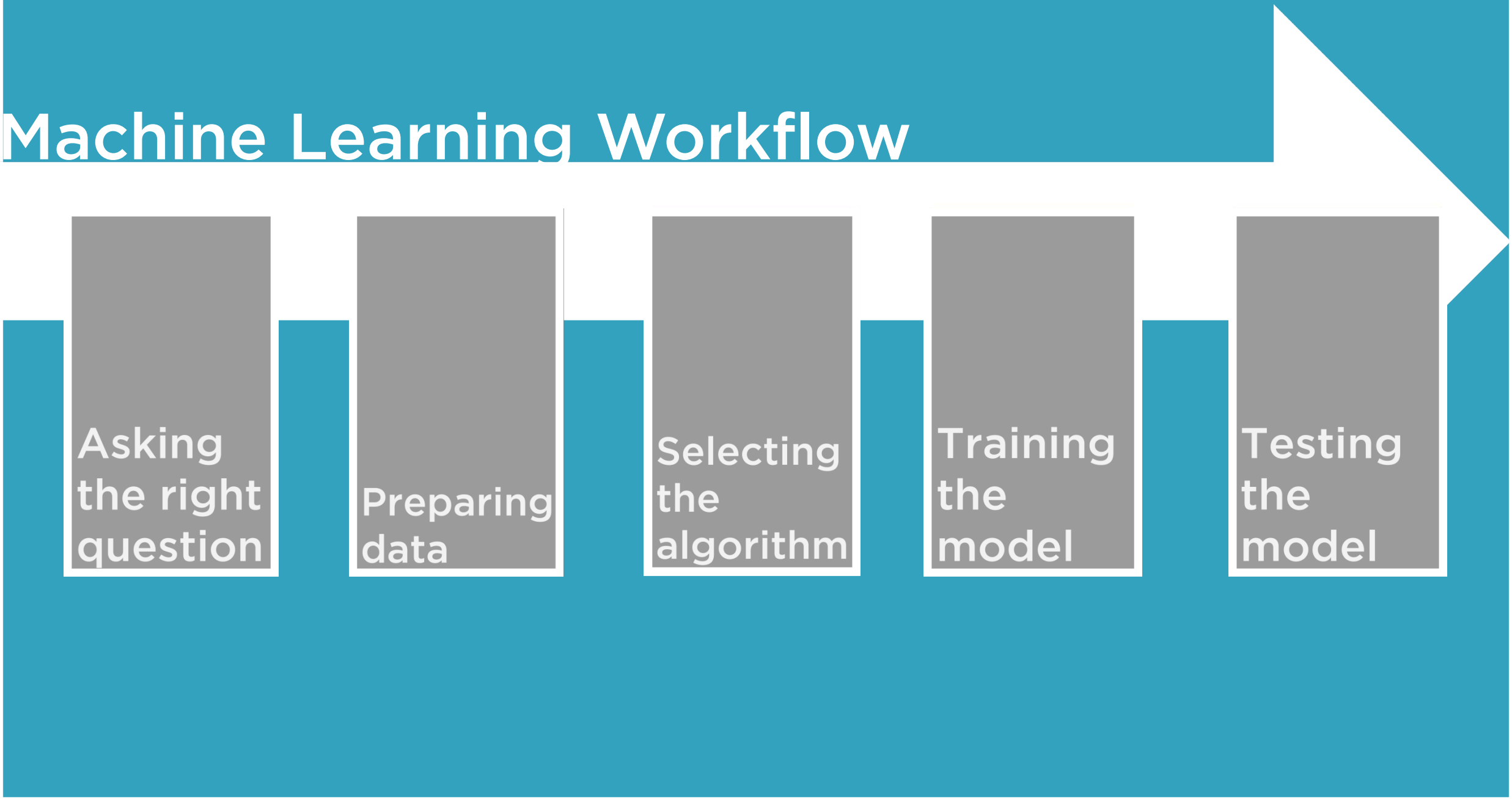
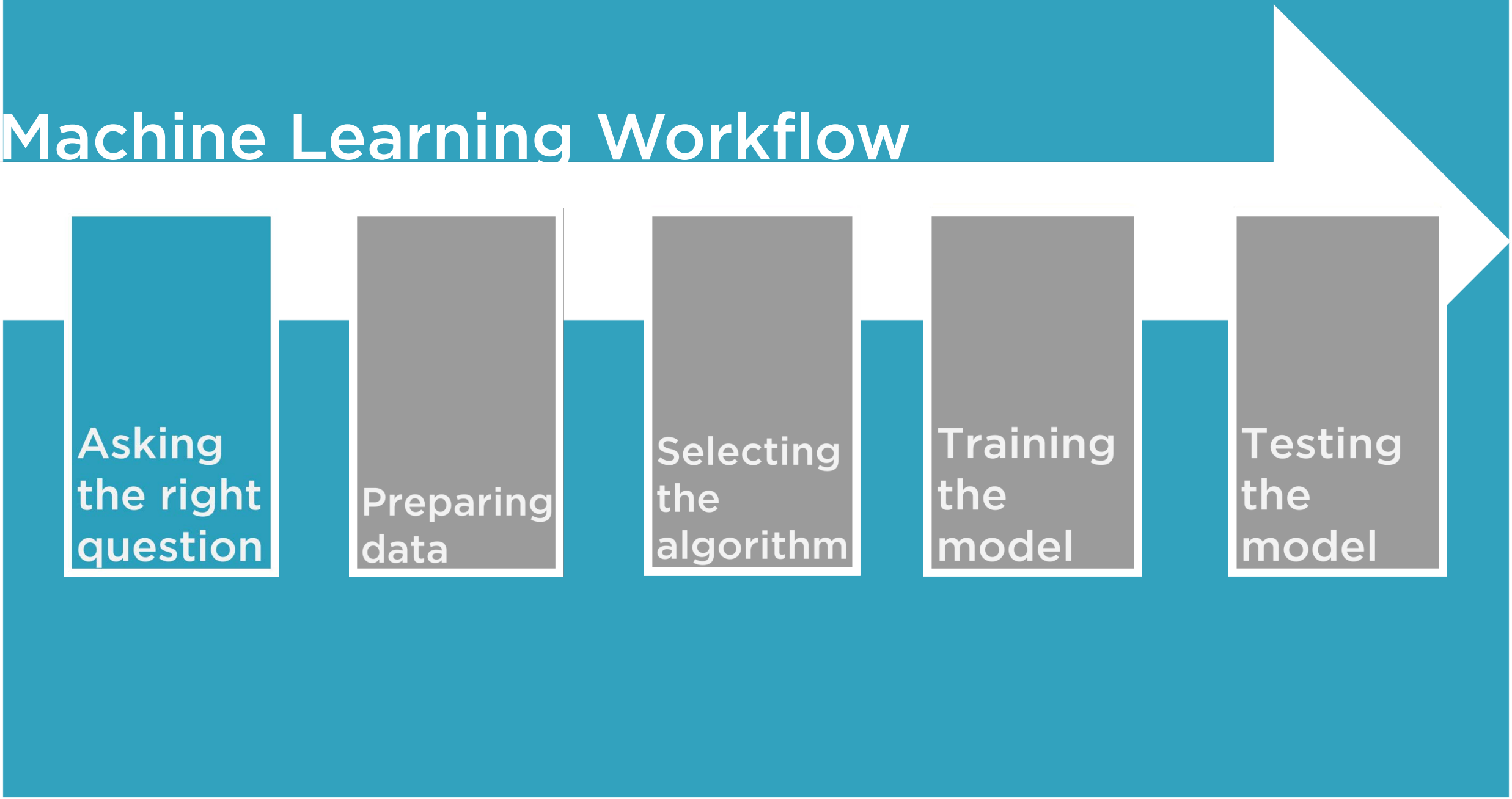
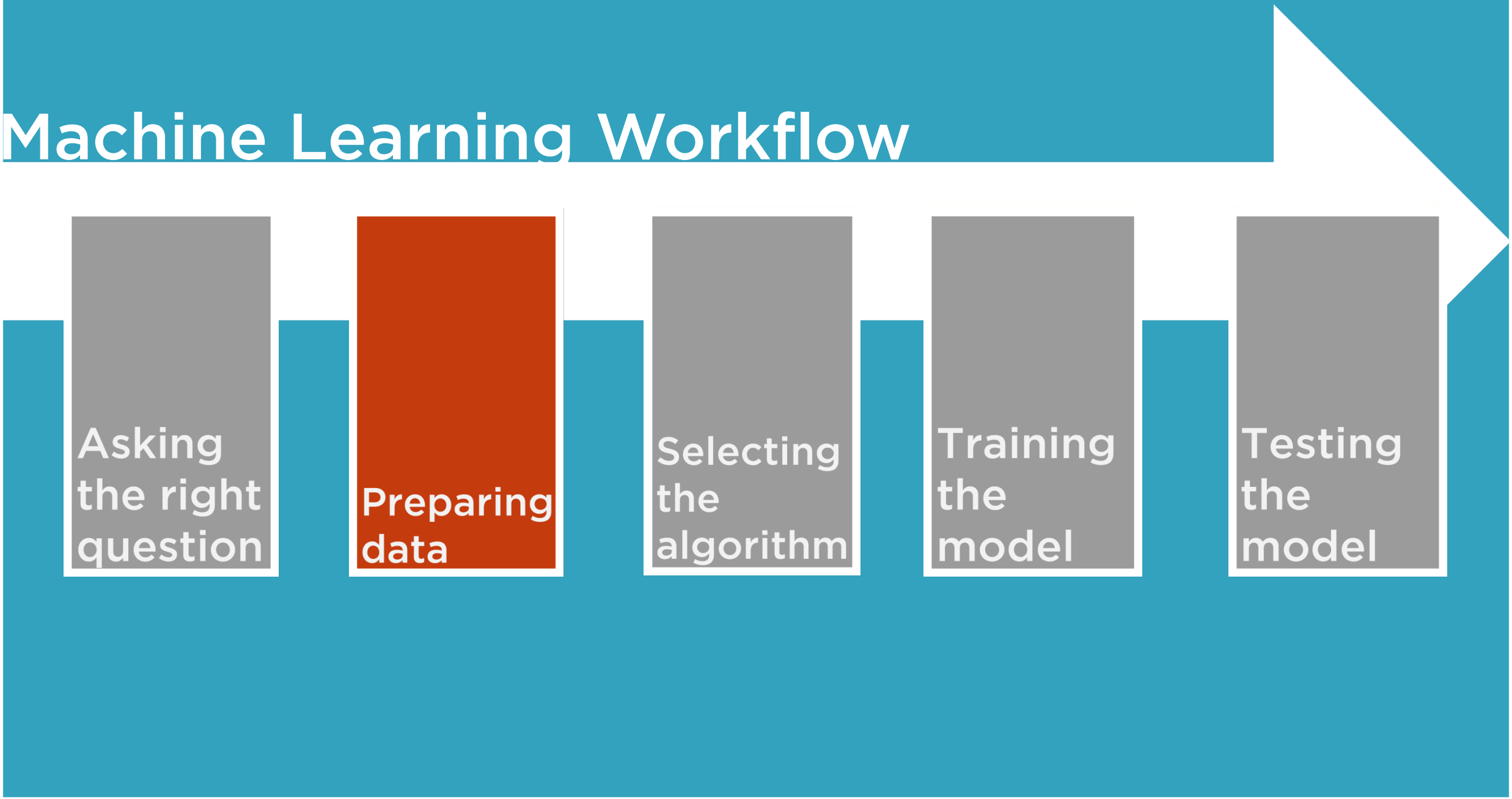
Machine Learning Workflow



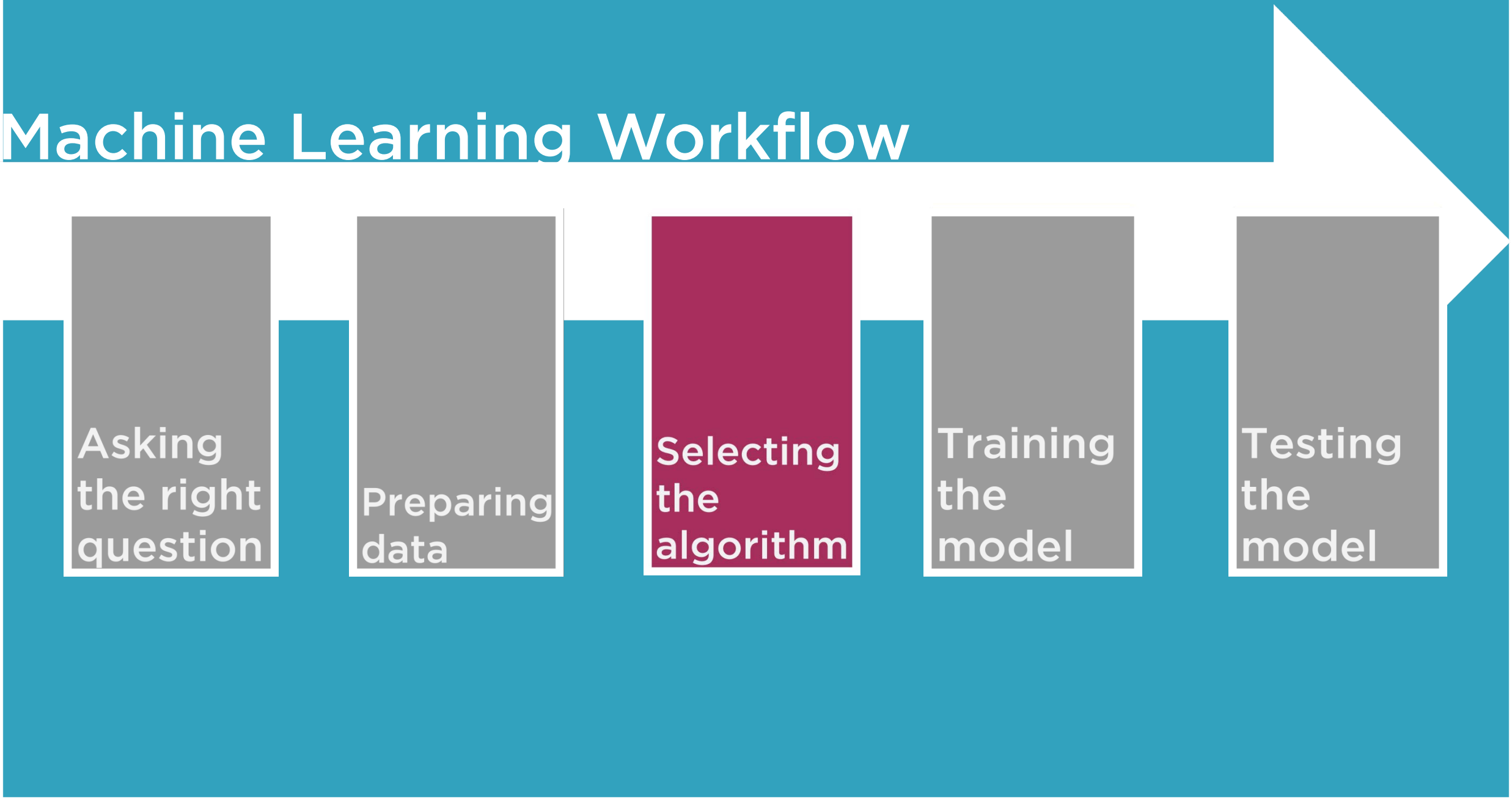
Machine Learning Workflow



Machine Learning Workflow



Machine Learning Workflow



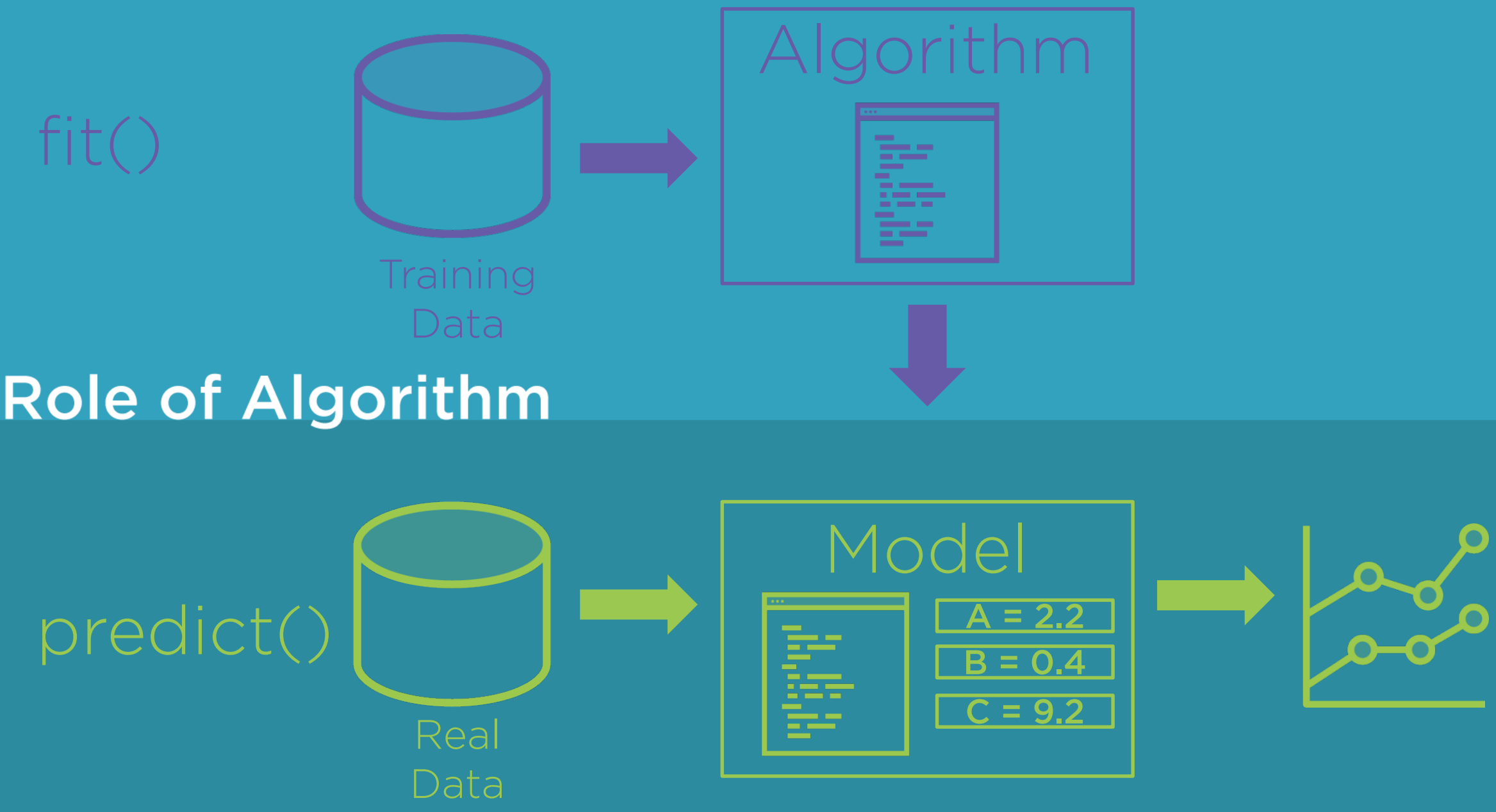
Overview

Role of algorithm

Perform algorithm selection

- Use solution statement to filter algorithms
- Discuss best algorithms
- Select one initial algorithm

Role of Algorithm



Over 50 algorithms

Algorithm Selection

Compare factors

Difference of opinions about which factors are important

You will develop your own factors

Algorithm Decision Factors

- Learning Type
- Result
- Complexity
- Basic vs enhanced

Learning Type

Learning Type

"Use the Machine Learning Workflow to process and transform Pima Indian data to create a prediction model. This model must predict which people are likely to develop diabetes with 70% or greater accuracy."

Learning Type

"Use the Machine Learning Workflow to process and transform Pima Indian data to create a prediction model. This model must predict which people are likely to develop diabetes with 70% or greater accuracy."

Prediction Model => Supervised machine learning

Over ~~50~~ 28 algorithms

Result Type

Regression

- Continuous values
- price = A * # bedroom
  + B * size
  + ...

Classification

- Discrete values
- small, medium, large
- 1-100, 101-200, 201-300
- true or false

Result Type

"... predict which people are likely to develop diabetes ..."

Result Type

"... predict which people are likely to develop diabetes ..."

Diabetes

Binary (TRUE/FALSE)

Algorithm must support classification

- Binary classification

Over ~~50~~ ~~28~~ 20 algorithms

Complexity

Keep it Simple

Eliminate “ensemble” algorithms

- Container algorithm
- Multiple child algorithms
- Boost performance
- Can be difficult to debug

Over ~~50~~ ~~28~~ ~~20~~ 14 algorithms

Enhanced vs. Basic

Enhanced

- Variation of Basic
- Performance improvements
- Additional functionality
- More Complex

Basic

- Simpler
- Easier to understand

Candidate Algorithms

Naive Bayes

Logistic Regression

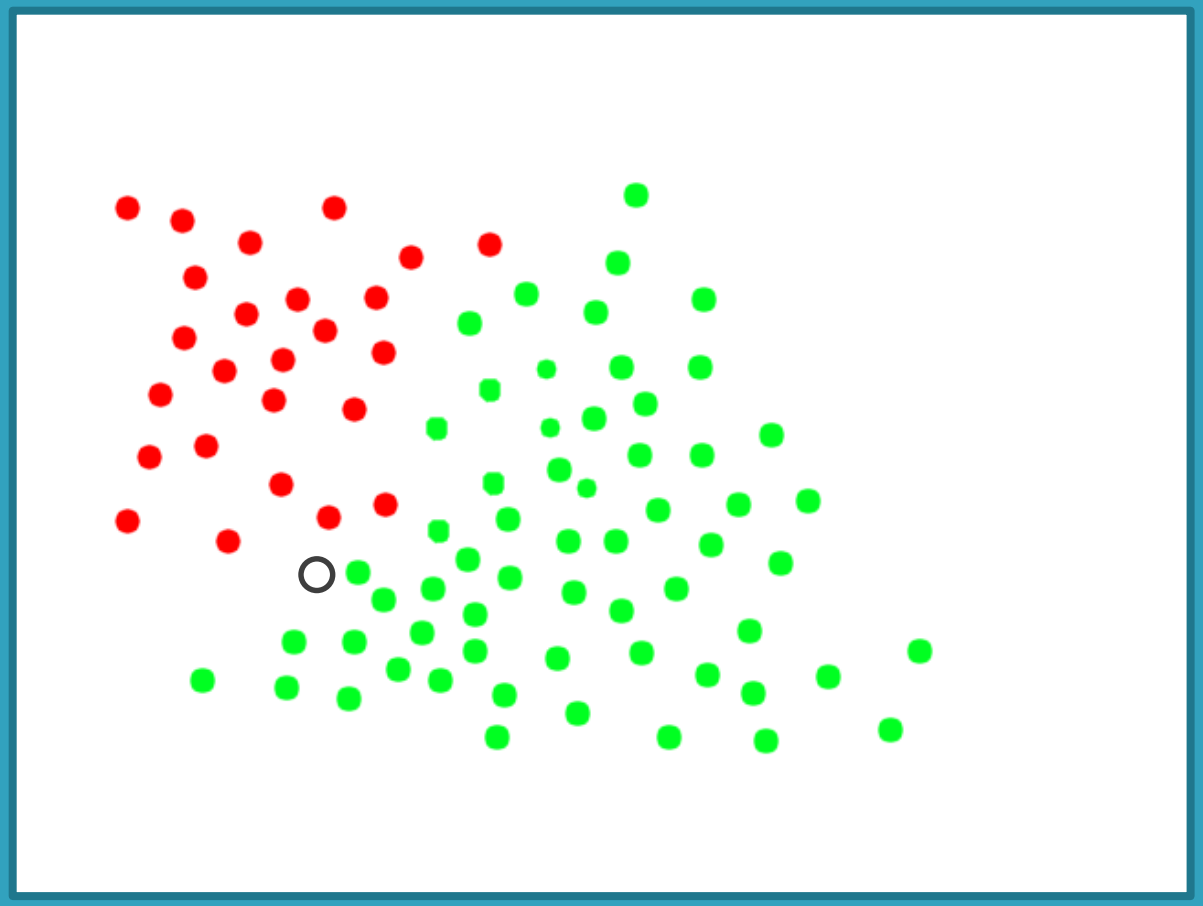
Decision Tree

Naive Bayes

Based on likelihood and probability

Every feature has the same weight

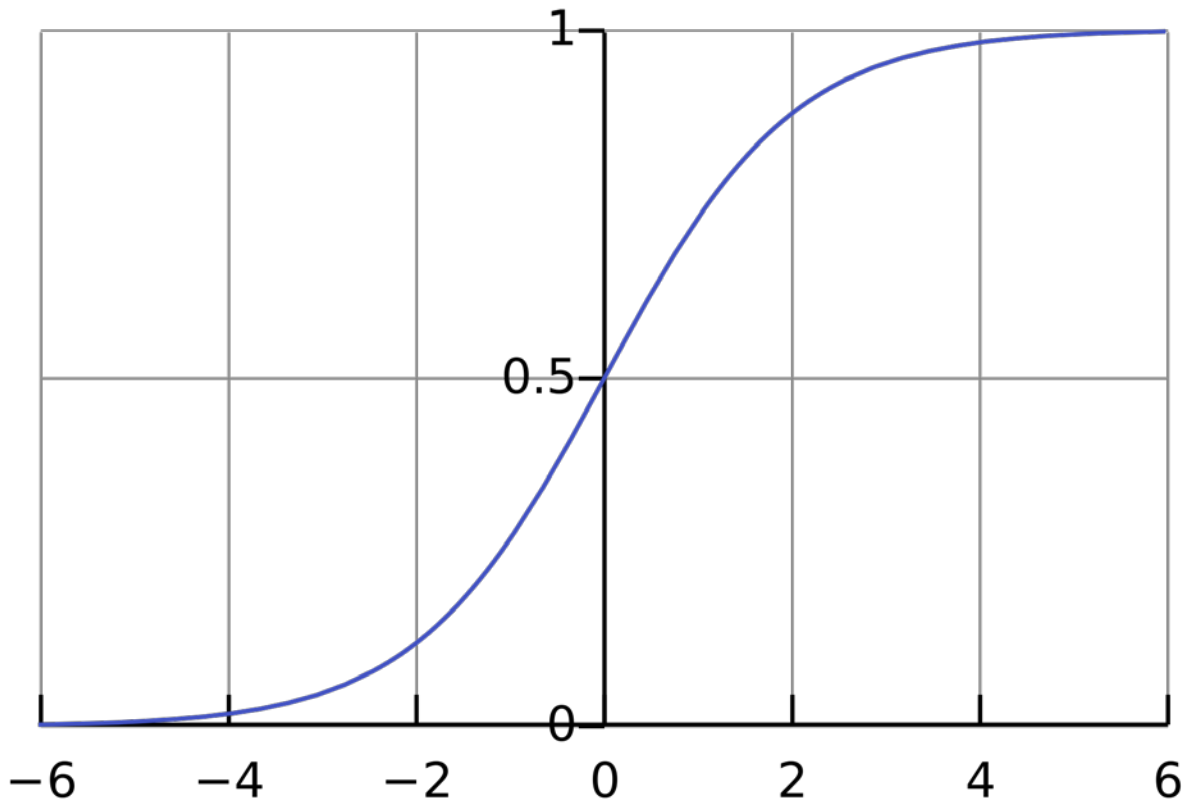
Requires smaller amount of data



Logistic Regression

Confusing name, binary result

Relationship between features are weighted

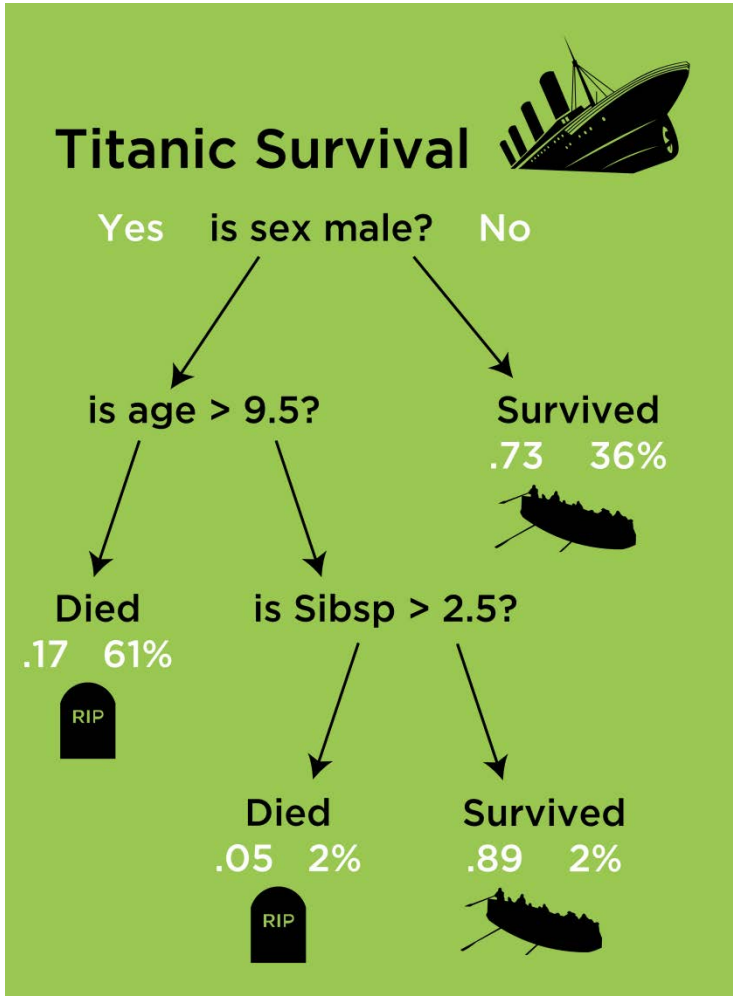


Decision Tree

Binary Tree

Node contains decision

Requires enough data to determine nodes and splits



Selected Algorithm

Naïve Bayes

Simple - easy to understand

Fast – up to 100X faster

Stable to data changes

Summary

Lots of algorithms available

Selection based on

- Learning = Supervised
- Result = Binary classification
- Non-ensemble
- Basic

Naïve Bayes selected for training

- Simple, fast, and stable