Towards Observability for Machine Learning Pipelines

Monitoring Streaming ML with Feedback Delays

Shreya Shankar, UC Berkeley 🐻
February 2022

Agenda

- Dealing with machine learning (ML) pipelines sucks
- Shift recap & existing methods
- Toy ML task introduction
- Monitoring challenges & solution ideas

Dealing with ML Pipelines Sucks 🤮

Production ML

An on-call engineer's biggest nightmare 😱

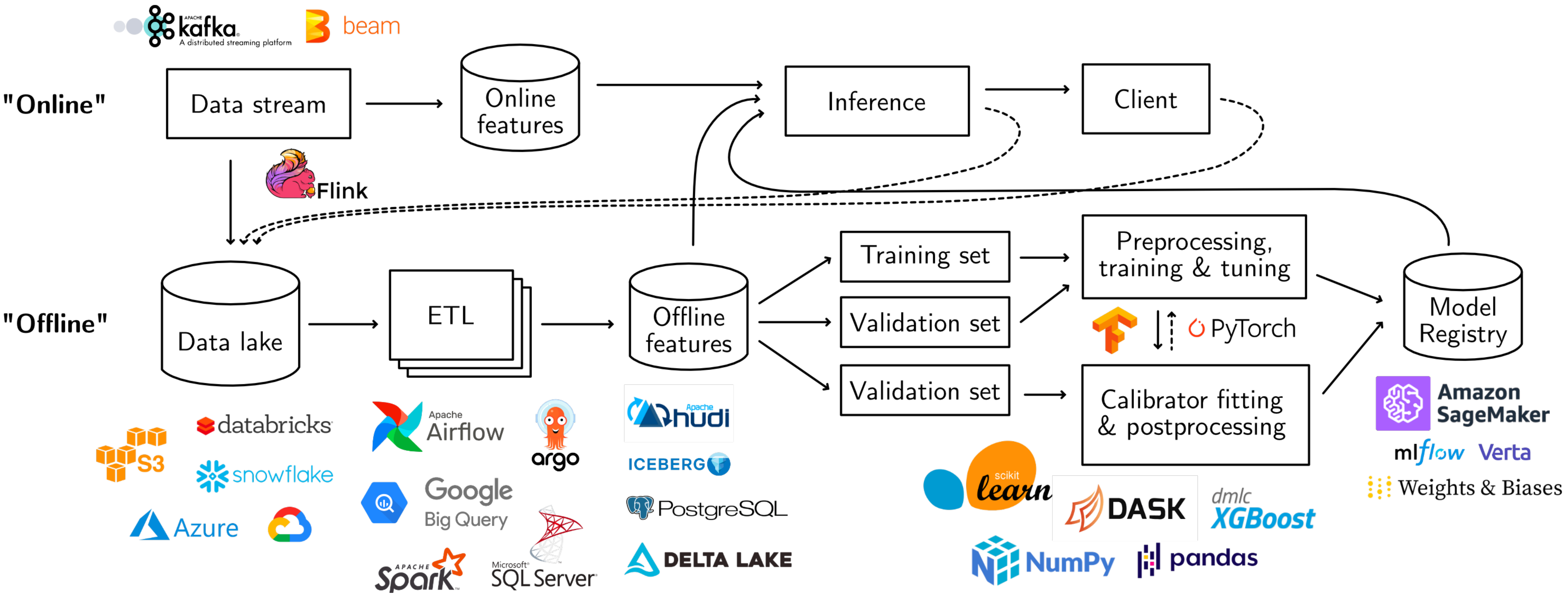


Figure 1: High-level architecture of a generic end-to-end machine learning pipeline. Logos represent a sample of tools used to construct components of the pipeline, illustrating heterogeneity in the tool stack. *Shankar et al. 2021*

Production ML

An on-call engineer's biggest nightmare 😱

- Many problems arise post-deployment
  - Corrupted upstream data
  - Model developer is on leave
  - Training assumptions don't hold in practice
  - Data "drifts" over time
  - And more...

Why Observability?

- Can't catch all bugs before they happen, but we want to *minimize downtime*
- We should:
    - Help engineers *detect* bugs
    - Help engineers *diagnose* bugs
- Need to support a wide variety of skill sets
    - Engineers, data scientists, etc.

Types of ML Data Management Solutions

Pre-training

- What do I need to start training a model?
- Feature stores, ETL pipelining, etc

Experiment Tracking

- What's the best model for a pipeline?
- `mlflow`, `wandb`, etc

Observability

- There's a bug in my pipeline; where is it?
- Real-time ML performance monitoring

Real-Time ML Performance Monitoring: Background

Why is this Hard?

Data "shifts"... 🤔

- Determining real-time performance requires *labels*
  - ...which are not always available post-deployment
- Is performance drop temporary (e.g., seasonal) or forever?
- Degenerate feedback loops
  - I.e., when predictions influence feedback (which labels are extracted from)

Shift Primer

Notation 🎶

- X is feature (covariate) space, Y is label space
- P(X): distribution of features
- P(Y): distribution of labels
- P(Y | X): distribution of labels given specific features
    - *This is what ML models are trying to learn!*

Shift Recap

Terminology 🧑‍🏫

- Covariate shift
  - $P(Y \mid X)$ is the same but $P(X)$ changes
- Concept shift
  - $P(Y \mid X)$ changes but $P(X)$ is the same

Existing Methods for Tackling Shift

Levels of sophistication 👸

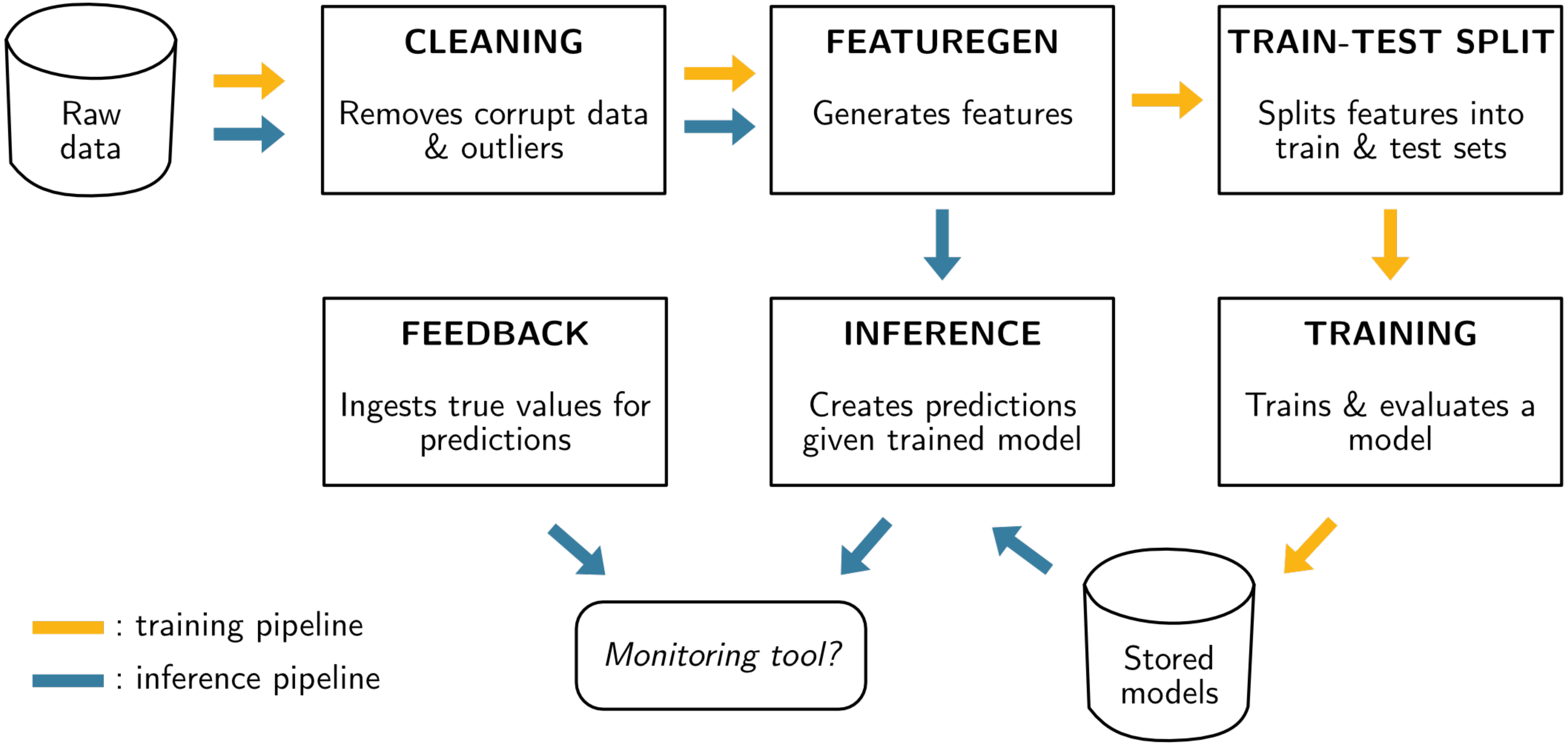
- Straw-man approach 🥤
  - Tracking means & quantiles of features and outputs
- "I took a stats class" approach 🧑‍🎓
  - Tracking MMD, KS & Chi-Square test statistics, etc
  - `alibi-detect`
- **Both approaches are label-unaware and don't use all the information we have. Can we do better?**

Toy ML Task: Running Example 🪀

Task familiarization 🚕

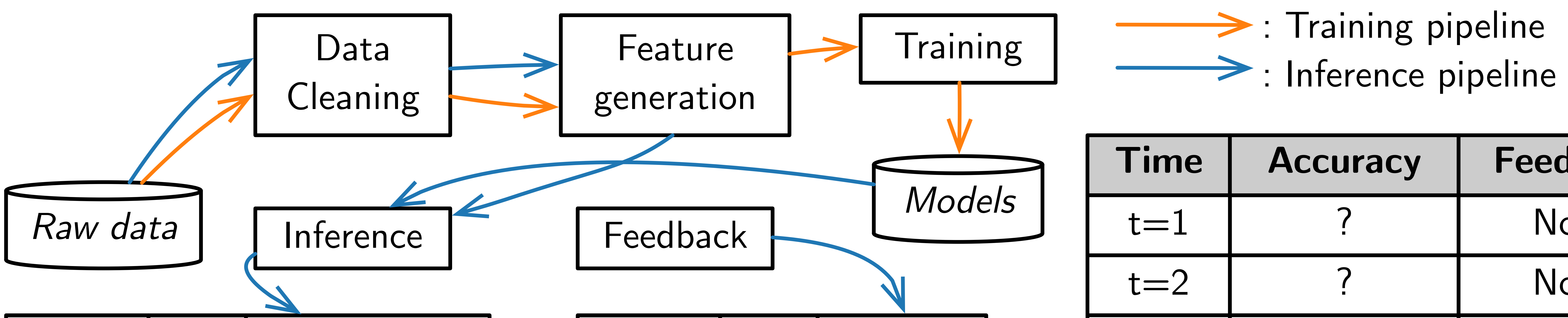
- Binary classification task: predict whether a passenger in a NYC taxi ride will give the driver a "reasonable" tip (>10% of fare)
- Using NYC Taxi & Limousine Commission public dataset
- Using `pd.DataFrame` and `sklearn` Random Forest Classifier
- Evaluating **accuracy**

Pipeline familiarization 👷



Pipeline familiarization 👷🏽

Feedback Delays Impact Accuracy



Time	ID	Prediction
t=1	A	0.7
t=2	B	0.6
t=3	C	0.9

$\bowtie$

Time	ID	Label
t=4	A	1
t=5	C	1
t=6	B	0

=

Time	Accuracy	Feedback
t=1	?	None
t=2	?	None
t=3	?	None
t=4	100%	Partial
t=5	100%	Partial
t=6	67%	Full

Shift Primer

Examples 🧑‍🏫

- X = features (e.g., location), Y = labels (high tip indicator)
- Covariate shift
  - More taxi rides in Midtown area around NYE 🎆
- Concept shift
  - Heavy construction in certain areas causes people to tip less 🚧

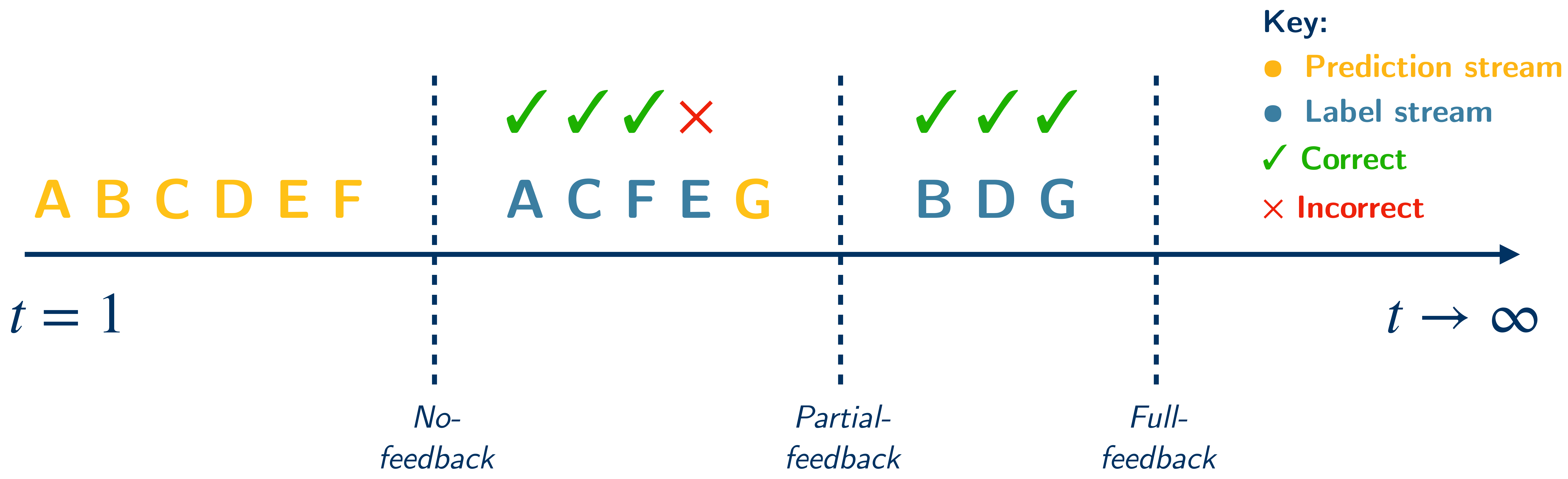
Real-Time ML Performance Monitoring: Challenges 👷

Challenge Tree 🌳

- "Coarse-grained" monitoring: *detecting* performance issues with label delays
  - Full-feedback, no-feedback, and partial-feedback cases
- "Fine-grained" monitoring: *diagnosing* performance issues
  - Teasing out engineering issues from data shift

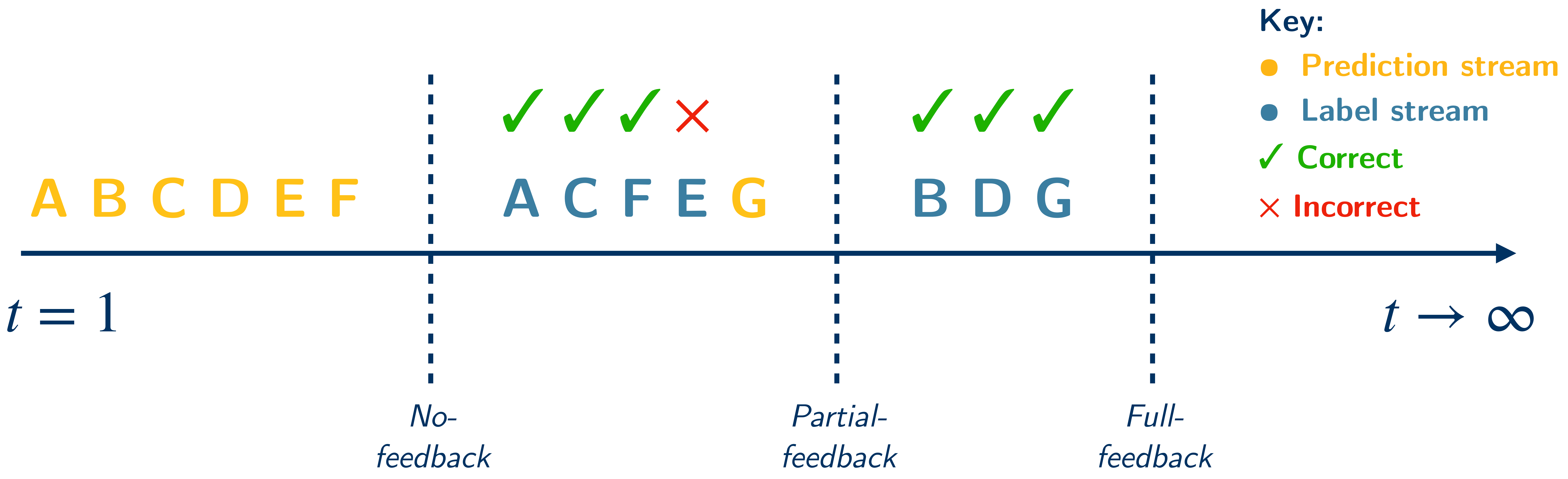
Coarse-grained Monitoring 🔎

***Detecting* performance issues: Feedback Delays**



Coarse-grained Monitoring 🔎

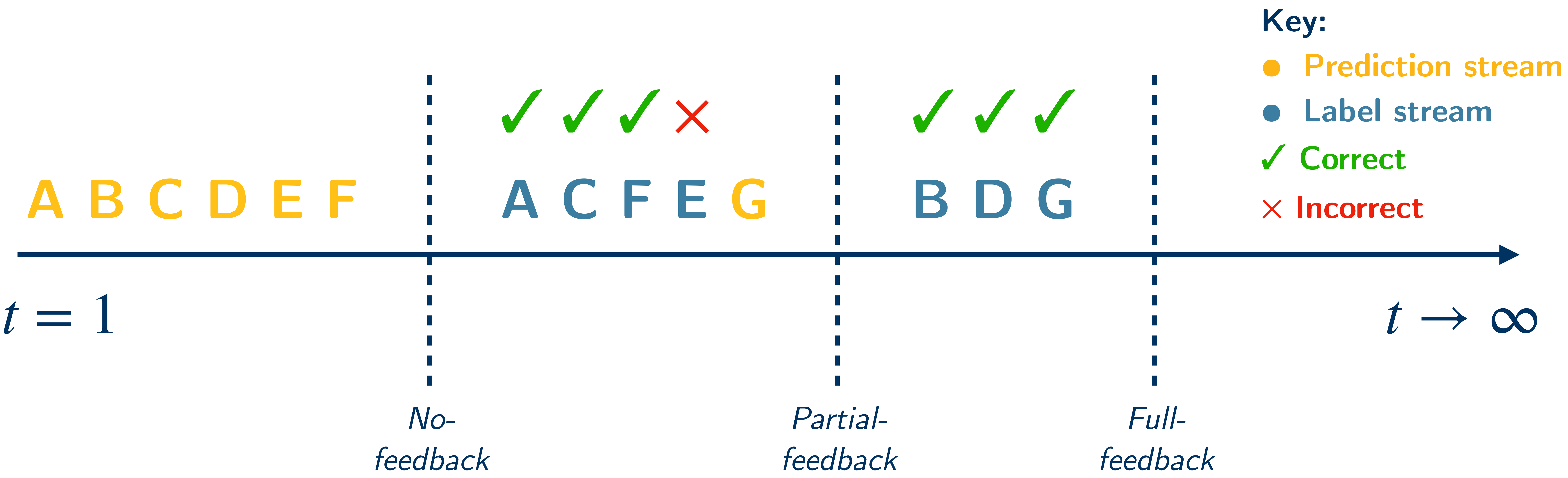
Detecting performance issues: Feedback Delays



Accuracy: ?? 🤔

Coarse-grained Monitoring 🔎

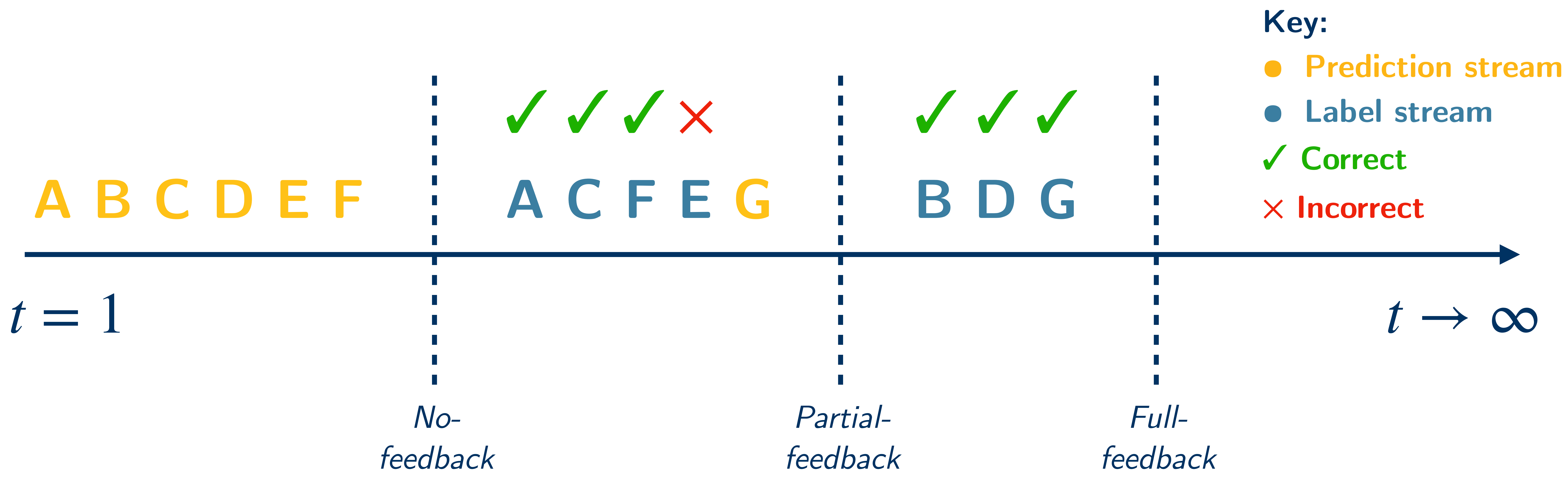
Detecting performance issues: Feedback Delays



Accuracy: 75% ?? 🤔

Coarse-grained Monitoring 🔎

Detecting performance issues: Feedback Delays



Accuracy: 86% 😀

Coarse-grained Monitoring 🔎

Detecting performance issues: full-feedback 📫

- # predictions made = # labels received
- Simplest case
  - 1) Do streaming join on predictions & feedback
  - 2) Compute accuracy on result
- What if...data is too large to fit in memory?

Coarse-grained Monitoring 🔎

Detecting performance issues: full-feedback 📫

- What if...data is too large to fit in memory?
  - Approximate streaming joins
  - Uniformly subsampling streams before joins yields *quadratically fewer* resulting tuples
- Idea: stratified subsampling
  - How to construct strata?

Coarse-grained Monitoring 🔍

Detecting performance issues: full-feedback 📫

- **Problem**: randomly subsampling predictions and labels before the join yields quadratically fewer samples to compute accuracy on
- **Solution**: stratified sampling
- How to construct strata/buckets?
  - **Want**: most accurate overall approximate accuracy
  - **Need**: buckets with similar prediction errors/losses

Coarse-grained Monitoring 🔎

***Detecting* performance issues: full-feedback** 📫

- **How to create dynamically evolving buckets with similar prediction errors/losses?**
- Solution ideas
  - Train decision tree to predict loss & use leaves as clusters
  - Frequent item-set or predicate search in loss "clusters"
- Lots of hyperparameters to decide 😔
- Need to constantly retrain bucket models?

Coarse-grained Monitoring 🔎

Detecting performance issues: no-feedback 📭

- Occurs immediately after deployment
- **Problem**: no labels
- **Solution**: importance-weight training bucket accuracy
    - Split train set into buckets
    - Create criteria for buckets
    - Determine training accuracy for each bucket

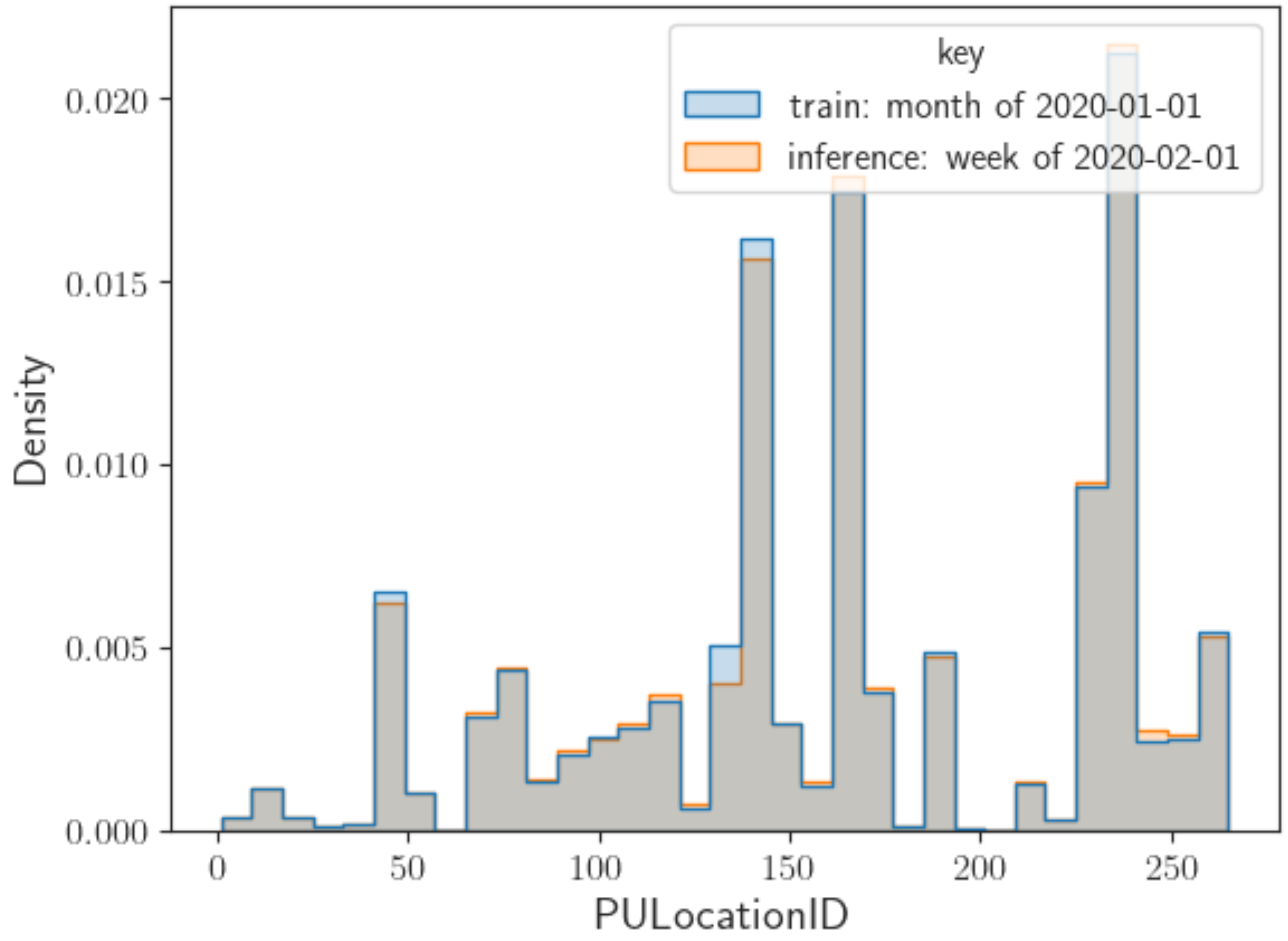
Coarse-grained Monitoring 🔎

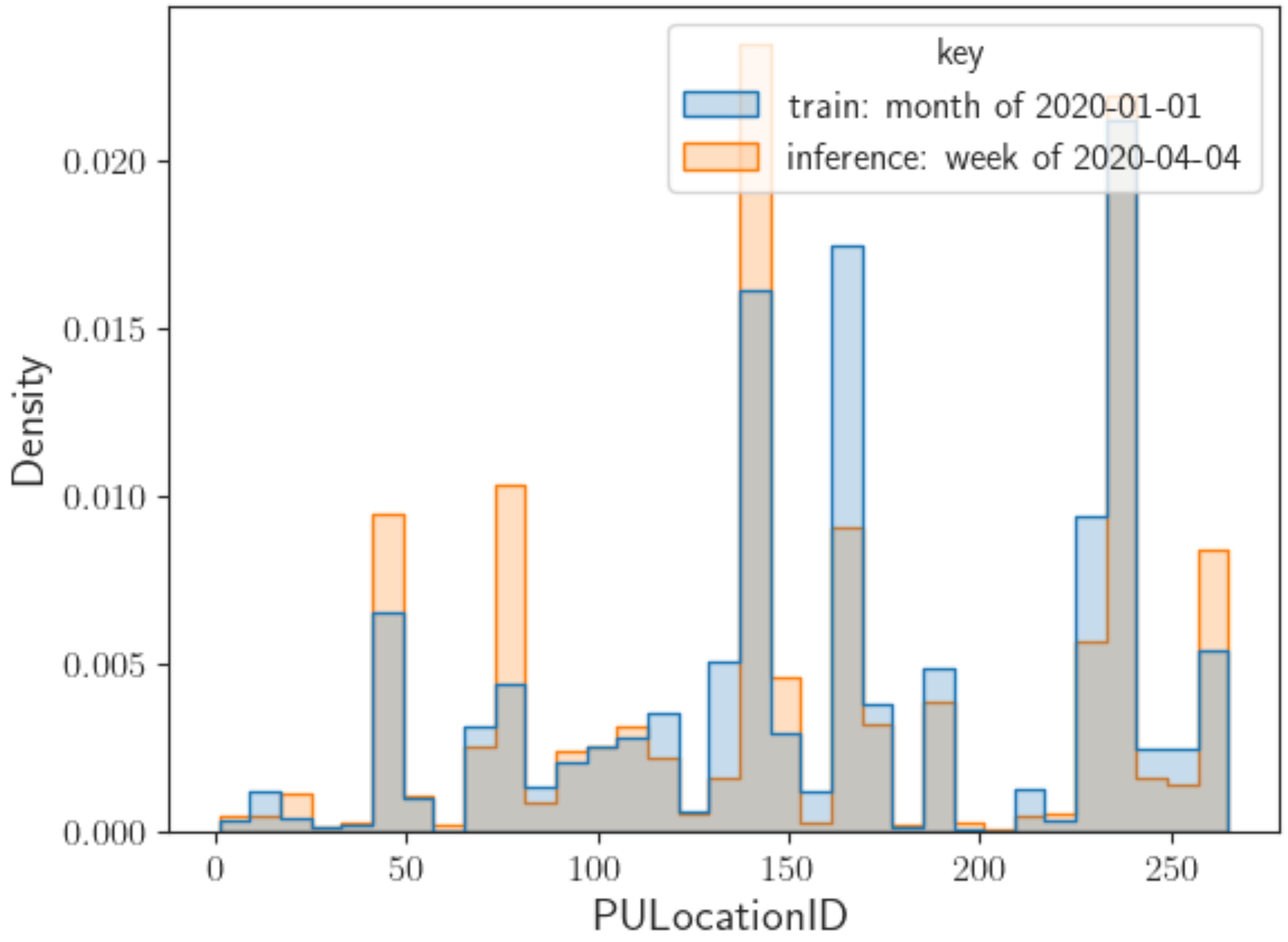
Detecting performance issues: no-feedback 📭

- At inference, classify data point (feature vector) into bucket
- Importance-weight bucket training accuracies by inference representation
- Example
  - Buckets FiDi and Midtown have accuracies of 80% and 50%
  - After deployment, we see 100 FiDi rides and 500 Midtown rides
  - Estimated accuracy $= 0.8 \times 100 + 0.5 \times 500 = \dfrac{80 + 250}{500} = \mathbf{55\%}$

Coarse-grained Monitoring 🔎

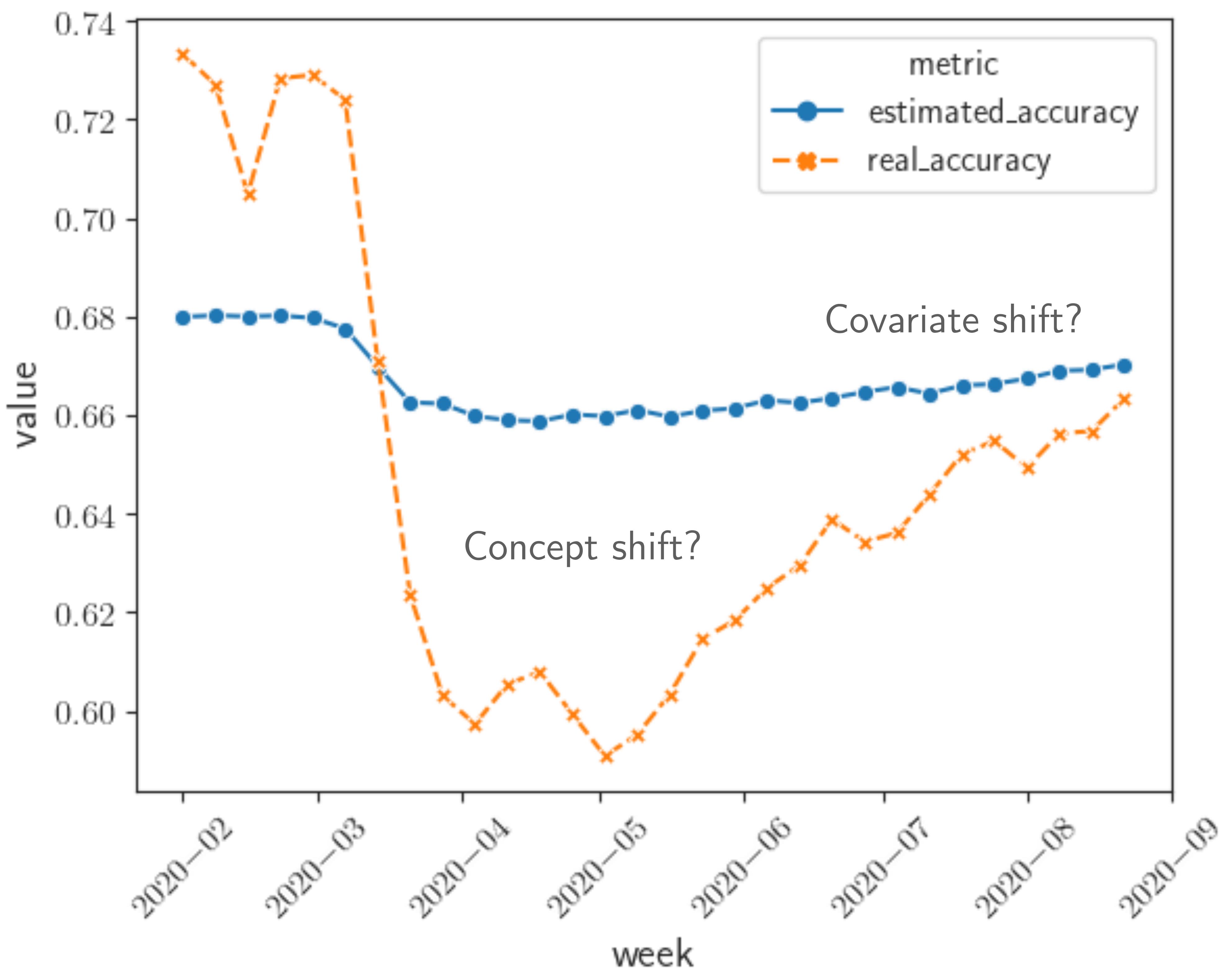
Detecting performance issues: no-feedback 📭





Coarse-grained Monitoring 🔎

Detecting performance issues: no-feedback 📭



Importance-weighted estimated vs real accuracy on a weekly basis.

Coarse-grained Monitoring 🔎

Detecting performance issues: no-feedback 📭

- **Solution**: importance-weight training bucket accuracy
- How to construct buckets?
  - **Want**: most accurate overall approximate accuracy
  - **Need**: buckets with good representation in train *and* inference sets

Coarse-grained Monitoring 🔎

Detecting performance issues: no-feedback 📭

- **How to create buckets with substantial representation in train sets?**
- Solution ideas
  - Clustering weighted combinations of train & inference sets
  - Hierarchical aggregate summaries (e.g., PASS from Liang et al.)
- Lots of hyperparameters to decide 😔
- Need to constantly recompute buckets

Coarse-grained Monitoring 🔎

Detecting performance issues: partial-feedback 📫

- Hybrid of full-feedback & no-feedback?
- Some data points have longer feedback delays than others
  - Delays aren't necessarily uniformly distributed
  - Why?
- **Additional problem**: identify groups of data points with similar feedback delays

Fine-grained Monitoring 🔎

***Diagnosing* performance issues: data quality issues** 🔧

- Instrument pipelines with data quality checks
  - Alert on missing data
  - Set upper and lower bounds for feature values
- Tedious to scale to 1000s of features 🙀
- Practitioners push DQ verification onto "shift" detection...

Fine-grained Monitoring 🔎

***Diagnosing* performance issues: towards retraining models** 🔁

- Using existing methods to compute shift doesn't work in practice
  - E.g., KS test has low p-values for O(1000) data points
  - Alert fatigue when monitoring every feature and output column
  - Seasonal & expected shifts
- **Idea: look into these statistics when coarse-grained approximated metrics are low**

Fine-grained Monitoring 🔎

***Diagnosing* performance issues: towards retraining models** 🔁

- Different shifts imply different retraining strategies, e.g.,
    - Covariate shift: augment some buckets in training
    - Concept shift: retrain on recent window
- **Research question: how to create self-adapting training sets?**

`mltrace`: Ongoing Work ⚙️

Ongoing Research Projects 🔬

- mltrace: lightweight, "bolt-on" ML observability tool in the making with projects in several research areas

Data Systems	Machine Learning	HCI
• Mitigating effects of feedback delays on real-time ML performance • Differential dataflow to compute streaming ML metrics quickly and efficiently at scale	• Creating streaming ML benchmarks • Building repository of tasks with "temporally evolving tabular data" (e.g. Ethereum gas price prediction)	• Interview study on best practices in CI / CD for ML • Visualizing large-scale data drift

Readings and Resources 📕

- Towards Observability for Machine Learning Pipelines
- The Modern ML Monitoring Mess
  - Rethinking Streaming ML Evaluation
  - Categorizing Post-Deployment ML Issues
  - Failure Modes in Existing Observability Tools
  - Research Challenges

- Contact: shreyashankar@berkeley.edu 📧