Машинное обучение: мониторинг моделей в production

Эмели Драль

Проектная работа

Весь объем работы можно разделить на три стадии:

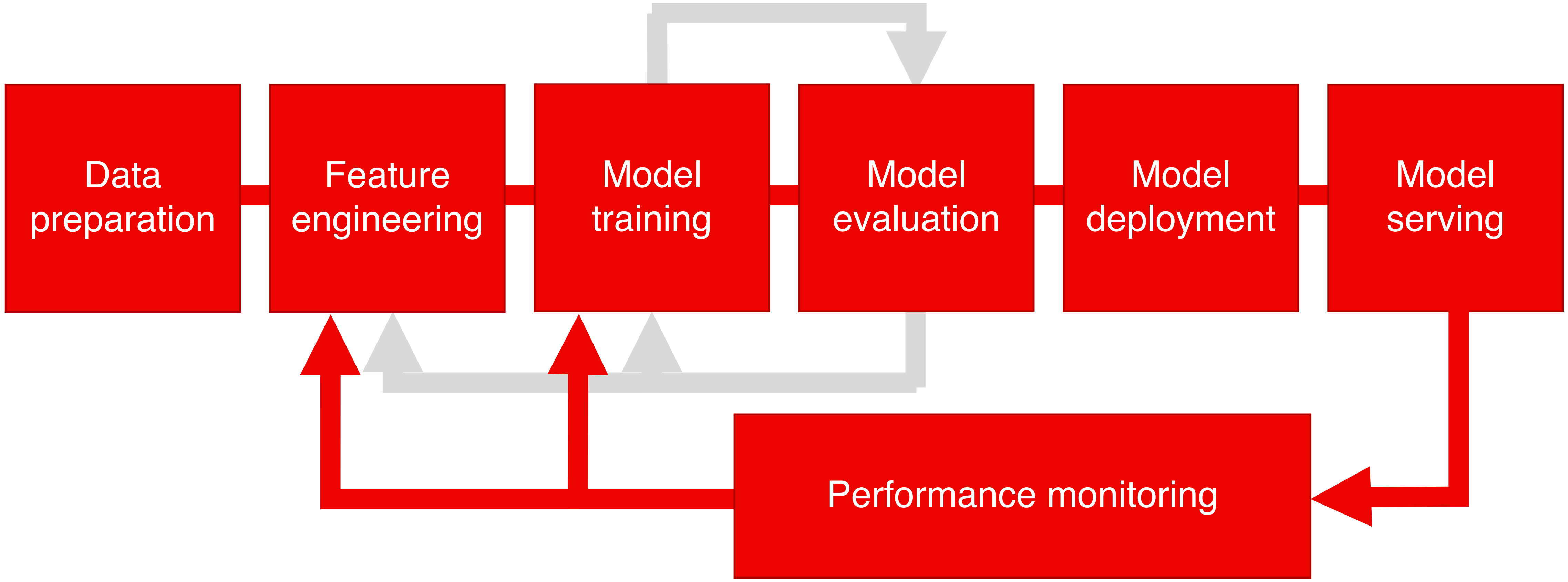
- Предпроектное исследование
- Работа над проектом
- **Работа после окончания проекта**

Мониторинг моделей в production

1. Что может пойти не так?
2. Структура мониторинга

Что может пойти не так?

Machine Learning Service Life Cycle



Data quality and integrity issues

Data processing issues

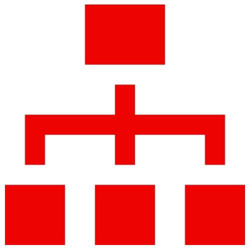
Broken pipelines, infrastructure updates, wrong source…

Data loss at the source

Broken sensor, logging error, database outage…

Data schema change

Change in the upstream system, external APIs, catalogue update…



Broken upstream model

One model's broken output = another model's corrupted feature

Example: data schema change

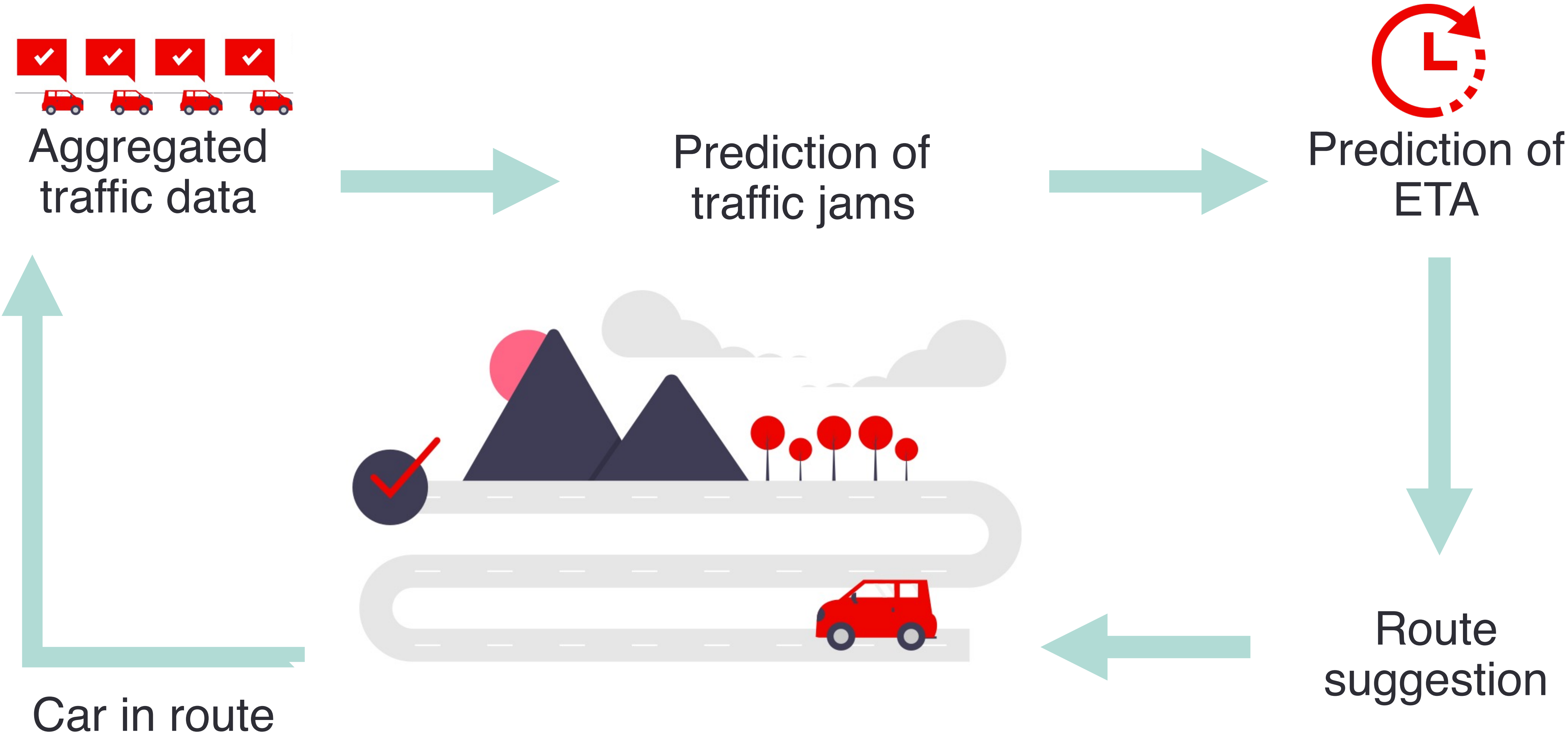
Cl_ID	Name	Type	Length	Status
#1229	######	card	2:27	solved
#1203	######	card	12:12	solved
#5661	######	account	8:06	solved
#8791	######	account	1:01	solved

BEFORE

Client ID	Client name	Call Type	Call Length	Channel preference	Status
#1229	######	card-lost	2:27	phone	solved
#1203	######	card-lost	12:12	phone	solved
#5661	######	account-balance	8:06	phone	solved
#8791	######	account-balance	1:01	email	solved

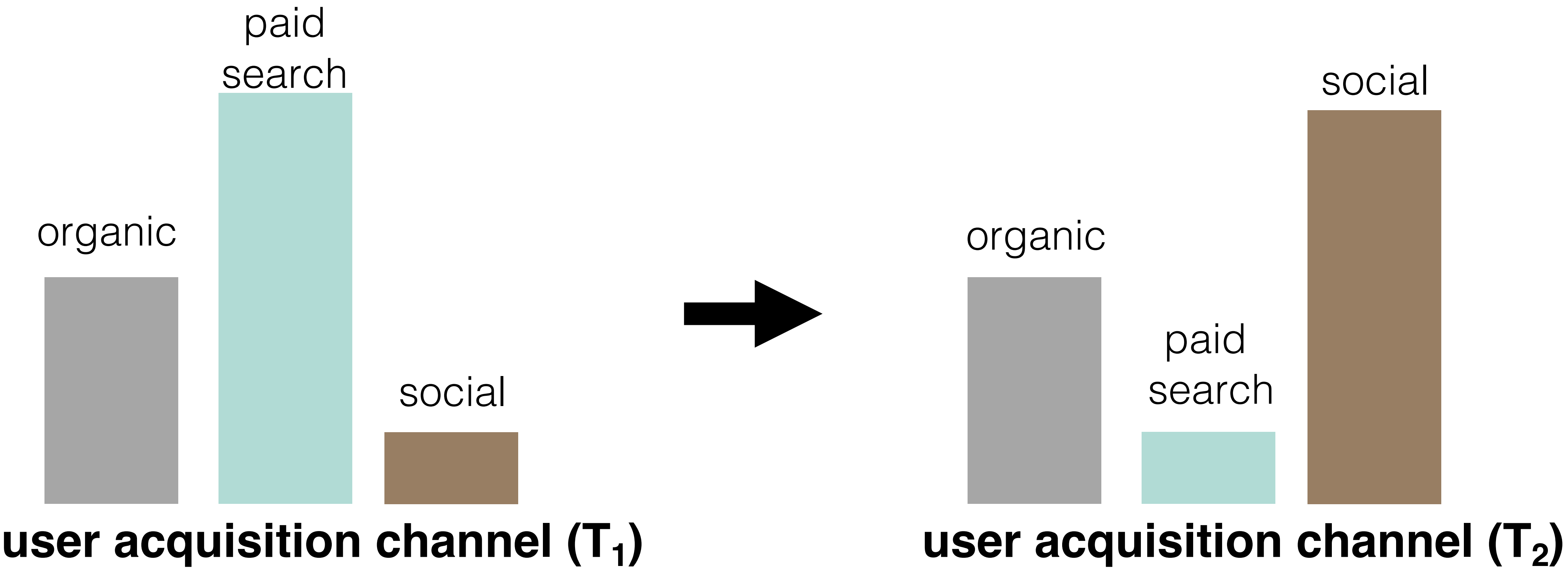
AFTER

Example: broken upstream model



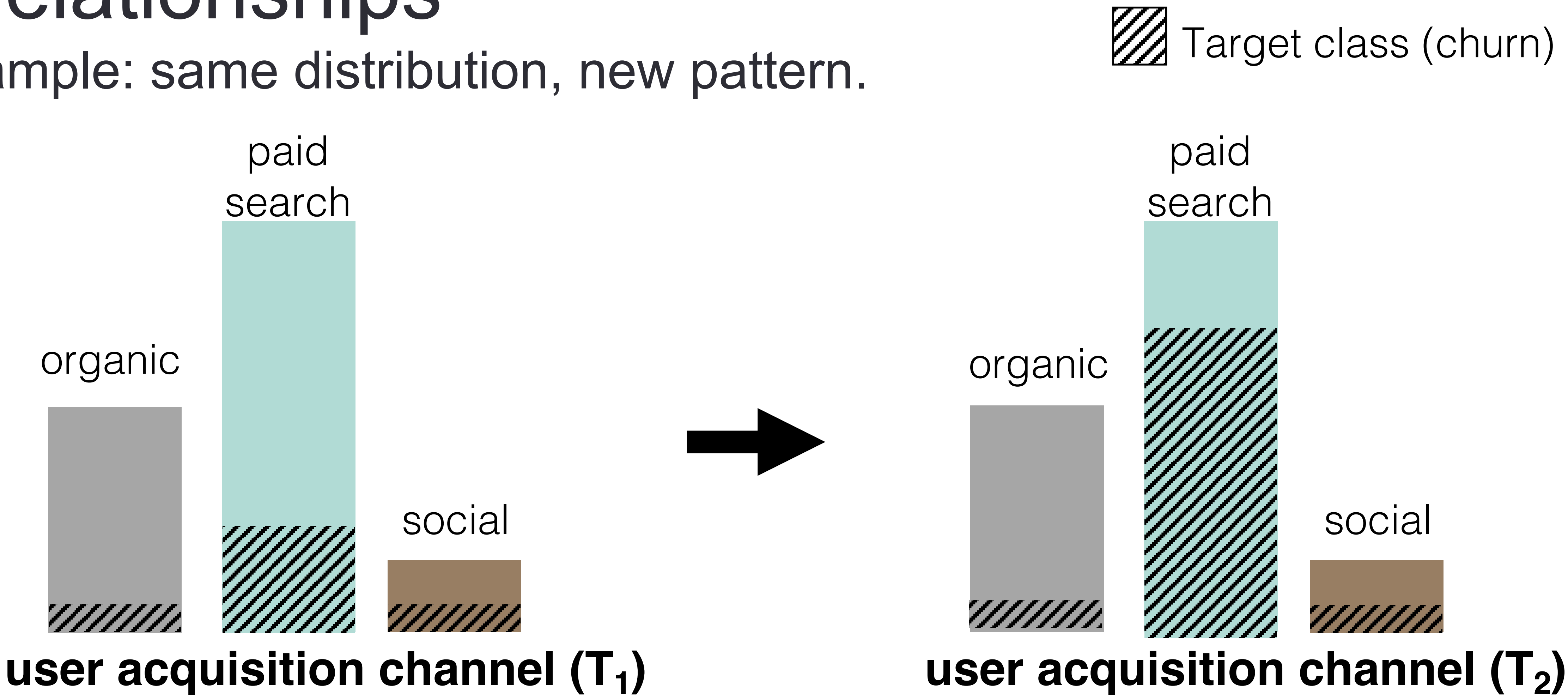
Data drift: change in feature distribution

Example: users come from a new channel.

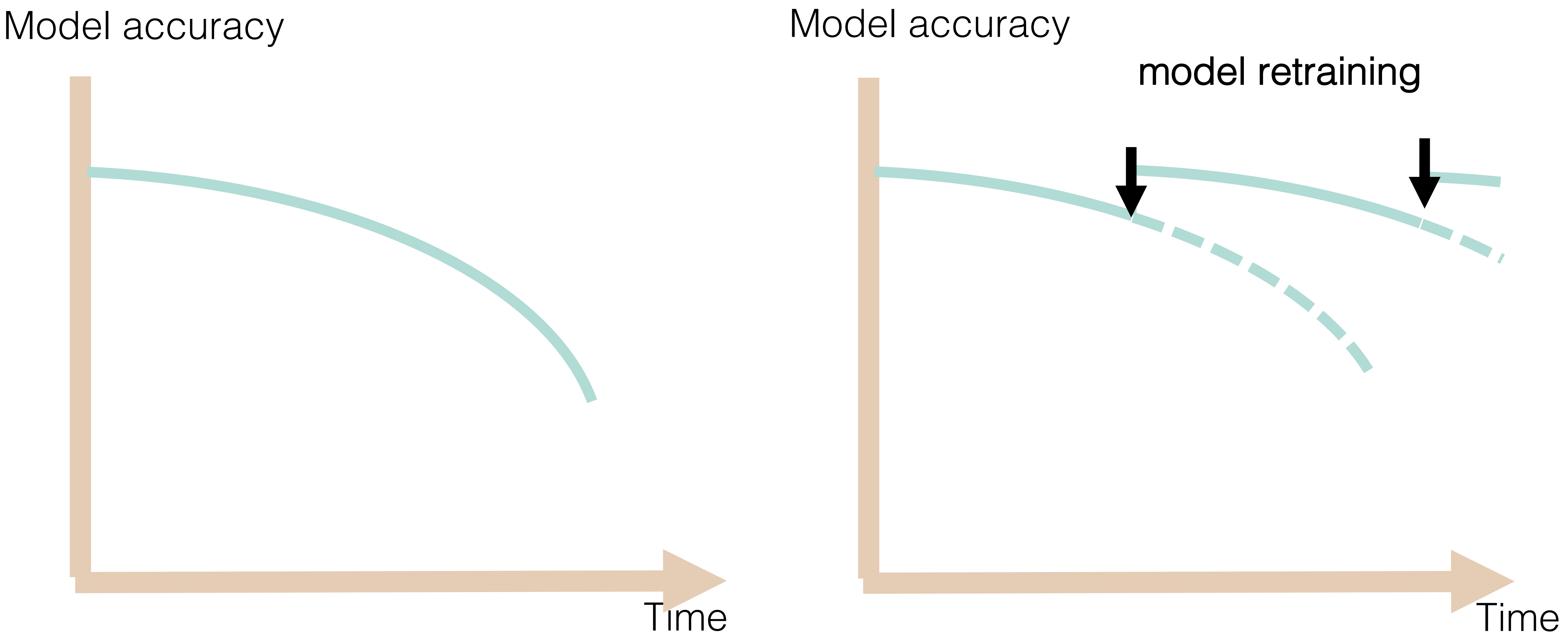


Concept drift: change in underlying relationships

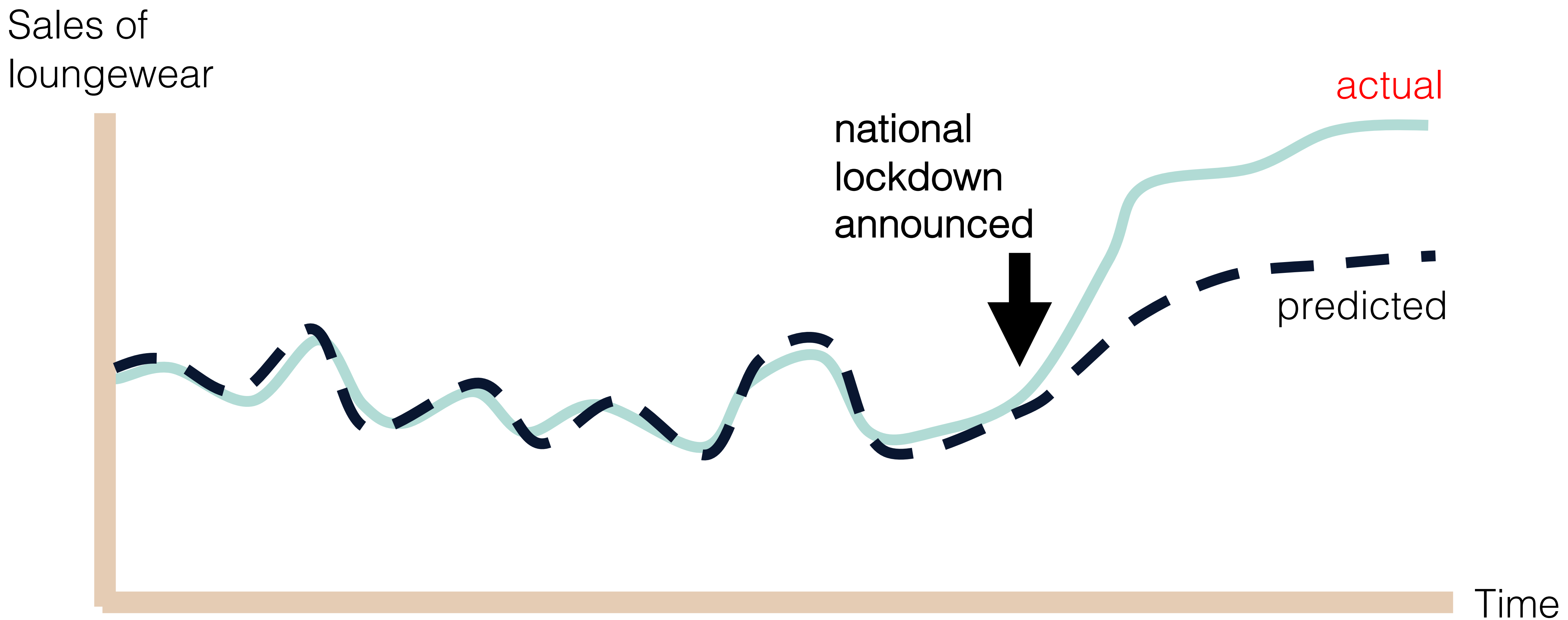
Example: same distribution, new pattern.



Gradual concept drift

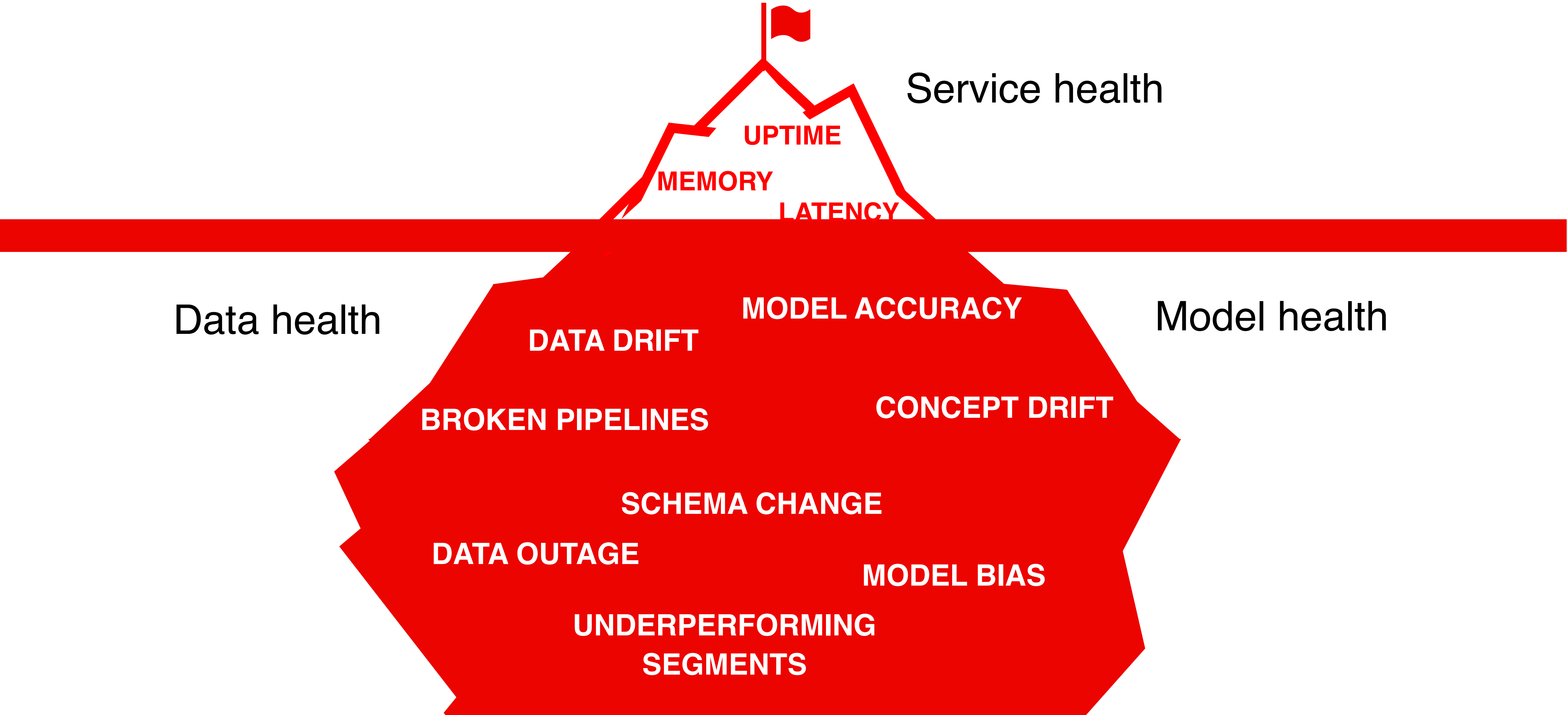


Sudden concept drift



Решение: мониторинг

How is machine learning monitoring different?

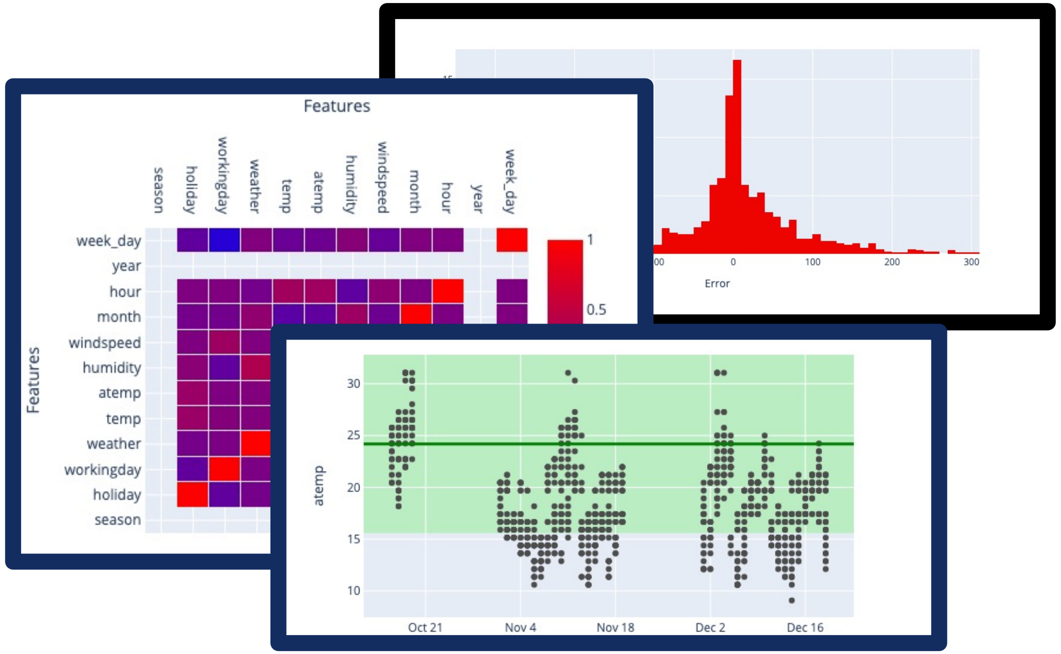


How is machine learning monitoring different?

67% **do not monitor their models**

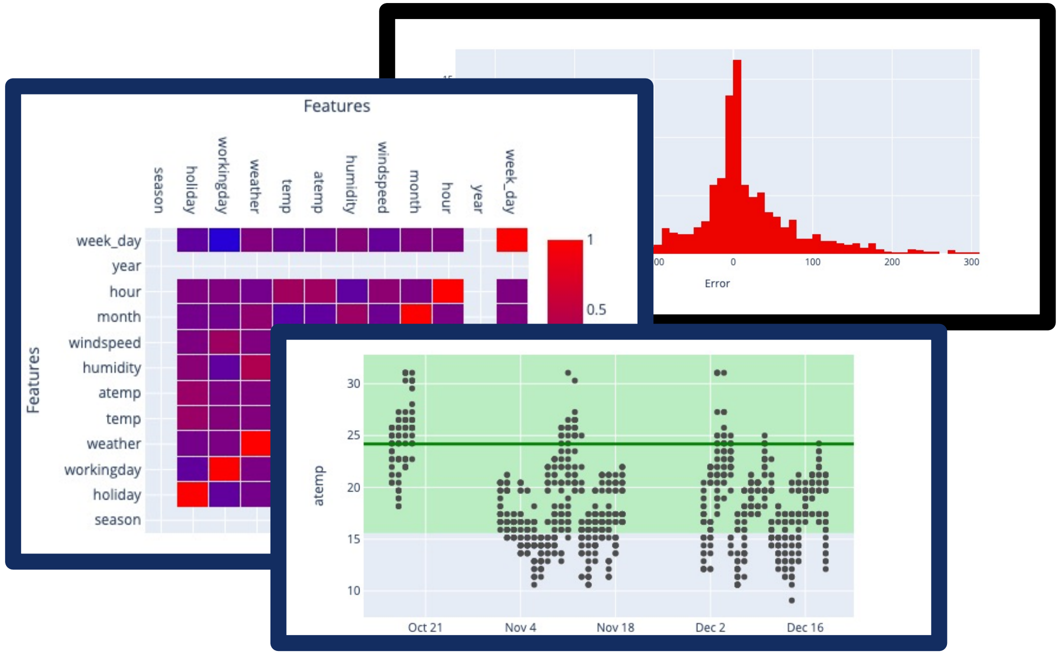
Source: report by FICO, 07/2020

How to monitor?

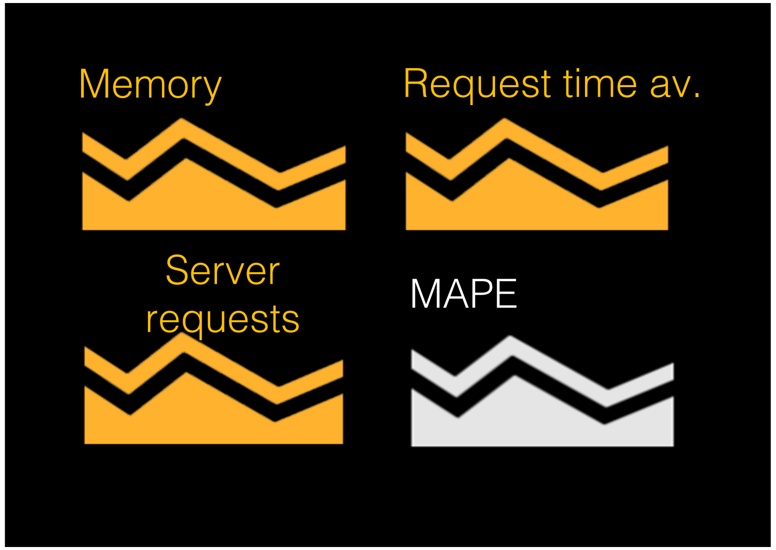


ML-focused Reports / Dashboards
(e.g. BI tools Tableau, Looker; or custom in Matplotlib, Plotly)

How to monitor?



ML-focused Reports / Dashboards
(e.g. BI tools Tableau, Looker; or custom in Matplotlib, Plotly)



Add ML metrics to service health monitoring
(e.g. Prometheus/Grafana)

Структура мониторинга

Monitoring approach: factors to consider

Use case importance

- Economic value
- Cost of error
- Risks

Complexity

- Data source diversity
- Pipeline complexity
- Batch / real-time
- Immediate / delayed response

Team resources

- Development resources

1.

Does it work?

- [x] **Service health**
- [ ] Model performance
- [ ] Data quality and integrity
- [ ] Data and concept drift

Model Calls: Start With Basics

2.

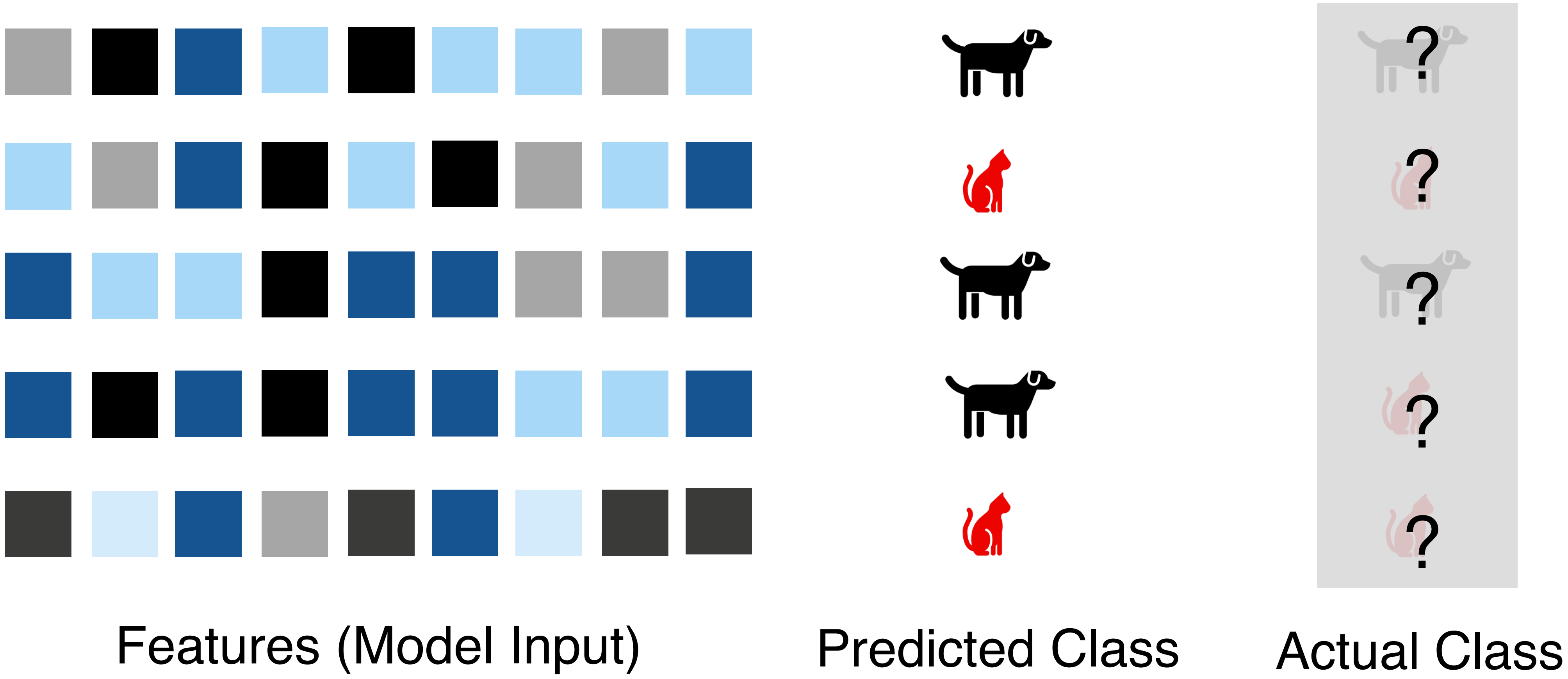
How it performs?

Did anything break?

- [x] Service health
- [x] **Model performance**
- [ ] Data quality and integrity
- [ ] Data and concept drift

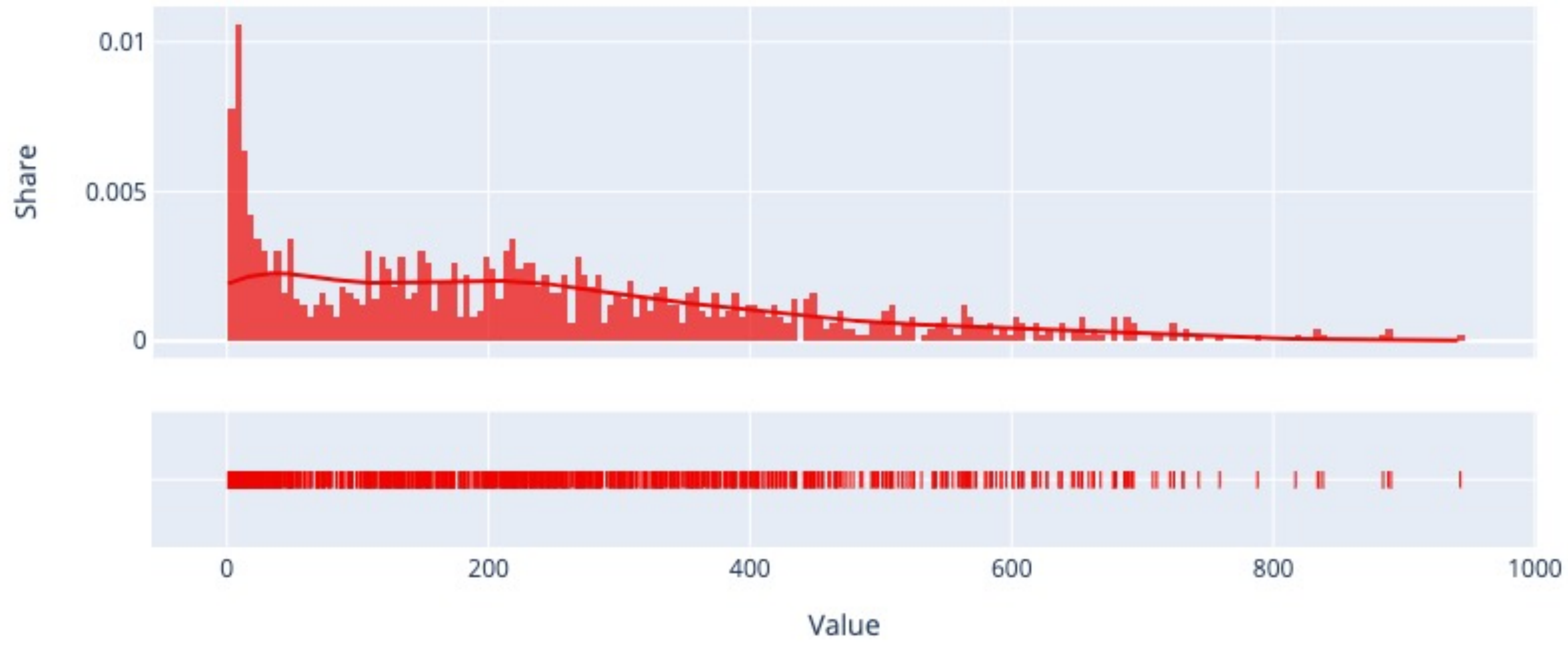
What if all we have are predictions?

Early monitoring when there is no ground truth



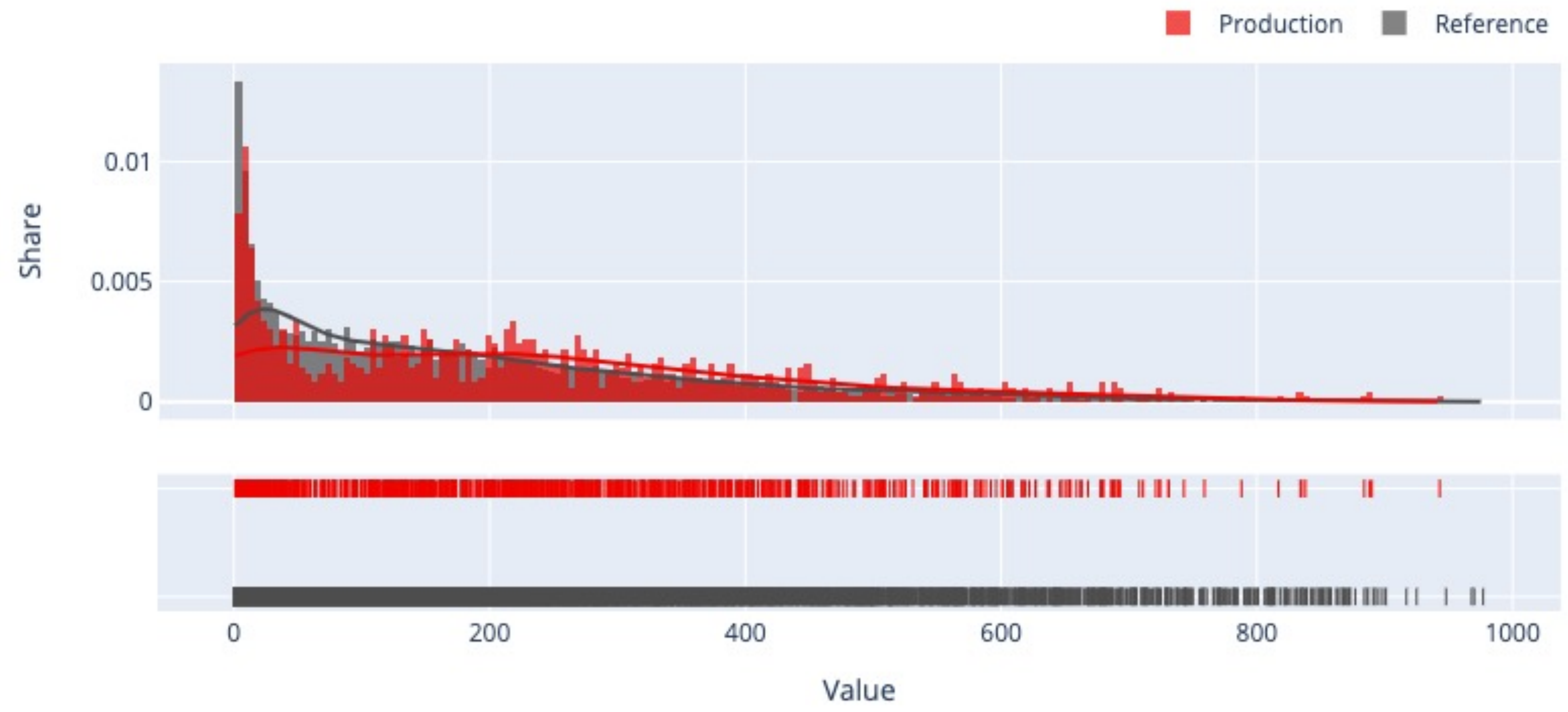
Model Output Distribution: Check Sanity and Ranges

If there is no immediate feedback loop



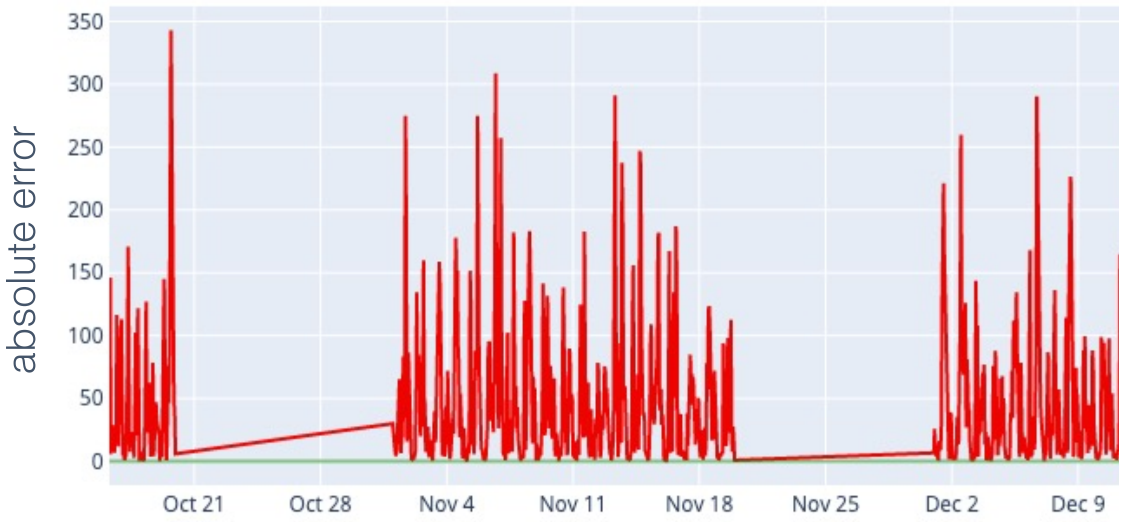
Model Output Distribution: Compare with Training

If you have some extra time



Model quality

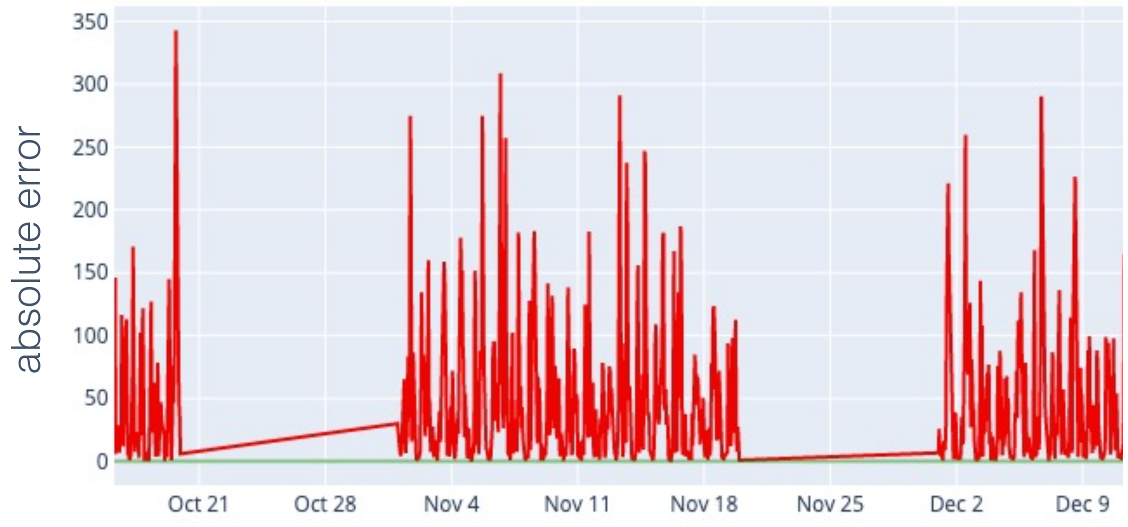
Ground truth is needed. Compare with results in hold-out to benchmark performance.



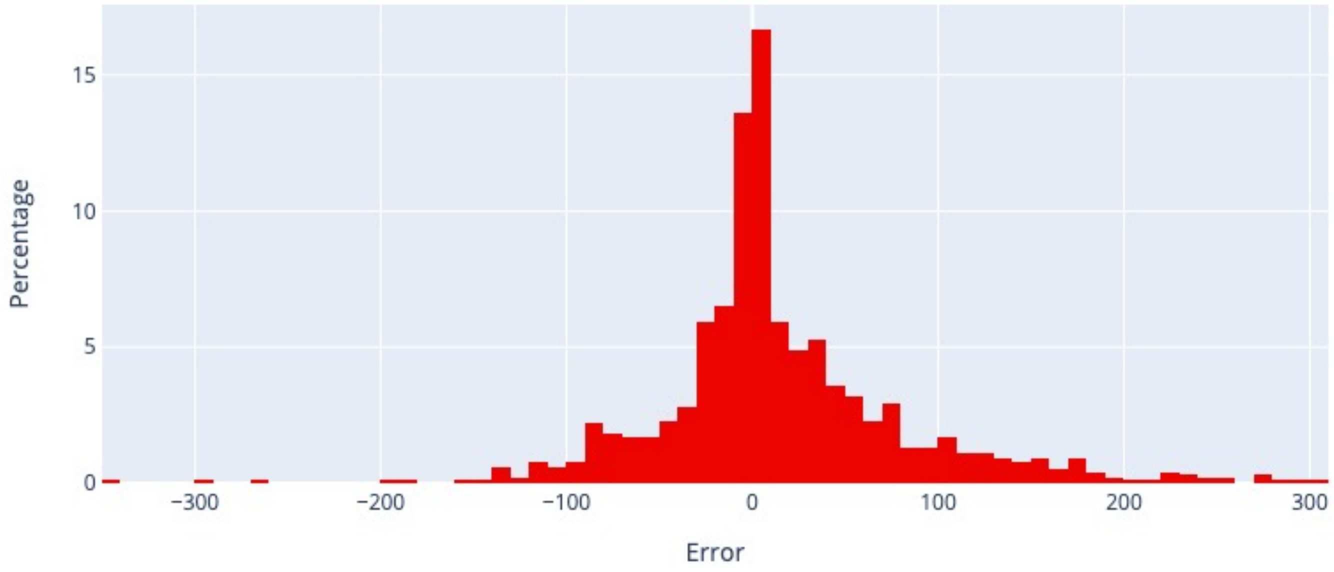
Interpretable metric **45.8** MAE

Model quality

Ground truth is needed. Compare with results in hold-out to benchmark performance.

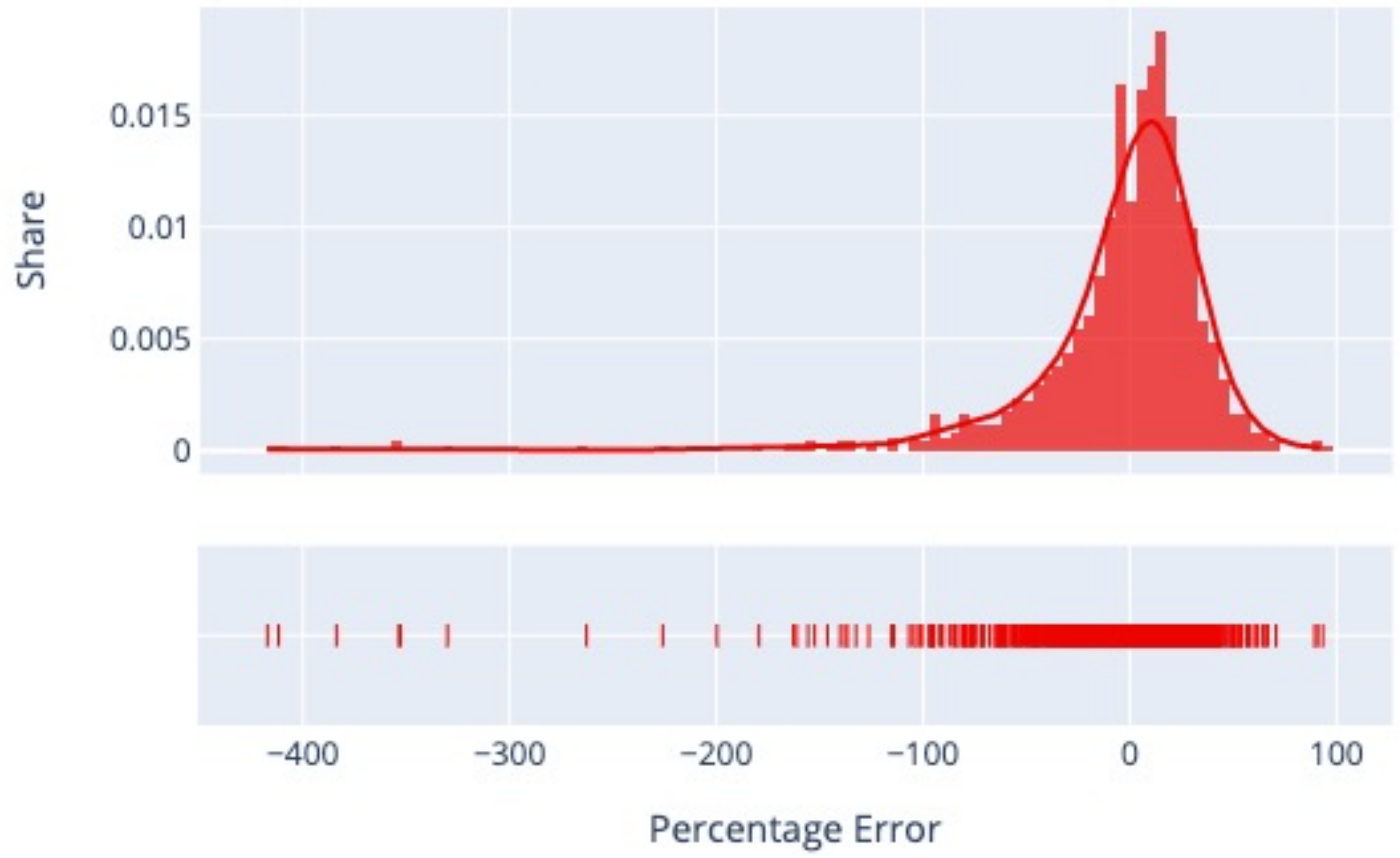


Interpretable metric **45.8** MAE



Error distribution

Perecentage Error Distribution: Check for Abnormalities



Business Metric: e.g. Share of Errors > 100

3.

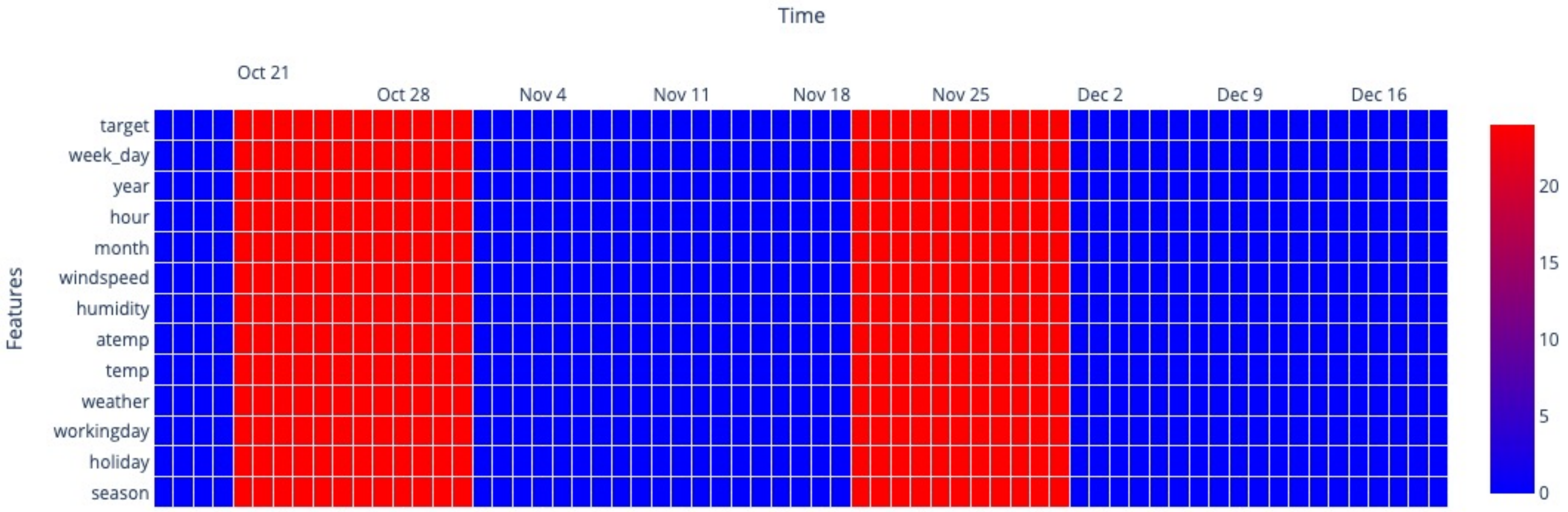
Where it breaks?

Where to dig further?

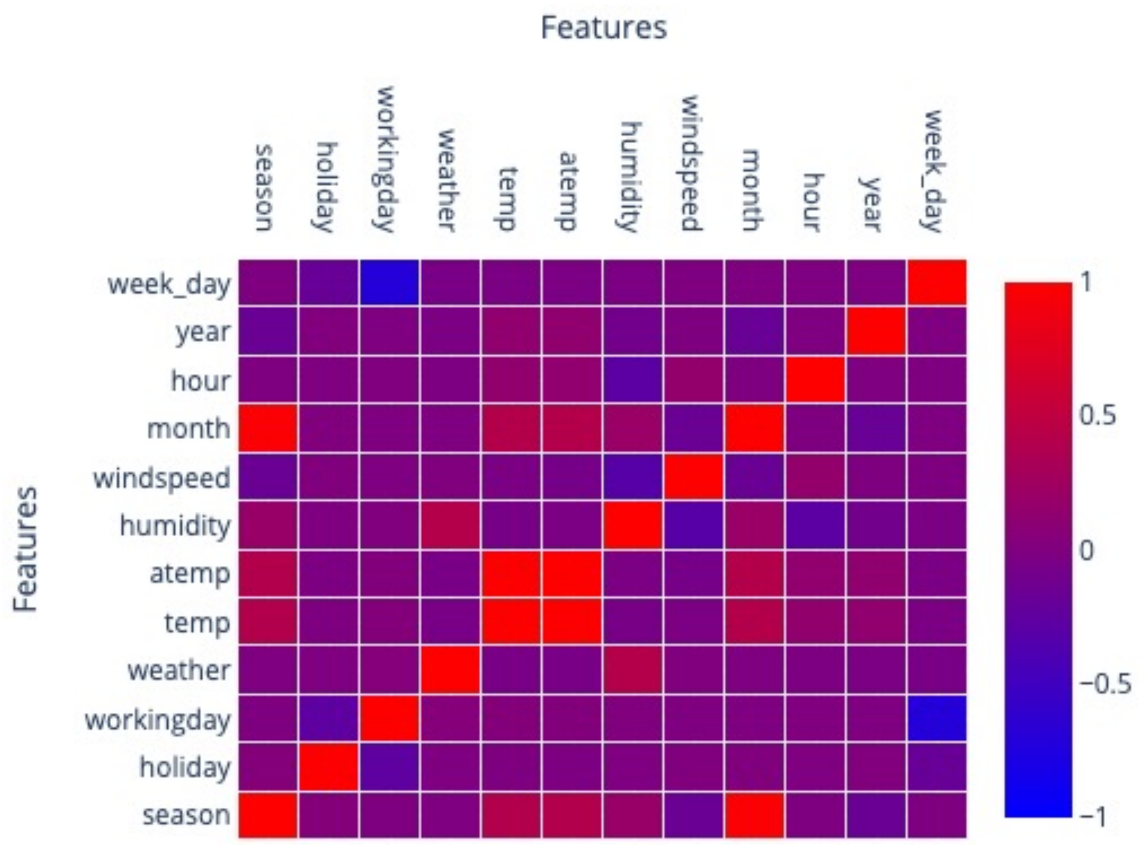
- [x] Service health
- [x] Model performance
- [x] **Data quality and integrity**
- [ ] Data and concept drift

Missing Data, Range Compliance, Type Compliance

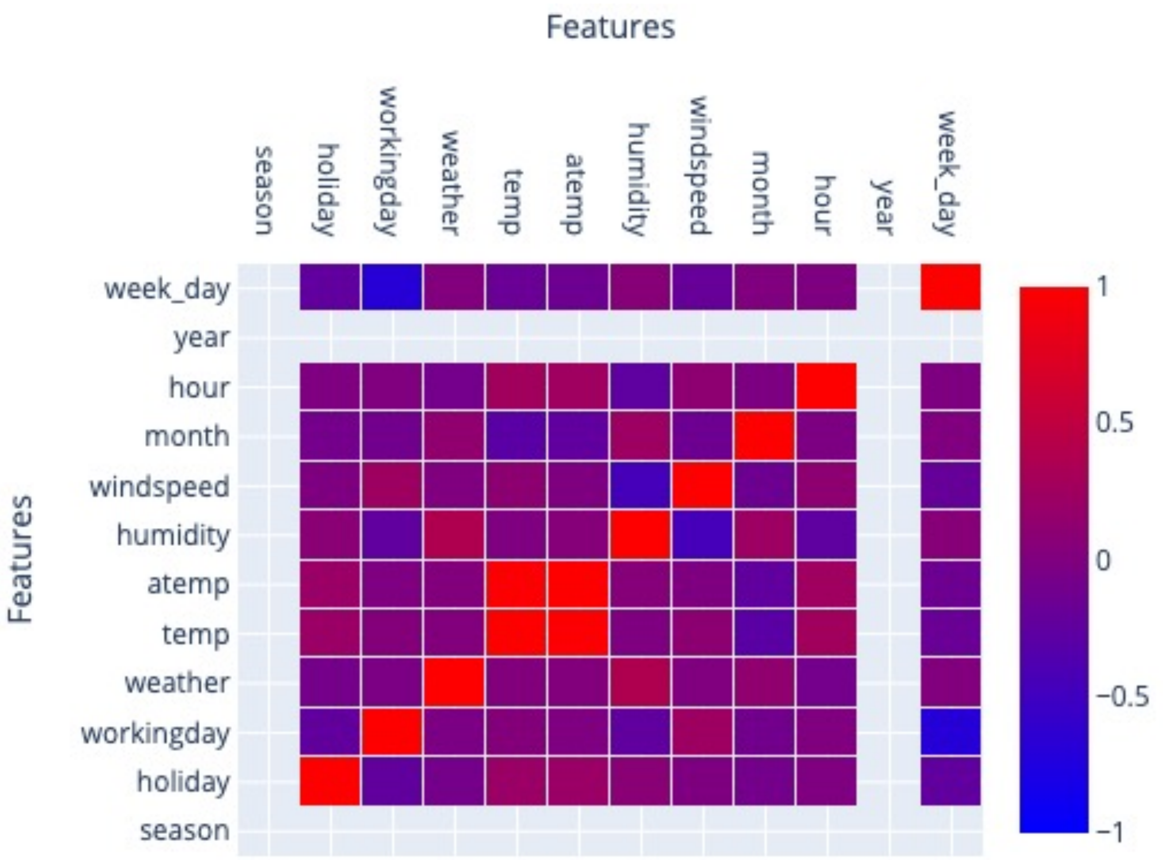
35%	0%	0%
missing data	out of range	type mismatch



Feature Correlation: Check for Changes

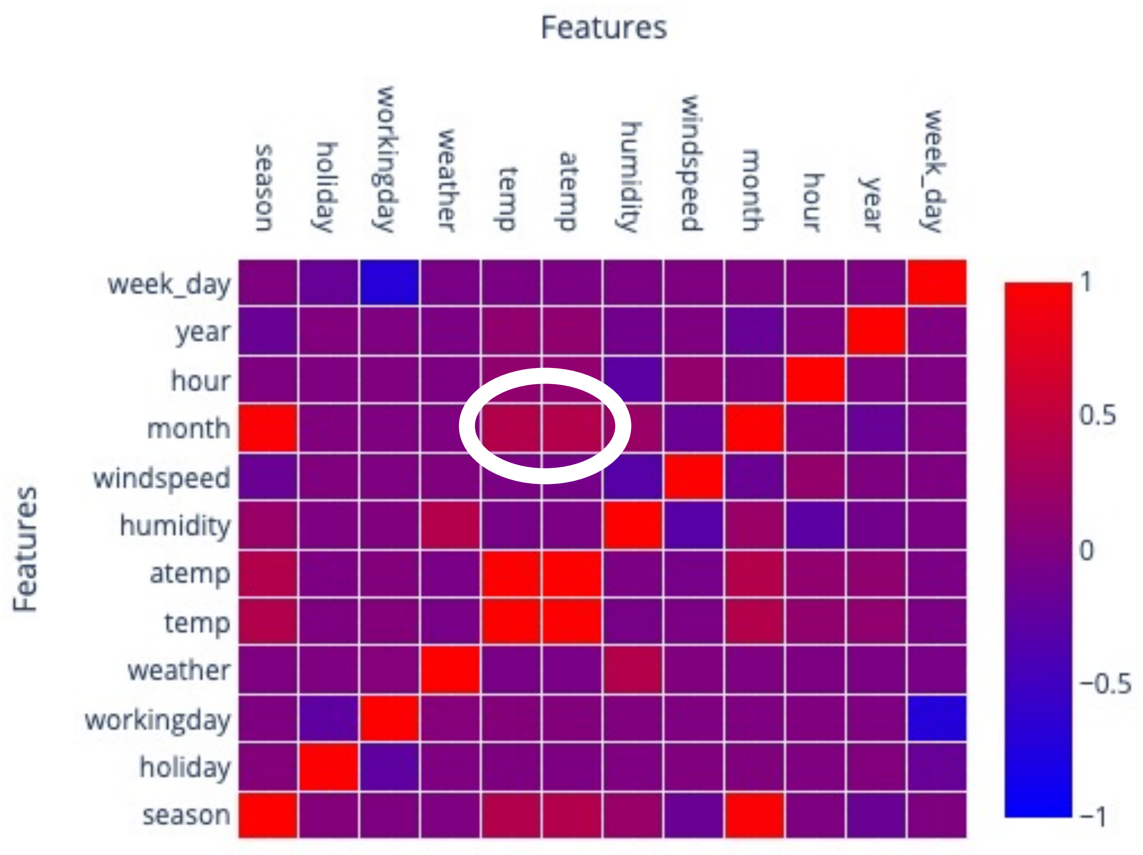


REFERENCE

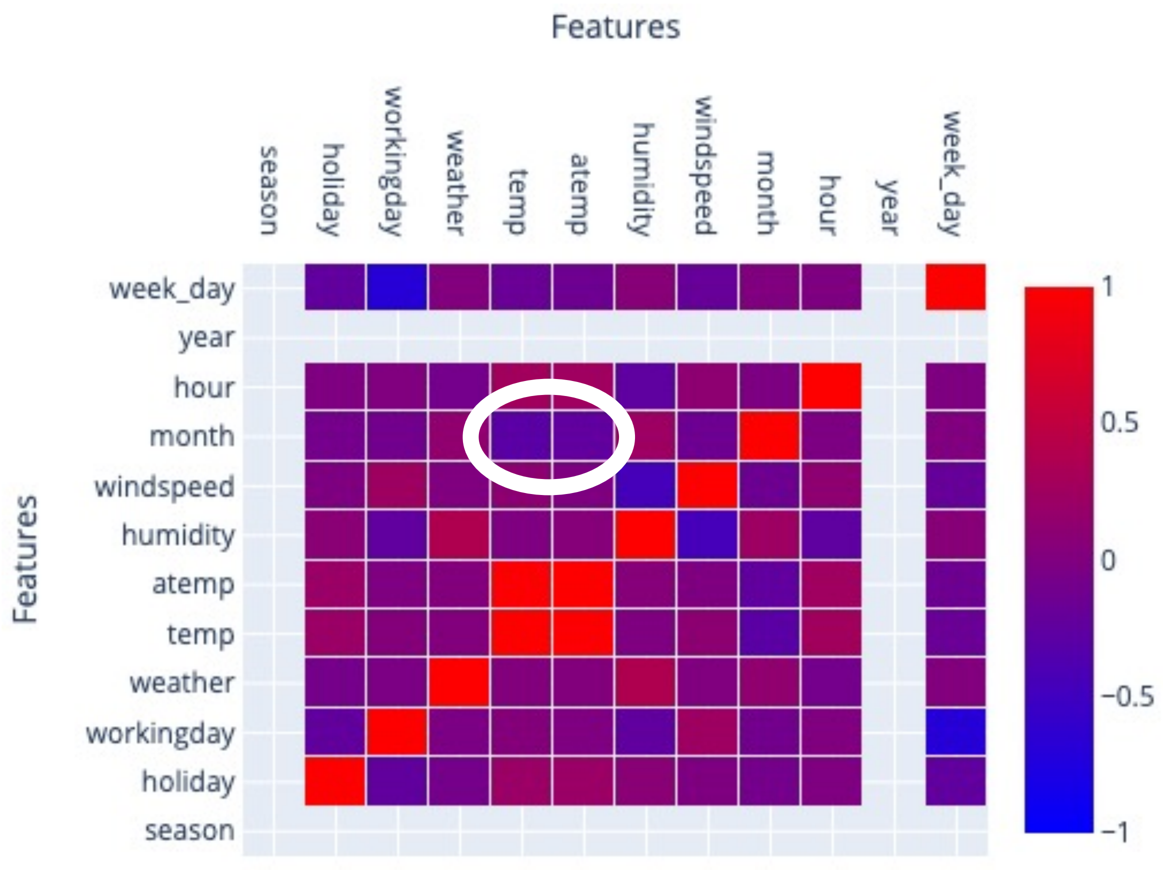


PRODUCTION

Feature Correlation: Check for Changes



REFERENCE



PRODUCTION

4.

Is model still relevant?

- ☑ Service health
- ☑ Model performance
- ☑ Data quality and integrity
- ☑ **Data and concept drift**

Why It Matters? Concept Drift.

1 / GRADUAL DRIFT

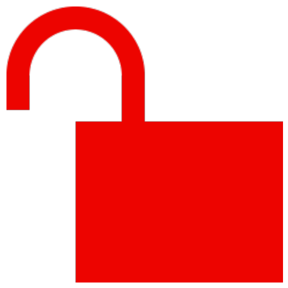
(model needs retraining / update)

New type of
fraud appeared

Equipment
wears out

Why It Matters? Concept Drift.

1 / GRADUAL DRIFT
(model needs retraining / update)



New type of fraud appeared

Equipment wears out

2 / SUDDEN DRIFT
(model is often rebuilt)

Grocery demand in pandemic

Unseen change in interest rate

Why It Matters? Concept Drift.

1 / GRADUAL DRIFT
(model needs retraining / update)

New type of fraud appeared

Equipment wears out

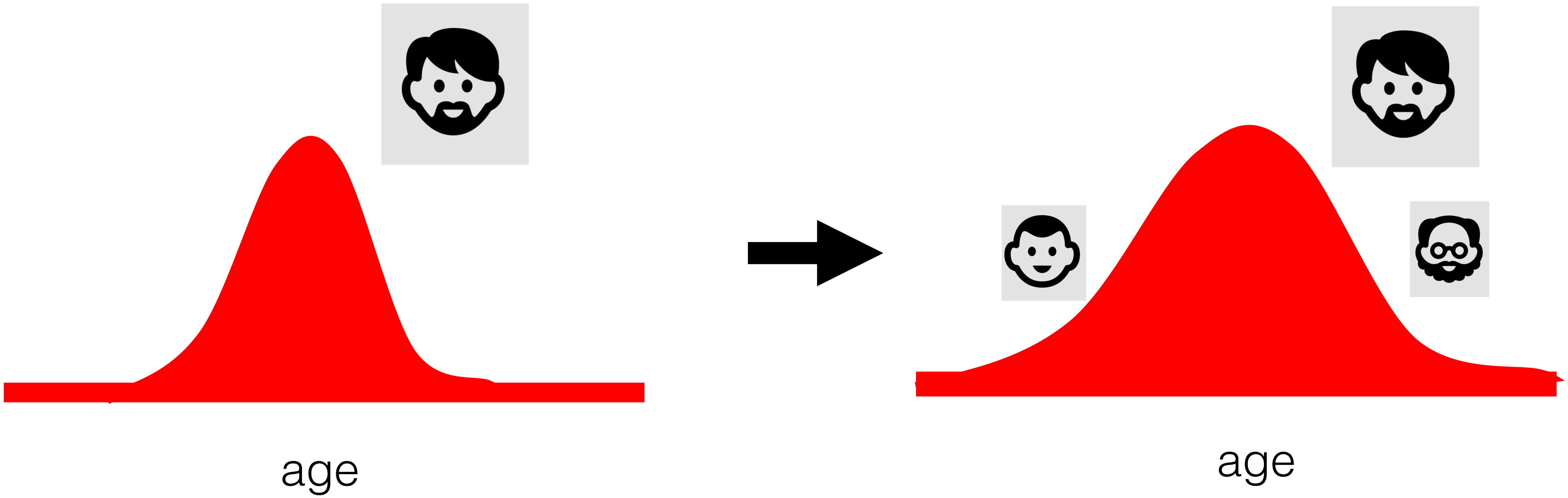
2 / SUDDEN DRIFT
(model is often rebuilt)

Grocery demand in pandemic

Unseen change in interest rate

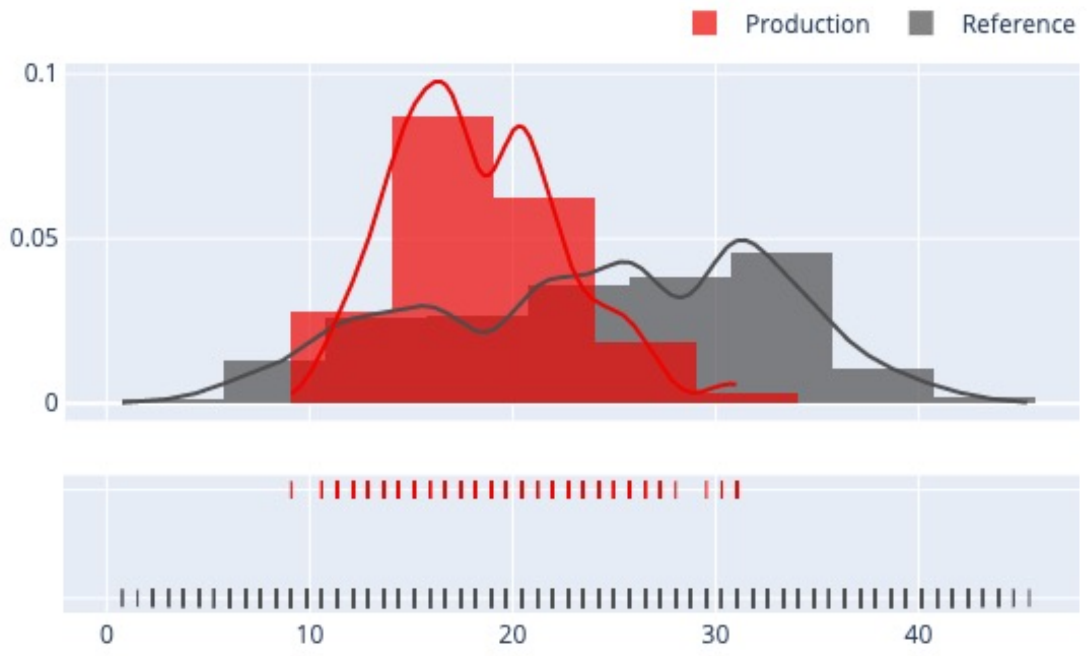
+ 3 / RECURRING DRIFT - unknown seasonality

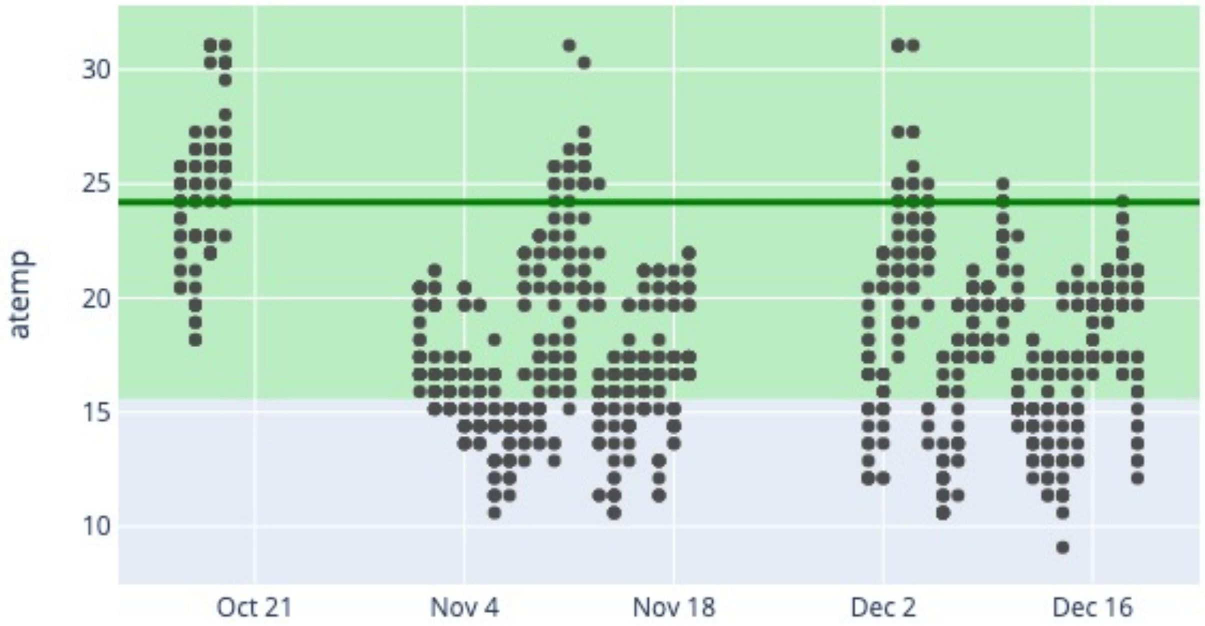
Why It Matters? Data Drift.



Feature Distribution And Statistics

Pragmatic approach: look only at key drivers. Check distribution visually & statistically.



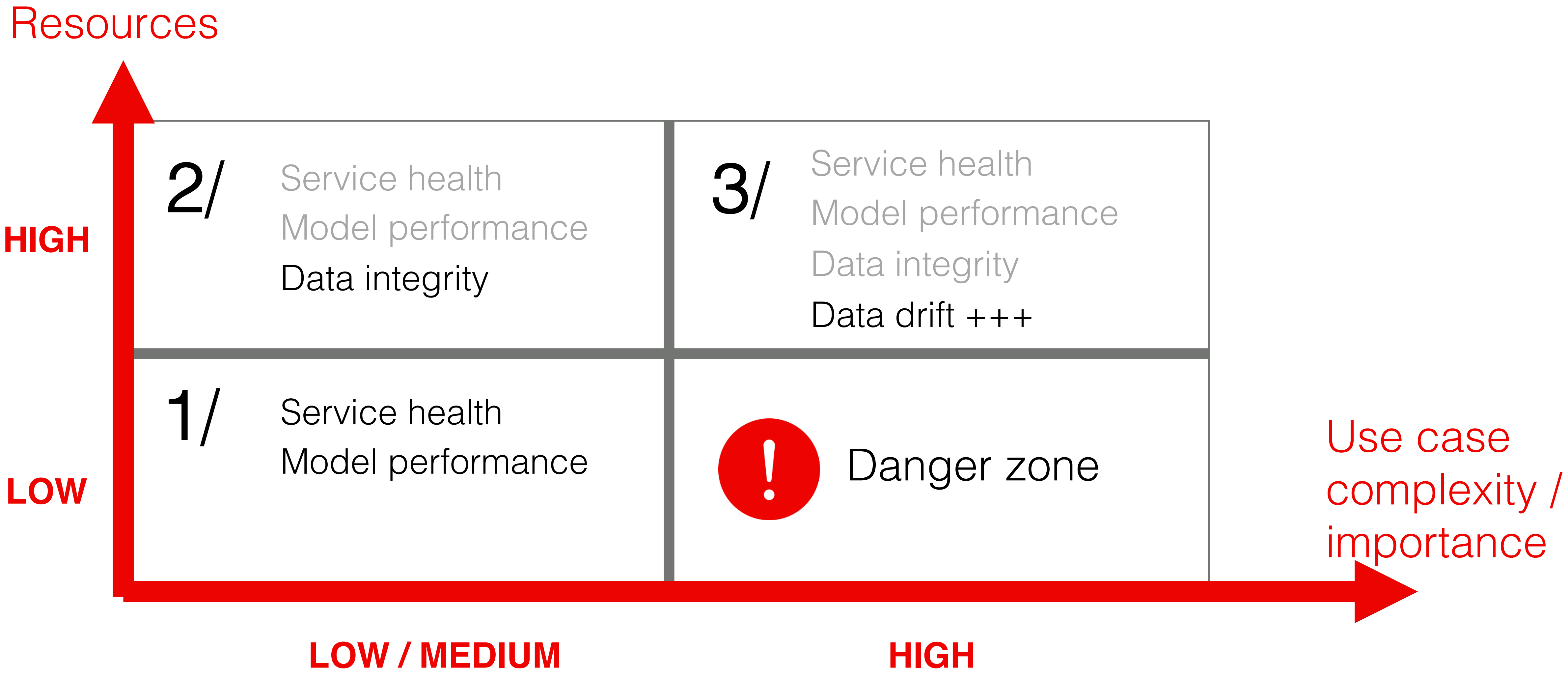


- **Example:** “feels like temperature” feature
- Model trained during summer, but applied in autumn

Comrehensive Monitoring: More Things to Look for

- [x] Service health
- [x] Model performance
- [x] Data quality and integrity
- [x] Data and concept drift
- [ ] Performance by segment
- [ ] Model bias / fairness
- [ ] Outliers
- [ ] Explainability

Pragmatic Approach: Summing Up



Машинное обучение: мониторинг моделей в production

Спасибо!

Эмели Драль