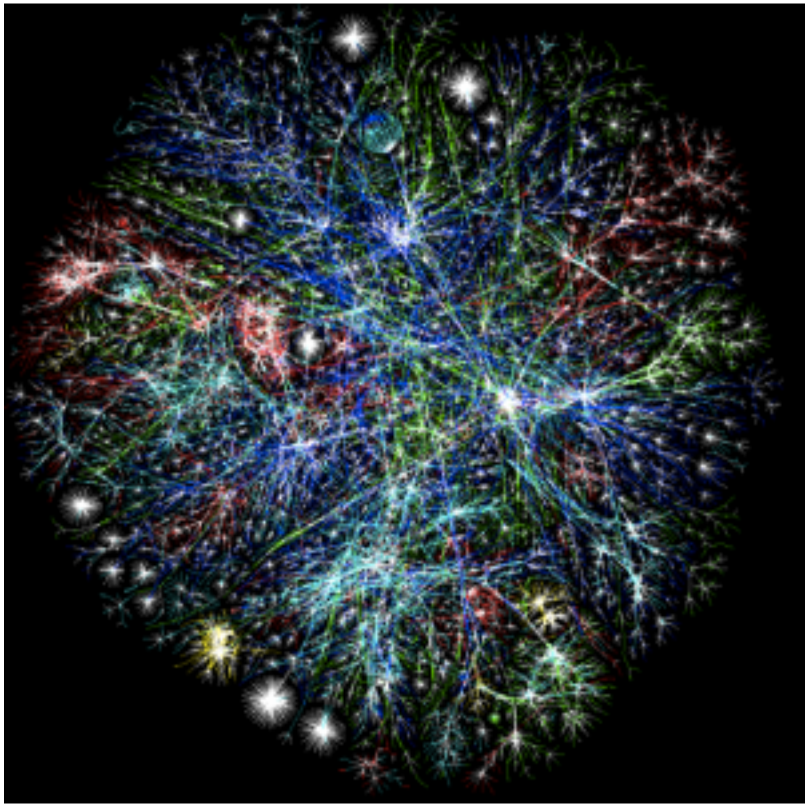
Global Challenges and “Big” Data



Credit: commons.wikipedia.org

Fred S. Roberts

Director of CCICADA

Rutgers University

Command, Control, and Interoperability Center for Advanced Data Analysis

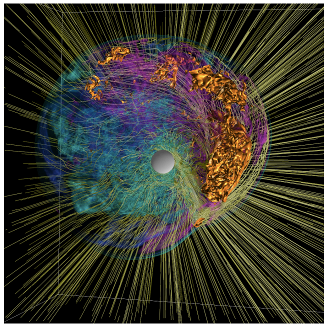
What is Big Data? And How has it Changed?

- Everyone is talking about ***Big Data***
- But what exactly is Big Data?
- Why is it considered so important?
- What about data has changed in the last 5 to 10 years?
- What challenges do we face in the next 10 years?
- How can we use it to address the big challenges we face?

http://www.stat.columbia.edu/~cook/movabletype/archives/data.jpg

What is Big Data?
And How has it Changed?

- ***Massive Data*** has a precise definition
  - Data not fitting into computer memory, thus requiring out of memory algorithms for solving complex problems.
- Big Data has no such definition.
- *Operational definition*: data so large that what to save is at question
  - In some cases, decisions on what to save need to be made instantaneously
  - E.g., astrophysical data



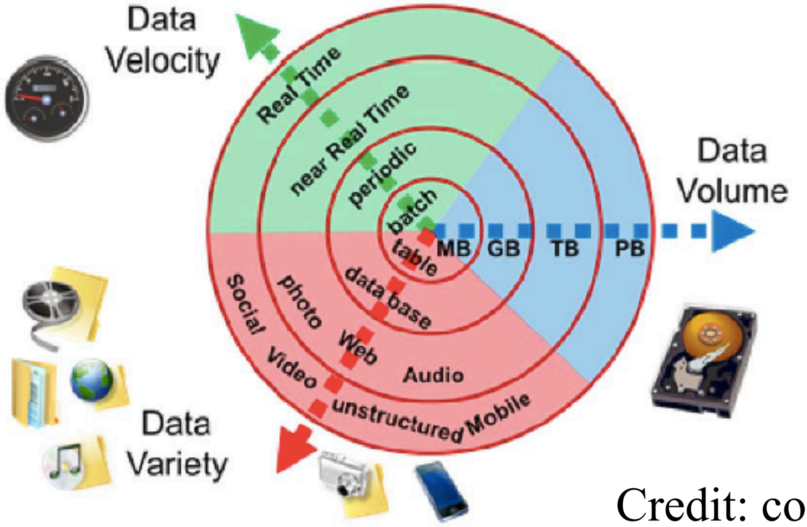
Credit: en.wikipedia.org

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

What is Big Data?
And How has it Changed?

- ***Big Data*** is sometimes described in terms of the three V's
  - ***Volume***
  - ***Variety***
  - ***Velocity***



Credit: commons.wikipedia.org

- It's not just the increase in any one of these factors that has created a challenge, but the concomitant increase in all three.

Command, Control, and Interoperability Center for Advanced Data Analysis

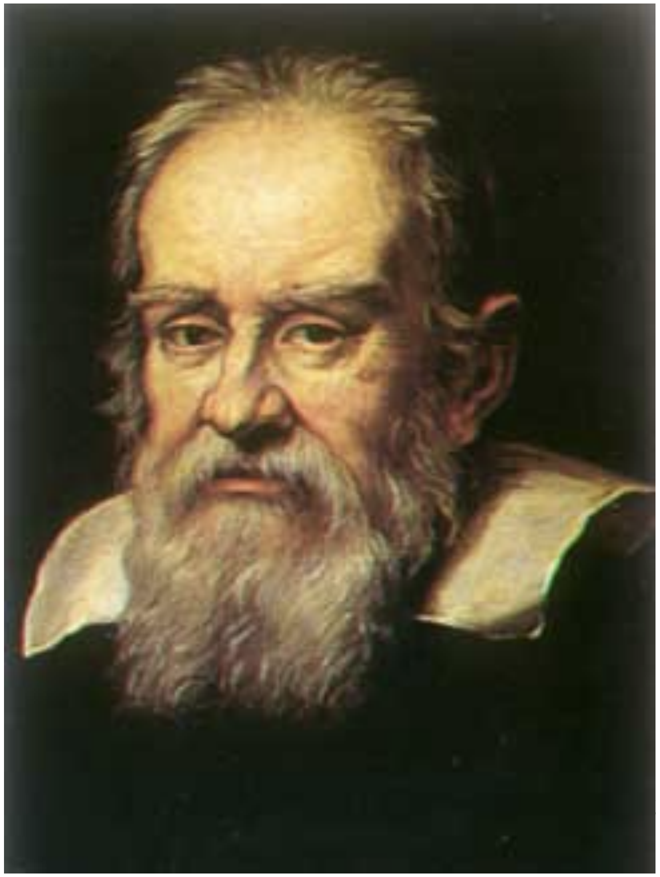
What is Big Data? And How has it Changed?

- It is not just the three V's that define Big Data
- It is something more difficult to define and capture: ***complexity***
  - Data today is very large, heterogeneous, interrelated, and complex
  - Data can be "dirty" (noisy)
  - Data can be "wide" (more variables than cases)
  - Data can be "fuzzy" (involving uncertainty)

Command, Control, and Interoperability Center for Advanced Data Analysis

What is Big Data? And How has it Changed?

- Let us remember: Data science is an old field
- Galileo Galilei was a data scientist – and not the first
- So what has changed?



Credit: en.wikipedia.org

What Leads to Big Data?

- Ever-increasing volumes of sensor data
- Ability to transmit data over ever-higher capacity networks
- Storage devices that can store and retrieve massive amounts of data
- Growing computing power
- The demand for faster solutions to complex problems
- Commercial and government applications

Wide Variety of Sources of Data

- News
- Text
- Audio
- Images
- Video
- LiDAR
- Geophysical analyses
- Sensors of all types
- GPS systems
- Smartphones and tablets

Credit: en.wikipedia.org

The remarkable variety of data sources present ***new challenges*** *for data science*

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

New or Increasing Challenges Facing Big Data

- Global environmental change
- Increasing frequency and severity of natural disasters
- Urbanization and the role of "smart cities"
- And much more

Photos from commons.wikimedia.org

Global Environmental Change

- The planet is constantly changing.
- But the pace of change has accelerated as a result of human activity:
  - Construction and deforestation change habitats
  - Over-fishing reduces wild populations
  - Fossil fuel combustion leads to atmospheric greenhouse gas buildup
  - Commerce and transport introduce non-native species.

Command, Control, and Interoperability
Center for Advanced Data Analysis

Global Environmental Change

- We need to:
  - Monitor global change to understand processes leading to change
  - Learn how to mitigate and adapt to its effects
  - Determine if we are meeting goals for our planet
  - Get early warning of dangerous trends

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

Global Environmental Change

- ***The Age of Observation***:
  - The unprecedented amount of data about health of the planet provides great opportunities but also poses immense challenges
  - How do we choose what to observe and what to save?
  - What are appropriate sampling and monitoring designs?
  - How to reconcile so many different variables with so many different spatiotemporal characteristics?

CCICADA
Command, Control, and Interoperability Center for Advanced Data Analysis

Global Environmental Change

- ***Metrics for Global Change***:
  - Measuring global change & making better policy decisions requires metrics of planetary health.
  - Example: How to measure biodiversity? The Convention on Biodiversity: goal of achieving a significant reduction in biodiversity loss by 2010. But how could we measure progress?
  - Example: How can we measure the impact of sea level rise for adaptation planning?

Photo: Fred Roberts

Photo: Fred Roberts

CCICADA
Command, Control, and Interoperability Center for Advanced Data Analysis

Global Environmental Change

- ***Effects of Global Change*:**
  - Goal is not so much to describe the many effects of global change as to understand:
    - ➢ Interface between change in one sector on another – e.g. in understanding Lyme Disease spread into Canada, need understand tick life cycles, bird migrations, climate change
    - ➢ Risk-based comparison of alternative adaptation and mitigation strategies – e.g., for control of invasive species or for more severe weather.

Command, Control, and Interoperability Center for Advanced Data Analysis

Natural Disasters

- No part of the world is impervious to natural disasters
  - Epidemics
  - Earthquakes
  - Floods
  - Hurricanes
  - Tornadoes
  - Wildfires
  - Tsunamis
  - Extreme temperatures
  - Drought
  - Oil spills

Nepal 2015: www.circleofblue.org

- Data science can help in predicting, monitoring, and responding to such events, and mitigating their effects.

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

Hurricanes

Hurricanes Harvey, Irma, Maria hit US in 2017

Maria – Puerto Rico

Photo credit: Yale Daily News

Harvey – Houston, Texas

Photo credit: Wikimedia Commons

Irma - Key West Florida

Photo credit: Wikimedia Commons

Command, Control, and Interoperability Center for Advanced Data Analysis

Hurricanes

SuperStorm Sandy Hits New Jersey Oct. 29, 2013

My backyard

My Neighborhood

Photo credits: Fred Roberts

Command, Control, and Interoperability Center for Advanced Data Analysis

Hurricanes

SuperStorm Sandy Hits New Jersey Oct. 29, 2013

NJ Shore – from Jon Miller

Hurricanes

- Will data science help us plan for the future?
  - What subways will be flooded?
  - How can we protect against such flooding?

Hurricanes

- Will data science help us plan for the future?
  - What power plants or other facilities on shore areas will be flooded?
  - Do we have to move them?

Hurricanes

- Will data science help us plan for the future?
  - How can we get early warning to citizens that they need to evacuate?
  - How can we plan such evacuations effectively?

Hurricanes

- Will data science help us plan for the future?
  - How can we plan placement of utility lines to minimize down time?

Photo credits: Fred Roberts

Hurricanes

- Will data science help us plan for the future?
  - How can we plan for getting people back on line after a storm?

Bringing in help from far away

Photo credits: Fred Roberts

Command, Control, and Interoperability Center for Advanced Data Analysis

Hurricanes

- Will data science help us plan for the future?
  - –How can we set priorities for cleanup?

Photo credits: Fred Roberts

Command, Control, and Interoperability Center for Advanced Data Analysis

Floods

- Which flood mitigation projects to invest in?
  - Buyouts
  - Better flood warning systems
  - “Green infrastructure” (cisterns & rain barrels)
  - Pervious concrete
  - Etc.

Raritan River flood
Bound Brook, NJ
August 2011

Floods

- This requires data-driven/Model-driven Decision Support
- Data-driven. Assemble data about:
  - Precipitation (duration, amount)
  - Antecedent conditions (soil moisture content, ground cover, seasonality)
  - River gage levels
  - Flood maps
  - Property damage data – payouts from Federal Emergency Management Agency

Disease Events

- Newly emerging diseases can threaten the health of millions of people.
- The 1918 influenza epidemic killed 50 million people around the world; WW I killed 16 million
- Great concern about the potential for a new influenza outbreak of similar proportions

Source: Dartmouth Medicine 2006

Command, Control, and Interoperability
Center for Advanced Data Analysis

Disease Events

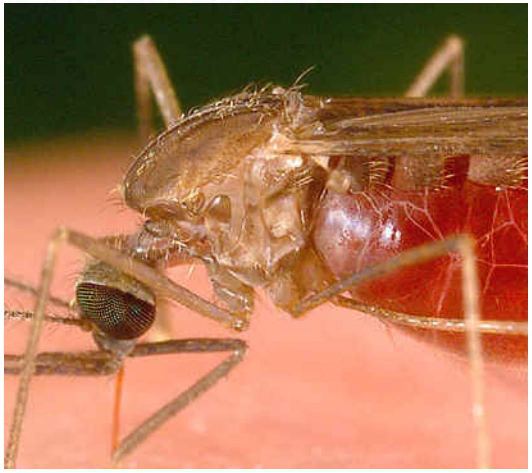
- Modern transportation systems make it much easier for diseases to spread around the world – e.g., Ebola
- Deliberate introduction of diseases by bioterrorists is a serious concern
- Climate change leads to diseases appearing in places they have not appeared in before – e.g., malaria in the highlands of Kenya & potential for malaria in the US

HEALTH ADVISORY: EBOLA
Recently in West Africa?
Watch for fever, headaches, and body aches in the next 3 weeks.
3 WEEKS
If you get sick, call a doctor.
Tell the doctor where you traveled.
For more information: visit www.cdc.gov/travel or call 800-CDC-INFO.
CDC

Disease Events: Data Science

- Syndromic surveillance
- Monitoring
  - Prescriptions for drugs
  - Hits on medical advice websites
  - Hospital admissions
  - School absences
  - Etc.

Dengue Fever



Malaria mosquito

CCICADA
Command, Control, and Interoperability
Center for Advanced Data Analysis

Oil Spills

- The Deepwater Horizon oil spill Gulf of Mexico (2010):

Oil Spills

- With climate change, more vessel traffic in the Arctic and possibility of offshore oil drilling
- Increased risk of spills from vessels or drilling
- Arctic challenges: resource allocation in advance in case of oil spill
  - Necessary because of long transit times, lack of infrastructure, remote locations, lack of roads, distant airlift
  - *"The lack of infrastructure and oil spill response equipment in the U.S. Arctic is a significant liability in the event of a large oil spill" (National Academies, 2014).*
  - Need data about existing equipment, transportation, communication, etc.

http://www.dhs.gov/xlibrary/photos/ snapshots/st-ada-250x250.jpg

“Smart Cities”

- In 1900, only 13% of the world’s population lived in cities.
- By 2050, it is predicted that 70% will.
- As rapid city expansion continues, data science can play key roles in shaping sustainable living environments

http://citadel.sjfc.edu/students/rnr00577/e-port/msti260/urban.htm

“Smart Cities”

“With recent advances in technology, we can infuse our existing infrastructures with new intelligence, … digitizing and connecting … [to] sense, analyze and integrate data, and respond intelligently to the needs of their jurisdictions. In short, we can revitalize them so they can become smarter and more efficient.” − IBM Smarter Planet Initiative

IBM SmarterPlanet. http://www.ibm.com/smarterplanet/us/en/?ca=v_smarterplanet, 2012.

Building a Smarter City and State

The Commonwealth of Massachusetts, The City of Boston and IBM are working together to transform the region's physical infrastructure, engage citizens, reduce costs and improve efficiency. Do you know where technology is at work where you live?

3
4
1
2
5
FINISH
6

1 **Buildings:**
The state of Massachusetts owns 72 million square feet of property. Software helps improve maintenance, space and management across public sector buildings.

2 **Traffic:**
Approximately 1.9 million commuters travel by car a day in Boston. Officials examine how Big Data technology makes transportation more efficient and reduce pollution.[1]

3 **Airport:**
Tens of millions of travelers pass through Logan Airport every year. Software helps the Port Authority better manage maintenance operations for equipment such as air conditioning, doors and escalators at Terminal A.[2]

4 **Physical Assets:**
Boston has more than 60,000 streetlights[3] and 13,000 fire hydrants.[4] Software helps city officials better manage and maintain physical assets.

5 **Special Events:**
More than half a million people attend events such as the Boston Marathon and July 4th fireworks every year. Software can integrate and visualize critical information across city departments including fire, police and emergency responders to help coordinate and plan special events.[5]

6 **Water:**
Massachusetts Water Resource Authority (MWRA) serves 2.5 million people in 61 communities.[6] Using software, MWRA decreased corrective maintenance and project work orders by 38 percent.

[1] "Boston ranked fifth most traffic-prone city in nation," Daily Free Press: http://dailyfreepress.com/2013/02/11/boston-ranked-fifth-most-traffic-prone-city-in-nation/
[2] About Logan International Airport: http://www.massport.com/logan-airport/about-logan/pages/default.aspx
[3] Street Lighting Division for the City of Boston
[4] Currents Newsletter, July-August 2011: http://www.bwsc.org/notices/public_notices/CUR_2011_4_JUL_AUG.pdf
[5] " BAA Offers Runner Defferral: Wait till Next year, Boston.com: http://www.boston.com/sports/marathon/articles/2012/04/17/baa_offers_runners_deferral_wait_till_next_year/
[6] About Massachusetts Water Resource Authority: http://www.mwra.state.ma.us/02org/html/whatis.htm

For more information, please visit: www.IBM.com/press/smarterplanet

The Role of Data in “Smart Cities”

- Some examples where data science is needed:
  - Smart systems to reduce congestion and pollution thru traffic prediction and optimization
  - Real-time rerouting of commuting passengers
  - Vehicle sharing systems
  - Energy Management
  - Water Management
  - Public health
  - Safety and security
  - Keeping citizens informed of municipal services (especially during disasters)

The Role of Data in "Smart Cities"

- Challenge: Find ways to use data to:
  - Make smarter, more livable cities
  - Understand patterns driving human behavior
  - Understand causes of the state of the urban environment
  - Learn how to optimize our choices.

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

The Role of Data in "Smart Cities"

- Planning Urban Transportation Systems:
  - What routes to create or expand?
  - Determining route capacities
  - Good combinations of transportation modes
  - Determining location and capacity of stations for bike sharing, zipcars, electric vehicle charging stations
  - How sensitive is plan to origin-destination data?
  - What is effect of population growth?
  - Of changes in built environment?
  - How sensitive is plan to people's transportation mode preferences?

Credit: Francesco Calabrese
Smarter Cities Technical Centre
IBM – Ireland; USDA Farm Service
Agency; Mass GIS, Commonwealth
of Massachusetts EOEA

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
  - Example: Flash flood prediction
    - ➢ Rain gage networks
    - ➢ Radar
    - ➢ Satellite algorithms
    - ➢ Computer models of atmospheric processes
    - ➢ Hydrological models

Credit: en.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
  - Example: Earthquake prediction (still speculative) – fusing information from:
    - ➢ Changes in Vp/Vs (velocity of primary wave over velocity of secondary wave)
    - ➢ Spikes in concentration of gases such as radon
    - ➢ Seismic electric signals (geoelectric voltages)
    - ➢ Accelerating cumulative # of foreshocks
    - ➢ Anomalous animal behavior

Haiti; credit: commons.wikipedia.org

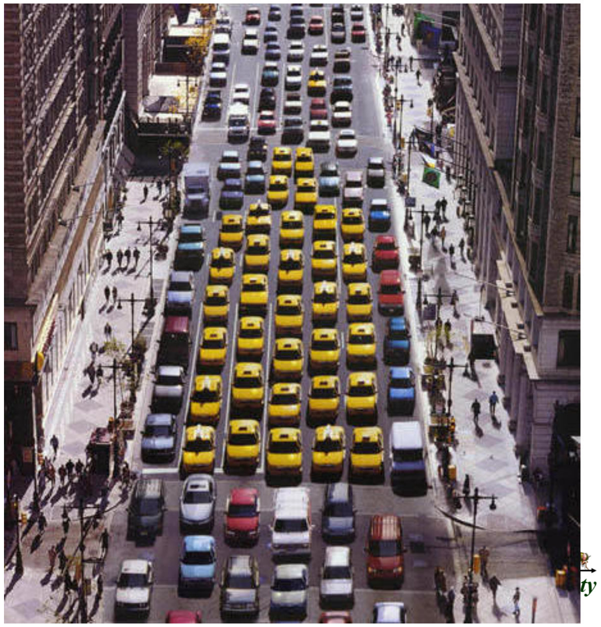
New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
  - Example: How to combine "hard" numerical readings of sensors monitoring emergency vehicle movements with "soft" natural language utterances of the driver and "tweets" of the public?

Credits: commons.wikipedia.org, flickr.com

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
- Example: Traffic Management in Smart Cities*
- Some data sci challenges: intelligent transportation systems:
  - Integrated fare management
  - Road usage charging
  - Traffic information management



*Most of ideas on traffic management taken from talk "Smart Cities – How can Data Mining and Optimization Shape Future Cities," by Francesco Calabrese of IBM Ireland at DIMACS workshop on Smart Cities, Paris, Sept. 2011

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
- Real-time road traffic management:
  - Key role of sensors
  - Monitor actual traffic situation (volumes, speeds, incidents)
  - Control or influence the flow using that information to:
    - ➢ Reduce traffic congestion
    - ➢ Deal with incidents
    - ➢ Provide accurate information to drivers and authorities
    - ➢ Grant proper authority/routing to emergency vehicles

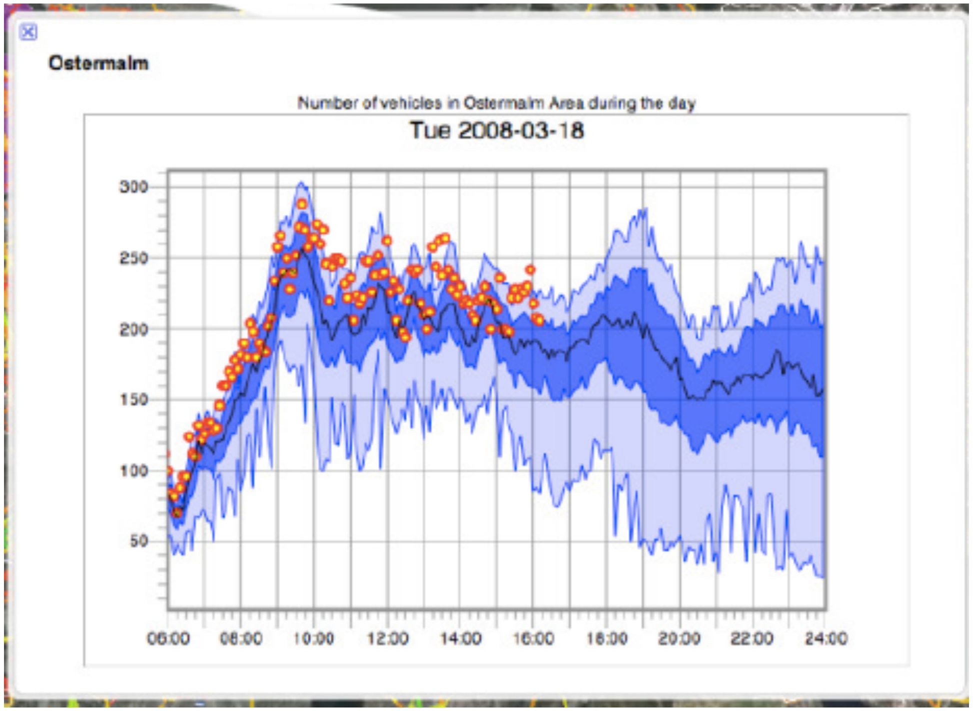
Credit: Francesco Calabrese
Smarter Cities Technical Centre
IBM - Ireland

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
- Real-time road traffic management: For sensor data, GPS data is key.
- Data Sci challenges:
  - GPS data needs to be related to underlying network (road or rail) by map matching algorithms that are computationally expensive
  - GPS data sampled at irregular intervals, possibly with large gaps – requires advanced analytics to reconstruct GPS trajectories
  - GPS data inaccurate, needs “cleaning”

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
- Real-time road traffic management
- Other sensor data sci challenges:
  - Real-time speed estimation
  - Estimated heading
  - Real-time traffic information



Credit: Francesco Calabrese
Smarter Cities Technical Centre
IBM - Ireland

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
- Real-time traffic management: Understanding human transit demands/needs: for real-time or planning purposes
- Data sci challenges:
  - Analyze transit needs in short and long term
  - Help citizens to navigate the city
  - Design adaptive urban transportation systems
  - Detect and predict travel demand
  - Offer real-time alternative routings
  - Improve event planning and management: predict effect of an event on urban transportation

Credit: Francesco Calabrese Smarter Cities Technical Centre IBM - Ireland

New Challenges for Data Science

- **Fusion Challenge**: Fusing information from multiple media or sources
- Understanding human transit demands/needs: Example: Closure of Pulaski Skyway in NJ.
  - Key commuter route into NYC
  - Traffic modeling – NJ Dept. of Transportation
  - Leads to idea of employers shifting work hours
  - Leads to idea of usage of shoulder on NJ Turnpike
  - But: need real-time rerouting through signs, social media, etc.
  - Data science challenge: what if everyone responds to same routing guidance???

Credit: Wikipedia

New Challenges for Data Science

- **Decision Support Challenge**
  - Today’s decision makers have available to them remarkable new technologies, huge amounts of information, ability to share information at unprecedented speeds and quantities.
- **Decision Support Challenge**: These tools and resources will enable better decisions if we can surmount some of the major challenges
  - Data often incomplete or unreliable or distributed, and involves great uncertainty
  - Many sources of data need to be fused into a good decision, often in a remarkably short time

Credit: www.bluediamondgallery.com

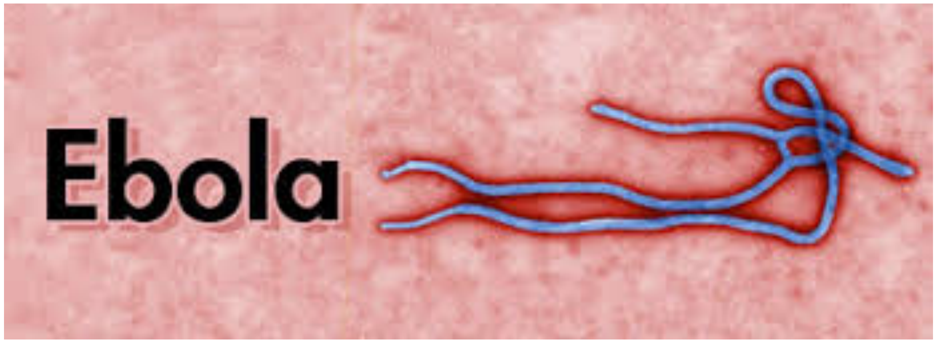
New Challenges for Data Science

- **Decision Support Challenge:** These tools and resources will enable better decisions if we can surmount some of the major challenges
  - Interoperating/distributed decision makers and decision-making devices need to be coordinated
  - Decisions must be made in dynamic environments based on partial information
  - There is heightened risk due to extreme consequences of poor decisions
  - Decision makers must understand complex, multidisciplinary problems

New Challenges for Data Science

- **Decision Support Challenge**
- Case in point: The 2014 Ebola outbreak in West Africa
- Outbreak reminded us: The world is ill-prepared for a severe disease epidemic.
- The risk of future global severe infectious disease outbreaks in an increasingly connected world is greater than ever.

Kidshealth.com



New Challenges for Data Science

- **Decision Support Challenge**
- Case in point: The 2014 Ebola outbreak in West Africa
- The successful fight to contain the outbreak was helped by application of data analysis and mathematical models
- Accurately predicted how and where the disease was spreading and how to contain it
- Data allowed decision makers to understand things like: how many beds and lab tests would be needed — and where and when to deploy them

Credit: Doctors Without Borders

New Challenges for Data Science

- **Decision Support Challenge**
- Case in point: The 2014 Ebola outbreak in West Africa
- Important to the success of the Ebola containment: the sheer and unprecedented magnitude of epidemiological data made available.
  - Online to researchers and modelers
  - By the World Health Organization and health ministries of the most affected countries.
- Though modelers have analyzed ongoing epidemics before, such as the 2003 SARS epidemic and 2009 Swine Flu pandemic, they did not have access to such rich sources of data.

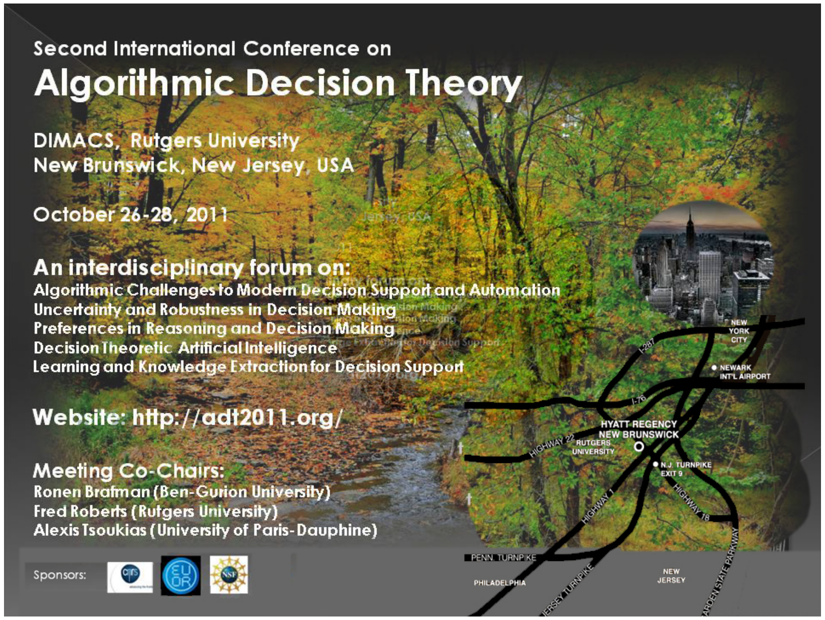
New Challenges for Data Science

- **Decision Support Challenge**
- Case in point: The 2014 Ebola outbreak in West Africa
- Data fed into models showed we could stop this outbreak if 70 percent of Ebola cases could be placed in Ebola treatment units, had effective isolation, and had safe burials.

Safe burial practices were key to the containment of Ebola in West Africa. *Photo credit: UNMEER, Flickr, Creative Commons*

New Challenges for Data Science

- **Decision Support Challenge**
  - Allow comparison of array of alternative solutions
  - Using data to make decisions is not new
  - Big data has led to using many different techniques to make better decisions
- Resulting new field: Algorithmic Decision Theory



Second International Conference on

Algorithmic Decision Theory

DIMACS, Rutgers University
New Brunswick, New Jersey, USA

October 26-28, 2011

An interdisciplinary forum on:
Algorithmic Challenges to Modern Decision Support and Automation
Uncertainty and Robustness in Decision Making
Preferences in Reasoning and Decision Making
Decision Theoretic Artificial Intelligence
Learning and Knowledge Extraction for Decision Support

Website: http://adt2011.org/

Meeting Co-Chairs:
Ronen Brafman (Ben-Gurion University)
Fred Roberts (Rutgers University)
Alexis Tsoukias (University of Paris-Dauphine)

Sponsors:

Command, Control, and Interoperability
Center for Advanced Data Analysis

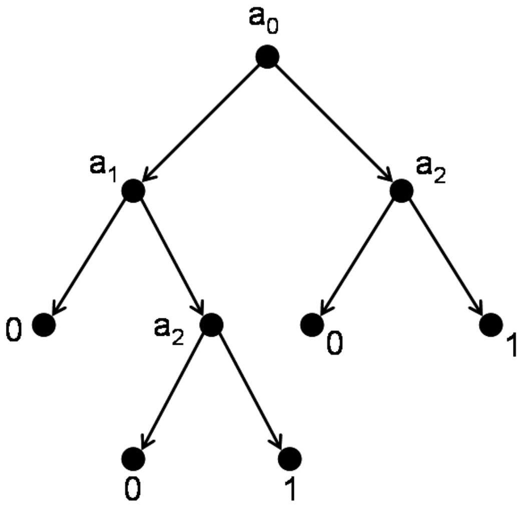
New Challenges for Data Science

- **Combinatorial Explosion Challenge**
  - Big data allows comparison of array of alternative solutions
  - However, the number of alternatives is often so large that we cannot take all into account in a timely way
  - We may not even be able to express all possible preferences among alternatives – too many alternatives
    - Example: “composite” auctions lead to “NP-complete” allocation problems; determining the “winner” can be computationally intractable

New Challenges for Data Science

- **Combinatorial Explosion Challenge**
  - Example: container inspection at ports
    - ➢ Sequential diagnosis: tests one at a time; next test chosen based on outcome of previous test
    - ➢ Represent possible tests as binary decision trees
    - ➢ Find "optimal" BDT
    - ➢ With 5 possible tests there are 263,515,920 possible BDTs

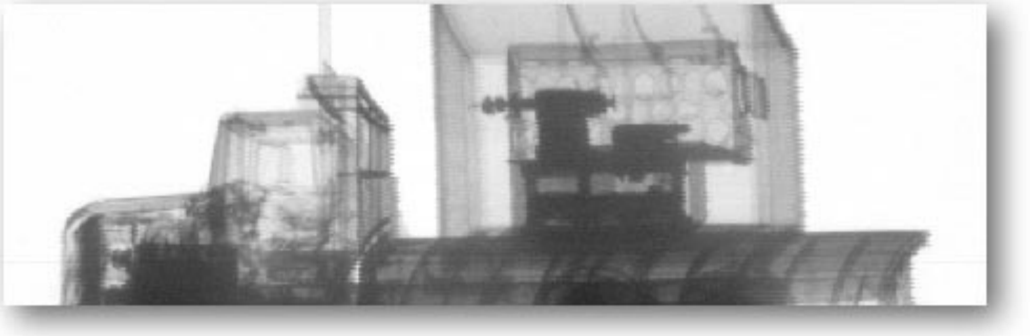
Credit; en.wikipedia.org



Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Combinatorial Explosion Challenge**
  - Example: Comparing performance of nuclear detection algorithms
  - Many relevant factors:
    - ➢ Type of Special Nuclear Material
    - ➢ Shielding
    - ➢ Masking
    - ➢ Altitude
    - ➢ Humidity
    - ➢ Temperature
    - ➢ Vehicle speed
  - Each has several values
  - Too many combinations to test all

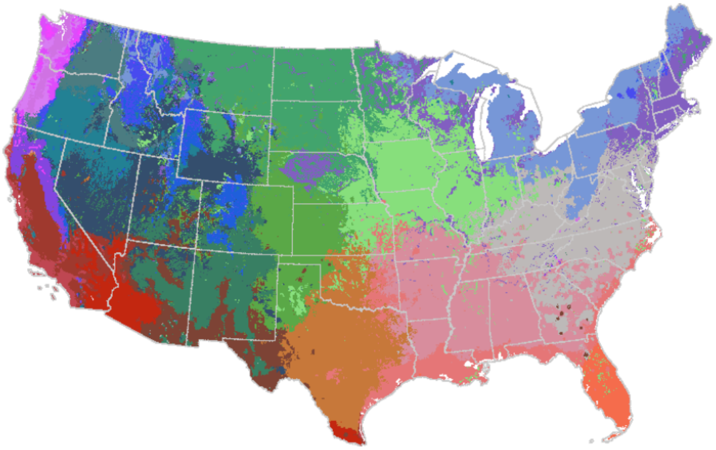


Credit: en.wikipedia.org

CCICADA
Command, Control, and Interoperability
Center for Advanced Data Analysis

New Challenges for Data Science

- **Combinatorial Explosion Challenge**
  - Example: Environmental Monitoring
  - National Ecological Observatory Network (NEON) collecting data at 20 sites across the U.S.
    - ➢ Goal: get a continent-wide picture of the impacts of climate change, land use change and invasive species on natural resources, and biodiversity



Credit: William Hargrove, U.S. Forest Service.

CCICADA
Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Combinatorial Explosion Challenge**
  - Example: Environmental Monitoring
    - How choosing 20 sites?
      - Divide the country into 8 million patches
      - For each patch, collect 9 pieces of information about its ecology and climate
      - Cluster the patches
      - Choose representative patch for each cluster
      - Better would be to use 100 pieces of information
      - But: combinatorially impossible

New Challenges for Data Science

- **Real-time Analytics Challenge**
  - How to make decisions based on data arriving so fast humans cannot absorb it?
    - Example: Power grid
      - Status upgrades used to be every 2-4 seconds, now 10 times a second
      - Rate too rapid for human alone to absorb anomaly in time to act
      - Need software agents to act on behalf of humans

Credit: commons.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Real-time Analytics Challenge**
  - How to make decisions based on data arriving so fast humans cannot absorb it?
    - ➢ Example: Dutch flower auctions
    - ➢ Flowers very perishable; need quick decisions
    - ➢ Typical transaction takes ~ 4 seconds
    - ➢ Information technology allows complex auctions with many bidders
    - ➢ Even determining the winner can be computationally intractable (NP-hard)

Credit: en.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Streaming Data Challenge for Graphs & Networks**
  - Data such as IP traffic level, access logs, command logs arise from rapidly evolving graphs & networks
  - Situational awareness requires us to translate the data into large, interpretable, & manageable graphs
    - ➢ Graphs that can be monitored to detect local changes that may not have a visible effect on global metrics

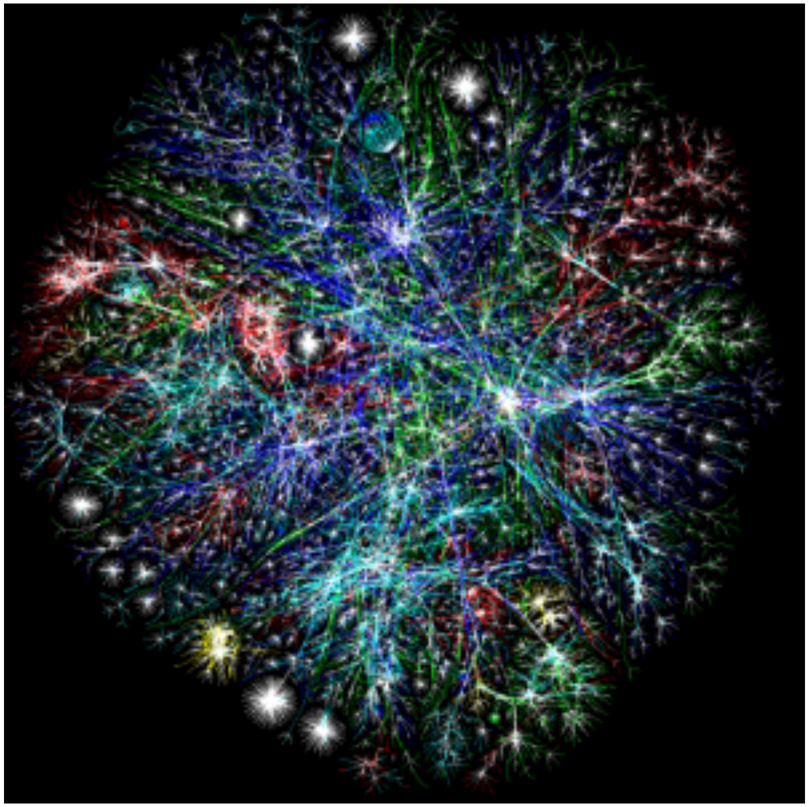
Credit: en.wikipedia.org

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Streaming Data Challenge**: New algorithms needed to deal with large and possibly massive graphs streaming in real time



Credit: commons.wikipedia.org

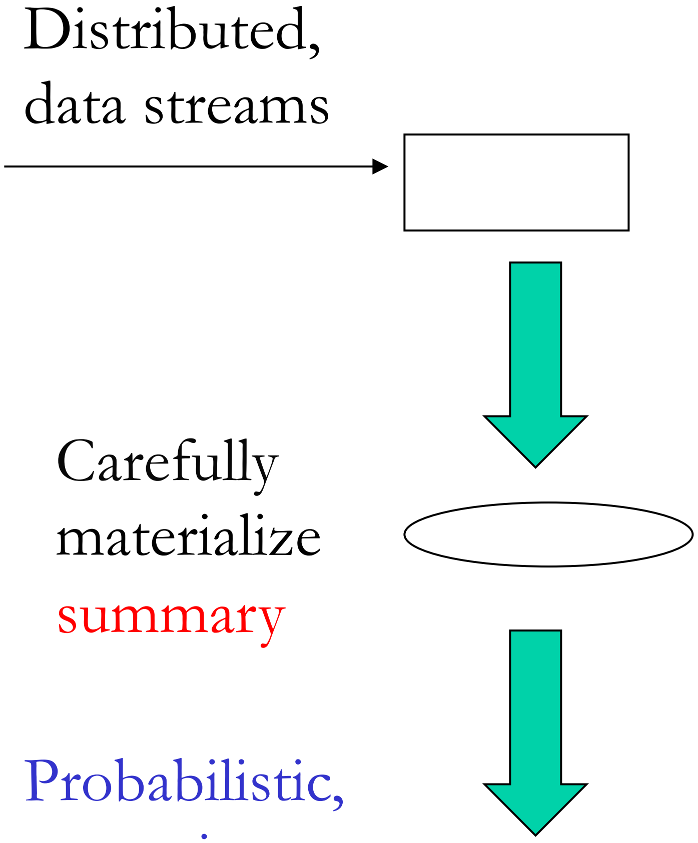
Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Data Summarization**

  Challenge: How to summarize data without being able to store individual items, in a way that allows one to uncover patterns from the summaries?

  - Data is gone, only summaries remain
  - Identifying patterns might not have been in areas of interest at time summaries are produced.
  - Can we use the summaries to get at causality, to aid in post-event mitigation or prevention of future events?



Distributed, data streams

Carefully materialize summary

Probabilistic, approximate ad hoc queries & historic analyses

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Vulnerabilities Challenge**
- Modern society is critically dependent on Big Data
  - Manufacturing and production
  - Power and water systems
  - Financial systems

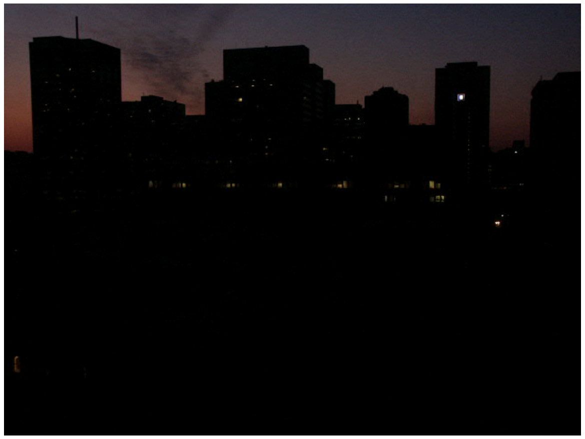
Credit: en.wikipedia.org

Credit: commons.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Vulnerabilities Challenge**
- Modern society is critically dependent on Big Data
- Vulnerabilities are ever present
  - Cyber attacks
  - Cascading failures
  - Rapid spread of anomalies



NYC Blackout 2003
Credit: en.wikipedia.org

Credit: www.flickr.com

Command, Control, and Interoperability
Center for Advanced Data Analysis

New Challenges for Data Science

- **Vulnerabilities Challenge**
- The very ability to utilize and benefit from large amounts of data creates vulnerabilities
  - Electronic medical records lead to hospitals being subject to "ransomware"

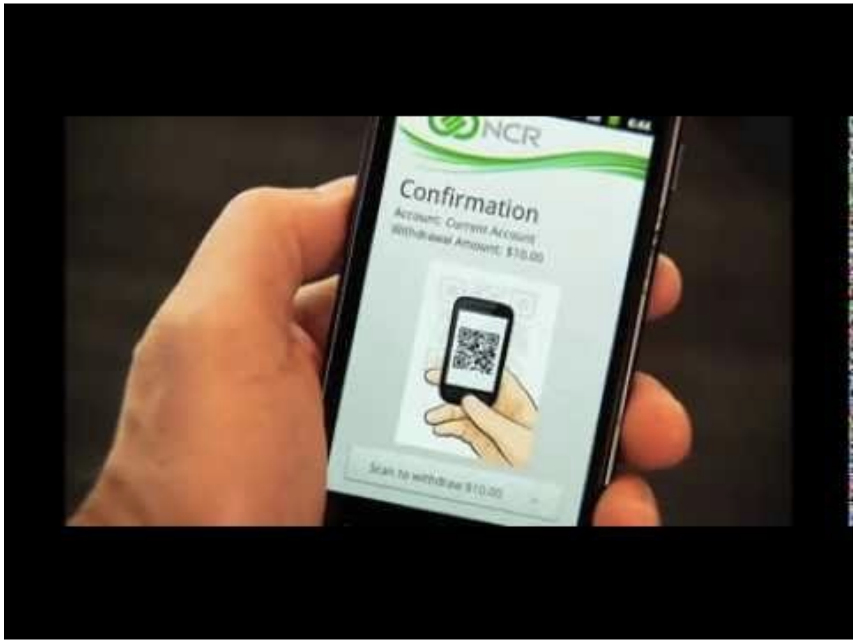
Surgeries in Hospitals Postponed Because of Ransomware

Credit: Community.spiceworks.com

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Vulnerabilities Challenge**
- The very ability to utilize and benefit from large amounts of data creates vulnerabilities
  - Ability to do banking from anywhere we travel leads to identity theft



Credit: www.youtube.com

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Vulnerabilities Challenge**
- The very ability to utilize and benefit from large amounts of data creates vulnerabilities
  - Example: Cyber-physical systems are vulnerable
  - Our cars are now computers on wheels, yet we can already hack into them and "take" control
  - Hacking into a Prius:

Credit: npr.org

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science

- **Vulnerabilities Challenge**
- The very ability to utilize and benefit from large amounts of data creates vulnerabilities
  - Example: Big data allows self-driving cars
  - But those cars can get into accidents

Recent crash of Tesla:
Credit: en.wikipedia.org

CCICADA
Command, Control, and Interoperability Center for Advanced Data Analysis

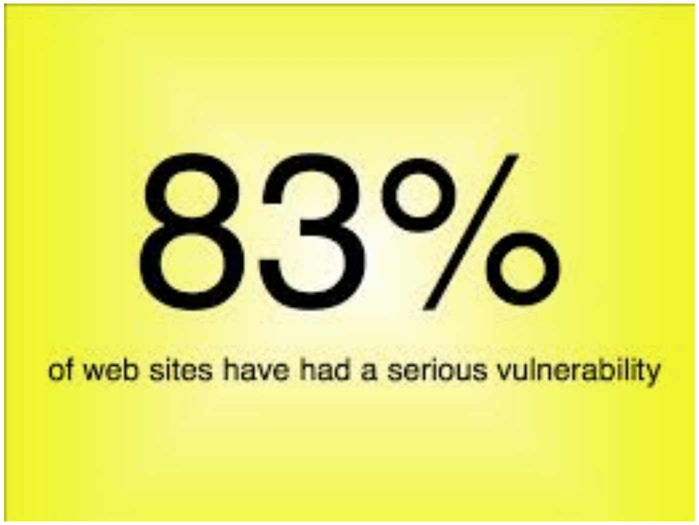
New Challenges for Data Science

- **Vulnerabilities Challenge**
- The very ability to utilize and benefit from large amounts of data creates vulnerabilities
  - Example: Oil drilling rigs can operate effectively thanks to dynamic positioning systems
  - However, hackers have tilted an oil rig, putting it out of business for days

Credit: www.peakoil.net

New Challenges for Data Science

- **Vulnerabilities Challenge**: How do we identify vulnerabilities caused by usage of data? How do we develop tools for monitoring and minimizing such vulnerabilities?



83%

of web sites have had a serious vulnerability

Credit: www.flickr.com

CCICADA

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science: Information from Data

- A key challenge is to aggregate data from multiple sources with potentially questionable quality and credibility and obtain useful “information” as a result.
- Turning to challenges related to getting “information” from data.

Credit: www.flickr.com

New Challenges for Data Science: Information from Data

- **Information Access Challenge**: How to develop high-accuracy information search and access capabilities
  - Google already does this
  - But what are the next new ideas?
  - One approach: develop special “extraction” technology combined with machine learning to learn the “story” being told across multiple dimensions of time and space.

New Challenges for Data Science: Information from Data

- **Information Distillation**

  Challenge: How to make inferences and draw hypotheses from large amounts of data, when data seldom exists in the form most suited for analysis?

  – Application: how to define "normal" in order to detect departure from normal?

  – Example: what is "normal" seismic activity?

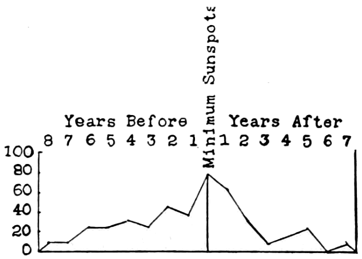


Figure 2. Simple Frequency - Sayles.

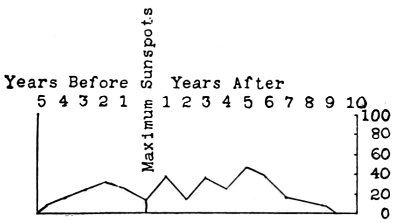


Figure 6. Simple Frequency - Sayles.

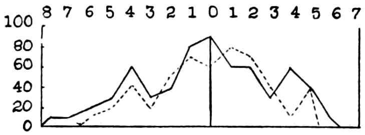


Figure 3. Simple Frequency - Jensen.

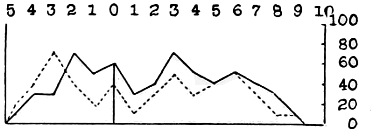


Figure 7. Simple Frequency Jensen

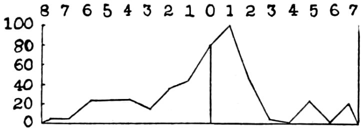


Figure 4. Intensity - Sayles.

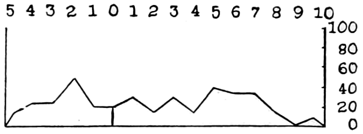


Figure 8. Intensity - Sayles.

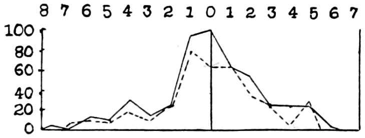


Figure 5. Intensity - Jensen.

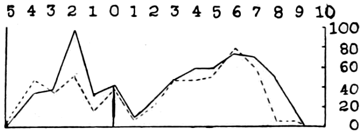


Figure 9. Intensity - Jensen.

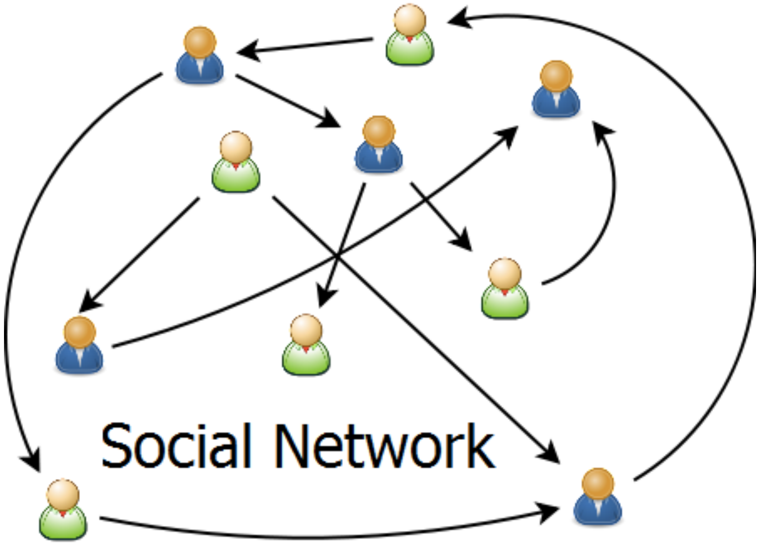
Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science: Information from Data

- **Information Storage & Management Challenge:** How to create very large-volume databases that support data homogenization across various sources?
  - Application: Data evolves, reflecting changing points of view, opinions, environmental conditions
  - How do you follow the development dynamics of adversarial views on a topic, an interest in a technology, or an opinion?

New Challenges for Data Science: Information from Data

- **Information Storage & Management Challenge:** How to create very large-volume databases that support data homogenization across various sources?
  - Example: Can you predict evolving connections in social networks?



Credit: commons.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science: Information from Data

- **Information Storage & Management Challenge**
- Case in Point: Logistics: Supply chains/stockpiles
- What supplies are needed during an emergency?
  - Water, Food, Fuel, Generators, Chainsaws?
- How and where can we stockpile them?
- What are good methods for getting these to those who need them in an efficient way?
- How can we better utilize the private sector?

New Challenges for Data Science: Information from Data

- **Information Storage & Management Challenge**
- Case in Point: Logistics: Supply chains/stockpiles
- How can we tell who needs what kinds of goods during an emergency?
- How can we locate stockpiles so as to be “agile” in allocating the resources when needed?
- E.g.: U.S. Centers for Disease Control & Prevention strategic national stockpile of medicines for emergencies: how do we decide what medicines to include, how many doses, where to keep them?

Source: cdc.gov

New Challenges for Data Science: Information from Data

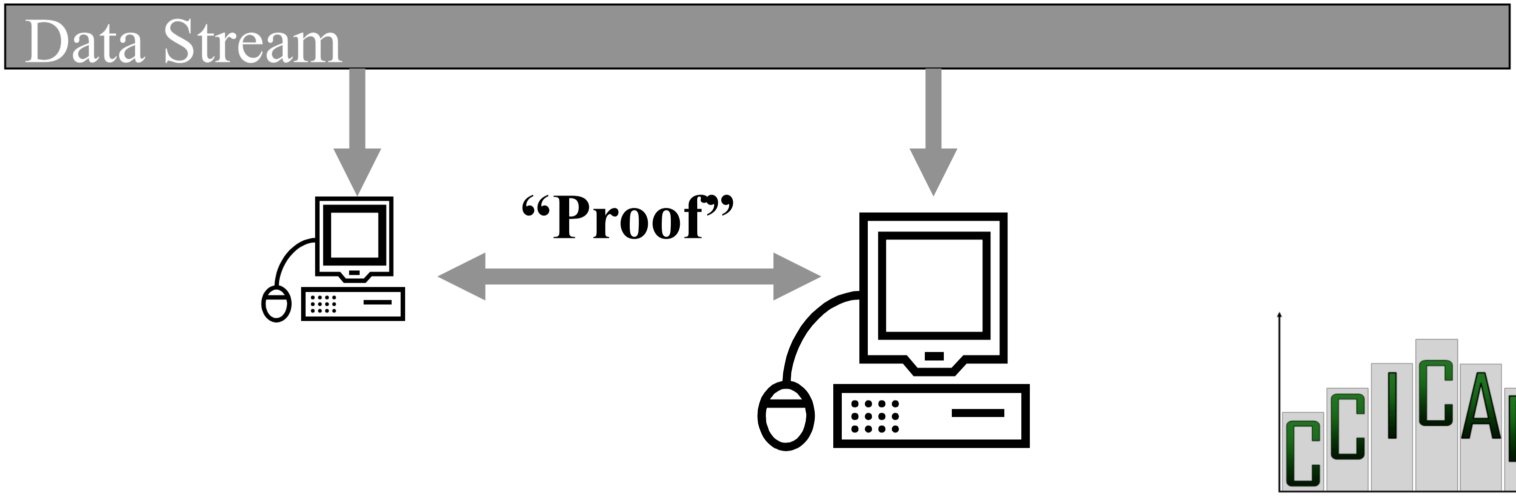
- **New Architectures Challenge:** As much data has grown too large to reside in one location, need new architectures
- Big change in this direction: increasing emphasis on use of "the Cloud" to do computations, store data

Credit: commons.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science: Information from Data

- As more computation is outsourced to a potentially untrusted third party party (“the cloud”), it is now necessary to seek assurances that computations are performed correctly as claimed
- ***“Proof systems***” can give the necessary assurance, but prior work on them is not sufficiently scalable or practical



Data Stream

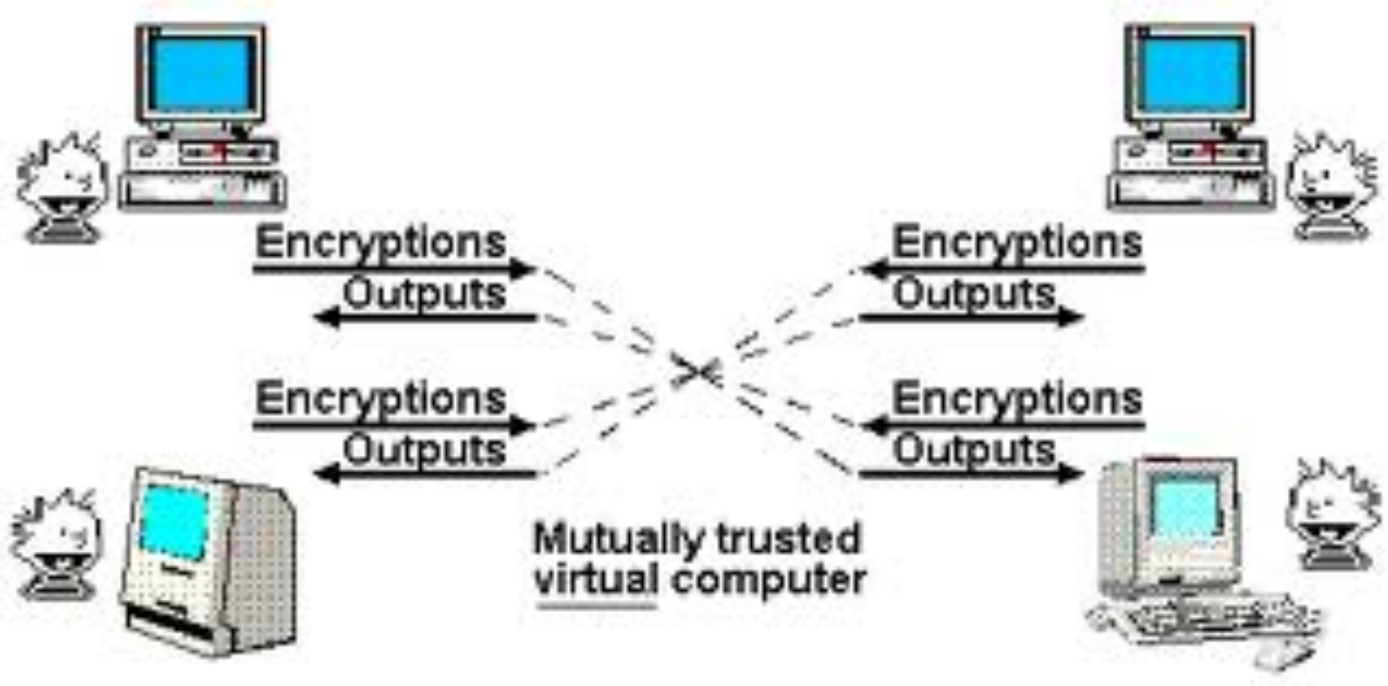
“Proof”

New Challenges for Data Science: Information from Data

- **Information Sharing Challenge:** Information sharing requires appropriately safeguarding both systems and information; selecting the most trusted information sources; and maintaining secure systems even in hostile settings
  - Example: “Secure Multiparty Computation” is a theoretical area aiming at allowing parties to jointly compute something over their inputs while keeping those inputs private

New Challenges for Data Science: Information from Data

- Secure multiparty computation is a "model" for secure information sharing.



Command, Control, and Interoperability Center for Advanced Data Analysis

New Challenges for Data Science: Information from Data

- **Trustworthiness Challenge:** To utilize the vast amounts of information available to us, we have to understand what sources we can trust
  - Example: Emergency situation; lots of data as to damage, physical needs, information needs, etc. What to trust?
  - Need precise definitions of factors contributing to trustworthiness: accuracy, completeness, bias

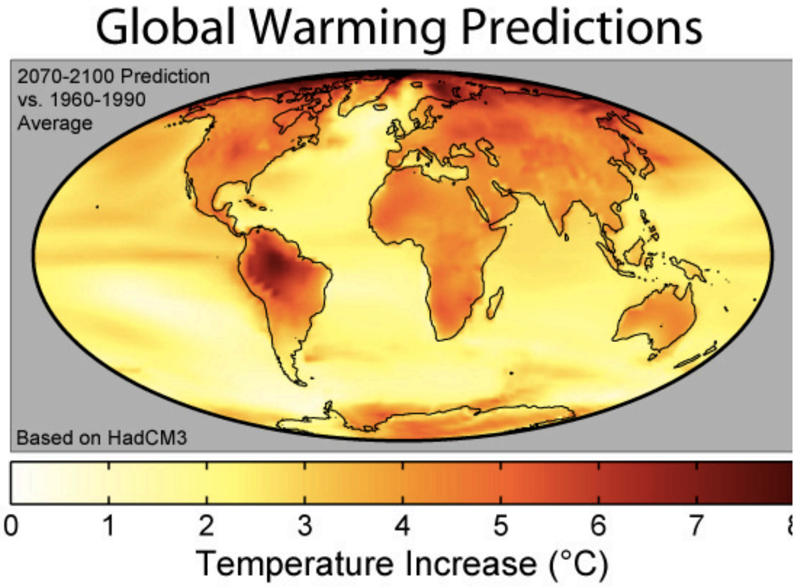
Japanese Earthquake & Tsunami; credits: commons.wikipedia.org and www.flickr.com

New Challenges for Data Science: Information from Data

- In building decision-supporting models, uncertainty arises from parameter values, model relationships, recorded observations, conflicting sources
- **Uncertainty Quantification Challenge**: How best present levels of uncertainty and best resolve conflicting predictions?

New Challenges for Data Science: Information from Data

- **Uncertainty Quantification Challenge**: How best present levels of uncertainty and best resolve conflicting predictions?
  - How to develop consensus when different models lead to at least seemingly different conclusions?
  - Example: Climate models



Credit: commons.wikipedia.com

Closing Comment

- It doesn't matter how big or small a dataset is.
- What matters is what we can do with the data.

Credit: www.flickr.com

Command, Control, and Interoperability
Center for Advanced Data Analysis