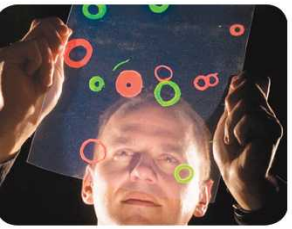
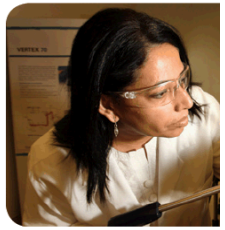
SAND2013-4736C





Sandia's Data Science Research Challenge

Ann N. Campbell, PhD
William E. Hart, PhD

June 12, 2013

Sandia National Laboratories

Exceptional service in the national interest

U.S. DEPARTMENT OF ENERGY

NNSA

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Overview

Sandia is doing strategic planning to identify focus areas:

- Mission Challenge: An obstacle to achieving a *national security mission* that is appropriate for a national security laboratory to address.
- Research Challenge: A science or engineering obstacle to solving a *mission challenge* that is appropriate for a national security laboratory to address.

Data Science was recently identified as a cross-cutting capability that would impact a wide variety of *candidate mission challenges*.

What is Data Science?

Sandia National Laboratories

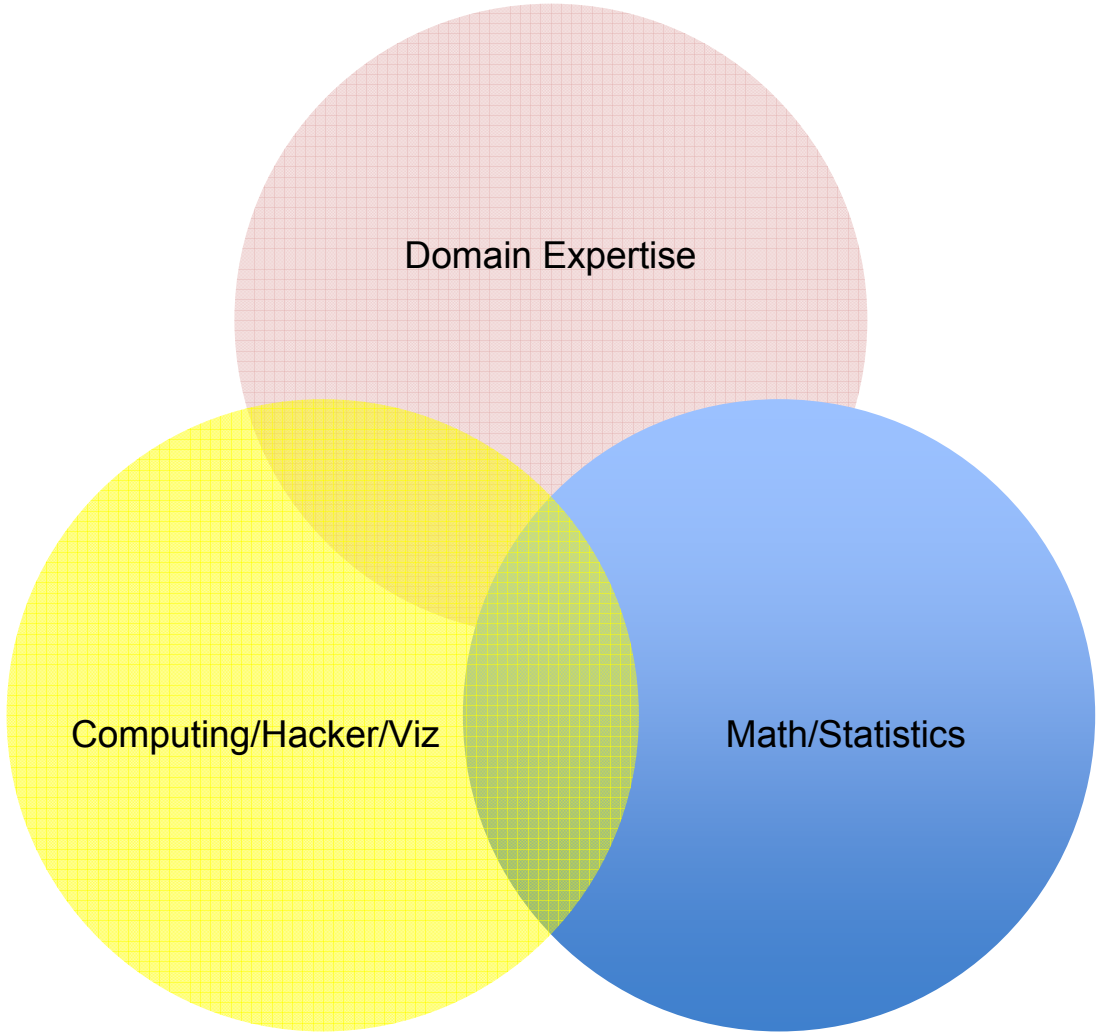
Data Science is the practice of

- Discovering what we don't know from data
- Obtaining predictive, actionable insight from data
- Creating *Data Products* that have mission impact
- Communicating relevant stories from data
- Making high-confidence decisions

Data Science is often used interchangeably with *business analytics*.

What is a Data Scientist?

Sandia National Laboratories



Why Data Science Now?

Sandia National Laboratories

Explosion of large, structured, unstructured and semi-structured data

- Scientific data, simulation data, business data, social data, …
- Management of *Big Data* is a key dimension of Data Science

Powerful computing platforms are commonly available

- Desktop PCs as powerful as previous HPCs
- Commodity clusters are affordable
- Large storage devices are cheap

Big Data has proven to have commercial value

- Can detect patterns that are not apparent in smaller data sets
- Can track and predict patterns in human behavior

History/Context at Sandia

Sandia National Laboratories

Data analysis is widely performed at Sandia

- Cybersecurity, bioinformatics, infrastructure security, WFO, sensor analysis, NW assessments

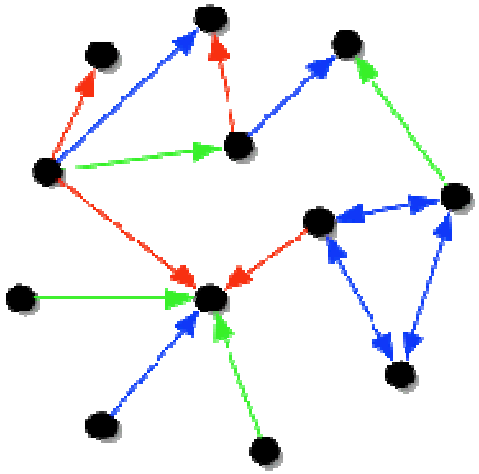
Sandia has distinguishing technical capabilities that support data science

- Graph algorithms
- Discrete math
- Scalable spectral methods
- Sparse tensor decomposition

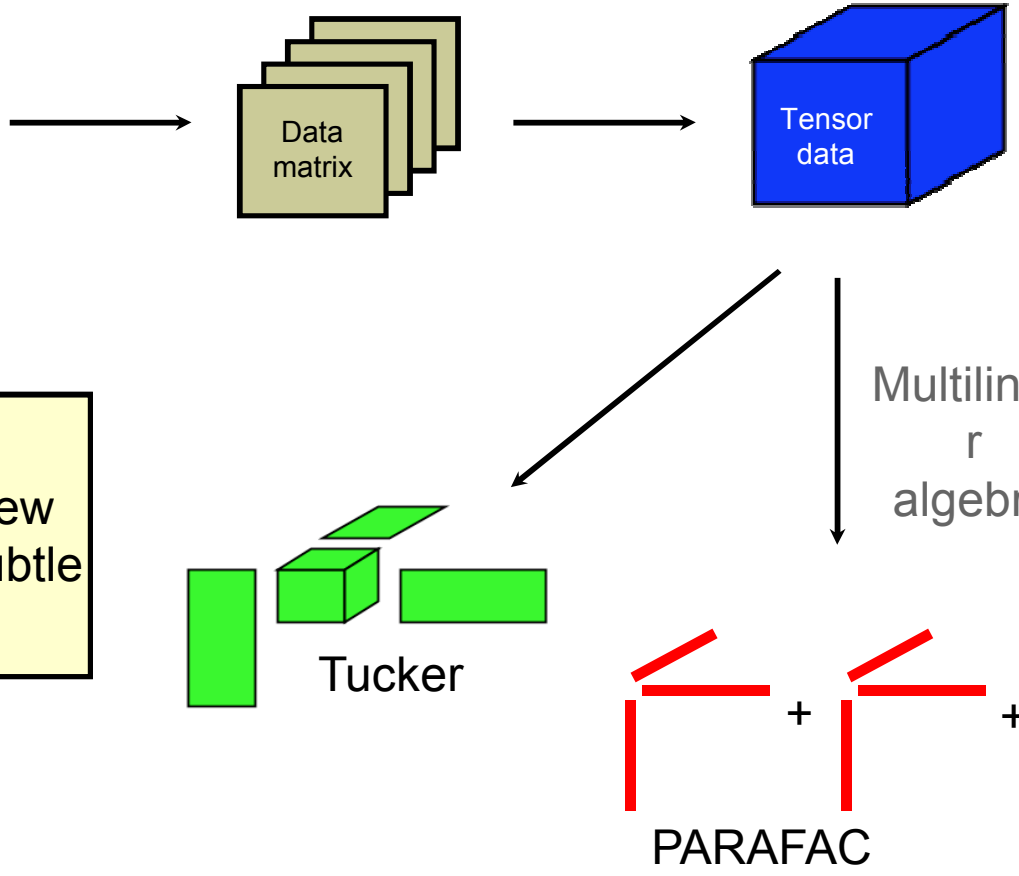
Sandia's LDRD program has made significant data science research investments

- NGC (2008-2010): analysis of social network and graph data
- PANTHER GC (2013-): analysis of image-like data

Tensor Analysis

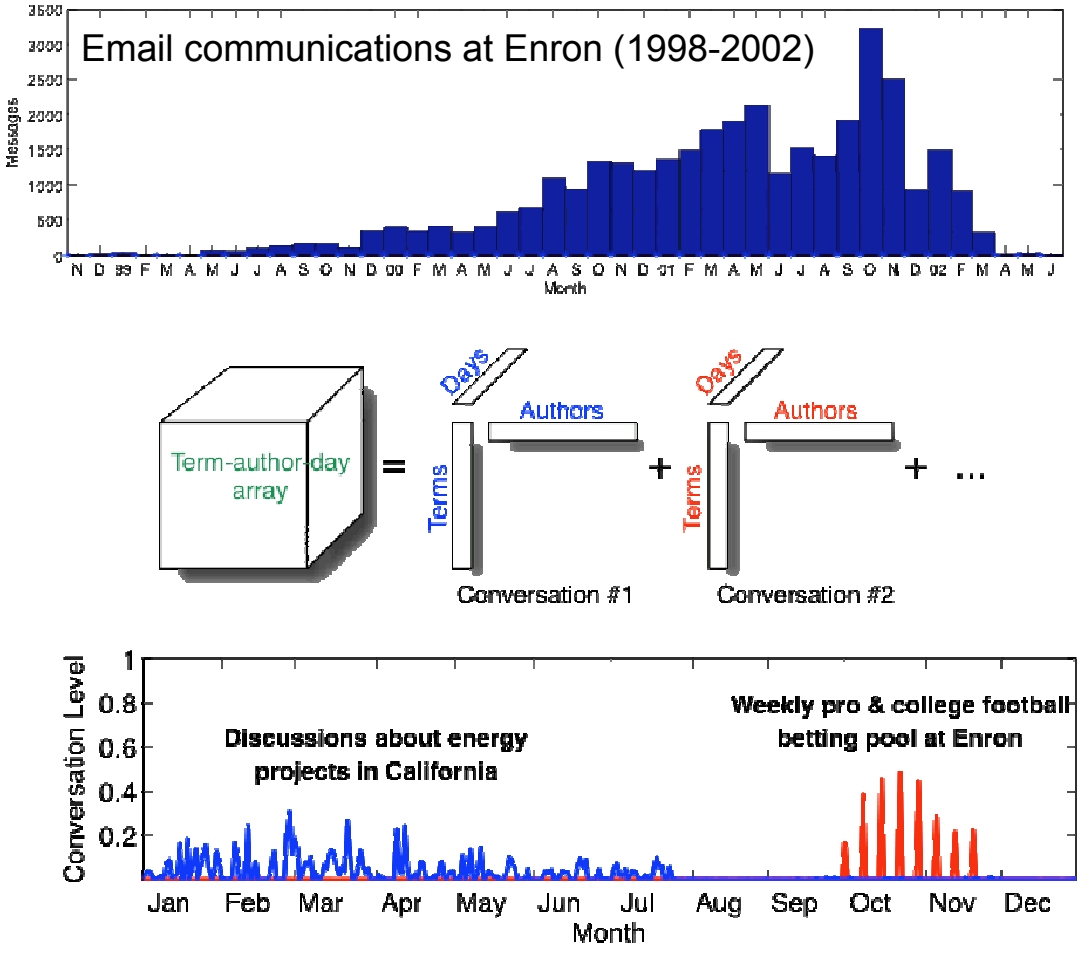


Build a “data array” such that there is a data matrix for each link type.



Tensor analysis offers more explanatory power: uncovers new latent information and reveals subtle relationships

Factoring Sparse Tensor Data

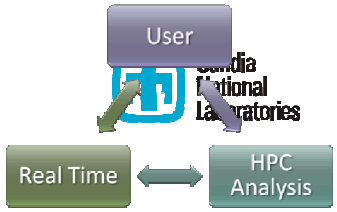


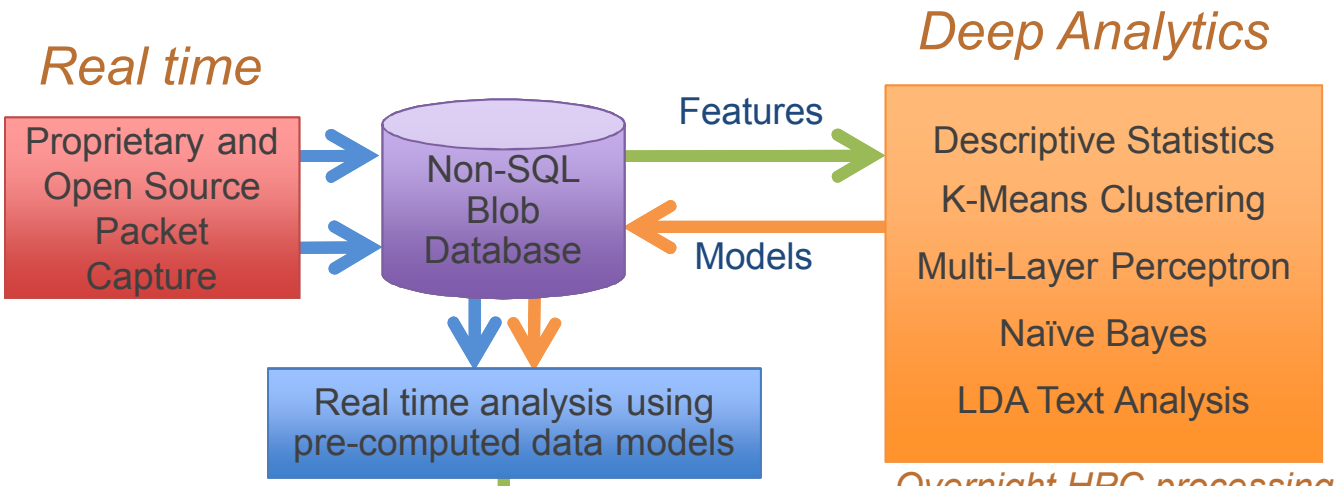
Tensor factors characterize data structure

Tensor Toolbox

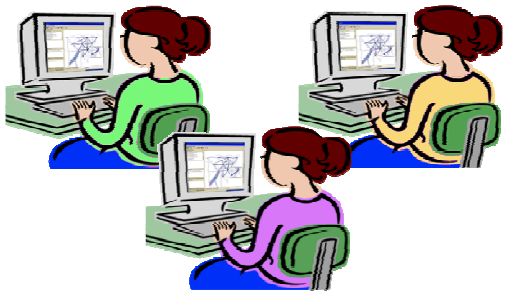
- Developed by Sandia
- Unique software for sparse multi-dimensional data
- Sparse calculations for decompositions
- Parallel decomposition techniques being developed (MPI, threading, map-reduce)

Hybrid Methods Overview



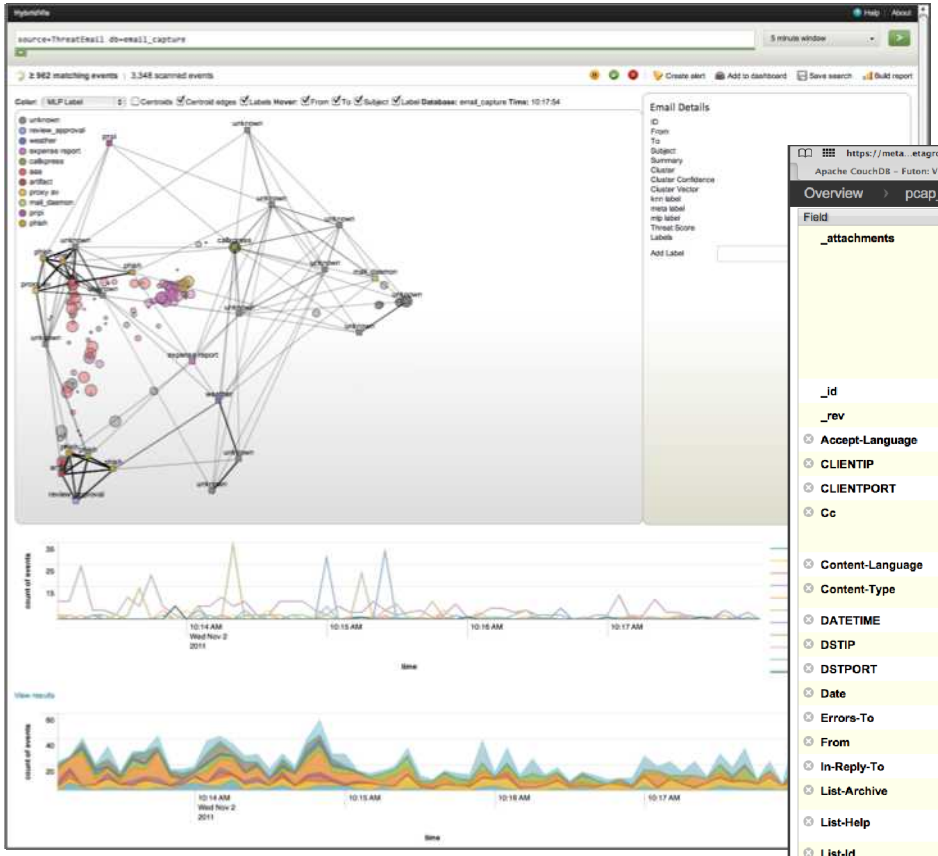


User/Web Interfaces

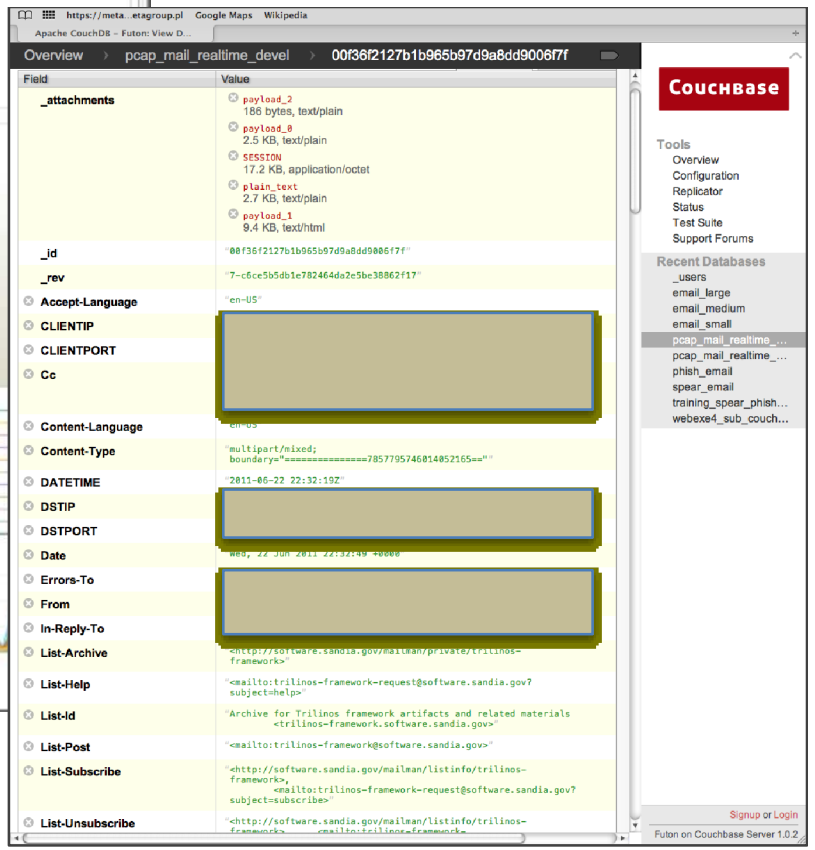


Web interfaces support cyber-team collaboration

Web Delivery of HPC Analysis



Realtime display in Splunk™ web interface

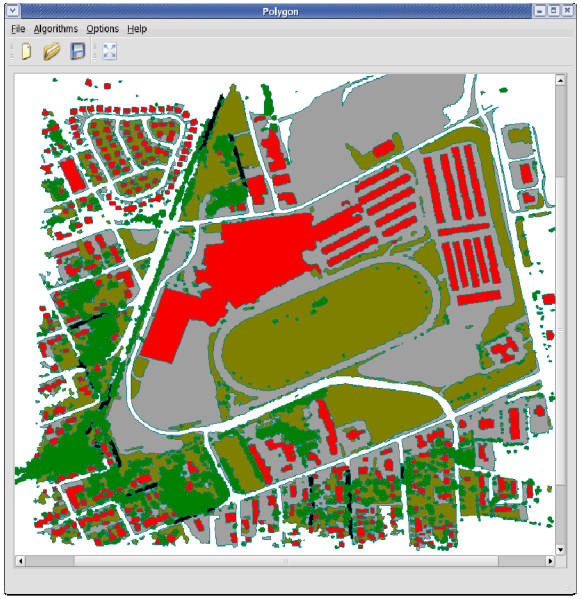


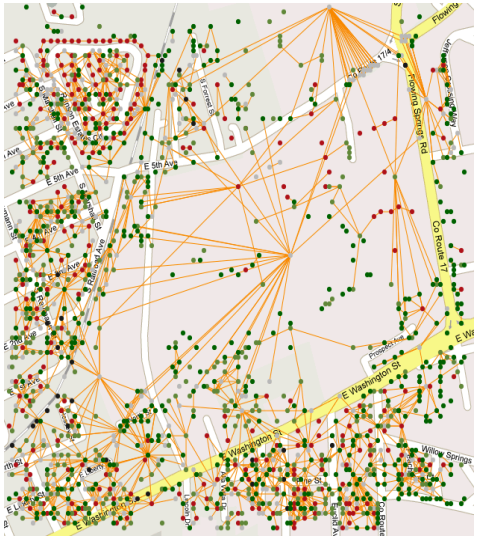
Artifacts stored in CouchDB for easy one click drilldown to get detailed information.

Activity Tracking via Semantic Graphs

Sandia National Laboratories

Concept: Don't store images as pixels; store as groups of objects and their relationships





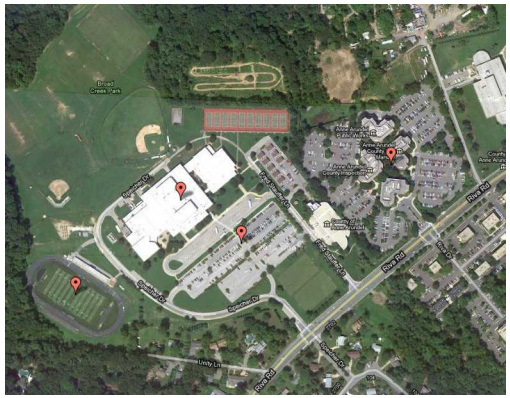
Can then search for analyst-like patterns

- "Find a large building near water that is not near any other building"
- "Find a small building that has a large parking lot near it"

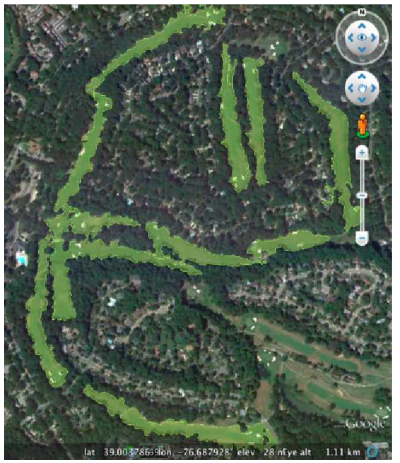
Creates a computer-driven pipeline that can cue analysts towards specified patterns. (NA-22 funding)

Examples

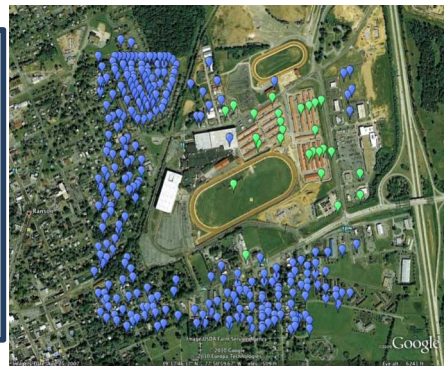
Found public high schools in suburban Maryland (large building next to parking lot and field)



Found golf courses (collections of thin, grassy areas with trees nearby)



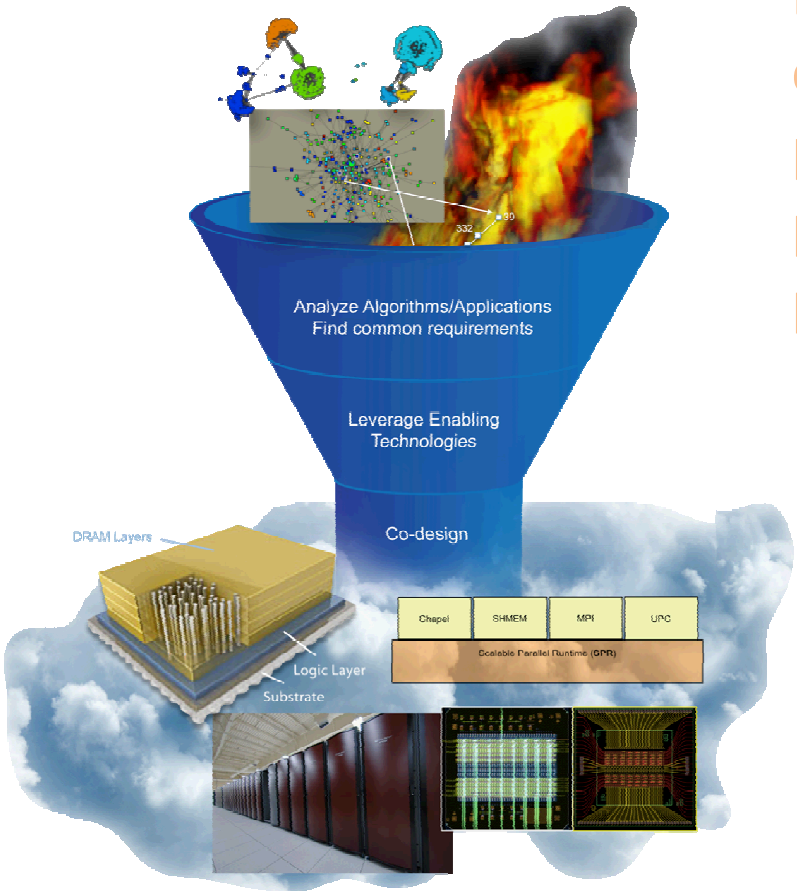
Separated distinct land use classes using *unsupervised* learning (suburbs vs. horse track facilities)



Advantages

- Queries are “analyst-like”
- Not sensitive to orientation or small pixels changes
- Orders of magnitude less data in search

Extreme-scale Computing Grand Challenge



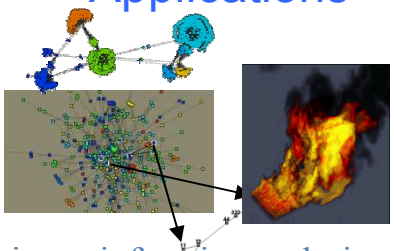
Develop key system architecture concepts that will enable a unified physics/data analytics computing platform or determine what gaps prevent it

Extreme-scale computing will be a key capability in meeting Sandia's national security missions. Understanding where application requirements for the two mission areas overlap will drive development of a set of common components to efficiently support both areas.

XGC Leverages Full System Expertise

Sandia National Laboratories

Applications

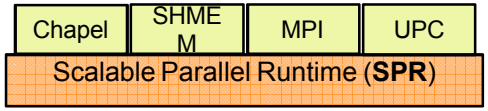


Miniapps informing codesign

Concurrency management primitives
Communication primitives

Application characteristics
Software primitives evaluations

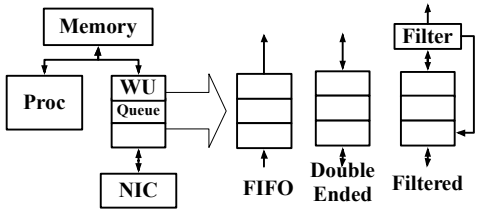
System Software



Task based runtime integrated with communication stack

Application characteristics
Architecture evaluations

Architecture



Hardware capabilities
Hardware architecture

Data movement architectures to alleviate bottlenecks

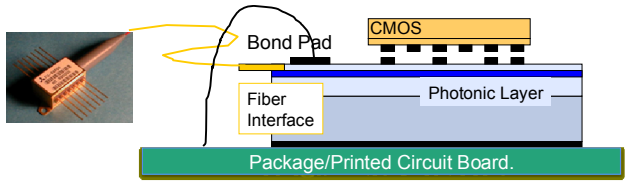
Memory hierarchies
Compute characteristics

Software capabilities
Software primitives

Novel architectures using
Si Photonics and evaluation

Latch-to-latch Si Photonics
Hardware characteristics

Photonics



Photonic/electronic integration working towards a 100 fJ/bit link energy

Full system approach allows each level of the system architecture to inform the others. Advances within each level consider the others and are directed at improving overall system efficiency.

Anomaly Detection in Cyber Data

Sandia National Laboratories

VAST 2009 Challenge

Goal: Identify an insider threat in a cyber environment.
Data: Header information from 115,414 network events over 1 month.

Source IP	Access Date/Time	Destination IP	Socket	Req Size	Resp Size
37.170.100.38	01/01/08 09:40 AM	37.170.100.200	80	7063	49591
37.170.100.38	01/01/08 09:43 AM	37.157.76.124	80	5171	434285
. . .	. . .	. . .	. . .	. . .	. . .

Approach: Use a probabilistic semantic analysis to identify unusual patterns.

- Header attributes treated as features for analysis
- Latent Dirichlet Allocation(LDA) used to generate a soft clustering of network events into 100 groups

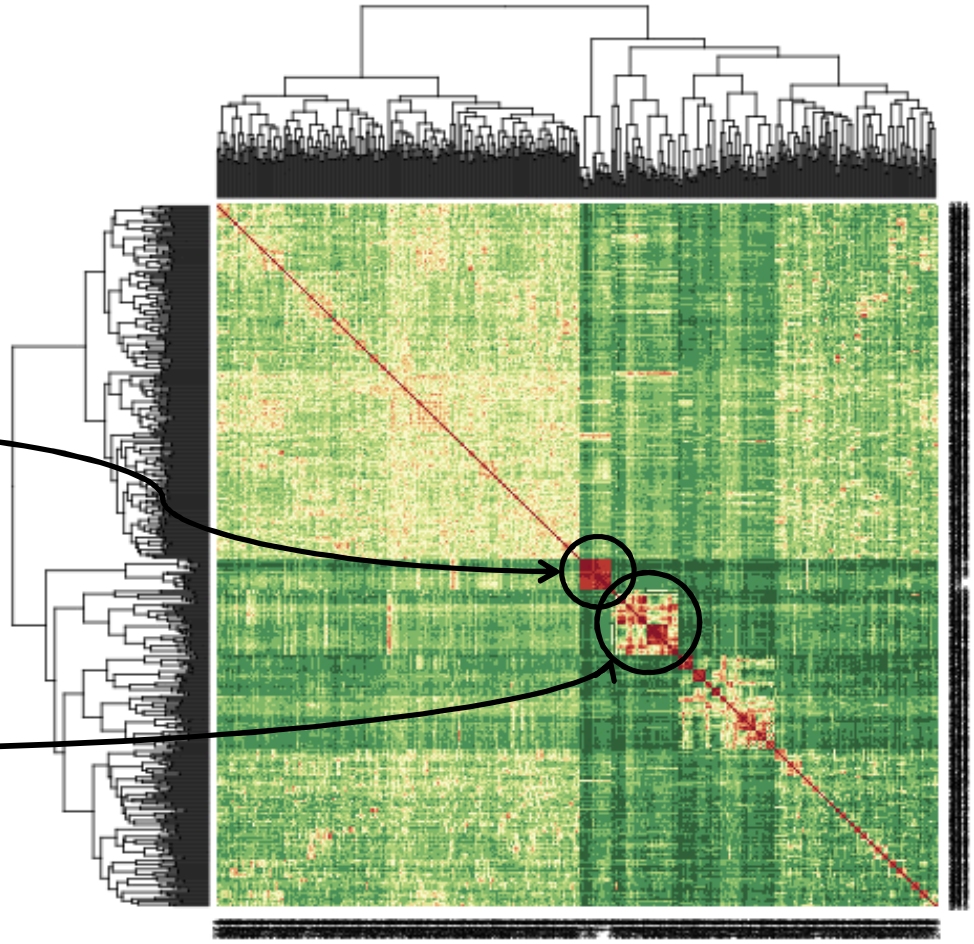
VAST Contest Results

Sandia National Laboratories

Cluster analysis of LDA

The 18 target events in this dataset

Another anomalous cluster of weakly related events



Research Challenge

Sandia National Laboratories

Goal: Develop scalable techniques for data analysis that enable human analysts to rapidly identify, characterize, and respond to key signatures buried in complex, heterogeneous data and information.

Vision/Impact: Sandia is an acknowledged leader for providing end-to-end solutions for collecting and effectively leveraging data to provide differentiating decision support to the national security community.

Mission Context

Large data sets are increasingly common in many mission areas

- Sensor data produced by DoD and IC assets
- Physical experiments and computational simulations
- Data streams from cyber, energy and other critical infrastructures

Data Science can derive new insights from this data

- Detect anomalies, model patterns of life, identify leading indicators, characterize threat signatures, etc.

Data analysis is becoming a bottleneck in some missions

- We cannot hire enough analysts to meet future demands

Why Sandia?

Sandia National Laboratories

We have a good start on "owning the signatures"

- We create sensor technologies that support critical national security missions.
- We are deeply involved with analyzing data coming from these systems, and in some cases we "own" the data.
- We effectively leverage the strong synergy between the sensor system and the data to drive successive development of each domain.

Data Science expertise is a critical enabler for meeting the needs of our mission partners.

- The national security community comes to Sandia because our technical strength in data analytics enables solutions to very challenging problems
- A focus on Data Science will enable SNL to continue to attract top talent

Issues/Challenges

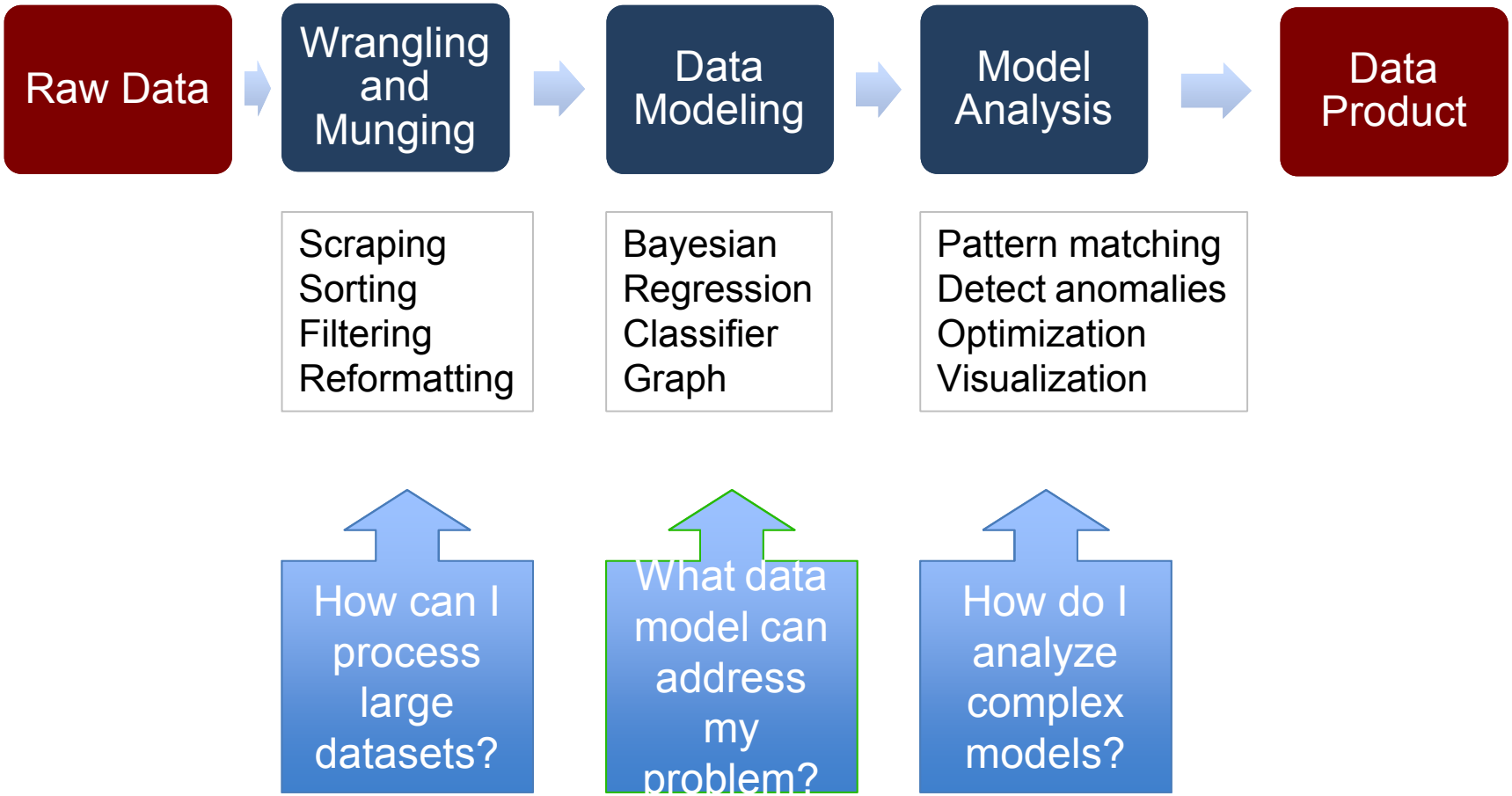
- Does Sandia's national security mission create a distinguishing role for Sandia in Data Science?
- What type of research investments will enable Sandia to work on a wide range of Data Science applications?
- What are cross-cutting research challenges that reflect Sandia's national security mission?
- What institutional investments are needed to foster Data Science R&D?
- How do we motivate mission-centric organizations to develop/promote cross-cutting Data Science capabilities?

Appendix

Where's the Research?

Sandia National Laboratories

Steps in a canonical data analysis:



What is Big Data?

- Volume
- Variety
- Velocity
- Etc.

What is a Good Data Scientist?

Sandia National Laboratories

- Is skeptical, curious. Has inquisitive mind
- Knows ML, Stats, Proby
- Applies scientific method. Runs experiments
- Is good at Coding and Hacking
- Able to deal with IT Data Engineering
- Knows how to build data products
- Able to find answers to known unknowns
- Tells relevant business stories from data
- Has Domain Knowledge

Data Science Principles

Sandia National Laboratories

- Socio-Technical Systems are complex
- Data is never at rest
- Data is dirty
- There is not single version of the truth
- Data munging and data wrestling > 70% of time
- Simplifications. Reduction. Distillation.
- Curiosity. Empiricism. Skepticism.

DS and Staffing

Sandia National Laboratories

We need …

- Hard scientists with math/stats backgrounds and an appreciation of the value of data
- Hackers that are willing to dig into data sources and understand IT technology
- People that can understand and value what a user needs
- Entrepreneurial individuals that could start their own company
- People that can convey complex data to a nontechnical audience through visual and anecdotal stories that reflect new insights