IOWA STATE UNIVERSITY

Department of Computer Science – Department of Statistics – Department of Mathematics – Department of Computer and Electrical Engineering

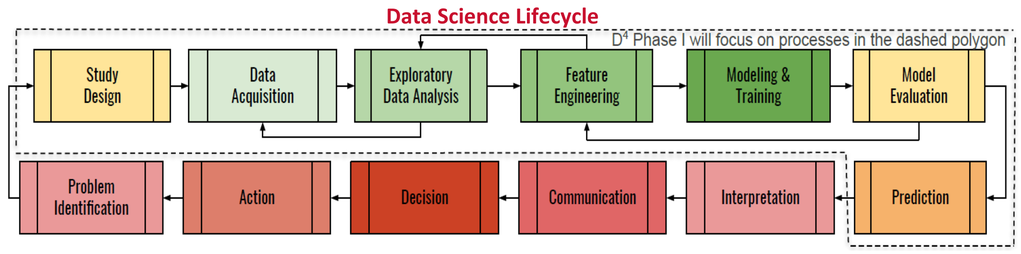
D4 (Dependable Data Driven Discovery) Institute

P.I.: Hridesh Rajan;
Co-P.I.s: Pavan Aduri, Chinmay Hegde, Daniel Nettleton, and Eric Weber;
Senior Personnel: Michael Catanzaro, Jia (Kevin) Liu, Henry Schenck, Vinodchandran N. Variyam, Namrata Vaswani, Lily Wang, and Zhengyuan Zhu

Goals: Dependable Discovery

- Advancing the foundations of dependable data driven discovery.
- Data science (DS) lifecycle is dependable, if there is a rigorous basis for justifiably trusting its output.
- Dependability is critical because unreliable discoveries can have catastrophic impacts.

Data Science Lifecycle

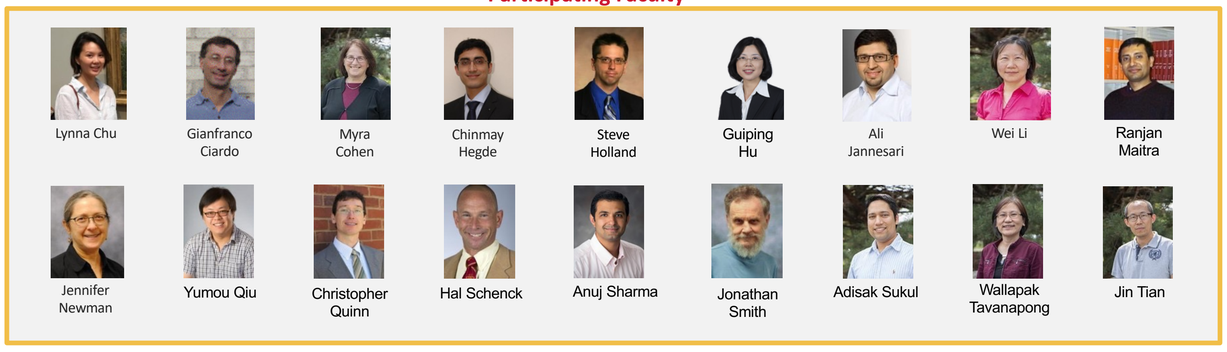


- D4 Institute is developing an overall framework for dependable data driven discovery
  - Parts: risks (What?), measures (How to quantify?), and mechanisms (How to mitigate?)
- Broader Impacts Include:
  - Establish and sustain cross-sector collaborations.
  - Create a hub for sharing data science expertise.
  - Educate researchers and practitioners in theoretical, applied and ELSEI (ethical, legal, social, economic, impacts) aspects.

D4 Team: CS, STAT, Math, and EE

Leadership

Participating Faculty



D4 Framework: Risks, Measures,and Mechanisms

Complexity

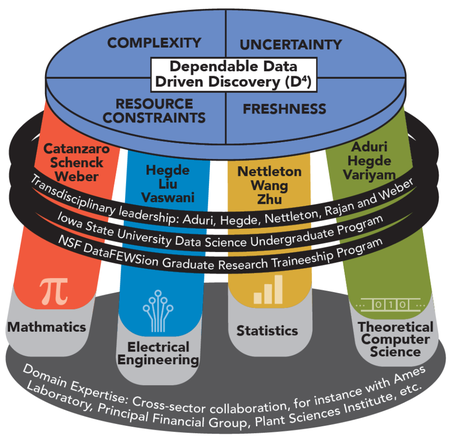
- Risk: modern data can be thought of as sample from unknown probability distribution, the sample spaces of these probability distributions are very large, and explicit description of these distributions is unavailable.
- Measures: sample complexity (# of observations) and computational resources (time and memory) taken by the statistical tests/algorithms.

We are exploring new research directions by postulating that many data distributions that arise in practice are in fact produced by computationally efficient processes.

Uncertainty

- Risk: For the most part, predictions that result from machine learning algorithms come with no information about how far each prediction may be from the true response targeted for prediction. Decision makers need more than a best guess to act appropriately when formulating actions.
- Measures: predictive distribution and uncertainty associated with results.

We are developing split-conformal inference to estimate probability distribution without increasing computational complexity and exploring a spectral representation of DNN.



Resource Constraints

- Risk: Learning and inferring of data distributions over graphical models is NP-hard, can be accelerated using parametric assumptions on shape and size of subgraphs, but naturally occurring graphs have irregular and asymmetric regions. All newer algorithms are extremely compute-expensive, incurring either quadratic ($O(n^2)$) running time or worse in terms of the size of the graph.
- Measures: time and space complexity.

We are designing resource-efficient algorithms, in particular algorithms with quasi-linear runtime, for various naturally occurring graph learning and inference problems.

Freshness

- Risk: In certain applications, such as automatic robotic control, remote healthcare and medical support, high-frequency trading, tracking social media trending topics, delayed/stale data can produce misleading predictions.
- Measures: Age of Information (AOI), at time t, the AoI at the destination is defined as t-u(t), where u(t) is the time-stamp of the latest received update at the destination, i.e., the time at which it was acquired at the source.

We are exploring holistic data sampling and transportation along with processing solutions to ensure data freshness in large-scale information networks.

Midwest Big Data Summer School

About

This year's Midwest Big Data Summer School was held between May 16-19, 2022. This summer school, the sixth iteration of what is shaping into a tradition for the Department of Computer Science, consisted of a four-day intensive curriculum designed to challenge and enrich attendees' knowledge of data science. Primarily aimed at early career researchers and practitioners to bolster data-driven research and development skills.

Objectives

- Learn about the data science pipeline: Data Acquisition, Data Preprocessing, Exploratory Data Analysis, Descriptive Data Analysis, Visualization and Communication, and Ethical Issues and Standards in Data Science.
- Combine these learned materials with keys and techniques used by statisticians, computer scientists, and data science researchers and apply them to real-world scenarios.

Sponsors

CCC Computing Community Consortium Catalyst
Midwest Big Data Hub
Iowa State University College of Liberal Arts and Sciences
Iowa State University Department of Statistics
workiva
Iowa State University Department of Computer Science
KINGLAND
Iowa State University Office of the Vice President for Research

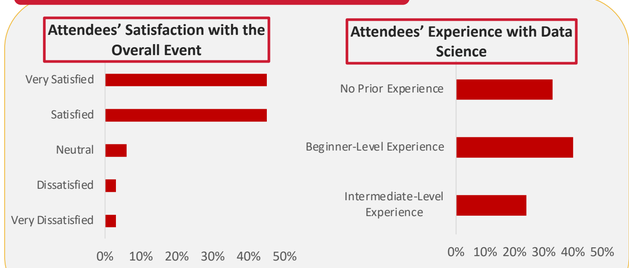
Overview of the Week

Monday	Tuesday	Wednesday	Thursday
Introduction to Data Science	A Hands-on Introduction to Deep Learning	Supervised Machine Learning Methods	Practical Data Science and Machine Learning on the Cloud

Tracks

Data Science for Public Good	Datamining and Machine Learning	Digital Agriculture	Foundations of Data Science

Past Attendee Data



Data collected from 144 participants, spanning 19 universities and 9 other organizations.

Team Building: TADS Lunch-n-Learn

About

TADS Lunch-n-Learn is held every Thursday each week, to provide a platform for building camaraderie and a shared vocabulary in theoretical and applied data science researchers.

Objectives

- Host a wide variety of presentations and talks to showcase recent results, upcoming funding opportunities, etc.
- Help participants meet up with other like-minded individuals to research funding opportunities and collaborate on papers and enquiry.
- Open discussion after presentation encourages participants to share their own thoughts or get their questions answered.

Past Presenters

Tamal Dey, Purdue University, *Computational Topology and Data Analysis: A New Way of Looking at Data.*

Yuliya Lierler, UN-Omaha, *Answer Set Programming and Automatic Optimization Methods in its Realm.*

Vinod Variyam, University of NE-Lincoln, *Learning and Sampling of Atomic Interventions from Observations.*

Lev Reyzin, UI-Chicago *Differential Privacy, Adaptive Data Analysis, and Free Speedups via Sampling.*

Aurore Delaigle, Melbourne University, *Estimating a Covariance Function from Fragments of Functional Data.*

Ness Shroff, Ohio State University, *Delay Optimality in Network Clouds via Load Balancing.*

Eduardo Blanco, North Texas University, *Towards Deeper Natural Language Understanding.*

Andrew McGregor, UM-Amherst, *Recent Results on Cycle Counting in the Data Stream Model.*

Vipin Kumar, Minnesota University, *Physics Guided Machine Learning: A New Framework for Accelerating Scientific Discovery.*

Helen Zhang, Arizona University, *Sparse and Smooth Function Estimation in Reproducing Kernel Hilbert Spaces.*

Stephen Wright, UW-Madison, *Second-order Methods for Nonconvex Optimization with Complexity Guarantees.*

Jane Cleland-Huang, Notre Dame, *Human-Drone Partnerships in Emergency Response.*

This work was supported in part by the US National Science Foundation under grant CCF-19-34884.
All opinions are of the authors and do not reflect the view of the NSF.