PSY9511: Seminar 1

Introduction to machine learning

Esten H. Leonardsen

August 18, 2025

UNIVERSITY
OF OSLO

Outline

Plan for the day

- Round of introductions
- Course information
- Introduction to machine learning
- Presentation of assignment 1

Teacher

Esten Høyland Leonardsen

- Master's degree in Informatics: Programming and Networks.
- Experience as a data scientist and programmer from the industry and various start-ups.
- PhD in Psychology, deep learning applied to neuroimaging data.
- Post-doc at the section for Cognitive psychology, Neuroscience and Neuropsychology.
- Chief Scientific Officer at baba.vision.
- Interests: Deep learning, explainable artificial intelligence, mental health, neuroimaging.

Students

What I want to know about you

- What's your name?
- What institution/faculty/department/section are you from?
- What's your research project about?
- Do you have experience with machine learning and/or programming?
- What do you hope to learn from this course? (e.g. specific applications in your research, a theoretical understanding of machine learning, following and contributing to the public discourse, a future job in data science, ...).

About the course

Canvas

- All relevant announcements will be made on Canvas (e.g. changes to assignments, lectures, interesting reading material etc.).
- Lecture slides and notebooks from live coding will be put on Canvas before/after a lecture.

About the course

Curriculum

- The course uses the book ”An Introduction to Statistical Learning”, available as a PDF online.
  - Only some chapters will be used, they are posted on Canvas under each Lecture module.
  - Although we won’t be relying much on the exercises I **highly recommend** that you look into them yourselves.
- I will add some scientific publications to the curriculum list as we go, depending on your preferences and interests.

About the course

Assignments

- The course has no final exam, but six mandatory assignments you need to pass.
  - Mostly practical coding, with some reflection.
  - Given with a **hard** deadline, unless there is a good reason for an extension.
  - Can be delivered multiple times based on feedback (but the first must be in time for the original deadline).
- Exercises 1-4 and 6 are mostly small and related to specific content of the preceding lecture, while 5 is a bit larger.
- **You should hand in runnable code (e.g. a Jupyter notebook, a python script, an R script, Rmarkdown etc.), not code copied into a Word document or a pdf.**

About the course

Generative artificial intelligence (e.g. ChatGPT)

- You are allowed (and even encouraged) to use generative AI in the assignments, but it is helpful for me when I correct the exercises to know when and where you have used it.
- Be critical, you should be able to understand and explain **all** the code you hand in.
- It is your responsibility, both in this course and elsewhere, to ensure you don't break any rules with regards to privacy or copyright. If you are uncertain, GPT UiO is a relatively safe alternative.

About the course

- No mandatory attendance, but I recommend attending.
  - Exercise solutions will be (partially) coded at the beginning of the lectures.
- Goal is to demonstrate central parts of the underlying theory in an intuitive manner.
- ∼2 hours of lecturing, ∼1 hour for individual work/help with assignments.
  - I will attempt to make lectures interactive, and do live coding where possible.

About the course

Course plan

1. Introduction to machine learning
2. Basics of regression and classification
3. Variable selection and regularization
4. Model selection, validation, and testing
5. Non-linearity: Splines and tree-based methods
6. Unsupervised learning
7. Deep learning and image analysis
8. Language processing

Introduction to machine learning

UNIVERSITY
OF OSLO

Introduction

Key terminology:

- Statistical learning: A set of tools (often called models) for finding patterns in data.

Introduction

Key terminology:

- Statistical learning: A set of tools (often called models) for finding patterns in data.
- Machine learning: Approximately the same as statistical learning.
  - More common among practitioners with a computer science or engineering background.
  - Often has a focus on prediction (as opposed to understanding).
  - Often uses very large datasets.
  - More pragmatic in nature(?).

Introduction

Key terminology:

- Statistical learning: A set of tools (often called models) for finding patterns in data.
- Machine learning: Approximately the same as statistical learning.
  - More common among practitioners with a computer science or engineering background.
  - Often has a focus on prediction (as opposed to understanding).
  - Often uses very large datasets.
  - More pragmatic in nature(?).
- Supervised learning: We know what task we want the model to solve.
- Unsupervised learning: We don't know what task we want the model to solve (or we don't have the data needed to solve it).

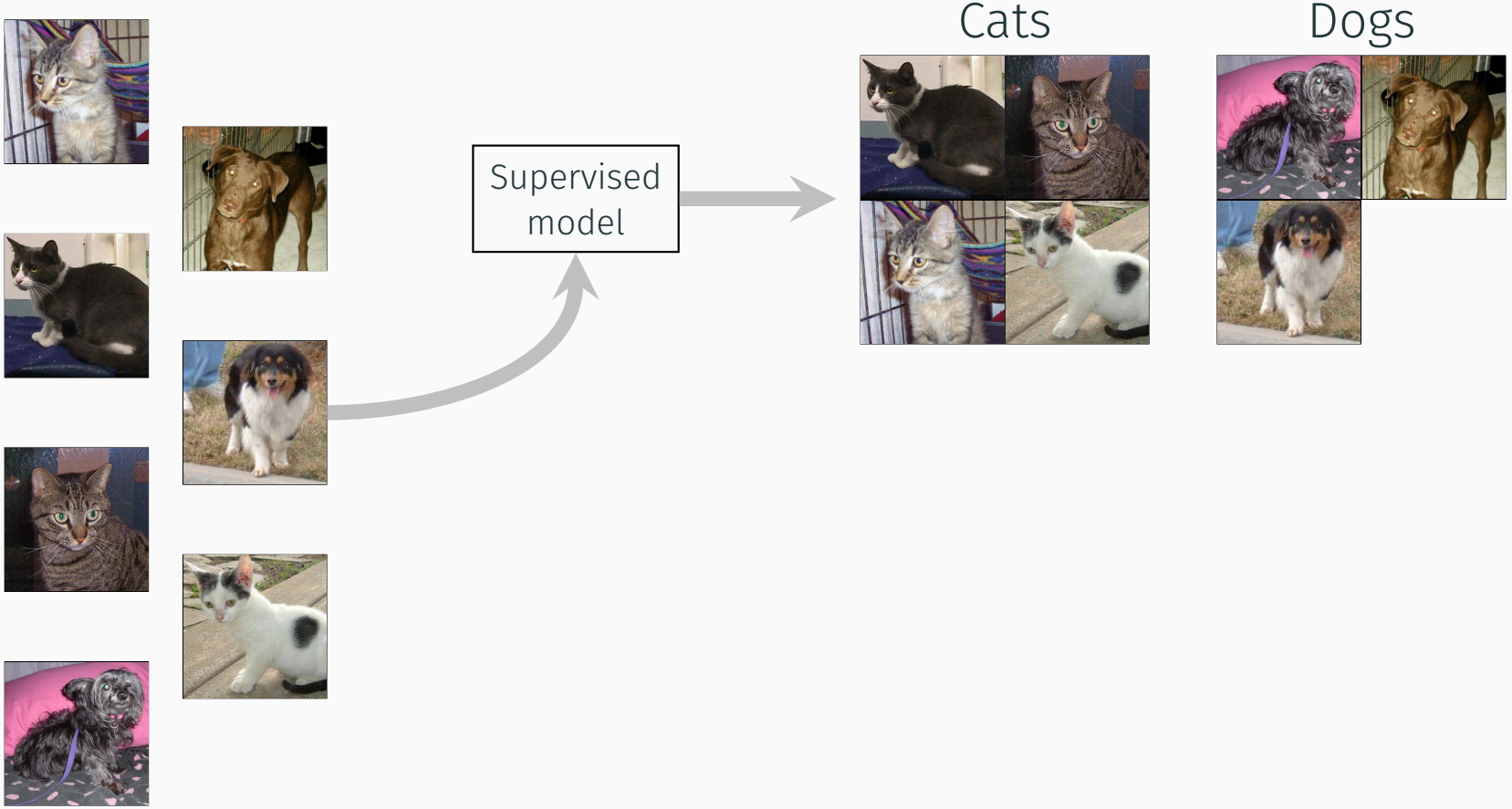
Introduction



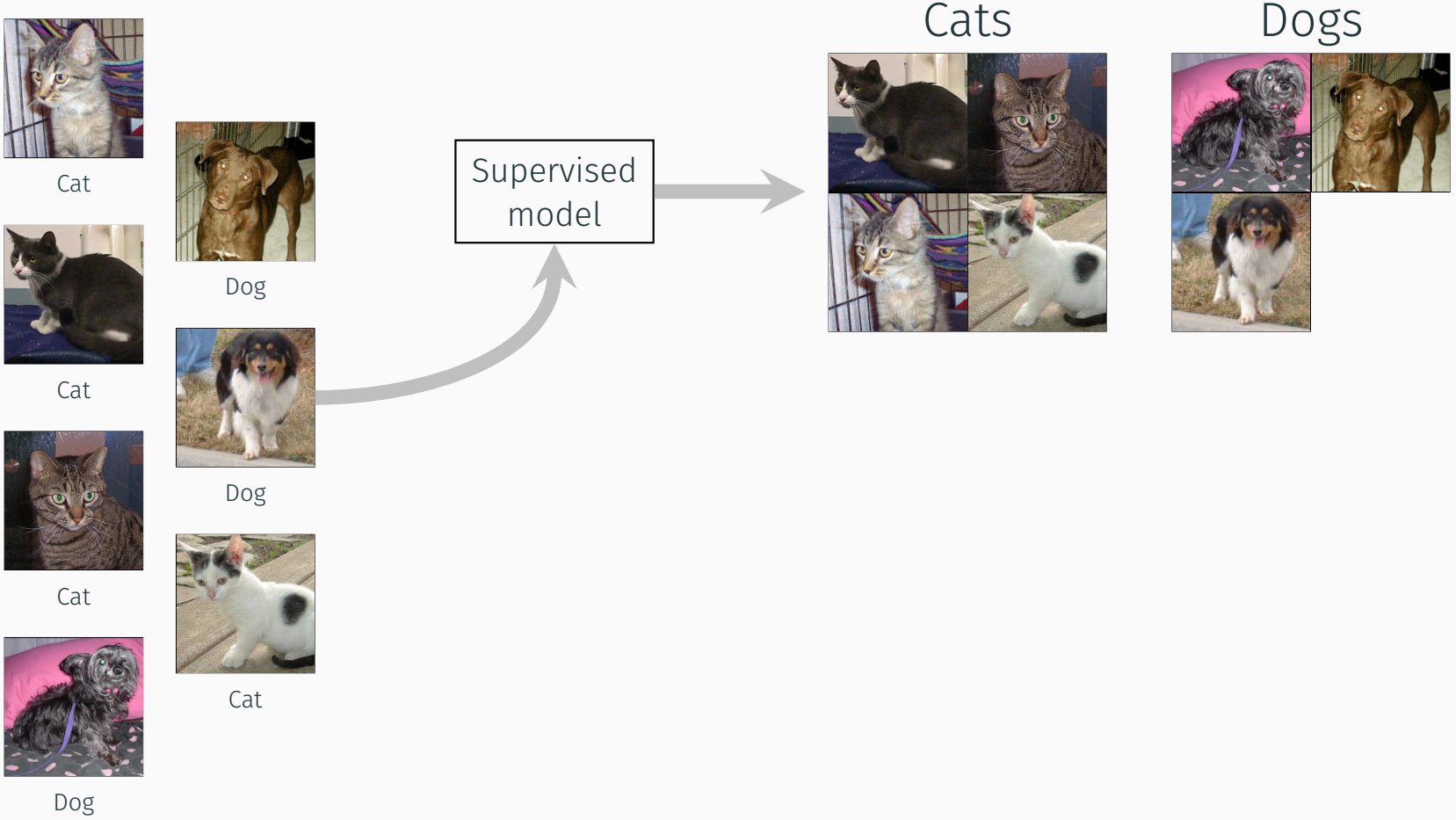




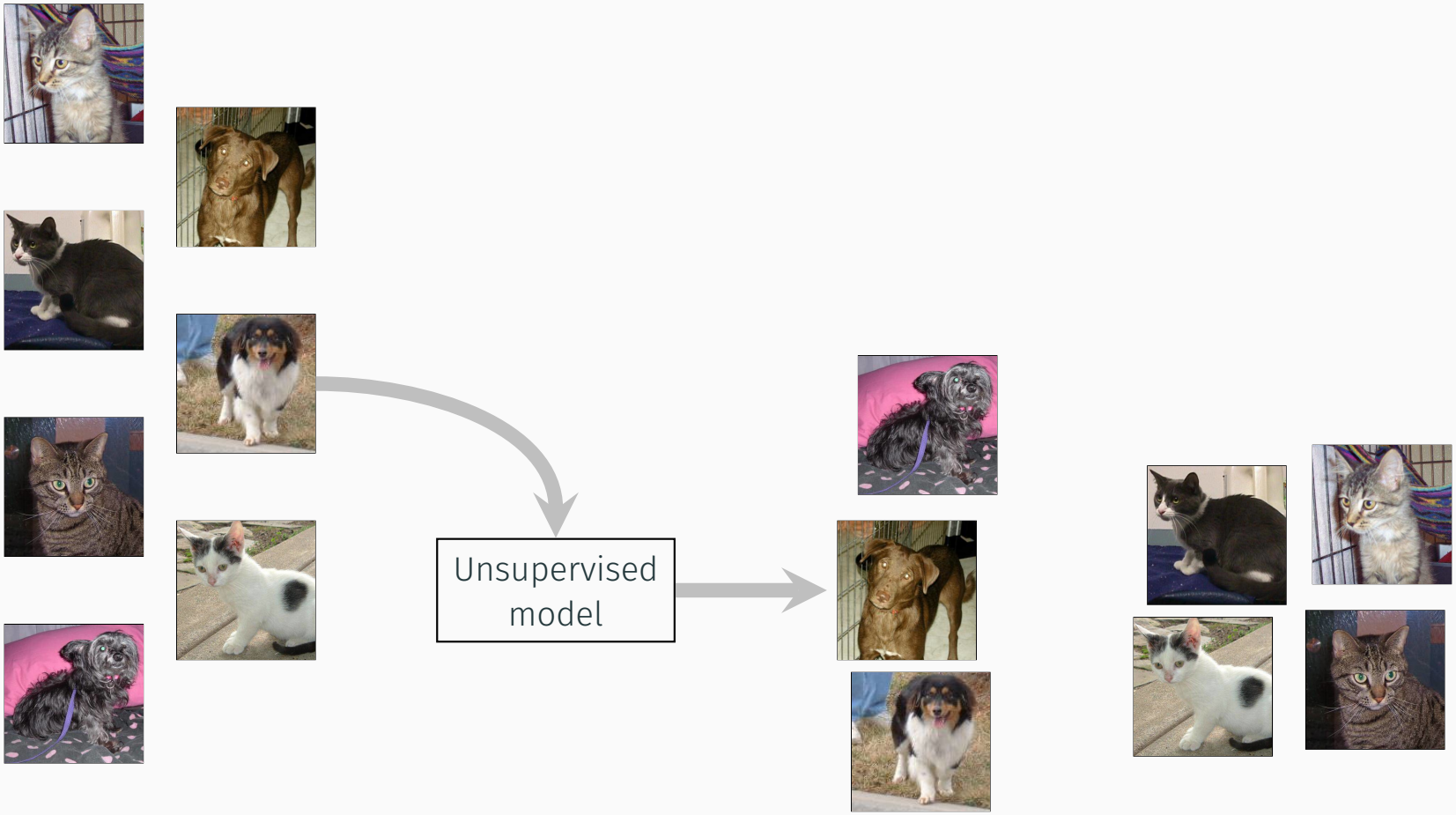
Introduction



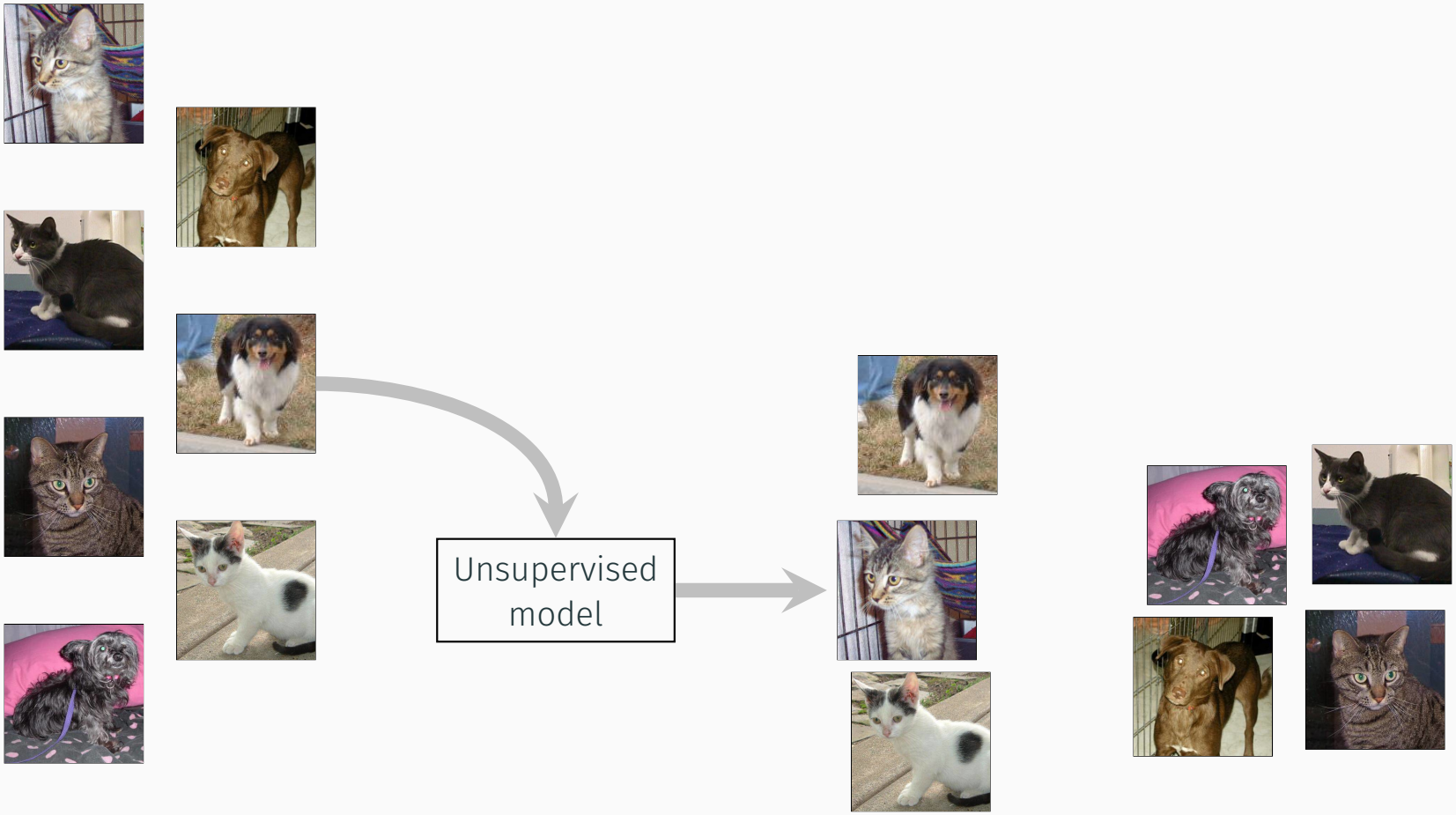
Introduction



Introduction



Introduction



Introduction

Supervised model

Unsupervised model

Introduction: Supervised learning

name	year	cylinders	horsepower	weight	mpg
Chevrolet Chevelle Malibu	1970	8	130	3504	18
Buick Skylark 320	1980	4	165	3693	15
Plymouth Satellite	1971	8	150	3436	18
AMC Rebel SST	1975	4	150	3433	16
Ford Torino	1978	8	140	3449	17

Prerequisites

- A dataset representing a given population

Introduction: Supervised learning

name	year	cylinders	horsepower	weight	mpg
Chevrolet Chevelle Malibu	1970	8	130	3504	18
Buick Skylark 320	1980	4	165	3693	15
Plymouth Satellite	1971	8	150	3436	18
AMC Rebel SST	1975	4	150	3433	16
Ford Torino	1978	8	140	3449	17

Prerequisites

- A dataset representing a given population

Introduction: Supervised learning

name	year	cylinders	horsepower	weight	mpg
Chevrolet Chevelle Malibu	1970	8	130	3504	18
Buick Skylark 320	1980	4	165	3693	15
Plymouth Satellite	1971	8	150	3436	18
AMC Rebel SST	1975	4	150	3433	16
Ford Torino	1978	8	140	3449	17

Prerequisites

- A dataset representing a given population

Introduction: Supervised learning

name	year	cylinders	horsepower	weight	mpg
Chevrolet Chevelle Malibu	1970	8	130	3504	18
Buick Skylark 320	1980	4	165	3693	15
Plymouth Satellite	1971	8	150	3436	18
AMC Rebel SST	1975	4	150	3433	16
Ford Torino	1978	8	140	3449	17

Prerequisites

- A dataset representing a given population
- A response-variable y that we want to predict

Introduction: Supervised learning

name	year	cylinders	horsepower	weight	mpg
Chevrolet Chevelle Malibu	1970	8	130	3504	18
Buick Skylark 320	1980	4	165	3693	15
Plymouth Satellite	1971	8	150	3436	18
AMC Rebel SST	1975	4	150	3433	16
Ford Torino	1978	8	140	3449	17

Prerequisites

- A dataset representing a given population
- A response-variable y that we want to predict
- A set of predictors X that we can use to predict y

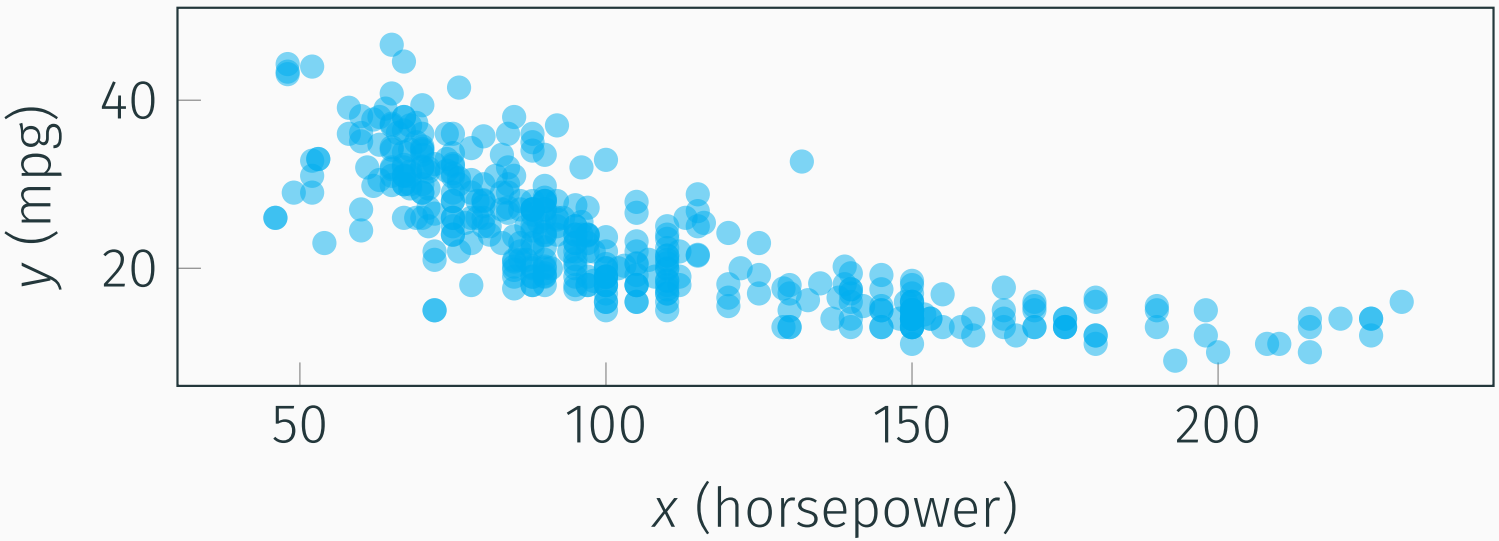
Introduction: Supervised learning

name	year	cylinders	horsepower	weight	mpg
Chevrolet Chevelle Malibu	1970	8	130	3504	18
Buick Skylark 320	1980	4	165	3693	15
Plymouth Satellite	1971	8	150	3436	18
AMC Rebel SST	1975	4	150	3433	16
Ford Torino	1978	8	140	3449	17

Prerequisites

- A dataset representing a given population
- A response-variable y that we want to predict
- A set of predictors X that we can use to predict y
- An **assumed** relationship between X and y that can be described by an unknown function f, such that $y = f(X) + \epsilon$

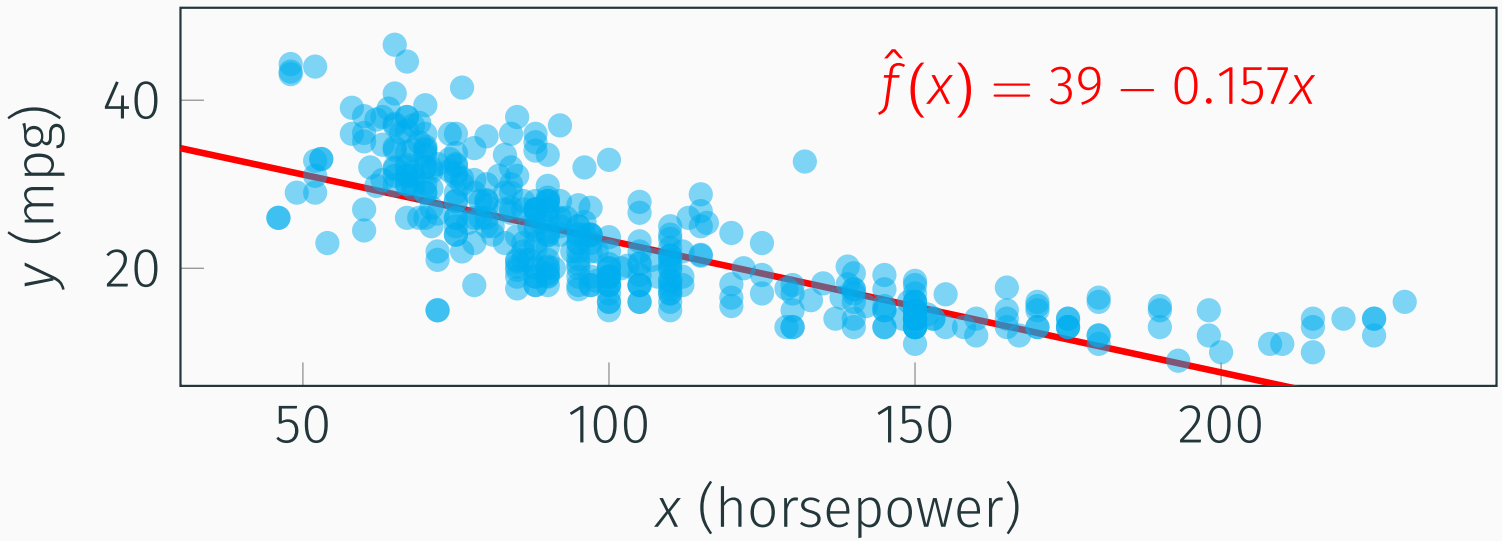
Introduction: Supervised learning



Estimation (or training the model)

- We have assumed that $y = f(X) + \epsilon$, but don't know f

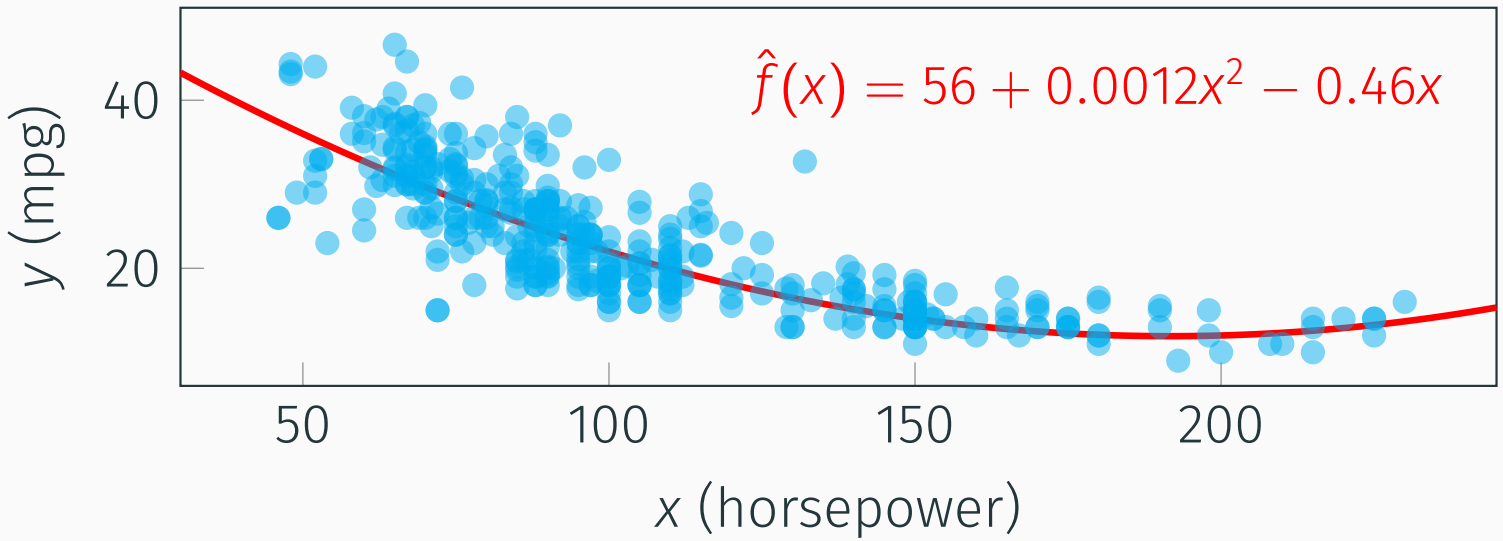
Introduction: Supervised learning



Estimation (or training the model)

- We have assumed that $y = f(X) + \epsilon$, but don't know f
- We produce an estimate $\hat{f}$

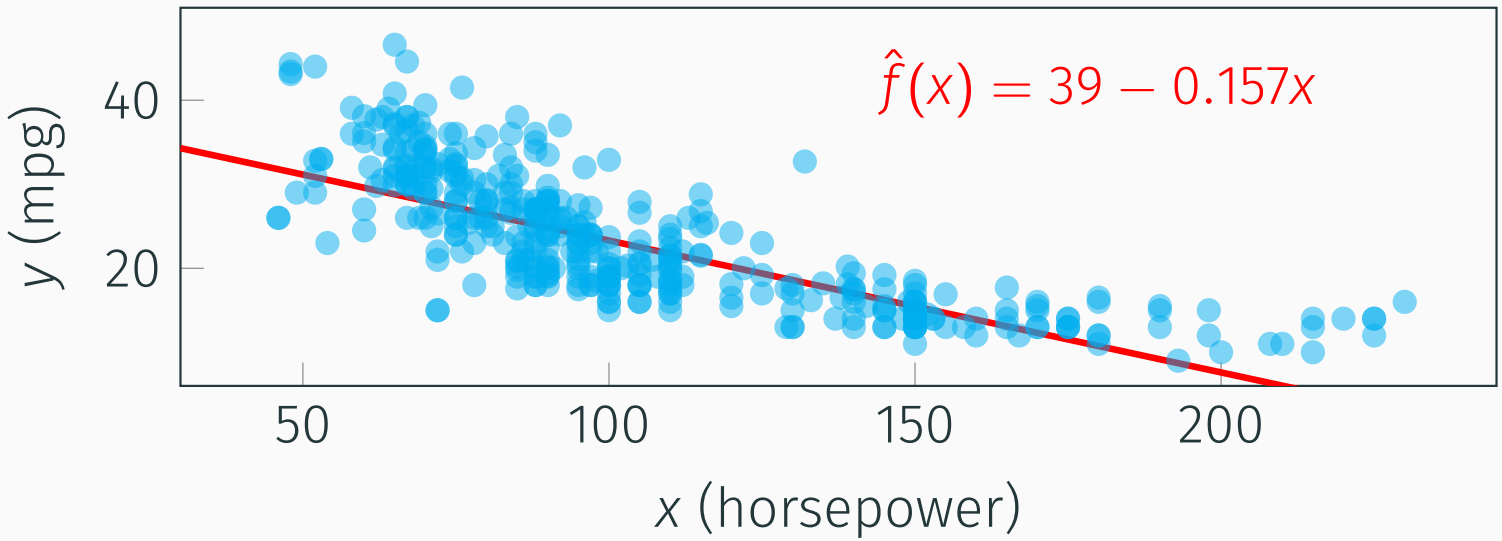
Introduction: Supervised learning



Estimation (or training the model)

- We have assumed that $y = f(X) + \epsilon$, but don't know f
- We produce an estimate $\hat{f}$

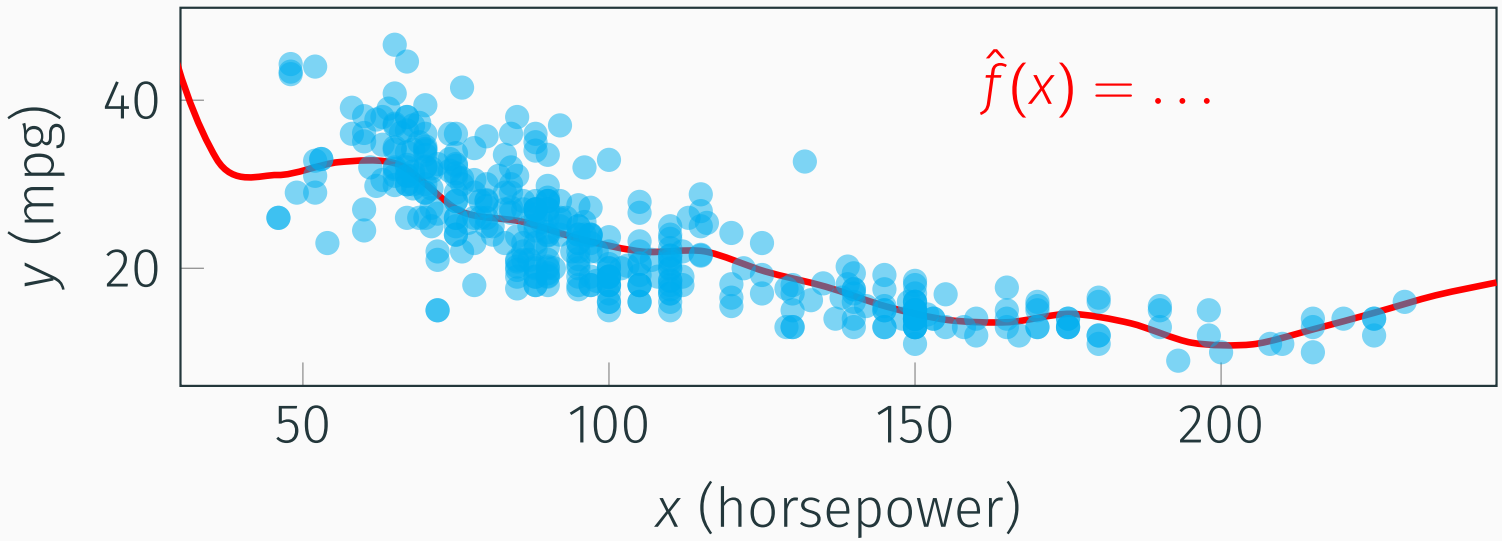
Introduction: Supervised learning



Estimation (or training the model)

- We have assumed that $y = f(X) + \epsilon$, but don't know f
- We produce an estimate $\hat{f}$
- Parametric models: $\hat{f}$ has a simple form
  - $\hat{f}(x) = \beta_0 + \beta_1 x$

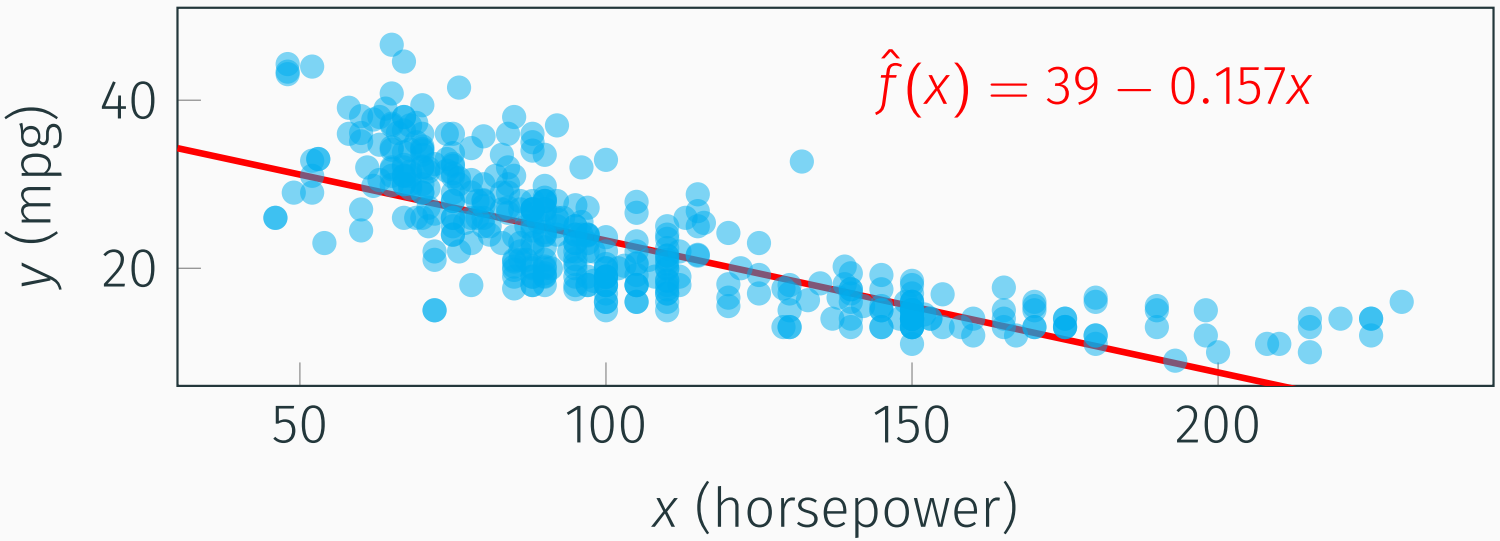
Introduction: Supervised learning



Estimation (or training the model)

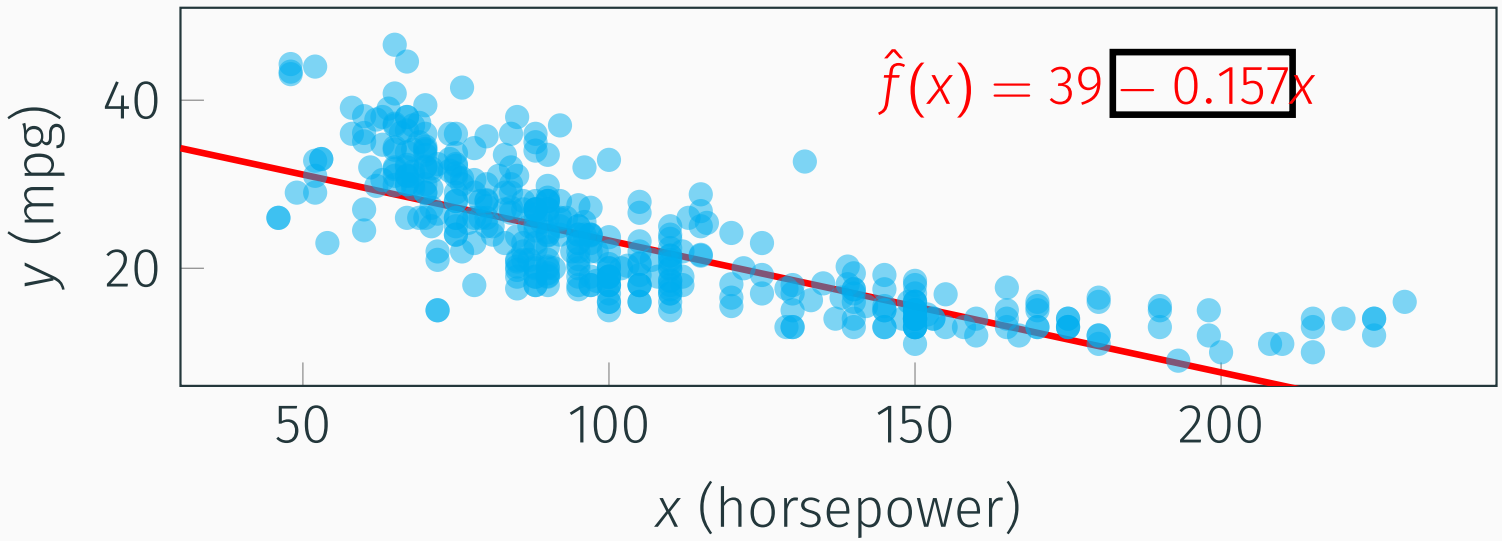
- We have assumed that $y = f(X) + \epsilon$, but don't know f
- We produce an estimate $\hat{f}$
- Parametric models: $\hat{f}$ has a simple form
  - $\hat{f}(x) = \beta_0 + \beta_1 x$
- Non-parametric models: $\hat{f}$ relies directly on the data

Introduction: Supervised learning



Inference: Understanding the relationship between the predictors and the response

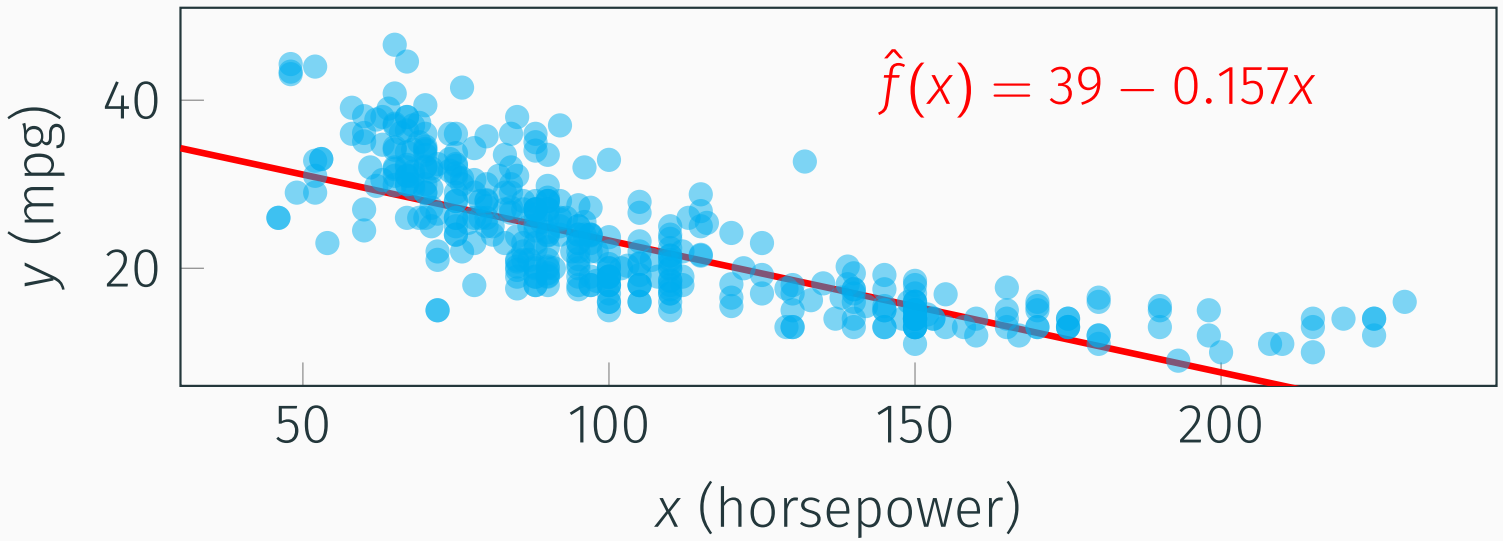
Introduction: Supervised learning



Inference: Understanding the relationship between the predictors and the response

- How does individual features relate to the response?

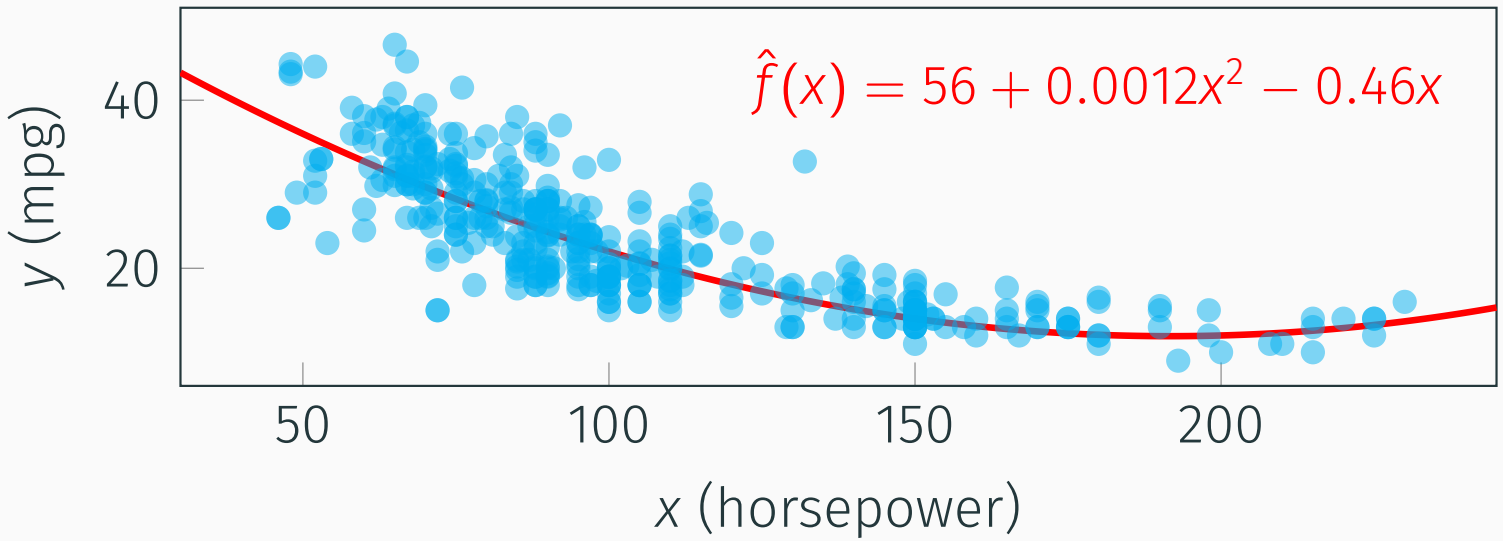
Introduction: Supervised learning



Inference: Understanding the relationship between the predictors and the response

- How does individual features relate to the response?
- What is the *functional form* of the relationship?

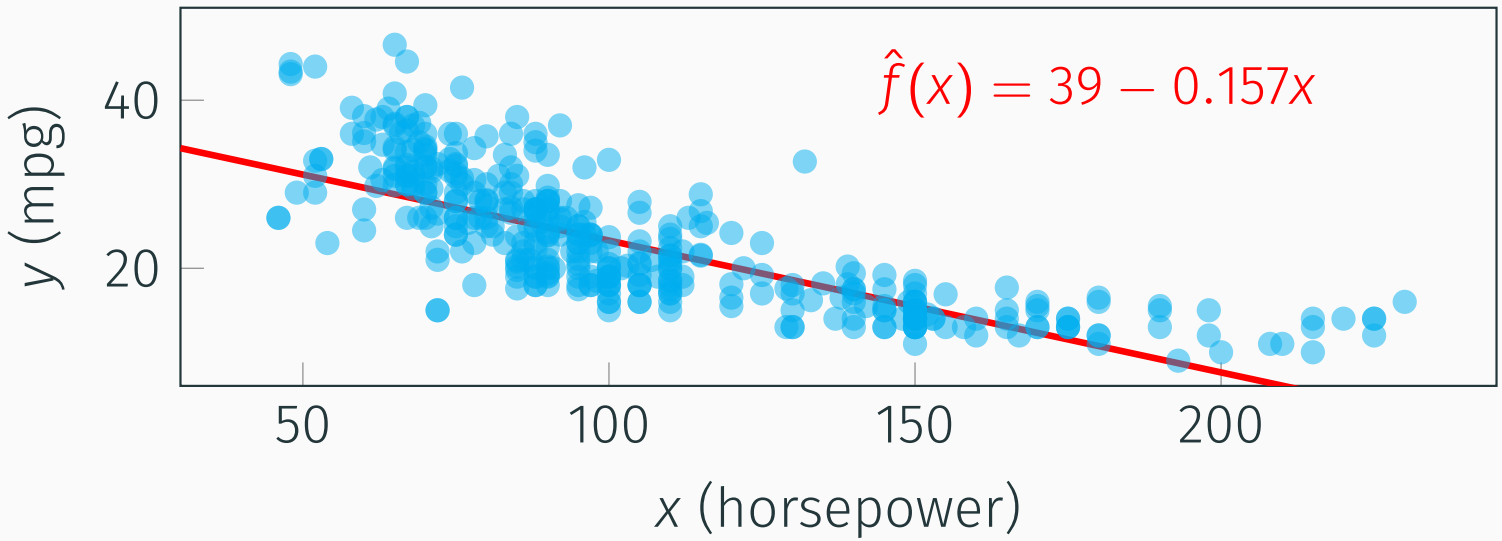
Introduction: Supervised learning



Inference: Understanding the relationship between the predictors and the response

- How does individual features relate to the response?
- What is the *functional form* of the relationship?

Introduction: Supervised learning



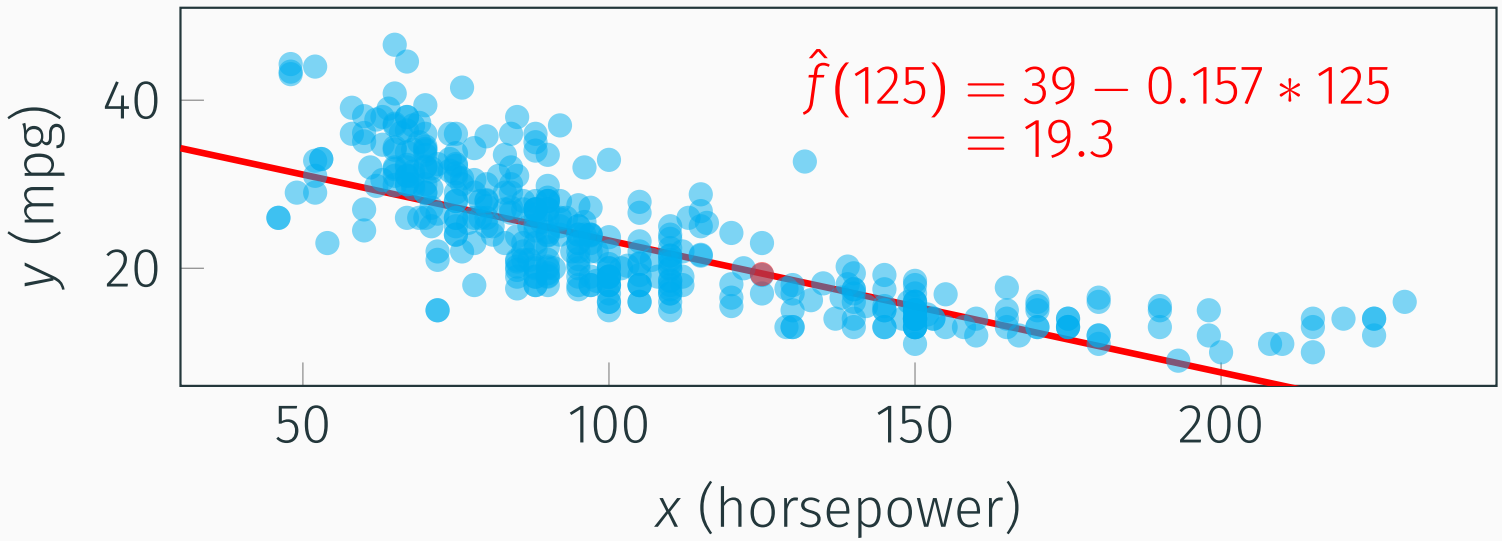
Inference: Understanding the relationship between the predictors and the response

- How does individual features relate to the response?
- What is the *functional form* of the relationship?

Prediction: Predicting the response for new observations

- Plugging new values X into $\hat{f}(X)$

Introduction: Supervised learning



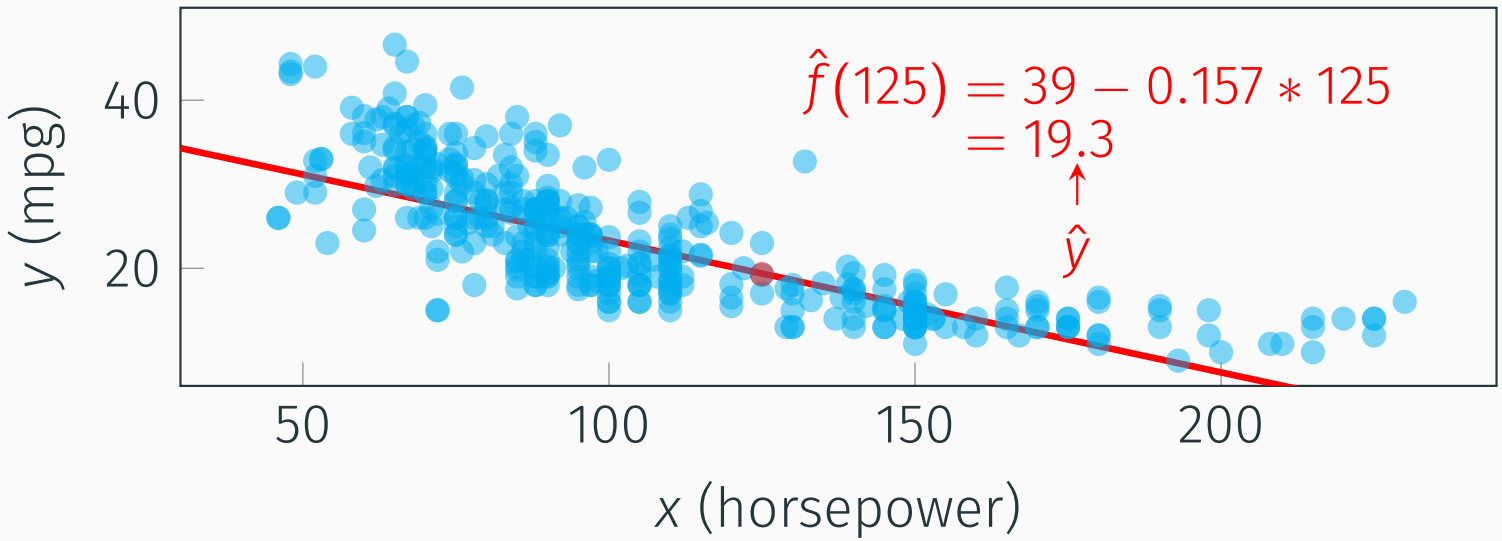
Inference: Understanding the relationship between the predictors and the response

- How does individual features relate to the response?
- What is the *functional form* of the relationship?

Prediction: Predicting the response for new observations

- Plugging new values X into $\hat{f}(X)$

Introduction: Supervised learning



Inference: Understanding the relationship between the predictors and the response

- How does individual features relate to the response?
- What is the *functional form* of the relationship?

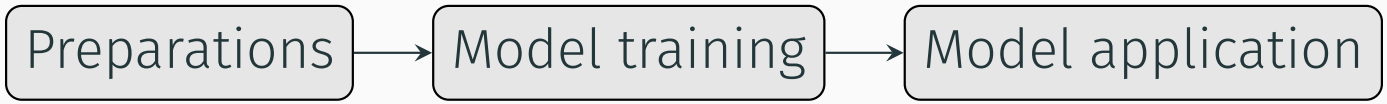
Prediction: Predicting the response for new observations

- Plugging new values X into $\hat{f}(X)$

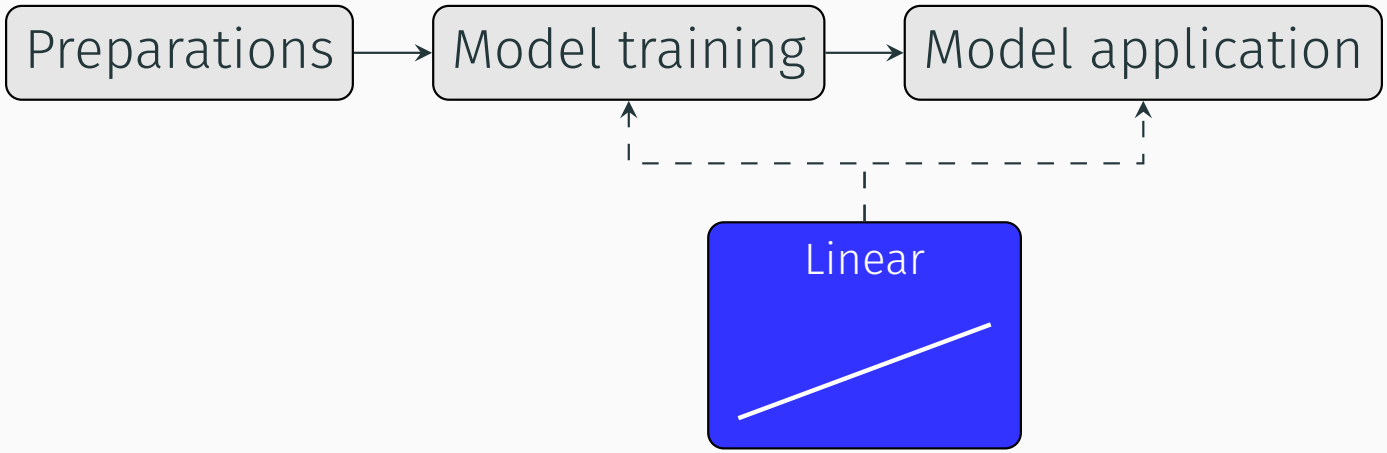
Introduction: Supervised learning

http://localhost:8888/tree

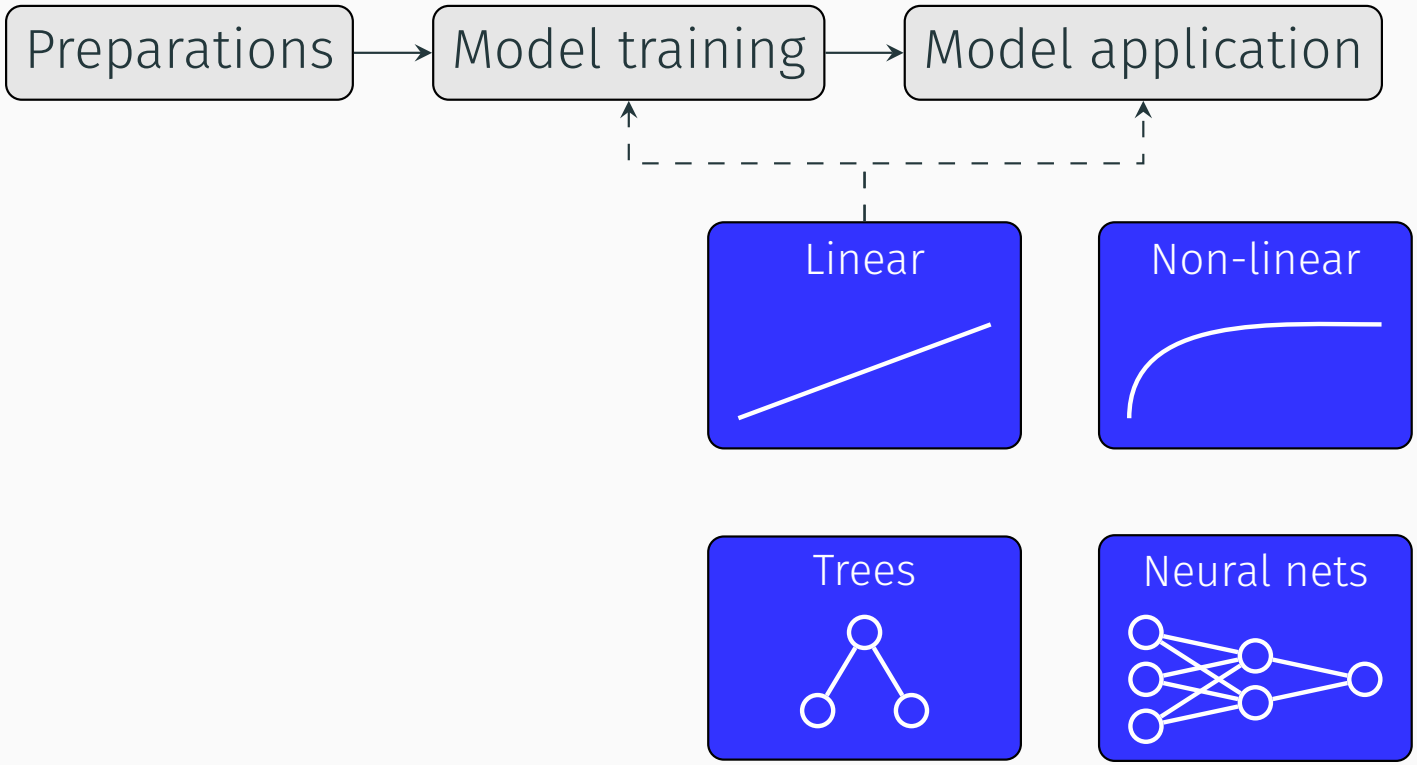
Introduction: Functional interfaces



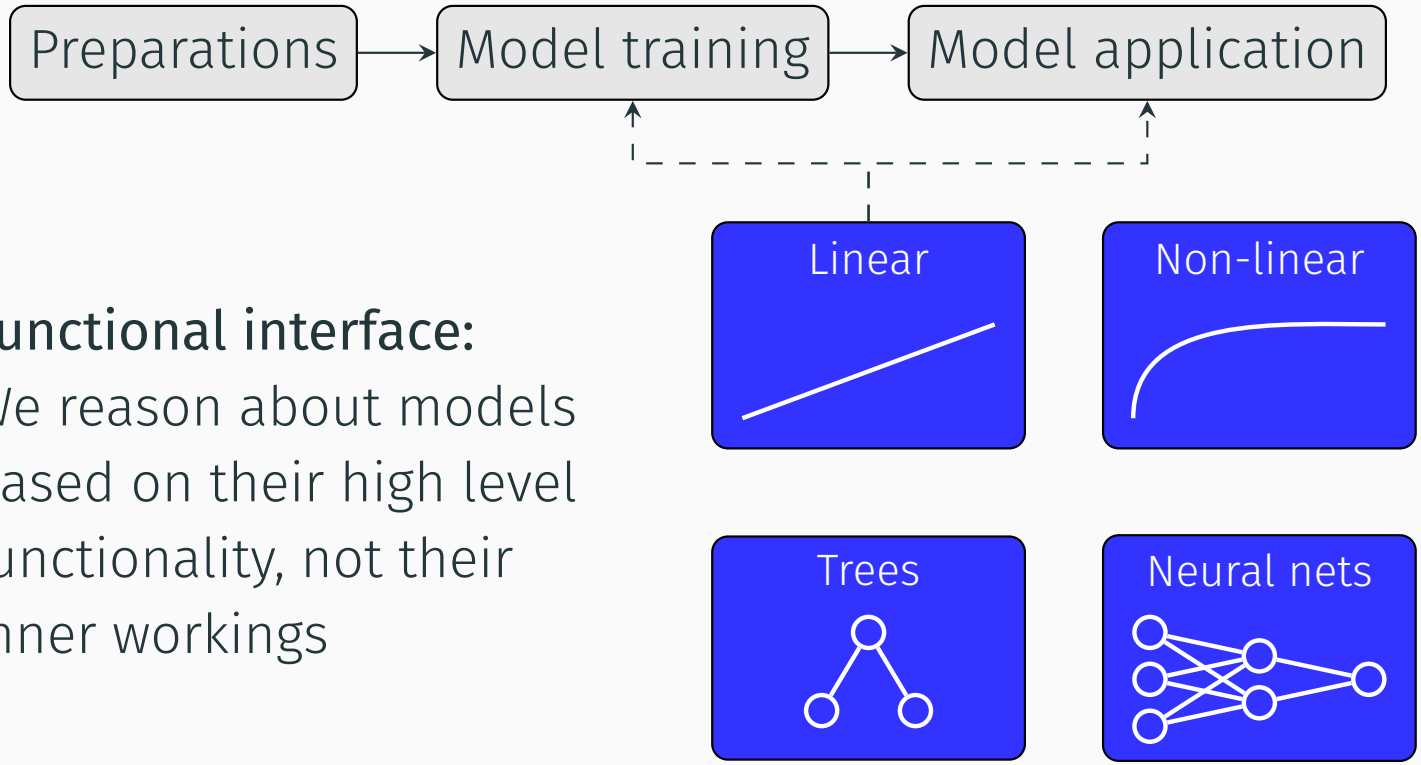
Introduction: Functional interfaces



Introduction: Functional interfaces

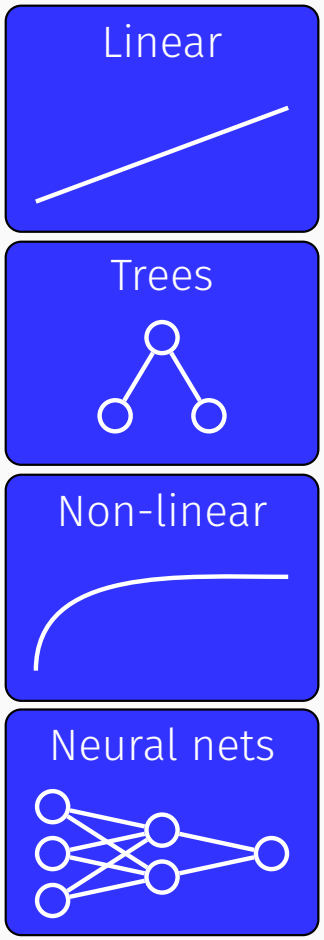


Introduction: Functional interfaces

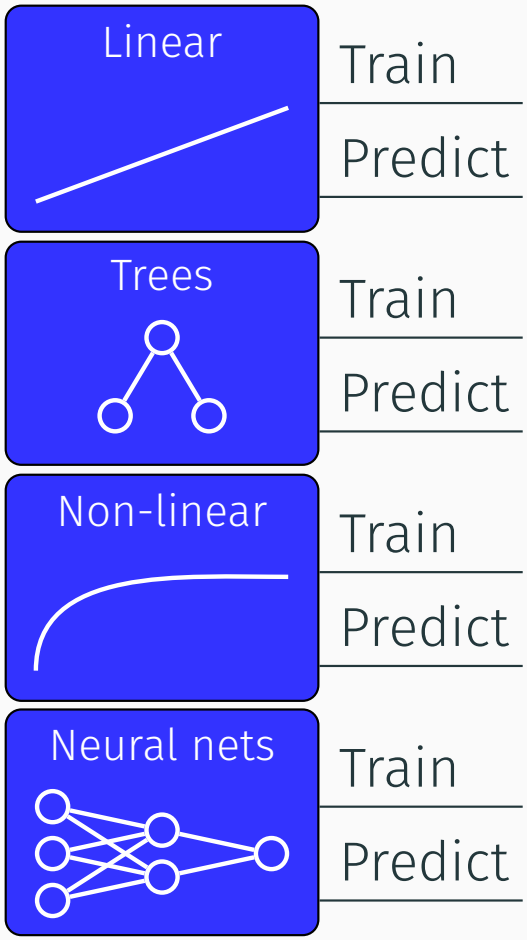


Functional interface:
We reason about models based on their high level functionality, not their inner workings

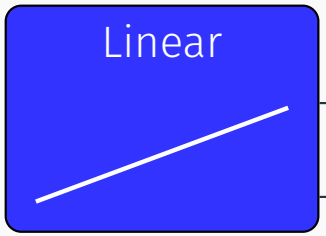
Introduction: Functional interfaces



Introduction: Functional interfaces

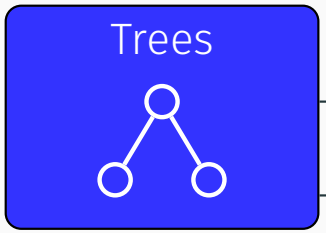


Introduction: Functional interfaces



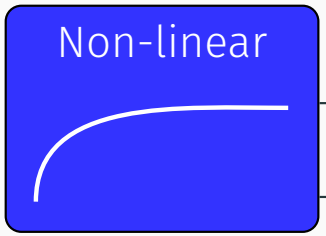
fit(X, y)

predict(X)



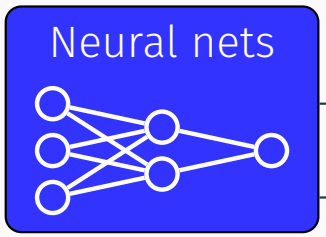
fit(X, y)

predict(X)



fit(X, y)

predict(X)



fit(X, y)

predict(X)

Introduction: Functional interfaces

https://scikit-learn.org/stable/developers/develop.html

Introduction: Functional interfaces

http://localhost:8888/tree

Introduction: Functional interfaces

Supervised learning: We train a model to solve a problem we already know, and that we already have labels for

- Although a bunch of things can vary (e.g. how we preprocess the data, the type of model, how we measure performance), the overall pattern very often remains the same:
  1. Data preparation
  2. Train the model
  3. Apply the model

Introduction: Model performance

Regression

Classification

Introduction: Model performance

Regression

y
18
15
18
16
17

Classification

y
cat
cat
dog
cat
dog

Introduction: Model performance

Regression

y	$\hat{y}$
18	15.3
15	16.1
18	17.2
16	16.8
17	19.5

Classification

y
cat
cat
dog
cat
dog

Introduction: Model performance

Regression

y	$\hat{y}$
18	15.3
15	16.1
18	17.2
16	16.8
17	19.5

Mean squared error (MSE):

$$\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

Classification

y
cat
cat
dog
cat
dog

Introduction: Model performance

Regression

y	$\hat{y}$	
18	15.3	$(18 - 15.3)^2 = 7.29$
15	16.1	$(15 - 16.1)^2 = 1.21$
18	17.2	$(18 - 17.2)^2 = 0.64$
16	16.8	$(16 - 16.4)^2 = 0.16$
17	19.5	$(17 - 19.5)^2 = \underline{6.25}$
		3.11

Mean squared error (MSE):

$$\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

Classification

y
cat
cat
dog
cat
dog

Introduction: Model performance

Regression

y	$\hat{y}$	
18	15.3	$(18 - 15.3)^2 = 7.29$
15	16.1	$(15 - 16.1)^2 = 1.21$
18	17.2	$(18 - 17.2)^2 = 0.64$
16	16.8	$(16 - 16.4)^2 = 0.16$
17	19.5	$(17 - 19.5)^2 = 6.25$
		3.11

Mean squared error (MSE):

$$\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

Classification

y
cat
cat
dog
cat
dog

Introduction: Model performance

Regression

y	$\hat{y}$	
18	15.3	$(18 - 15.3)^2 = 7.29$
15	16.1	$(15 - 16.1)^2 = 1.21$
18	17.2	$(18 - 17.2)^2 = 0.64$
16	16.8	$(16 - 16.4)^2 = 0.16$
17	19.5	$(17 - 19.5)^2 = 6.25$
		3.11

Mean squared error (MSE):

$$\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

Classification

y	$\hat{y}$
cat	cat
cat	dog
dog	dog
cat	cat
dog	cat

Introduction: Model performance

Regression

y	$\hat{y}$	
18	15.3	$(18 - 15.3)^2 = 7.29$
15	16.1	$(15 - 16.1)^2 = 1.21$
18	17.2	$(18 - 17.2)^2 = 0.64$
16	16.8	$(16 - 16.4)^2 = 0.16$
17	19.5	$(17 - 19.5)^2 = 6.25$
		3.11

Mean squared error (MSE):

$$\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

Classification

y	$\hat{y}$
cat	cat
cat	dog
dog	dog
cat	cat
dog	cat

Accuracy:

$$\frac{1}{n}\sum_{i=0}^{n}\mathbb{1}(y = \hat{y})$$

Introduction: Model performance

Regression

y	$\hat{y}$	
18	15.3	$(18 - 15.3)^2 = 7.29$
15	16.1	$(15 - 16.1)^2 = 1.21$
18	17.2	$(18 - 17.2)^2 = 0.64$
16	16.8	$(16 - 16.4)^2 = 0.16$
17	19.5	$(17 - 19.5)^2 = 6.25$
		3.11

Mean squared error (MSE):

$$\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

Classification

y	$\hat{y}$	
cat	cat	cat=cat $\implies 1$
cat	dog	cat≠dog $\implies 0$
dog	dog	dog=dog $\implies 1$
cat	cat	cat=cat $\implies 1$
dog	cat	dog≠cat $\implies 0$
		0.60

Accuracy:

$$\frac{1}{n}\sum_{i=0}^{n}\mathbb{1}(y = \hat{y})$$

Introduction: Model performance

http://localhost:8888/tree

Introduction: Model performance

mean_squared_error(y, predictions)

Introduction: Model performance

mean_squared_error(y, predictions)
mean_absolute_error(y, predictions)
median_absolute_error(y, predictions)
r2_score(y, predictions)
⋮

https://scikit-learn.org/stable/api/sklearn.metrics.html

Introduction: Model performance

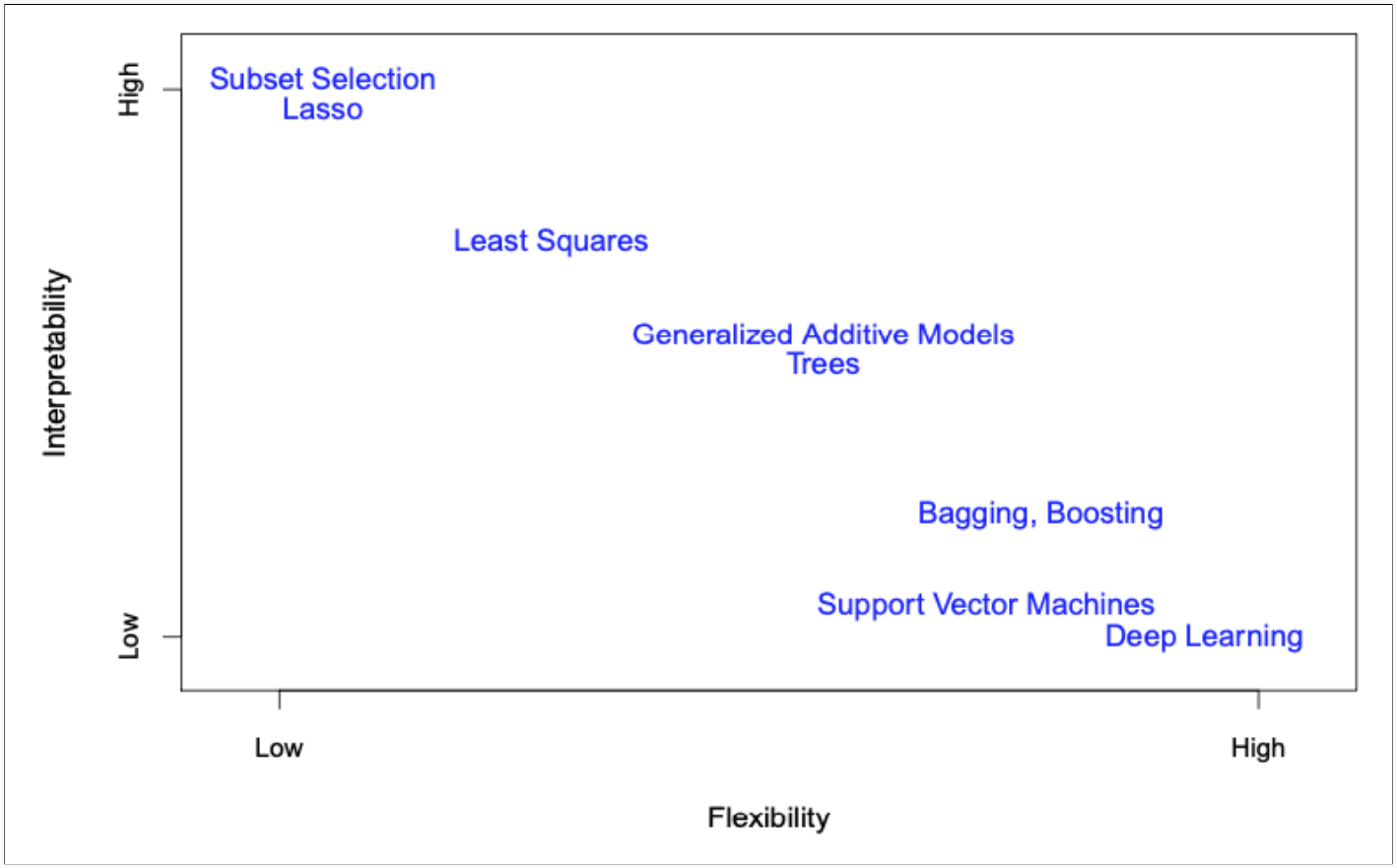
http://localhost:8888/tree

Introduction: Model performance

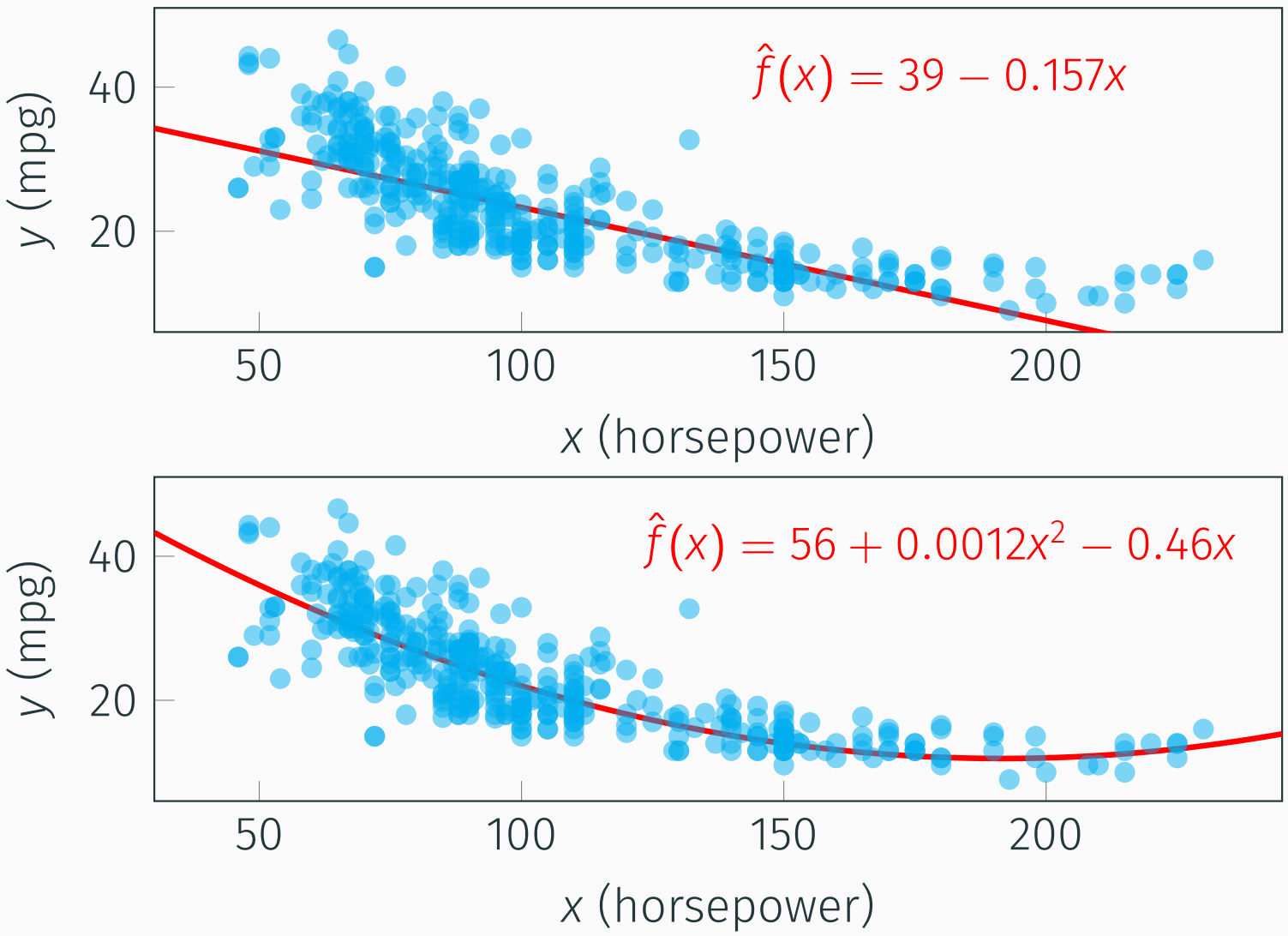
We generally divide prediction problems into two types, regression and classification

- Regression problems deal with continuous outputs
- Classification problems deal with categorical outputs
- There is a range of metrics to evaluate model performance, each suitable for different purposes

Introduction: The Bias-Variance Trade-off



Introduction: The Bias-Variance Trade-off



Introduction: The Bias-Variance Trade-off

Model performance will depend on the dataset we use to calculate the performance metrics

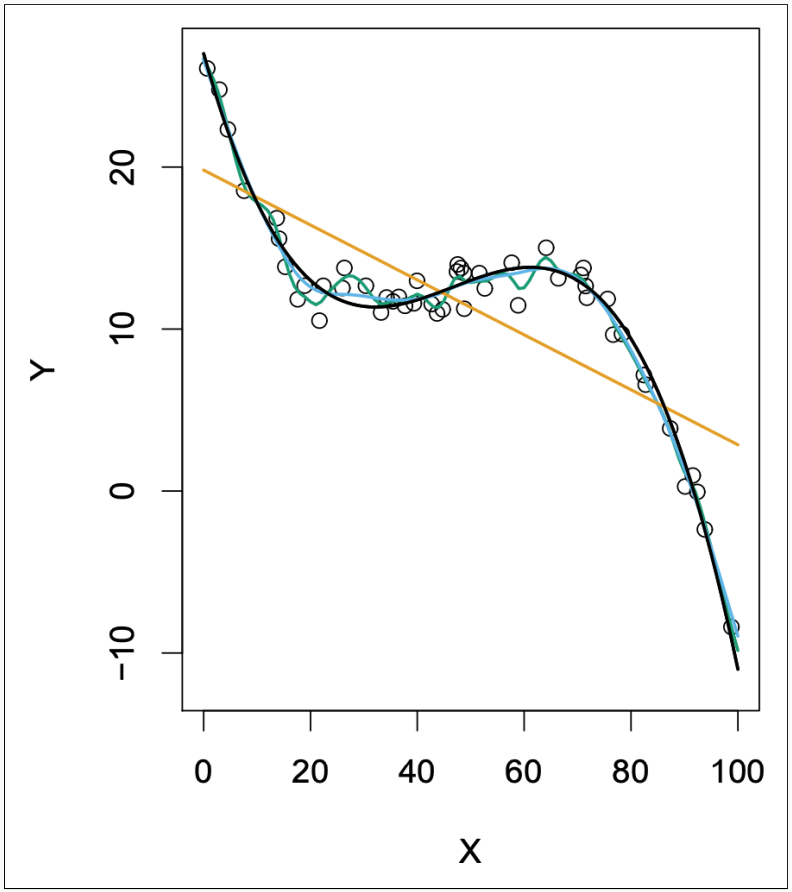
- Training set: The data we use to estimate the model
  - With a sufficiently flexible model we can **always** achieve 0 error in the training set
- Test set: Data held-out from the training set such that it remains unseen by the model
  - Performance in the test set is indicative of how well the model generalizes to new data (almost always worse than in the training set)
  - If our model performs well in new data, we can assume that it accurately describes the relationship between the predictors and the response in the **general case**

Introduction: The Bias-Variance Trade-off

How can our model perform poorly?

- Underfitting: The model is too simple to capture the relationship between the predictors and the response
  - High error in both the training and test set
- Overfitting: The model is too complex and captures noise in the training set
  - Low error in the training set, high error in the test set

Introduction: The Bias-Variance Trade-off



Introduction: The Bias-Variance Trade-off

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Var}(\hat{f}(x)) + \left[\mathrm{Bias}(\hat{f}(x))\right]^2 + \mathrm{Var}(\epsilon)$$

Introduction: The Bias-Variance Trade-off

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Var}(\hat{f}(x)) + \left[\mathrm{Bias}(\hat{f}(x))\right]^2 + \mathrm{Var}(\epsilon)$$

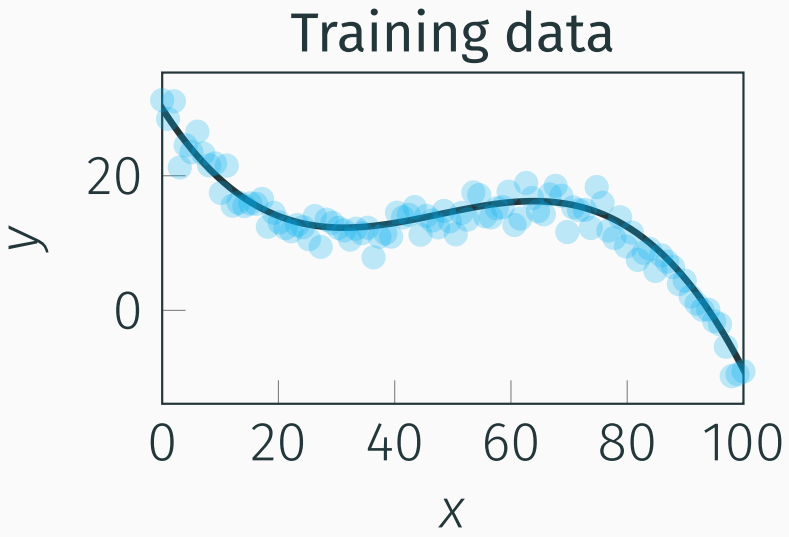
Irreducible error

Introduction: The Bias-Variance Trade-off

$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Var}(\hat{f}(x)) + \left[\mathrm{Bias}(\hat{f}(x))\right]^2 + \mathrm{Var}(\epsilon)$$

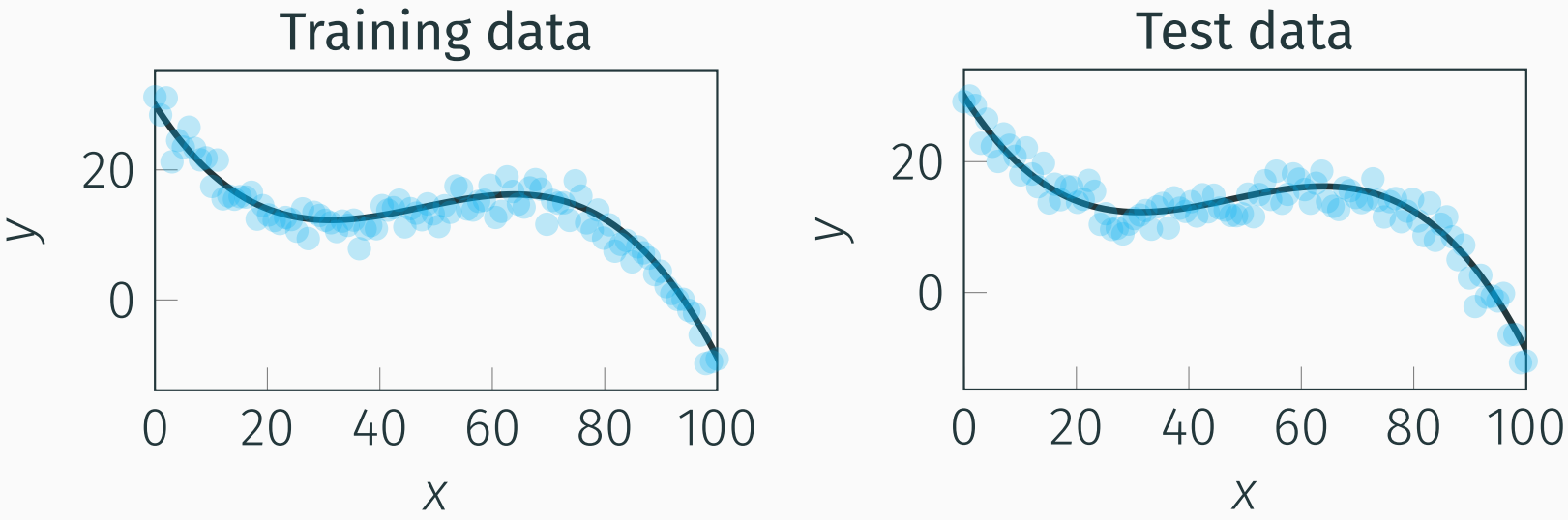
Variance Bias

Introduction: The Bias-Variance Trade-off



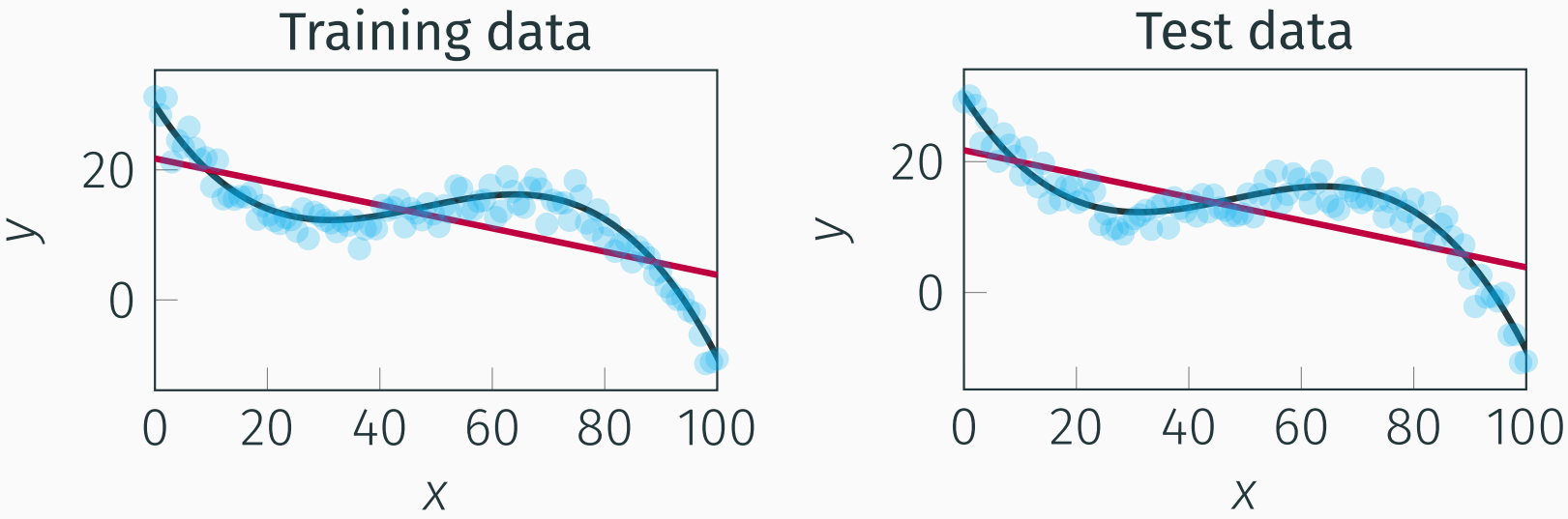
$$f(x) = -0.000226x^3 + 0.032262x^2 - 1.3543x + 30 + \epsilon$$

Introduction: The Bias-Variance Trade-off



$$f(x) = -0.000226x^3 + 0.032262x^2 - 1.3543x + 30 + \epsilon$$

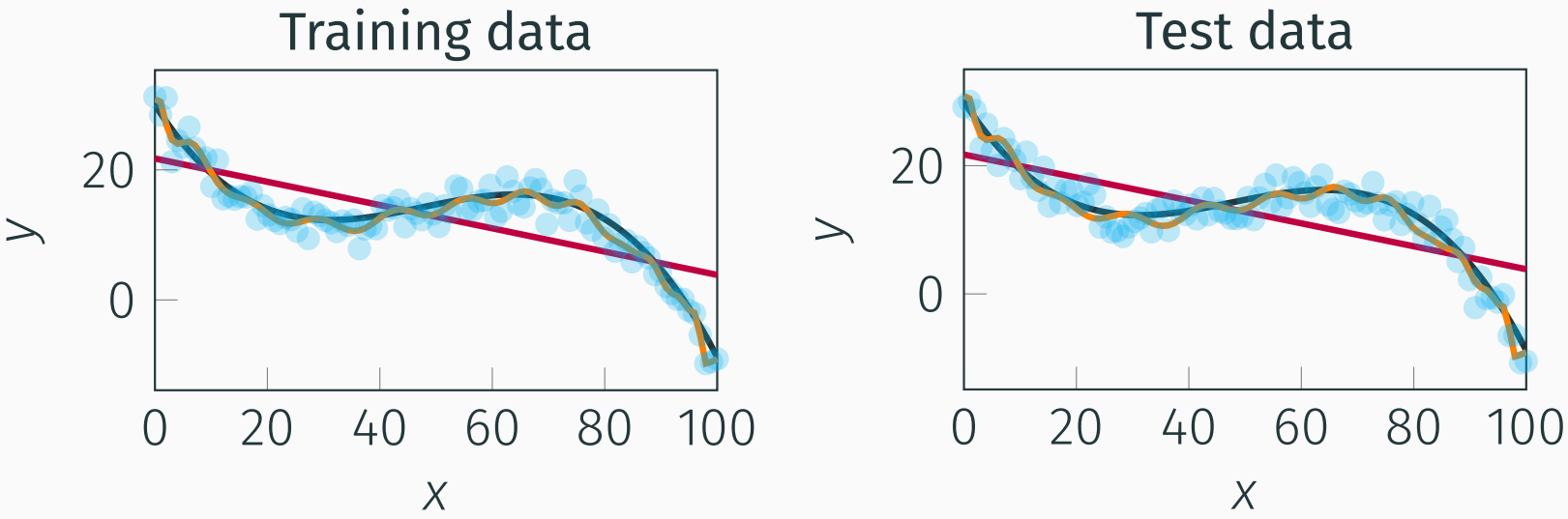
Introduction: The Bias-Variance Trade-off



$$f(x) = -0.000226x^3 + 0.032262x^2 - 1.3543x + 30 + \epsilon$$

$$\hat{f}_0(x) = -0.17x + 21.74$$

Introduction: The Bias-Variance Trade-off



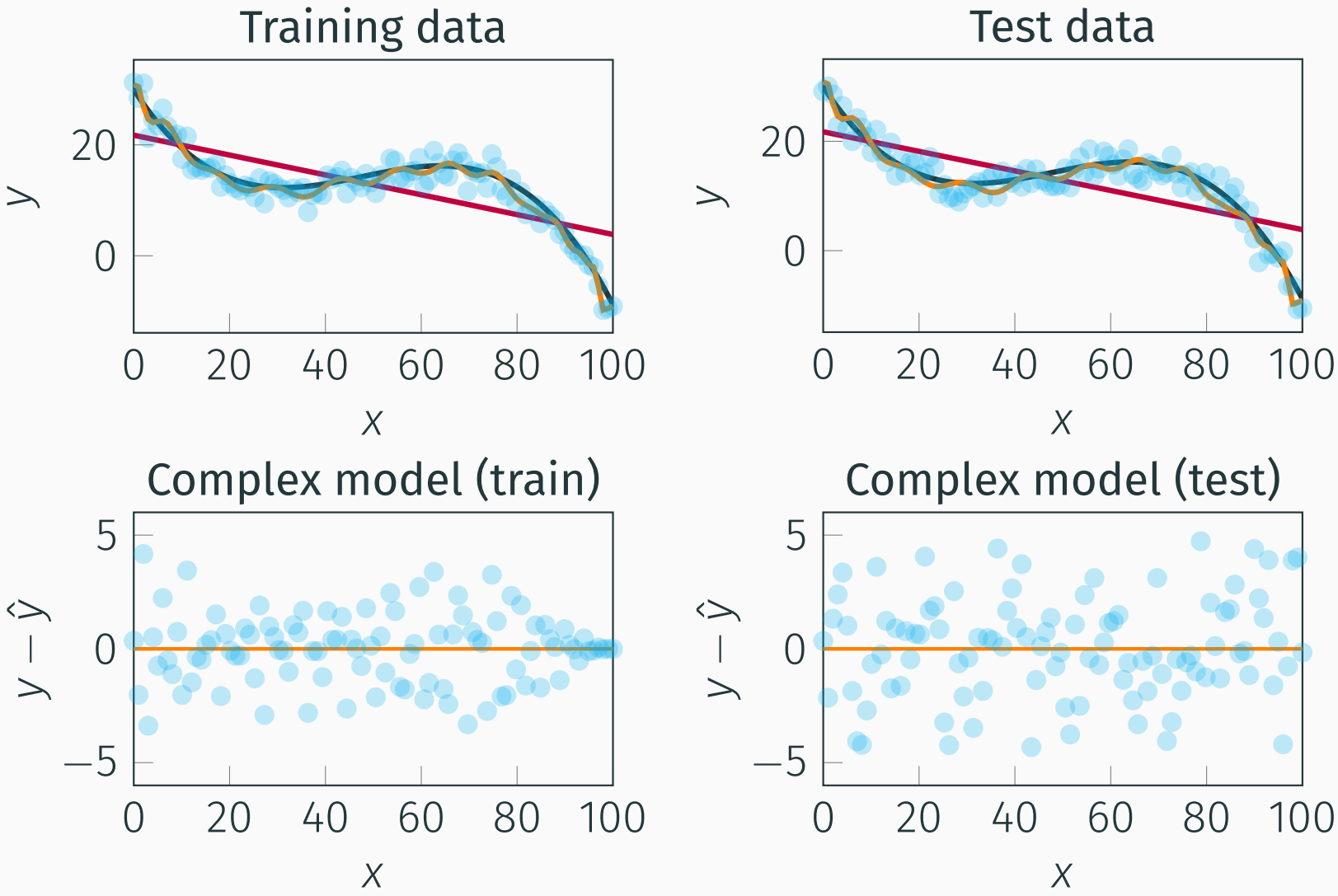
$$f(x) = -0.000226x^3 + 0.032262x^2 - 1.3543x + 30 + \epsilon$$

$$\hat{f}_0(x) = -0.17x + 21.74$$

$$\hat{f}_1(x) = 1.32 * 10^{-142}x^{80} - 2.18 * 10^{-140}x^{79} + \ldots$$

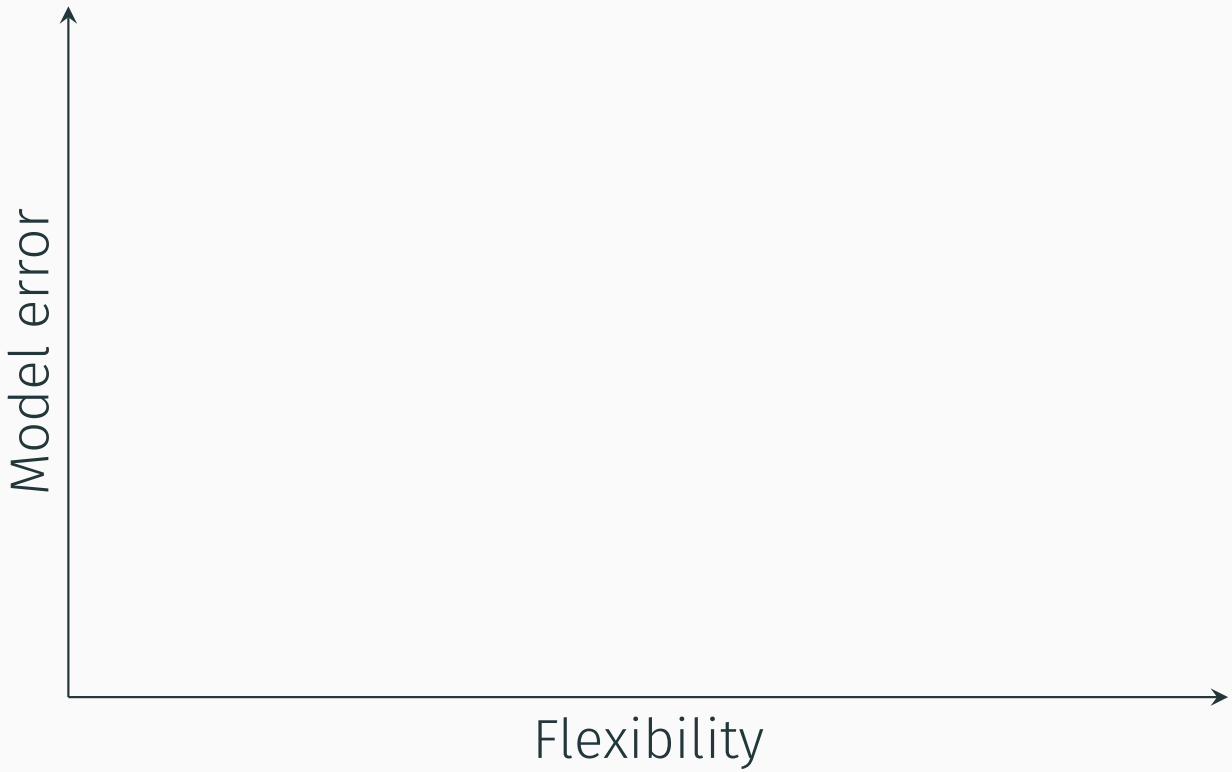
Introduction: The Bias-Variance Trade-off

Introduction: The Bias-Variance Trade-off

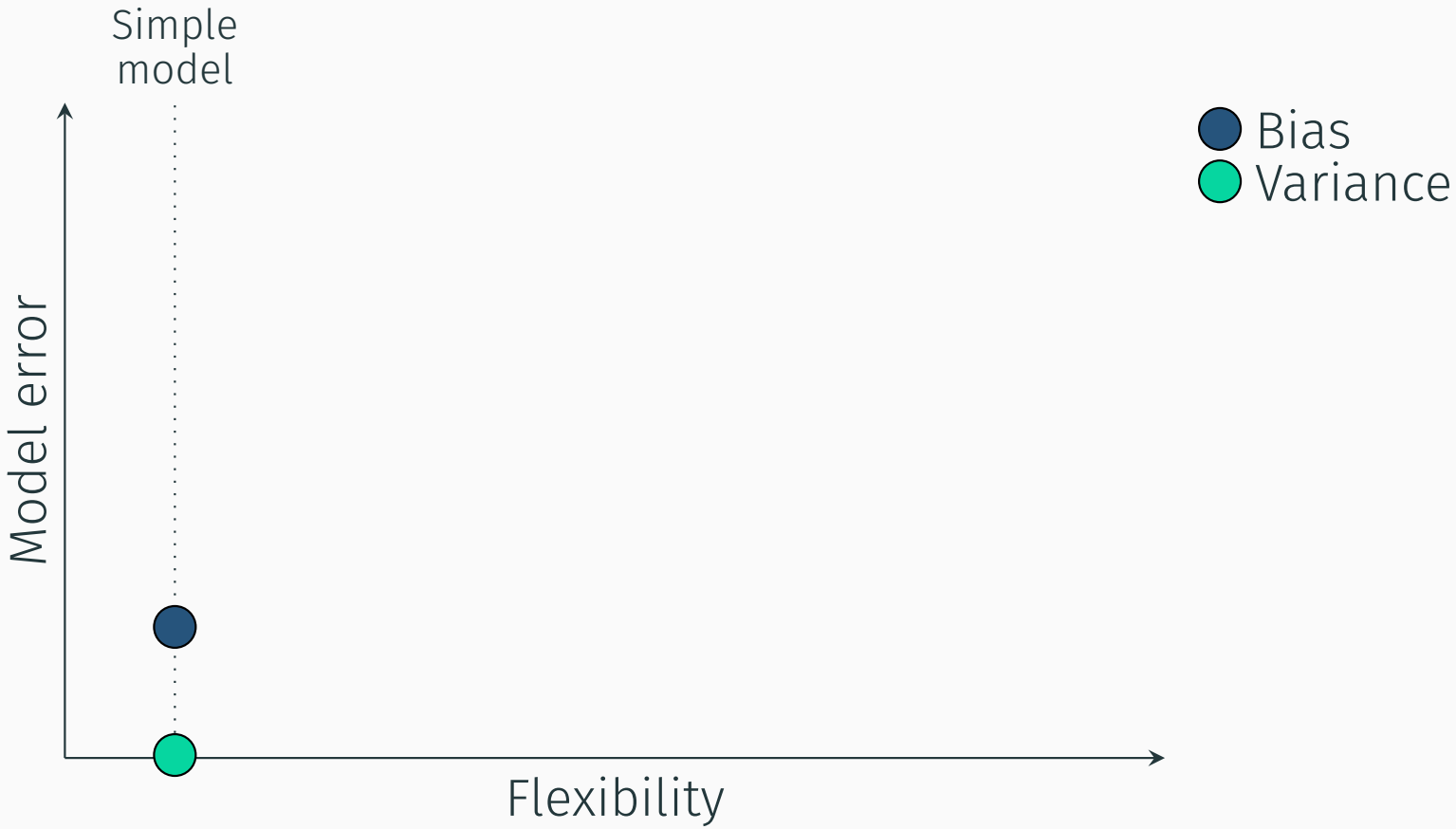


Introduction: The Bias-Variance Trade-off

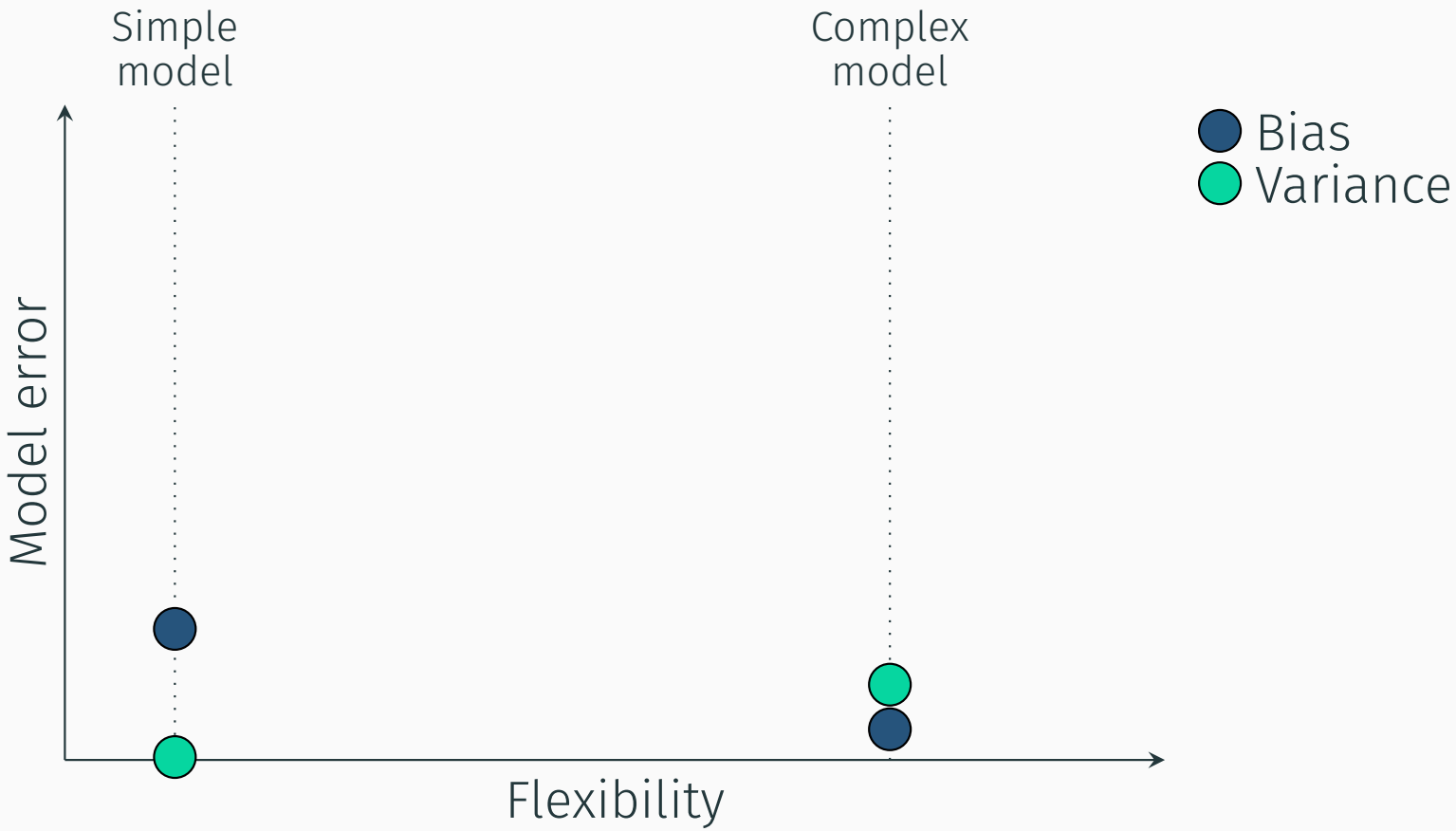
Introduction: The Bias-Variance Trade-off



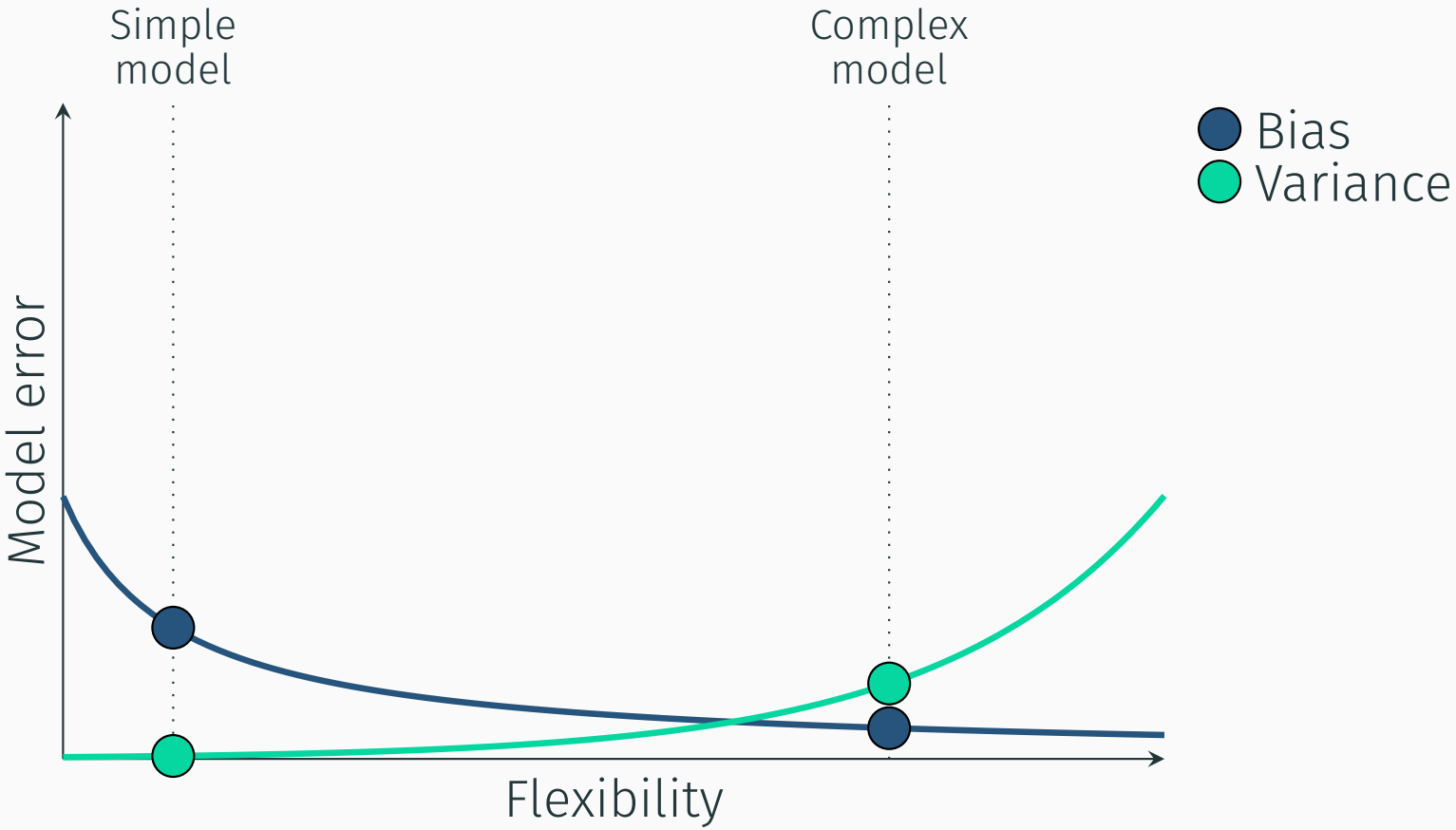
Introduction: The Bias-Variance Trade-off



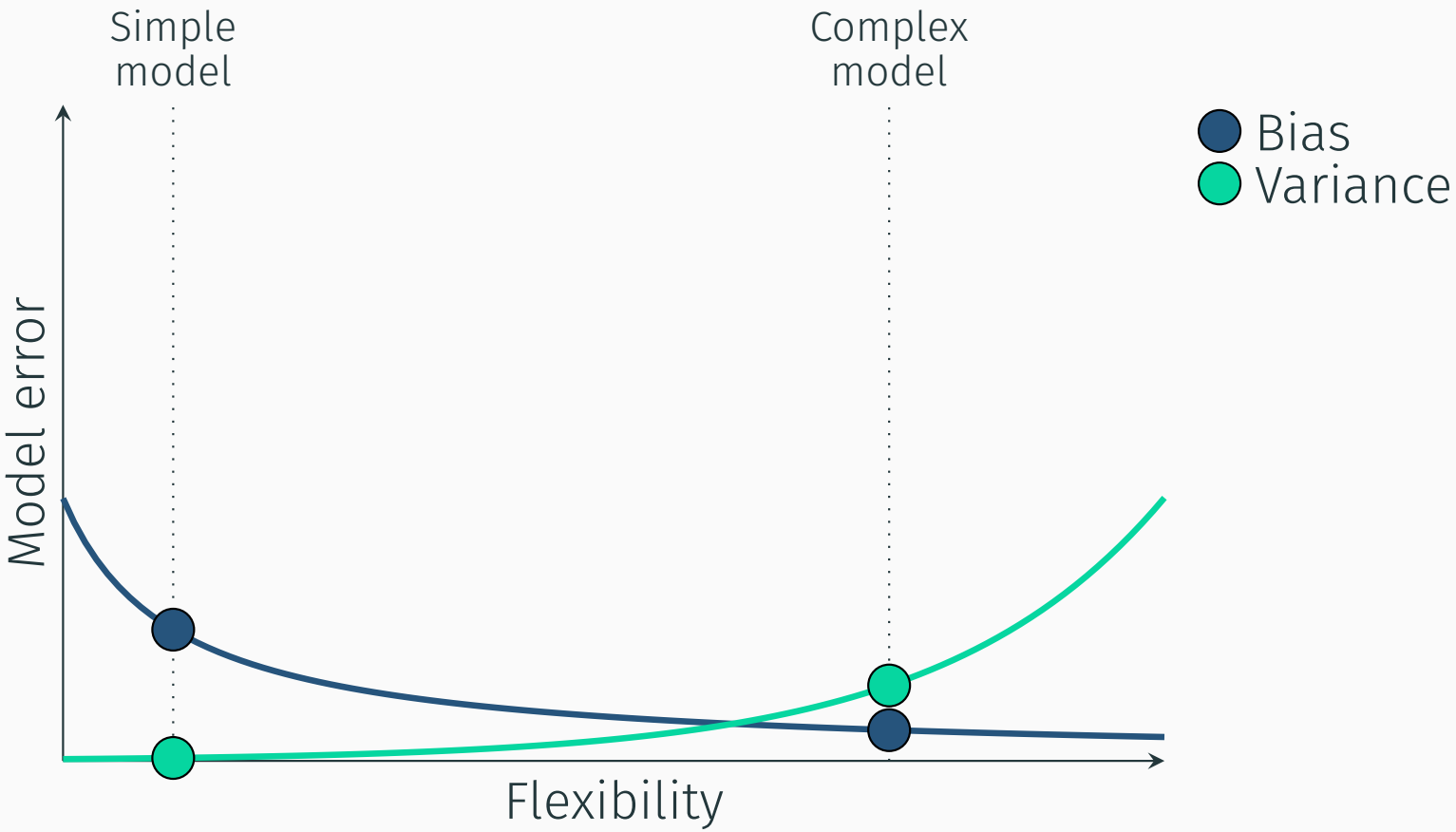
Introduction: The Bias-Variance Trade-off



Introduction: The Bias-Variance Trade-off

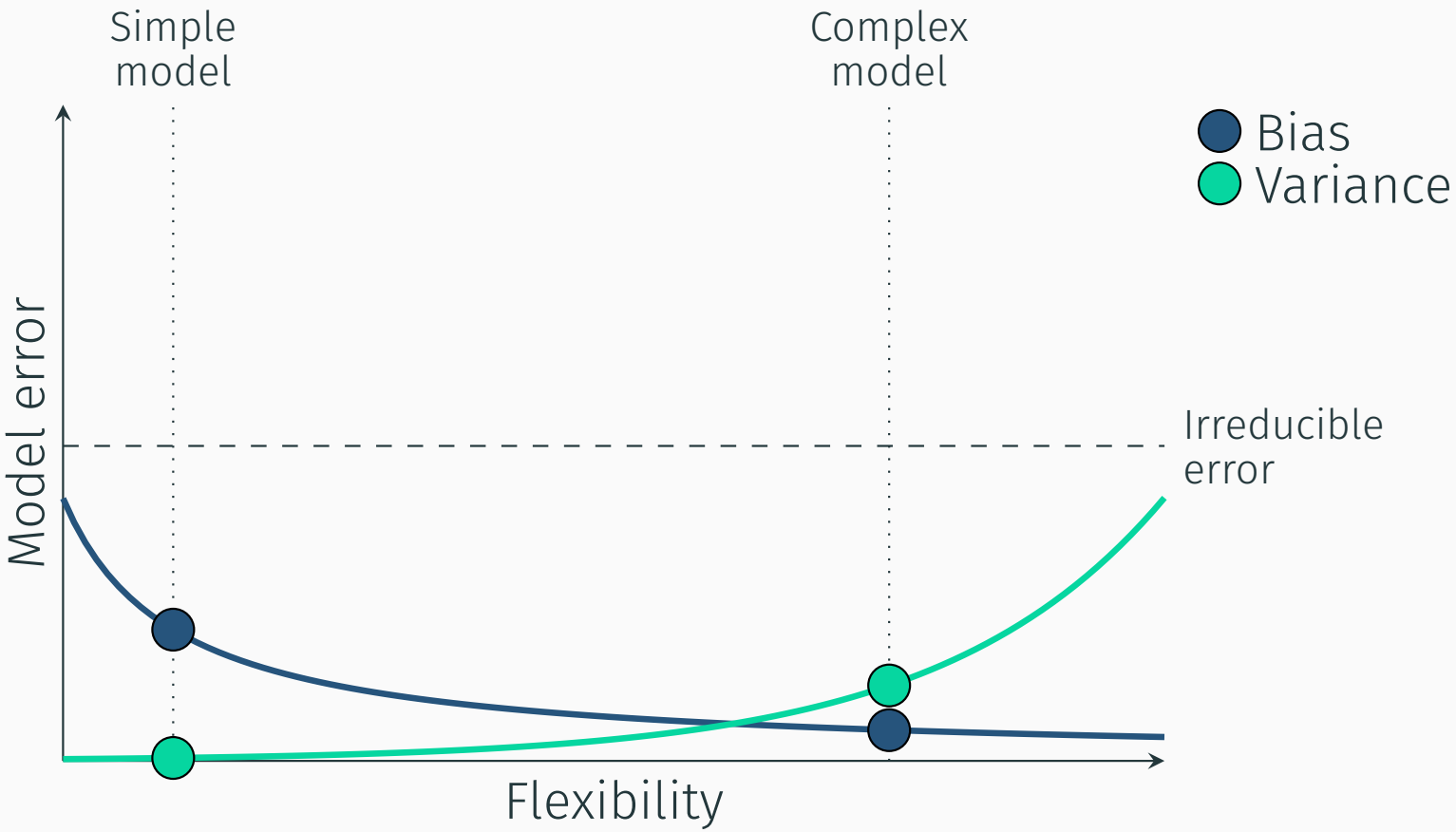


Introduction: The Bias-Variance Trade-off



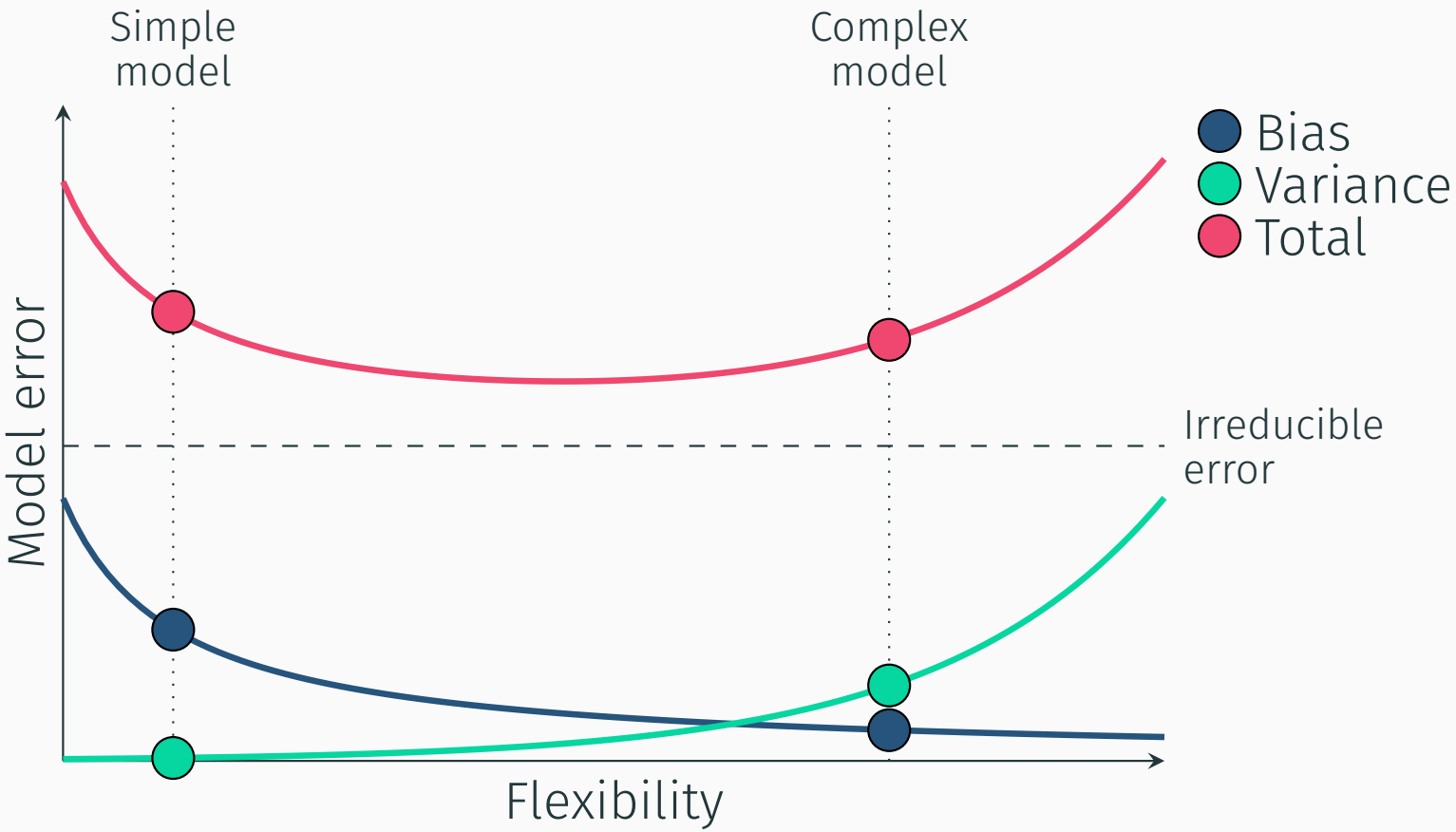
$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Var}(\hat{f}(x)) + \left[\mathrm{Bias}(\hat{f}(x))\right]^2 + \mathrm{Var}(\epsilon)$$

Introduction: The Bias-Variance Trade-off



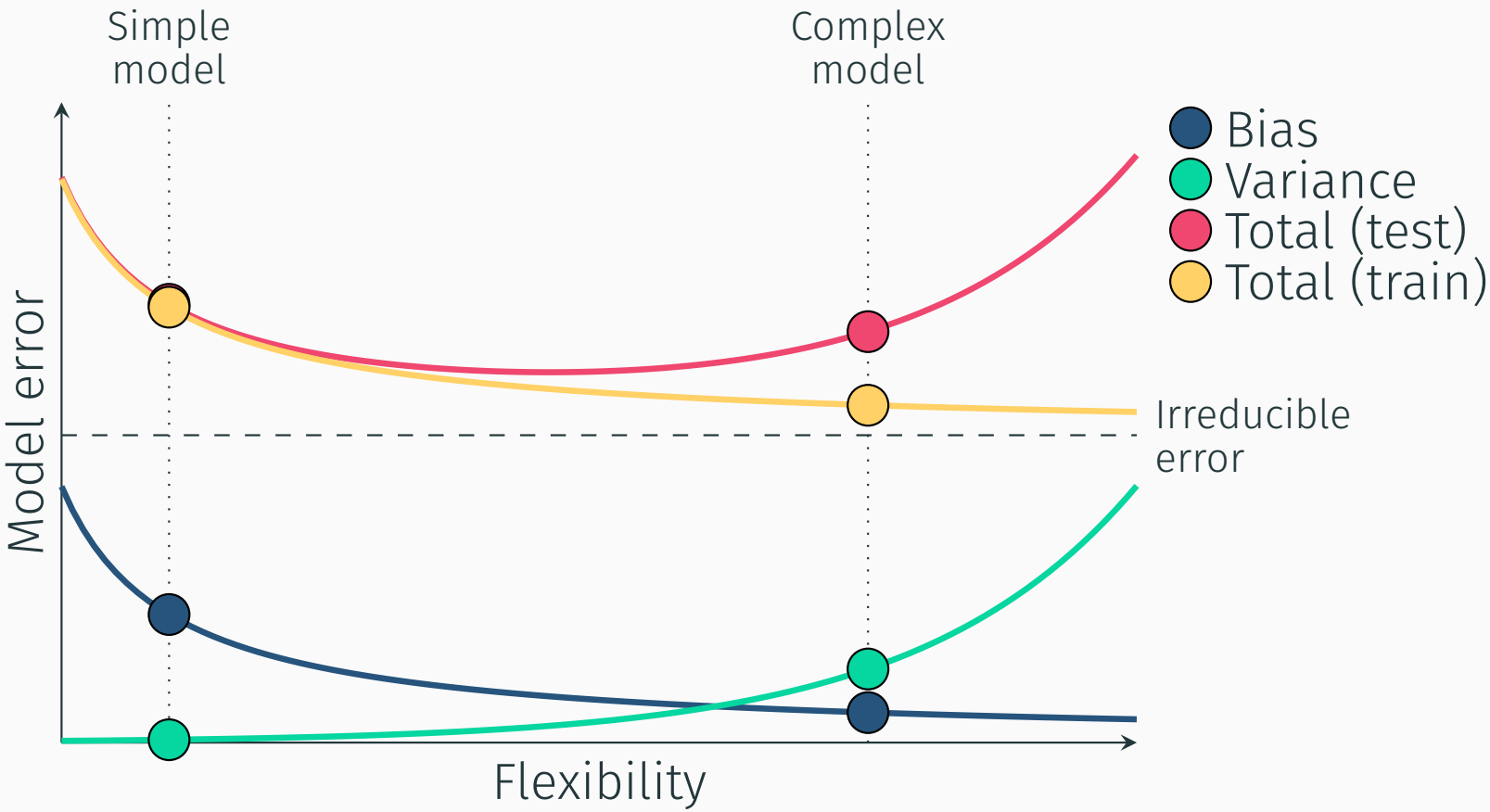
$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Var}(\hat{f}(x)) + \left[\mathrm{Bias}(\hat{f}(x))\right]^2 + \mathrm{Var}(\epsilon)$$

Introduction: The Bias-Variance Trade-off

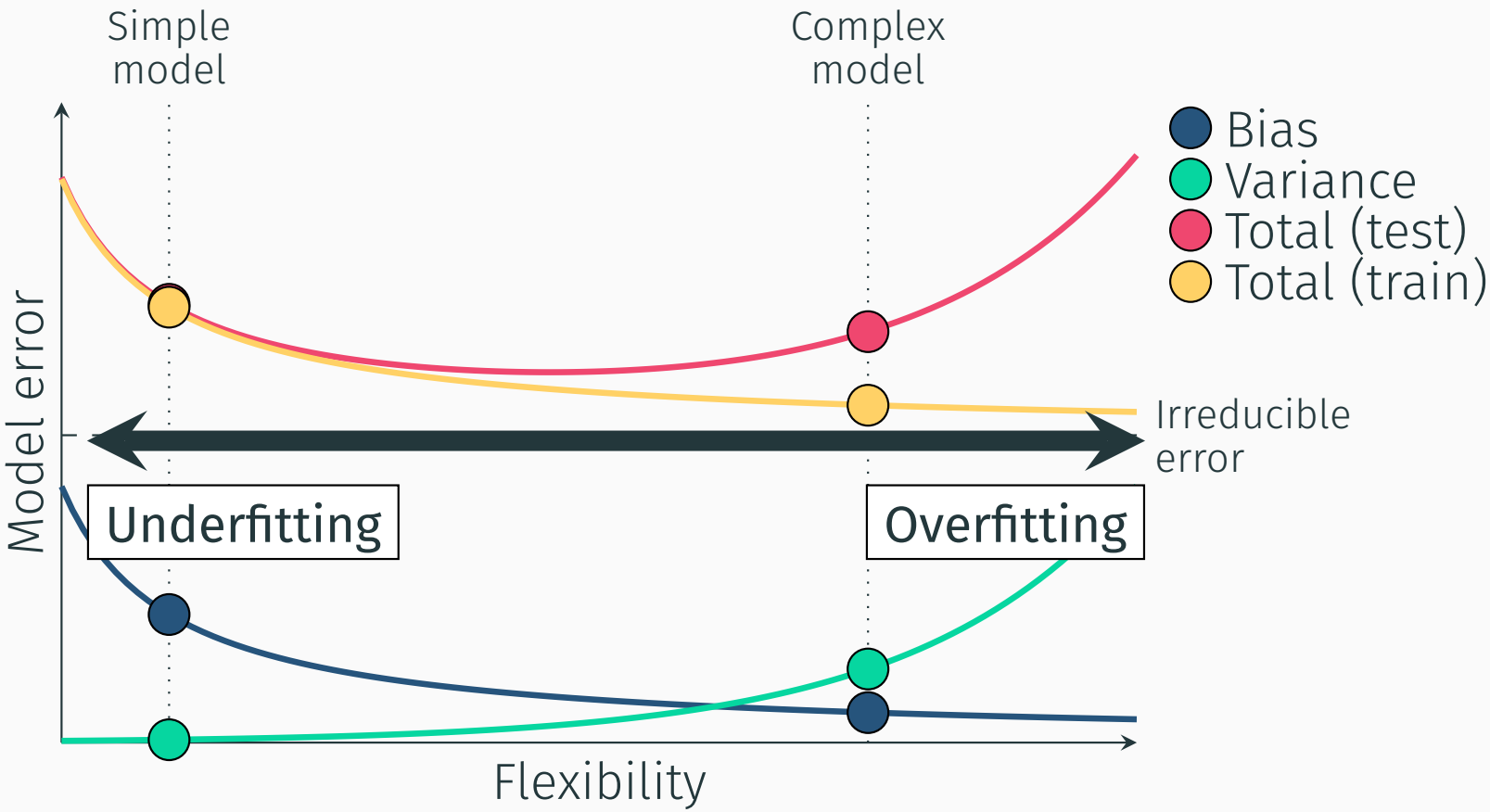


$$\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathrm{Var}(\hat{f}(x)) + \left[\mathrm{Bias}(\hat{f}(x))\right]^2 + \mathrm{Var}(\epsilon)$$

Introduction: The Bias-Variance Trade-off



Introduction: The Bias-Variance Trade-off

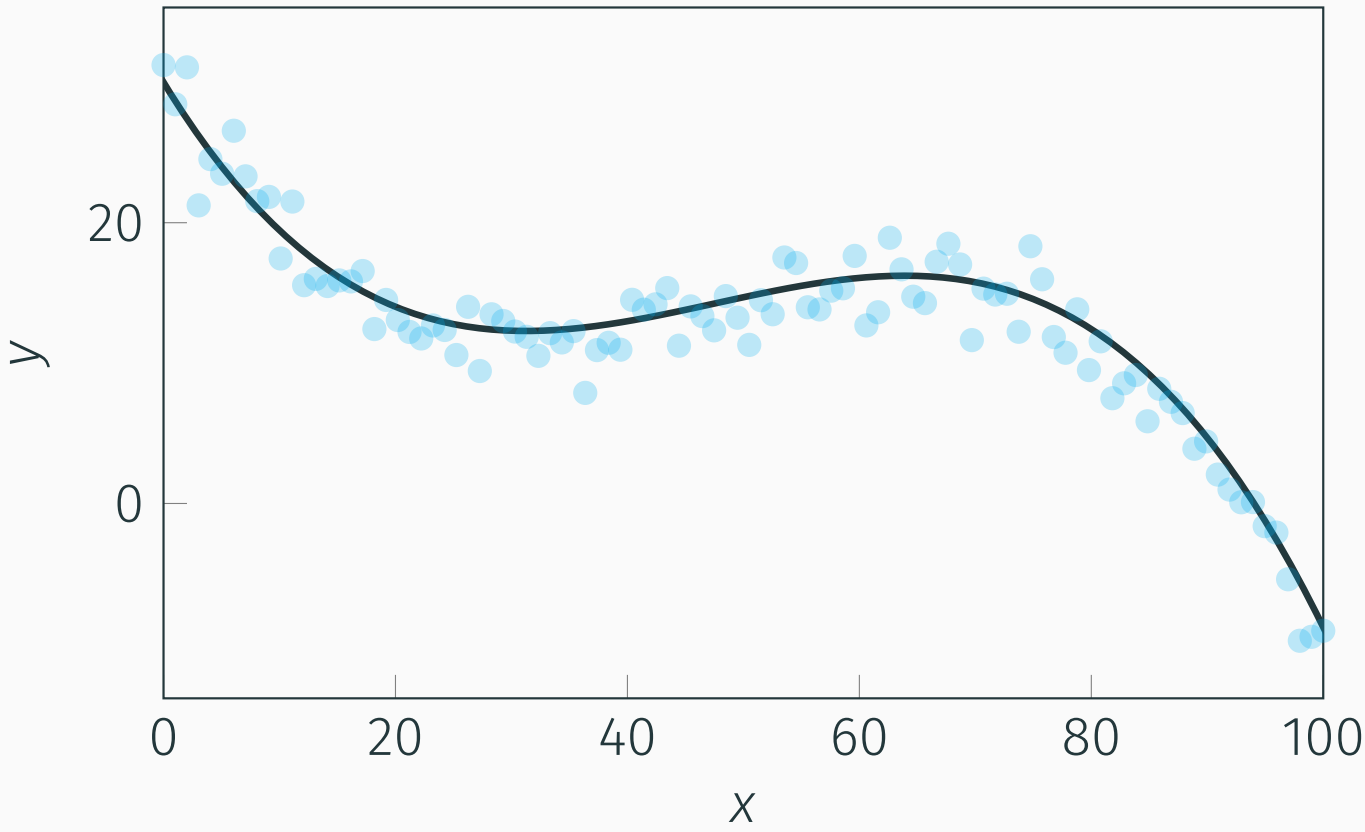


Introduction: The Bias-Variance Trade-off

The bias-variance trade-off lets us reason about why our model performs poorly

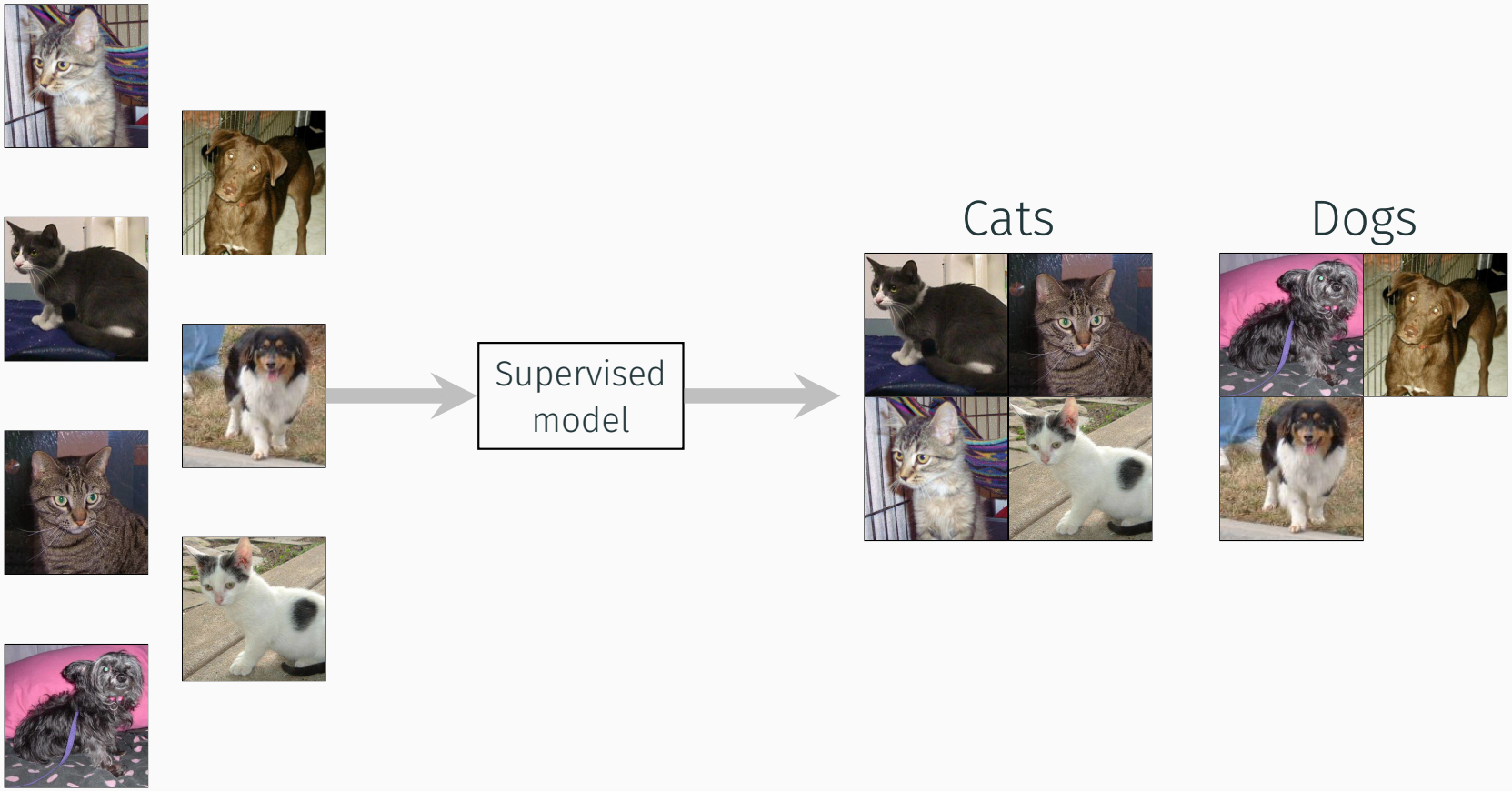
- If our model is too simple it will have high bias and low variance, which we can recognize by a high error in both the training and test set
- If our model is too complex it will have low bias and high variance, which we can recognize by a low error in the training set and a high error in the test set
- **This is way easier in theory than in practice**

Introduction: The Bayes Classifier



$$y = f(x) + \epsilon$$

Introduction: The Bayes Classifier

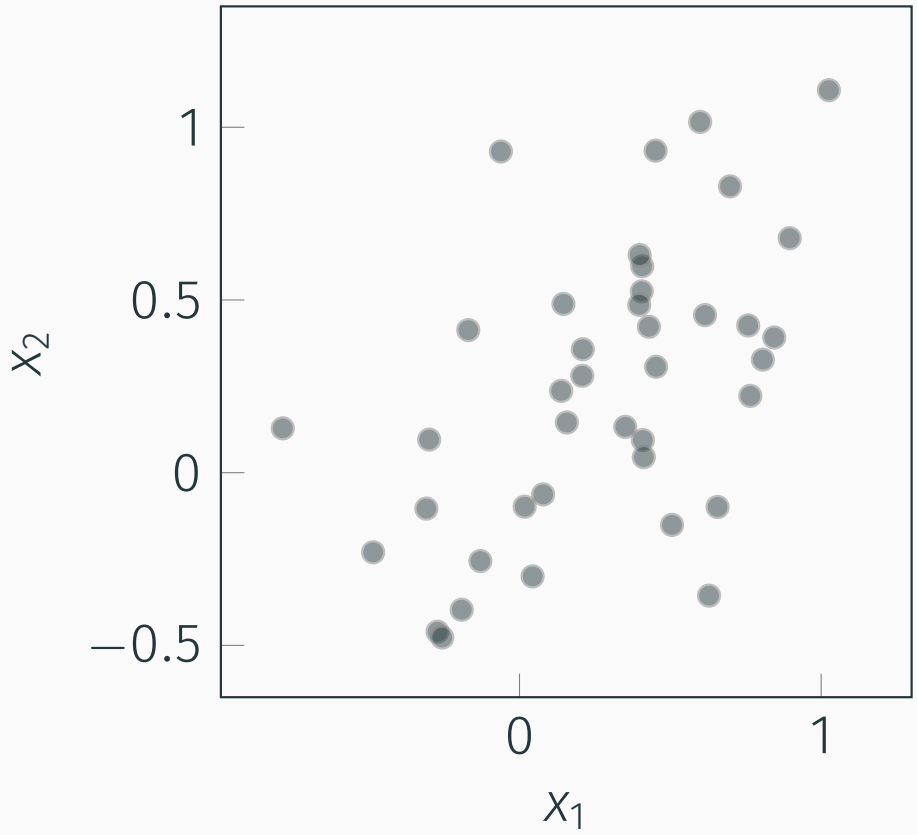


Introduction: The Bayes Classifier

x_1	x_2	y
-0.06	0.17	0
1.16	1.20	1
1.12	1.60	1
0.69	0.28	0
1.25	1.11	1

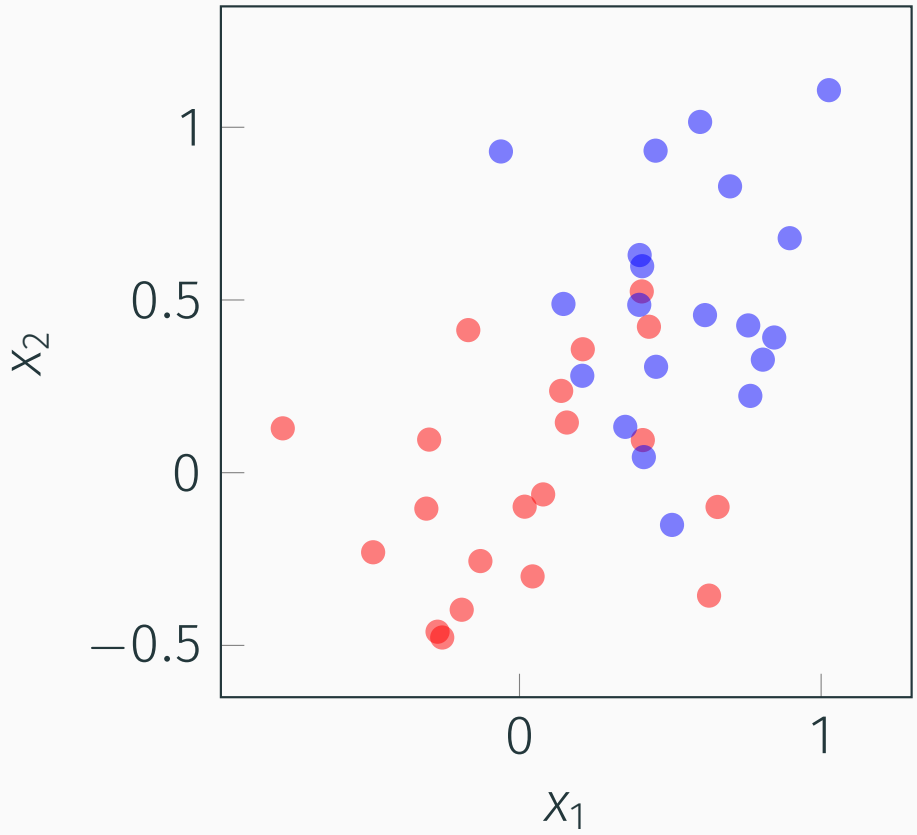
Introduction: The Bayes Classifier

x_1	x_2	y
-0.06	0.17	0
1.16	1.20	1
1.12	1.60	1
0.69	0.28	0
1.25	1.11	1



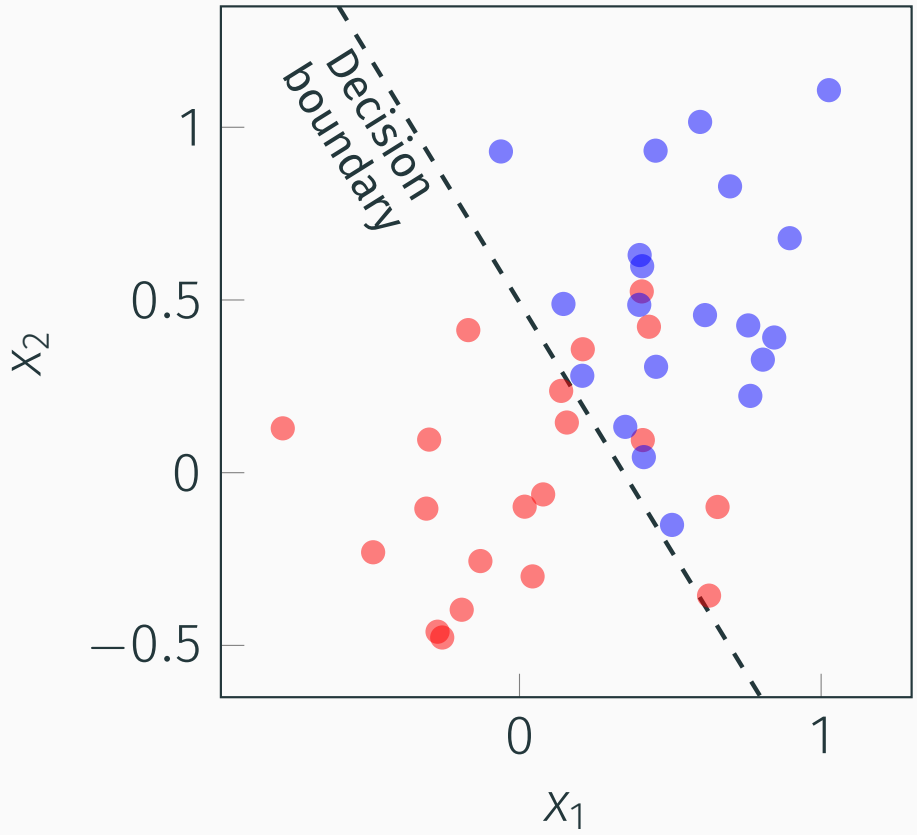
Introduction: The Bayes Classifier

x_1	x_2	y
-0.06	0.17	0
1.16	1.20	1
1.12	1.60	1
0.69	0.28	0
1.25	1.11	1



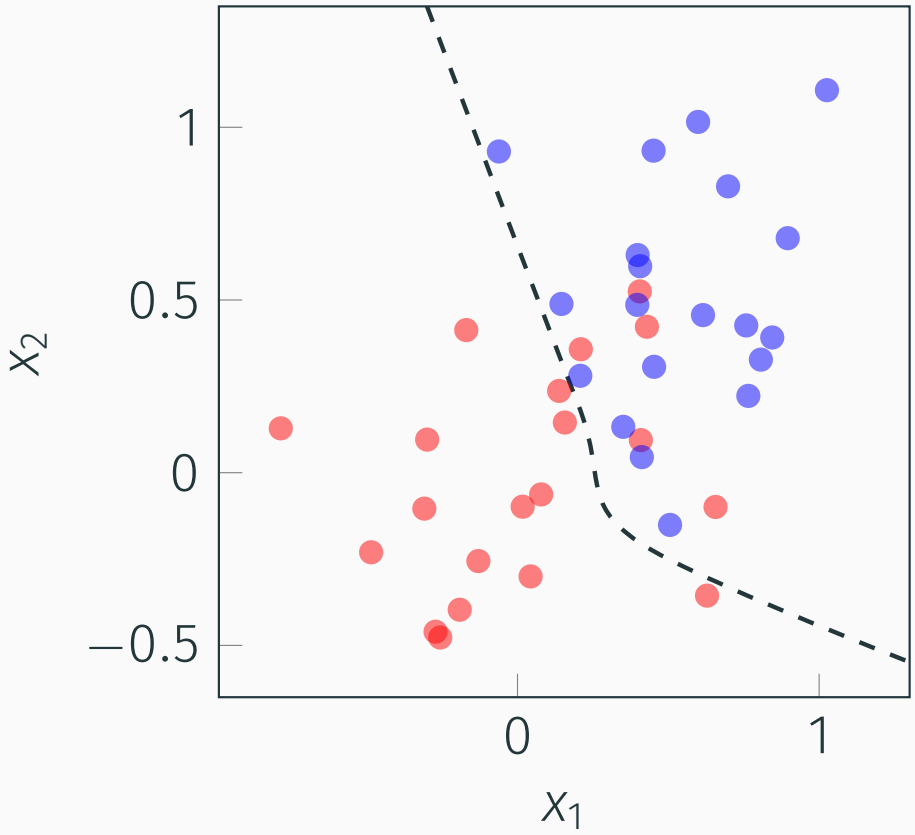
Introduction: The Bayes Classifier

x_1	x_2	y
-0.06	0.17	0
1.16	1.20	1
1.12	1.60	1
0.69	0.28	0
1.25	1.11	1



Introduction: The Bayes Classifier

x_1	x_2	y
-0.06	0.17	0
1.16	1.20	1
1.12	1.60	1
0.69	0.28	0
1.25	1.11	1



$$y = \underset{k \in \{0,1\}}{\operatorname{argmax}} \, Pr(y = k|x)$$

Introduction: The Bayes Classifier

Classification can be seen as the task of finding a decision boundary that best separates our classes in a high-dimensional vector space representing our predictors

- The Bayes classifier is the best classifier that can be theoretically achieved (although in practice we have to approximate it using data)
- The Bayes error rate quantifies the error of the best possible classifier

Assignment 1

Follow the steps in Chapter 2.3 of Introduction to Statistical Learning to make sure your system is set up correctly, whether you want to use Python or R. Then do the following exercises to make sure you understand:

- Create a vector of 100 standard normally distributed numbers and visualize them with a histogram.
- Show rows 5, 8, 9, and 10 of the Auto dataset.
- Show the last three columns of the Auto dataset.
- Show all cars with five cylinders in the Auto dataset.