Common Model Infrastructure

XLDB 2018

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

• The Context
• The Opportunity
• The Questions – How to proceed…what are the challenges?
The Context

• NSF Big Ideas
  – 6 distinct research ideas

• Convergence
  – Deep integration of knowledge, techniques, and expertise
  – To form new and expanded frameworks for addressing scientific and societal challenges and opportunities.
  – Merging of distinct and diverse approaches into a unified whole to foster new paradigms or domains

• Ubiquity of data
  – ML, DL, data analytics models—combined with science-based models
The Technical context

• A number of schemes for capturing model information
  • PMML: Predictive Model Markup Language
  • PFA: Portable Format for Analytics
  • SBML: Systems Biology Markup Language
  • SED-ML: Simulation Experiment Description Markup Language

• Guidelines for recording information
  – Minimum Information About a Simulation Experiment, MIASME

• Resources
  – BioModels database, TensorFlow Hub
The Opportunity

• Big Ideas and Accelerators ➔ Project / Goal orientation
  – Not typical for NSF (which generally focuses on foundational/basic research)

• Budget commitment
  – Each Big Idea is allocated $30M/year
  – HDR is allocated an additional $30M/year for a “Convergence Accelerator”
    • Required to find matching $20M/year from industry, other agencies
The Opportunity…

• Big Ideas would benefit from shared model repositories
  – Build more complex models by composing other models; reduce unnecessary duplication of effort

• Existing NSF programs would benefit
  – E.g., Smart & Connected Communities; Secure and Trustworthy Cyberspace; Neuroscience programs; etc…

• Help with reproducibility
  – Several studies / workshops / papers on reproducibility
The Questions

• What should be the Research and Infrastructure Agenda?
• A data science research and infrastructure agenda in model discovery and reuse
  – Methodologies – for how to record experiments
  – Standards – to use in recording information
  – Tools - to help implement standards
  – Resources – to allow for storing and sharing of models, datasets
  – Also, model transparency, interpretability, reproducibility
The Questions: Challenges

• Incentives for…
  – “Doing it right the first time” – Use of proper methodology and standards
  – “Doing it right the second time” – Reuse
Upcoming events

• ACM SIGMOD 2nd Workshop on Data Management for End-to-End Machine Learning (DEEM), June 15, Houston, TX, [http://deem-workshop.org/](http://deem-workshop.org/)
  – Workshop chairs: Sebastian Schelter, Stephan Seufert, Amazon Research; Arun Kumar, UC San Diego