Welcome to Data and Society
CSCI 6370 (Grads) / 4370 (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%202016/Data%20Course%20-%202016.html
Today (1/29/15)

• Why Data and Society?

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

• Lecture 1

• Break

• Grading Specifics – Data Roundtable

• 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

New possibilities for innovation / new challenges for infrastructure

Privacy and rights

New modes of social and community interaction, organization

Fran Berman, Data and Society, CSCI 4370/6370
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 and infrastructure

• 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:

• Phil Bourne, NIH
• Bulent Yener, RPI CS Professor
<table>
<thead>
<tr>
<th>Section Theme</th>
<th>Date</th>
<th>First “half”</th>
<th>Second “half”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Section 1: The Data Ecosystem -- Fundamentals</strong></td>
<td>January 29</td>
<td>Class introduction; Digital data in the 21st Century (L1)</td>
<td>Data Roundtable / Fran</td>
</tr>
<tr>
<td></td>
<td>February 5</td>
<td>Data Stewardship and Preservation (L2)</td>
<td>L1 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>February 12</td>
<td>Data-driven Science (L3)</td>
<td>L2 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>February 19</td>
<td>Future infrastructure – Internet of Things (L4)</td>
<td>L3 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>February 26</td>
<td>Section 1 Exam</td>
<td>L4 Data Roundtable / 5 students</td>
</tr>
<tr>
<td><strong>Section 2: Data and Innovation – How has data transformed science and society?</strong></td>
<td>March 4</td>
<td>Paper assignment description</td>
<td>Section 1 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>March 11</td>
<td>Data and Health: Phil Bourne guest lecture (L5)</td>
<td>Section 2 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>March 18</td>
<td>Spring Break / no class</td>
<td></td>
</tr>
<tr>
<td></td>
<td>March 25</td>
<td>Data and Entertainment (L6)</td>
<td>L5 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>April 1</td>
<td>Big Data Applications (L7)</td>
<td>L6 Data Roundtable / 5 students</td>
</tr>
<tr>
<td><strong>Section 3: Data and Community -- Social infrastructure for a data-driven world</strong></td>
<td>April 8</td>
<td>Data in the Global Landscape (L8) Section 2 paper due</td>
<td>L7 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>April 15</td>
<td>Digital Rights (L9)</td>
<td>L8 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>April 22</td>
<td>Bulent Yener Guest Lecture, Data Security (L10)</td>
<td>L9 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>April 29</td>
<td>Digital Governance and Ethics (L11)</td>
<td>L10 Data Roundtable / 5 students</td>
</tr>
<tr>
<td></td>
<td>May 6</td>
<td>Section 3 Exam</td>
<td>L11 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>
Course Information
http://www.cs.rpi.edu/~bermaf/Data%20Course%202016/Data%20Course%2020-%202016.html

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

• Reference and Roundtable materials will be on the web
  – Embedded reference materials in the lecture will be given by URL.
    Lectures will be on the web.

• Reference and reading materials may be tested on the Section exams
How you’ll be graded

Student grades are computed from:

• **2 section exams** (20 points each)

• **1 section paper** (20 points):
  – Undergrads: 7 page research paper on an approved topic
  – Grads: 10 page research paper on an approved topic

• **Class participation** (10 points)

• **1 Op-Ed** (10 points, due before April 29)

• **2 Data Roundtable reviews / presentations** (roughly 1 per section, 10 points each)

Time permitting, one “do-over” op-ed or data roundtable may be accepted. More information on whether that is possible will be provided after Section 1.
More about grading
(additional grading specifics for Data Roundtable later today)

- **Data Roundtables:**
  - Students are responsible for scheduling their Data Roundtables and ensuring that both are done. Information about Data Roundtables will be given during Lecture 1.

- **Op-Eds, Papers:**
  - Information about Op-Eds will be given during Lecture 2. Information about the Paper Assignment will be given on or before March 4.

- **Class engagement / attendance:**
  - Students are expected to attend 14/15 out of the class meetings. 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.

- **There will be a slightly different workload for grad students and undergrads**
  - Section 2 paper lengths 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, school, 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 much data is there and where does it come from?
  – How does data vary?
  – What does the “data universe” look like?

• Data Transformation -- The Information Age
What is Digital Data?

• Wikipedia: “Digital data, in information theory and information systems, are discrete, discontinuous representations of information or works, as contrasted with continuous, or analog signals which behave in a continuous manner, or represent information using a continuous function.

• Although digital representations are the subject matter of discrete mathematics, the information represented can be either discrete, such as numbers and letters, or it can be continuous, such as sounds, images, and other measurements.”
Digital data comes from everywhere

Entertainment

Education

Health

Commerce

Research

Physical Infrastructure and Smart Systems

Communication / Community
All data are 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
- **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 dispersed and shared
- **Variability**: Extent to which data points differ from each other.

How we access, manage, use, store and preserve data also varies widely

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

- **SIZE / SCALE:**
  - *Small-scale* (KBs, 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 (DIKW)

- **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.*

... Where is the life we have lost in living?

Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information? …

First recorded instance of DIKW in 1934 poem “Choruses from the Rock” by T.S. Eliot.
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 all the digital data come from?

- Digital information comes from many sources: computers, RFIDs and sensors, scientific and other instruments, imaging devices and surveillance, 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
- Areas experiencing increasing data analysis and use:
  - Surveillance footage
  - Embedded and medical devices
  - Entertainment and social media
  - Consumer images
  - Enterprise transactional data; data processing
  - Internet of Things and smart applications
Huge growth in data from the “Internet of Things”

• Major growth spurts in the digital universe:
  – Film $\to$ digital technology
  – Analog functions monitoring and managing physical world $\to$ digital functions involving communications and software telemetry
  – Analogue TV $\to$ 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, roughly 30 billion connected devices in the digital universe

Figure: http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
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 5% of data is “target-rich”. This is predicted to grow to 10+% 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.
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 4370/6370
Digital universe profiles


Type of Information in the Cloud in 2020


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

Portion of DU Needing Protection

EXAMPLES:
- Corporate financial data
- Personally identifiable information (PII)
- Medical records
- User account information

Portion Protected

Portion Not Protected

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.

The information about you is much greater than the information you create yourself. This is called your Digital Shadow. And it’s growing continuously.

Need for data and IT-savvy professionals having tremendous impact on the workforce

From McKinsey Report on Big Data:
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

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 18\textsuperscript{th} / early 19\textsuperscript{th} century.
  – Hand production $\rightarrow$ machines, new chemical manufacturing, new iron production processes
  – Improved efficiency of water power and the increased use of steam power
  – Wood and bio-fuels $\rightarrow$ coal
  – England $\rightarrow$ 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

*Planet image: NASA; Quote from "Our Common Future" http://www.un-documents.net/our-common-future.pdf*
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
- **Used by global organizations including the United Nations Global Compact Cities Programme, The World Association of Metropolises, World Vision, and others.**
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)

- Previous IDC Reports: http://www.emc.com/leadership/digital-universe/index.htm#Archive
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: 7 page research paper on an approved topic
  - Grads: 10 page research paper on an approved topic
- **Class participation** (10 points)
- **1 Op-Ed** (10 points, due before April 29)
- **2 Data Roundtable reviews / presentations** (roughly 1 per section, 10 points each)
Grading Detail – Data Roundtables

Do 2 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 will be given in class and can be found on the class website.

• All written reviews and pptx slides must be turned in before 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.

Fran Berman, Data and Society, CSCI 4370/6370
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 Backdrop:
   - 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. Your thoughts / assessment:
   - If the article is about a report, article, paper or other source material, did they do an accurate job of representing it?
   - What should be done next to explore this issue?
   - 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”?
- Is the “thoughts/assessment” 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 / backdrop?
   • 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?
   • Where would you go from here if you were interested in this topic?

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?
• Is the presentation compelling?
• Does the presentation tell an interesting story?
• Did the speaker use the timeframe effectively?
• 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

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

2. What are the data issues / backdrop?
   • 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?
   • Where would you go from here if you were interested in this topic?
Data Roundtable 1

Internet of Things brings new era of weather forecasting, Computerworld, 1/4/16

You are driving down the road and your smartphone or vehicle alert system announces that black ice is 2.5 miles ahead. As you move closer there’s a countdown: 500 feet, 300 feet, and now urgently: Warning. Warning -- 100 feet to black ice.

How can you know the location of black ice so precisely?

Weather systems will collect data from other vehicles on that same road. The vehicles will wirelessly transmit road condition and weather data. Vehicles’ data points will include barometric pressure, air temperature, windshield wiper settings, and vehicle stability control, or the amount of differential rotation.
The Internet of Things

[Wikipedia] The Internet of Things (IoT) is the network of physical objects or "things" embedded with electronics, software, sensors and network connectivity that enables those objects to collect and exchange data. Each thing is uniquely identifiable and can interoperate within an integrated infrastructure.

The IoT creates opportunities to develop cyber-physical systems that enable "smart applications"
IoT-driven predictions

- IoT-based forecasting can enhance safety and productivity

- “Smart driving”: Potential life-saving benefits to using vehicle sensor data to improve weather information and alert systems for drivers.
  - On average, nearly 6,000 people are killed and over 445,000 people are injured in weather-related crashes each year (Federal Highway Administration)

- “Smart farming”: Use of soil, atmospheric and other conditions, combined with government and private sources for weather data being used to help farmers make better decisions about crops, resource usage
  - Food production will need to increase 70% on the same amount of arable land to feed roughly 9.6 giga-people in 2050

Image from http://www.defensivedriversdiscount.com/driving-on-black-ice/
Precision forecasting ➔
better information on road conditions ➔
safer driving

How it could be done:

• Integrated systems will combine weather data with data collected from other vehicles on the same road
  – Vehicles will wirelessly transmit road condition and weather data
  – Vehicles’ data points could include barometric pressure, air temperature, windshield wiper settings, vehicle stability control, amount of differential rotation between wheels indicating slipper conditions, etc.
  – “Vehicle internet” will permit cars to communicate directly with one another

• Information will be combined, analyzed and reported in real-time to provide precision “now-casting” to warn of hazards

Precision forecasting ➔ precision agriculture ➔ more productivity

- Precision Agriculture aims to optimize field-level management by using comprehensive information about soil, crops, weather, pricing, resources, etc. to improve decision making and productivity [Wikipedia]
  - Data sources include
    - Sampling
    - Remote sensing
    - Proxy detection (e.g. measurement by in-vehicle sensors)
    - Aerial or satellite remote sensing

- More data giving rise to precision forecasting that can “mitigate the risk of weather” and also to capitalize on information about agricultural “micro-climates”
  - Remote monitoring stations (e.g. by Schneider Electronic) being used to measure atmospheric and ground conditions such as soil moisture
  - Agriculture data being combined with government and private sources of weather data used to help develop hourly forecasts to help farmers make better tactical decisions about pesticides, water, fertilizer, etc.
Making better weather predictions

• What used to happen:
  – Before sensor and wireless technology, weather data primarily came from airport weather stations and ships at sea
  – Weather observers recorded conditions such as wind direction and speed, temperature, pressure and pressure trends, dew point and sky cover
  – Professionals plotted data and transmitted it for forecast analysis
  – Forecasters used 100,000’s of weather plots

• Today ++
  – Employment of sensor and wireless technology: More data points, more metadata, more frequent updates, more real-time data. Billions of reference points.
  – More focused analysis and modeling with local climate modeling, more computing horsepower
  – Greater use of terrestrial sensors, satellites, radar, etc.
  – Greater coupling of local environments with large-scale models
Quality of data → Quality of results

Key IT issues:

• **Appropriate Metadata:** Metadata needed to help identify which instruments were used to gather the data and their accuracy so the people who run the forecasting models “can pick and choose data sets that are going to be of value”, William Mahoney, NCAR
  
  – Metadata from different sources should follow same community practices / models and be represented in consistent units.

• **Data Quality and Consistency:** Quality of vehicle data can vary depending on where sensors are located (e.g. proximity to the engine), etc.
  
  – No standards for optimizing sensor placement at present.

• **Adequate Data Models:** Increasing amounts of sensor data means trying to capture the physics correctly at finer resolutions.
  
  – Physical inter-relationships between urban environments, lakes, rivers, streams and other conditions all influence micro-climates.
  
  – Crowd-sourced models need to account for number and integrity of information sources

• **Access to Information:** Variation in local policy or practice may restrict use of information and change the quality of results.
  
  – Who owns information and who has the right to use it?
Precision Forecasting providing opportunities for broader goods and services

**Schneider Electric** – local weather devices

**Precision Weather** – Forecasting as a private service

**Dark Sky** – precision nowcasting app with radar, satellite, sensor, gov't. and crowd-sourced data
References

- **Precision Agriculture**, *Wikipedia*, [https://en.wikipedia.org/wiki/Precision_agriculture](https://en.wikipedia.org/wiki/Precision_agriculture)
- **Precision Weather website**, [http://precisionweather.com/services](http://precisionweather.com/services)
- **The Future of Agriculture – Smart Farming**, *Forbes*, [http://www.forbes.com/sites/federicoguerrini/2015/02/18/the-future-of-agriculture-smart-farming/#2715e4857a0b41f93a0a337c](http://www.forbes.com/sites/federicoguerrini/2015/02/18/the-future-of-agriculture-smart-farming/#2715e4857a0b41f93a0a337c)
References


• **Precision Agriculture**, Wikipedia, [https://en.wikipedia.org/wiki/Precision_agriculture](https://en.wikipedia.org/wiki/Precision_agriculture)

• **Precision Weather website**, [http://precisionweather.com/services](http://precisionweather.com/services)

• **The Future of Agriculture – Smart Farming**, Forbes, [http://www.forbes.com/sites/federicoguerrini/2015/02/18/the-future-of-agriculture-smart-farming/#2715e4857a0b41f93a0a337c](http://www.forbes.com/sites/federicoguerrini/2015/02/18/the-future-of-agriculture-smart-farming/#2715e4857a0b41f93a0a337c)

• Info from Fran
The Internet of Things

[Wikipedia] The Internet of Things (IoT) is the network of physical objects or "things" embedded with electronics, software, sensors and network connectivity that enables those objects to collect and exchange data. Each thing is uniquely identifiable and can interoperate within an integrated infrastructure.

The IoT creates opportunities to develop cyber-physical systems that enable "smart applications"
IoT-driven predictions

- **IoT-based forecasting can enhance safety and productivity**

- **“Smart driving”:** Potential life-saving benefits to using vehicle sensor data to improve weather information and alert systems for drivers.
  - On average, nearly 6,000 people are killed and over 445,000 people are injured in weather-related crashes each year (Federal Highway Administration)

- **“Smart farming”:** Use of soil, atmospheric and other conditions, combined with government and private sources for weather data being used to help farmers make better decisions about crops, resource usage
  - Food production will need to increase 70% on the same amount of arable land to feed roughly 9.6 giga-people in 2050

Image from http://www.defensivedriversdiscount.com/driving-on-black-ice/
Precision forecasting → better information on road conditions → safer driving

How it could be done:

• Integrated systems will combine weather data with data collected from other vehicles on the same road
  – Vehicles will wirelessly transmit road condition and weather data
  – Vehicles’ data points could include barometric pressure, air temperature, windshield wiper settings, vehicle stability control, amount of differential rotation between wheels indicating slipper conditions, etc.
  – “Vehicle internet” will permit cars to communicate directly with one another

• Information will be combined, analyzed and reported in real-time to provide precision “now-casting” to warn of hazards

Precision forecasting → precision agriculture → more productivity

- Precision Agriculture aims to optimize field-level management by using comprehensive information about soil, crops, weather, pricing, resources, etc. to improve decision making and productivity [Wikipedia]
  - Data sources include
    • Sampling
    • Remote sensing
    • Proxy detection (e.g. measurement by in-vehicle sensors)
    • Aerial or satellite remote sensing

- More data giving rise to precision forecasting that can “mitigate the risk of weather” and also to capitalize on information about agricultural “micro-climates”
  - Remote monitoring stations (e.g. by Schneider Electronic) being used to measure atmospheric and ground conditions such as soil moisture
  - Agriculture data being combined with government and private sources of weather data used to help develop hourly forecasts to help farmers make better tactical decisions about pesticides, water, fertilizer, etc.

Fran Berman, Data and Society, CSCI 4370/6370
Making better weather predictions

• What used to happen:
  – Before sensor and wireless technology, weather data primarily came from airport weather stations and ships at sea
  – Weather observers recorded conditions such as wind direction and speed, temperature, pressure and pressure trends, dew point and sky cover
  – Professionals potted data and transmitted it for forecast analysis
  – Forecasters used 100,000’s of weather plots

• Today ++
  – Employment of sensor and wireless technology: More data points, more metadata, more frequent updates, more real-time data. Billions of reference points.
  – More focused analysis and modeling with local climate modeling, more computing horsepower
  – Greater use of terrestrial sensors, satellites, radar, etc.
  – Greater coupling of local environments with large-scale models
Quality of data → Quality of results

Key IT issues:

• **Appropriate Metadata**: Metadata needed to help identify which instruments were used to gather the data and their accuracy so the people who run the forecasting models “can pick and choose data sets that are going to be of value”, William Mahoney, NCAR
  
  – Metadata from different sources should follow same community practices / models and be represented in consistent units.

• **Data Quality and Consistency**: Quality of vehicle data can vary depending on where sensors are located (e.g. proximity to the engine), etc.
  
  – No standards for optimizing sensor placement at present.

• **Adequate Data Models**: Increasing amounts of sensor data means trying to capture the physics correctly at finer resolutions.
  
  – Physical inter-relationships between urban environments, lakes, rivers, streams and other conditions all influence micro-climates.
  
  – Crowd-sourced models need to account for number and integrity of information sources

• **Access to Information**: Variation in local policy or practice may restrict use of information and change the quality of results.
  
  – Who owns information and who has the right to use it?
Precision Forecasting providing opportunities for broader goods and services

**Schneider Electric** – local weather devices

**Precision Weather** – Forecasting as a private service

**Dark Sky** – precision nowcasting app with radar, satellite, sensor, govt. and crowd-sourced data
References

- Precision Agriculture, Wikipedia, [https://en.wikipedia.org/wiki/Precision_agriculture](https://en.wikipedia.org/wiki/Precision_agriculture)
- Precision Weather website, [http://precisionweather.com/services](http://precisionweather.com/services)
- The Future of Agriculture – Smart Farming, Forbes, [http://www.forbes.com/sites/federicoguerrini/2015/02/18/the-future-of-agriculture-smart-farming/#2715e4857a0b41f93a0a337c](http://www.forbes.com/sites/federicoguerrini/2015/02/18/the-future-of-agriculture-smart-farming/#2715e4857a0b41f93a0a337c)
- Info from Fran

Fran Berman, Data and Society, CSCI 4370/6370
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?

   • Precision forecasting can improve safety and productivity
   • Driving: weather-related accidents can be reduced by using more precise information to avoid hazards. Agriculture: greater predictability can help farmer’s utilize resources more efficiently and economically
   • IoT emerging as a useful tool to create smarter and more effective environments

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?

   • More data will necessitate better metadata and more precise analysis algorithms
   • Data software infrastructure, IoT devices and computational platforms will need to support crowd-sourcing, real-time data and very heterogeneous data sources. Adequate data infrastructure will be critical to support accurate forecasting.
   • Stewardship and preservation of relevant data will be critical. Rights will need to be sorted out -- who will own the data?
3. Your thoughts / assessment:

• *If the article is about a report, article, paper or other source material, did they do an accurate job of representing it?*
• *What should be done next to explore this issue?*
• *Did you like the article? Why or why not?*

• *Article could have included more detail and substance about precision forecasting: what kinds of infrastructure and algorithms are needed, why is it hard to deal with heterogeneous data, what are some of the issues around rights and ownership?*
• *I would have liked to see more specifics about how both the driving example and precision agriculture work.*
• *Based on the information in the article, I did some exploration about private sector precision forecasting, which is an interesting area for services.*
• *I liked the article but thought it was a little lightweight*
Your turn: Readings for Lecture 1 Data Roundtable (February 5)

Readings for Lecture 2 Data Roundtable (February 12)


• “Preserving the Internet”, CACM (January 2016), http://cacm.acm.org/magazines/2016/1/195738-preserving-the-internet/fulltext (Dan L.)
Next time: Data Stewardship and Preservation