APPLIED ROOTS

Data Quality & Verification @ Uber, Netflix & Amazon

APPLIED ROOTS

Active Learning

↳ Participation & Discussion

Let's solve it together...

APPLIED ROOTS

Sources:

1. https://eng.uber.com/monitoring-data-quality-at-scale/ (2020) ← Uber DQM
2. https://databricks.com/session_na20/an-approach-to-data-quality-for-netflix-personalization-systems (2020)
3. https://www.amazon.science/publications/automating-large-scale-data-quality-verification (2018)

APPLIED ROOTS

Scale @ all companies

~ 1 - 10PB

10K - 100K tables

100s of Millions of events

1000's of metrics

APPLIED ROOTS

(Q) what platform(s) & techniques would you use to process internet-scale data?

APPLIED ROOTS

Data Quality impacts all downstream tasks

↓

Data Analysis

Biz decision-making

ML & DL models

Infra-costs

APPLIED ROOTS

Let's map Data Quality to an ML-problem?

Any Suggestions?

APPLIED ROOTS

Data Quality

- Time-series of metrics (detect anamolies)
  - KPI
  - utilization
  - Table
    - # rows
    - column-level
      - missing values
      - min; max; unique-values
      - ...

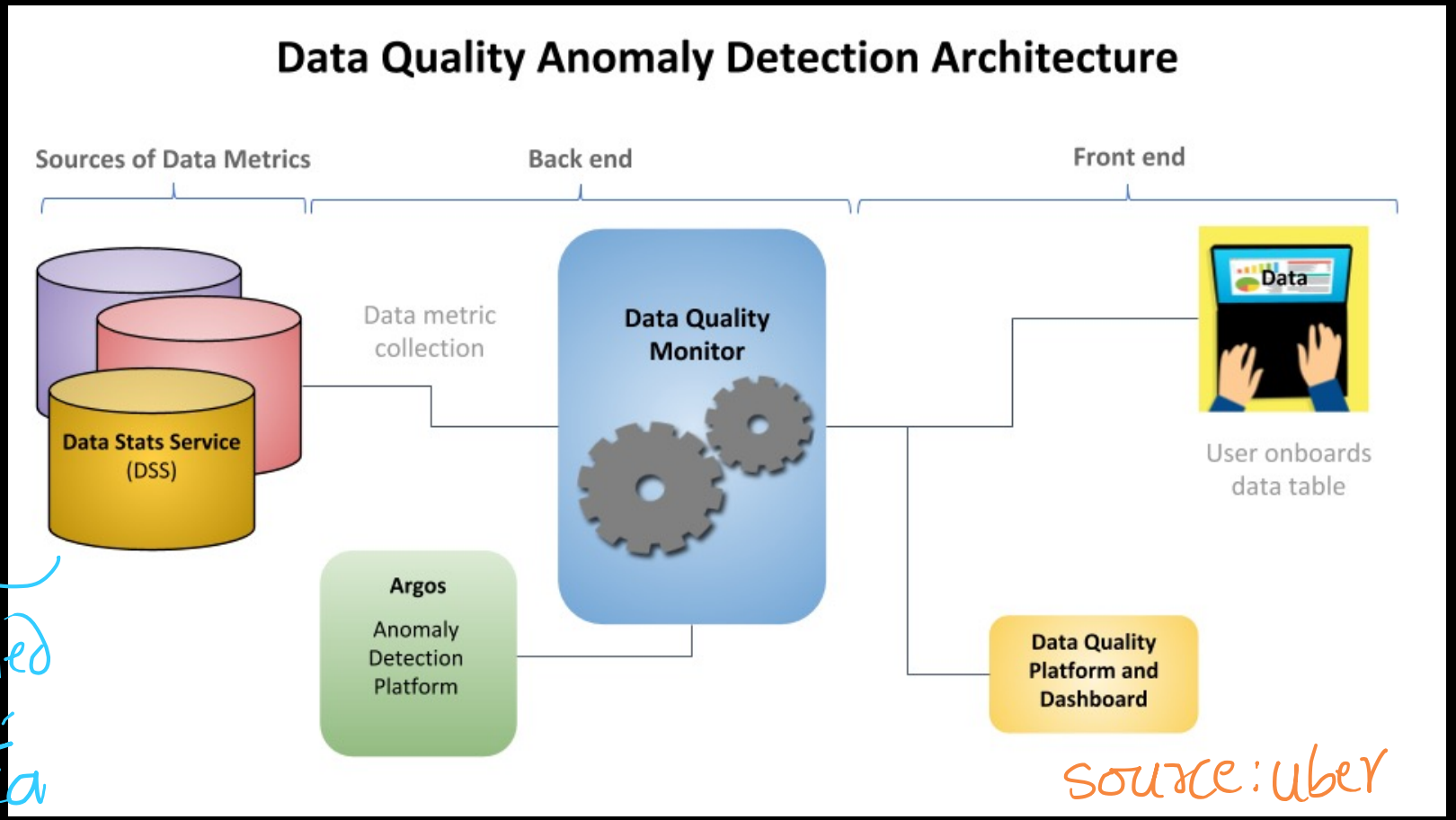
APPLIED ROOTS

— Data Anamolies $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$

(Q) what techniques can we employ here ?

APPLIED ROOTS

Uber DQM



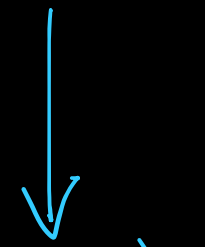
date partitioned

Hive & Vertica

source: uber

APPLIED ROOTS

DSS: multi-dimensional time series output



(DQM) → PCA & project to lower-dim

→ Time-series anamoly detection (using forecasting)

APPLIED ROOTS

(Q) why do you PCA is helpful here?

APPLIED ROOTS

many metrics are highly correlated

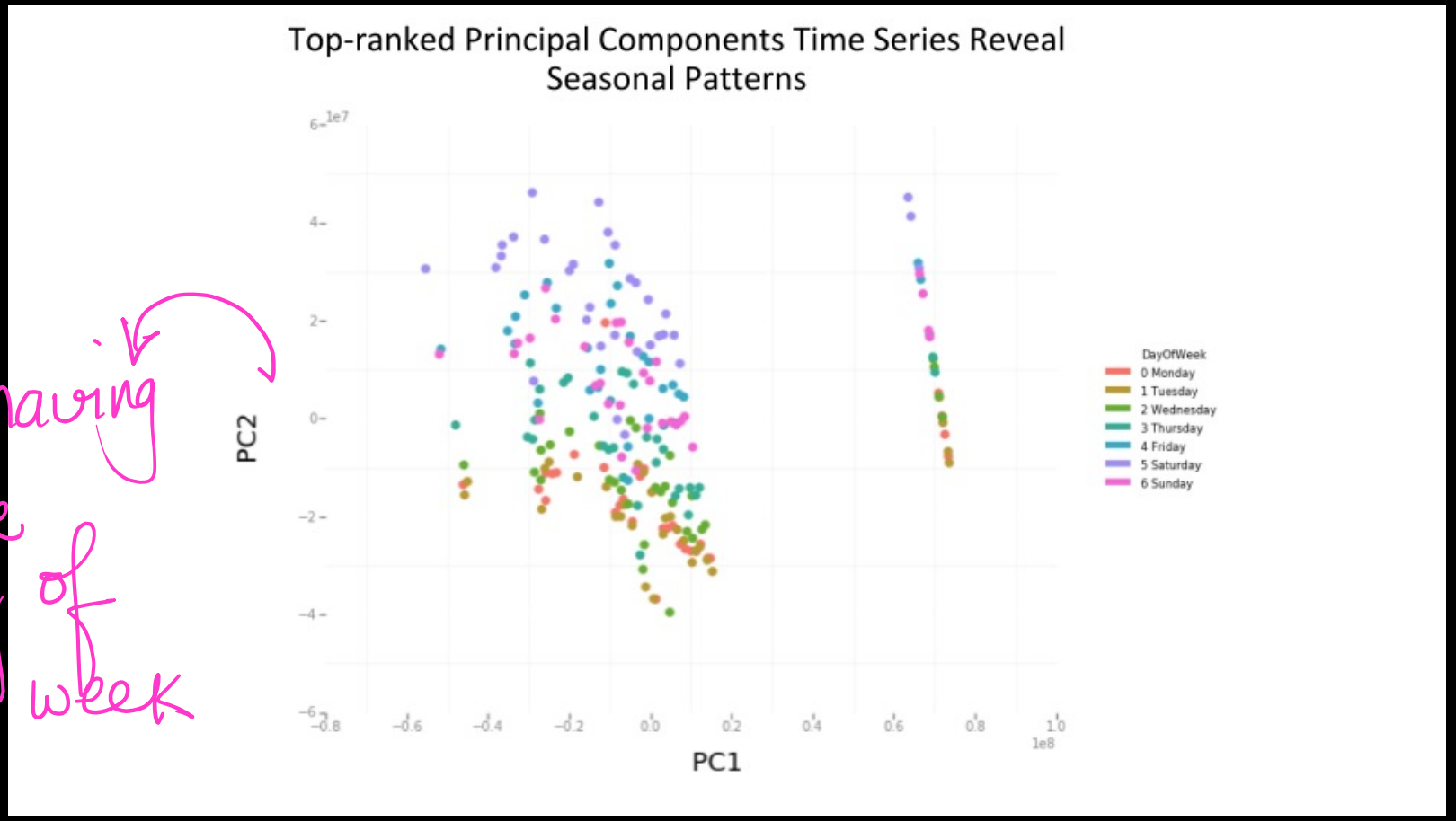
↓

eg: trip duration & trip distance

Uber: 90% of variability explained by the top-PC.

APPLIED ROOTS

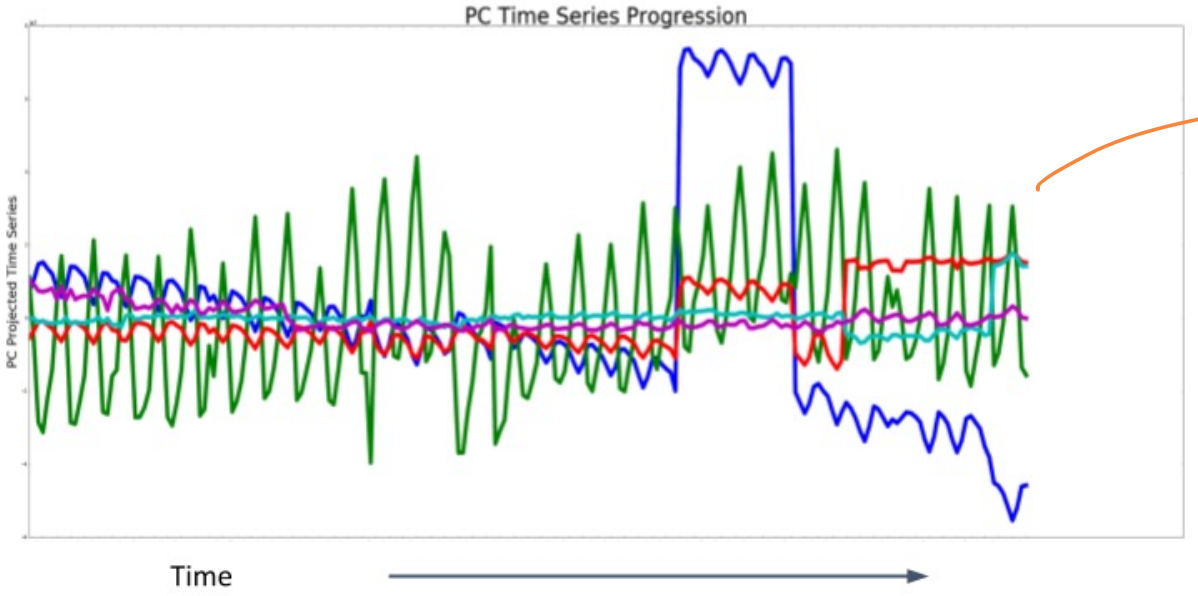
source: uber



behaving like day of week

APPLIED ROOTS

Principal Component Time Series Shows State Changes and Seasonalities



Top 5-PCs across time

APPLIED ROOTS

(Q) why not perform time-series anamoly detection on raw metrics ?

APPLIED ROOTS

Analyzing top-5 PCs

vs

Analyzing 100's of metrics

APPLIED ROOTS

Time-series anamoly detection

- via forecasting
- Holt-winters model
  - exponential moving average
  - Seasonality
  - Trends

https://otexts.com/fpp2/holt-winters.html

APPLIED ROOTS

(Q) why exponential smoothing ?

APPLIED ROOTS

recent data is given more weightage

↓

fast growing startup with sudden changes

APPLIED ROOTS

Seasonality → weekends; evening-hours; holidays ...

Trend → Typically to capture broader growth over time.

APPLIED ROOTS

Lesson: Use the 'simplest' model that works.

→ easy to interpret

↳ do NOT need DL everywhere

APPLIED ROOTS

Table-level scores

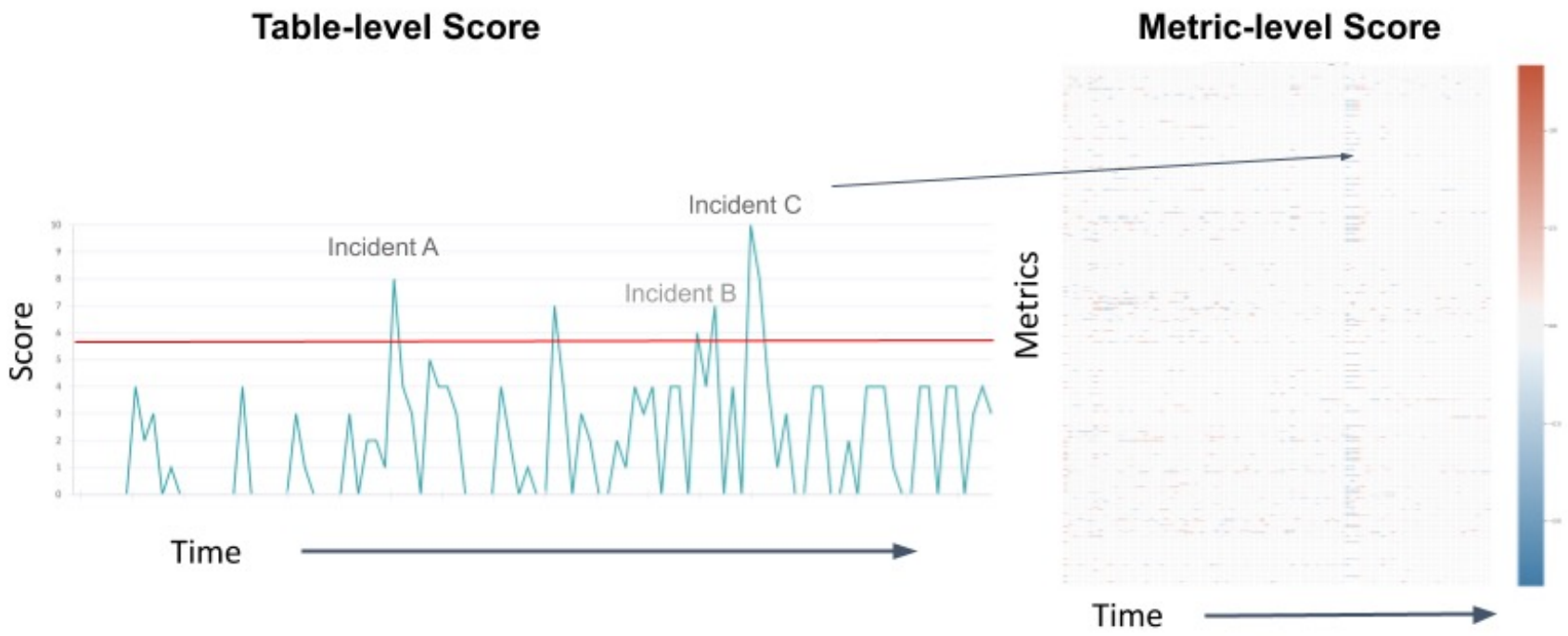
↳ Time-series of data in a table

→ using top PC's with equal weightage

→ anomaly detection

source: Uber

Anomaly Capture with Table and Metric-Level Scores



APPLIED ROOTS

DQM: Pyspark & Hive

↳ (what better platform than this?)

Vertica (Columnar NoSQL database)

Figure 5. One of our next steps towards making alerts more intelligent is leveraging data table lineage information. In fact, we have observed strong correlation between data table quality and lineage as the clustering of table-level quality scores over time can reconstruct table ancestry. This is validated in practice as we see related tables have common root causes when they degrade in data quality.

Source : Uber

APPLIED ROOTS

Source: Netflix.com

Source ↓

https://databricks.com/session_na20/an-approach-to-data-quality-for-netflix-personalization-systems

(slides & video)

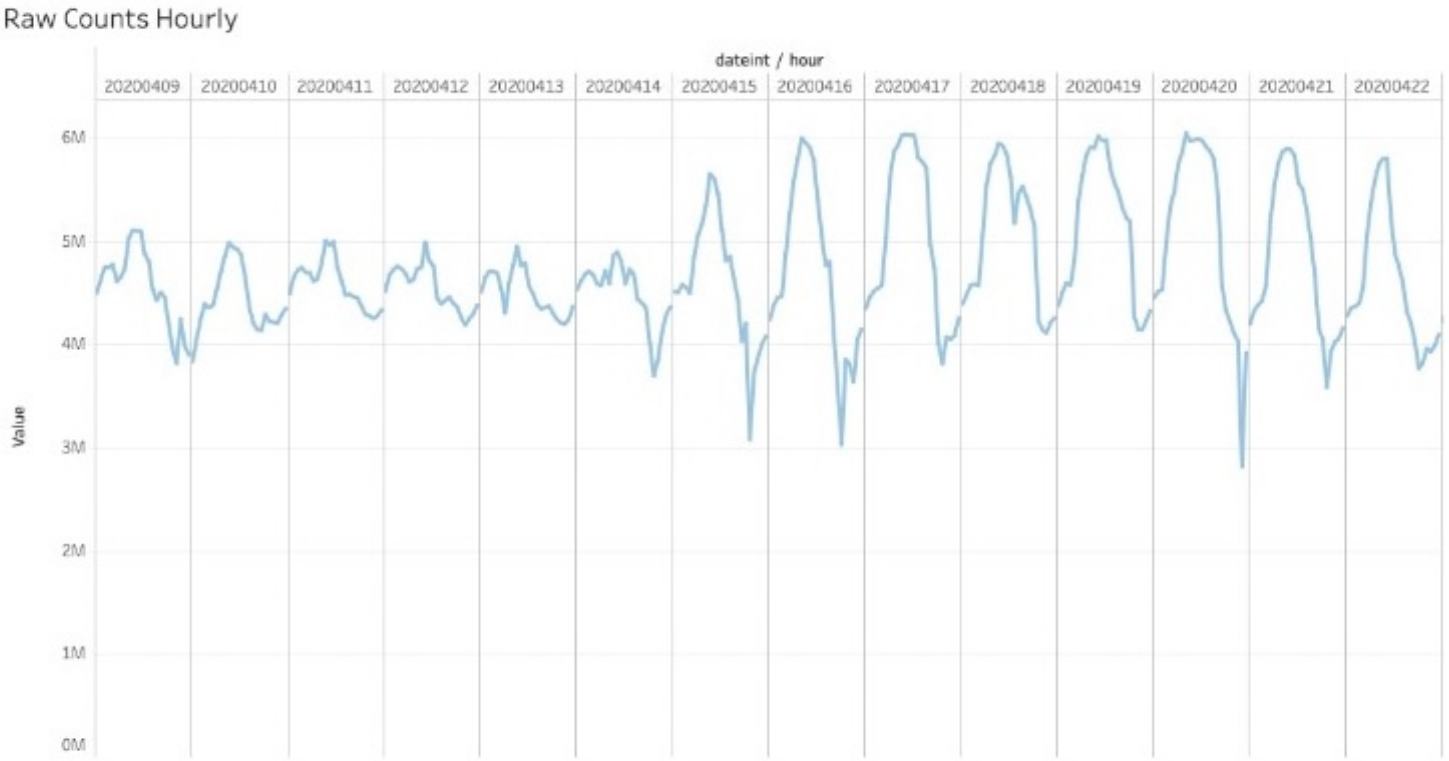
Source: Netflix.com

Netflix PI Historical Fact Store

- Manages more than 10PB of data
- Manages hundreds of attributes
- More than 1 Billion rows flow through the ETL pipelines per day
- This data is used by several machine learned models and algorithms that enable Personalization

APPLIED ROOTS

Bad Data Example 1: Drift



APPLIED ROOTS

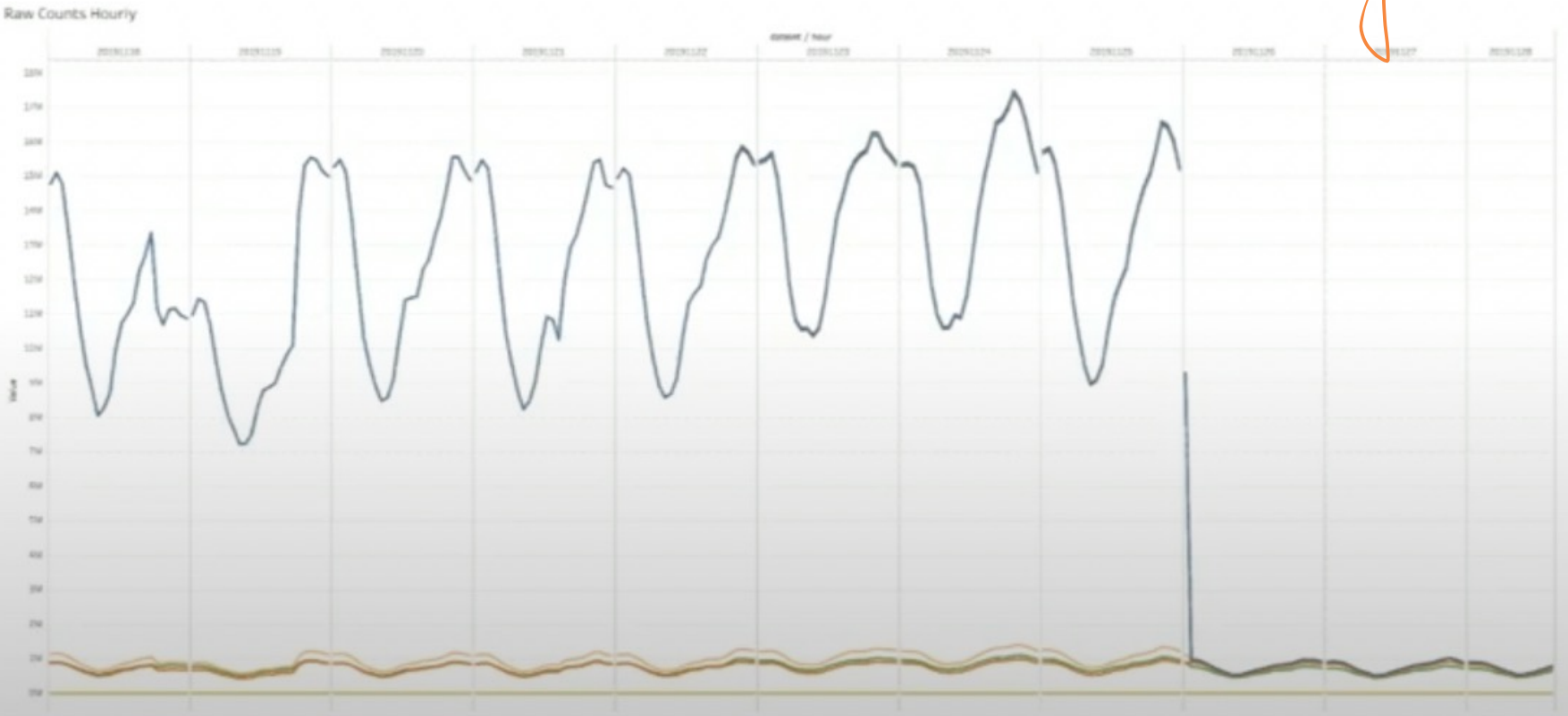
Source: Netflix.com

Bad Data Example 2: Drastic changes

APPLIED ROOTS

Very important but often ignored

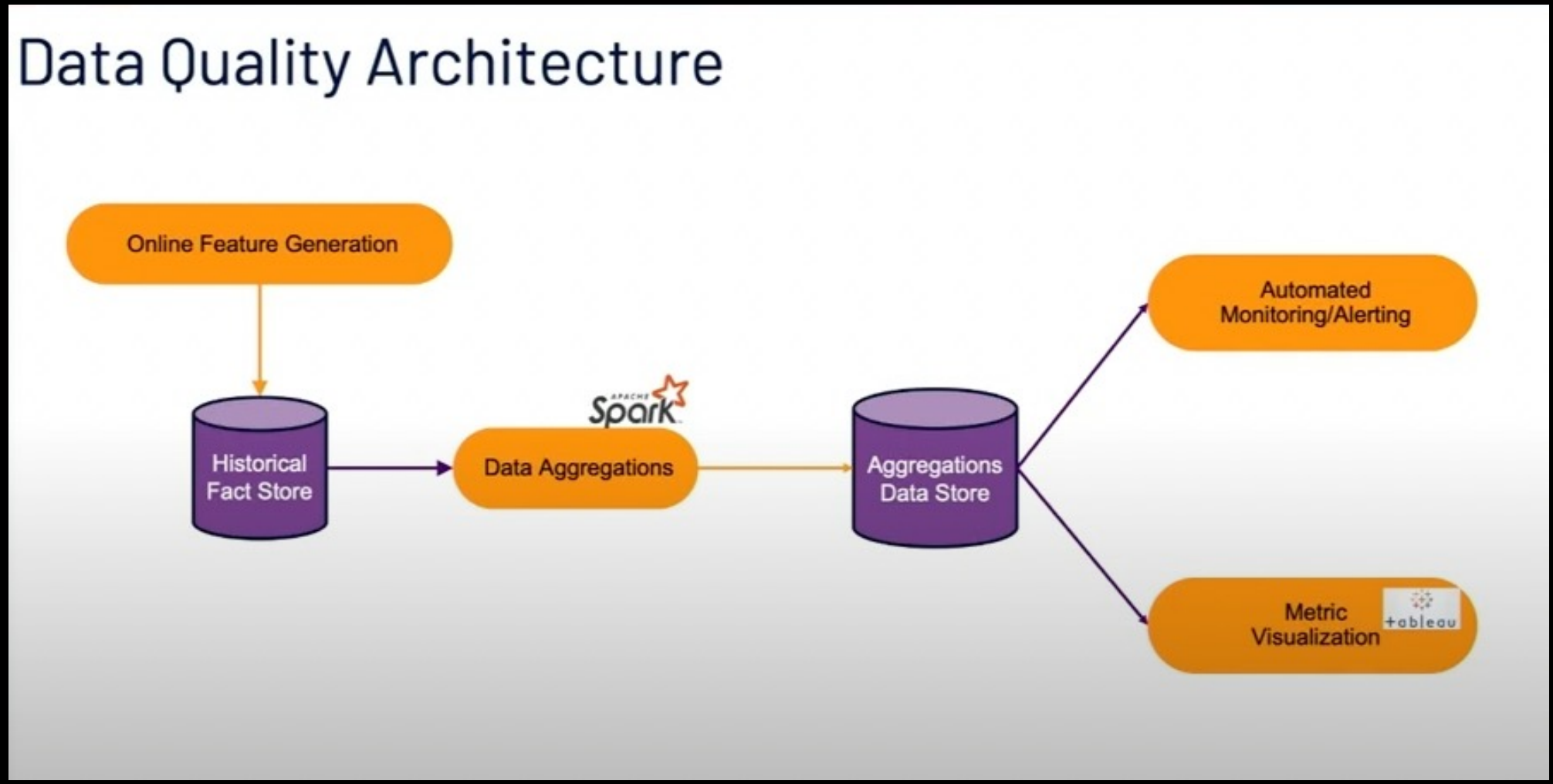
Bad Data Example 3: Under Utilization



Source: Netflix.com

APPLIED ROOTS

Source: Netflix.com



APPLIED ROOTS

Source: Netflix.com

Aggregations

Date (DD/MM/YYYY)	Number Of Null Plays	Median Play Duration (seconds)	99th Percentile Play Duration (seconds)
1/1/2020	3	360	450
1/2/2020	20	270	780
1/3/2020	1	150	570

APPLIED ROOTS

Source: Netflix.com

Automated Monitoring:

↳ distribution - matching of attributes

APPLIED ROOTS

(Q) Any suggestions ?

APPLIED ROOTS

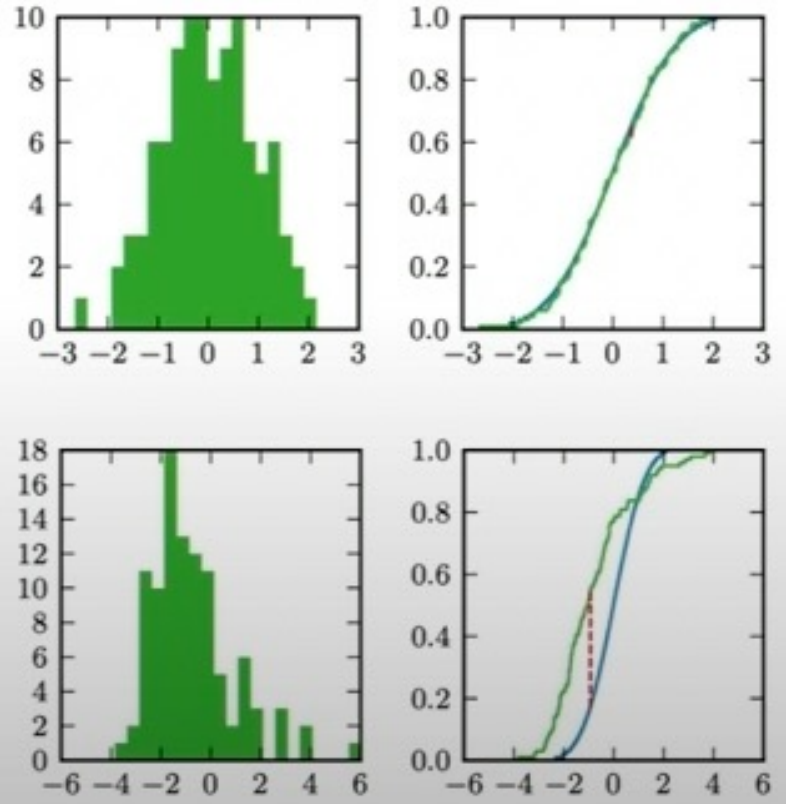
Source: Netflix.com

Distribution Checks – Statistical Test

Kolmogorov-Smirnov statistical test[1]

$$F_n(a) = \frac{1}{n} \sum_i \mathbb{I}(x_i < a)$$

$$D_n = \max_x |F_n^1(x) - F_n^2(x)|$$



[1]Source: MIT 6.S085

APPLIED ROOTS

Source: Netflix.com

Audits

```
from pyspark.sql import SQLContext
import AuditRunner, Audit

df = SQLContext.sql('SELECT * FROM ml.aggregations WHERE dateint = 20200303')

# Audit Pseudo code
auditRunner = AuditRunner().add(
    Audit (df,  'num_null_plays',      { 'upper_bound': 10 },              { 'alert': 'pager' } ),
    Audit (df,  'num_null_plays',      { 'normal_distribution': True },    { 'alert': 'dev_notification' } ),
    Audit (df,  'avg_play_duration',   { 'query_anomaly_detection': 'rad' }, { 'alert': 'dev_notification' } )
    # ...
)
```

Thresholds

distribution matching

Robust Anamoly Detcn

APPLIED ROOTS

Netflix RAD:

https://netflixtechblog.com/rad-outlier-detection-on-big-data-d6b0494371cc

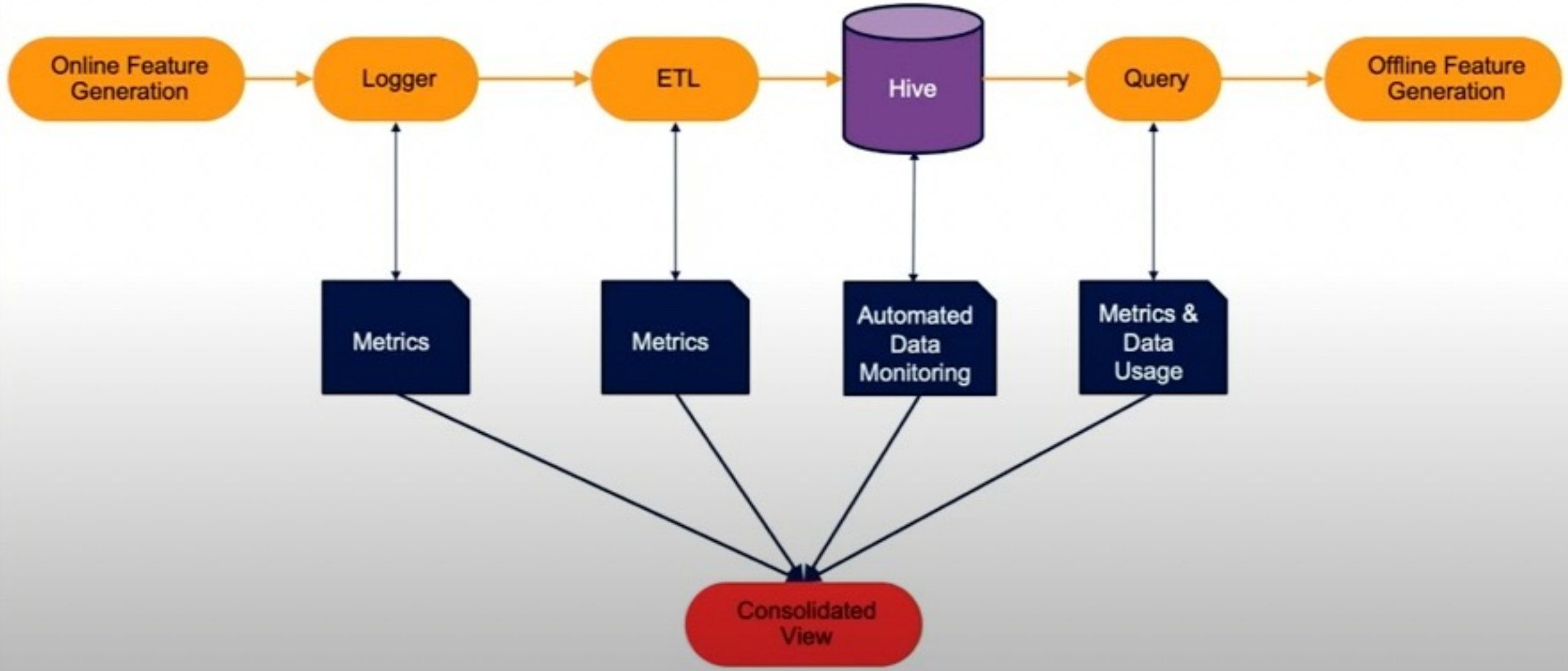
↳ Robust PCA (repeatedly calculate SVD and apply thresholds)

APPLIED ROOTS

Source: Netflix.com

Plain old visualizations

Data Quality Checkpoints



APPLIED ROOTS

Next Session : Amazon's Data Quality System