The Past, Present, and Future of Data Science Education

Kirk Borne
@KirkDBorne  http://kirkborne.net
George Mason University
School of Physics, Astronomy, & Computational Sciences
Outline

• Research and Application
• My POV: Data Science Relationship to Big Data
• Data Science Programs at Mason (GMU):
  – Past, Present, and Future – PhD
  – Past and Future – BS, and an undergraduate minor
  – Future – MS professional masters degree
• Challenges and Reflections

http://kirkborne.net/
Outline

• Research and Application
  • My POV: Data Science Relationship to Big Data
  • Data Science Programs at Mason (GMU):
    – Past, Present, and Future – PhD
    – Past and Future – BS, and an undergraduate minor
    – Future – MS professional masters degree
• Challenges and Reflections
Astronomy Example

• Before we look at Data Science…

• … Let us look at an astronomy example …

• The LSST (Large Synoptic Survey Telescope)

• … Mason is a partner institution and our scientists are involved with the science, data management, and education programs of the LSST
LSST = Large Synoptic Survey Telescope

http://www.lsst.org/

8.4-meter diameter primary mirror = 10 square degrees!

(mirror funded by private donors)

Hello!
LSST = Large Synoptic Survey Telescope
http://www.lsst.org/

(mirror funded by private donors)
8.4-meter diameter primary mirror = 10 square degrees!

In the President’s budget for FY 2014

Hello!
LSST = Large Synoptic Survey Telescope
http://www.lsst.org/

8.4-meter diameter primary mirror = 10 square degrees!

- 100-200 Petabyte image archive
- 20-40 Petabyte database catalog
**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 in time and space:**
- **When?** ~2022-2032
- **Where?** Cerro Pachon, Chile

**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 Summary
http://www.lsst.org/

- 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
The LSST Data Challenges

10,000,000 events every night
100 PB image archive
50 billion object database
20 PB science catalog
Mason (GMU) is an **LSST** member institution

**Borne** is chairman of the LSST Astroinformatics and Astrostatistics research team

http://www.lsst.org/

@KirkDBorne

Architect’s design of LSST Observatory


Outline

• Research and Application

• My POV : Data Science Relationship to Big Data

• Data Science Programs at Mason (GMU):
  – Past, Present, and Future – PhD
  – Past and Future – BS, and an undergraduate minor
  – Future – MS professional masters degree

• Challenges and Reflections
What is Data Science?

It is a collection of mathematical, computational, scientific, and domain-specific methods, tools, and algorithms to be applied to Big Data for discovery, decision support, and data-to-knowledge transformation.

- Statistics
- Data Mining (Machine Learning) & Analytics (KDD)
- Data & Information Visualization
- Semantics (Natural Language Processing, Ontologies)
- Data-intensive Computing (e.g., Hadoop, Cloud, …)
- Modeling & Simulation
- Metadata for Indexing, Search, & Retrieval
- Advanced Data Management & Data Structures
- Domain-Specific Data Analysis Tools
What is Big Data?

From Wikipedia:

• Big Data refers to any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

• The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
Definitions of Big Data

From Wikipedia:

• Big Data refers to any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

• The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

My suggestion:

• Big Data refers to “Everything, Quantified and Tracked!”

• According to the standard (Wikipedia) definition, even the Ancient Romans had Big Data! That’s ridiculous!

– See my article “Today's Big Data is Not Yesterday's Big Data” at: http://bit.ly/1aXb7hD
Definitions of Big Data

From Wikipedia:
- Big Data refers to any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

My suggestion:
- **Big Data refers to “Everything, Quantified and Tracked!”**
- The challenges do not change – but their scale, scope, scariness, discovery potential do change!
- Examples:
  - Big Data Science Projects
  - Social Networks
  - IoT = Internet of Things
  - M2M = Machine-to-Machine
Data-Oriented Discovery

• Experiments can now be run against the big data collection
• Hypotheses are inferred, questions are posed, experiments are designed & run, results are analyzed, hypotheses are tested & refined!
• This is the 4\textsuperscript{th} Paradigm of Science
• This is Data Science
• This is especially (and correctly) true if the data collection is the “full” data set for a given domain:
  – astronomical sky surveys, human genome (the 1000 Genomes Project), social networks, large-scale simulations, earth observing system, ocean observatories initiative, banking, retail, national security, cybersecurity, … and the list goes on and on …
Outline

• Research and Application
• My POV: Data Science Relationship to Big Data
• **Data Science Programs at Mason (GMU):**
  – **Past, Present, and Future** – PhD
  – Past and Future – BS, and an undergraduate minor
  – Future – MS professional masters degree
• Challenges and Reflections
CSI Graduate Program at Mason

• [http://spacs.gmu.edu/content/academic-programs](http://spacs.gmu.edu/content/academic-programs)
• CSI = Computational Science & Informatics
  – CSI graduate program has existed at GMU since 1992
  – Over 200 PhD’s graduated in past 20 years
  – Approximately 95 students currently enrolled
  – About 10% of students end with M.S. (Masters in Computational Science)
  – We have a Graduate Certificate in Computational Techniques and Applications (non-degree professional certification program)
  – Note that there is no specific Data Science concentration in CSI.
  – However, students can enroll in other departments to study specific X-Informatics disciplines:
    • Geoinformatics (including Geospatial Intelligence)
    • Bioinformatics
    • Health Informatics
    • X-informatics = Application of Data Science to discipline X
CSI Graduate Program at Mason

- [http://spacs.gmu.edu/content/academic-programs](http://spacs.gmu.edu/content/academic-programs)
- CSI = Computational Science & Informatics
  - Students can choose a concentration from several choices:
    - Computational Astrophysics
    - Space Science
    - Computational Physics
    - Computational Fluid Dynamics
    - Computational Statistics
    - Computational Learning
    - Computational Mathematics
    - Computational Materials Science (Physical Chemistry)
  - A student may “create” their own concentration, such as one of these previously approved concentrations:
    - Computational Finance
    - Remote Sensing
    - Computational Economics
CSI Graduate Program at Mason

- http://spacs.gmu.edu/content/academic-programs
- CSI = Computational Science & Informatics

Students must complete 4 core courses from this set of 5:
- CSI 700 Numerical Methods
- CSI 701 Foundations of Computational Science
- CSI 702 High Performance Computing
- CSI 703 Scientific and Statistical Visualization
- CSI 710 Scientific Databases

There are also many electives (at least 5 additional CSI courses are required, plus concentration science electives).

For example:
- Data Mining, Knowledge Mining, Computational Learning, Statistical Learning, Computational Statistics, Statistical Graphics, Data Exploration, etc.
Example Course Syllabus: CSI 710 – Scientific Databases

- CSI 710 Scientific Databases (taught by K.B.) – lectures include:
  - Relational Databases: Modeling, Schemas, Normalization, SQL
  - Scientific Databases, Big Data in Science, The 4th Paradigm
  - E-Science, Ontologies, Semantic E-Science, X-Informatics
  - Distributed Data, Federated Data, Virtual Observatories
  - Citizen Science with Big Data
  - Scientific Data Mining I
  - Scientific Data Mining II
  - Astroinformatics and Astro databases
  - Bioinformatics and Bio databases
  - Geoinformatics and Geo databases
  - Health Informatics
  - Online Science (Jim Gray’s KDD-2003 lecture)
  - Intelligent Archives of the Future
Outline

• Research and Application
• My POV: Data Science Relationship to Big Data
• Data Science Programs at Mason (GMU):
  – Past, Present, and Future – PhD
  – Past and Future – BS, and an undergraduate minor
  – Future – MS professional masters degree
• Challenges and Reflections
CDS Undergraduate Program at Mason

• http://spacs.gmu.edu/content/academic-programs
• CDS = Computational and Data Sciences
  – Undergraduate B.S. degree program at Mason since 2007
  – Currently in “hiatus”, pending program modifications
  – Originally, students could choose the general CDS degree, or else choose one of these concentrations:
    • Physics
    • Chemistry
    • Biology
  – A student could “create” their own concentration. For example:
    • Environmental Science
  – Anticipated modifications to the program – only 2 emphasis areas (maximizing student’s ability to “create” their own domain-specific course of study, in addition to a small set of required courses):
    • Modeling and Simulation
    • Data Science
CDS Undergraduate Program at Mason

- [http://spacs.gmu.edu/content/academic-programs](http://spacs.gmu.edu/content/academic-programs)
- CDS = Computational and Data Sciences
  - The DATA SCIENCE component of the curriculum was developed with the support of a grant (2007) from the NSF (National Science Foundation):
    - CUPIDS = *Curriculum for an Undergraduate Program In Data Sciences*
  - **Primary Goal**: *to increase student’s understanding of the role that data plays across the sciences as well as to increase the student’s ability to use the technologies associated with data acquisition, mining, analysis, and visualization.*
  - **Objectives** – students are trained:
    - … to access large distributed data repositories
    - … to conduct meaningful inquiries into the data
    - … to mine, visualize, and analyze the data
    - … to make objective data-driven inferences, discoveries, and decisions
CDS Undergraduate Program at Mason

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences

Core courses that students can choose from:
- CDS 101 – Introduction to Computational Data Sciences
- CDS 130 – Computing for Scientists
- CDS 251 – Introduction to Scientific Programming
- CDS 301 – Scientific Information and Data Visualization
- CDS 302 – Scientific Data and Databases
- CDS 401 – Scientific Data Mining
- CDS 410 – Modeling and Simulations I
- CDS 411 – Modeling and Simulations II

Additional required courses include Math, Statistics, Computer Science, Physics I and II, plus courses in student’s chosen science concentration
CDS Undergraduate Program at Mason

- http://spacs.gmu.edu/content/academic-programs
- CDS = Computational and Data Sciences

We are also “infiltrating” the entire undergraduate program at Mason through 3 of our courses that satisfy university General Education graduation requirements for all students at the university:
  - CDS 101 – Introduction to Computational Data Sciences
    - Satisfies GMU’s Natural Science requirement
  - CDS 130 – Computing for Scientists
    - Satisfies GMU’s I.T. requirement
  - CDS 151 – Data Ethics
    - Satisfies GMU’s Ethics requirement

- CDS undergraduate minor (open to any student at the university)
  - 12 CDS credits (focus on Data Science or Modeling & Simulation)
  - …plus one additional science class
Example of Learning Objectives:
*CDS 401 – Scientific Data Mining*

- Be able to explain the role of data mining within scientific knowledge discovery.
- Be able to describe the most well known data mining algorithms and correctly use data mining terminology.
- Be able to express the application of statistics, similarity measures, and indexing to data mining tasks.
- Identify appropriate techniques for classification and clustering applications.
- Determine approaches used for mining large scientific databases (e.g., genomics, virtual observatories).
- Recognize techniques used for spatial and temporal data mining applications.
- Express the steps in a data mining project (e.g., cleaning, transforming, indexing, mining, analysis).
- Analyze classic data mining examples and use cases, and assess the applicatio of different data mining techniques.
- Effectively prepare data for mining.
- Effectively use software packages for data exploration, visualization, and mining.
Outline

• Research and Application
• My POV: Data Science Relationship to Big Data

• **Data Science Programs at Mason (GMU):**
  – Past, Present, and Future – PhD
  – Past and Future – BS, and an undergraduate minor
  – Future – MS professional masters degree

• Challenges and Reflections
This will take effect in Fall 2014:  http://goo.gl/0L4gNo

• This program will be a restructuring of our existing MS in Computational Science (i.e., what our website says today is not the way the program will look starting in Fall 2014)

• The changes include:
  – focus on serving the professional community (not the academic research community)
  – focus on meeting workforce skills demands (especially for Data Scientists)
  – deployment of 3 new Areas of Emphasis that our “customers” are demanding:
    • Data Science
    • Transportation Safety  (associated with the new National Center for Collision Safety and Analysis within our school)
    • Modeling and Simulation
Outline

• Research and Application
• My POV : Data Science Relationship to Big Data
• Data Science Programs at Mason (GMU):
  – Past, Present, and Future – PhD
  – Past and Future – BS, and an undergraduate minor
  – Future – MS professional masters degree

• Challenges and Reflections
Challenges and Reflections

- **Attracting students**: after we started, we realized that no student ever says “I want to be a data scientist when I grow up.” (… not true any more!)
- **Visibility**: most other science departments are not aware of the importance of our courses for their majors. (Note: Biology and Neuroscience now require our Science Computing course.)
- **Scientific computing course**: this was identified 3 years ago as a necessary course to attract students … we now have this course, and it is very popular (nearly 200 students each semester, and growing…)
- **Program evolution**: our journey (PhD to BS to MS) reflects the way the “Big Data” and “Data Science” world has evolved – from academic research, to general education, to meeting workforce demands here and now, but then it will come back to the non-graduate program focus ...
- **Future expectations for Data Science Education**: the general education focus will become essential, spreading to K-12 (eventually)
Data Science Education: 2 Perspectives

• **Data Science in Education** – introduce data in all learning settings:
  • Informatics (Data Science) enables transparent reuse and analysis of data in inquiry-based classroom learning.
  • Learning is enhanced when students work with real data and information (especially online data) that are related to the topic (any topic) being studied.
  • http://serc.carleton.edu/usingdata/ (“Using Data in the Classroom”)
  • http://www.oceansofdata.org/ (EDC’s Oceans of Data Institute)

• **An Education in Data Science** – students are specifically trained:
  • … to access large distributed data repositories
  • … to conduct meaningful inquiries into the data
  • … to mine, visualize, and analyze the data
  • … to make objective data-driven inferences, discoveries, and decisions
There are many programs now...

See Data Informed’s Map of University Programs in Big Data Analytics at:
http://data-informed.com/bigdata_university_map/
The KirkDBorne Ultimatum**

Data Literacy for all!

[ **no connection to Robert Ludlum’s “Bourne Ultimatum” ]

http://kirkborne.net/