Ontology Summit 2012

Track 3 Summary Report
Challenge: Ontology and Big Data

Co-Chairs

Mary Brady

Ernie Lucier

Thursday, April 12, 2012

Track 3 – Challenge
Ontology and Big Data

Mission:

- Identify appropriate objectives for an Ontology and Big Data challenge
- Prepare problem statements, identify the organizations and people to be advocates, and identify the resources necessary to complete a challenge

Engage the community in designing ontology solutions to benefit BIG DATA applications

Track 3 Challenge
Ontology and Big Data

<u>Goal</u>:

Meet Big Data Challenges via Ontology

- Advance ontology and semantic web technologies
- Identify challenges that will increase applications and accelerate adoption

Session 1: Panelists

Presenter	Organization	Topic
Dr. Barry Smith	University of Buffalo, SUNY	How BIG DATA might benefit from Ontology and why it usually fails
Chris Musialek (for Dr. Jeanne Holm)	Data.gov	Data.gov datasets (>400,000) that could benefit from ontology
Bryan Thompson, Mike Personick	SYSTAP, LLC	Managing scale in ontological systems
James Kirby	Naval Research Lab	Ontology for Software Production

Session 2: Panelists

Panelist	Organization	Topic
Dr. Tim Finin, Dr. Anupam Yoshi	UMBC	Making the Semantic Web Easier to Use
Kyoung-Sook Kim	NICT	Use Cases of Cyber Physical Data Cloud
Mike Folk		HDF5
Mario Paolucci	FuturICT	Global Participatory Computing for Our Complex World
Dr. Ursula R. Kattner	NIST	Materials Genome: Data Standards
Edin Muharemagic	LexisNexis	HPCC, Machine Learning

Current State (Ontology)

- Ontology may tame big data, drive innovation, facilitate the rapid exploitation of information, contribute to long-lived and sustainable software, and improve Complicated Systems Modeling.
- Ontology promises to:
  - Achieve global data standards, meanings, knowledge representation
  - Reduce complexity and costs
  - Improve agility
  - Allow reasoning and inferencing capabilities
- But, there is a growing ontology base to choose from…without much regard for standardization.
- Recommendation: Develop ontologies in the same field in a coordinated fashion to ensure that there is exactly one ontology for each subdomain, e.g., the Gene Ontology

Current State (BIG DATA)

- BIG DATA – Data Drives Decisions
  - Commerce, Financial, and Homeland Security success stories in mining BIG DATA.
    - Amazon => suggest possible purchases
    - Credit card companies differentiate between fraudulent and legitimate purchases
    - Financial Analysts predict investments
    - Homeland Security monitors is constantly analyzing purchases to predict individual's future buying habits
  - BIG DATA environments vary => Google Map/Reduce, HADOOP, LexisNexis HPCC, machine learning, appearance of higher-order languages
  - Important to consider entire "big data stack" and consider use of ontology at multiple levels (storage, feature identification and correlation, large-scale data integration, etc.)
  - Large-scale, national priority applications could learn from these applications areas; all could benefit from integrated ontology and machine learning approaches to provide global standards, meaning, knowledge representation

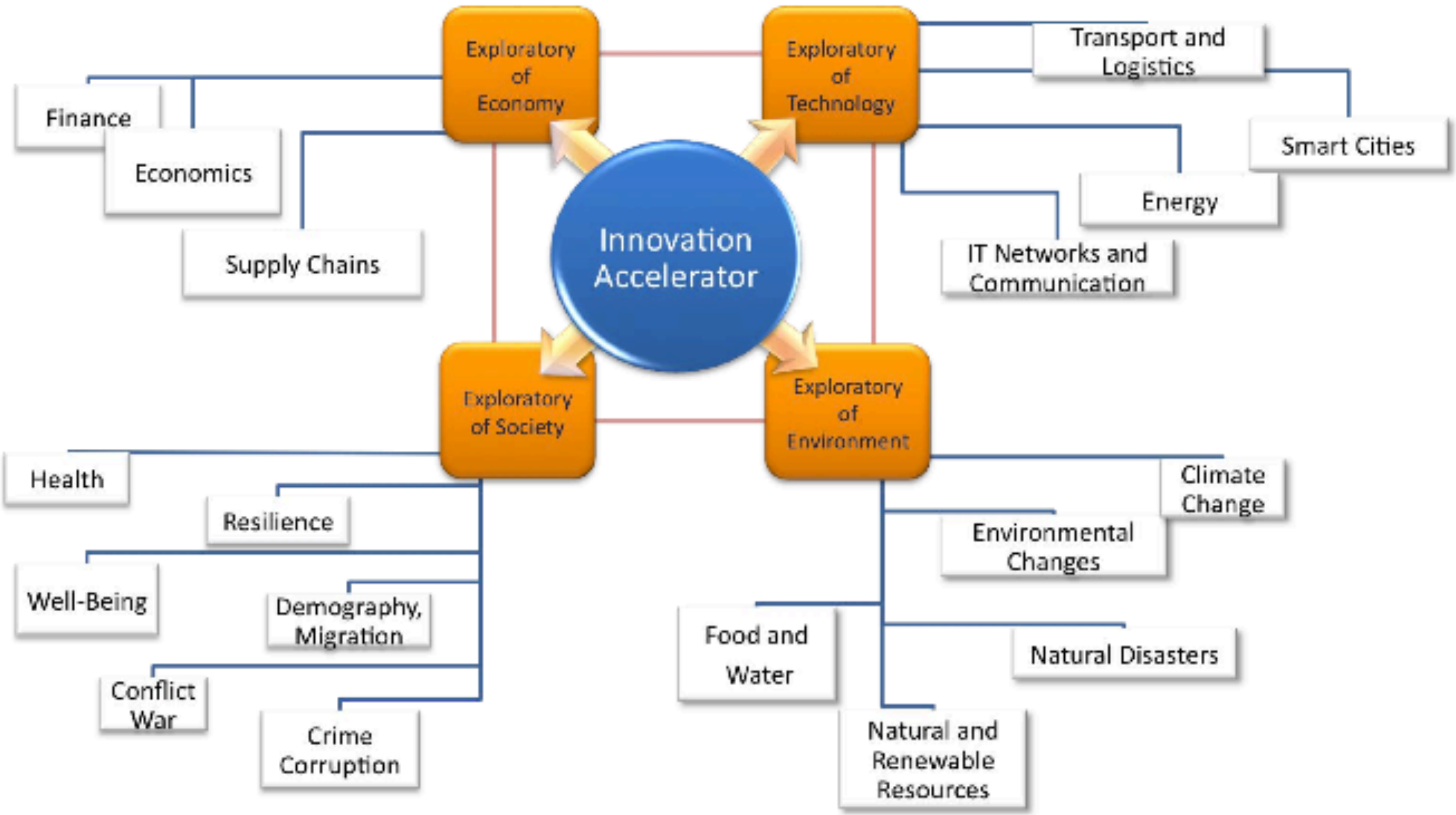
BIG DATA Applications

- DATA.GOV
- FuturICT
- Materials Genome: Data Standards
- Cyber-Physical Systems

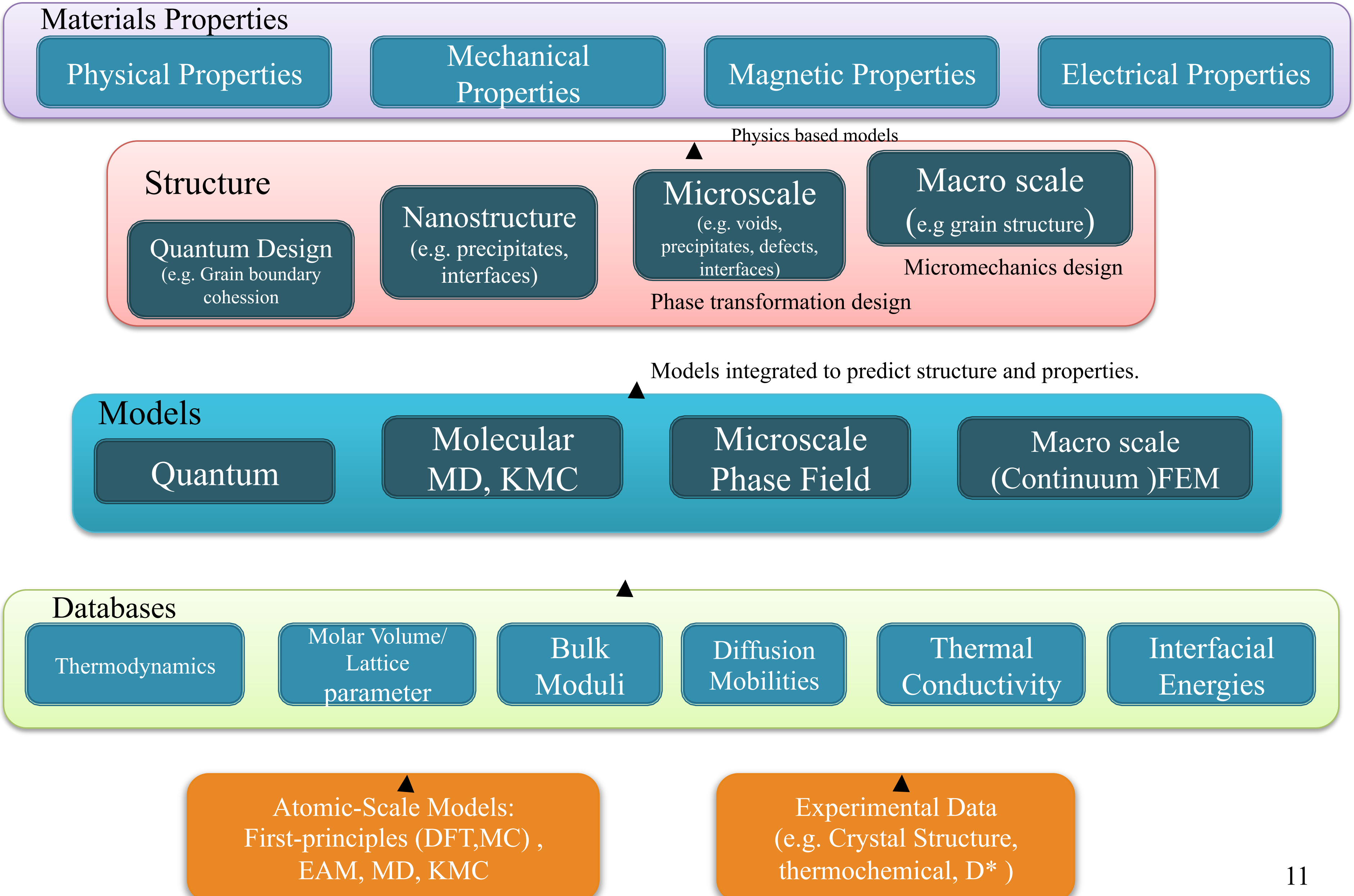
Data.gov

- Data.gov is **Driving Innovation by Creating a Data Ecosystem**
  - **Gather data** from many places and give it freely to developers, scientists, and citizens
    - Bring data up and out of government to the public
    - Make data accessible and linked
  - **Connect the community by** finding solutions to allow collaboration through social media, events, platforms
    - Create communities to understand and apply data
  - **Provide an infrastructure** built on standards
  - **Encourage technology developers** to create apps, maps, and visualizations of data that empower people's choices
    - Provide simple ways to visualize the data
    - Connect and collaborate with small businesses, industry, and academia to drive innovation
  - **Gather more data** and connect more people

FuturICT Observatories and Exploratories

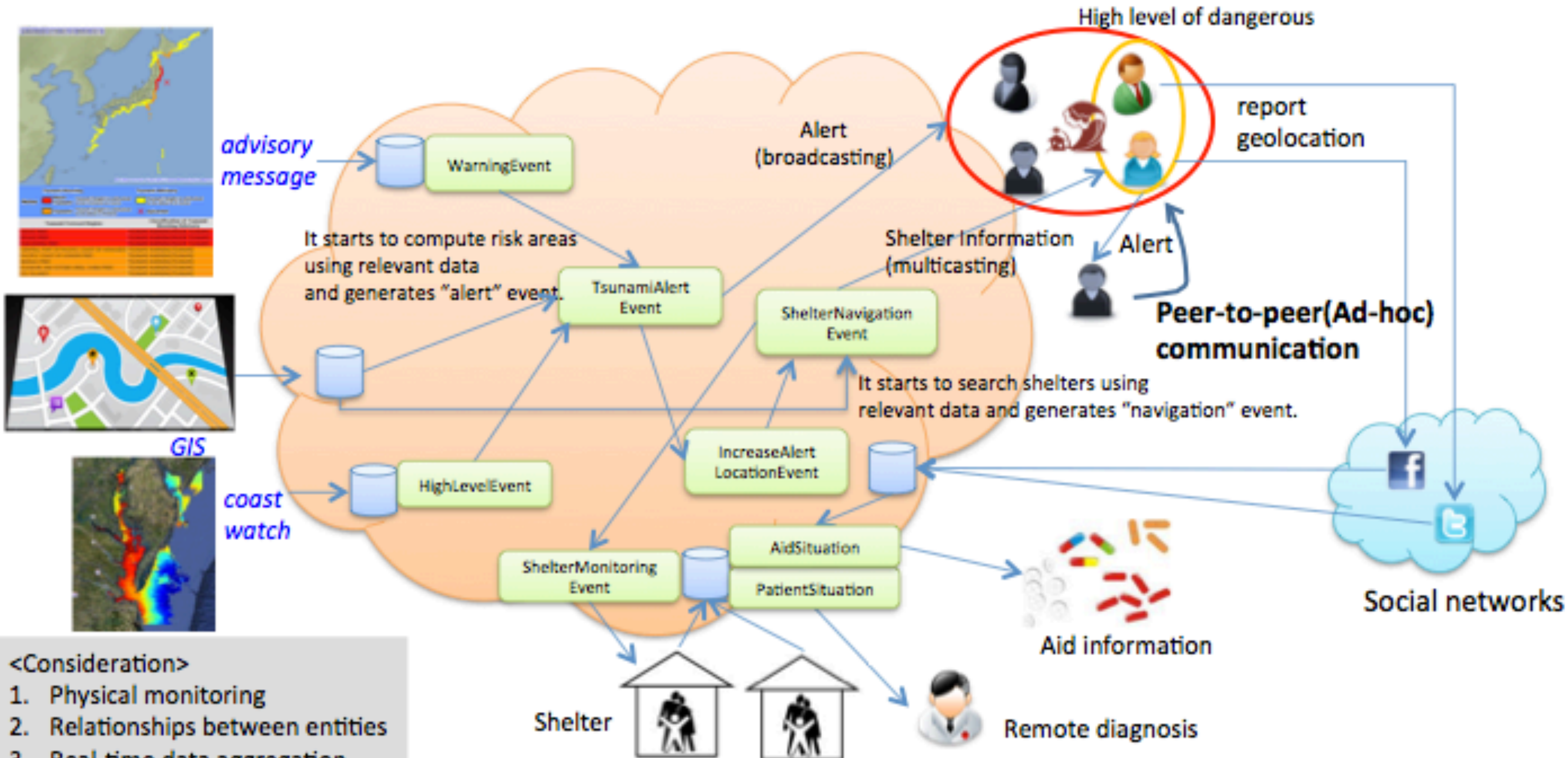


Materials Genome Initiative



Cyber-Physical and the Cloud

- Globally monitoring and locally fencing (safe and rapid evacuation)



<Consideration>
1. Physical monitoring
2. Relationships between entities
3. Real-time data aggregation
4. Situation analysis
5. ...

*Japan experience: the shelter assessment system was up and running **two weeks** after the disaster.

BIG DATA Challenge: Considerations

- Ontology has great promise for BIG DATA, but must have concerted standard efforts, similar to Gene Ontology, to be successful at a large scale.

- Promising technology at each layer that should be considered for ontology use – storage, domain ontology, linked data, integration between domains, etc.

- Methods to build on existing infrastructure rather than re-vamping?

- Methods to address learning curve:
  - Education of future ontologists – topic of last year's summit
  - What can we learn from other efforts?
    - Security, sysadm – over time, moved from system internals => certificate programs
    - BIG DATA platforms – emphasis on creating high-order languages that remove complexity of underlying hw/sw stack from user
    - Similar paradigm for ontologists?

BIG DATA Challenge: Goals

- Increase:
  - Awareness of ontology technology among programmers/database managers
  - Number of qualified personnel to facilitate the growth of the ontology technologies
- Accelerate agencies' adoption of semantic and ontology capabilities through improved implementation methodologies
- Create a cross-culture of domain scientists, engineers, computer scientists, solution providers to:
  - Ameliorate any mismatch between those with data and those with the skills to analyze it
  - Enable scientists and engineers to make maximum use of big data
  - Enable scientists and engineers to understand the potential of ontology-based systems integration
  - Enable ontologists to understand scientists and engineers needs

Big Data Challenge: Basic Principles

- Heterogeneous collections of data to become more homogeneous and searchable "on the fly" or "at first presentation"
- Involves more than one agency (could specify the agencies) and the resulting application/tool could be easily generalized for use by multiple agencies.
- Incorporates agency mission statements
- Involves more than one data set, of which:
  - At least one must be a "big data" data set (as defined… see data set summary)
  - At least one must be an active or streaming data set (this could be a requirement, or an option)
- Promotes Data to Knowledge to Action