香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

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

- **<u>Background and Motivation</u>**
- Technical Challenges
- Our Reasearch 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.

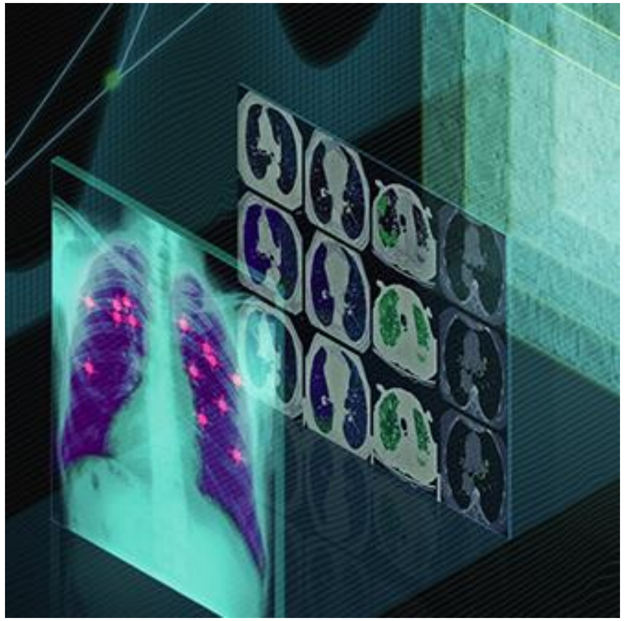
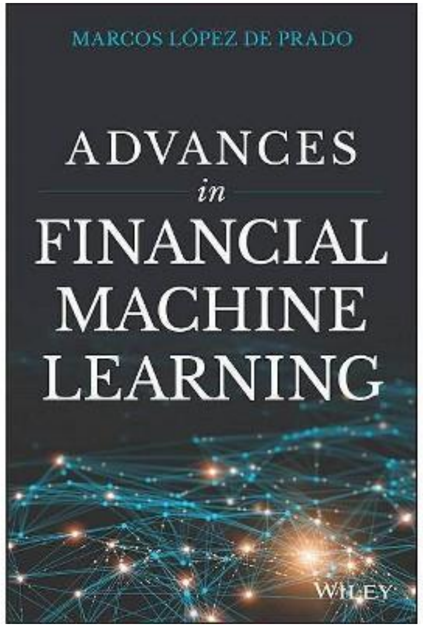


Image Recognition

Natural Language Processing



MARCOS LÓPEZ DE PRADO

ADVANCES *in* FINANCIAL MACHINE LEARNING

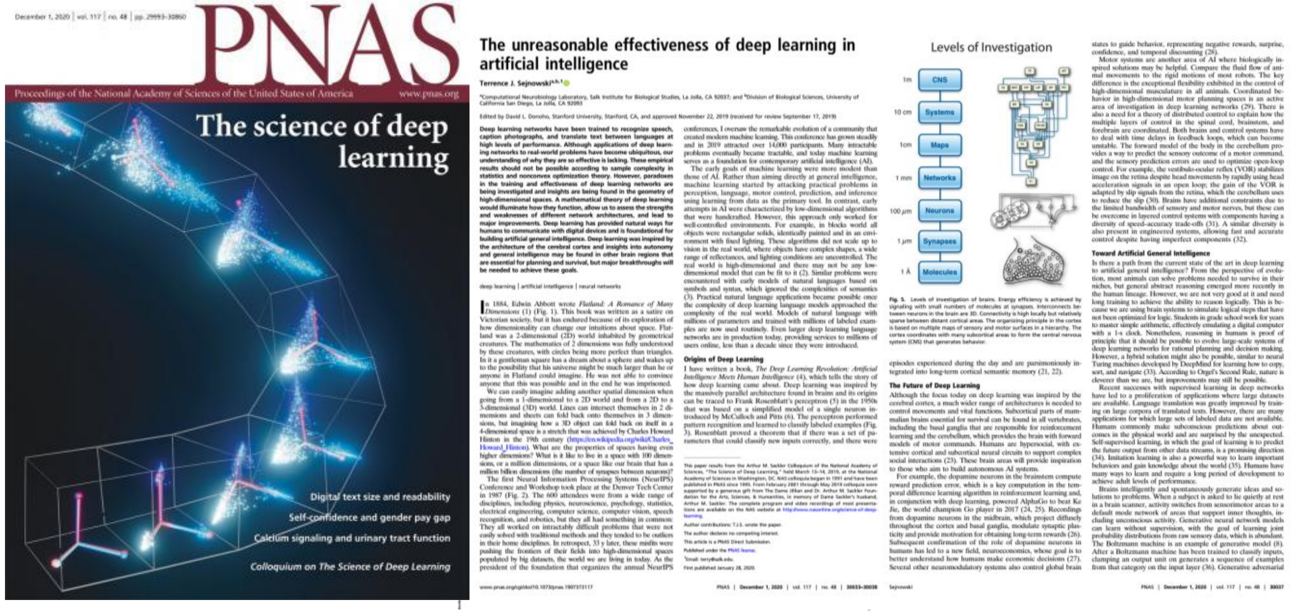
WILEY

Smart Finance

Intelligent Transportation

Background: Data is the new oil

- The first secret of DL's success: *big data*


December 1, 2020 | vol. 117 | no. 48 | pp. 29993–30860

PNAS

Proceedings of the National Academy of Sciences of the United States of America www.pnas.org

The science of deep learning

Digital text size and readability

Self-confidence and gender pay gap

Calcium signaling and urinary tract function

Colloquium on The Science of Deep Learning

The unreasonable effectiveness of deep learning in artificial intelligence

Terrence J. Sejnowski[a,b,1]

[a]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; and [b]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved November 22, 2019 (received for review September 17, 2019)

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

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

Motivation: Why data management for DL?

✓ : related to data management

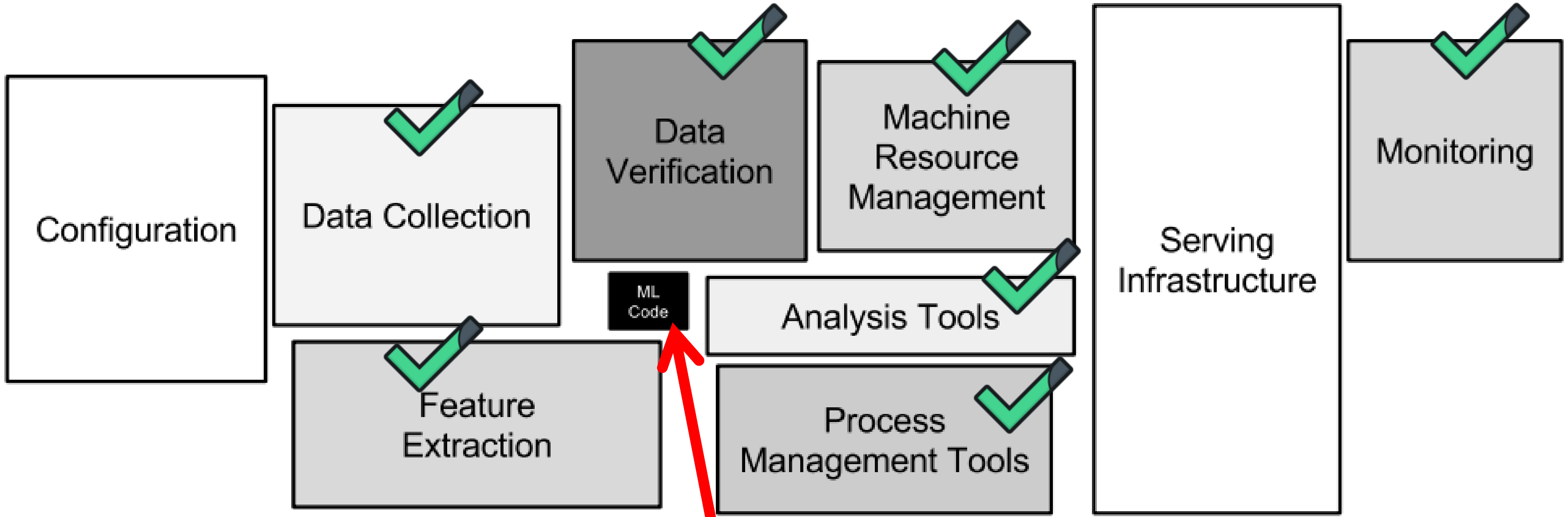
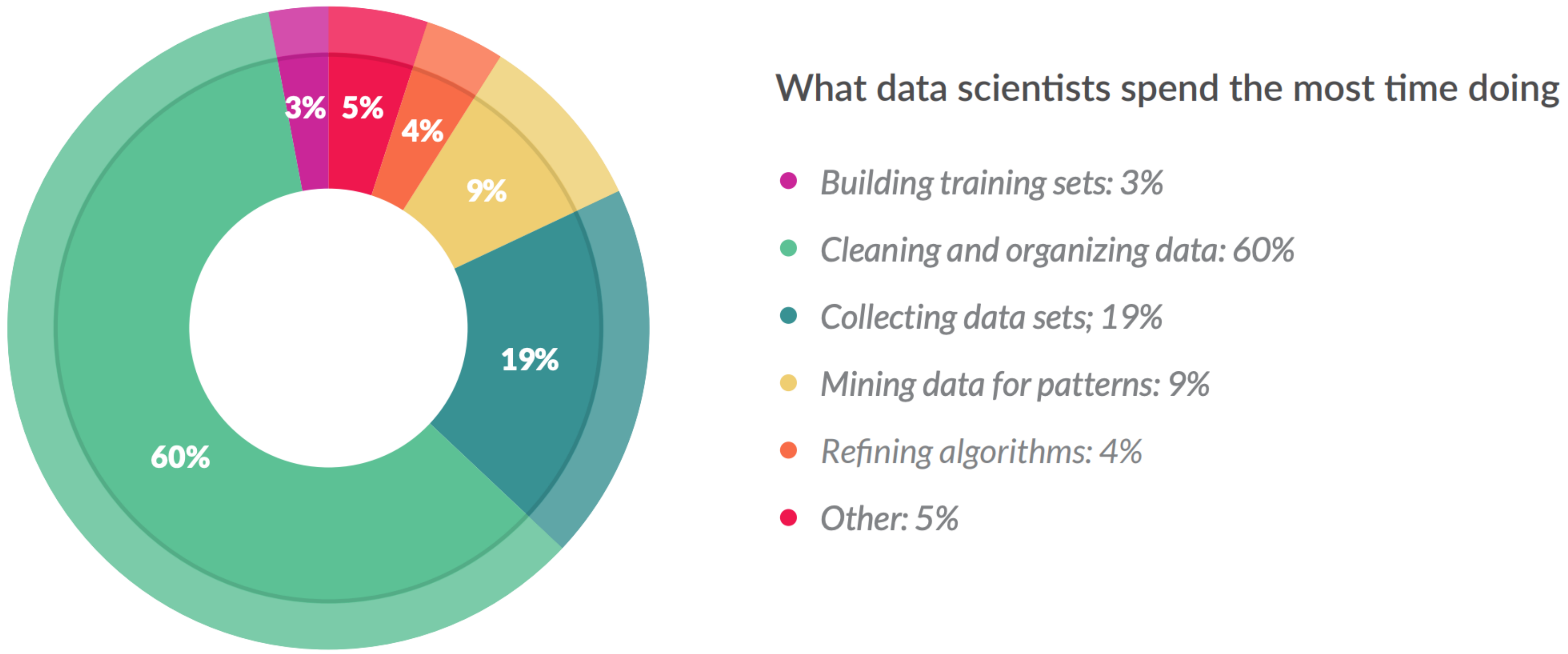


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

"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 DL?



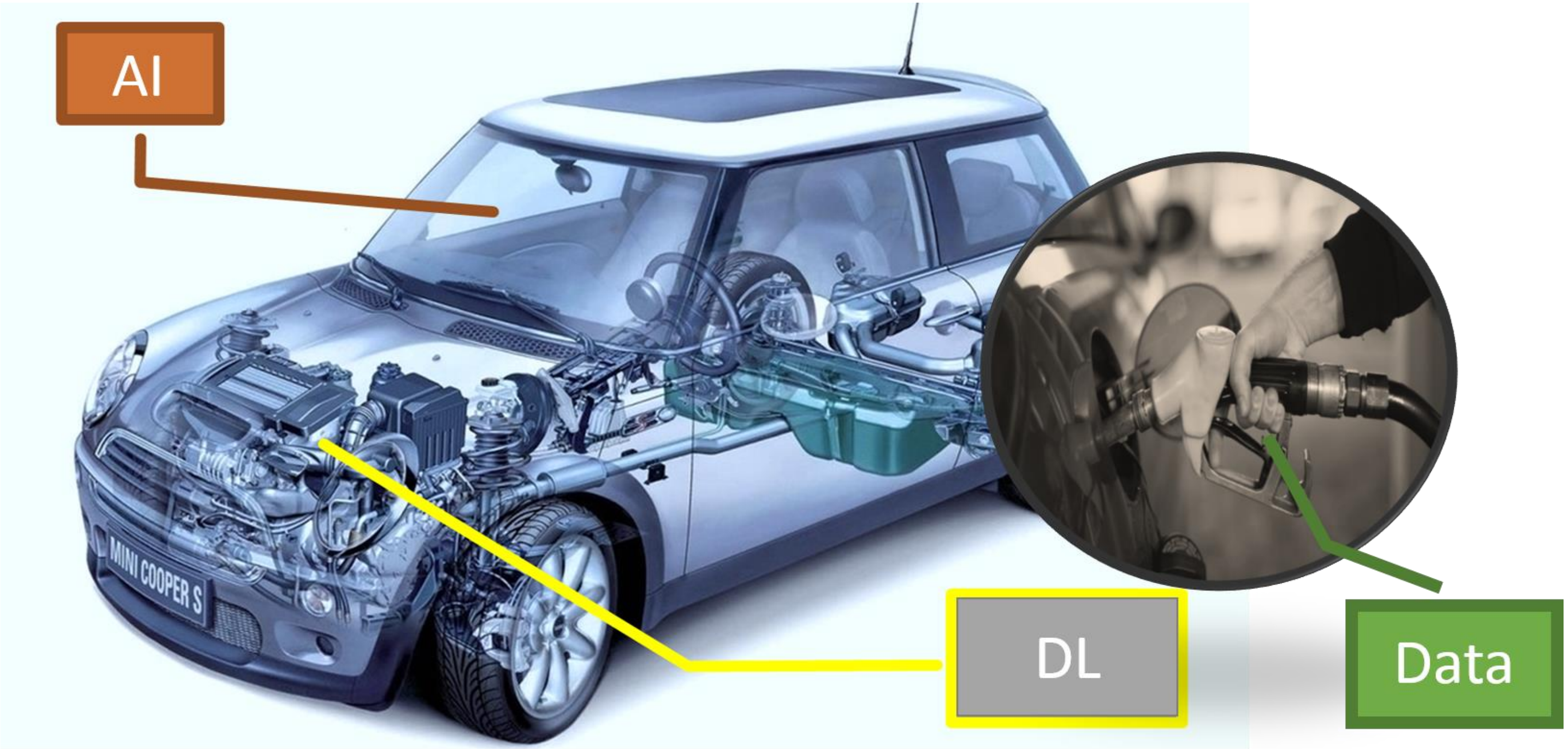
What data scientists spend the most time doing

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

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Background: Big Data, Deep learning and AI

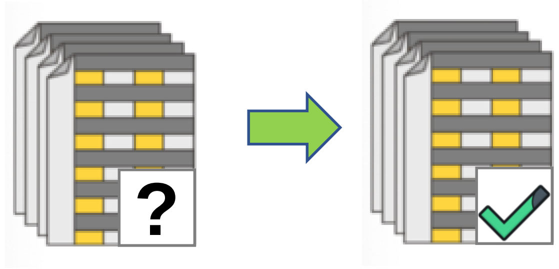
- If data is viewd as the oil, DL is the engine and AI is the car.



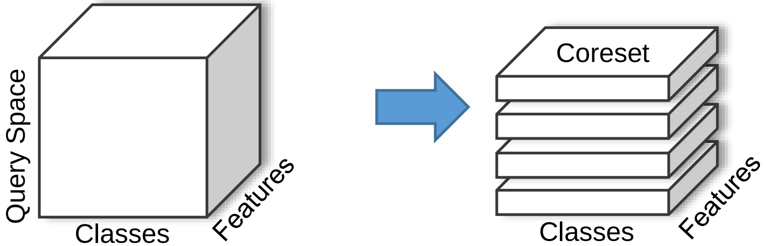
Challenges: data management for DL

#1: Data Preparation for DL → #2: Optimized Training in DL → #3: Result Validation and Explanation in DL

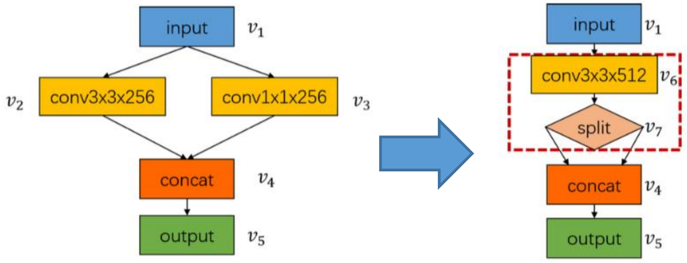
1.1 Data Extraction and Integration



1.2 Semi-automatic Data Labeling

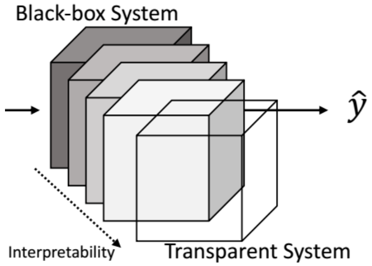


2.1 Coreset Selection and Compression



2.2 Graph and Operator Optimization

3.1 Result Validation



3.2 Result Explanation