Welcome to Data and Society
CSCI 6963 (Grads) / 4967 (Undergrads)

• Professor: Dr. Fran Berman

• Office: AE 218, 276-3794

• Office Hours: Friday 1-2 or by appointment (send email to bermaf@rpi.edu)

• Course website (linked off Fran’s RPI web page): http://www.cs.rpi.edu/~bermaf/Data%20Course.html
Today (1/30/15)

• Why Data and Society?

• Intro – about this course
  – Syllabus and grading expectations
  – Learning objectives and expectations
  – Why are you here?

• Lecture 1

• Grading Specifics – Data Roundtable

• Break

• Data Round Table (Fran)
Data-driven innovation is a priority nationally, internationally, and in all sectors.
There’s more to the data story than technology

Policy and regulatory issues

Workforce evolution

Privacy and rights

New possibilities for innovation / new challenges for infrastructure

New modes of social and community interaction, organization
Data and Society – about this course

• This course will provide a broad snapshot of the data-driven world
  – We’ll skim the sea of interesting data stuff, but we won’t / can’t include everything
  – We’ll focus more on societal issues than technical issues
  – The course should provide a complement to the material in the ITWS Data Science, Web Science, Data Analytics and other courses

• The course will be structured to
  – Increase your engagement with material
  – Evolve your professional communication and assessment skills
  – Help you develop as a “data-literate” professional

Course structure:
• Section 1: The Data Ecosystem -- Fundamentals
• Section 2: Data and Innovation – How data has transformed science, commerce, and life
• Section 3: Data and Community – Social infrastructure for a data-driven world

Guest Speakers this Semester:
• Colin Bodel, CTO, Time Inc.
• Mike Schroepfer, CTO, Facebook
• Bulent Yener, RPI CS Professor

Fran Berman, Data and Society, CSCI 4967/696
<table>
<thead>
<tr>
<th>Section 1: The Data Ecosystem -- Fundamentals</th>
<th>January 30</th>
<th>Class introduction; Digital data in the 21st Century (L1)</th>
<th>Data Roundtable / Fran</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>February 6</td>
<td>Data Stewardship and Preservation (L2)</td>
<td>L1 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>February 13</td>
<td>Data and Computing (L3)</td>
<td>L2 Data Roundtable / 6 students</td>
</tr>
<tr>
<td></td>
<td>February 20</td>
<td>Colin Bodel, Time Inc. CTO Guest Lecture and Q&amp;A</td>
<td>L3 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section 2: Data and Innovation – How has data transformed science and society?</th>
<th>February 27</th>
<th>Section 1 Exam</th>
<th>Data and the Health Sciences (L4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>March 6</td>
<td>Paper preparation / no class</td>
<td></td>
</tr>
<tr>
<td></td>
<td>March 13</td>
<td>Data and Entertainment (L5)</td>
<td>L4 Data Roundtable / 6 students</td>
</tr>
<tr>
<td></td>
<td>March 20</td>
<td>Big Data Applications (L6)</td>
<td>L5 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section 3: Data and Community – Social infrastructure for a data-driven world</th>
<th>April 3</th>
<th>Data in the Global Landscape (L7) Section 2 paper due</th>
<th>L6 Data Roundtable / 6 students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>April 10</td>
<td>Bulent Yener Guest Lecture, Data Privacy / Bad guys on the Internet (L8)</td>
<td>L7 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>April 17</td>
<td>Data and the Workforce (L9)</td>
<td>L8 Data Roundtable / 6 students</td>
</tr>
<tr>
<td></td>
<td>April 24</td>
<td>Mike Schroepfer, Facebook CTO Guest Lecture and Q&amp;A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>May 1</td>
<td>Data Futures (L10)</td>
<td>L9 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>May 8</td>
<td>Section 3 Exam</td>
<td>L10 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>
Course Information
http://www.cs.rpi.edu/~bermaf/Data%20Course.html

• Course website (above) will have all up-to-date information and materials.
  – Syllabus may evolve slightly

• Course Readings
  – Provide background and information for the lectures and issues
  – Will be given every week. Most source materials and readings for each week provided on the course web page
  – Readings eligible for Data Roundtable reviews/presentations will be specified in class and on the web

• Source and reading materials may be tested on the Section exams – relevant material will be indicated
How you’ll be graded

Student grades are computed from:

- **2 section exams** (20 points each)
- **1 section paper** (20 points):
  - Undergrads: 6-8 page research paper on an approved Section 2 topic
  - Grads: NSF-style 10 page mini-proposal on an approved Section 2 topic
- **Class participation** (10 points)
- **3 Data Roundtable reviews / presentations** (roughly 1 per section, 10 points each)

Students can obtain up to 5 points extra credit by doing an Op-Ed (due any time up to May 1)
More about grading
(additional grading specifics for each component later today)

- **Data Roundtables:**
  - Students are responsible for scheduling their Data Roundtables and ensuring that all 3 are done.
  - Students are responsible for self-organizing so that each student presents a **different** Roundtable article.

- **Class engagement / attendance:**
  - Students are expected to attend 90+ % of the class meetings (12/13).
    Attendance will be taken in class
  - Engagement grade: 5% attendance, 5% class participation

- **Exams** will be primarily in essay format. You’re responsible for anything covered in class and in the relevant readings.

- **Extra credit:** Up to 5 points extra credit can be earned by doing an Op-Ed by May 1.

- **There will be a slightly different workload for grad students and undergrads**
  - Section 2 paper / mini-proposal requirements are different.
  - In writing and presentations, each student will be assessed at a level appropriate to their educational level (undergrad or grad)
## Learning Objectives and Outcomes

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop greater data literacy</td>
<td>Be able to understand and explain the role that data plays as well as its limitations in various areas of research, commerce and modern life.</td>
</tr>
<tr>
<td>Develop critical thinking skills around data</td>
<td>Be able to read, understand, assess, and discuss data-oriented professional and popular publications and articles.</td>
</tr>
<tr>
<td>Develop communication skills around data</td>
<td>Be able to advance an evidence-based argument about data, data cyberinfrastructure and data-oriented efforts to both knowledgeable specialists within the field as well as non-specialists.</td>
</tr>
</tbody>
</table>
Academic Integrity

• Student-teacher relationships are built on trust. For example, students must trust that teachers have made appropriate decisions about the structure and content of the courses they teach, and teachers must trust that the assignments that students turn in are their own. Acts, which violate this trust, undermine the educational process. The Rensselaer Handbook of Student Rights and Responsibilities defines various forms of Academic Dishonesty and you should make yourself familiar with these.

• In this class, all assignments that are turned in for a grade must represent the student’s own work. In cases where help was received, or teamwork was allowed, a notation on the assignment should indicate your collaboration. If references or other materials are used, they should be cited. Submission of any assignment that is in violation of this policy will result in a penalty.

• If found in violation of the academic dishonesty policy, students may be subject to two types of penalties. The instructor administers an academic (grade) penalty, and the student may also enter the Institute judicial process and be subject to such additional sanctions as: warning, probation, suspension, expulsion, and alternative actions as defined in the current Handbook of Student Rights and Responsibilities. If you have any question concerning this policy before submitting an assignment, please ask for clarification.
Why are you here?

1. Name, major, year, grad or undergrad?
2. What do you expect to be doing after you finish your degree?
3. Why did you take this course?
4. What do you hope to get out of this course?
5. What areas / topics in the data landscape are of most interest to you?
6. What is the coolest recent thing you’ve heard about digital data?
Lecture 1: Data and Society
Lecture 1 Outline

• Some basics
  – How does data vary?
  – How much data is there and where does it come from?

• Data Transformation -- The Information Age
Data comes from everywhere

- Entertainment
- Health
- Education
- Commerce
- Physical Infrastructure and Smart Systems
- Research

Communication / Community
All data is not alike

- **Volume**: amount of data, number of files
- **Velocity**: Rate at which data flows into an organization as well as speed of the “feedback loop” (can the data be where you want it when you want it)
- **Variety**: Diversity of data types and sources (“messiness” of using, combining, managing data)
- **Value**: importance of the data
- **Volatility**: how quickly data changes, how long the data is useful for

Sources:  
- [http://strata.oreilly.com/2012/01/what-is-big-data.html](http://strata.oreilly.com/2012/01/what-is-big-data.html)  

- **Validity**: legitimacy / accuracy of sources
- **Viscosity**: resistance to flow in the volume of data (improved infrastructure, management, and technologies can reduce viscosity)
- **Virality**: how quickly the data is disperse and shared
- **Variability**: Extent to which data points differ from each other. (commonly used measures of variability: range, mean, variance and standard deviation)
How we access, manage, use, store and preserve data also varies widely

- **RETENTION TIMEFRAME:**
  - Short-term (few months, years) to long-term (decades, centuries, …)

- **SIZE / SCALE:**
  - Small-scale (GBs, MBs) to large-scale / “big” (TBs, PBs, EBs)

- **PREPARATION:**
  - Well-tended (curated, sufficient metadata, cleaned and filtered) to poorly tended (flat files, insufficient metadata)

- **POLICY / REGULATION RESTRICTIONS:**
  - Subject to more restrictive policy and regulation (e.g. HIPAA) vs. subject to less restrictive policy and regulation

- **LIFE CYCLE PLANNING:**
  - Has a data management and / or sustainability plan vs. ad hoc approach

- **COMMUNITY ACCESSIBILITY:**
  - Shared with others in the community vs. kept private; Curated and organized using community standards vs. ad hoc or home-grown approaches
Meaning and context increase the impact of data: Data, Information, Knowledge, Wisdom

- **Data** = Qualitative or quantitative values at the lowest level of abstraction
- **Information** = Data and its associated meaning
- **Knowledge** = Theoretical or practical understanding of information
- **Wisdom** = The quality of having experience, knowledge and good judgment

Considerable overlap and many definitions ...

Many articles use digital data and digital information interchangeably. For the most part, we will too.
How Much Data is There?

- There won’t be an exaflop supercomputer until the end of the decade.
- We have had exabytes of data for at least 10 years and hit a zettabyte in 2009-2010.
Digital universe doubling every two years

How big is ... (roughly)

- **A byte:** ~ 1 character
- **A megabyte:** ~ 1 small novel
- **A gigabyte:** 50 gigabytes → ~ 1 floor of books
- **A terabyte:** 10 Terabytes → ~ printed collection of the Library of Congress
- **A petabyte:** 2 Petabytes → ~ All US academic research libraries
- **An Exabyte:** 5 exabytes → all words ever spoken by human beings recorded in text
- **A Zettabyte:** 42 zettabytes → ~ all words ever spoken by human beings digitized as audio

Source: http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html
Where does the data come from?

• Digital information comes from many sources: computers, RFIDs and sensors, scientific and other instruments, imaging devices, cell phones, etc.

• Most of the digital universe is transient – unsaved Netflix streams, temporary routing information in networks, sensor signals discarded when no alarms go off, etc.

• 2014: Digital universe = 1.7 MB/minute for every person on earth (2+ gigapeople ...)

• Areas experiencing increasing data analysis and use:
  – Surveillance footage
  – Embedded and medical devices
  – Entertainment and social media
  – Consumer images
  – Enterprise transactional data; data processing
Huge growth in data from the “Internet of Things”

- Major growth spurts in the digital universe:
  - Film $\rightarrow$ digital technology
  - Analog functions monitoring and managing physical world $\rightarrow$ digital functions involving communications and software telemetry
  - Analogue TV $\rightarrow$ digital TV
  - Increasing data from embedded systems
- IDC estimates that
  - In 2014, things in the digital universe approaching 200 billion, 10% (20 billion) of those wired and communicating with the Internet
  - In 2020, things in the digital universe roughly 30 billion devices will be connected to the Internet

Huge growth in mobile “things” – connected and otherwise

Finally, a good portion of the digital universe will be generated by mobile devices and people – from 17% in 2013 to 27% in 2020 – but the percentage of mobile “things” in the IoT will be more than 75% by 2020.

More growth trends in the Digital Universe

• Growth of the Internet (> 1 billion users) and broadband availability

• Conversion of formerly analog information to digital

• Falling prices and increased performance for digital devices; ability to store more information and share it in standard formats

• Rise of automation, data-intensive, graphics-intensive, and “smart” applications

• Rise of data centers, cloud computing, social networks

• Regulations mandating new archiving and privacy protection rules

• Increased computerization of business, education, entertainment, etc.
Which data is useful?

• Data is useful when we know something about it – what it represents, where it was collected, what units are being used, etc.? Metadata a critical part of the data universe that makes data useful.

• IDC estimates tremendous growth in useful data from 2014 to 2020

Which data is “valuable”?

• Value: IDC considers “target rich” data as data that is easy to access, transformative, real-time, have intersection synergy (multiple attributes), large footprint (affect many things).

• In 2014, IDC estimates that 6% of “useful” data (1.5% of the total digital universe) is “target-rich”. This is predicted to grow to 11% in 2020.


Note that “value” is in the eye of the beholder. This is IDC’s take on this. More perspectives in Lecture 2.
All digital data cannot be stored

- 2007 was the “crossover year”: Began to generate more digital data than storage to keep it.
- In 2013, current storage capacity could hold just 33% of the digital universe.
- By 2020, current capacity will be able to store less than 15% of the digital universe.

Source: IDC, 2008
2010 Update on the Storage Gap:
By 2020, more than twice as much information will be created as storage available

Figure 5: The Emerging Gap
Information Creation > Storage Available

Source: IDC Digital Universe Study, sponsored by EMC, May 2010
Fran Berman, Data and Society, CSCI 4967/6963
Digital universe profiles


<table>
<thead>
<tr>
<th>Region</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>32%</td>
</tr>
<tr>
<td>Western Europe</td>
<td>32%</td>
</tr>
<tr>
<td>China</td>
<td>19%</td>
</tr>
<tr>
<td>India</td>
<td>13%</td>
</tr>
<tr>
<td>Rest of the</td>
<td>4%</td>
</tr>
</tbody>
</table>

Total: 2,837 EB

Type of Information in the Cloud in 2020

- Computers, Phones, Consumer Electronics: 46.7%
- Entertainment: 35.1%
- Embedded and Medical: 8.5%
- Surveillance: 9.7%

Total: 14,996 EB


IDC, 2014

Information Security: Much of the Data that Needs to Be Protected Is Not Yet Protected

- Examples: Camera phone photos, digital video streaming, public website content, open source data

Portion of DU Not Needing Protection: 57%

Portion of DU Needing Protection: 43%

Portion Protected: 48%

Portion Not Protected: 52%

Source: IDC, 2014
The “Digital Shadow”

- Less than half of your digital footprint is related to individual actions – taking pictures, making VoIP calls, uploading files, etc.
- The rest of your digital footprint is “ambient” content and metadata related to you: surveillance images, banking records, medical records, information about your web searches and behavior in social networks, etc.
Who can collect, has rights to, and can use information about you, and under what conditions is the subject of national discussions world-wide.
Need for data and IT-savvy professionals having tremendous impact on the workforce


140,000–190,000 more deep analytical talent positions, and 1.5 million more data-savvy managers needed to take full advantage of big data in the United States.
How did IDC calculate their estimations?

• Forecasts developed for 40+ classes of devices and/or applications that can capture or create digital information

• Estimate annual usage and number of times a unit of information is replicated, either to share or store.

• Analysis based on previous IDC research, information capture and workload characteristics, surveys, studies, etc.

• Data adjusted for geographic region, kind of device, kind of information, etc.

Data Transformation: The Information Age
The Information Age

• “The **Information Age** (also known as the **Computer Age**, **Digital Age**, or **New Media Age**) is a period in human history characterized by the **shift from traditional industry that the industrial revolution brought through industrialization, to an economy based on information computerization**. The onset of the Information Age is associated with the Digital Revolution, just as the Industrial Revolution marked the onset of the Industrial Age.” *Wikipedia*
How did the Industrial Revolution Transform the World?

• Transition to new manufacturing processes in late 18th / early 19th century.
  – Hand production → machines, new chemical manufacturing, new iron production processes
  – Improved efficiency of water power and the increased use of steam power
  – Wood and bio-fuels → coal
  – England → Western Europe, US

• Major turning point, almost every aspect of daily life influenced in some way
Technological Innovation during the Industrial Revolution

- New technological capabilities and emerging needs had a transformative effect on
  - Work opportunities and workforce needs
  - National and international priorities
  - Economic, cultural, social, and political structures
  - Leading sectors (manufacturing, health, energy) and new areas for innovation and impact. Broad ripple effect from both.

Jacquard Loom – Precursor to the Programmable Computer

- **Jacquard loom** invented by Joseph Marie Jacquard and first demonstrated in 1801.

- Loom controlled by punch cards for the purpose of manufacturing textiles with complex patterns.
  - Rows of holes were punched on each card corresponded to one row of the design.

- Loom serves as an important conceptual precursor in the development of computer programming

Social Innovation during the Industrial Revolution

- **Economic transformation**
  - Better standard of living
  - Better agricultural practices, housing, food supplies
  - Less expensive clothing and consumer goods

- **Urbanization**
  - Rise of factories and modern cities
  - Change in employment options

- **Social policy**
  - Child Labor laws
  - Growth in trade unions

*Cottonopolis* is a name given to the city of Manchester, in England. It denotes a metropolis of cotton and cotton mills, as inspired by Manchester's status as the international centre of the cotton and textile processing industries during this time.

Engraving by Edward Goodall (1795-1870), original title *Manchester, from Kersal Moor* after a painting of W. Wylde. Wikipedia (cropped from original)
Fast forward to the Information Age

• We are experiencing a transformation analogous to the Industrial revolution

• New technological capabilities and emerging needs again having a transformative effect on
  – Work opportunities and workforce needs
  – National and international priorities
  – Economic, cultural, social, and political structures
  – Leading sectors and new areas for innovation and impact. Broad ripple effect from both.
Transformative Potential of Data: Emerging Technologies

Exascale computing →
more compute and data at all tiers in the Branscomb Pyramid.

New breakthroughs in power and computer architectures required.

Smart Devices, Sensor Networks →
More data-enabled devices and approaches drive crowd-sourced, real-time, and other aggregation applications

Information-Driven Analysis →
X-informatics and X-analytics enable new targets for data-driven research and decision-making models
Transformative Potential of Data: Massive-scale coordination, inclusion, access

Greater access → Greater participation, “democratization” possible

High quality, on-line education → On-line / on-site education solutions have the potential to transform higher education

Greater transparency, management, monitoring → More measurement, transparency, monitoring possible
Social Impacts – adequate legal, regulatory, and policy underpinnings for data needed

- How do you maintain personal freedom and sufficient privacy / control over your information?

- What are your rights?
  - What do you own?
  - What can you distribute?
  - What can you charge for?

- What / whom do you trust?
  - Your data?
  - Your respondent?
  - Your hardware?
  - Your system / software?
What happens when digital data becomes the vehicle for progress in the Information Age?

• What kind of infrastructure is needed to support the access, management, use and re-use of digital data today and tomorrow? (Section 1) -- **How do we create a useful data ecosystem?** How do we make it sustainable?

• How is digital data being used to drive new innovation? (Section 2) – **How do we make the most out of data?**

• What social and community constructs are needed to realize data’s potential? (Section 3) – **What’s needed for a digitally responsible society?**
Towards a sustainable data ecosystem

*Sustainable development:* "development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

*Our Common Future, U.N. Brundtland Commission*

- **Key components**
  - Ecological sustainability
  - Cultural / institutional sustainability
  - Economic sustainability
  - Political sustainability
What contributes to sustainability?

- **Circles of Sustainability** developed to assess and understand sustainability. Used
  - For managing projects directed towards socially sustainable outcomes
  - To assess the sustainability of cities and urban settlements


Fran Berman, Data and Society, CSCI 4967/6963
Next time: What infrastructure is needed to support digital stewardship and preservation?
Lecture 1 Sources (not already in text) (pdfs or links on course website)

• You are responsible for material on Lecture 1 slides and in red for Exam 1

• L1 Sources
  
    -- responsible for entire report (all 5 sections)

  – Previous IDC Reports: http://www.emc.com/leadership/digital-universe/index.htm#Archive

  – “It’s official: 2014 was the hottest year in recorded history”, Wonkblog, Washington Post, January 16
    http://www.washingtonpost.com/blogs/wonkblog/wp/2015/01/16/its-official-2014-was-the-hottest-year-in-recorded-history/ (Data Roundtable)
Break
Data Roundtable Grading Specifics
How you’ll be graded

Student grades are computed from:

• **2 section exams** (20 points each)
• **1 section paper** (20 points):
  – Undergrads: 6-8 page research paper on an approved Section 2 topic
  – Grads: NSF-style 10 page mini-proposal on an approved Section 2 topic
• **Class participation** (10 points)
• **3 Data Roundtable reviews / presentations** (roughly 1 per section, 10 points each)

Students can obtain up to 5 points extra credit by doing an **Op-Ed** (due any time up to May 1)
Grading Detail – Data Roundtables

Do 3 of these, 10 points each, one in each Section.

• Grade distribution:
  – Written review: 3 points on content of review, 2 points on writing. Reviews should be 3-4 typed pages (12 pt. font).
  – Oral presentation: 3 points on presentation slides, 2 points on presentation style.

• Roundtable sources should come from the designated materials on the class website

• Students are responsible for self-organizing so that each student presents a different Roundtable article.

• All written reviews must be turned in at the beginning of the class during which you do your oral presentation. Please send a copy of the presentation slides and a .pdf of the review to bermaf@rpi.edu.
Grading Detail – Data Roundtable Written Review

Each **written review** (3-4 pages) should include:

1. **Succinct summary of the points of the article**
   - What is the article about? What issues does it focus on?
   - Why are these issues interesting/important?
   - What is newsworthy in the article and why?

2. **Data Issues:**
   - How does digital data play a role in the article?
   - What is the “data backstory”, i.e. what data infrastructure, policy, practice, etc. is needed to be there for the data to play its role?

3. **Next steps for exploration**
   - Where you would go next to find out more about these issues. What would you read?
   - What would be appropriate/interesting areas for exploration based on this article?

4. **Your thoughts:**
   - If the article is about a report, article, paper or other source material, did they do an accurate job of representing it?
   - Did you like the article? Why or why not?

**Written review Grading Metrics:**

**Content (3%):**
- Does the review content demonstrate a clear understanding of the material?
- Are the main points and issues clearly described?

**Writing (2%):**
- Is the review well-organized and readable by non-specialists?
- Does the review “tell a story”?
- Are the “next steps for explanation” and “thoughts” section thought-provoking and interesting?
Grading Detail – Data Roundtable Oral Presentation

Oral presentation components
(10 minutes presentation + 5 minutes Q&A):

1. **Summary:**
   - What is the article about?
   - What is the point of view presented in the article?

2. **What are the data issues?**
   - How is data used to support the article’s point of view? Does it succeed in doing this?
   - What is the “data backstory”, i.e. what data infrastructure, policy, practice, etc. needed to be there for the data to play its role?

3. **How is the article useful in a broader context?**
   - What questions arise from reading this article?
   - How can the article “takeaways” be applied in other settings?

Note: You may need to read additional publications, websites for your presentations and reviews

Oral Presentation Grading Metrics:

**Talk (3%):**
- Does the speaker understand and communicate well about their topic?
- Does the presentation tell an interesting story?
- Is the speaker well prepared for questions?

**Slides (2%):**
- Are the slides well-organized and informative?
- Do the slides help tell the story?
- Are the slides visually interesting?
Data Roundtable Presentation format

Summary:
• What is the article about?
• What is the point of view presented in the article?

What are the data issues?
• How is data used to support the article’s point of view? Does it succeed in doing this?
• What is the “data backstory”, i.e. what data infrastructure, policy, practice, etc. needed to be there for the data to play its role?

How is the article useful in a broader context?
• What questions arise from reading this article?
• How can the article “takeaways” be applied in other settings?
### Data Round Table 1

**It's official: 2014 was the hottest year in recorded history**

*By Chris Mooney, January 16*

#### Land & Ocean Temperature Percentiles Jan–Dec 2014

By NOAA’s National Climatic Data Center

Data Source: GSN0-M version 2.2 & ERNST version 3.1

National Oceanic and Atmospheric Administration

---

**Washington Post, Wonkblog, January 16**

Fran Berman, Data and Society, CSCI 4967/6963
Key Results

• NASA and NOAA declared 2014 was hottest year ever measured (measurements go back to 1880)
  – Japan Meteorological Agency also pronounced 2014 the hottest year in its records (going back to 1891)
  – Berkeley Earth: 2014 is nominally the hottest year, but statistically tied with 2005 and 2010
  – U. of Alabama Huntsville: 2014 the third warmest year for “this part of the planet”

• 10 warmest years in instrumental record (except 1998) have occurred since 2000

• NASA: First year since 1997 that the record warmest year was not an “El Nino” year at the beginning of the year.
  – El Nino = irregularly occurring and complex series of climatic changes affecting the equatorial Pacific region and beyond every few years, characterized by the appearance of unusually warm, nutrient-poor water off northern Peru and Ecuador, typically in late December.
What increasing temperature looks like
Why was it so warm?

• NASA / NOAA say new record driven the world ocean – sea surface temperatures were warmer; land temperatures were not uniformly
  – Eastern U.S. considered a “cold anomalie” 😊
  – Calculations reflect Earth’s average temperature. Aggregate increase is 1.4 degrees F since 1880.

• NASA scientists believe that increase in temperature driven by increase in carbon dioxide and other human emissions into the planet’s atmosphere.
  – Majority of warming happened in last 3 decades
  – Local weather patterns can cause “noise” so increased temperature is not linear, but overall trend is increasing temperature
How the temperature is calculated

• Data analysis incorporates 6300 surface temperature measurements from
  – land-based weather stations
  – ship- and buoy-based observations of sea surface temperatures
  – Antarctic research stations
  – Satellites, air-bourne and land-based observation campaigns

• Raw data analyzed using an algorithm that takes into account varied space and urban heating effects that could skew the calculation

• Model evaluates globally-averaged temperature across land and ocean surfaces

• Model validated on measurements from 1951 to 1980, applied to current sensor measurements
Key IT issues

- **Correctness of data:**
  - Sensor readings must be within reasonable tolerance for accuracy
  - Distribution of sensor readings must be taken into account in analysis
  - Sensor readings must be interoperable – different units or definitions for metadata must be harmonized

- **Accuracy of the analysis algorithm:**
  - Computational analysis must produce reasonably correct results
  - Model must be tested on prior data to ensure it aligns with known results in order to be credible for assessments and predictions

- **Stewardship and preservation of raw data**
  - Data must be curated, accessible for current and future analysis
  - Processes, SW, HW, and organizations must be in place to ensure stewardship and preservation of evolving raw and derived data collections
Broader Impact

• **Warming trend used to support scientific claims about global warming**
  – NASA claims that new record is what you would expect to see on a warming planet
  – “If you’ve got a long-term warming trend, you’re going to get new records every so often – in fact on a pretty regular basis. This is what you’re expecting, and this is going to continue to happen because the underlying rate of global warming really hasn’t changed.” Gavin Schmidt, NASA

• **Warming trend impacts planetary environment and ecosystems**
  – Warming trend has tremendous impact on planetary ecosystems, who depend on specific climate conditions.
    • Agricultural yields can vary with temperature change, increased floods or droughts, etc.
    • Range and health of animal populations and their risk for parasites, disease, predators, etc. may vary
References

• “It’s official: 2014 was the hottest year in recorded history”, Wonkblog, Washington Post, January 16
  http://www.washingtonpost.com/blogs/wonkblog/wp/2015/01/16/its-official-2014-was-the-hottest-year-in-recorded-history/

• “State of the Climate: Global Analysis for December 2014”, NOAA National Climatic Data Center, published online January 2015,
  http://www.ncdc.noaa.gov/sotc/global/2014/12


• EPA website: http://www.epa.gov/climatechange/impacts-adaptation/
## Data Round Table 1

*It’s official: 2014 was the hottest year in recorded history*

By Chris Mooney, January 16

Washington Post, Wonkblog, January 16

---

#### Land & Ocean Temperature Percentiles Jan–Dec 2014

**NOAA’s National Climatic Data Center**

Data Source: GHAT-M version 2.2.2 & ER饬T version 3.6

![Temperature Map](image)

It is official: According to both NASA and the National Oceanic and Atmospheric Administration, the year 2014 was the hottest ever measured.

---

1. Quiz: Can you name these cities just by looking at their subway maps?
2. President Obama finally has his Piketty moment
3. Obama’s new tax plan would go straight after Mitt Romney’s money
4. Traveling to Europe is about to get a whole lot cheaper
5. There are only three countries in the world where your boss is more likely to be a woman
Key Results

• NASA and NOAA declared 2014 was hottest year ever measured (measurements go back to 1880)
  – Japan Meteorological Agency also pronounced 2014 the hottest year in its records (going back to 1891)
  – Berkeley Earth: 2014 is nominally the hottest year, but statistically tied with 2005 and 2010
  – U. of Alabama Huntsville: 2014 the third warmest year for “this part of the planet”

• 10 warmest years in instrumental record (except 1998) have occurred since 2000

• NASA: First year since 1997 that the record warmest year was not an “El Nino” year at the beginning of the year.
  – El Nino = irregularly occurring and complex series of climatic changes affecting the equatorial Pacific region and beyond every few years, characterized by the appearance of unusually warm, nutrient-poor water off northern Peru and Ecuador, typically in late December.
What increasing temperature looks like

Start here

Fran Berman, Data and Society, CSCI 4967/6963
Why was it so warm?

- NASA / NOAA say new record driven the world ocean – sea surface temperatures were warmer; land temperatures were not uniformly
  - Eastern U.S. considered a “cold anomalie” 😊
  - Calculations reflect Earth’s average temperature. Aggregate increase is 1.4 degrees F since 1880.

- NASA scientists believe that increase in temperature driven by increase in carbon dioxide and other human emissions into the planet’s atmosphere.
  - Majority of warming happened in last 3 decades
  - Local weather patterns can cause “noise” so increased temperature is not linear, but overall trend is increasing temperature
How the temperature is calculated

• Data analysis incorporates **6300 surface temperature measurements** from
  – land-based weather stations
  – ship- and buoy-based observations of sea surface temperatures
  – Antarctic research stations
  – Satellites, air-borne and land-based observation campaigns

• Raw data analyzed using an algorithm that takes into account varied space and urban heating effects that could skew the calculation

• **Model evaluates globally-averaged temperature across land and ocean surfaces**

• Model validated on measurements from 1951 to 1980, applied to current sensor measurements
Key IT issues

• Correctness of data:
  – Sensor readings must be within reasonable tolerance for accuracy
  – Distribution of sensor readings must be taken into account in analysis
  – Sensor readings must be interoperable – different units or definitions for metadata must be harmonized

• Accuracy of the analysis algorithm:
  – Computational analysis must produce reasonably correct results
  – Model must be tested on prior data to ensure it aligns with known results in order to be credible for assessments and predictions

• Stewardship and preservation of raw data
  – Data must be curated, accessible for current and future analysis
  – Processes, SW, HW, and organizations must be in place to ensure stewardship and preservation of evolving raw and derived data collections
Broader Impact

• **Warming trend used to support scientific claims about global warming**
  
  – NASA claims that new record is what you would expect to see on a warming planet
  
  – “If you’ve got a long-term warming trend, you’re going to get new records every so often – in fact on a pretty regular basis. This is what you’re expecting, and this is going to continue to happen because the underlying rate of global warming really hasn’t changed.” Gavin Schmidt, NASA

• **Warming trend impacts planetary environment and ecosystems**
  
  – Warming trend has tremendous impact on planetary ecosystems, who depend on specific climate conditions.
    
    • Agricultural yields can vary with temperature change, increased floods or droughts, etc.
    
    • Range and health of animal populations and their risk for parasites, disease, predators, etc. may vary
References

• “It’s official: 2014 was the hottest year in recorded history”, Wonkblog, Washington Post, January 16
  http://www.washingtonpost.com/blogs/wonkblog/wp/2015/01/16/its-official-2014-was-the-hottest-year-in-recorded-history/

• “State of the Climate: Global Analysis for December 2014”, NOAA National Climatic Data Center, published online January 2015,
  http://www.ncdc.noaa.gov/sotc/global/2014/12


• EPA website: http://www.epa.gov/climatechange/impacts-adaptation/

• Information in black from Fran
Written Review Talking Points

1. Succinct summary of the points of the article
   • What is the article about? What issues does it focus on?
   • Why are these issues interesting/important?
   • What is newsworthy in the article and why?
   • NASA and NOAA declared 2014 was hottest year ever measured (measurements go back to 1880)
   • NASA and NOAA believe that record driven by sea surface temperature rather than land
   • NASA / NOAA Scientists see increasing temperature trend as evidence for global warming

2. Data Issues:
   • How does digital data play a role in the article?
   • What is the “data backstory”, i.e. what data infrastructure, policy, practice, etc. is needed to be there for the data to play its role?
   • Temperature data calculated from sensors measuring land and sea surface
   • NASA/NOAA analyses corroborated by Japanese Meteorological Society and Berkeley Earth (who think that 2014 is within the margin of error with other years)
   • Analysis model must be validated and accurate.
   • Organizational stewardship and preservation critical to maintain temperature data for future use
Written Review Talking Points

3. Next steps for exploration

• Where you would go next to find out more about these issues. What would you read?
• What would be appropriate/interesting areas for exploration based on this article?
• NASA and NOAA press releases and more in-depth reports, other articles about 2014 temperatures, scientific papers on the data set and the analytical model
• Further exploration:
  – How accurate is the model? How accurate is the data? How does this impact the measurements?
  – If temperatures follow this trend, what are specific impacts on planetary ecosystems and when are they likely to happen?

4. Your thoughts:

• If the article is about a report, article, paper or other source material, did they do an accurate job of representing it?
• Did you like the article? Why or why not?
• Article could have included more detail about where the measurements came from and how global temperature was calculated.
• Article coupled politics of global warming with temperature increase but the science was a bit skimpy
• I would have liked to see more specifics about the data science and analysis in the article.
Your turn: Eligible readings for Lecture 1 Data Roundtable (February 6)


Eligible readings for Lecture 2 Data Roundtable (February 13)

- Princeton Single-Pay Storage model (2 students):
  - “DataSpace: A Funding and Operational Model for LongTerm Preservation and Sharing of Research Data” White Paper (August, 2010), [http://dataspacenr.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf](http://dataspacenr.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf) (Lars Olsson)
Next time:  Data Stewardship and Preservation