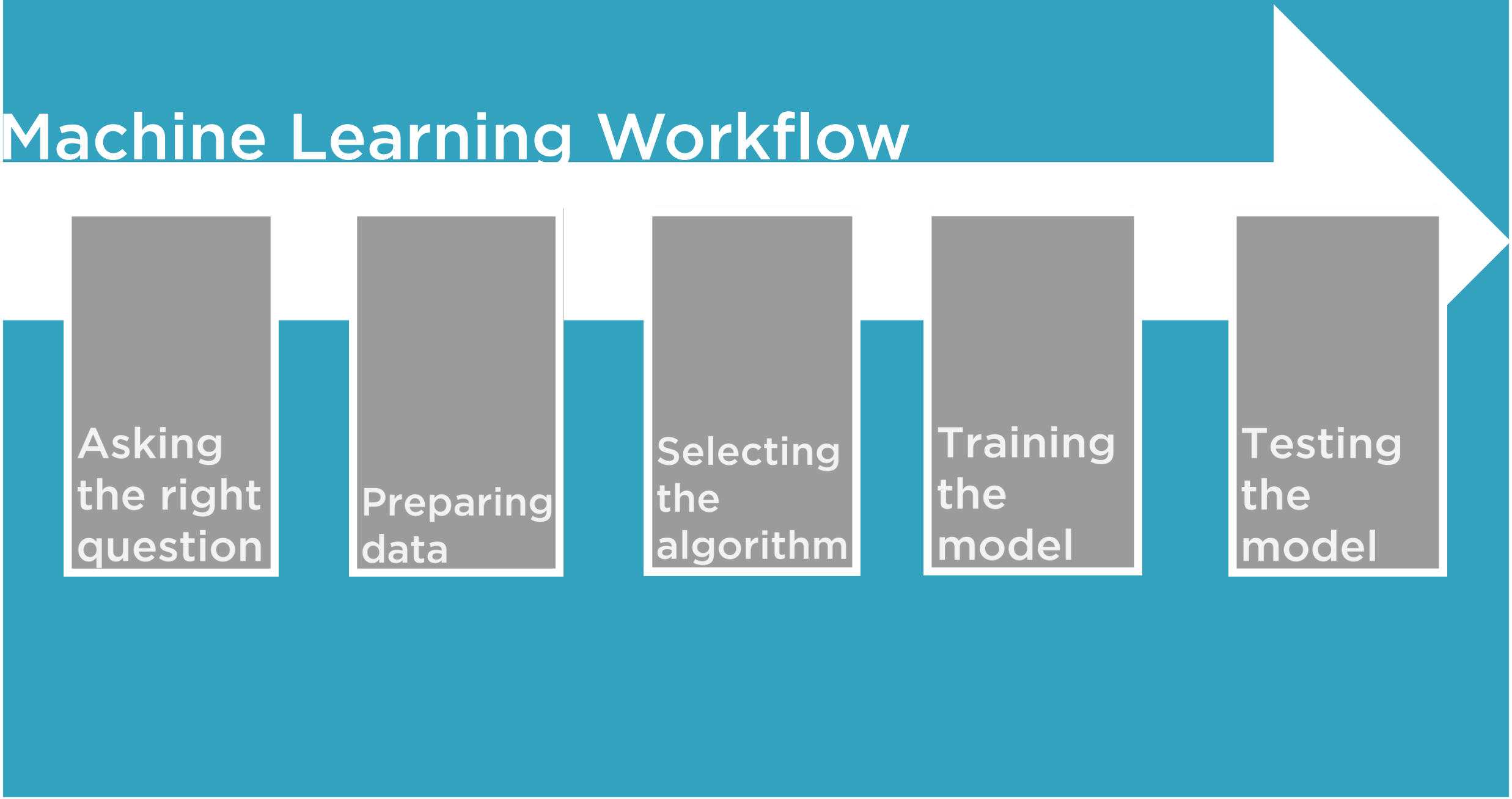
Training The Model

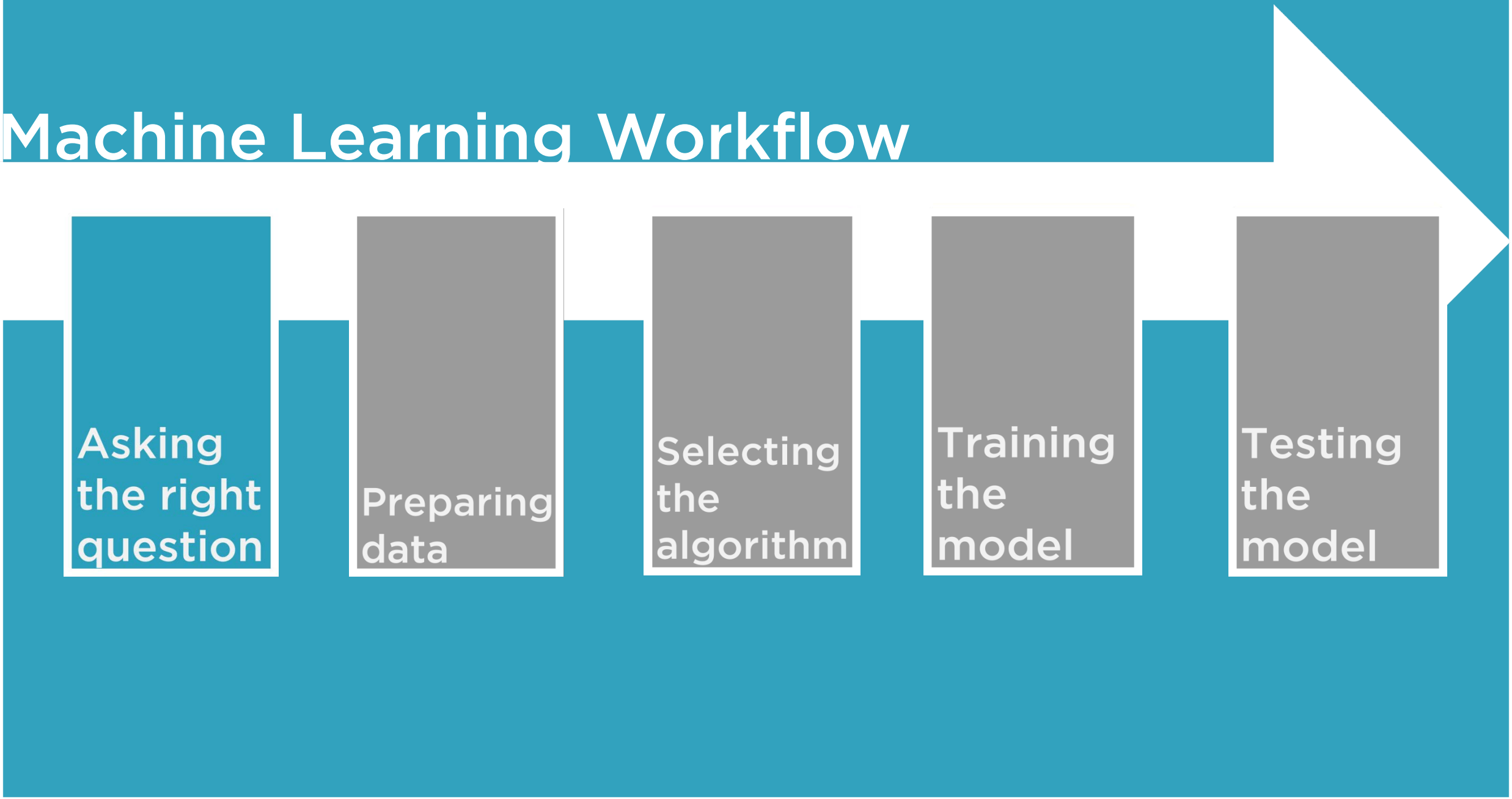
Jerry Kurata
CONSULTANT

@jerrykur www.insteptech.com

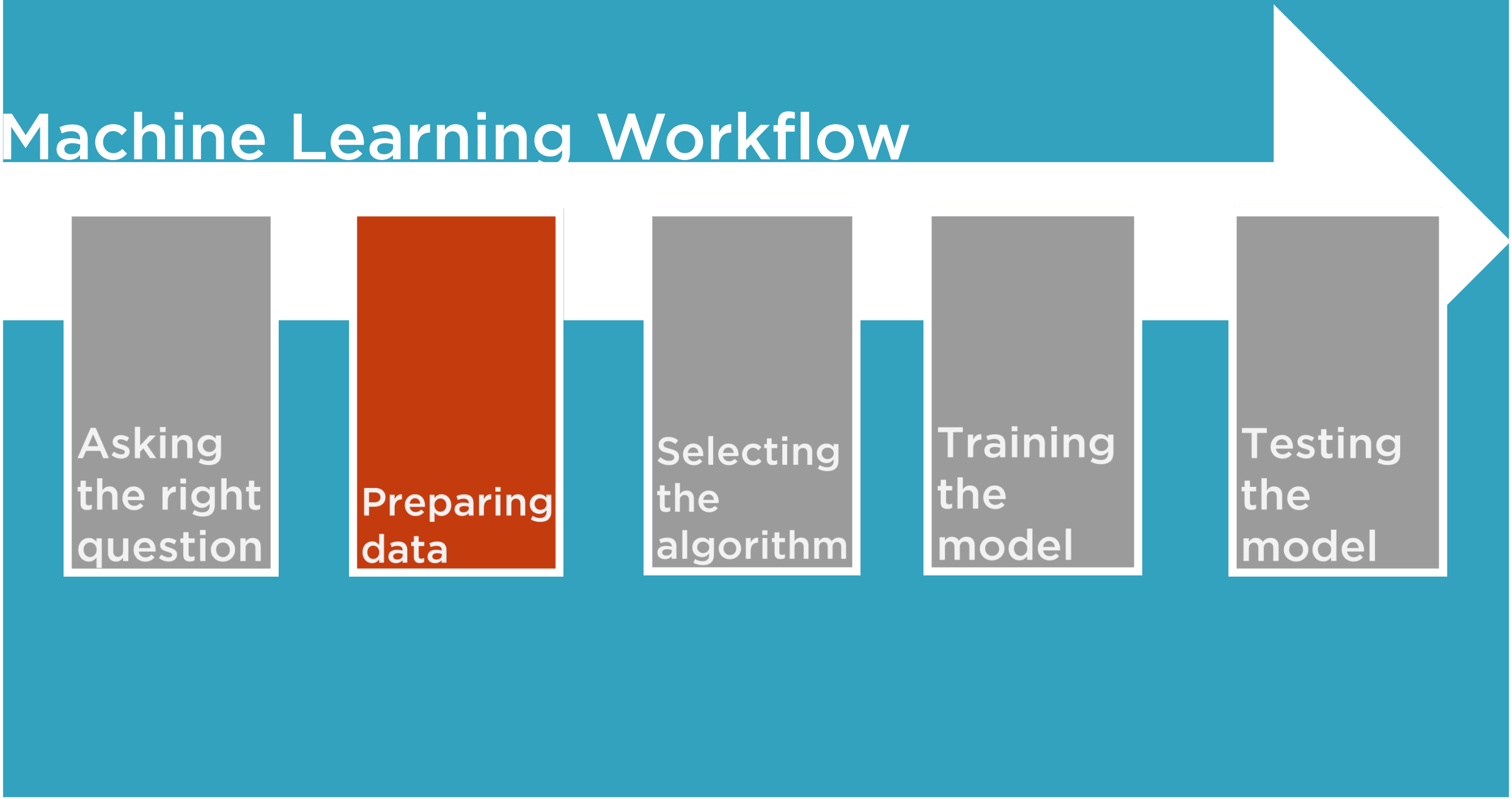
Machine Learning Workflow



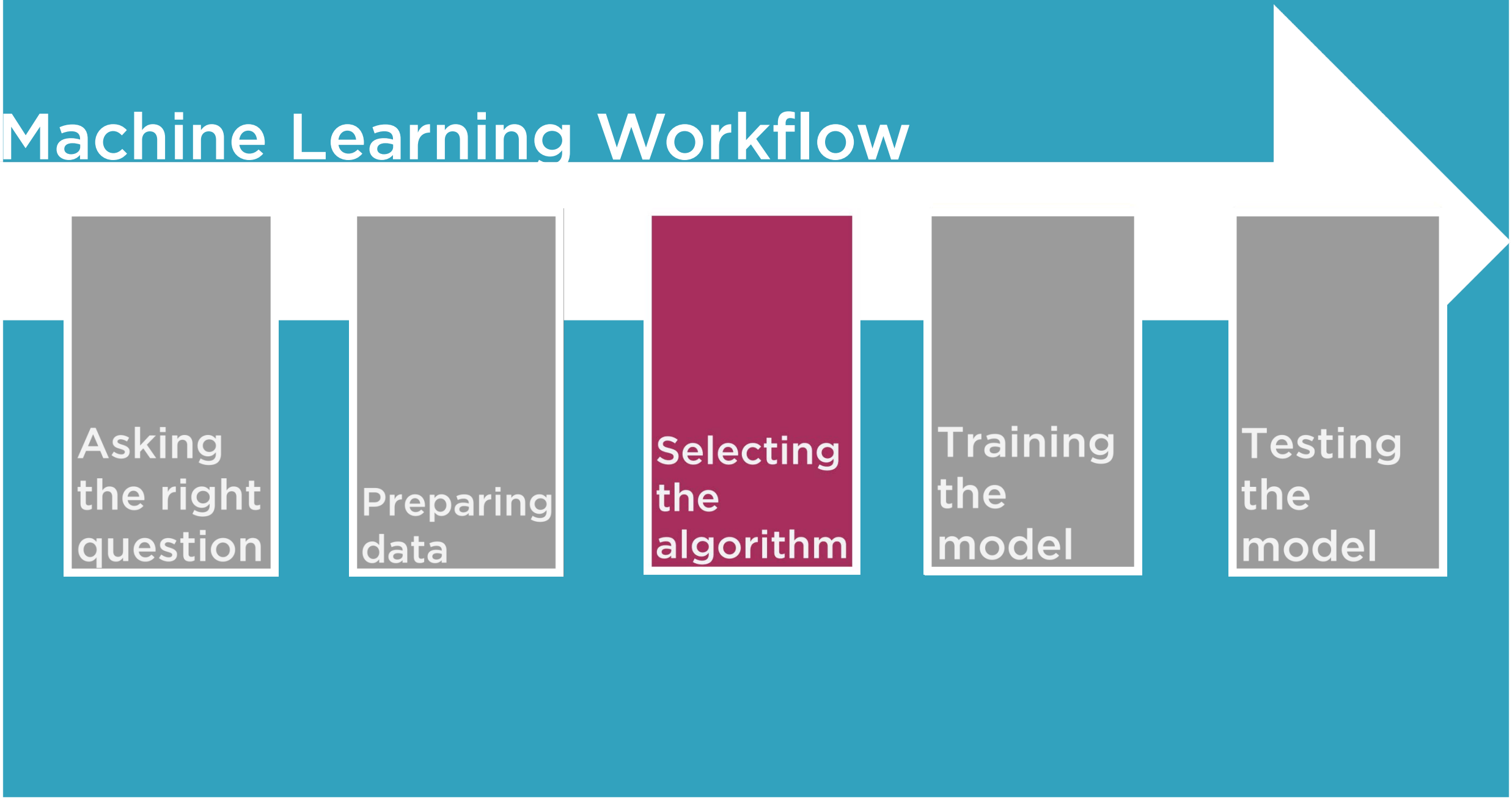
Machine Learning Workflow



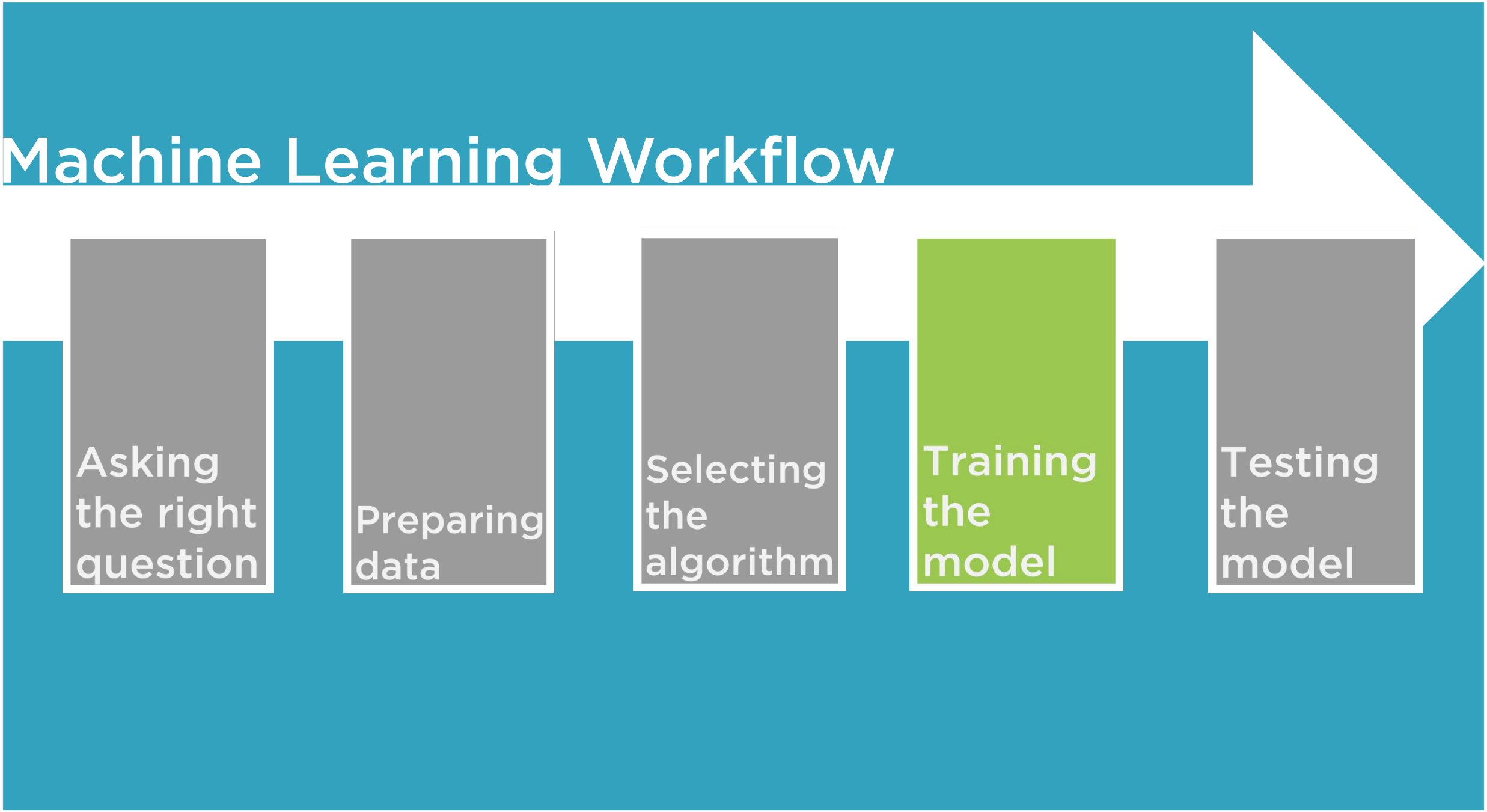
Machine Learning Workflow



Machine Learning Workflow



Machine Learning Workflow



Overview

Understand the training process

Scikit-learn package

Train algorithm with Diabetes data

Machine Learning Training

Letting specific data teach a Machine Learning algorithm to create a specific prediction model.

Machine Learning Training

Letting specific data teach a Machine Learning algorithm to create a specific prediction model.

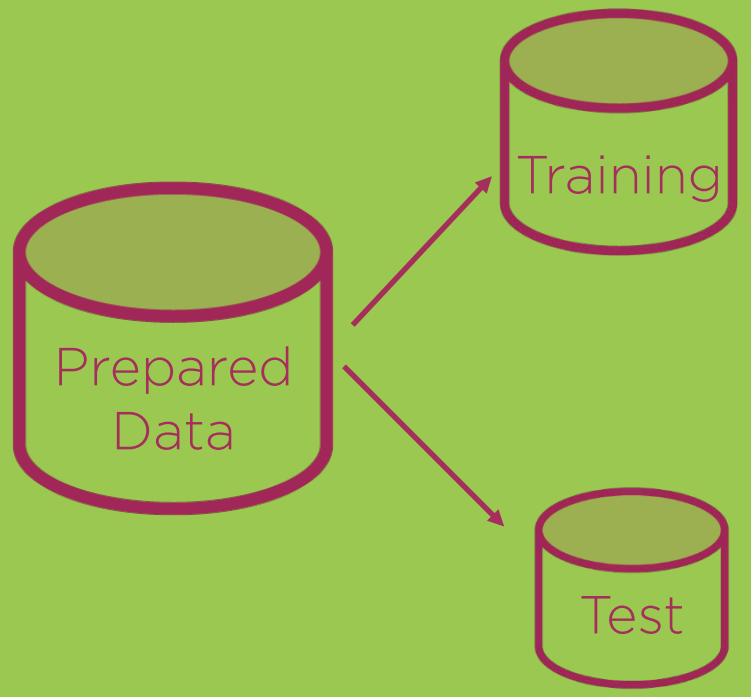
Why retrain?

New data => better predictions

Verify training performance with new data

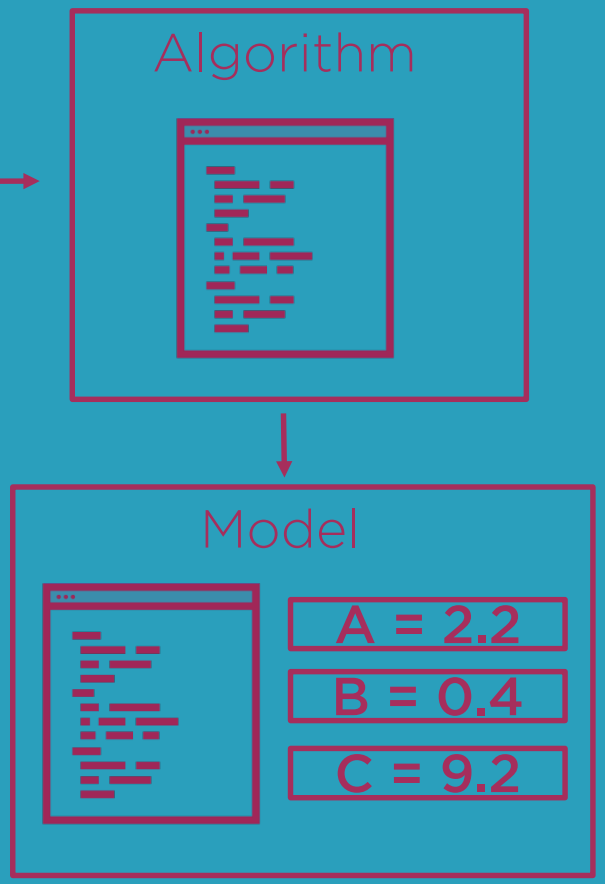
Training Overview

Split Data



70% Training
30% Testing

Train Model



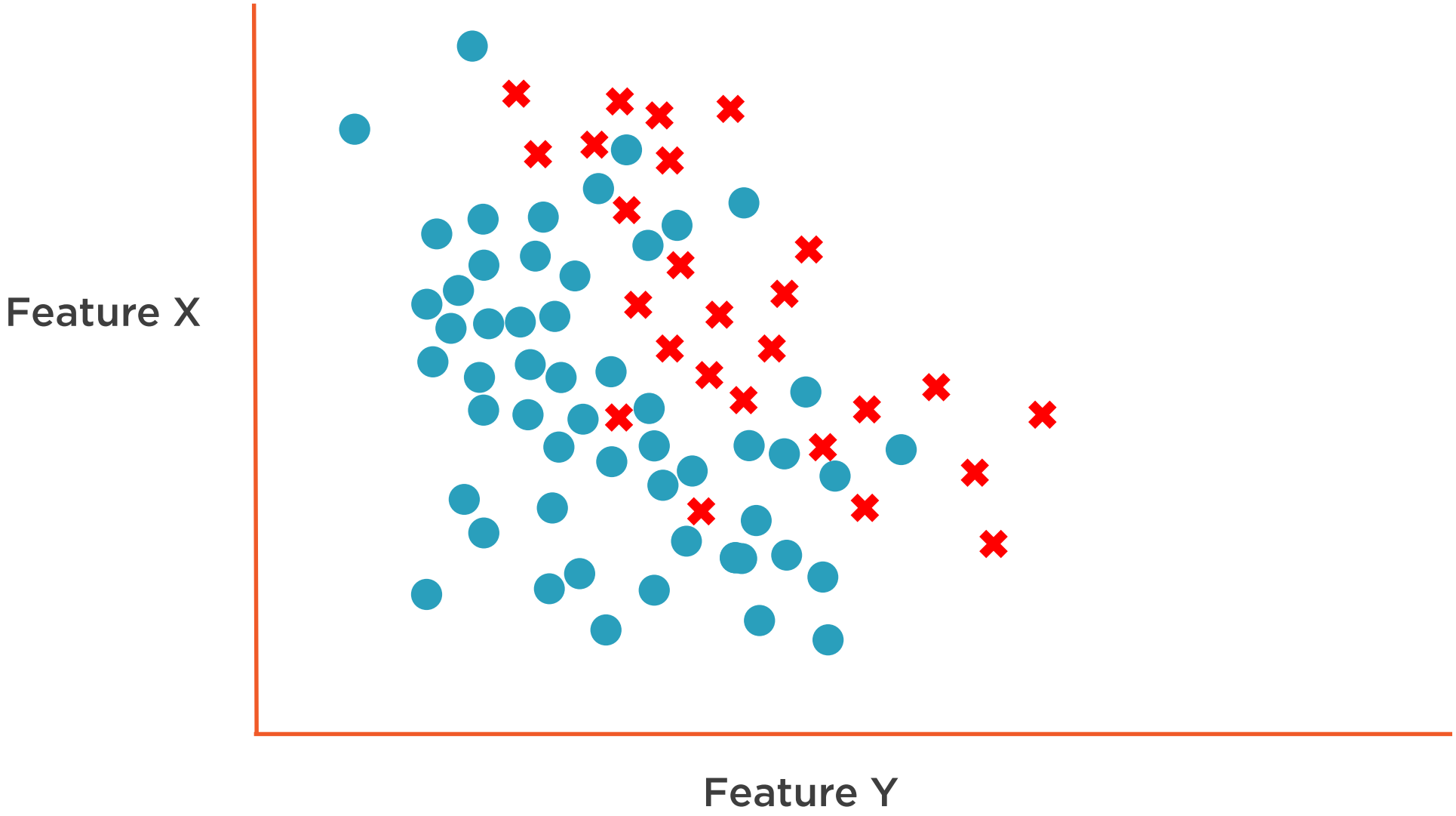
Evaluate Model

Training Goal

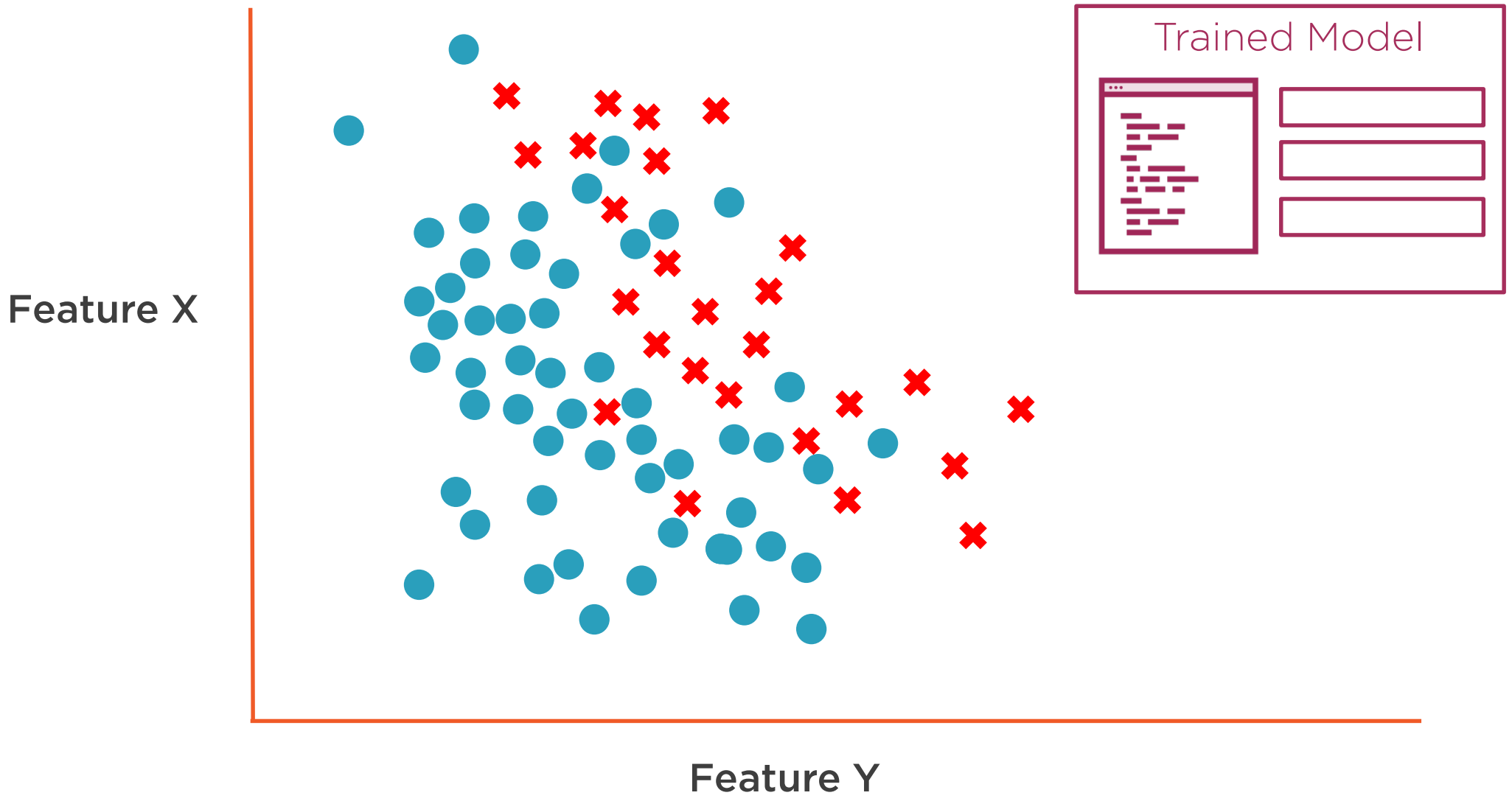
Hypothetical Data

Not Diabetes Data

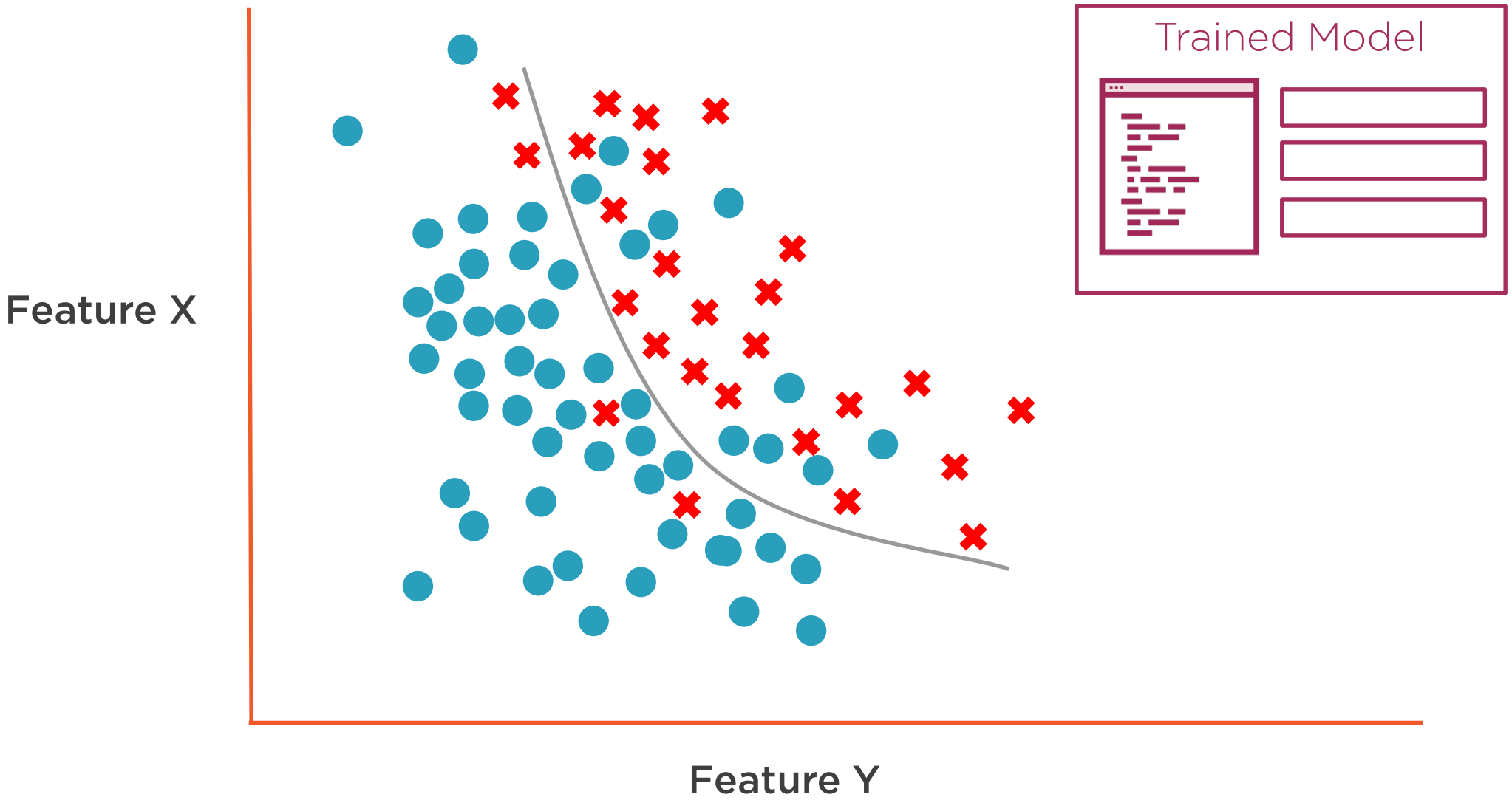
Training Goal



Training Goal



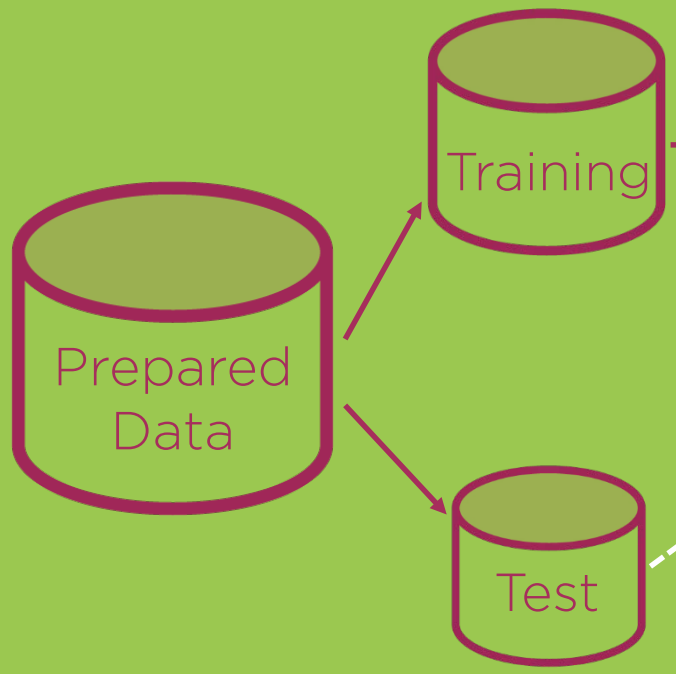
Training Goal



What about the test data?

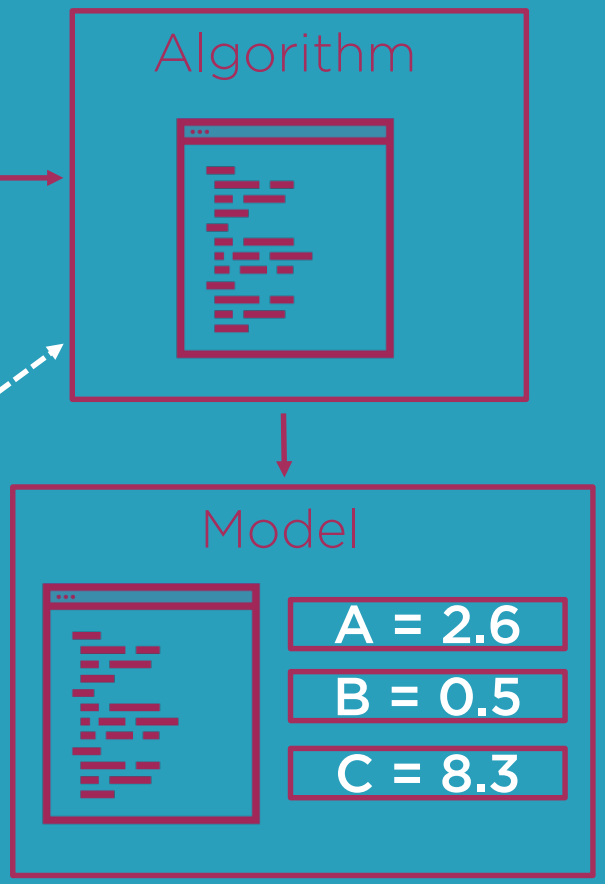
Training Overview

Split Data



70% Training
30% Testing

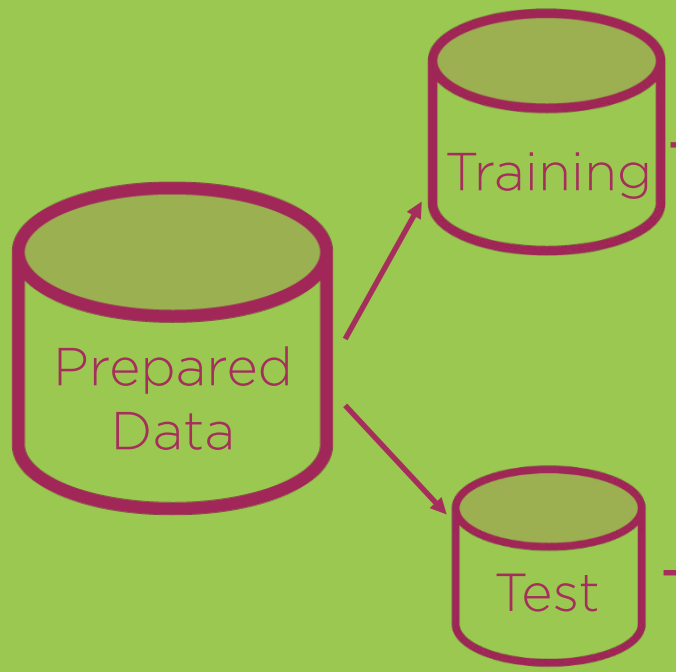
Train Model



Real World Model Performance

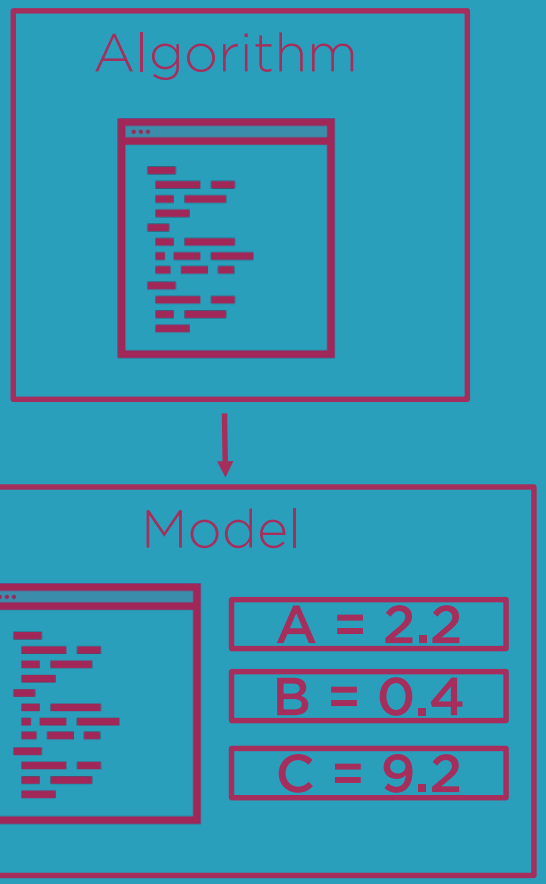
Training Overview

Split Data



70% Training
30% Testing

Train Model



Evaluate Model

Selecting Training Features

We want minimum features (columns)

Selected features

- # of Pregnancies
- Glucose Concentration
- Blood Pressure
- Skin Thickness
- Insulin Level
- Body Mass Index
- Diabetes Predisposition
- Age

Python Training Tip

Don’t rewrite from scratch

scikit-learn has training functions

Scikit-learn library

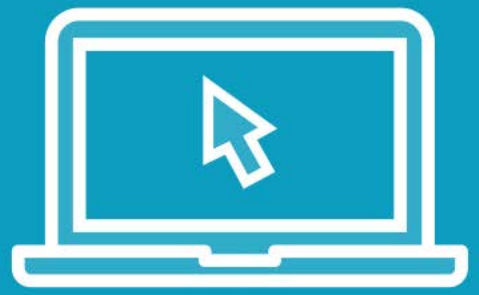
Designed to work with NumPy, SciPy and Pandas

Toolset for training and evaluation tasks

- Data splitting
- Pre-processing
- Feature selection
- Model training
- Model tuning

Common interface across algorithms

Demo



Split data into training and test data sets

Perform post split data preparation

Train with initial algorithm

Missing Data

Common Problem

Options

- Ignore
- Drop observations (rows)
- Replace values (Impute)

Data numbers

- 768 rows
- 374 missing insulin values
- Can we ignore/delete almost 50% of data?

Imputing Options

Replace with mean, median

Replace with expert knowledge derived value

Using mean imputing

Summary

Reviewed training process

Used Python to split data

- Utilized the scikit-learn methods with NumPy and Pandas data structures

Reasoned about missing data

- Used mean imputation

Trained the initial Naïve Bayes model