Data and Society

Lecture 4: Data and Science

2/10/17

Earthquakes, LHC, Data Science
Announcements 2/10

• Wednesday class February 15, starts at 8 a.m.

  – Plan to return draft on February 17

• Paper instructions (draft due on March 15) today after lecture.

• Discussion article for next week (Friday, February 17). Please Read:
Today (2/10/17)

• Lecture 4: Data and Science
  – Geoscience
  – High Energy Physics
  – Data Science

• Paper Instructions

• Break

• 4 Student Presentations
<table>
<thead>
<tr>
<th>Wednesday Section</th>
<th>Friday Lecture</th>
<th>First Half of Class</th>
<th>Second Half of Class</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 25: NO class</td>
<td>January 27</td>
<td>L2: Big data applications / Data and the election; Data and Target; Discussion</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>February 1: 6 presentations</td>
<td>February 3</td>
<td>L3: Data and Health / PDB, Precision Medicine; Discussion</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>February 8: NO class</td>
<td>February 10</td>
<td>L4: Data and Science / Earthquakes, LHC; Paper Instructions</td>
<td>4 Presentations</td>
<td>Op-Ed Draft Due</td>
</tr>
<tr>
<td>February 15: 6 presentations</td>
<td>February 17</td>
<td>L5: Data Cyberinfrastructure; Discussion</td>
<td>4 Presentations</td>
<td>Op-Ed Draft Due</td>
</tr>
<tr>
<td>February 22: 6 presentations</td>
<td>February 24</td>
<td>L6: Data Stewardship and Data Preservation; Discussion</td>
<td>4 Presentations</td>
<td>Op-Ed Final Due</td>
</tr>
<tr>
<td>March 1: NO class</td>
<td>March 3</td>
<td>NO class</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>March 8: 6 presentations</td>
<td>March 10</td>
<td>L7: Data Futures – Internet of Things; Discussion</td>
<td>4 Presentations</td>
<td>Paper Draft Due</td>
</tr>
<tr>
<td>March 15: Spring Break</td>
<td>March 17</td>
<td>Spring Break</td>
<td>4 presentations</td>
<td></td>
</tr>
<tr>
<td>March 22: NO class</td>
<td>March 24</td>
<td>L8: Data rights and policy / U.S. and EU; Discussion</td>
<td>4 presentations</td>
<td></td>
</tr>
<tr>
<td>March 29: 6 presentations</td>
<td>March 31</td>
<td>Op-Ed Pecha-Kucha</td>
<td>4 presentations</td>
<td>Paper Draft Due</td>
</tr>
<tr>
<td>April 5: NO class</td>
<td>April 7</td>
<td>NO class</td>
<td>4 presentations</td>
<td></td>
</tr>
<tr>
<td>April 12: 4 presentations</td>
<td>April 14</td>
<td>Hilary Mason Guest Lecture</td>
<td>4 presentations</td>
<td>Final Paper Due</td>
</tr>
<tr>
<td>April 19: 4 presentations</td>
<td>April 21</td>
<td>L9: Data and Ethics; Discussion</td>
<td>4 presentations</td>
<td></td>
</tr>
<tr>
<td>April 26: 6 presentations</td>
<td>April 28</td>
<td>Paper Pecha-Kucha</td>
<td>4 presentations</td>
<td></td>
</tr>
</tbody>
</table>
Lecture 3: Data and Science
Modeling, simulation, analysis a critical tool in addressing science and societal challenges

What is the potential impact of Global Warming?

How will natural disasters effect urban centers?

What plants work best for biofuels?

Can we accurately predict market outcomes?

Is there life on other planets?

What therapies can be used to cure or control cancer?

Fran Berman, Data and Society, CSCI 4370/6370
Data paired with other technologies to enable scientific discovery

Protein analysis and modeling of function and structures

Storage and Analysis of Data from the CERN Large Hadron Collider

Real-time disaster response

Big data, Big compute

Big data, Big BW

Data (more BYTES)

ComputE (more FLOPS)

NETWORK (more BW)

Cosmology

Development of biofuels

Seti@home, MilkyWay@Home, BOINC

Fran Berman, Data and Society, CSCI 4370/6370
Data-driven Science

• Earthquake Simulations

• Large Hadron Collider Data Analysis
Earthquake Simulation

Some information in this section courtesy of Amit Chourasia and Yifeng Cui, SDSC and SDSU
Earthquake Simulation

Background:

- Earth constantly evolving through the movement of “plates”

- Using plate tectonics, the Earth's outer shell (lithosphere) is posited to consist of seven large and many smaller moving plates.

- As the plates move, their boundaries collide, spread apart or slide past one another, resulting in geological processes such as earthquakes and tsunamis, volcanoes and the development of mountains, typically at plate boundaries.
Why Earthquake Simulations are Important

• If we understand how earthquakes can happen, we can
  – Predict which places might be hardest hit
  – Reinforce bridges and buildings to increase safety
  – Prepare police, fire fighters and doctors in high-risk areas to increase their effectiveness

• Information technologies drive more accurate earthquake simulation
What would be the impact of an earthquake on the lower San Andreas fault?
Simulation decomposition strategy leverages parallel high performance computers

- Southern California partitioned into “cubes” then mapped onto processors of high performance computer
- Data choreography used to move data in and out of memory during processing

Builds on data and models from the Southern California Earthquake Center, Kinematic source (from Denali) focuses on Cajon Creek to Bombay Beach
TeraShake Simulation (2000’s)

Simulation of Southern of 7.7 earthquake on lower San Andreas Fault

- Physics-based dynamic source model – Anelastic Wave Propagation Model code -- simulation of mesh of 1.8 billion cubes with spatial resolution of 200 m

- Simulated first 3 minutes of a magnitude 7.7 earthquake, 22,728 time steps of 0.011 second each

- Simulation for TeraShake 1 and 2 simulations generated 45+ TB data

Fran Berman, Data and Society, CSCI 4370/6370
Under the surface

TeraShake2

SDSC

Fran Berman, Data and Society, CSCI 4370/6370
Resources must support a complicated orchestration of computation and data movement.

240 procs on SDSC Datastar, 5 days, 1 TB of main memory

Continuous I/O 2GB/sec

47 TB output data for 1.8 billion grid points

Parallel file system

Data parking

“Fat Nodes” with 256 GB of DataStar used for pre-processing and post visualization

10-20 TB data archived a day

Finer resolution simulations require even more resources.
TeraShake scaled to run on petascale architectures
Terashake and Data

**Data Management**
- 10 Terabytes moved per day during execution over 5 days
- Derived data products registered into SCEC digital library (total SCEC library had 168 TB)

**Data Post-processing:**
- *Movies* of seismic wave propagation
- Seismogram formatting for interactive on-line analysis
- *Derived data:*
  - Velocity magnitude
  - Displacement vector field
  - Cumulative peak maps
  - Statistics used in visualizations

---

**TeraShake Resources**

**Computers and Systems**
- 80,000 hours on IBM Power 4 (DataStar)
- 256 GB memory p690 used for testing, p655s used for production run, TeraGrid used for porting
- 30 TB Global Parallel file GPFS
- Run-time 100 MB/s data transfer from GPFS to SAM-QFS
- 27,000 hours post-processing for high resolution rendering

**People**
- 20+ people for IT support
- 20+ people in domain research

**Storage**
- SAM-QFS archival storage
- HPSS backup
- SRB Collection with 1,000,000 files
Application Evolution

- TeraShake → PetaSHA, PetaShake, CyberShake, M8, etc. at SCEC and SDSC

- Evolving applications improving
  - Resolution
  - Models and algorithms
  - Simulation accuracy
  - Features
  - Workflow and simulation efficiency and performance, etc.

- **M8**: Regional scale wave propagation simulation using realistic 3D earth model, dynamic earthquake source, simulating ground motions at frequencies of interest to building engineers

Image: https://scec.usc.edu/scecpedia/M8
Application Evolution

• Recent 7.8M simulation of M8 won the Nvidia Global Impact award in 2015:  
  [link](http://ucsdnews.ucsd.edu/pressrelease/sdsc_researchers_win_nvidias_2015_global_impact_award/)
  - *Seismogram* = Record written by a seismograph in response to ground motions produced by an earthquake, explosion, or other ground-motion sources. [USGS]
  - *Blue Waters* → 13 PF peak performance
  
• More cool visualizations:  
  [link](http://visservices.sdsc.edu/projects/scec/m8/1.0/)

<table>
<thead>
<tr>
<th></th>
<th>Terashake (2000’s)</th>
<th>M8 (2010’s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Magnitude</strong></td>
<td>7.7</td>
<td>8</td>
</tr>
<tr>
<td><strong>Inner Scale</strong></td>
<td>200m</td>
<td>40m</td>
</tr>
<tr>
<td><strong>Resolution of terrain grid</strong></td>
<td>1.8 billion mesh points</td>
<td>436 billion mesh points</td>
</tr>
<tr>
<td><strong>Time steps</strong></td>
<td>20,000 (.012 sec/step)</td>
<td>160,000 (.0023 sec/step)</td>
</tr>
<tr>
<td><strong>Cores / TF</strong></td>
<td>240 processors / DataStar (&lt;= 10 TF)</td>
<td>18.6K nodes / 220 TF (sustained)</td>
</tr>
<tr>
<td><strong>Runtime</strong></td>
<td>18,000 CPU hours / 4 days</td>
<td>5.3M CPU hours / 24 hours</td>
</tr>
</tbody>
</table>
SCEC 2017-2021 Science Priorities

The strategic framework for the SCEC5 Science Plan is cast in the form of five basic questions of earthquake science:

1. How are faults loaded on different temporal and spatial scales?
2. What is the role of off-fault inelastic deformation on strain accumulation, dynamic rupture, and radiated seismic energy?
3. How do the evolving structure, composition and physical properties of fault zones and surrounding rock affect shear resistance to seismic and aseismic slip?
4. How do strong ground motions depend on the complexities and nonlinearities of dynamic earthquake systems?
5. In what ways can system-specific studies enhance the general understanding of earthquake predictability?

These questions cover the key issues driving earthquake research in California, and they provide a basis for gauging the intellectual merit of proposed SCEC5 research activities.

(From https://www.scec.org/research/scec5_objectives.html)
Large Hadron Collider Data Analysis
The Large Hadron Collider (LHC)

- LHC is the world’s most powerful particle collider.
- LHC’s goal is to allow physicists to test the predictions of different theories of particle physics, high-energy physics, (in particular the properties of the Higgs Boson) and the large family of new particles predicted by supersymmetric theories.
- LHC contains seven detectors, each designed for a different kind of research. LHC built near Geneva between 1998 and 2008 in collaboration with over 10,000 scientists and engineers from over 100 countries.
- LHC lies in a 17 mile circumference tunnel beneath the France-Switzerland border.
- LHC collisions produce 10’s of PBs of data per year.
  - Subset of data analyzed by distributed grid of 170+ computers in 36 countries

A collider is a type of a particle accelerator with two directed beams of particles.

In particle physics colliders are used as a research tool: they accelerate particles to very high kinetic energies and let them impact other particles.

**Analysis of the byproducts of these collisions gives scientists good evidence of the structure of the subatomic world and the laws of nature governing it.**

Many of these byproducts are produced only by high energy collisions, and they decay after very short periods of time. Thus many of them are hard or near impossible to study in other ways.
What happens at CERN?

- Accelerators create particle collisions
  - Protons circulate at close to the speed of light
  - 10’s of millions of collisions every second
  - Collisions recreate the conditions of the first moments of the universe

- Detectors study collisions and the thousands of particles emerging from them.

- Worldwide network of computers filters, records and processes the data from the collisions
  - LHC computing grid processes PBs of data each year

- Physicists throughout the world analyze the data

Information from http://home.cern/
Higgs and Beyond

“A major goal of Run1 of the LHC was to find evidence of the Higgs boson…”

Its discovery was announced on July 4 2012.

After a prolonged downtime to prepare for running at almost twice the energy of Run1, the LHC restarted in 2015 and will run (mainly during the summer months) for another 3 years before yet another upgrade.

Major goals of Run2 are to search “beyond the standard model”, including searches for Dark Energy and Dark Matter.

Information from Jamie Shiers, CERN
Worldwide LHC Computing Grid

This map shows registered WLCG sites currently in operation.

Tier 2 sites
- AT | HEPHY-UIBK
- AT | Hephy-Vienna
- AU | Australia-ATLAS
- BE | BEgrid-ULB-VUB

... 145 more

Tier-0 sites
- CH | CERN Data Centre, Tier-0
- HU | Wigner Research Centre for Physics, T...

Tier-1 sites
- CA | TRIUMF-LCG2
- DE | FZK-LCG2
- ES | PIC
- FR | IN2P3-CC

... 9 more

Image from http://wlcg.web.cern.ch/
LHC – Stewardship and Preservation Challenges

• Significant volumes of high energy physics data are thrown away “at birth” – i.e. via very strict filters (aka triggers) before writing to storage. To a first approximation, all remaining data needs to be preserved for a few decades.
  
  – LHC data particularly valuable as reproducibility of experiments is tremendously expensive and almost impossible to achieve

• Tier 0 and 1 sites currently provide bit preservation at scale
  
  – Data more usable and accessible when services coupled with bit preservation
  
  – In the process of “self certification” according to ISO 16363 of the Tier0 and Tier1 sites.

• CERN also developing an advanced Data Management Plan that will be updated roughly annually based on “an intelligent super-set” of the Horizon 2020, DoE and NSF guidelines – with a few of our own for good measure.

*Slide adapted from Jamie Shiers, CERN 2016*
Post-collision

After the collisions have stopped

> Finish the analyses! But then what do you do with the data?
  - Until recently, there was no clear policy on this in the HEP community
  - It’s possible that older HEP experiments have in fact simply lost the data

> Data preservation, including long term access, is generally not part of the planning, software design or budget of an experiment
  - So far, HEP data preservation initiatives have been in the main not planned by the original collaborations, but rather the effort a few knowledgeable people

> The conservation of tapes is not equivalent to data preservation!
  - “We cannot ensure data is stored in file formats appropriate for long term preservation”
  - “The software for exploiting the data is under the control of the experiments”
  - “We are sure most of the data are not easily accessible!”

*Slide adapted from Jamie Shiers, CERN 2016*
Data: Outlook for HL-LHC

- The LHC – including all foreseen upgrades – will run until circa 2040. By that time between **10 and 100 EB** of data will have been gathered.

- These data (the uninteresting stuff has already been discarded) should be preserved for a number of decades.

- Very rough estimate of a new RAW data per year of running using a simple extrapolation of current data volume scaled by the output rates.
  - To be added: derived data (ESD, AOD), simulation, user data...
  - **At least 0.5 EB / year (x 10 years of data taking)**
Data Science – an emerging field

- [Wikipedia] **Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).

*Slide adapted from Rob Rutenbar, UIUC*
Evolving Data Science:
How do we develop a roadmap for data science?

• Research
  – Who is doing data science research? Who is doing research enabled by data science? How can this be fostered/facilitated? What are the programmatic needs?

• Curriculum & pedagogy
  – Who teaches foundations? Where? To whom?

• Infrastructure
  – What is needed to support data sci?

• Applications
  – How do we apply data science to drive new discovery and innovation?

• Verticals
  – How will data science disrupt the end-to-end computer stack

• Ethics & society
  – What informs use and collection of data? What are its social impacts?

Fran Berman, Data and Society, CSCI 4370/6370
Data science research: Many challenges in optimizing the potential of the data analysis pipeline [big data whitepaper]

Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.
Challenges in the pipeline

- **Data acquisition and recording challenges:**
  - Collection, data integrity, curation, metadata

- **Information extraction and cleaning challenges:**
  - How to extract and structure, how to clean, how to deal with error

- **Data integration, aggregation and representation challenges:**
  - Machine and human understandable, interoperability

- **Query processing, data modeling and analysis challenges:**
  - Meaningful mining, modeling, analysis, statistics
  - Scaling, error estimation, prediction accuracy

- **Interpretation challenges:**
  - Visualization, reproducibility, provenance, etc.
Challenges in the pipeline – cross-cutting issues

- **Heterogeneity challenges:**
  - Interoperability, harmonization of error, effective representation, flexible DB design

- **Scale challenges:**
  - Extensible storage and cloud technologies; I/O and query performance,
  - Data security, replication

- **Timeliness challenges:**
  - Real-time analysis, common vs. custom searches

- **Privacy/policy challenges:**
  - Support for privacy, differential privacy and sharing mechanisms, policy
  - Clarification of rights, ownership, access

- **Human collaboration challenges:**
  - Harmonization of input from multiple human experts and shared exploration of results
  - Crowd-sourcing approaches that facilitate correction of errors.
Evolving Data Science: Where should federal funding agencies invest?

NSF CISE Current Data Science Investment Strategy

**Foundational Research:**
- to develop new techniques and technologies to derive knowledge from data

**Cyberinfrastructure:**
- to manage, curate, and serve data to domain research communities

**Education & Workforce Development:**
- for a growing emerging discipline

**Collaborations:**
- to support interdisciplinary science and communities

Fran Berman, Data and Society, CSCI 4370/6370
Foundational question to guide evolution of data science: What kind of discipline is data science?

- Statistics?
- Machine learning?
- Hybrid discipline?
  - X-analytics
  - X-informatics (e.g. data expansion of bio-informatics)
  - data + X (e.g. data equivalent of computational biology)
- Discipline on its own?

Maybe all of the above, but ... what mechanisms can be used to grow and invest in the discipline to maximize its potential?

*Slide adapted from Rob Rutenbar, UIUC*
Target Areas for Recommendations to NSF for a Data Science Agenda (2016)

- National Data Science Research Agenda
- National Data Science Educ+Training Agenda
- Data Sets & Data Infrastructure for Research and Education

*Slide adapted from Rob Rutenbar, UIUC*
Recommendations: Promote a National Data Science Research Agenda

• **R1. CREATE DATA SCIENCE RESEARCH CENTERS.**
  - Focus / re-focus a set of subsequent calls for Science & Technology Centers (STCs) and Engineering Research Centers (ERCs) as Data Research Centers.

• **R2. INVEST IN RESEARCH INTO DATA SCIENCE INFRASTRUCTURE THAT FURTHERS EFFECTIVE DATA SHARING, DATA USE, AND LIFE CYCLE MANAGEMENT.**
  - Develop programs that focus attention on critical problems (privacy, inference, provenance, etc.) that remain obstacles to the use of data at scale.

• **R3. SUPPORT RESEARCH INTO EFFECTIVE REPRODUCIBILITY.**
  - Develop research programs that support computational reproducibility and computationally-enabled discovery, as well as cyberinfrastructure that supports reproducibility.

• **R4. FUND RESEARCH INTO MODELS THAT UNDERLIE EVIDENCE-BASED DATA POLICY AND DECISION MAKING.**
  - Invest in the development of models that can be used to support data-related policy, regulation, and strategic investment.

• **R5. EXPAND FUNDING INTO DEEP LEARNING, SMART ENVIRONMENTS, AND OTHER AI-EMPOWERED AREAS AND THEIR USE IN DATA-DRIVEN APPS.**
  - Continue to invest in both the foundations of “embodied intelligence” research and in their use by applications and within domains.
Recommendations:
Support Increased Data Science Education and Training

- **E.1 SUPPORT THE DESIGN AND DEVELOPMENT OF DS PEDAGOGY & CURRICULA.**
  - Make data a central focus for large cross-disciplinary education research centers, course and complete curricula, and modes of pedagogy.

- **E.2. TARGET EXISTING OR NEW PROGRAMS TO DEVELOP PROGRAMS TO DEVELOP DS CURRICULA AT EPSCOR AND MINORITY SERVING INSTITUTIONS.**
  - Include curricula and training programs that assist students at participating institutions to be competitive for Data Science internships in the private sector and fellowships in the academic sector.

- **E.3. ENCOURAGE “DATA INCUBATOR PROGRAMS” OR OTHER PRIVATE/PUBLIC PARTNERSHIPS THAT PROVIDE STUDENTS/FACULTY ACCESS AND OPPORTUNITIES FOR ENGAGEMENT WITH FACULTY/INDUSTRY/NON-PROFITS WITH REAL PROBLEMS.**
  - Develop programs that train a sophisticated Data Science workforce that is more prepared to address Data Science challenges at the innovation frontier.

- **E.4 SUPPORT PHD AND POSTDOC FELLOWSHIPS IN DATA SCIENCE.**
  - These are important to address current and future research and workforce needs.
Recommendations: Create Accessible Data Infrastructure to Support Research and Education

- **I.1. SUPPORT THE CREATION, ACQUISITION, AND PUBLIC DEPLOYMENT OF A BROAD PORTFOLIO OF REALISTIC, STATE-OF-THE-ART, AT-SCALE DATA SETS FOR ACADEMIC RESEARCHERS.**
  - Create collections of educational data sets and materials and make them available via publicly accessible repositories, libraries and stewardship environments. Link them to the publications they support.

- **I.2. DEVELOP AND DEPLOY DS INFRASTRUCTURE TO SUPPORT CUTTING EDGE RESEARCH / EDU.**
  - Invest in both national and institutional infrastructure to support emerging Data Science research and education programs.

- **I.3. CREATE NATIONAL EFFORT TO WORK WITH INSTITUTIONAL LIBRARIES AND DOMAIN REPOSITORIES TO DEVELOP SUSTAINBLE MODELS OF DATA STEWARDSHIP AND PRESERVATION OF VALUED SPONSORED RESEARCH DATA.**
  - Work with the community to develop and ramp into sustainable institutional economic models that enable participating organizations to provide open access research data stewardship options to the community over the long term.

- **I.4. ESTABLISH BEST PRACTICE GUIDELINES FOR DETERMINING THE BALANCE OF DATA SCIENCE INVESTMENT THAT SHOULD BE FOCUSED ON RESEARCH VS. ENABLING INFRASTRUCTURE.**
  - Look to other organizations, institutions, and sectors to establish a rough guideline about the distribution of investment to maximize the effectiveness of research and education efforts.

- **I.5. INSTRUMENT FASTLANE AND RESEARCH.GOV TO GATHER KEY INFORMATION ABOUT THE SCALE AND STATE OF SPONSORED RESEARCH DATA ACCESS AND STEWARDSHIP TO INFORM EVIDENCE-DRIVEN INVESTMENT AND POLICY.**
  - Gather statistics for sponsored research that describe the amount, characteristics, and stewardship of the data and allow linking between data and scholarly publication.
Recommendations: Invest in Areas that are increasingly expanding Data Science

Data-Enabled Policy & Evidence-Based Decisions

Data and the Internet of Things

Data and the End of Moore’s Law

• **N.1. STRENGTHEN THE CURRENT NSF INTERNET OF THINGS (IoT)-RELATED AGENDA.**
  – Focus in particular on critical areas such as data security, data privacy, smart systems, infrastructure supporting governance and policy, ethics, etc.

• **N.2. SUPPORT FUNDAMENTAL RESEARCH ON RADICAL HARDWARE AND SOFTWARE SYSTEM ARCHITECTURES THAT TARGET EMERGING DATA-INTENSIVE TASKS.**
  – Focus on research that promotes innovative data analysis methods for next-gen scenarios and the evolution or re-thinking of the fundamental computing platforms on which they might best be executed.

*Slide adapted from Rob Rutenbar, UIUC*
Lecture Materials (not already on slides)

- Southern California Earthquake Center,  [http://www.scec.org/](http://www.scec.org/)
- LHC, [www.wikipedia.com](http://www.wikipedia.com)
- San Diego Supercomputer Center, [www.sdsc.edu](http://www.sdsc.edu)
- “Challenges and Opportunities with Big Data” a community white paper developed by leading researchers across the U.S.,  [http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf](http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf)
Paper Assignment

• Paper draft: 16 points

• Final paper: 16 points
  – Undergrads: 5 page research paper on a data-related topic (does not include references)
  – Grads: 7 page research paper on a data-related topic (does not include references)

• Paper draft due March 10 before class
  – Bring hardcopy to class
Research Paper Structure

Specs

• Paper: 5 pages (undergrad) or 7 pages (grad), 1.5 spaced, 12 pt font

• Focus of paper should be an area of science or society that has been transformed by the availability of digital data

• General outline:
  — Description of the area/problem and the role of data in transforming it
  — Specifics on how data has made the transformation possible
  — Specifics on the kind of data tools, skills, policy, infrastructure, etc. that is needed to support this transformation
  — Conclusion and thoughts about the potential of data in this area in the future

• Paper should include adequate references and bibliography (not included in the page count).
  — If you use material from a source, reference it in the bibliography.
  — If you copy material from a source, put it in quotes and reference it in the bibliography.
Paper Grading Metrics (16 points total)

Content (8 points):
• Does the paper content provide adequate depth and evidence to describe the transformation of an area by digital data?
• Does the paper include appropriate specifics on innovation and infrastructure?
• Are the references reasonable and adequate?

Writing (8 points):
• Is the paper well-organized, readable by non-specialists, and credible to specialists?
• Does the writing tell a story? Is it articulate and compelling?
• Is the paper well-structured with the main points backed up by evidence?

If you want to check your topic or approach, please come to office hours (Friday 1-2, AE 218) or make an appt. with Fran (bermaf@rpi.edu)

This can save you time and help you meet the expectations of the assignment
Paper Draft and Paper Final

• **Paper draft due:** March 10
• **Paper draft returned:** March 31
• **Final paper due:** April 14

• **Important note:** If your draft research paper is strong, you may choose to not turn in the final research paper and double your draft grade (research paper grade = 2 X draft research paper grade). *You must make this decision and let Fran know before April 14.*
Break
Presentations
Presentation Articles for February 15

February 15 (data and science)

- “Big Data: Astronomical or Genomical?” PLOS Biology, [Patrick C]
  http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

- “Big data shows how what we buy affects endangered species,” Science Daily, [Priyanka K]
  https://www.sciencedaily.com/releases/2017/01/170104103604.htm

- “Neuroscience: Big brain, big data”, Scientific American, [Stephen N]
  https://www.scientificamerican.com/article/neuroscience-big-brain-big-data/

- “The Australian Square Kilometre Array Pathfinder Finally Hits the Big Data Highway”, Phys. Org., [James N]

- “Eat, Sleep, Repeat: Crowdsourcing the data of a baby’s typical day”, Discover, [Aditya S]

- “Computer Science Technique Helps Astronomers Explore the Universe” Inside Science, [Brandon T]
Presentation Articles February 17

• February 17 (data and cyberinfrastructure)


Presentations for February 22

- Feb 22 (data stewardship and preservation)

Fran Berman, Data and Society, CSCI 4370/6370
Feb 24 (data stewardship and preservation)


Class on Wednesday. Class next Friday.

• Next Friday: Data Cyberinfrastructure; Discussion

• Read for February 17 Discussion:

Presentation Articles February 10

- February 10:
  - “Crowdsourcing: For the Birds”, NY Times, [Dan S]
  - “African Elephant Numbers Plummet 30 Percent, Landmark Study Finds,” National Geographic, [Deborah A]