Machine learning validation

Evaluating ML models and avoiding leakage

Jesper Dramsch

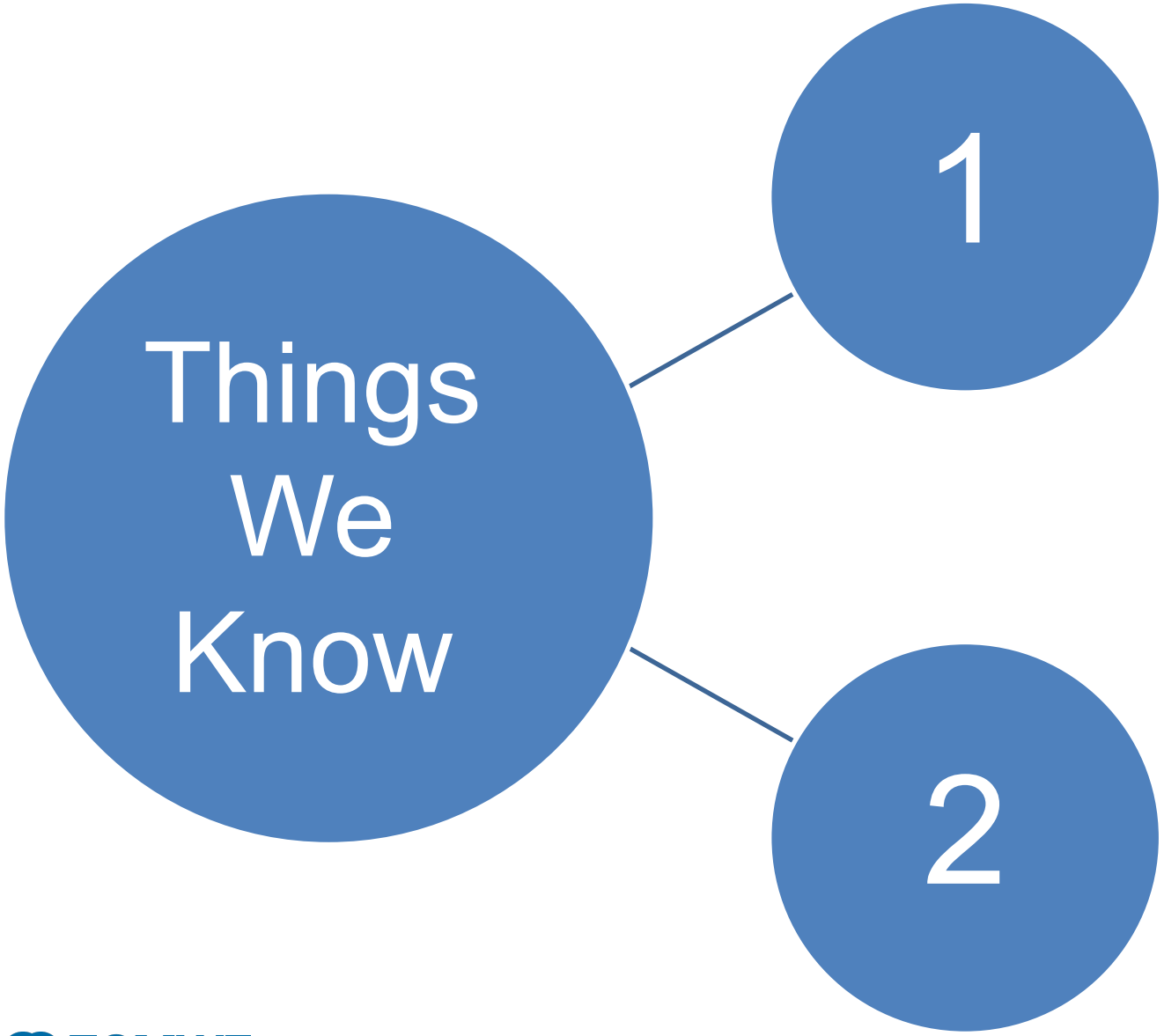
ECWMF

Jesper.Dramsch@ecmwf.int

ECMWF

© ECMWF March 19, 2024

Motivation



- Memorise Data
- Retrieve

- Derive Mechanisms
- Generalize

Motivation


●●○○○ AT&T LTE
3:23 PM
80%
Back
Robbie
Contact
Tue, May 3, 1:21 PM
What's 6+7?
Donut
6 + 7 = 13
Got it 6+7=13
What's 6+7?
13
What's 13+9?
13
iMessage
Send

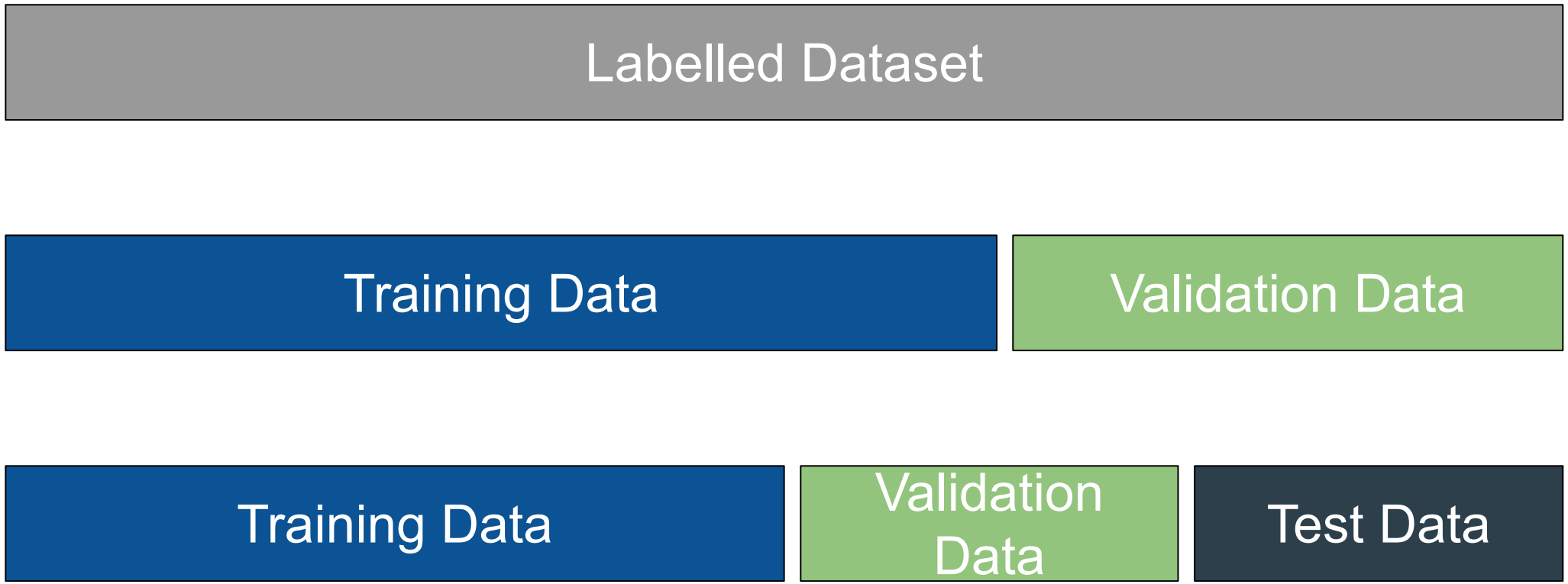
How do we ensure our **models work** on **unseen data** in the future?

Outline

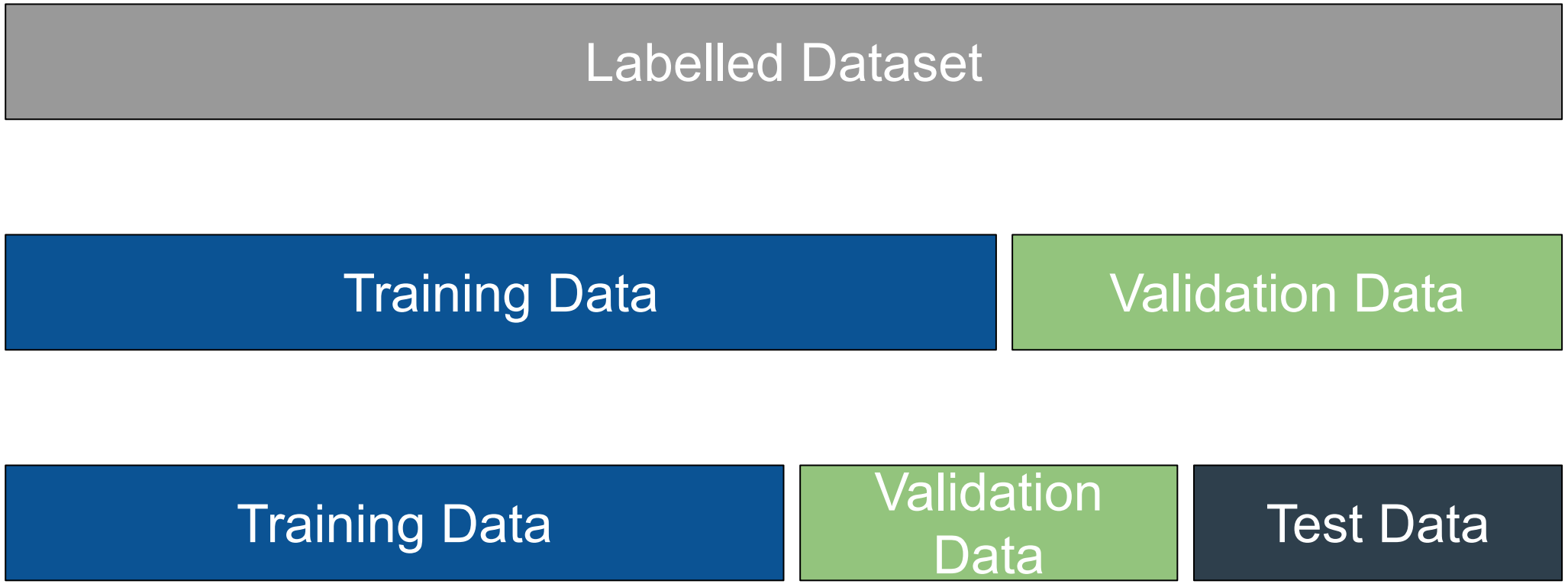
- Basic Validation Strategies
- Imbalanced and Heterogeneous Data
- Correlated and Connected Data
- Data, Target, and Concept Drift
- Practical considerations in Snooping and Data Leakage
- Baseline Methods and Model Interpretation

Basic Validation Strategies

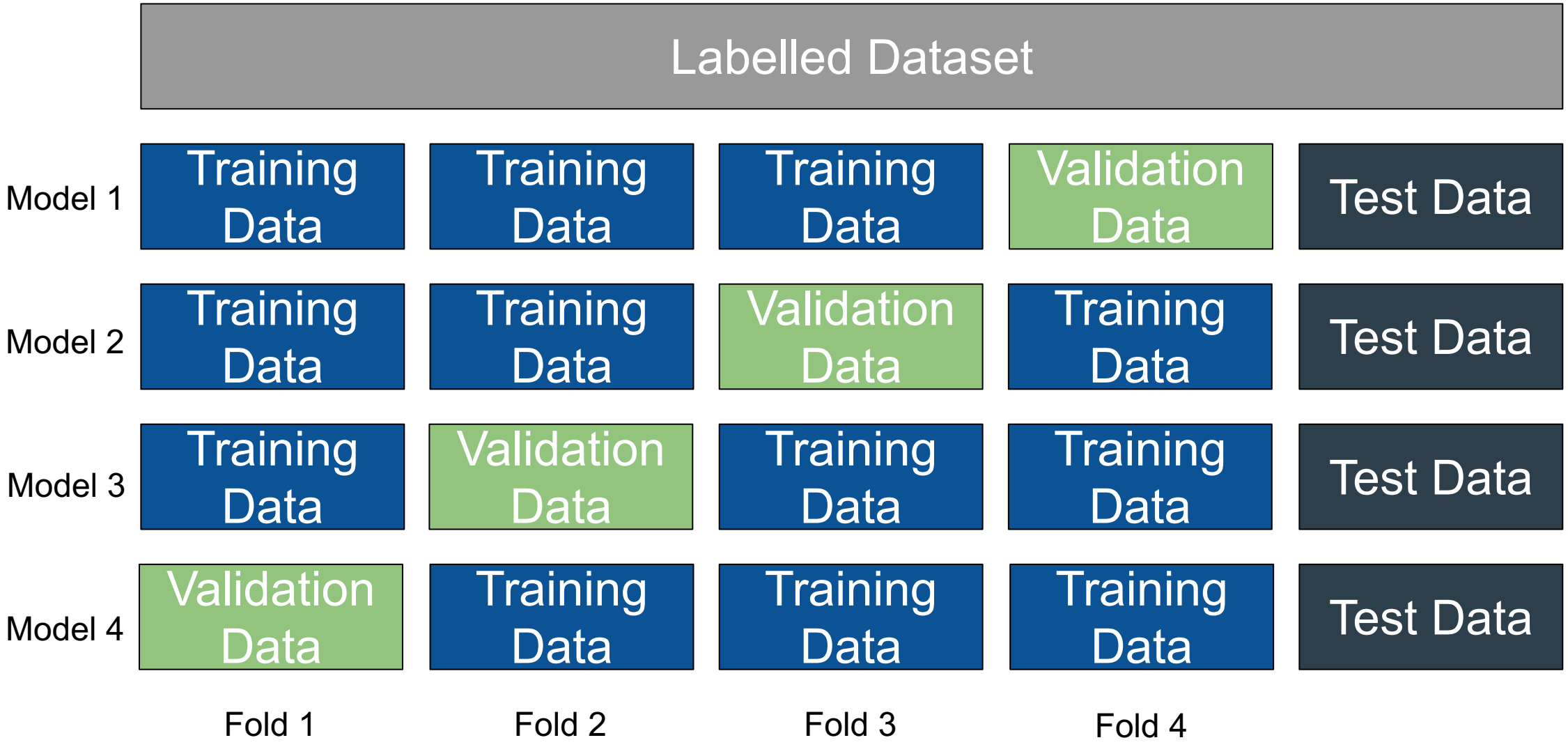
Obtaining Data to Test On



Validation on Small Dataset

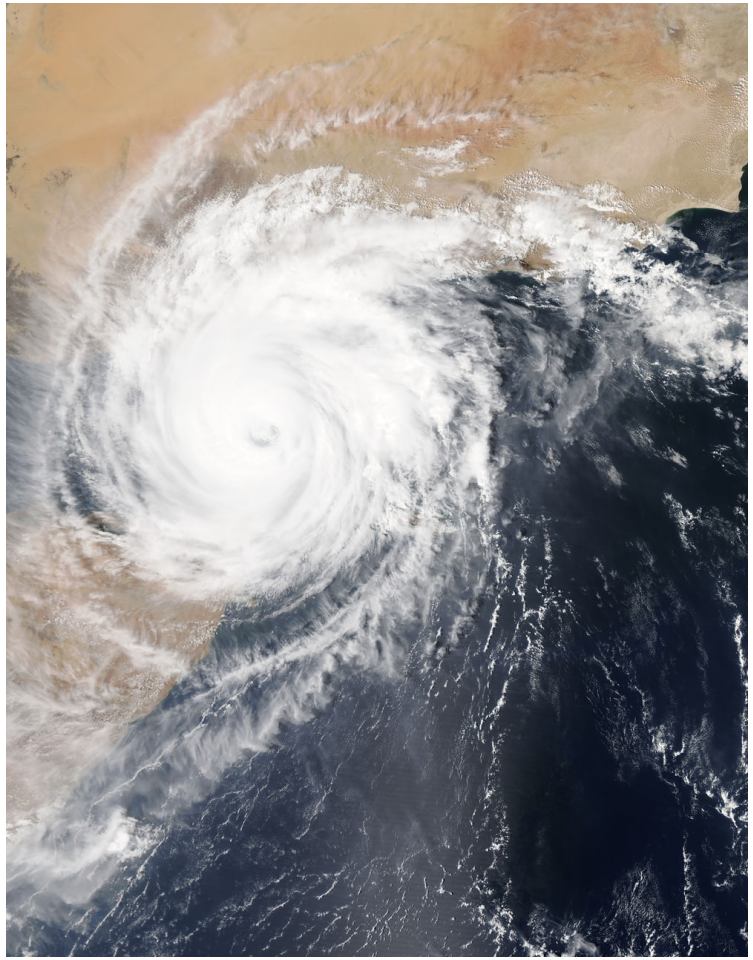


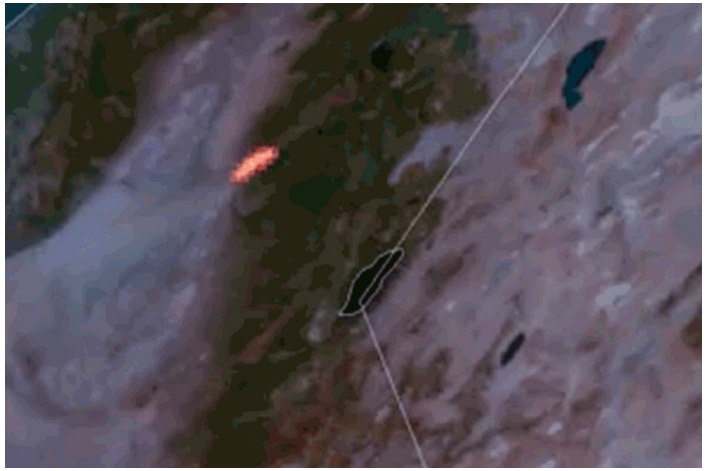
Cross-Validation



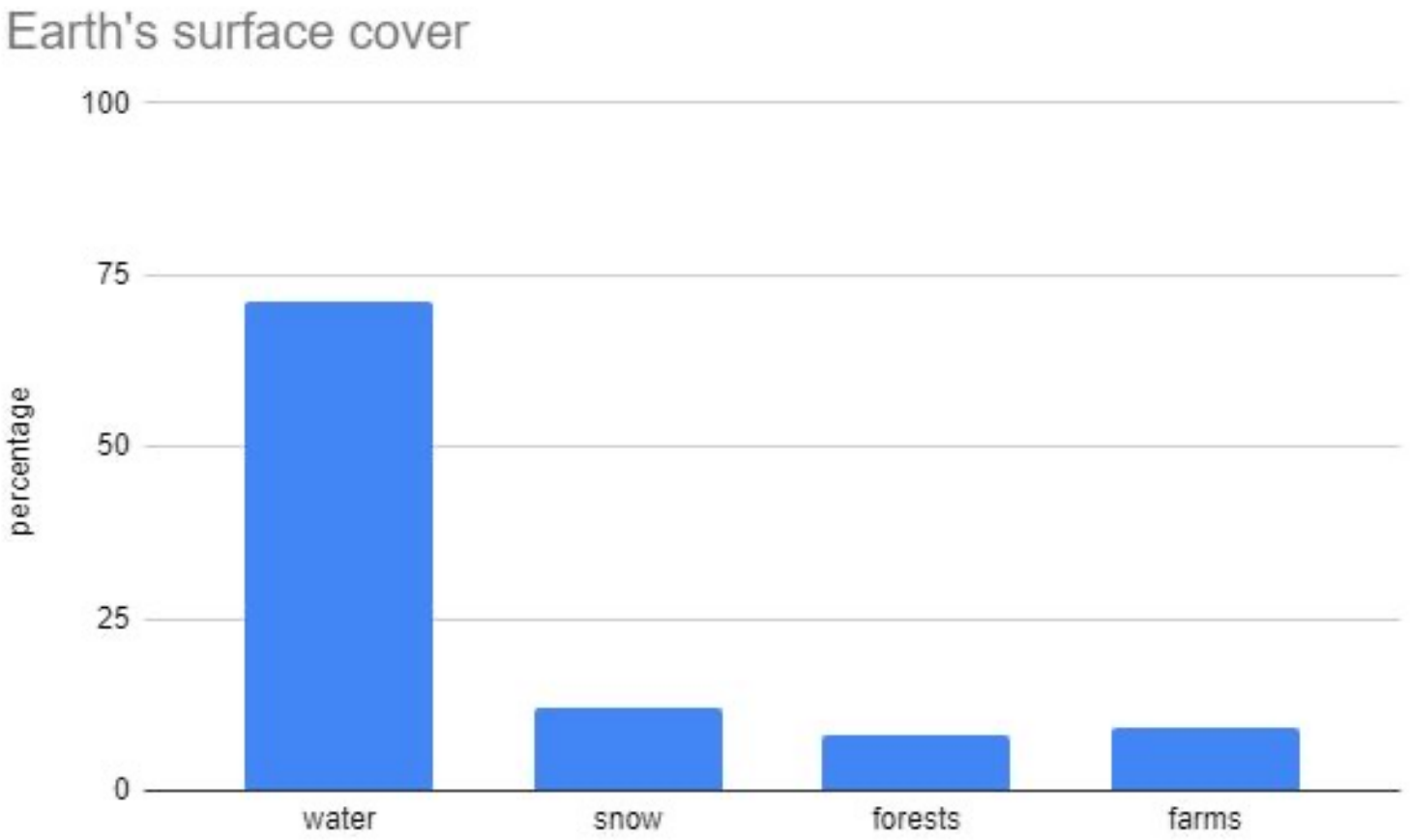
Imbalanced and Heterogeneous Data

Examples for Imbalanced Data

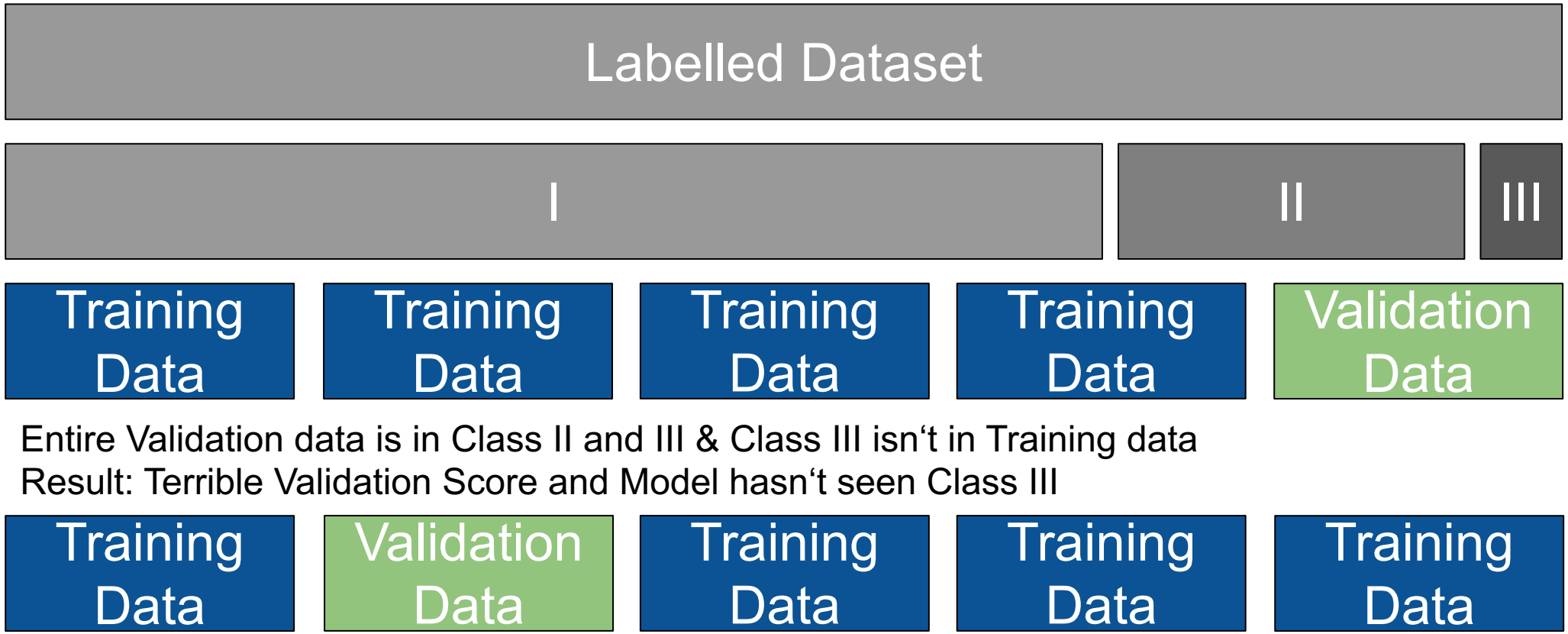




Class Imbalance



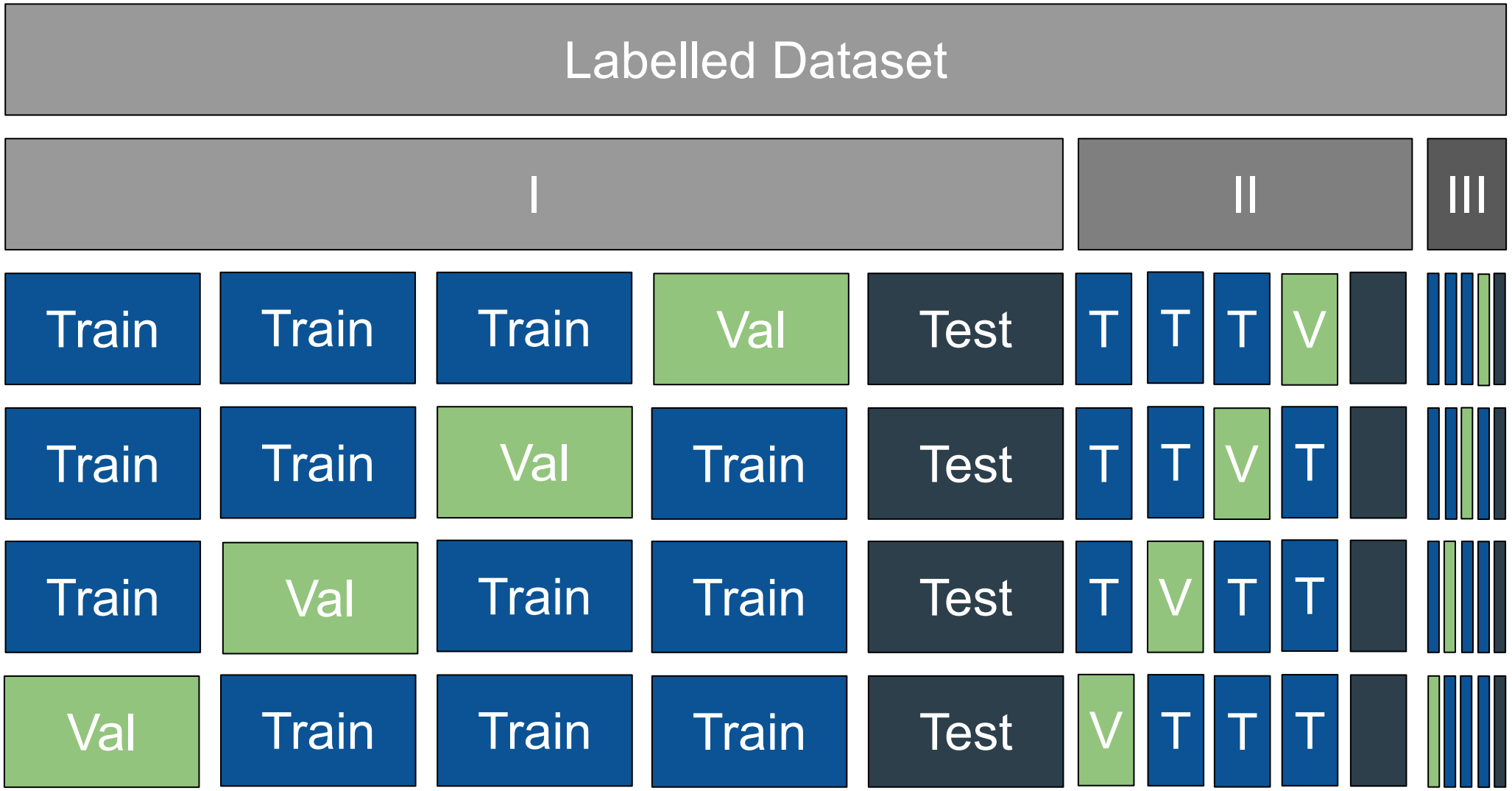
Why not use Random Sampling like before?



Entire Validation data is in Class II and III & Class III isn‘t in Training data
Result: Terrible Validation Score and Model hasn‘t seen Class III

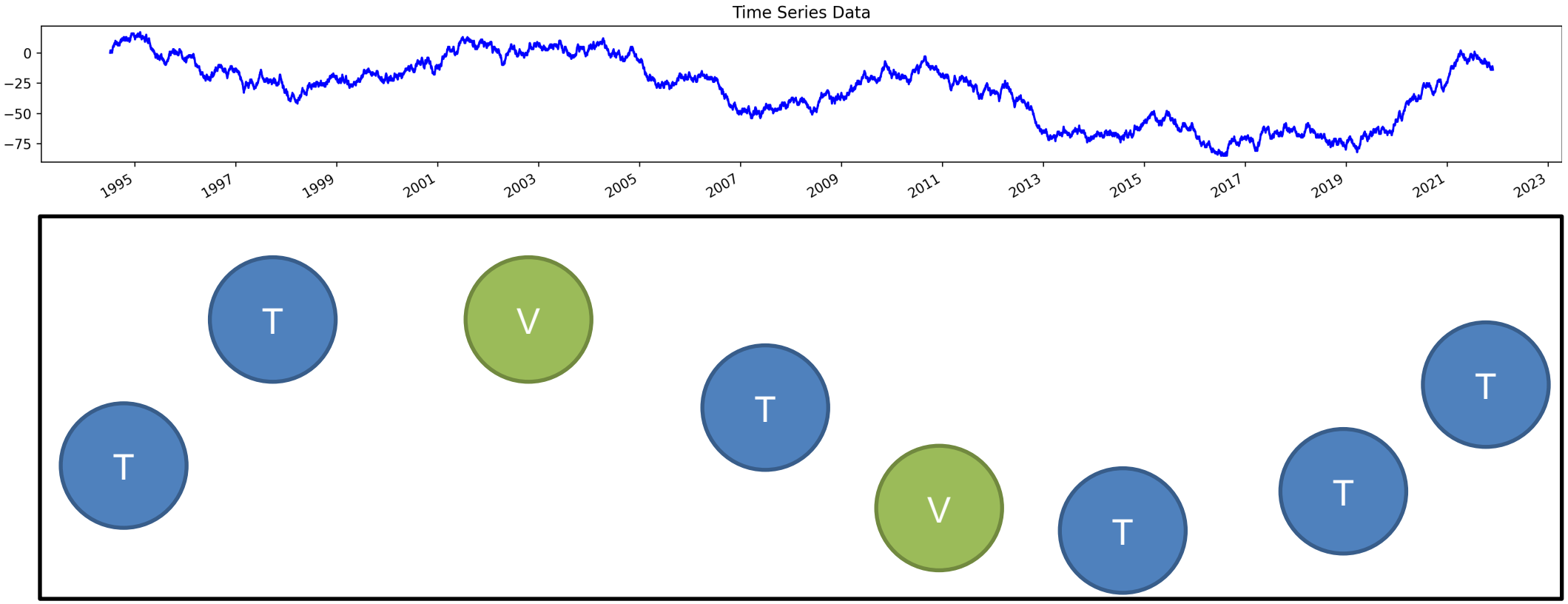
Entire Validation data is in Class I
Result: Great Validation Score but no validation of Class II & III at all

Stratification for Imbalanced Data



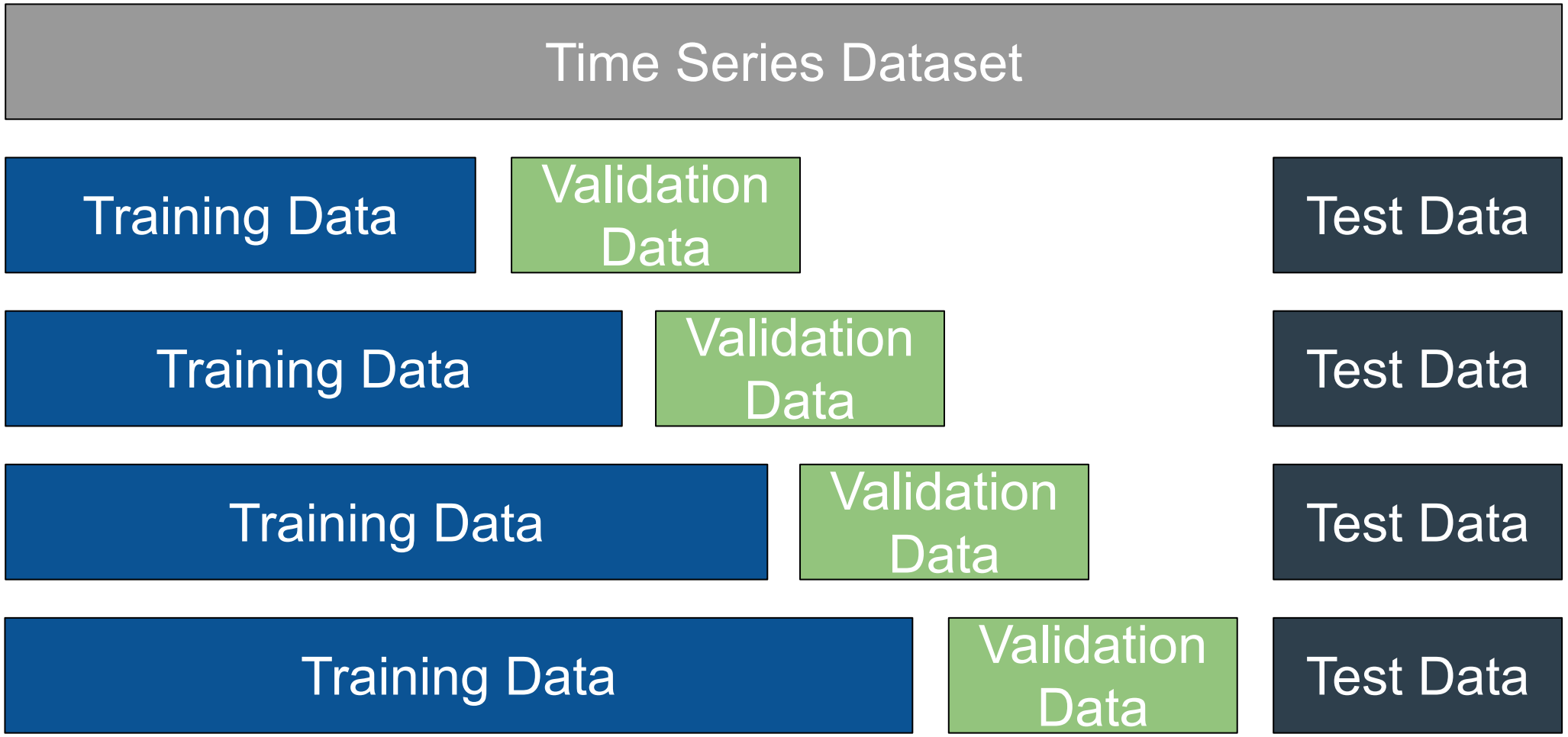
Correlated and Connected Data

Time Series Data



- Random Splits on Time Series Data equates to Interpolation
- Bad on standard time series problems
- Devastating on forecasting problems

Validation on Time Series Data



Validation of Geospatial Data

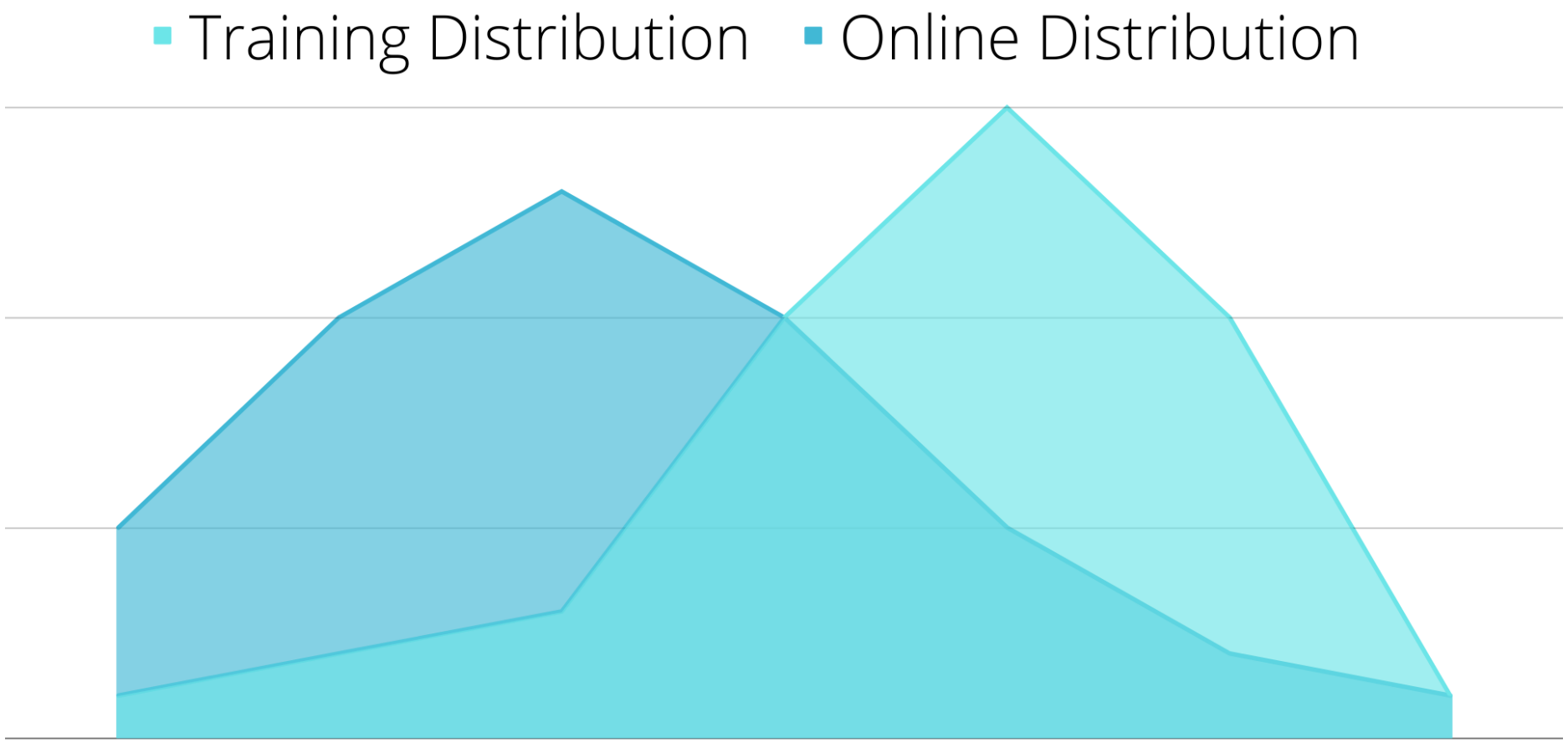
- Geospatial Data Examples
  - Stations
  - Satellite Data
  - Weather Radar
- Geospatial Data is spatially correlated
- Problems with random split of data:
  - Clustering of Validation Locations
  - Overlap of Validation and Training Locations

Validation of Geospatial Data

- Geospatial Data Examples
  - Stations
  - Satellite Data
  - Weather Radar
- Geospatial Data is spatially correlated
- Problems with random split of data:
  - Clustering of Validation Locations
  - Overlap of Validation and Training Locations

Data, Target, and Concept Drift

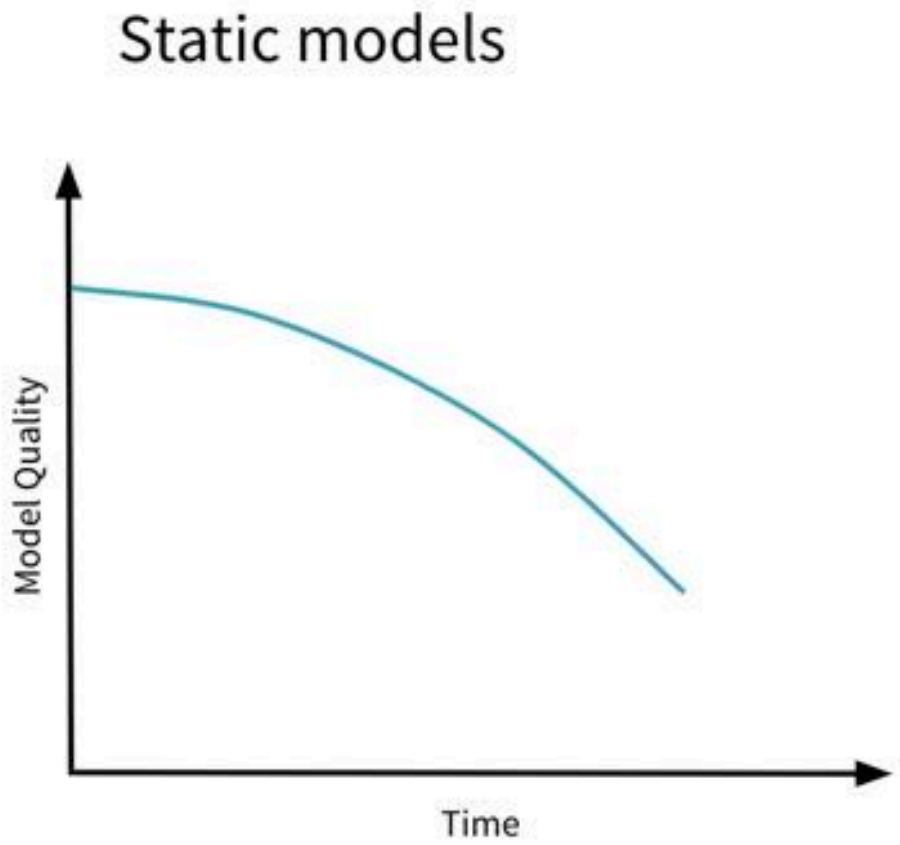
Data Drift

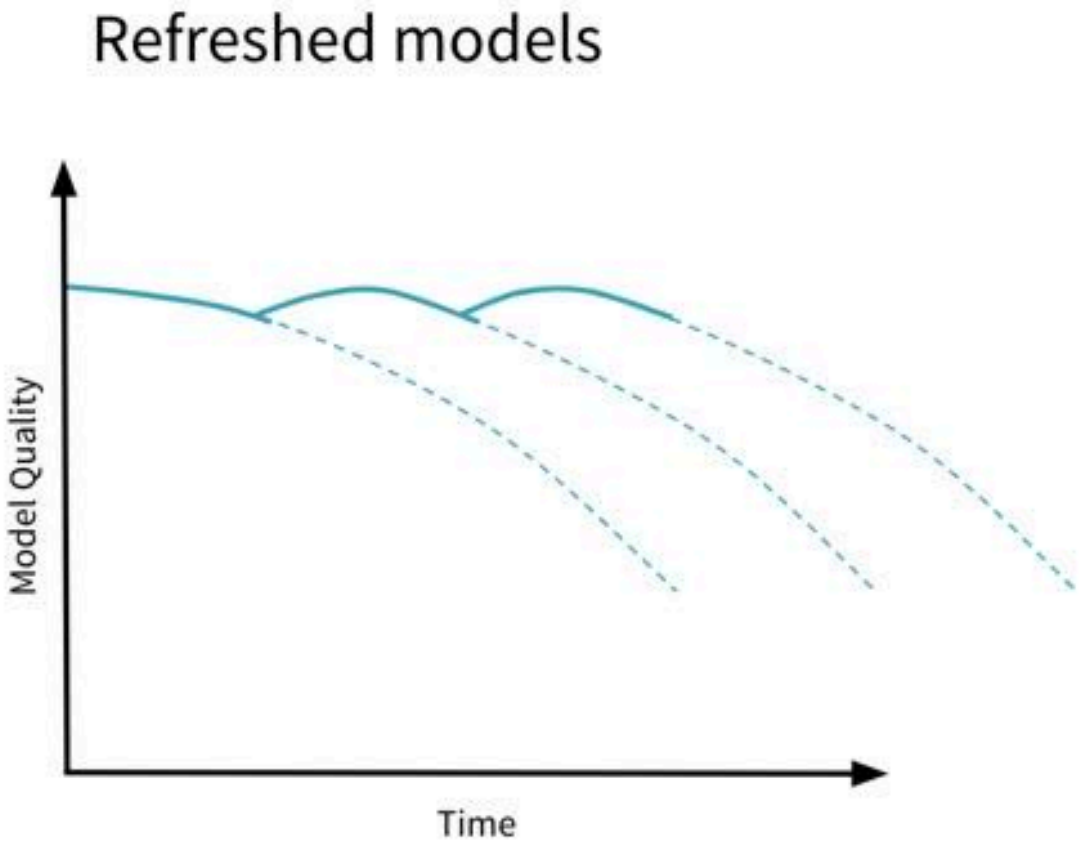


Data Drift

- Changes of Input Data
  - Also called covariate shift
- Examples:
  - Change in global temperature
  - Users of social media platform growing older
- Mitigation Strategies:
  - Measure distribution of input data
    - Continuous: Kolmogorov-Smirnov test
    - Categorical: Chi-squared test
  - Periodic retraining of models in production

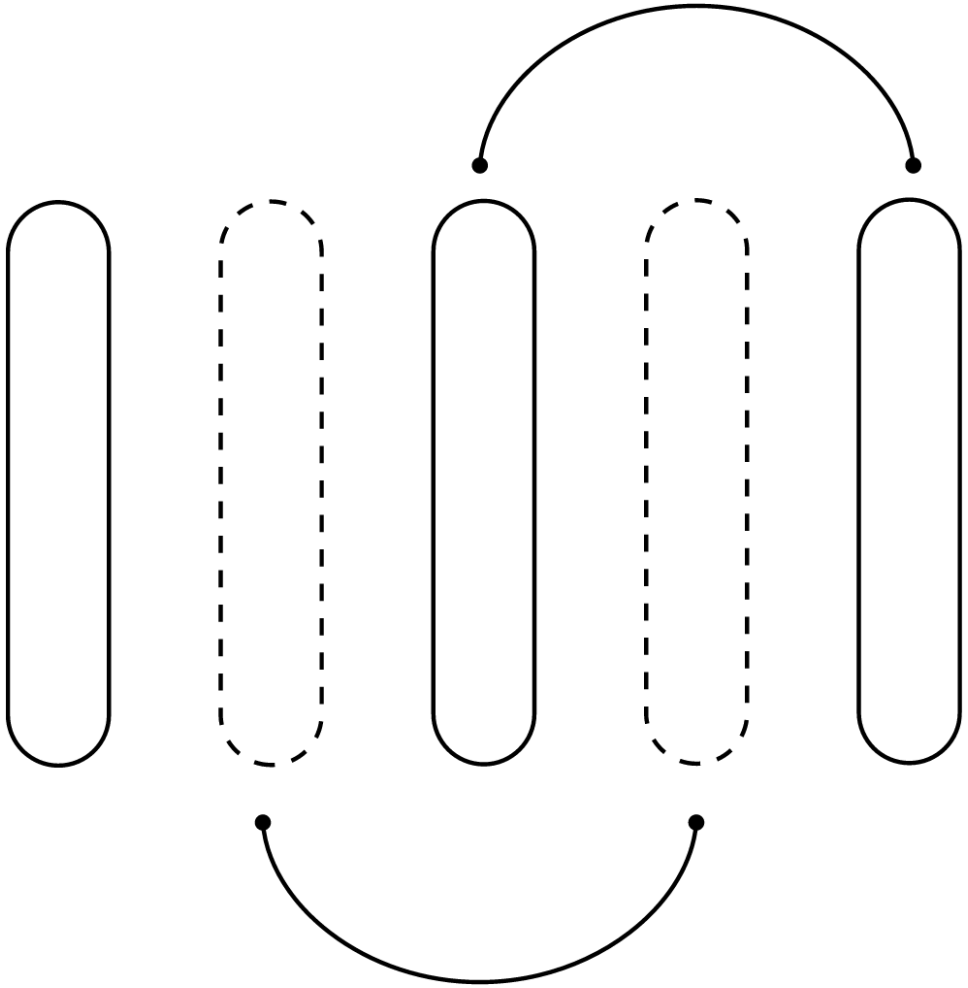
Why periodic retraining?





Source: Databricks

Target Drift



Target Drift

- Changes in Target Data
- Examples:
  - Natural changes, e.g. change of coastlines through erosion
  - Change of Categories, e.g. Regulatory Changes
- Mitigation Strategies:
  - Risk Assessment for Category Changes
  - Flexible Production Pipelines
  - Monitoring of Target Data Distribution
  - Retraining New Model on New Target Definition

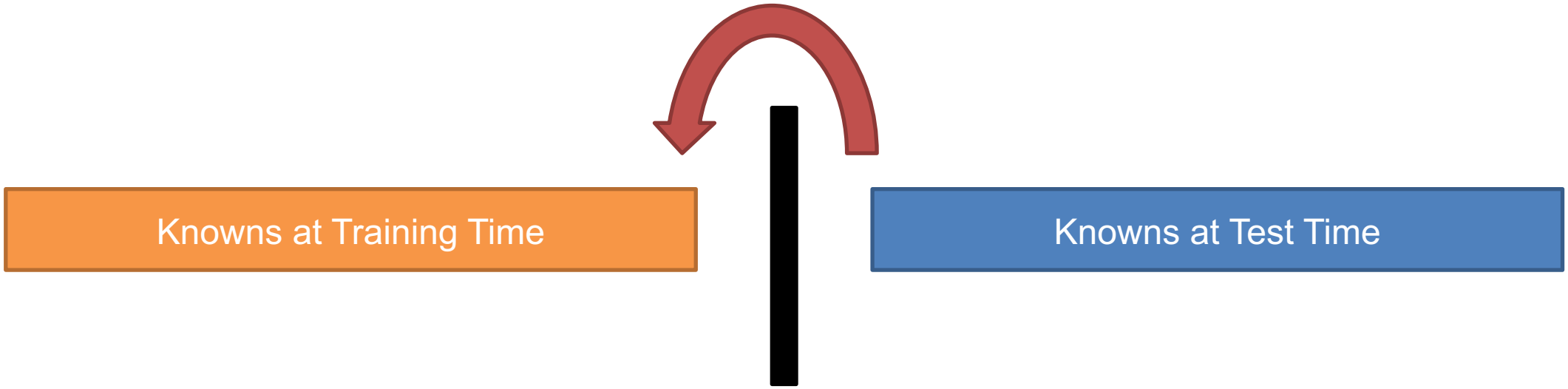
Concept Drift

Concept Drift

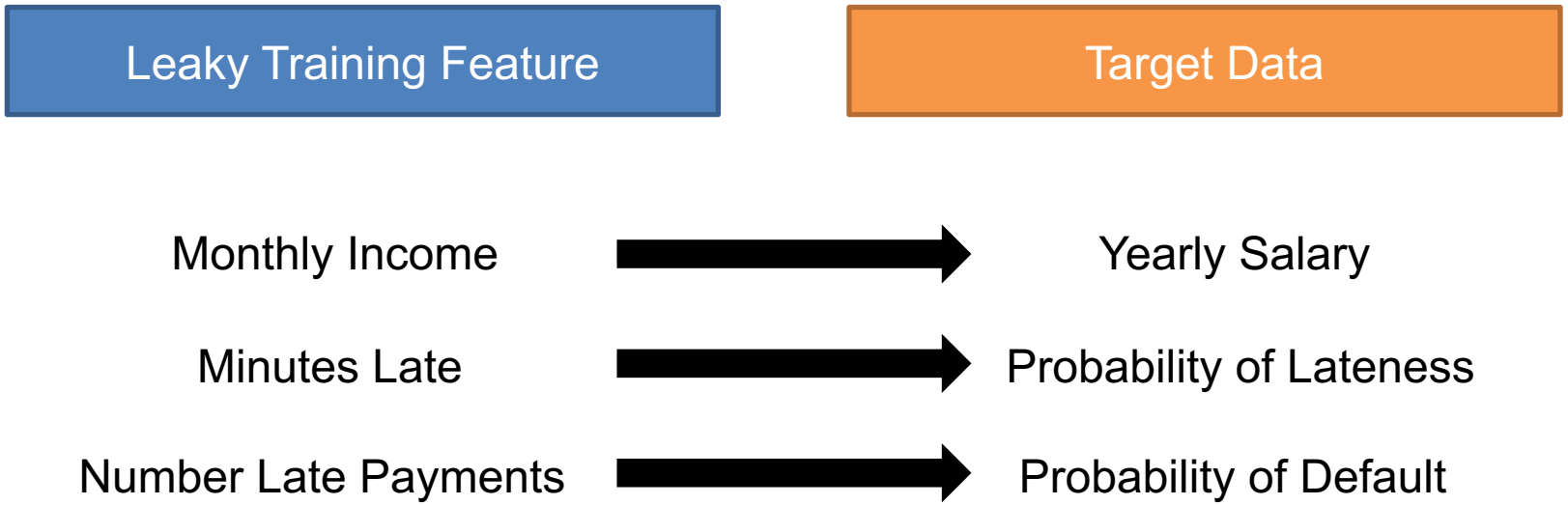
- Change of Relationship between Input and Output data
- Example:
  - Rayleigh Scattering to Mie-Scattering
  - Shopping Behaviour in April 2020
- Mitigation Strategy:
  - Automatic Monitoring of Model Performance
  - Set Up Alerts
  - Be Prepared to Take Model Offline

Practical considerations in Snooping and Data Leakage

Data Leakage “Anachronisms”

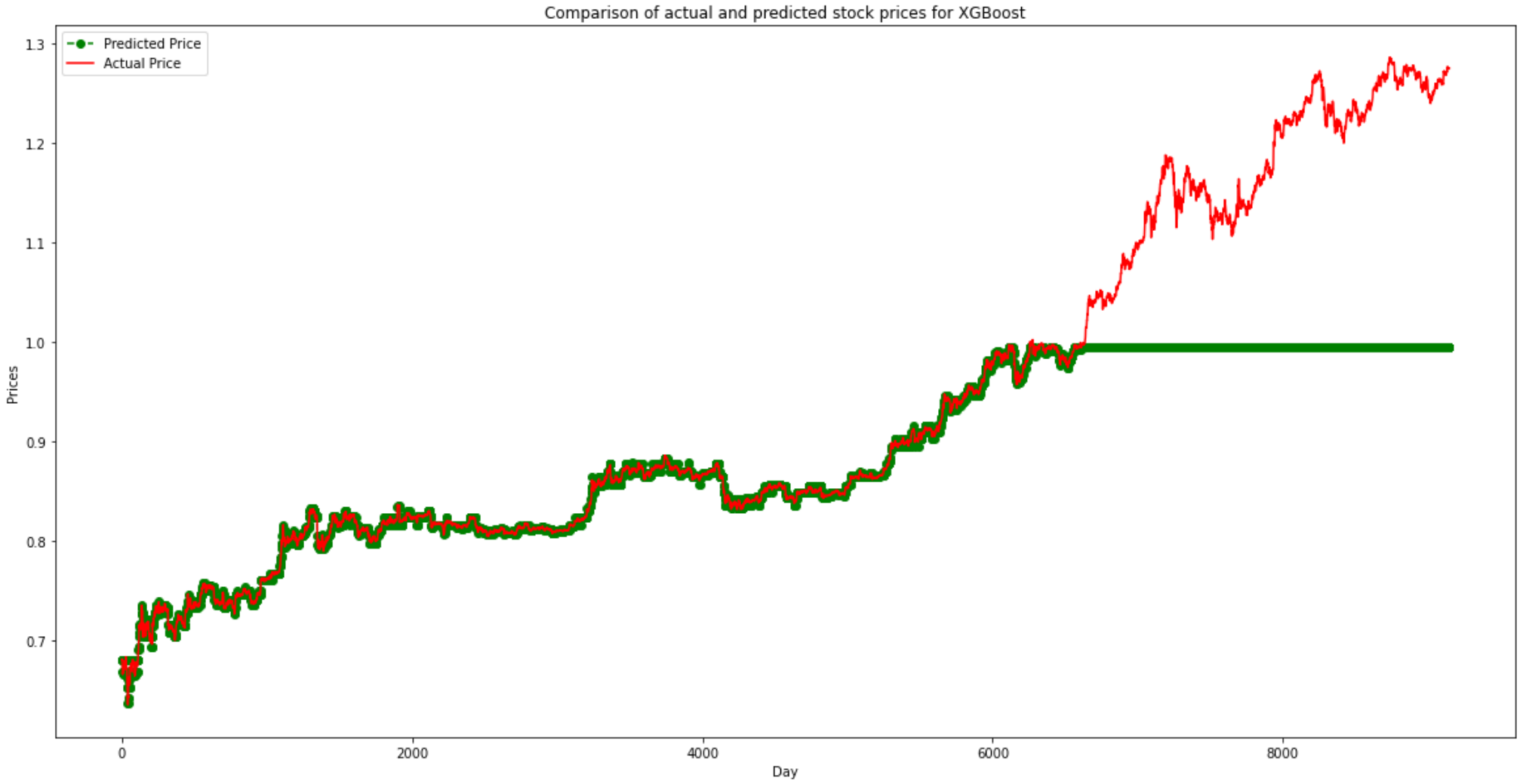


Examples of Leaking Features in Tabular Data



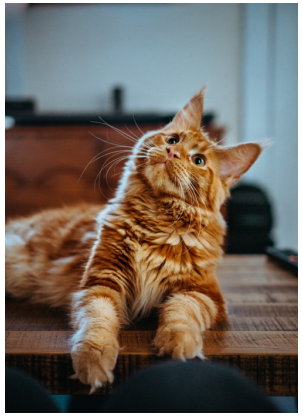
Possible Error: Image Artifacts

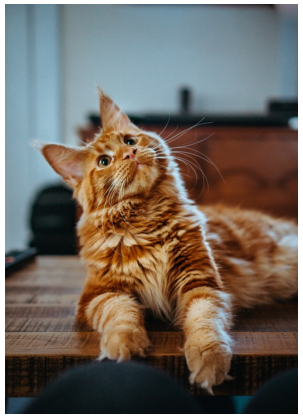
Possible Error: Normalizing on Test Data



Possible Error: Data Augmentation & Duplicates in Training & Test Set

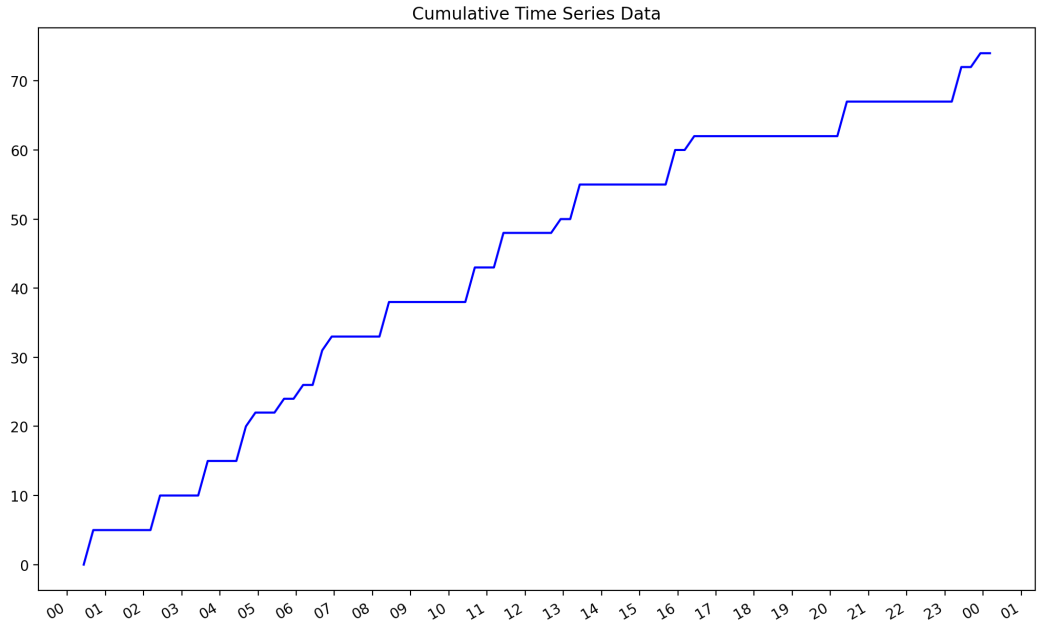
Training Data





Test Data

Possible Error: Cumulative Data with Temporal Leakage

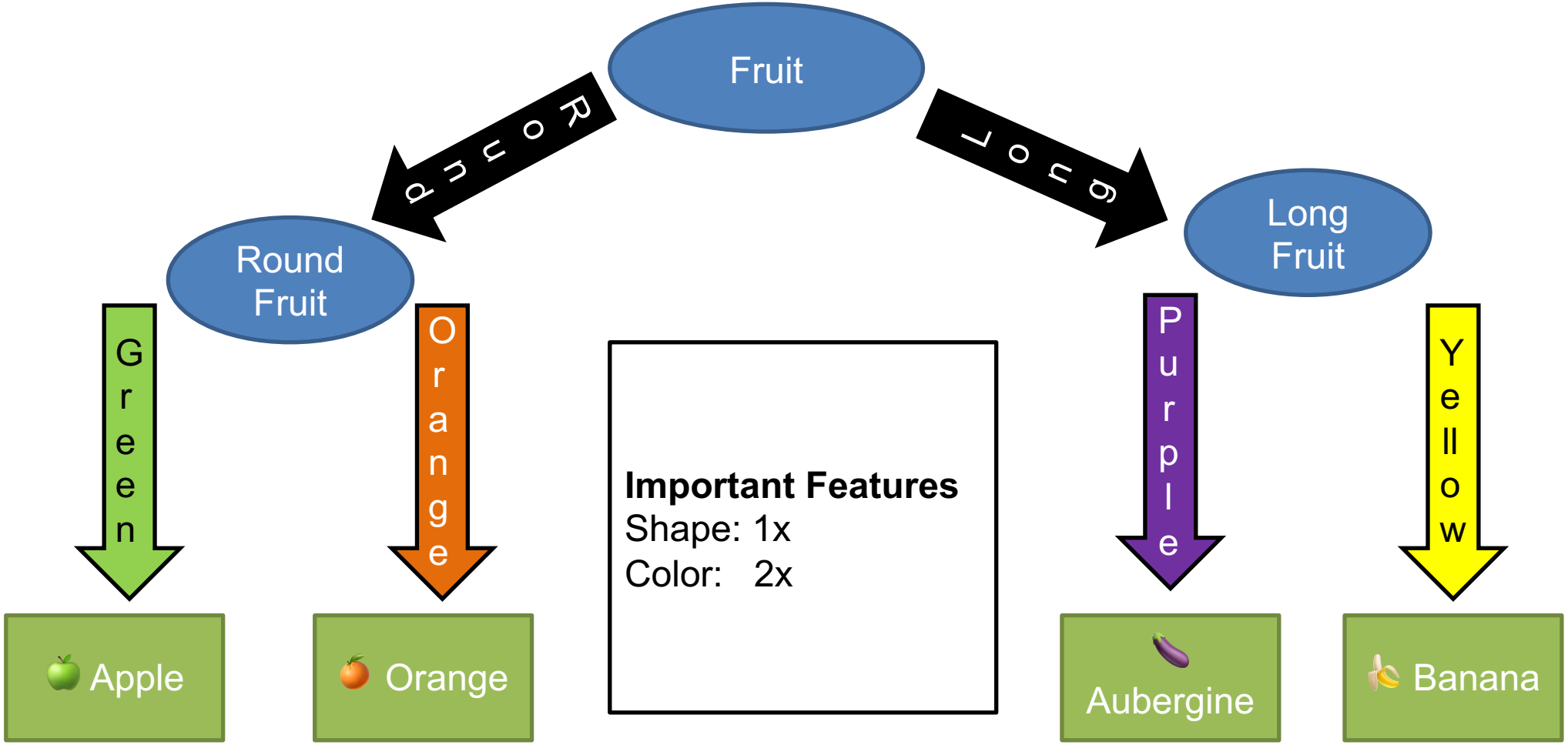


Baseline Methods and Model Interpretation

Baseline Models

- Build Machine Learning Baselines
  - Linear Regression
  - Random Forests
  - SVM
- Consult with Domain Experts
  - Compare with existing Model Performance
  - Use Explainability to Discuss Model Interpretation
- Verify against Benchmarks and Established Models

Feature Importance for Model Interpretation

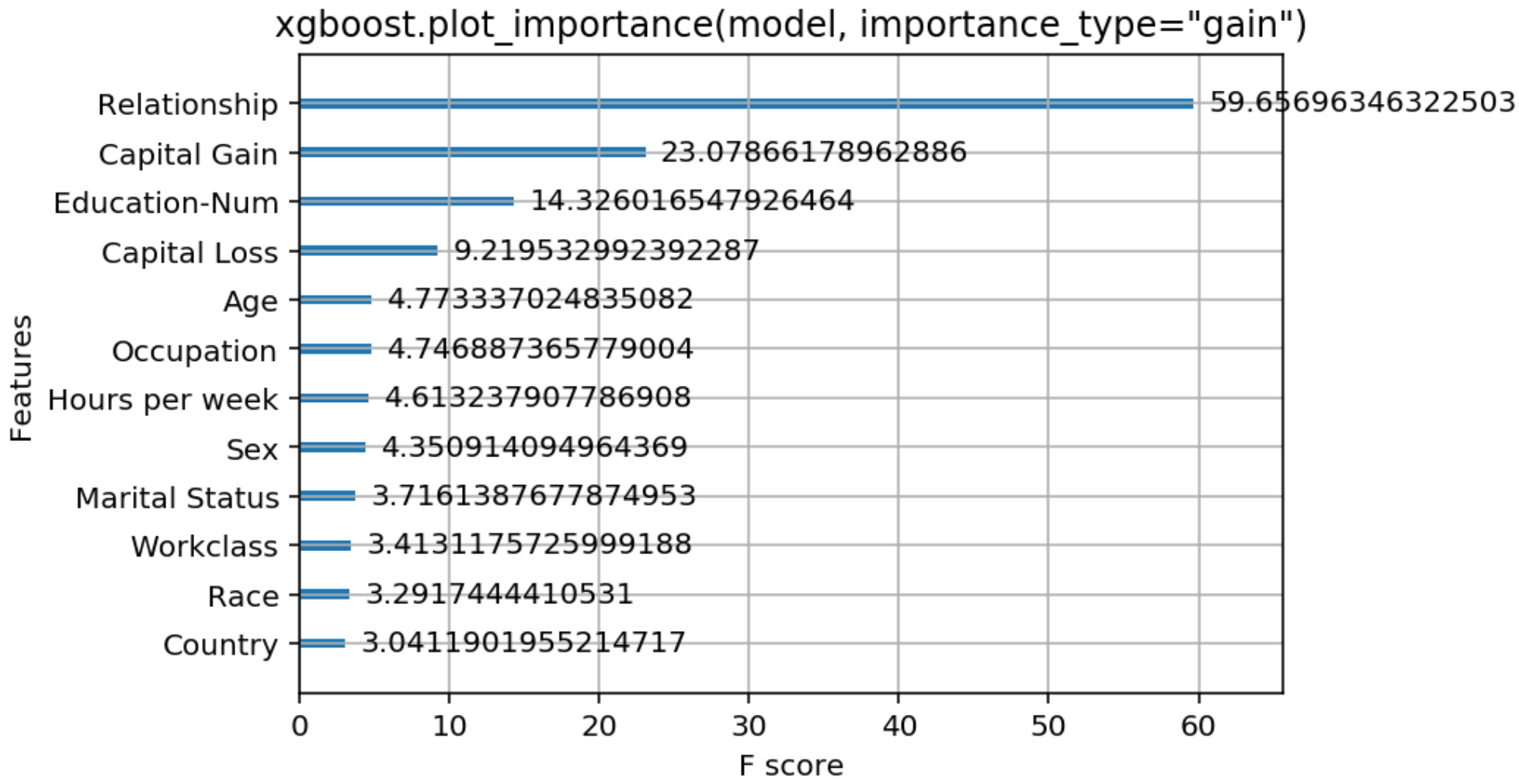


Feature Importance for Model Interpretation

Iteratively Scramble Feature

Evaluate Importance

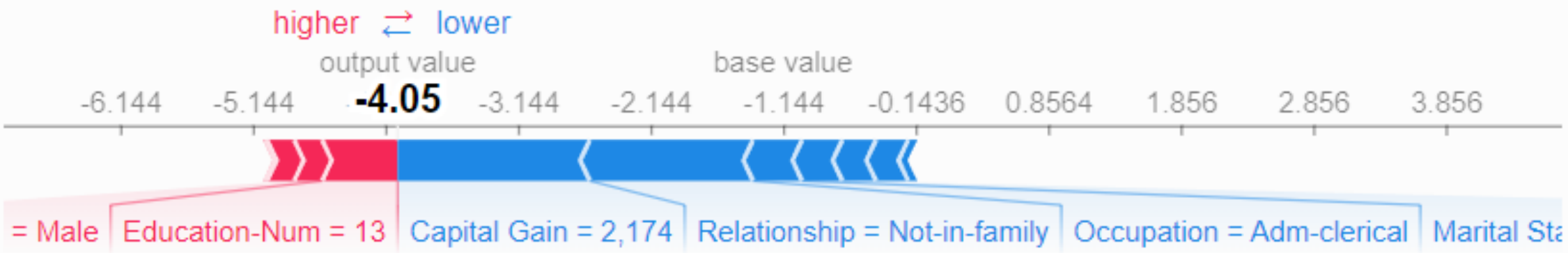
XGBoost Feature Importance



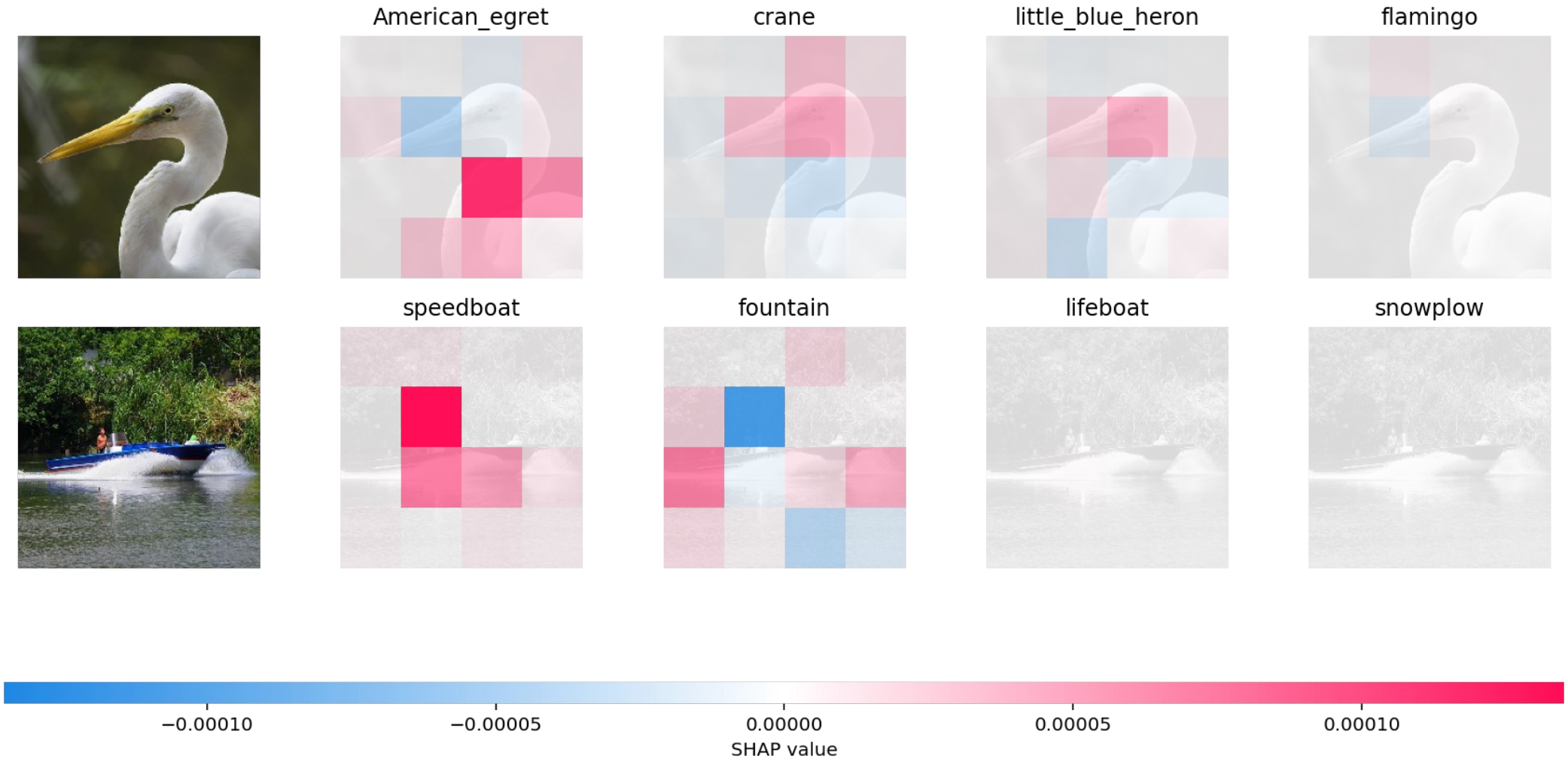
Visualizing Single Predictions

```
[8]: shap.force_plot(explainer.expected_value, shap_values[0,:], X_display.iloc[0,:])
```

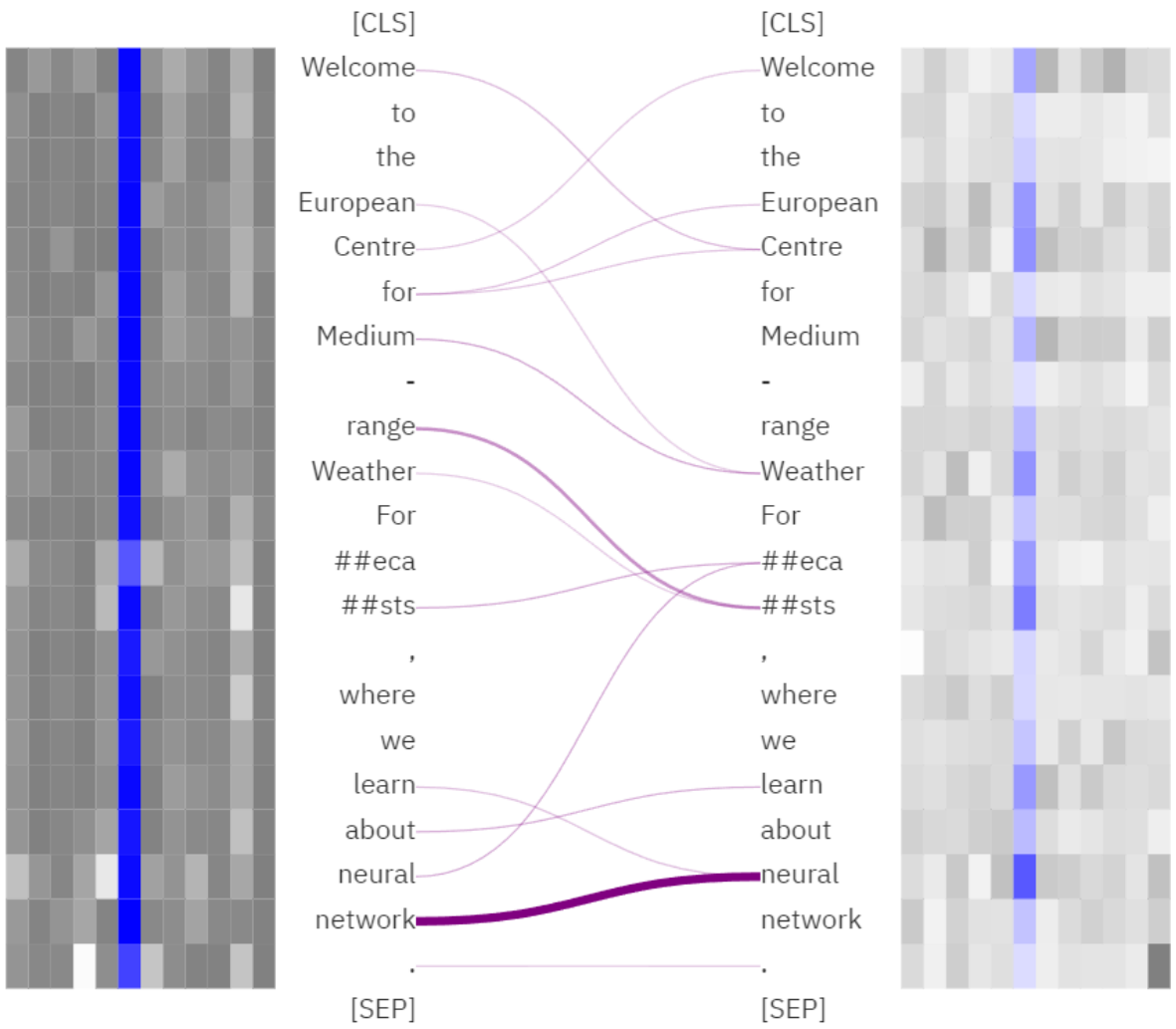
[8]:



Saliency Maps



Explore Attention Maps



What We Learned

- Generalization & Overfitting
- Random Splits into Training, Validation & Test Set
- Grouped Splitting of Spatially Correlated Data
- Time Series Splits of Temporally Correlated Data
- Various Modes of Drift
- Possible Pitfalls During Pre-Processing and Data Preparation
- Baseline Models
- Model Interpretability