Big Data and Machine Learning

👩‍💻 🤖 👨‍🏫

EDRS Open Science Pre-Conference Workshop

Shirley B. Wang September 17, 2021

Today's Topics

Provide a **conceptual introduction** to predictive modeling/machine learning.

Introduce common **terms** in machine learning.

Understand similarities and differences between **inference** and **prediction**.

Describe methods for increasing **rigor, reproducibility, and transparency** of machine learning approaches for big data.

Discuss methods for incorporating **open science** into the machine learning workflow.

Slides adapted from the Pittsburgh Summer Methodology Series Applied Machine Learning in R course, developed by Jeffrey Girard and Shirley Wang.

Conceptual Introduction

What is machine learning?

The field of machine learning (ML) is a **branch of computer science**.

ML researchers **develop algorithms** with the capacity to **learn from data**.

When algorithms learn from (i.e., are **trained on**) data, they create **models**.[1]

ML algorithms are often used to create **predictive models**.

The goal will be to **predict unknown values** of important variables **in new data**

Note that this differs from traditional inferrential statistics[2], which aims to **understand** and **explain** phenomena rather than **predict** it.

[1] ML models are commonly used for prediction, data mining, and data generation.
[2] For an excellent overview of explanation vs. prediction in psychology, I recommend Yarkoni & Westfall (2017).

Signal and Noise

A Delicate Balance

Any data we collect will contain a mixture of **signal** and **noise**

- The "signal" represents informative patterns that generalize to new data
- The "noise" represents distracting patterns specific to the original data

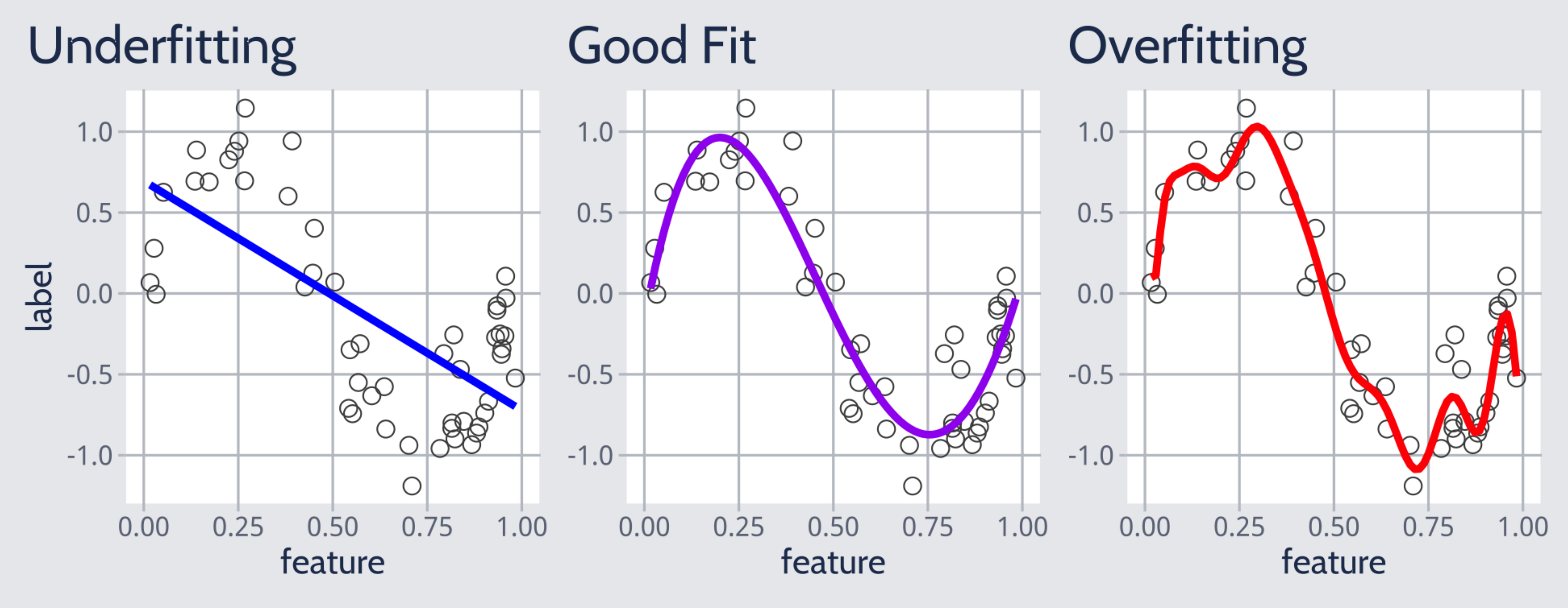
We want to capture as much signal and as little noise as possible

More complex models will allow us to capture **more signal** but also **more noise**

Overfitting: If our model is too complex, we will capture unwanted noise

Underfitting: If our model is too simple, we will miss important signal

Model Complexity



A Super Metaphor

What makes machine learning so amazing is its **ability to learn complex patterns**.

However, with this great power and flexibility comes the looming **danger of overfitting**.

Overfitting reduces **generalizability** and **reproducibility**. Thus, much ML research aims to detect and counteract overfitting.

For detection, we need two sets of data:

Training set: used to learn relationships

Testing set: used to evaluate performance

Bias-Variance Tradeoff

In ML, **bias** is a lack of predictive accuracy in the original data (the "training set")

In ML, **variance** is a lack of predictive accuracy in new data (the "testing set")

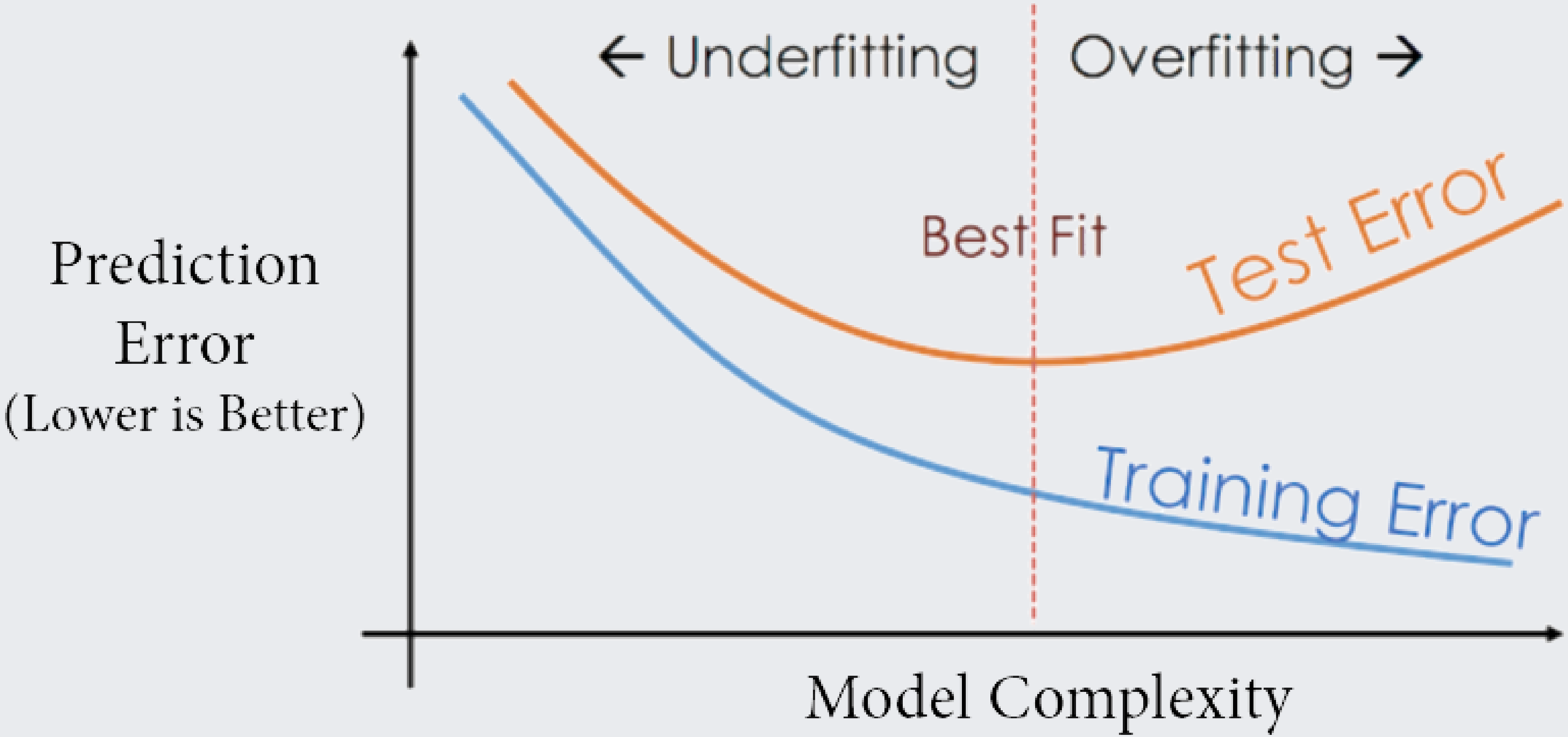
An ideal predictive model would have both low bias and low variance

However, there is often an inherent **trade-off between bias and variance**[1]

We want to find the model that is **as simple as possible** but **no simpler**

[1] To increase our testing set performance, we often need to worsen our performance in the training set.

A Graphical Explanation of Overfitting



A Meme-based Explanation of Overfitting

Countering Overfitting

Cross-Validation

There are some clever algorithmic tricks to prevent overfitting

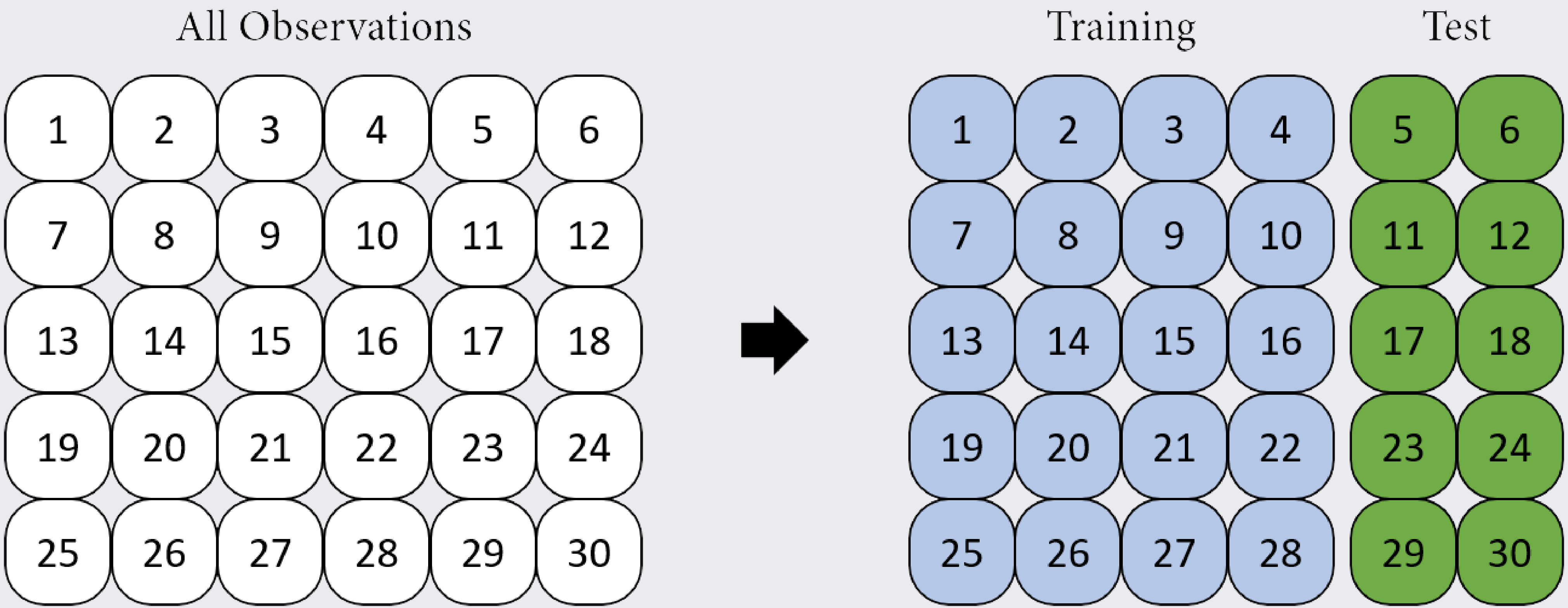
- For example, we can penalize the model for adding complexity

The main approach, however, is to use **cross-validation**:

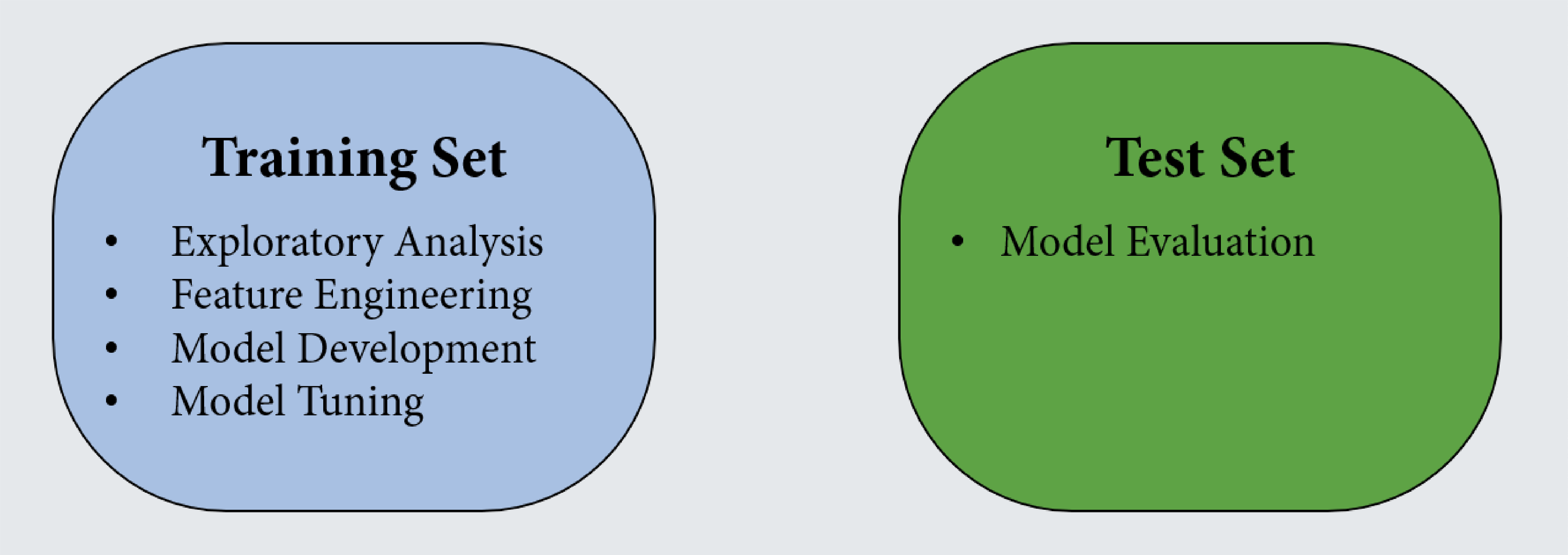
- Multiple **fully independent** sets of data are created (by subsetting or resampling)
- Some sets are used for training (and tuning) and other sets are used for testing
- **Model evaluation is always done on data that were not used to train the model**
- This way, if performance looks good, we can worry less about variance/overfitting

Caution: We still need to consider whether the original data was representative!

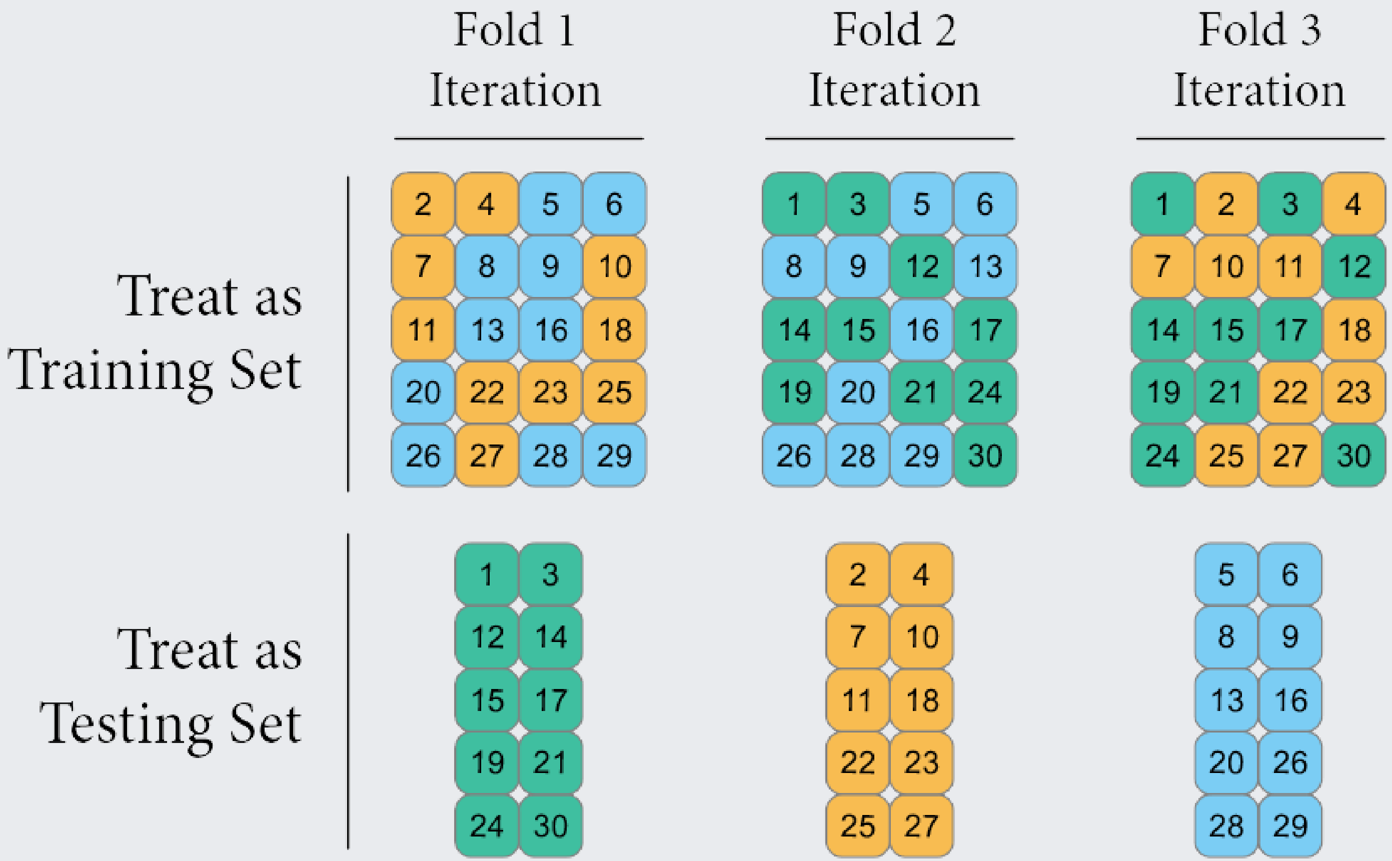
Holdout Cross-Validation



Holdout Cross-Validation

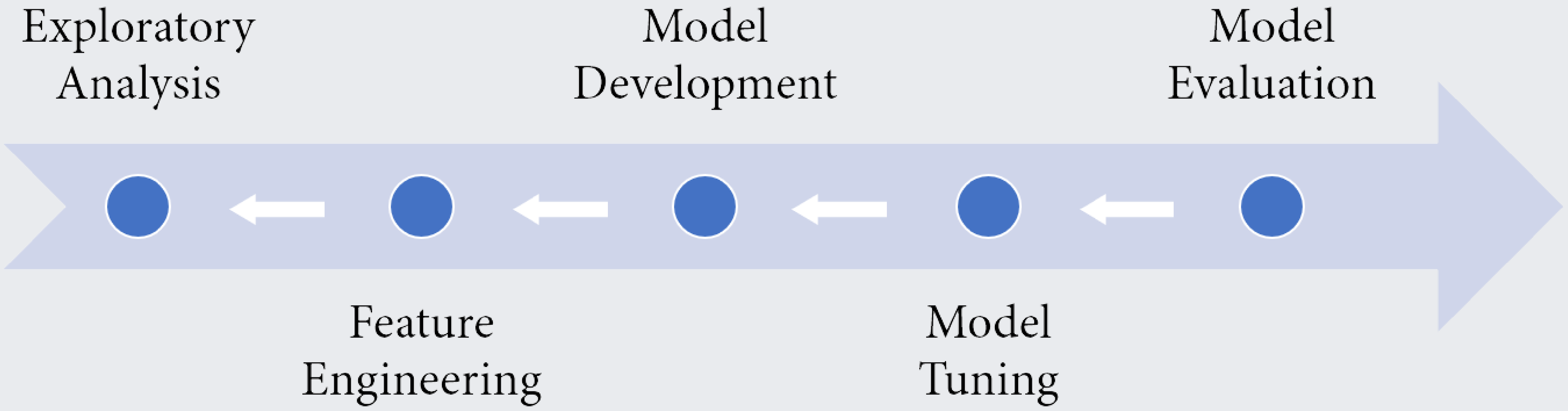


k-fold Cross-Validation



Modeling Workflow

Typical ML Workflow



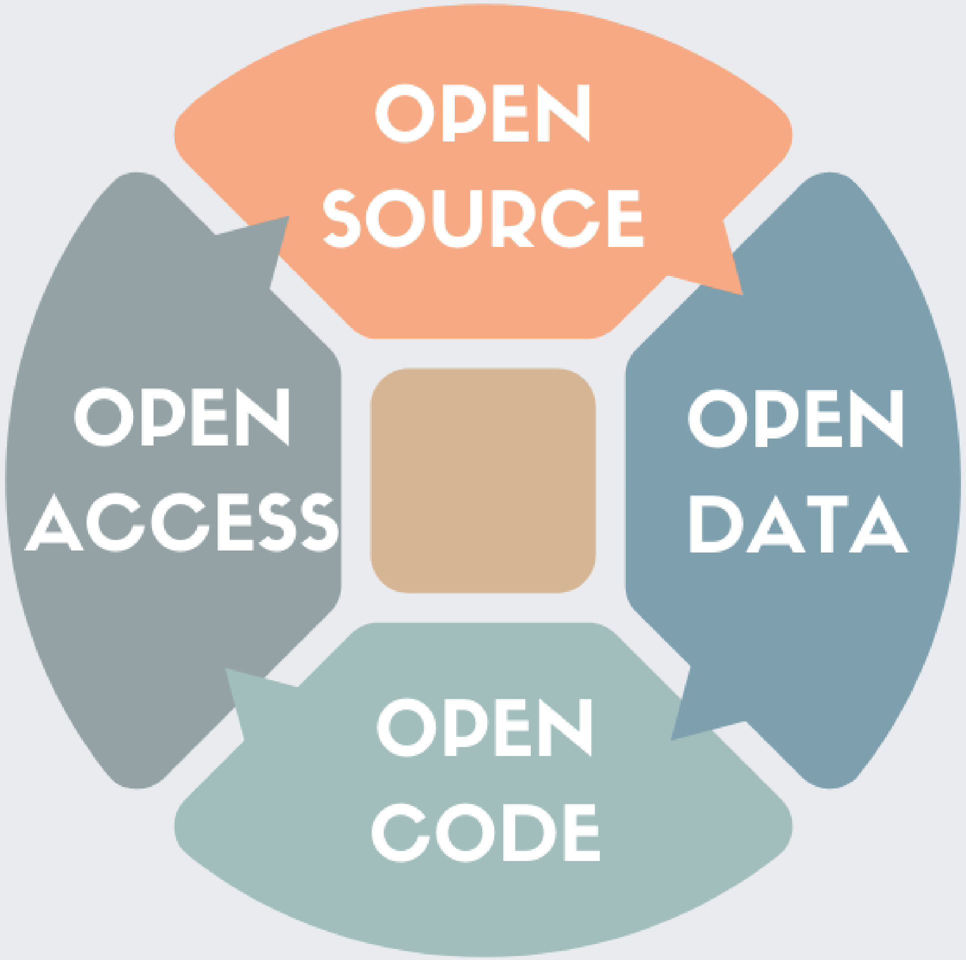
Incorporating Open Science into the ML Workflow

The development and evaluation of machine learning models should be **open**, **transparent**, and **reproducible**.

Preregistration of some methods (e.g., cross-validation procedure, model evaluation metrics, specific algorithms) can be useful.

Be transparent about **model limitations**, including limits to generalizability.

Consider potential for **harm**. Does the model combat or entrench societal injustices?



Thank you!

Shirley Wang
PhD Candidate, Harvard University
Clinical Psychology
Computational Science & Engineering

Email: shirleywang@g.harvard.edu

Twitter: @ShirleyBWang

Website: shirleywang.rbind.io

All materials available on github. For more, see https://pittmethods.github.io/appliedml/.