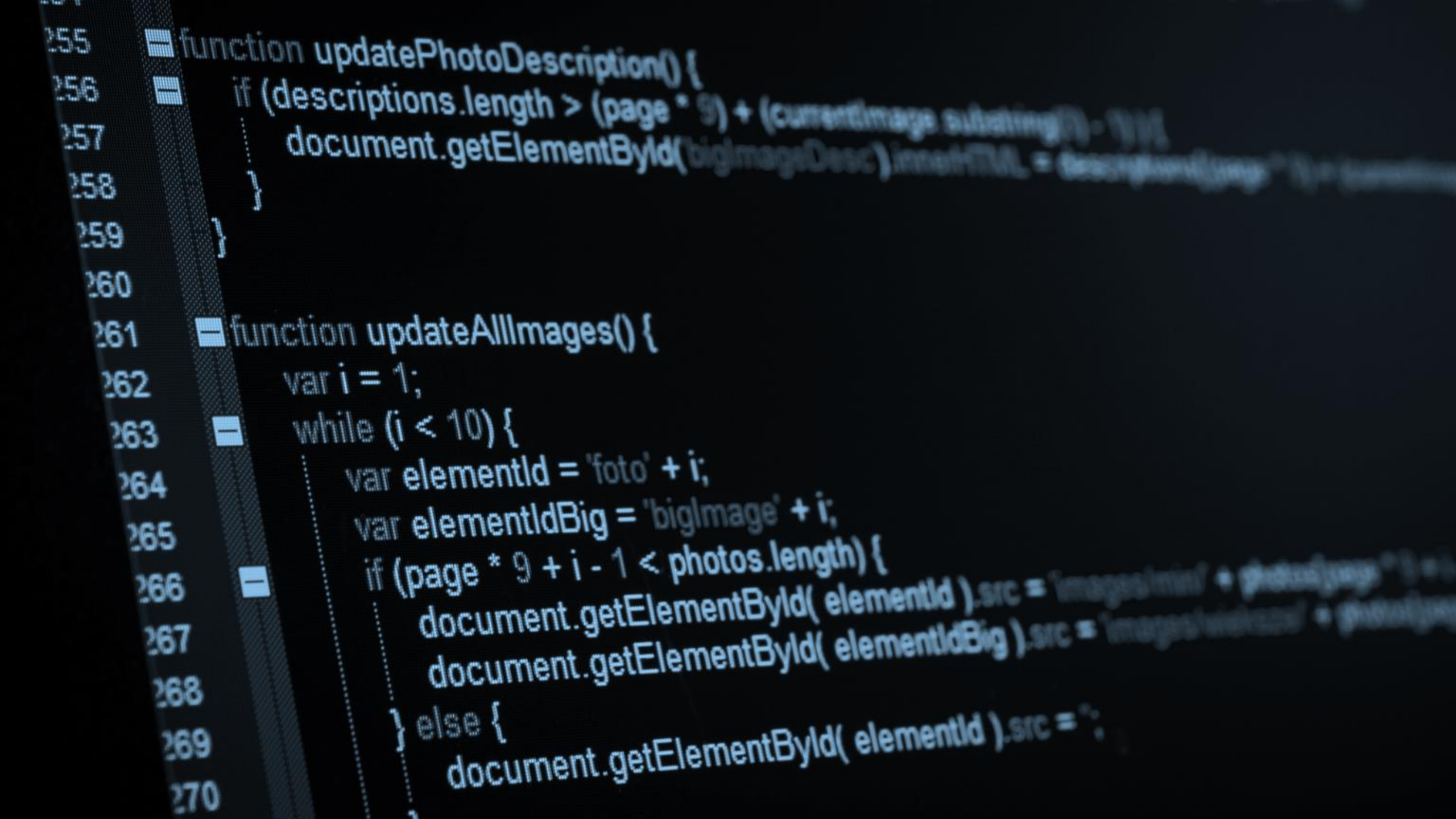
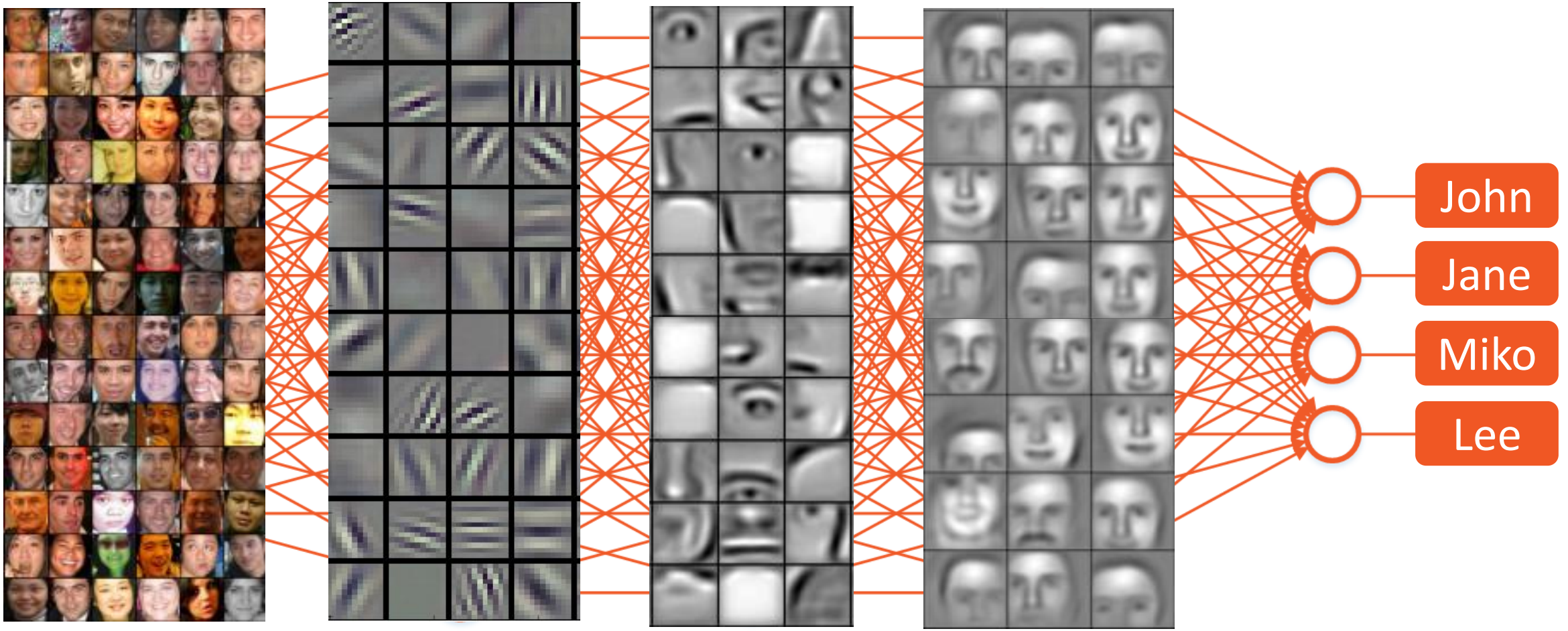
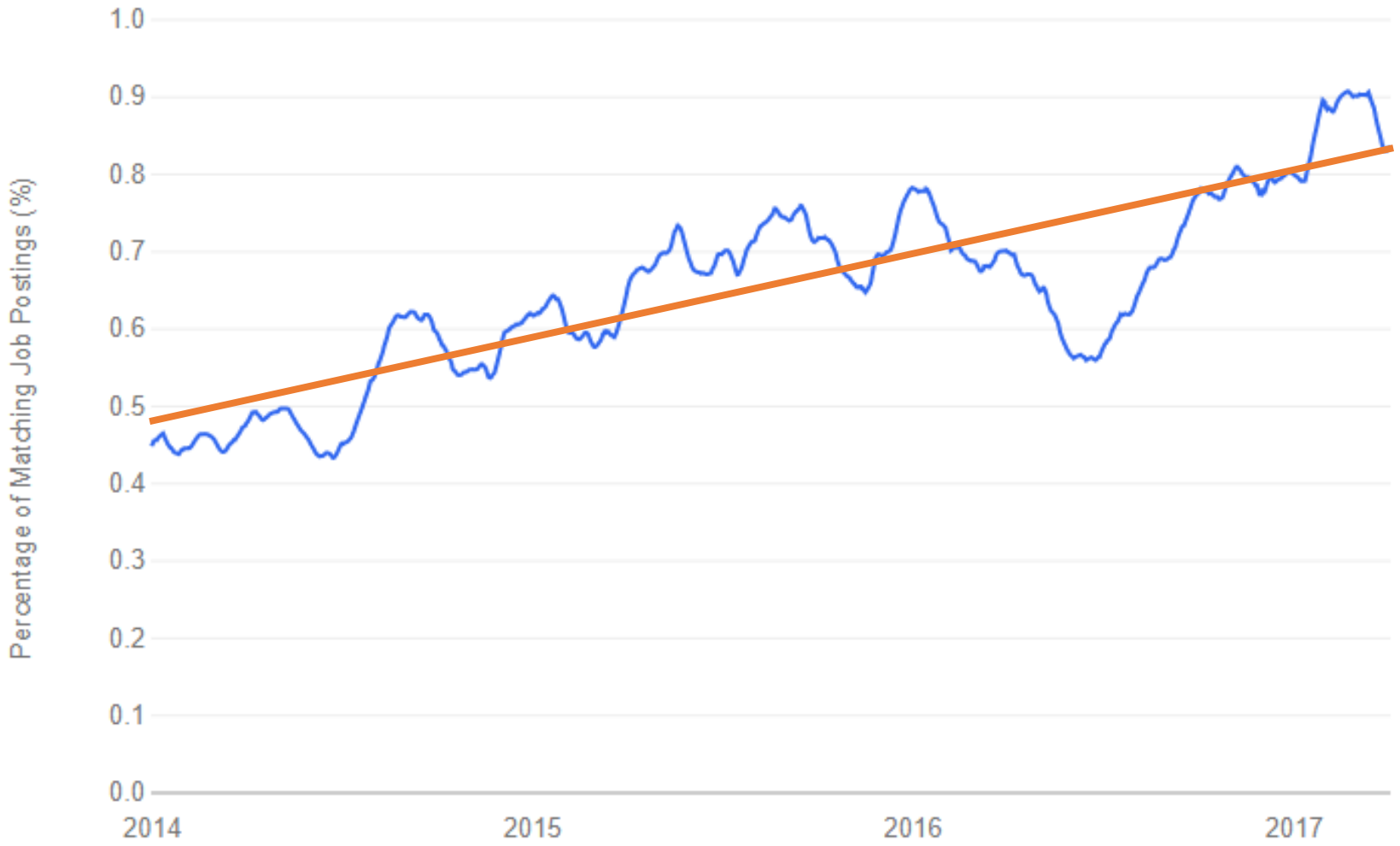
Practical Machine Learning with R

@MatthewRenze


255 function updatePhotoDescription() {
256 if (descriptions.length > (page * 9) +
257 document.getElementById(
258 }
259 }
260
261 function updateAllImages() {
262 var i = 1;
263 while (i < 10) {
264 var elementId = 'foto' + i;
265 var elementIdBig = 'bigImage' + i;
266 if (page * 9 + i - 1 < photos.length) {
267 document.getElementById(elementId).
268 document.getElementById(elementIdBig).
269 } else {
270 document.getElementById(elementId).src = '';

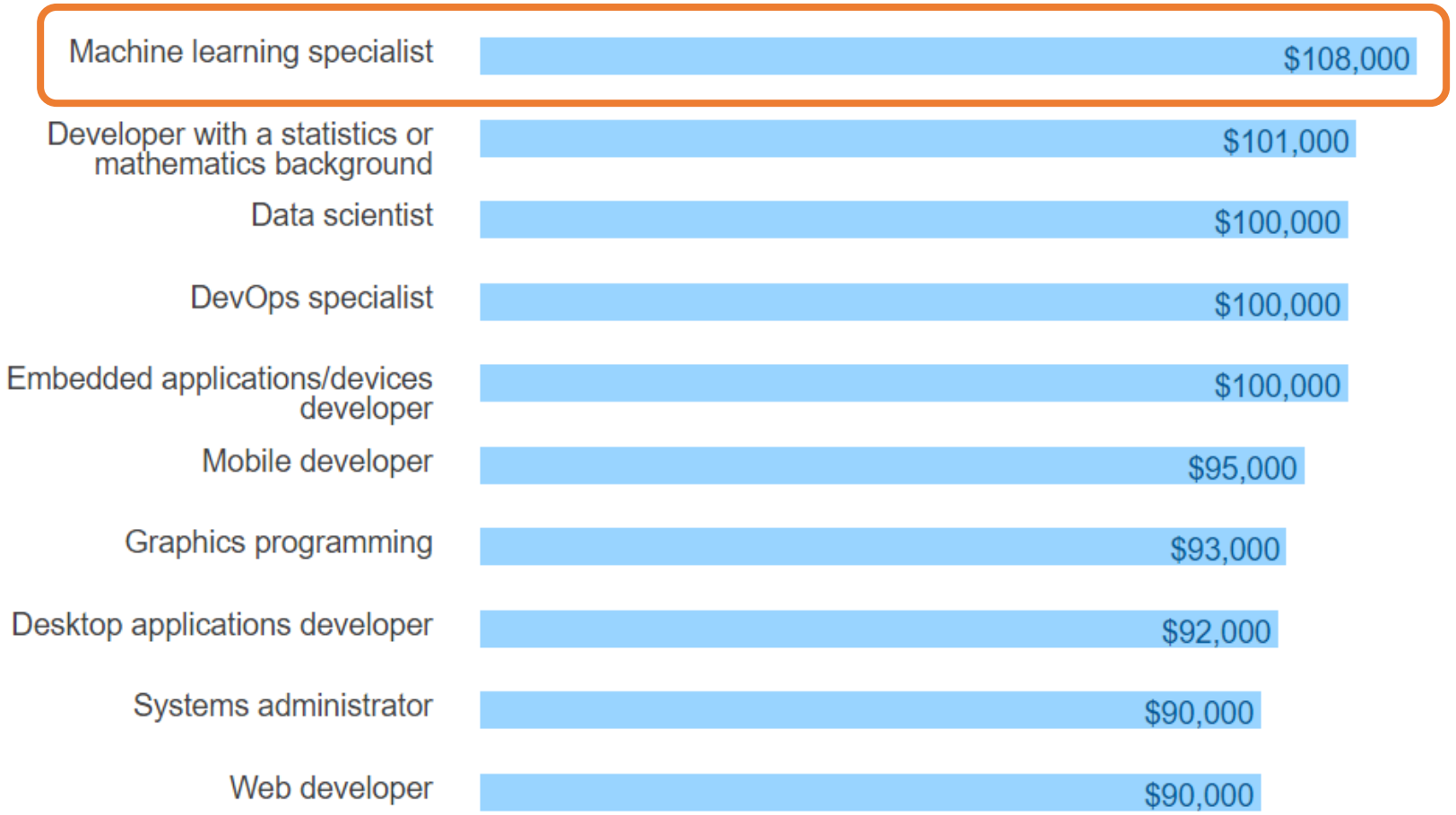

John
Jane
Miko
Lee

Job Postings for Machine Learning



Source: Indeed.com

Average Salary by Job Type (USA)



Source: Stack Overflow 2017

Overview

1. Introduction to ML
2. Introduction to R
3. Classification
4. Regression
5. Clustering
6. ML in Practice

About Me

Data Science Consultant

Education

- B.S. in Computer Science
- B.A. in Philosophy

Community

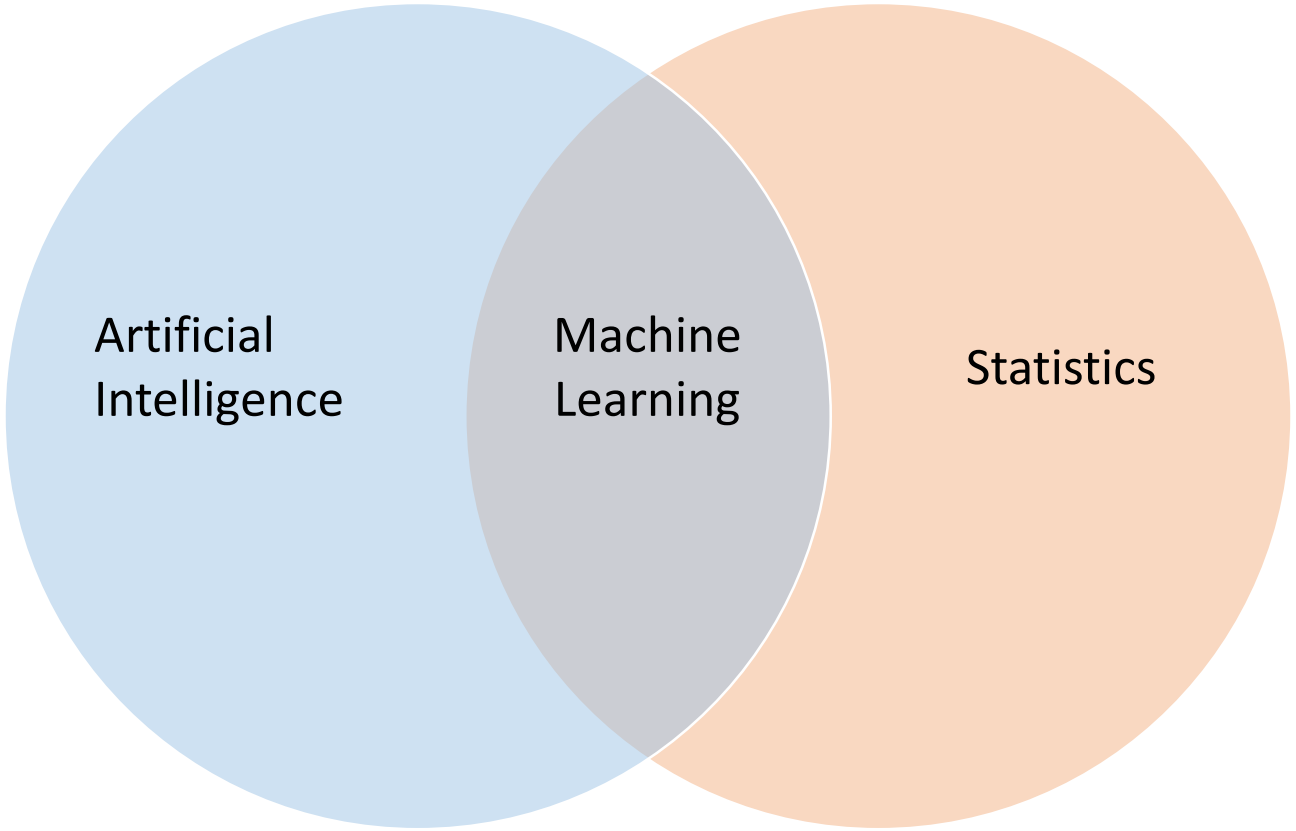
- Public Speaker
- Pluralsight Author
- Microsoft MVP
- ASPInsider
- Open-source Software

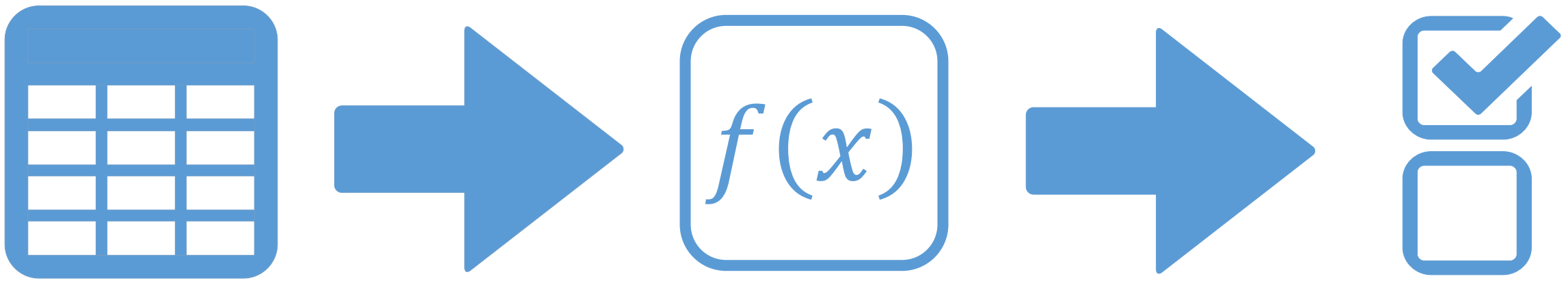
How Does This Apply to Me?

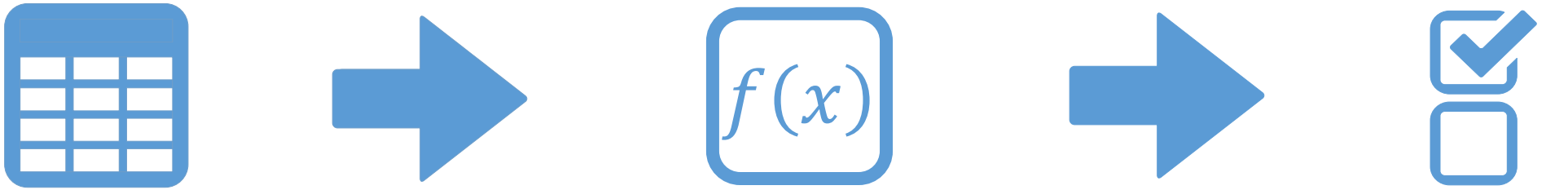
- ☑ Make decisions using data
- ☑ Make predictions using data
- ☑ Make recommendations using data
- ☑ Write code that does all these things

Introduction to Machine Learning

What is Machine Learning?


Artificial
Intelligence
Machine
Learning
Statistics

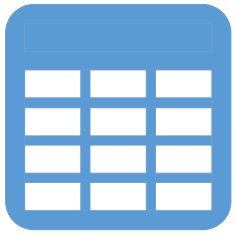

$f(x)$



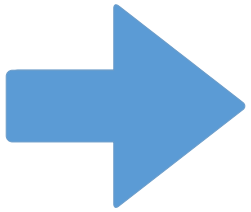
Data

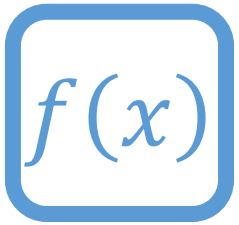
Function

Prediction

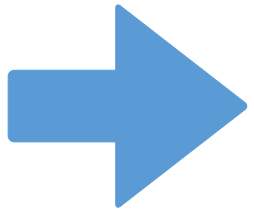


Data





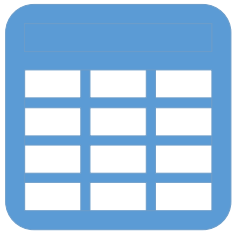
Function



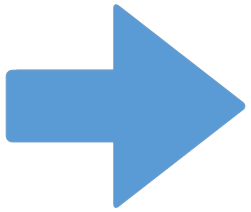
Prediction

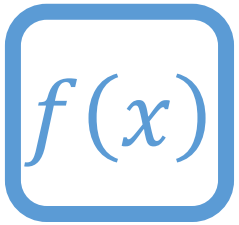
Cat

Dog

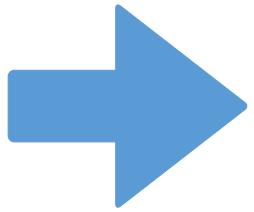


Data





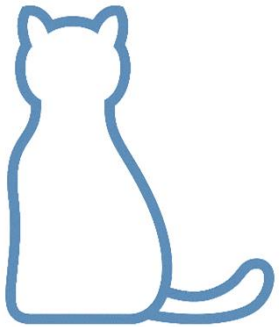
Function



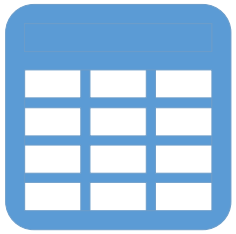
Prediction

Cat

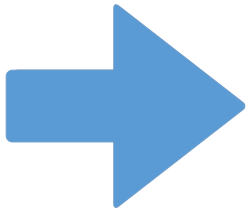
Dog

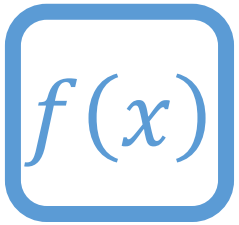


Is cat?

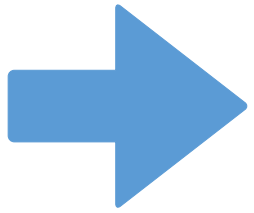


Data





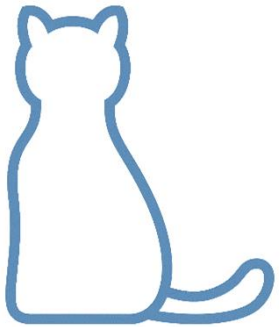
Function



Prediction

Cat

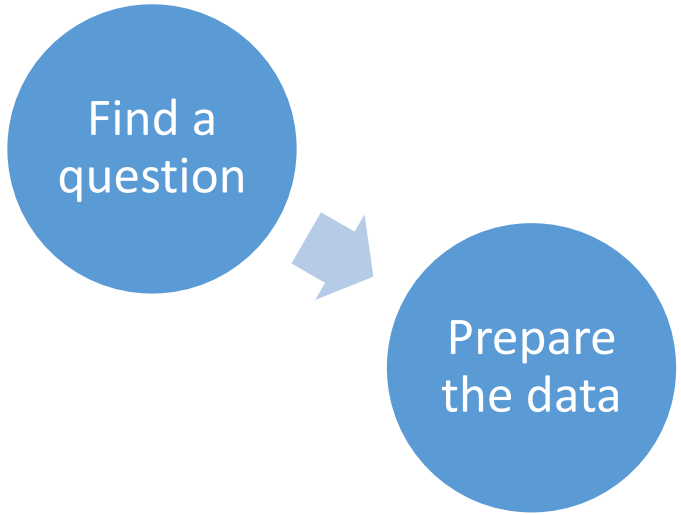
Dog

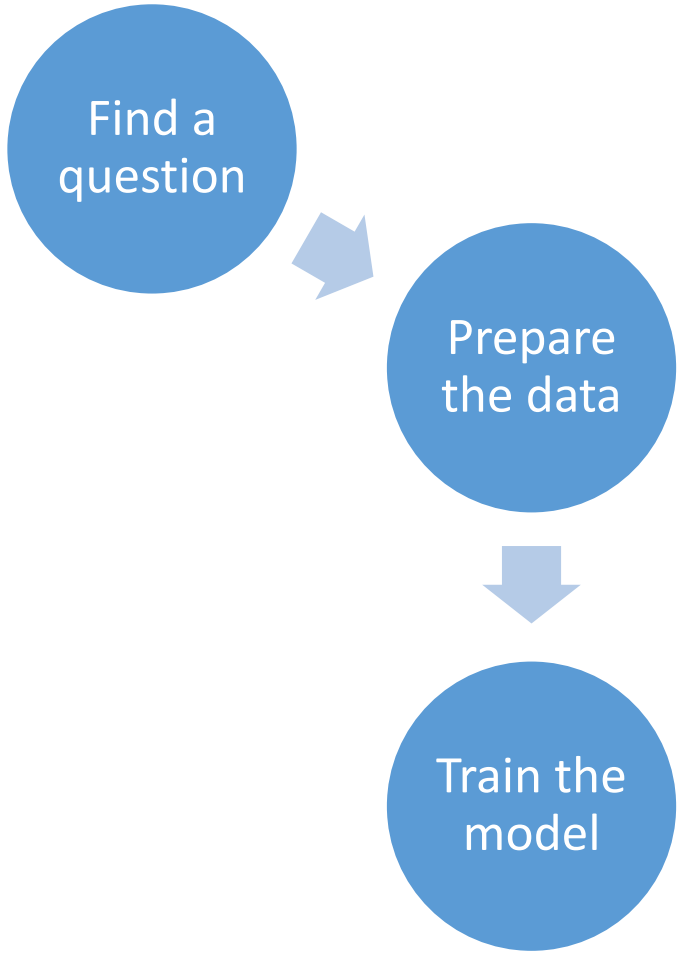


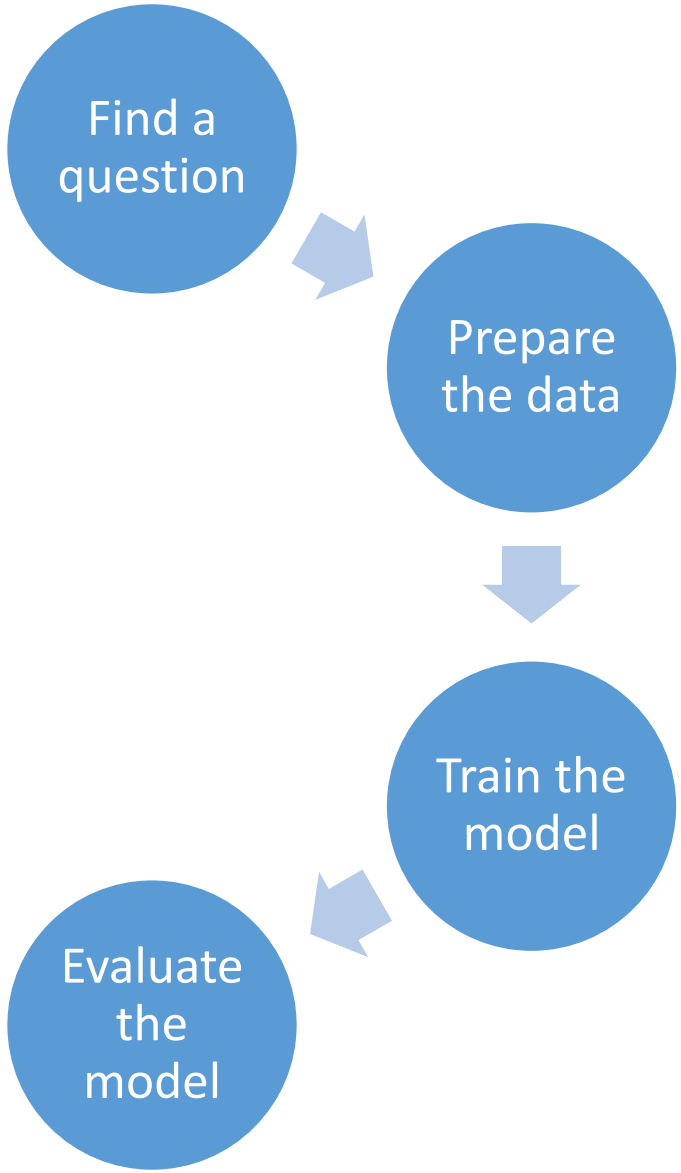
Is cat?

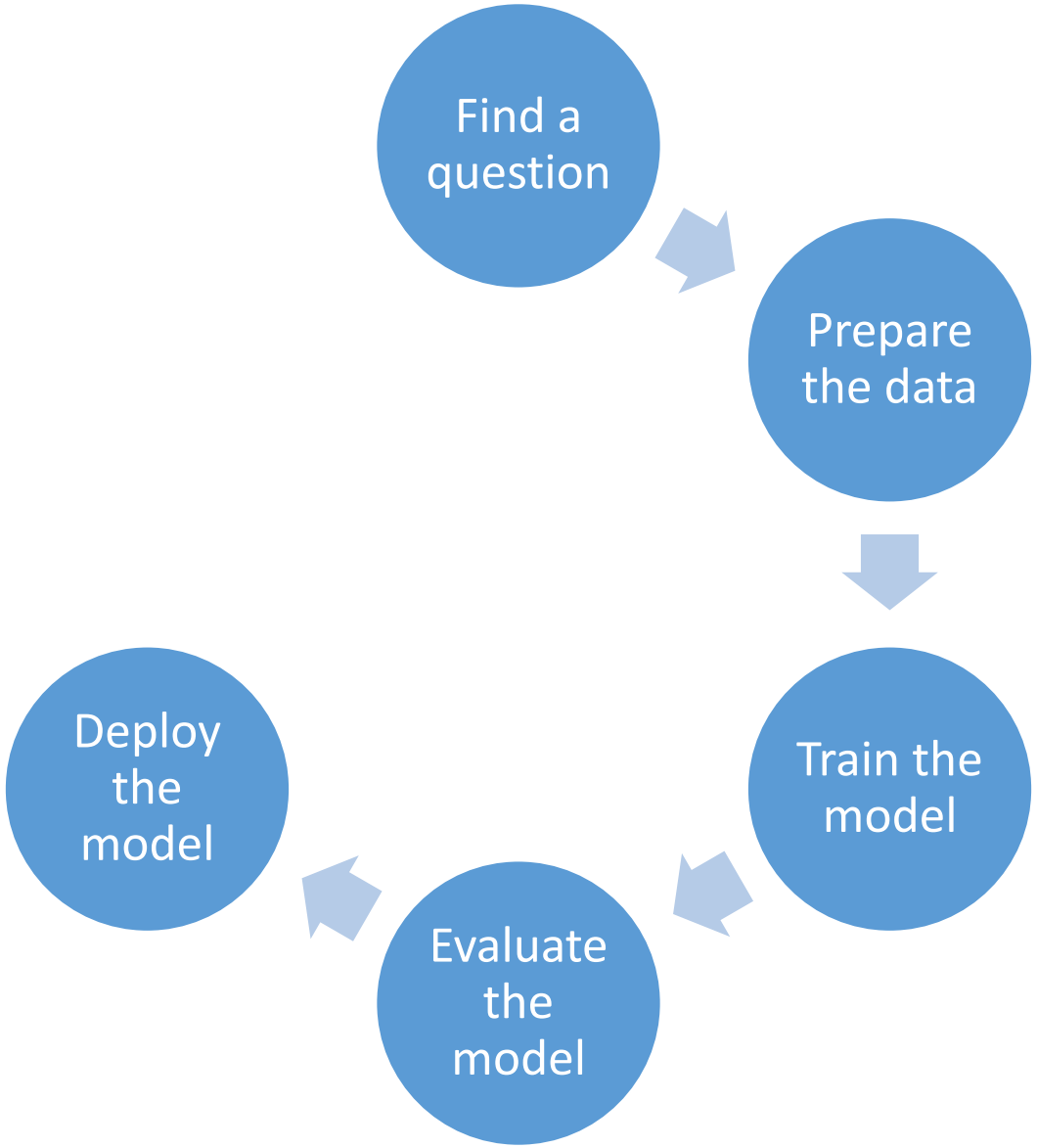
Yes

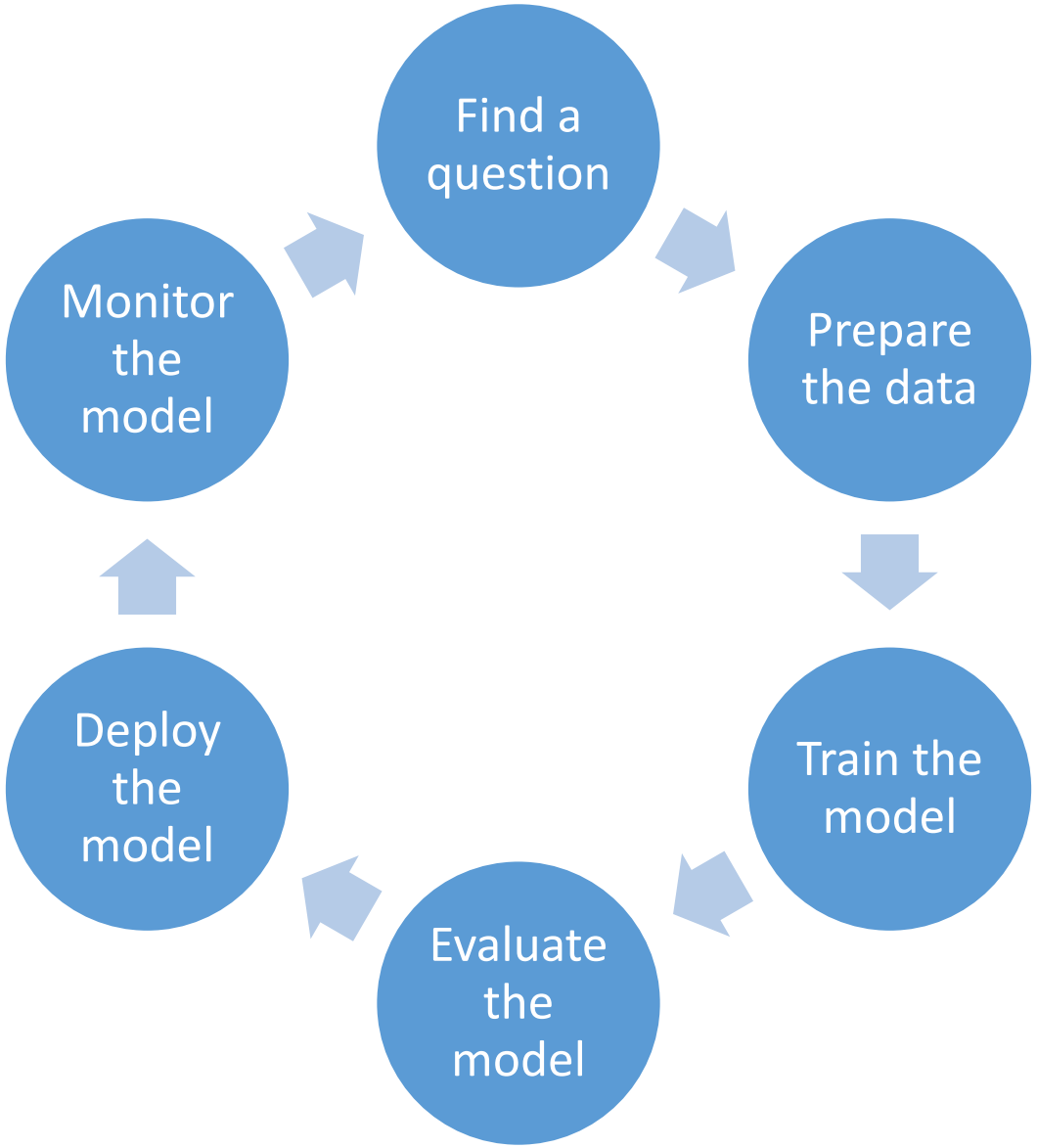
Find a question

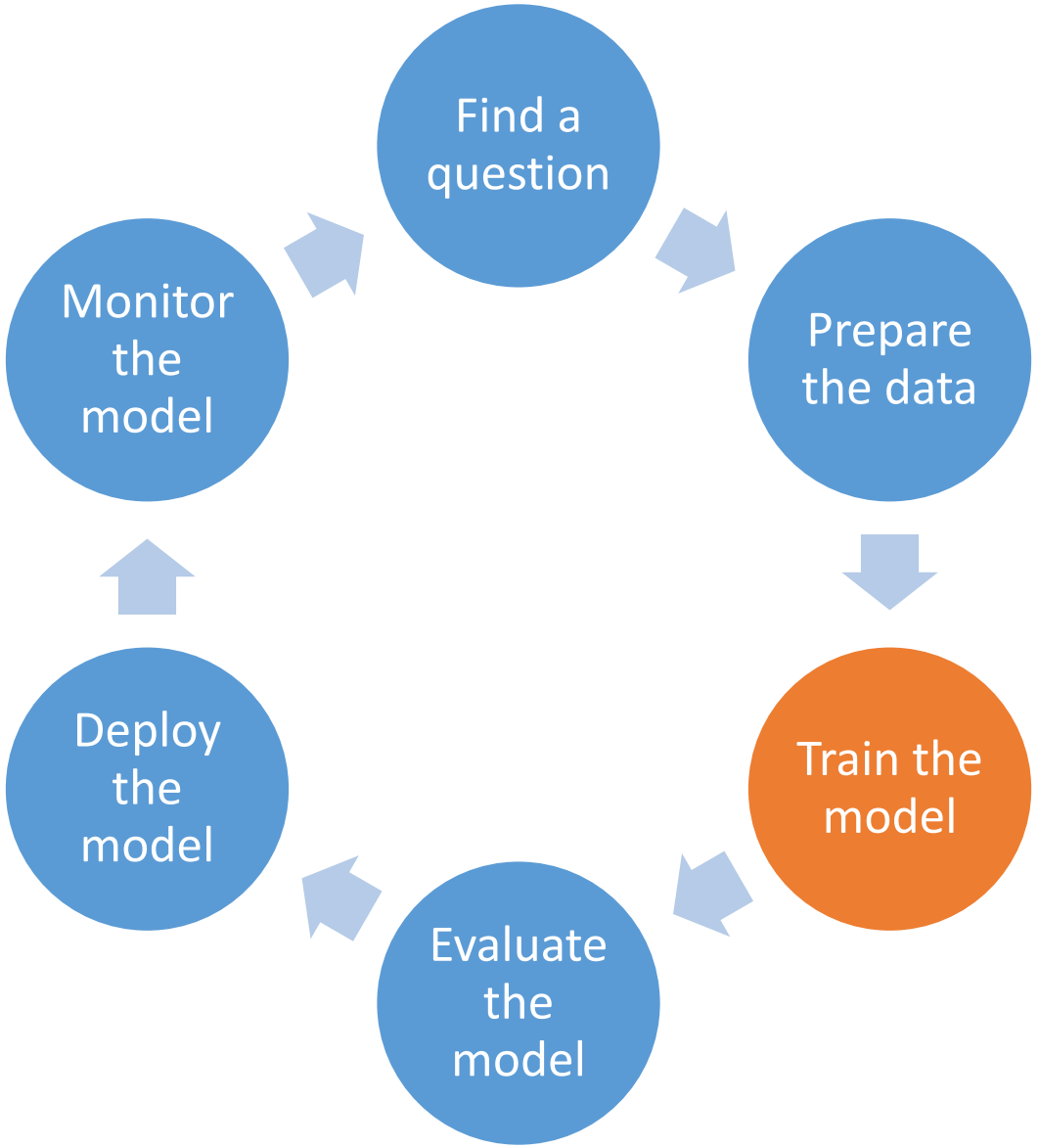

Find a question
Prepare the data

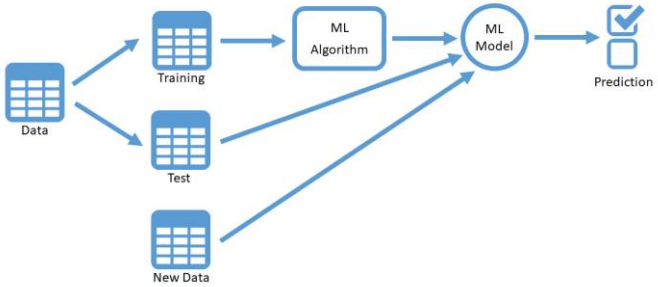

Find a question
Prepare the data
Train the model

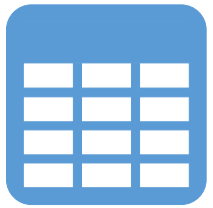

Find a question
Prepare the data
Train the model
Evaluate the model


Find a question
Prepare the data
Train the model
Evaluate the model
Deploy the model

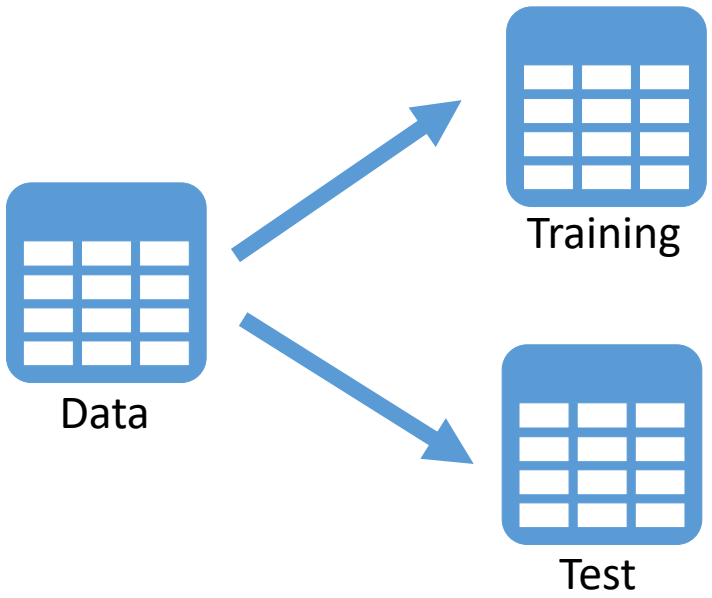

Find a question
Prepare the data
Train the model
Evaluate the model
Deploy the model
Monitor the model

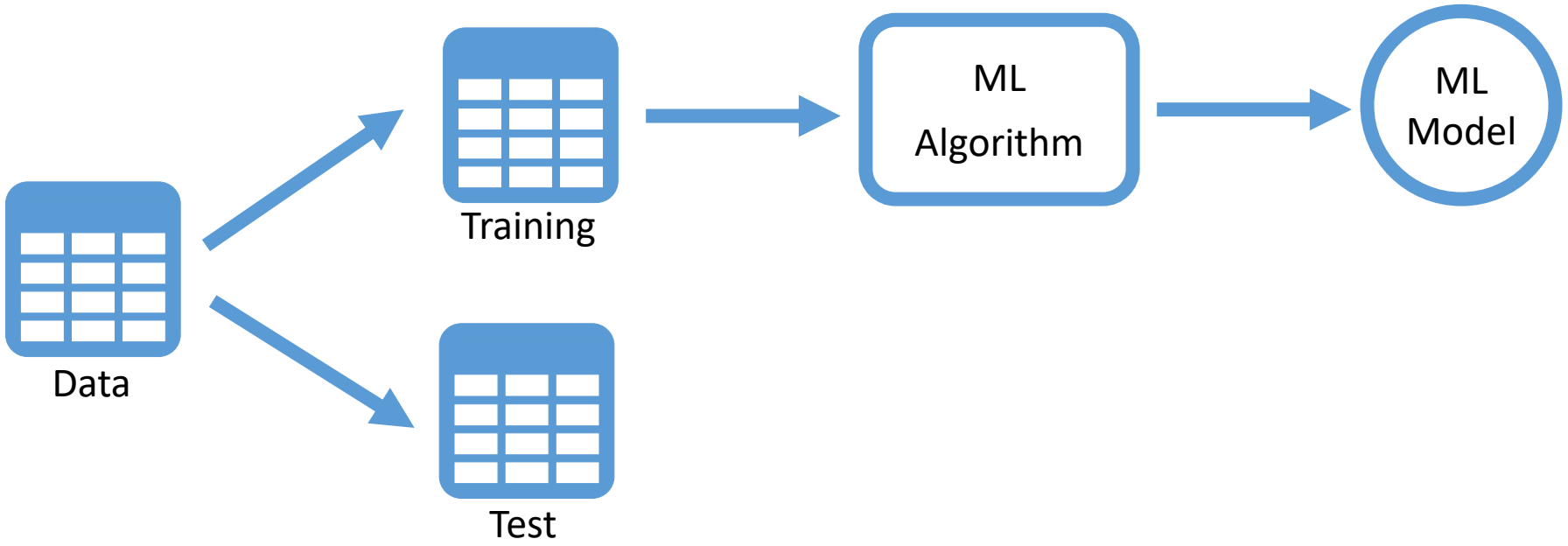

Find a question
Prepare the data
Train the model
Evaluate the model
Deploy the model
Monitor the model

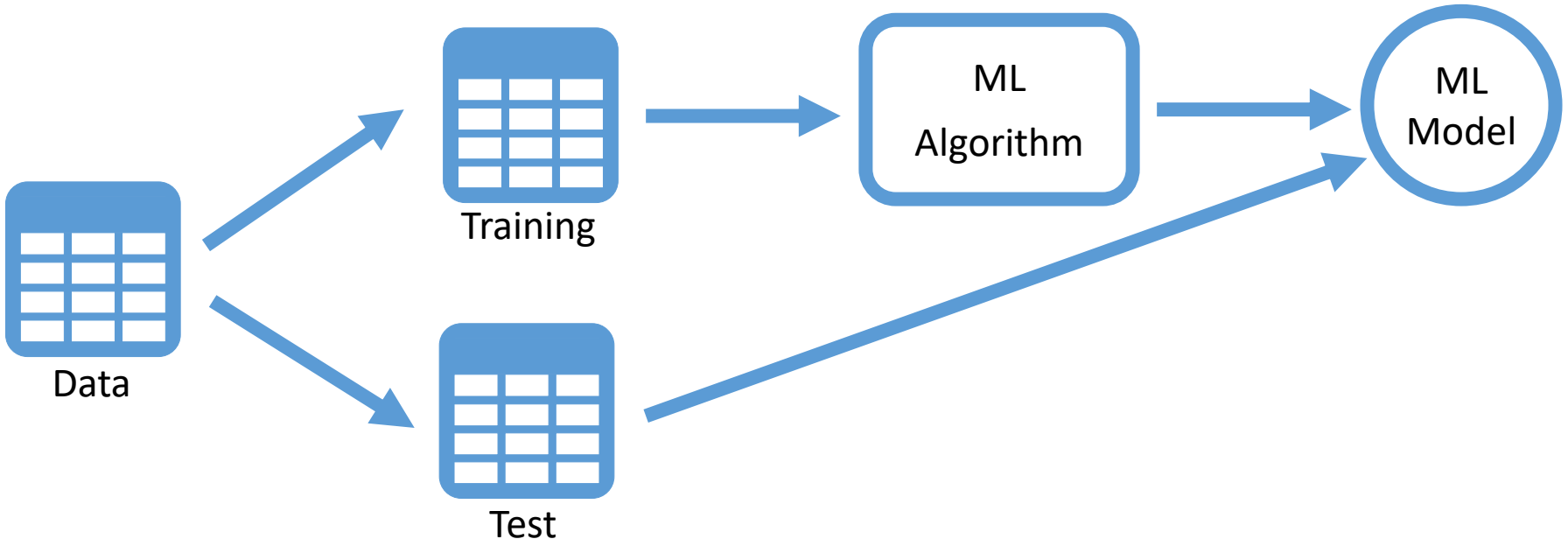

Data
Training
Test
New Data
ML Algorithm
ML Model
Prediction

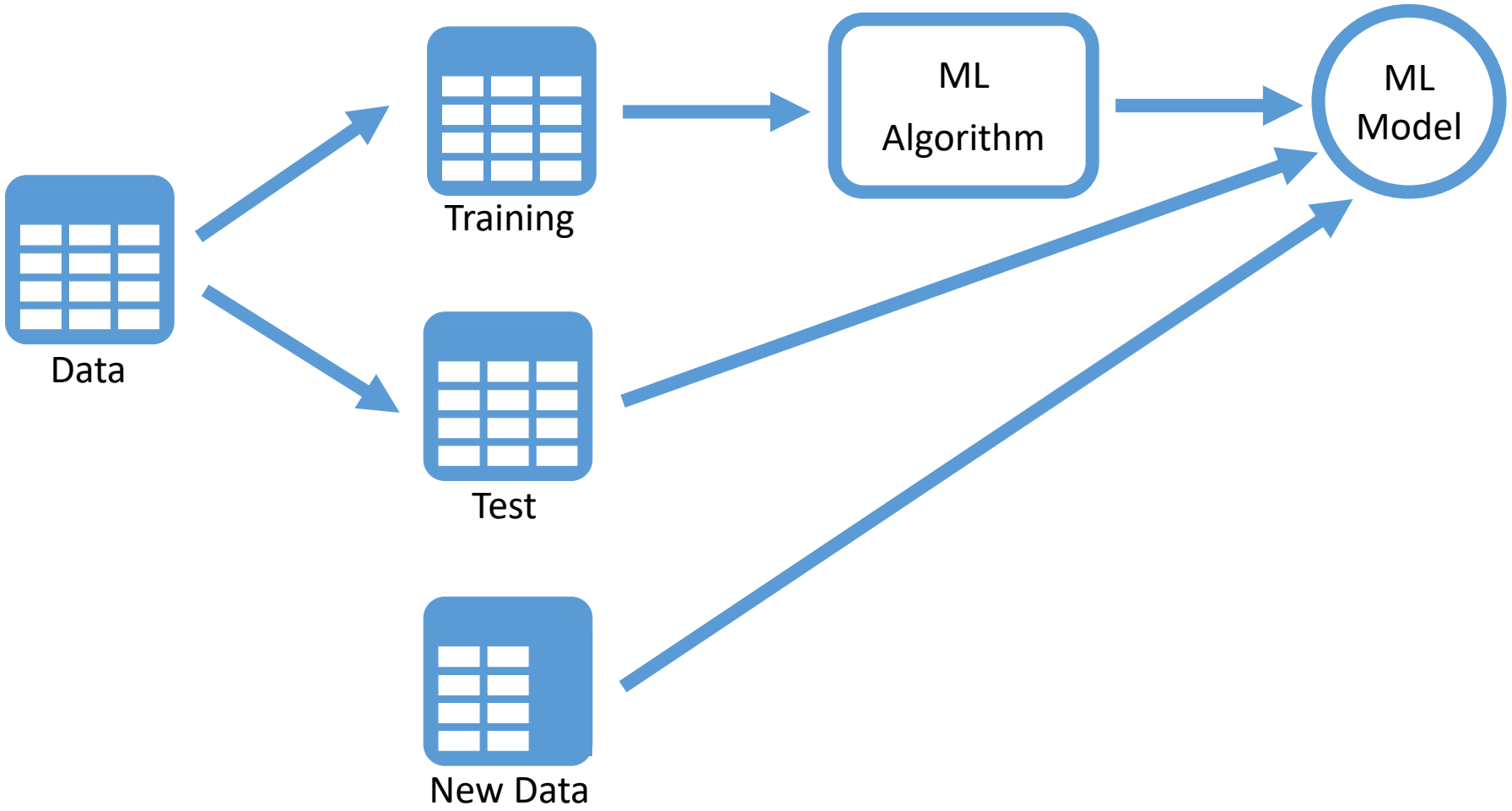


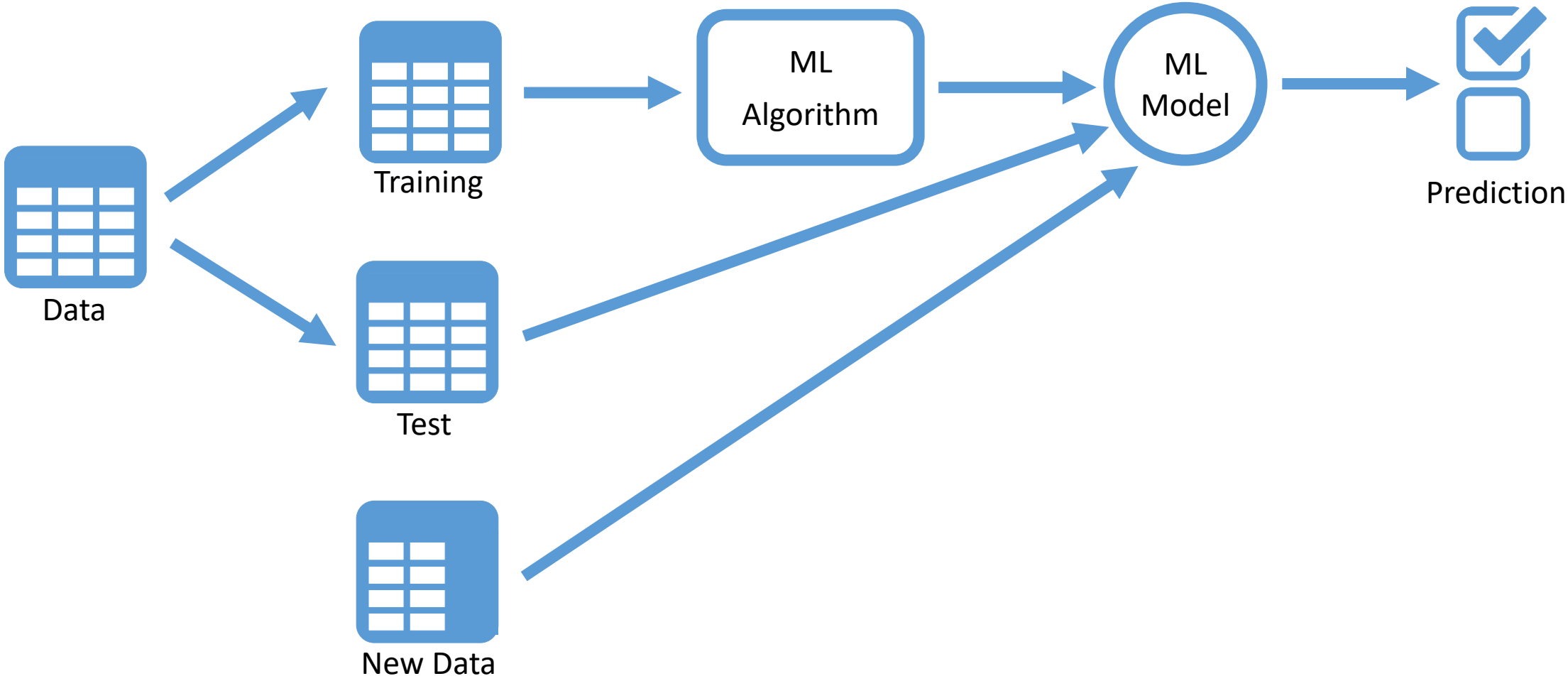
Data


Training
Data
Test

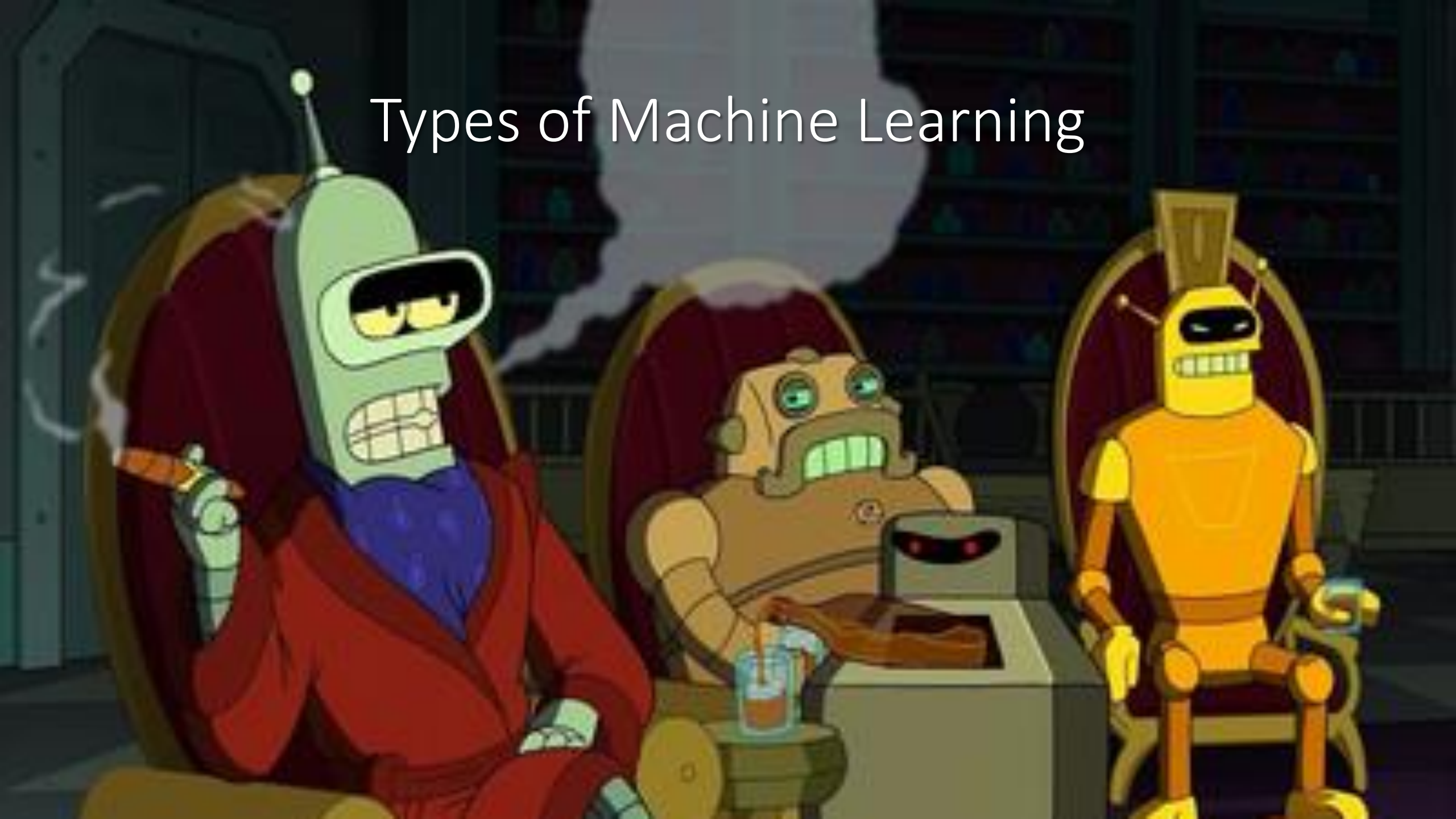

Data
Training
Test
ML
Algorithm
ML
Model


Data
Training
Test
ML
Algorithm
ML
Model

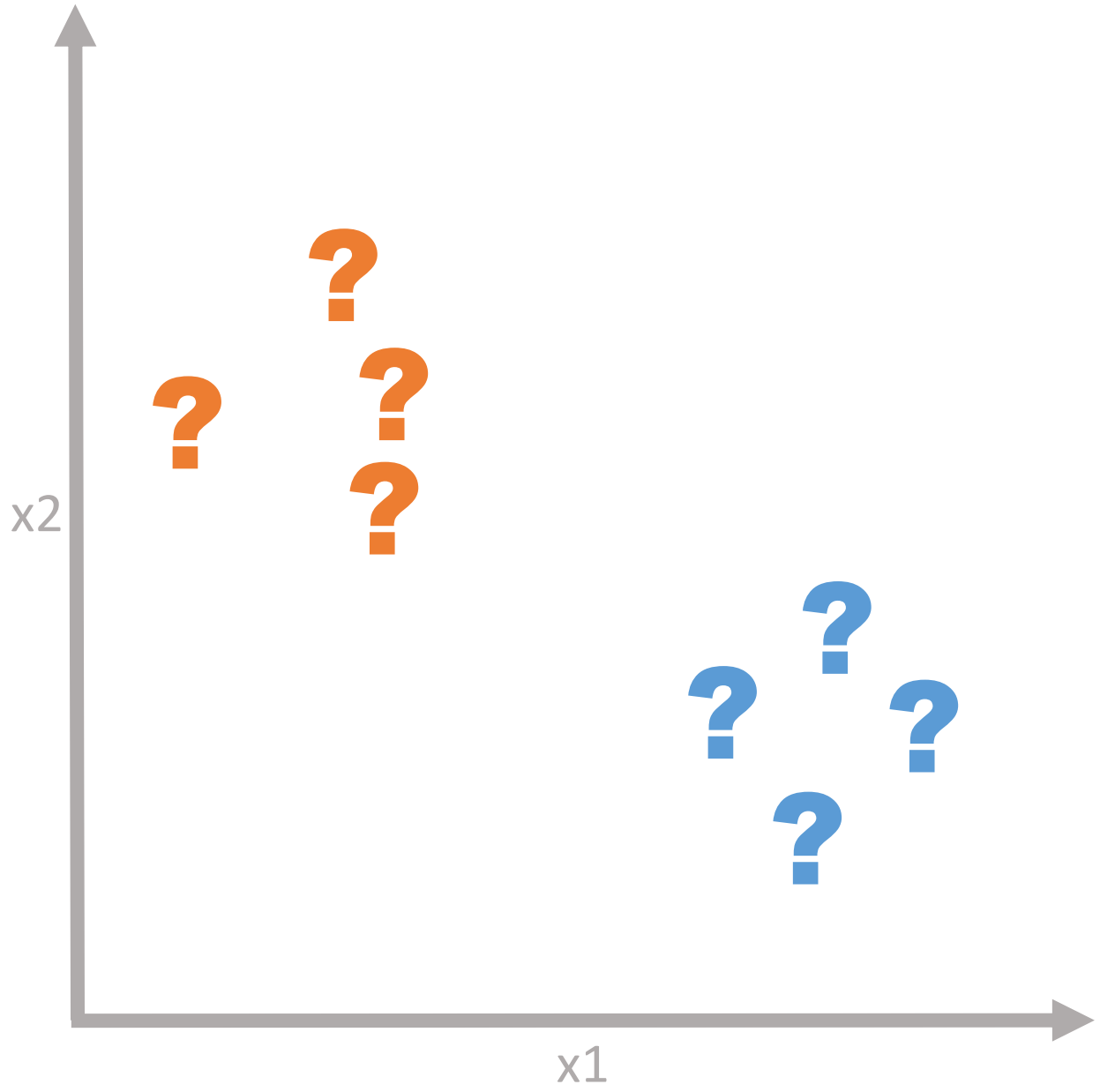

Data
Training
Test
New Data
ML
Algorithm
ML
Model

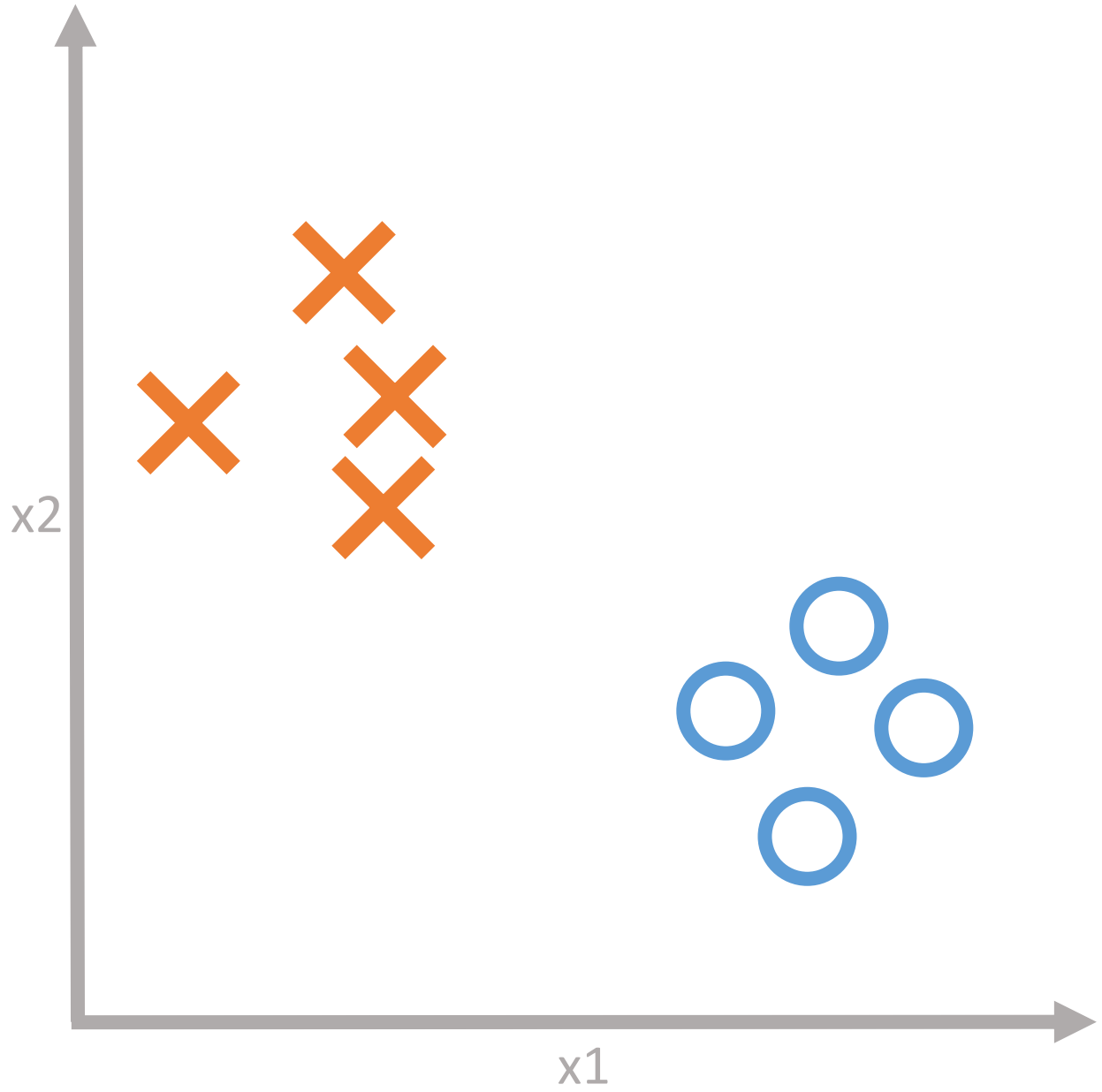

Data
Training
Test
New Data
ML
Algorithm
ML
Model
Prediction

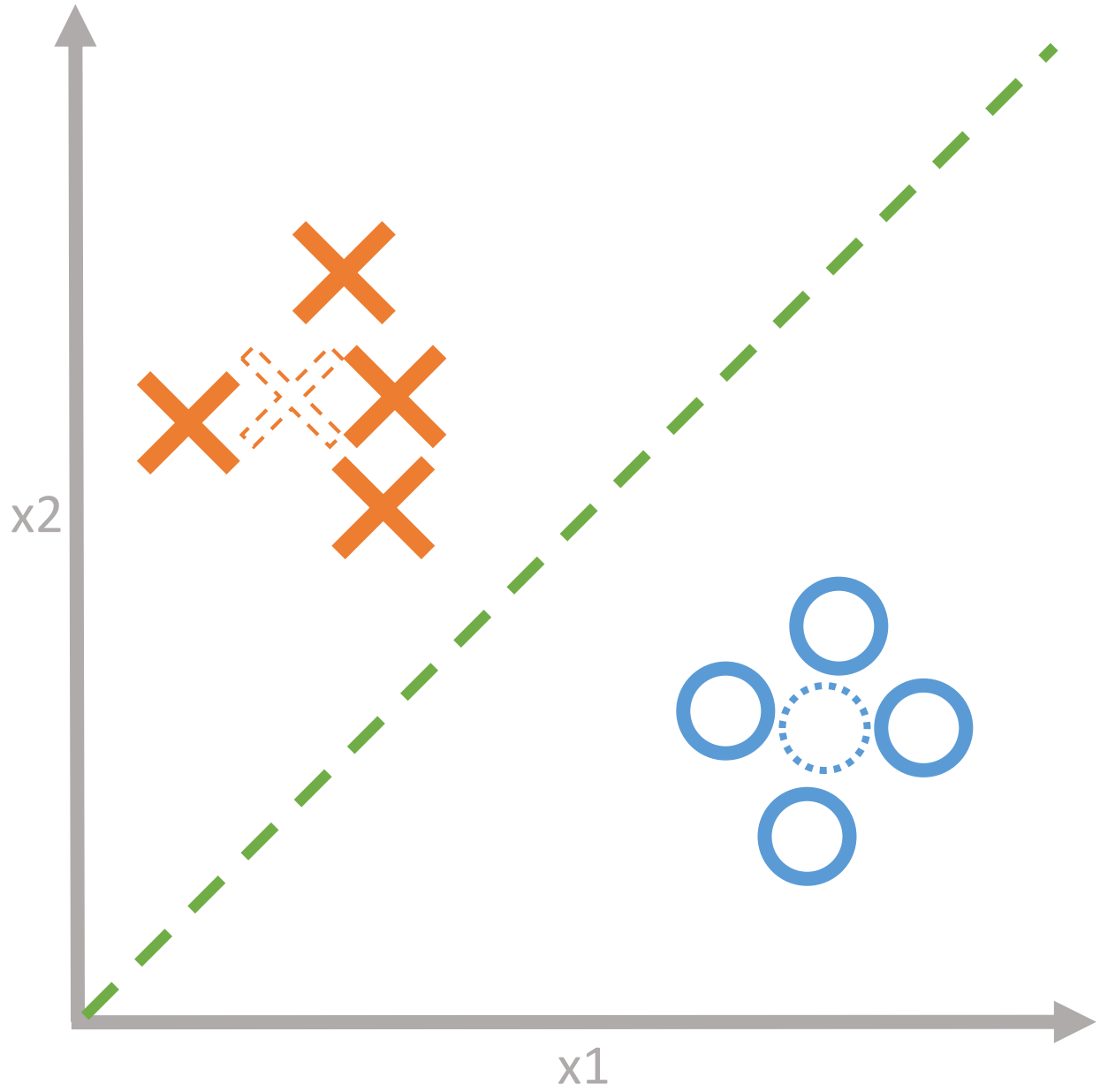
Types of Machine Learning



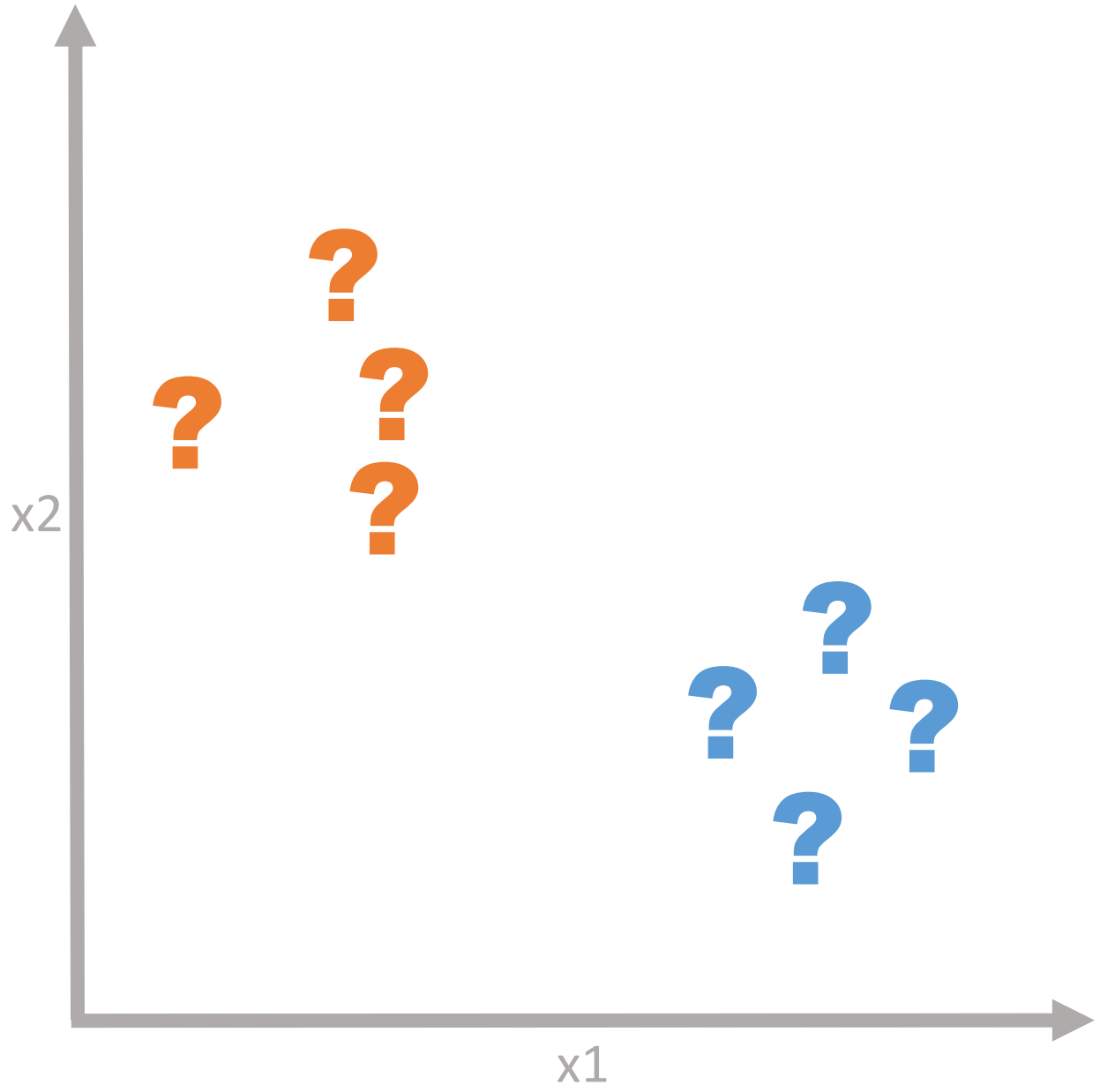
Supervised Learning

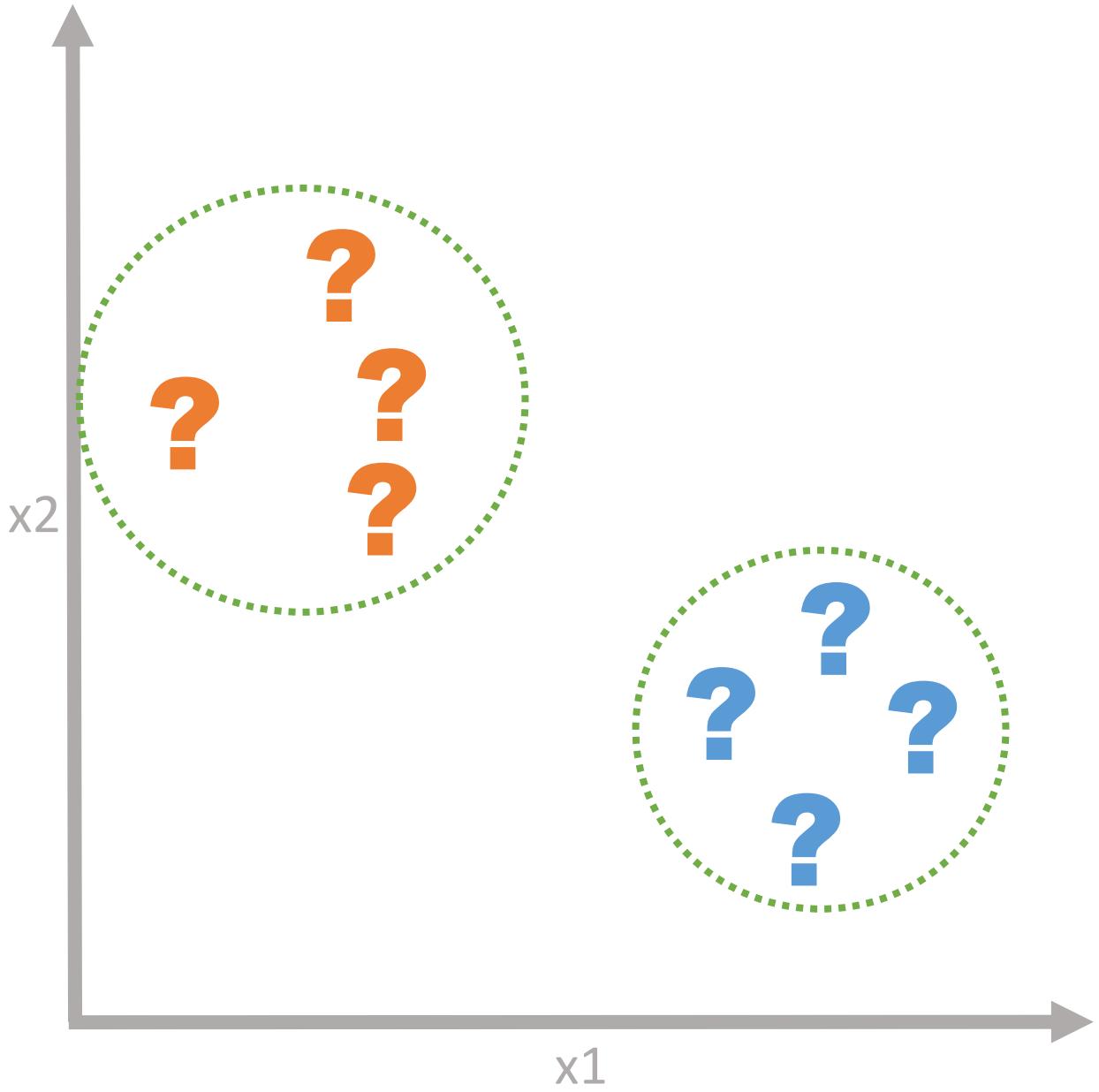

x2
x1

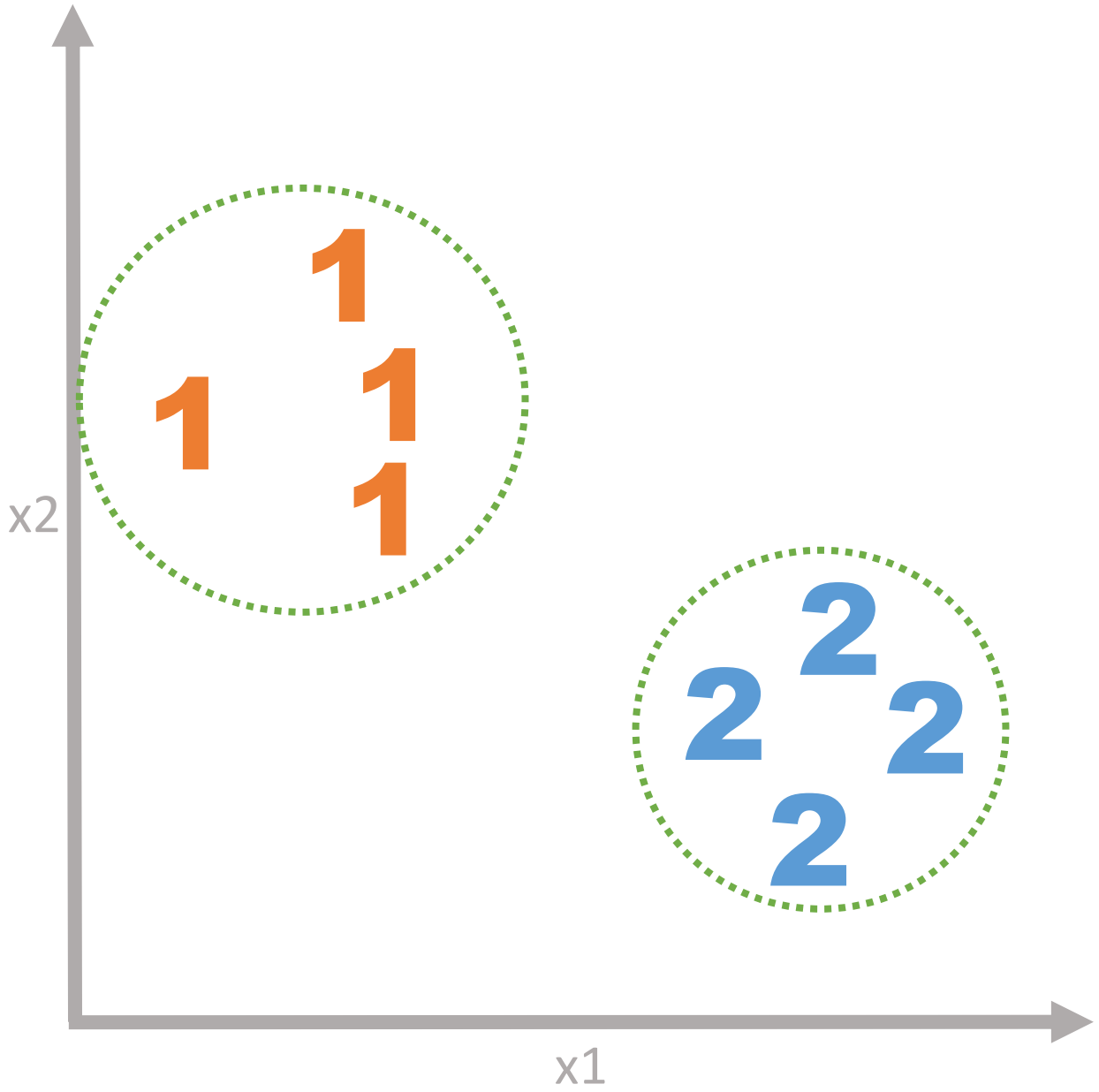

x2
x1

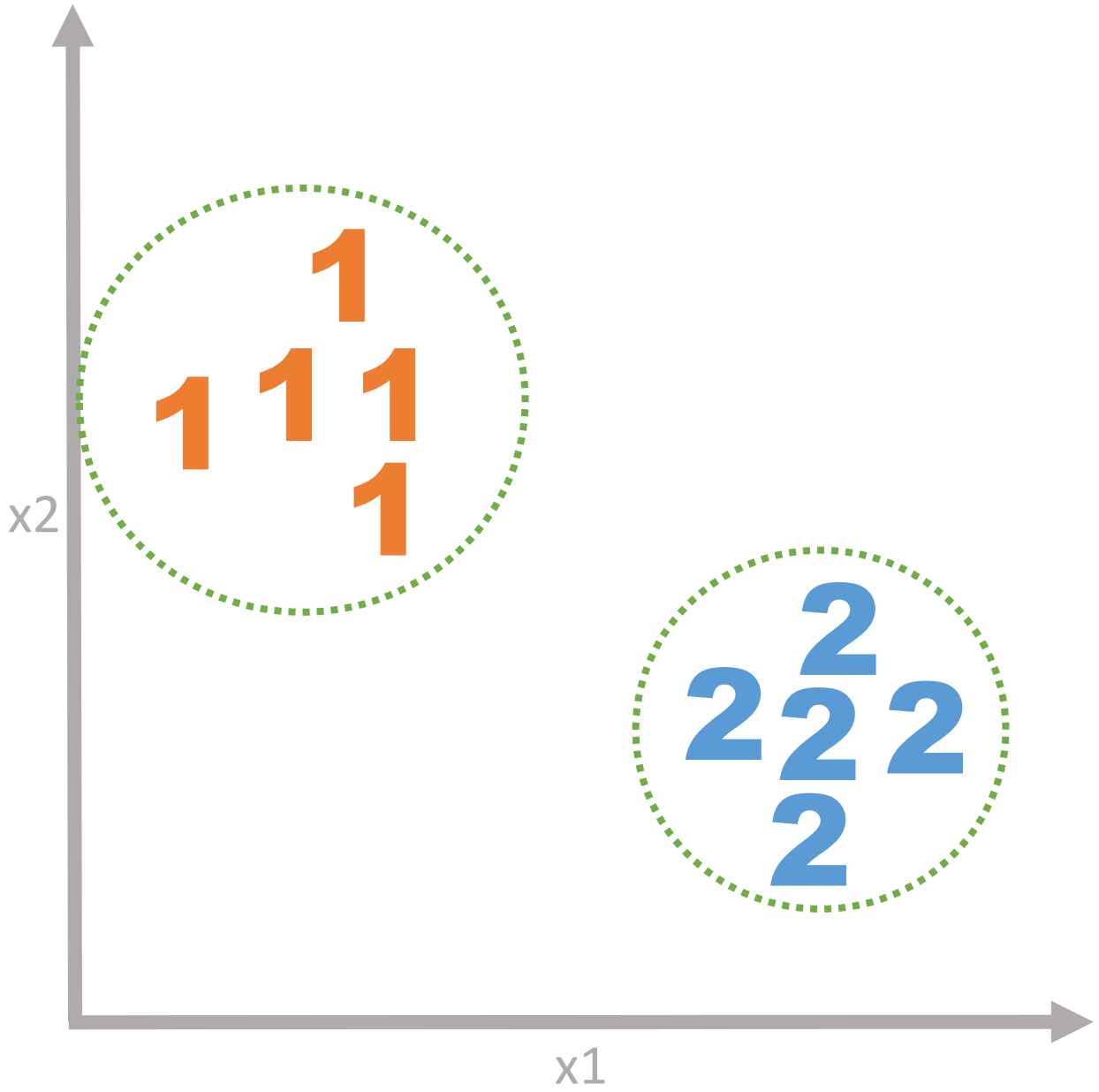

x2
x1

Unsupervised Learning

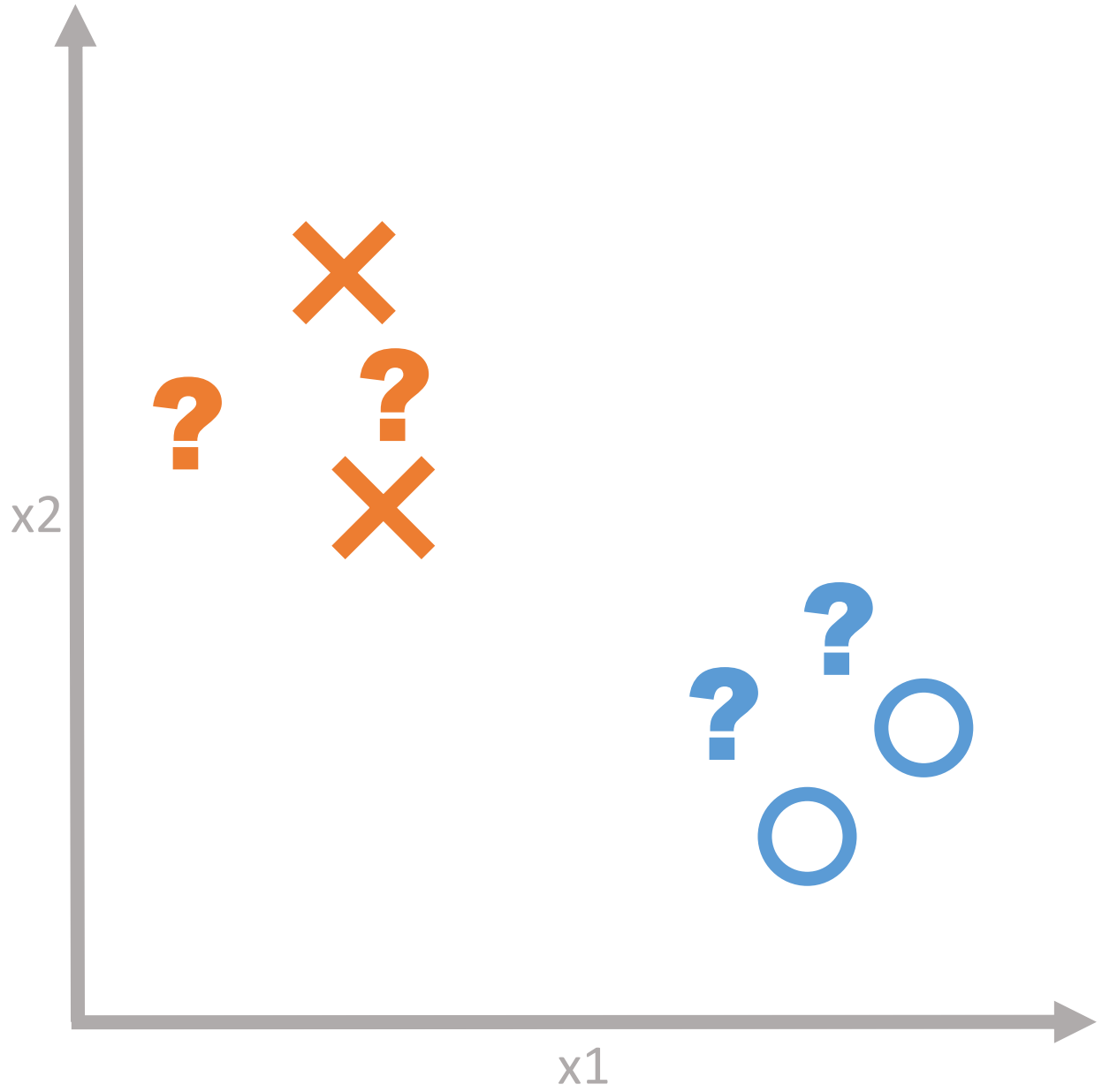

x2
x1

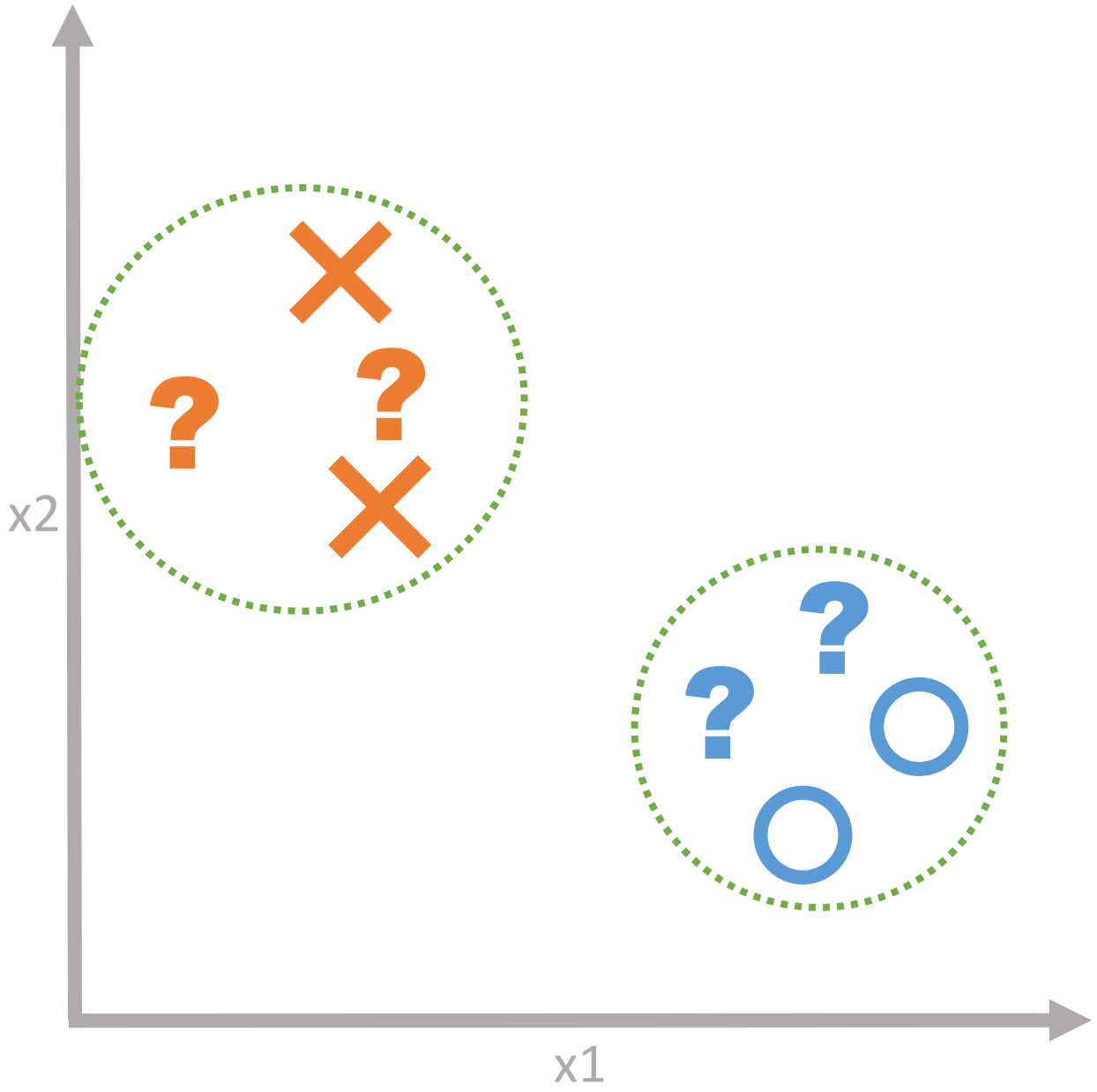

x2
x1

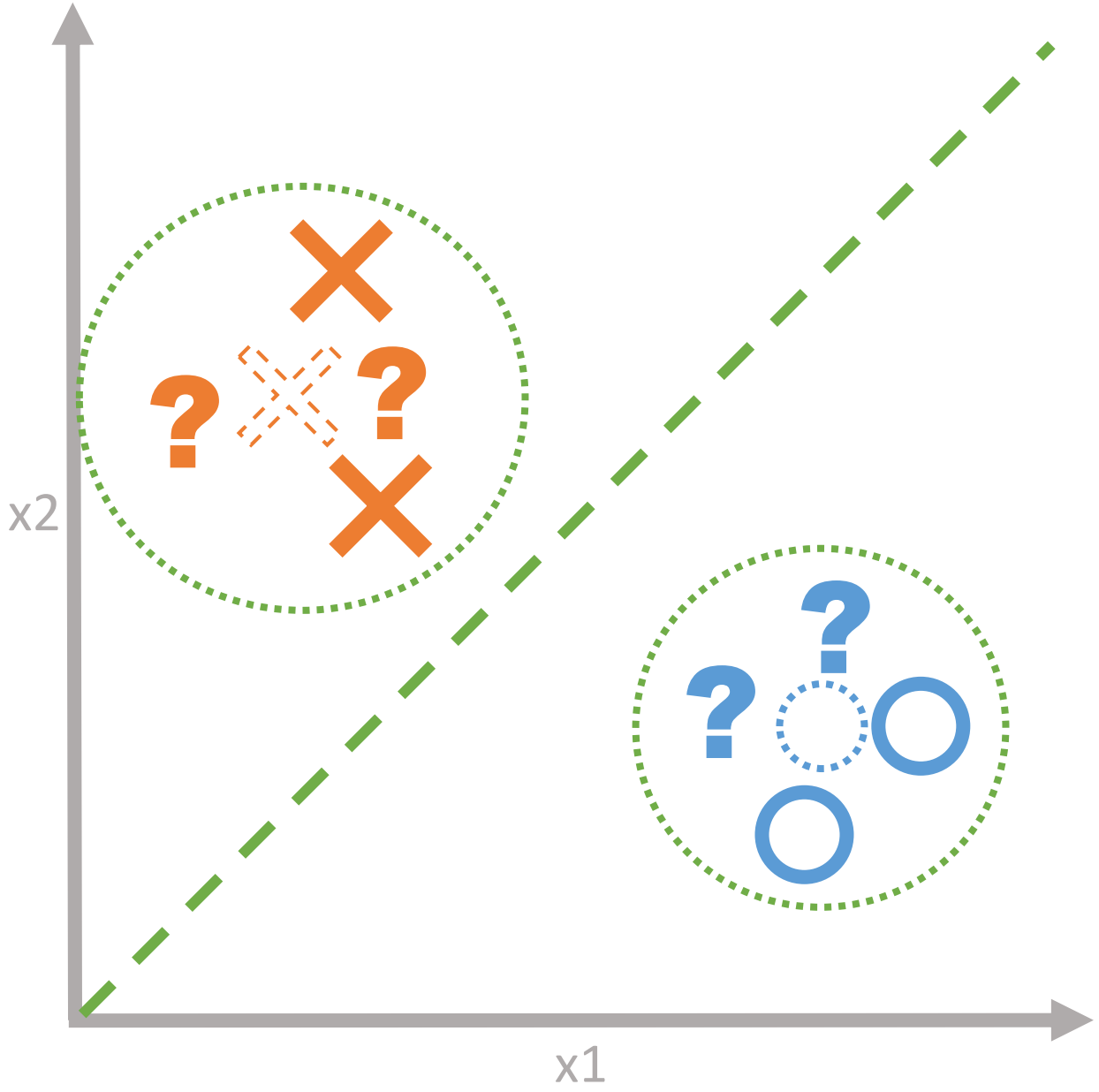

x2
x1
1
1
1
1
2
2
2
2


x2
1
1 1 1
1
2
2 2 2
2
x1

Semi-supervised Learning

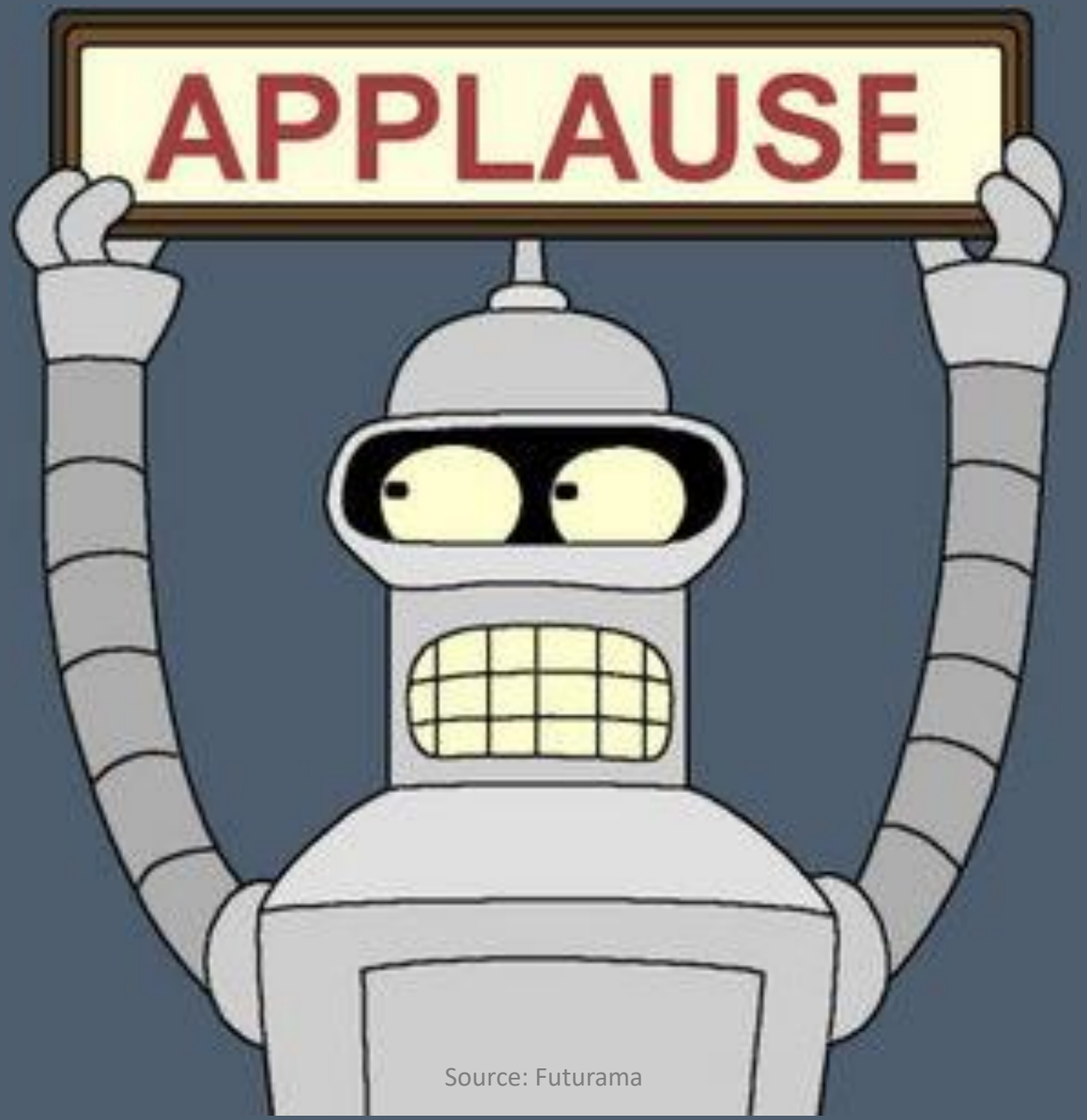

x2
x1


x2
x1


x2
x1

What Can Machine Learning Do?


$f(x)$
1.23



Source: Futurama

Introduction to R

What is R?

Open source

Language and environment

Numerical and graphical analysis

Cross platform

What is R?

Active development

Large user community

Modular and extensible

9000+ extensions


FREE


FREE

RGui (64-bit)

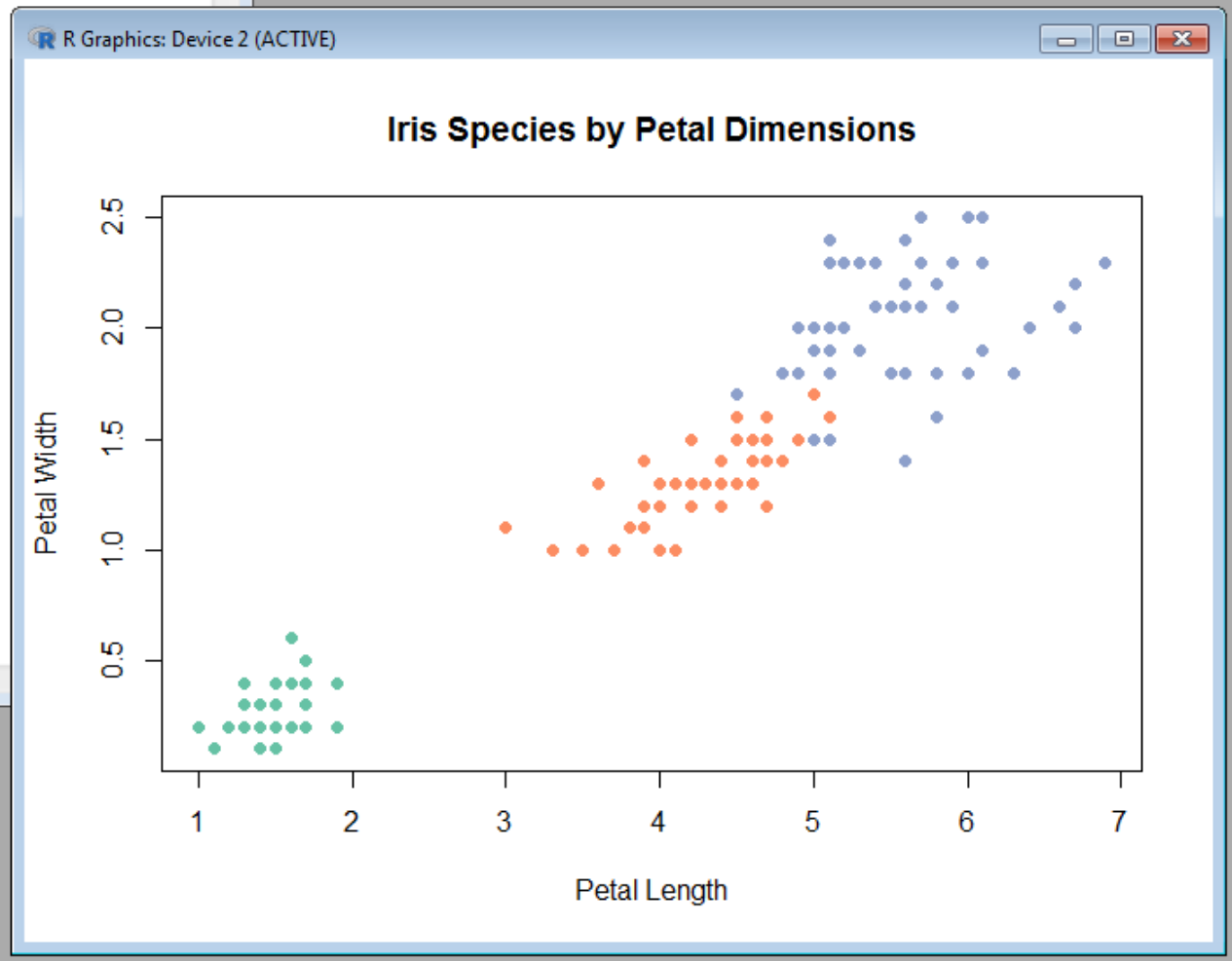
File History Resize Windows

R Console

```
> # Create a plot of species by dimension
> plot(
+     x = iris$Petal.Length,
+     y = iris$Petal.Width,
+     pch = 19,
+     col = palette[as.numeric(iris$Species)],
+     main = "Iris Species by Petal Dimensions",
+     xlab = "Petal Length",
+     ylab = "Petal Width")
>
> # Create a frequency table of species
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
>
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
>
> # Get the correlation coefficient
> cor(
+     x = iris$Petal.Length,
+     y = iris$Petal.Width)
[1] 0.9628654
```

R Graphics: Device 2 (ACTIVE)



RStudio

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function | Addins | Project: (None)

Script.R

Source on Save | Run | Source

```
16 # Create a frequency table of species
17 table(iris$Species)
18
19 # Get the average petal length
20 mean(iris$Petal.Length)
21
22 # Get the correlation coefficient
23 cor(
24     x = iris$Petal.Length,
25     y = iris$Petal.Width)
```

21:1 (Top Level) R Script

Console ~/

```
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
> # Get the correlation coefficient
> cor(
+ x = iris$Petal.Length,
+ y = iris$Petal.Width)
[1] 0.9628654
>
```

Environment History

Import Dataset | List

Global Environment

Data

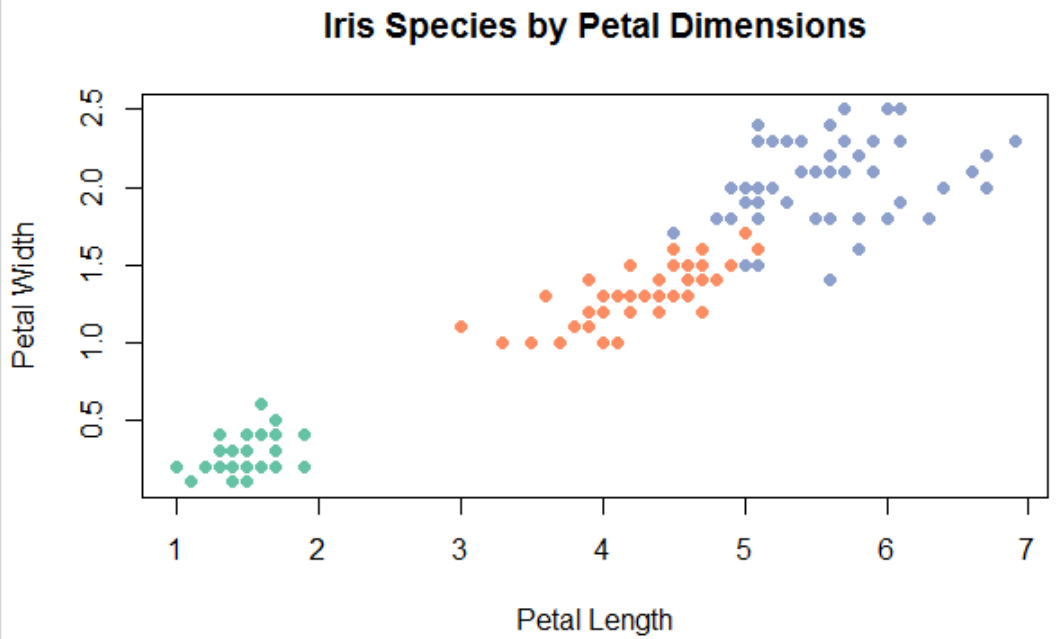
iris	150 obs. of 5 variables

Values

palette	chr [1:3] "#66C2A5" "#FC8D62" "#8DA0C…

Files Plots Packages Help Viewer

Zoom | Export | Publish



Script.R - Microsoft Visual Studio

File Edit View NCrunch Project Debug Team Tools Architecture Test ReSharper R Tools Analyze Window Help

Matthew Renze

Quick Launch (Ctrl+Q)

Attach...

Script.R

```
        main = "Iris Species by Petal Dimensions",
        xlab = "Petal Length",
        ylab = "Petal Width")

# Create a frequency table of species
table(iris$Species)

# Get the average petal length
mean(iris$Petal.Length)

# Get the correlation coefficient
cor(
    x = iris$Petal.Length,
    y = iris$Petal.Width)
```

100 %

R Interactive

Attach Debugger

```
> # Create a frequency table of species
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
> # Get the average petal length
> mean(iris$Petal.Length)
[1] 3.758
> # Get the correlation coefficient
> cor(
+     x = iris$Petal.Length,
+     y = iris$Petal.Width)
[1] 0.9628654
>
```

100 %

Variable Explorer

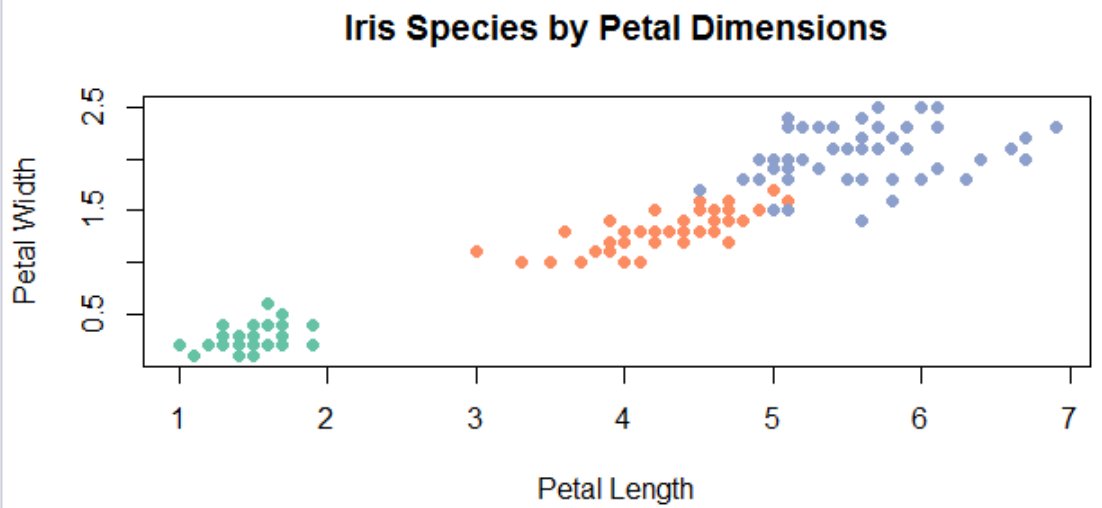
Search

.GlobalEnv

Name	Value	Class	Type
iris	150 obs. of 5 variables	data.frame	list
palette	chr [1:3] "#66C2A5" "#FC8D62" "#8DA0CB"	character	character

Variable Explorer | R History

R Plot



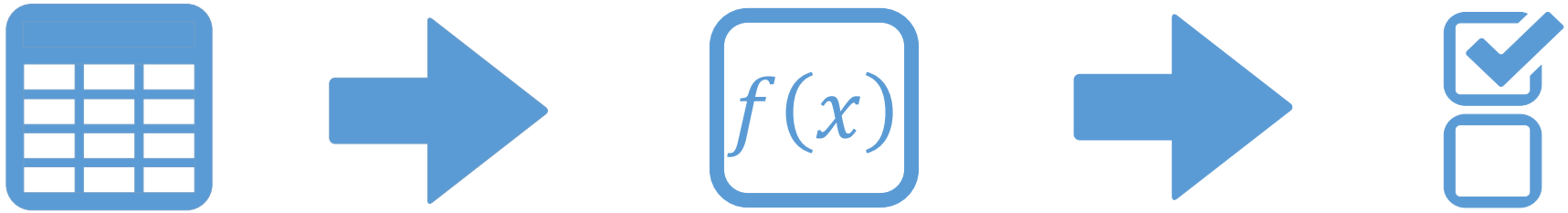
Solution Explorer | R Plot | R Package Manager | R Help

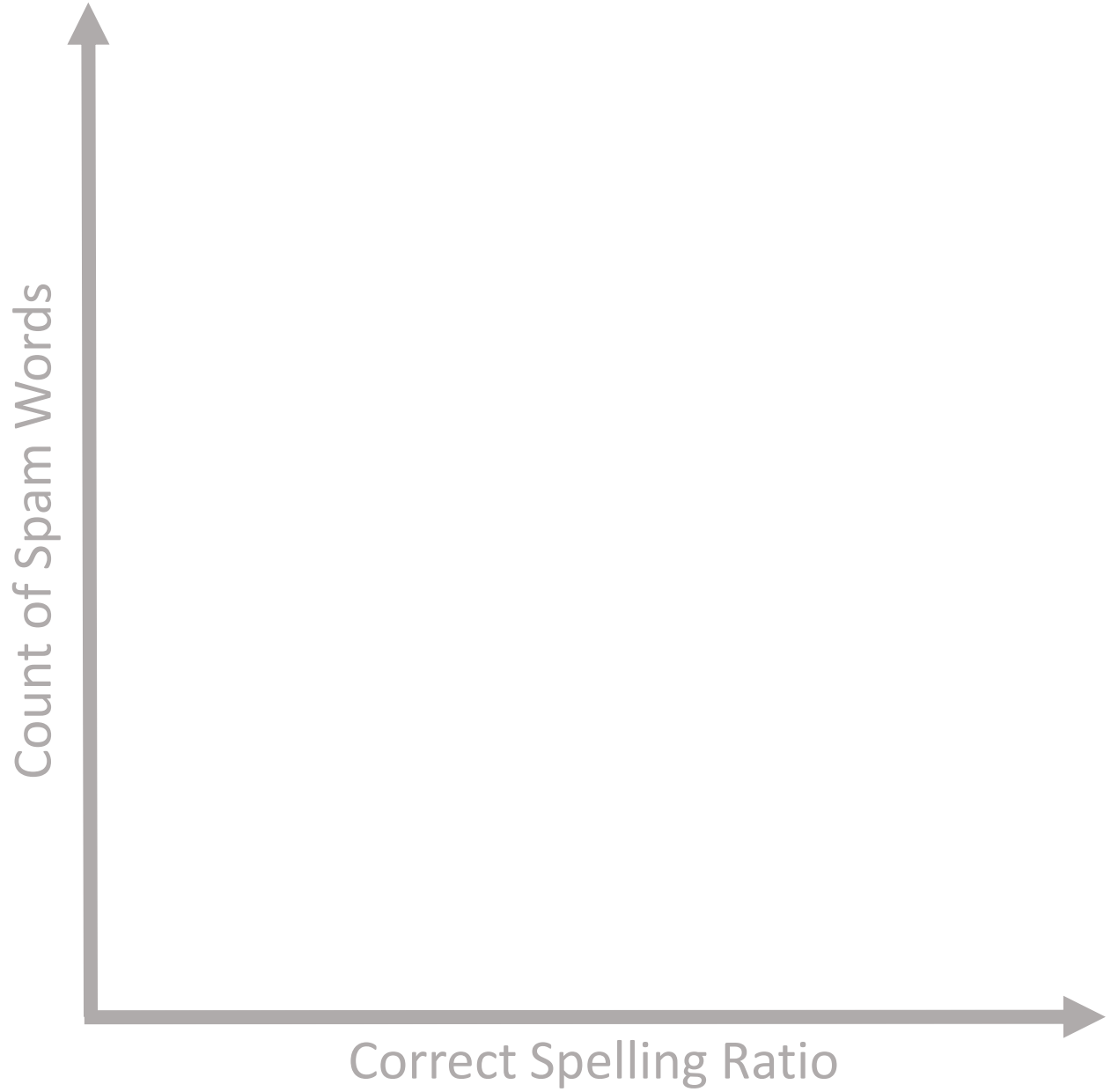
Error List | Output | Azure App Service Activity

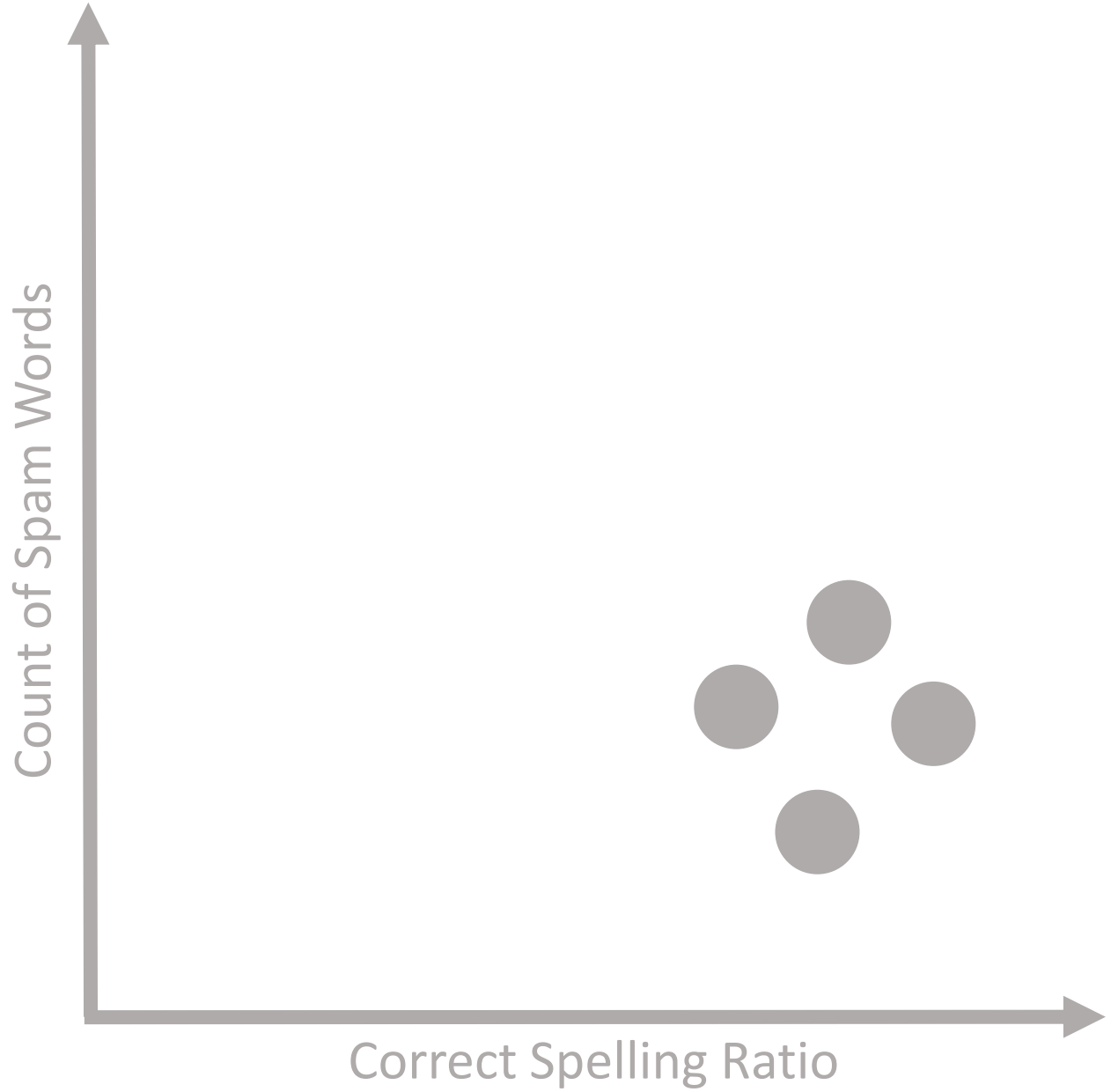
Ready | Ln 30 | Col 1 | Ch 1 | INS | 7 | 0 | Root | master

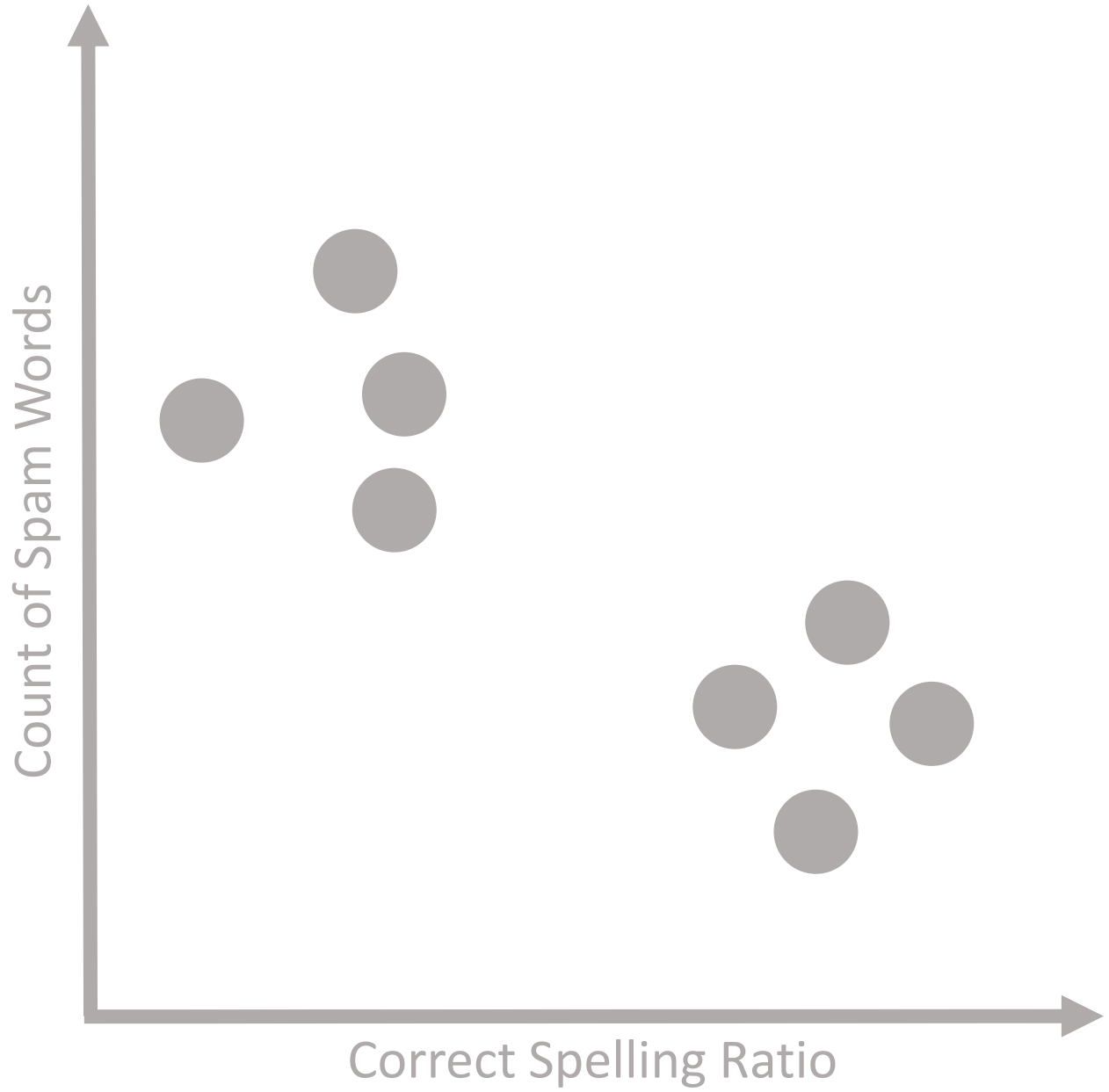
Code Demo

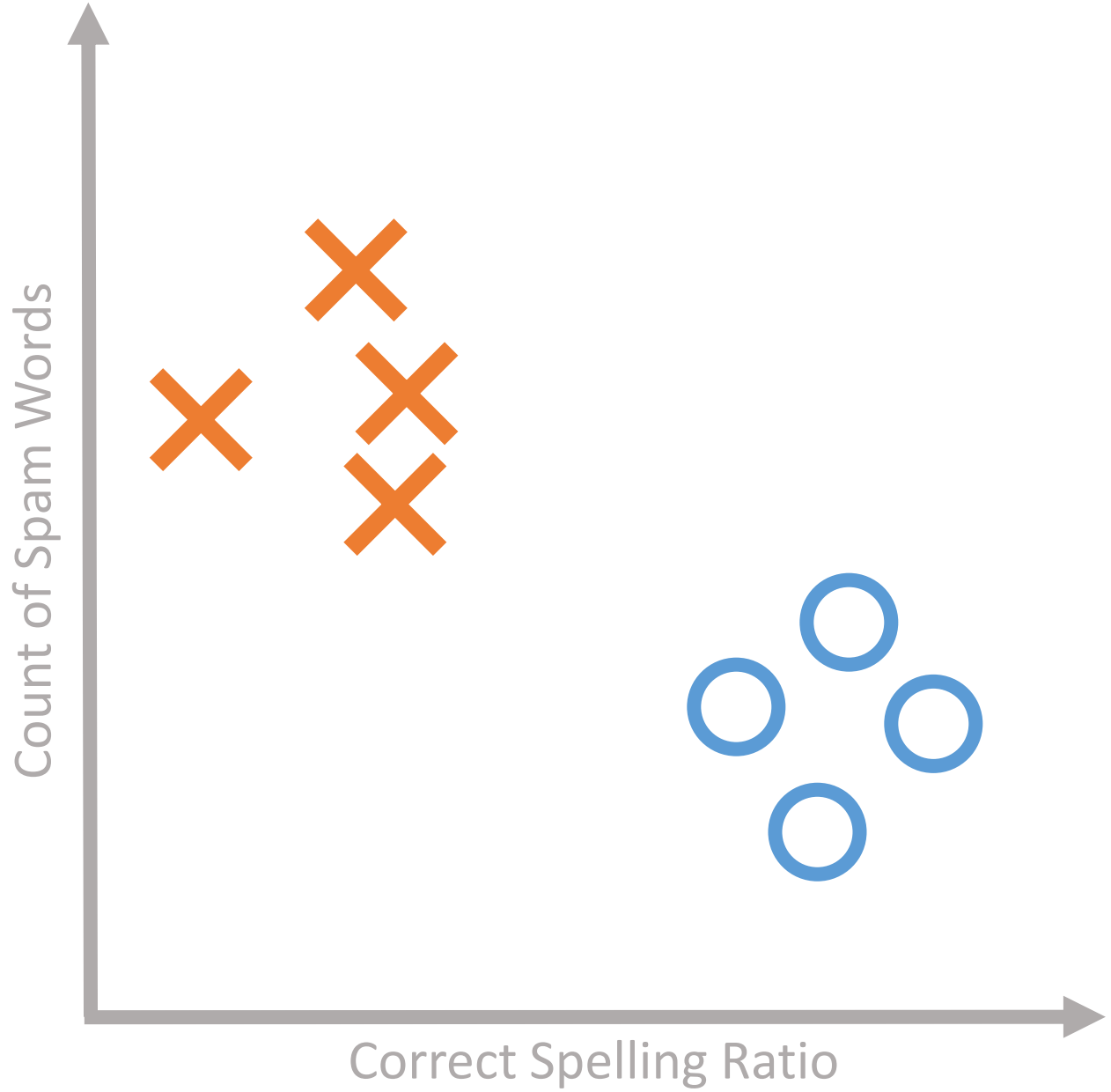
Classification

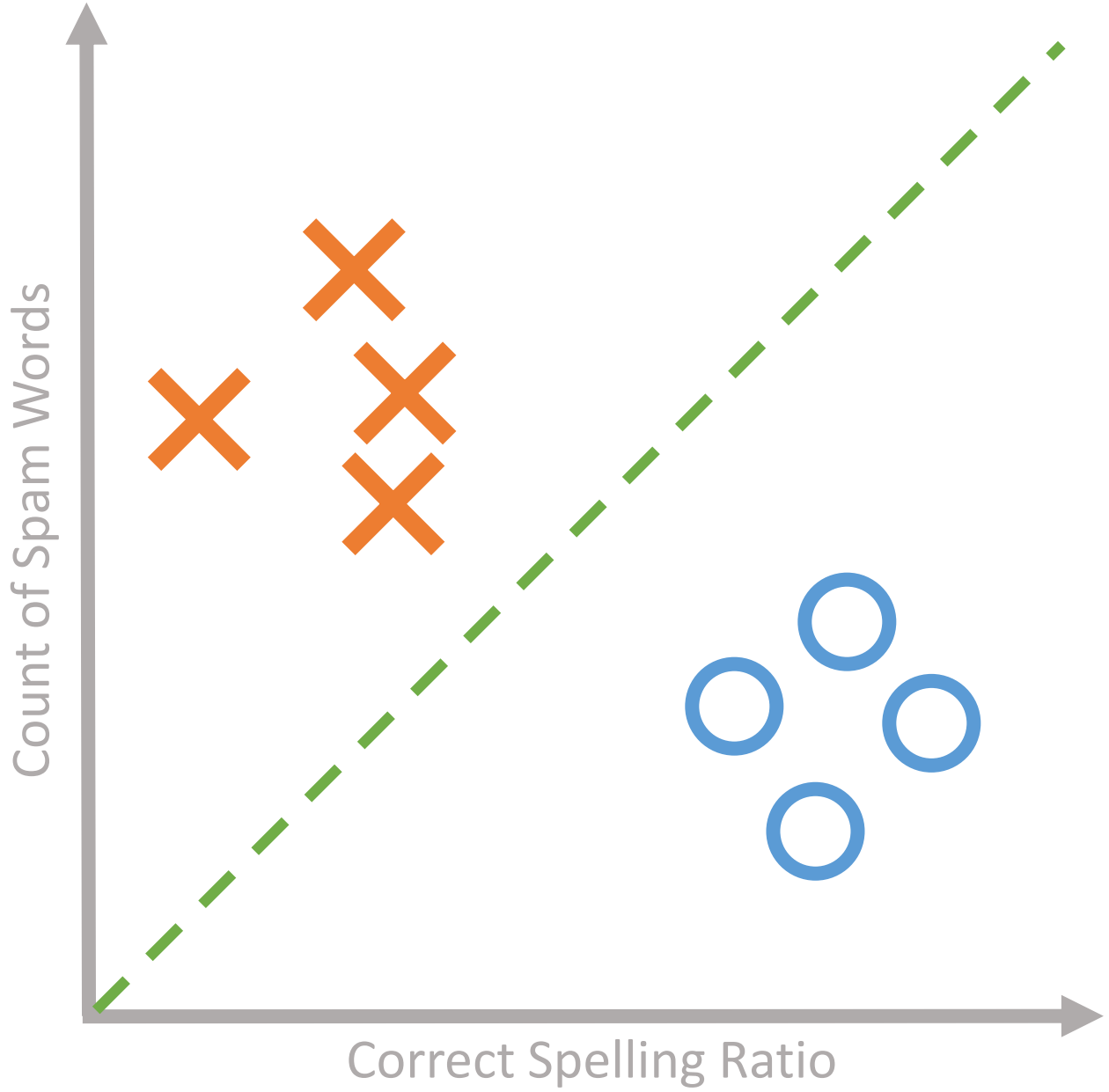

$f(x)$

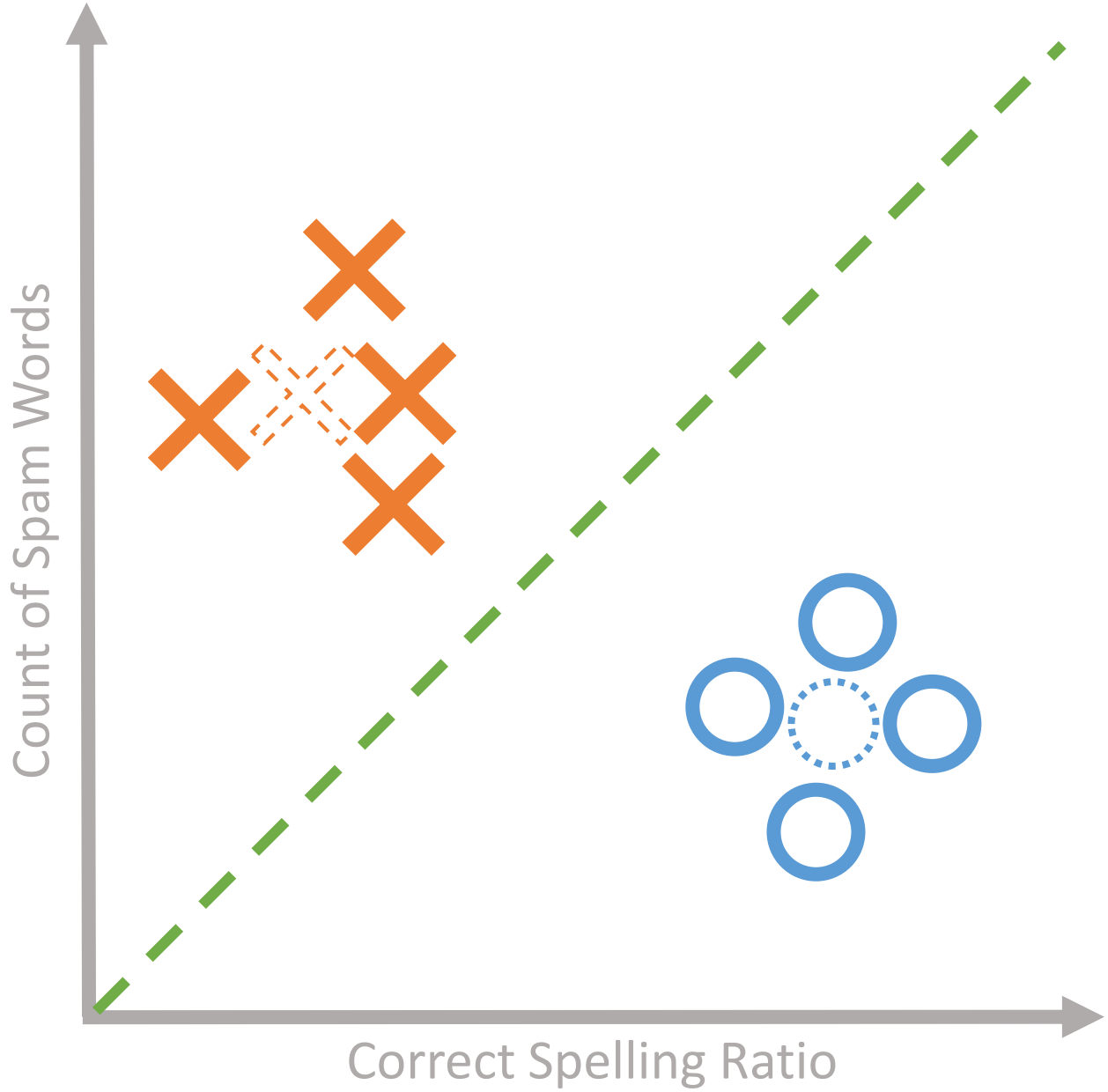

Count of Spam Words
Correct Spelling Ratio


Count of Spam Words
Correct Spelling Ratio


Count of Spam Words
Correct Spelling Ratio


Count of Spam Words
Correct Spelling Ratio


Count of Spam Words
Correct Spelling Ratio


Count of Spam Words
Correct Spelling Ratio

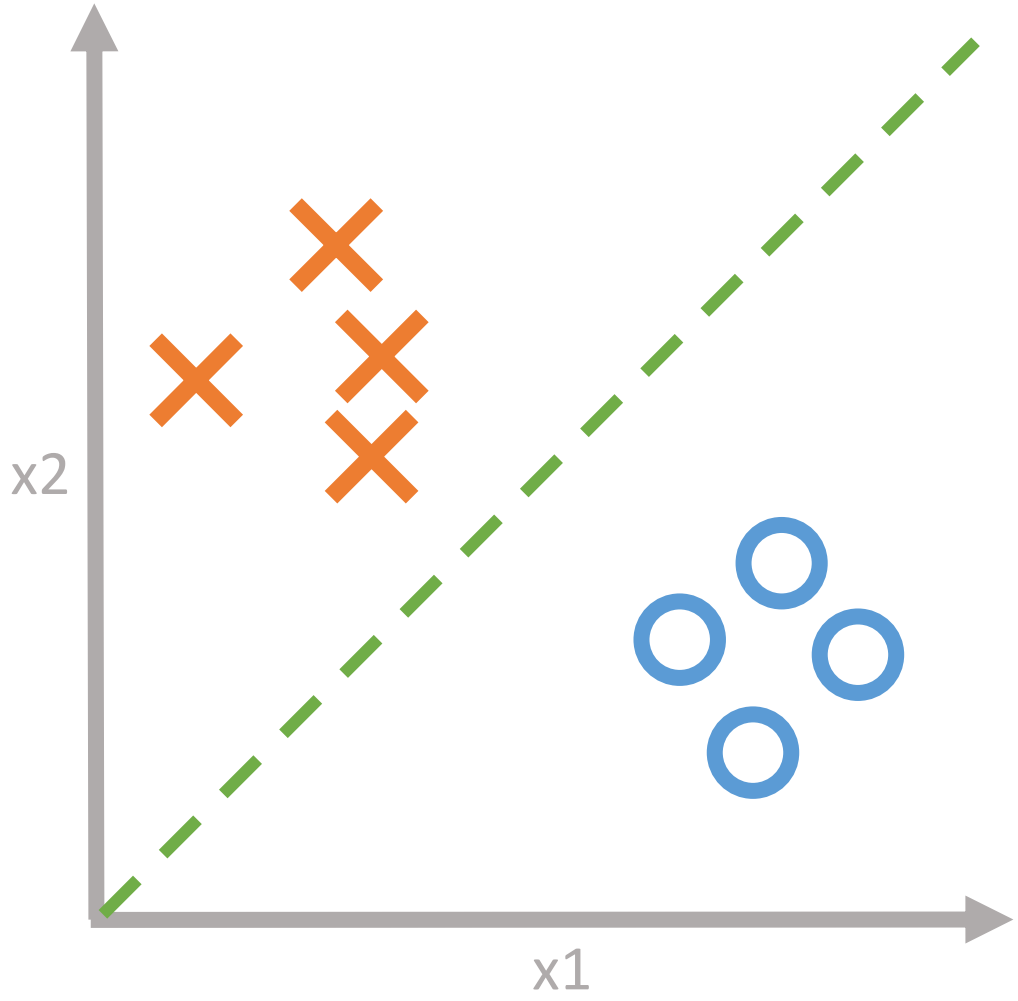
Classification Algorithms

Decision Tree Classifier

Naïve Bayes Classifier

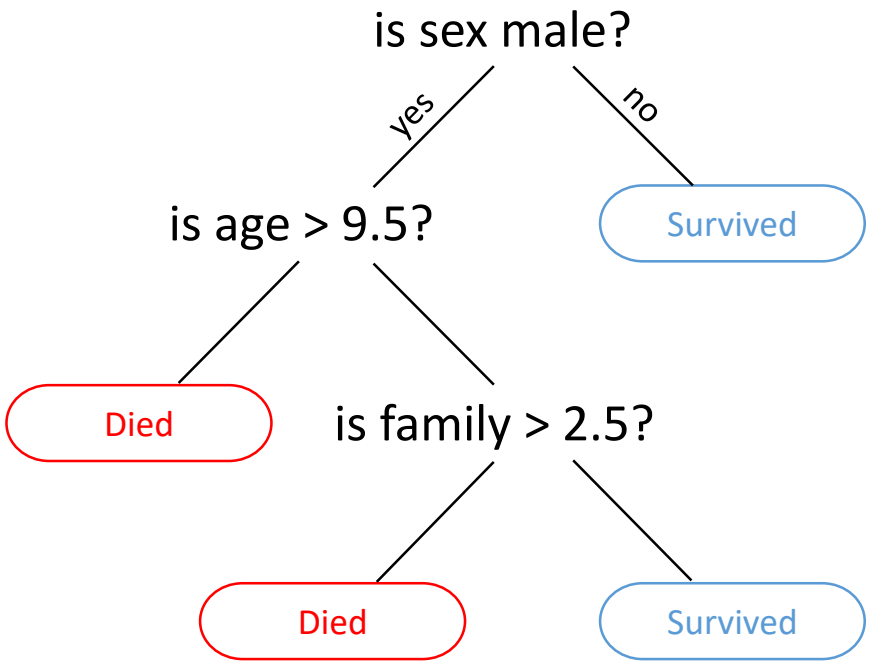
Support Vector Machine

Neural Network



Decision Tree Classifier

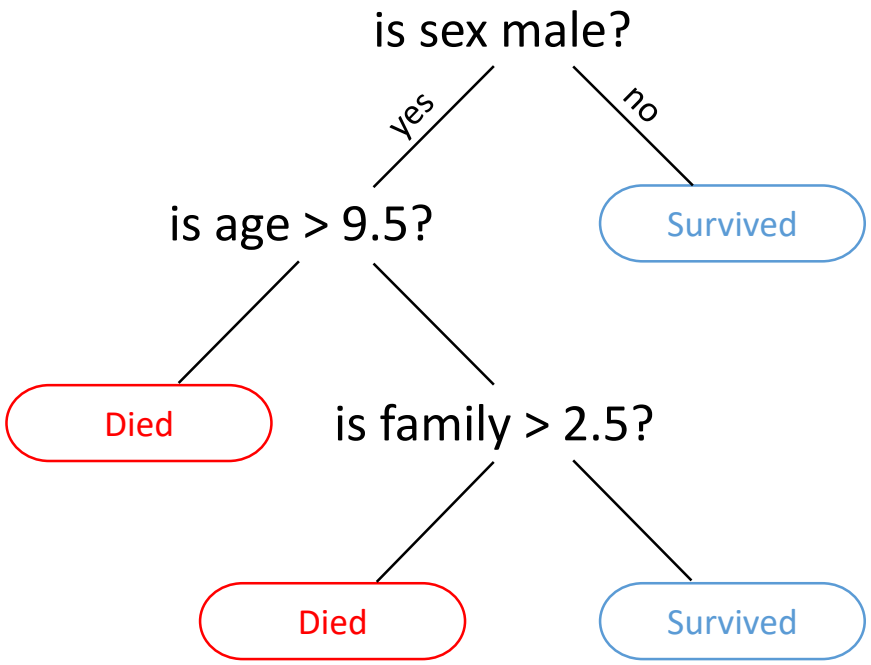
Supervised learning



Decision Tree Classifier

Supervised learning

Tree of decisions

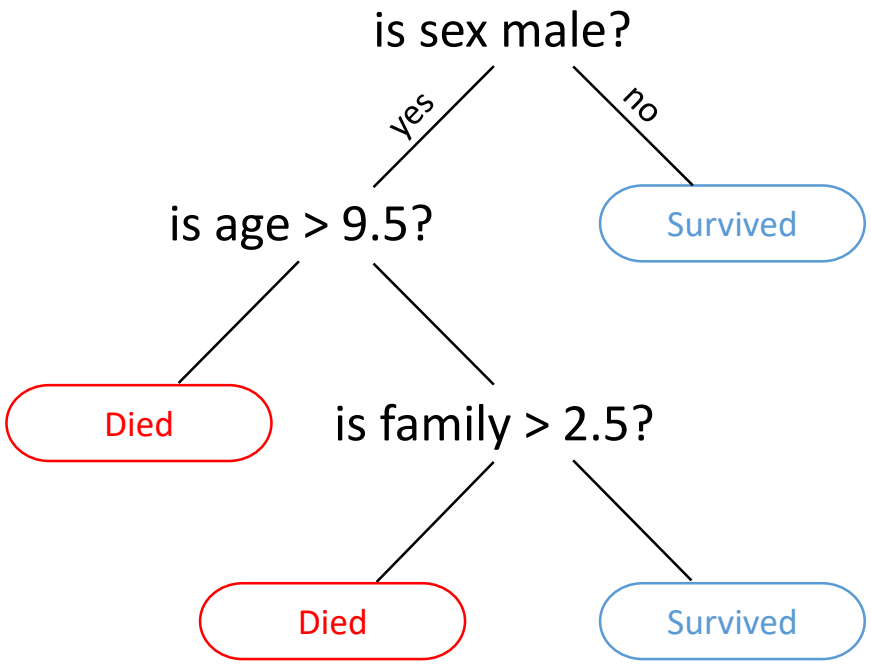


Decision Tree Classifier

Supervised learning

Tree of decisions

Easy to understand



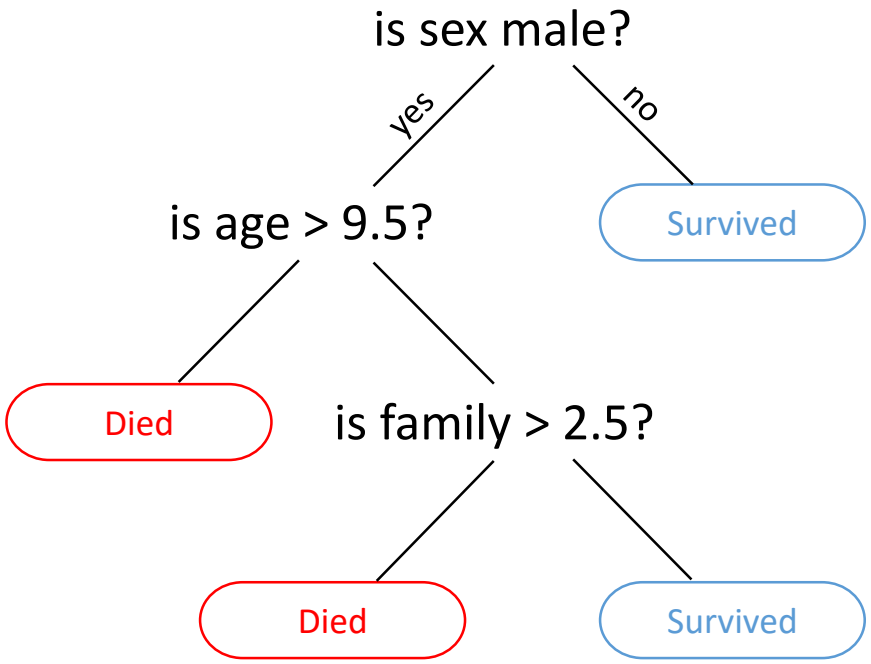
Decision Tree Classifier

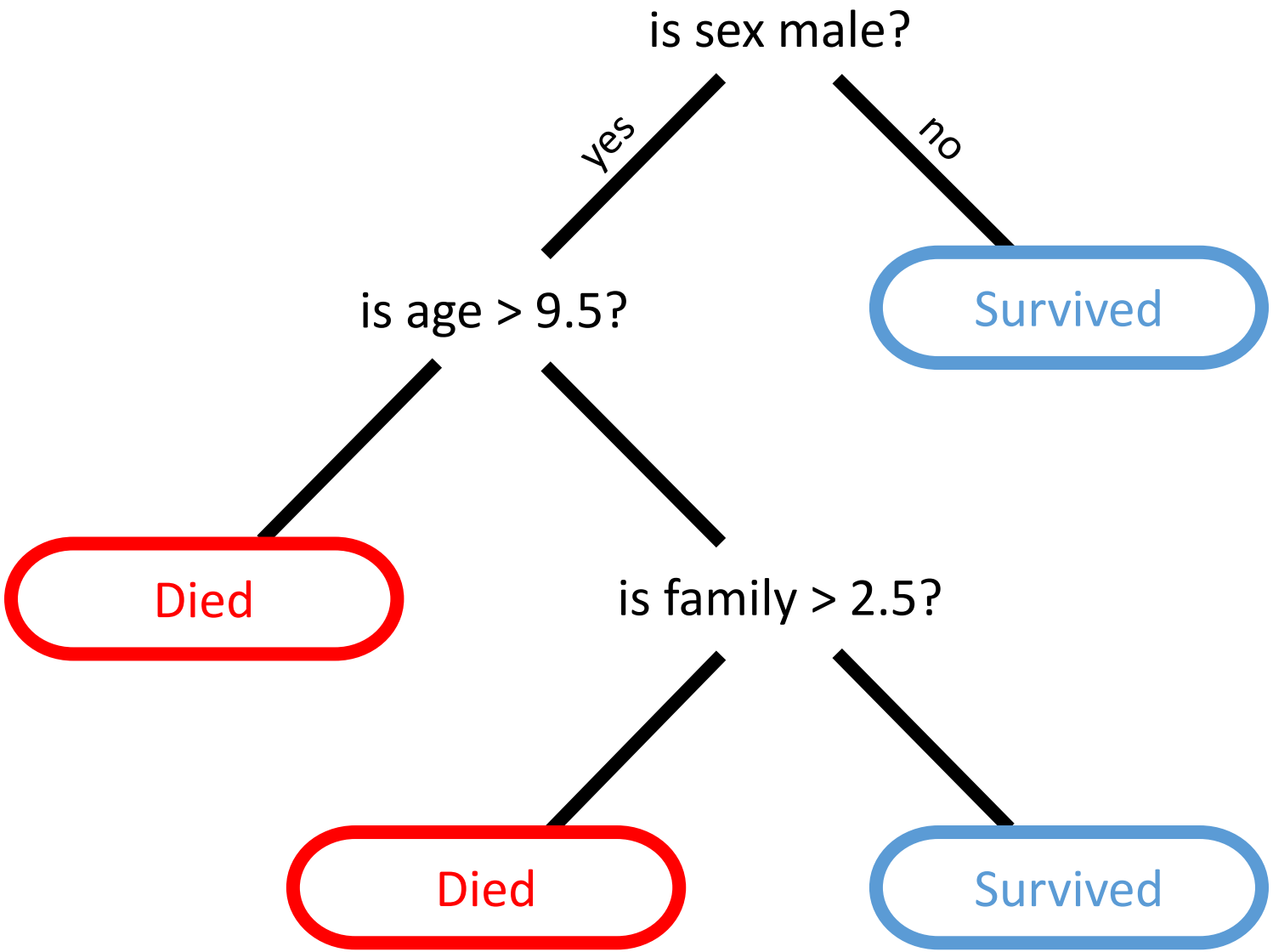
Supervised learning

Tree of decisions

Easy to understand

Transparent




is sex male?
yes
no
is age > 9.5?
Survived
Died
is family > 2.5?
Died
Survived

Iris Data Set

Iris Setosa

Iris Versicolor

Iris Virginica

Photos by Radomił Binek, Danielle Langlois, and Frank Mayfield

Iris Data Set

Fisher's Iris Data				
Species	Petal Length	Petal Width	Sepal Length	Sepal Width
setosa	1.1	0.1	4.3	3
setosa	1.4	0.2	4.4	2.9
setosa	1.3	0.2	4.4	3
setosa	1.3	0.2	4.4	3.2
setosa	1.3	0.3	4.5	2.3
…		…	…	…

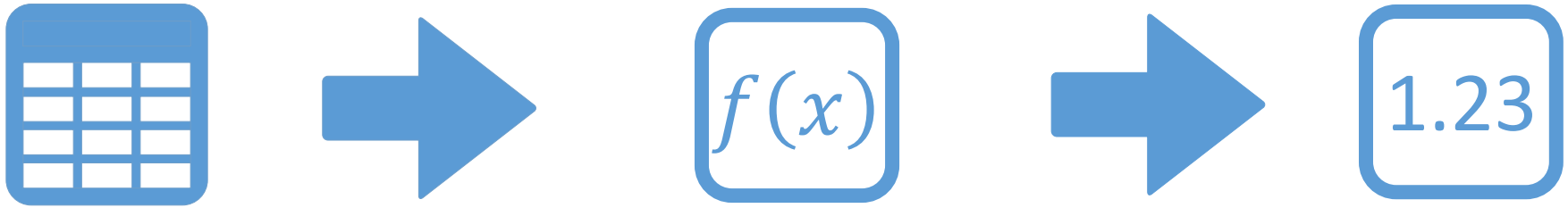
Classification Demo

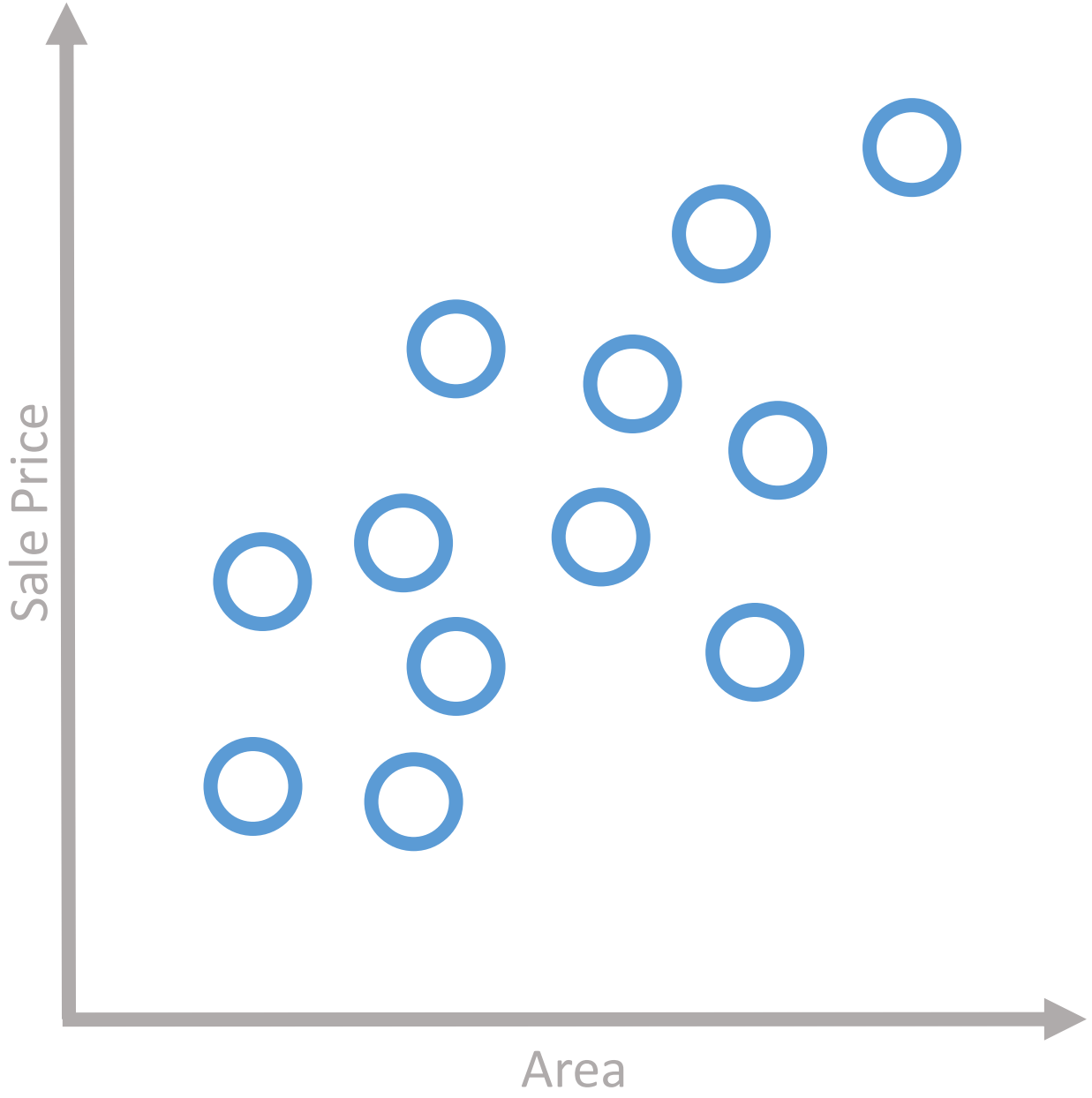
Goal: Predict species based on petal and sepal measurements

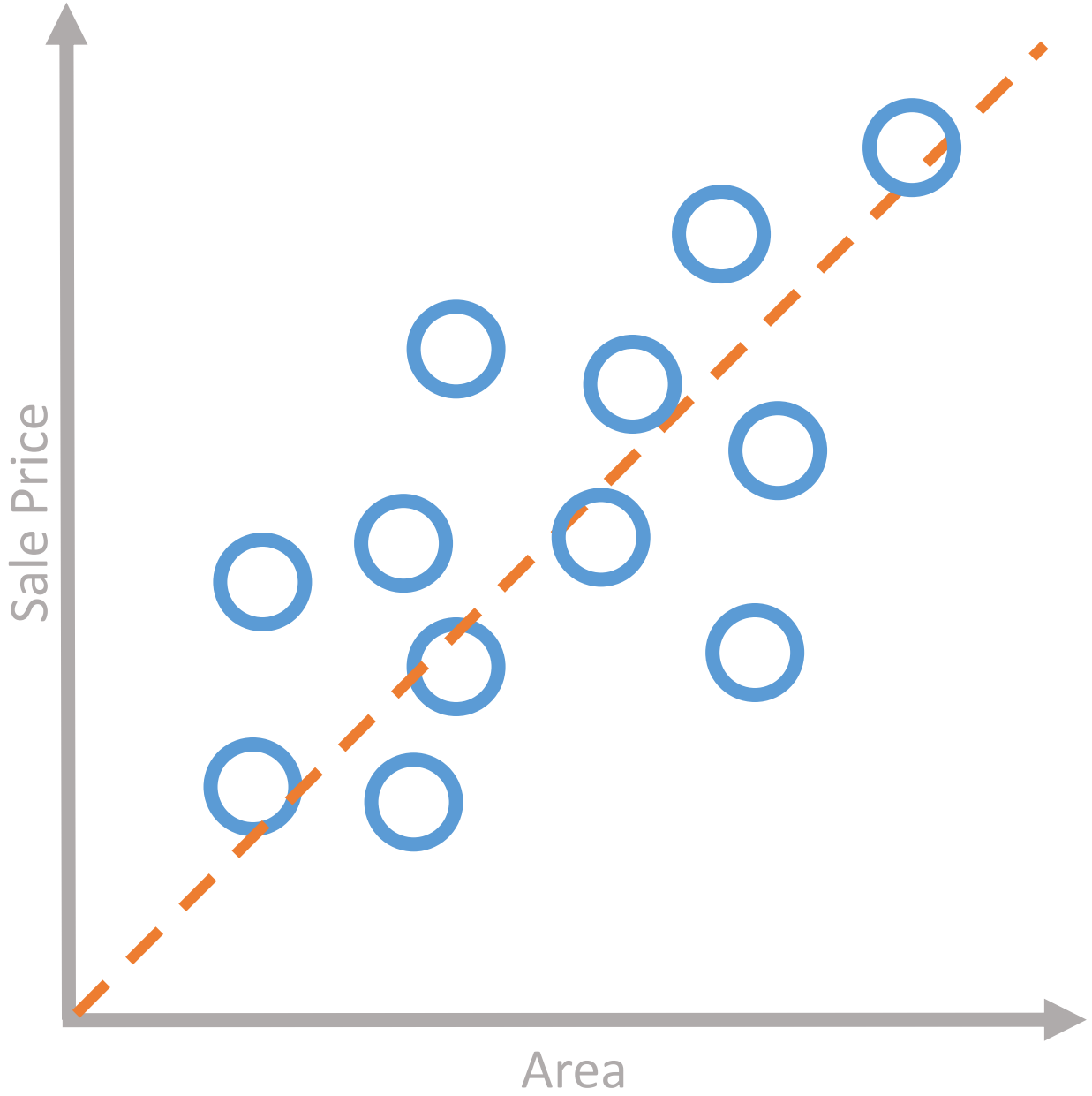
Real-World Examples

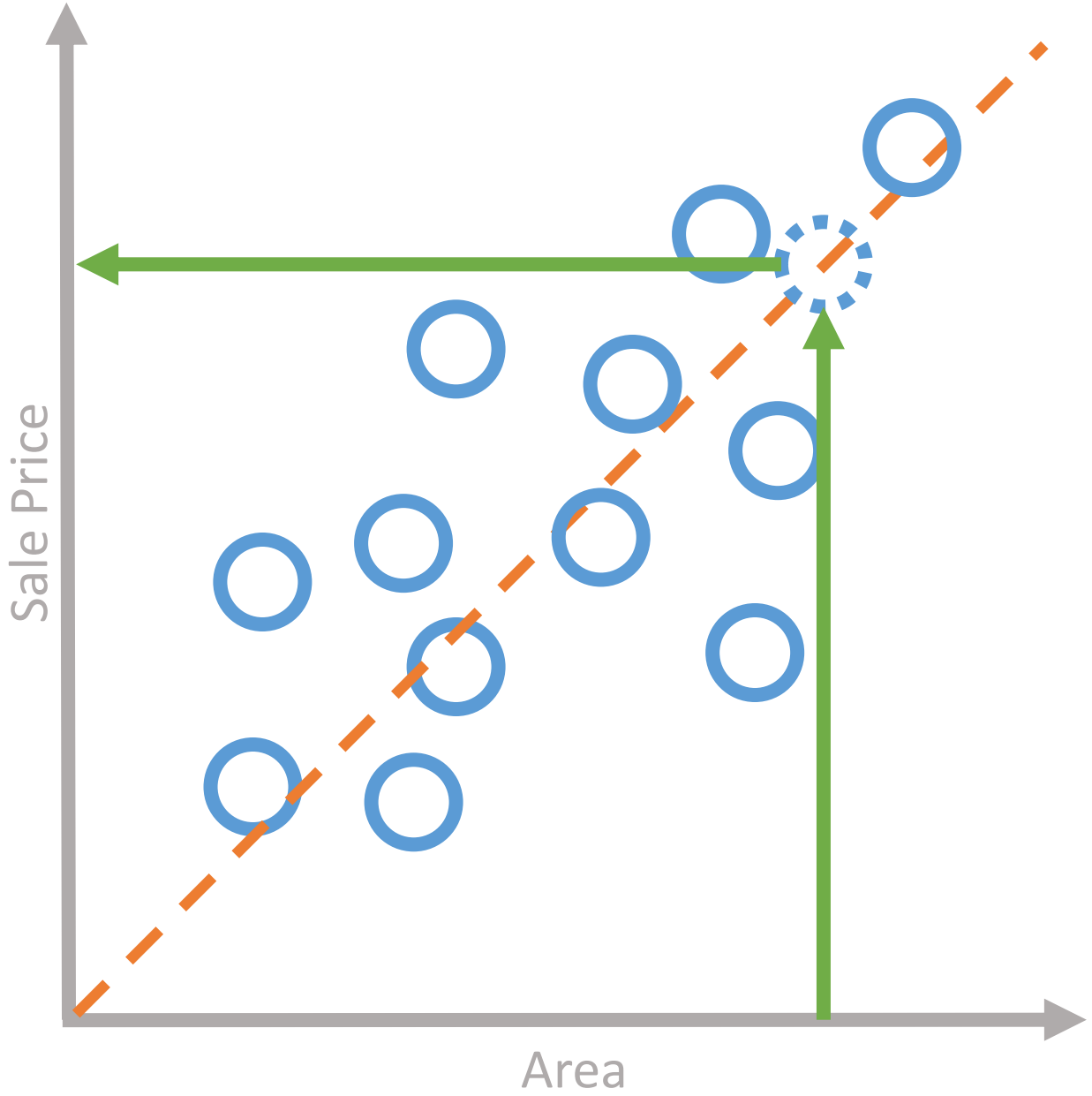
- Should we approve this loan?
- Will this customer buy from us?
- Should we replace this part?
- Does this person have cancer?

Regression


$f(x)$
1.23


Sale Price
Area


Sale Price
Area


Sale Price
Area

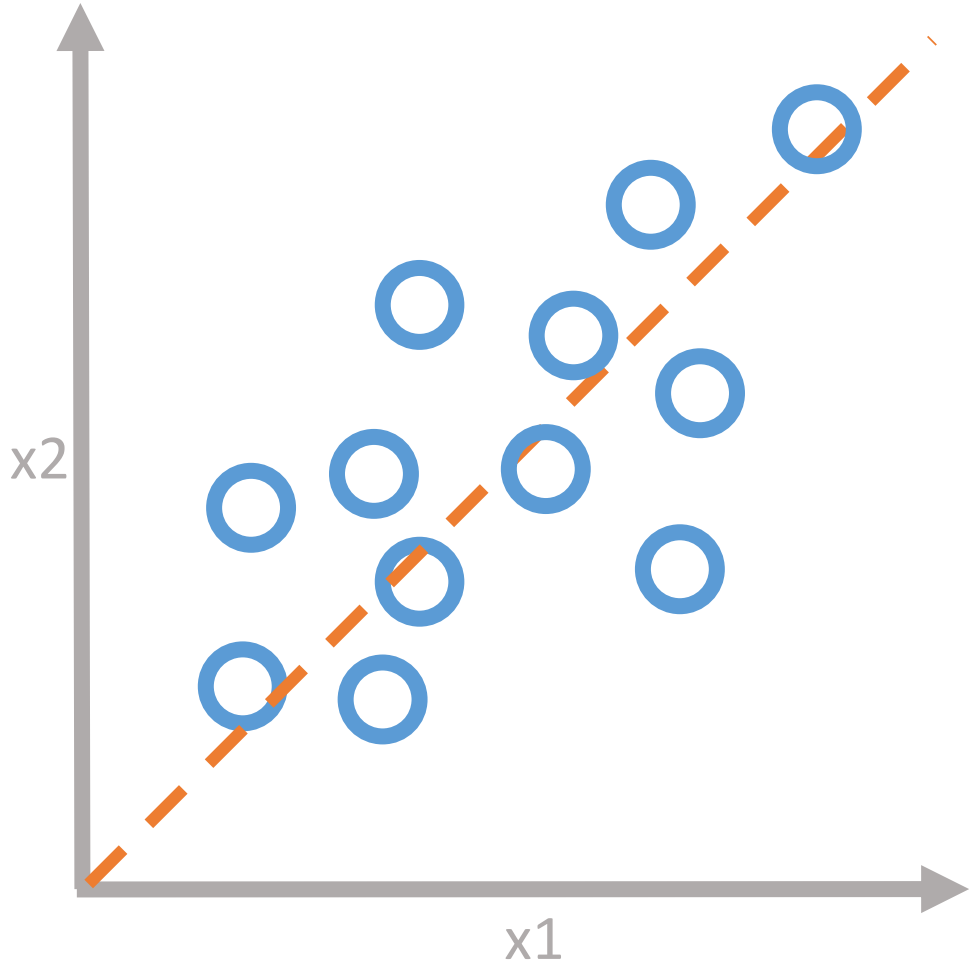
Regression Algorithms

Linear Regression

Polynomial Regression

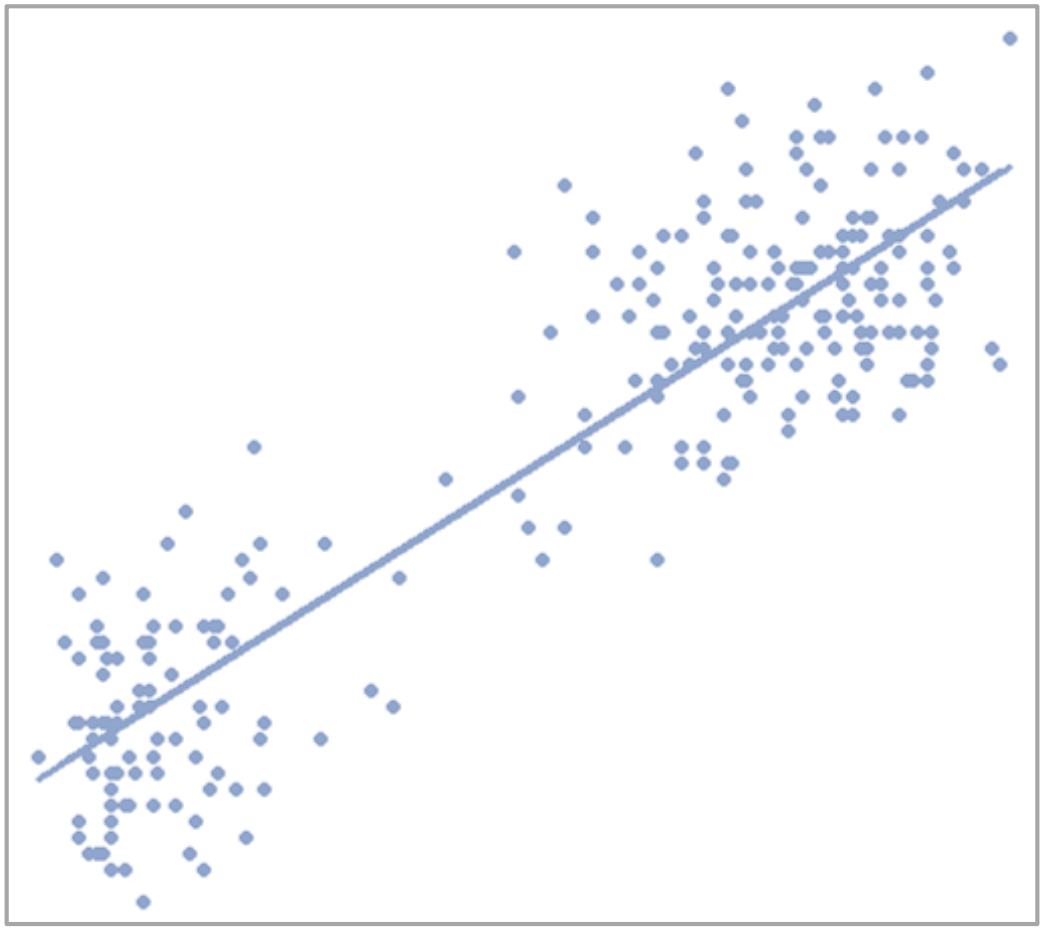
Lasso Regression

ElasticNet Regression



Simple Linear Regression

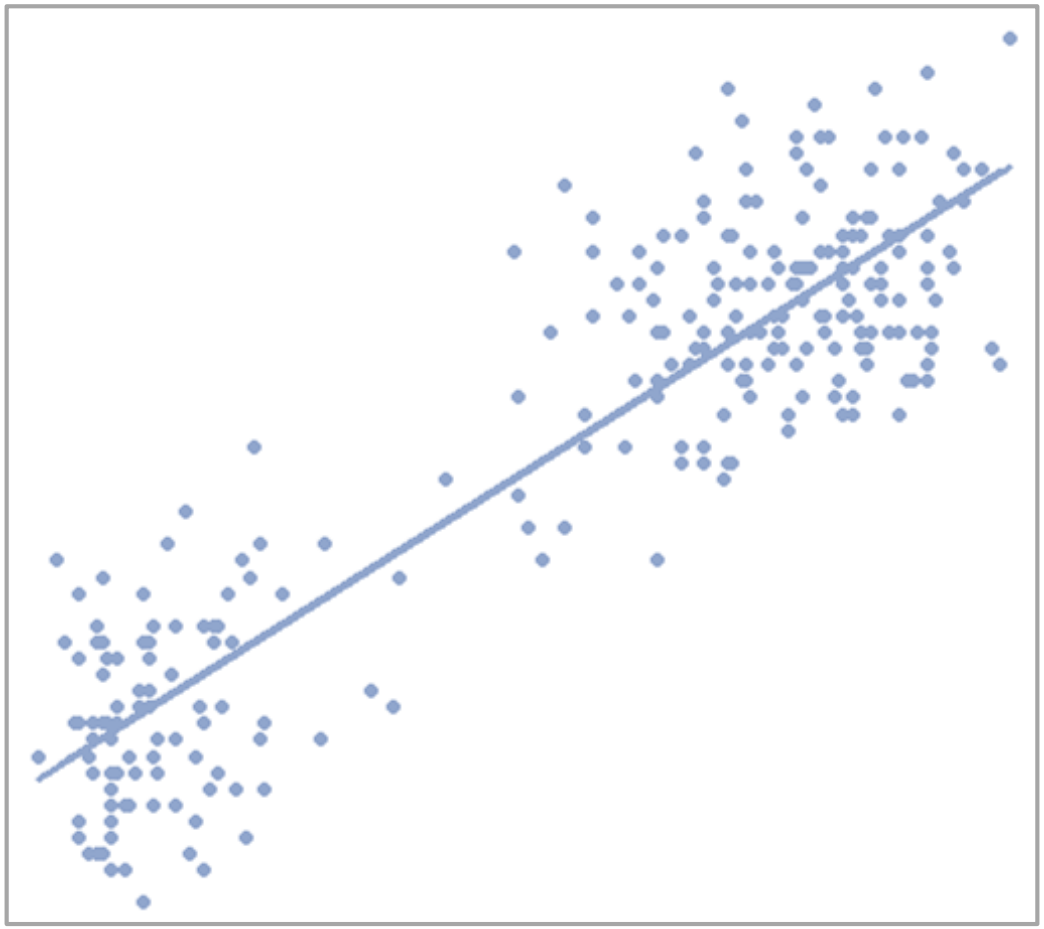
Relationship



Simple Linear Regression

Relationship

Linear model

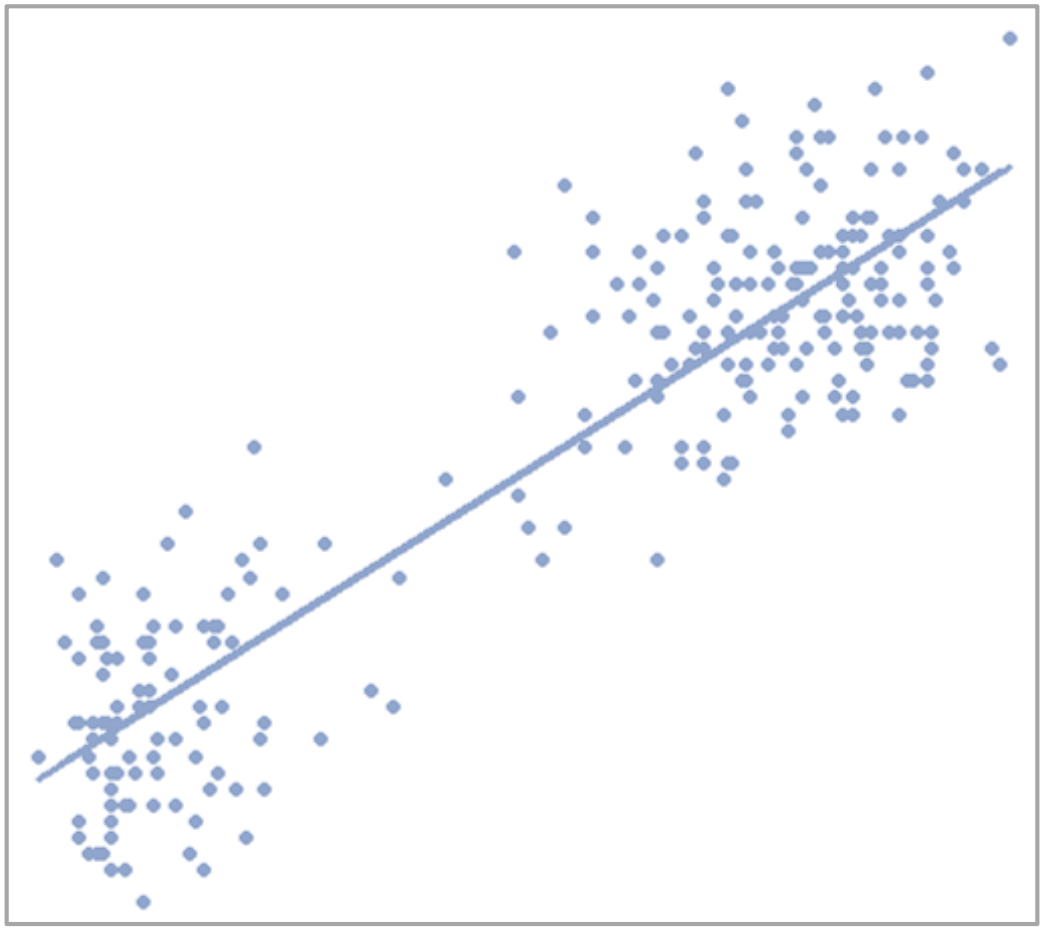


Simple Linear Regression

Relationship

Linear model

Explanatory variable



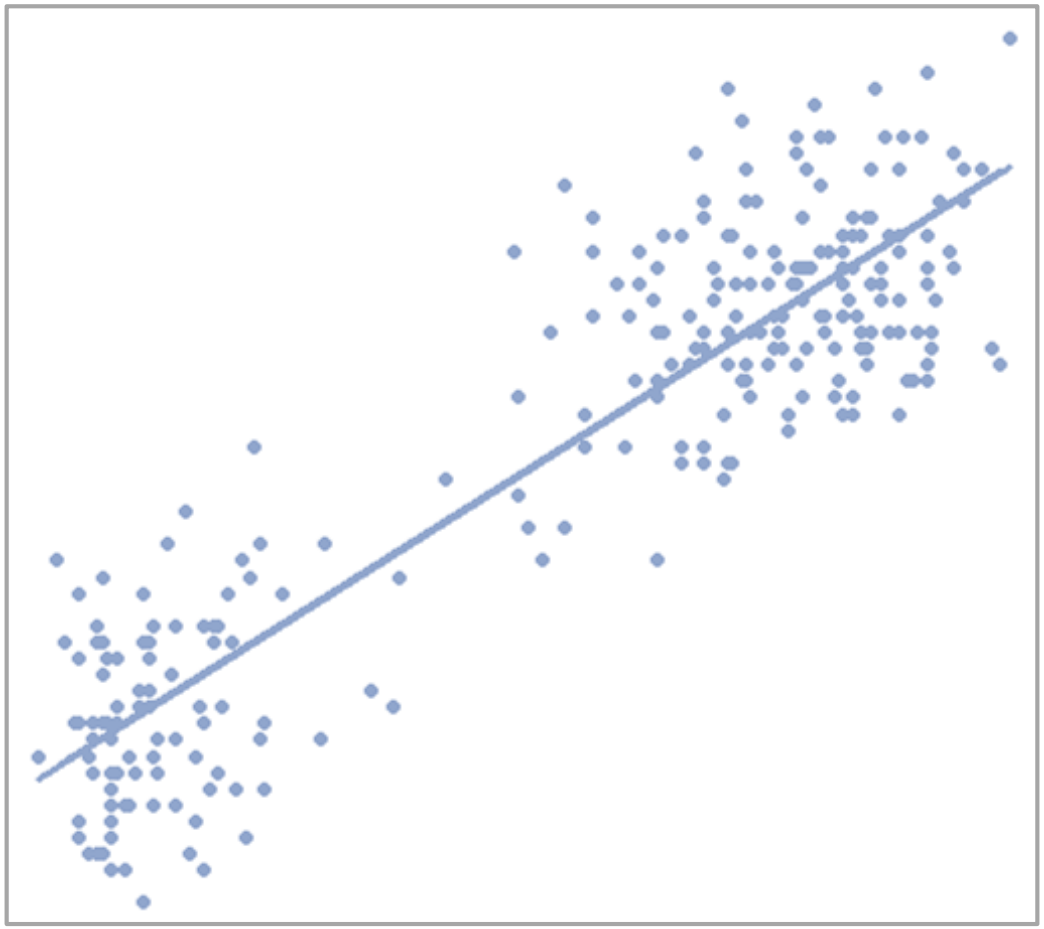
Simple Linear Regression

Relationship

Linear model

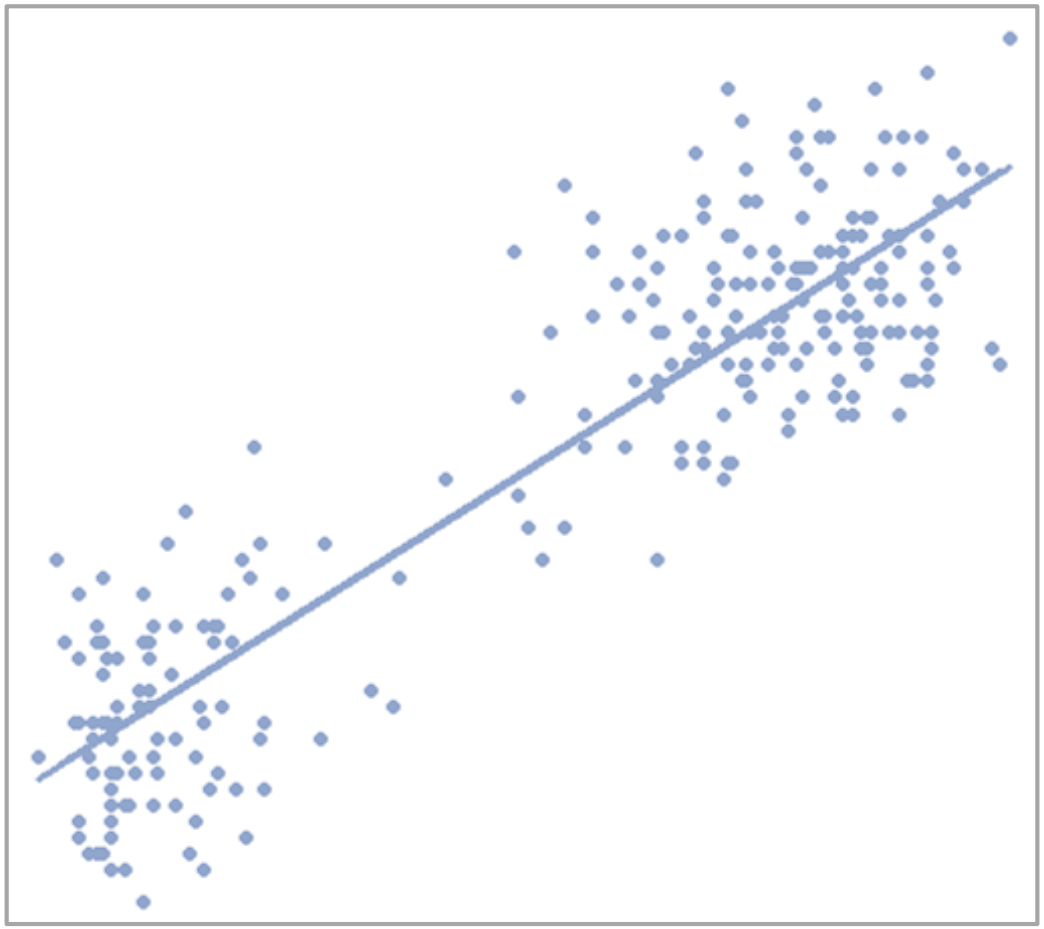
Explanatory variable

Outcome variable



Simple Linear Regression

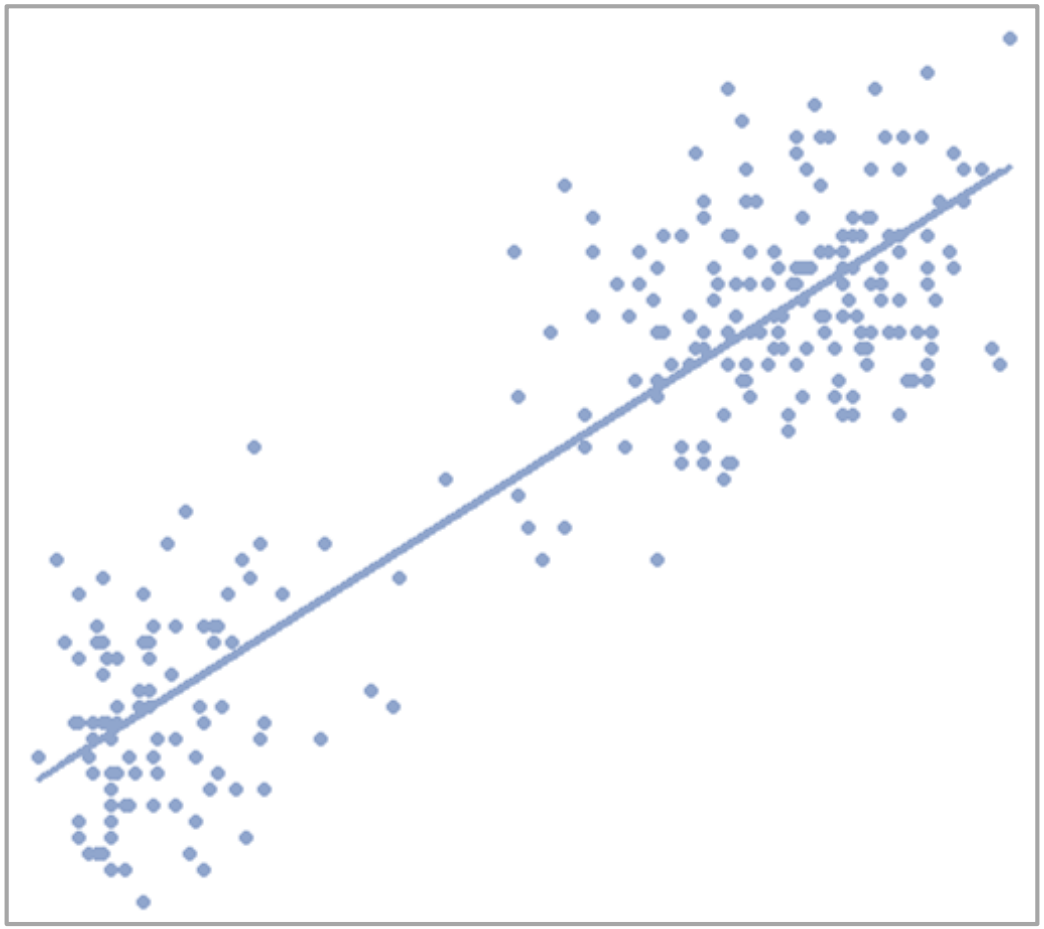
Linear predictor function



Simple Linear Regression

Linear predictor function

$y = m \cdot x + b$

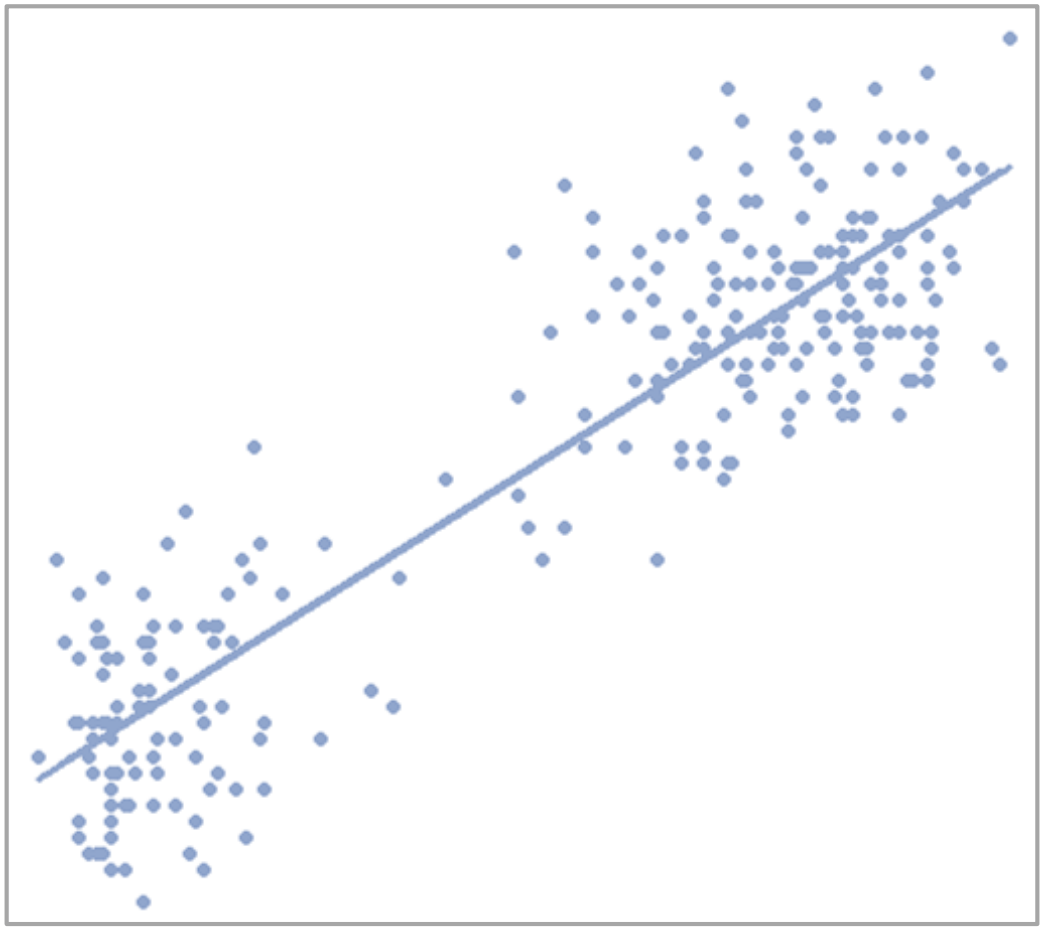


Simple Linear Regression

Linear predictor function

$y = m \cdot x + b$

Parameters estimated



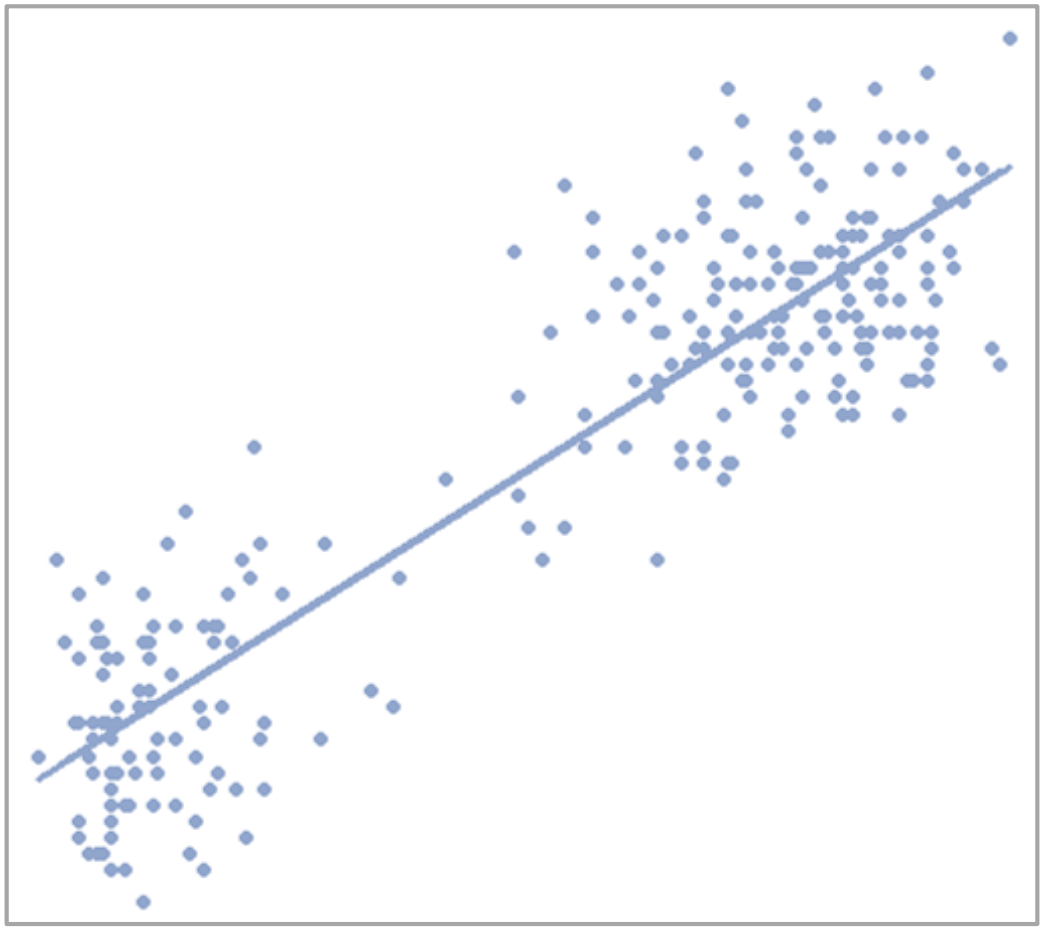
Simple Linear Regression

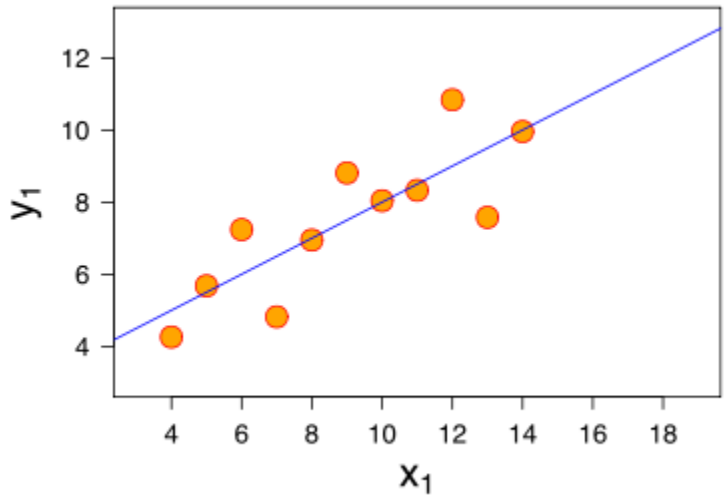
Linear predictor function

$y = m \cdot x + b$

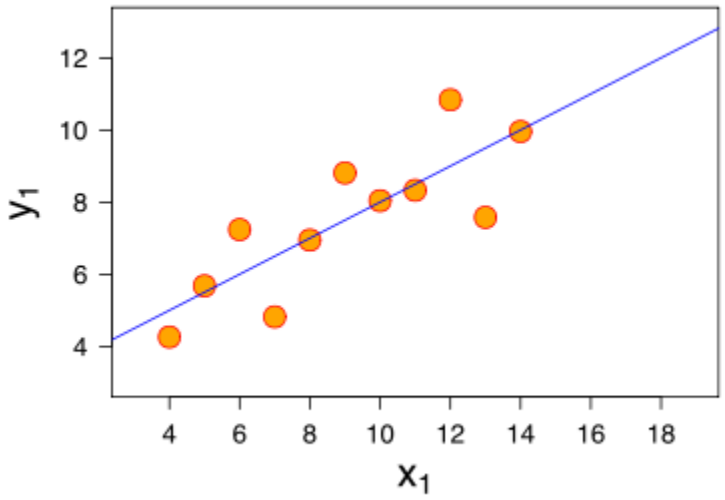
Parameters estimated

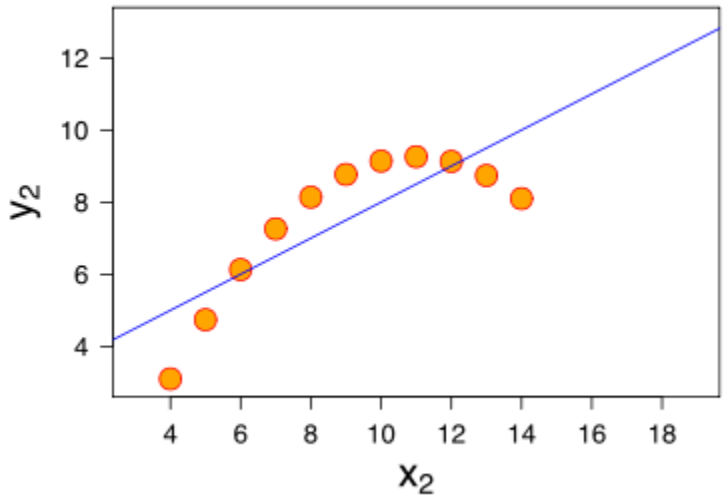
Relies on assumptions



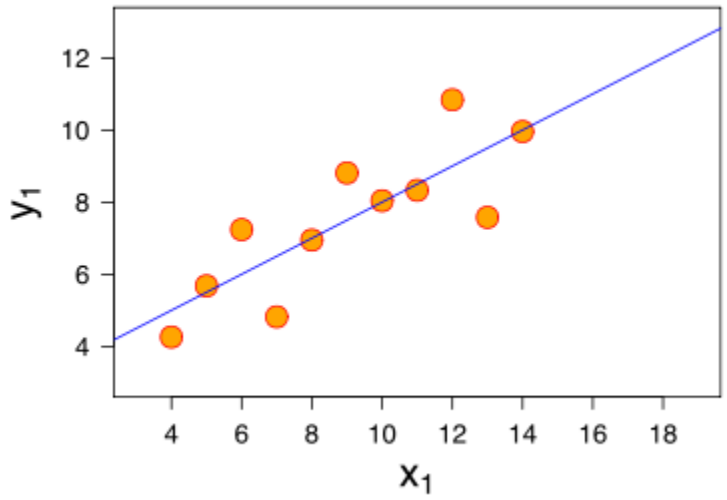


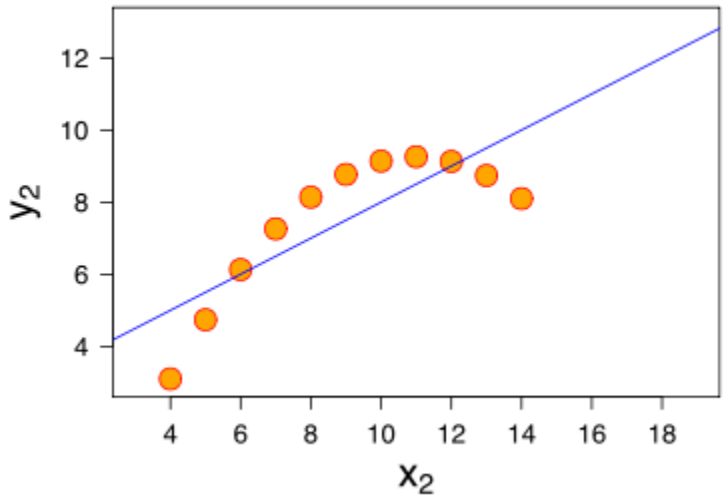
Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

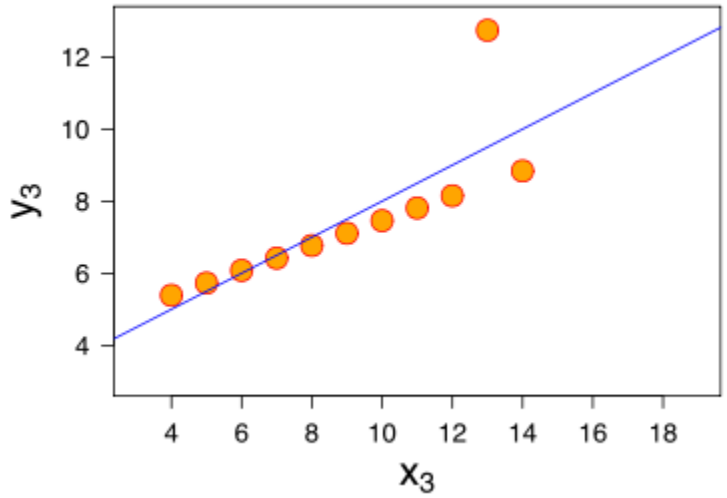




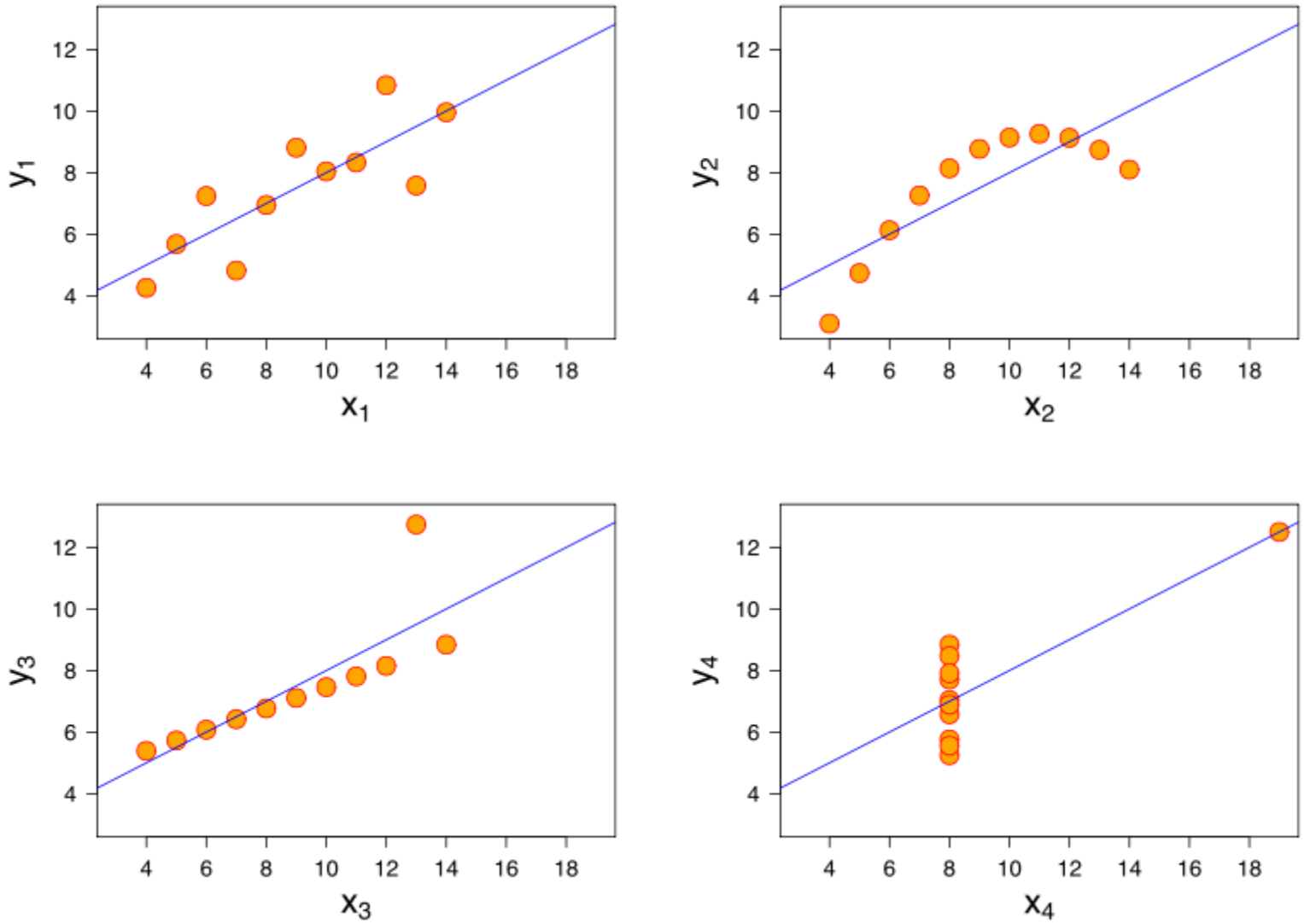
Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet







Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet



Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

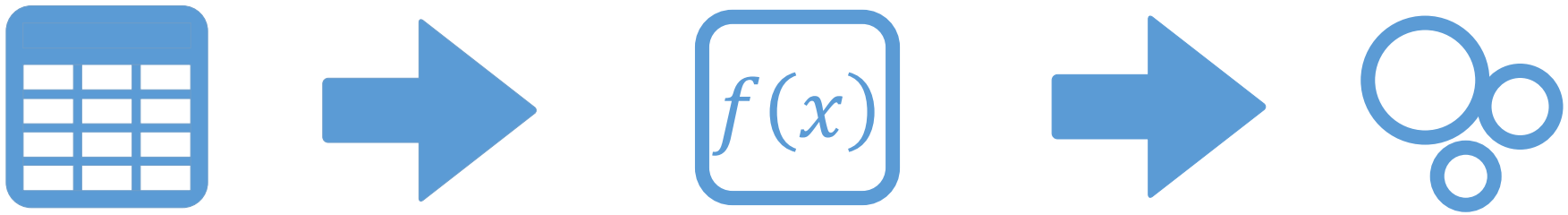
Regression Demo

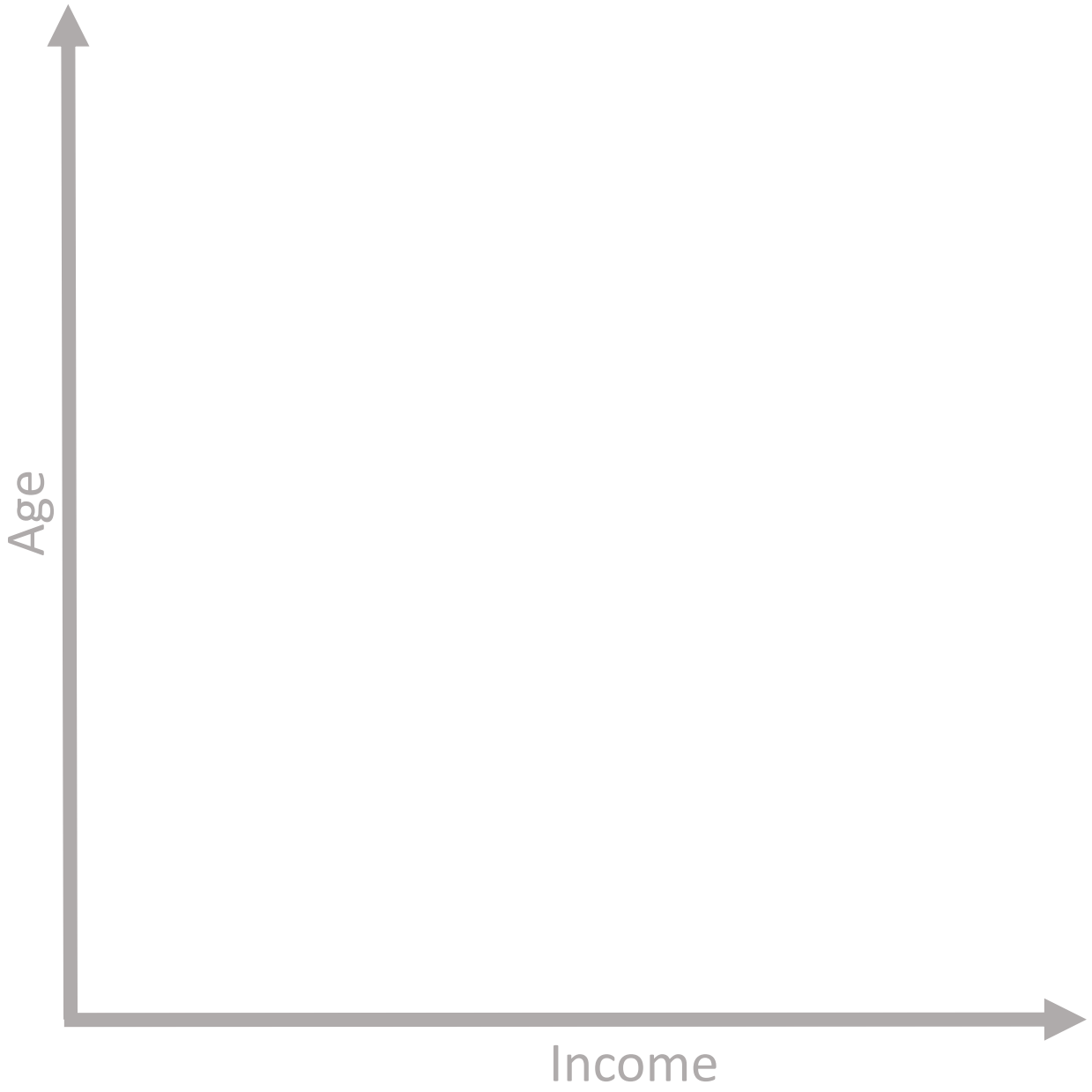
Goal: Predict petal width
based on petal length

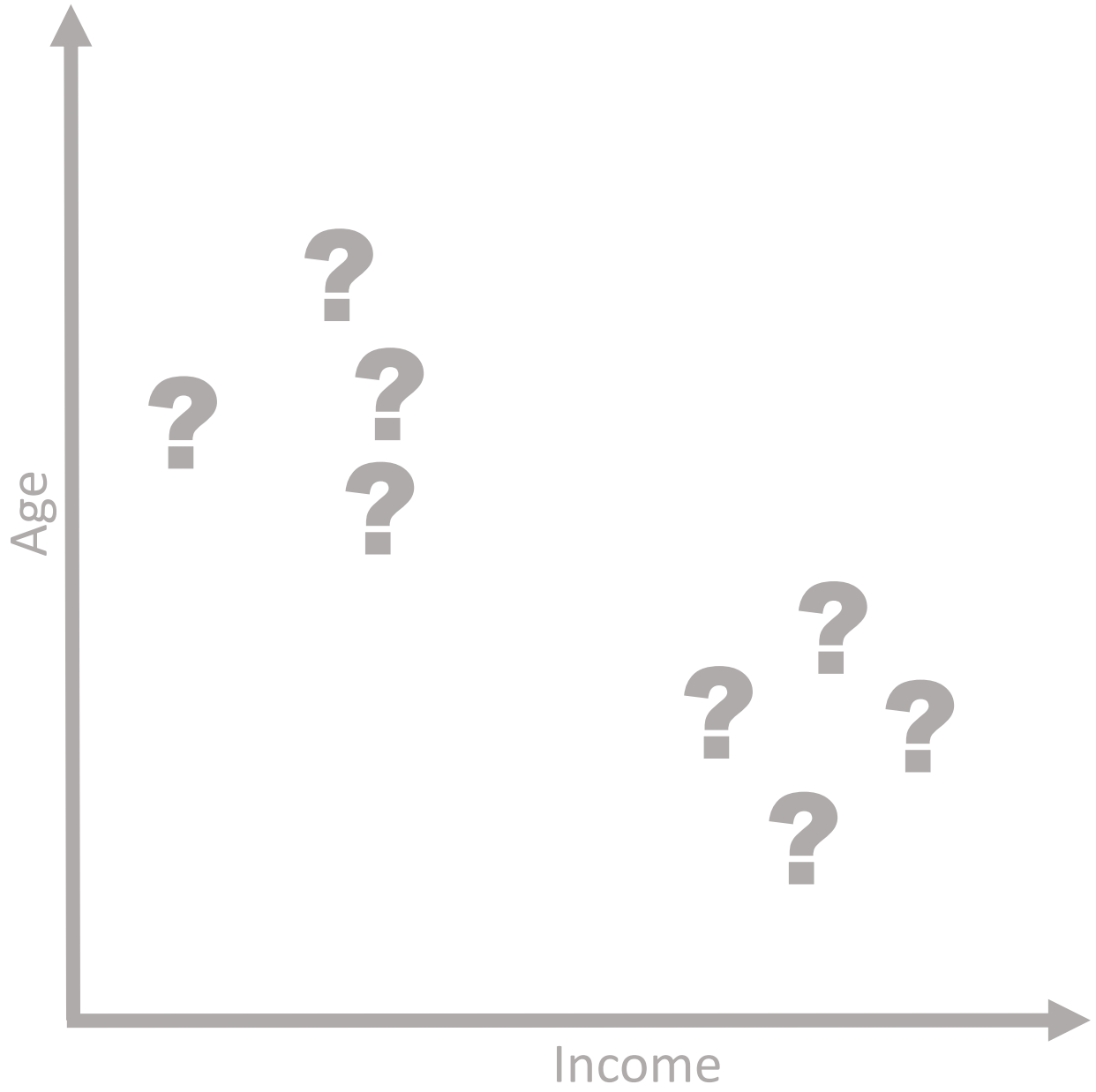
Real-World Examples

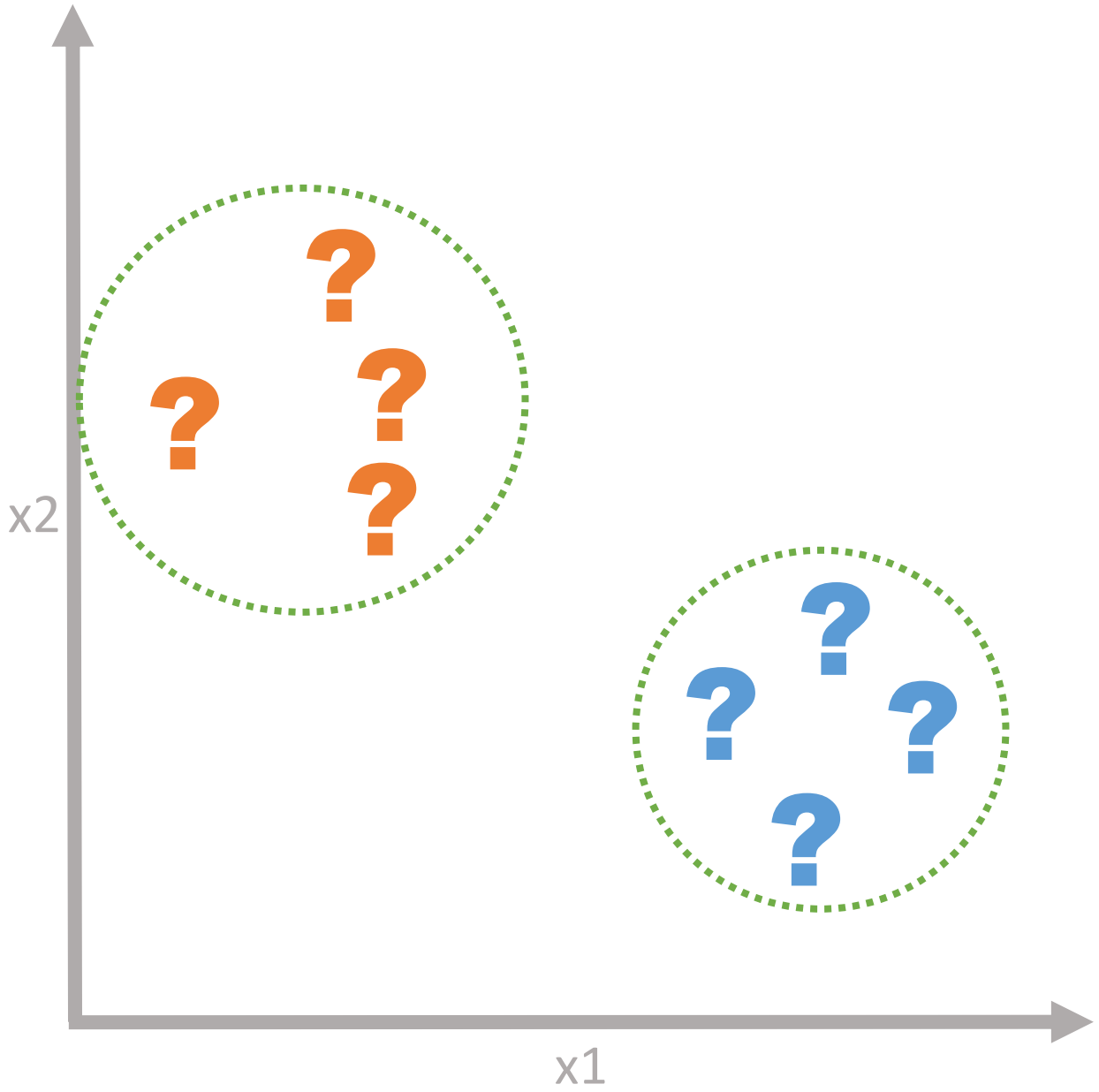
- How much profit will we make?
- What will the price be tomorrow?
- How many will this person buy?
- How long until this part fails?

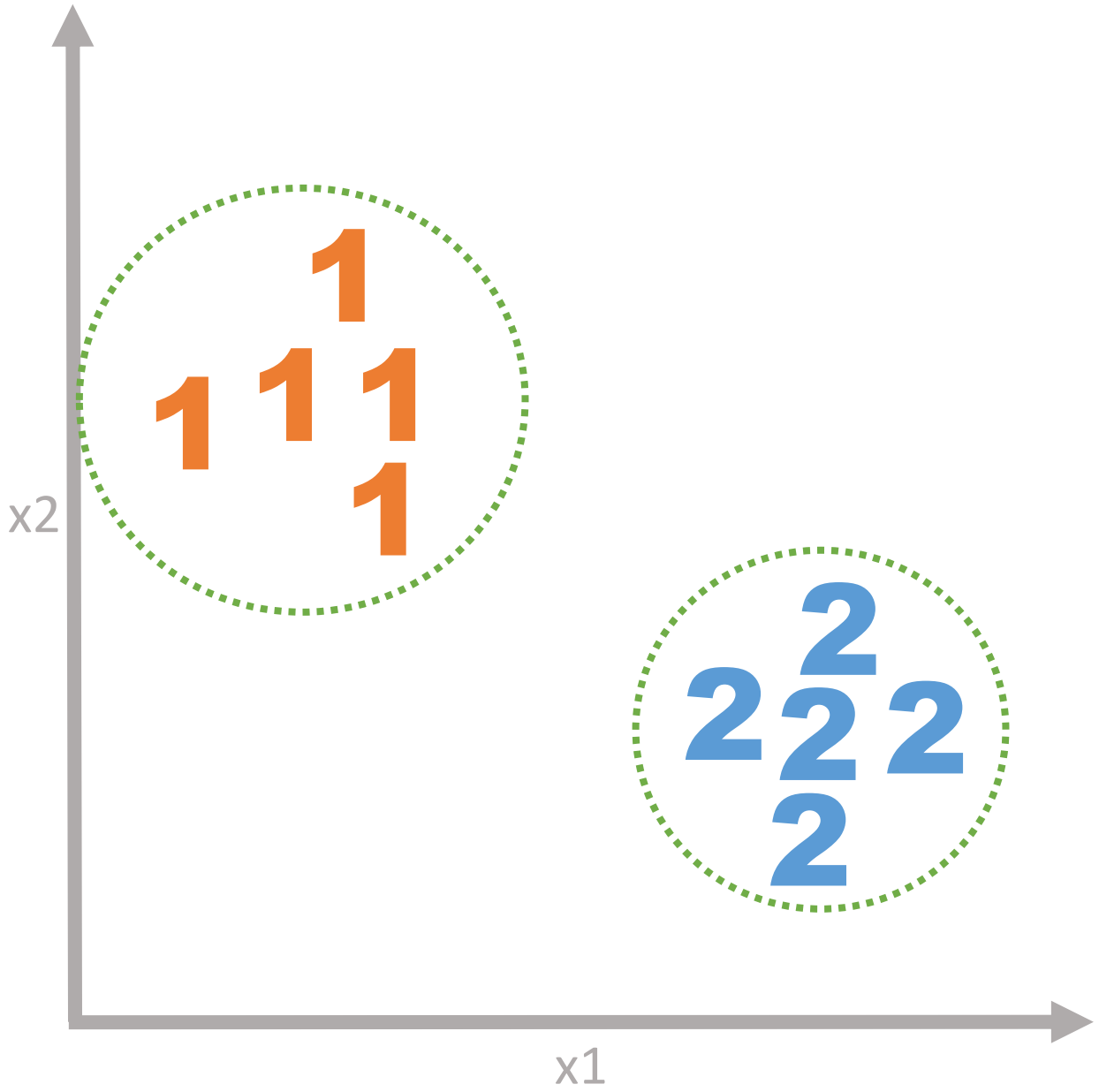
Clustering


f(x)


Age
Income


Age
Income


x2
x1

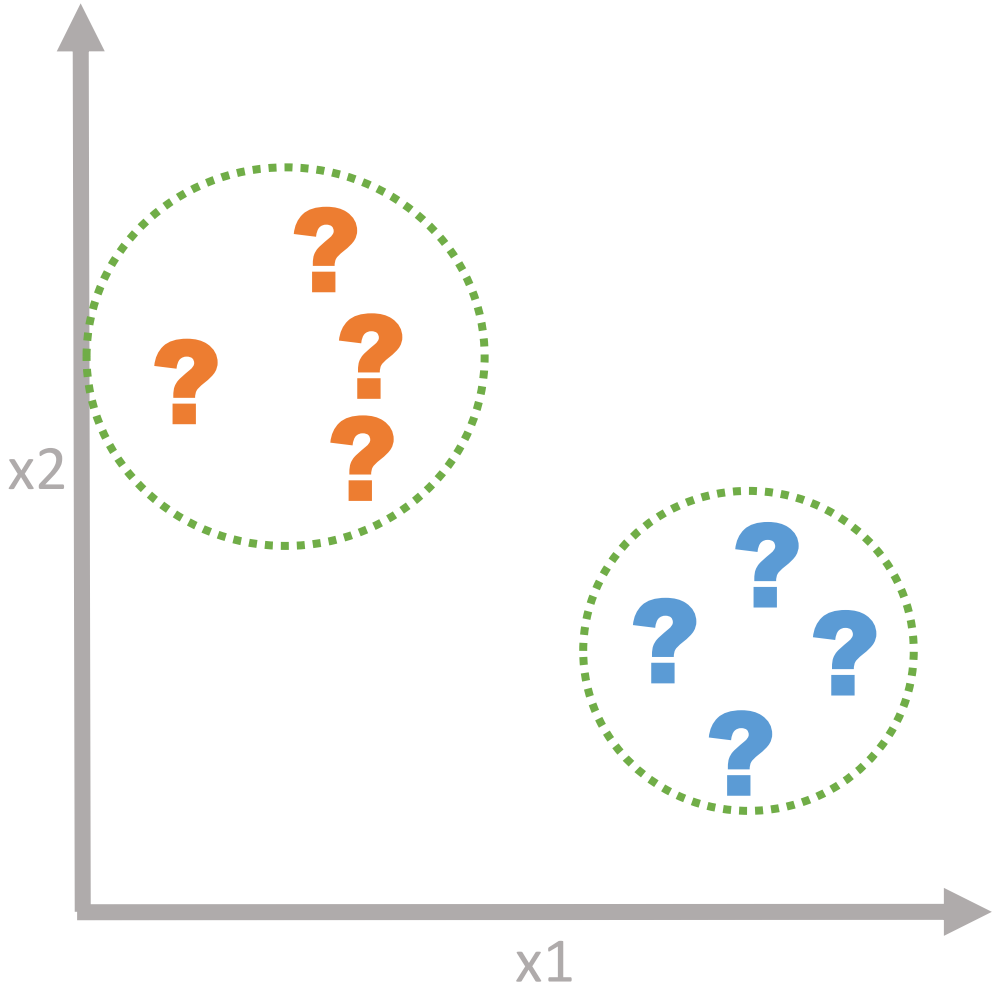

x2
1
1 1 1
1
2
2 2 2
2
x1

Clustering Algorithms

K-means

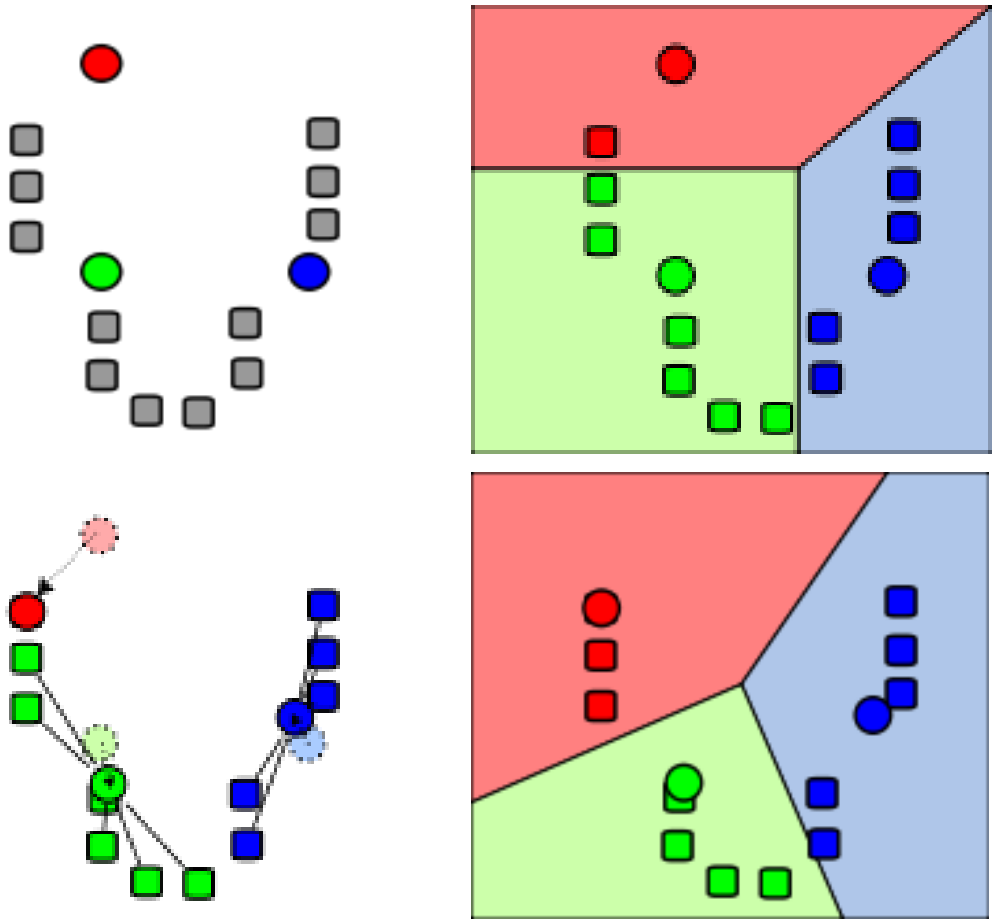
Hierarchical clustering

Expectation maximization



k-Means Clustering

Unsupervised learning

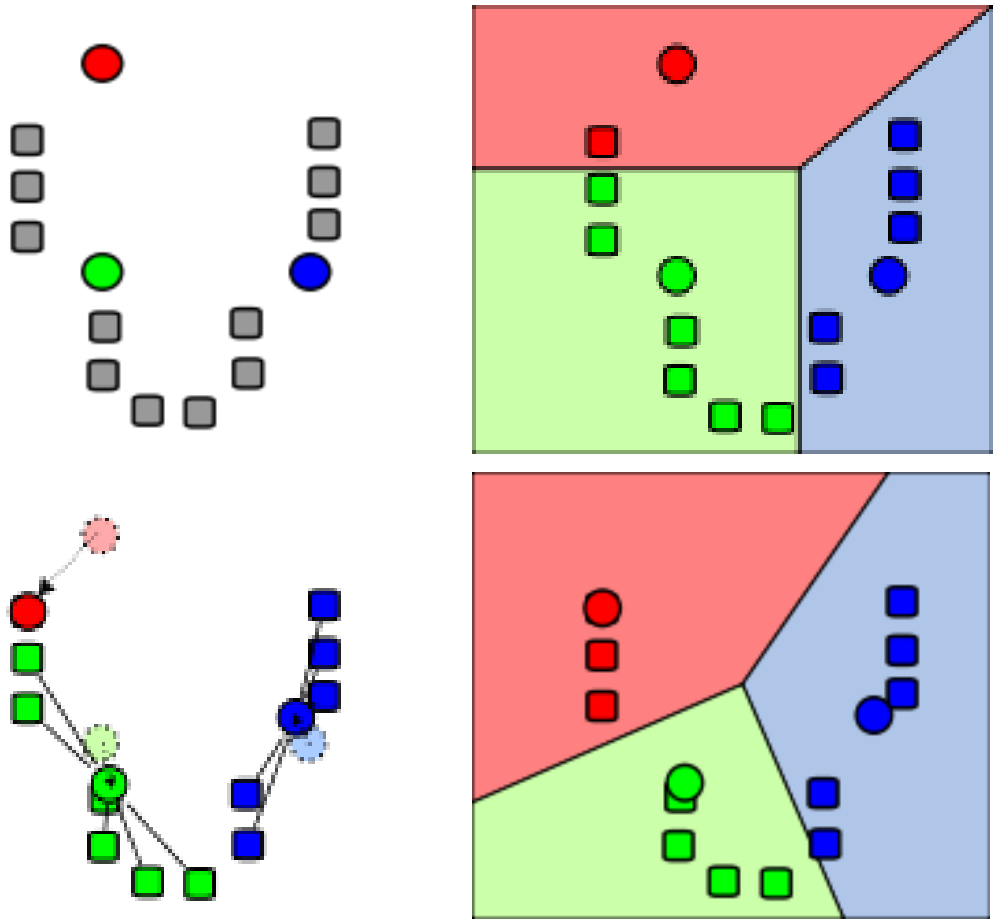


Source: Wikipedia

k-Means Clustering

Unsupervised learning

Specify k (# of clusters)



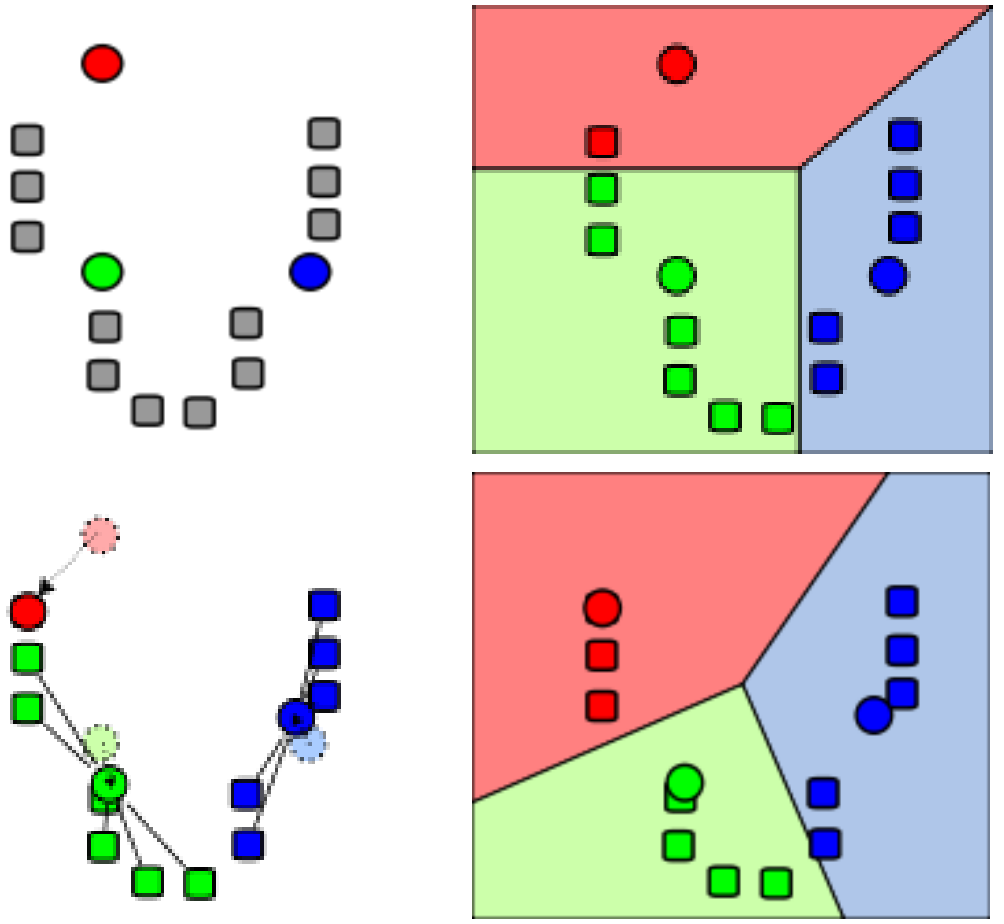
Source: Wikipedia

k-Means Clustering

Unsupervised learning

Specify k (# of clusters)

Algorithm finds centers



Source: Wikipedia

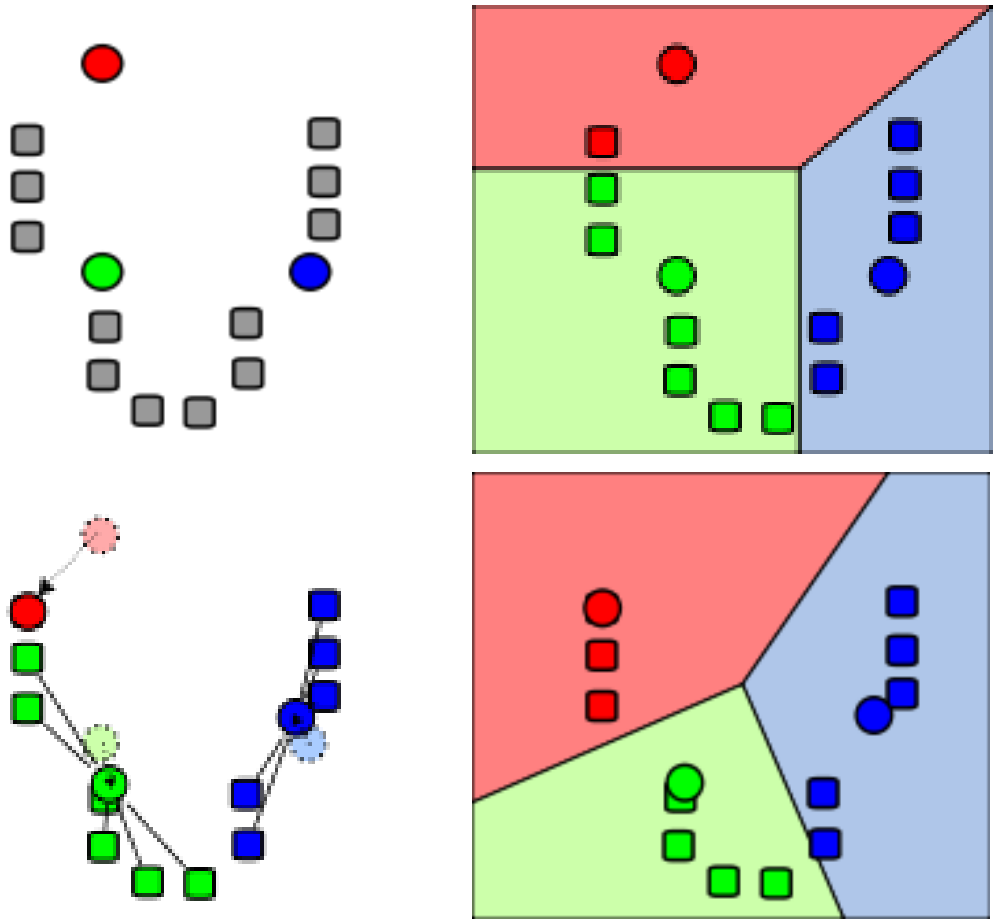
k-Means Clustering

Unsupervised learning

Specify k (# of clusters)

Algorithm finds centers

Random restarts



Source: Wikipedia

Clustering Demo

Real-world Examples

- Market segmentation
- Document classification
- Recommendation systems
- Market basket analysis

Beyond the Basics


This is just the tip of the iceberg!

Robust Models

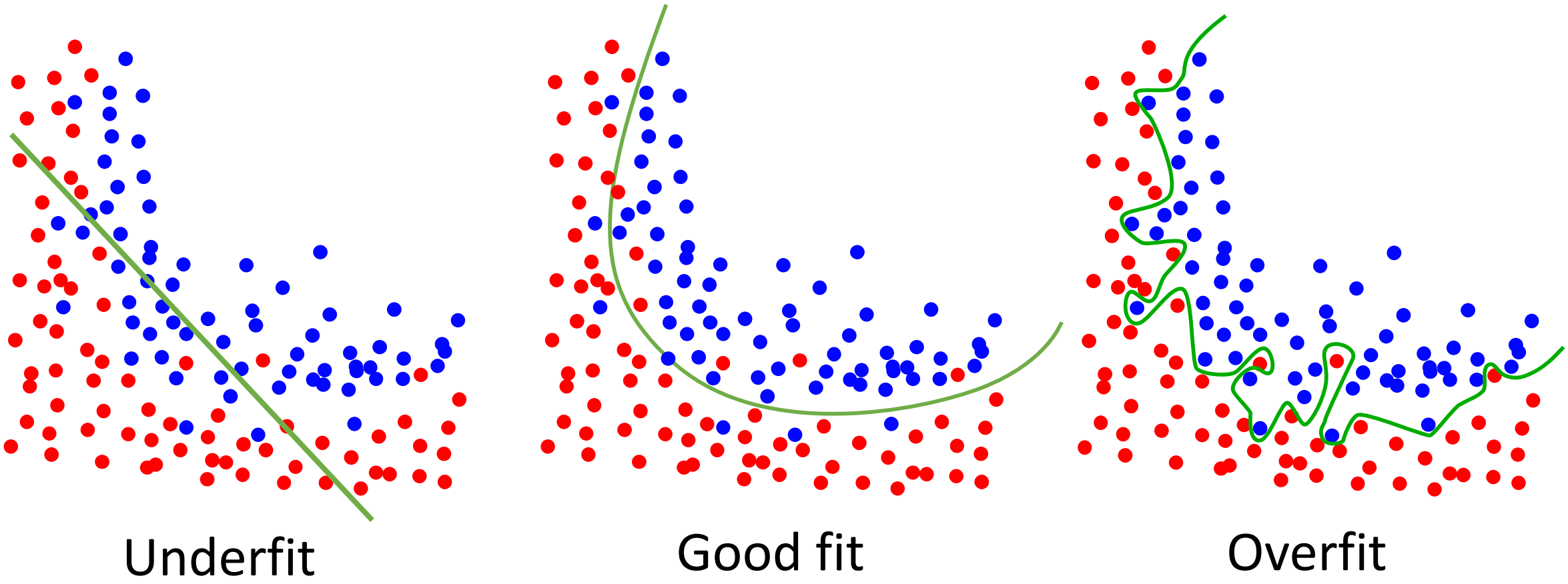
Cleaning and Transforming Data

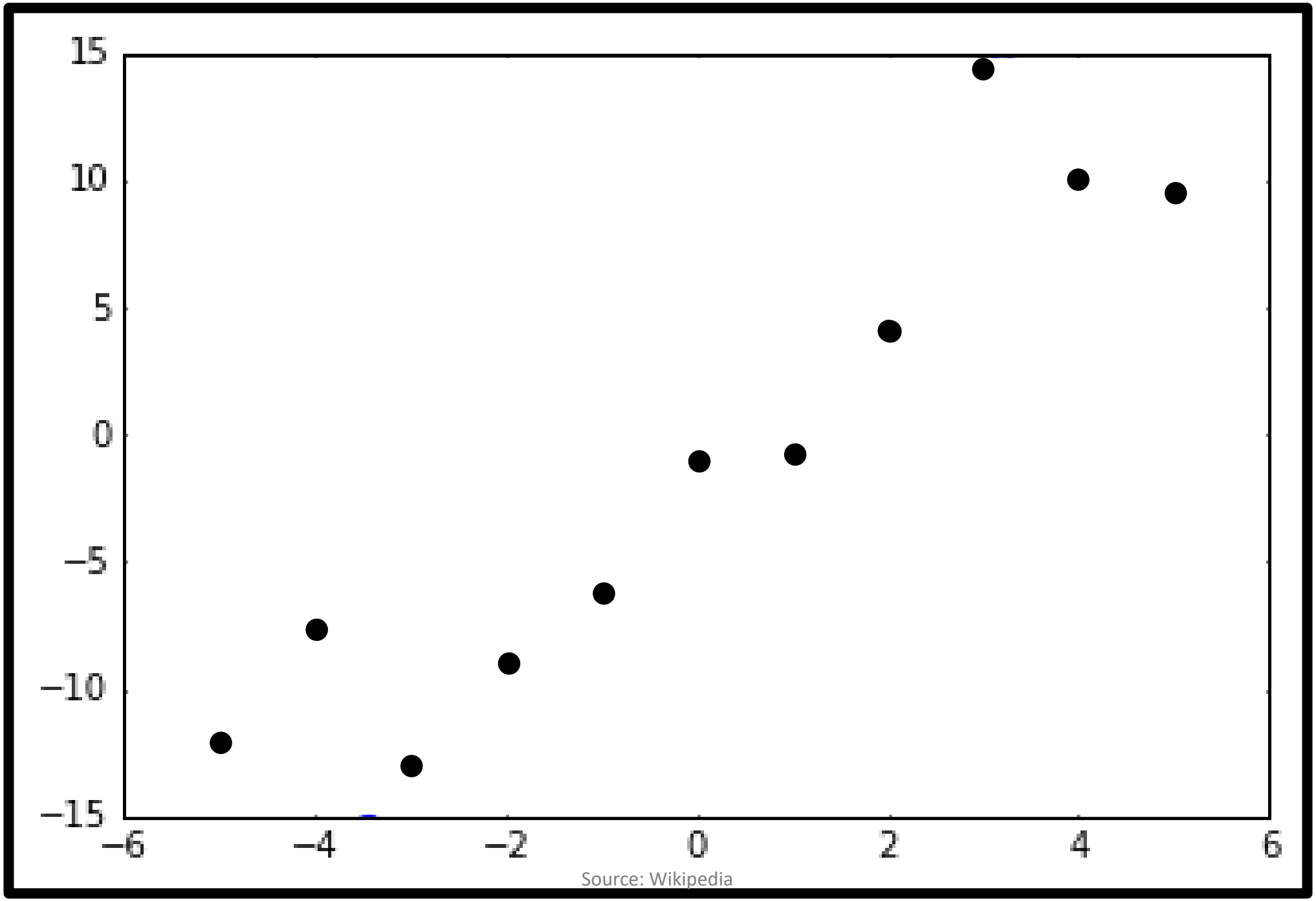
Data are messy

80% of work

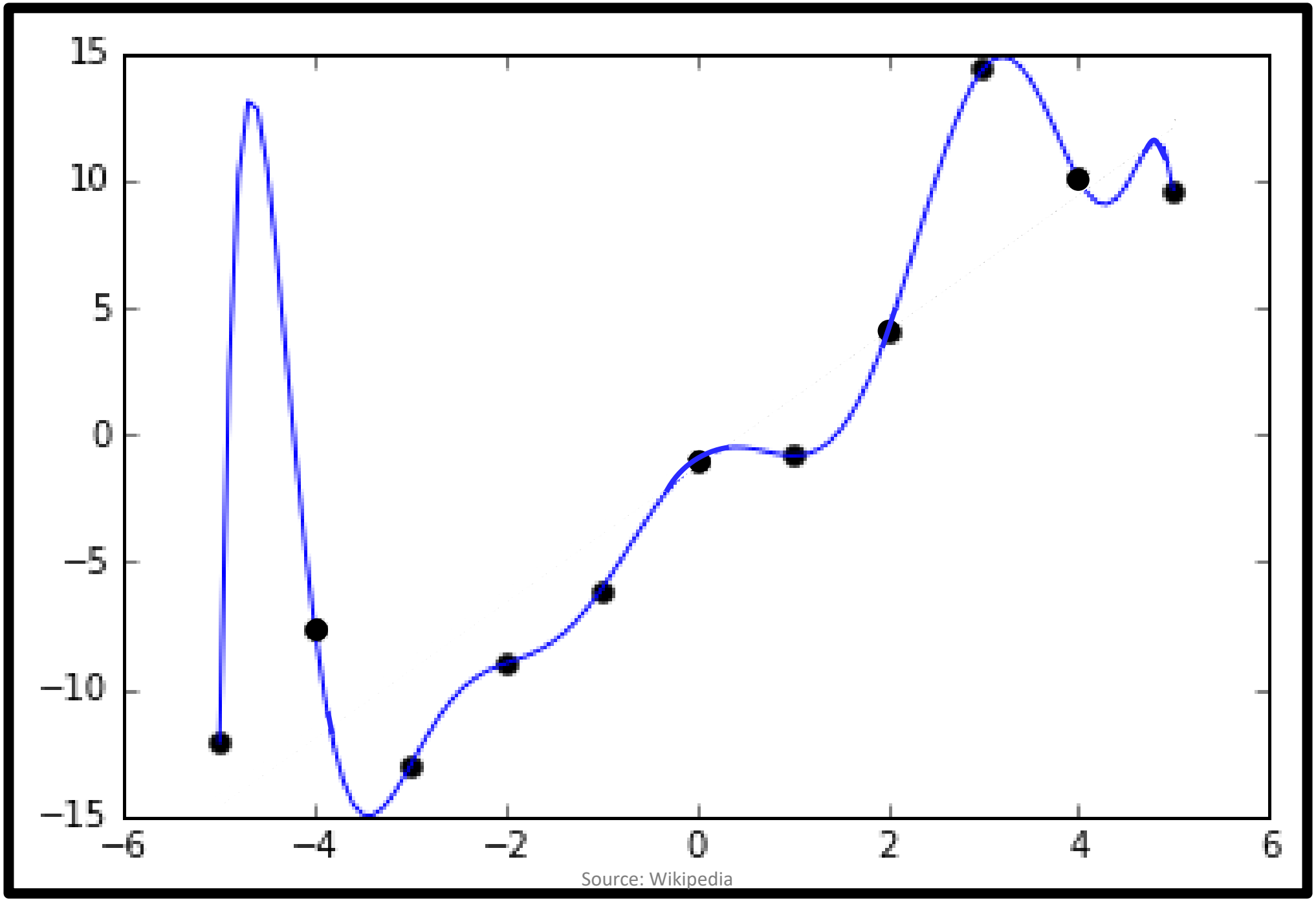
R helps a lot

Record all steps

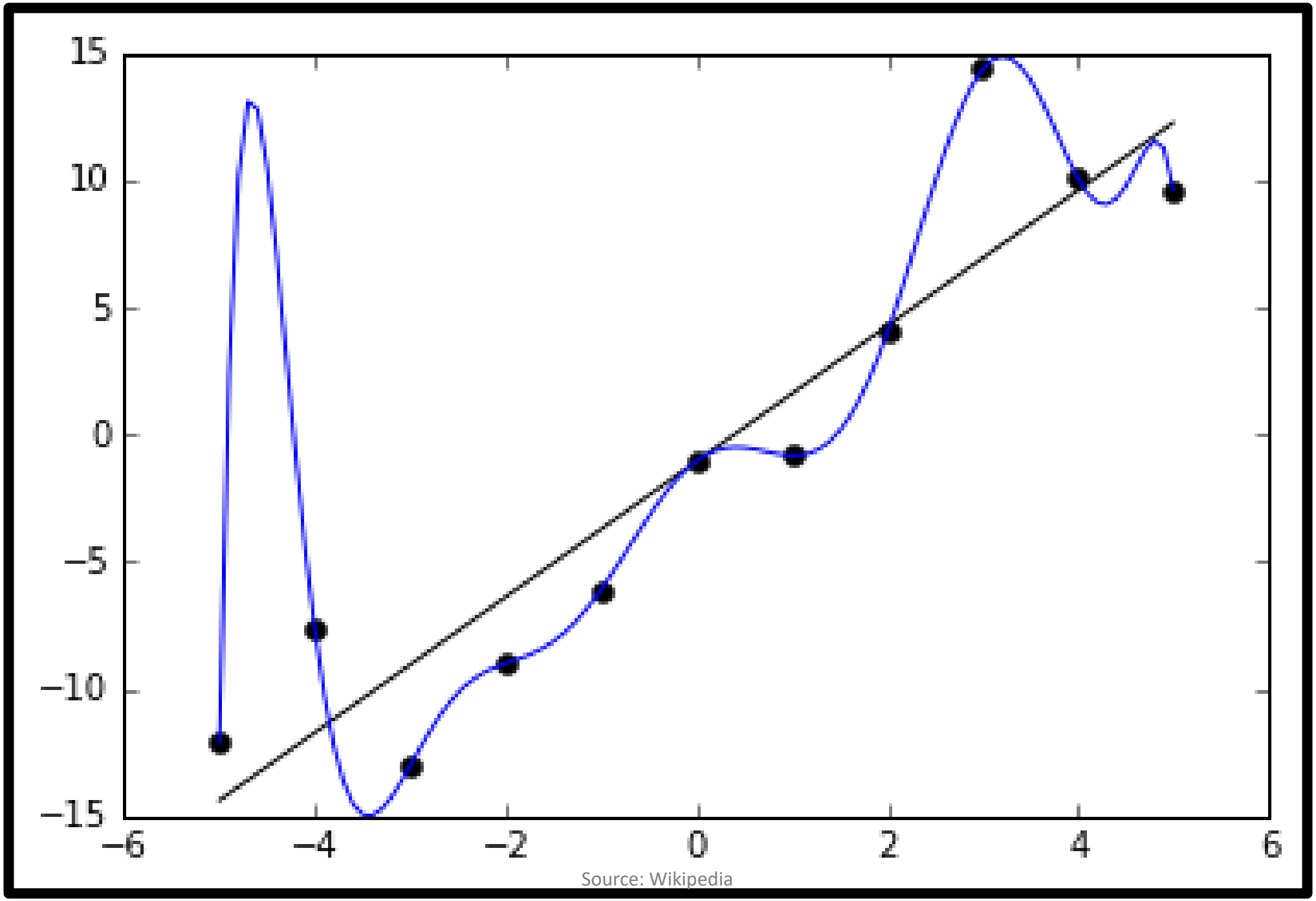

Underfit
Good fit
Overfit



Source: Wikipedia



Source: Wikipedia



Source: Wikipedia

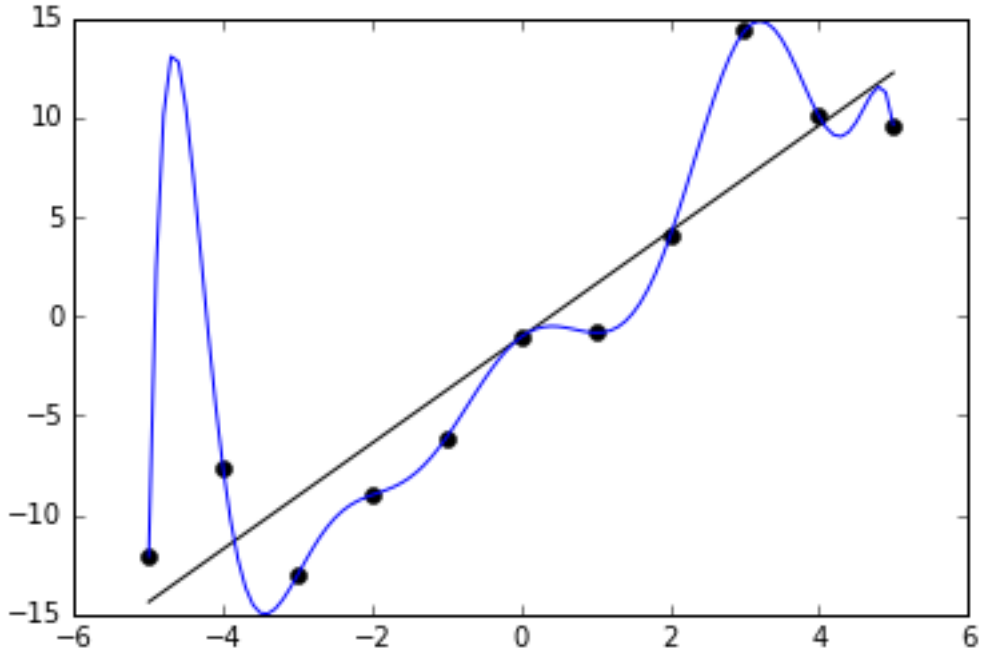
Regularization Techniques

Early stopping

Pruning (trees)

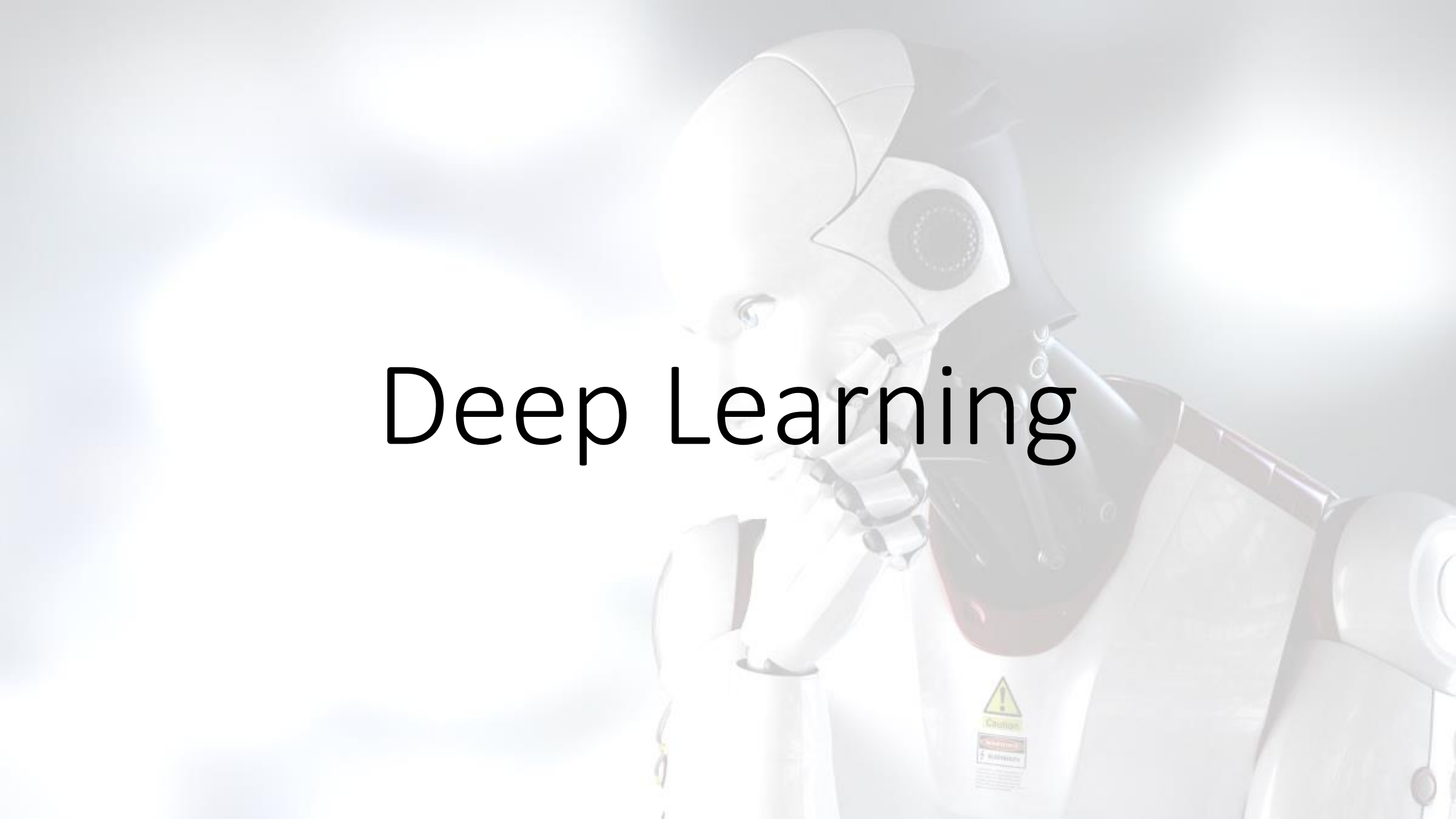
Adding noise

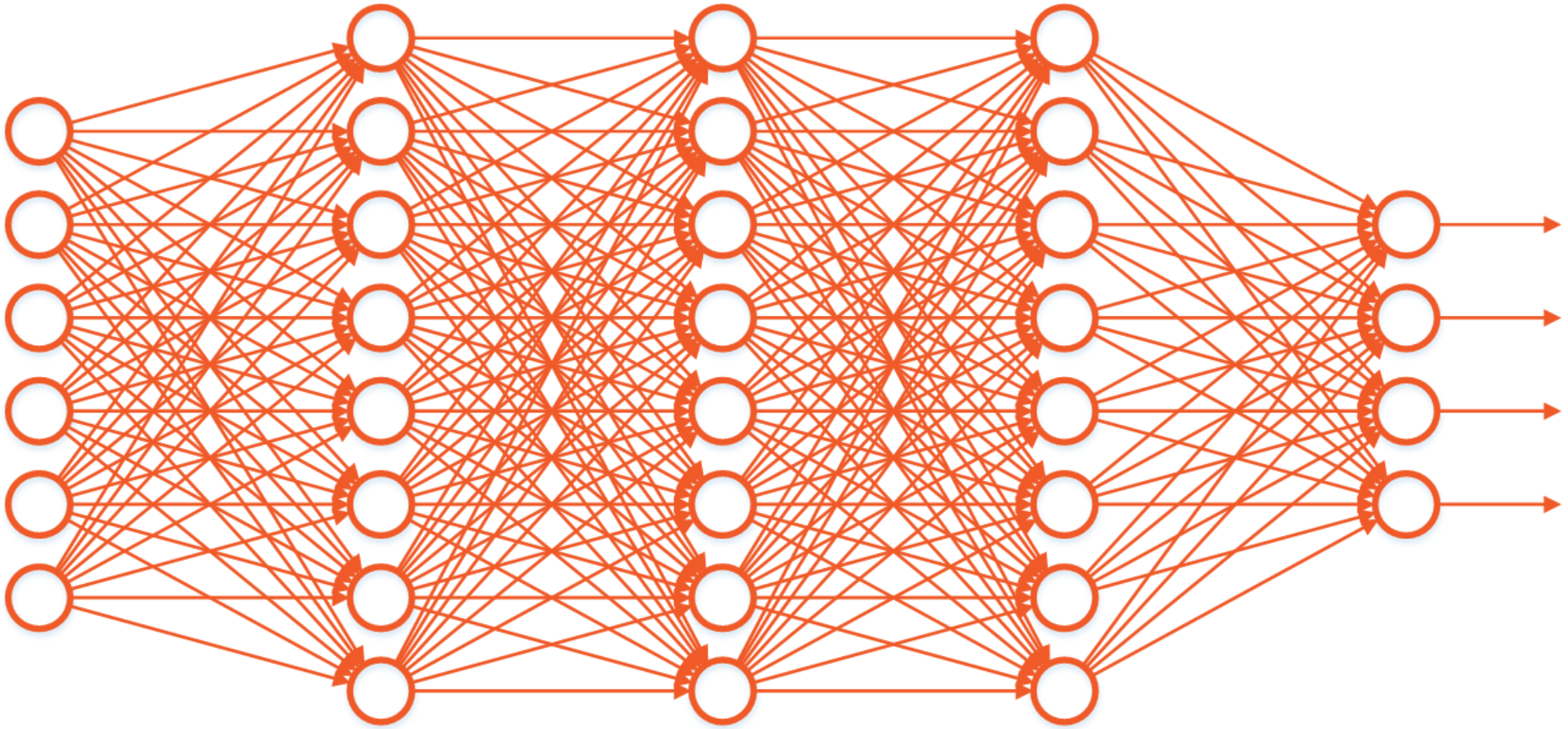
Parameter tuning

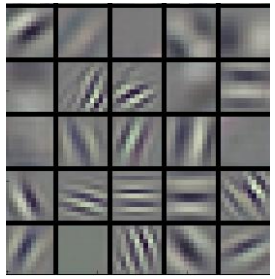


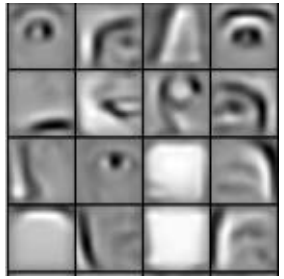
Source: Wikipedia

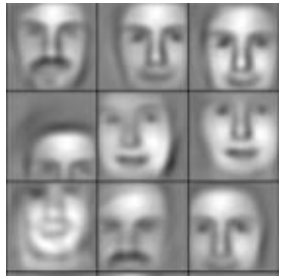
Deep Learning





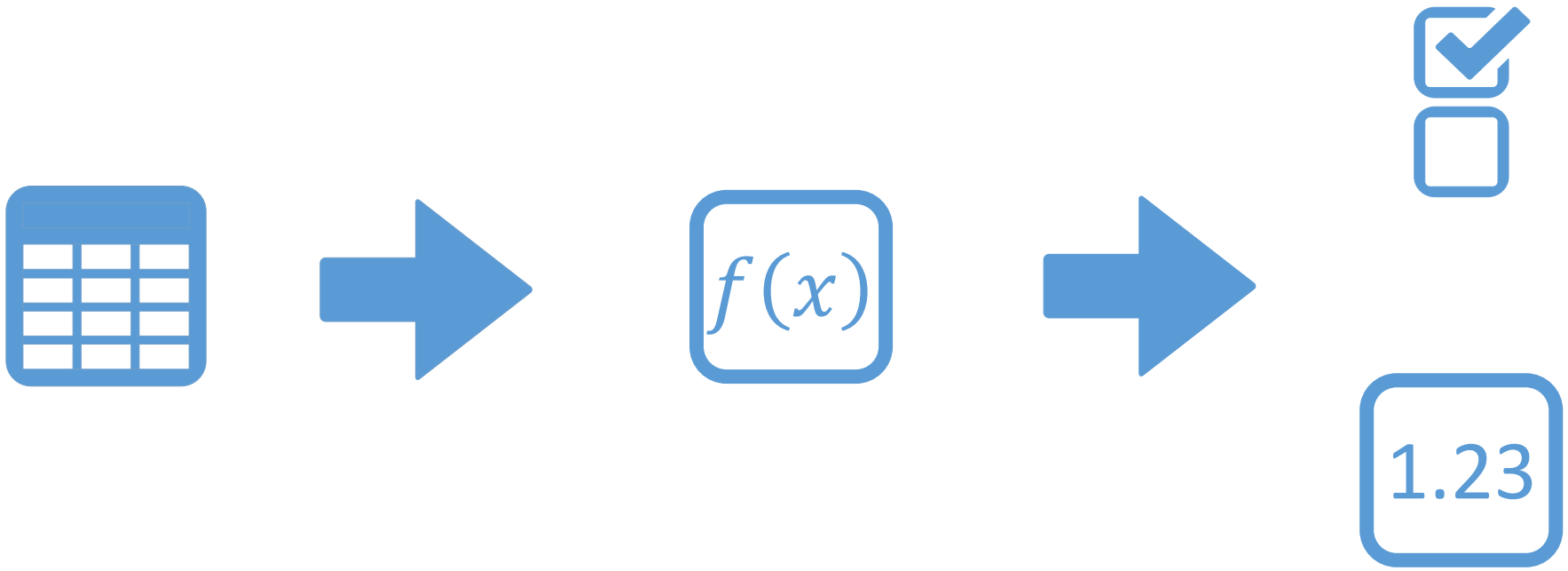


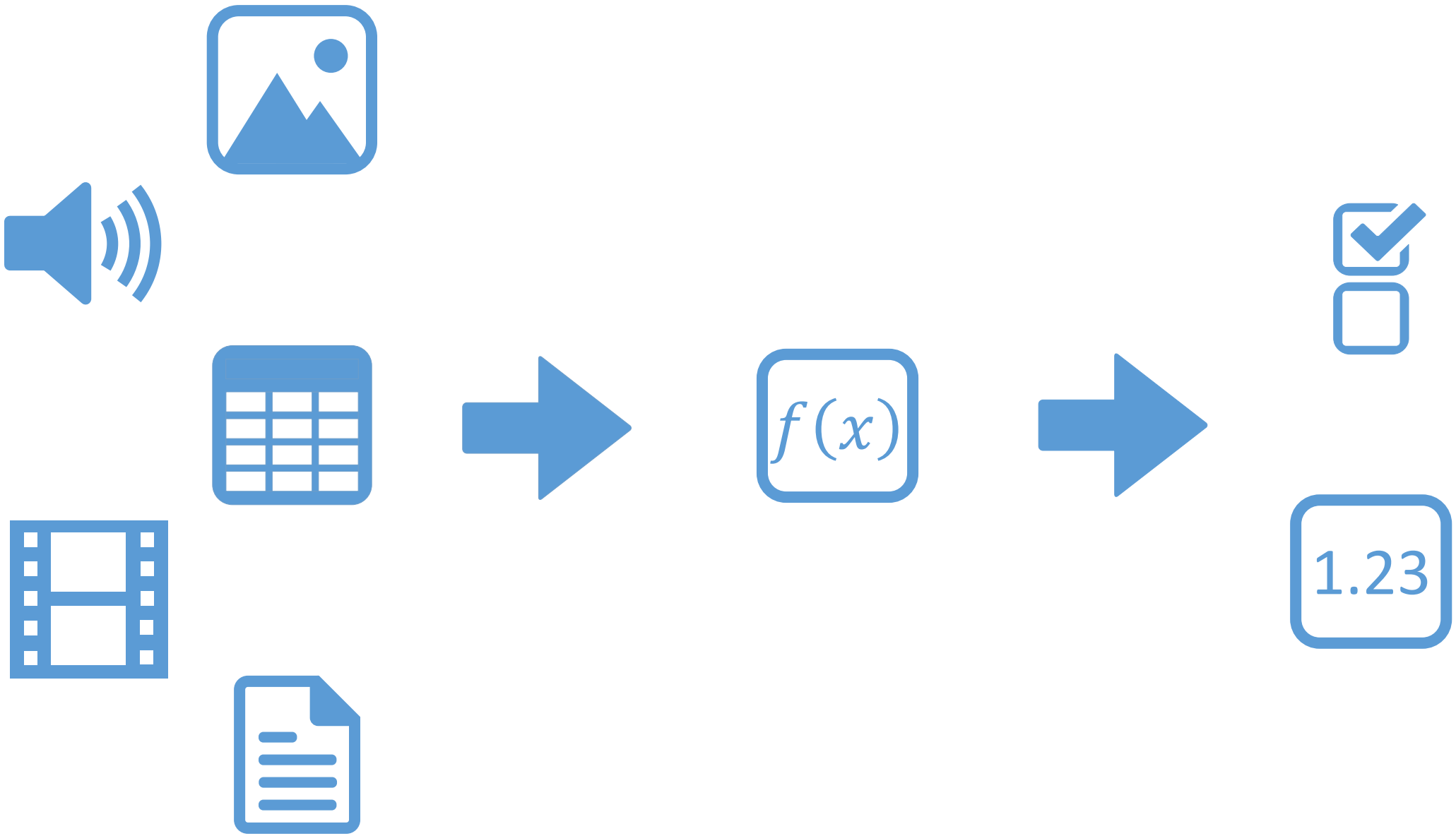


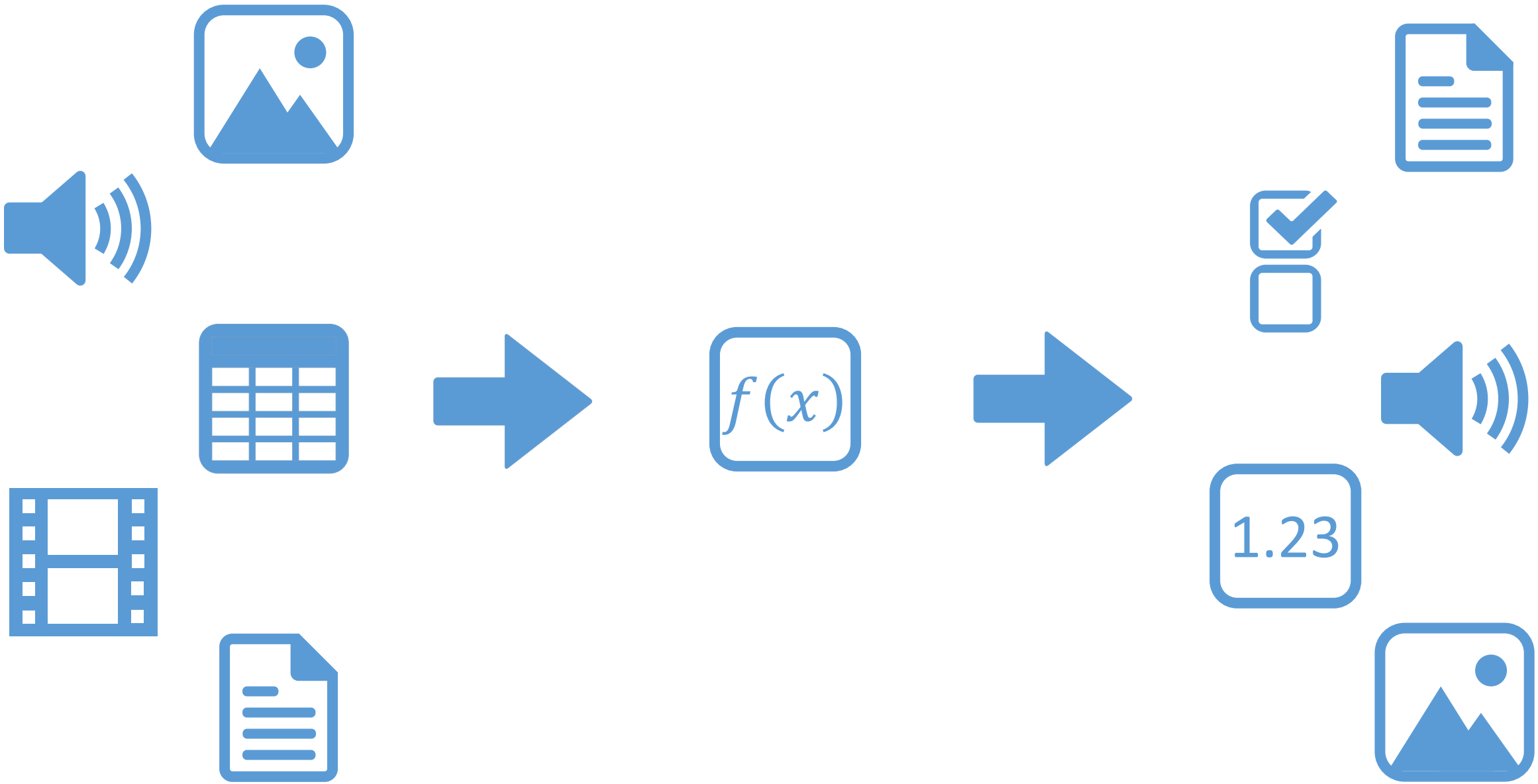



John
Miko
Jane
Lee


Caution
WARNING
ELECTRICITY


$f(x)$
1.23


$f(x)$
1.23


$f(x)$
1.23


FaceApp

Where to Go Next

Pluralsight: https://www.pluralsight.com

Coursera: https://www.coursera.org

Data Camp: https://www.datacamp.com

Tensorflow: http://playground.tensorflow.org

My Website

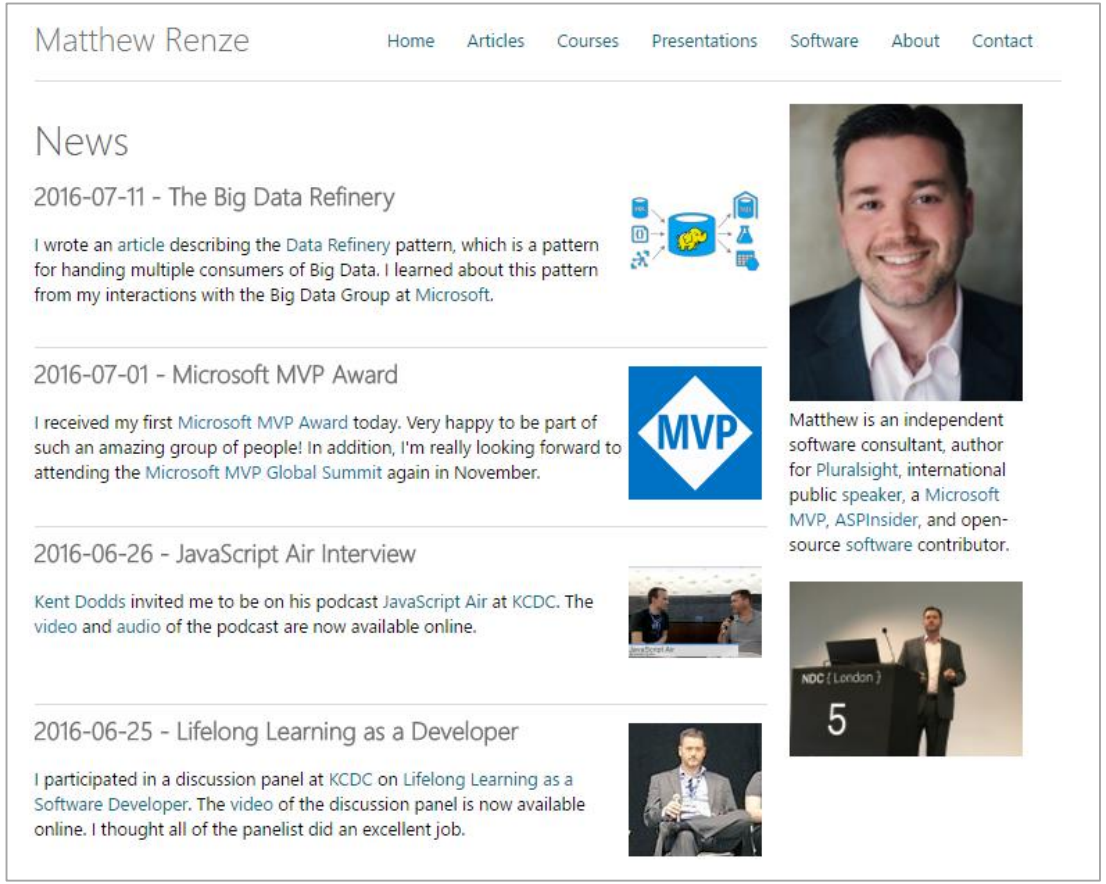
Articles

Presentations

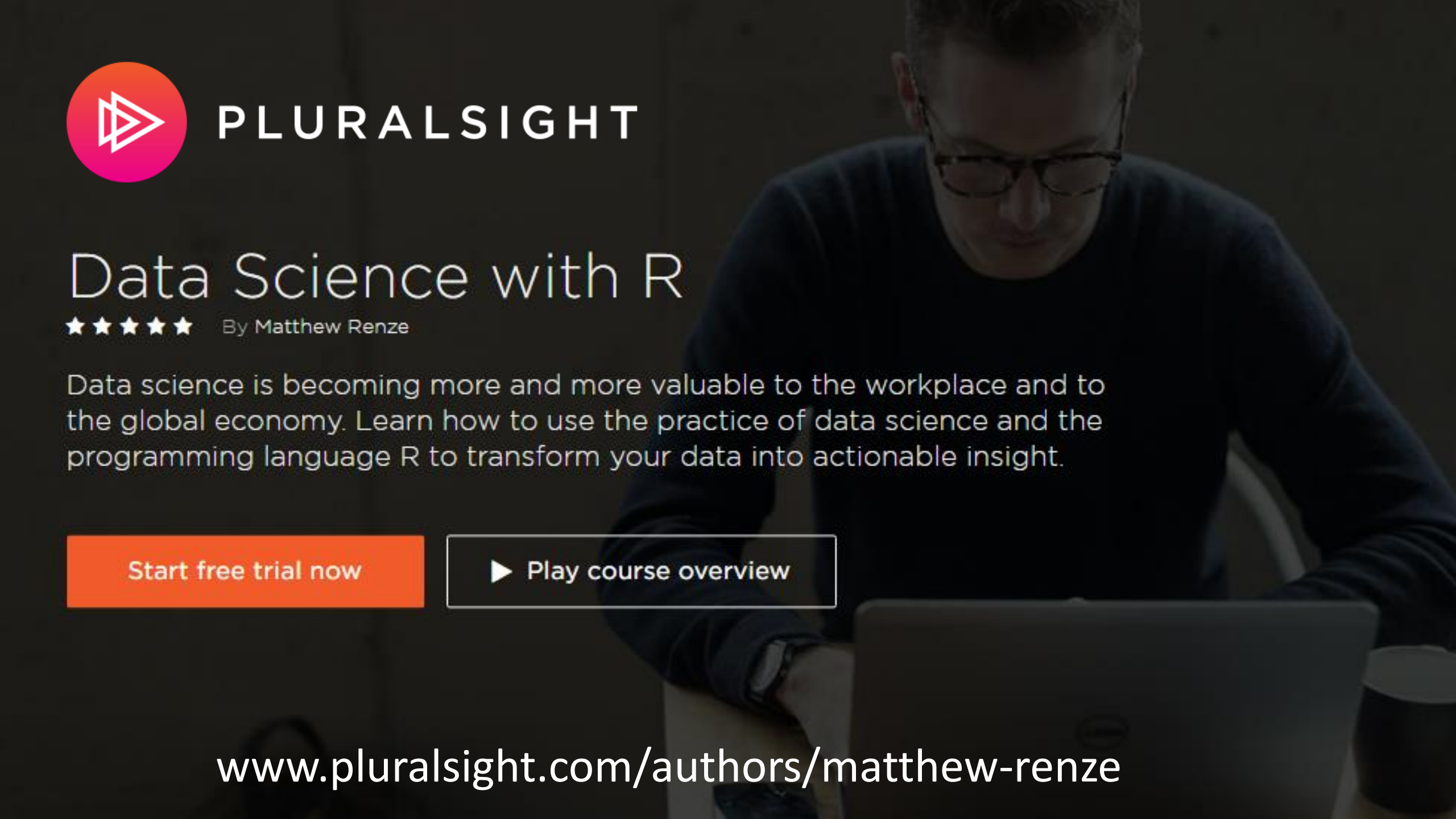
Source Code

Videos

Workshops



www.matthewrenze.com



www.pluralsight.com/authors/matthew-renze

Conclusion

Conclusion

1. Introduction to ML
2. Introduction to R
3. Classification
4. Regression
5. Clustering
6. ML in Practice

Feedback

Very important to me!

What did you like?

What could I improve?

Contact Info

Matthew Renze

Data Science Consultant

Renze Consulting

Twitter: @matthewrenze

Email: matthew@matthewrenze.com

Website: www.matthewrenze.com

Thank You! :)