Big Data, Enormous Opportunity

Ed Lazowska
Bill & Melinda Gates Chair in Computer Science & Engineering
University of Washington

The 27th Elliott Organick Memorial Lectures
University of Utah

April 2014
Today

- What’s all the fuss about?
- Jim Gray’s “fourth paradigm”: smart discovery / data-intensive discovery / eScience
- My personal story, and the story of the UW eScience Institute
- Three science examples: survey astronomy, environmental metagenomics, neuroscience
- The NYU / Berkeley / UW “Data Science Environments” project
- Entrepreneurial potential
- Some non-science examples
What is “big data”? 

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

(Dan Ariely)
Exponential improvements in technology and algorithms are enabling the “big data” revolution

• A proliferation of sensors
  – Think about the sensors on your phone

• More generally, the creation of almost all information in digital form
  – It doesn’t need to be transcribed in order to be processed

• Dramatic cost reductions in storage
  – You can afford to keep all the data

• Dramatic increases in network bandwidth
  – You can move the data to where it’s needed
• Dramatic cost reductions and scalability improvements in computation
  – With Amazon Web Services, or Google App Engine, or Microsoft Azure, 1000 computers for 1 day costs the same as 1 computer for 1000 days
• Dramatic algorithmic breakthroughs
  – Machine learning, data mining – fundamental advances in computer science and statistics
• Ever more powerful models producing ever-increasing volumes of data that must be analyzed
The “big data” revolution is what actually puts the “smarts” in “smart everything”

- Smart homes
- Smart cars
- Smart health
- Smart robots
- Smart crowds and human-computer systems
- Smart interaction (virtual and augmented reality)
- Smart discovery (exploiting the data deluge)
Smart discovery / data-intensive discovery / eScience

Jim Gray
Microsoft Research
Observation

Experiment

Theory

Credit: John Delaney, University of Washington
Observation

Experiment

Theory
Observation

Experiment

Theory
Observation
Experiment
Theory
Computational Science
Nearly every field of discovery is transitioning from “data-poor” to “data-rich”

- Massive volumes of data from sensors and networks of sensors

Apache Point telescope, SDSS

80TB of raw image data (80,000,000,000,000 bytes) over a 7 year period
Large Synoptic Survey Telescope (LSST)

15TB/day
(2 SDSS’s each week),
100+PB in its 10-year lifetime

400mbps sustained data rate between Chile and NCSA
Large Hadron Collider

700MB of data per second, 60TB/day, 20PB/year
Illumina HiSeq 2000 Sequencer
~1TB/day

Major labs have 25-100 of these machines
Regional Scale Nodes of the NSF Ocean Observatories Initiative

1000 km of fiber optic cable on the seafloor, connecting thousands of chemical, physical, and biological sensors
The Web

~1.2B Facebook users

~~750M websites

~~~200B web pages
Point-of-sale terminals
eScience is about the *analysis* of data

- The automated or semi-automated extraction of knowledge from massive volumes of data
  - There’s simply too much of it – and it’s too complex – to explore manually

- It’s not just a matter of volume – it’s “the 3 V’s”:
  - Volume
  - Velocity (rate)
  - Variety (dimensionality / complexity)
eScience utilizes a spectrum of computer science techniques and technologies

- Sensors and sensor networks
- Backbone networks
- Databases
- Data mining
- Machine learning
- Data visualization
- Cluster computing at enormous scale (the cloud)
- Collaboration and crowd sourcing
eScience will be pervasive

• Simulation-oriented computational science has been transformational, but – honestly – it has been a niche
  – As an institution (e.g., a university), you didn’t need to excel in order to be competitive

• eScience capabilities must be broadly available in any institution
  – If not, the institution will simply cease to be competitive
“From data to knowledge to action”

• The ability to extract knowledge from large, heterogeneous, noisy datasets – to move “from data to knowledge to action” – lies at the heart of 21st century discovery
• To remain at the forefront, researchers in all fields will need access to state-of-the-art eScience methodologies and tools
• These methodologies and tools will need to advance rapidly, driven by the requirements of discovery
• eScience is driven more by intellectual infrastructure (human capital) and software infrastructure (shared tools and services – digital capital) than by hardware
My personal story, and the story of the UW eScience Institute

Early 1980s

Late 1990s
UW eScience Institute

• “All across our campus, the process of discovery will increasingly rely on researchers’ ability to extract knowledge from vast amounts of data... In order to remain at the forefront, UW must be a leader in advancing these techniques and technologies, and in making [them] accessible to researchers in the broadest imaginable range of fields.”
This was not as broadly obvious in 2005 as it is today

• But we asked UW’s leading faculty, and they told us!
  – *From the get-go, this has been a bottom-up, needs-based, driven-by-the-scientists effort!*
Strategies

• “Long tail”

• “Flip the influentials”
Multiple modes of interaction, multiple time scales

Communication (events)

Incubation (projects)

Collaboration (partnerships)

1-2 weeks and down

1-2 quarters

1-2 years and up
• Focus on tools, but recognize and avoid the common failure modes of cyberinfrastructure projects

- Reactive, ad hoc, one-off, “hero” efforts
- Address one application
- No leverage; doesn’t scale

- “Uber-system”
- Over-abstraction
- Tries to meet so many needs, it winds up meeting none well

The sweet spot: bottom-up, needs-based, driven-by-the-scientists … and “just general enough” to achieve leverage
• A variety of individuals ... a variety of careers and career paths
  – Faculty
  – Research Scientists
  – Software Professionals
  – Postdocs
  – Graduate and Undergraduate Students

  translation
  robustness
  the next generation – the real agents of cultural change
• On the methodology side, seek faculty in “Pasteur’s Quadrant”
• Across-the-board, strive to create “Pi-shaped” scholars
• Resurrect the water cooler!
AstroDB: Cosmology at Scale

Andrew Connolly (Astronomy)
Magda Balazinska (CSE)
• In cosmology there is a growing tension between theory and data
  – Universe is made up of dark energy (68%), dark matter (27%), and other stuff (5%)
  – The physics of dark energy is unknown and there are no firm detections of dark matter particles
  – We will provoke this tension through observations and large scale surveys (as the signals are small)
The Large Synoptic Survey Telescope

- Survey half the sky every 3 nights (1000-fold increase in data vs. Sloan Digital Sky Survey)
- Enabled by a 3.2 Gigapixel camera with a 3.5 degree field
- 15 TB/night (100 PB over 10 years), 20 billion objects, and 20 trillion measurements
How do we do science at petabyte scale?

Science questions ...

• Finding the unusual
  – Supernova, GRBs
  – Probes of Dark Energy

• Finding moving sources
  – Asteroids and comets
  – Origins of the solar system

• Mapping the Milky Way
  – Tidal streams
  – Probes of Dark Matter

• Measuring shapes of galaxies
  – Gravitational lensing
  – The nature of Dark Energy
How do we do science at petabyte scale?

Science questions … map to computational questions

- Finding the unusual
  - Supernova, GRBs
  - Probes of Dark Energy
- Finding moving sources
  - Asteroids and comets
  - Origins of the solar system
- Mapping the Milky Way
  - Tidal streams
  - Probes of Dark Matter
- Measuring shapes of galaxies
  - Gravitational lensing
  - The nature of Dark Energy
- Finding the unusual
  - Anomaly detection
  - Density estimations
- Finding moving sources
  - Tracking algorithms
  - Kalman filters
- Mapping the Milky Way
  - Clustering techniques
  - Correlation functions
- Measuring shapes of galaxies
  - Image processing
  - Data intensive analysis
Role of microbes in marine ecosystems

Ginger Armbrust (Oceanography)
Bill Howe (Computer Science & Engineering + eScience Institute)

Microbial community visualized with DNA stain
Challenges:
1) Integration across different data types
2) Distributed and remote labs
SQLShare: Database-as-a-Service for Science

SQLShare is a database service aimed at removing the obstacles to using relational databases: installation, configuration, schema design, tuning, data ingest, and even application design. You simply upload your data and immediately start querying it.
Integrating across physics, biology, and chemistry

Query across data sets in real-time
“not just faster...different!”

Dan Halperin,
Research Scientist, eScience Institute

Konstantin Weitz
Graduate student, CSE
Connecting across distributed labs

SeaFlow instrument
  ↓
Ship computer
  ↓
Processed data
  ↓
am automated
Cloud computer
  ↓
SQLShare
  ↓
Web display
  ↓
Collaborator computers

Other ship data streams

Completely automated

Email ship

August, 2013
Devices + Neuroscience + Data Science

Tom Daniel (Biology)

How do natural systems make decisions?

How do they manage massive data flow?
What features do animals extract to solve problems?

Complex environments

Neural activity

How is information synthesized to drive decisions?

Motor activity

Behavioral output

How does action affect subsequent sensation?

How do muscles work together to perform actions?
These scientists are involved because their science can only succeed if there is a major cultural shift within universities and a major change in the way we approach discovery.
## Faculty core team

### Data science methodology

<table>
<thead>
<tr>
<th>Name</th>
<th>Position/Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cecilia Aragon</td>
<td>Human Centered Design &amp; Engr.</td>
</tr>
<tr>
<td>Magda Balazinska</td>
<td>Computer Science &amp; Engineering</td>
</tr>
<tr>
<td>Emily Fox</td>
<td>Statistics</td>
</tr>
<tr>
<td>Carlos Guestrin</td>
<td>CSE</td>
</tr>
<tr>
<td>Bill Howe</td>
<td>CSE</td>
</tr>
<tr>
<td>Jeff Heer</td>
<td>CSE</td>
</tr>
<tr>
<td>Ed Lazowska</td>
<td>CSE</td>
</tr>
</tbody>
</table>

### Biological sciences

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom Daniel</td>
<td>Biology</td>
</tr>
<tr>
<td>Bill Noble</td>
<td>Genome Sciences</td>
</tr>
<tr>
<td>Randy LeVeque</td>
<td>Applied Mathematics</td>
</tr>
<tr>
<td>Werner Stuetze</td>
<td>Statistics</td>
</tr>
</tbody>
</table>

### Physical sciences

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andy Connolly</td>
<td>Astronomy</td>
</tr>
<tr>
<td>John Vidale</td>
<td>Earth &amp; Space Sciences</td>
</tr>
<tr>
<td>Josh Blumenstock</td>
<td>iSchool</td>
</tr>
<tr>
<td>Mark Ellis</td>
<td>Geography</td>
</tr>
<tr>
<td>Tyler McCormick</td>
<td>Sociology, CSSS</td>
</tr>
<tr>
<td>Thomas Richardson</td>
<td>Statistics, CSSS</td>
</tr>
</tbody>
</table>

### Environmental sciences

<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ginger Armbrust</td>
<td>Oceanography</td>
</tr>
<tr>
<td>Bill Noble</td>
<td>Genome Sciences</td>
</tr>
<tr>
<td>Mark Ellis</td>
<td>Geography</td>
</tr>
<tr>
<td>Tyler McCormick</td>
<td>Sociology, CSSS</td>
</tr>
<tr>
<td>Thomas Richardson</td>
<td>Statistics, CSSS</td>
</tr>
</tbody>
</table>
Faculty core team

Data science methodology
Cecilia Aragon
Human Centered Design & Engr.
Magda Balazinska
Computer Science & Engineering
Emily Fox
Statistics
Carlos Guestrin
CSE
Bill Howe
CSE
Jeff Heer
CSE
Ed Lazowska
CSE

Biological sciences
Tom Daniel
Biology
Bill Noble
Genome Sciences

Physical sciences
Andy Connolly
Astronomy
John Dale
Earth & Space Sciences

Environmental sciences
Ginger Armbrust
Oceanography

Social sciences
Josh Blumenstock
iSchool
Mark Ellis
Geography
Tyler McCormick
Sociology, CSSS

Randy LeVeque
Applied Mathematics
Werner Stuetzle
Statistics

12 Departments / Colleges

5 Schools / Colleges
A 5-year, $37.8 million cross-institutional collaboration
Goals

• **Do** breakthrough science
  – In Scientific Theme Areas
  – In Data Science Methodology areas

• **Enable** breakthrough science
  – Through new tools and methods
  – Through changing the process of discovery and driving cultural changes

• **Establish a “virtuous cycle”**
UW “Flagship Activities”

- Establish two new roles: Data Science Fellows and Data Scientists
- Establish a new graduate program in data science (NSF IGERT)
- Establish an “Incubator” seed grant program
- Establish a campus-wide community around reproducible research
- Establish a “Data Science Studio”
- Establish a research program in “the data science of data science”

Each of these is essential

None of these has been possible
The rising tide that lifts all boats

- PIs on major proposals
- + eScience Institute Steering Committee
- + Participants in February 7 Campus-Wide Data Science poster session
Commercial Uptake of Research

Project: Intelligent systems to transform, clean and integrate data without programming (Jeff Heer)

Now commercialized via Trifacta, a venture-backed company that has raised over $16M

Project: Novel languages for creating expressive and effective data visualizations (Jeff Heer)

Data-Driven Documents (D3.js) now the de facto standard for web-based visualization. Used by The New York Times, Square, and hundreds of others
Project: Huge-scale machine learning accessible to all (Carlos Guestrin)

Now open-sourced via **GraphLab.org** and commercialized via **GraphLab.com**

Project: Database-as-a-service for open data analytics (Bill Howe)

**SQLShare** – widely-used freeware
Non-science examples of “big data in action”

- Collaborative filtering
• Fraud detection
Secret government surveillance of American citizens

The New York Times

Drug Agents Use Vast Phone Trove, Eclipsing N.S.A.’s

By SCOTT SHANE and COLIN MOYNIHAN
Published: September 1, 2013 | 285 Comments

For at least six years, law enforcement officials working on a counternarcotics program have had routine access, using subpoenas, to an enormous AT&T database that contains the records of decades of Americans’ phone calls — parallel to but covering a far longer time than the National Security Agency’s hotly disputed collection of phone call logs.

The Hemisphere Project, a partnership between federal and local drug officials and AT&T that has not previously been reported, involves an extremely close association between the government and the telecommunications giant.

“Hemisphere Project”
• 26 years of records of every call that passed through an AT&T switch
• New records added at a rate of 4B/day
• Price prediction
Hospital re-admission prediction

Background
With payers implementing penalties for readmissions, it's critical to start addressing readmission risks today. Solutions that help healthcare enterprises understand last month's readmissions are no longer sufficient. Payers need to know which patients in their hospitals today are at risk for being readmitted within 30 days of discharge, so they can take action and address those risks before the patient walk out the door.

Amalga helps healthcare organizations proactively identify inpatient and Emergency Department (ED) patients at risk for readmissions and help them take action to avoid preventable readmissions, reduce costs, and deliver higher quality care – today and tomorrow.

Overview
By using predictive modeling technologies, Amalga can help reduce preventable readmissions by enabling healthcare delivery organizations to:

- Effectively define and monitor patient groups across the enterprise
- Use data collected in Amalga to predict readmission probability based on a given hospital's historical data
- Proactively manage at-risk patients throughout their stay and at discharge
- Access patterns in key indicators to identify and address root causes of readmissions

Features and Benefits
With Amalga, organizations can:

- Actively identify and track patient groups
- Integrate disparate systems and identify patient cohorts based on key characteristics
- Use predictive modeling technologies to help identify patients at risk for readmission
- Analyze readmission patterns and monitor 90-day inpatient and 72-hour ED readmissions
- View simplified reports that support identification and addressing of root causes sooner

Avoid preventable readmissions, reduce costs and deliver higher quality care
• Travel time prediction under specific circumstances
• Coaching / play calling in all sports
• Speech recognition

Siri.  Beta
Your wish is its command.

Siri lets you use your voice to send messages, schedule meetings, place phone calls, and more. Ask Siri to do things just by talking the way you talk. Siri is so easy to use and does so much, you’ll keep finding more and more ways to use it.
• **Machine translation**
  
  – Speech -> text
  
  – Text -> text translation
  
  – Text -> speech in speaker’s voice

http://www.youtube.com/watch?v=Nu-nlQqFCKg&t=7m30s
7:30 – 8:40
• Presidential campaigning

Inside the Secret World of the Data Crunchers Who Helped Obama Win

By Michael Scherer | Nov 07, 2012 | 268 Comments
• Electoral forecasting

DATA MINING

Nate Silver’s Sweep Is a Huge Win for ‘Big Data’

The data utopia awaits.
By Nitasha Tiku 11/07 11:10am
We’re at the dawn of a revolutionary new era of discovery and of learning