



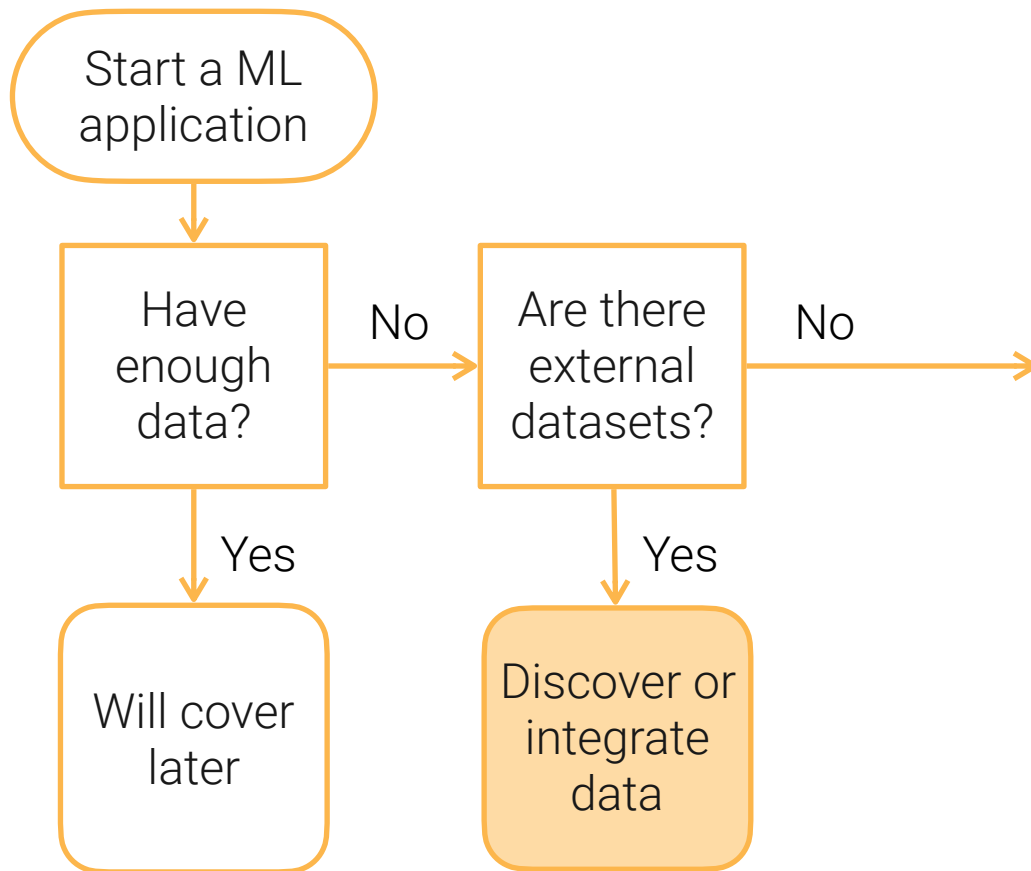
CS 329P : Practical Machine Learning (2021 Fall)

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 hyperparameter tuning algorithm
 - E.g. Large scale datasets for a very big deep neural network

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

A screenshot of the 'Datasets' website interface. The header is teal with the word 'Datasets' in white and '4,711 machine learning datasets' below it. Below the header is a navigation bar with buttons for 'Filters', 'List', 'Gallery', and 'Best match'. The main content area is divided into two sections: 'Filter by Modality' and 'Filter by Task'. The 'Filter by Modality' section shows a list of modalities with their respective counts: Images (1492), Texts (1328), Videos (481), Audio (234), Medical (172), and 2D (136). The 'Filter by Task' section shows a list of tasks with their respective counts: Question Answering (252), Semantic Segmentation (188), Object Detection (152), Image Classification (132), Language Modelling (117), and Reading Comprehension (89).

| Filter by Modality | |
|--------------------|------|
| Images | 1492 |
| Texts | 1328 |
| Videos | 481 |
| Audio | 234 |
| Medical | 172 |
| 2D | 136 |

| Filter by Task | |
|-----------------------|-----|
| Question Answering | 252 |
| Semantic Segmentation | 188 |
| Object Detection | 152 |
| Image Classification | 132 |
| Language Modelling | 117 |
| Reading Comprehension | 89 |

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 project 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

Table 1 ●

| | | |
|---|--|--|
| | | |
| 1 | | |
| 2 | | |

Inner Join ○○

| | | | |
|---|--|--|--|
| | | | |
| 1 | | | |
| | | | |

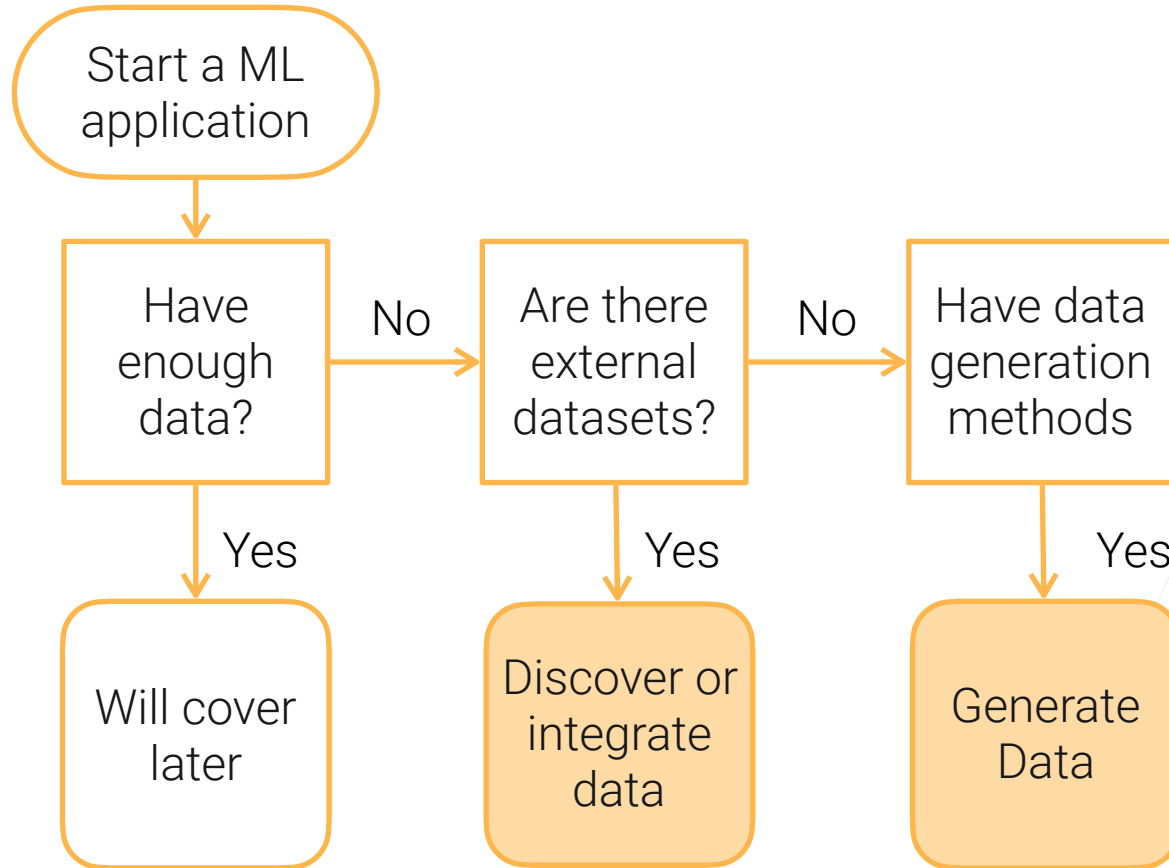
Table 2 ●

| | | |
|---|--|--|
| | | |
| 1 | | |
| 3 | | |
| 4 | | |

Left Join ○○

| | | | |
|---|--|--|--|
| | | | |
| 1 | | | |
| 2 | | | |
| | | | |

Flow Chart for Data Acquisition

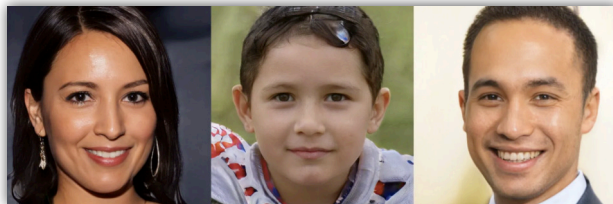


Generate Synthetic Data



- Use GANs

Faces



<https://thispersondoesnotexist.com/>

Furnitures in living rooms



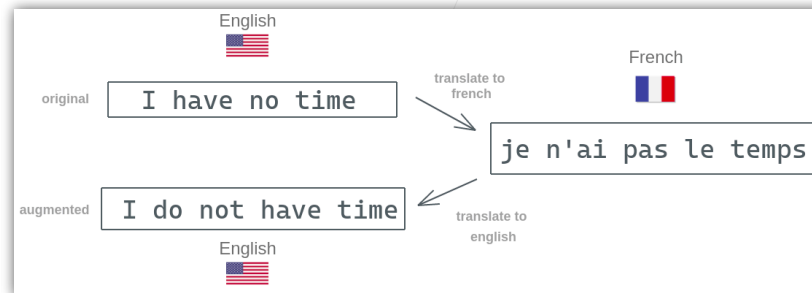
Gadde et al., ICCV'21

- Simulation
- Data augmentations

Image augmentation



Back Translation



<https://amitnesh.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