Eric & Wendy Schmidt

Data Science For Social Good

Summer Fellowship

THE UNIVERSITY OF CHICAGO

Identifying and Influencing Students at Risk of Not Finishing High School

Muskingum Valley Educational Service Center


Zhe Zhang


Johanna Torrence


Jacqueline Gutman


Xiang Cheng


Ali Vanderveld


Kevin Wilson


Chad Kenney

Improving Student Success & Reducing Dropouts

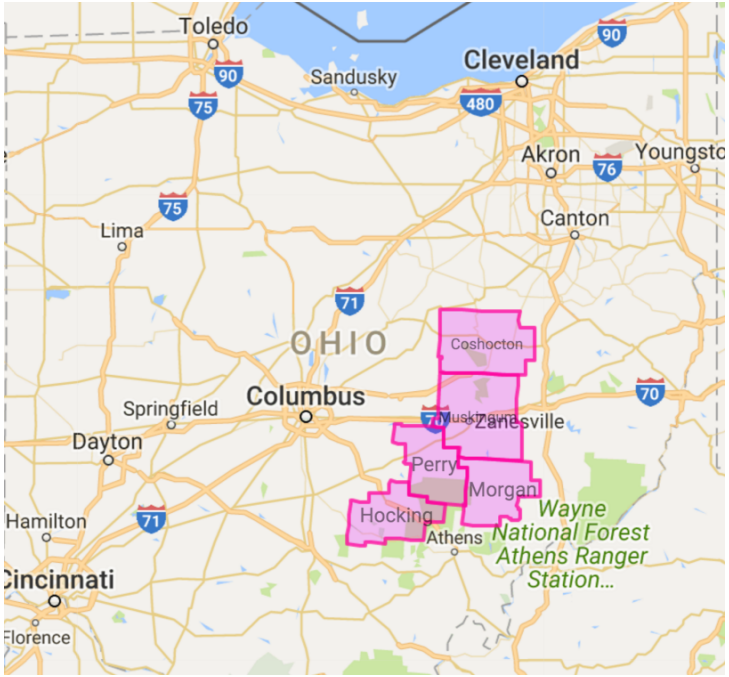
1.2 Million - annual dropouts in U.S.

30% - rate of unemployment

34% - lower annual income than non-dropouts

75% - of Ohio inmate population

Partnering with MVESC



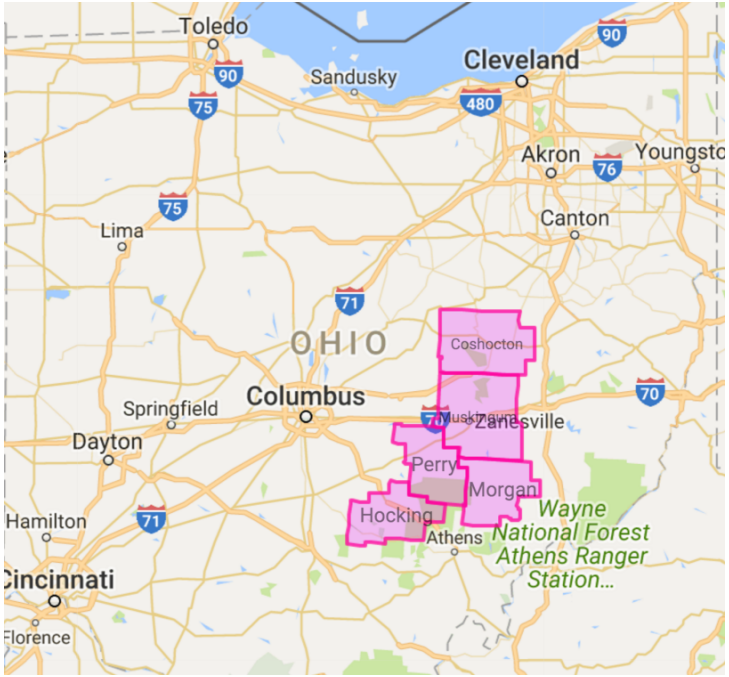
Data Science For Social Good

MuskingumValley Educational Service Center

16 School Districts

30,897 Students

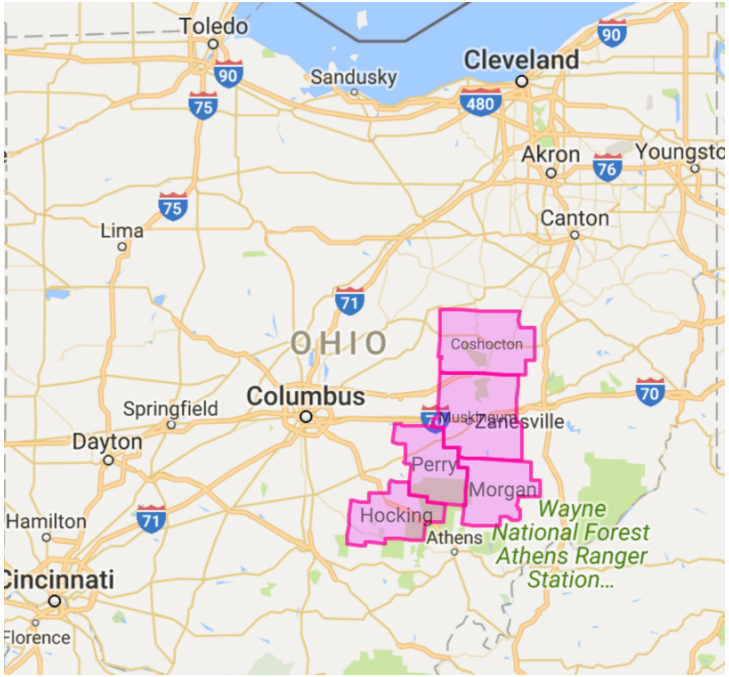
Partnering with MVESC



MuskingumValley Educational Service Center

16 School Districts

Partnering with MVESC



MuskingumValley Educational Service Center

16 School Districts

Deliverables

Analysis

Understanding from MVESC data resources

Key Features

Distributions

Correlations

Prediction

Validation & Action

Key	Risk		Top 3 Risk Factors		
student	score	level	factor1	factor2	factor3
23214	9.87	High	GPA	Disability	Absence
00621	8.62	High	Age	Absence	Mobility
18762	8.59	Med	Disability	GPA	Absence

Pipeline Code

Reproducibility & Improvement

- Source Codes
- Generated Data
- Documentations

Data from MVESC

Students

- Demographics
- Grades
- Test Scores
- Attendance
- Graduation Date
- Withdrawal
- Mobilities
- Intervention
- IEP

Teachers

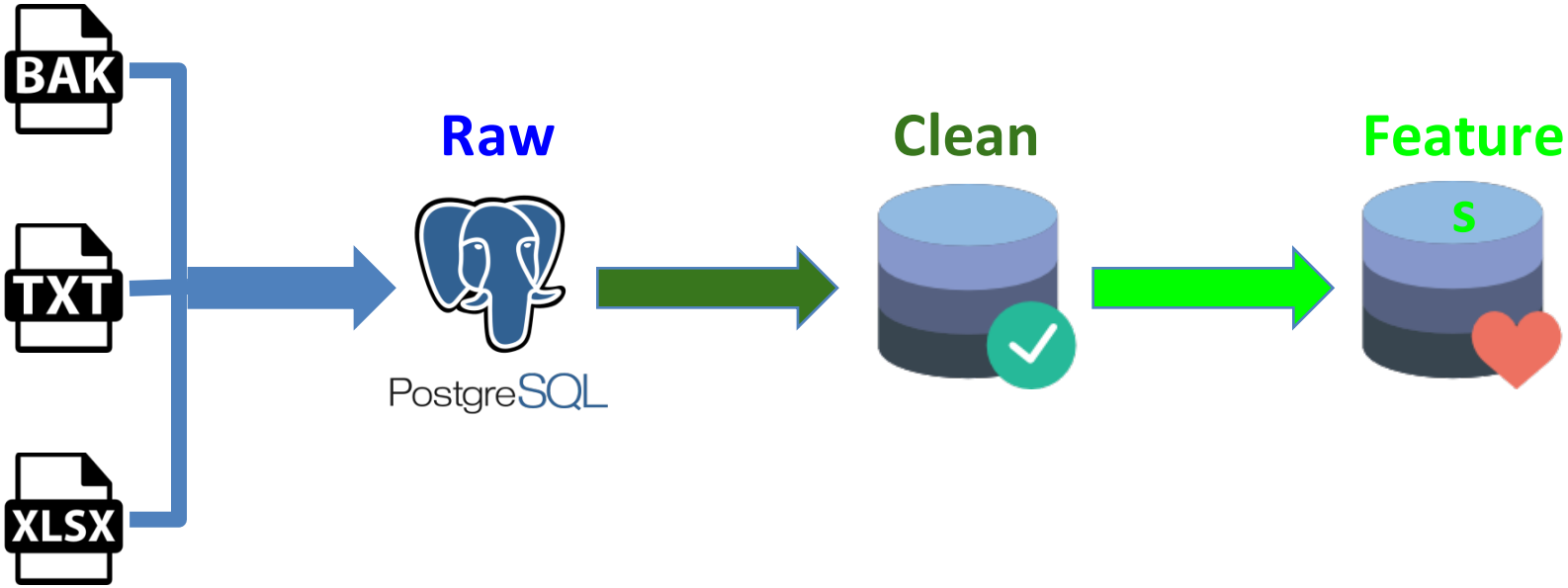
- Courses
- Students
- School Term

Schools & Districts

- District Rating
- IRN
- Mobility Rate
- Graduation Rate

Icons designed by Freepik and distributed by Flaticon

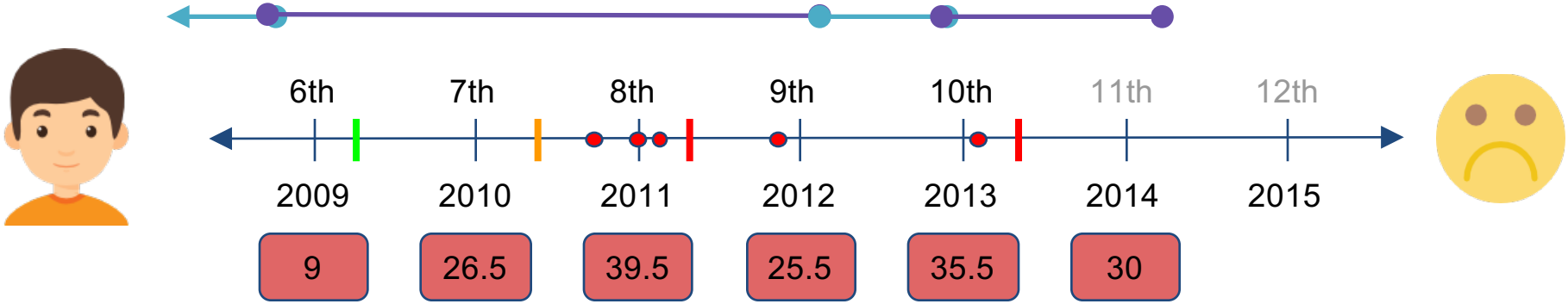
Project Pipeline


BAK
TXT
XLSX
Raw
PostgreSQL
Clean
Feature
S

ETL **Stories** Outcomes Features Validation Results

- Moved in 6th, 9th, and 10th grade
- Free/reduced lunch
- Learning disability
- GPA 1.3

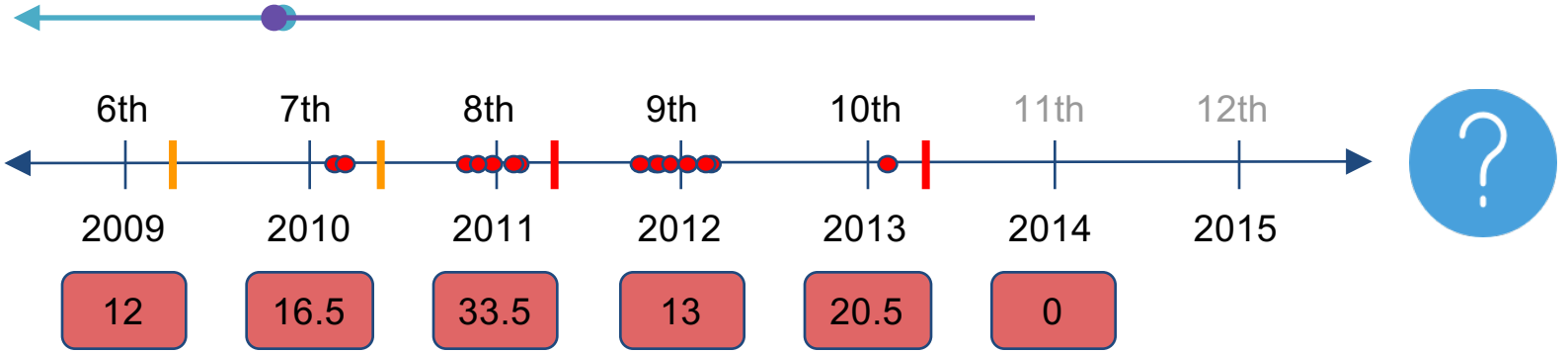
- Tests performance declined
- 166 absences
- Received OSS 5 times
- Dropped out in 11th Grade



ETL **Stories** Outcomes Features Validation Results

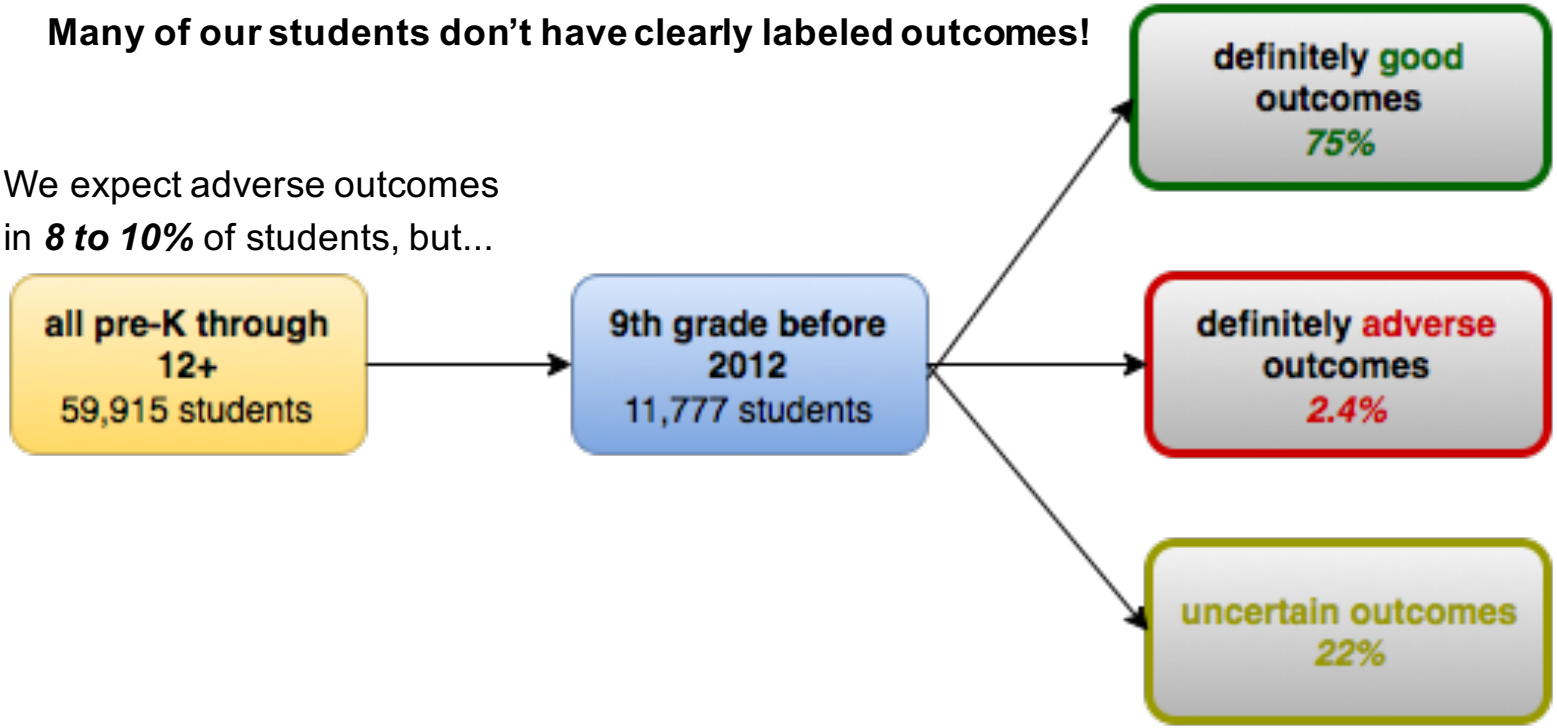
- Moved in 7th grade
- Free/reduced lunch
- No disability and not gifted
- No grades reported

- Poor Test Performance
- 95.5 absences
- 14 disciplinary incidents (2 ISS, 2 OSS)
- Transferred out of state



Many of our students don't have clearly labeled outcomes!

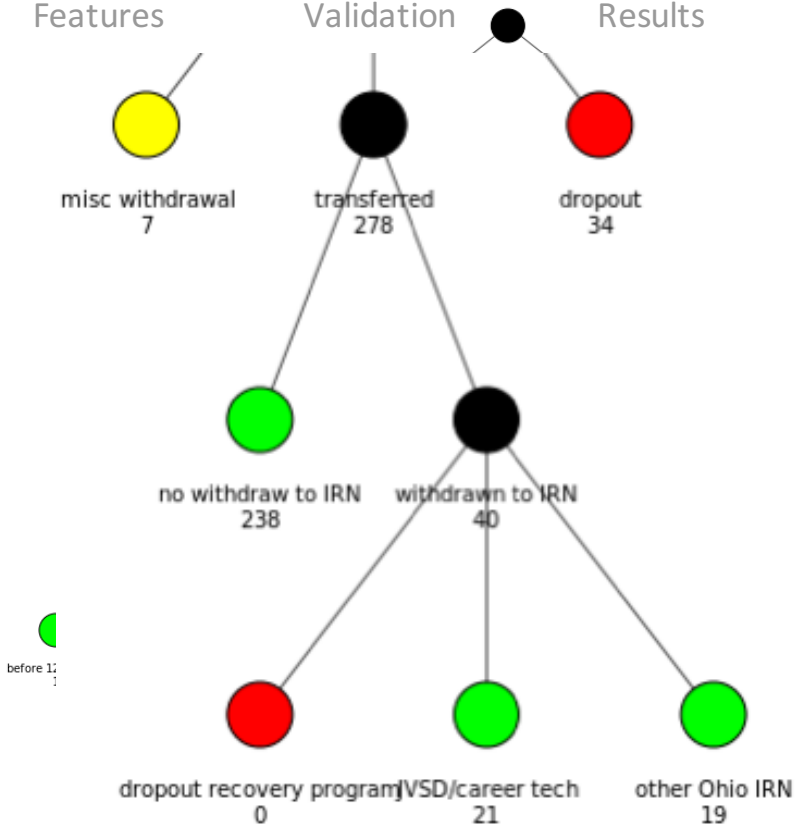
We expect adverse outcomes in ***8 to 10%*** of students, but...



Approach 1
Provide fine-grained labels and experiment with various label-to-outcome mappings in the pipeline

- Revise labels and outcome mappings in iterative exploratory process

Approach 2
Use a semi-supervised approach treating uncertain outcomes as unlabeled data



ETL | Stories | Outcomes | **Features** | Validation | Results

Demographics

- Gender
- Ethnicity
- Age

Snapshots

- Disability
- Discipline
- Gifted
- Disadvantaged

Mobility

- By Address
- By City
- By District

GPA & Scores

- GPA
- Scores
- # Passed
- # Failed

Attendance

- Absence
- Tardy
- Unexcused
- Consecutive

Intervention

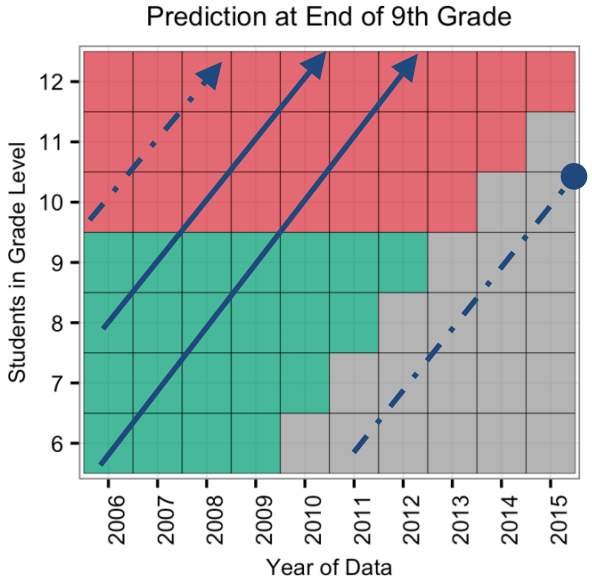
- Academic
- Special Ed
- Sports/Clubs
- Placement
- Therapies

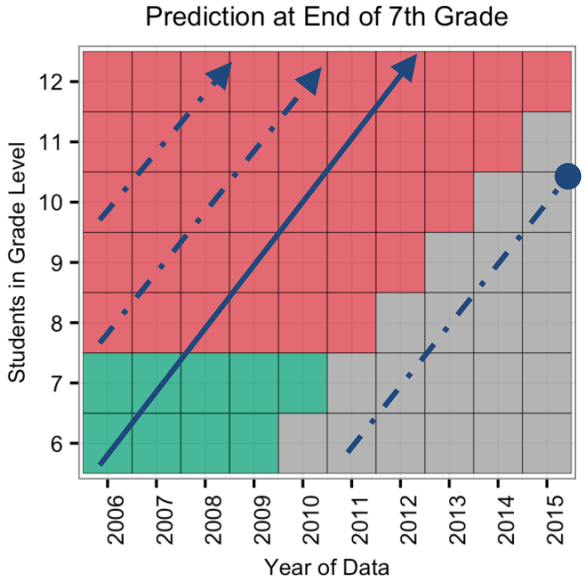
Choice 1: Time of Prediction

Not used; Occurs after prediction

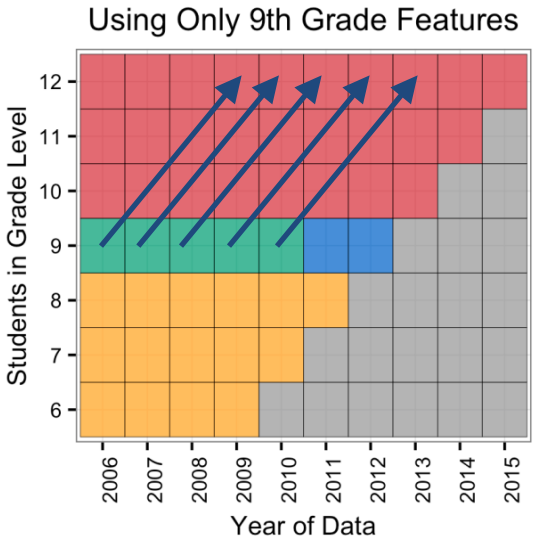
Use for machine learning

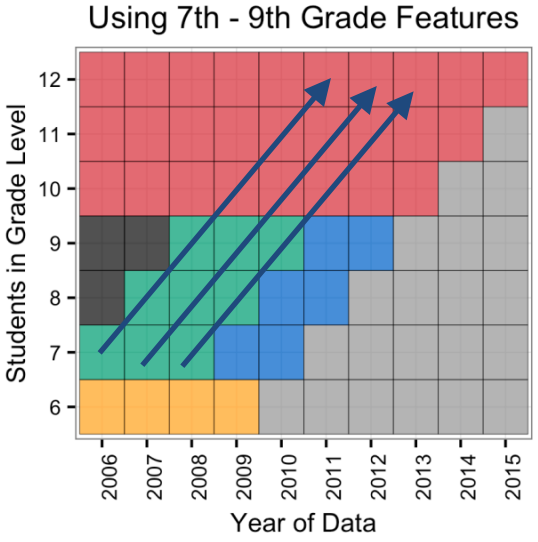
No outcome labels

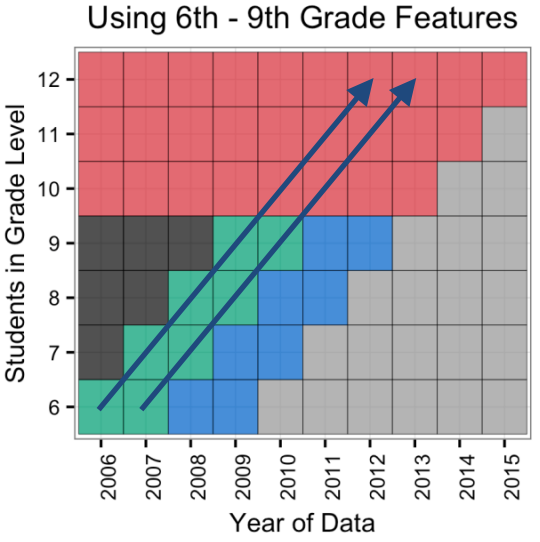




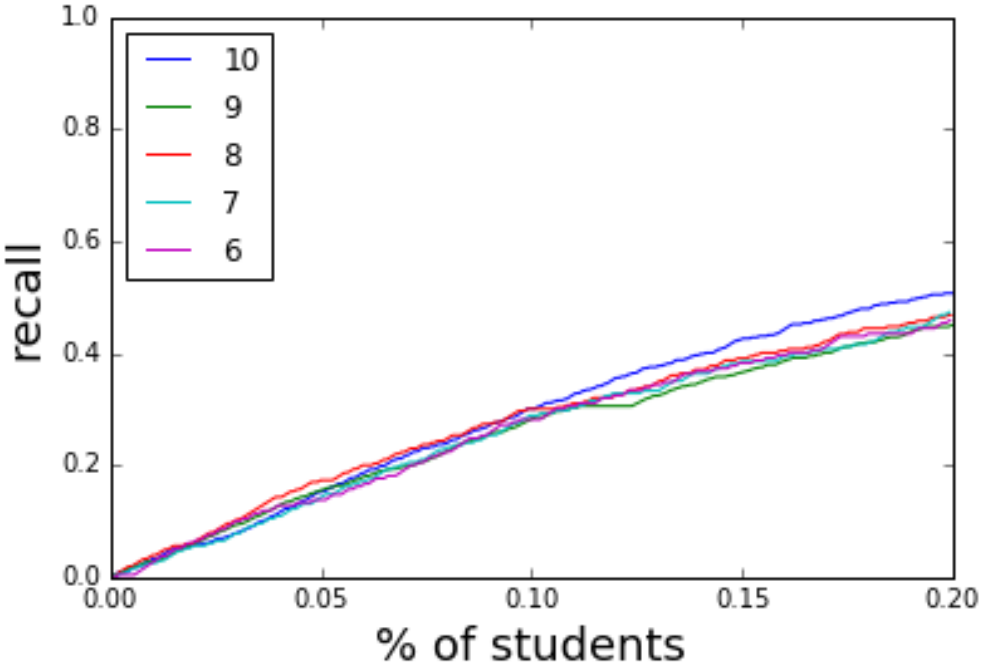
Choice 2: Grade Range of Features





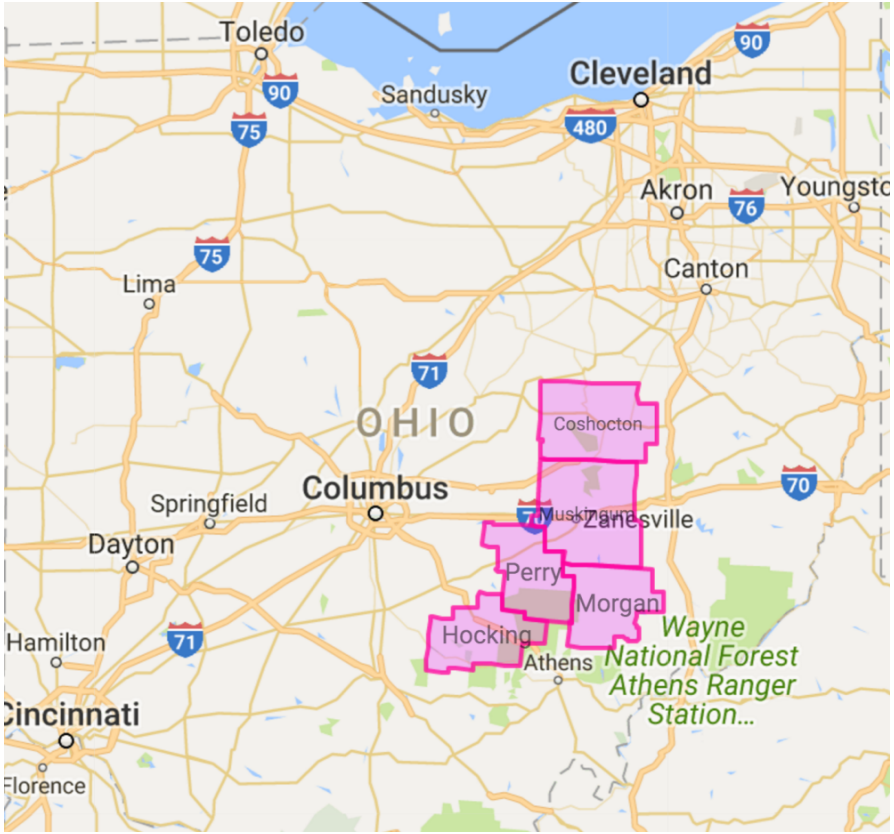


Random Forest Preliminary Model



Questions?

Key	Risk		Top 3 Risk Factors		
student	score	level	factor1	factor2	factor3
23214	9.87	High	GPA	Disability	Absence
00621	8.62	High	Age	Absence	Mobility
18762	8.59	Med	Disability	GPA	Absence

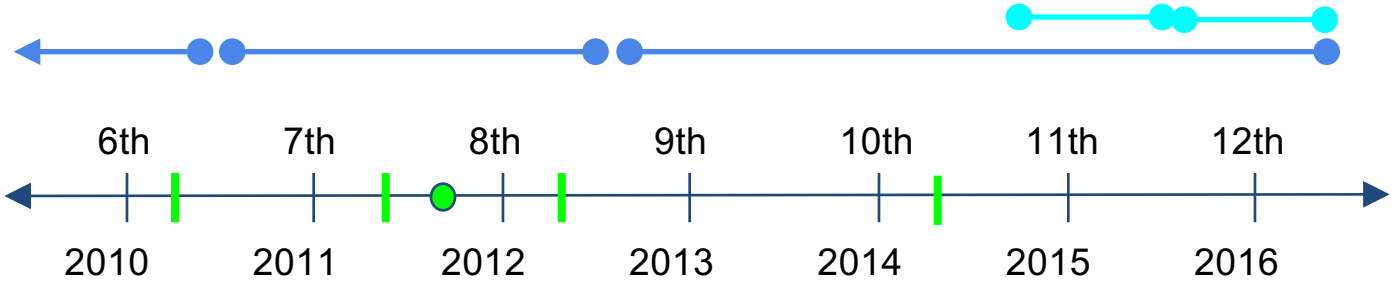


Icons designed by Freepik and distributed by Flaticon

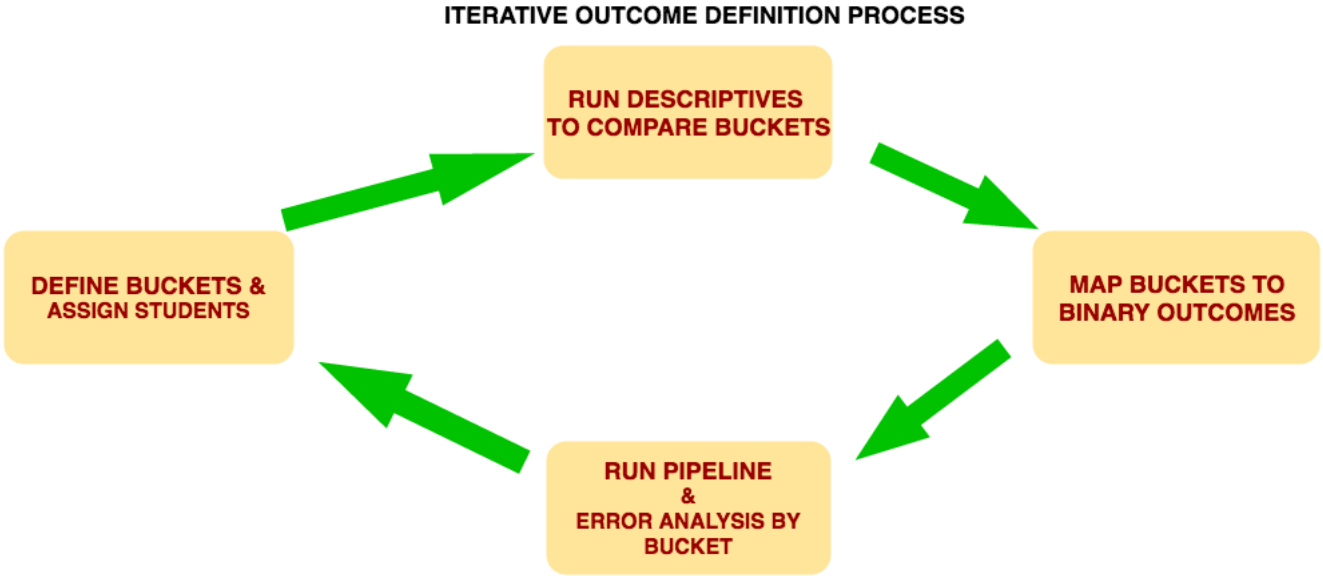
Extra Slides Following

- Never moved
- Accelerated or Advanced on all OAA exams
- Straight As
- Gifted
- No free/reduced lunch

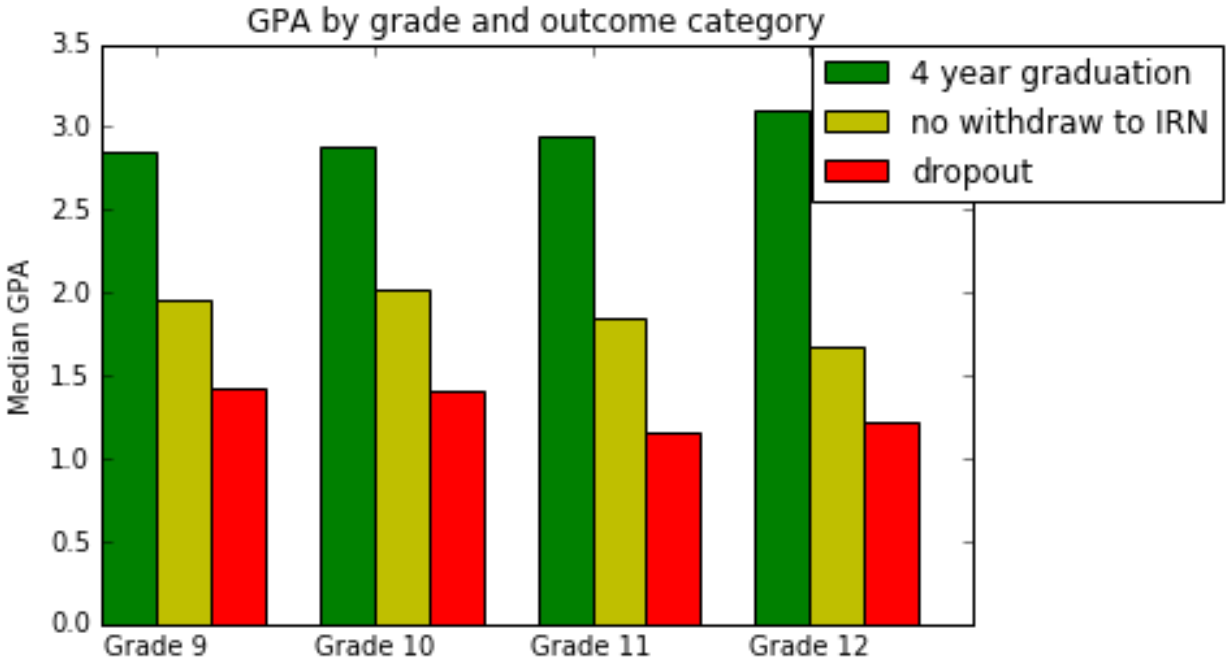
- Few unexcused absences
- One disciplinary incident
- Part time at Central Ohio Technical College
- Full time at OSU Newark Campus
- Graduates!



Flexible specification of outcome variable definition in pipeline

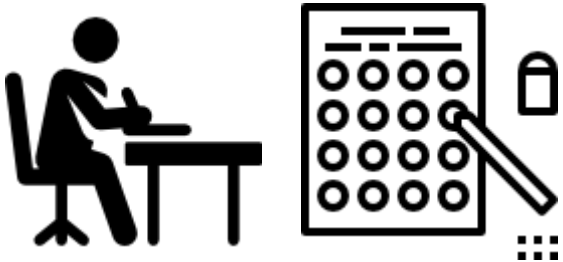


Do uncertain buckets of students look more like graduates or dropouts?



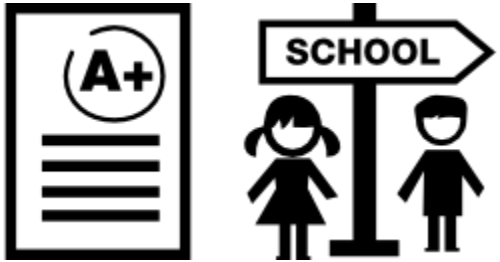
ETL | Stories | Outcomes | **Features** | Validation | Results

Ohio Standardized Tests



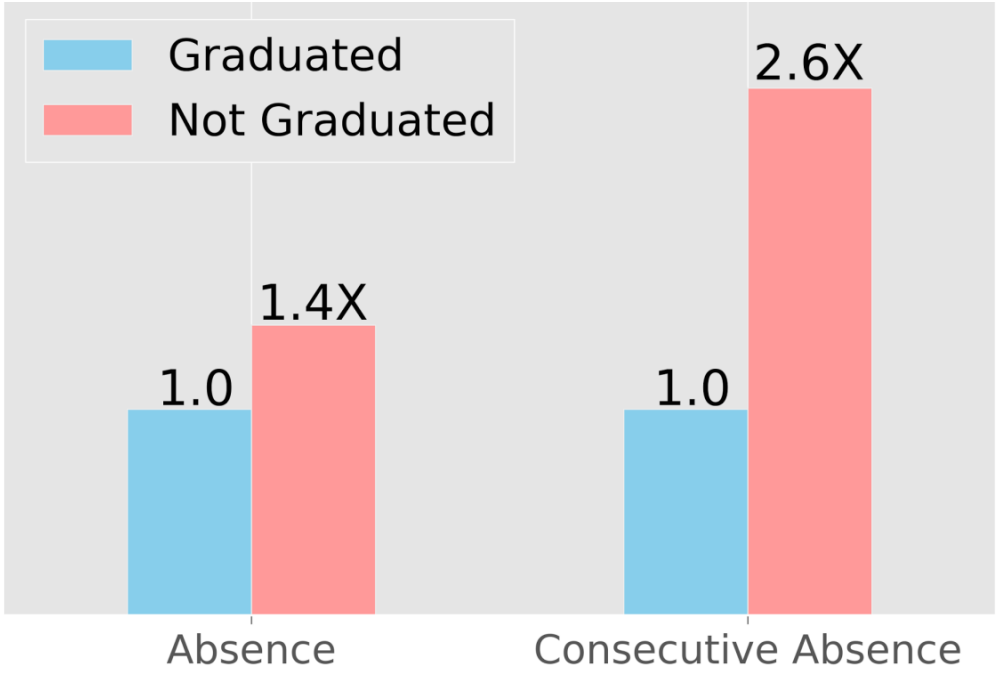
- Scores normalized by cohort
- Percentile
- Number of Std Deviations from Mean

School Grades



- Normalized by district / school
- GPA
- Hierarchical features

Absence Features



ETL | Stories | Outcomes | **Features** | Validation | Results

Absence Features

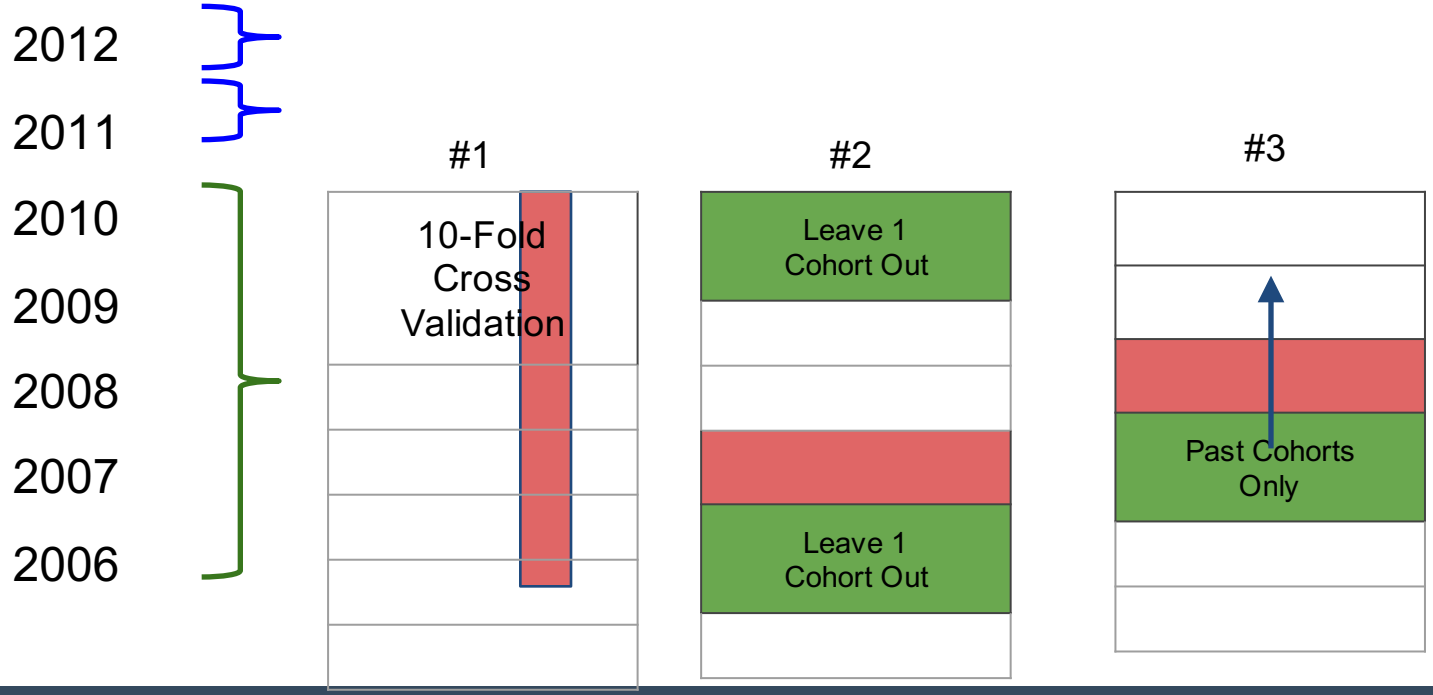
Aggregated Attendance

- Absence
- Unexcused Absence
- Tardy
- Unexcused Tardy
- Medical

Daily Attendance

- Consecutive Absence
- Consecutive Tardy

9th Grade Cohorts:

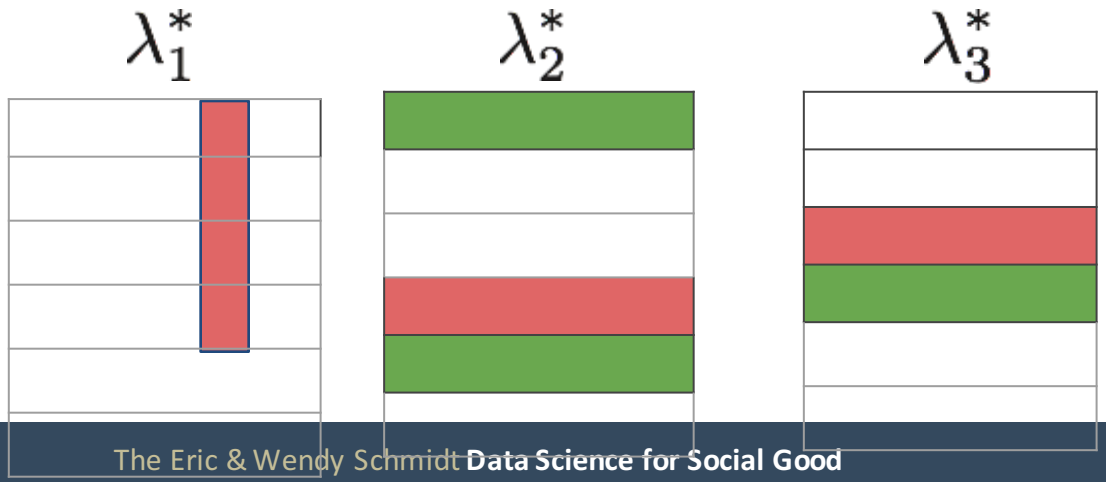


9th Grade Cohorts:

2012
2011
2010
2009
2008
2007
2006

Performance Estimate on 2012 Cohort

Test on 2011 Cohort = λ^*




```
model_classes_selected: [logit, DT, SVM]
user_description: initial results to use in the deep dive 6/26
file_save_name: deep_dive_1_year
write_predictions_to_database: False
random_seed: 2187

cohort_grade_level_begin: cohort_9th
prediction_grade_level: 9 # predicting at the end of this grade

model_test_holdout: temporal_cohort
parameter_cross_validation_scheme: leave_cohort_out
validation_criterion: custom_precision_10

cohorts_held_out: [2008]
cohorts_training: [2006, 2007]

features_included:
  demographics: [ethnicity, gender]
  grades: [gpa_gr_9]
  mobility: [n_addresses_to_gr_9, n_cities_to_gr_9, n_districts_to_gr_9]
  snapshots: [disadvantagement_gr_9,
              disability_gr_9,
              district_gr_9,
              gifted_gr_9,
              iss_gr_9,
              oss_gr_9,
              limited_english_gr_9,
              special_ed_gr_9,
              status_gr_9,
              days_absent_gr_9,
              days_absent_unexcused_gr_9,
              discipline_incidents_gr_9]

outcome_name: definite

missing_impute_strategy: median_plus_dummies
feature_scaling: robust
```

```
cohort_grade_level_begin: cohort_9th
```

```
prediction_grade_level: 9
```

```
outcome_name: definite
```

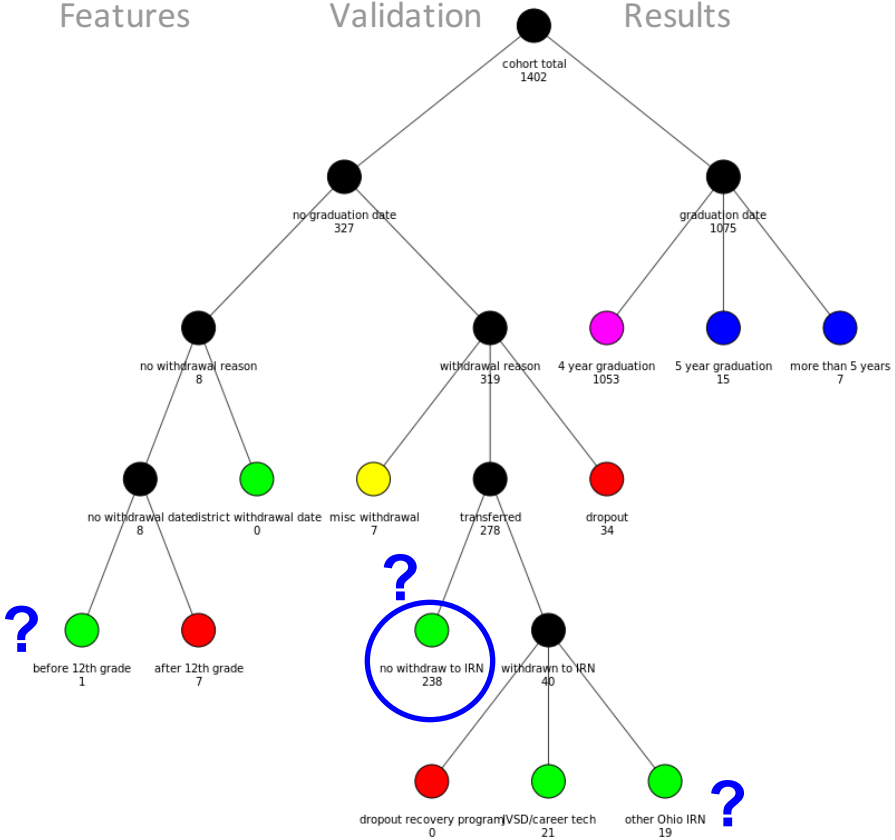
Approach 1

Provide fine-grained labels and experiment with various label-to-outcome mappings in the pipeline

- Revise labels and outcome mappings in iterative exploratory process

Approach 2

Use a semi-supervised approach treating uncertain outcomes as unlabeled data



Mid school-year moves

Between school-year moves

STUDENT MOBILITY

Change of schools only

Change of school and address

ADVERSE OUTCOMES