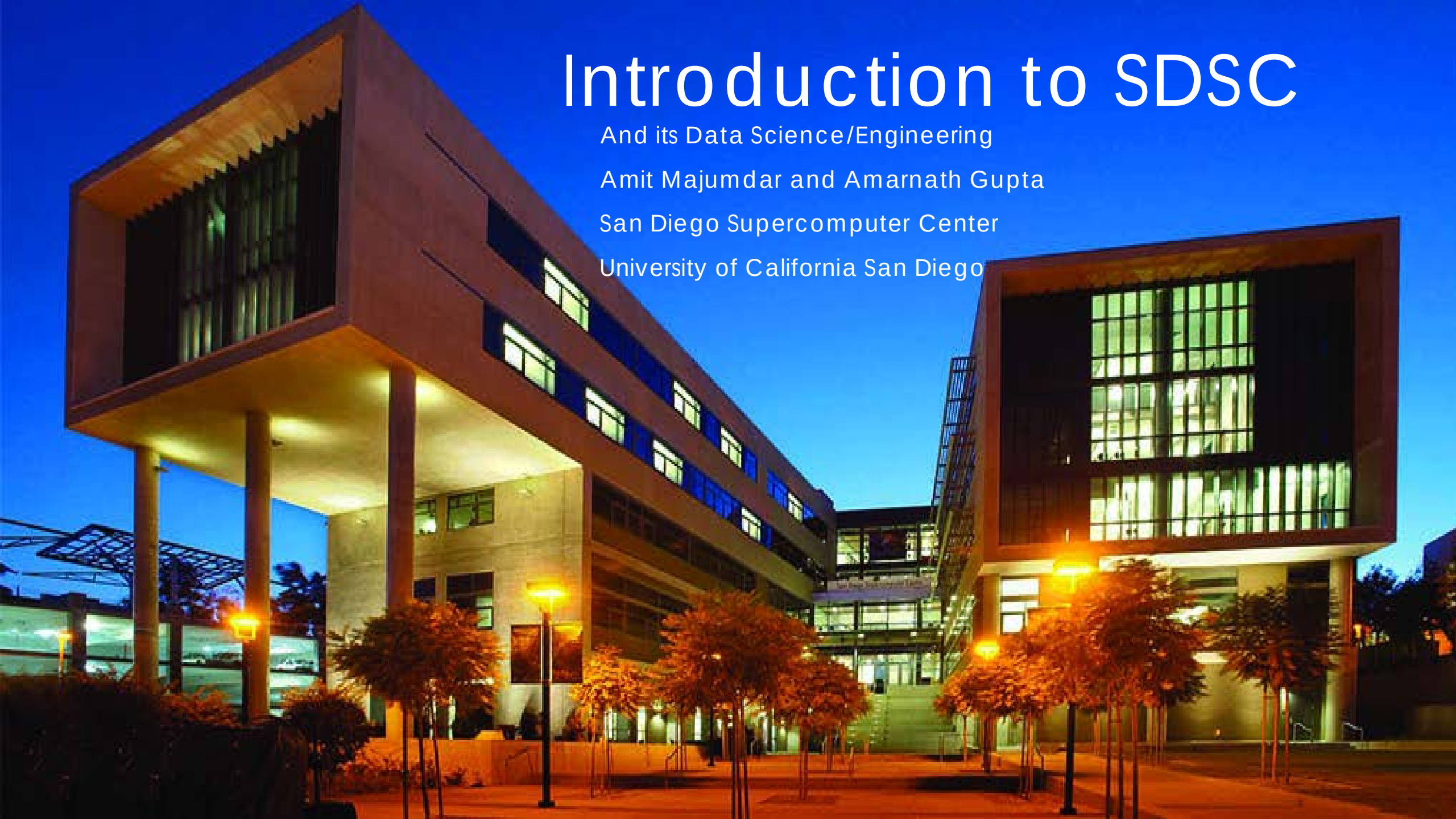
One day Introductory Workshop on Big Data Methods (on HPC)

- **MORNING**
- Introduction of San Diego Supercomputer Center - Data Science (and HPC)
- Introduction to SDSC HPC Systems - Comet and Gordon hardware (including simple hands-on job submission)
- Introduction to Python
- Big Picture of Data Science Approaches and Tools
- Description of Hadoop (including hands on with Hadoop)
- **AFTERNOON**
- Introductory Spark for Scientific Computing (including hands on with Spark)
- Introduction of MongoDB with Pymongo
- Introduction of neo4j **
- Brief Discussion of Advanced Spark for Scientific Computing
- ** remote presentation



Introduction to SDSC

And its Data Science/Engineering

Amit Majumdar and Amarnath Gupta

San Diego Supercomputer Center

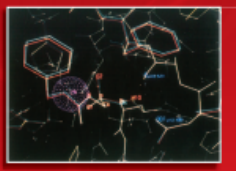
University of California San Diego

SAN DIEGO SUPERCOMPUTER CENTER at UC SAN DIEGO

SDSC

30 years

of Turning Data to Discovery



In pioneering efforts in drug design, Paul Bash, et. al., using SDSC supercomputers, determine free energies of solvation for proteins and nucleic acids, and relative free energies for binding, published in *Science*.

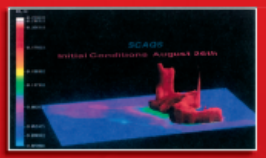
1987

1989

Atmospheric Carbon Dioxide from Fossil Fuels

Beginning in the late 1950s, Charles Keeling from UC San Diego's Scripps Institution of Oceanography (SIO) continuously collected data on the distribution of carbon dioxide around the globe. In 1989, Keeling—with SIO colleagues Stephen Piper and Robert Bacastow, using SDSC resources—constructed a three-dimensional computer model of the terrestrial carbon cycle that took advantage of the data collected by Keeling. The model was the first to confirm the importance of fossil fuel combustion in loading

1991



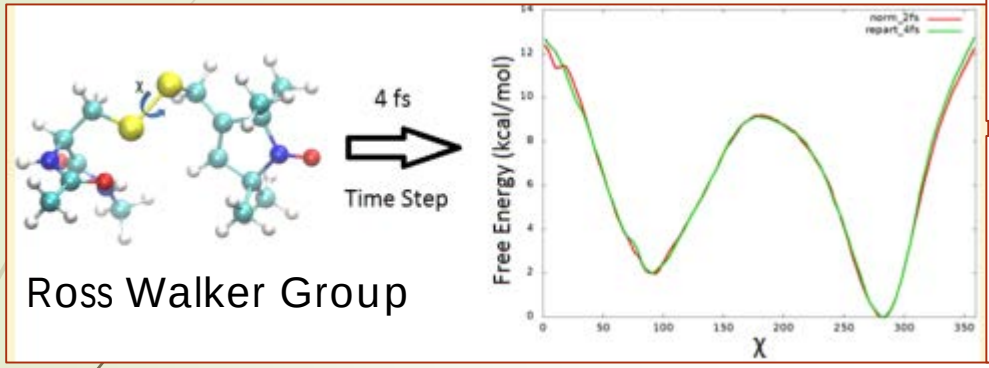
UC receives a 3-year, $15 million grant from DEC to develop an advanced information and data management system to increase the productivity of researchers studying global change – called Project Sequoia.

1998

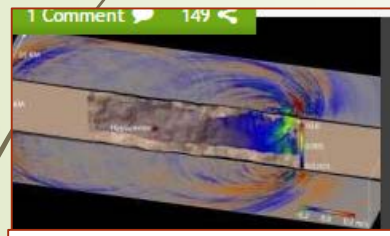
With large-scale computer simulations run at SDSC, researchers led by J. Andrew McCammon at UCSD show how one of the fastest enzymes in the world, acetylcholinesterase, does its work; results are published in the *Proceedings of the*

Science and Computing at SDSC

Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning



Ross Walker Group



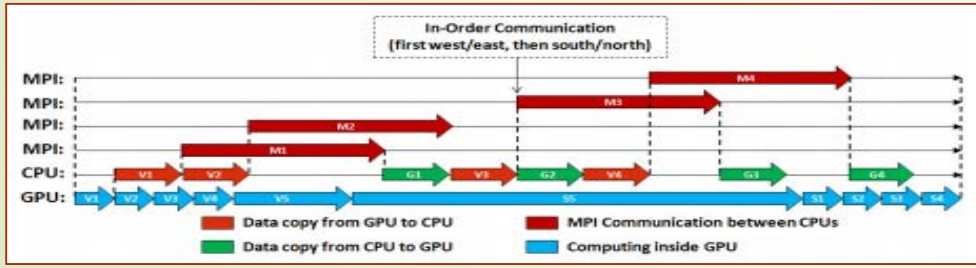
CORPORATE RESPONSIBILITY

San Diego Supercomputer Center's Quake Research Wins $150,000 Global Impact Award

By Tonie Hansen on March 16, 2015

Multi-GPU Implementation of a 3D Finite Difference Time Domain Earthquake Code on Heterogeneous Supercomputers

Jun Zhou[a,b,*], Yifeng Cui[a], Efecan Poyraz [a,b], Dong Ju Choi [a], Clark C. Guest [b]



Fast construction of nanosecond level snapshots of financial markets

Jiading Gai[1], Dong Ju Choi[2], David O'Neal[3], Mao Ye[1] and Robert S. Sinkovits[2,*,†]

Symbol	Wall time (s) original code	Wall time (s) modified code	Speedup
SWN	8400	128	66x
AMZN	55,200	437	126x
AAPL	129,914	1145	113x

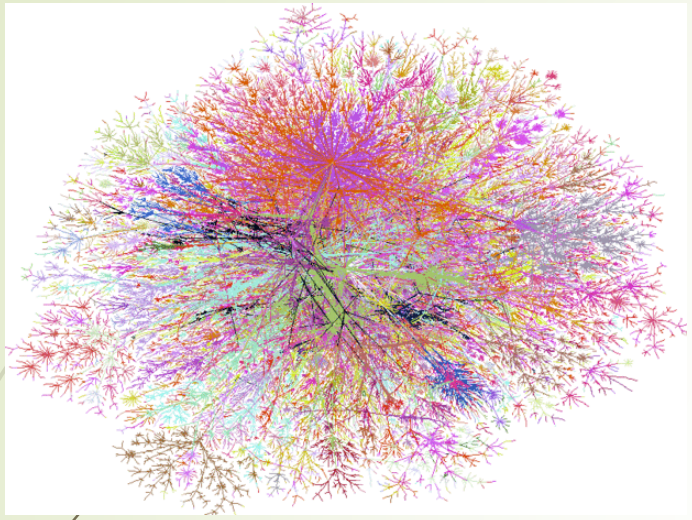
Group-based variant calling leveraging next-generation supercomputing for large-scale whole-genome sequencing studies

Kristopher A. Standish[1,2], Tristan M. Carland[2], Glenn K. Lockwood[3], Wayne Pfeiffer[3], Mahidhar Tatineni[3], C Chris Huang[4], Sarah Lamberth[4], Yauheniya Cherkas[4], Carrie

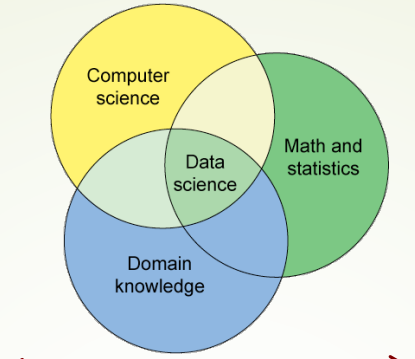
Step	Tool	Memory per command (GB)	Cores per command	Commands per node
Map	BWA	32	8	2
Bam	Samtools	4	1	16
Merge	Samtools	4	1	16
Sort	Samtools	4	1	16
MarkDuplicates	PicardTools	7	2	8
TargetCreator	GATK	7	2	8
IndelRealigner*	GATK	12	3	5
BaseRecalibrator	GATK	30	8	2
PrintReads*	GATK	30	8	2
HaplotypeCaller	GATK	60	16	1

*Smaller memory allocation and more samples per node may prove more computationally efficient

Definition of Data Science is Evolving



BIG DATA



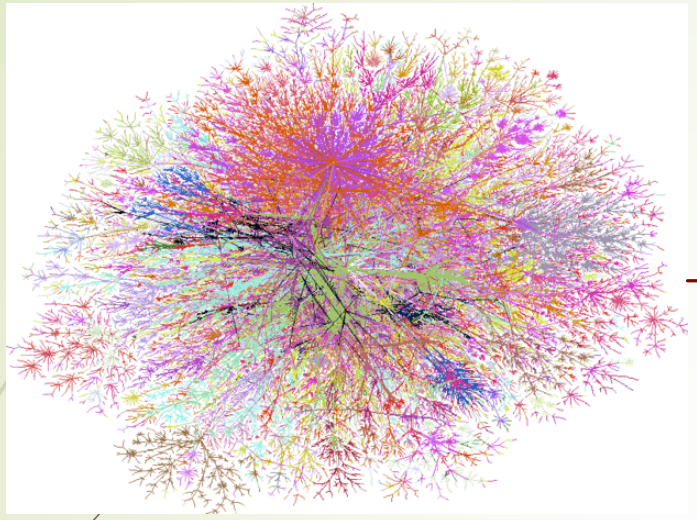
ON-DEMAND COMPUTE

Allows for data-enabled decision making at scale

Many current and future applications with dynamic and measurable impact!

APPLICATION-SPECIFIC KNOWLEDGE and QUESTIONS

?

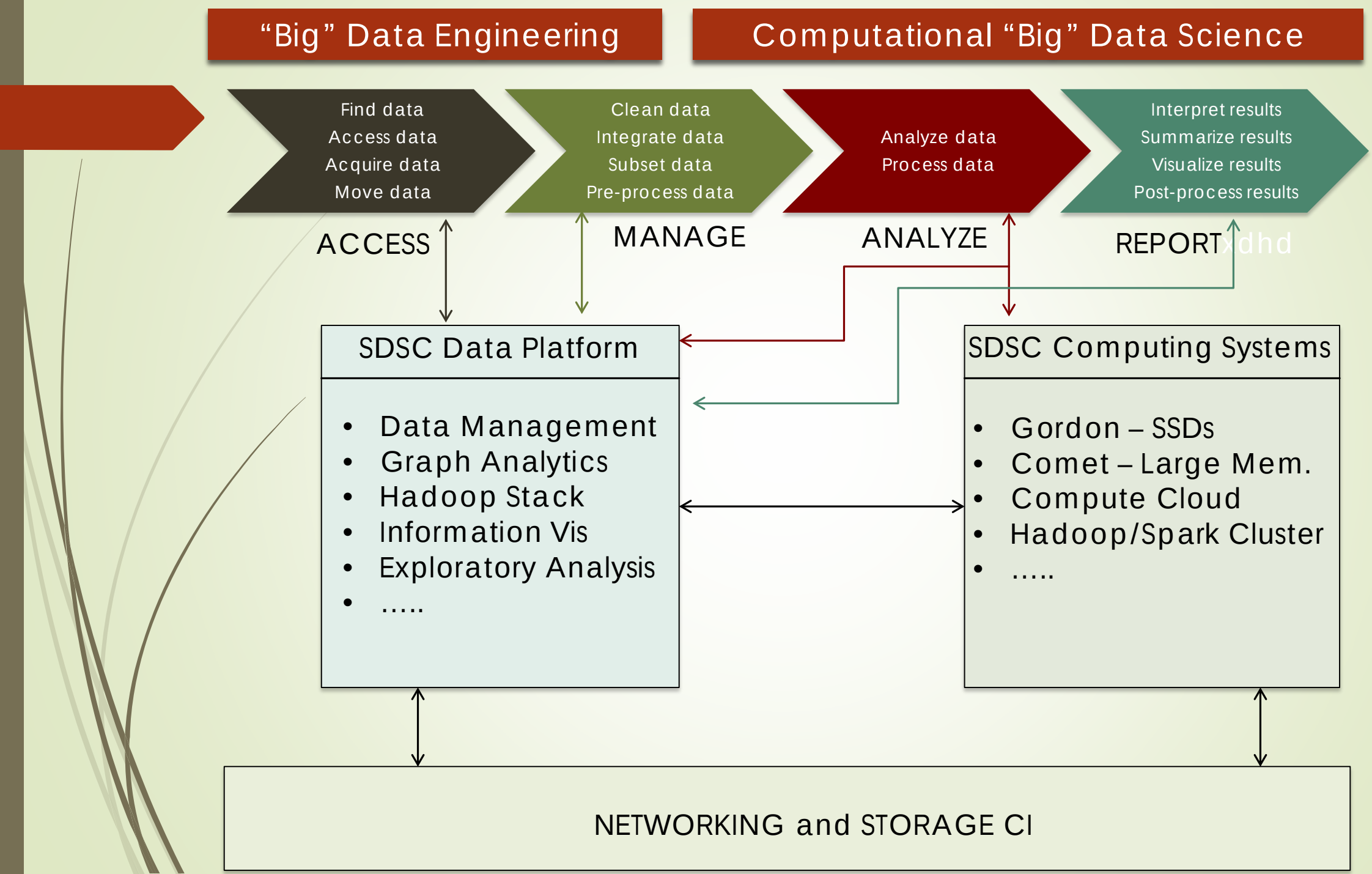


BIG DATA

ON-DEMAND COMPUTE

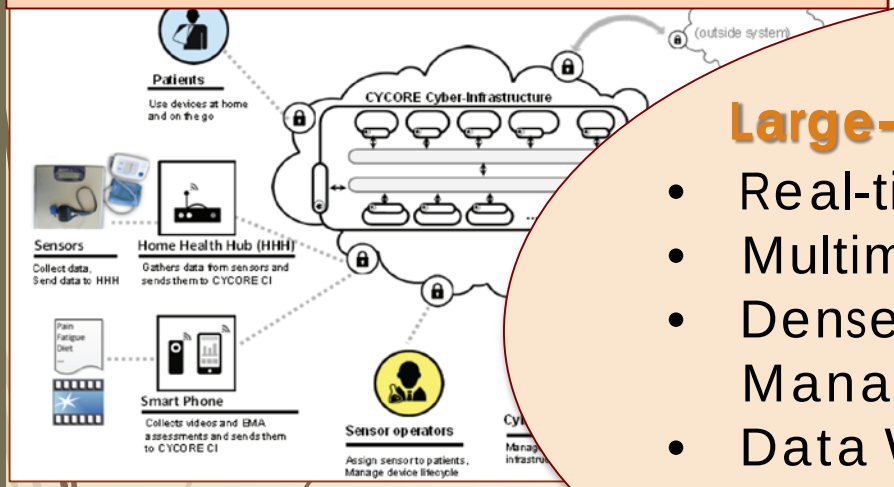
Allows for data-enabled decision making at scale, using data management, statistics, data mining, graph analytics, etc.

Requires support for experimental work by a multidisciplinary group of experts and dynamic scalability on many platforms!


"Big" Data Engineering
Computational "Big" Data Science
Find data
Access data
Acquire data
Move data
Clean data
Integrate data
Subset data
Pre-process data
Analyze data
Process data
Interpret results
Summarize results
Visualize results
Post-process results
ACCESS
MANAGE
ANALYZE
REPORT
SDSC Data Platform
• Data Management
• Graph Analytics
• Hadoop Stack
• Information Vis
• Exploratory Analysis
• .....
SDSC Computing Systems
• Gordon – SSDs
• Comet – Large Mem.
• Compute Cloud
• Hadoop/Spark Cluster
• .....
NETWORKING and STORAGE CI

CLDS
Center for Large-scale Data Systems Research

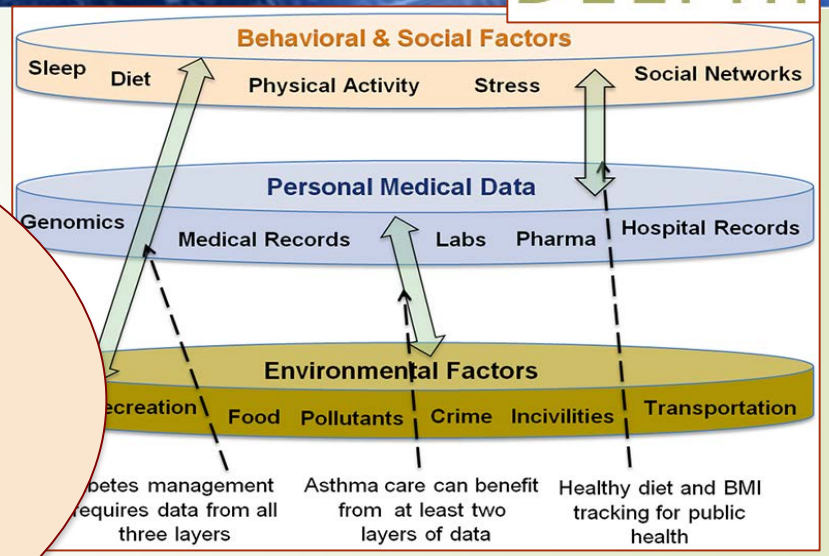
Sensing Health for Treatment Effectiveness

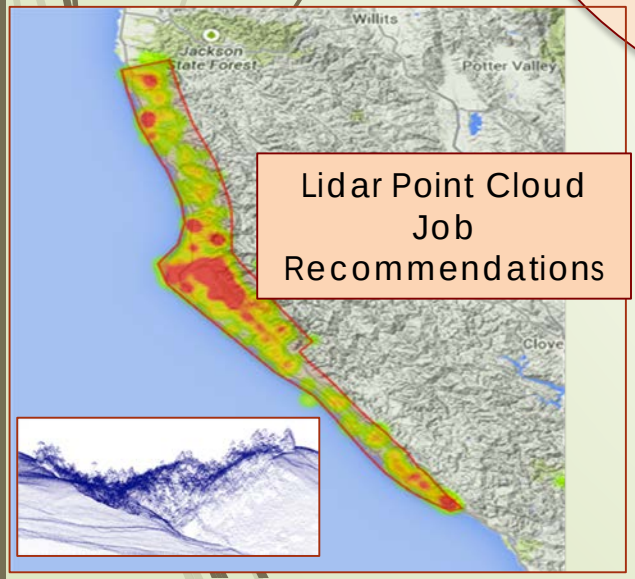


Large-Scale Data Infrastructure

- Real-time Data Management
- Multimedia Data Acquisition
- Dense Data (e.g., LIDAR) Management with DBMS
- Data Warehousing
- Systems Integration

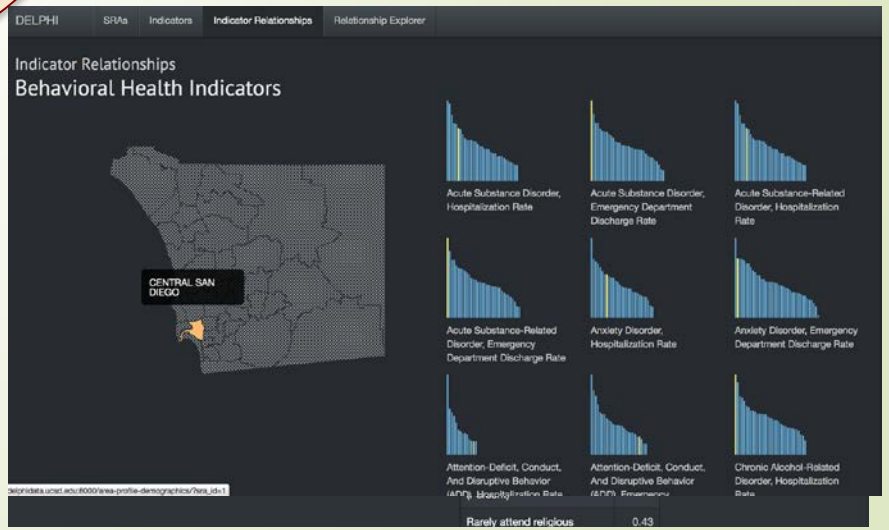
DELPHI





Lidar Point Cloud Job Recommendations

Monitoring Wild Life with Videos

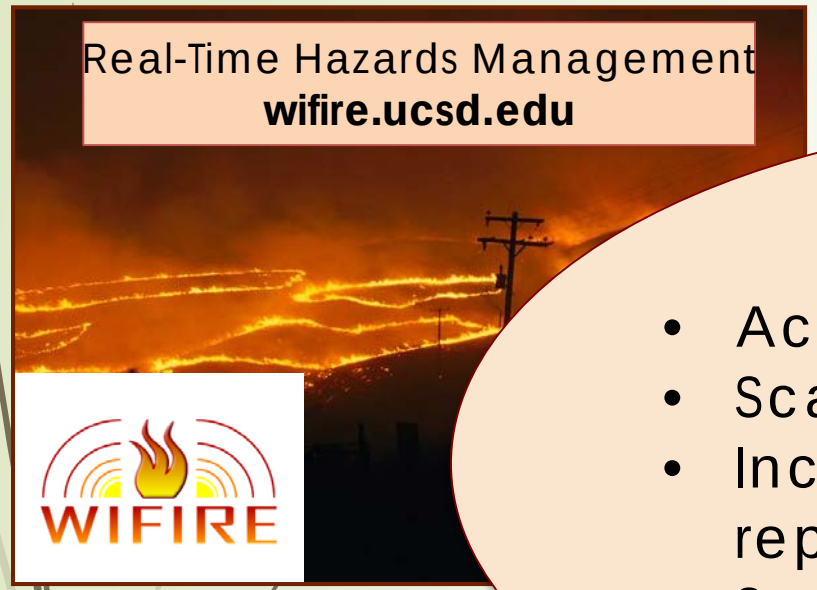


Workflows for Data Science (WorDS) Center at SDSC

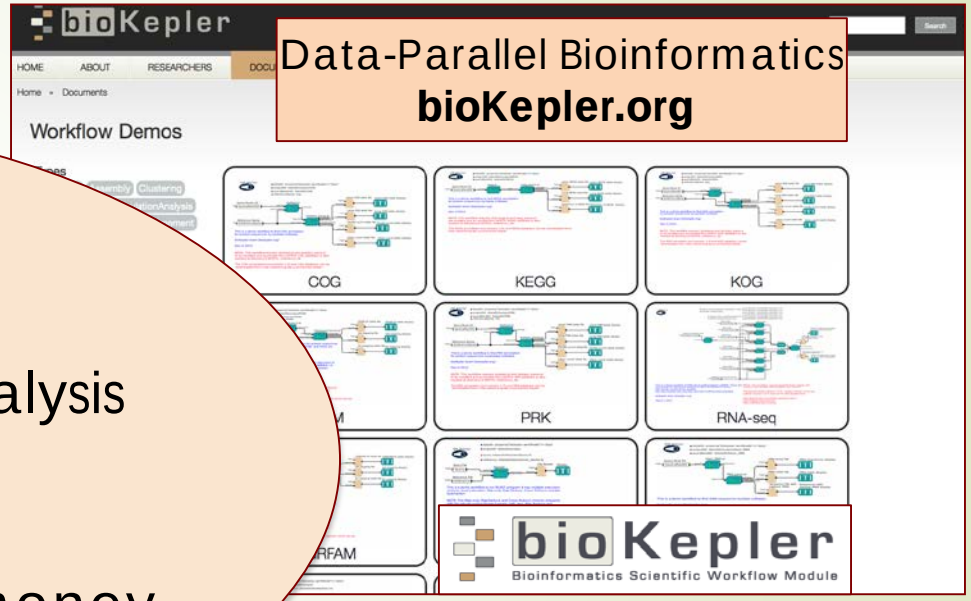
- Programmable, Reusable and Reproducible Scalability -

WorDS.sdsc.edu

Real-Time Hazards Management
wifire.ucsd.edu



Data-Parallel Bioinformatics
bioKepler.org

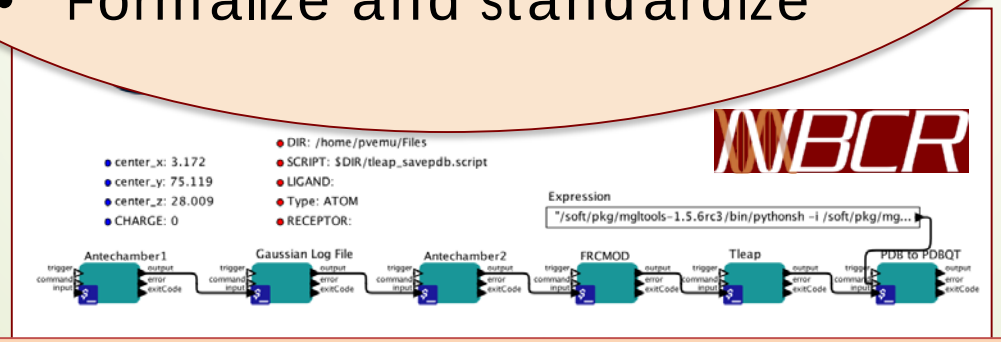


WorDS Center

- Access and query data
- Scale computational analysis
- Increase reuse and reproducibility
- Save time, energy and money
- Formalize and standardize

kepler-project.org

Kepler



Scalable Automated Molecular Dynamics and Drug Discovery
nbcr.ucsd.edu

Big Data

WIFIRE

NSF

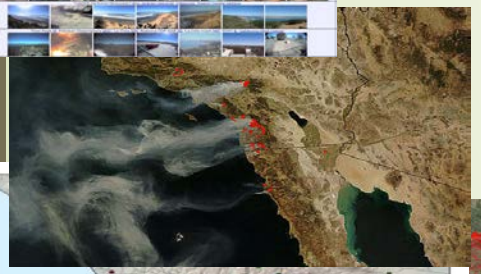
Monitoring Visualization Fire Modeling

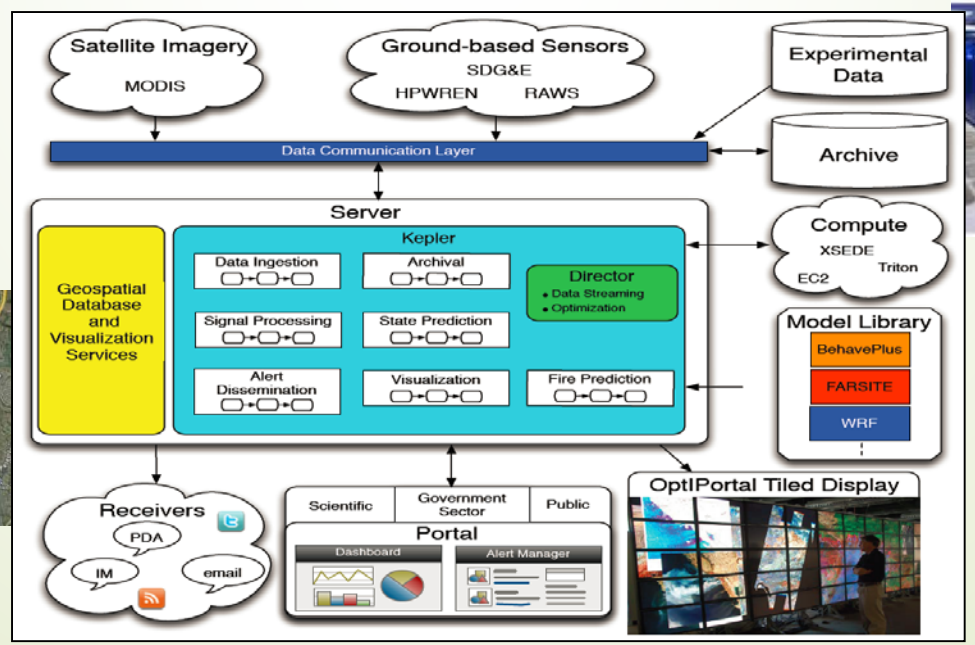
Best Data-Intensive System (End User focused) HPCwire Editors' Choice Awards 2014

Best Application of Big Data in HPC HPCwire Readers' Choice Awards 2014

Best Application of Big Data in HPC HPCwire Editors' Choice Awards 2014

wifire.ucsd.ed

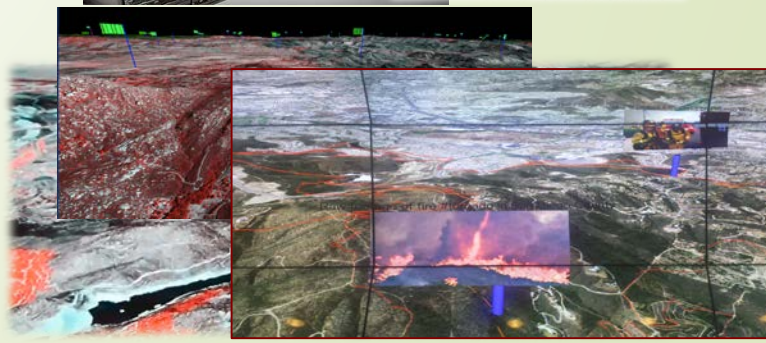




hadoop

NVIDIA TESLA

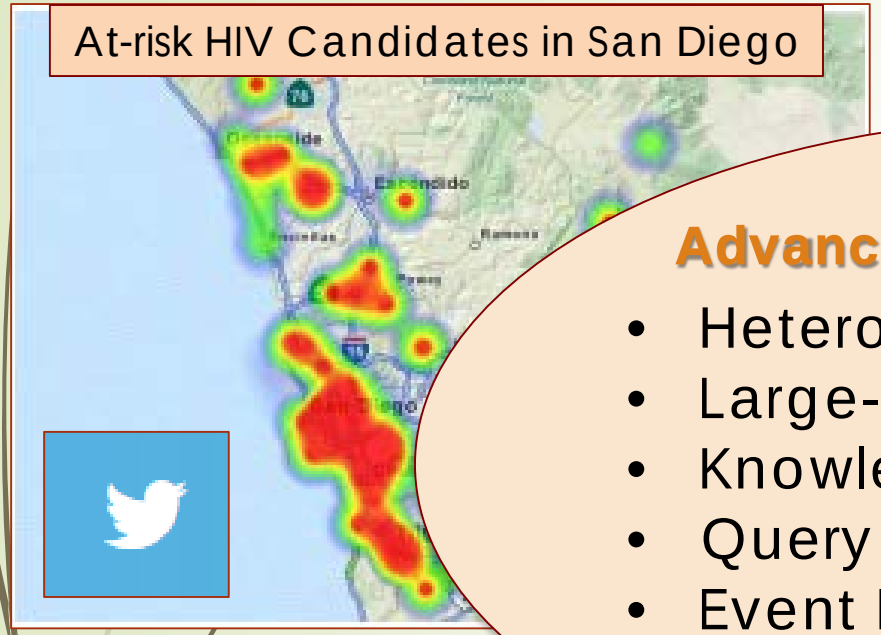
amazon web services



AWESOME: A Workbench for Exploration of Social Media

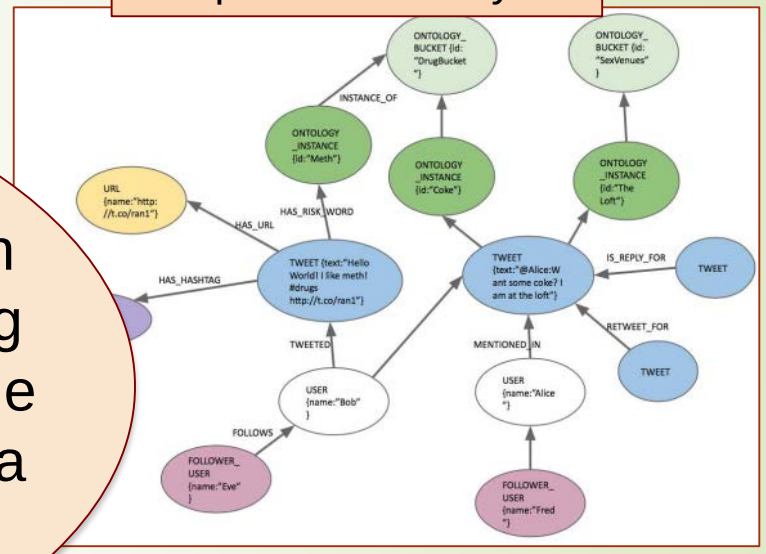
- A Polystore-Based Big Data Platform for Social Media Data -

At-risk HIV Candidates in San Diego



AQP.sdsc.edu/AWESOME

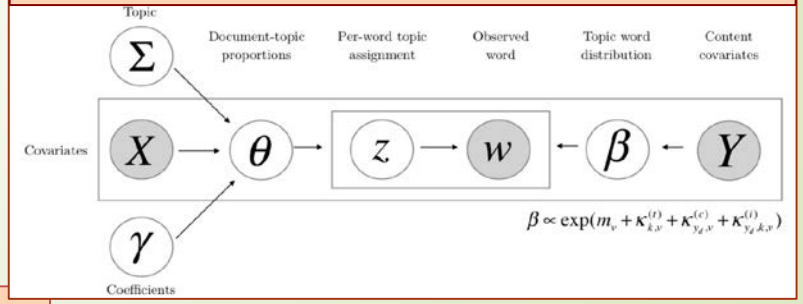
Graph Data Analytics



Advanced Query Processing Lab

- Heterogeneous Data Integration
- Large-Scale Ontology Processing
- Knowledge-based Search Engine
- Query Processing on Graph Data
- Event Modeling with Spatiotemporal Data

Parallelized Structural Topic Model



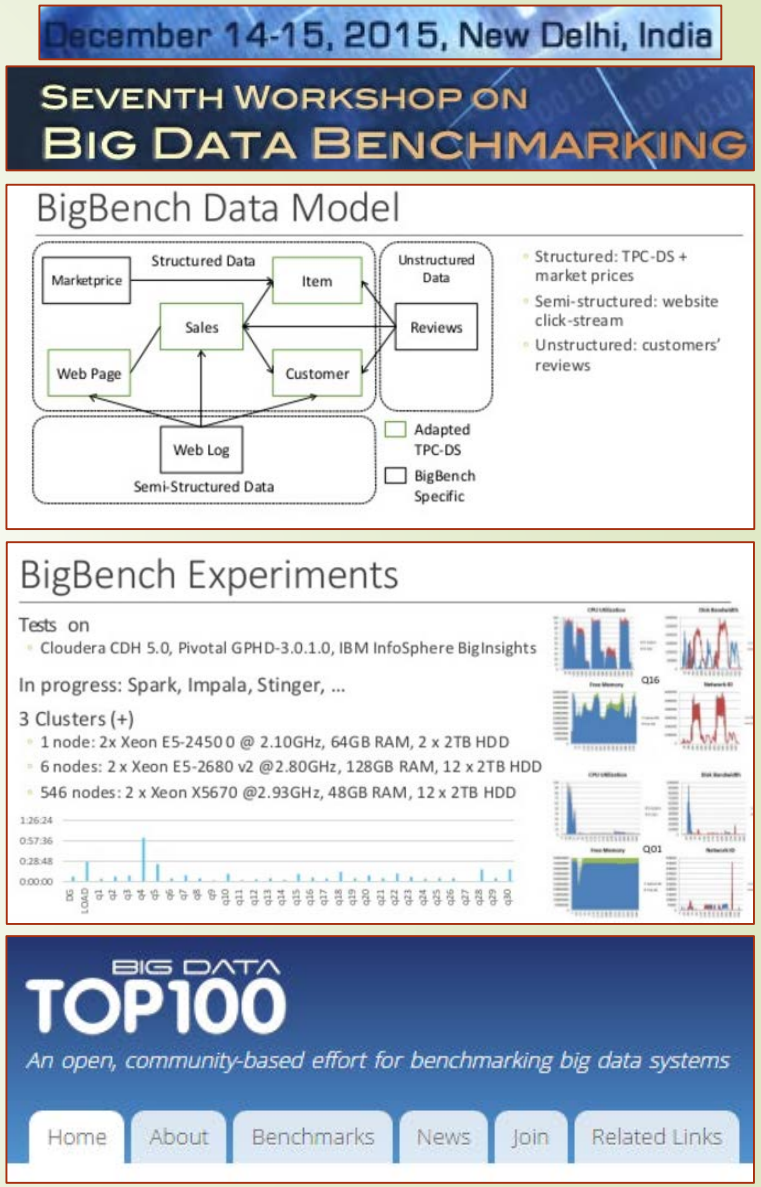
AsterixDB
more engine less trunk

NIF Information Framework *a Semantic Search Engine*
Find what you're looking for, *Faster.*
What can NIF do for me?

Information Discovery using Domain Knowledge
www.neuinfo.org

Big Data Benchmarking

- Initiated via an NSF Clue grant
  - Spinoff from the NSF Open Topography project
  - Test idea of storing all data in a scientific data archive in Hadoop, while "caching" working set data in a relational database
- New TPC benchmarks are introduced
  - BigBench – a new big data benchmarking effort with the industry
- Initiated new ideas
  - Deep Analytics Pipeline model for benchmarking Big Data/Data Science workloads
  - Creation of a BigData Top100 List
  - Referenced in NSF solicitation on Benchmarks of Realistic Scientific Application Performance of Large-Scale Computing Systems (BRAP)
- Collaborations
  - Open to collaboration in this area, e.g. identifying applications scenarios and defining corresponding benchmarks
  - Designing challenges and competitions—perhaps with industry collaboration



Chaitan Baru, baru@sdsc.edu

Convergence of Big Data and HPC – Research

May be you will have your own research ideas on this topic of convergence of Big Data and HPC

- Infiniband HPC clusters ⟷ RDMA based Apache Hadoop, Spark
- Co-processors (eg GPU, Intel MIC) ⟷ Big data platforms needs faster processing for complex analytics
- Big Data leads to public Cloud ⟷ HPC Cloud
- New mathematical, statistical algorithms for Big Data ⟷ Will need to be implemented on HPC
- High energy physics, astronomy, genomics ⟷ Doing big data + HPC for a long time
- How to inject analytics intensive workload into ⟷ HPC systems
- Hadoop and HPC ⟷ Lustre and HDFS

Thank you