

---

# CMPT 733

# Data Preparation

Instructor

Zhengjie Miao

Course website

<https://coursys.sfu.ca/2025sp-cmpt-733-gl/pages/>

Source

based on slides by Jiannan Wang

# Outline

---

1. Data Preparation Overview
2. Data Preparation Tasks

# Data Preparation Is **Still** the Bottleneck!!!

2014

The New York Times

## *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*

Yet far too much handcrafted work — what data scientists call “data wrangling,” “data munging” and “data janitor work” — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

<https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>

2020

 ANACONDA

## The State of Data Science 2020 Moving from hype toward maturity

We were disappointed, if not surprised, to see that data wrangling still takes the lion's share of time in a typical data professional's day. Our respondents reported that almost half of their time is spent on the combined tasks of data loading and cleansing. Data

<https://www.anaconda.com/state-of-data-science-2020>

# Trend: Data Prep about 38% of effort

2022

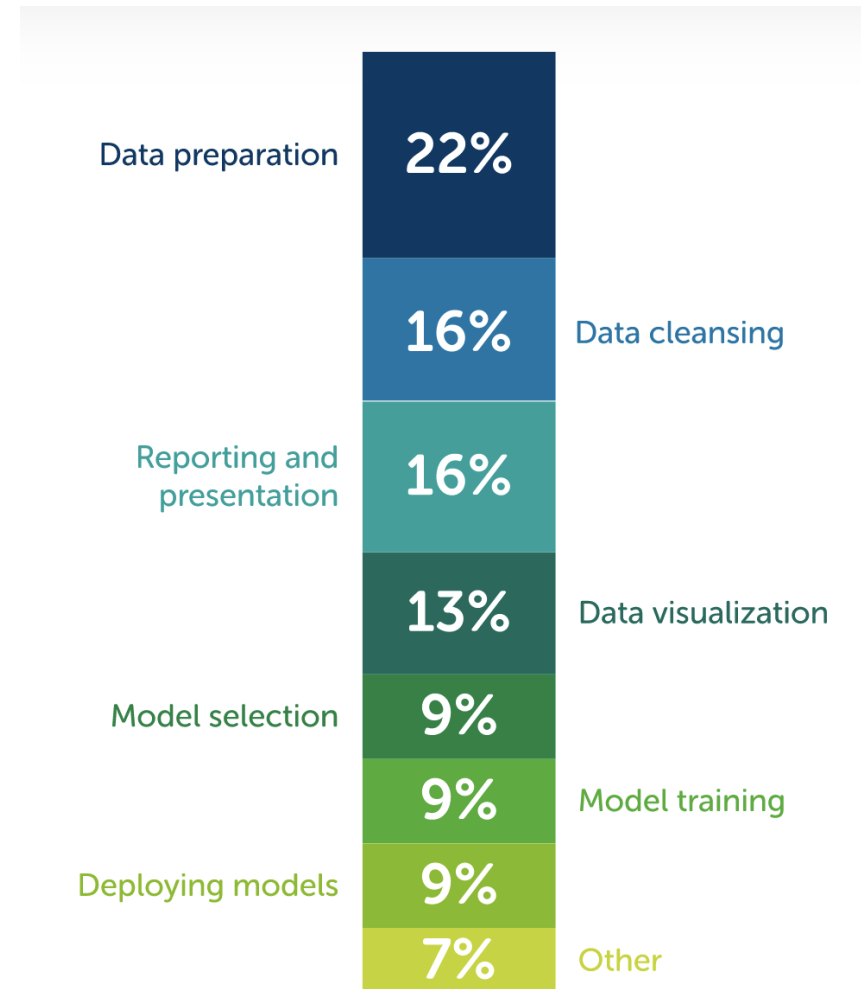


## DATA PROFESSIONALS AT WORK

### How do data scientists spend their time?

Data professionals spend their time on a variety of tasks that require diverse technical and non-technical skills. Respondents indicated they spend about 37.75% of their time on data preparation and cleansing. Beyond preparing and cleaning data, interpreting results remains critical. **Data visualization** (12.99%) and demonstrating data's value through reporting and presentation (16.20%) are essential steps toward making data actionable and providing answers to critical questions. Working with models through selection, training, and deployment takes about 26.44% of respondents' time (-8.56% YoY).

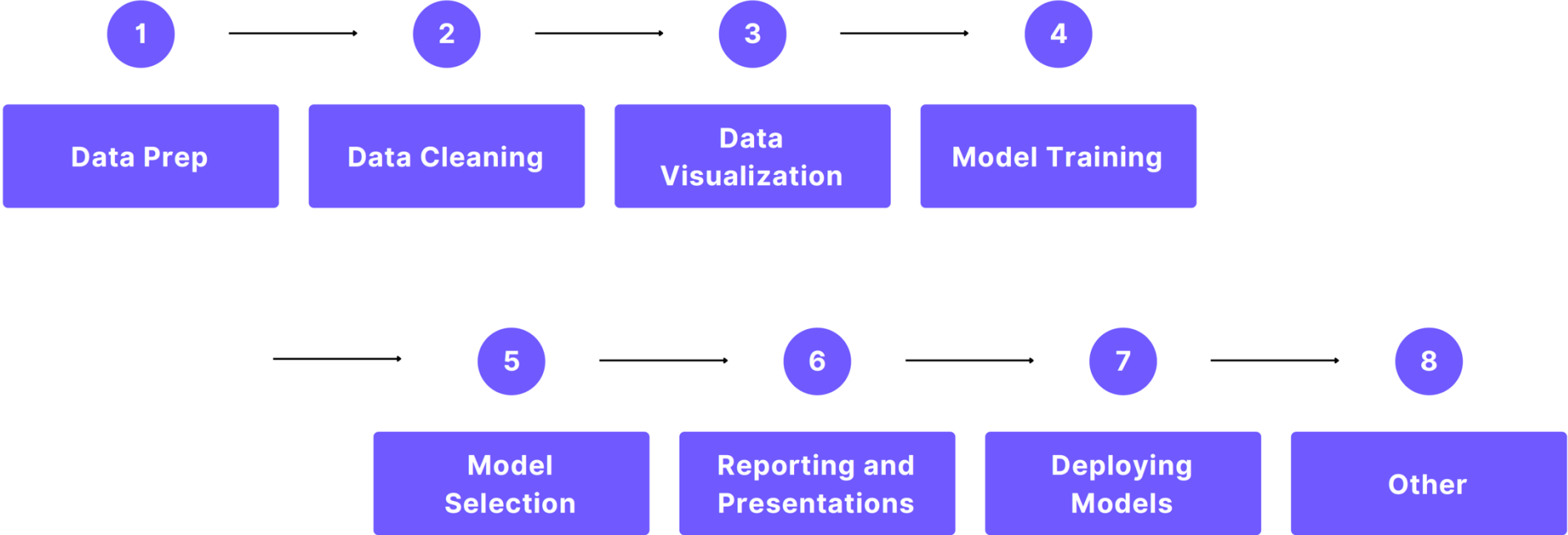
<https://www.anaconda.com/state-of-data-science-2022>



n = 1,966

# Trend in 2023

Thinking about your current role, what tasks are most time consuming? (Responses ranked from most to least time consuming)



n=1,071

# Trend in 2023

Generative AI has been in the news for some time now, with the introduction of high-performing large language models (LLMs). In our data science practitioner track, **40% of respondents say their companies are working on internal generative AI tools, such as LLMs.** While there are many conversations about the ethics and

**The majority (63%) of data science practitioners say they're using generative AI the same amount or more this year compared to 2022.** Respondents who report using these tools and techniques in their work most commonly use them for content creation (e.g., text or image generation) and data cleaning, visualization, and analysis. Less common use cases include automating tasks and writing and debugging code.

# Why Is Data Preparation Hard?



Collection



Cleaning



Integration



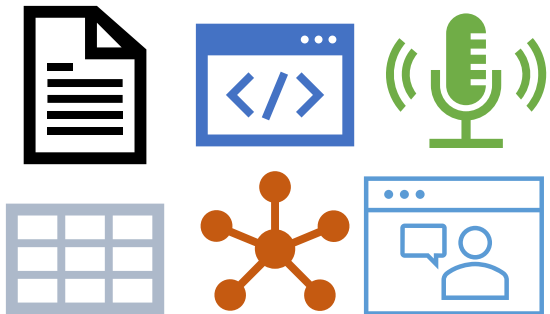
Analysis

How much time is spent on preparation?

1. **Too many small problems** (e.g., standardize date, dedup address, etc)
2. Humans have **different levels of expertise** (in data science and programming)
3. **Domain specific** (finance, social science, healthcare, economics, etc.)

# Data come from many sources and in a variety of formats

Source Data  
(Raw Text, HTML,  
Tweeter, Tiktok,...)



Structured	Semi-structured	Raw/Unstructured
CSV	JSON	Text
TSV	XML	Images
Excel	HTML	Audio
Dataframe	Python Pickle	Video
SQL	MongoDB	
...	...	



# Data are noisy

---

- Data entry errors
- Measurement errors
- Extraction errors
- Format conversion errors
- Non-uniform collection/sampling
- ...

# Human-in-the-loop Data Preparation

---

- Three Directions
- Spreadsheet GUI
- Workflow GUI
- Notebook GUI

# Spreadsheet GUI

CUSTOMER ANALYSIS > customer Random

Pattern Details CONTRACT\_END

#	IMSI	CONTRACT_END	CONTRACT_START	#	SUBSCRIBER_AGE	RBC	STATUS
310T - 310.26T		Jan 2013 - Dec 2016	Jan 2000 - Dec 2014	0 - 15		2 Categories	
310170226812721		6/4/16	7/29/09				ACTIVE
310160900766700		3/28/15	10/6/13	1			ACTIVE
310170546822541		9/23/16	1/9/07	7			ACTIVE
310005432849230		5/29/15	2/14/01	13			ACTIVE
310026939721905		9/11/15	9/18/10	4			ACTIVE
310026015466952		8/27/15	3/13/06	8			ACTIVE
310170484724861		1/16/16	5/11/04				ACTIVE
310170765640471		05-Jul-2011	9/11/06	4			INACTIVE
310260310245556		12/24/15	3/28/01	13			ACTIVE
310150834295817		3/6/15	7/26/00	14			ACTIVE
310160464252516		9/25/15	4/4/04	10			ACTIVE
310120438750772		4/30/16	9/8/04	10			ACTIVE
310260195729676		1/16/15	1/3/04	11			ACTIVE
310026261822880		8/13/13	11/23/08	4			INACTIVE
310005667082048		8/4/16	10/22/14				ACTIVE
310170836020164		1/22/15	10/19/14	0			ACTIVE
310160772267782		11/21/15	12/28/14				ACTIVE
310170116249240		27-Sep-2011	2/9/09				INACTIVE
310026110612337		5/29/15	3/29/05	9			ACTIVE
310260681676970		11/17/16	5/21/07	7			ACTIVE
310004436630316		9/15/16	7/24/11				ACTIVE
310120423699542		2/27/15	6/29/11	3			ACTIVE
310120773194729		4/28/16	6/15/04	10			ACTIVE
310030295859214		2/7/15	3/24/12	2			ACTIVE
310012150088547		13-Jan-2009	12/10/05	3			INACTIVE
310120387060694		10/1/16	10/25/11	3			ACTIVE

19 Columns 20,000 Rows 8 Data Types

Show only affected  Rows

12.65k

5.37k

Trifacta

# Workflow GUI

The screenshot displays the Alteryx Workflow Designer interface. The top menu bar includes File, Edit, View, Options, and Help. Below it is a toolbar with various tool icons. The main workspace is divided into three sections: **Input Data**, **Data Preparation**, and **Data Blend**.

- Input Data:** Two data sources are connected: **Customers CRM** and **Transactions AWS**.
- Data Preparation:** The CRM data flows through a **Crosstab** tool and a **Filter** tool. The AWS data flows through a **Formula / Calculate Fields** tool and a **Summarize / Pivot Table** tool.
- Data Blend:** The outputs from the Filter and Summarize tools are joined using a **Join** tool. The output of the Join tool then flows through a **VLOOKUP** tool. A **Visualitytics** tool is also connected to the workflow.

On the left side, a **Profile** panel for the **ZIP** field is visible, showing a bar chart of frequency and a data quality summary. The data quality summary indicates 0.0% NOT OK, 0.0% NULL, 0.0% EMPTY, and 100.0% OK. Below the summary is a table with the following data:

Field	Data Type	Size	Non-Nulls	Uniques	Nulls	Blanks	Values with Leading Whitespace	Values with Trailing Whitespace	Shortest (Non-Blank) Length	Average Length
ZIP	V_String	255	1735	86	0	0	0	0	5	5.0

At the bottom, a **Results - Browse (41) - Input** panel shows a table with 14 fields and 1,735 records displayed. The table columns are: Record #, Customer ID, Address, City, Customer Segment, First Name, Last Name, Responder, State, Store Number, Suite, ZIP, and Cu.

Record #	Customer ID	Address	City	Customer Segment	First Name	Last Name	Responder	State	Store Number	Suite	ZIP	Cu
1	5	5360 Zuni St	Denver	Home Office	LINDA	TREVINO	No	CO	100	[Null]	80221	5
2	6	1599 Williams St.	Denver	Home Office	H	MACK	No	CO	106	[Null]	80218	6
3	7	12066 E Lake Cir	Greenwood Village	Home Office	MARISSA	LATTA	No	CO	105	[Null]	80111	7
4	8	7225 S Gaylord St	Centennial	Home Office	PHYLLIS	WALKER	No	CO	101	[Null]	80122	8
5	9	4497 Cornish Way	Denver	Home Office	VIVIAN	GAULDEN	No	CO	105	[Null]	80239	9

alteryx

# Notebook GUI

jupyter dataprep Last Checkpoint: 9 minutes ago (unsaved changes)

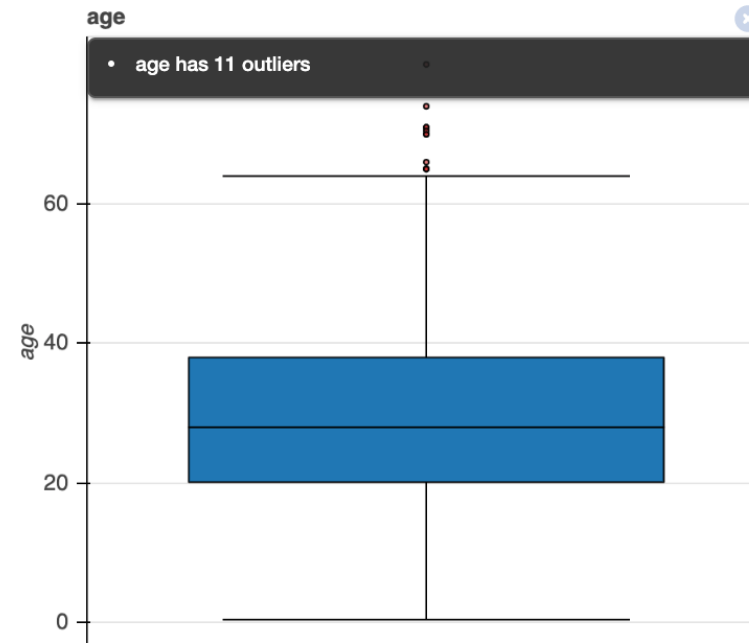
File Edit View Insert Cell Kernel Widgets Help

📁 + 🔍 📄 ⬆️ ⬆️ ▶️ Run 🛑 ↺ ▶️ Code ▾ 🖨️

```
In [7]: from dataprep.eda import plot
        from dataprep.datasets import load_dataset, get_dataset_names

        df = load_dataset("titanic")
        plot(df, "age")
```

Out[7]: Stats Histogram KDE Plot Normal Q-Q Plot **Box Plot**



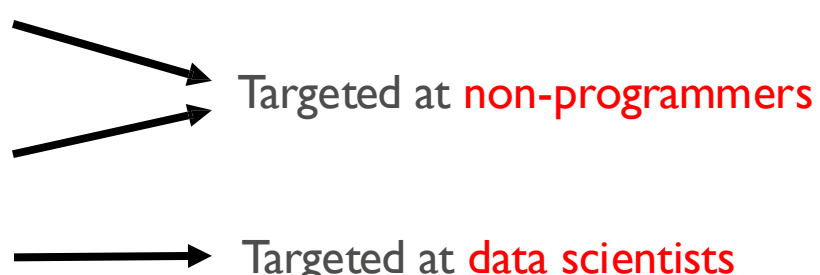
# Which Direction To Go?

---

“ Data Prep Market was valued at USD 4.02 Billion in 2024 and is projected to reach **USD 16.12 Billion by 2031**, growing at a **CAGR of 19% from 2024 to 2031** ”

Source: <https://www.verifiedmarketresearch.com/product/data-prep-market/>

## Three Directions

- Spreadsheet GUI
  - Workflow GUI
  - Notebook GUI
- Targeted at **non-programmers**
- Targeted at **data scientists**
- 

# Data Preparation Tasks

---

## Data Collection

- Where to collect
- How to collect

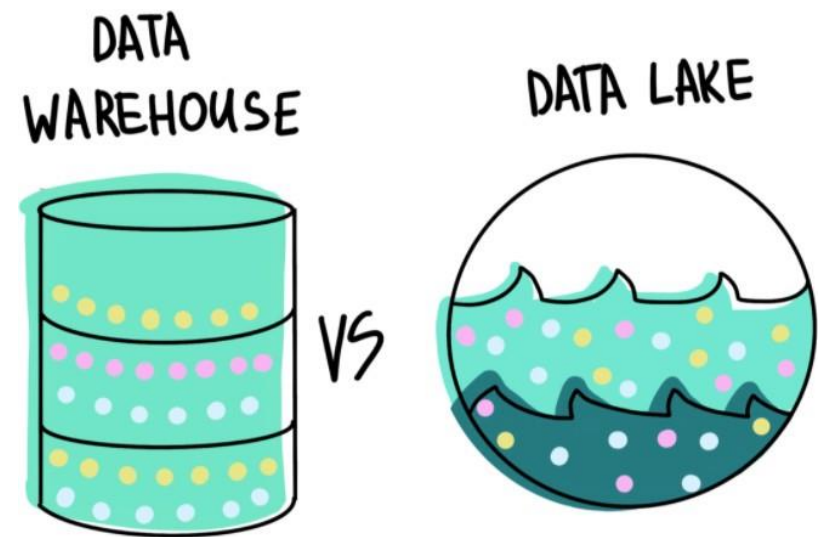
## Data Cleaning

## Data Integration

# Where to Collect?

## Internal Data

- Data Warehouse (Tabular Data)
- System Logs (Text Files)
- Documents (Word, Excel, PDF)
- Multimedia Data (Video, Audio, Image)



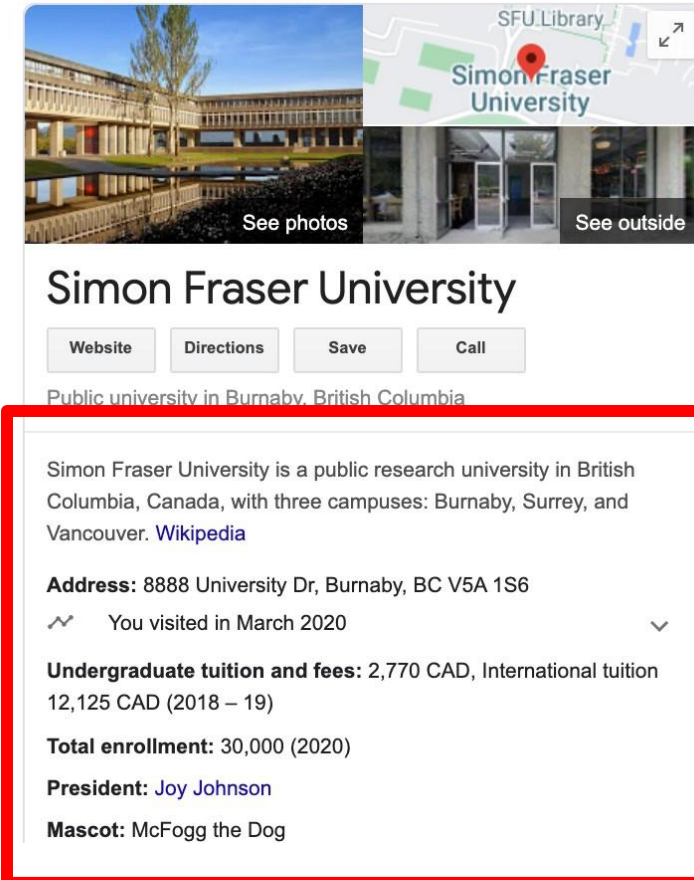
[Aside: [AWS Data Lake on S3](#)]



# Where to Collect?

## External Data

- Web Pages
- Web APIs (<https://github.com/public-apis/public-apis>)
- Open Data ([data.vancouver.ca](http://data.vancouver.ca), [www.data.gov](http://www.data.gov))
- Knowledge Graph (Wikidata, Freebase)



Simon Fraser University

Public university in Burnaby, British Columbia

Simon Fraser University is a public research university in British Columbia, Canada, with three campuses: Burnaby, Surrey, and Vancouver. [Wikipedia](#)

**Address:** 8888 University Dr, Burnaby, BC V5A 1S6

📍 You visited in March 2020

**Undergraduate tuition and fees:** 2,770 CAD, International tuition 12,125 CAD (2018 – 19)

**Total enrollment:** 30,000 (2020)

**President:** Joy Johnson

**Mascot:** McFogg the Dog

# Data Classification

## Data

### Structured

Team	W	L	Pct	GB	Conf	Div	Home	Away	L10	Strk
1  Rockets	20	4	.833	-	11-2	3-2	8-3	12-1	9-1	W9
2  Warriors	21	6	.778	0.5	9-4	2-1	8-3	13-3	8-2	W6
3  Spurs	19	8	.704	2.5	9-4	4-1	13-2	6-6	8-2	W4
4  Timberwolves	16	11	.593	5.5	13-5	4-1	9-4	7-7	6-4	W2

### Semi-structured

CLE - James Layup Shot: Missed	06:48 CLE
CLE - James Rebound (Off:1 Def:0)	06:46 CLE
CLE - James Reverse Layup Shot: Made (2 PTS)	06:45 CLE 9-15
Stoppage: Out-of-Bounds	06:29

### Unstructured

#### Is LeBron breaking the aging curve?

Kevin Pelton  
ESPN Staff Writer 5:10 AM PT

During his 15th NBA season, [Cleveland Cavaliers](#) star [LeBron James](#) is performing at a level that echoes the prime that saw him win four MVPs.

As James nears his 33rd birthday later this month, his performance at that age stands up to any of his predecessors, including [Michael Jordan's](#) 1995-96 season that produced an MVP, a then-record 72 wins and a championship. (Because James entered the NBA directly out of high school, NBA experience isn't the best way to compare how he's aging to his peers. After all, Jordan's 15th year was actually his final one in the NBA at age 40.)

# Challenges

- Data Discovery

- How to find related data?

- Domain knowledge
- Information retrieval skills

- Data Privacy

- How to protect user privacy?

- Data masking
- Differential privacy

- Security

- How to avoid a data breach?

- Follow data access rules
- Encrypt highly confidential data

# Getting Data

---

- From CSV Files
- From JSON Files
- From the Web
- From HDFS
- From Databases
- From S3
- From Web APIs

# Load Data From CSV Files

CSV is a file format for storing tabular data

```
rankings.csv ×
1 Team,Win,Loss,Win%
2 Houston Rockets,20,4,0.833
3 Golden State Warriors,21,6,0.778
4 San Antonio Spurs,19,8,0.704
5 Minnesota Timberwolves,16,11,0.593
6 Denver Nuggets,14,12,0.538
7 Portland Trail Blazers,13,12,0.52
8 New Orleans Pelicans,14,13,0.519
9 Utah Jazz,13,14,0.481
10
```

## Reading CSV File (pandas library)

```
import pandas as pd
df = pd.read_csv('rankings.csv')
```

# Load Data From JSON Files

JSON is a file format for storing nested data (array, dict)

```
players.json *
1 {
2   "Kobe Bryant" :{
3     "Born": "08/23/1978",
4     "Number": ["8", "24"],
5     "Team": ["Los Angeles Lakers"]
6   },
7   "Michael Jordan":{
8     "Born": "02/17/1963",
9     "Number": ["23"],
10    "Team": ["Chicago Bulls", "Washington Wizards"]
11  }
12 }
```

## Reading JSON File (pandas Libarary)

```
import pandas as pd
df=pd.read_json("players.json")
```

# Web Scraping

---

- Open web pages
  - urllib2 (<https://docs.python.org/2/library/urllib2.html>)
  - request (<http://docs.python-requests.org/en/master/>)
- Parse web pages
  - BeautifulSoup (<https://www.crummy.com/software/BeautifulSoup/>)
  - lxml (<http://lxml.de/>)
- Putting everything together
  - Scrapy (<https://scrapy.org/>)

# Before you scrape

---

Check to see if CSV, JSON, or XML version of an HTML page are available – better to use those

Check to see if there is a Python library that provides structured access (e.g., dataprep)

Check that you have permission to scrape

From [“Deb Nolan. Web Scraping & XML/Xpath”](#)



# If you do scrape

---

- Be careful to not to overburden the site with your requests
- Test code on small requests
- Save the results of each request so you don't have to repeat the request unnecessarily
- CAPTCHA



From "[Deb Nolan. Web Scraping & XML/Xpath](#)"

# Example Application Scenario: Experiential Hotel Search

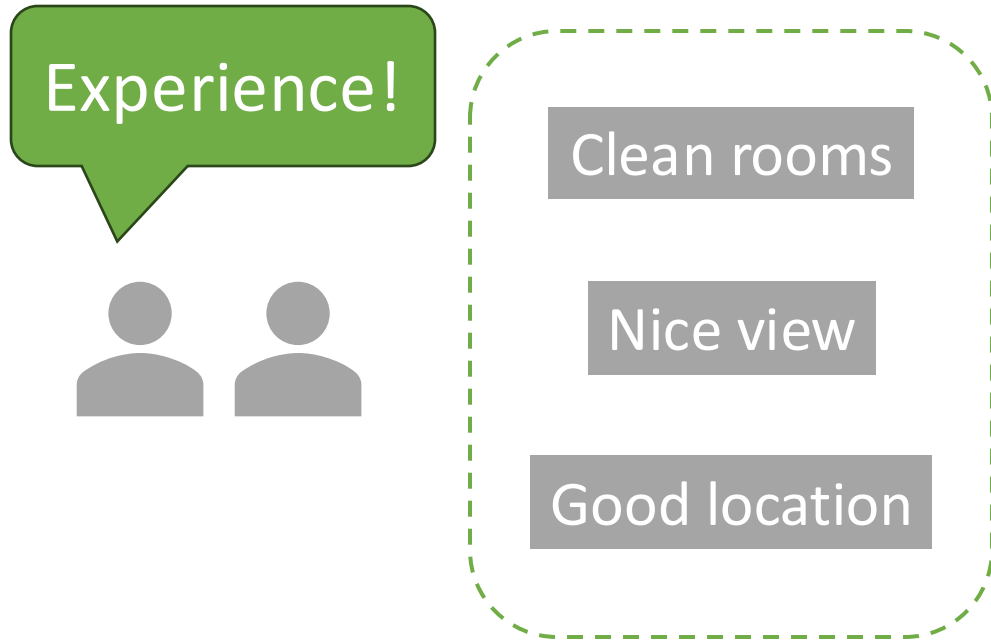
---



Hotel that is less than \$300 per night, has very clean room, and has a nice view of the city in Vancouver

Disclaimer: this is an imaginary application, not about how the real-world examples are built

# Where can we get those subjective attributes?



Reviewed: January 16, 2024

9.0

It was an exceptional stay. Highly recommended.

😊 · It was so close to most attractions and everything is almost of walking distance. The room they gave us had an excellent **view**. Customer service was top-notch. We were checked in 3 hours prior to check-in time and after checking out they stored our luggages so we can still walk in the area without carrying our luggages around. Maintenance attended to our issue right away when something broke, cleaning staff was friendly.

😞 · None

👍 Helpful    🗨️ Not helpful

Reviewed: July 8, 2023

8.0

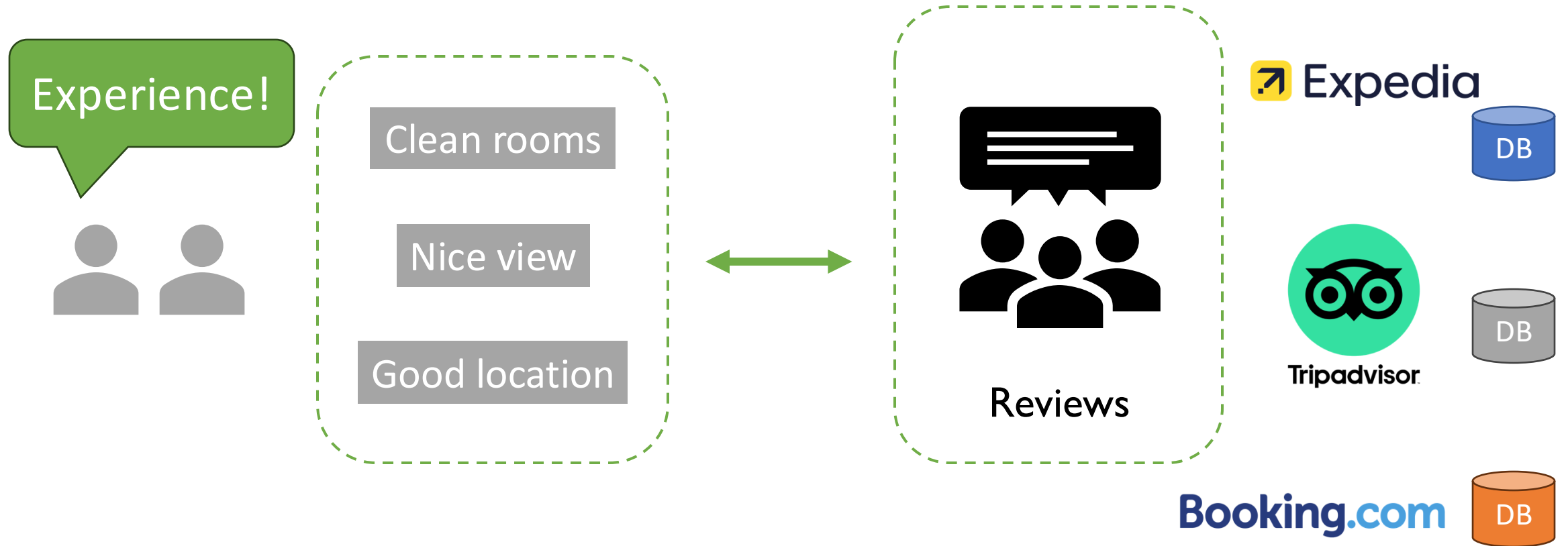
Very satisfied

😊 · Staff is very nice. Hotel is clean. Nice **views**.

😞 · Upon arrival they had not given me the right room I had booked with booking.com. But they fixed it immediately once I brought it up.

👍 Helpful    🗨️ Not helpful

# Experiential Search



# Preparing data for the search application

---

- Web scraping
- Information extraction (perhaps will cover in future lectures)
- Data cleaning
- Data integration
- Data annotation

# Outline

---

- Data Collection
- Data Cleaning
  - Dirty Data Problems
  - Data Cleaning Tools
  - Example: Outlier Detection
- Data Integration

# Data Cleaning

NEWS

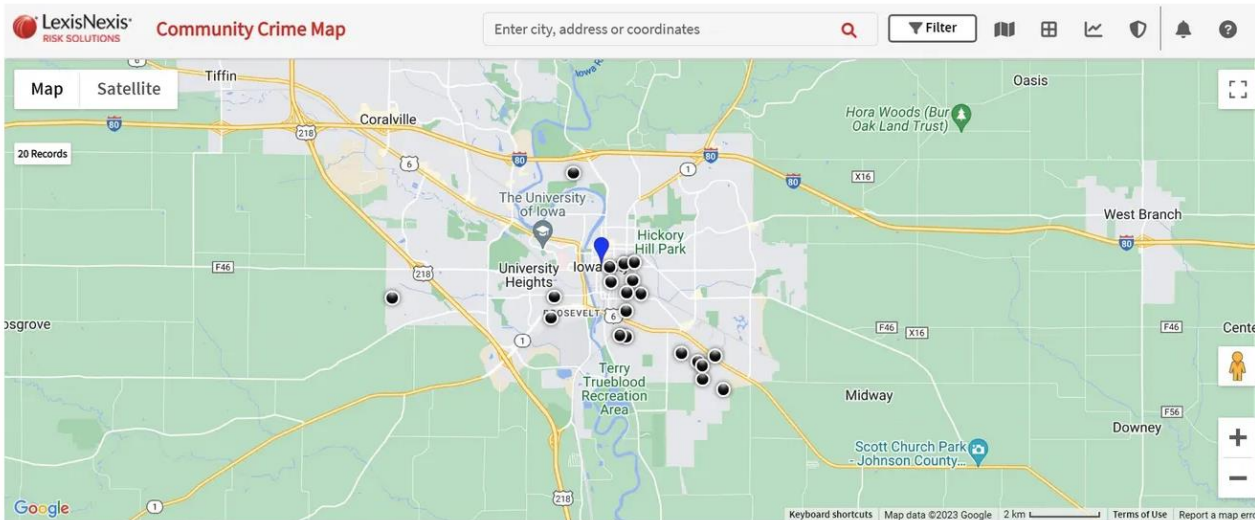
## Iowa City crime map shows 20 homicides in 2023. It's wrong.



**George Shillcock**

Iowa City Press-Citizen

Published 9:37 a.m. CT March 17, 2023



Because of a technical glitch, Iowa City's community crime map is showing that 20 homicides occurred in the city between Jan. 1, 2023 and March 1, 2023. They didn't. *George Shillcock*

He explained in an email that there had been a large-scale reclassification of National Incident-Based Reporting System crime types in 2022. The city has an automated system that uses the crime types when it exports information about criminal activity from its records system to the LexisNexis site that manages the mapping.

Somewhere along the way, Hermiston said, the export process, relying on old codes, failed to capture the crime classifications properly. So suddenly the map reflected what looked like a brutal murder spree.

<https://www.press-citizen.com/story/news/2023/03/17/iowa-city-crime-map-glitch-mistakenlhy-shows-20-homicides-in-2023/69959003007/>

# Dirty Data Problems

---

- 1) Parsing text into fields (separator issues)
- 2) Missing required field (e.g. no SIN)
- 3) Different representations (iphone 2 vs iphone 2<sup>nd</sup> generation)
- 4) Fields too long (get truncated)
- 5) Formatting issues – especially dates
- 6) Primary key violation (two people with the same SIN)
- 7) Redundant Records (exact match or other)
- 8) Outliers (age = 120)

- Adapted from Stanford Data Integration Course



# Data Cleaning --- Error Detection

	Name	City	Star Rating	Minimum Price	Zip
t <sub>1</sub>	Sheraton Vancouver - Wall Centre	<b>Burnaby</b>	★ ★ ★ ★	\$260	V6Z 2R9
t <sub>2</sub>	<b>TheBurrardHotel</b>	Vancouver		\$200	V6Z 1Y7
t <sub>3</sub>	Element Metrotown	Burnaby	★ ★ ★	<b>\$3000</b>	V5H 2A7
t <sub>4</sub>	Embarc Hotel	Vancouver	★ ★ ★ ★	\$180	V6Z 2R9

Typos

Duplicated values

Outliers

Missing values

Constraint violation

# Another Annoying Example

Transactions	Amount	Currency	Mode
10 Dollars, credit	10	Dollars	credit
5Euros - debit	5	Euros	debit
30 Pesos, credit	30	Pesos	credit
1 dollar, credit	1	Dollars	credit
7,00 euros, debit	7	Euros	debit
credit 20 dollars	20	Dollars	credit
21 pesos	21	Pesos	?
debit Fifty pesos and 10 euros	50	Pesos	debit
	10	Euros	debit



From Brandon Fain (Duke)

# Data Cleaning Tools

## Python

- [Missing Data](#) (Pandas)
- [Deduplication](#) (Dedup)

## OpenRefine

- Open-source Software (<http://openrefine.org>)
- OpenRefine as a Service ([RefinePro](#))

## Data Wrangler

- The Stanford/Berkeley Wrangler research project
- Commercialized ([Trifacta](#))

Not Many Tools.  
That's why we are building DataPrep  
(<http://dataprep.ai>)

```
1 import pandas as pd
2 from dataprep.clean import clean_country
3 df = pd.DataFrame({"country": ["USA", "country: Canada", " France ",
4                               "233", " tr "]})
4 clean_country(df, "country")
```

	country	country_clean
0	USA	United States
1	country: Canada	Canada
2	France	France
3	233	Estonia
4	tr	Turkey

# Outlier Detection

---

The ages of employees in a US company

1 20 21 21 22 26 33 35 36 37 39 42 45 47 54 57 61 62

$$\text{Mean} = \frac{1}{n} \sum_{i=1}^n x_i = 37$$

$$\text{Stddev} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \text{mean})^2} = 16$$

$$[37 - 2 * 16, 37 + 2 * 16] = [4, 70]$$

# Outlier Detection

The ages of employees in a US company

1 20 21 21 22 26 33 35 36 37 39 42 45 47 54 57 61 62 400

$$\text{Mean} = \frac{1}{n} \sum_{i=1}^n x_i = 56$$

$$\text{Stddev} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \text{mean})^2} = 83 \quad [56 - 2 * 83, 56 + 2 * 83] = [-109, 221]$$

# Outlier Detection

---

The ages of employees in a US company

1 20 21 21 22 26 33 35 36 37 39 42 45 47 54 57 61 62 400

$$\text{Median} = \text{median}(X) = 37$$

$$[37 - 2 * 15, 37 + 2 * 15] = [7, 67]$$

$$\text{MAD} = \text{median}(X - \text{median}(X)) = 15$$

# Data Preparation Tasks

---

- Data Collection
- Data Cleaning
- Data Integration
  - Data Integration Problem
  - Three Steps (Schema Matching, Entity Resolution, Data Fusion)
  - Example: Entity Resolution

# Google Thinks I'm Dead

(I know otherwise.)

Me

could be me...?

Rachel Abrams  
American writer

Rachel Abrams was an American writer, editor, and artist. She was the wife of Elliott Abrams. [Wikipedia](#)

**Born:** January 2, 1951

**Died:** June 7, 2013

**Spouse:** Elliott Abrams (m. 1980–2013)

**Parents:** Midge Decter

**Children:** Sarah Abrams, Jacob Abrams, Joseph Abrams

People also search for



Adam Sroka · 2nd  
I build data and AI platforms for the energy sector...  
21h · 🌐

+ Follow

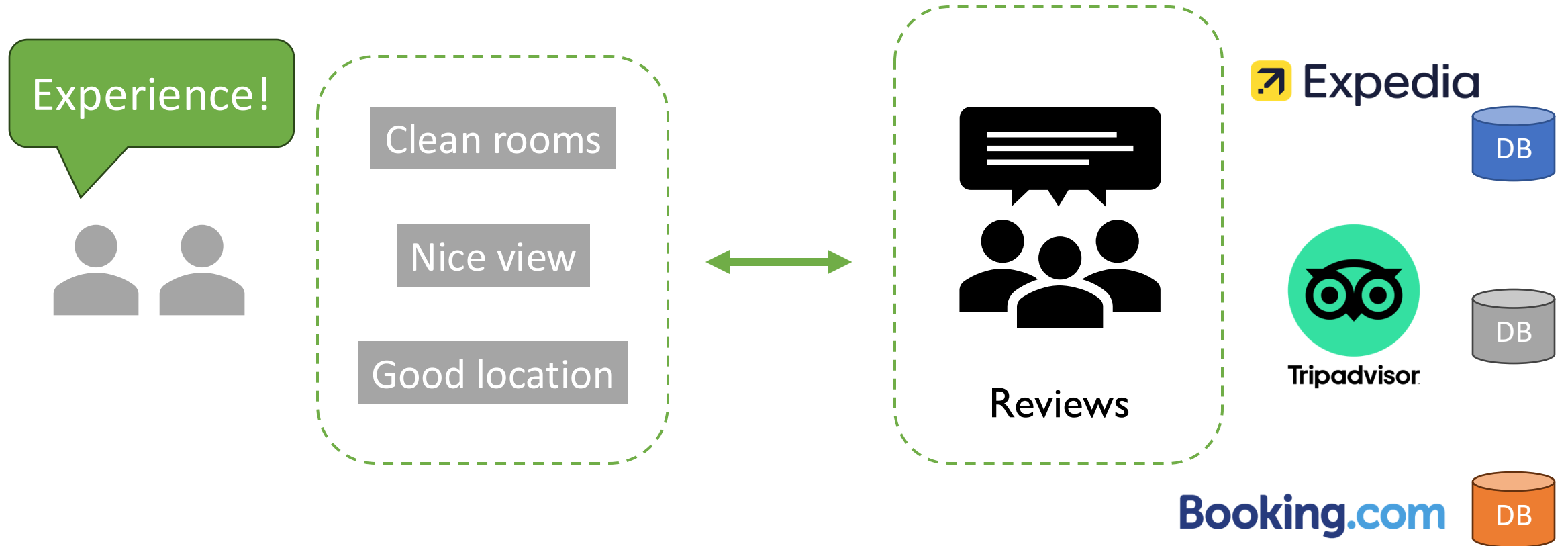
Management: We have great datasets

The datasets:

```
[ 'St. Albans',  
  'St. Albans',  
  'St Albans',  
  'St. Ablans',  
  'St. albans',  
  'St. Alans',  
  'S. Albans',  
  'St.. Albans',  
  'S. Albnas',  
  'St. Albnas',  
  'St. Al bans',  
  'St. Algans',  
  'Sl. Albans',  
  'St. Allbans',  
  'St, Albans',  
  'St. Alban',  
  'St. Alban']
```

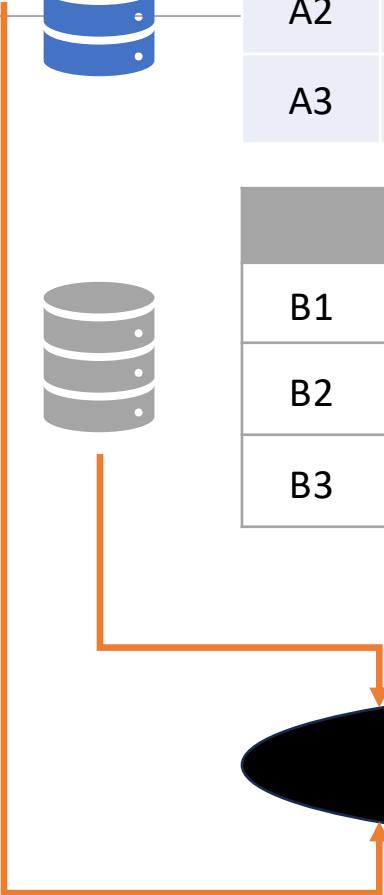
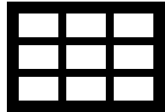


# Experiential Search



	Name	Address	Tel
A1	Sheraton Vancouver - Wall Centre	1000 Burrard Street , Vancouver	(604) 331-1000
A2	The Burrard Hotel	1100 Burrard Street, Vancouver	(604) 681-2331
A3	Delta Hotels BCC	4331 Dominion Street, Burnaby	(604) 453-0750

	Zip	Hotel	Phone
B1	V4K 0B2	Delta Hotels by Marriott Vancouver Delta	+1 604-382-8222
B2	V6Z 1Y7	The Burrard	+1 800-663-0366
B3	V5G 1C7	Delta Hotels by Marriott Burnaby Conference Centre	+1 604-453-0750



# Data Integration: Three Steps

---

## Schema Mapping

- Creating a global schema
- Mapping local schemas to the global schema

## Entity Resolution

- You will learn this in detail later

## Data Fusion

- Resolving conflicts based on some confidence scores

## Want to know more?

- Anhai Doan, Alon Y. Halevy, Zachary Ives. [Principles of Data Integration](#). Morgan Kaufmann Publishers, 2012.

# Schema Mapping

---

- Approaches:
  - View & logic-based: mapping between sources and the global schema
  - Learning to match:
    - Classify the semantic relation of attribute pairs
    - Cluster attributes
  - Universal schema
    - Fill in each cell by latent probabilities instead of one-to-one column mapping
    - Allow overlap, subset/superset, etc.
  - With the rise of LMs
    - Simply concatenating all values could work for certain tasks



	Name	Address	Tel
A1	Sheraton Vancouver - Wall Centre	1000 Burrard Street , Vancouver	(604) 331-1000
A2	The Burrard Hotel	1100 Burrard Street, Vancouver	(604) 681-2331
A3	Delta Hotels BCC	4331 Dominion Street, Burnaby	(604) 453-0750






	Zip	Hotel	Phone
B1	V4K 0B2	Delta Hotels by Marriott Vancouver Delta	+1 604-382-8222
B2	V6Z 1Y7	The Burrard	+1 800-663-0366
B3	V5G 1C7	Delta Hotels by Marriott Burnaby Conference Centre	+1 604-453-0750

Integrated Data



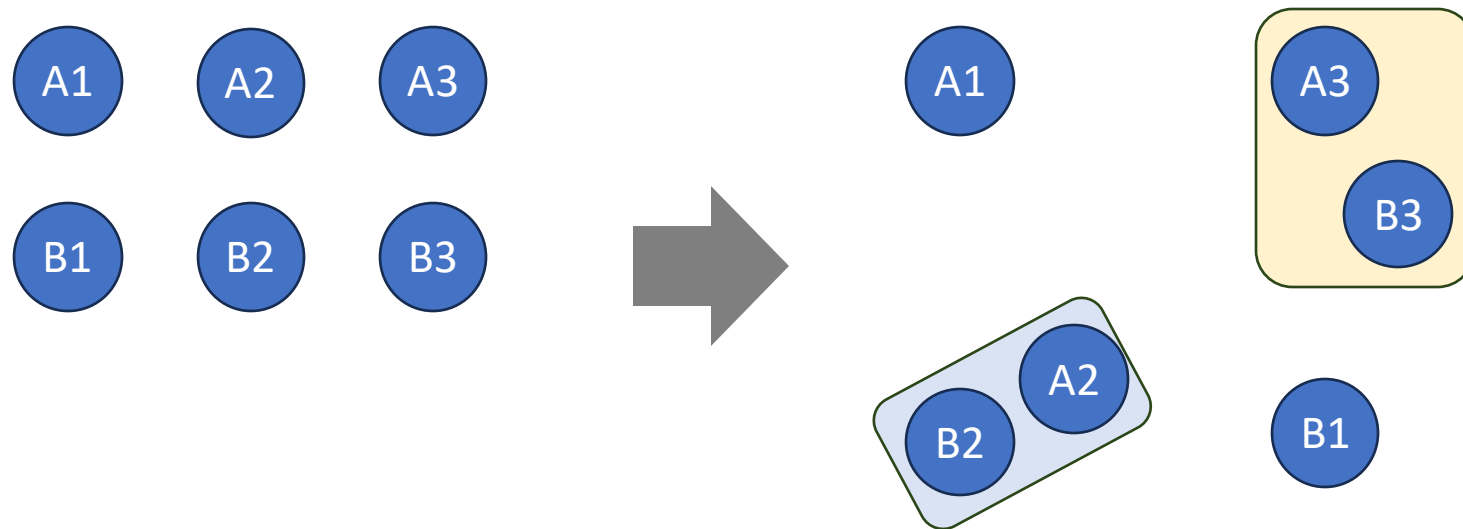
	Name	Address	Phone	Zip
H1	Sheraton Vancouver - Wall Centre	1000 Burrard Street , Vancouver	(604) 331-1000	
H2	The Burrard Hotel	1100 Burrard Street, Vancouver	+1 800-663-0366	V6Z 1Y7
H3	Delta Hotels BCC	4331 Dominion Street, Burnaby	+1 604-453-0750	V5G 1C7
H4	Delta Hotels by Marriott Vancouver Delta		+1 604-382-8222	V4K 0B2

# Another Example of Entity Resolution

	<p>Apple iPad 2 <b>MC775LL/A</b> Tablet (64GB Wifi + AT&amp;T 3G Black) <b>NEWE</b></p> <p>Apple iPad XX6LL/A Tablet (64GB, Wifi + AT&amp;T 3G, Black) NEWEST MODEL</p>	<p><b>\$660</b> and up (3 stores)</p> <p><input type="checkbox"/> Compare (Share and Compare)</p>
	<p>Apple iPad 2 <b>MC775LL/A</b> 9.7" LED 64 GB Tablet Computer - Wi-Fi - 3G ...</p> <p><b>Brand</b> Apple - <b>Weight</b> 1.40 lb - <b>Screen size</b> 9.70 in</p> <p>There's more to it. And even less of it. Two cameras for FaceTime and HD video recording. The dual-core A5 chip. The same 10-hour battery life. All in a thinner, lighter design.... <a href="#">more...</a></p>	<p><b>\$642</b> and up (10 stores)</p> <p><input type="checkbox"/> Compare (Share and Compare)</p>
	<p><b>Black iPad 8gb</b></p> <p>The iPad 2 is the second and current generation of the iPad, a tablet computer designed, developed and marketed by Apple. It serves primarily as a platform for audio-visual media... <a href="#">more...</a></p>	<p><b>\$599</b> eCRATER</p> <p><input type="checkbox"/> Compare (Share and Compare)</p>

# Output of Entity Resolution

- Also referred to as record linkage and entity matching
- Identify records in a (or more) dataset(s) representing the same entity



# Entity Resolution Techniques

## Similarity-based

- Similarity Function  $Jaccard(r, s) = \left| \frac{r \cap s}{r \cup s} \right|$
- Threshold (e.g., 0.8)

$$Jaccard(r1, r2) = 0.9 \geq 0.8$$

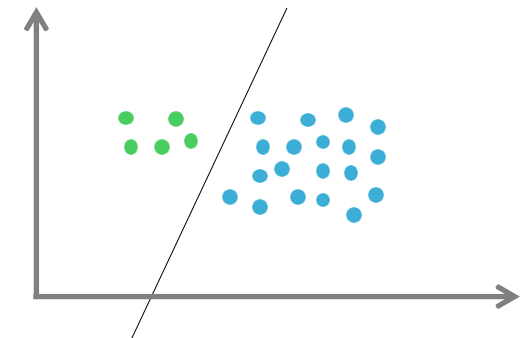
Matching

$$Jaccard(r4, r8) = 0.1 < 0.8$$

Non-matching

## Learning-based

- Represent a pair of records as a feature vector





# Similarity-based

---

Suppose the similarity function is Jaccard.

Problem Definition

Given a table  $T$  and a threshold  $\theta$ , the problem aims to find all record pairs  $(r, s) \in T \times T$  such that  $\text{Jaccard}(r, s) \geq \theta$

The naïve solution needs  $n^2$  comparisons

# Filtering-and-Verification

---

## Step 1. Filtering/blocking

- Removing obviously dissimilar pairs

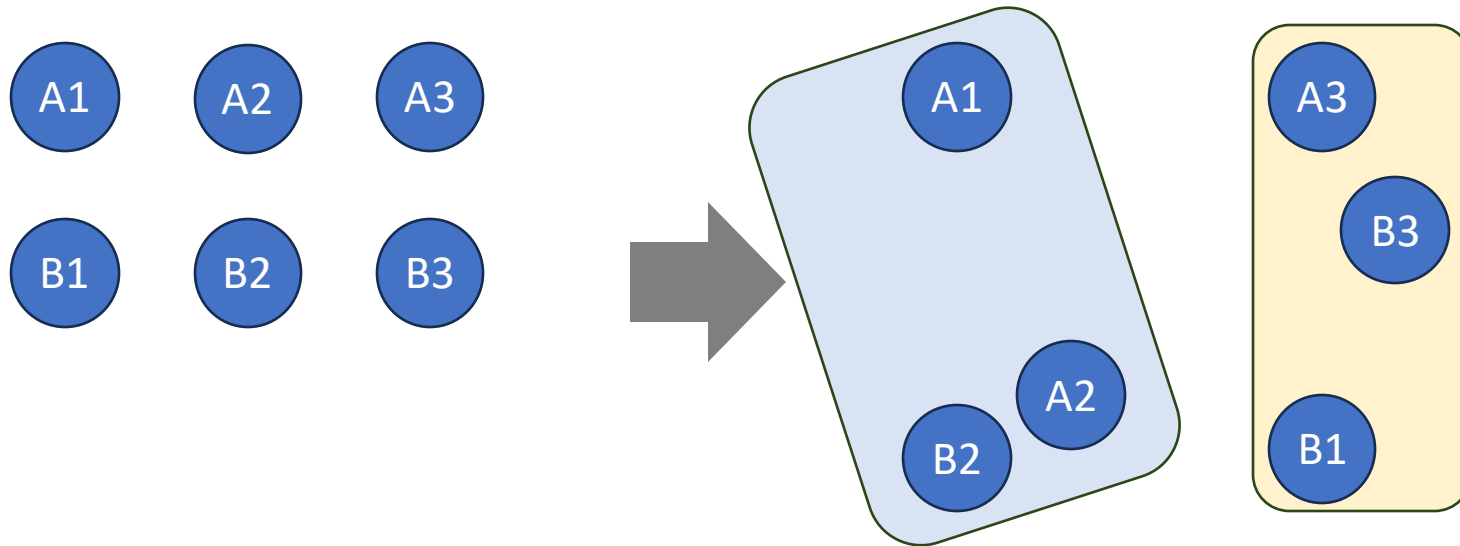
## Step 2. Verification/pairwise matching

- Computing Jaccard similarity only for the survived pairs

# Filtering/blocking

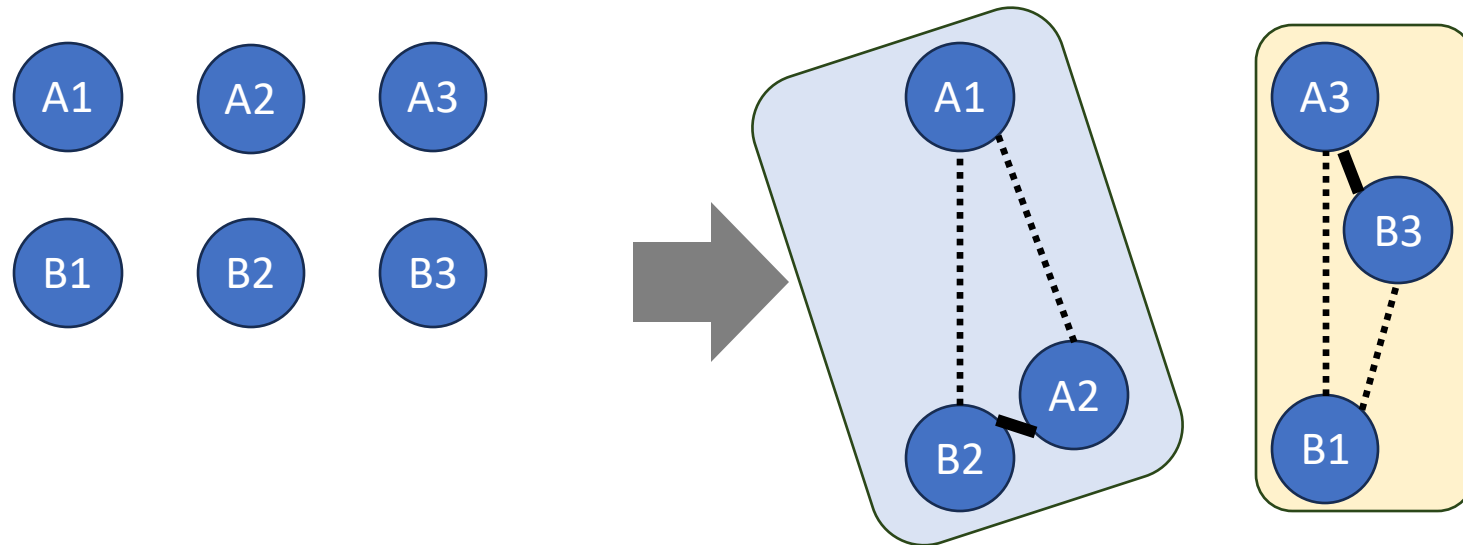
---

- Removing obviously dissimilar pairs



# Verification/Pairwise Matching

- Computing similarity only for the survived pairs



# How Does Filtering Work?

---

What are “obviously dissimilar pairs”?

- Two records are obviously dissimilar if they do not share any word.
- In this case, their Jaccard similarity is zero, thus they will not be returned as a result and can be safely filtered.

How can we efficiently return the record pairs that share at least one word?

- To help you understand the solution, let's first consider a simplified version of the problem, which assumes that each record only contains one word

# A simplified version

Suppose each record has only one word. Write an SQL query to do the filtering.

r <sub>1</sub>	Apple
r <sub>2</sub>	Apple
r <sub>3</sub>	Banana
r <sub>4</sub>	Orange
r <sub>5</sub>	Banana

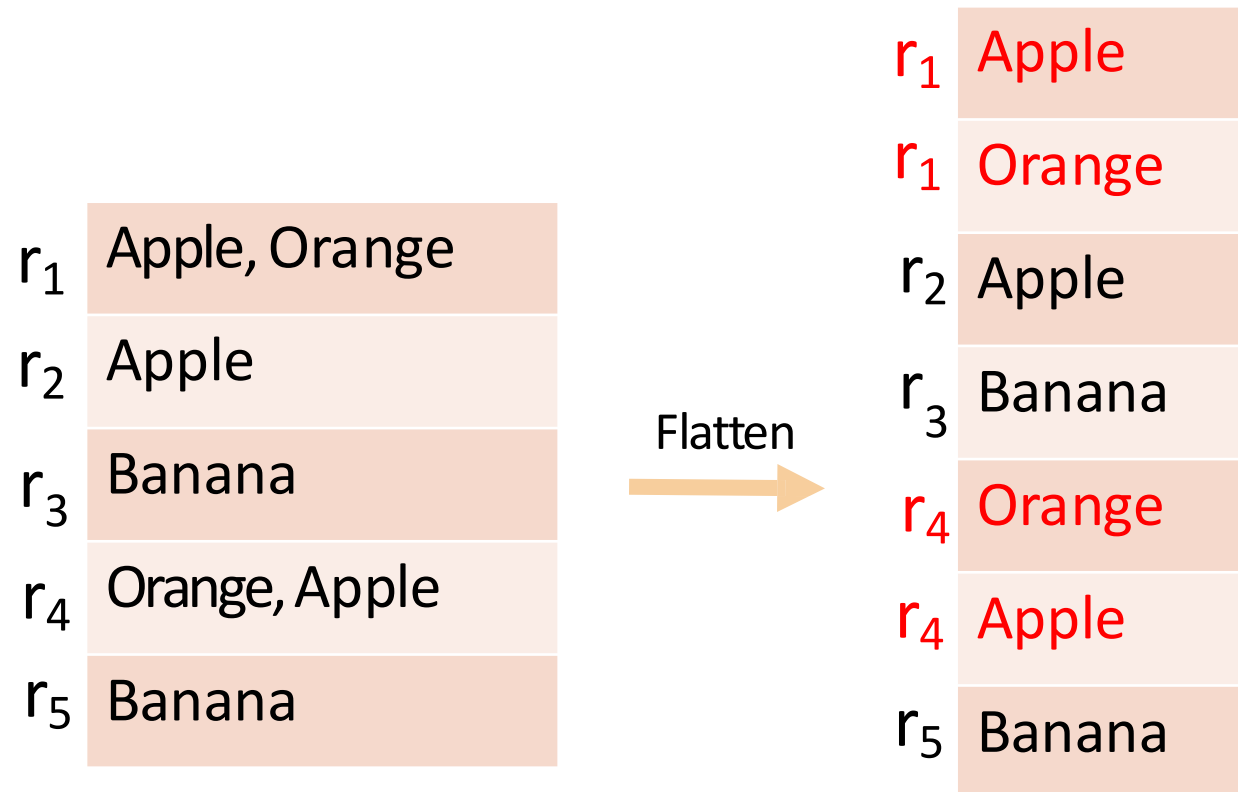
```
SELECT T1.id, T2.id
FROM Table T1, Table T2
WHERE T1.word = T2.word and T1.id < T2.id
```

Does it require  $n^2$  comparisons?

Output: (r<sub>1</sub>, r<sub>2</sub>), (r<sub>3</sub>, r<sub>5</sub>)

# A general case

Suppose each record can have multiple words.



1. This new table can be thought of as the inverted index of the old table.
2. Run the previous SQL on this new table and remove redundant pairs.

# Not satisfied with efficiency?

---

## Exploring stronger filter conditions

- Filter the record pairs that share **zero** token
- Filter the record pairs that share **one** token
- ....
- Filter the record pairs that share **k** tokens

## Challenges

- How to develop efficient filter algorithms for these stronger conditions?

Jiannan Wang, Guoliang Li, Jianhua Feng.

[Can We Beat The Prefix Filtering? An Adaptive Framework for Similarity Join and Search.](#) SIGMOD 2012:85-96.



# Not satisfied with result quality?

---

## TF-IDF

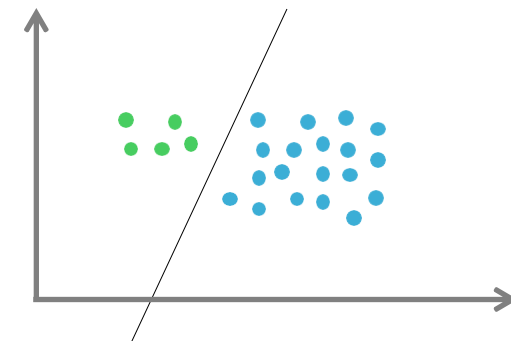
- Use weighted Jaccard:  $WJaccard(r, s) = \frac{wt(r \cap s)}{wt(r \cup s)}$

## Crowdsourcing

- Ask human to decide whether two records are matching or not

## Learning-based

- Model entity resolution as a classification problem



# Human-in-the-loop: Crowdsourcing

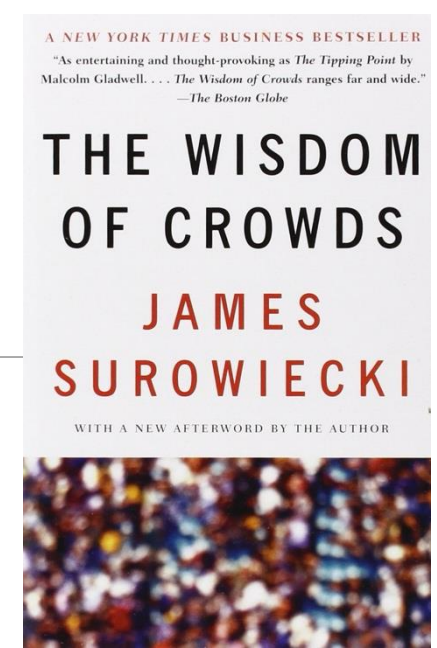
CMPT 884: Human-in-the-loop Data Management (SFU, Fall 2016)

<https://sfu-db.github.io/cmpt884-fall16/>

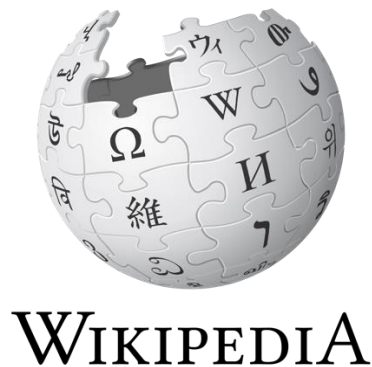
# The Wisdom of Crowds

What does it mean?

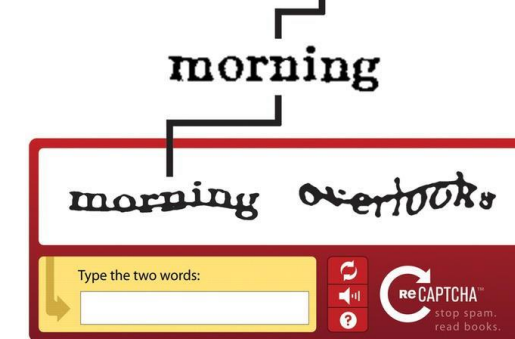
- Two heads are better than one



Some famous examples



The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.



# Amazon Mechanical Turk

500K+ workers\*



Get Started with Amazon Mechanical Turk



Create Tasks

Human intelligence through an API. Access a global, on-demand, 24/7 workforce.

Create a Requester account

or



Make Money

Make money in your spare time. Get paid for completing simple tasks.

Create a Worker account

amazonmechanical turk Artificial Intelligence Your Account HITs Qualifications 473,182 HITs available now

All HITs | HITs Available To You | HITs Assigned To You

Find HITs containing that pay at least \$ 0.00 for which you are require Master Qu

Timer: 00:00:00 of 2 minutes Want to work on this HIT? Accept HIT Total Earned: Unava Total HITs Submitted: 0

Identify if two receipts are the same  
Requester: Jon Brelig  
Qualifications Required: None  
Reward: \$0.01 per HIT  
HITs Available: 1  
Duration: 2 minu

person  
dog  
chair

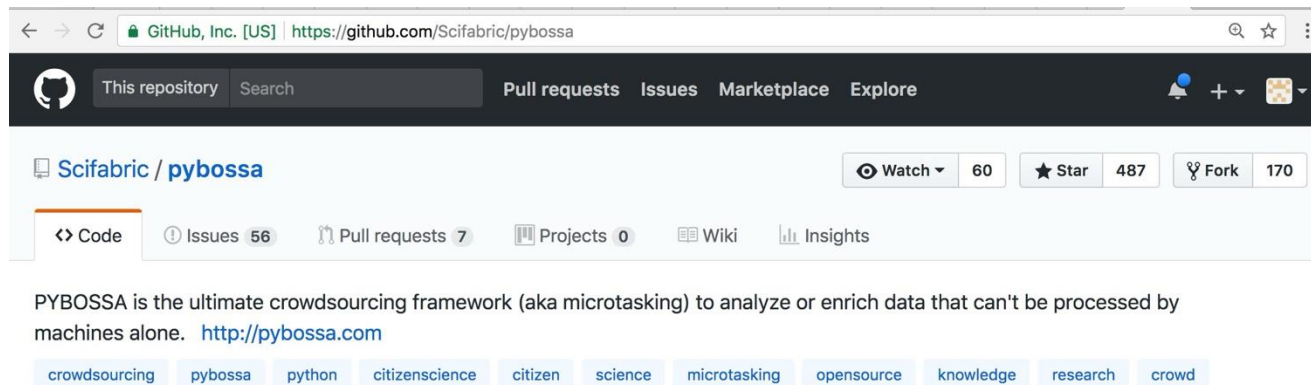
\* <https://requester.mturk.com/tour>

# Crowdsourcing may not work 😞

## What if your data is confidential?

- E.g., Medical Data, Customer Data

## Internal Crowdsourcing Platform



The screenshot shows a web browser window displaying the GitHub repository page for Scifabric/pybossa. The browser's address bar shows the URL <https://github.com/Scifabric/pybossa>. The repository page includes a search bar, navigation links for Pull requests, Issues, Marketplace, and Explore, and a notification bell. The repository name is Scifabric / pybossa, with 60 Watchers, 487 Stars, and 170 Forks. Below the repository name, there are tabs for Code, Issues (56), Pull requests (7), Projects (0), Wiki, and Insights. The description states: "PYBOSSA is the ultimate crowdsourcing framework (aka microtasking) to analyze or enrich data that can't be processed by machines alone. <http://pybossa.com>". At the bottom, there are several topic tags: crowdsourcing, pybossa, python, citzenscience, citizen, science, microtasking, opensource, knowledge, research, and crowd.

# Crowdsourcing may not work 😞

What if your data is so big?

- E.g., Label **10 million** images

Crowdsourcing may not work 😞

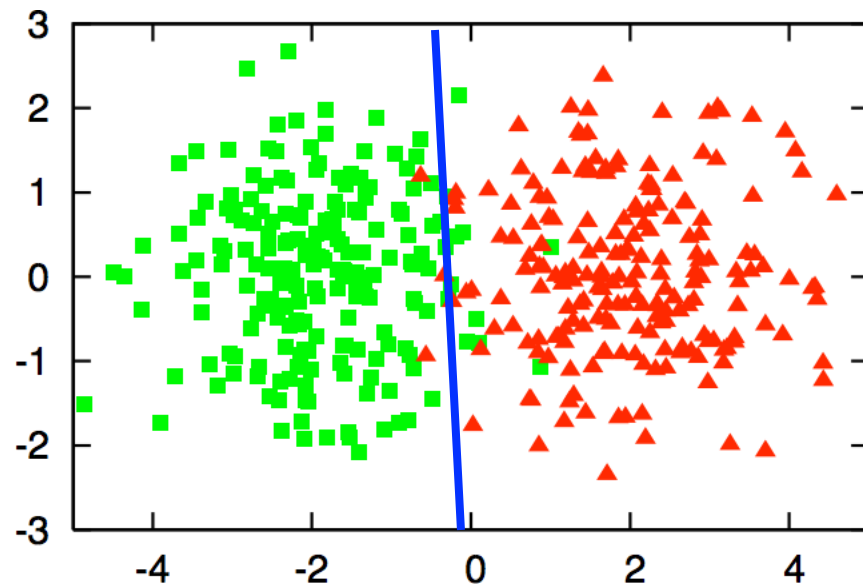
What if your data is so big?

- E.g., Label 10 million images

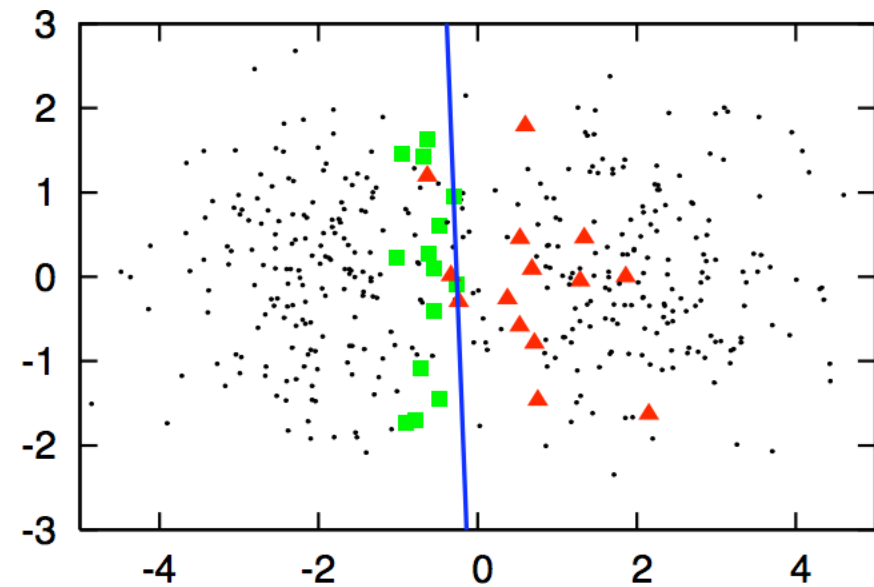
# Human-in-the-loop: Active Learning

# Active Learning

## Supervised Learning



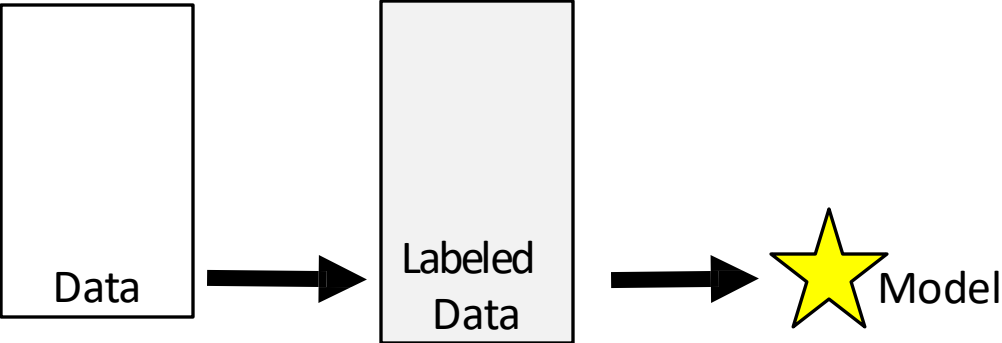
## Active Learning



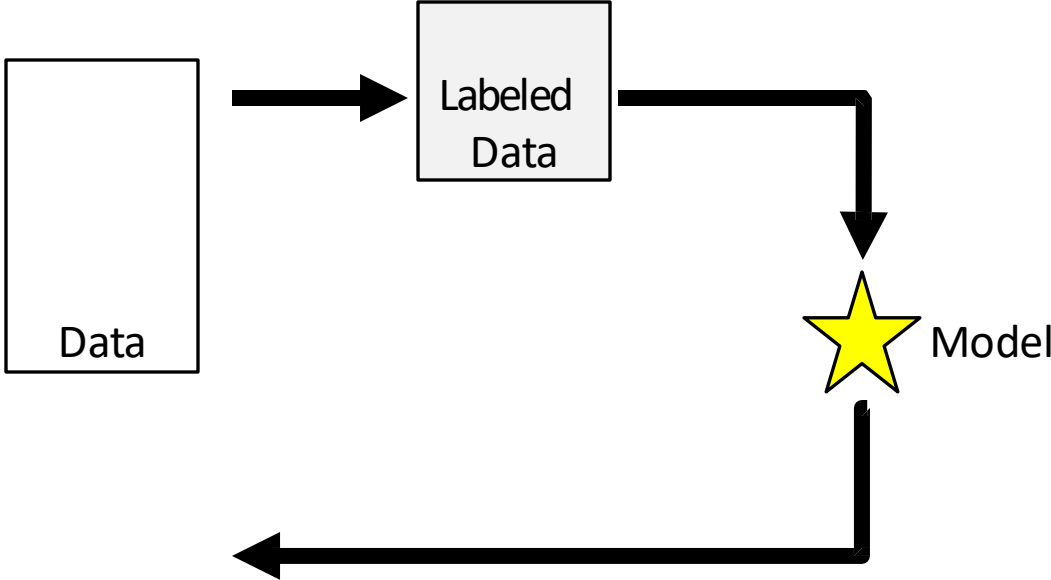


# Workflow

## Supervised Learning



## Active Learning



# Query Strategy

---

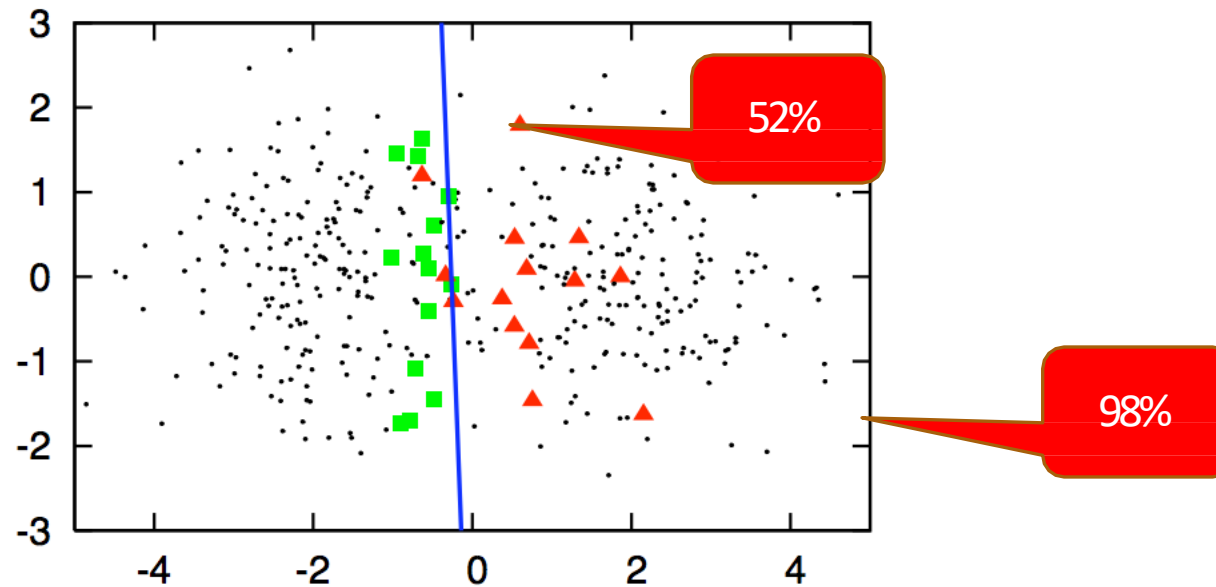
- Which data points should be labeled?

- Uncertain Sampling
- Query-By-Committee
- Expected Error Reduction
- Expected Model Change
- Variance Reduction
- Density-Weighted Methods

- Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison 52.55-66 (2010): 11.

# Uncertain Sampling

Pick up most uncertain datapoints to label



Logistic Regression

- `predict_proba(X)`

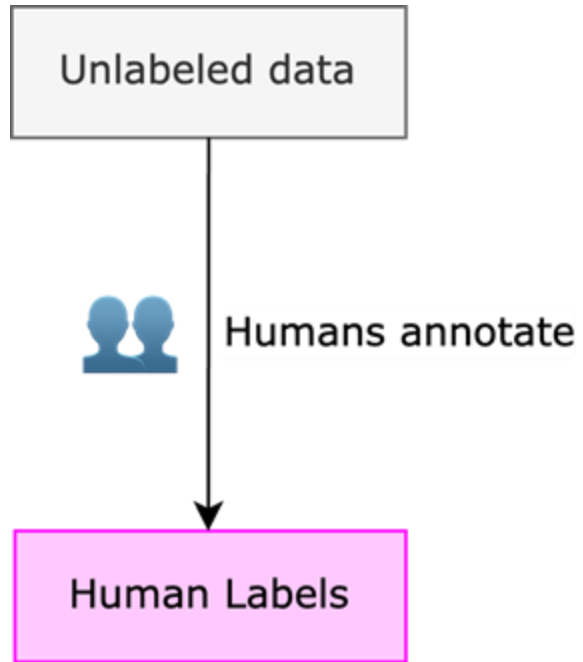
Crowdsourcing may not work 😞

What if your data is so big?

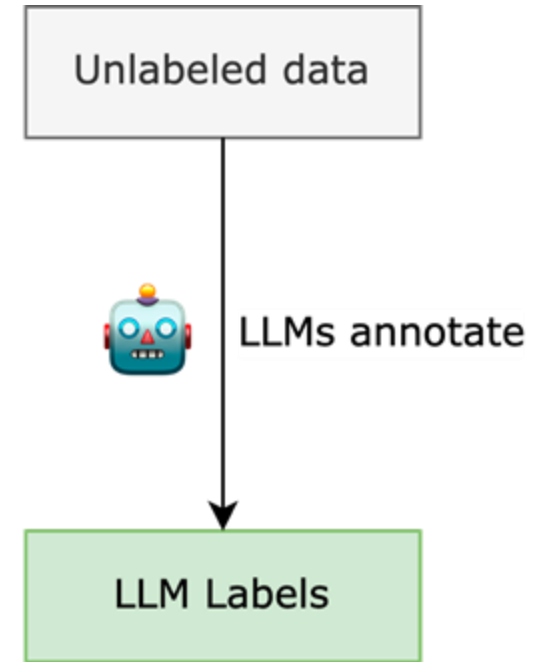
- E.g., Label 10 million images

**Human-in-the-loop: fully  
automated with LLM?**

# Perhaps not there yet

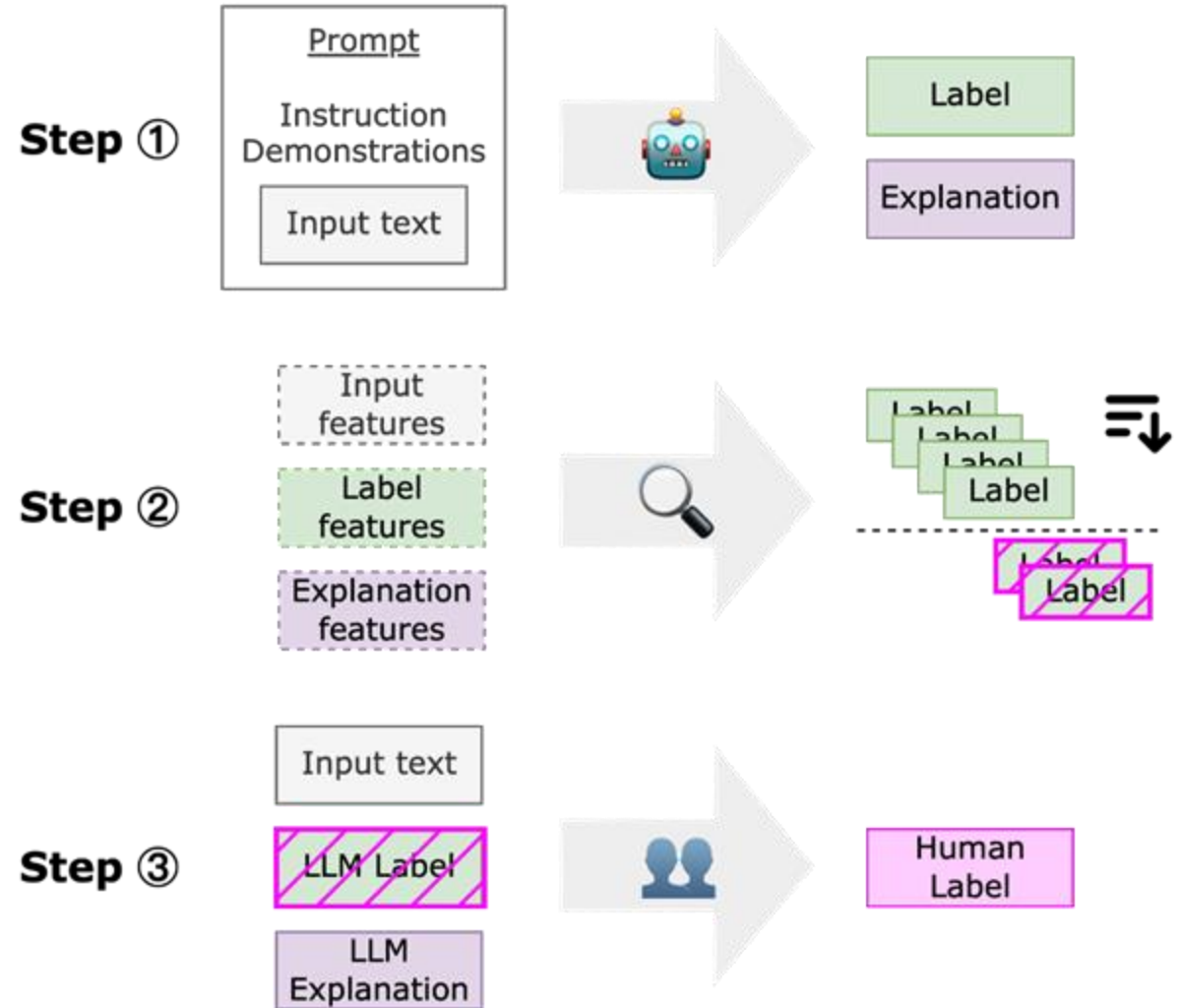
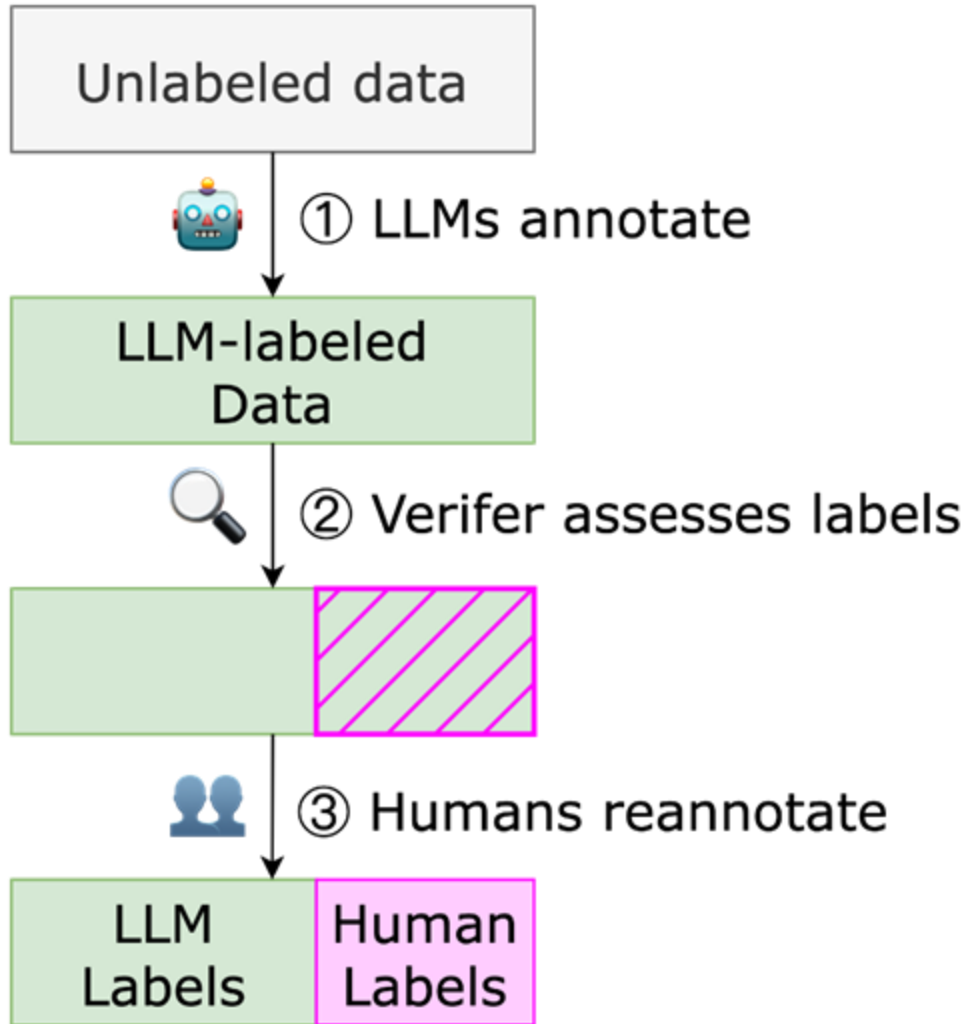


- ✗ time consuming
- ✗ not scalable
- ✗ costly



- ✓ good zero-shot and few-shot learners for many NLP tasks
- ✓ much cheaper
- ✗ performance varies across tasks, labels, and instances

# Human-LLM Collaborative Annotation Framework



Wang et al. "Human-LLM collaborative annotation through effective verification of LLM labels." CHI 2024

# Summary

## Preppin' Data

A weekly challenge to help you learn to prepare data and use Tableau Prep

<https://preppindata.blogspot.com/>

---

### Data Collection

- Where to collect, How to Collect

### Data Cleaning

- Dirty Data Problems, Data-cleaning tools

### Data Integration

- Schema Mapping, Entity Resolution, Data Fusion

### Entity Resolution

- Similarity-based, Crowdsourcing, Active Learning, LLM