

Driving AI Quality

Achieving high performance with
ML testing and monitoring

Anupam Datta

Co-Founder, Chief Scientist and President, TruEra
Former Professor, Carnegie Mellon University

Managing ML Model Performance is a huge problem



Zillow shuts down
home-buying business;
\$500M write off

Eric Ulken

**Generative
AI brings
wrongness
at scale**

Generative AI can get more
wrong, faster -
with less transparency

**MIT
Technology
Review**

Weird pandemic behavior
messes with AI models

What we often hear

Pre-deployment

1. We peer review our models
2. We A/B test our models

Post-deployment

1. We frequently retrain our models
2. We have alerts



**World changes
over time!**

Why is frequent retraining not sufficient?

! Does not address:


- Model quality problems stemming from **concept and data drift**
 - e.g., Zillow house price estimation failure
- **Data pipeline issues**, a common source of model quality problems
 - e.g., data pipeline using stale product codes

Why is alerting alone not sufficient?

! Gaps

- Too many **false alerts** lead to alert fatigue
- Alerts do not guide **debugging** the issue

Agenda

1. **Observe and iterate**
 2. Debug rapidly
 3. Monitor LLMs
- 

Fundamental #1: Observe & Iterate

Monitor

Track performance
& alert at scale



Debug

Quickly pinpoint &
root cause issues

Iterate

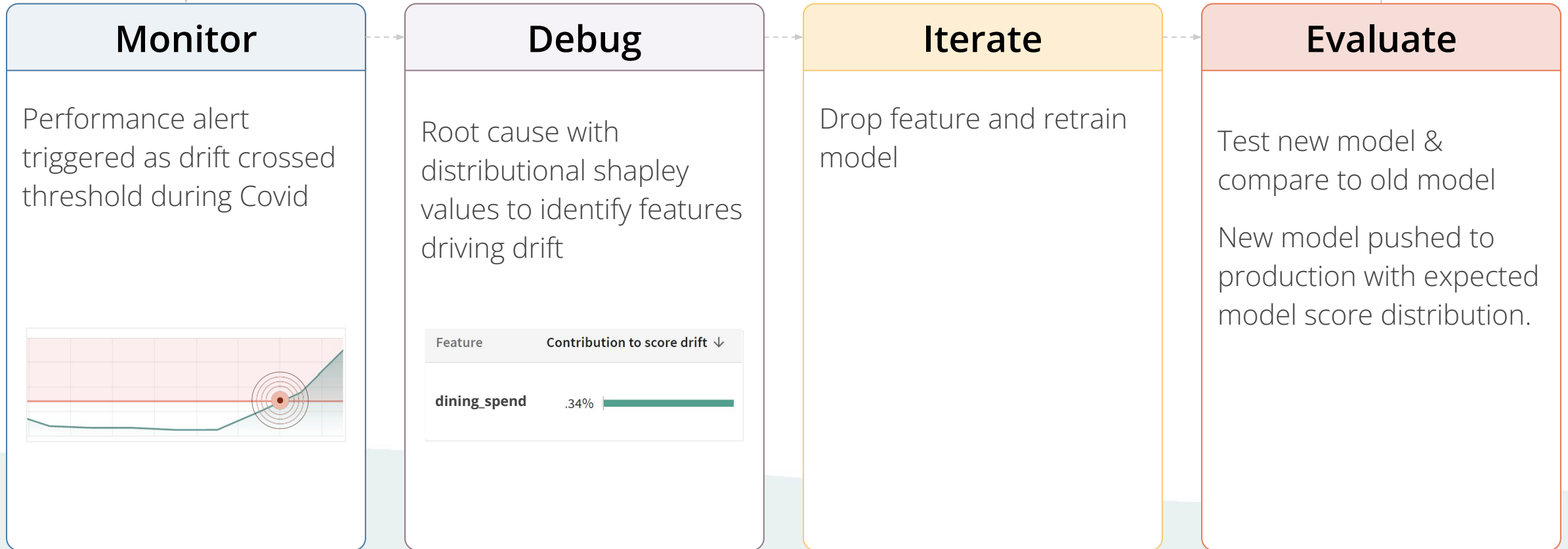
Inform retraining
& improvement

Evaluate

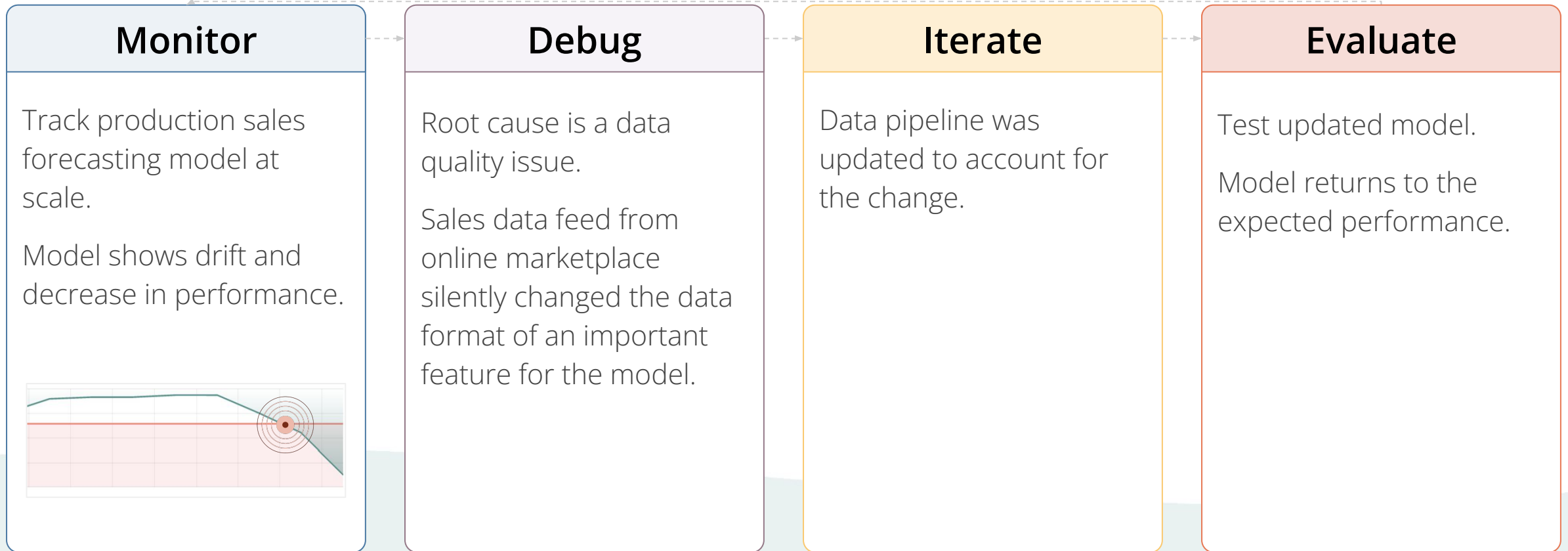
Test & verify quality
pre-launch

Address drift, data pipeline bugs, actionable alerting


Example: Addressing concept drift



Example: Addressing data pipeline issue



Agenda

1. Observe and iterate
 - 2. Debug rapidly**
 3. Monitor LLMs
- 

Track and alert at scale



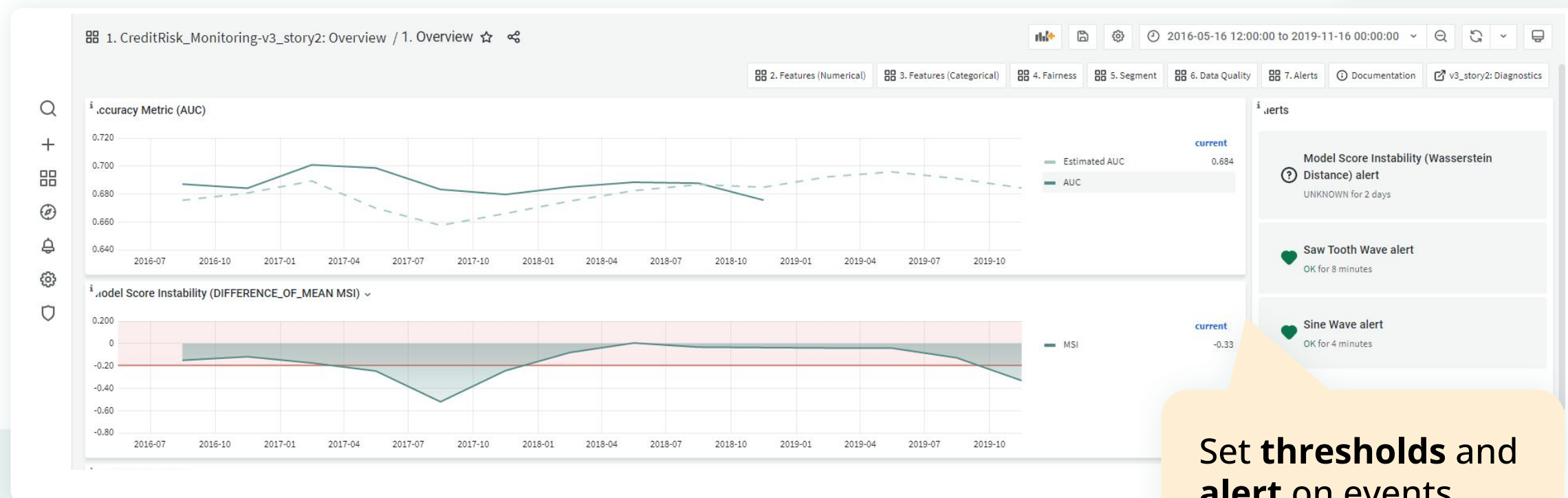
truera

Sample **metrics** across **segments**

- **Performance:** Accuracy, NDCG
- **Drift:** Difference of Means, Wasserstein Distance
- **Fairness:** Impact ratio, Equality of Opportunity

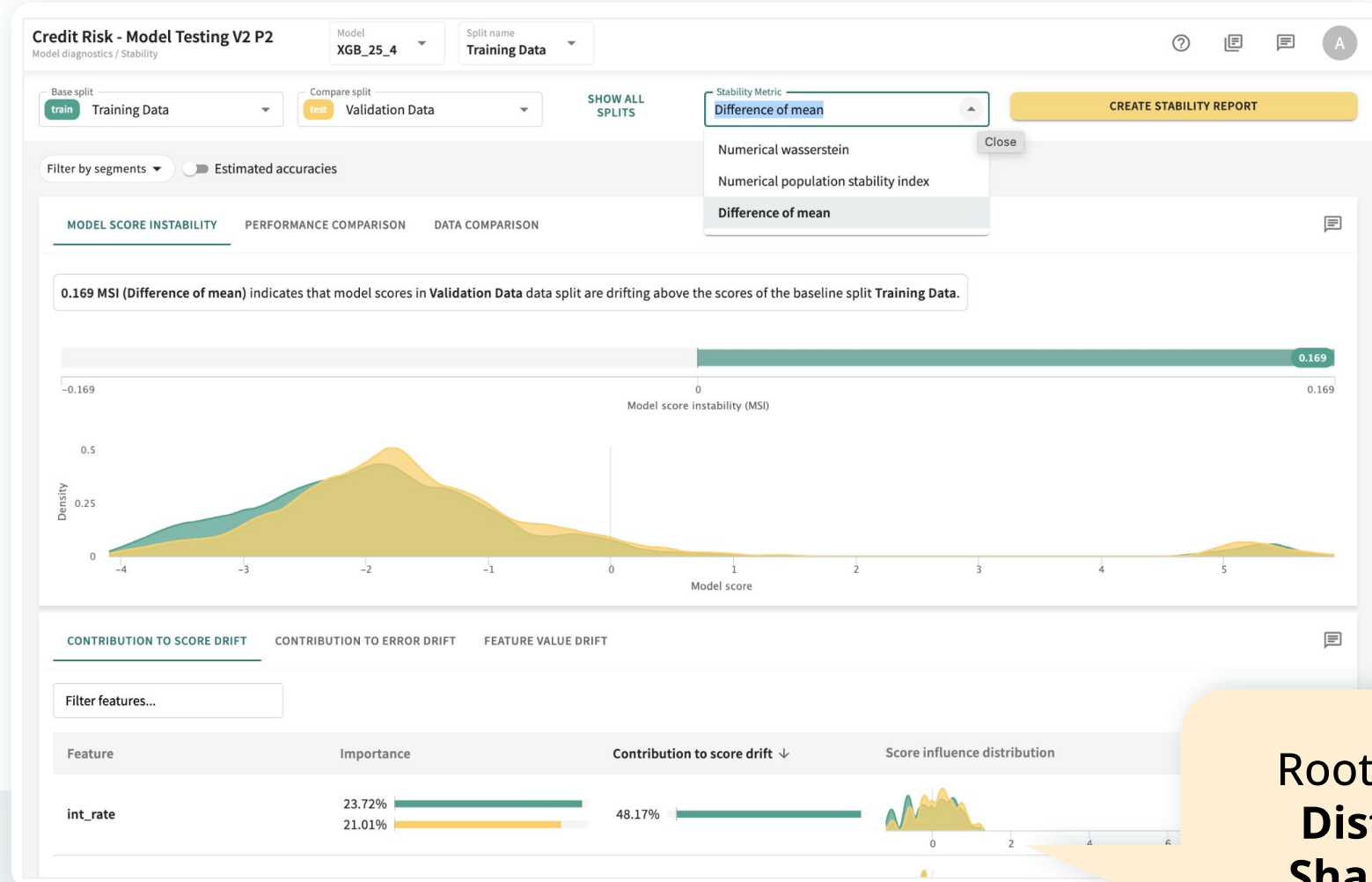
Scale with **streaming metrics & time series data store**

 **Time Series Datastore**



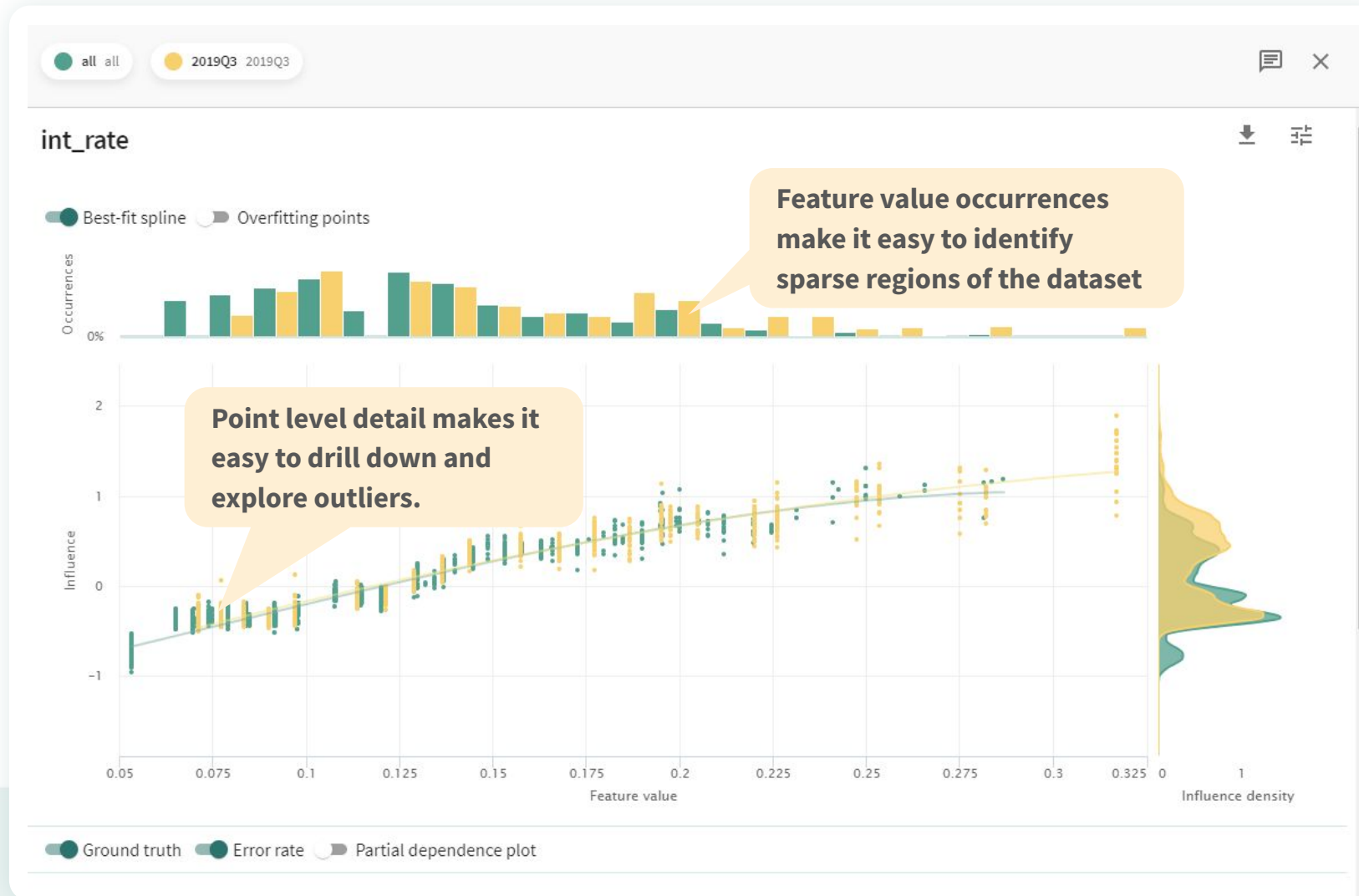
Set **thresholds** and **alert** on events

Debug with root causes



Root causes with
**Distributional
Shapley Values**

Debug with root causes



Debug common causes of model issues

Shifting data/labels

- data drift
- concept drift

Insufficient data/labels

- Underrepresentation
- Features are not rich enough for the problem

Incorrect data/labels

- Training data/labels are incorrect
- Test data/labels are incorrect

External bugs/changes

- Pipeline failures
- “State of the world” has changed
- Code changes

Model is under/overfitting

- Model is under/overparameterized
- Model has not “converged” (not trained long enough)

Model is not aligned to performance metric

- Training procedure not reflective of/optimized for performance criteria

Test and debug your ML models

The screenshot shows the Truera web application interface. At the top, there is a dark blue header with the 'truera' logo on the left and a help icon on the right. Below the header, a navigation bar contains 'MY PROJECTS' and 'TUTORIALS' (which is highlighted with a green underline). The main content area displays four tutorial cards arranged in a grid. Each card has a title, a 'Goals' section with a red icon, a list of four steps, and an 'OPEN TUTORIAL' button at the bottom. The cards are: 1. 'Starter Example - Drift' with 'Tabular' and 'Regression' tags, a teal decorative arc, and goals including stability tests and drift mitigation. 2. 'Starter Example - Fairness' with 'Tabular' and 'Classification' tags, a yellow-green decorative arc, and goals including fairness evaluation and mitigation. 3. 'Starter Example - Explainability' with 'Tabular' and 'Classification' tags, a blue decorative arc, and goals including feature importance and influence sensitivity. 4. 'Starter Example - Performance' with 'Tabular' and 'Regression' tags, an orange decorative arc, and goals including performance tests and mitigation.

truera

MY PROJECTS TUTORIALS

Starter Example - Drift Tabular Regression

Goals

1. Set up and view the results of stability tests.
2. Debug the root cause of drift.
3. Mitigate drift using techniques including data augmentation and feature removal.
4. Retest the new model and confirm the effectiveness of the mitigation strategy.

OPEN TUTORIAL

Starter Example - Fairness Tabular Classification

Goals

1. Train and ingest a credit-worthiness model and evaluate its fairness.
2. Diagnose potential root causes of disparate impact.
3. Mitigate fairness issues using techniques including feature removal and class rebalancing.
4. Retest the new model and confirm the effectiveness of the mitigation strategy.

OPEN TUTORIAL

Starter Example - Explainability Tabular Classification

Goals

1. Understand Global and Local Feature Importance
2. Interpret Influence Sensitivity Plots

OPEN TUTORIAL

Starter Example - Performance Tabular Regression

Goals

1. Set up and view the results of performance and feature importance tests.
2. Diagnose the root cause(s) of performance issues.
3. Mitigate these issues using techniques including regularization and model pruning.
4. Retest the new model and confirm the effectiveness of the mitigation strategy.

OPEN TUTORIAL

<https://app.truera.net/>

Agenda

1. Observe and iterate
2. Debug rapidly
- 3. Monitor LLMs**

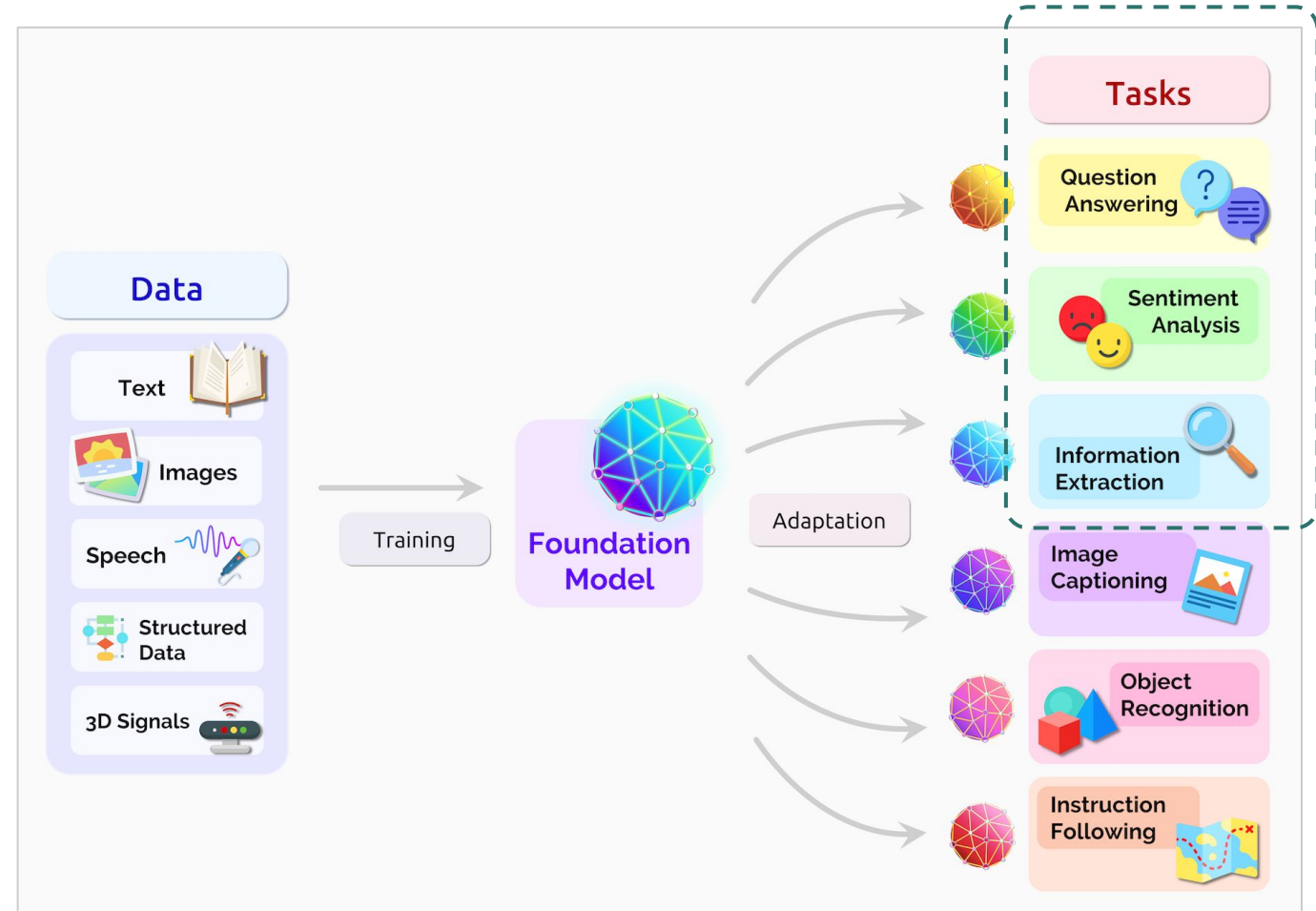
Foundation models are everywhere

Sample language foundation models

- BERT, RoBERTa, DistillBERT,
- ChatGPT, GPT-3, LaMDA, BLOOM

Adaptation uses:

- your data to
- build a model for
- your specific task



Examples of language model use cases

State of play

- BERT-style models are extensively used in practice
- GPT-style model use starting to take off

Traditional use cases

- Customer engagement analytics
- Search & recommenders
- Document processing
- Sales engagement analytics
- Conduct risk (e.g., trader misconduct, misselling)
- ...

Generative use cases

- Customer service chat-bots
- Internal help desks (e.g., HR, IT)
- Marketing copy
- Sales reach outs & follow ups
- Text summaries
- ...

Monitoring LLMs at scale

Model Performance: Satisfaction



How do you measure quality?

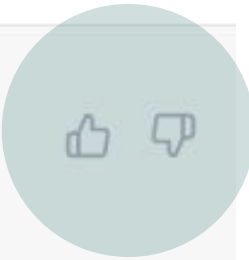


This isn't a great service.

Sentiment
analysis on
prompts



I'm sorry to hear that you feel that way. As an AI language model, my goal is to provide helpful and informative responses to your questions or concerns. Can you please provide more specific feedback about what you find lacking in my service? This will help me understand how I can improve and better meet your needs.



Track user
responses

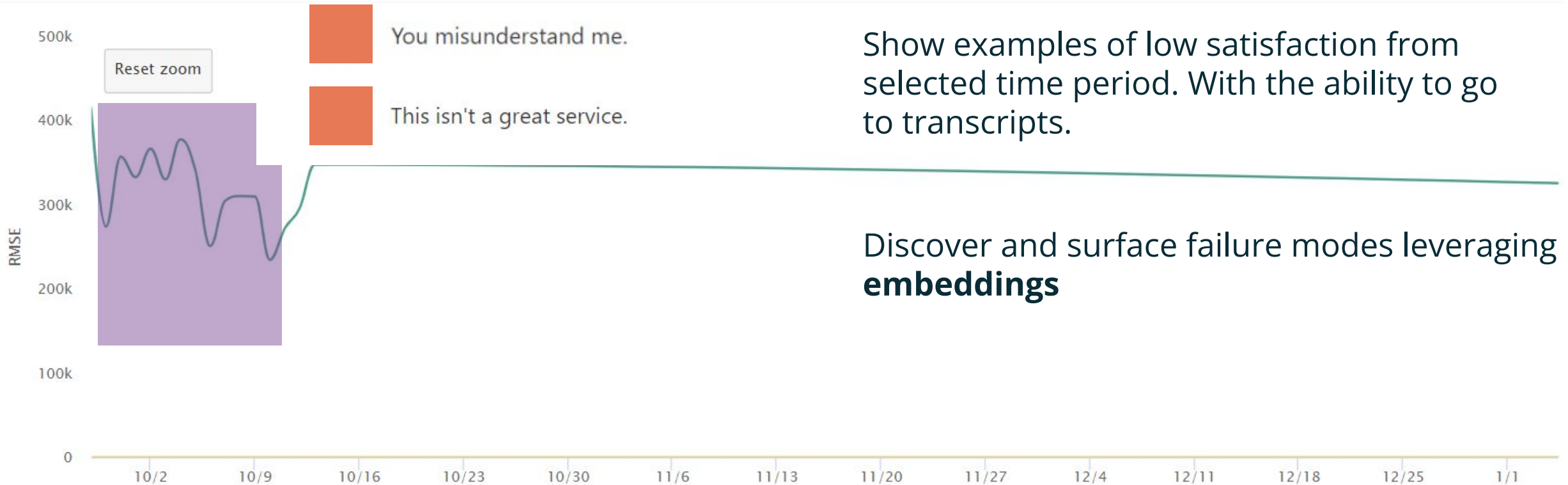


You misunderstand me.

More generally: Create a notion of **feedback function** (like labeling functions) to monitor on an ongoing basis.

How do you debug?

Model Performance: Satisfaction



Show examples of low satisfaction from selected time period. With the ability to go to transcripts.

Discover and surface failure modes leveraging **embeddings**

Key takeaways

1. Observe and iterate
2. Debug model performance, drift, fairness issues rapidly
3. Monitor LLMs with feedback functions

Thank you!

Test and debug your ML models with TruEra.
Free signup at <https://app.truera.net/>

TruEra booth C50: for demos,
teddy bears and tech deep dives.

