Introduction to Machine Learning

MIT Center for
Transportation & Logistics

ctl.mit.edu

Introduction

Motivating questions

- What is machine learning?
- When do we need machine learning instead of using other techniques, such as regression?
- What are the different classes of machine learning algorithms?
- How do we use machine learning approaches to make inferences about new data?

Review of Regression

Predictions with linear regression

- Linear regression uses the values of one or more variables to make a prediction about the value of an outcome variable:
  - Input variables are known as independent variables
  - Output variable is known as dependent variable

- Example: predict the price of a car based on its weight, horsepower, and other characteristics

- Example: predict the profit of an order based on price of items, number of items, shipment distance, shipment weight, and other characteristics

Regression implementation and outputs

- Linear regression output includes coefficients for each independent variable, which is a measure of how much each independent variable contributes to the prediction of the dependent variable

- Linear regression output also includes metrics to assess how well the model fits the data
  - A model that fits the data well is more likely to make accurate predictions about new, unseen data

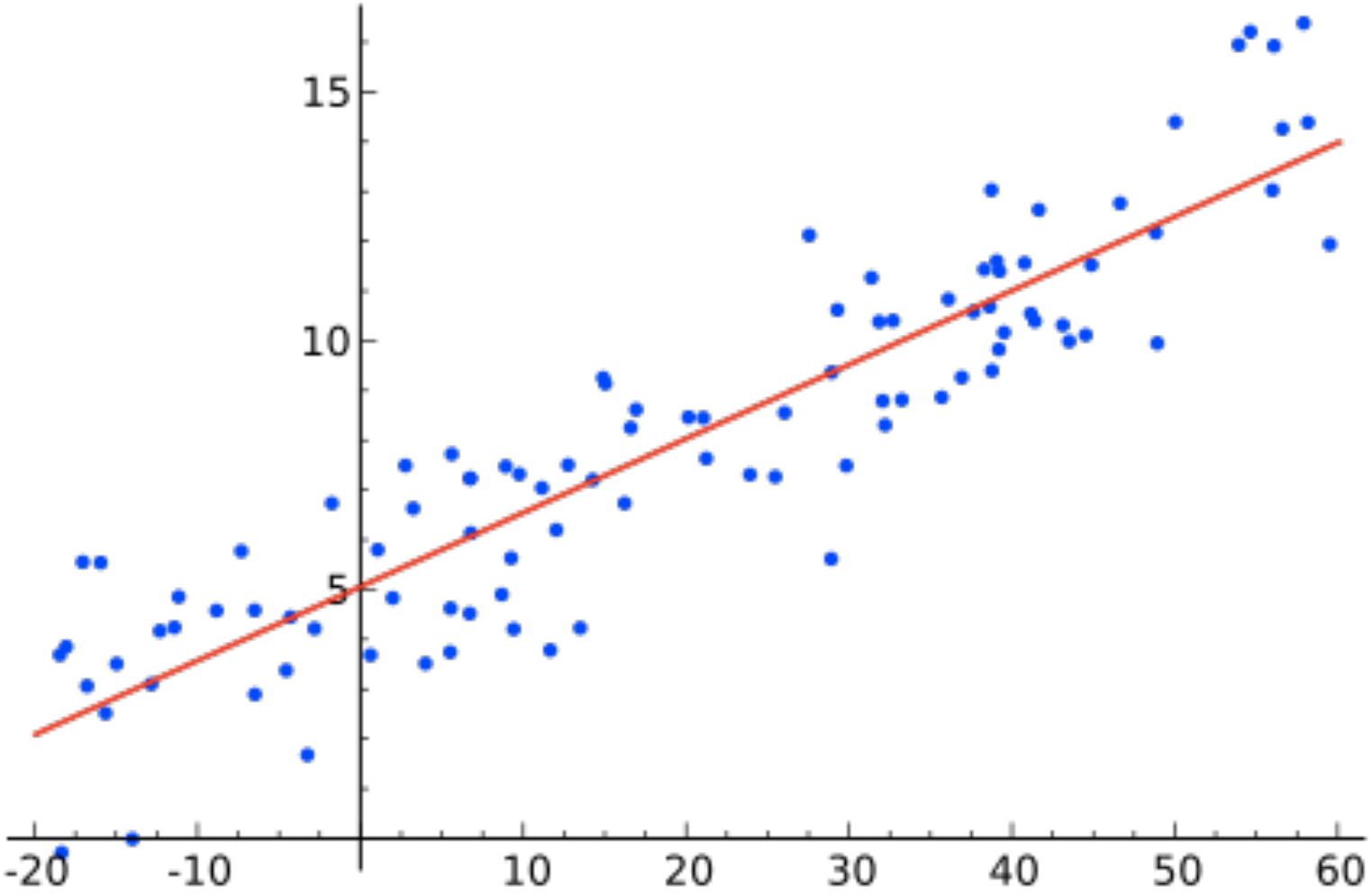
Linear regression on auto data to predict price

- The magnitudes of the estimated coefficients tell us which variables are most important for predicting price
- Negative coefficients are associated with variables which correspond to lower prices and vice versa
- Particular makes and features of engine design seem to be the most important features for predicting price

	Variable	Coeff Est ▲	Std Error	p
1	engine-type=dohcv	-25415.721	11058.794	0.200
2	num-of-cylinders=twelve	-4461.400	5831.537	0.600
3	engine-type=ohcv	-3498.092	2004.519	0.500
4	fuel-system=spfi	-2306.008	1574.598	0.600
5	num-of-cylinders=five	-2263.062	863.017	0.700
6	bore	-2188.458	3725.545	0.600
7	make=mitsubishi	-2101.298	1285.167	0.500
8	make=peugot	-1828.653	1687.988	0.400
9	make=dodge	-1813.266	953.866	0.700
10	make=plymouth	-1811.204	1022.126	0.600
11	make=chevrolet	-1418.635	1183.732	0.800
12	engine-type=dohc	-1289.387	1456.378	0.600
13	fuel-system=4bbl	-1227.601	1330.483	1.000
14	stroke	-1197.918	1173.616	0.700
15	make=mercury	-1029.732	774.592	0.900

How good is this regression model?

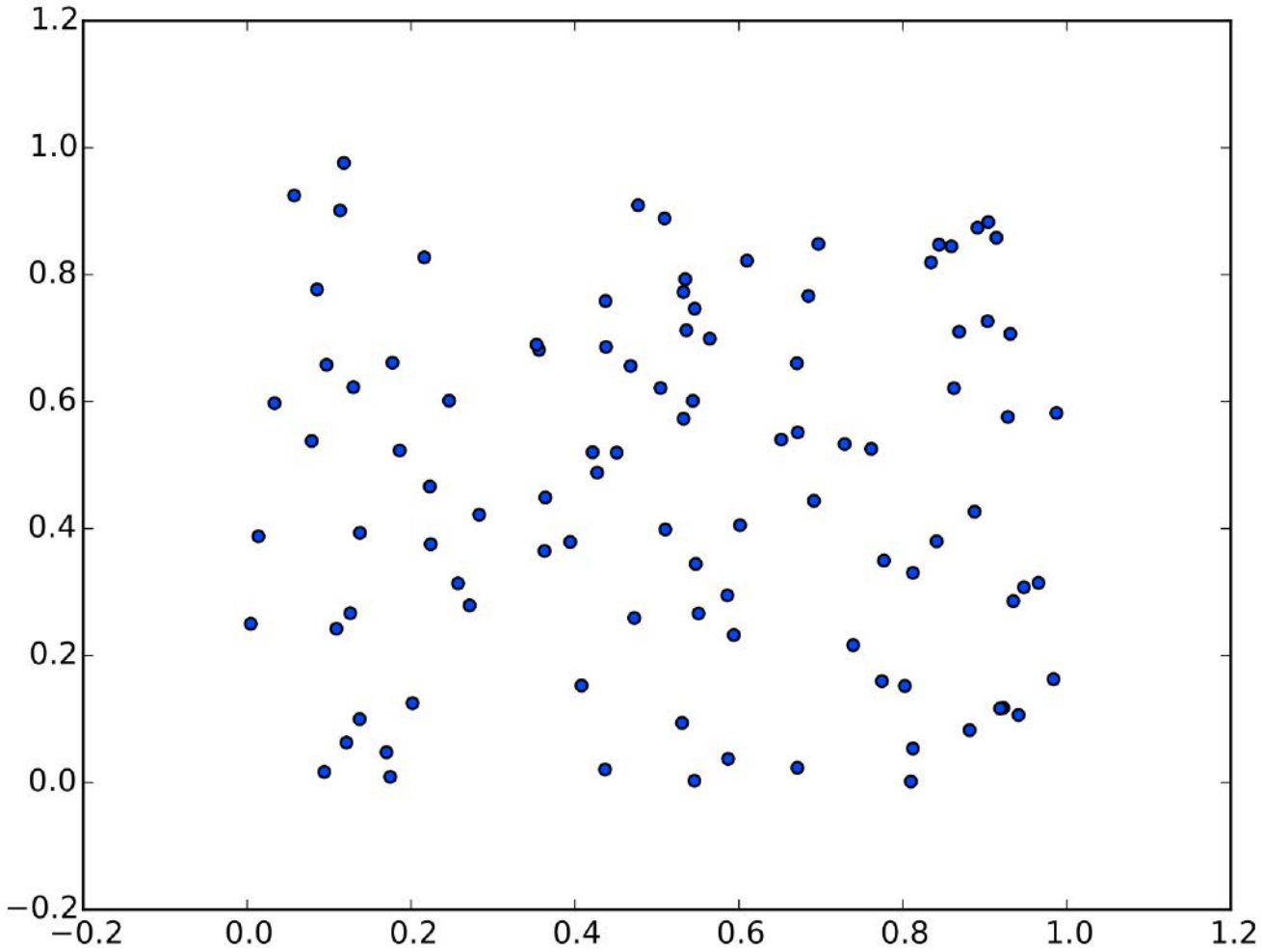
- Regression does a good job here, it would likely do a good job of predicting new, unseen data



Source: Wikipedia

How good is this regression model?

- Regression would not do a good job here



Source: Wikimedia

Key points from lesson

- Regression models estimate the importance of one or more independent variables in terms of predicting the output variable
- Using coefficients calculated from historic data, a regression model can be used make predictions about the value of the outcome variable for new records
- This estimation and prediction approach is the same as what is used when applying more advanced machine learning algorithms

Motivating demonstration

r2d3 demonstration

- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Overview of Machine Learning Algorithms

Classes of algorithms

- Classification: assigning records to pre-defined discrete groups
- Clustering: splitting records into discrete groups based on similarity
  - Groups are not known a priori
- Regression: predicting value of a continuous or discrete variable
- Association learning: observing which values appear together frequently

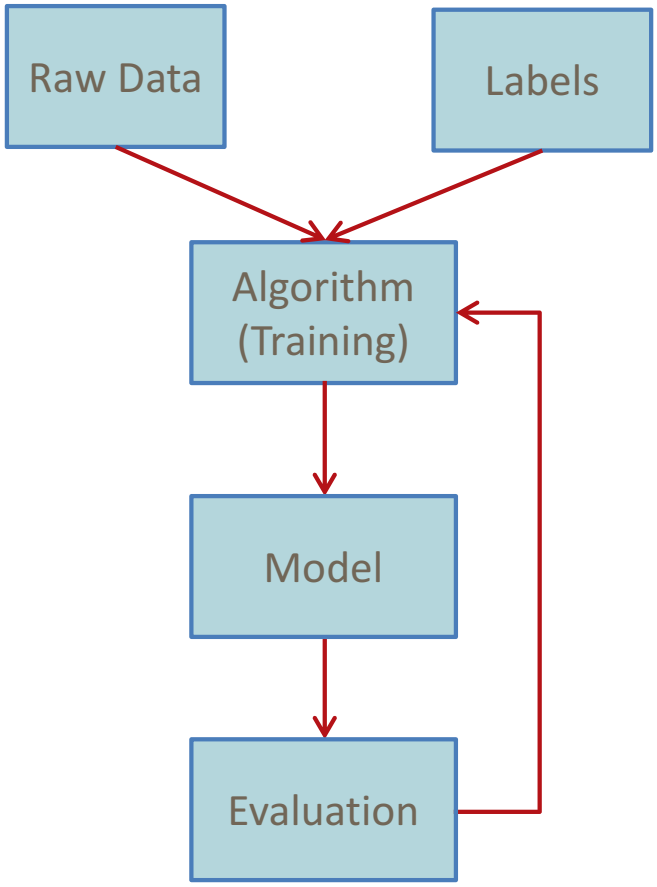
Supervised vs. Unsupervised

- Supervised learning
  - Correct answer, label, is known in the training data
  - Label is applied by a person or already exists
  - Labeled data are used to train an algorithm using feedback
  - Apply or test the trained model on new, unseen data to predict the label

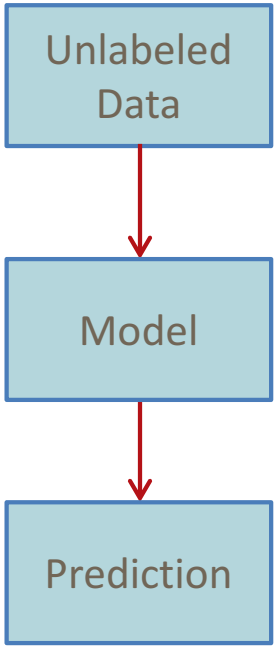
- Unsupervised learning
  - Finds previously unknown patterns in the data without any labels or guidance
  - No training/testing/validation process because correct answer is unknown

Supervised learning workflow

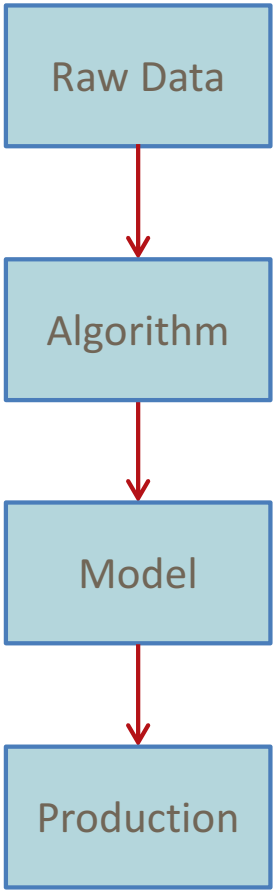
Training



Making Predictions

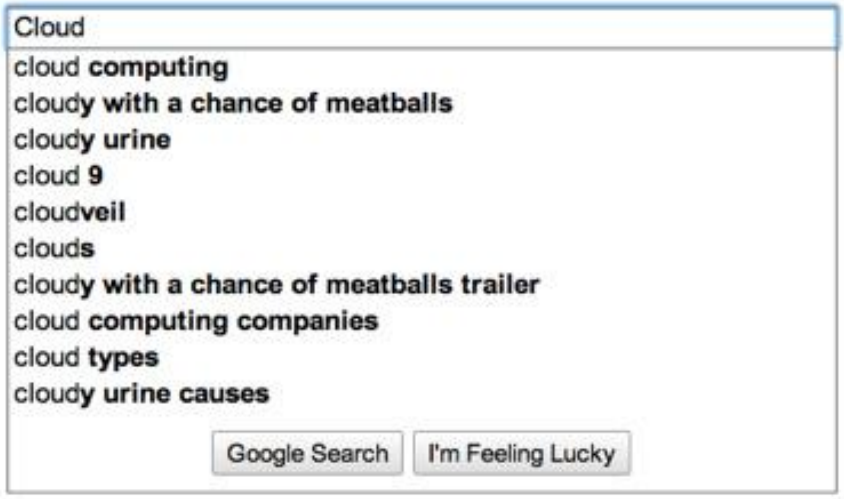


Unsupervised learning workflow



Unsupervised examples

- Fraud and anomaly detection
- Association learning



Source: Wikimedia and flickr

Key points from lesson

- Supervised learning uses outcome variables, known as labels, for each record to identify patterns in the input variables or features related to the outcome variable

- In unsupervised learning, the outcome variable values are unknown, therefore relationships among the input variables are used to identify patterns or clusters of records

- Machine learning algorithms are primarily used to make predictions or learn about new, unlabeled data

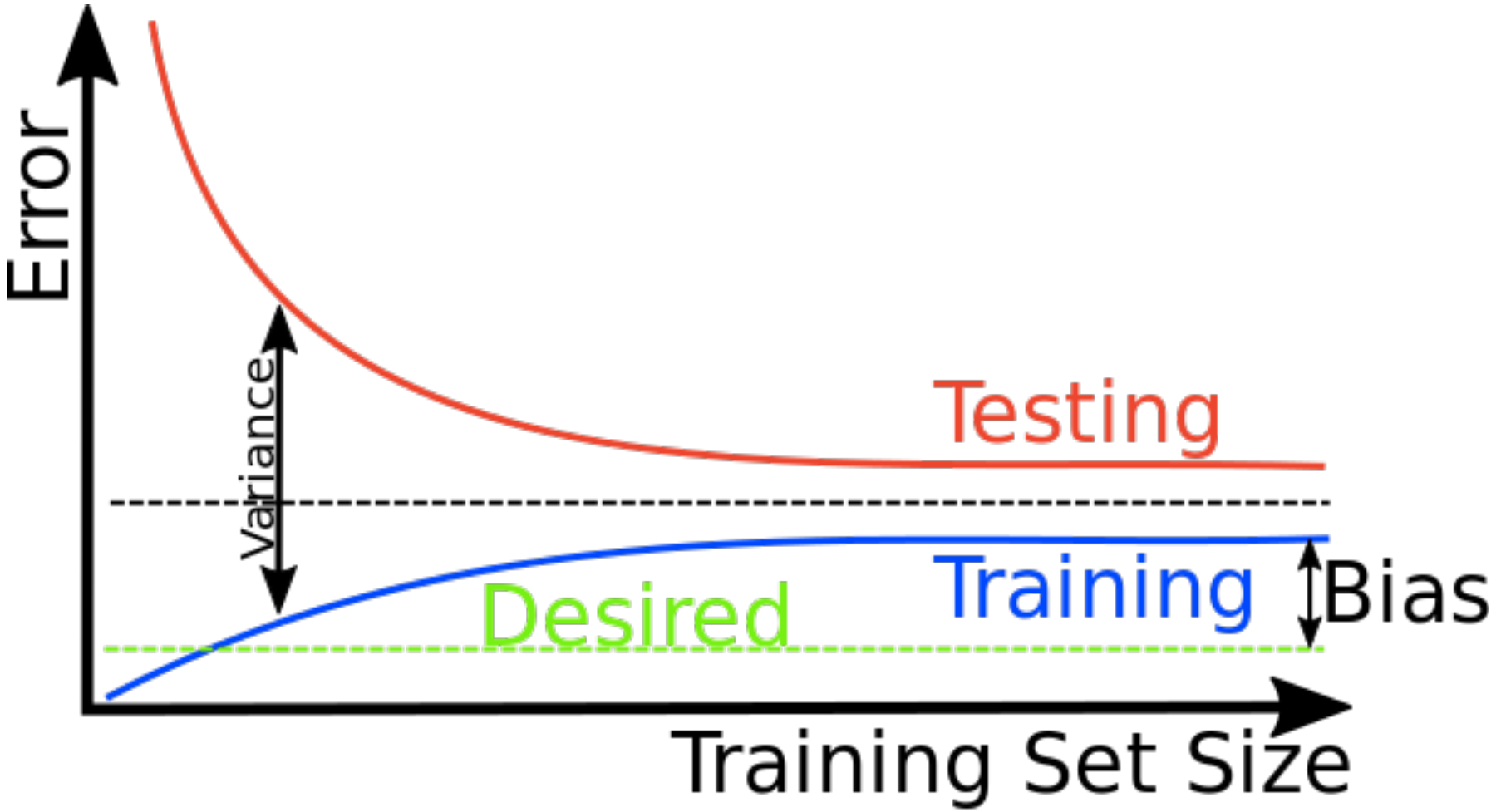
Model quality

Bias and variance

- Bias quantifies the lack of ability of a model to capture the underlying trend in the data
- Variance quantifies a model's sensitivity to small changes in the underlying dataset
- Ideally want low bias and low variance, but there is a tradeoff between the two quantities
- More complex models decrease bias but tend to increase variance

Representative training sample

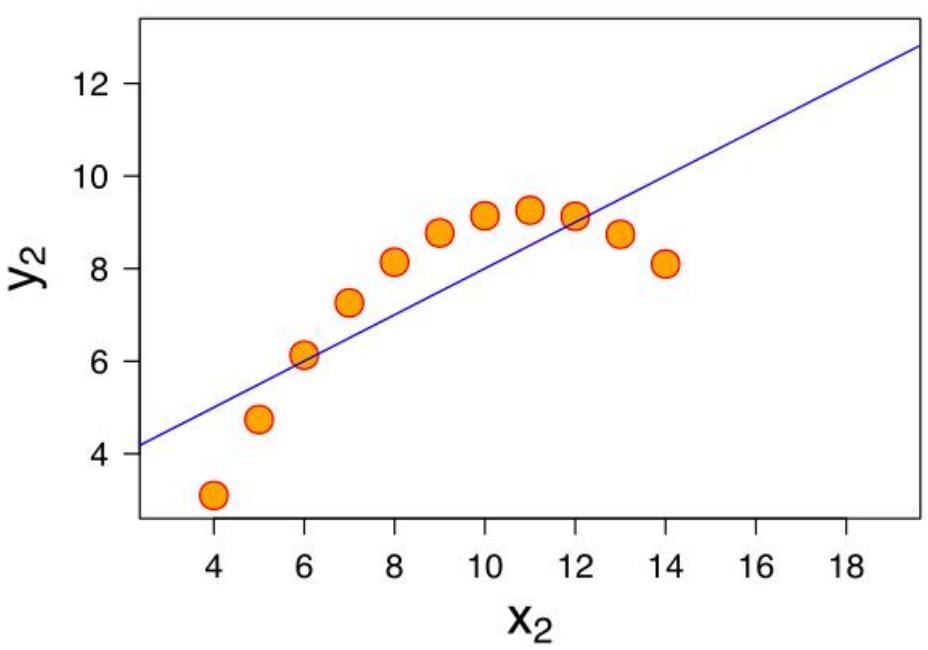
- A large and representative sample of the labeled data should be used to train the model, the remainder is used for testing

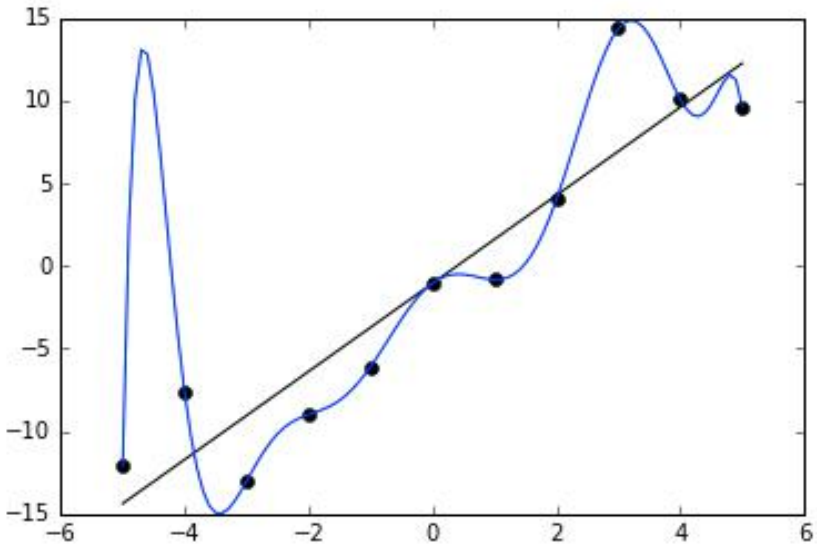


Source: Wikimedia

Overfitting versus underfitting

- Underfitting - model is too simple, high bias and low variance
- Overfitting – model is too complex, low bias and high variance



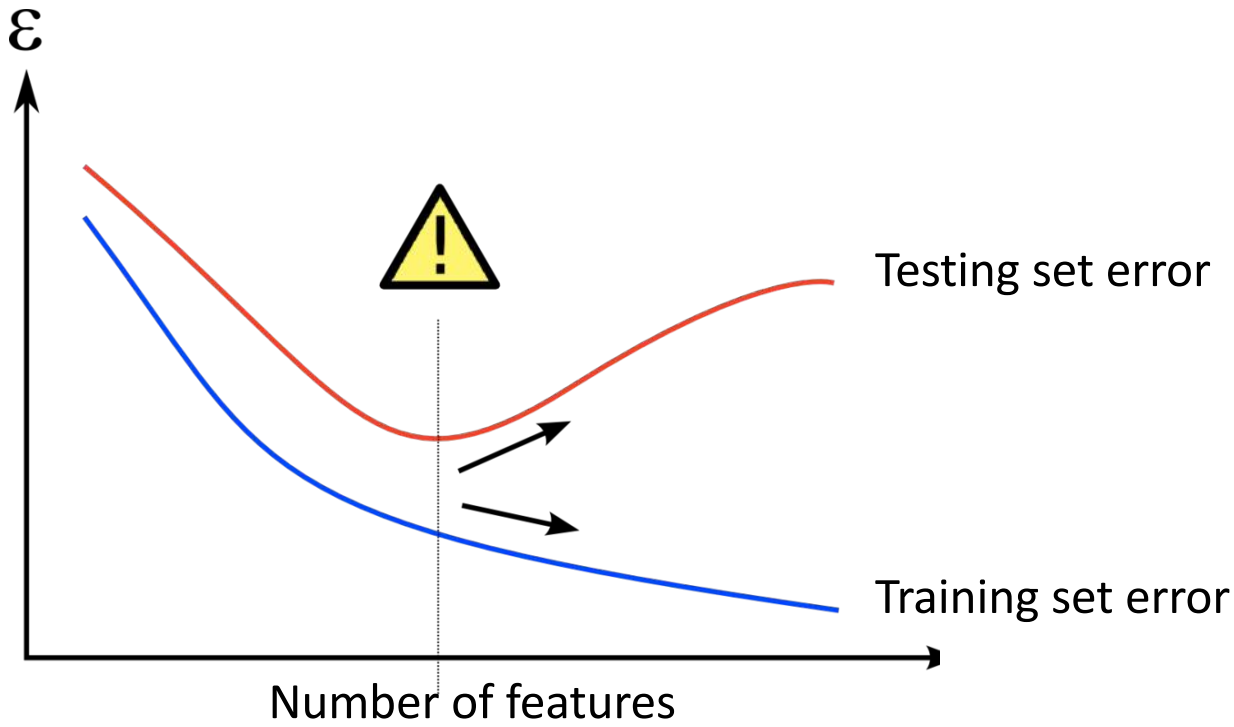


- Want something that is just right, generalizable to new data

Source: Wikimedia

Overfitting and model complexity

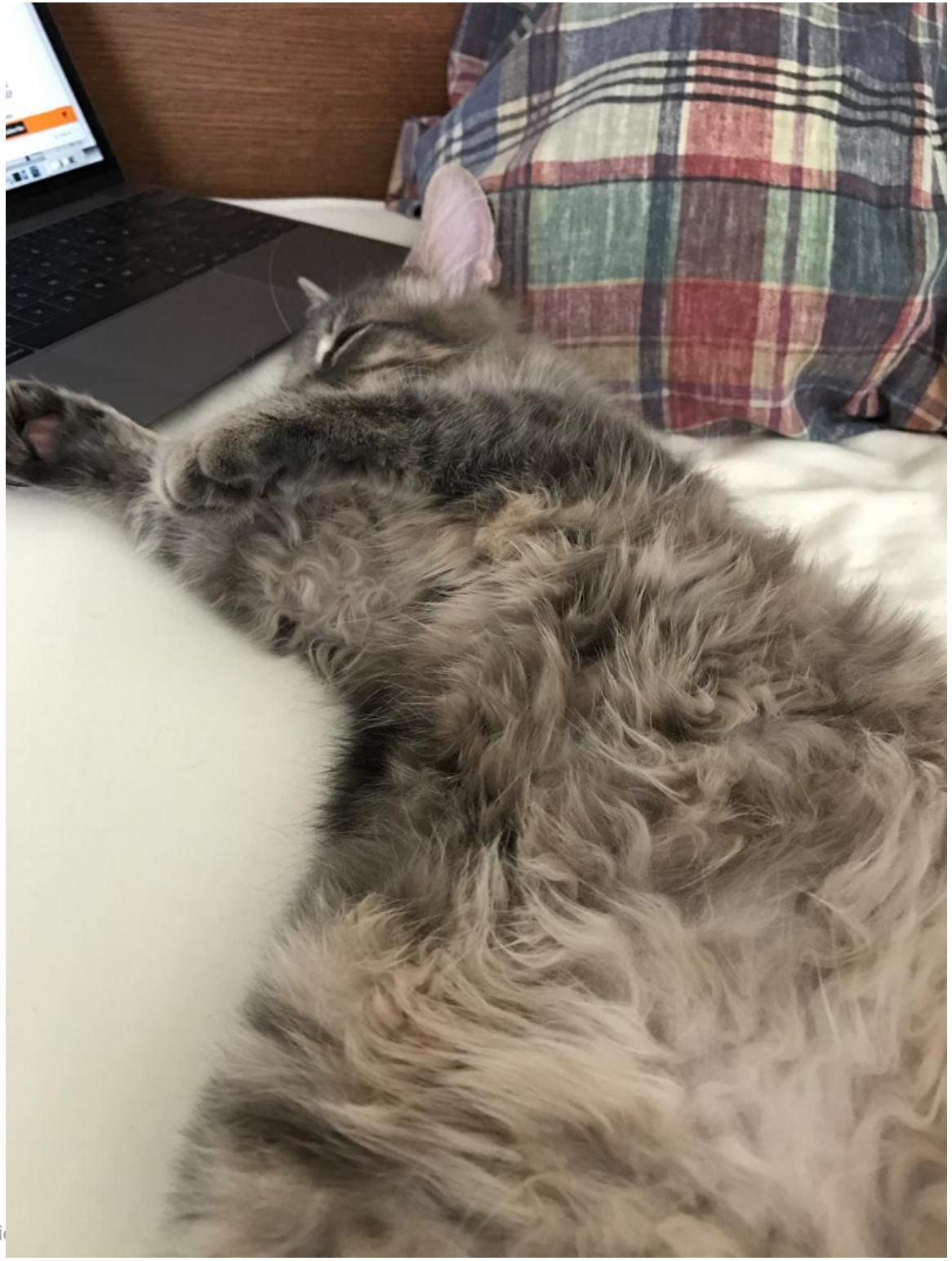
- Overfitting is a more common pitfall
- Tempting to make models fit better by adding more features
- Results in a model that is incapable of generalizing beyond the training data



Source: Wikimedia

Key points from lesson

- Machine learning models should be trained on an unbiased set of data that is representative of the variance in the overall dataset
- If there is a bias in the training data or if too many features are included in a model, the model is at risk of being overfit
- In overfit models, the coefficients, known as parameters, will not be generalizable enough to make good predictions for new records



Machine Learning in Action

http://en.akinator.com/

- https://en.wikipedia.org/wiki/File:Credit_card.jpg
- https://www.flickr.com/photos/cote/4346500259
- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- http://en.akinator.com/