Knowledge Networks, ModelCommons, and Beyond…
Harnessing the Data Revolution

Chaitan Baru
Senior Advisor, Data Science Research Initiatives
Office of Research Affairs
San Diego Supercomputer Center
Halicioglu Data Science Institute
UC San Diego
What we may hear today about the Needs and Opportunities for the SCEC community

- Needs: Big Compute, Big I/O, Big data
- Possible needs/opportunities: *Hybrid* computing – HPC and Cloud, HPC in Cloud?
- Possible opportunities: Machine learning, data integration

![Image: A generic ML workflow that guides many applications: (1) data collection, (2) preprocessing, (3) model training, (4) model evaluation, and (5) production.](image-url)
SCEC Opportunities in the context of NSF programs

- [PS my background: was Senior Advisor for Data Science, CISE Directorate, NSF, 2014-2018, engaged in the HDR Big Idea.]

- Returning to NSF as Senior Advisor, NSF Convergence Accelerator for the Open Knowledge Network, Aug 2019

- Note: Speaking here on behalf of myself!

- **Big Ideas**
  - HDR
  - Quantum Leap

- **Big Idea Opportunities?**
  - Mid-scale Research Infrastructure
  - Growing convergent research
NSF Convergence Research

- Growing Convergence Research solicitation, NSF 19-551
  - May 8, 2019, Feb 3, 2020
- Convergence research...
  - A means for solving vexing, complex research problems focusing on societal needs. Entails integrating knowledge, methods, and expertise from different disciplines and forming novel frameworks to catalyze scientific discovery and innovation.
- Characteristics...
  - Research driven by specific and compelling problem arising from deep scientific questions or pressing societal needs.
  - Deep integration across disciplines. New frameworks, paradigms or even disciplines can form sustained interactions across multiple communities.
Convergence Research...

• From inception, *intentionally* brings together intellectually diverse researchers and stakeholders to frame the research questions, develop effective ways of communicating across disciplines and sectors, adopt common frameworks for their solution, and, when appropriate, develop a new scientific vocabulary.

• Research teams practicing convergence aim at developing sustainable relationships that may not only create solutions to the problem that engendered the collaboration, but also develop novel ways of framing related research questions and open new research vistas.
NSF HDR Big Idea

- **TRIPODS: Transdisciplinary Research In Principles of Data Science**
  - Phase 1: 12 projects
  - Phase 2: 2-3 large, national centers

- **HDR Institutes: Convergent research in data-driven topics**
  - Phase 1 Planning grants based on Ideas Labs, Framework proposals
  - Phase 2: A few large, national centers

- **HDR Data Science Corps: Data science education**

- **HDR Convergence Accelerator for the Open Knowledge Network**
  - Phase 1: Upto $1M for 9 months for planning
  - Phase 2: Upto $4M for 2 years for implementations
  - Total: About $40-60M / year
NSF Convergence Accelerator
Track A1: Open Knowledge Network (OKN)

• OKN objectives:
  • NSF seeks to create an OKN to drive innovation across all areas of science and engineering, and unleash the power of data and artificial intelligence to achieve scientific discovery and economic growth.
  • Create a nonproprietary shared knowledge infrastructure.
  • Build public-private cooperation and engage convergence teams from all areas of data science and science and engineering domains to create a shared, open infrastructure.

• Phase 1
  • Supports multidisciplinary and multi-institutional teams needed to identify the development paths for an OKN,
  • Emphasis on exploiting publicly available data sets, including U.S. Government data
  • Teams may address "horizontal" challenges that apply to all domains, addressing challenges such as developing the underlying representation of facts, querying services that perform reasoning tasks with the data, or developing secured access capabilities.
  • Or, "vertical" challenges specific to different topical domains such as geosciences, education, smart health, finance, and manufacturing.
  • A participatory design approach that considers the needs and perspectives of the many user communities will be essential.
What *is* a knowledge network?

- A network of (people) expertise?
- A (distributed, scalable) technical infrastructure for knowledge representation, question-answering, deduction,…?
- A representation of domain knowledge for a given domain?
- In what ways is SCEC already a knowledge network?
Convergence Accelerator: Future Topics

• NSF 19-065
  • Dear Colleague Letter: Request for Information on Future Topics for the NSF Convergence Accelerator, due June 24, 2019.

• SCEC actions
  • See if there are any opportunities to collaborate with any of the funded Phase 1 projects, and join them more fully in Phase 2.
  • See if there are opportunities to participate in workshops, etc related to future Convergence Accelerator topics
“ModelCommons”: Infrastructure for sharing data and models

- Current modeling practices are rather *ad hoc*, often depending upon the experience and expertise of individual data scientists and types of pre-processing used, which may be specific to domains.
- There is a need for cataloging, sharing and discovering the models.
- Different application domains/disciplines may use similar models and modeling tools, yet, sharing is limited.
- Modeling results often have poor reproducibility; information on when/how a model works, and when it may fail, is oftentimes not clearly recorded.
- Model provenance and the original intent behind the knowledge discovery process should be well-recorded.
- Many predictive analytics algorithms are not transparent to end-users.
- The need for cataloging analytics procedures and for model management has emerged as a key issue for the machine learning community.
Activities in this area

• **Industry:** Tensorflow Hub (Google), MLFlow (DataBricks), AutoML, …

• **Common Model Infrastructure**
  - Workshop at upcoming IEEE International Conference on Data Mining (ICDM), Nov 8-11, 2019, Beijing, [https://sites.google.com/view/cmi-icdm2019/home](https://sites.google.com/view/cmi-icdm2019/home)

• **DLHub: Model and Data Serving for Science**
  - Ryan Chard, Zhuozhao Li, Kyle Chard, Logan Ward, Yadu Babuj, Anna Woodard, Steven Tuecke, Ben Blaiszik, Michael J. Franklin, and Ian Foster, ANL and University of Chicago, Chicago, IL.
  - Share, publish, verify, reproduce, and reuse models, and address concerns related to model reproducibility
  - Implement scalable and low-latency serving capabilities to leverage parallel and distributed computing resources to democratize access to published models through simple web interfaces