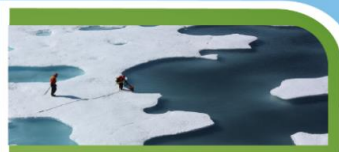
Harnessing the Data Revolution: A Perspective from the NSF

Chaitan Baru
Senior Advisor for Data Science, CISE
National Science Foundation

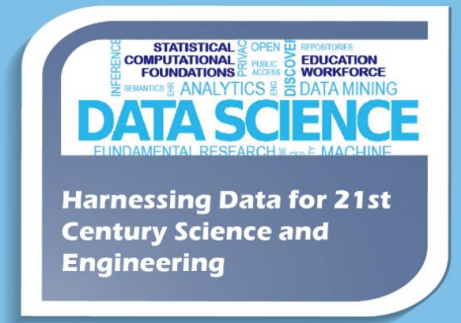
NSF "Big Ideas"

- 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

Looking Ahead: Ten Big Ideas



Navigating the New Arctic



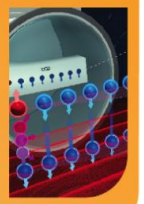
Harnessing Data for 21st Century Science and Engineering

Work at the Human-Technology Frontier: Shaping the Future

RESEARCH IDEAS

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

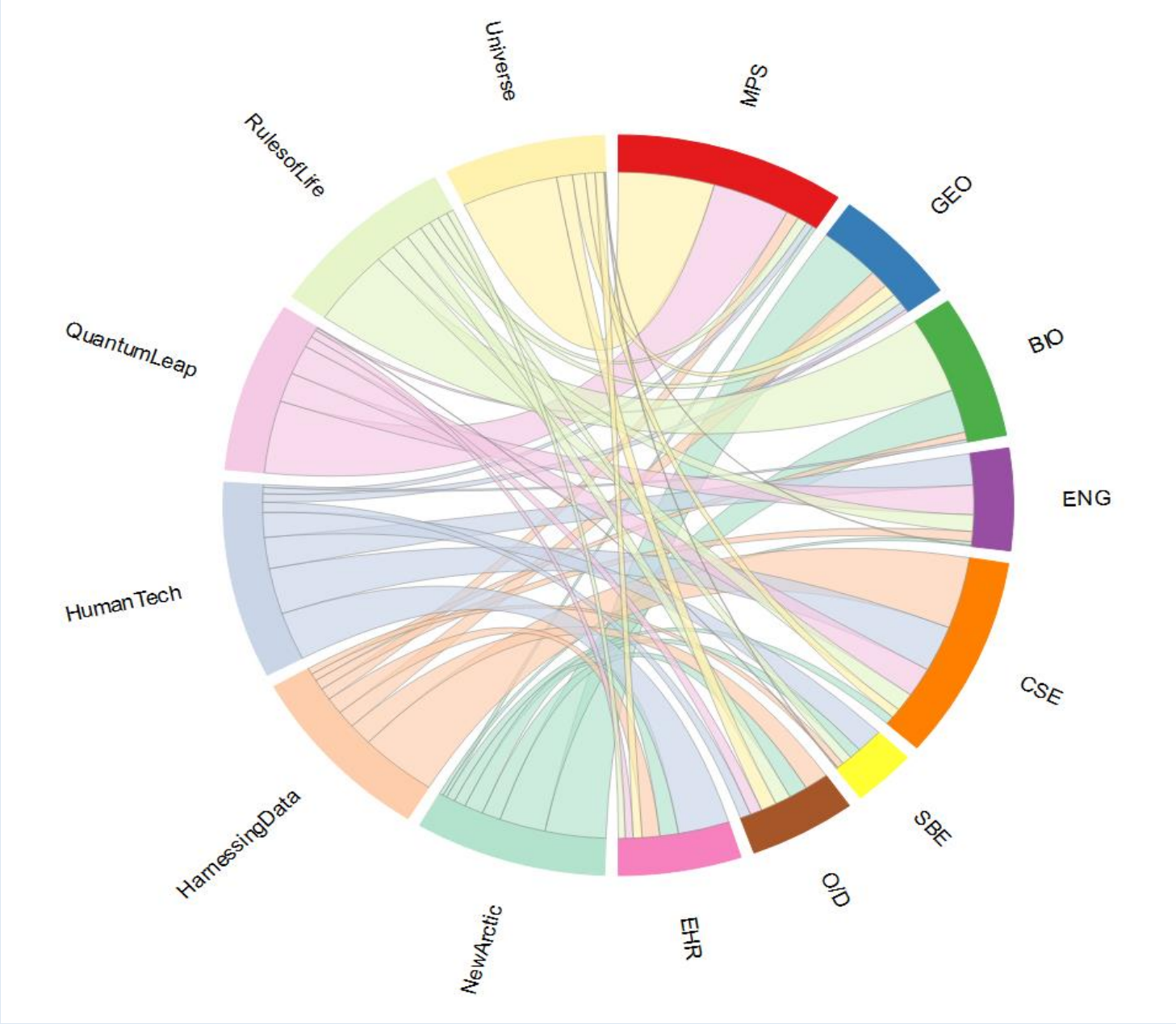
Growing Convergent Research at NSF

NSF-Includes: Enhancing Science and Engineering through Diversity

Mid-scale Research Infrastructure

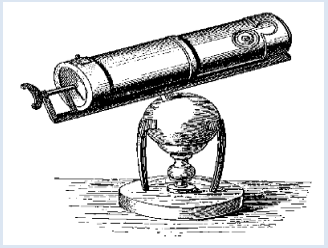
NSF 2050: Seeding Innovation

Interconnectedness of Big Ideas



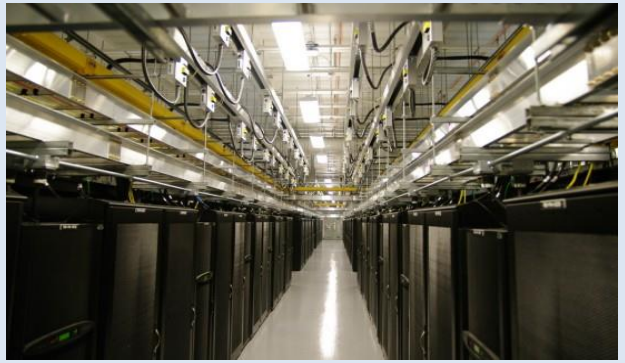
Paul Morris OD/OIA, NSF

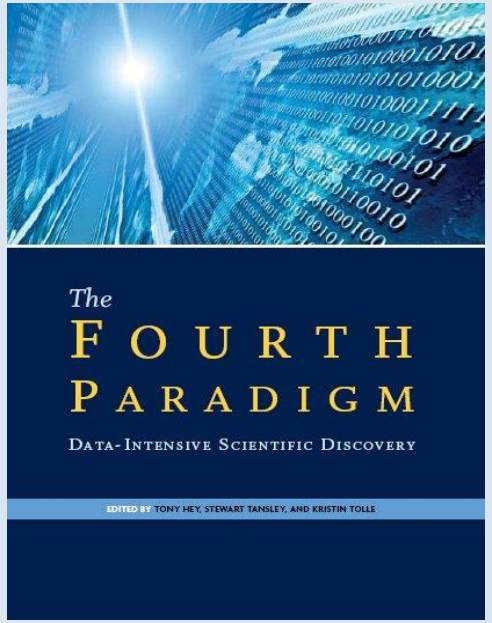
Harnessing the Data Revolution



$$\oint \mathbf{E}\cdot d\mathbf{A} = \frac{q_{enc}}{\varepsilon_0}$$
$$\oint \mathbf{B}\cdot d\mathbf{A} = 0$$
$$\oint \mathbf{E}\cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$$
$$\oint \mathbf{B}\cdot d\mathbf{s} = \mu_0\varepsilon_0\frac{d\Phi_E}{dt} + \mu_0 i_{enc}$$

$$\frac{\partial \rho}{\partial t}+\frac{\partial}{\partial x_i}(\rho u_i)=S_m$$
$$\frac{\partial}{\partial t}(\rho u_i)+\frac{\partial}{\partial x_j}(\rho u_i u_j)=$$
$$-\frac{\partial p}{\partial x_i}+\frac{\partial \tau_{ij}}{\partial x_j}+\rho g_i+F_i$$





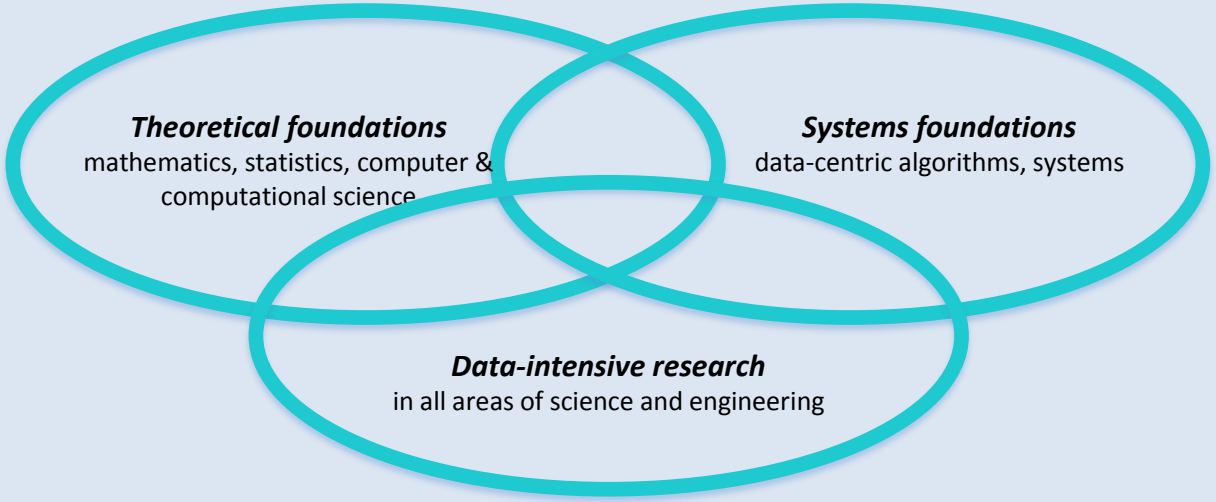
Experimental → Theoretical → Computational → Data

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



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

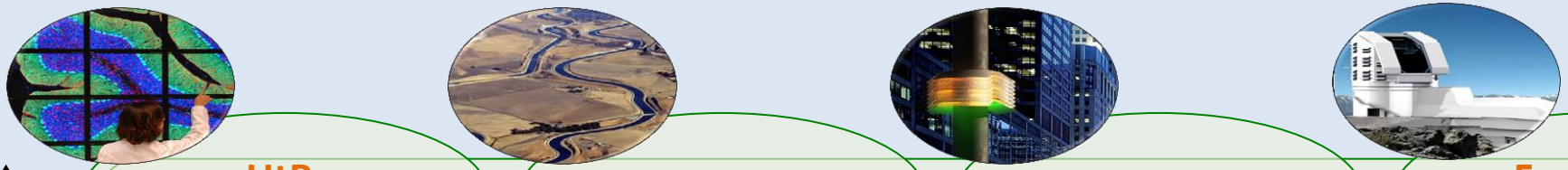
Advanced cyberinfrastructure ecosystem
for accelerating data-intensive research, including large-scale facilities

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



UtB, NBO

INFEWS

S&CC

Facilities, MREFC

Applications: Gateways

New Data Services: Discovery, Access, Deep Analytics, Semantics

Core services, e.g. authentication, distributed st

–Open Data

–Open System

–Open Source

Storage and Compute Infrastructure

Campus | Regional | Commercial | National | International

National/International Research and Education Network

Increasing interdisciplinary sharing

Increasing disciplinary emphasis

Governance, policy, sustainability

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

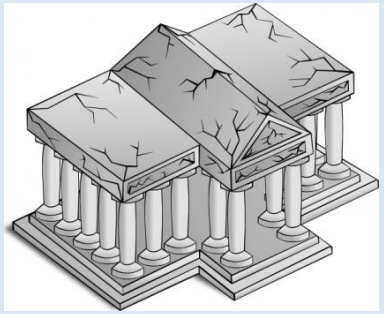
The Data Ecosystem

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

Independent researcher data

Large, community data repositories

Small group databases



Institution-level 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