Harnessing the Data Revolution: A Perspective from the NSF

Chaitan Baru
Senior Advisor for Data Science, CISE
National Science Foundation
NSF “Big Ideas”

- Catalyze interest, investment in fundamental research
- Generate broad public appeal and attract partnerships with industry, private foundations, and academia
- Cutting-edge research agenda, appropriate for NSF
- Process: directorate discussions, white papers, AD retreat, subsequent collaborative refinement

Better understanding the changing Arctic is one item on a new list of big ideas that should shape the National Science Foundation's work.

NSF director unveils big ideas, with an eye on the next president and Congress

By Jeffrey Mervis | May 10, 2016, 3:30 PM
Looking Ahead: Ten Big Ideas

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

**Process Ideas**
- Growing Convergent Research at NSF
- NSF-Includes: Enhancing Science and Engineering through Diversity
- Mid-scale Research Infrastructure
- NSF 2050: Seeding Innovation
Interconnectedness of Big Ideas
Harnessing the Data Revolution

**Vision:** move beyond isolated, standalone approaches for data science, services and infrastructure towards a cohesive, federated, national-scale approach that will harness the data revolution and transform U.S. science, engineering, and education over the next decade and beyond

And, **beyond** the 4th Paradigm... “Found Data”, “Data Exhaust”, social media, images, clickstreams, transactions, text collections...
Harnessing Data: Themes

Research across all NSF Directorates

- **Theoretical foundations**: mathematics, statistics, computer & computational science
- **Systems foundations**: data-centric algorithms, systems
- **Data-intensive research**: in all areas of science and engineering

Innovative educational pathways, grounded in an education-research-based framework

Advanced cyberinfrastructure ecosystem for accelerating data-intensive research, including large-scale facilities
NSF Transdisciplinary Research in Principles of Data Science, TRIPODS: Phase I

- From the TRIPODS solicitation, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505347
  - Proposals for TRIPODS Institutes should demonstrate plans to address the following important factors
    - ...Data provenance, **reproducibility**, privacy, and algorithmic fairness are all fundamental topics that Institutes should actively investigate. These areas are important for foundational research to make impacts beyond academic environments...

- Letter of Intent due: Jan 4-19, 2017
- Full proposals due: March 1-15, 2017
A vision for research cyberinfrastructure

Architecting an open national data infrastructure

New Data Services: Discovery, Access, Deep Analytics, Semantics
Core services, e.g. authentication, distributed storage

Enabling and accelerating science drivers, including NSF initiatives & facilities

Privacy-preserving data sharing: major challenge

Governance, policy, sustainability

Increasing interdisciplinary sharing
Increasing disciplinary emphasis

Storage and Compute Infrastructure

Campus Regional Commercial National/International Research and Education Network

Open Data
Open System
Open Source

UtB, NBO
INFEWS Applications: Gateways
S&CC
Facilities, MREFC

Campus Regional Commercial National

National/International Research and Education Network

Increasing interdisciplinary sharing
Increasing disciplinary emphasis

Infrastructure

Applications: Gateways

National/International Research and Education Network

Privacy-preserving data sharing: major challenge

Enabling and accelerating science drivers, including NSF initiatives & facilities
Data Science Infrastructure, Systems, Engineering

- Infrastructure: Storage + Data + Basic Data Services
  - The “`Hello, World` for Big Data” problem [courtesy: RV Guha]
- Not Just Queries: Building multiple interfaces to data
  - Question/Answering
  - Storytelling
  - Dialogs
- ML “Systems”
  - Going from successful verticals to generic systems
- Reproducibility
  - Test and Verification
The Data Ecosystem

Data Services for discovery, access, and integration of information across disparate, distributed information sources

- Independent researcher data
- Small group databases
- Institution-level Repositories
- Large, community data repositories
Developing an Open Knowledge Network

- Entities, Facts, Questions, Answers
  - Recent NITRD meeting. ~20 attendees from industry, academia, agencies.
- Motivation—We are moving towards:
  - More natural human-data interfaces; people interact naturally with information
  - Networked data and information infrastructure capable of supporting integration of information from multiple, independent—often disparate—information sources
- Creating the *Semantic Information Infrastructure* for the Future
  - Use of such an infrastructure will be enabled by “simple” interfaces, e.g. question/answer
- Multiple interfaces
  - Machine learning and knowledge representation can be instrumental in creating systems that support question/answer and dialog-based interactions, and offer comprehensive explanations for data and model-driven, machine-based decisions.
TOKeN: The Open Knowledge Network

- **Vision:**
  - Create an open web-scale knowledge network to foster research and innovation on an entire class of new applications that leverage data, context, and inferences from data.
  - Planning a follow-on workshop, in next few months...
Thanks!

- Chaitan Baru, cbaru@nsf.gov