DB for AI:
Data Management for Deep Learning

Lei Chen
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Outline

• **Background and Motivation**
• Technical Challenges
• Our Research Studies
  • Knowledge Extraction and Labelling
  • Graph substitutions on DNN computation graphs
  • Explainable Recommendation and Explainable GNN
• Beyond DB for AI
Background: DL applications are ubiquitous

- DL has made a huge **success** over the past years.
Background: Data is the new oil

• The first secret of DL’s success: *big data*

“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 DL?

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

Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.
Motivation: Why data management for DL?

“80% of ML users’ time/effort (often more) spent on data issues!”
Background: **Big Data, Deep learning and AI**

- If data is viewed as the oil, DL is the engine and AI is the car.
Challenges: data management for DL

#1: Data Preparation for DL
1.1 Data Extraction and Integration
1.2 Semi-automatic Data Labeling

#2: Optimized Training in DL
2.1 Coreset Selection and Compression
2.2 Graph and Operator Optimization

#3: Result Validation and Explanation in DL
3.1 Result Validation
3.2 Result Explanation