Yuqing Yao

Data Quality Assurance in Credit Card Fraud Prevention

PACIFIC NW SOFTWARE QUALITY CONFERENCE

THE FUTURE IS NOW

PNSQC.ORG

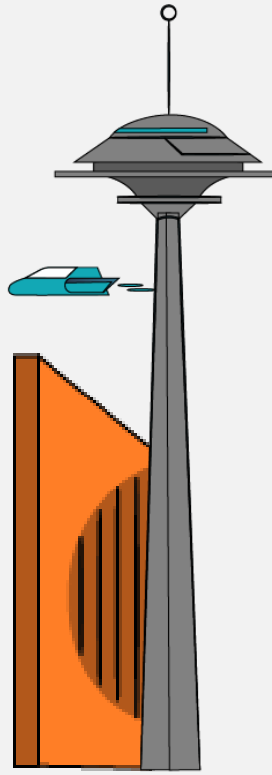
OCTOBER 14-16 2024

Meet the author

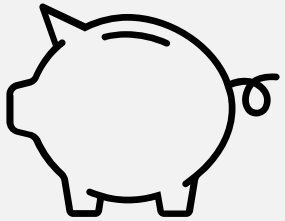
- Senior **Fraud Analyst** at Klarna
- Passionate about **data quality**
- **Hobbies**
  - organize AI/ML developer meetups
  - boardgames
  - Toastmasters

Outline

- **Why** should we care about **data quality** in Credit Card Fraud Prevention?
- **What** data quality **problems** do we encounter?
- **How** shall we **solve** the problems?



Why should we care about data quality in Credit Card Fraud Prevention?



Credit Card Fraud Prevention is important

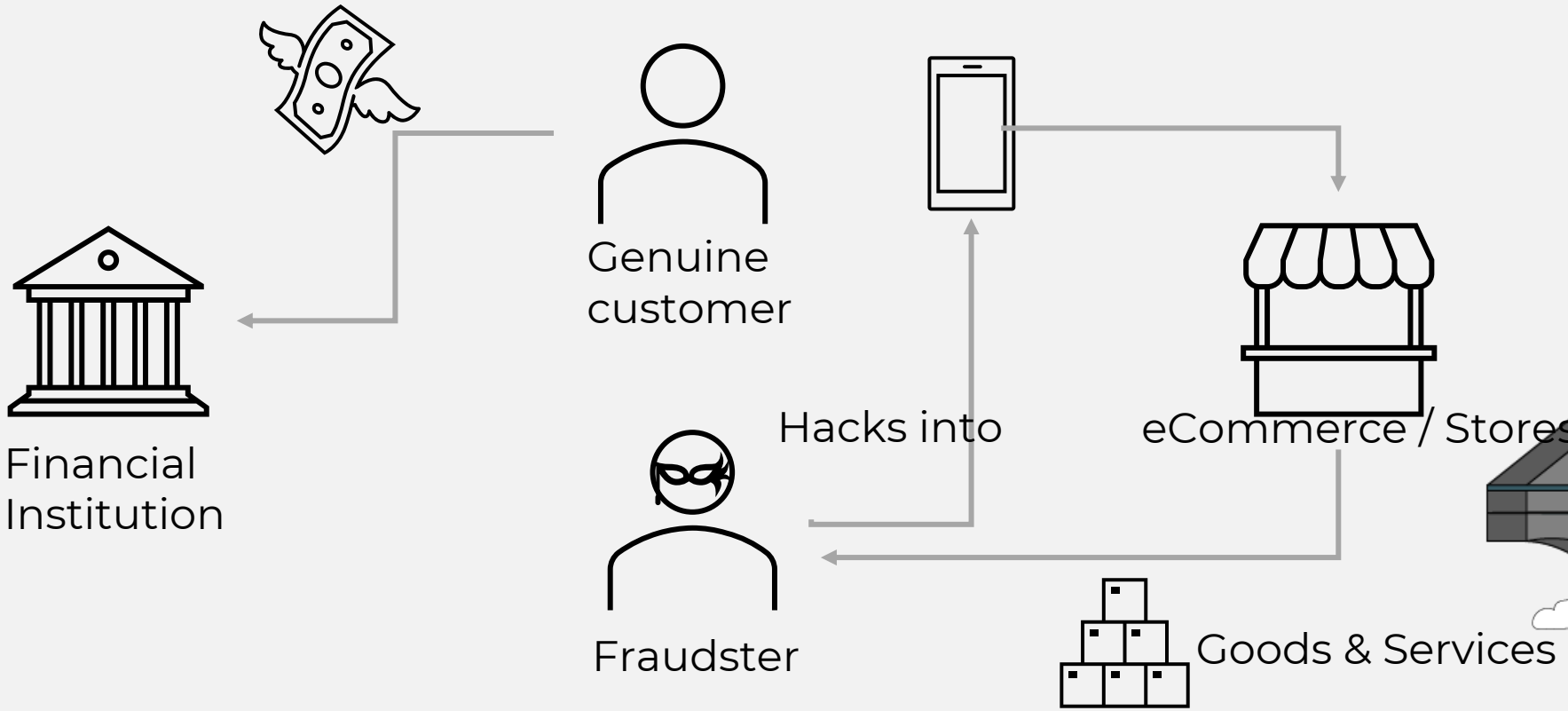
- **$10 billion** in consumer losses from credit card frauds in the US ~ **150k US households**' annual expenses

This domain heavily relies on data

- Behavioral analysis
- Fraud attacks
- Investigations

Yuqing Yao

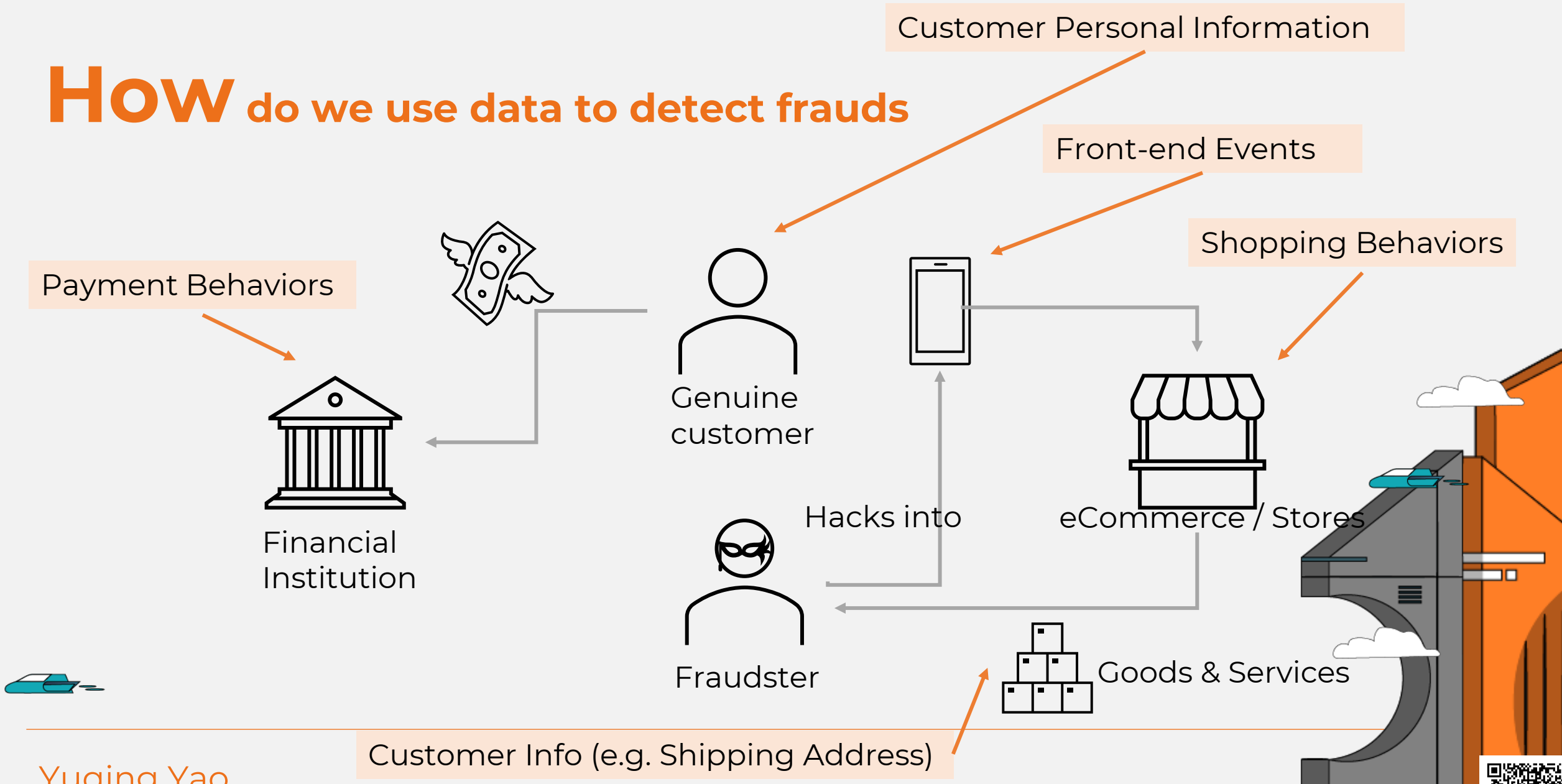
How are frauds conducted



Yuqing Yao

Data Quality Assurance in Credit Card Fraud Prevention

How do we use data to detect frauds



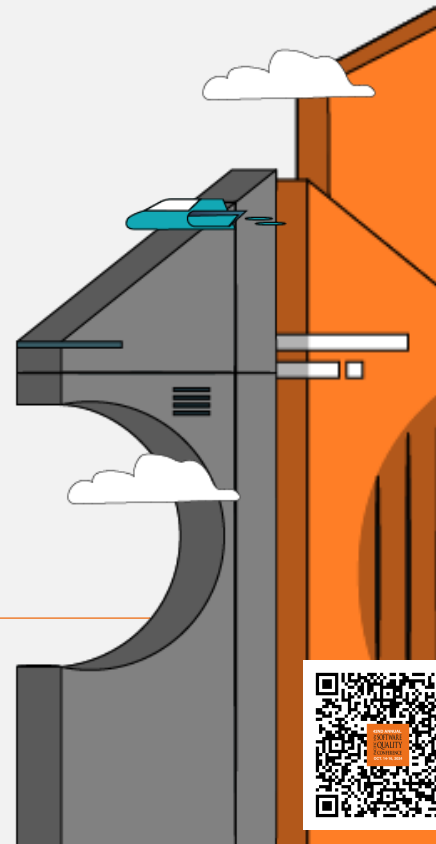
Yuqing Yao

Data Quality Assurance in Credit Card Fraud Prevention

What data quality problems do we encounter?

- **Types** of data used
  - Customer Personal Information
  - Transactions
  - Front-end Events
  - Behaviors
  - External Data Sources
  - Disputes

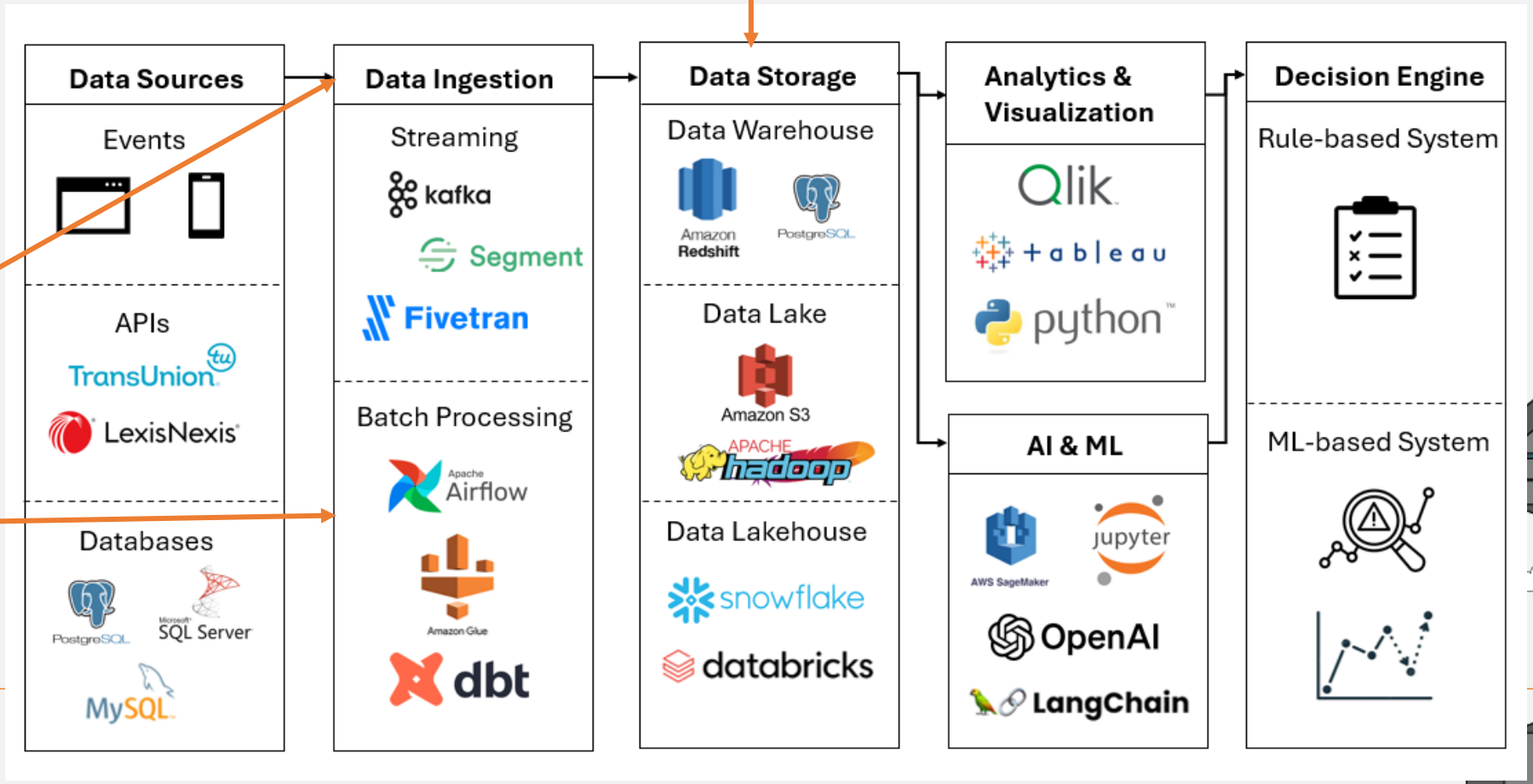
By Nature of Data
• Imbalanced Data
• Wrong Labels
• Incomplete Data



What data quality problems do we encounter?

Data Silos

- **Data Stack** – Procedures in Credit Card Fraud Prevention

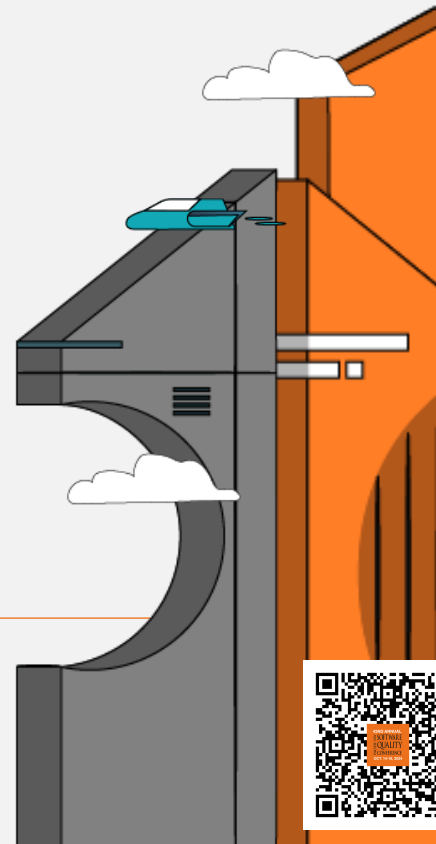


Latency

Wrong Calculations

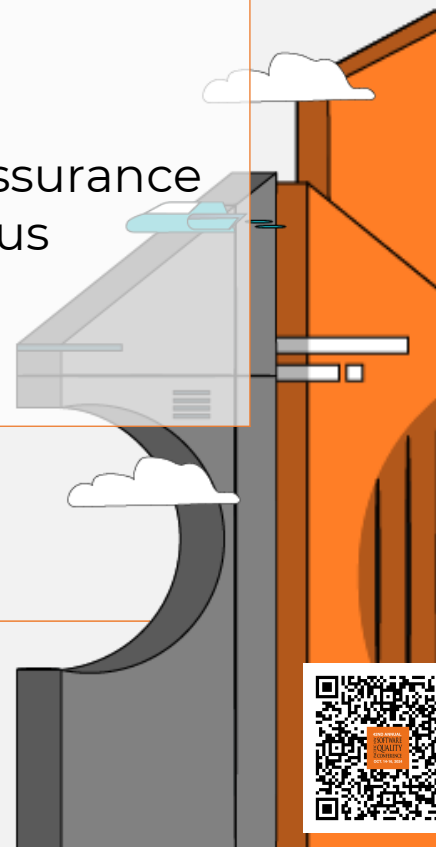
What data quality problems do we encounter?

- By Nature of the Task
  - Afterthought of the Business
  - Data Quality Assurance could be Tedious



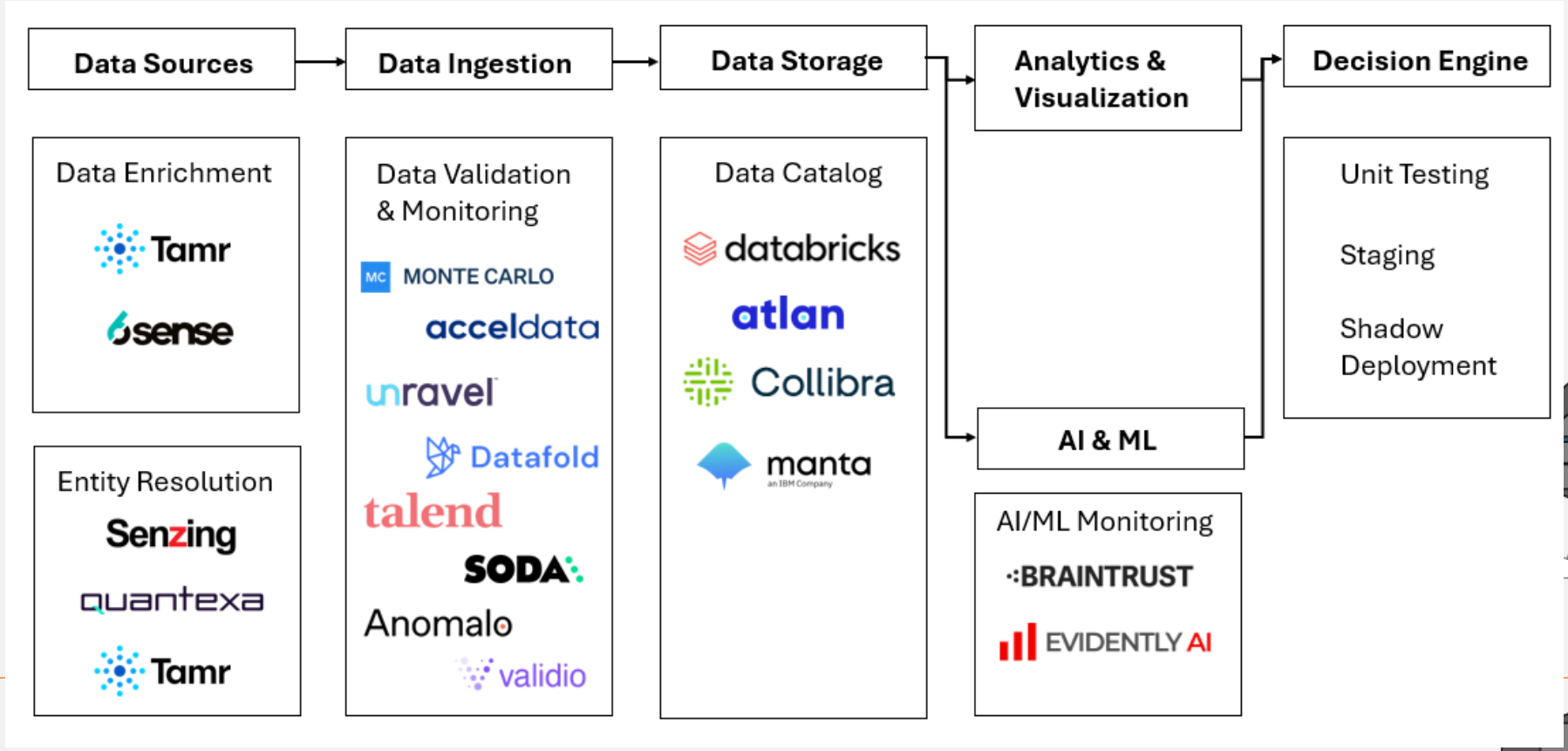
Recap of data quality problems

By Nature of Data	By Procedure	By Nature of the Task
• Imbalanced Data	• Latency	• Afterthought of the Business
• Wrong Labels	• Wrong Calculations	• Data Quality Assurance could be Tedious
• Incomplete Data	• Data Silos	



How shall we solve the problems?

- Solutions / Tools



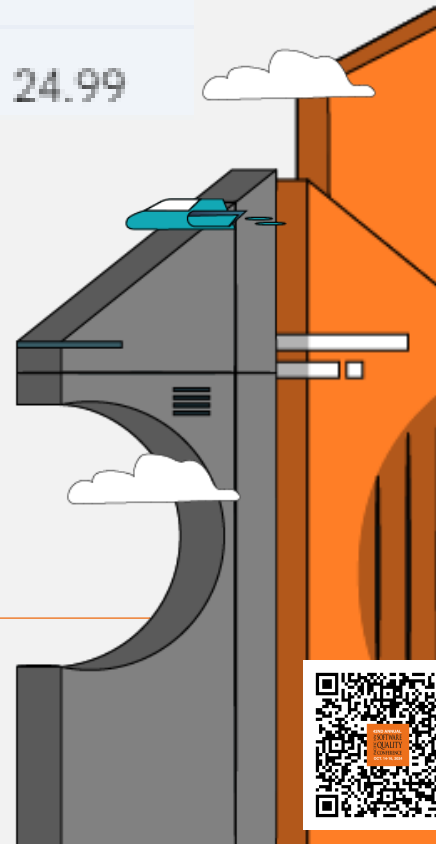
Yuqing Yao

Data Quality Assurance in Credit Card Fraud Prevention

Example 1 Wrong labels

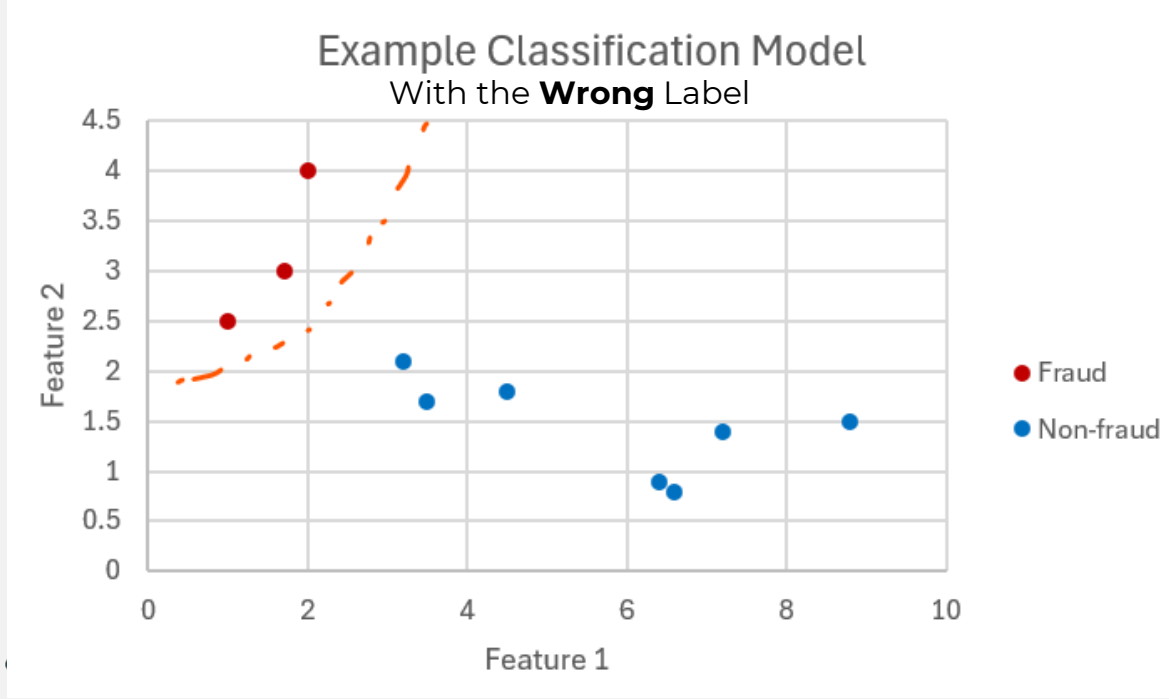
- What is the problem?

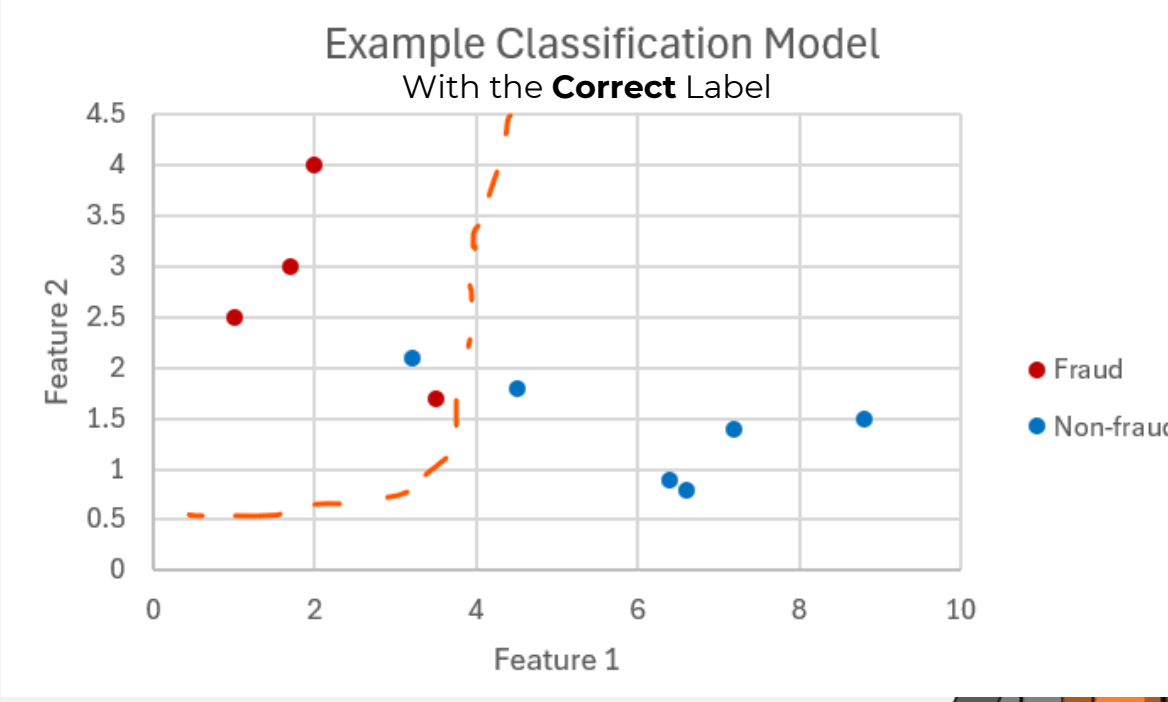
customer	transaction_date	dispute_date	dispute_reason	amount
John Smith	2024-01-01	2024-01-03	NULL	24.99



Example 1 Wrong labels

- Why is it a problem
  - Misleads supervised learning or statistical analysis

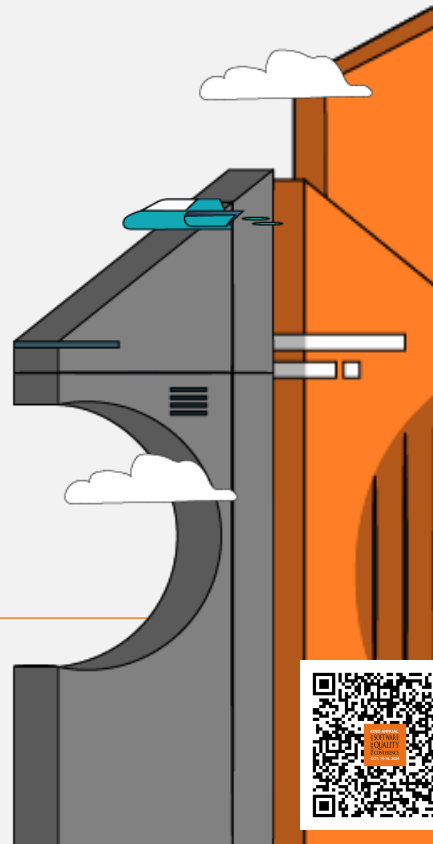




Example 1 Wrong labels

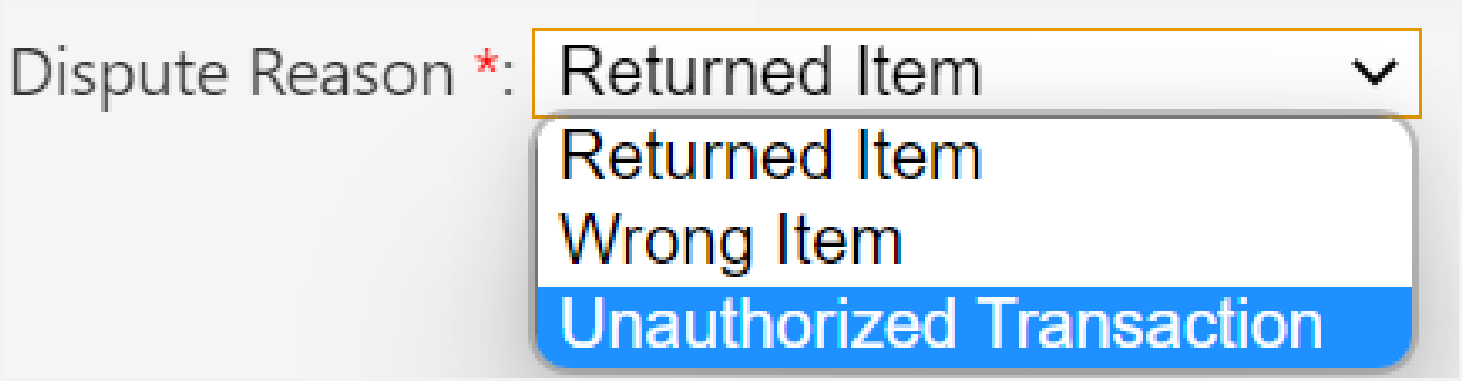
- What caused the problem?
  - No data validation during entry

Dispute Reason:



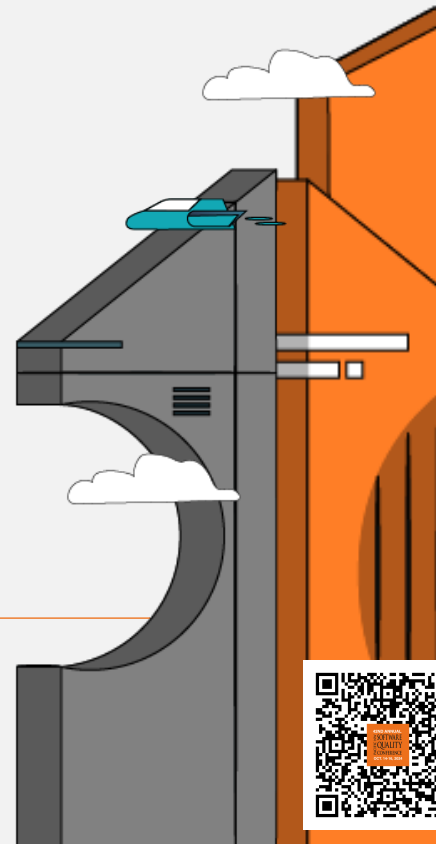
Example 1 Wrong labels

- What's the solution
  - Design the workflow to enforce data entry
  - Data validation during entry



Example 1 Wrong labels

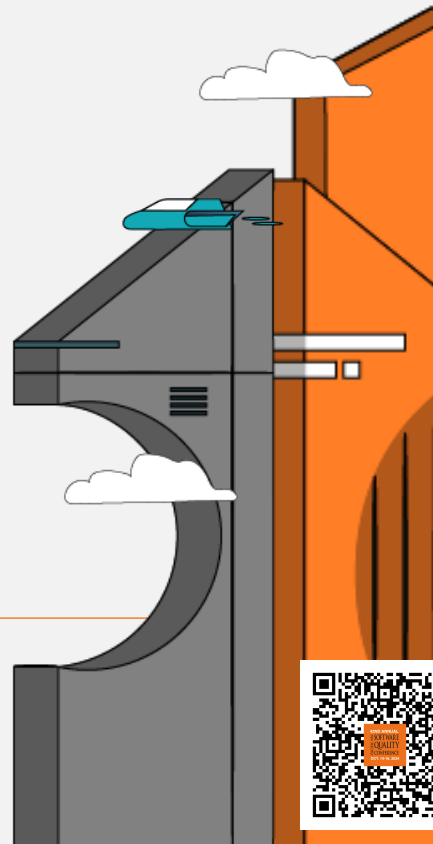
- What's result
  - Higher rate of data being populated
  - More standardized values



Example 2 Abnormal values

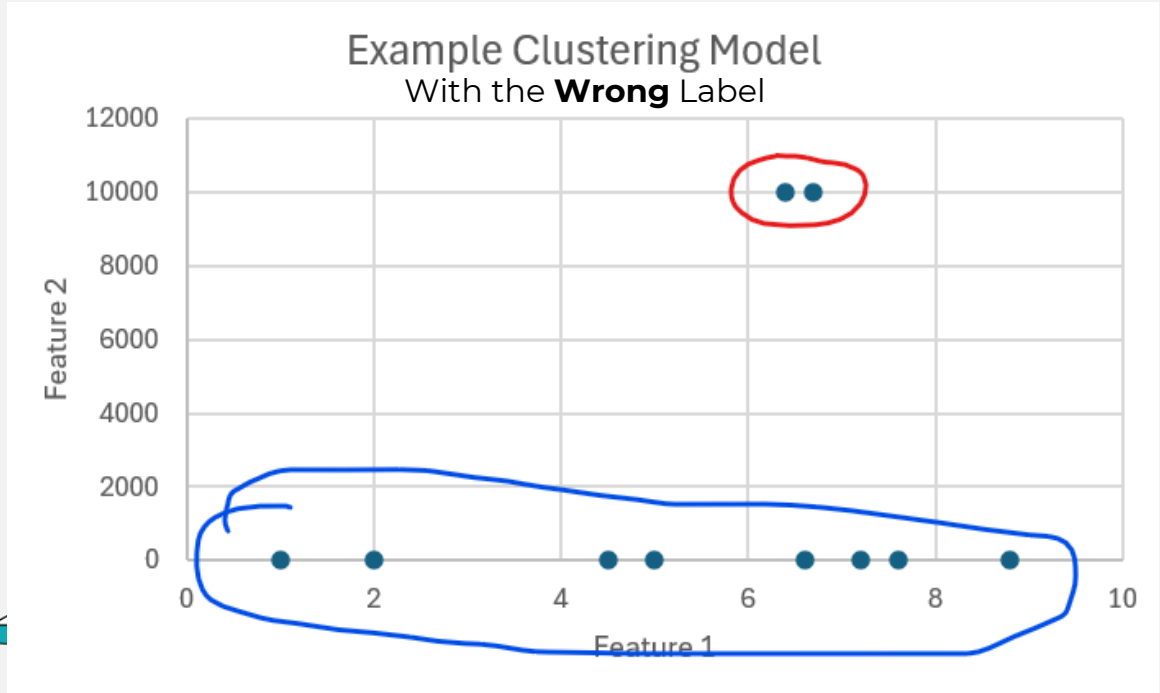
- What is the problem?
  - Certain attribute suddenly has more occurrences of extreme values (e.g. 9999)

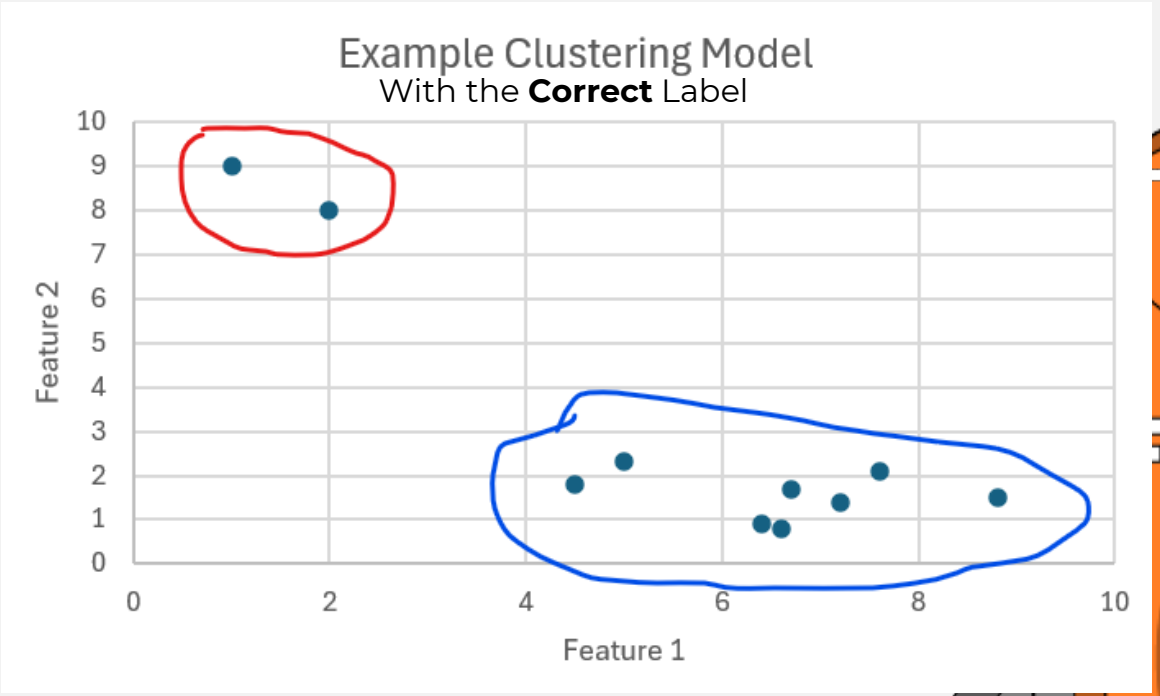
Feature 1	Feature 2
2	8
1	9
5	2.3
7.6	2.1
6.7	9999
4.5	1.8
6.4	9999
6.6	0.8
7.2	1.4
8.8	1.5



Example 2 Abnormal values

- Why is it a problem
  - Misleads the unsupervised learning models



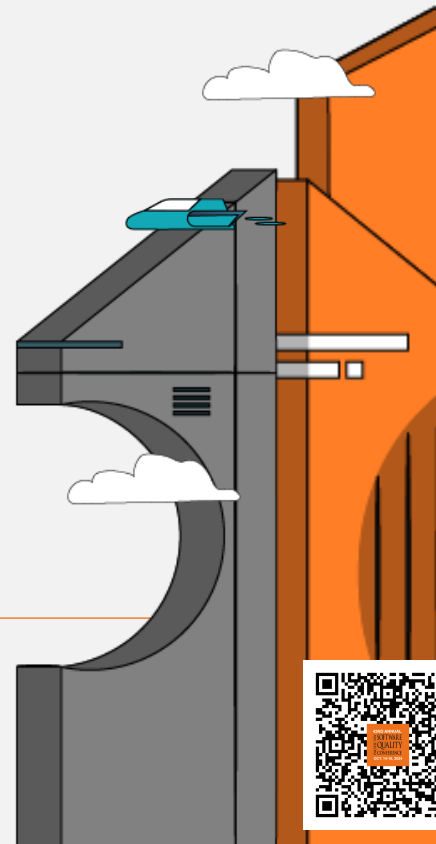


Yuqing Yao

Data Quality Assurance in Credit Card Fraud Prevention

Example 2 Abnormal values

- What caused the problem?
  - Vendor changed its missing value definition
  - Upstream team changed how they fill in missing values

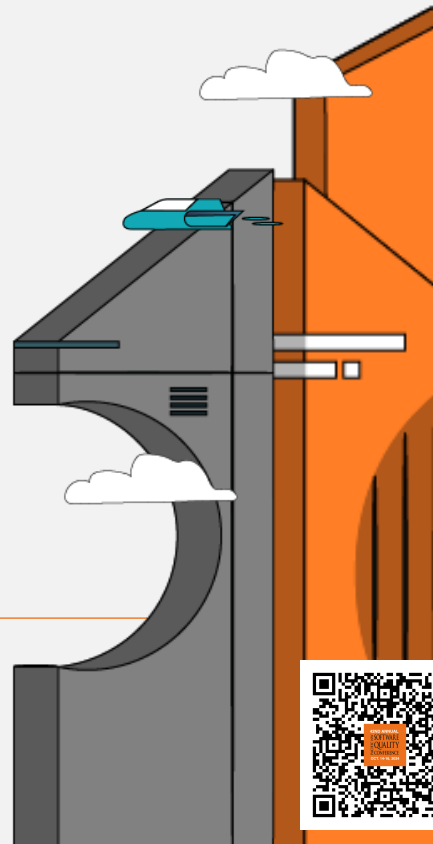


Example 2 Abnormal values

- What's the solution
  - Set up monitoring based on business logic and corresponding alerts

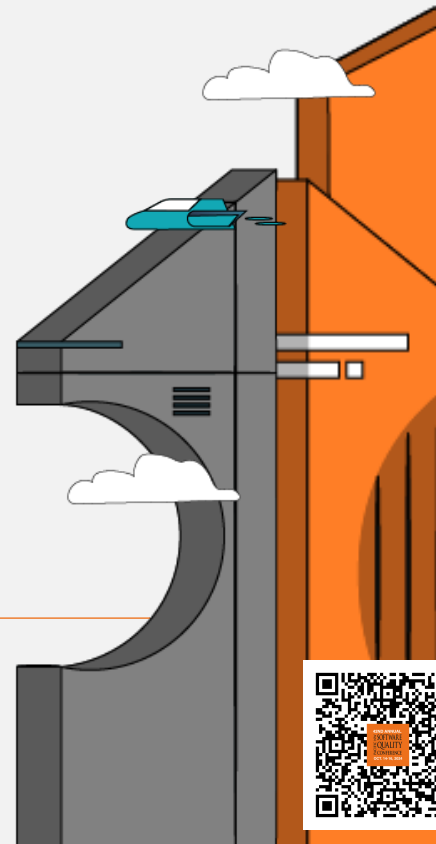
Definition Name	Condition	Action
Column A	X% values over 2 SD (last 1 month)	Alert

Alert
Warning
Recover



Example 2 Abnormal values

- What’s result
  - Discover data errors before they contaminates the model training
  - Fix the problems in time



Example 3 Disconnected data

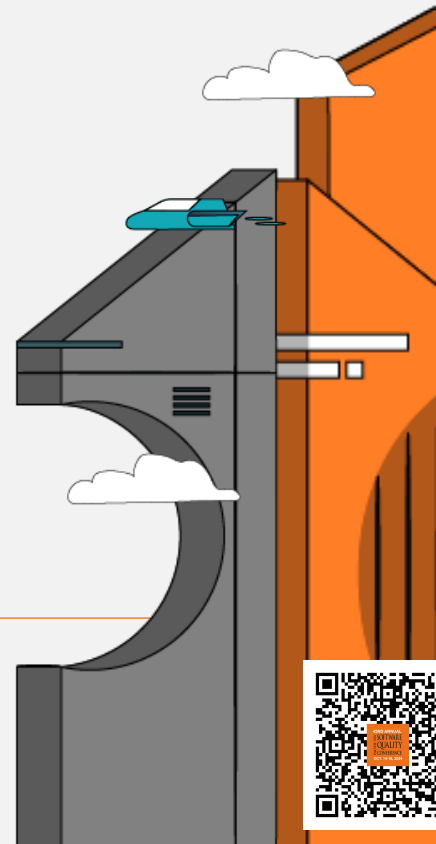
- What is the problem?
  - Customer created different accounts using different set of contact info

Name	John Smith
Email	js95@gmail.com
Phone	+14159999999
Address	123 ABC St, San Francisco, CA

Name	John K.P. Smith
Email	johnisawesome@gmail.com
Phone	+14159999999
Address	123 ABC St, San Francisco, CA

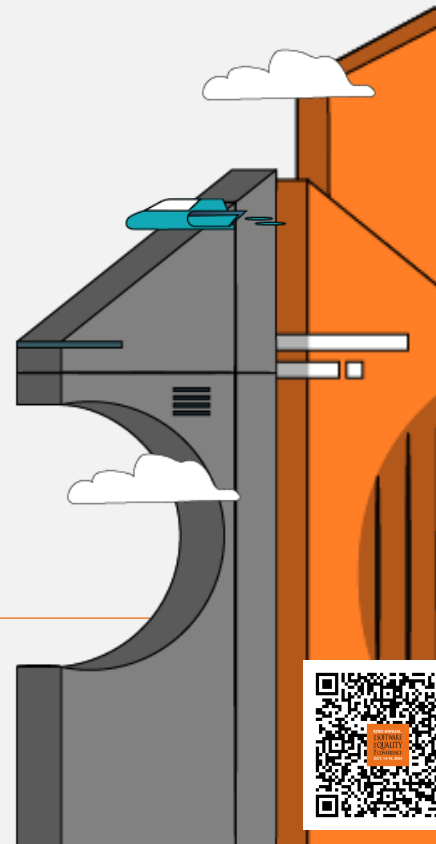
Example 3 Disconnected data

- Why is it a problem
  - These records are physically the same person, but digitally treated as different people
    - Mistakenly upsample or downsample frauds
    - Miss out fraud rings



Example 3 Disconnected data

- What caused the problem?
  - Data collected from different sources (Marketing / Log-in / vendors)
  - How records are created naturally cause this problem



Example 3 Disconnected data

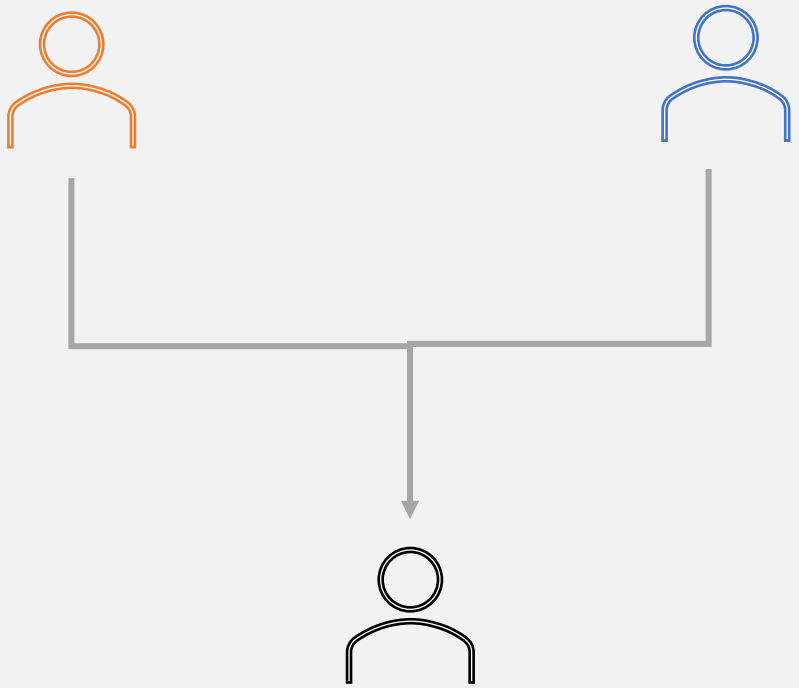
- What's the solution
  - Entity resolution

John Smith

js95@gmail.com

+14159999999

123 ABC St, San Francisco, CA



John K.P. Smith

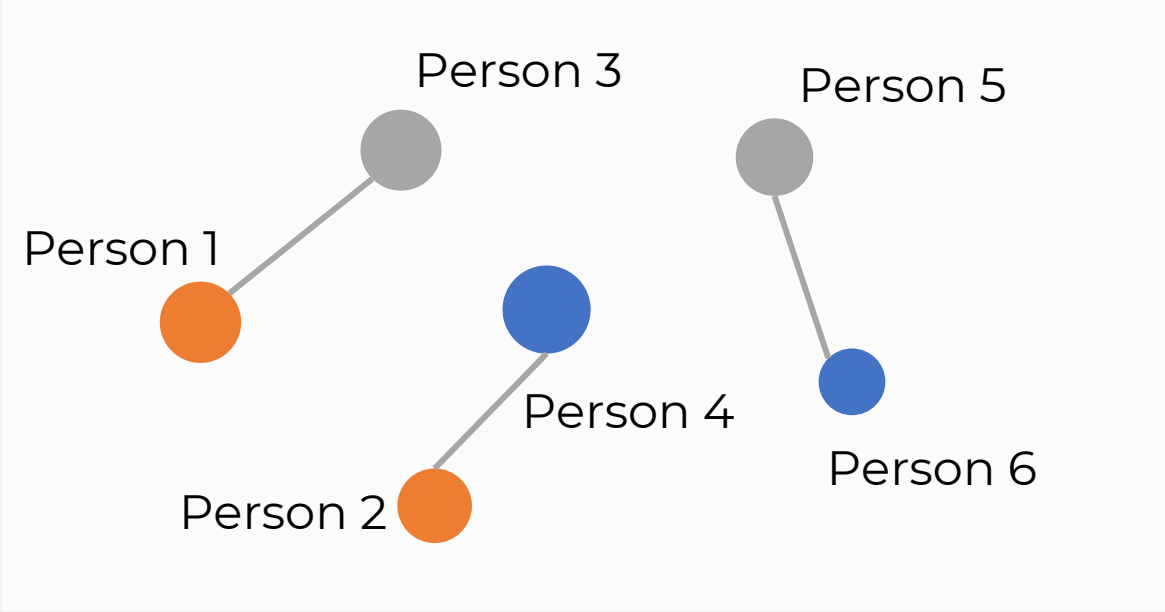
johnisawesome@gmail.com

+14159999999

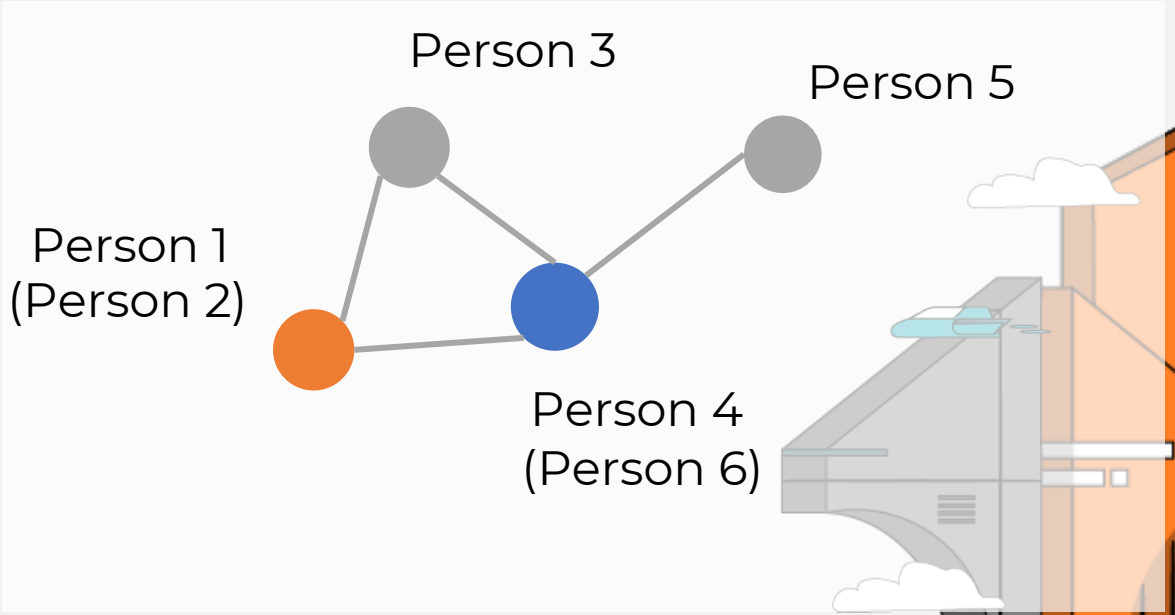
123 ABC St, San Francisco, CA

Example 3 Disconnected data

- What's the result
  - Discover new fraud rings and unreported frauds



Before Entity Resolution

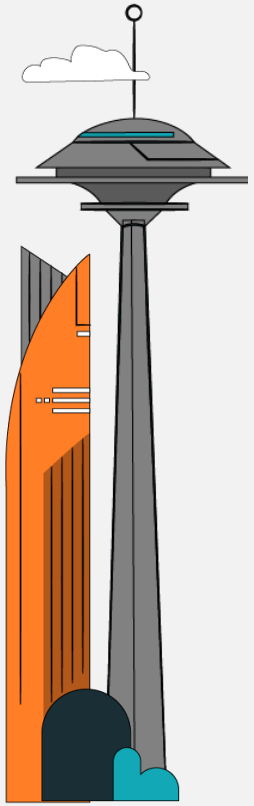


After Entity Resolution

Recap of examples

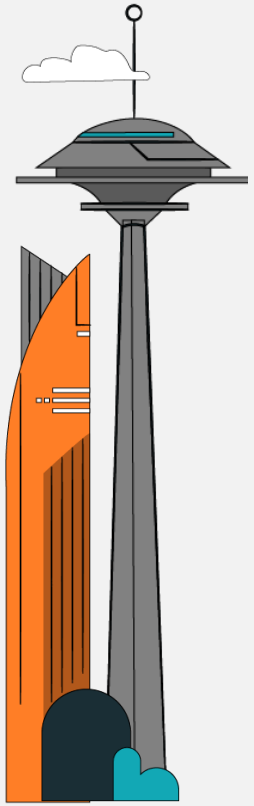
	Wrong labels	Abnormal values	Disconnected data
Impact	Misleads supervised learning or statistical analysis	Misleads the unsupervised learning	Mistakenly upsample or downsample frauds; Miss out fraud rings
Cause	No data validation during entry	Upstream processes changed filler values	Data collected from different sources
Solution	Data validation + Better UI	Set up monitoring based on business logic and corresponding alerts	Entity resolution
Result	More standardized values	Discover and fix data errors	Discover new fraud rings and unreported frauds

Conclusion and Future Improvements



- **Conclusion**
  - Data quality is **important** in Credit Card Fraud Prevention because each piece of missing info contributes to solving the case
  - Data issues can be **caused by** the nature of the data, procedures involved, and the nature of the task
  - **Solutions** of the data issues lie in using good tools and controls during each step of conducting Credit Card Fraud Prevention, including data validation, entity resolution, data enrichment, data monitoring, and data catalogs

Conclusion and Future Improvements



- **Future Improvements**
  - Use real-world **case studies** of FinTech companies to show the impact of applying the best practices of data quality assurance, and how that prevents losses
  - Discuss **how AI is used** in Credit Card Fraud Prevention and how **AI observability** should be conducted in this field

42ND ANNUAL
PACIFIC NW SOFTWARE QUALITY CONFERENCE
OCT. 14-16, 2024

THE FUTURE IS NOW

PNSQC.ORG

OCTOBER 14-16 2024