

1st Annual Commonwealth Computational Summit

Harnessing the Data Revolution

Chaitan Baru
 Senior Advisor for Data Science
 Computer and Information Science and Engineering Directorate

National Science Foundation



1st Annual Commonwealth Computational Summit, Oct 17, 2017

1

NSF “Big Ideas”

RESEARCH IDEAS



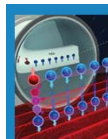
Harnessing Data for 21st Century Science and Engineering

Work at the Human-Technology Frontier: Shaping the Future

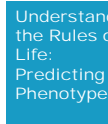


Navigating the New Arctic

Windows on the Universe: The Era of Multi-messenger Astrophysics



The Quantum Leap: Leading the Next Quantum Revolution



Understanding the Rules of Life: Predicting Phenotype

PROCESS IDEAS

Mid-scale Research Infrastructure



NSF 2050: Seeding Innovation



Growing Convergent Research at NSF



NSF-INCLUDES: Enhancing Science and Engineering through Diversity

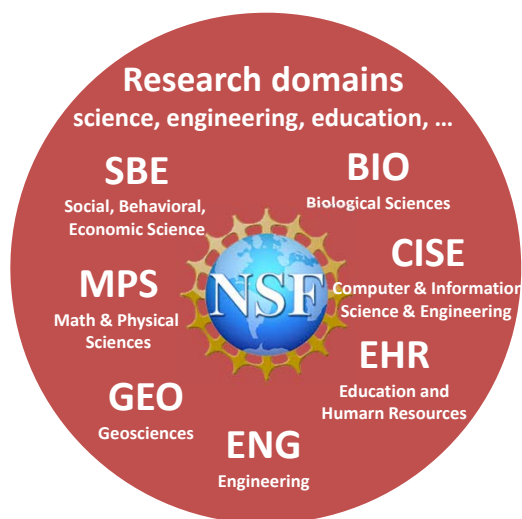


1st Annual Commonwealth Computational Summit, Oct 17, 2017

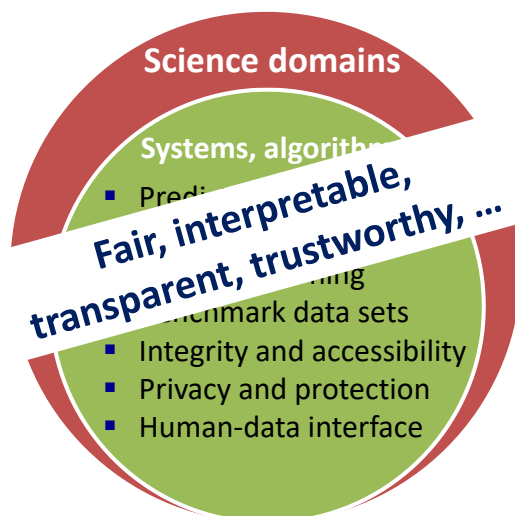
2

2

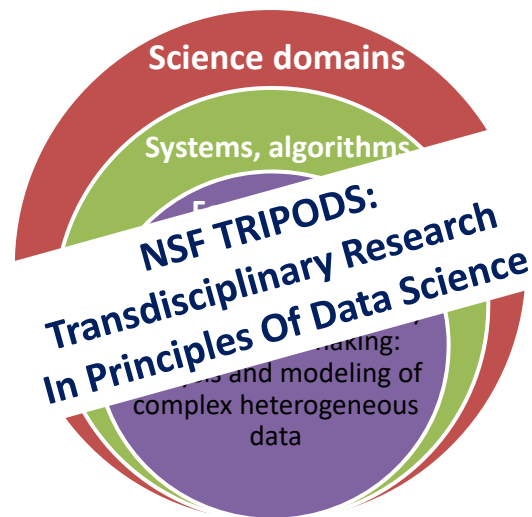
Harnessing the Data Revolution: Domains



Harnessing the Data Revolution: Systems



Harnessing the Data Revolution: Foundations



Data Science Foundations: TRIPODS

- TRIPODS: Transdisciplinary Research In Principles Of Data Science
 - Required close collaboration among CS, Math, Stats
 - **Phase I:** 3 years, 12 “Proto centers”
 - **Possible Phase II:** ~3 Large, national centers (subject to availability of funding).
 - Important: Connection with applications domains, industry



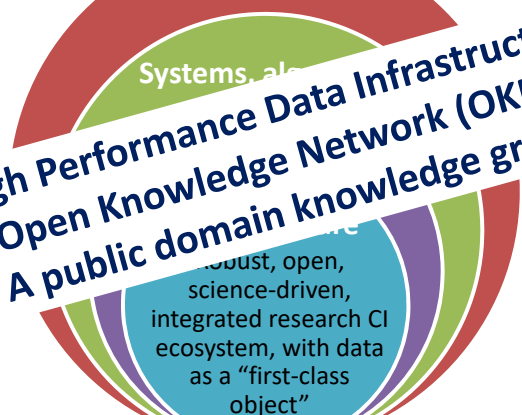
NSF TRIPODS Projects

TRIPODS Kickoff PI Meeting,
Oct 12-13, Alexandria, VA

1. [UA-TRIPODS: Building Theoretical Foundations for Data Sciences](#): Hao Zhang, University of Arizona
2. [Foundations of Model Driven Discovery from Massive Data](#): Jeffery Brock, Brown University (Convergence and EPSCoR co-funding)
3. [Berkeley Institute on the Foundations of Data Analysis](#): Michael Mahoney, University of California, Berkeley
4. [TRIPODS: Towards a Unified Theory of Structure, Incompleteness and Uncertainty in Heterogeneous Graphs](#): Lise Getoor, University of California, Santa Cruz
5. [From Foundations to Practice of Data Science and Back](#): John Wright, Columbia University
6. [TRIPODS: Data Science for Improved Decision-Making: Learning in the Context of Uncertainty, Causality, Privacy, and Network Structures](#): Kilian Weinberger, Cornell University (Convergence co-funding)
7. [Transdisciplinary Research Institute for Advancing Data Science \(TRIAD\)](#): Xiaoming Huo, Georgia Institute of Technology
8. [Collaborative Research: TRIPODS Institute for Optimization and Learning](#): Katya Scheinberg, Lehigh University; Han Liu, Northwestern University; Francesco Orabona, State University of New York at Stony Brook
9. [Institute for Foundations of Data Science \(IFDS\)](#): Piotr Indyk, Massachusetts Institute of Technology
10. [Topology, Geometry, and Data Analysis \(TGDA@OSU\): Discovering Structure, Shape, and Dynamics in Data](#): Tamal Dey, The Ohio State University
11. [Algorithms for Data Science: Complexity, Scalability, and Robustness](#): Sham Kakade, University of Washington
12. [Institute for Foundations of Data Science](#): Stephen Wright, University of Wisconsin-Madison (Convergence co-funding)



Harnessing the Data Revolution: Cyberinfrastructure

- 
- High Performance Data Infrastructure
 - Open Knowledge Network (OKN):
A public domain knowledge graph
- robust, open,
science-driven,
integrated research CI
ecosystem, with data
as a "first-class
object"



Open knowledge network

- An open **web-scale** knowledge network
 - Suggested by the community: Andrew Moore, CMU; Ramnathan Guha; et al.
- Semantically-linked concepts, data
 - To foster research on an entire class of new applications leveraging data, context, and inferences from data
- Question/answer interfaces, dialog-based interactions, explanatory/story-telling interfaces
- Joint academia, industry, government workshops
 - July 2016, Washington, DC
 - Feb 2017, Sunnyvale, CA
- [Oct 4,5, 2017, National Library of Medicine, Bethesda, Maryland](#)
 - System architecture/software; data representation/curation; research related to representation and use of massive knowledge graphs
 - Domains discussed: Biomedical, Finance, Geoscience, Manufacturing



CISE DCL on Cloud Computing

- Expand upon partnership initiated with Google, Microsoft
- Capitalize on the cloud providers of-the-art, on-demand, cloud computing
- Benefits of partnership
 - Access by CISE research and education community to a range of useful resources and services —scalable storage; real-time analytics; streaming data services; state-of-the-art compute nodes;
 - Ability to experiment with real datasets where scale and performance are key considerations. Exploit resources offered by the cloud providers



Community workshop on Cloud Computing and CS Research and Education



Harnessing the Data Revolution: Education

Science domains



- Envisioning the Data Science Discipline: The Undergraduate Perspective, National Academy of Sciences, study/workshops
 - <https://www.nap.edu/catalog/24886/> (Interim Report)
- Keeping Data Science Broad: Negotiating the Digital and Data Divide, Oct.31-Nov.1, 2017, Atlanta, GA (Renata Rawlings-Goss, GaTech)
- Convergence/HDR Workshop: Social Science Insights for 21st Century Data Science Education (SSI) (PI: Cathryn Carson, co-Pis: David Culler, Saul Perlmutter, Berkeley)



Putting it all Together: Translational Data Science

Application of data science techniques, tools in science and other applications domains

- 1st Workshop on Translational Data Science, June 26-27, 2017, U.Chicago (Robert Grossman, Chicago; Raghu Machiraju, OSU)
- 2nd Workshop on TDS, November 13-14, UC Berkeley (David Culler, Berkeley)
- 3rd Workshop on TDS, planned for March 2018 (Kathy McKeown, Columbia; Juliana Freire, NYU; Deborah Estrin, Cornell Tech)



Data Science Corps

Getting your hands dirty with data!

- **VISION**

- Provide practical experiences, teach new skills, and offer teaching opportunities in data science to U.S. data scientists and data science students, in the service of science and society

- **MISSION:**

- Enable U.S. data scientists and data science students to obtain practical experience with data-intensive applications;
- Promote a better understanding of the power of data, and the role that data can play in addressing issues at the local, regional, national, and international levels;
- Teach data literacy and provide basic training in data science to the existing workforce in communities, organizations, and institutions at the local, state, national, and international levels



1st Data Science Corps Workshop, Dec 7-8, 2017, McCourt School of Public Policy, Georgetown University

Data Science Corps

Data Science Corps Volunteers / Volunteer Organizations

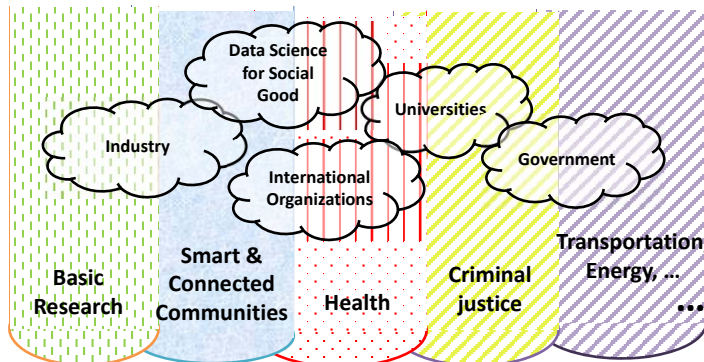
Data Science Corps Projects / Project Organizations

Professionals
from Industry, NGOs



Students
from Academic Institutions

Edu
Institutions
Undergrad / Grad
Universities;
4-year Colleges;
Community Colleges;
Online Programs;
etc.



Thank You!

- cbaru@nsf.gov

