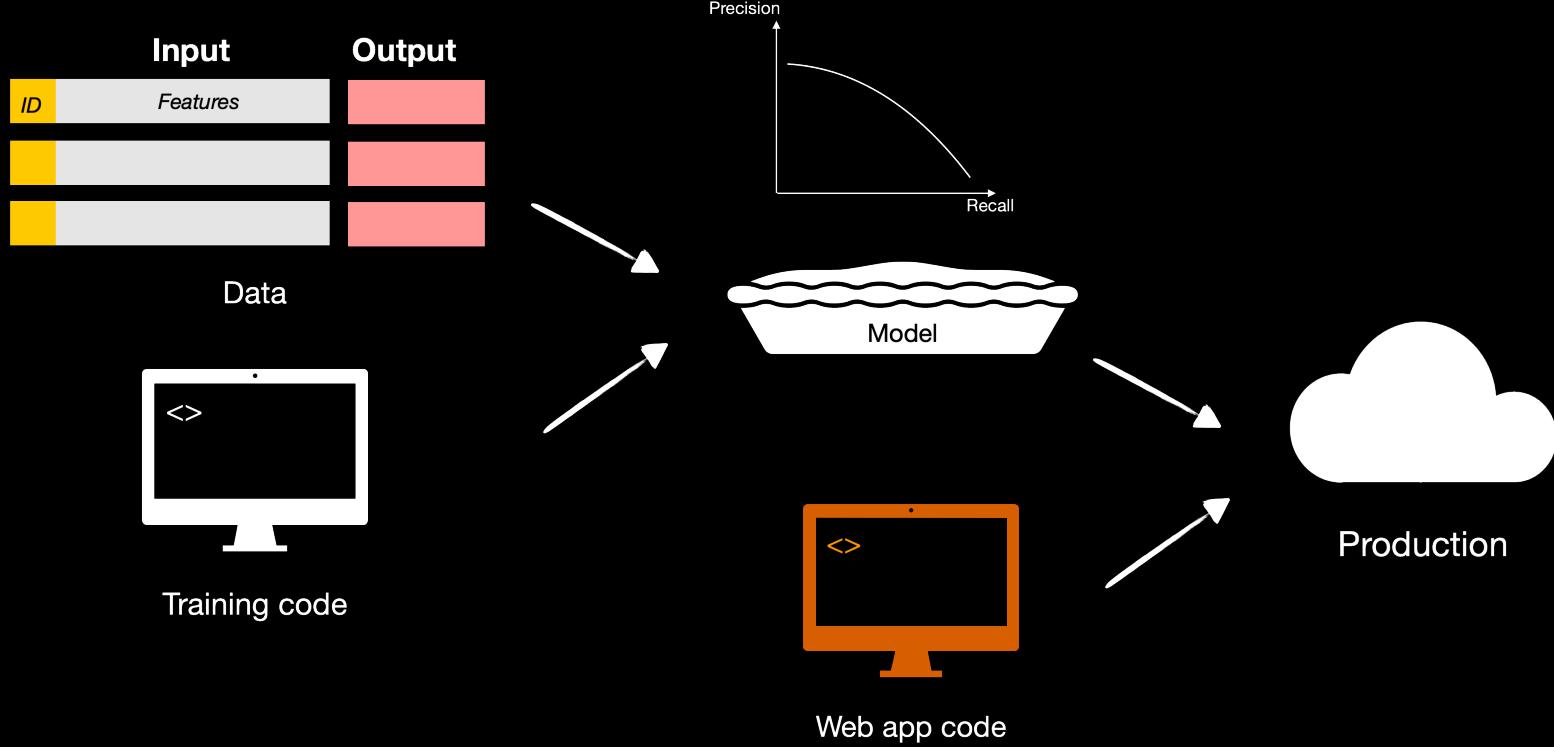


# Data Acquisition and Management (for ML)

Jin Guo  
SOCS McGill University



What's missing?

The background features a surreal landscape with floating spheres of various sizes and colors (yellow, blue, black) against a dark, textured ground. In the center-right, a white sheep stands inside a transparent sphere resting on a green patch of grass. The surrounding environment is composed of large, soft-edged, curved shapes in shades of brown, yellow, and blue.

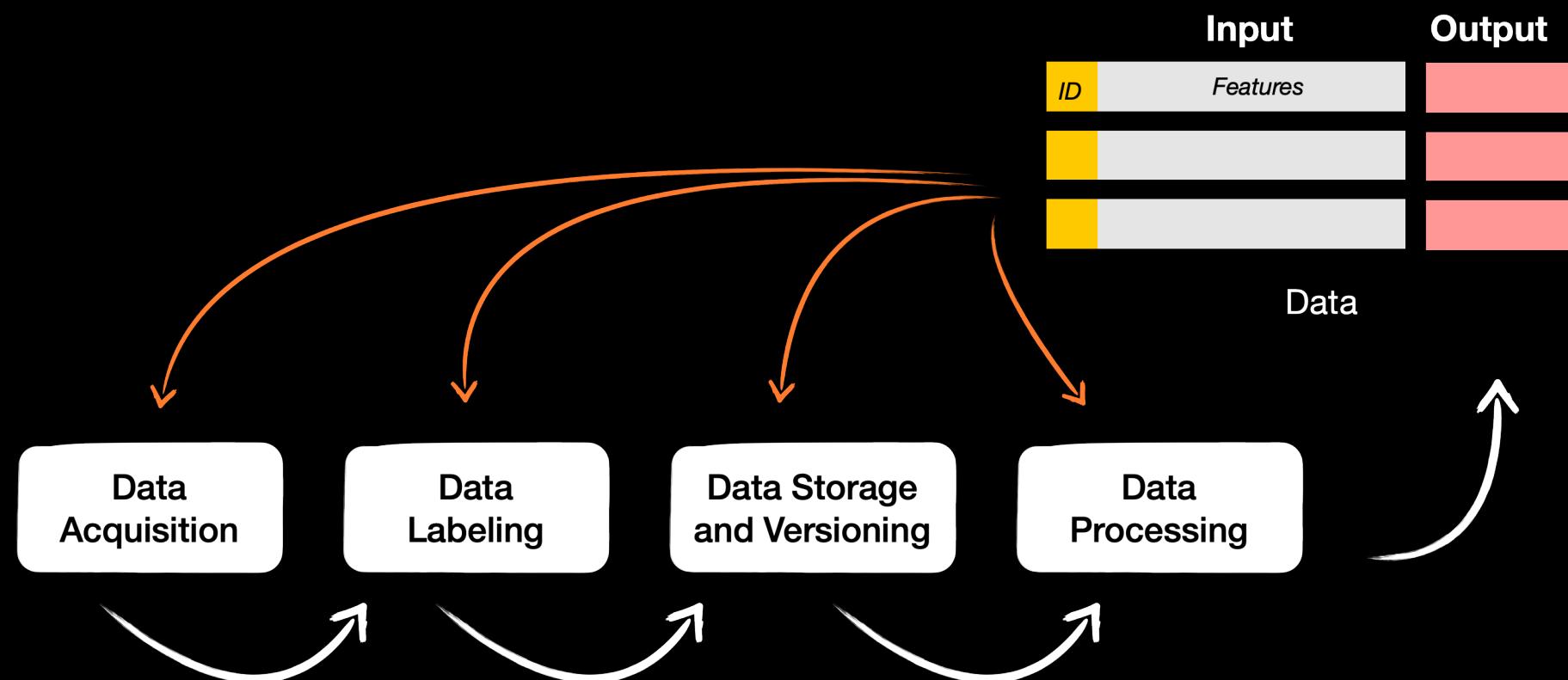
What do we know about the training data?

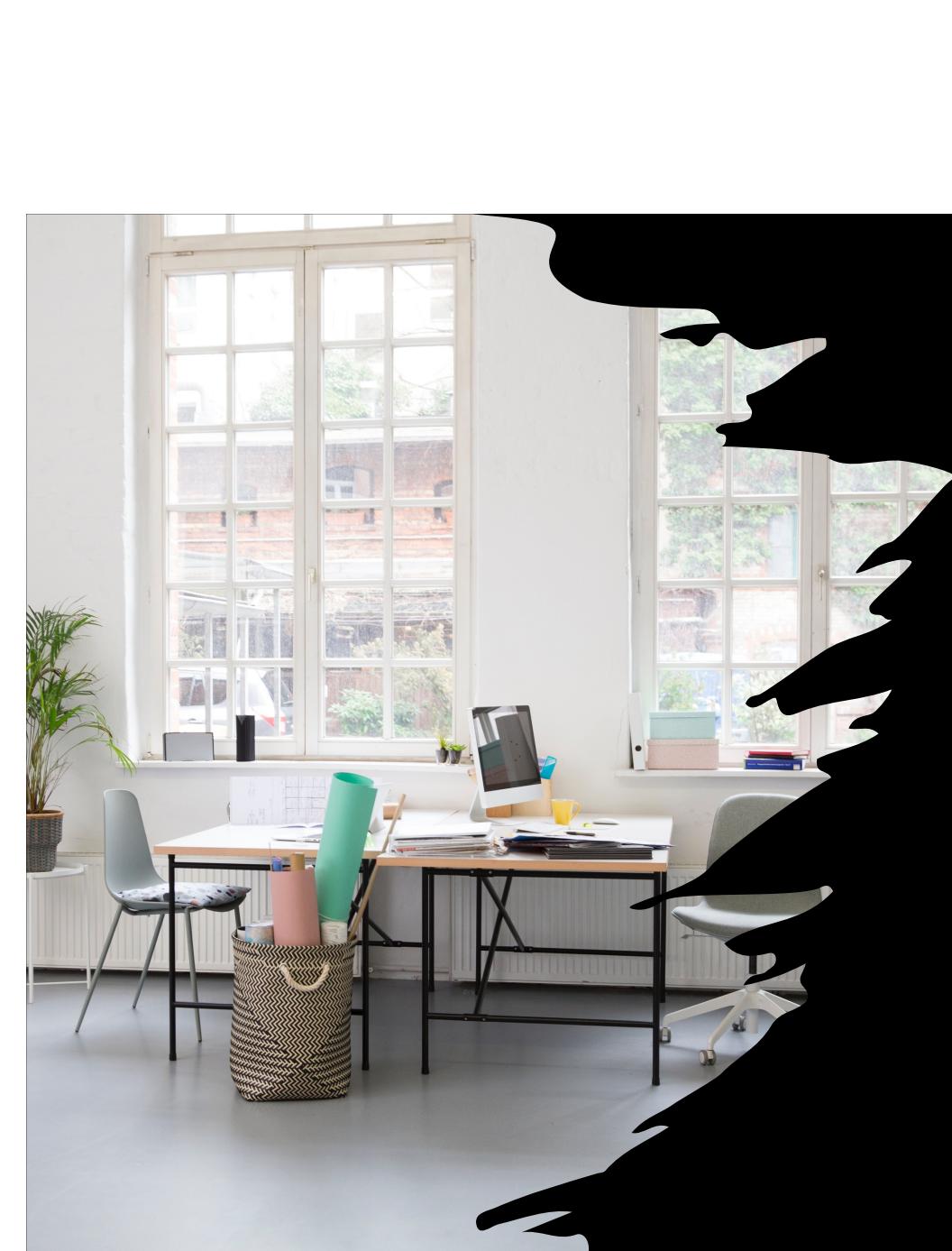
# Assumptions about Data

- Data sources
- Meaning
- Data type
- Data quality



# Data Pipeline

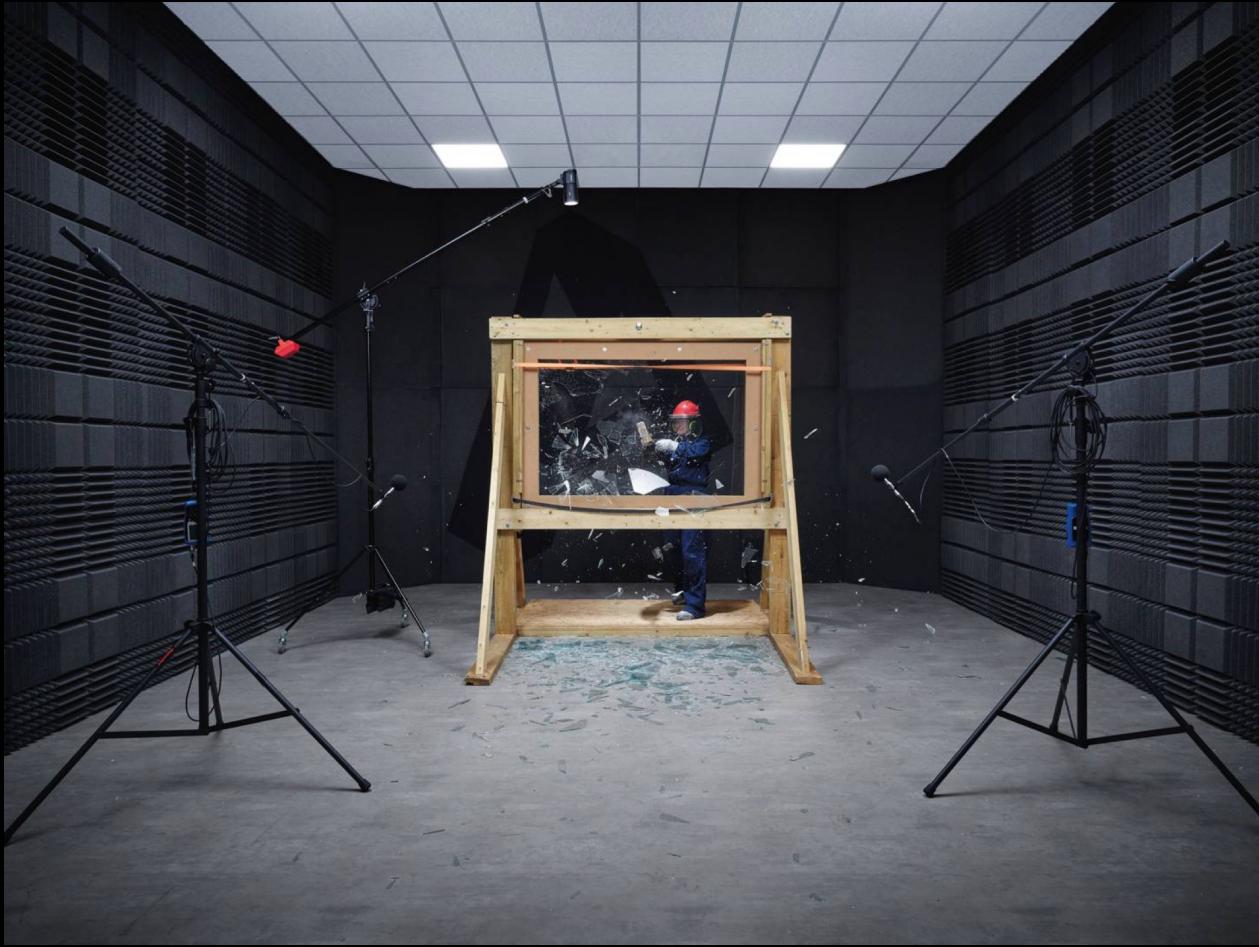




## Activity 1:

Consider if you plan to build a sound recognition (recognize family members, doorbells, fire alarms, etc.) application for the smart home environment.

- What kind of data would you collect?
- How do you plan to collect such data?
- What kind of problems you might encounter?



Credit: Greg White

Source: <https://www.wired.co.uk/article/audio-analytics-sound-map>

# Data Acquisition

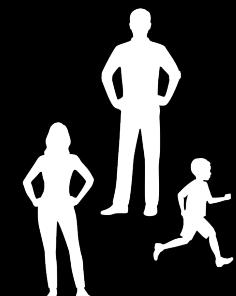
*Other challenges for each strategy?*



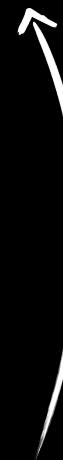
Public Available Data



Cost (Money and Time)



"Real" Data



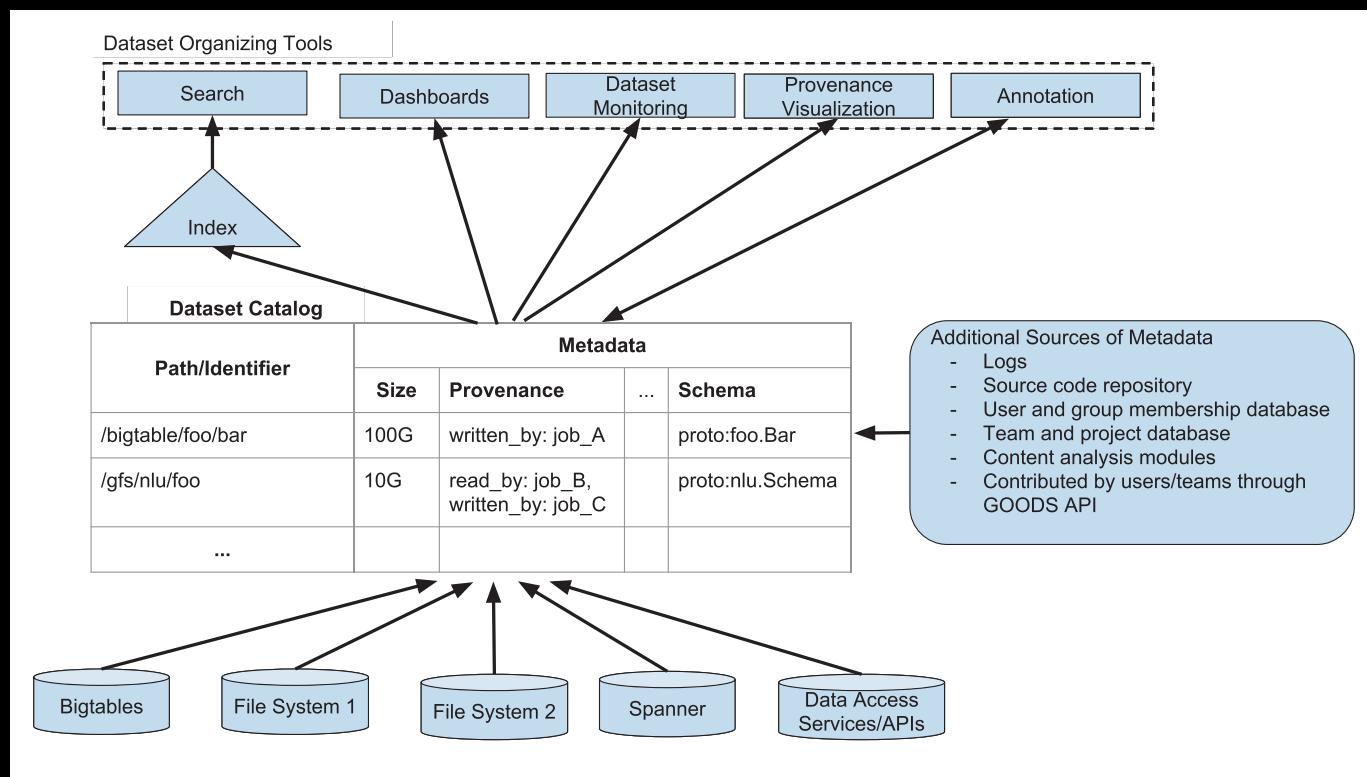
Relatedness

# Type of Data Sources

- User Input Data
- System-Generated Data
  - Logs*
  - Telemetry*

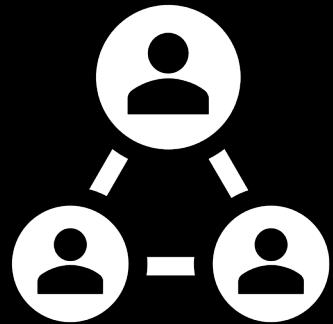
# Search within Enterprise Data

Example:  
Google Dataset  
Search (Goods)



Halevy, Alon, et al. "Goods: Organizing google's datasets." Proceedings of the 2016 International Conference on Management of Data. 2016.

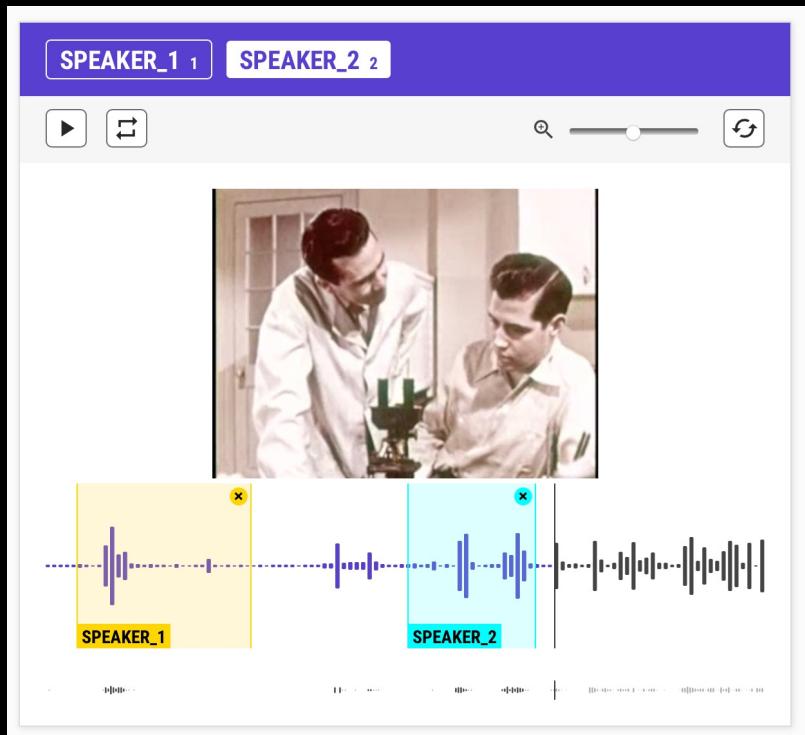
# Data Labeling



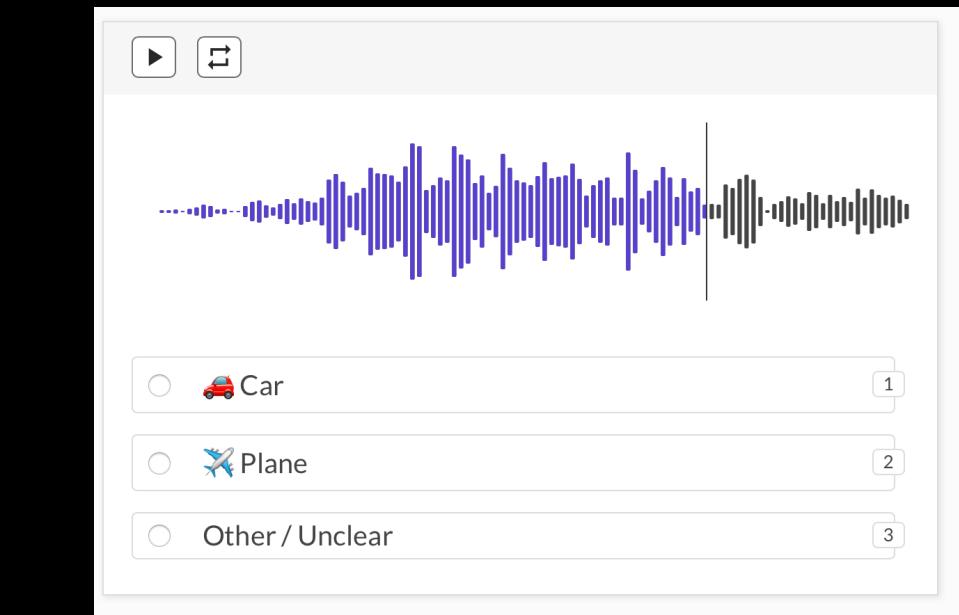
- Recruit Annotator
- Outsourcing
- Crowdsourcing

For the audio you have collected in Activity 1, what kind of information do you plan to provide to the annotators to ensure of quality of the labeling?

# Data Labeling



# prodigy



# Data Labeling



**Vibration**

Hierarchy      > Channel, environment and background > Noise > Vibration Try another category

Description	The general class of sounds caused by a periodic mechanical oscillation.
URI	<a href="http://en.wikipedia.org/wiki/Vibration">http://en.wikipedia.org/wiki/Vibration</a>
Examples	Three small spectrograms showing different types of vibration patterns. Each spectrogram has a play button icon and a timestamp: 0:00/0:04, 0:00/0:03, and 0:25/0:25.

Response type	Meaning
Present and predominant	The type of sound described is <i>clearly present</i> and <i>predominant</i> . This means there are no other types of sound, with the exception of low/mild background noise.
Present but not predominant	The type of sound described is <i>present</i> , but the audio clip also <i>contains other salient types of sound and/or strong background noise</i> .
Not present	The type of sound described is <i>not present</i> in the audio clip.
Unsure	<i>I am not sure</i> whether the type of sound described is present or not.

# Data Labeling

- Natural Labels

*E.g., recommender systems*

Google Map Route Planning

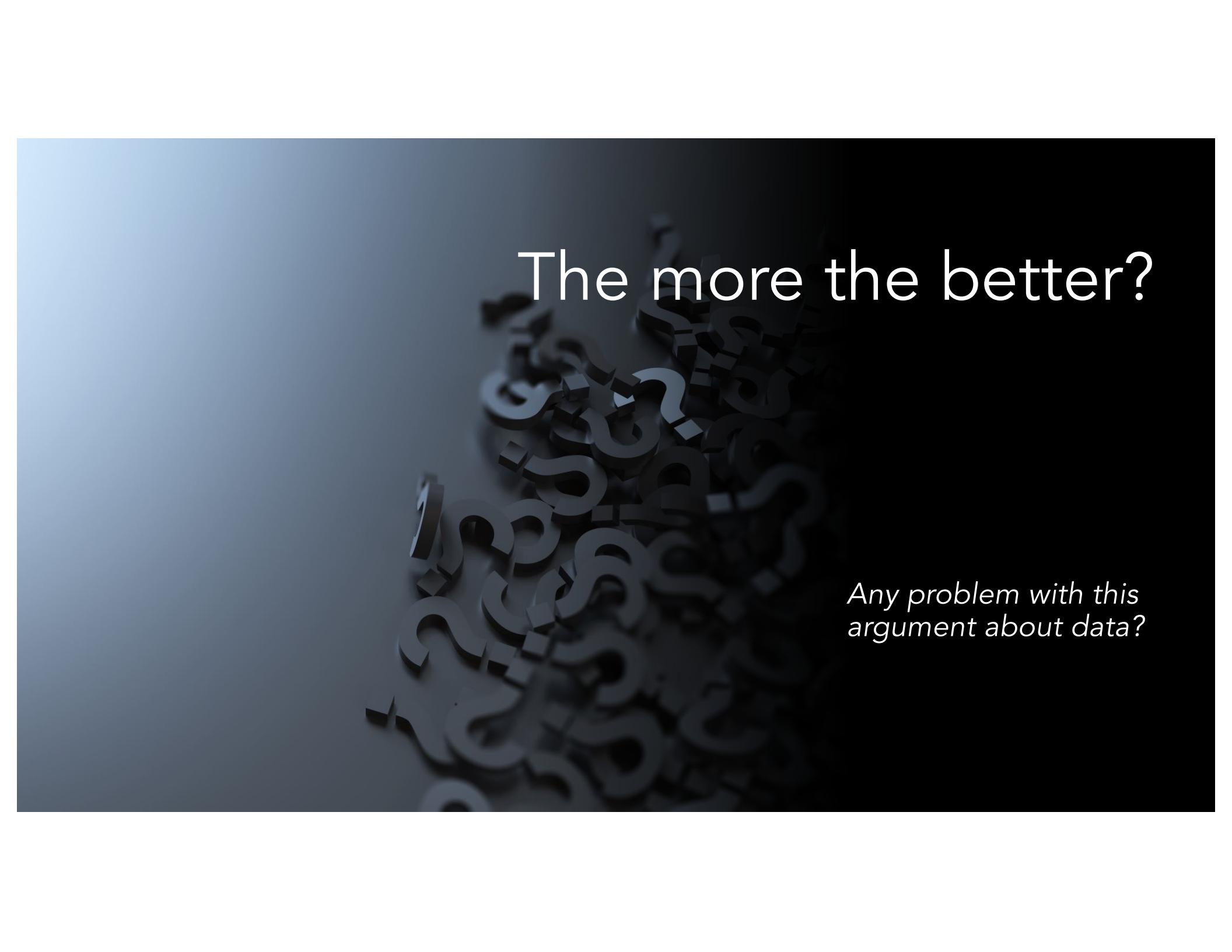
Newsfeed Recommendation

Tweeter Content Ranking

# Data Augmentation

- Speech/Video
  - Change speed, add pause, etc.
- Vision
  - Make perturbation on original image
- Text

Testing VS Training?



# The more the better?

*Any problem with this argument about data?*

# Predicting the risk of child abuse?

Use features: receiving county health or mental health treatment; being reported for drug or alcohol abuse; accessing supplemental nutrition assistance program benefits, cash welfare assistance, or Supplemental Security Income, etc.

## A Child Abuse Prediction Model Fails Poor Families

Why Pittsburgh's predictive analytics misdiagnoses child maltreatment and prescribes the wrong solutions

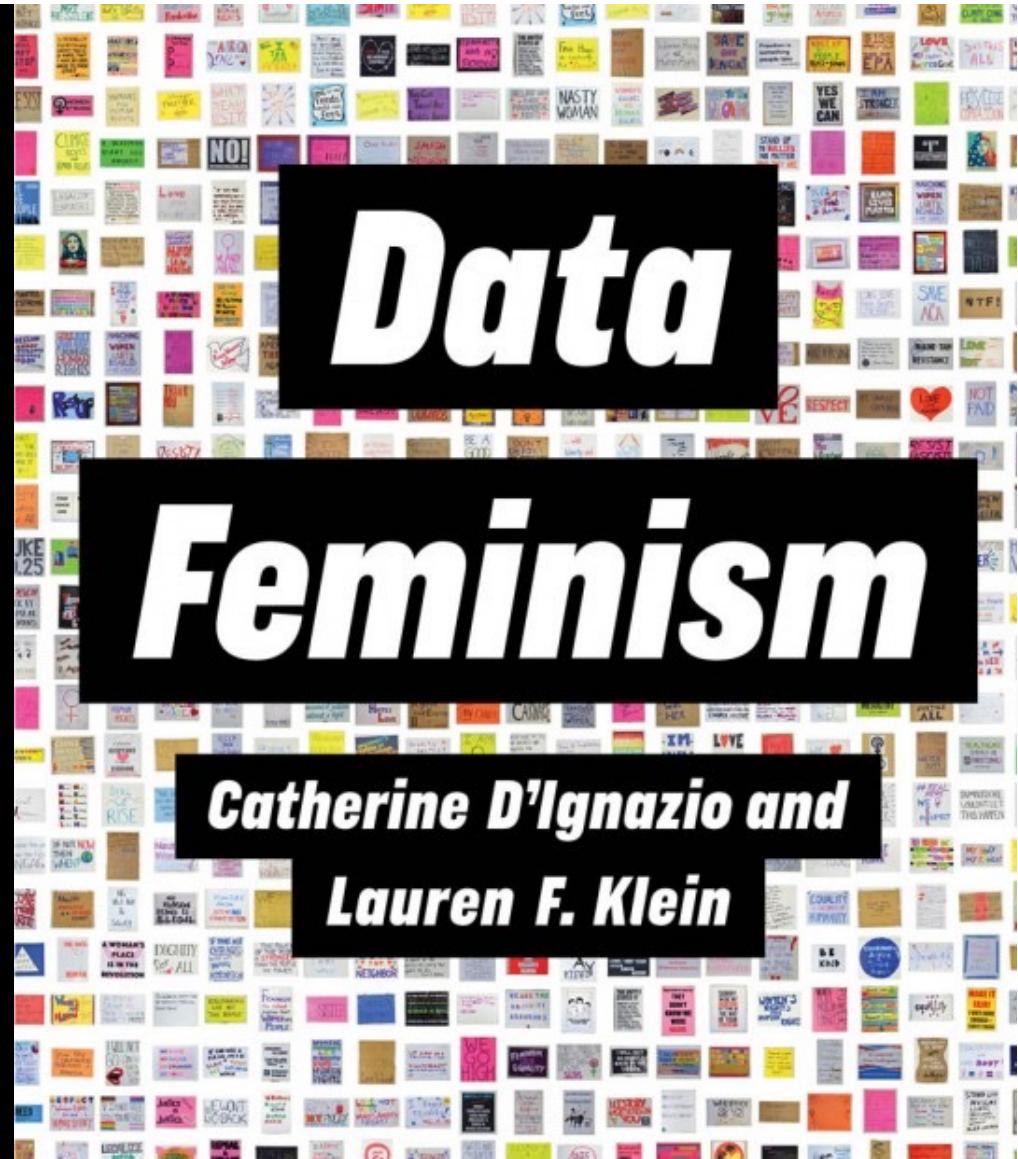
*"Allegheny County has an extraordinary amount of information about the use of public programs. But the county has no access to data about people who do not use public services. Parents accessing private drug treatment, mental health counseling, or financial support are not represented in DHS data. "*

The problem of oversampling the poor

# Data Sampling Types

- Nonprobability Sampling
  - Random Sampling
  - Stratified Sampling
  - ...
- What constitutes your population?*

- Who is doing the work of data science (and who is not)?
- Whose goals are prioritized in data science (and whose are not)?
- And who benefits from data science (and who is either overlooked or actively harmed)?



# Data Storage

- Data Models
  - Relational databases
  - SQL
  - Not versioned

```
dvdrental=# select title, release_year, length, replacement_cost from film
dvdrental=#   where length > 120 and replacement_cost > 29.50
dvdrental=#   order by title desc;
      title      | release_year | length | replacement_cost
-----+-----+-----+-----+
    West Lion      |      2006 |     159 |        29.99
  Virgin Daisy      |      2006 |     179 |        29.99
 Uncut Suicides      |      2006 |     172 |        29.99
   Tracy Cider      |      2006 |     142 |        29.99
  Song Hedwig      |      2006 |     165 |        29.99
Slacker Liaisons      |      2006 |     179 |        29.99
  Sassy Packer      |      2006 |     154 |        29.99
 River Outlaw      |      2006 |     149 |        29.99
 Right Cranes      |      2006 |     153 |        29.99
 Quest Mussolini      |      2006 |     177 |        29.99
 Poseidon Forever      |      2006 |     159 |        29.99
 Loathing Legally      |      2006 |     140 |        29.99
 Lawless Vision      |      2006 |     181 |        29.99
 Jingle Sagebrush      |      2006 |     124 |        29.99
 Jericho Mulan      |      2006 |     171 |        29.99
 Japanese Run      |      2006 |     135 |        29.99
 Gilmore Boiled      |      2006 |     163 |        29.99
 Floats Garden      |      2006 |     145 |        29.99
 Fantasia Park      |      2006 |     131 |        29.99
Extraordinary Conquerer      |      2006 |     122 |        29.99
 Everyone Craft      |      2006 |     163 |        29.99
  Dirty Ace      |      2006 |     147 |        29.99
 Clyde Theory      |      2006 |     139 |        29.99
Clockwork Paradise      |      2006 |     143 |        29.99
Ballroom Mockingbird      |      2006 |     173 |        29.99
(25 rows)
```

By BernardoSulzbach - Own work, CC BY-SA 4.0,  
<https://commons.wikimedia.org/w/index.php?curid=73867425>

# Data Storage

- Data Models
  - Relational databases
  - SQL
  - Not versioned

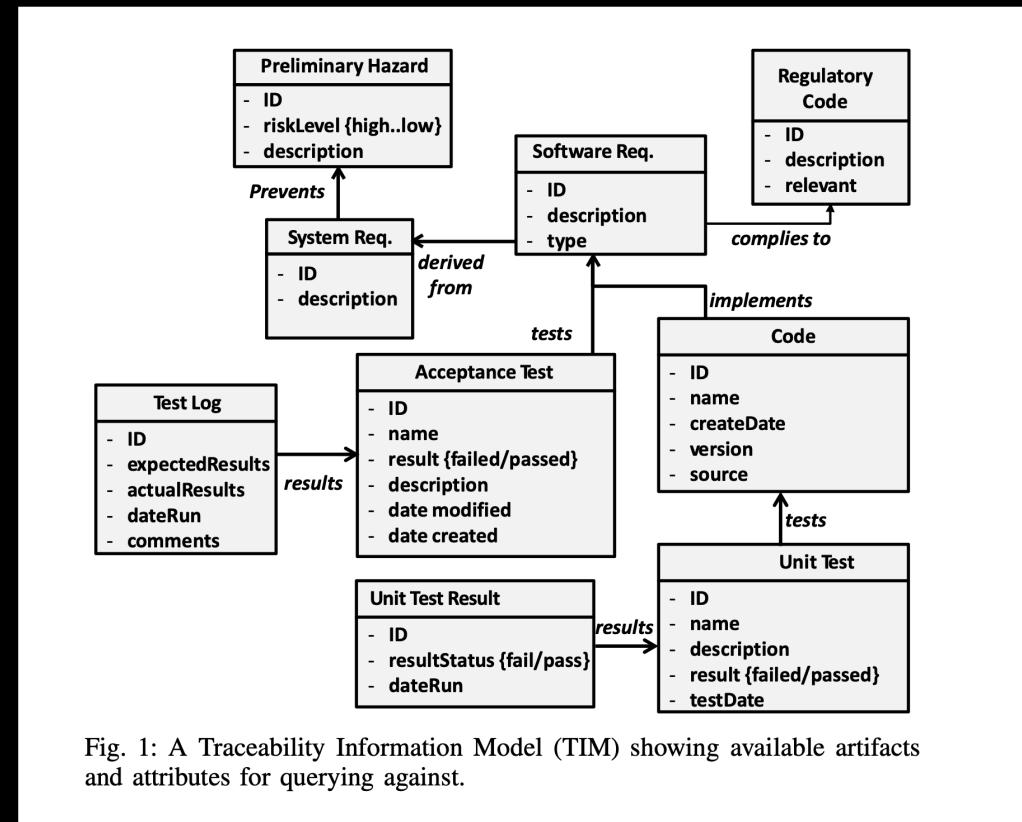
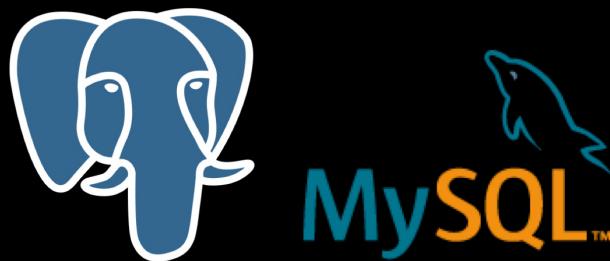


Fig. 1: A Traceability Information Model (TIM) showing available artifacts and attributes for querying against.

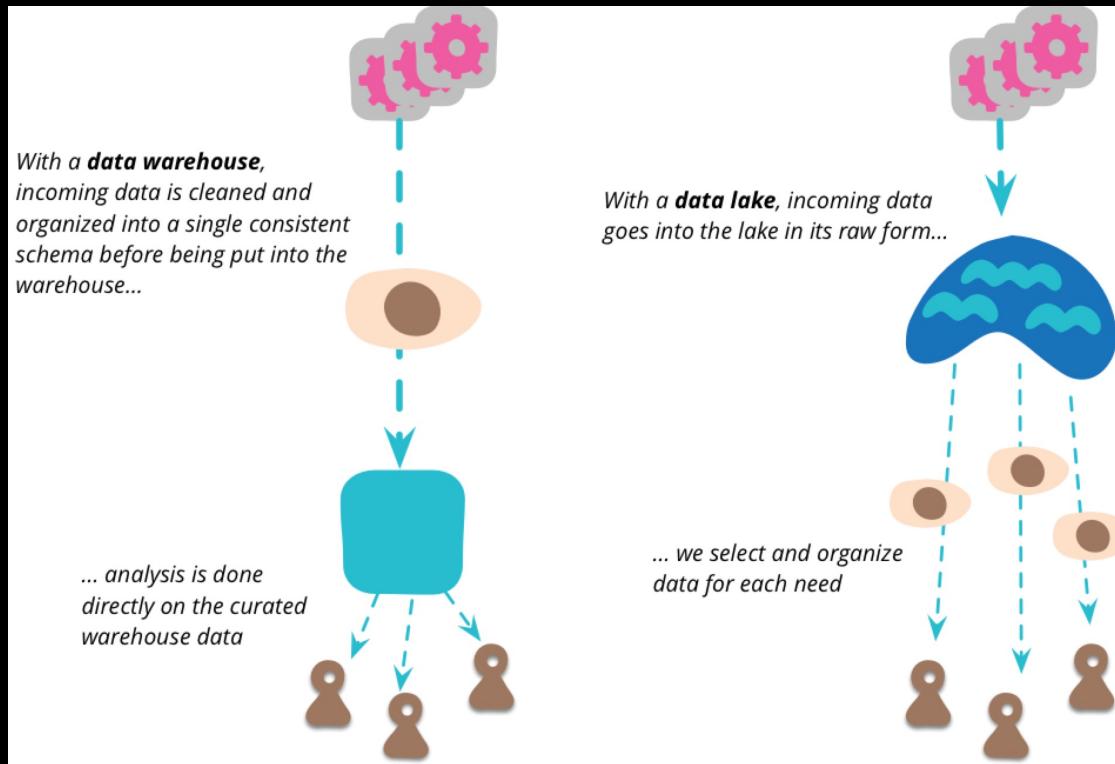
Lin, Jinfeng, et al. "Tiqi: A natural language interface for querying software project data." 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2017.

# Data Storage



- Data Models
  - NoSQL (not only SQL)
  - JSON documents, Key-value: key-value pairs, Wide-column: tables with rows and dynamic columns, Graph: nodes and edges
  - Faster Query, flexible data models

# Data Lake VS Data Warehouse



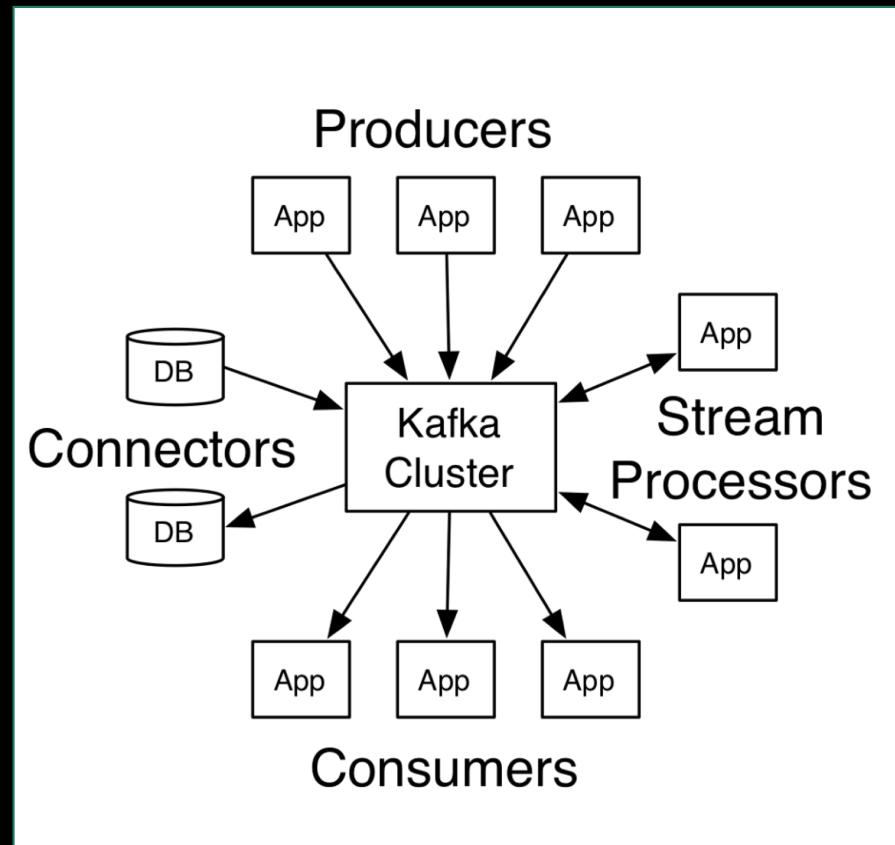
By Martin Fowler: <https://martinfowler.com/bliki/DataLake.html#footnote-mart-v-warehouse>

# Dataflow modes

- Databases
- Services using requests such as the requests, such as REST APIs.  
*Request Driven*
- Through a real-time transport like Apache Kafka  
*Event Driven*

# Event Streaming

## Apache Kafka



On Wednesdays:

Software Requirement