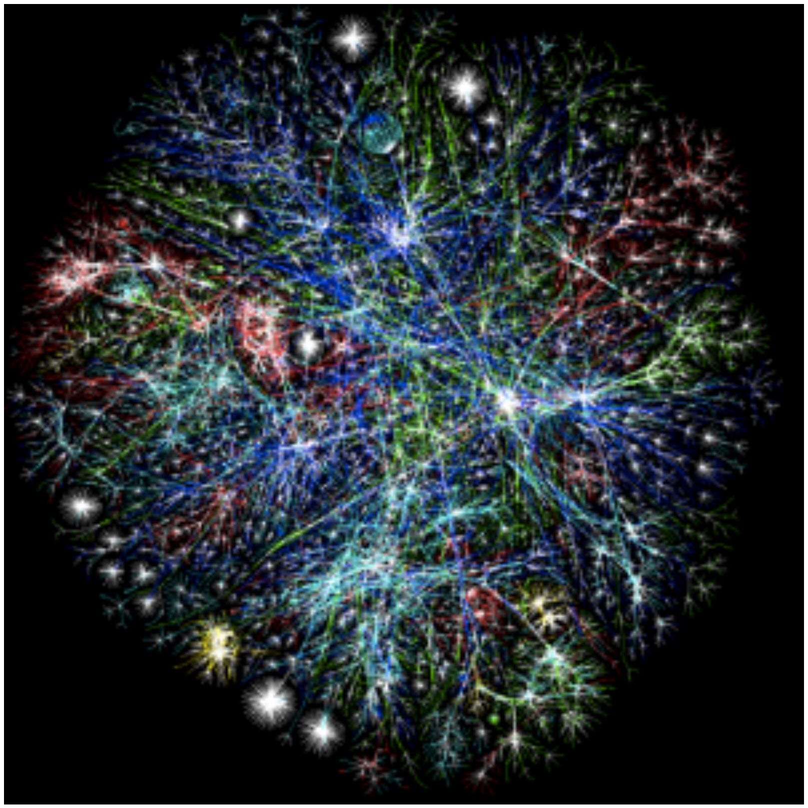
What is Big Data?
And How has it Changed?



Fred S. Roberts
Director of
CCICADA
Rutgers University

Credit: commons.wikipedia.org

Command, Control, and Interoperability
Center for Advanced Data Analysis

What is Big Data?
And How has it Changed?

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

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?

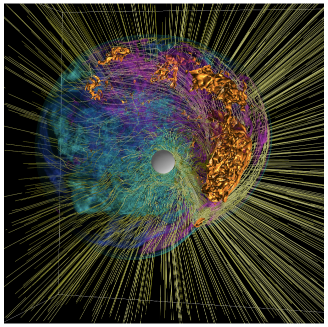
What challenges do we face in the next 10 years?

Command, Control, and Interoperability Center for Advanced Data Analysis

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

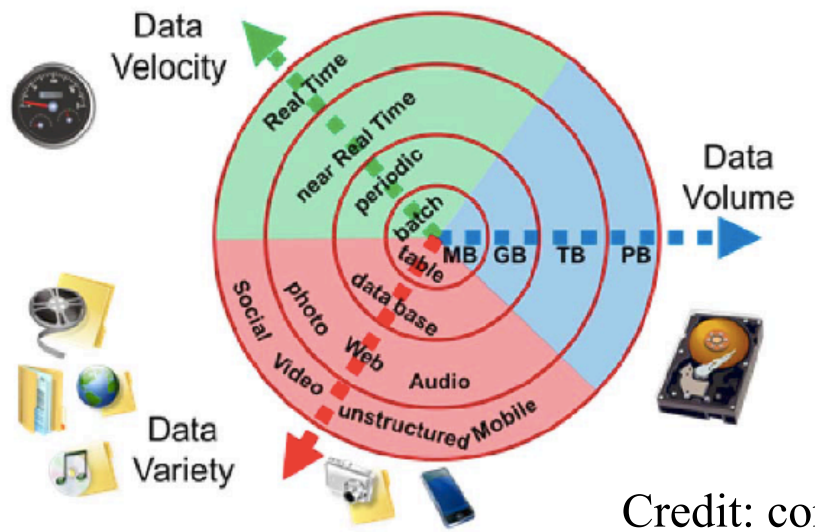
Credit: en.wikipedia.org



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.

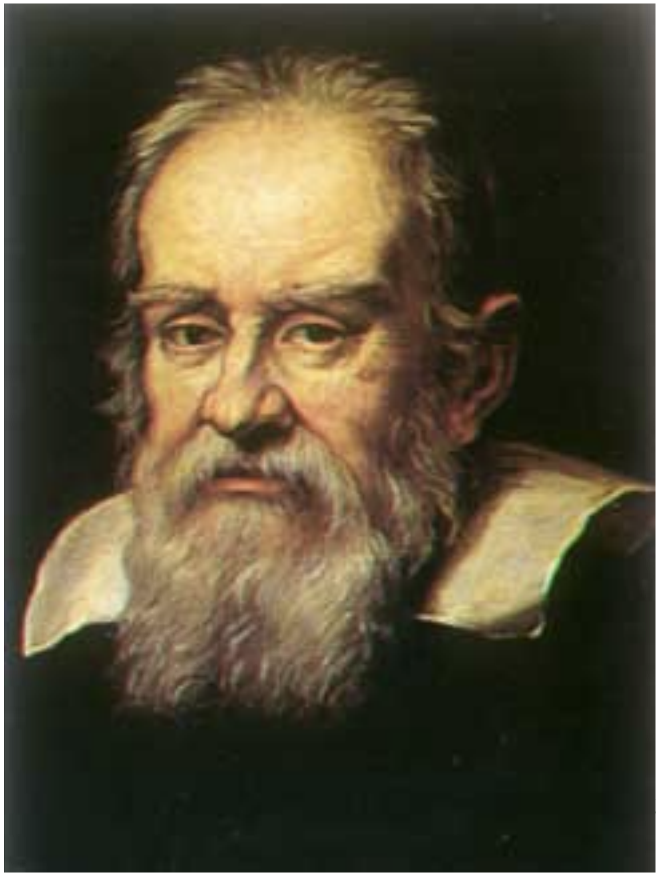
Command, Control, and Interoperability Center for Advanced Data Analysis

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)

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

Command, Control, and Interoperability Center for Advanced Data Analysis

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*

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

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

Credit: en.wikipedia.org

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

Resulting Challenges

- **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

Resulting Challenges

- **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?

Command, Control, and Interoperability Center for Advanced Data Analysis

Credits: commons.wikipedia.org, flickr.com

Resulting Challenges

- **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

Resulting Challenges

- **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

Resulting Challenges

- **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

Command, Control, and Interoperability Center for Advanced Data Analysis

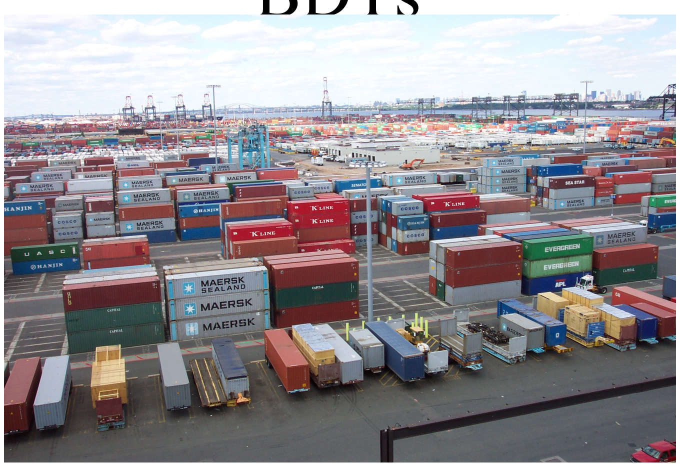
Resulting Challenges

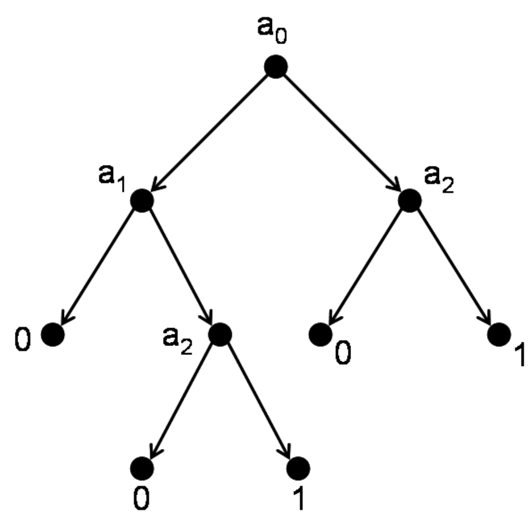
- **Combinatorial Explosion Challenge**
  - Big data allows comparison of array of alternative solutions
  - However, the number of alternatives is often so large that we cannot take all into account in a timely way
  - We may not even be able to express all possible preferences among alternatives –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

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

- **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



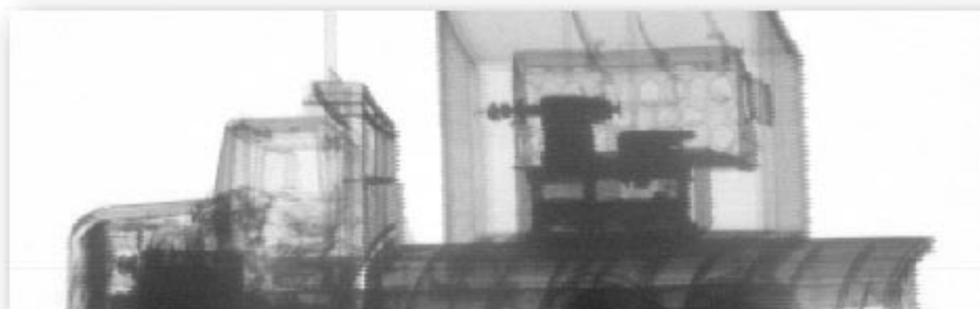


 Credit; en.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

- **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

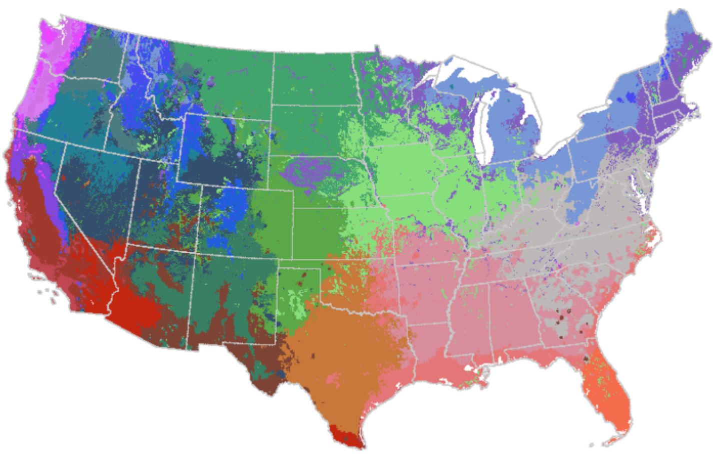


Credit: en.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

- **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.

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

- **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

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

- **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

Resulting Challenges

- **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

Resulting Challenges

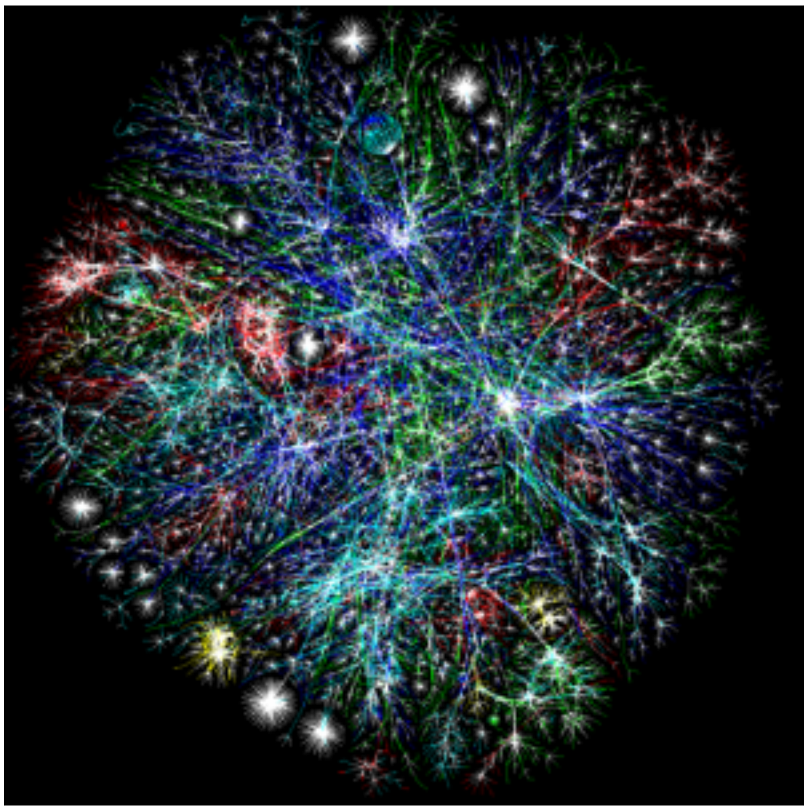
- **Streaming Data Challenge for Graphs & Networks**
  - Data such as IP traffic level, access logs, command logs arise from rapidly evolving graphs & networks
  - Situational awareness requires us to translate the data into large, interpretable, & manageable graphs
    - ➢ Graphs that can be monitored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

Credit: en.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

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



Credit: commons.wikipedia.org

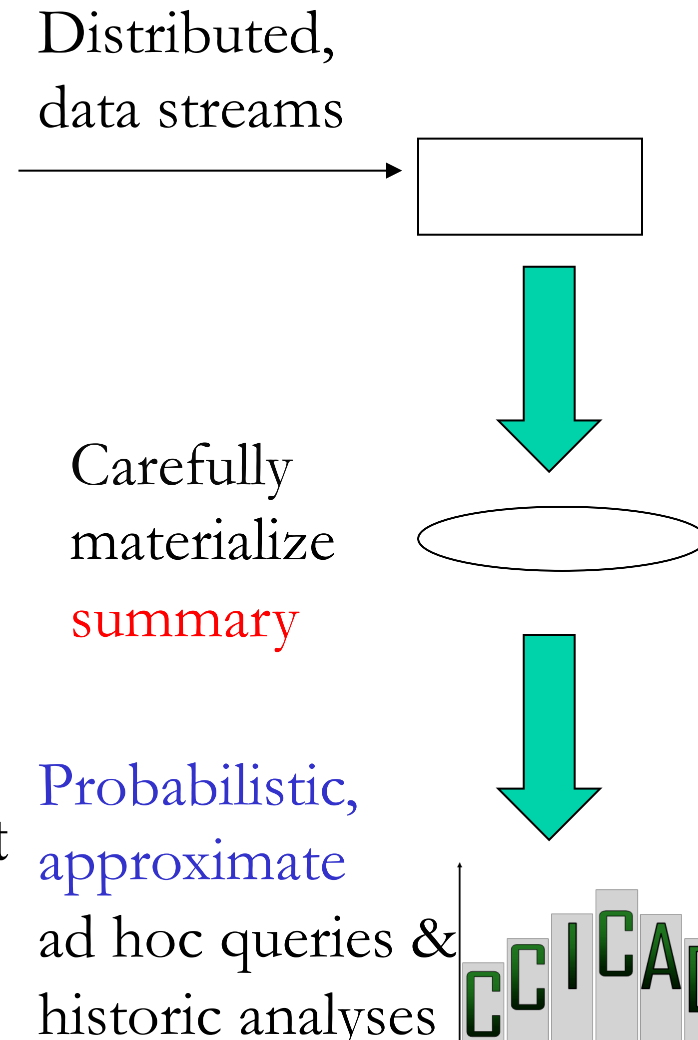
Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

- **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?



Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

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

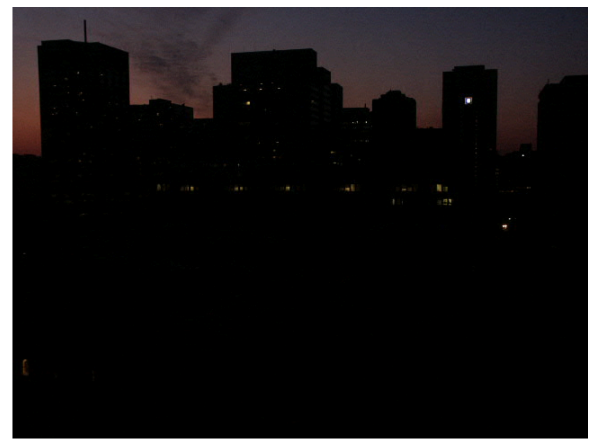
Credit: en.wikipedia.org

Credit: commons.wikipedia.org

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

- **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

Resulting Challenges

- 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

Resulting Challenges

- 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

Resulting Challenges

- 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

Resulting Challenges

- 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

Resulting Challenges

- 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

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges

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

Credit: www.flickr.com

Command, Control, and Interoperability Center for Advanced Data Analysis

Resulting Challenges: 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

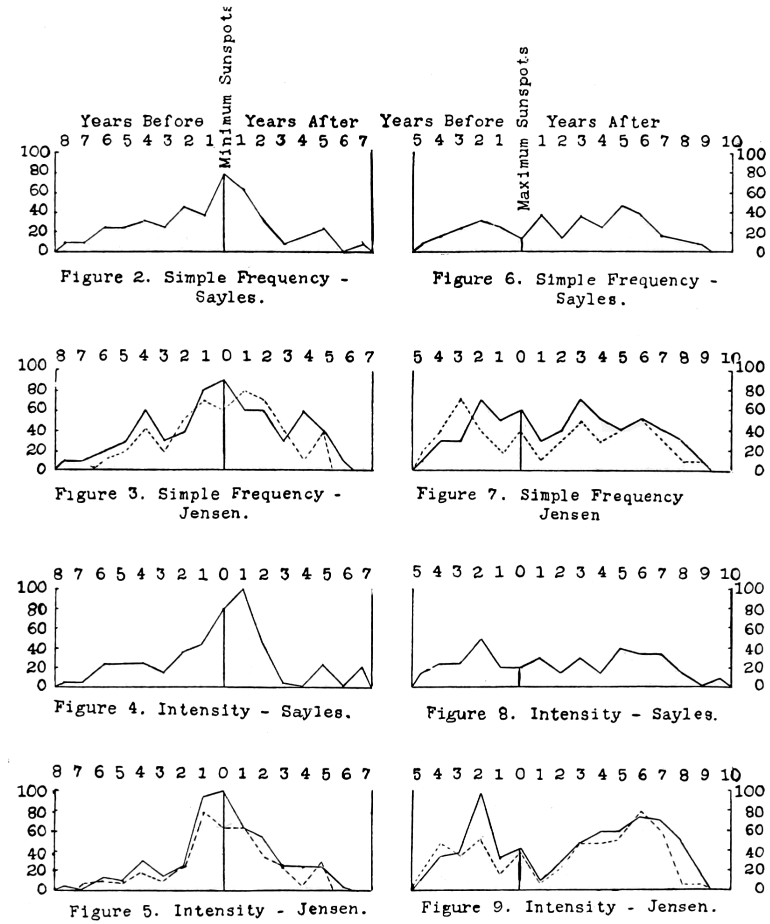
Information from Data

- **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.

Command, Control, and Interoperability Center for Advanced Data Analysis

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?



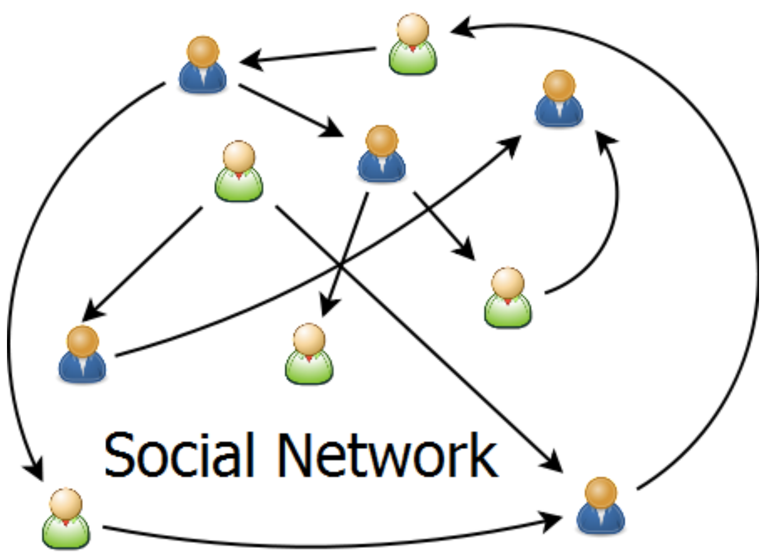
Command, Control, and Interoperability Center for Advanced Data Analysis

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?

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

Information from Data

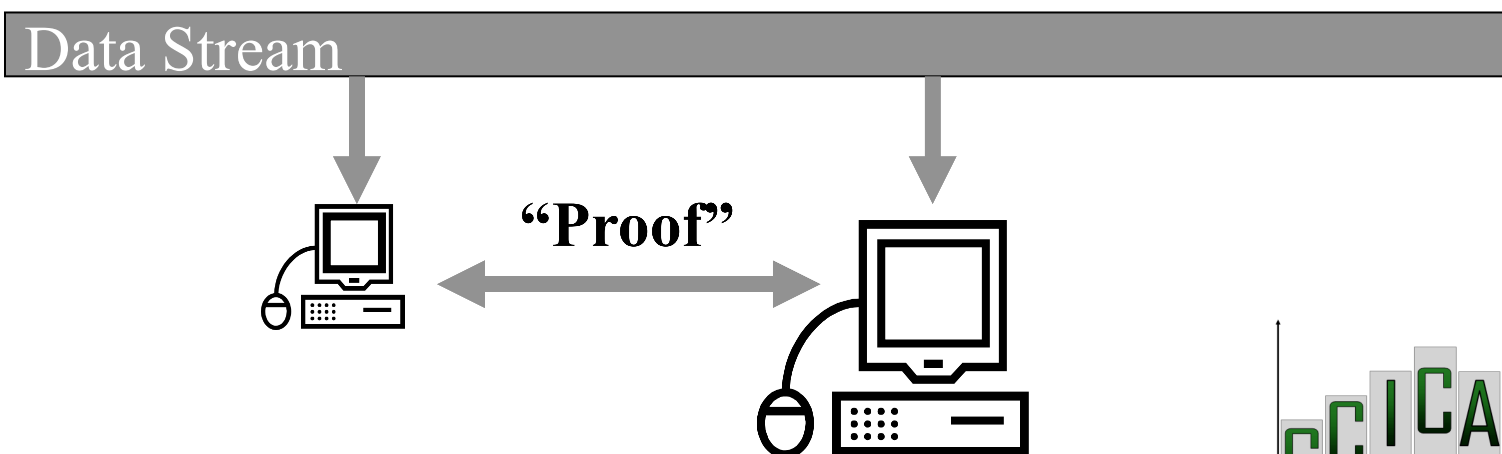
- **New Architectures Challenge:** As much data has grown too large to reside in one location, need new architectures.
- Big change in this direction: increasing emphasis on use of "the Cloud" to do computations, store data

Credit: commons.wikipedia.org

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

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.



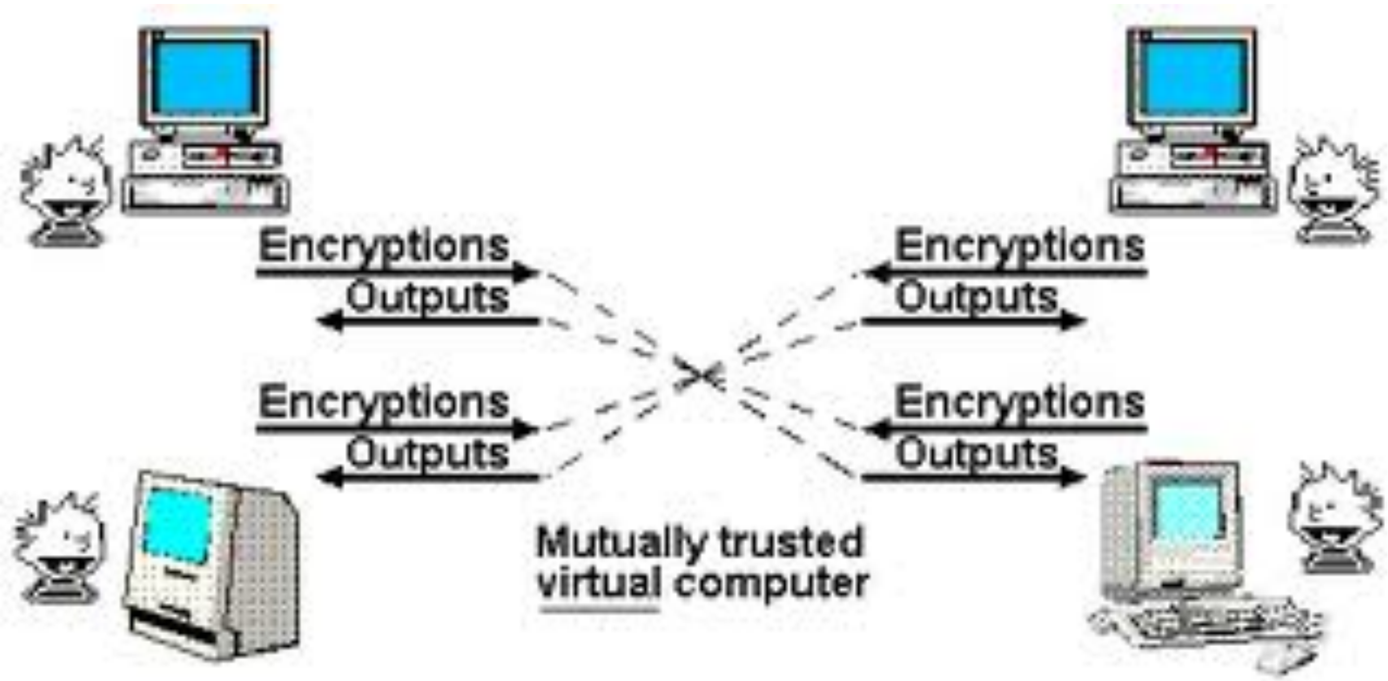
Command, Control, and Interoperability Center for Advanced Data Analysis

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.

Information from Data

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



Command, Control, and Interoperability Center for Advanced Data Analysis

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

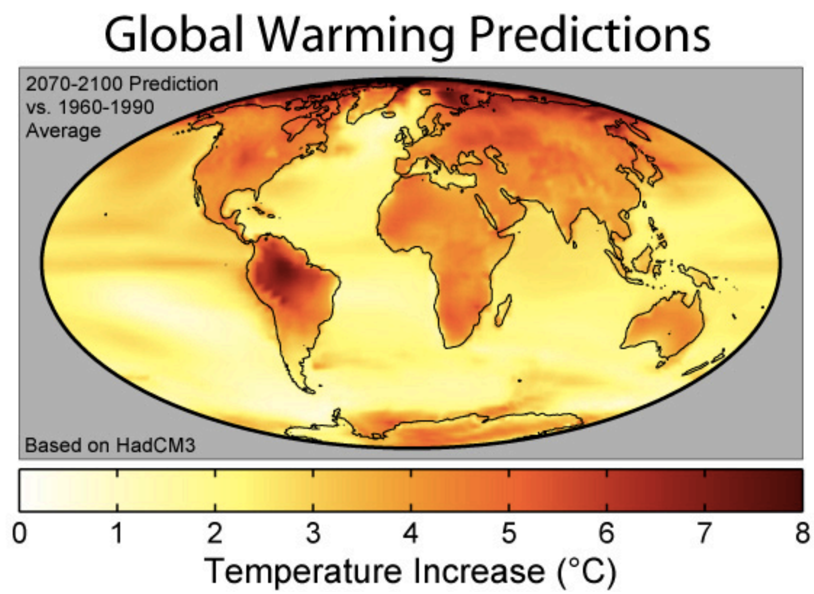
Command, Control, and Interoperability Center for Advanced Data Analysis

Information into Knowledge

- 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?

Information into Knowledge

- **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