Data and Knowledge as Infrastructure

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

Motivation…Easy access to data

- The ‘Hello World’ problem (courtesy: R.V. Guha)
  - Access a 1PB (or, 100TB, or 10TB?) dataset
  - Create a subset of 10TB
  - Perform an operation (statistical computation)
  - Print the result
  - Do this as a homework problem by next class session
  - In a class of 500 students…
- Dataset size is not important; could be about accessing multiple, heterogeneous data sources, …

Motivation…Better access to data

- Why can't I talk to my data?
  - Natural (natural language) interfaces to data
  - And talk to my data about other data…?
- Story Telling
  - Need to be able to tell stories *about* your data
    - Milind Kamkolkar, CDO, Sanofi, hired journalists as his first hires as a CDO. From MIT CDOIQ meeting, July 12-14, 2017
  - Want to tell stories *with* data

Motivation…Data in an interlinked world

- NITRD Big Data Interagency Working Group Workshop on Metrics for Digital Data Repositories, July 2017
  - An observation: One of the evaluation criteria for data repositories should be about how well they are “networked” to other data

NSF "Big Ideas"

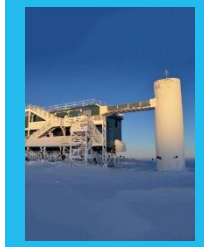
RESEARCH IDEAS

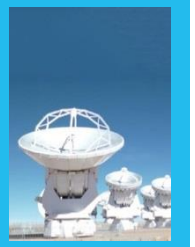
Harnessing Data for 21st Century Science and Engineering

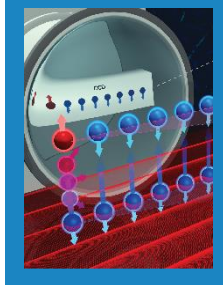
Work at the Human-Technology Frontier: Shaping the Future

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

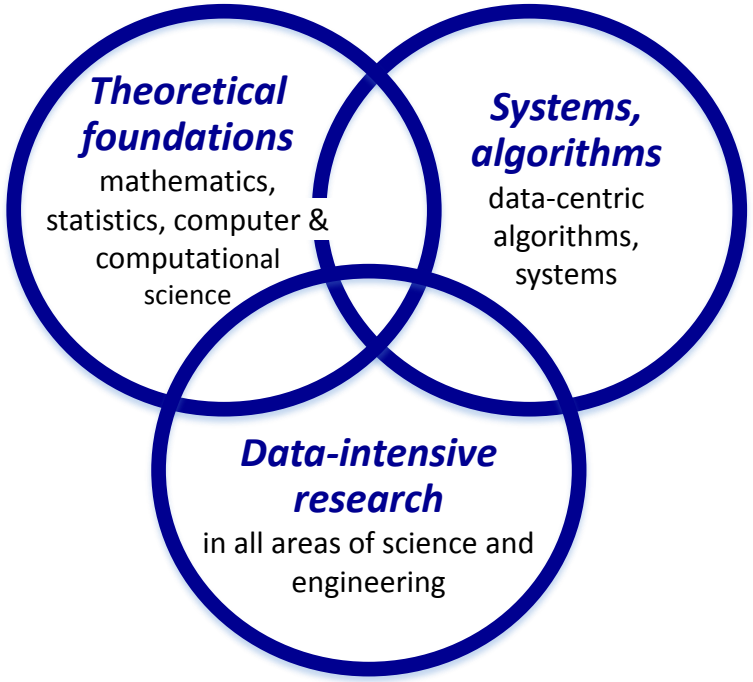
NSF 2050: Seeding Innovation

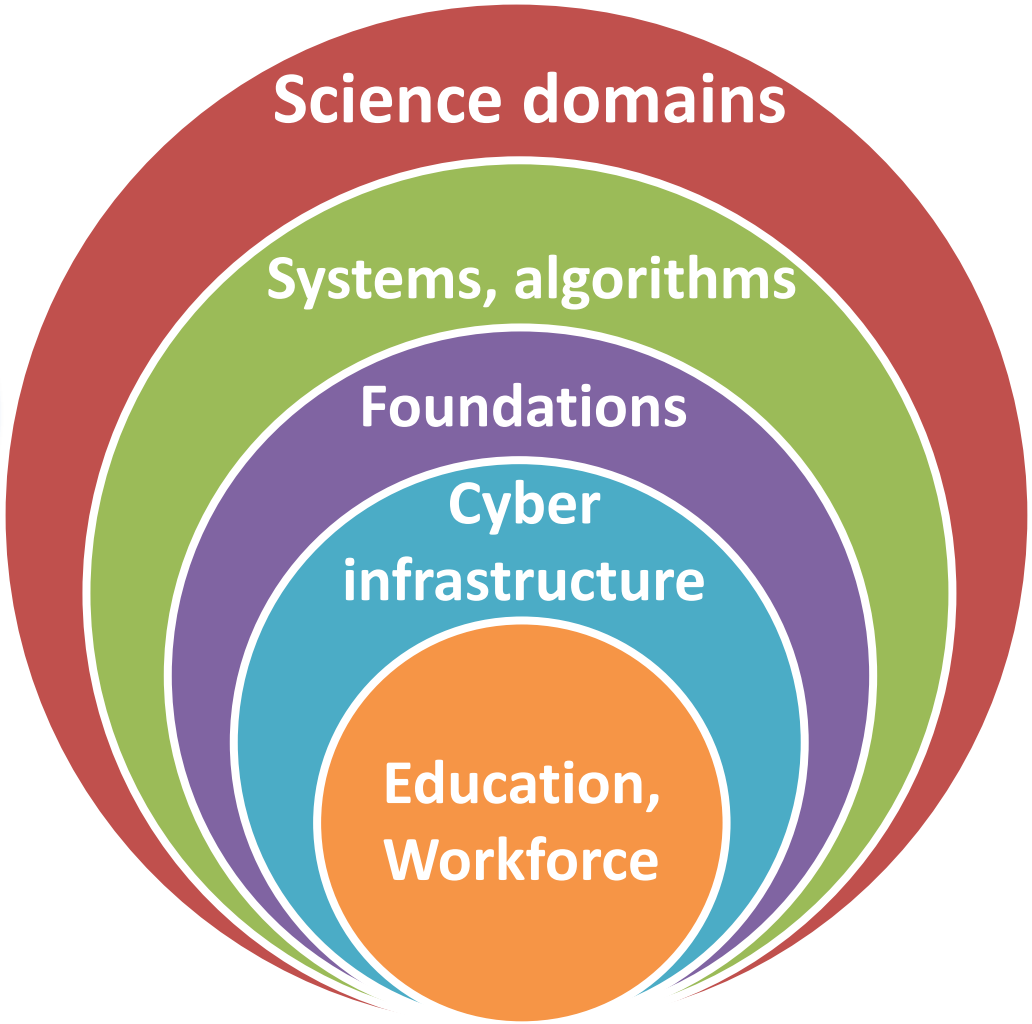
Growing Convergent Research at NSF

NSF-INCLUDES: Enhancing Science and Engineering through Diversity

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

Motivation for Knowledge Infrastructure

- Foster research on a class of new applications leveraging data, context, and inferences from data
- Support integrative analysis and interpretation of multimodal data
- Develop advanced applications, e.g.:
  - Question/answer interfaces
  - Dialog-based interactions
  - Explanatory/story-telling interfaces

Past/Current Related NSF Efforts

- Research on
  - creation of knowledge bases (representation, performance)
  - creation of ontologies
  - knowledge extraction
  - knowledge aggregation
  - reasoning …

Example NSF projects - 1

- **Knowledge Graph Mining for Financial Risk Analytics**, PI: Mohammed Zaki, 2017
  - a "financial risk" knowledge graph from textual and semantic features mined from the publicly available annual and quarterly reports filed with the SEC; and textual data from news articles and credit assessment reports.

- **Developing the Next Generation of Community Financial CyberInfrastructure for Monitoring and Modeling Financial Eco-Systems and for Managing Systemic Risk**, PI: Louiqa Raschid, 2013
  - Financial entity identification data challenges 2016, 2017
  - In collaboration with NIST and OFR, https://ir.nist.gov/dsfin
  - Creation of multiple open source graph datasets using SEC filings—in collaboration with IBM Almaden.

Example NSF projects - 2

- **From Data to Knowledge: Extracting and Utilizing Concept Graphs in Online Environments**, PI: Cornelia Caragea, 2016
  - Explore construction of scholarly knowledge graphs by combining evidence from multiple resources, in an open information extraction framework;
  - Design and develop novel algorithms for detection and analysis of interesting and previously unknown connections between concepts, to enforce knowledge discovery on the Scholarly Web;
  - Investigate the utility of scholarly knowledge graphs in a question answering system

Example NSF projects – 3

- **Scalable Probabilistic Inference for Large Knowledge Bases**, PI: Dan Suciu, 2016
  - Use of database technology to support construction of knowledge bases/graphs
- **Efficient Query Processing over Large Probabilistic Knowledge Bases**, PI: Daisy Zhe Wang, 2015
  - Infer missing knowledge from large-scale knowledge bases
- **Fusion of Heterogeneous Networks for Synergistic Knowledge Discovery**, PI: Philip Yu, 2015
  - Effective transfer of relevant knowledge across “partially aligned” networks—depends upon the relatedness of the different networks, and also the target applications/uses

Example NSF projects - 4

- **Constructing Knowledge Bases by Extracting Entity-Relations and Meanings from Natural Language via "Universal Schema”**, PI: Andrew McCallum, 2015
  - Automated knowledge base (KB) construction from natural language

- **Knowledge Graph Query Processing and Benchmarking**, PI: Xifeng Yan, 1528175
  - Provide a standardized way to fairly and comprehensively evaluate different knowledge graph query algorithms;
  - Improve understanding of existing query engines;
  - Advance the area by providing a common benchmarking framework

Example NSF projects - 5

- **Using Knowledge Resources to Improve Information Retrieval, PI: Jamie Callan, 2014**
  - Examines how to use knowledge bases to improve IR tasks such as *ad hoc* search
  - Some of the work was performed in conjunction with Allen Institute for Artificial Intelligence's Semantic Scholar search engine.
  - Link documents and queries to the KB through entities…which improves the representation of the query and document, leading to more accurate ranking.
  - **KG4IR**: The First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis, in conjunction with ACM SIGIR 2017, Tokyo, Japan, August 11, 2017

Science and Ontologies

- Many efforts across sciences, especially Biomedical, Biology, Ecology, in developing and using ontologies
- Some significant effort in other domains, e.g. astronomy, hydrology, some areas of engineering
- More recent efforts in other domains, e.g. materials science, social science, education research, …

Recent related meetings

- Community and inter-agency meetings
- *Entities, Facts, Questions, Answers: Building the Foundations for Semantic Information Processing*
  - July 2016, Washington, DC
- *TOKeN: The Open Knowledge Network*
  - February 27th, Sunnyvale, CA
- *Workshop on Creating an Open Knowledge Network*
  - October 4-5, 2017, National Library of Medicine, Bethesda, MD,
  - Attendees from academia, industry, govt
  - Participation by NSF, NIH, DARPA, NIST, NASA