Model Selection (and Validation)

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Reminders

This week:

- Wednesday Session
  - Won’t have one this week, but will talk through update 1 on Thursday
- If you didn’t read it for today, read for Thursday: Cross-Validation Strategies

Coming up next week:

- Monday: Project Update 2 (will be posted on canvas soon)
- Tuesday: Weekly Feedback Form

Reminder: Office Hours

Instructor Office Hours (course content, general questions, etc)

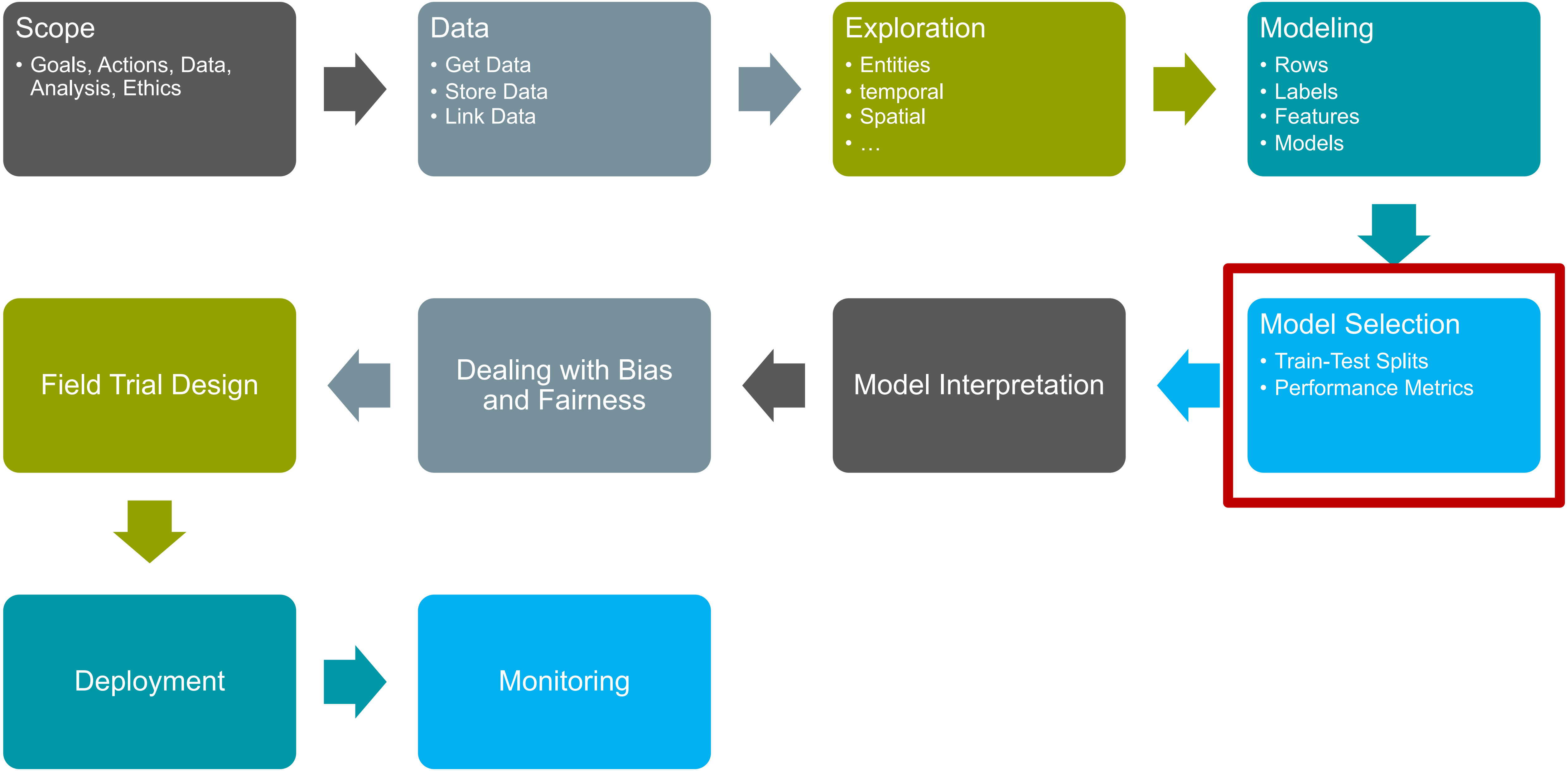
- Rayid (in GHC 8023): Tuesdays 12-1, Wednesdays 2-3
- Kit (in GHC 8018): Wednesdays 11-12, Thursdays 12-1

Infrastructure and Tech Setup Office Hours with the TAs:

- Riyaz (8th floor GHC, by printers): Mondays 12-1, Fridays 10-11
- Abhishek (8th floor GHC, by printers): Mondays 11-12, Fridays 2-3

Plan for the week

- What you should be discussing this week within your team
  - Finalizing your analytical formulation and baselines to compare against
  - Validation strategy
- What you should be building this week
  - V0 of your ML pipeline
  - Training/validation logic
  - Implemented Baselines



Scope

- Goals, Actions, Data, Analysis, Ethics

Data

- Get Data
- Store Data
- Link Data

Exploration

- Entities
- temporal
- Spatial
- …

Modeling

- Rows
- Labels
- Features
- Models

Model Selection

- Train-Test Splits
- Performance Metrics

Model Interpretation

Dealing with Bias and Fairness

Field Trial Design

Deployment

Monitoring

How to solve a prediction problem

- Define and Create Rows (unit of prediction)
- Define and Create Label (outcome/target variable – what event and when?)
- Define and Create Features (features/predictors)
- Create Training and Validation/Test Sets
- Train model(s) on Training Set(s)
- Validate model(s) on Validation/Test Set(s)
- Select “best” model

What is the goal of model selection?

What is the goal of model selection?

- You’ve run a large number of different types of models varying model types, hyperparameters, features, … (other decisions in the pipeline)

- Now we need to
    - Understand what types of models are effective under what circumstances
    - Decide which one(s) to use in the **future**

What do we need our selected model to do?

What do we need our selected model to do?

- Perform well
  - What metric?
  - Compared to what?
- Generalize
  - To what?

What should the model we select generalize to?

What do need to know to perform model selection

- Deployment scenario
  - Model Selection Methodology
  - Metric(s) (need to match your initial goals)
- Comparison with baselines (to know if you’re effective)

How do we select a model that does that?

Model Selection - Methodology

- In-sample
- Out of sample
- Multiple Out-of-sample (Hold-out) Splits
- Cross Validation
    - Leave one out (LOO)
    - K fold
- Temporal Holdouts
- Spatial Holdouts
- Other Holdouts?

Scenarios

1: We want to track news coverage of epidemic related topics. We have tagged a small corpus (n=1000) of news articles from Jan 2019 to September 2020 and now have a stream of new incoming articles every day. We want to deploy a system that tracks the intensity of coverage by media outlet going forward.

What should our model generalize to?

What is a training set?

What is the corresponding validation set?

Scenarios

2: We want to track news coverage of epidemic related topics. We have tagged a small corpus (n=1000) of news articles from Jan 2019 to September 2020 and now have a stream of new incoming articles every day. We want to deploy a system that tracks intensity of coverage by media outlet over the last 2 years as well as going forward.

What should our model generalize to?

What is a training set?

What is the corresponding validation set?

Scenarios

3: We want to predict whether there will be an increase in epidemic related articles in the media during the next week.

What should our model generalize to?

What is a training set?

What is the corresponding validation set?

Scenarios

4: We want to predict whether there will be an increase in epidemic related articles in the media during the next month.

What should our model generalize to?

What is a training set?

What is the corresponding validation set?

Parameters

- How far back to go when training models? (max training history)
    - To the beginning of time (expanding training window)?
    - Fixed history (rolling training window)?
    - Something else?
    - How far back do you get your features from?
- How much to move forward from train-validation pair 1 to train-validation pair 2?
    - A day?
    - A month?
    - Something else?

Other considerations

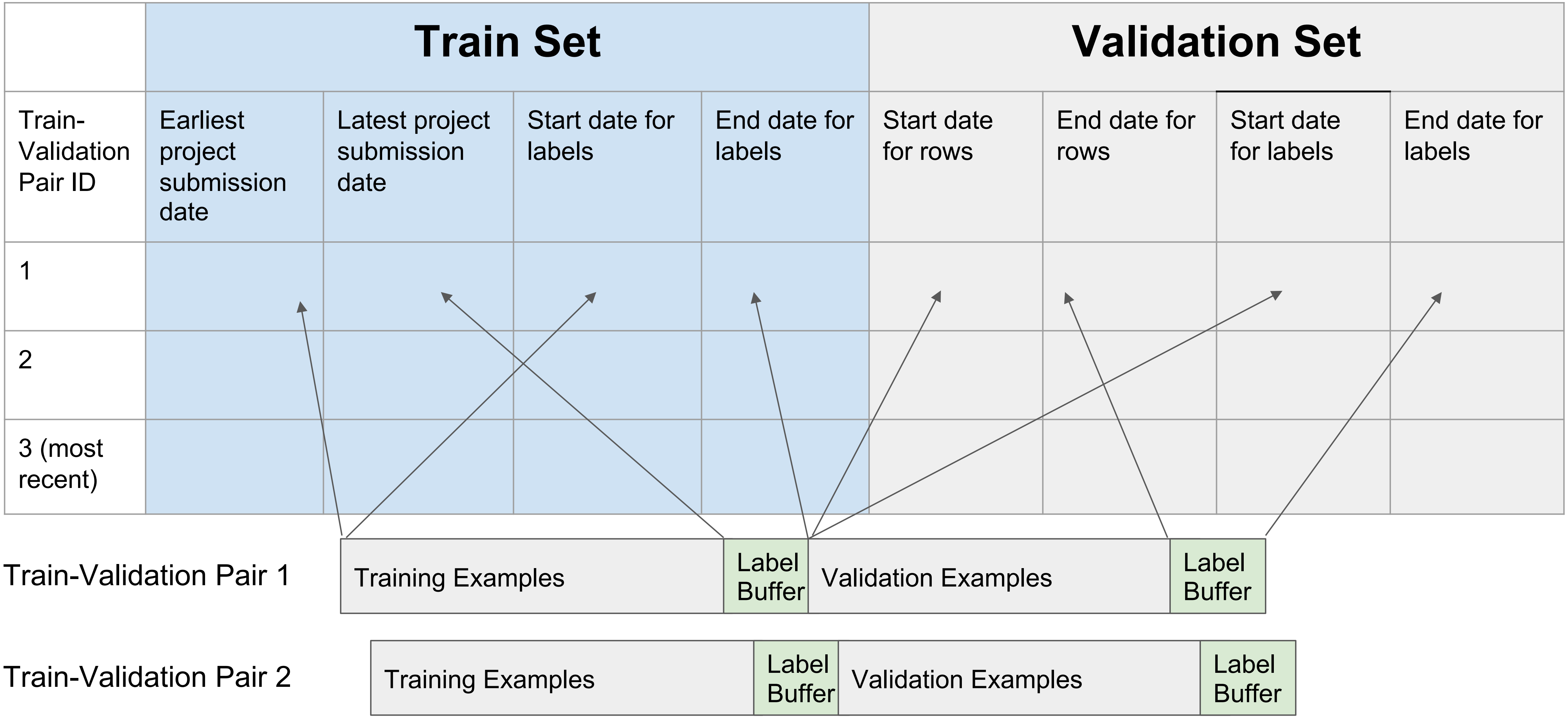
- If making repeated predictions about the same entity at different times, how often should an entity be repeated in the training data?
  - In an event-based deployment setup?
  - In a “take action at regular-ish intervals” deployment?
- What about in the validation set?

Some tips

- Set up validation set(s) to match deployment scenarios (and constraints)

- Set up training set(s) any way we want but match data (both features and labels) available at training time
  - Making sure labels are not censored based on label period
  - Sampling (if helpful)
  - Data collection and update lag

Train Validation Pairs



	Train Set				Validation Set			
Train-Validation Pair ID	Earliest project submission date	Latest project submission date	Start date for labels	End date for labels	Start date for rows	End date for rows	Start date for labels	End date for labels
1								
2								
3 (most recent)								

Train-Validation Pair 1: Training Examples | Label Buffer | Validation Examples | Label Buffer

Train-Validation Pair 2: Training Examples | Label Buffer | Validation Examples | Label Buffer

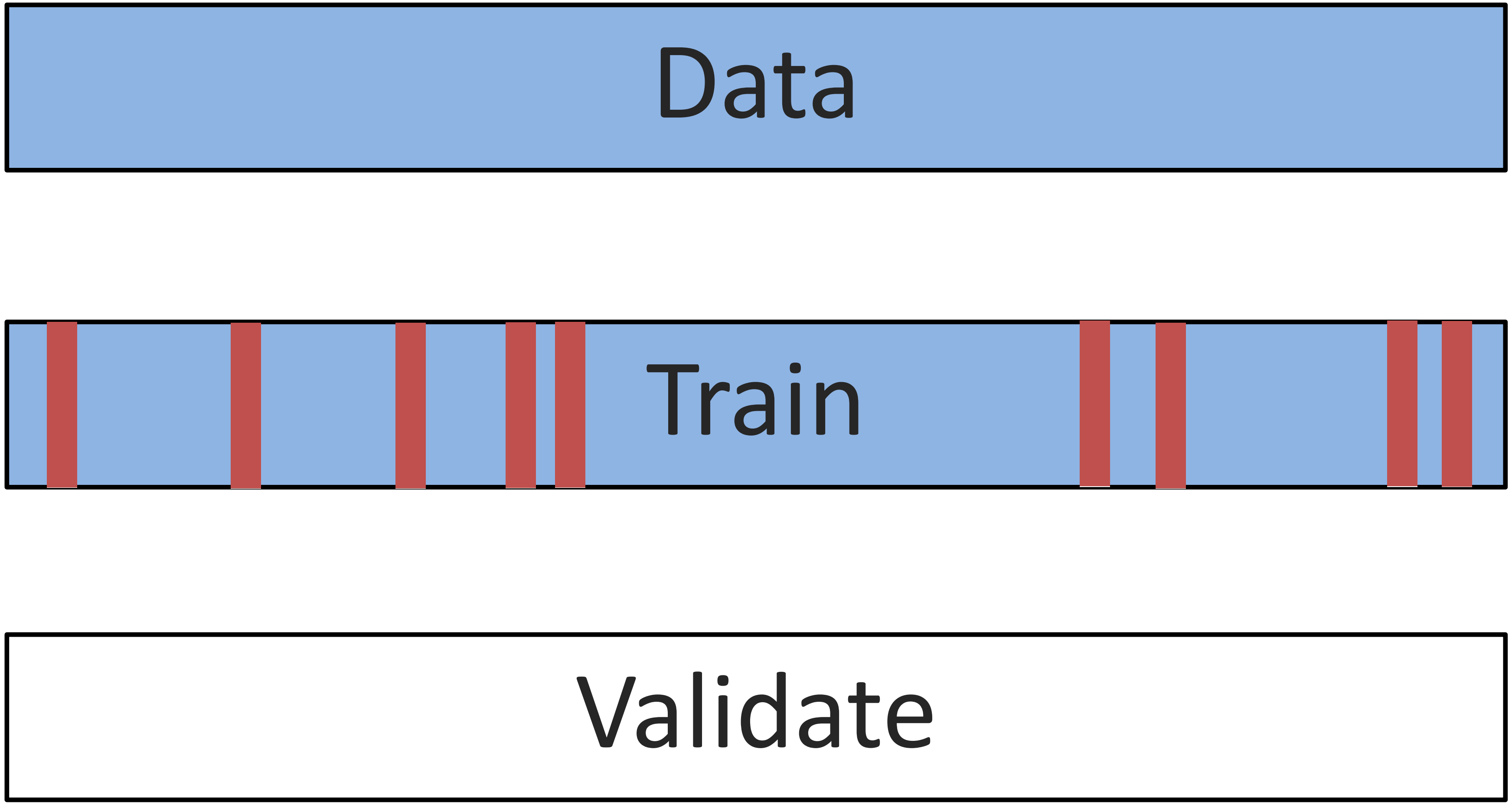
In-sample

Data

Train

Validate

Out-of-Sample

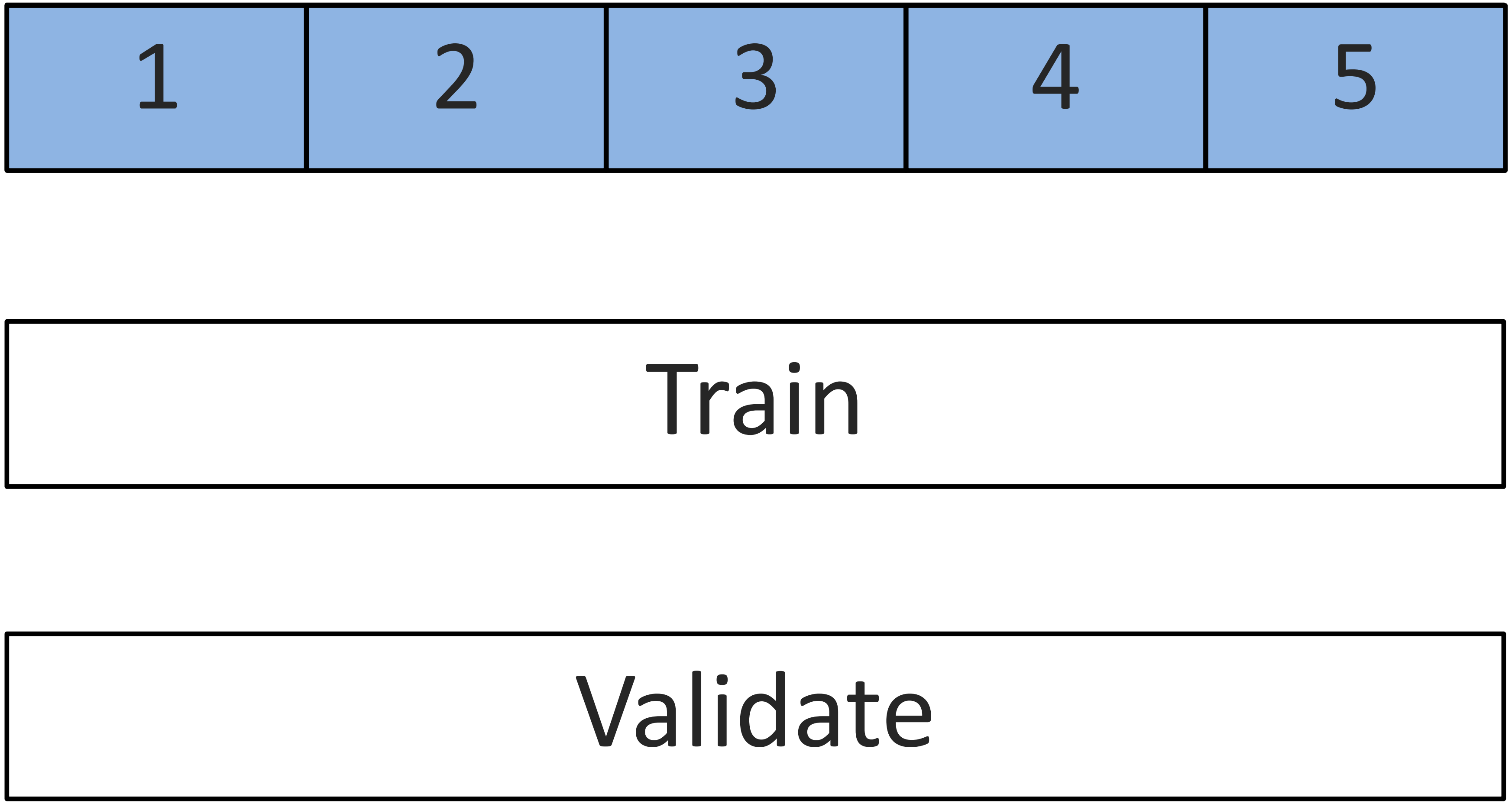


Data

Train

Validate

N-fold Cross-Validation

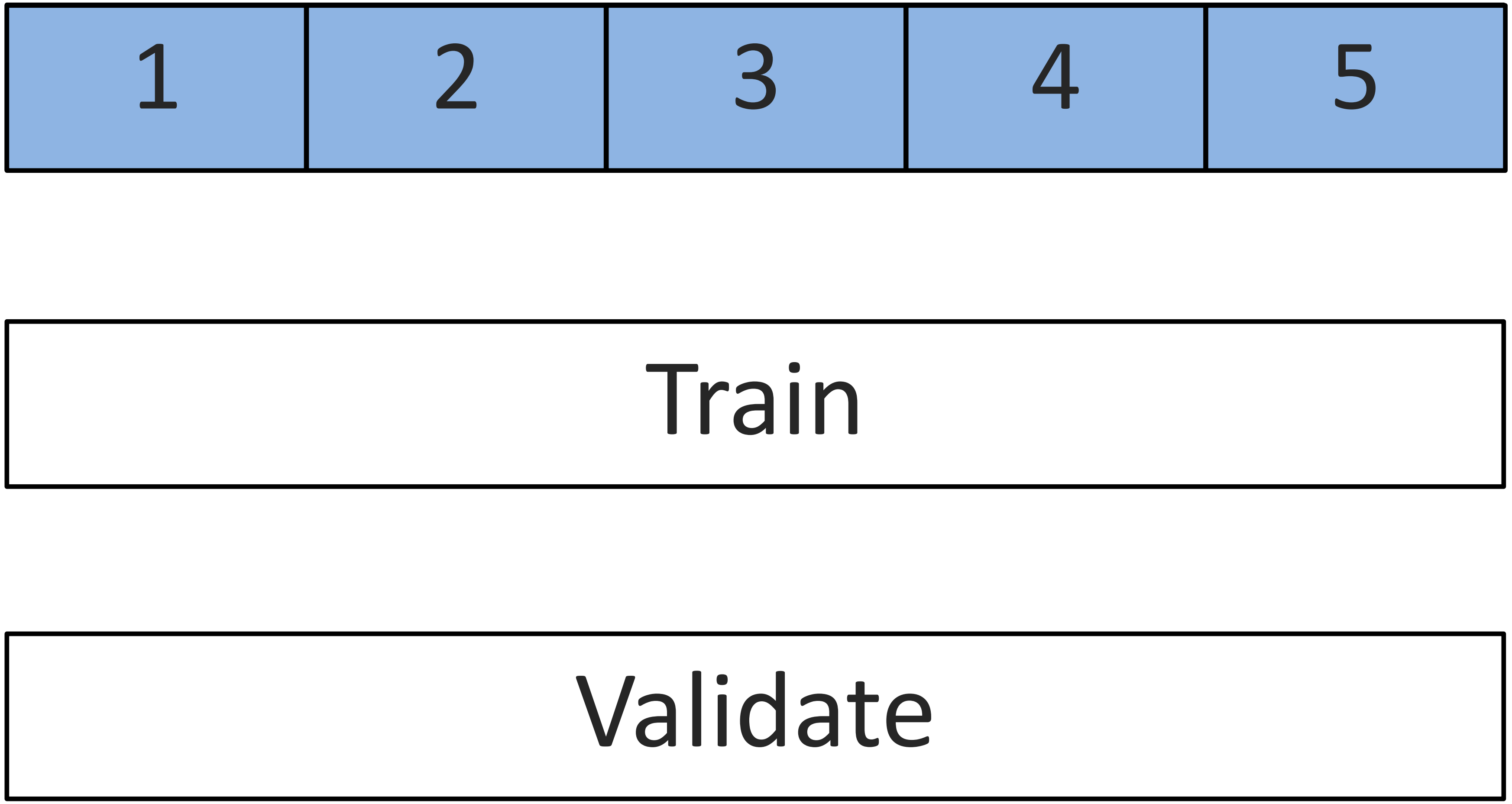


1 2 3 4 5

Train

Validate

N-fold Cross-Validation

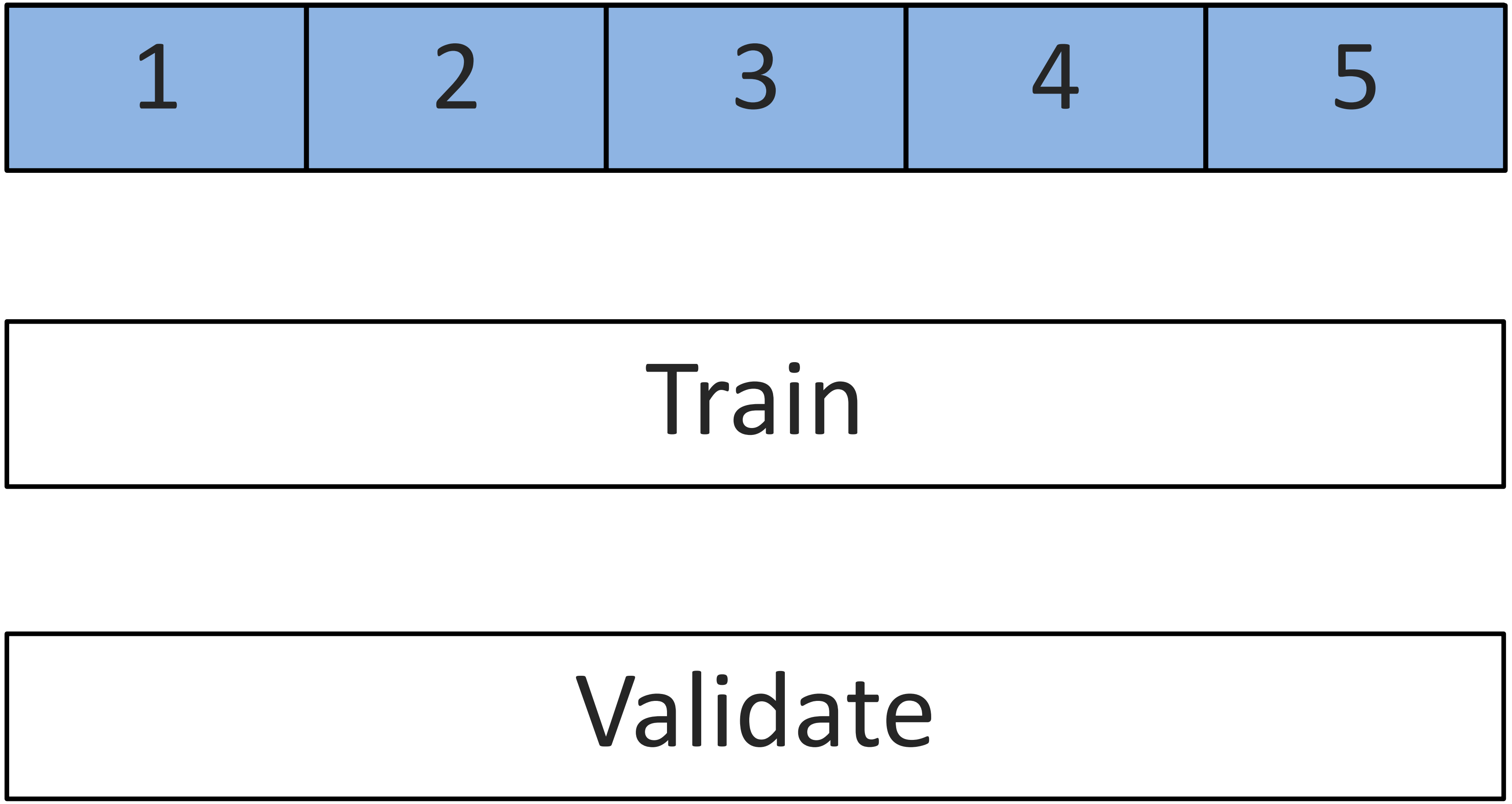


1 2 3 4 5

Train

Validate

N-fold Cross-Validation

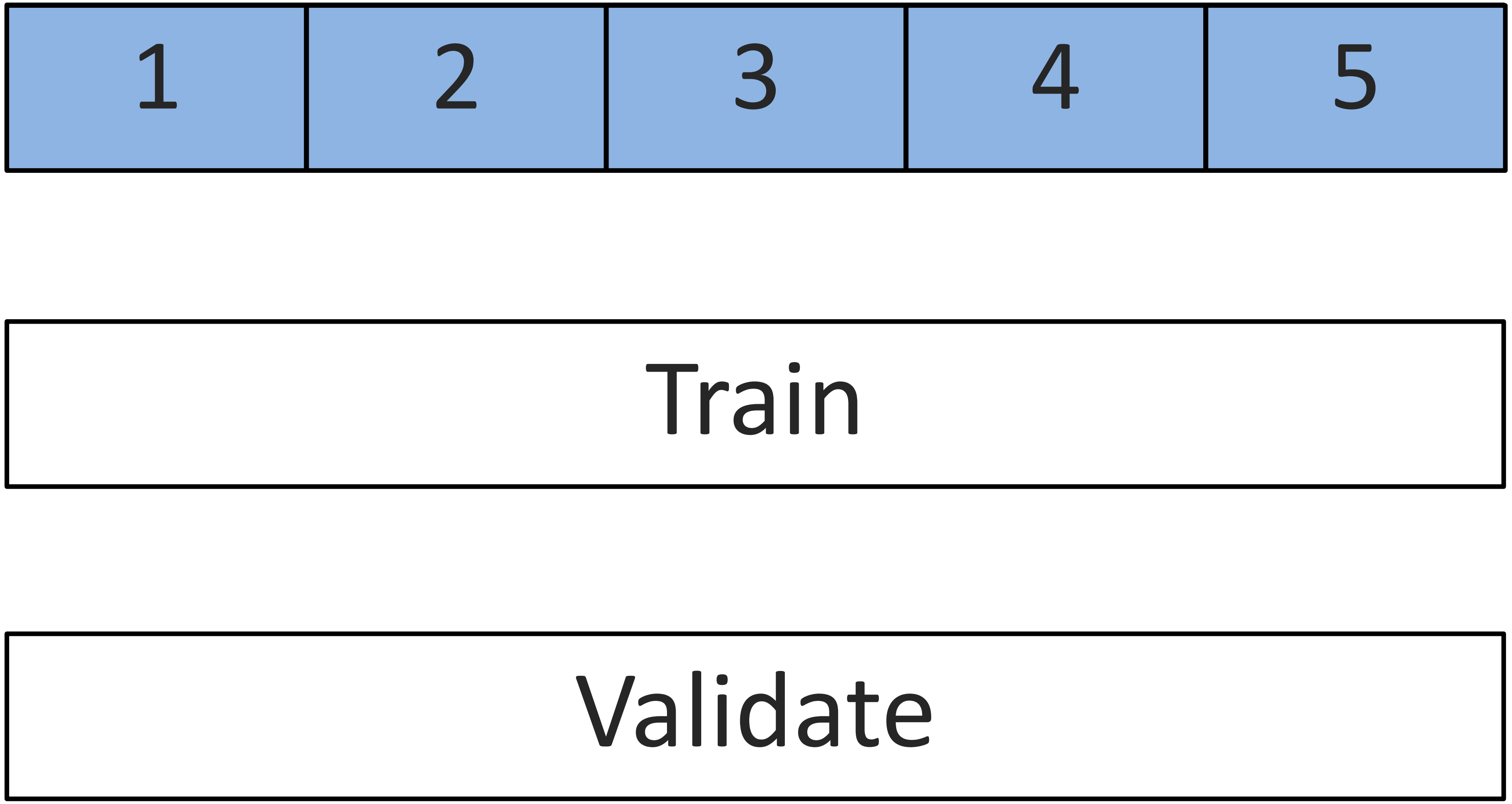


1 2 3 4 5

Train

Validate

N-fold Cross-Validation

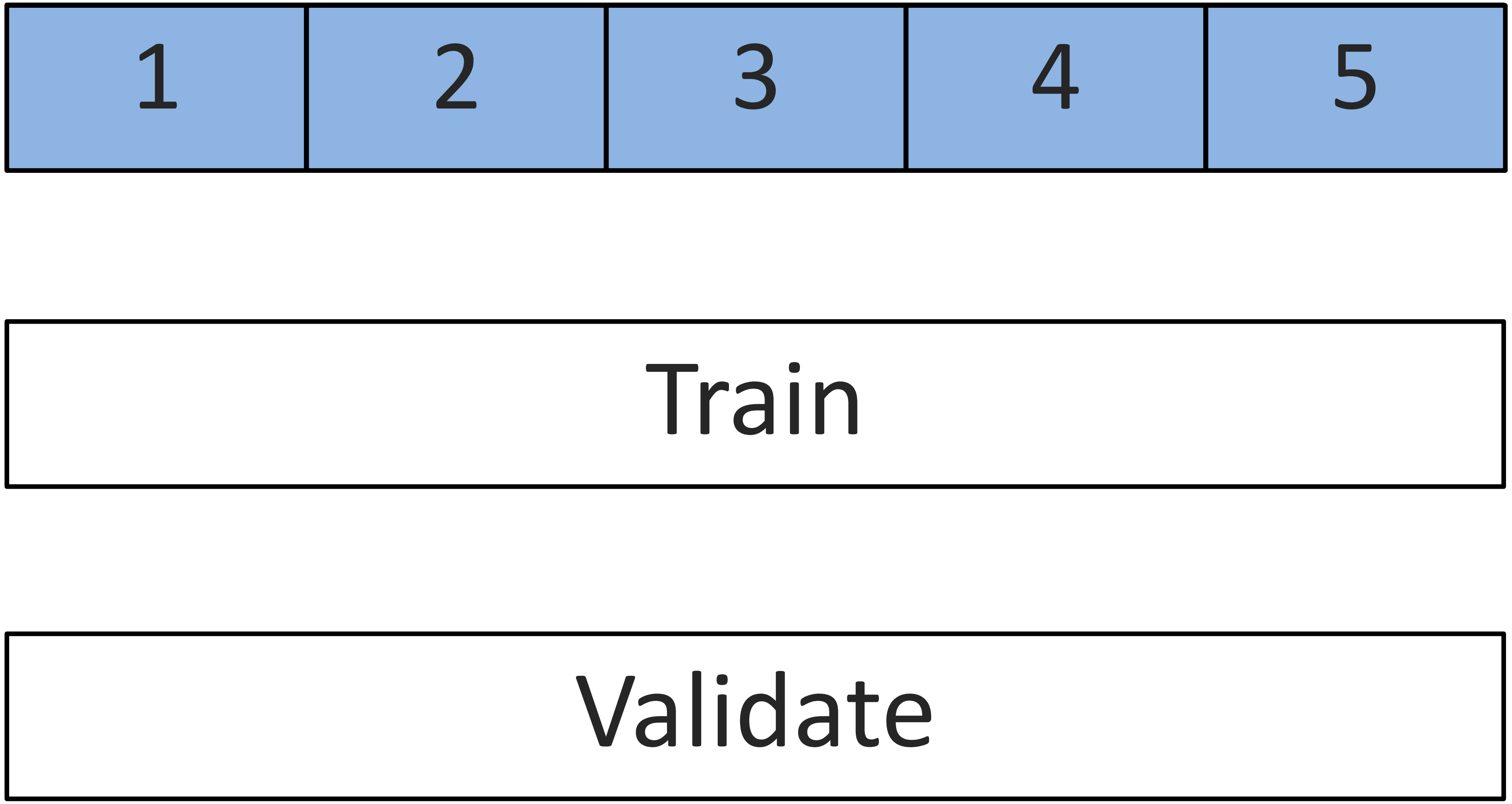


1	2	3	4	5

Train

Validate

N-fold Cross-Validation

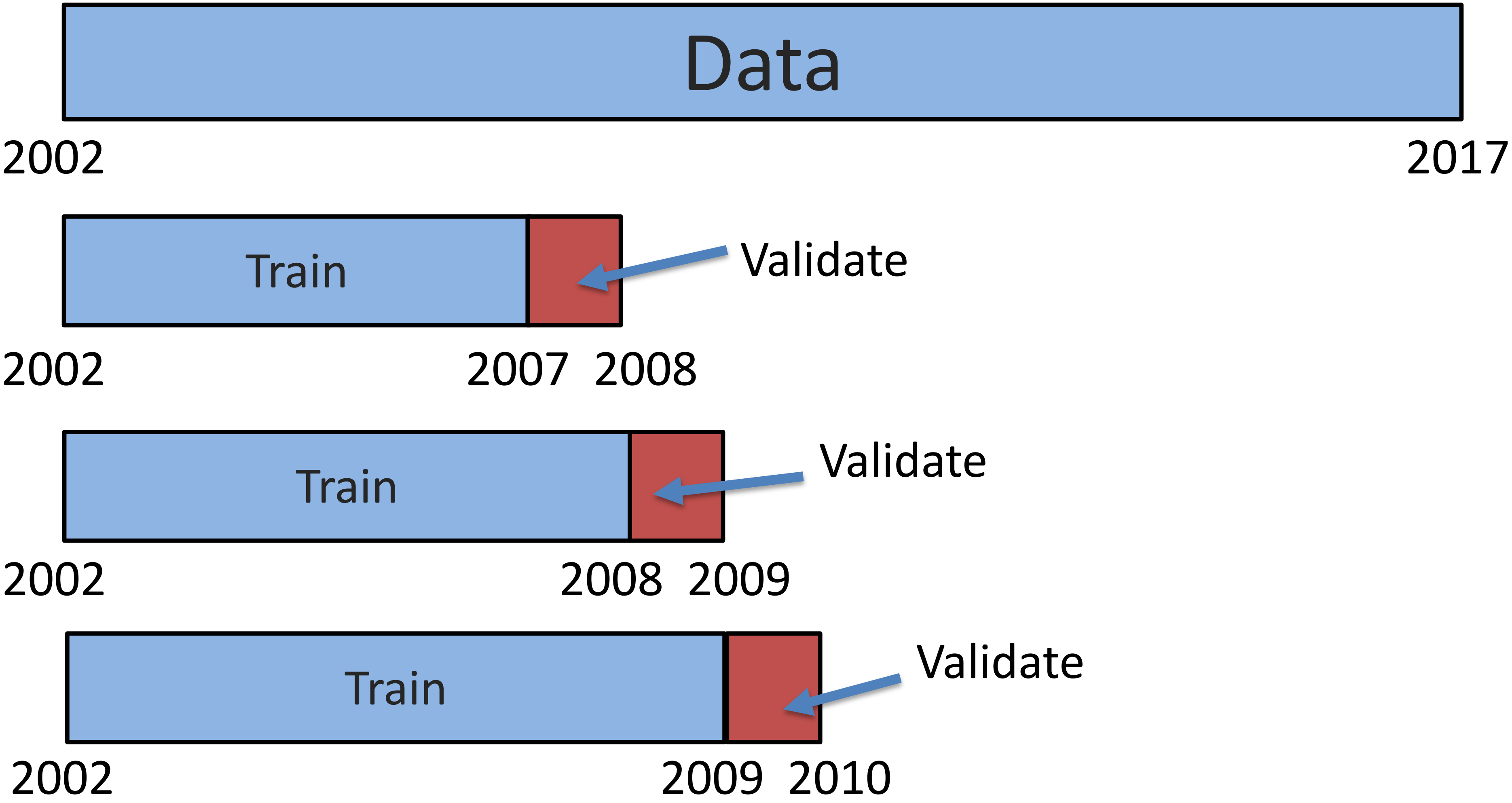


1	2	3	4	5

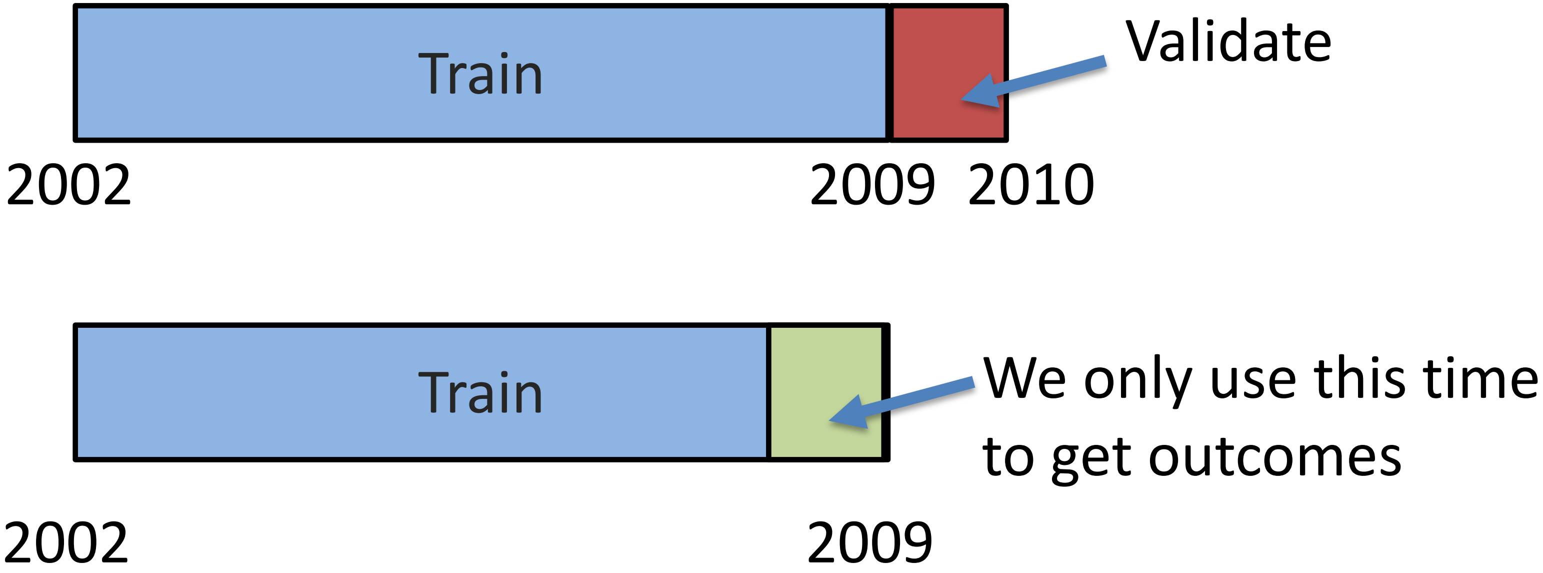
Train

Validate

Temporal Holdouts

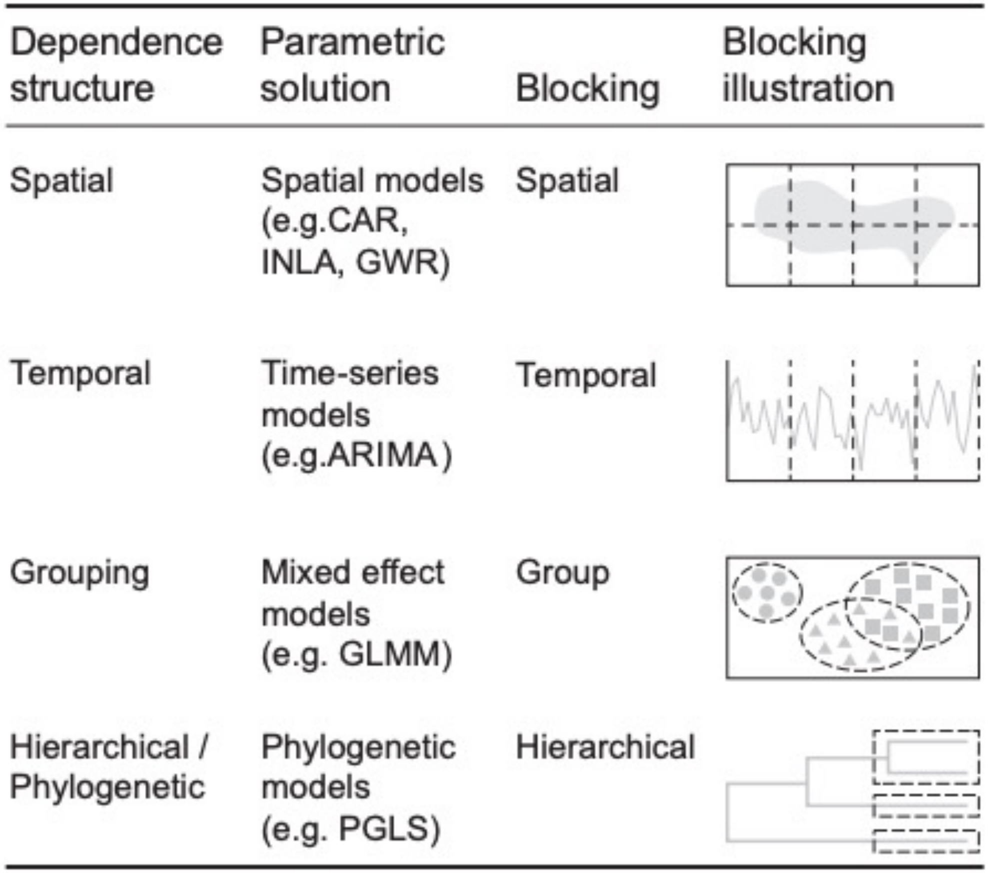


Training - Time splits



Evaluation - Methodology

- In-sample
- Out of sample
- Multiple Out-of-sample (Hold-out) Splits
- Cross Validation
  - Leave one out (LOO)
  - K fold
- Holdouts Using Structure of Data
  - Temporal
  - Spatial, Hierarchical, etc.



Dependence structure	Parametric solution	Blocking	Blocking illustration
Spatial	Spatial models (e.g.CAR, INLA, GWR)	Spatial	
Temporal	Time-series models (e.g.ARIMA)	Temporal	
Grouping	Mixed effect models (e.g. GLMM)	Group	
Hierarchical / Phylogenetic	Phylogenetic models (e.g. PGLS)	Hierarchical	

Figure 1. Examples of dependence structures, parametric solutions to parameter estimation, and the associated blocking approaches for cross-validation to increase reliability of prediction error estimates.