COMP/EECE 7/8745 Machine Learning

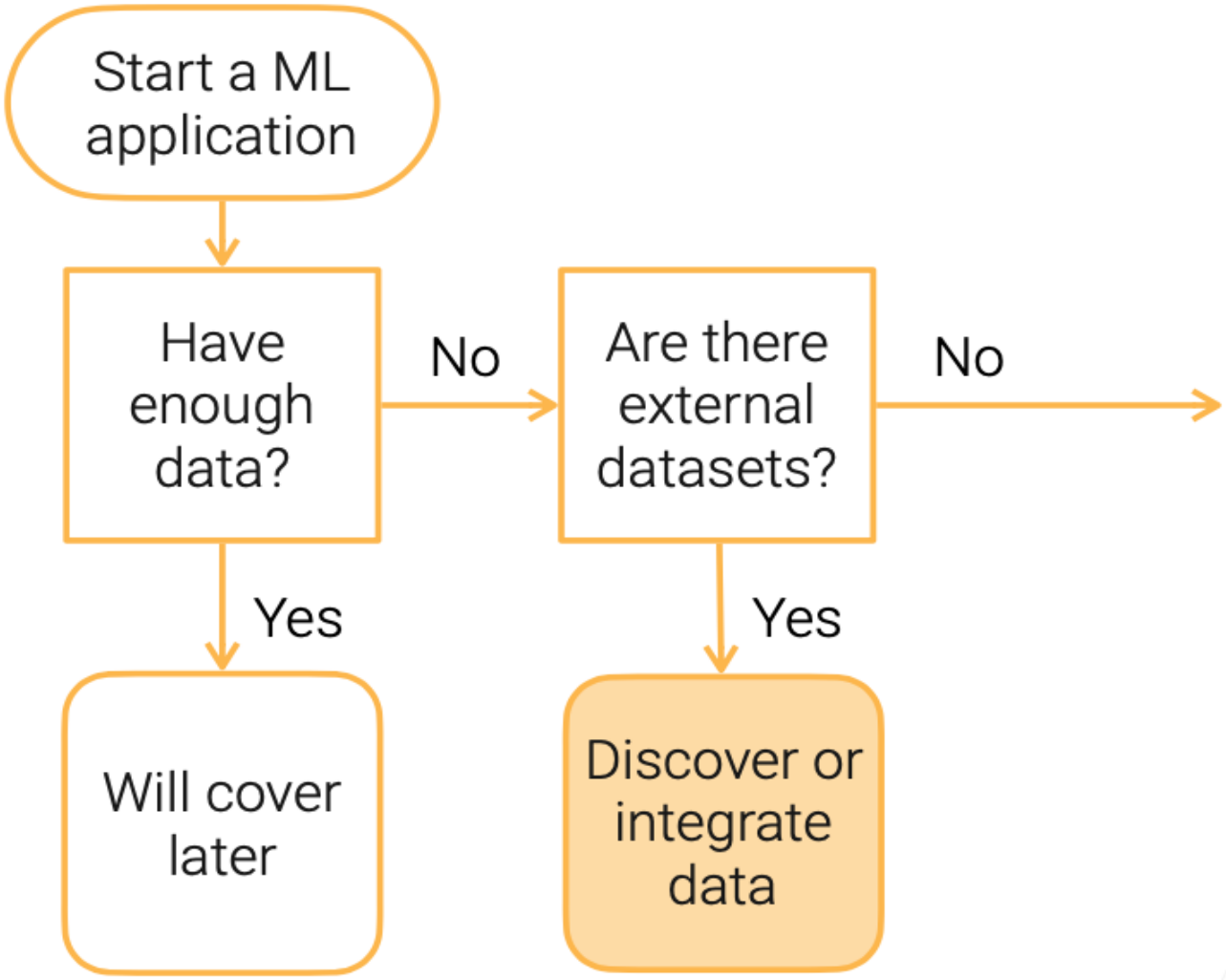
Topics:

Data preparations:

- Data acquisition and labeling
- Web scraping
- Why data preprocessing and clearing?
- Data integration, transformation, and reduction
- Discretization

Md Zahangir Alom
Department of Computer Science
University of Memphis, TN

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

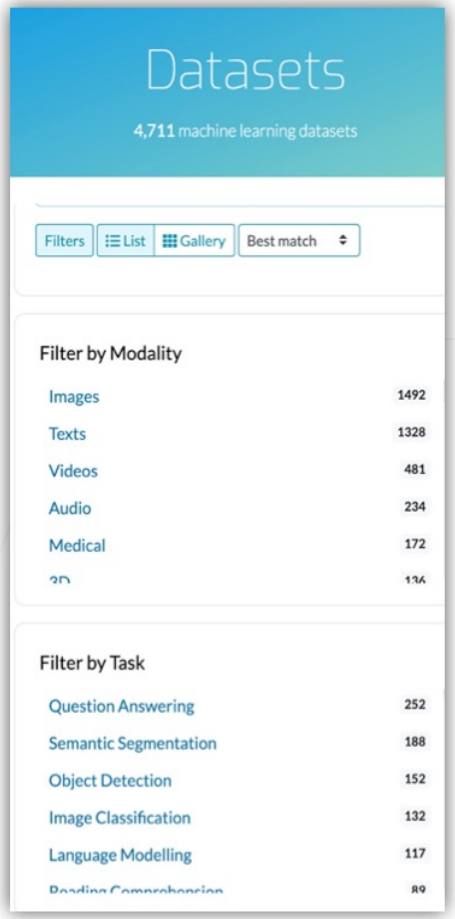
Popular ML datasets

- 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

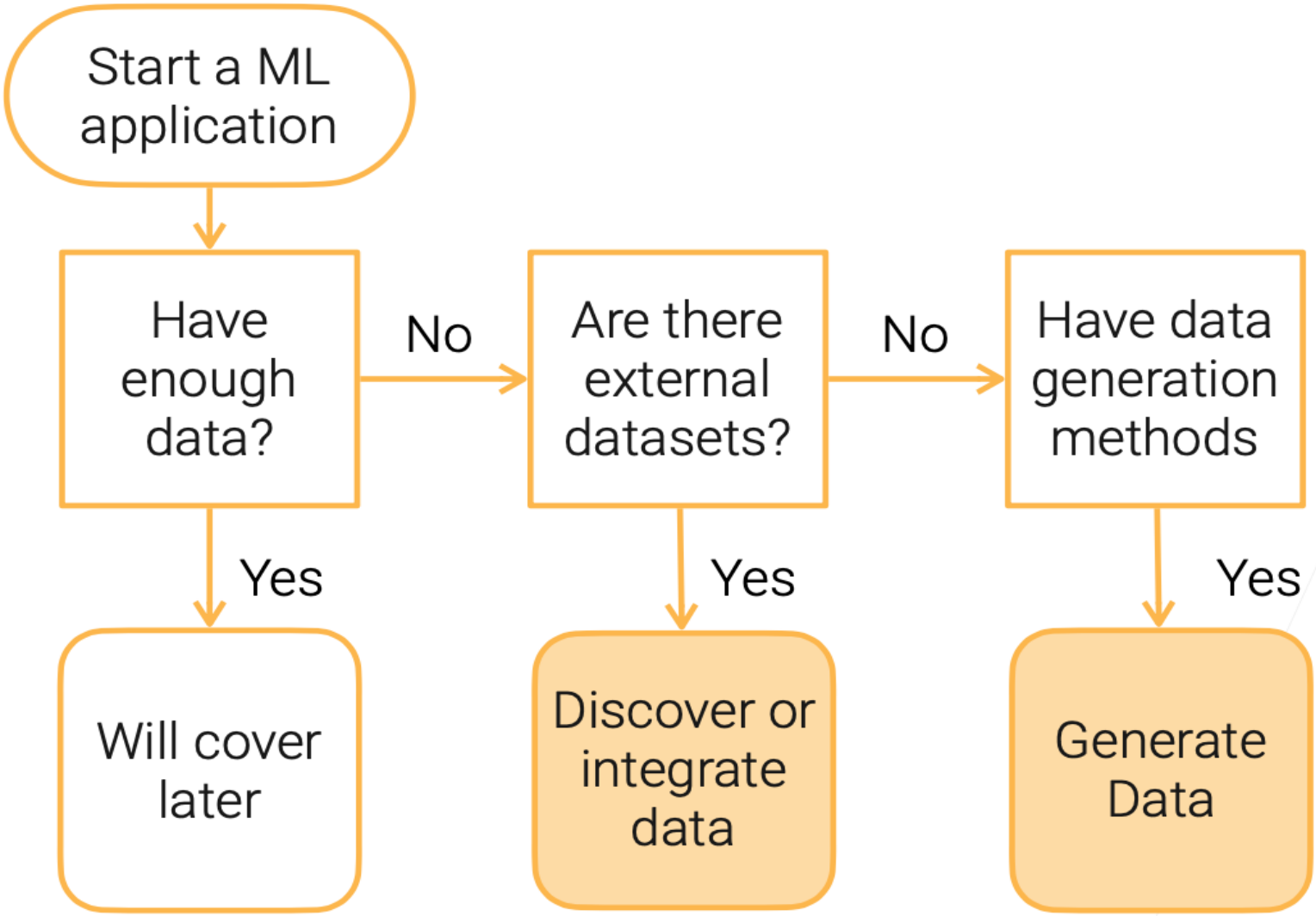


Where to Find Datasets

	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,...

Flow Chart for Data Acquisition



Stanford CS 329P (2021 Fall)

Generate Synthetic Data

- Use GANs

Faces

https://thispersondoesnotexist.com/

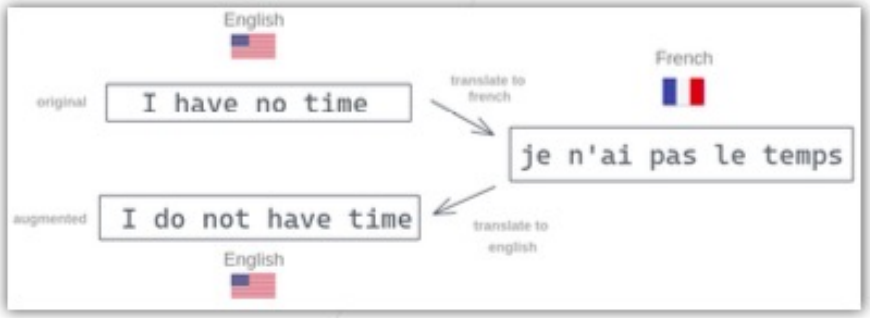
Furnitures in living rooms

Gadde et al., ICCV'21

- Simulation
- Data augmentations

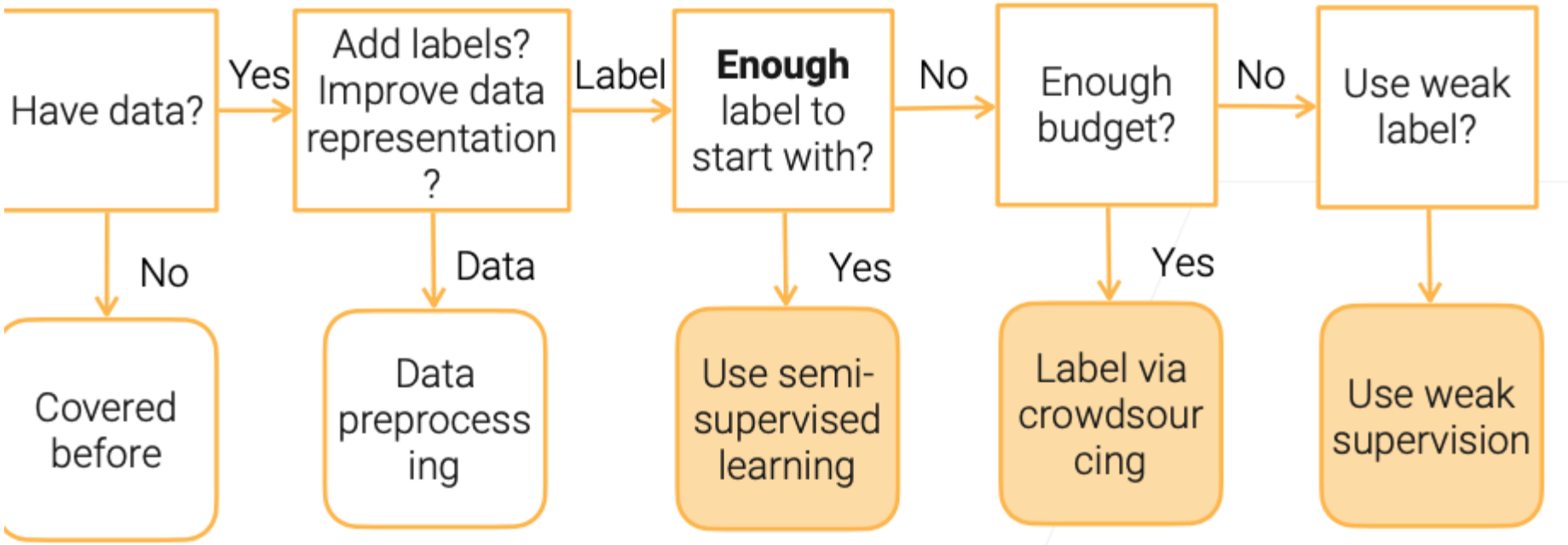
Image augmentation

Back Translation



https://amitness.com

Flow Chart for Data Labelling

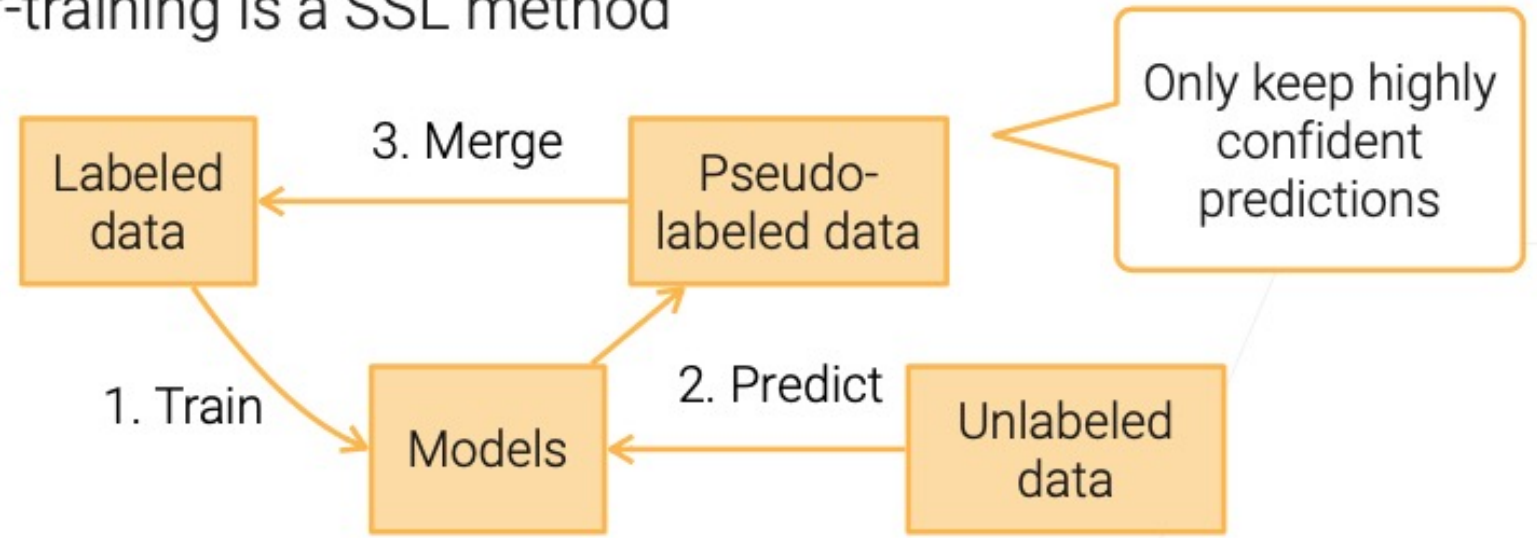


Semi-Supervised Learning (SSL)

- Focus on the scenario where there is a small amount of labeled data, along with large amount of unlabeled data
- Make assumptions on data distribution to use unlabeled data
  - **Continuity assumption:** examples with similar features are more likely to have the same label
  - **Cluster assumption:** data have inherent cluster structure, examples in the same cluster tend to have the same label
  - **Manifold assumption:** data lie on a manifold of much lower dimension than the input space

Self-training

- Self-training is a SSL method



- We can use expensive models
  - Deep neural networks, model ensemble/bagging

Label through Crowdsourcing

- ImageNet labeled millions of images through Amazon Mechanical Turk. It took several years and millions dollars to build
- According to Amazon SageMaker Ground Truth, the estimated price of using Amazon Mechanical Turk:

Image/text classification	$0.012 per label
Bounding box	$0.024 per box
Semantic segmentation	$0.84 per image

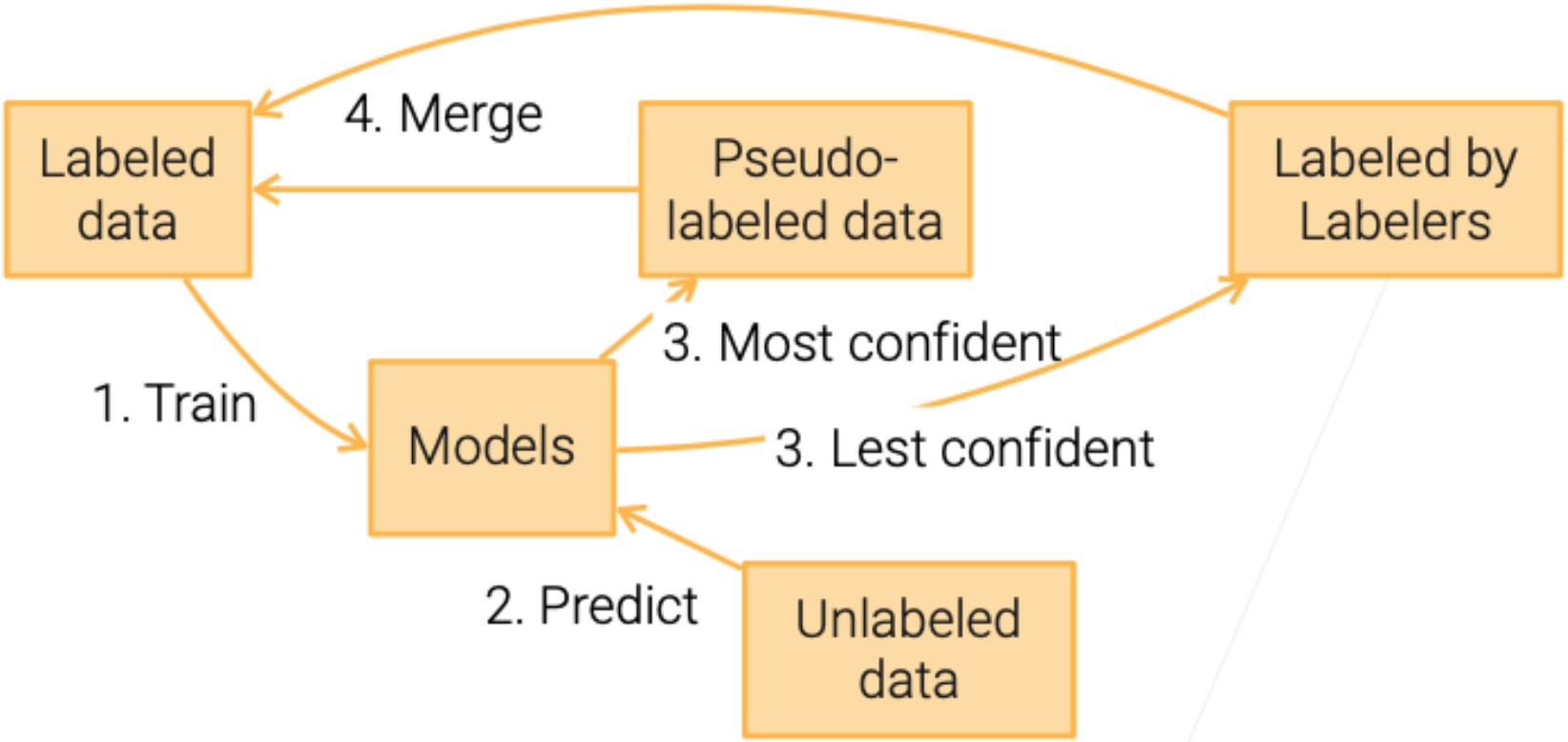
User interaction

- Example of user instruction and labeling task (MIT Place365)

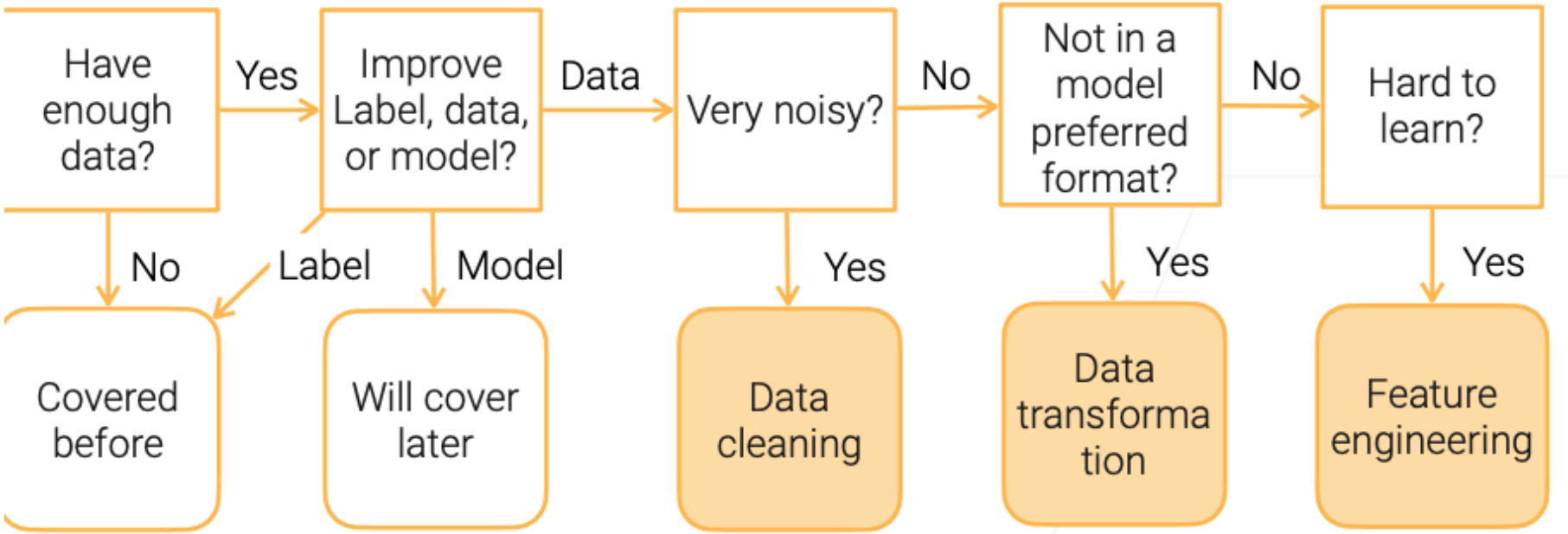
b)

Active Learning + Self-training

- These two methods are often used together



Flow chart for data preprocessing



Web Scraping

- The goal is to extract data from website
  - Noisy, weak labels, can be spammy
  - Available at scale
  - E.g. price comparison/tracking website
- Many ML datasets are obtained by web scraping
  - E.g. ImageNet, Kinetics
- Web crawling VS scrapping
  - Crawling: indexing whole pages on Internet
  - Scraping: scraping particular data from web pages of a website

Image credit: Aaron Zappia

Tools

- "curl" often doesn't work
  - Website owners use various ways to stop bots
- Use headless browser: a web browser without a GUI
- You need a lot of new IPs, easy to get through public clouds
  - In all IPv4 IPs, AWS owns 1.75%, Azure 0.55%, GCP 0.25%

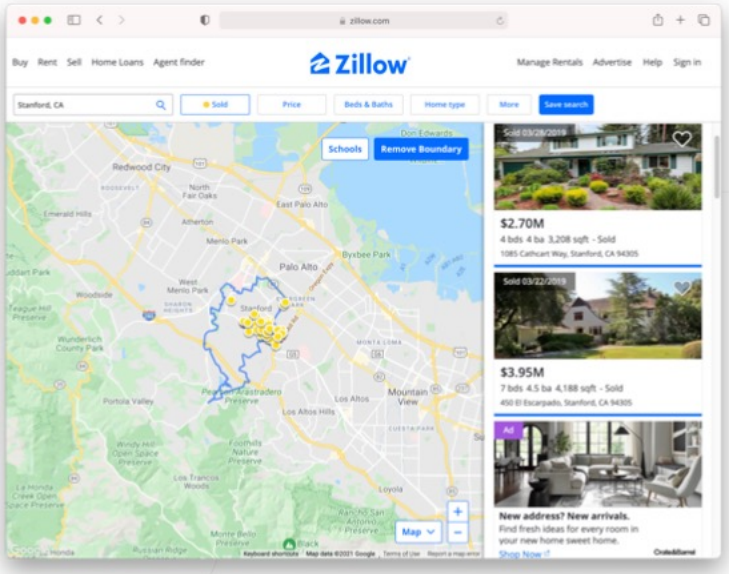
```
from selenium import webdriver

chrome_options = webdriver.ChromeOptions()
chrome_options.headless = True
chrome = webdriver.Chrome(
  chrome_options=chrome_options)

page = chrome.get(url)
```

Case Study

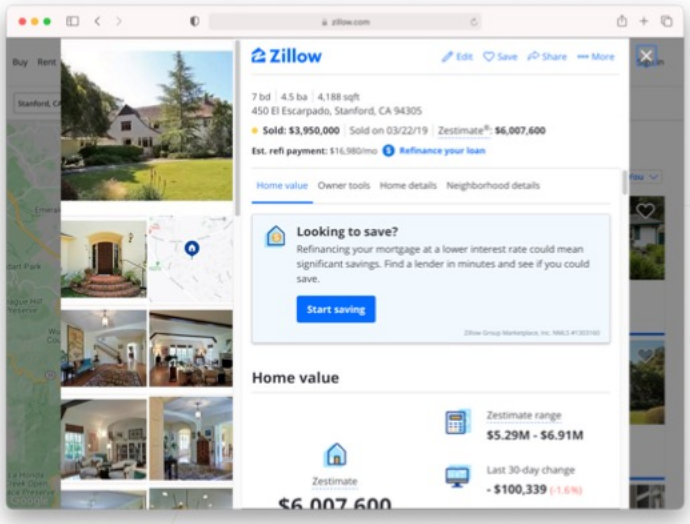
- Query houses sold in near Stanford
  - https://www.zillow.com/stanford-ca/sold/
  - https://www.zillow.com/stanford-ca/sold/2-p/
  - ...
- You can replace the city and state in the URL for other places



Craw individual pages

- Get the house IDs from the index pages

```
page = BeautifulSoup(open(html_path, 'r'))
links = [a['href'] for a in page.find_all(
          'a', 'list-card-link')]
ids = [l.split('/')[-2].split('_')[0]
       for l in links]
```

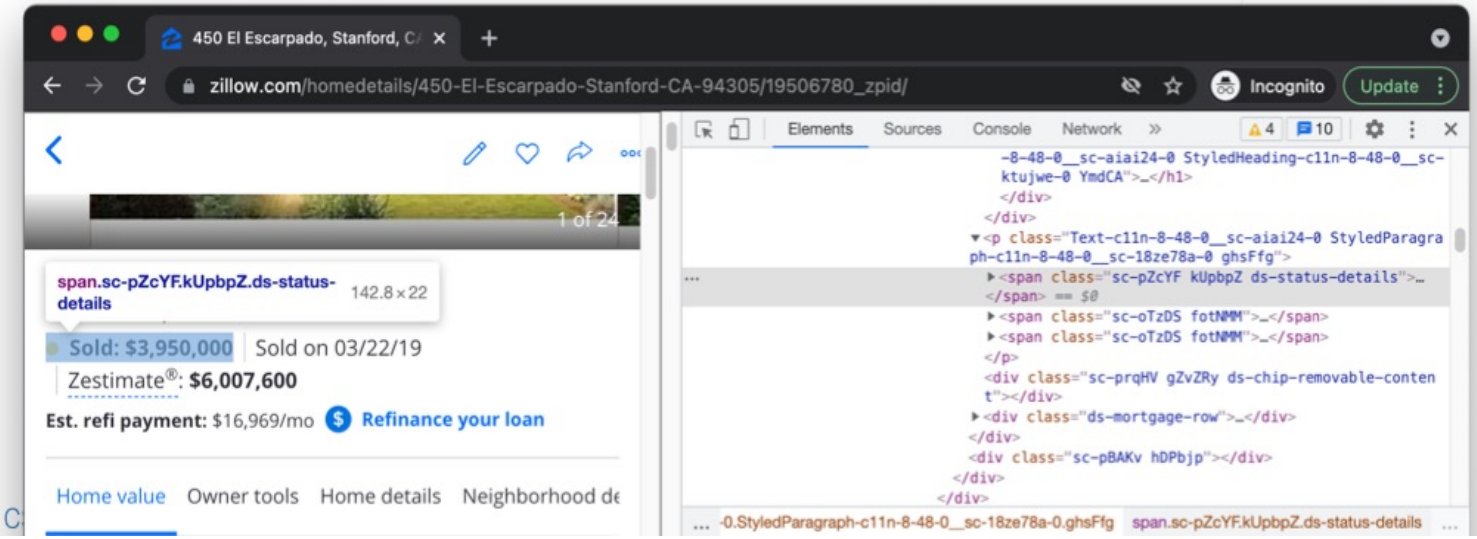


- The house detail page by ID
  - https://www.zillow.com/homedetails/19506780_zpid/

Extract data

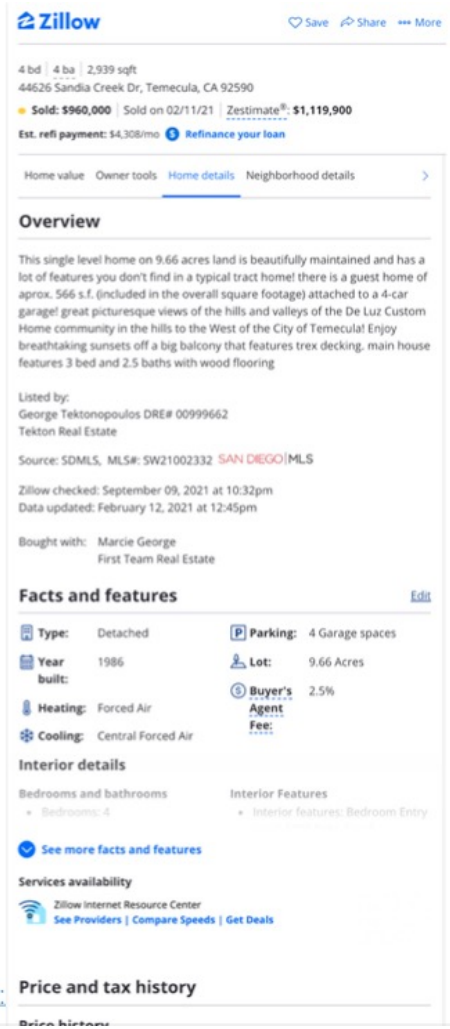
- Identify the HTML elements through **Inspect**

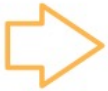
```
sold_items = [a.text for a in page.find(
  'div', 'ds-home-details-chip').find('p').find_all('span')]
for item in sold_items:
    if 'Sold:' in item:
        result['Sold Price'] = item.split(' ')[1]
    if 'Sold on' in item:
        result['Sold On'] = item.split(' ')[-1]
```



Extract data

- Repeat the previous process to extract other field data





```
{'Id': '18173100',
 'Address': '44626 Sandia Creek Dr,',
 'Sold Price': '$960,000',
 'Sold On': '02/11/21',
 'Summary': "This single level home on 9.66 acres la
 'Type': 'SingleFamily',
 'Year built': '1986',
 'HOA': '$70 annually',
 'Lot': '420,790 sqft',
 'Bedrooms': '4',
 'Bathrooms': '4',
 'Full bathrooms': '3',
 '1/2 bathrooms': '1',
 'Main level bathrooms': '3',
 'Association name': 'De Luz Ranchose',
 'Tax assessed value': '$784,118',
 'Annual tax amount': '$12,036',
 'Listed On': '1/5/2021',
 'Listed Price': '$975,000',
 'Last Sold On': '3/5/2008',
 'Last Sold Price': '$840,000',
 'Elementary School': 'Murrieta Elementary School',
 'Elementary School Score': '7',
 'Elementary School Distance': '4.9',
 'Middle School': 'Thompson Middle School',
 'Middle School Score': '6',
 'Middle School Distance': '5.5',
 'High School': 'Murrieta Valley High School',
 'High School Score': '8',
```

Cost

- Use AWS EC2 t3.small (2GB memory, 2 vCPUs, $0.02 per hour)
  - 2GB is necessary as the browser needs a lot memory, CPU and bandwidth are usually not an issue
  - Can use spot instance to reduce the price
- The cost to crawl 1M houses is $16.6
  - The speed is about 3s per page,
  - 8.3 hours if using 100 instances
  - The extra cost includes storage, restart instances when IP is banned

Crawl Images

- Get all image URLs

```
p = r'https:\\/\\/photos.zillowstatic.com\\/fp\\/([\d\w\-\_]+).jpg'
ids = [a.split('-')[0] for a in re.findall(p, html)]
urls = [f'https://photos.zillowstatic.com/fp/{id}-
uncropped_scaled_within_1536_1152.jpg' for id in ids]
```

- A house listing has ~20 images
  - The crawling cost is still reasonable: ~$300
  - Storing these images is expensive: ~$300 per month
    - You can reduce the image resolutions, or send data back

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
- Consult a lawyer if you are doing it for profit

Data preprocessing

Why data preprocessing?

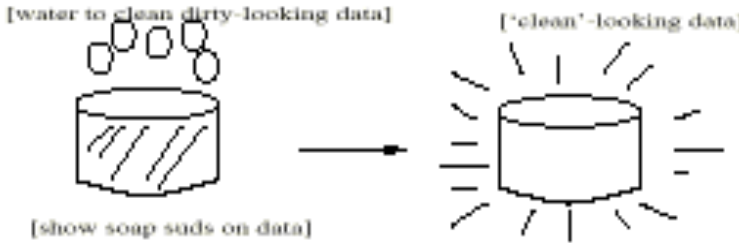
- Data in the real world is dirty
  - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - **noisy:** containing errors or outliers
  - **inconsistent:** containing discrepancies in codes or names
- **No quality data, no quality mining results!**
  - Quality decisions must be based on quality data
- A multi-dimensional measure of data quality:
  - A well-accepted multi-dimensional view:
    - accuracy, completeness, consistency, timeliness, believability, value added, interpretability, accessibility
  - Broad categories:
    - intrinsic, contextual, representational, and accessibility.

Major tasks in data preprocessing

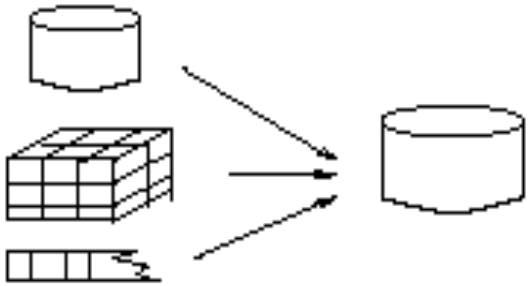
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, files, or notes
- Data transformation
  - Normalization (scaling to a specific range)
  - Aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Data discretization: with particular importance, especially for numerical data
  - Data aggregation, dimensionality reduction, data compression, and generalization

Forms of data preprocessing

Data Cleaning



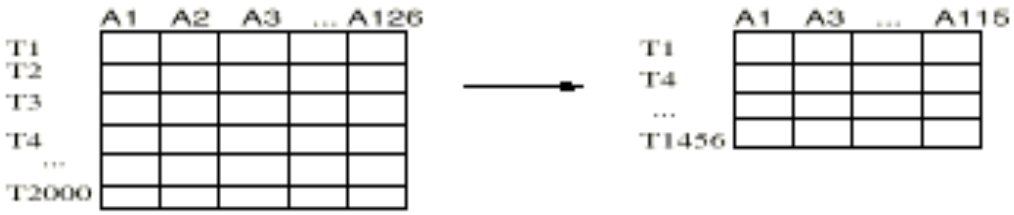
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- **Data is not always available**
  - E.g., **many tuples have no recorded value for several attributes**, such as **customer income in sales data**
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
- Missing data may **need to be inferred**

How to handle missing data?

- **Ignore the tuple:** usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- **Fill in the missing value manually: tedious + infeasible?**
  - Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
  - Use the attribute mean to fill in the missing value
  - Use the attribute mean for all samples of the same class to fill in the missing value: smarter
  - Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

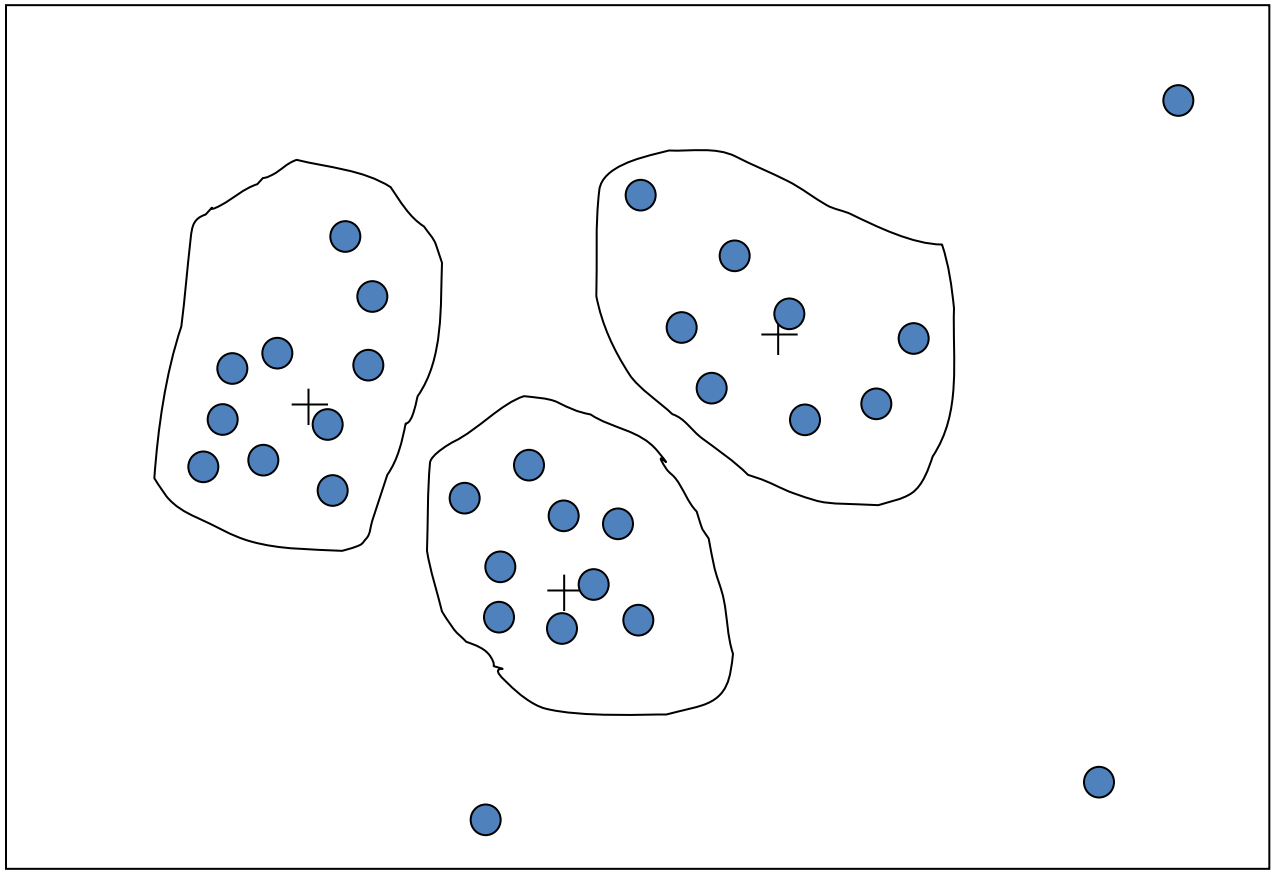
Noisy Data

- Q: What is noise?
  - A: Random error in a measured variable.
  - Incorrect attribute values may be due to
    - faulty data collection instruments
    - data entry problems
    - data transmission problems
    - inconsistency in naming convention
  - Other data problems which requires data cleaning
    - duplicate records
    - incomplete data
    - inconsistent data

How to handle noisy data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - used also for discretization
- Clustering
  - detect and remove outliers
- Semi-automated method: combined computer and human inspection
  - detect suspicious values and check manually
- Regression
  - smooth by fitting the data into regression functions (next week)

Cluster Analysis



Data integration

- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units, different currency

Data transformation

- **Normalization:** scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

Particularly useful for classification (Distance measurements, NN classification, etc)

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where *j* is the smallest integer such that Max(| v' |)<1

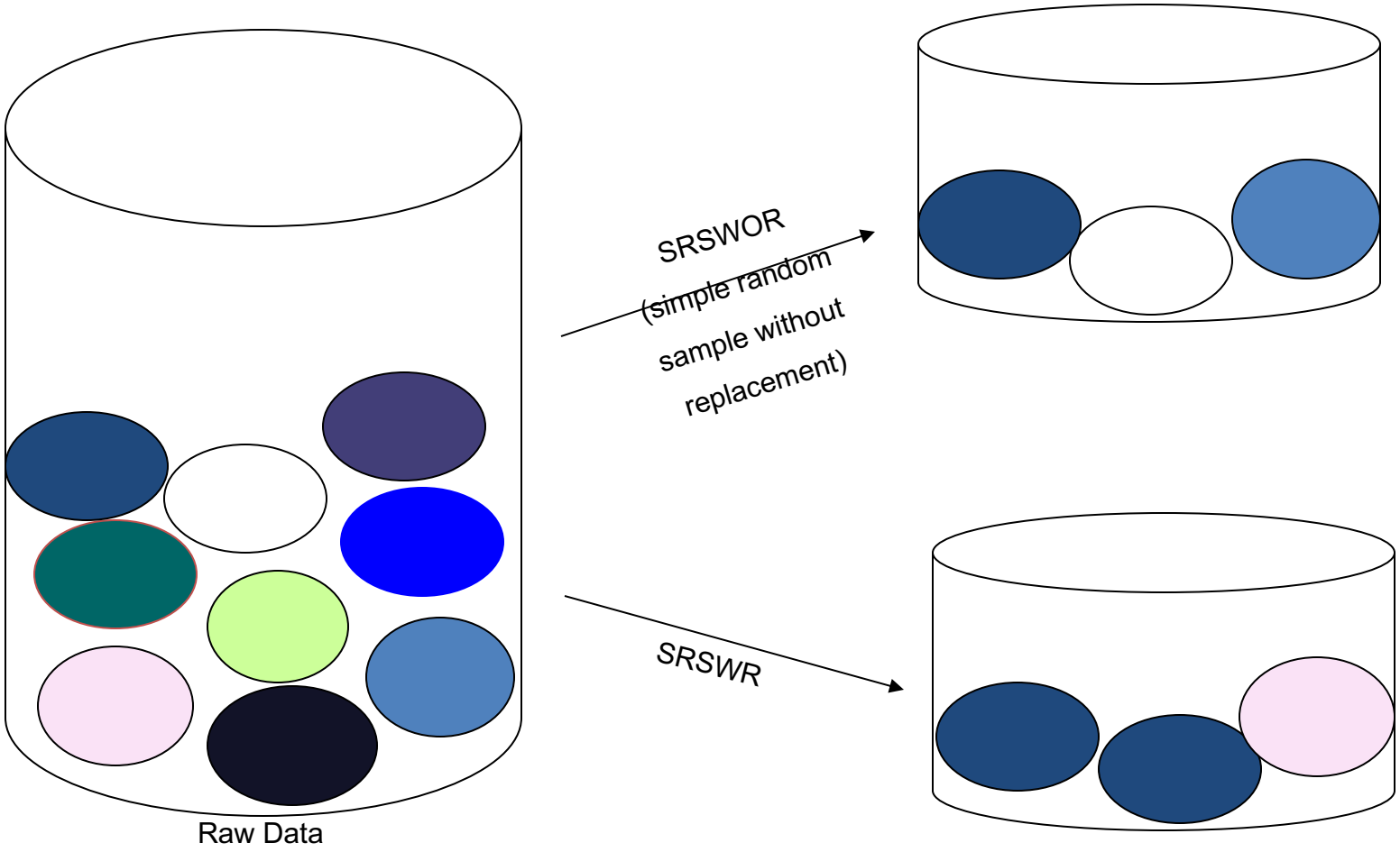
Data reduction

- Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
  - Dimensionality reduction
  - Data compression
  - Numerosity reduction
  - Discretization and concept hierarchy generation

Sampling

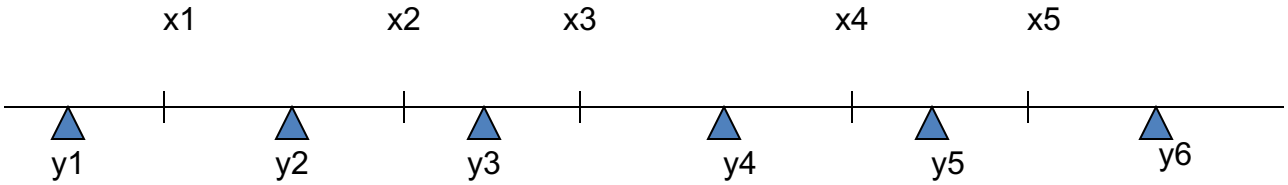
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Cost of sampling: proportional to the size of the sample, increases linearly with the number of dimensions
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- **Develop adaptive sampling methods**
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Sampling may not reduce database I/Os (page at a time).

Sampling


SRSWOR
(simple random
sample without
replacement)
SRSWR
Raw Data

Discretization/Quantization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization/Quantization:
  - ☛divide the range of a continuous attribute into intervals



  - Some classification algorithms only accept categorical attributes.
  - **Reduce data size by discretization**
  - Prepare for further analysis

Summary

- Data acquisition and preparation
- Data preprocessing includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- What's next:
  - Model validation and evaluation