The NIST IAD Data Science Research Program (DSRP)

Bonnie Dorr
March 17-18, 2016
Data Science: Domain-Independent Solutions

Data

Traffic Data

Algorithms

Traffic Route Recommender Algorithms

Result

Improved Traffic Route Recommendations

Movie Ratings

Movie Recommender Algorithms

Improved Movie Recommendations
Data Science: Domain-Independent Solutions

Data

Traffic Data

Input Data\textsubscript{New}

Movie Ratings

Algorithms

Generalized Algorithms and Data Processing Techniques

Result

Improved Traffic Route

Improved Output\textsubscript{New}

Improved Recommendations
Goal: Improve the measurement and standards of the generalized algorithms and data processing techniques of data science
Data Science Research Program

**Evaluation & Metrology**
- Evaluation Series for Data Science
- Evaluation paradigms and metrics
- Evaluation-Driven Research
- Datasets, Tasks, Analysis Tools

**Standards**
- Standards development
- Big Data Best Practices
- Working Groups

**Compute Infrastructure**
- Data science cluster for evaluation
- Agile System Architecture
- Benchmarking Tools
- Novel T&E Approaches

**Community Outreach**
- Build Community of Interest
- Technical Symposiums
- Focus on Generalized DS Problems
Evaluation and Metrology

- Evaluation Series for Data Science (see next talk)
- Development of evaluation paradigms and metrics
- Evaluation-Driven Research
- Datasets, Tasks, Analysis Tools
Data Science Standards

- NIST Big Data Public Working Group (NBD-PWG)
- Standards form a framework for data scientists to work with data and infrastructure that involves swappable components
- Provision of capability to apply problems in data science more generally
Infrastructure: Data Science Cluster for Evaluation

- NIST data science cluster serves as an infrastructure that allows for easy deployment and reconfiguration of computational and storage needs.
- Provides a benchmarking framework to collect metrics on software systems submitted by participants.
- Reproducibility of given problems on different compute architectures.
Community Outreach

- NIST aims to use the DSRP to bring researchers together from multiple disciplines
- Evaluations and Symposia provide a way to encourage multidisciplinary collaboration across different research communities.

**Examples:**
- NIST Data Science Symposium (Mar’14)
- NIST Data Science Evaluation Series Workshop (Mar’16)
- NIST Data Science Evaluation Series Pilot Workshop (Fall’16)
DSRP Recap

- New DSRP and DSE series
  - Drive performance
  - Measure state of the art
  - Study generalizability of algorithms
  - Improve quality of DS and data analysis
- Measurement at both component and system level
- DSRP and DSE development is ongoing
- Feedback is welcome, please be interactive!