

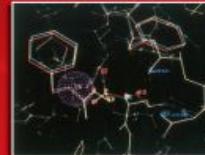
The background image shows a modern architectural complex at night. On the left, a large building with a dark, angular facade and a glass-enclosed entrance is brightly lit from within. To its right is a lower building with a light-colored, textured facade and several small trees in front. The sky is a deep blue, suggesting twilight or night. The overall atmosphere is clean and professional.

Introduction to SDSC

And its Data Science/Engineering

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



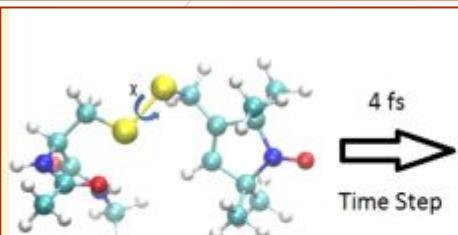
UC receives a 3-year, \$15 million grant from DEC to develop an advanced information and data management system to increase the productivity of researchers studying global change – called Project Sequoia.

1998

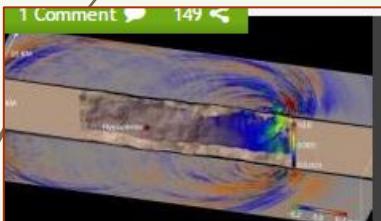
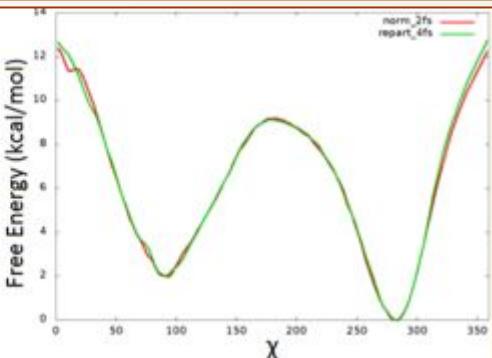
With large-scale computer simulations run at SDSC, researchers led by J. Andrew McCammon at UCSD show how one of the fastest enzymes in the world, acetylcholinesterase, does its work; results are published in the *Proceedings of the*

Science and Computing at SDSC

Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning



Ross Walker Group

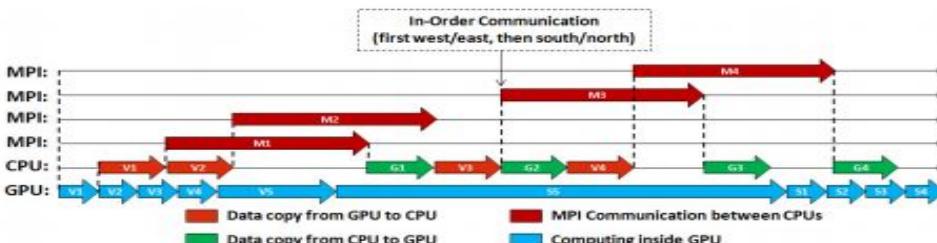


CORPORATE RESPONSIBILITY San Diego Supercomputer Center's Quake Research Wins \$150,000 Global Impact Award

By Tonic 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 Gai¹, Dong Ju Choi², David O'Neal³, Mao Ye¹ 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

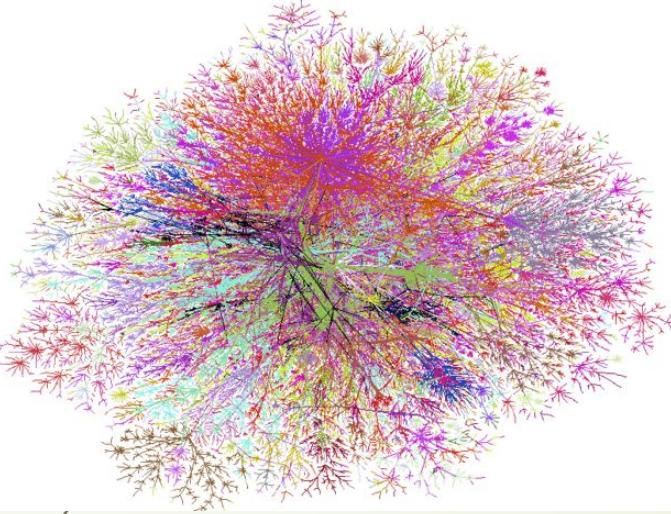
Group-based variant calling leveraging next-generation supercomputing for large-scale whole-genome sequencing studies

Kristopher A. Standish^{1,2}, Tristan M. Carland², Glenn K. Lockwood³, Wayne Pfeiffer³, Mahidhar Tatineni³, C Chris Huang⁴, Sarah Lamberth⁴, Yauheniya Cherkas⁴, Carrie

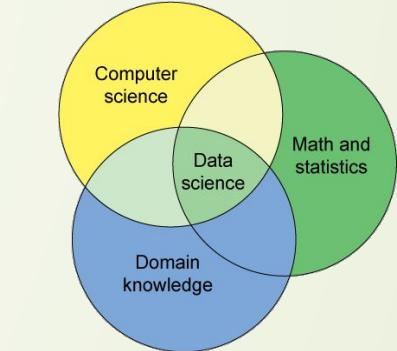
Step	Tool	Memory per command (GB)	Cores per command	Commands per node
Map	BWA	32	8	2
Bam	Samtools	4	1	16
Merge	Samtools	4	1	16
Sort	Samtools	4	1	16
MarkDuplicates	PicardTools	7	2	8
TargetCreator	GATK	7	2	8
IndelRealigner*	GATK	12	3	5
BaseRecalibrator	GATK	30	8	2
PrintReads*	GATK	30	8	2
HaplotypeCaller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

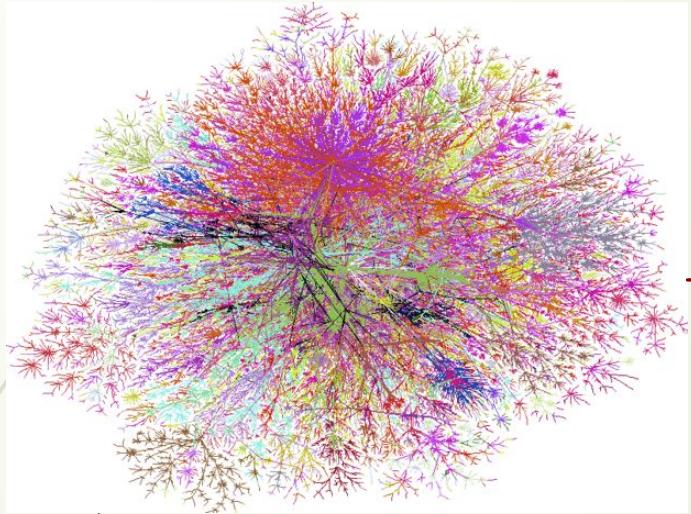


ON-DEMAND COMPUTE

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



amazon
web services

Google Cloud Platform

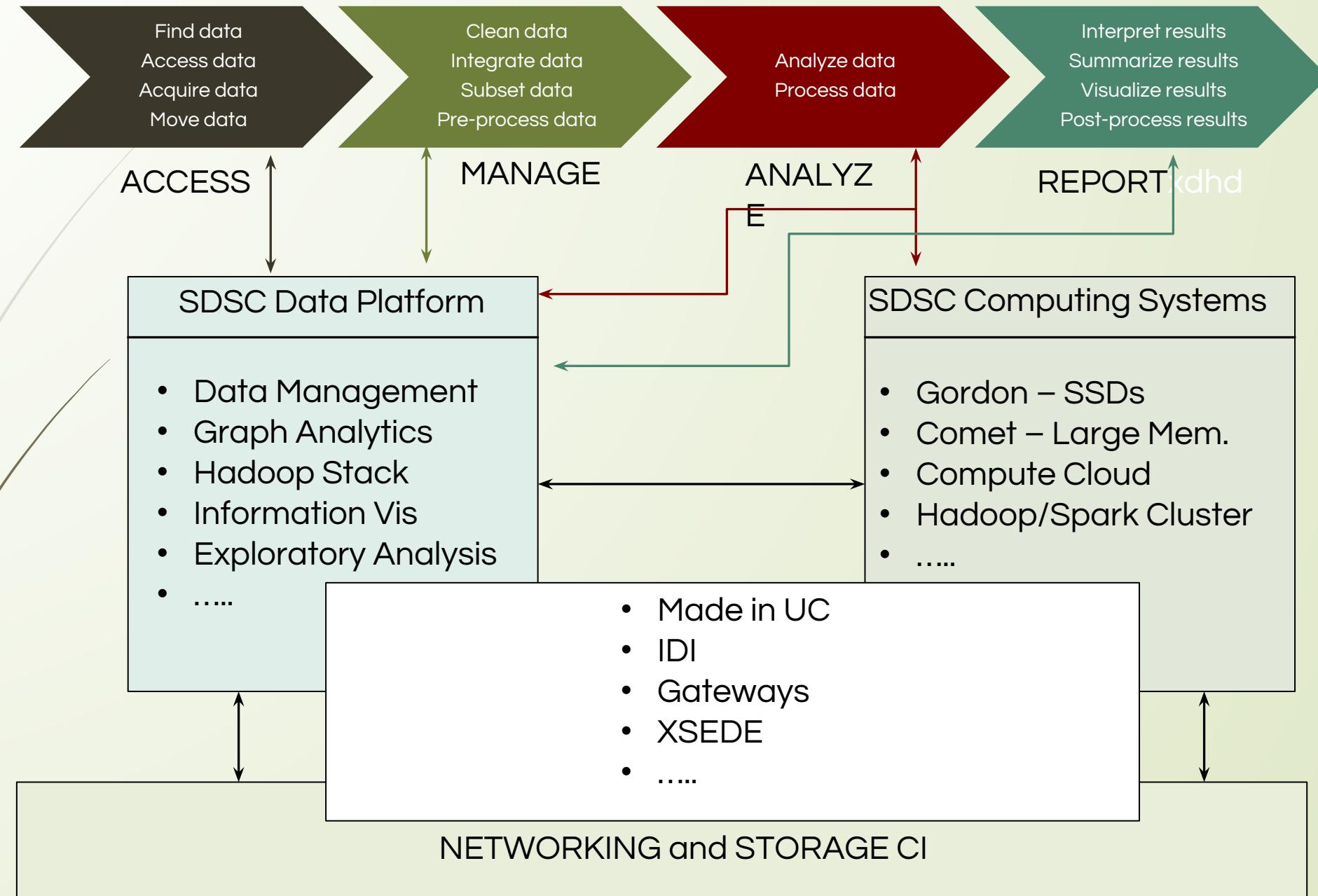
ON-DEMAND COMPUTE

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

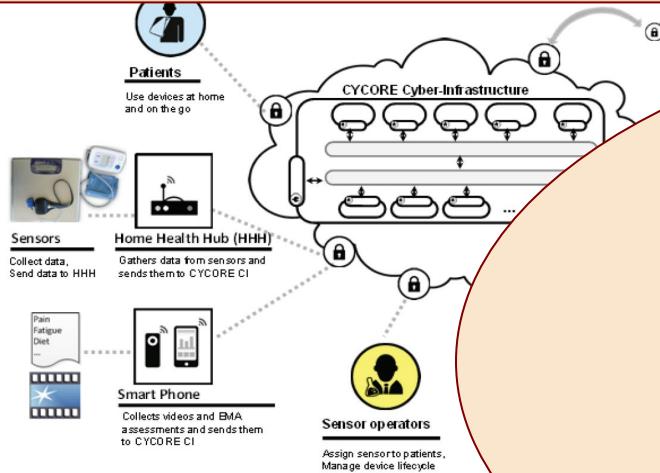
Requires support for **experimental work** by a multidisciplinary group of experts and **dynamic scalability** on many platforms!

"Big" Data Engineering

Computational "Big" Data Science



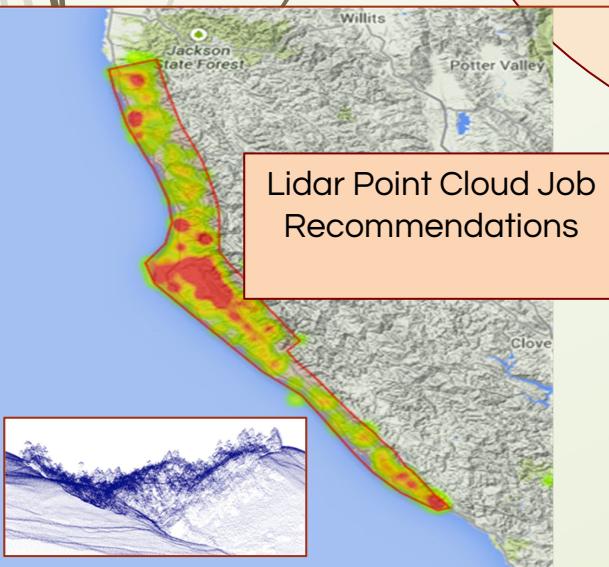
Sensing Health for Treatment Effectiveness



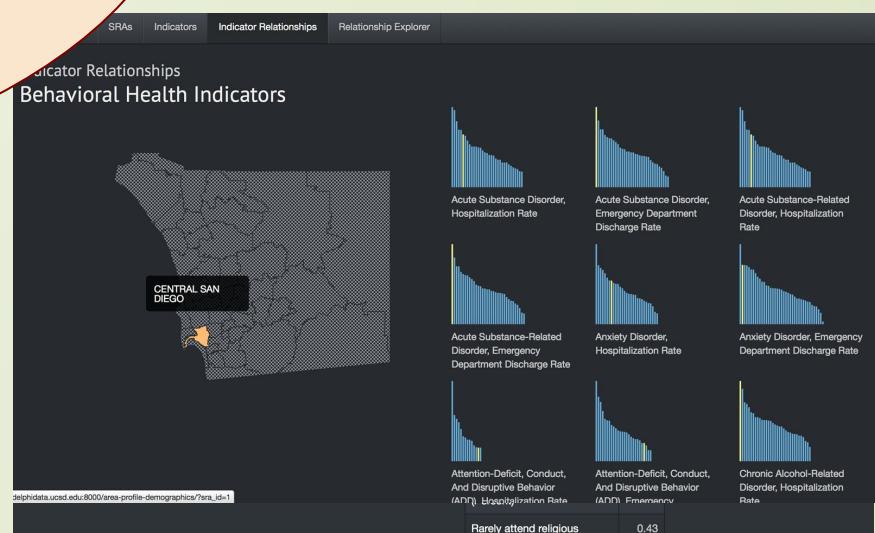
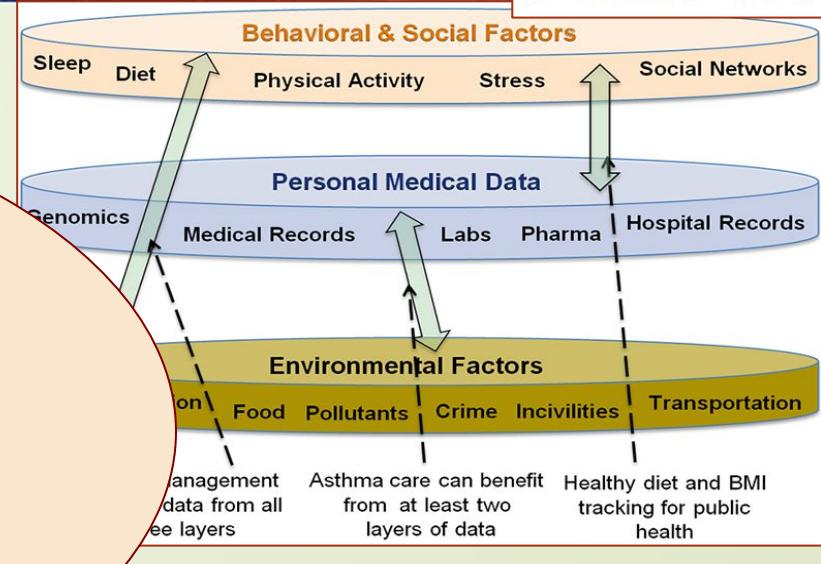
Large-Scale Data Infrastructure

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

Lidar Point Cloud Job Recommendations

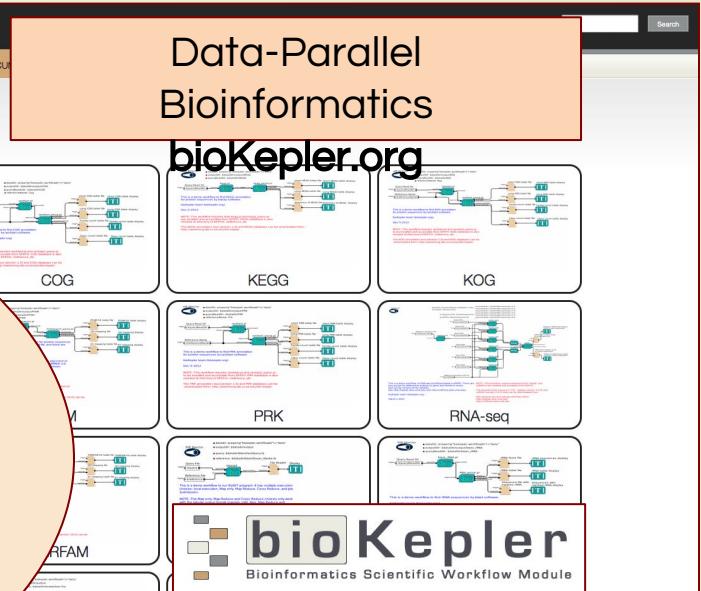
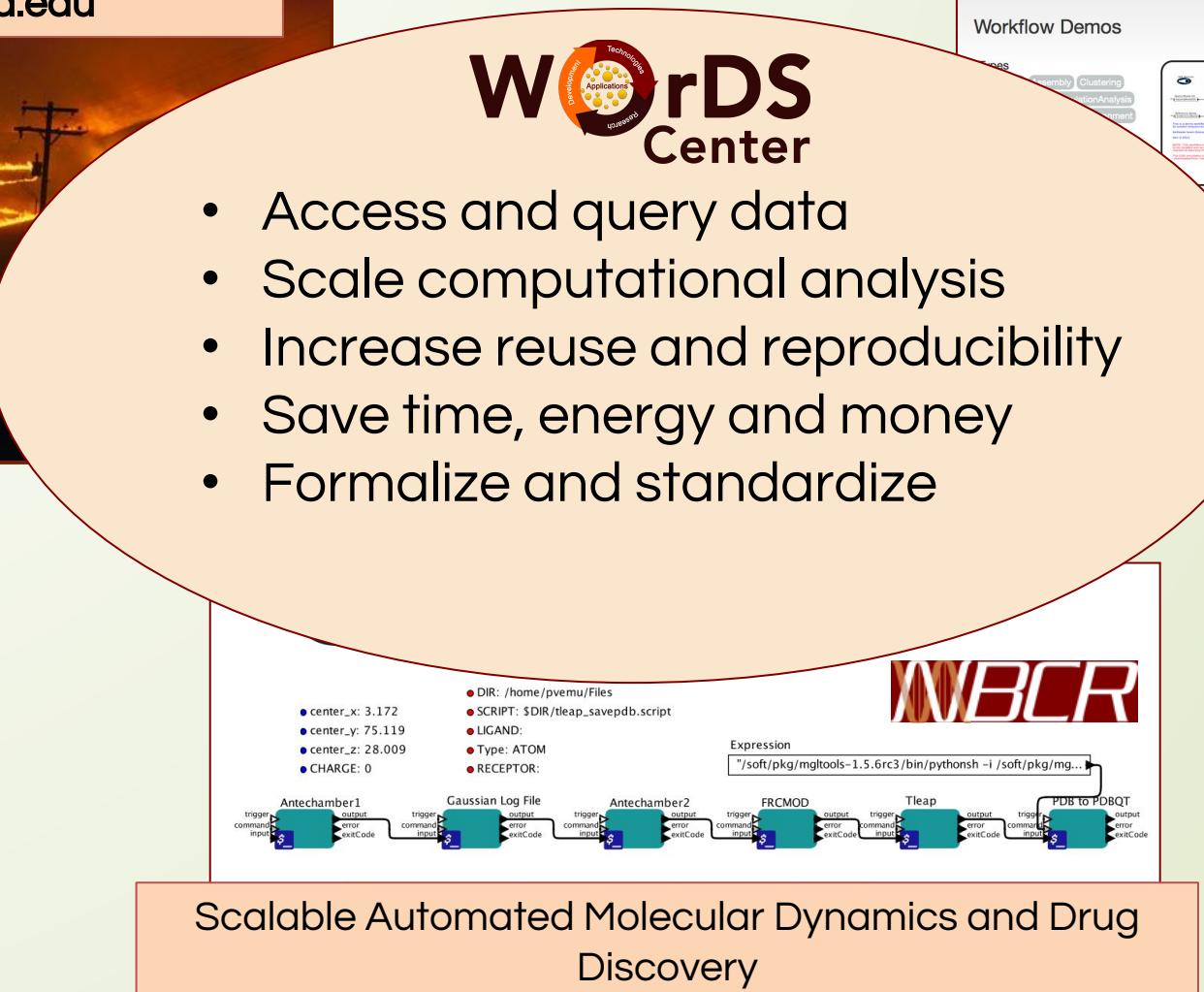


Monitoring Wild Life with Videos



Workflows for Data Science (WorDS) Center at SDSC

- Programmable, Reusable and Reproducible Scalability -



Best Data-Intensive System
(End User focused)



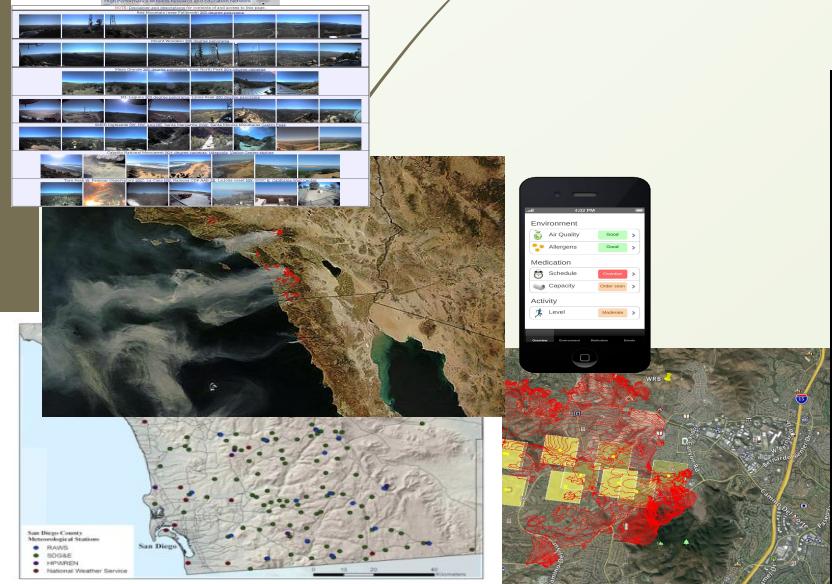
Best Application of Big Data in HPC



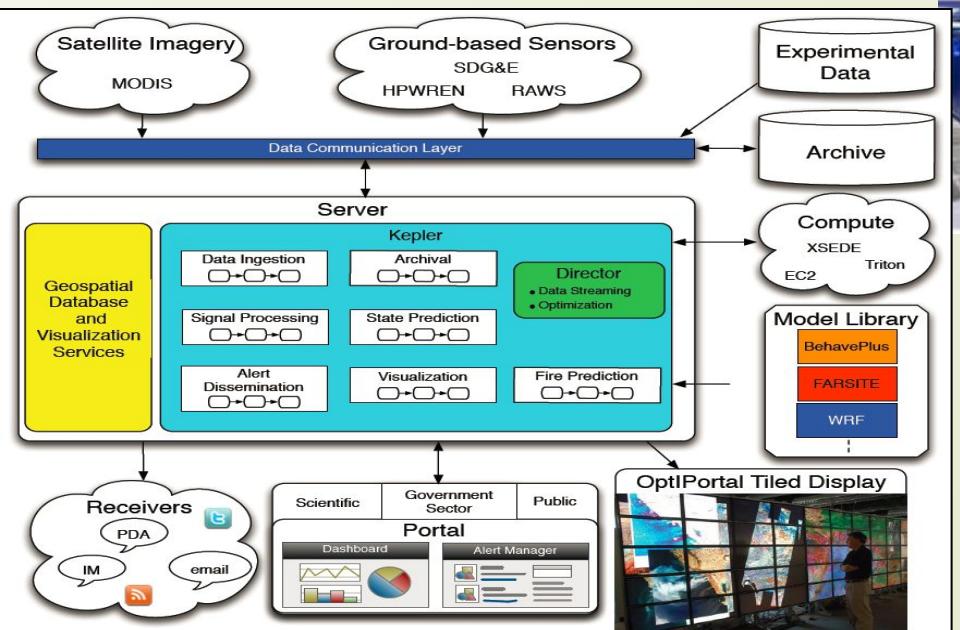
Best Application of Big Data in HPC



Big Data



Monitoring Visualization Fire Modeling



wifire.ucsd.edu

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



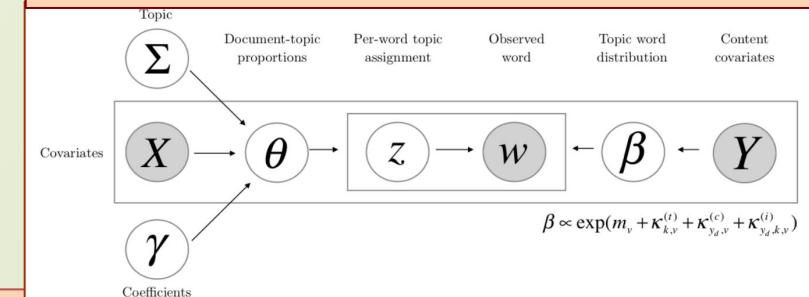
NIF Information Framework *a Semantic Search Engine*

Find what you're looking for,
Faster.

What can NIF do for me?

Information Discovery using Domain Knowledge
www.neuinfo.org

Parallelized Structural Topic Model



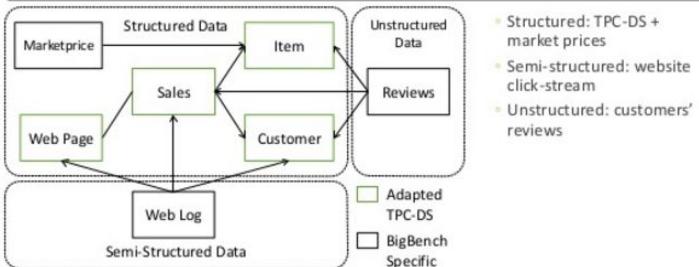
Big Data Benchmarking

- 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 workload
 - 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

December 14-15, 2015, New Delhi, India

SEVENTH WORKSHOP ON BIG DATA BENCHMARKING

BigBench Data Model



BigBench Experiments

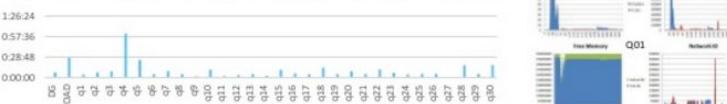
Tests on

Cloudera CDH 5.0, Pivotal GPHD-3.0.1.0, IBM InfoSphere BigInsights

In progress: Spark, Impala, Stinger, ...

3 Clusters (+)

- 1 node: 2x Xeon E5-2450 v0 @ 2.10GHz, 64GB RAM, 2 x 2TB HDD
- 6 nodes: 2x Xeon E5-2680 v2 @ 2.80GHz, 128GB RAM, 12 x 2TB HDD
- 546 nodes: 2x Xeon X5670 @ 2.93GHz, 48GB RAM, 12 x 2TB HDD



BIG DATA TOP100

An open, community-based effort for benchmarking big data systems

[Home](#) [About](#) [Benchmarks](#) [News](#) [Join](#) [Related Links](#)

Chaitan Baru, baru@sdsc.edu