Intro to Machine Learning (CS436/CS580L)

Lecture 2: Model Representation & Cost Function

Xi Peng, Fall 2018

Thanks to Tom Mitchell, Andrew Ng, Ben Taskar, Carlos Guestrin, Eric Xing, Hal Daume III, David Sontag, Jerry Zhu, and Tina Eliassi-Rad for some slides & teaching material.

This Class

- Graphical Example
  - Classification
  - Regression
  - Clustering
- Model representation
- Loss function

Median

Types of Machine Learning

- Supervised Learning (SL)
- Unsupervised Learning (UL)
- Reinforcement Learning (RL)

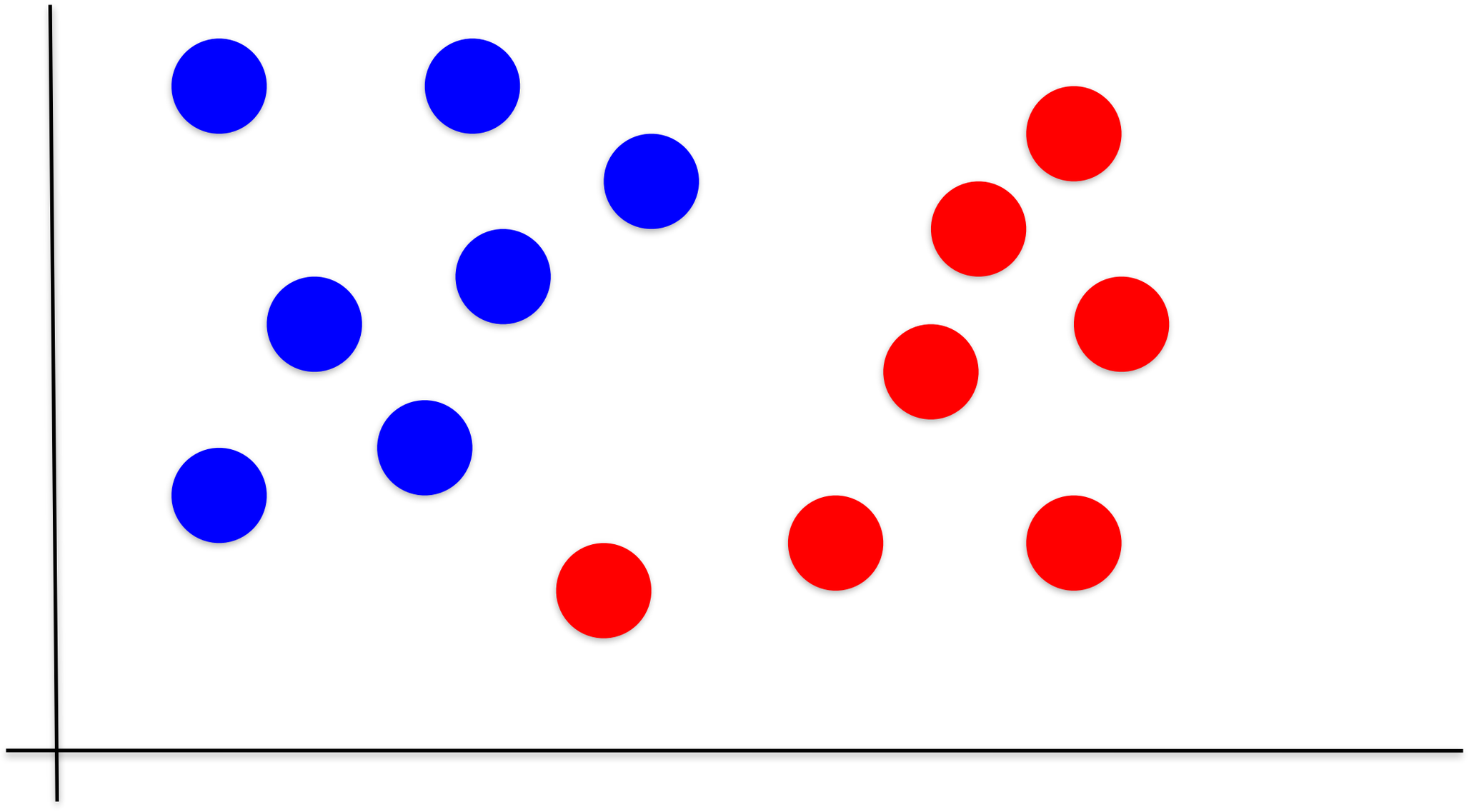
Types of Machine Learning

Supervised Learning

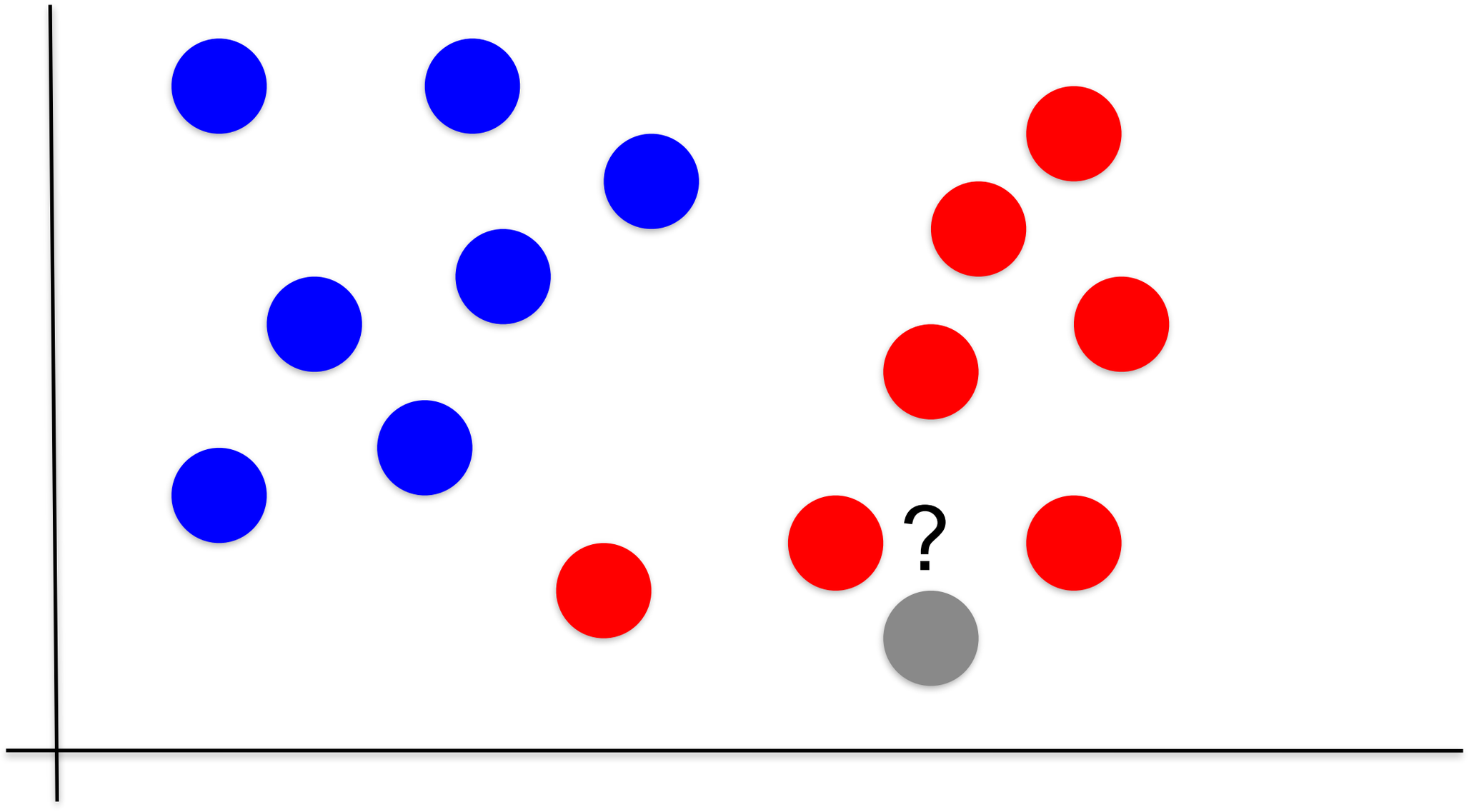
- Observe *n* training examples $D=\{x_i, y_i\}$
- Learn a function, *h*, mapping any input *x to output y*: $h(x) \approx y$

- Classification: predict a small number of discrete-valued outputs
- Regression: predict a continuous-valued output
- Ranking
- …

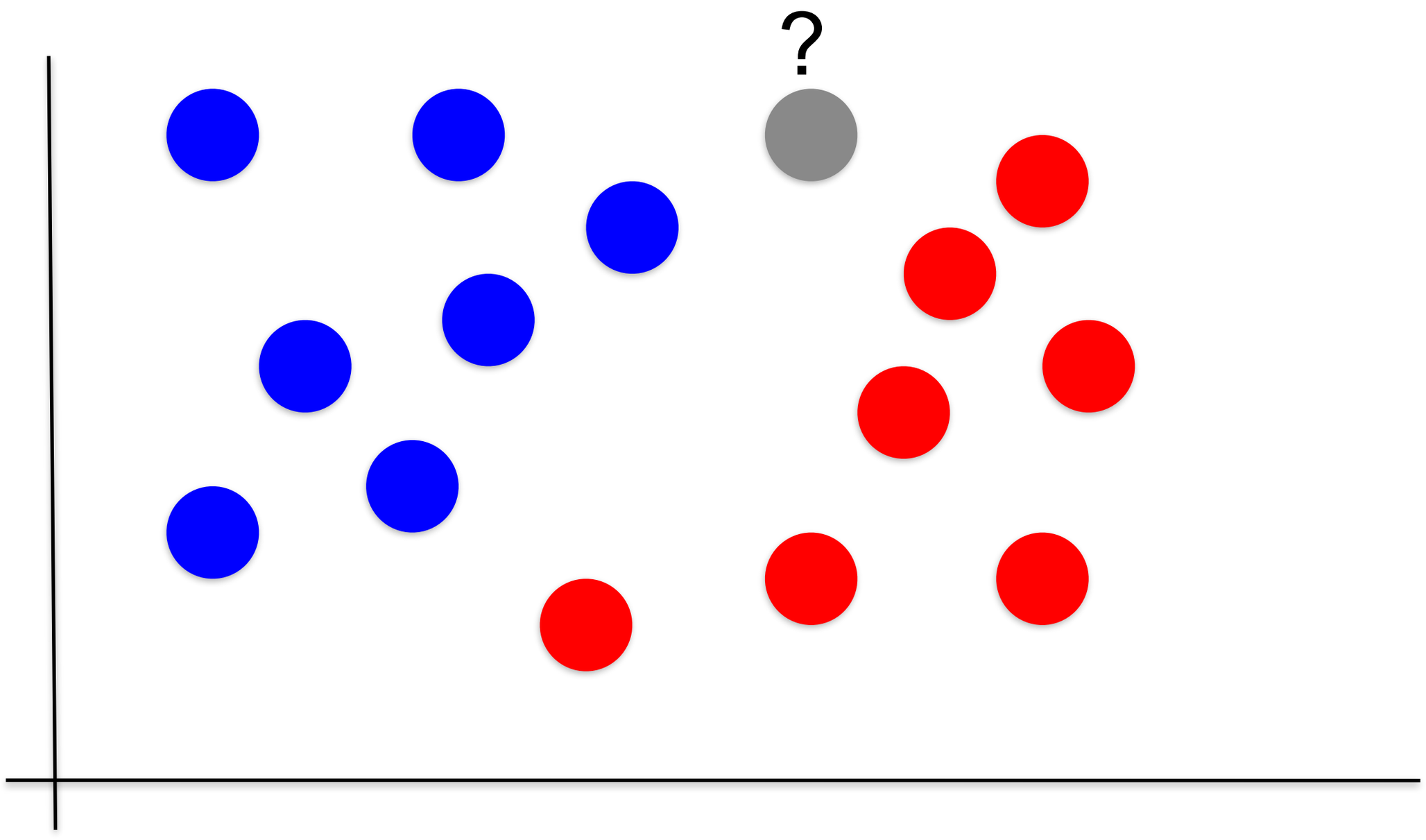
Graphical Example of Classification



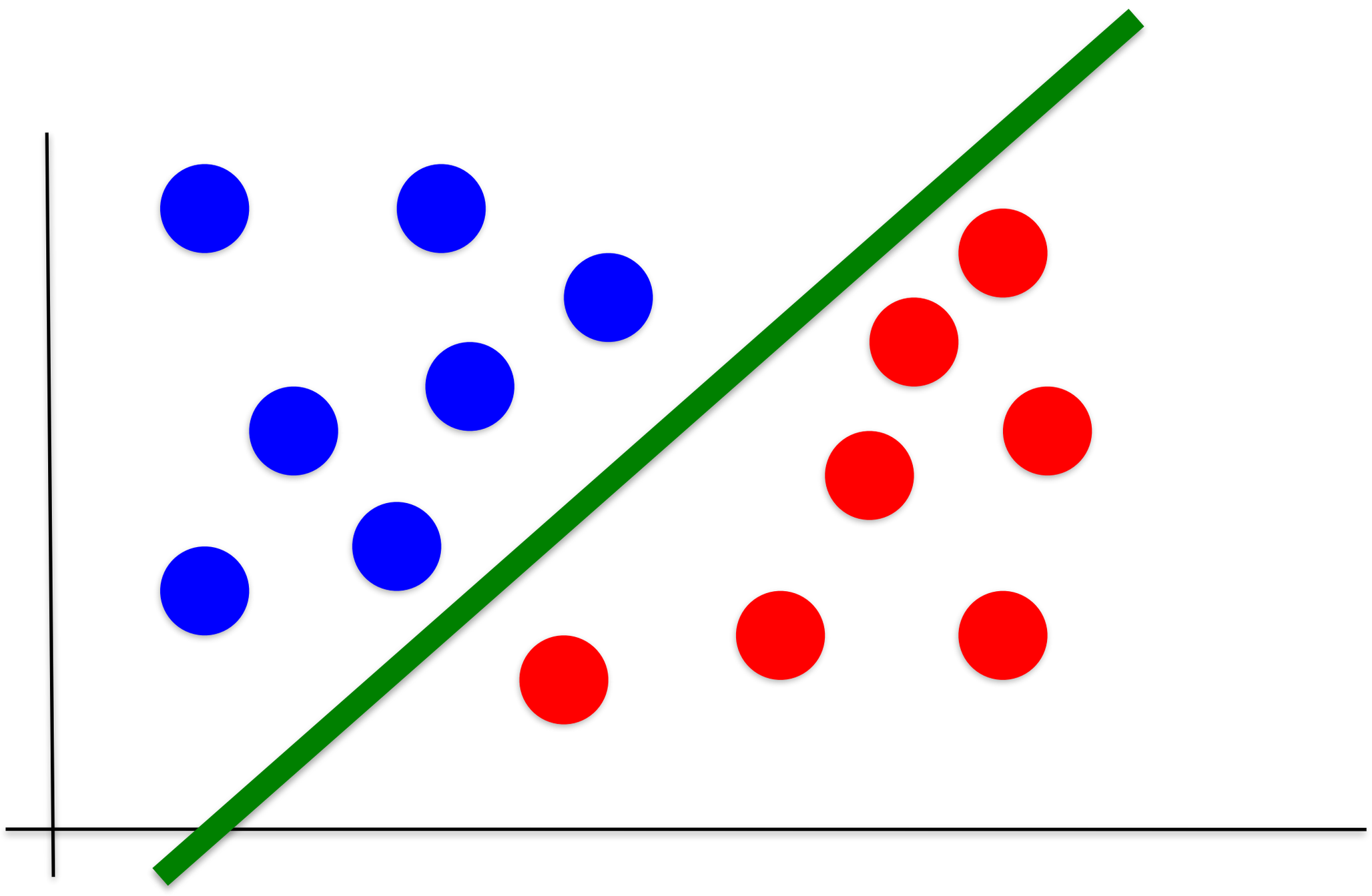
Graphical Example of Classification



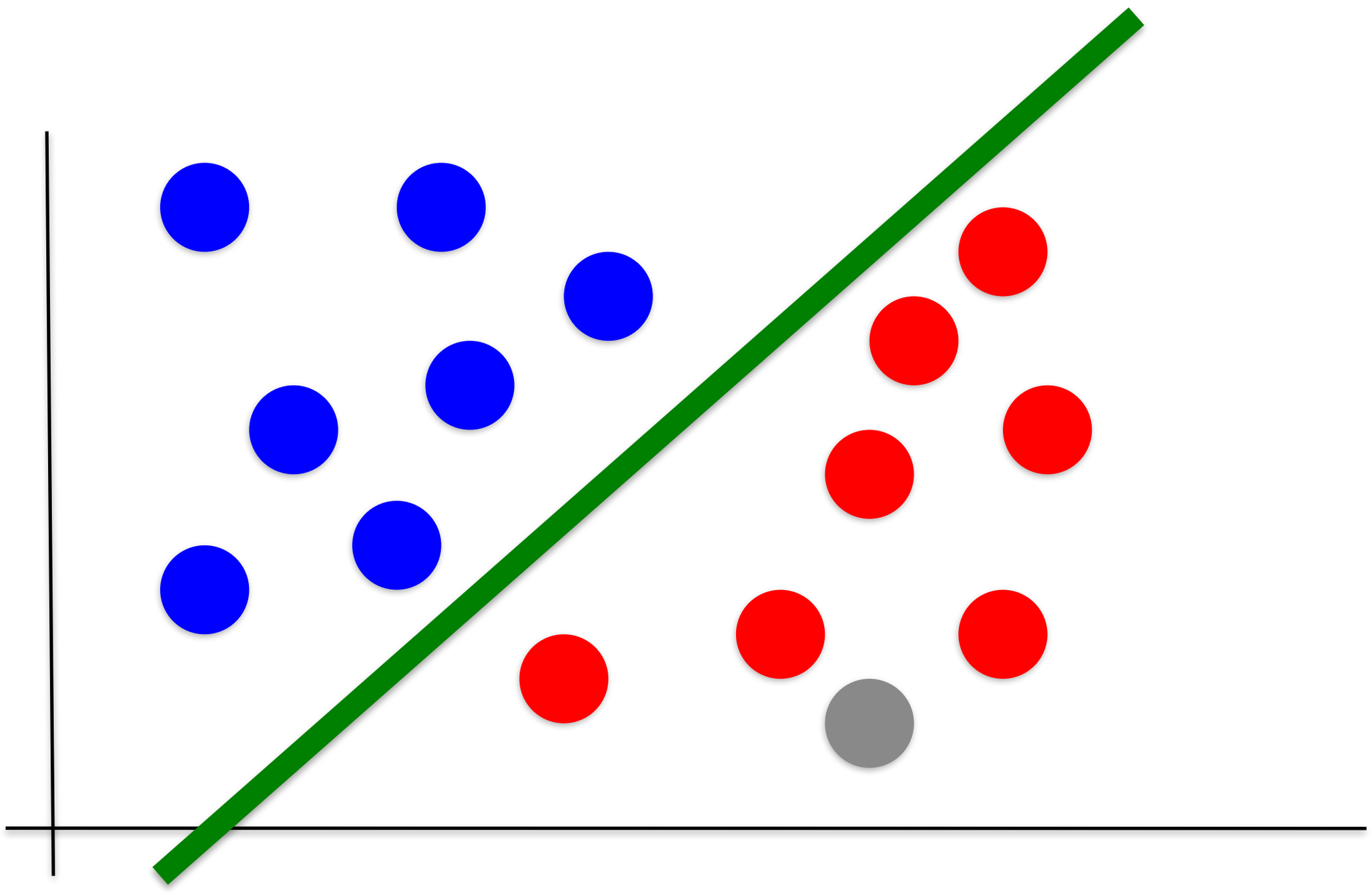
Graphical Example of Classification



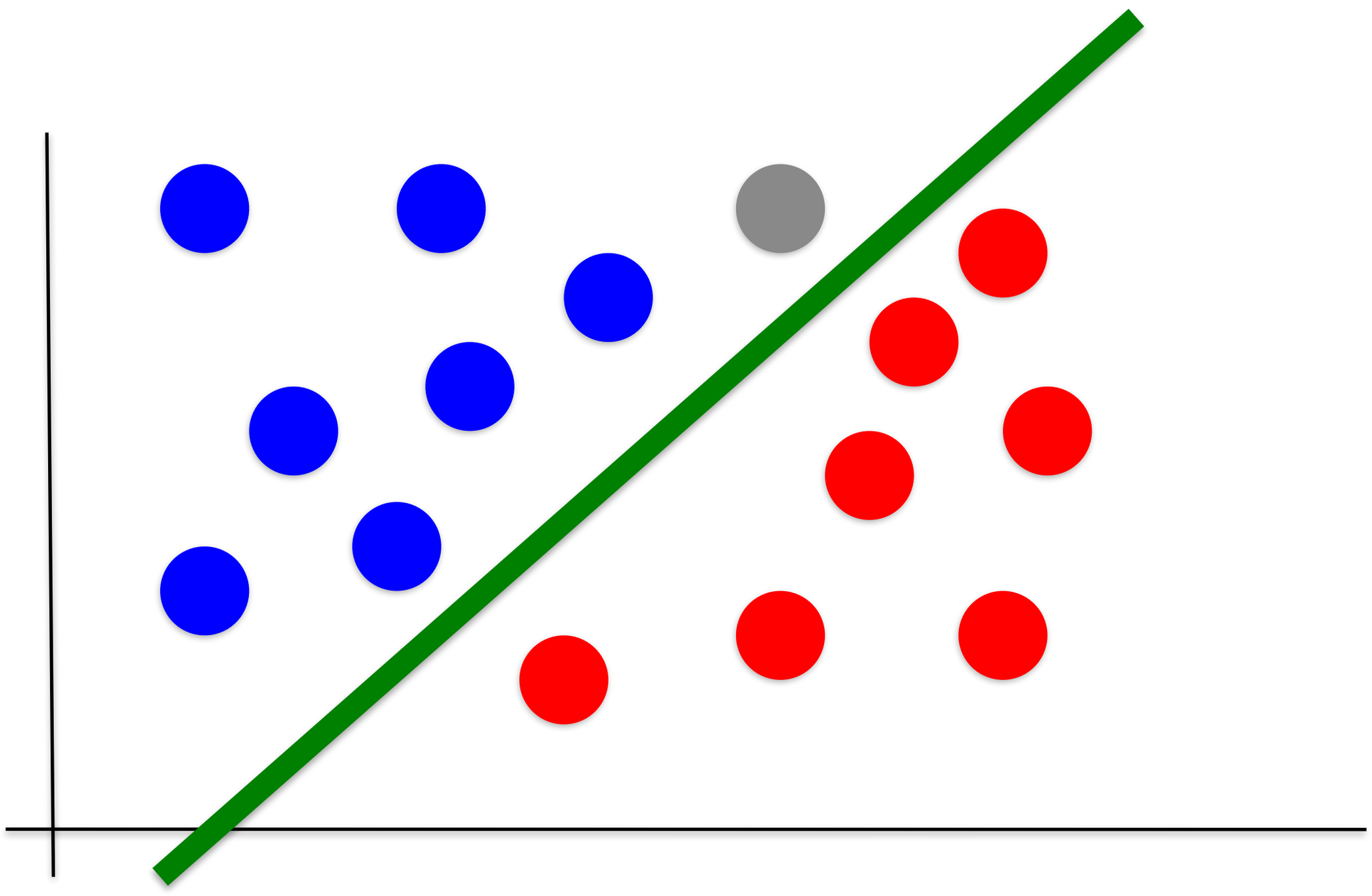
Graphical Example of Classification



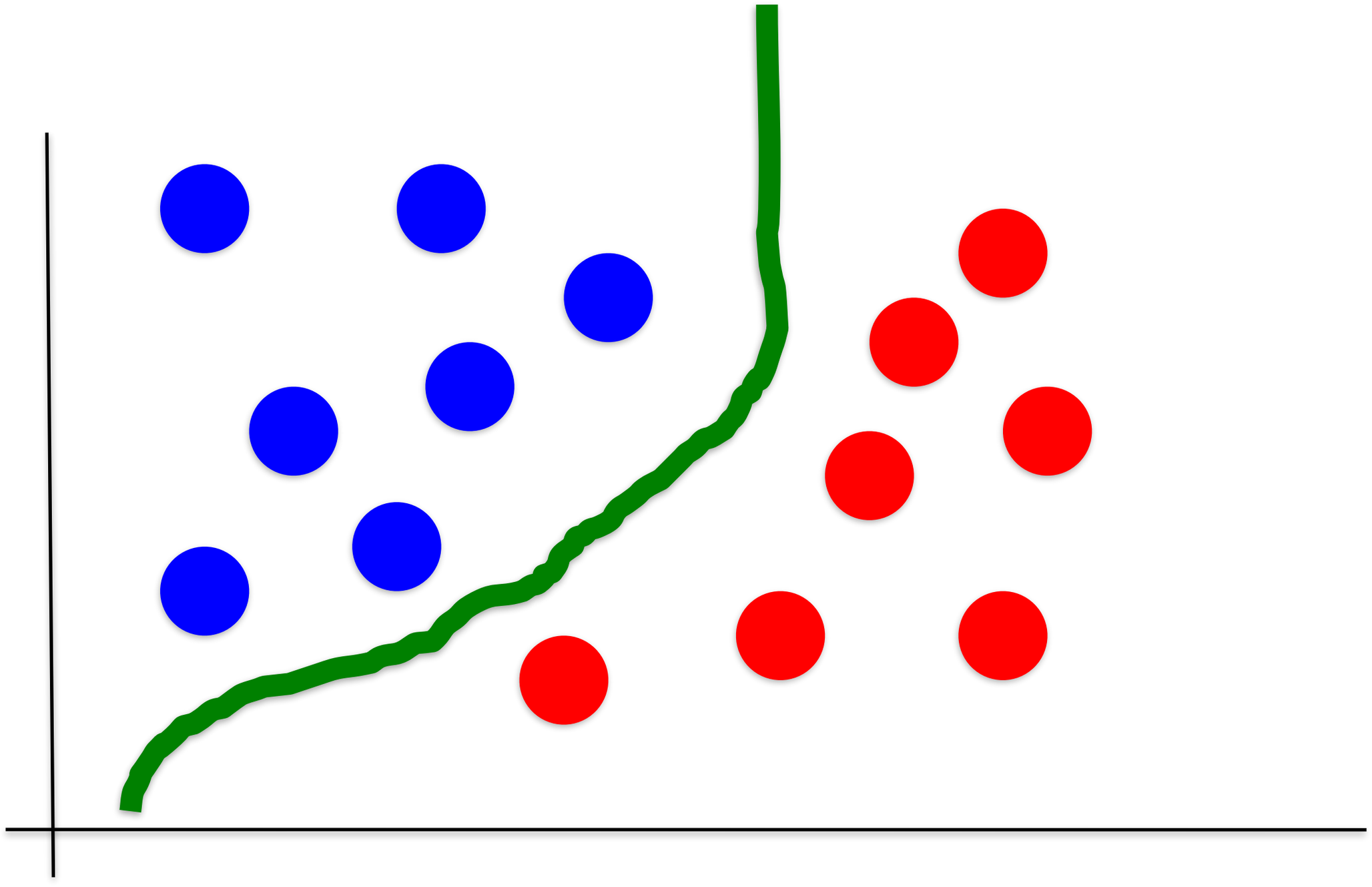
Graphical Example of Classification



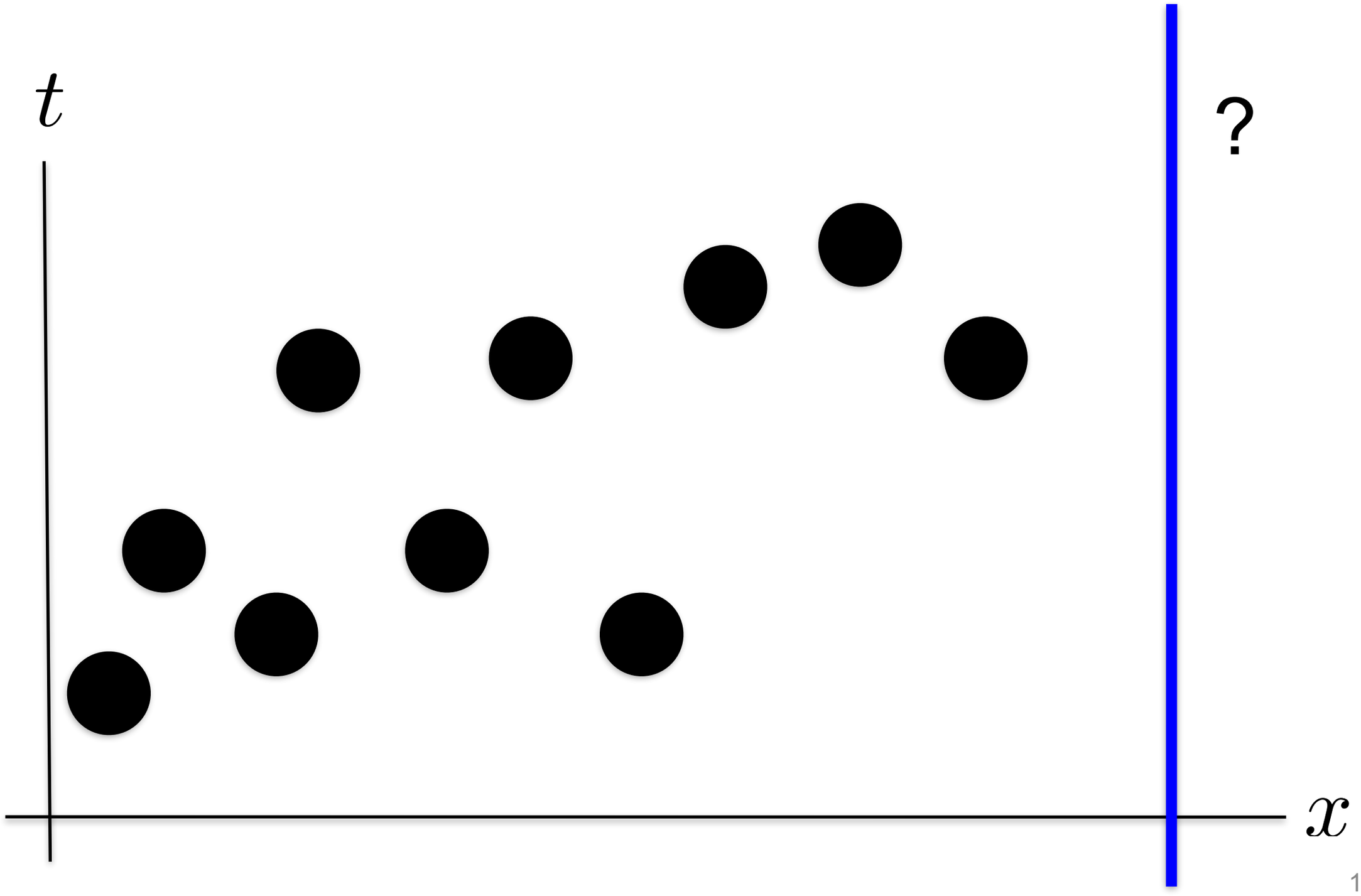
Graphical Example of Classification



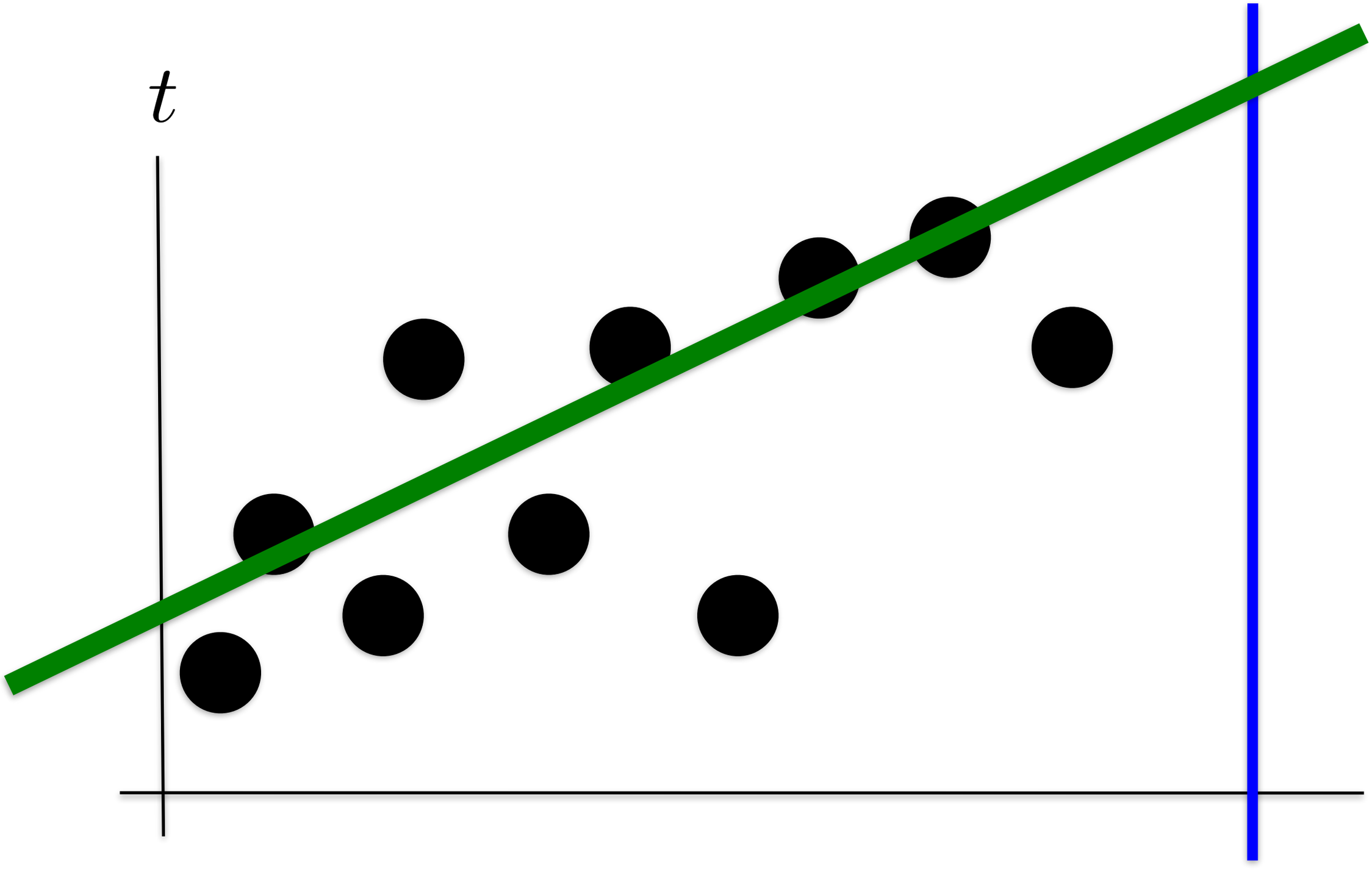
Decision Boundaries



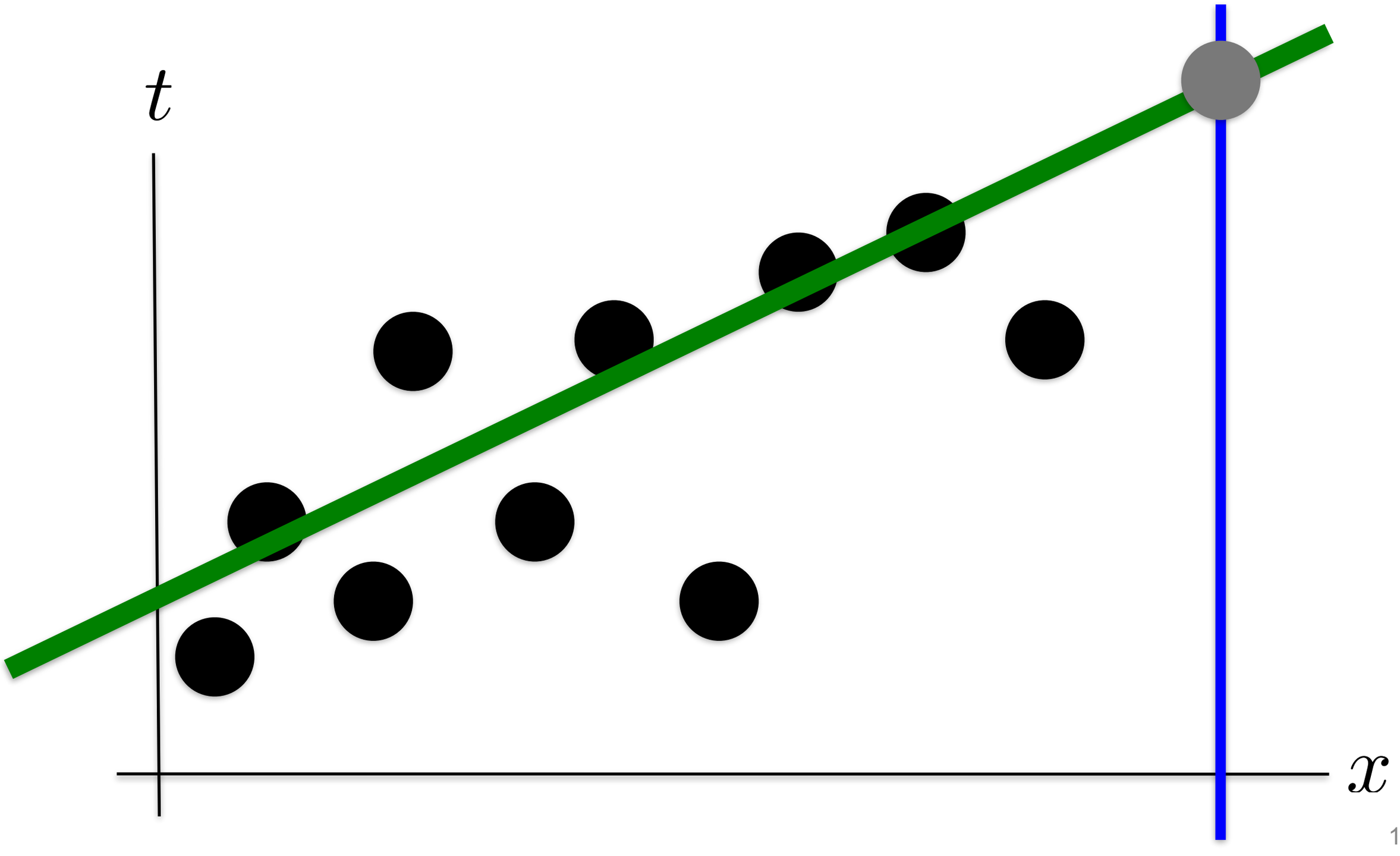
Graphical Example of Regression



Graphical Example of Regression



Graphical Example of Regression



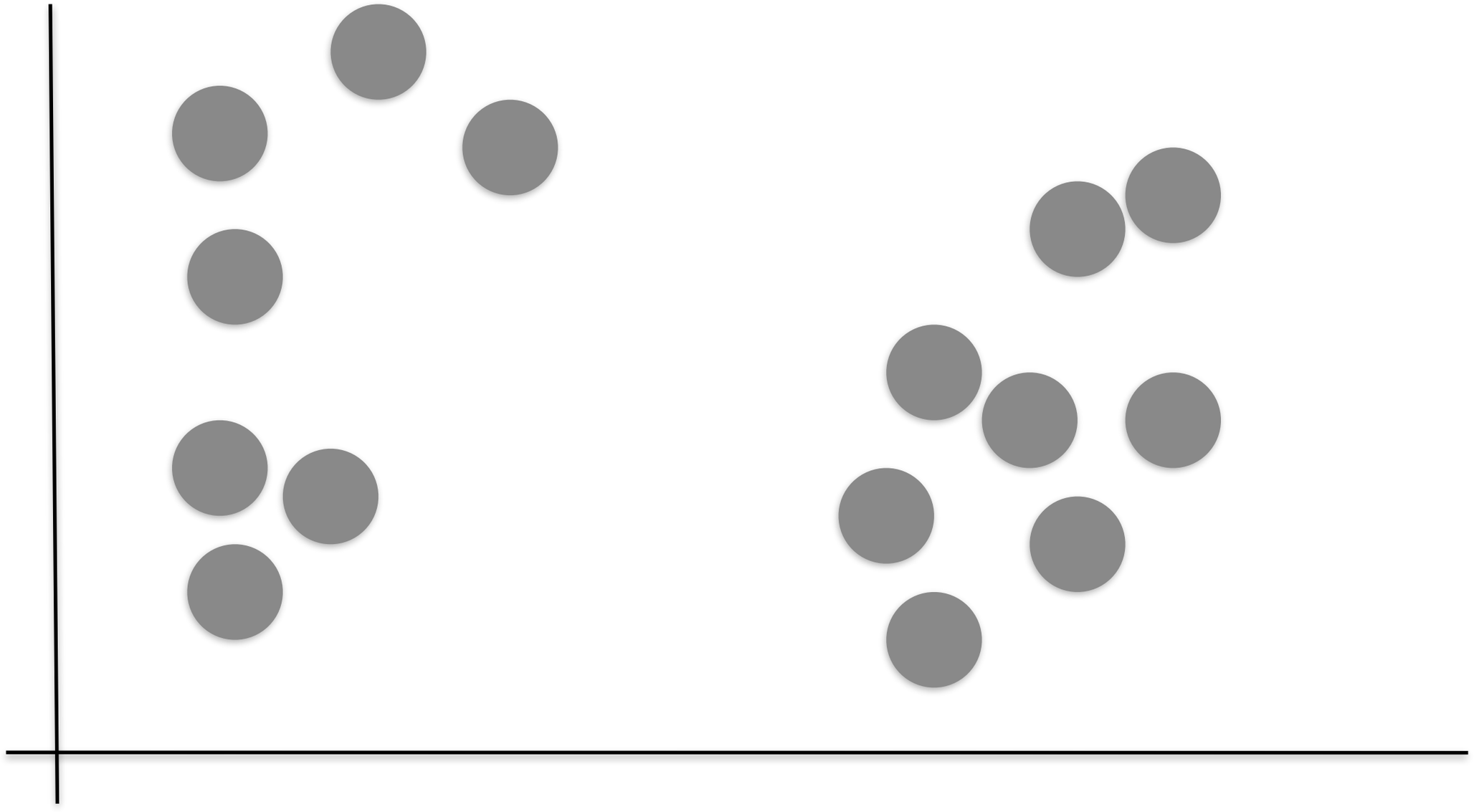
Types of Machine Learning
Unsupervised Learning

- Observe *n* examples $D = \{x_i\}$
- Clustering: Find a well-separated partition of example $D=\{D_1, D_2, \ldots, D_k\}$
- Dimensionality reduction: Find a low-distortion, low-dimensional projection $y = P\,x$, where the number of dimensions of *y* is less than (sometimes much less than) the number of dimensions of *x*
- The goal is to replace the high-complexity description of with a lower-complexity one

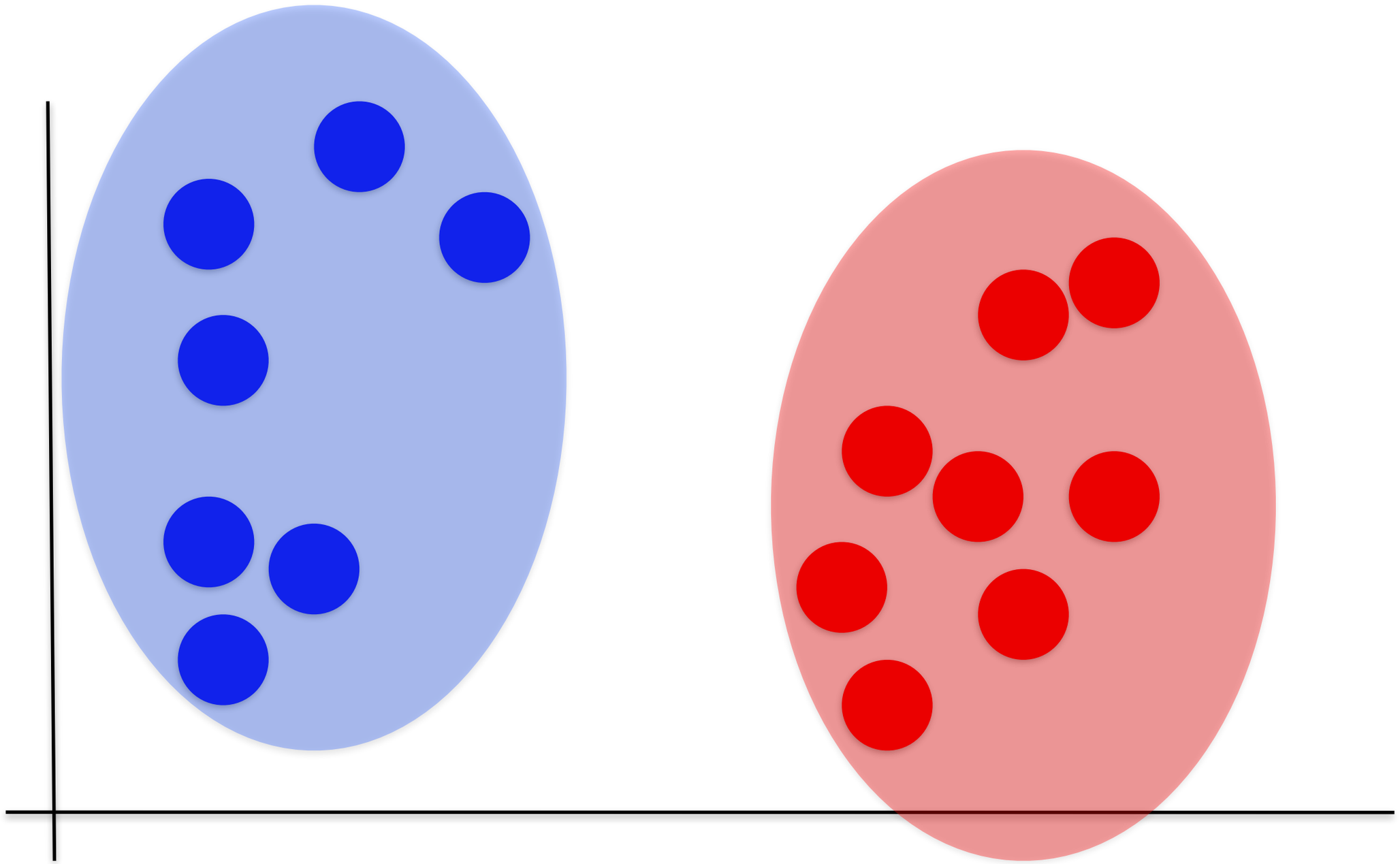
Clustering

- Clustering is an **unsupervised** learning task.
  - There is no target value to shoot for.
- Identify groups of “similar” data points, that are “dissimilar” from others.
- **Partition** the data into groups (clusters) that satisfy these constraints
  1. Points in the same cluster should be **similar.**
  2. Points in different clusters should be **dissimilar.**

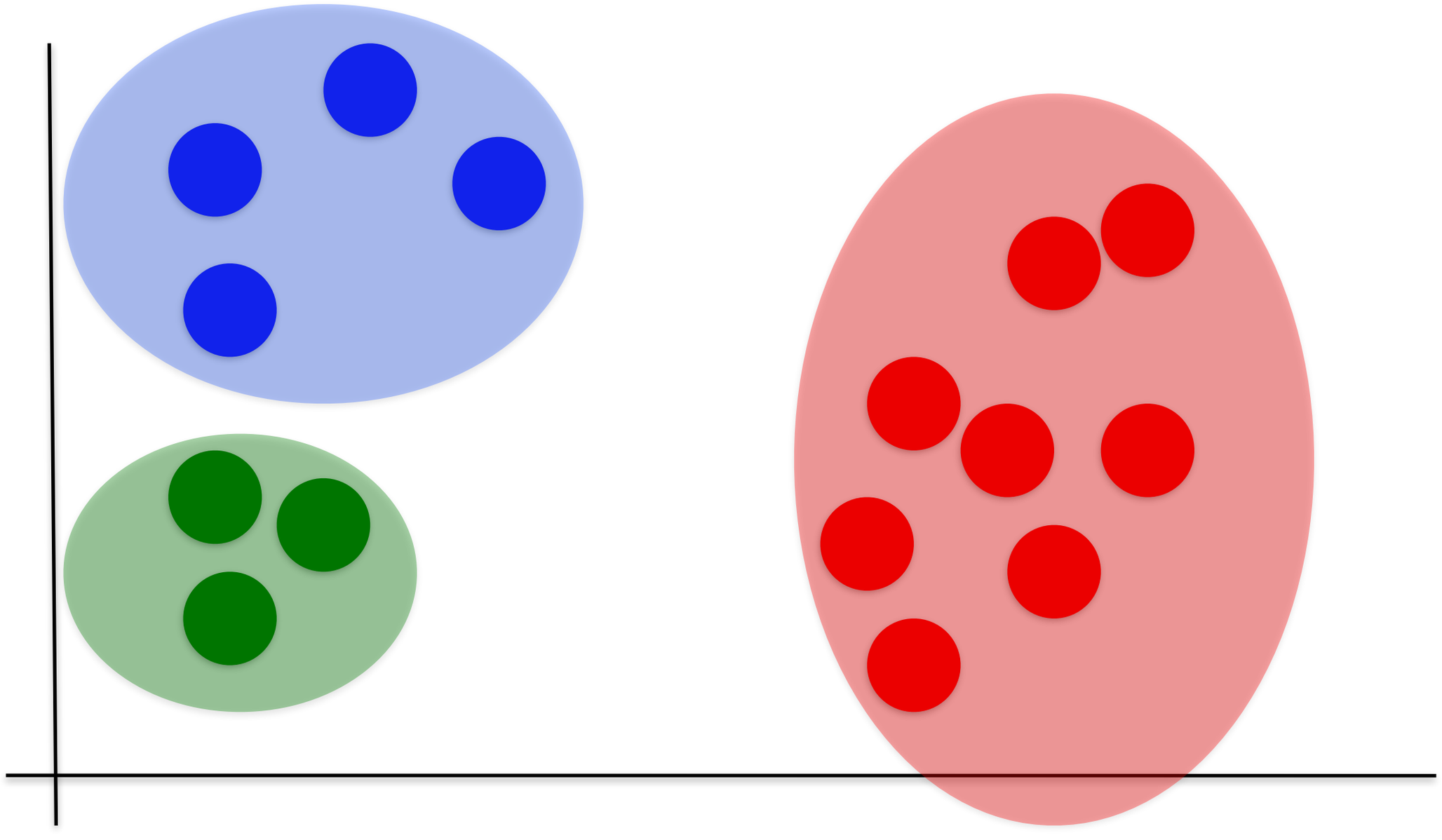
Graphical Example of Clustering



Graphical Example of Clustering

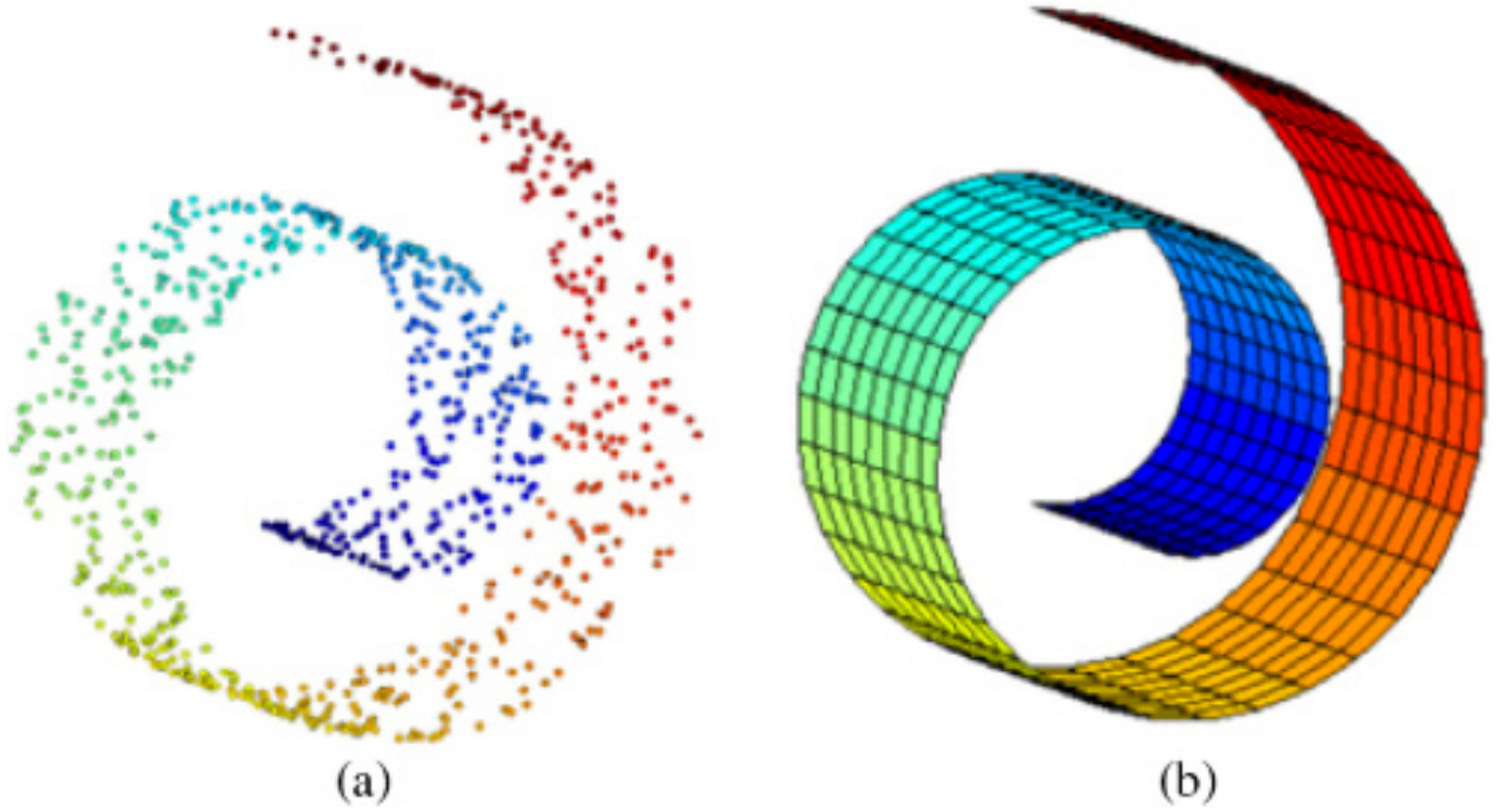


Graphical Example of Clustering

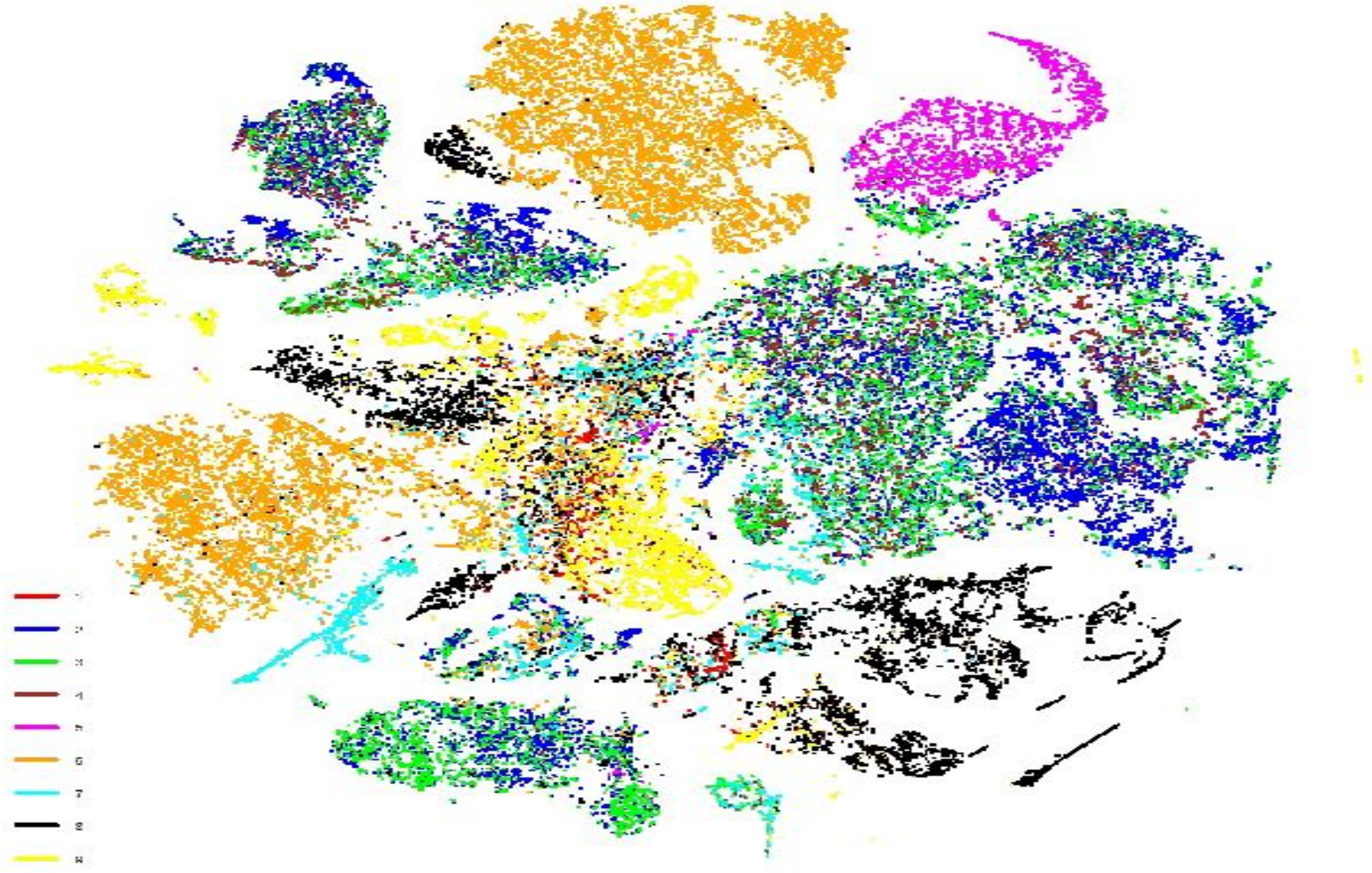


Dimension Reduction

- High dimension → low dimension
  - Dimension: number of features
- Principle:
  - Preserve some quantity of the original data
- Goal:
  - People can “see” the data in eyes
  - Some model can actually work on them



Data Manifold

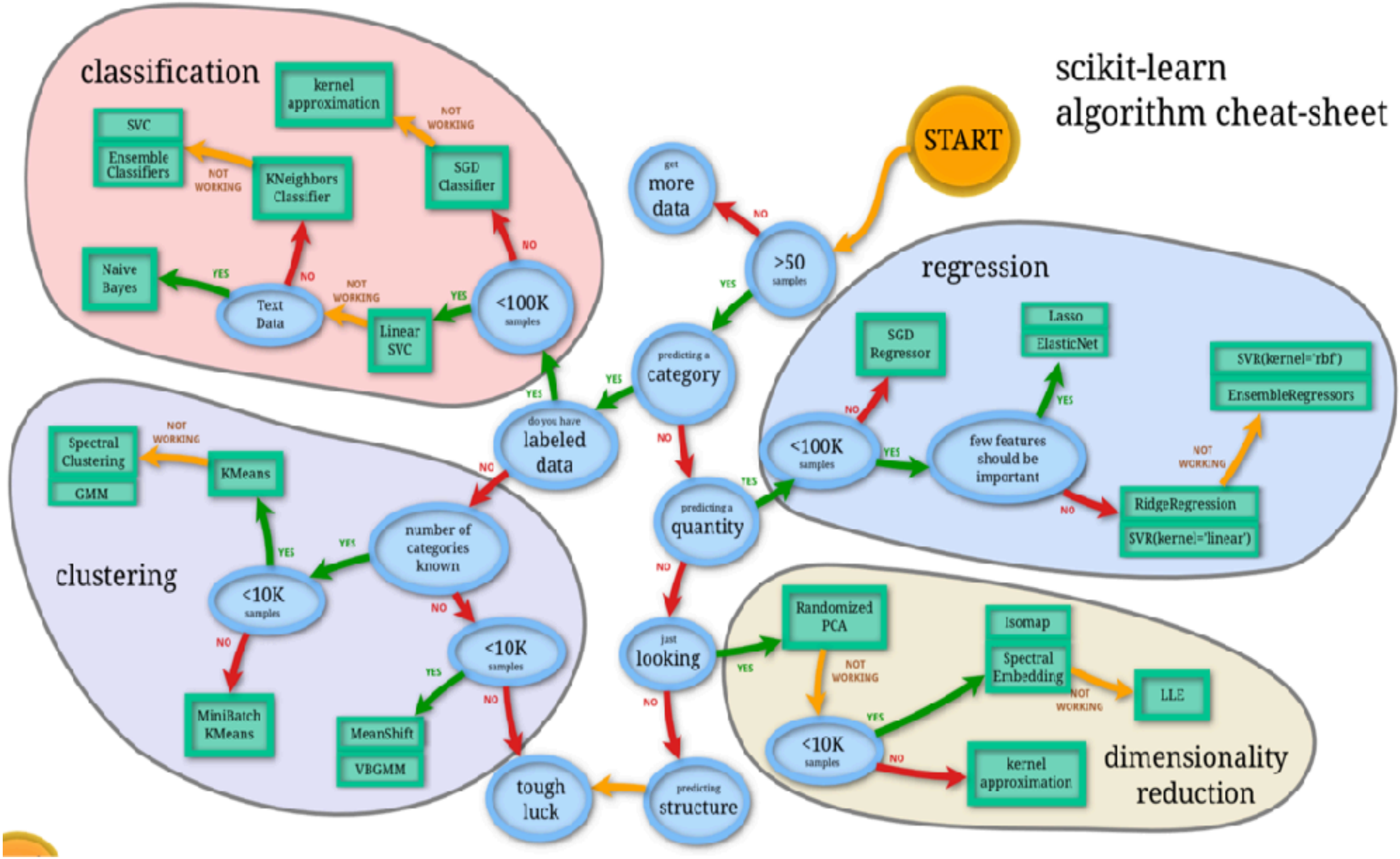


https://www.kaggle.com/c/otto-group-product-classification-challenge/forums/t/13122/visualization

Kaggle competition, otto group product, 93 features, 200K products (data), 10 categoreis

Types of Machine Learning

- Supervised Learning (SL)
- Unsupervised Learning (UL)
- Reinforcement Learning (RL)



Python scikit learn
http://scikit-learn.org/stable/tutorial/machine_learning_map/

Model Representation

Task: predict the house price based on the size of house.

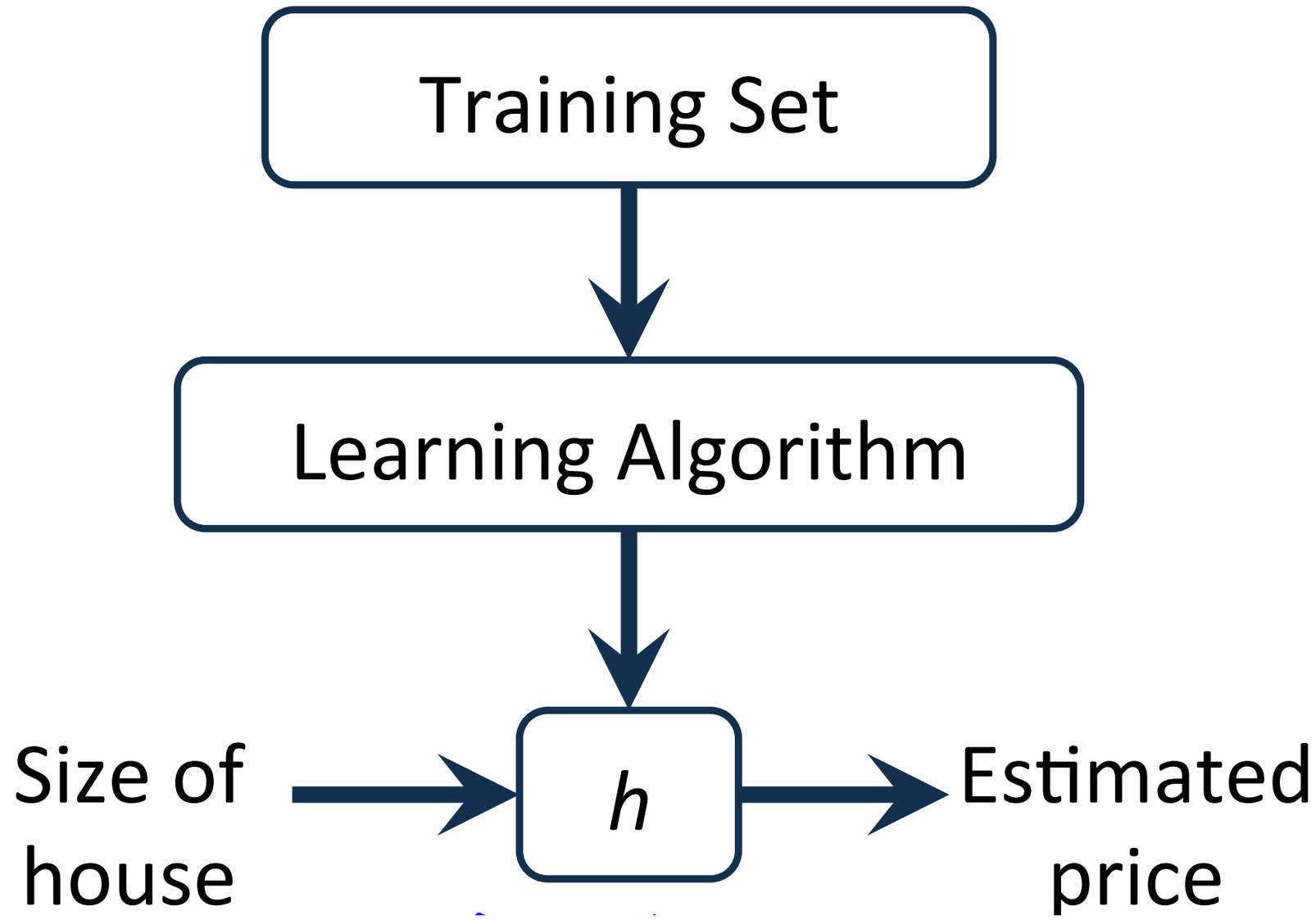
Training Set

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

x → **y**

- Learn a function, *h*, mapping any input *x to output y*: $h(x) \approx y$

Model Representation



Model Representation

Training Set

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Model Representation

Training Set

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

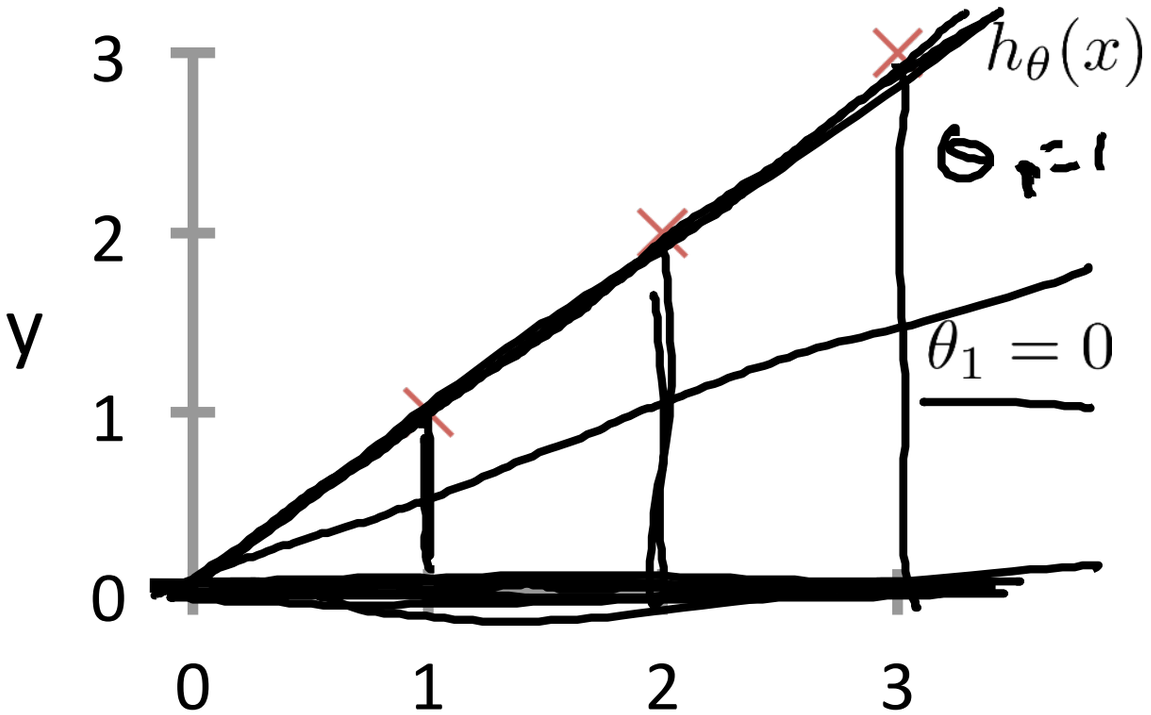
Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

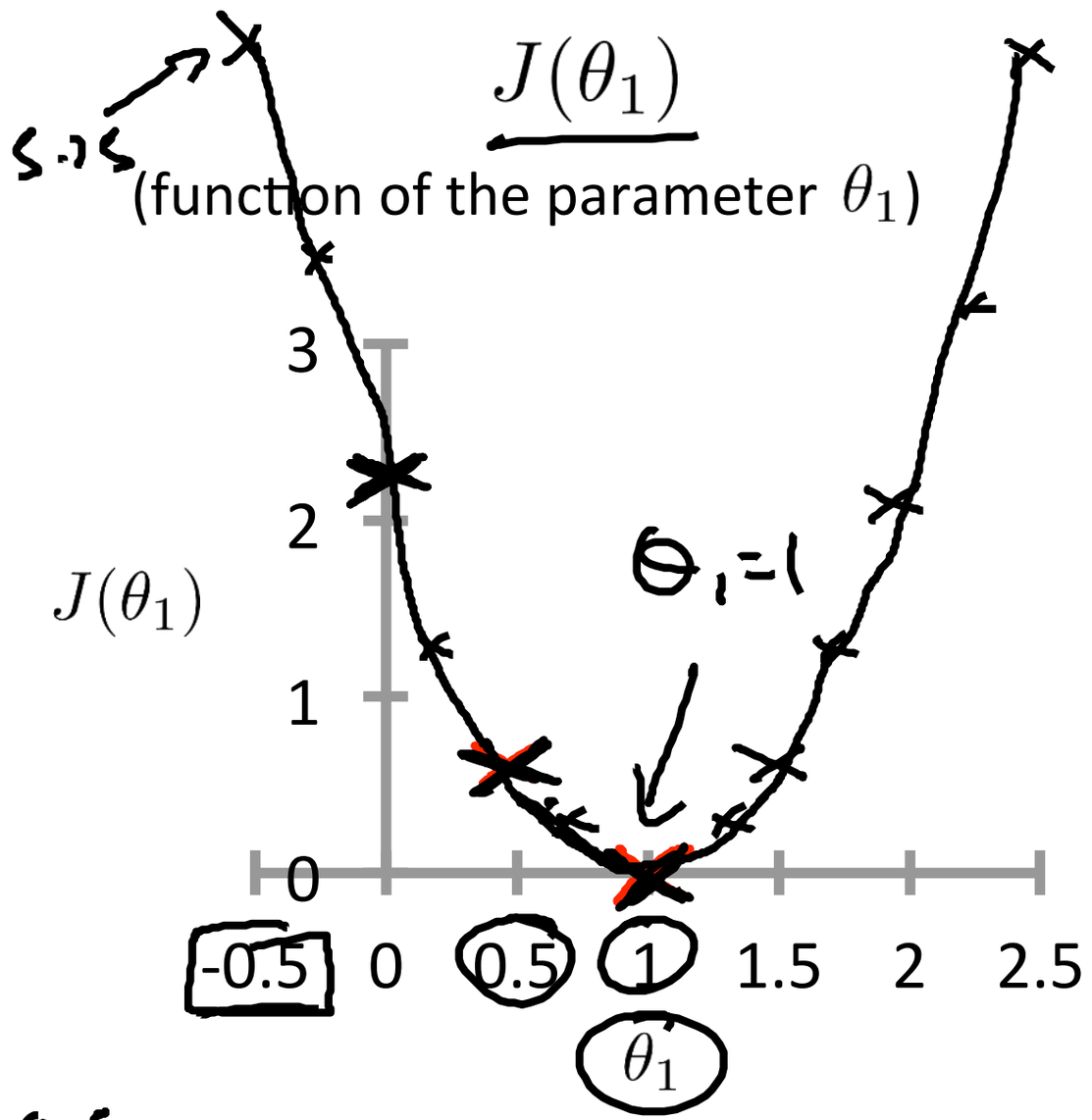
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

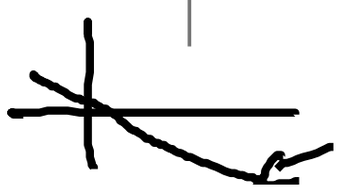
(function of the parameter θ_1)



$J(0) = \frac{1}{2m}(1^2 + 2^2 + 3^2)$

$= \frac{1}{6} \cdot 14 \approx 2.3$

$h(x) = -0.5 \times$

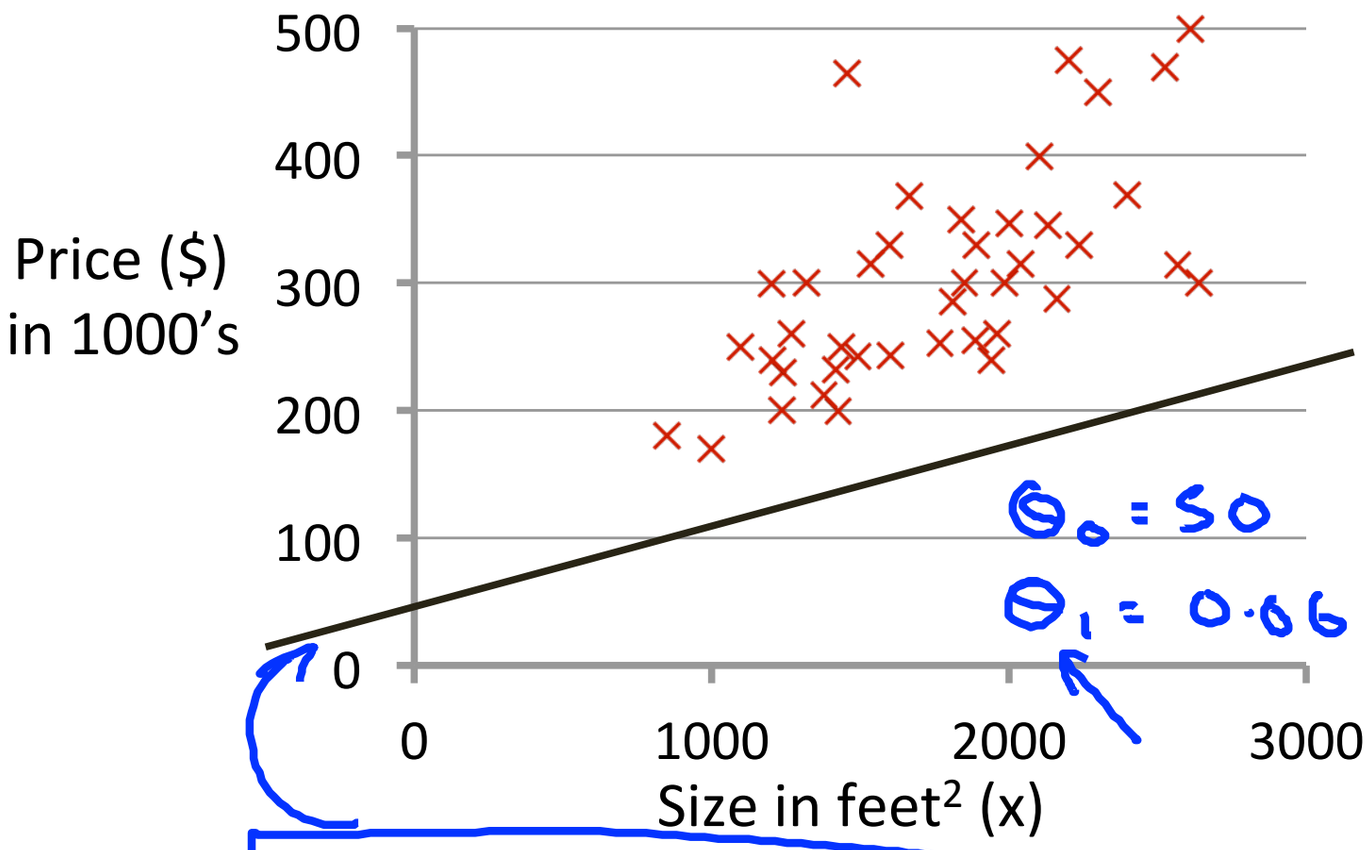


Minimize $J(\theta_1)$

θ_1

$h_\theta(x)$

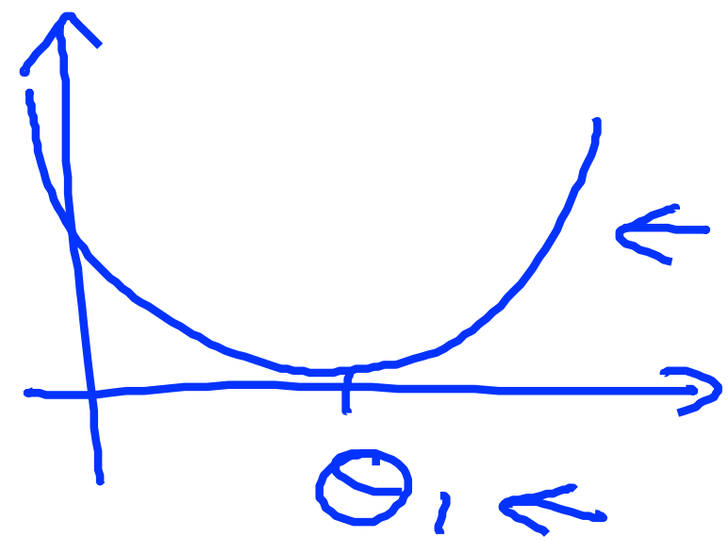
(for fixed θ_0, θ_1, this is a function of x)



$$h_\theta(x) = 50 + 0.06x$$

$J(\theta_0, \theta_1)$

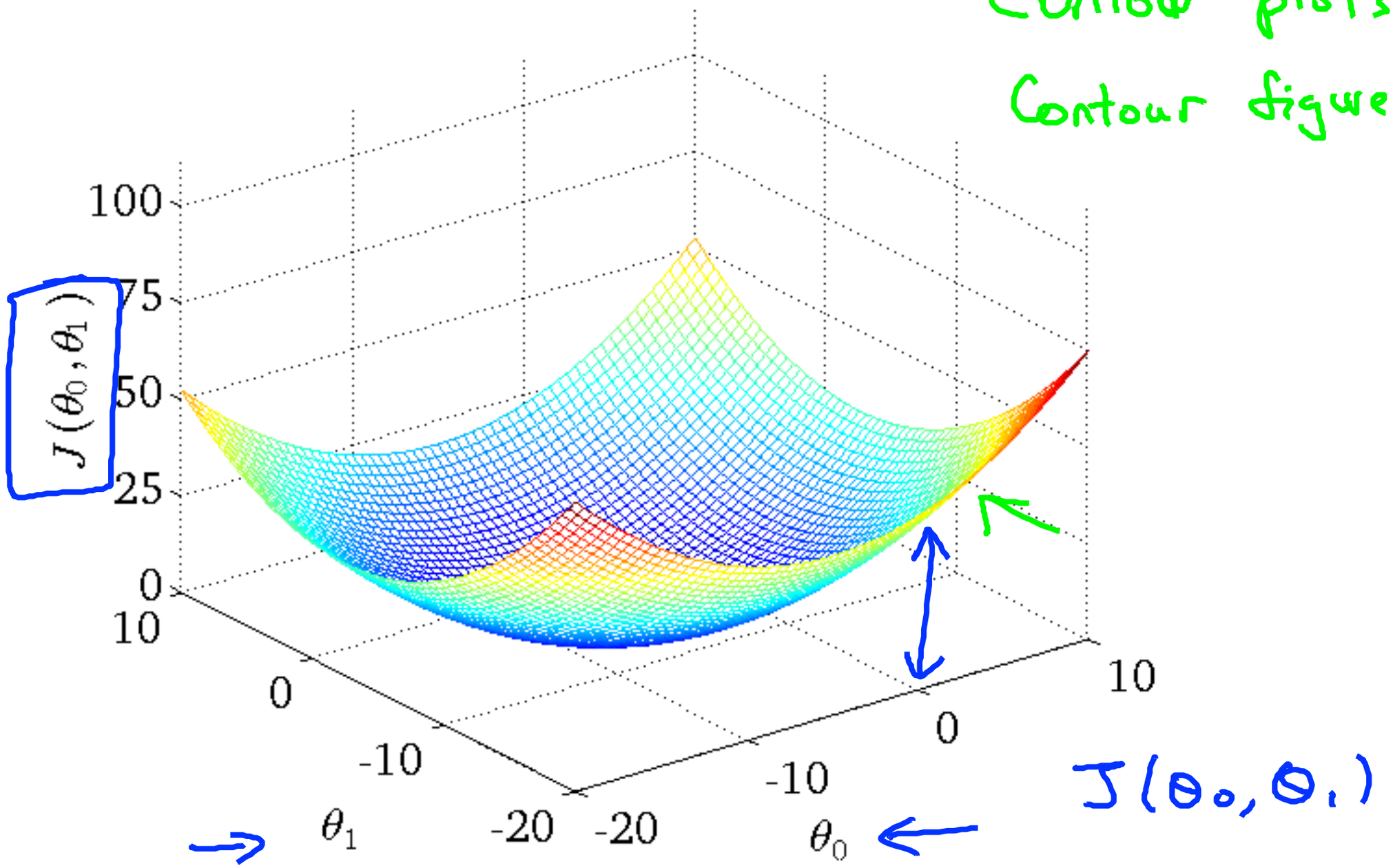
(function of the parameters θ_0, θ_1)



θ_0, θ_1

Contour plots

Contour figures -



$J(\theta_0, \theta_1)$

This Class

- Graphical Example
  - Classification
  - Regression
  - Clustering
- Model representation
- Loss function

Median

Next Class

- Linear Regression
- Maximum Likelihood Estimation (MLE)
- Gradient Decent
- Overfitting & Underfitting

Hard

To Do

- Read “Bishop”: Ch 2 & Ch 3.
- Read “Statistics” and “Linear Algebra” resources.
- Class project.