Wentworth
INSTITUTE OF TECHNOLOGY

AIR FORUM 2023

New Retention Insights: National Student Student Clearinghouse & Machine Learning Learning

Steven Sherrin, PhD Wentworth Institute of Technology

What’s This Presentation About?

- I will introduce a machine learning tool that allows institutions to assess retention of students of varying demographic, financial, and academic qualities - comparing results to peer institutions.

- I will show how the tool can be used to assess or benchmark retention for any group of students*.

- I will demonstrate how to use data mining techniques to discover new retention insights.

What Does The Tool Do?

Data Prep

- Clean and analyze National Student Clearinghouse enrollment data
- Integrate NSC data with College Scorecard data

Machine Learning

- Prepare and run machine learning models
- Select the best machine learning models
- Ensure models are fair and accurate

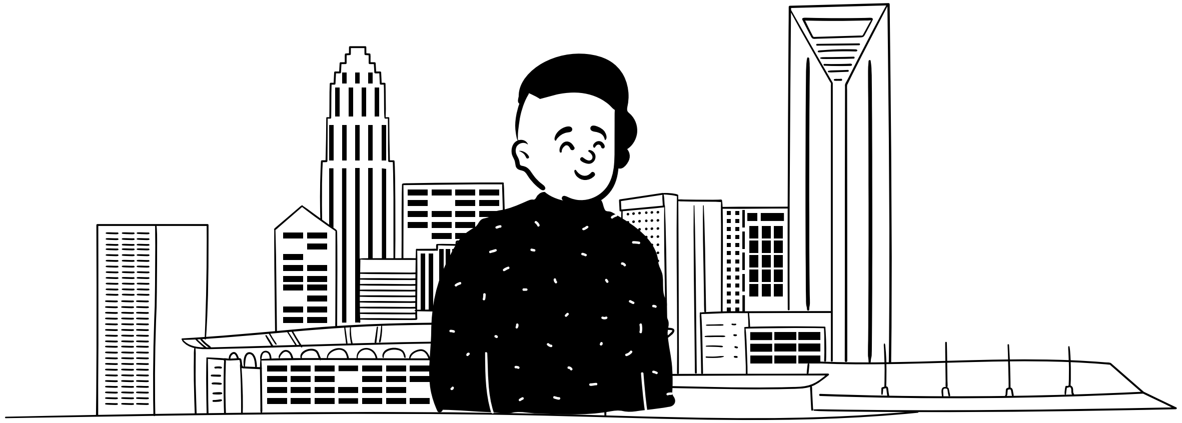
Data Insights

- Use machine learning models to compare retention rates by institution and demographic groups
- Share data mining tools to discover new retention insights

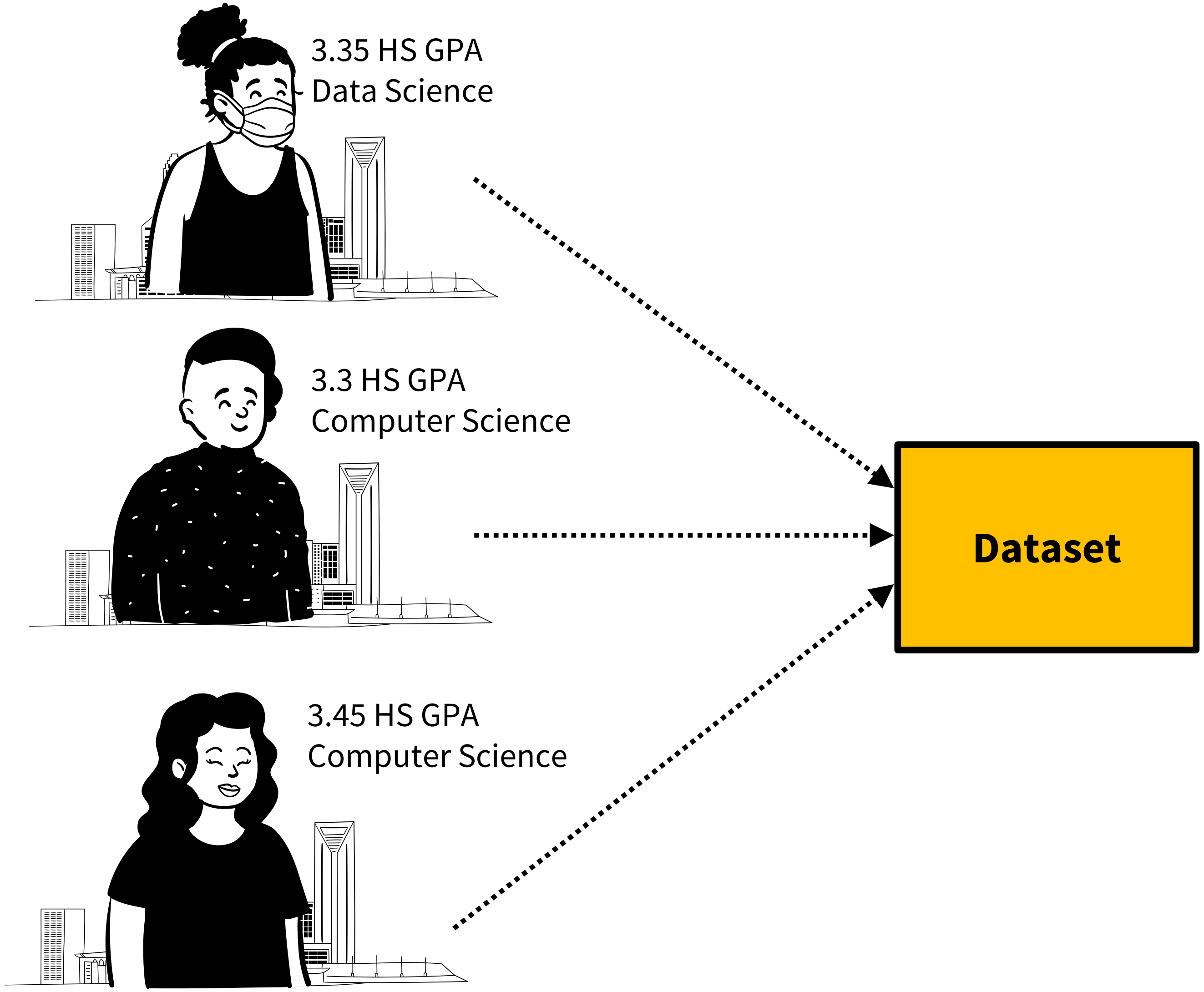
Conceptual Approach

Example

Female STEM Students
At AIR University



AIR University



3.35 HS GPA
Data Science

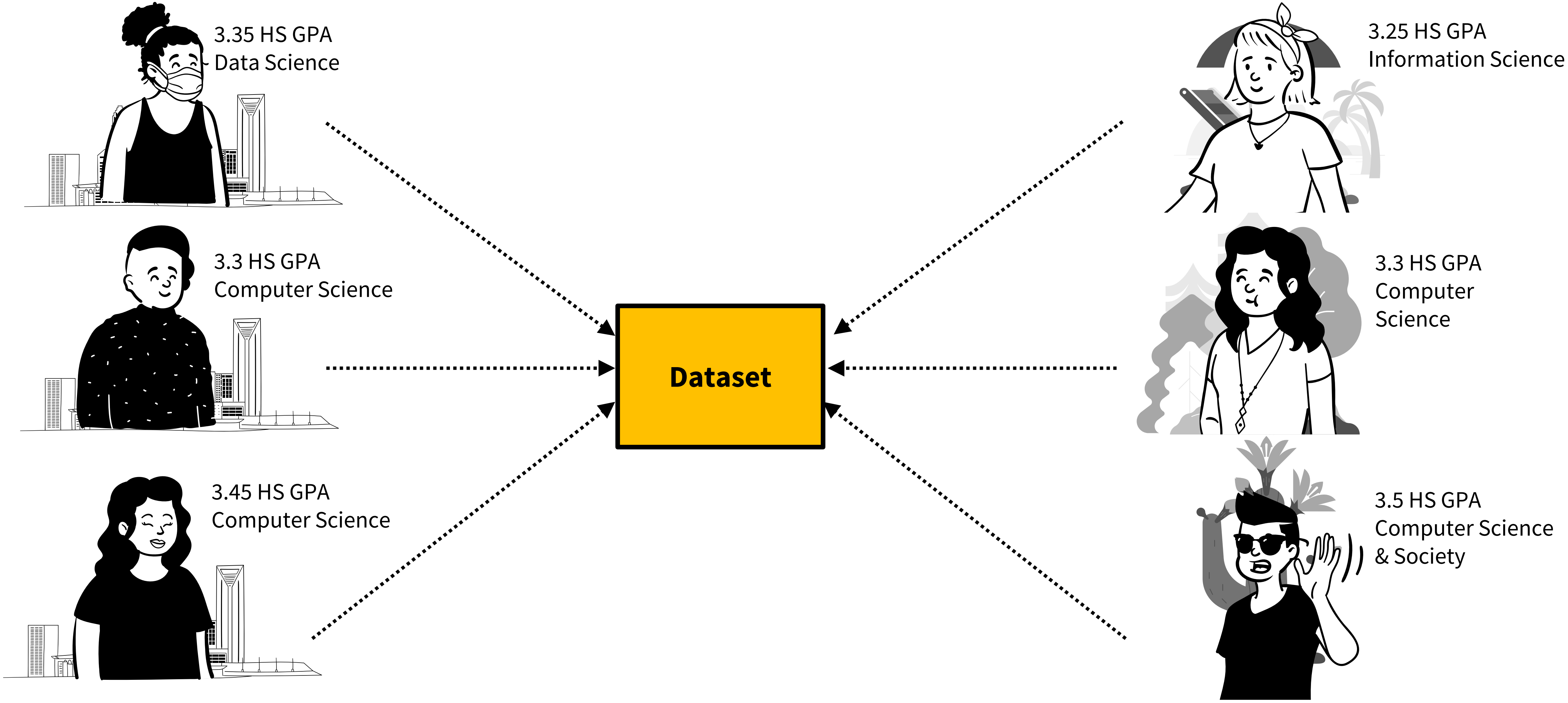
3.3 HS GPA
Computer Science

3.45 HS GPA
Computer Science

Dataset

AIR University

Other Institutions



3.35 HS GPA
Data Science

3.3 HS GPA
Computer Science

3.45 HS GPA
Computer Science

Dataset

3.25 HS GPA
Information Science

3.3 HS GPA
Computer
Science

3.5 HS GPA
Computer Science
& Society

Wentworth

Data & Methods

Data

Admissions

E.g. HS GPA, geographic, socioeconomic data

Enrollment

Which institution student(s) chose to attend

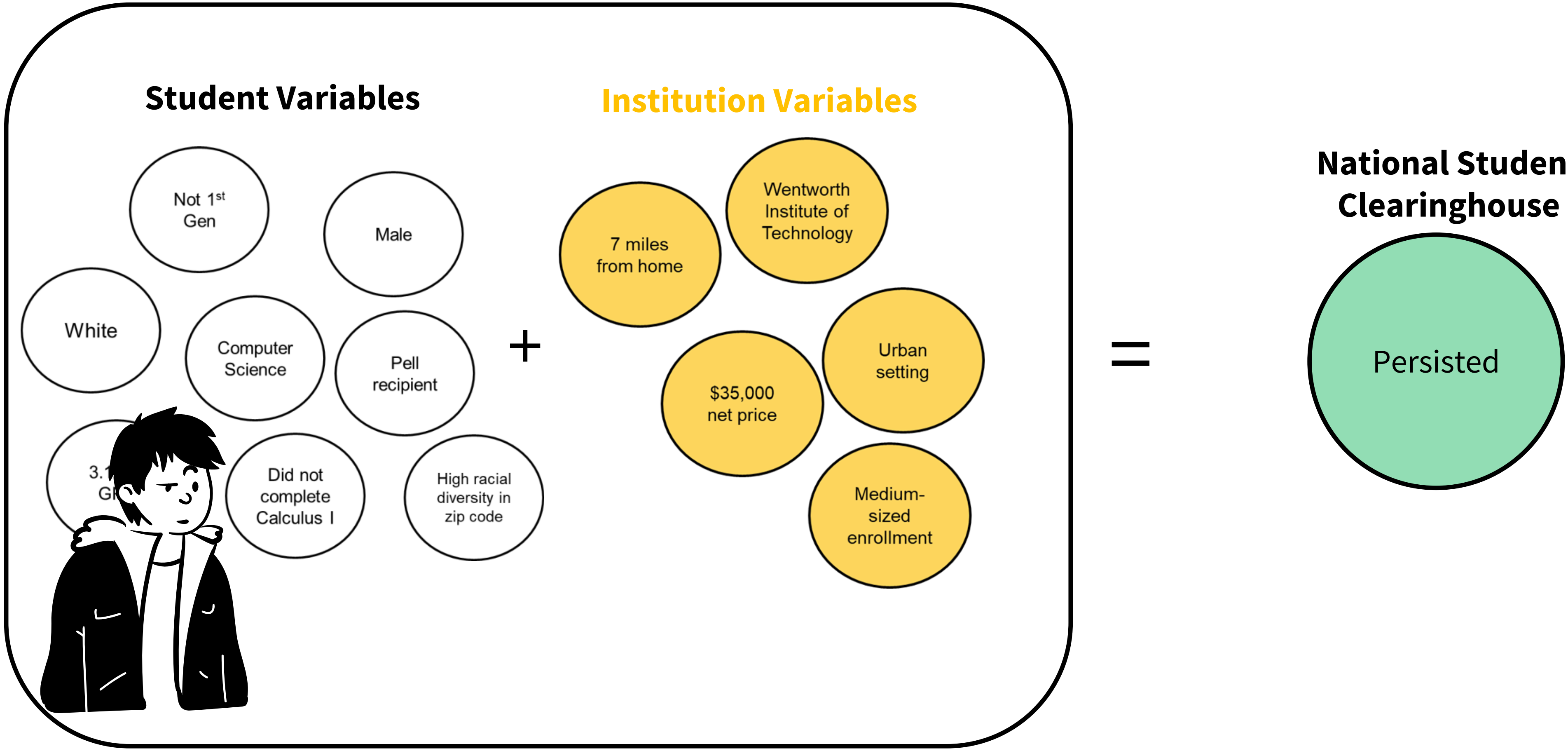
Enrollment Outcomes

E.g. Fall-to-fall retention

SCAN ME

Link to GitHub

Integrating Data Sources



Student Variables

Not 1st Gen

Male

White

Computer Science

Pell recipient

3.1 GP

Did not complete Calculus I

High racial diversity in zip code

+

Institution Variables

7 miles from home

Wentworth Institute of Technology

Urban setting

$35,000 net price

Medium-sized enrollment

=

National Student Clearinghouse

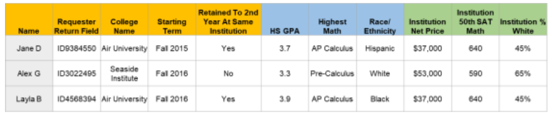
Persisted

SCAN ME

Link to GitHub

Analytic Plan

Data



Name	Requester Return Field	College Name	Starting Term	Retained To 2nd Year At Same Institution	HS GPA	Highest Math	Race/ Ethnicity	Institution Net Price	Institution 50th SAT Math	Institution % White
Jane D	ID9384550	Air University	Fall 2015	Yes	3.7	AP Calculus	Hispanic	$37,000	640	45%
Alex G	ID3022495	Seaside Institute	Fall 2016	No	3.3	Pre-Calculus	White	$53,000	590	65%
Layla B	ID4568394	Air University	Fall 2016	Yes	3.9	AP Calculus	Black	$37,000	640	45%

- Simple (10 predictors)
- Complex (50+ predictors)

Algorithms

- Generalized additive mixed mixed models (GAMM)
- Gradient boosting (GBM) with cross-validation
  - 100s of models tested tested via hyperparameter tuning. tuning.
- Ensemble stacking of models models

Predictions

- Test model performance performance
- Examine model bias
- Ensure model fairness
- Predictions

SCAN ME

Link to GitHub

Wentworth

Results

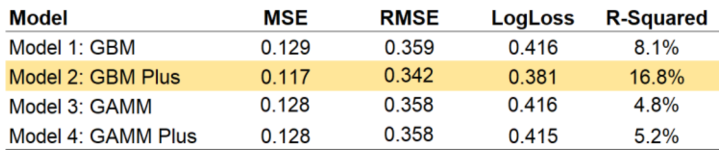
Wentworth

Model Validation

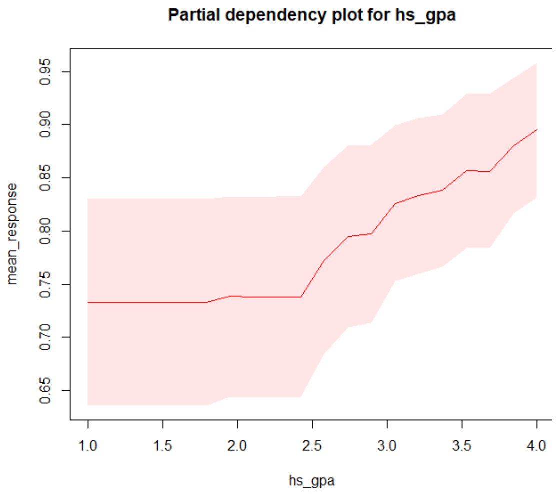
1 Compare Models

2 Understand Models

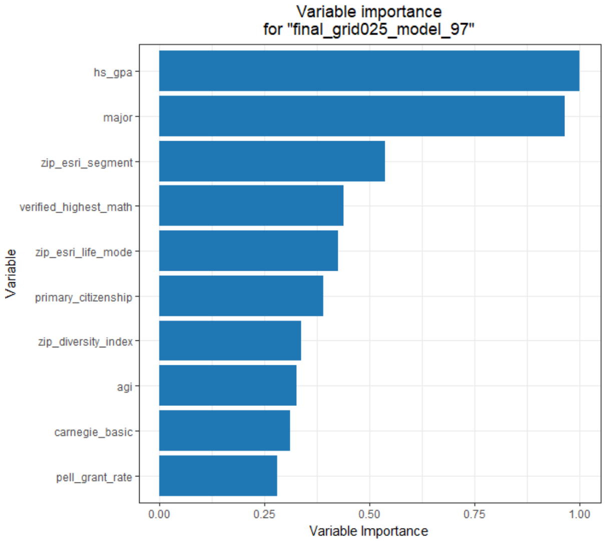
3 Examine Model Bias



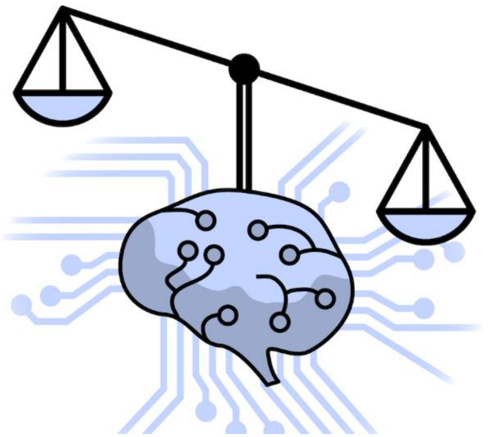
Model	MSE	RMSE	LogLoss	R-Squared
Model 1: GBM	0.129	0.359	0.416	8.1%
Model 2: GBM Plus	0.117	0.342	0.381	16.8%
Model 3: GAMM	0.128	0.358	0.416	4.8%
Model 4: GAMM Plus	0.128	0.358	0.415	5.2%



Partial dependency plot for hs_gpa



Variable importance
for "final_grid025_model_97"



Wentworth

Model Results

Example results

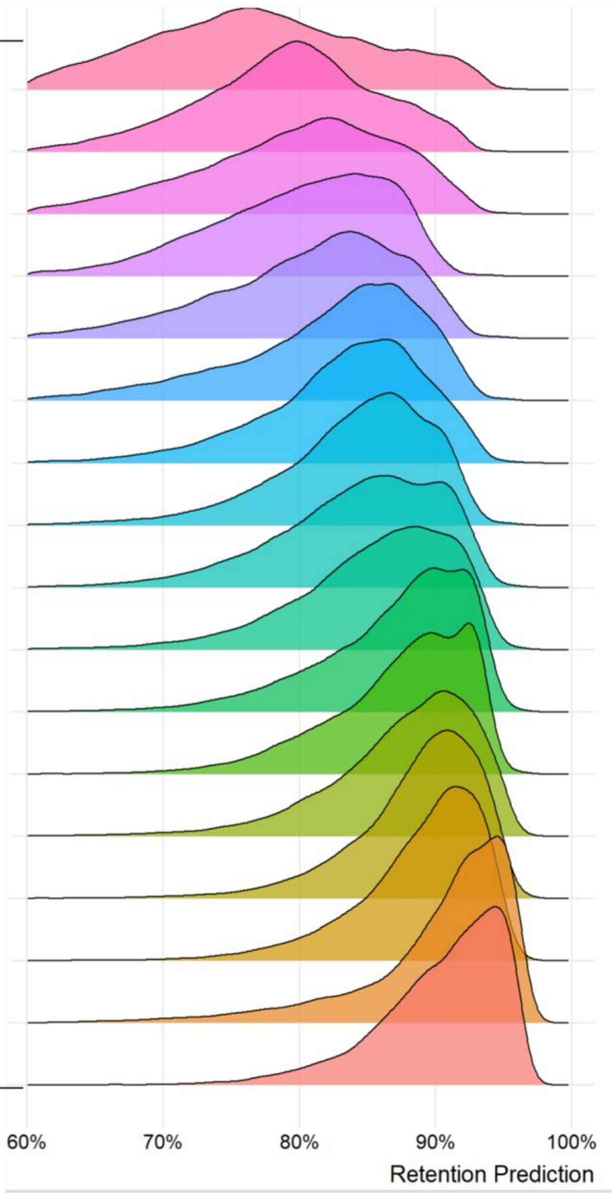
Question #1

“Overall, how are we doing?”

Wentworth

Overall

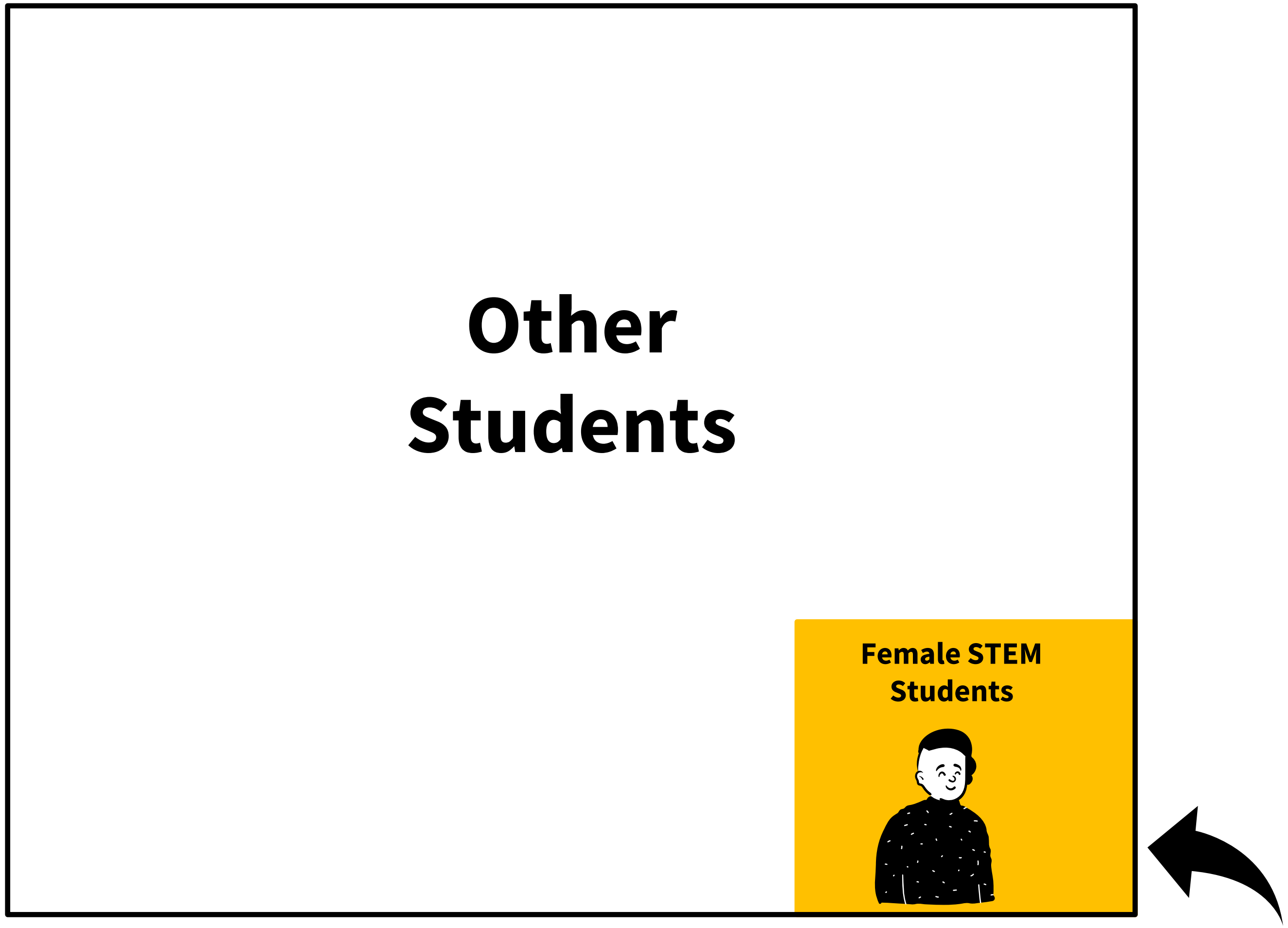
Institution	Predicted Retention Rate	Difference from Average Institution
Institution #1	74.7%	-8.6%
Institution #2	77.8%	-5.5%
Institution #3	78.1%	-5.2%
Institution #4	78.6%	-4.7%
Institution #5	79.2%	-4.0%
Institution #6	81.0%	-2.3%
Institution #7	82.2%	-1.1%
Institution #8	83.0%	-0.3%
Institution #9	83.2%	-0.1%
Institution #10	84.4%	1.2%
Institution #11	86.0%	2.7%
Institution #12	86.3%	3.0%
Institution #13	86.8%	3.5%
Institution #14	88.0%	4.7%
Institution #15	88.0%	4.7%
Institution #16	88.9%	5.6%
Institution #17	89.5%	6.2%



Question #2

“How are we doing in retaining [student group]?”

Female STEM Students



Other Students

Female STEM Students

Analyze predictions for only these students.

Female STEM Students

versus

Male STEM Students

Or, compare predictions between two student groups.

Female STEM Students

versus

Male STEM Students

Female STEM Students

versus

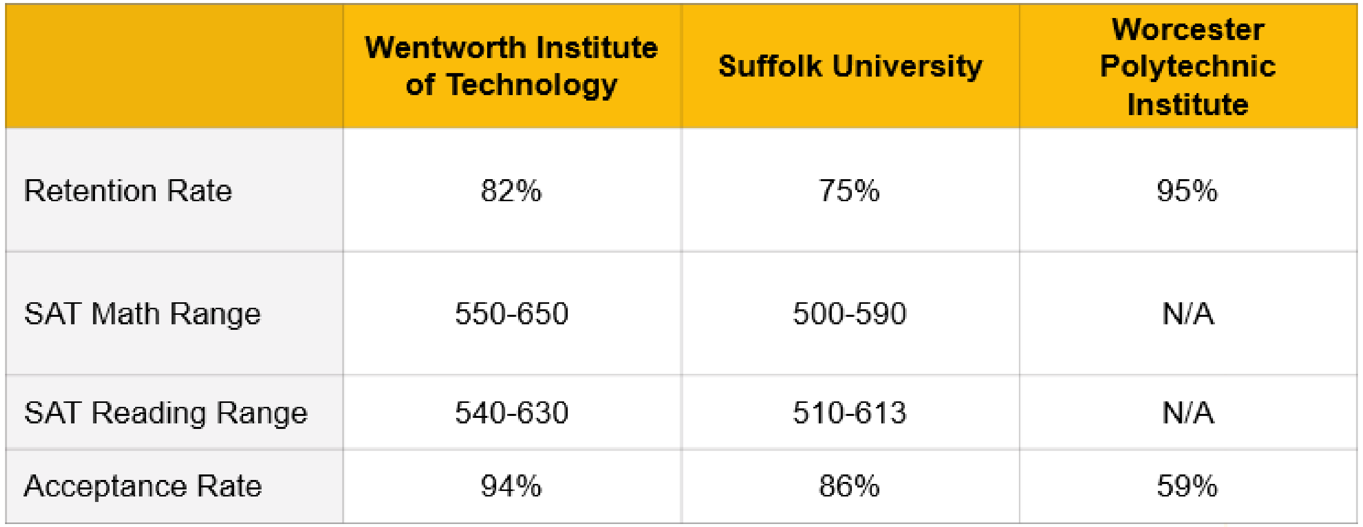
Male STEM Students

Or, compare predictions between two student groups at **two different institutions**.

Question #3

“How do we compare our results to schools with higher (or lower) academic selectivity?”

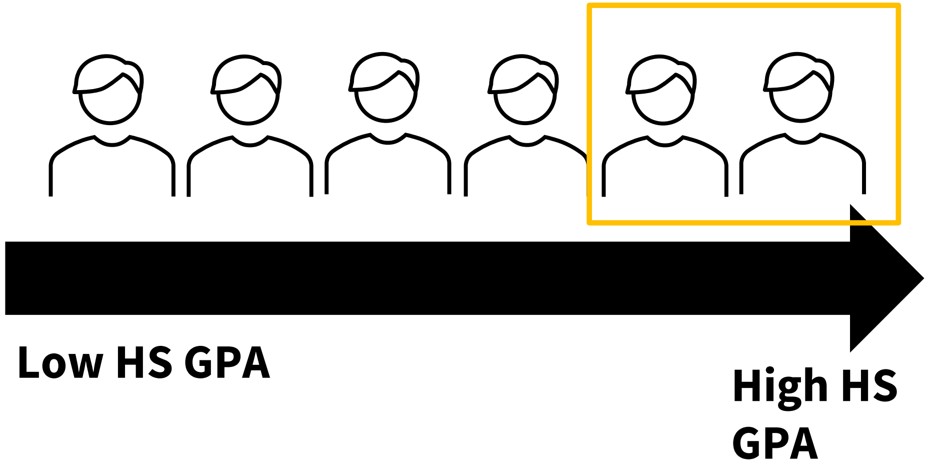
Comparing Institutions by Student Group



	Wentworth Institute of Technology	Suffolk University	Worcester Polytechnic Institute
Retention Rate	82%	75%	95%
SAT Math Range	550-650	500-590	N/A
SAT Reading Range	540-630	510-613	N/A
Acceptance Rate	94%	86%	59%

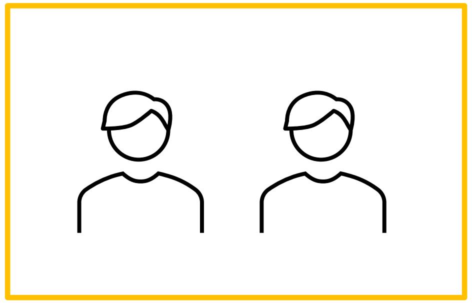
You can’t use this. It’s comparing apples to oranges!

Selecting Comparisons



Select students who share attributes in common with students from the institution(s) you want to compare to.

Example



“Our model predicts that among students with high GPAs (3.5 or above), our institution [over- or under-] performs in retention, compared to WPI and other “elite” competitor schools, by ____%.”

Question 4

“Did your super-fancy model tell us anything we don’t know/discover anything new?”

Discovering New Insights

Use **data mining** techniques to discover new insights.

Example:

- Surrogate modeling of machine learning model.
- A decision tree that predicts the **biggest predicted retention differences** at your institution compared to other institutions.

Summary

Summary

- This tool enables institutions to assess retention of students of varying demographic, financial, and academic qualities.
- Can be used to assess retention for *any* group of students (*with sufficient sample size).
- Exploring model can uncover previously unknown retention insights.

Wentworth
INSTITUTE OF TECHNOLOGY

Thank You!

Steven Sherrin, Ph.D.
sherrins@wit.edu

Thank You!

Statistics,
Methods, & Code

Do You Want To
Collaborate?

Connect With Me