Why Data Science?

Stanley C. Ahalt, PhD
Director, Renaissance Computing Institute
Professor of Computer Science, UNC-Chapel Hill
Director, Biomedical Informatics Service, NC TraCS,
UNC School of Medicine

ncds | THE NATIONAL CONSORTIUM for DATA SCIENCE

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci
RENAISSANCE COMPUTING INSTITUTE

Presentation Outline

1. Context
2. Why Data Science?
3. Meeting the Challenges and Opportunities of Data Science
4. The Mathematics of Data
5. The Economics and Ethics of Data
6. Possible Approaches
7. National Consortium for Data Science (NCDS)
8. Conclusion

Context

- Why Data Science?
- Meeting the Challenges and Opportunities of Data Science
- The Mathematics of Data
- The Economics and Ethics of Data
- Possible Approaches to the Data Challenge
- NCDS
- Conclusion

RENCI's Mission

- Be a **leader** in cyberinfrastructure (CI) research, development and deployment.
- Be an essential CI **partner** for research teams, industry, government
- Data is central to all we do. Underlying theme: ***Data to Decisions***

Triangle Campuses Snapshot

UNC	Duke	NCSU	Top Departments/ Schools UNC	Top Departments/ Schools-Duke	Top Departments/ Schools-NCSU
18,579 undergrads	6,504 undergrads	25,176 undergrads	Info and Library Sciences	Medicine	Engineering
10,558 grad/ professional	7,744 grad/ professional	9,591 grads	Public Health	Law	Textiles
3,518 faculty	1,770 faculty	2,068 faculty	Business	Business	Computer Science
UNC, Duke & NCSU Research Funding by Agency			Medicine/ Pharmacy	Environmental	Agricultural & Bio Engineering
NIH: $320.5M	NIH: $405.5M	NIH: $15.1M	Computer Science	Divinity	Veterinary Medicine
NSF: $20.5M	NSF: $36.3M	NSF: $27M			
DOE: $2.8M	DOE: $9.8M	DOE: $4.8M			
DOD: $7M	DOD: $30.3M	DOD: $10.1M			

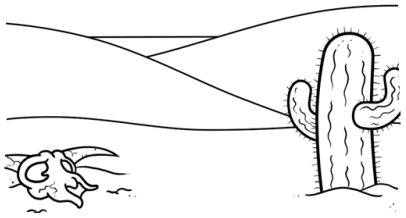
RENCI's latest role: Bridge Big Data's "Valley of Death" For Our Stakeholders

Data Producers

Universities
Healthcare
Industry
Government
Financial

Data

Tools, visualization, analysis

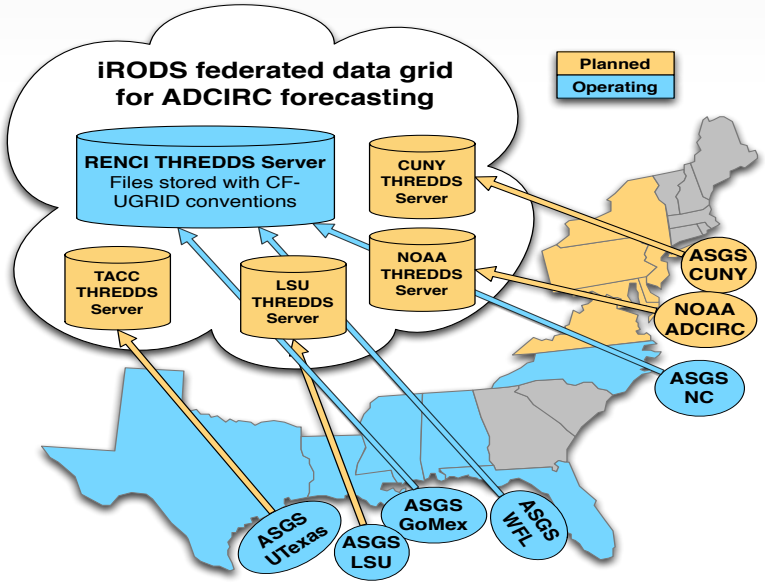


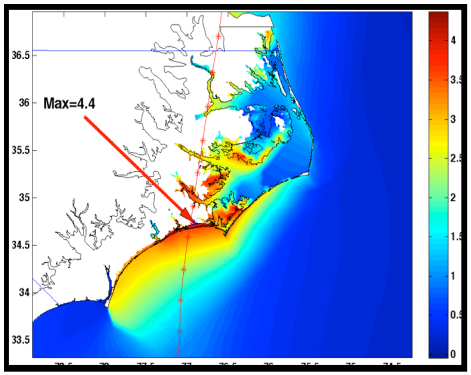
Data Consumers

Scientists
Citizens
Caregivers
Business
Government

The Challenge: the data exists, but tools are needed to enable at-scale use by data consumers.

Storm Surge Forecasting (ADCIRC)





Sandy (2012) and Irene (2011) flooding forecasts used by

- National Hurricane Center in Miami
- US Coast Guard Atlantic Command
- Regional National Weather Service Offices
- State and local emergency managers

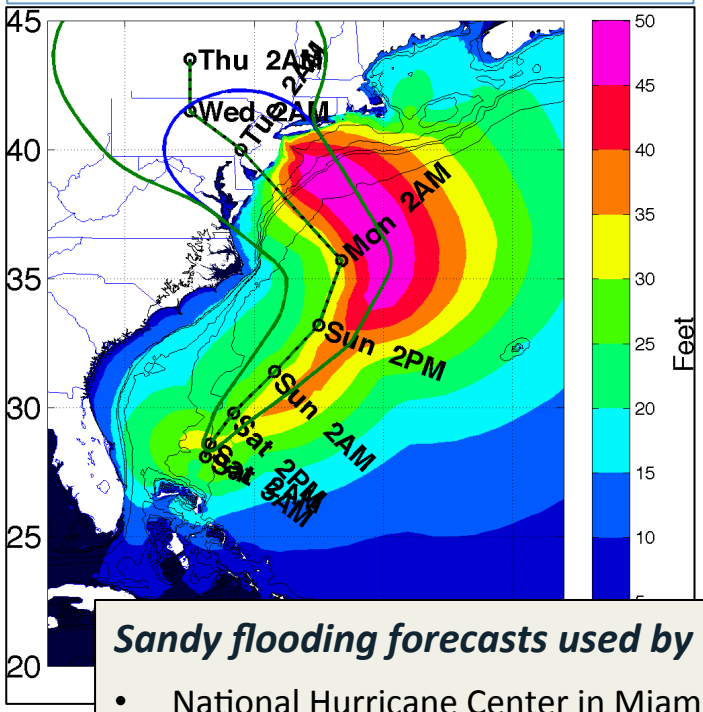
Winner, DHS Science & Technology Impact Award, 2012 and IDC HPC Innovation Excellence Award 2013

- System uses **US NSF/NARA** funded iRODS**, NOAA NOS** gauge data, **USGS** data, **US DHS/FEMA** collected high-water mark, meteorological forecasts from **NOAA's NCEP** and **NHC**
- Very large pre-existing datasets; provides early guidance information, available about 10 minutes after official NHC forecast storm advisory
- **US DHS-funded research activity** through the DHS Coastal Hazards Center of Excellence at the University of North Carolina at Chapel Hill

ADCIRC/North Carolina Forecast System

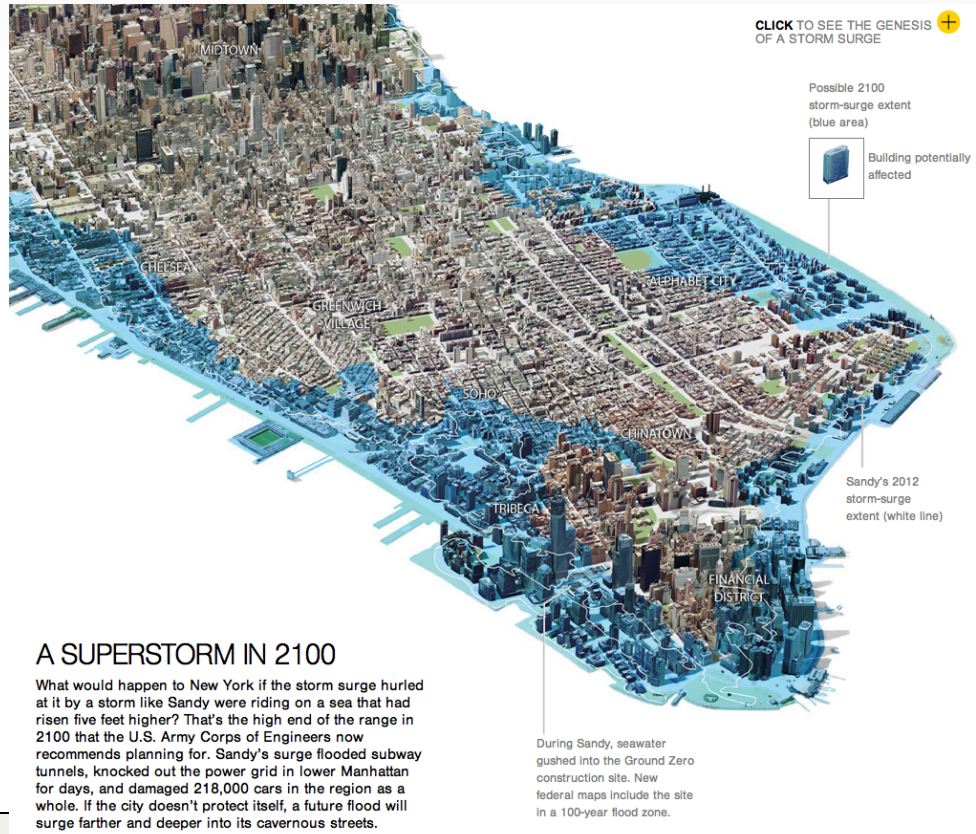
Hurricane Sandy 2012

Forecast track and Wave Heights for Hurricane Sandy Advisory 20



Sandy flooding forecasts used by

- National Hurricane Center in Miami
- Regional National Weather Service Offices
- NC Division of Emergency Management
- NC County Emergency Operations
- ***US Coast Guard Atlantic Command***



Above: RENCI helped interpret data from the NOAA *SLOSH* model to show what Sandy-like storm surge could look like with sea levels five feet higher. Image was a center spread in National Geographic Magazine (September 2013)

NCGENES - Clinical Genomics

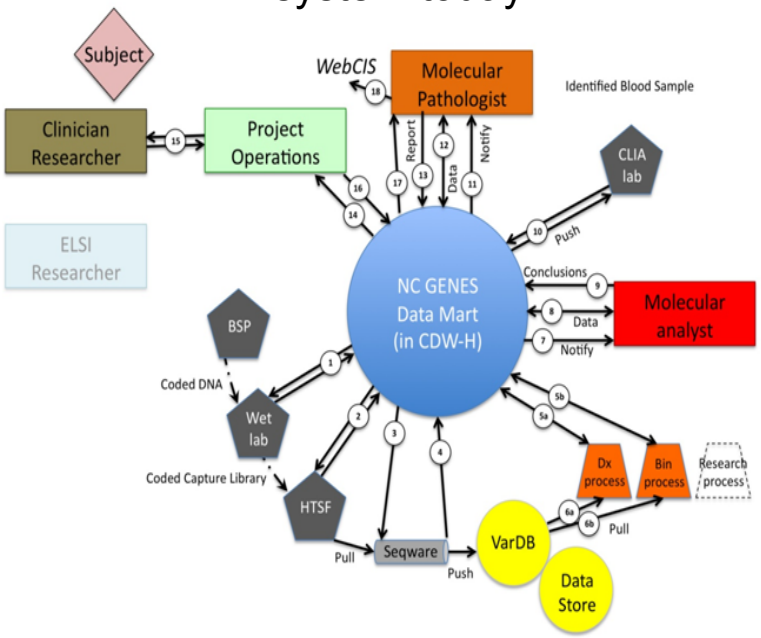
Today:

- NIH prototype to evaluate the *ethical* and *social* challenges of genomic sequencing in clinical care
- Big Data to clinically-relevant knowledge ('Clinical bins')
- Over 100 patients in the system today…

Tomorrow:

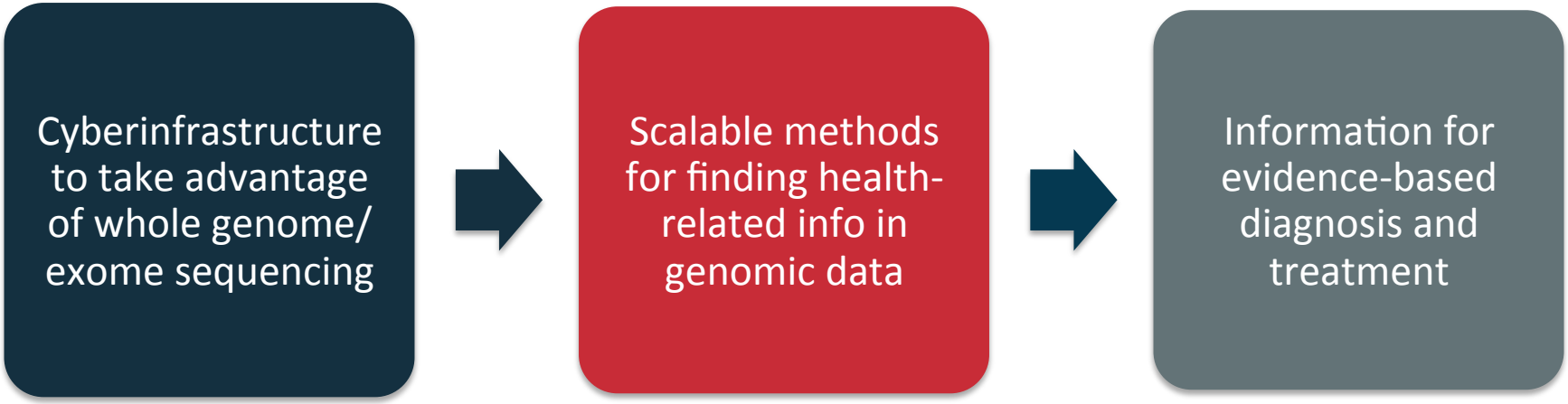
- 100M+ genomes scattered throughout the health care system
- We face a multitude of data challenges before we realize the potential of genomics in healthcare…

North Carolina Genetic Evaluation by Next-generation Exome Sequencing

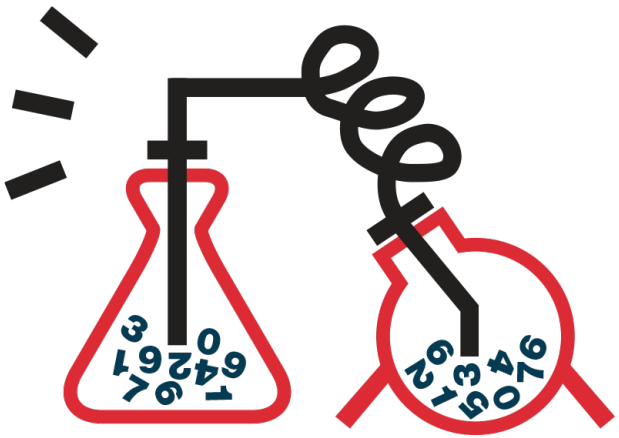


Criteria:	*Loci with Clinical Utility*	*Loci with Clinical Validity*			*Loci with Unknown Clinical Implications*	Loci with important reproductive implications
Bins:	**Bin 1** Genes, which when mutated, result in high risk of clinically actionable condition	**Bin 2A** Low risk incidental information	**Bin 2B** Medium risk incidental information	**Bin 2C** High risk incidental information	**Bin 3** All other loci	**Bin R** Carrier status for severe AR disease

NC GENES: The Vision



A Step toward Personalized Medicine



- Context
- **Why Data Science?**
- Meeting the Challenges and Opportunities of Data Science
- The Mathematics of Data
- The Economics and Ethics of Data
- Possible Approaches to the Data Challenge
- NCDS
- Conclusion

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

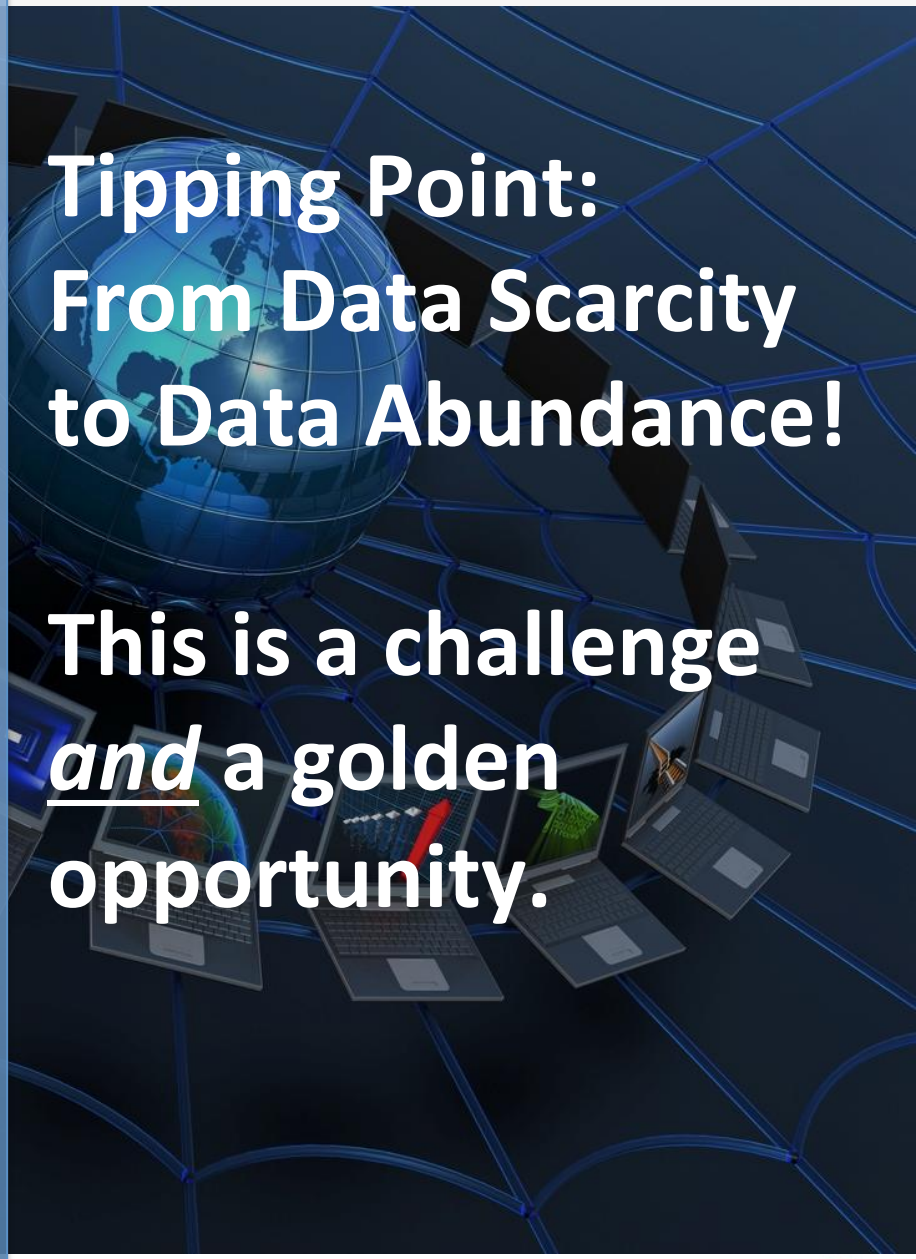
renci RENAISSANCE COMPUTING INSTITUTE

Why Data Science? ABUNDANCE

Percentage of worldwide digital data created in the last two years?

90%

Since 2010 we have been creating as much data every two days as was previously created in all of history up to 2003.



Tipping Point: From Data Scarcity to Data Abundance!

This is a challenge *and* a golden opportunity.

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE renci

From Compute-Centric to Data-centric Research

From a Molehill To a Mountain

A timeline of digital data

Everyone knows that the amount of digital data has exploded in recent years. But it isn't always easy to put today's mountain of data in perspective.

To illustrate the data boom, here's a timeline with snapshots of some of the largest corporate data collections from each decade since the Univac computer was the hot new thing. —*Matthew Kassel*

2010s | **100 Petabytes**

Facebook Is Facebook's data hoard bigger than Google's today? Some say yes, some say no. According to Facebook, its user content makes up more than 100 petabytes of stored photos and video. And analyzing that data generates about 500 terabytes of new information every day—more than 2½ times the size of a '90s Wal-Mart data cache.

2000s | **25 Petabytes**

Google The explosion of the Internet in the '90s set the stage for Web-based companies to emerge in the following decade as the global leaders in big data.

1990s | **180 Terabytes**

Wal Mart Wal-Mart in this decade became the largest American bricks-and-mortar retail operation, and, it is believed, had the biggest commercial data warehouse in the world. To put this number in perspective, experts have estimated that Amazon.com in the late '90s had

Close

bytes of stored data.

1980s | **450 Gigabytes**

CitiCorp's NAIB In the 1980s, banks were at the forefront of data growth, with ATMs coming into vogue and banks focusing on collecting and analyzing transaction data for all their businesses.

1970s | **80 Gigabytes**

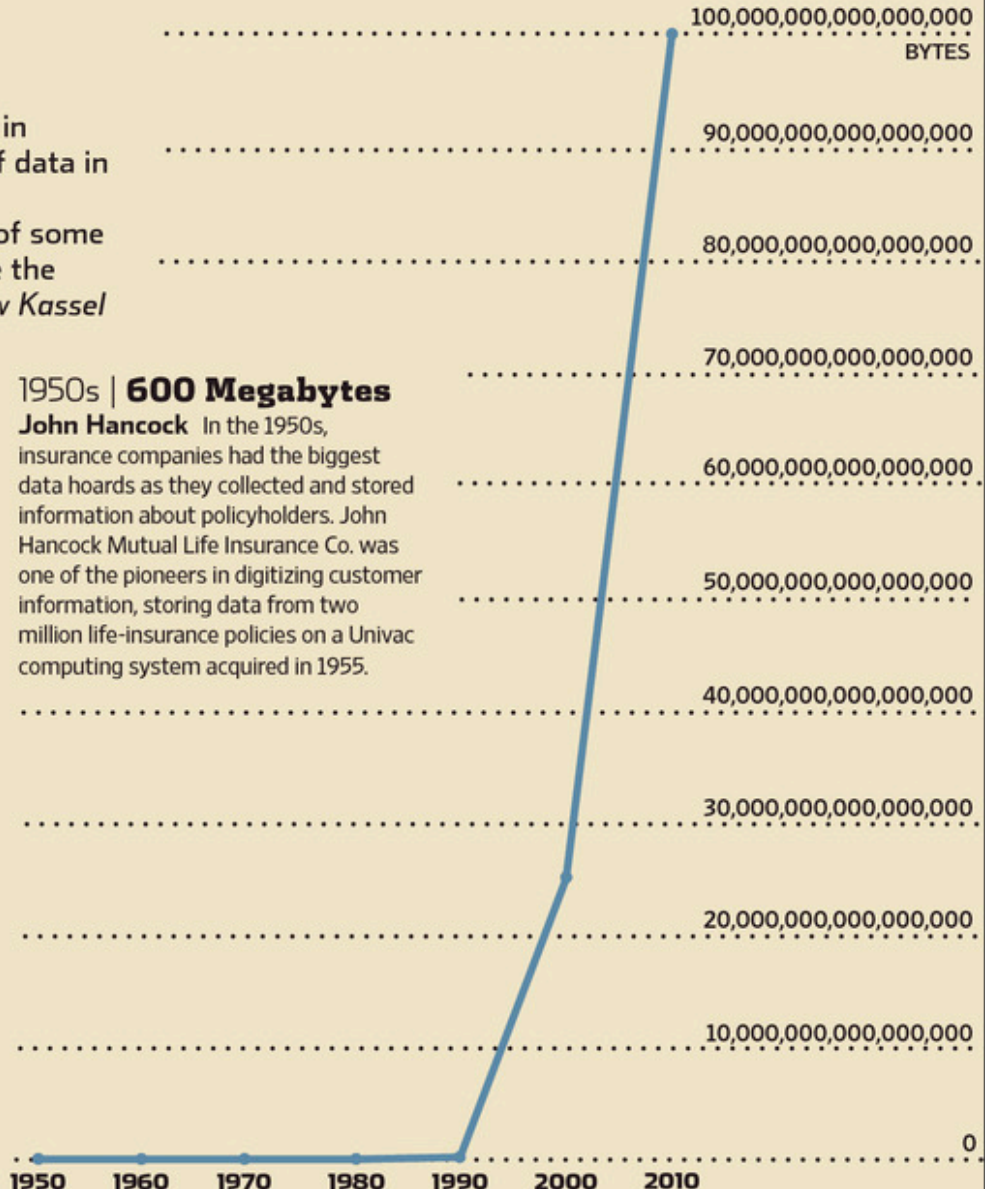
Federal Express Cosmos FedEx's Cosmos system, introduced in the 1970s, allowed the company to scan and track its huge volume of packages being shipped around the world.

1960s | **807 Megabytes**

American Airlines Sabre In the 1960s, American Airlines developed Sabre, a flight-reservation system built around one of the largest IBM computing systems available. Sabre was one of the first online computing systems, allowing the airline to keep track of an immense matrix of reservations, flight schedules and seat inventories.

1950s | **600 Megabytes**

John Hancock In the 1950s, insurance companies had the biggest data hoards as they collected and stored information about policyholders. John Hancock Mutual Life Insurance Co. was one of the pioneers in digitizing customer information, storing data from two million life-insurance policies on a Univac computing system acquired in 1955.



Source: Wall Street Journal, Special Report on Big Data, March 11, 2013

Importance Driven by Technology

- The Internet made it easy to move, share, and find data:
  - *"information wants to be free," and it wants to be expensive*
- Faster processors, more and cheaper storage capacity:
  - *Creating, processing, storing data is easier, clouds have accelerated this trend.*
- Sensors and the explosion of real-time data:
  - *More than **1 trillion** sensors now connected to the Web*
  - *Example: Google I/O 2013 conference deployed hundreds of sensors to collect ambient data*
- The Internet of Things = an explosion of data created by connected devices, not people.
- Biological data: sequencing/medicine could produce **50EBs** of data/year.

Big Data, Big Results

- **Express Scripts:**
  - 1 billion pharmacy insurance claims analyzed and used to drive patients to more cost-effective mail order prescriptions
  - Predictive modeling of 400 factors to find patients at risk for non-adherence to subscriptions (a $317 billion/year problem).
- **UPS:**
  - Analyzing continuous streams of sensor data from thousands of delivery trucks eliminated 5.3M miles from routes, reduced engine idling time by 10M minutes, saved 650,000 gallons of fuel, reduced carbon emissions by + 6,500 metric tons.

- **Intel:**
  - Analysis of massive data and application of predictive algorithms helped ID potential high-sale resellers (result: +$20M in potential new sales).
  - Manufacturing predictive analytics reduced microprocessor testing time (result: $3M saved during proof of concept period. $30M savings expected by 2014).

Source: CIO, July 15, 2013

How big is the opportunity?

- **$300B** potential annual value to US healthcare—more than total annual healthcare spending in Spain.
  - *McKinsey Global Institute, May 2011*
- €250B potential annual value to Europe's public sector administration.
  - *McKinsey Global Institute, May 2011*
- Energy savings of 1% in gas-powered plants – savings of **$68B over 15 years.**
  - *Industrial Internet: Pushing the Boundaries of Minds and Machines, GE, Nov. 12, 2012*
- Companies using data-directed decision making **boost productivity by 5-6%.**
  - *Cukier, K., Data, data everywhere, The Economist, Feb. 25, 2010*
- Jobs: demand for data-related administrators and software developers projected **to grow by ~32% in US by 2020.**
  - *Occupational Outlook Handbook, 2012-2013, US Bureau of Labor Statistics*

Big Data Jobs: The Opportunity

- ***Globally:***
  - Big Data and analytics jobs expected to exceed **4 million** by 2015. *(source: icrunchdata Big Data Jobs Index)*
- ***Nationally***:
  - Big data **job postings up 63%** on icruchdata job site.*(source: icrunchdata.com)*
  - **1.9M new big data jobs** by 2015, but only 1/3 will be filled due to lack of trained talent *(source: Gartner, October 2012)*
  - Each big data job will create 3 additional jobs. *(source: Gartner, 2012)*
  - Demand for data-related administrators and software developers projected **to grow by ~32%** in US by 2020 (*source: Occupational Outlook Handbook, 2012-2013, US Bureau of Labor Statistics*
  - **$300B** potential annual value to US healthcare—more than total annual healthcare spending in Spain (source: *McKinsey Global Institute, May 2011)*

Challenges: Big Data Talent Shortage

- **78 percent** of 2012 survey respondents said there is a big data talent shortage *(The Big Data London Group in Raywood, 2012)*
- **70 percent** of survey respondents noted a knowledge gap between data workers and managers/CIOs *(The Big Data London Group in Raywood, 2012)*
- **60 percent** of survey respondents say it's difficult to find big data professionals *(NewVantage Partners 2012)*
- **50 percent** of survey respondents have difficulty finding and hiring business leaders and managers who understand how to apply big data *(NewVantage Partners 2012)*

Big data experts need skills in:

- Advance analytics and predictive analysis
- Complex event processing
- Rule management
- Business intelligence tools
- Data integration

Big data scientists need the skills of their IT predecessors, plus a solid computer science background (knowledge apps, modeling, statistics, analytics, math), business savvy, and the ability to communicate their findings.

More Big Data Challenges

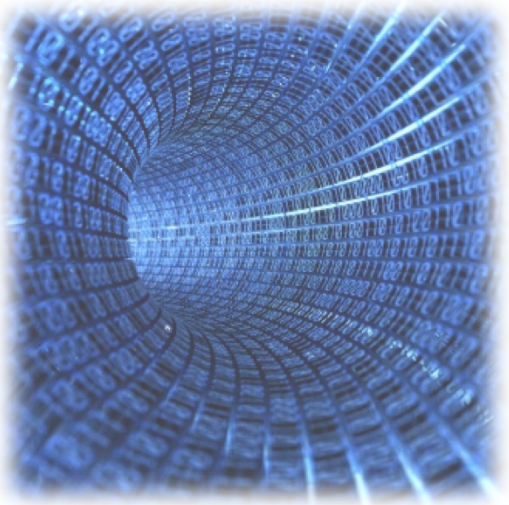
- **Data is:**
  - Handled in different ways
  - Captured in different formats
- No methodologies for measuring data's value

- Big Data must be defined
- Methodologies must be established to measure its value

Defining "Big" Data

The Five Vs:

- **Volume**: The Large Hadron Collider discards 99.999% of its data *because the data cannot be processed!*
- **Velocity**: Retail transactions, communications, industrial sensor data, demand real-time analysis and action.
- **Variety**: Health data includes images, test results, medical histories, doctor's notes.
- **Veracity:** Data quality essential for discovery and informed decision making
- **Value**: How important or rare is the data, and what do we keep and for how long?



Data use cases are heterogeneous

- Importance of each V varies, even within same data set

Data management and analytics hardware and expertise are expensive

- Can be barriers to entry, especially for small businesses and new researchers

The “Big” in Big Data

“Big Data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight, discovery and process optimization.”

Source: Beyer & Laney 2012

Like “artificial intelligence” the definition is a moving target.

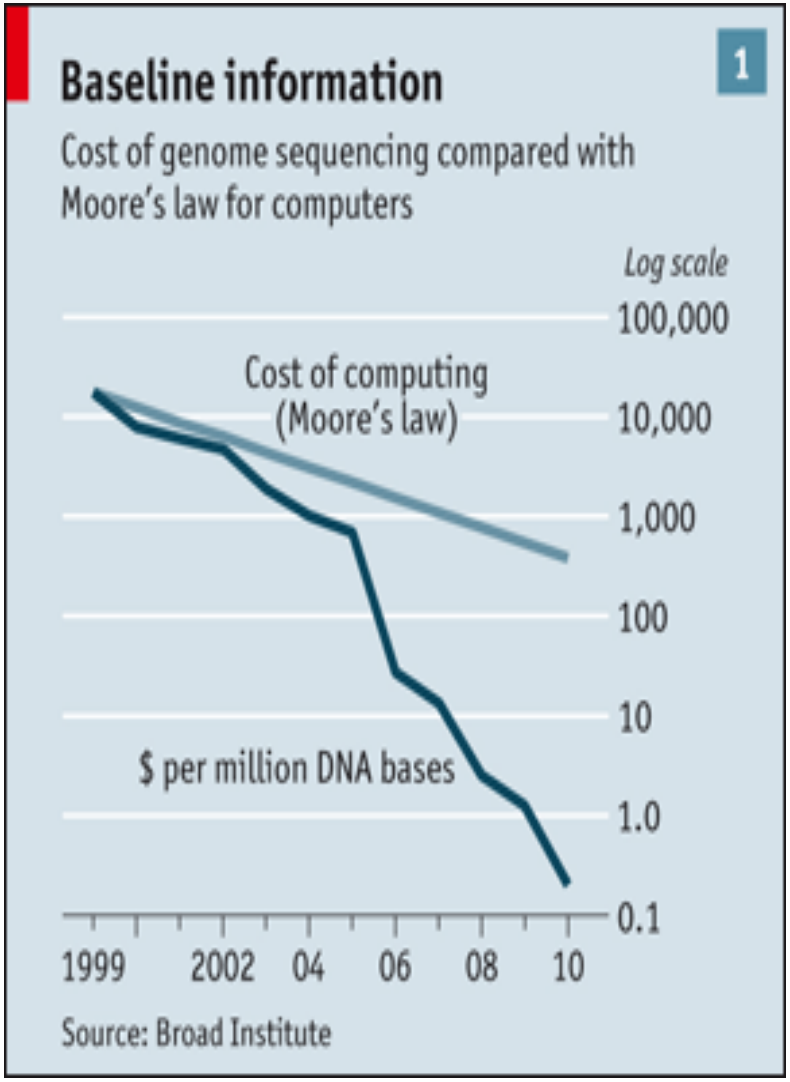


Image courtesy Chris Bizon. RENCI

Defining Data Science

Data Science:
Systematic study of organization and use of digital data for:

- ❑ research discoveries,
- ❑ decision-making, and
- ❑ the data-driven economy.

What Is a Data Scientist?

"Good data scientists will not just address business problems, they will pick the right problems that have the most value to the organization.

-*IBM*

Data scientists "must be able to take data sets, model them mathematically, and understand the math required to build those models. And they must be able to find insights and tell stories from that data. That means asking the right questions."

-*Hilary Mason, Wall Street Journal, in Rooney 2012*

Data Science has a history!

1962:

ANNALS OF MATHEMATICAL STATISTICS

previous :: next

The Future of Data Analysis

John W. Tukey

Source: Ann. Math. Statist. Volume 33, Number 1 (1962), 1-67.

First Page: Show

Related Works:

See Correction: John W. Tukey. Correction Notes: Correction to "The Future of Data Analysis". Ann. Math. Statist., Volume 33, Number 2 (1962), 812--812.
Project Euclid: euclid.aoms/1177704604

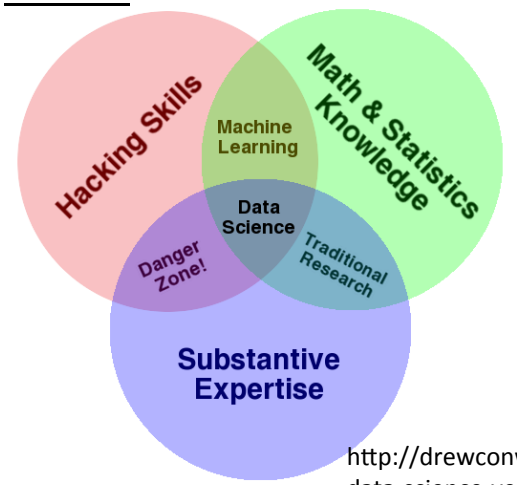
2002:

ICSU CODATA DATA SCIENCE Journal

2007:

复旦大学 1905 数据科学研究中心
Research Center for Dataology and DataScience

2010:



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

2013:

Forbes

TECH | 5/28/2013 @ 9:09AM | 23,872 views

A Very Short History Of Data Science

http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/2/

- Context

Why Data Science?

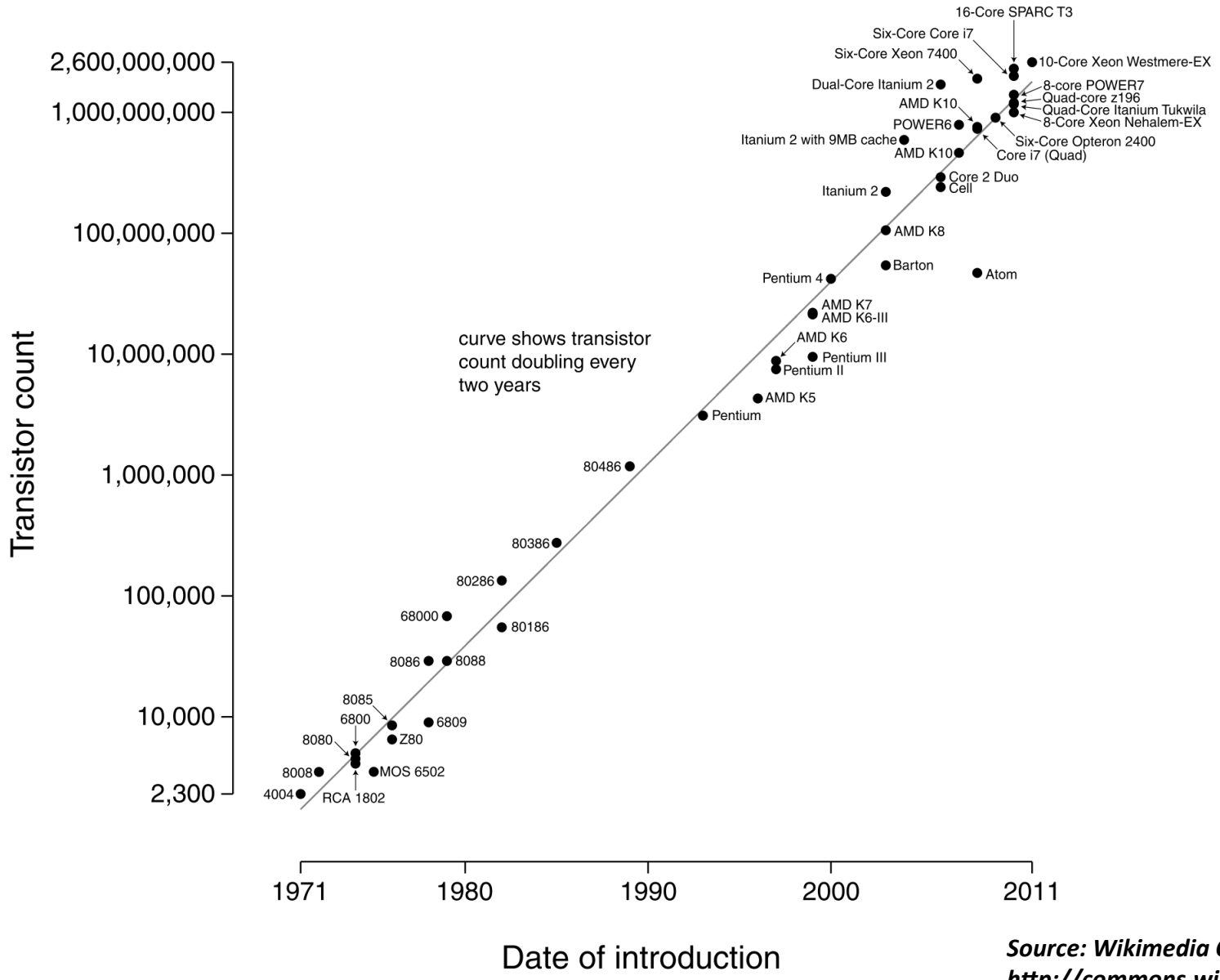
- **Meeting the Challenges and Opportunities of Data Science**
- The Mathematics of Data
- The Economics and Ethics of Data
- Possible Approaches to the Data Challenge
- NCDS
- Conclusion

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci RENAISSANCE COMPUTING INSTITUTE

Microprocessor Transistor Counts 1971-2011 & Moore's Law



Source: Wikimedia Commons, http://commons.wikimedia.org

A Moore's Law for Data?

Moore's Law: capabilities (value!) of computer hardware will double every 18 months.

➡ ***Is the same true for data?***

Data expected to grow 64%/year, some categories (particle accelerators and DNA Sequencer data) are growing much faster

➡ ***As we increase the volume of data, are we increasing its value?***

Some Think Data is Very Valuable

National Science Foundation
WHERE DISCOVERIES BEGIN

QUICK LINKS

SEARCH

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT NSF | FASTLANE

News

- News
- News From the Field
- For the News Media
- Special Reports
- Research Overviews
- NSF-Wide Investments
- Speeches & Lectures
- *NSF Current* Newsletter
- Multimedia Gallery
- News Archive

News by Research Area

- Arctic & Antarctic
- Astronomy & Space
- Biology

Email Print

Media Advisory 13-013

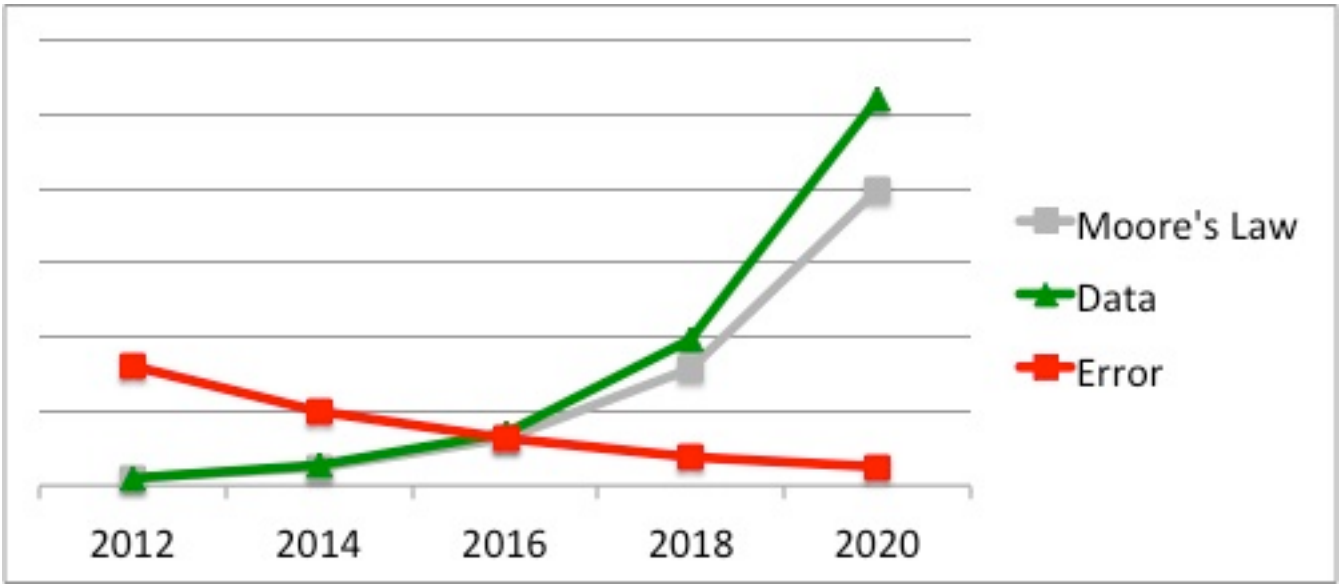
Data: a resource more valuable than gold?

US-hosted meeting of international group to address data sharing across global and disciplinary boundaries

The international Research Data Alliance convenes in Washington, D.C., Sept. 16-18.
Credit and Larger Version

How to Value Data

"For big data, Moore's Law means better decisions."
-Ion Stoica, UC Berkeley



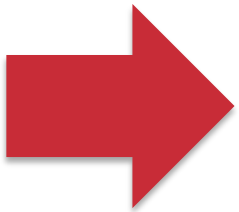
- Goal: extract value from data (usually decisions)
- Data is growing faster than Moore's Law (~64% per year)
- Possible solution: approximate answer using subsets of data, so volume of data is exploited to bound error.

Projects at Berkeley AMPLab

- Bag of Little Bootstraps (BLB): Bootstrapping big data
  - *A simple and powerful means of assessing the quality of estimators.*
- BlinkDB
  - *A massively parallel, approximate query engine for running interactive SQL queries on large volumes of data*

Is More Always Better?

More data points, more data sets, more time steps, more complete analysis



More accurate conclusions, more discoveries, more data into knowledge and action

"We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where scientists cannot."

-Chris Anderson, Wired, June 2008

Other Views of the Data Deluge

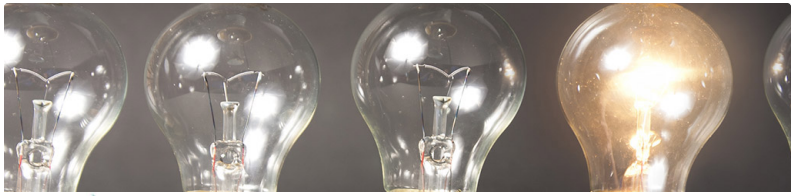
The Data Deluge is the current Wave of the Future … The problem is that when "waves of the future" show up they often wash away a number of worthy things and leave a number of questionable items littering the beach."

-George Andrews, Notices of the AMS, August 2012.

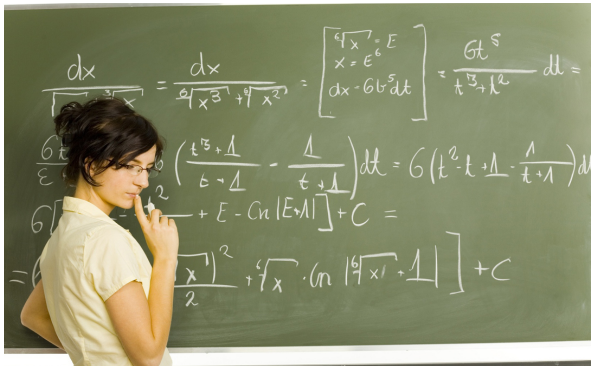
Andrews Asks:

- In an age when it is possible to collect, analyze draw conclusions from data without models—should we let go of the scientific method of hypothesize, model and test?
- What will this mean for education?
- Does it ignore how humans learn?

"I fear that one of the unintended consequences is the unstated assumption that nothing is trustworthy if it is not supported by data. "



- Context
- Why Data Science?
- Meeting the Challenges and Opportunities of Data Science
- **The Mathematics of Data**
- The Economics and Ethics of Data
- Possible Approaches to the Data Challenge
- NCDS
- Conclusion



ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci RENAISSANCE COMPUTING INSTITUTE

Algorithm Developer Tool Gap

Scalable Databases (triple stores)

Cassandra
HBASE
Google Bigtable
accumulo

Legacy Tools

Java
php
perl
python
Ruby

Distributed Mobile Display Devices

Algorithm Developer Needs
Graphs + Strings + Numbers
Composable Mathematics

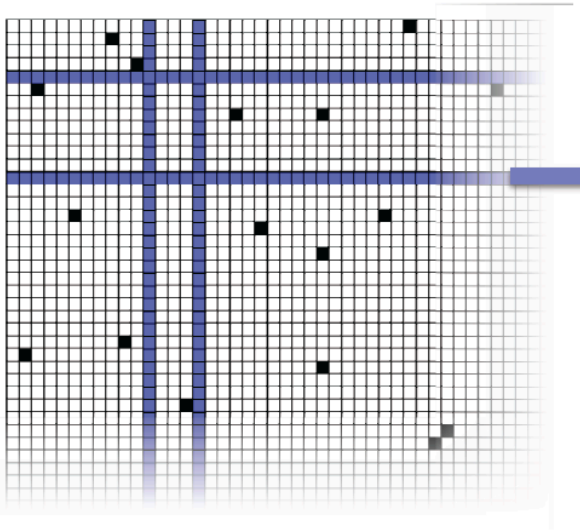
Tool Requirements
Multi-dimensional arrays
Operator overloading
Sparse linear algebra

- Scalable databases provide low latency access to vast stores of data
- Mobile devices allows data to be viewed anywhere
- Legacy tools not intended for big data algorithm development

Source: Kepner et al., 2012, Lincoln Laboratory, MIT

D4M: "Databases For Matlab"

Triple Store
Distributed Database

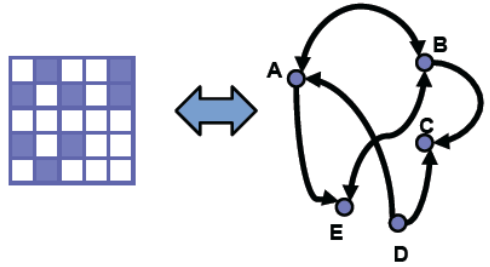


Triple store are high performance distributed databases for heterogeneous data

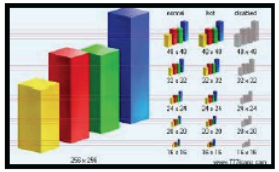
D4M
Dynamic Distributed Dimensional Data Model

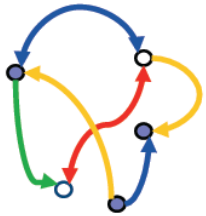
Query:
Alice
Bob
Cathy
David
Earl

Associative Arrays
Numerical Computing Environment



A D4M query returns a sparse matrix or graph from Accumulo …





…for statistical signal processing or graph analysis in MATLAB

- **D4M binds Associative Arrays to Triple Store, enabling rapid prototyping of data-intensive cloud analytics and visualization**

Source: Kepner et al., 2012, Lincoln Laboratory, MIT

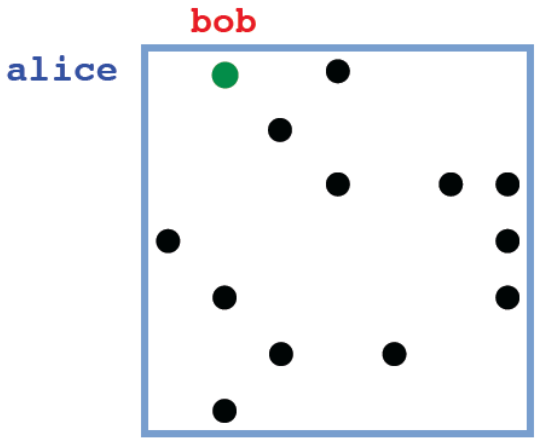
Multi-Dimensional Associative Arrays

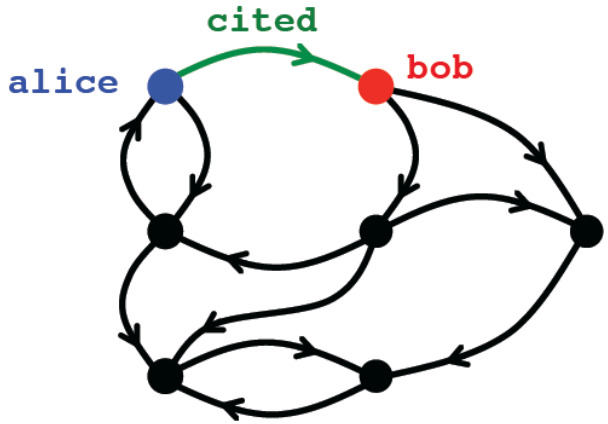
- **Extends associative arrays to 2D and mixed data types**

```
A('alice ','bob ') = 'cited '
```
or
```
A('alice ','bob ') = 47.0
```

- **Key innovation: 2D is 1-to-1 with triple store**

```
('alice ','bob ','cited ')
```
or
```
('alice ','bob ',47.0)
```





- **Associative arrays unify four viewpoints into one concept**

Source: Kepner et al., 2012, Lincoln Laboratory, MIT

Composable Associative Arrays

- **Key innovation: mathematical closure**
  - **All associative array operations return associative arrays**
- **Enables composable mathematical operations**

```
A + B      A - B      A & B      A|B      A*B
```

- **Enables composable query operations via array indexing**

```
A('alice bob ',:)     A('alice ',:)     A('al* ',:)

A('alice : bob ',:)   A(1:2,:)          A == 47.0
```

- **Simple to implement in a library (~2000 lines) in programming environments with: 1st class support of 2D arrays, operator overloading, sparse linear algebra**

> - **Complex queries with ~50x less effort than Java/SQL**
> - **Naturally leads to high performance parallel implementation**

Source: Kepner et al., 2012, Lincoln Laboratory, MIT

Technology Comparison

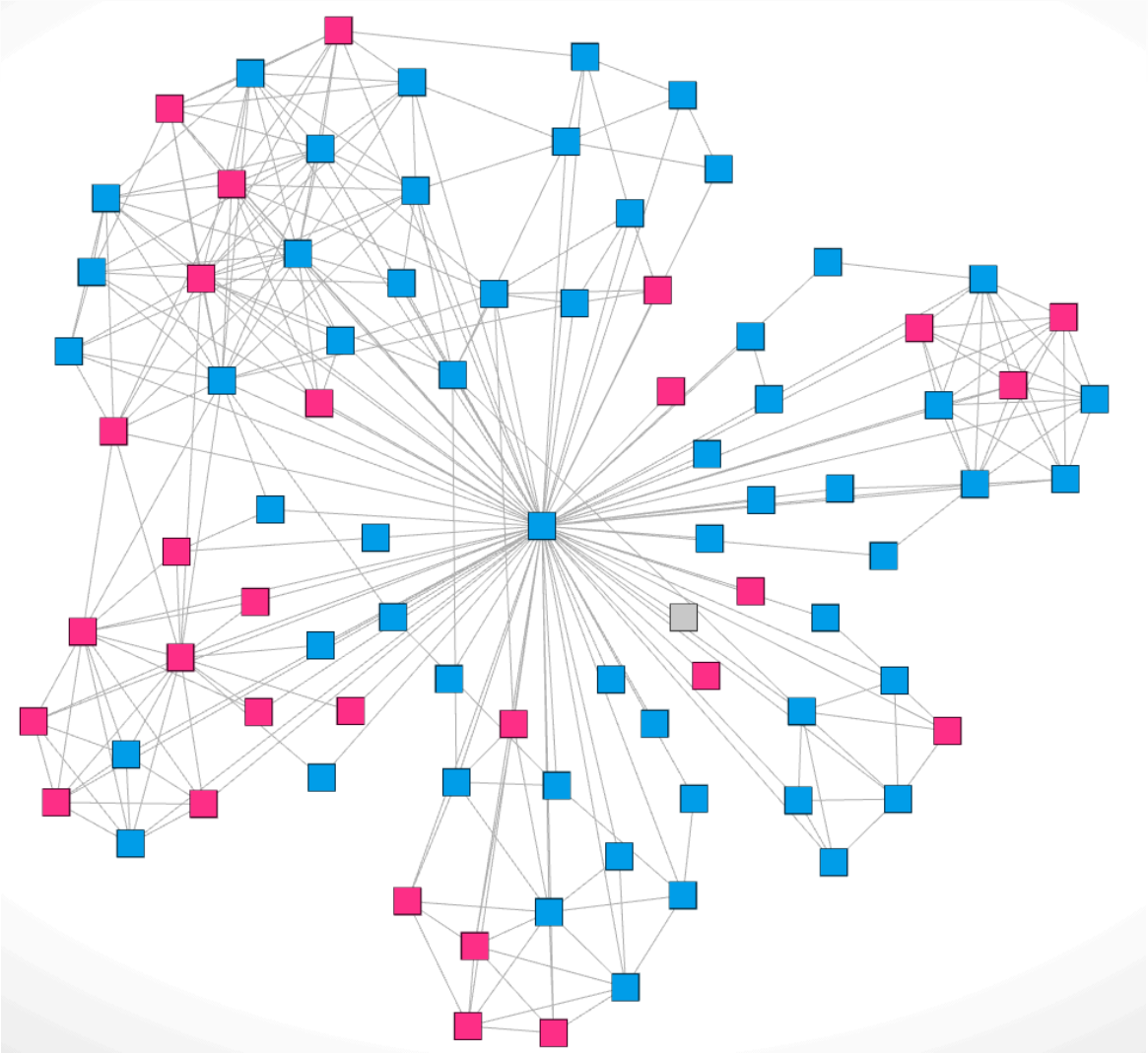
Feature	Perl	SQL	HBase	Linda	Comm BLAS	Tensor Toolbox	UPC	VSIPL++	pMatlab	D4M
Associative Array										
1D	X	X	X	X						X
2D	X	X	X	X						X
String key/value	X	X	X	X						X
Numeric key/value		X	X	X						X
Composable query										X
Composable compute					X	X	X	X	X	X
Tuple Store			X							X
Parallel Client			X	X	X		X	X	X	X
Distributed array				X	X		X	X	X	X

- **D4M features can be found across a wide range of technologies**
- **D4M uniquely combines these features into a composable language for algorithm development**

Source: Kepner et al., 2012, Lincoln Laboratory, MIT

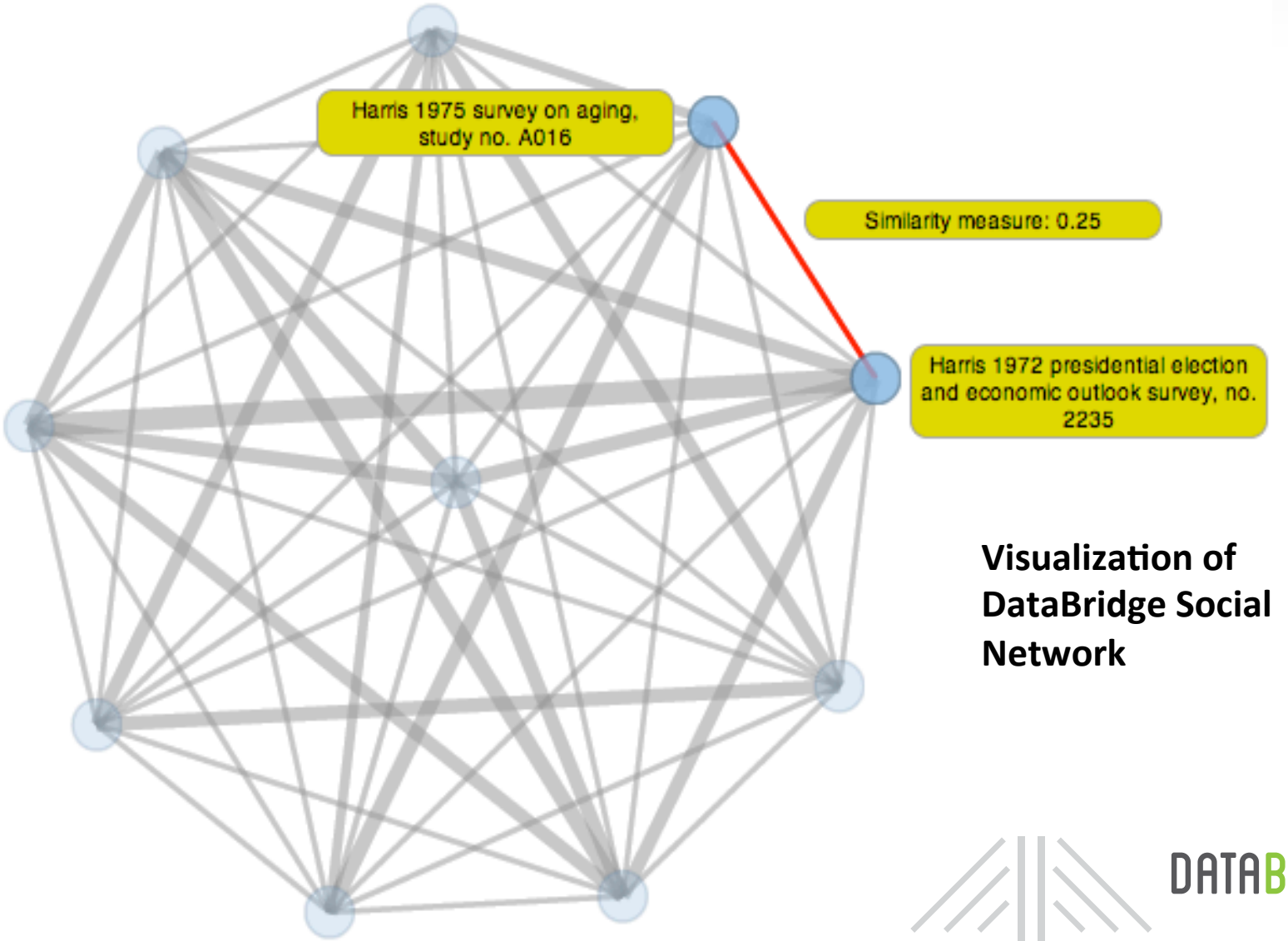
SQL (System Query Language), UPC (Universal Parallel C), VSIPL++ (Vector, Signal and Image Processing Library)

The DataBridge: A Social Network for Data



Visualization of Facebook's Social Network

The DataBridge: A Social Network for Data



Visualization of DataBridge Social Network

DATABRIDGE

DataBridge Strategy and Team

- **Strategy:** *Construct a multi-dimensional sociometric network for data that examines how we: Evaluate similarities of data sets, detects the resulting set of similarities, provide query interfaces on resulting multi-dimensional network?*
- **Collaborators:**
  - Odum Institute, *UNC-Chapel Hill*
  - Population Informatics Research Group, *UNC-Chapel Hill, Texas A & M University*
  - iLab, *North Carolina A&T University*
  - Institute for Quantitative Social Science, *Harvard University*

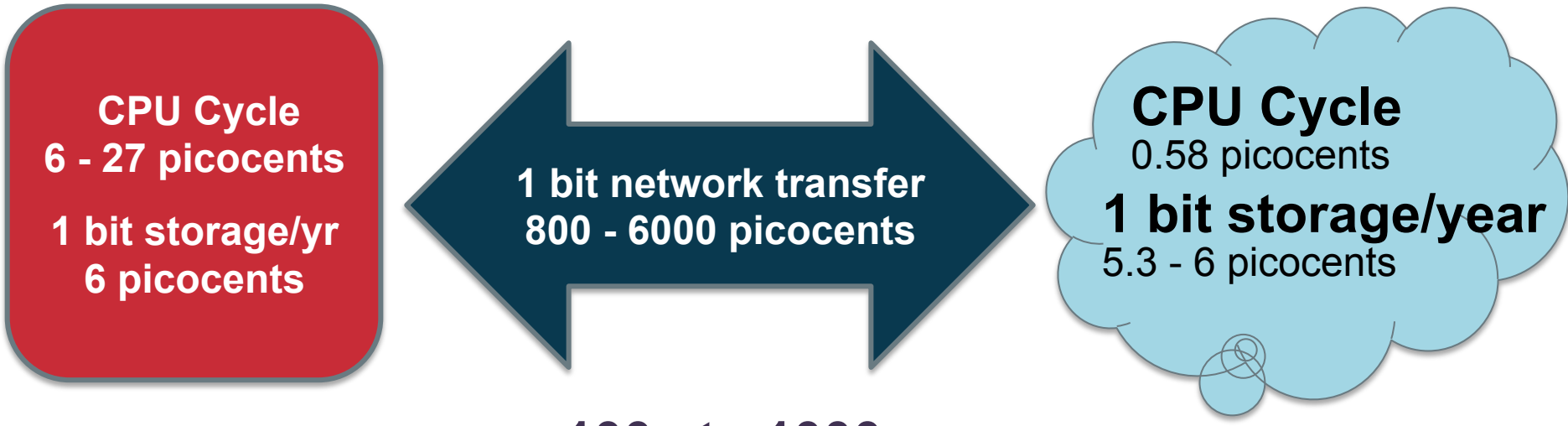
Winner: Best Paper Award, ASE/IEEE Conference on Big Data, September 2013

- Context
- Why Data Science?
- Meeting the Challenges and Opportunities of Data Science
- The Mathematics of Data
- **The Economics and Ethics of Data**
- Possible Approaches to the Data Challenge
- NCDS
- Conclusion

The Economics of Volume …

The Current Transmission Gap

Cycles are 10x to 50x cheaper in the cloud!



Site-based

100x to 1000x more costly

Cloud-based

Adapted from: Radu Sion, Stony Brook University, 2009

A pico-cent($) is approx equivalent to a pico-cent(€)

What can we do about this?

What drives the cost of moving data? Energy!

- **New data centers:**
  - **Maiden, NC,** 4,356,000 sq ft. Apple, 20MWatt, RENEWABLE ENERGY, 20MWatt Solar Array!
  - **Champaign-Urbana, Il**, NCSA, Blue Waters facility
  - **Boulder Colorado**, 152,000 sq ft. US Department of Energy National Renewable Energy Lab's Energy Systems Integration Facility (ESIF).
  - **Tokyo Japan**, Tokyo Institute of Technology, according to the Green500 list, uses HP ProLiant servers to operate the world's most energy efficient production petascale supercomputer.
  - **Reykjavik Iceland**, in a data center built by Verne Global one of the first occupants, a consortium of Nordic universities,
- All of these are being built around **ONE CONSTRAINT: ENERGY**
- A quick look at the power usage of proposed Exascale systems explains why:
  - Today's best process :70 picojoules per floating point operation (FLOP). By 2020: 5-10 picojoules per FLOP.
  - Thus: 2020 Exascale system would require 5-10 megawatts to perform 1 ExaFLOP of calculations.
- **BUT!**
  - The energy cost of moving two 64 bit operands in and out of the processor is estimated at 1000 to 3000 picojoules per FLOP
- Thus, approx 1 gigawatt would be required for an Exascale system, well outside the capabilities of even the world's largest data centers

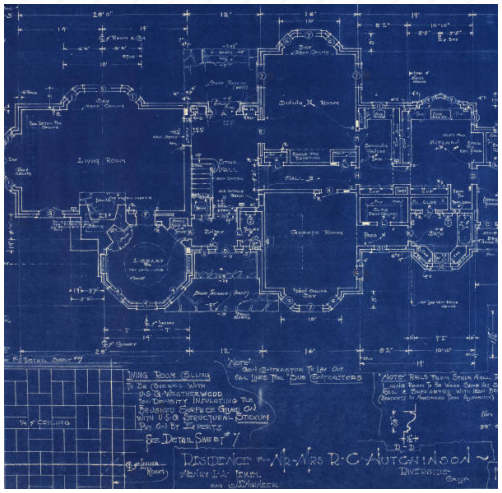
ISC HPC Blog, **The Worldwide Quest for Energy Efficient Supercomputing**, Posted: 05-02-2012 16:05
Other misc. sources.

Ethical issues are critical:

- Data accrues more quickly than consensus on how to interpret it
- ID of life-changing information now possible:
  - *e.g., general genomic screening can find risk of early onset Alzheimer's and other genetic diseases*
- Data isn't perfect: in medicine, false negatives and positives are possible, some data is distressing:
  - *Who should have access to it? Is it right to keep distressing data from patients when nothing can be done?*

Data Ethics and the Law

- Currently are no protections for informational property rights
  - *US generally allows covert collection of DNA evidence in criminal cases*
- Distinctions must be made between privacy, confidentiality and security.
  - *Different meanings to technologists, lawyers, clinical researchers, etc.*
- Should genomic data be treated the same as other personal data?



- Context
- Why Data Science?
- Meeting the Challenges and Opportunities of Data Science
- The Mathematics of Data
- The Economics and Ethics of Data
- **Possible Approaches to the Data Challenge**
- NCDS
- Conclusion

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

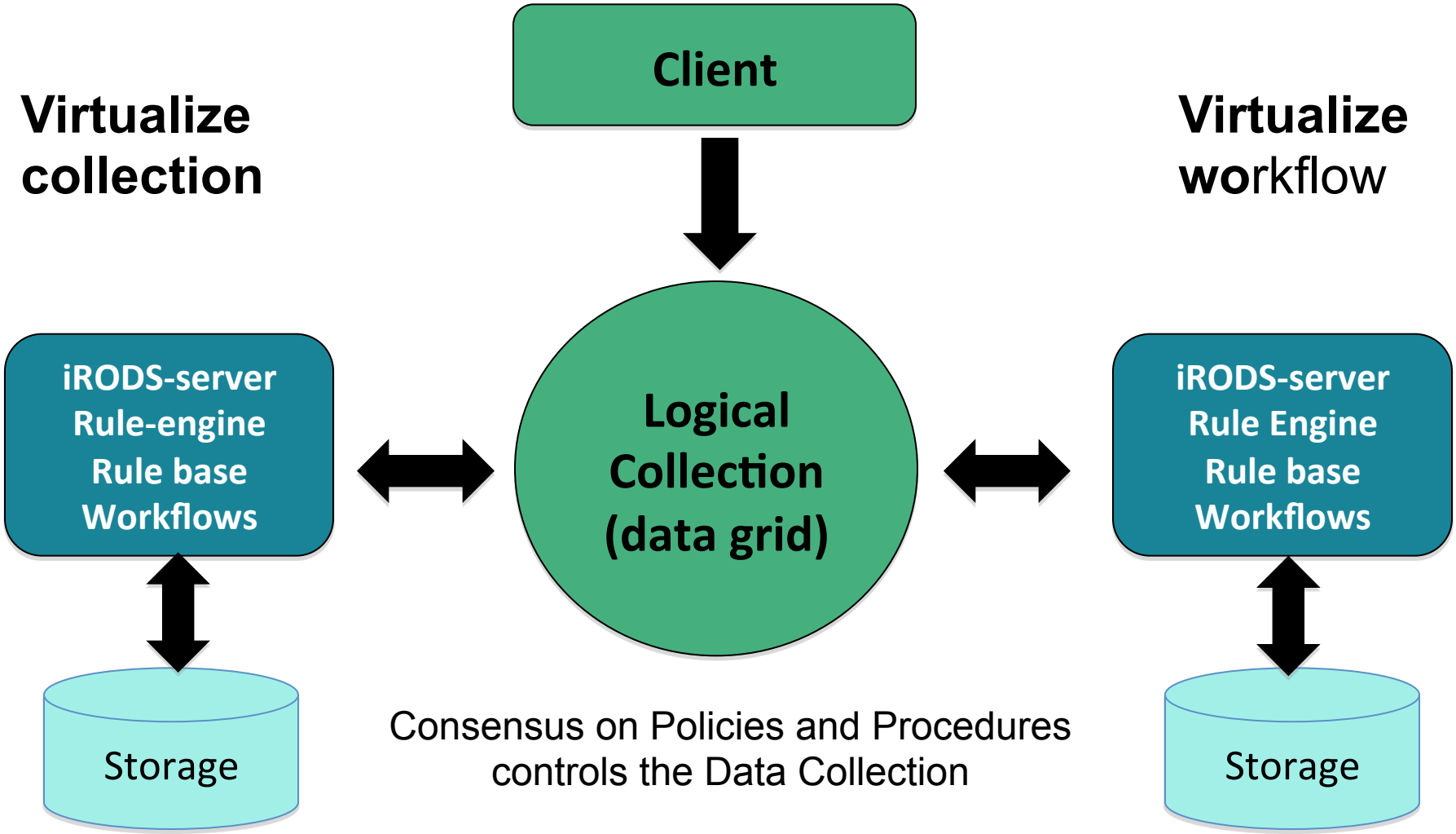
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci RENAISSANCE COMPUTING INSTITUTE

Coming to Grips With Big Data

- **Computation:** Reproducible data-driven research through integration of analysis workflows with data management.
- **Distribution:** Execution of procedures at storage location.
- **Validation:** Sophisticated technologies for verifying assessment criteria.
- **Virtualization:** Distributed data access via networks to a federated system that may be geographically decentralized.

Policy-based Data Management

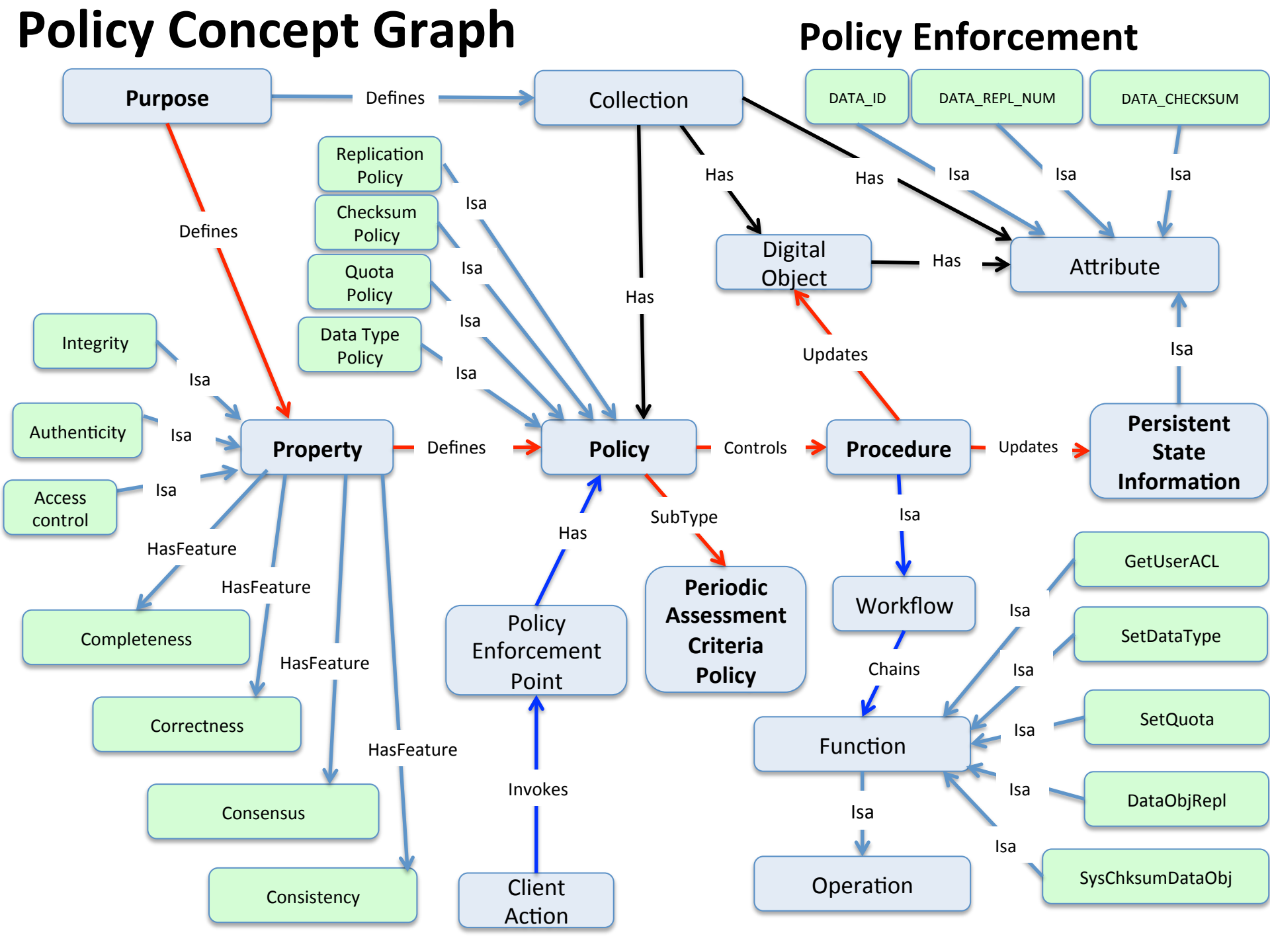


Policy-Based Data Management

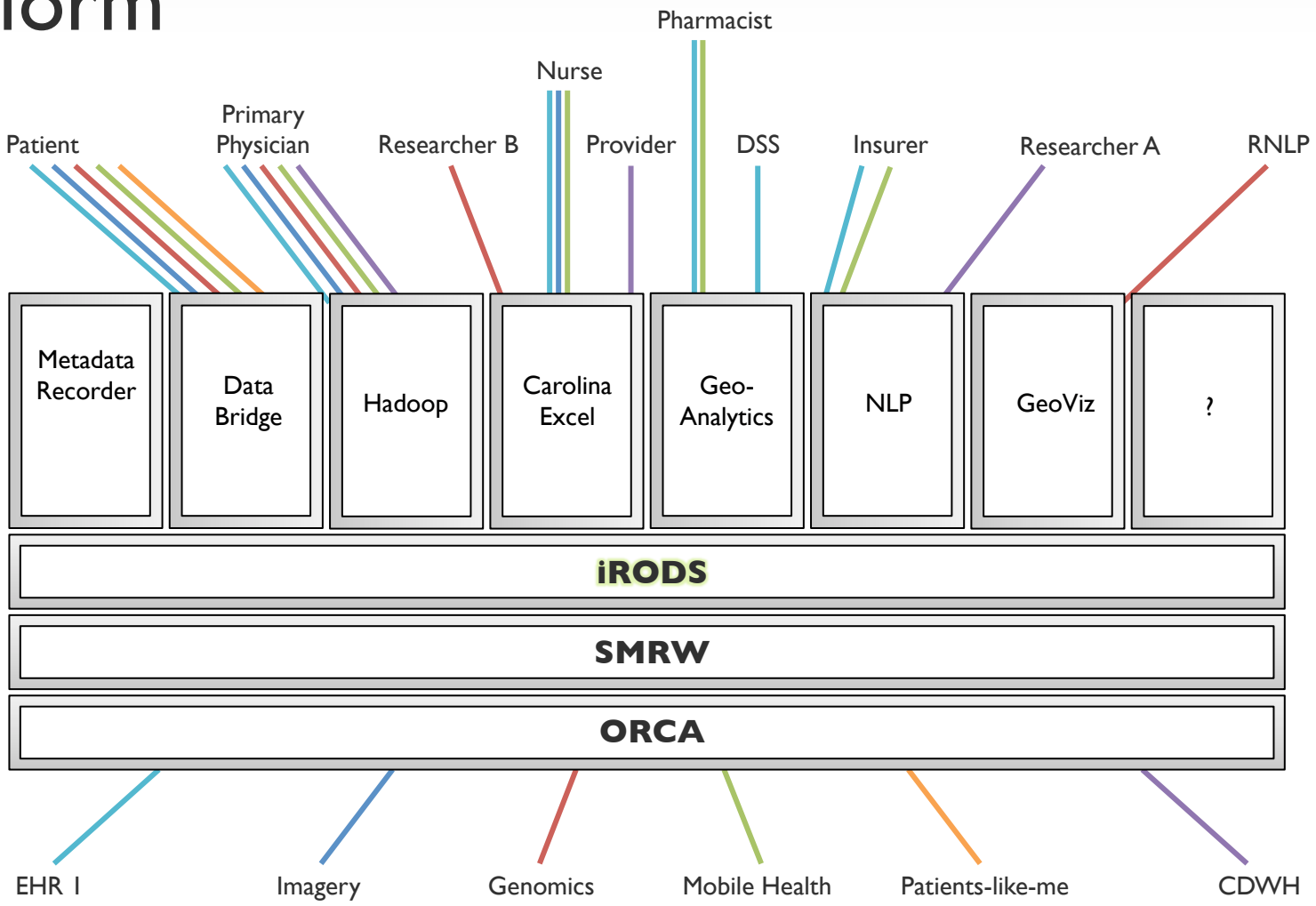
- **Purpose:**
  - *The reason a collection is assembled*
- **Properties**
  - *Attributes needed to ensure the purpose*
- **Policies**
  - *Enforce and maintain collection properties*
- **Persistent state information**
  - *Results of applying procedures*
- **Property assessment criteria**
  - *Validation that state information conforms to desired purpose*
- **Federation**
  - *Controlled sharing of logical name spaces*

These are the necessary elements for collection management

Policy Concept Graph



Initial Technology: Next-Gen Research Platform



- Context
- Why Data Science?
- Meeting the Challenges and Opportunities of Data Science
- The Mathematics of Data
- The Economics and Ethics of Data
- Possible Approaches to the Data Challenge
- **NCDS**
- Conclusion

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci RENAISSANCE COMPUTING INSTITUTE

The National Consortium for Data Science

www.data2discovery.org

- **Mission:** Secure US role as leaders in data science research & education, position US industry to use the power of data to drive economic growth
- **Vision:** Focused multi-sector, multidisciplinary data science community to solve big data challenges and drive the field forward
- **Goals:**
  - **Engage** broad communities of data experts
  - **Coordinate** data science research priorities that span disciplines and industries
  - **Facilitate** development education & training programs
  - **Support** development of technical, ethical & policy standards
  - **Apply** NCDS expertise to data challenges in science, business and government

NCDS is a strategic approach to data science and big data opportunities

NCDS Founding Members

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci

CISCO

GE

NC STATE UNIVERSITY

IBM

sas

Drexel UNIVERSITY

Duke UNIVERSITY

RTI INTERNATIONAL

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

UNCCHARLOTTE

NetApp

TEXAS A&M UNIVERSITY

MCNC
Connecting North Carolina's Future Today

NIEHS
National Institute of Environmental Health Sciences

NCDS Components

- **Data Observatory**
  - Shared, distributed infrastructure housing large organized research data; platform for data science education
- **Data Laboratory**
  - R&D into critical tools and techniques for data science
- **Data Fellows program**
  - Seed grants for faculty and post-docs to work on consortium-approved projects; NCDS review panel will evaluate proposals
  - Industry internships for graduate students
  - Visiting industry data scientists at member universities
- **Data Science Events**
  - Leadership Summits (Spring)
  - Outreach events and speakers (Fall and Spring)

ncds

NCDS Foundations

- Shared, distributed infrastructure will be the foundation for the NCDS Data Observatory and a Data Laboratory, a virtual lab providing access to tools and infrastructure needed to test techniques for storing, sharing, analyzing, transforming, and visualizing data.

Year-one Focus

- Create initial sets of federated data collections.
- Document and integrate set of initial tools
- Pilot a data science education platform comprised of compute, storage and data management tools for classroom use
- Target data-intensive courses across multiple disciplines
- Offer 2-3 courses, expand in subsequent years
- Data sets and tools/software to be contributed by NCDS members
- Distribute hosting model

www.data2discovery.org/data-observatory

NCDS Data Science Faculty Fellow Program

- Will foster private-public relationships, engage future data scientists, bridge gaps between research and practice, create NCDS-sponsored scholarship

Timeline

Mid September: RFP released

November 1: Proposal due

November 15: Notification of acceptance

Year-one Focus

- **Seed grant** approach to fund initial cadre of Fellows from NCDS academic member campuses
- **Teaming with an NCDS member** encouraged, but not required; potential for future collaboration part of review criteria
- Funds used for **course buy-outs, summer salary, graduate student support**, conference travel and modest infrastructure costs
- Target: **3-5 awards in year 1, $30K each**

www.data2discovery.org/data-fellows

Support provided by UNC General Administration to offer fellowships to all UNC System campuses

First NCDS Leadership Summit

Data to Discovery: Genomes to Health, April 23 – 24, 2013

- Keynote address: **Dr. Eric Green**, Director, National Human Genome Research Institute,
- First in annual **Leadership Summits** on big data issues in targeted domains.
- **Purpose:** Focused discussion by top data and domain scientists to elicit key data problems and opportunities
- **Final Product:** White Paper on data challenges and opportunities in genomic science. *Summary version under review for publication by a major scientific journal.*

Next Leadership Summit:
Working Title: Sustainability in the 21st Century: "Big Data for Smaller Carbon Footprints"
April 2014, Chapel Hill, NC

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE renci

iRODS Enterprise Distribution

Production version of iRODS

- Managed by iRODS Consortium
- DICE recommended distribution

3.0 now released

- 3.01 due Fall 2013
- 3.4 due Winter 2013-14

irods
enterprise data management

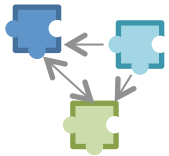
- Binary releases, open source
- Focus on testing and reliability
- Pluggable framework

iRODS Enterprise Distribution: a pluggable framework for data grid technology

Core is a substrate upon which new functionality may be added via seven interfaces. The core is designed to be a small, stable broker of extensible services.

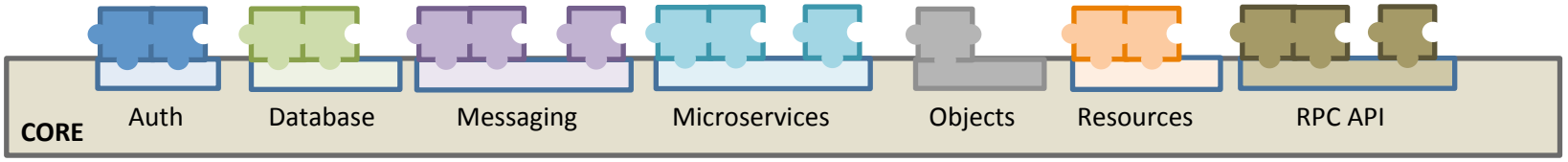
Interfaces for Extensibility:
Authentication, Database, Messaging, Microservices, Objects, Resources, RPC API

Plugins extend the functionality of E-iRODS relevant to a given interface. They are self contained, dynamically loadable, and could be proprietary.



Includes a plugin dependency model. Plugins may be inter-dependent and provide new functionality via multiple plugins.

A Bundle of plugins can provide a set of features to support newly created first-class objects within iRODS such as Tickets or Workflows.

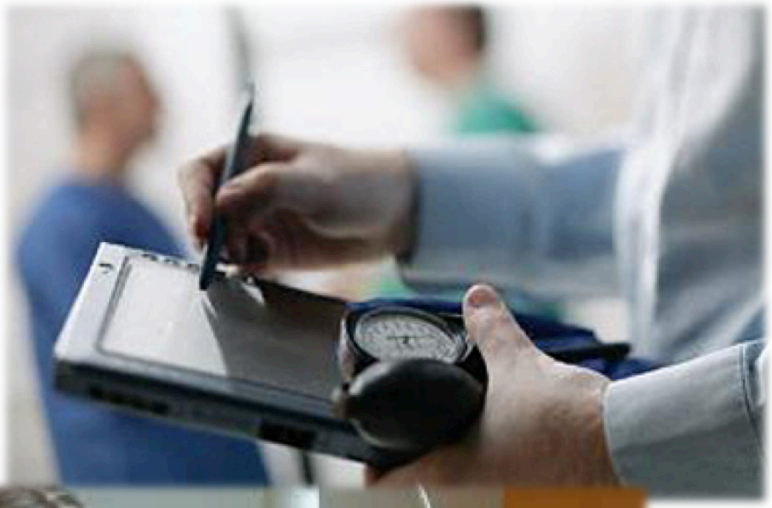


Secure Medical Workspace

A secure "virtual desktop" where researchers can work with sensitive data

1. Safeguard Protected Health Information (PHI) data
2. Enable medical and translational research

Key Technology Across Domains



- Context
- Why Data Science?
- Meeting the Challenges and Opportunities of Data Science
- The Mathematics of Data
- The Economics and Ethics of Data
- Possible Approaches to the Data Challenge
- NCDS
- **Conclusion**

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

renci RENAISSANCE COMPUTING INSTITUTE

Developing Data Science Will:

- ***Develop*** the next generation of data science experts and leaders
- ***Create*** strategies, practices, and scientific methods for understanding data
- ***Enable*** more collaborations among data and domain scientists, business, academia and government
- ***Assist*** those who are struggling to collect, analyze, manage and use data
- ***Establish*** methodologies for measuring the value and impact of data

Questions?