An Update on Data Science at LLNL

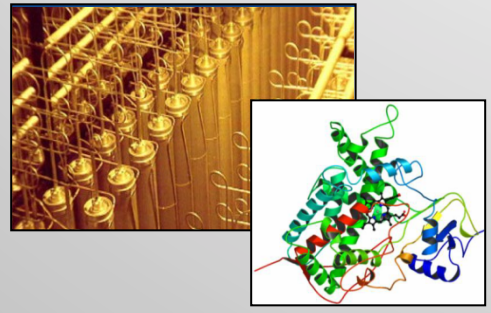
James M. Brase
Deputy Associate Director, Computation
Lawrence Livermore National Laboratory

Lawrence Livermore National Laboratory

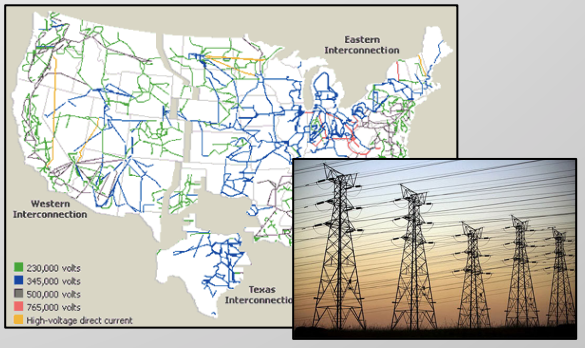
LLNL-PRES-XXXXXX
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

LLNL priorities increasingly focus on the security and resilience of complex networks and information systems

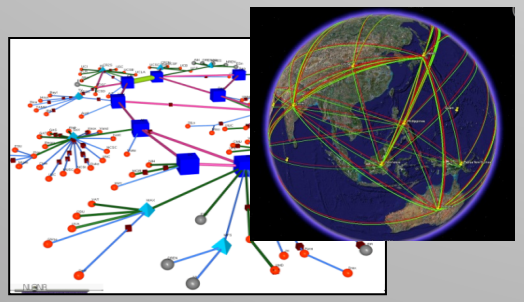
Detection and neutralization of global asymmetric threats



Resilience and security of critical infrastructure



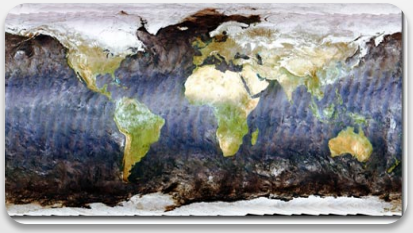
Managing risk in information and communications



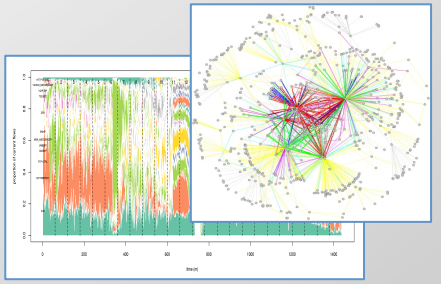
Complex physical and social systems

The Data Science Initiative is building the foundations needed for these missions

Critical complex systems

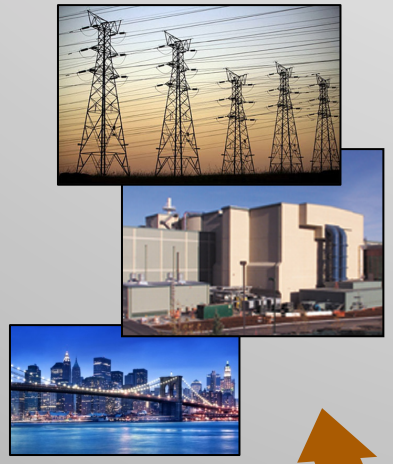


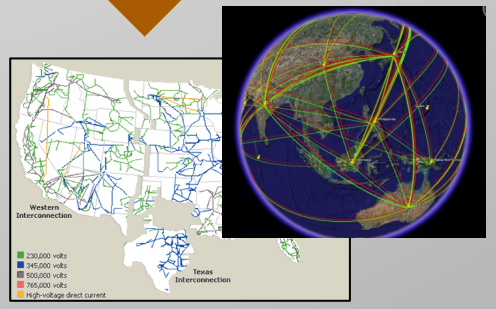
Managing large, distributed data



Discovering patterns of behavior in the data

New data-intensive computing architectures





Creating predictive analytic and simulation models

Applications in
- Detection and intervention
- Resilience and security
- Policy impacts

We face challenges that will require new combinations of large-scale data analytics with modeling and simulation

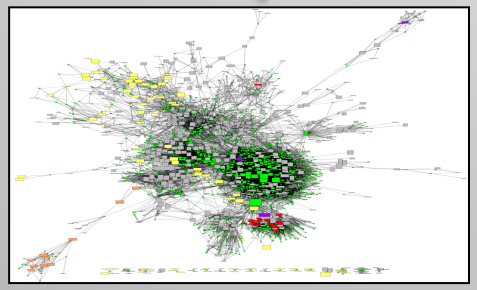
Data-driven analytics

Integrated science-driven analytics and simulation

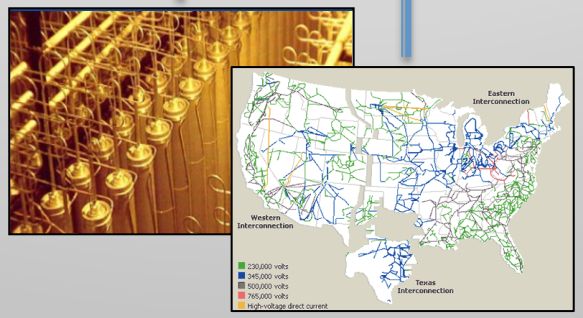
Multi-scale simulation

Data >> model parameters

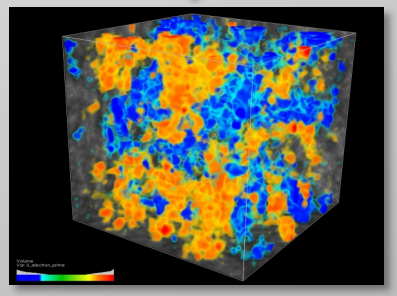
Model parameters >> Data



Cybersecurity analytics



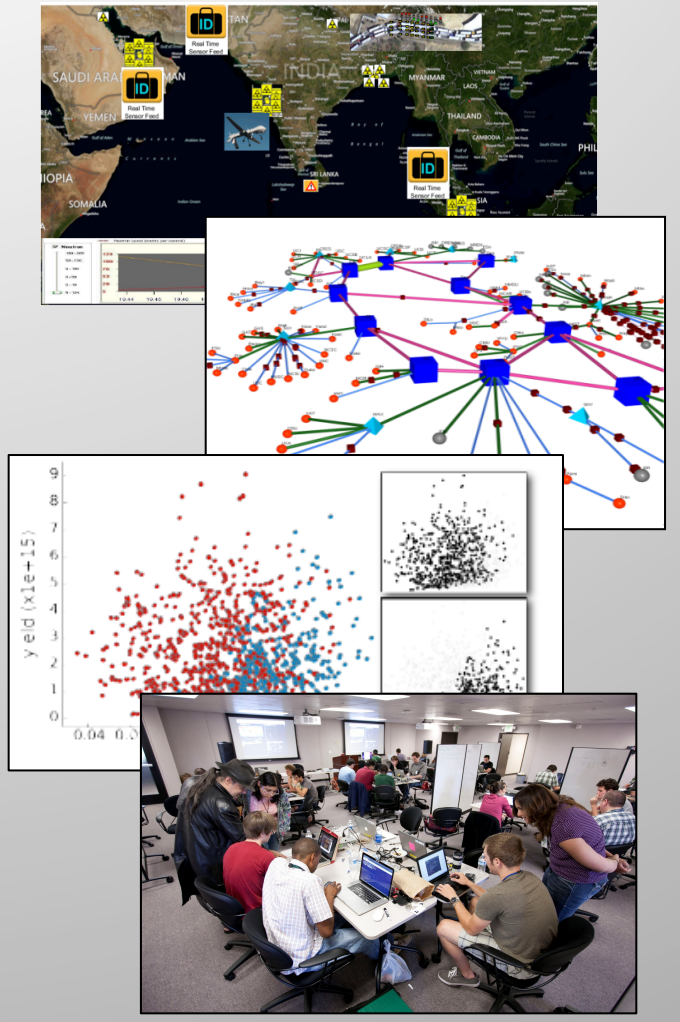
Nonproliferation and critical infrastructure protection



Material properties simulations

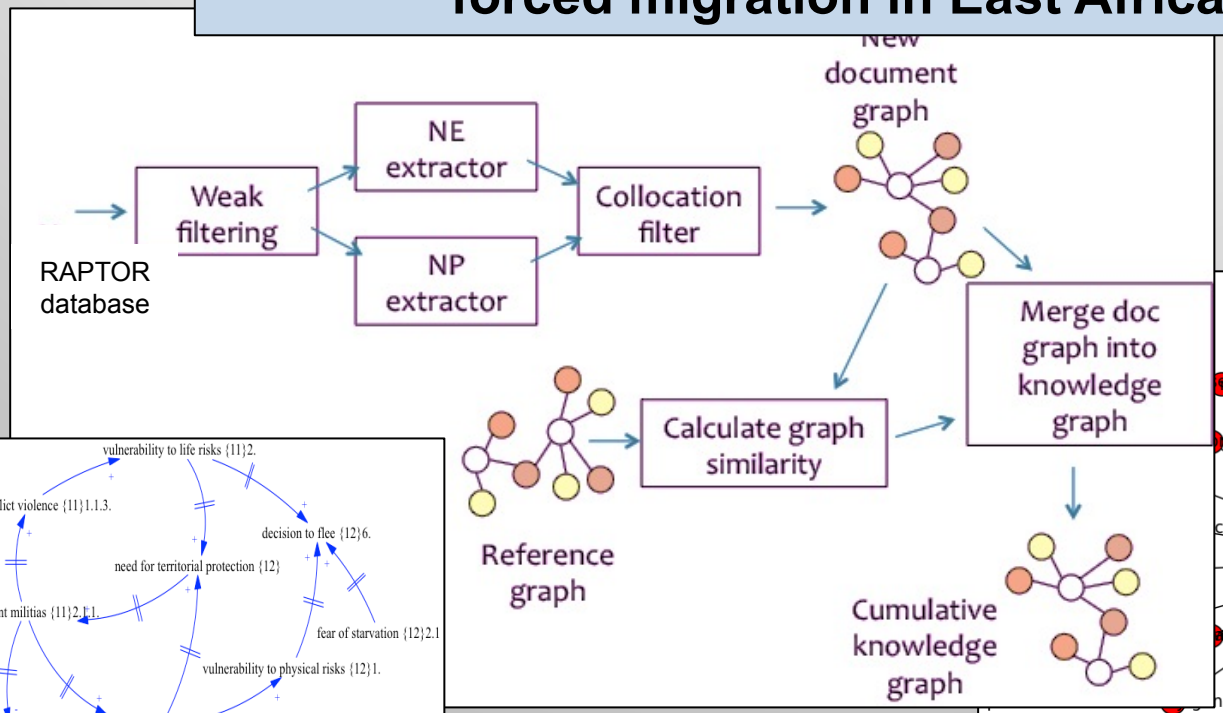
LLNL has defined a set of mission priorities that drive our investments

- **Science-based situational awareness - a cross-cutting theme in Global Security**
  - **Counter-proliferation / CWMD / intel analytics**
  - **Social science and policy applications – learning and simulating complex systems**
- **Genomic, clinical analytics, and agent-based simulation focused on predictive biology and biosecurity**
- **Next-generation analytics for large-scale physical simulation – in-situ analytics and uncertainty quantification**
- **Cyber mapping, analysis, and simulation**
- **Increasing and broadening student programs and Bay Area outreach**

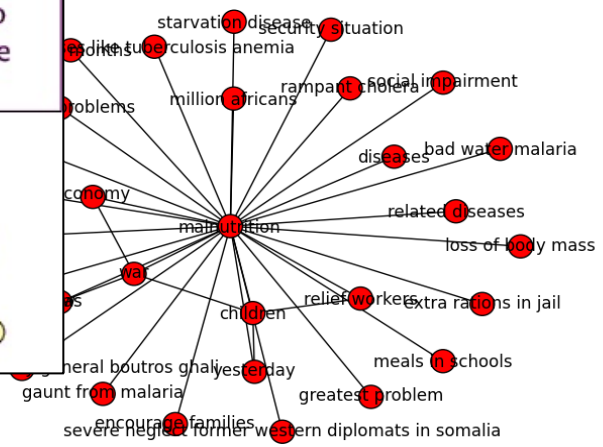


A partnership with Georgetown on computational social science provides a testbed for analytic methods

Can we forecast areas of increased probability of forced migration in East Africa?

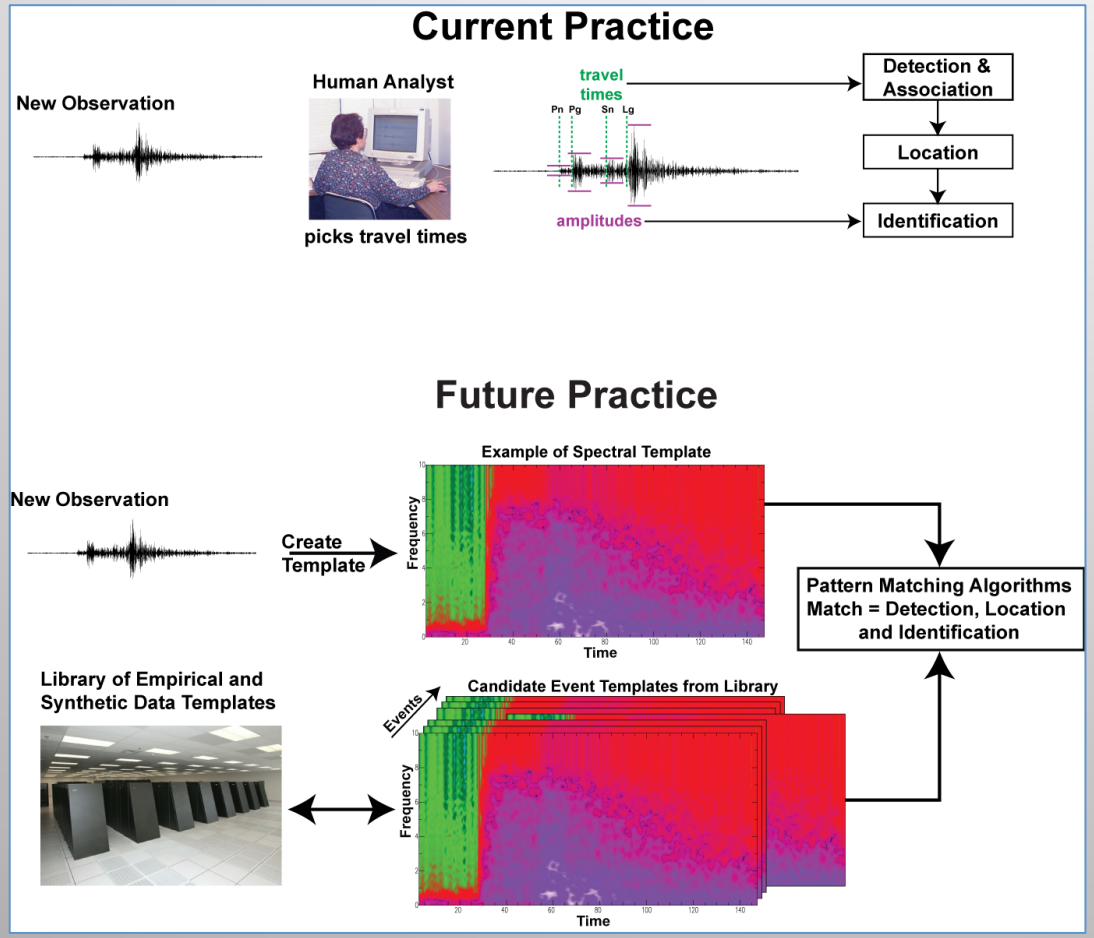


Querying the cumulative graph for concepts linked to "malnutrition"



Causal loop diagrams developed by Georgetown anthropologists

New methods in scalable time series analysis for Global Nuclear Explosion Monitoring – increasing detection performance at increased event rate



Historic focus on large events ~ 100 events / day

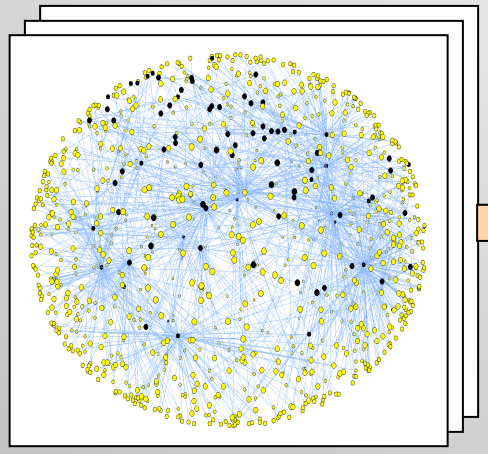
- Runs on small local cluster
- Requires many highly trained human analysts
- Only portions of the signal utilized

Reduced threshold ~ 10,000 events / day

- Large-scale data and HPC capabilities
- Full signal-bandwidth time series analysis
- Increasing analytic complexity: classification and anomaly detection

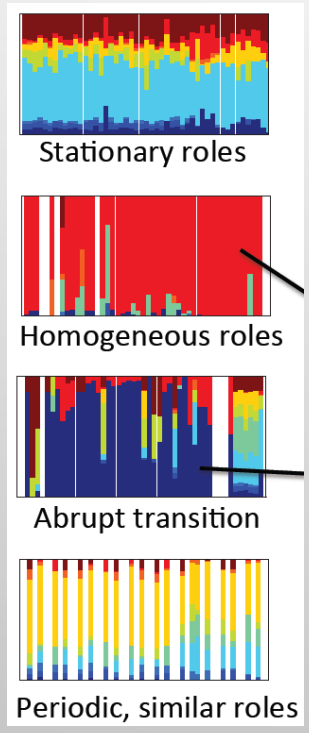
Cyber mapping and activity models for improved behavior prediction and anomaly detection

Dynamic IP-IP graph

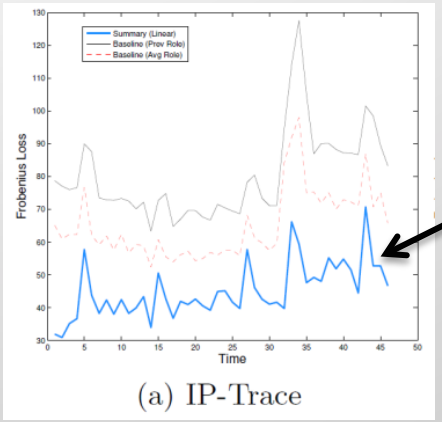


Host roles are local characteristics of the IP-IP graph structure e.g. "center of star", end node, …

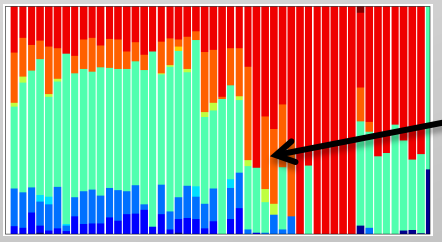
Host role discovery



Learning Markov models for behavior forecasting



Reduced prediction error using host roles

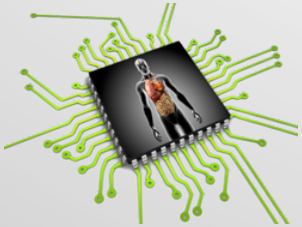


Anomaly Detection in host role distribution

Ryan Rossi, Brian Gallagher, Jennifer Neville, Keith Henderson. Modeling Dynamic Behavior in Large Evolving Graphs. ACM International Conference on Web Search and Data Mining (WSDM), 2013.

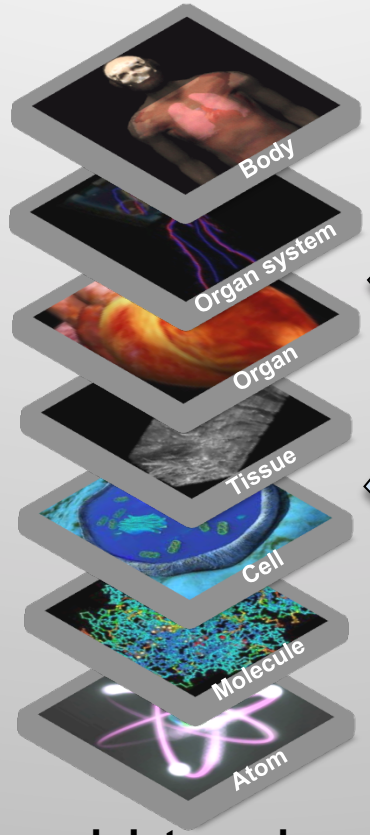
Data-driven multiscale computational models will transform development and delivery for both biosecurity and health care

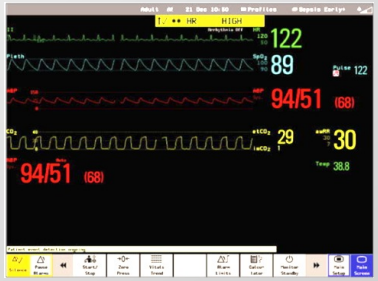
New sensor technologies



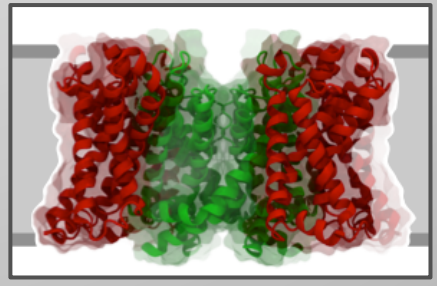
Real-time predictive analytics

High-performance computing





Simulation-based critical care



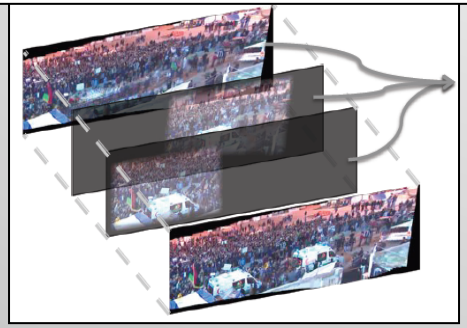
Rapid computational drug design

Individualized models driven by new data, analytics, and computing

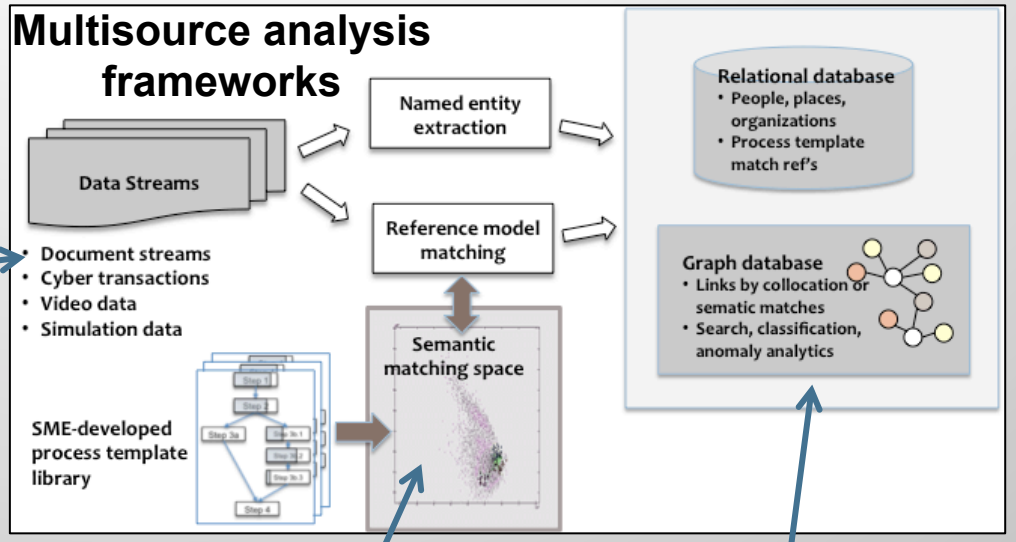
Shared data and model space – ultimately over the population

R&D area 1: Science-based models for pattern discovery

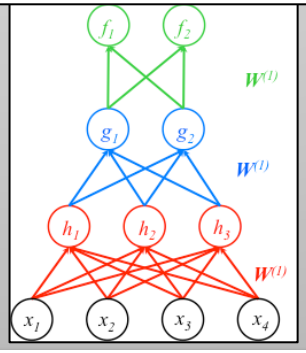
Image, temporal, and audio features in video



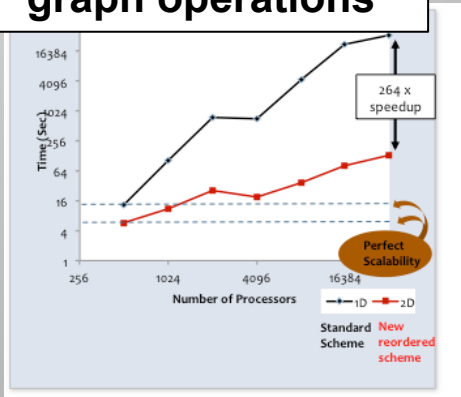
Multisource analysis frameworks



Deep learning neural networks



High-performance graph operations

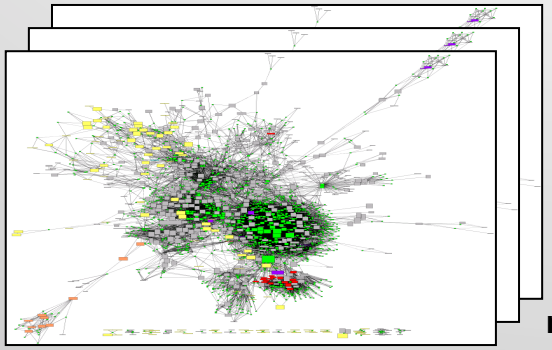


Projects

Continuous Network Cartography – C. Matarazzo
Directed and Hypergraph Analysis – V. Henson
Video Analytics – D. Poland
Coupled segmentation of CT images – T. Bremer
Situational Awareness Tools – D. Buttler
Large-scale Deep Learning– B. Chen

R&D area 2: Simulation as an analytic tool

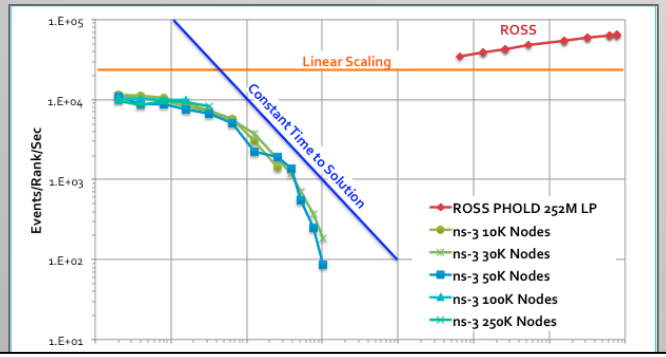
Complex at-scale analytics



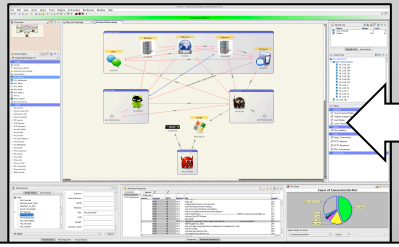
Patterns in the data drive the predictive model

Predictive model drives data and analysis

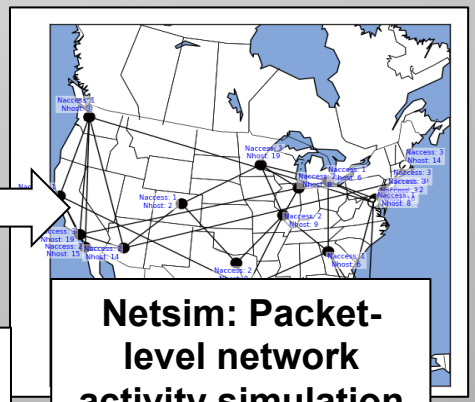
Continuous and discrete simulation



Extreme parallel discrete event simulation P. Barnes, D. Jefferson

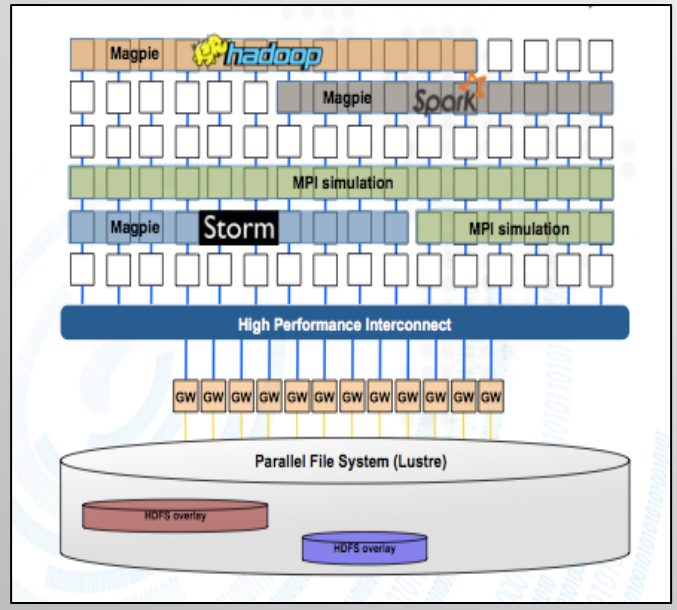


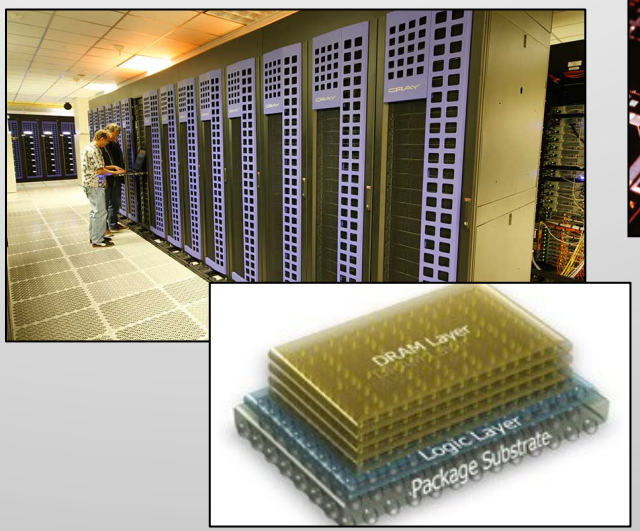
Continuous Network Topography - C. Matarazzo

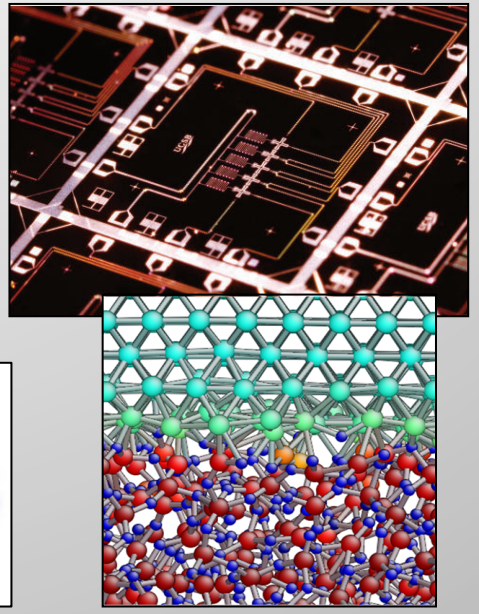


Netsim: Packet-level network activity simulation

R&D area 3: High-performance computing for data analytics







Integrating data-intensive computing into HPC architectures
R. Goldstone

Memory intensive architectures
M. Gokhale

Simulation of quantum computing materials, devices, and systems
V. Lordi

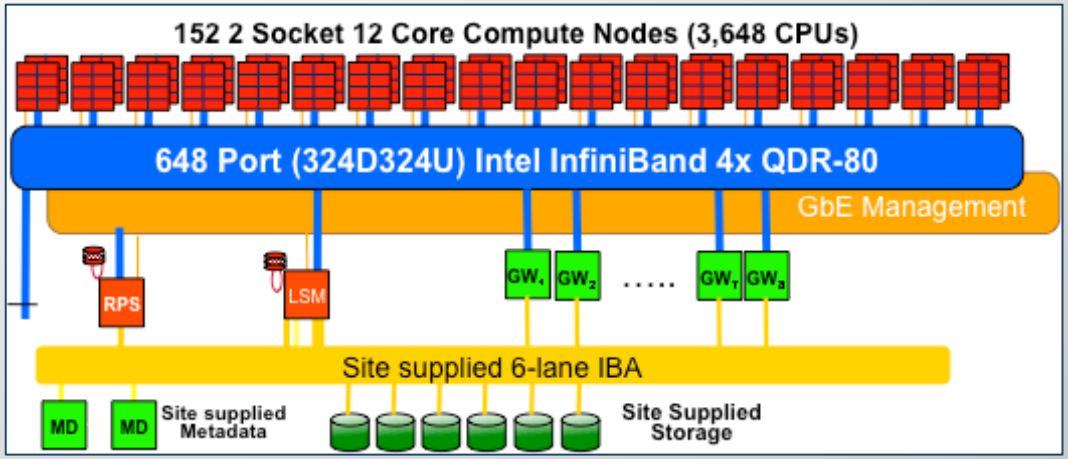
An Intel-Cray collaboration– “Catalyst” for data analytics

Uses the HPCIC partnership model to bring in a cost-shared formidable 150 TF/s data analytics computer

Technical focus on NVRAM layers in memory hierarchy supporting 24 core node – prototyping analytics in new environment

Initial applications will focus on

- **Prototyping exascale simulation analysis architectures**
- **Bioinformatics algorithms**
- **Graph analytics**



324 nodes
7,776 cores
149.3 TF/s, 41.5 TB DRAM, 281.6 TB NVRAM, IB QDR-80

Over 5GB DRAM & 36GB NVRAM per core

Workforce is central - building the next generation of our data science workforce

CyberDefenders – 35 students and 3 faculty in 2013

Contributed to 16 staff, 4 post-doc, 5 transfer hires over two years

Strategy

- Increase student program funding
- Create public face for LLNL data science
- Connect to leading R&D efforts
- University partnerships to establish pipelines
- Provide unique analytic resources
- Provide cross-training resources

Data science tutorial series

- Machine learning
- Neural nets and deep learning
- Analytics in R
- Scala/Spark programming
- Parallel discrete event simulation (broadcast to Sandia)