Harnessing the Data Revolution

Chaitan Baru (cbaru@nsf.gov)
Senior Advisor for Data Science
Directorate for Computer and Information Science and Engineering
National Science Foundation

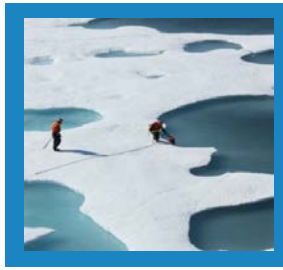
Panel on Big Data and Cyberinfrastructure:
Building Jurisdictional Capabilities to Align with NSF Strategies Panel
EPSCoR Meeting on: Expanding Networked Communities to Advance Research Competitiveness
May 2018

NSF "Big Ideas"

RESEARCH IDEAS

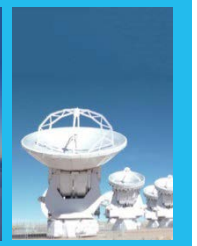
Harnessing Data for 21st Century Science and Engineering

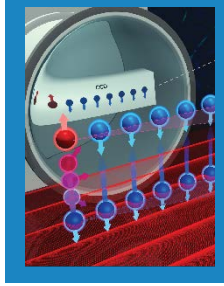
Work at the Human-Technology Frontier: Shaping the Future



Navigating the New Arctic

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



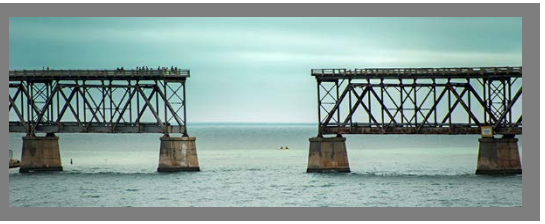


The Quantum Leap: Leading the Next Quantum Revolution

Understanding the Rules of Life: Predicting Phenotype

PROCESS IDEAS

Mid-scale Research Infrastructure



NSF 2050: Seeding Innovation

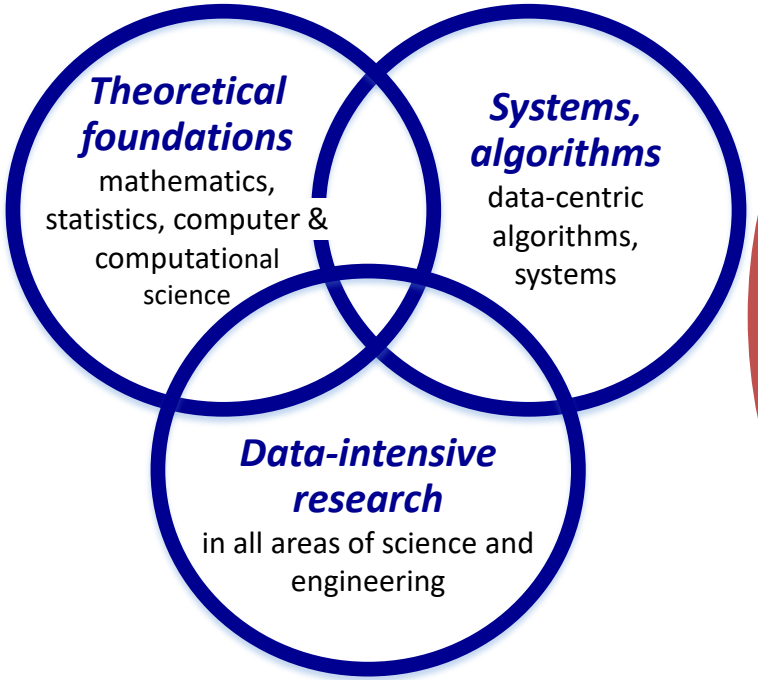
Growing Convergent Research at NSF

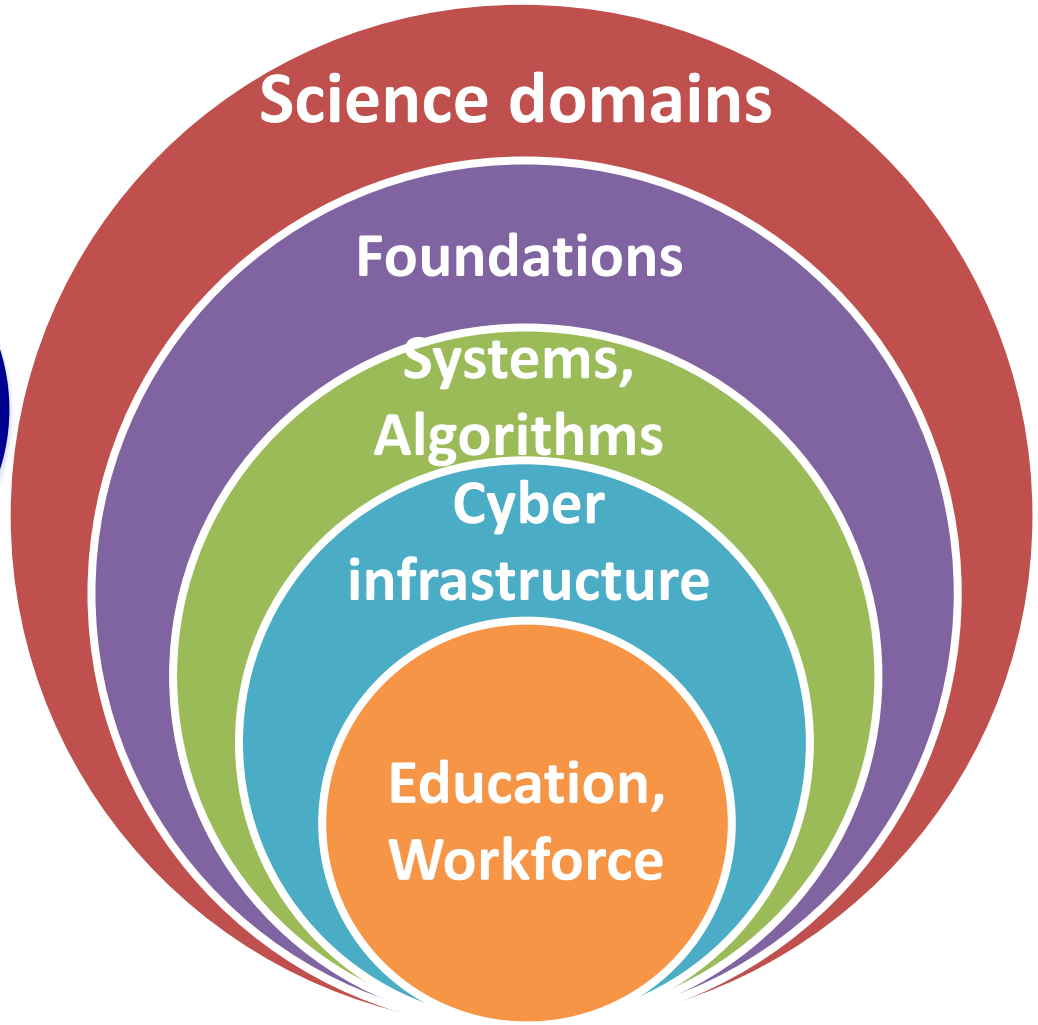
NSF-INCLUDES: Enhancing Science and Engineering through Diversity

*"Engage NSF's research community in the pursuit of **fundamental research in data science and engineering**, the development of a cohesive, federated, national-scale approach to **research data infrastructure**, and the development of **a 21st-century data-capable workforce**."*

Harnessing the Data Revolution: five themes

Research across all NSF Directorates





Educational pathways

Innovations grounded in an education-research-based framework

Advanced cyberinfrastructure

Accelerating data-intensive research

Foundations: Transdisciplinary Research in Principles of Data Science (TRIPODS) NSF 16-615

Foundations

***Workshop on Theoretical Foundations of Data Science*, April 2016**

- Design, enable center-scale collaboration on theoretical research among mathematicians, statisticians, computer and computational scientists

TRIPODS Phase I: develop capacity, thematic directions for traditional center-like activities. 12 awards

- long term research, education, workforce transcending disciplinary boundaries

A foundation for convergent research in data science

TRIPODS+X: *Partnerships between Science and Engineering Fields and the NSF TRIPODS Institutes*

Foundations

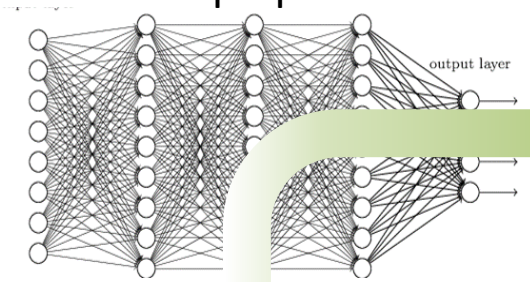
- Expand scope of the TRIPODS program by engaging researchers across other NSF disciplines and the TRIPODS research teams in collaborative activities.
- Foster relationships between researchers in science & engineering domains and foundational data scientists by leveraging existing NSF investments in the TRIPODS organizations.
- Focus on data-driven research challenges motivated by applications in one or more science and engineering domains or other activities aimed at building robust data science communities.

Fair, Interpretable, Transparent, and Trustworthy Data Science

Systems, Algorithms

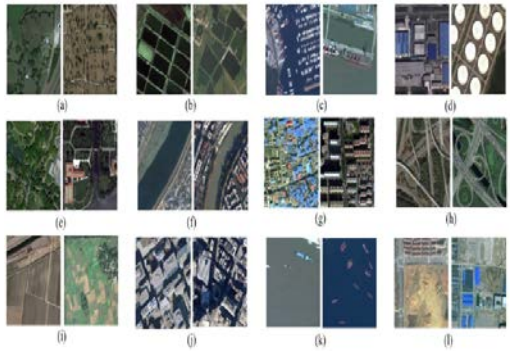
Today's ML systems: opaque, predictive interpretation hard

Trained opaque classifier



prediction

Image training data

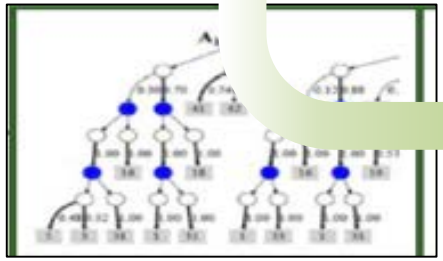


new image(s)

Accuracy?

Explainability?

- Why this prediction?
- Why not another?
- How to correct errors?



prediction

Fair, interpretable, transparent, trustworthy classifier

Fair, Interpretable, Transparent, and Trustworthy Data Science

Systems, Algorithms

Big Data
Volume 5 Number 2, 2017
© Mary Ann Liebert, Inc.
DOI: 10.1089/big.2016.0048

ORIGINAL ARTICLE

Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification

Opinion | OP-ED CONTRIBUTOR

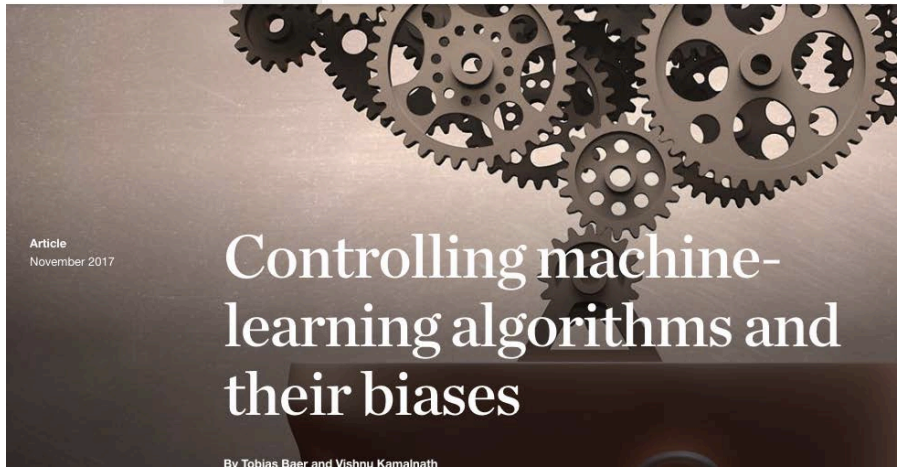
When an Algorithm Helps Send You to Prison

By ELLORA THADANEY ISRANI OCT. 26, 2017

SundayReview | OPINION

Artificial Intelligence's White Guy Problem

McKinsey&Company Risk | Our Insights | How We Help Clients | Our People | Contact Us



MIT Technology Review

Intelligent Machines

Biased Algorithms Are Everywhere, and No One Seems to Care

The big companies developing them show no interest in fixing the problem.

by Will Knight July 12, 2017

Open Knowledge Network:
A National Semantic Data Infrastructure

Systems, Algorithms

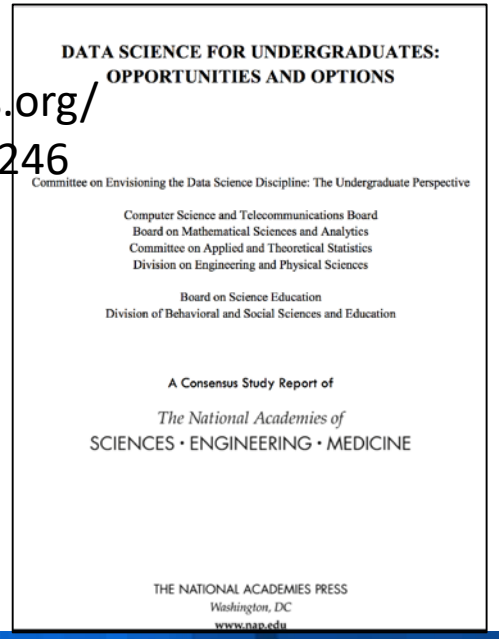
- An open **web-scale** knowledge network
- Semantically-linked concepts, data
  - To foster research on an entire class of new applications leveraging data, context, and inferences from data
- Question/answer interfaces, dialog-based interactions, explanatory/story-telling interfaces
- Joint academia, industry, government workshops
  - July 2016, Washington, DC
  - Feb 2017, Sunnyvale, CA
  - Oct 4,5, 2017, National Library of Medicine, Bethesda, Maryland
- Interest from industry and multiple agencies
  - E.g., NIH Data Translator; DOE Precision Medicine, Cognitive Supercomputing, and Exascale

HDR Education and Workforce

Education, Workforce, Outreach

- *NASEM Study/Report on Envisioning the Data Science Discipline: The Undergraduate Perspective, May 2018*
- NSF Workshop on *Keeping Data Science Broad: Negotiating the Digital and Data Divide*, Oct, 2017
  - Part of *Keeping Data Science Broad* series, NSF South Big Data Hub

http://sites.nationalacademies.org/cstb/currentprojects/cstb_175246



DATA SCIENCE FOR UNDERGRADUATES: OPPORTUNITIES AND OPTIONS

Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective

Computer Science and Telecommunications Board
Board on Mathematical Sciences and Analytics
Committee on Applied and Theoretical Statistics
Division on Engineering and Physical Sciences

Board on Science Education
Division of Behavioral and Social Sciences and Education

A Consensus Study Report of

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

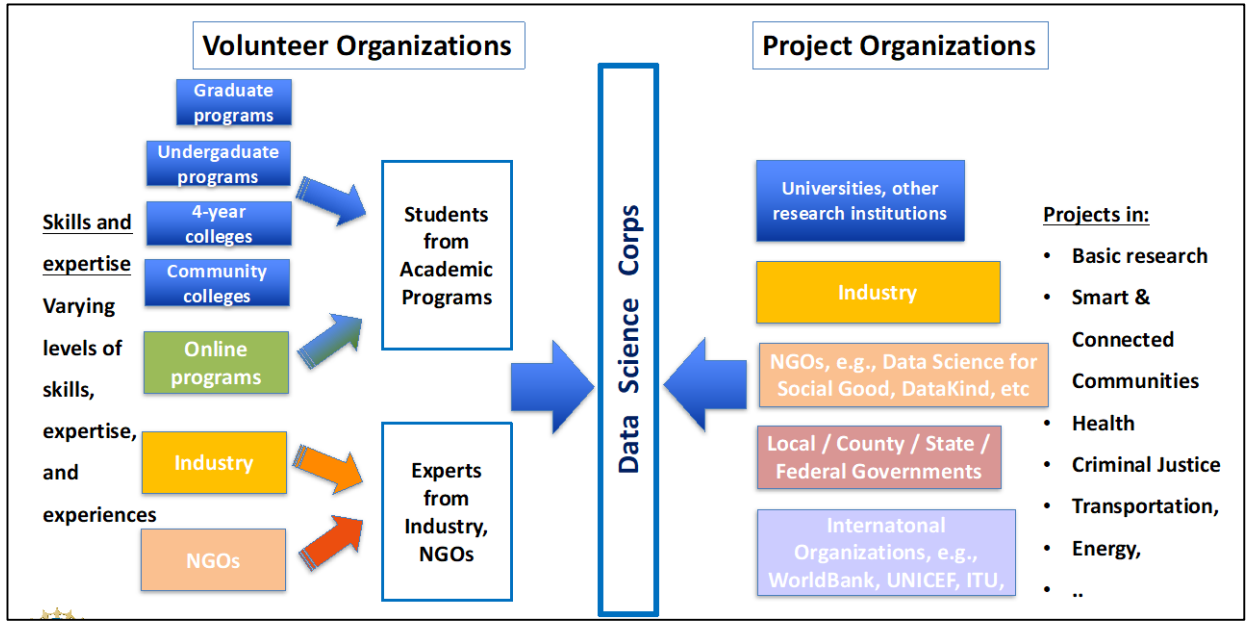
http://bit.ly/KeepingDataScienceBroad_Report

HDR: Data Science Corps

Education, Workforce, Outreach

Data Science Corps: helping to unleash the power of data at the local, state, national, and international levels in the service of science and society by providing practical experiences, teaching new skills, and offering teaching opportunities in data science to U.S. data scientists and data science students.

- Data Science Corps Workshop, Dec 2017, Georgetown Univ



Workshop on "Taking Data Science to America's Emerging Workforce"

Education, Workforce, Outreach

Theme: Can the land grant institution system, and its Cooperative Extension program, be used as a model for disseminating data science through the rural areas and small towns of America?

- June 5-6, Raleigh, NC
- Organized by Marc Hoit, NCSU; Alan Blatecky, UNC Chapel Hill

National Data Cyberinfrastructure

- Vision should span NSF; campuses (CIOs); public clouds (industry), etc.
- NSF Workshop on *Enabling Computer and Information Science and Engineering Research and Education in the Cloud*, Jan 8-9, 2018, Alexandria, VA
- Data Ecosystem includes:
  - Data collection/generation platforms
  - Management of “active” data, including analysis/ML/…
  - Data Repositories
  - Data Archives
- Open, extensible platforms
  - Sharing of data…but also models, semantic information,
  - Open Knowledge Network, model sharing, open storage platforms, …

Cyber infrastructure