Genomes Galore – Big Data Analytics for High Throughput DNA Sequencing

Srinivas Aluru, Iowa State University (Email: aluru@iastate.edu)

Driving Grand Challenges

- Identification of complex disease traits
- Detection of biological threats
- Microbial studies and human health
- Plant genotype to phenotype

⋮

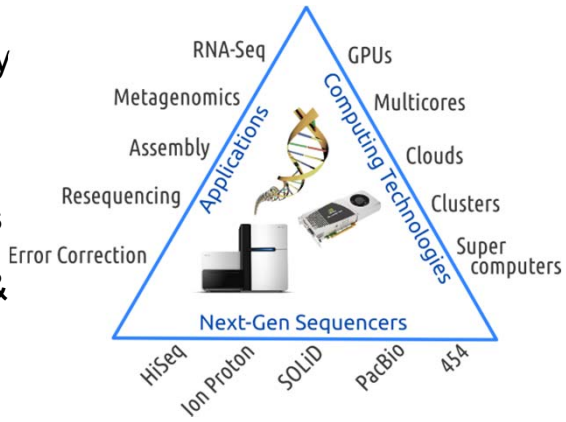
⋮

The Big Data Challenge

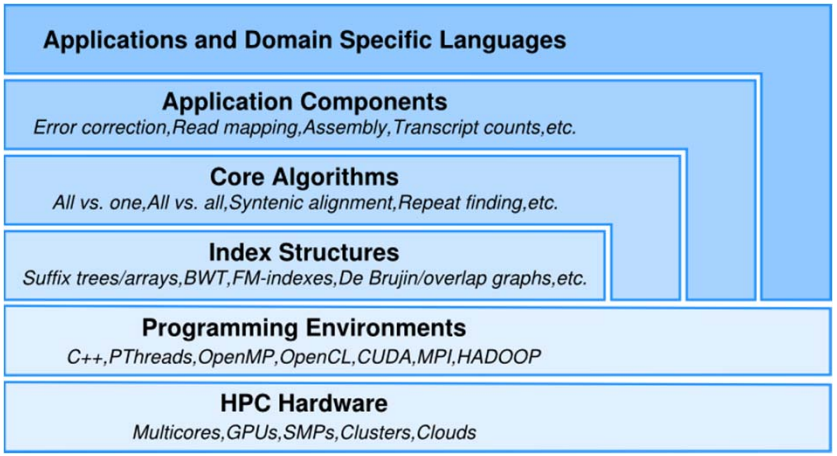
Then (2005)	Now
	
ABI 3700 96 ~800 bp reads 76.8 X 10^3 bases ~$1 per kilo base	Illumina Hiseq 2500 6 billion 100 bp reads 600 X 10^9 bases ~$1 per 200 million bases

Vision and Goals

- Empower community migration to HPC
- Preserve ability to create new solutions
- Target researchers & software developers



Research and Dissemination Approach



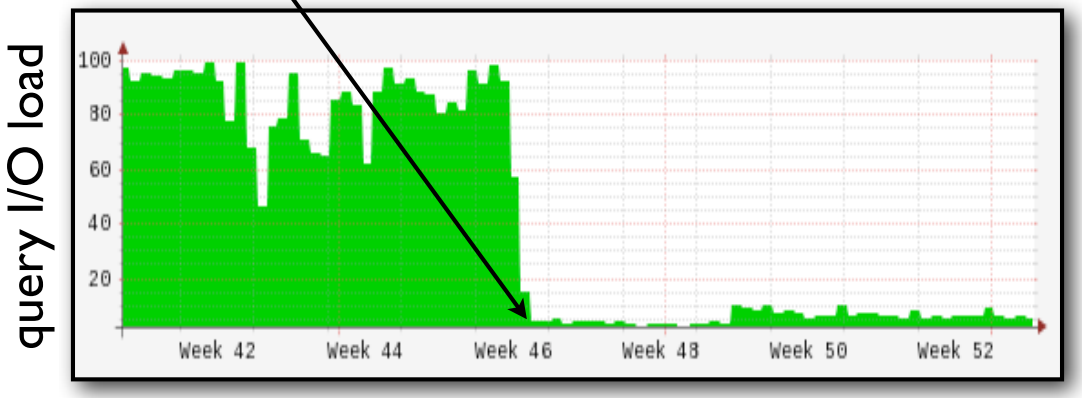
NSF IIS-1247716; Team: Iowa State University, Stanford University, U. of Michigan, and Virginia Tech

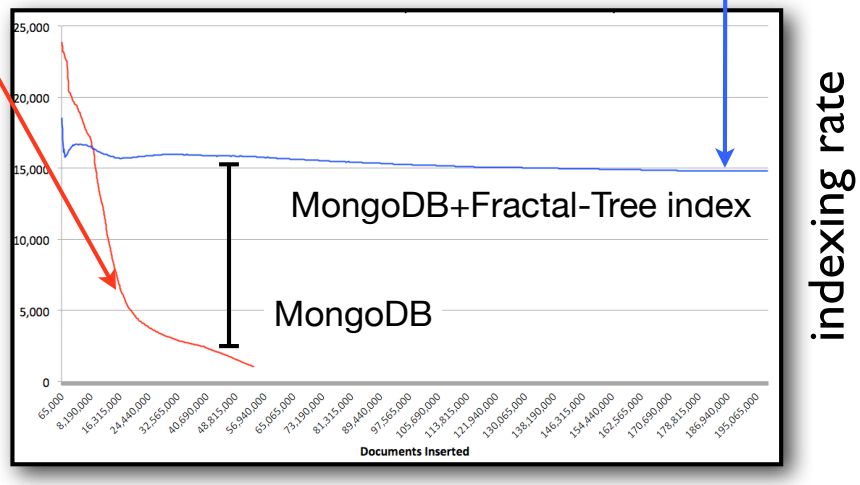
Write Optimization: The Future of Big Data Storage

Adding the right index leads to faster queries.

Index maintenance has been notoriously slow.

Write-optimized Fractal Tree indexes ingest data 10x-100x faster.





Write optimization impact on SSDs:

- Better wear-out.
- Better compression.

NSF Big Data grant: Algorithmics of write-optimization.

NSF: Supported basic research.
NSF & DOE: Supported tech transfer.

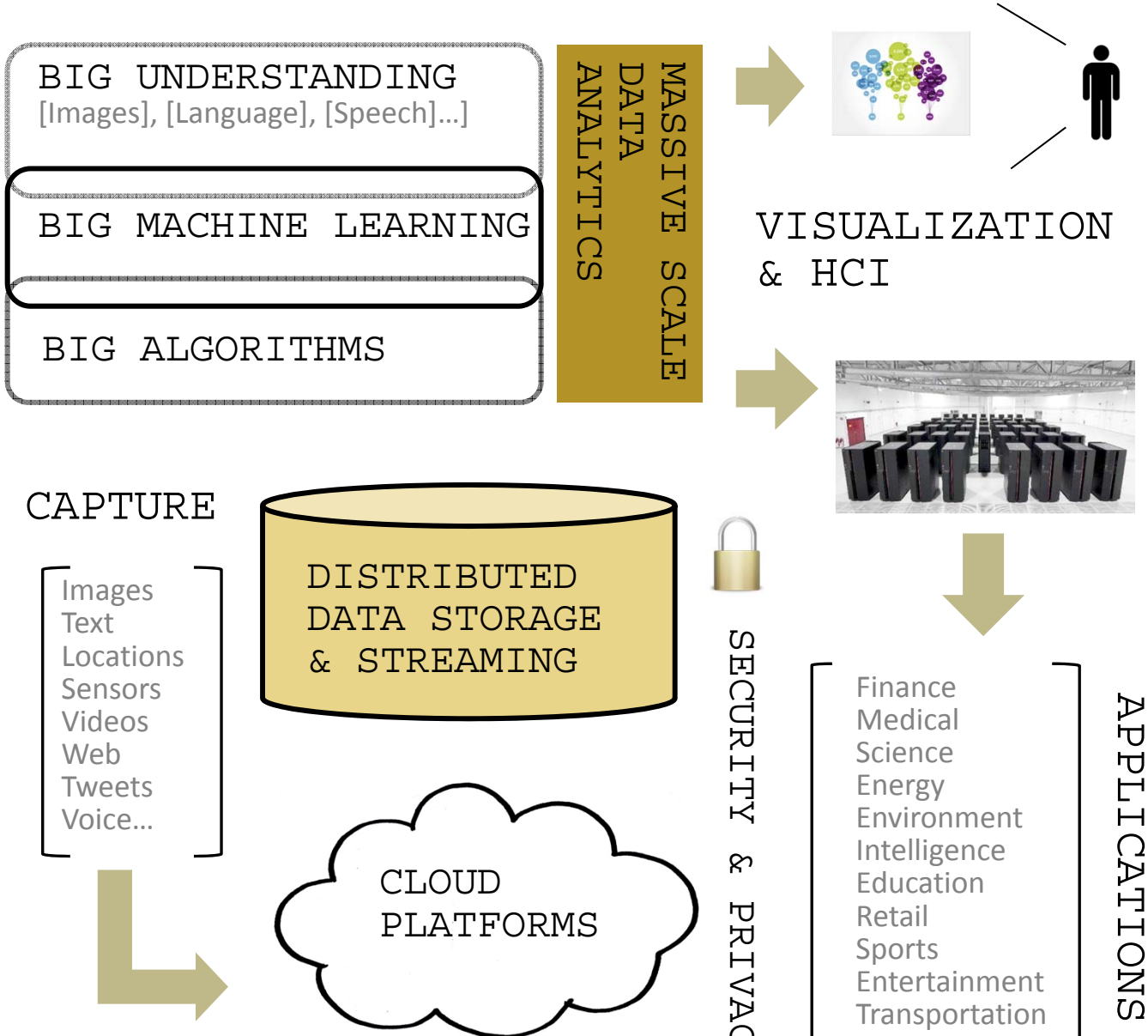
MIT Big Data Initiative at CSAIL

Mission. The goal of the MIT Big Data Initiative at CSAIL is to collaborate with industry and government to identify and develop new technologies needed to solve the next generation data challenges which will require the ability to scale well beyond what today's computing platforms, algorithms, and systems can provide.

We want to enable individuals and organizations to truly leverage Big Data by developing tools and platforms that are reusable, scalable and easy to deploy across multiple application domains.

Industry Members.

AIG, Alior Bank, BT, EMC², facebook, Huawei, intel, Microsoft, Quanta Computer, SAP, Samsung, Thomson Reuters



bigdata @CSAIL

MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

http://bigdata.csail.mit.edu/

Big Data Science for the Masses: start small, Think Big

@KirkDBorne, George Mason University

Visualize This:

A sea of Data (sea of CDs)

This is the CD Sea in Kilmington, England

(600,000 CDs ~ 300 TB).

"Big Data" is different!

We need more Data Scientists in order to discover the unknown unknowns in BIG DATA collections more efficiently and more effectively.

1) Informatics in Education

Work with data in all learning settings:

- Informatics (Data Science) enables transparent reuse and analysis of data in inquiry-based classroom learning.
- Learning is enhanced when students work with real data and information (especially online data) that are related to the topic (any topic) being studied.
- http://serc.carleton.edu/usingdata/ ("Using Data in the Classroom")

2) An Education in Informatics

Students are specifically trained to:

- access & query large distributed data repositories;
- conduct meaningful inquiries into data;
- mine, visualize, and analyze the data;
- make objective data-driven inferences, discoveries, and decisions; and
- communicate "stories" through data.

Provider

Individuals/Patients

Better Care Coordination
& Patient Engagement

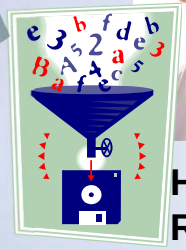
EHR

mHealth App
Remote Device
PHR, etc.

Self-Generated
Health
Information
Exchange

Payer

Lower Costs



Health
Researcher

Comparative
Effectiveness
Research

CDC

Improved
Public
Health
Surveillance

Pharma

Development of
Safer & More
Effective Drugs

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

RTI
INTERNATIONAL

OPEN SCIENCE FOR BIG DATA

The Dataverse Network® Project

Find, Share, Cite, Reuse, Reproduce Research

NOW

Large # of small data sets

Data ↔ Publication

Persistent data citation; open access; multiple formats

COMING SOON

Small # of very large data sets

- Citation of a subset
- Provenance
- Real-time Explore
- Combine Data
- Crowdsourcing

Downloadable, fixed-size data files (KB – GB)

Data + Metadata

Largest collection of social science data (>700,000 files); astronomy data

Big Data (TB - EB)

Frequently Updated — FOREVER DATA

Real-time Streaming — MapR

Privacy Sensitive — MyDataCan™

Examples:

Connectome

Pan-STARRS

Health

Education

Social Media

Crosas, Sweeney, Waldo, Seltzer, Goodman, Dominici, King, Pfister, Lewis Mostak, Knowles-Barley, Jones, Bar-Sinai

thedata.org

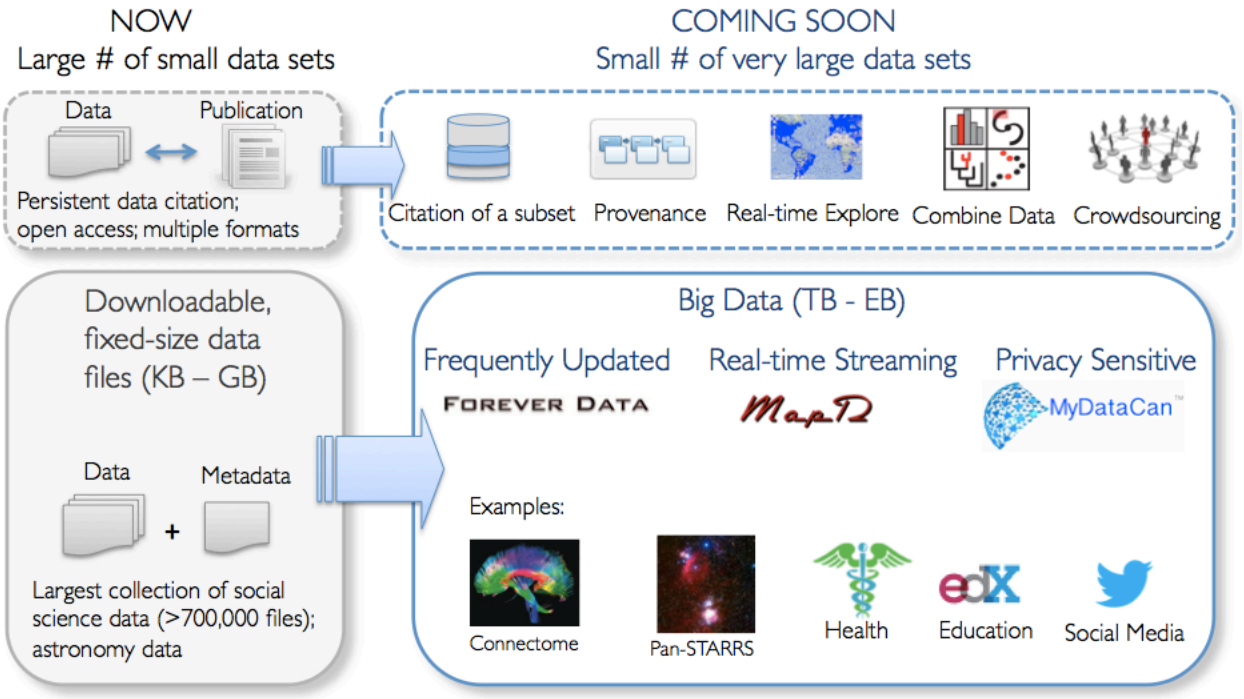
IQSS HARVARD UNIVERSITY

Dataverse for Big Data

Mercè Crosas, Latanya Sweeney, Gary King

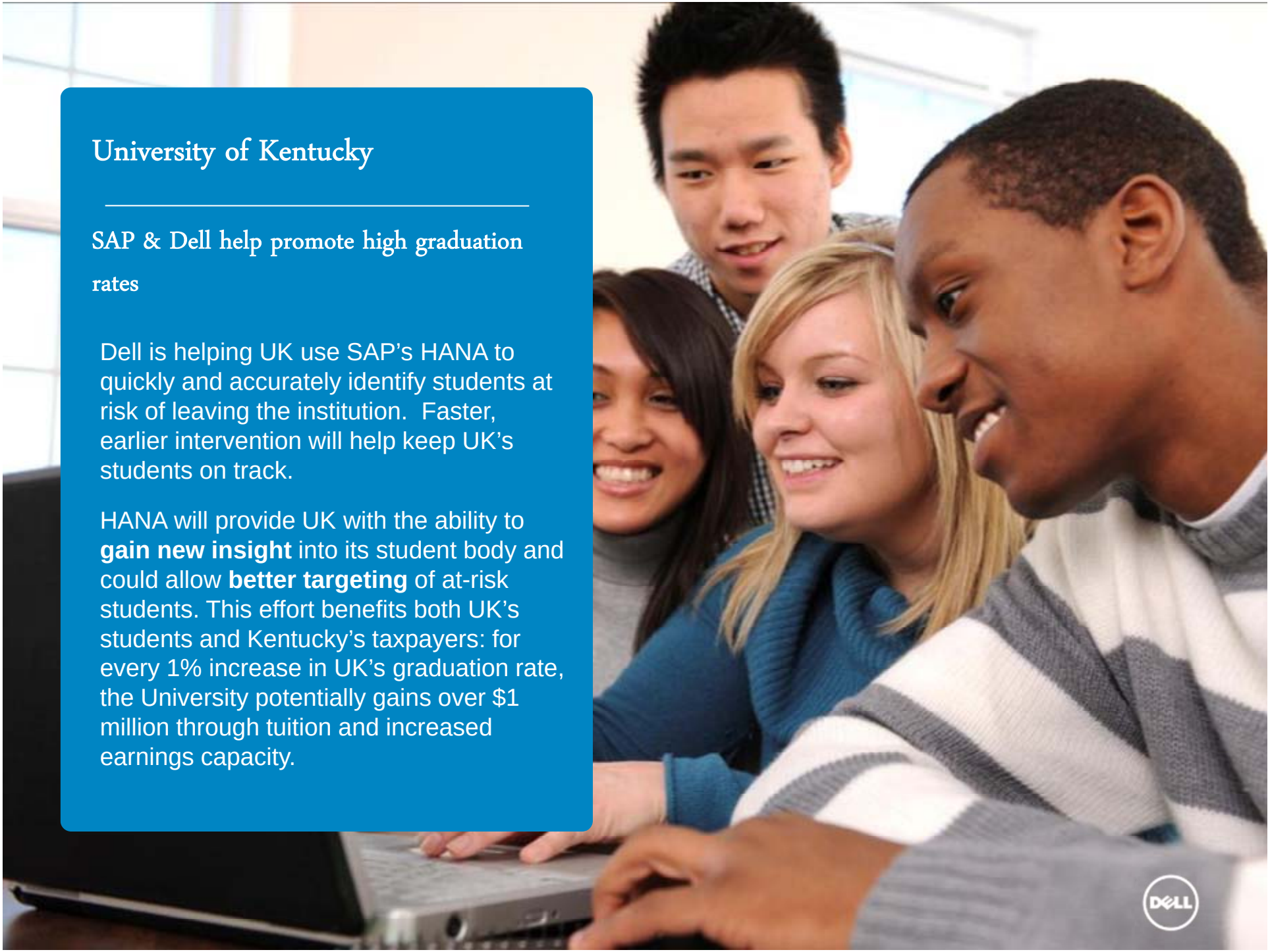
Harvard University[1]

The Dataverse for Big Data will make very large data sets shareable, citable and reusable, and facilitate reproducibility of research emerging from this new generation of data. Up to now, most scientific research comes from a large number of small data sets. For these fixed-size, downloadable data sets, the Dataverse already provides a solution to find, share, cite and reuse them, having the largest repository of social science data in the world and replicated across institutions worldwide. However, a small, but increasing, number of very large data sets are now being used for research, offering a rich new source of information. One of the challenges with these large, frequently updated or streaming data sets is that they are no longer downloadable to the researcher's computer. Another challenge is that the amount of data cannot be any longer curated and analyzed by the few that collect them, and therefore a substantial fraction of the data remains unexplored. And a third challenge is that often those data sets contain privacy sensitive data. The Dataverse for Big Data will resolve these challenges by providing an open science platform where all researchers will be able to explore, analyze and query (real-time, in some cases) big data sets, persistently cite a subset of the data, get the provenance trail of a data set, use privacy and secure tools to explore sensitive data, and when needed, use crowdsourcing tools to bring additional person-hours to help curate and analyze the data.



The Dataverse platform will integrate with other technologies: Forever Data – a system to harvest frequently updated data sets-, MapD - a database with very fast queries to render streaming data in real-time-, and MyDataCan – a system to securely stored personal data that eventually can be donated to science. This new open science platform will enable distribution of Big Data from a variety of scientific fields. Initially, it will make accessible nanometer-scale images of the brain from the Connectome project, astronomy observations from nearby approaching objects from the Pan-STARRS project, health data from hospital claims, student usage data collected by massive online open courses such as EdX, and social media data.

[1] Primary contact: Mercè Crosas, mcrosas@iq.harvard.edu **thedata.org**



University of Kentucky

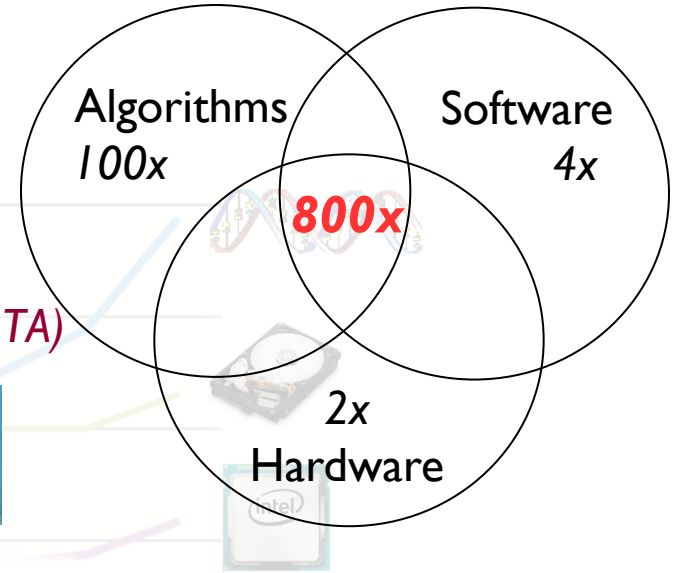
SAP & Dell help promote high graduation rates

Dell is helping UK use SAP's HANA to quickly and accurately identify students at risk of leaving the institution. Faster, earlier intervention will help keep UK's students on track.

HANA will provide UK with the ability to **gain new insight** into its student body and could allow **better targeting** of at-risk students. This effort benefits both UK's students and Kentucky's taxpayers: for every 1% increase in UK's graduation rate, the University potentially gains over $1 million through tuition and increased earnings capacity.

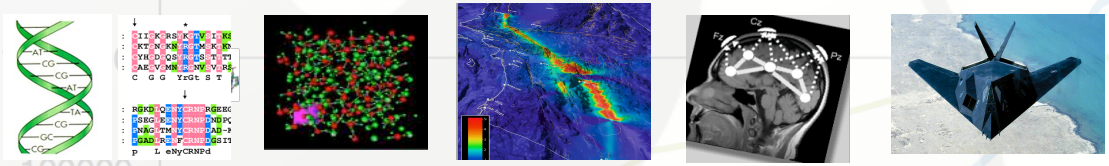
Synergistic Co-Design for BIG DATA

Prof. Wu Feng, Virginia Tech

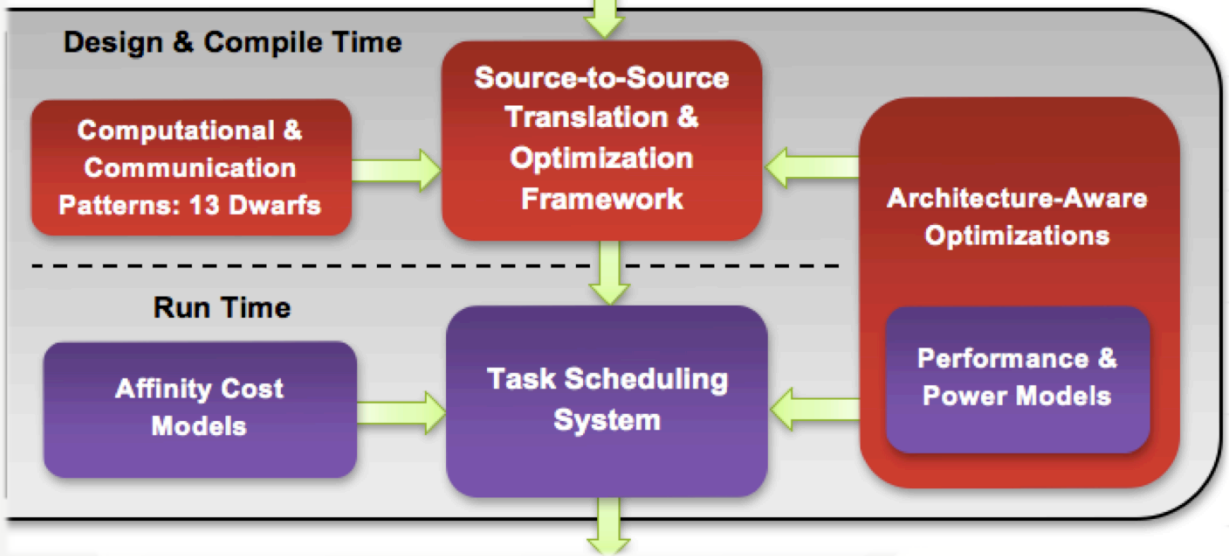


- Changing Landscape
  - … from *FLOPS ("old HPC") to bytes ("new HPC" → BIG DATA)*

Apps



Software Ecosystem



Hardware

Co-Design Exemplars

1. *ParaMEDIC: Parallel Metadata Environment for Distributed I/O & Computing (yrs → mins)*
   → Find missing genes
   http://archive.isgtw.org/?pid=1000811

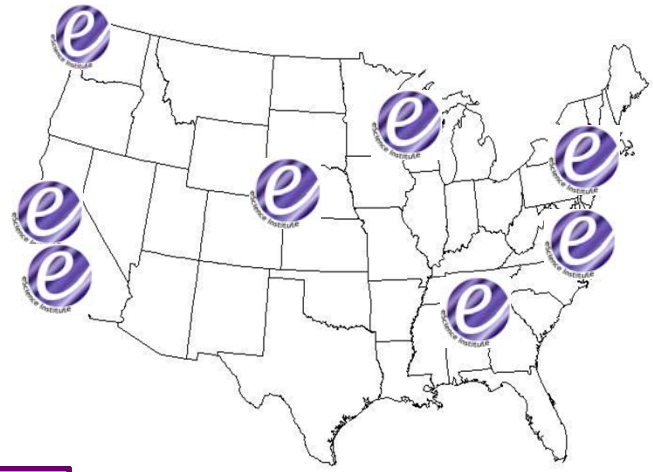
2. *Molecular Modeling*
   → Rational drug design
   http://www.youtube.com/watch?v=zPBFenYg2Zk

3. *Temporal Data Mining of Brain*
   http://synergy.cs.vt.edu/pubs/papers/feng-temporal-data-mining-gpu-gems-2011.pdf

Incubator

- Seed grants to students and postdocs
- Rotating staff from science **and** industry
- An evolving portfolio of reusable services
- A network of cross-boundary **partnerships**
- *Produce digital capital **and** human capital*

2018



2013

Data Science Incubator

2008

eScience Institute

Some local observations:

- Big data work exposes common ground
- Every job is becoming "data scientist"
- More π-shaped people!
- Democratization to the long tail is key
- *Industry and research aren't too different*

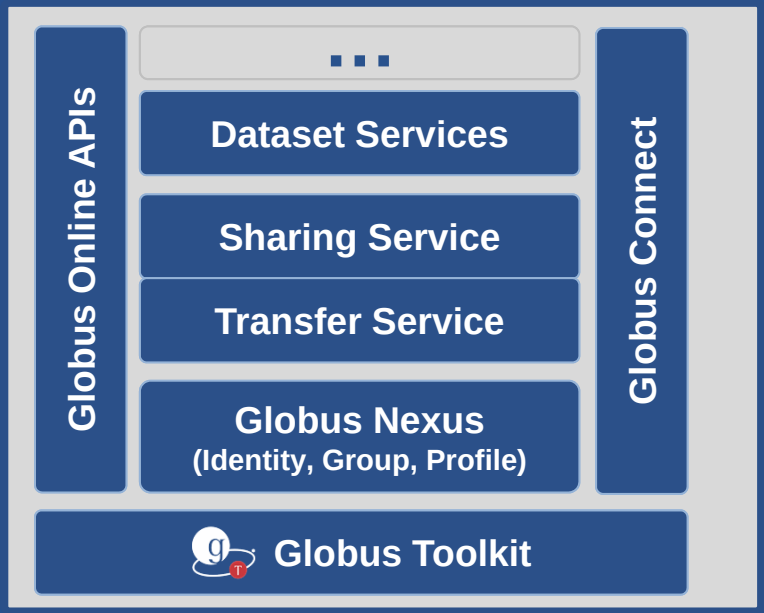
MITRE's Big Data Analytics Activities

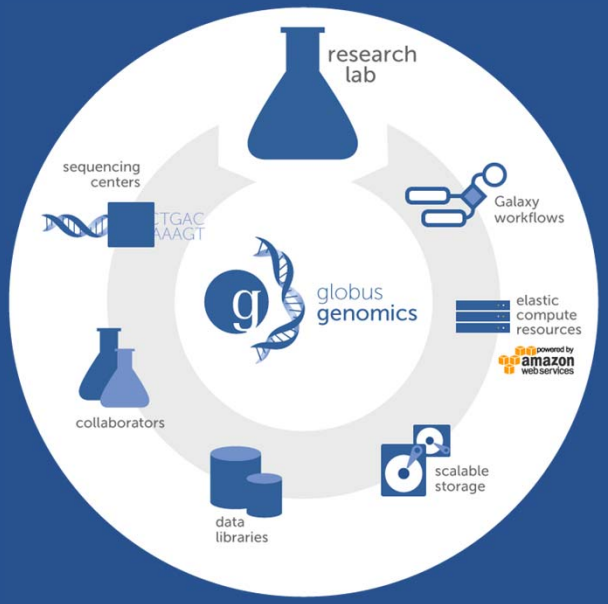
- The MITRE Corporation is a not for profit company that runs multiple FFRDCs, chartered *in the public interest*
- We have expertise in the full range of big data technologies, including high performance computing, complex event processing, parallel relational databases and analytic cloud computing
- The MITRE Innovation Program invests in game-changing research: cyber security, healthcare informatics, counter fraud, enhancing intelligence analysis, and core computational approaches for big data
  - Our researchers partner with government, academia, industry
- The MITRE Institute runs a series of internal training courses on big data and cloud computing
- MITRE's Enterprise Computing Environment provide on-demand computing and storage services to our projects
- We work across missions and agencies

See also http://www.mitre.org/news/digest/advanced_research/06_12/data_analytics.html

It should be trivial for all researchers to: **Collect, Organize, Move, Sync, Share, Analyze, Annotate, Publish, Search, Backup, & Archive** BIG DATA …but in reality it's very challenging

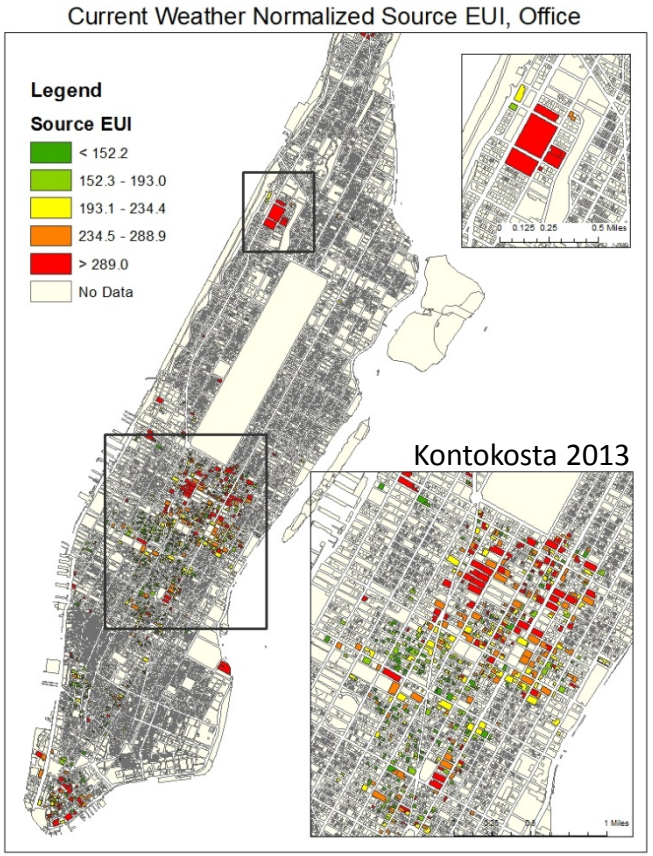
Globus Online uses SaaS approaches to address this challenge and make advanced research data management capabilities broadly accessible



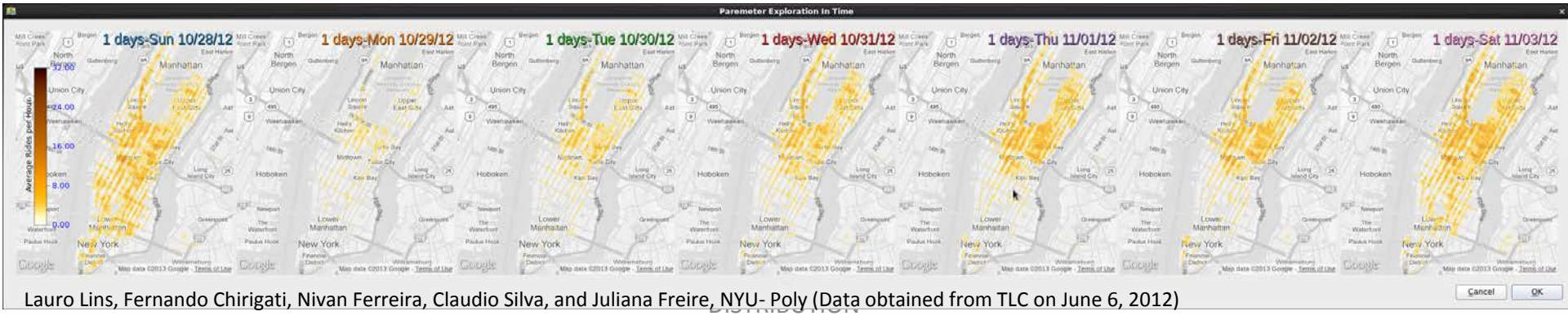


Big Data management +
NGS Analysis pipeline +
Cloud Computing Infrastructure =
Flexible, scalable, easy-to-use genomics analysis for all biologists

CENTER FOR URBAN SCIENCE+PROGRESS



The **Center for Urban Science and Progress (CUSP)** is a unique public-private research center that uses New York City as its **laboratory and classroom** to help cities around the world become more **productive, livable, equitable, and resilient**. CUSP **observes, analyzes, and models cities** to optimize outcomes, prototype new solutions, formalize new tools and processes, and develop new expertise/experts. These activities will make CUSP the world's leading authority in the emerging field of "**Urban Informatics**."



Lauro Lins, Fernando Chirigati, Nivan Ferreira, Claudio Silva, and Juliana Freire, NYU- Poly (Data obtained from TLC on June 6, 2012)

Confidentiality and Data Access in the Use of Big Data: Theory and Practical Approaches

Editors: Stefan Bender, Julia Lane, Helen Nissenbaum, Victoria Stodden

Goals

Identify ways in which vast new sets of data on human beings can be collected, integrated, and analyzed to improve urban systems and quality of life while protecting confidentiality.

Provide both a theoretical and practical foundation which cities across the world can draw from in establishing their data access rules and data security procedures.

Authors

Steve Koonin CUSP, Andrew Gelman, Columbia, Mark Hansen, UCLA, Alessandro Acquisti, Carnegie Mellon University, Helen Nissenbaum, NYU, Kathy Strandberg, NYU, Victoria Stodden, Columbia, Alan Karr, NISS, Jerry Reiter, Duke University, John Wilbanks, Sage Bionetworks and Kauffman Foundation, Sandy Pentland, MIT, Carl Landwehr, George Washington University, Peter Elias, Warwick

Sponsors

ASA
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics

CUSP
CENTER FOR URBAN SCIENCE+PROGRESS

Bundesagentur für Arbeit

Explore & Understand

Link & Explore

Robert Wood Johnson Foundation

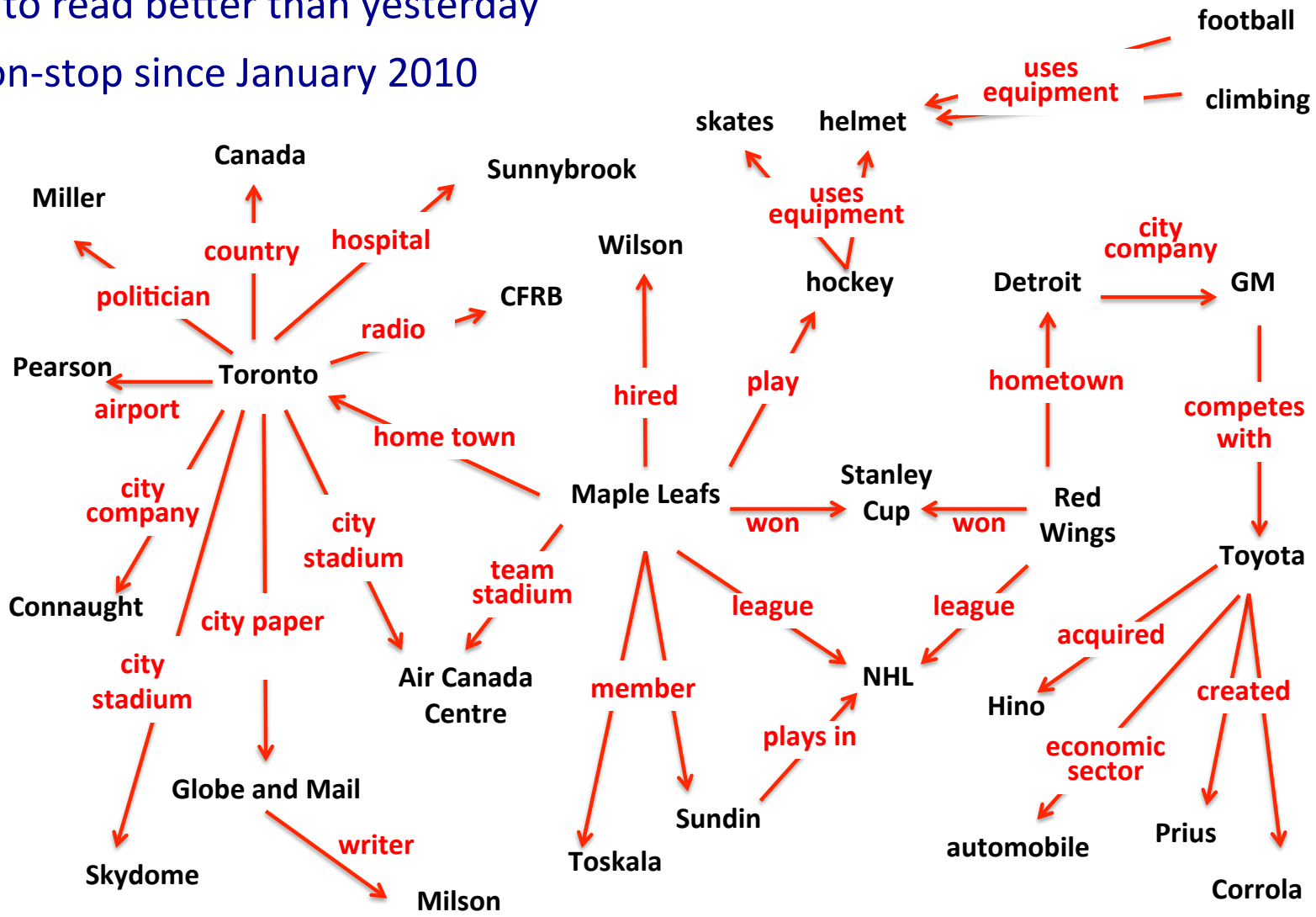
NELL: Learning to Read the Web

Tom Mitchell, CMU
http://rtw.ml.cmu.edu

Each day NELL:

1. Reads more facts from web
2. Learns to read better than yesterday

Learning non-stop since January 2010



Fragment of NELL's ever-growing Knowledge Base, currently containing ~50 Million beliefs

Projects of Interest

Microsoft

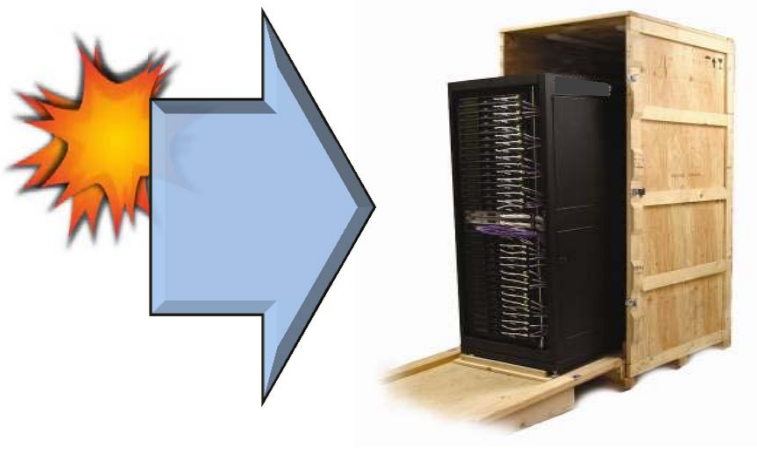
- The Internet of (Orderly) Things
  - Exploitation of digital device information through cloud capture and storage of sensors in industrial or utility automation/control; consumer product telematics; etc.
  - Cloud CEP and historical colorations in 'real enough time' to intervene in process optimization
  - Automotive Telematics: Predictive analytics for maintenance, electric car/utility optimization
  - Industrial Process Automation: Correlate quality with control parameters and process monitoring data; optimization at the system, beyond device or process level
- The Internet of (Disorderly) Things
  - 'Unconventional', use-case or vertical domain-specific, hard to search and analyze data.
  - Satellite crop images; mammograms and tumors; petroleum reservoir inferences; tweets and opinion data; resume verbatims; email analysis; website log files; genomics.
  - Healthcare: Genomic & image analytics, brain mapping, Pharma 3.0 clinical trial optimization
  - Labor Pool: Supply/demand matching
- Large Scale Analytical Computation
  - Interactive, exploratory work leading to scaled 'production systems' which incorporate models and algorithms
  - Reference data (material science, econometric data, government statistics, etc.) needed as part of the analytical process.
  - Consumerization of portfolio performance attribution
  - On demand, evaluate energy impact of design alternatives for architectural designs, structural analysis, kinematic simulation.

SOLUTIONS FOR BIG DATA

Comprehensive

- HPC
- Cloud
- Storage
- Services

Technology Fusion



Leadership

- Speed
- Scale
- Efficiency

Proven

- NASA Earth Exchange powered by SGI ICE
- SGI UV2 -- 64 TB of Shared Memory, 1 O/S
- 8% of Hadoop Clusters run on SGI
- 600 PBs shipped in 2012

www.SGI.com

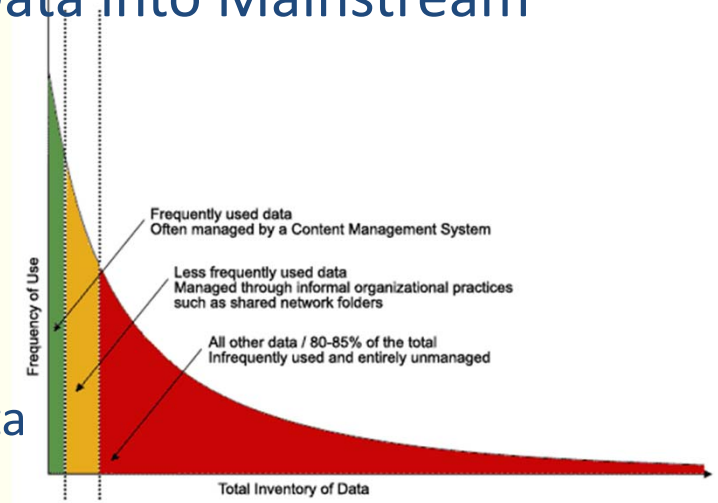
Data Bridge: Solving the First & Last Mile Problems in Big Data

First Mile: Bringing the Long-tail of Science Data into Mainstream

Last Mile: Automation of Linking, Clustering and Discovering Heterogeneous Data

Last Mile: Discover Interesting Relationships

Data Bridge: NSF-funded Big Data Project



- Apply Socio-metric Network Analysis (SNA) to data
- Link through Multi-dimensional vectors
  - Similar to, but for data:
- Explore Relationships between Data, Users, Resources, Methods, Workflows, …
- Architecture: Extensible, Highly Distributed, Plug & Play Algorithms, MyVector
  - Very-loosely coupled Message-Oriented Middleware (using AMQP)
  - Built upon proven technologies: Integrated Rule Oriented Data Systems (iRODS), Dataverse Network

IBM Brings Big Data Skills to the Classroom and Workforce

In the public and private sectors, employers are seeking professionals who can uncover insights from Big Data. However, there is a skills gap. IBM is addressing the challenge.

- Early STEM Learning
- Higher Ed Partners
- Analytics Center
- Workforce Training
- Curriculum
- Case Competitions
- Faculty Grants
- Internships
- Watson Donation

IBM

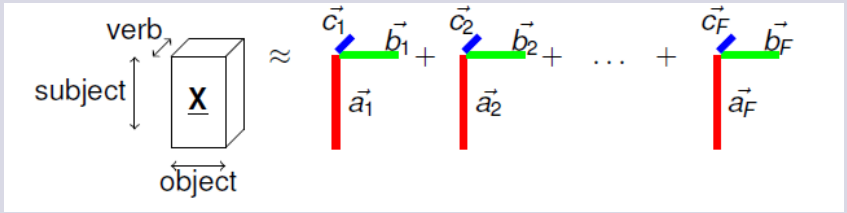
Carnegie Mellon University
NSF IIS-1247632: Big Tensor Mining: Theory, Scalable Algorithms, and Applications
University of Minnesota Driven to Discover℠

C. Faloutsos, N. Sidiropoulos, T. Mitchell, G. Karypis

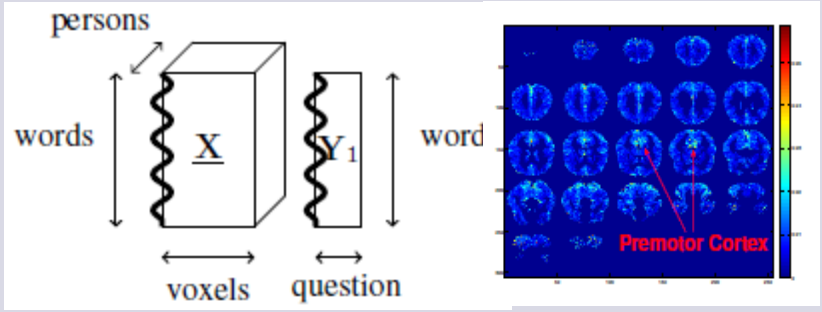
Analyzing (Big) Data Boxes: Multi-way CS of Tensors

Motivating applications

- NELL @ CMU / Tom Mitchell



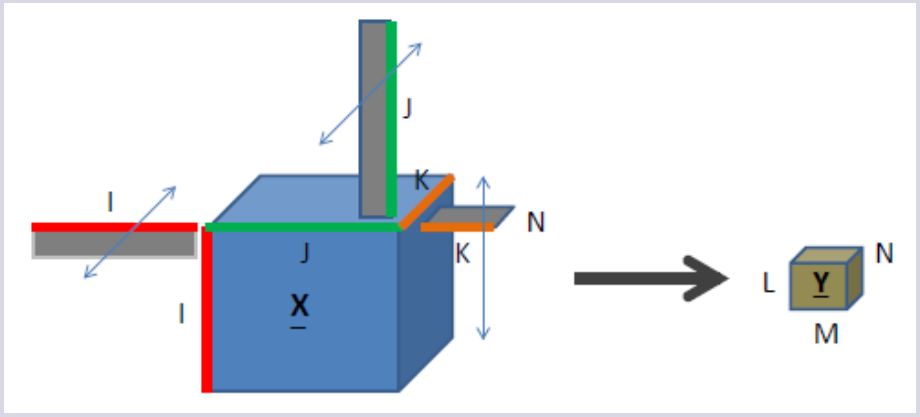
- Semantic analysis of brain fMRI



- Context-aware recommender systems (healthcare)

New analytical & computational tools

- Tensor does not fit in RAM – compress?



- Theorem: Latent factors sparse → can reduce down to LMN=O(F^3/2) and still recover the **big** red-blue-green factors exactly!
- Scalable coupled tensor-matrix computations (Hadoop)
- Reduced-rank tensor filtering

Myria: Foundations and Systems for Big Data Management

Q: Why are relational database so successful?
A: Because of Mathematical foundations + systems design

Myria = UW's Big Data Management project

- Studies both mathematical foundations and systems' design

- Mathematical foundations:
  - Communication = the new complexity parameter:
    How many rounds? How much data replication per round?
  - Results: tight rounds/replication tradeoff for `select-from-where` queries
  - Next steps: rounds/replication tradeoff for skew, iteration, aggregates
- Systems' design:
  - Old stuff = relational model, datalog, shared-nothing architecture
  - New features (model) = multi-way join operators, iteration, UDFs, UDAs
  - New features (system) = cloud service, SLA, multi-tenants, fault tolerance

NSF grant IIS-1247469, http://db.cs.washington.edu/myria/

White House Big Data Partners Workshop, May 3, 2013

Paul F. Uhlir
puhlir@nas.edu and *www.nas.edu/brdi*

Projects of the NRC Board on Research Data & Info, and US CODATA:

- ***Developing Data Attribution and Citation Practices and Standards***
- ***Future Career Opportunities and Educational Requirements for Digital Curation***

- ***BRDI Competition on Management of Research Data***
- ***Legal Interoperability of Research Data***
- ***Building a Biomedicine Knowledge Network***
- ***Strategies for Sharing of Earth Observation Data Between China and the United States***
- ***Sharing of Data on Human Subjects***

- ***Grand Challenges for Scholarly Communication***
- ***Sustainability Strategies for Publicly Funded Scientific Databases***
- ***Decadal Survey of Research Priorities for Data Management***
- ***Hosting 2016 CODATA –WDS Data Summit in Washington, DC***

Analytical Approaches to Massive Data Analysis

Eli Upfal – Brown University

- Big data isn't always the right data
- Data Mining/Machine Learning output isn't always the significant, relevant output **How do we evaluate analytics results?**

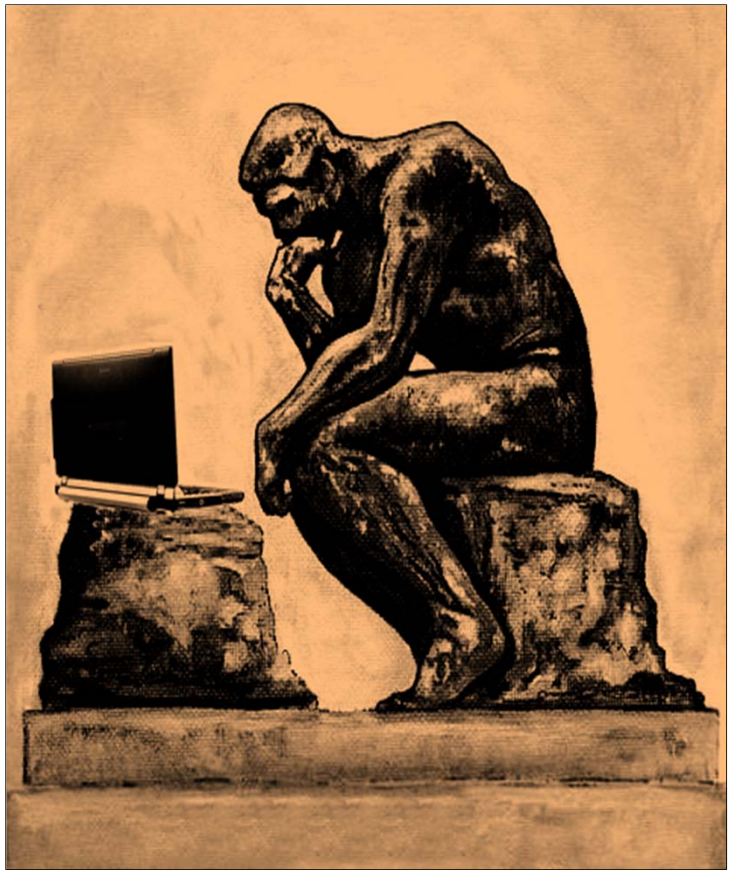
- Statistics vs. Machine Learning – formal statistical inference (p-value, confidence level) vs. computationally practical solutions
- Goal: **Efficient data analysis with rigorous statistical guarantees**

Recent work: Efficient and statistically sound tools for questions like:

- Is it a discovery or noise? Multi-hypothesis testing, FDR, VC-dim..
- Is the hypothesis wrong or the sample too small? Minimum sample size for detecting complex structures

Genomic applications

Integrating Humans, Machines and Networks: A Global Review of Data-to-Decision Technologies



THE NATIONAL ACADEMIES

National Ground Intelligence Center, U.S. Army

A review of current research:

- Cognitive science
- Data analytics
- Decision science
- Machine learning
- Natural language processing
- Neuroscience
- Sensing and perception
- Software agents and multi-agent systems

A multinational review:

- Singapore
- Germany
- Turkey
- Etc.

Big Data ✚ Cyber Security

Justin Zhan - iLAB at North Carolina A&T State University

C:\fromchaostoknowledge.exe

www.dynamicdataanalytics.org

The Center for Dynamic Data Analytics

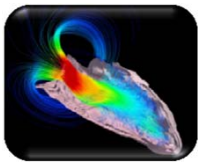
cdda
center for dynamic data analytics

"From Chaos to Knowledge"

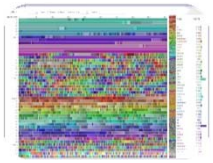
Fact Sheet

- Established 2010 as an NSF I/UCRC between Rutgers and SUNY Stony Brook
- Conducts innovative research with strong industrial applications
- Funded by NSF and Industrial Memberships
- 20+ Active Projects, 30+ Researchers

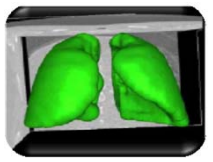
BIOMEDICINE



DATA MINING



VISUALIZATION



National Science Foundation
WHERE DISCOVERIES BEGIN

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Stony Brook University
The State University of New York

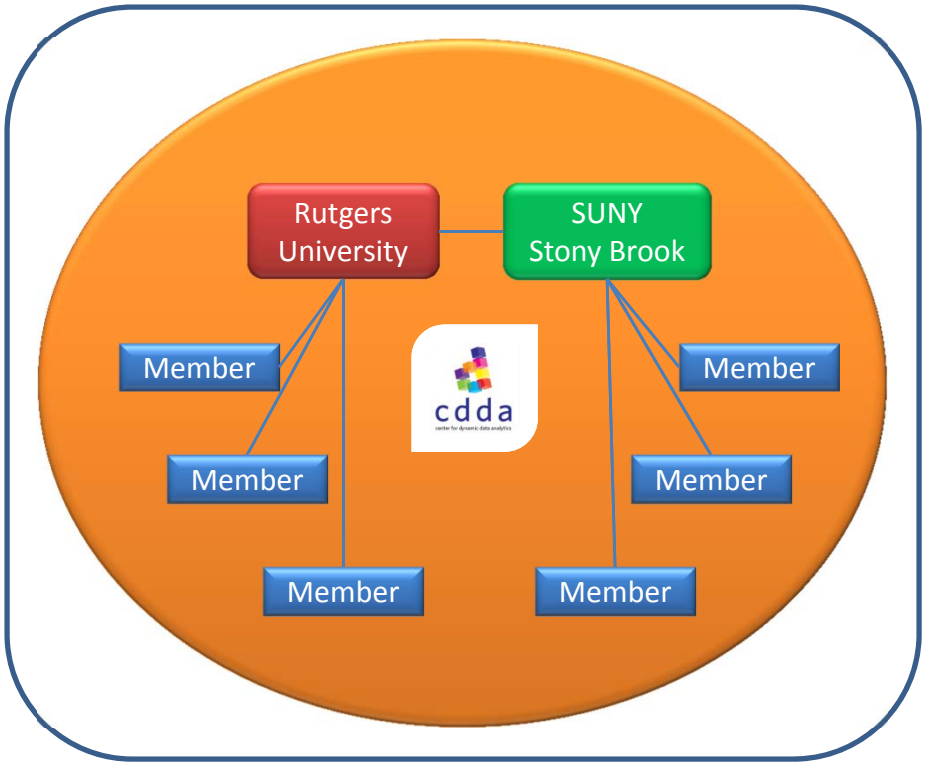
C:\fromchaostoknowledge.exe

www.dynamicdataanalytics.org

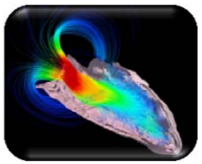
The Center for Dynamic Data Analytics

cdda
center for dynamic data analytics

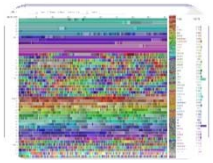
"From Chaos to Knowledge"



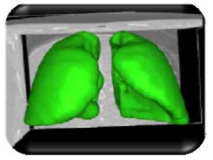
BIOMEDICINE



DATA MINING



VISUALIZATION



National Science Foundation
WHERE DISCOVERIES BEGIN

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Stony Brook University
The State University of New York

www.dynamicdataanalytics.org

The Center for Dynamic Data Analytics

cdda
center for dynamic data analytics

"From Chaos to Knowledge"

Stony Brook University

- Reality Deck – Immersive Gigapixel Display for Big Data Visual Analytics
- Big Data Ingestion Bottleneck Reduction
- Fast and Reliable Information-Theoretic Anomaly Detection and Description
- Knowledge Representation, Acquisition and Querying of Institutional Knowledge
- In Search of Styles in Language: Identifying Deceptive Information and Author Profiles
- Acquiring and Analyzing Imagery for Medical Diagnosis
- Volumetric Rendering and 3D UI for Mobile Health Applications

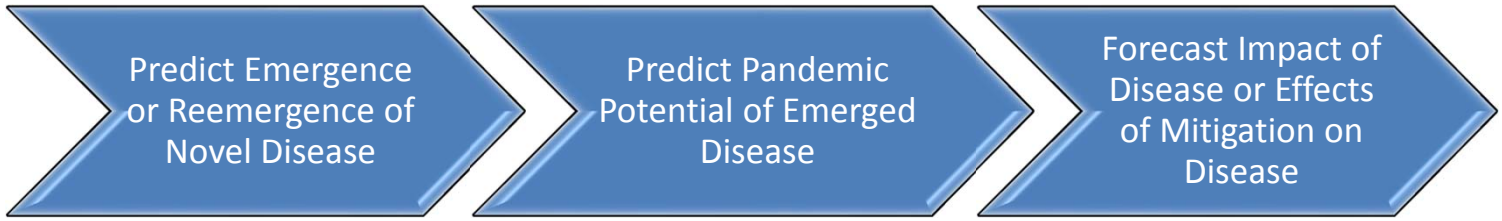
National Science Foundation
WHERE DISCOVERIES BEGIN

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Stony Brook University
The State University of New York

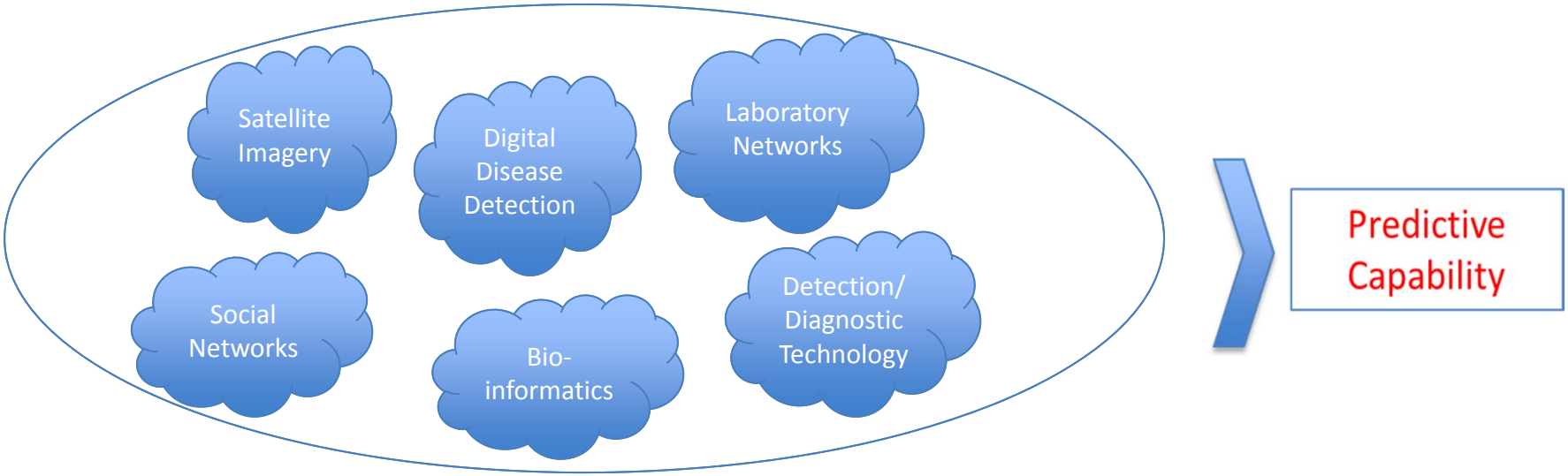
National Weather Service for Infectious Disease

Predicting the Next Pandemic

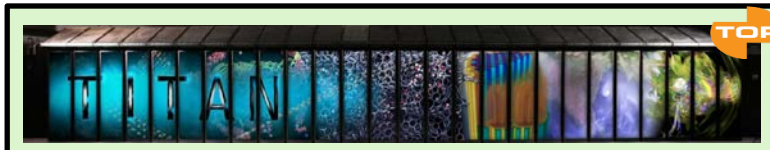


Objectives:

- Create Consortium to enhance multi-sector collaboration and examine how big data can be used to predict pandemics;
- Identify current capabilities, gaps, data requirements, and analytic needs necessary for development of a predictive capability; and
- Develop pilot project to demonstrate feasibility.



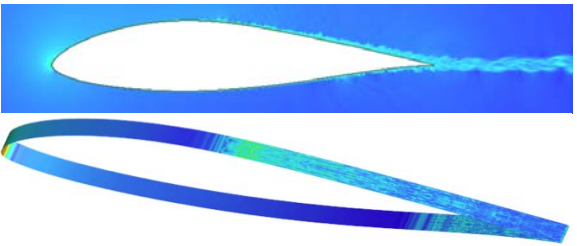
DOE Leadership Computing supports Big Data



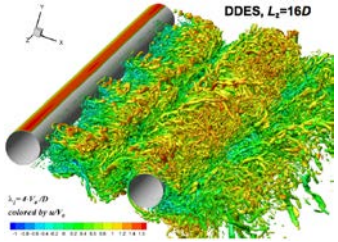
- Peak performance of 27.1 Petaflops
- 18,688 Hybrid Compute Nodes with 16-Core AMD Opteron CPU and NVIDIA Tesla "K20x" GPU and 32 + 6 GB memory
- 200 Cabinets; 710 TB total system memory; 8.9 MW peak power

- Peak performance of 10 Petaflops
- 49,152 Compute Nodes each with 16 GB memory and 6-Core Power PC A2 CPU with 64 Hardware Threads and 16 Quad FPUs
- 56 Cabinets; 786 TB total system memory; 4.8 MW peak power

- Innovative and Novel Computational Impact on Theory and Experiment
  - Leadership Computing for Open Science (ALCF, OLCF)
  - Small number of projects (about 50) and users (about 800) with computationally intensive peer reviewed projects to advance science, speed innovation, and strengthen industrial competitiveness.
  - Also provides Big Compute for Big Data

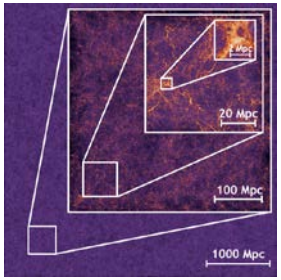


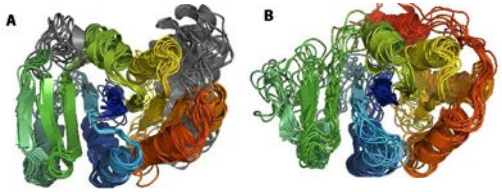
Large-Eddy Simulation of Wind-turbine Airfoil for aerodynamics/ aero acoustic noise reduction. GE captured laminar-to-turbulent transition using 200 million grid points



BOEING Simulation of turbulence created by aircraft landing gear. Calculated the noise caused by two cylinders placed in tandem in an air stream using 60 million grid points with overlapping grids. Contributes to the design of safe and quiet technologies.

Outer Rim Simulation at 2.2 Gyr. The outer image shows the full volume 1.1Trillion particle simulation

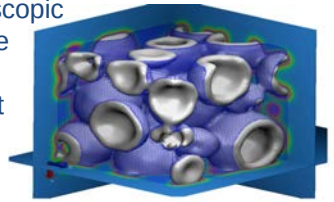




Protein folding Structure of ALG13
(A) computationally (Rosetta)
(B) experimentally (NMR)

Simulation of the complex interactions of billions of atoms to determine how tiny submicroscopic structures impact the characteristics of the ingredients in soaps, detergents, lotions and shampoos to accelerate development of many consumer goods, foods, and fire control materials.

P&G



U.S. DEPARTMENT OF ENERGY | Office of Science

NEW CALL NOW OPEN – www.doeleadershipcomputing.org
Proposals due June 28, 2013