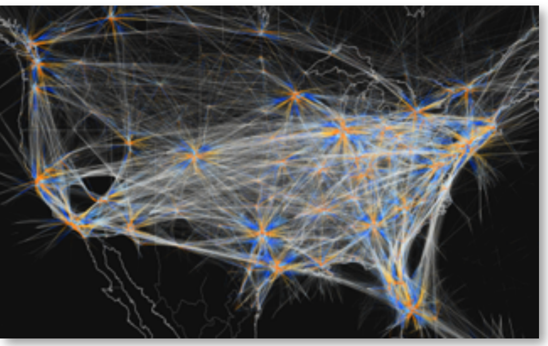
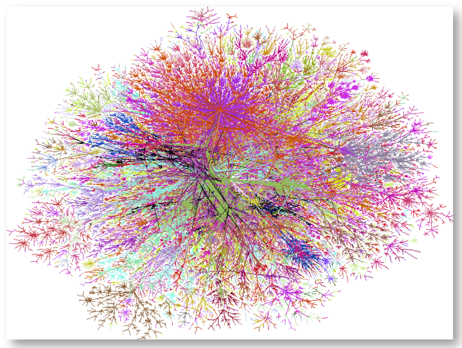
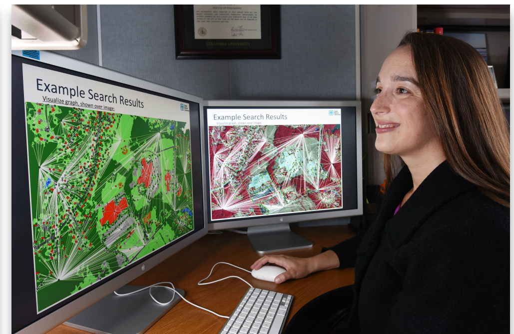
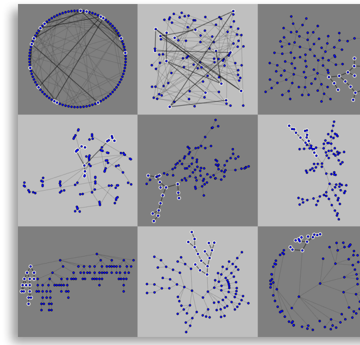
Exceptional service in the national interest

Sandia National Laboratories









Increasing Coherence Between Simulation and Data Analytics

Chesapeake Large Scale Data Analytics Conference
Annapolis, MD
October 25, 2016

Rob Leland
Vice President, Science & Technology
Chief Technology Officer
Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

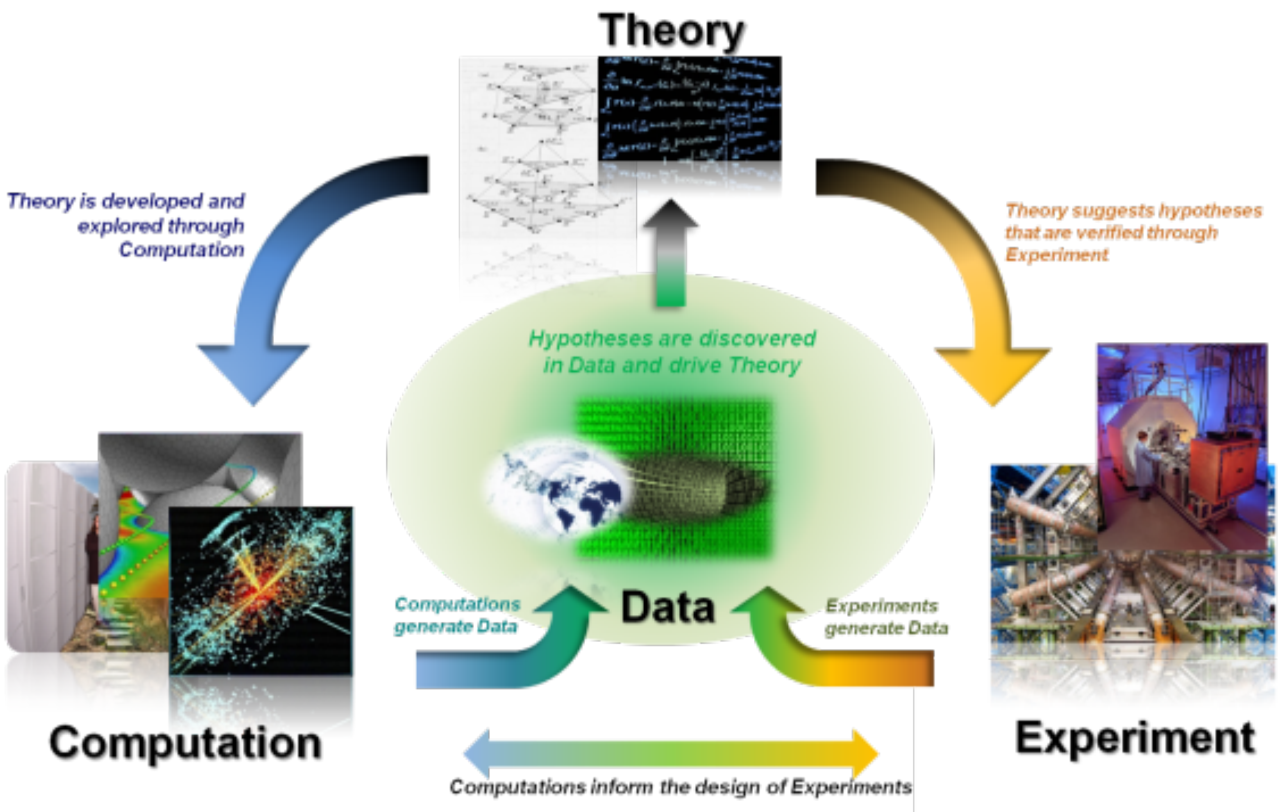
Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

SAND2016-10762 C

Outline

- **A tale of two visions**
- **Some background**
- **A charge from the National Strategic Computing Initiative**
- **Answers to three key questions**
  - Why is an increasing coherence between simulation and analytics important?
  - What is really meant by “increasing coherence” between the two?
  - How might coherence be furthered in practice?
- **A unifying vision**

Vision 1: From a scientific perspective



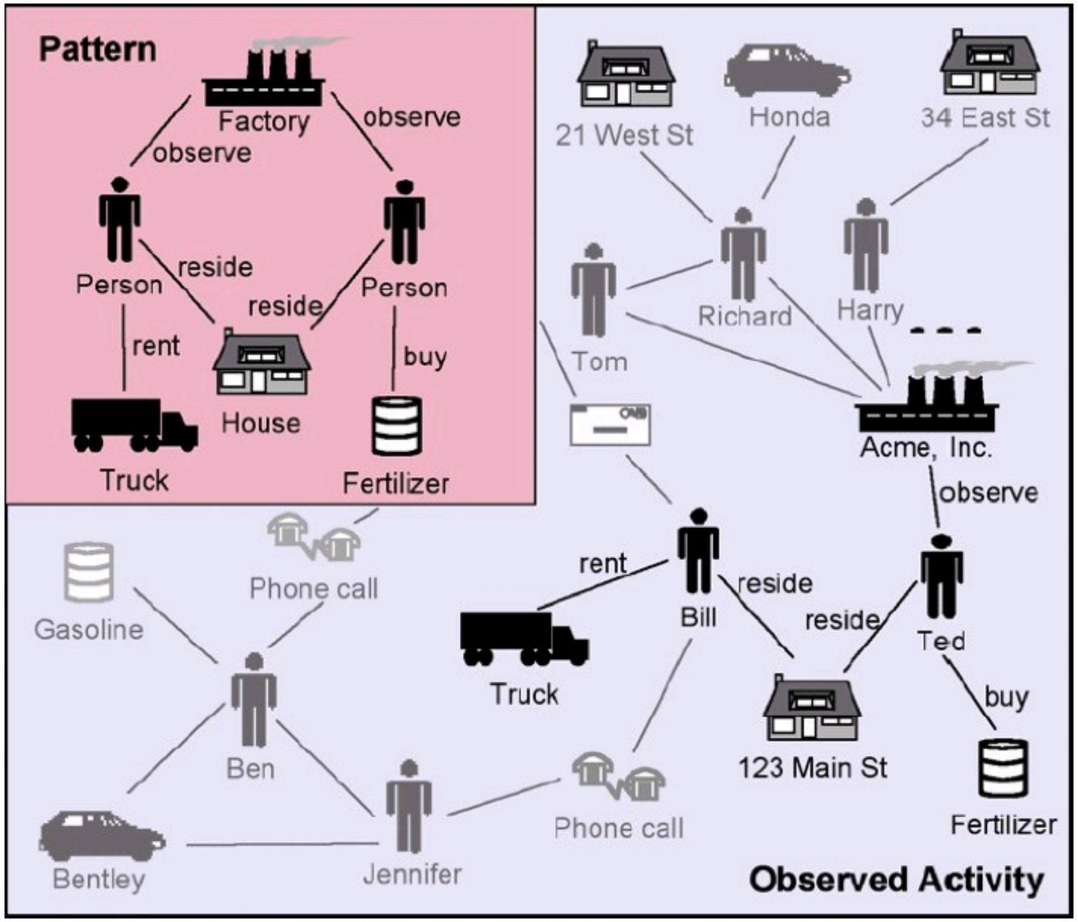
Data analysis complements theory, experiment, and computation

From *The Fourth Paradigm: Data-Intensive Scientific Discovery* by Jim Gray

Vision 2: From a national security perspective

Graph matching example of data analytics

A key analytic primitive -- used to find a specific instance of an abstract pattern of interest

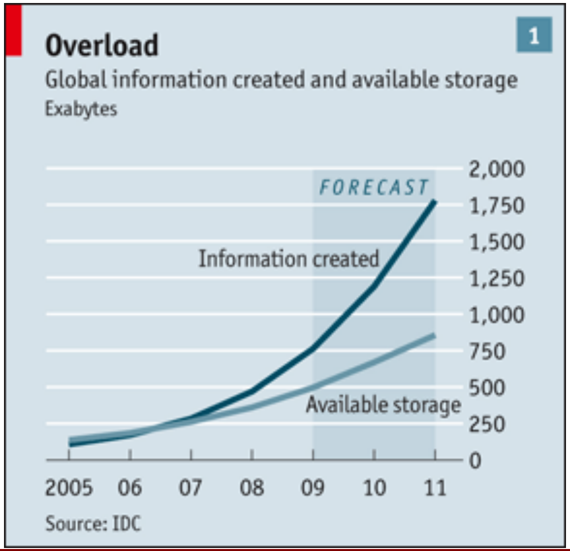


From Coffman, Greenblatt, and Marcus, *Graph-Based Technologies for Intelligence Analysis,* Communications of the ACM, 47, March 2004.

Some background

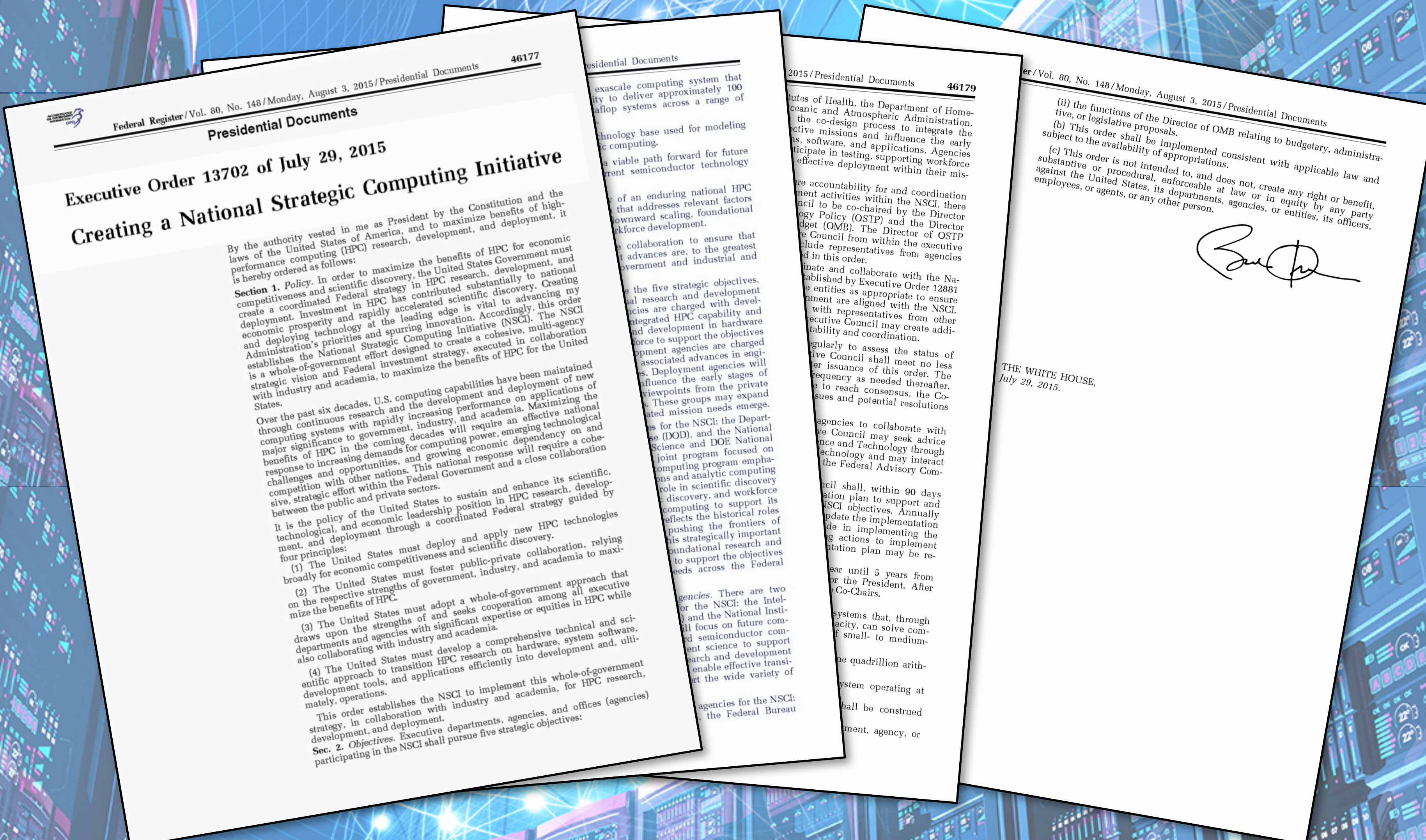
- **Simulation**
  - Computations to understand physical phenomena or conduct engineering
- **Large Scale Data Analytics (LSDA)**
  - Data Analytics = Discovering meaningful patterns in data
  - Large Scale = Requiring leading-edge processing and storage capabilities
- **LSDA is increasing in importance**
  - Pervasive
    - Commerce, finance, health care, science, engineering, national security, ...
  - Lasting societal significance
    - Internet search, genomics, climate modeling, Higgs particle, ...
- **LSDA is getting "harder"**
  - Captured data growing exponentially with time
  - Individual analysis becoming more sophisticated
  - More people examining more data more frequently
  - Aggregate work growing much faster than Moore's Law

The Economist:



National Strategic Computing Initiative (NSCI)

Sandia National Laboratories



Federal Register / Vol. 80, No. 148 / Monday, August 3, 2015 / Presidential Documents 46177

Presidential Documents

Executive Order 13702 of July 29, 2015

Creating a National Strategic Computing Initiative

By the authority vested in me as President by the Constitution and the laws of the United States of America, and to maximize benefits of high-performance computing (HPC) research, development, and deployment, it is hereby ordered as follows:

Section 1. *Policy*. In order to maximize the benefits of HPC for economic competitiveness and scientific discovery, the United States Government must create a coordinated Federal strategy in HPC research, development, and deployment. Investment in HPC has contributed substantially to national economic prosperity and rapidly accelerated scientific discovery. Creating and deploying technology at the leading edge is vital to advancing my Administration's priorities and spurring innovation. Accordingly, this order establishes the National Strategic Computing Initiative (NSCI). The NSCI is a whole-of-government effort designed to create a cohesive, multi-agency strategic vision and Federal investment strategy, executed in collaboration with industry and academia, to maximize the benefits of HPC for the United States.

Over the past six decades, U.S. computing capabilities have been maintained through continuous research and the development and deployment of new computing systems with rapidly increasing performance on applications of major significance to government, industry, and academia. Maximizing the benefits of HPC in the coming decades will require an effective national response to increasing demands for computing power, emerging technological challenges and opportunities, and growing economic dependency on and competition with other nations. This national response will require a cohesive, strategic effort within the Federal Government and a close collaboration between the public and private sectors.

It is the policy of the United States to sustain and enhance its scientific, technological, and economic leadership position in HPC research, development, and deployment through a coordinated Federal strategy guided by four principles:

(1) The United States must deploy and apply new HPC technologies broadly for economic competitiveness and scientific discovery.

(2) The United States must foster public-private collaboration, relying on the respective strengths of government, industry, and academia to maximize the benefits of HPC.

(3) The United States must adopt a whole-of-government approach that draws upon the strengths of and seeks cooperation among all executive departments and agencies with significant expertise or equities in HPC while also collaborating with industry and academia.

(4) The United States must develop a comprehensive technical and scientific approach to transition HPC research on hardware, system software, development tools, and applications efficiently into development and, ultimately, operations.

This order establishes the NSCI to implement this whole-of-government strategy, in collaboration with industry and academia, for HPC research, development, and deployment.

Sec. 2. *Objectives*. Executive departments, agencies, and offices (agencies) participating in the NSCI shall pursue five strategic objectives:

Federal Register / Vol. 80, No. 148 / Monday, August 3, 2015 / Presidential Documents

(ii) the functions of the Director of OMB relating to budgetary, administrative, or legislative proposals.

(b) This order shall be implemented consistent with applicable law and subject to the availability of appropriations.

(c) This order is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

THE WHITE HOUSE,
July 29, 2015.

NSCI Strategic Objectives

- (1) Accelerating delivery of a capable exascale computing system that integrates hardware and software capability to deliver approximately 100 times the performance of current 10 petaflop systems across a range of applications representing government needs.
- **(2) Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing.**
- (3) Establishing, over the next 15 years, a viable path forward for future HPC systems even after the limits of current semiconductor technology are reached (the "post-Moore's Law era").
- (4) Increasing the capacity and capability of an enduring national HPC ecosystem by employing a holistic approach that addresses relevant factors such as networking technology, workflow, downward scaling, foundational algorithms and software, accessibility, and workforce development.
- (5) Developing an enduring public-private collaboration to ensure that the benefits of the research and development advances are, to the greatest extent, shared between the United States Government and industrial and academic sectors.

Q1: Why is increasing coherence between simulation and analytics important?

- **For simulation**
  - HPC simulation must ride on some commodity curve
  - Larger market forces behind analytics
  - Can exploit commodity component technology from analytics
- **For analytics**
  - Large Scale Data Analytics problems becoming ever more sophisticated
  - Requiring more coupled methods
  - Can exploit architectural lessons from HPC simulation
- **For both: Integration of simulation and analytics in the same workflow**
  - Automation of analysis of data from simulation
  - Creation of synthetic data via simulation to augment analysis
  - Automated generation and testing of hypothesis
  - Exploration of new scientific and technical scenarios
  - ...

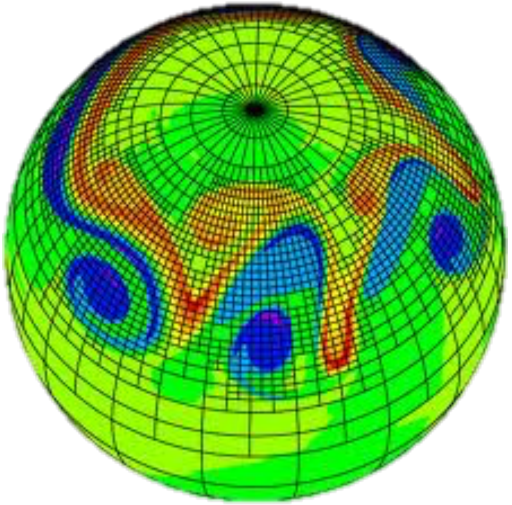
Mutual inspiration, technical synergy, and economies of scale in the creation, deployment, and use of HPC resources

A challenge because simulation and analytics differ in many respects …

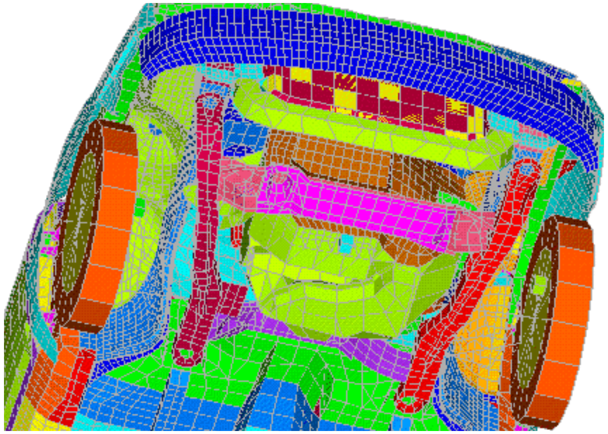
Data structures describing simulation and analytics differ

Graphs from simulations may be irregular, but have more locality than those derived from analytics

Computational Simulation of physical phenomena:

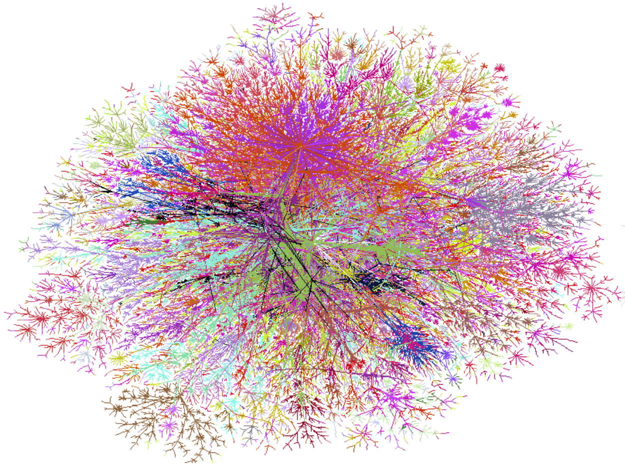


Climate modeling

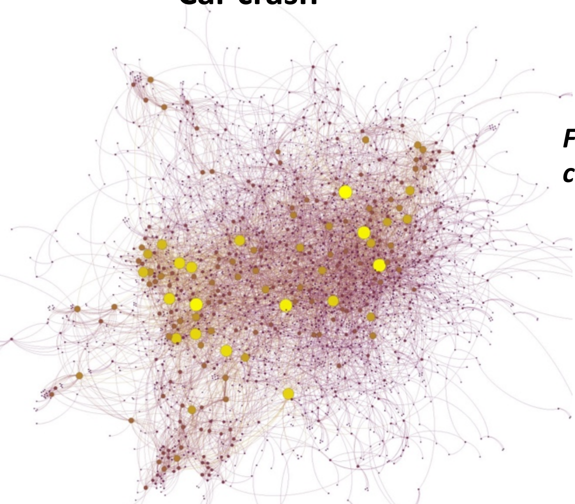


Car crash

Large Scale Data Analytics:



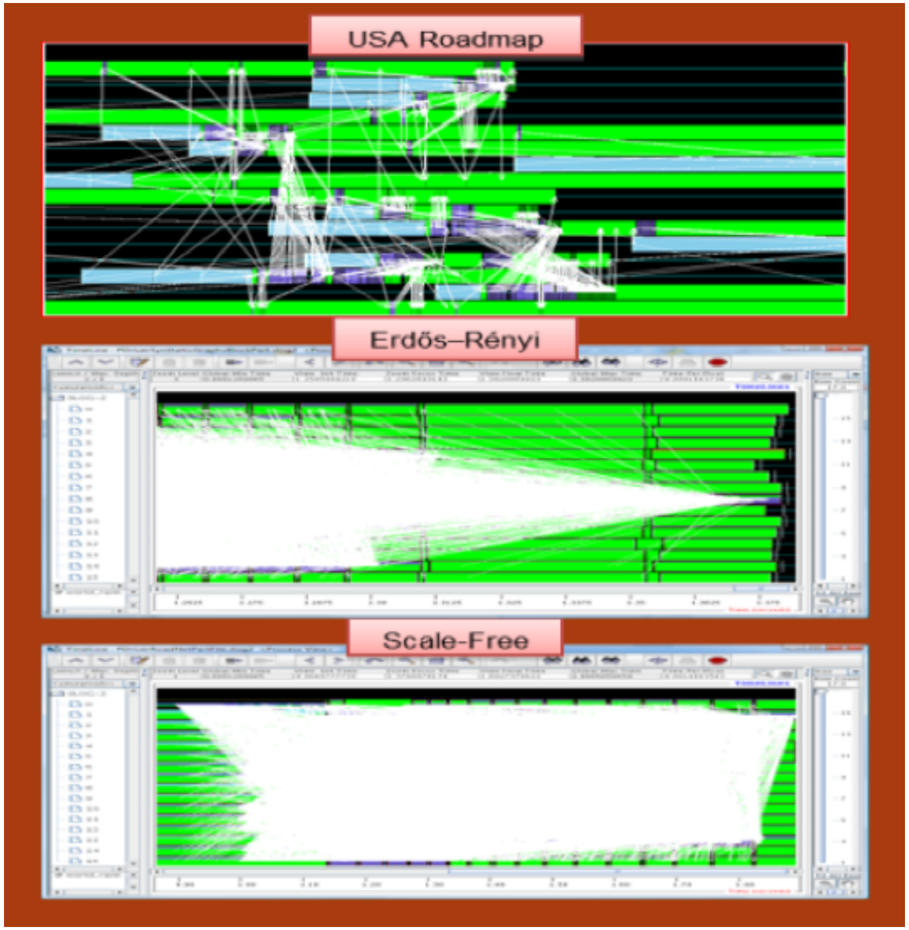
Internet connectivity



Yeast protein interactions

Figures from Leland et. al. courtesy of Yelick, LBNL.

Computation and communication patterns differ



Black = time spent computing
Green = time spent communicating
White = time spent waiting for data to be communicated

The U.S. roadmap, which has spatial locality and is thus most similar of the three in structure to computational patterns that would arise in typical physical simulations.

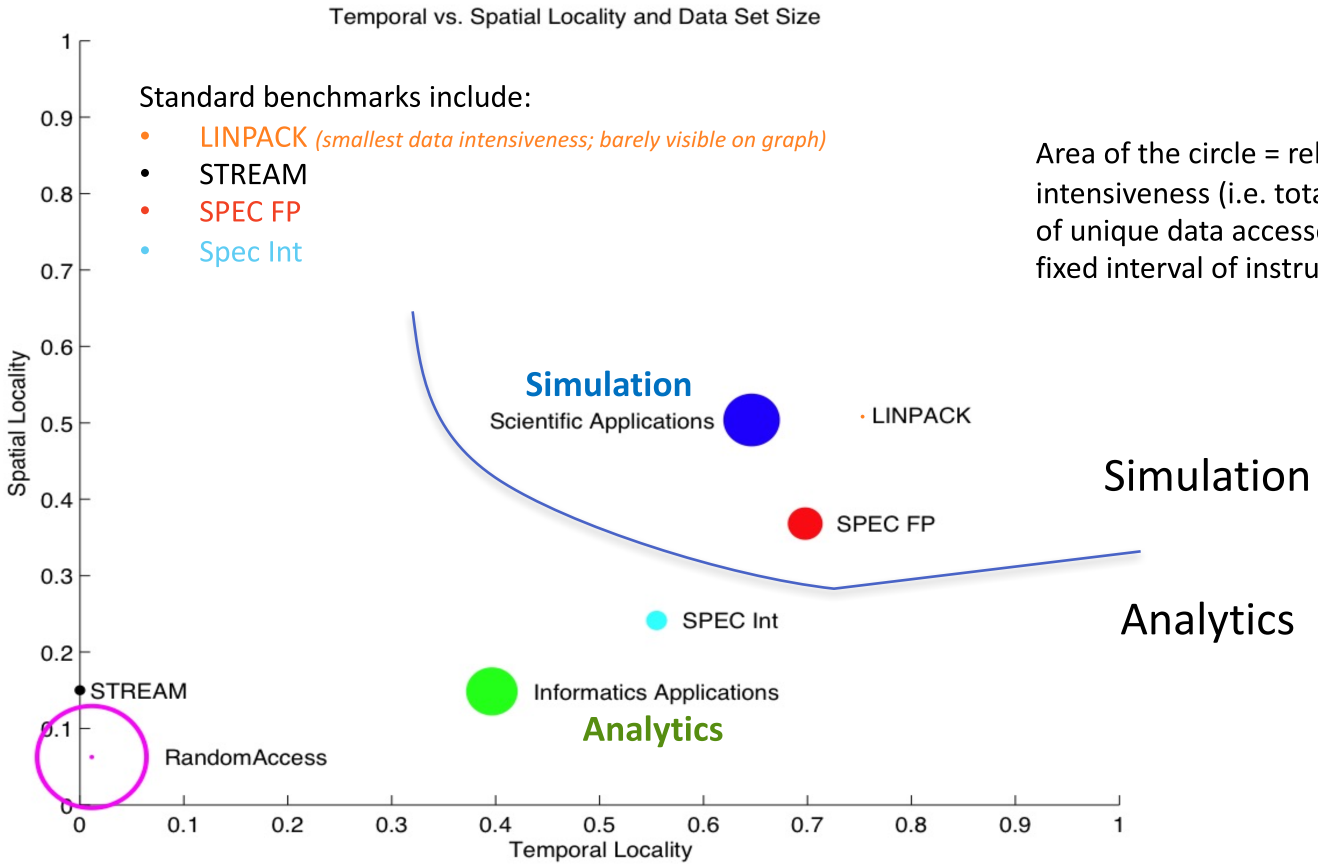
The *Erdős-Rényi* graph, a well-studied example in graph theory work.

A scale-free graph, an example more reflective of real-world networks.

Figure from Leland et. al. courtesy of Johnson, PNNL.

Memory performance demands differ

A key differentiator in the performance of simulation and analytics



Area of the circle = relative data intensiveness (i.e. total amount of unique data accessed over a fixed interval of instructions)

Figure from Murphy & Kogge with adjustment to double radius of Linpack data point to make it visible.

Application code characteristics differ

Sandia National Laboratories

Contrasting properties:

Application code property	Simulation	Analytics
Spatial locality	High	Low
Temporal locality	Moderate	Low
Memory footprint	Moderate	High
Computation type	May be floating-point dominated*	Integer intensive
Input-output orientation	Output dominated	Input dominated

* Increasingly, simulation work has become less floating-point dominated

Q2: So what do we really mean by "increasing coherence" between simulation and analytics?

- **NOT one system ostensibly optimized for both simulation and analytics**
- **Greater commonality in underlying componentry and design principles**
- **Greater interoperability, allowing interleaving of both types of computations**

… A more common hardware and software roadmap between simulation and analytics

And yet, there is hope …

Simulation and analytics are evolving to become more similar in their architectural needs

Sandia National Laboratories

- **Current challenges for the LSDA community**
  - Data movement
  - Power consumption
  - Memory/interconnect bandwidth
  - Scaling efficiency

… similar to HPC simulation

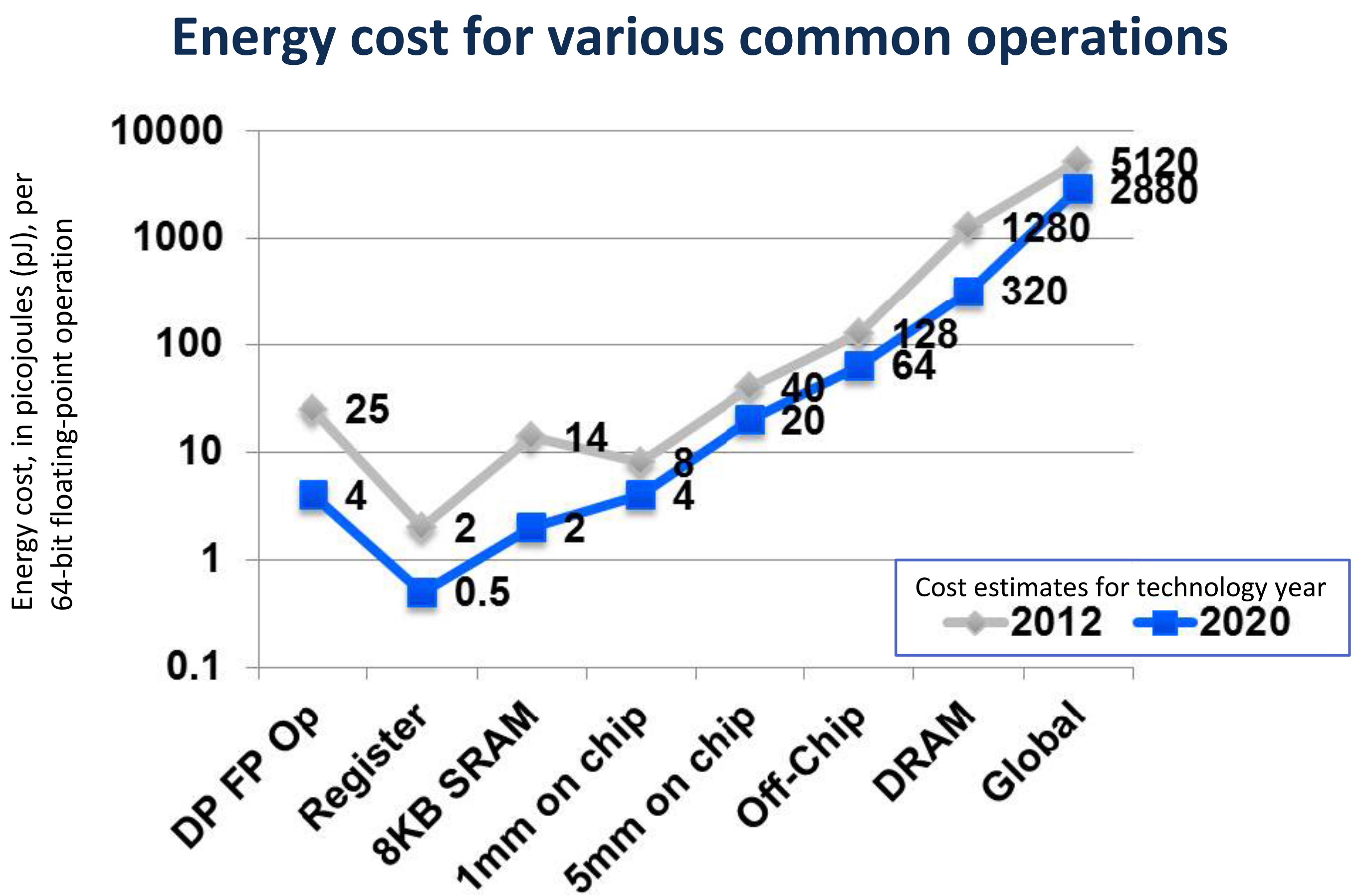
- **Instruction mix for Sandia's HPC engineering codes**
  - Memory operations 40%
  - Integer operations 40%
  - Floating point 10%
  - Other 10%

… similar to LSDA

- **Common design impacts of energy cost trends**
  - Increased concurrency (processing threads, cores, memory depth)
  - Increased complexity and burden on
    - system software, languages, tools, runtime support, codes

Energy cost of moving data is becoming dominant

Sandia National Laboratories



From Dan McMorrow, *Technical Challenges of Exascale Computing*, JSR-12-310, JASON, MITRE Corporation, April 2013.

Emerging architectural and system software synergies

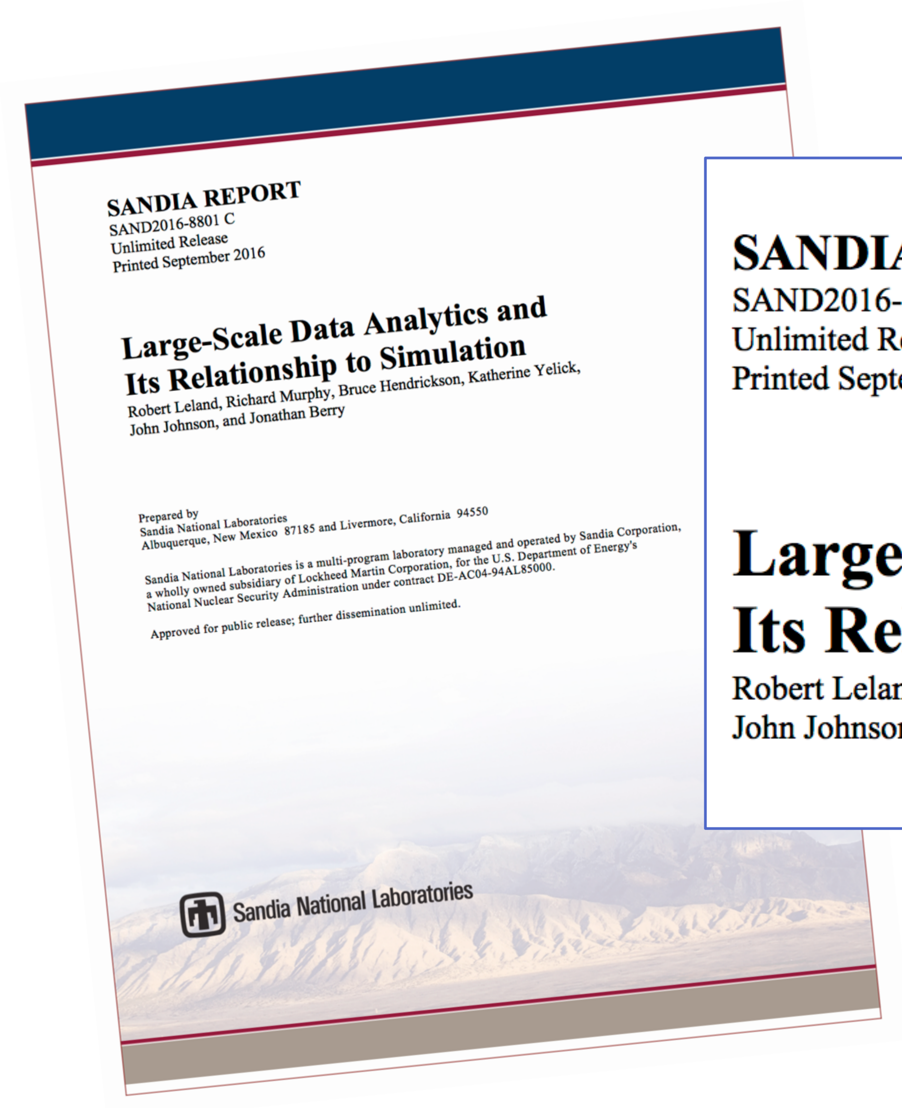
Similar needs:

Architectural Characteristic	Simulation	Analytics
Computation	Memory address generation dominated	Same
Primary memory	Low power, high bandwidth, semi-random access	Same
Secondary memory	Emerging technologies may offset cost, allowing much more memory	… require extremely large memory spaces
Storage	Integration of another layer of memory hierarchy to support checkpoint/restart	… to support out-of-core data set access
Interconnect technology	High bisection bandwidth, (for relatively coarse-grained access)	… (for fine-grained access)
System software (node-level)	Low dependence on system services, increasingly adaptive, resource management for structured parallelism	… highly adaptive, resource management for unstructured parallelism
System software (system-level)	Increasingly irregular workflows	Irregular workflows

Q3: How might coherence be furthered in practice?

- **Making it an element of national strategy**
  - Check via the NSCI
- **Building this in to exascale computing efforts**
  - Also a component of the NSCI
- **Communicating with and enlisting the technical communities concerned**
  - This forum and similar events
- **Further developing the vision**
  - Today’s dialogue session!

Acknowledgements



SANDIA REPORT
SAND2016-8801 C
Unlimited Release
Printed September 2016

Large-Scale Data Analytics and Its Relationship to Simulation

Robert Leland, Richard Murphy, Bruce Hendrickson, Katherine Yelick, John Johnson, and Jonathan Berry

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

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.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

SANDIA REPORT
SAND2016-8801 C
Unlimited Release
Printed September 2016

Large-Scale Data Analytics and Its Relationship to Simulation

Robert Leland, Richard Murphy, Bruce Hendrickson, Katherine Yelick, John Johnson, and Jonathan Berry

Additional references

- *The Economist,* "Data, Data, Everywhere," Feb 25th, 2010
- R. C. Murphy and P. M. Kogge, "On the Memory Access Patterns of Supercomputer Applications: Benchmark Selection and Its Implications," IEEE Transactions on Computers 56 (7, July 2007): 937–945.
- R. Murphy, "Power Issues," presentation to JASON 2012, June 2012.
- Peter Kogge (editor) et al., ExaScale Computing Study: *Technology Challenges in Achieving Exascale Systems.* DARPA, 2008.
- Dan McMorrow, *Technical Challenges of Exascale Computing*, JSR-12-310, JASON, MITRE Corporation, April 2013.
- Tony Hey, Stewart Tansley, and Kristin Tolle (editors), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.
- Jim Gray, *The Fourth Paradigm: Data-Intensive Scientific Discovery*