Data Science for VIVO

Story
Data Science for VIVO Application
Call for Applications

Story
Data Science for VIVO and the IV MOOC

Story
Datasets for the IV MOOC

Data Science for Drug Discovery, Health and Translational Medicine

Open Directory Project (ODP)
About DMOZ
The Republic of the Web
The Definitive Catalog of the Web
The Internet Brain
You Can Make a Difference
Join DMOZ
Further Information

Slides
Data Science for IV MOOC-Spotfire-Cover Page
Data Science for IV MOOC-Spotfire-VFLC
Data Science for IV MOOC-Spotfire-DSDHT Assignment 5
Data Science for IV MOOC-Spotfire-FAERS Statistics

Spotfire Dashboard
Fifth Annual VIVO Conference
Call for Papers
Abstract

Slides
Slide 1 Data Science for VIVO
Slide 2 Overview 1
Slide 3 Atlas of Science
Slide 4 Overview 2
Slide 5 Vivoweb.org
Slide 6 Indiana University CNS: Vivo Data and Visualizations
Slide 7 Indiana University CNS: VIVO Presentation Information
Slide 8 Indiana University CNS: 2.5 Sample Data Sets
Slide 9 Cyberinfrastructure for Network Science Center
Slide 10 Information Visualization MOOC
Slide 11 Pre-Questionnaire for the IVMOOC
Slide 12 Emails for the IVMOOC
Slide 13 IVMOOC - Schedule
Slide 14 IVMOOC - Week 5
Slide 15 IVMOOC – Student Locations
Slide 16 Indiana University CNS: Vivo Data and Visualizations Overview
Slide 17 The UCSD Map of Science
Slide 18 Published But Not the Raw Data
Slide 19 Indiana University CNS: Vivo Data and Visualizations Sample Data
Slide 20 Temporal Visualization: Grants Over Time
Slide 21 Statements I Disagree With!
Slide 22 Exercises and Data Sets
Slide 23 iNRN
Slide 24 Theory Unit Structure
Slide 25 Data Science for VIVO Knowledge Base
Slide 26 Scholarly Database: MEDLINE data Sets in Spotfire
Slide 27 VIVO Workshop August 2013 Selected Data Visualizations in Spotfire
Slide 28 SCI2: Full Sample Selected Data Visualizations in Spotfire

Spotfire Dashboard
Research Notes
Overview
Schedule
Pre-Questionnaire
Your Work Practices
Your Needs
Your Expertise
Unit 1 – Visualization Framework & Workflow Design
Theory
Welcome by Katy Börner (1:57)
Course Overview (11:36)
Visualization Framework (28:59)
Workflow Design (19:40)
Self-Assessment

Hands-on
Introduction by Ted Polley (:47)
Download Tool, Install, and Visualize Data with Sci2 (10:54)
Legend Creation with Inkscape (16:03)
Weekly Tip: Extend Sci2 by adding Plugins (3:13)

Homework

Unit 2 – “When”: Temporal Data

Theory
Welcome by Katy Börner (:39)
Exemplary Visualizations (9:46)
Overview and Terminology (16:38)
Workflow Design (19:41)
Burst Detection (14:14)
Self-Assessment

Hands-on

Semantic MEDLINE Query: mesothelioma

Unit 3 – "Where": Geospatial Data

Theory
Welcome by Katy Börner (:54)
Exemplary Visualizations (5:55)
Overview and Terminology (11:09)
Workflow Design (6:39)
Color (8:41)
Self-Assessment

Hands-on
Introduction by Ted Polley (:52)
Choropleth and Proportional Symbol Map (8:06)
Congressional District Geocoder (8:49)
Geocoding NSF Funding with the Generic Geocoder (11:33)
Weekly Tip: Memory Allocation (3:16)

Homework

EPA Waterways

Unit 4 – "What": Topical Data

Theory
Welcome by Katy Börner (1:13)
Exemplary Visualizations (9:09)
Overview and Terminology (10:15)
Workflow Design (18:11)
Design and Update of a Classification System: The UCSD Map of Science (Optional) (24:48)
Comparison of Text- and Linkage-Based Approaches (Optional) (9:01)
Self-Assessment

Hands-on
Introduction by Ted Polley (:54)
Mapping Topics and Topic Bursts in PNAS (9:31)
Word Co-Occurrence Networks with Sci2 (14:35)
Weekly Tip: Removing Files from the Data Manager (1:30)

Homework
Euretos BRAIN

Mid-Term
1: Visualization Framework and Workflow Design
2: "When": Temporal Data
3: "Where": Geospatial Data
4: "What": Topical Data

Unit 5 – “With Whom”: Trees

Theory
Welcome by Katy Börner (:42)
Exemplary Visualizations (6:51)
Overview and Terminology (5:46)
Workflow Design (9:09)
Algorithm Comparison (9:44)
Self-Assessment

Hands-on
Introduction by Ted Polley (1:04)
Visualizing Directory Structures (6:07)
Weekly Tip: Create your own TreeML (XML) Files (9:19)

Homework
Unit 6 – “With Whom”: Networks

Email

Theory
Welcome by Katy Börner (2:01)
Exemplary Visualizations (7:58)
Overview and Terminology (15:54)
Workflow Design (12:49)
Clustering (7:59)
Backbone Identification (7:15)
Error and Attack Tolerance (3:07)
Exhibit Map with Andrea Scharnhorst (Optional) (6:22)
Self-Assessment

Hands-on
Introduction by Ted Polley (1:04)
Co-Occurrence Networks: NSF Co-Investigators (12:18)
Directed Networks: Paper-Citation Network (8:01)
Bipartite Networks: Mapping CTSA Centers (10:27)
Weekly Tip: How to use Property Files (7:17)

Homework
NodeXL

Unit 7 – Dynamic Visualizations & Deployment

Theory
Welcome by Katy Börner (1:57)
Exemplary Visualizations (20:50)
Dynamics (7:56)
Hans Rosling’s Gapminder (4:27)
Deployment (21:07)
Color Perception and Reproduction (8:11)
The Making of AcademyScope (Optional) (8:05)
Self-Assessment

Hands-on
Introduction by Ted Polley (1:05)
Evolving Networks with Gephi (9:23)
Exporting Networks from Gephi with Zoom.it (formerly Seadragon)

Homework
Final Exam

Instructions

Framework & Workflow Design

“When, Where, and With Whom”: Temporal, Geospatial, and Network Analysis & Visualization

Search Query
Temporal Visualizations
Burst Analysis
Geospatial Visualization
Co-Funding Network

“What and With Whom”: Topical and Network Analysis & Visualization

Topical
“With Whom”: Trees
“With Whom”: Networks

Dynamic Visualizations & Deployment

List of Clients

<table>
<thead>
<tr>
<th>Project Title: Information Visualizations for Big Data in Drug Discovery, Health and Translational Medicine</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Client Name</strong></td>
</tr>
<tr>
<td><strong>Project goal/scientific or practical value</strong></td>
</tr>
<tr>
<td><strong>Information on dataset(s) to be used</strong></td>
</tr>
<tr>
<td><strong>Web-link to dataset(s)</strong></td>
</tr>
<tr>
<td><strong>Relevant publications, websites, etc.</strong></td>
</tr>
<tr>
<td><strong>Conditions under which students can publish results and/or add project results to their resume</strong></td>
</tr>
<tr>
<td><strong>Client Forum</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Project Title: Human Genome Project Documentary History: An Annotated Scholarly Guide to the HGP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Client Name</strong></td>
</tr>
<tr>
<td><strong>Project goal/scientific or practical value</strong></td>
</tr>
<tr>
<td><strong>Information on dataset(s) to be used</strong></td>
</tr>
<tr>
<td><strong>Web-link to dataset(s)</strong></td>
</tr>
<tr>
<td><strong>Relevant publications, websites, etc.</strong></td>
</tr>
<tr>
<td><strong>Conditions under which students can publish results and/or add project results to their resume</strong></td>
</tr>
<tr>
<td><strong>Client Forum</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Project Title: Federal Library Collection Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Client Name</strong></td>
</tr>
<tr>
<td><strong>Project goal/scientific or practical value</strong></td>
</tr>
<tr>
<td><strong>Information on dataset(s) to be used</strong></td>
</tr>
<tr>
<td><strong>Web-link to dataset(s)</strong></td>
</tr>
<tr>
<td><strong>Relevant publications, websites, etc.</strong></td>
</tr>
<tr>
<td><strong>Conditions under which students can publish results and/or add project results to their resume</strong></td>
</tr>
<tr>
<td><strong>Client Forum</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Project Title: Global Biotic Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Client Name</strong></td>
</tr>
<tr>
<td><strong>Project goal/scientific or practical value</strong></td>
</tr>
<tr>
<td><strong>Information on dataset(s) to be used</strong></td>
</tr>
<tr>
<td><strong>Web-link to dataset(s)</strong></td>
</tr>
<tr>
<td><strong>Relevant publications, websites, etc.</strong></td>
</tr>
<tr>
<td><strong>Conditions under which students can publish results and/or add project results to their resume</strong></td>
</tr>
<tr>
<td><strong>Client Forum</strong></td>
</tr>
<tr>
<td><strong>Interview</strong></td>
</tr>
</tbody>
</table>
Project Title: Synthesizing spatial diet data of fishes from the Gulf of Mexico

Client Name

Project goal/scientific or practical value

Information on dataset(s) to be used

Web-link to dataset(s)

Relevant publications, websites, etc.

Conditions under which students can publish results and/or add project results to their resume

Client Forum

Project Title: Knowledge Network Evolution

Client Name

Project goal/scientific or practical value

Information on dataset(s) to be used

Web-link to dataset(s)

Relevant publications, websites, etc.

Conditions under which students can publish results and/or add project results to their resume

Client Forum

Project Title: The Genealogy of Psychoanalysis

Client Name

Project goal/scientific or practical value

Information on dataset(s) to be used

Web-link to dataset(s)

Relevant publications, websites, etc.

Conditions under which students can publish results and/or add project results to their resume

Client Forum

Project Title: Evolution of Wikipedia's Category Structure

Client Name

Project goal/scientific or practical value

Information on dataset(s) to be used

Web-link to dataset(s)

Relevant publications, websites, etc.

Conditions under which students can publish results and/or add project results to their resume

Client Forum

Project Title: 30 Years of Alzheimer’s disease Research at NIA

Client Name

Project goal/scientific or practical value

Information on dataset(s) to be used

Web-link to dataset(s)

Relevant publications, websites, etc.

Conditions under which students can publish results and/or add project results to their resume
**Globalization of the United States, 1789-1861**

**Project Title:**

**Client Name**

**Project goal/scientific or practical value**

**Information on dataset(s) to be used**

**Web-link to dataset(s)**

**Relevant publications, websites, etc.**

**Conditions under which students can publish results and/or add project results to their resume**

---

**Week 8 - Mar. 18, 2014: Picking a Client**

**Forming a Client Group**

**Data Science for the Federal Big Data Initiative**

**Week 9 - Mar. 25, 2014: Project Ideas**

**Week 10 - Apr. 1, 2014: 1st Project Draft**

**Complete by Monday, Apr. 7, 2014 at 5pm EST**

**Week 11 - Apr. 8, 2014: 1st Project Draft**

**Complete by Monday, Apr. 14, 2014 at 5pm EST**

**Week 12 - Apr. 15, 2014: Peer Feedback**

**Complete by Monday, Apr. 21, 2014 at 5pm EST**

**Week 13 - Apr. 22, 2014: 2nd Project Draft**

**Complete by Monday, Apr. 27, 2014 at 5pm EST**

**Week 14 - Apr. 28, 2014: Project Submission Due**

**Complete by Monday, May 5, 2014 at 5pm EST**

**Week 15 - Presentations**

**Complete by Monday, May 5, 2014 at 5pm EST**

---

**Post-Questionnaire**

---

**Story**

**Data Science for VIVO Application**

To qualify applications must:

- Consume VIVO data (linked open data in the VIVO ontology) from more than one VIVO site. Response: This content is linked open data from many VIVO digital objects in many locations. The ontology comes from the index of the mashup of the diverse VIVO data sources and is expressed as the Table of Contents of the MindTouch page in a spreadsheet in Semantic Web RDF Linked Open Data Format (Subject, Object, Predicate) following the OMB Ontology and Ontologizing Memo guidance (see Summary below). This is how we are building a Data Science Data Publication Commons (see Story).
• Be accessible from a single persistent URL, accessible on the public Internet without authentication. Response: This is.
• Available as open source via a standard open source license. Response: Yes, MindTouch is open source via a standard open source license

Submissions must include:

• Names, academic degree(s), affiliations, and locations of all authors. Brief (500 words or less) description of the application and its value to scientists. Response: Dr. Brand L. Niemann, Director and Senior Data Scientist/Data Journalist, Semantic Community, Fairfax, Virginia 22030. See Paper Submitted to VIVO Conference 2014 below.
• A URL for the application. Response: http://semanticommunity.info/Data_Science/Data_Science_for_VIVO
• Instructions regarding operation. The judges will execute the application from the URL provided without assistance from the authors. Response: This MindTouch Wiki Page

Semantic Community uses Linked Data (http://linkeddata.org) standards for data access that supports the Resource Description Framework (RDF) and is a robust open source, open community space (using MindTouch) like DuraSpace with content that supports implementation, adoption, and development efforts around the world.

Semantic Community follows the guidance in the OMB Ontology and Ontologizing Memo - In Summary:
An Ontology:
• is a formal representation of meaning in an information system;
• creates the bridge between the internal world of the computer and the external world of people’s understanding;
• provides an inter lingua between disparate data sources and knowledge bases;
• allows us to build useful and usable systems for complex tasks in health care.

Remember:
• don’t try to divorce the Ontology from its application (the ‘universal ontology’)
• building and embedding an Ontology in a useful application has pitfalls that require judgment, experience, clarity of purpose, and resources.

Call for Applications

The 2014 VIVO Conference and the VIVO Apps and Tools working group are sponsoring a competition for applications using data coming from VIVO systems to support research communities.

Eligibility

Applications can be written in any language. To qualify applications must:
• Consume VIVO data (linked open data in the VIVO ontology) from more than one VIVO site.
• Be accessible from a single persistent URL, accessible on the public Internet without authentication.
• Available as open source via a standard open source license.

Criteria for Success

Applications will be judged on the following criteria:
1. Value to the research community.
2. Functionality: The application performs as described.
3. Presentation: The application is well organized and attractive.

Prizes
The first prize winner will have registration refunded, receive a cash prize and will be awarded a plaque at a recognition ceremony during the conference. Second and third prize winners will receive certificates.

Submissions
To submit your application, please send an email to Apps & Tools Working Group Chairs Ted Lawless (tlawless@brown.edu) and Christopher Barnes (cpb@ufl.edu) Applications are due July 31, 2014. Winners will be announced at the conference. Submissions must include:

- Names, academic degree(s), affiliations, and locations of all authors. Brief (500 words or less) description of the application and its value to scientists.
- A URL for the application.
- Instructions regarding operation. The judges will execute the application from the URL provided without assistance from the authors.

Additional Resources
VIVO uses Linked Data (http://linkeddata.org) standards for data access via Resource Description Framework (RDF). See the Linked Data site for more information regarding Linked Data, RDF and data processing.

VIVO enjoys a robust open source, open community space on DuraSpace at http://wiki.duraspace.org/display/VIVO with content that supports implementation, adoption, and development efforts around the world. The VIVO software and ontology are publicly available at https://github.com/vivo-project/VIVO.

For more information on the working group please check out our wiki: https://wiki.duraspace.org/display/Tools+Working+Group

---

**Story**

**Data Science for VIVO and the IV MOOC**

How did I come by this title and story?

While working on Euretos BRAIN for the March 4th Meetup of the Federal Big Data Working Group Meetup, I recalled that Dr. Barend Mons had presented at the 2010 VIVO Conference, so I decided to check back on the status of VIVO and especially the VIVO work at the Indiana University.

From the VIVO 2013 Conference, I found three very interesting developments:

- Indiana University CNS: Vivo Data and Visualizations which was a "Data Science" tutorial
- VIVO's Evolution from Semantic Web Application to DuraSpace; and

http://semanticommunity.info/Data_Science/Data_Science_for_VIVO
Updated: Fri, 18 Sep 2015 23:02:30 GMT
Powered by mindtouch
Indiana University's Evolution from VIVO Information Visualization to Data Science

IU now offers an Information Visualization MOOC (Massive Open Online Course) as follows:

- This course provides an overview about the state of the art in information visualization. It teaches the process of producing effective visualizations that take the needs of users into account.
- This year, the course can be taken for three Indiana University credits as part of the Online Data Science Program just announced by the School of Informatics and Computing. Students interested in applying to the program can find more information here.
- Everyone who registers gains free access to the Scholarly Database (26 million paper, patent, and grant records) and the Sci2 Tool (100+ algorithms and tools). My Note: I wondered if this was all of it or just a few thousand records at a time like I found in my previous work!

Interestingly, my previous work was Data Science for VIVO:

- Atlas of Science
- An Interface to a Digital Library of the Atlas of Science
  - 2010 Science and Engineering Indicators
  - Nobel Prizes
- Build VIVO in the Cloud
- NIH Workshop on Value Added Services for VIVO
  - Build the Scholarly Databases in the Cloud

So my goal became to use the "Data Science" tutorial data sets and the Information Visualization MOOC data sets in Spotfire and NodeXL to audit their results and report on them. In my professional opinion, Data Science is really about answering three basic questions:

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

in a manner that another data scientist can come along and readily reproduce or not-reproduce, whichever the truth may be. I expressed reservation about being able to do that with the IU VIVO work previously because of their use of copyrighted/proprietary data and/or lack of transparent scholarship.

I downloaded the VIVO 2013 August Workshop Data ZIP file and attached the files to this Wiki and mined the Information Visualization MOOC for data sets most of which come from the Scholarly Database.

The Schedule for the Information Visualization MOOC has been structured/visualized as a Knowledge Base for ease of reuse by linking the slides and extracting key points, links to data, and my data science commentary.

The Unit 2 Homework Exercise was the following:

- Select the “MEDLINE” dataset and enter the Search term “mesothelioma” in the title field:
- From the search results page, select download. Then, download the first thousand results and save the MEDLINE master table. It will download as a .csv file:
• Load the .csv file into Sci2 and conduct a burst analysis, similar to the one conducted in this week’s hands-on video. Once you have completed the analysis compare the results to the Wikipedia article for mesothelioma to see support for the visualization you have just created.

My result was I could not get all 5,597 "mesothelioma" hits, so I selected 2000 and got the following CSV files attached below:

• MEDLINE_author_table
• MEDLINE_co-author_table_(nwb_format)
• MEDLINE_master_table
• MEDLINE_MeSH_heading_table
• MEDLINE_MeSH_qualifier_table

Next I am going to visualize these tables along with the others mentioned above in a Spotfire Dashboard and also ask Dr. Tom Rindflesch to do the same for "mesothelioma" in his Semantic Medline. My Data Science for VIVO work is also documented in the Slides below.

I just tweeted:

Brand Niemann @bniemannsr 46s

#ivmooc See Data Science for VIVO http://semanticommunity.info/Data_Science/Data_Science_for_VIVO ... using TIBCOSpotfire and @MindTouch pic.twitter.com/BPZhTGDozG

Brand Niemann @bniemannsr 2h

#ivmooc See Data Science for VIVO http://semanticommunity.info/Data_Science/Data_Science_for_VIVO#Semantic_MEDLINE_Query:_mesothelioma ... using Semantic Medline medlineplus pic.twitter.com/3Yl8OAHRlB

Brand Niemann @bniemannsr 2h

#ivmooc See Data Science for VIVO http://semanticommunity.info/Data_Science/Data_Science_for_VIVO#EPA_Waterways ... temporal and spatial data (2.5 GB) TIBCOSpotfire (2.5 GB) pic.twitter.com/hVEvFFWU94

Brand Niemann @bniemannsr 1h

#ivmooc See Data Science for VIVO http://semanticommunity.info/Data_Science/Data_Science_for_VIVO#Euretos_BRAIN ... by Dr. Barden Mons showing sample nanopublication output pic.twitter.com/2HSMRQyQ5l

This data science work could also be viewed as creating a "data paper" for an upcoming workshop on The Semantics for e-science in an intelligent Big Data context:

Semantic publishing should be central to the openness that has been embraced by scholarly communication, e-science, data journalism, e-government and across many other domains. This openness implies deep changes in making the semantics of the data available for integration, consumption and analysis. Researchers are moving from a purely narrative based communication into a narrative supported by data; this support will be enhanced by
semantic descriptions. Such a shift makes an impact on all layers of scholarly publishing; data needs to be archived and kept readily available and interoperable. Scholars across many disciplines are involved in an important shift in their communication practices; reproducibility, smart data storage, intelligent use of the Web as a platform and not solely as a dissemination channel, business intelligence for e-science content and many others are currently a matter of debate in the academic community. Experimental data in scientific disciplines is a Big Data problem; how can we make effective use of scientific data, how should it be semantically represented, interlinked, reused, how can experiments in scientific publications be represented effectively? How can the gap between publications and data repositories be bridged?

See NodeXL and Sci2 for Data Science

The Data Science for VIVO and IV MOOC Data Paper (this MindTouch wiki) in a Data Browser (the Spotfire dashboard Web Player) are shown below based on the Excel spreadsheet data sets.

The presentation of the results will be given at the May 6th Federal Big Data Working Group Meetup.

There are twelve groups of students working on 10 client projects of which 2 do not seem to have data sets and 2 that require permission to use the data set. I am looking at the other data sets to see what they contain and will report on that next. See Story below.

Story

Datasets for the IV MOOC

The IV MOOC Datasets inventory below, and my research into them, suggests the first and the third are the most suitable for further data science work: So I started with data mining at the following:

Data Science for Drug Discovery, Health and Translational Medicine
## Open Directory Project (ODP), also known as "DMOZ"

<table>
<thead>
<tr>
<th>Group Name</th>
<th>Client Project</th>
<th>Web-link to dataset(s)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data in Drug Discovery, Health and Translational Medicine #1 and #2</td>
<td>Information Visualizations for Big Data in Drug Discovery, Health and Translational Medicine</td>
<td>Web-link to dataset</td>
<td>MOOC Class Assignment Example: [MQIC Patient Data Detailed Sample](XLSX &amp; CSV)</td>
</tr>
<tr>
<td>Data Science for the Federal Big Data Initiative</td>
<td>Federal Big Data Working Group Meetup and NIST</td>
<td>Web-link to dataset</td>
<td>My Group See Spreadsheet</td>
</tr>
<tr>
<td>Evolution of Wikipedia's Category Structure</td>
<td>Evolution of Wikipedia's Category Structure</td>
<td>Web-link to dataset</td>
<td>See PDF</td>
</tr>
<tr>
<td>Globalization of the United States, 1789-1861</td>
<td>Globalization of the United States, 1789-1861</td>
<td>Web-link to dataset</td>
<td>No data?</td>
</tr>
<tr>
<td>Knowledge Network Evolution</td>
<td>Knowledge Network Evolution</td>
<td>Web-link to dataset</td>
<td>No data?</td>
</tr>
<tr>
<td>Cold Spring Harbor Laboratory</td>
<td>Human Genome Project Documentary History: An Annotated Scholarly Guide to the HGP</td>
<td>Web-link to dataset</td>
<td>Downloaded 3 files: Two Word and 27 Text in a ZIP</td>
</tr>
<tr>
<td>Synthesizing spatial diet data of fishes from the Gulf of Mexico</td>
<td>Synthesizing spatial diet data of fishes from the Gulf of Mexico</td>
<td>Web-link to dataset</td>
<td>Registered for database access</td>
</tr>
<tr>
<td>The Genealogy of Psychoanalysis</td>
<td>The Genealogy of Psychoanalysis</td>
<td>Web-link to dataset</td>
<td>No data?</td>
</tr>
<tr>
<td>Visualization of Species Interaction Data</td>
<td>Global Biotic Interactions</td>
<td>Web-link to dataset</td>
<td>Original datasets (23)</td>
</tr>
<tr>
<td>Visualizing 30 Years of Alzheimer’s disease Research</td>
<td>30 Years of Alzheimer’s disease Research at NIA</td>
<td>Web-link to dataset</td>
<td>If you are a student taking &quot;Spring 2014 Information Visualization MOOC&quot; class taught by Prof. Katy Borner at Indiana University and are interested in the client project &quot;30 Years of AD Research at NIA&quot;, please</td>
</tr>
</tbody>
</table>
The results of my data mining are shown below in preparation for Spotfire analytics and visualizations.

**Data Science for Drug Discovery, Health and Translational Medicine**

Source: [http://dsdht.wikispaces.com/](http://dsdht.wikispaces.com/)

*To tweet about this course, please use the hashtag #dsdht and also#datascience (if there is room!)* My Note: See these.

**Course description**

With exploding healthcare costs, greater longevity and widespread health challenges of diabetes, obesity, cancer and cardiovascular disease, medicine and healthcare will be a primary scientific and economic focus for the remainder of this century. Informatics and data science offer the promise of a level of understanding of health, disease and treatment on a scale never before imagined. This course will address the big data techniques that are being used in the drug discovery, healthcare and translational medicine domains and will be organized around three questions: how can data science help researchers find new drugs and reuse old ones? How can data science help doctors treat patients better? And how can data science help us all lead healthier lives?

**Assignments**

Below are 5 assignments. We may add a sixth later. You should choose two of these assignments and submit your completed PDF's to the appropriate assignment on Oncourse (i.e. with the same assignment number) by April 21.

- **Assignment 1** - Machine learning using chemical compound and activity data  
  ◦ My Note: Downloaded 4 datasets, but cannot use these data

- **Assignment 2** - Analysis of microarray assay gene expression data  
  ◦ My Note: The raw data is available at resources/Assignment 2 folder on Oncourse includes the .CEL files and phenotype data. So this is restricted use data.

- **Assignment 3** - Network visualization of semantic disease-related data using Cytoscape / Sci2  
  ◦ My Note: The data is given is as a csv file on oncource resources/Assignment 3 folder.

- **Assignment 4** - Designing a data science health app  
  ◦ My Note: In this assignment, you will design a smartphone/tablet app (iPhone, iPad, Android) to help people find the most pressing information relevant to them and their conditions, including latest research findings and [http://semanticommunity.info/Data_Science/Data_Science_for_VIVO](http://semanticommunity.info/Data_Science/Data_Science_for_VIVO)
adverse event reports. Note that you do not need to implement the app, but to make a design for one. See below.

• **Assignment 5** - Mining Electronic Medical Record Data **My Note: Downloaded 2 datasets and can use these data**

**Assignment 4:** Here are some of the sources of data your might consider using if you want to make your app prototype more realistic (you can try downloading some relevant data for your prototype - but you don't have to do this):

• Recent publications in PubMed - perhaps using MESH classifications or keyword search. **My Note: We have worked with Semantic Medline**

• **Physician Reported Adverse Events** **My Note: There are data sets available here. See below.**

• Web searches, including Wikipedia (note there is a structured version of Wikipedia called dbPedia) **My Note: See ODP below.**

• Data extracted from social media (e.g. Twitter) **My Note: Can use NodeXL for this.**

Other sources that might be good for the app, but which you won't have access to for this prototype (although you can "make it up"):

• Electronic medical records data

• Structured data extracted from social media **My Note: Can use Recorded Future for this.**

Also found: If you understood the video how to perform predictive modeling then give a shot at this [KDD dataset](#) i.e (Prediction of Molecular Bioactivity for Drug Design -- Binding to Thrombin) and try to post the results.

**My Note: I have worked with the [KDD Cup](#) datasets, but I am not sure I worked with this one so check it out.**

**Are FAERS Data Available to the Public?**

Source: [http://www.fda.gov/Drugs/GuidanceCom...ts/default.htm](http://www.fda.gov/Drugs/GuidanceCom...ts/default.htm)

FAERS data are available to the public in the following ways:

• **FAERS Statistics:** provides numbers of reports that FDA has received for drug and therapeutic biologic products over the past ten years. **My Note: I captured these to a spreadsheet. See below.**

• **FAERS Data Files:** provides raw data consisting of individual case safety reports extracted from the FAERS database. A simple search of FAERS data cannot be performed with these files by persons who are not familiar with creation of relational databases. **My Note: See download information below**

• Individual case safety reports from the FAERS database can also be obtained by sending a Freedom of Information (FOI) request to FDA

**FDA Adverse Event Reporting System (FAERS) Statistics**

The FDA Adverse Event Reporting System (FAERS) contains over seven million reports of adverse events and reflects data from 1969 to the present. FDA moved data from our legacy system (AERS) on August 28, 2012.

Data from FAERS are presented here as summary statistics. These summary statistics cover data received over the last ten years. These data are presented at the individual report level; some of the numbers may reflect duplicate reporting due to factors such as follow-up reports received on a case or different persons reporting on the same patient case. We will update these data files each quarter; therefore, the most recent year displayed may contain only partial year data.
• **Reports Received and Reports Entered into FAERS by Year** [My Note: Spreadsheet table.](#)
Number of reports received by FDA and entered into AERS/FAERS by type of report since the year 2003.

• **Domestic and Foreign Reports by Year** [My Note: Spreadsheet table.](#)
Number of domestic (U.S.) and foreign reports in AERS/FAERS since the year 2003.

• **Reporting by Healthcare Providers and Consumers by Year** [My Note: Spreadsheet table.](#)
Number of reports in AERS/FAERS by type of reporter (Healthcare Professional [HCP] or consumer) since the year 2003.

• **Patient Outcomes by Year** [My Note: Spreadsheet table.](#)
Patient outcome(s) for reports in AERS/FAERS since the year 2003. Serious outcomes include death, hospitalization, life-threatening, disability, congenital anomaly and/or other serious outcome.

**FDA Adverse Event Reporting System (FAERS): Latest Quarterly Data Files**

The files listed on this page contain raw data extracted from the AERS database for the indicated time ranges and are not cumulative.

Users of these files need to be familiar with creation of relational databases using applications such as ORACLE®, Microsoft Office Access, MySQL® and IBM DB2 or the use of ASCII files with SAS® analytic tools.

A simple search of AERS/FAERS data cannot be performed with these files by persons who are not familiar with creation of relational databases. However, you can get a summary FAERS report for a product by sending a Freedom of Information Act (FOIA) request to FDA. You can also request individual case reports by submitting a FOIA request listing case report numbers.

• [General Instructions on How to Make a FOIA Request](#)
• [Instructions for Requesting Individual Case Reports](#)

The quarterly data files, which are available in ASCII or SGML formats, include:

• demographic and administrative information and the initial report image ID number (if available);
• drug information from the case reports;
• reaction information from the reports;
• patient outcome information from the reports;
• information on the source of the reports;
• a "README" file containing a description of the files.

**Older Quarterly Legacy AERS Data Files Page**

**FAERS Data Files**

Click on a link below to begin downloading.

• [FAERS_ASCII_2013q1.zip](#) (ZIP - 23.7MB) [My Note: Downloaded these and need to merge them in Spotfire](#)
  January - March 2013

• [FAERS_XML_2013q1.zip](#) (ZIP - 37.2MB)
  January - March 2013
Open Directory Project (ODP)

Also known as "DMOZ"

Source: http://www.dmoz.org/

About DMOZ

Source: http://www.dmoz.org/docs/en/about.html

DMOZ is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a passionate, global community of volunteers editors. It was historically known as the Open Directory Project (ODP).

The Republic of the Web

The web continues to grow at staggering rates. Automated search engines are increasingly unable to turn up useful results to search queries. The small paid editorial staffs at commercial directory sites can't keep up with submissions, and the quality and comprehensiveness of their directories has suffered. Link rot is setting in and they can't keep pace with the growth of the Internet.

Instead of fighting the explosive growth of the Internet, DMOZ provides the means for the Internet to organize itself. As the Internet grows, so do the number of net-citizens. These citizens can each organize a small portion of the web and present it back to the rest of the population, culling out the bad and useless and keeping only the best content.

The Definitive Catalog of the Web

DMOZ follows in the footsteps of some of the most important editor/contributor projects of the 20th century. Just as the Oxford English Dictionary became the definitive word on words through the efforts of volunteers, DMOZ follows in its footsteps to become the definitive catalog of the Web.

DMOZ was founded in the spirit of the Open Source movement, and is the only major directory that is 100% free. There is not, nor will there ever be, a cost to submit a site to the directory, and/or to use the directory's data. DMOZ data is made available for free to anyone who agrees to comply with our free use license.

The Internet Brain

DMOZ is the most widely distributed data base of Web content classified by humans. Its editorial standards body of net-citizens provide the collective brain behind resource discovery on the Web. DMOZ powers the core directory services
for the Web's largest and most popular search engines and portals, including Netscape Search, AOL Search, Google, Lycos, HotBot, DirectHit, and hundreds of others.

**You Can Make a Difference**

Like any community, you get what you give. DMOZ provides the opportunity for everyone to contribute.

**Signing up is easy**: choose a topic you know something about and join. Editing categories is a snap. We have a comprehensive set of tools for adding, deleting, and updating links in seconds. For just a few minutes of your time you can help make the Web a better place, and be recognized as an expert on your chosen topic.

**Join DMOZ**

1. Find a category that you would like to maintain.
2. Follow the Become an Editor link at the top of the category page.

Note that some categories do not have a Become an Editor link; you should find a more specific category which interests you, and apply there. Once you have joined, and gained some experience, you can apply for more general categories.

**Further Information**

- [General FAQ](#) - Want more info? Check out some general FAQs about us in our Help area.
- [Sites Using DMOZ Data](#) - Category listing of all sites, search engines, portals, etc. using the DMOZ RDF dump.
- [DMOZ in the News and Media](#) - News, articles and other reports about DMOZ.
- [Linking to DMOZ](#) - Show your support of the DMOZ by linking to us from your site.
- [Downloading DMOZ Data](#) - Get the Definitive Directory of the Web for your site. It's free and available to everyone.
  - [categories.txt](#) 06-Apr-2014 00:31 61M **My Note: Found in txt format and downloaded**
- [Free Use License](#) - License governing use of DMOZ data.
- [Social Contract](#) - Our commitment and promise to all net-citizens to keep DMOZ a free and open resource.

---

**Slides**
Data Science for IV MOOC-Spotfire-DSDHT Assignment 5

Data Science for IV MOOC-Spotfire-FAERS Statistics

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.
Fifth Annual VIVO Conference

The Fifth Annual VIVO Conference will be held August 6-8, 2014 at the Hyatt Regency on the shores of Lady Bird Lake situated in the heart of downtown Austin, Texas. The VIVO Conference creates a unique opportunity for people from across the country and around the world to come together to explore ways to use semantic technologies and linked open data to promote scholarly collaboration and research discovery.

The VIVO conference is an excellent opportunity to meet with VIVO team members from participating institutions, and it offers an open and collaborative environment to share ideas and discuss topics related to adoption and implementation of VIVO, VIVO-based tools and the opportunities created by advancing data sharing and team science. This year, the conference will be co-located with the Science of Team Science (SciTS) conference, and attendees will be able to freely attend both VIVO and SciTS presentations. Please visit http://www.vivoconference.org for more information.

Authors are invited to submit abstracts for poster, panel, and paper presentations related to the Topics of Interest for the Fifth Annual VIVO Conference in August. For a copy of the full Call for Papers, please click here. All submission must be submitted through EasyChair by Friday, April 4th 5:00 PM EST.

Call for Papers

VIVO was originally funded by Cornell University and the National Institutes of Health (U24 RR029822) and is currently a community supported incubator project under the DuraSpace umbrella (http://duraspace.org).

Over the past five years, a growing international movement of developers, researchers, administrators, funders, librarians and informaticians has converged around the vision of openly representing research and researchers via Linked Open Data. VIVO is helping to make this vision a reality through its community, through open software and the VIVO-ISF (Integrated Semantic Framework) ontology, and through a growing number of adopters and collaborators worldwide, across multiple knowledge domains.

The 2014 VIVO conference will explore how to participate in and best take advantage of the emerging Linked Open Data world encompassing and expanding our understanding of the rapidly growing network of information describing and interlinking researchers and research. How can we contribute? How will newly available data and the applications built around it change the future of research networking? How will the vision evolve into practice?

This year, the conference will be co-located with the Science of Team Science (SciTS) conference, and attendees will be able to freely attend both VIVO and SciTS presentations.

Topics of Interest

- Researcher collaboration and networking
- Implementation and adoption of VIVO and related systems that interoperate through shared ontologies and Linked Open Data
- Open research data and related issues in discovery, reuse, and attribution – including research dataset representation and linkages
- Open representations of research and implications for the research process, collaboration, and virtual research communities
• Team-based science
• Theory and practice of working with the VIVO-ISF Ontology and other modules of the Integrated Semantic Framework
• Applications that use VIVO-compliant Linked Open Data and/or SPARQL queries
• Visualization, analysis, and metrics
• Managing and discovering knowledge about researchers across institutional, disciplinary, and geographic boundaries
• The intersection of VIVO and international research information and identifier standards
• Policy perspectives, planning, and modelling for compliance and/or knowledge mobilization
• Perspectives on policy, research representation, and research impact, including questions of privacy, individual vs. institutional sourcing of data, and change over time
• Semantic Web development and extensions of the VIVO platform to reach the full Web community
• Development and extensions of the VIVO platform
• Use of VIVO data for evaluation and strategic forecasting for institutions and organizations

Submissions
Authors are invited to submit abstracts for poster, panel, and paper presentations related to the Topics of Interest. All paper submissions will be handled electronically. Authors can submit their abstract and the completed submission form at the conference submission site from https://www.easychair.org/conferences/?conf=vivo14

Abstracts will undergo the peer-review process and the reviewing process is strictly confidential. Proposals must be submitted as a one-page (8.5 x 11 inch) document [PDF or MS Word format]. Each submission should not exceed one page typeset in 12-point font and must include:

• The title of the submission and the names, academic degree(s), affiliations, and locations (city, state, and country) of all authors
• An abstract summarizing the submission and offering the reviewers a clear reflection of the contents and key points of the proposed presentation, panel, or poster.

Categories of Submissions
• Presentations: Papers are requested on topics related to those listed in the Topics of Interest.
• Authors will be considered for either a full-length presentation of 45 minutes, with 10 minutes for questions or for inclusion into a collection of presentations to be delivered during a single session. These shorter presentations will be 15 minutes long with 5 minutes for questions. Please indicate your preference (short or long talk) on your submission.
• Posters: Poster presentations offer an excellent opportunity to present preliminary research and projects and allow for stimulating dialog on the topic at hand.
• Panels: Panel presentations offer an opportunity to address a topic in a group format. Panels may be organized around a specific topic or may offer an innovative approach that cuts across multiple topic areas, technologies, experiences, or disciplines. There will usually be no more than four panelists in a given panel session.

Important Dates
Accepted papers will be presented by their author(s) and will be published in the conference proceedings. The proceedings are provided free of charge to conference attendees. For further details or inquiries, please contact Program Chair, Dean Krafft at conference2014@vivoweb.org.
• Abstracts due: April 4, 2014
• Decision: May 2, 2014

About VIVO
VIVO is an open source, open ontology, open process platform for hosting information about the interests, activities and accomplishments of scientists and scholars. VIVO supports open development and integration of science and scholarship through simple, standard semantic web technologies. Learn more at http://vivoweb.org.

Abstract
Source: https://www.easychair.org/conference...1;track=102151

Data Science for VIVO and the IV MOOC

Dr. Brand Niemann
Director and Senior Data Scientist
Semantic Community

http://semanticommunity.info
http://www.meetup.com/Federal-Big-Data-Working-Group/
http://semanticommunity.info/Data_Science/Federal_Big_Data_Working_Group_Meetup

Semantic Community supports VIVO by using an open source, open ontology, open process platform for information about the interests, activities and accomplishments of scientists and scholars much like DuraSpace. Semantic Community hosts and co-organizes the Federal Big Data Working Group Meetup which features Semantic Data Science Teams doing open development and integration of science and scholarship through simple, standard semantic web technologies.

Three very interesting developments from the VIVO 2013 Conference led to this presentation:
• Indiana University CNS: Vivo Data and Visualizations which was a "Data Science" tutorial
• VIVO’s Evolution from Semantic Web Application to DuraSpace; and
• Indiana University’s Evolution from VIVO Information Visualization to Data Science

My previous work was Data Science for VIVO:
• Atlas of Science
• An Interface to a Digital Library of the Atlas of Science
  ◦ 2010 Science and Engineering Indicators
  ◦ Nobel Prizes
• Build VIVO in the Cloud
• NIH Workshop on Value Added Services for VIVO
  ◦ Build the Scholarly Databases in the Cloud
So my goal became to use the "Data Science" tutorial data sets and the Information Visualization MOCC data sets in Spotfire and NodeXL to audit their results and report on them. In my professional opinion, Data Science is really about answering three basic questions:

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

in a manner that another data scientist can come along and readily reproduce or not-reproduce, whichever the truth may be. I expressed reservation about being able to do that with the IU VIVO work previously because of their use of copyrighted/proprietary data and/or lack of transparent scholarship.

Our March 4th Meetup of the Federal Big Data Working Group Meetup, featured Dr. Barend Mons, who keynoted at the 2010 VIVO Conference, presenting his work on Euretos BRAIN which uses semantic web technologies on biomedical data to produce nanopublications. One of the key semantic technologies he uses is the Semantic Medline - YarcData Graph Appliance Application developed by the Federal Big Data Working Group Meetup. He mentioned his recent recommendations for the future direction of VIVO in his presentation.

This presentation will show my work as a data scientist/data journalist to structured/visualized the Information Visualization MOCC as a Knowledge Base for ease of reuse by linking the slides and extracting key points, links to data, and my data science commentary. This data science work should also be viewed as creating a "data paper" for an upcoming workshop on The Semantics for e-science in an intelligent Big Data context from which cardinal assertions can be extracted for nanopublications.

Keywords

Semantic Community is like DuraSpace
Federal Big Data Working Group Meetup features Semantic Data Science Teams
Previous Data Science work for VIVO
Dr. Barend Mons Euretos BRAIN and nanopublications
Data Papers for Semantic e-science
Information Visualization MOCC Knowledge Base and Visualizations

Slides

Slides
Slide 1 Data Science for VIVO

Slide 2 Overview 1

Overview 1

- Data Fair Port Unconference 2014:
  - Discussed at Federal Big Data Working Group Meetup March 4th on Joint NSF-NIH Biomedical Big Data Research
  - Dr. Barend Mons referred to ORCID (Open Researcher and Contributor ID) VIVO in October 2012 Presentation
  - See Dr. Barend Mons 2010 VIVO Conference Video "The Next Step in Knowledge Evolution; Colonization of Brains..."

- My Previous Work Was Data Science for VIVO:
  - Atlas of Science (See next slide)
  - An Interface to a Digital Library of the Atlas of Science
    - 2010 Science and Engineering Indicators
      - Nobel Prizes
    - Build VIVO in the Cloud
    - NIH Workshop on Value Added Services for VIVO
      - Build the Scholarly Databases in the Cloud

Slide 3 Atlas of Science

Web Player
Slide 4 Overview 2

Overview 2

- VIVO’s Evolution from Semantic Web Application to DuraSpace:
  - VIVO is an open community, an information model, and an open source semantic web application.
  - VIVO is joining DuraSpace as the VIVO Project, to realize a shared vision through an established legal and financial framework compatible with VIVO’s own vision and goals.
  - DuraSpace is an independent 501(c)(3) non-profit born from a vision to help save our shared scholarly, scientific, and cultural record.

- Indiana University’s Evolution from VIVO Information Visualization to Data Science:
  - This course provides an overview about the state of the art in information visualization. It teaches the process of producing effective visualizations that take the needs of users into account.
  - This year, the course can be taken for three Indiana University credits as part of the Online Data Science Program just announced by the School of Informatics and Computing. Students interested in applying to the program can find more information here.
  - Everyone who registers gains free access to the Scholarly Database (26 million paper, patent, and grant records) and the Sci2 Tool (100+ algorithms and tools).

Slide 5 Vivoweb.org

http://www.vivoweb.org/
Slide 6 Indiana University CNS: Vivo Data and Visualizations


Slide 7 Indiana University CNS: VIVO Presentation Information

http://wiki.cns.iu.edu/display/PRES/...on+Information
Slide 8 Indiana University CNS: 2.5 Sample Data Sets

http://sci2.wiki.cns.iu.edu/display/...ample+Datasets

Slide 9 Cyberinfrastructure for Network Science Center

http://cns.iu.edu/
Slide 10 Information Visualization MOOC

http://ivmooc2014.appspot.com/home?
Pre-Questionnaire for the IVMOOC

Emails for the IVMOOC

- IVMOOC: Welcome to the Sci2 Tool
  - Thank you for registering for the Information Visualization MOOC. As part of this course, you will be using the Science of Science (Sci2) tool, an open source data mining and visualization software package developed by the team that runs the course.
  - We have created an account on the Sci2 website for you. Please use the link below to activate the account.
- IVMOOC: SOAP Account Confirmation
  - Thank you for registering for the Information Visualization MOOC. As part of this course, you will be asked to use data from the Scholarly Database, an open data source maintained by the team that runs the course. Our records show that your e-mail address is already registered for the Scholarly Database. Please take a moment to visit the below site and confirm that you remember your login information.
- IVMOOC: Welcome to the Information Visualization MOOC
  - Thank you for registering for the Information Visualization (IVMOOC) course. This email is to confirm receipt of your registration information. Unit one of the course will open on January 28, 2014, with another Unit opening each week. The course introduction posted on week one will include all the information about what you will need to do in order to earn the Mozilla Open Badge.
  - You should receive two emails in addition to this one that contain your login information for the Sci2 Tool and Scholarly Database, two free tools that you will be using throughout the course.

IVMOOC - Schedule

http://ivmooc2014.appspot.com/course
Slide 14 IVMOOC - Week 5

http://en.wikipedia.org/wiki/Grading_on_a_curve
http://ivmooc.cns.iu.edu

IVMOOC - Week 5

- Many of you did a great job on the Mid-Term! All students who submitted the Mid-Term will receive a pointer to the correct answers shortly. For your information, the Mid-Term will be graded on a curve, see:
  - http://en.wikipedia.org/wiki/Grading_on_a_curve
- Week 5 materials are now available at:
  - http://ivmooc.cns.iu.edu
  - Enjoy analyzing and visualizing trees.
Slide 15 IVMOOC – Student Locations

ICMOOC – Student Locations

Geospatial Visualization (Proportional Symbol Map) of 2014 IVMOOC Student Locations

Legend

Slide 16 Indiana University CNS: Vivo Data and Visualizations Overview

http://nrm.cns.iu.edu

Indiana University CNS:
Vivo Data and Visualizations Overview

Presentation Overview

Part One
• Introduction
• VIVO Visualization Review
• VIVO Data Storage and Retrieval
  • Ontology
  • Basic SPARQL
  • Data Cleaning
• Introduction to Sos2
• Temporal Bar Graph
• Break

Part Two
• Analyzing and visualizing VIVO data
  • Burst Detection
  • Map of Science
  • Network
• The next level of aggregated data
  • PRIN: http://nrm.cns.iu.edu/
  • VIVO: http://nrm.cns.iu.edu/
  • VIVO: Search
• Continuing Education and Opportunities

Slide 17 The UCSD Map of Science

http://www.plosone.org/article/info%.../pone.0039464
The UCSD Map of Science

Originally created in 2006 and updated in 2011.

Created 554 clusters of journals called sub-disciplines based on citations and keywords, divided into 13 disciplines.

Maps over 25,000 journals from Scopus and Thomson Reuters’ Web of Science as of 2011 update.

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0039464

Published But Not the Raw Data

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0039464
Slide 19 Indiana University CNS: Vivo Data and Visualizations Sample Data

Indiana University CNS: Vivo Data and Visualizations Sample Data

My Note: These are all very small compared to the UCD Map of Science and Scholarly Database!

Slide 20 Temporal Visualization: Grants Over Time

Temporal Visualization: Grants Over Time

My Note: Not Impressed With This!

Slide 21 Statements I Disagree With!

http://vivo-netsci.cns.iu.edu/vivo12...lay/Grant78086
http://vivo-netsci.cns.iu.edu/vivo12...Grant78086.rdf
Statements I Disagree With!

- Visualization brings large amounts of data together in a single graphic to allow for greater understanding.
- All data is stored in the same format, via triples (subject, predicate, and object).
- There is no table structure to guide data collection in a triple store. Instead, you have the ontology. The ontology defines how data is linked together, and is critical when it comes to retrieving that data.
- With some understanding of the ontology, we're ready to start writing SPARQL queries to get a data table for the visualization.
- VIVO does not store information based on the name, as we would consider it. Instead, it assigns every entity a Uniform Resource Identifier (or URI).
- Here is the core data behind this, see the problem?
- Excel is generally the one that most people are familiar with, but it is rather limited in the ability to replicate work easily.
- In the end, your final output can only be as good as the data that went into it.

Slide 22 Exercises and Data Sets

Exercises and Data Sets

- Burst Detection:
  - Use publication_result.csv
  - Normalize publication title by using Lowercase, Tokenize, Stem, Stopword Text plugin.
  - Save as burst.csv
- Map of Science:
  - Use publication_result.csv
  - The Journal Naming Convention plugin tries to normalize these names and link them to lookup tables within Sc2 to create the best matching possible.
- Network:
  - Use publication_result.csv
  - Extract the Co-Author network using Extract Co-Occurrence Network plugin.
  - Visualize Using GUESS and Gephi. My Note: Use NodeXL

Slide 23 iNRN

http://nrn.cns.iu.edu/#
Slide 24 Theory Unit Structure

http://ivmooc.cns.iu.edu/slides/01-schedule.pdf

Theory Unit Structure

• Each theory unit comprises:
  – Examples of best visualizations
  – Visualization goals
  – Key terminology
  – General visualization types and their names
  – Workflow design
    • Read data
    • Analyze
    • Visualize
  – Discussion of specific algorithms

http://ivmooc.cns.iu.edu/slides/01-schedule.pdf

Slide 25 Data Science for VIVO Knowledge Base

http://semanticommunity.info/Data_Science/Data_Science_for_VIVO
Data Science for VIVO Knowledge Base

Slide 26 Scholarly Database: MEDLINE data Sets in Spotfire

Scholarly Database: MEDLINE Data Sets in Spotfire
Slide 27 VIVO Workshop August 2013 Selected Data Visualizations in Spotfire

Slide 28 SCI2: Full Sample Selected Data Visualizations in Spotfire

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

Course Slide:

Client Work: Using Drupal Marketplace (peer review)

Client work is peer reviewed via online forum.

Tutorial Slide:

Try to visualize the following networks using the grant_result.csv data:
Co-PI Network using PI column http://wiki.cns.iu.edu/pages/viewpag...pagId=2785284

Co-occurrence word network using title column http://wiki.cns.iu.edu/pages/viewpag...urrenceNetwork

Work on bipartite network such as PI to Agency network, Grant to PI network, etc. http://wiki.cns.iu.edu/pages/viewpag...pagId=2785293

Overview

Source: http://ivmooc2014.appspot.com/home

This course provides an overview about the state of the art in information visualization. It teaches the process of producing effective visualizations that take the needs of users into account.

This year, the course can be taken for three Indiana University credits as part of the Online Data Science Program just announced by the School of Informatics and Computing. Students interested in applying to the program can find more information here.

Among other topics, the course covers:

- Data analysis algorithms that enable extraction of patterns and trends in data
- Major temporal, geospatial, topical, and network visualization techniques
- Discussions of systems that drive research and development.

Everyone who registers gains free access to the Scholarly Database (26 million paper, patent, and grant records) and the Sci2 Tool (100+ algorithms and tools).

Please watch the introduction video to learn more.

My Note: Watched video - 10 years of data mining and visualization work, plug and play macrosopes, Type of Analysis vs. Level of Analysis, Needs-Driven Workflow Design. Complete the Profiles for Team assignments.
Schedule

Source: http://ivmooc2014.appspot.com/course

Pre-Questionnaire

Source: https://docs.google.com/spreadsheet/...VYdUE6MA#gid=0

My Note: Completed the pre-questionnaire

Please answer the questions below to help us understand your work practice, needs, and experience.

* Required

Enter your Google ID (the one you’re using for this MOOC) *

Your Work Practices

In your daily work, what Datasets do you use? Open Government Data

In your daily work, what Software/Tools do you use? MindTouch, Excel, and Spotfire

In your daily work, what Hardware do you use (desktop, laptop, iPad, PDA), mostly online or offline? All and mostly online

Have you designed a temporal visualization before?  
Yes  
No

Have you designed a geospatial visualization before?  
Yes  
No

Have you designed a topical visualization before?  
Yes  
No

Have you analyzed or visualized a network before?  
Yes  
No

Your Needs

What visualizations are/would be most helpful for your daily decision making? See my Data Stories

What research questions or practical questions would you like to answer? See my Data Stories
What Software/Tool functionality do you miss/need? None

What would you most like to learn in the Information Visualization MOOC? How to use this in my Data Science Tutorials and Classes

Your Expertise

What data scale type is ‘day of the week’?
categorical (nominal)
ordinal
interval
ratio

What data scale type is ‘Fahrenheit temperature’?
categorical (nominal)
ordinal
interval
ratio

When performing time series analysis, what does it mean to use ‘cumulative’ time slices?
Every row in the original data table is in exactly one time slice.
Selected rows are in multiple time slices.
Every row in a time slice is in all later time slices.

Which of the below steps is NOT relevant for text normalization in preparation for topical analysis and visualization?
Lowercase words
Remove stop words
Stemming
Extract co-occurrence network
Tokenization

What visualization should be used to show how many observations of a certain value have been made?
Line graph
Temporal bar graph
Scatter plot
Histogram

What map type is best to visualize population density?
Choropleth map
Proportional symbol map

Unit 1 – Visualization Framework & Workflow Design

Theory
Welcome by Katy Börner (1:57)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=1&lesson=1

My Note: Watched Video

Course Overview (11:36)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=1&lesson=2

SLIDES: http://ivmooc.cns.iu.edu/slides/01-schedule.pdf (PDF)

Instructors

Katy Börner – **Theory Parts**, Instructor, Professor at SLIS
David E. Polley – **Hands-on Parts**, CNS Staff, Research Assistant with MIS/MLS, Teaches & Tests Sci2 Tool **My Note**: I use Spotfire
Scott B. Weingart – **Client Work**, Assistant Instructor, SLIS PhD student

Unit Structure
The course and each unit has three components:
Theory: Videos and Slides
Self-Assessment (not graded)
Hands-on: Videos and Slides & Wiki pages with workflows
Homework (not graded)
Client Work: Using Drupal Marketplace (peer review) **My Note**: I use MindTouch

Theory Unit Structure
Each theory unit comprises:

- Examples of best visualizations
- Visualization goals
- Key terminology
- General visualization types and their names
- Workflow design
  - Read data
  - Analyze
  - Visualize
- Discussion of specific algorithms

Twitter: “ivmooc” **My Note**: I tweeted

Book, Web Sites, and References **My Note**: I have read Atlas of Science.
**Visualization Framework (28:59)**


**SLIDES:** [http://ivmooc.cns.iu.edu/slides/01-framework.pdf](http://ivmooc.cns.iu.edu/slides/01-framework.pdf) (PDF)

### Type of Analysis vs. Level of Analysis

<table>
<thead>
<tr>
<th>Statistical Analysis/Profiling</th>
<th>Individual person and their expertise profiles</th>
<th>Larger labs, centers, universities, research domains, or states</th>
<th>All of NSF, all of USA, all of science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Analysis (When)</td>
<td>Funding portfolio of one individual</td>
<td>Mapping topic bursts in 20 years of PNAS</td>
<td>113 years of physics research (1)</td>
</tr>
<tr>
<td>Geospatial Analysis (Where)</td>
<td>Career trajectory of one individual</td>
<td>Mapping a state’s intellectual landscape</td>
<td>PNAS publications</td>
</tr>
<tr>
<td>Topical Analysis (What)</td>
<td>Base knowledge from which one grant draws</td>
<td>Knowledge flows in chemistry research</td>
<td>VxOrd/Topic maps of NIH funding</td>
</tr>
<tr>
<td>Network Analysis (With Whom?)</td>
<td>NSF Co-PI network of one individual</td>
<td>Co-author network</td>
<td>NIH’s core competency</td>
</tr>
</tbody>
</table>


**My Note:** Are any of the data sets publicly available for these examples?

**Workflow Design (19:40)**


**SLIDES:** [http://ivmooc.cns.iu.edu/slides/01-workflow.pdf](http://ivmooc.cns.iu.edu/slides/01-workflow.pdf) (PDF)
Visualization Types (Reference Systems)

1. Charts: No reference system—e.g., Wordle.com, pie charts
2. Tables: Categorical axes that can be selected, reordered; cells can be color coded and might contain proportional symbols.
   Special kind of graph.
3. Graphs: Quantitative or qualitative (categorical) axes. Timelines, bar graphs, scatter plots.

Data Overlays

Given a reference system, also called base map (see Visualization Types), data record attributes can be used to
1. Modify base map—e.g., distort area sizes (cartogram) and/or to visually encode base map areas (color by #life expectancy)
2. Place data records and visually encode nodes.
3. Place data record linkages and visually encode links.
Aggregations such as cluster boundaries or network backbones are encoded using steps 1-3 at different (semantic) zoom levels.
In addition, there is commonly a title, labels, legend, explanatory text, and author info.

Data Scale Types

**Categorical (nominal):** A categorical scale, also called nominal or category scale, is qualitative. Categories are assumed to be nonoverlapping.

**Ordinal:** An ordinal scale, also called sequence or ordered, is a qualitative. It rank-orders values representing categories based on some intrinsic ranking but not at measurable intervals.

**Interval:** An interval scale, also called value scale, is a quantitative numerical scale of measurement where the distance between any two adjacent values (or intervals) is equal but the zero point is arbitrary.

**Ratio:** A ratio scale, also called proportional scale, is a quantitative numerical scale. It represents values organized as
an ordered sequence, with meaningful uniform spacing, and has a true zero point.

**Data Scale Types - Examples**

**Categorical**: Words or numbers constituting the names and descriptions of people, places, things, or events.

**Ordinal**: Days of the week, degree of satisfaction and preference rating scores (e.g., using a Likert scale), or rankings such as low, medium, high.

**Interval**: Temperature in degrees or time in hours. Spatial variables such as latitude and longitude are interval.

**Ratio**: Physical measures such as weight, height, (reaction) time, or intensity of light; number of published papers, co-authors, citations.

**My Note**: I followed my Data Science process and method (see The CRISP Data Mining Process below) and use my Spotfire 6 for Data Science Users Guide to explain how to work with data and document it.

Figure 2-2.png
Self-Assessment

LINK: http://ivmooc2014.appspot.com/activities/unit=1&lesson=4

Week 5 - Feb. 25, 2014: 'With Whom': Trees My Note: This seems strange!

Hands-on

Introduction by Ted Polley (:47)


Download Tool, Install, and Visualize Data with Sci2 (10:54)


WIKI: http://wiki.cns.iu.edu/display/SCI2Tool%2C+Uninstall

My Note: I did not use this because I use Spotfire to create and visualize the data ecosystem.

Legend Creation with Inkscape (16:03)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=1&lesson=7

WIKI: http://wiki.cns.iu.edu/display/SCI2Tool%2C+Publication

My Note: I did not use this because I use Spotfire to create and visualize the data ecosystem.

Weekly Tip: Extend Sci2 by adding Plugins (3:13)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=1&lesson=8

WIKI: http://wiki.cns.iu.edu/display/SCI2Tool%2C+and+Save+Data

My Note: I did not use this because I use Spotfire to create and visualize the data ecosystem.

Homework

LINK: http://ivmooc2014.appspot.com/activities/unit=1&lesson=8

For the first homework assignment you should go to the "My Profile" link in the navigation menu of the course site and set up a profile. You will be prompted to login again with the Google ID you used to sign up for this course. From there you will need to fill out your profile with as much information as you can. The more you share about yourself, the easier it will be to connect with your fellow students in this course to form teams for the client work. Start to form groups of 4-5 students to begin the client work.
Make sure to share your Twitter and Flickr ID’s in your profile so you can share the visualizations you create and the insight you gain from this course. If you do not already have a Twitter and Flickr account, go ahead and set one up for this course.

My Note: I did this - see http://ivmooc.cns.iu.edu/forums/user/705

Unit 2 – “When”: Temporal Data

Theory

Welcome by Katy Börner (:39)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=2&lesson=1

My Note: Watched Video

Exemplary Visualizations (9:46)


SLIDES: http://ivmooc.cns.iu.edu/slides/02-exemplary.pdf (PDF)

My Note: Exemplary, but easily understood without the data story and reproduced without the data.

Overview and Terminology (16:38)


SLIDES: http://ivmooc.cns.iu.edu/slides/02-terminology.pdf (PDF)

Temporal Analysis and Visualization Goals

Main goals:
• Understand temporal distribution of a dataset—e.g., first and last time point, any zero or missing values, trends, growth, latency to peak, decay rate.
• Forecasting—i.e., predicting future values of the time-series variable.

Patterns in Time-Series Data
• Trends: General tendency such as steadily shifting, stabilizing, or cyclic—e.g., increase in spam email. Frequently, some form of filtering is applied to reduce noise and to make patterns more salient—e.g., averaging using window of a certain width or curve approximation/fitting.
• Seasonality: Repetitive and predictable movement around a trend line—e.g., cyclic variations of flu infections, crops harvested, construction workers employed per month, emails received at night/day.
• Burst Analysis: Identification of sudden bursts of activity (e.g., right before a deadline) or in response to external events—e.g., disasters.
Terminology
A time series is a sequence of events/observations which are ordered in one dimension—time. Time-series data can be continuous—i.e., there is an observation at every instant of time—or discrete—i.e., observations exist for regularly or irregularly spaced intervals.

General Visualization Types
Graphs
• Line graph: Show trends over time.
• Stacked graph: See individual and total trends.
• Temporal bar graph: Show begin, end, and properties of events—see Hands-on session.
• Scatter plot: See relationships—e.g., correlations between two data variables.
• Histogram: How many observations of a certain value have been made.

Geomap: Understand change over time in geospatial distribution, see Unit 3.
Topic map: Understand change over time in topical distribution, see Unit 4.
Network Graph: Understand change over time in topical distribution, see Units 5 & 6.

My Note: Why not provide some basic data sets and explore what one can learn from the data by trying various visualization types. What about tables and statistics as data visualizations?

Workflow Design (19:41)
VIDEO: http://ivmooc2014.appspot.com/unit?unit=2&lesson=4
SLIDES: http://ivmooc.cns.iu.edu/slides/02-workflow.pdf (PDF)

Read Data
Data Repositories:
• Gapminder data, http://www.gapminder.org/data/
(Time Series)
• Scholarly Database, http://sdb.cns.iu.edu
• IBM ManyEyes Datasets, http://www958.ibm.com/software/data/...yeyes/datasets (350,976 )
• Eurostat Data Market, http://datamarket.com

Data Formats:
• TXT
• XLS, CSV
• Databases

My Note: There are certainly more and better sources of data than these
Data Preprocessing
Filtering—e.g., time slicing (see next slide)
Test for and remove large spikes in the data, but report this.

Normalization
• Deduplication
• Unit conversion
• Adjust (dollars) for inflation
• Adjust for time zones

Integration/interlinkage of different data sources
Classification/aggregation

My Note: There is much more to data preprocessing than this.

Top-250 Movies from IMDb
Copy from web page at http://www.imdb.com/chart/top (on Nov 15, 2012), save as text file, open in Excel or other table editing program: Is there a correlation between Rating and #Votes? Any trends over time? Recent movies more popular?

My Note: This is an interesting data set to look at.

Relevant Tools
• TimeSearcher from HCIL supports the visual exploration of time-series data http://www.cs.umd.edu/hcil/timesearcher/
• Tableau, http://www.tableausoftware.com

My Note: These are certainly not all the "relevant tools"!

Burst Detection (14:14)


SLIDES: http://ivmooc.cns.iu.edu/slides/02-burst.pdf (PDF)

Kleinberg’s burst-detection algorithm identifies sudden increases in the frequency of words. Given time-stamped text, it identifies words that burst.

My Note: This is new to me. I use Recorded Future for detection.

Text Normalization (see Topical Analysis)
Sample text: “Emergence of Scaling in Random Networks”
• Lowercase: The example text becomes “emergence of scaling in random networks.”
• Tokenize: The text blob is split into a list of individual words. The example text becomes “emergence|of|scaling|in|random|networks.”
• Stem: Common or low-content prefixes and suffixes are removed to identify the core concept. The example
text becomes "emerg[of]scale[in]random[network]."

- Stopword: Low-content tokens like "of" and "in" are removed (see the complete stopword list). The example text becomes "emerg[scale][random][network]."

My Note: Is Burst Detection just a subset of Natural Language Processing?

Self-Assessment

LINK: http://ivmooc2014.appspot.com/activi...nit=2&lesson=5

My Note: There is nothing here.

Hands-on

Introduction by Ted Polley


Temporal Bar Graph: NSF Funding Profiles (9:44)


WIKI: http://wiki.cns.iu.edu/pages/viewpag...pageld=2200061

Burst Detection in Publication Titles (11:18)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=2&lesson=8

WIKI: http://wiki.cns.iu.edu/pages/viewpag...pageld=2785326

Weekly Tip: Sci2 Log Files (3:51)


WIKI: http://wiki.cns.iu.edu/display/SCI2T.../3.7+Log+Files

My Note: I think my hands-on with Spotfire is much more effective.

Homework

LINK: http://ivmooc2014.appspot.com/activi...nit=2&lesson=9

This week you should login to the Scholarly Database (http://sdb.cns.iu.edu) with the password you set during the registration process at the beginning of this course. Select the "MEDLINE" dataset and enter the Search term "mesothelioma" in the title field:

Scholarly Database. Conducting a Search Screenshot
From the search results page, select download. Then, download the first thousand results and save the MEDLINE master table. It will download as a .csv file:

Scholarly Database Downloading Results

Load the .csv file into Sci2 and conduct a burst analysis, similar to the one conducted in this week’s hands-on video. Once you have completed the analysis compare the results to the Wikipedia article for mesothelioma to see support for the visualization you have just created.

Report insight via Twitter, making sure to use #ivmooc.

My Note: I did this.

Semantic MEDLINE Query: mesothelioma


Source: Medline, Semantic Medline, Dr. Tom Rindflesch, NIH/NLM/LHC trindflesch@mail.nih.gov

Most Recent: 500 citations,

Start Date: 01/01/1900,

End Date: 11/30/2013,

3169 predications extracted.

Summarized for Substance Interactions:

An overview of current research on mesothelioma.

For example

Diagnosis:
Hyaluronic Acid ASSOCIATED_WITH Malignant Pleural Mesothelioma

PMID:24161718

Date of Publication: Dec 2013

Title: Pleural effusion hyaluronic acid as a prognostic marker in pleural malignant mesothelioma.

Treatment:

Pemetrexed TREATS Malignant Pleural Mesothelioma

PMID:24023280

Date of Publication: Sep 2013

Title: The value of pemetrexed for the treatment of malignant pleural mesothelioma: a comprehensive review.

Etiology:

PMID:23624653

Date of Publication: Jul 2013

Title: Overexpression of Numb suppresses tumor cell growth and enhances sensitivity to cisplatin in epithelioid malignant pleural mesothelioma

Unit 3 – “Where”: Geospatial Data

Theory

Welcome by Katy Börner (:54)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=3&lesson=1

My Note: Watched Video

Exemplary Visualizations (5:55)


SLIDES: http://ivmooc.cns.iu.edu/slides/03-exemplary.pdf (PDF)

My Note: Impact of Air Travel on Global Spread of Infectious Diseases - Vittoria Colizza, Alessandro Vespignani, and Elisha F. Hardy - 2007 appears to have a "built-in" data story.
Overview and Terminology (11:09)


SLIDES: http://ivmooc.cns.iu.edu/slides/03-terminology.pdf (PDF)

Thematic Map Types
Classification according to content:
• Physio-geographical maps: geological, geophysical, meteorological, soils, vegetation
• Socio-economic maps: historical, political, population, economy, cultural, voting, epidemics
• Technical maps: navigation, cadastre (shows boundaries and ownership of land parcels), civil engineering

General Map Types
Emphasis on location:
• General reference maps
• Topographic maps
• Thematic maps

Focus here is on thematic maps that emphasize the spatial distribution of one or more geographic variables.

Representation of Geospatial Data
• Addresses
• US Zip codes, see http://benfry.com/zipdecode
• US Census blocks
• US Congressional districts
• US States
• Countries
• Latitude/Longitude

Terminology
• Geocode: Location of a record (e.g., address, census tract, postal code, geographic coordinates).
• Geographic coordinates: Locations on the surface of the Earth expressed in degrees of latitude and longitude.
• Geodesic: The shortest distance between two points on the surface of a spheroid.
• Great Circle: Shortest distance between two points on Earth—i.e., a circular line which runs around the Earth at its fattest point.
• Gazetteers: Lists of geographic places and their coordinates, along with other information such as area, population, and cultural statistics used to geocode—see Yahoo! Geocoder in Hands-on.

<table>
<thead>
<tr>
<th>Proportional symbol map</th>
<th>Choropleth map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Represents data variables by symbols that are sized, colored, etc. according to their amount. Data is (or can be) aggregated at points within areas. Do NOT use for</td>
<td>Represents data variables such as densities, ratios, or rates by proportionally colored or patterned areas. Each artificial collection unit is called a chronogram and</td>
</tr>
<tr>
<td>densities, ratios, or scales, which should be rendered as choropleth map.</td>
<td>has a distinctive color or shading.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Heat (isopleth) maps</strong> represent continuous data variable values by colors. While choropleth maps color predefined regions, heat maps might show color-based contour lines that connect points of equal value or value-by-area maps.</td>
<td><strong>Cartograms</strong> are not drawn to scale. Instead, they distort geographical areas in proportion to data values. Familiarity with regions is necessary. Mostly used for world, continental, and country maps.</td>
</tr>
<tr>
<td><strong>Flow maps</strong> show the paths that (in)tangible objects take to get from one geospatial place to another. Variables such as capacity or maximum speed are encoded proportionally by line width or color.</td>
<td><strong>Space-time cubes</strong> Display entities, locations, and events over time.</td>
</tr>
</tbody>
</table>


**My Note:** I have analyzed the 2012 Presidential Campaign data and found that maps tell only part of the story.

**Workflow Design (6:39)**


**SLIDES:** [http://ivmooc.cns.iu.edu/slides/03-workflow.pdf](http://ivmooc.cns.iu.edu/slides/03-workflow.pdf) (PDF)

**Read Data**

Data Repositories:
- Digging into Data list of repositories, [http://www.diggingintodata.org/Repos...7/Default.aspx](http://www.diggingintodata.org/Repos...7/Default.aspx)
- Scholarly Database, [http://sdb.cns.iu.edu](http://sdb.cns.iu.edu)

**Data Formats:**
- Vector—e.g., Shape files, PS
- Raster—e.g., TIFF, JPEG
- Tabular—e.g., in CSV
- Software specific

**Preprocess Data**
- Thresholding
- Unification, see Gazetters
- Aggregation, see Aggregation
Example: “Mapping the Diffusion of Information among Major U.S. Research Institutions”

Need a compromise between maintaining geographic identity and statistical significance. For example, should IU be represented as one or eight campuses? IUB has two ZIP codes—represent IUB by one or two places?


My Note: I would use the Mapping section of my Spotfire 6 for Data Science Users Guide to explain how to work with spatial data and document it.

Relevant Tools

• List of GIS software, also open software, http://en.wikipedia.org/wiki/List_of_...stems_software
• Data-Driven Documents (D3), http://d3js.org
• Tableau, http://tableausoftware.com
• IBM’s Many Eyes, http://www958.ibm.com/software/data/cognos/manyeyes

My Note: Again I use the Mapping section of my Spotfire 6 for Data Science Users Guide

Color (8:41)


SLIDES: http://ivmooc.cns.iu.edu/slides/03-color.pdf (PDF)

Color

Use to
• convey importance or attract attention to specific symbols
• label, categorize, compare
• imitate reality (e.g., blue lakes in maps)
• generate emotions—orange and red are perceived as warm and active while blue, purple are cold and passive.

Do NOT use
• for displaying the layout of objects in space
• how they are moving, or
• what their shapes are.

Color Brewer: http://colorbrewer2.org

My Note: I am not familiar with this, but rely on the Color section of my Spotfire 6 for Data Science Users Guide

Self-Assessment

LINK: http://ivmooc2014.appspot.com/activi...nit=3&lesson=5
**My Note:** There is nothing here.

**Hands-on**

**Introduction by Ted Polley:**


**Choropleth and Proportional Symbol Map:**


*WIKI:* [http://wiki.cns.iu.edu/pages/viewpag...pageld=3901879](http://wiki.cns.iu.edu/pages/viewpag...pageld=3901879)

**Congressional District Geocoder:**


*WIKI:* [http://wiki.cns.iu.edu/display/SCI2T...trict+Geocoder](http://wiki.cns.iu.edu/display/SCI2T...trict+Geocoder)

**Geocoding NSF Funding with the Generic Geocoder:**


*WIKI:* [http://wiki.cns.iu.edu/display/CISHELL/Geocoder](http://wiki.cns.iu.edu/display/CISHELL/Geocoder)

**Weekly Tip: Memory Allocation:**


*WIKI:* [http://wiki.cns.iu.edu/display/SCI2T...ory+Allocation](http://wiki.cns.iu.edu/display/SCI2T...ory+Allocation)

*My Note:* I think my hands-on with Spotfire is much more effective.

**Homework**

*LINK:* [http://ivmooc2014.appspot.com/activi...it=3&lesson=10](http://ivmooc2014.appspot.com/activi...it=3&lesson=10)

For this assignment, download some NSF data of your choosing from the Scholarly Database ([http://sdb.cns.iu.edu](http://sdb.cns.iu.edu)) and create a geospatial visualization of your choosing. It does not have to be very complicated; it can be as simple as geocoding the addresses in an NSF file and using the proportional symbol map to map NSF funding.

Once you have a visualization that you are satisfied with you can upload the image to Flickr or Twitter.

Don’t forget to tag the image with #ivmooc.
My Note: I did this prior to the course.

**EPA Waterways**

Source: [http://semanticommunity.info/Data_Science/EPA_Waterways](http://semanticommunity.info/Data_Science/EPA_Waterways)

Data Science for Business: Specific Example, Brand Niemann, Director and Senior Data Scientist, Semantic Community, working for a client to create a 2.5 GB Spotfire file of all of [EPA Waterways](http://semanticommunity.info/Data_Science/EPA_Waterways) data.

![Image of EPA Waterways data](http://semanticommunity.info/Data_Science/EPA_Waterways)

---

**Unit 4 – “What”: Topical Data**

**Theory**

Welcome by Katy Börner (1:13)


My Note: Watched Video

**Exemplary Visualizations (9:09)**


Reference


My Note: Where is the data to learn from these visualizations?
Overview and Terminology (10:15)


SLIDES: http://ivmooc.cns.iu.edu/slides/04-terminology.pdf (PDF)

Representations of Topical Data
• Charts: Word cloud
• Tables: GRIDL
• Graphs: MDS plots, circular visualization, Crossmaps
• Geospatial maps: SOM maps
• Network graphs: Tree visualizations, word co-occurrence networks, concept maps, science map overlays

Topical Analysis and Visualization Goals
Main goals are to understand
• Topical distribution of a dataset, e.g., what topics are covered and how much.
• How topics emerge, merge, split, or die.
• Bursts of topics, see Unit 2 on ‘Temporal Analysis’
• Topical change over time, see Unit 2

Topical analyses at different levels of aggregation are common.
Analyses may range from micro to macro—e.g.
• single documents (micro), journal/book volumes, scientific disciplines (macro), or
• single individuals (micro), institutions, or countries (macro)

Terminology
Text: A sequence of written or spoken words.
Text corpus: A large and structured set of texts (e.g., tweets, emails, books).
Topic: A noun phrase that expresses what a sentence is about.
N-gram: A subsequence of n items (e.g., phonemes, syllables, letters, words) from a given sequence.
Stop words: Very commonly used words (e.g., a, and, in) that are excluded from topical analysis.
Stemming: Process for reducing inflected (or sometimes derived) words to their stem, base, or root form.
Synonymy: Words or phrases alike in meaning or significance (e.g., happy, joyful, elated or close, shut).
Polysemy: The same word having many meanings (e.g., bank, crane).

GRIDL, developed at HCIL, uses categorical and hierarchical axes to support categorical zooming.

Concept maps are network graphs that show the relationships among concepts.

My Note: Again is this a subset of Natural Language Processing and what about Topic Maps?

Workflow Design (18:11)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=4&lesson=4
Read Data

Data Repositories:
- Digging into Data list of repositories, [http://www.diggingintodata.org/Repos...7/Default.aspx](http://www.diggingintodata.org/Repos...7/Default.aspx)
- Google ngrams datasets, text from millions of books scanned by Google
- Scholarly Database, [http://sdb.cns.iu.edu](http://sdb.cns.iu.edu)

Major Data Formats:
- TXT
- CSV

Preprocessing—Text Normalization

Sample text: Emergence of Scaling in Random Networks
- Lowercase: The example text becomes "emergence of scaling in random networks."
- Tokenize: The text blob is split into a list of individual words. The example text becomes "emergence|of|scaling|in|random|networks."
- Stem: Common or low-content prefixes and suffixes are removed to identify the core concept. The example text becomes "emerg|of|scale|in|random|network."
- Stopword: Low-content tokens like "of" and "in" are removed (see the complete stopword list). The example text becomes "emerg|scale|random|network."
- Identification of synonymy and polysemy.

Topical Analysis

- Frequency analysis
- Clustering/Classification
- Sentiment analysis
- Burst analysis, see Unit 1
- Dimensionality reduction, see ARIST chapter

Using a Dictionary and Thesaurus

Sorted by relevance, occurrences, select ‘geography’ words

My Note: I need to look at this tool and data source

Chart Example: Word Cloud

Wordle.net of Titles – create your own at [http://wordle.net](http://wordle.net)
Layout: Oval space filling; frequent words are closer to center
Type font size: Word frequency
Font color: No meaning, but different colors help legibility

**My Note:** I need to look at this tool

**Co-Occurrence Network of IMDb Movie Title Words**

**My Note:** I need to work with these data

**Relevant Tools**
- TextAnalyzer, [http://textalyser.net](http://textalyser.net)
- TexTrend (OSGi/CIShell compatible), [http://textrend.org](http://textrend.org)
- VOSviewer, [http://vosviewer.com](http://vosviewer.com)
See many more at [http://www.kdnuggets.com/software/text.html](http://www.kdnuggets.com/software/text.html)

**My Note:** I need to look at these tools

**Design and Update of a Classification System: The UCSD Map of Science (Optional) (24:48)**


**SLIDES:** [http://ivmooc.cns.iu.edu/slides/04-science-map.pdf](http://ivmooc.cns.iu.edu/slides/04-science-map.pdf) (PDF)

**My Note:** This is all about creating this map, but the data is not provided to actually do it!

**Design and Update of a Classification System: The UCSD Map of Science**
1. Original map
2. Initial Update Using Scopus
3. Final Updated Map
4. Validation
5. Applications

Deployment:
The UCSD map of science data is available at http://sci.cns.iu.edu/ucsdmap/

Data:
The 2010 UCSD map of science and classification system covers ten years (2001-2010) of data from Thomson Reuters’ Web of Science and eight years (2001-2008) of Elsevier’s Scopus, specifically the fractional assignment of about 25,000 journal names to 554 subdisciplines grouped into 13 disciplines of science.

The counts for major record types are given here:
1. 13 disciplines with labels and color codes
2. 554 subdisciplines with x, y positions and size
3. 15,849 journals captured by 5-year map
4. 25,258 journals captured by 10-year map
5. 13,520 journal names used by Thomson Reuters
6. 22,253 journal names used by Scopus
7. 21,630 Scopus journal ID numbers
8. 19,988 ISSN numbers
9. 66,759 terms

See Data Dictionary in Supplement 2 in http://www.plosone.org/article/info%...l.pone.0039464

UCSD map table schema

Sci2 Tool Usage at National Institutes of Health
Sci2 Tool now supports Web services and serves as a visual interface to publically available NIH RePORT Expenditure and Results RePORTER)/ RePORTER data provided by NIH.
My Note: I also know that NIH uses Spotfire extensively from a previous story I did.

Comparison of Text- and Linkage-Based Approaches (Optional) (9:01)


SLIDES: http://ivmooc.cns.iu.edu/slides/04-comparison.pdf (PDF)

Comparing the Accuracy of Text-Based Similarity Measures Using Five Analytical Techniques

Example: document-document relatedness

• Cosine similarity using term frequency-inverse document frequency vectors (tf-idf cosine)
• Latent semantic analysis (LSA)
• Topic modeling
• Two Poisson-based language models:
  – BM25
  – PMRA (PubMed Related Articles).


My Note: The work that Dr. Barend Mons and others are doing now seems more relevant to helping scientists published their data and reuse other's scientific data.

Self-Assessment

LINK: http://ivmooc2014.appspot.com/activi...nit=4&lesson=6

My Note: There is nothing here.

Hands-on

Introduction by Ted Polley (:54)


Mapping Topics and Topic Bursts in PNAS (9:31)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=4&lesson=8

WIKI: http://wiki.cns.iu.edu/pages/viewpag...pageId=1245862

Word Co-Occurrence Networks with Sci2 (14:35)

Weekly Tip: Removing Files from the Data Manager (1:30)


WIKI: http://wiki.cns.iu.edu/display/SCI2T...and+Save+Data

My Note: I think my hands-on with Spotfire is much more effective.

Homework

LINK: http://ivmooc2014.appspot.com/activi...it=4&lesson=10

This week you should select a database on the Scholarly Database (http://sdb.cns.iu.edu) (NSF, NIH, MEDLINE etc.) and search for a keyword of your choice in the title field. Take the first 1000 records and create a co-word occurrence of the words that appear in the title.

It will probably be helpful to run DrL. See the respective hands-on video from this week for guidance. Upload the image to Flickr or report any insight using Twitter. Don’t forget to use #ivmooc!

My Note: I did this prior to the course. I am not sure of the value of the co-word occurrence.

Euretos BRAIN

Source: http://semanticommunity.info/Data_Science/Euretos_BRAIN

Dr. Barend Mons

Mid-Term

My Note: Signed Up To Late
1: Visualization Framework and Workflow Design

Define and distinguish between the three levels of analysis:
-- micro, meso, and macro

Understand the interaction between the key elements of the iterative workflow design:
-- stakeholders, reading data, analyzing data, visualizing data, deploying result

Understand for which kinds of data each of these analysis types are used:
-- statistical, temporal, geospatial, topical, network

Define and distinguish between these visualization types:
-- charts, tables, graphs, geospatial maps, network graphs

Be familiar with the elements of visually encoded data, including:
-- base map, overlaid data, visually encoded data

Classify and use these and other graphical variable types:
-- position, form, color, texture, optics

Define and use the four data scale types:
-- categorical, ordinal, interval, ratio

2: “When”: Temporal Data

Understand time series and be able to distinguish between the two binary types:
-- discrete, continuous

Know the key aspects of time slicing:
-- resolution, type (disjoint, overlapping, and cumulative), calendar alignment

Understand the various types of trends over time, including:
-- increasing, decreasing, stable, cyclic

Understand the meaning and use of different types of temporal analysis:
-- trends, patterns (i.e. seasonality), correlations, bursts

Know the meaning and use of burst detection analysis

3: “Where”: Geospatial Data

Identify and use the various types of geographic maps, including:
-- reference, topographic, and thematic

Know the three categories of thematic maps:
-- physio-geographical, socio-economic, technical

Be familiar with terms of cartography and geography:
-- geocode, geographic coordinates set, geodesic, great circle, gazetteers

Distinguish between common formats:
-- vector, raster, tabular

Be able to distinguish and classify between these data types:
-- qualitative, quantitative

Understand the definition, characteristics, and use of the three color properties:
-- value, hue, saturation
Define, distinguish between, and identify the use of four common color schemes:
   -- binary, diverging, sequential, qualitative

4: “What”: Topical Data

Be familiar with the terminology of topical analysis, including:
   -- text, text-corpus, topics, n-gram, synonymy, polysemy
Classify and understand the use of steps in data preprocessing:
   -- lowercase, tokenize, stem, remove stop words
Know the role played by dimensionality reduction in topical analysis
Understand the role and implementation of word co-occurrence analysis

Visualization Types
Understand and be able to identify the visualization types presented so-far, including:
   -- Cartogram, choropleth map, isopleth (heat) map, flow map, proportional symbol map, space-time cube maps, cross map, stacked graph, line graph, temporal bar chart, histogram, scatter plot, word cloud

Unit 5 – “With Whom”: Trees

Theory

Welcome by Katy Börner (:42)
VIDEO: http://ivmooc2014.appspot.com/unit?unit=5&lesson=1

My Note: Watched Video

Exemplary Visualizations (6:51)
SLIDES: http://ivmooc.cns.iu.edu/slides/05-exemplary.pdf
Treemap View of 2004 Usenet Returnees - Marc Smith & Danyel Fisher - 2005
My Note: Now see use of Treemaps
Examining the Evolution & Distribution of Patent Classifications - Daniel O. Kutz, Katy Börner & Elisha Hardy - 2004
My Note: Now see multiple visualizations with data story like I do.

Overview and Terminology (5:46)

http://semanticommunity.info/Data_Science/Data_Science_for_VIVO
Updated: Fri, 18 Sep 2015 23:02:30 GMT
Powered by mindtouch™
Sample Trees
Hierarchies
- File systems and web sites
- Organizational charts
- Categorical classifications
- Similarity and clustering

Branching processes
- Genealogy and lineages
- Phylogenetic trees

Decision processes
- Indices or search trees
- Decision trees
- Tournaments

Source & samples: [http://www.graphics.stanford.edu/~ha.../todrawatree/](http://www.graphics.stanford.edu/~ha.../todrawatree/)

Terminology

Tree Types
Rooted tree: Has designated root node.
Unrooted tree: No designated root node.
Binary tree: Each node has at most two child nodes.
Balanced tree: Rooted tree whose subtrees differ in height by no more than one and the subtrees are balanced, too.
Sorted tree: Children of each node have a designated order (not necessarily based on their value) and can be referred to specifically.
Node Properties
In-degree of a node is the number of edges arriving at that node.
Out-degree of a node is the number of edges leaving that node.
The root is the only node in the tree with In-degree = 0.
All the leaf nodes have Out-degree = 0.

Depth of a node is the length of the path from the root to the node. Root node is at depth zero.

Each node can have additional properties—e.g., in a family tree, each person has a name, age, gender, hair/eye color, etc.

Tree Properties
Size: Number of nodes.
Height (or depth of tree): Length of the path from the root to the deepest node in the tree.
Example:
Binary tree of size 9 and depth 3. Unbalanced and not sorted.

Workflow Design (9:09)
VIDEO: http://ivmooc2014.appspot.com/unit?unit=5&lesson=4
SLIDES: http://ivmooc.cns.iu.edu/slides/05-workflow.pdf (PDF)

Read Data
Sample Data:
• Stanford Large Network Dataset Collection, http://snap.stanford.edu/data/
• Tore Opsahl’s Datasets, http://toreopsahl.com/datasets/
• Sci2 Datasets, http://sci2.wiki.cns.iu.edu/display/...AL/2.5+Sample+

Data Formats
Network Formats
• GraphML (*.xml or *.graphml)
• XGMML (*.xml)
• Pajek .NET (*.net)
• NWB (*.nwb)

Other Formats
• Pajek Matrix (*.mat)
• TreeML (*.xml)
• Edgelist (*.edge)
• CSV (*.csv)

Tree Analysis
Extract relevant subtrees
Calculate node and edge properties—e.g., in- and out-degrees
Calculate tree properties
Sort tree
Compare trees

Visualization Goals
Representing hierarchical data
• Structural information
• Content information

Objectives
• Efficient space utilization
• Comprehension
• Interactivity
• Esthetics

Visualization Types
• Tree view
• Tree map
• Radial tree

Relevant Tools
• GUESS
• Gephi
• Cytoscape
30+ more are at http://sci2.wiki.cns.iu.edu/8.2+Netw...nd+Other+Tools

Algorithm Comparison (9:44)


SLIDES: http://ivmooc.cns.iu.edu/slides/05-algorithms.pdf (PDF)

Algorithm Comparison
• Radial Tree Layout
• Treemap Layout

Radial Tree Layout
• All nodes lie in concentric circles that are focused in the center of the screen.
• Nodes are evenly distributed.
• Branches of the tree do not overlap.

My Note: I prefer Treemaps because I think they are easy to create and understand

Treemap Layout

My Note: I like this simple diagram!

Strengths
• Utilizes 100% of display space.
• Shows nesting of hierarchical levels.
• Represents node attributes (e.g., size and age) by area size and color.
• Scalable to data sets of a million items.

Weaknesses
• Size comparison is difficult.
• Labeling is a problem.
• Cluttered display.
• Difficult to discern boundaries.
• Shows only leaf content information.

My Note: I like the simplicity of this section, but think more information is needed about graph databases and I rely on the Treemap section of my Spotfire 6 for Data Science Users Guide

Self-Assessment

LINK: http://ivmooc2014.appspot.com/activi...nit=5&lesson=5

My Note: This is a test

1. What tree visualization(s) would be most effective to visualize space consumed by files in a file system?
   
   Tree view
   Radial tree
   Tree map
   Check Answer

2. What tree visualization(s) would be most effective to visualize decision trees?
   
   Tree view
   Radial tree
   Tree map
   Check Answer

3. What tree visualization(s) would not be effective to visualize a family genealogy?
   
   Tree view
   Radial tree
   Tree map
   Check Answer

Given the tree shown here,
Tree with nine nodes
Identify node attributes:

4. In-degree of node D

**In-degree of node D is 1**
Check Answer  Show Answer

5. Out-degree of node D

**Out-degree of node D is 2**
Check Answer  Show Answer

6. Depth of node D

**The depth of node D is 2**
Check Answer  Show Answer

Plus identify major attributes for this tree:

7. Tree size

**The size of this tree is 9**
Check Answer  Show Answer

8. Tree height

**The height of this tree is 3**
Check Answer  Show Answer

9. Is this tree rooted?
Yes
No
Check Answer

10. Is this tree balanced?
Yes
No
Check Answer

11. Is this a binary tree?
Yes
No
Check Answer

12. Is this tree sorted?
Yes
No

Hands-on

Introduction by Ted Polley (1:04)

Visualizing Directory Structures (6:07)
WIKI: http://wiki.cns.iu.edu/pages/viewpag...agId=12322311

Weekly Tip: Create your own TreeML (XML) Files (9:19)
VIDEO: http://ivmooc2014.appspot.com/unit?unit=5&lesson=8
WIKI: http://wiki.cns.iu.edu/pages/viewpag...agId=12322311

My Note: We have one from Semantic Medline (see above) (XML)

Homework
LINK: http://ivmooc2014.appspot.com/activi...nit=5&lesson=8
For the first homework assignment you should go to the “My Profile” link in the navigation menu of the course site and set up a profile. You will be prompted to login again with the Google ID you used to sign up for this course. From there you will need to fill out your profile with as much information as you can. The more you share about yourself, the easier it will be to connect with your fellow students in this course to form teams for the client work. Start to form groups of 4-5 students to begin the client work.

Make sure to share your Twitter and Flickr ID’s in your profile so you can share the visualizations you create and the insight you gain from this course. If you do not already have a Twitter and Flickr account, go ahead and set one up for this course.

**My Note:** We have already formed the [Federal Big Data Working Group](http://semanticommunity.info/Data_Science/Data_Science_for_VIVO) with Data Science Teams presenting work on Data Science Problems for Business and Science

**My Note:** I got a different result when I went back to this

For this week you should visualize a directory on your computer that you don’t mind sharing an image of. You can also create a directory if you want. Visualize the hierarchical structure of this directory using one of the methods covered this week in the hands-on videos. Report any insights gained and share the image via Twitter using #ivmooc.

---

**Unit 6 – “With Whom”: Networks**

**Email**

We’ve very much enjoyed seeing your visualizations and comments over the last month and a half - keep up the great work!

Many of you had questions about networks from the Weeks 4 & 5; hopefully this week will answer your questions. Week 6 material is now online, please go and watch the videos and do the homework. Also, behind the scenes, we’re preparing a list of clients with data whom you will be working with in the coming weeks; we look forward to seeing the projects that will come out of it.

**My Question: With What: Networks of Real Network Data or of Calculated Network Data**
Theory

Welcome by Katy Börner (2:01)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=6&lesson=1

My Note: Watched Video

Exemplary Visualizations (7:58)


SLIDES: http://ivmooc.cns.iu.edu/slides/06-exemplary.pdf (PDF)

Reference:

The Human Connectome - Eugen Ludwig, Josef Klingler, Patric Hagmann & Olaf Sporns – 2008

Science-Related Wikipedian Activity - Bruce W. Herr II, Todd Holloway, Katy Börner, Elisha F. Hardy & Kevin Boyack - 2007


My Note: These are all 2007 except for one 2011. Is the data to reproduce these available?

Overview and Terminology (15:54)


SLIDES: http://ivmooc.cns.iu.edu/slides/06-terminology.pdf (PDF)

Network Analysis Examples and Goals
• Natural Networks: Neuronal, cell signaling, food webs
• Social Networks: Friendship, business, communication, collaboration networks
• Technological: Water networks, power grid, Internet, WWW

Importance of network thinking:
• Food webs might completely disassemble if just one species goes extinct.
• Weak ties and brokers are extremely important in professional networks.
Relevant Research Disciplines
The study of networks has a long tradition in
• graph theory and discrete mathematics
• sociology
• communication research
• bibliometrics/scientometrics
• webometrics/cybermetrics
• biology
• physics

Today, it is conducted in mathematics, statistics, physics, social network analysis, economics, information science, and computer science, etc.

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Mathematics / Physics</th>
<th>Statistics/Social Network Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminology</td>
<td>Adjacency matrix</td>
<td>Sociomatrix</td>
</tr>
<tr>
<td></td>
<td>Average shortest path length or Diameter</td>
<td>Characteristic path length</td>
</tr>
<tr>
<td></td>
<td>Clustering coefficient</td>
<td>Fraction of transitive triples</td>
</tr>
</tbody>
</table>

Representations of Network Data

Representations of Network Data cont.

Matrix representation and visual layout

Undirected Graphs
- Fully connected
- Partially Connected

Directed Graphs
- Partially Connected

http://semanticommunity.info/Data_Science/Data_Science_for_VIVO
Updated: Fri, 18 Sep 2015 23:02:30 GMT
Powered by mindtouch™
Terminology
Network (or graph) is composed of nodes (or vertices) and links (or edges).

Nodes can be
- **Isolated**: Unconnected to other nodes.
- **Labeled**: Have labels attributes—e.g., weights.

Edges can be
- **Undirected** (symmetric) or **directed** (nonsymmetric).
- **Labeled**: Have labels attributes—e.g., weights.
- **Signed**: Be positive and negative (friend/foe, trust/distrust).

Networks can be
- **Labeled**: Network contains labels (weights, attributes) on nodes and/or edges.
- **Temporal**: For each node/edge we know the time when it appeared in the network.
- **Undirected**: Relations between pairs of nodes are symmetric or **directed** (also called digraph) with directed links.
- **Multigraph**: Network has multiple edges between a pair of nodes.
- **Bipartite**: Nodes can be divided into two disjoint sets U, V such that every link connects a node in U to one in V.
- **Signed**: Edges can be positive and negative (friend/foe, trust/distrust).

Network Measurements
- Node and Link Properties
- Network Properties
- Statistical Properties
- Network Types

Reference:

Node and Link Properties
- **Isolated node**: Not connected to any other node.
- **Degree of a node**: Number of links connected to it.
- **Betweenness centrality of a node**: Number of shortest paths between pairs of nodes that pass through a given node.
- **Betweenness centrality of a link**: Number of shortest paths among all possible node pairs that pass through a given link.
- **Shortest path length**: Lowest number of links to be traversed to get from nodes i to j.

Network Properties
**Number of**
- Nodes
- Isolated nodes
- Edges
- Self-loops

**Diameter**: Longest of all shortest paths among all possible node pairs in a network—i.e., #links to be traversed to
interconnect the most distant node pairs.

**Density**: Ratio of the number of edges in the network to the square of the total number of nodes.

**Number of**

- **Strongly connected components**: There is a directed path from each node in the network to every other node.

- **Weakly connected components**: Maximal subgraph in which all pairs of vertices are reachable from one another—disregards link directions.

- **Clustering coefficient**: Measures the average probability that two neighbors of the node i are also connected.

**Statistical Properties**

- **Node degree distribution** $P(k)$ of an undirected graph is defined as the probability that any randomly chosen node has degree $k$.

  Power law distribution $P(k)$ for a scale free network

**Network Properties and Network Types**

- **Average clustering coefficient** measures the average probability that two neighbors of the node i are also connected.

- **Average path length**: Average number of steps along the shortest paths for all possible pairs of network nodes.

**Network Types**

- **Small World**: Road maps, food chains, electric power grids, networks of brain neurons, telephone call graphs, and social influence networks.

- **Scale-free**: World-Wide Web, the Internet, social networks, airline.

**Workflow Design (12:49)**


**Read Data**

**Sample Data:**

- UCINet datasets, [https://sites.google.com/site/ucinetsoftware/datasets](https://sites.google.com/site/ucinetsoftware/datasets)
- Pajek Datasets, [http://pajek.imfm.si/doku.php?id=data:index](http://pajek.imfm.si/doku.php?id=data:index)
- CASOS Datasets, [http://www.casos.cs.cmu.edu/computat...tools/datasets](http://www.casos.cs.cmu.edu/computat...tools/datasets)
Major Data Formats:

Input:
Network Formats
- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net)
- NWB (*.nwb)
Scientometric Formats
- ISI (*.isi)
- Bibtex (*.bib)
- Endnote Export Format (*.enw)
- Scopus csv (*.scopus)
- NSF csv (*.nsf)
Other Formats
- Pajek Matrix (*.mat)
- TreeML (*.xml)
- Edgelist (*.edge)
- CSV (*.csv)

Output:
Network File Formats
- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)
Image Formats
- JPEG (*.jpg)
- PDF (*.pdf)
- PostScript (*.ps)

Preprocess Data
- Unification, see Unit 4
- Network extraction
- Delete isolate nodes
- Remove self-loops
- Threshold—e.g., extract nodes/edges above or below value
- Merge two networks
- Pathfinder network scaling, see Backbone Identification
Preprocessing—Network Extraction
Weighted, undirected co-occurrence network.

Paper Authors References Year

Unweighted, directed bimodal network.

Paper Authors References Year

Unweighted, directed network of two types.
Calculate node degrees.

Paper Authors References Year

Unweighted, directed paper-citation network.
Arcs go from papers to references.

Paper Authors References Year

Unweighted, directed bipartite network.

Paper Authors References Year

WRONG!!!

Network Analysis
Calculate
• Node and link properties
• Network properties
• Statistical properties
• Network types

• Extract relevant subtrees

• Calculate error and attack tolerance
• Compute clusters and backbones, see later parts in this unit

Visualization Goals
Representing hierarchical data
• Structural information
• Content information

Objectives
• Efficient space utilization
• Comprehension
• Interactivity
• Aesthetics

**Example: Collaboration Network**
• Random layout
• Circular layout
• Generalized Expectation-Maximization (GEM) layout

First, determine node layout
Then add nodes and linkages

For GEM layout, color- and size-code nodes and links and add a legend.

**Aesthetic Criteria for Graph Drawing**
• Maximize symmetry
• Evenly distributed nodes
• Uniform edge lengths
• Minimized edge crossings
• Orthogonal drawings
• Minimize area, bends, slopes, angles
• Maximize consistent flow direction (in directed networks)

Optimization criteria may be relaxed to speed up layout process.

**Visualizing Large Networks**
Discover landmark nodes based on
• Existing node attributes—e.g., frequency of access
• Connectivity (hubs & authorities)
• Depth in a hierarchy

strong (and weak) links

Identify backbone, see Backbone Identification

Show clusters, see Clustering

**Interacting With Networks**
Modify focusing parameters while continuously providing visual feedback and updating display (fast computer response).
• Conditioning: filter, set background variables and display foreground parameters
• Identification: highlight, color, shape code
• Parameter control: line thickness, length, color legend, time slider, and animation control
• Navigation: bird's-eye view, zoom, and pan
• Information requests: mouse over or click on a node to retrieve more details or collapse/expand a subnetwork
Clustering (7:59)


SLIDES: http://ivmooc.cns.iu.edu/slides/06-clustering.pdf (PDF)

Goal

Clustering of graphs (also called community detection) aims to identify modules using the information encoded in the graph topology—i.e., geometric position and spatial relations. Results might be visualized using graphic variable types (see Unit 1) or additional cluster boundaries.

Different Clustering Approaches

- Divisive algorithms: detect inter-community nodes or links and remove them from the network—e.g., using betweenness centrality thresholding.
- Agglomerative algorithms: merge similar nodes/communities recursively—e.g., Blondel community detection.
- Optimization methods: maximization of an objective function.


See examples on subsequent slides.

Calculate node betweenness centrality (BC) and delete high BC nodes

Collaboration network of Vespignani and Barabási

Nodes are sized and color coded by betweenness centrality
Betweenness Centrality (BC) measures a node's centrality, load, or importance in a network. It equals the number of shortest paths from all nodes to all others that pass through that node. That is, nodes with higher 'betweenness' occur on more paths between other nodes.

BC of a node \( n \) in a network graph \( G:=(N, L) \) with \( N \) nodes is computed as follows:

- For each pair of nodes \( (s, t) \), compute the shortest paths between them.
- For each pair of nodes \( (s, t) \), determine the fraction of shortest paths that pass through the node in question (here, node \( n \)).
- Sum this fraction over all pairs of nodes \( (s, t) \).

In the network to the right, node BC is represented by hue (from red=0 to blue=max).
See also http://en.wikipedia.org/wiki/Centrality

**Agglomerative clustering using Blondel Community Detection**
Algorithm reads a network and calculates additional attributes for each node (up to three community levels). Links are not modified.


**Blondel Community Detection**
- Aims to partition a network into communities of densely connected nodes, with the nodes belonging to different communities being only sparsely connected.
- The quality of communities within a cluster partition is measured by the modularity of the partition. Modularity is a scalar value between -1 and 1 that measures the density of links within communities as compared to links between communities.
- Modularity can be used as an objective function to arrive at the best communities but also to compare different methods.

**Blondel Community Detection Initialization**
Algorithm starts with a weighted network of N nodes.
Each node is assigned to a different community—i.e., \#nodes = \#communities.

**Phase I**
For each node i we consider the neighbors j of i and calculate the gain of modularity that would take place by removing i from its community and by placing it in the community of j. The node i is then placed in the community for which this gain is maximum (in case of a tie we use a breaking rule), but only if this gain is positive. If no positive gain is possible, i stays in its original community. This process is applied repeatedly and sequentially for all nodes until a local maxima of the modularity is attained—i.e., when no individual move can improve the modularity.

**Phase II**
Found communities are aggregated in order to build a new network of communities. Here, weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities. Links between nodes of the same community lead to self-loops for this community in the new network.

The passes are repeated iteratively until no increase of modularity is possible.

Backbone Identification (7:15)


**SLIDES:** [http://ivmooc.cns.iu.edu/slides/06-backbone.pdf](http://ivmooc.cns.iu.edu/slides/06-backbone.pdf) (PDF)

**Goal**
- Identify and visualize the main structure of a network.
- Apply if the network is too dense (looks like a spaghetti ball).
- Link weights are used to identify important connections.
- Careful with directed links.
Terminology
Backbone: The chief support of a system; the mainstay.
In networks: The part of a network that handles the major traffic and/or has the highest-speed transmission paths.

Example: Road Network

All Streets map visualization of all the 240 million individual road segments in the U.S.

Example: Road Network Backbone
Wikipedia features the U.S. Interstate Highways within the 48 contiguous states.

**Different Backbone Identification Approaches**

- Use existing node/edge properties to identify and delete superfluous links.
- Keep top-n highest weight edges per node (used in DrL).
- Calculate **Pathfinder Network Scaling** network.

See examples on subsequent slides.

**Use existing node/edge properties**

Here, number of times co-authored

---

**Keep top-n highest weight edges per node**

DrL is a force-directed, highly scalable graph layout developed by Shawn Martin and colleagues at Sandia National Laboratories.

Original NW

Nodes: 247, Edges: 795

Reduced NW using top-5

Nodes: 247, Edges: 579
Calculate Pathfinder Network Scaling Network
**Pathfinder Network Scaling**

Takes a similarity or distance matrix as input and extracts a network that preserves only the most important links. Relies on the triangle inequality to eliminate redundant or counterintuitive links. That is, given two nodes connected by multiple paths, only the path is preserved that has a greater weight defined via the Minkowski metric.

Two parameters influence the result:

- **q**: defines the length of a path examined up to which the triangle inequality must be maintained. A network of N nodes can have a maximum path length of q=N-1. With q=N-1 the triangle inequality is maintained throughout the entire network.
- **r**: defines the metric used for computing the distance of paths—the Minkowski distance. It is a real number between 1 and infinity, inclusive.

The higher r or q, the fewer links in the respective PFnet(q, r). The PFnet(n − 1, ∞) has the minimum number of links.

Error and Attack Tolerance (3:07)


Reference:

Generate Three Networks (see Network Types)
- Random Networks using the model by Erdös and Rényi, 1957.
- Small World Networks using the model by Watts and Strogatz, 1998.
- Scale-Free Networks using the model by Barabasi and Albert, 2000.

Error Tolerance: Delete 100 random nodes
Attack Tolerance: Delete top-100 highest degree nodes

Example: Topological resilience to targeted attacks
Two networks are studied:
• Scale-free Internet Router level network
• Erdős and Rényi random network

Scale-free network is more fragile. Removal density as low as $g=0.05$ suffices to fragment entire network.
Exhibit Map with Andrea Scharnhorst (Optional) (6:22)
VIDEO: http://ivmooc2014.appspot.com/unit?unit=6&lesson=8

Self-Assessment
My Note: Nothing

Hands-on

Introduction by Ted Polley (1:04)
VIDEO:

Co-Occurrence Networks: NSF Co-Investigators (12:18)
WIKI: http://wiki.cns.iu.edu/pages/viewpag...pageld=2785284

Directed Networks: Paper-Citation Network (8:01)
WIKI: http://wiki.cns.iu.edu/pages/viewpag...pageld=2200066

Bipartite Networks: Mapping CTSA Centers (10:27)
WIKI: http://wiki.cns.iu.edu/pages/viewpag...pageld=2785293

Weekly Tip: How to use Property Files (7:17)
WIKI: http://wiki.cns.iu.edu/display/SCI2T...roperty+Files1

Homework
My Note: Nothing
Unit 7 – Dynamic Visualizations & Deployment

Dear IVMOO students,

It has been great watching your visualizations and watching you learn for the main instructional section of the course, but our last week of lessons, on dynamic visualizations, is here, as is the final exam. The final must be taken by March 17 at 5pm EST if you want to get credit for the course.

Over the next few weeks, though we will still be posting some lessons, you will primarily be forming teams to work with clients for the remaining 7 weeks. We will post instructions on how to do so next week.

Best of luck on the final - make sure you are using a computer which can use the Sci2 Tool. Some of you have had trouble with the tool because of international settings (in many countries, decimal numbers like 3.05 are represented as 3,05) - if this is causing issues, changing the country or region settings of your computer, temporarily, to the U.S., should fix this issue. If other issues arise, please email us and we will get back to you as soon as possible.

Make sure you give yourself a few hours to take the test, as there are quite a few questions, and it may take some time to finish. As with the midterm, keep in mind that your eventual score will probably increase once we have applied the curve and corrected any fill-in-the-blank questions that may have been answered correctly but marked wrong.

Best of luck, and keep in mind, the final covers the entire course.

Until next week,
Scott, Katy, and the IVMOO team
Welcome by Katy Börner (1:57)

VIDEO: http://ivmooc2014.appspot.com/unit?unit=7&lesson=1

My Note: Watched Video

Exemplary Visualizations (20:50)


SLIDES: http://ivmooc.cns.iu.edu/slides/07-exemplary.pdf (PDF)

Interactive Visualizations
Desktop
• Tableau
• Sci2/GUESS
• Gephi
• Cytoscape
Large Touch Displays
• Illuminated Diagram display
• NASCast
Online (also hand-held mobile)
• Sci2 Web services
• MapSustain & Gene Therapy
• NIH Topic Map
• VIVO National Researcher Network

Europe Raw Cotton Imports in 1858, 1864 and 1865 - Charles Joseph Minard - 1866

Mapping the Evolution of Co-Authorship Networks
Ke, Visvanath & Börner. 2004. Won 1st prize at the IEEE InfoVis Contest.

Illuminated Diagram Display on display at the Smithsonian in DC.
http://scimaps.org/exhibit_info/#ID

Sci2 Tool Web Services
As visual interface to publically available NIH RePORT Expenditure and Results RePORTER)/ RePORTER data provided by NIH.

My Note: Look at this

http://mapsustain.cns.iu.edu
http://kongch.cns.iu.edu/genetherapy/geomap.html
Geospatial Analysis (Where) A geospatial map of the US is used to show where what science is performed by whom.

Topical Analysis (What) Science map overlays show where a person, department, or university publishes most in the world of science.

Other Examples
• Max Planck Research Networks, http://max-planck-research-networks.net/
• NOAA’s Science on a Sphere, http://www.sos.noaa.gov
• Disney, http://disney.go.com/disneyinteractivestudios
• UCSB’s AlloSphere, http://www.allosphere.ucsb.edu

Dynamics (7:56)


SLIDES: http://ivmooc.cns.iu.edu/slides/07-interactive.pdf (PDF)

Dynamics
• Different Types of Dynamics
• Time-Slicing Data, see also Hands-on
• Visualization Formats
• (Non-Sequential) Story Telling

Different Types of Dynamics
Data attributes change:
Use, e.g., temporal graphs (also called chronological graphs) to show changing properties or derivative statistics, see Unit 2.

Data and data attributes change:
Overlay dynamic data on static basemaps/reference systems (chart, graph, geomap, or network graph).

Data and reference system change:
Use dynamic basemap with dynamic data overlays—e.g., world map with changing political boundaries and annual migration trajectories or evolving collaboration networks.

Time Slicing Data, see Unit 2

Resolution: Milliseconds, seconds, minutes, hours, days, weeks, fortnights (fourteen days/two weeks), months, quarters, years, decades, and centuries.

Type:
• Disjoint: Every row in the original table is in exactly one time slice.
• Overlapping: Selected rows are in multiple time slices.
• Cumulative: Every row in a time slice is in all later time slices.

Alignment with calendar: If first event is June 7th, 2006, and yearly slices are chosen, then the first slice will be from
• No: June 7th, 2006, to June 6th, 2007

Time Slicing Data—Issues

Outliers:
Identify and deal with outliers: e.g., web page gets Slashdotted—a popular website links to a smaller site causing a massive increase in traffic analogous to a denial-of-service attack.

Seasonality:
Many datasets show the impact of day/night, winter/summer, and other cycles.

Select best frame length:
• Too short: Few data records are visible—e.g., networks might have many isolated nodes.
• Too long: Too many data records are visible—e.g., network is a spaghetti ball.

Visualization Formats
• One static image
• Multiple static images
• Animations which can be started, stopped, fast-forwarded, or rewound interactively.
• Interactive services that support "overview, filter, and details on demand" functionality.

Tell Non-Sequential Stories
“Slides serve up small chunks of promptly vanishing information in a restless one-way sequence.”
Beautiful Evidence, Edward Tufte, Graphics Press, 2006, p. 160. My Note: This is why I put them in graphic two-way sequence
“Overview first, zoom and filter, details on demand.” My Note: This is why I use Spotfire
The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations, Ben Shneiderman, 1996.

Print solutions:
• Use panels to sequence narratives from left to right, top to bottom.
• Number sections of the display sequentially.
• Suggest visual pathways, e.g., by using arrows.

Tell stories that are
• Simple as possible but not simpler
• Seamless in their integration of words and images
• Sequential, as a narrative
• Informative
• True
• Contextual (past, present, future)
• Familiar (know your audience)
• Concrete
• Personal
• Emotional
• Actionable
See Hans Rosling’s Gapminder for an excellent example.
**Hans Rosling’s Gapminder (4:27)**


Wealth & Health of Nations

Gapminder World Map – Mattias Lindgren – 2010

Watch [http://www.youtube.com/watch?v=jbkSRLYSojo](http://www.youtube.com/watch?v=jbkSRLYSojo)

Dissecting the Gapminder Graph
- 2-axis reference system (x-axis money, y-axis health)
- Country money and health: Location
- Country population: Circle size (quantitative)
- Country continent: Circle color hue (qualitative)
- Country name: Label

Country location and circle size (but not color) change over time.

**Deployment (21:07)**


**Deployment**
- Image Resolution
- Display Ratio
- Data Formats
- Image Generation/Capture Options
- Image Display Options
- The Ultimate Display

**Image Resolution**

Pixel, or picture element, is the smallest area that a device can read or write.
Voxel, or volume element, is the smallest volume that a device can read or write.
Dots per inch (DPI) is the number of dots/pixels in a one-inch line.
Resolution is measured in DPI. Common values are
- 72 DPI for the Web
- 300 DPI for prints
- 1200 DPI for scanners

The higher the DPI, the higher the resolution and the file size.
Mega Pixels (MP) equal 1,000,000 pixels.
Display Ratio
The quantitative relation between the width and height of a display.

Letter (8.25” x 11”) 4 : 3
Movies 4 : 3 or 16 : 9
TV 4 : 3 or 16 : 9
Most displays 4 : 3 or 16 : 9
Laptop screens 4 : 3 or 16 : 9
YouTube 16 : 9

Examples
Camera
2981 x 1677 = 5MP, 16:9 ratio

Letter size paper + 300 DPI laser printer
8.5 x 11 inch at 300 DPI, 8.5 MP

iPad3
2048 x 1536 pixels on 5.8 x 7.76 inch, 264 DPI

iPhone4S
960 x 640 pixels, 326 DPI

TV Screens
• VGA: 640 x 480, 4:3 ratio
• PAL/SECAM: 768 x 576, 4:3 ratio
• HD 720: 1280 x 720, 16:9 ratio

Example: 5MP Camera Photo
Aspect ratio: 16:9
Resolution: 2981 x 1677=5,000,000 pixels or 5 MP

File size
Bits Color scale Size
8-bit grayscale 4.77 MB
8-bit RGB color 14.3 MB
16-bit RGB color 28.6 MB - more common
32-bit RGB color 57.2 MB
Down sampled to 72 DPI Web resolution reduces the file size by a factor of four.

Can be printed in
300 DPI at a size of 9.9 x 5.6 inches (25.2 x 14.2 cm)

Can be displayed on HD Screen in full resolution:
2.5 HD screens (each with a resolution of 1920 x 1080 = 2,073,600)

Data Formats
Vector
Stored as geometric description that can be rendered at any size.

- Postscript .ps
- Scalable Vector Graphics .svg
- MS Power Point .ppt

**Raster**

Stored as grid of pixels.

- .jpg
- .tiff
- .gif
- .bmp
- .png

**Image Generation/Capture Options**

**Image generation**

- Render into file: Size and resolution only restricted by file/disk size

**Image capture**

- Screen capture: 72 DPI, pixel size and ratio depend on screen
- Camera
- Scanner

**Super-high-resolution images:** Combine multiple images—e.g., Photopic Sky Survey is a **5,000 MP** photograph of the entire night stitched together from 37,440 exposures. Requires 1000 times more space to print or display than a 5 MP image.

[http://skysurvey.org](http://skysurvey.org)

**Image Display Options**

**Static**

- 2D printout—e.g., on paper
- 3D printouts

**Interactive**

- 2D digital displays: Hand-held devices, desktop and laptops, large displays
- 3D digital displays—e.g., CAVE

**Super-high-resolution displays:** Combine multiple displays

**Combination**

- Illuminated Diagram display: printout with projected data overlays
  See examples on subsequent slides.

**2D Printouts—e.g., on Paper**

- Are cheap—no computer hardware/software/expertise costs
- Offer high resolution—a map the size of a 4 x 6 foot (1.2 x 1.8m) dining table in 300 DPI print quality can display more than 310 MP
- Fast—no boot up time
- Easy to transport and deploy—no outlet needed
• Can be easily explored and annotated (e.g., using a pen) by a single viewer or by a team
• Durable—archival paper prints stored in a dry, dark room are likely to be readable in 500 years

3D Printout
• Can be created manually or using computers
• Using plastics, resins, or metals
• Different resolutions
• Single or multi-color
From http://norikoambe.com

2D Digital Displays
• Computer, laptop, tablet, and phone displays come in different sizes, resolutions, interactivity, and prices. In 2012, high resolution displays might reach 10 MP.
• Super-High-Resolution Displays compile multiple displays into a tiled display wall, the walls of a room (CAVE), on the surface of a globe, etc.
From http://pti.iu.edu/avl

Super High Resolution Displays
• Davos Studio Room uses 5 x 16 = 80 modules, each with 128 x 128 pixels—i.e., 2048 x 640 = 1.3 MP on 7.90 x 2.56m wall, http://www.youtube.com/watch?v=7MUaR24tYJ8
• IU’s IQ-Wall uses 12 highresolution monitors with a total of 12.5 MP

Ingo Gunther’s WorldProcessor globe design now shown on the Giant Geo Cosmos OLED display at the Museum of Emerging Science and Innovation in Tokyo, Japan.

Combination: Illuminated Diagram display

Science maps in “Expedition Zukunft” science train (12 coaches, 300 m long) visiting 62 cities in 7 months. Opening was on April 23rd, 2009, and attended by German Chancellor Merkel. http://www.expedition-zukunft.de

The Ultimate Display
Would effectively match human visual perception:
• Resolution equals visual acuity of the human eye
• Wide viewing angle that also stimulates peripheral vision
• High brightness and color brilliance
• High update rate
• Supports stereoscopy
But differs from Sutherland’s vision in which a computer “can control the existence of matter.”

Relevant Software
• http://Zoom.it (formerly Seadragon), see Unit 7: Hands-on
• http://gigapan.org
• http://www.openzoom.org open source toolkit for the Adobe Flash Platform support the sharing interactive exploration (zoom and pan) of large images.

**Color Perception and Reproduction (8:11)**


SLIDES: http://ivmooc.cns.iu.edu/slides/07-color-production.pdf (PDF)

**Terminology**

*Color Space* is a mathematical model for describing color.

Examples:
- RGB (red, green, blue) - additive
- CMYK (cyan, magenta, yellow, black) - subtractive
- HSB (hue, saturation, brightness)
- HSV (hue, saturation, value)
- HLS (hue, lightness, saturation)

Some models are additive, others are subtractive.

**Example of Additive RGB Color Mixing Space**

Emitting surfaces (e.g., computer screens) use three primary colors: red, green, and blue. Only colors that fall within the triangle defined by these three colors can be produced.

**Color Perception**

**Example of Subtractive Mixing**

**Example of Additive Mixing**

My Note: See PDF
Color Reproduction

• Reproducing the very same colors across multiple hardware platforms (printers, displays) is critical if color encodes data values.
• Non-emitting surfaces such as paper absorb and reflect the light that hits their surface.
• The type of paper used (coated, uncoated, and matte stock), will affect the appearance of colors, as will paper color.
• Standard matching systems—e.g., the PANTONE MATCHING SYSTEM that ships with Adobe and many other products as well as most printers—are used to ensure color consistency.

http://www.pantone.com

The Making of AcademyScope (Optional) (8:05)


Self-Assessment

LINK: http://ivmooc2014.appspot.com/activi...nit=7&lesson=7

What is the resolution of a standard

1. Scanner?
   - 72 dpi
   - 300 dpi
   - **600 dpi**
   - Check Answer

2. Printer?
   - 72 dpi
   - **300 dpi**
   - 600 dpi
   - Check Answer

3. Web Page?
   - **72 dpi**
   - 300 dpi
   - 600 dpi
   - Check Answer

4. High-definition television (HDTV) provides about how many times more pixels than standard-definition television (SD)?
   - 3
   - **5**
5. The ‘Ultimate Display’ would not have the following feature:

- Resolution that equals the visual acuity of the human eye
- Wide viewing angle that also stimulates peripheral vision
- High brightness and color brilliance
- High update rate
- **Smell**

Check Answer

6. What technique would help users examine local details without losing the global structure?

- Focus & context
- Brushing & linking
- Local & global

Check Answer

7. What percentage of users abandon a website that takes more than 3 seconds to load?

- 20%
- 40%
- 60%

Check Answer

**Hands-on**

**Introduction by Ted Polley (1:05)**


**Evolving Networks with Gephi (9:23)**


**Exporting Networks from Gephi with Zoom.it (formerly Seadragon)**


WIKI: [https://gephi.org/plugins/seadragon/](https://gephi.org/plugins/seadragon/)
**Homework**

Take the network you created for the homework in Week 6 and load it into Gephi. Then export the resulting visualization with Seadragon using the steps covered this week during the hands-on session. Take a screenshot of your visualization and post it to Twitter using #ivmooc.

---

**Final Exam**

**Instructions**

- You will only be able to take the exam once.
- You must be on a computer that can run Sci2 in order to complete this exam.
- **Honor Code:** Academic integrity requires that participants take credit only for ideas and efforts that are their own. Participants must not use unauthorized assistance, materials, information, or study aids in any assessments. Participants must not use another person as a substitute in taking assessments. Participants must not take any credit for a team project unless the participant has made a fair and substantial contribution to the group effort. Lastly, participants must not intentionally or knowingly help or attempt to help another participant to commit an act of academic misconduct.

**My Note:** I used all the above materials by searching with Google Chrome Find for the answers and their sources. In effect this page becomes "the visualization of the IV MOOC"!

**My Note:** Caution - I have no way of knowing if my answers below are correct and copying for the credit course would be a violation of the IU Student Policy.

**Framework & Workflow Design**


1. Zip codes are an example of which type of data-scale?
   - Nominal
   - Ordinal
   - Interval
   - Ratio

2. Age is an example of which type of data-scale?
   - Nominal
   - **Ordinal**
   - Interval
   - Ratio

3. Winners in a race are an example of which type of data-scale?
   - Nominal
   - **Ordinal**
   - Interval
   - Ratio
4. Celsius temperature is an example of which type of data scale?
   - Nominal
   - Ordinal
   - Interval
   - Ratio

   See: [Hans Rosling's Gapminder (4:27)](#)

5. Which of the following color schemes should be used to encode qualitative data?
   - Color value (lightness)
   - Color hue (tint)
   - Color saturation (intensity)

6. Which of the following should be used to encode quantitative data?
   - Shape
   - Size
   - Enclosure

“*When, Where, and With Whom*: Temporal, Geospatial, and Network Analysis & Visualization

Search Query

The following awards were matching the search query ‘Information Visualization’ ([Click here to see the file](#)). Using either the Sci2 tool or another program in the course, such as Microsoft Excel, analyze the data. Report immediate results and the properties of the final network. Note that NSF changed its file format recently-please load the data in ‘Standard csv format’. We have already done the preprocessing of the raw data to save you time when taking the exam. You should be able to load the file into Sci2 and perform the steps necessary to answer the questions.

See: [IVMOOCFinal-Spotfire](#)

7. Number of awards that match the search query ‘Information Visualization’:
   - Answer: 10

8. Title of award with highest 'AwardedAmountToDate':
   - Answer: DLI Phase 2: Informedia-II: Integrated Video Information Extraction and Synthesis for Adaptive Presentation and Summarization from Distributed Libraries

9. Earliest start date, in format YYYY:
   - Answer: 1990

10. Latest expiration date, in format YYYY:
    - Answer: 2012
Temporal Visualizations

See: IVMOOCFinal-Spotfire

11. Render a temporal bar graph visualization. How many unique ‘NSFOrganization’ exist?
Answer: 10

12. What is the short-term, highly funded grant after 1998?
Answer: DLI Phase 2: Informedia-II: Integrated Video Information Extraction and Synthesis for Adaptive Presentation and Summarization from Distributed Libraries

Burst Analysis

My Note: This did not seem that important to me

Lowercase, Tokenize, and Stopword the ‘Title’ words. Next, identify what bursts exist for tokens in titles using the columns ‘StartDate’ and ‘Title’. Set the ‘Density Scaling’ parameter to 3.0.

13. Token label:
Answer:

14. Starting Year:
Answer:

15. Ending Year:
Answer:

Geospatial Visualization

See: IVMOOCFinal-Spotfire

Using Aggregate Data, identify all unique geolocations and sum up the number of their total AwardedAmountToDate. Using the Generic Geocoder algorithm and the column ‘OrganizationZip’, identify latitude and longitude information for each award. Next, create a proportional symbol map that shows the sum of AwardedAmountToDate per unique geolocation.
16. Report the number of unique geolocations based on ‘OrganizationZip’:
Answer: 45

For the geolocation with the highest amount of award funding, report (remember: use Aggregate Data and then Excel or a basic text editor to view the file):

17. AwardedAmountToDate:
Answer: 4,541,968

18. Latitude:
Answer: 32.23

19. Longitude:
Answer: -110.95

20. Count (total number of awards at location):
Answer: 4

21. Where is the highest concentration of awards by total number of awards?
   East coast
   West coast

Co-Funding Network
See: [IVMOCFinal-Spotfire](#)

Some awards are funded by multiple NSF programs. Extract the co-funding (co-occurrence) network from the ‘Program(s)’ column.

22. Number of unique program nodes:
Answer: 46

23. Number of co-occurrence links:
Answer: 8
24. Number of weakly connected components (a weak component is defined as a maximal subgraph in which all pairs of vertices in the subgraph are reachable from one another in the underlying undirected subgraph.)
   Answer: 38

25. Number of nodes in largest component:
   Answer: 4

26. Identify the two programs that have the strongest co-funding relationship. List their names separated by a semicolon (;):
   Answer: HUMAN COMPUTER INTER PROGRAM; COLLABORATIVE SYSTEMS/HUMAN COMPUTER INTER PROGRAM

27. Number of times those two programs co-funded:
   Answer: 1

“My Note: This did not seem that important to me

Topical

The Sci2 Wiki table of contents (ToC) looked like this in early 2014. Use Excel or another tool to reformat it into a .csv file with two columns. Load it into the Sci2 tool, Lowercase, Tokenize, and Stopword the 'Title' words and run 'Extract Word Co-Occurrence Network'. Report immediate results and the properties of the final network.

28. Number of data records (ToC entries):
   Answer:

29. Number of unique tokens after text normalization:
   Answer:

30. Number of nodes in the co-occurrence network:
   Answer:
31. Number of edges in the co-occurrence network: Answer:

32. Size of largest connected component: Answer:

List the top 3 tokens with the highest degree.

33. Highest frequency token label: Answer:

34. Highest frequency token degree: Answer:

35. Second highest frequency token label: Answer:

36. Second highest frequency token degree: Answer:

37. Third highest frequency token label: Answer:

38. Third highest frequency token degree: Answer:

39. Weight of strongest co-occurrence link: Answer:

“With Whom”: Trees

Given the tree shown here,
See: [Self-Assessment](http://semanticommunity.info/Data_Science/Data_Science_for_VIVO)

### Identify node attributes:

40. In-degree of node B:
   **Answer:** 1

41. Out-degree of node B:
   **Answer:** 1

42. Depth of node B:
   **Answer:** 1

### Identify tree attributes:

43. Size:
   **Answer:** 13

44. Height:
   **Answer:** 3

45. Is this a rooted tree?
   - **Yes**
   - **No**
46. Is this a sorted tree?
   Yes
   No

47. Is this a balanced tree?
   Yes
   No

48. Is this a binary tree?
   Yes
   No

“With Whom”: Networks

Given the network shown here,

![Network Diagram](image)

See: Overview and Terminology (15:54)

**Identify node attributes:**

49. Degree of node E:
   **Answer:** 4

50. Degree of node G:
   **Answer:** 2

51. Label of isolated node:
   **Answer:** K

**Identify network attributes:**

52. Size:
   **Answer:** 15

53. Number of edges:
   **Answer:** 14
54. Number of components (a component is defined as a maximal subgraph in which all pairs of vertices in the subgraph are reachable from one another in the underlying undirected subgraph.):
   Answer: 3

55. Size of largest connected component:
   Answer: 9

56. Density of giant component:
   Answer: 1/9

57. Diameter of giant component:
   Answer: 5

58. Is this a directed network?
   Yes
   No

59. Is this a weighted network?
   Yes
   No

60. Is this a fully connected network?
   Yes
   No

61. Is this a signed network?
   Yes
   No

62. Is this a labeled network?
   Yes
   No

63. Is this a multigraph network?
   Yes
   No

See: Clustering (7:59)

64. Blondel community detection is a
   Divisive Algorithm
   Agglomerative Algorithm
   Optimization Method

See: Error and Attack Tolerance (3:07)
65. Which network type has the lowest tolerance to targeted attacks:
   - Lattice
   - Random
   - Small World
   - Scale-Free

**Dynamic Visualizations & Deployment**

**See:** [Deployment (21:07)]

The resolution of a camera, scanner, printer, or monitor is commonly measured in dots per inch (DPI)—the number of dots in a one-inch line. Given a two-foot wide (24 inch, about 61cm) timeline visualization printed in 300 DPI, what width in inches (rounded to nearest full integer) would this timeline have in maximum resolution on a:

66. 72 DPI projector wall: \(72 \times 24 / 300 = 5.76\)
   **Answer:** 6

67. 200 DPI printout:
   **Answer:** 16

68. DPI printout:
   **Answer:** 64

69. How many iPads Airs with retina display (each with a resolution of \(2048 \times 1536 = 3,145,728\) pixels) would it take to view a photo taken with its camera at a resolution of \(2981 \times 1677 = 5,000,000\) pixels (5 megapixels)?
   - 1
   - 1.5 \(5/3.1 = 1.59\)
   - 2.5
   - 5

**List of Clients**

Source: [http://ivmooc.cns.iu.edu/forums/clients](http://ivmooc.cns.iu.edu/forums/clients)

**Project Title:** Information Visualizations for Big Data in Drug Discovery, Health and Translational Medicine

**Client Name**

David Wild, Big Data in Drug Discovery, Health and Translational Medicine MOOC

**Project goal/scientific or practical value**

Students working on this project are given the opportunity to analyze and visualize scientific funding and publications environment, with a focus on academic and corporate research organizations that working in the area of drug-discovery and translational medicine. The goal of the project is to produce information visualizations related to Drug Discovery and Big data applications. Student will identify key concepts, key researchers and organizations, important publications, and/
or grant or patent profiles for researchers and organizations related to drug discovery and translational medicine. Students can then use the publication/patent/grant/twitter, etc. data to identify trends over time, bursts of activity, key institutions/geolocations, analyze co-author and other networks, etc. If there are programmers in your team then they might be able to generate interactive visualization, such as a map like http://mapsustain.cns.iu.edu that lets others explore the space of recent work as relevant to researchers and instructors interested in locating relevant data and resources. The practical and scientific value of this project comes in producing original research in this field, with all of the practical difficulties that come with implementing a replicable work-flow to collect and analyze a data set, and determine the appropriate visualization techniques that best answer your research questions that your visualizations will represent.

Information on dataset(s) to be used


Web-link to dataset(s)

Background materials for Big Data and Drug Discovery

Relevant publications, websites, etc.


Conditions under which students can publish results and/or add project results to their resume

There are no conditions for publication at this time.

Client Forum

Information Visualizations for Big Data in Drug Discovery, Health and Translational Medicine

Project Title: Human Genome Project Documentary History: An Annotated Scholarly Guide to the HGP

Client Name

Cold Spring Harbor Laboratory
Project goal/scientific or practical value

Cold Spring Harbor Laboratory (CSHL) is seeking support from the National Library of Medicine (NLM) to prepare an annotated scholarly guide to the international Human Genome Project (HGP). The tentative title is: Human Genome Project Documentary History: An Annotated Scholarly Guide to the HGP. The guide will be of value to biomedical researchers and historians of medicine and the life sciences, as well as bioethicists and public health officials. This idea came out of an international meeting on the history of the Human Genome Project held in May 2012 at Cold Spring Harbor Laboratory (CSHL). Participants included scientists, administrators, authors, publishers, filmmakers, historians, and funders. At that meeting, scientists gave presentations on different aspects of the HGP, and the discussion that followed centered on how to best present the history of the HGP to different audiences. The goal of the annotated scholarly guide is to provide scholars with a tool that may be used as a starting point for research on the history of the HGP. The Human Genome Project (HGP) was a 13-year research project carried out in more than 20 laboratories around the world, including labs located in the U.S., UK, Japan, Germany, France, and China. The goal of the HGP was to discover all 20,000 to 25,000 human genes, determine the sequence of the 3 billion DNA subunits contained in the human chromosomes or "genome," and make this information available for further study. The HGP has contributed to improving human health by enabling a better understanding of the molecular basis of various diseases. A study released in 2011 by Battelle shows that the HGP also had extensive economic impacts. Because it was such a large project, distributed over labs in six countries, and because it spanned the time before and after widespread use of the World Wide Web (1990-2003), the HGP presents a challenging topic for scholarly research.

Information on dataset(s) to be used

The data set consists of bibliographic records for key individuals associated with the Human Genome Project, dating from 1977-2003. The search was done in EndNote Online using the database Web of Science (TS). We searched individually by author (see attached list) for papers with genom* in the Title/Keywords/Abstract, limiting the year of publication to 1977-2003 (see attached screen shot). Of the results returned, we selected only those papers that had been cited at least once. (We do have a data set containing the papers that had not been cited at least once, if you need that.) We have not screened the data at all to disambiguate author names or to determine if papers are truly relevant. We are hoping that the analysis will help us to zero in on the relevant authors and papers.

Web-link to dataset(s)

Human Genome Project bibliographic records.

Relevant publications, websites, etc.

http://projectreporter.nih.gov/proje...&icde=19500167

Conditions under which students can publish results and/or add project results to their resume

The only thing we ask is that the raw data not be published before we complete our project. However, it would be okay to post any visualizations that the students create, as long as you send us the links and the project titles, so we can publicize them, too.

Client Forum

Human Genome Project Documentary History:

http://semanticommunity.info/Data_Science/Data_Science_for_VIVO

Updated: Fri, 18 Sep 2015 23:02:30 GMT

Powered by mindtouch

115
Project Title: Federal Library Collection Analysis

Client Name
Steve Short

Project goal/scientific or practical value
Federal libraries provide information services to vast constituencies who seek information across the spectrum of intellectual and creative endeavors. Collections in federal libraries reflect this broad information content and federal librarians now manage a large inventory of resources and data describing the collections as part of the library collections. In the current environment, federal libraries are being asked to better leverage their collections, reduce their space requirements based on the shift from tangible print to electronic resources, identify cost savings where possible, and increase collaboration across the federal government. Currently, there is not an established process for analyzing library resources across the federal government. FEDLINK sponsored a research pilot project to better understand the requisite processes for analyzing bibliographic holdings and overlap among federal libraries. The sharing of Science, Technology, Engineering, and Mathematics (STEM) collections has been a topic of great interest in FEDLINK discussions about shared collection management. STEM collections were therefore the starting point for comparative analysis of federal library holdings. The pilot was started in the summer of 2012 and included a small number of federal libraries with significant holdings in the STEM fields. Eleven libraries provided bibliographic records for use in this pilot project.

Information on dataset(s) to be used
I included screenshots in the file to explain the data structure. For purposes of this sample, I only included OCLC and ISSN match tables. The screenshot for the OCLC match applies to the ISSN table. In addition to the detail, there is a summary table.

Web-link to dataset(s)
STEM Bibliographic records

Relevant publications, websites, etc.
http://calhoun.nps.edu/public/bitstr...pdf?sequence=1

Conditions under which students can publish results and/or add project results to their resume
We would like to be able to review and approve public communications about results. Also attributions should included the Library of Congress.

Client Forum
Federal Library Collection Analysis

Project Title: Global Biotic Interactions
Client Name
Jorrit Poelen

Project goal/scientific or practical value
The mission of this project is to find efficient ways to normalize and integrate species-interaction data. By making this data readily available, GloBI will enable researchers and enthusiasts to answer questions about localized, one-to-one species interactions and big-picture changes in species interactions over time. For example, GloBI can answer which species an Angel Shark (Squatina squatina) eats in the Gulf of Mexico, or return the results of a query for the number of Angel Sharks feeding in the Gulf of Mexico between 2005 and 2010. At time of writing, the Encyclopedia of Life (http://eol.org) and GoMexSI (http://gomexsi.tamucc.edu) use the GloBI infrastructure to access species interaction data.

Information on dataset(s) to be used
At time of writing (Jan 2014), GloBI's spatio-temporal species interaction datasets include about 470k interactions across 21k source taxa (e.g. predators, parasites) and 14k target taxa (e.g. predator, prey) across about 3800 locations (marine, fresh water and terrestrial) around the world.

Web-link to dataset(s)
Species Interaction Data

Relevant publications, websites, etc.

Conditions under which students can publish results and/or add project results to their resume
Minimal conditions are to reference the GloBI project. I'd like to make sure that those that contributed to the GloBI project are easy to find.

Client Forum
Global Biotic Interactions

Interview
Global Biotic Interactions Introduction
Project Title: Synthesizing spatial diet data of fishes from the Gulf of Mexico

Client Name
Dr. James Simons

Project goal/scientific or practical value
The goal of the project is to take spatially explicit, yet very heterogeneous diet data of fishes in the Gulf of Mexico, synthesize it, and then prepare visualizations of that data. We are currently working with one species for which we have completed data entry. This would be very useful to be able to quickly glance at a map of a particular species and get a sense of differences in its diet across space.

Information on dataset(s) to be used
The data are being housed at the Gulf of Mexico Species Interaction (GoMexSI) webpage, with the database going by the same name. The webpage was launched on 3 Sep 2013 and there is still much to be done and much data to be added, but there are approximately 70,000 interactions currently in the database. The data can be explored though the webpage or downloaded to a csv file. Currently the detailed data about the diets are not available on the download, only the record of an interaction, ie predator-prey. In addition, location info and a few other items are in the download.

Web-link to dataset(s)
GoMexSi Project website

Relevant publications, websites, etc.
A published paper that provides a lot of detail about the project and the database can be found at the following web link: http://ccs.tamucc.edu/wp-content/upl...l-BMS-2013.pdf I can provide other reference materials or students are welcome to do searches on key words to find other reference material. I also have a final report on the diet data and spatial food webs of king mackerel that I can share with interested students. It is not posted anywhere at present.

Conditions under which students can publish results and/or add project results to their resume
I would like to know if the students are planning to publish anything using the data. If appropriate I would like to be a co-author, and at the very least we would like the website/database to be attributed, and sources of the data as well.

Client Forum
Synthesizing spatial diet data of fishes from the Gulf of Mexico

Project Title: Knowledge Network Evolution

http://semanticommunity.info/Data_Science/Data_Science_for_VIVO
Updated: Fri, 18 Sep 2015 23:02:30 GMT
Powered by mindtouch
Client Name
Peng-hui Lyu

Project goal/scientific or practical value
Trace the history of the evolution for a given concept.

Information on dataset(s) to be used
Complex Networks database

Web-link to dataset(s)

Relevant publications, websites, etc.

Conditions under which students can publish results and/or add project results to their resume
Students who involved in information visualization and knowledge networks.

Project Title: The Genealogy of Psychoanalysis

Client Name
Michael Clifford

Project goal/scientific or practical value
The goal of this project is to establish the skeleton for a future website, “The Genealogy of Psychoanalysis.” This website will be a family tree of psychoanalysts and (relatedly) of psychoanalytic ideas. We will use genograms (widely used in the mental health professions), which are family trees with symbols that visually show the characteristics of people in the family tree (i.e., gender, sexual orientation, number and dates of marriages, full- or half- sibling, bipolar, alcoholic, etc.), as well as the quality of relationship between people in the family tree (i.e., merged, hostile, estranged, etc.). For this project, new genogram symbols need to be developed (i.e., analyst-patient relationship, supervisor-supervisee relationships, etc.). This website could revolutionize the discipline of the history of psychoanalysis.
Psychoanalysis is unique in the influence of personal relationships on the development of the field. It matters who a psychoanalytic theorist's own analyst, teachers, and supervisors were, and the Genealogy of Psychoanalysis could visualize previously unrecognized relationships between an analyst's professional background, as well as how ideas have developed throughout the history of the discipline. Accumulating the data for the website will be crowd-sourced to the international community of psychoanalysts, and as the director of the website, I will review the data to ensure it meets the standards of the project.

Information on dataset(s) to be used

I will supply information to make the genograms for the project. That is, I will cull the information necessary from standard works in the field, and email it to my team. For example, I will provide information on Freud's grandparents, parents, siblings, teachers, mentors, first patients, etc., as the raw data to build the genograms.

Web-link to dataset(s)

Relevant publications, websites, etc.

The data which would be visualized in this project is embedded as text in the biographies and autobiographies of leading figures in the history of psychoanalysis. There is also information embedded as text in articles published in psychoanalytic journals. Finally, there appear to be no on-line sites comparable this project.

Conditions under which students can publish results and/or add project results to their resume

Conditions: I do not request to be a co-author, but I would like my name and the title of the project to be listed.

Client Forum

The Genealogy of Psychoanalysis

Project Title: Evolution of Wikipedia's Category Structure

Client Name

Knowledge Space Lab
Project goal/scientific or practical value

The Knowledge Space Lab project conducted a research on the evolution of the Wikipedia category system, and produce snapshots of two kind of networks: network of links between category pages; and networks of links between Wikipedia article pages. In the course of the project, a map was produced in which the category network of 2008 is compared with a library classification network[http://scimaps.org/maps/map/design_v...ergence__127/]. What was never done in the project is to visualize the extracted 43 snapshots in time for both of these networks. We would like to invite you to this visualization challenge. Visualizing the growth of the networks should give us insights such as - How does the category system change over time, i.e. which categories are added to which top categories, which categories do shift most (i.e. are changed from one top category to other)? - which areas of the category network grow most? - to which extent a consolidation takes place – increase in connectivity – emergence of quasi hierarchical structures? [see the article in Advances of Complex Systems] - Can we detect any burst in the terms of the labels? The size of network is another challenge. Since the active use of categories, the number of category pages explodes, and so do the links between them. Still there are interesting phases of quasi-stability and growth alternating with turbulent phases of reorganization. You might need to make decisions to extract parts of the network in order to render a ‘readable’ visualization. At the end of the dataset description you will find pointers to snapshots, which might be particularly interesting to visualize.

Information on dataset(s) to be used

The files have names like "enwiki-20080103-pages-meta-history-ALL_categorylinks_vksrun2_build_YYYY-MM-DDT00_00_00Z.txt" and are in Pajek NET format. They contain snapshots of the Wikipedia category structure at specified dates (YYYY-MM-DDT00_00_00Z in the file name). The network contains all article and category pages and all category links of Wikipedia at a given moment. Article pages are normal Wikipedia pages (such as "Earth" or "Ludwig_van_Beethoven") that usually contain information on the topic. Category pages are special Wikipedia pages that basically serve as lists of article pages and other category pages (for example "Category:History" or "Category:Physics"). The special pages (such as portals ("Portal:Physics"), talk pages ("Talk:Physics") and others (ex. "Special:Export_page") are not included in the snapshots. The network contains category links, meaning links that lead from an article page or a category page to another category page. For example the article page "Ludwig_van_Beethoven" has links to category pages such as "Category:Ludwig_van_Beethoven", "Category:Romantic_composers", "18th-century_German_people", etc. These links are category links. Another example of category links are links between category pages. For instance, the category page "Category:Romantic_composers" has category links to "Category:Classical_composers", "Category:Romantic_music" and "Category:19th-century_composers"; note that the article page about Beethoven is not the same page as the category page devoted just to him, and many categories have accompanying article pages or vice versa). The snapshots do not contain any pagelinks. Pagelinks are links between article pages., All other type of links found in category and article pages are disregarded as well. For example links to article pages in different languages, links to external pages, links to talk pages, etc. have been eliminated while preparing the snapshots. Vertices are labeled by text compounded of two numbers. In Wikipedia, all pages have a unique id, i.e. 17914 is id of the page named “Ludwig_van_Beethoven” and 5992558 is id of “Category:Ludwig_van_Beethoven” is a numbering system that reports the status of any given page: The number “0” means that the page does not exist. The number “1” refers to “unknown page type” The number “2” denotes an article-page The number “3” is a category-page. The number “4” refers to special pages. The snapshots contain only page- types with the number of “2” and “3” (if other types are encountered, they can be treated as noise and ignored completely without real impact on the whole dataset). The data has an accompanying file with the name "wiki_names.txt", which contains the names of all Wikipedia pages (articles and categories) in the format "<id> <name>", one entry per text line (id and name are separated by whitespace). The id is the unique Wikipedia id of the page (same as id in the snapshots). This file can be used to obtain actual page names. Additionally, the category structure can be considered to be “rooted” at certain category pages, if we are interested in topical categories. There may be technical categories, such as “Category:Contents” that does not belong to any other category, and contain some technical categories such as “Category:Wikipedia_administration” and “Category:Help” or “Category:Portals”, that do not sort the articles into topical groups, but also contain “Category:Articles” that actually contains the root “Category:Main_topic_classifications”. Following roots have been identified: 2004-05-30 to 2004-06-13 : suitable root has not been found, but it doesn't mean it doesn't exist 2004-06-13 to 2006-10-08 : Category:Fundamental (id 722262)
Web-link to dataset(s)

Evolution of Wikipedia Dataset

Relevant publications, websites, etc.


Conditions under which students can publish results and/or add project results to their resume


Client Forum

Evolution of Wikipedia's Category Structure

Project Title: 30 Years of Alzheimer’s disease Research at NIA

Client Name

Li Shen

Project goal/scientific or practical value

Goals: Create a detailed visual interactive map of Alzheimer’s disease (AD) research literature for the last 30 years. This map should enable us to see (1) how research areas are connected to each other; (2) in which sub area most of the
research has taken place in terms of number of publications; (3) a chronological view to see how research has evolved over time; (4) how National Institute of Aging (NIA) contributes to the literature (in terms of numbers of publication and or citation) in these areas; and (5) the correlation between what NIA funded and what added to the AD knowledge from these funds. Value: Provide a detailed map of AD research over the years. Map NIH/NIA funding to the research areas. Reveal how NIA funding might have guided AD research.

Information on dataset(s) to be used

1. ISI Web of Science: http://ulib.iupui.edu/node/9061
2. Scopus: http://www.scopus.com/
3. NIH RePORTER: http://projectreporter.nih.gov/reporter.cfm

Web-link to dataset(s)
Mapping AD Research

Relevant publications, websites, etc.
http://www.nia.nih.gov/

Conditions under which students can publish results and/or add project results to their resume
Co-authorship; Approval of results by client before publication; Communication during the life of the project

Client Forum
30 Years of AD Research at NIA

Project Title: Globalization of the United States, 1789-1861

Client Name
Konstantin Dierks

Project goal/scientific or practical value
This is intended to be a mapping visualization, generating geo-referenceable historical maps of the way the world, and specific world regions, looked in the 18th and 19th century. It could be year-specific or animated across years. It could focus on one variable or multiple variables. It would have clickable data. Its value is in generating geo-referenceable historical maps of the world and of the various world regions, to hide the accuracy of GIS underneath a display of
historical maps (since modern maps are inappropriate to the past). Its value is in animating the data across time, and layering it across variables. And its value is in being not a mere visualization, but clickable to enable users to pursue their own research angles. These GIS techniques would be applicable to anyone mapping variables onto historical maps.

**Information on dataset(s) to be used**

I have provided datasets for three variables: military, diplomacy, and treaty. • globalizing geography of military actions of United States, 1798-1871 • globalizing geography of U.S. diplomatic missions, 1779-1868 • globalizing geography of U.S. treaties, 1778-1868

**Web-link to dataset(s)**

Professor Dierks website

**Relevant publications, websites, etc.**

the most pertinent is: http://dsl.richmond.edu/historicalatlas/ scroll down to see link to georectified maps, animated maps, and clickable maps these are limited to the United States, whereas my project is global. this page does have a world basemap: http://dsl.richmond.edu/historicalatlas/166/a/ I am not yet interested in layering historic maps, as this seems a separate technical challenge that has been done. I am more interested displaying historical maps with modern GIS capabilities underlying them; this has not been done, as far as I know.

**Conditions under which students can publish results and/or add project results to their resume**

I would aim to be generous, but I would have to approve to make sure that the idea cannot be purloined. Interns have briefly worked on this, and given presentations about their experience. Adding results to a c.v. seems perfectly necessary, but publication creates risk.

**Client Forum**

Globalization of the United States, 1789-1861

**Week 8 - Mar. 18, 2014: Picking a Client**


**Forming a Client Group**

As a unique feature of the IVMOOC, you are invited to apply your new data analysis and visualization skills to a real world project. Specifically, you will use the IVMOOC Forum to self-select a team of four to five students. See existing groups at http://ivmooc.cns.iu.edu/forums/groups. Each team will work on one of the projects listed at http://ivmooc.cns.iu.edu/forums/clients. You will apply or extend existing tools, e.g., the Sci2 Tool, to serve the specific information needs of a real world client. The result is a visualization and a write-up documenting your work.

Most projects will have a data selection, data cleaning and preparation, data analysis, and data visualization phase followed by a detailed examination, discussion, and documentation of results. **Final results need to be submitted by Monday, April 28, 2014 at 5pm EST** (in about 6 weeks):
The first step down this road is to look through the client project descriptions to find one you like. You must make sure you have your full profile filled out on the forums. Once you do, visit http://ivmooc.cns.iu.edu/forums/groups to either create a group, or join one, making sure that each group has 4-5 members. Fewer members will make it hard to finish on time; more members will make it hard to coordinate work. Review the projects listed at http://ivmooc.cns.iu.edu/forums/clients and select exactly ONE for your group. Starting the following week, your group will automatically have a subforum in which you can communicate and collaborate. Make sure your groups are formed by Monday, March 24th.

My Response: See http://ivmooc.cns.iu.edu/forums/node/408

**Data Science for the Federal Big Data Initiative**


is mining NSF Funding Opportunities in Data Science: http://semanticommunity.info/Data_Science/NSF_Funding_Opportunities_in_Data_Science
to work on the Open Government Policy: http://semanticommunity.info/An_Open_Data_Policy


and Project Open Data: http://semanticommunity.info/An_Open_Data_Policy/Project_Open_Data

Our Data Science Teams are making presentations every two weeks to answers the questions:
Where did you get the data?,
Where did you store the data?, and
What were your results?

The recent Data Science Symposium at NIST: http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013

identified the NIST Research Library as an excellent source of digital archives: http://www.nist.gov/digitalarchives

I would like to help a data science team work on this and make a presentation.

You would be welcome to join our Meetup tonight on Graph Databases in person or remotely if someone could handle the SKYPE.

Dr. Brand Niemann
Director and Senior Data Scientist
Semantic Community
http://semanticommunity.info
http://www.meetup.com/Federal-Big-Data-Working-Group/
http://semanticommunity.info/Data_Science/Federal_Big_Data_Working_Group_Meetup

**Week 9 - Mar. 25, 2014: Project Ideas**

Source: http://ivmooc.cns.iu.edu/forums/groups
My Note: I do not know why my project is listed here. I have several emails from students saying that my proposed project is not with one of the original clients.

<table>
<thead>
<tr>
<th>Group Name</th>
<th>Client Project</th>
<th>Description</th>
<th>Competencies Needed</th>
<th>Members</th>
</tr>
</thead>
</table>
| Big Data in Drug Discovery, Health and Translational Medicine #1          | Information Visualizations for Big Data in Drug Discovery, Health and Translational Medicine | Produce visualizations about
1. Research topics
2. Key People
3. Funding
in the field of Drug Discovery.                                                                                           | Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert | 5       |
<p>| Big Data in Drug Discovery, Health and Translational Medicine #2          | Information Visualizations for Big Data in Drug Discovery, Health and Translational Medicine | Goal: To produce information visualizations related to Drug Discovery and Big data applications. Identify key concepts, key researchers and organizations, important publications, and/or grant or patent profiles for researchers and organizations related to drug discovery and translational medicine | Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert | 5       |</p>
<table>
<thead>
<tr>
<th>Group Name</th>
<th>Client Project</th>
<th>Description</th>
<th>Competencies Needed</th>
<th>Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolution of Wikipedia's Category Structure</td>
<td>Evolution of Wikipedia's Category Structure</td>
<td>The challenge is to visualize extracted snapshots in time (43-off) for network of links between Wikipedia article or category pages and Wikipedia category pages. If you are interested in joining this group, read this paper from Suchecki et al.: <a href="http://arxiv.org/pdf/1203.0788.pdf">http://arxiv.org/pdf/1203.0788.pdf</a> Members based in Germany, Israel, the Netherlands and Moscow</td>
<td>Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert</td>
<td>5</td>
</tr>
<tr>
<td>Group Name</td>
<td>Client Project</td>
<td>Description</td>
<td>Competencies Needed</td>
<td>Members</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Globalization of the United States, 1789-1861</td>
<td>Globalization of the United States, 1789-1861</td>
<td>A mapping visualization, generating georeferenceable historical maps of the way the world, and specific world regions, looked in the 18th and 19th century. This looks like a challenging, but fascinating, project! Chandra Guglik -- Eastern time zone</td>
<td>Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert</td>
<td>5</td>
</tr>
<tr>
<td>Knowledge Network Evolution</td>
<td>Knowledge Network Evolution</td>
<td>Project goal/scientific or practical value: Trace the history of the evolution for a given concept. Client - Penghui Lyu - Wuhan, China UTC/GMT +8 Joe Kelly (Lansing, MI, USA) - Eastern Time Zone UTC/GMT -4 Yulia Markova (Moscow, Russia) - UTC/GMT +4 Andrea Pavlick - Eastern Time Zone</td>
<td>Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert</td>
<td>7</td>
</tr>
<tr>
<td>Group Name</td>
<td>Client Project</td>
<td>Description</td>
<td>Competencies Needed</td>
<td>Members</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Local Group</td>
<td>Human Genome Project Documentary History: An Annotated Scholarly Guide to the HGP.</td>
<td>Group for people in or near Indiana. Members: Mark Ciganovic, Ben Fulton, Kate Garlock, David Peters, Sarah Soliman, Omar Sosa Tzec</td>
<td>Data Mining, Visualization Expert</td>
<td>6</td>
</tr>
<tr>
<td>Synthesizing spatial diet data of fishes from the Gulf of Mexico</td>
<td>Synthesizing spatial diet data of fishes from the Gulf of Mexico</td>
<td>This group is now full Members: Shannon Sofian (Central Time Zone, US), Jessica Jones (Central Time Zone, US), Attila Csapo, Sushant Pritmani, Giuseppe Margiotta</td>
<td>Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert</td>
<td>6</td>
</tr>
<tr>
<td>Group Name</td>
<td>Client Project</td>
<td>Description</td>
<td>Competencies Needed</td>
<td>Members</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-----------------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td><strong>The Genealogy of Psychoanalysis</strong></td>
<td><strong>The Genealogy of Psychoanalysis</strong></td>
<td>Looking for members to work on this exciting project!</td>
<td>Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert</td>
<td>7</td>
</tr>
<tr>
<td><strong>Visualization of Species Interaction Data</strong></td>
<td><strong>Global Biotic Interactions</strong></td>
<td></td>
<td>Data Manager, Data Mining, Designer, Programmer, Technical Writer, Usability Expert, Visualization Expert</td>
<td>5</td>
</tr>
<tr>
<td><strong>Visualizing 30 Years of Alzheimer's disease Research</strong></td>
<td><strong>30 Years of Alzheimer's disease Research at NIA</strong></td>
<td>snip: Create a detailed visual interactive map of Alzheimer's disease (AD) research literature for the last 30 years.</td>
<td>Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert</td>
<td>8</td>
</tr>
<tr>
<td>Group Name</td>
<td>Client Project</td>
<td>Description</td>
<td>Competencies Needed</td>
<td>Members</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>---------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Visualizing Federal Library Collections (VFLC)</td>
<td>Federal Library Collection Analysis</td>
<td>This group is for anyone, who has diverse backgrounds and is interested in analyzing and visualizing the collections in the federal libraries.</td>
<td>Data Manager, Data Mining, Designer, Librarian, Programmer, Project Manager, Technical Writer, Usability Expert, Visualization Expert</td>
<td>9</td>
</tr>
</tbody>
</table>

**Week 10 - Apr. 1, 2014: 1st Project Draft**

Source: [http://ivmooc.cns.iu.edu/project_instructions.html](http://ivmooc.cns.iu.edu/project_instructions.html)

**Complete by Monday, Mar. 31, 2014 at 5pm EST**

**Detail Project**
Discuss what analyses and visualizations you plan to run. Post questions to the forum. Compile hand sketches (simply draw on paper, photograph or scan, upload to Flickr or Twitter, and share link via the forum) to communicate envisioned results to other team members and/or to your client for feedback.

To use the forum for team-specific client work, simply go to ‘Student-Client Discussion’ on [http://ivmooc.cns.iu.edu/forums/forum](http://ivmooc.cns.iu.edu/forums/forum), click on the sub-forum for the your project group, which will be found in the forum for the project you plan to work on, create a ‘New topic’ for your team, and use subtopics to organize your discussion and to share your results. Please note: your sub-forum will take time to be set up after you create a new group.

**Submit via Forum**
Compile a write-up (see sample at [http://ivmooc.cns.iu.edu/docs/ProjectForm.pdf](http://ivmooc.cns.iu.edu/docs/ProjectForm.pdf)) (PDF) comprising:

2. Title of planned visualization
3. Team/Author names and affiliations
4. Description of visualization goal/need, hand-sketch of the envisioned visualization, and discussion why this project is important
5. Discussion of related work. Here you will have to do a literature/web search to find out what approaches/visualizations already exist. You don’t want to re-invent but to build on existing work.
Create a subtopic named ‘Write-up #1-#5’ that points to all your results from that week so that others can comment.

#1: Project Title
#2: Title of Planned Visualization
#3: Authorship
1st Author Name 2nd Author Name 3rd Author Name
Affiliation of 1st Author Affiliation of 2nd Author Affiliation of 3rd Author
1@domain.name 2@domain.name 3@domain.name
#4: Description of visualization goal/need, hand-sketch of the envisioned visualization, and discussion why this project is important.
Text …
#5: Discussion of related work. Here you will have to do a literature/web search to find out what approaches/visualizations already exist. You don’t want to re-invent but to build on existing work.
Text …
#6: Simple statistics of the data sets used, e.g., number of entities, major entity attributes, etc.
Text …
#7: Data analysis/visualization (algorithms) applied and resulting visualizations.
Text and links to visualization(s).
#8: Discussion of key insights gained from the analysis/visualization.
Text …
#9: What problems surfaced during validation and how does your redesign resolve them?
Text and links to final visualization(s).
#10: Discussion of challenges and opportunities
Text …
#11: Acknowledgements
Thank people who provided data, tools, or resources or that helped you with this project, e.g., your client or colleagues/friends that participated in the validation studies.
#12: References

Week 11 - Apr. 8, 2014: 1st Project Draft

Source: http://ivmooc.cns.iu.edu/project_instructions.html

Complete by Monday, Apr. 7, 2014 at 5pm EST

First Project Draft – Data Analysis & Visualization
Use the best tool(s) and run different workflows with alternative parameter settings. Keep a log of all runs—alogous to ‘Console’ printouts in Sci2—so that you can describe the best workflow in the final write-up. Explore at least two alternative visualizations of your data. Make sure each visualization has a legend. My Note: Yes
Submit via Forum
Revise the existing write-up, making changes from any comments you’ve received and discussion you’ve had with your group (see items #1-5 above). Add to these items:

6. Simple statistics of the data sets used, e.g., number of entities, major entity attributes, etc.
7. Data analysis/visualization (algorithms) applied and resulting visualizations
8. Discussion of key insights gained from the analysis/visualization

Create a subtopic named ‘Write-up #1-#8’ that points to all your results from that week so that others can comment.

Week 12 - Apr. 15, 2014: Peer Feedback

My Note: About 100 students looked at this project, about 10 students expressed interest by email, but only 4 signed up possibly because I was not one of the original clients. One student told me that my work was "really interesting and praiseworthy".

The principal clients: NIST, NSF, and NIH have been very complimentary. One senior NIST official said: "Thanks again for your effort in putting this program together"! One senior NSF official attended the Federal Big Data Working Group Meetup where this work was mentioned and was very complimentary and several NIH senior officials have also been very complimentary of the Joint NIH-NSF Big Biomedical Data Meetup that we had recently featuring work on "data papers", which Data Science for VIVO and the IV MOOC and NIST Scientific Data for Data Science are.

Complete by Monday, Apr. 14, 2014 at 5pm EST

Source: http://ivmooc.cns.iu.edu/project_instructions.html

First Project Draft – Validation & Redesign
Validate the data analysis/visualization results by having your friends, colleagues, and if possible your client(s) comment on them. Do they understand the visualization (elements)? What do they miss? In addition, validate your results by comparing them with related work if applicable (see item #6).

Redesign the visualizations you have included so far. Optimize analysis workflows and visualizations to address problems identified during evaluation. Use Photoshop, Inkscape, or other programs to enhance the expressiveness and legibility of your visualization(s).

Submit via Forum
Revise last week’s write-up taking into account discussion and feedback (see items #1-8 above). Add two more points this week:

9. What problems surfaced during validation and how does your redesign resolve them?
10. Discussion of challenges and opportunities. Address complexity and scaling issues, desirable modifications and extensions of your work. You have only four weeks for this project restricting the amount of work that can be done considerably. I would like to know what promising avenues you see for future work.

In addition to these two final points, create a visualization or small set of visualizations that pull everything together, that represents your data and the story you are trying to tell. My Note: Yes
Create a subtopic named ‘Write-up #1-#10’ that points to all your results from that week so that others can comment.

Keep in mind this draft will be reviewed by your peers in other groups, so make it legible and professional!

**Week 13 - Apr. 22, 2014: 2nd Project Draft**

Source: [http://ivmooc.cns.iu.edu/project_instructions.html](http://ivmooc.cns.iu.edu/project_instructions.html)

**Complete by Monday, Apr. 21, 2014 at 5pm EST**

**Peer Feedback**

Your assignment this week will be to read through all of the group project first drafts, excluding your own, and to comment on them in the ‘Write-up #1-#10’ subtopic of that particular group. Please keep your comments constructive, and make sure they address all 5 points in the grading rubric at the bottom of this page. **My Note: NEED TO DO THIS**

**Week 14 - Apr. 28, 2014: Project Submission Due**

Source: [http://ivmooc.cns.iu.edu/project_instructions.html](http://ivmooc.cns.iu.edu/project_instructions.html)

**Complete by Monday, Apr. 27, 2014 at 5pm EST:**

**Second Project Draft – Visualization & Write-Up**

Put all the elements together: Use text from your 10-item write-up and the most insightful visualizations you created to compile a visualization that "tells it all." Browse the maps from [http://scimaps.org/exhibit_info](http://scimaps.org/exhibit_info) and other visualization sited for inspiration on how to layout text and images in a synergistic manner. Use the feedback given to you by your peers to improve your work.

Review your 10-item write-up. It should explain the data analyses/visualizations you did. Make sure that if somebody else reads this explanation s/he is able to obtain the very same results you got. Update items as needed. Make sure the write-up is internally consistent. Use your peer-feedback. If relevant, add an ‘Acknowledgment’ section in which you thank people who provided resources or helped you with this project. Also provide a ‘References’ section with complete data, software, literature references to all work you use or discuss.

Please aim for quality and not quantity. The final documentation should print on no more than four pages on a letter size printout. A sample form is at [http://ivmooc.cns.iu.edu/docs/ProjectForm.pdf](http://ivmooc.cns.iu.edu/docs/ProjectForm.pdf).

**Week 15 - Presentations**

Source: [http://ivmooc.cns.iu.edu/project_instructions.html](http://ivmooc.cns.iu.edu/project_instructions.html)

**Complete by Monday, May 5, 2014 at 5pm EST:**

**Finalize Everything**

Discuss your project with your team and, if applicable, your client. Ask for peer reviews from your fellow students if you
think someone or some group may have particular insight into your project. Make sure everything is professionally presented, error-free, and fit for production. Create your final draft.

Official Hand-In

Upload your finalized visualization to Flickr with the #ivmooc tag (we use Flickr here because of its ability to easily host large image files). Make your final write up available online as Google document or PDF file. Create a subtopic named ‘Client Project Results’ that points to the image shared via Flickr and to the final write-up.

Grading Rubric:
Your project results will be evaluated based on:

- Quality of data selection, cleaning, preparation, and documentation
- Appropriate selection of tools, algorithms, workflows, and parameter values
- Quality of data analysis, visualization results, and discussion of insights gained
- Completeness and quality of validation and redesign
- Overall quality of content, including the accuracy and completeness of information, the expressiveness and clarity in communication of ideas using text and visualizations, and the appropriateness of references to/attribution(s) for the work of others

This is a group assignment, so please work as a group. Try to find the strength of each team member and distribute and organize team work so that you maximize the result. Form, storm, norm, and perform!

My Project Results:

- NIST Scientific Data for Data Science
- Data Science for VIVO and the IV MOOC

for the May 6th Federal Big Data Working Group Meetup

Post-Questionnaire