Small Business Act (15 USC § 637(d)), it is Government policy to enable small business and small disadvantaged business concerns to be considered fairly as subcontractors to organizations performing work as prime contractors or subcontractors under Government contracts, and to ensure that prime contractors and subcontractors carry out this policy. If applicable, prepare a subcontractor plan in accordance with FAR 19.702(a) (1) and (2). The plan format is outlined in FAR 19.704.

If this section is not applicable, state “NONE.”

xiii. Appendix B: If desired, include a brief bibliography with links to relevant papers, reports, or resumes. Technical papers may be included. This section is optional, and the linked materials will not be evaluated as part of the proposal review.

b. Volume 2 - Cost Proposal
This volume is mandatory and must include all the listed components. No page limit is specified for this volume.
The cost proposal should include a spreadsheet file (.xls or equivalent format) that provides formula traceability among all components of the cost proposal. The spreadsheet file must be included as a separate component of the full proposal package. Costs must be traceable between the prime and subcontractors/consultants, as well as between the cost proposal and the SOW.
Pre-award costs will not be reimbursed unless a pre-award cost agreement is negotiated prior to award.
i. **Cover Sheet**: Include the same information as the cover sheet for Volume 1, but with the label “Proposal: Volume 2.”

ii. **Cost Summary**: Provide a single-page summary broken down by fiscal year listing cost totals for labor, materials, other direct charges (ODCs), indirect costs (overhead, fringe, general and administrative (G&A)), and any proposed fee for the project. Include costs for each task in each year of the project by prime and major subcontractors, total cost and proposed cost share, if applicable. Include any requests for Government-furnished equipment or information with cost estimates (if applicable) and delivery dates.

iii. **Cost Details**: For each task, provide the following cost details by month. Identify any cost sharing. Include supporting documentation describing the method used to estimate costs.

   1. **Direct Labor**: Provide labor categories, rates and hours. Justify rates by providing examples of equivalent rates for equivalent talent, past commercial or Government rates or Defense Contract Audit Agency (DCAA) approved rates.

   2. **Indirect Costs**: Identify all indirect cost rates (such as fringe benefits, labor overhead, material overhead, G&A, etc.) and the basis for each.

   3. **Materials**: Provide an itemized list of all proposed materials, equipment, and supplies for each year including quantities, unit prices, proposed vendors (if known), and the basis of estimate (e.g., quotes, prior purchases, catalog price lists, etc.). For proposed equipment/information technology (as defined in FAR 2.101) purchases equal to or greater than $50,000, include a letter justifying the purchase. Include any requests for Government-furnished equipment or information with cost estimates (if applicable) and delivery dates.

   4. **Travel**: Provide a breakout of travel costs including the purpose and number of trips, origin and destination(s), duration, and travelers per trip.

   5. **Subcontractor/Consultant Costs**: Provide above info for each proposed subcontractor/consultant. Subcontractor cost proposals must include interdivisional work transfer agreements or similar arrangements.

   6. **ODCs**: Provide an itemized breakout and explanation of all other anticipated direct costs.

   The proposer is responsible for the compilation and submission of all subcontractor/consultant cost proposals. Proposal submissions will not be considered complete until the Government has received all subcontractor/consultant cost proposals.

   Proprietary subcontractor/consultant cost proposals may be included as part of Volume 2 or emailed separately to BigMechanism@darpa.mil. Email messages must include “Subcontractor Cost Proposal” in the subject line and identify the principal investigator, prime proposer organization and proposal title in the body of the message.

iv. **Proposals Requesting a Procurement Contract**: Provide the following information where applicable.
(1) **Proposals for $700,000 or more**: Provide “certified cost or pricing data” (as defined in FAR 2.101) or a request for exception in accordance with FAR 15.403.

(2) **Proposers without a DCAA-approved cost accounting system**: If requesting a cost-type contract, provide the DCAA Pre-award Accounting System Adequacy Checklist to facilitate DCAA’s completion of an SF 1408. The checklist may be found at [http://www.dcaa.mil/preaward_accounting_checklist.html](http://www.dcaa.mil/preaward_accounting_checklist.html)

**v. Proposals Requesting an Other Transaction for Prototypes (845 OT) agreement**: Proposers must indicate whether they qualify as a nontraditional Defense contractor, have teamed with a nontraditional Defense contractor, or are providing a one-third cost share for this effort. Provide information to support the claims.

Provide a detailed list of milestones including: description, completion criteria, due date, and payment/funding schedule (to include, if cost share is proposed, contractor and Government share amounts). Milestones must relate directly to accomplishment of technical metrics as defined in the solicitation and/or the proposal. While agreement type (fixed price or expenditure based) will be subject to negotiation, the use of fixed price milestones with a payment/funding schedule is preferred. Proprietary information must not be included as part of the milestones.

**2. Proprietary and Classified Information**

DARPA policy is to treat all submissions as source selection information (see FAR 2.101 and 3.104) and to disclose the contents only for the purpose of evaluation. Restrictive notices notwithstanding, during the evaluation process, submissions may be handled by support contractors for administrative purposes and/or to assist with technical evaluation. All DARPA support contractors performing this role are expressly prohibited from performing DARPA-sponsored technical research and are bound by appropriate nondisclosure agreements.

**a. Proprietary Information**

Proposers are responsible for clearly identifying proprietary information. Submissions containing proprietary information must have the cover page and each page containing such information clearly marked. Proprietary information must not be included in the proposed schedule, milestones, or SOW.

**b. Classified Information**

DARPA anticipates that most submissions received under this solicitation will be unclassified; however, classified submissions will be accepted. Classified submissions must be appropriately and conspicuously marked with the proposed classification level and declassification date. Use classification and marking guidance provided by the DoD Information Security Manual (DoDM 5200.1, Volumes 1-4) and the National Industrial Security Program Operating Manual (DoD 5220.22-M). When marking information previously classified by another Original Classification Authority (OCA), also use the applicable security classification guides. Classified submissions must indicate the classification level of not only the submitted materials, but also the anticipated classification level of the award document.
If a proposer believes a submission contains classified information (as defined by Executive Order 13526), but requires DARPA to make a final classification determination, the information must be marked and protected as though classified at the appropriate classification level (as defined by Executive Order 13526). Submissions requesting DARPA to make a final classification determination shall be marked as follows:

“CLASSIFICATION DETERMINATION PENDING. Protect as though classified ____________________ [insert the recommended classification level, e.g., Confidential, Secret, or Top Secret].”

Proposers submitting classified proposals or requiring access to classified information during the lifecycle of the project shall ensure all industrial, personnel, and information system processing security requirements (e.g., facility clearance, personnel security clearance, certification and accreditation) are in place and at the appropriate level, and any foreign ownership control and influence issues are mitigated prior to submission or access. Proposers must have existing, approved capabilities (personnel and facilities) prior to award to perform research and development at the classification level proposed. Additional information on these subjects is at http://www.dss.mil.

Classified submissions will not be returned. The original of each classified submission received will be retained at DARPA, and all other copies destroyed. A destruction certificate will be provided if a formal request is received by DARPA within 5 days of notification of non-selection.

If a determination is made that the award instrument may result in access to classified information, a DD Form 254, “DoD Contract Security Classification Specification,” will be issued by DARPA and attached as part of the award. A DD Form 254 will not be provided to proposers at the time of submission. For reference, the DD Form 254 template is available at http://www.dtic.mil/dtic/pdf/formsNguides/dd0254.pdf.

C. Submission Dates and Times

Proposers are warned that submission deadlines as outlined herein are strictly enforced. Note: some proposal requirements may take from 1 business day to 1 month to complete. See the proposal checklist in Section VIII.C for further information.

DARPA will acknowledge receipt of complete submissions via email and assign control numbers that should be used in all further correspondence regarding submissions. Note: these acknowledgements will not be sent until after the due date(s) as outlined herein.

Failure to comply with the submission procedures outlined herein may result in the submission not being evaluated.

1. Proposals

The proposal package—full proposal (Volume 1 and 2) and, as applicable, encryption password, proprietary subcontractor cost proposals, classified appendices to unclassified proposals—must be submitted per the instructions outlined herein and received by DARPA no later than March 18, 2014, at 1200 noon (EST). Submissions received after this time will not be reviewed.
D. Funding Restrictions

Not applicable.

E. Other Submission Requirements

1. Unclassified Submission Instructions

Proposers must submit all parts of their submission package using the same method; submissions cannot be sent in part by one method and in part by another method, nor should duplicate submissions be sent by multiple methods. Email submissions will not be accepted.

a. Proposals Requesting a Procurement Contract or Other Transaction

DARPA/I2O will employ an electronic web-based upload submission system for UNCLASSIFIED proposals seeking a procurement contract or OT under this solicitation. For each proposal submission, proposers must complete an online cover sheet in the DARPA/I2O Solicitation Submission System (https://www.i2osupport.csc.com/aa/index.asp). Upon completion of the online cover sheet, a confirmation screen will appear which includes instructions on uploading the proposal.

If a proposer intends to submit more than one proposal, a unique user ID and password must be used in creating each cover sheet or subsequent uploads will overwrite previous ones. Once each upload is complete, a confirmation will appear and should be printed for the proposer’s records.

All uploaded proposals must be zipped with a WinZip-compatible format and encrypted using 256-bit key AES encryption. Only one zipped/encrypted file will be accepted per submission. Submissions which are not zipped/encrypted will be rejected by DARPA. At the time of submission, an encryption password form (https://www.i2osupport.csc.com/baa/password.doc) must be completed and emailed to BigMechanism@darpa.mil with the word “PASSWORD” in the subject line of the email. Failure to provide the encryption password will result in the submission not being evaluated.

Since proposers may encounter heavy traffic on the web server, they should not wait until the day proposals are due to fill out a cover sheet and upload the submission. Technical support for web server/submission issues may be directed to BAATechHelp@darpa.mil. Technical support is typically available during regular business hours (9:00 AM – 5:00 PM ET, Monday – Friday).

b. Proposals Requesting a Grant or Cooperative Agreement

Proposers requesting grants or cooperative agreements may submit proposals through one of the following methods: (1) mailed directly to DARPA; or (2) electronic upload per the instructions at http://www.grants.gov/applicants/app...or-grants.html. Proposers choosing to mail proposals to DARPA must include one paper copy and one electronic copy of the full proposal package.

Grants.gov requires proposers to complete a one-time registration process before a proposal can be electronically submitted. If proposers have not previously registered, this process can take between three business days and four weeks if all steps are not completed in a timely manner. See the Grants.gov user
guides and checklists at http://www.grants.gov/web/grants/app...resources.html and http://www.grants.gov/documents/19/1...gChecklist.pdf for further information.

Once Grants.gov has received an uploaded proposal submission, Grants.gov will send two email messages to notify proposers that: (1) their submission has been received by Grants.gov; and (2) the submission has been either validated or rejected by the system. It may take up to two business days to receive these emails. If the proposal is rejected by Grants.gov, it must be corrected and re-submitted before DARPA can retrieve it (assuming the solicitation has not expired). If the proposal is validated, then the proposer has successfully submitted their proposal and Grants.gov will notify DARPA. Once the proposal is retrieved by DARPA, Grants.gov will send a third email to notify the proposer. The proposer will then receive an email from DARPA acknowledging receipt and providing a control number.

To avoid missing deadlines, proposers should submit their proposals to Grants.gov in advance of the proposal due date, with sufficient time to complete the registration and submission processes, receive email notifications and correct errors, as applicable.

Technical support for Grants.gov submissions may be reached at 1-800-518-4726 and support@grants.gov.

2. Classified Submission Instructions

Classified materials must be submitted in accordance with the guidelines outlined herein and must not be submitted electronically by any means, including the electronic web-based system or Grants.gov, as described above. Classified submissions must be transmitted per the classification guidance provided by the DoD Information Security Manual (DoDM 5200.1, Volumes 1-4) and the National Industrial Security Program Operating Manual (DoDM 5220.22-M). If submissions contain information previously classified by another OCA, proposers must also follow any applicable SCGs when transmitting their documents. Applicable classification guide(s) must be included to ensure the submission is protected at the appropriate classification level.

a. Confidential and Collateral Secret Information

Classified information at the Confidential or Secret level must be submitted by one of the following methods:

− Hand carried by an appropriately cleared and authorized courier to DARPA. Prior to traveling, the courier must contact the DARPA Classified Document Registry (CDR) at 703-526-4052 to coordinate arrival and delivery.

or

− Mailed by U.S. Postal Service Registered Mail or Express Mail.

All classified information will be enclosed in opaque inner and outer covers and double wrapped. The inner envelope must be sealed and plainly marked with the assigned classification and addresses of both sender and addressee. The inner envelope must be addressed to:
The outer envelope must be sealed without identification as to the classification of its contents and addressed to:

Defense Advanced Research Projects Agency
Security and Intelligence Directorate, Attn: CDR
675 North Randolph Street
Arlington, VA 22203-2114

b. Top Secret (TS) Information
TS information must be hand carried, by appropriately cleared and authorized courier(s), to DARPA. Prior to traveling, the courier(s) must contact the DARPA CDR at 703-526-4052 for instructions.

c. Special Access Program (SAP) Information
SAP information must be transmitted by approved methods only. Prior to submission, contact the DARPA Special Access Program Control Office at 703-526-4052 for instructions.

d. Sensitive Compartmented Information (SCI)
SCI must be transmitted by approved methods only. Prior to submission, contact the DARPA Special Security Office at 703-526-4052 for instructions.

V. APPLICATION REVIEW INFORMATION

A. Evaluation Criteria

Proposals will be evaluated using the following criteria listed in descending order of importance: Overall Scientific and Technical Merit; Potential Contribution and Relevance to the DARPA Mission; Clarity of Claims and Evaluation Plans; and Cost Realism.

− Overall Scientific and Technical Merit: The proposed technical approach is feasible, achievable, complete and supported by a proposed technical team that has the expertise and experience to accomplish the proposed tasks. The task descriptions and associated technical elements are complete and in a logical sequence, with all proposed deliverables clearly defined such that a viable attempt to achieve project goals is likely as a result of award. The proposal identifies major technical risks and clearly defines feasible mitigation efforts.

The Big Mechanism program is classified as fundamental research, so successful proposals will demonstrate scientific and technical merit by articulating an innovative vision of how Big Mechanisms will improve how the science of very complicated systems is done. This vision will be well-supported by prior research. (Proposers are asked to include previous publications, and the References sections of their proposals do not count against page limits.) Additional refinement of this evaluation criterion is provided below:
• The extent to which proposed work advances the state of the art in a chosen area (e.g., advancing the semantics of NLP, or advancing curation methods, or advancing consistency checking, etc.);
• The extent to which proposed work will take advantage of extant resources, particularly prior knowledge in the form of ontologies, databases and so on; and,
• The clarity of proposers’ visions of the roles of their technologies in an integrated Big Mechanism system, if they propose component technologies rather than an entire system.

− Potential Contribution and Relevance to the DARPA Mission: The potential contributions of the proposed project are relevant to the national technology base. Specifically, DARPA’s mission is to maintain the technological superiority of the U.S. military and prevent technological surprise from harming national security by sponsoring revolutionary, high-payoff research that bridges the gap between fundamental discoveries and their application. This includes considering the extent to which any proposed intellectual property restrictions will potentially impact the Government’s ability to transition the technology.

− Clarity of Claims and Evaluation Plans: Proposals must concisely state the research and technology claims of the proposal, and describe clearly how the proposed approach will be evaluated. Evaluation plans need to be executable, with a clear method and data analysis plan.

− Cost Realism: The proposed costs are based on realistic assumptions, reflect a sufficient understanding of the technical goals and objectives of the solicitation, and are consistent with the proposer’s technical/management approach (to include the proposed SOW). The costs for the prime and subcontractors/consultants are substantiated by the details provided in the proposal (e.g., the type and number of labor hours proposed per task, the types and quantities of materials, equipment and fabrication costs, travel and any other applicable costs).

B. Review and Selection Process

DARPA policy is to ensure impartial, equitable, and comprehensive proposal evaluations and to select proposals that meet DARPA technical, policy, and programmatic goals.

Qualified Government personnel will conduct a scientific and technical review of each conforming proposal and (if necessary) convene panels of experts in the appropriate areas. Subject to the restrictions set forth in FAR 37.203(d), input on technical aspects of the proposals may be solicited by DARPA from non-Government consultants/experts who are strictly bound by appropriate nondisclosure agreements/requirements.

The review process identifies proposals that meet the established criteria and are, therefore, selectable for negotiation of funding awards by the Government. Selections under this solicitation will be made to proposers on the basis of the evaluation criteria listed in Section V.A. Proposals that are determined to be selectable will not necessarily receive awards. Selections may be made at any time during the period of solicitation.

Proposals are evaluated individually, not rated competitively against other proposals because they are not submitted in accordance with a common work statement. For purposes of evaluation, a proposal is defined to be the document and supporting materials as described in Section IV. DARPA’s intent is to review submissions as soon as possible after they arrive; however, submissions may be reviewed periodically for administrative reasons.
Failure to comply with the submission procedures may result in the submission not being evaluated. No submissions, classified or unclassified, will be returned. After proposals have been evaluated and selections made, the original of each proposal will be retained at DARPA. Hard copies will be destroyed.

VI. AWARD ADMINISTRATION INFORMATION

A. Selection Notices

After proposal evaluations are complete, proposers will be notified as to whether their proposal was selected for funding as a result of the review process. Notification will be sent by email to the technical and administrative POCs identified on the proposal cover sheet. If a proposal has been selected for award negotiation, the Government will initiate those negotiations following the notification.

B. Administrative and National Policy Requirements

1. Intellectual Property

Proposers should note that the Government does not own the intellectual property of technical data/computer software developed under Government contracts; it acquires the right to use the technical data/computer software. Regardless of the scope of the Government’s rights, performers may freely use their same data/software for their own commercial purposes (unless restricted by U.S. export control laws or security classification). Therefore, technical data and computer software developed under this solicitation will remain the property of the performers, though DARPA will have a minimum of Government Purpose Rights (GPR) to software developed through DARPA sponsorship.

To the greatest extent feasible, proposers should not include background proprietary software and technical data as the basis of their proposed approach. If proposers desire to use proprietary software or technical data or both as the basis of their proposed approach, in whole or in part, they should: 1) clearly identify such software/data and its proposed particular use(s); 2) explain how the Government will be able to reach its program goals (including transition) within the proprietary model offered; and 3) provide possible nonproprietary alternatives in any area that might present transition difficulties or increased risk or cost to the Government under the proposed proprietary solution.

Proposers expecting to use, but not to deliver, commercial open source tools or other materials in implementing their approach may be required to indemnify the Government against legal liability arising from such use.

All references to "Unlimited Rights" or "Government Purpose Rights" are intended to refer to the definitions of those terms as set forth in the Defense Federal Acquisition Regulation Supplement (DFARS) 227.

a. Intellectual Property Representations

All proposers must provide a good faith representation of either ownership or possession of appropriate licensing rights to all other intellectual property to be used for the proposed project. Proposers must provide a short summary for each item asserted with less than unlimited rights that describes the nature of the restriction and the intended use of the intellectual property in the conduct of the proposed research.
b. Patents

All proposers must include documentation proving ownership or possession of appropriate licensing rights to all patented inventions to be used for the proposed project. If a patent application has been filed for an invention, but it includes proprietary information and is not publicly available, a proposer must provide documentation that includes: the patent number, inventor name(s), assignee names (if any), filing date, filing date of any related provisional application, and summary of the patent title, with either: (1) a representation of invention ownership, or (2) proof of possession of appropriate licensing rights in the invention (i.e., an agreement from the owner of the patent granting license to the proposer).

c. Procurement Contracts

- Noncommercial Items (Technical Data and Computer Software): Proposers requesting a procurement contract must list all noncommercial technical data and computer software that it plans to generate, develop, and/or deliver, in which the Government will acquire less than unlimited rights and to assert specific restrictions on those deliverables. In the event a proposer does not submit the list, the Government will assume that it has unlimited rights to all noncommercial technical data and computer software generated, developed, and/or delivered, unless it is substantiated that development of the noncommercial technical data and computer software occurred with mixed funding. If mixed funding is anticipated in the development of noncommercial technical data and computer software generated, developed, and/or delivered, proposers should identify the data and software in question as subject to GPR. In accordance with DFARS 252.227-7013, “Rights in Technical Data - Noncommercial Items,” and DFARS 252.227-7014, “Rights in Noncommercial Computer Software and Noncommercial Computer Software Documentation,” the Government will automatically assume that any such GPR restriction is limited to a period of 5 years, at which time the Government will acquire unlimited rights unless the parties agree otherwise. The Government may use the list during the evaluation process to evaluate the impact of any identified restrictions and may request additional information from the proposer, as may be necessary, to evaluate the proposer’s assertions. Failure to provide full information may result in a determination that the proposal is not compliant with the solicitation. A template for complying with this request is provided in Section IV.B.1.a.xii.(5).

- Commercial Items (Technical Data and Computer Software): Proposers requesting a procurement contract must list all commercial technical data and commercial computer software that may be included in any noncommercial deliverables contemplated under the research project, and assert any applicable restrictions on the Government’s use of such commercial technical data and/or computer software. In the event a proposer does not submit the list, the Government will assume there are no restrictions on the Government’s use of such commercial items. The Government may use the list during the evaluation process to evaluate the impact of any identified restrictions and may request additional information from the proposer to evaluate the proposer’s assertions. Failure to provide full information may result in a determination that the proposal is not compliant with the solicitation. A template for complying with this request is provided in Section IV.B.1.a.xii.(5).

d. Other Types of Awards

Proposers responding to this solicitation requesting an award instrument other than a procurement contract shall follow the applicable rules and regulations governing these various award instruments, but in all cases should appropriately identify any potential restrictions on the Government’s use of any intellectual property contemplated
under those award instruments in question. This includes both noncommercial items and commercial items. The Government may use the list as part of the evaluation process to assess the impact of any identified restrictions, and may request additional information from the proposer, to evaluate the proposer’s assertions. Failure to provide full information may result in a determination that the proposal is not compliant with the solicitation. A template for complying with this request is provided in Section IV.B.1.a.xii.(5).

2. Human Subjects Research (HSR)

All research selected for funding involving human subjects, to include the use of human biological specimens and human data, must comply with Federal regulations for human subject protection. Further, research involving human subjects that is conducted or supported by the DoD must comply with 32 CFR 219, “Protection of Human Subjects” and DoD Instruction 3216.02, “Protection of Human Subjects and Adherence to Ethical Standards in DoD-Supported Research.”

Institutions awarded funding for research involving human subjects must provide documentation of a current Assurance of Compliance with Federal regulations for human subject protection, such as a Department of Health and Human Services, Office of Human Research Protection Federal Wide Assurance. All institutions engaged in human subject research, to include subcontractors, must have a valid Assurance. In addition, all personnel involved in human subject research must provide documentation of completion of HSR training.

For all research that will involve human subjects in the first year or phase of the project, the institution must submit evidence of or a plan for review by an institutional review board (IRB) as part of the proposal. The IRB conducting the review must be the IRB identified on the institution’s Assurance of Compliance. The protocol, separate from the proposal, must include a detailed description of the research plan, study population, risks and benefits of study participation, recruitment and consent process, data collection, and data analysis. The designated IRB should be consulted for guidance on writing the protocol. The informed consent document must comply with 32 CFR 219.116. A valid Assurance of Compliance with human subjects protection regulations and evidence of appropriate training by all investigators and personnel should accompany the protocol for review by the IRB.

In addition to a local IRB approval, a headquarters-level human subjects administrative review and approval is required for all research conducted or supported by the DoD. The Army, Navy, or Air Force office responsible for managing the award can provide guidance and information about their component’s headquarters-level review process. Confirmation of a current Assurance of Compliance and appropriate human subjects protection training is required before headquarters-level approval can be issued.

The time required to complete the IRB review/approval process will vary depending on the complexity of the research and the level of risk to study participants. The IRB approval process can last 1 to 3 months, followed by a DoD review that could last 3 to 6 months. Ample time should be allotted to complete the approval process. DoD/DARPA funding cannot be used toward HSR until all approvals are granted.

3. Animal Use

Award recipients performing research, experimentation, or testing involving the use of animals shall comply with the rules on animal acquisition, transport, care, handling, and use as outlined in:
For projects anticipating animal use, proposals should briefly describe plans for Institutional Animal Care and Use Committee (IACUC) review and approval. Animal studies in the program will be expected to comply with the “Public Health Service Policy on Humane Care and Use of Laboratory Animals.”

All award recipients must receive approval by a DoD-certified veterinarian, in addition to IACUC approval. No animal studies may be conducted using DoD/DARPA funding until the U.S. Army Medical Research and Materiel Command (USAMRMC) Animal Care and Use Review Office (ACURO) or other appropriate DoD veterinary office(s) grant approval. As a part of this secondary review process, the recipient will be required to complete and submit an ACURO Animal Use Appendix.

4. Export Control

Per DFARS 225.7901, all procurement contracts, OTs and other awards (as deemed appropriate), resultant from this solicitation will include the DFARS Export Control clause (252.225-7048).

5. Electronic and Information Technology

All electronic and information technology acquired through this solicitation must satisfy the accessibility requirements of Section 508 of the Rehabilitation Act (29 USC § 794d) and FAR 39.2. Each project involving the creation or inclusion of electronic and information technology must ensure that: (1) Federal employees with disabilities will have access to and use of information that is comparable to the access and use by Federal employees who are not individuals with disabilities; and (2) members of the public with disabilities seeking information or services from DARPA will have access to and use of information and data that is comparable to the access and use of information and data by members of the public who are not individuals with disabilities.

6. Employment Eligibility Verification

Per FAR 22.1802, recipients of FAR-based procurement contracts must enroll as Federal contractors in E-verify and use the system to verify employment eligibility of all employees assigned to the award. All resultant contracts from this solicitation will include the clause at FAR 52.222-54, “Employment Eligibility Verification.” This clause will not be included in grants, cooperative agreements, or OTs.

7. System for Award Management (SAM) Registration and Universal Identifier Requirements

Unless the proposer is exempt from this requirement, as per FAR 4.1102 or 2 CFR 25.110, as applicable, all proposers must be registered in the SAM and have a valid DUNS number prior to submitting a proposal. All proposers must provide their DUNS number in each proposal they submit. All proposers must maintain an active SAM registration with
current information at all times during which they have an active Federal award or proposal under consideration by DARPA. Information on SAM registration is available at http://www.sam.gov. Note that new registrations can take an average of 7-10 business days to process in SAM. SAM registration requires the following information:

- DUNS number
- TIN
- CAGE Code. If a proposer does not already have a CAGE code, one will be assigned during SAM registration.
- Electronic Funds Transfer information (e.g., proposer’s bank account number, routing number, and bank phone or fax number).

8. Reporting Executive Compensation and First-Tier Subcontract Awards

Per FAR 4.1403, FAR-based procurement contracts valued at $25,000 or more will include the clause at FAR 52.204-10, “Reporting Executive Compensation and First-Tier Subcontract Awards." A similar award term will be used in grants and cooperative agreements.

9. Updates of Information Regarding Responsibility Matters

Per FAR 9.104-7(c), all contracts valued at $500,000 or more, where the contractor has current active Federal contracts and grants with total value greater than $10,000,000, will include FAR clause 52.209-9, “Updates of Publicly Available Information Regarding Responsibility Matters."

10. Representation by Corporations Regarding Unpaid Delinquent Tax Liability or a Felony Conviction under Any Federal Law

In accordance with section 101(a)(3) of the Continuing Appropriations Resolution, 2013 (Pub. L. 112-175), sections 8112 and 8113 of Division C and sections 514 and 515 of Division E of the Consolidated and Further Continuing Appropriations Act, 2013 (Pub. L. 113-6), none of the funds made available by either Act for DoD use may be used to enter into a contract with any corporation that: (1) has any unpaid Federal tax liability that has been assessed, for which all judicial and administrative remedies have been exhausted or have lapsed, and that is not being paid in a timely manner pursuant to an agreement with the authority responsible for collecting the tax liability, where the awarding agency is aware of the unpaid tax liability, unless the agency has considered suspension or debarment of the corporation and made a determination that this further action is not necessary to protect the interests of the Government; or (2) was convicted of a felony criminal violation under any Federal law within the preceding 24 months, where the awarding agency is aware of the conviction, unless the agency has considered suspension or debarment of the corporation and made a determination that this action is not necessary to protect the interests of the Government. Each proposer must complete and return the representations outlined in Section IV.B.1.a.xii.(8) with their proposal submission.

11. Cost Accounting Standards (CAS) Notices and Certification

Per FAR 52.230-2, any procurement contract in excess of $700,000 resulting from this solicitation will be subject to the requirements of the Cost Accounting Standards Board (48 CFR 99), except those contracts which are exempt as
specified in 48 CFR 9903.201-1. Any proposer who submits a proposal which, if accepted, will result in a CAS-compliant contract, must include a Disclosure Statement as required by 48 CFR 9903.202. The disclosure forms may be found at http://www.whitehouse.gov/omb/procurement_casb.

12. Controlled Unclassified Information (CUI) on Non-DoD Information Systems

CUI refers to unclassified information that does not meet the standard for National Security Classification but is pertinent to the national interests of the United States or to the important interests of entities outside the Federal Government and under law or policy requires: (1) protection from unauthorized disclosure, (2) special handling safeguards, or (3) prescribed limits on exchange or dissemination. All non-DoD entities doing business with DARPA are expected to adhere to the following procedural safeguards, in addition to any other relevant Federal or DoD specific procedures, for submission of any proposals to DARPA and any potential business with DARPA:

- Do not process DARPA CUI on publicly available computers or post DARPA CUI to publicly available webpages or websites that have access limited only by domain or Internet protocol restriction.
- Ensure that all DARPA CUI is protected by a physical or electronic barrier when not under direct individual control of an authorized user and limit the transfer or DARPA CUI to subcontractors or teaming partners with a need to know and commitment to this level of protection.
- Ensure that DARPA CUI on mobile computing devices is identified and encrypted and all communications on mobile devices or through wireless connections are protected and encrypted.
- Overwrite media that has been used to process DARPA CUI before external release or disposal.

13. Safeguarding of Unclassified Controlled Technical Information (This applies only to FAR-based awards)

Per DFARS 204.7303, DFARS 252.204-7012, Safeguarding of Unclassified Controlled Technical Information, applies to this solicitation and all FAR-based awards resulting from this solicitation.

C. Reporting

1. Technical and Financial Reports

The number and types of technical and financial reports required under the contracted project will be specified in the award document, and will include, as a minimum, monthly financial status reports and a yearly status summary. A final report that summarizes the project and tasks will be required at the conclusion of the performance period for the award. The reports shall be prepared and submitted in accordance with the procedures contained in the award document.

2. Representations and Certifications

In accordance with FAR 4.1201, prospective proposers shall complete electronic annual representations and certifications at http://www.sam.gov.
3. Wide Area Work Flow (WAWF)

Unless using another means of invoicing, performers will be required to submit invoices for payment directly at https://wawf.eb.mil. If applicable, WAWF registration is required prior to any award under this solicitation.

4. i-Edison

Award documents will contain a requirement for patent reports and notifications to be submitted electronically through the i-Edison Federal patent reporting system at http://s-edison.info.nih.gov/iEdison.

VII. AGENCY CONTACTS

DARPA will use email for all technical and administrative correspondence regarding this solicitation.
- Technical POC: Paul Cohen, Program Manager, DARPA/I2O
- Email: BigMechanism@darpa.mil
- Mailing address:
  DARPA/I2O
  ATTN: DARPA-BAA-14-14 675 North Randolph Street Arlington, VA 22203-2114
- I2O Solicitation Website: http://www.darpa.mil/Opportunities/S...citations.aspx

VIII. OTHER INFORMATION

A. Frequently Asked Questions (FAQs)

Administrative, technical, and contractual questions should be sent via email to BigMechanism@darpa.mil. All questions must include the name, email address, and the telephone number of a point of contact.

DARPA will attempt to answer questions in a timely manner; however, questions submitted within 7 days of closing may not be answered. If applicable, DARPA will post FAQs to http://www.darpa.mil/Opportunities/S...citations.aspx.

B. Submission Checklist

The following items apply prior to proposal submission. Note: some items may take up to 1 month to complete.

<table>
<thead>
<tr>
<th>Check Mark</th>
<th>Item</th>
<th>BAA Section</th>
<th>Applicability</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obtain DUNS number</td>
<td>IV.B.1.a.i</td>
<td>Required of all proposers</td>
<td>The DUNS Number is the Federal Government’s contractor identification code for all procurement-related activities. See <a href="http://fedgov.dnb.com/webform/index.jsp">http://fedgov.dnb.com/webform/index.jsp</a> to...</td>
<td></td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA

Updated: Sat, 19 Sep 2015 00:07:55 GMT

Powered by mindtouch™ 83
<table>
<thead>
<tr>
<th>Obtain Taxpayer Identification Number (TIN)</th>
<th>IV.B.1.a.i</th>
<th>Required of all proposers</th>
<th>request a DUNS number. Note: requests may take at least one business day.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A TIN is used by the Internal Revenue Service in the administration of tax laws. See <a href="http://www.irs.gov/businesses/small/...=96696.00.html">http://www.irs.gov/businesses/small/...=96696.00.html</a> for information on requesting a TIN. Note: requests may take from 1 business day to 1 month depending on the method (online, fax, mail).</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Register in the System for Award Management (SAM)</td>
<td>VI.B.7</td>
<td>Required of all proposers</td>
<td>The SAM combines Federal procurement systems and the Catalog of Federal Domestic Assistance into one system. See <a href="http://www.sam.gov">www.sam.gov</a> for information and registration. Note: new registrations can take an average of 7-10 business days. SAM registration requires the following information: -DUNS number -TIN -CAGE Code. A CAGE Code identifies companies doing or wishing to do business with the Federal Government. If a proposer does not already have a</td>
</tr>
<tr>
<td>Requirements</td>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>----------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>CAGE code, one will be assigned during SAM registration.</td>
<td></td>
<td>-Electronic Funds Transfer information (e.g., proposer’s bank account number, routing number, and bank phone or fax number).</td>
<td></td>
</tr>
<tr>
<td>Register in E-Verify</td>
<td>VI.B.6</td>
<td>Required for proposers requesting procurement contracts</td>
<td></td>
</tr>
<tr>
<td>E-Verify is a web-based system that allows businesses to determine the eligibility of their employees to work in the United States. See <a href="http://www.uscis.gov/e-verify">http://www.uscis.gov/e-verify</a> for information and registration.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensure representations and certifications are up to date in the SAM</td>
<td>VI.B.7</td>
<td>Required of all proposers</td>
<td></td>
</tr>
<tr>
<td>Federal provisions require entities to represent/certify to a variety of statements ranging from environmental rules compliance to entity size representation. See <a href="http://www.sam.gov">http://www.sam.gov</a> for information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensure eligibility of all team members (primes, subcontractors, consultants)</td>
<td>III</td>
<td>Required of all proposers</td>
<td></td>
</tr>
<tr>
<td>Verify there are no potential OCIs. Verify eligibility, as applicable, for FFRDCs and Government entities.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Register at Grants.gov</td>
<td>IV.E.1.b</td>
<td>Required for proposers requesting grants or cooperative agreements</td>
<td></td>
</tr>
<tr>
<td>Grants.gov requires proposers to complete a one-time registration process before a proposal can be electronically submitted.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
submitted. If proposers have not previously registered, this process can take between three business days and four weeks if all steps are not completed in a timely manner. See the Grants.gov user guides and checklists at http://www.grants.gov/web/grants/app...resources.html for further information.

The following items apply as part of the submission package:

<table>
<thead>
<tr>
<th>Check Mark</th>
<th>Item</th>
<th>BAA Section</th>
<th>Applicability</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Encryption password</td>
<td>IV.E.1.a</td>
<td>Required of proposers using the DARPA/I2O electronic web-based BAA submission system.</td>
<td>Email to <a href="mailto:BigMechanism@darpa.mil">BigMechanism@darpa.mil</a></td>
</tr>
<tr>
<td></td>
<td>Volume 1 (Technical and Management Proposal)</td>
<td>IV.B.1</td>
<td>Required of all proposers</td>
<td>Conform to stated page limits and formatting requirements. Include all requested information.</td>
</tr>
</tbody>
</table>
|            | Appendix A | IV.B.1.a.xiii | Required of all proposers | - Team member identification 
- Government/FFRDC team member proof of eligibility 
- Organizational conflict of interest affirmations 
- Intellectual |
<table>
<thead>
<tr>
<th>Volume 2 (Cost Proposal)</th>
<th>IV.B.1.b</th>
<th>Required of all proposers</th>
</tr>
</thead>
</table>

- Human subjects research
- Animal use
- Subcontractor plan, if applicable
- Unpaid delinquent tax liability/felony conviction representations
- CASB disclosure, if applicable

- Cover Sheet
- Cost summary
- Cost details by task
- Include costs for direct labor, indirect costs/rates, materials/equipment, subcontractors/consultants, travel, other direct costs
- Include basis of estimates.
- Travel cost estimate to include purpose, departure/arrival destinations, and sample airfare.
- Itemized list of material and equipment items to be purchased including vendor quotes or engineering estimates for material and equipment exceeding $50,000
- If applicable, provide
- Subcontractor cost proposals
- Consultant agreements, teaming agreements or letters of intent
- List of milestones
for 845 OTA agreements
- Cost spreadsheet file (.xls or equivalent format)

Footnotes

1
Information on award instruments can be found at http://www.darpa.mil/Opportunities/C...anagement.aspx.

2
The DUNS number is used as the Government's contractor identification code for all procurement-related activities. Go to http://fedgov.dnb.com/webform/index.jsp to request a DUNS number (may take at least one business day). See Section VI.B.8. for further information.

3
See http://www.irs.gov/businesses/small/...=96696,00.html for information on requesting a TIN. Note, requests may take from 1 business day to 1 month depending on the method (online, fax, mail).

4
A CAGE Code identifies companies doing or wishing to do business with the Federal Government. See Section VI.B.7 for further information.

5

6
http://www.dtic.mil/whs/directives/c...df/321602p.pdf

7
http://www.hhs.gov/ohrp

8
http://grants.nih.gov/grants/olaw/olaw.htm
References


DARPA Open Catalog

Source: http://www.darpa.mil/opencatalog/
Introduction

Welcome to the DARPA Open Catalog, which contains a curated list of DARPA-sponsored software and peer-reviewed publications. DARPA funds fundamental and applied research in a variety of areas including data science, cyber, anomaly detection, etc., which may lead to experimental results and reusable technology designed to benefit multiple government domains.

The DARPA Open Catalog organizes publically releasable material from DARPA programs, beginning with the XDATA program in the Information Innovation Office (I2O). XDATA is developing an open source software library for big data. DARPA has an open source strategy through XDATA and other I2O programs to help increase the impact of government investments.

DARPA is interested in building communities around government-funded software and research. If the R&D community shows sufficient interest, DARPA will continue to make available information generated by DARPA programs, including software, publications, data and experimental results. Future updates are scheduled to include components from other I2O programs such as Broad Operational Language Translation (BOLT) and Visual Media Reasoning (VMR).

The DARPA Open Catalog contains two tables:

- The Software Table lists performers with one row per piece of software. Each piece of software has a link to an external project page, as well as a link to the code repository for the project. The software categories are listed; in the case of XDATA, they are Analytics, Visualization and Infrastructure. A description of the project is followed by the applicable software license. Finally, each entry has a link to the publications from each team's software entry.
- The Publications Table contains author(s), title, and links to peer-reviewed articles related to specific DARPA programs.

Program Manager:
Dr. Christopher White
christopher.white@darpa.mil

Report a problem: opencatalog@darpa.mil

The content below has been generated by organizations that are partially funded by DARPA; the views and conclusions contained therein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

Software

See: Excel
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aptima Inc.</td>
<td>Network Query by Example</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/Aptima/pattern-matching.git">https://github.com/Aptima/pattern-matching.git</a></td>
<td>stats</td>
<td>Hadoop MapReduce-over-Hive based implementation of network query by example utilizing attributed network pattern matching.</td>
<td>ALv2</td>
</tr>
<tr>
<td>Boeing/Pitt Publications</td>
<td>SMILE-WIDE: A scalable Bayesian network library</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/SmileWide/main.git">https://github.com/SmileWide/main.git</a></td>
<td>stats</td>
<td>SMILE-WIDE is a scalable Bayesian network library. Initially, it is a version of the SMILE library, as in SMILE With Integrated Distributed Execution. The general approach has been to provide an API similar to the existing API SMILE developers use to build &quot;local,&quot; single-threaded applications. However, we provide</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Carnegie Mellon University Publications</td>
<td>Support Distribution Machines</td>
<td>Analytics</td>
<td></td>
<td></td>
<td></td>
<td>&quot;vectorized&quot; operations that hide a Hadoop-distributed implementation. Apart from invoking a few idioms like generic Hadoop command line argument parsing, these appear to the developer as if they were executed locally.</td>
<td>BSD</td>
</tr>
</tbody>
</table>

Python implementation of the nonparametric divergence estimators described by Barnabas Poczos, Liang Xiong, Jeff Schneider (2011). Nonparametric divergence estimation with applications to machine learning on distributions. Uncertainty in Artificial...
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuum Analytics</td>
<td>Blaze</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/ContinuumIO/blaze.git">https://github.com/ContinuumIO/blaze.git</a></td>
<td>Blaze is the next-generation of NumPy. It is designed as a foundational set of abstractions on which to build out-of-core and distributed</td>
<td>BSD</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
</tbody>
</table>

algorithms over a wide variety of data sources and to extend the structure of NumPy itself. Blaze allows easy composition of low level computation kernels (C, Fortran, Numba) to form complex data transformations on large datasets. In Blaze, computations are described in a high-level language (Python) but executed on a low-level runtime (outside of Python), enabling the easy mapping of high-level expertise to data without sacrificing

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch™
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuum Analytics</td>
<td>Bokeh</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/ContinuumIO/statsbokeh.git">https://github.com/ContinuumIO/statsbokeh.git</a></td>
<td>low-level performance. Blaze aims to bring Python and NumPy into the massively-multicore arena, allowing it to leverage many CPU and GPU cores across computers, virtual machines and cloud services.</td>
<td>BSD</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>--------------------------</td>
<td>-----------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Continuum Analytics</td>
<td>CDX</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/ContinuumIO/cdx.git">https://github.com/ContinuumIO/cdx.git</a></td>
<td>stats</td>
<td>high-performance interactivity over large data to thin clients.</td>
<td>BSD</td>
</tr>
<tr>
<td>Continuum Analytics</td>
<td>Numba</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/numba/numba.git">https://github.com/numba/numba.git</a></td>
<td>stats</td>
<td>Software to visualize the structure of large or complex datasets / produce guides that help users or algorithms gauge the quality of various kinds of graphs &amp; plots.</td>
<td>BSD</td>
</tr>
</tbody>
</table>

Numba is an Open Source NumPy-aware optimizing compiler for Python sponsored by Continuum Analytics, Inc. It uses the LLVM compiler infrastructure to compile Python syntax to machine code.
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>It is aware of NumPy arrays as typed memory regions and so can speed-up code using NumPy arrays. Other, less well-typed code is translated to Python C-API calls effectively removing the &quot;interpreter&quot; but not removing the dynamic indirection. Numba is also not a tracing just in time (JIT) compiler. It compiles your code before it runs either using runtime type information or type information you provide in the decorator. Numba is a mechanism</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------</td>
<td>------------</td>
<td>------------------------</td>
<td>---------------</td>
<td>-----------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Continuum Analytics and Indiana University Publications</td>
<td>Abstract Rendering</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/JosephCottam/stats">https://github.com/JosephCottam/stats</a> AbstractRendering.git</td>
<td></td>
<td>Information visualization rests on the idea that a meaningful relationship can be drawn between pixels and data. This is most often mediated by geometric entities (such as circles, squares and text) but always involves pixels eventually to display. In most systems, the pixels are tucked away under levels of abstraction.</td>
<td>BSD</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
</tbody>
</table>

Abstract Rendering takes the opposite approach: expose the pixels and gain powerful pixel-level control. This pixel-level power is a complement to many existing visualization techniques. It is an elaboration on rendering, not an analytic or projection step, so it can be used as an epilogue to many existing techniques. In standard rendering, geometric objects are projected to an image and represented on that image's discrete pixels. The source

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA

Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuum Analytics and Indiana</td>
<td>Stencil</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/JosephCottam/statsStencil.git">https://github.com/JosephCottam/statsStencil.git</a></td>
<td></td>
<td>space is an abstract canvas that contains logically continuous geometric primitives and the target space is an image that contains discrete colors. Abstract Rendering fits between these two states. It introduces a discretization of the data at the pixel-level, but not necessarily all the way to colors. This enables many pixel-level concerns to be efficiently and concisely captured.</td>
<td>BSD</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td>-------------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>University Publications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>visualization specification at a higher-level.</td>
<td></td>
</tr>
<tr>
<td>Data Tactics Corporation</td>
<td>Circuit</td>
<td>Infrastructure</td>
<td></td>
<td>2014-07</td>
<td></td>
<td>Go Circuit reduces the human development and sustenance costs of complex massively-scaled systems nearly to the level of their single-process counterparts. It is a combination of proven ideas from the Erlang ecosystem of distributed embedded devices and Go's ecosystem of Internet application development. Go Circuit extends the reach of Go's linguistic environment to multi-host/multi-process applications.</td>
<td>ALv2</td>
</tr>
</tbody>
</table>

Go Circuit reduces the human development and sustenance costs of complex massively-scaled systems nearly to the level of their single-process counterparts. It is a combination of proven ideas from the Erlang ecosystem of distributed embedded devices and Go's ecosystem of Internet application development. Go Circuit extends the reach of Go's linguistic environment to multi-host/multi-process applications.
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Tactics Corporation</td>
<td>Vowpal Wabbit</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/JohnLangford/stats/vowpal_wabbit.git">https://github.com/JohnLangford/stats/vowpal_wabbit.git</a></td>
<td></td>
<td>The Vowpal Wabbit (VW) project is a fast out-of-core learning system sponsored by Microsoft Research and (previously) Yahoo! Research. Support is available through the mailing list. There are two ways to have a fast learning algorithm: (a) start with a slow algorithm and speed it up, or (b) build an intrinsically fast learning algorithm. This project is about approach (b), and it's reached a state where it may be useful to others as a platform for</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>Draper Laboratory Publications</td>
<td>Analytic Activity Logger</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/draperlab/xdatalogger.git">https://github.com/draperlab/xdatalogger.git</a></td>
<td></td>
<td>Analytic Activity Logger is an API that creates a common message passing interface to allow heterogeneous software components</td>
</tr>
</tbody>
</table>

research and experimentation. There are several optimization algorithms available with the baseline being sparse gradient descent (GD) on a loss function (several are available). The code should be easily usable. Its only external dependence is on the boost library, which is often installed by default.
<table>
<thead>
<tr>
<th>Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgia Tech / GTRI Publications</td>
<td>libNMF: a high-performance library for nonnegative matrix factorization and hierarchical clustering</td>
<td>Analytics</td>
<td>2014-07</td>
<td>Pending</td>
<td>LibNMF is a high-performance, parallel library for nonnegative matrix factorization on both dense and sparse matrices</td>
<td>ALv2</td>
<td></td>
</tr>
<tr>
<td>Harvard and Kitware, Inc. Publications</td>
<td>LineUp</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/Caleydo/org.caley...neup.demos.git">https://github.com/Caleydo/org.caley...neup.demos.git</a></td>
<td>LineUp is a novel and scalable visualization technique that uses bar charts. This interactive technique supports</td>
<td>BSD</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>the ranking of items based on multiple heterogeneous attributes with different scales and semantics. It enables users to interactively combine attributes and flexibly refine parameters to explore the effect of changes in the attribute combination. This process can be employed to derive actionable insights as to which attributes of an item need to be modified in order for its rank to change. Additionally, through integration of slope graphs, LineUp can also be used to compare multiple</td>
<td></td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA

Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch™
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvard and Kitware, Inc. Publications</td>
<td>LineUp Web</td>
<td>Visualization</td>
<td></td>
<td>2014-07</td>
<td>2014-06</td>
<td>alternative rankings on the same set of items, for example, over time or across different attribute combinations. We evaluate the effectiveness of the proposed multi-attribute visualization technique in a qualitative study. The study shows that users are able to successfully solve complex ranking tasks in a short period of time.</td>
<td>BSD</td>
</tr>
</tbody>
</table>

LineUpWeb is the web version of the novel and scalable visualization technique. This interactive

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA

Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch™
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lindenstrauss random projection (JL); a faster version of JL based on fast transform techniques; sparse techniques that can be applied in time proportional to the number of nonzero matrix entries; and methods for approximating kernel functions and Gram matrices arising in nonlinear statistical modeling problems. We have a library of such sketching techniques, built using MPI in C++ and callable from Python, and are applying the library to</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>-----------------------------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>--------------------------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Kitware, Inc.</td>
<td>Tangelo</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/Kitware/tangelo.git">https://github.com/Kitware/tangelo.git</a></td>
<td>stats</td>
<td>igraph provides a fast generation of large graphs, fast approximate computation of local graph invariants, fast parallelizable graph embedding. API and Web-service for batch processing graphs across formats.</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------</td>
<td>----------</td>
<td>------------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>MDA Information Systems, Inc., Jet Propulsion Laboratory, USC/ Information Sciences Institute</td>
<td>OODT</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://svn.apache.org/repos/asf/oodt/">https://svn.apache.org/repos/asf/oodt/</a></td>
<td>stats</td>
<td>web server architecture that cleanly separates your web applications (pure Javascript, HTML, and CSS) and web services (pure Python). This software is bundled with some great tools to get you started.</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------</td>
<td>---------------</td>
<td>------------------------</td>
<td>--------------------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>MDA Information Systems, Inc., Jet Propulsion Laboratory, USC/ Information Sciences Institute</td>
<td>Wings</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/varunratnakar/stats/wings.git">https://github.com/varunratnakar/stats/wings.git</a></td>
<td></td>
<td>WINGS provides a semantic workflow system that assists scientists with the design of computational experiments. A unique feature of WINGS is that its workflow representations incorporate semantic constraints about datasets and workflow components, and are used to create and validate workflows and to generate metadata for new data products. WINGS submits workflows</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>----------</td>
<td>------------------------</td>
<td>------------</td>
<td>-----------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>MIT-LL Publications</td>
<td>Information Extractor</td>
<td>Analytics</td>
<td>2014-07</td>
<td>Pending</td>
<td></td>
<td>to execution frameworks such as Pegasus and OODT to run workflows at large scale in distributed resources.</td>
<td>ALv2</td>
</tr>
<tr>
<td>MIT-LL Publications</td>
<td>Julia</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/JuliaLang/julia.git">https://github.com/JuliaLang/julia.git</a></td>
<td>stats</td>
<td>Julia is a high-level, high-performance dynamic programming language for technical computing, with syntax that is familiar to users of other technical computing environments. It provides a sophisticated compiler, distributed parallel execution, numerical accuracy,</td>
<td>MIT/GPL/LGPL/BSV</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>MIT-LL Publications</td>
<td>Query By Example (Graph QuBE)</td>
<td>Analytics</td>
<td></td>
<td>2014-07</td>
<td>2014-02-15</td>
<td>Query-by-Example (Graph QuBE) on dynamic transaction graphs.</td>
<td>ALv2</td>
</tr>
<tr>
<td>MIT-LL Publications</td>
<td>SciDB</td>
<td>Infrastructure</td>
<td></td>
<td>2014-07</td>
<td></td>
<td>Scientific Database for large-scale numerical data.</td>
<td>GPLv3</td>
</tr>
<tr>
<td>Next Century Corporation</td>
<td>Neon Visualization Environment</td>
<td>Visualization</td>
<td></td>
<td>2014-07</td>
<td></td>
<td>Neon is a framework that gives a datastore agnostic way for visualizations to query data and perform simple operations on that data such as filtering, aggregation, and transforms. It is divided</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>into two parts, neon-server and neon-client. Neon-server provides a set of RESTful web services to select a datastore and perform queries and other operations on the data. Neon-client is a javascript API that provides a way to easily integrate neon-server capabilities into a visualization, and also aids in 'widgetizing' a visualization, allowing it to be integrated into a common OWF based ecosystem.</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Next Century Corporation</td>
<td><strong>Ozone Widget Framework</strong></td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/ozoneplatform/owf.git">https://github.com/ozoneplatform/owf.git</a></td>
<td></td>
<td>Ozone Widget Framework provides a customizable open-source web application that assembles the tools you need to accomplish any task and enables those tools to communicate with each other. It is a technology-agnostic composition framework for data and visualizations in a common browser-based display and interaction environment that lowers the barrier to entry for the development of big data visualizations and enables efficient exploration.</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Oculus Info Inc. Publications</td>
<td>Aperture Tile-Based Visual Analytics</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/oculusinfo/aperture-tiles.git">https://github.com/oculusinfo/aperture-tiles.git</a></td>
<td></td>
<td>New tools for raw data characterization of 'big data' are required to suggest initial hypotheses for testing. The widespread use and adoption of web-based maps has provided a familiar set of interactions for exploring abstract large data spaces. Building on these techniques, we developed tile based visual analytics that provide browser-based interactive visualization of billions of data points.</td>
<td>MIT</td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch™
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oculus Info Inc. Publications</td>
<td>ApertureJS</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/oculusinfo/aperturejs.git">https://github.com/oculusinfo/aperturejs.git</a></td>
<td></td>
<td>ApertureJS is an open, adaptable and extensible JavaScript visualization framework with supporting REST services, designed to produce visualizations for analysts and decision makers in any common web browser. Aperture utilizes a novel layer based approach to visualization assembly, and a data mapping API that simplifies the process of adaptable transformation of data and analytic results into visual forms and properties. Aperture vizlets can be easily</td>
<td>MIT</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Oculus Info Inc. Publications</td>
<td>Influent</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/oculusinfo/influent.git">https://github.com/oculusinfo/influent.git</a></td>
<td>stats</td>
<td>Influent is an HTML5 tool for visually and interactively following transaction flow, rapidly revealing actors and behaviors of potential concern that might otherwise go unnoticed. Summary visualization of transactional patterns and actor characteristics, interactive link expansion and dynamic entity clustering enable Influent to operate effectively</td>
<td>MIT</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Oculus Info Inc. Publications</td>
<td><strong>Oculus Ensemble Clustering</strong></td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/oculusinfo/ensemble-clustering.git">https://github.com/oculusinfo/ensemble-clustering.git</a></td>
<td>stats</td>
<td>at scale with big data sources in any modern web browser. Influent has been used to explore data sets with millions of entities and hundreds of millions of transactions.</td>
<td>MIT</td>
</tr>
</tbody>
</table>

Oculus Ensemble Clustering is a flexible multi-threaded clustering library for rapidly constructing tailored clustering solutions that leverage the different semantic aspects of heterogeneous data. The library can be used on a single machine using...
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phronesis</td>
<td>bigalgebra</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td></td>
<td>multi-threading or distributed computing using Spark.</td>
<td>ALv2</td>
</tr>
<tr>
<td></td>
<td>Phronesis</td>
<td>bigmemory</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td></td>
<td>Bigmemory is an R package to create, store, access, and manipulate massive matrices. Matrices are allocated to shared memory and may use memory-mapped files. Packages biganalytics, bigtabulate, synchronicity, and bigalgebra provide</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Phronesis</td>
<td>flexmem</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/kaneplusplus/flexmem.git">https://github.com/kaneplusplus/flexmem.git</a></td>
<td>stats</td>
<td>Flexmem is a general, transparent tool for out-of-core (OOC) computing in the R programming environment. It is launched as a command line utility, taking an application as an argument. All memory allocations larger than a specified threshold are memory-mapped to a binary file. When data are not needed, they are stored on disk. It is both process- and thread-safe.</td>
<td>ALv2</td>
</tr>
<tr>
<td>Phronesis</td>
<td>laputa</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/kaneplusplus/laputa.git">https://github.com/kaneplusplus/laputa.git</a></td>
<td>stats</td>
<td>Laputa is a Python package</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Phronesis</td>
<td>stat_agg</td>
<td>Analytics</td>
<td>2014-07</td>
<td></td>
<td></td>
<td>that provides an elastic, parallel computing foundation for the stat_agg (statistical aggregates) package.</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------------------------------</td>
<td>----------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Raytheon</td>
<td><strong>Content and Context-based Graph Analysis:</strong> NILS, Network Inference of Link Strength</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/plamenbbn/XDATA.git">https://github.com/plamenbbn/XDATA.git</a></td>
<td>stats</td>
<td>Network Inference of Link Strength will take any text corpus as input and quantify the strength of connections between any pair of entities. Link strength probabilities are computed via shortest path.</td>
<td>ALv2</td>
</tr>
<tr>
<td>Raytheon BBN</td>
<td><strong>Content and Context-based Graph Analysis:</strong> PINT, Patterns in Near-Real Time</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/plamenbbn/XDATA.git">https://github.com/plamenbbn/XDATA.git</a></td>
<td>stats</td>
<td>Patterns in Near-Real Time will take any corpus as input and quantify the strength of the query match to a SME-based process model, represent process model as a Directed Acyclic Graph</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------------------</td>
<td>----------</td>
<td>------------------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
</tbody>
</table>
| Royal Caliber Publications | **GPU based Graphlab style**  
**Gather-Apply-Scatter (GAS) platform for quickly implementing and running graph algorithms** | Analytics | 2014-07               | https://github.com/RoyalCaliber/vertexAPI2.git | stats     | Allows users to express graph algorithms as a series of Gather-Apply-Scatter (GAS) steps similar to GraphLab. Runs these vertex programs using a single or multiple GPUs - demonstrates a large speedup over GraphLab. | ALv2    |
<p>| Scientific Systems Company, Inc., MIT, and University of Louisville | <strong>BayesDB</strong> | Analytics | 2014-07               | <a href="https://github.com/mit-probabilistic/BayesDB.git">https://github.com/mit-probabilistic/BayesDB.git</a> | stats     | BayesDB is an open-source implementation of a predictive database table. It provides predictive extensions to SQL that enable users to query the database. | ALv2    |</p>
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crosscat</td>
<td>Analytics</td>
<td></td>
<td>2014-07</td>
<td></td>
<td>implications of their data --- predict missing entries, identify predictive relationships between columns, and examine synthetic populations --- based on a Bayesian machine learning system in the backend.</td>
<td>ALv2</td>
</tr>
</tbody>
</table>

CrossCat is a domain-general, Bayesian method for analyzing high-dimensional data tables. CrossCat estimates the full joint distribution over the variables in the table from the data via approximate inference in a hierarchical, Bayesian machine learning system.

https://github.com/mit-probabilistic/stats/crosscat.git
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>nonparametric Bayesian model, and provides efficient samplers for every conditional distribution. CrossCat combines strengths of nonparametric mixture modeling and Bayesian network structure learning: it can model any joint distribution given enough data by positing latent variables, but also discovers independencies between the observable variables.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sotera Defense Solutions, Inc.</td>
<td>ARIMA</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/Sotera/rhipe-arima">https://github.com/Sotera/rhipe-arima</a></td>
<td></td>
<td>Hive and RHIEP implementation of an ARIMA analytic.</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------</td>
<td>----------</td>
<td>------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Inc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>algorithm to find highly correlated vectors using an approximation algorithm.</td>
<td></td>
</tr>
<tr>
<td>Publications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Recursive algorithm to remove nodes from a network where degree centrality is 1.</td>
<td>ALv2</td>
</tr>
<tr>
<td>Sotera</td>
<td>Leaf Compression</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/Sotera/leaves-">https://github.com/Sotera/leaves-</a></td>
<td>stats</td>
<td>Giraph/Hadoop implementation of a distributed version of the Louvain community detection algorithm.</td>
<td>ALv2</td>
</tr>
<tr>
<td>Solutions,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inc.</td>
<td>Page Rank</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/Sotera/page-">https://github.com/Sotera/page-</a></td>
<td>stats</td>
<td>Sotera Page Rank is a Giraph/Hadoop implementation of a distributed version of the Page Rank algorithm.</td>
<td>ALv2</td>
</tr>
<tr>
<td>Publications</td>
<td></td>
<td></td>
<td></td>
<td>rank.git</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sotera</td>
<td>Spark MicroPath</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/Sotera/aggregate-">https://github.com/Sotera/aggregate-</a></td>
<td>stats</td>
<td>The Spark implementation of the</td>
<td>ALv2</td>
</tr>
<tr>
<td>Defense</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solutions,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch™
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inc. Publications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>micropath analytic.</td>
<td></td>
</tr>
<tr>
<td>Sotera Defense Solutions, Inc. Publications</td>
<td>Zephyr</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="http://github.com/Sotera/zephyr">http://github.com/Sotera/zephyr</a></td>
<td>stats</td>
<td>Zephyr is a big data, platform agnostic ETL API, with Hadoop MapReduce, Storm, and other big data bindings.</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Stanford University - Boyd Publications</td>
<td>Proximal Operators</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/cvxgrp/proximal.git">https://github.com/cvxgrp/proximal.git</a></td>
<td>ALv2</td>
<td>This library contains sample implementations of various proximal operators in Matlab. These implementations are intended to be pedagogical, not the most performant. This code is associated with the paper Proximal Algorithms by Neal Parikh and Stephen Boyd.</td>
<td></td>
</tr>
<tr>
<td>Stanford University - Boyd Publications</td>
<td>QCML (Quadratic Cone Modeling Language)</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/cvxgrp/qcml.git">https://github.com/cvxgrp/qcml.git</a></td>
<td>ALv2</td>
<td>Seamless transition from prototyping to code generation. Enable ease and expressiveness of convex optimization across scales with little</td>
<td></td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch™
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SCS (Self-dual Cone Solver)</td>
<td>Analytics</td>
<td></td>
<td></td>
<td></td>
<td>Implementation of a solver for general cone programs, including linear, second-order, semidefinite and exponential cones, based on an operator splitting method applied to a self-dual homogeneous embedding. The method and software supports both direct factorization, with factorization caching, and an indirect method, that requires only the operator associated with the problem data and its adjoint. The</td>
<td>ALv2</td>
</tr>
</tbody>
</table>

Stanford University - Boyd Publications

https://github.com/cvxgrp/scs.git

Updated: Sat, 19 Sep 2015 00:07:55 GMT

Powered by mindtouch
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford University - Hanrahan Publications</td>
<td>imMens</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/StanfordHCI/imMens.git">https://github.com/StanfordHCI/imMens.git</a></td>
<td>stats</td>
<td>imMens is a web-based system for interactive visualization of large databases. imMens uses binned aggregation to produce summary visualizations that avoid the shortcomings of standard sampling-based approaches. Through data decomposition methods</td>
<td>BSD</td>
</tr>
<tr>
<td><strong>XDATA Team</strong></td>
<td><strong>Software</strong></td>
<td><strong>Category</strong></td>
<td><strong>Instructional Material</strong></td>
<td><strong>Code</strong></td>
<td><strong>Dev Stats</strong></td>
<td><strong>Description</strong></td>
<td><strong>License</strong></td>
</tr>
<tr>
<td>----------------</td>
<td>--------------</td>
<td>--------------</td>
<td>----------------------------</td>
<td>----------</td>
<td>---------------</td>
<td>-----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Stanford University - Hanrahan Publications</td>
<td><strong>RHIPE: R and Hadoop Integrated Programming Environment</strong></td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/saptarshiguha/statsRHIPE.git">https://github.com/saptarshiguha/statsRHIPE.git</a></td>
<td>(to limit data transfer) and GPU computation via WebGL (for parallel query processing), imMens enables real-time (50fps) visual querying of billion+ element databases.</td>
<td>ALv2</td>
<td></td>
</tr>
</tbody>
</table>

In Divide and Recombine (D&R), big data are divided into subsets in one or more ways, forming divisions. Analytic methods, numeric-categorical methods of machine learning and statistics plus visualization methods, are applied to each of the subsets of a division.
Then the subset outputs for each method are recombined. D&R methods of division and recombination seek to make the statistical accuracy of recombinations as large as possible, ideally close to that of the hypothetical direct, all-data application of the methods. The D&R computational environment starts with RHIPE, a merger of R and Hadoop. RHIPE allows an analyst to carry out D&R analysis of big data wholly from within R, and use any of the thousands of methods.
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford University - Hanrahan Publications</td>
<td>Riposte</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/jtalbot/riposte.git">https://github.com/jtalbot/riposte.git</a></td>
<td>stats</td>
<td>Riposte is a fast interpreter and JIT for R. The Riposte VM has 2 cooperative subVMs for R scripting (like Java) and for R vector computation (like APL). Our scripting code has been 2-4x faster in Riposte than in R's recent bytecode interpreter. Vector-heavy code is 5-10x faster. Speeding up R can greatly increases the analyst's efficiency.</td>
<td>BSD</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------</td>
<td>-----------</td>
<td>------------------------</td>
<td>-----------------------</td>
<td>-----------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Stanford University - Hanrahan Publications</td>
<td>trelliscope</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/hafen/trelliscope.git">https://github.com/hafen/trelliscope.git</a></td>
<td></td>
<td>Trellis Display, developed in the 90s, also divides the data. A visualization method is applied to each subset and shown on one panel of a multi-panel trellis display. This framework is a very powerful mechanism for all data, large and small. Trelliscope, a layer that uses datadr, extends Trellis to large complex data. An interactive viewer is available for viewing subsets of very large displays, and the software provides the capability to sample subsets of panels.</td>
<td>BSD</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>-------------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Stanford University - Olukotun Publications</td>
<td>SNAP</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/snap-stanford/snap">https://github.com/snap-stanford/snap</a></td>
<td>stats</td>
<td>Stanford Network Analysis Platform (SNAP) is a general purpose network analysis and graph mining library. It is written in C++ and easily scales to massive networks with hundreds</td>
<td>BSD</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>Lyra</td>
<td>Visualization</td>
<td>2014-07</td>
<td>2014-02</td>
<td></td>
<td>of millions of nodes, and billions of edges. It efficiently manipulates large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges.</td>
<td>BSD</td>
</tr>
<tr>
<td>Stanford, University of Washington, Kitware, Inc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lyra is an interactive environment that makes custom visualization design accessible to a broader audience. With Lyra, designers map data to the properties of graphical marks to author expressive visualization designs without writing code. Marks can</td>
<td></td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYSTAP, LLC</td>
<td>bigdata</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://bigdata.svn.sourceforge.net/svnroot/bigdata/stats">https://bigdata.svn.sourceforge.net/svnroot/bigdata/stats</a></td>
<td>Bigdata enables massively</td>
<td>GPLv2</td>
<td></td>
</tr>
</tbody>
</table>

Lyra also provides a data pipeline interface for iterative, visual specification of data transformations and layout algorithms. Visualizations created with Lyra are represented as specifications in Vega, a declarative visualization grammar that enables sharing and reuse.
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYSTAP, LLC</td>
<td>mpgraph</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="http://svn.code.sf.net/p/mpgraph/code/">http://svn.code.sf.net/p/mpgraph/code/</a></td>
<td>stats</td>
<td>Mpgraph enables massively parallel graph processing on GPUs and many core CPUs. The approach is based on the GraphLab 2.1 Gather Apply Scatter (GAS) API. Execution is available on GPUs, Intel Xenon Phi (aka MIC), and multi-core GPUs.</td>
<td>ALv2</td>
</tr>
</tbody>
</table>

parallel graph processing on GPUs and many core CPUs. The approach is based on the decomposition of a graph algorithm as a vertex program. The initial implementation supports an API based on the GraphLab 2.1 Gather Apply Scatter (GAS) API. Execution is available on GPUs, Intel Xenon Phi (aka MIC), and multi-core GPUs.
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trifacta (Stanford, University of Washington, Kitware, Inc. Team)</td>
<td>Vega</td>
<td>Visualization</td>
<td>2014-07</td>
<td><a href="https://github.com/trifacta/vega.git">https://github.com/trifacta/vega.git</a></td>
<td>stats</td>
<td>decomposition of a graph algorithm as a vertex program. The initial implementation supports an API based on the GraphLab 2.1 Gather Apply Scatter (GAS) API. Execution is available on GPUs, Intel Xenon Phi (aka MIC), and multi-core GPUs.</td>
<td>BSD</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>UC Davis</td>
<td>Gunrock</td>
<td>Analytics</td>
<td>2014-07</td>
<td><a href="https://github.com/gunrock/">https://github.com/gunrock/</a> gunrock.git</td>
<td>stats</td>
<td>Gunrock is a CUDA library for graph primitives that refactors, integrates, and generalizes best-of-class GPU implementations of breadth-first search, connected components, and betweenness centrality into a unified code base useful for future development of high-performance GPU graph primitives.</td>
<td>ALv2</td>
</tr>
<tr>
<td>University of California, Berkeley Publications</td>
<td>BDAS</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td>N/A</td>
<td>BDAS, the Berkeley Data Analytics Stack, is an open source software stack that integrates software components being built</td>
<td>ALv2, BSD</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>University of California, Berkeley Publications</td>
<td>BlinkDB</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/sameeragarwaltats/blinkdb.git">https://github.com/sameeragarwaltats/blinkdb.git</a></td>
<td>by the AMPLab to make sense of Big Data.</td>
<td>BlinkDB is a massively parallel, approximate query engine for running interactive SQL queries on large volumes of data. It allows users to trade-off query accuracy for response time, enabling interactive queries over massive data by running queries on data samples and presenting results annotated with meaningful error bars. To achieve this, BlinkDB</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>uses two key ideas: (1) An adaptive optimization framework that builds and maintains a set of multi-dimensional samples from original data over time, and (2) A dynamic sample selection strategy that selects an appropriately sized sample based on a query's accuracy and/or response time requirements. We have evaluated BlinkDB on the well-known TPC-H benchmarks, a real-world analytic workload derived from Conviva Inc. and</td>
<td></td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>University of California, Berkeley Publications</td>
<td>Mesos</td>
<td>Infrastructure</td>
<td></td>
<td></td>
<td></td>
<td>are in the process of deploying it at Facebook Inc.</td>
<td>ALv2</td>
</tr>
<tr>
<td>University of California, Berkeley Publications</td>
<td>Shark</td>
<td>Infrastructure</td>
<td></td>
<td></td>
<td></td>
<td>Apache Mesos is a cluster manager that provides efficient resource isolation and sharing across distributed applications, or frameworks. It can run Hadoop, MPI, Hypertable, Spark, and other applications on a dynamically shared pool of nodes.</td>
<td>ALv2</td>
</tr>
</tbody>
</table>

Apache Mesos is a cluster manager that provides efficient resource isolation and sharing across distributed applications, or frameworks. It can run Hadoop, MPI, Hypertable, Spark, and other applications on a dynamically shared pool of nodes.

Shark is a large-scale data warehouse system for Spark that is designed to be compatible with Apache Hive. It can...
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of California, Berkeley Publications</td>
<td>Spark</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/mesos/spark.git">https://github.com/mesos/spark.git</a></td>
<td></td>
<td>Apache Spark is an open source cluster computing system that aims to make data</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>analytics both fast to run and fast to write. To run programs faster, Spark offers a general execution model that can optimize arbitrary operator graphs, and supports in-memory computing, which lets it query data faster than disk-based engines like Hadoop. To make programming faster, Spark provides clean, concise APIs in Python, Scala and Java. You can also use Spark interactively from the Scala and Python shells to rapidly</td>
<td></td>
</tr>
</tbody>
</table>

http://semanticommunity.info/Data_Science/A_Data_Science_Big_Mechanism_for_DARPA
Updated: Sat, 19 Sep 2015 00:07:55 GMT
Powered by mindtouch
<table>
<thead>
<tr>
<th>XDATA Team</th>
<th>Software</th>
<th>Category</th>
<th>Instructional Material</th>
<th>Code</th>
<th>Dev Stats</th>
<th>Description</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of California, Berkeley Publications</td>
<td>Tachyon</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/amplab/tachyon.git">https://github.com/amplab/tachyon.git</a></td>
<td>stats</td>
<td>Tachyon is a fault tolerant distributed file system enabling reliable file sharing at memory-speed across cluster frameworks, such as Spark and MapReduce. It achieves high performance by leveraging lineage information and using memory aggressively. Tachyon caches working set files in memory, and enables different jobs/queries and frameworks to access cached files at memory speed. Thus, Tachyon</td>
<td>BSD</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>University of Southern California Publications</td>
<td>goffish</td>
<td>Infrastructure</td>
<td>2014-07</td>
<td><a href="https://github.com/usc-cloud/goffish.git">https://github.com/usc-cloud/goffish.git</a></td>
<td>stats</td>
<td>The GoFFish project offers a distributed framework for storing timeseries graphs and composing graph analytics. It takes a clean-slate approach that leverages best practices and patterns from scalable data analytics such as Hadoop, HDFS, Hive, and Giraph, but with an emphasis on performing native analytics on graph (rather than tuple) data</td>
<td>ALv2</td>
</tr>
<tr>
<td>XDATA Team</td>
<td>Software</td>
<td>Category</td>
<td>Instructional Material</td>
<td>Code</td>
<td>Dev Stats</td>
<td>Description</td>
<td>License</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>----------</td>
<td>------------------------</td>
<td>------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Boeing/Pitt</td>
<td>An Empirical Comparison of Bayesian Network Parameter</td>
<td>Boeing/Pitt</td>
<td>Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Publications

See: [Excel](#)

<table>
<thead>
<tr>
<th>XData Team</th>
<th>Title</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boeing/Pitt</td>
<td>An Empirical Comparison of Bayesian Network Parameter</td>
<td><a href="http://d-scholarship.pitt.edu/19109/">http://d-scholarship.pitt.edu/19109/</a></td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Continuum Analytics and Indiana University</td>
<td>Abstract Rendering: Out-of-core Rendering for Information Visualization</td>
<td>To appear in SPIE: Visualization and Data Analysis (VDA) 2014</td>
</tr>
<tr>
<td>Draper Laboratory</td>
<td>Measuring the value of big data exploitation systems: quantitative, non-subjective metrics with the user as a key component</td>
<td><a href="http://pjim.newschool.edu/issues/2014/01/">http://pjim.newschool.edu/issues/2014/01/</a></td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>IBM Research</td>
<td>Highly Scalable Linear Time Estimation of Spectrograms - A Tool for Very Large Scale Data Analysis</td>
<td>To appear</td>
</tr>
<tr>
<td>IBM Research</td>
<td>Sketching Structured Matrices for Faster Nonlinear Regression Haim Avron</td>
<td>To appear</td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Oculus Info Inc.</td>
<td>Interactive Data Exploration with ‘Big Data Tukey Plots’</td>
<td><a href="http://www.oculusinfo.com/assets/pdfs/papers/Submitted_Oculus_Big_Data_Scatter_Plot_EDA.pdf">http://www.oculusinfo.com/assets/pdfs/papers/Submitted_Oculus_Big_Data_Scatter_Plot_EDA.pdf</a></td>
</tr>
<tr>
<td>Oculus Info Inc.</td>
<td>Louvain Clustering for Big Data Graph Visual Analytics</td>
<td><a href="http://www.oculusinfo.com/assets/pdfs/papers/Submitted_Oculus_Big_Data_Louvain_Clustering.pdf">http://www.oculusinfo.com/assets/pdfs/papers/Submitted_Oculus_Big_Data_Louvain_Clustering.pdf</a></td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Oculus Info Inc.</td>
<td>Visual Thinking Design Patterns</td>
<td><a href="http://www.oculusinfo.com/assets/pdfs/papers/Ware_Et_Ai_VTDP_2013.pdf">http://www.oculusinfo.com/assets/pdfs/papers/Ware_Et_Ai_VTDP_2013.pdf</a></td>
</tr>
<tr>
<td>Royal Caliber</td>
<td>VertexAPI2 - A Vertex-Program API for Large Graph Computations on the GPU</td>
<td><a href="http://www.royal-caliber.com/vertexapi2.pdf">http://www.royal-caliber.com/vertexapi2.pdf</a></td>
</tr>
<tr>
<td>Scientific Systems Company, Inc., MIT, and University of Louisville</td>
<td>Advanced Machine Learning and Statistical Inference Approaches for Big Data Analytics and Information Fusion</td>
<td>To appear</td>
</tr>
<tr>
<td>Sotera Defense Solutions, Inc.</td>
<td>Correlation Using Pair-wise Combinations of Multiple Data Sources and Dimensions at Ultra-Large Scales</td>
<td>To appear</td>
</tr>
<tr>
<td>Stanford University - Hanrahan, Purdue, PNNL</td>
<td>EDA and ML - A Perfect Pair for Large-Scale Data Analysis</td>
<td><a href="http://ml.stat.purdue.edu/gaby/MLandEDAforBigData.pdf">http://ml.stat.purdue.edu/gaby/MLandEDAforBigData.pdf</a></td>
</tr>
<tr>
<td><strong>XData Team</strong></td>
<td><strong>Title</strong></td>
<td><strong>Link</strong></td>
</tr>
<tr>
<td>----------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>Stanford University - Olukotun</td>
<td>Forge: Generating a High Performance DSL Implementation from a Declarative Specification</td>
<td><a href="http://dl.acm.org/citation.cfm?id=2517220">http://dl.acm.org/citation.cfm?id=2517220</a></td>
</tr>
<tr>
<td>Stanford University - Olukotun</td>
<td>On the precision of social and information networks</td>
<td><a href="http://doi.acm.org/10.1145/2512938.2512955">http://doi.acm.org/10.1145/2512938.2512955</a></td>
</tr>
<tr>
<td>The New School</td>
<td>Big Data and Knowledge Discovery Through Metapictorial Visualization</td>
<td><a href="https://www.dropbox.com/sh/ea6ya5cnpnxreuk/q6h5q76tfY/Big_Data_and_Knowledge_Discovery_Metapictorial_Visualization_Parsons.pdf">https://www.dropbox.com/sh/ea6ya5cnpnxreuk/q6h5q76tfY/Big_Data_and_Knowledge_Discovery_Metapictorial_Visualization_Parsons.pdf</a></td>
</tr>
<tr>
<td>The New School</td>
<td>Data Visualization for Big Data (Goranson)</td>
<td><a href="https://www.dropbox.com/sh/ea6ya5cnpnxreuk/Ae6tpC9L30/Data_Visualization_for_Big_Data_Parsons.pdf">https://www.dropbox.com/sh/ea6ya5cnpnxreuk/Ae6tpC9L30/Data_Visualization_for_Big_Data_Parsons.pdf</a></td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>A General Bootstrap Performance Diagnostic</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/a-general-bootstrap-performance-diagnostic/">https://amplab.cs.berkeley.edu/publication/a-general-bootstrap-performance-diagnostic/</a></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/blinkdb-queries-with-bounded-errors-and-bounded-response-times-on-very-large-data/">https://amplab.cs.berkeley.edu/publication/blinkdb-queries-with-bounded-errors-and-bounded-response-times-on-very-large-data/</a></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>Bolt-on Causal Consistency</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/bolt-on-causal-consistency/">https://amplab.cs.berkeley.edu/publication/bolt-on-causal-consistency/</a></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>Carat: Collaborative Energy Diagnosis for Mobile Devices</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/carat-sensys/">https://amplab.cs.berkeley.edu/publication/carat-sensys/</a></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>Discretized Streams: Fault-Tolerant Streaming Computation at Scale</td>
<td><a href="http://dl.acm.org/citation.cfm?doid=2517349.2522737">http://dl.acm.org/citation.cfm?doid=2517349.2522737</a></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>GraphX: A Resilient Distributed Graph System on Spark</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/graphx-grades/">https://amplab.cs.berkeley.edu/publication/graphx-grades/</a></td>
</tr>
<tr>
<td>XData Team</td>
<td>Title</td>
<td>Link</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>Shark: SQL and Rich Analytics at Scale</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/shark-sql-and-rich-analytics-at-scale/">https://amplab.cs.berkeley.edu/publication/shark-sql-and-rich-analytics-at-scale/</a></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>Sparrow: Distributed, Low Latency Scheduling</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/sparrow-distributed-low-latency-scheduling/">https://amplab.cs.berkeley.edu/publication/sparrow-distributed-low-latency-scheduling/</a></td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>The Case for Tiny Tasks in Compute Clusters</td>
<td><a href="https://amplab.cs.berkeley.edu/publication/the-case-for-tiny-tasks-in-compute-clusters/">https://amplab.cs.berkeley.edu/publication/the-case-for-tiny-tasks-in-compute-clusters/</a></td>
</tr>
</tbody>
</table>