



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

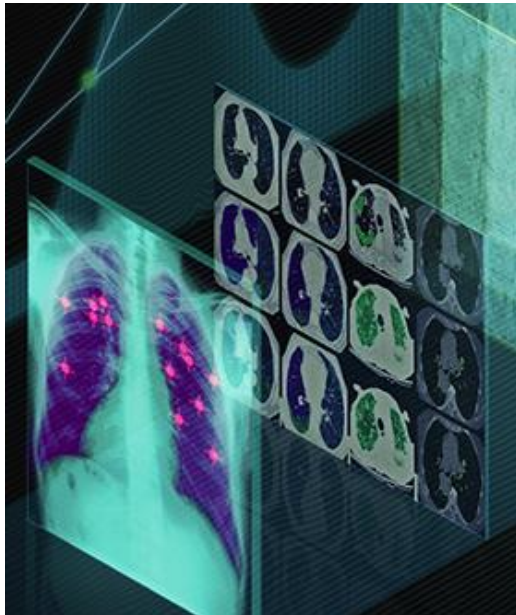
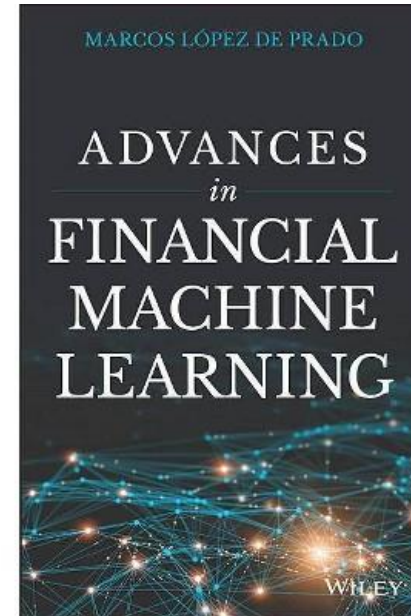


Image
Recognition



Large Language
Model



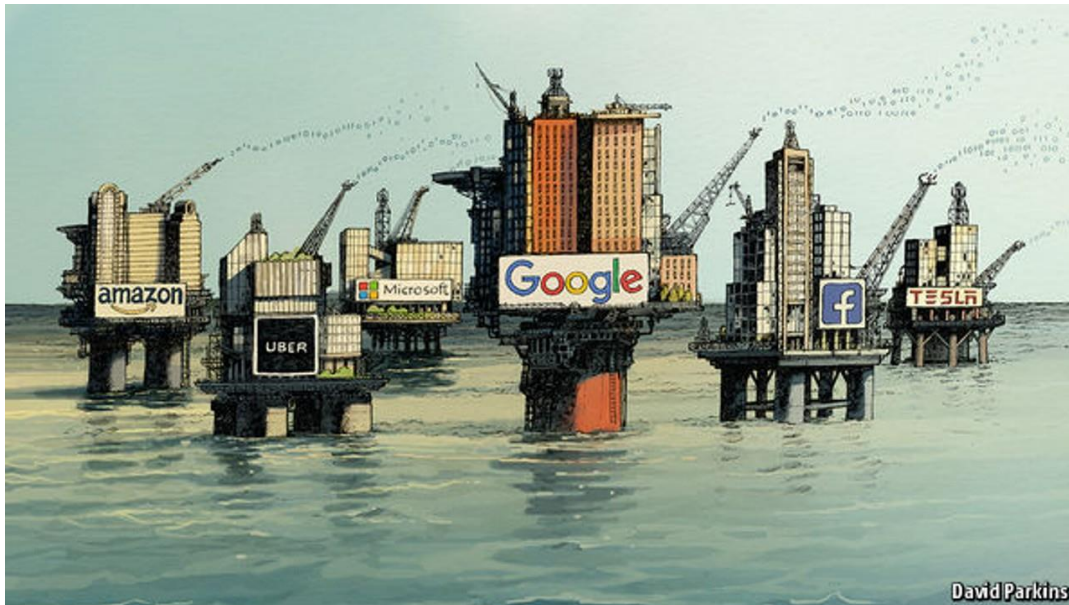
Smart
Finance



Intelligent
Transportation

Background: Data is the New Oil

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



PNAS
The unreasonable effectiveness of deep learning in artificial intelligence
Terrence J. Sejnowski^{1,2*}

¹Department of Neurobiology, University of California, San Diego, La Jolla, CA 92037, and ²Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92037

Edited by David L. Donchin, Stanford University, Stanford, CA, and approved November 22, 2019 (received for review September 11, 2019)

Deep learning networks have been trained to recognize speech, caption photographs, and translate text between languages at high levels of performance. Although applications of deep learning networks to real-world problems have become ubiquitous, our understanding of why they are so effective is lacking. These remarkable results should not be possible according to simple complexity arguments in the learning and effectiveness of deep learning networks are being investigated and insights are being found in the geometry of high-dimensional spaces. A mathematical theory of deep learning is being developed that explains how they function, allow us to assess the strengths and weaknesses of different network architectures, and lead to major improvements. Deep learning has provided natural ways for humans to communicate with digital entities and is foundational for building artificial general intelligence. Deep learning was inspired by the architecture of the cerebral cortex and taught into autonomy and general intelligence may be found in other brain regions that are essential for planning and learned, but major breakthroughs will be needed to achieve these goals.

Origins of Deep Learning
In 1984, Edwin Abbott wrote *Flatland: A Romance of Many Dimensions* (1) (Fig. 1). This book was written as a satire on Victorian society, but it has endured because of its exploration of low dimensionality and its implications about space. Flatland was a 2-dimensional (2D) world inhabited by geometrical creatures. The mathematics of 2 dimensions was fully understood by these creatures, with circles being more perfect than triangles. In a 2D world, a square has a direct edge to its center, and it is possible that this square might be much larger than the circle in Flatland could imagine. He was not able to conceive of anything that this was possible and in the end he was imprisoned. We can easily imagine adding another spatial dimension when going from a 1-dimensional to a 2D world and from a 2D to a 3-dimensional (3D) world. Lines can intersect themselves in 2 dimensions, but imagine how a 3D object can fold back on itself in a 4-dimensional space a stretch that was achieved by Charles Howard Hinton in the 19th century (https://en.wikipedia.org/wiki/Charles_Howard_Hinton). What are the properties of space having one higher dimension? What is it like to live in a space with 10 dimensions or a million dimensions, or a space that our brain that has a million billion dimensions, the number of synapses between neurons? The first Neural Information Processing Systems (NeurIPS) Conference, the premier conference in artificial intelligence (AI) and machine learning took place at the Denver Tech Center in 1987 (Fig. 2). The 600 attendees were from a wide range of disciplines, including physics, neuroscience, psychology, statistics, electrical engineering, computer science, computer vision, speech recognition, and robotics, but they all had something in common: They all worked on intractably difficult problems that were not only solved by traditional methods and they tended to be outliers in their own disciplines. In retrospect, 35 years, these results were pushing the frontiers of their fields into high-dimensional spaces predicted by the dreamers, the world we are living in today. As the president of the foundation that organizes the annual NeurIPS

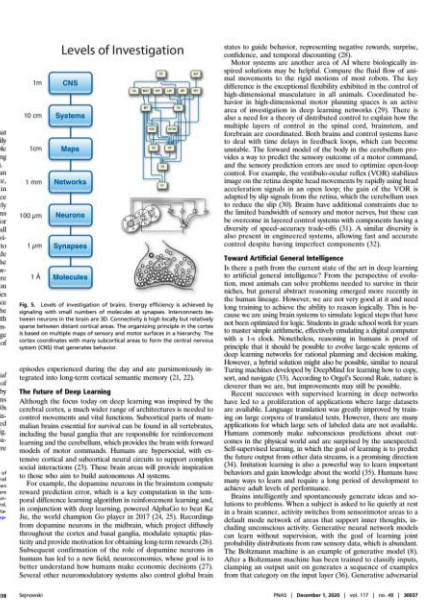
conferences, I review the remarkable evolution of a community that moved machine learning. This conference has grown steadily and in 2019 attracted over 14,000 participants. Many intractable problems eventually became tractable, and today machine learning serves as a foundation for contemporary artificial intelligence (AI). The early goals of machine learning were more modest than those of AI. Rather than aiming directly at general intelligence, machine learning started by attacking practical problems in perception, language, motor control, prediction, and inference using learning from data as the primary goal. In contrast, early attempts in AI were characterized by low-dimensional algorithms that were handcrafted. However, this approach only worked for well-controlled environments. For example, in checkers, which did not scale up to chess in the real world, where pieces have complex shapes, a wide range of reflections, and lighting conditions are uncontrolled. The real world is high-dimensional and there can be an almost infinite dimensional model that can be fit to it (2). Similar problems were encountered with early models of natural language based on syntax and syntax, which ignored the complexities of semantics (3). Practical natural language applications became possible once the complexity of deep learning models approached the complexity of the real world. Models of natural language with millions of parameters and trained with millions of labeled examples are now used routinely. Even larger deep learning language models are in production today, providing services to millions of users online, but their scale since they were introduced.

Origins of Deep Learning
I have written a book, *The Deep Learning Revolution: Artificial Intelligence Meets Human Intelligence* (4), which tells the story of how deep learning came about. Deep learning was inspired by the massively parallel architecture found in brains and its origin can be traced to Frank Rosenblatt's perceptron (5) in the 1950s that was based on a simplified model of a single neuron introduced by McCulloch and Pitts (6). The perceptron performed pattern recognition and learned to classify labeled examples (Fig. 3). Rosenblatt proved a theorem that if there was a set of parameters that could classify new inputs correctly, and there were

¹The main results from the author of the book *The Deep Learning Revolution: Artificial Intelligence Meets Human Intelligence* (4) are available on the author's website at www.sejnowski.com. ²For a list of attendees, see the book's website at www.sejnowski.com. ³For a list of attendees, see the book's website at www.sejnowski.com. ⁴For a list of attendees, see the book's website at www.sejnowski.com. ⁵For a list of attendees, see the book's website at www.sejnowski.com. ⁶For a list of attendees, see the book's website at www.sejnowski.com.

Author contributions: T.J.S. wrote the paper.
The author declares no competing interest.
This article is a U.S. Government work and, as such, is in the public domain in the United States of America.

PNAS | December 1, 2020 | vol. 117 | no. 48 | 30853-30858
www.pnas.org/cgi/doi/10.1073/pnas.1902321117

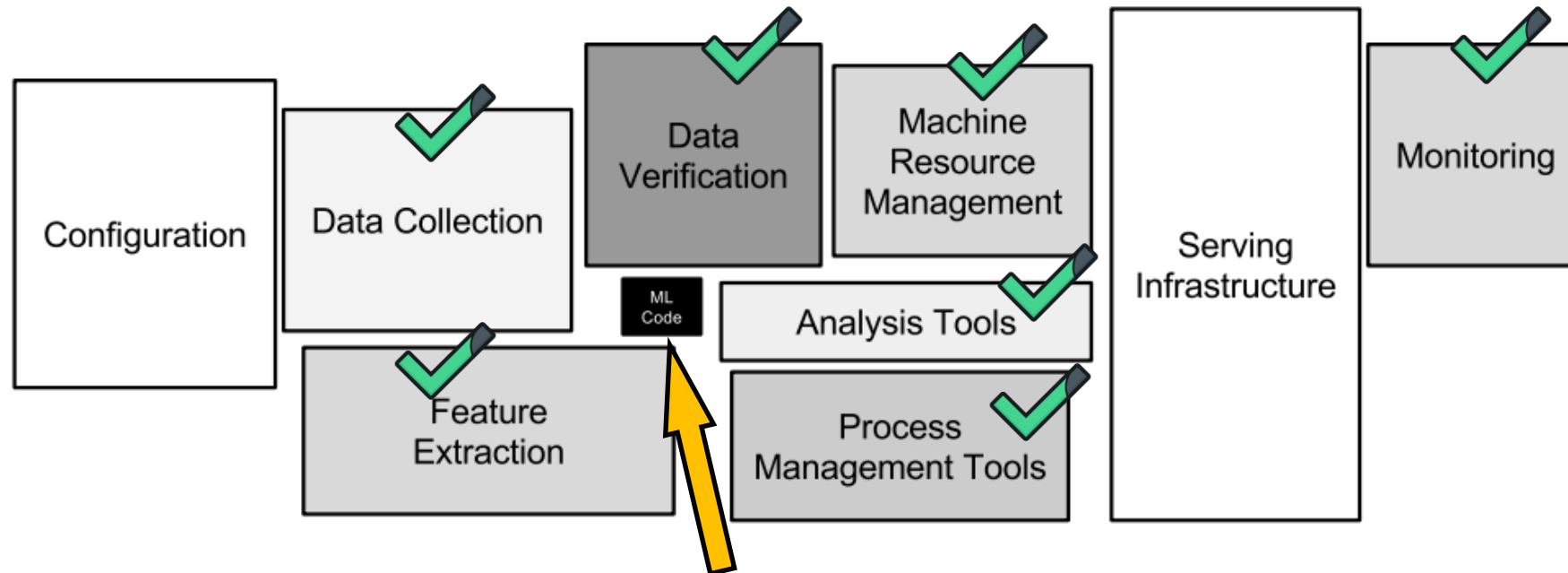


“The world’s most valuable resource is no longer oil, but data”. -- The Economist, 2017

“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?

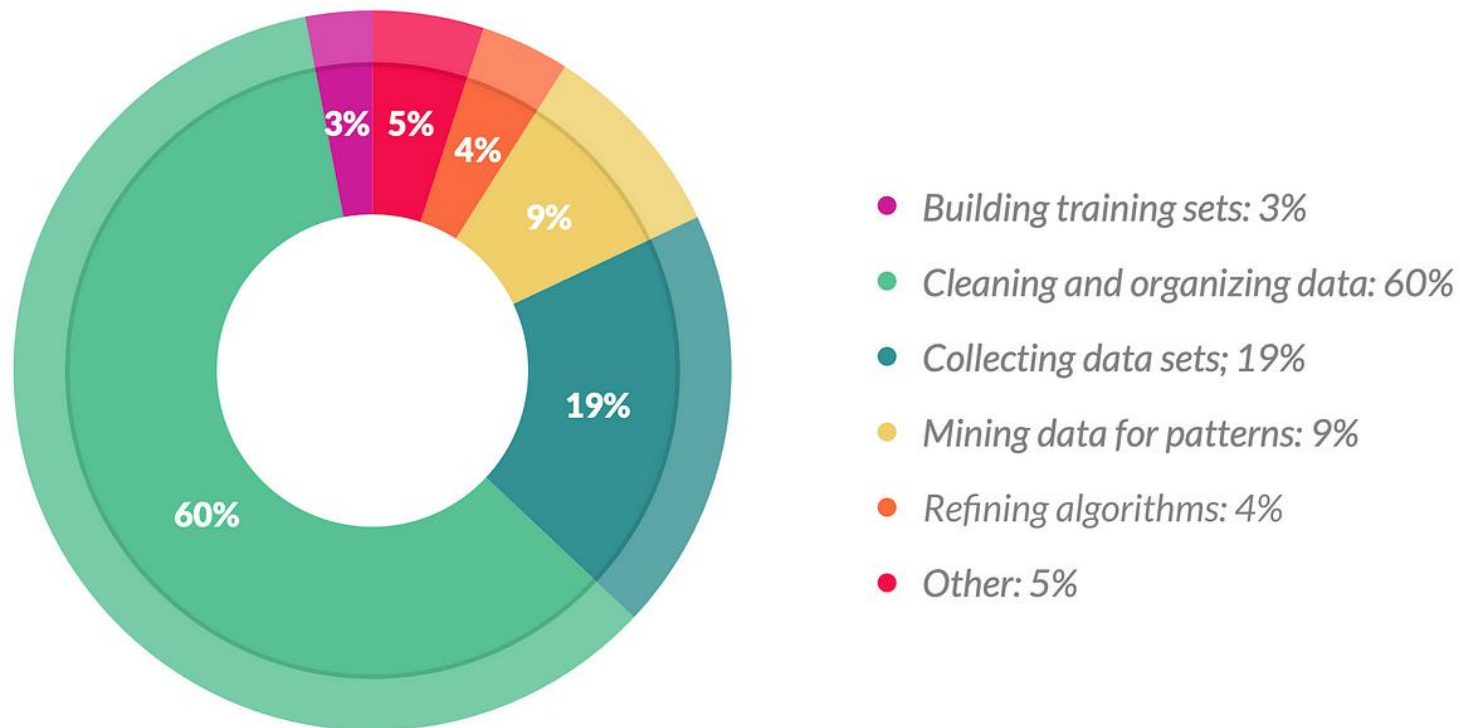
✓ : related to data management



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

Motivation: Why Data Management for AI?

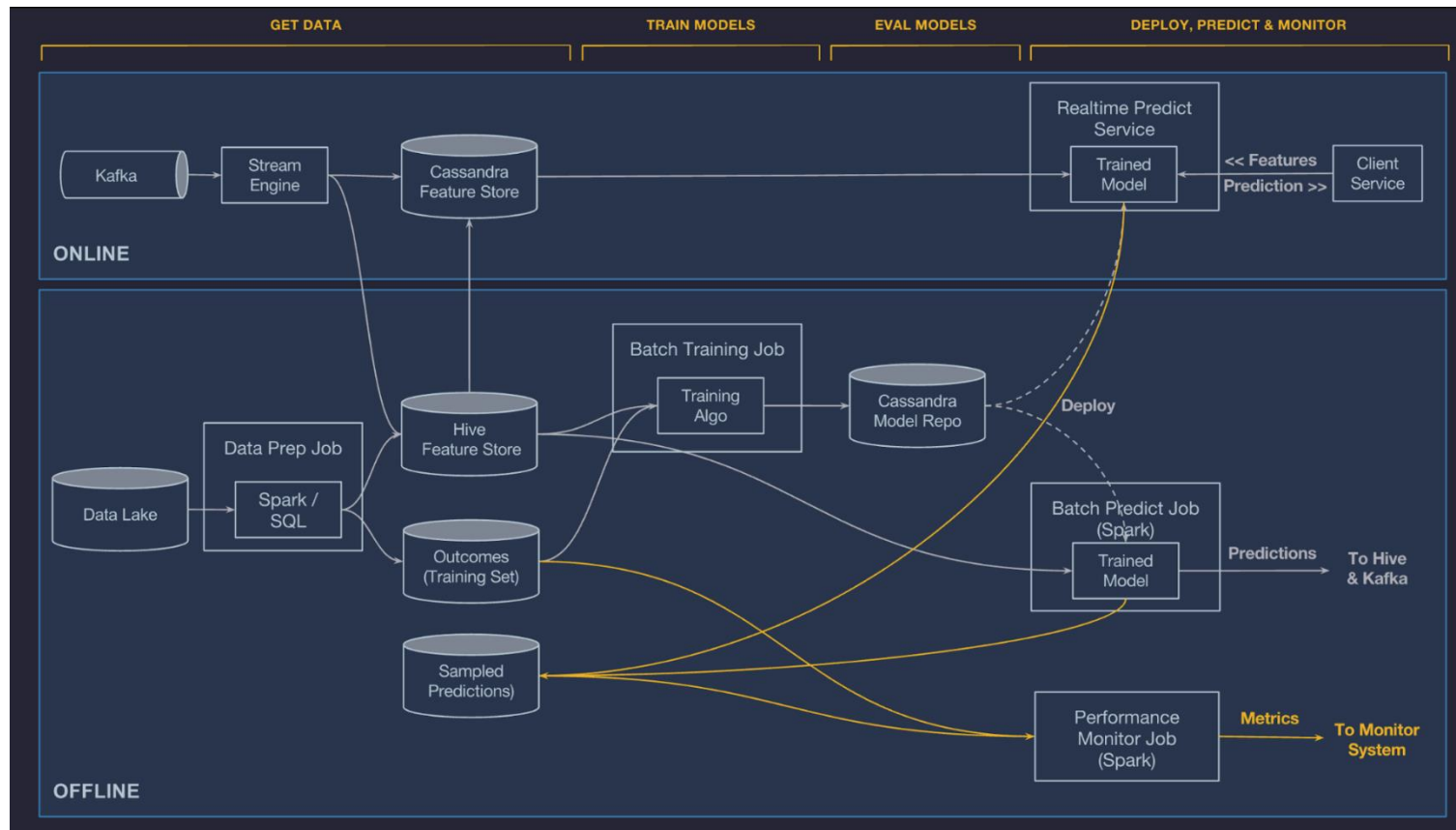
- What data scientists spend the most time doing?



*“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

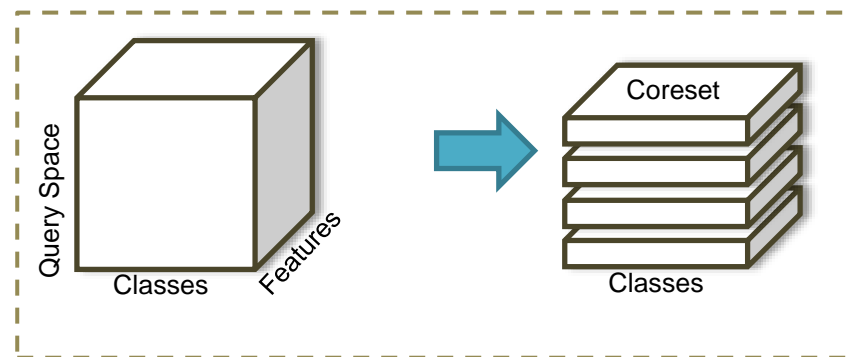
- Background and Motivation
- **Technical Challenges**
- Our Recent Research
- Beyond DB for AI
- Summary

Challenges: Data Management for AI

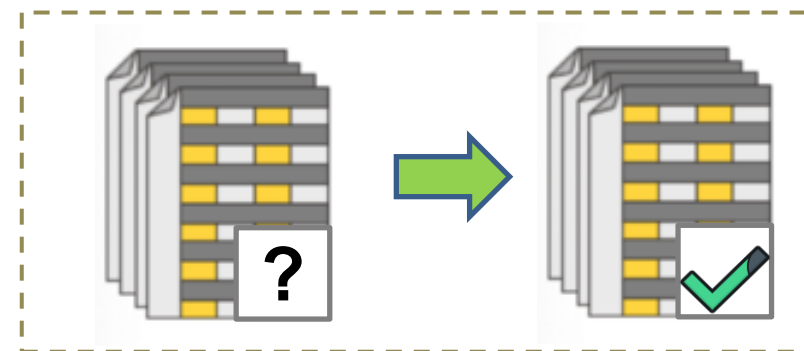
- “Data management ...”: How to organize, query, scale, and manage the analysis of large and complex datasets?
- “... for AI”: three fundamental challenges
 - Data preparation
 - Optimized model training and inference
 - Model validation and explanation

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
 -



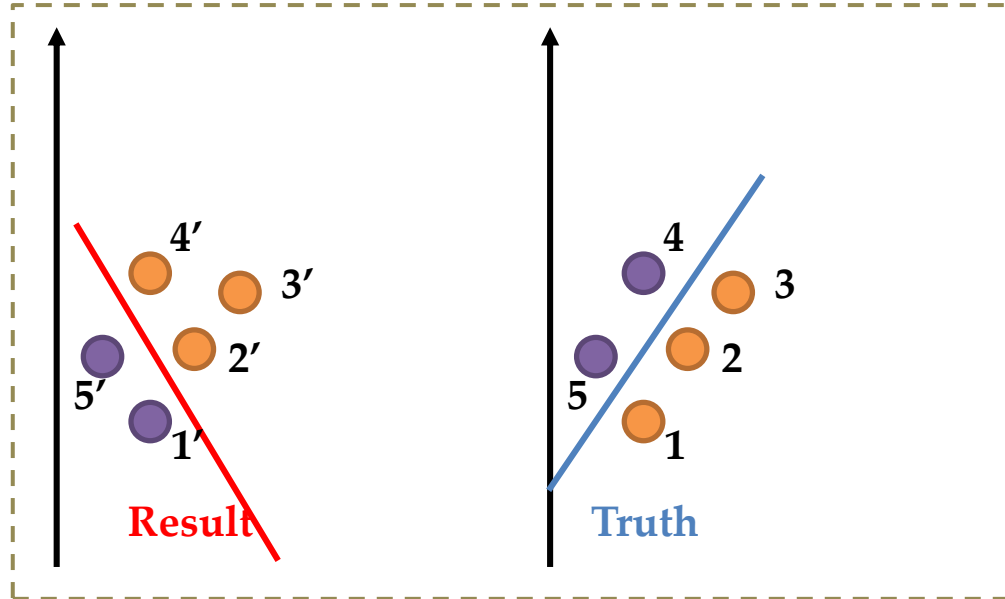
Coreset Selection and Compression



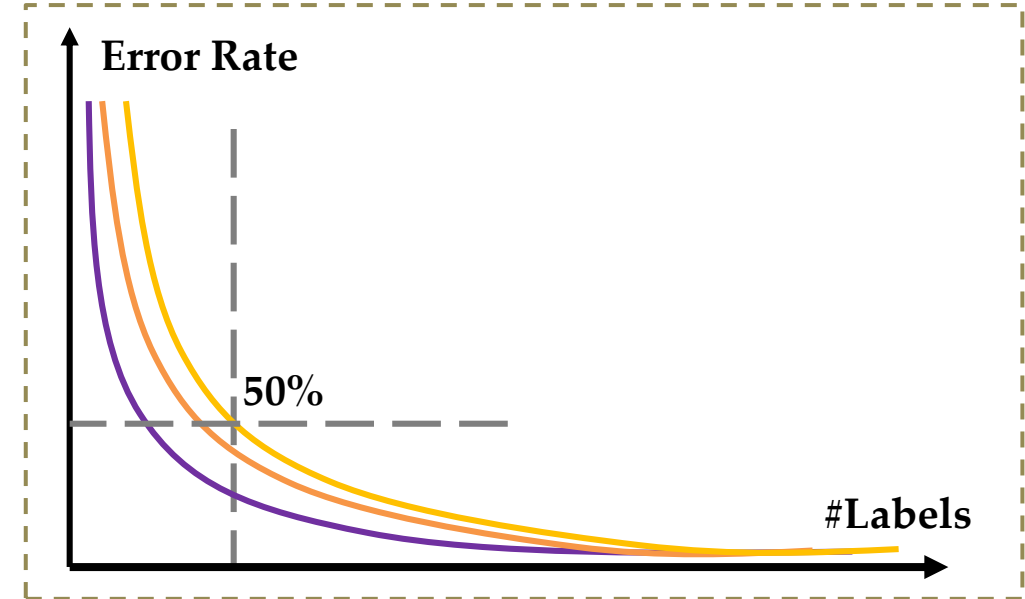
Semi-automatic Data Labeling

Challenge 1: Data Preparation

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



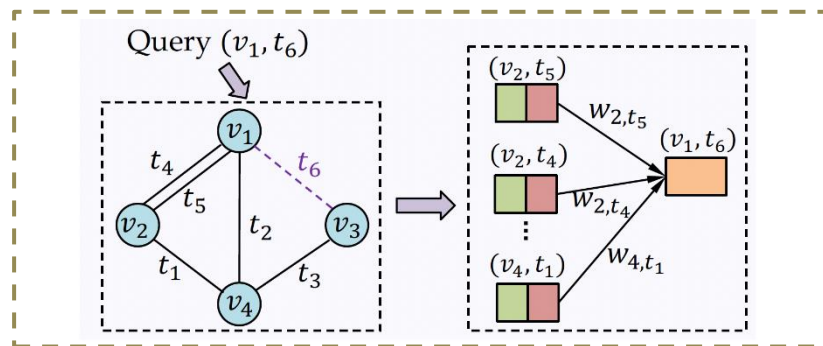
Noisy data can result in low accuracy



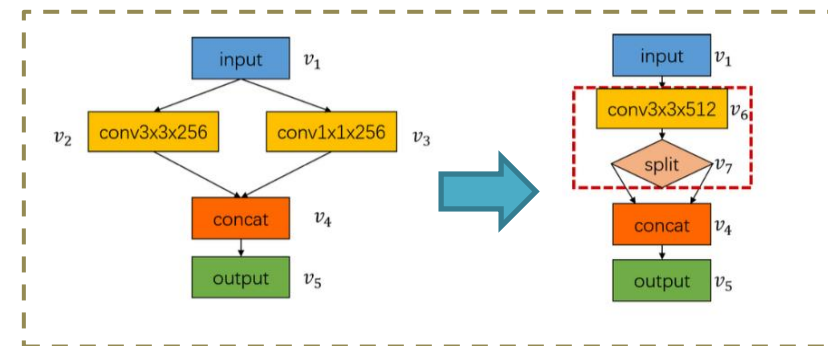
Insufficient data can result in low accuracy

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 |

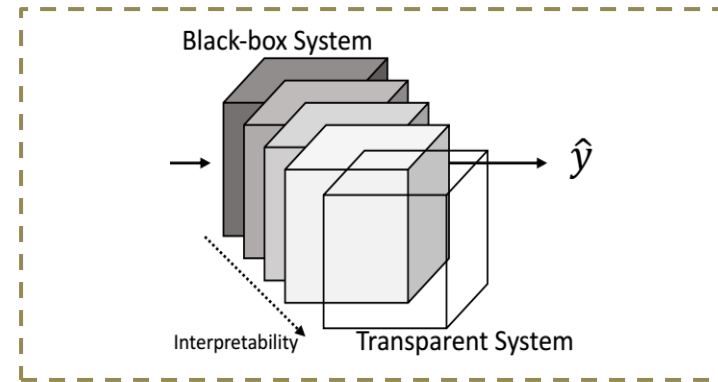
Sam Altman stated that *“the cost of training GPT-4 was more than \$100 million”*.

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

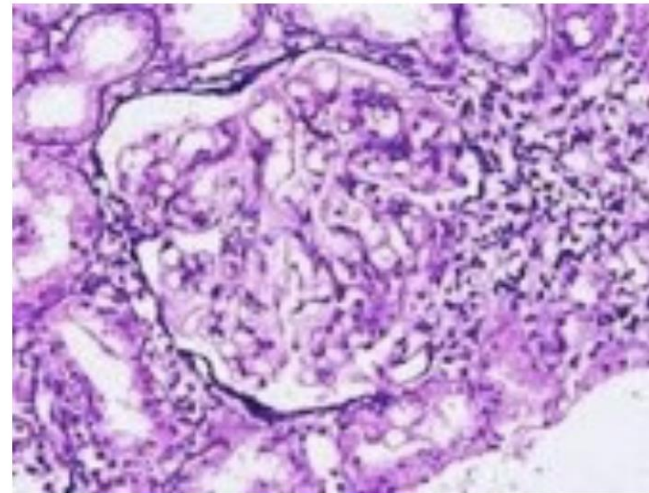
Challenge 3: Model Validation and Explanation

- 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 ...”*

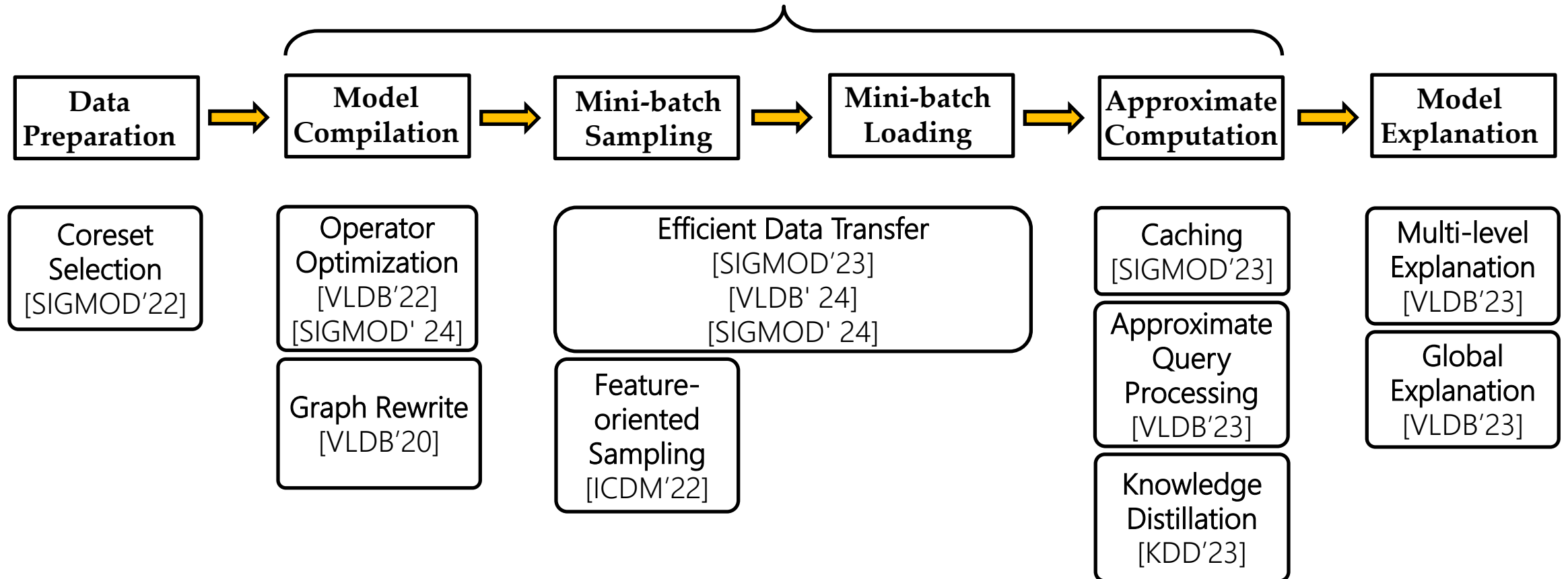


Outline

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

Overview of Our Research

Optimized Model Training and Inference

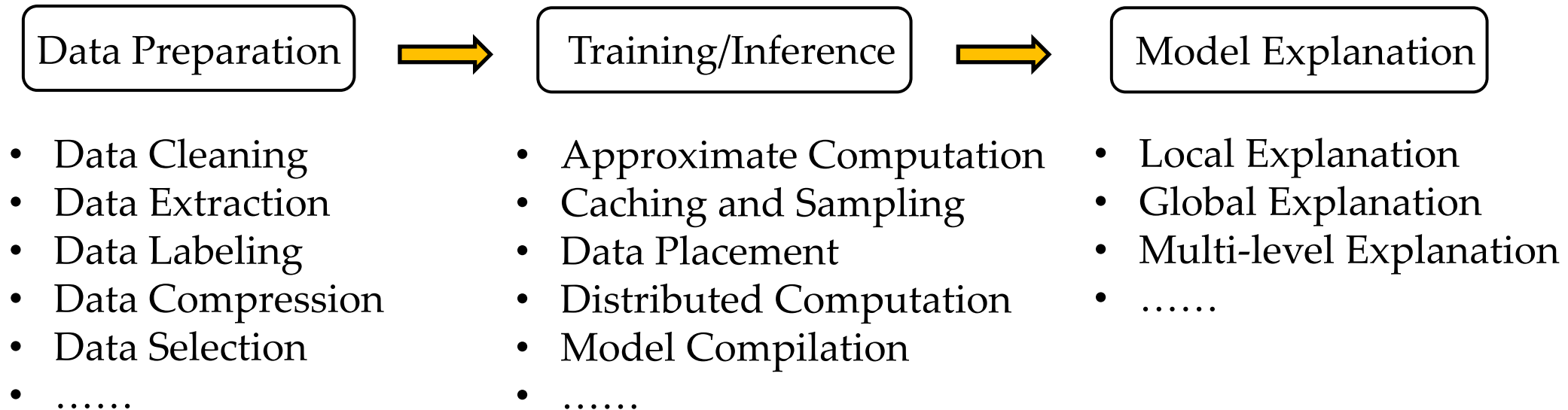


Publications

- 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



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



安必平医疗数据智能联合实验中心

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

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