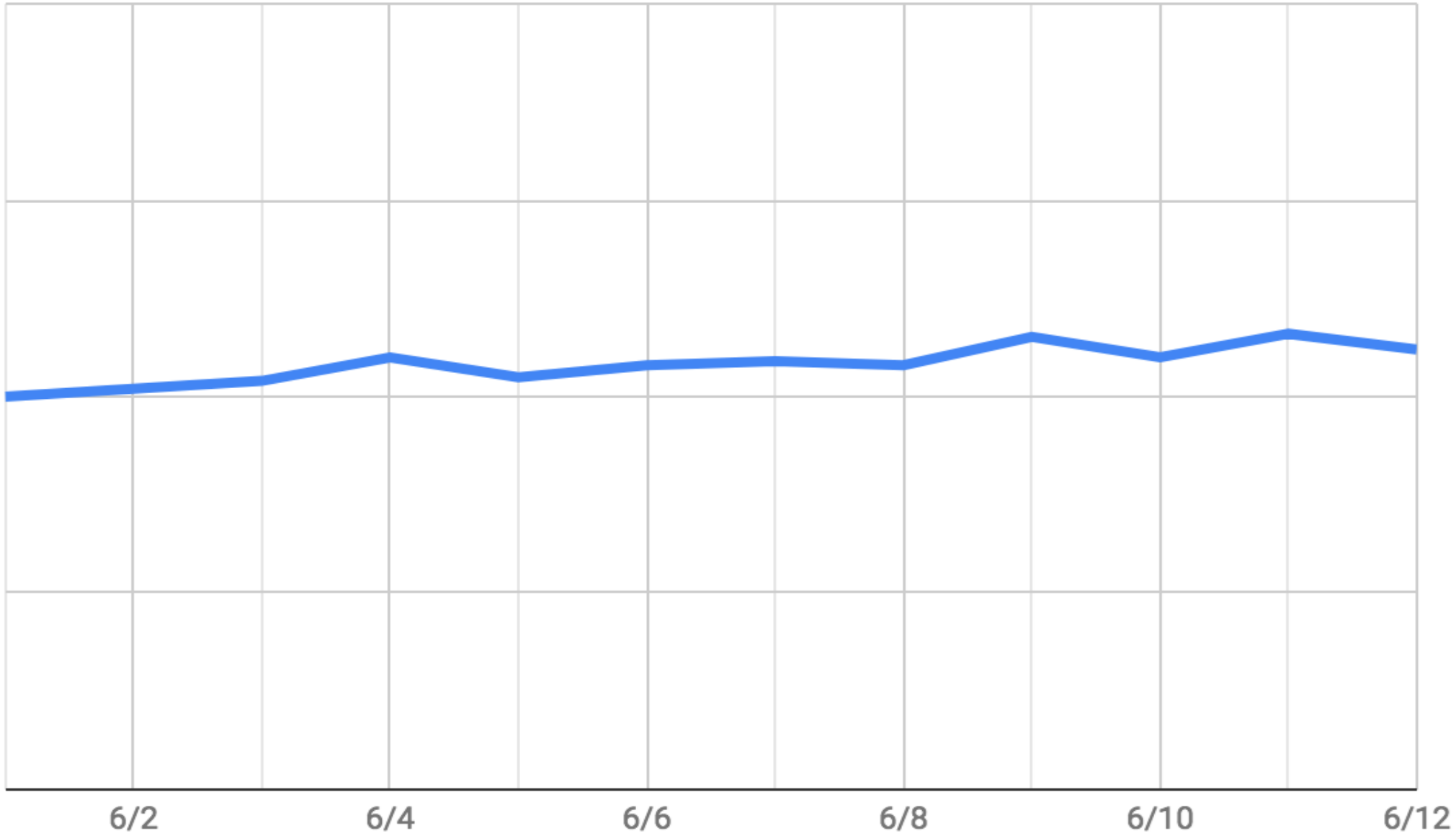
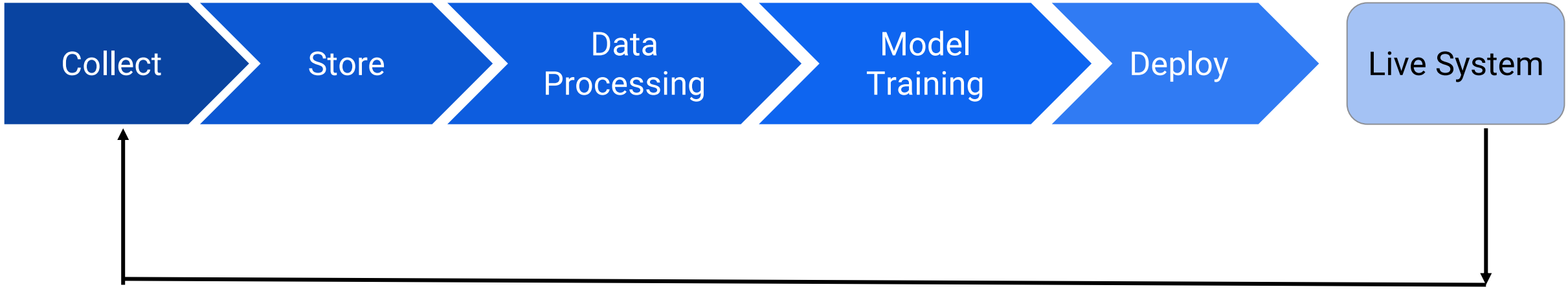
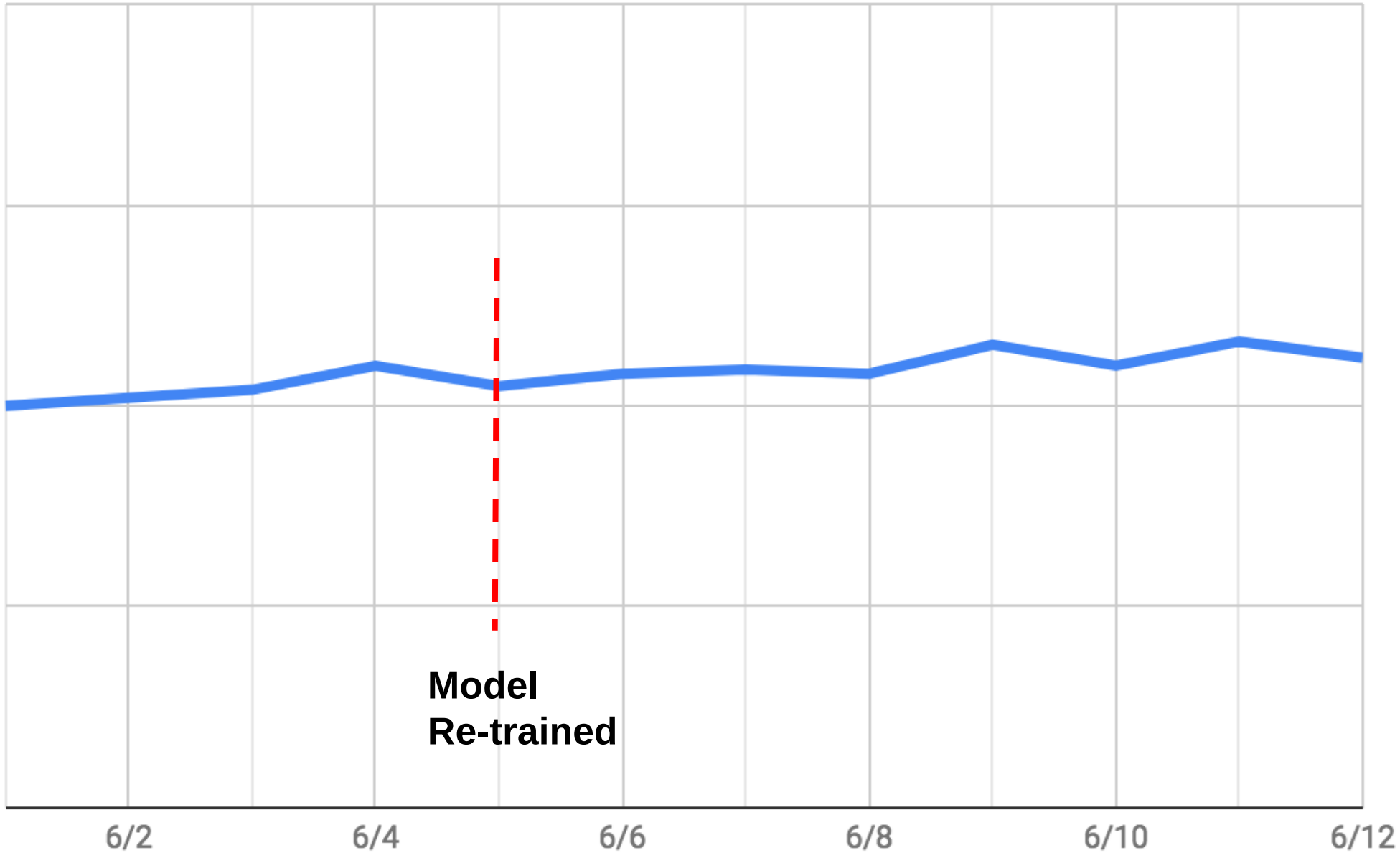
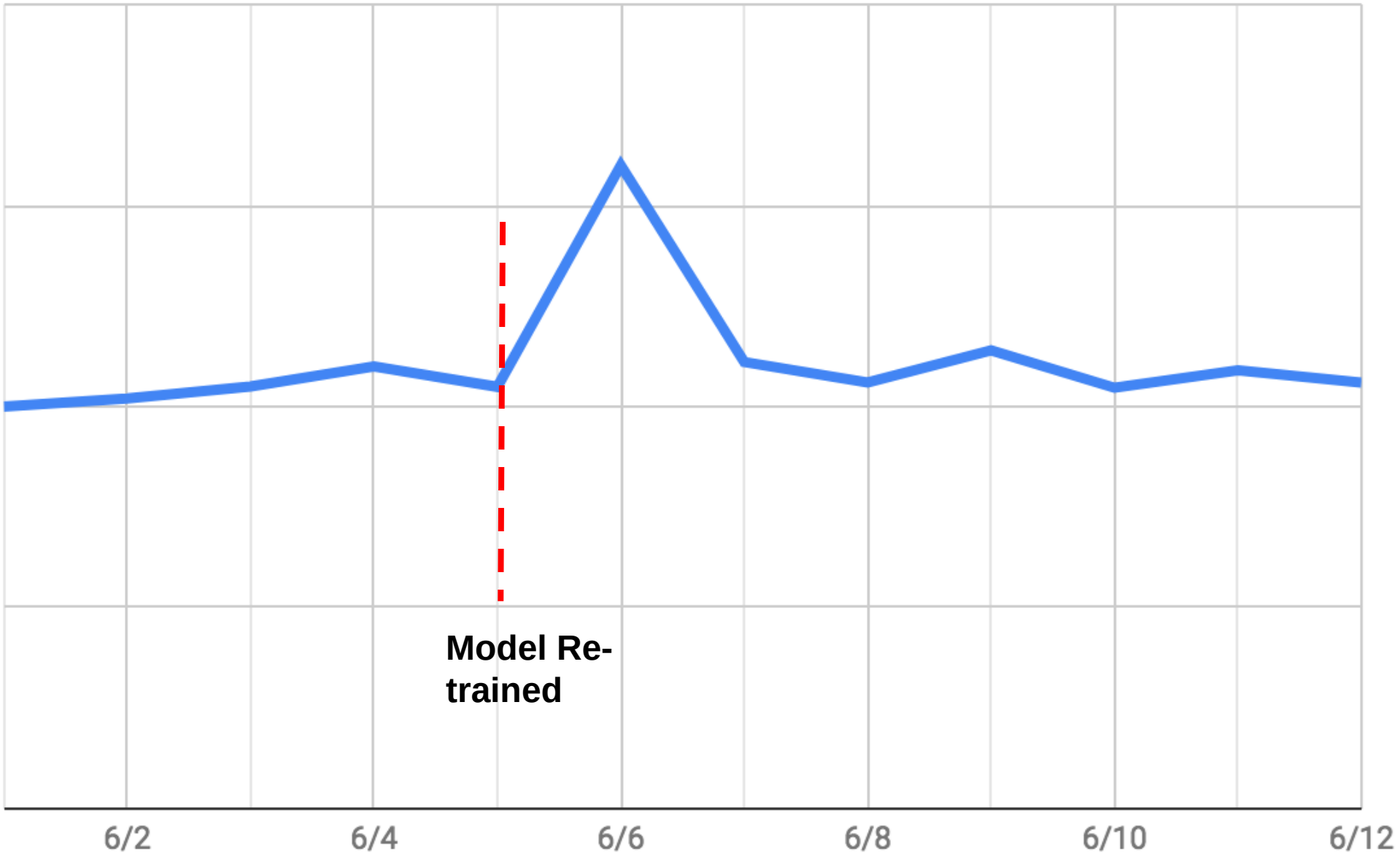
# Re-Training ML Models without Data Auditing is like Skydiving without Parachutes

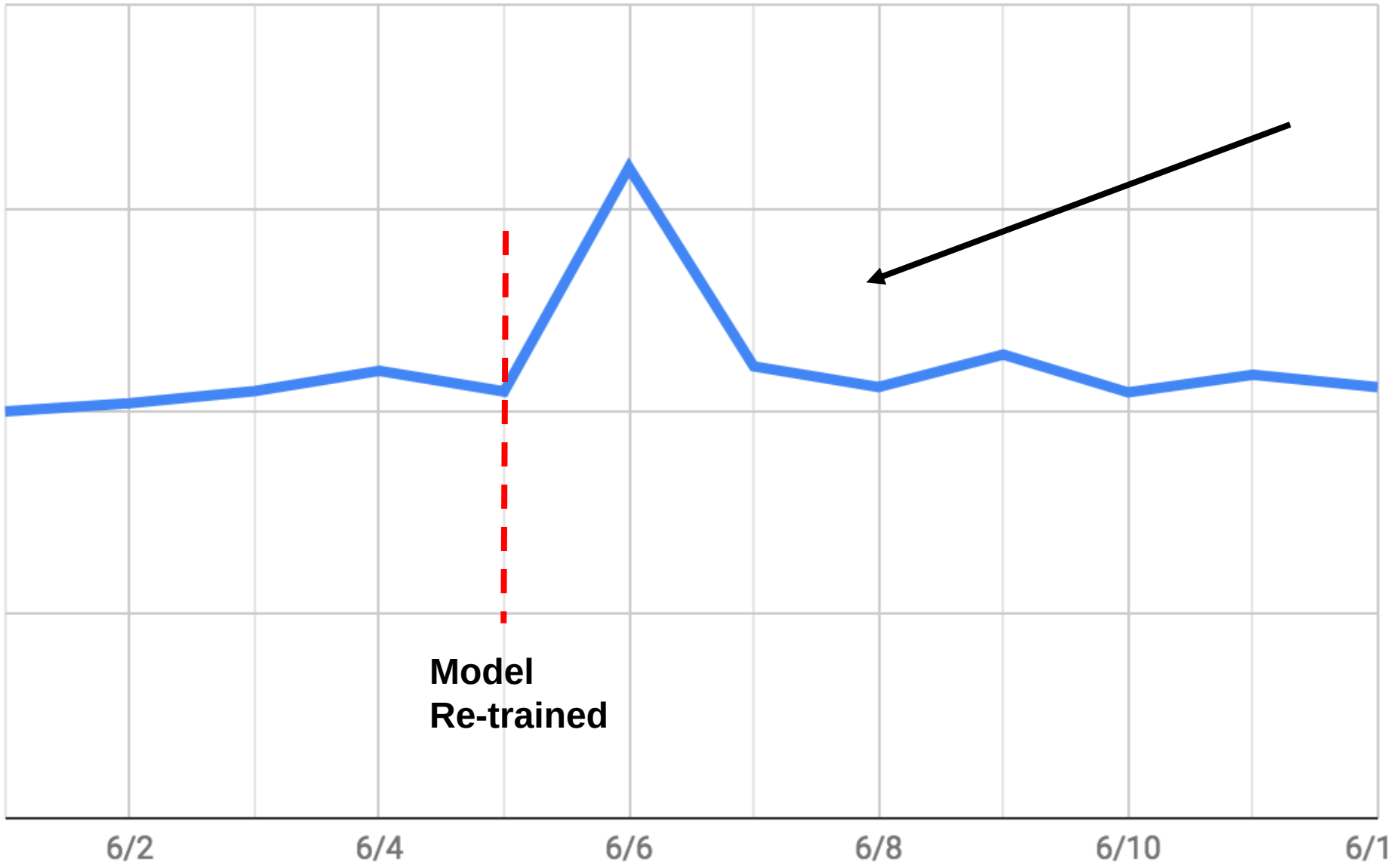
*Sandeep Uttamchandani*
*Chief Data Architect & Director of Engineering,*
*Small Business & Self Employed Group*
*Intuit*


6/2
6/4
6/6
6/8
6/10
6/12


Collect
Store
Data Processing
Model Training
Deploy
Live System


Model
Re-trained
6/2
6/4
6/6
6/8
6/10
6/12


Model Re-trained
6/2
6/4
6/6
6/8
6/10
6/12

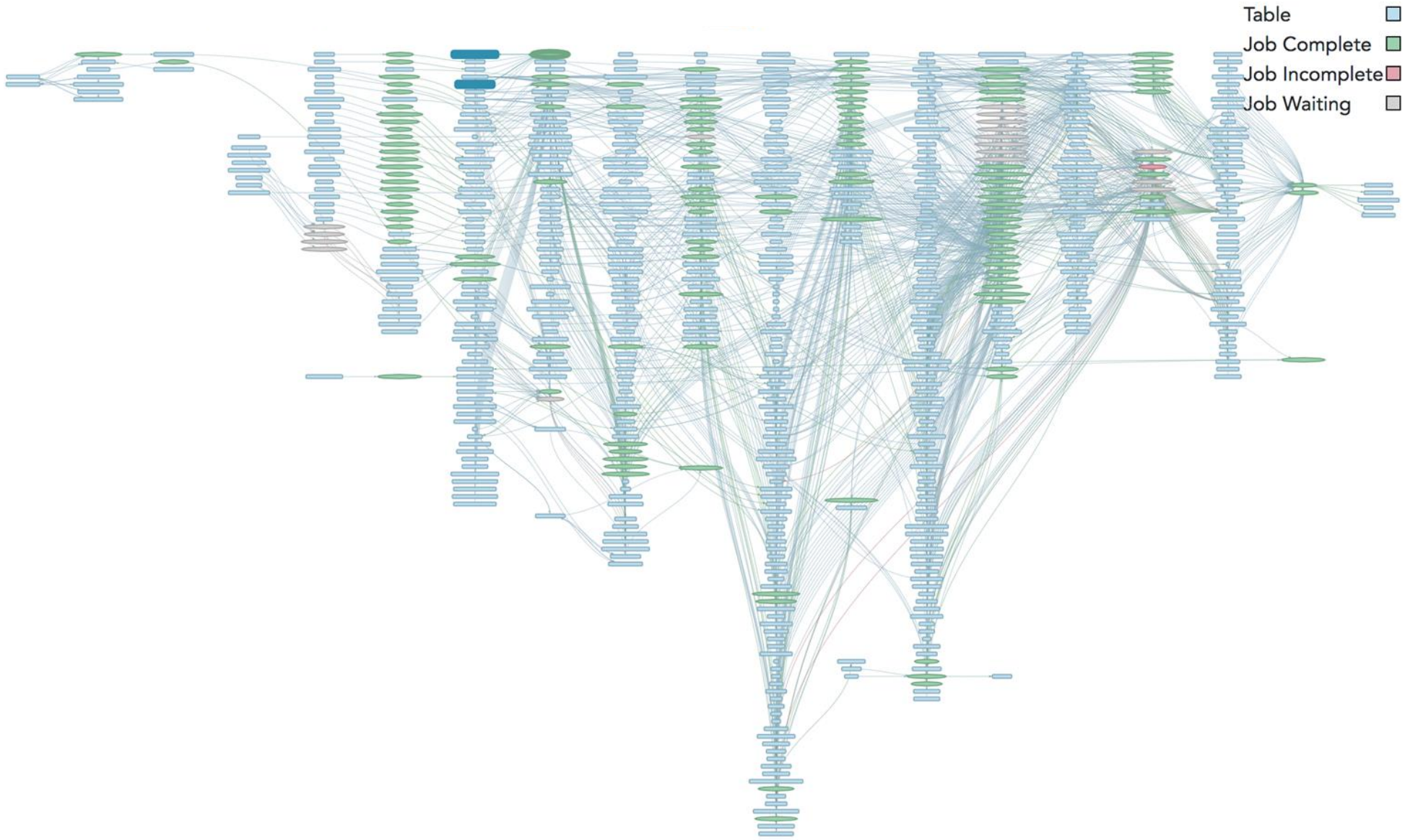



**Real?**

**OR**

**Issue?**

# Data Pipeline Jungle



# Real-world Data Pipeline Issues

## Data Source Issues

- Table inconsistencies
  a. Illegitimate values
  b. Missing values
  c. Duplicate Primary keys
- Hard deletes
- Bulk inserts
- Missing updates to CDC column

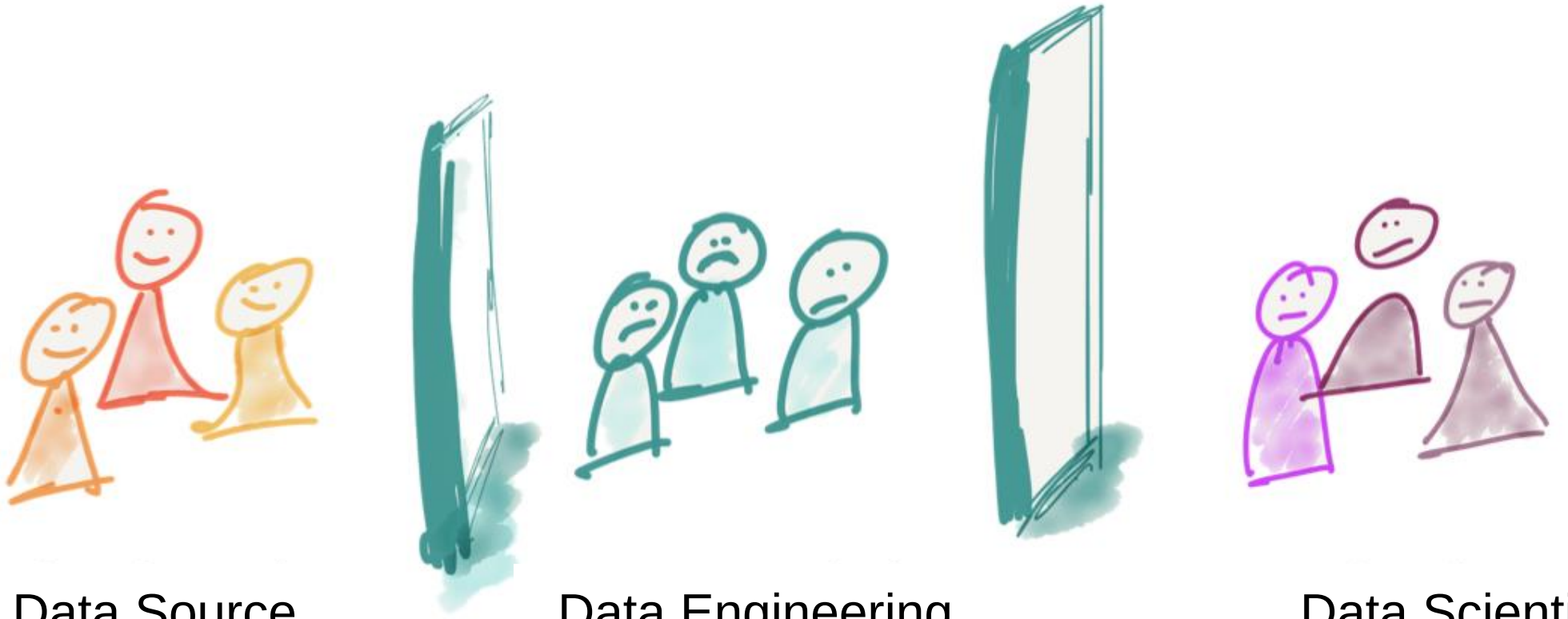
## Data Ingestion Issues

- Uncoordinated upstream changes
  a. Volume of data
  b. Change in schema
  c. Change in meaning of data
  d. Upgrade of platform
- No CDC for large tables leading to delayed availability
- Errors in ETL logic
- Timezone inconsistencies
- Duplicate or null records due to ingestion errors

## Referential Integrity Issues

- Data elements have different data types and/or meaning in different sources
- Inconsistent data element enums
- Heuristic ID Correlation
- Uncoordinated schema changes
- Dropped updates across data sources

# #NotMyProblem

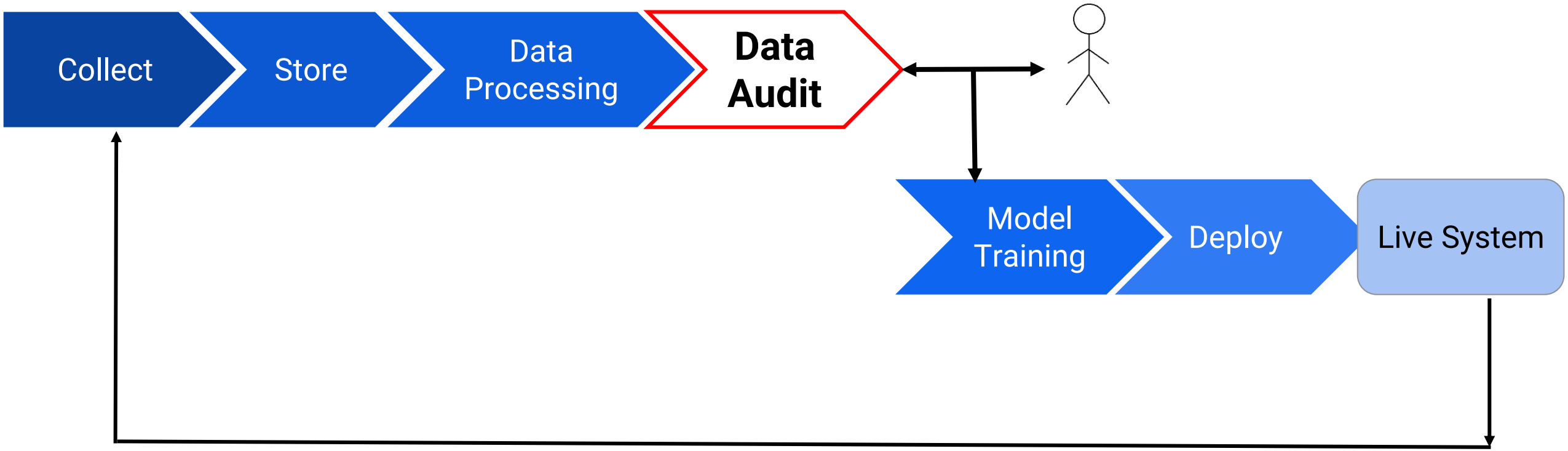


Data Source Teams

Data Engineering Team

Data Scientists/ Analysts

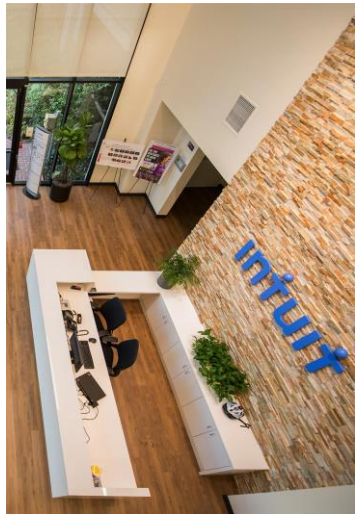
Image Credit: https://martinfowler.com/articles/data-monolith-to-mesh.html

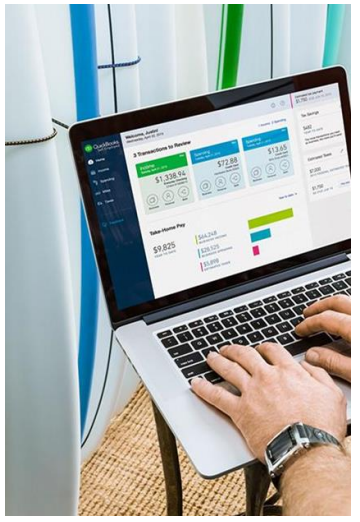

Collect
Store
Data Processing
Data Audit
Model Training
Deploy
Live System

powering prosperity

Who we are:

| 1983 | 1993 | 8,900 | 20 | $6B FY18 | 50M |
|---|---|---|---|---|---|
| Founded | IPO | Employees | Locations | Revenue | Customers |



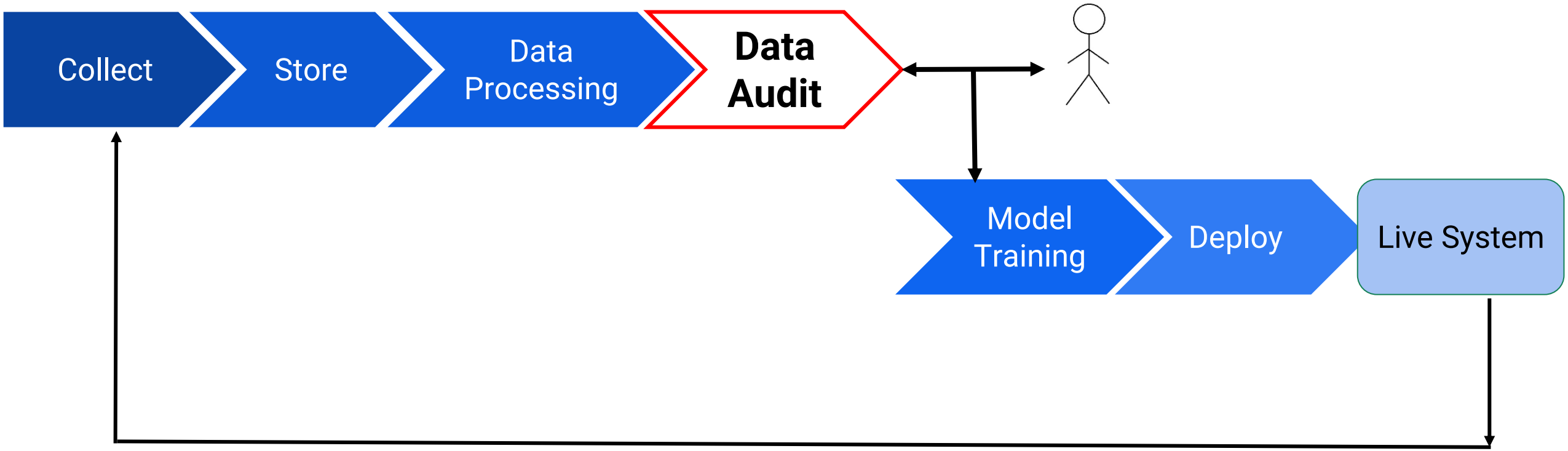



Who we serve:
Consumers
Small businesses
Self-employed
intuit
powering prosperity
turbotax
quickbooks
proconnect
mint

intuit powering prosperity

turbotax quickbooks proconnect mint

# Powering Prosperity Around the World


Collect
Store
Data Processing
Data Audit
Model Training
Deploy
Live System

# Data Audit Process

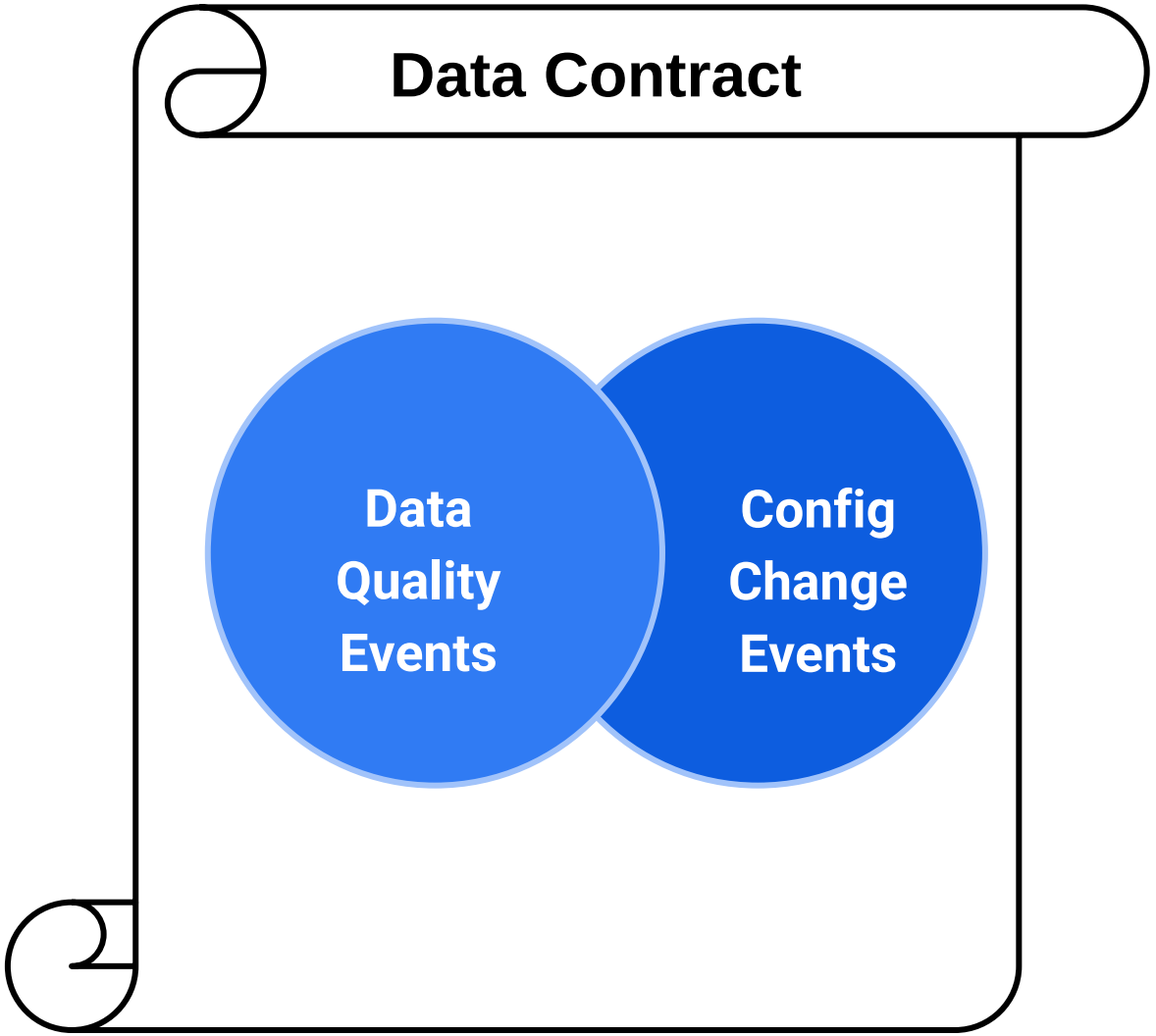
Data Source Teams

Data Engg Team

**Data Contract**

Data Scientists/ Analysts

# Data Audit Process



# Data Quality Events

### Single Column

- Cardinalities
- Patterns & Data types
- Value distributions
- Domain classification

### Multi-Column

- Correlations
- Association rules
- Clustering
- Outliers
- Summaries & sketches

### Cross-DB dependencies

- Unique column combinations
- Inclusion dependencies
- Functional dependencies

Reference: Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2015. Profiling relational data: a survey. The VLDB Journal 24, 4 (August 2015), 557-581.

# Data Profiling

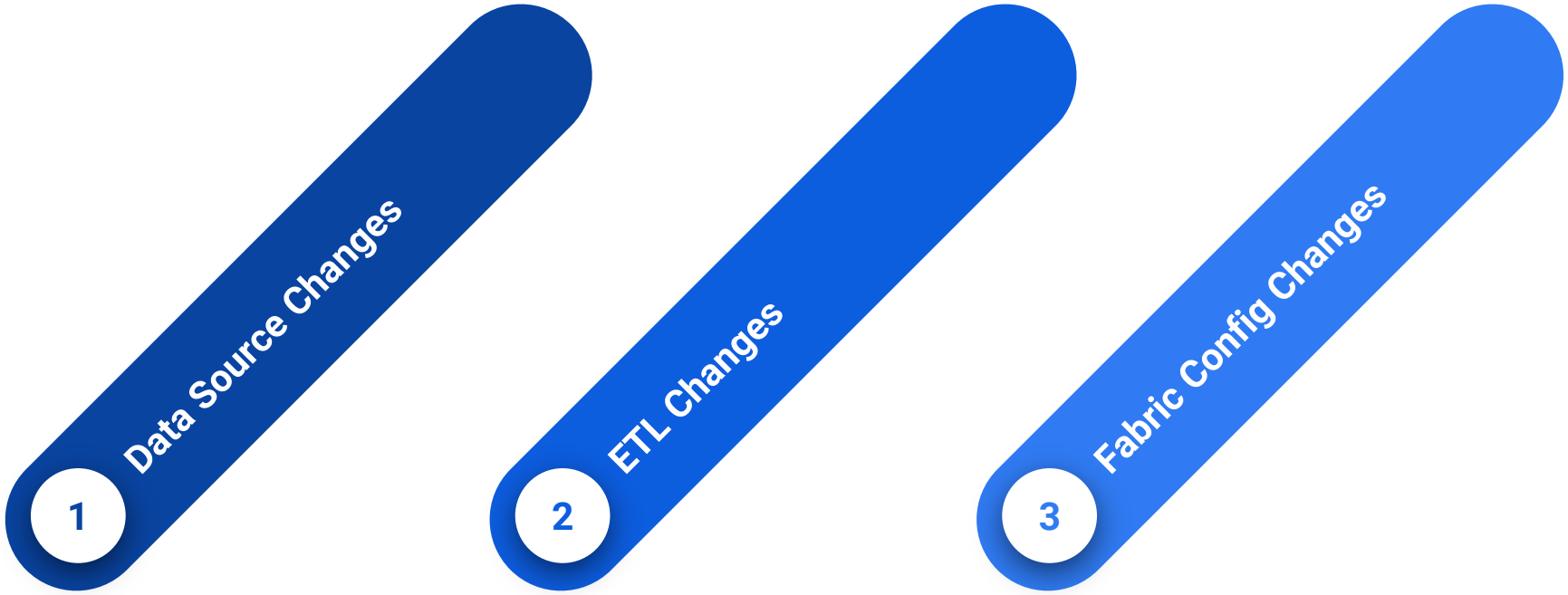
## Single Column

- Cardinalities
- Patterns & Data types
- Value distributions
- Domain classification

| Category | Task | Description |
|---|---|---|
| Cardinalities | num-rows | Number of rows |
| | value length | Measurements of value lengths (minimum, maximum, median, and average) |
| | null values | Number or percentage of null values |
| | distinct | Number of distinct values; sometimes called "cardinality" |
| | uniqueness | Number of distinct values divided by the number of rows |
| Value distributions | histogram | Frequency histograms (equi-width, equi-depth, etc.) |
| | constancy | Frequency of most frequent value divided by number of rows |
| | quartiles | Three points that divide the (numeric) values into four equal groups |
| | first digit | Distribution of first digit in numeric values; to check Benford's law |
| Patterns, data types, and domains | basic type | Generic data type, such as numeric, alphabetic, alphanumeric, date, time |
| | data type | Concrete DBMS-specific data type, such as varchar, timestamp, etc. |
| | size | Maximum number of digits in numeric values |
| | decimals | Maximum number of decimals in numeric values |
| | patterns | Histogram of value patterns (Aa9...) |
| | data class | Semantic, generic data type, such as code, indicator, text, date/time, quantity, identifier |
| | domain | Classification of semantic domain, such as credit card, first name, city, phenotype |

Reference: Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. 2015. Profiling relational data: a survey. The VLDB Journal 24, 4 (August 2015), 557-581.

# Config Change Events



# Data Contract Clauses

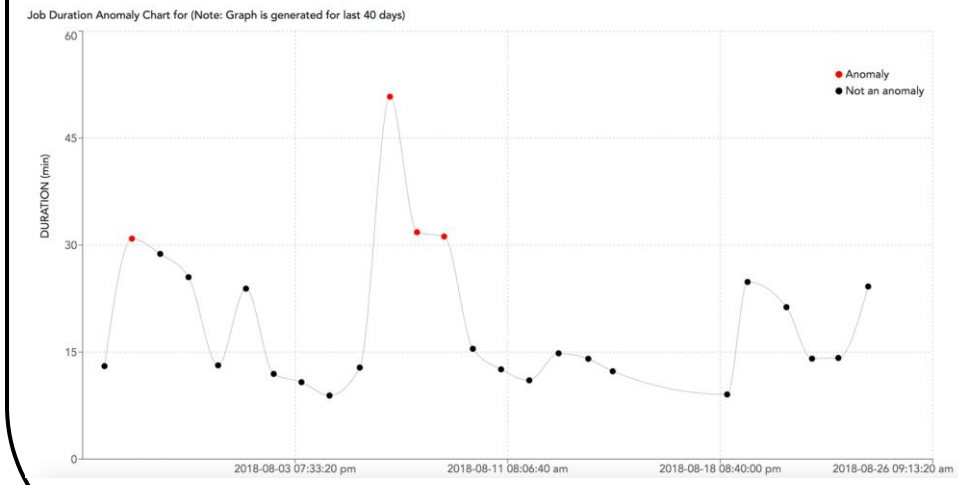
## Absolute Threshold Rules

**If current_state >= threshold_value**
**Then <Alert>**

| | **Example** |
|---|---|
| Operational Rule | If job_retries > 3, then fail job |
| Data Value Rule | If col_val > 2$\sigma$, then alert |

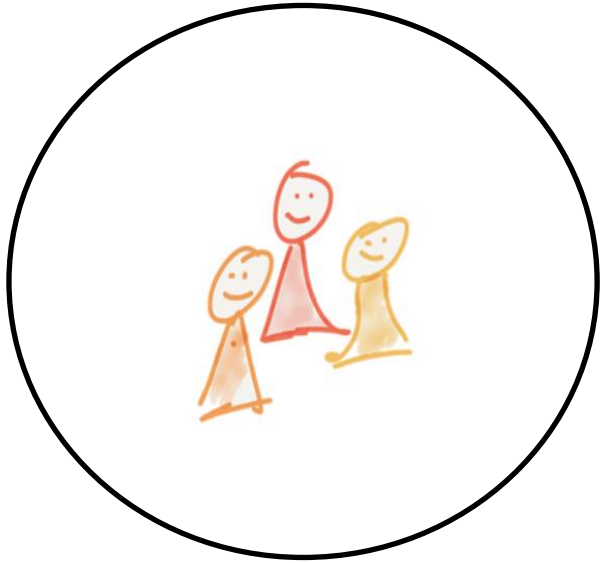
## History-based Anomaly Rules

**If current_state <*differs*> history_state**
**Then <Alert>**



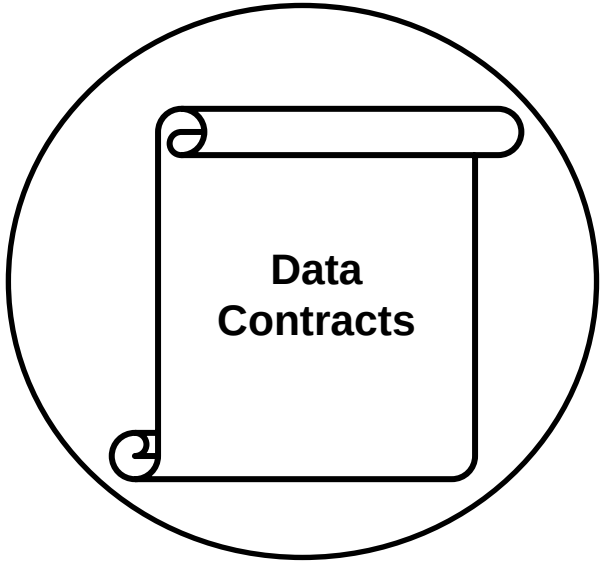
# Holistic Approach to Data Audit for ML Training

People



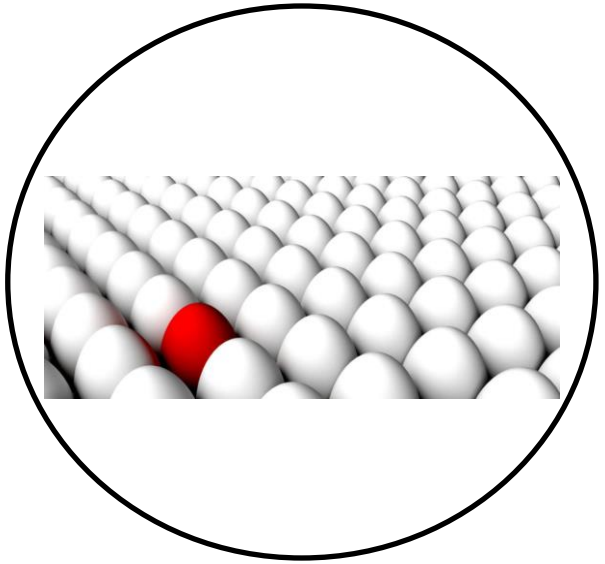
+

Process




+

Technology



**We are hiring! Come join us!**

https://careers.intuit.com/