Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
  - Creating training data
  - Data cleaning
- Part IV. Conclusions and research direction

DI and ML: A Natural Synergy

- Data integration is one of the oldest problems in data management

- Transition from logic to probabilities revolutionized data integration
  - Probabilities allow us to reason about inherently noisy data
  - Similar to the AI-revolution in the 80s [https://vimeo.com/48195434]

- Modern machine learning and deep learning have the power to streamline DI

DI and ML: A Natural Synergy

- Data is bottleneck of modern ML and AI applications

- DI-related methods and algorithms have revolutionized the way supervision is performed.
    - Weak supervision signals are integrated into training datasets

- Data integration solutions (e.g., data cataloging solutions) can lead to cheaper collection of training data and more effective data enrichment

Opportunities for DI

One System vs. An Ecosystem: Every RBMS is a monolithic system. This paradigm has failed for DI. Tools for different DI tasks are prevalent. We need abstractions and execution frameworks for such ecosystems.

Humans-in-the-loop: DI tasks can be very complex. Is weak supervision the right approach to inject domain knowledge? What about quality evaluation?

Multi-modal DI: ML-based DI has focused on structured data with the exception of DI over images using crowdsourcing and some recent efforts that target textual data. DL is the de facto solution to reasoning about high dimensional data. Can is help develop unified DI solutions for visual, textual, and structured data?

Efficient Model Serving: This means efficient model serving. Many compute-intensive operations such as normalization and blocking are required. Featurization may also rely on compute-heavy tasks (e.g., computing string similarity). What is the role of pipelining and RDBMS-style optimizations?

Opportunities for ML

Data Catalogs: Data augmentation relies on data transformations performed on data records in a single dataset. How can we leverage data catalogs and data hubs to enable data augmentation go beyond a single dataset?

Valuable Data for ML applications: Our community has focused on assessing the value of data [Dong et al., VLDB’12, Koutris et al., JACM 2015]. These ideas are not pervasive to ML but if ML is to become a commodity [Jordan, 2018] we need methods to reason about the value of data.

DI for Benchmarks: Increasing efforts on creating manually curated benchmarks for ML. Current efforts rely on manual collection and curation. How can we leverage meta-data and existing DI solutions to automate such efforts?

“How reliable are our current measures of progress in machine learning?”
Do CIFAR-10 Classifiers Generalize to CIFAR-10?, Ben Recht et al., 2018

MLPerf

DI & ML as Synergy

- **ML for effective DI: AUTOMATION, AUTOMATION, AUTOMATION**
    - Automating DI tasks with training data
    - Ensemble learning and deep learning provide promising solutions
    - Better understanding of semantics by neural network

- **DI for effective ML: DATA, DATA, DATA**
    - The software 2.0 stack is data hungry
    - Create large-scale training datasets from different sources
    - Cleaning of data used for training

Thank you!