ONLINE MASTERS IN DATA SCIENCE

DSC 208R - Data Management for Analytics

Data Collection and Governance

Arun Kumar

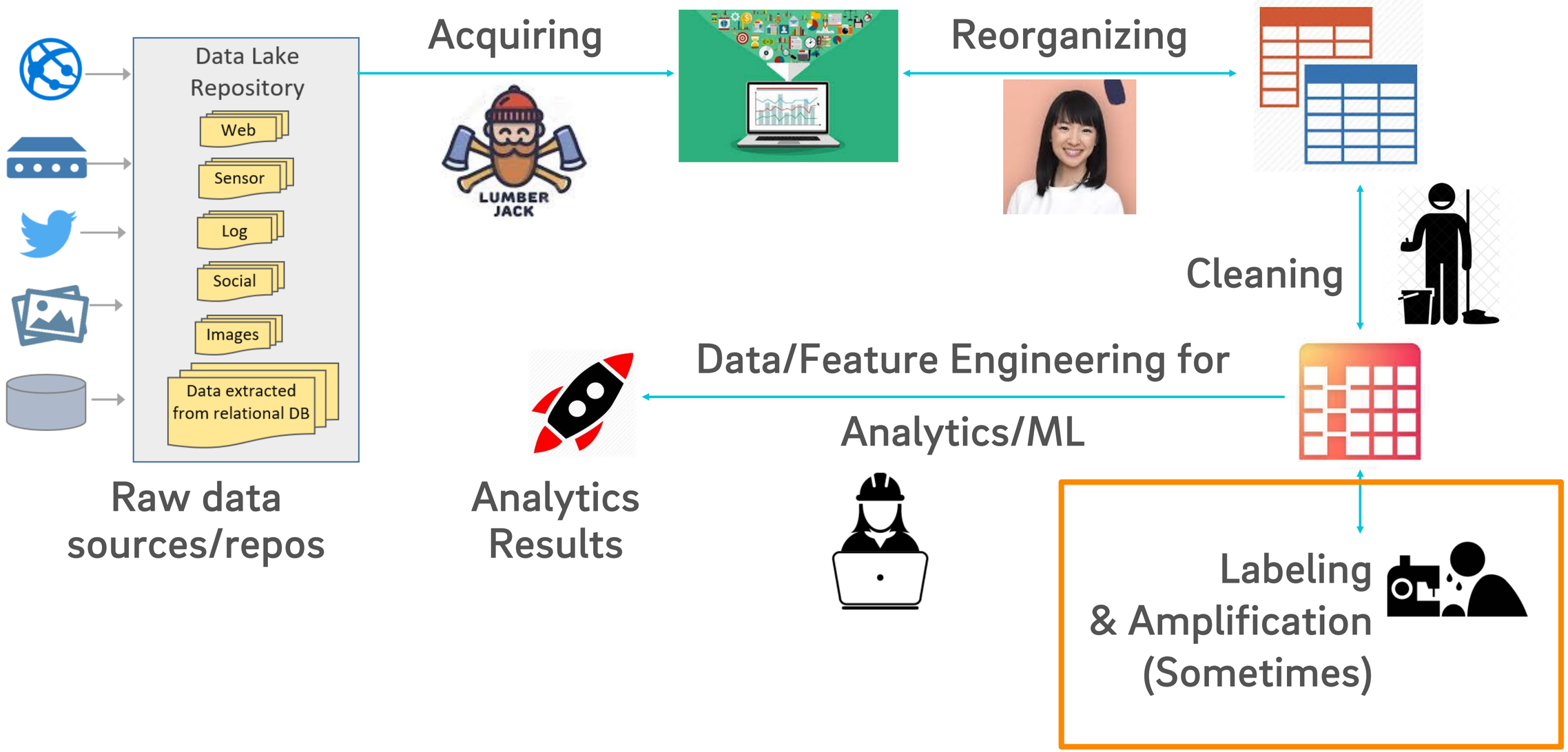
UC San Diego

COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

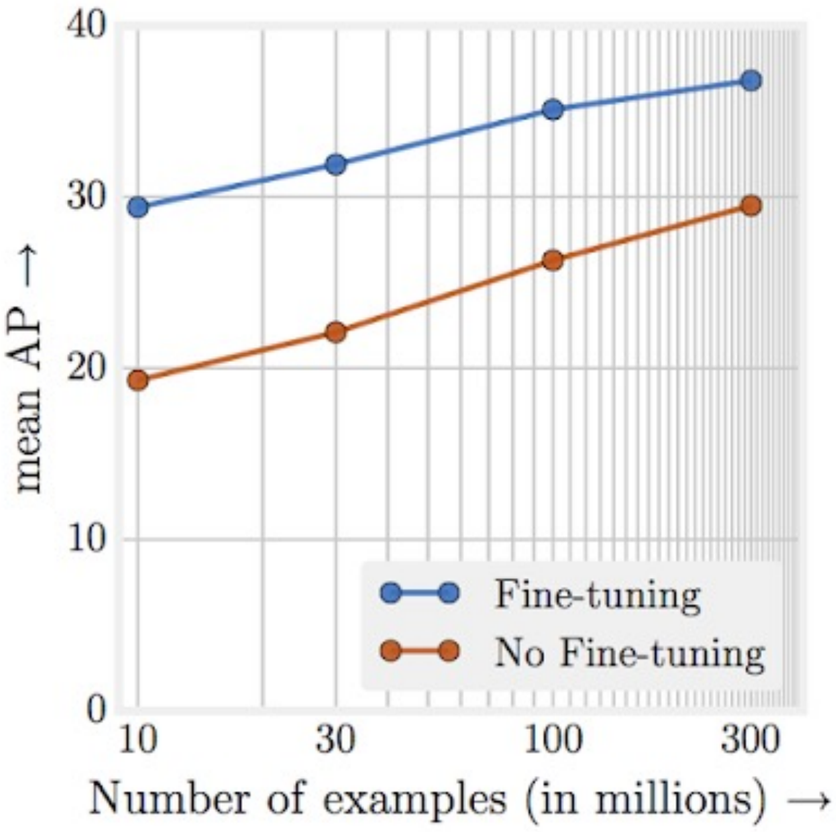
- Overview
- Data Organization and File Formats
- Data Acquisition
- Data Reorganization and Preparation
- **Data Labeling and Amplification**
- Data Governance and Privacy

Data Labeling



Data Labeling

- Most recent AI successes due to *supervised* ML
  - Large dataset is not enough—need *labeled* datasets, i.e., pairs of (input, output) examples



Object detection performance when pre-trained on different subsets of JFT-300M from scratch.
x-axis is the dataset size in log-scale, y-axis is the detection performance in mAP@[.5,.95] on COCO-minival subset.

https://ai.googleblog.com/2017/07/revisiting-unreasonable-effectiveness.html

Data Labeling

- **Labeling:** Process of annotating an example (raw or featurized) with *ground truth* label for a given prediction task
    - Notion of "label" is a prediction task-specific and data type-specific; can be almost any data structure!

Q: What is a label for this image?

- Dog (object recognition)
- Couch (object recognition)
- Shiba Inu (dog breed classifier)
- Yes (meme classifier!)
- Dog w/bounding box (obj.detection)
- Highlight dog (segmentation)

Data Labeling: Application Need

WRT sources of labels, 3 kinds of prediction applications:

1. Data-generating **process offers labels naturally**

   E.g.: Customer churn prediction, forecasting

2. Product/service **users offer labels (in)directly**

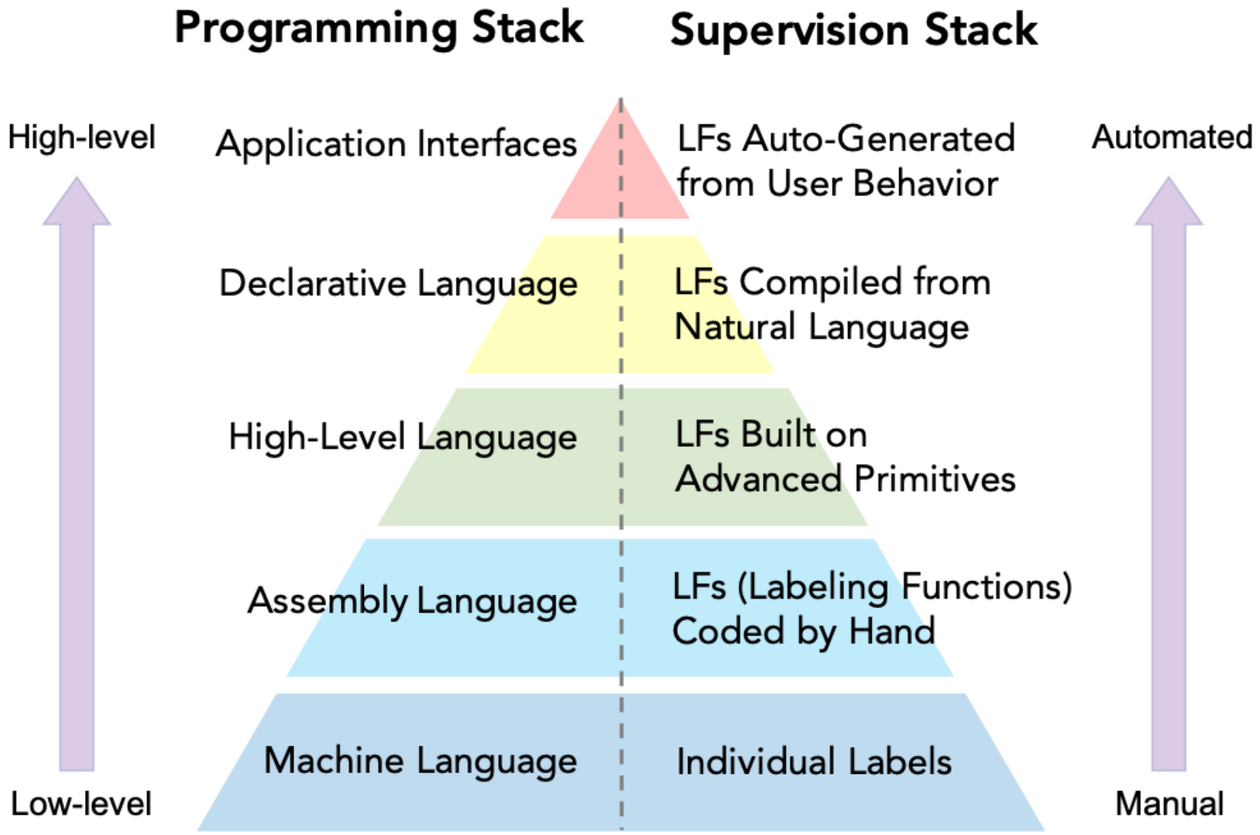
   E.g.: Email spam filters, online advertising, product recommendations, photo tagging, web search

3. Need **application-specific extra effort** for labels

   E.g.: Radiology, self-driving cars, species classification, video surveillance, machine translation, knowledge base construction, document summarization

Programmatic Labeling

- **Basic Idea:** Instead of manually labeling each example, write programs/rules/heuristics that encode some domain intuition to label examples en masse



http://cidrdb.org/cidr2019/papers/p58-ratner-cidr19.pdf

Pros

- Improved labeling productivity
- Likely lower costs

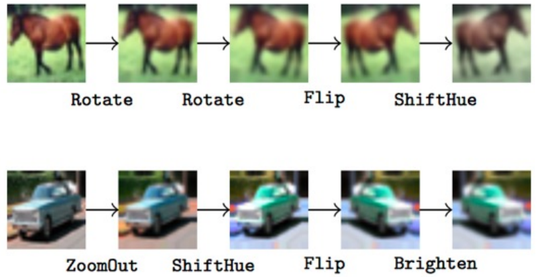
Cons

- Need to write code
- Less reliable accuracy
- Unclear if complex prediction outputs supportable

Amplification of Labeled Data

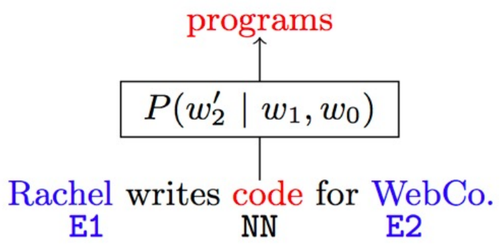
- **Label-preserving transforms** are common, esp. in vision

Images



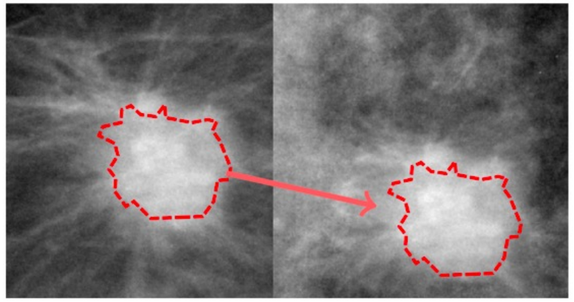
- Rotations
- Scaling / Zooms
- Brightness
- Color Shifts
- Etc…

Text



- Synonymy
- Positional Swaps
- Etc…

Medical



Domain-specific transformations.
Ex:

1. *Segment tumor mass*
2. *Move*
3. *Resample background tissue*
4. *Blend*

- **Synthesis** sometimes possible in robotics/sci./eng.
  - Physical laws-based, simulation-based, etc.
  - Tricky; needs knowledge of underlying data distr.