Big Data Science for Astronomy & Space

Story
Story
Story
Slides
Slide 1 Big Data for Government Symposium
Slide 2 Big Data in Space and Earth Sciences
Slide 3 Astronomy Example
Slide 4 LSST = Large Synoptic Survey Telescope 1
Slide 5 LSST = Large Synoptic Survey Telescope 2
Slide 6 LSST = Large Synoptic Survey Telescope 3
Slide 7 LSST Key Science Drivers: Mapping the Dynamic Universe
Slide 8 LSST Key Science Drivers: Observing Strategy
Slide 9 LSST Summary
Slide 10 The LSST Big Data Challenges
Slide 11 What is Big Data Science good for? 1
Slide 12 Characteristics of Big Data Science
Slide 13 Rationale for Big Data Science
Slide 14 Characterizing and Exposing the Big Data Hype: 3 V's or ? 1
Slide 15 Characterizing and Exposing the Big Data Hype: 3 V's or ? 2
Slide 16 Insufficient Variety: stars & galaxies are not separated in this parameter
Slide 17 Sufficient Variety: stars & galaxies are separated in this parameter
Slide 18 The 3 important D's of Big Data Variety
Slide 19 This graph says it all …
Slide 20 Data-Driven Discovery
Slide 21 Correlation Discovery
Slide 22 Initial impression 1
Slide 23 Initial impression 2
Slide 24 Novelty Detection = Surprise Discovery!
Slide 25 Novelty Detection = improved discovery
Slide 26 Association Discovery between events on the Sun and around the Earth
Slide 27 Association Discovery = finding interesting co-occurring associations
Slide 28 Knowledge Discovery for Multi-source Data
Slide 29 Space Weather Example
Slide 30 What is Big Data Science good for? 2
Slide 31 Case Study - Mars Rovers 1
Slide 32 Case Study - Mars Rovers 2
Slide 33 Mars Rover 3
Slide 34 Big Data Science meets Big Science Simulations
Slide 35 The Big Picture of Big Data in Space and Earth Sciences

Spotfire Dashboard
Research Notes
Web Addresses

TAEM Interview with Dr. Kirk Borne of George Mason University

TAEM-Professor Borne, please tell our readers about your educational training for your fields
TAEM-You are also a member of your university’s SPACS program. Please tell us about it and the goals that it has set forth
TAEM-What part do you play in its agenda?
TAEM-Your specialties are listed as a Data Scientist, Astrophysicist, Big Data Science Consultant, and Public Speaker
TAEM-Please tell us in detail about Transdisciplinary Data Science.
TAEM-How alike are Big Data and Large Data Bases, and how do they interact?
TAEM-How does Data Mining and knowledge discovery figure into these?
TAEM-We understand that you basically invented the idea of astro-informatics
TAEM-We understand that your findings involving Large Database Astronomy has pinpointed Groups and Clusters of Galaxies
TAEM-What advice and support can you give to NASA for its future programs of space exploration?
TAEM-What programs are you working on that would pertain to this, and what research projects are you planning on for the future?
TAEM-What information can you give us so that the many students who read our publication learn more about you and the school’s programs?
TAEM-Dr. Borne, it has been a sincere honor to be able to interview you for our magazine

Sample Project Summary

Intellectual merit
Broader impacts

CV Dr. Kirk D. Borne

Title
Professional Preparation
Appointments
Related Products
**Big Data Science for Astronomy and Space**

Astronomy is the way man devised to originally understand the universe and space is what man now used to describe everything that is out there.

The Large Sky Survey Telescope (LSST), which is the primary NSF project for the decade, will be the most powerful astronomical imaging device ever built. One of the reasons for LSST is to track objects 100 m objects that could wipe out cities.

Interestingly, a group of former astronauts announce recently an effort to accomplish the same goal by space means, not a telescope.

**Far More Asteroids Have Hit The Earth Than We Thought, Astronauts Say**

Bad news, earthlings. A former NASA scientist says it’s mere happenstance that an Armageddon-style asteroid hasn’t hit a densely populated area in the last few years.

On Tuesday, the B612 Foundation, which is devoted to preventing the next deep impact, will present data from a nuclear-weapons test warning satellite showing that far more asteroids have hit earth in the past few years than previously thought, the organization announced on its website.

The data, collected from a nuclear missile detection system that picks up large blasts on earth, shows that since 2001, asteroids have caused 26 explosions on the scale of an atomic bomb.

“This data shows that asteroid impacts are NOT rare, but actually 3-10 times more common than we previously thought,” Ed Lu, one of the astronauts working on the project, said in a statement. "The fact that none of these asteroid impacts shown in the video was detected in advance is proof that the only thing preventing a catastrophe from a 'city-killer' sized asteroid is blind luck."

The silver lining? Scientists are working to deflect any future space rocks from our planet.

Lu, along with fellow ex-astronauts Tom Jones and Bill Anders, has been attempting to develop a better asteroid early-warning system, the Sentinel Infrared Space Telescope, which they hope will become "the principal means by which nearly all asteroid discoveries will be made." In an interview with Wired, Lu explained that the telescope will work by scanning the sky in infrared, which will allow it to calculate the trajectory and velocity of asteroids.

NASA has made efforts to track asteroids, but they haven’t been as aggressive as the B612 scientists would like. In 1998, NASA established the Near-Earth Object Program Office to detect potentially hazardous comets and asteroids.
March, the agency announced a contest for scientists to develop asteroid-detecting algorithms, a year after a meteorite explosion in Russia made international headlines.

You can read more about B612 and its mission from Slate's Phil Plait here.

---

**Story**

**Semantic Data Science Harvard’s ADS System**

**ADS** is "the" repository of all astronomy pubs since recorded history, going back to Ptolemy, including all journals today.

Using journal collections from multiple disciplines is a good plan.

They are funded by NASA and they are doing excellent funded semantics work already on their collection.

The NASA ADS team has their own linked data (articles + data sets) approach that is well developed and mature (and deeply discipline-specific).

Find the ADS presentation I attended in this knowledge base.

---

**Story**

**Dr. Kirk Borne of George Mason University**

Mine Tweets - See Book Developing Analytical Talent

See [Workshops on Extremely Large Databases](#)

See recent articles: Kirk Borne is blogging and consulting for MapR (also see [Glossary](#) for letter Y).

Creating Linked Data in RDF format in spreadsheets for input to Spotfire, Semantic Insights, and YarcData from publications, journals, books, Web pages, etc.

- [http://www.esa.int/ESA](http://www.esa.int/ESA)
- [http://www.esa.int/spaceinvideos/Videos/2012/06/ESA_-_Space_to_Relax_Stellar_Works_of_Art](http://www.esa.int/spaceinvideos/Videos/2012/06/ESA_-_Space_to_Relax_Stellar_Works_of_Art)
- "Linking Digital Humanities Content to the Web of Data" [http://t.co/76htOVEJtv](http://t.co/76htOVEJtv)

The published scientific data will be shared via the World Wide Web (WWW) using our Data Management Plan for sharing semantically relevant data. This directly benefits the global community by enabling rapid collaboration and enrichment of our novel engineering techniques with a fuller spectrum of analysis to include intellectual, descriptive, exploratory, inferential, predictive, causal, and mechanistic data analytics. For example, as our innovations with asteroid mapping deliver new data and discoveries, we will be able to cross reference these new resources with rich media in the Sloan Digital Sky Survey (www.sdss.org) to enable better educational and encyclopedic documentation for the non-scientific community.
Slide 1 Big Data for Government Symposium

http://www.ttcus.com

Slide 2 Big Data in Space and Earth Sciences

@KirkDBorne
Slide 3 Astronomy Example

Astronomy Example
Let us begin with a space science example …

The LSST (Large Synoptic Survey Telescope)

Slide 4 LSST = Large Synoptic Survey Telescope 1

http://www.lsst.org/

Slide 5 LSST = Large Synoptic Survey Telescope 2

http://www.lsst.org/
Slide 6 LSST = Large Synoptic Survey Telescope

http://www.lsst.org/
LSST Key Science Drivers: Mapping the Dynamic Universe

- Solar System Inventory (moving objects, NEOs, asteroids, census & tracking)
- Nature of Dark Energy (distant supernovae, weak lensing, cosmology)
- Optical transients (of all kinds, with alert notifications within 60 seconds)
- Digital Milky Way (proper motions, parallaxes, star streams, dark matter)

LSST in time and space:
- When? ~2022-2032
- Where? Cerro Pachon, Chile

Observing Strategy:
One pair of images every 40 seconds for each spot on the sky, then continue across the sky continuously every night for 10 years (~2022-2032), with time domain sampling in log(time) intervals to capture dynamic range of transients.

- LSST (Large Synoptic Survey Telescope):
  - Ten-year time series imaging of the night sky – mapping the Universe!
  - ~10,000,000 events each night – anything that goes bump in the night!
  - Cosmic Cinematography! The New Sky! @ http://www.lsst.org/

LSST Summary

http://www.lsst.org/
Slide 10 The LSST Big Data Challenges

The LSST Big Data Challenges

- 3-Gigapixel camera
- One 6-Gigabyte image every 20 seconds
- 30 Terabytes every night for 10 years
- 100-Petabyte final image data archive anticipated – all data are public!!
- 20-Petabyte final database catalog anticipated
- Real-Time Event Mining: ~10 million events per night, every night, for 10 yrs
  – Follow-up observations required to classify these
- Repeat images of the entire night sky every 3 nights: Celestial Cinematography
Slide 11 What is Big Data Science good for?

What is Big Data Science good for?

- Discovery
- Data-driven Decision Support

Slide 12 Characteristics of Big Data Science

Characteristics of Big Data Science

- The emergence of Data Science and Data-Oriented Science (the 4th paradigm of science).
- A complete data collection on any complex domain (e.g., Earth, the Universe, or the Human Body) has the potential to encode the knowledge of that domain, waiting to be mined and discovered.
- We call this "Xinformatics": addressing the D2X (Data-to-Knowledge) Challenge in any discipline X using Data Science.
- Examples: Astroinformatics, Bioinformatics, Geoinformatics, Climate Informatics, Ecological Informatics, Biodiversity Informatics, Environmental Informatics, Health Informatics, Medical Informatics, Neuroinformatics, Crystal Informatics, Cheminformatics, Discovery Informatics, and more.
Slide 13 Rationale for Big Data Science

Rationale for Big Data Science

- If we collect a thorough set of parameters (high-dimensional data) for a complete set of items within our domain of study, then we would have a "perfect" statistical model for that domain.
- In other words, Big Data becomes the model for a domain X = we call this X-informatics.
- Anything we want to know about that domain is specified and encoded within the data.
- The goal of Big Data Science is to find those encodings, patterns, and knowledge nuggets.

See article: Big-Data Vision? Whole-population analytics
- From Reporting to Prediction
- From hindsight & oversight to insight & foresight
- From DATA to INFORMATION to KNOWLEDGE to UNDERSTANDING

Slide 14 Characterizing and Exposing the Big Data Hype: 3 V’s or ? 1

http://bit.ly/1hH6sB9

Characterizing and Exposing the Big Data Hype: 3 V’s or ?

- If the only distinguishing characteristic was that we have lots of data, we would call it "Lots of Data" (or a Tonnabytes!)
- Big Data characteristics: the 3+n V’s =
  1. **Volume** (lots of data = "Tonnabytes")
  2. **Variety** (complexity, curse of dimensionality, many formats)
  3. **Velocity** (high rate of data and information flow, real-time, incoming!)
  4. **Veracity** (necessary & sufficient data to test many hypotheses)
  5. **Validity** (data quality, governance, master data management)
  6. **Value** (= the all-important V)
  7. **Variability** (dynamic, evolving, spatiotemporal data, time series)
  8. **Venue** (distributed, heterogeneous, multiple platforms/owners)
  9. **Vocabulary** (ontologies, semantics, schema, data models, …)
  10. **Vagueness** (confusion over the meaning of Big Data, tools, methods, …)

Slide 15 Characterizing and Exposing the Big Data Hype: 3 V’s or ? 2

http://bit.ly/1hH6sB9
Slide 16 Insufficient Variety: stars & galaxies are not separated in this parameter

http://bit.ly/1hH6sB9
Slide 17 Sufficient Variety: stars & galaxies are separated in this parameter

Slide 18 The 3 important D’s of Big Data Variety

Slide 19 This graph says it all …

http://www.cs.princeton.edu/courses/...BrunnerDPS.pdf
This graph says it all ...

3 Steps to Discovery - Data Mining your Big Data

- **Unsupervised Learning**: Cluster Analysis - partition the data items into clusters, without bias, ignoring any initially assigned categories = Class Discovery!
- **Supervised Learning**: Classification - for each new data item, assign it to a known class (i.e., a known category or cluster) = Predictive Power Discovery!
- **Semi-supervised Learning**: Outlier/Novelty Detection - identify data items that are outside the bounds of the known classes of behavior = Surprise Discovery!

---

### Data-Driven Discovery

(KDD: Knowledge Discovery from Data)

1. Class Discovery
2. Correlation Discovery
3. Novelty Discovery
4. Association Discovery
   - Finding unusual (improbable) co-occurring associations

---

**Benefits of very large datasets:**
- Best statistical analysis of “typical” events
- Automated search for “rare” events
Slide 21 Correlation Discovery

Correlation Discovery: Fundamental Plane for 156,000 cross-matched Sloan-2MASS Elliptical Galaxies: plot shows variance captured by first two Principal Components as a function of local galaxy density.

Reference: Borne, Dutt, Gannella, Kargupta, & Griffin 2008

low (Local Galaxy Density) high

% of variance captured by PC1+PC2

Slide 22 Initial impression 1

Initial impression is that the data are extended in only one direction (principal component)
Slide 23 Initial impression 2

Initial impression is that the data are extended in only one direction (principal component)

But, a cut across the width of the data distribution reveals a high Kurtosis ... hence there are 2 Classes!

(an example of ICA = Independent Component Analysis)

Slide 24 Novelty Detection = Surprise Discovery!

Novelty Detection = Surprise Discovery!

- **Outlier detection:** (unknown unknowns)
  - Anomaly Detection, Deviation Detection, Surprise Discovery, Novelty Discovery: Finding objects and events that are outside the bounds of our expectations (outside known clusters)
  - Finding new, rare, one-in-a-million(billion)(trillion) objects and events
  - These may be real scientific discoveries or garbage
  - Outlier detection is therefore useful for:
    - Anomaly Detection – is the detector system working?
    - Data Quality Assurance – is the data pipeline working?
    - Novelty Discovery – is my Nobel prize waiting?
  - How does one optimally find outliers in 10^2-D parameter space? or in interesting subspaces (in lower dimensions)?
  - How do we measure their “interestingness”?
Slide 25 Novelty Detection = improved discovery

Novelty Detection = improved discovery of rare objects across multiple databases

Slide 26 Association Discovery between events on the Sun and around the Earth

Association Discovery between events on the Sun and around the Earth
Slide 27 Association Discovery = finding interesting co-occurring associations

Slide 28 Knowledge Discovery for Multi-source Data
Slide 29 Space Weather Example

Slide 30 What is Big Data Science good for? 2

What is Big Data Science good for?

- Discovery
- Data-driven Decision Support
Slide 31 Case Study - Mars Rovers 1

Slide 32 Case Study - Mars Rovers 2
Slide 33 Mars Rover 3

Mars Rover = Decision Science Engine
(enabling data-to-decisions = DTD)

- Decisions are based on data mined, prior experience, new knowledge, and the set of learned rules.
- Rover acts autonomously, without human intervention, in Deep Space environment.
- Actions are driven by mining actionable data from all sensors.

Slide 34 Big Data Science meets Big Science Simulations

- Cosmology (colliding galaxies: crash science)
- Climate Science
- Fusion Science
- Vehicle Safety (colliding cars: crash science)
- Digital Manufacturing
- Aircraft, Ship, and Automotive Design
- Multiphysics, Turbulence, Energy systems, etc. ...

Characterize, measure, and track massive data outputs for: deviations, anomalies, emergent behavior & patterns, "events", signals of changes in system stationarity, ...
- Enabling Discovery and Data-Driven Decision-making
The Big Picture of Big Data in Space and Earth Sciences

Knowledge Discovery
- BD2K = Big Data-to-Knowledge  [http://bd2k.nih.gov]
- Class discovery: predictive power discovery
- Association discovery
- Correlation discovery
- Novelty discovery: surprise!

Data-driven Decision Support
- Data-to-Decisions (DTD)
- Predictive & Prescriptive Analytics
- The Last Mile Challenge (Actionable Intelligence)
- Decision Science-as-a-Service™ [http://www.apptiwa.com]

Big ROI (Return On Innovation) !!!

Spotfire Dashboard

Research Notes

Web Addresses

http://spacs.gmu.edu/profile/kirk-borne/


http://spacs.gmu.edu/wp-content/uplo...enData2012.pdf

TAEM Interview with Dr. Kirk Borne of George Mason University

TAEM- The Arts and Entertainment’ Magazine’s publisher, Joseph J. O’Donnell, issued a challenge in the December 15th issue of our publication to start a ‘grass roots movement’ to support NASA. This support is spreading over the academic world and its start has taken place in the George Mason University faculty and student body. The challenge has centered on not only on Support of NASA, but to give the agency ideas for space exploration for its future programs.

Dr. Kirk Borne, of GMU, is a Data Scientist and Astrophysicist, and is one of the many professors from the school that has stepped forward to offer insights into what can be achieved.

**TAEM-Professor Borne, please tell our readers about your educational training for your fields**

KB- My undergraduate B.S. degree was in Physics at Louisiana State University, with a lot of math and some astronomy. My goal was to study astronomy in graduate school, so the math and physics coursework was essential. I loved all of those topics, and astronomy gave me the opportunity to study them all. I went to graduate school at Caltech, receiving a PhD in astronomy in 1983. I studied under some of the great astronomers of that era. It was a fantastic experience. In the years since then, I worked on NASA’s Hubble Space Telescope project for 10 years and at NASA’s Astronomy Data Center within the Space Science Data Operations Office at the Goddard Space Flight Center for another 10 years, and I have now been at George Mason University since 2003. All of my research and my work experiences at NASA always involved working with scientific data – this led me to the field of Data Science, which is the application of data methods and algorithms to the study of any discipline.
TAEM-You are also a member of your university’s SPACS program. Please tell us about it and the goals that it has set forth

KB- SPACS is the School of Physics, Astronomy, and Computational Sciences. This is a unique program among universities in the US. Our faculty and students focus on a wide range of research problems involving physics, astronomy, computational science, and data science.

TAEM-What part do you play in its agenda?

KB- I have helped to develop the Data Science curriculum within the school. In that capacity, I am the undergraduate advisor for students in the Computational & Data Sciences B.S. degree program. I also advise many graduate students within our CSI (Computational Science & Informatics) PhD program. I teach courses related to Data Science, including Scientific Databases, Computational Data Science, Data Mining, and Data Ethics. In addition to advising and teaching in this program, I also carry out data science research – mostly in astronomy, but covering many other fields.

TAEM-Your specialties are listed as a Data Scientist, Astrophysicist, Big Data Science Consultant, and Public Speaker

Please describe your capacities in these venues and how they are connected.

KB- I have many years of experience working with data, databases, and data science methodology (including data mining, statistics, and visualization). This experience includes teaching and research, but it also has led me to assist, advise, and consult with other organizations and federal agencies regarding their data activities. This has gained me some notoriety, so I receive many (10 to 20) invitations to speak at conferences and universities worldwide each year on the topic of Data Science, specifically “Big Data”. My two most amazing experiences in this capacity are these: first, in 2001, I was asked to brief the US President on data mining; and second, in 2011, I was the conference keynote speaker at the Medicare and Medicaid Statistics Conference. I never imagined such experiences when I was focusing only on my astronomy research years ago.

TAEM-Please tell us in detail about Transdisciplinary Data Science.

KB- Transdisciplinary is different from multi-disciplinary or interdisciplinary in that it refers to the fact that Data Science transcends traditional discipline boundaries. I can work with financial experts, climate scientists, agriculture specialists, criminologists, drug safety organizations, and library staff on data issues without necessarily requiring me to learn their field or requiring them to learn astronomy. The language of Data Science (databases, data, metadata, statistics, visualization, data mining) transcends those discipline-specific concepts. Data Science enables productive, meaningful, and enlightening research experiences across discipline boundaries for everyone involved. For me personally, I like to think of myself as a Transdisciplinary Data Scientist because my research on data mining algorithms, data structures, data management methods, and statistics are applicable to almost any discipline.

http://semanticommunity.info/?title=Data_Science/Big_Data_Science_for_Astronomy_%26_Space
Updated: Sat, 19 Sep 2015 07:33:50 GMT
Powered by mindtouch™
TAEM-How alike are Big Data and Large Data Bases, and how do they interact?

KB- Big Data is a concept that conveys many meanings and implications to different audiences. It includes large databases, but it includes tons of other things, including large data collections that are not in databases (such as Internet blogs, social network postings, news reports, online video content, images, audio, publications, articles, and anything else that is in digital or non-digital form). Large databases are only a small subset of Big Data collections. The Big Data concept also conveys additional meanings, such as the challenges associated with discovery, access, mining, analysis, and interpretation of massive data collections. These challenges include the large data volume, but also the complexity of the data (as indicated above, the large variety of data types), as well as the enormous rate at which data are being generated in the world today. The data production rate doubles every year, which means that the world will have at least 1000 times more data ten years from now, one million times more data 20 years from now, and so on. One estimate of the total amount of data created in the world each and every day is three exabytes, which is three million terabytes or three billion gigabytes. This is more than a thousand times all of the information in all of the books and journals in all of the libraries in the world that have ever been published in the history of humanity. This is the amount that we create each new day this year – we will create roughly this same amount of data every 90 seconds ten years from now, and roughly this same amount every 1/10 of a second 20 years from now! We must teach the next generation how to handle, deal with, cope with, and make use of this information flood. Big Data therefore refers to the combination of all of these challenges and issues (both technological and human).

TAEM-How does Data Mining and knowledge discovery figure into these?

KB- The collection of large data into archives and databases is useless unless you intend to use the data for something; and that “something” is discovery of patterns, trends, correlations, features, outliers, anomalies, and unexpected “knowledge” nuggets hidden within these enormous data sets (i.e., finding the proverbial needle in the haystack). That is what we call data mining, which is also called Knowledge Discovery from Data. This process is also referred to as “Learning from Data”. When it is applied to making decisions about future behaviors of consumers, or systems, or physical processes, data mining is called “Predictive Analytics”. Yes, people use data mining to predict the future. There are numerous movies and TV shows that highlight these techniques – the techniques are real (even though the shows are fictional). Finding new knowledge is what science is all about. The amazing thing now is that businesses, agencies, sports teams, grocery stores, entertainers, social networking sites, and everyone else have realized that they too can discover new knowledge (e.g., predict behaviors or outcomes) from their data collections: ticket sales, purchases, buying patterns, behavior patterns, system logs, and more.
TAEM—We understand that you basically invented the idea of astro-informatics

Please tell our readers about this and some of the Big Data issues that faces science today.

KB—Informatics is the application of Data Science to a specific discipline, though we sometimes simply define Informatics as Data Science. Some science-specific informatics disciplines are well established, including Bioinformatics and Geoinformatics. It occurred to me about 10 years ago that astronomy’s big (and growing) data collections would require a similar methodological subdiscipline of the field of astronomy—I called this Astroinformatics. I used this word for many years without much uptake by the astronomy community. I published a journal paper and another paper for the National Academy of Sciences on Astroinformatics 3 years ago—now, it is a very commonly used term, there are Astroinformatics conferences every year, and there are Astroinformatics committees within the major astronomy professional societies (in the US and internationally). I am a member of all of these committees. It is very exhilarating to see how far the field has come in such a short time. The real reasons for these informatics approaches to science are the same reasons mentioned earlier: scientists are generating huge amounts of data from their experiments, we want to explore these data collections as effectively and efficiently as possible, we want to discover all of the knowledge that is hidden in these data, and we want to unlock the mysteries of the world and Universe around us that our huge experiments can now reveal to us.

![Cosmic Epochs](image)

TAEM—We understand that your findings involving Large Database Astronomy has pinpointed Groups and Clusters of Galaxies

How does Massive Databases and Large Sky Surveys make this possible?

KB—My current research is focused on outlier detection, which I prefer to call Surprise Discovery—finding the unknown unknowns and the unexpected patterns in the data. These discoveries may reveal data quality problems (i.e., problems with the experiment or data processing pipeline), but they may also reveal totally new astrophysical phenomena: new types of galaxies or stars or whatever. That discovery potential is huge within the huge data collections that are being generated from the large astronomical sky surveys that are taking place now and will take place in the coming decades. I haven’t yet found that one special class of objects or new type of astrophysical process that will win me a Nobel Prize, but you never know what platinum-plated needles may be hiding in those data haystacks.
TAEM-What advice and support can you give to NASA for its future programs of space exploration?

KB- In the context of data science, the most important lesson is that the data generated from all NASA missions be made openly available to the research community in useful and self-explanatory ways, to facilitate new and interesting uses of the data for research and discovery. It is also helpful if these data are annotated or tagged with rich metadata, which will further enable integration and fusion of data products from multiple missions, thus enabling far greater discovery potential. One of the primary functions of metadata is to provide a short-hand condensed representation of the data product. This helps to address some of the challenges associated with Big Data: making the data more manageable and conveniently usable. We are already familiar with such hierarchical data structures – anyone who has used Google or Bing Maps experiences this – the map of the world is initially presented to you in very low resolution mode, but the resolution getting higher and higher as you drill down to some specific location – you are finally able to view your backyard or some destination at the highest resolution available from some satellite image collection, but you are definitely not viewing the whole Earth at one time at that same high resolution. Exploratory data analysis makes use of such hierarchical data structures, which NASA missions should generate for their science data users.

TAEM-What programs are you working on that would pertain to this, and what research projects are you planning on for the future?

KB- I have been working on Citizen Science projects, such as Galaxy Zoo and Zooniverse.org, in which volunteer citizens tag and annotate our scientific data products (e.g., images, or time series, or model outputs, or whatever) with descriptive characteristics. This characterization of complex data products becomes part of the metadata associated with that data product, thus enabling linkages between different data products and discovery of new patterns, trends, correlations, and behaviors. I envision extending this to other projects in the future, examining the role of non-experts in metadata generation (characterization) and conducting exploratory research into the “best” pattern recognition methods for discovering interesting, surprising, and informative features in large science data collections.

TAEM-What information can you give us so that the many students who read our publication learn more about you and the school’s programs?

KB- I invite others to check out some of the research and academic programs within Mason’s SPACS school at http://spacs.gmu.edu and to check out my own research and teaching interests at http://classweb.gmu.edu/kborne . If you are a Twitter user, you can follow me there as I actively tweet about Big Data, Data Science, and Astronomy under the handle @KirkDBorne.
TAEM-Dr. Borne, it has been a sincere honor to be able to interview you for our magazine

We have discovered that, like yourself, the faculty at George Mason University is a virtual well of information and that your school has been one of the premier sources of one of the most well trained student bodies in the nation. We want to thank you for your time and look forward to talking again with you in the very near future. TAEM

Sample Project Summary

Source: [Word](http://semanticommunity.info/?title=Data_Science/Big_Data_Science_for_Astronomy_%26_Space)

Updated: Sat, 19 Sep 2015 07:33:50 GMT

Powered by [MindTouch](http://semanticommunity.info/?title=Data_Science/Big_Data_Science_for_Astronomy_%26_Space)
gravitational lenses can assist in understanding some of the most fundamental questions about the universe, with impact far beyond the scope of this project.

2) The algorithms will be applied to the already-available SDSS data and the future LSST, but the algorithms and source code will become publicly available and will provide solutions to other large-scale surveys such as the Dark Energy Survey (DES) and Pan-STARRS, which are also among the most important scientific ventures of the decade. Since data analysis is the main bottleneck of digital sky surveys, the developed methods will allow better utilization of the power of these important projects.

3) Due to the volume and quality of SDSS and especially LSST data, it is likely that the proposed project will lead to high-impact discoveries as described in the proposal. In addition to their astrophysical importance, these discoveries and the computational methods that enable them can promote the field of Astroinformatics and increase the awareness of the Computer Science community to this emerging field.

4) The project will include the participation of graduate and undergraduate students, who will contribute to the development of the code, as well as the practical experiments and performance evaluation. This will expose graduate and undergraduate Computer Science students to interdisciplinary research, and will provide them with the opportunity to take part in Computer Science projects that aim at practical solutions to fundamental scientific questions.

Key Words: Astroinformatics; Computational Astrophysics; Digital Sky Surveys; Big data.

CV Dr. Kirk D. Borne
Source: PDF

Title

Professor of Astrophysics & Computational Science
School of Physics, Astronomy, & Computational Sciences
George Mason University, Fairfax, VA 22030
Phone: (703) 993-8402, E-mail: kborne@gmu.edu
WWW: http://classweb.gmu.edu/kborne/

Professional Preparation

1975 B.S., Physics, Summa Cum Laude, Louisiana State University
1980 M.S., Astronomy, Caltech, Pasadena
1983 Ph.D., Astronomy, Caltech, Pasadena (Advisor: James E. Gunn)
1981-83 Postdoctoral Fellow, Astronomy, University of Michigan
1983-85 Carnegie Fellow, Dept. of Terrestrial Magnetism, CIW
Appointments

2011- Professor of Astrophysics and Computational Science, GMU

2003-2011 Associate Professor of Astrophysics and Computational Science, GMU

2002-2007 Adjunct Associate Professor, UMUC Graduate School, Database Systems Technologies Program

2005-2007 Program Manager, SSDOO (Space Science Data Operations Oce) Project, QSS Group Inc., NASA/GSFC

1999 Sabbatical Visitor, STScI

1995-2002 Astrophysics Department Manager, Astrophysics Data Facility and Astronomical Data Center, Raytheon ITSS / Hughes STX, NASA/GSFC


1981-1983 Teaching Fellow, Dept. of Astronomy, University of Michigan

Related Products


Other Significant Products


---

Synergistic Activities

Senior Science Personnel, National Virtual Observatory Project

Chairman, LSST Informatics and Statistics Science Collaboration Team

Member, ISI Executive Committee for Astrostatistics

Science Organizing Committee, Conference on Intelligent Data Understanding (2010, 2011)

Science Organizing Committee, Statistical Challenges in Modern Astronomy (SCMA 2011)

---

Collaborators and Co-Editors (last 4 years)

S.Arora (UMBC), T.Axelrod (LSSTC), J.Babu (PSU), A.C.Becker (U.Wash), J.Becla (SLAC), K.Bhaduri (UMBC), T.Boroson (NOAO), K.Bowyer (UND), D.L.Burke (SLAC), D.Carr (GMU), A.Chang (MIT), C.Clauser (NOAO), A.Connolly (U.Wash), K.Cook (LLNL), D.Darg (Oxford), K.Das (SGT Inc.), M.DeMaria (NOAO), G.Djorgovski (Caltech), H.Dutta (Columbia), T.Eastman (GSFC), E.Feigelson (PSU), T.Finin (UMBC), L.Fortson (Adler Planetarium and Univ. Minnesota), P.Fox (RPI), S.Fung (NASA), H.Ferguson (STScI), J.Gentle (GMU), C.Giannella (Mitre Corporation), D.K.Gilmore (SLAC), M.Graham (Caltech), A.Gray (Georgia Tech), E.Grayzeck (NASA), J.Green (NASA), W.Grin (UMBC), L.Hall (U.South Florida), T.Hamilton (Shawnee St. U.), R.Hanisch (STScI), Z.Ivezic (U.Wash), S.Jacoby (LSSTC), R.L.Jones (U.Wash), A.Joshi (UMBC), M.Juric (IAS), S.M.Kahn (SLAC), H.Kargupta(UMBC), L.Kerschberg(GMU), K.-T.Lim (SLAC), J.Lin (GMU), C.Lintott(Oxford), Z.Liu (GMU), T.Loredo (Cornell), J.Lotz (NOAO), R.Lupton(Princeton), A.Mahabal(Caltech), T.Mahule(UMBC), T.Matheson(NOAO), T.McGlynn (NASA), R.McGuire (NASA) J.Miller (NASA), D.G.Monet (USNO-FS), T.Narock (NASA,UMBC), O.Olmedo (GMU), W.Pence (NASA), P.Pinto (U.Arizona), A.Poland (GMU), M.J.Raddick (JHU), S.Ridgway (NOAO), N.Samatova (NCSU/ORNL), A.Saha (NOAO), D.Sawyer (NASA), B.Sesar (U.Wash), R.Shaw (NOAO), E.Shaya (UMD), A.Smith (Oxford), C.W.Stubbs (CFA/Harvard), V.Sugumaran (Oakland U.), D.Sun (GMU), A.Szalay (JHU), J.A.Tyson (UCDavis), K.Wagsta (JPL), J.Wallin (GMU), H.Wechsler (GMU), E.Wegman (GMU), R.Weigel (GMU), R.Williams (Caltech), R.Wol (UMBC), R.Yang (GMU), J.Zhang (GMU), X.Zhu (UMBC)
Graduate and Postdoctoral Advisors and Advisees

Graduate students: T. Boggs, R.Dun, D.Ghoshal, C.Grieg, G.Jacobs, P.Nayak, A.Vedachalam; PhD advisor: J.Gunn (Princeton); Postdoctoral sponsors: D.Richstone (U.Michigan), V.Rubin (DTM-CIW)