

# NYU FRE 7773 - Week 5

---

*Machine Learning in Financial Engineering*

Ethan Rosenthal

# Case Study: Fraud Detection

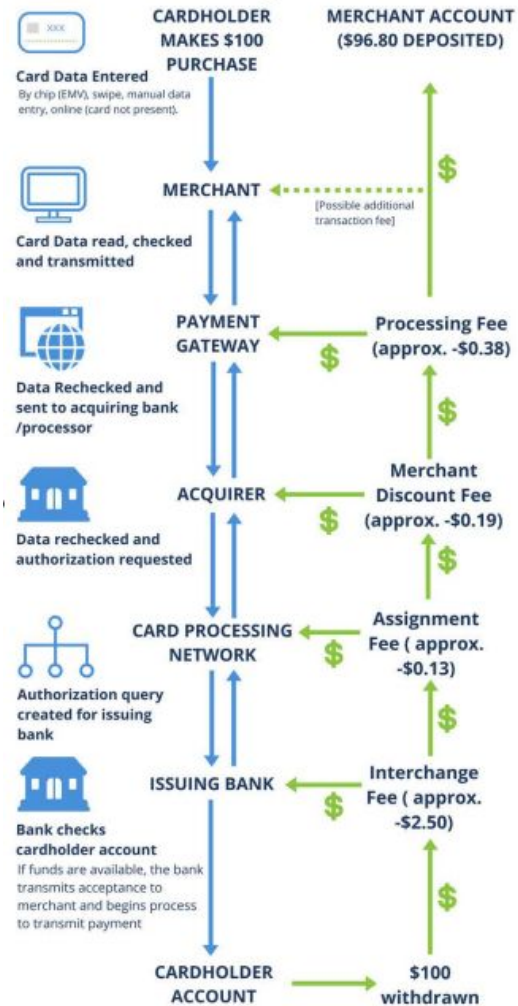
---

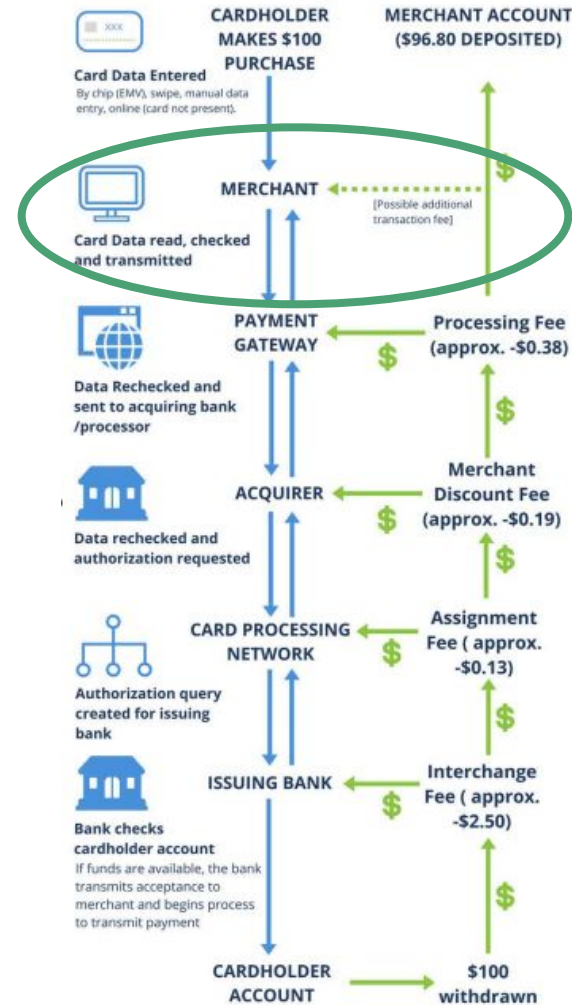
*Machine Learning in Financial Engineering*  
Ethan Rosenthal

Risk

---







## Square Launches On-Demand and Instant Payments for Square Payroll Customers



Square introduced two new features this week to help its customers using Square Payroll – as well as their employees – manage their cash flow.

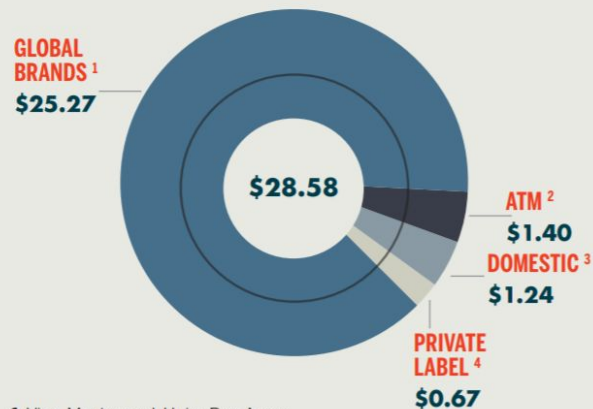
Fraud

---



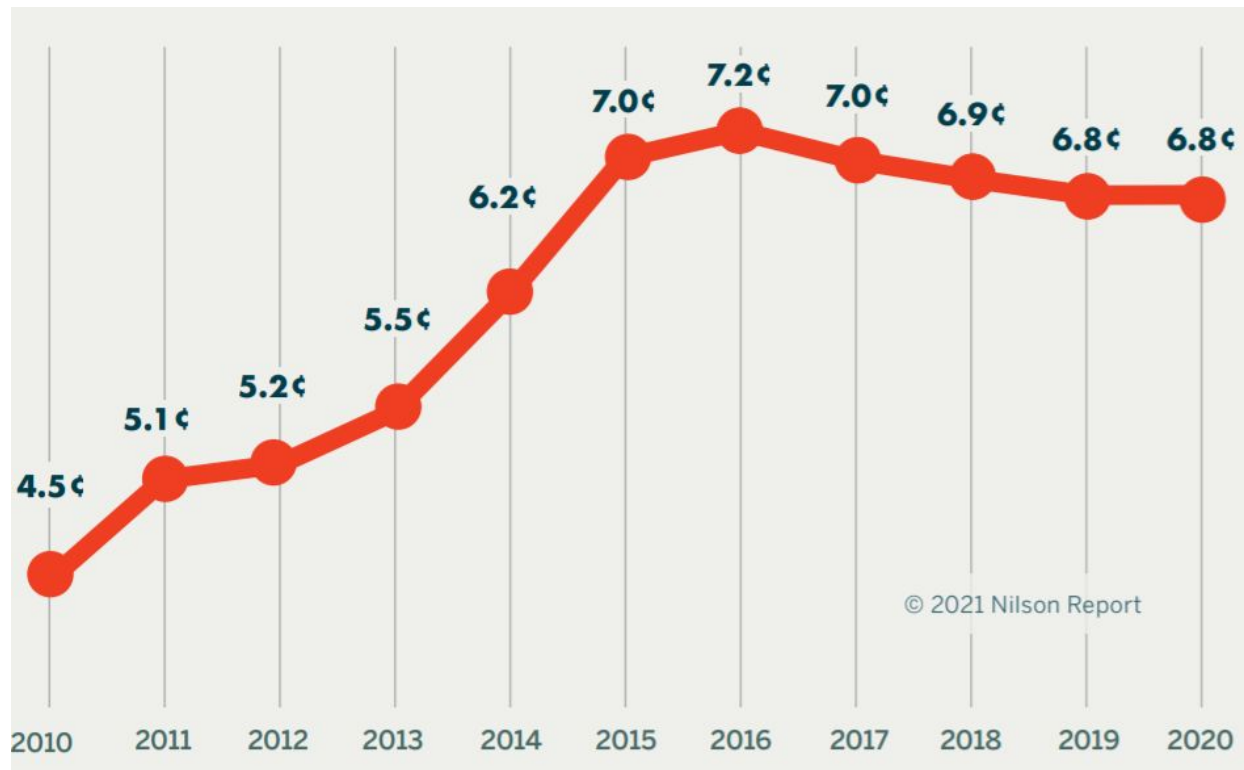
## Fraud by Type of Card

Billions in 2020



**1** Visa, Mastercard, UnionPay, Amex, Diners/Discover, JCB. **2** From transactions processed outside of global networks. **3** Elo, RuPay, Interac, Cartes Bancaire, Mir and 88 others. **4** Includes store, gasoline, airlines, medical, ACH debit, prepaid etc.

©2021 The Nilson Report



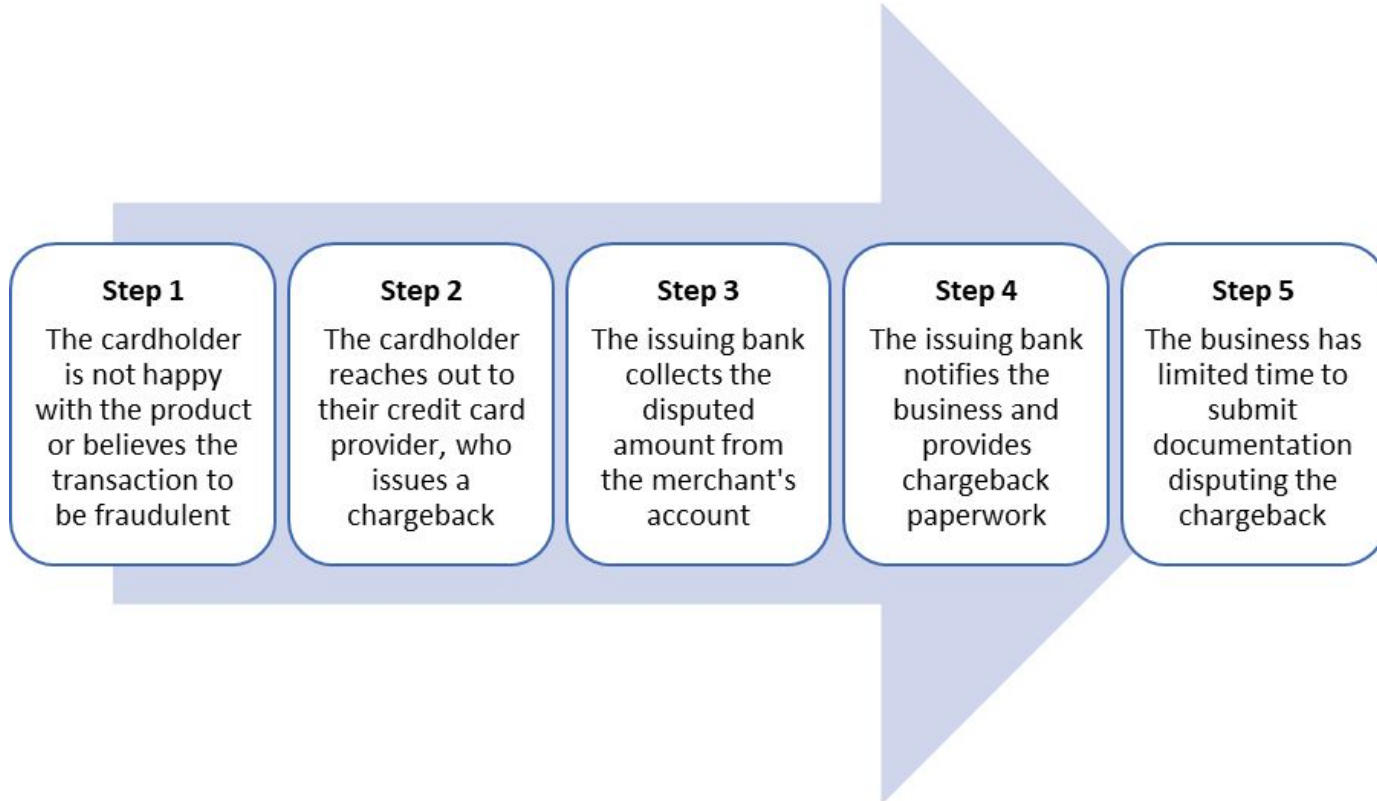
© 2021 Nilson Report

## Card Fraud Projected through 2030

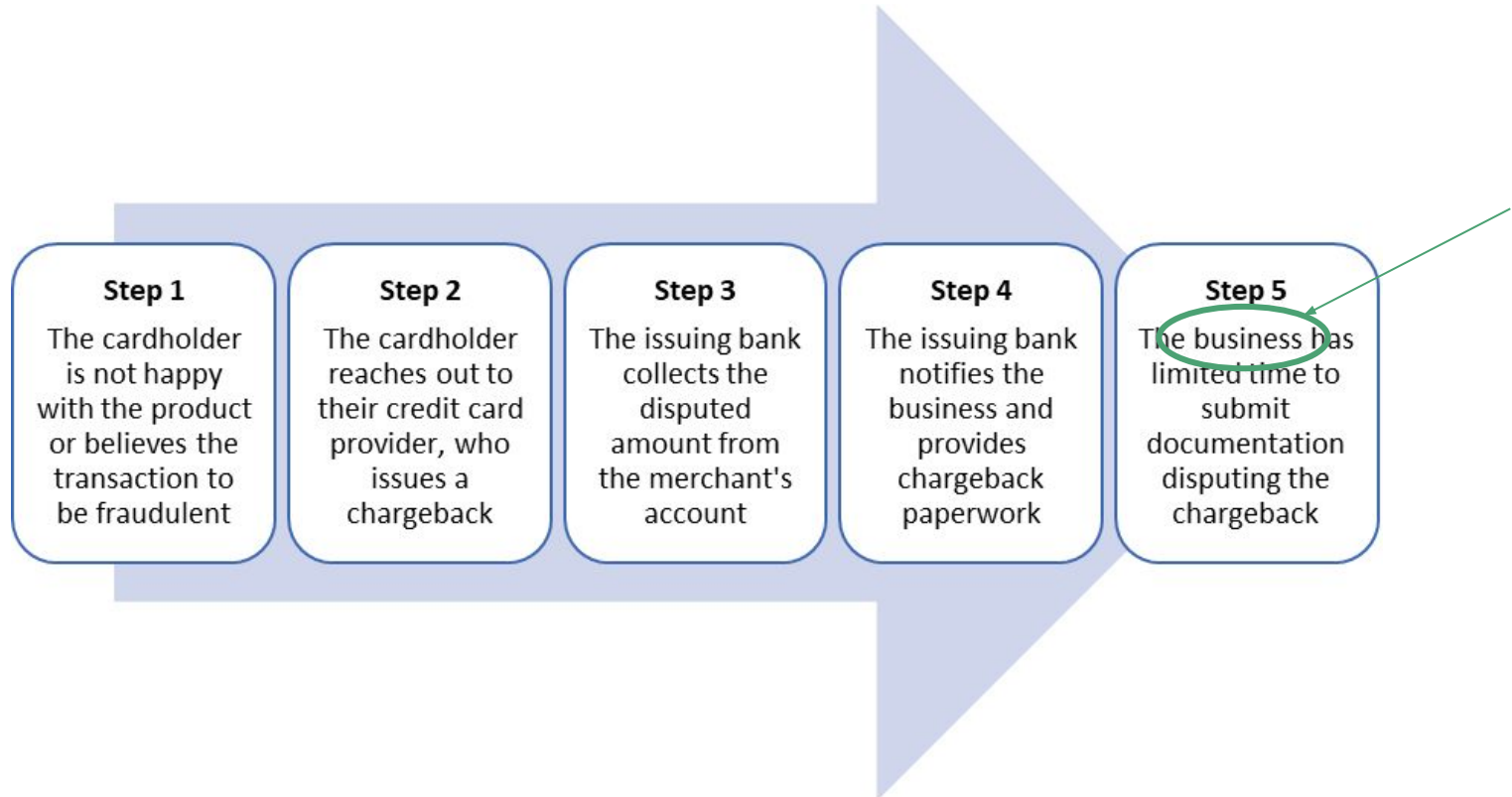
|      | Total Volume | Fraud   | Cents per    |
|------|--------------|---------|--------------|
| YEAR | (TRIL.)      | (BIL.)  | \$100 VOLUME |
| 2020 | \$41.962     | \$28.58 | 6.81         |
| 2021 | \$47.229     | \$32.20 | 6.82         |
| 2022 | \$50.868     | \$34.36 | 6.75         |
| 2023 | \$54.061     | \$36.13 | 6.68         |
| 2024 | \$57.323     | \$38.07 | 6.64         |
| 2025 | \$60.583     | \$39.89 | 6.58         |
| 2026 | \$64.038     | \$41.73 | 6.52         |
| 2027 | \$67.570     | \$43.76 | 6.48         |
| 2028 | \$71.221     | \$45.54 | 6.39         |
| 2029 | \$75.111     | \$47.50 | 6.32         |
| 2030 | \$79.140     | \$49.32 | 6.23         |

© 2021 Nilson Report

# Chargebacks



# Chargebacks

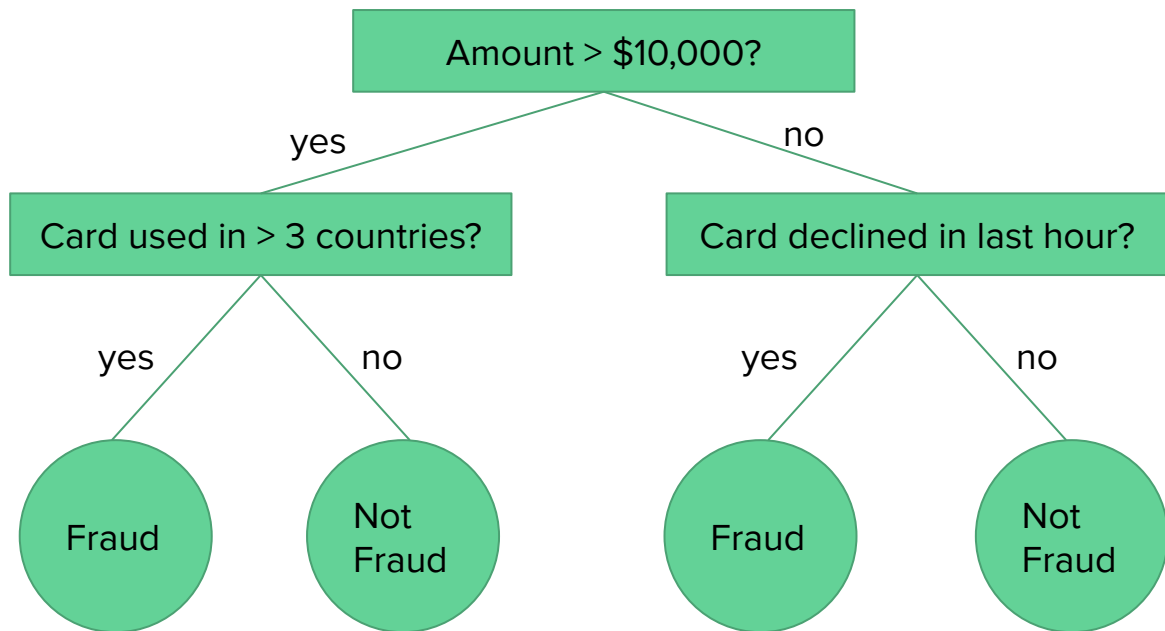


# Fraud Detection

---

# Rules

- A manual decision tree
- Ideally, high precision, low recall
- Quick and dirty.
- Lots of manual work and tuning.

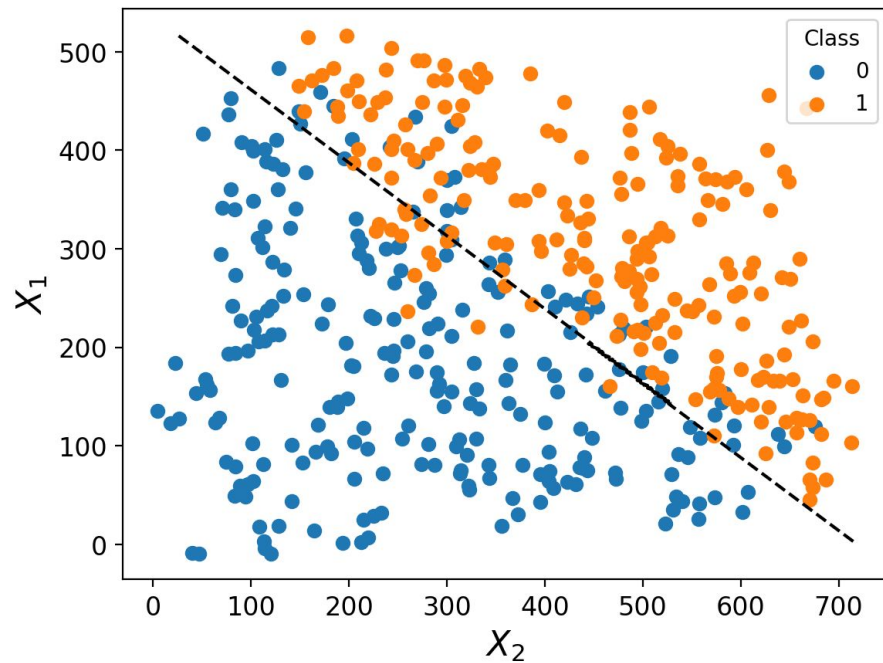


# Machine Learning Models

- Binary classification: fraud vs. not fraud
- Sample -> often an **event** (e.g. payment transaction, bank withdrawal).
- **X** -> event features + auxiliary features
- **y** -> was event “associated” with fraud or not
- Train model on historical events, predict on real time events.

# Machine Learning Models

- Logistic regression is simple and can be extremely fast for low latency applications.

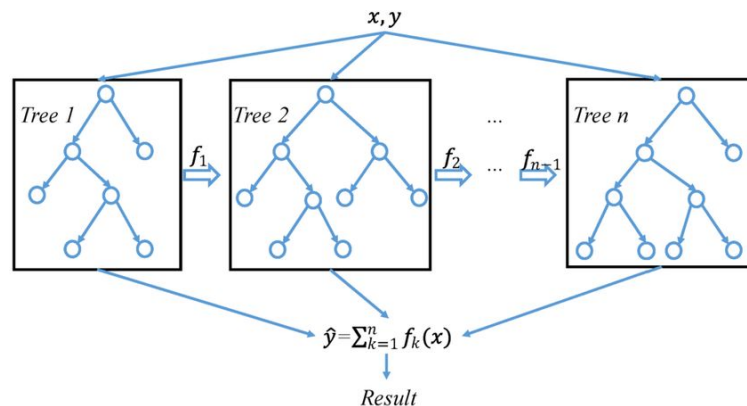




# Machine Learning Models

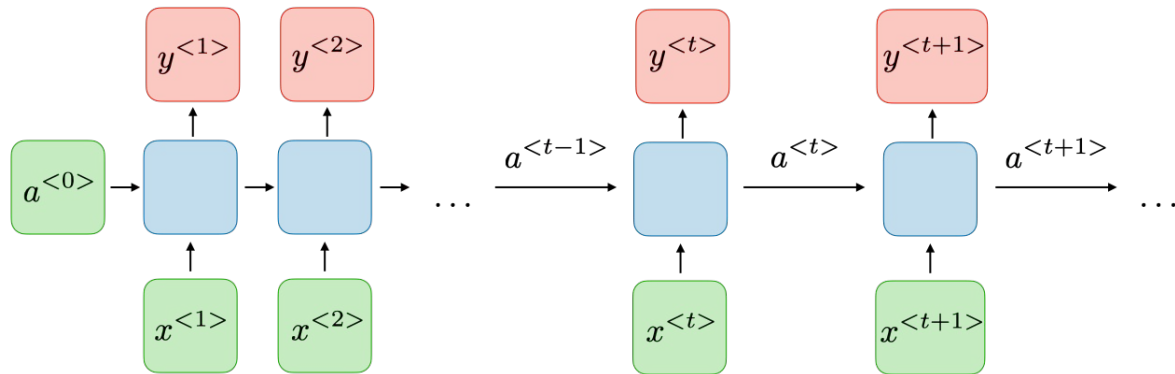
- Tree-based models work particularly well.
- Domain naturally lends itself to if/else statements.
- Easy to tune.
- Relatively fast to train and predict.
- Somewhat interpretable.

*dmlc*  
**XGBoost**



# Machine Learning Models

- Deep learning models can handle sequences of events.
- Much trickier to model and tune.



TensorFlow

 PyTorch

# Fraud Detection Difficulties

---

KEANU REEVES DENNIS HOPPER SANDRA BULLOCK

GET  
READY FOR  
RUSH  
HOUR.



**SPEED**

TWENTIETH CENTURY FOX PRESENTS A MARK GORDON PRODUCTION KEANU REEVES DENNIS HOPPER SANDRA BULLOCK "SPEED"  
JOE MORTON AND JEFF DANIELS EDITOR MARK MANOVNA PRODUCED BY JOHN WRIGHT, A.S.C. EXECUTIVE PRODUCERS JACKSON LACROIX PRODUCED BY ANDREJA BARTKOWIAK EXECUTIVE PRODUCERS JEFF DANIELS WRITTEN BY GRAHAM YOST  
DIRECTED BY MARK GORDON  
R RESTRICTED  
PARENTS STRONGLY CAUTIONED  
MPAA RATING  
JUNE 10  
20TH CENTURY FOX  
© 1994 TWENTIETH CENTURY FOX

# Speed – Realtime Features

- Payment Amount
- Transaction Country
- Credit Card Brand
- IP Address
- Tip Percentage



# Speed – Realtime Features

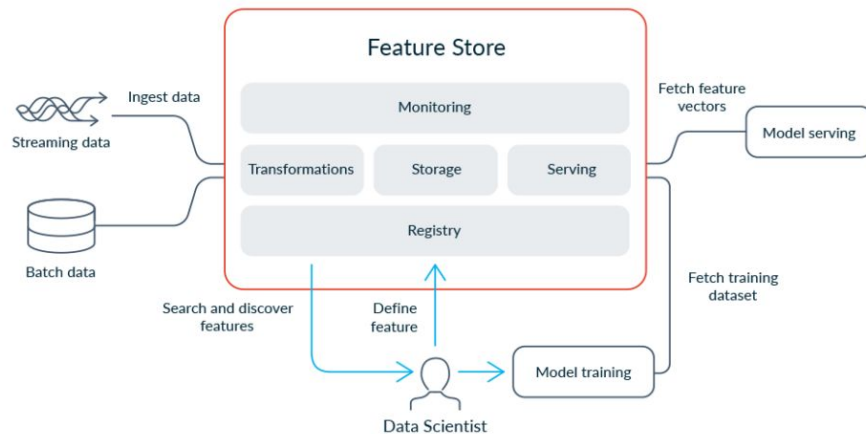
- Standard deviation of milliseconds between transaction attempts over the last 30 seconds.



# Speed – Realtime Features

Feature Stores are systems for ingesting streams of events and converting them into features, *quickly*.

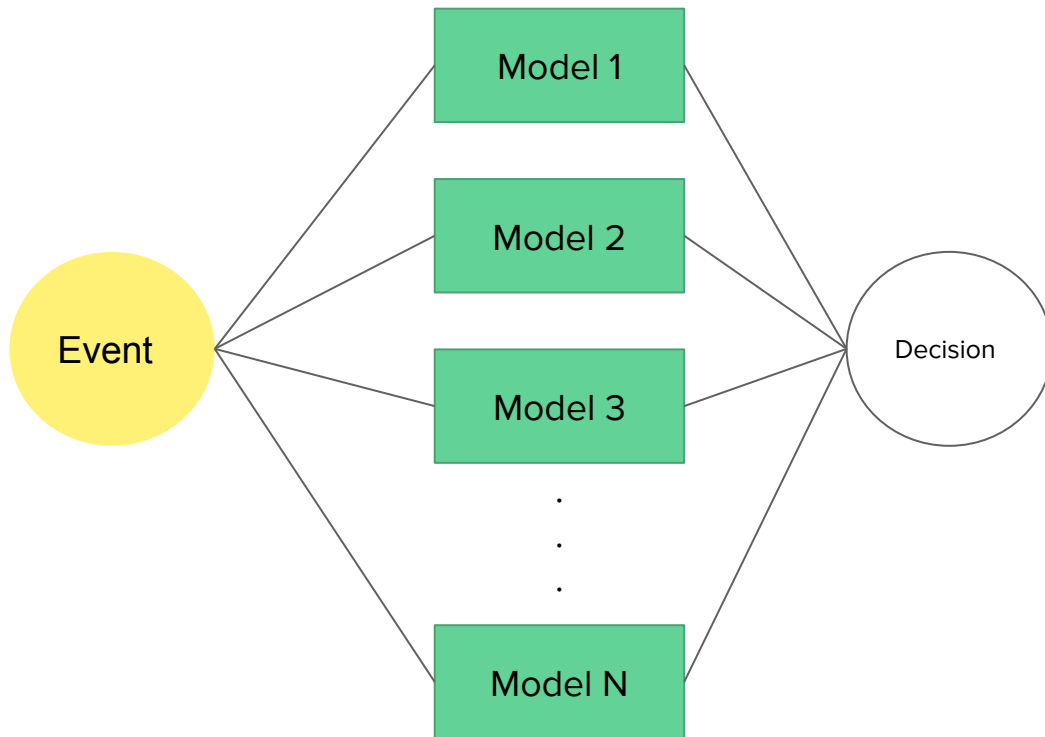
Tradeoff between feature *accuracy* and *latency*.



# tection

# Speed – Realtime Detection

- Thousands of rules and models for a given event.
- Limited time that each can take.
- Sync vs. async workflows.



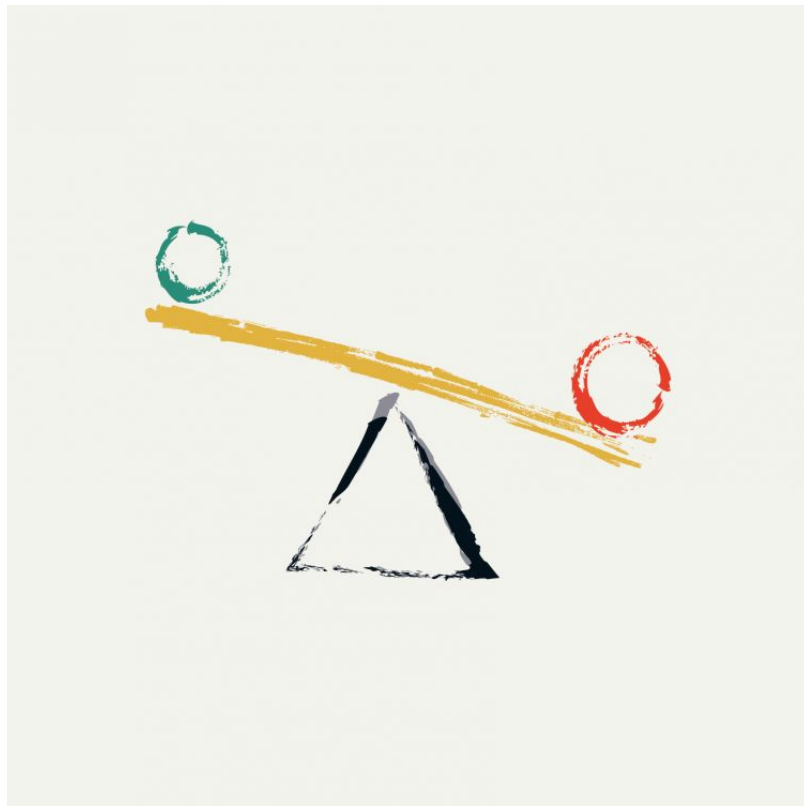


# Fraud Decisioning

---

# Tradeoffs

- Automation vs. Confidence
  - Block high fraud likelihood transactions.
  - Manually review when not sure.
- Friction vs. Risk
  - Verify via SMS vs. providing bank documents.
- The goal is to “cost more” than the alternative.
- **False Positives vs. False Negatives**





**lil du bois** 🌸  
@\_lildubois

...

well, well, well, if it isn't the consequences of my own actions

1:25 AM · Sep 29, 2018 · Twitter for Android

---

**133.7K** Retweets   **5,164** Quote Tweets   **297K** Likes

# Training Models

---

# Machine Learning Models

- Binary classification: fraud vs. not fraud
- Sample -> often an **event** (e.g. payment transaction, bank withdrawal).
- **X** -> event features + auxiliary features
- **y** -> was event “associated” with fraud or not
- Train model on historical events, predict on real time events.

# Machine Learning Models

- Binary classification: fraud vs. not fraud
- Sample -> often an **event** (e.g. payment transaction, bank withdrawal).
- **X -> event features + auxiliary features**
- **y** -> was event “associated” with fraud or not
- Train model on historical events, predict on real time events.

# Speed – ~~Realtime~~ Historical Features

- Standard deviation of milliseconds between transaction attempts over the last 30 seconds **at the time the Fraud model would make its prediction.**



# Speed – ~~Realtime~~ Historical Features

- Standard deviation of milliseconds between transaction attempts over the last 30 seconds **at the time the Fraud model would make its prediction.**
- **Don't leak the future into the past!**





# Speed – ~~Realtime~~ Historical Features

- Standard deviation of milliseconds between transaction attempts over the last 30 seconds **at the time the Fraud model would make its prediction.**
- **Don't leak the future into the past!**
- **Historical features must match realtime features (“train/test skew”)**



# Machine Learning Models

- Binary classification: fraud vs. not fraud
- Sample -> often an **event** (e.g. payment transaction, bank withdrawal).
- **X** -> event features + auxiliary features
- **y** -> was event “associated” with fraud or not
- Train model on historical events, predict on real time events.

# Life of Ground Truth

Fraudulent  
Payment  
Transaction

$y = 1$

Chargeback



Good  
Payment  
Transaction

$y = 0$



# Life of Ground Truth

Fraudulent  
Payment  
Transaction

$y = 1$

?????

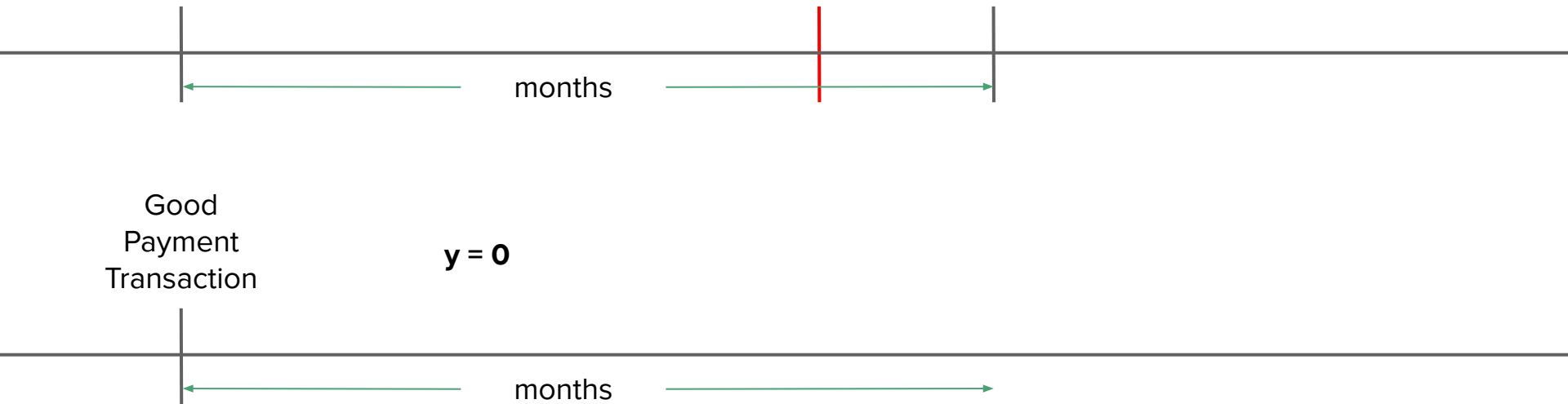
Chargeback

months

Good  
Payment  
Transaction

$y = 0$

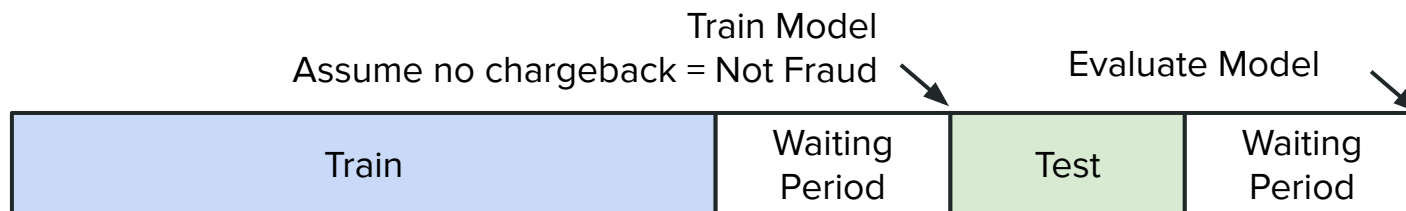
months

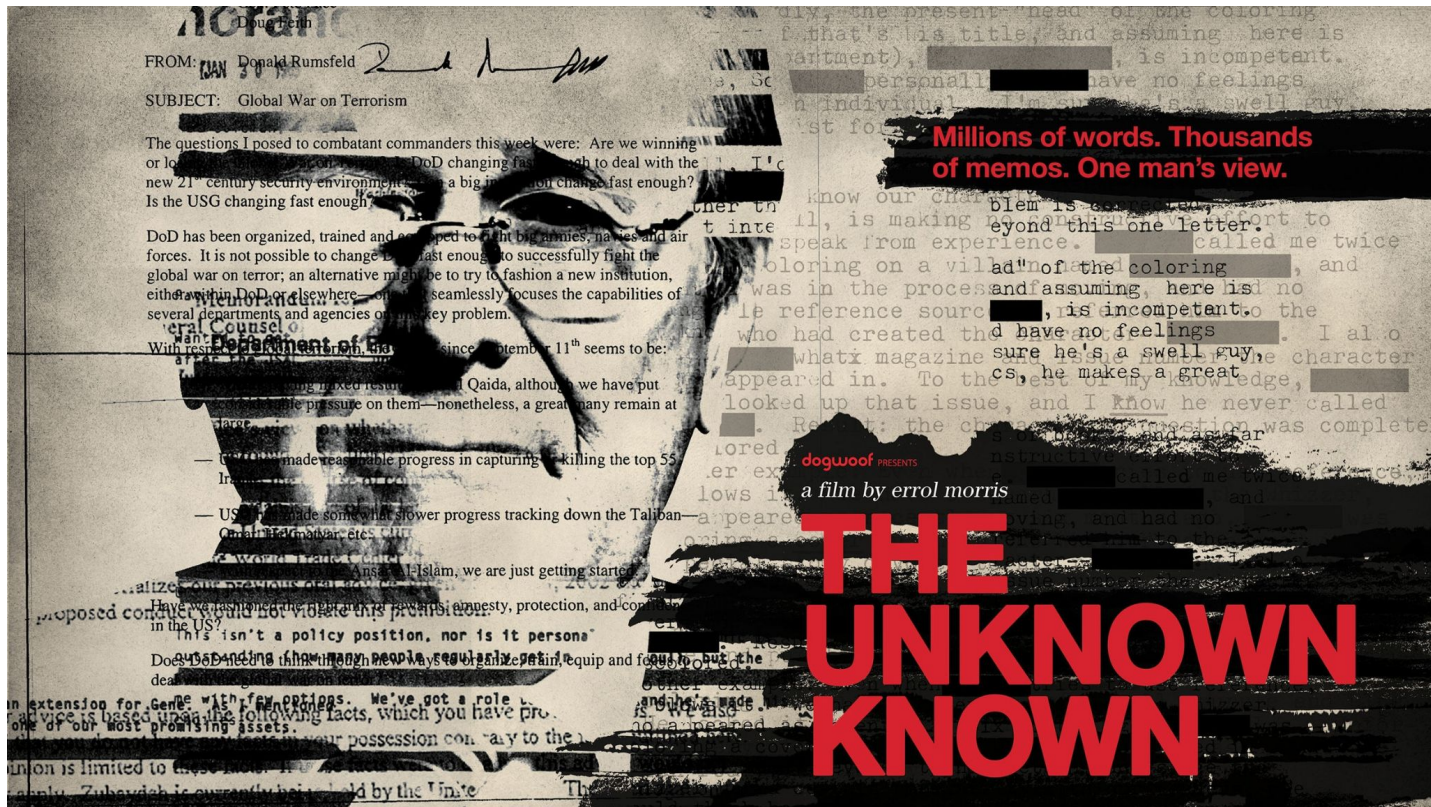


# Train/Test Split



# Train/Test Split





FROM: DONALD RUMSFELD

SUBJECT: Global War on Terrorism

The questions I posed to combatant commanders this week were: Are we winning or losing? Is the DoD changing fast enough to deal with the new 21st century security environment? Is a big institutional change fast enough? Is the USG changing fast enough?

DoD has been organized, trained and equipped to fight big armies, navies and air forces. It is not possible to change fast enough to successfully fight the global war on terror; an alternative might be to try to fashion a new institution, either within DoD or elsewhere, one that seamlessly focuses the capabilities of several departments and agencies on the key problem.

With respect to the war on terror, the DoD since September 11th seems to be:

- making mixed results on Al Qaeda, although we have put considerable pressure on them—nonetheless, a great many remain at large.
- made reasonable progress in capturing or killing the top 55 leaders of Al Qaeda.
- USG made somewhat slower progress tracking down the Taliban, Osama bin Laden, etc.

As we move forward with Islam, we are just getting started.

Have we tried to protect this against, honesty, protection, and control in the US?

This isn't a policy position, nor is it persona.

DoD has not done this in many years, and it is a key part of the war on terror.

in extension for Gen. We've got a role to play in the following facts, which you have provided.

one of our most promising assets. Your possession can pay to the

union is limited to these assets. It is not a key part of the war on terror.

currently, Zubovitch is currently being held by the White

Millions of words. Thousands of memos. One man's view.

dogwoof PRESENTS  
a film by errol morris

THE  
UNKNOWN  
KNOWN

# Life of Ground Truth

Fraudulent  
Payment  
Transaction

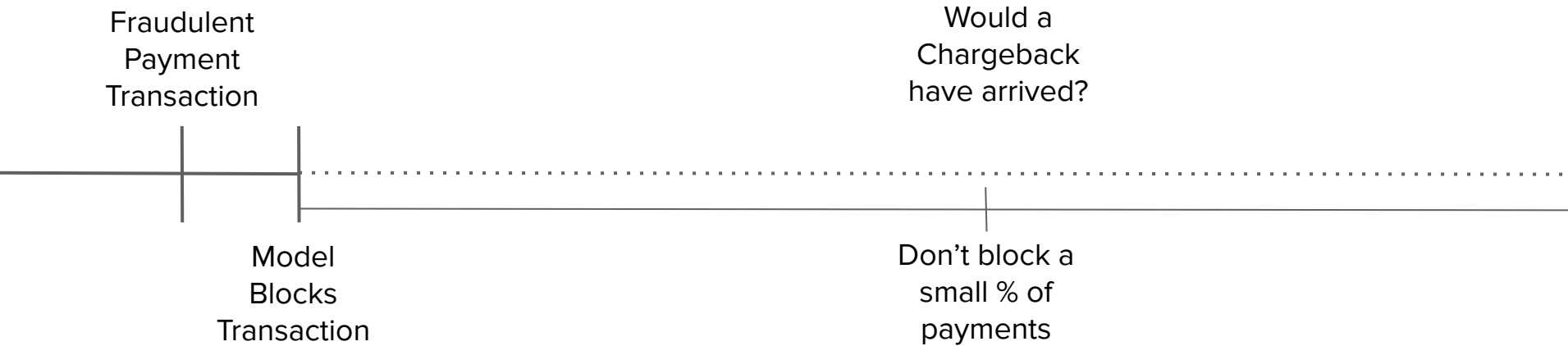
Would a  
Chargeback  
have arrived?

Model  
Blocks  
Transaction





# Life of Ground Truth



# Optimization

---

# What are we optimizing for?

## False Positive

- Miss out on a sale
- Churn – long term risk
- Difficult to measure!

## False Negative

- Straightforward financial loss
- Could be unbounded

## Capacity

- Limited capacity for human review
- Capacity has a cost

# What are we optimizing for?

## False Positive

## Precision

- Miss out on a sale
- Churn – long term risk
- Difficult to measure!

## False Negative

## Recall

- Straightforward financial loss
- Could be unbounded

## Capacity

## Support / Positive Prediction Rate

- Limited capacity for human review
- Capacity has a cost

# What are we optimizing for?

## False Positive

## Precision

- Miss out on a sale
- Churn – long term risk
- Difficult to measure!

## Capacity

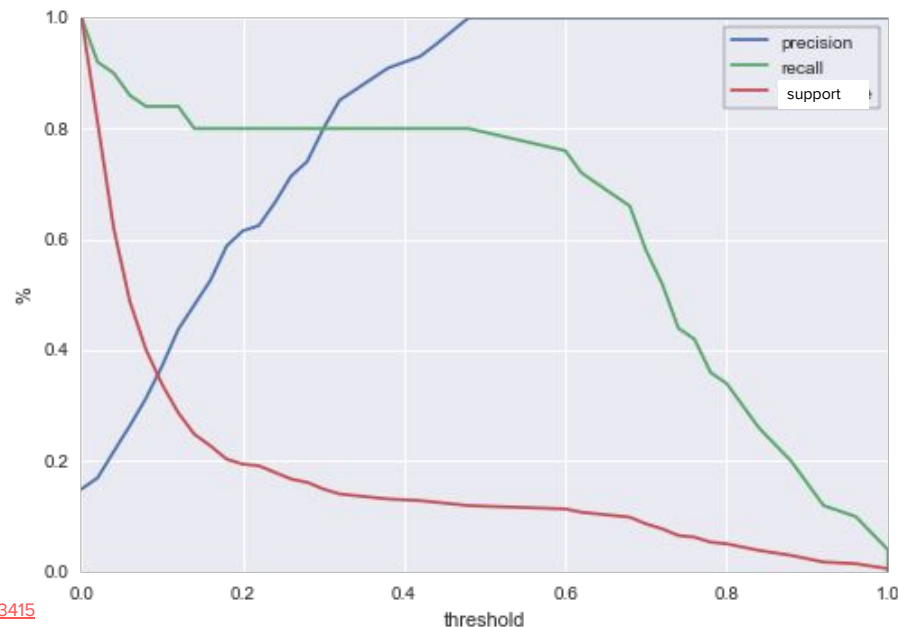
## Support / Positive Prediction Rate

- Limited capacity for human review
- Capacity has a cost

## False Negative

## Recall

- Straightforward financial loss
- Could be unbounded



# What are we optimizing for?

## False Positive

## Precision

- Miss out on a sale
- Churn – long term risk
- Difficult to measure!

## Capacity

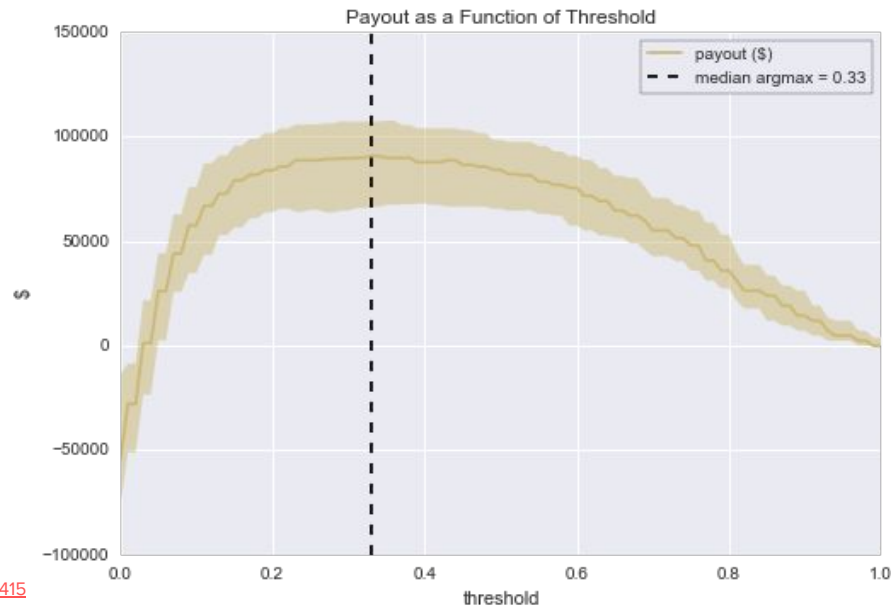
## Support / Positive Prediction Rate

- Limited capacity for human review
- Capacity has a cost

## False Negative

## Recall

- Straightforward financial loss
- Could be unbounded

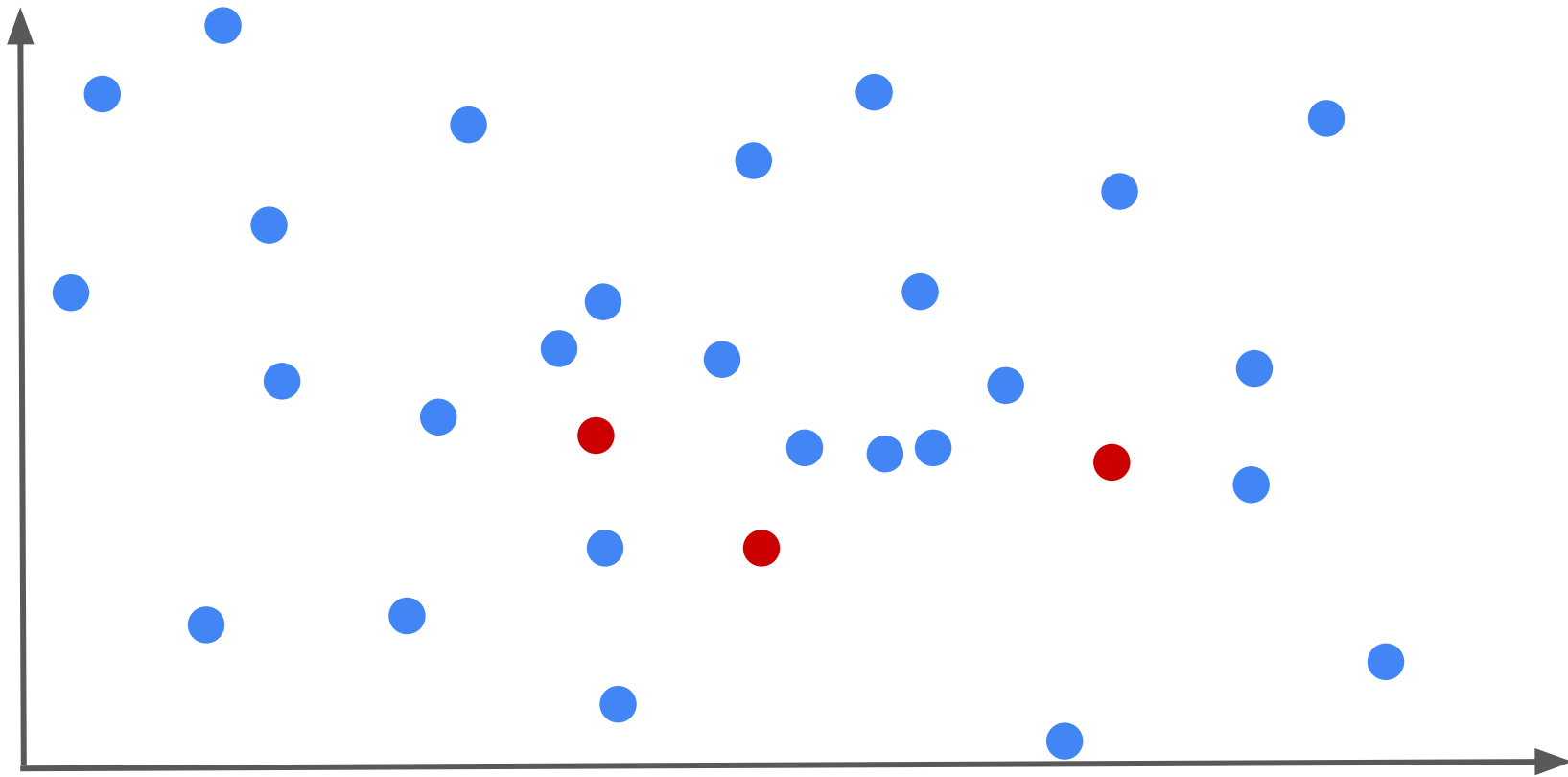


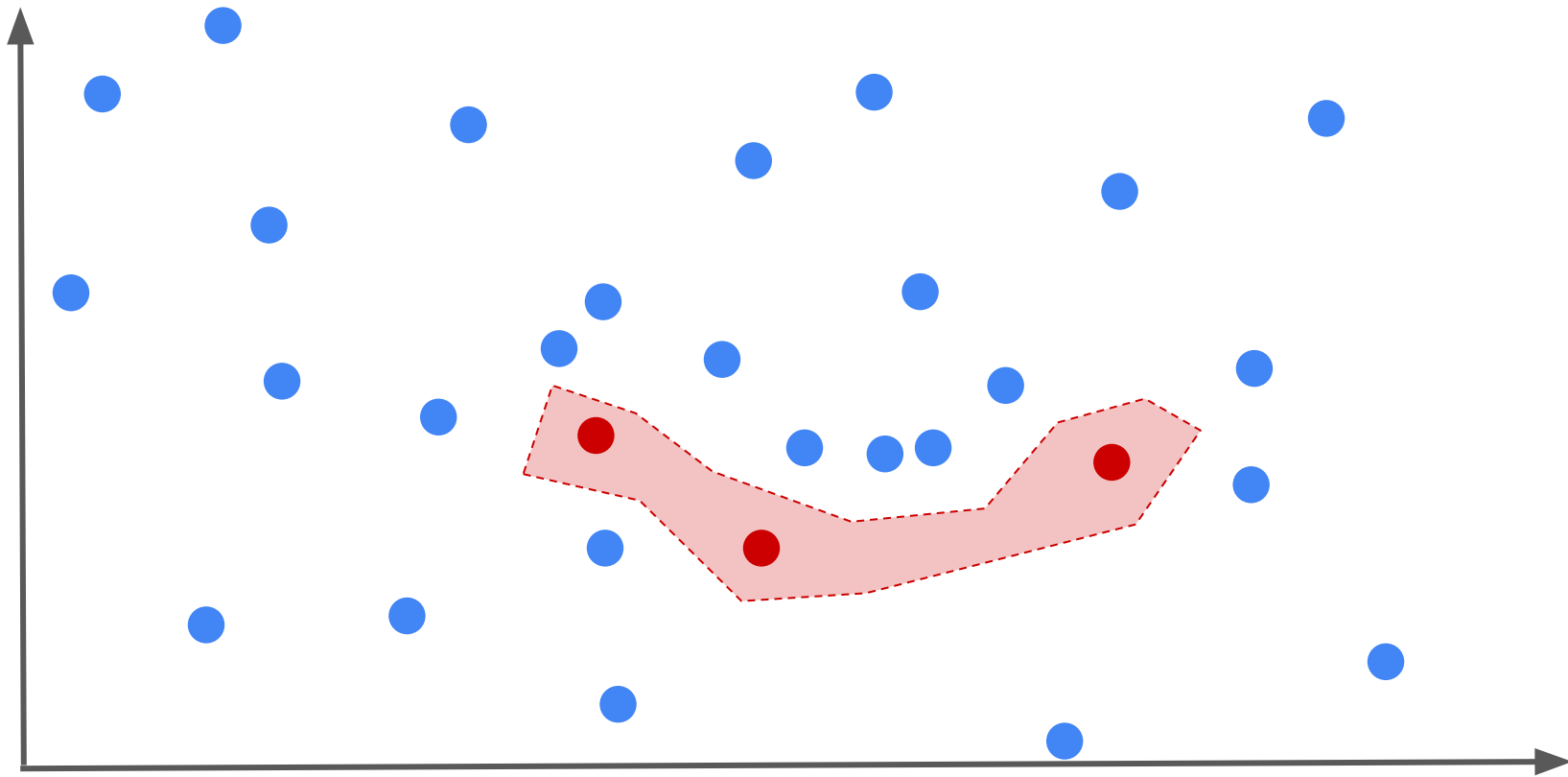
# Class Imbalance

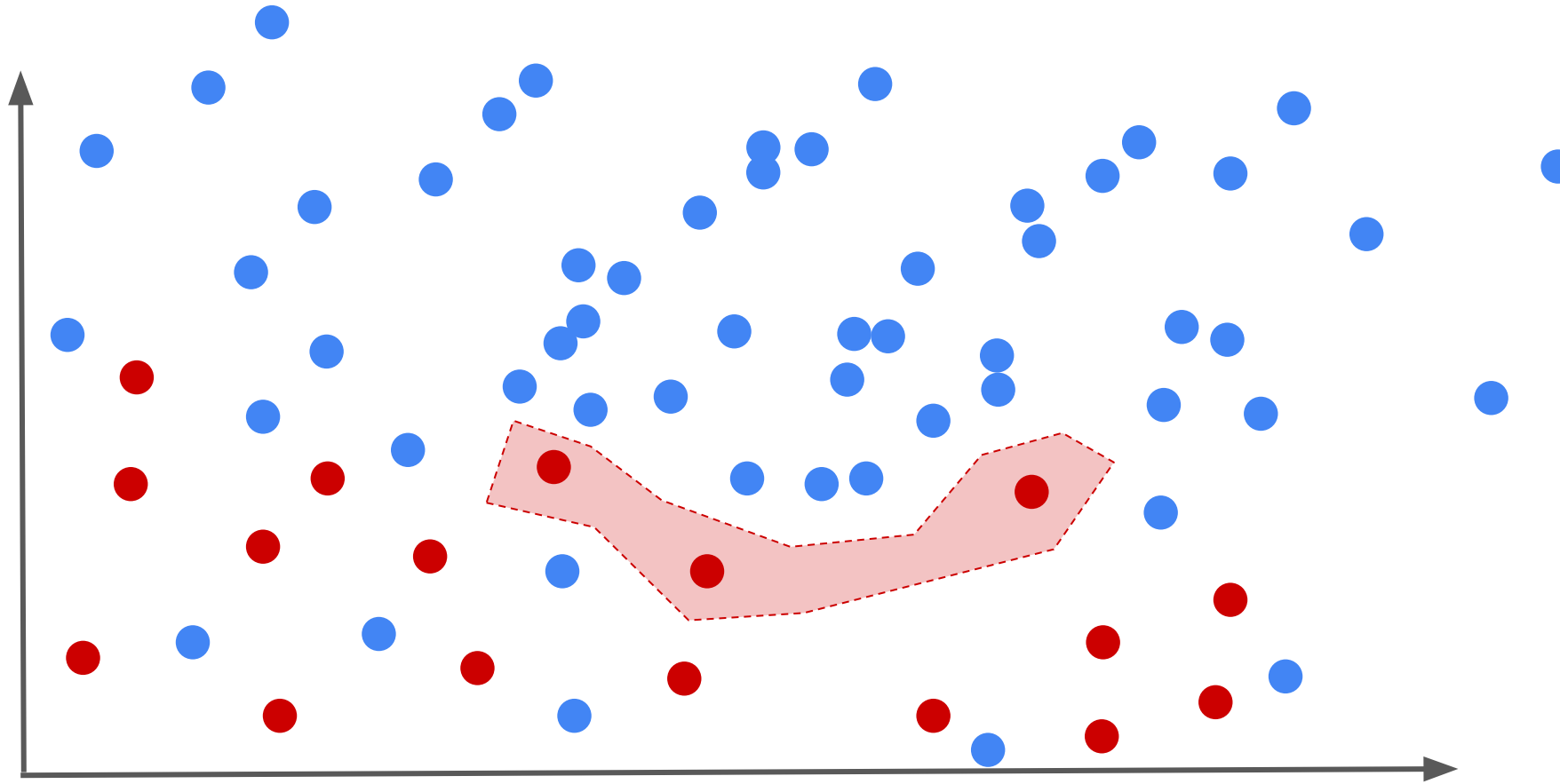
---

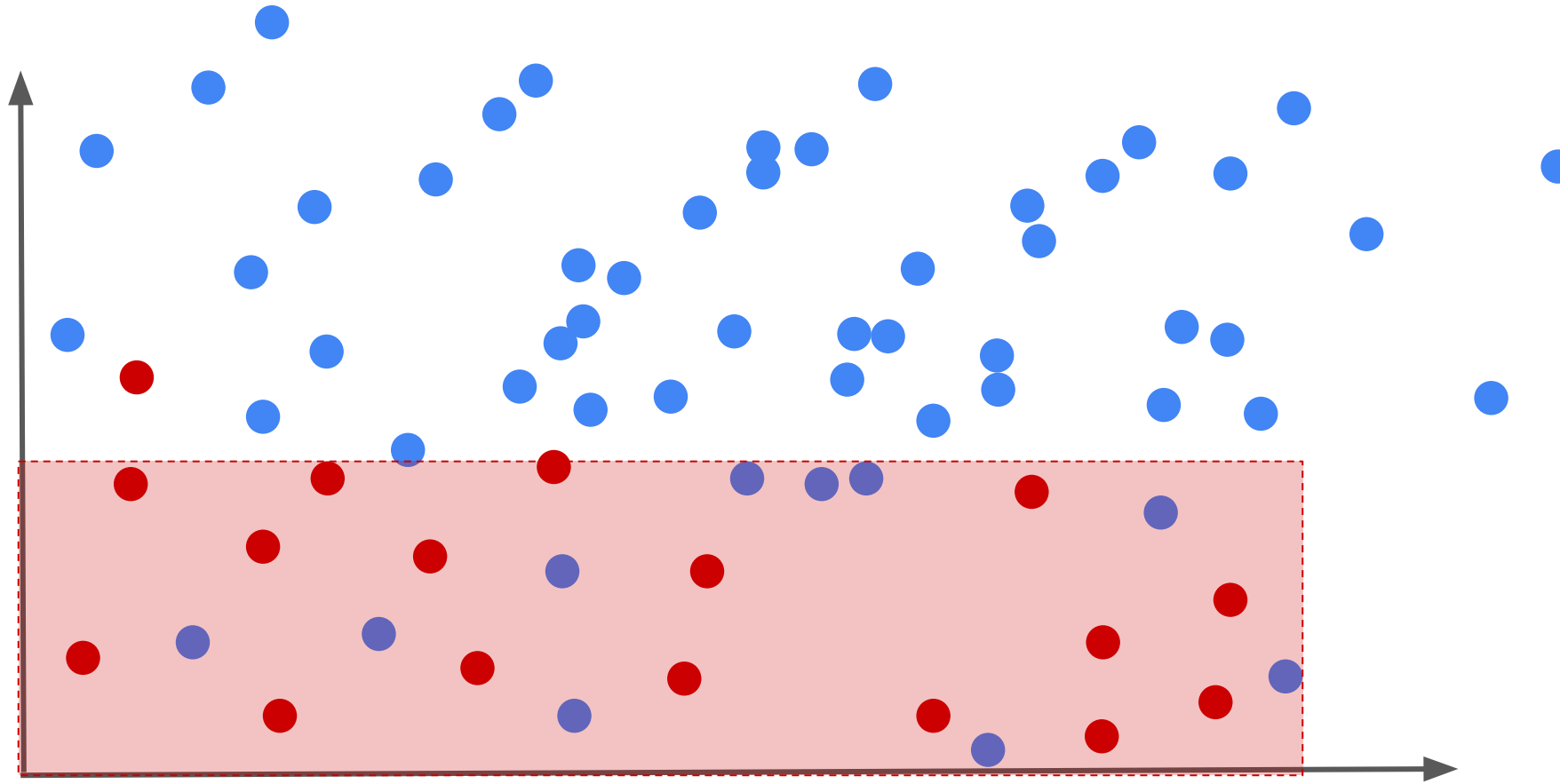
You need a lot of data





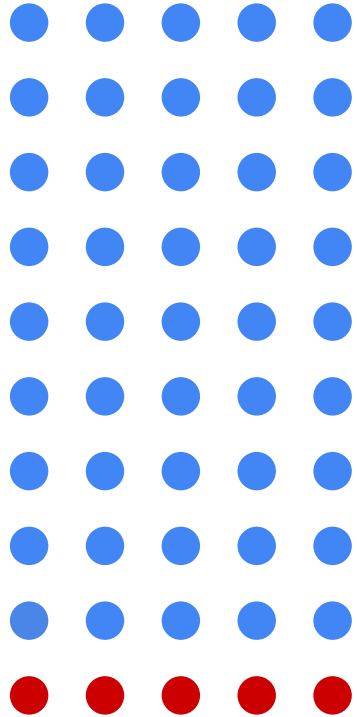




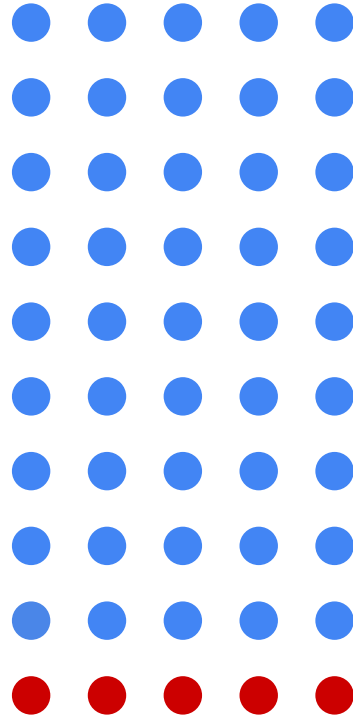


# Stratified Sampling

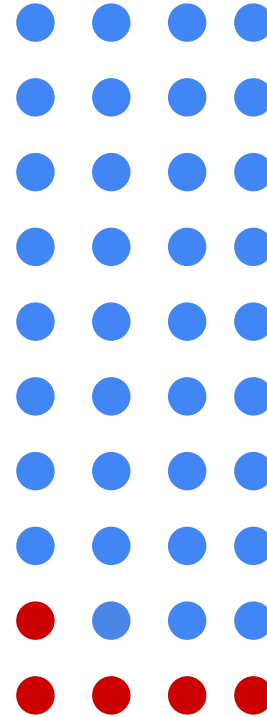
Full Dataset:  
50 samples, 10% Red



Full Dataset:  
50 samples, 10% Red



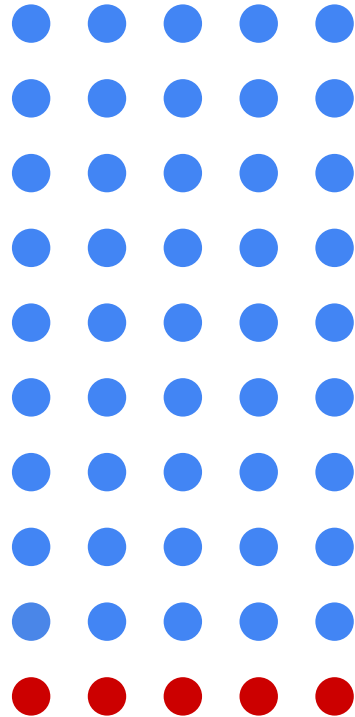
Training  
40 samples, 12.5% Red



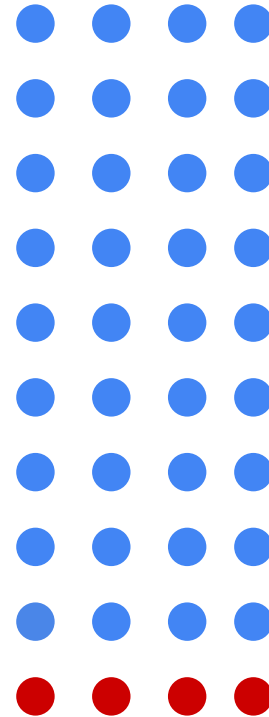
Test  
10 samples, 0% Red



Full Dataset:  
50 samples, 10% Red



Training  
40 samples, 10% Red



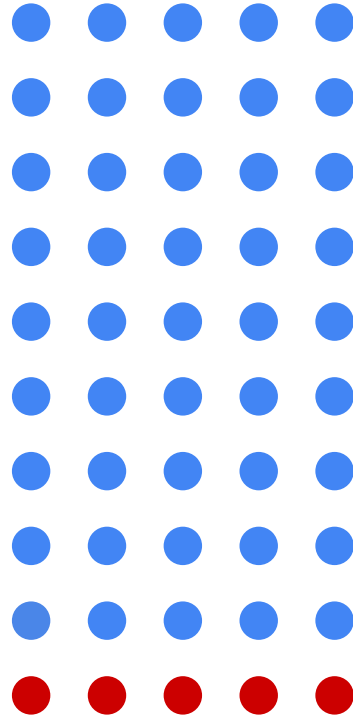
Test  
10 samples, 10% Red



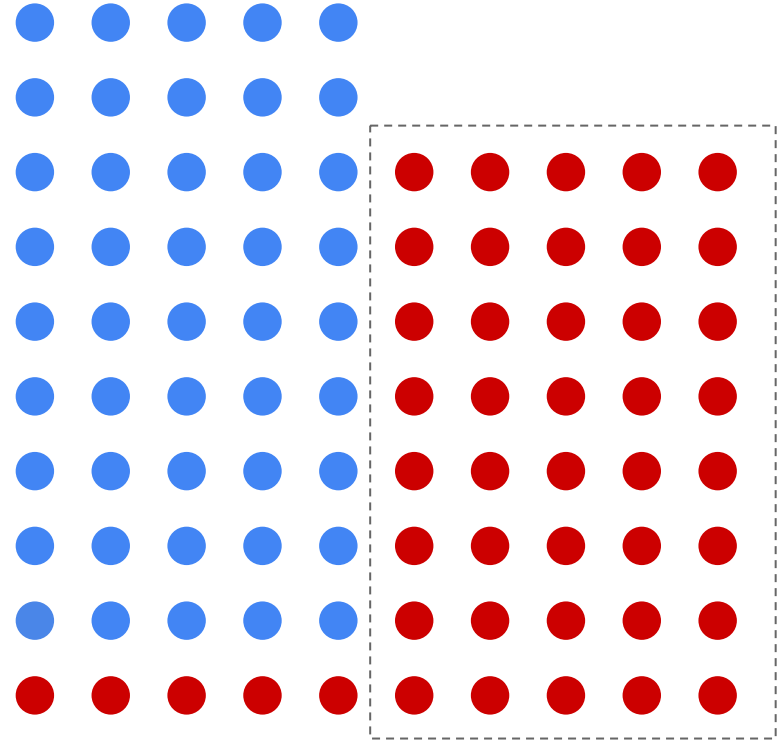


Upsampling

Full Dataset:  
50 samples, 10% Red

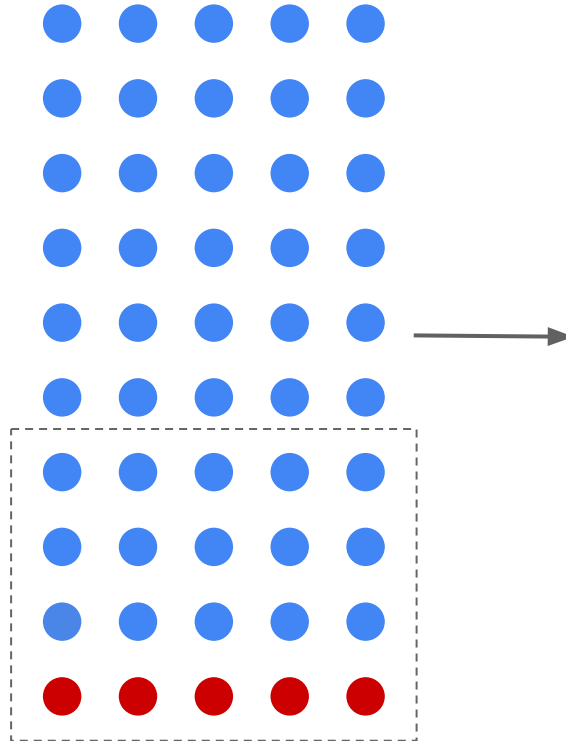


Upsampled Dataset:  
100 samples, 50% Red

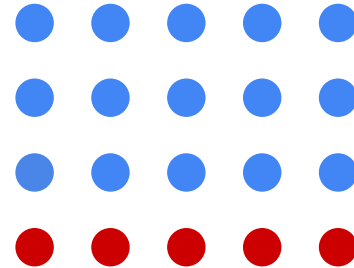


Downsampling

Full Dataset:  
50 samples, 10% Red

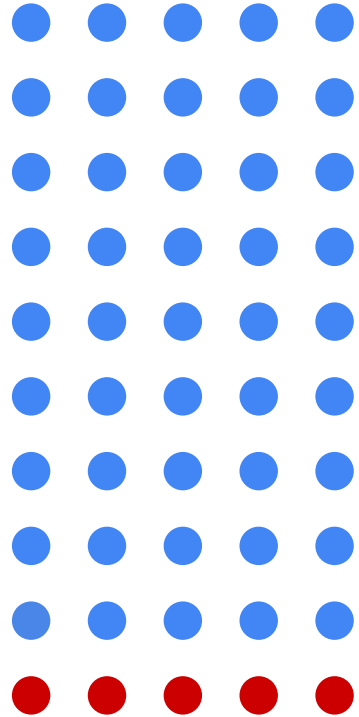


Downsampled Dataset:  
40 samples, 25% Red

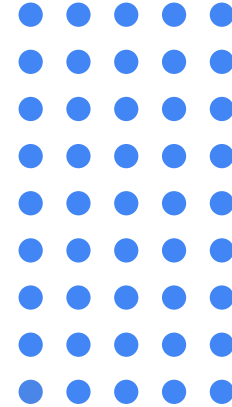


# Class Weighting

Full Dataset:  
50 samples, 10% Red



Class Weighted Dataset



Metrics

# Imbalanced Metrics

- Accuracy is not good!
- Sampling -> weighted metrics
- Uncertainty