Deep Learning for Healthcare

3. Machine Learning Basics

Prof. Jimeng Sun

ILLINOIS

Outline

Supervised learning

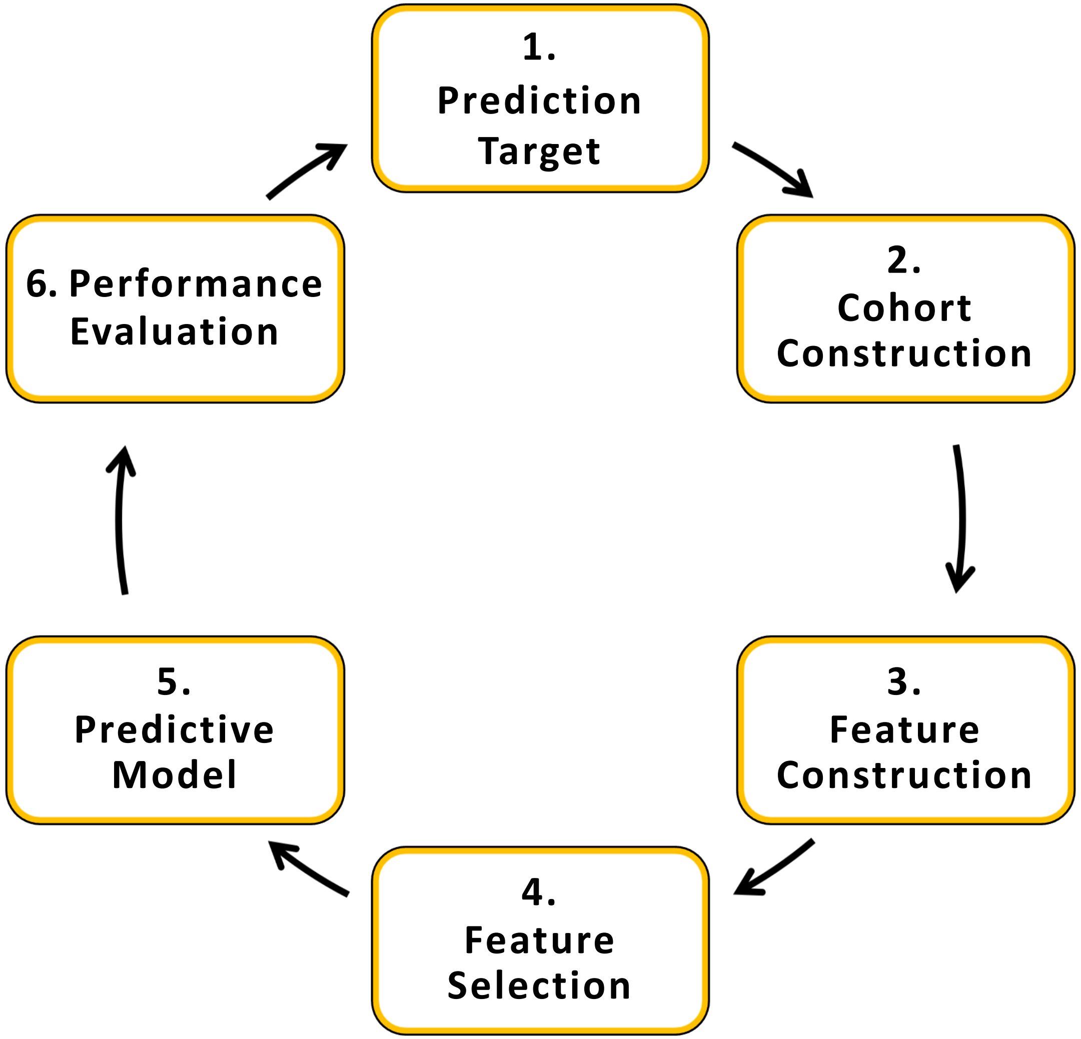
Unsupervised learning

Evaluation metrics

Supervised learning

- Predictive Modeling Pipeline
- Gradient and Stochastic gradient descent

Predictive Modeling Pipeline



1. Prediction Target

2. Cohort Construction

3. Feature Construction

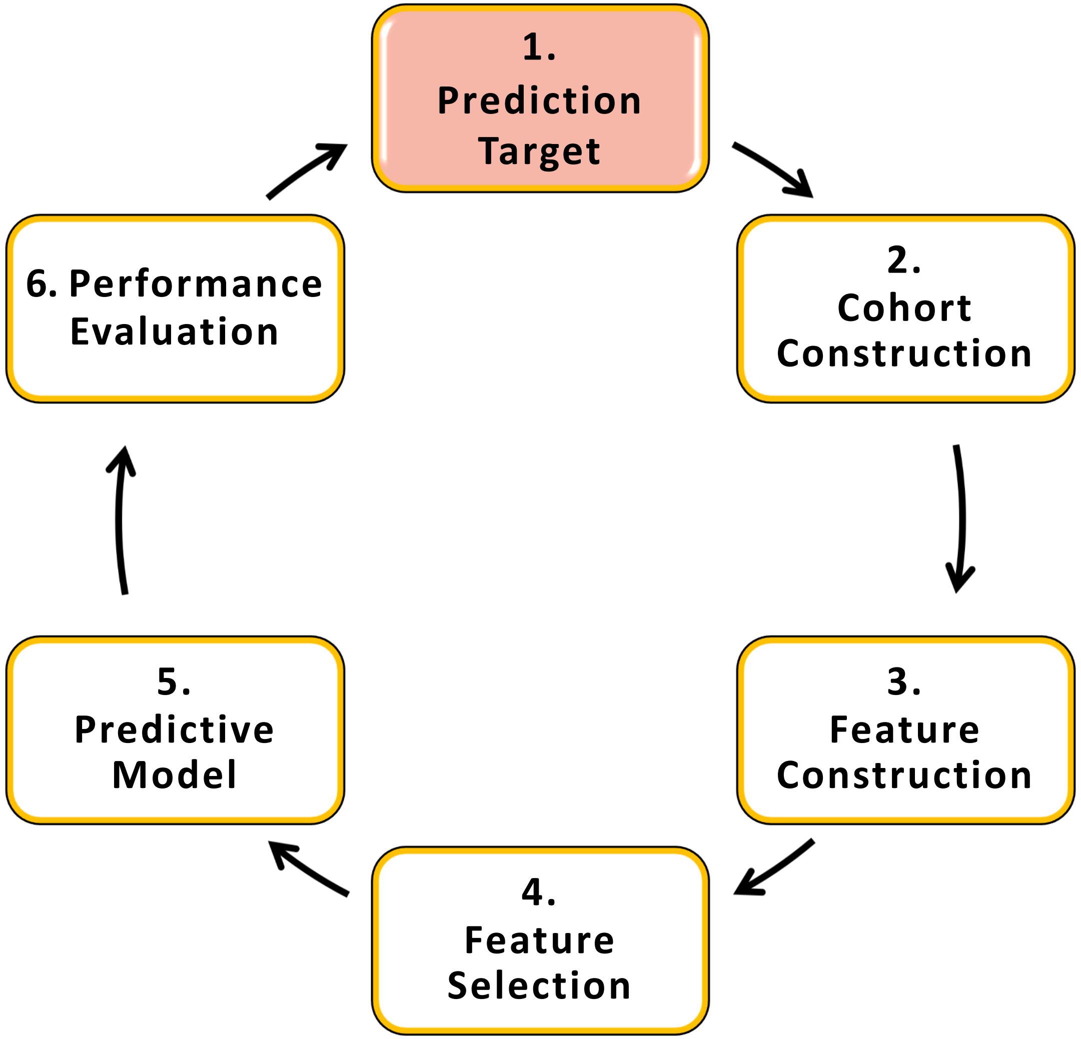
4. Feature Selection

5. Predictive Model

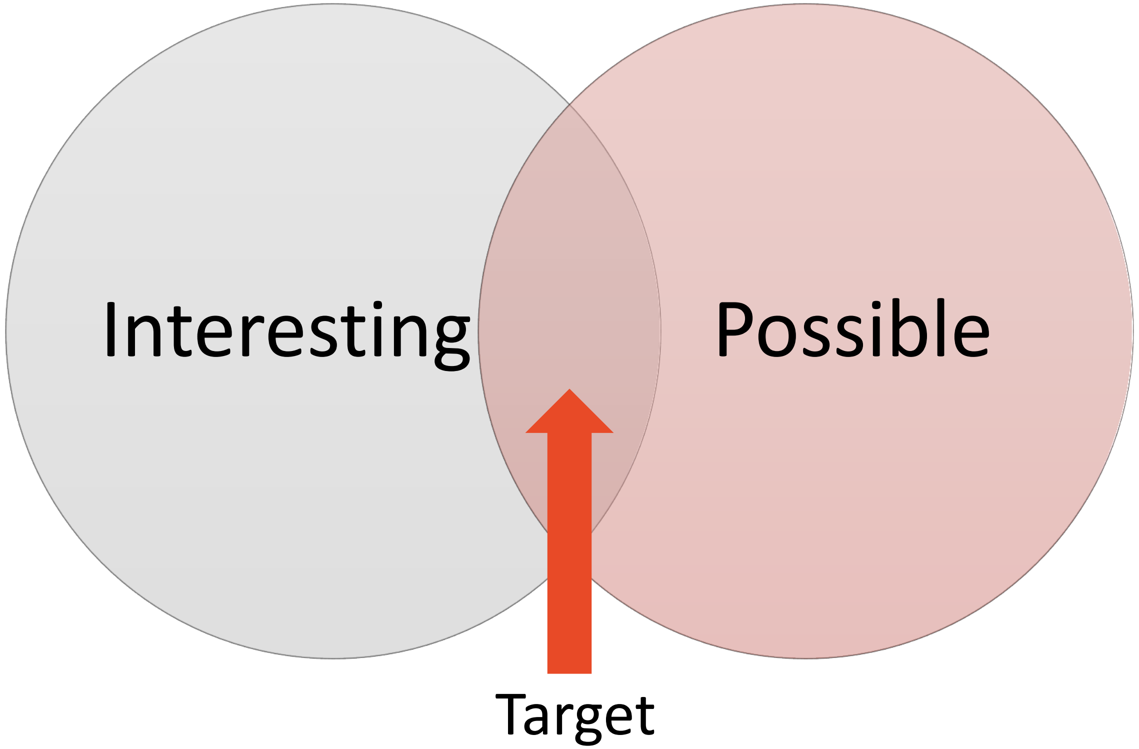
6. Performance Evaluation

ILLINOIS

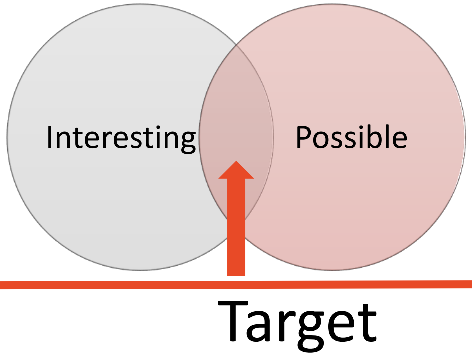
Predictive Modeling Pipeline



Prediction Target

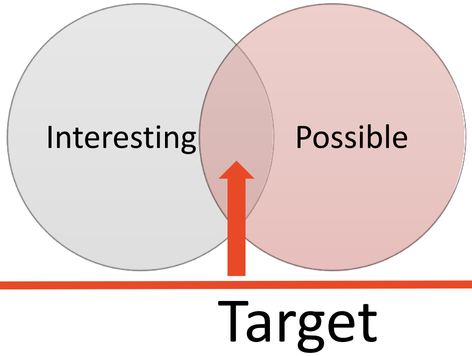


How do you know the target is interesting?



- Talk to domain experts
- Read domain publications
- General common sense metrics
  - High cost
  - Long time
  - Bad quality

How do you know the target is possible?

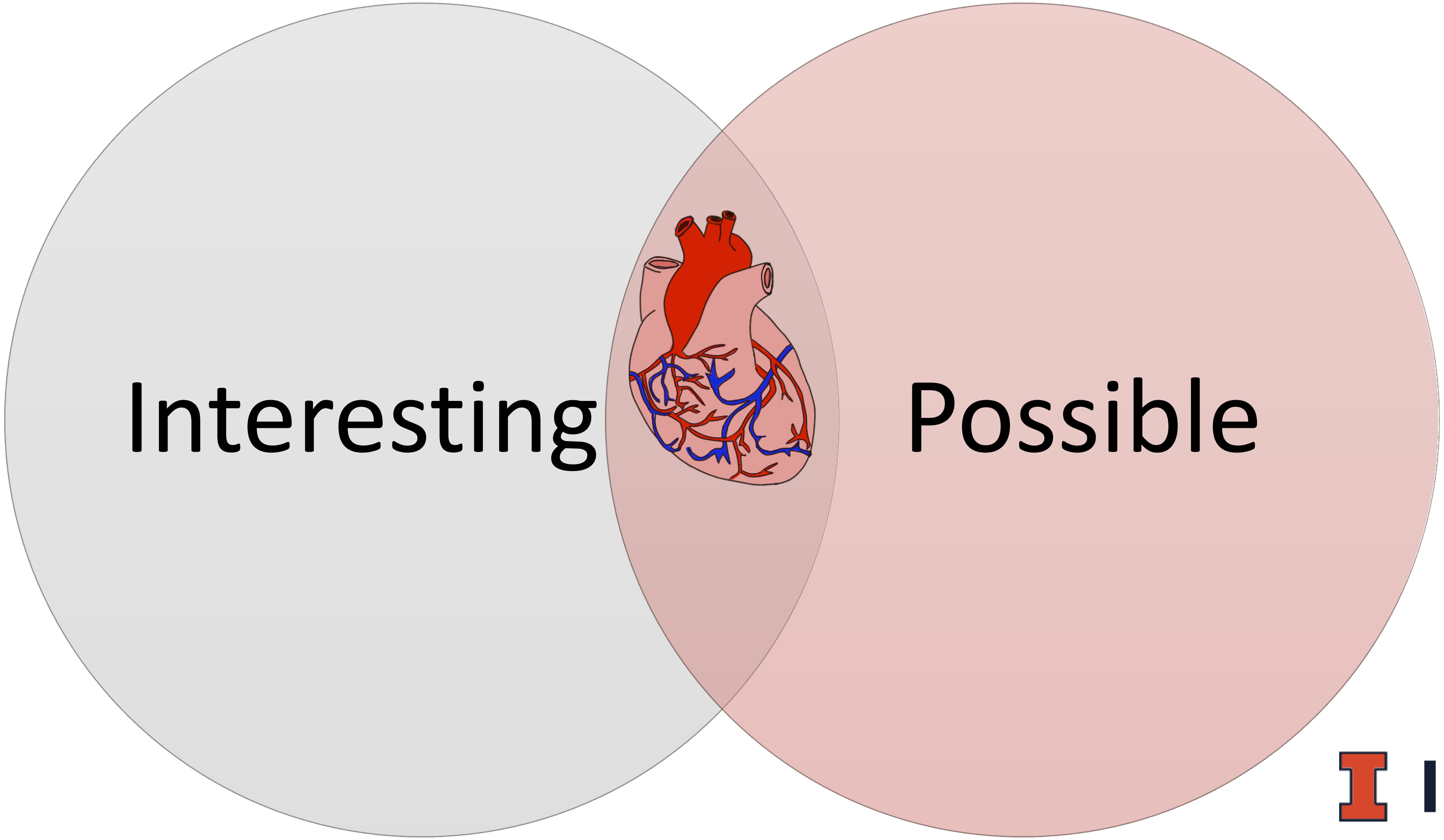


- Human performance

- Experience from similar projects

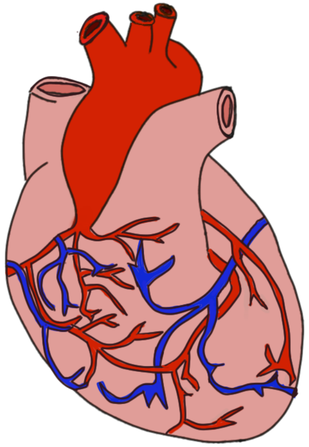
- Results from prior publications ?

Prediction Target



Heart Failure Quiz

How many new cases of heart failure occur each year in



□ A. 17,000

□ B. 260,000

☑ C. 550,000

□ D. 1,250,000

Motivations For Early Detection Of Heart Failure

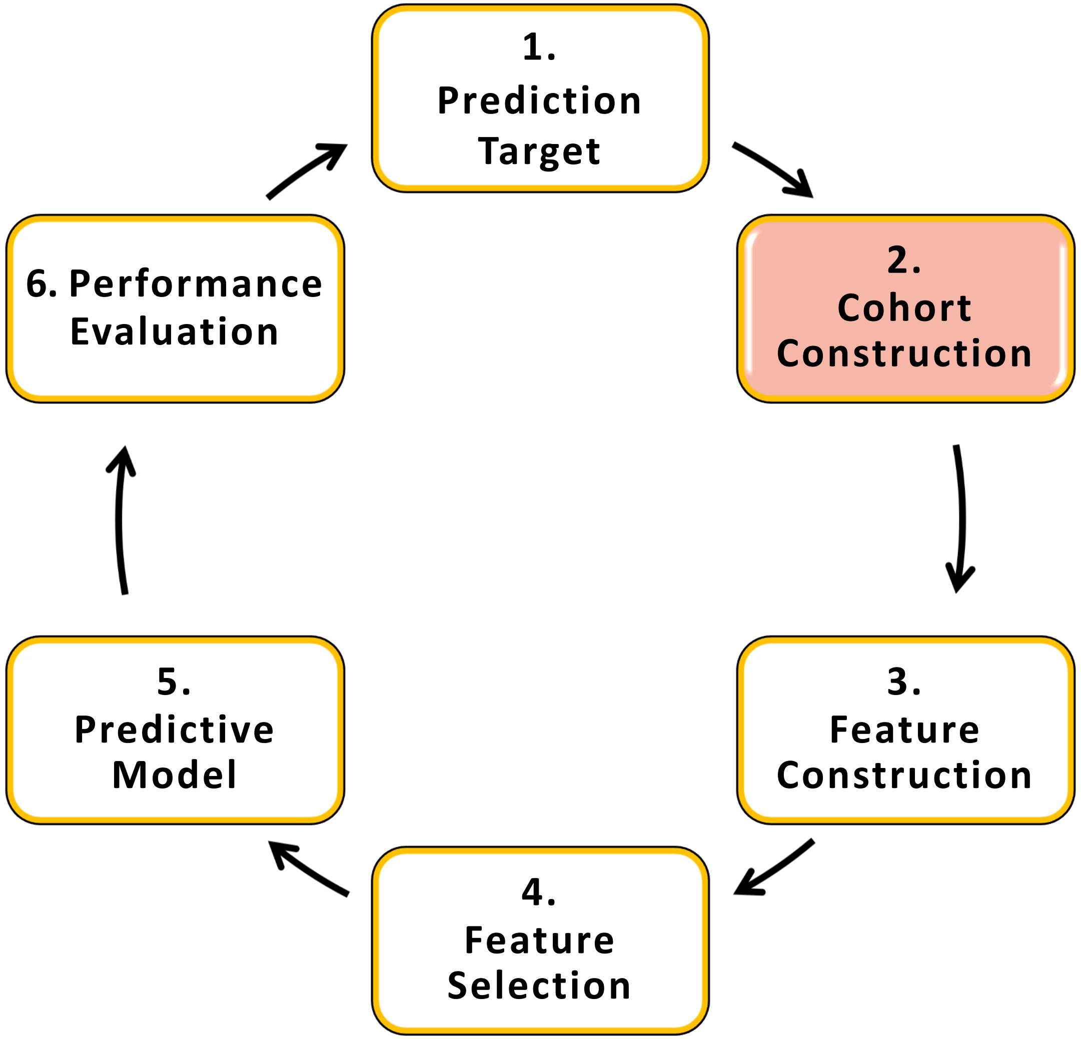
Heart failure is a complex disease.

Reduces cost and hospitalization.

Improves existing clinical guidelines of HF prevention.

Early intervention can slow down disease progression.

Predictive Modeling Pipeline



1. Prediction Target

2. Cohort Construction

3. Feature Construction

4. Feature Selection

5. Predictive Model

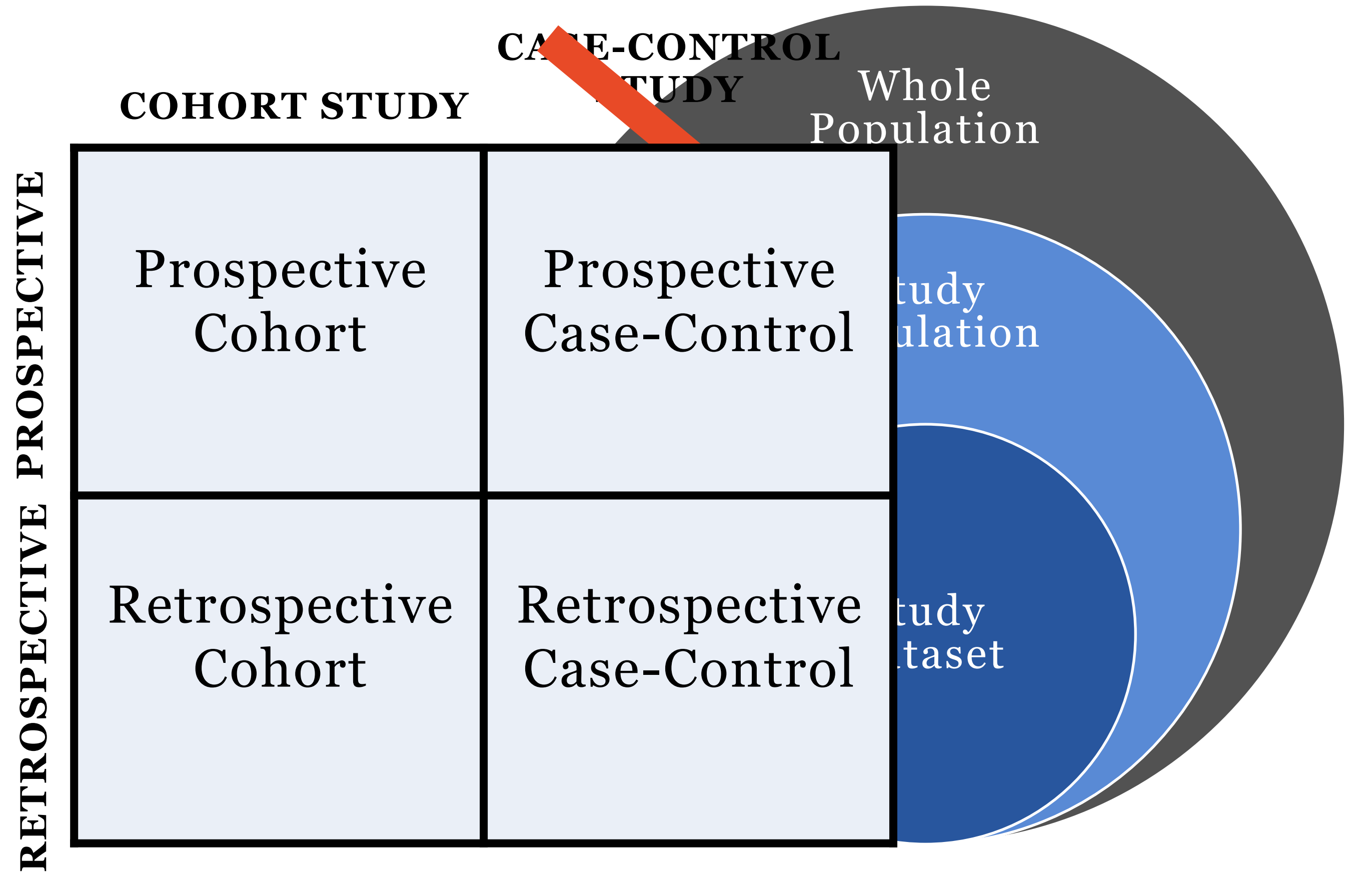
6. Performance Evaluation

Why do we need cohort construction?

- General ML practice:
  - Give a dataset => build a model
- Healthcare predictive modeling
  - Create a dataset (cohort, feature) => build a model
- Reasons:
  - Avoid obvious models: e.g., age predicts mortality
  - Focus on the population of interests
  - Data acquisition cost

COHORT CONSTRUCTION - STUDY DESIGN

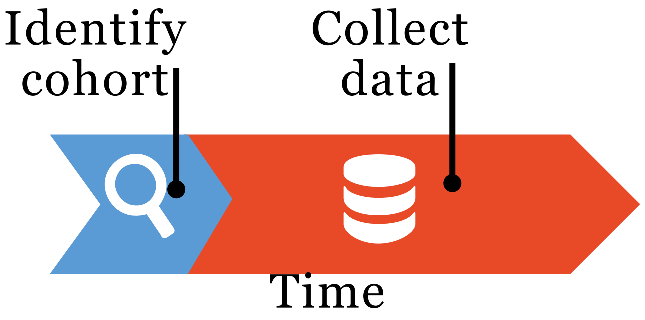
DEFINE



	COHORT STUDY	CASE-CONTROL STUDY
PROSPECTIVE	Prospective Cohort	Prospective Case-Control
RETROSPECTIVE	Retrospective Cohort	Retrospective Case-Control

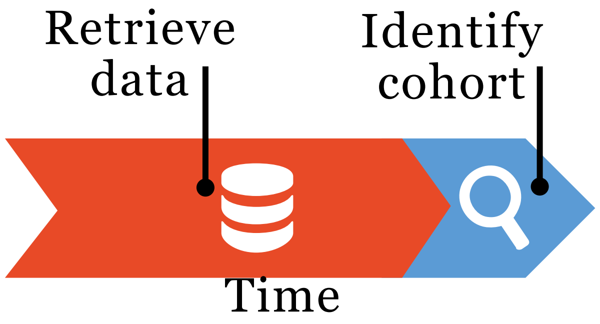
PROSPECTIVE VS. RETROSPECTIVE

PROSPECTIVE



- Identify the study population P
- Then start to collect data on P

RETROSPECTIVE



- Identify the study population P
- Trace back the historical data of P

QUIZ: PROSPECTIVE VS. RETROSPECTIVE

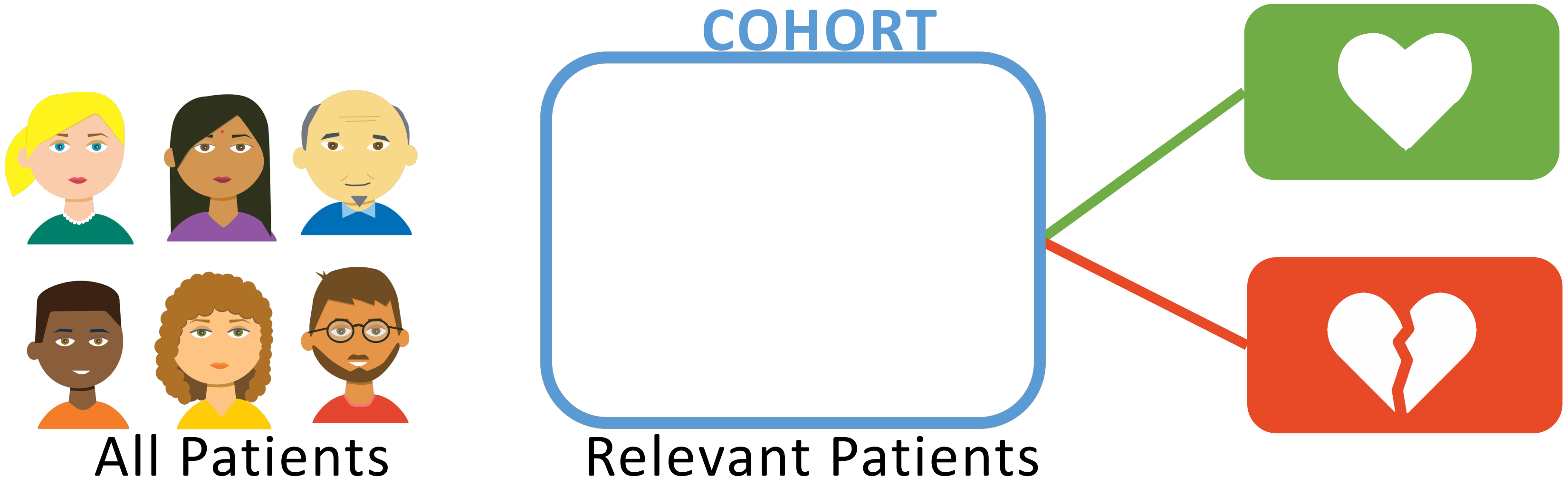
Property	Prospective Study	Retrospective Study
More noise in the data	□	☑
More expensive	☑	□
Takes a longer time	☑	□
Common on large dataset	□	☑

Cohort Study

Select a group of patients who are exposed to the risk

TARGET: Heart Failure Readmission

- COHORT: all HF patients discharged from hospital
- KEY: define the right inclusion/exclusion criteria

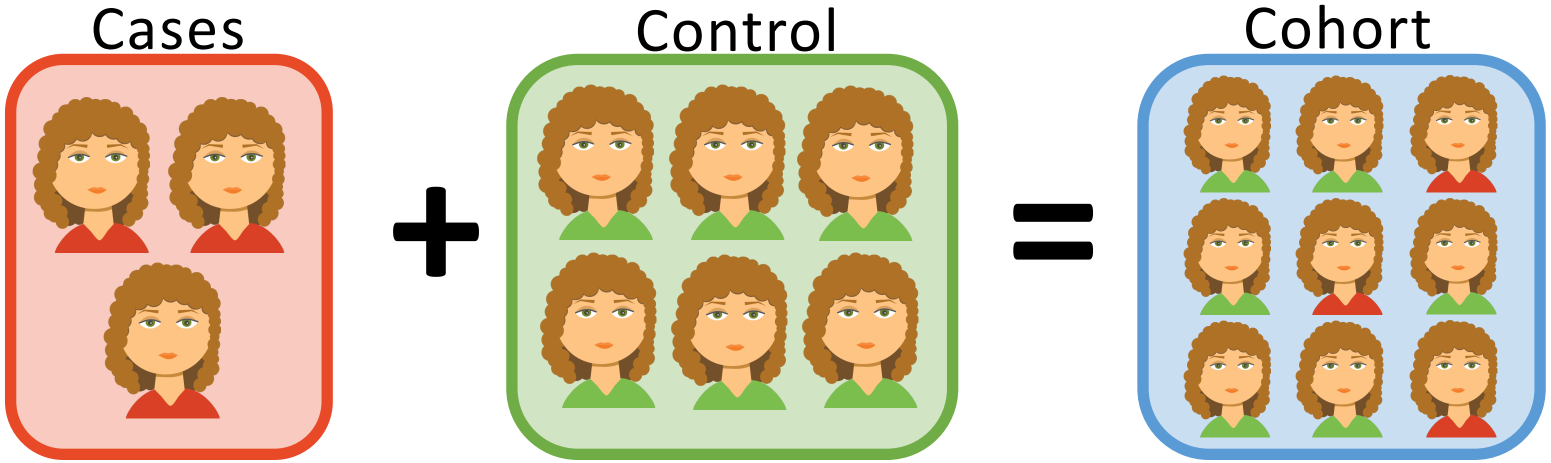


Case-control Study

CASES: patients with positive outcome (have the disease)

CONTROLS: patients with negative outcome (healthy) but otherwise similar

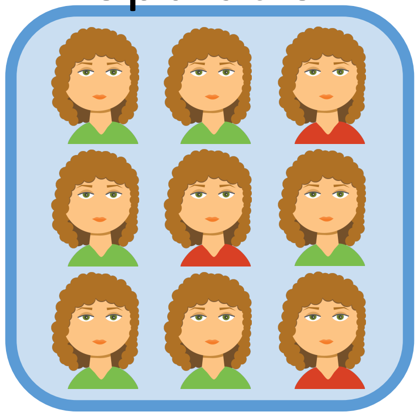
KEY: matching criteria between cases and controls

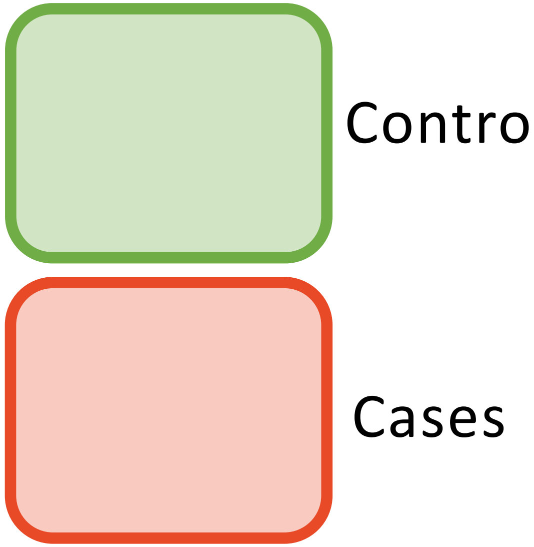


Example Of Case-control Study

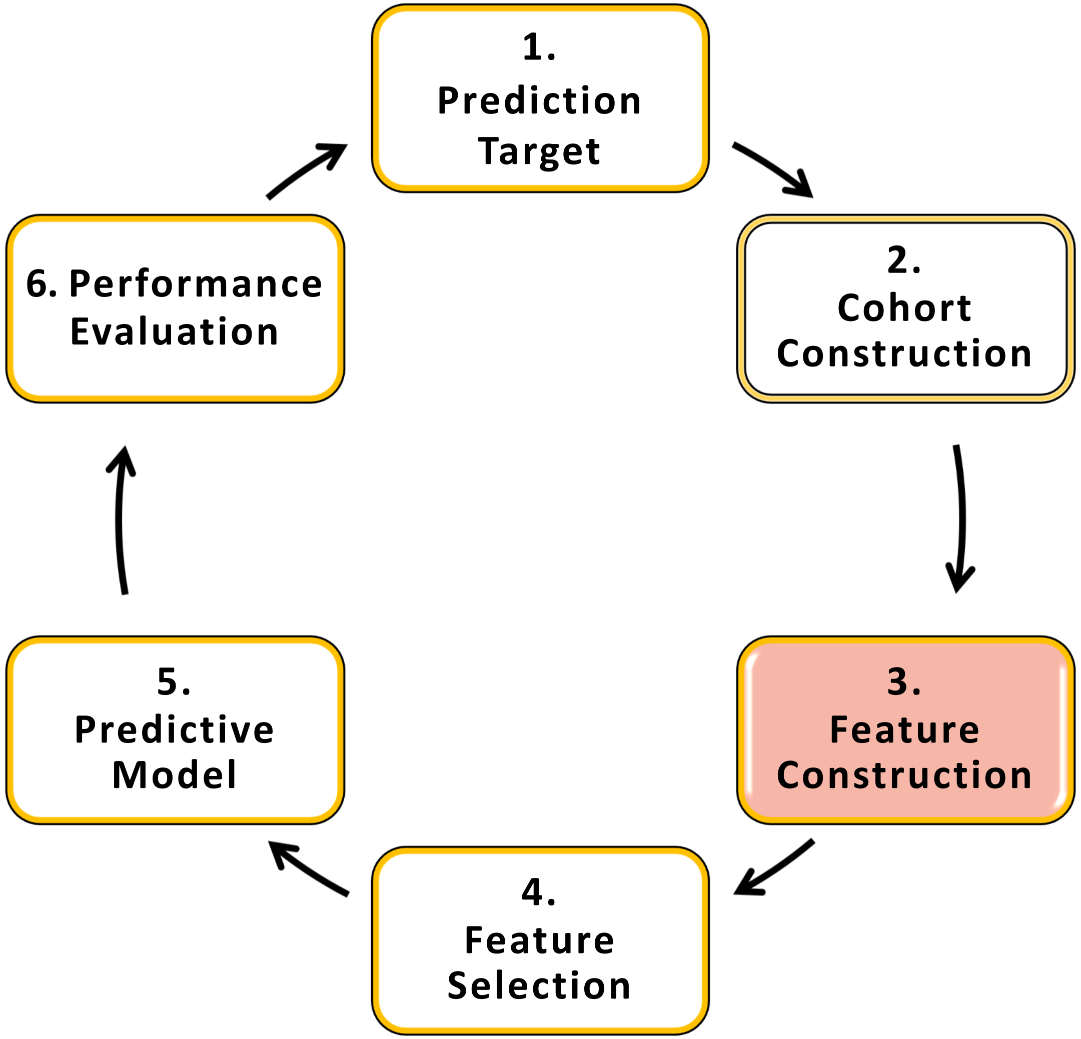
- Goal: Predict Heart Failure cases against control patients
- Population: 50,625 Patients
  - Case Patients: 4,644
  - Controls: 45,981 (matched on age, gender and clinic)

Population

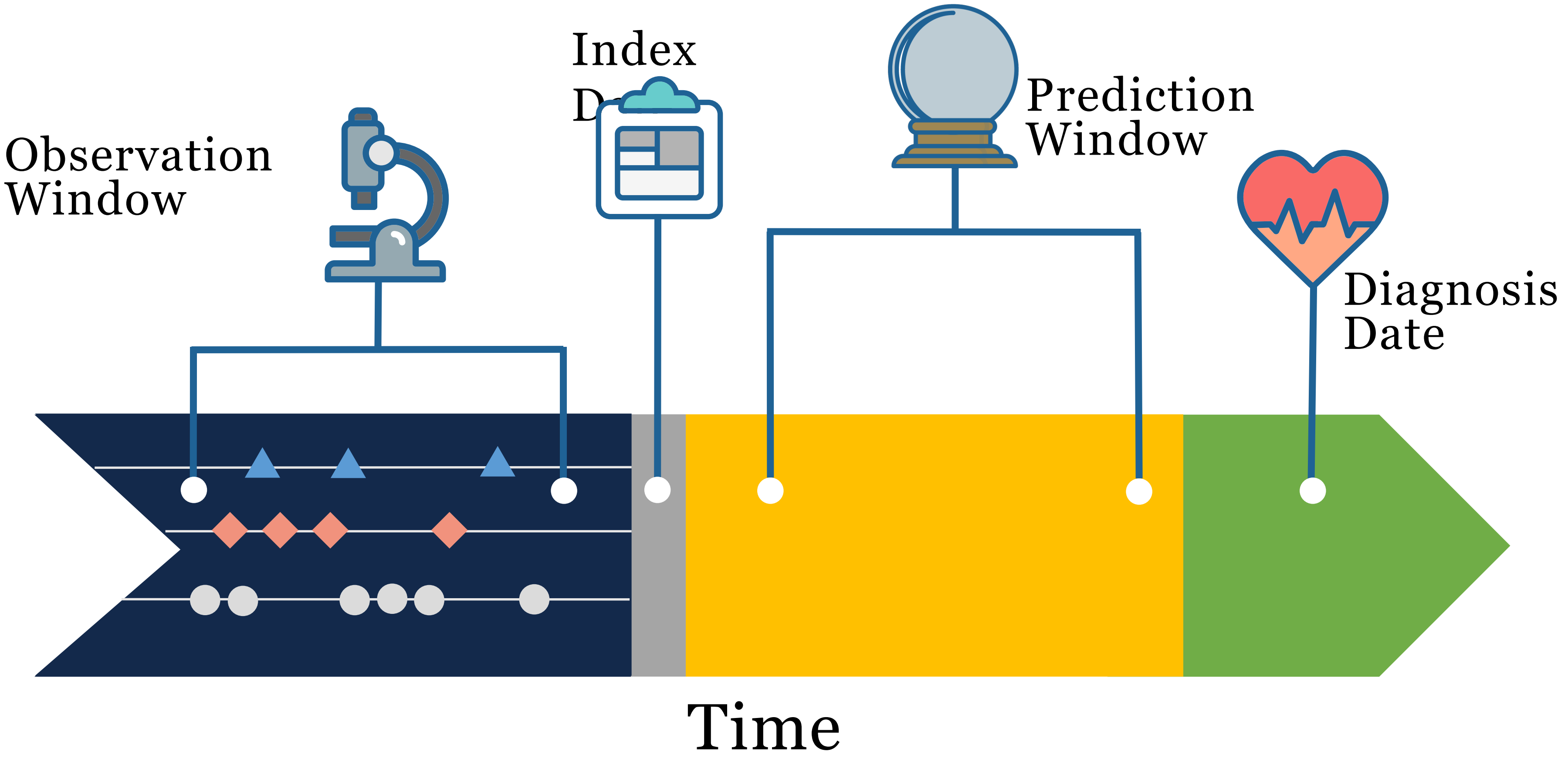




Predictive Modeling Pipeline

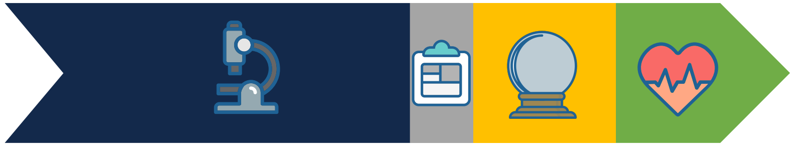


Feature Construction

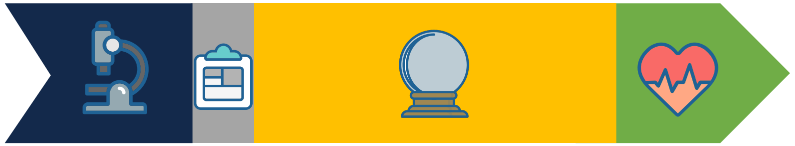


Feature Construction Quiz 1

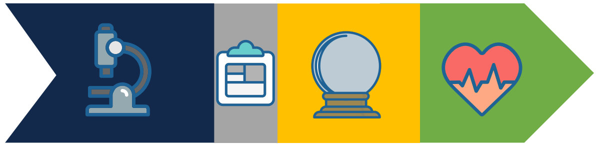
Which one of these timelines is the easiest for modeling?



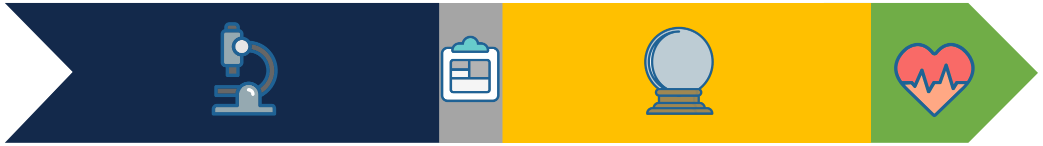
A. ☑



B. □



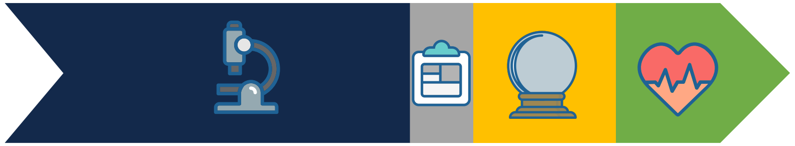
C. □



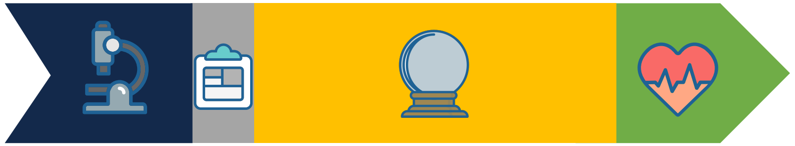
D. □

FEATURE CONSTRUCTION QUIZ 2

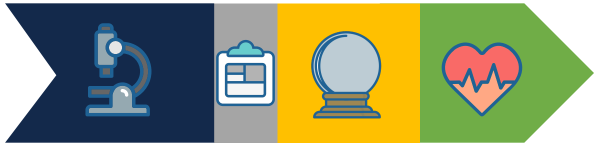
Which one of these timelines is the most useful model?



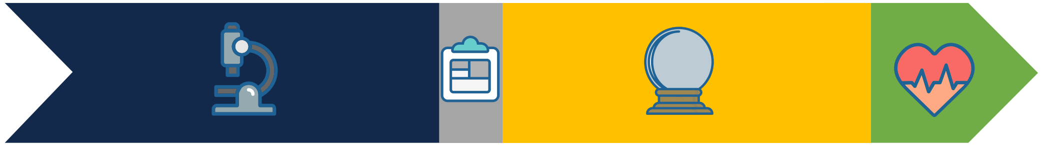
A. □



B. ☑

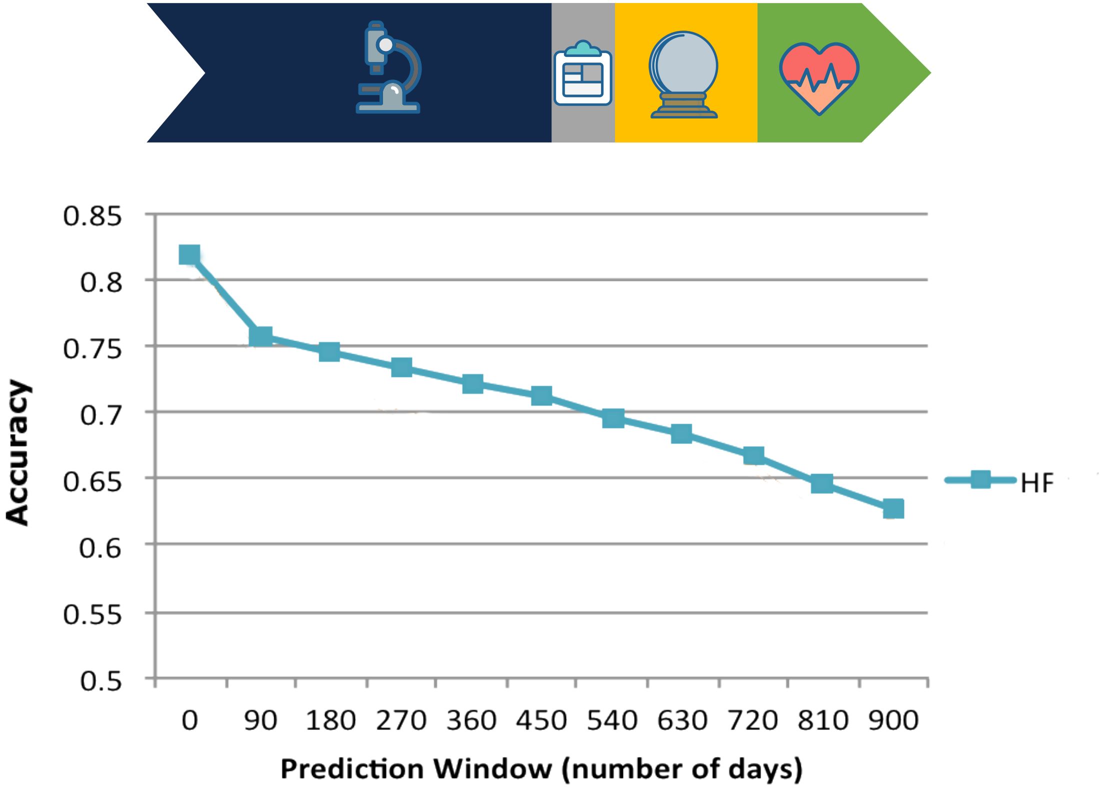


C. □



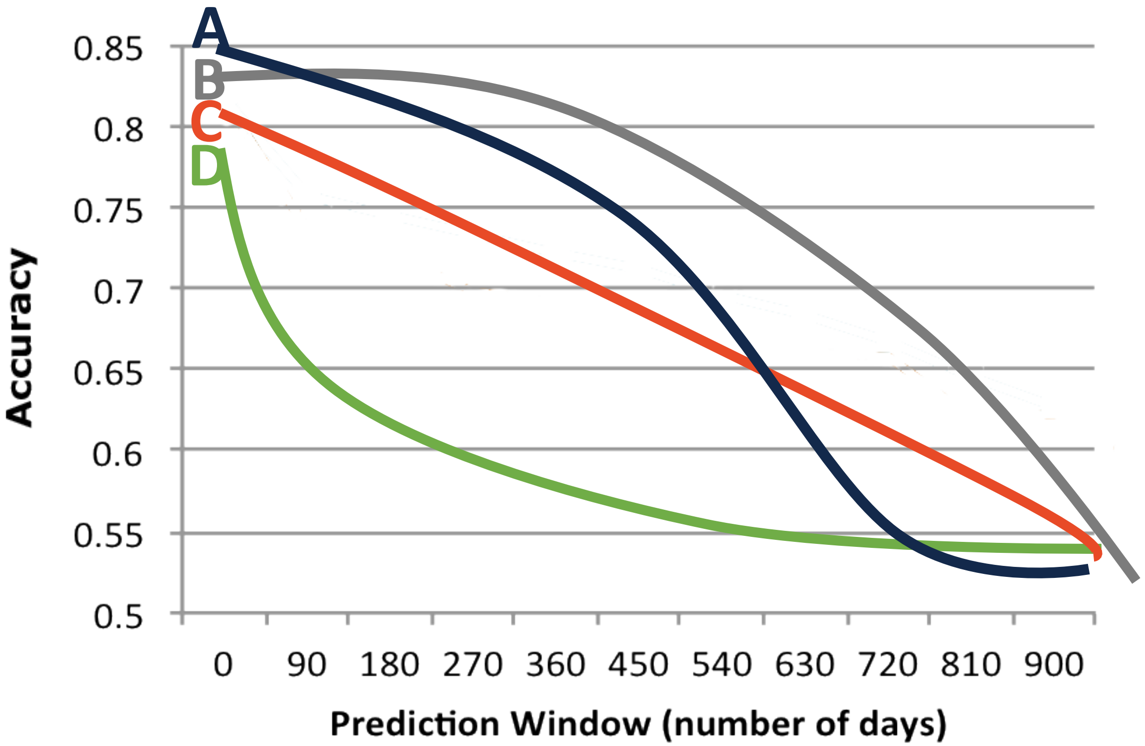
D. □

Prediction Performance On Different Prediction Windows



Prediction Window Quiz

Which of these options is the most desirable prediction curve?



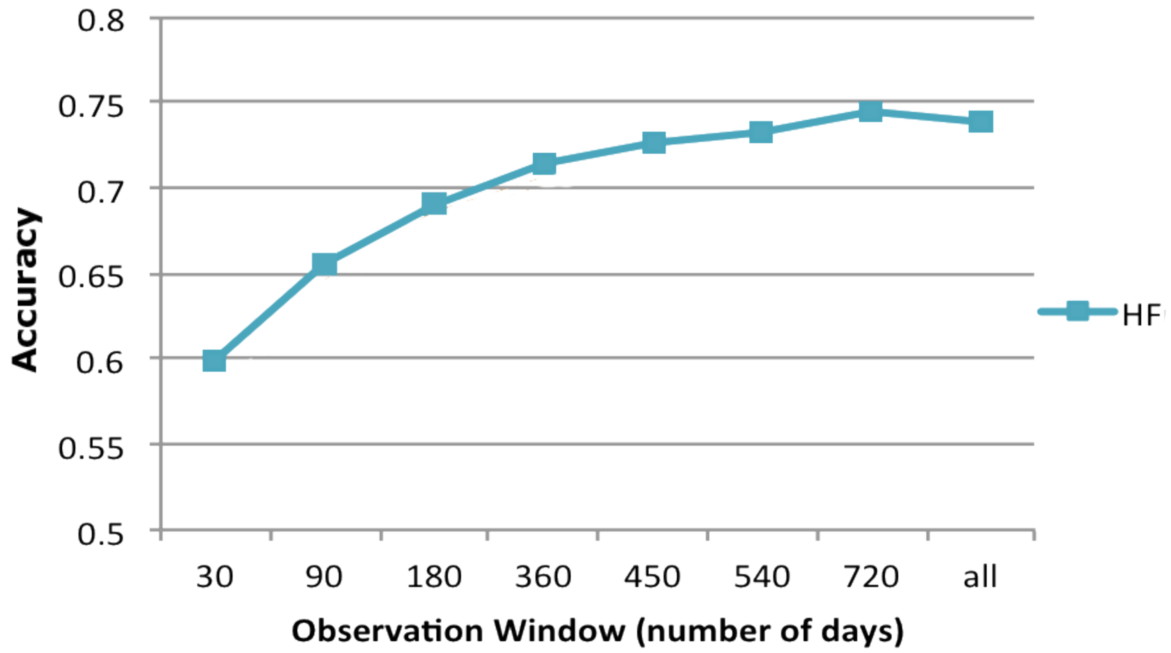
□ A

☑ B

□ C

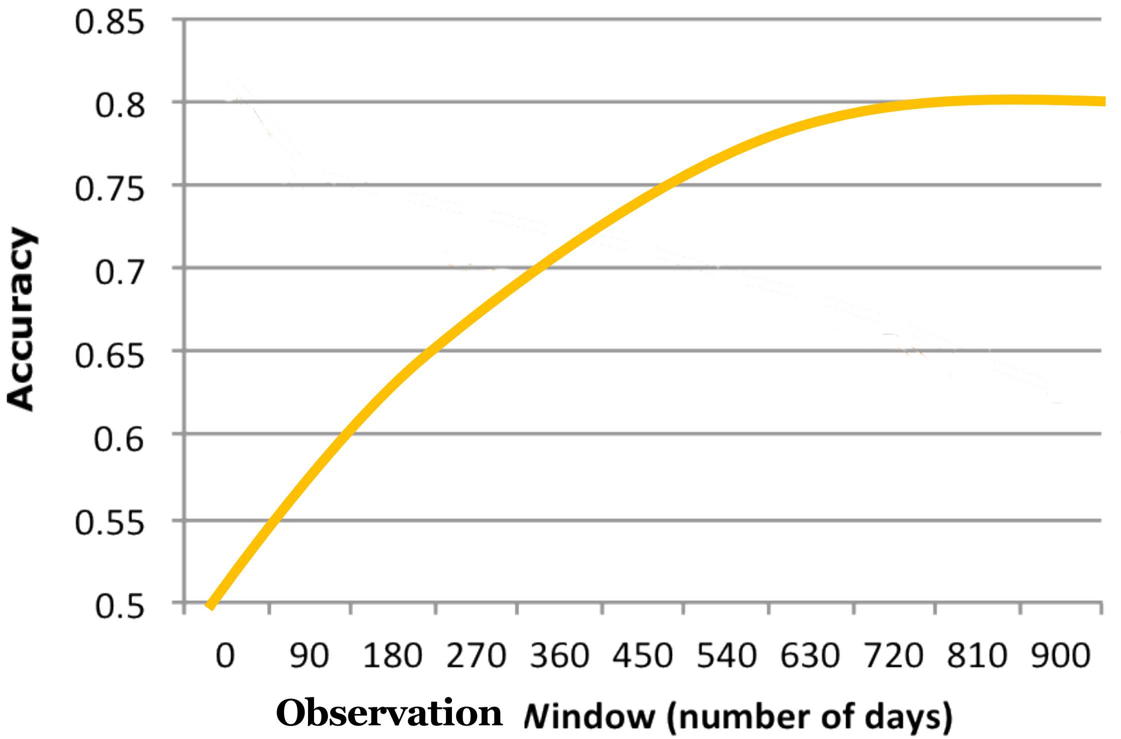
□ D

Prediction Performance On Different Observation Windows



Observation Window Quiz

What is the optimal observation window?



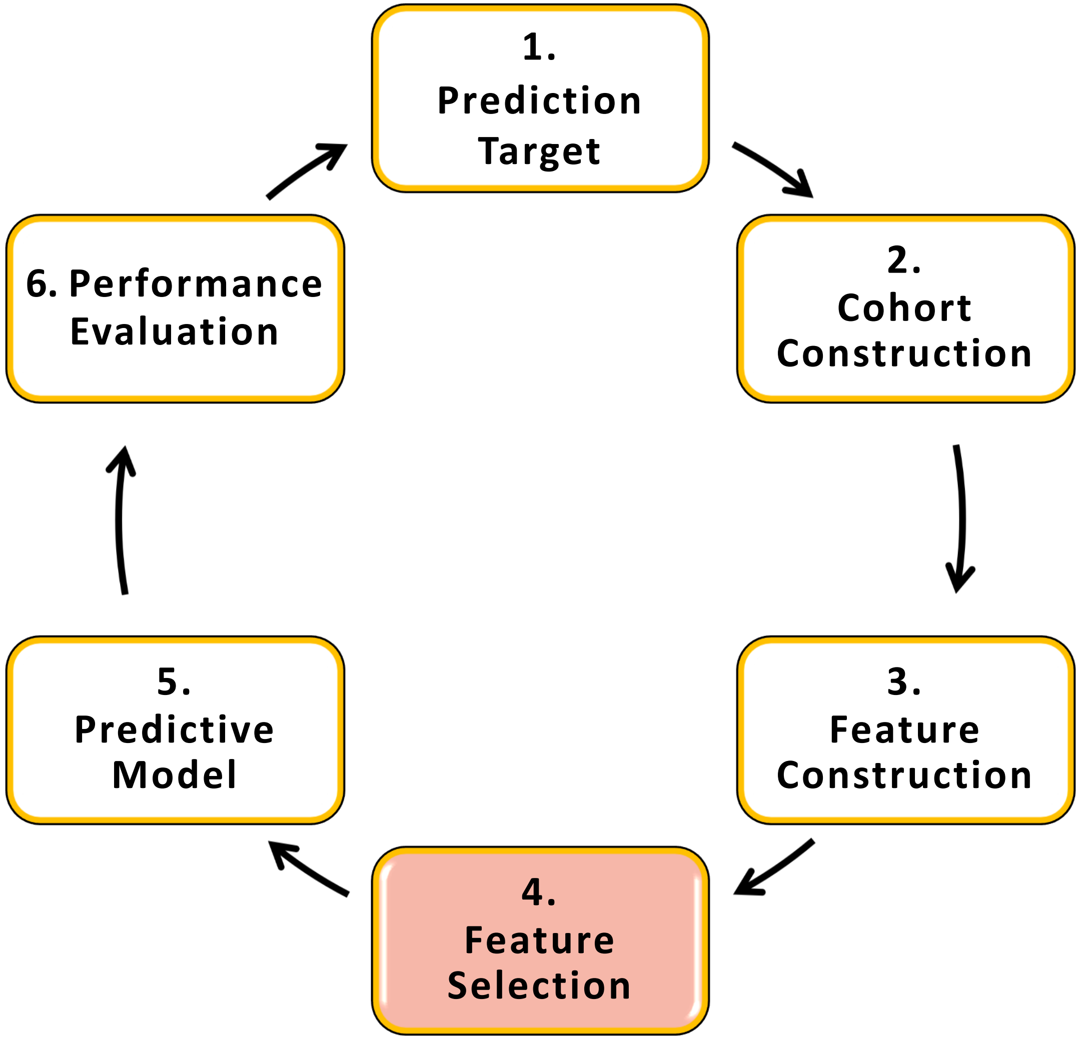
□ A. 90 days

□ B. 270 days

☑ C. 630 days

□ D. 900 days

Predictive Modeling Pipeline



1. Prediction Target

2. Cohort Construction

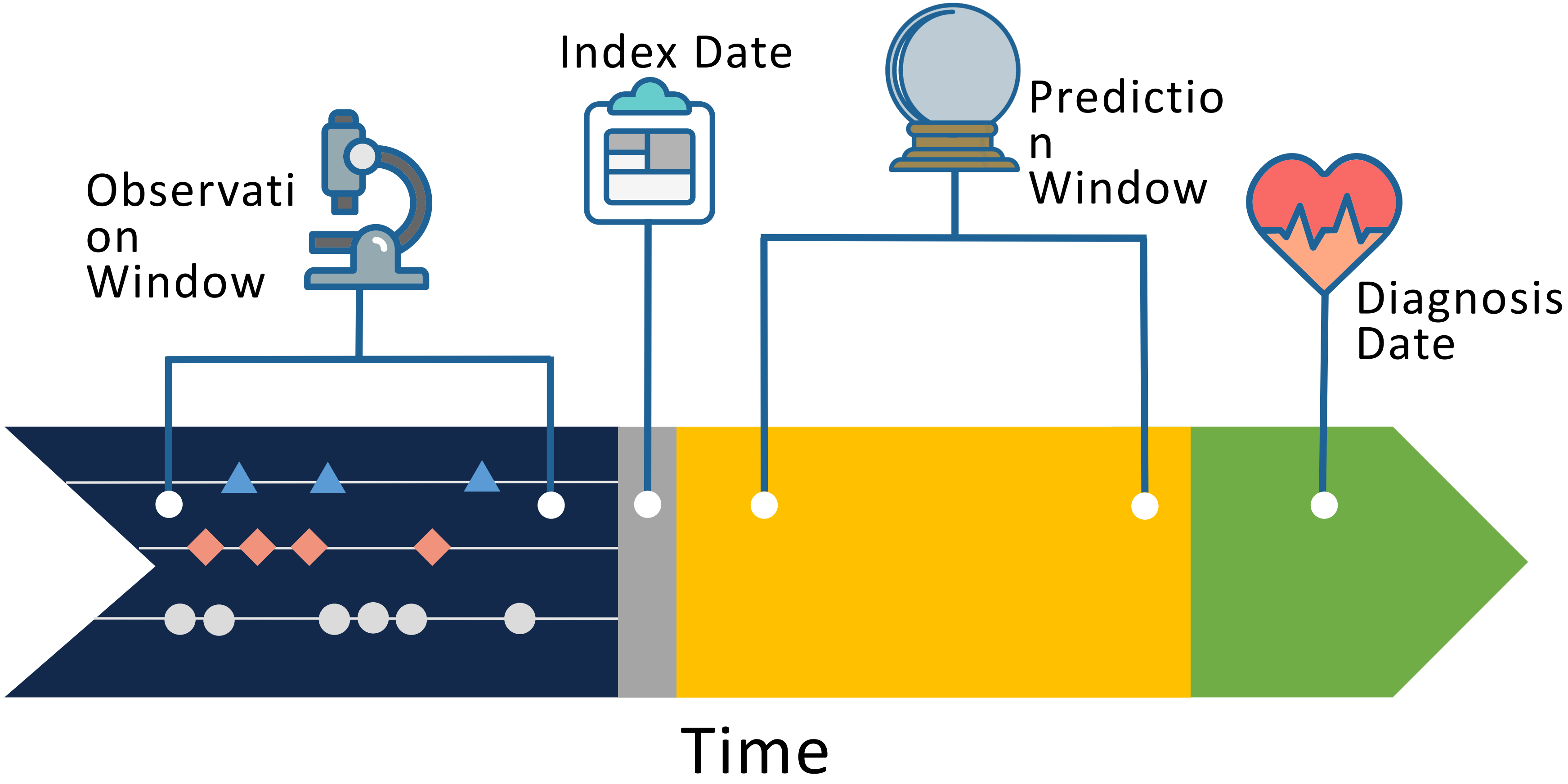
3. Feature Construction

4. Feature Selection

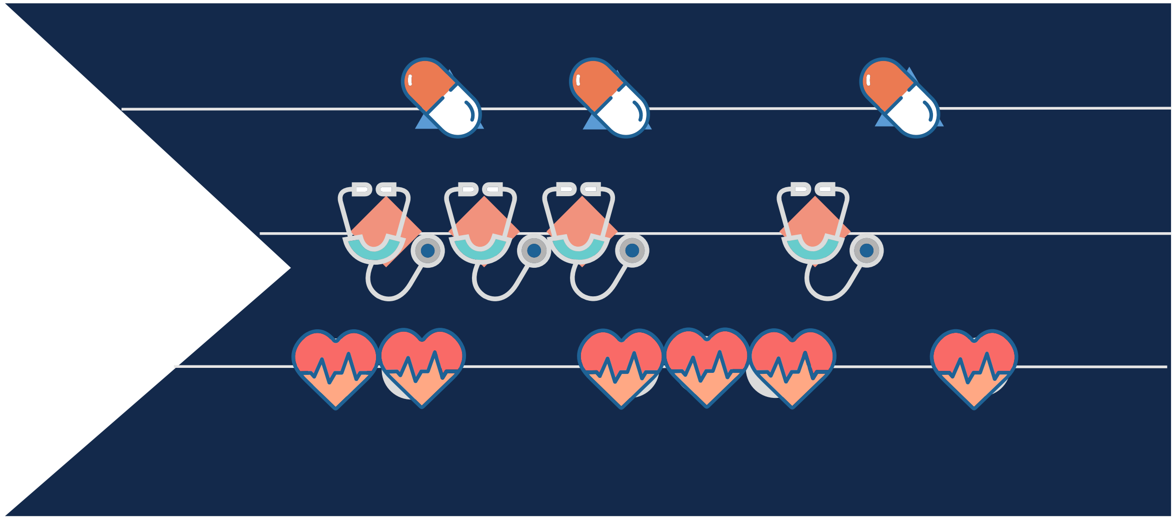
5. Predictive Model

6. Performance Evaluation

Feature Selection



Feature Selection



Feature Types

- Demographics
- Diagnosis
- Lab result
- Symptoms
- Medications
- Vitals

Feature Selection

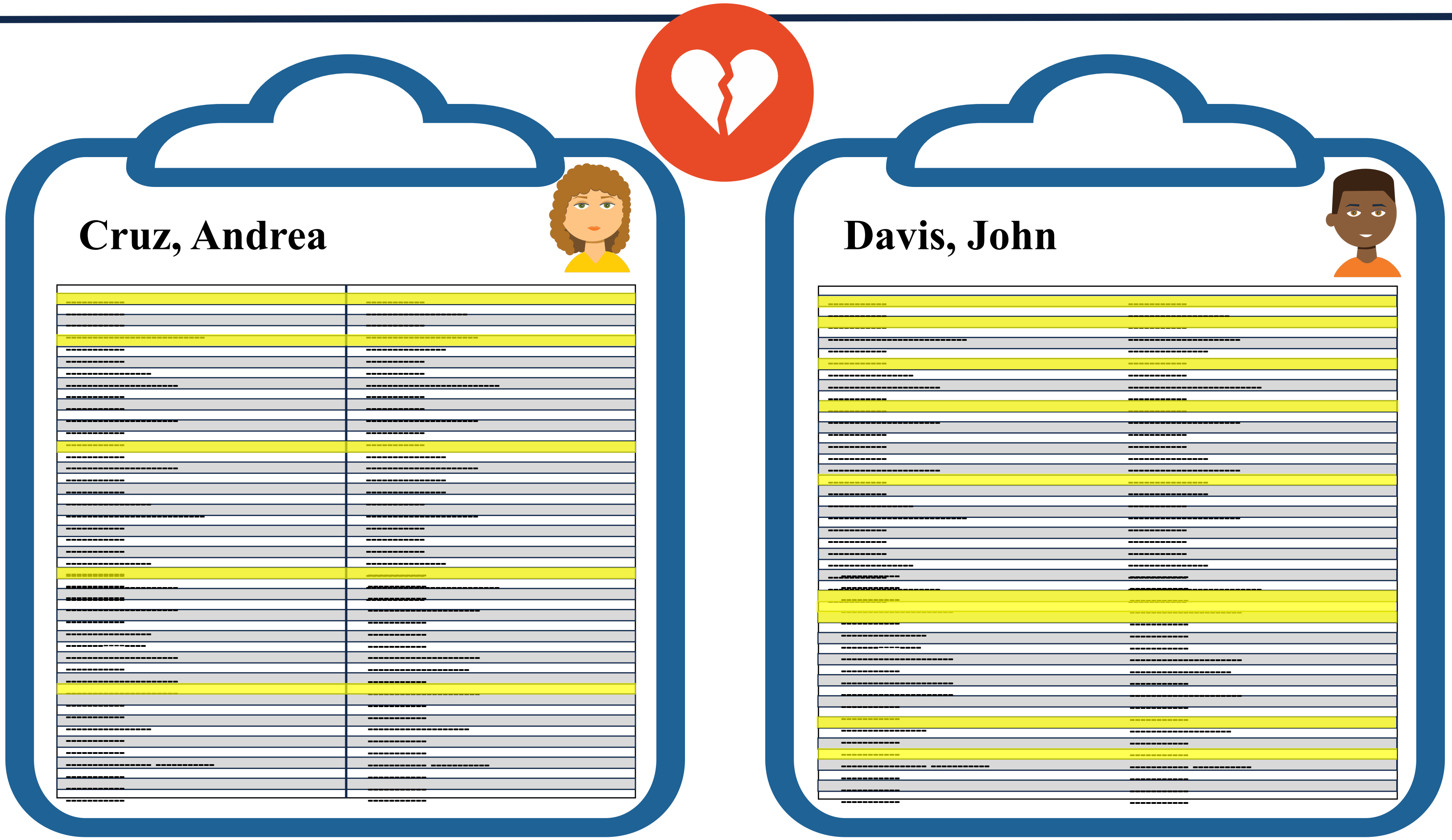
Cruz, Andrea

Age	34
Sex	F
Race	White
Blood Pressure	114/72
Diabetes, Type II	YES
Hypertension	NO

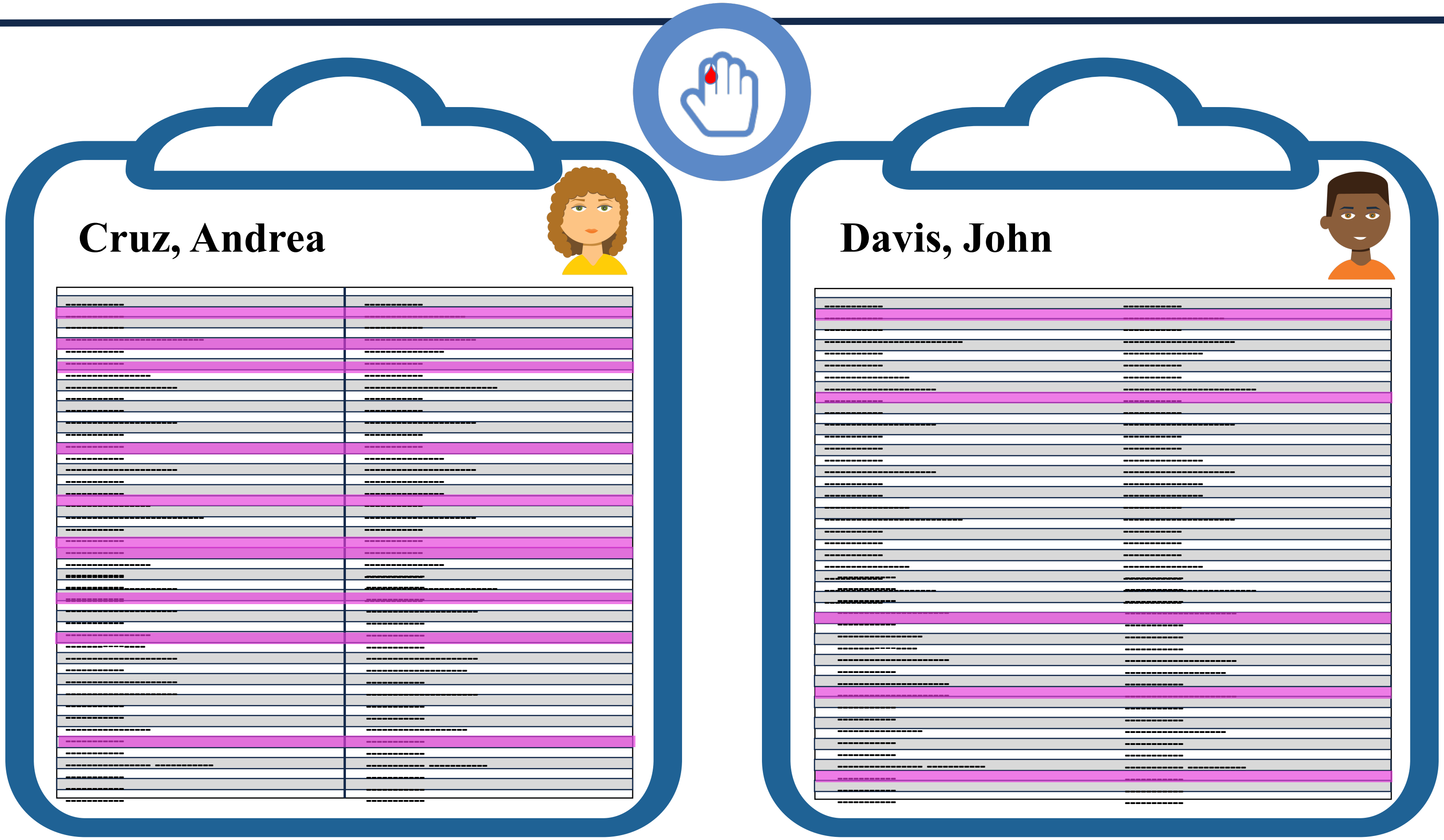
Davis, John

Age	67
Sex	M
Race	African American
Blood Pressure	160/100
Diabetes, Type II	NO
Hypertension	YES

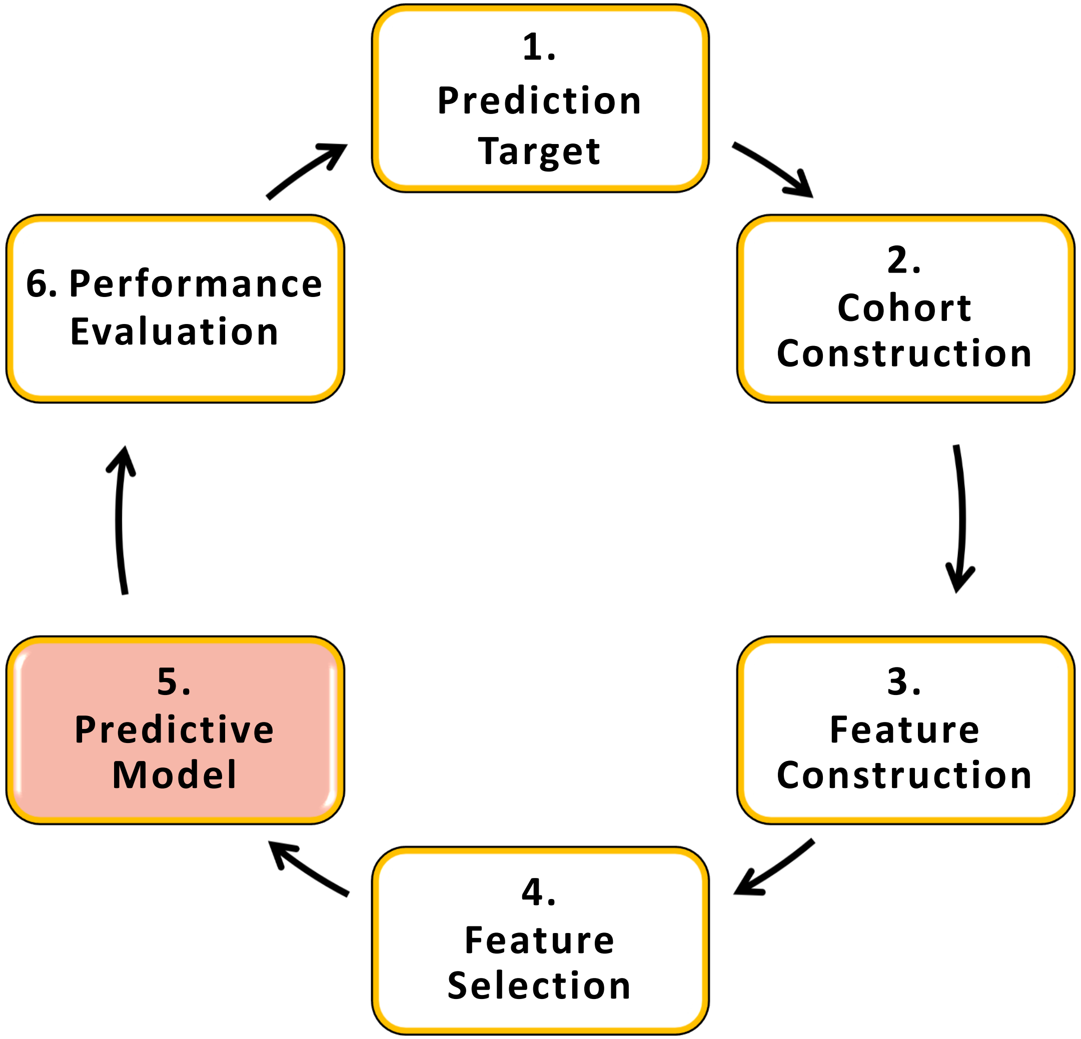
Feature Selection



Feature Selection



Predictive Modeling Pipeline



1. Prediction Target

2. Cohort Construction

3. Feature Construction

4. Feature Selection

5. Predictive Model

6. Performance Evaluation

Predictive Models

$$\overset{\text{Target}}{y} = f(\underset{\text{Features}}{x}) + \overset{\text{Error}}{e}$$

REGRESSION

- Target y is continuous
- Popular Methods
  - Linear Regression
  - Generalized Additive Models

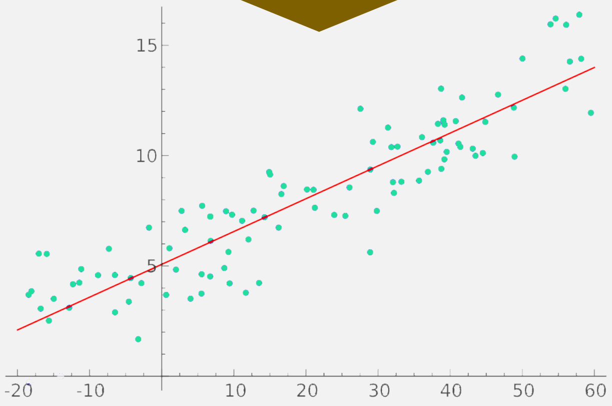
CLASSIFICATION

- Target y is categorical
- Popular Methods
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree, random forest
  - Neural networks

REGRESSION

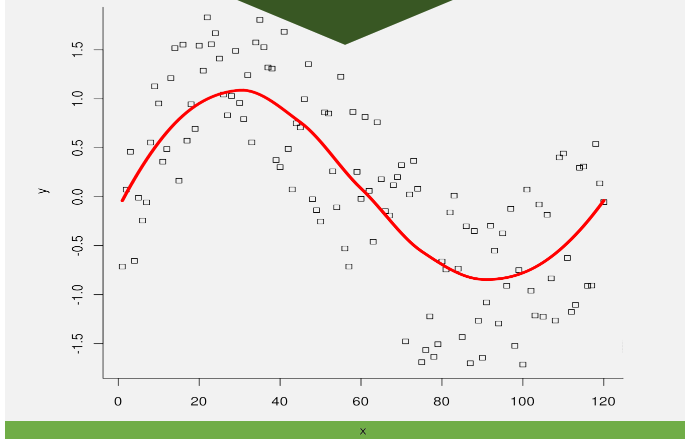
LINEAR REGRESSION

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + e$$

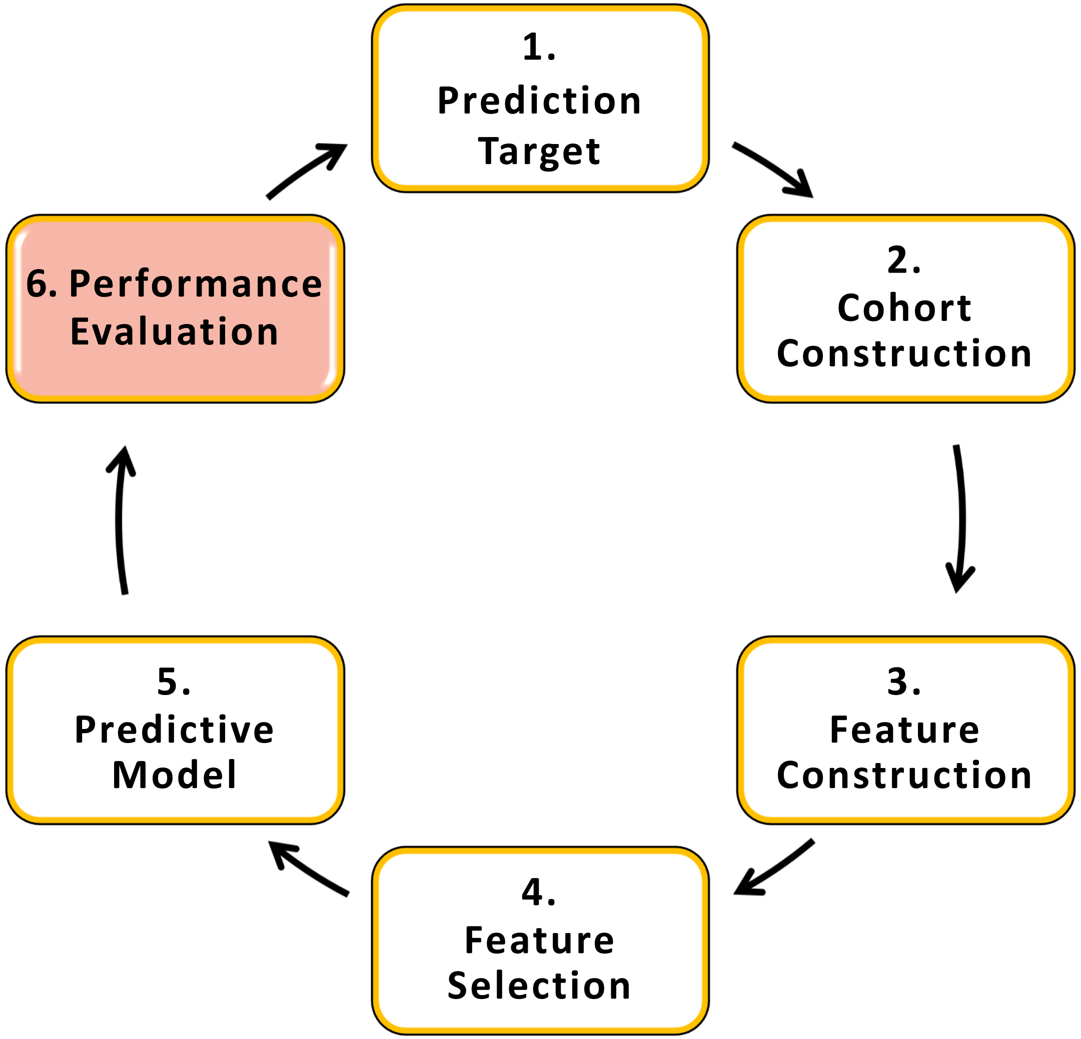


GENERALIZED ADDITIVE MODEL

$$y = \beta_0 + \sum_{i=1}^{n} f_i(x_i) + e$$



Predictive Modeling Pipeline



1. Prediction Target

2. Cohort Construction

3. Feature Construction

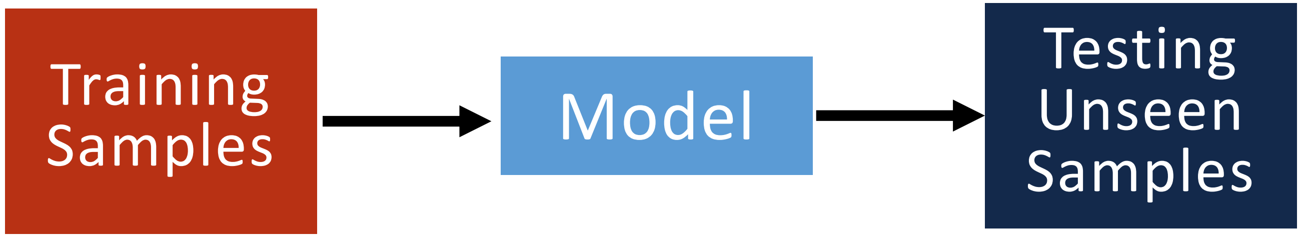
4. Feature Selection

5. Predictive Model

6. Performance Evaluation

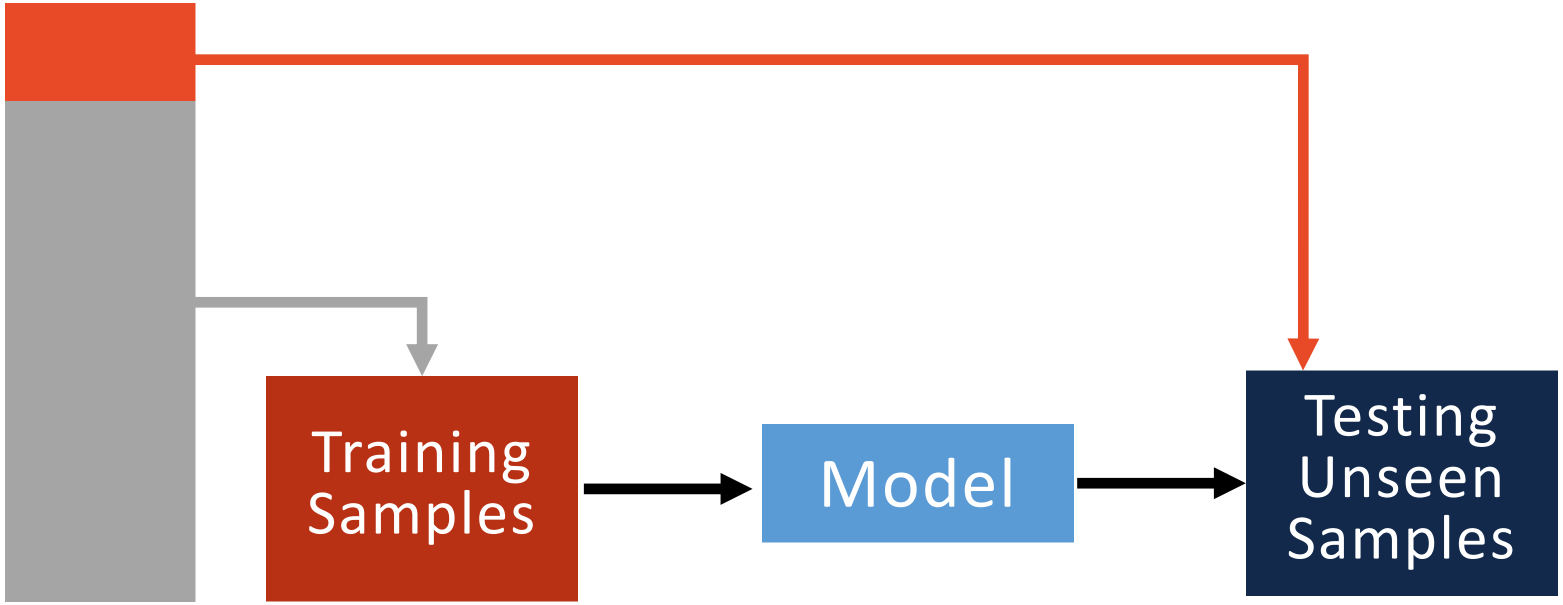
EVALUATION

- Training error is NOT very useful
- Testing error is the key metric
- Approach:
  - Cross-validation (CV)

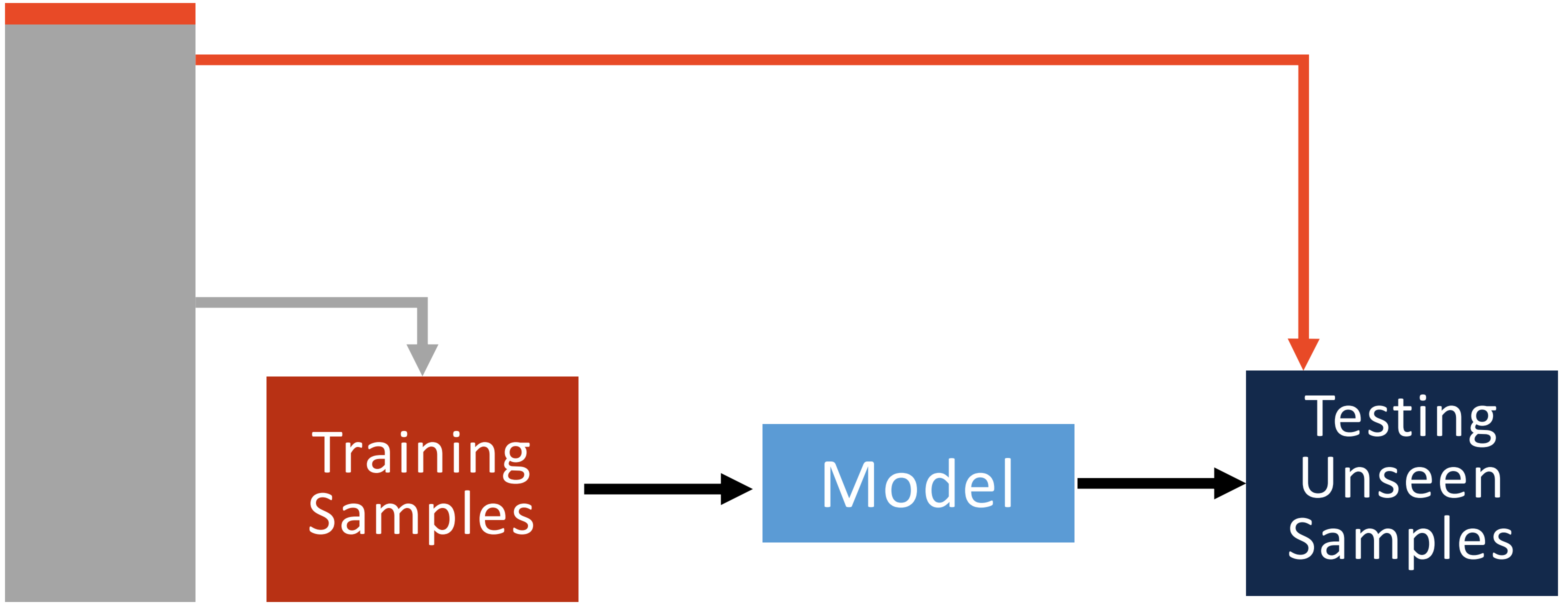


Cross-validation (Cv)

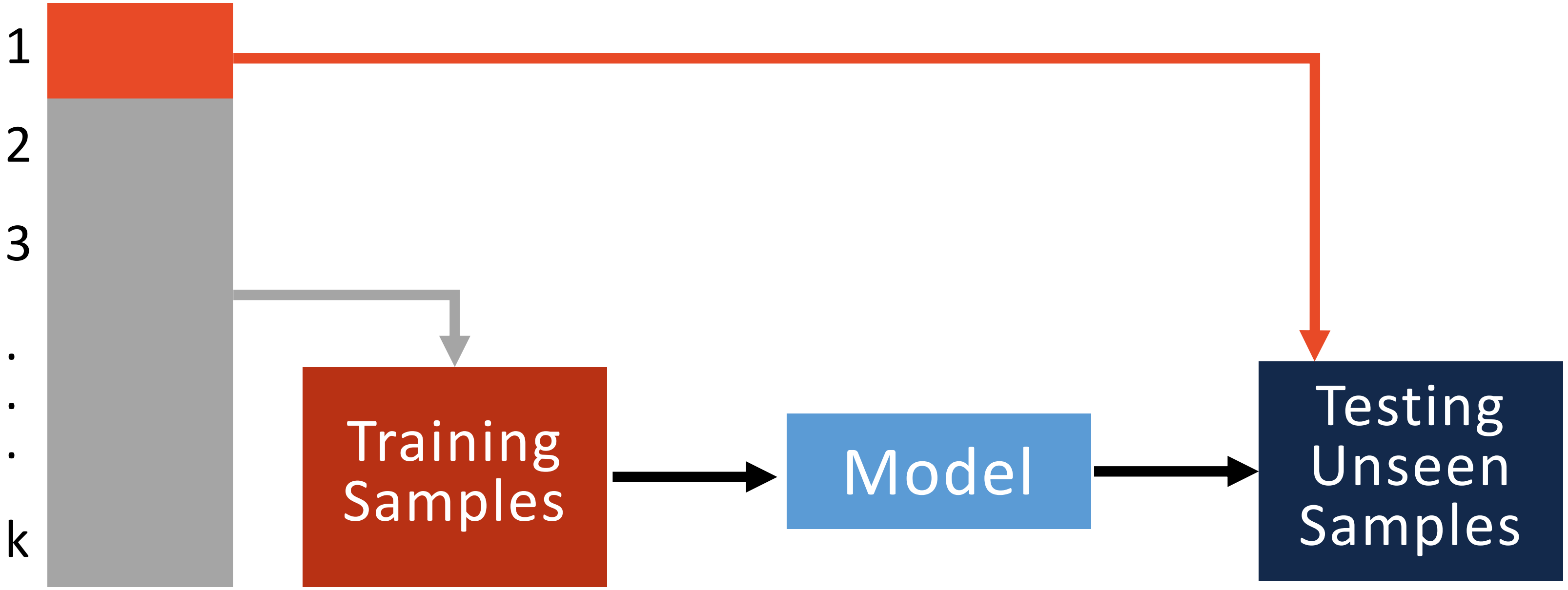
- Leave-1-out CV
- K-fold CV
- Randomized CV



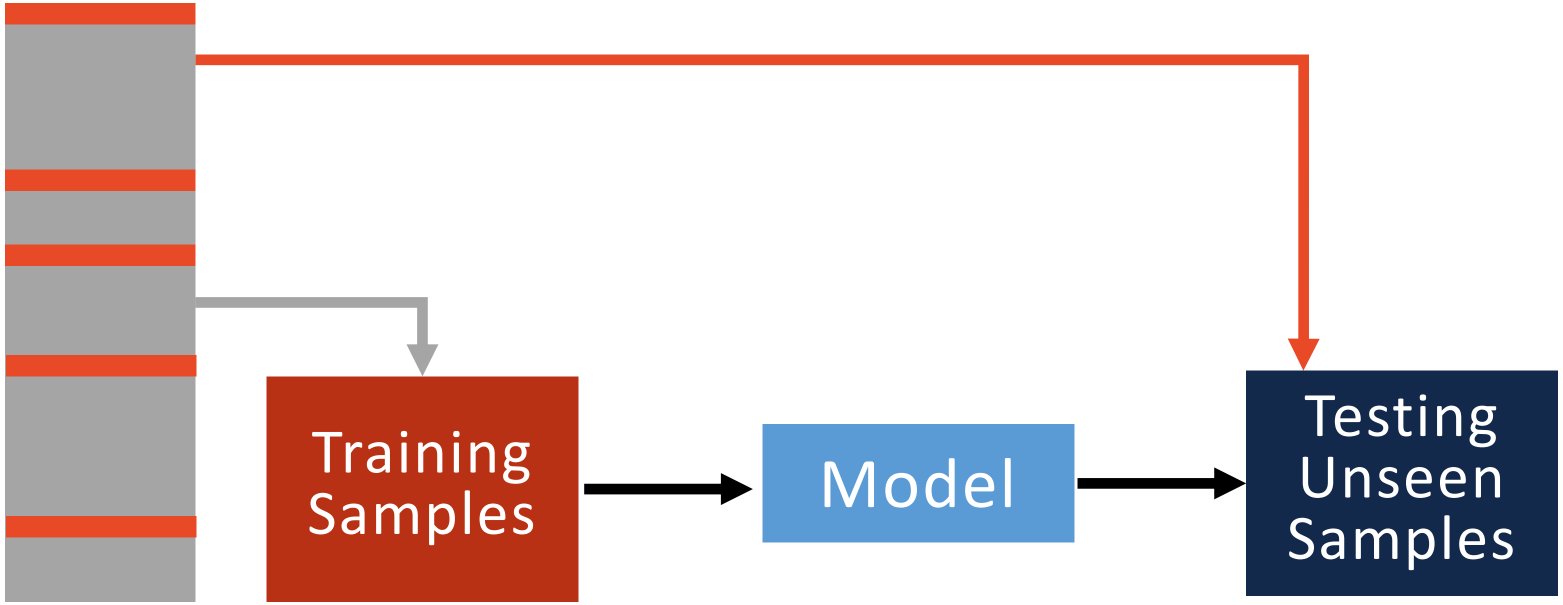
Leave-1-out Cv



K-fold Cv

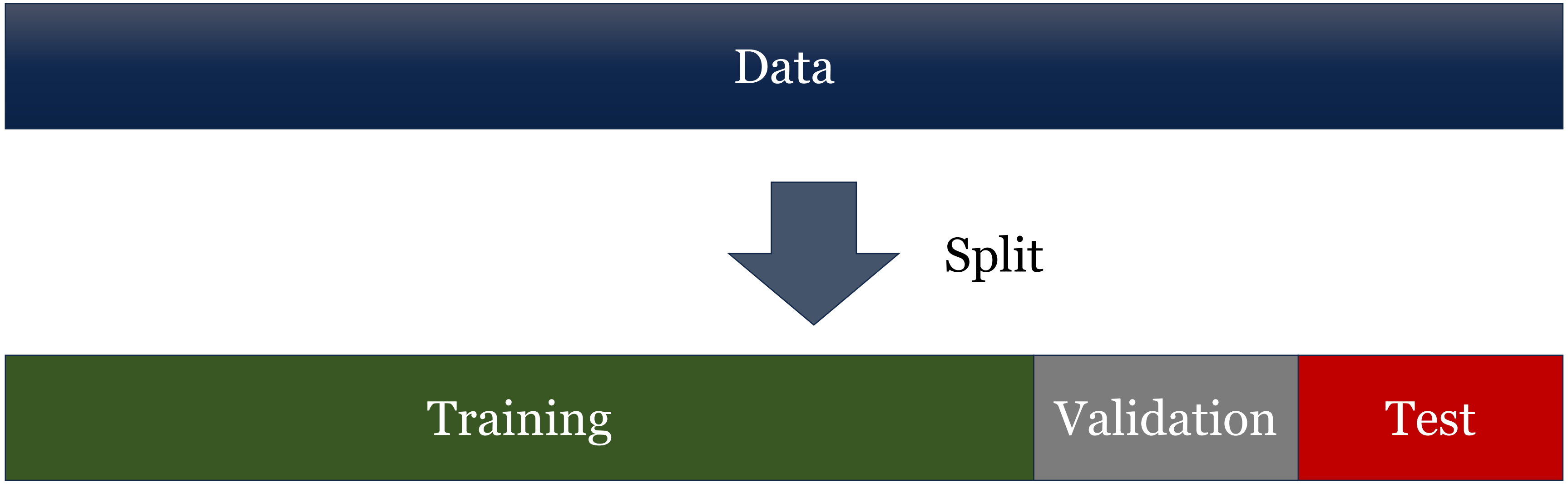


Randomized Cv



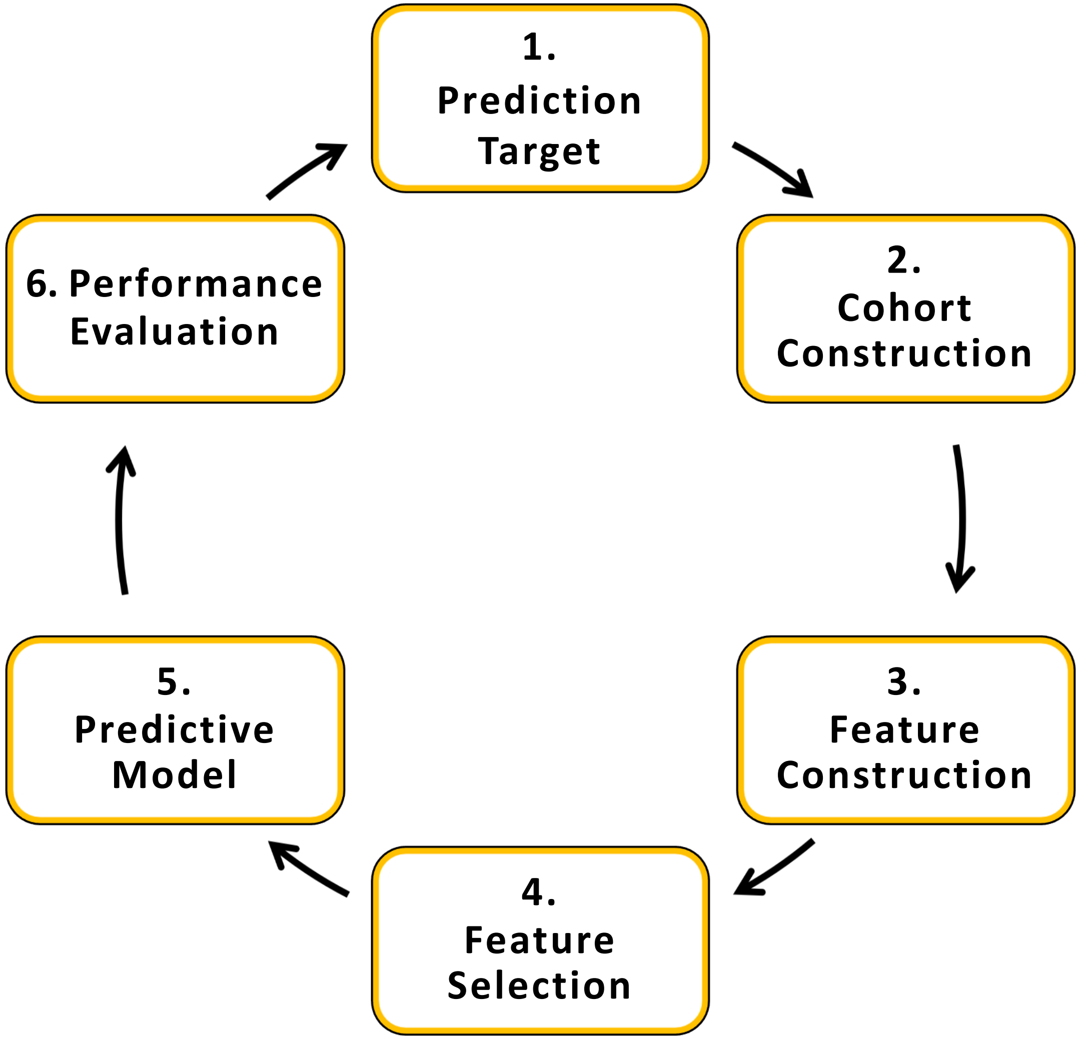
ILLINOIS

Best practice of predictive modeling evaluation



- Validation and test sets can be small,
  - but should be similar to each other
- Training data can be flexible
  - High volume is preferred
  - Some low quality data can be allowed

Predictive Modeling Pipeline


1. Prediction Target
2. Cohort Construction
3. Feature Construction
4. Feature Selection
5. Predictive Model
6. Performance Evaluation