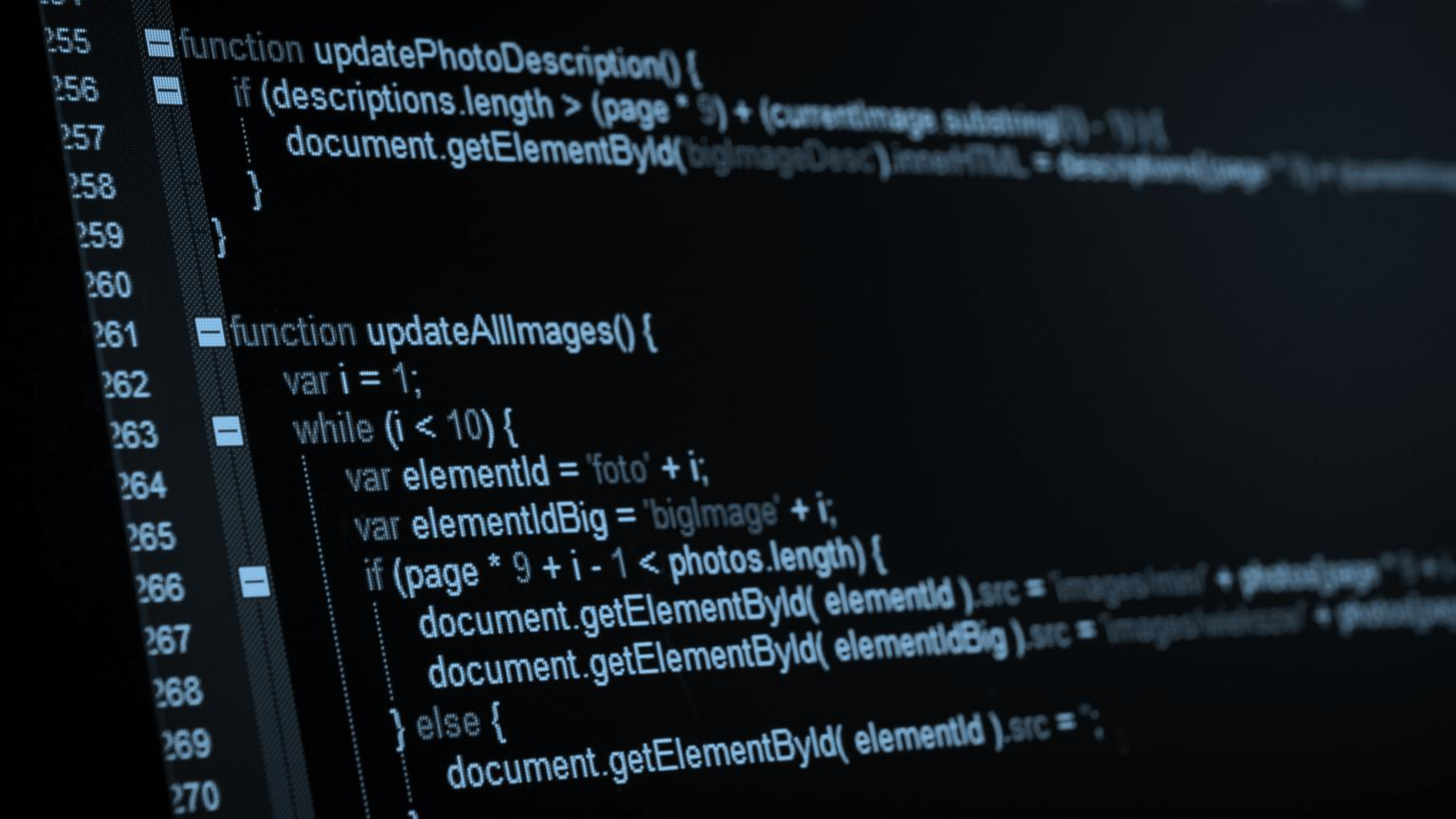
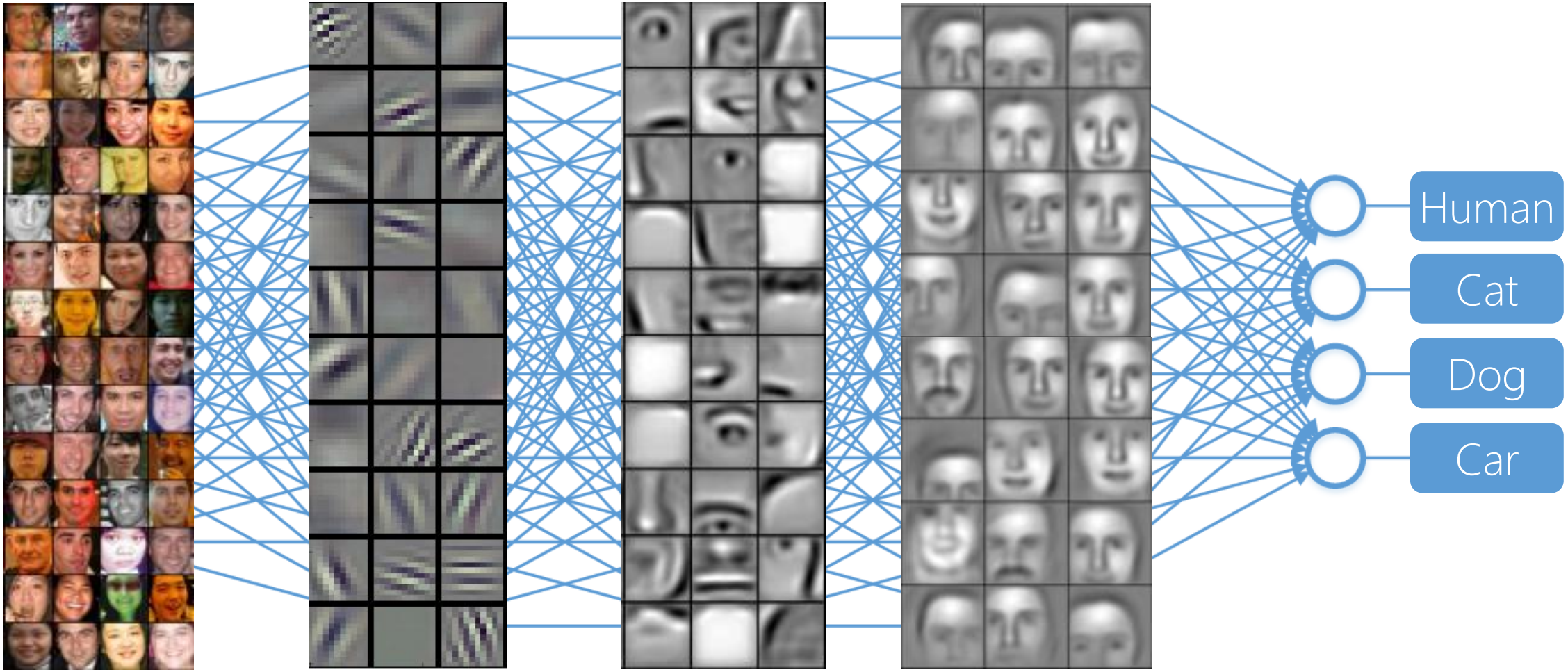
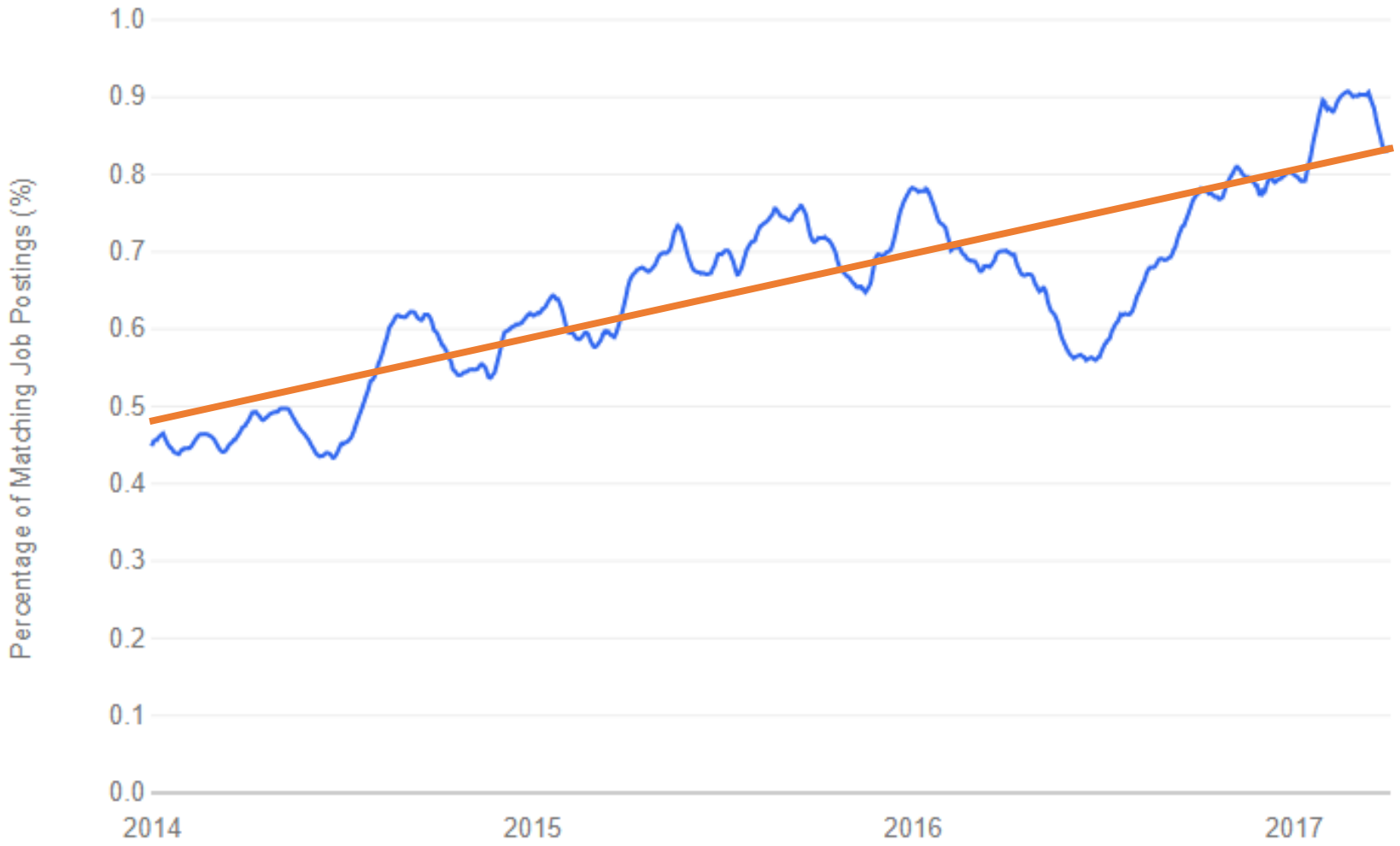
Practical Machine Learning with R

@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 = '';

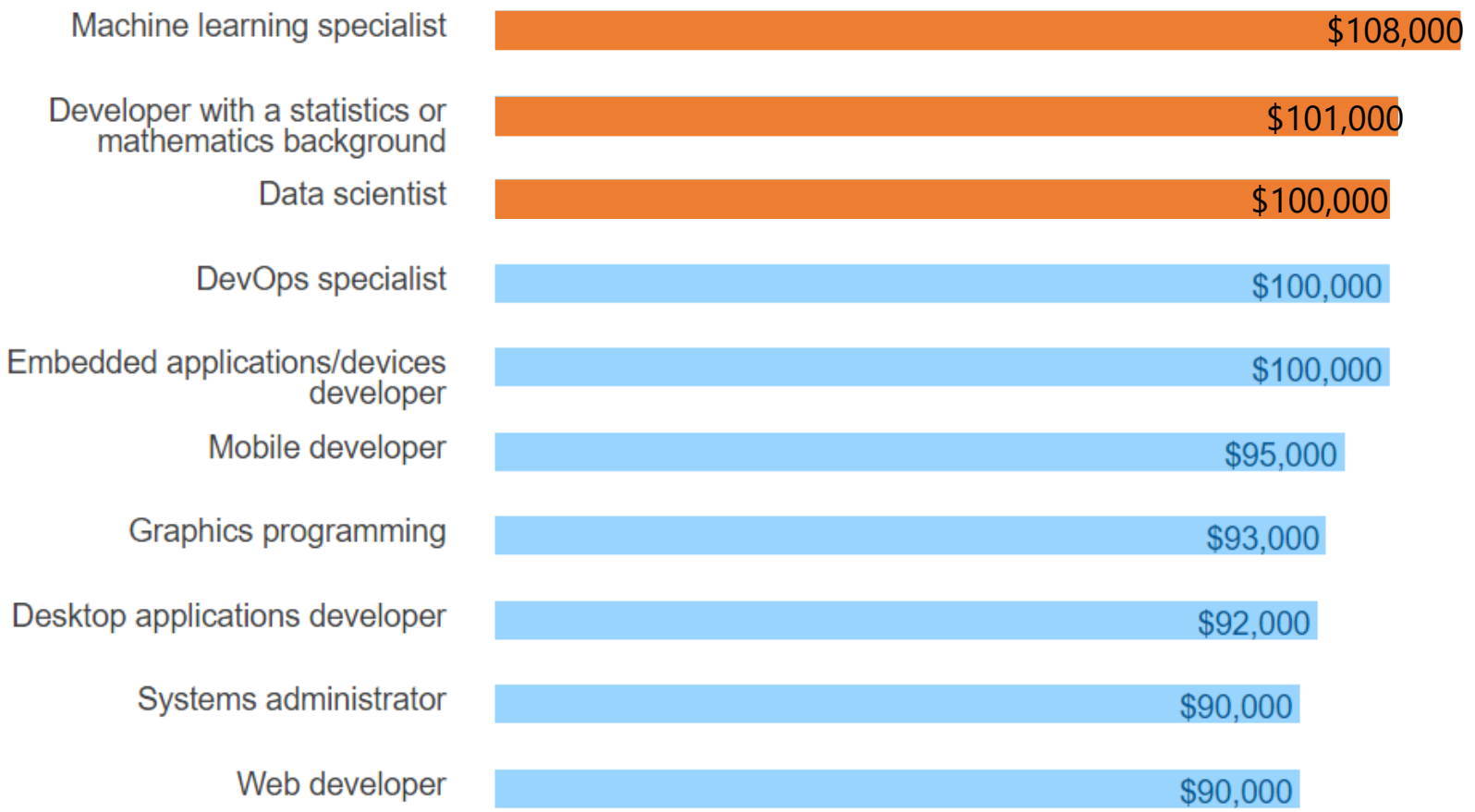

Human
Cat
Dog
Car

Job Postings for Machine Learning



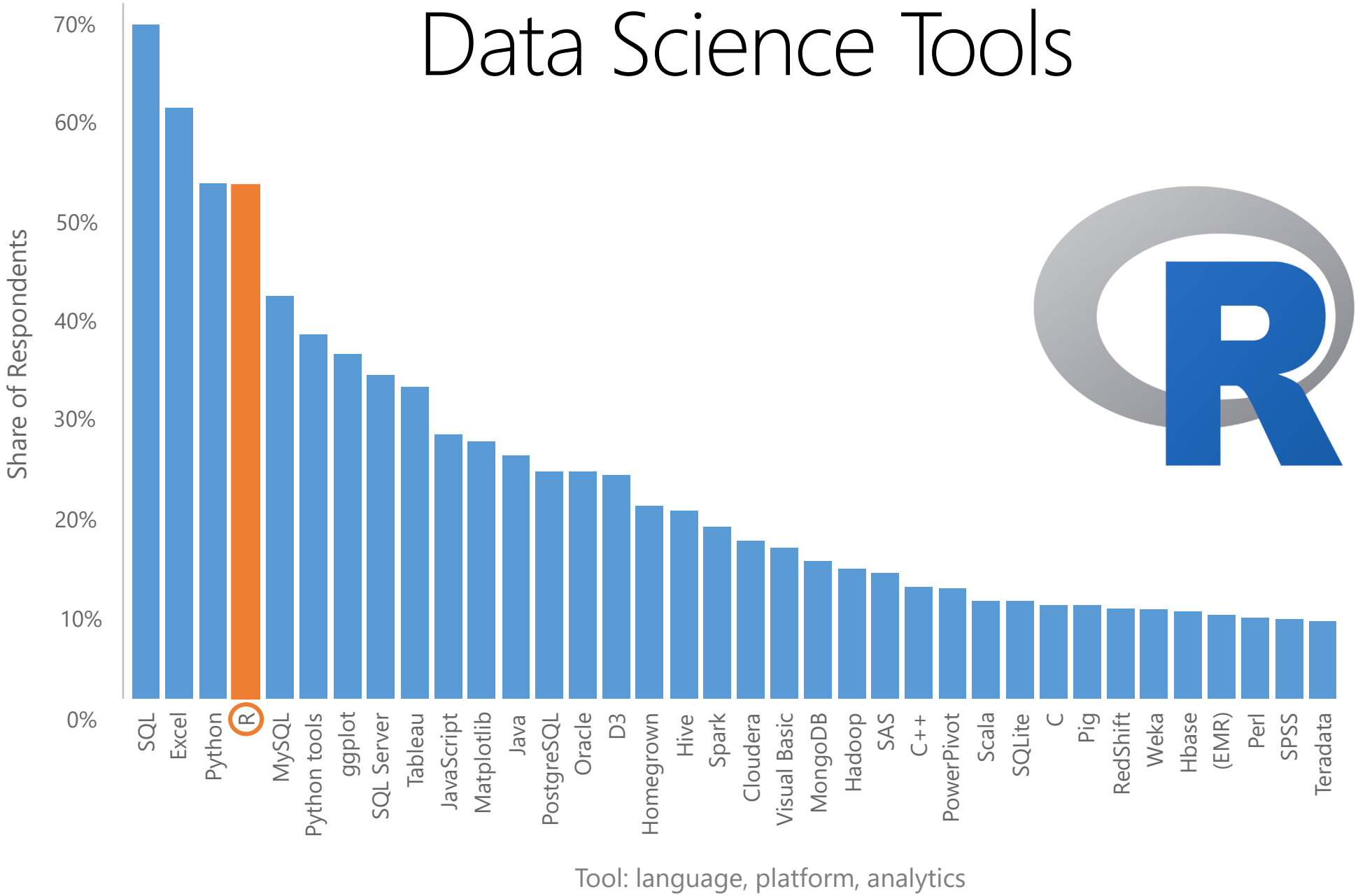
Source: Indeed.com

Average Salary by Job Type (USA)



Source: Stack Overflow 2017

Data Science Tools



Source: O'Reilly 2015 Data Science Salary Survey

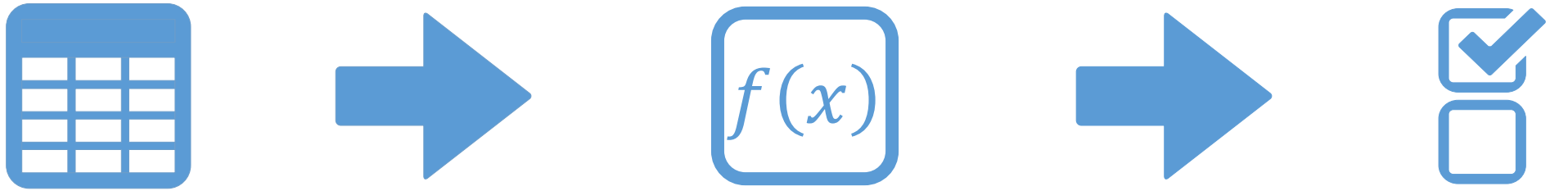
Overview

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

How Does This Apply to Me?

- ☑ Make decisions using data
- ☑ Make predictions using data
- ☑ Make recommendations using data
- ☑ Automate these with code

Conceptual Model



About Me

Data Science Consultant

Education

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

Community

- Public speaker
- Pluralsight author
- Microsoft MVP
- Open source

Schedule

Lectures (15 min)

Demos (10 min)

Labs (30 min)

Breaks (5 min)

Logistics

Pairing for labs is optional

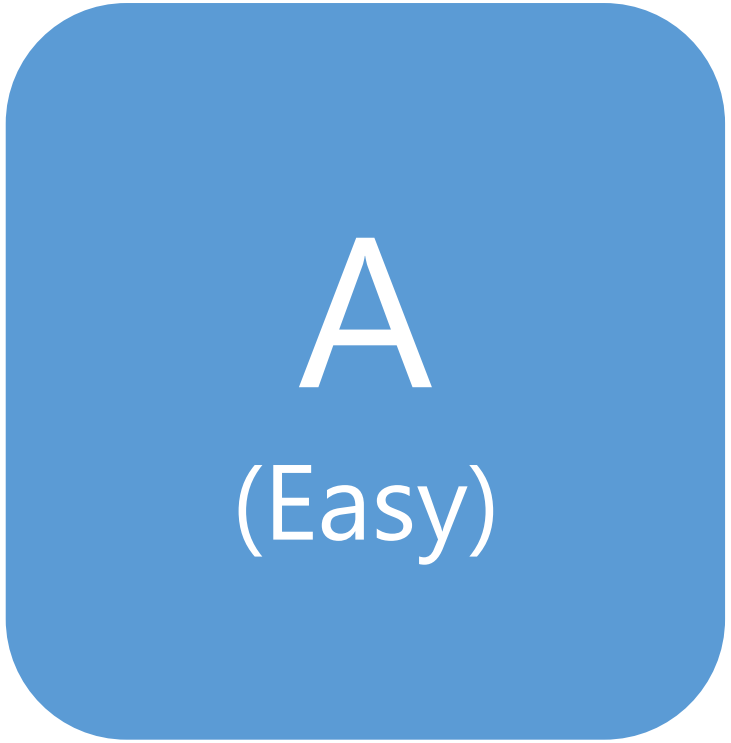
Ask questions if needed

Come and go as needed

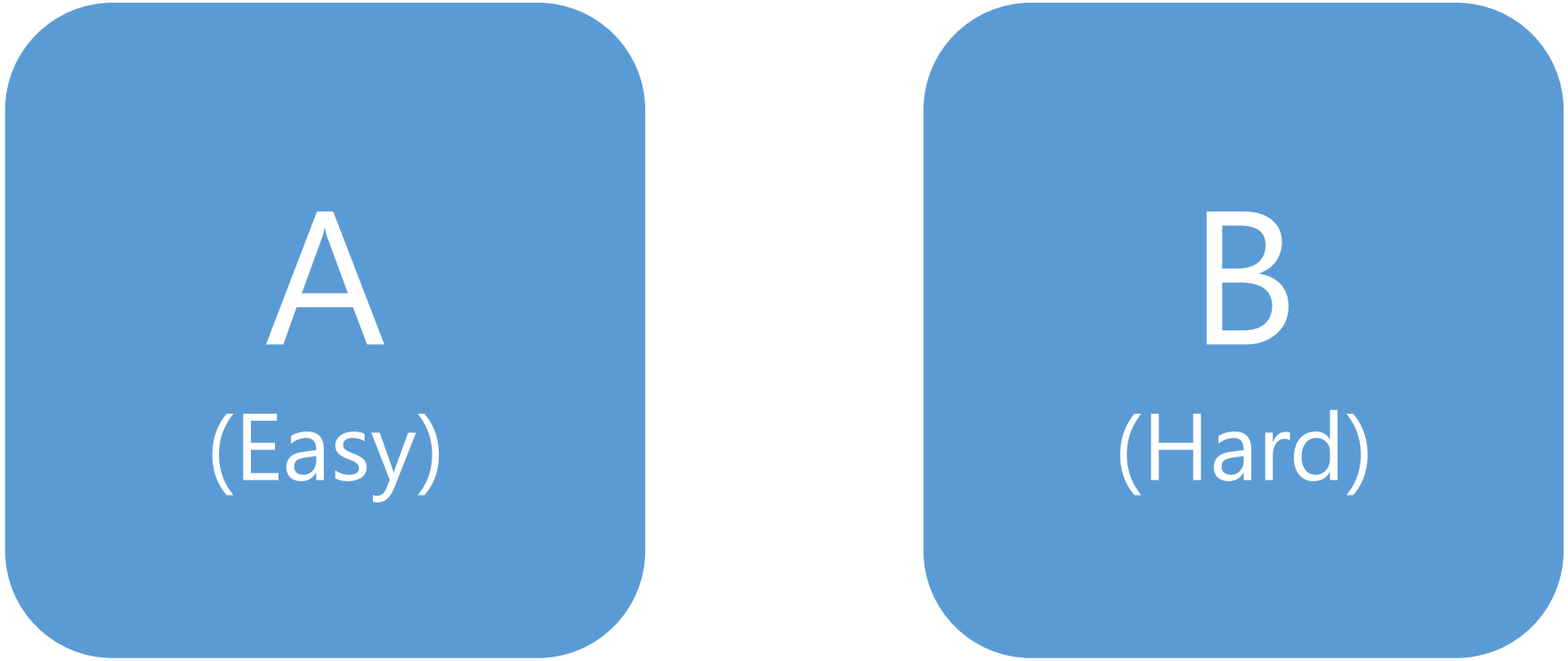
Feedback forms at the end

Labs

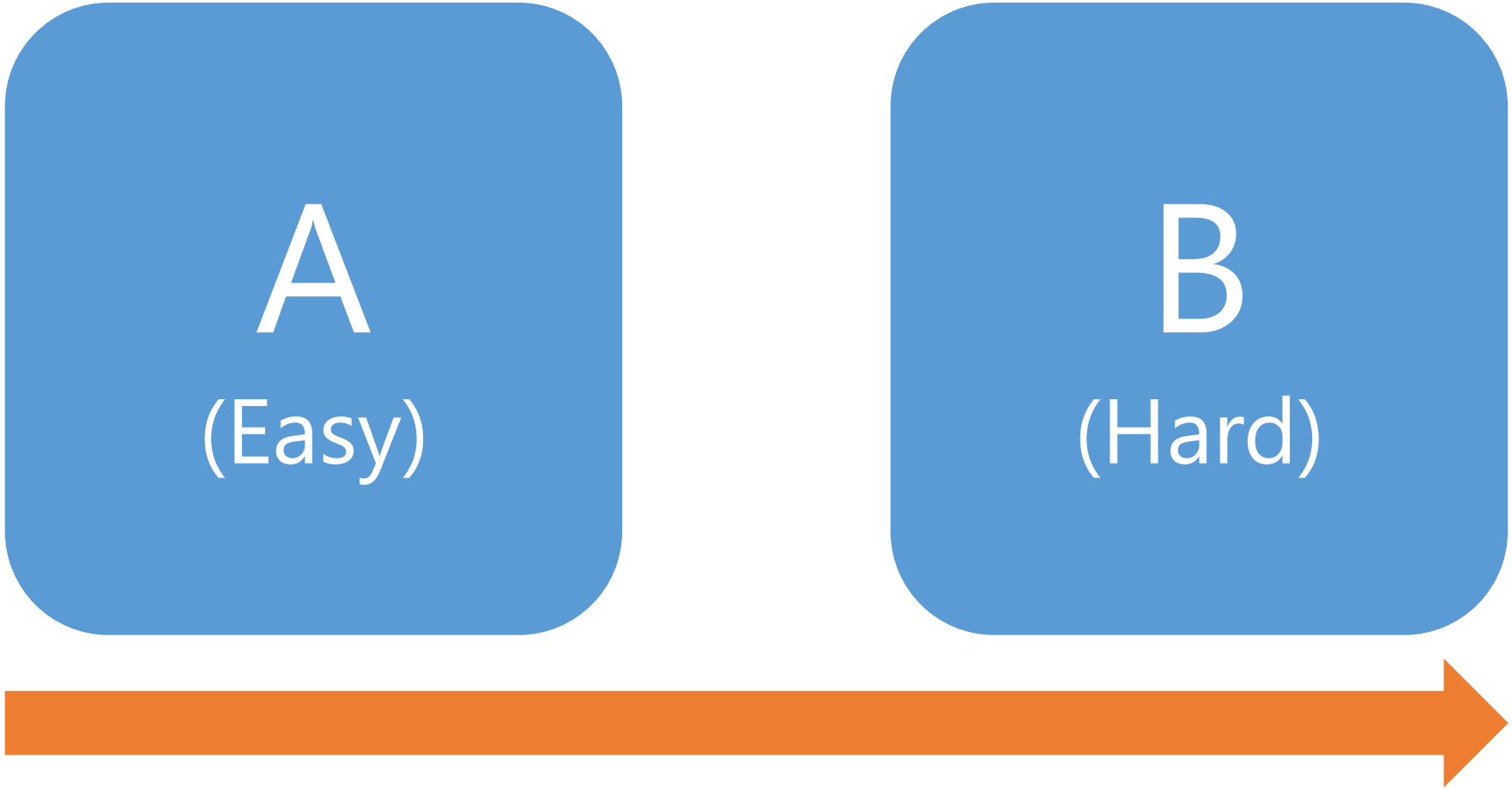
Labs



Labs



Labs

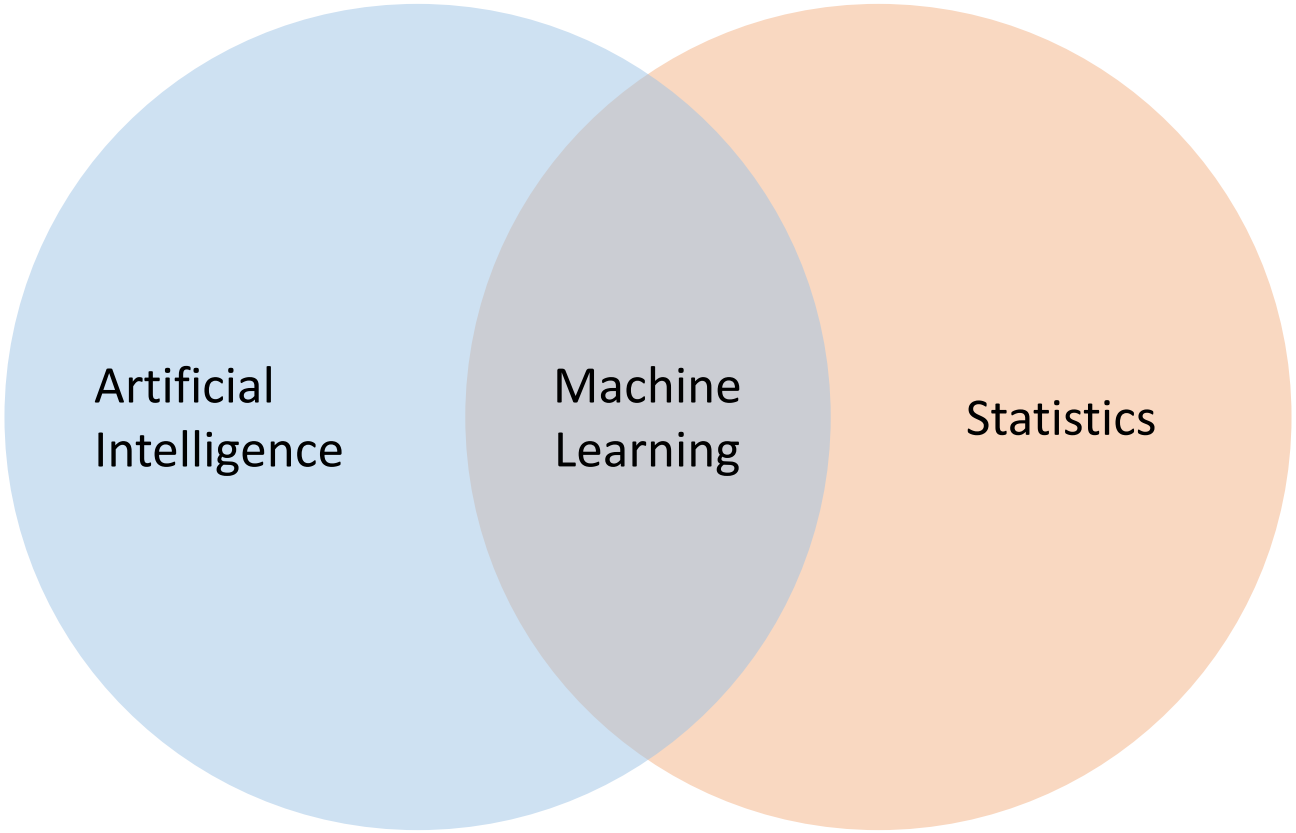


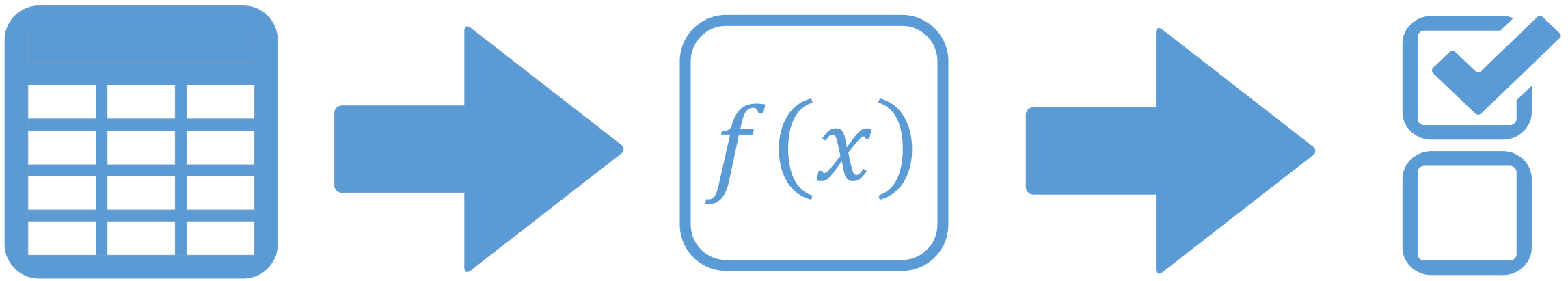
Workshop URL

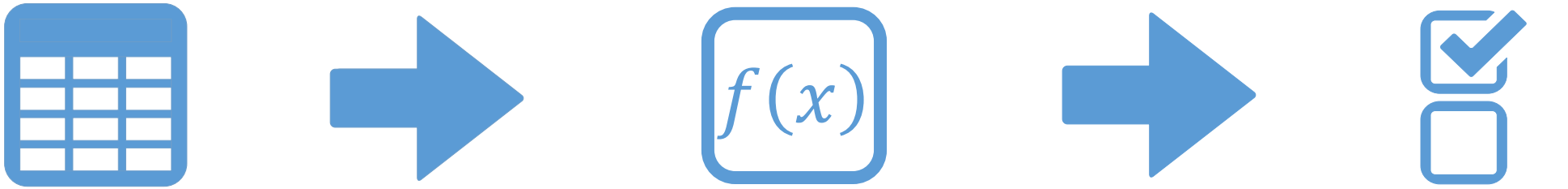
http://www.matthewrenze.com/workshops/practical-machine-learning-with-r/

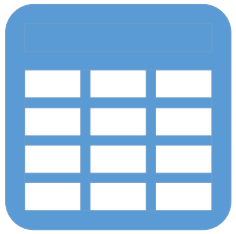
Introduction to Machine Learning

What is machine learning?

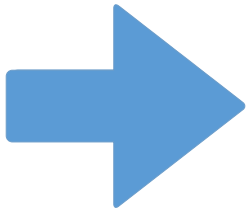

Artificial
Intelligence
Machine
Learning
Statistics

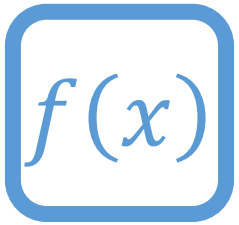

$f(x)$


$f(x)$
Data
Function
Prediction

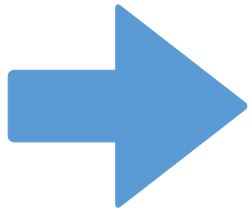


Data

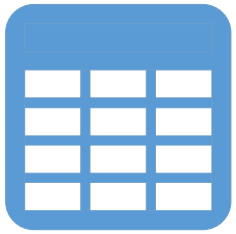




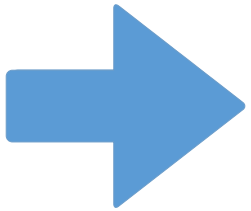
Function

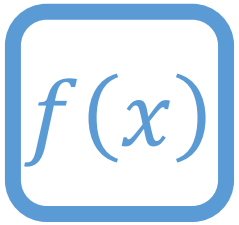


Prediction

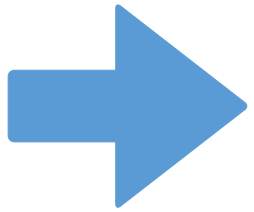


Data





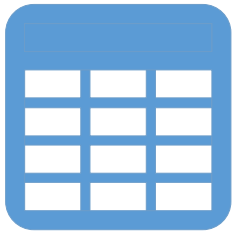
Function



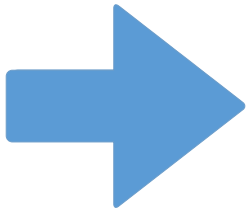
Prediction

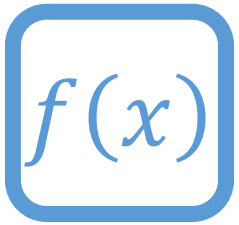
Cat

Dog

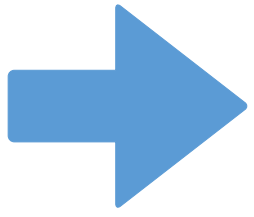


Data





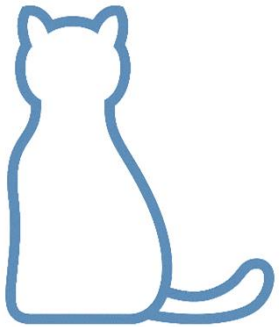
Function

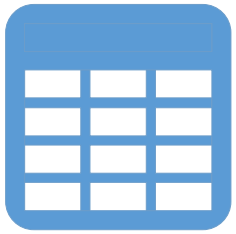


Prediction

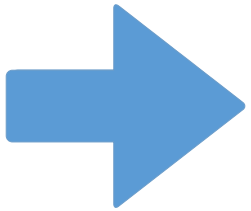
Cat

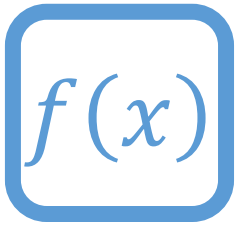
Dog



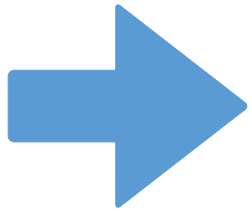


Data





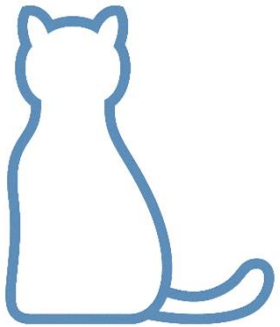
Function



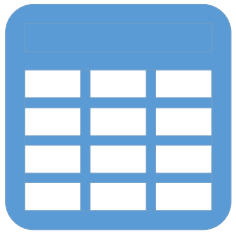
Prediction

Cat

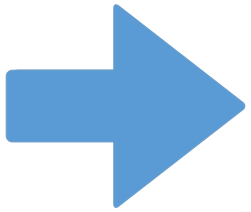
Dog

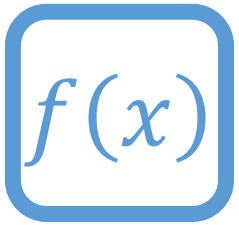


Is cat?

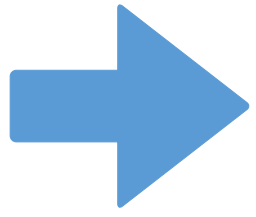


Data





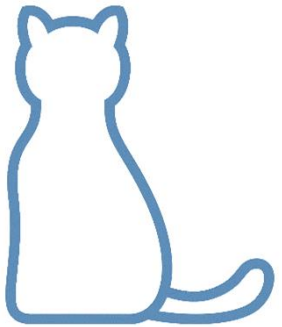
Function



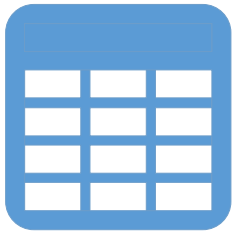
Prediction

Cat

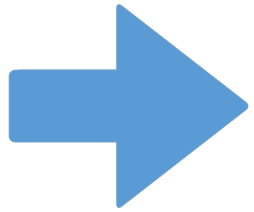
Dog

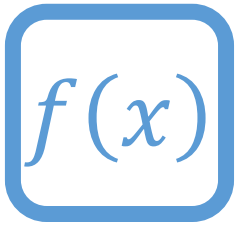


Is cat?

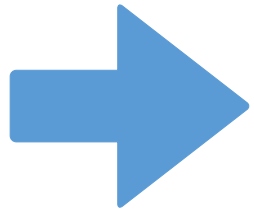


Data





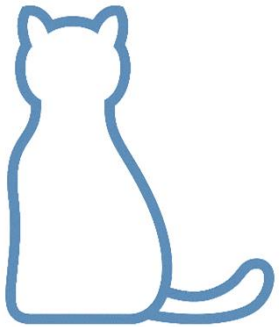
Function



Prediction

Cat

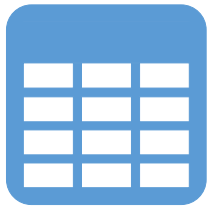
Dog



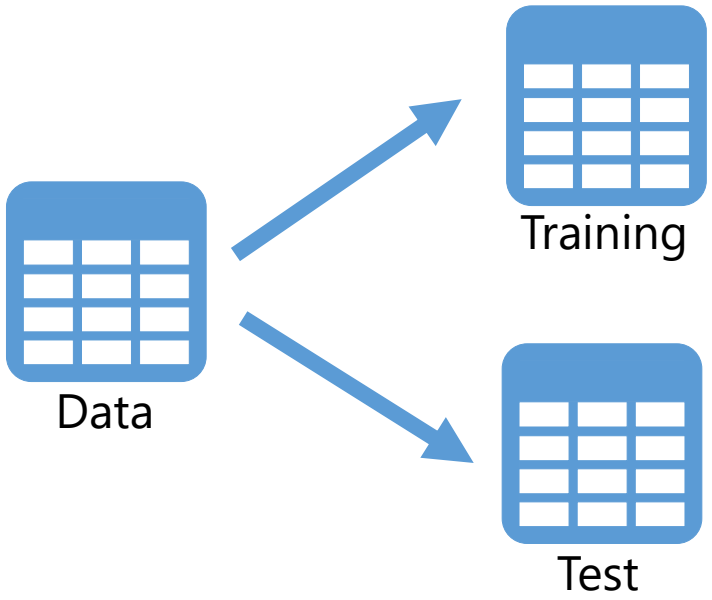
Is cat?

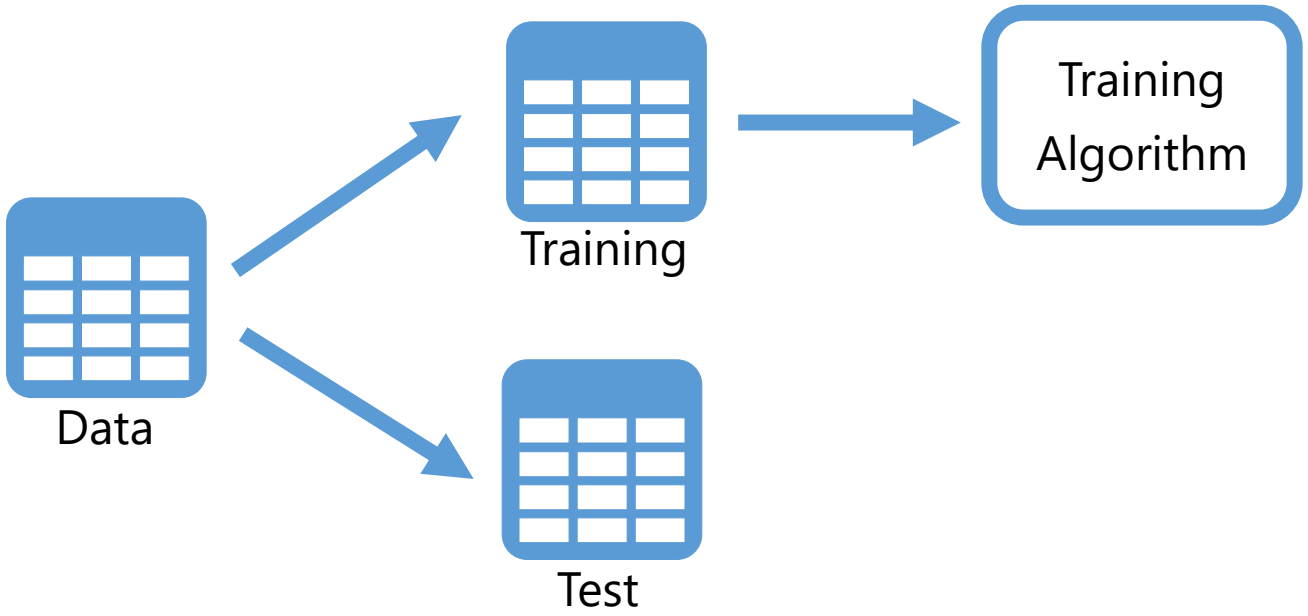
Yes

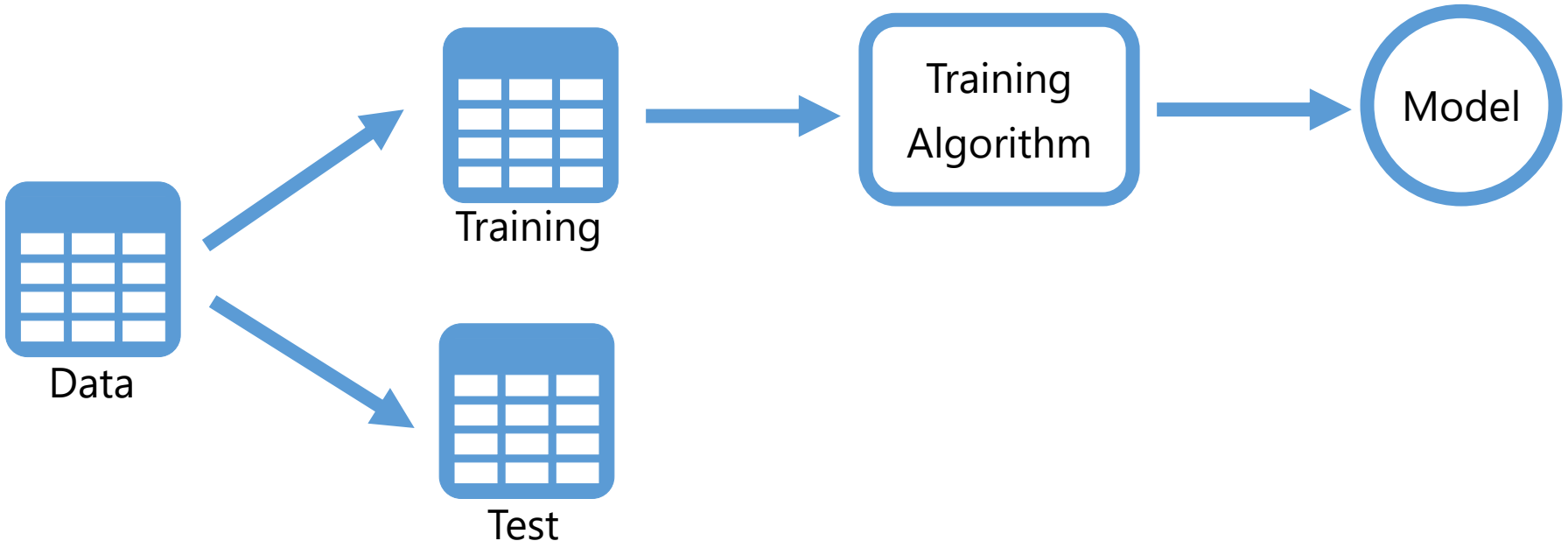
How does machine learning work?

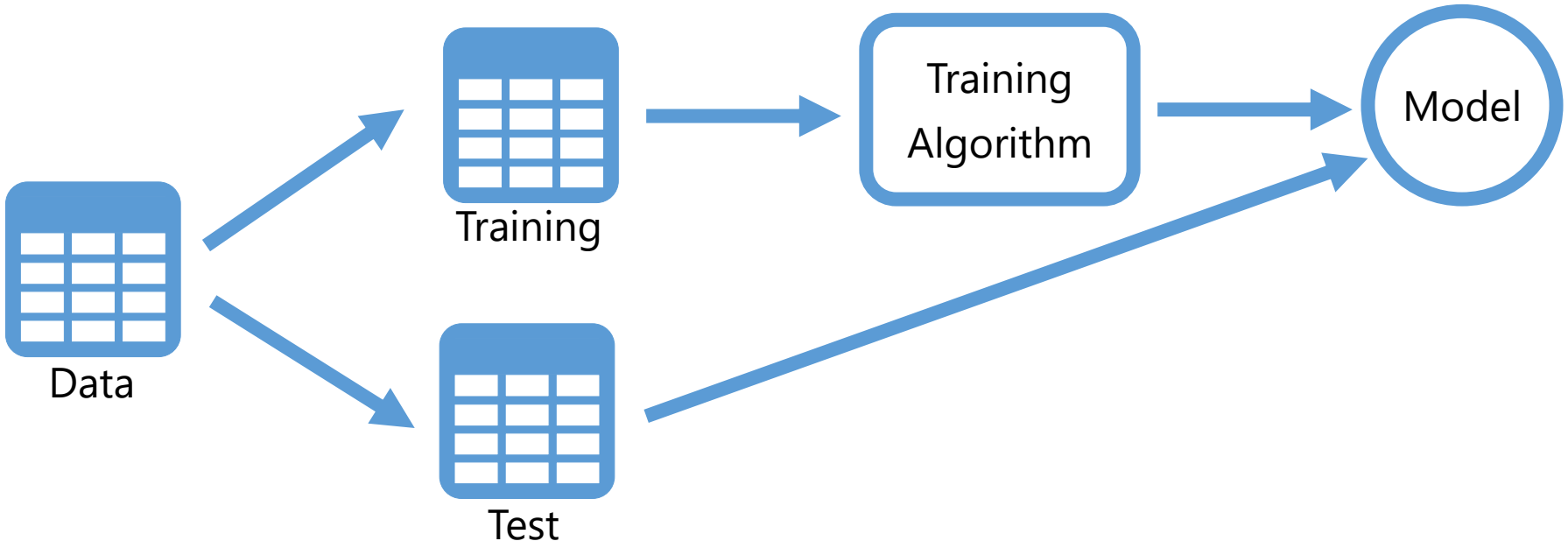


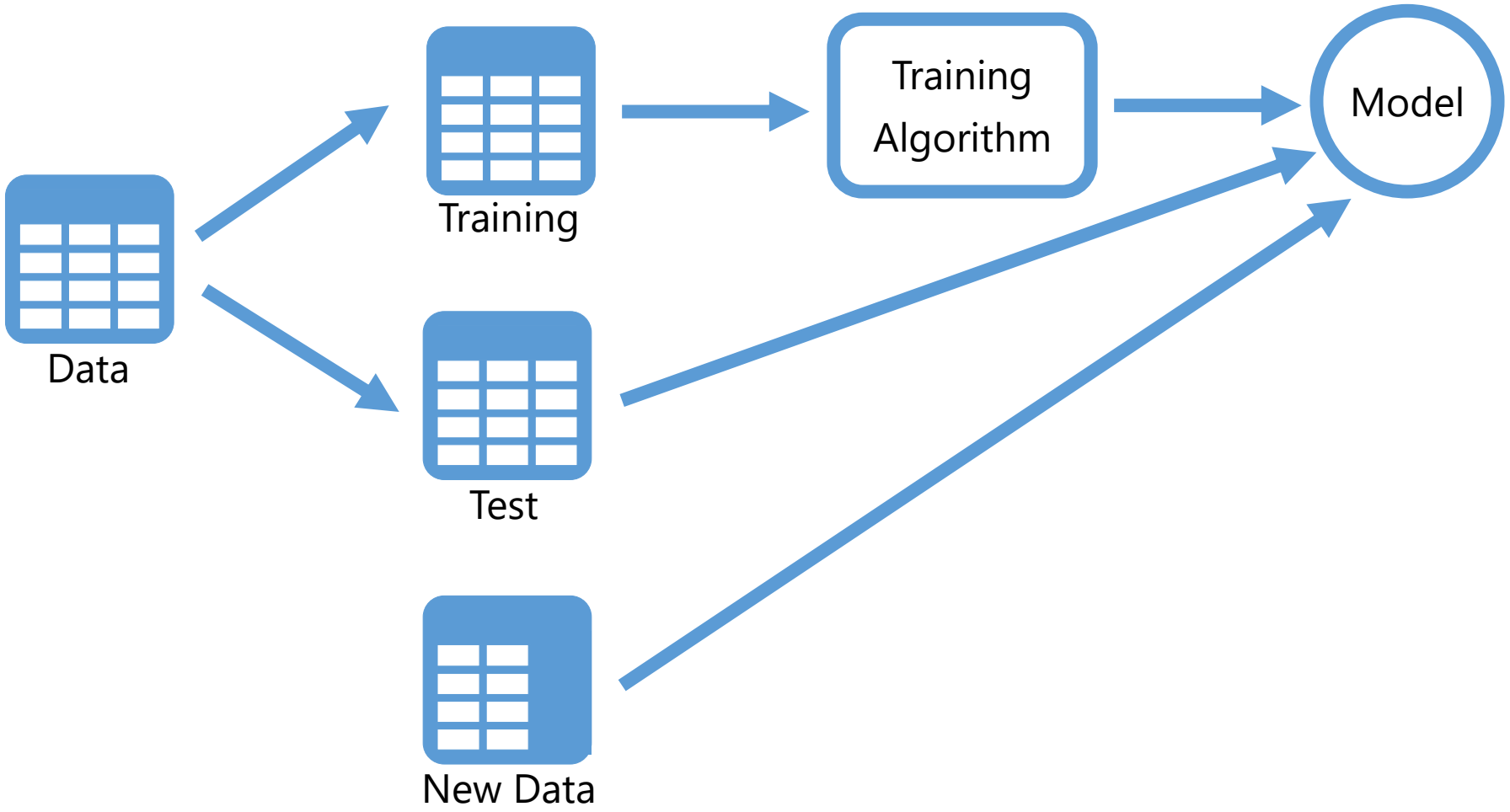
Data

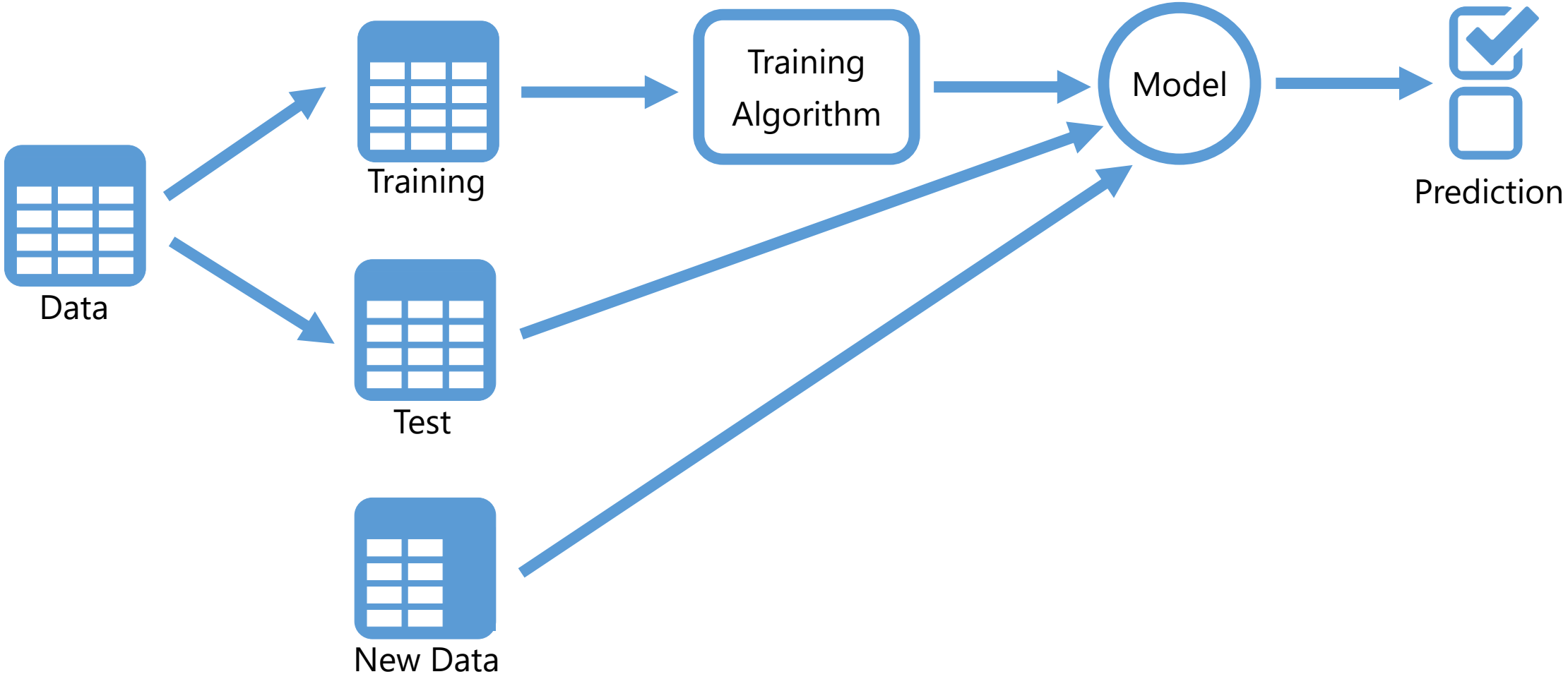

Training
Data
Test

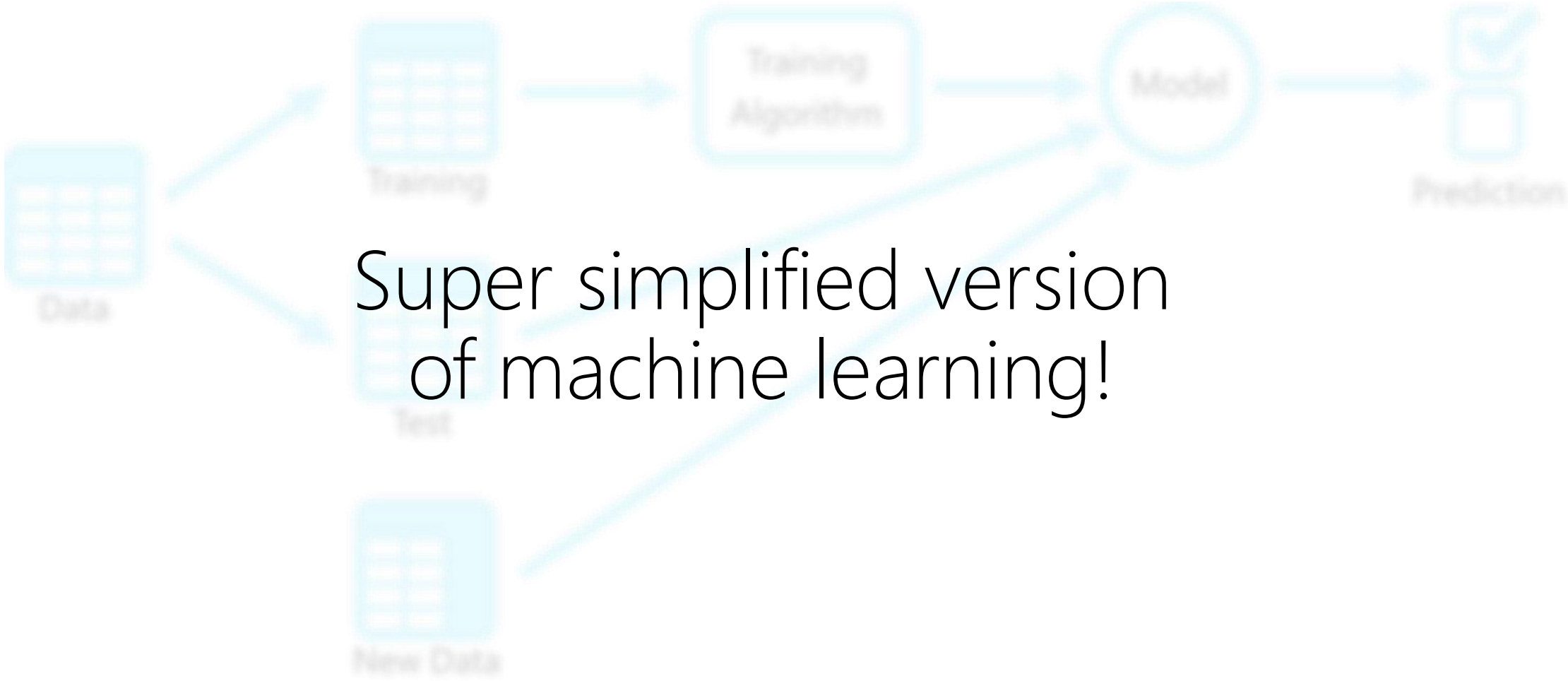

Training
Algorithm
Training
Data
Test


Data
Training
Test
Training
Algorithm
Model


Data
Training
Test
Training
Algorithm
Model


Data
Training
Test
New Data
Training
Algorithm
Model

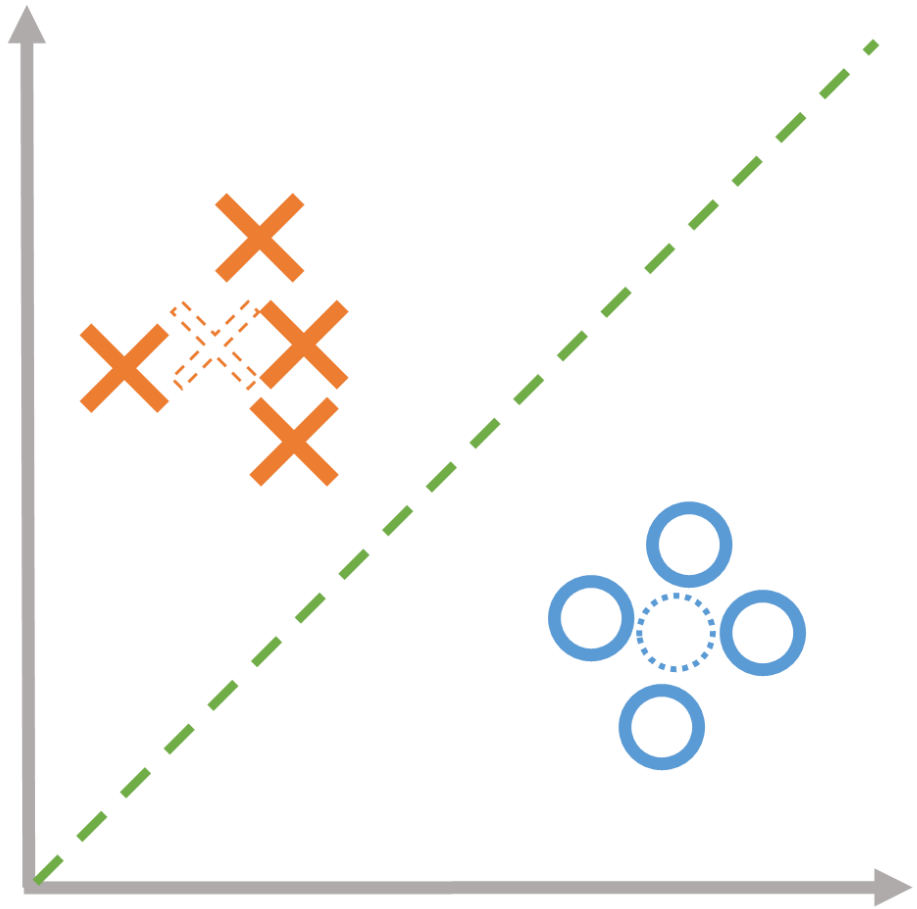

Data
Training
Test
New Data
Training
Algorithm
Model
Prediction



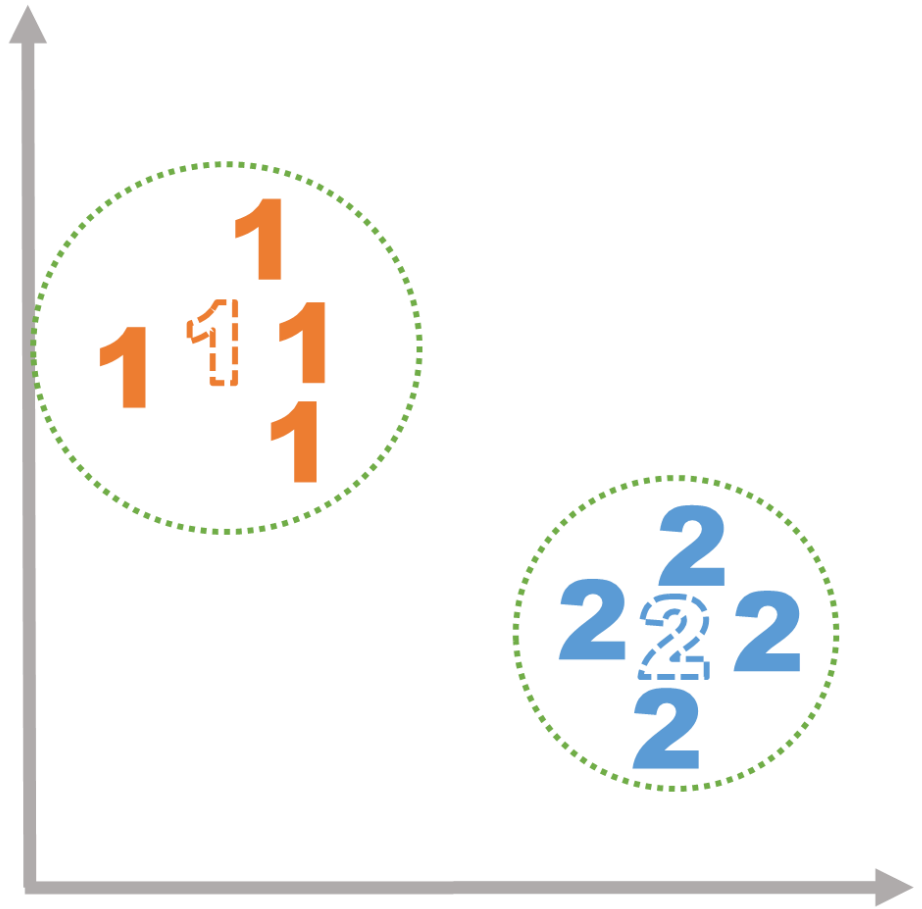
Super simplified version of machine learning!

What types of machine learning exist?

Types of Machine Learning

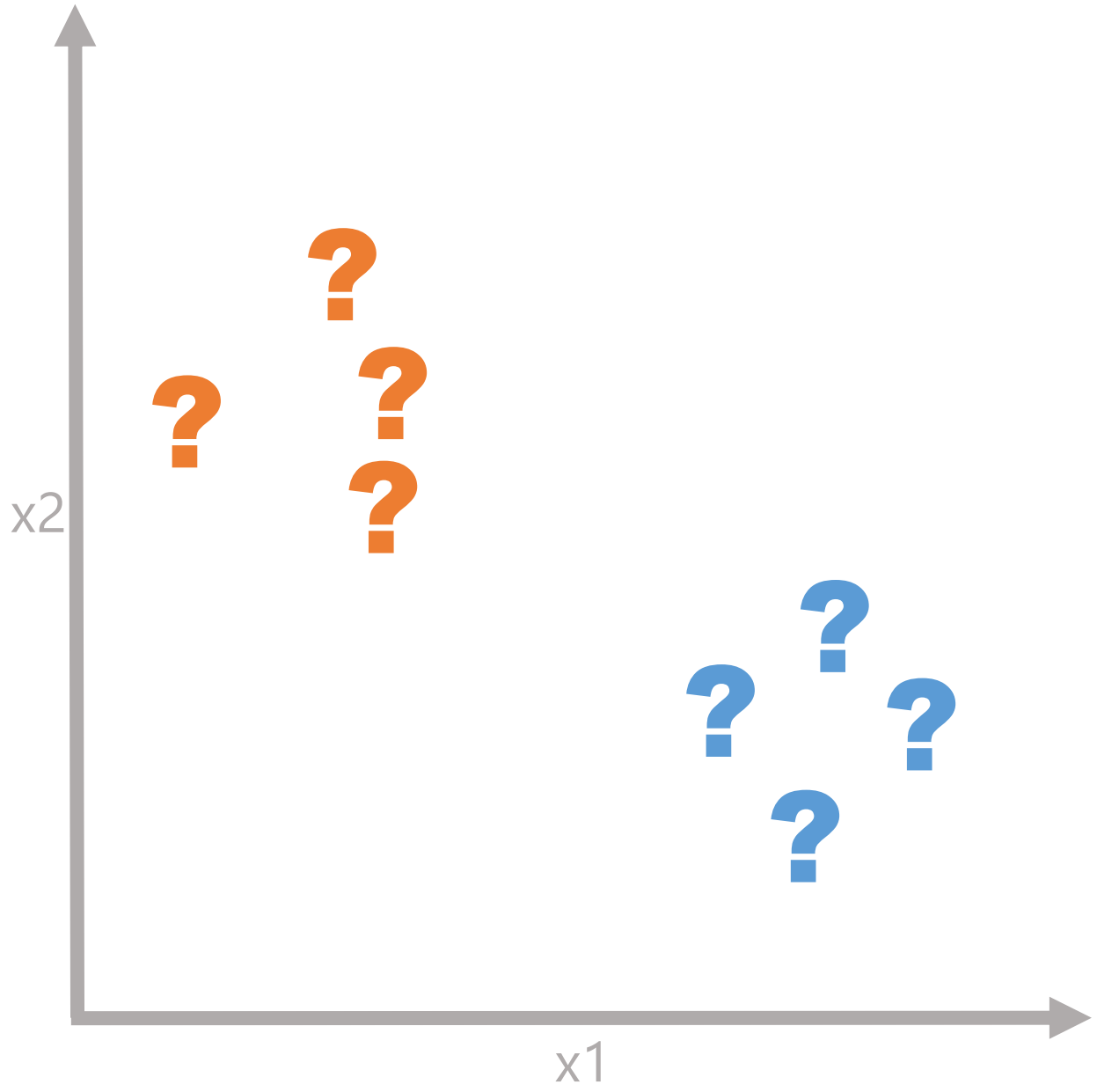


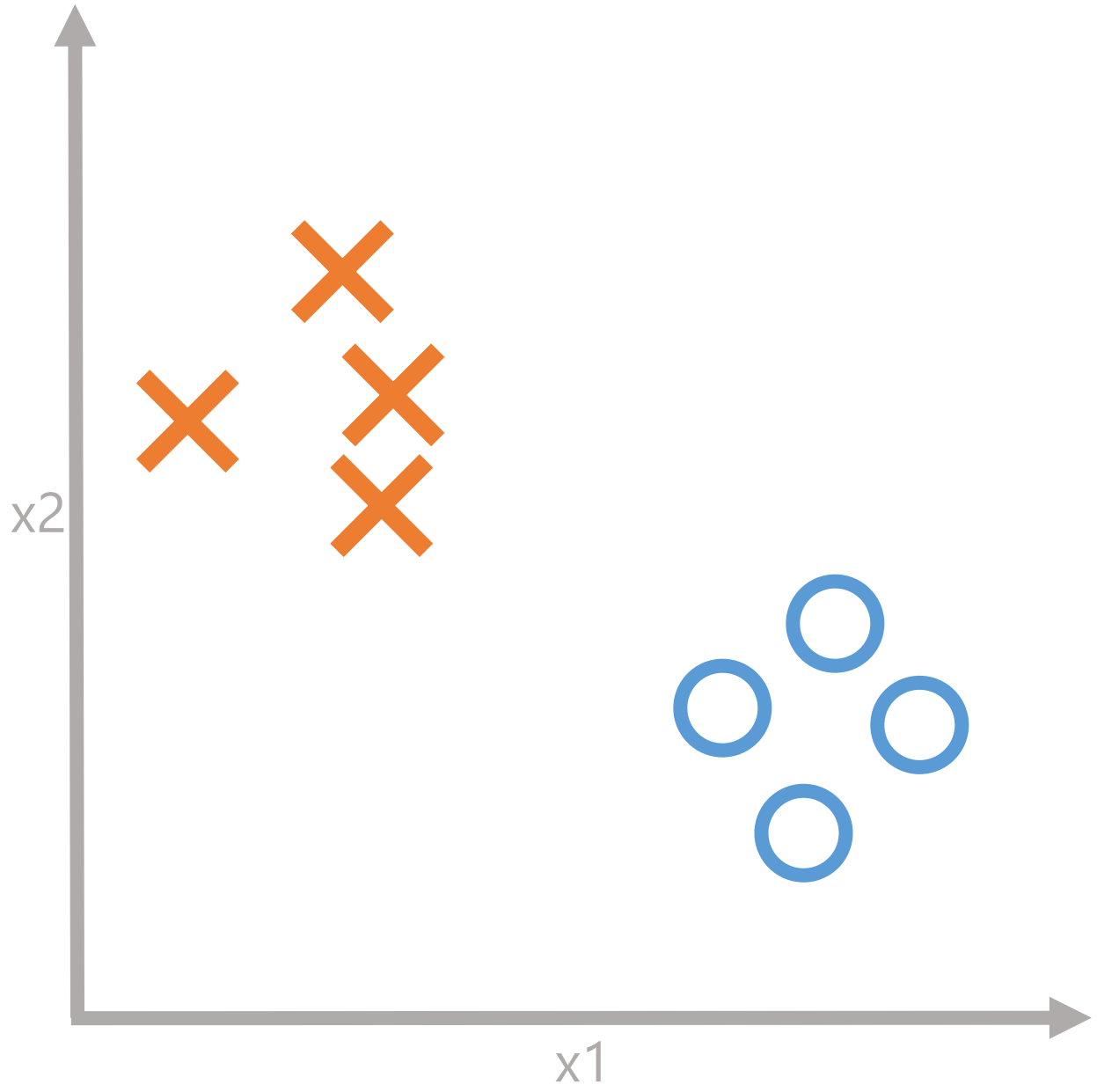
Supervised Learning

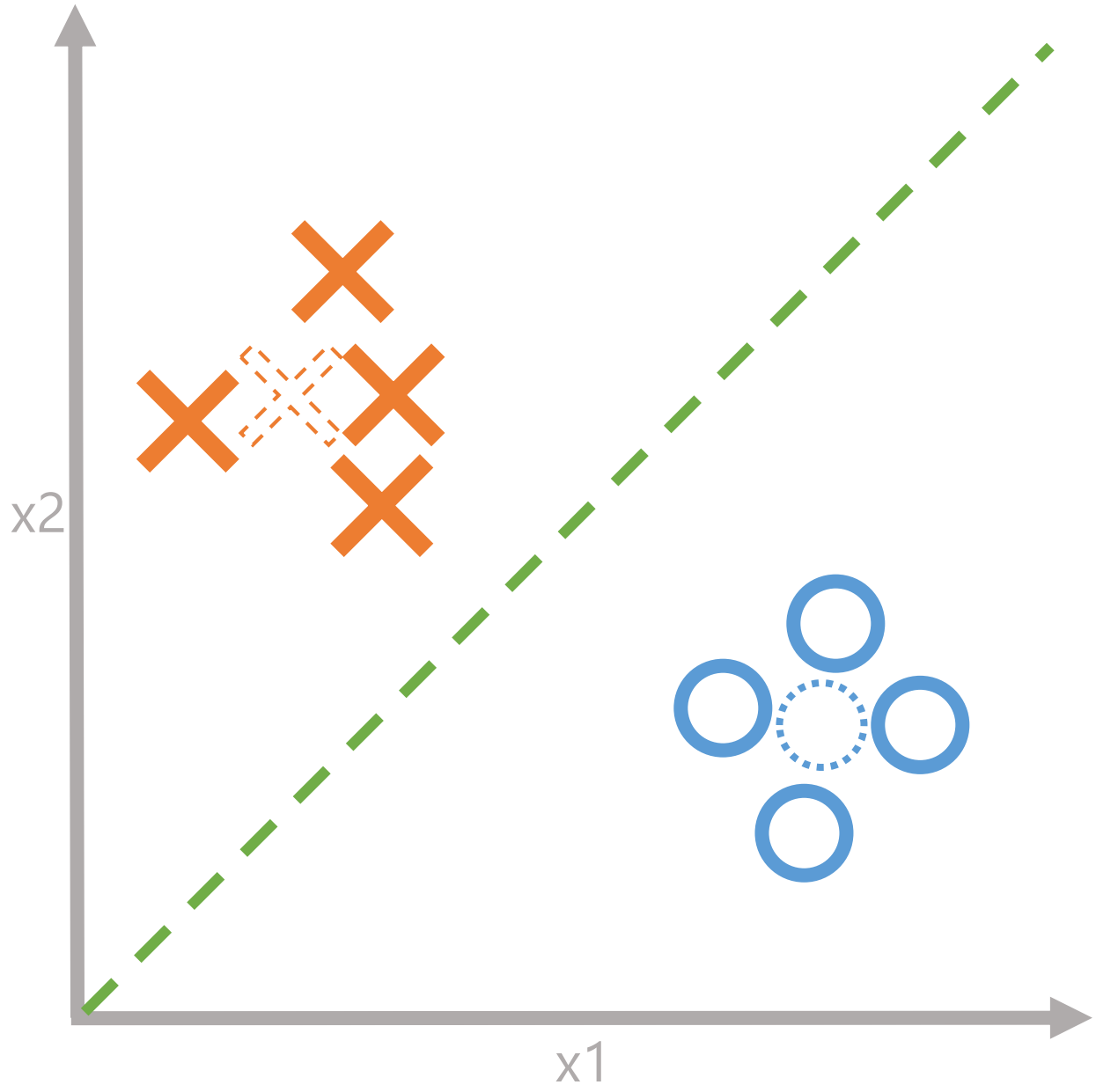


Unsupervised Learning

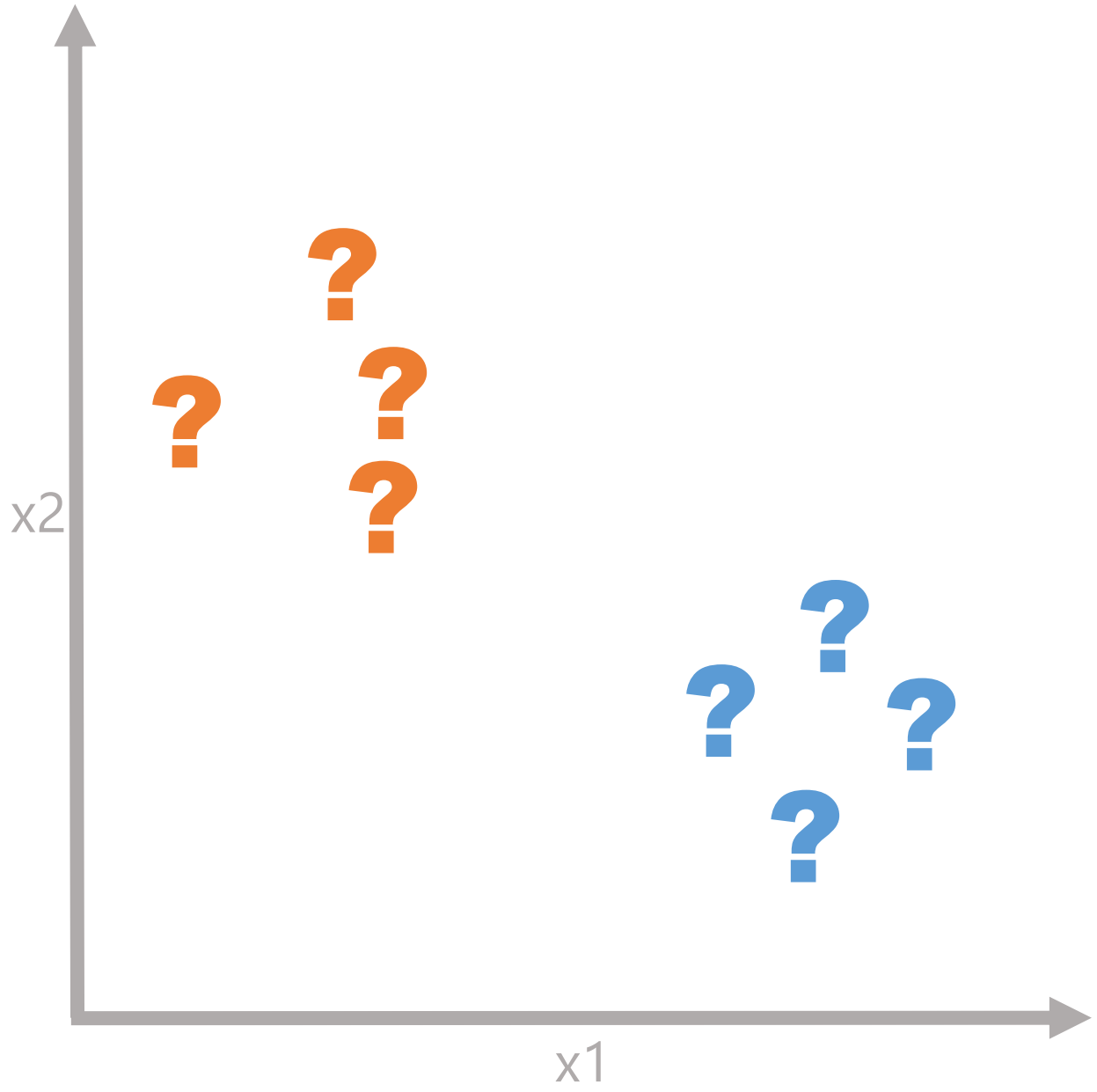
Supervised Learning

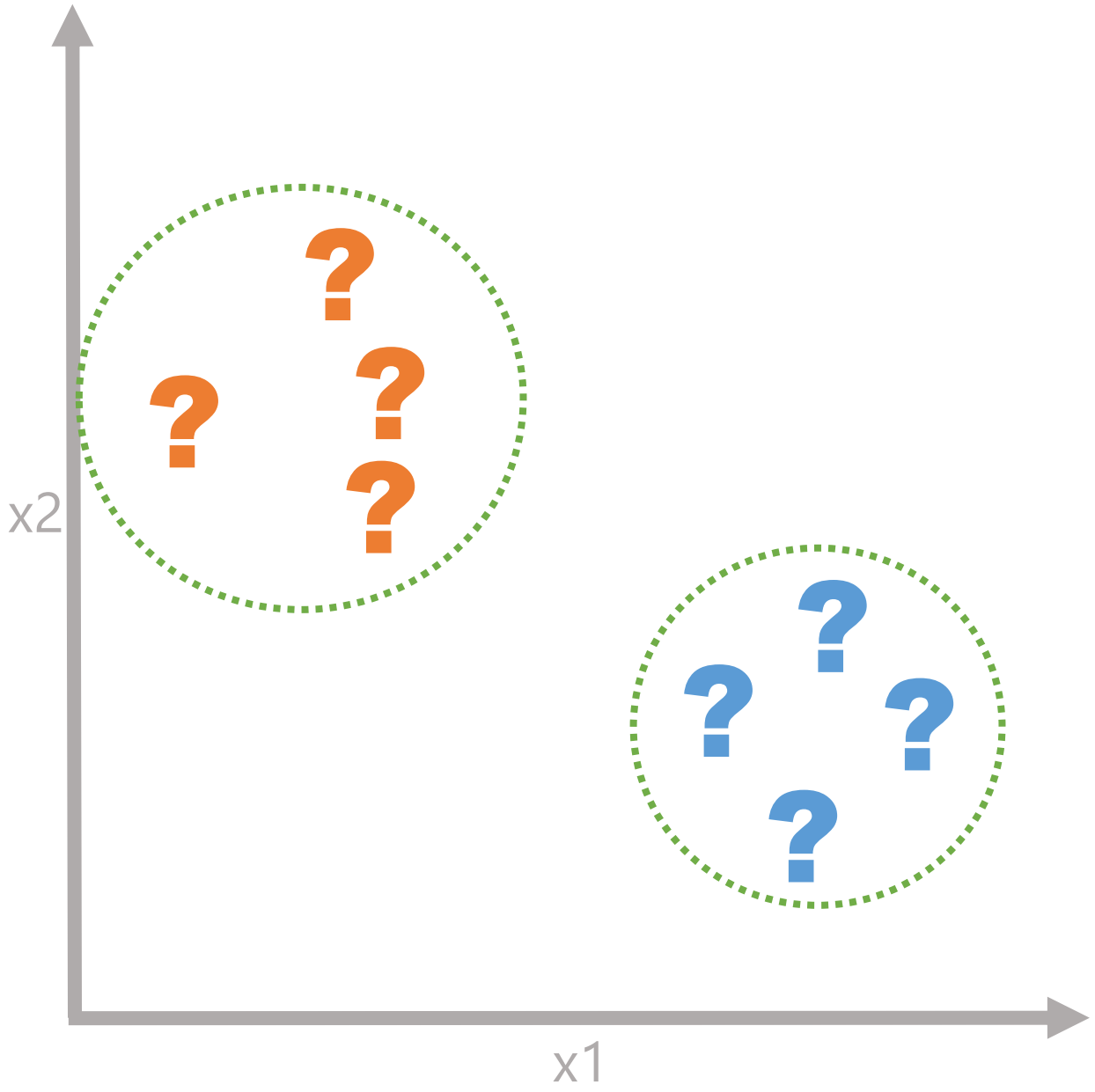

x2
x1

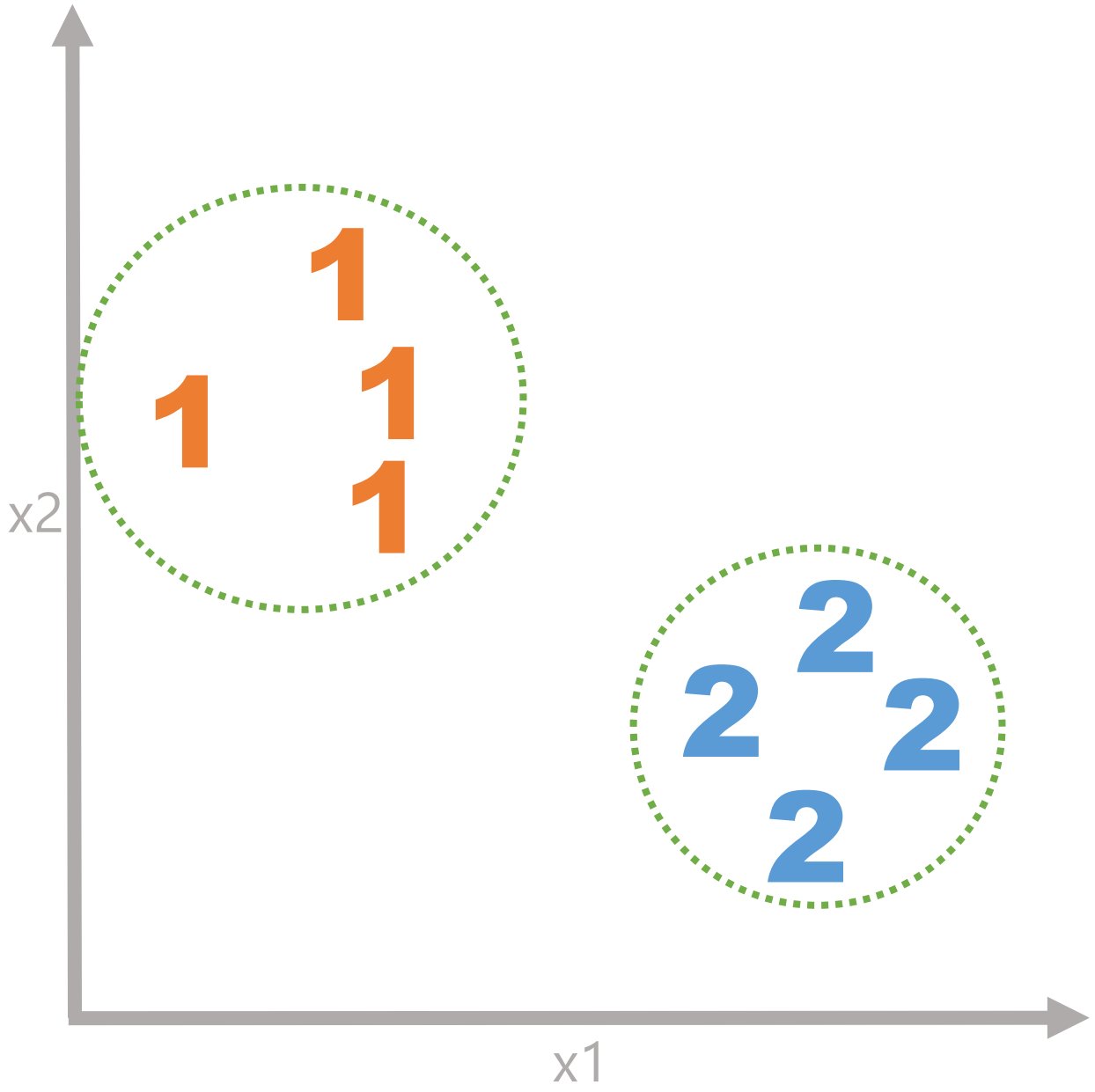

x2
x1

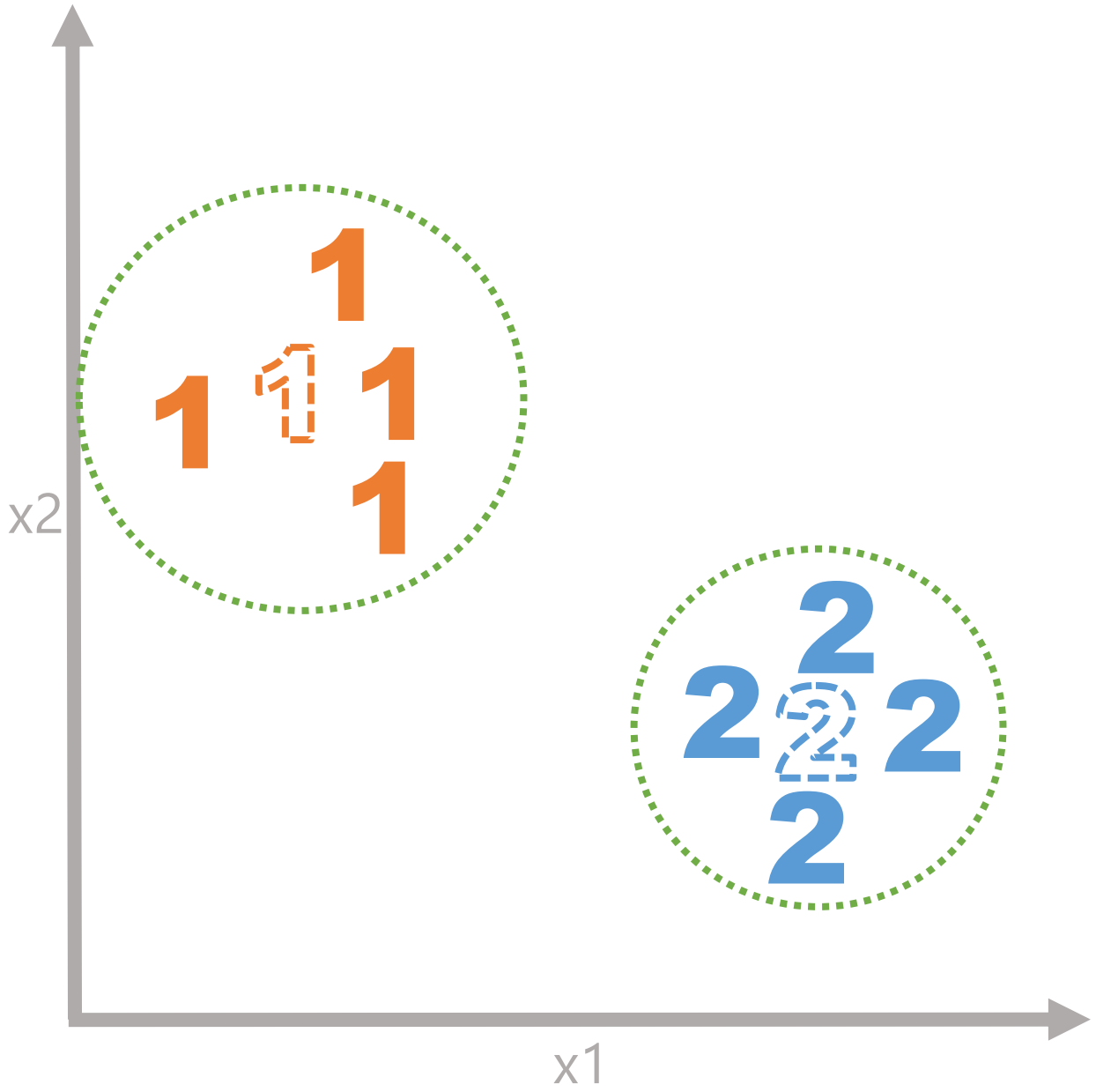

x2
x1

Unsupervised Learning

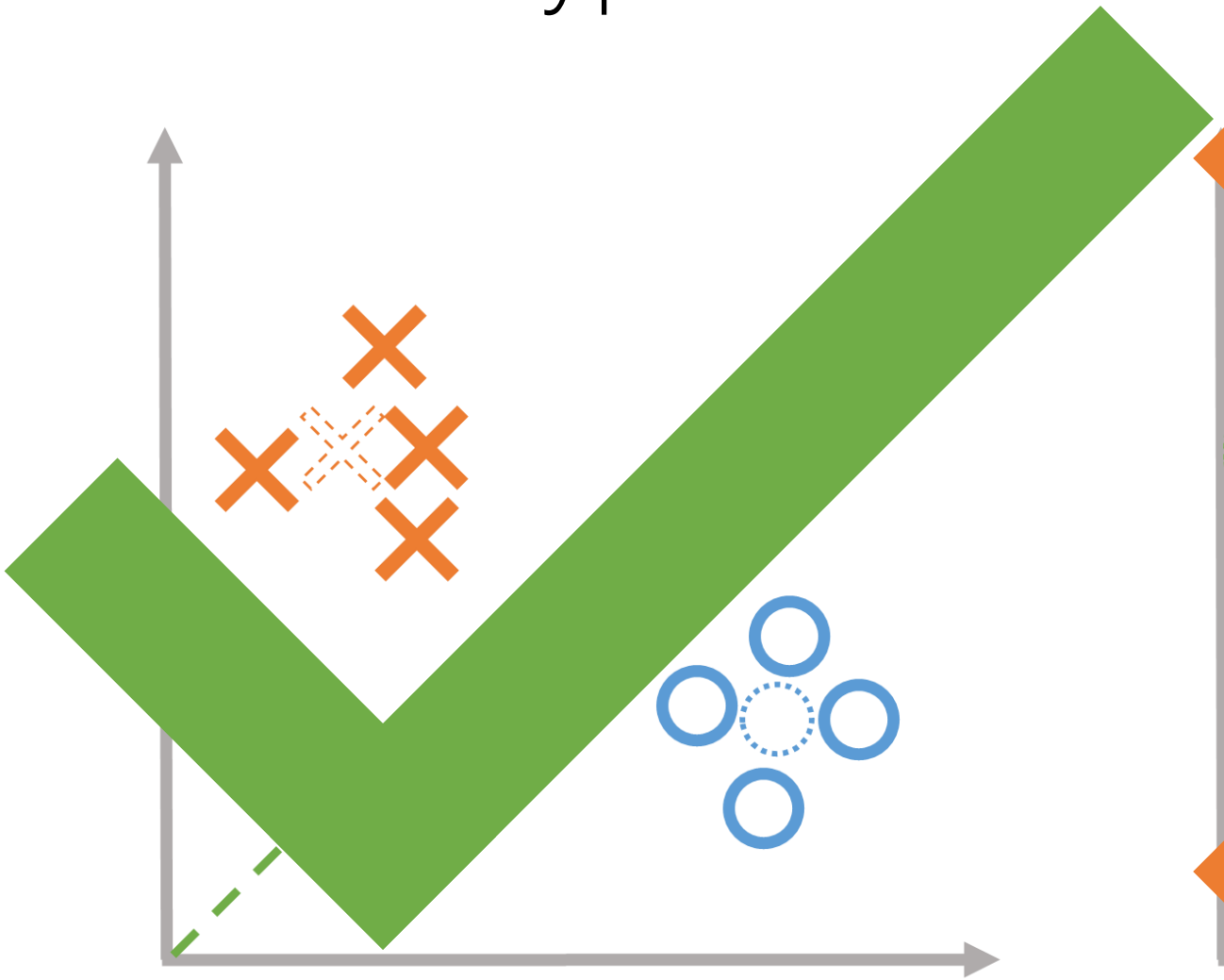

x2
x1


x2
x1

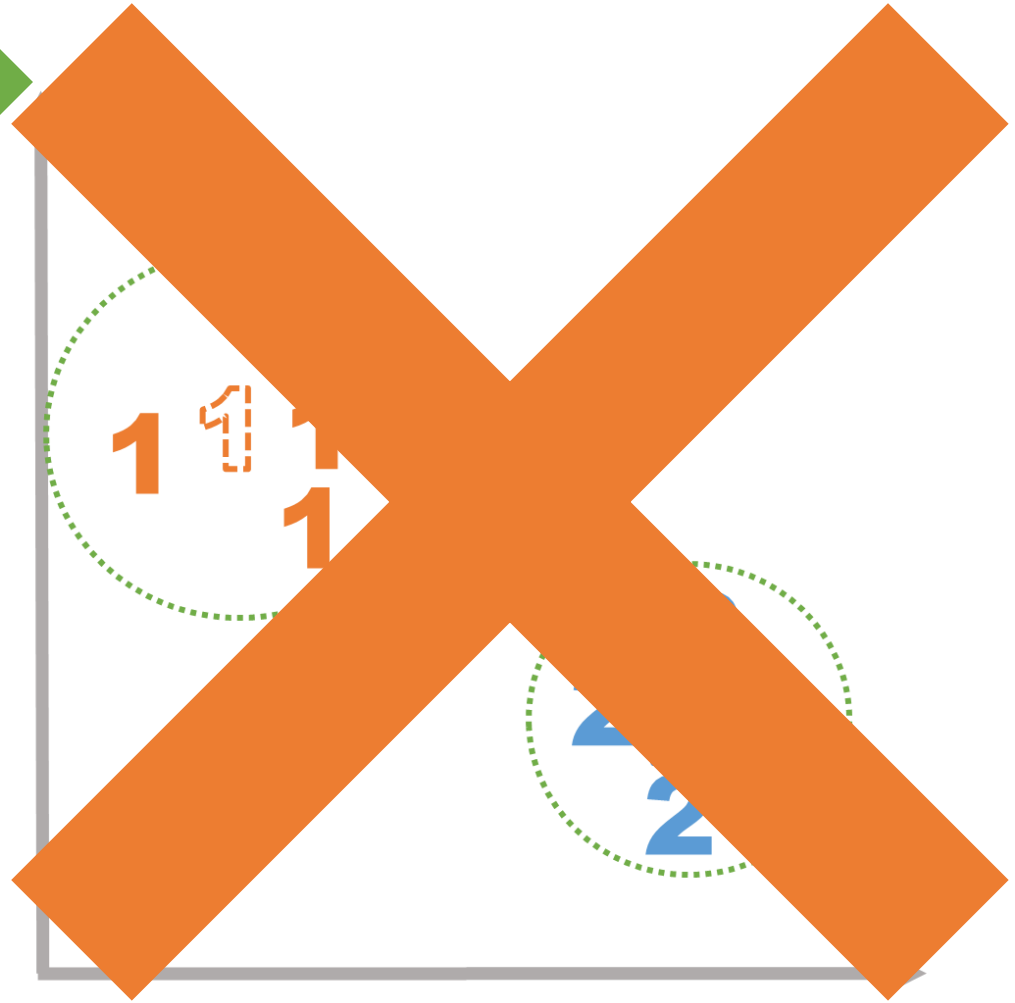

x2
x1
1
1
1
1
2
2
2
2


x2
x1
1
1
1
1
2
2
2
2

Types of Machine Learning

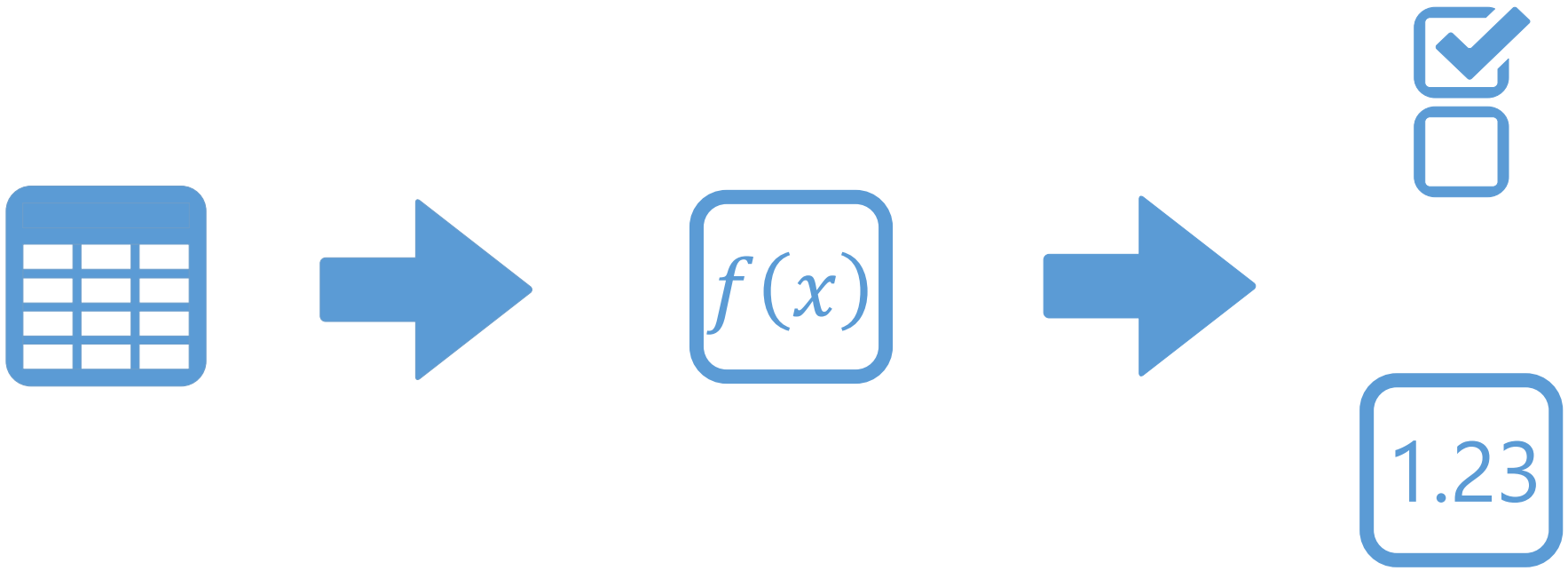


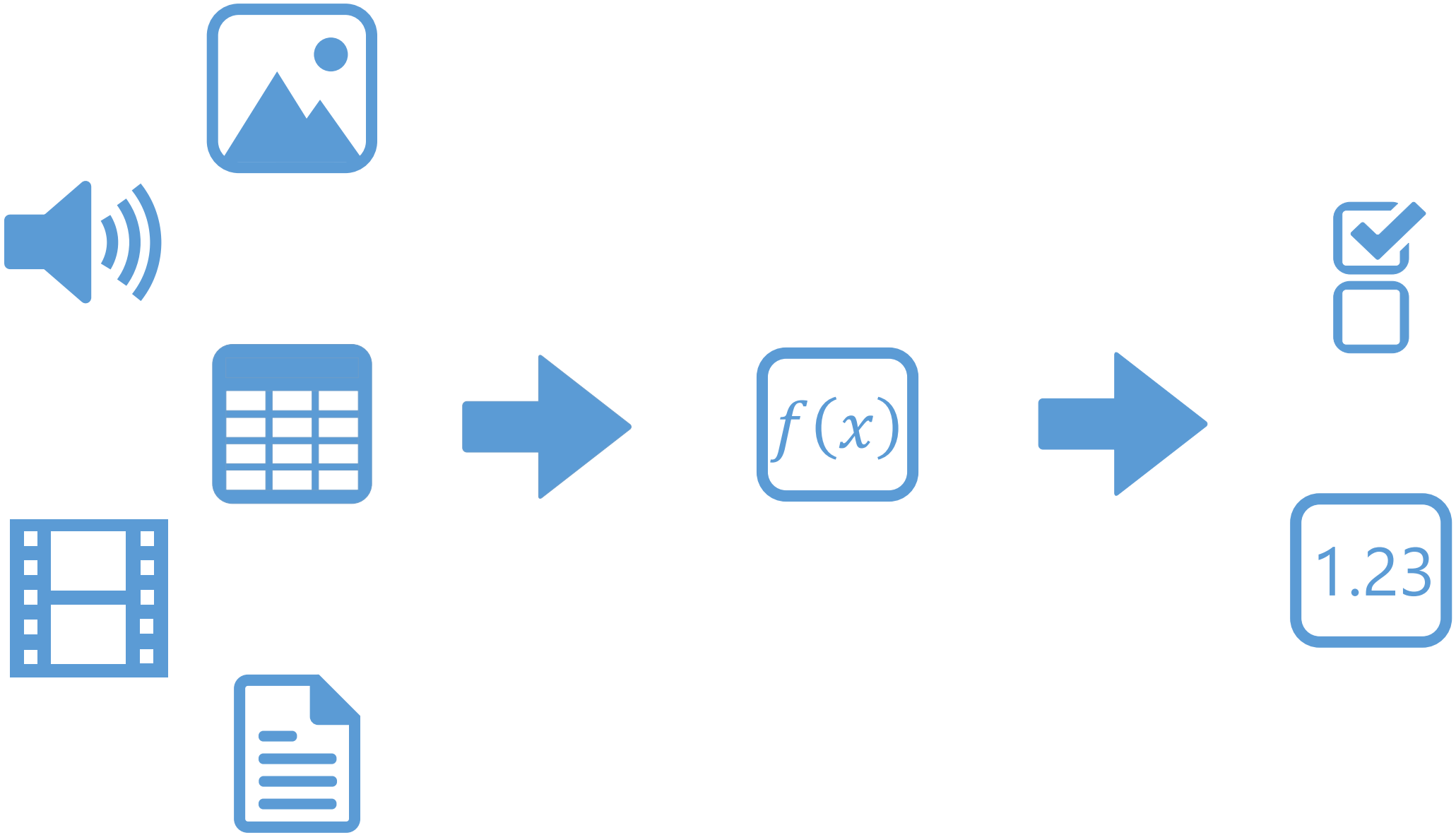
Supervised Learning

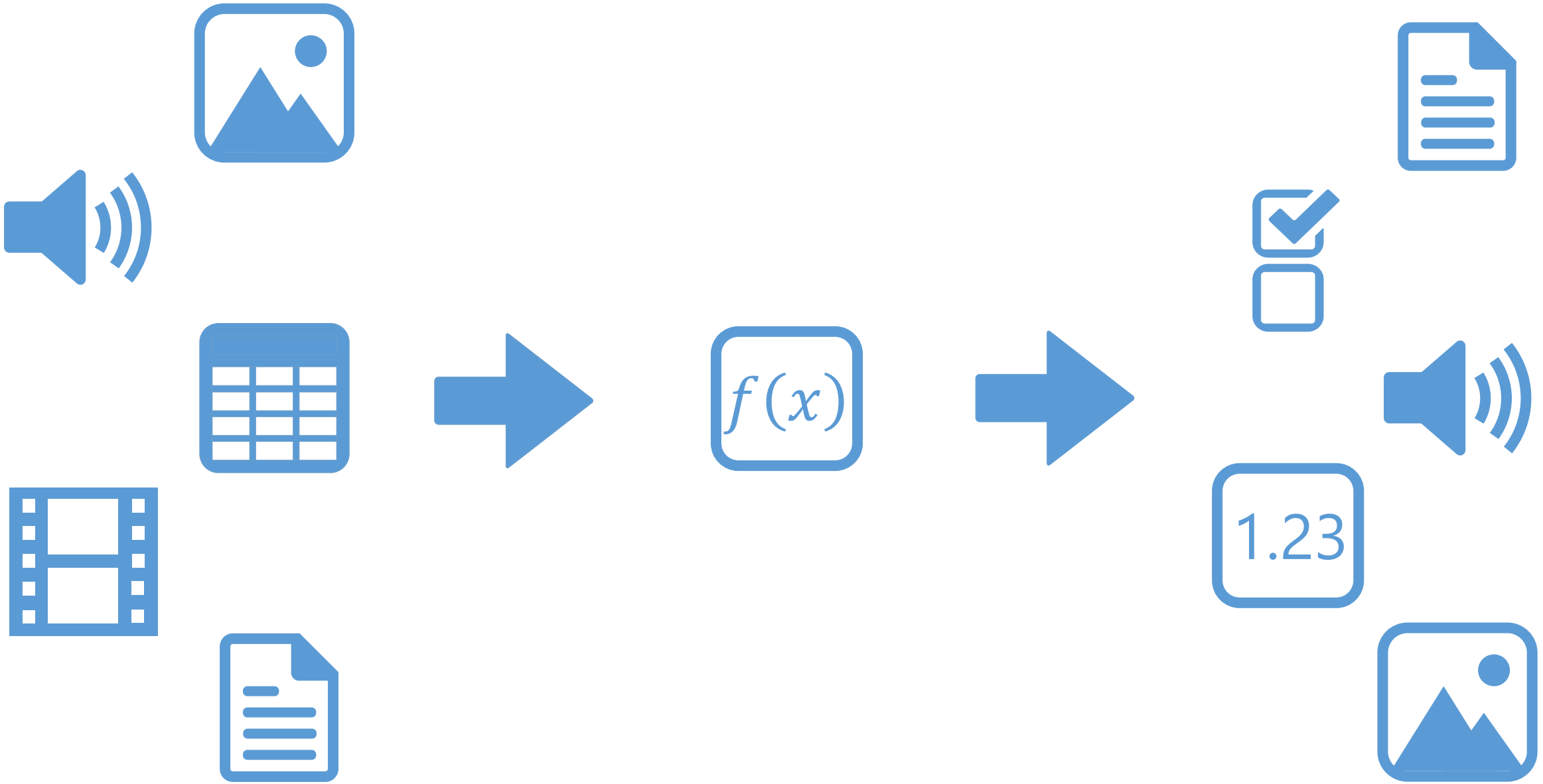


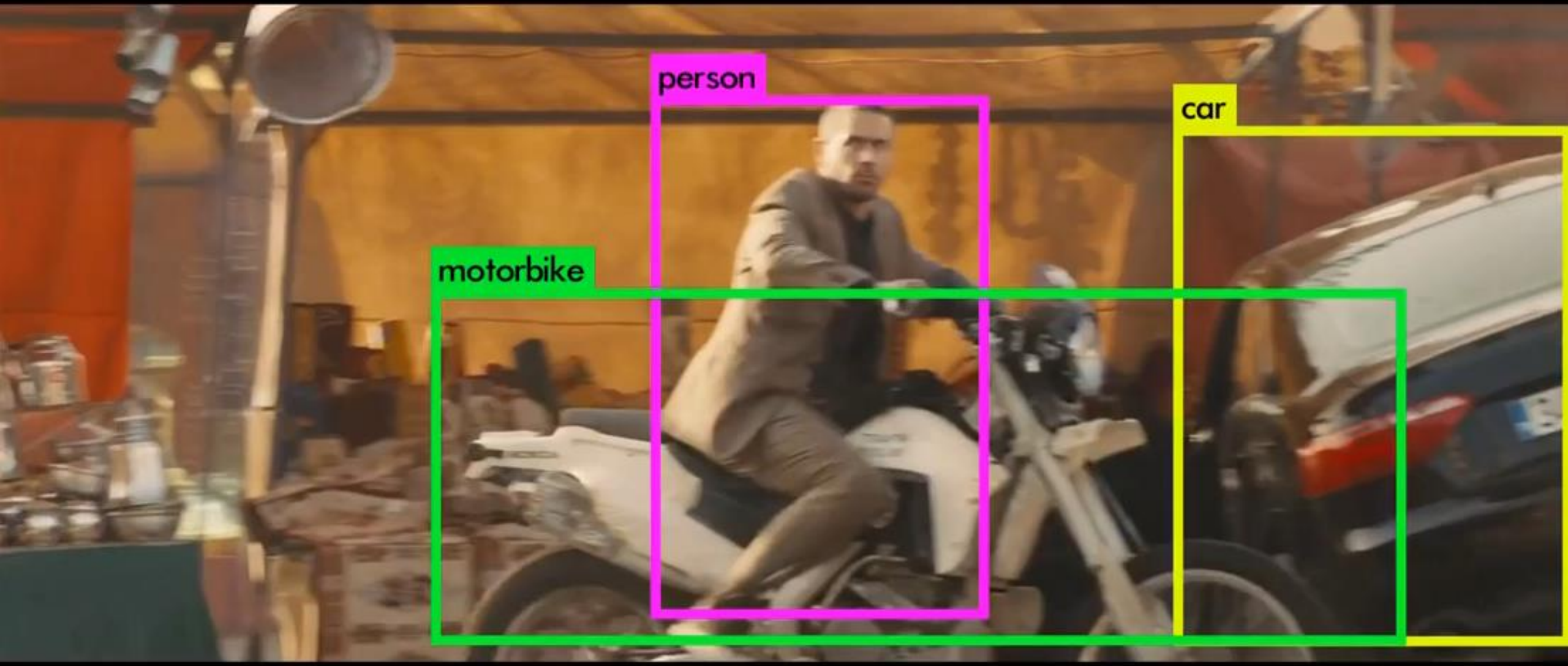
Unsupervised Learning

What can machine learning do?


$f(x)$
1.23


$f(x)$
1.23


$f(x)$
1.23

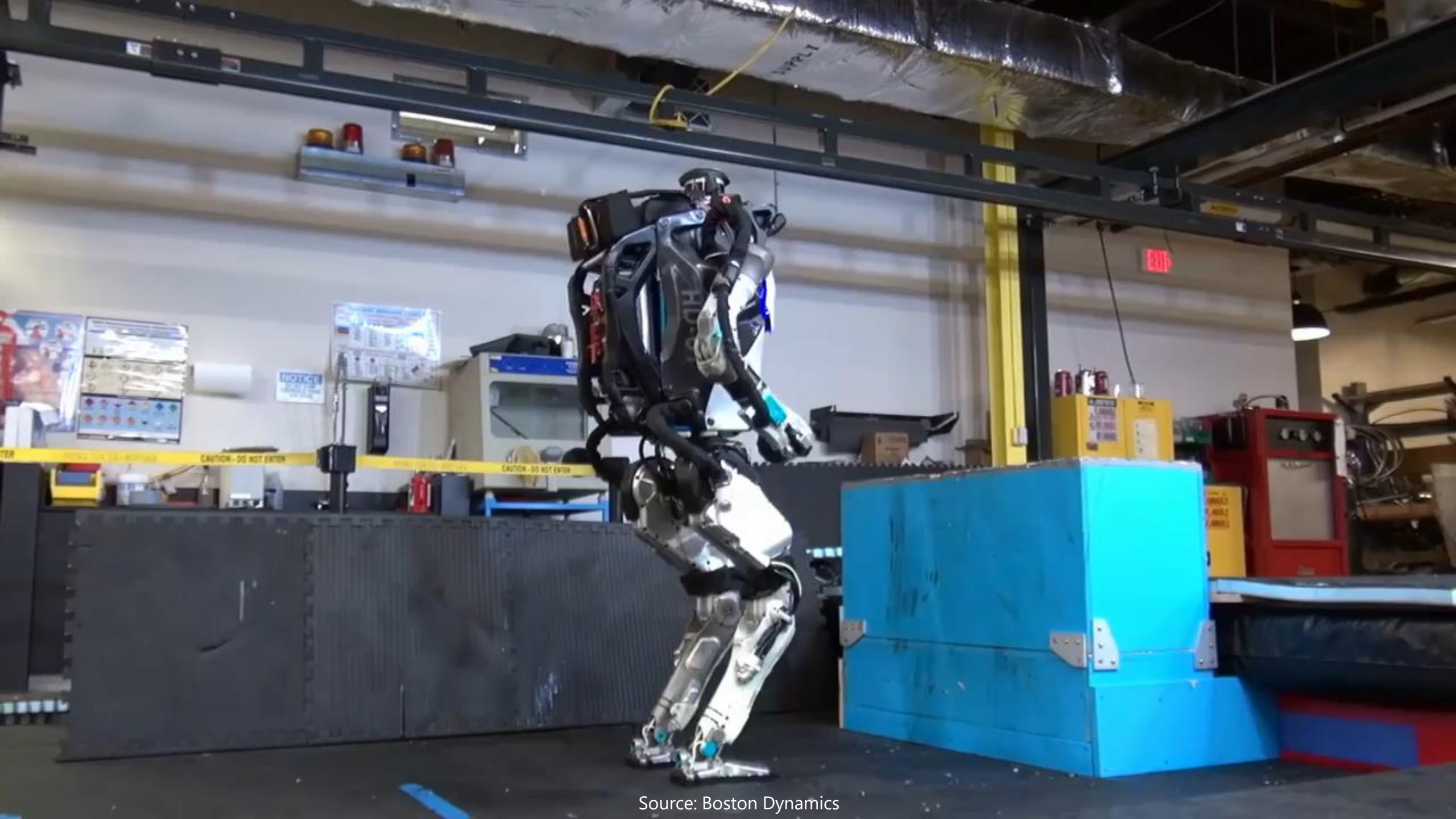


Source: YOLO: Real-Time Object Detection

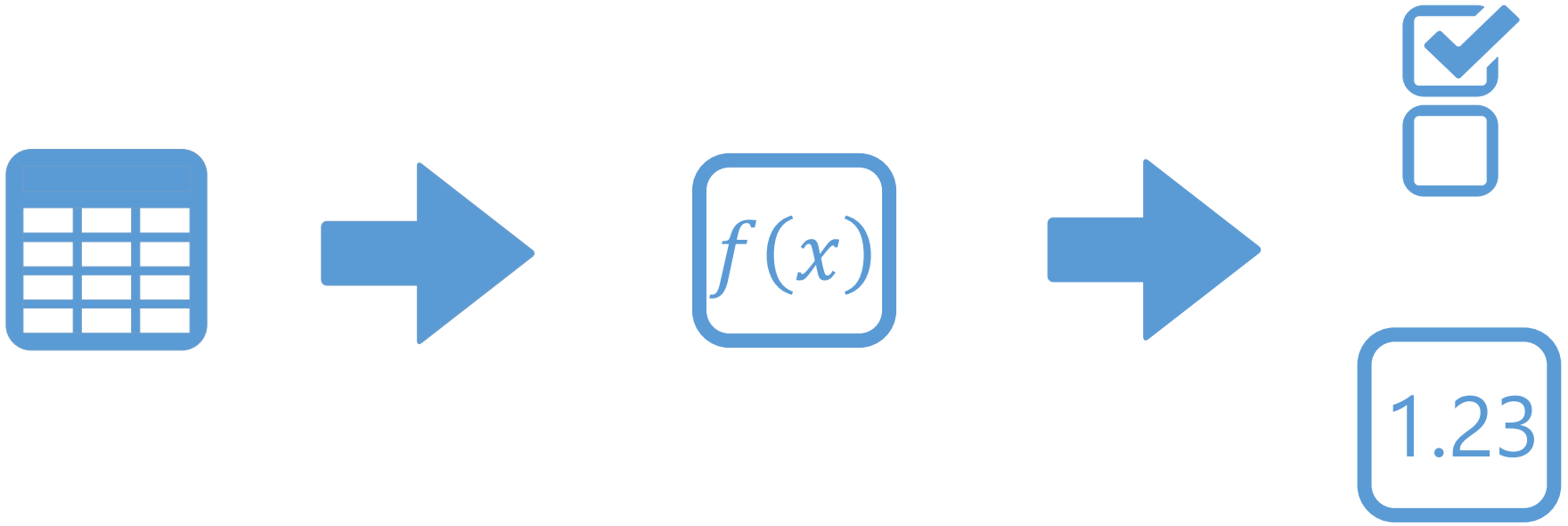

Source: Nvidia

Source: http://grail.cs.washington.edu/projects/AudioToObama/

Source: Google Deep Mind



Source: Boston Dynamics


$f(x)$
1.23

Disclaimer

Introduction to R

What is R?

Open source

Language and environment

Numerical and graphical

Cross platform

What is R?

Active development

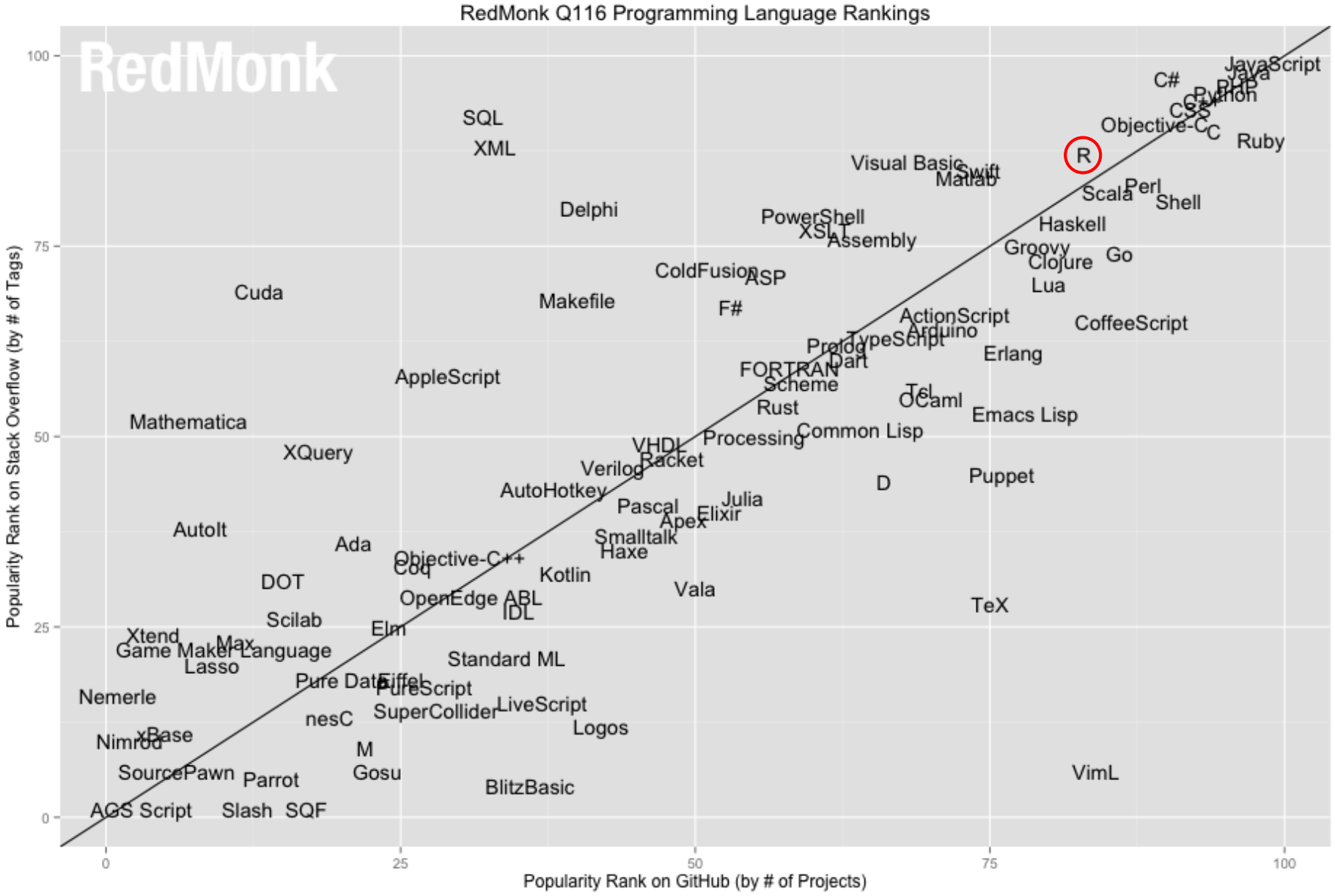
Large user community

Modular and extensible

9000+ extensions

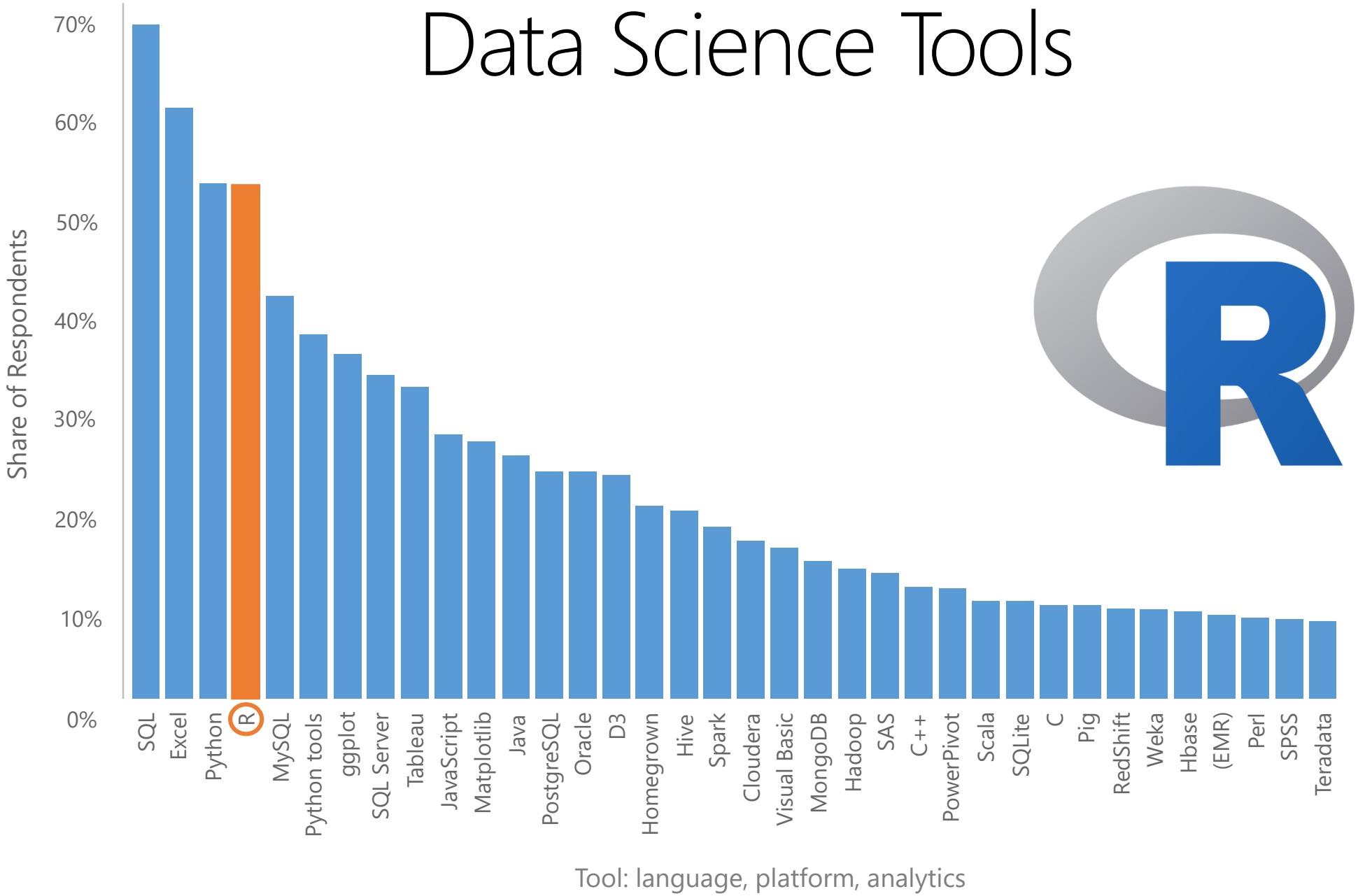

FREE


FREE



Source: http://redmonk.com/sogrady/2016/07/20/language-rankings-6-16/

Data Science Tools



Source: O'Reilly 2015 Data Science Salary Survey

RGui (64-bit)

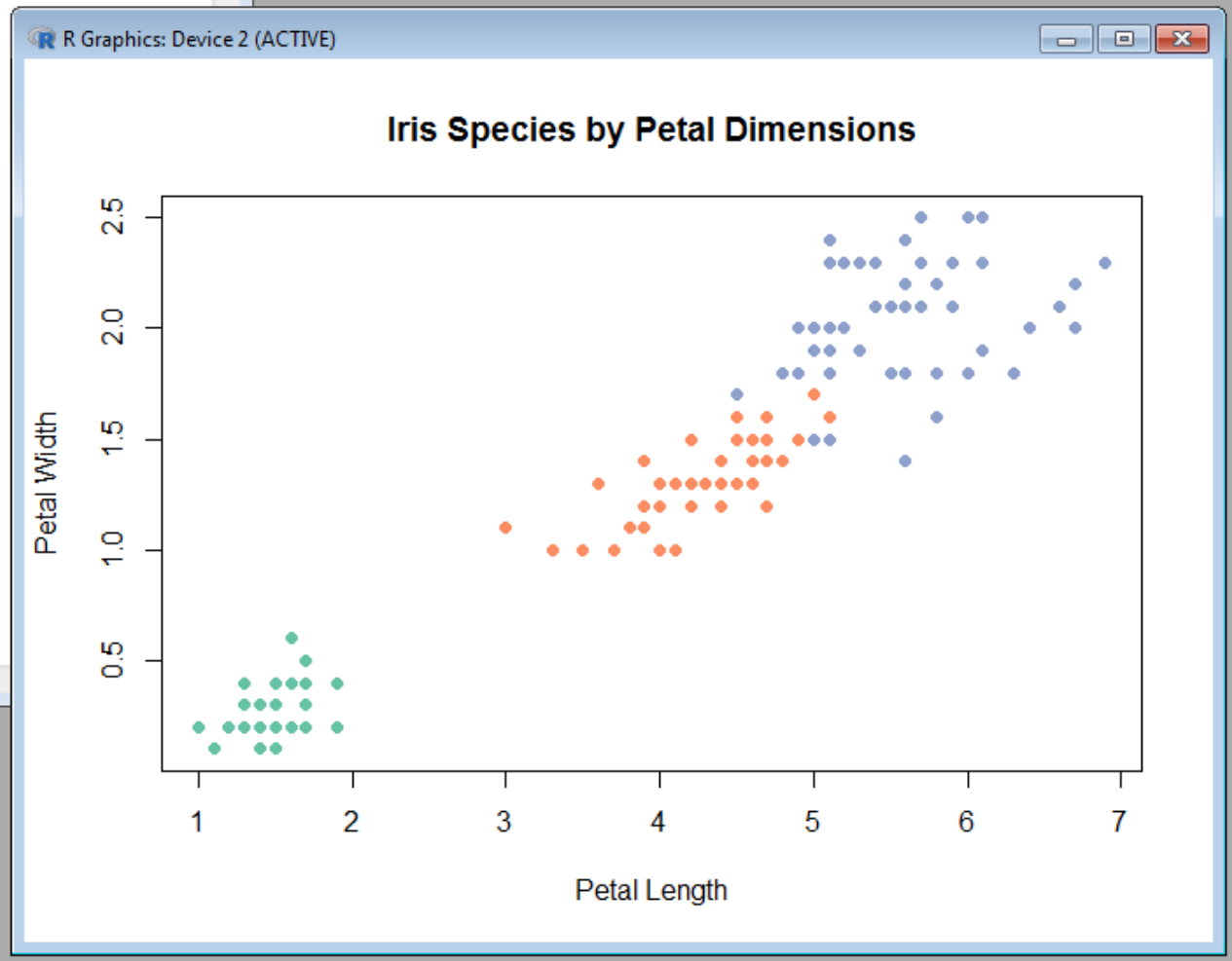
File History Resize Windows

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

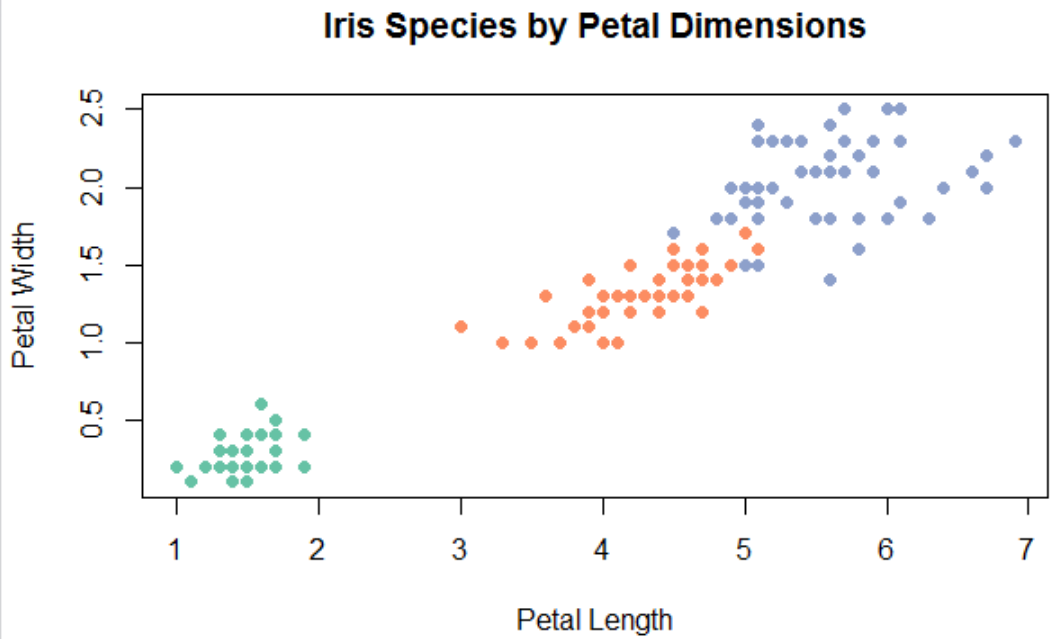
iris	150 obs. of 5 variables

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

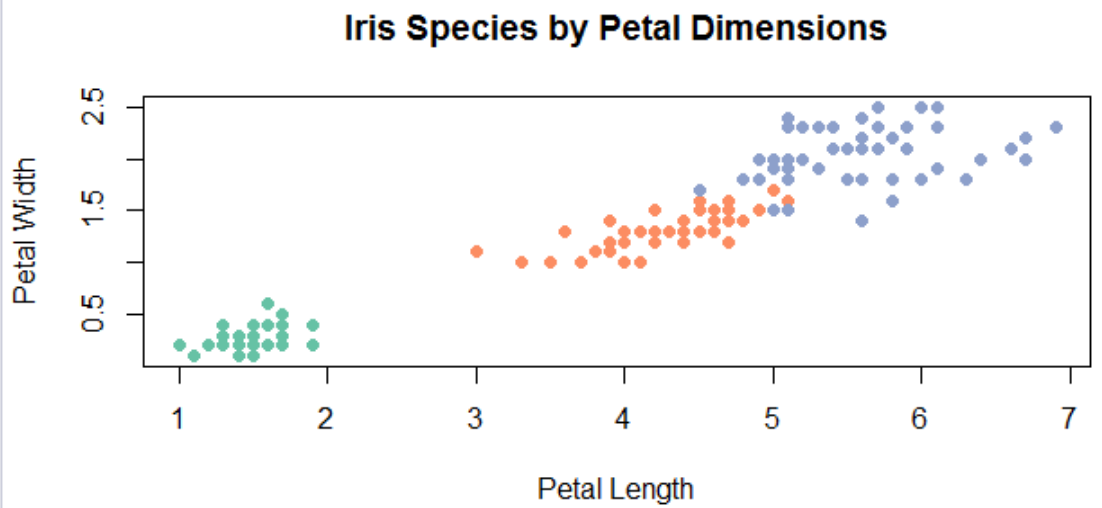
Search

.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

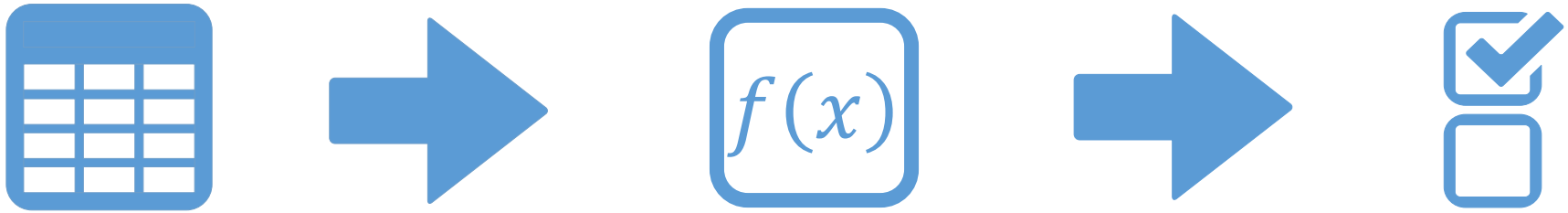
Error List | Output | Azure App Service Activity

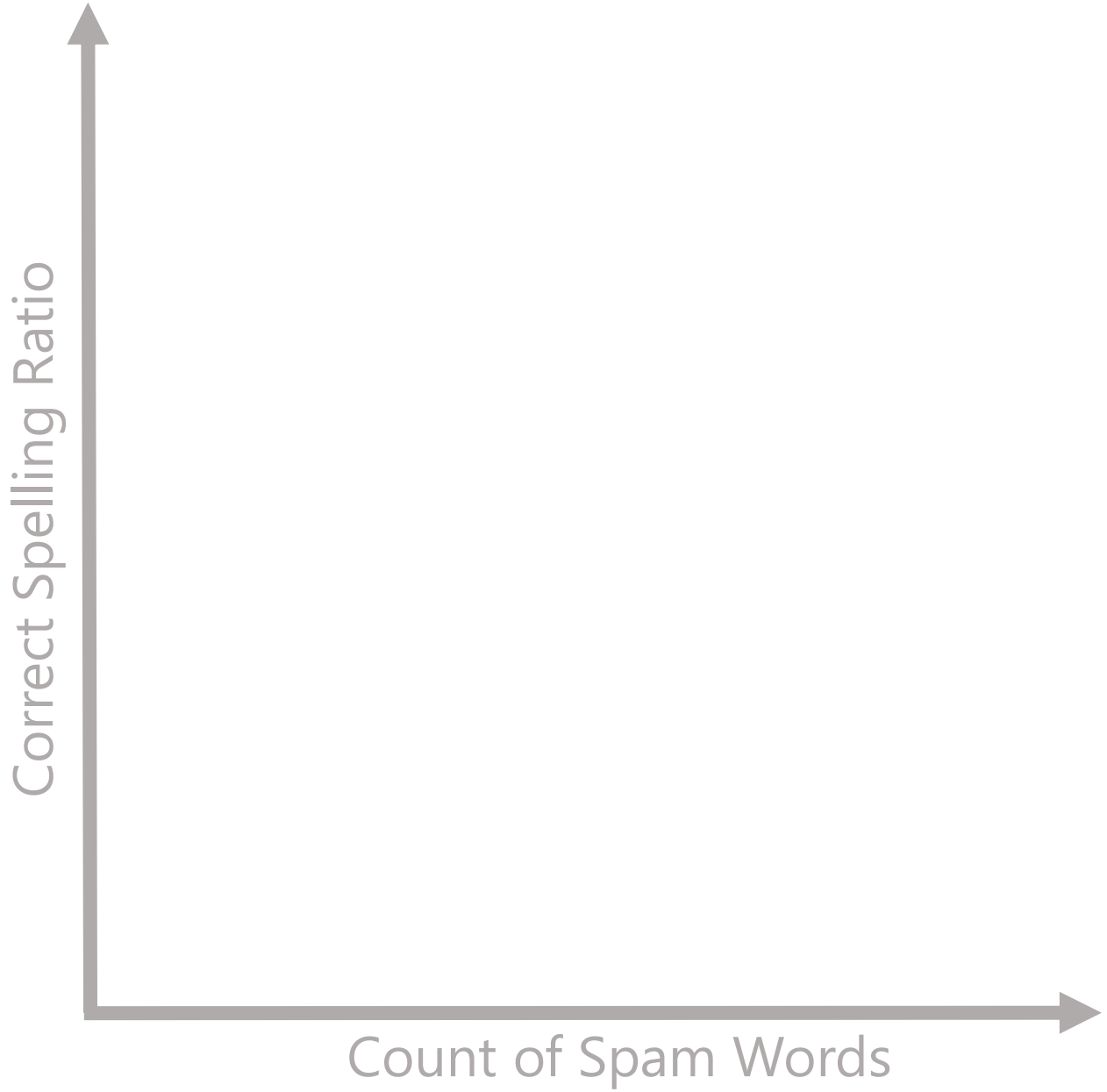
Ready | Ln 30 | Col 1 | Ch 1 | INS | 7 | 0 | Root | master

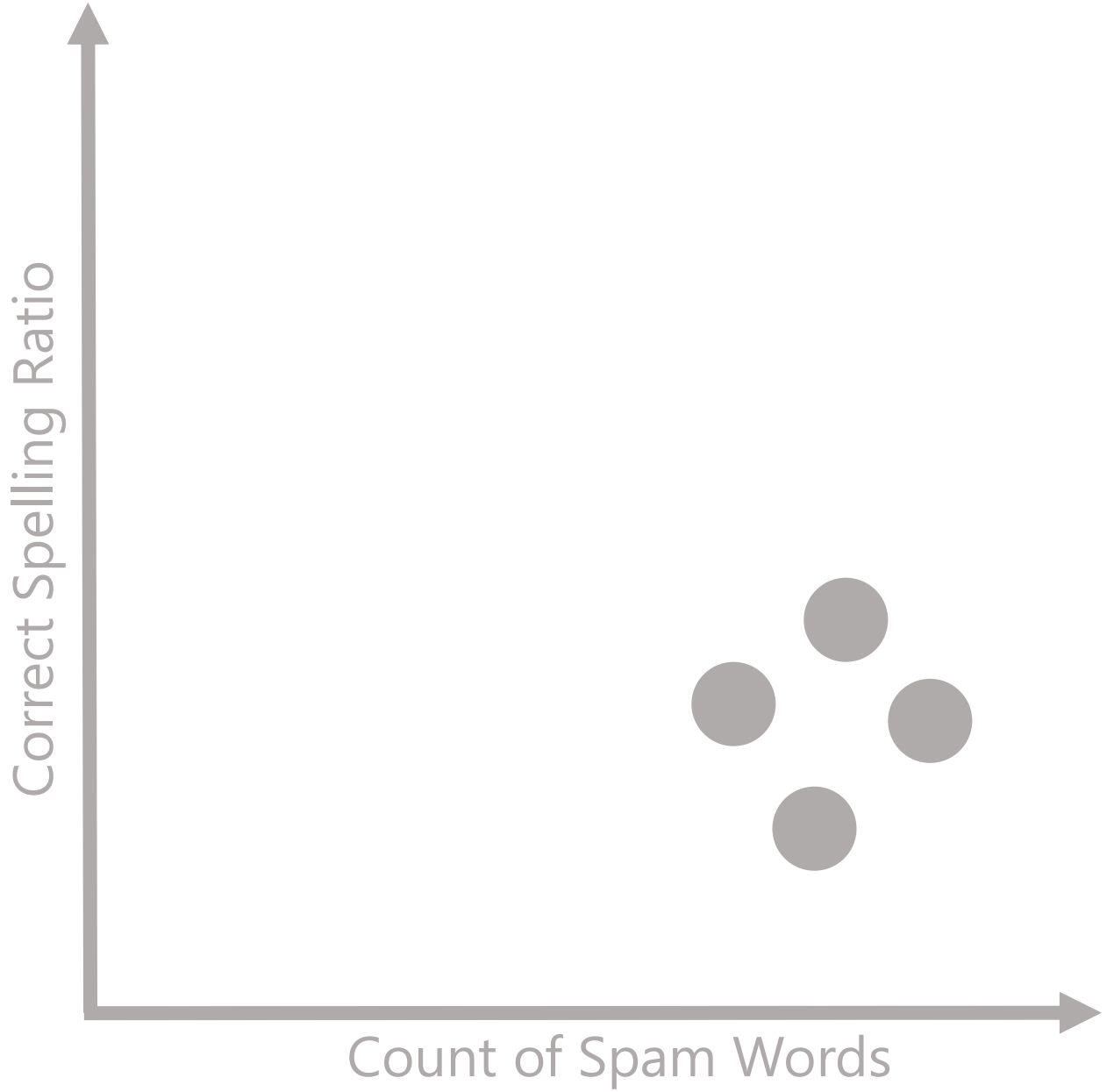
Demo 1
R Language Basics

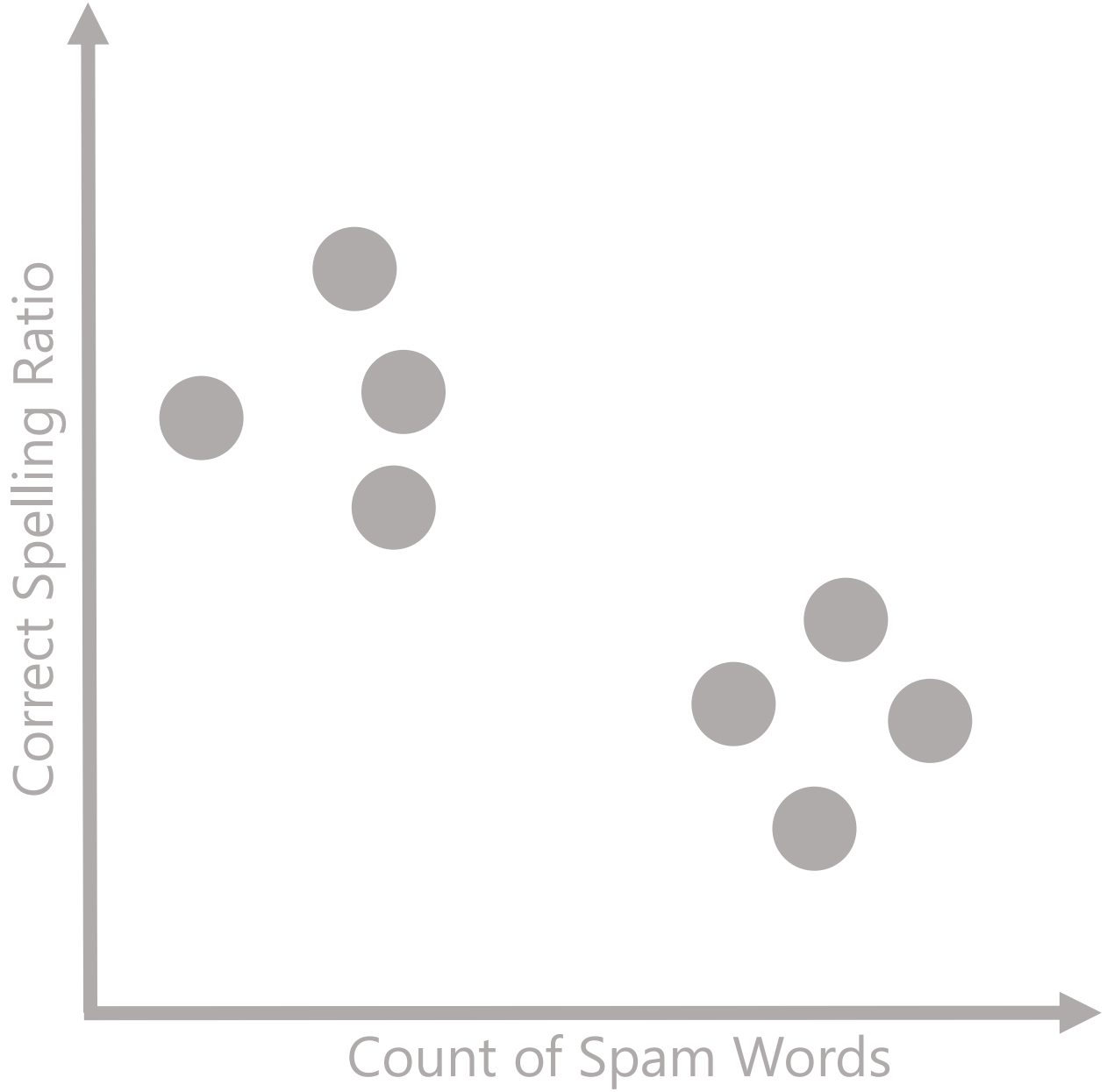
Lab 1
R Language Basics

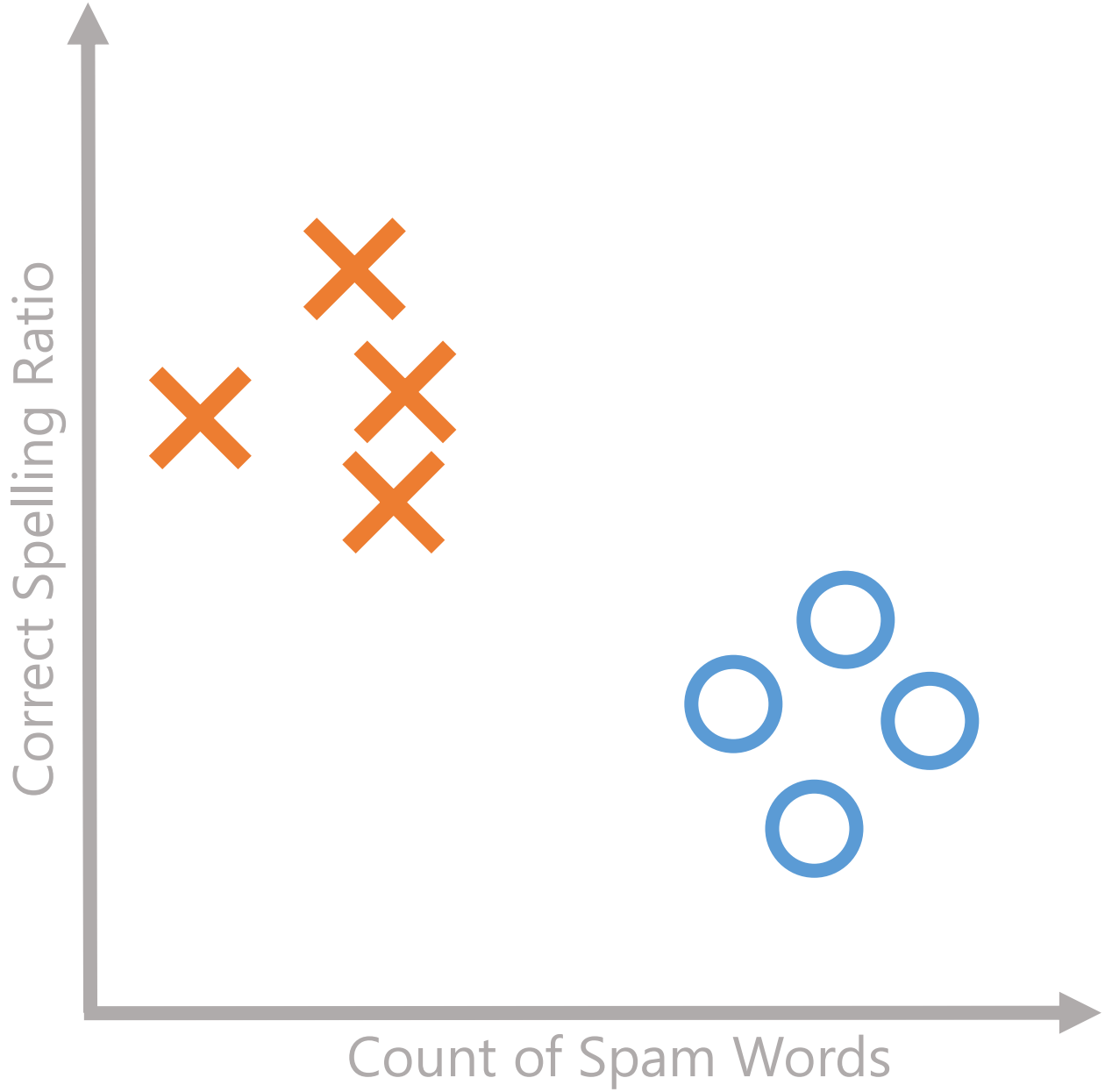
Classification

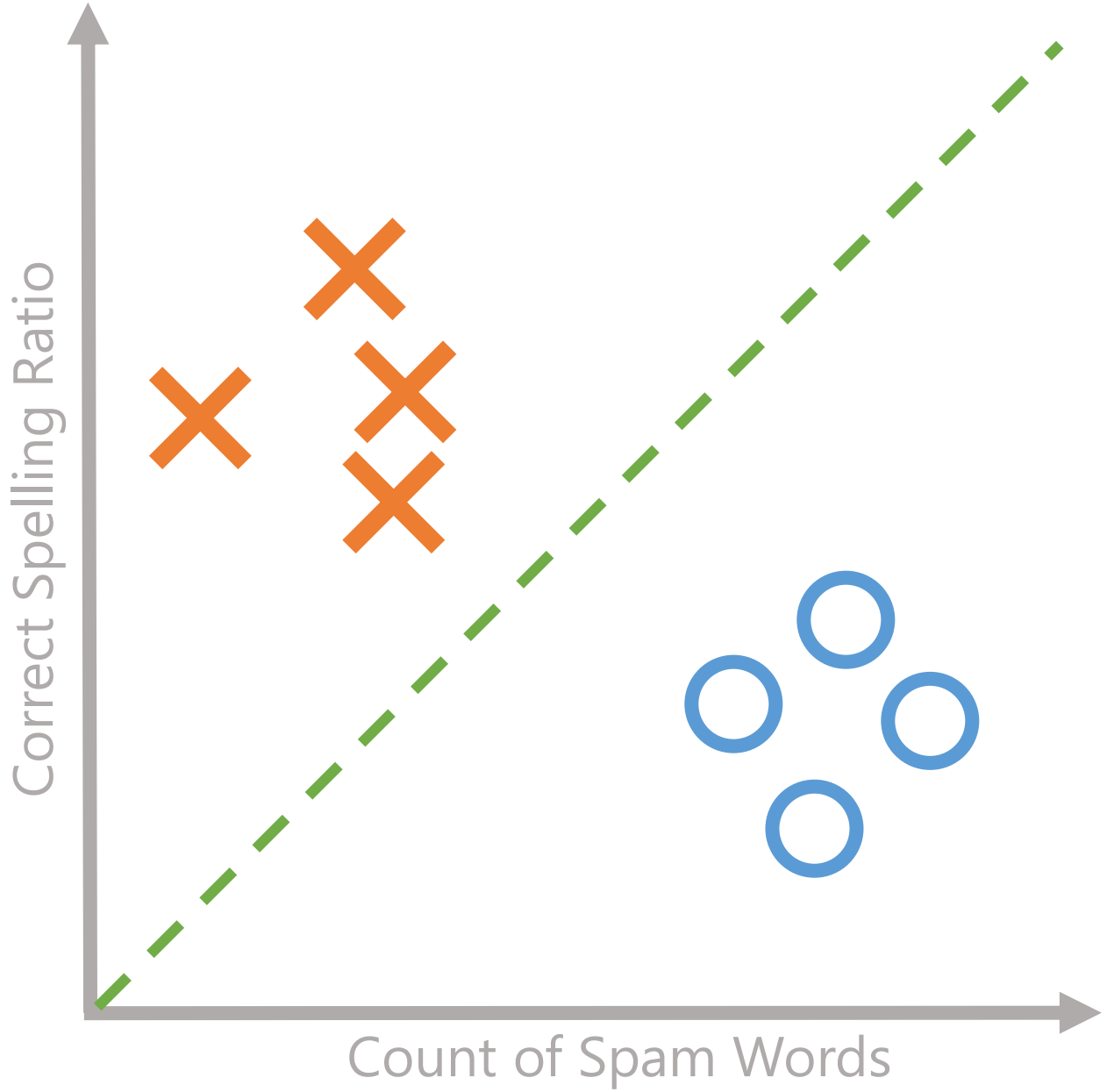

$f(x)$

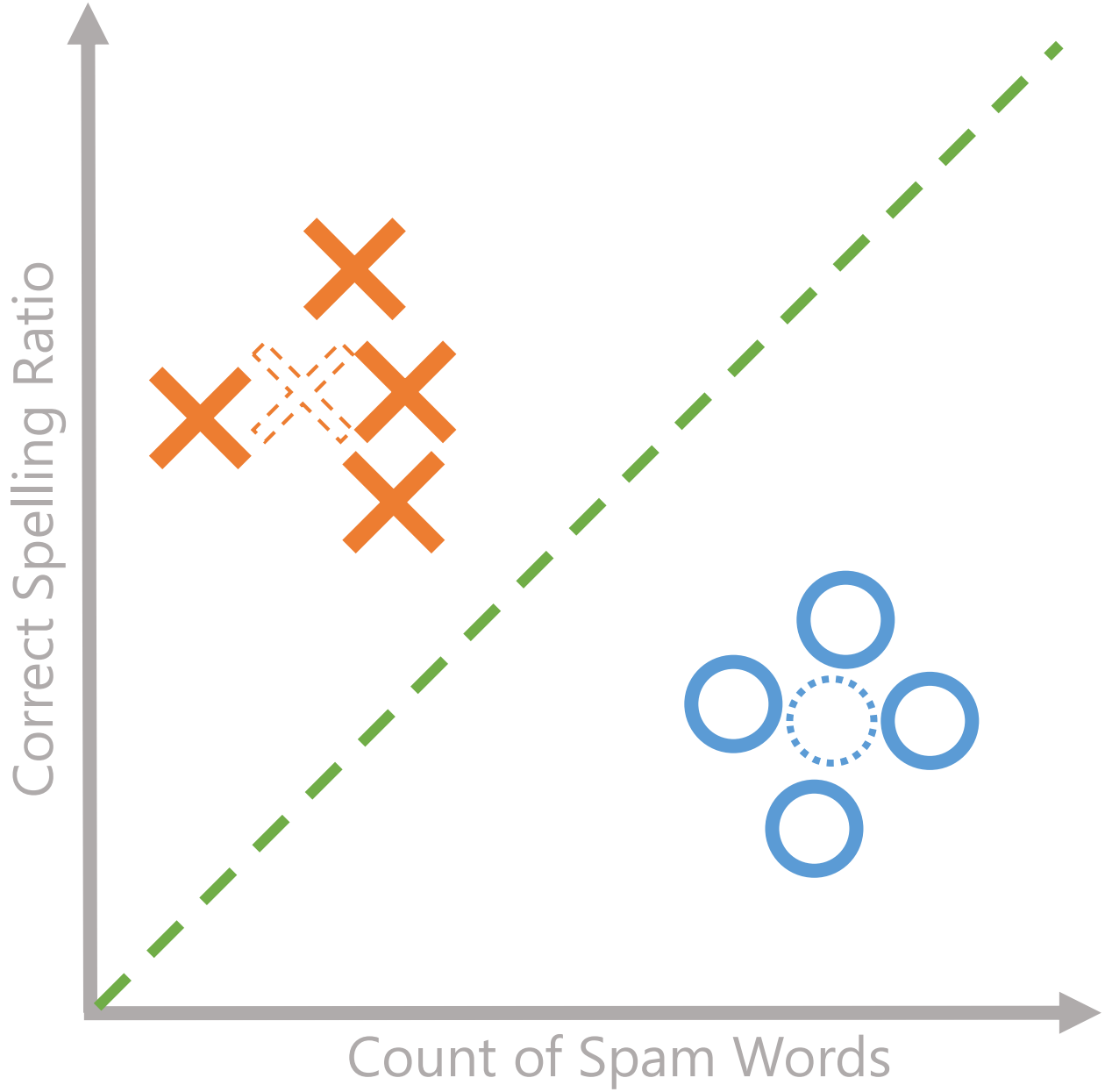

Correct Spelling Ratio
Count of Spam Words


Correct Spelling Ratio
Count of Spam Words


Correct Spelling Ratio
Count of Spam Words

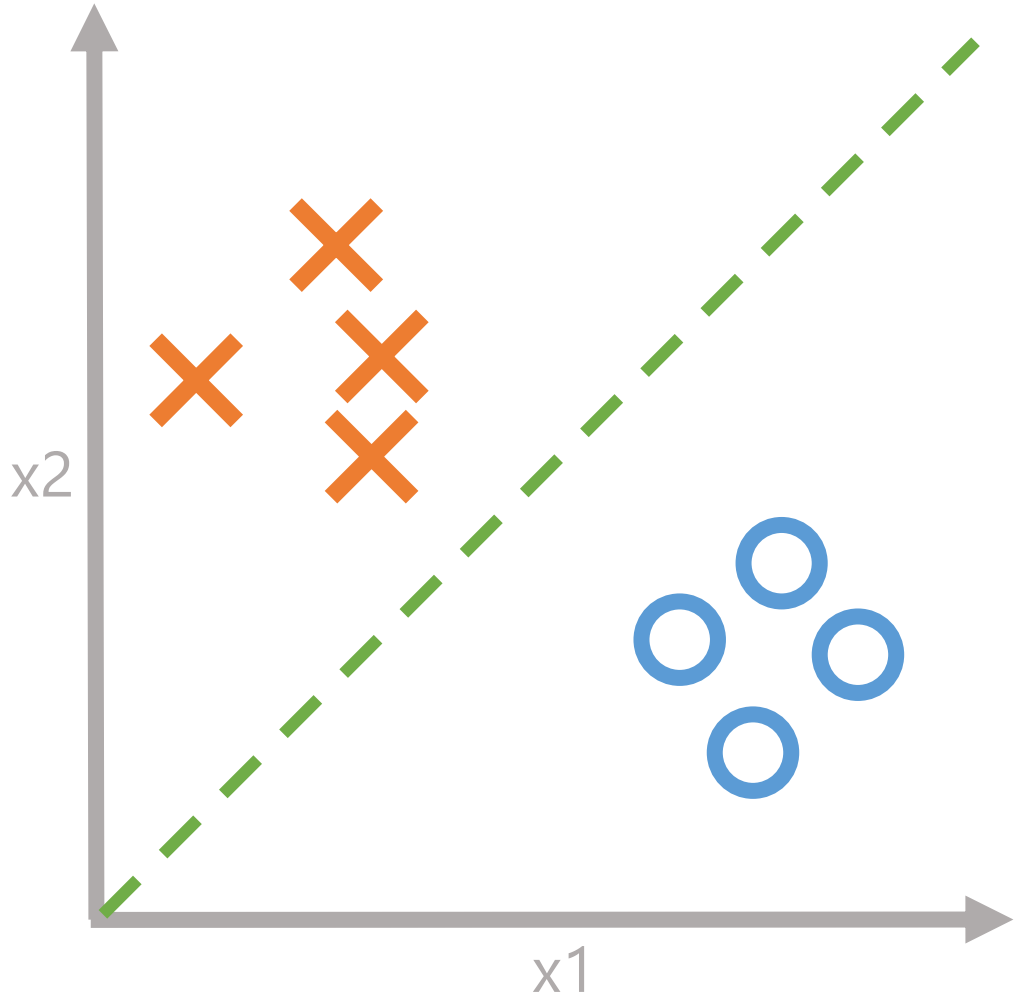

Correct Spelling Ratio
Count of Spam Words


Correct Spelling Ratio
Count of Spam Words

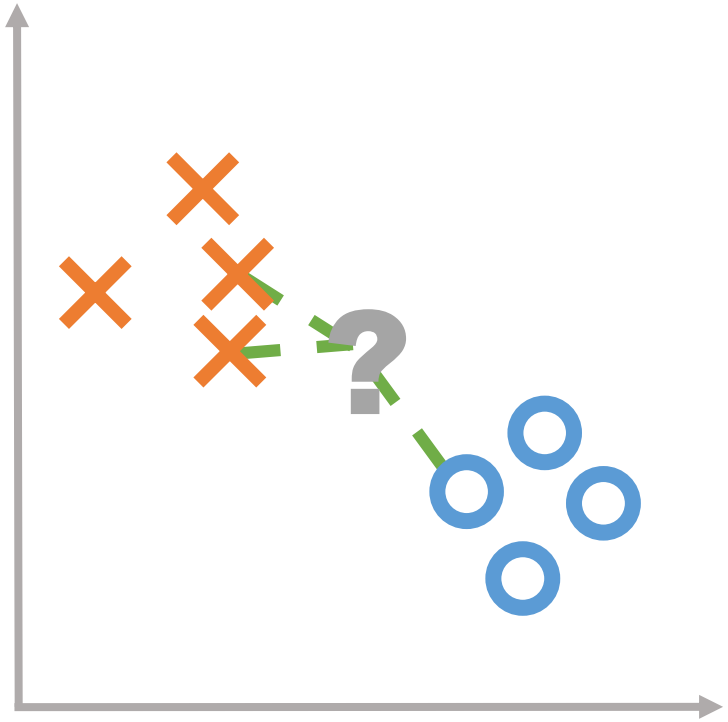

Correct Spelling Ratio
Count of Spam Words

Classification Algorithms

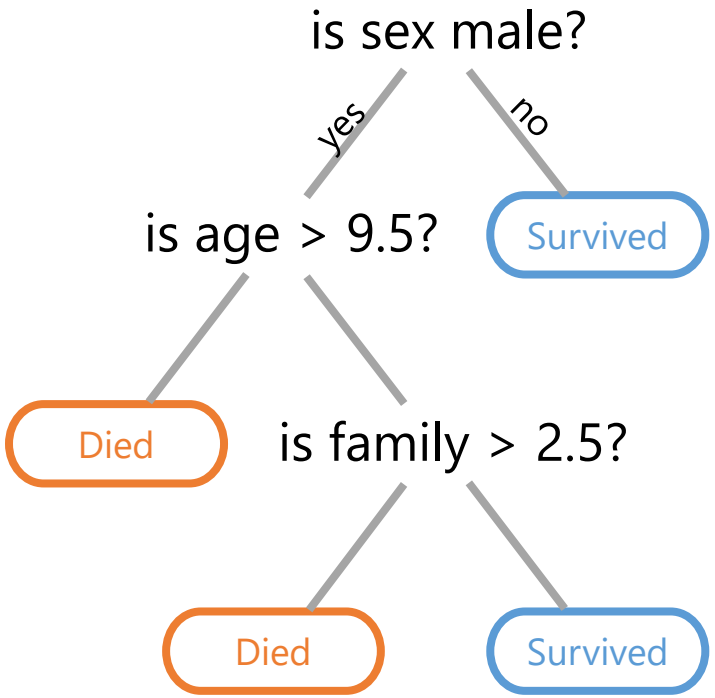
k-Nearest Neighbors
Decision Tree Classifier
Naïve Bayes Classifier
Support Vector Machine
Neural Network Classifier



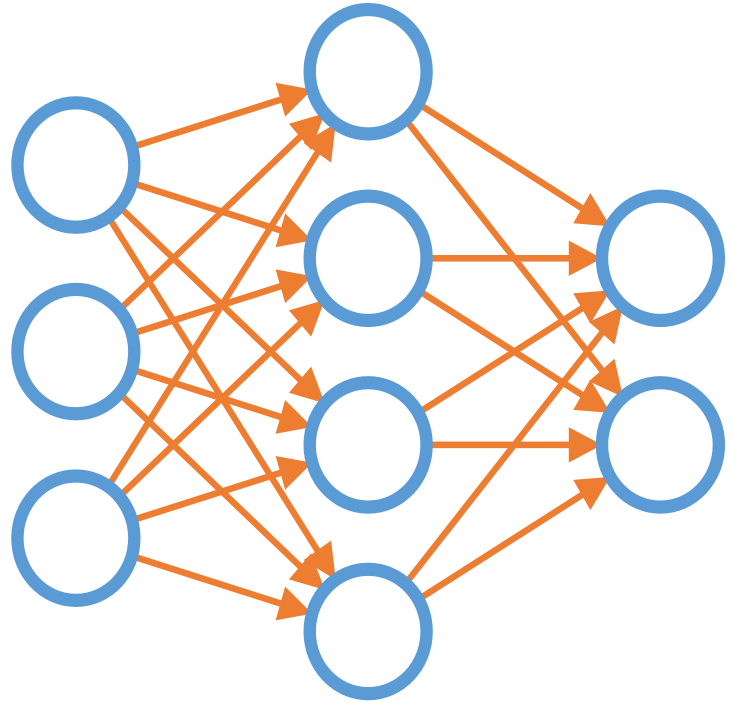
Classification Algorithms



k-Nearest Neighbors



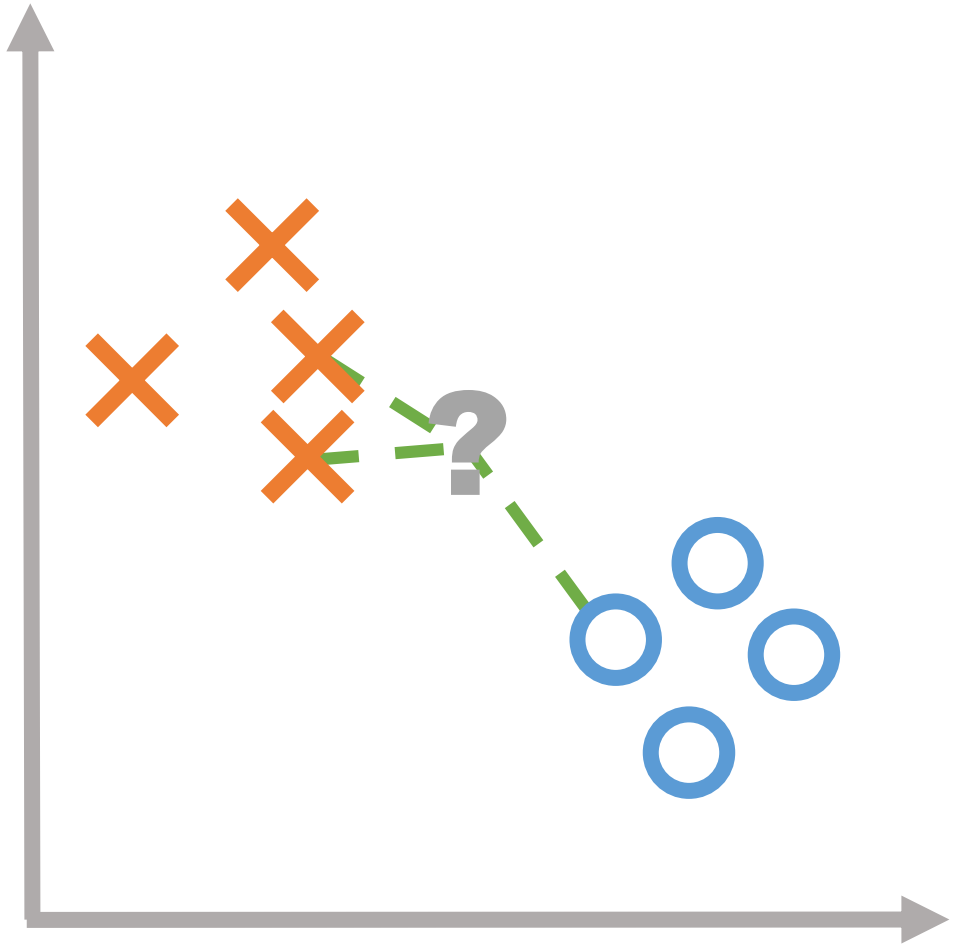
Decision Tree



Neural Network

K-Nearest Neighbors Classifier

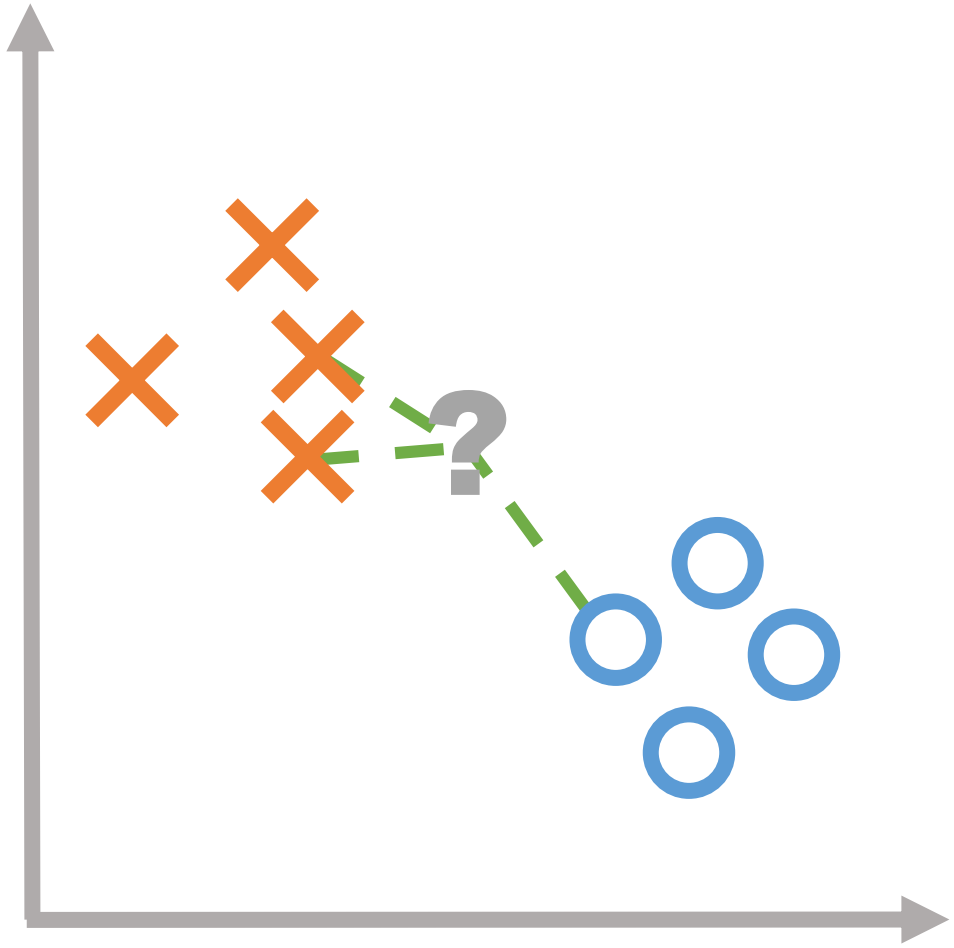
Supervised learning



K-Nearest Neighbors Classifier

Supervised learning

Uses class of neighbors

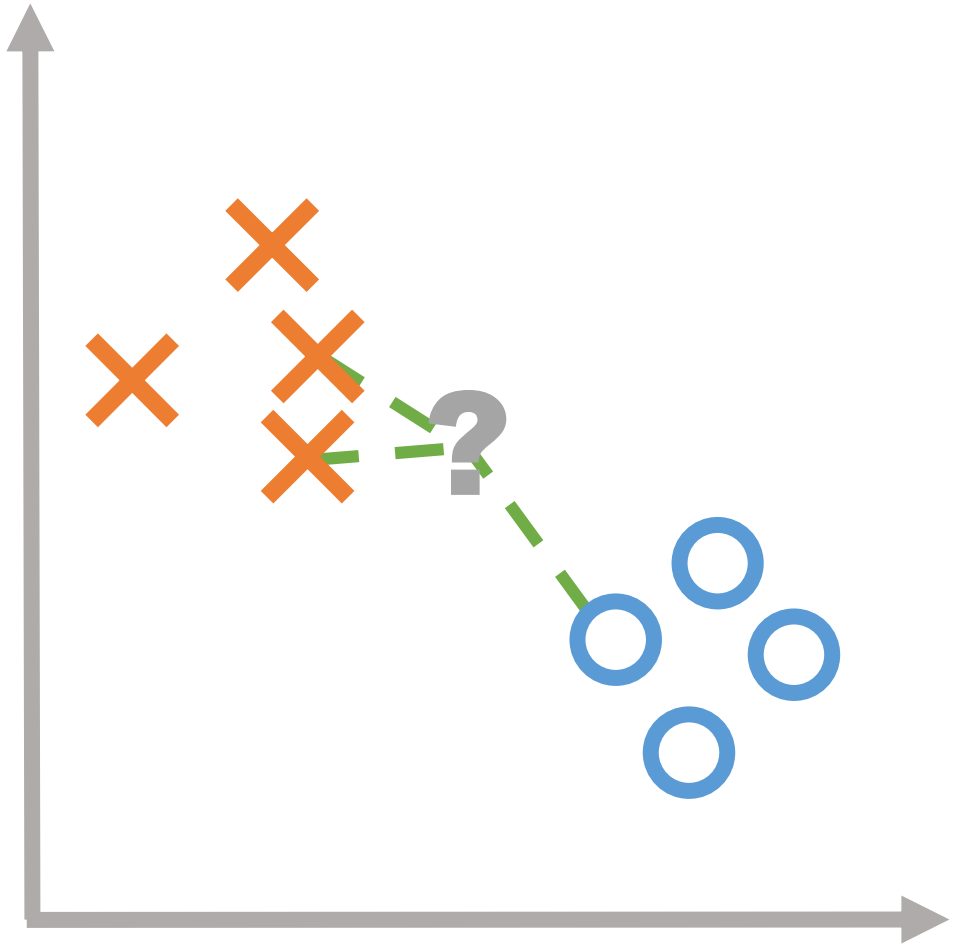


K-Nearest Neighbors Classifier

Supervised learning

Uses class of neighbors

k specifies how many



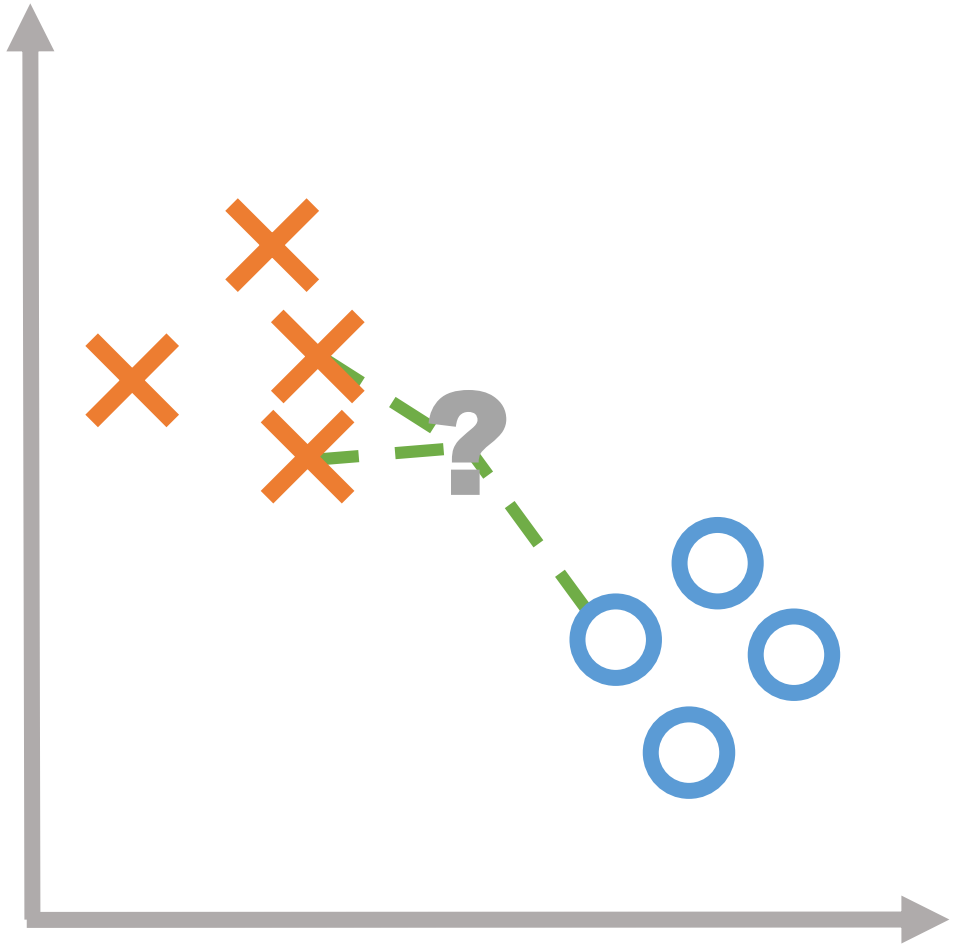
K-Nearest Neighbors Classifier

Supervised learning

Uses class of neighbors

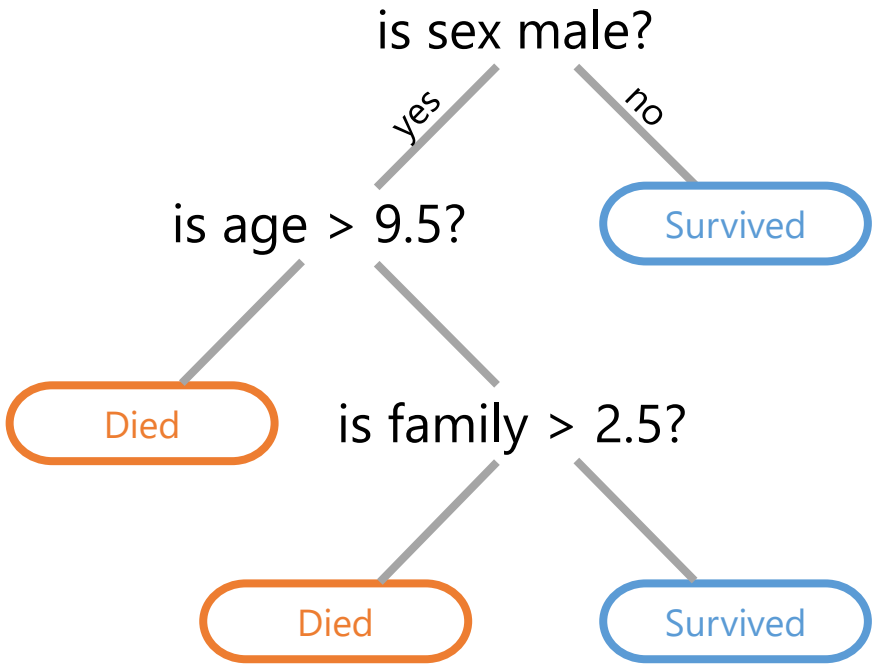
k specifies how many

Simple and easy



Decision Tree Classifier

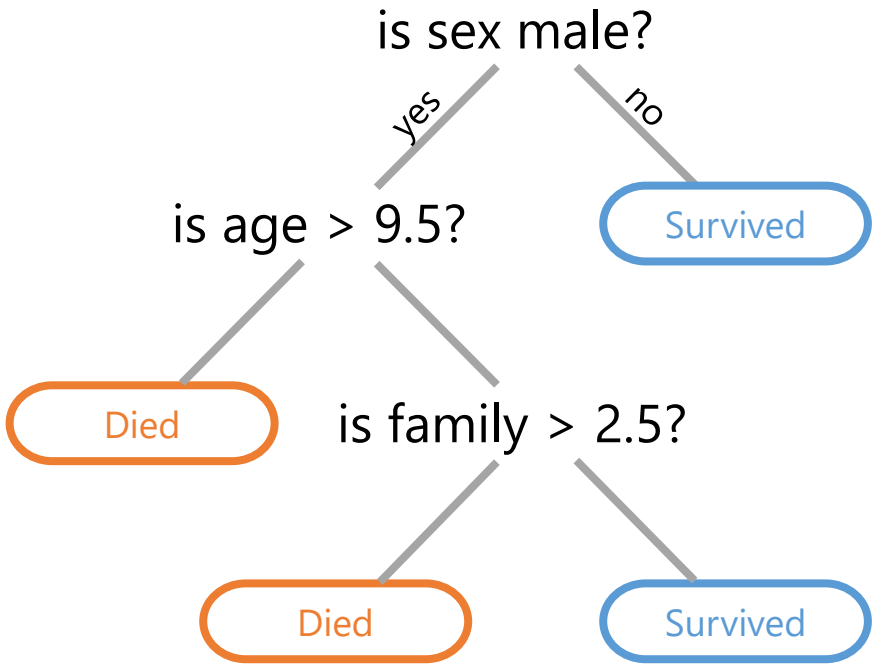
Supervised learning



Decision Tree Classifier

Supervised learning

Tree of decisions

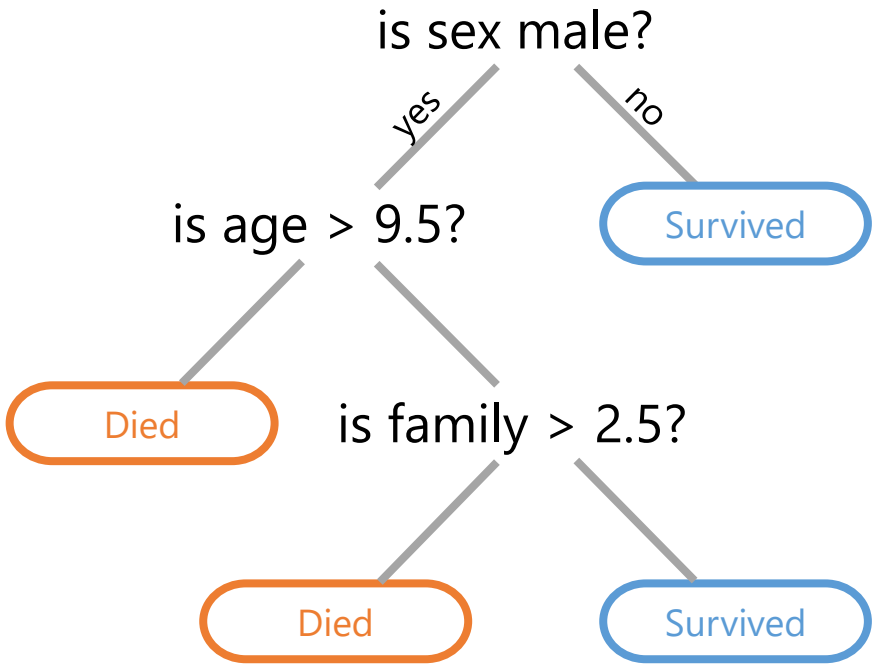


Decision Tree Classifier

Supervised learning

Tree of decisions

Information gain



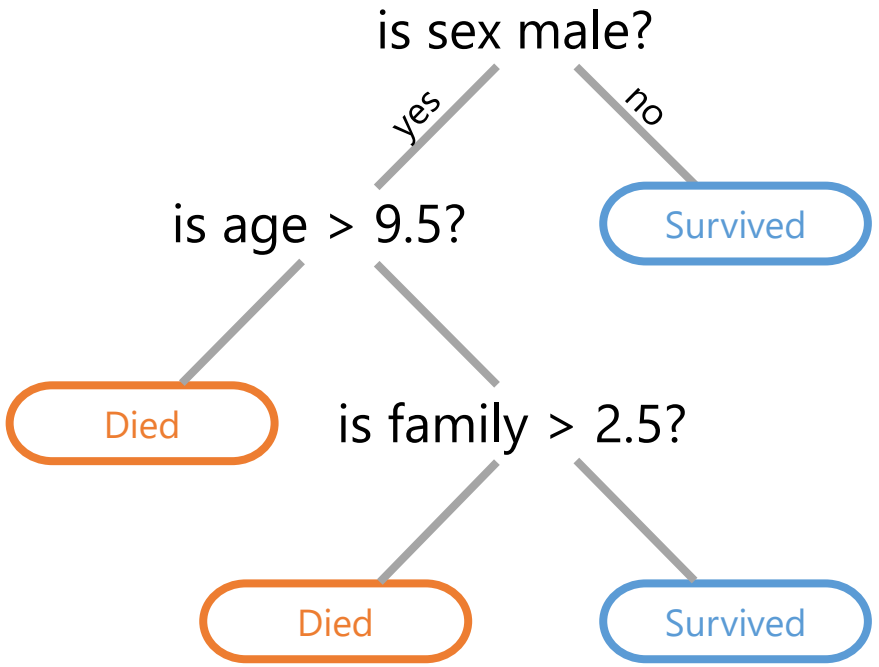
Decision Tree Classifier

Supervised learning

Tree of decisions

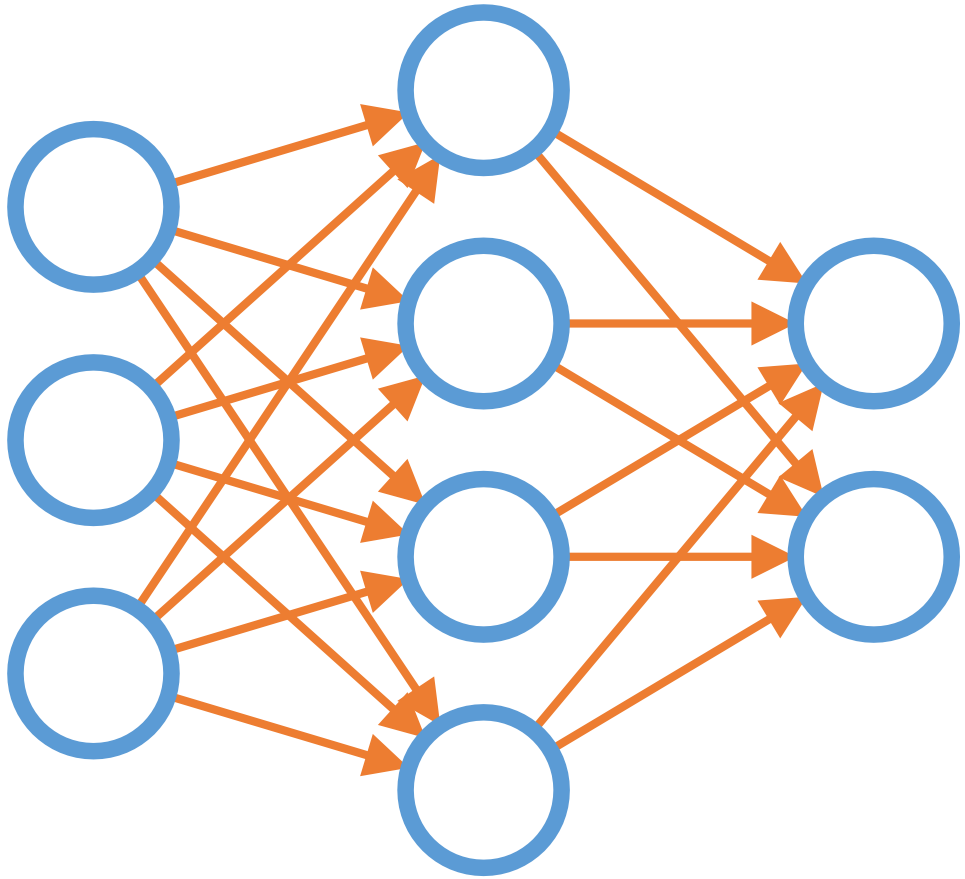
Information gain

Simple and easy



Neural Network Classifier

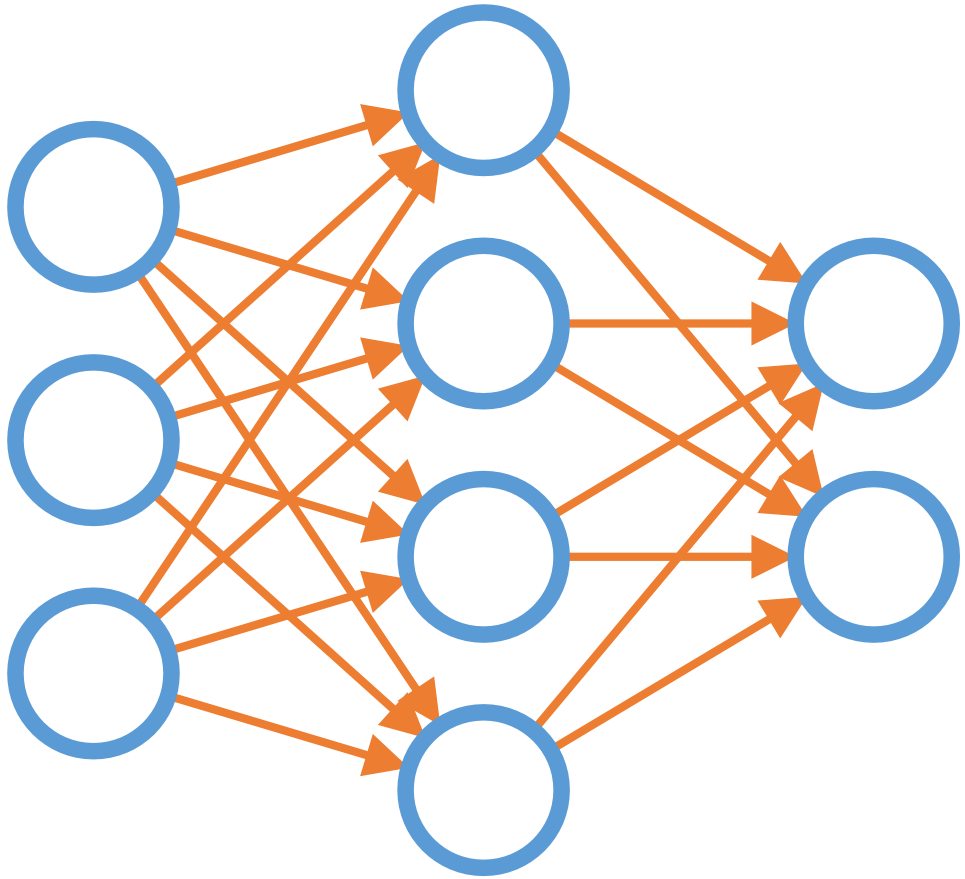
Supervised learning



Neural Network Classifier

Supervised learning

Neurons in a brain

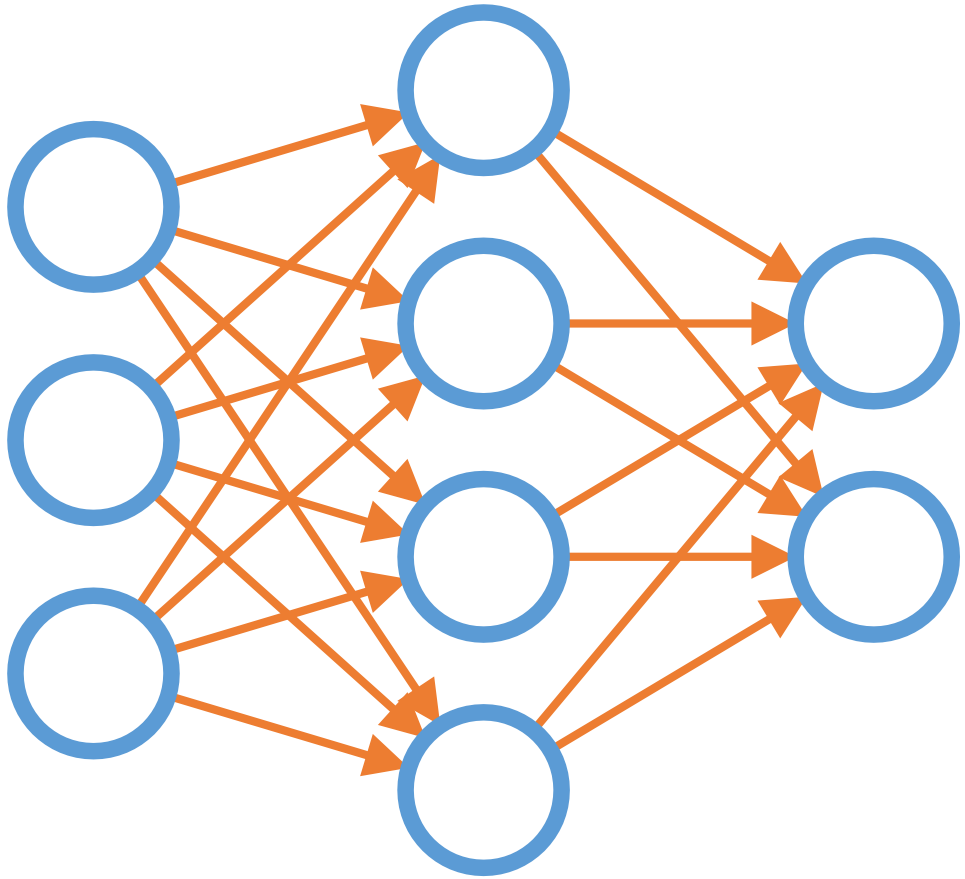


Neural Network Classifier

Supervised learning

Neurons in a brain

Complex



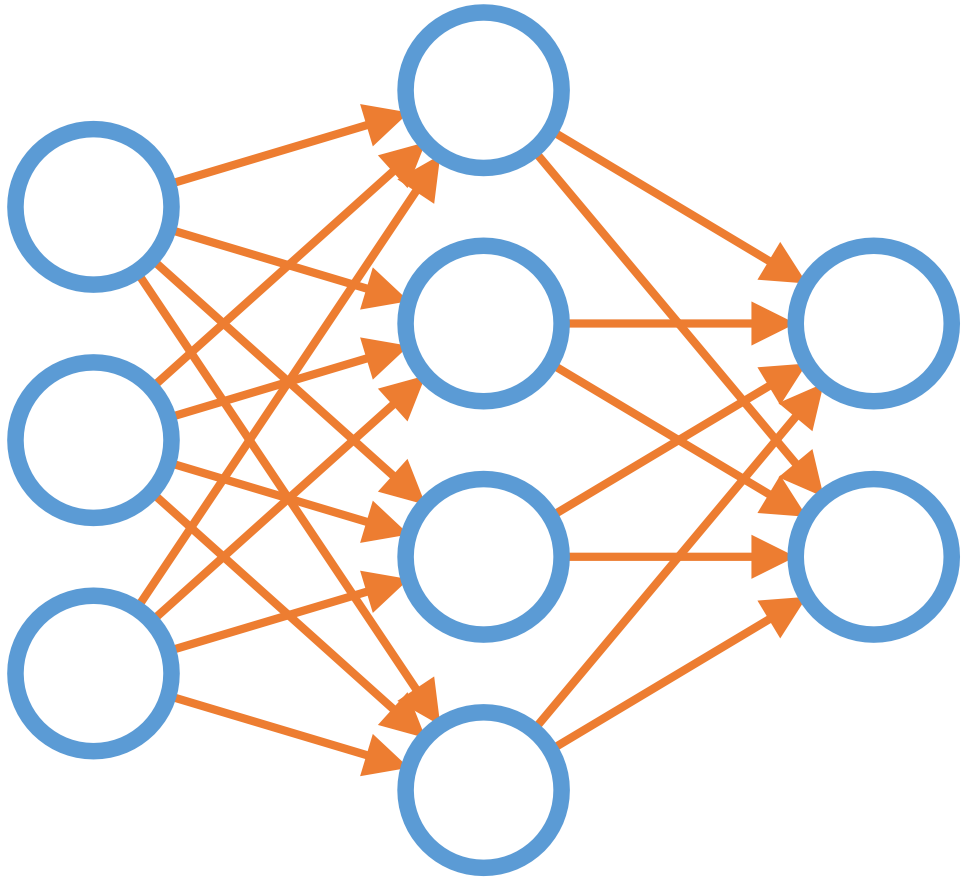
Neural Network Classifier

Supervised learning

Neurons in a brain

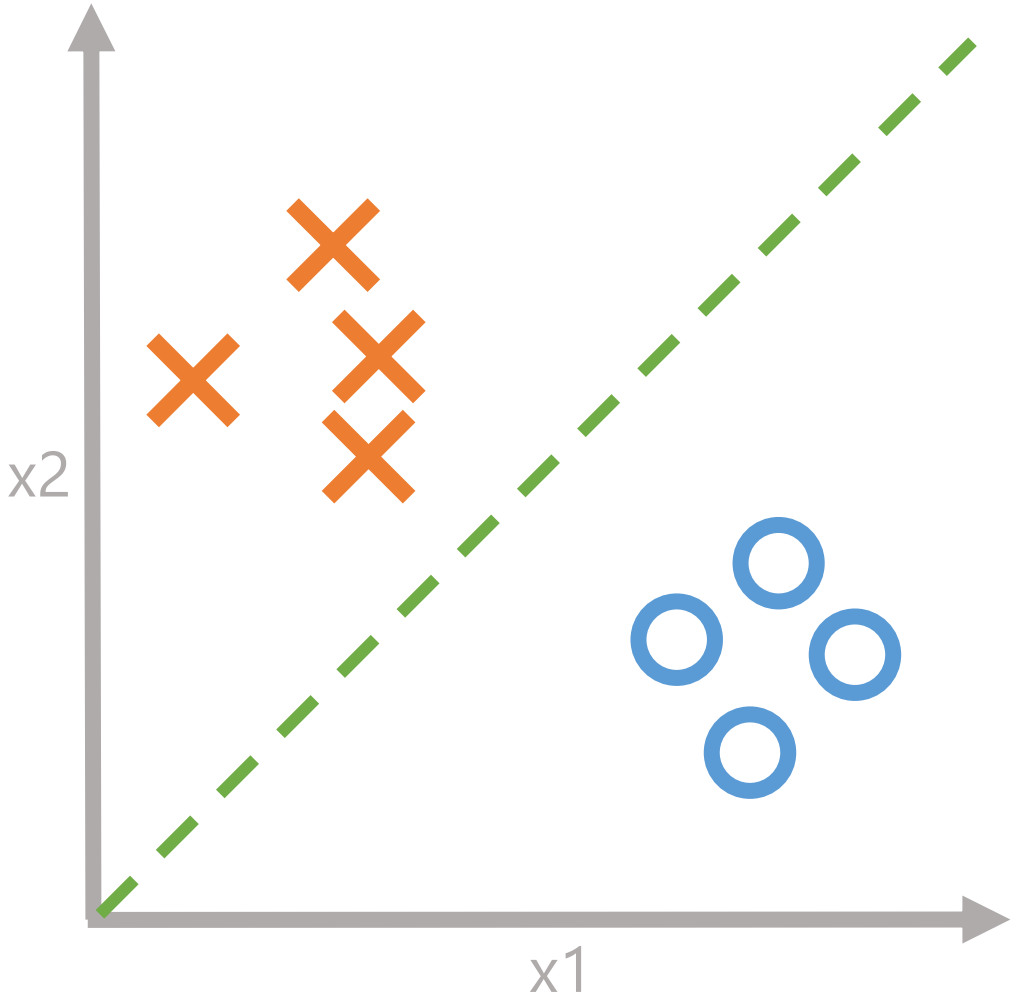
Complex

Not transparent



Real-World Examples

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



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
…		…	…	…

Demo 2 - Classification

Goal: Predict species based on petal and sepal measurements

Insurance Policy Data Set

Insurance Policy Risk							
Gender	State	State Rate	Height	Weight	BMI	Age	Risk
Male	MA	0.01	184	67.8	20.0	77	High
Male	VA	0.14	163	89.4	33.6	82	High
Female	NY	0.09	170	81.2	28.1	31	Low
Male	TN	0.12	175	99.7	32.6	39	Low
Female	FL	0.11	184	72.1	21.3	68	High
…		…	…	…	…	…	…

Insurance Policy Data Set

Insurance Policy Rates							
Gender	State	State Rate	Height	Weight	BMI	Age	Rate
Male	MA	0.01	184	67.8	20.0	77	0.33
Male	VA	0.14	163	89.4	33.6	82	0.87
Female	NY	0.09	170	81.2	28.1	31	0.01
Male	TN	0.12	175	99.7	32.6	39	0.02
Female	FL	0.11	184	72.1	21.3	68	0.15
…		…	…	…	…	…	…

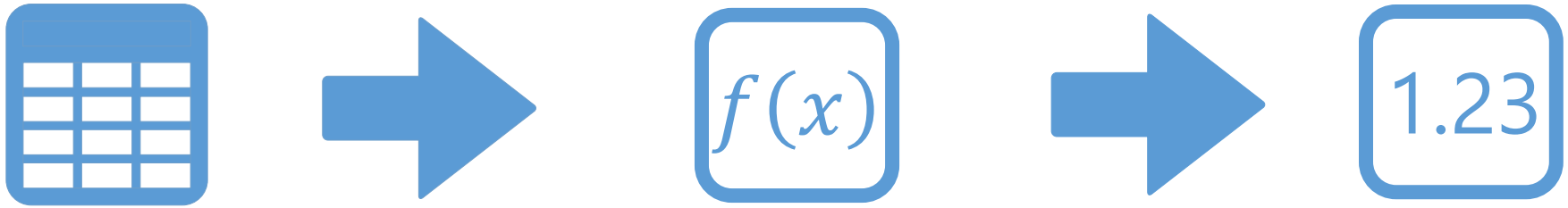
Lab 2A – Classification (Easy)

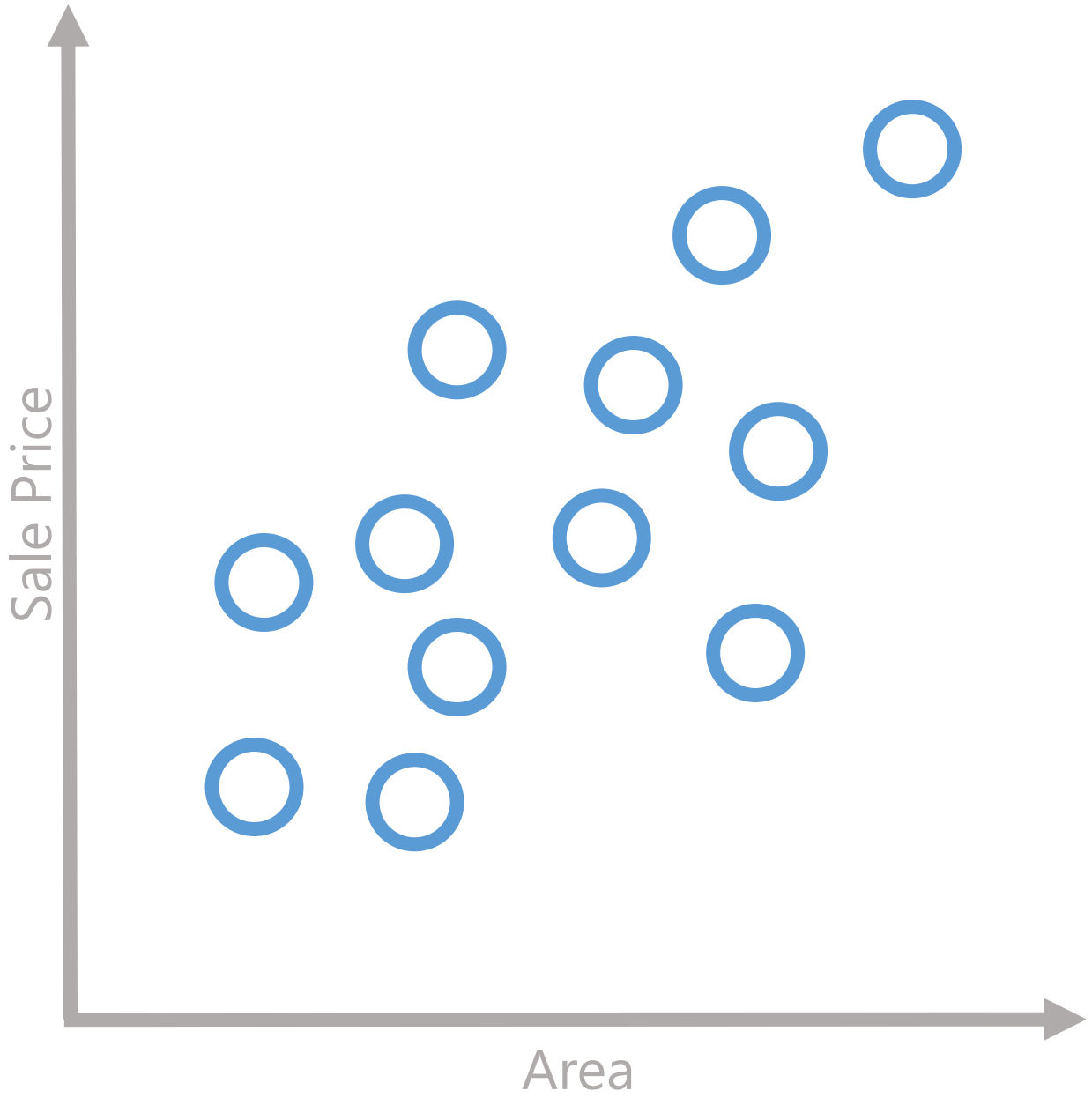
Goal: Predict species based on petal and sepal measurements

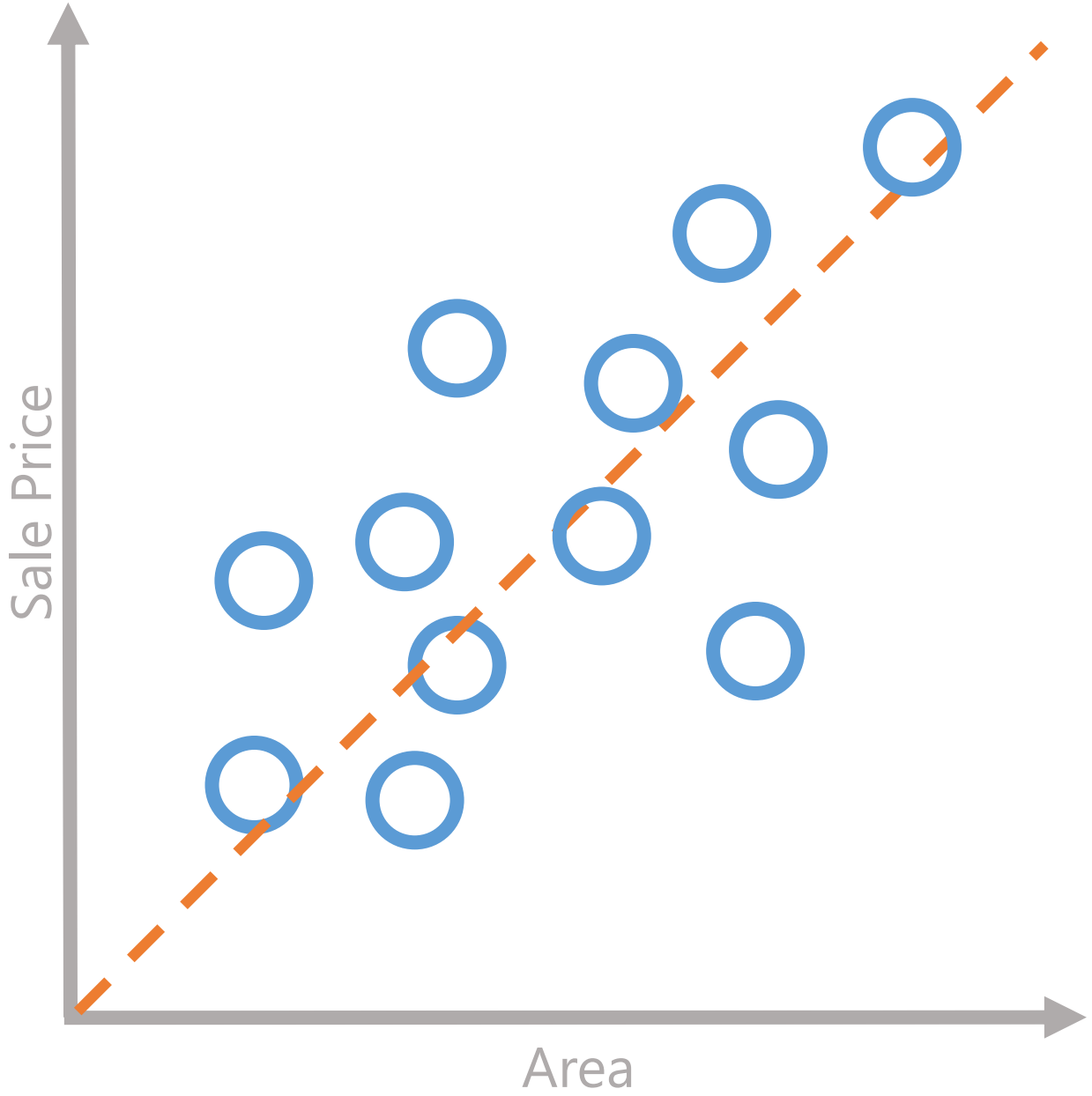
Lab 2B – Classification (Hard)

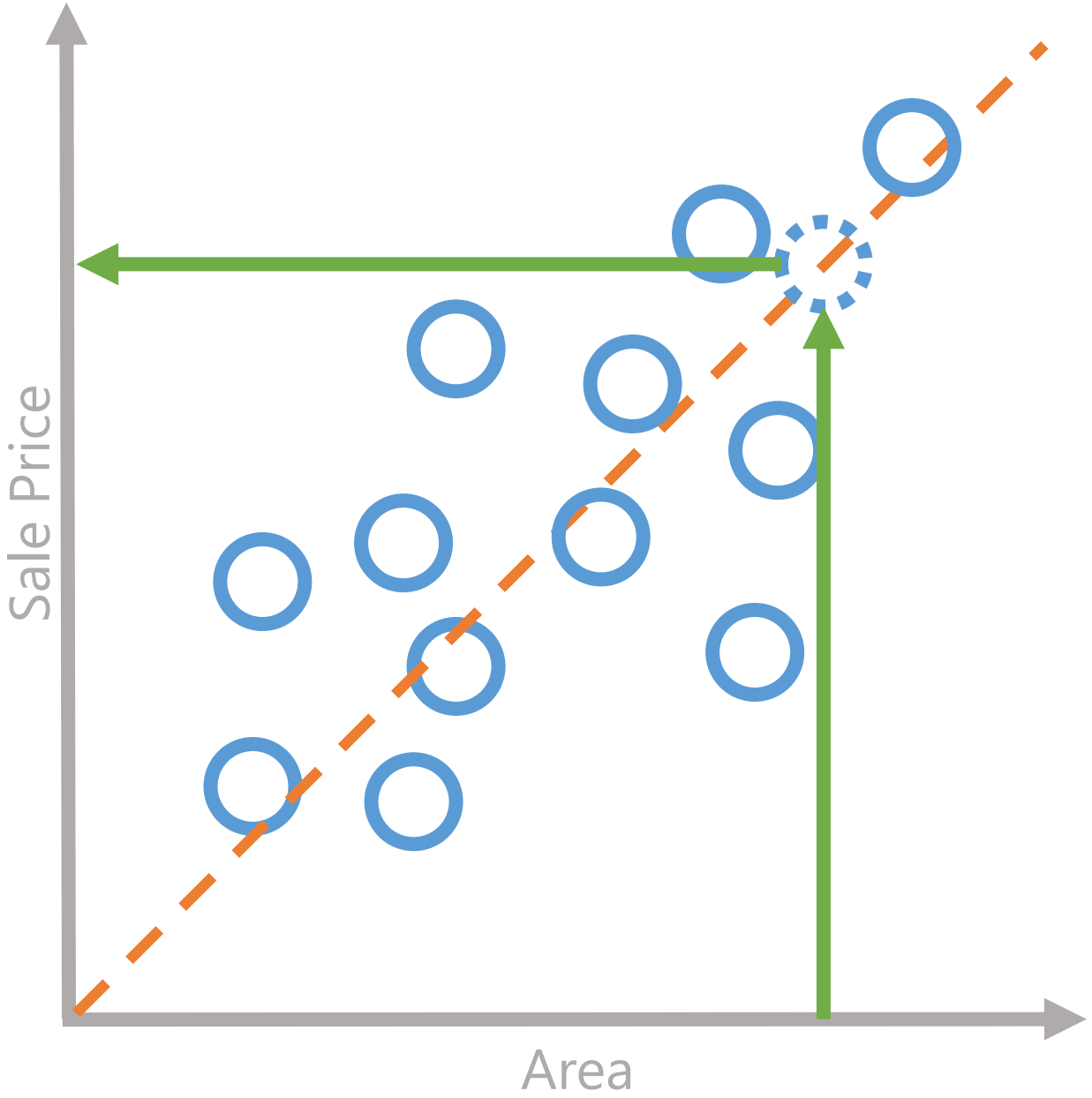
Goal: Predict the risk of
an insurance policy

Regression


$f(x)$
1.23


Sale Price
Area


Sale Price
Area


Sale Price
Area

Regression Algorithms

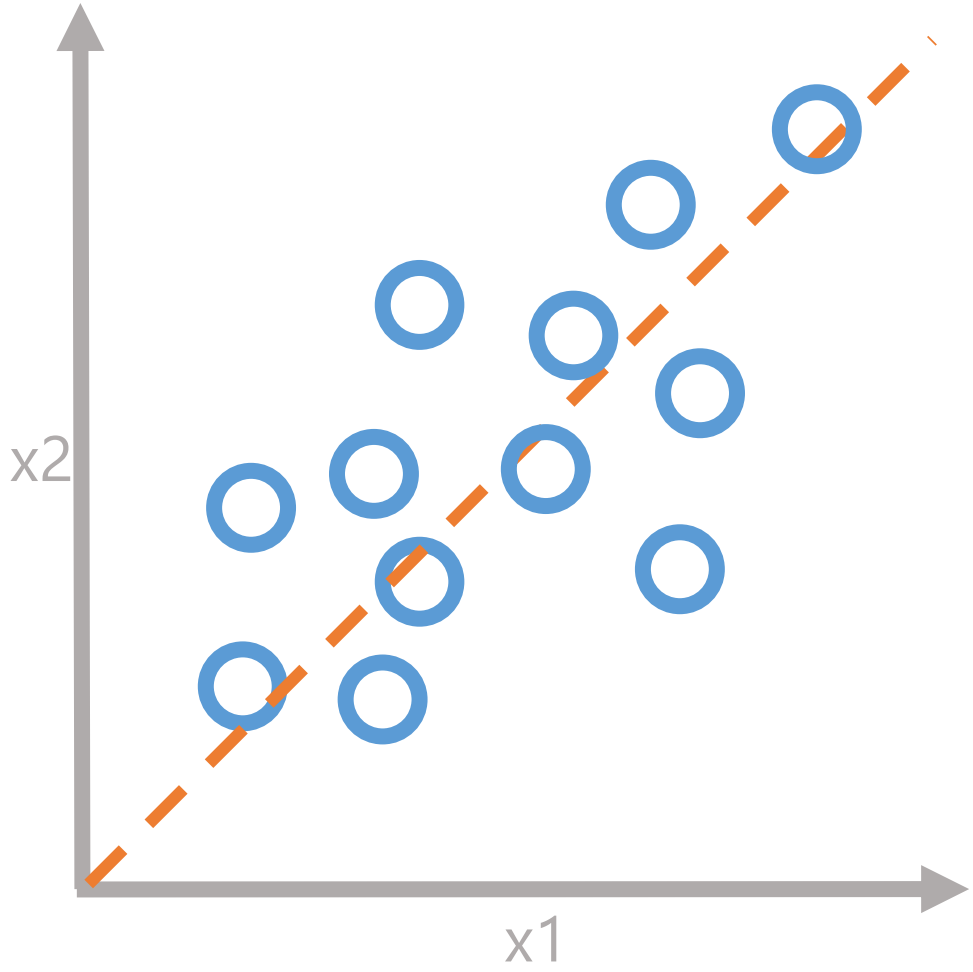
Linear Regression

Polynomial Regression

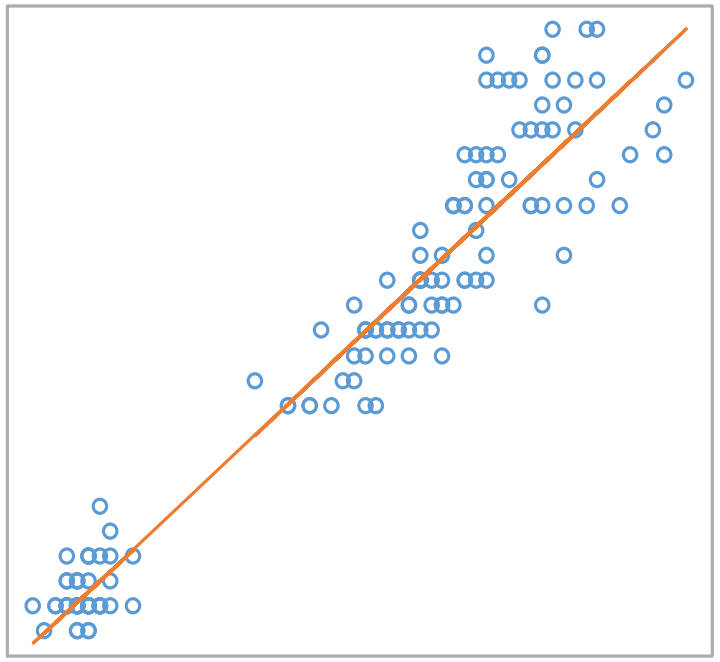
Lasso Regression

ElasticNet Regression

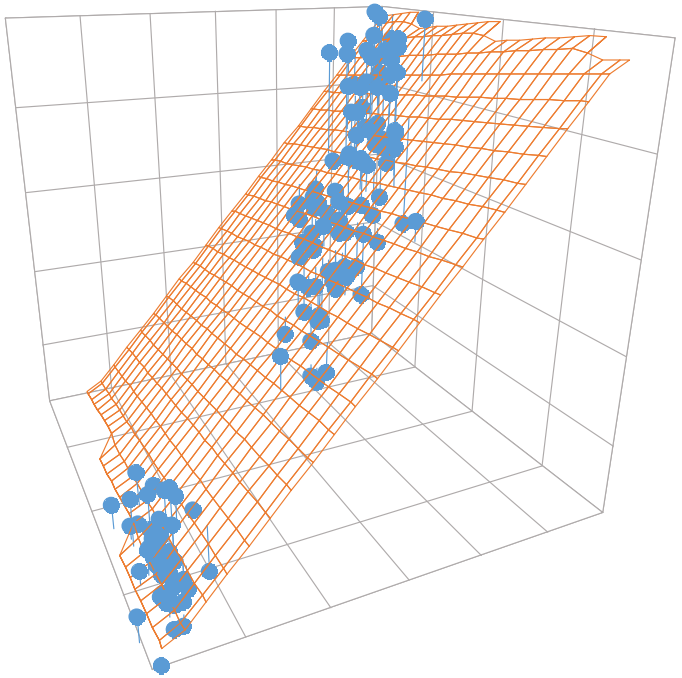
Neural Network Regression



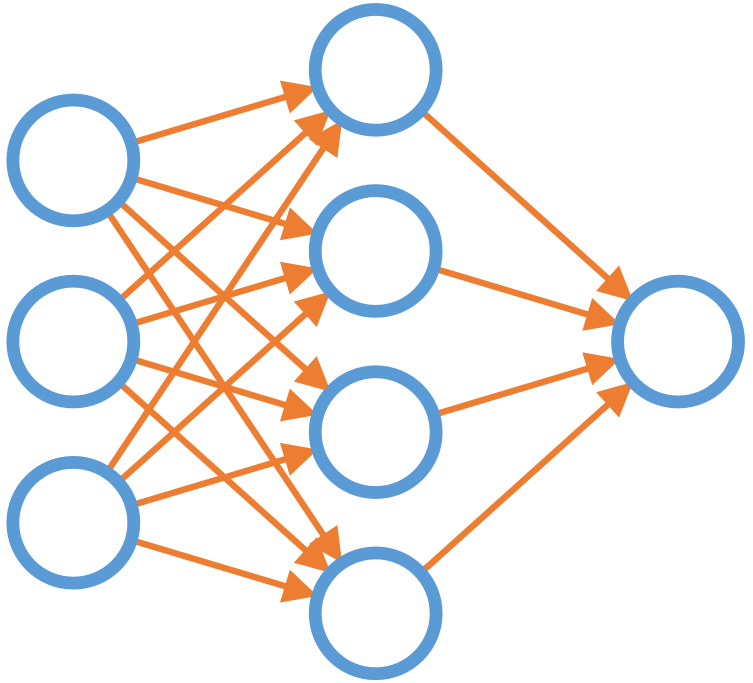
Regression Algorithms



Simple Linear



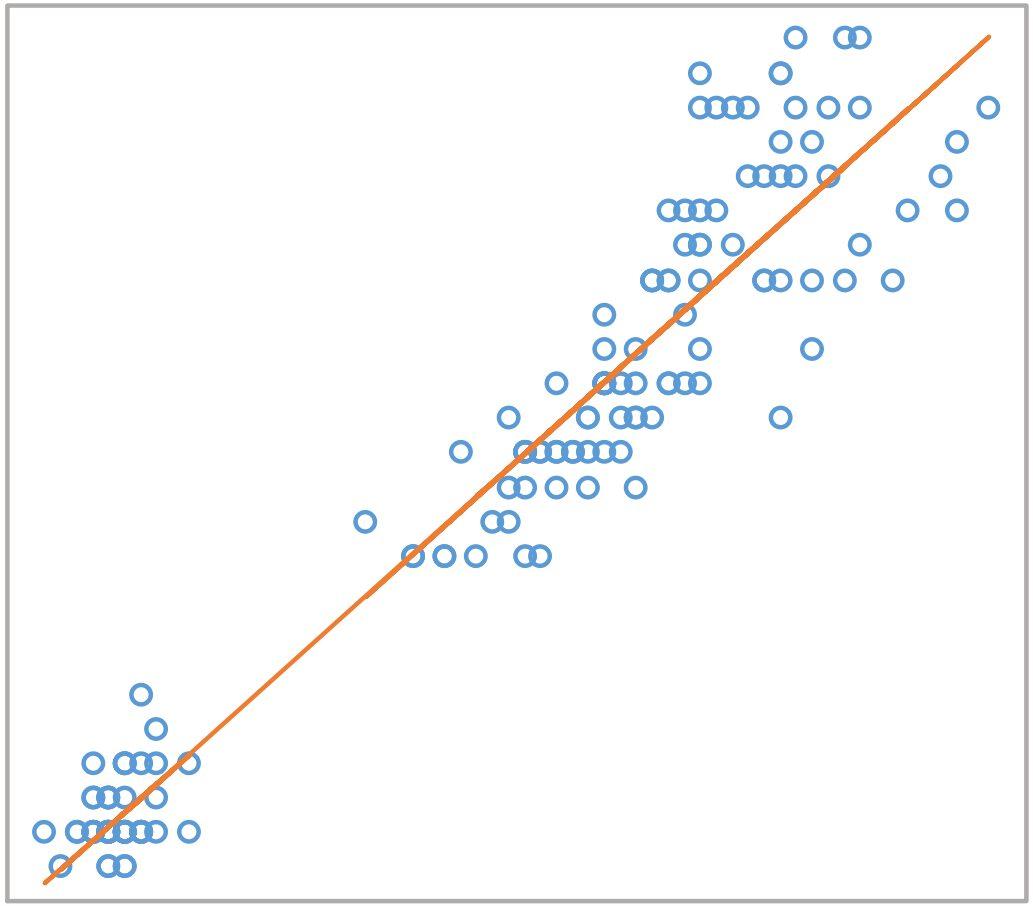
Multiple Linear



Neural Network

Simple Linear Regression

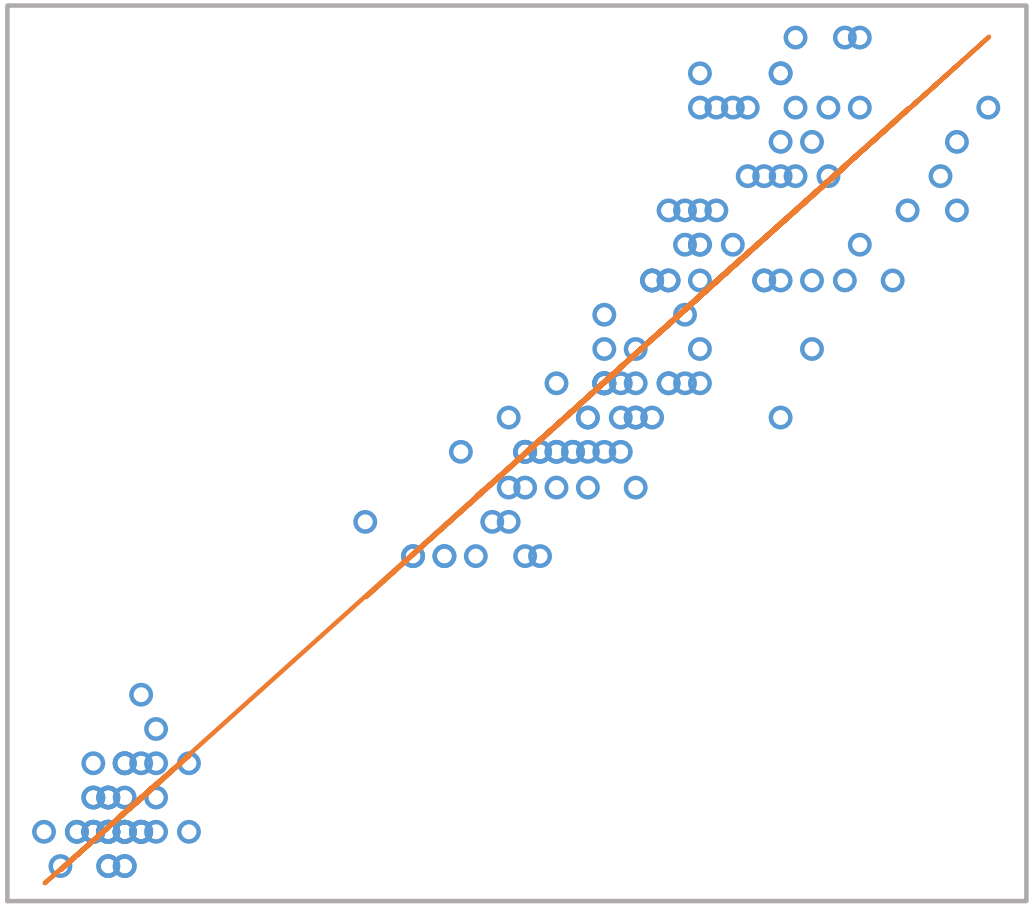
Relationship



Simple Linear Regression

Relationship

Linear model

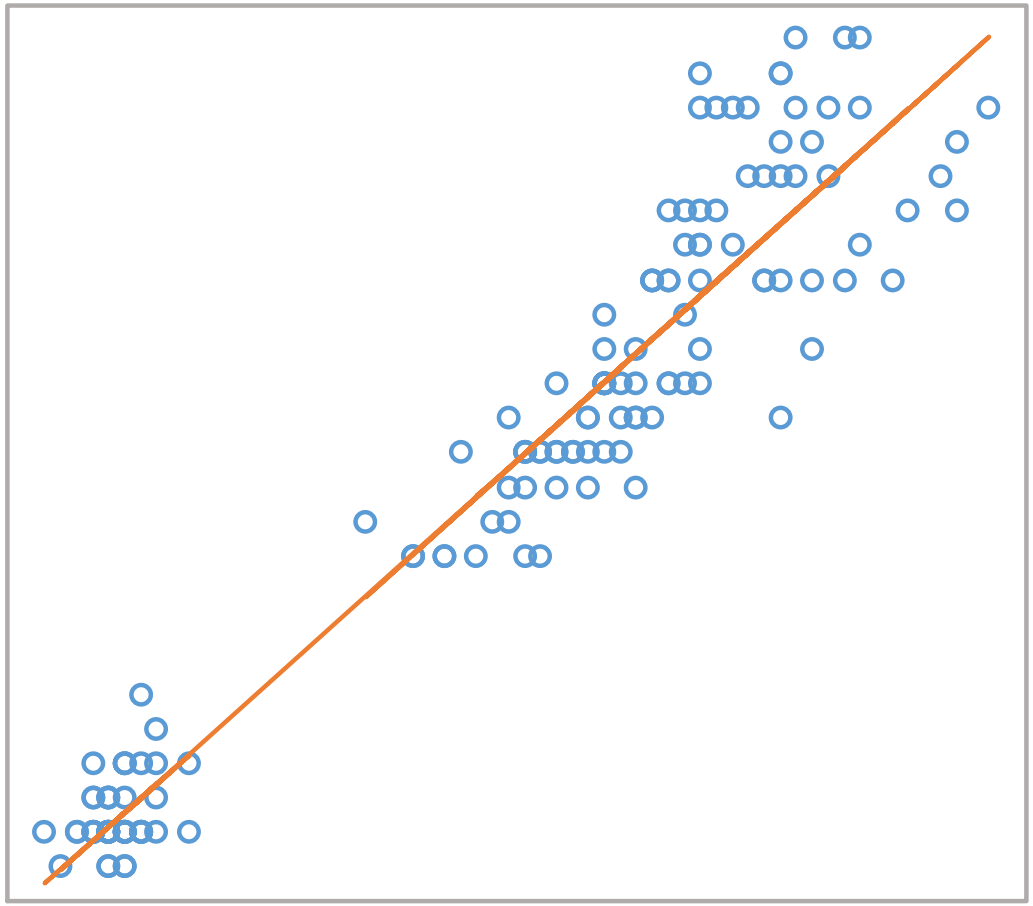


Simple Linear Regression

Relationship

Linear model

$y = m \cdot x + b$



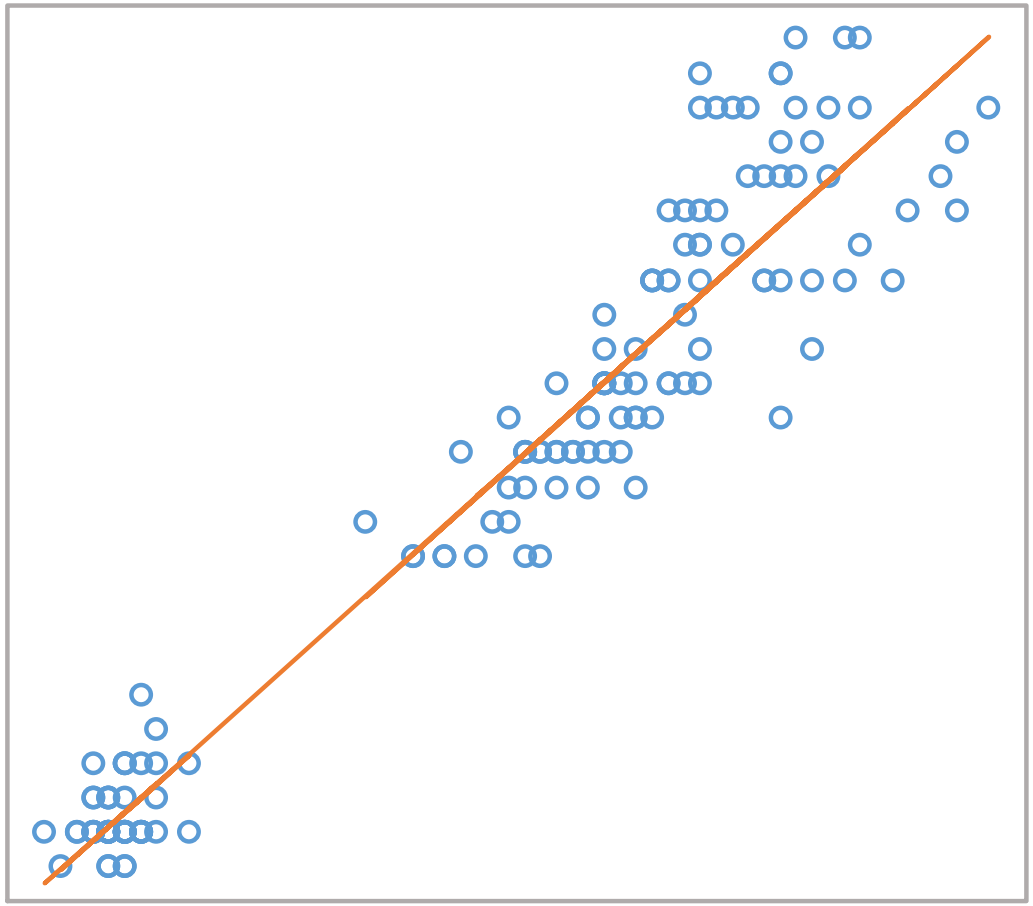
Simple Linear Regression

Relationship

Linear model

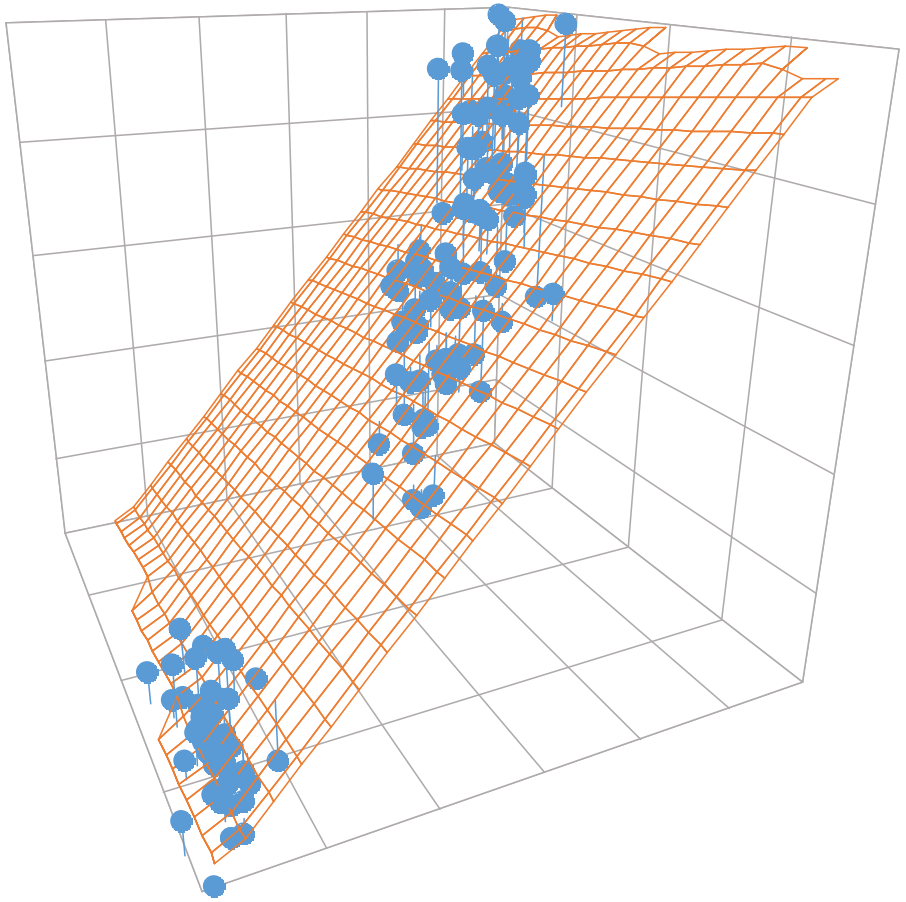
$y = m \cdot x + b$

Parameters estimated



Multiple Linear Regression

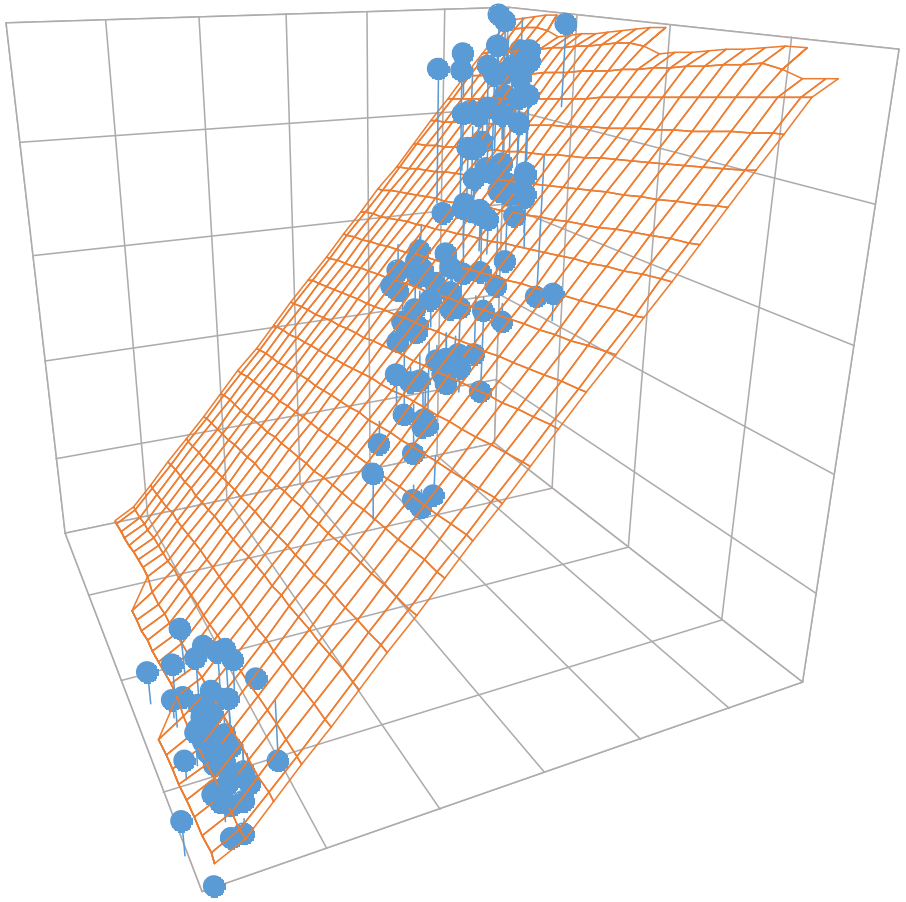
Similar to SLR



Multiple Linear Regression

Similar to SLR

Multiple variables

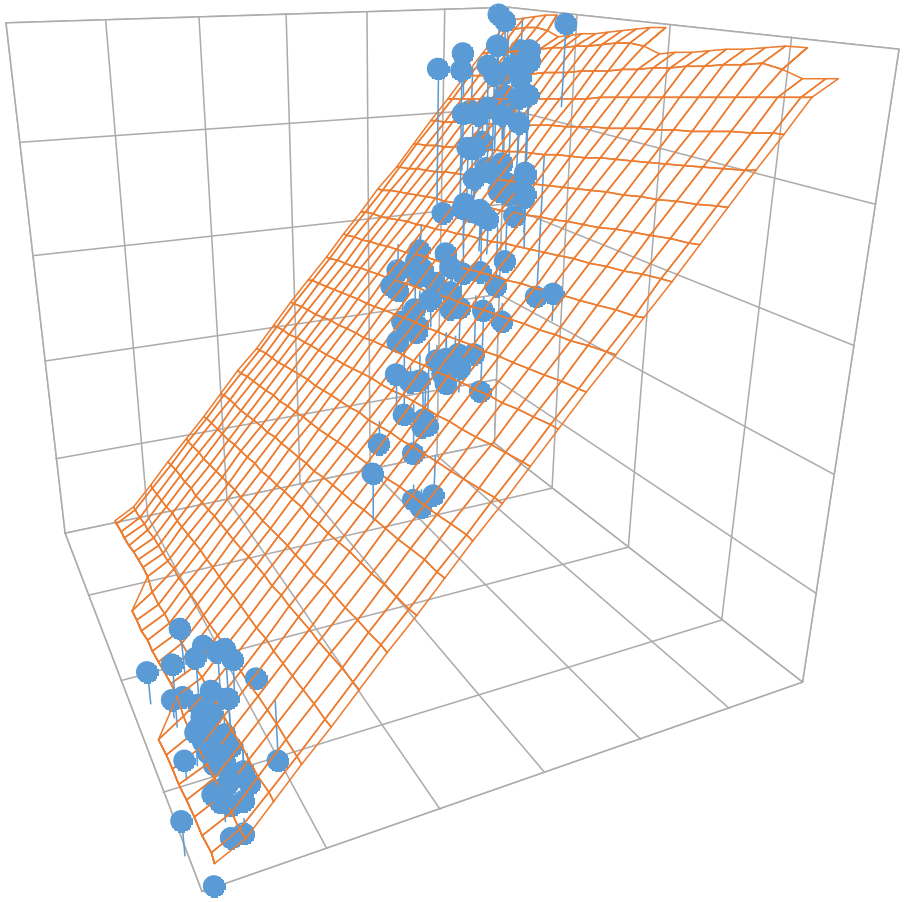


Multiple Linear Regression

Similar to SLR

Multiple variables

Multiple slopes



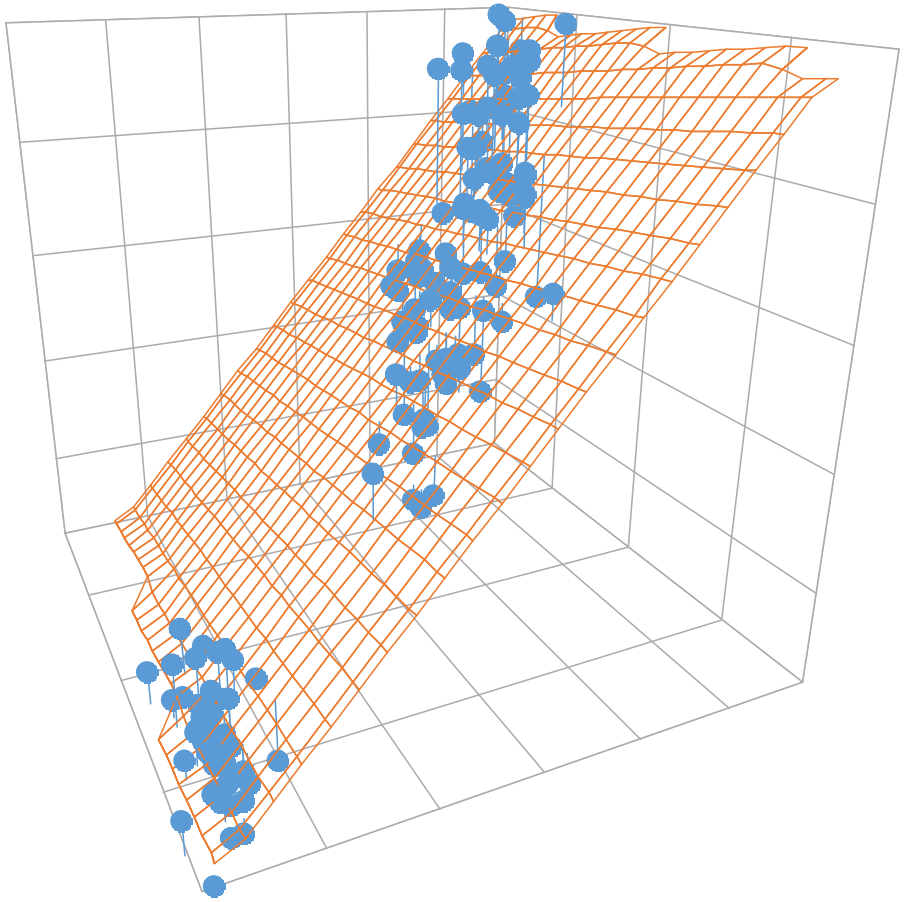
Multiple Linear Regression

Similar to SLR

Multiple variables

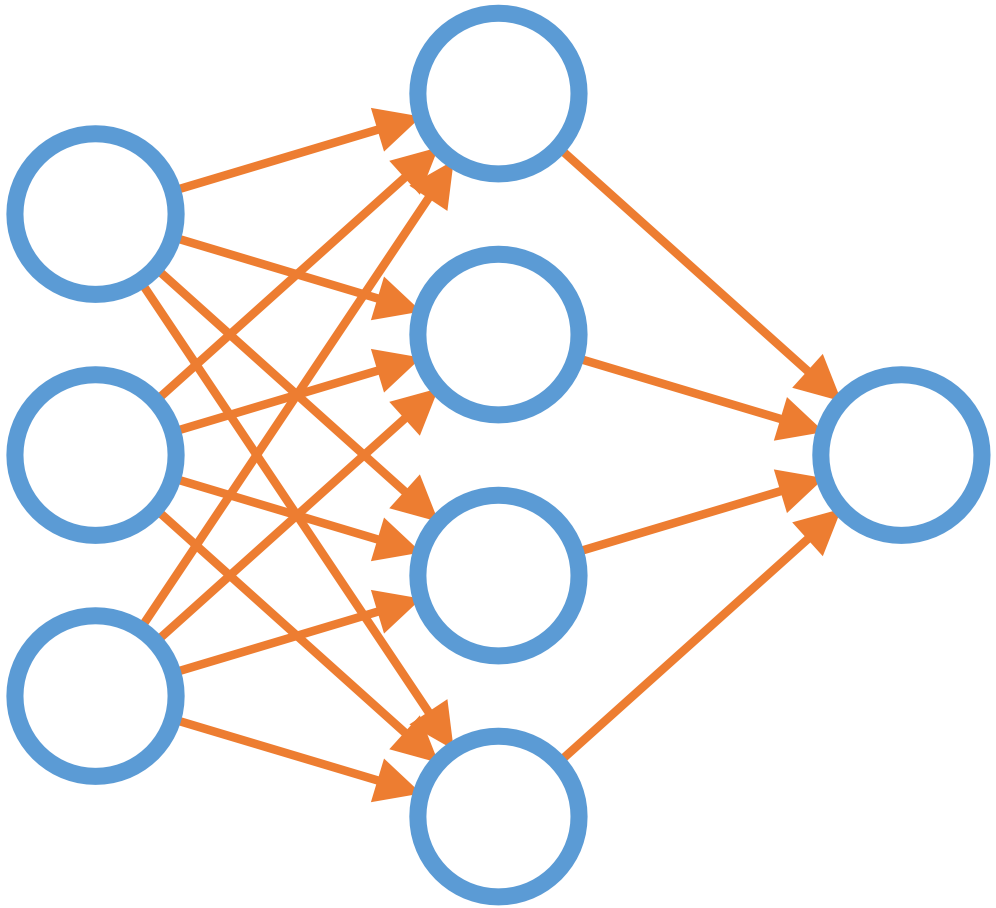
Multiple slopes

Categorical variables



Neural Network Regression

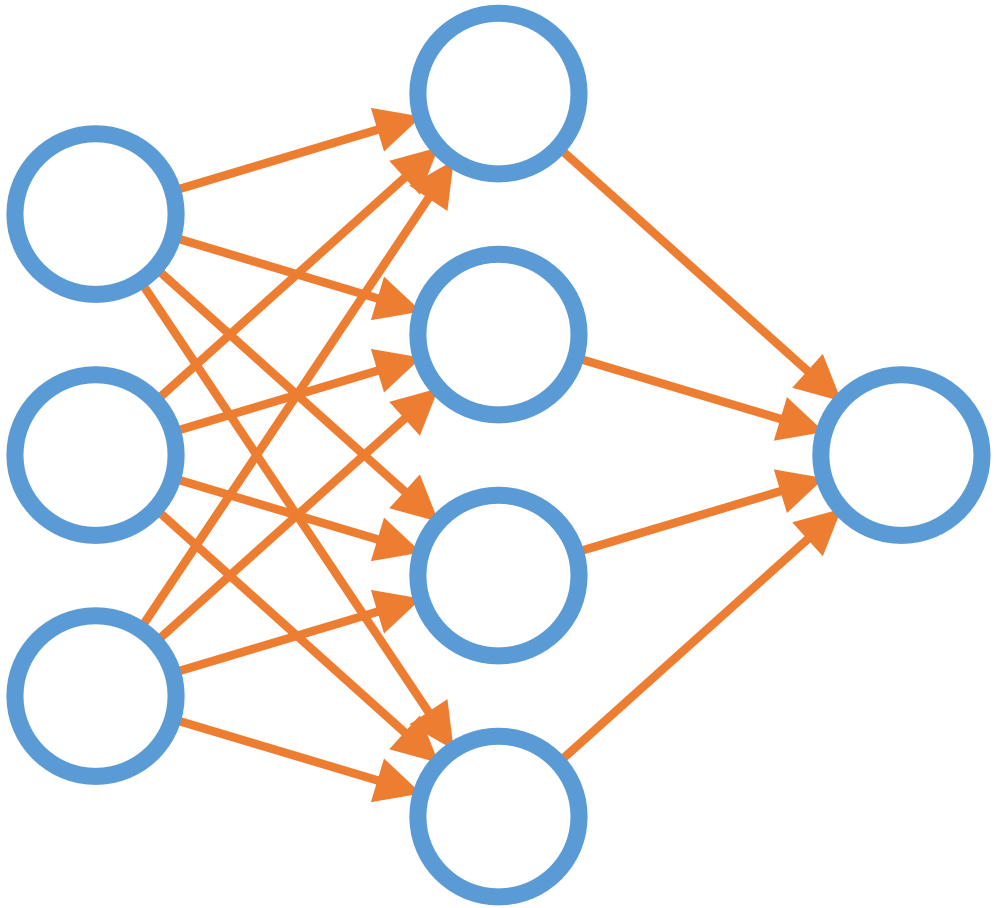
Similar to NN classifier



Neural Network Regression

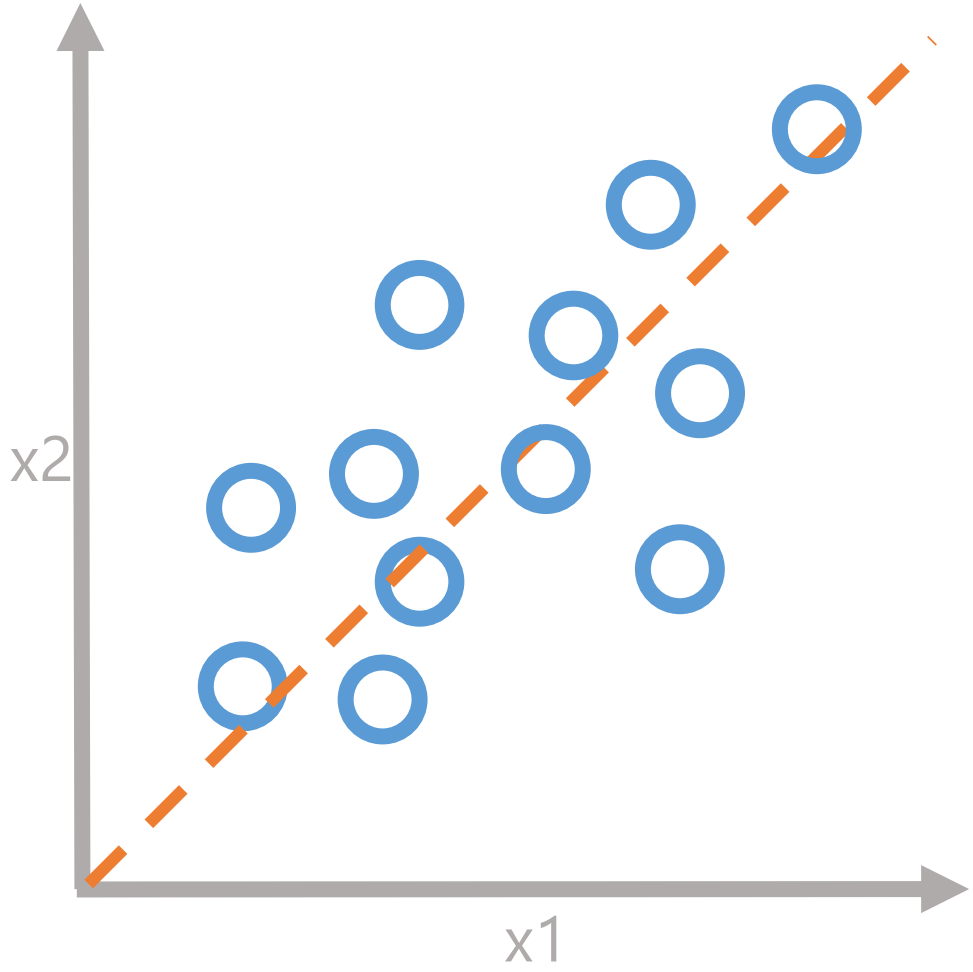
Similar to NN classifier

Numeric output



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?



Demo 3 - Regression

Goal: Predict petal width
based on petal length

Lab 3A – Regression (Easy)

Goal: Predict petal width

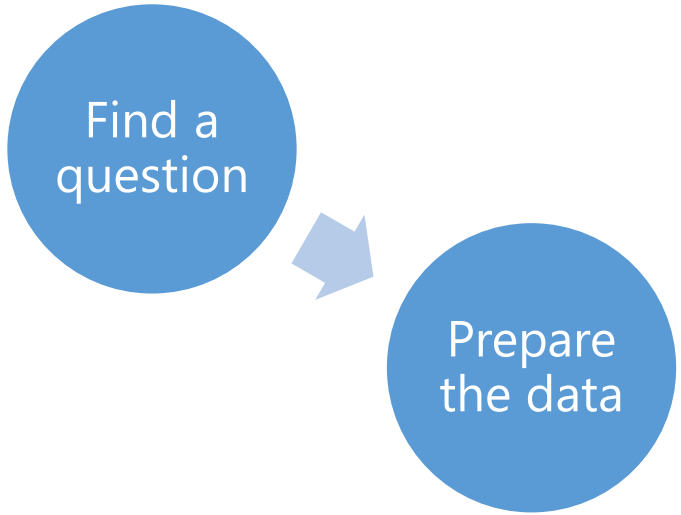
Lab 3B – Regression (Hard)

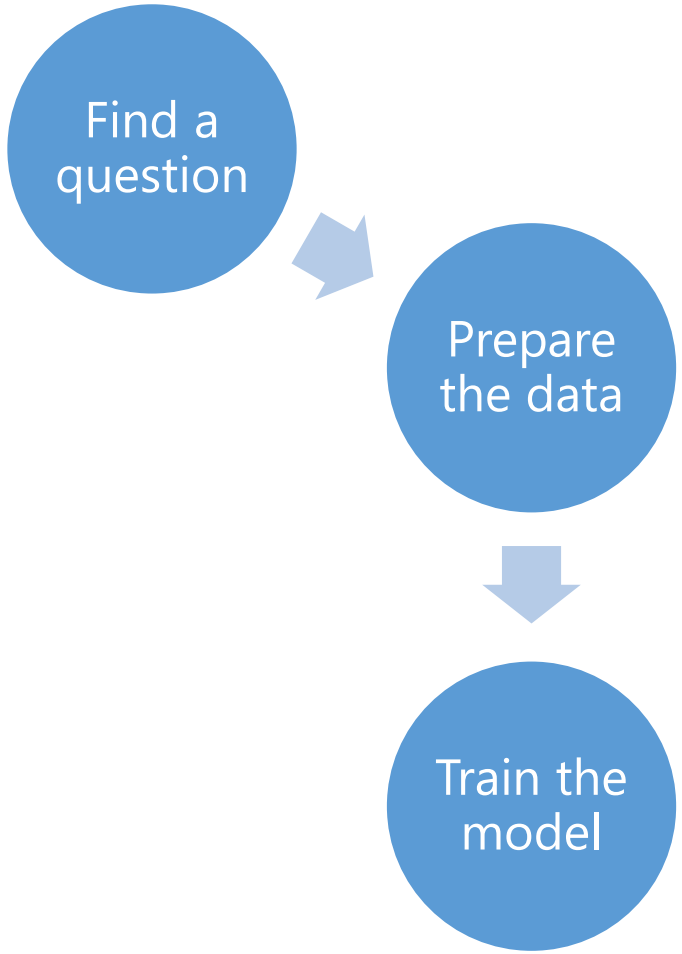
Goal: Predict mortality rate

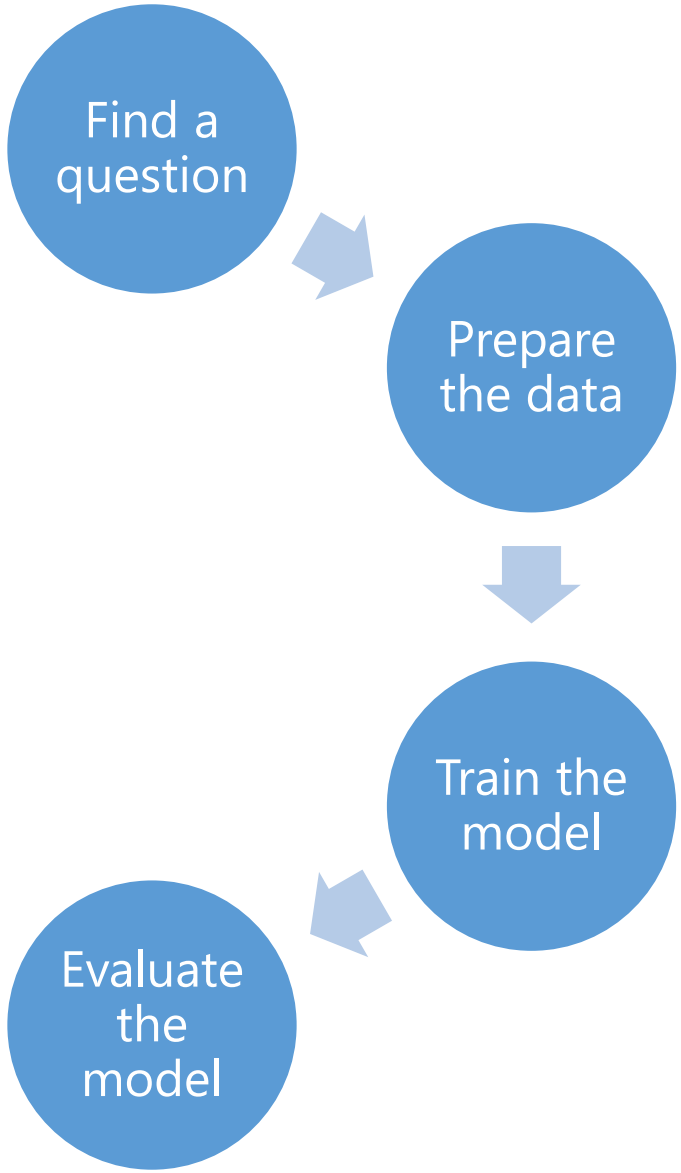
ML in Practice

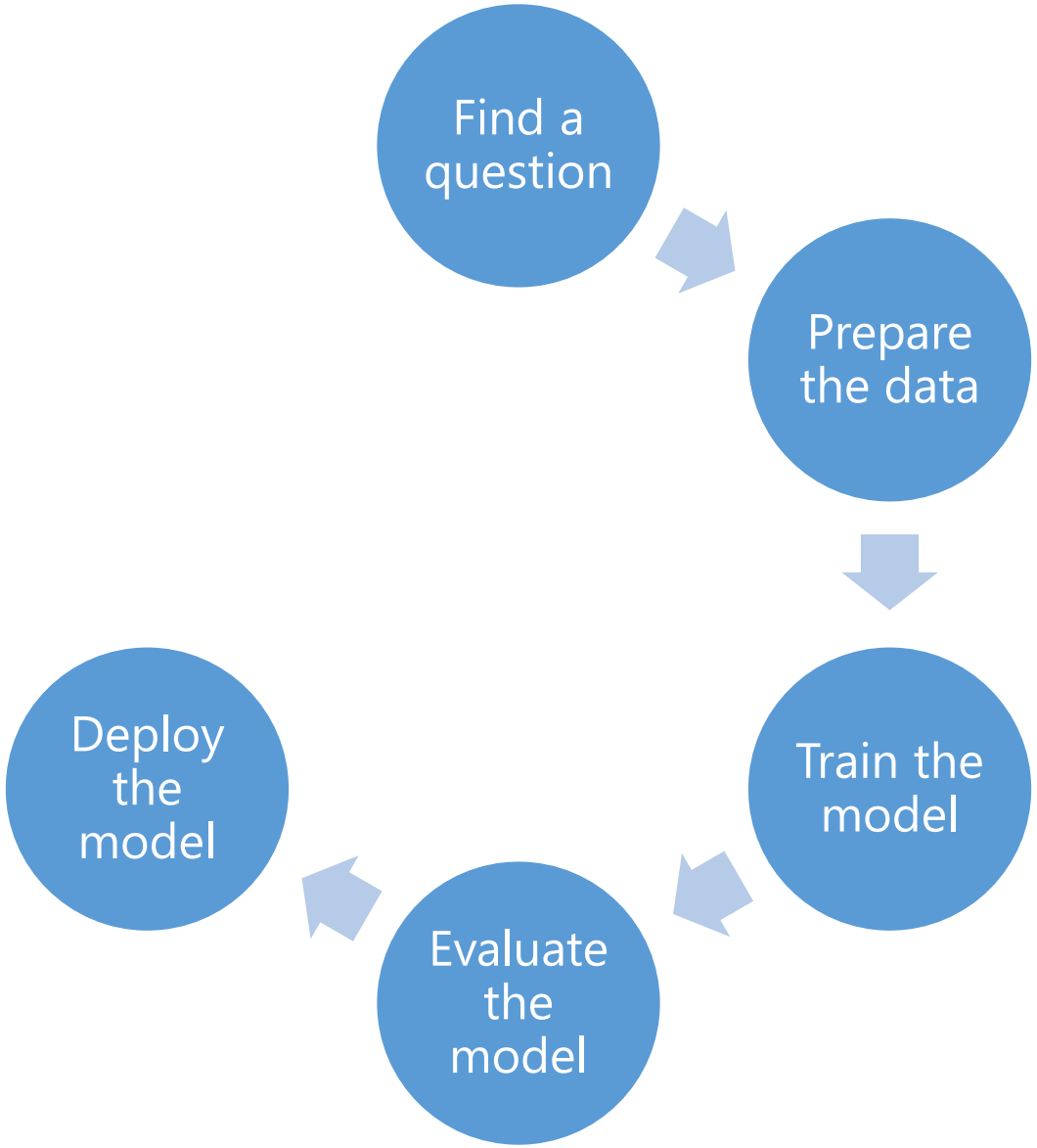
What is the machine learning process?

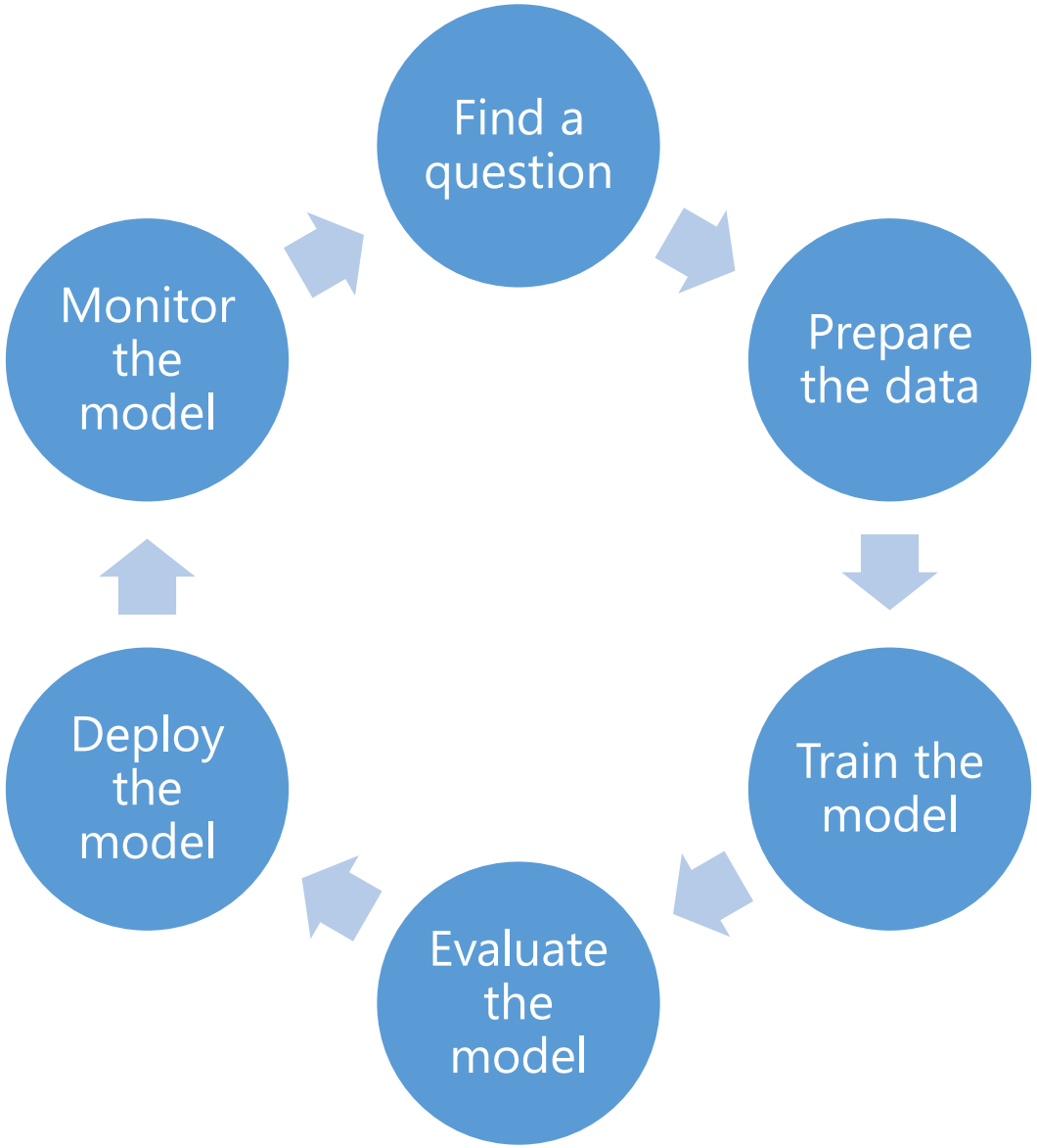
Find a question

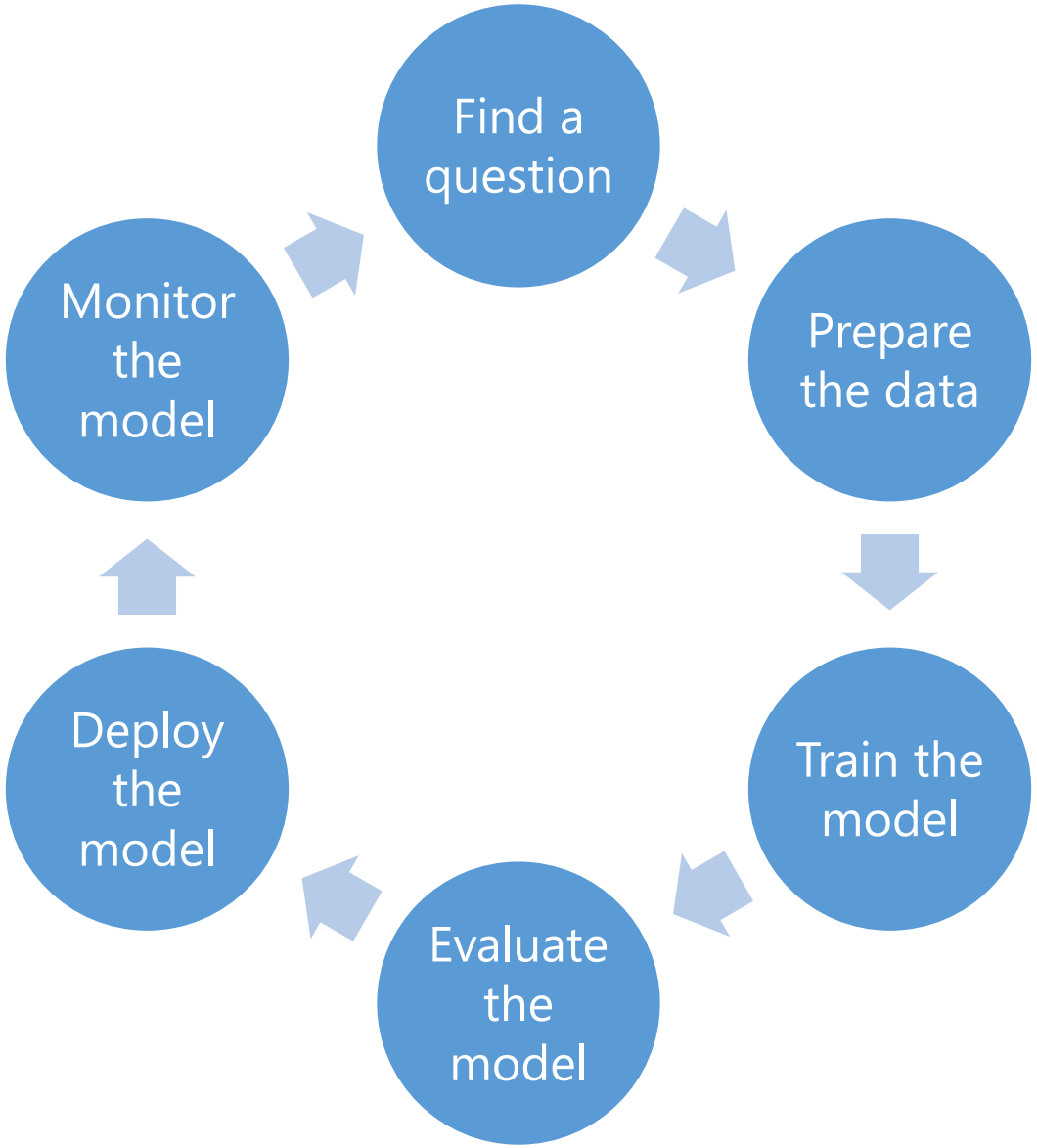

Find a question
Prepare the data

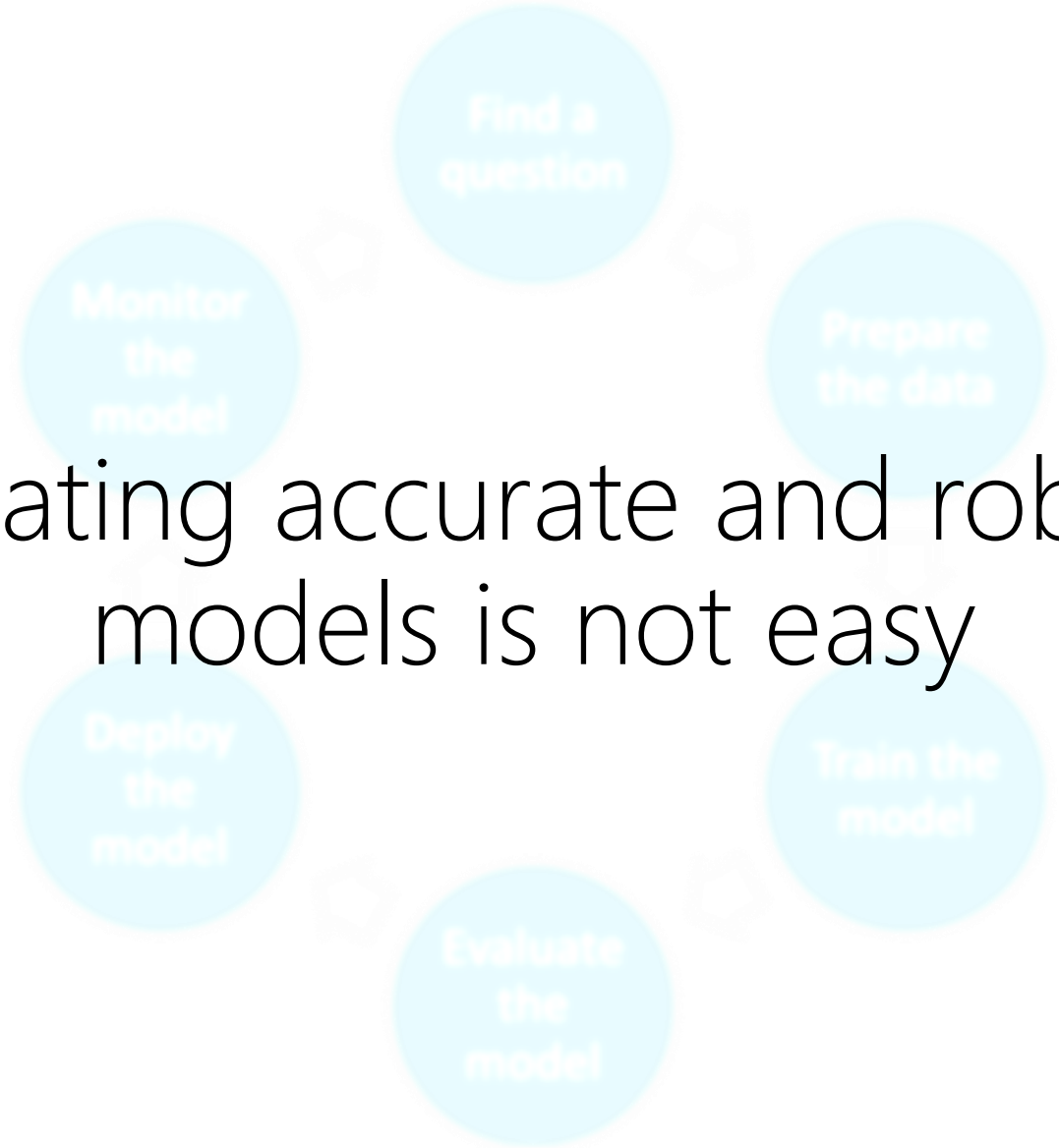

Find a question
Prepare the data
Train the model


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

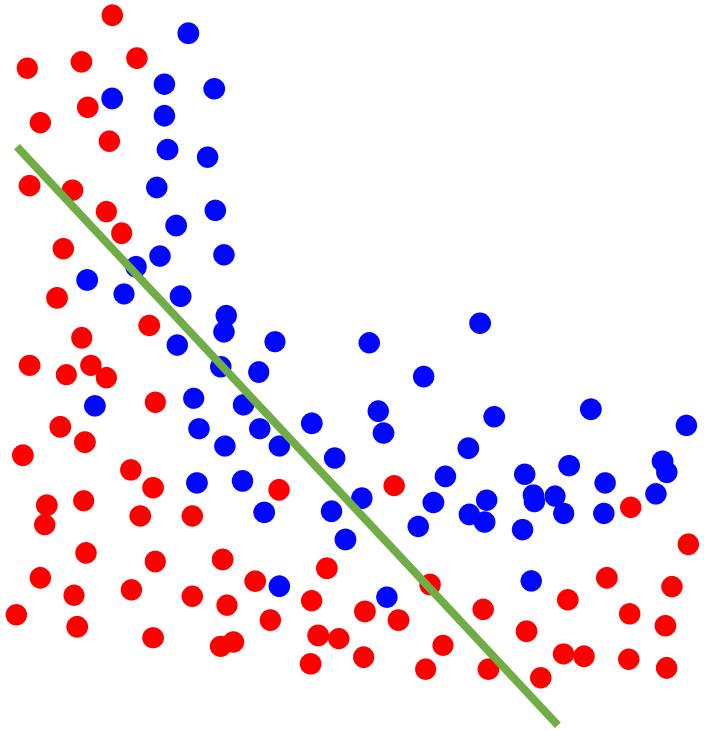

Find a question
Prepare the data
Train the model
Evaluate the model
Deploy the model
Monitor the model



Creating accurate and robust models is not easy

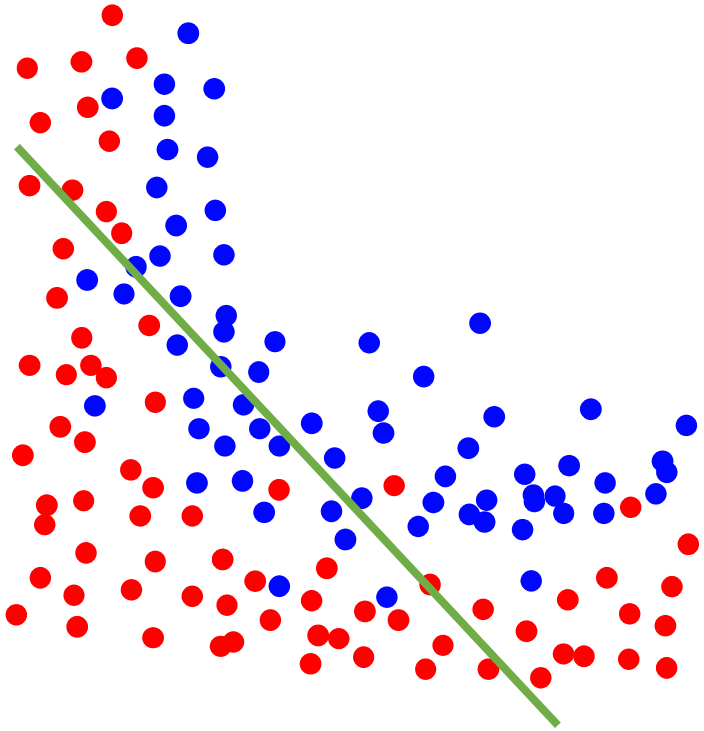
Goodness of Fit

Goodness of Fit

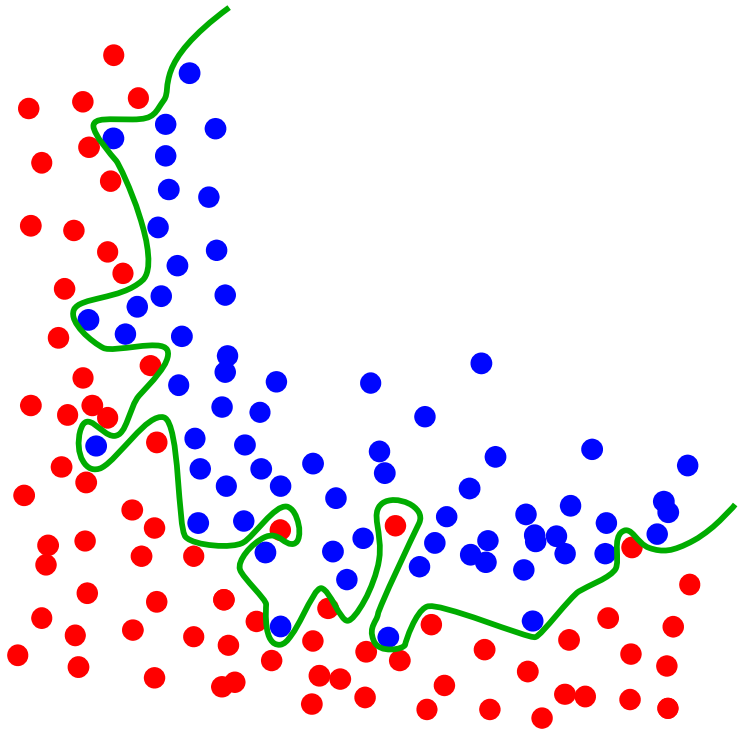


Underfit

Goodness of Fit

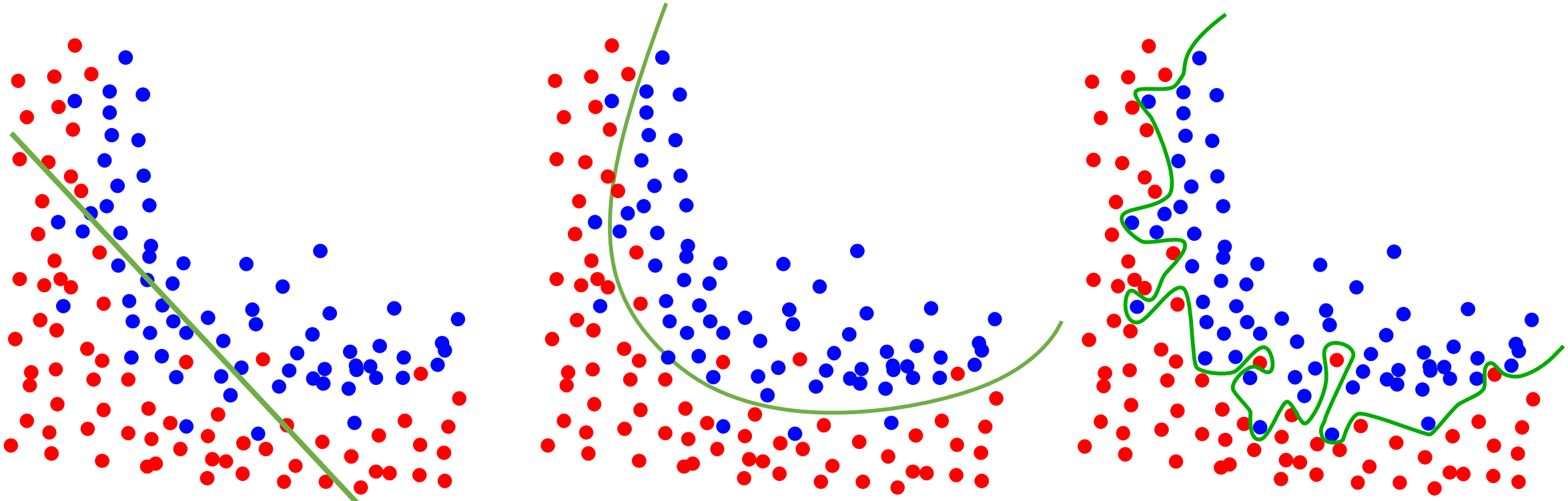


Underfit



Overfit

Goodness of Fit



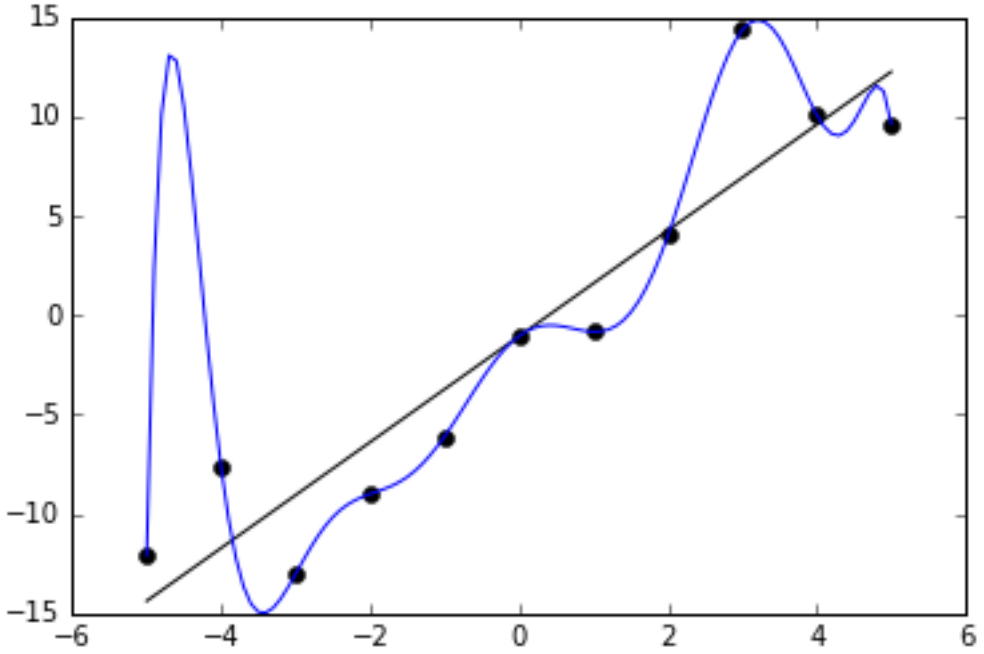
Underfit

Good fit

Overfit

Regularization Techniques

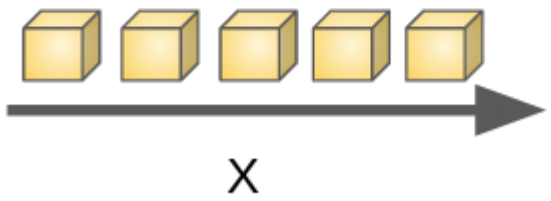
Early stopping
Pruning (trees)
Adding noise
Parameter tuning
Complexity penalties



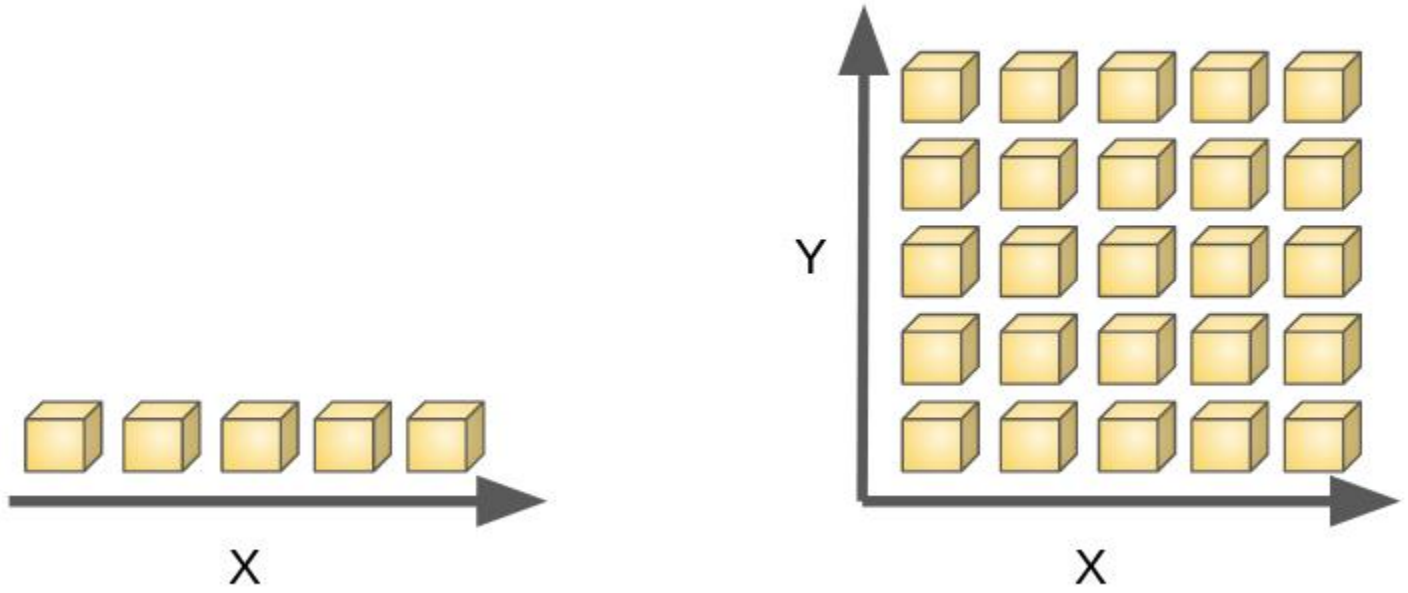
Source: Wikipedia

Curse of Dimensionality

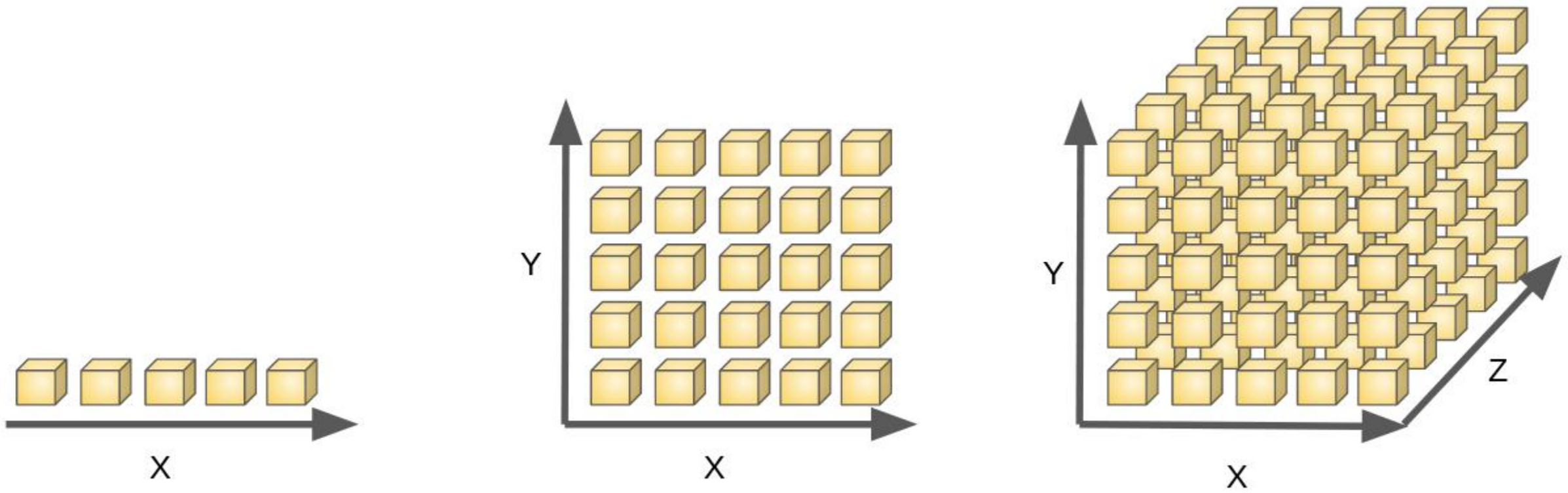
Curse of Dimensionality



Curse of Dimensionality



Curse of Dimensionality




Movie Break

Demo 5 – ML in Practice

Goal 1: Predict Titanic survivors
Goal 2: Deploy Iris classifier to production

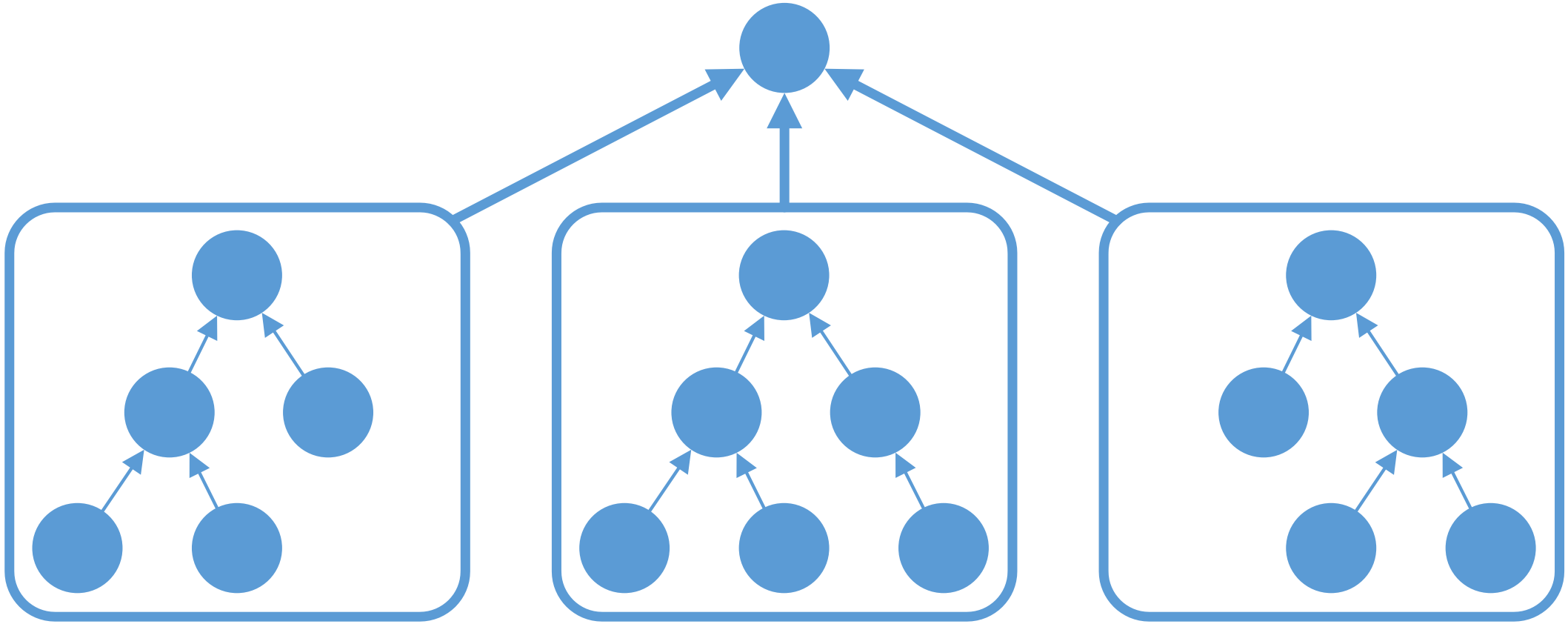
Lab 5A – ML in Practice (Easy)

Goal: Predict survivors
of the Titanic

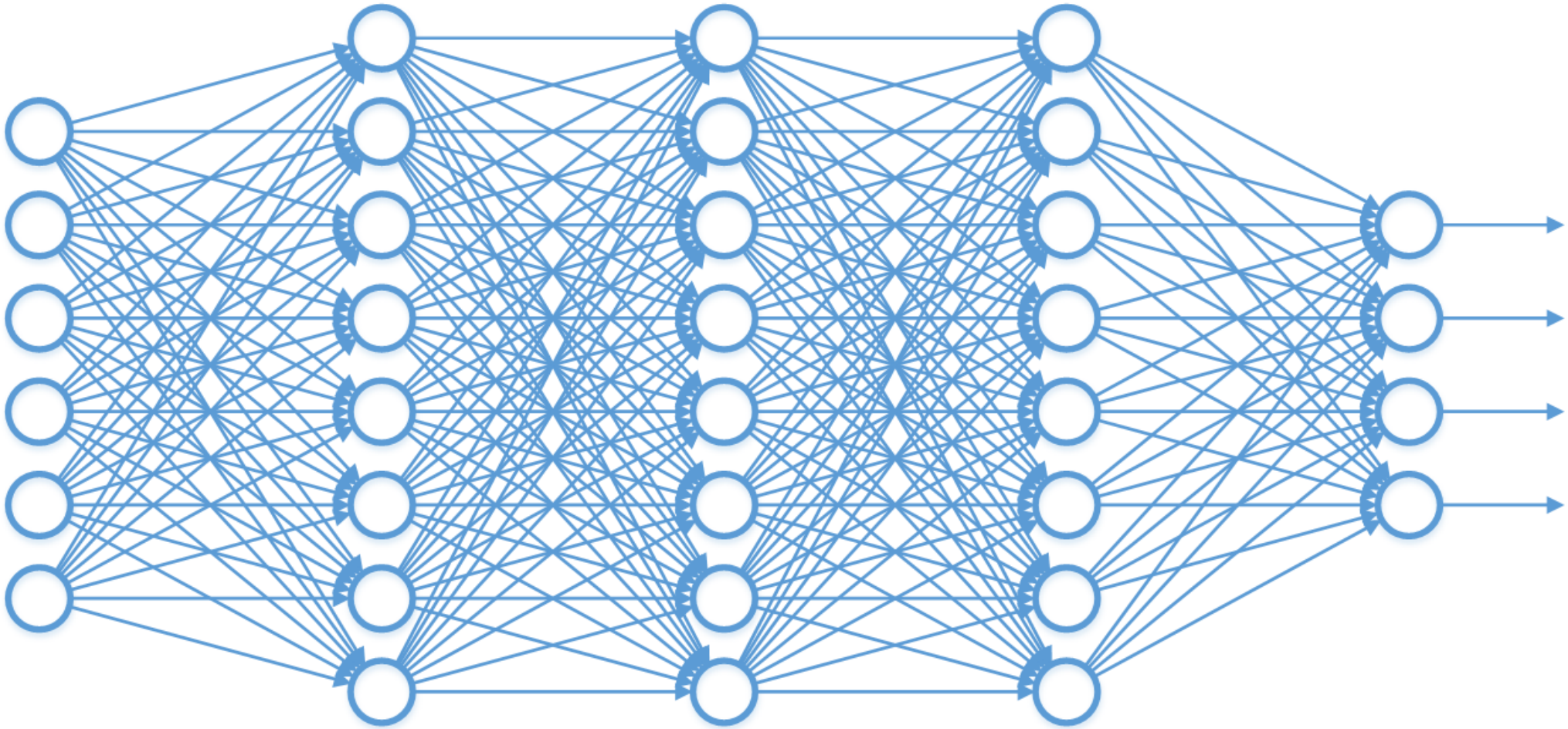
Conclusion


This is just the tip of the iceberg!

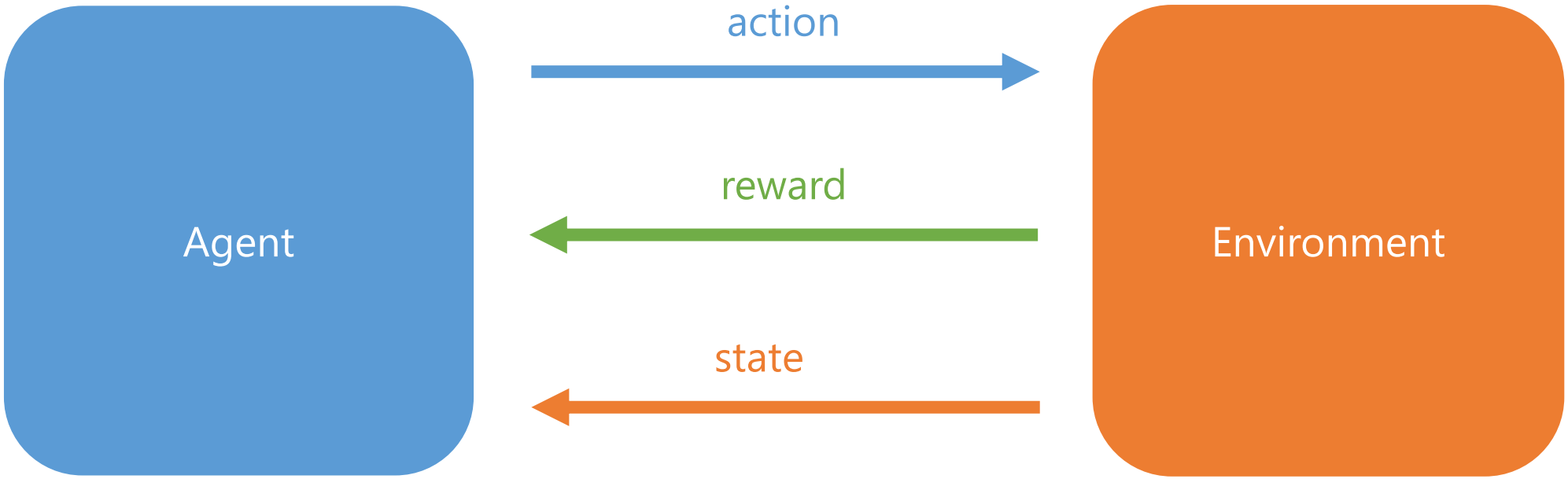
Ensemble Learning



Deep Learning



Reinforcement Learning



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

Pluralsight Courses

Data Science: The Big Picture

Data Science with R

Exploratory Data Analysis with R

Data Visualization with R (3-part)

https://www.pluralsight.com/authors/matthew-renze

Matthew Renze

Home Articles Courses Presentations Software About Contact

News

2017-08-25 - Invitation to Speak at Devoxx Morocco

Very excited to announce that I've been invited to give a keynote in Casablanca at Devoxx Morocco in November. My keynote presentation will be on Artificial Intelligence.

2017-08-16 - Invitation to Speak at Microsoft Ignite

I've been invited to speak at Microsoft Ignite in Orlando, Florida in September. This will be my first time speaking at Ignite. Talks will include both Data Science and Machine Learning with R.

2017-08-14 - Dev on Fire Interview

Matthew is a data science consultant, author for Pluralsight, international public speaker, a Microsoft MVP, ASPInsider, and open-source software contributor.

www.matthewrenze.com

Data Science and Machine Learning

Feedback

Very important to me!

What did you like?

What could I improve?

Conclusion

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

Contact Info

Matthew Renze
Data Science Consultant
Renze Consulting

Twitter: @matthewrenze
Email: matthew@matthewrenze.com
Website: www.matthewrenze.com

Thank You! :)