Data Management for Deep Learning

Prof. Lei CHEN

Data Science and Analytics (DSA) Thrust - Information Hub
The Hong Kong University of Science and Technology (GZ)

Outline

- Background and Motivation
- Technical Challenges
- Our Recent Research
- Beyond DB for AI
- Summary

Background: AI Applications are Ubiquitous

- AI has made a huge success over the past years

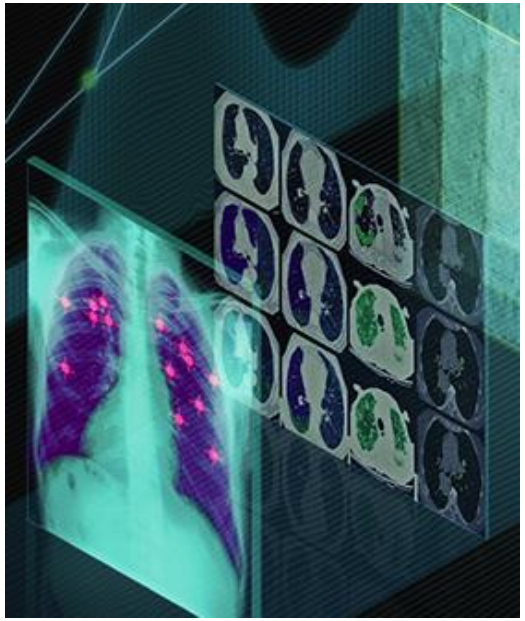
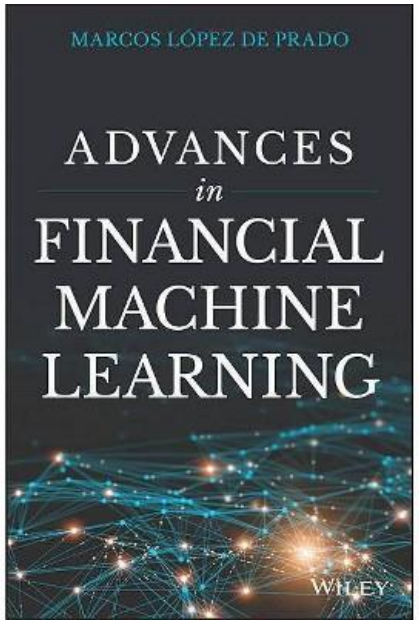


Image Recognition

Large Language Model



Smart Finance

Google Maps

Uber

DiDi

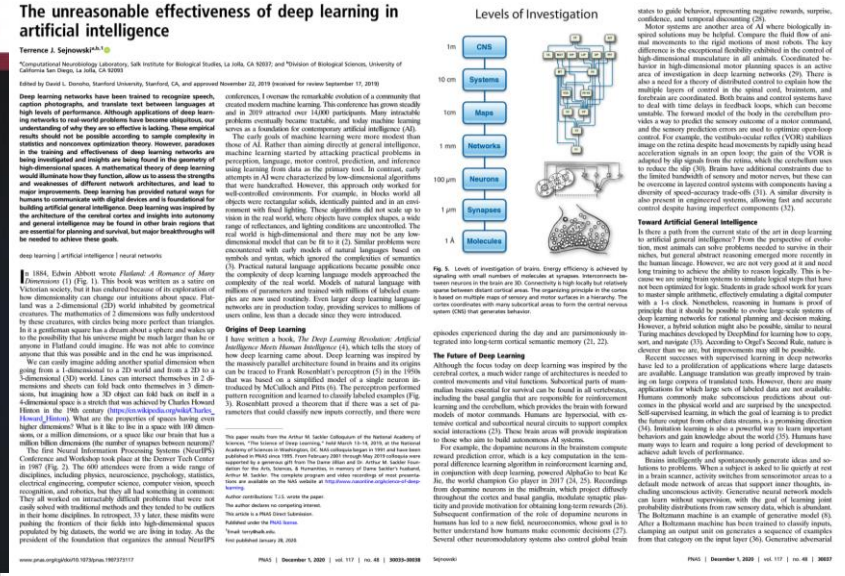
More than a journey

Intelligent Transportation

Background: Data is the New Oil

- The first secret of AI's success: big data

"The world's most valuable resource is no longer oil, but data". -- The Economist, 2017


The unreasonable effectiveness of deep learning in artificial intelligence

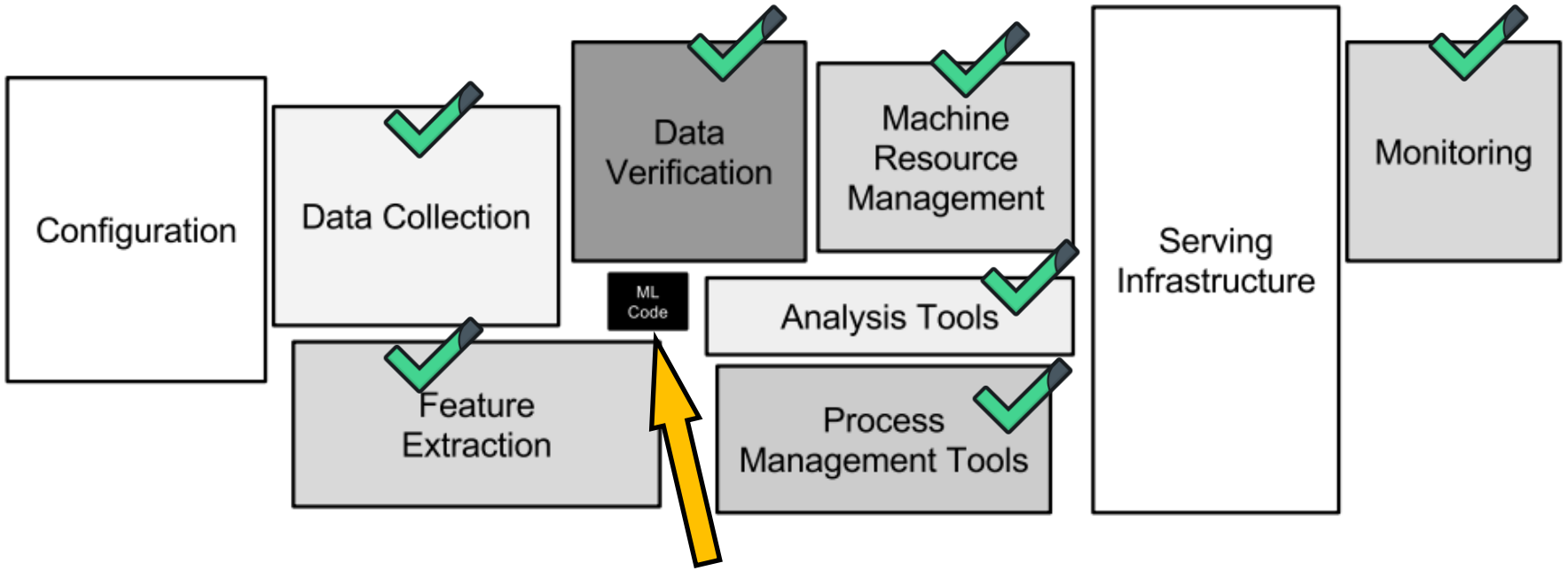
Terrence J. Sejnowski

Levels of Investigation

"Recent successes in deep networks have led to a proliferation of applications where large datasets are available". -- Terrence J. Sejnowski, in PNAS 2020

Motivation: Why Data Management for AI?

✓: related to data management

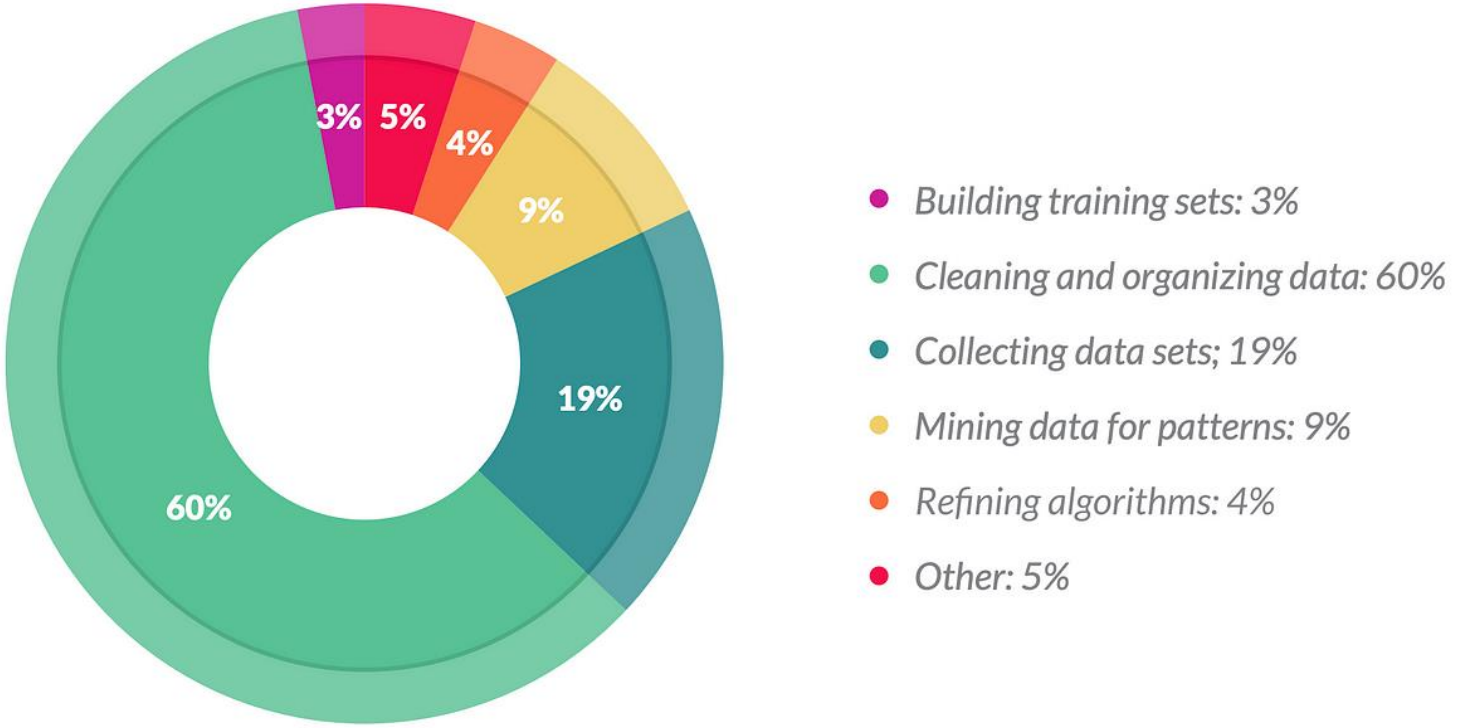


"In Google, only a tiny fraction of the code in many ML systems is actually devoted to learning."

D. Sculley, Gary Holt, Daniel Golovin et al. Hidden Technical Debt in Machine Learning Systems. In NeurIPS 2015.

Motivation: Why Data Management for AI?

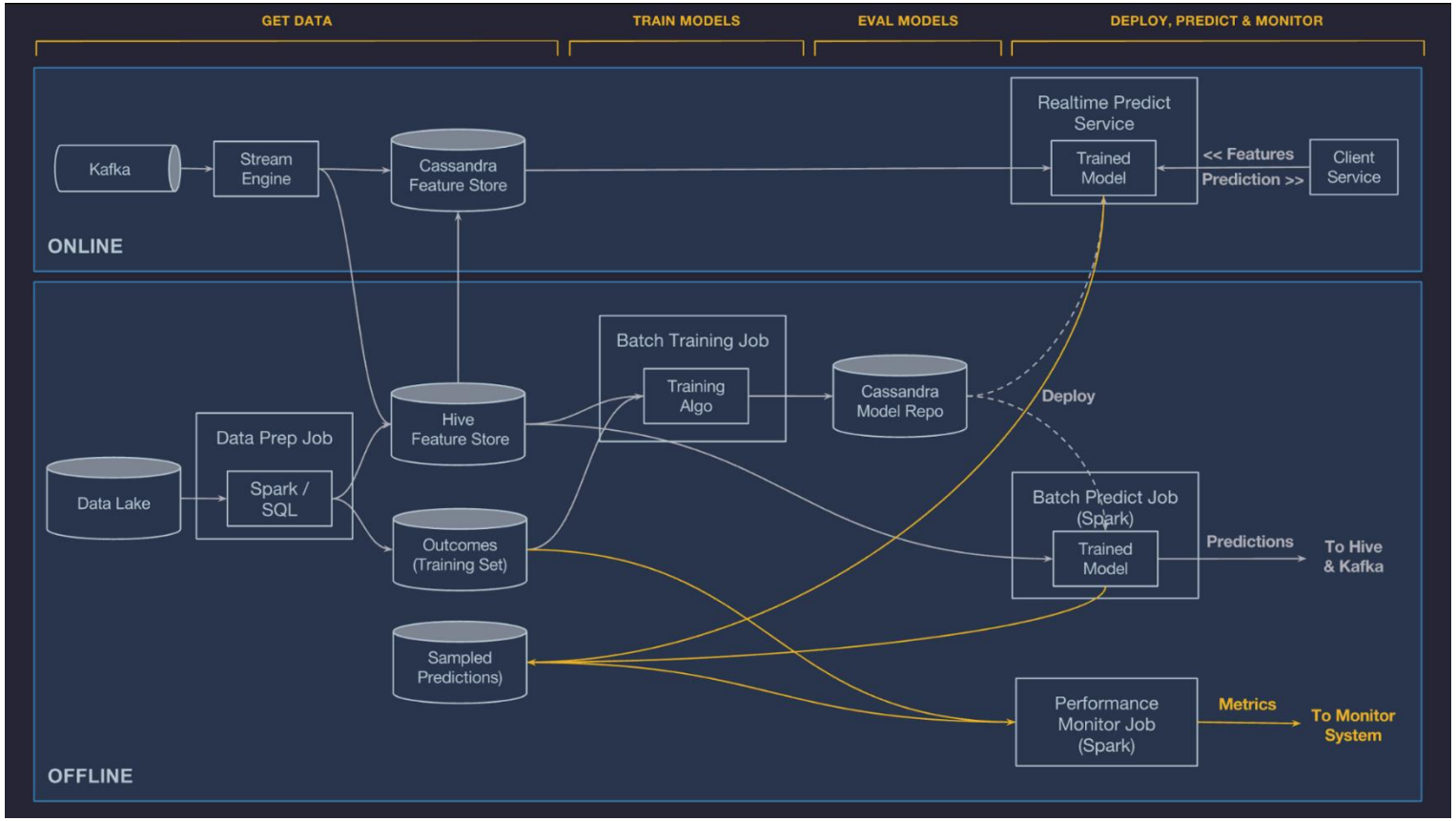
- What data scientists spend the most time doing?



"80% of ML users' time/effort (often more) spent on data issues!"

Motivation: Why Data Management for AI?

- Michelangelo: Uber's Machine Learning Platform



"Building and managing data pipelines is typically one of the most costly pieces of a complete machine learning solution."

Benefits of Data Management for AI

- Key concerns in AI:
  - Accuracy
  - Runtime efficiency

- Additional key practical concerns in AI systems:
  - Scalability (and efficiency at scale)
  - Usability
  - Manageability
  - Developability

} *Long-standing concerns in the DB systems world!*

Can often trade off accuracy a bit to gain on the rest!

Outline

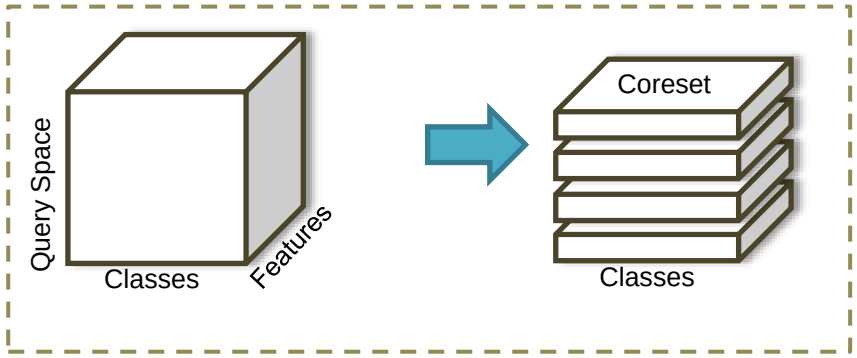
- Background and Motivation
- Technical Challenges
- Our Recent Research
- Beyond DB for AI
- Summary

Challenges: Data Management for AI

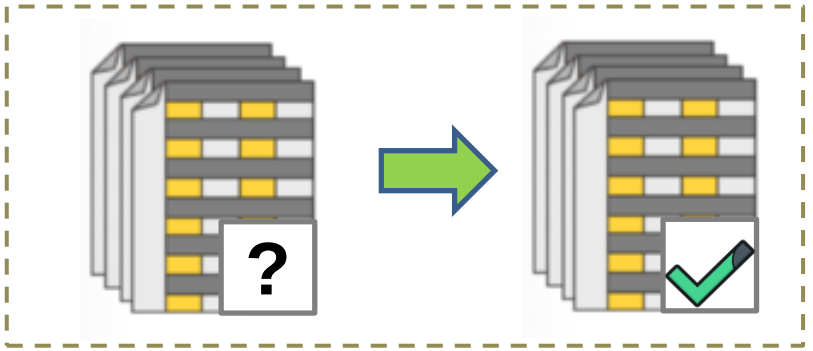
- "Data management …": How to organize, query, scale, and manage the analysis of large and complex datasets?
- "… for AI": three fundamental challenges
  - Data preparation
  - Optimized model training and inference
  - Model validation and explanation

Challenge 1: Data Preparation

- Why is data preparation crucial for AI?
  - The success of DL relies on massive high-quality data
- Key issues in data preparation
  - Data integration
  - Data labeling
  - Data selection
  - ……



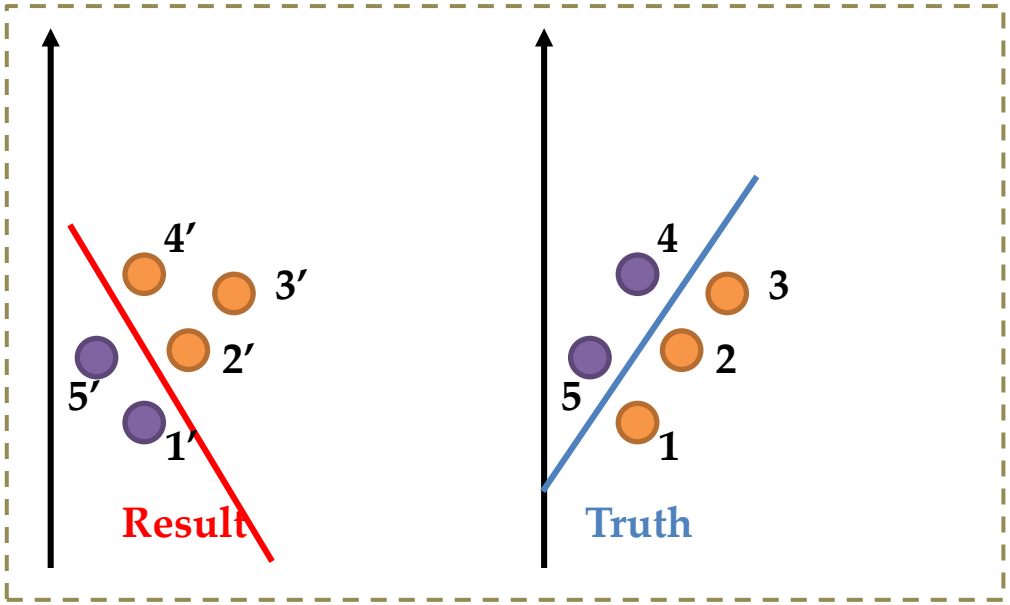
Coreset Selection and Compression



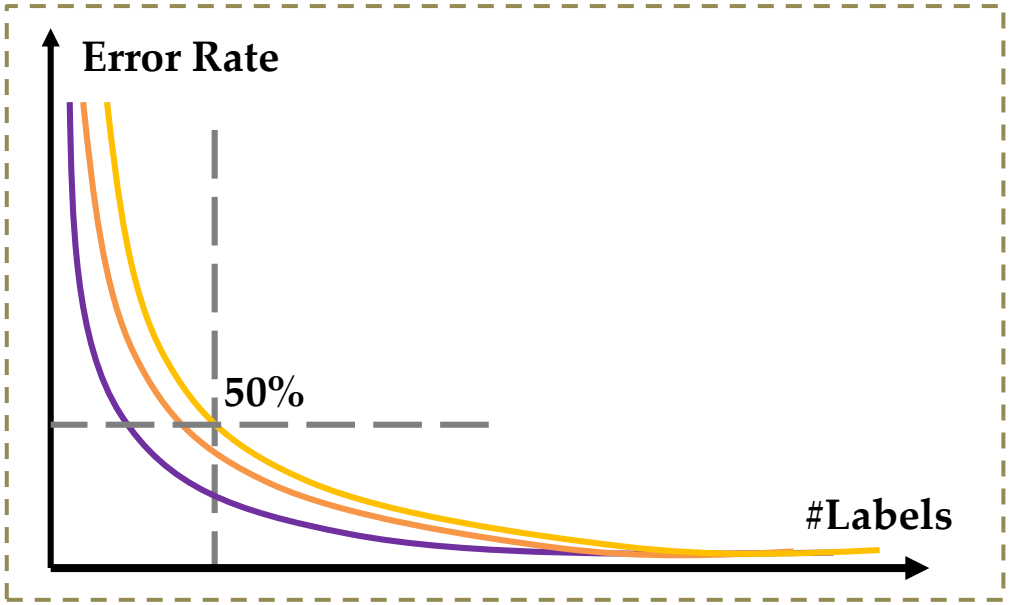
Semi-automatic Data Labeling

Challenge 1: Data Preparation

- What will happen in AI when data preparation is poor?
  - Eg, noisy data, insufficient data, etc.



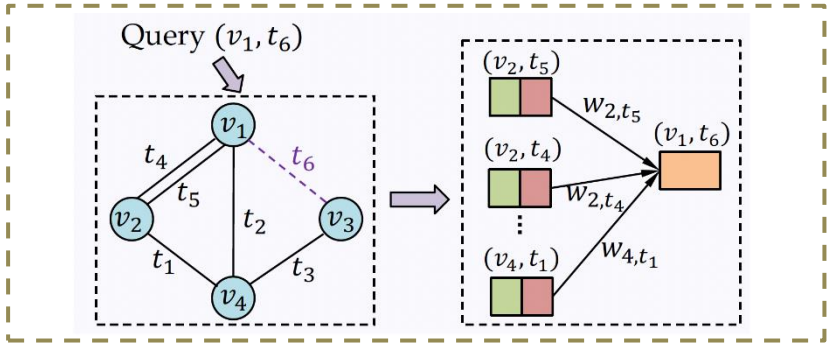
Noisy data can result in low accuracy



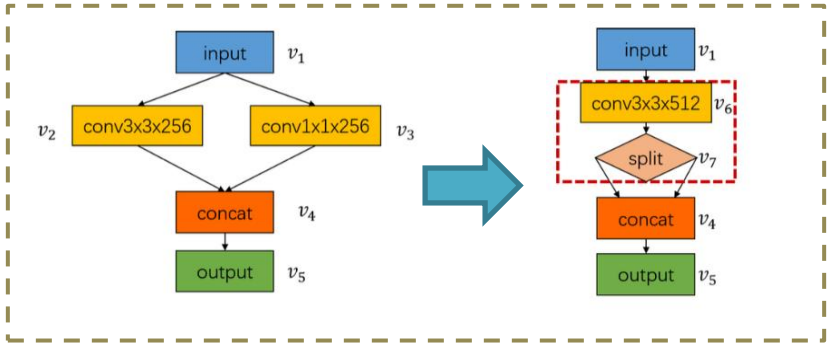
Insufficient data can result in low accuracy

Challenge 2: Optimized Training and Inference

- Why is optimized training and inference crucial for AI?
  - Optimized execution improves the efficiency and scalability of AI models
- Key issues in optimized training and inference
  - Approximate computation
  - Data placement and scheduling
  - Graph and operator optimization
  - ……



Approximate Query Processing



Graph and Operator Optimization

Challenge 2: Optimized Training and Inference

- What will happen if there is too much data or the model size is too large?
  - Heavy computation and strong hardware requirements
  - Data access can become the computational bottleneck
  - Slow model training and inference

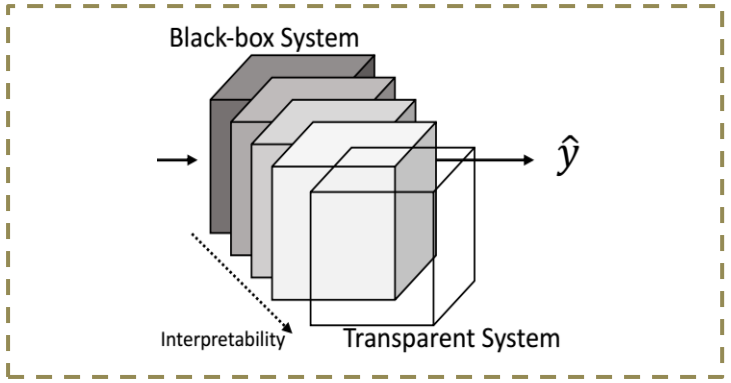
	GPT-3 large	LLaMa
Vocabulary size	50,257	32,000
Sequence length	2048	2048
Parameters in the largest model trained	175B	65B
Tokens in the training dataset	300B	1 – 3T
Number of GPUs	10,000 V100 GPUs	2048 A100GPUs
Training time	One month	21 days

Sam Altman stated that *"the cost of training GPT-4 was more than $100 million"*.

Challenge 3: Model Validation and Explanation

- Why is model validation and explanation crucial for AI?
  - Model validation ensures the effectiveness of the AI model
  - Model explanation improves the transparency of the AI model

Result Validation



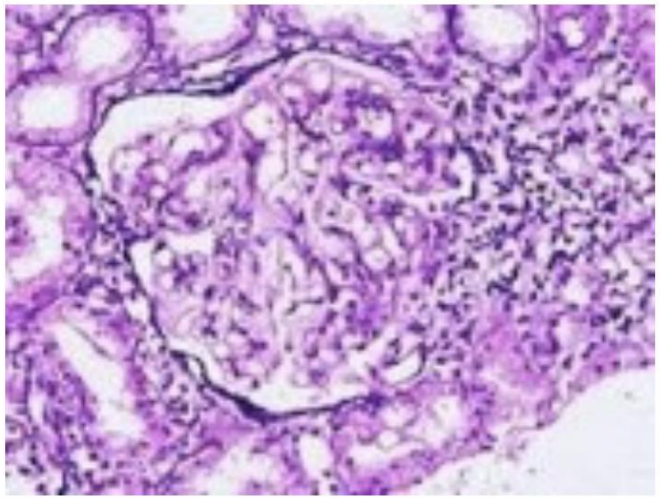
Result Explanation

Challenge 3: Model Validation and Explanation

- What will happen when AI model is a black box?
  - Wrong decision can be dangerous for critical systems

"Autonomous car crashes, because it wrongly recognizes …"

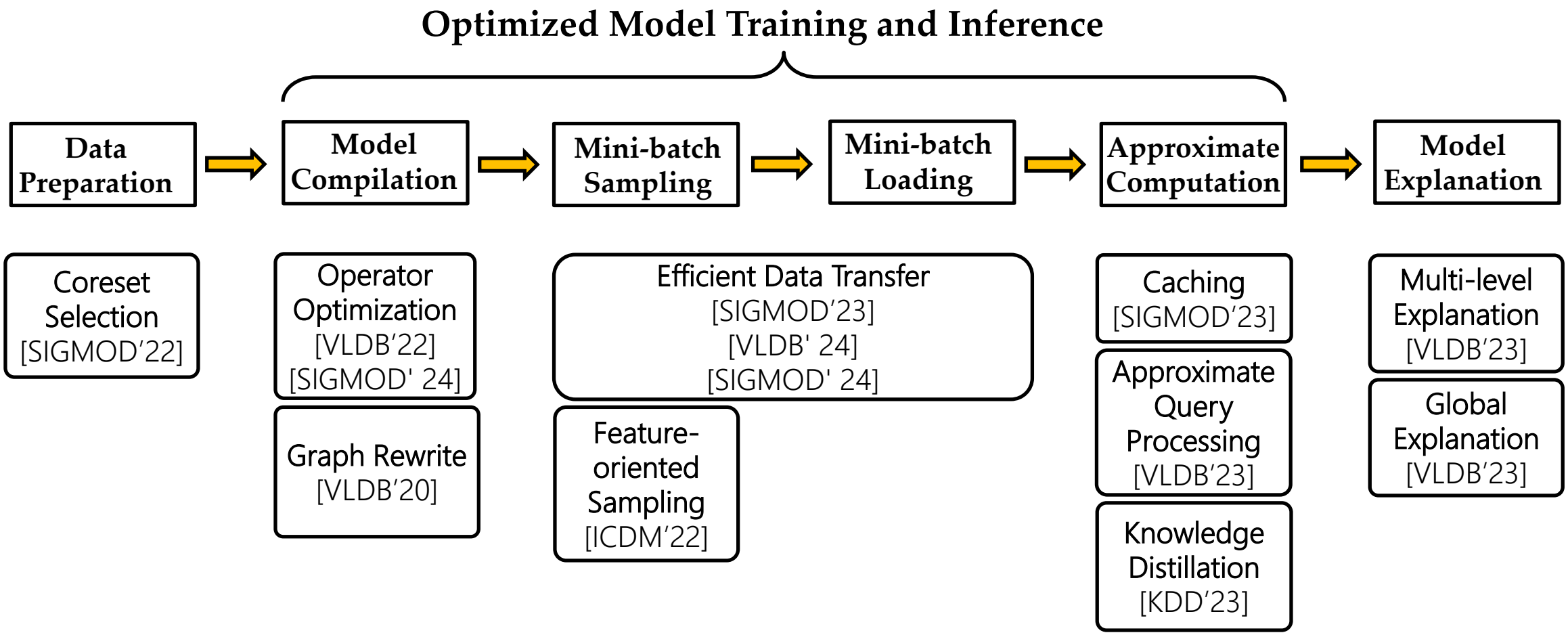
"AI medical diagnosis system misclassifies patient's disease …"



Outline

- Background and Motivation
- Technical Challenges
- Our Recent Research
- Beyond DB for AI
- Summary

Overview of Our Research



Publications

- SIGMOD (5):
  - Camel: Managing Data for Efficient Stream Learning. [SIGMOD'22]
  - Orca: Scalable Temporal Graph Neural Network Training with Theoretical Guarantees. [SIGMOD'23]
  - DUCATI: A Dual-Cache Training System for Graph Neural Networks on Giant Graphs with the GPU. [SIGMOD'23]
  - STile: Searching Hybrid Sparse Formats for Sparse Deep Learning Operator Automatically. [SIGMOD'24]
  - SIMPLE: Efficient Temporal Graph Neural Network Training at Scale with Dynamic Data Placement. [SIGMOD'24]
- VLDB (7):
  - Optimizing DNN Computation Graph using Graph Substitutions. [VLDB'20]
  - ETO: Accelerating Optimization of DNN Operators by High-Performance Tensor Program Reuse. [VLDB'22]
  - SANCUS: Staleness-Aware Communication-Avoiding Full-Graph Decentralized Training in Large-Scale Graph Neural Networks. [VLDB'22] (Best Research Paper Award 2022)
  - Zebra: When Temporal Graph Neural Networks Meet Temporal Personalized PageRank. [VLDB'23]
  - On Data-Aware Global Explainability of Graph Neural Networks. [VLDB'23]
  - HENCE-X: Toward Heterogeneity-Agnostic Multi-Level Explainability for Deep Graph Networks. [VLDB'23]
  - ETC: Efficient Training of Temporal Graph Neural Networks over Large-scale Dynamic Graphs. [VLDB'24]
- KDD (1):
  - Narrow the Input Mismatch in Deep Graph Neural Network Distillation. [KDD'23]
- ICDM (1):
  - Feature-Oriented Sampling for Fast and Scalable GNN Training. [ICDM'22]

Summary

- Three fundamental research areas in data management for AI
  - Data preparation => massive high-quality data
  - Optimized model training and inference => efficiency and scalability
  - Model validation and explanation => effectiveness and transparency

- Data Cleaning
- Data Extraction
- Data Labeling
- Data Compression
- Data Selection
- ……

- Approximate Computation
- Caching and Sampling
- Data Placement
- Distributed Computation
- Model Compilation
- ……

- Local Explanation
- Global Explanation
- Multi-level Explanation
- ……

Laboratory

Theme Labs

- Big Data Institute (BDI)

Joint Labs

- Metaverse Joint Innovation Laboratory
- HKUST (GZ) -Tencent SSV Innovation Joint Lab for Incl
- HKUST (GZ) - Ambiping Joint Medical Data Lab
- HKUST (GZ) - Chuanglin Graph Data Joint Laboratory

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

安必平医疗数据智能联合实验中心

HKUST(GZ)-LBP Medical Data Intelligence Joint-Lab

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

DB4AI @ HKUST

- Members


Yiming Li


Jingzhi Fang


Ge Lv


Xin Zhang

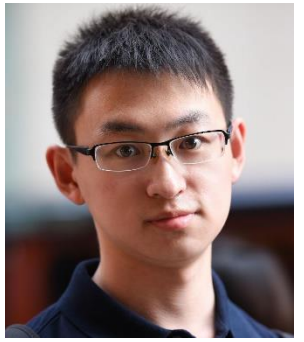

Qiqi Zhou


Shihong Gao

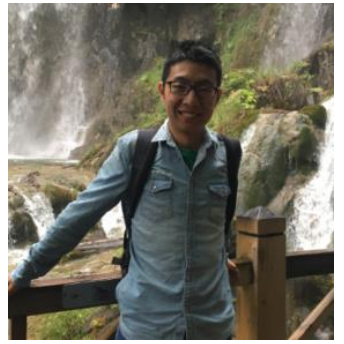
- Collaborators


Yanyan Shen


Yingxia Shao


Yue Wang


Yuxiang Zeng


Qiyu Liu

VLDB 2024 Welcome to **Guangzhou**

August 26-30, 2024