Stanford University

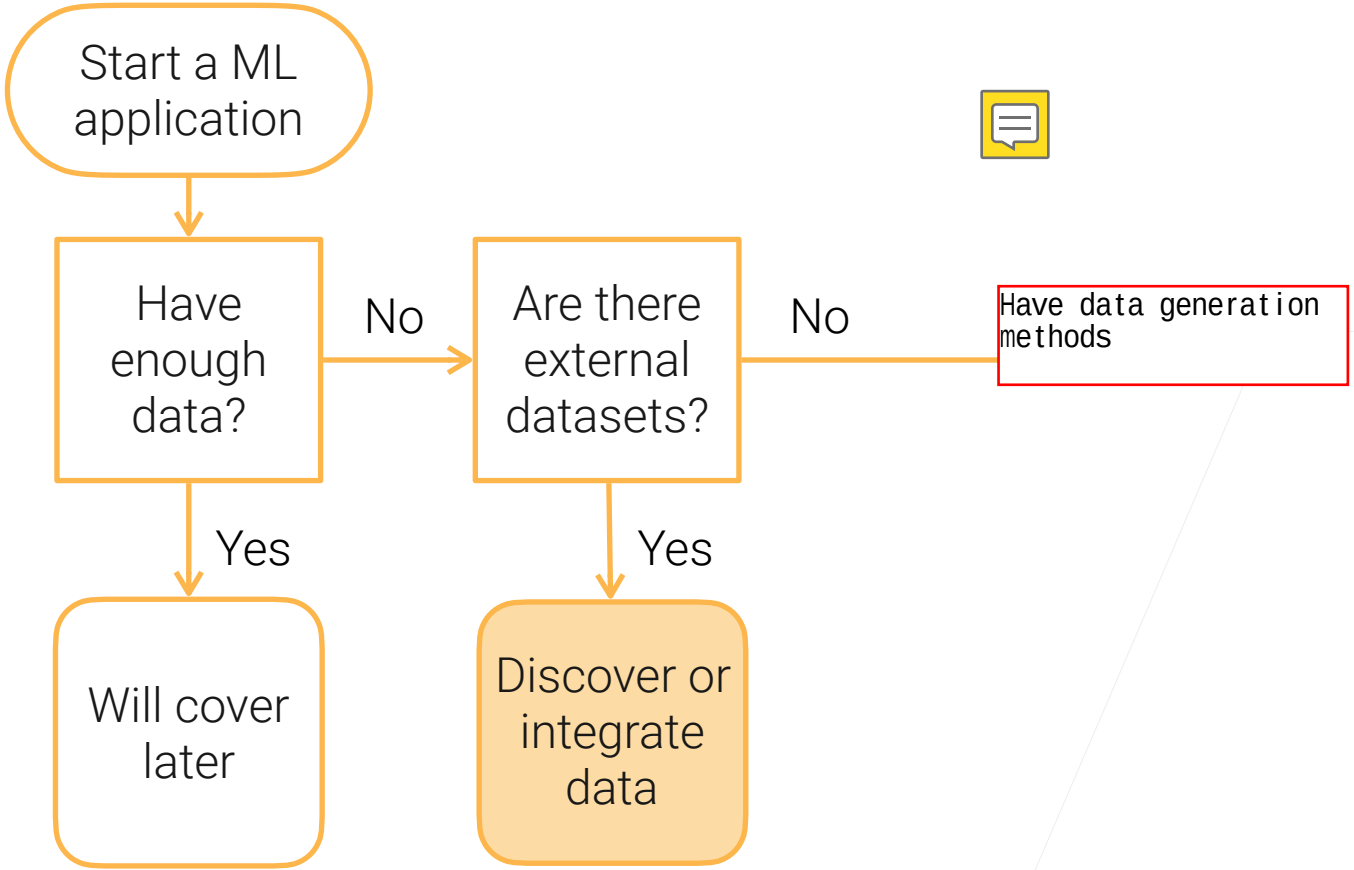
CS 329P : Practical Machine Learning (2021 Fall)

1.2 Data Acquisition

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

Flow Chart for Data Acquisition



Discover What Data is Available

- Identify existing datasets
- Find benchmark datasets to evaluate a new idea
  - E.g. A diverse set of small to medium datasets for a new hyper-parameter tuning algorithm
  - E.g. Large scale datasets for a very big deep neural network

```
Collect new data
  E.g. driving videos covering different driving scene
```

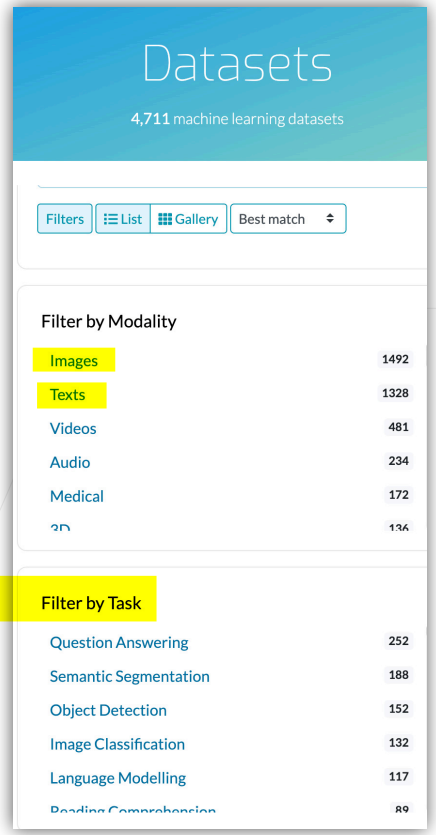
Popular ML datasets

- MNIST: digits written by employees of the US Census Bureau
- ImageNet: millions of images from image search engines
- AudioSet: YouTube sound clips for sound classification
- LibriSpeech: 1000 hours of English speech from audiobook
- Kinetics: YouTube videos clips for human actions classification
- KITTI: traffic scenarios recorded by cameras and other sensors
- Amazon Review: customer reviews and from Amazon online shopping
- SQuAD: question-answer pairs derived from Wikipedia

More at https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Where to Find Datasets

- Paperswithcodes Datasets: academic datasets with leaderboard
- Kaggle Datasets: ML datasets uploaded by data scientists
- Google Dataset search: search datasets in the Web
- Various toolkits datasets: tensorflow, huggingface
- Various conference/company ML competitions
- Open Data on AWS: 100+ large-scale raw data
- Data lakes in your own organization



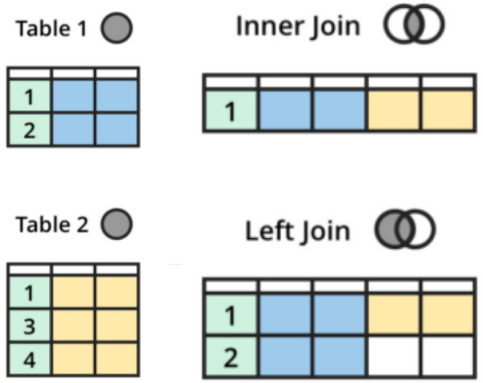
Datasets Comparison

	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

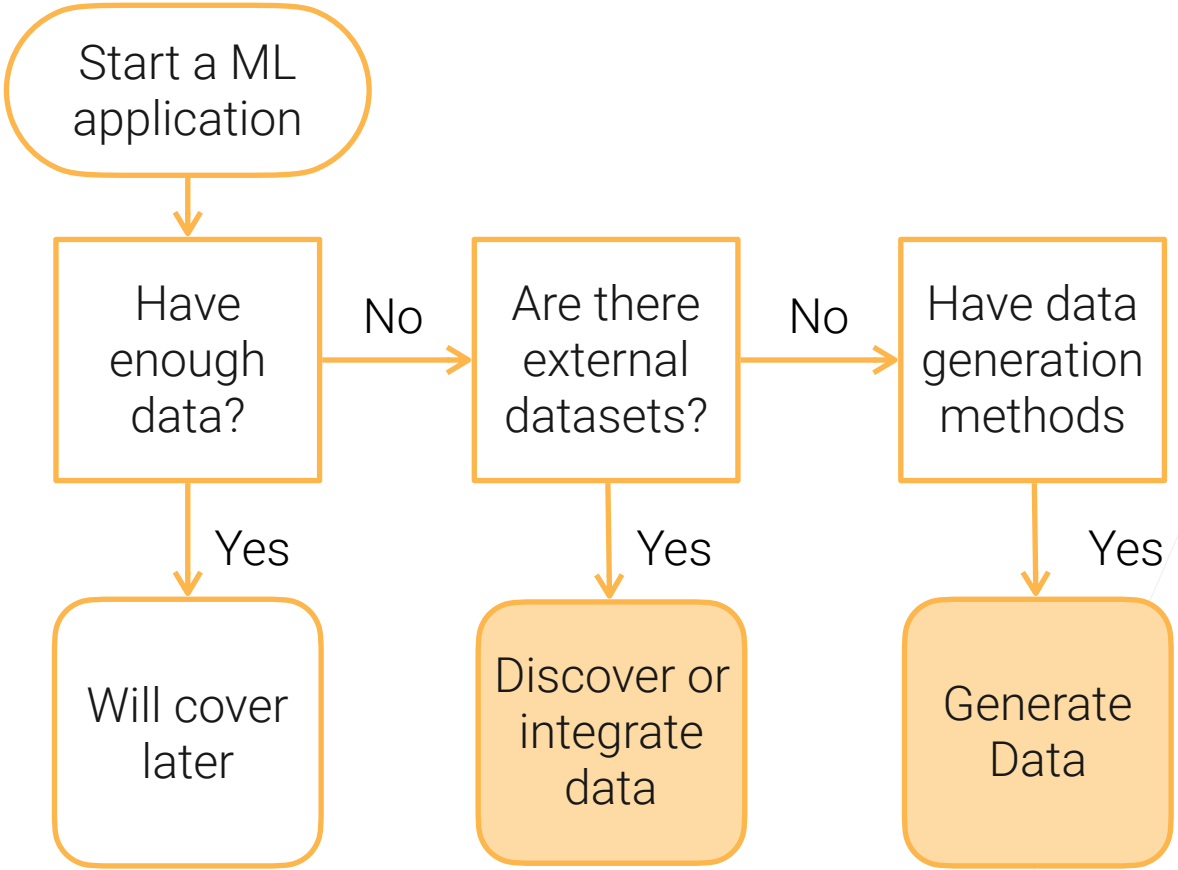
- You often need to deal with raw data in industrial settings
- Data curation can be a big projection involving multiple teams. Processing pipeline, storage, legal issue, privacy,...

Data Integration

- Combine data from multiple sources into a coherent dataset
- Product data is often stored in multiple tables
  - E.g. a table for house information, a table for sales, a table for listing agents
- Join tables by keys, which are often entity IDs
- Key issues: identify IDs, missing rows, redundant columns, value conflicts



Flow Chart for Data Acquisition



Generate Synthetic Data

- Use GANs

Faces

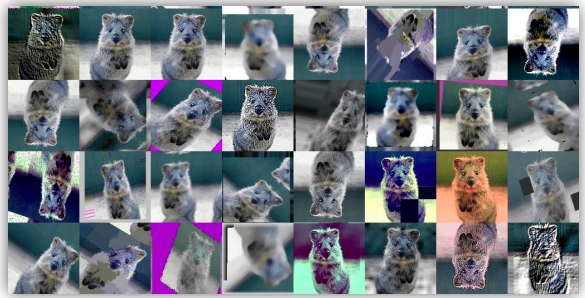
https://thispersondoesnotexist.com/

Furnitures in living rooms

Gadde et al., ICCV'21

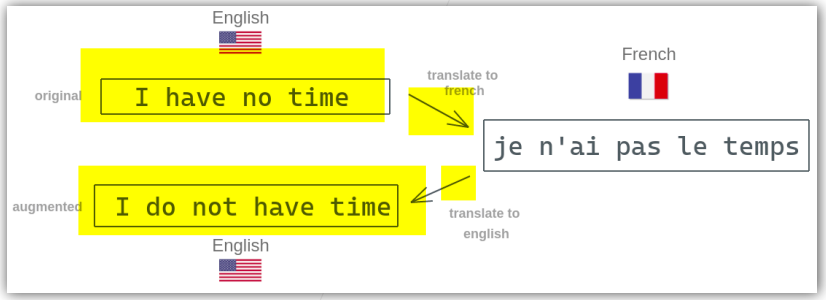
- Simulation
- Data augmentations

Image augmentation



https://github.com/aleju/imgaug

Back Translation



https://amitness.com

Summary

- Finding the right data is challenging
- Raw data in industrial settings VS academic datasets
- Data integration combines data from multiple sources
- Data augmentation a common practice
- Synthesizing data is getting popular