October 6th, 2022

Introduction to Machine Learning Part #2

Yusen He, PhD - DASIL Data Scientist

Prof. Julia Bauder – Director of DASIL

Martin Pollack – DASIL Post-Bac Fellow

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Intro to Machine Learning Part #2

AGENDA

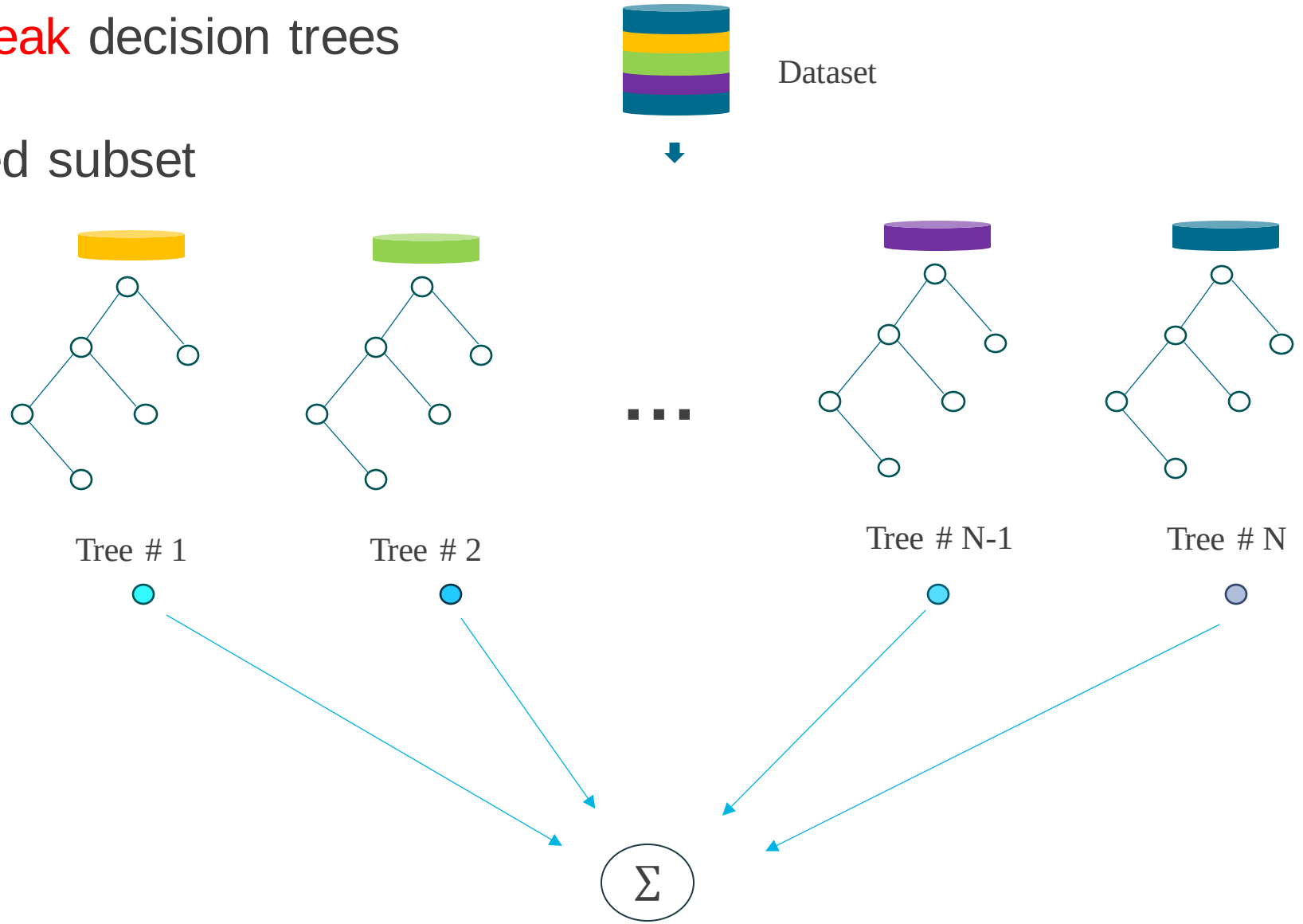
Bagging & Boosting

Model Evaluation

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

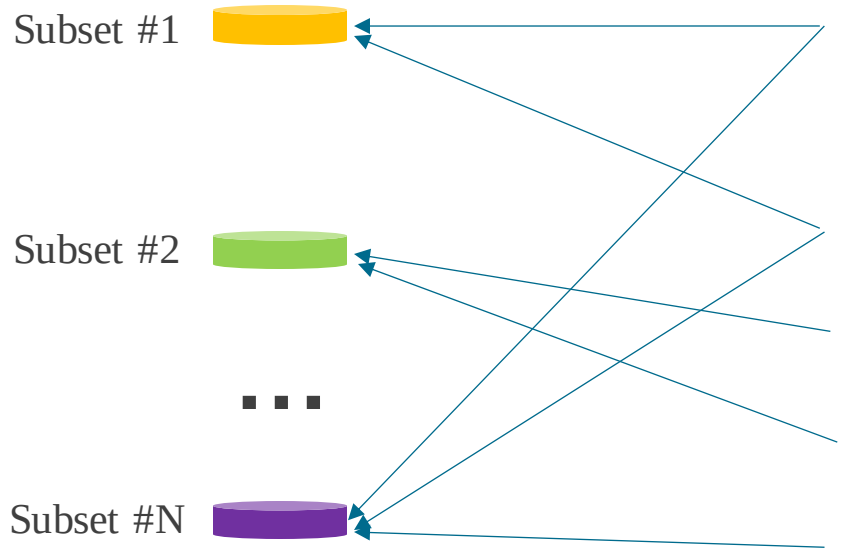
Advanced Tree Algorithms – Random Forest

- Bagging: A set of weak decision trees
  - Randomly sampled subset
  - With replacement
  - One tree per set



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Advanced Tree Algorithms – Random Forest



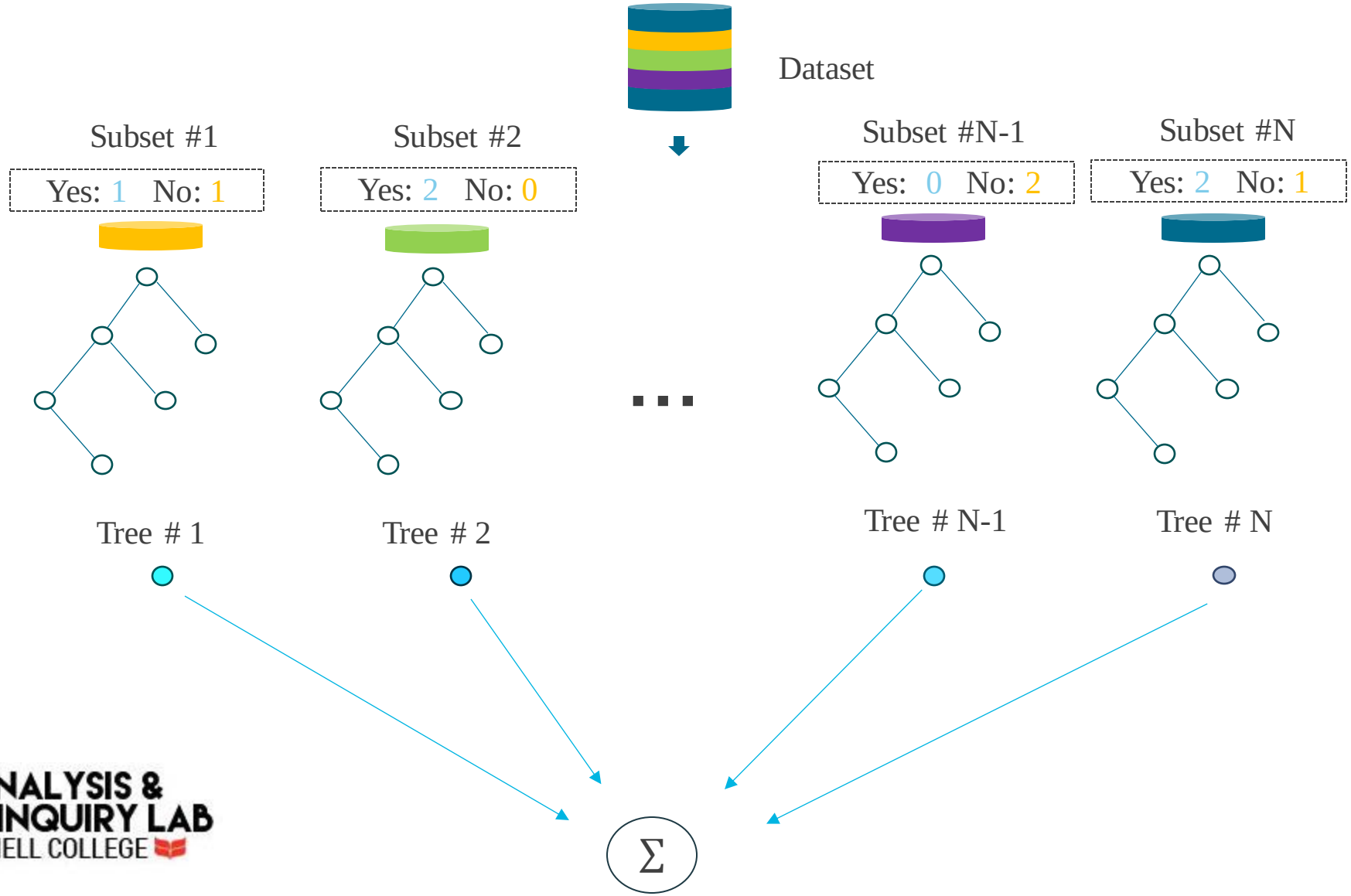
Predictor variable

Target

ID	Country	Gender	Age	Leave
1	Scottland	Male	Old	No
2	Scottland	Male	Old	No
3	England	Male	Young	Yes
4	Wales	Male	Old	Yes
5	Wales	Female	Old	Yes
6	Wales	Female	Young	No

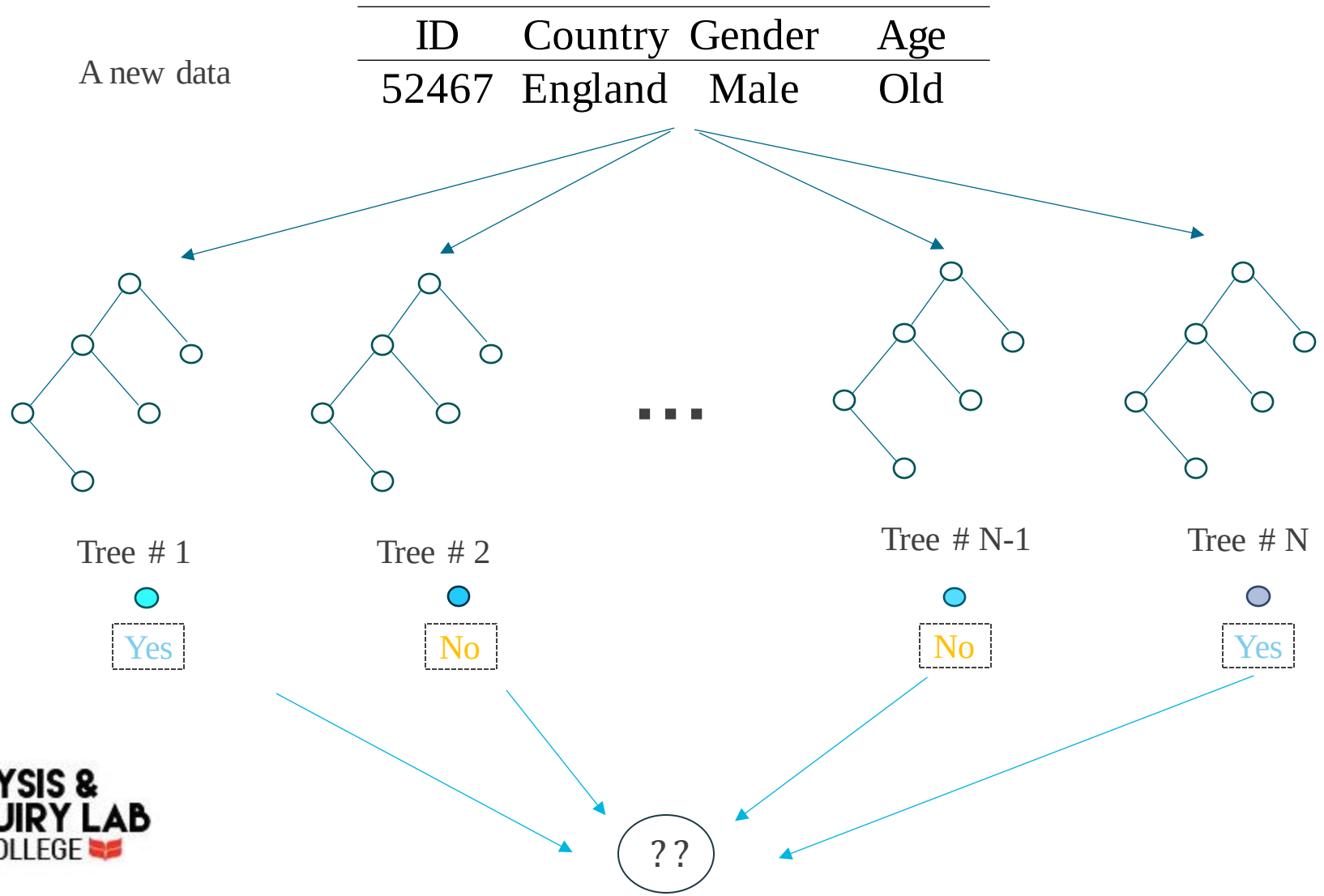
DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Advanced Tree Algorithms – Random Forest



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Advanced Tree Algorithms – Random Forest



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

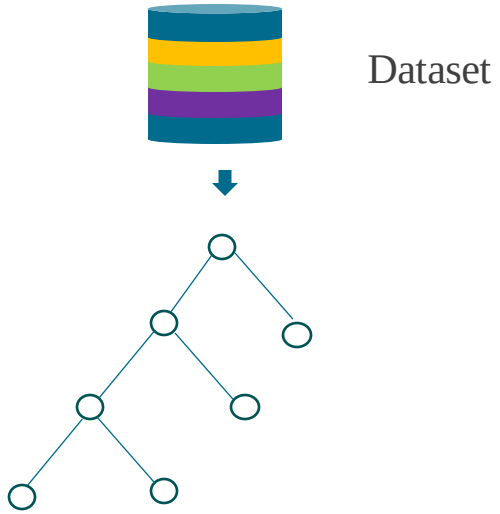
Advanced Tree Algorithms – XGBoosting

- Boosting: Stacked strong decision trees
  - Focus on one dataset
  - Focus more on errors
  - Overfitting

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Advanced Tree Algorithms – XGBoosting

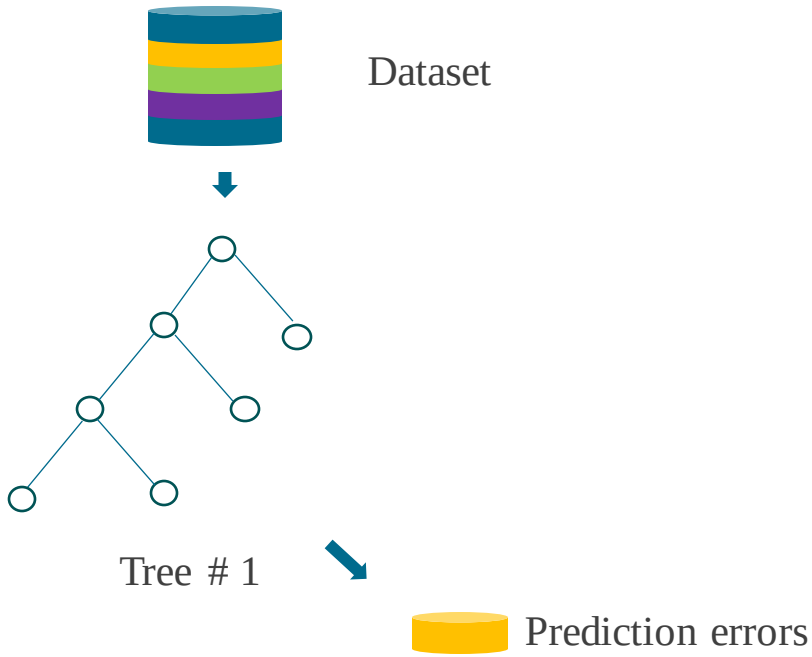
- Step 1: Fit a decision tree



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Advanced Tree Algorithms – XGBoosting

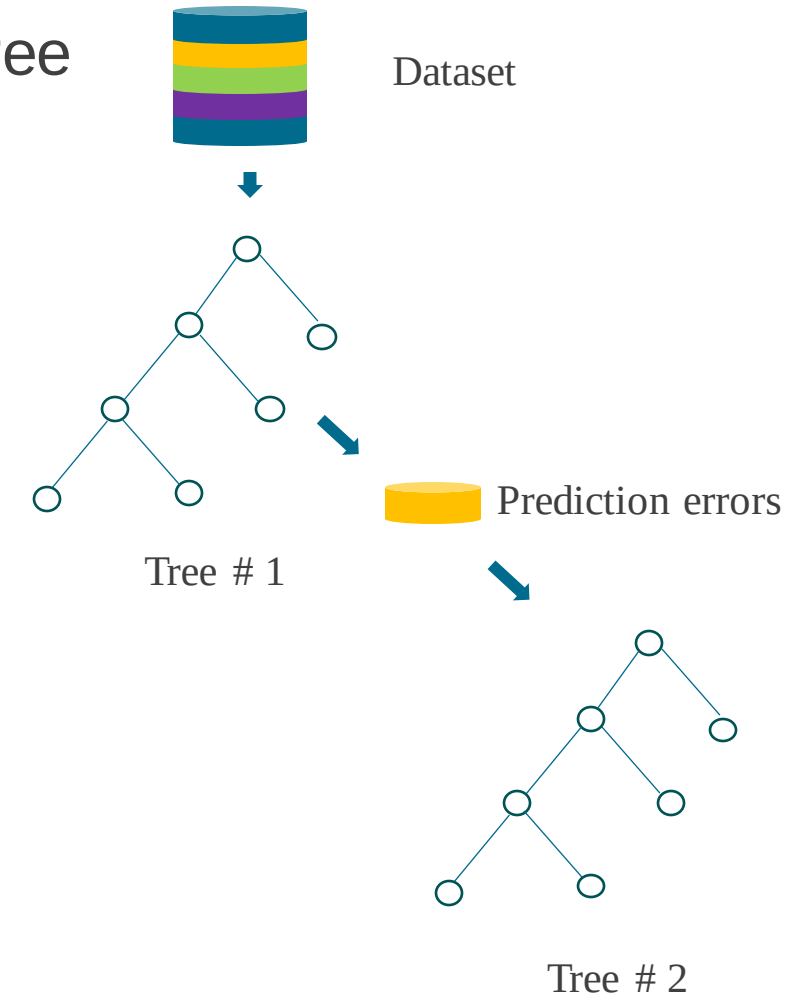
- Step 2: Compute prediction errors



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Advanced Tree Algorithms – XGBoosting

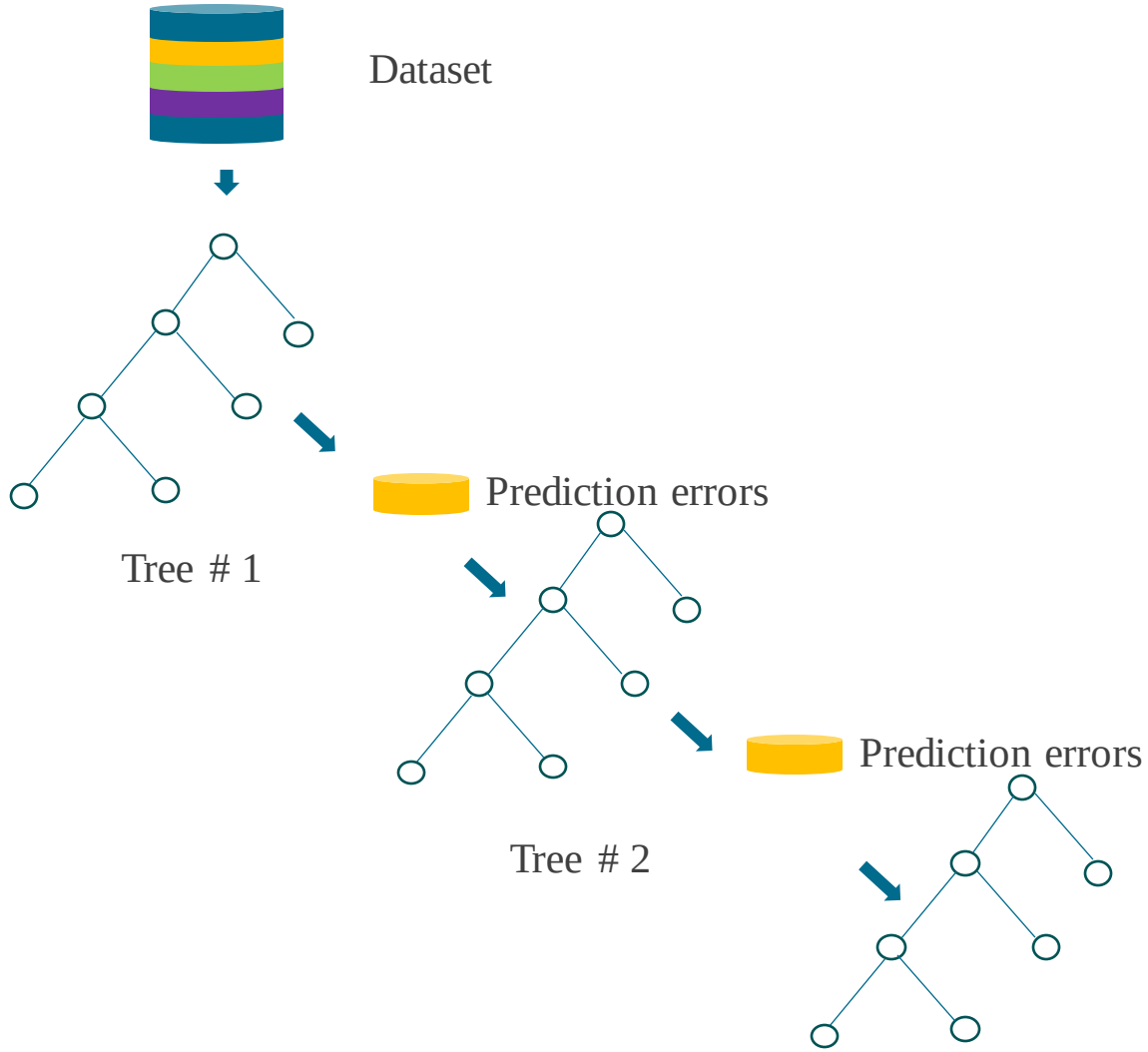
- Step 3: Stacked with another decision tree
  - Focus on error only



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Advanced Tree Algorithms – XGBoosting

- Step N: Stack trees iteratively



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Pop Quiz!

- Talk to your neighbor about the difference between *bagging* and *boosting*.

Which one is more likely to overfitting?

Which one is likely to underfitting?

Which one has higher computational complexity?

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Intro to Machine Learning Part #2

AGENDA

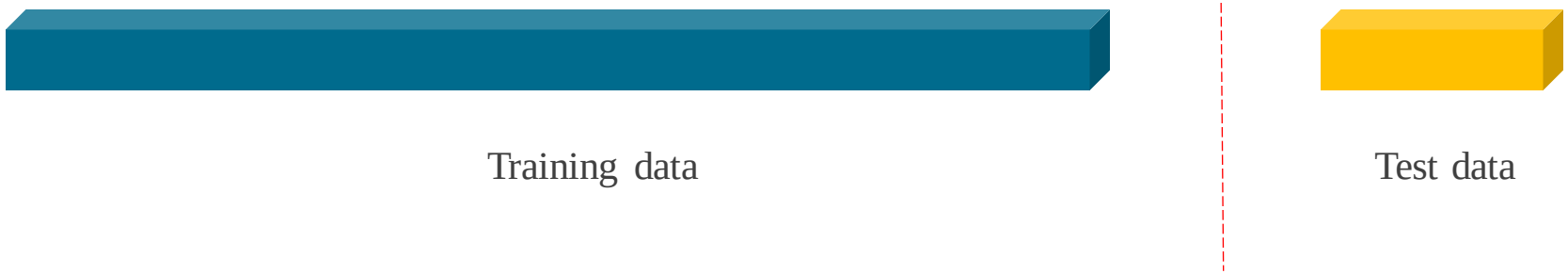
Bagging & Boosting

Model Evaluation

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Training / Testing Dataset Split

- Training dataset versus test dataset



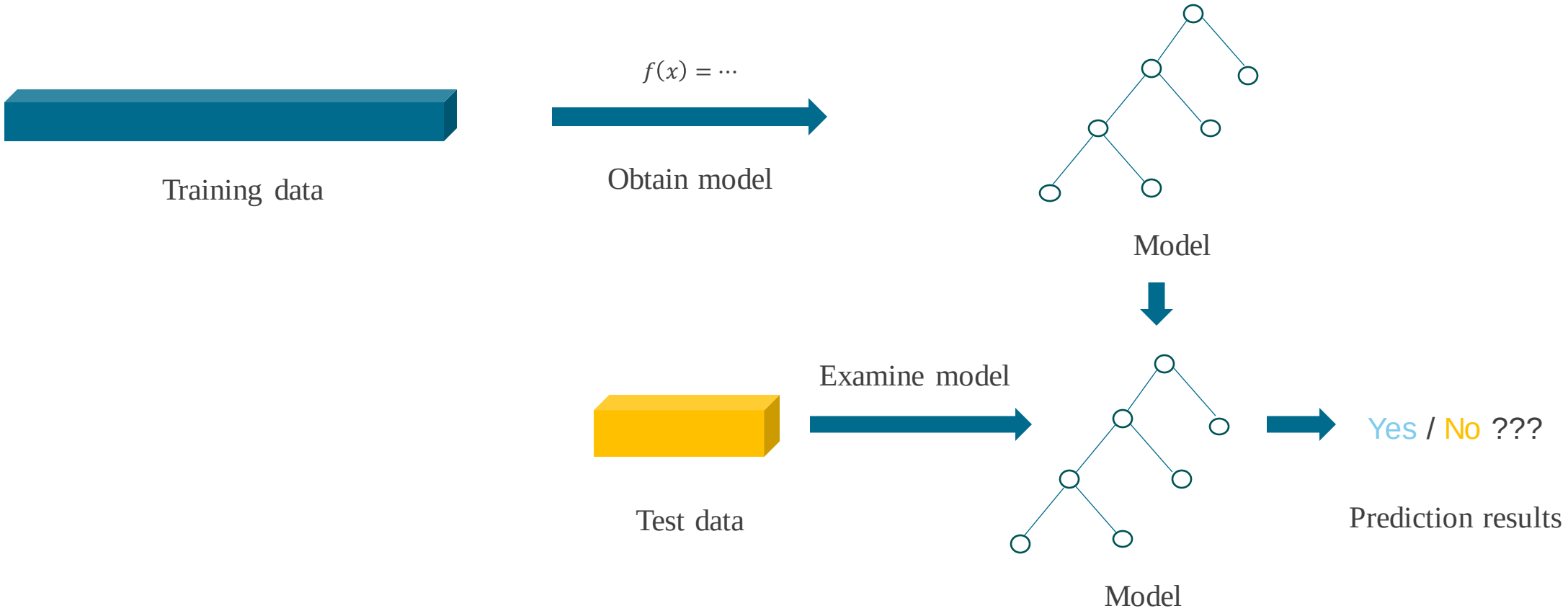
Independent of each other

No overlap

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Training / Testing Dataset Split

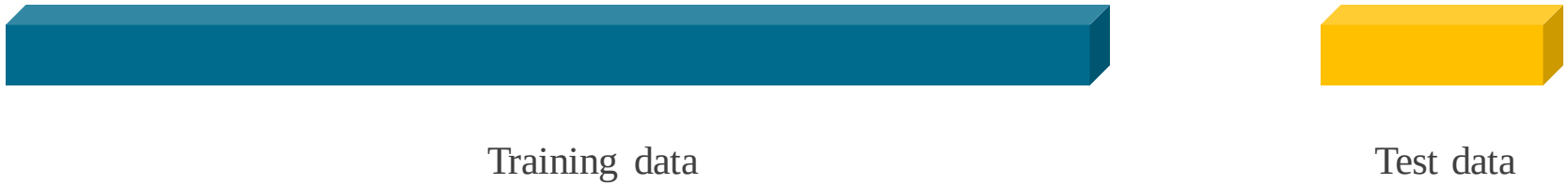
- Training dataset versus test dataset



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Training / Testing Dataset Split

- Single train-test split

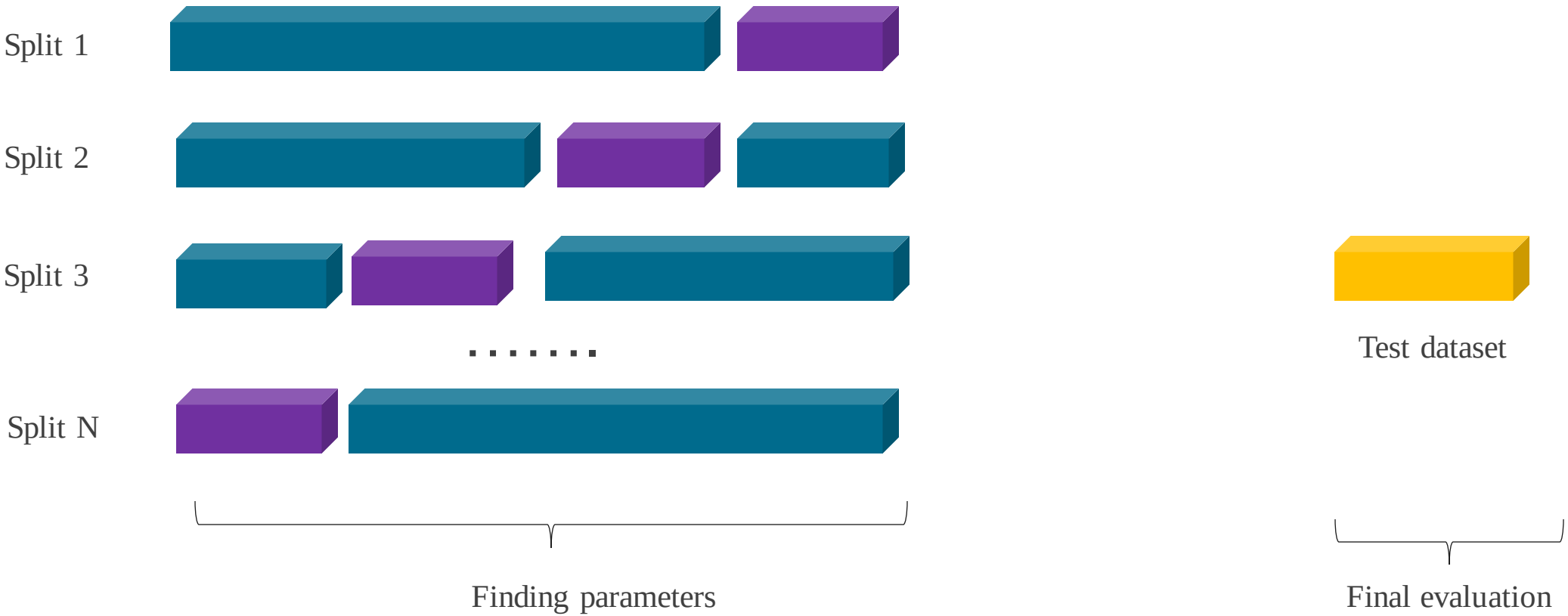


- Single train-test split with validation dataset

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Training / Testing Dataset Split

- Cross-validation



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Measurement Metrices - Classification

- Confusion Matrix:

		actual value: *p*	actual value: *n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
total		P	N	

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Measurement Metrices - Classification

- Confusion Matrix:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Sensitivity = \frac{TP}{\text{Actual positive}} = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{\text{Actual negative}} = \frac{TN}{TN + FP}$$

		actual value: *p*	actual value: *n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Measurement Metrices - Classification

- Discuss:
  - If I am designing a COVID-19 test, is having a higher sensitivity or specificity more important?

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

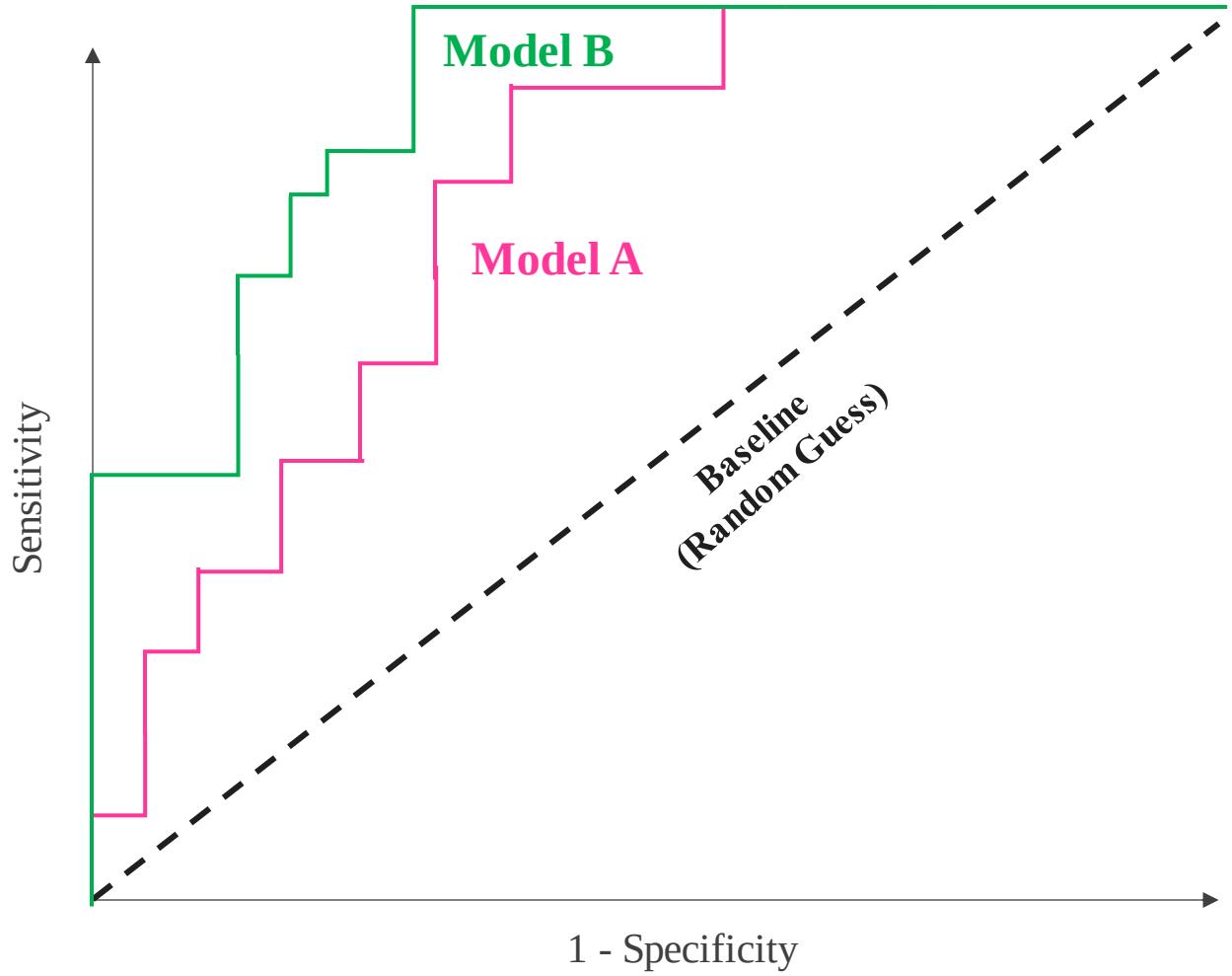
$$Sensitivity = \frac{TP}{\text{Actual positive}} = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{\text{Actual negative}} = \frac{TN}{TN + FP}$$

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Measurement Metrices - Classification

- Area Under the ROC Curve (AUC)



DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Measurement Metrices - Classification

- Area Under the ROC Curve
  - Measures how well the model can separate the groups
  - Shows sensitivity/specificity for each decision boundary between 0 and 1
  - Higher AUC -> better model

DATA ANALYSIS & SOCIAL INQUIRY LAB
GRINNELL COLLEGE

Q & A