Machine learning validation

Evaluating ML models and avoiding leakage

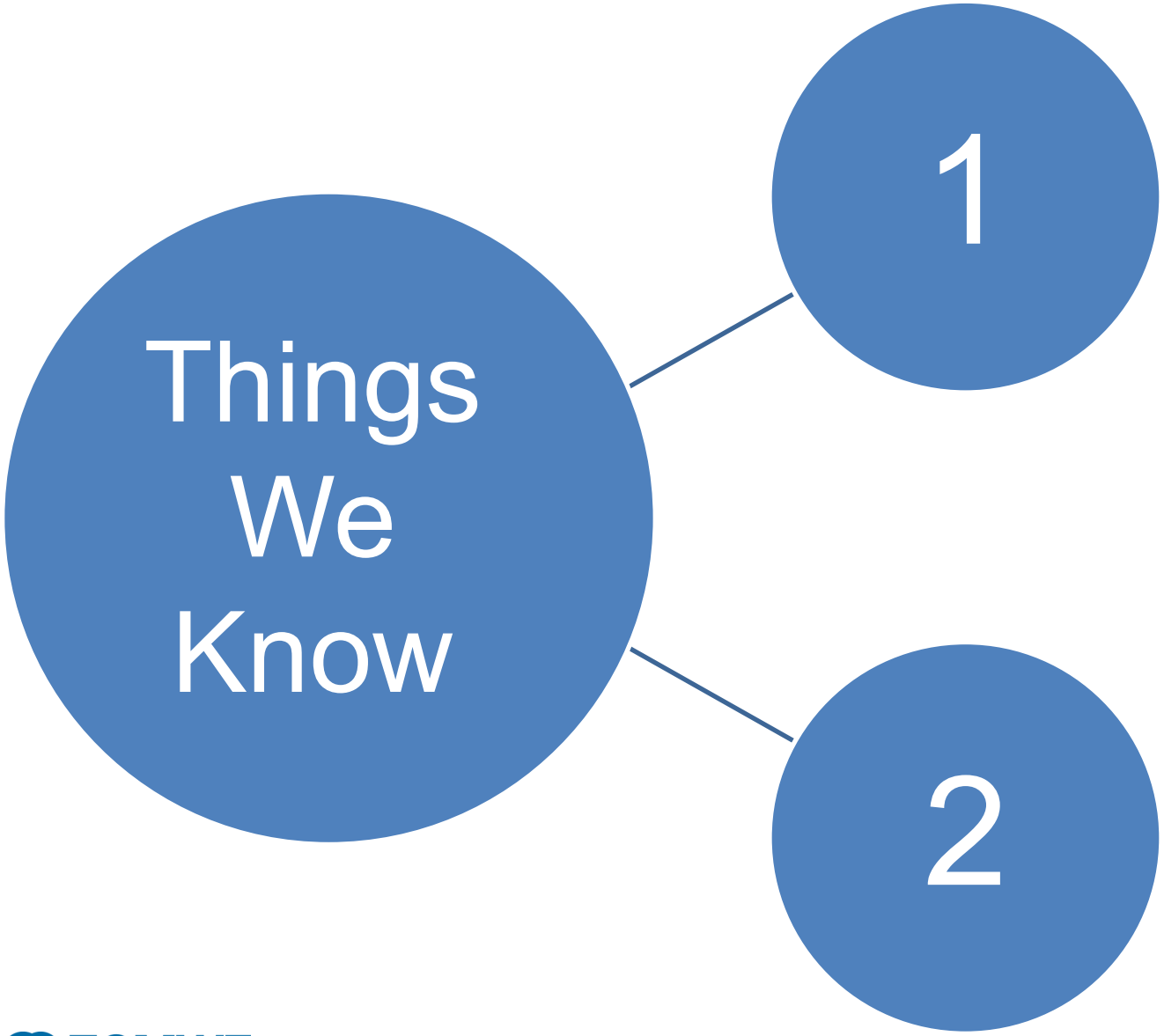
Julian Kuehnert, Jesper Dramsch

ECWMF Bonn

julian.kuehnert@ecmwf.int

ECMWF

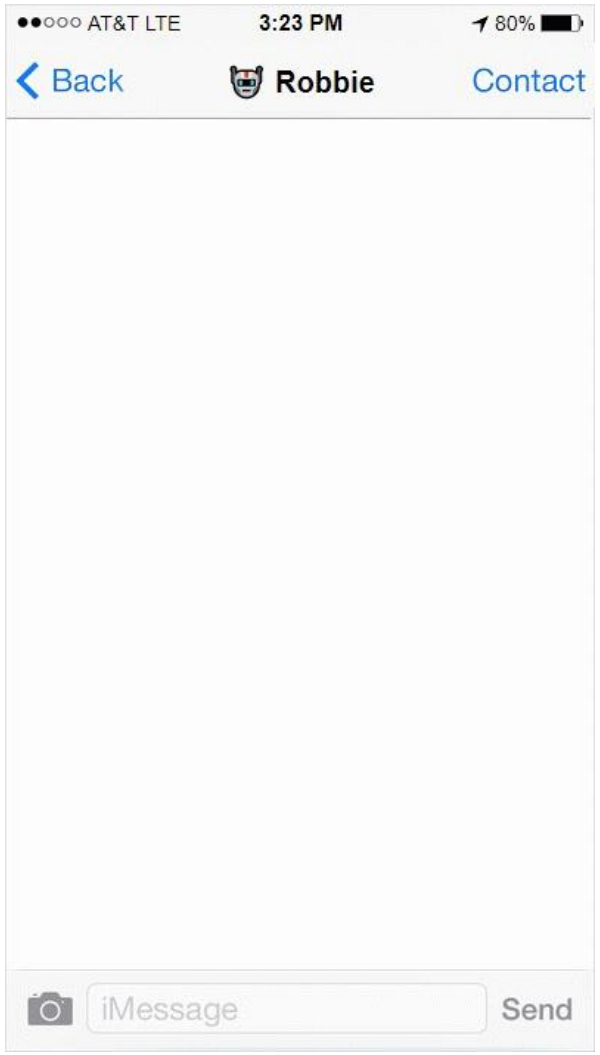
Motivation



- Memorise Data
- Retrieve

- Derive Mechanisms
- Generalize

Motivation



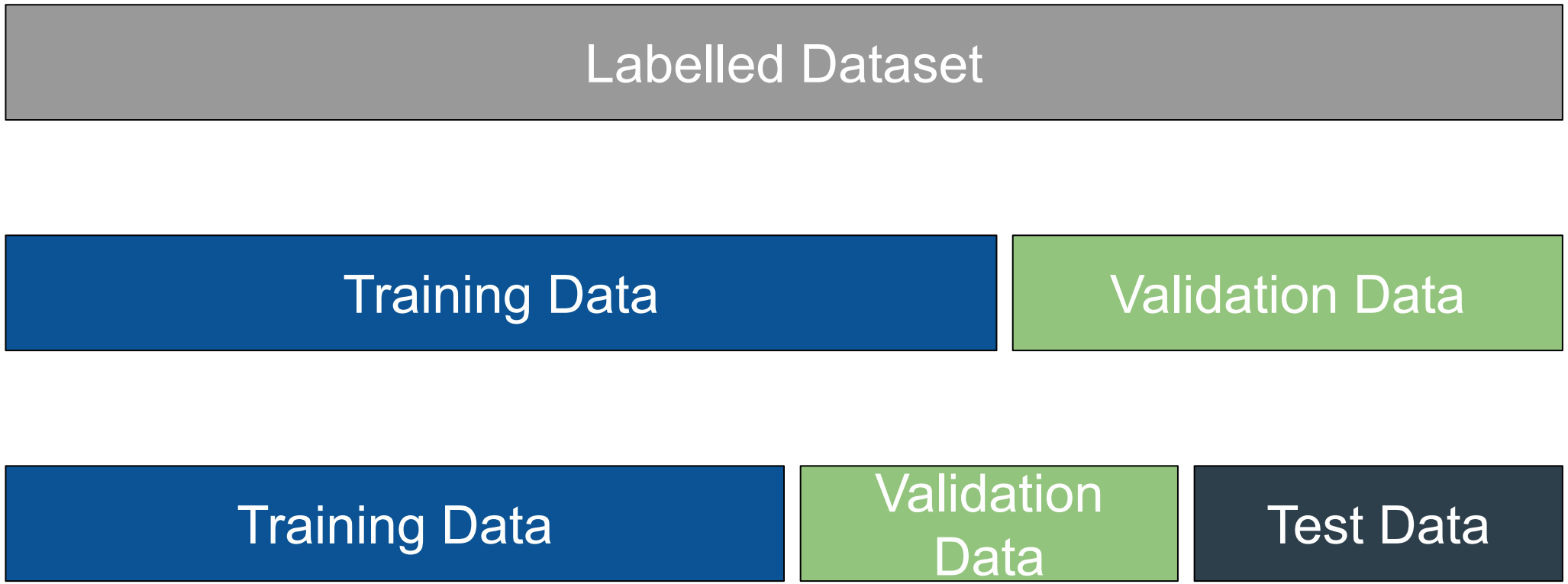
How do we ensure our **models work** on **unseen data** in the future?

Outline

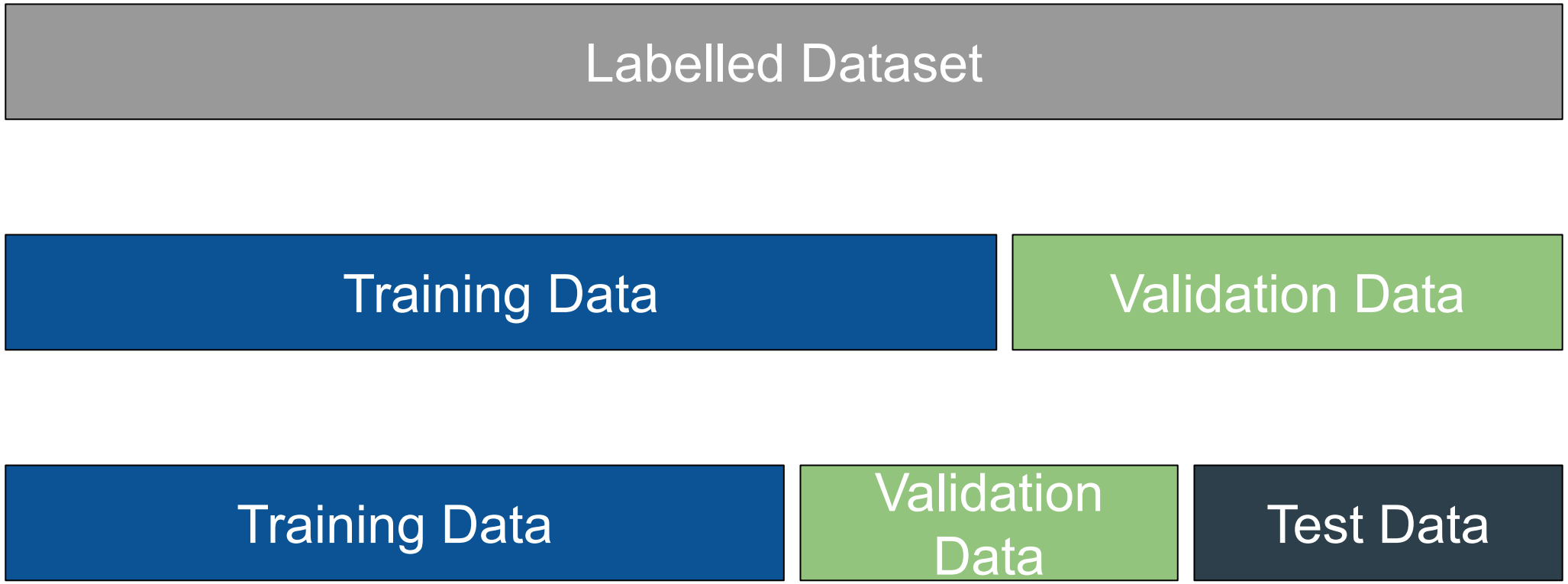
- Basic Validation Strategies
- Imbalanced and Heterogeneous Data
- Correlated and Connected Data
- Data, Target, and Concept Drift
- Baseline Methods and Model Verification
- Practical considerations in Snooping and Data Leakage

Basic Validation Strategies

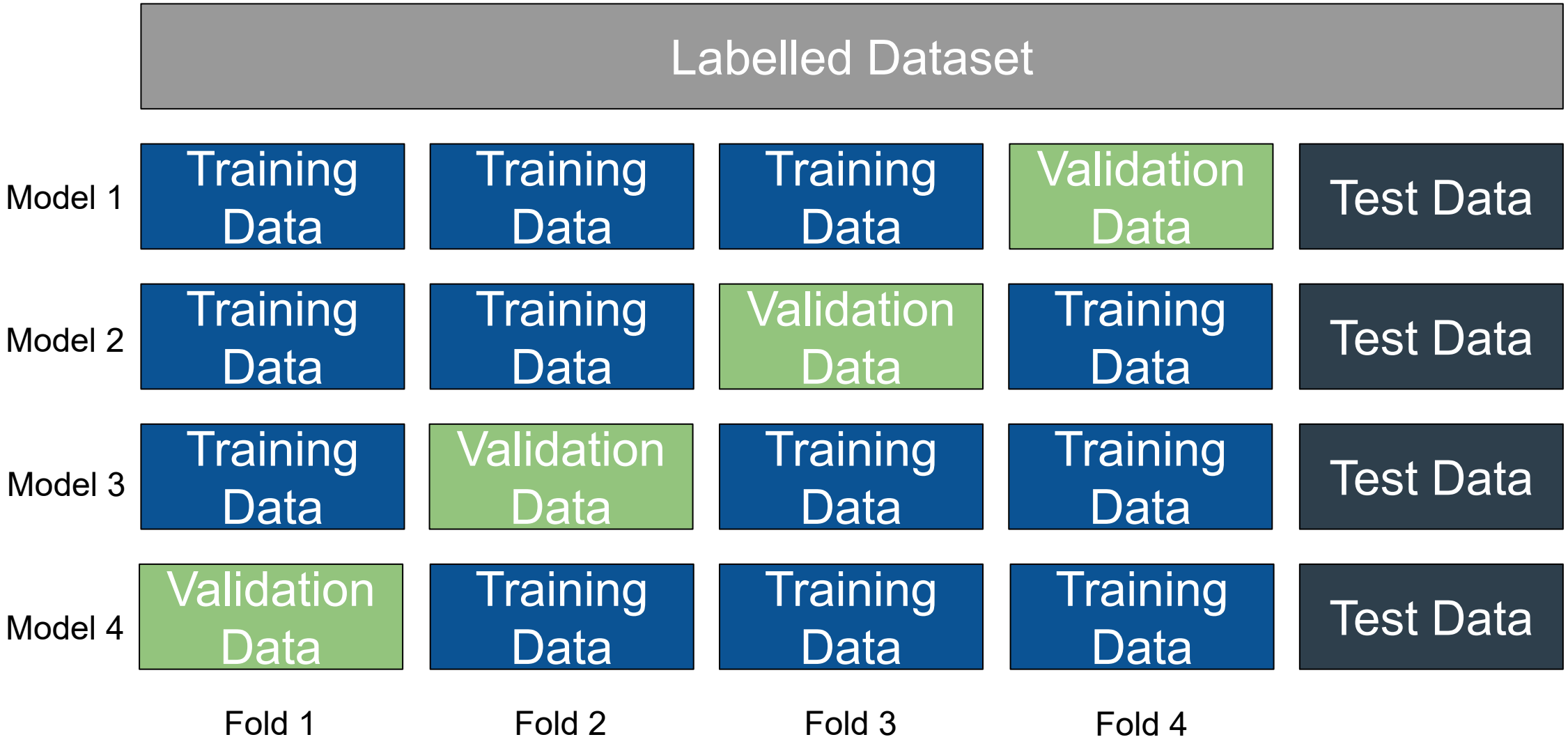
Obtaining Data to Test On



Validation on Small Dataset

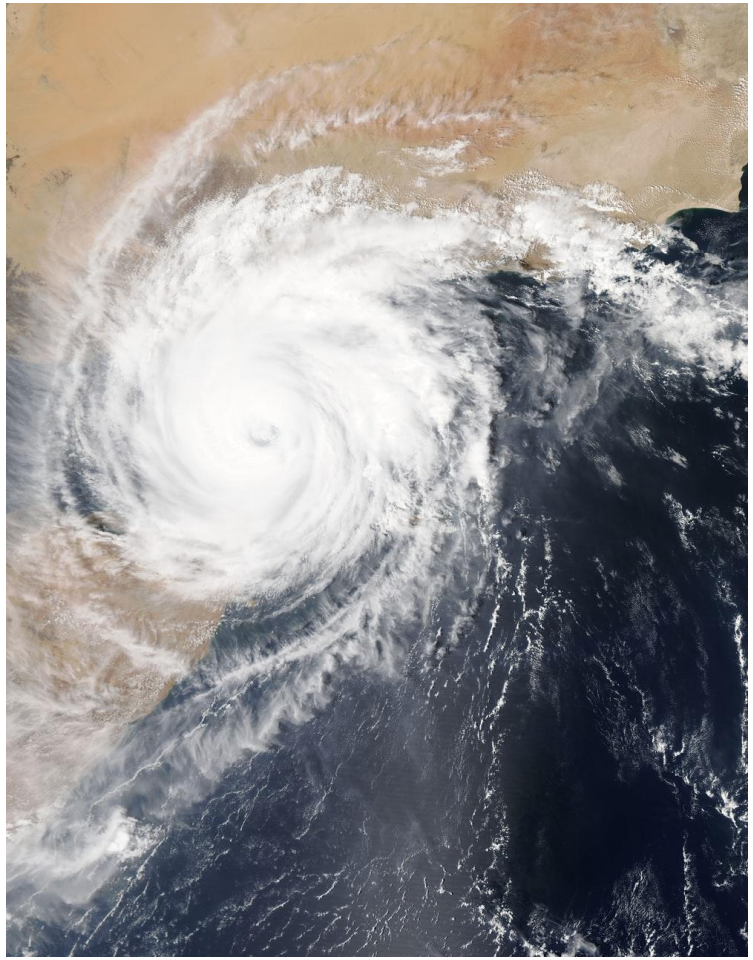


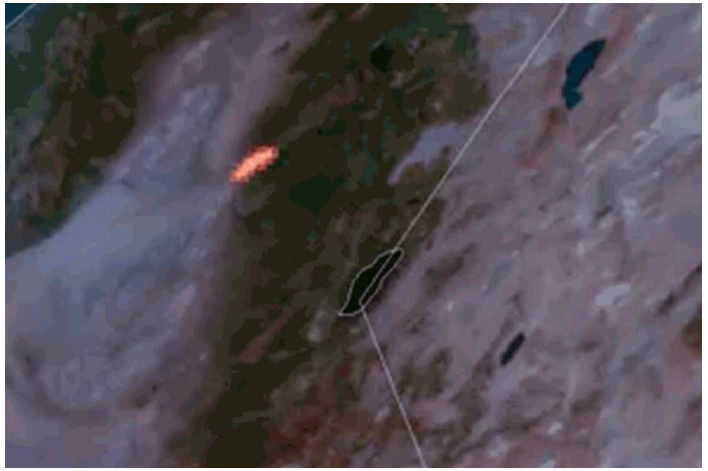
Cross-Validation



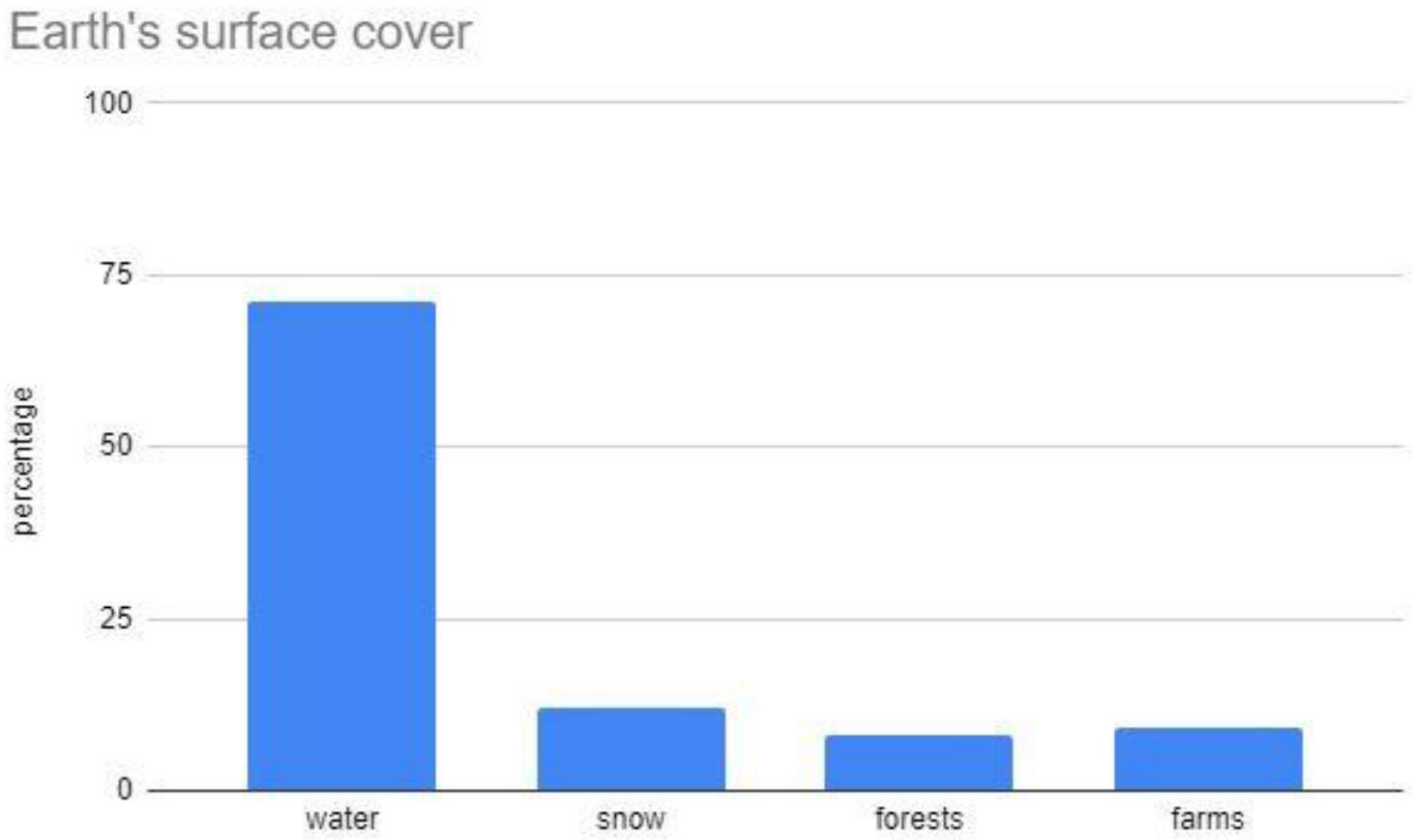
Imbalanced and Heterogeneous Data

Examples for Imbalanced Data

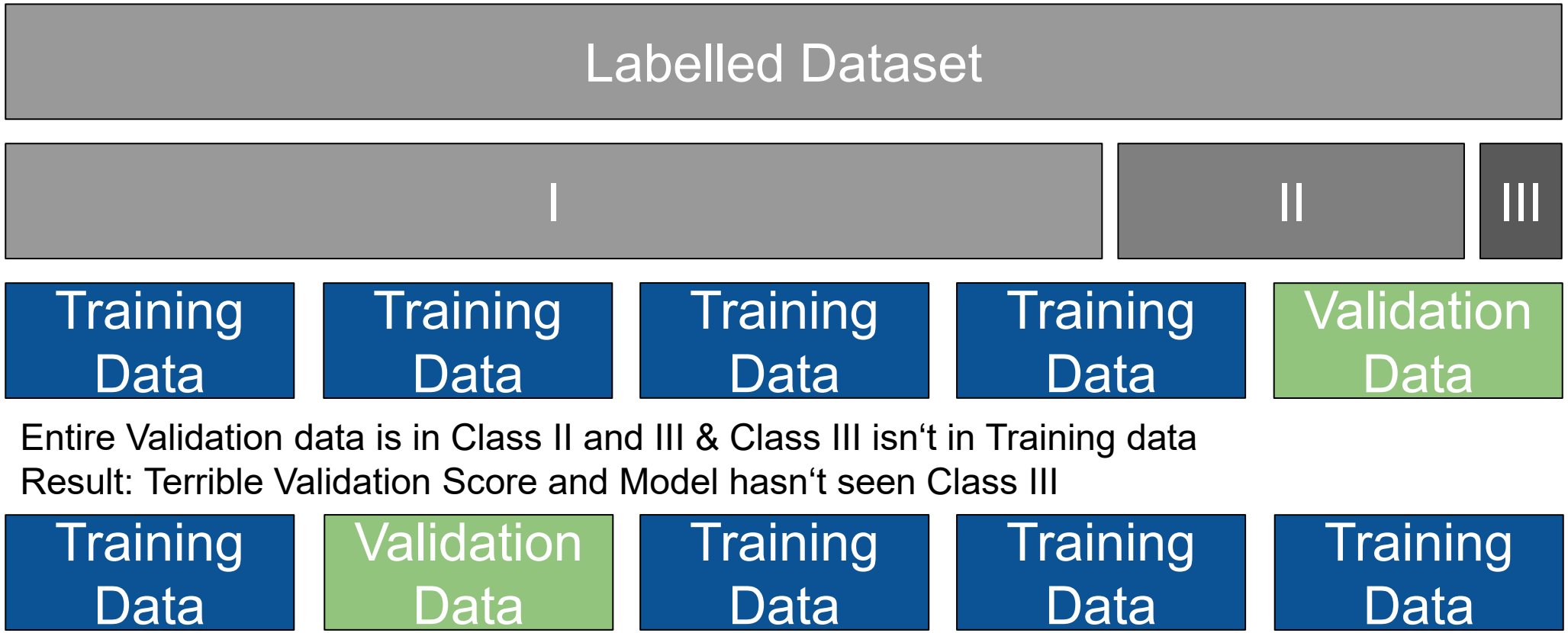




Class Imbalance



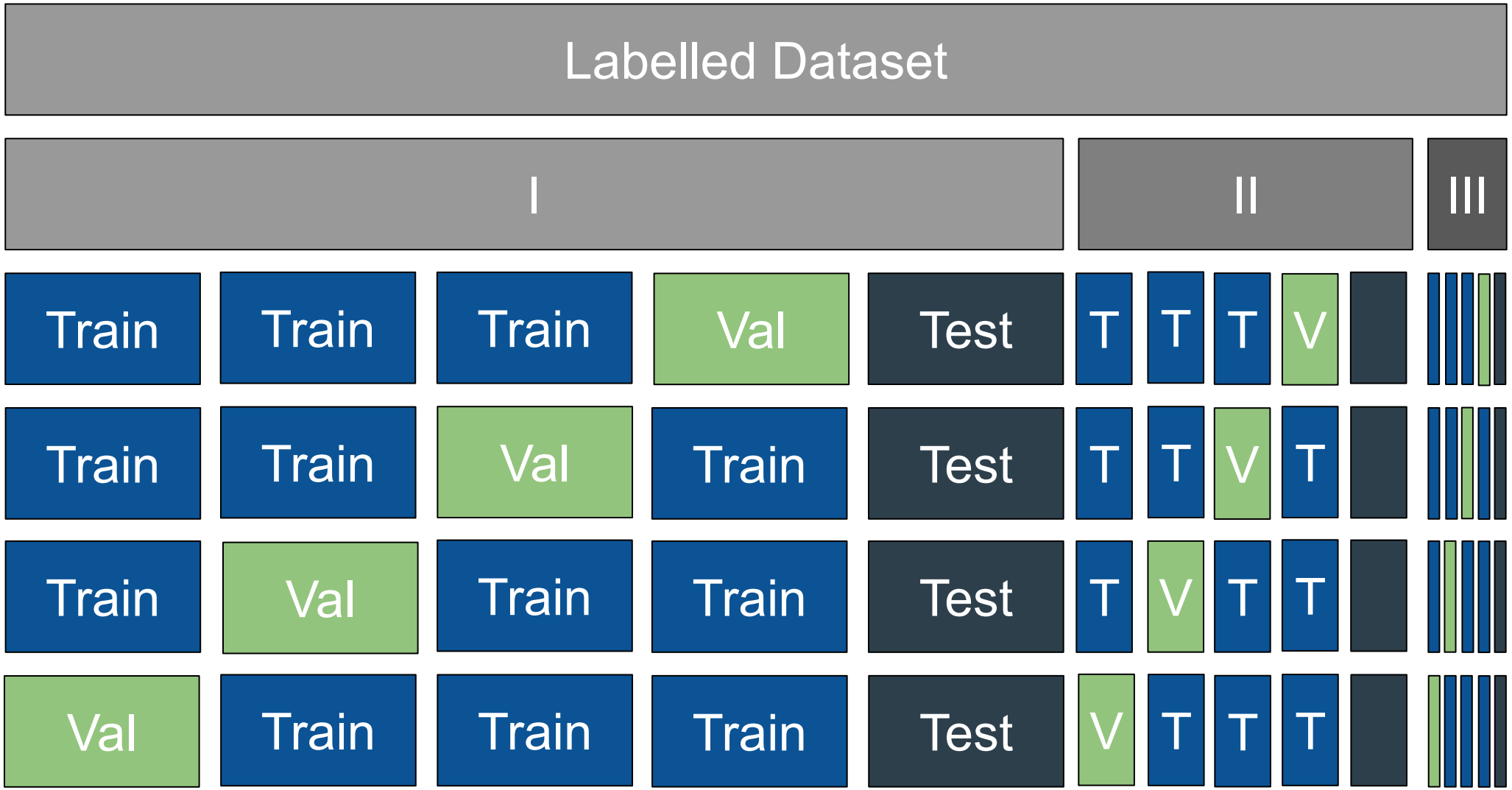
Why not use Random Sampling like before?



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

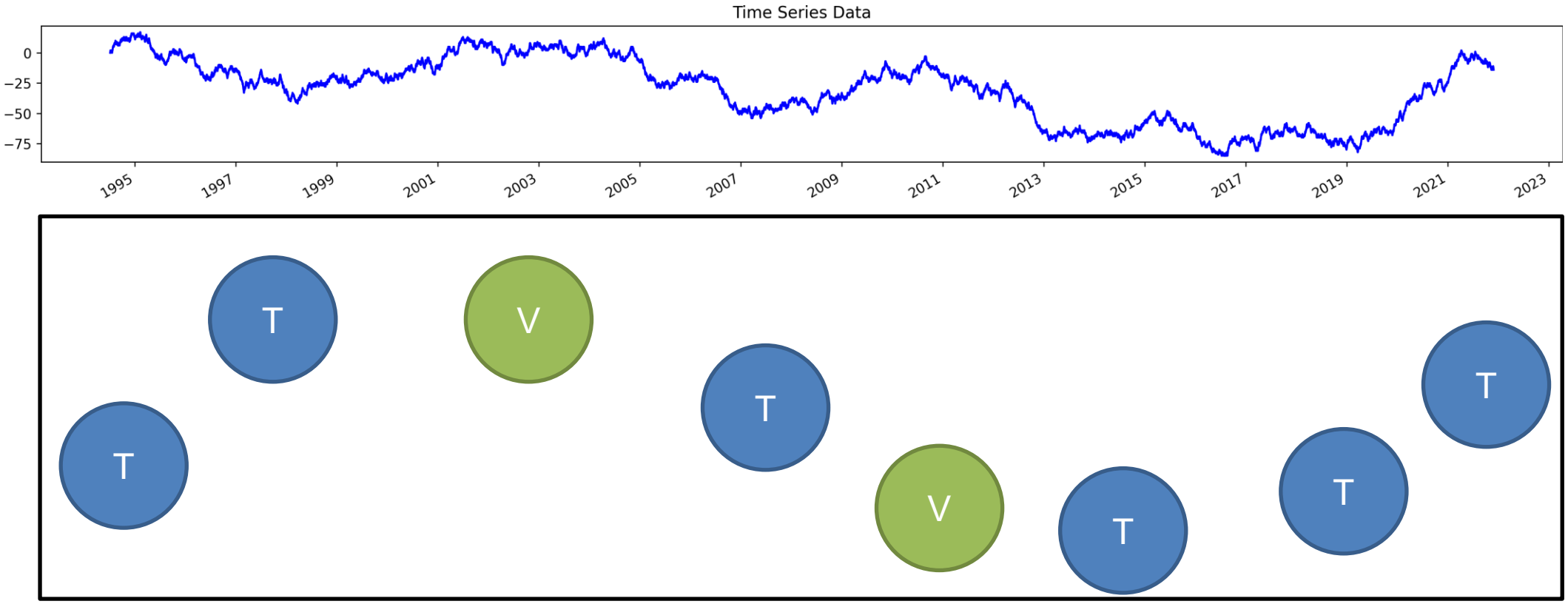
Entire Validation data is in Class I
Result: Great Validation Score but no validation of Class II & III at all

Stratification for Imbalanced Data



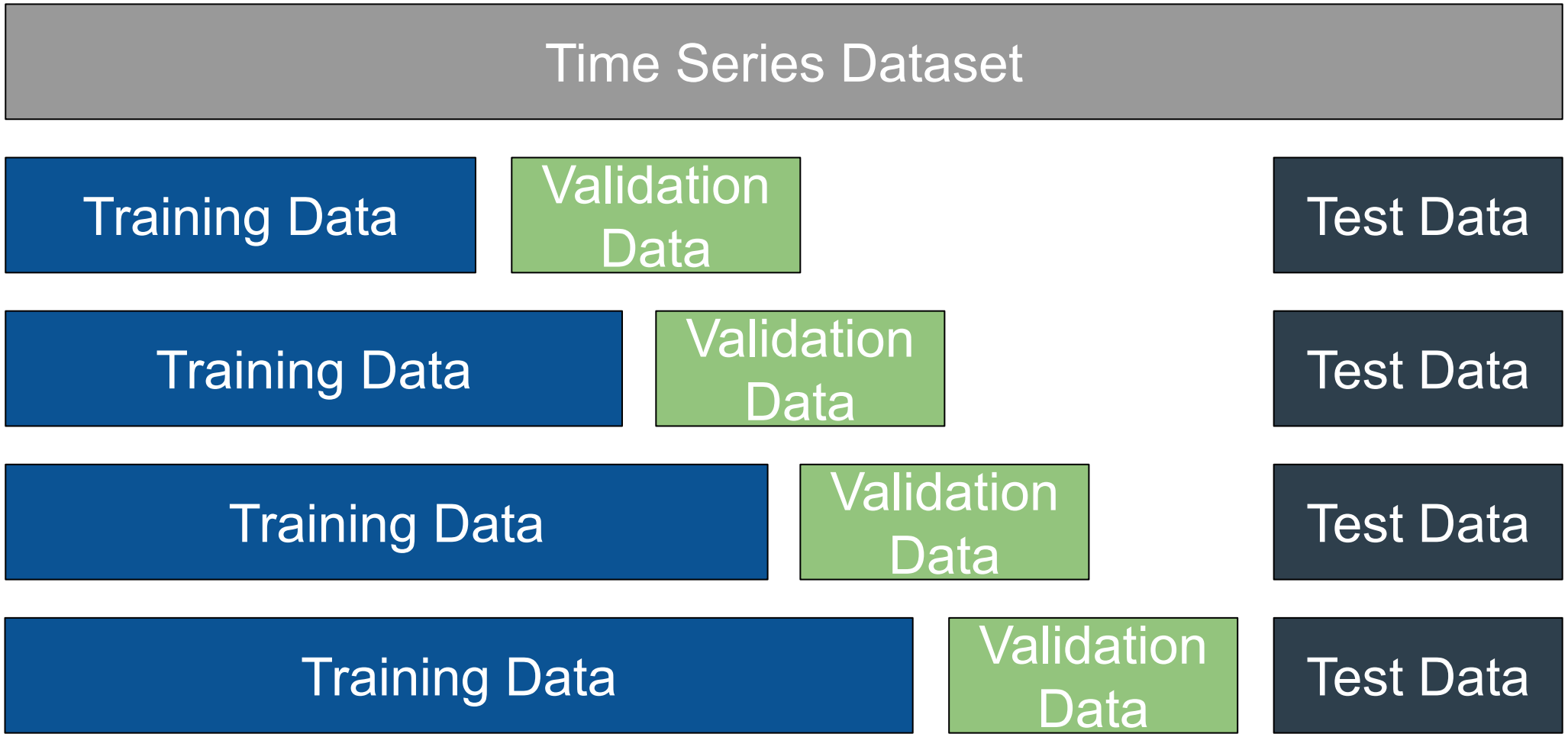
Correlated and Connected Data

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

- Shifts in Input Data
- Examples
  - Global Temperature through Climate Change
  - Land Cover Change through Urbanisation
- Mitigation Strategies
  - Monitoring of Input Data Distribution
    - Continuous, e.g. Kolmogorov-Smirnov test
    - Categorical, e.g. Chi-squared test
  - Automatic Retraining of ML Models
    - Define Threshold for Monitored Metrics
    - Implement periodic retraining

Target Drift

- Shifts in the Target / Label Data
- Examples
  - Reclassification from Human Labellers
  - Regulatory Changes
- Mitigation Strategies
  - Monitoring of Output Data Distribution
  - Automatic Retraining of ML Models
  - Anticipate Class Changes if Probable
    - Set Up Pipelines for Label adaption
    - Make it Easy to Change Label Processing

Concept Drift

- Shift in underlying „connection“ between input data and labels
- Example
  - Shopping Behaviour in 2020
  - Radiation Models using Rayleigh Scattering then changing Wavelengths
- Devastating for Machine Learning
- Mitigation Strategies
  - Monitor raw Model Metrics
  - Set up Alerts for Deterioration
  - Be prepared to take Model out of Production

“The true Test Set is in Production.”

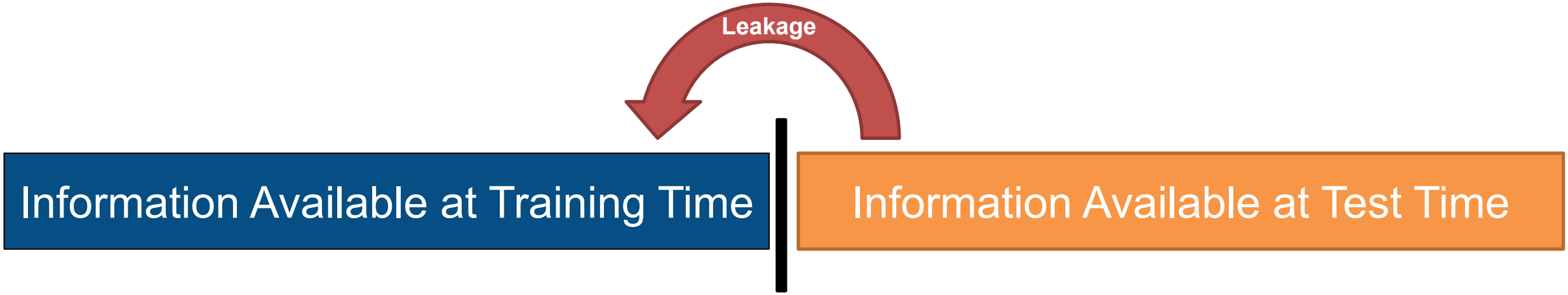
Baseline Methods and Model Verification

Baseline Model Verification

- Compare Machine Learning Results to known Baseline
- Build Baseline ML Models to assert Overfitting
- Consult Domain Experts with Feature Importances
- Test on Completely Unseen Data

Practical considerations in Snooping and Data Leakage

Data Leakage



Example from Pascal-VOC image prediction

Data Leakage Examples

- Labels or Proxy Features for Labels in Training Data
  - „Monthly Salary“ in Training to predict label „Yearly Salary“
  - „Num Late Payments“ in customer data to model a loan decision
- Normalisation on Validation & Test Data
- Duplicate Rows in Training and Test sets
  - Data Augmentation
  - Oversampling or SMOTE on Imbalanced Data
- Manual Preprocessing of Entire Data based on Known Properties

Conclusion

Conclusion

- Split Data Immediately into
  - Training
  - Validation
  - Test
- Try Cross-validation for Best Results
- Beware of Correlated Data when Splitting
  - Time Series & Geospatial Data are Always Correlated
- Use MLOps to Mitigate Drift
- Build Baseline Models and Consult Experts
- Beware of Data Leakage