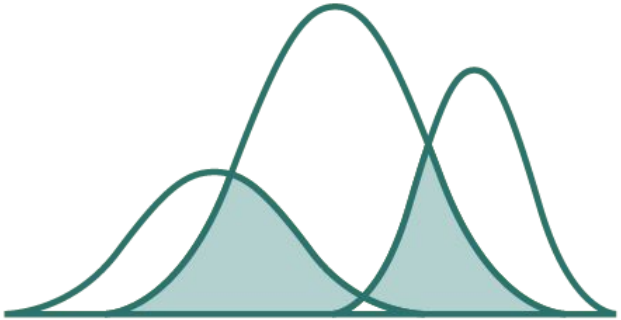
AI Monitoring & Explainability: The Critical Hidden Connection

Anupam Datta

Co-Founder, President, Chief Scientist

TruEra

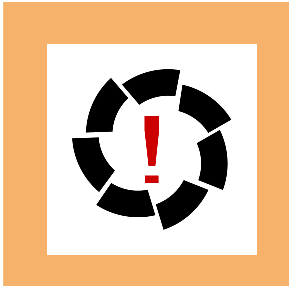


What people think ML Monitoring is like…

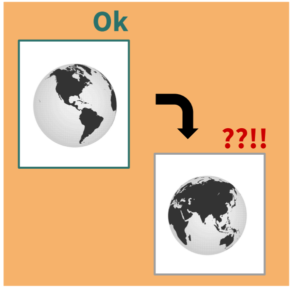
and what it's actually like.

A lot can go wrong.

Data Bugs



Unforeseen Changes



New, untrained use cases

Shifting concepts & behavior

Adversarial attacks

The harsh reality of ML.

The moment you put a model in production, it goes on a wild ride.

So monitoring is key.

Monitoring is not that easy today.
Data Science and ML Ops teams struggle to minimize ML risk.

There's a wild goose chase going on.

How can I better understand how my models are working?

How do I identify real problems with the model?

What is the problem's root cause? How can I debug quickly?

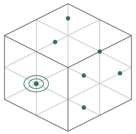
How can I make monitoring work easily in my environment?

Teams struggle with:

- Visibility and observability
- Diagnosis and actionability
- Complex environments and workflow (diverse models, diverse stakeholders)

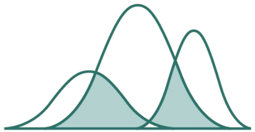
Monitoring Requirements

Fast, precise, and complete.



Broad coverage of model & data quality metrics

Fast, precise debugging

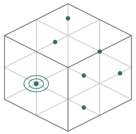


Easy to deploy and scale

AI Monitoring & Explainability:
The Critical Hidden Connection

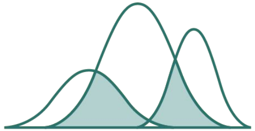
Focus Today: Monitoring Requirements

Fast, precise, and complete.



Model Drift & Performance Metrics

Fast, precise debugging with root cause analysis



Easy to deploy and scale

AI Monitoring & Explainability: The Critical Hidden Connection

Outline

- Overview
  - Why does drift happen?
  - What are different kinds of drift?
  - What is consequential drift?
- How to identify drift?
  - Measures
  - Challenges
- How to mitigate drift?
- Monitoring

Overview of Drift

Overview of Drift

Bikes used to look like this

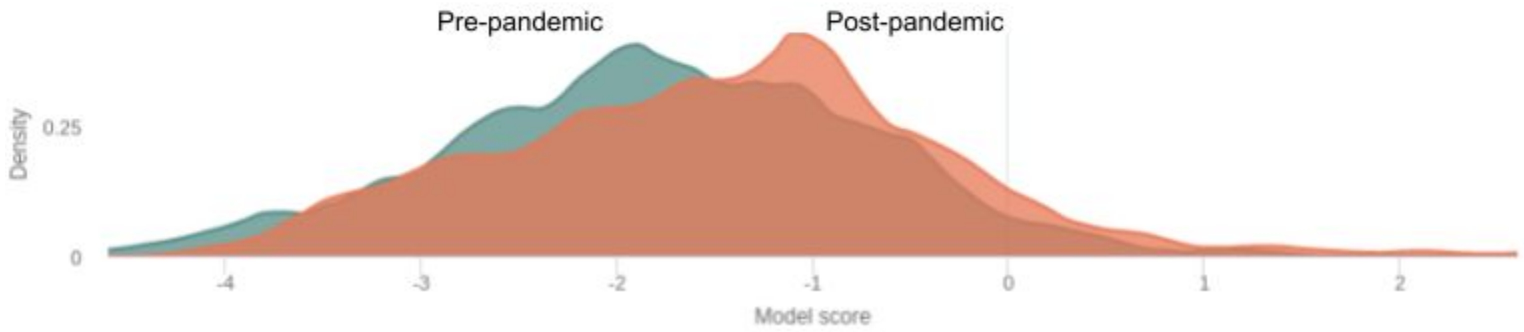
… now they look like this

Will an ML model trained on images like the left continue to work well?

Overview of Drift

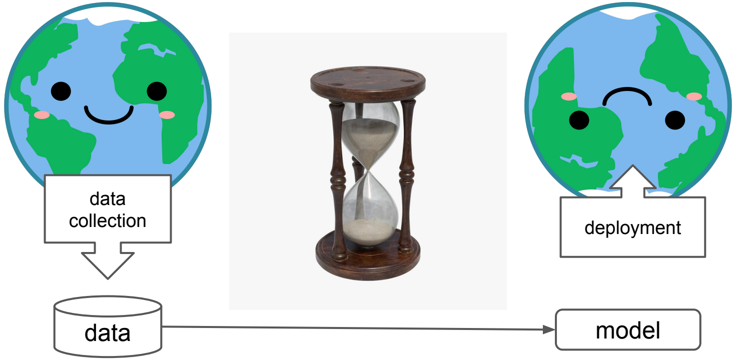
This is similar to what happened to models with Covid.

Example: risk scoring model. Lower model score shows lower risk.



Will an ML model trained on pre-pandemic data continue to work well?

Overview:
Why does drift happen?



Data quality issues

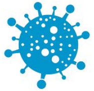
Examples: NaN

- Broken feature pipelines

The External World Has Changed

Examples:

- The pandemic
- Housing market fluctuations



Zillow

Model Applied to a New Context

Example:

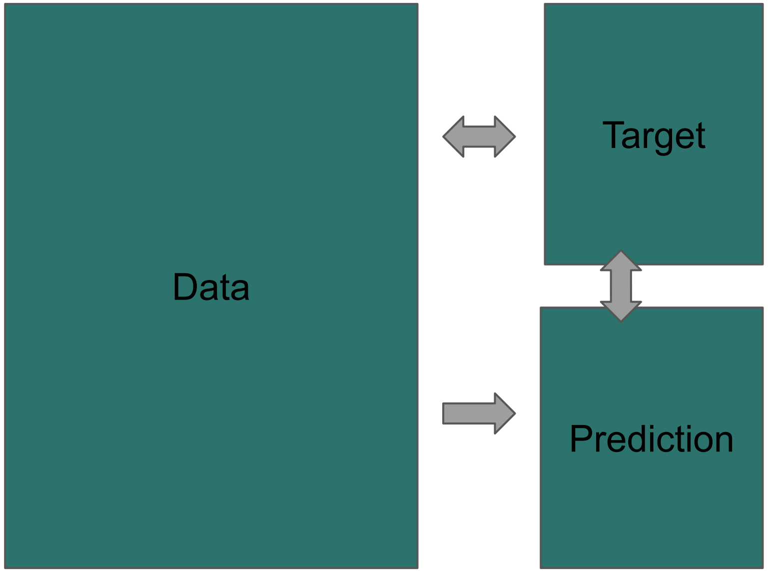
- Model trained on Wikipedia applied to news articles

Collected Training Data Is Different

Example:

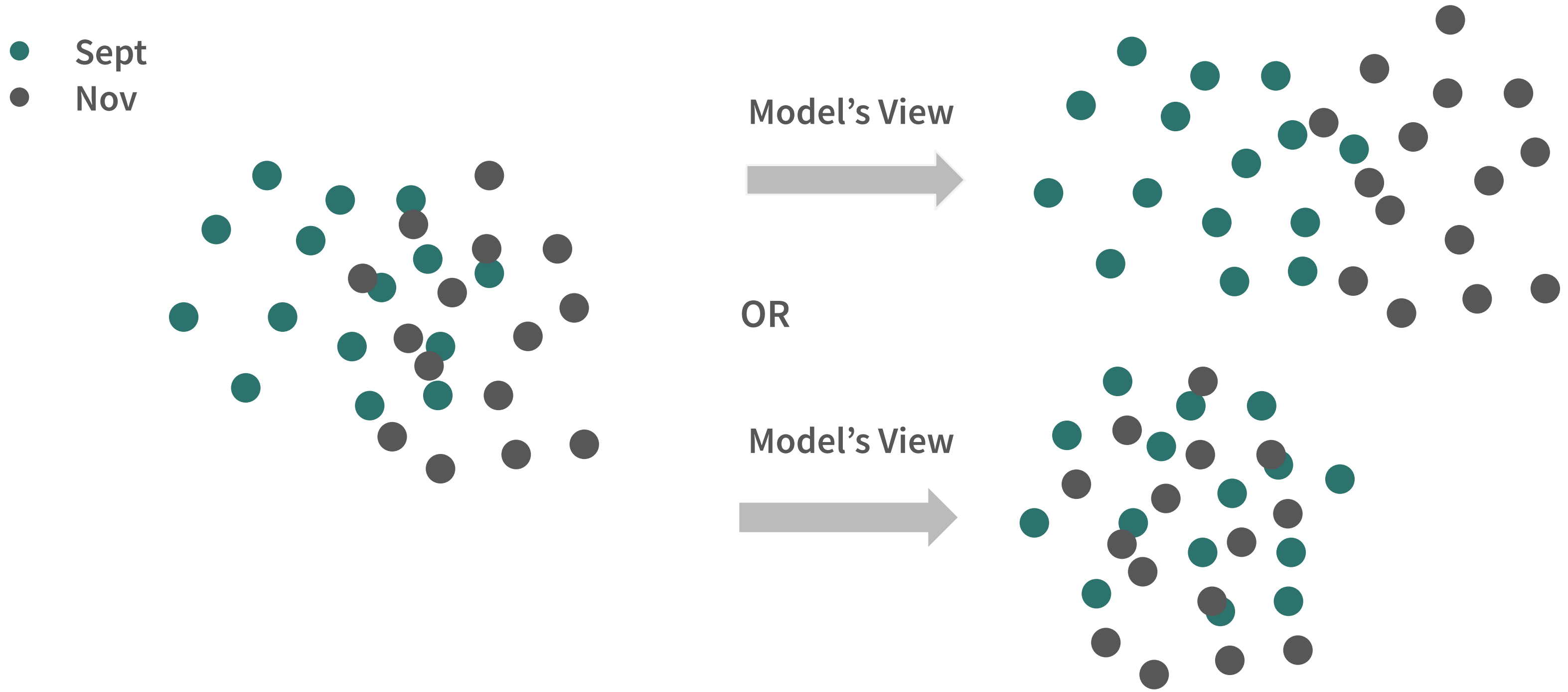
- For credit decisions, labels are only available for approved applicants
- Impact of your models on the data

Overview: What are the Different Kinds of Drift?



1. Data drift
    a. Covariate shift -- drift in input features
    b. Concept drift -- drift in relationship between input and target
2. Model decay -- performance loss due to data drift
3. "Prediction shift" -- drift in model predictions

Overview: Which Drifts are Consequential and Why?



High-dimensional data always drifts (curse of dimensionality)

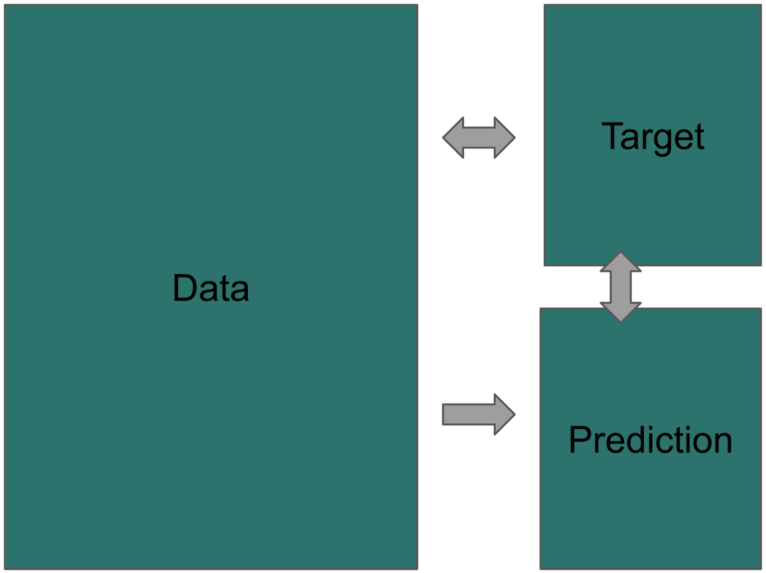
… but not necessarily in ways that affect the model

How to identify drift?

Standard Approaches To Measuring Drift

- Measure model performance in deployment

- Compare distributions of
  - Ground Truth
  - Single input features
  - Prediction
  - Full data sets



Challenges with Standard Approaches

- Measure model performance in deployment
- Compare distributions of
  - Ground Truth
  - Single input features
  - Prediction
  - Full data sets

Don’t have ground truth in many cases.

Does a 5% shift in feature 28 matter?

Why is the prediction shifting?

Curse of dimensionality

How to mitigate drift?

example scenarios

Key Takeaway: How to mitigate drift?

Blind model retraining is often not the best answer to counter drift.

Step 1: Understand root causes of drift:

Where is it happening?
When is it happening?
How much is there?
What is causing it?

Monitoring & Explainability – The Critical Hidden Connection!

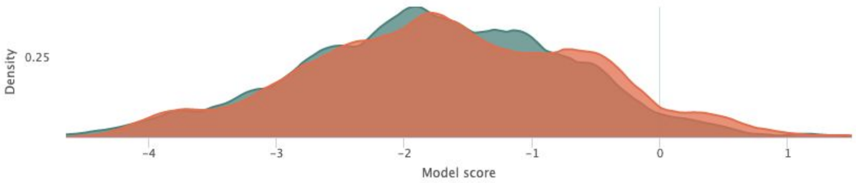
Step 2: Understanding the root cause of drift leads to targeted ways to address drift

How to mitigate drift?

Is the drift caused by an unstable feature?

- Identify and address cause (of prediction drift).
  - Remove a feature without retraining (i.e. replace with mean/mode).
  - Remove a feature and retrain with existing data.

Feature	Importance	Feature value drift ↓	Feature values distribution
annual_inc	1.95% 2.14%	19156.559	
revol_bal	2.65% 2.95%	6702.749	
total_bc_limit	3.68% 3.92%	759.644	



drift in input features → drift in model output

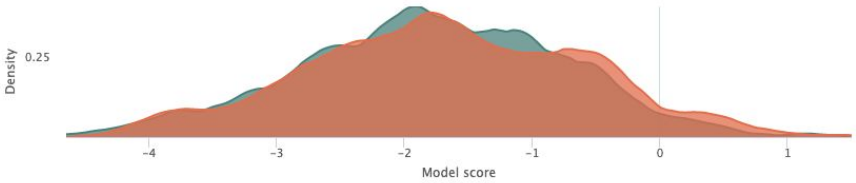
How to mitigate drift?

Is the drift caused by an unstable feature?

- Explainability technology under the hood
  - Feature importances based on Shapley Values, gradients & more

Topic for
Q & A

Feature	Importance	Feature value drift ↓	Feature values distribution
annual_inc	1.95% 2.14%	19156.559	
revol_bal	2.65% 2.95%	6702.749	
total_bc_limit	3.68% 3.92%	759.644	

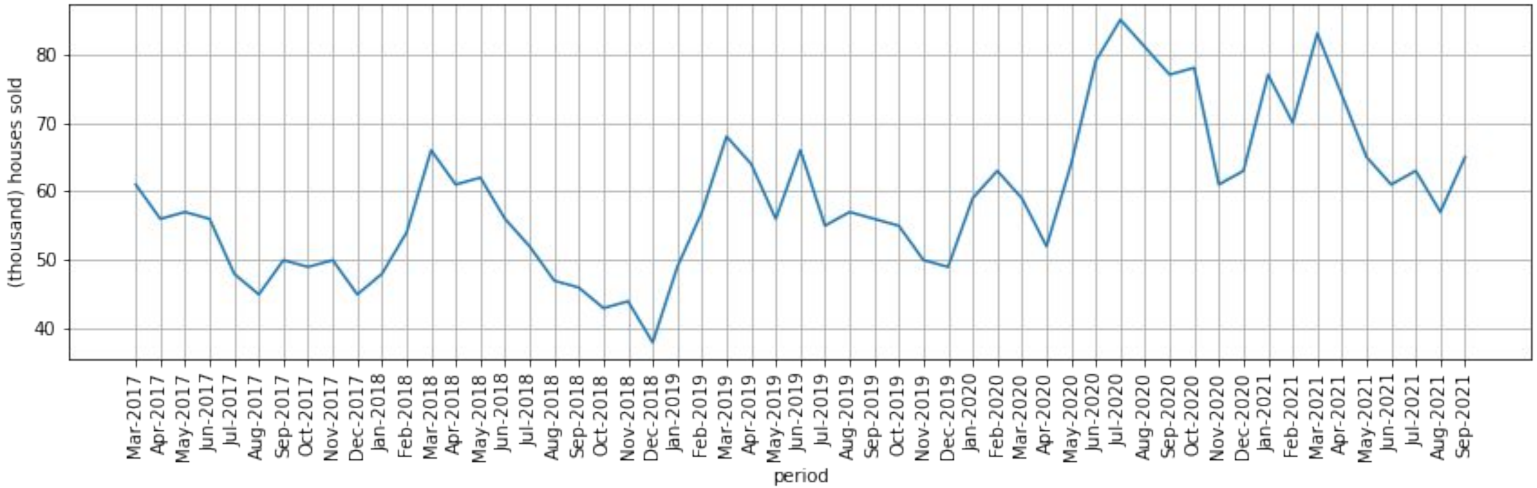


drift in input features → drift in model output

How to mitigate drift?

Is the drift periodic or learnable?

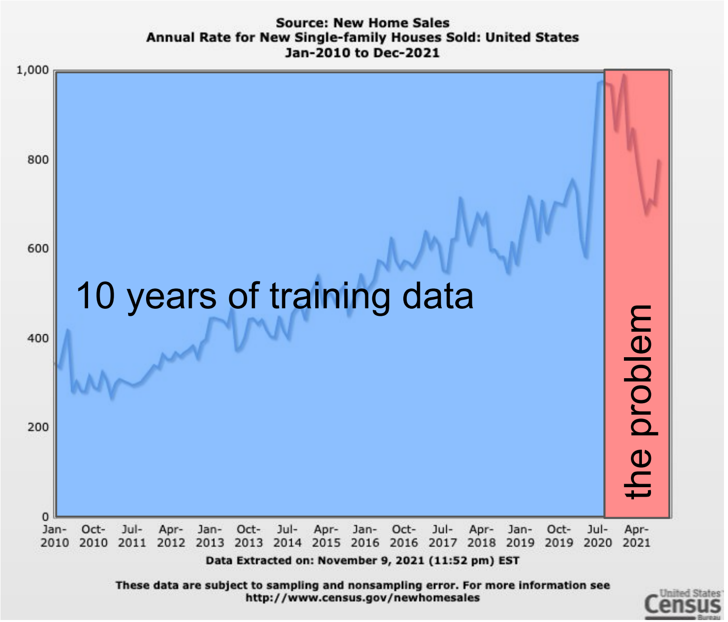
- Concept drift --> Covariate drift with feature engineering
  - Add features to learn periodic change over time.
  - Add indicators of effects of unexpected events ("is-covid" vs "unemployment-rate")
  - Might not need labeling additional data.



How to mitigate drift?

Is the drift sudden relative to training period?

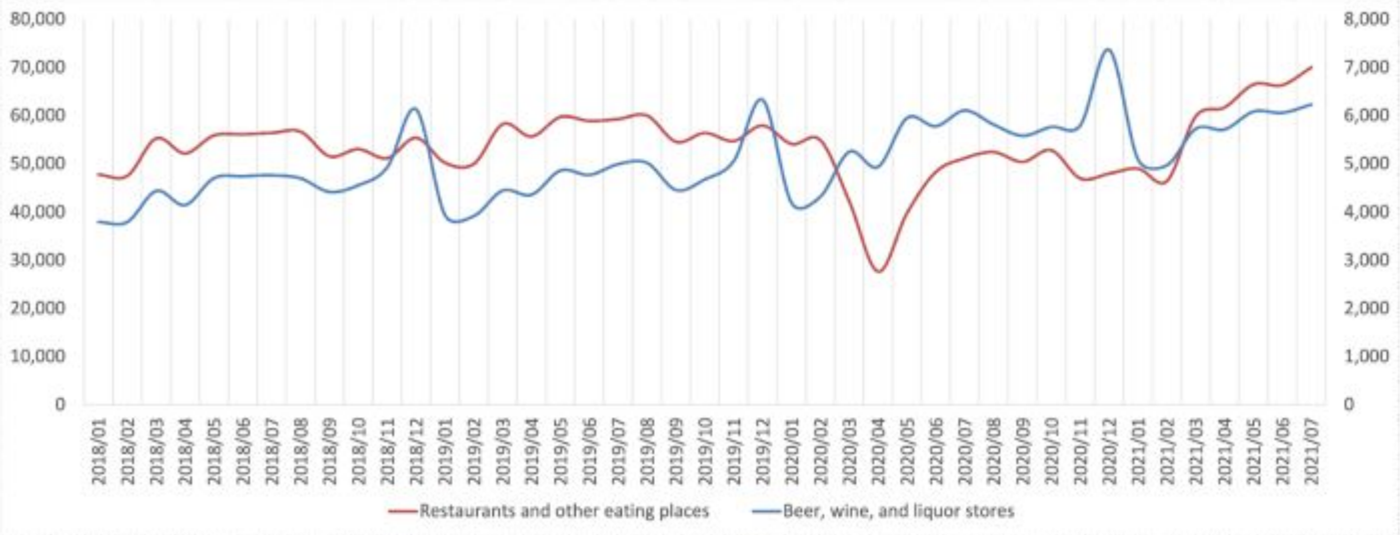
- Drift period may be too insignificant for a retrained model to pick up on it.
- Options:
  - Upweight recent data.
  - Fine tune model with recent data.
  - Identify new features that can help generalize to newer data
    - Example: newer data might be characterized by lower interest rates which might not have been predictive before



What can we do about drift?

Is the drift periodic or learnable?

- Concept drift --> Covariate drift with feature engineering
  - Add features to learn periodic change over time.
  - Add indicators of effects of unexpected events ("is-covid" vs "unemployment-rate")
  - Might not need labeling additional data.



How to mitigate drift?

Is the drift significant enough? Is it affecting model outputs? Is it affecting performance?

- No action may be needed.
  - It might be the case that the model has shifted in a way that is still reasonable.
  - Also needs understanding the root cause of drift.

ML Monitoring

ML Monitoring involves computing drift on data or metrics over time

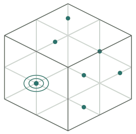
- Track drift over time
  - Basics: Feature Data, Predictions
    - If available: Ground truth, Accuracy
    - Consequences: Influences, MSI, etc
- Set alerts if drift above specific threshold
- Run automated root cause analysis
- Mitigate

Takeaways

- Overview
  - Data drift can happen due to a variety of internal and external causes.
  - Not all drift impacts the model
  - Important to identify consequential drift
- How to identify drift?
  - Different classes of metrics to capture different types of drift: features, ground truth, model output, relationships
  - How to use TruEra to identify root causes of drift
- How to mitigate drift?
  - Not just retrain: Important to understand type and root cause of drift in order to mitigate
  - Retraining, adding features, feature engineering, fixing data quality, and more

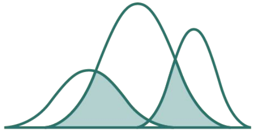
Focus Today: Monitoring Requirements

Fast, precise, and complete.



Model Drift & Performance Metrics

Fast, precise debugging with root cause analysis

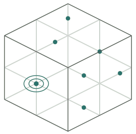


Easy to deploy and scale

AI Monitoring & Explainability:
The Critical Hidden Connection

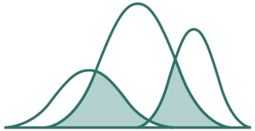
Monitoring Requirements

Fast, precise, and complete.



Broad coverage of model & data quality metrics

Fast, precise debugging



Easy to deploy and scale

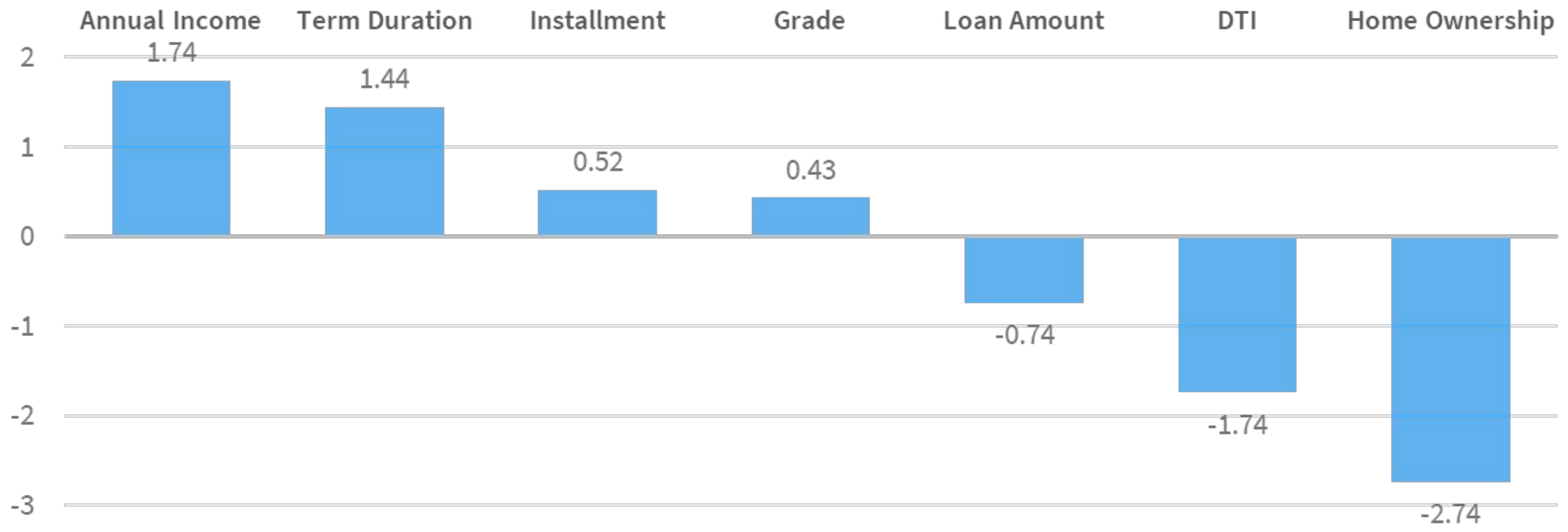
truera

Thank you!

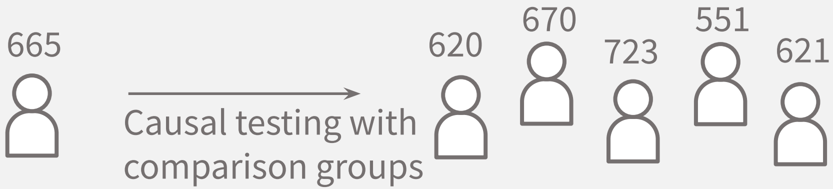
Q&A Time

Appendix: Explainability Methods

Input Feature Importance for a Tree Model



Elements of Explanation Methods

1	QUERY DEFINITION	Why does the model: • have a score of 665 for Jane • have disparate impact • deny Jane
2	OUTPUT COMPARISON	
3	SUMMARIZATION	Of 665, 133 is accounted for by DTI, -45 by income, etc. (Aumann) Shapley accurate estimation

Integrated Gradient

Shapley Value → Aumann Shapley → Integrated Gradient

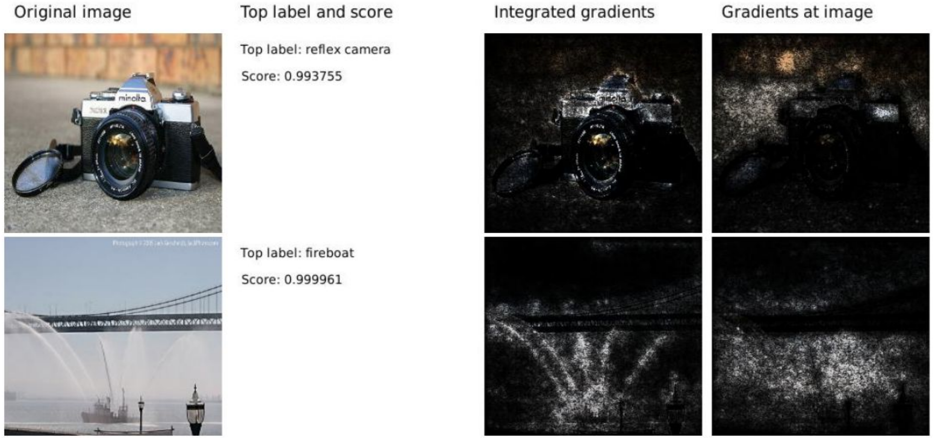
continues features

differentiable output

[Sundararajan et al. ICML 2017]

Integrated Gradient is the **only** path method that satisfies

- Symmetry
- Dummy
- Efficiency(Completeness)
- Additivity

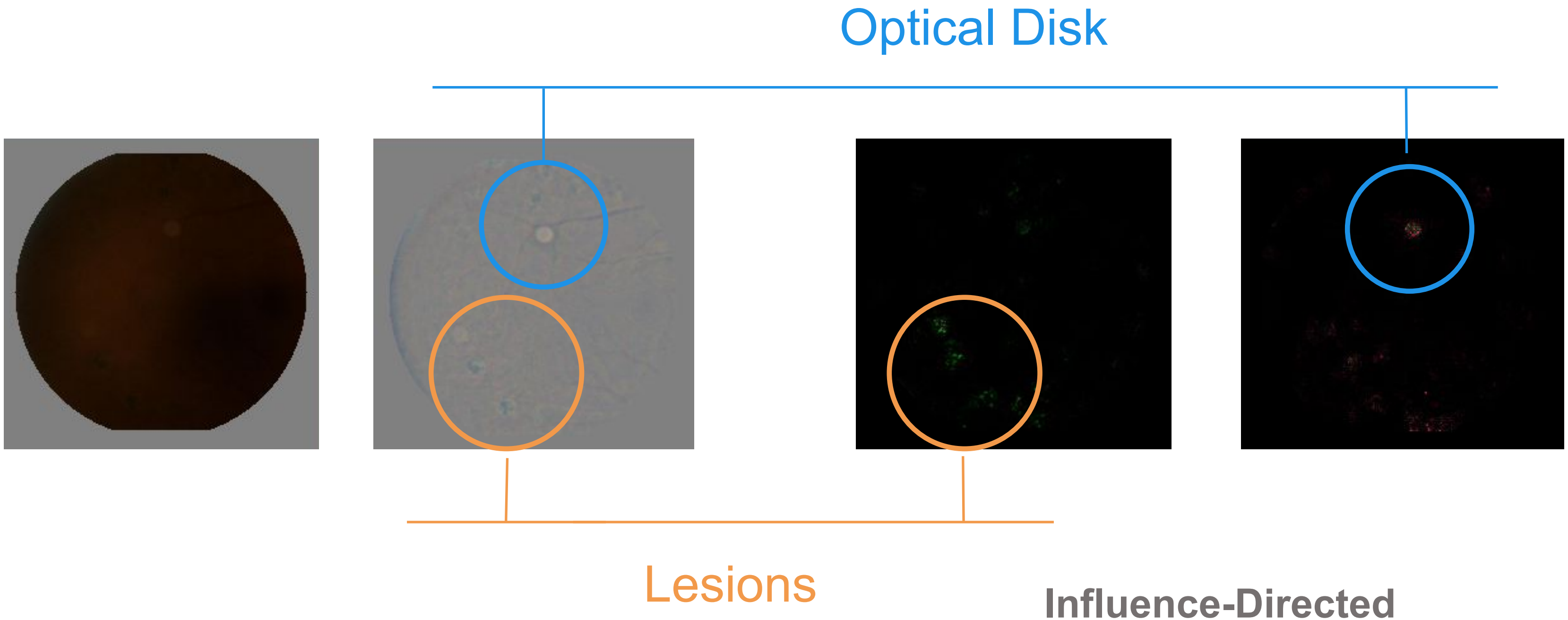


What Makes Orlando Bloom Orlando Bloom?

Internal explanation for a deep network

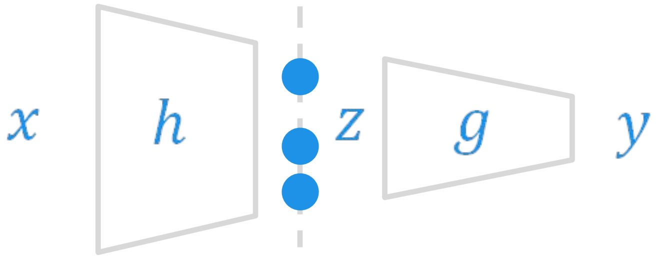
Influence-Directed Explanations
Leino, Sen, Fredrikson, Datta, Li, ITC ‘18

Detecting Diabetic Retinopathy Stage 5



Influence-Directed Explanations
Leino, Sen, Fredrikson, Datta, Li 2018

Requirements for “Good” Explanations



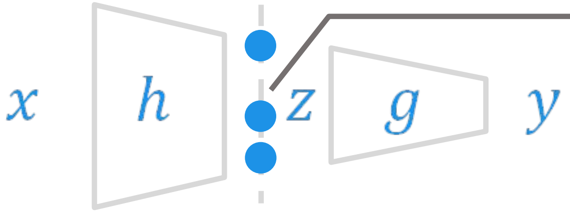
Causal	Succinct	Distributional Faithfulness
Identify features that are causing model predictions	A “few” features explain model predictions	Model is fed “familiar” inputs

Influence-Directed Explanations

Leino, Sen, Fredrikson, Datta, Li, ITC ‘18

Distributional Influence

Influence = average gradient over distribution of interest



$$y = F(x) = g(z), z = h(x)$$

$$\mathrm{I}_j^s(F, P) := \int_{x \in \mathcal{X}} \frac{\partial g(z)}{\partial z_j} P(x) dx$$

z_j

Gradient

For input x
[note z = h(x)]

Weighted by probability
of input x

Influence-Directed Explanations

Leino, Sen, Fredrikson, Datta, Li, ITC ‘18