深度學習實作與應用
Deep learning and its applications

Project Announcement &
Data processing

IM5062, Spring 2024

Some slides from Qingqing Huang, Mu Li, Alex Smola

黃意婷

Syllabus

Week	Date	Topic	備忘
1	2/20	Course introduction	
2	2/27	Basic Neural Network (I): from regression to neural networks. Regression, perceptron, forward propagation, activation functions	
3	3/5	Basic Neural Network (II): neural networks, activation functions	
4	3/12	Basic Neural Network (III): Loss functions, gradient descent, backward propagation.	HW1 announce
5	3/19	Basic Neural Network (III): optimizers, evaluation metrics, regularization.	
6	3/26	Convolutional Networks: Architectures, convolution / pooling layers	HW2 announce
7	4/2	Guest Lecturer (1): 中研院資訊所 王建堯博士	
8	4/9	Recurrent Neural Networks: RNN, GRU, LSTM	
9	4/16	Midterm	
10	4/23	Project Proposal	

Syllabus

Week	Date	Topic	備忘
11	4/30	Sequence to sequence learning: encoder-decoder, attention mechanism	
12	5/7	Guest Lecturer (2): 柏駿資本管理公司 杜勇正博士	HW3 announce
13	5/14	Transformer: Attention is all you need, BERT, GPT	
14	5/21	Guest Lecturer (3):八維智能 陳珮華 營運長	
15	5/28	Project presentation	
16	6/4	Project presentation	

Talk 後繳交的心得 1%*3

- 針對本週演講主題描述心得，填寫於NTU Cool作業區
  - 如有發問的同學：請簡述你的問題與獲得到的答案
  - 其他同學：請寫下400-600個字的心得

- Deadline:
  - 每次演講後的下週一 5pm

Final project: 30%

- 3~4 people in a team
- Novel problem: 5%
- Data collection and validation: 5%
- Novel approach: 5%
- Results w/o comparisons: 5% (w/ 5%)
- Proposal/Presentation/Report: 5%

Proposal 內容

- Task definition:
    - Input
    - Output
- Data collection
    - 如果使用公開的資料集，須說明來源、細節與比較對象
    - 自行蒐集的資料及：
        - 整體 Data數量與分布 (分別說明X和Y)
        - 如何切 training/validation/testing
    - Data validation & Preprocessing
- Approach:
    - 預計使用的方法與緣由
- Evaluation:
    - Metrics
    - 預計比較的對象

期末報告

- 書面資料格式
  - Introduction
  - Method
    - Task Definition
    - Approach
  - Data collection and validation
  - Evaluation
    - Metrics
    - Baselines
    - Results (Numerical & Case Study)
  - Conclusion

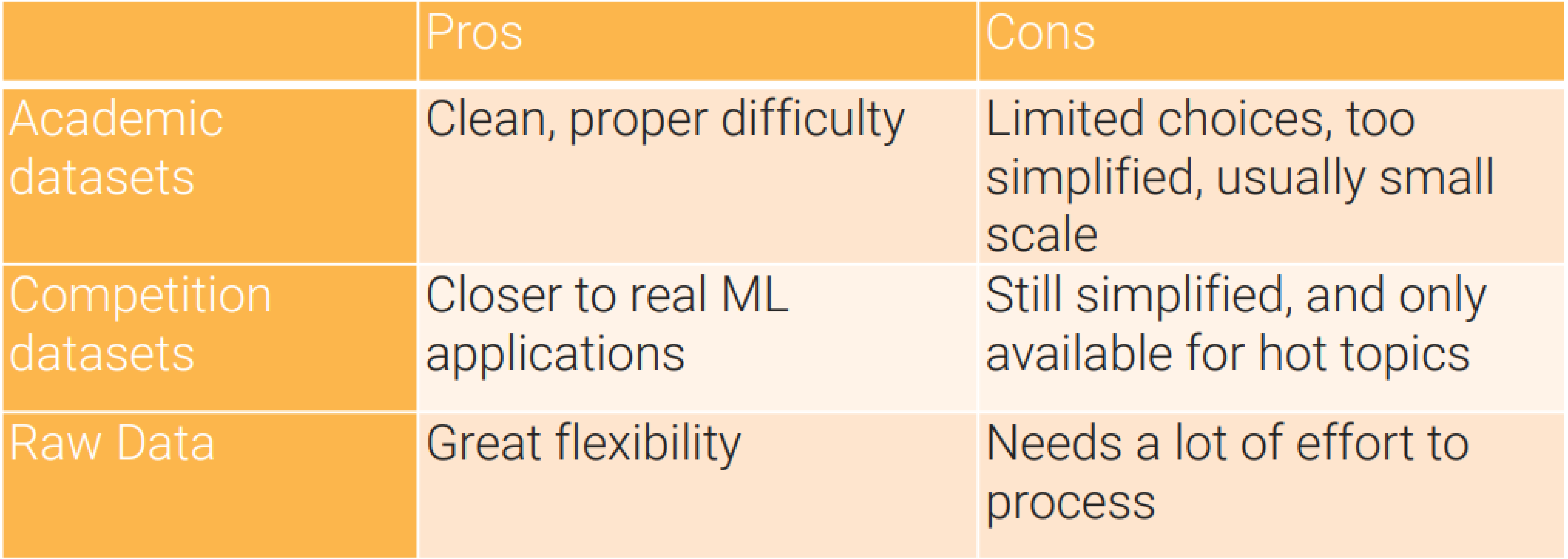
繳交文件與截止日期

- 4/8 中午12PM以前填寫分組(3~4人)名單
- 4/22 公告報告順序
- 4/23 Proposal:
  - 不須繳交檔案
  - 報告10-15分鐘
- 5/28, 6/4 Project presentation:
  - 繳交投影片、5 page 書面資料
    - 格式：word和tex檔 、pdf檔
  - 報告20-25分鐘

Open-source datasets

- Popular datasets
  - MNIST: digits written by employees of the US Census Bureau
  - ImageNet: millions of images from image search engines
- More Image, Text, Audio, etc. data at https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
- Search
  - Kaggle: https://www.kaggle.com/datasets
  - Google dataset search: https://datasetsearch.research.google.com/

Dataset comparison



	Pros	Cons
Academic datasets	Clean, proper difficulty	Limited choices, too simplified, usually small scale
Competition datasets	Closer to real ML applications	Still simplified, and only available for hot topics
Raw Data	Great flexibility	Needs a lot of effort to process

Make dataset on your own

- Web crawling VS scrapping
    - Crawling: indexing whole pages on Internet
    - Scraping: scraping particular data from web pages of a website

- Legal Considerations
    - Web scraping isn’t illegal by itself
    - But you should
        - NOT scrape data have sensitive information (E.g. private data involving username/password, personal health/medical information)
        - NOT scape copyrighted data (E.g. YouTube videos, Flickr photos)
        - Follow the Terms of Service that explicitly prohibits web scraping

Labelling

- Enough data/label
    - Notice the data distribution

- Not Enough data/label
    - Crowdsourcing: leverage global labelers to manually label data
        - Quality Control:
            - Sending the same task to multiple labelers, then determine the label by majority voting
            - Improve: prune low-quality labelers

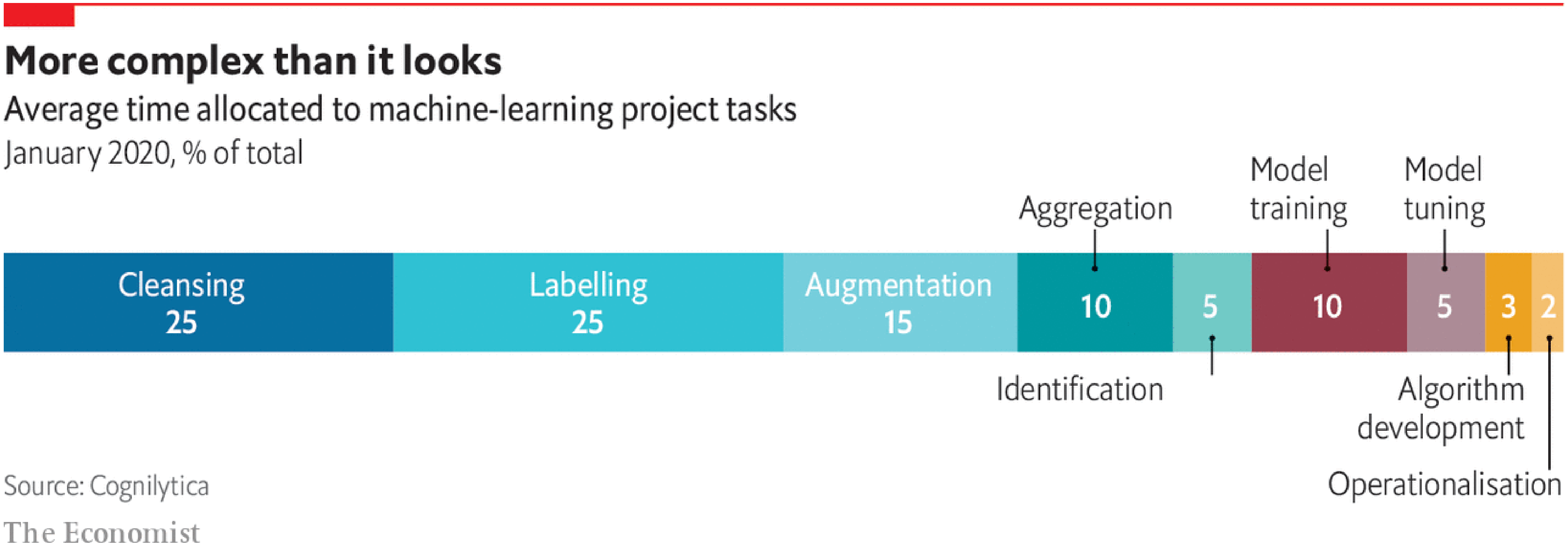
    - Data programming: heuristic programs to assign noisy labels
        - Domain specific heuristics to assign labels
        - Keyword search, pattern matching, third-party models

Data cleaning

- Outliers: data values that significantly deviate from other observations
- Rule violations: data values violate integrity constraints such as “Not Null” and “Must be unique” and “Non negative”
- Pattern violations: data values violate syntactic and semantic constraints such as formatting, misspelling

Data transformation

- Normalization:
  - Real value features:
    - Min-max
    - Z-score
    - Log scaling
  - Text:
    - Stemming
    - Lemmatization
    - Tokenization



More complex than it looks

Average time allocated to machine-learning project tasks

January 2020, % of total

Source: Cognilytica

The Economist

Roh, Yuji, Geon Heo, and Steven Euijong Whang. "A survey on data collection for machine learning: a big data-ai integration perspective." *IEEE Transactions on Knowledge and Data Engineering* (2019).