https://xkcd.com/1425

Data Science & Visualization

Basic Statistics

Gabriela Molina León
molina@uni-bremen.de

Institute for Information Management Bremen
Information Management Group (AGIM)

University of Bremen

Sessions & Deliverables

Date	Lecture (10:00-11:30)	Practical (11:45-13:15)
08.04.24	Introduction to Data Science	*Python Introduction*
15.04.24	Basic Statistics & Supervised Learning	*Practical Statistics + Supervised*
22.04.24	Unsupervised Learning	*Practical Unsupervised*
29.04.24	Introduction to Data Visualization	Guest Lecture on NLP by Oxana Vitman
06.05.24	Exploratory Data Analysis	Data Science & Vis Presentation

19.07.24 Deadline Final Report

Note that for the slots marked red, you are expected to prepare presentations.
For the slots marked blue, you are expected to bring a computer.
In the slots marked purple, you are not expected to come to the classroom.

Next week: Unsupervised Learning

- I will send you the materials on Monday
  - Lecture: Video
  - Tutorial: Jupyter notebook
- Focus on
  - Clustering
  - Dimensionality reduction

This class is about applying computational methods to real-world problems.

The goal is not only to learn basic concepts, but also to enable you to make a difference.

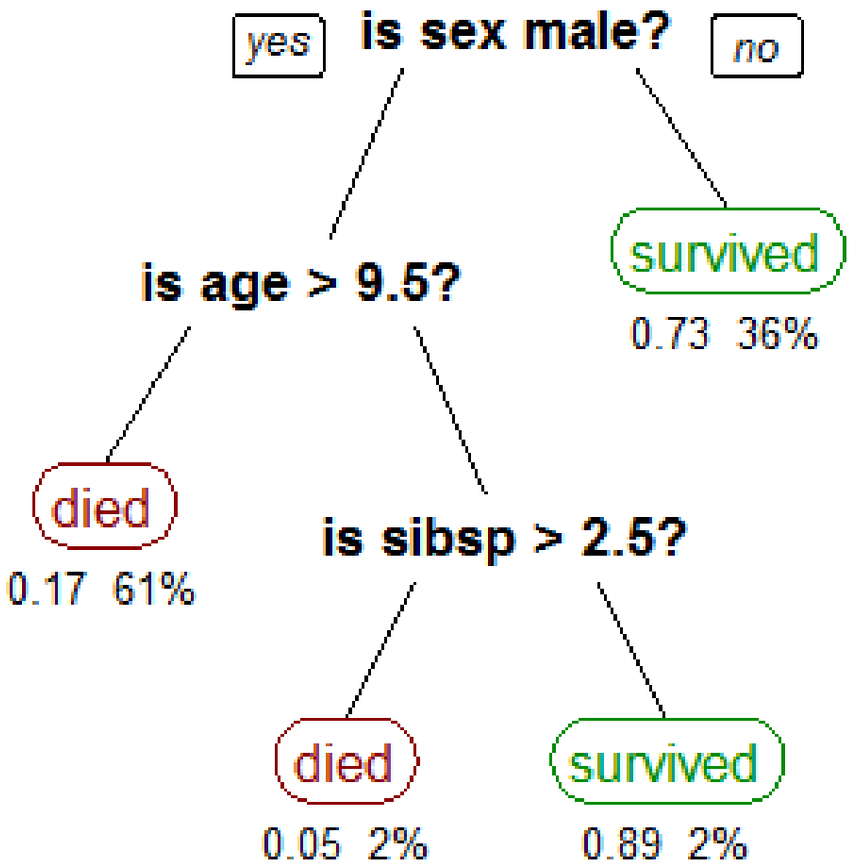
Titanic Statistics

Paramount Home Entertainment/AP Photo

Exploratory Data Analysis: Goals

- Analyse the dataset to answer your questions using **statistical** methods and data visualization
- Build a machine learning model to make predictions of the future

Titanic Dataset



What to look at first?

- Descriptive statistics
- Five-number summary (mean/median, min, max, q1, q3)
- Histograms
- Box and Whiskerplots, also known as box plots

Five-number summary

- mean
  - the sum of the values divided by the count of non - missing observations
- median
  - the number exactly in the middle of an ordered list of numerical values
- minimum
- maximum
- first quartile (q1)
- third quartile (q3)
- also: mode
  - mode is the most frequent value in a dataset

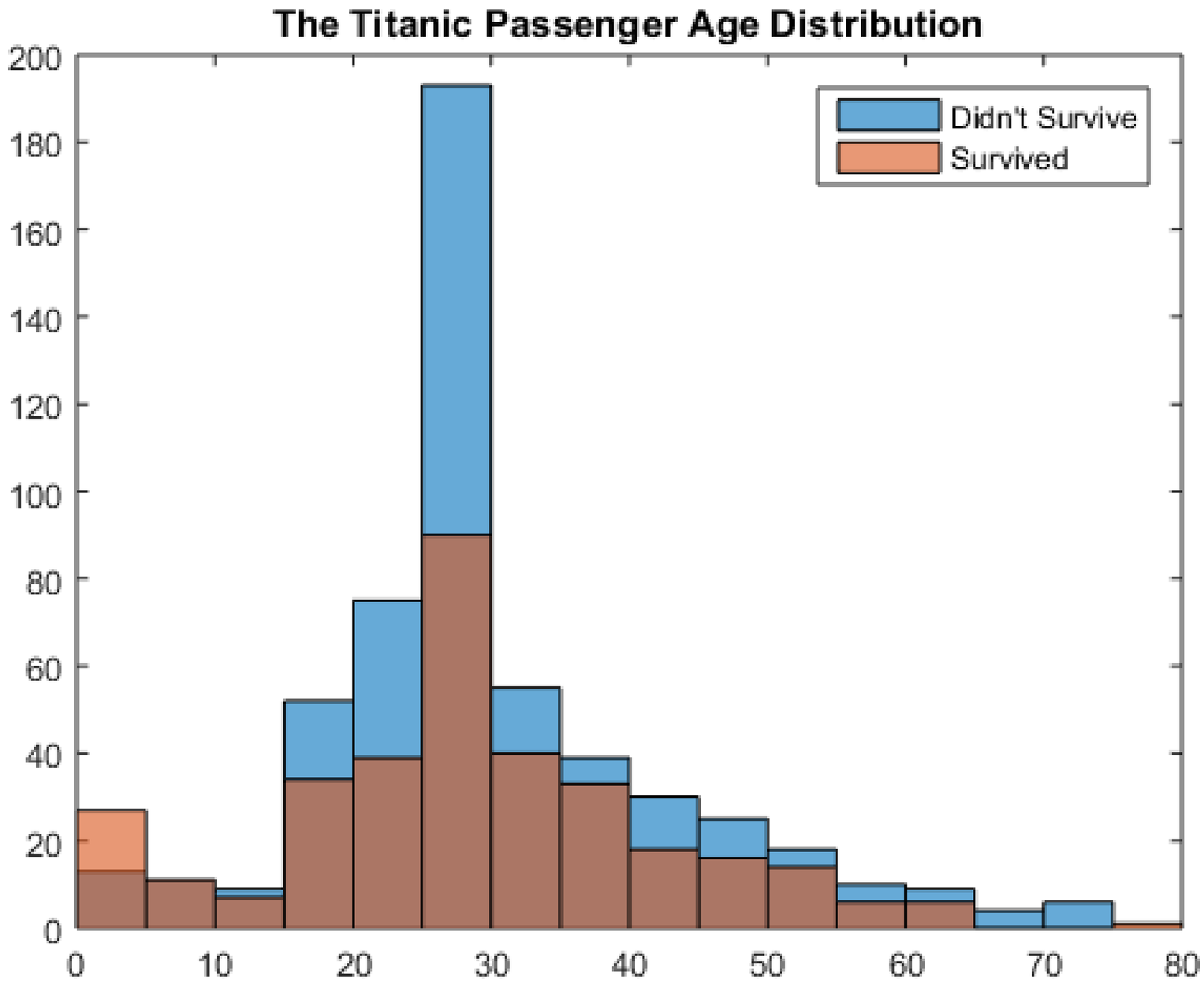
Quartiles

- cut points that split a distribution in equal sizes
- first quartile: 0.25 (25% of the dataset)
- second quartile: 0.50 == Median
- third quartile: 0.75

Standard deviation

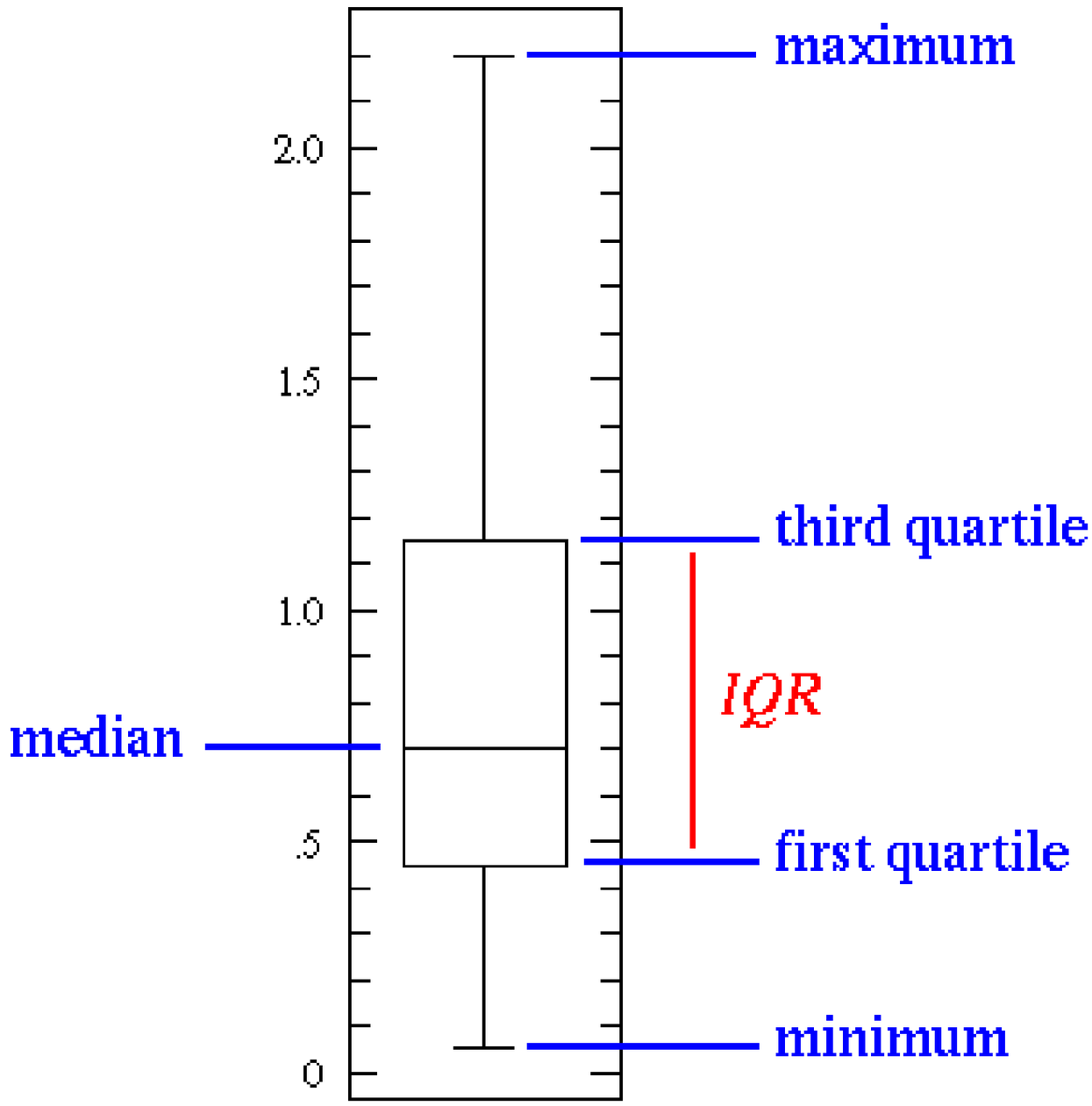
- a measure of dispersion
- in the same unit as the values
- a high standard deviation indicates the data points are spread over a wide range of values

Histograms



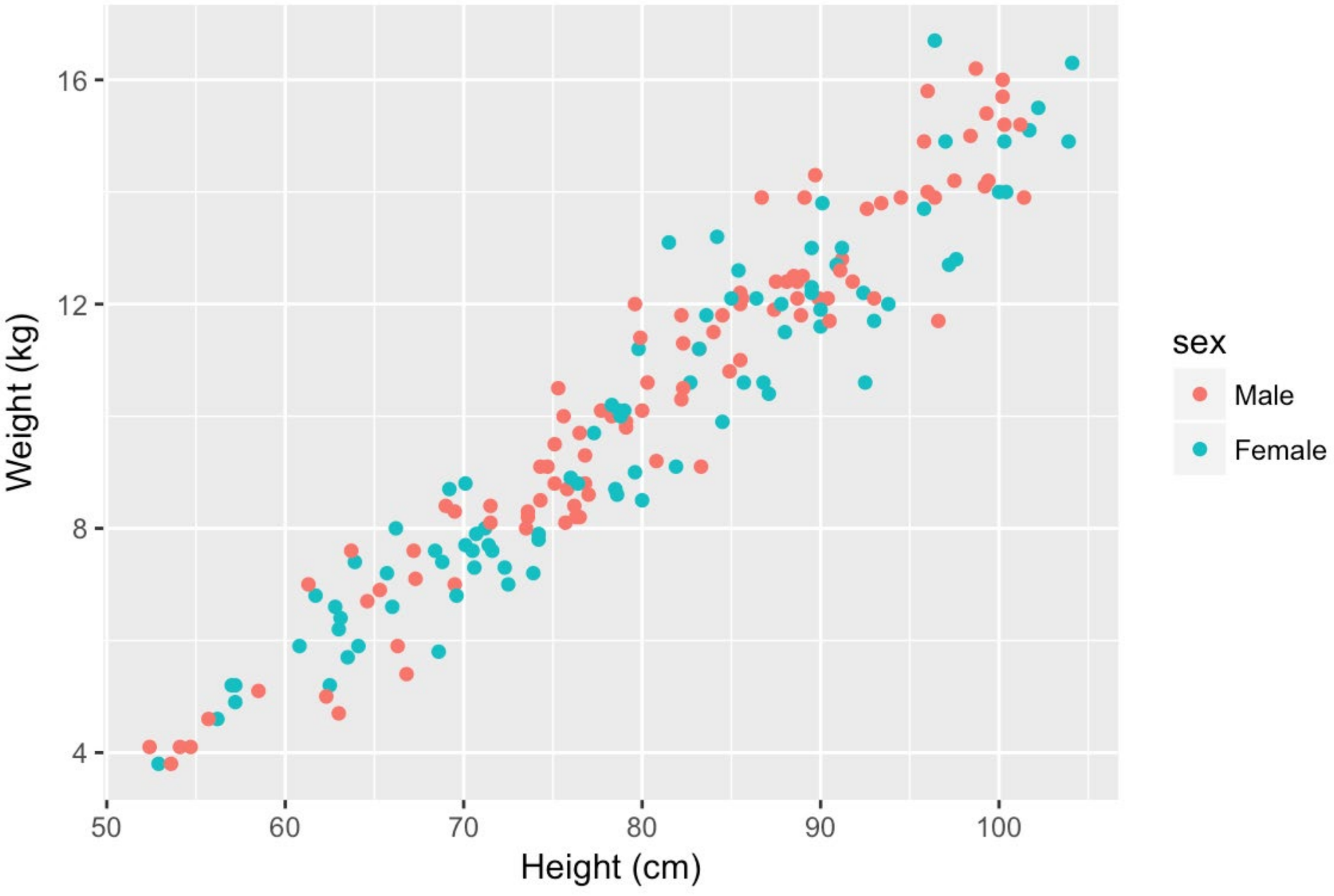
https://blogs.mathworks.com/loren/2015/06/18/getting -started-with-kaggle-data-science-competitions/

Box plots



http://www.physics.csbsju.edu/stats/box2.html

Scatterplots



https://bookdown.org/rdpeng/RProgDA/basic-plotting-with-ggplot2.html

Imputation: Creating values

- in the Titanic dataset, the Age feature has 177 null values
- to replace these NaN values, we can assign them the mean age of the dataset
- problem: we can't assign a 4 year kid with the mean age that is 29 years
- Solution: Use additional information

Example: Titanic Data

2.1 Download and load the data.

We download the dataset from Kaggle website. We will first list all files we downloaded.

In [2]:
```
!ls datasets/
```

```
data_description.txt  submission.csv        train.csv
gender_submission.csv test.csv
```

Now let's start by looking at the data description from Kaggle.

In [3]:
```
!cat datasets/data_description.txt
```

```
Data Dictionary:
survival        Survival                    0 = No, 1 = Yes
pclass          Ticket class                1 = 1st, 2 = 2nd, 3 = 3rd
sex             Sex
Age             Age in years
sibsp           # of siblings / spouses aboard the Titanic
parch           # of parents / children aboard the Titanic
ticket          Ticket number
fare            Passenger fare
cabin           Cabin number
embarked        Port of Embarkation     C = Cherbourg, Q = Queenstown, S = Southampton
```

Source: github.com/msznajder/titanic_survival_analysis_kaggle

Example: Titanic Data

In [5]:
```
titanic_train.head()
```

Out[5]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Source: github.com/msznajder/titanic_survival_analysis_kaggle

Example: Titanic Data

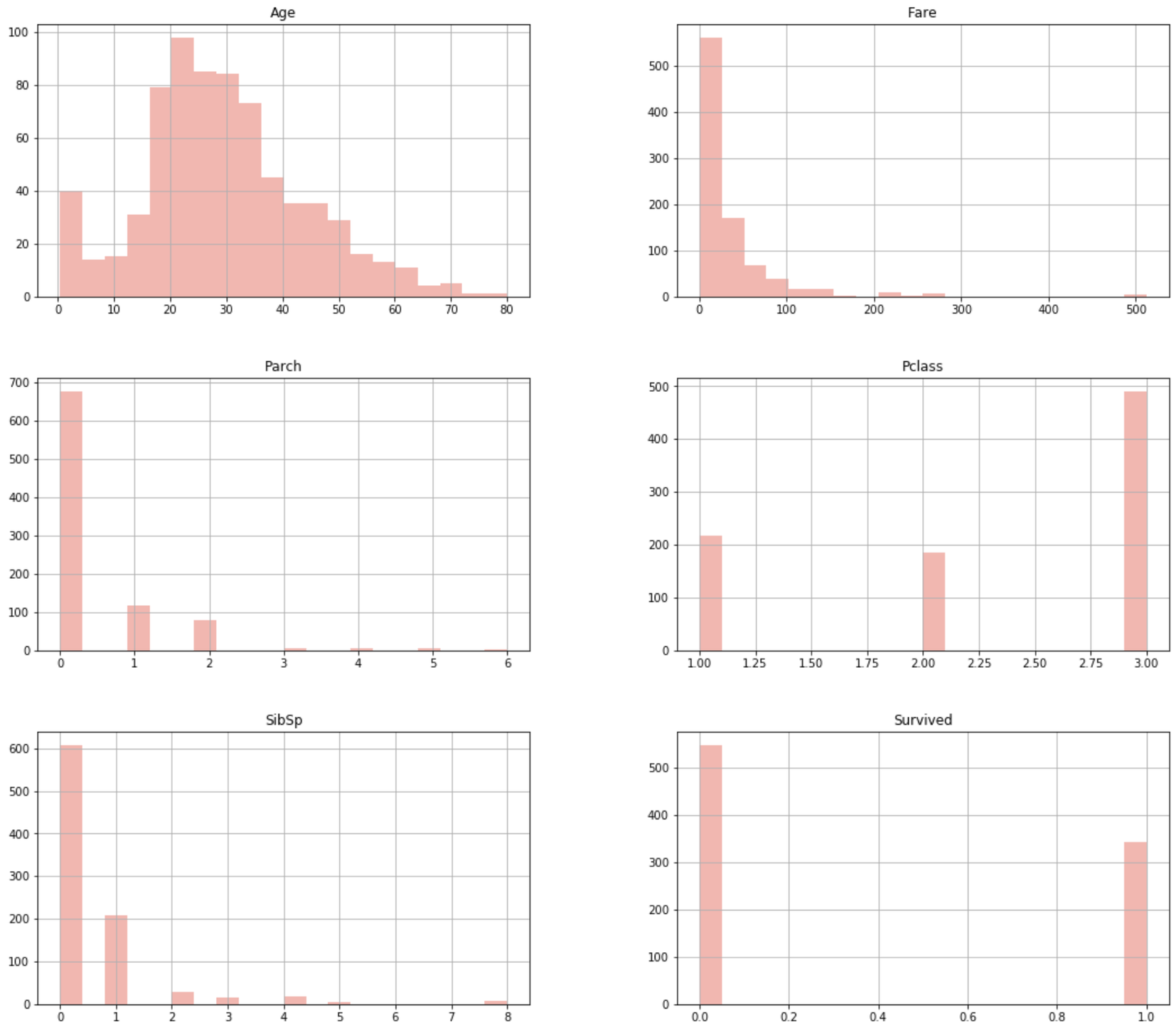
In [8]:
```
titanic_train.describe()
```

Out[8]:

	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Example: Titanic Data

```
titanic_train.hist(bins=20, figsize=(18, 16), color="#f1b7b0")
```



Example: Titanic Data

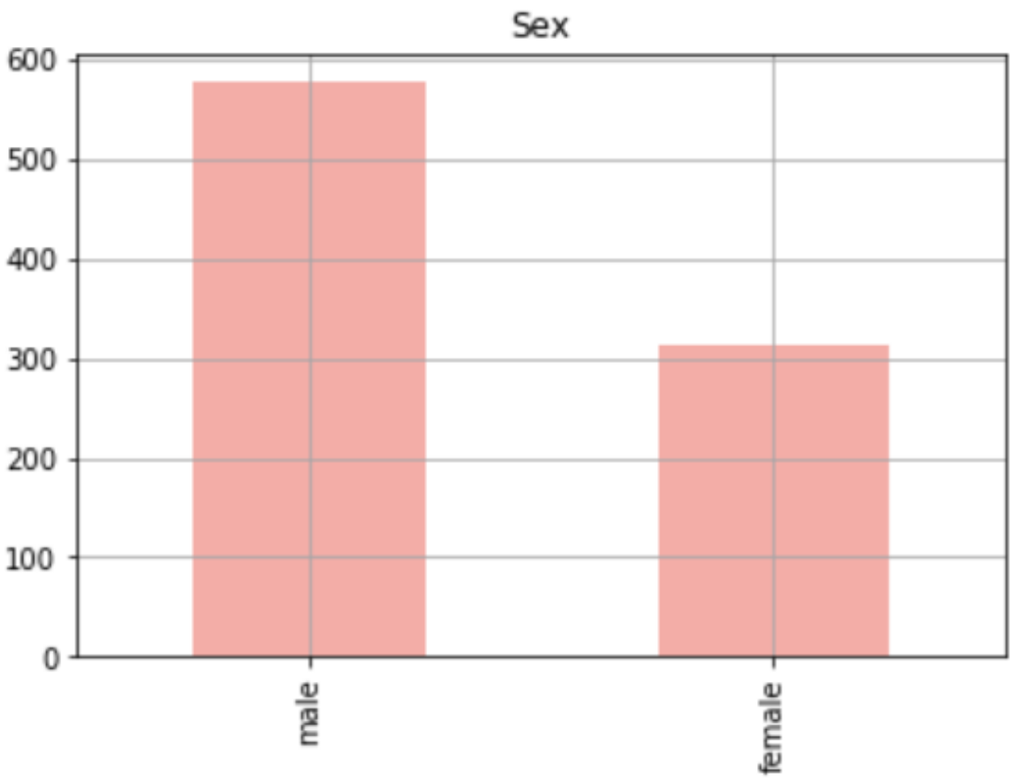
```
In [12]: titanic_train["Ticket"].value_counts()
```

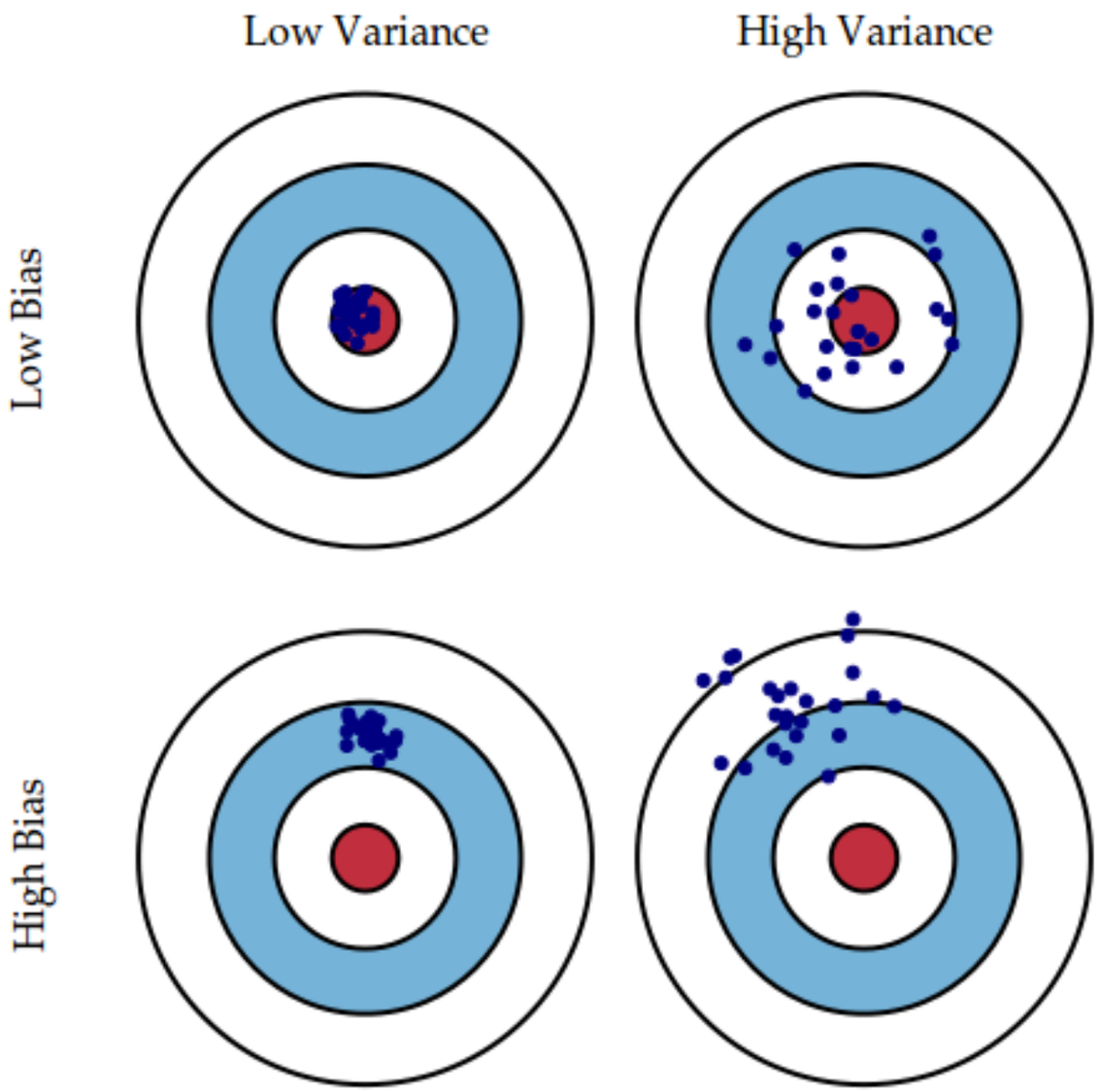
```
Out[12]: CA. 2343            7
         1601                7
         347082              7
         347088              6
         3101295             6
         CA 2144             6
         S.O.C. 14879        5
         382652              5
         LINE                4
         4133                4
         2666                4
         W./C. 6608          4
         PC 17757            4
         113781              4
```

Example: Titanic Data

```
In [15]: titanic_train["Sex"].value_counts().plot(kind='bar', figsize=(6, 4), grid=True, color="#f1b7b0", t
         itle="Sex")
```

```
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x11c9d58d0>
```





Data Science & Visualization

Supervised Learning

Gabriela Molina León
molina@uni-bremen.de

Institute for Information Management Bremen
Information Management Group (AGIM)

University of Bremen

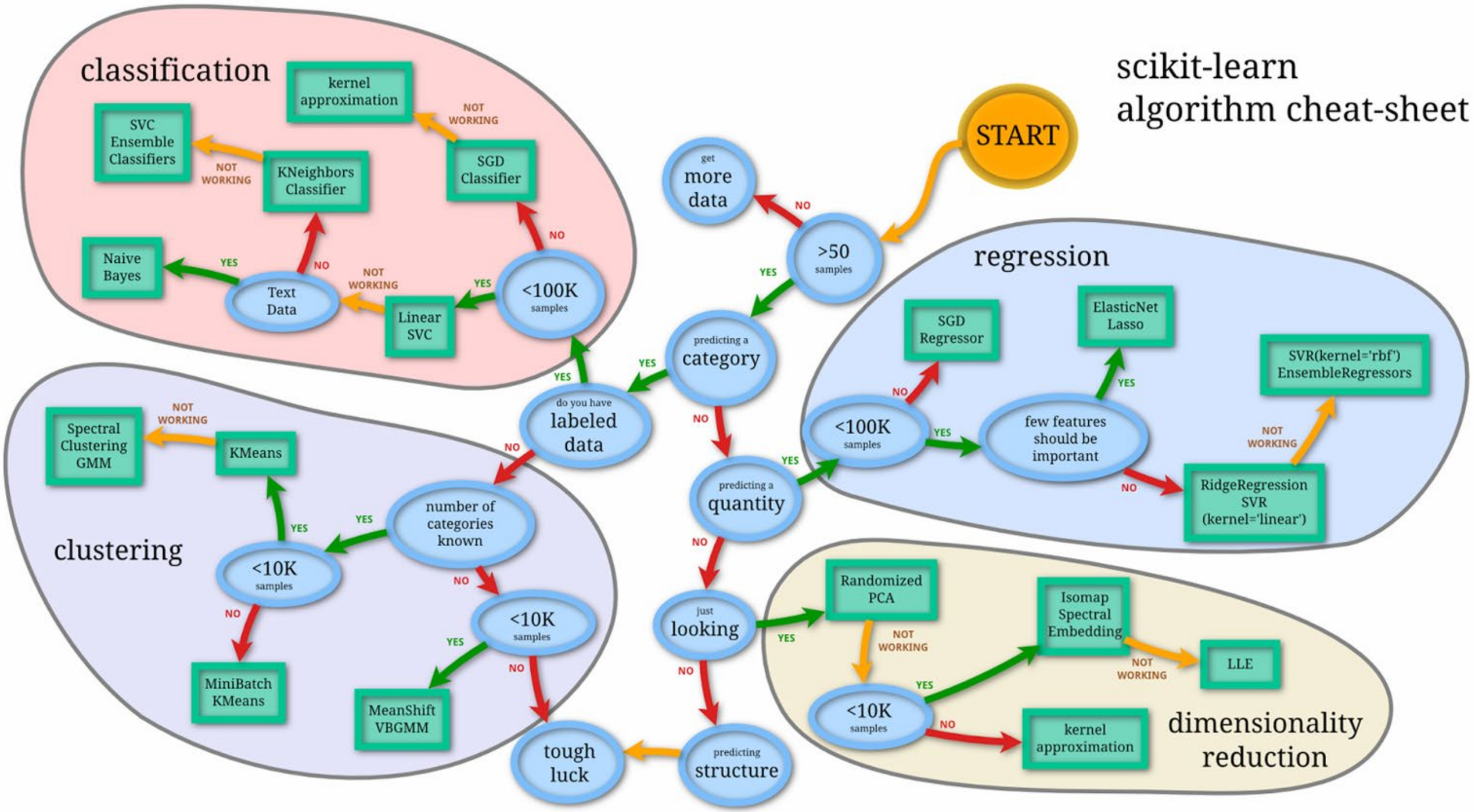
Supervised learning

- We have features (a X) and we have classes (a Y)
- The goal is prediction

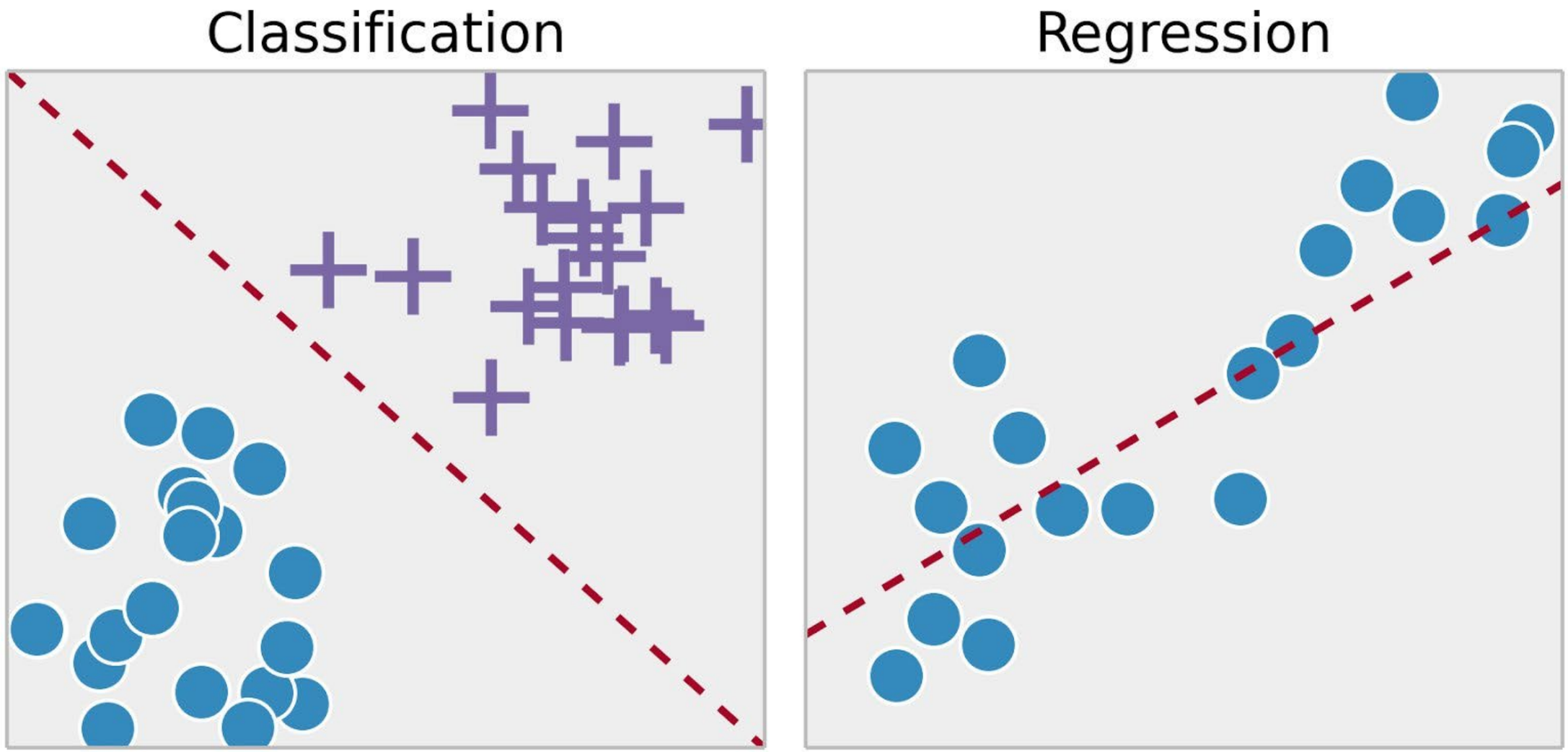
Unsupervised learning

- we have features (a X)
- The goal is exploration

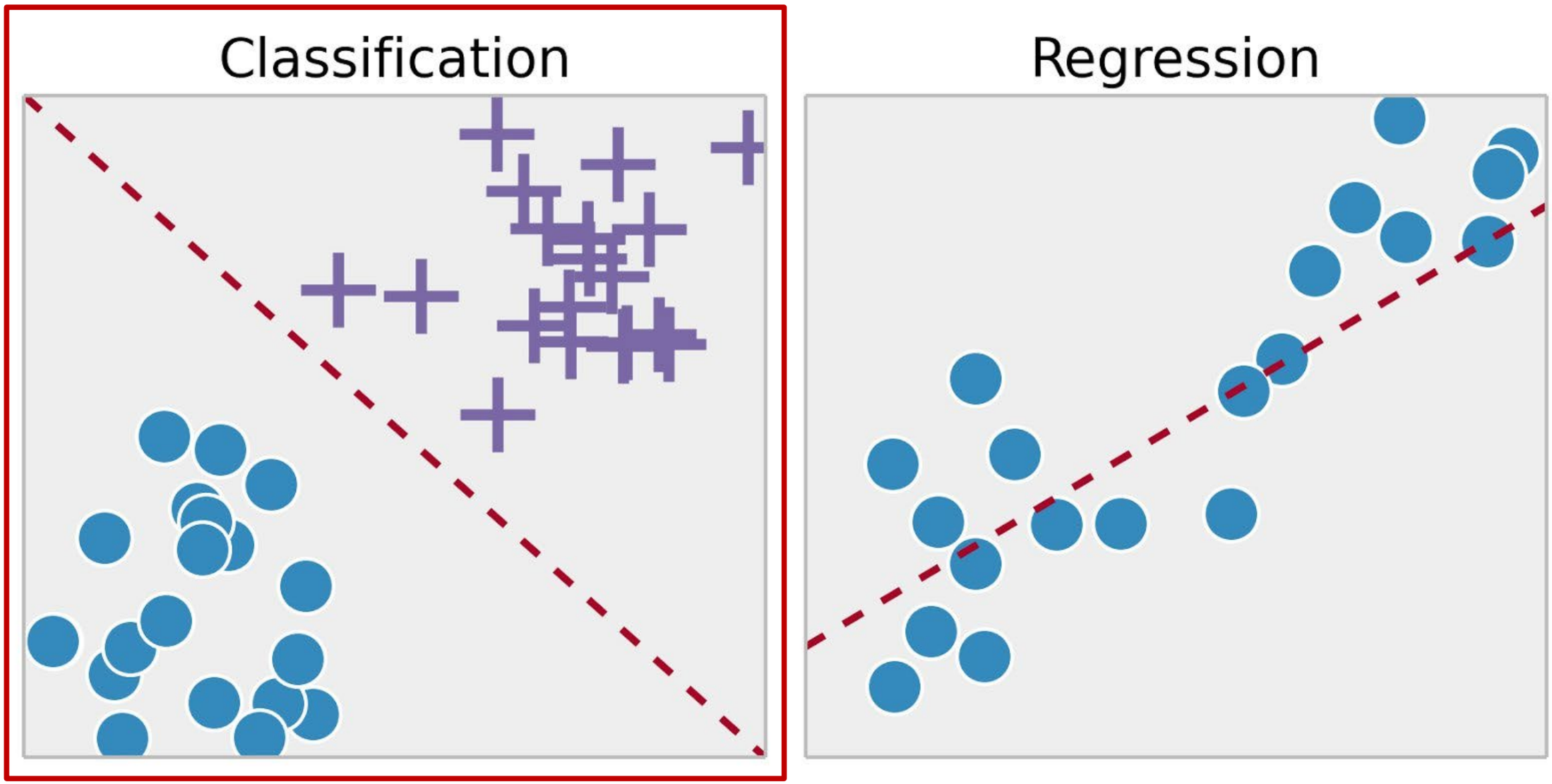
scikit-learn algorithm cheat-sheet



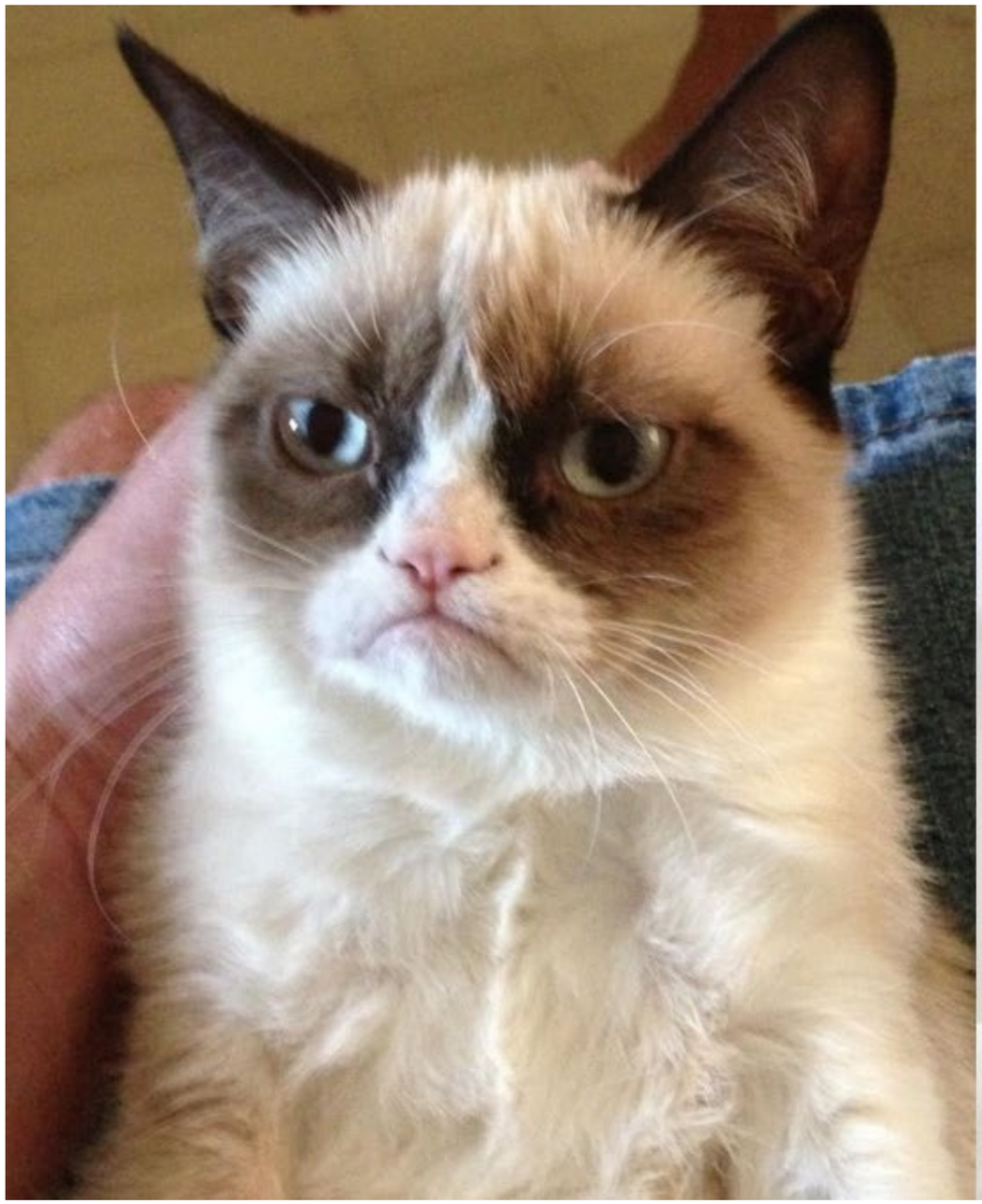
Supervised learning



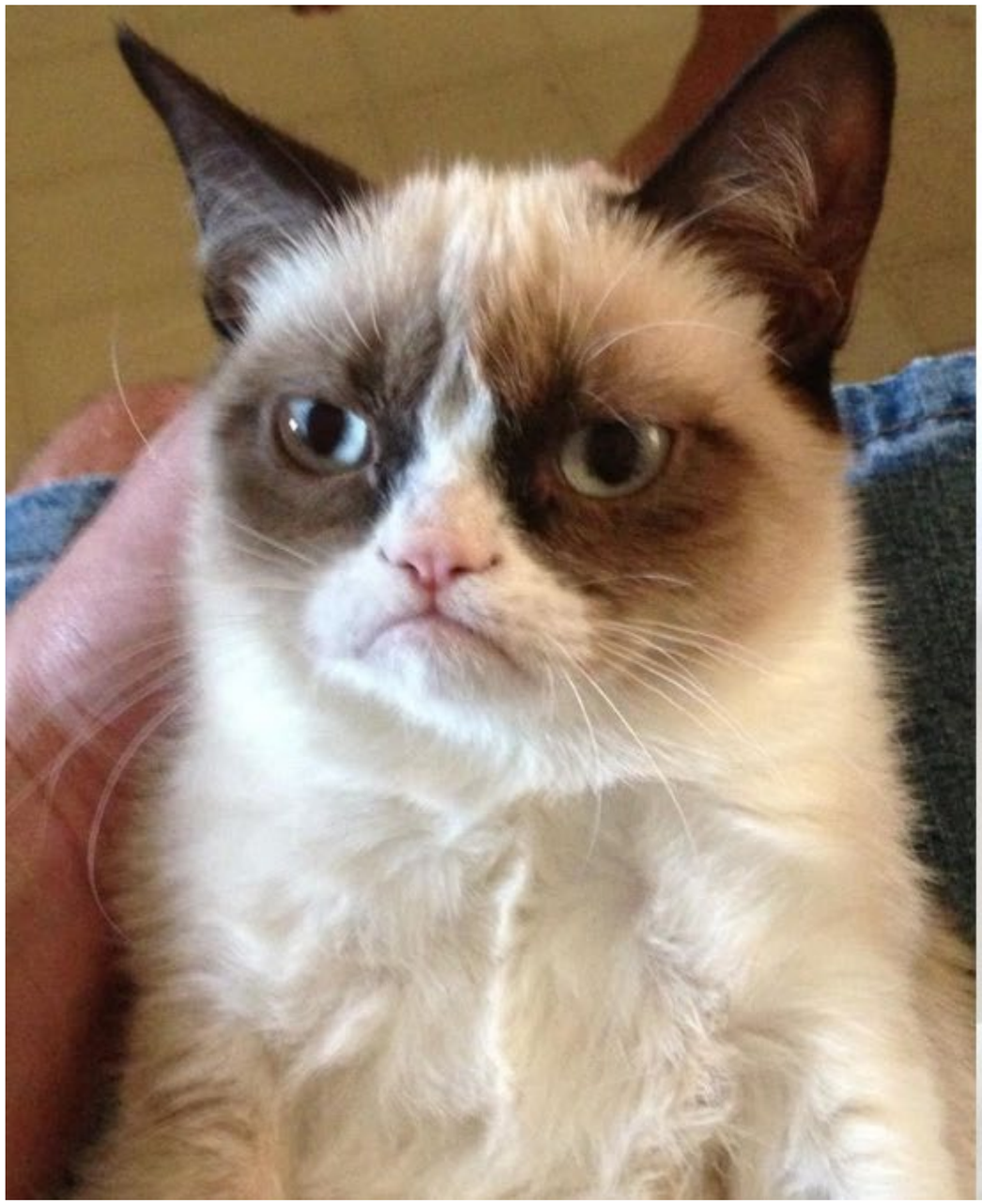
Supervised learning



Classification is simple, right?



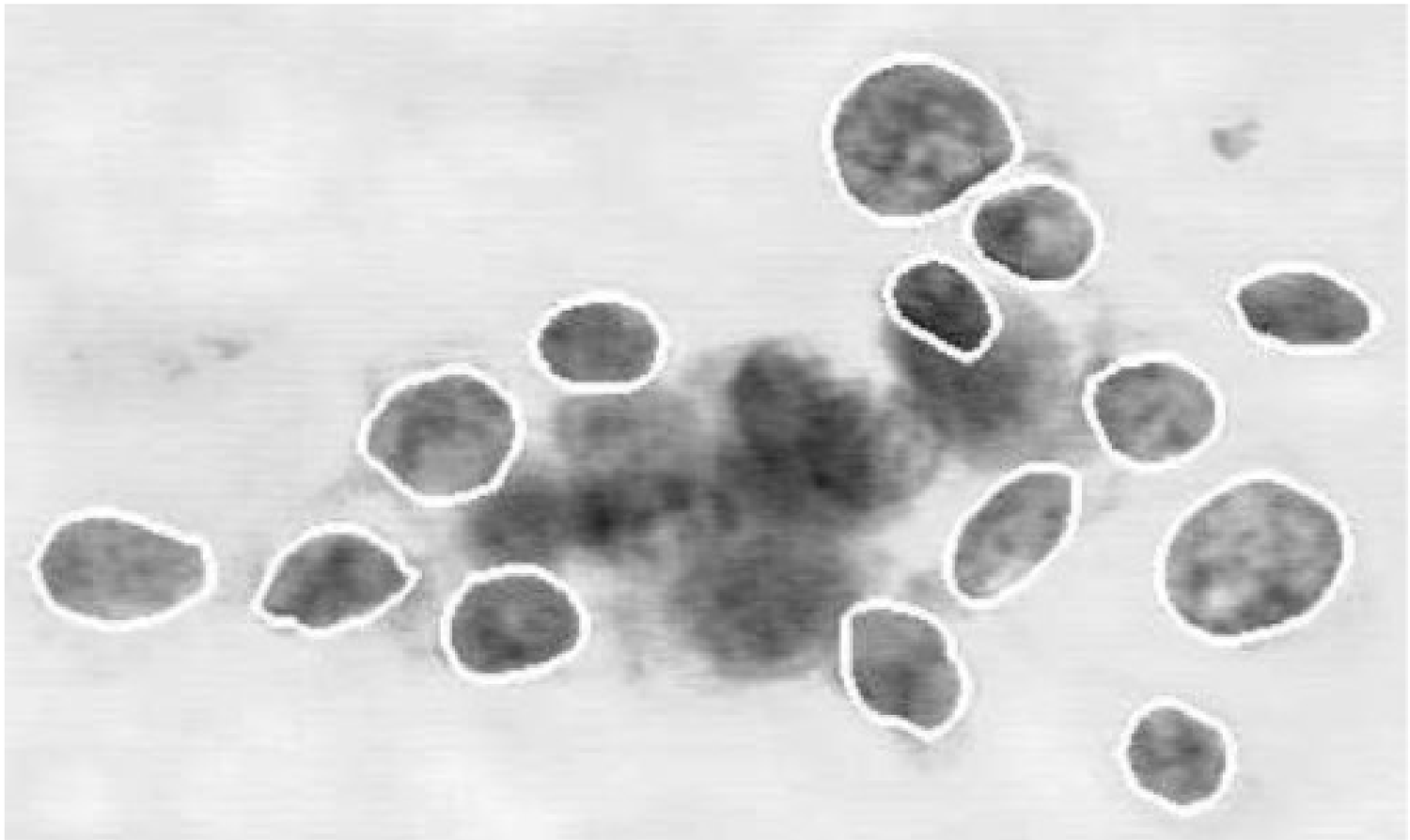
Classification is simple, right?



Animal / **Cat** / Light Fur

Animal / **Dog** / Light Fur

Breast Cancer



Is the tumor

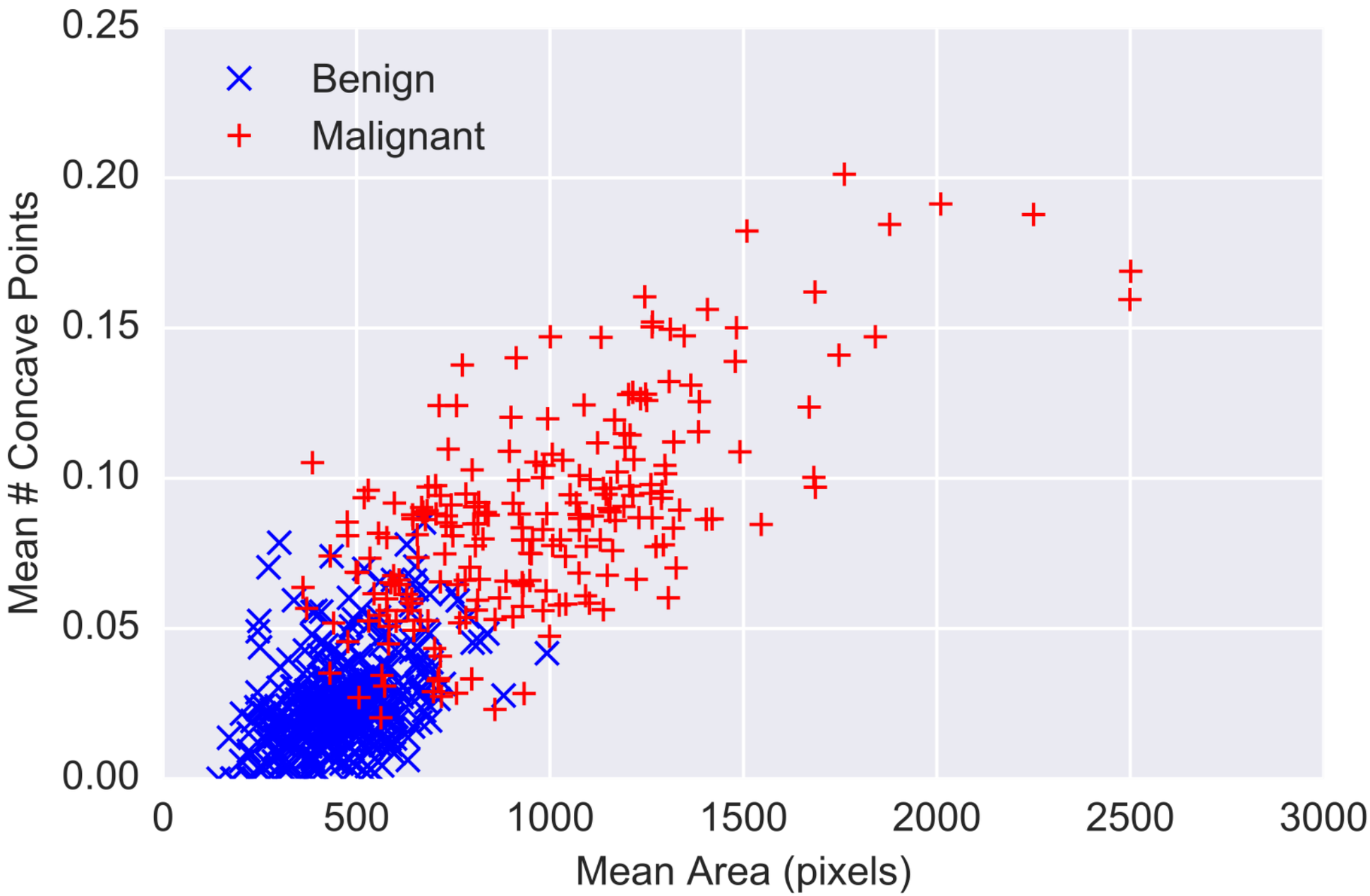
benign

or

malignant?

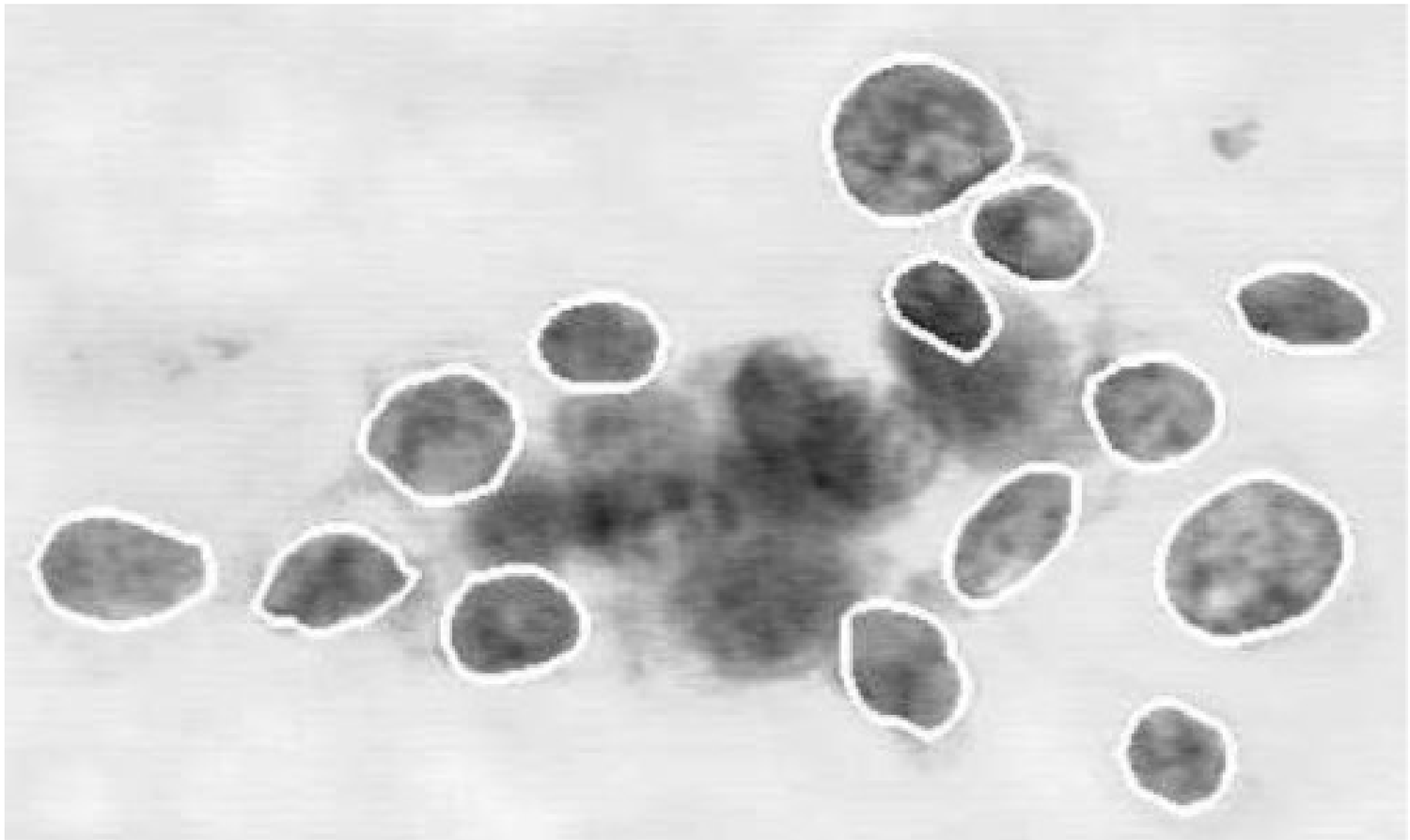
Source: J. Zico Kolter, 15-388/688, Practical Data Science

Breast Cancer



Source: J. Zico Kolter, 15-388/688, Practical Data Science

Breast Cancer



Is the tumor

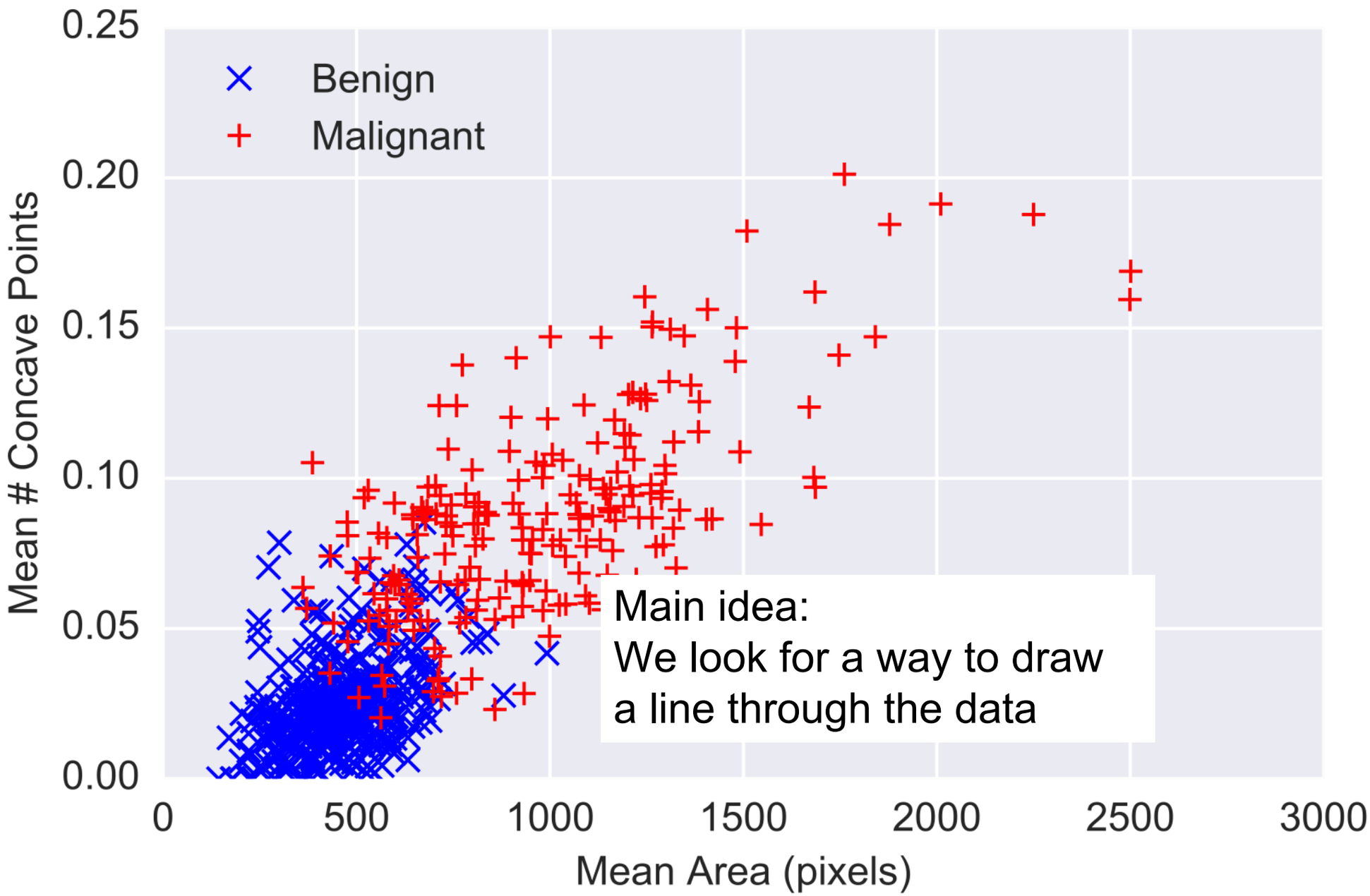
benign

or

malignant?

Source: J. Zico Kolter, 15-388/688, Practical Data Science

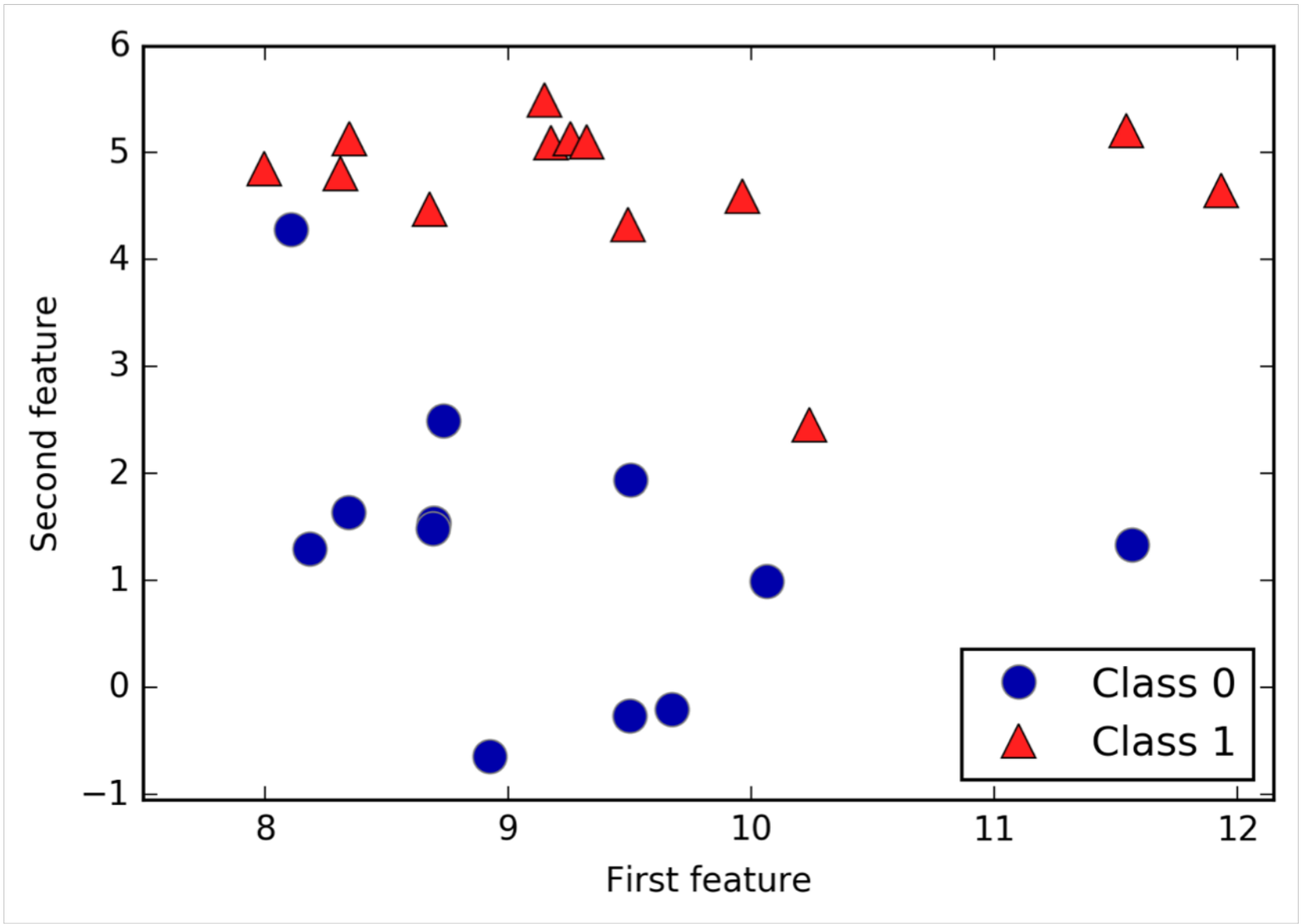
Breast Cancer



Source: J. Zico Kolter, 15-388/688, Practical Data Science

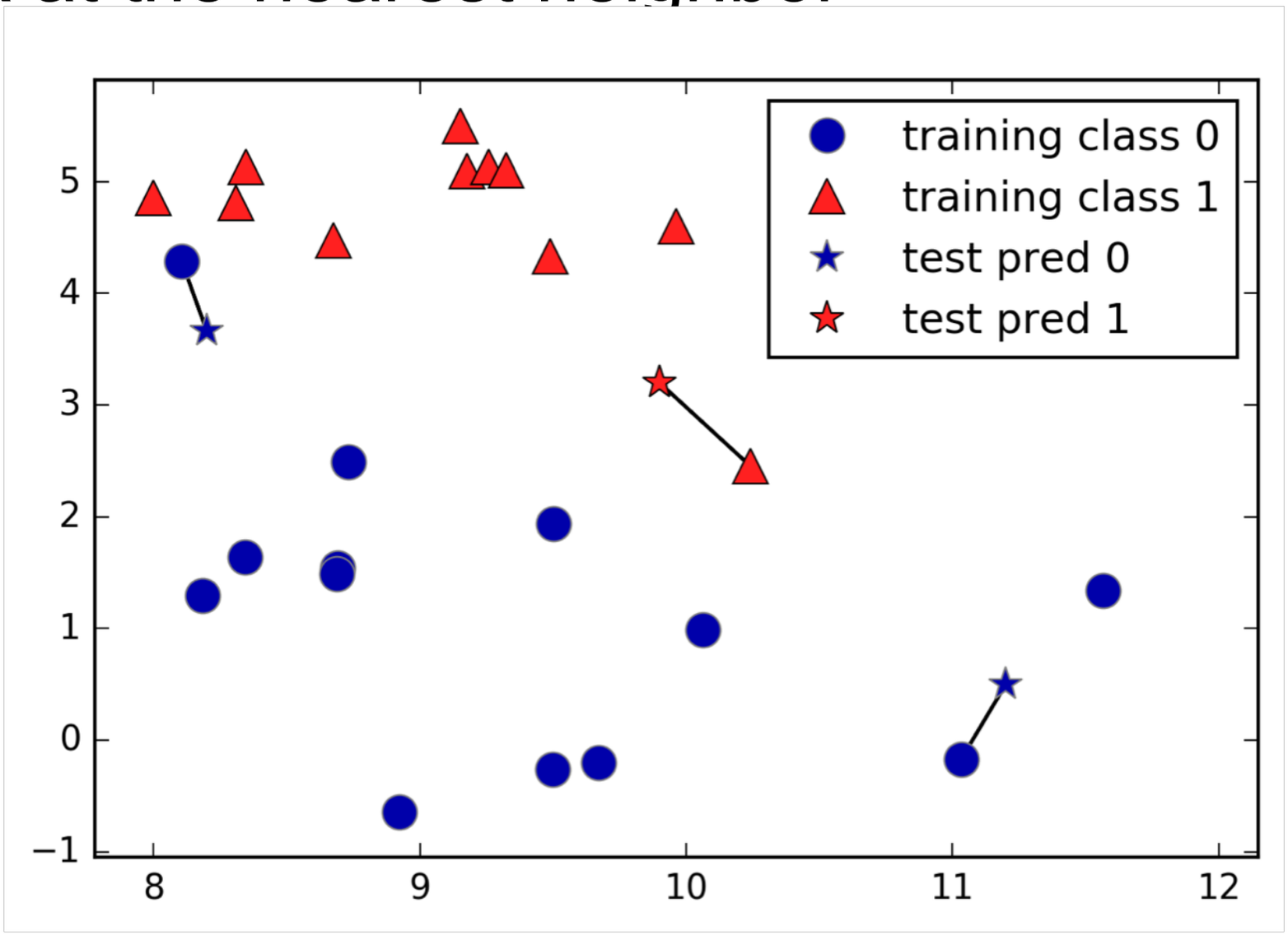
Classification

We want to predict the **class**: is it 0 or 1?



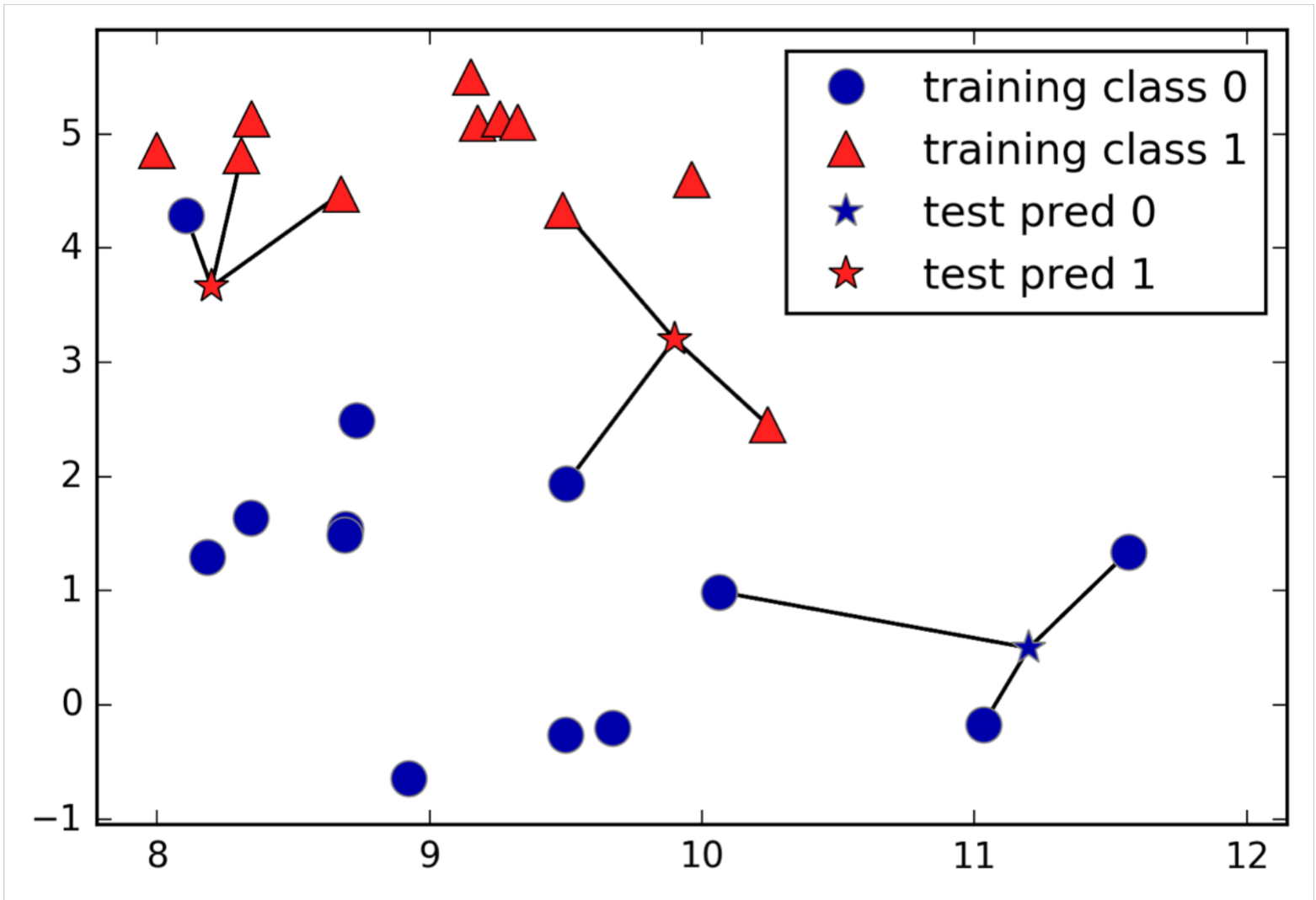
Classification: 1 -NN

We look at the nearest neighbor

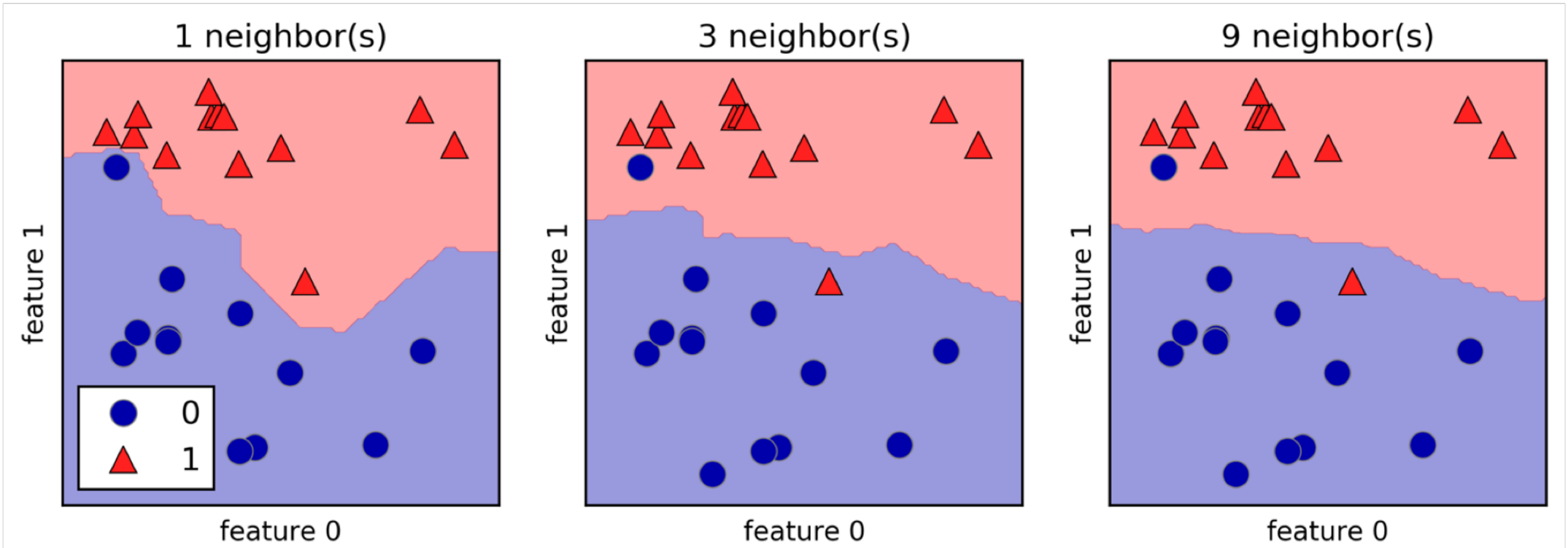


Classification: 3 -NN

We look at the **three** nearest neighbors

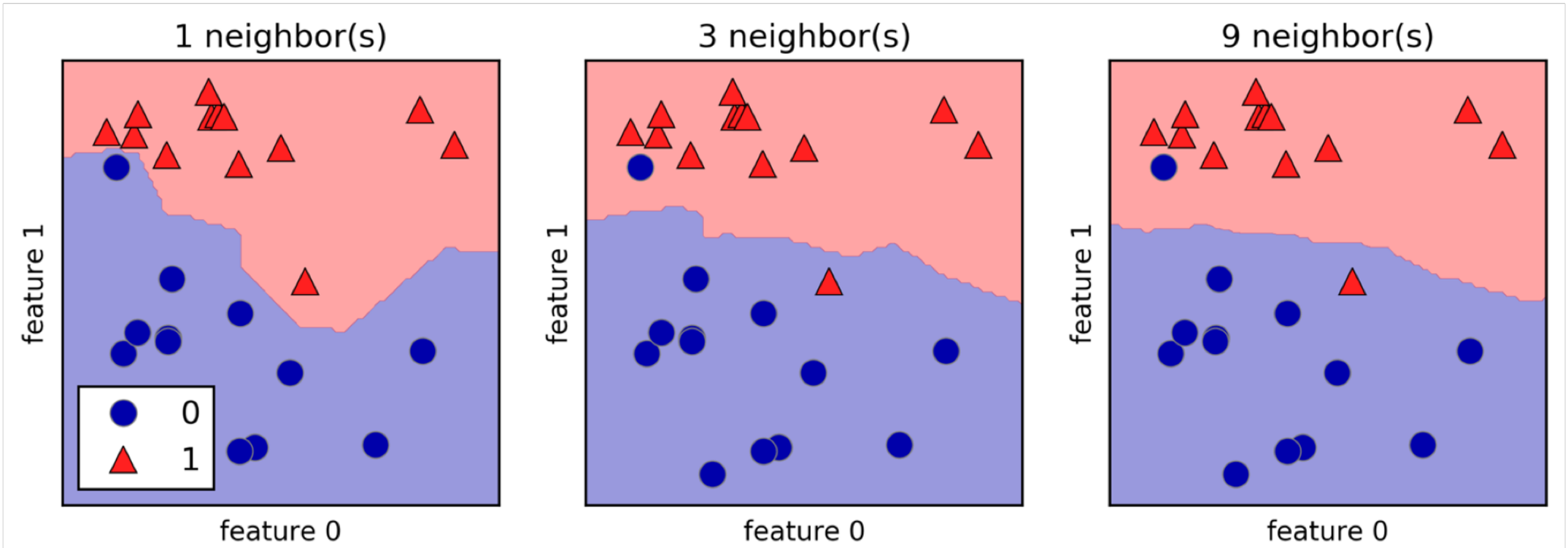


Nearest Neighbors: Different Ks



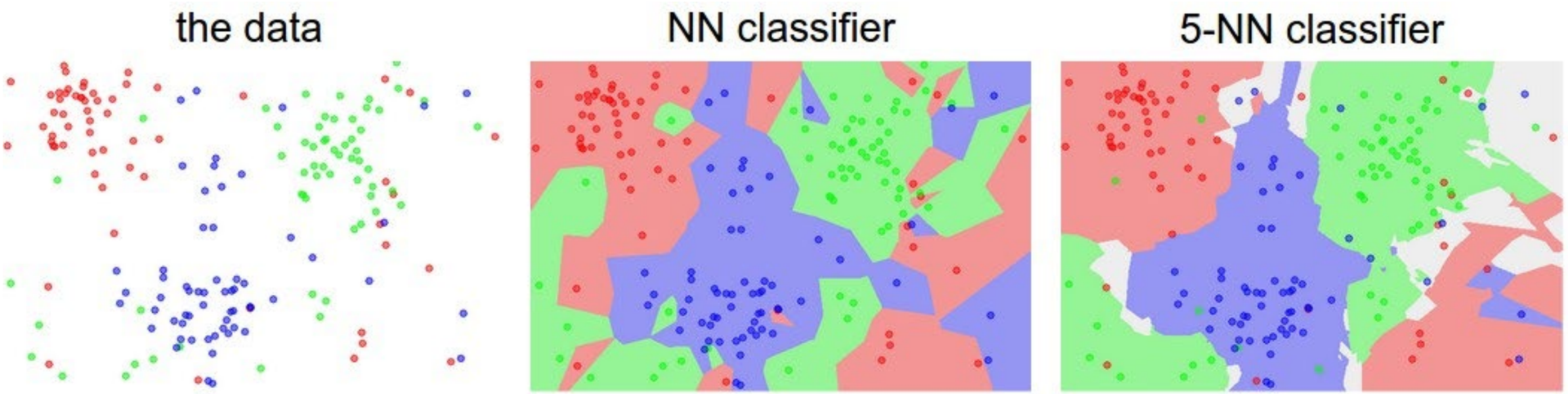
It is better to use odd numbers to do a **majority** vote.

Nearest Neighbors: Different Ks



If you use even numbers (2, 4, etc.), you introduce more randomness to your model. The system will resolve a draw by picking **randomly** .

Nearest Neighbors: Different Ks



Classification: Definitions

- Predict discrete-valued quantity y
- Binary classification:

  $y \in \{-1, +1\}$

  For example, $y \in$ {Malignant, Benign}

- Multiclass classification:

  $y \in \{1,2, \ldots, k\}$

  For example, $y \in$ {Cat, Dog, Mouse, … , Bird, Rabbit}

Classification: Definitions

- **c** is the concept to learn

  $c(x) \rightarrow 0/1,\ x \in X$

- **h** is a hypothesis, the result of the learning ("guessed c")

  $h(x) \rightarrow 0/1,\ x \in X$

- **H** is the hypotheses space, all conceivable hypotheses (before the data arrives)

  $h \in H$

- **D** is the set of available training data

  $D \subseteq X$

Classification: Definitions

- Positive example

  $x : c(x) = 1, x \in D$

- Negative example

$x : c(x) = 0, x \in D$

Classification example

- Sky $\in \{\text{Sunny}, \text{Cloudy}, \text{Rainy}\}$
- Temp $\in \{\text{Warm}, \text{Cold}\}$
- Wind $\in \{\text{Windy}, \text{Calm}\}$
- Humid $\in \{\text{Humid}, \text{Dry}\}$

Thus, the total number of possible weather conditions is:

$$|X| = 3 \cdot 2 \cdot 2 \cdot 2 = 24$$

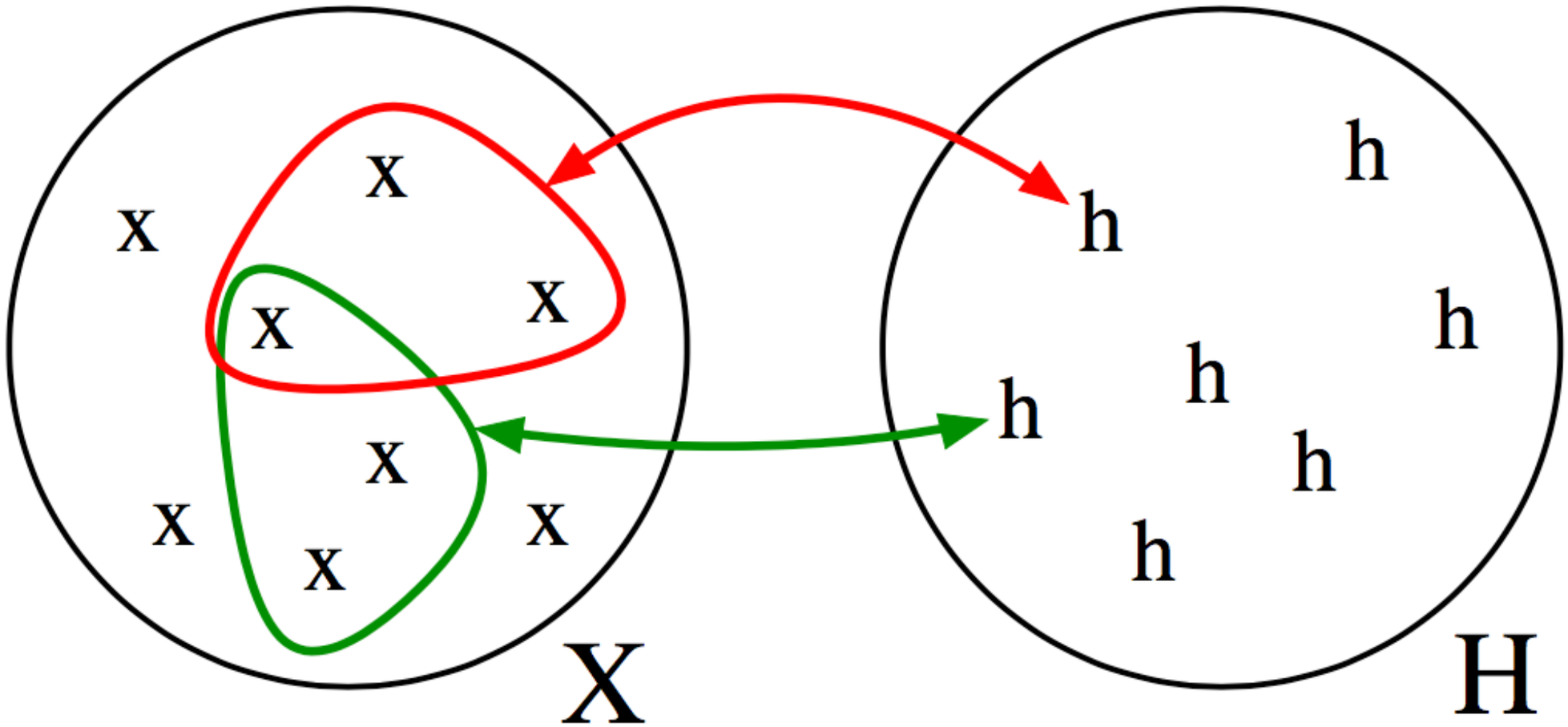
Classification example

Some typical training samples:

If c(x) → Nice or Bad

- <Sunny, Warm, Windy, Dry> → Nice
- <Sunny, Warm, Windy, Humid> → Nice
- <Rainy, Cold, Windy, Humid> → Bad
- <Sunny, Warm, Calm, Humid> → Nice

Hypotheses Space



How many hypotheses can we choose from?

- $|H| = 2^{|X|}$

In the weather example, the total number of possible weather conditions was

- $|X| = 24$
- $|H| = 2^{24} = 16777216$

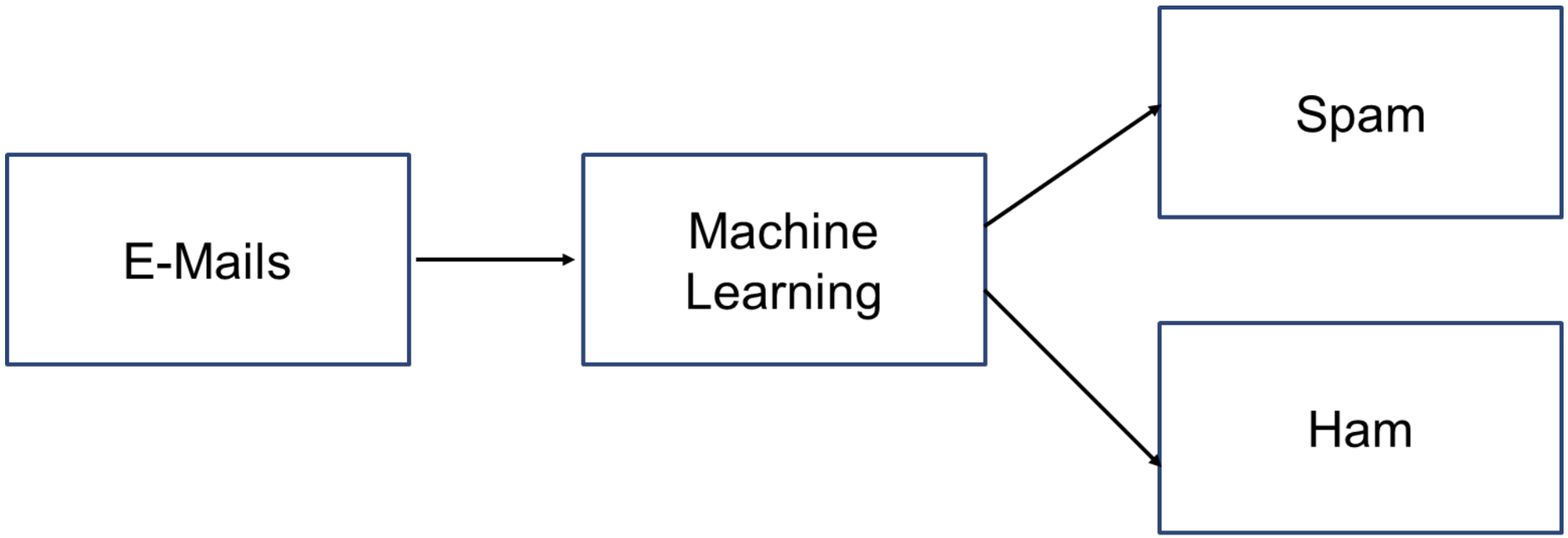
How many hypotheses can we choose from?

- $|H| = 2^{|X|}$

In the weather example, the total number of possible weather conditions was

- $|X| = 24$
- $|H| = 2^{24} = 16777216$
- **Thus, it is necessary to make restrictions!**

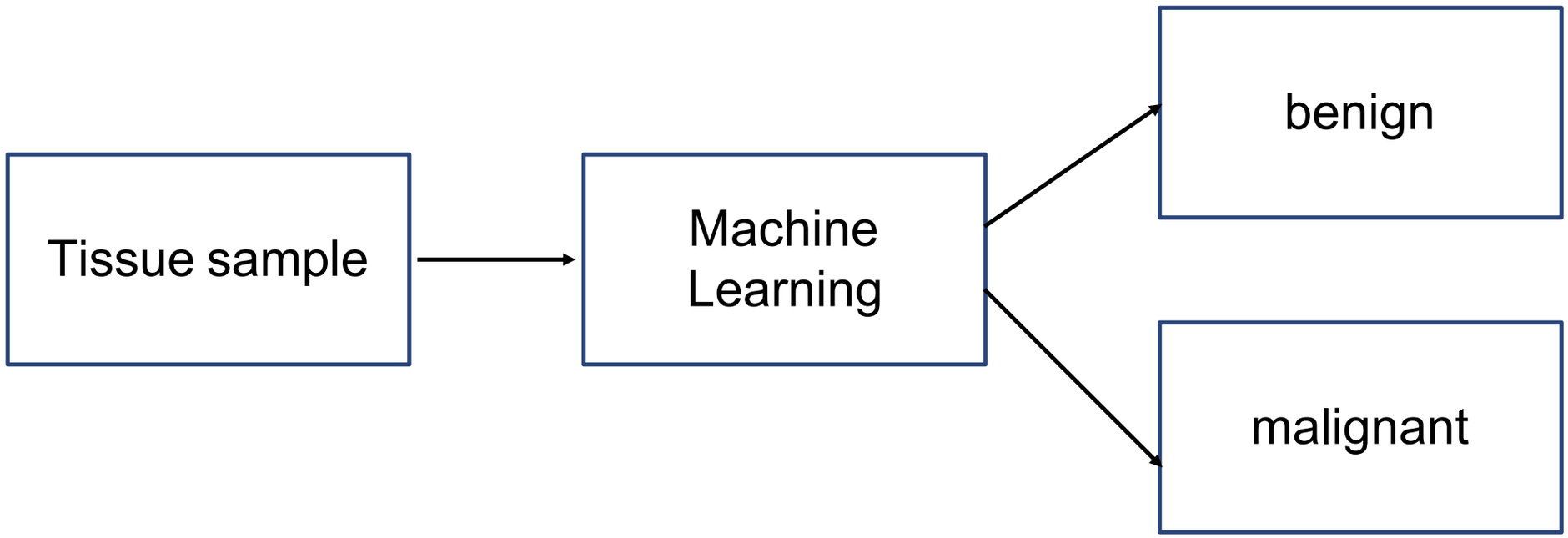
Example: Spam Filter



X

y

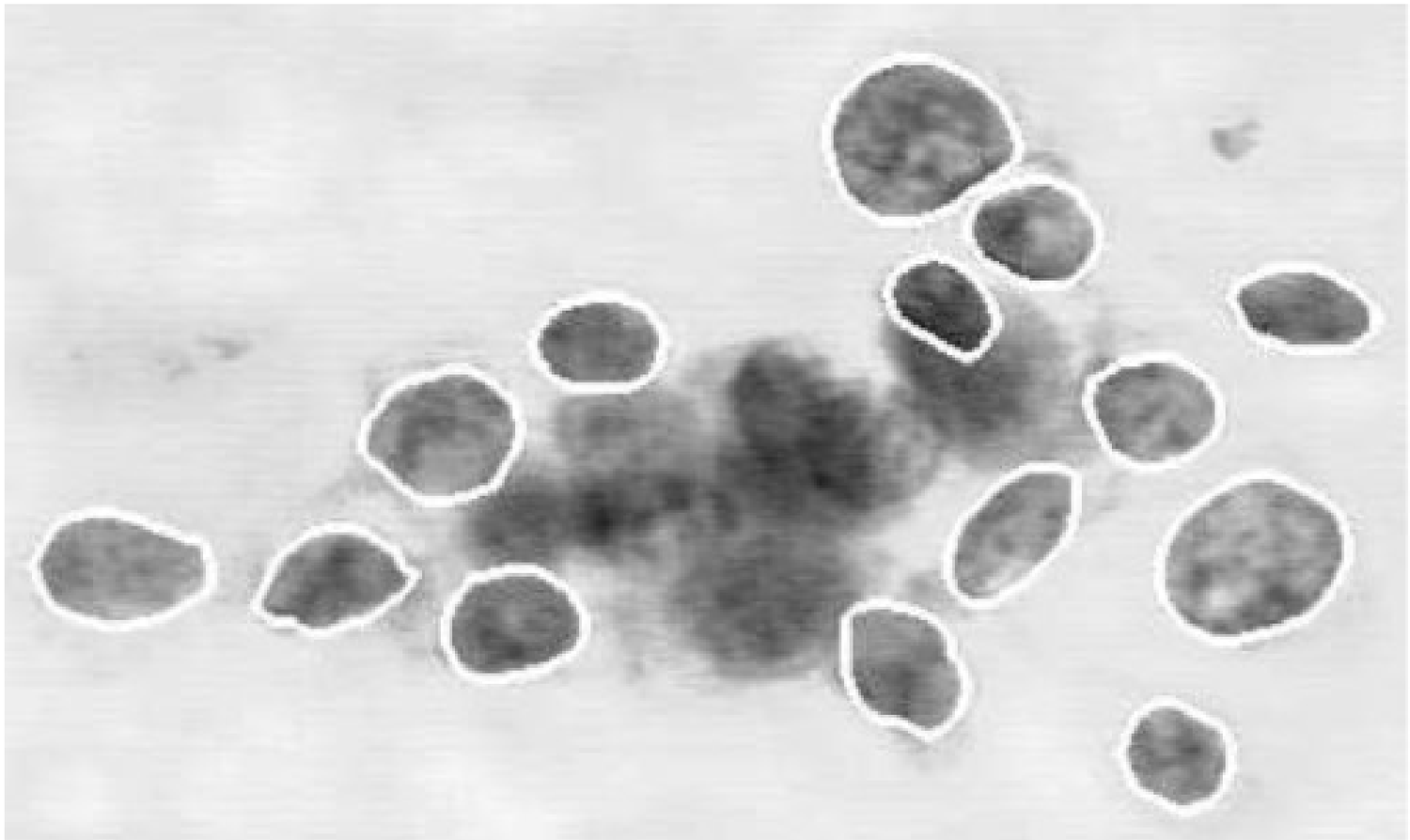
Example: Breast Cancer



X

y

Breast Cancer



Is the tumor

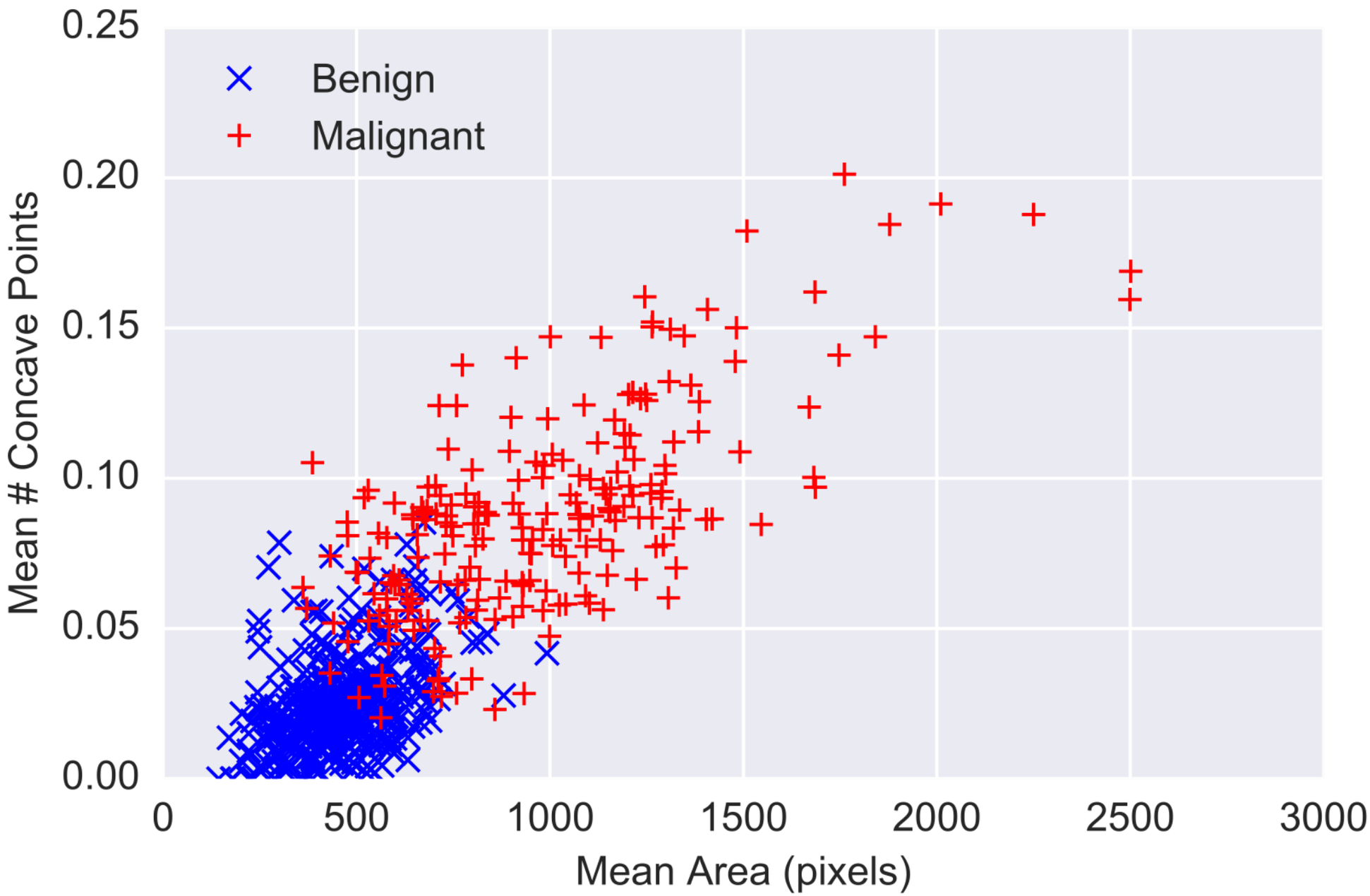
benign

or

malignant?

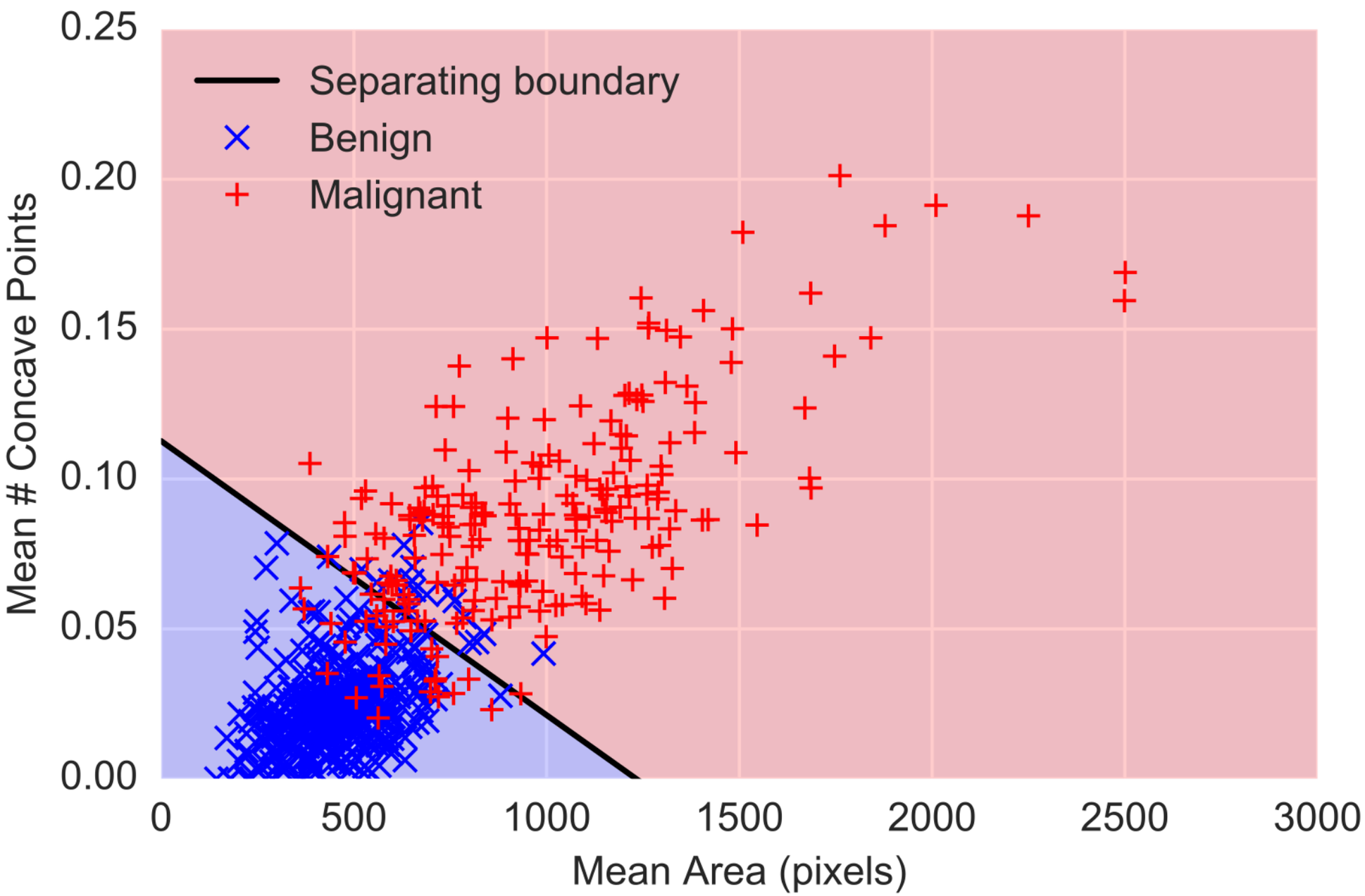
Source: J. Zico Kolter, 15-388/688, Practical Data Science

Breast Cancer



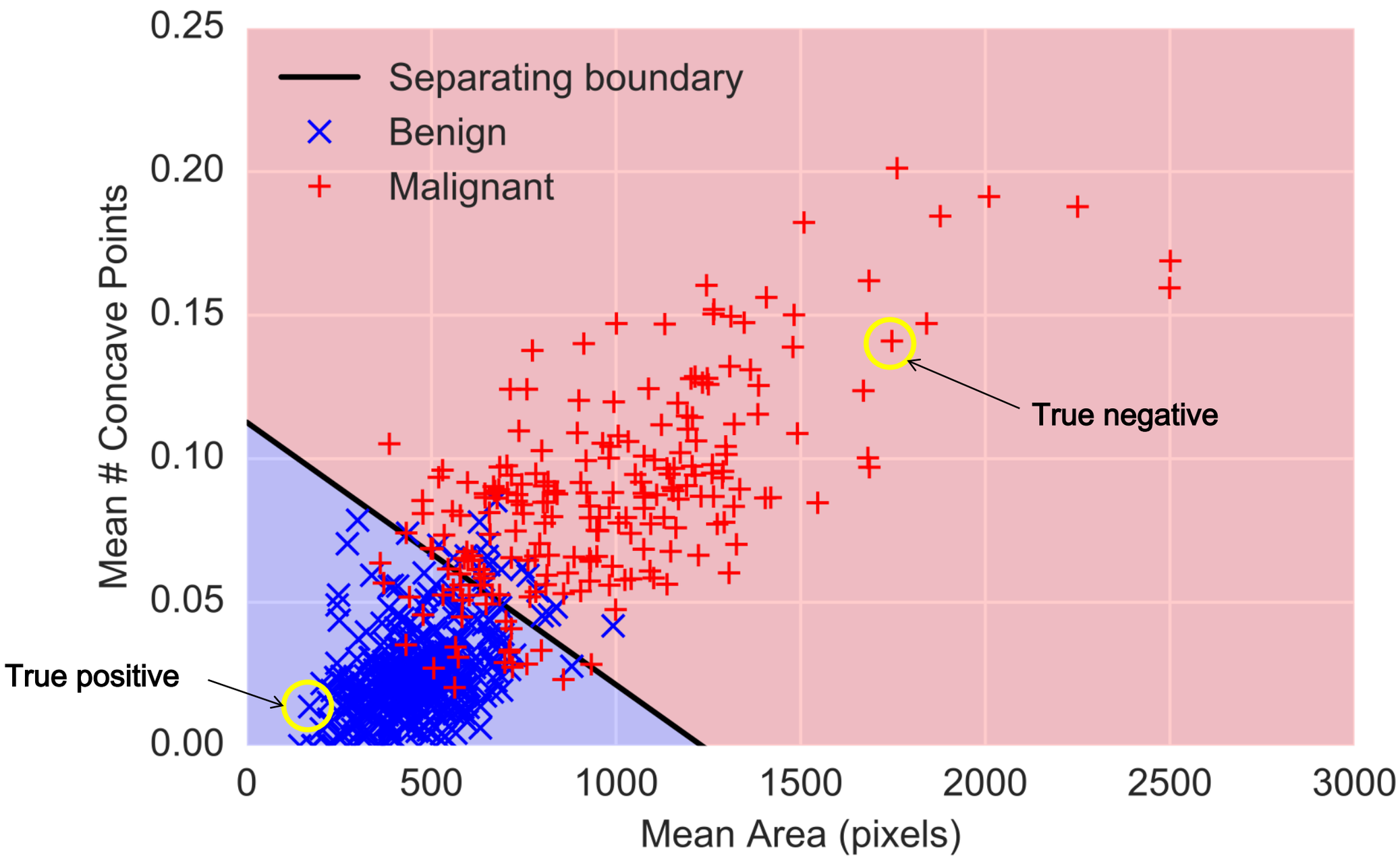
Source: J. Zico Kolter, 15-388/688, Practical Data Science

Breast Cancer



Source: J. Zico Kolter, 15-388/688, Practical Data Science

Breast Cancer



Source: J. Zico Kolter, 15-388/688, Practical Data Science

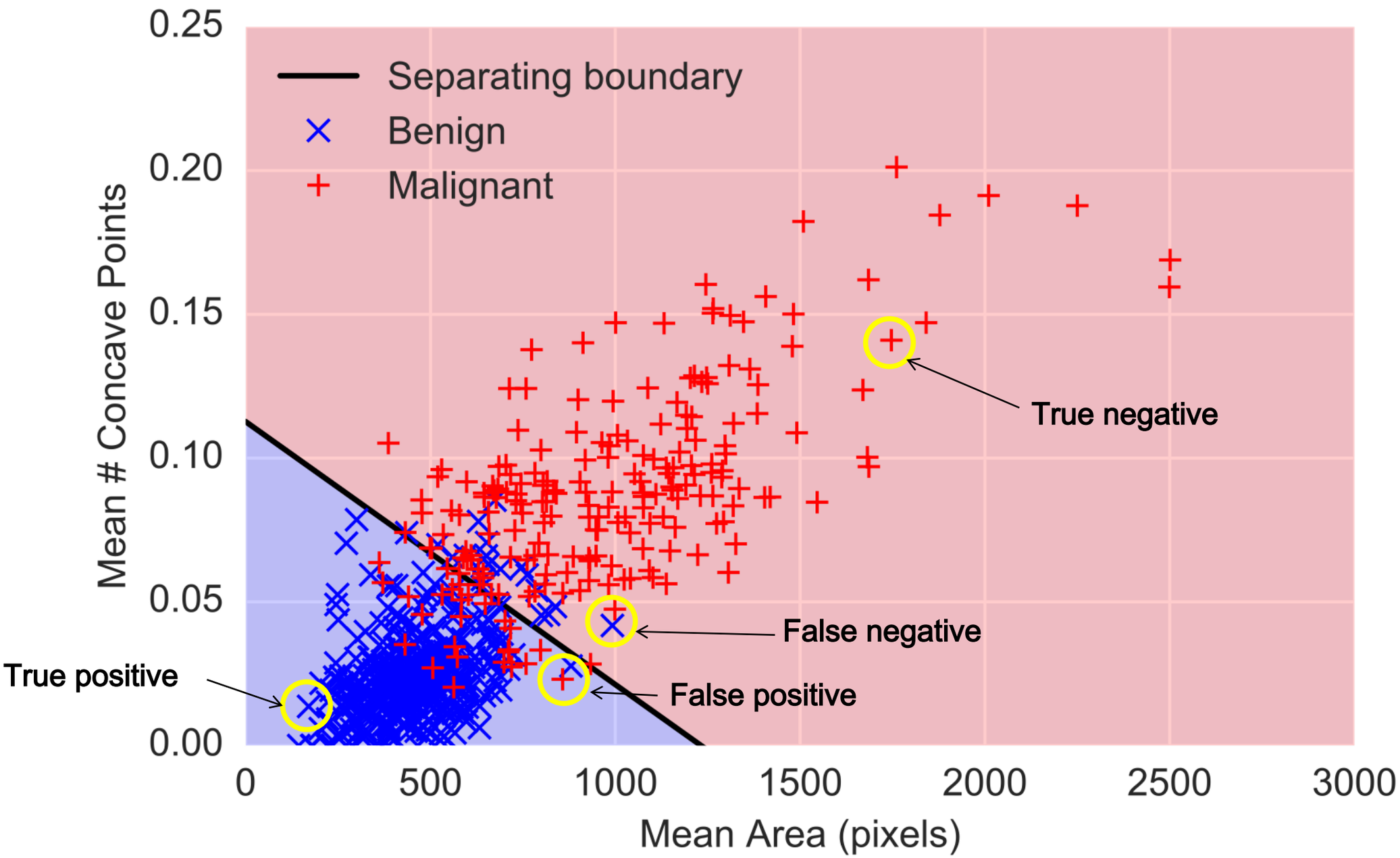
Accuracy

For the cancer example with {benign, malignant}

- True Positives (TP) = elements correctly predicted as benign
- True Negatives (TN) = elements correctly predicted as malignant

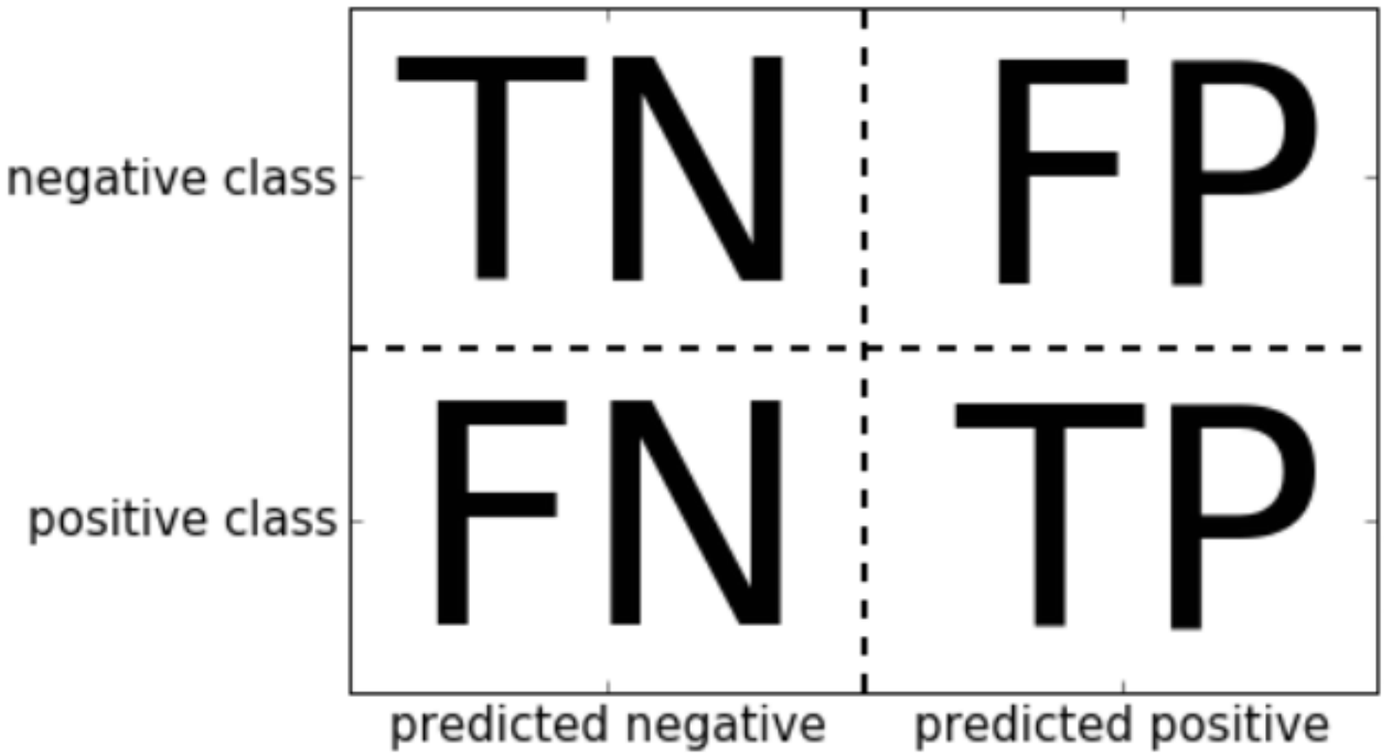
- Accuracy = **(TP + TN) / all elements**

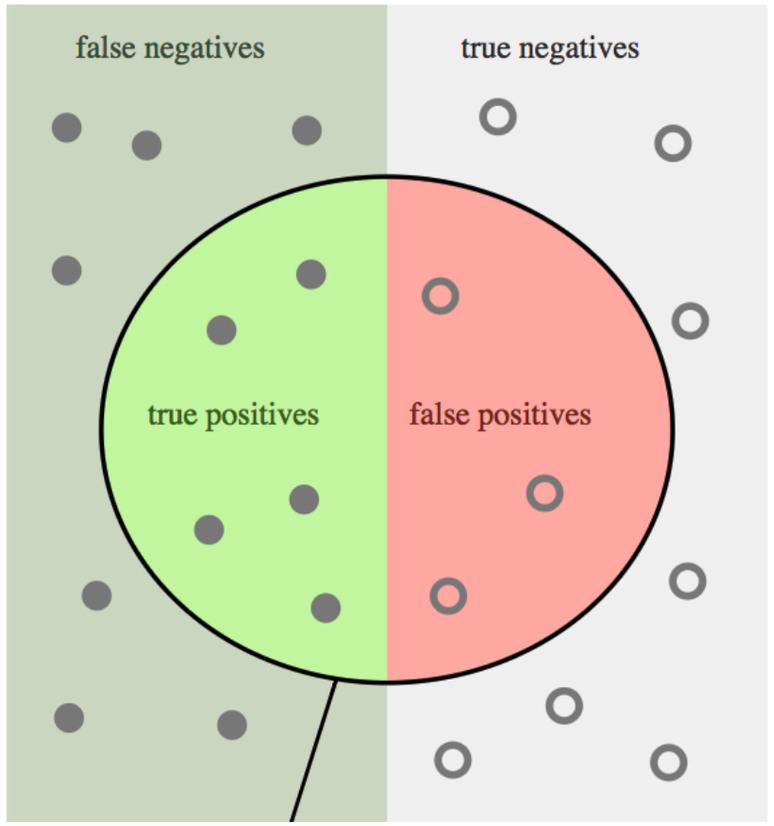
Breast Cancer



Source: J. Zico Kolter, 15-388/688, Practical Data Science

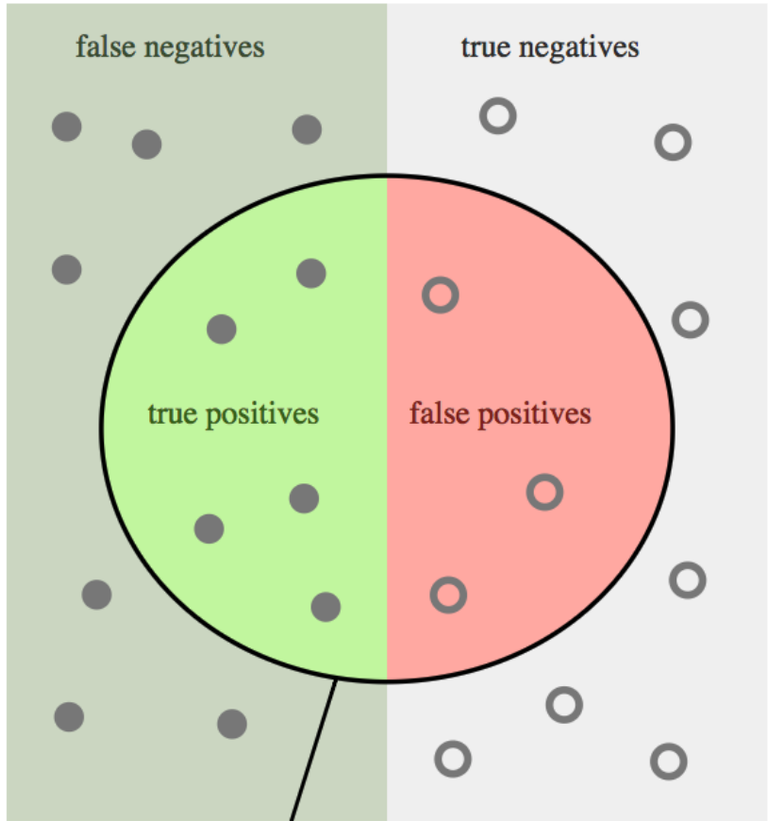
Confusion Matrix





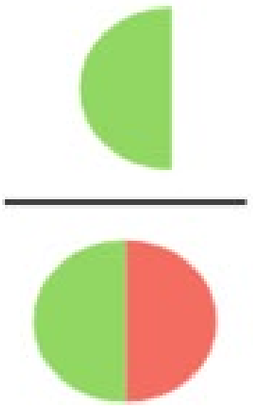
https://en.wikipedia.org/wiki/Precision_and_recall

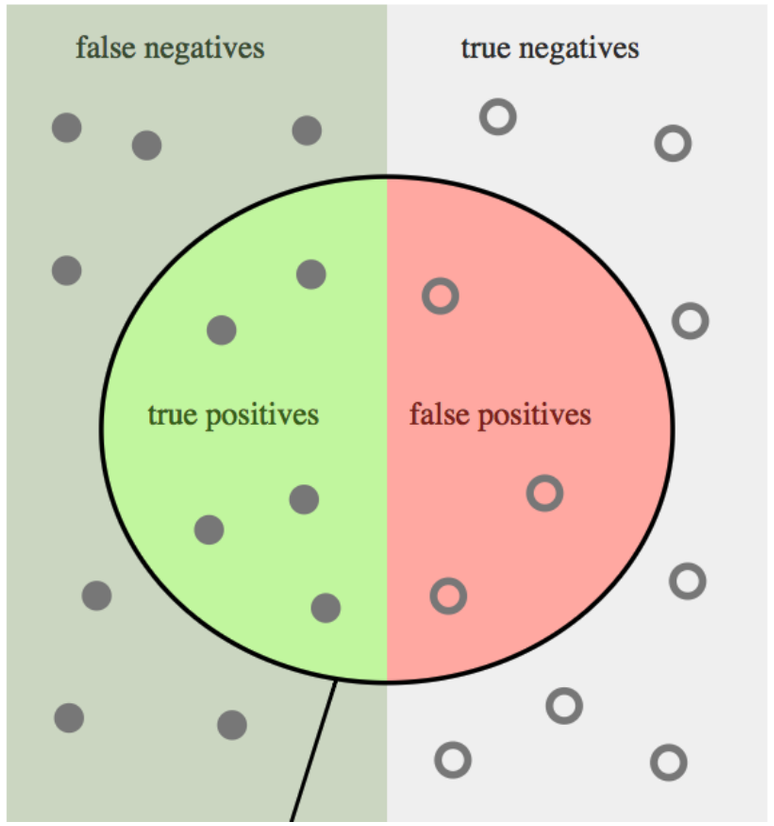
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$



The “positives” refer to the value you choose as **relevant** .
In the breast cancer example, we took “benign” as the relevant class.

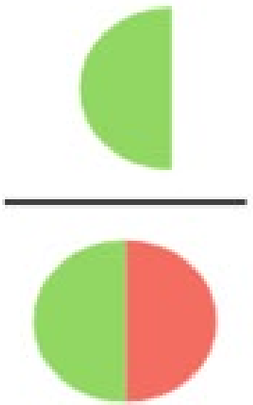
$$\text{Precision} = \frac{TP}{TP+FP}$$

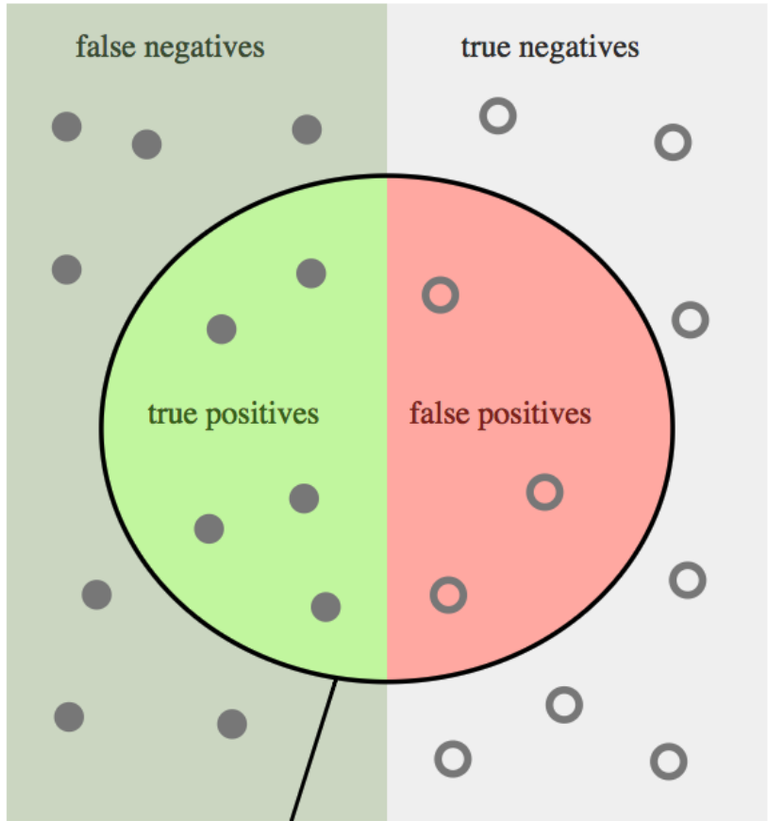




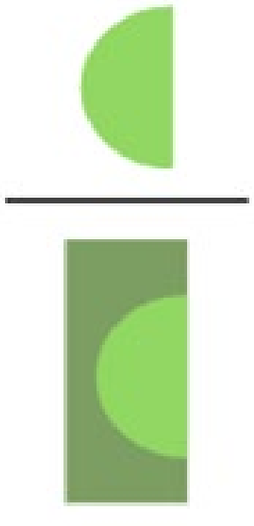
The “positives” refer to the value you choose as **relevant** .
In the breast cancer example, we took “benign” as the relevant class.

$$\text{Precision} = \frac{TP}{TP+FP}$$

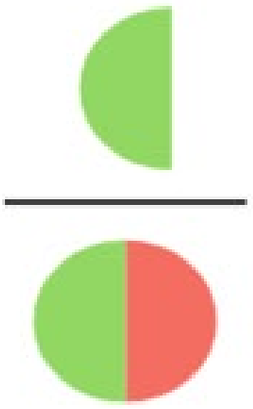


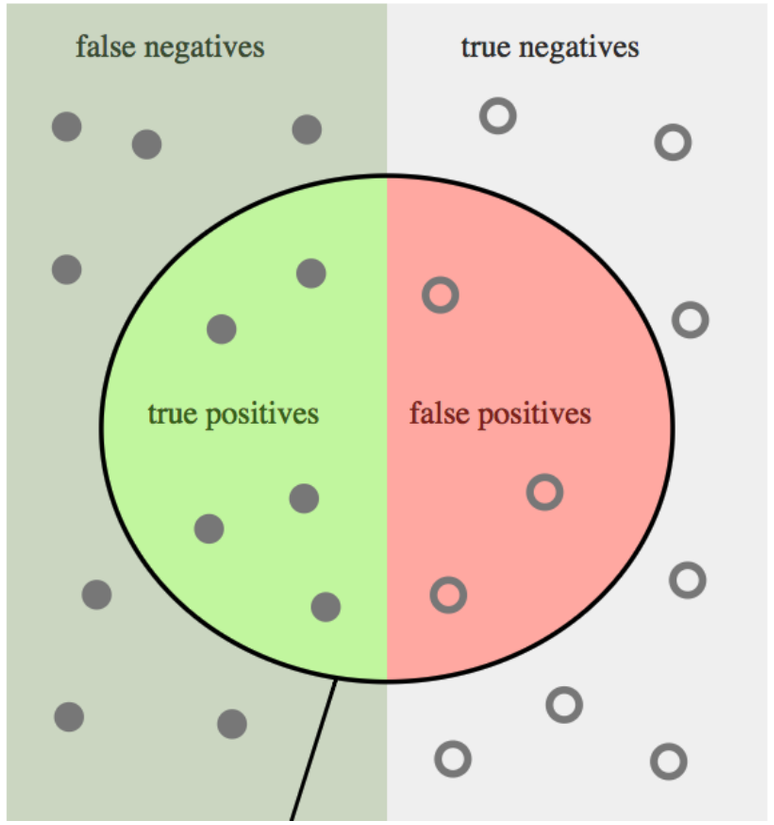


$$\text{Recall} = \frac{TP}{TP+FN}$$



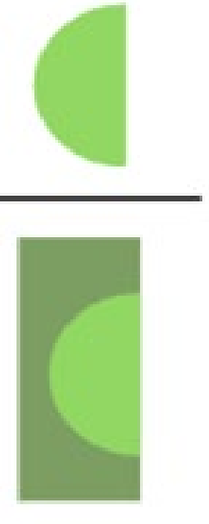
$$\text{Precision} = \frac{TP}{TP+FP}$$



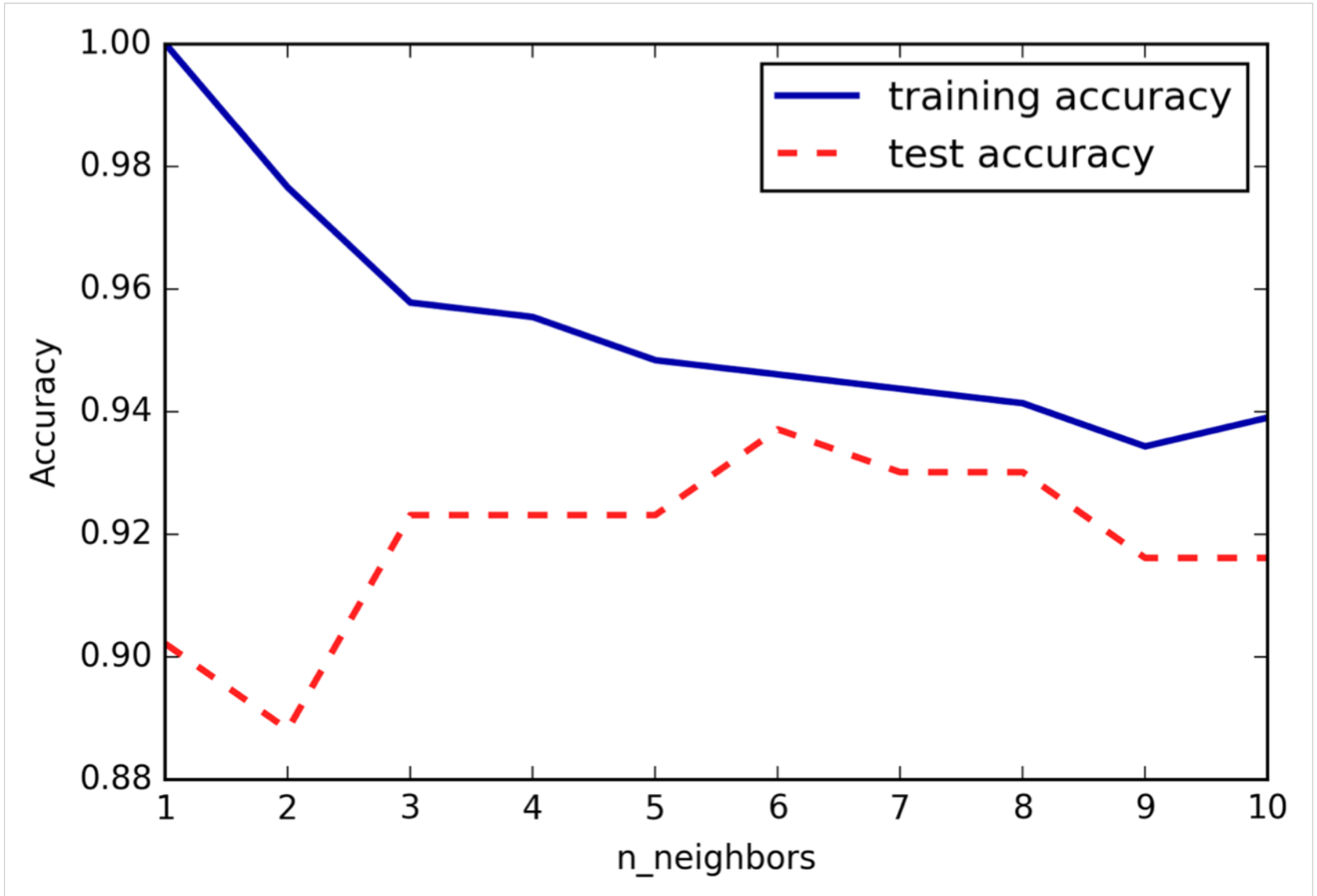


$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

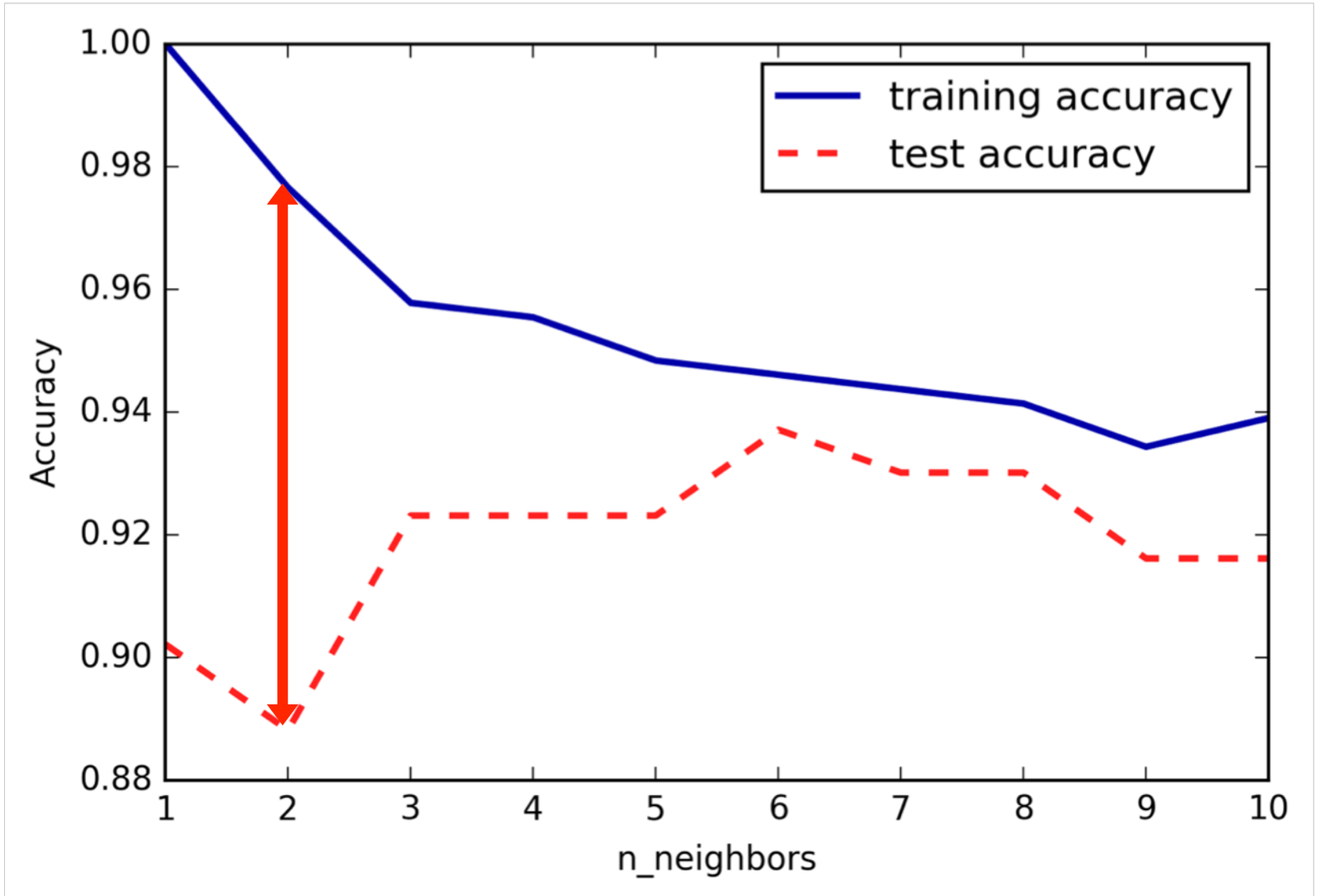


Accuracy



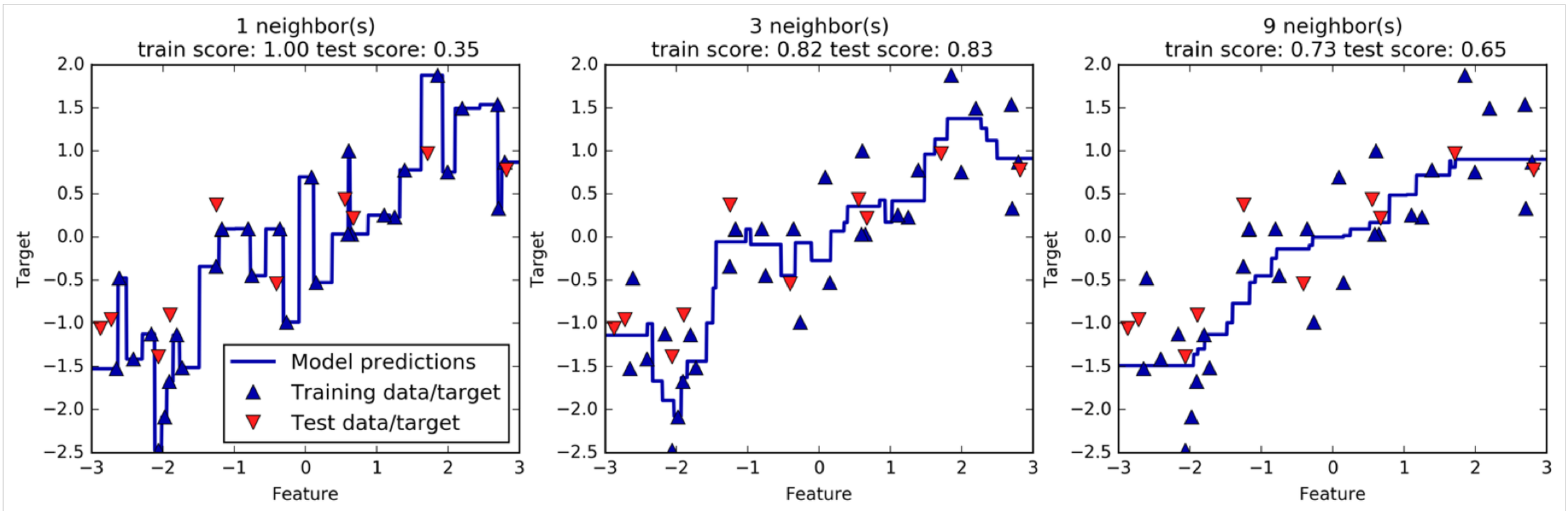
We separate our dataset in *training data* and *test data*.
We train the model with the training data (e.g., 80%) and use the testing data (e.g., 20%) to estimate the generalization capability of our model.

Accuracy



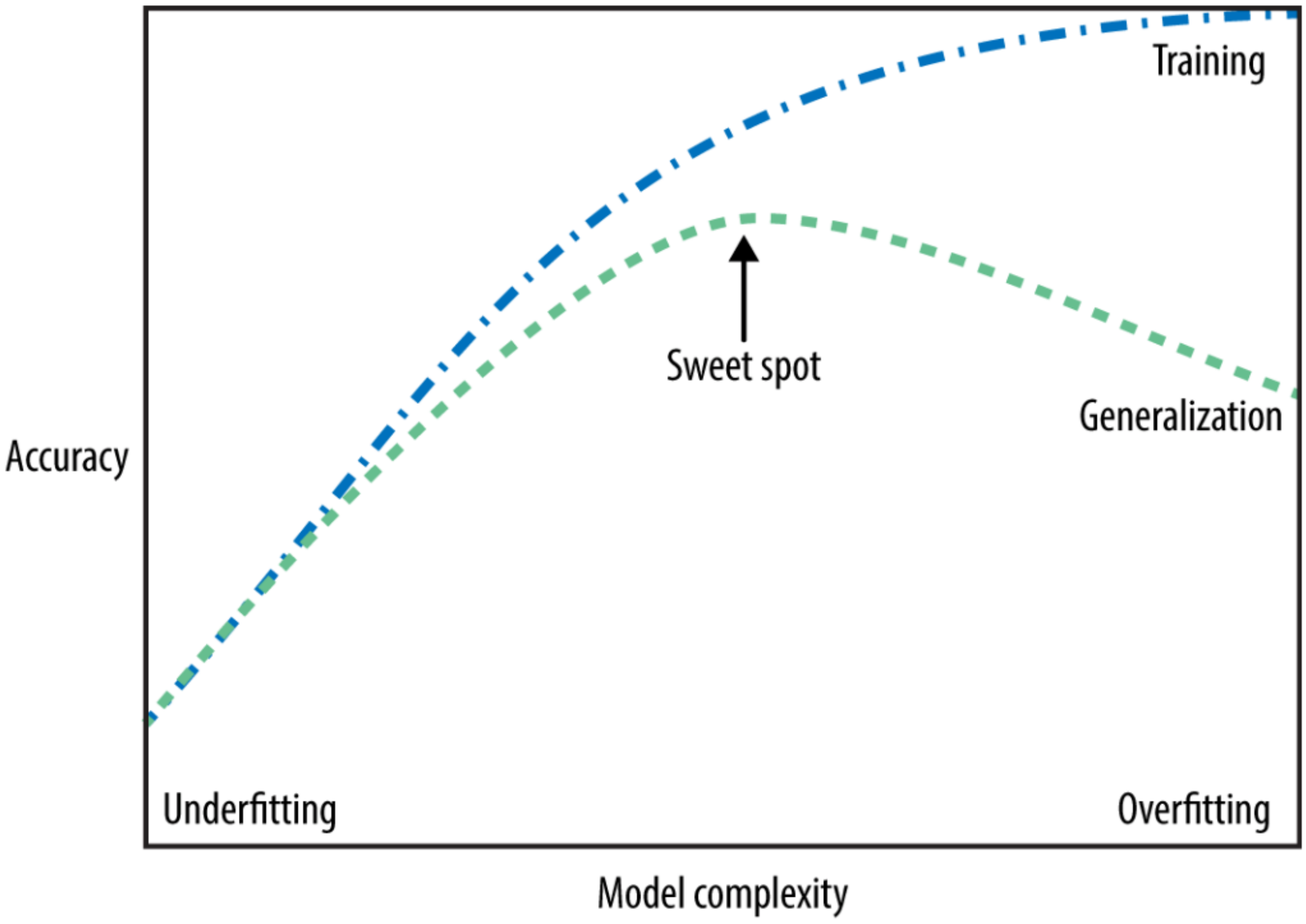
Generali zation error (out-of-sample error)

Accuracy



Overfitting
(too sensitive)

Underfitting
(not learning enough)


Training
Sweet spot
Generalization
Accuracy
Underfitting
Overfitting
Model complexity

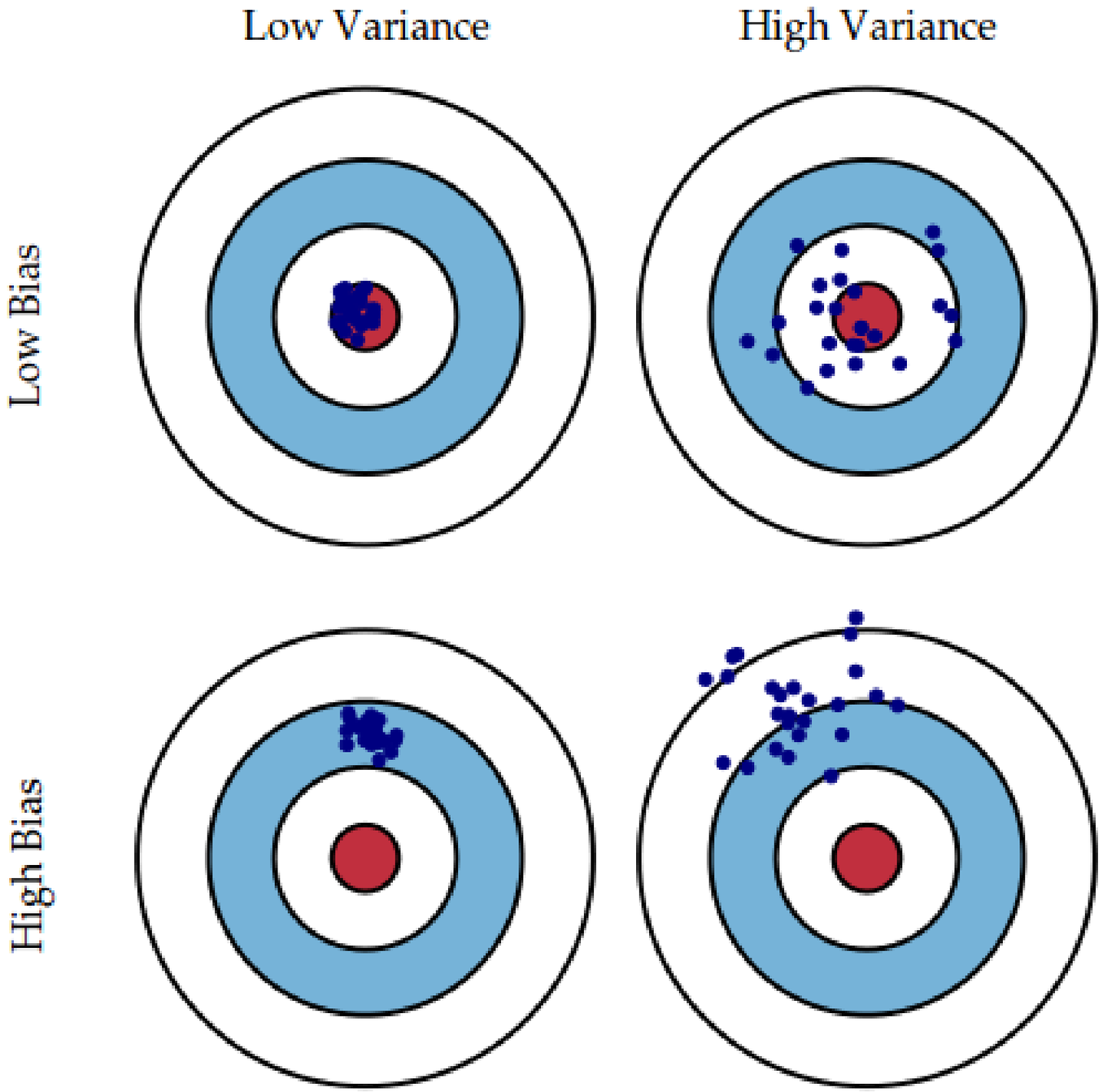
Bias

- Error from erroneous assumptions in the learning algorithm
- It leads to **underfitting**

Variance

- Error from sensitivity to small fluctuations in the training set
- It leads to **overfitting**

Bias-Variance Tradeoff

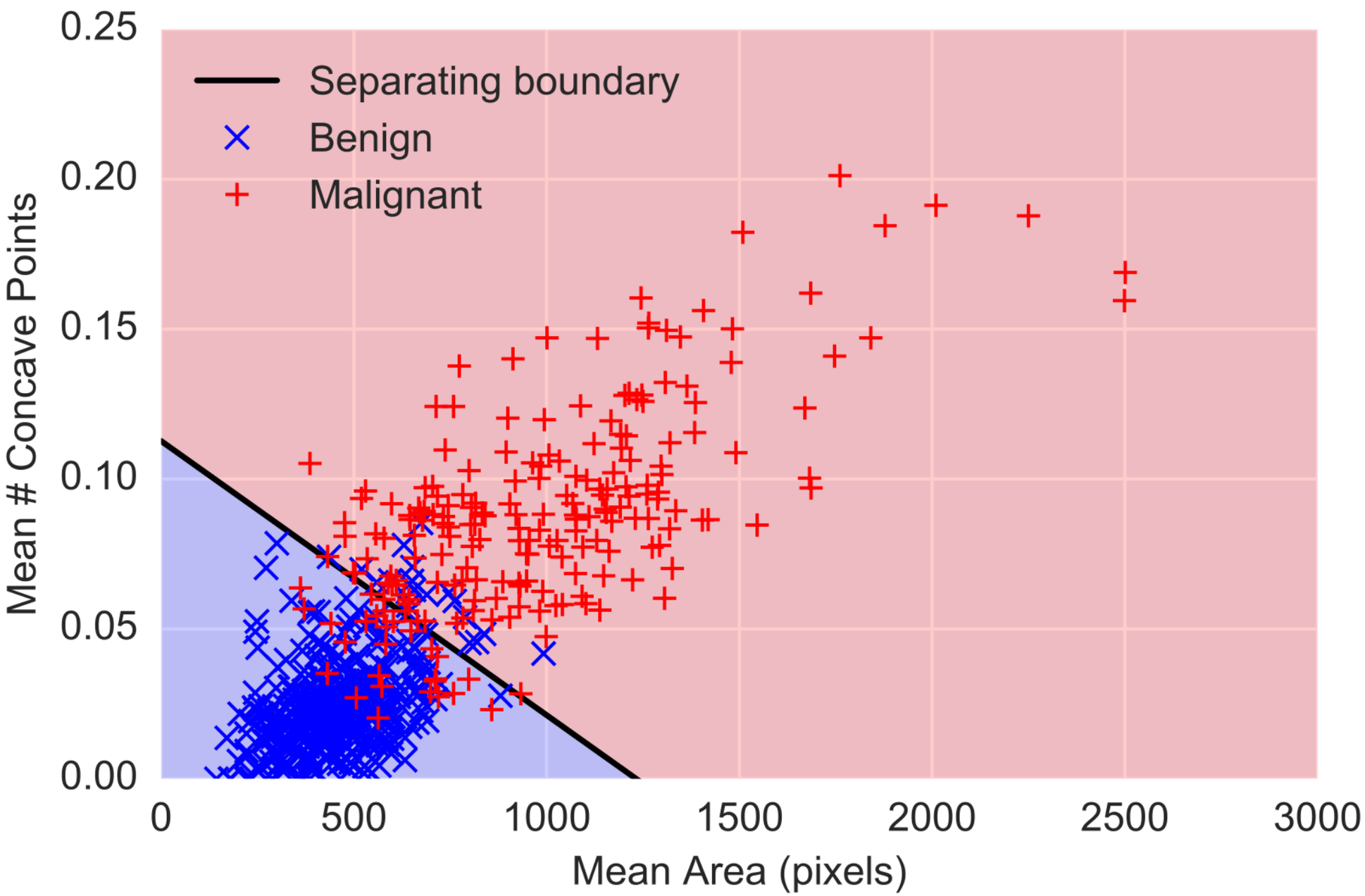


- **Bias:** Error from erroneous assumptions in the learning algorithm

- **Variance:** Error from sensitivity to small fluctuations in the training set

Source: ghetmir.blogspot.de/2013/04/bias -variance-tradeoff-in-human.html

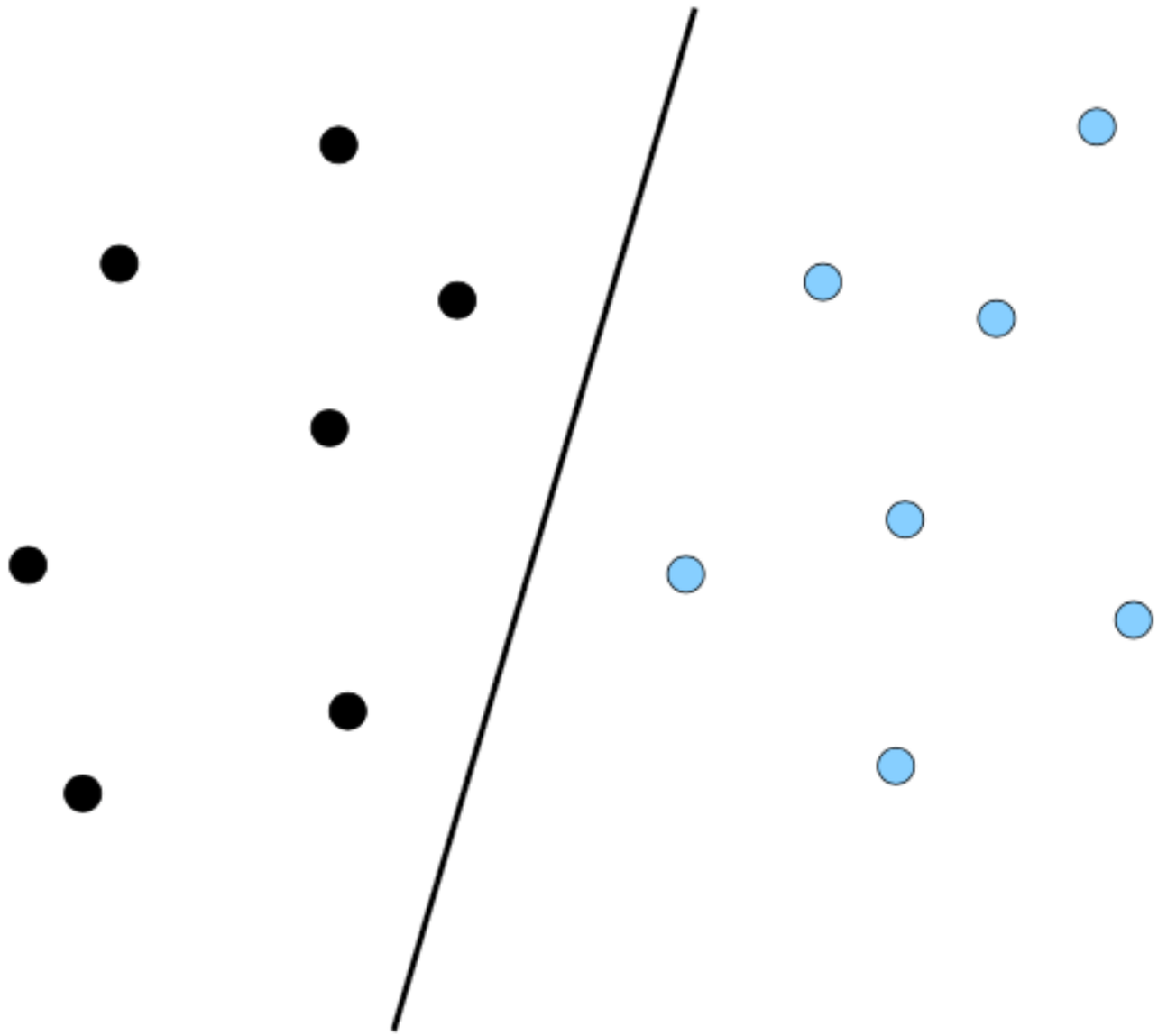
Breast Cancer



Source: J. Zico Kolter, 15-388/688, Practical Data Science

Separating your dataset in 2 dimensions?

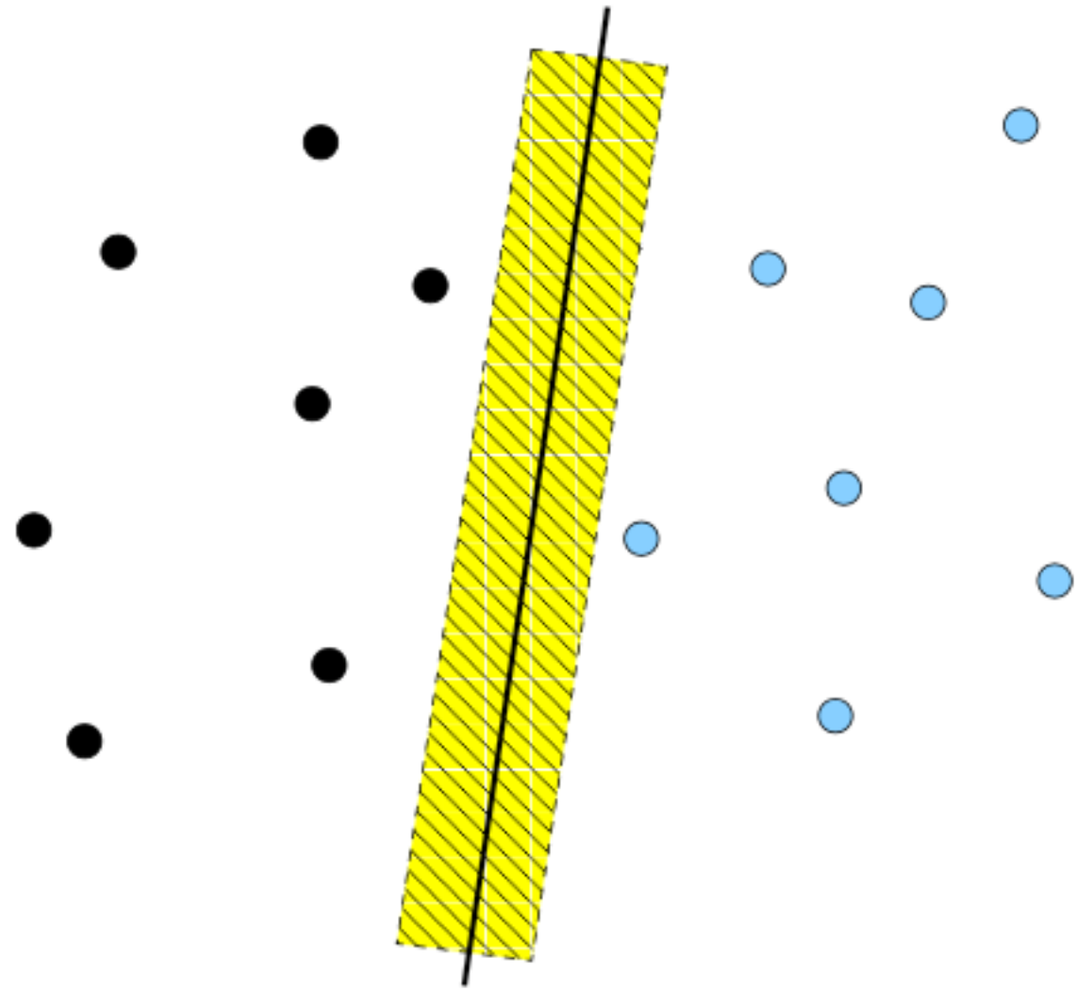
- Line



The Idea

Separating Hyperplane

$$\vec{w}^T\vec{x} + b = 0$$



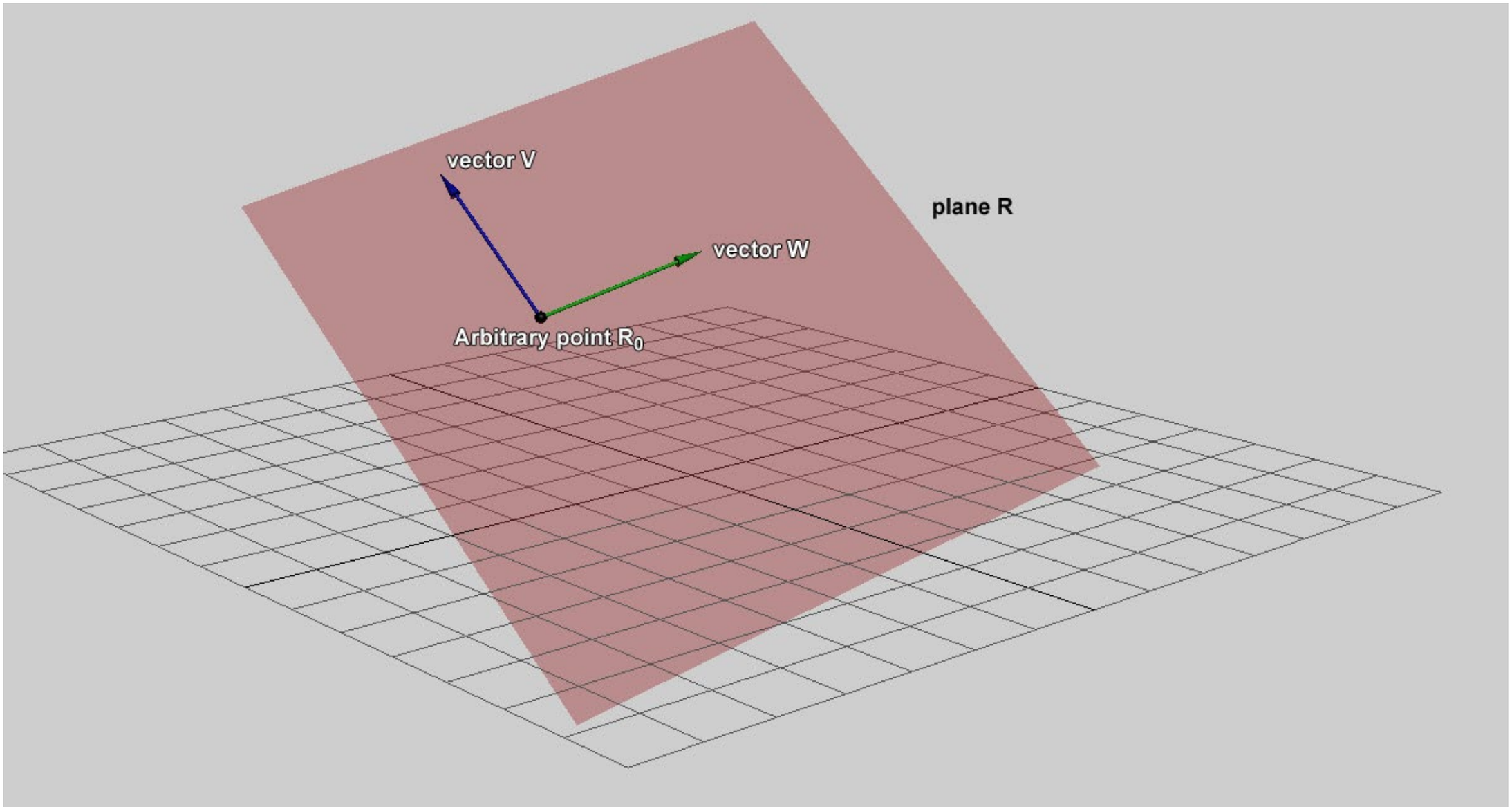
Separating Hyperplane

$$\vec{w}^T\vec{x} + b \geq 1 \qquad \text{when } t = 1$$

$$\vec{w}^T\vec{x} + b \leq -1 \qquad \text{when } t = -1$$

Separating your dataset in 3 dimensions?

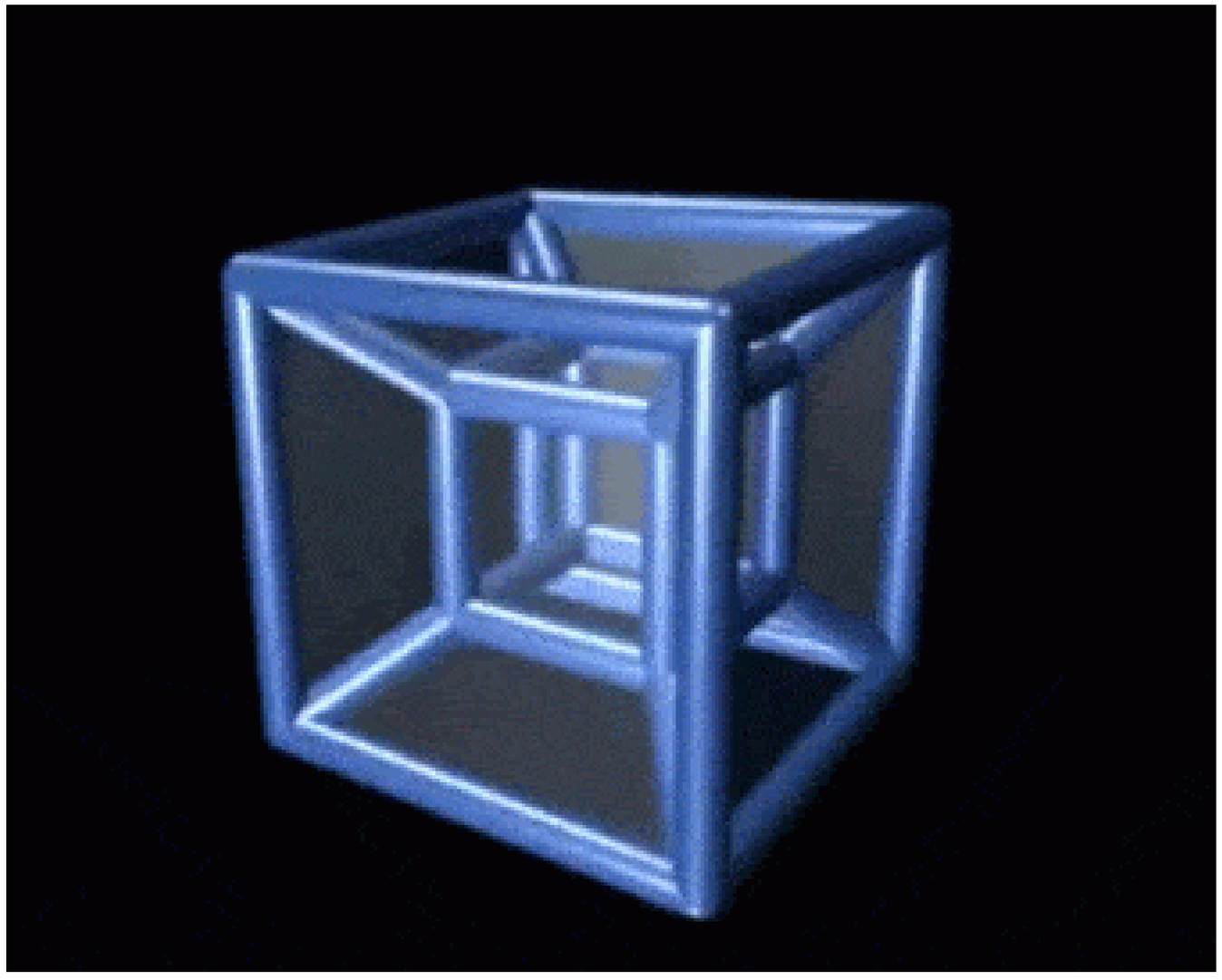
- Plane



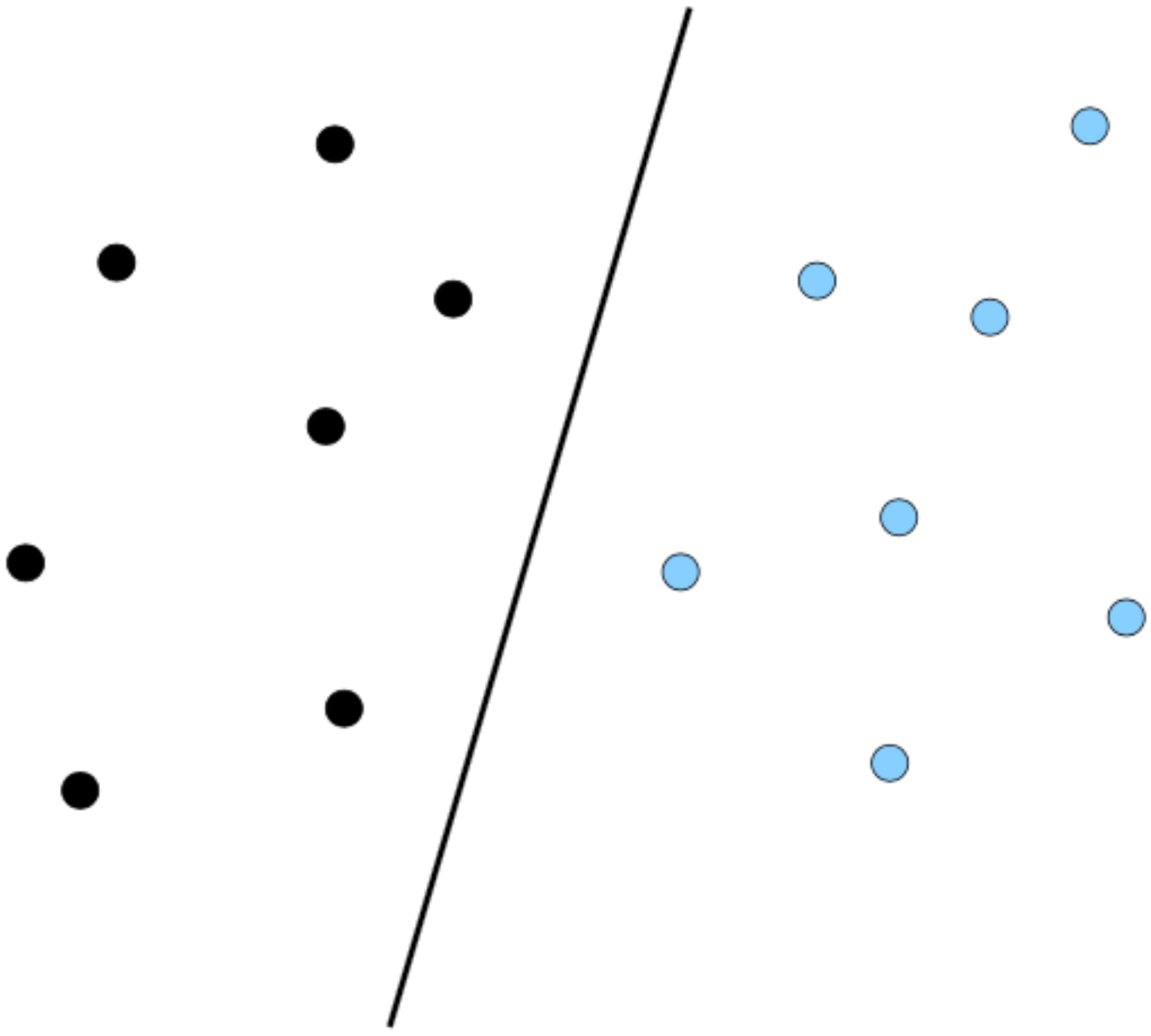
https://en.wikipedia.org/wiki/Plane_(geometry)

Separating your dataset in 4+ dimensions?

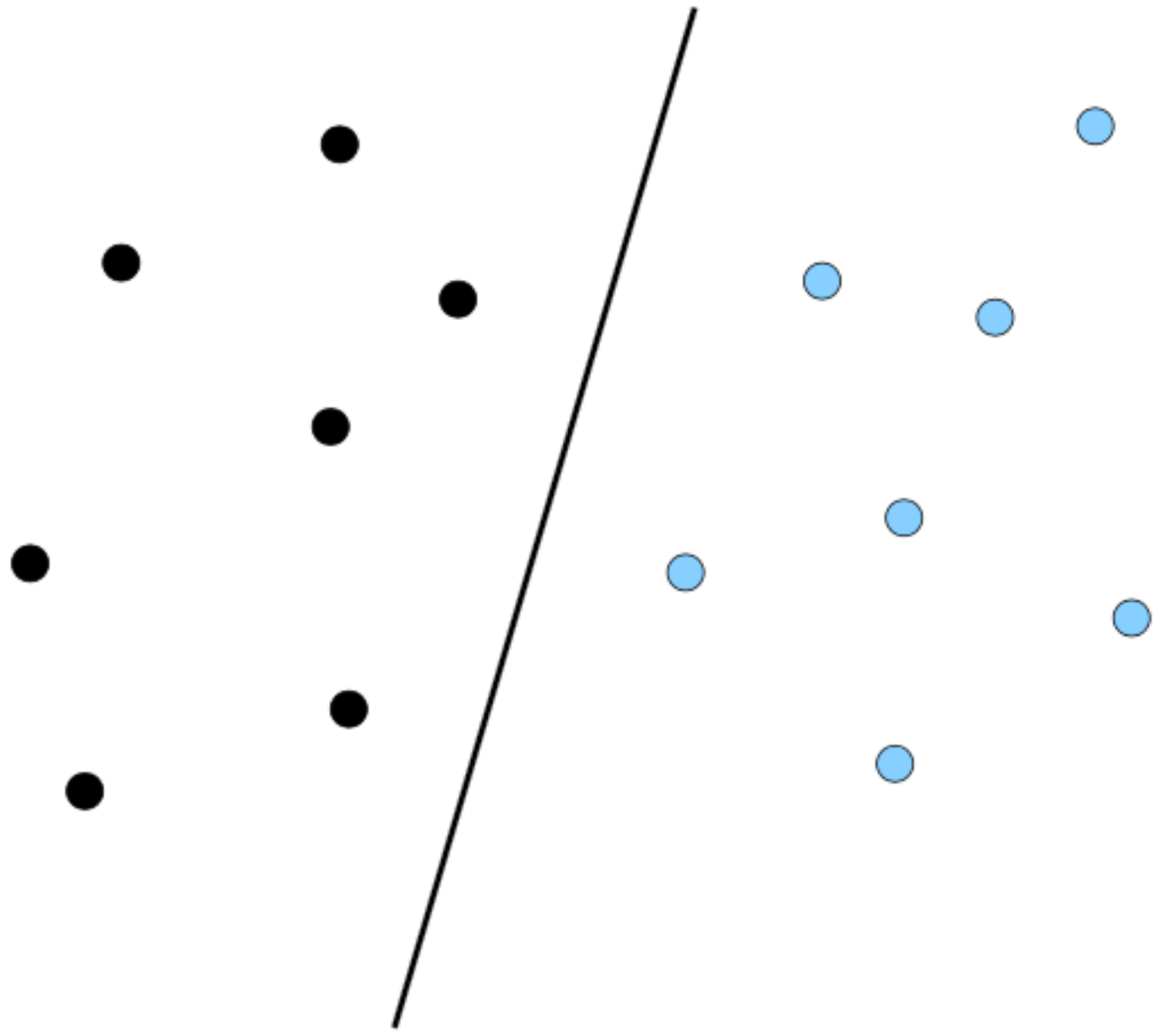
- Hyperplane



The Goal



The Goal

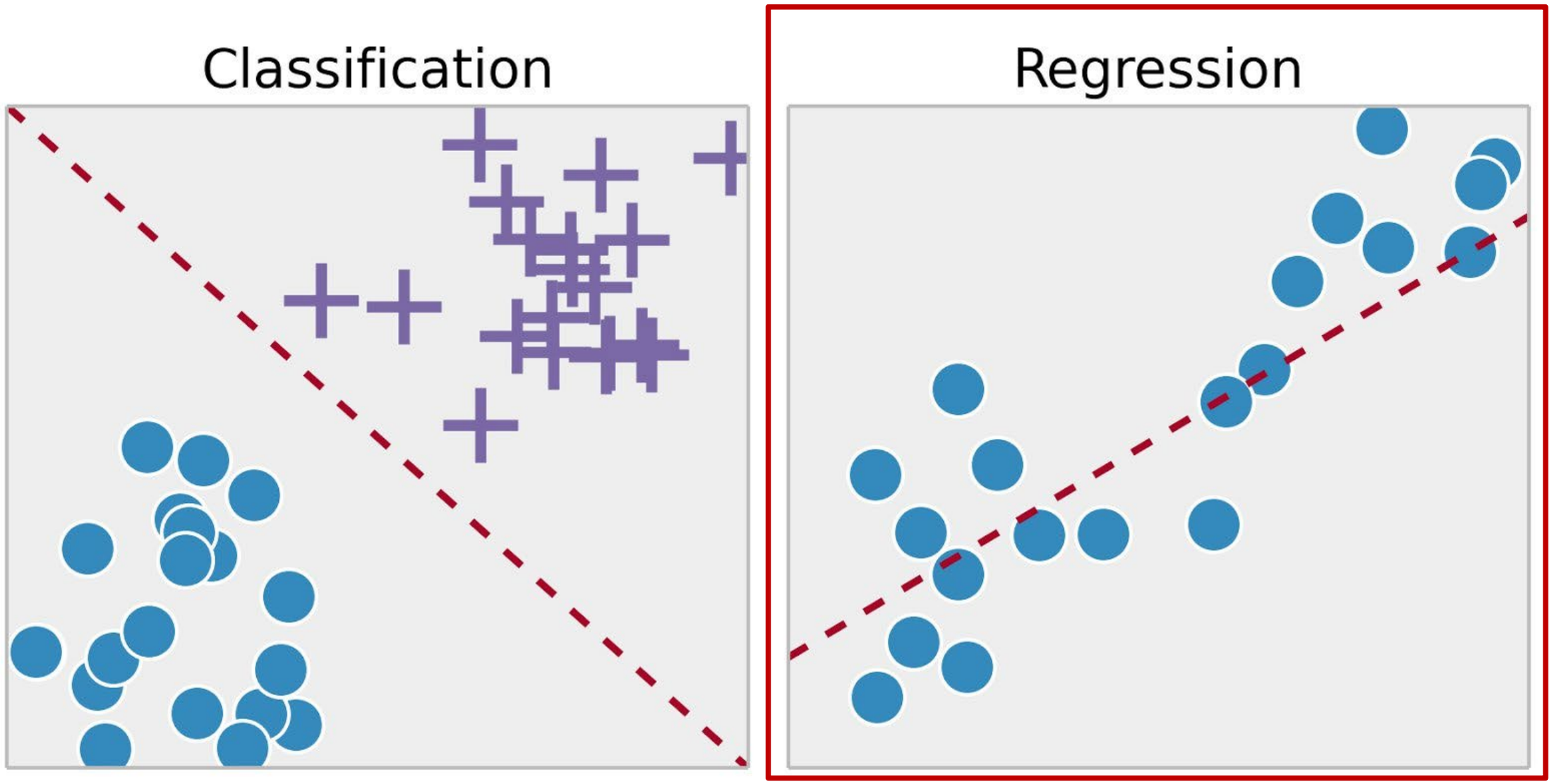


$$\vec{w}^T\vec{x} + b = 0$$

The Problem:
Many solutions, structural risk

$$\vec{w}^T\vec{x} + b = 0$$

Supervised learning



Regression

- Main goal: Predict a value
- Classification helps to distinguish between categories
- Regression helps you to predict numerical values

Housing In Boston

http://wonderfulengineering.com/mit -wallpapers -backgrounds -massachusetts -institute -of-technology/

Example: Housing in Boston

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000

Source: medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471#.4jrjf7vto

Example: Housing in Boston

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	???

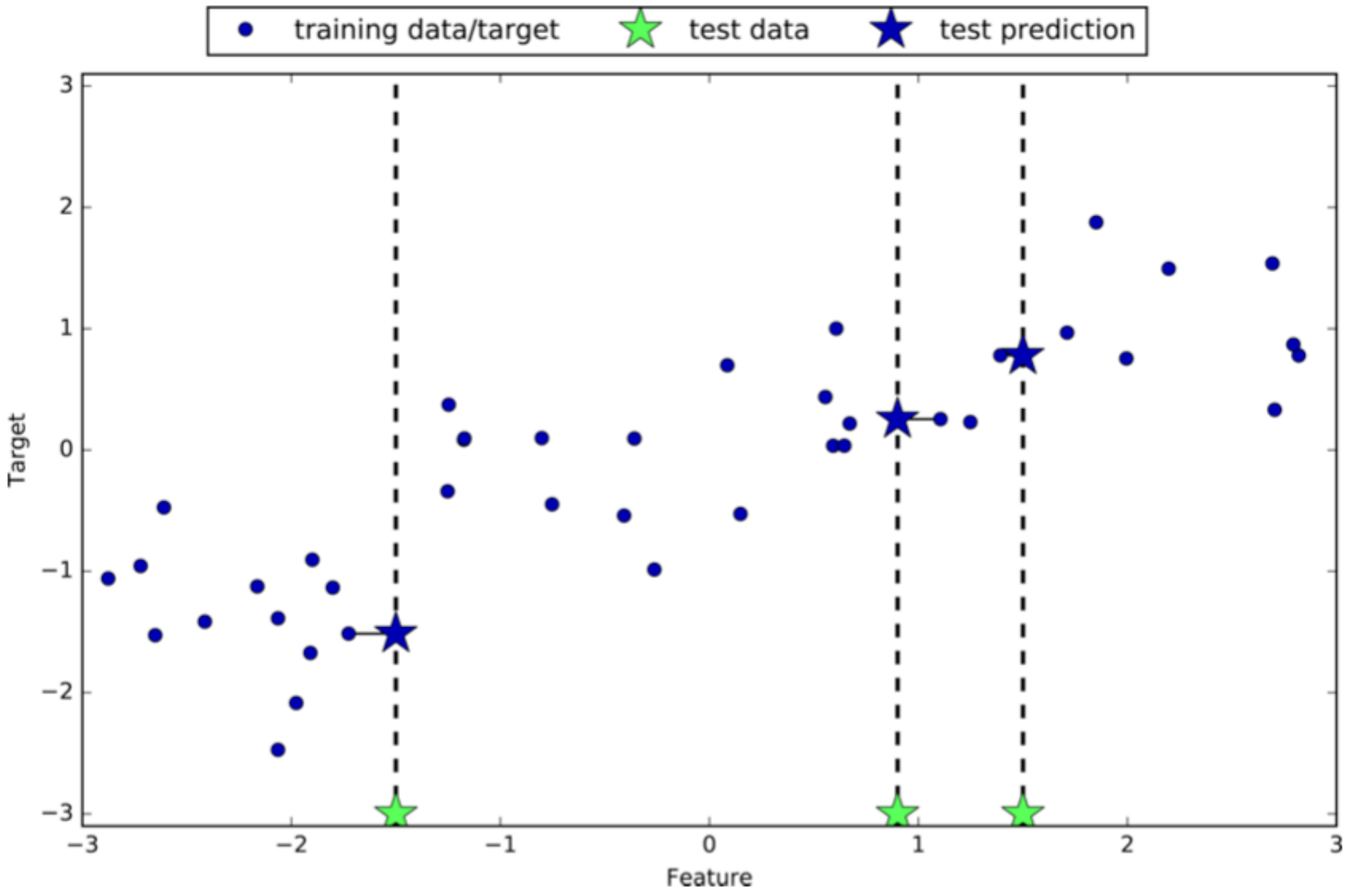
Source: medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471#.4jrjf7vto

Example: Housing in Boston

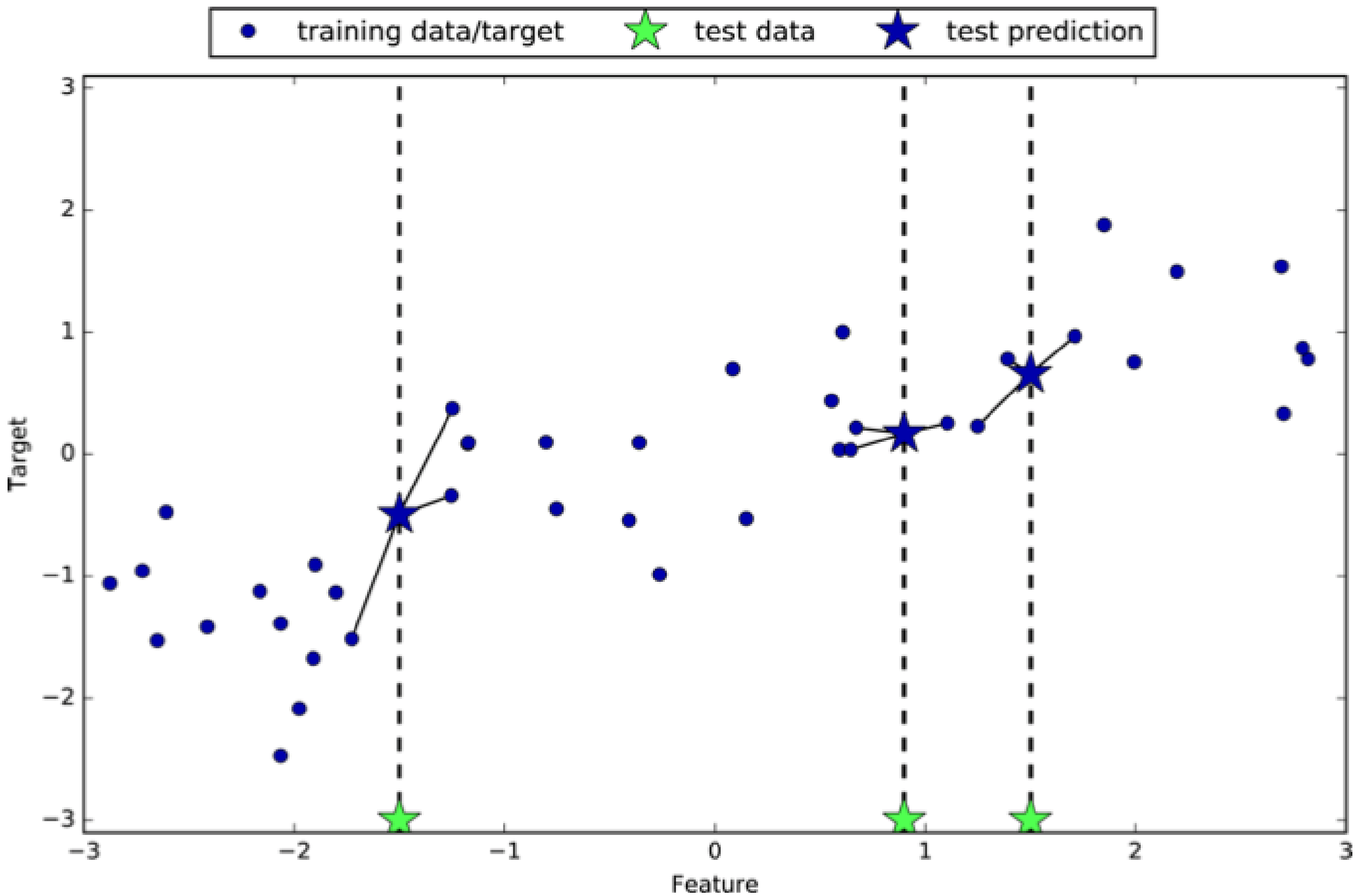
Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	$250,000
2	800	Hipsterton	$300,000
2	850	Normaltown	$150,000
1	550	Normaltown	$78,000
4	2000	Skid Row	$150,000

Source: medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471#.4jrjf7vto

1-Nearest Neighbor Regression



3-Nearest Neighbors Regression



Evaluate Your Model

```
from sklearn.metrics import mean_squared_error

print mean_squared_error(
            y_test, clf.predict( X_test ) )
```

Evaluate Your Model

```
from sklearn.metrics import mean_squared_error

print mean_squared_error(
            y_test, clf.predict( X_test ) )
```

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

https://en.wikipedia.org/wiki/Mean_squared_error

Evaluate Your Model

```
from sklearn.metrics import mean_squared_error

print mean_squared_error(
            y_test, clf.predict( X_test ) )
```

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

https://en.wikipedia.org/wiki/Mean_squared_error

Evaluate Your Model

```
from sklearn.metrics import mean_squared_error

print mean_squared_error(
            y_test, clf.predict( X_test ) )
```

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

https://en.wikipedia.org/wiki/Mean_squared_error

Evaluate Model

```
from sklearn.metrics import mean_squared_error

print mean_squared_error(
            y_test, clf.predict( X_test ) )
```

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

https://en.wikipedia.org/wiki/Mean_squared_error

Evaluate Model

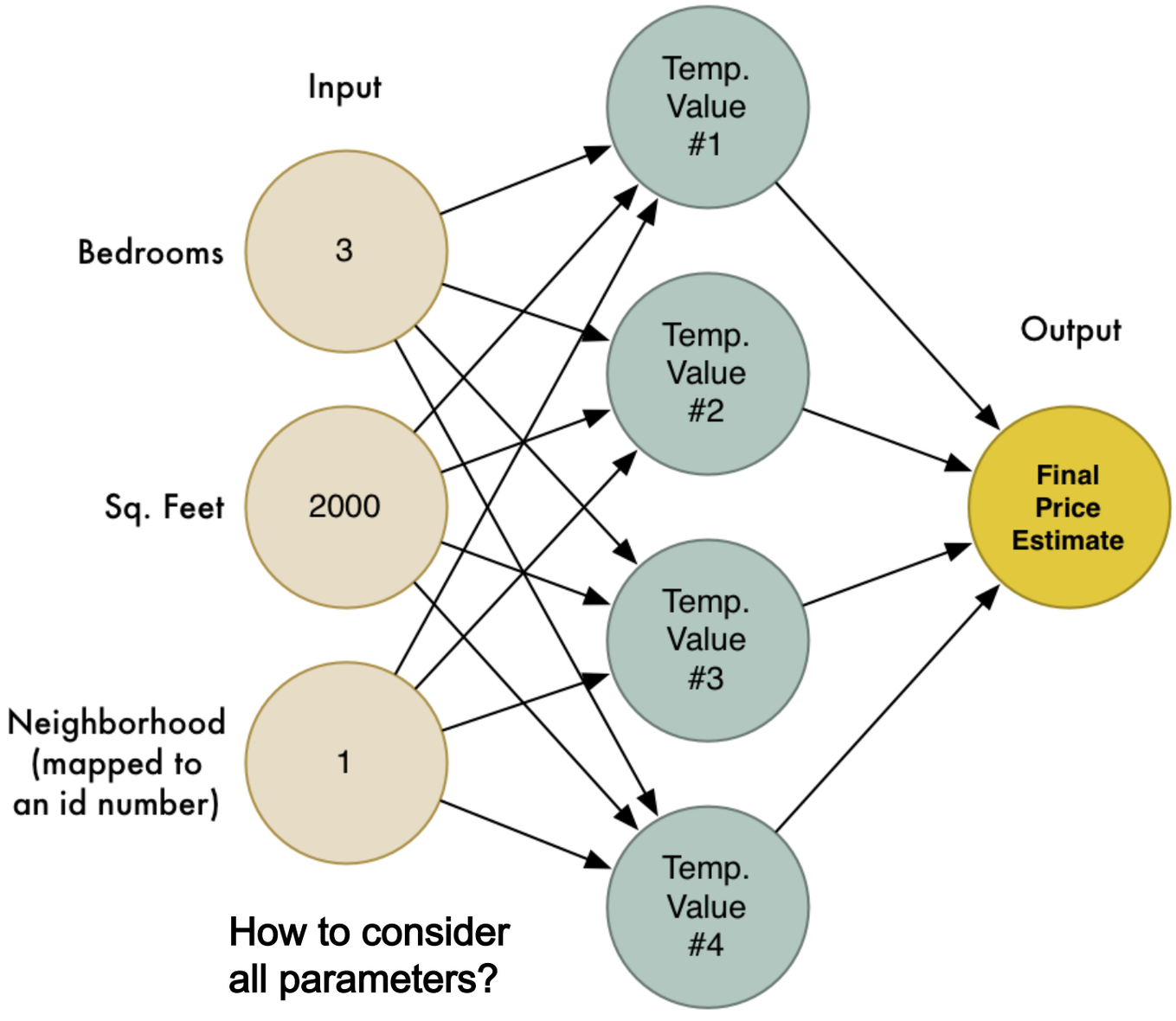
```
from sklearn.metrics import mean_squared_error

print mean_squared_error(
            y_test, clf.predict( X_test ) )
```

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

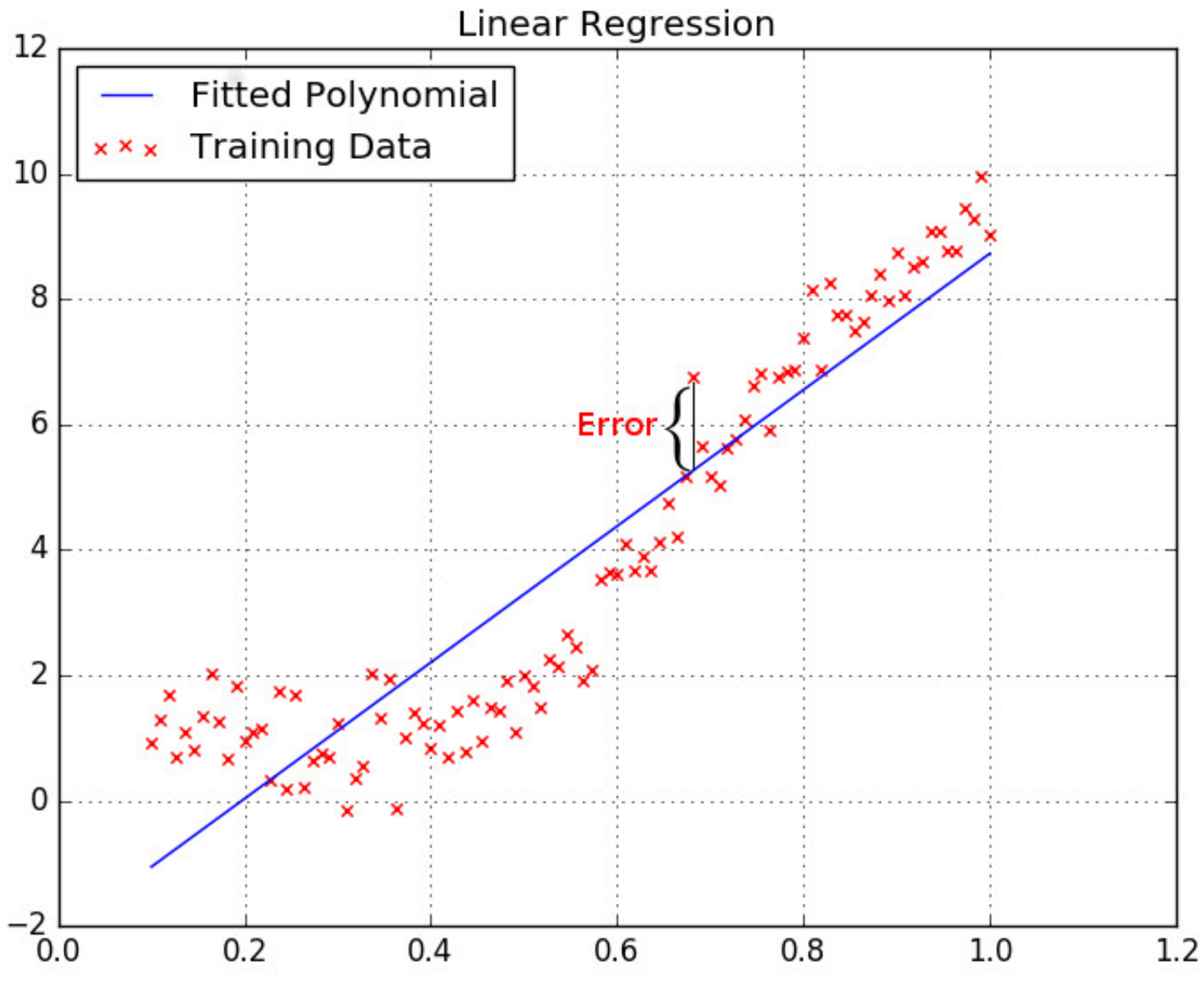
https://en.wikipedia.org/wiki/Mean_squared_error

Example: Housing in Boston



Source: medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471#.4jrjf7vto

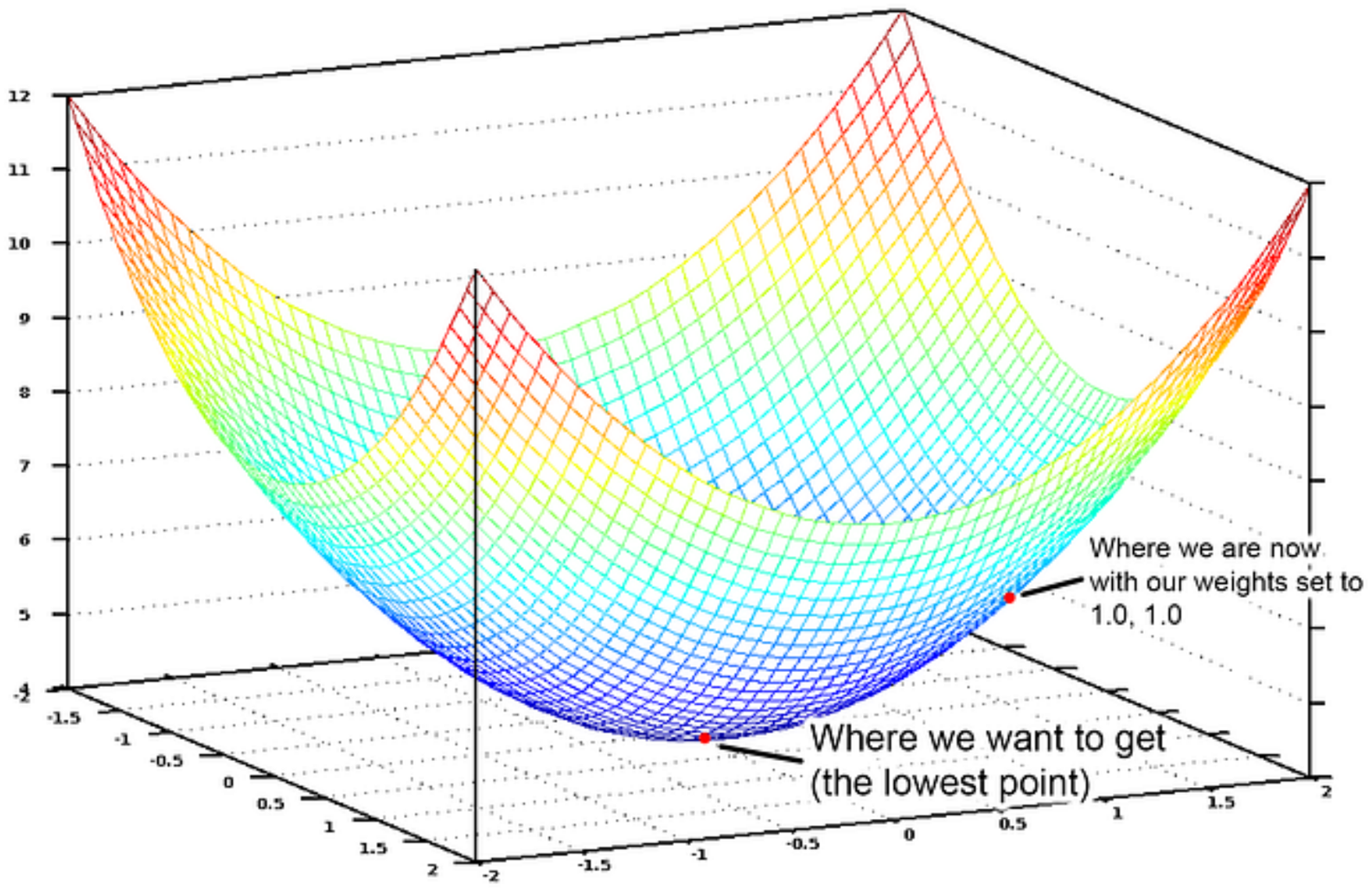
Another Method: Linear Regression



Source: https ://www.cs.toronto.edu/~frossard/post/linear_regression/

Gradient Descent

- It uses gradients (derivatives) to update the parameter values until minimizing the error



Data Science & Vis Presentation

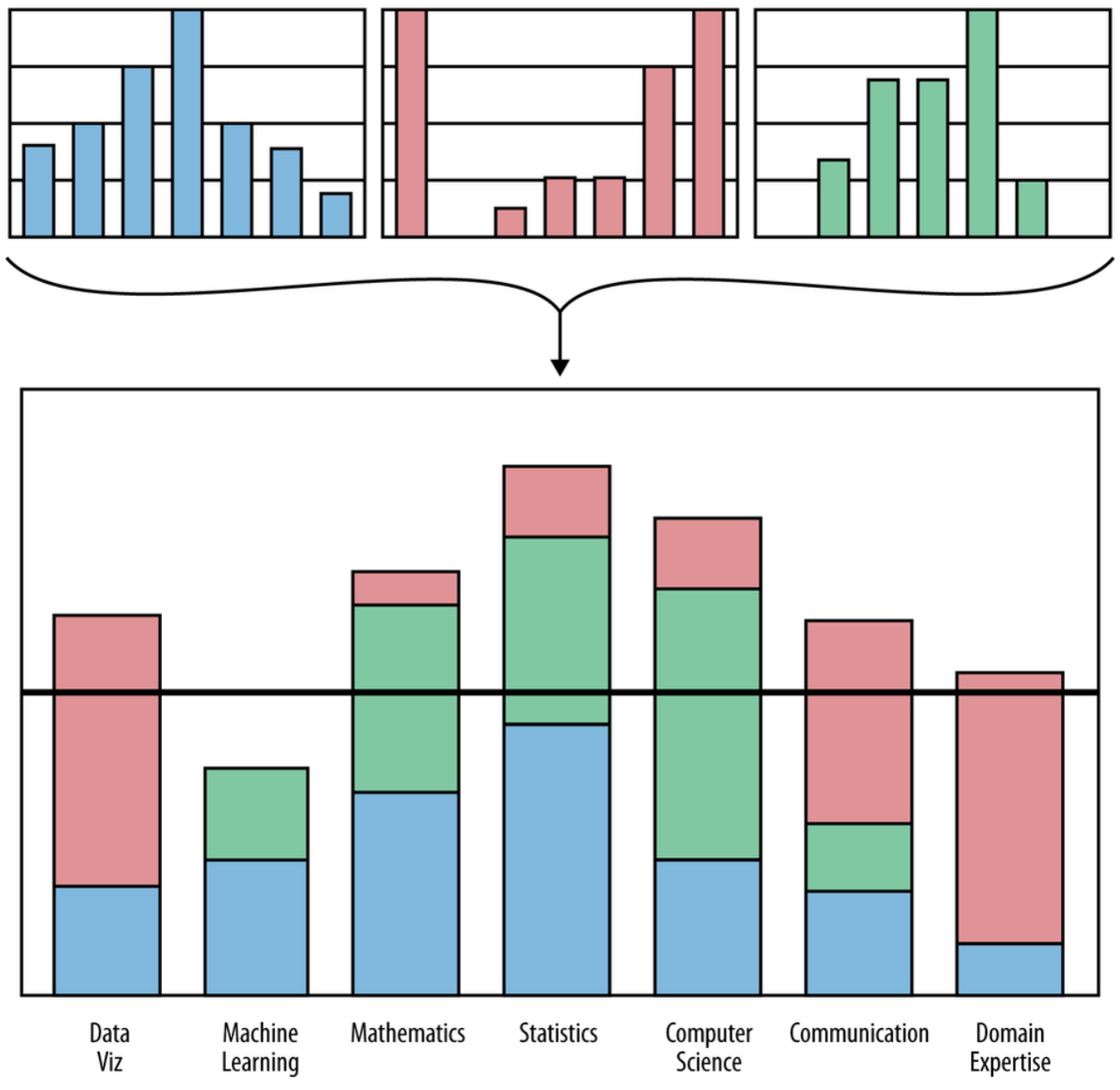
- **Deadline for topic: 22.04.2024 23:59**
- **Date of the presentation: 06.05.2024**
- Submit the topic via email to me (**molina@uni-bremen.de**)
- If you have another paper in mind that you would like to present, please let us know – if we approve it, you can present it

Data Science & Vis Presentation

- **Every group member needs to present.**
- If you can’t make it in person, talk to your team and make it possible in hybrid format.
- I will send you the paper list today.

Groups for Presentation (20% of your grade)

- Who is already part of a group?
- Who not yet?



Next week: Unsupervised Learning

- I will send you the materials on Monday
  - Lecture: Video
  - Tutorial: Jupyter notebook
- Focus on
  - Clustering
  - Dimensionality reduction

Next week: Unsupervised Learning

- If you do not have a group by then, I strongly encourage you to come and organize one
- If you already have a group, you can still come and coordinate your group work