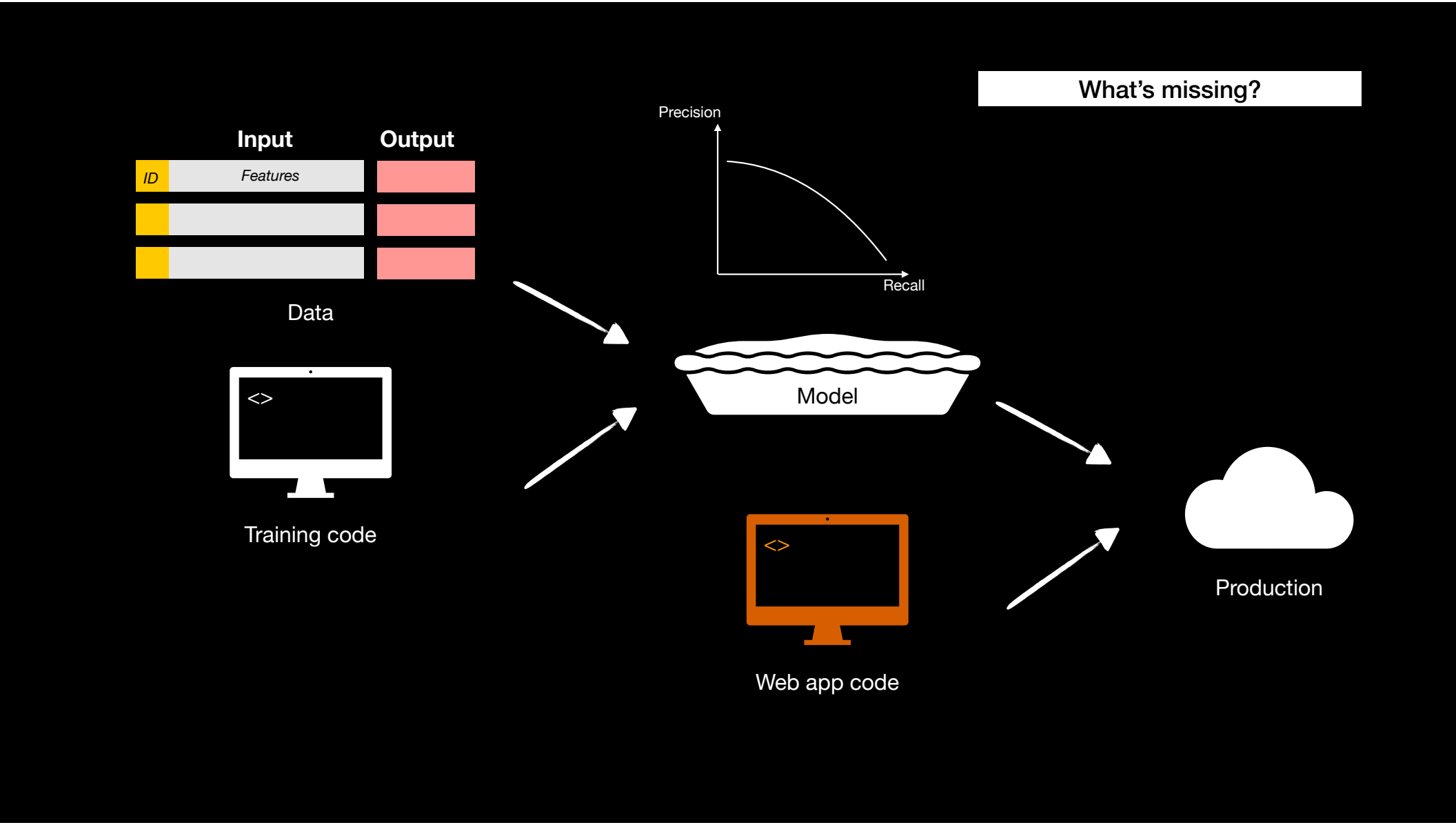
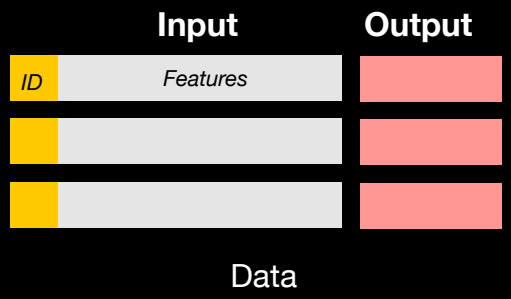
Data Acquisition and Management

Jin Guo Sept 29, 2020

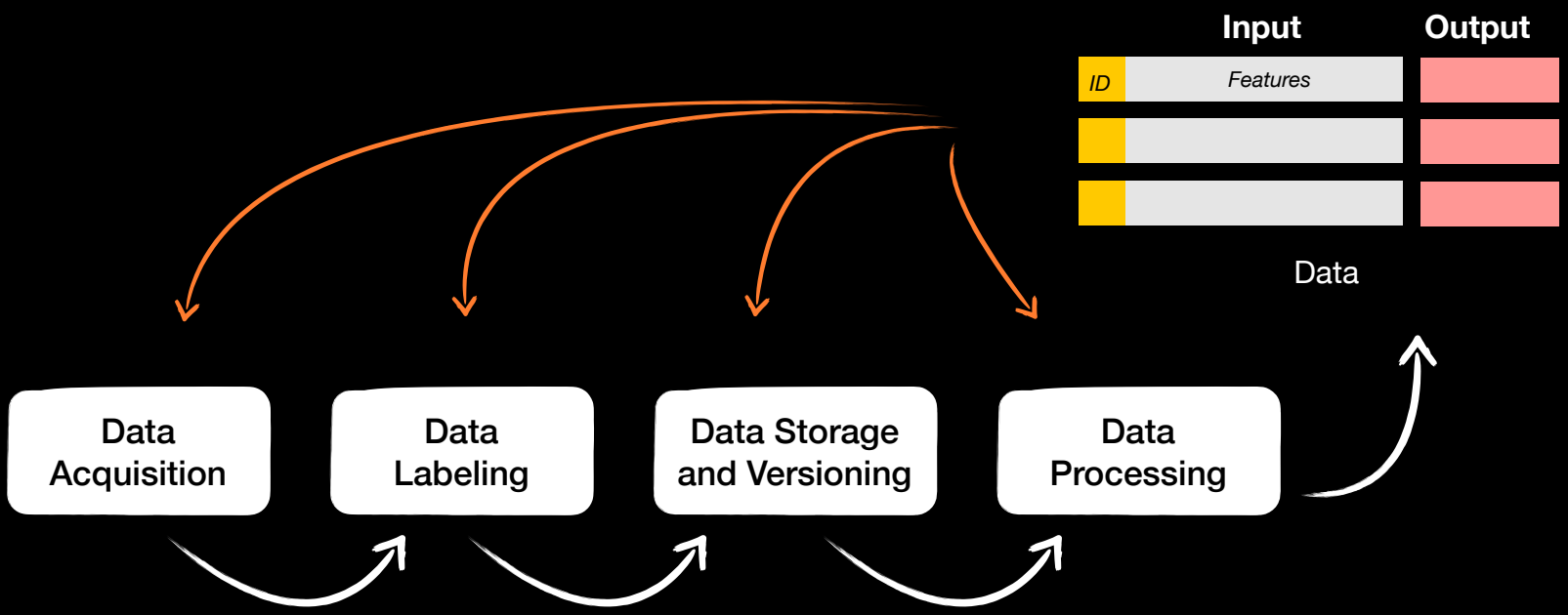

What's missing?
Input
Output
ID
Features
Data
Precision
Recall
Model
<>
Training code
<>
Web app code
Production

Assumptions about Data

- Data sources
- Meaning
- Data type
- Data quality



Data Pipeline



Data Engineers

Data Acquisition

Other challenges for each strategy?

Cost (Money and Time)

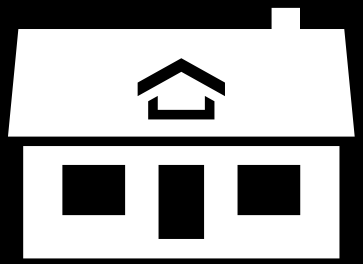
Public Available Data

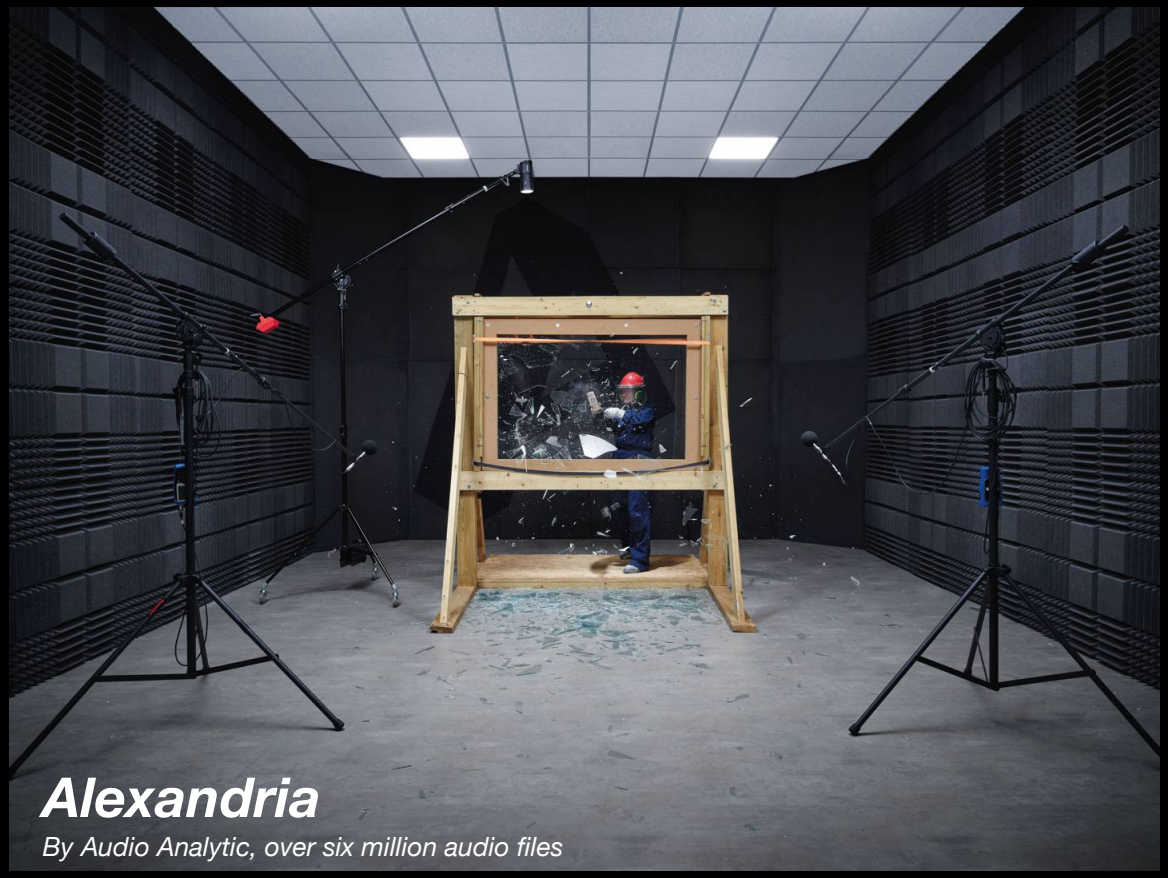
“Real” Data

Relatedness

Activity 1:

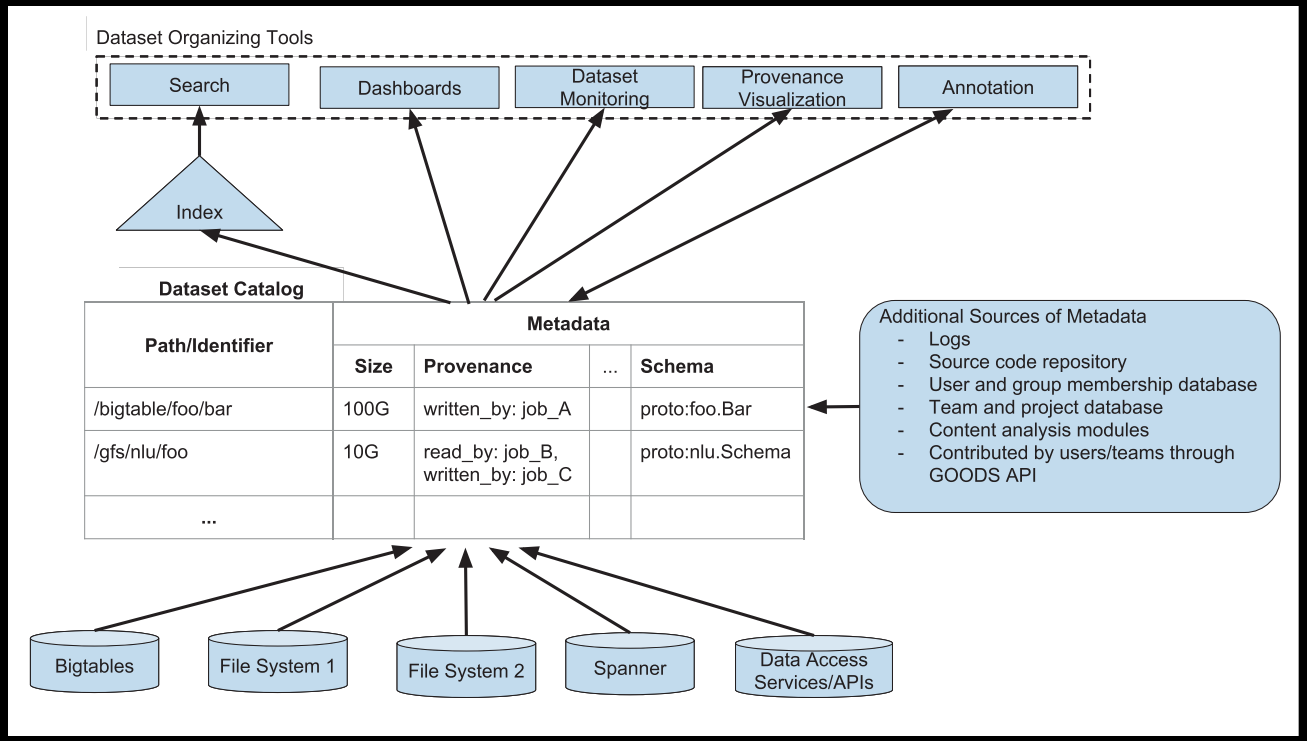
- Consider if you plan to build a **sound recognition** (recognize family members, doorbells, fire alarms, etc.) application for the smart home environment.
  - What kind of data would you collect?
  - How do you plan to collect such data?
  - What kind of problems you might encounter?





Credit: Greg White
Source: https://www.wired.co.uk/article/audio-analytics-sound-map

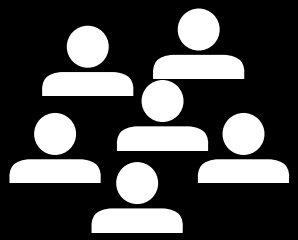
Search within Enterprise Data



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

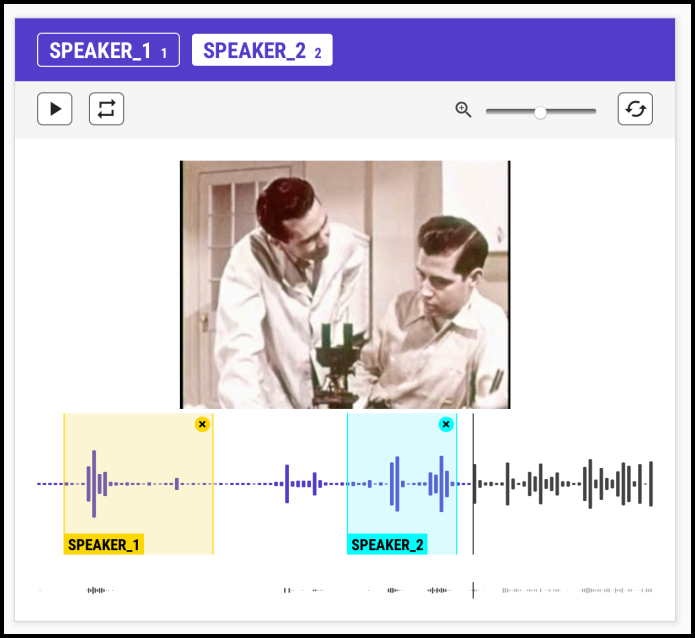


Recrute Annotator
Outsourcing
Crowdsourcing

- 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

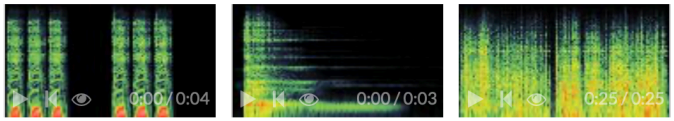


Data Labeling

freesound

Vibration

Try another category

Hierarchy	> Channel, environment and background > Noise > Vibration
Description	The general class of sounds caused by a periodic mechanical oscillation.
URI	http://en.wikipedia.org/wiki/Vibration
Examples	

Response type	Meaning
Present and predominant	The type of sound described is ***clearly present*** and ***predominant***. 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 ***present***, but the audio clip also ***contains other salient types of sound and/or strong background noise***.
Not present	The type of sound described is ***not present*** in the audio clip.
Unsure	***I am not sure*** whether the type of sound described is present or not.

Data Augmentation

Testing VS Training?

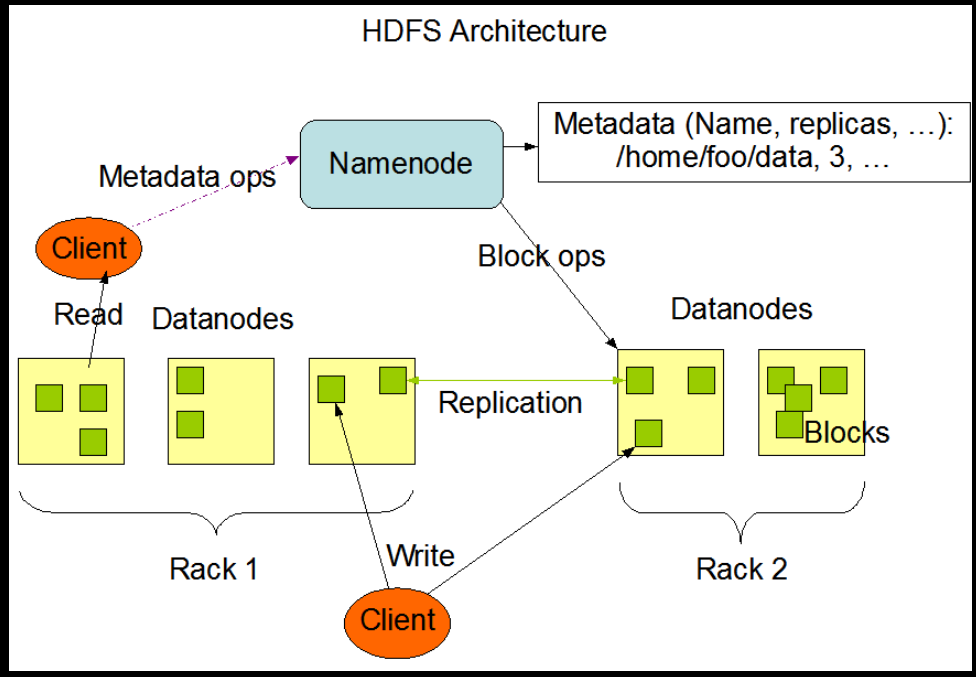
- Speech/Video
  - Change speed, add pause, etc.
- Vision
  - Make variations on original AI
- Tabular
  - Delete cells
- Text

Synthetic Dataset

- Simulated Dataset
  - Driving and robotics
- GAN

Data Storage

- Filesystem
  - Can be networked and distributed (e.g.HDFS)
    - Fault tolerant
    - High throughout



Data Storage

- Object Storage
  - Amazon S3, Google Cloud, etc.
    - Scalable
    - Unstructed data, customizable metadata
    - Versioning

Data Storage

- Databases
  - 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)
```

Data Storage

- Databases
  - Relational databases
  - SQL
  - Not versioned

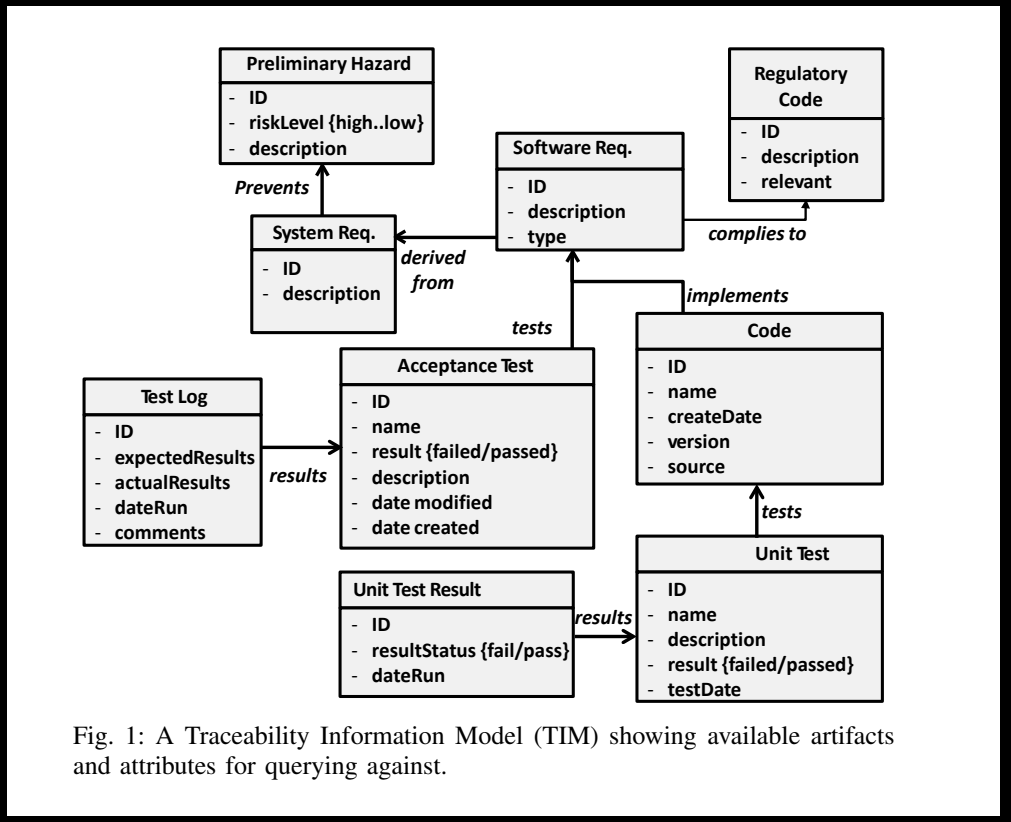
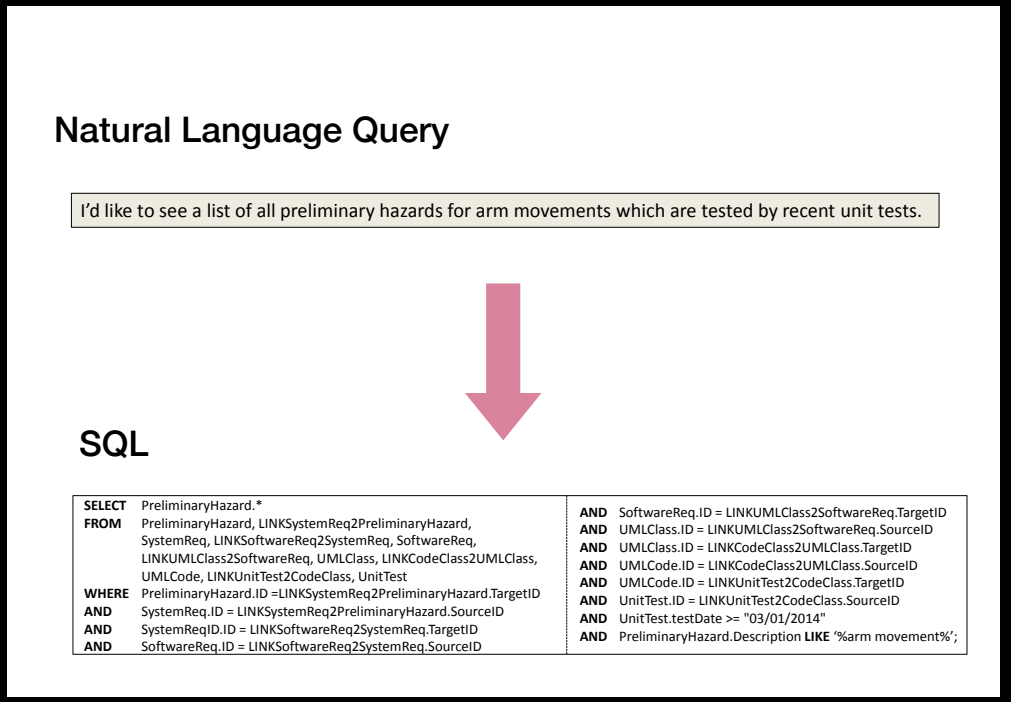


Fig. 1: A Traceability Information Model (TIM) showing available artifacts and attributes for querying against.

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

- Databases
  - Relational databases
  - SQL
  - Not versioned

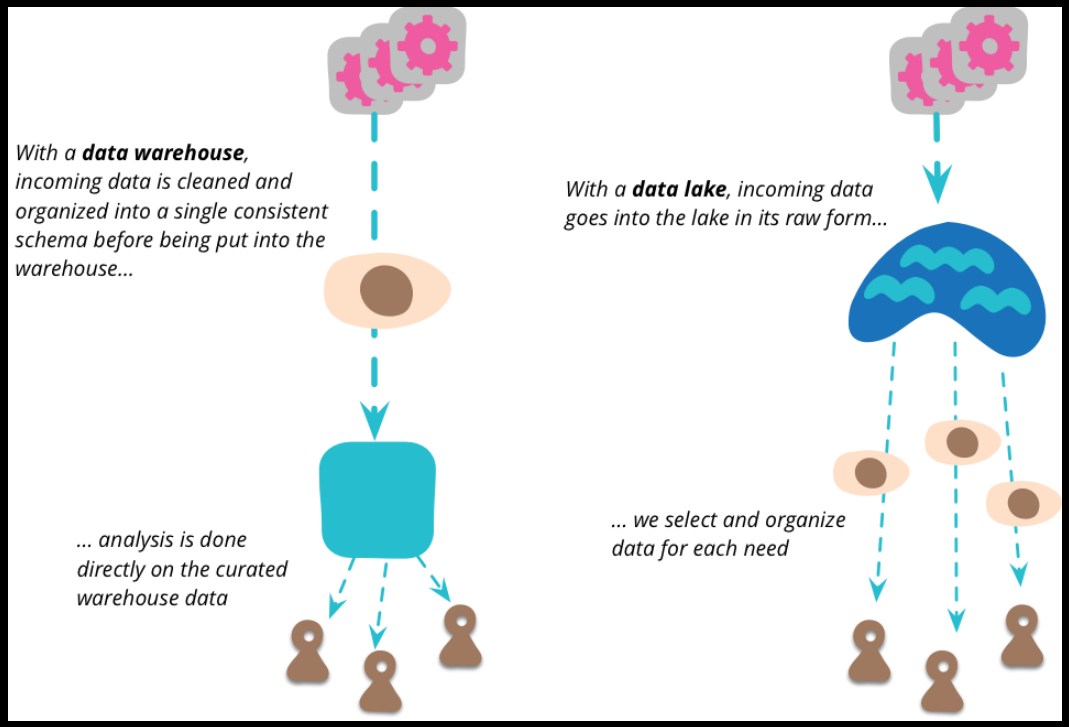


Lin, Jinfeng, et al. "Tiqi: A natural language interface for querying software project data." *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2017.

Data Storage

- Databases
  - 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



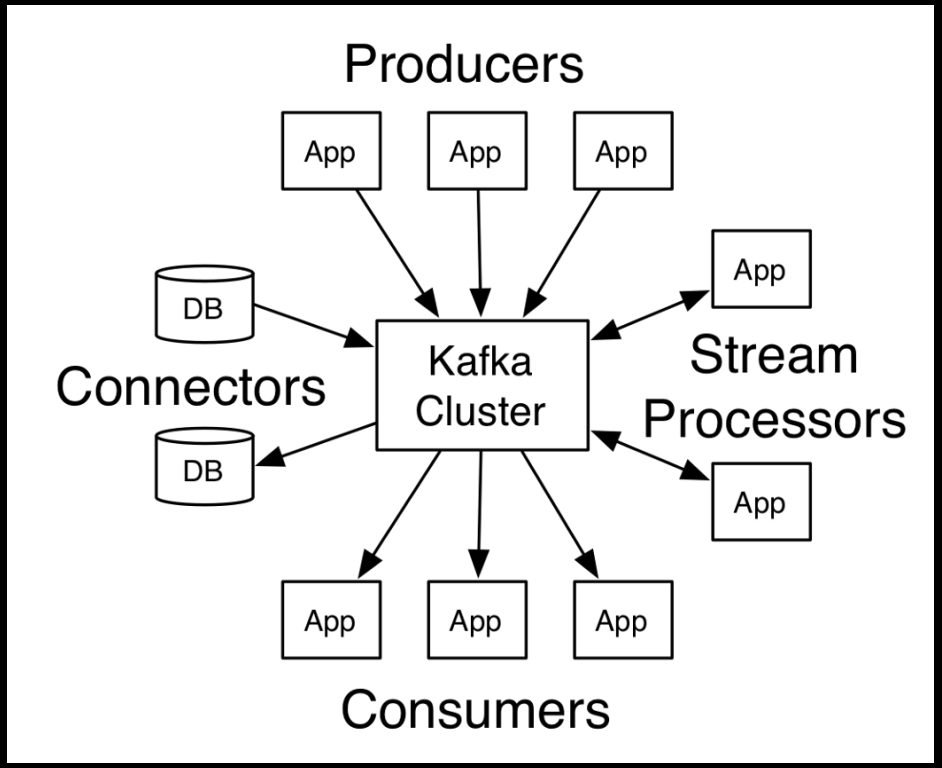
By Martin Fowler: https://martinfowler.com/bliki/DataLake.html#footnote-mart-v-warehouse

Data Processing

- Batch
  - A collection of data points that have been grouped together within a specific time interval
- Stream
  - Continuous data in real time

Event Streaming

- Apache Kafka



Data Pipeline Testing

- Regression test
- Health Check: job has succeeded
- Correctness (e.g. fake and original data)
- Latency (time for the data pipeline to complete)
- Time Series Metrics

Later in this course

- Data versioning (in the context of ML engineering)
- Data specification and validation