

Big Data, Social Networks, and Analytics

(Big Data Science at ORNL)

Presented at:
University of Memphis Cyber Security
Expo

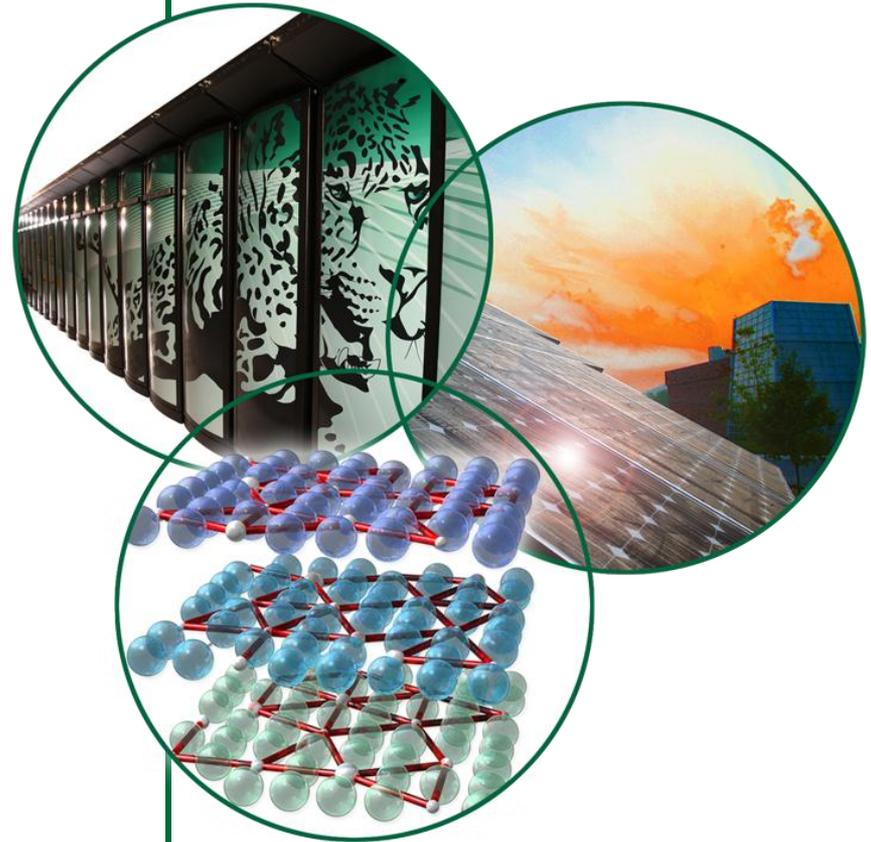
Memphis, TN
presented by Robert K. Abercrombie, Ph.D.

October 19, 2012

Authors:
Bob G. Schlicher and
Robert K. Abercrombie



The submitted manuscript has been authored by a contractor of the U.S. Government under contract DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.



Agenda for Today's Presentation

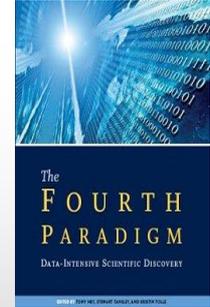
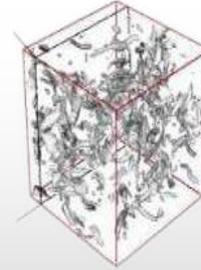
- The Changing Nature Of Research
- Data Explosion
- Business of Big Data
- ORNL Computing and Data Infrastructure
- Accessing and Disseminating Data
- Information Platforms Based on Social Media Features
- Examples of Big Data Projects at ORNL
 - Scientometrics and Analytics Cloud
 - Text Analysis for Analysts
 - Fusion for Prediction of Population Distributions
 - Extreme Scale Visual Analytics for Climate Science
 - Analytics for Biological Data



The Changing Nature Of Research



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



Experiment

**Thousand
Years Ago**

*Description of
natural
phenomena*

Theory

**Last Few
Hundred Years**

*Newton's laws,
Maxwell's
equations...*

Computation

**Last
Few Decades**

*Simulation of
complex
phenomena*

Data

Today and the Future

*Unify theory,
experiment, and
simulation with large
multidisciplinary data*

*Using data exploration
and data mining
(from instruments,
sensors, humans...)*

Distributed Communities

Caution to the Fourth Paradigm

“There are three kinds of lies: lies, damned lies, and statistics.”

“Every generalization is false, including this one.”

- Mark Twain

The thrill of Human Scientific Discover must not be lost on computerized methods and data-intensive scientific research.

Big Data = Volume, Variety and Velocity

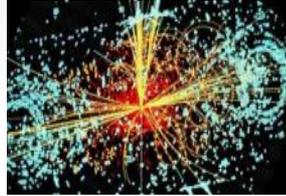


The Data Explosion

Experiments



Simulations



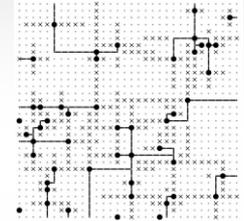
Archives



Social Media



Sensors



Information Technology

The Challenge
Enable Discovery

Deliver the capability to mine, search, and analyze this data in near real time

Petabytes
Exabytes
Zettabytes

The Response

Science itself is evolving

Volumes and Rates

- **Published Papers/Patents at ORNL** 7 TB
- **Library of Congress Text** 20 TB
- **Amazon** 42 TB
- **ChoicePoint** 250 TB
- **ORNL Scientometrics Cloud** 260 TB
- **AT&T** 323 TB
- **US Government** 848 TB (`09)
- **US Discrete Manufacturing Companies** 966 TB (`09)

Archives



Volumes and Rates

- **Twitter Updates** **400 M/d**
- **Facebook**
 - Likes/Comments **2.7 B/d**
 - Shared Contents **30 B/m**
- **World Emails** **419 B/d**
- **YouTube**
 - Storage **76 PB/yr.**
 - Traffic **16.2 EB/yr.**
- **World Social Media** **1.8 ZB (x2 every 2 yrs.)**

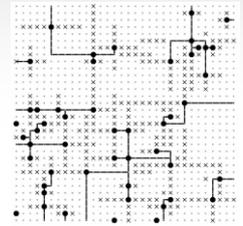
Social Media



Volumes and Rates

- **2.5 m Telescope** **200 GB/d**
- **Ion Mobility Spectroscopy** **10 TB/d**
- **X-ray Photon Correlation Spectroscopy**
3D X-ray Diffraction Microscopy **24 TB/d**
- **Boeing 737 cross-country flight** **240 TB**
- **Personal Location Data** **1 PB/yr.**
- **Astrophysics Data** **10 PB (2014)**
- **Square Kilometer Array** **480 PB/d**

Sensors



The Business of Big Data

- **\$300 billion annual value of big data for the U.S. health care system, two-thirds of which would come in reduced expenditures (McKinsey).**
- **\$165 billion worth of value for big clinical data (McKinsey).**
- **966 petabytes data stored by discrete manufacturing companies in the U.S. during 2009; 848 petabytes of data stored by government in the same year (McKinsey).**
- **By 2020, IT departments will have 10 times more servers and 50 times more data to look after than they do now.**

The Business of Big Data (cont.)

- The U.S. will face shortages of:
 - between 140,000 and 190,000 individuals with “deep analytical skills” capable of working with very large data sets;
 - between 300,000 and 400,000 skilled technicians and support staff;
 - about 1.5 million “data-savvy“ managers and analysts. (McKinsey)

"big data" Job Trends



Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows relative growth for jobs we find matching your search terms.

[Find "big Data" jobs](#)

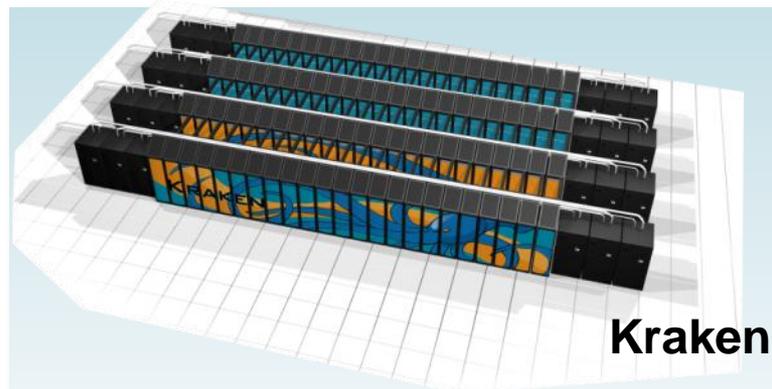
ORNL Computing and Data Infrastructure

Today, we have one of the world's most powerful computing facilities



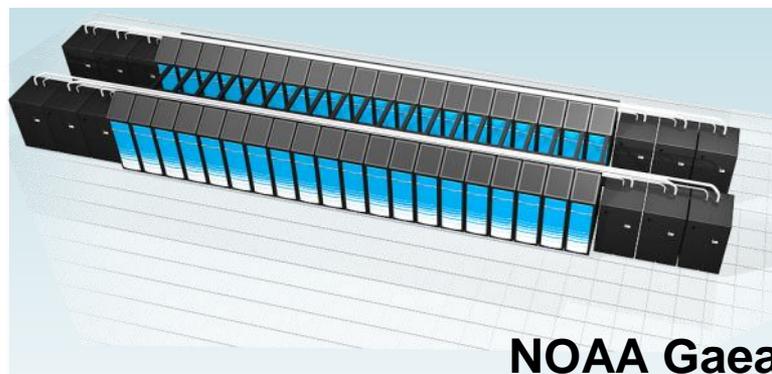
Jaguar

Peak performance	3.3 PF/s
Memory	600 TB
Disk bandwidth	> 240 GB/s
Square feet	5,000
Power	5.2 MW



Kraken

Peak performance	1.17 PF/s
Memory	147 TB
Disk bandwidth	> 50 GB/s
Square feet	2,300
Power	3.5 MW



NOAA Gaea

Peak Performance	1.1 PF/s
Memory	240 TB
Disk Bandwidth	104 GB/s
Square feet	1,600
Power	2.2 MW



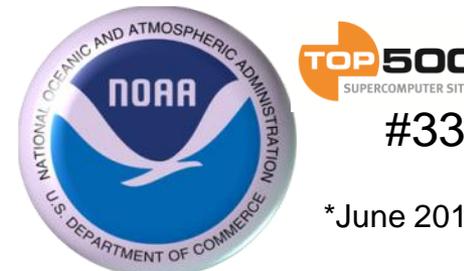
#6*

Dept. of Energy's most powerful computer



#21*

National Science Foundation's most powerful computer



#33*

*June 2012

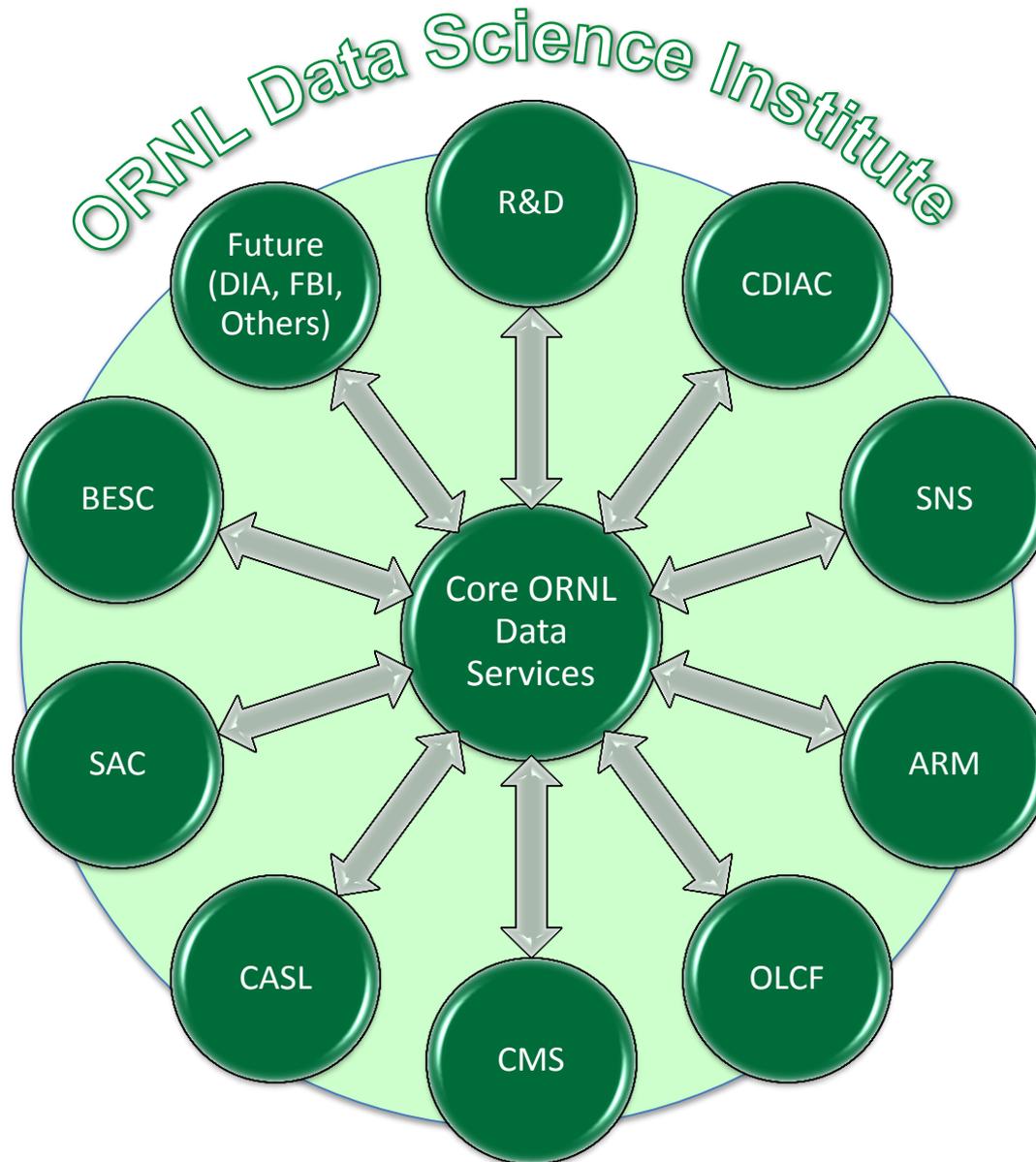
National Oceanic and Atmospheric Administration's most powerful computer



Data Infrastructures at ORNL (representative)

- National Geospatial Agency (NGA) storage (LandScan™)
 - http://www.ornl.gov/sci/landscan/landscan_data_avail.shtml
- Oak Ridge Leadership Computing Facility (OLCF) storage
 - <http://www.olcf.ornl.gov/>
- Spallation Neutron Source (SNS) infrastructure –
 - Accelerating Data Acquisition, Reduction, and Analysis (ADARA)
 - <http://neutrons.ornl.gov/facilities/SNS/>
 - Guinness World Record - Most Powerful pulsed Neutron Source
 - <http://www.guinnessworldrecords.com/records-5000/most-powerful-pulsed-spallation-neutron-source/>
- Scientometrics cluster
 - Intelligence Advanced Research Projects Activity (IARPA)
 - <http://www.iarpa.gov/Programs/ia/FUSE/fuse.html>

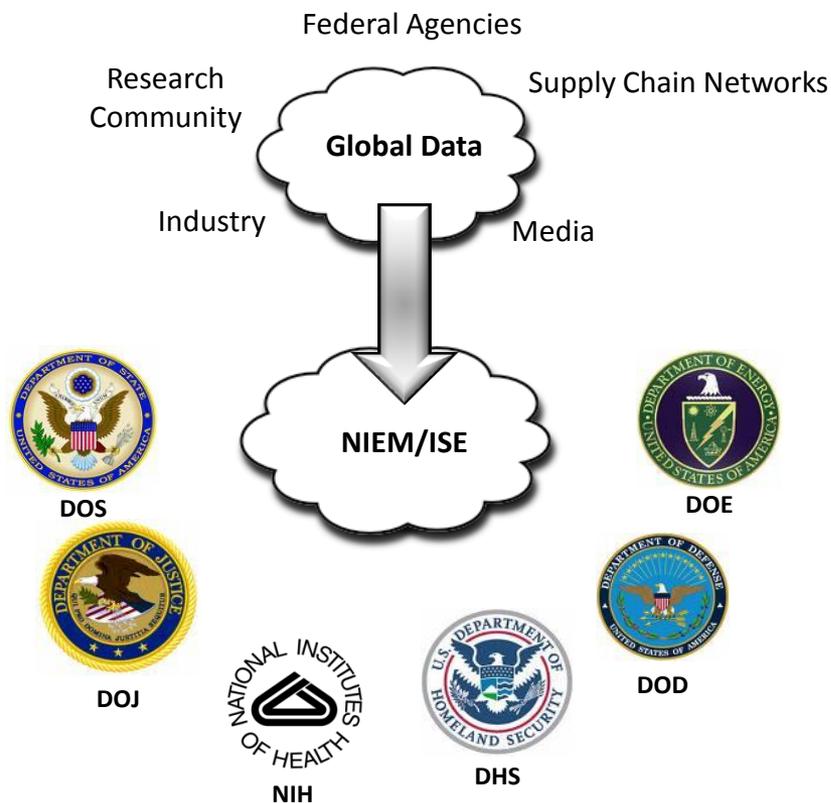
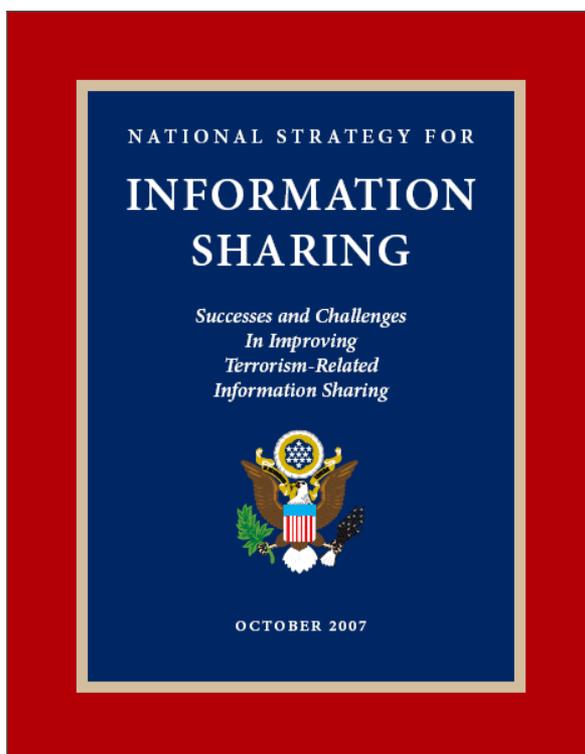
Providing Expertise in Data For Mission



Accessing and Disseminating Data

Data Sharing Challenges

- Carry out successfully a discovery process across data held by industry, academia, and government agencies to address critical national missions.



Information Platforms Based on Social Media Features

- Reference by URL
- Contributed content
- Reputation system
- Tagging + search
- User-defined mash-ups
- Publish-subscribe



VERDE Electric Grid Status
(Real-Time Grid Awareness)



Tracking 2.0
(Cradle-to-grave tracking)



Sensorpedia
(The “Wikipedia of Sensors”)



Knowledge Discovery Framework
(National Biomass Distributions)

Open Government Initiative



- Transparency promotes accountability
- Participation allows people to contribute ideas
- Collaboration encourages cooperation within government and with industry



Data.gov Communities

An Official Web Site of the United States Government | Tuesday, January 18, 2011 | Text: A- A+ A | Share

DATA.GOV / HEALTH | Login | Sign Up

WELCOME TO THE HEALTH DATA COMMUNITY
You've found a public resource designed to bring together high-value datasets, tools, and applications using data about health and health care to support your need or better knowledge and to help you to solve problems. These datasets and tools have been gathered from agencies across the Federal government with the goal of improving health for all Americans. Check back frequently because the site will be updated as more datasets and tools become available.
[More Information](#)

HEALTH INDICATORS WAREHOUSE
[VIEW MORE](#)

Home | **Develop** | Data/Tools | Apps Expo | Other Data Sites | Blogs | Forums | Search This Community

Data.gov » All Communities

Challenges

2010 Health 2.0 DEVELOPER CHALLENGE

Pellentesque libero mauris, mollis at sollicitudin ut, mattis in lorem. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Sed lobortis congue elementum. Pellentesque elit tortor, feugiat eu iaculis et, portitor at turpis Vivamus vehicula nibh vel massa elementum blandit.

What's New

Learn about our newest high-value datasets.

- [ClinicalTrials.gov: linking patients to medical research](#)
- [Genetics home reference](#)
- [MedlinePlus Health Topic Web Service](#)
- [PubChem](#)

Recent Blog Post

Test blob 17jan 15:16 V4
Posted on: 1/17/2011
Test blob 17jan 15:16 V4

Community

Restore the Gulf ✓

Open Data ✓

Semantic Web ✓

Health ✓

Law ✓

Energy ✓

Education

Ocean

Research and Development

Public Safety

Human rights

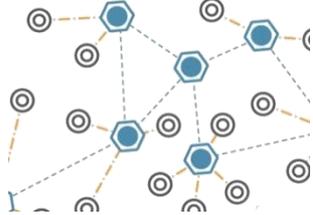
+ *many more...*



Architecture Concept for Big Data

Sources

Sensor Networks



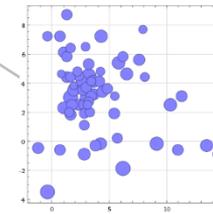
Architectures

Tagging



Tools

Analytics



Distributed Global Databases



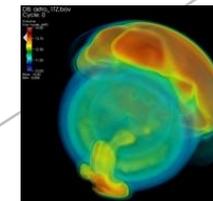
Faceted Classification Systems

- Classify by TYPE**
 - Type A
 - Type B
 - Type C
- Classify by ZONE**
 - Zone A
 - Zone B
 - Zone C
- Classify by TIME**
 - Time A
 - Time B
 - Time C

Facet

Value

Visualization



End User



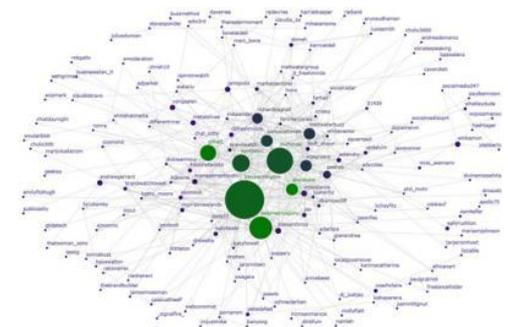
Example Big Data Projects at ORNL

Scientometrics Analytics Cloud: Mission and Vision

Vision: Provide world-class analytics for 'Big Data' problems to enrich policy decision-makers in Intelligence, Defense for global security, Transportation, and finance.

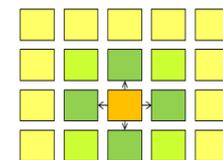


Mission: Design and deliver quality solutions for 'Big Data' problems that address ever-increasing complexity and data scale by applying established, state-of-the-art, and cutting-edge data analytics.



Analytics Cloud Today: hosting 100+ Million raw documents of published scientific papers (EiSevier, Web of Science) and patents

- 200+ commercial (private) and academic users
- Seven core Teams investigating “Technology Predictive Emergence”
- 192 cores, 360 Terabytes disk
- Moving to tiered, high-speed data I/O of HPC measures



Big Data Text Analysis for Analysts

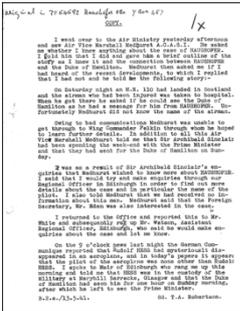
10,000 raw intelligence documents per day



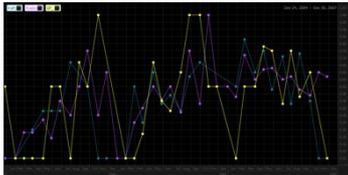
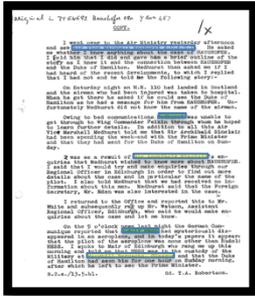
1,000 Unread documents



100 Read documents

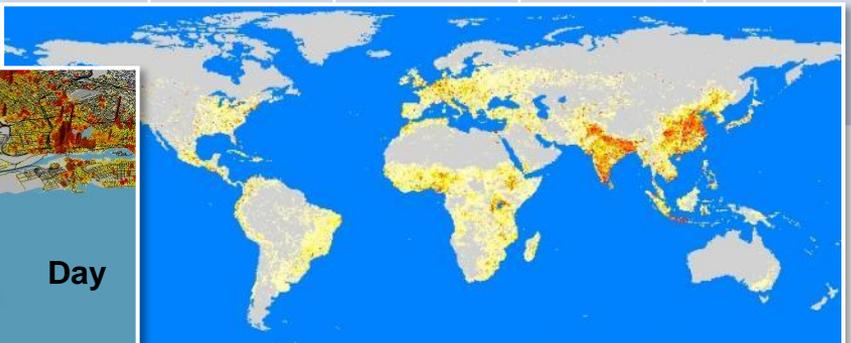
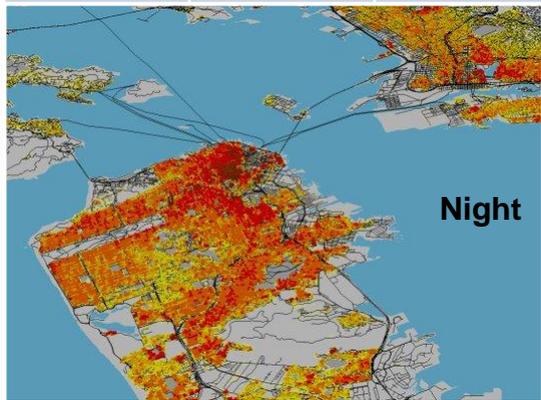


20 Tagged documents



Big Data Fusion for Prediction of Population Distributions

Population	Road	Railroads	Land cover/land use	Slope	Academic institutions	Prisons	Hospitals	Business employment	Imagery
<ul style="list-style-type: none"> Census Polygons Tract-to-tract worker flow BLS quarterly updates 	<ul style="list-style-type: none"> VMAP TeleAtlas Multinet TIGER; 	<ul style="list-style-type: none"> 1:100K national railway network NTAD 	<ul style="list-style-type: none"> Geocover MODIS National Land Cover Data (NLCD) State GIS 	<ul style="list-style-type: none"> DTED LiDAR National Elevation Data (NED) 	<ul style="list-style-type: none"> Department of Education HSIP Schools ESRI GDT 	<ul style="list-style-type: none"> National Jail Census HSIP Prisons 	<ul style="list-style-type: none"> American Hospital Association (AHA) 	<ul style="list-style-type: none"> InfoUSA Pitney Bowes Dunn and Bradstreet 	<ul style="list-style-type: none"> Earth-Viewer Terra Server Google



LandScan Global

- Spatial resolution of 30 arc seconds (~1km)
- Ambient population (average of 24 hours)
- Remote sensing based global data modeling and mapping



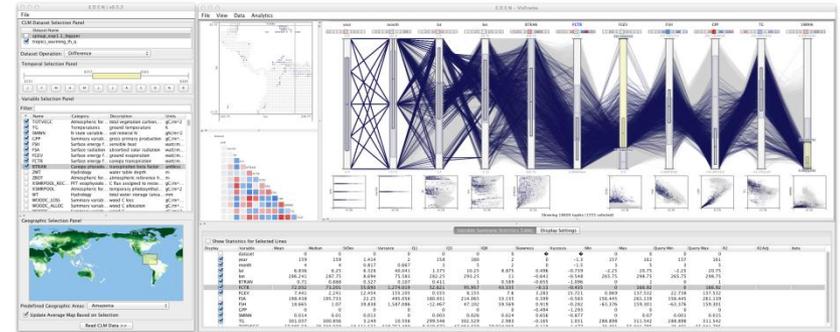
LandScan USA

- Spatial resolution of 3 arc seconds (~90m)
- Nighttime and daytime population
- Integration of infrastructure and activity databases



Extreme Scale Visual Analytics for Climate Science

- Interactive visual analysis of global model ensembles at extreme scales
 - 100s TBs, 300+ variables, multi-decadal
- Highlights significant associations to effectively guide the scientist to insight
- Data summarization for hyper-dimensional datasets via an intelligent user interface.
- Online linkage to HPC platforms (Jaguar) for statistical analytics.
- Reduces knowledge discovery timelines by several orders of magnitude.
- Funded by DOE Office of Science BER



Environmental Data analysis ENvironment (EDEN)

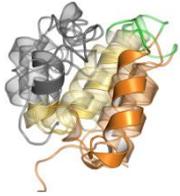


1000 simulations, 81 parameters, 7 output variables
11,520 x 3072 (35 million) pixels

Steed et al., "Practical Application of Parallel Coordinates for Climate Model Analysis". In *Proceedings of the International Conference on Computational Science*, pp. 877-886, 2012.

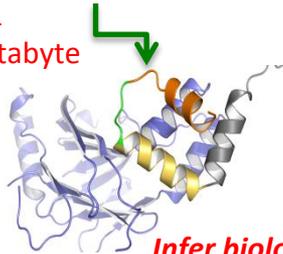
Big-Data Analytics for Biology

Biological Data

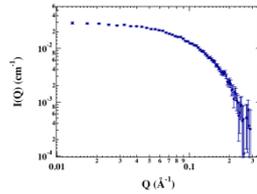


High resolution
all-atom
simulations
(Jaguar/Titan)

> 1
petabyte

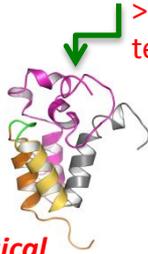


Infer biological
function?

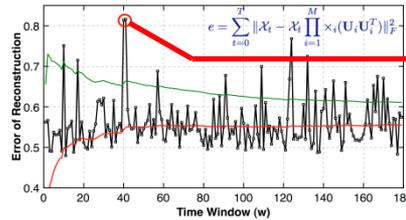


Sparse
experimental
read-outs
(SNS)

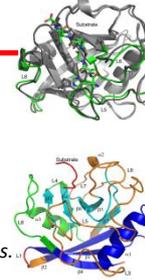
> 100
terabytes



Online data analytics for anomaly/event detection

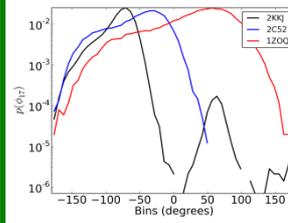


Ramanathan, A. et al, *Proteins* (2012), in press.
Ramanathan, A., et al, *J. Comp. Bio.* (2009)

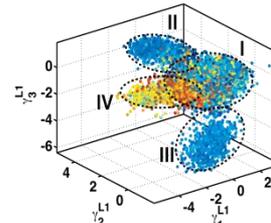


- Enable analysis of high volume and velocity data
- Tensor representation (multi-way dependencies)
- Error of reconstruction (e) identifies anomalies
- Dynamically cluster regions that exhibit similar spatio-temporal patterns
- **Applications:** social media tracking for bio-surveillance; control molecular simulations

Exploiting long-tailed behavior for feature extraction



Ramanathan, A. et al, *PLoS One* (2011), e15827
Savol, A.J. et al, *Bioinformatics* (2011): 27 (13), i52-i60
Burger, V. et al, *Pacific Symposium on Biocomputing* (2012)

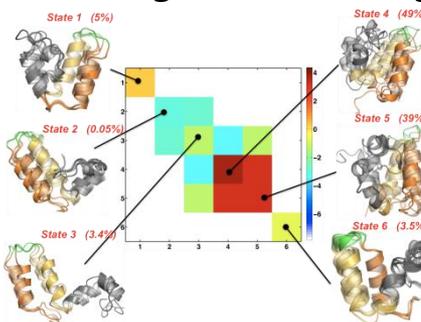


- Unsupervised learning for discovering statistical structure of high-dimensional data
- Higher-order statistics to extract biologically relevant features in data
- **Applications:** enabling drug-discovery using feature spaces learned from atomistic simulations; perceptual dimensions of odor recognition; tumor detection in mammograms

Challenges:

- Insights from high dimensional and noisy datasets
- Integrate sparse experimental observations to reconstruct high resolution details
- Common across genomics, imaging, cheminformatics

Clustering and visual organization of high-dimensional data



- Visualization of high-dimensional data in a meaningful low-dimensional space to ease interpretation
- **Applications:** Interpreting the conformational landscape of proteins; High-throughput virtual screening of drugs; enabling disease diagnostics using high resolution video microscopy

Contact: Arvind Ramanathan, ramanathana@ornl.gov

Oak Ridge National Laboratory: Meeting the challenges of the 21st century



Bob G. Schlicher, Robert K. Abercrombie, Ph.D.
Email: schlicherbg@ornl.gov, abercrombier@ornl.gov
Phone: (865) 574-4988, (865)241-6537
<http://www.ornl.gov/~abe>