The Emerging Data Science Landscape

Chaitan Baru
Senior Advisor for Data Science, CISE
National Science Foundation
cbaru@nsf.gov

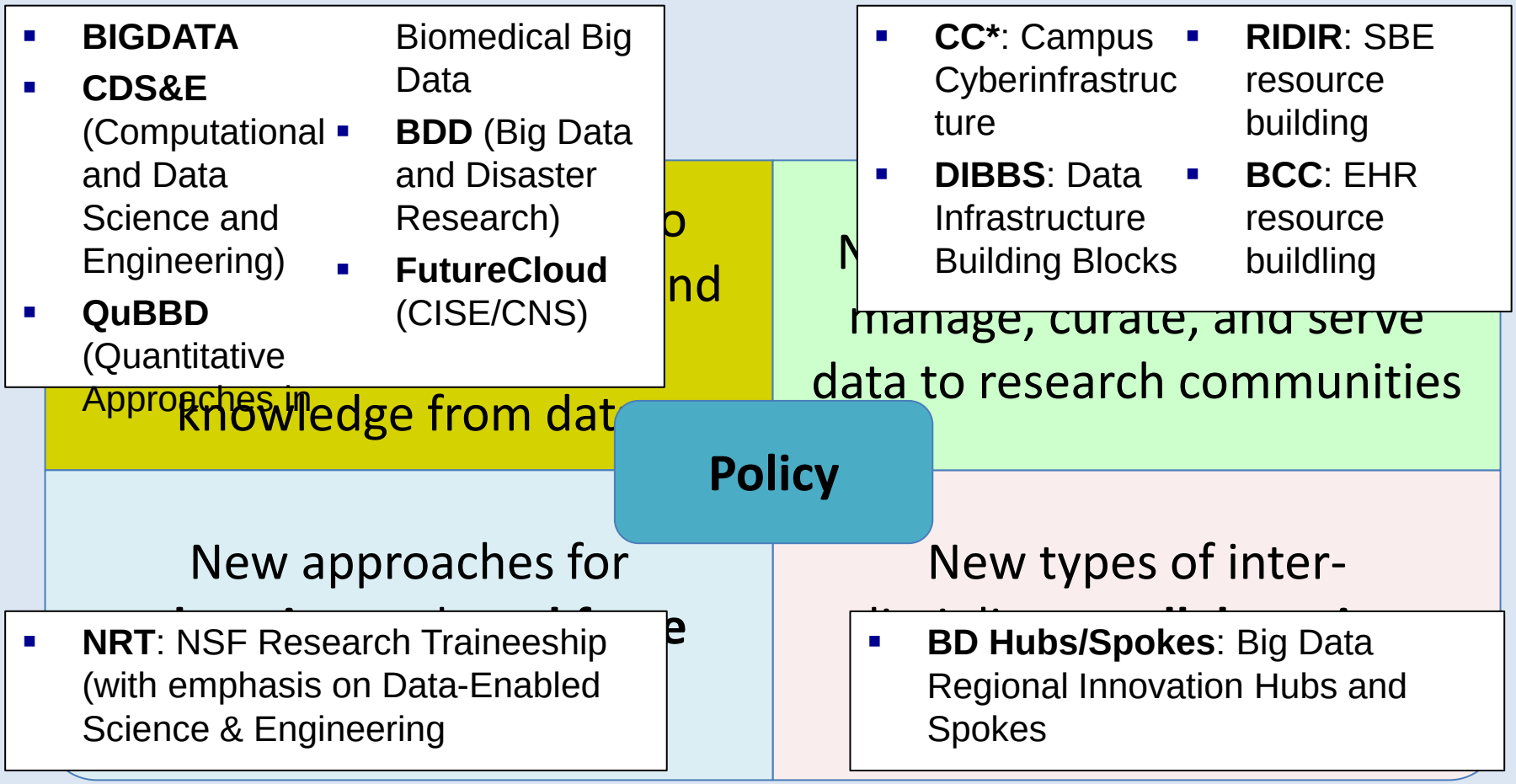
The Emerging Landscape

- Data Science and the “domains”
  - Natural Sciences
  - Social Sciences
  - Engineering
  - Medicine
  - Finance
- More data → Data-intensive science
  - The *Fourth Paradigm*
    - Hypothesis generation, discovery, broad inferences from data…within one dataset, one experiment
- “Data-by-design” vs Data “Reuse”
  - Cross-discipline, heterogeneous data integration

Data Science (and Engineering) as a discipline

- Principles and systematic approaches for the full data lifecycle
  - Collection, cleaning, metadata, management, curation, use, analysis, preservation, …
- Use of statistical and machine learning techniques
- Combining CS and Stats
  - Training in Stats, incorporation of statistical methods and approaches in CS techniques, technologies
  - Incorporation of computational notions, computational complexity in Stats
- Ethical considerations

NSF's Big Data / Data Science Programs



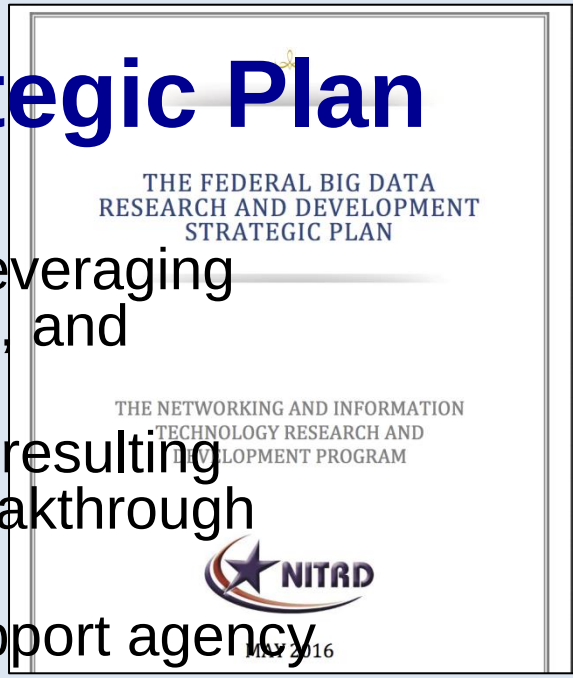
- **BIGDATA**
- **CDS&E** (Computational and Data Science and Engineering)
- **QuBBD** (Quantitative Approaches in Biomedical Big Data
- **BDD** (Big Data and Disaster Research)
- **FutureCloud** (CISE/CNS)

- **CC***: Campus Cyberinfrastructure
- **DIBBS**: Data Infrastructure Building Blocks
- **RIDIR**: SBE resource building
- **BCC**: EHR resource buildling

- **NRT**: NSF Research Traineeship (with emphasis on Data-Enabled Science & Engineering

- **BD Hubs/Spokes**: Big Data Regional Innovation Hubs and Spokes

Federal Big Data R&D Strategic Plan



1. **Create next-generation capabilities**: by leveraging emerging Big Data foundations, techniques, and technologies
2. **Understand trustworthiness of data** and resulting knowledge, for better decisions, enable breakthrough discoveries, and take confident action
3. **Build Big Data cyberinfrastructure** to support agency missions and innovation
4. Increase the value of data through **sustainable preservation** and the sharing of infrastructure and policies
5. Understand **Privacy, Security, and Ethics** in Big Data collection, sharing, and use
6. Improve the national landscape for Big Data **education and training**
7. **Create and enhance connections** in the national Big Data innovation ecosystem

NSF "Big Ideas"

RESEARCH IDEAS

- Harnessing Data for 21st Century Science and Engineering
- Shaping the new Human – Technology Frontier
- Understanding the Rules of Life: Predicting Phenotype
- The Quantum Leap: Leading the Next Quantum Revolution
- Navigating the New Arctic
- Windows on the Universe: The Era of Multi-messenger Astrophysics

PROCESS IDEAS

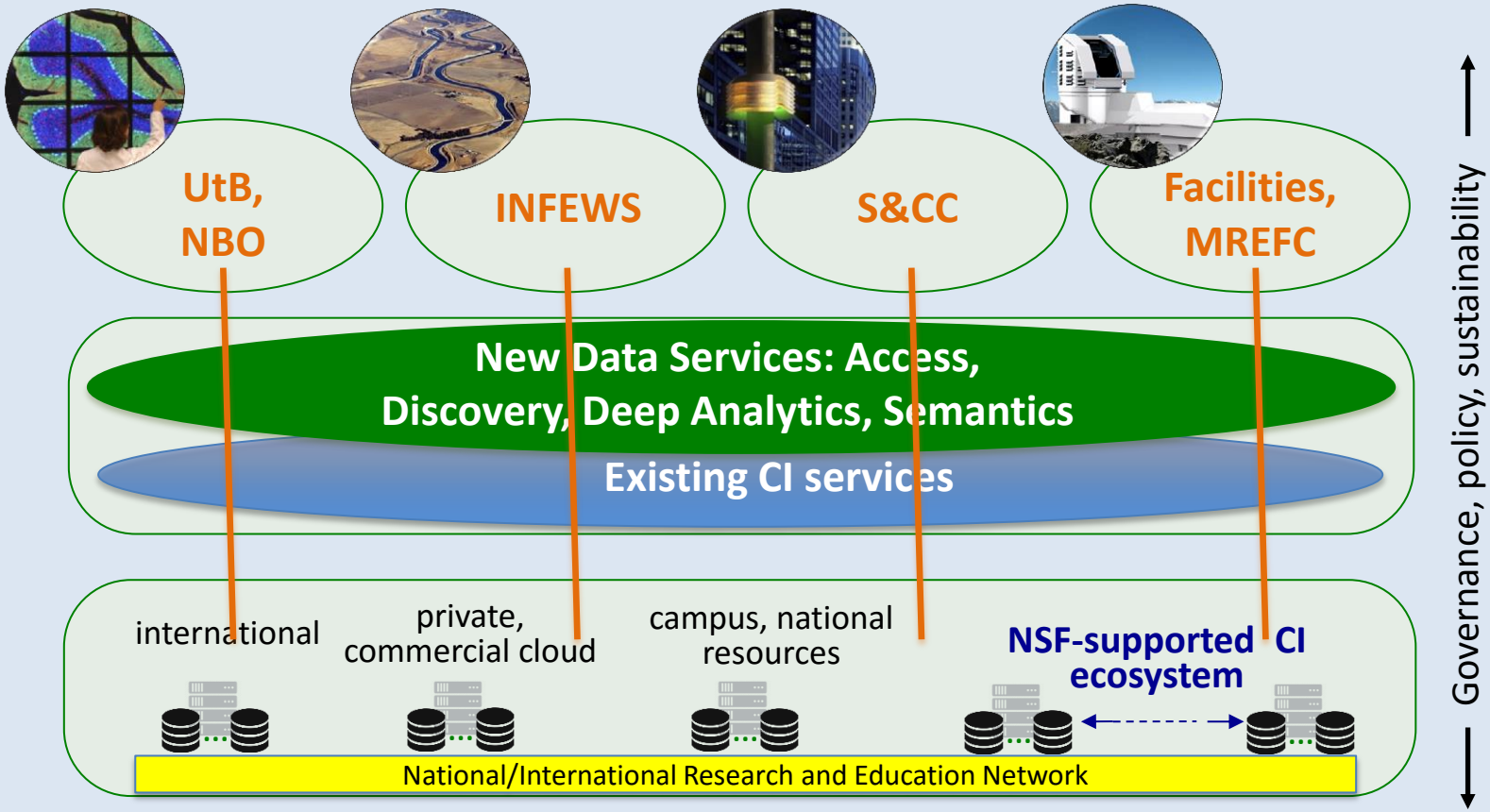
- Growing Convergent Research at NSF
- Mid-scale Research Infrastructure
- NSF 2050

*Video of NSB presentation and discussion is at:
http://www.tvworldwide.com/events/nsf/160505/globe_show/default_go_archive.cfm?gsid=2957&type=flv&test=0&live=0
(the presentation/discussion starts about 20 minutes into this video)

A vision for research cyberinfrastructure

Architecting an open national data infrastructure

Enabling and accelerating science drivers, including NSF initiatives & facilities



Some Related Events / Activities

- CRA – Computing Research Assocation statement on role of CS in Data science
- Recent meetings
  - **US-UK Health Data Science Workshop**, March 1-2, 2016, NIH Campus, Bethesda, MD. Hosted by Stanford University, in conjunction with the Research Councils of the United Kingdom (RCUK), NIH, NSF
  - NSF **BIGDATA PI Meeting**, April 20-21, Arlington, VA
  - NSF Workshop on **TFODS: Theoretical Foundations of Data Science**, April 28-30, 2016
  - CATS **Workshop on Causal Inference from Big Data**, Washington DC, June 2016
- Upcoming events
  - NAS Workshops on **Envisioning the Data Science Discipline: The Undergraduate Perspective**, Washington DC, 2016
  - NITRD Workshop on **Metrics for Assessing the Value of Digital Data Repositories,** Washington DC, 2016

Backup

Big Data / Data Science
Community Building

Foundational research to develop new techniques and technologies to derive knowledge from data

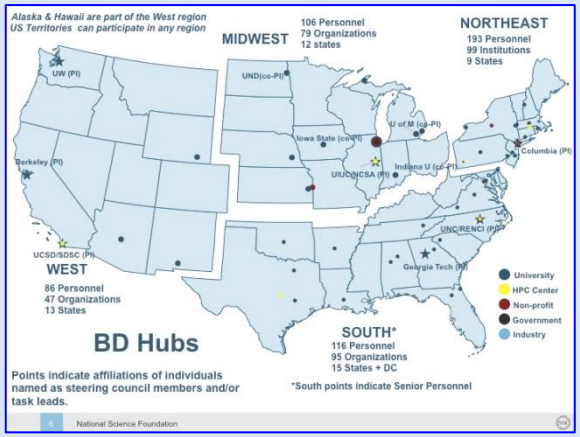
New **cyberinfrastructure** to manage, curate, and serve data to research communities

Policy

New approaches for **education** and **workforce development**

Community building

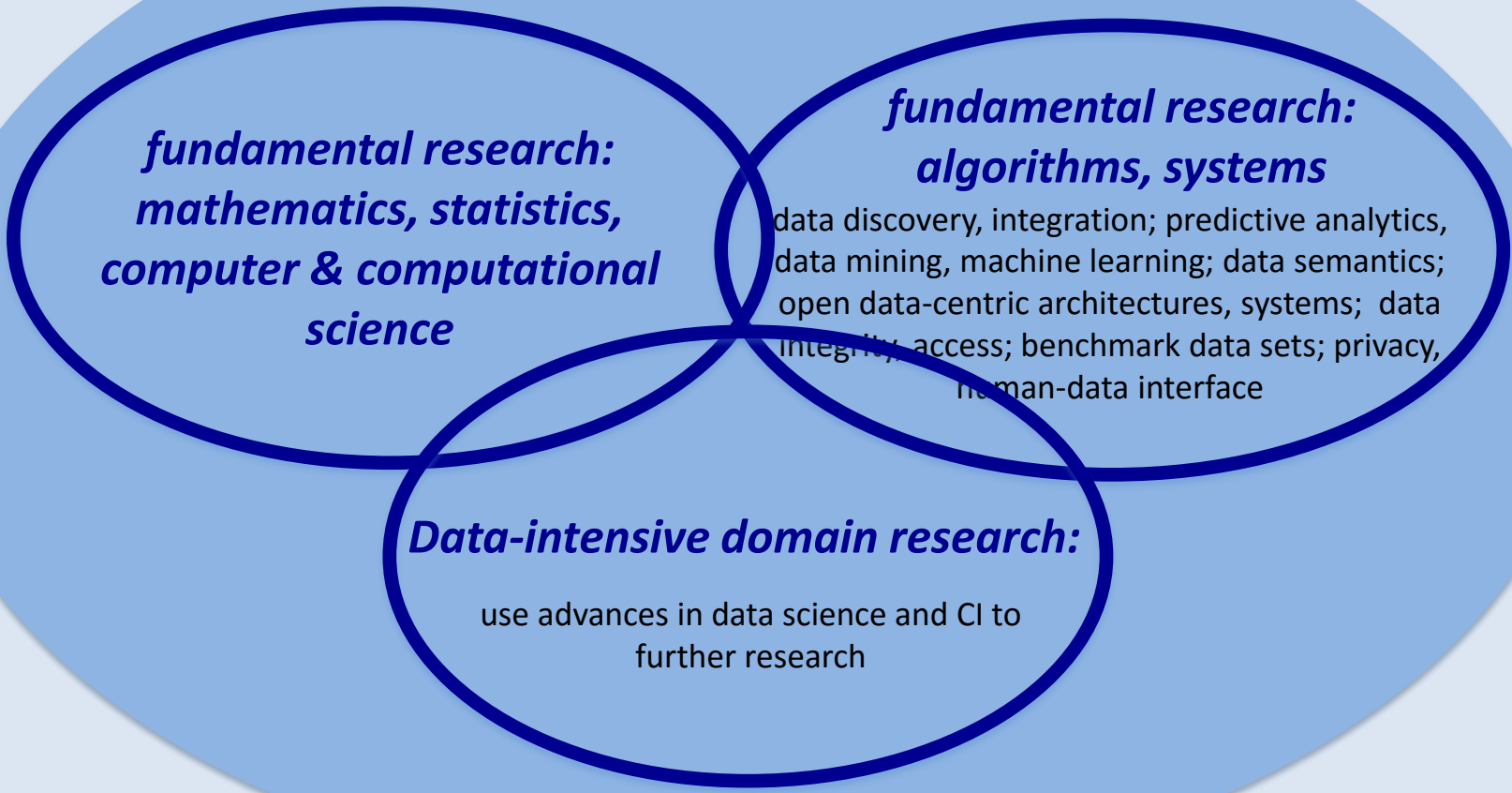
- Big Data Regional Innovation Hubs and Spokes (BD Hubs/Spokes)



BD Spokes, Planning Grants:
Soon to be funded

Harnessing the Data Revolution

Embodiment of innovations in robust, comprehensive, open, science-driven, CI ecosystem: accelerating data-intensive research, including large-scale facilities



Development, evaluation of innovative learning opportunities, educational pathways: grounded in an education-research-based framework