

Practical Machine Learning with R

@MatthewRenze

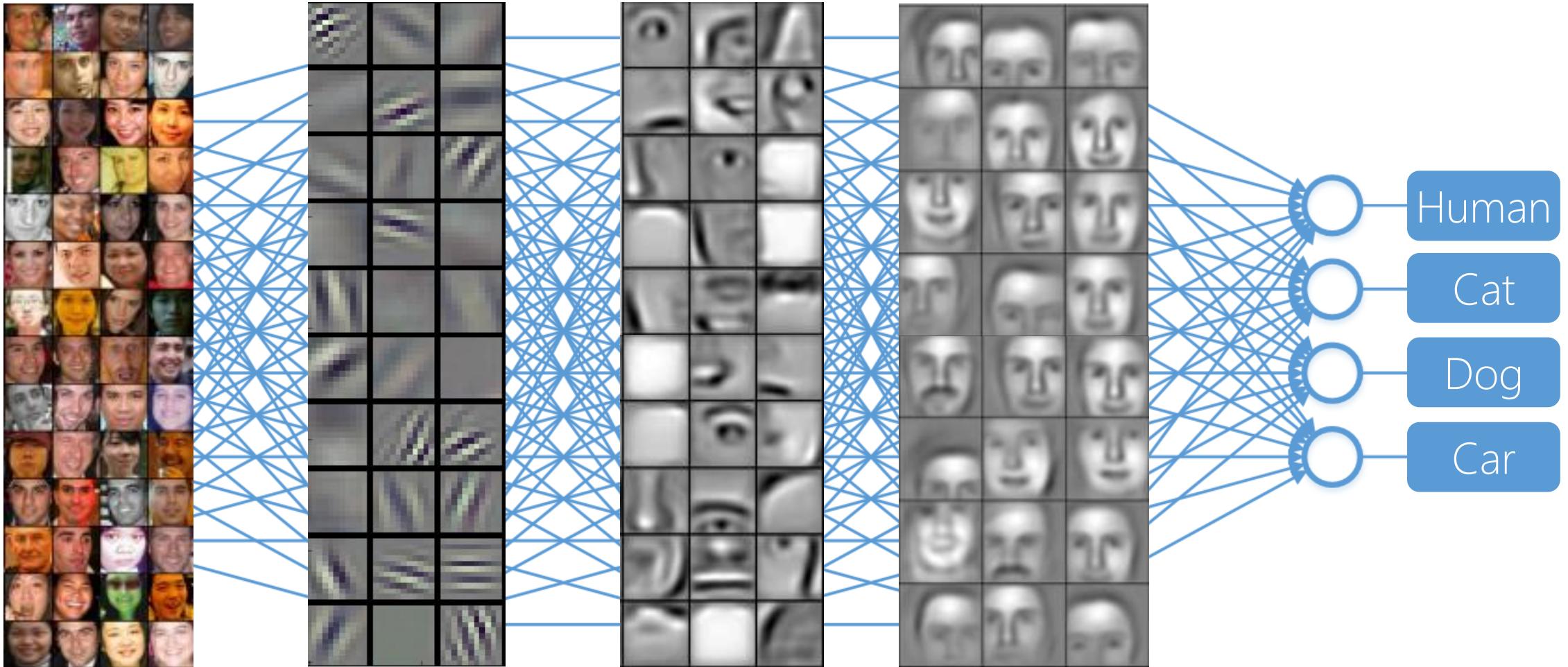
#CodeMash



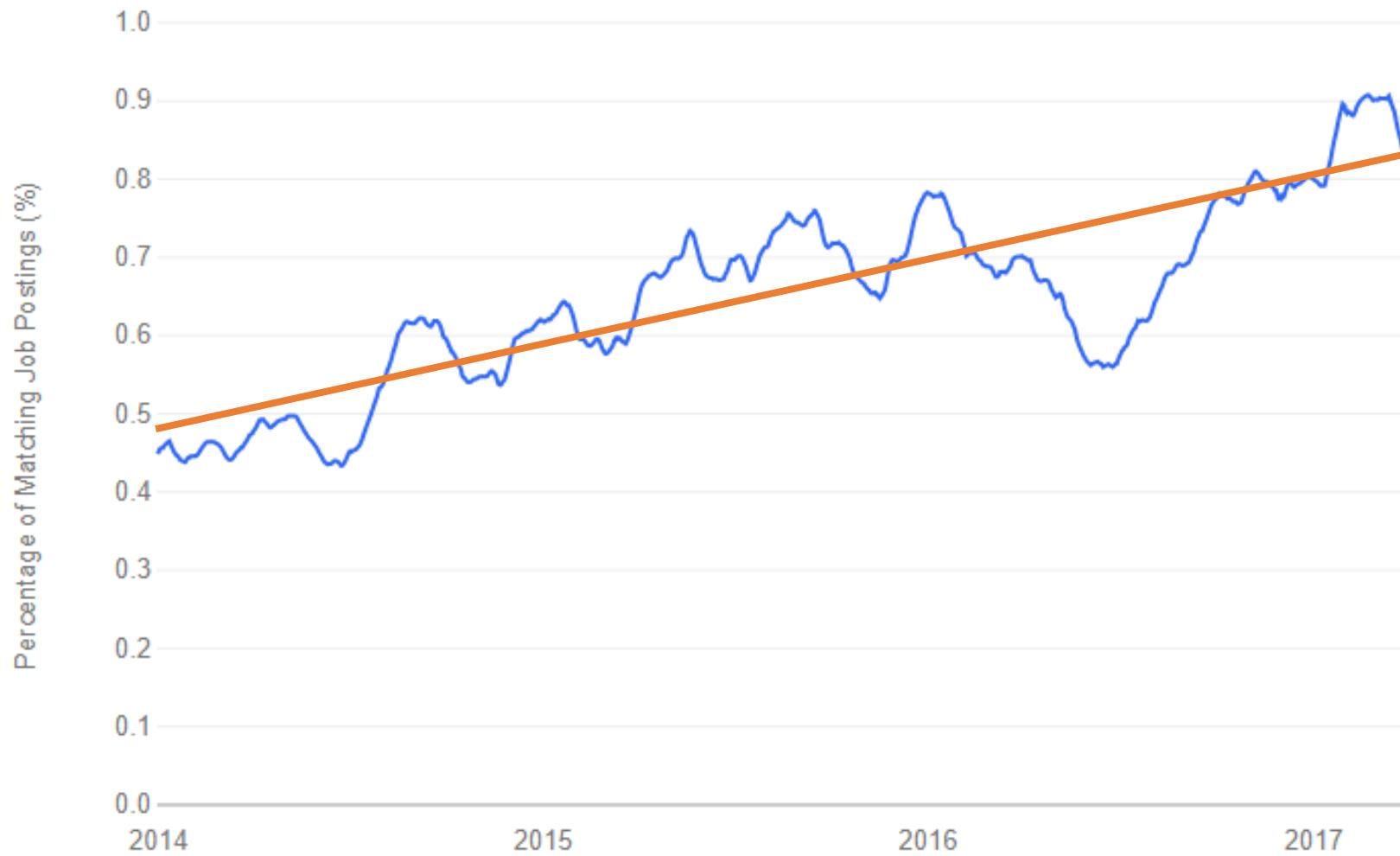


```
function updatePhotoDescription() {
    if (descriptions.length > (page * 9) + (currentImage - 1)) {
        document.getElementById('bigImageDesc').innerHTML = descriptions[currentImage - 1];
    }
}

function updateAllImages() {
    var i = 1;
    while (i < 10) {
        var elementId = 'foto' + i;
        var elementIdBig = 'bigImage' + i;
        if (page * 9 + i - 1 < photos.length) {
            document.getElementById(elementId).src = 'image/min/' + photos[i - 1];
            document.getElementById(elementIdBig).src = 'image/big/' + photos[i - 1];
        } else {
            document.getElementById(elementId).src = '';
        }
        i++;
    }
}
```

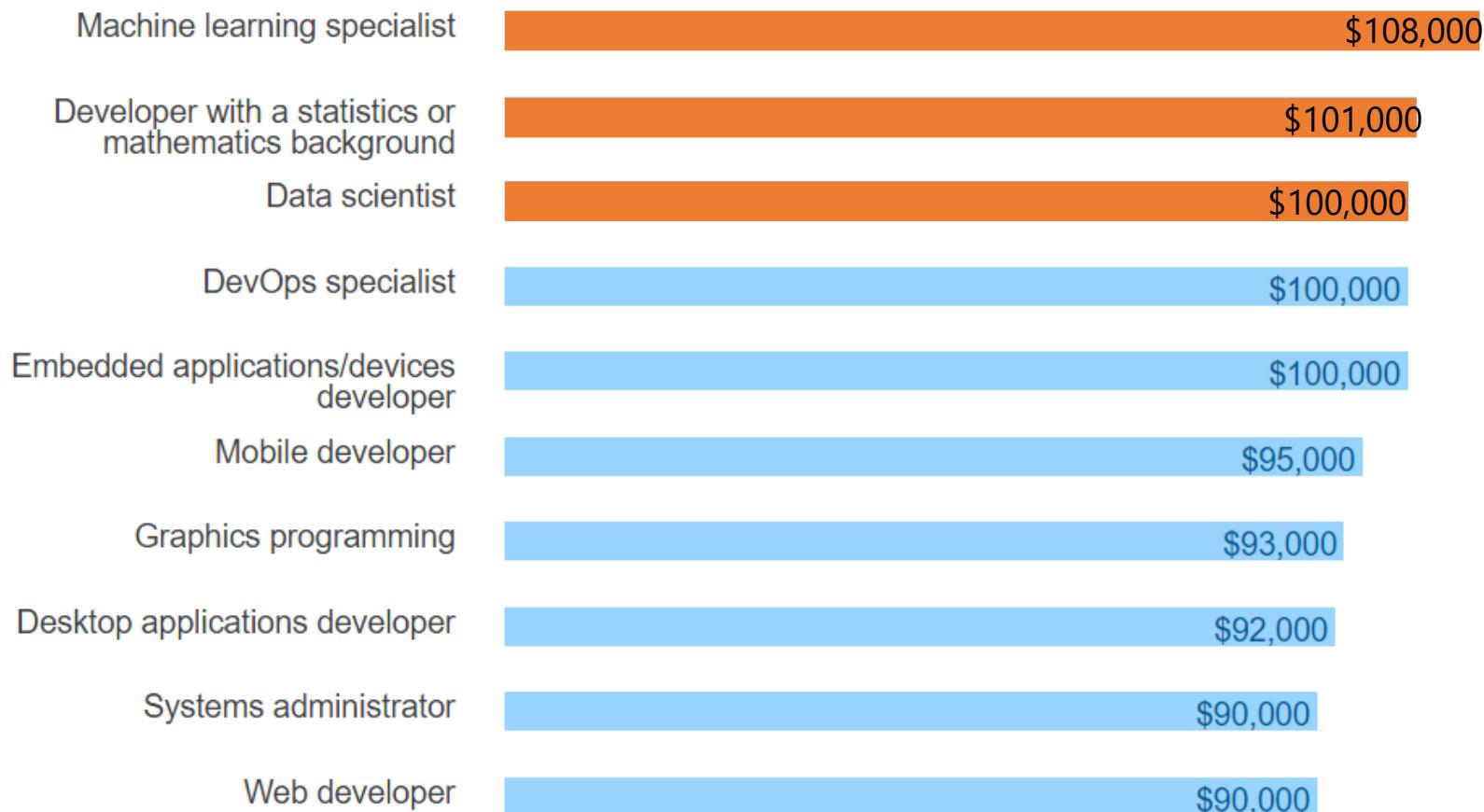


Job Postings for Machine Learning

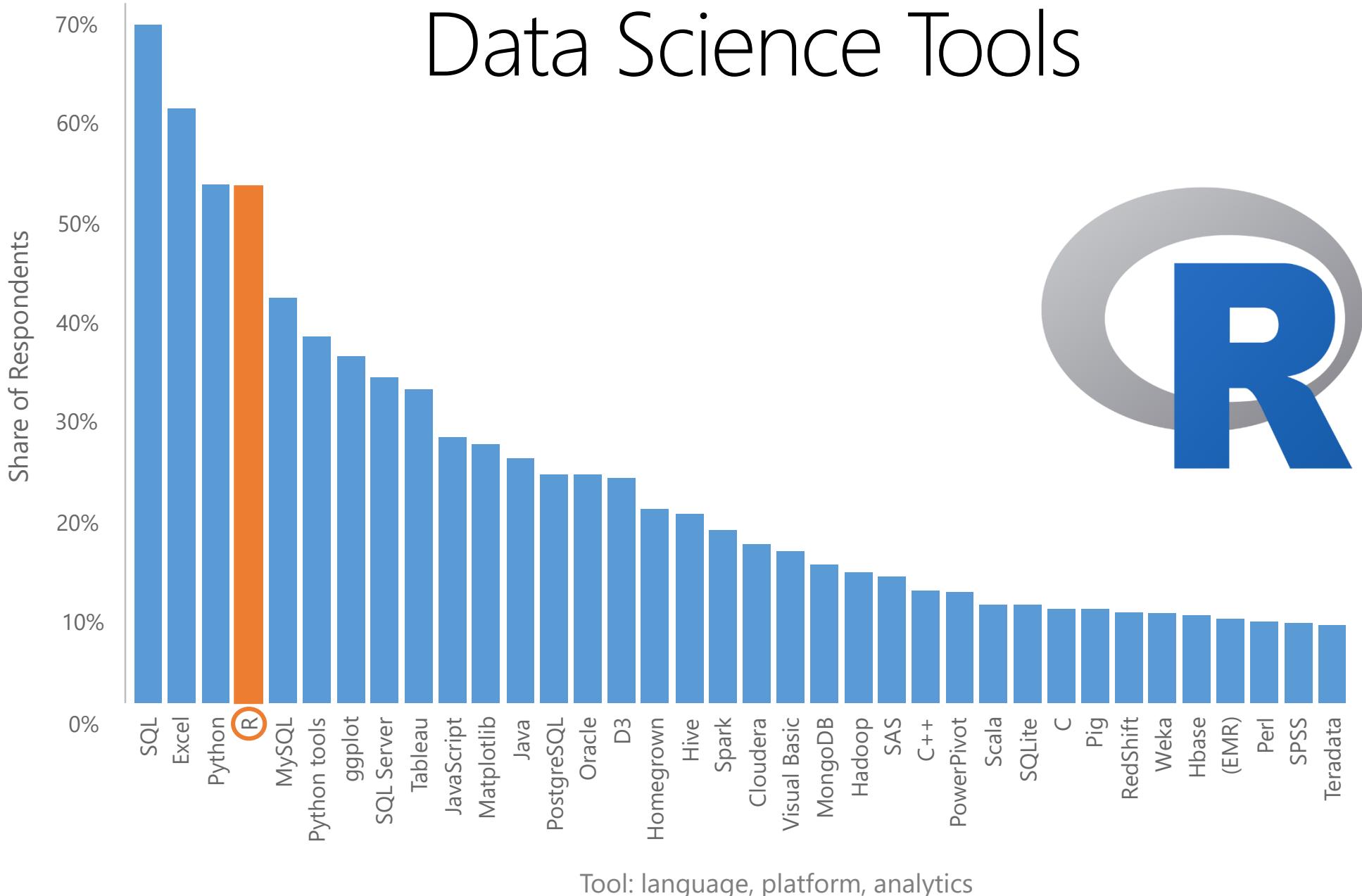


Source: Indeed.com

Average Salary by Job Type (USA)



Data Science Tools



Source: O'Reilly 2015 Data Science Salary Survey



TR

Overview

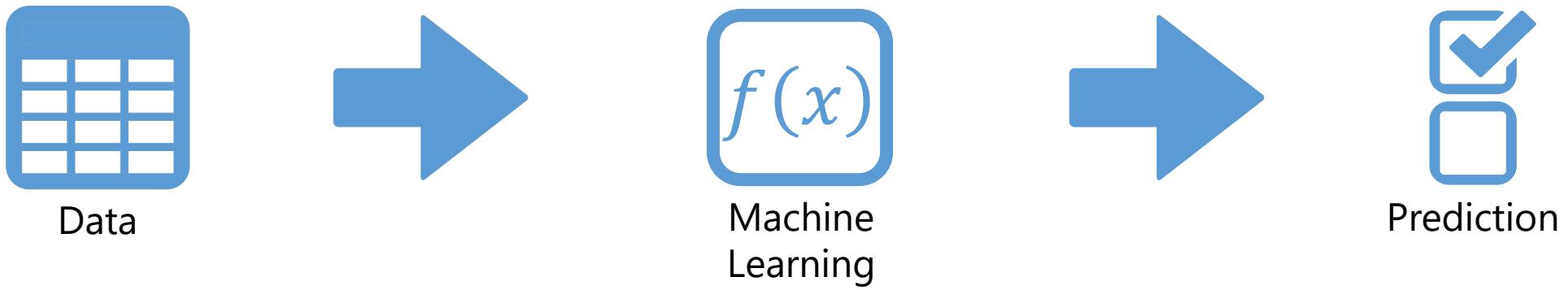
1. Introduction to ML
2. Introduction to R
3. Classification
4. Regression
5. ML in Practice



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

IOWA STATE
UNIVERSITY



Schedule

Lectures (10 min)

Demos (10 min)

Labs (20 min)

Breaks (5 min)

Logistics

Pairing for labs is optional

Ask questions if needed

Come and go as needed

Feedback at the end

Labs

Labs

A
(Easy)

Labs

A

(Easy)

B

(Hard)

Labs

A
(Easy)

B
(Hard)

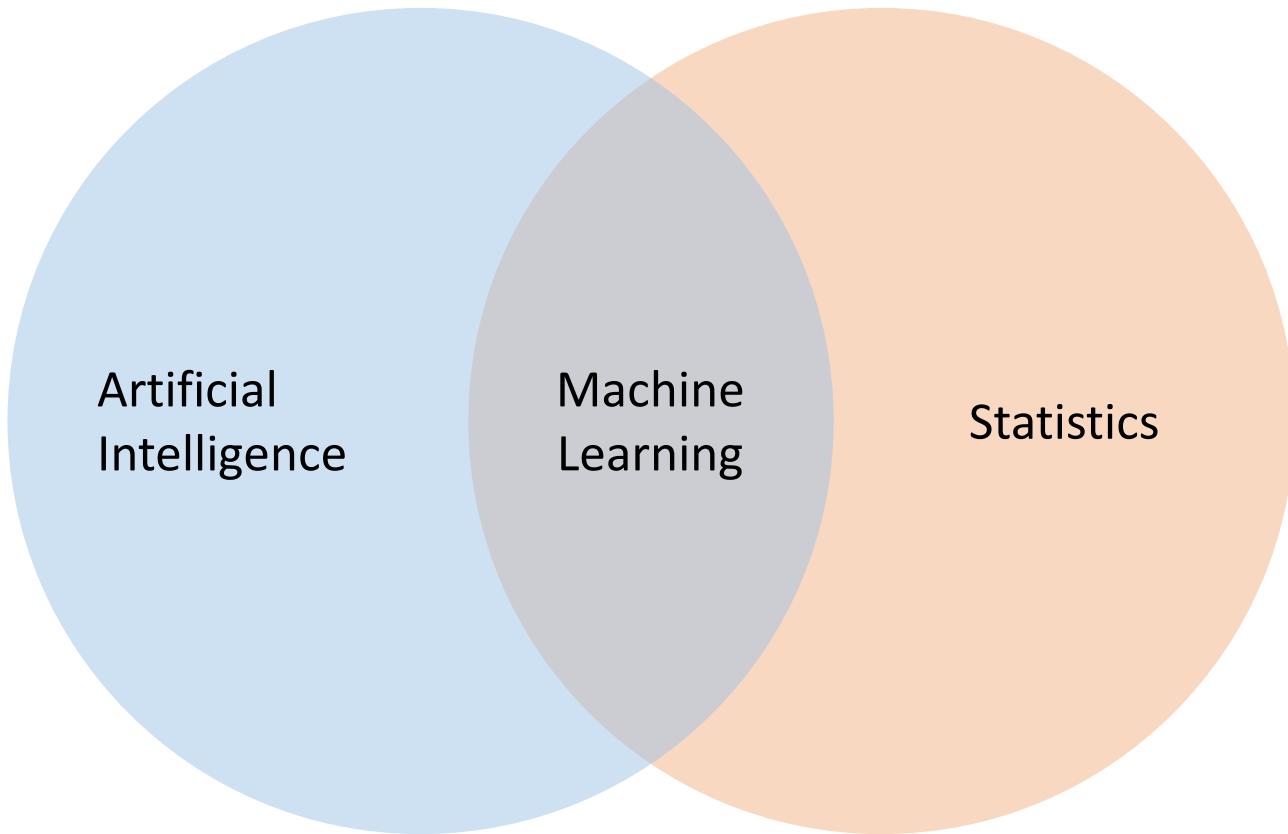


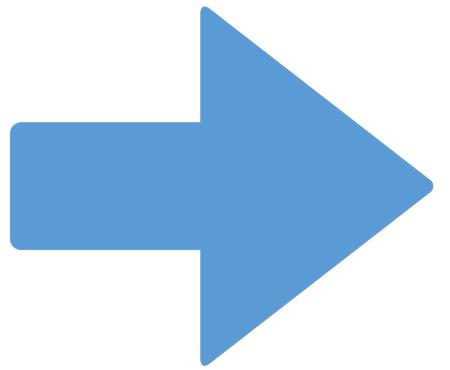
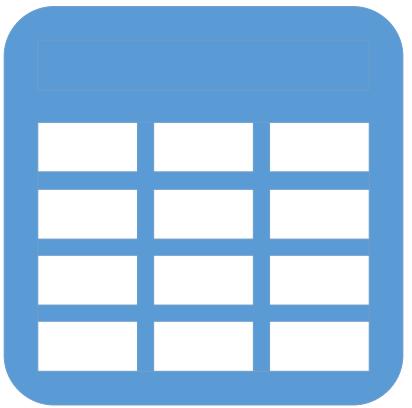
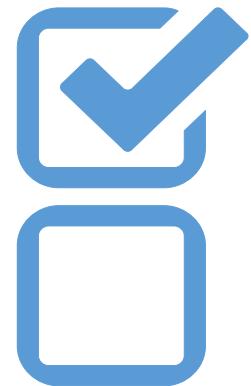
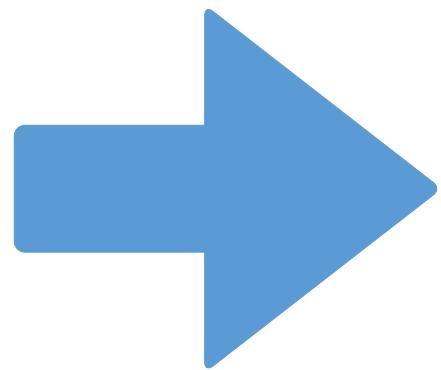
Workshop URL

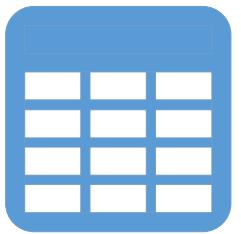
<http://www.matthewrenze.com/workshops/practical-machine-learning-with-r/>

Introduction to Machine Learning

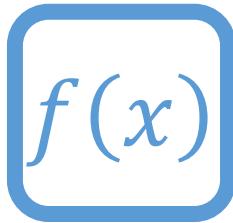
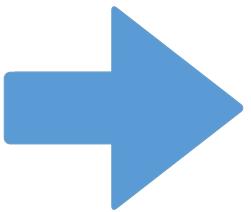
What is machine learning?



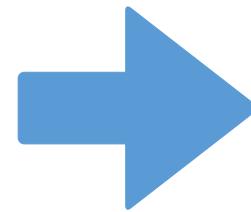
 $f(x)$ 



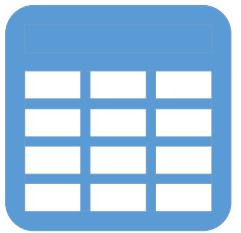
Data



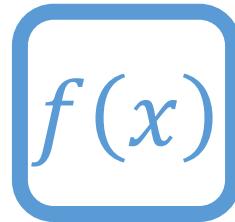
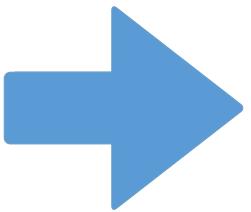
Function



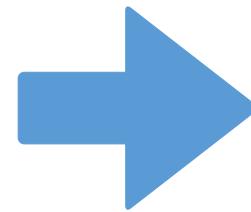
Prediction



Data

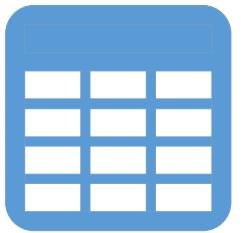


Function

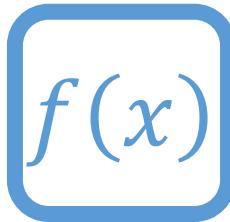
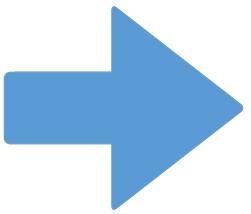


Prediction

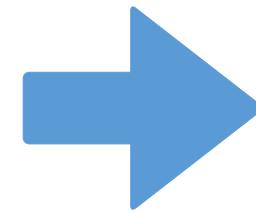




Data



Function



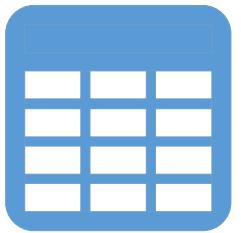
Prediction



Cat



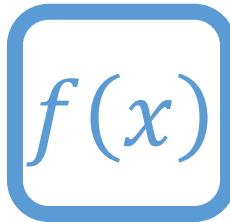
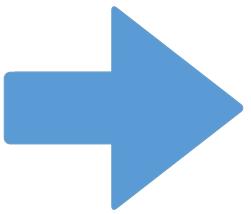
Not cat



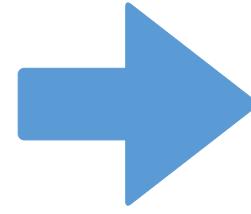
Data



Cat



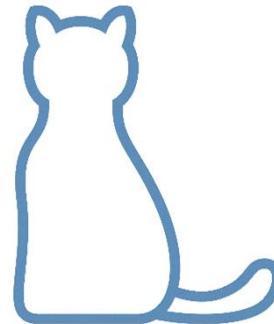
Function

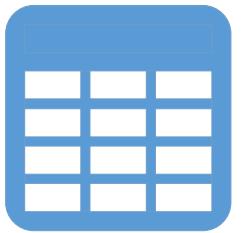


Prediction



Not cat

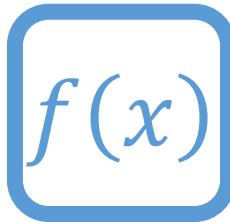
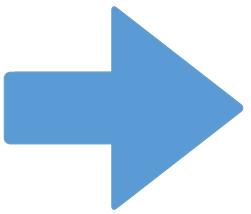




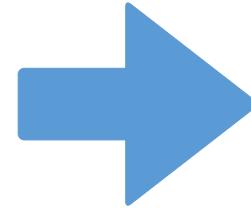
Data



Cat



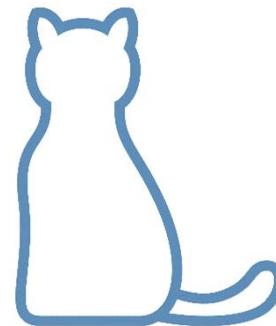
Function



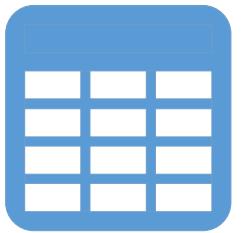
Prediction



Not cat



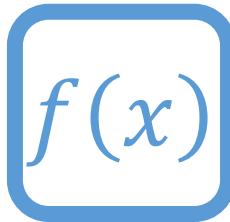
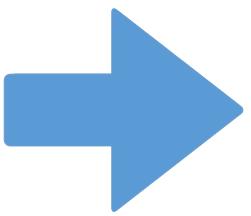
Is cat?



Data



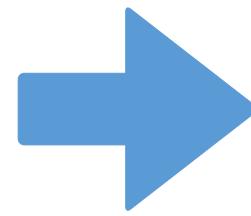
Cat



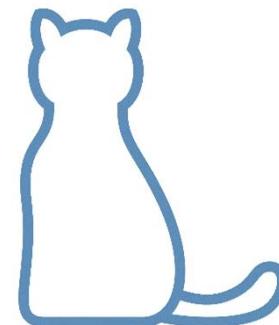
Function



Not cat

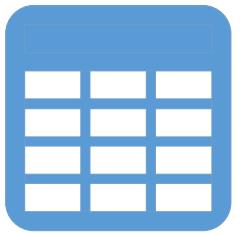


Prediction



Is cat?

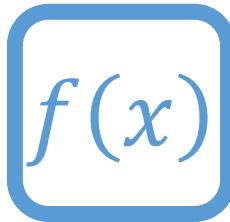
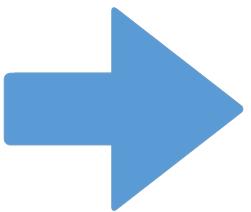




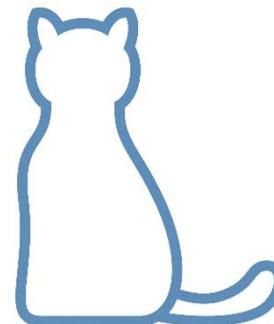
Data



Cat



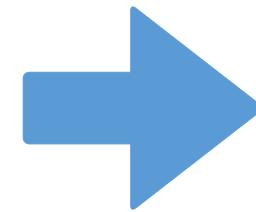
Function



Is cat?



Not cat



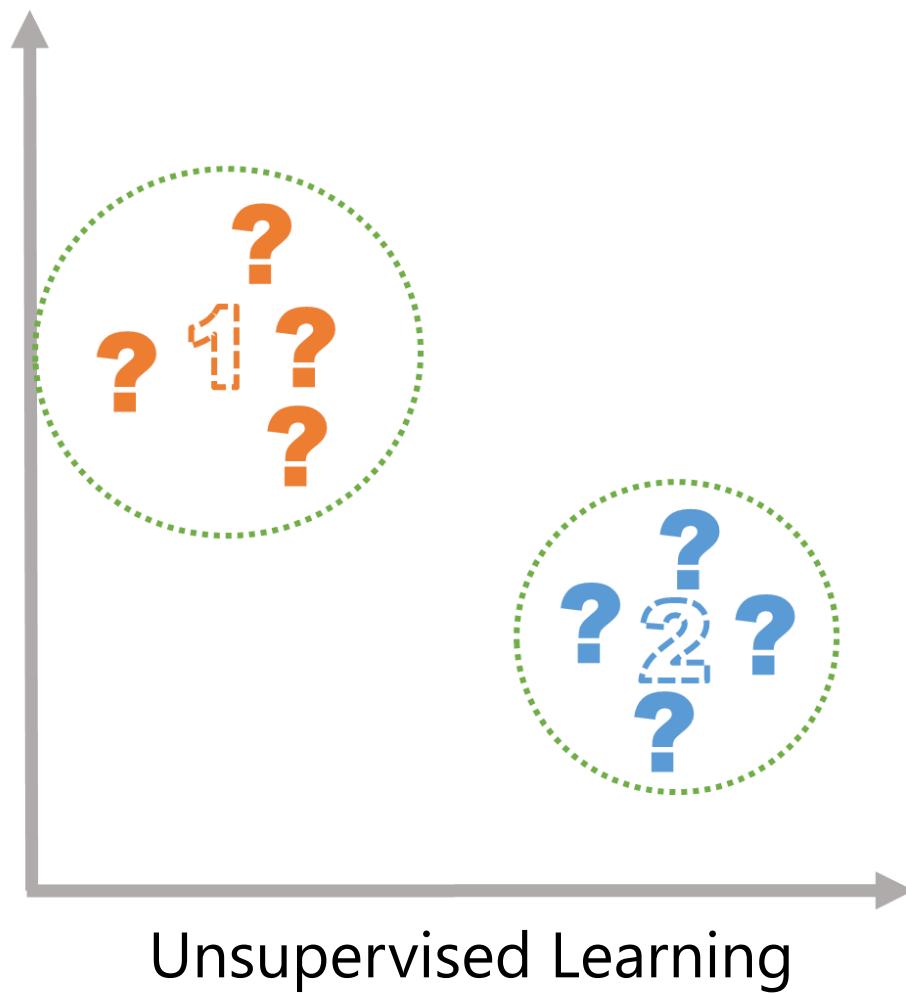
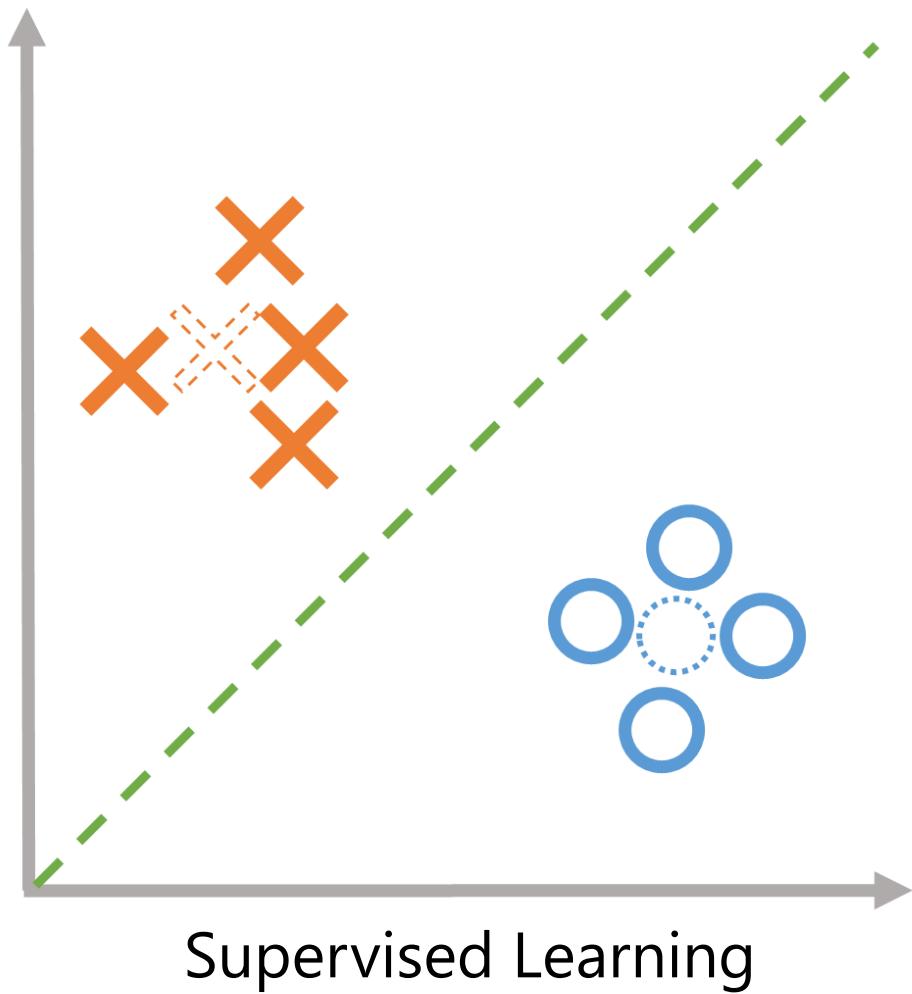
Prediction



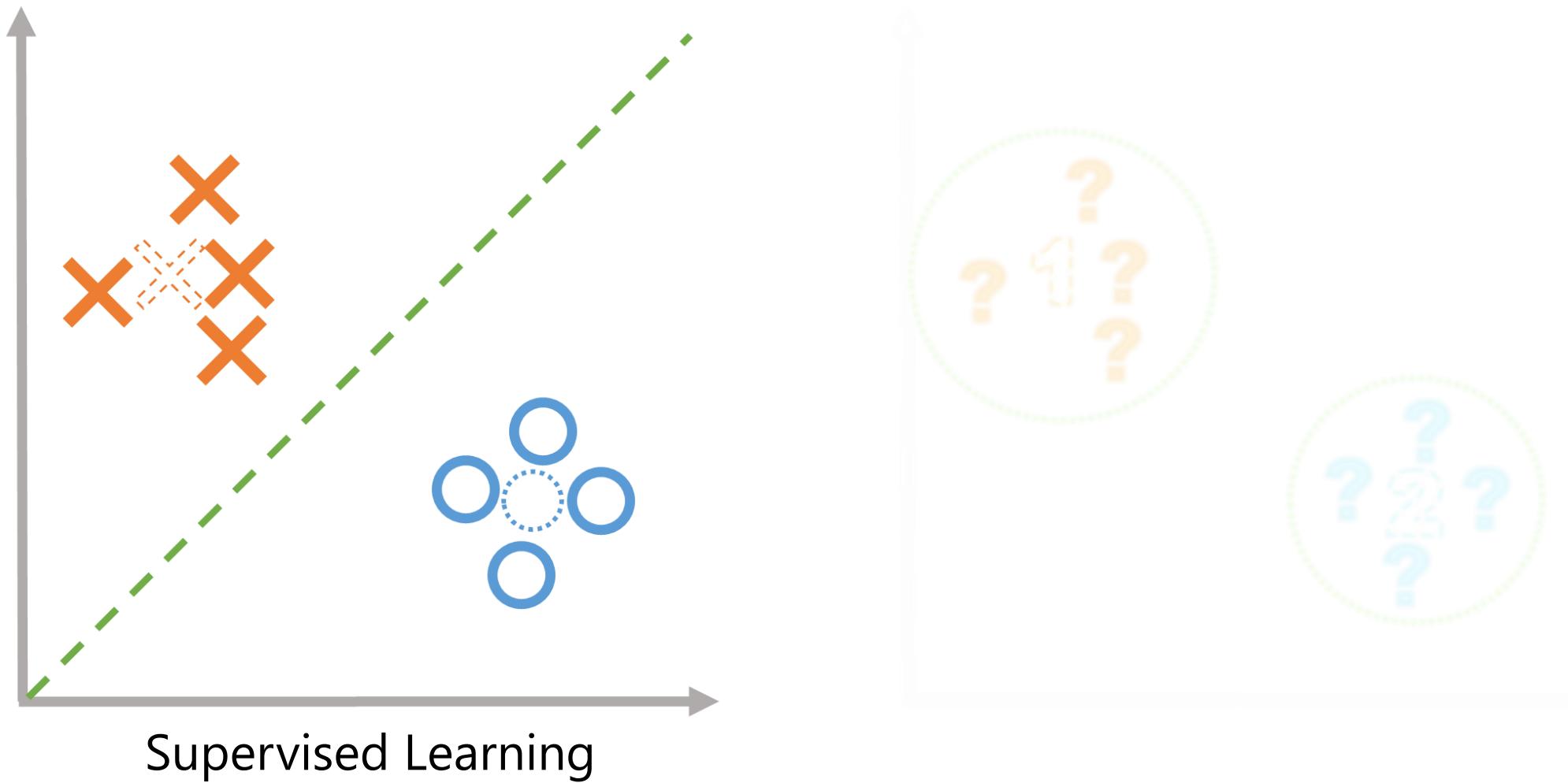
Yes

What types of machine learning exist?

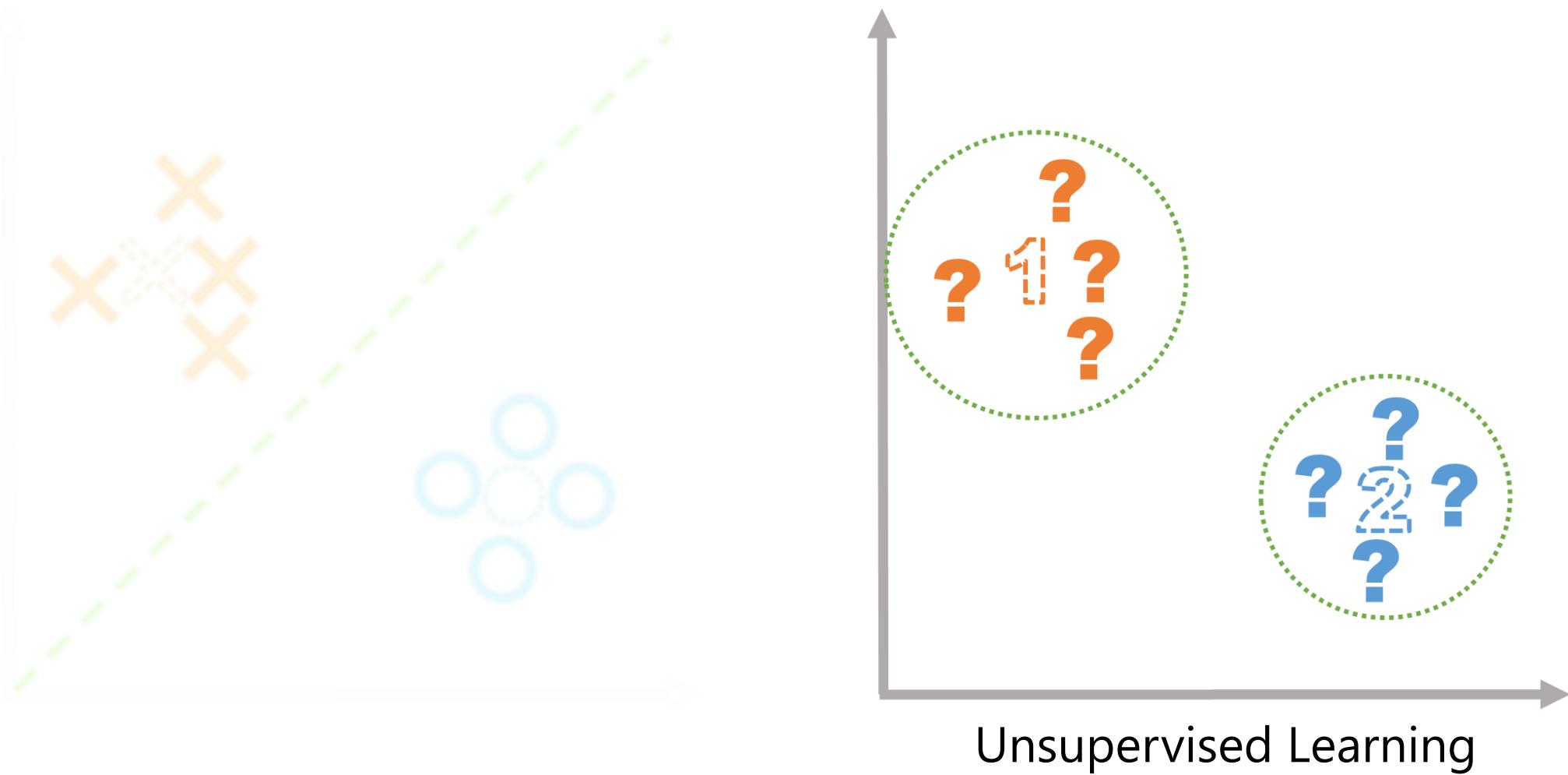
Types of Machine Learning



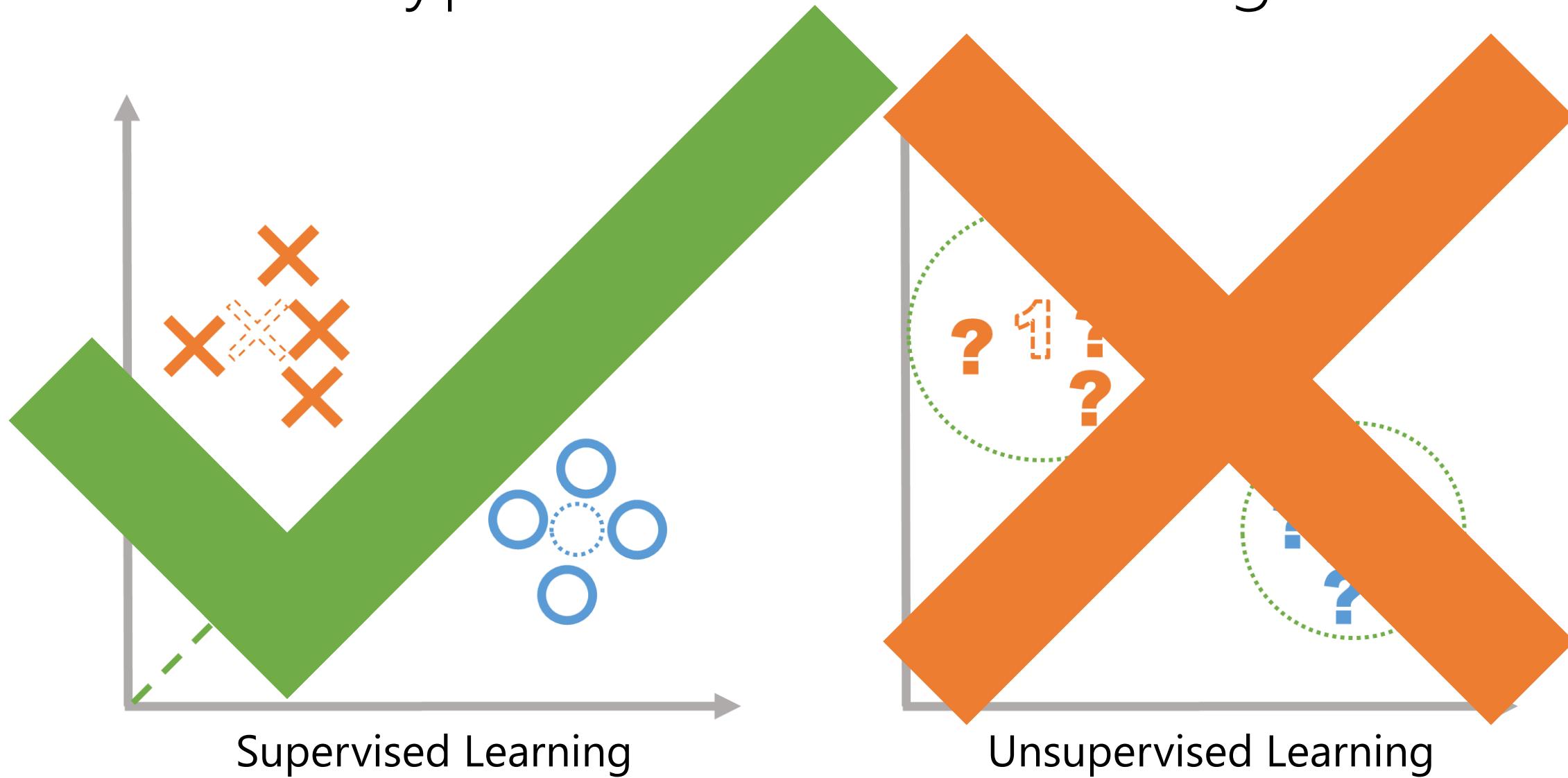
Types of Machine Learning



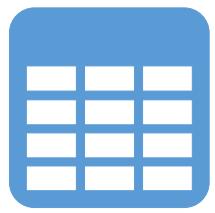
Types of Machine Learning



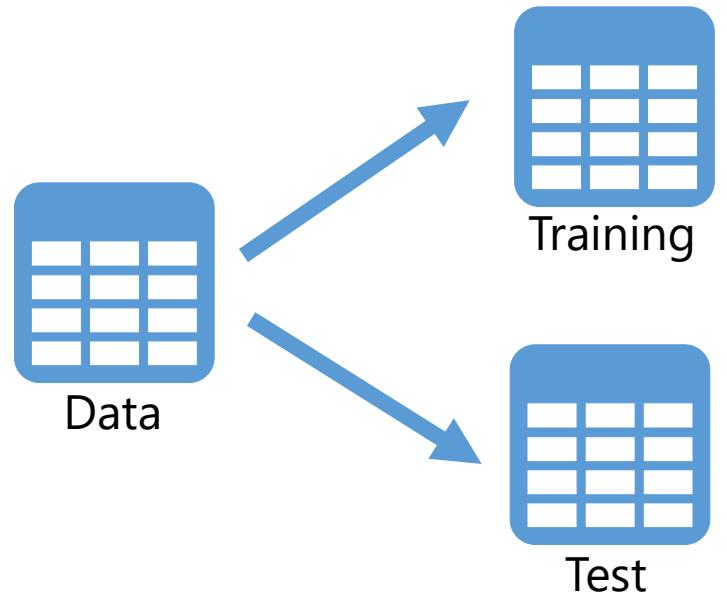
Types of Machine Learning

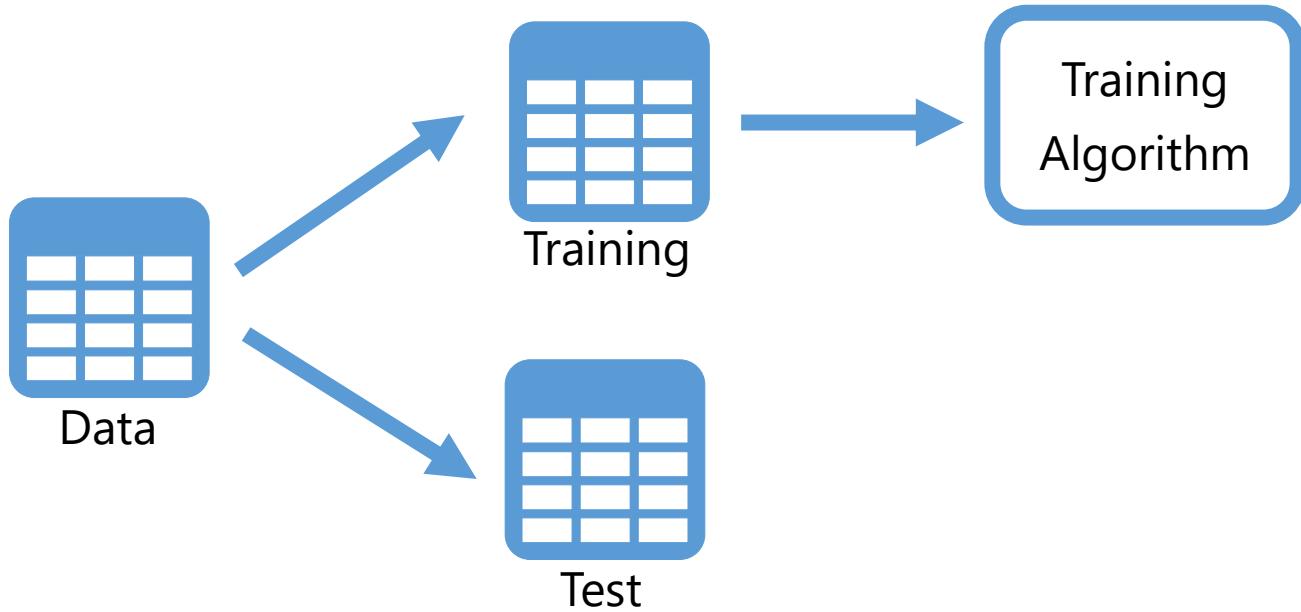


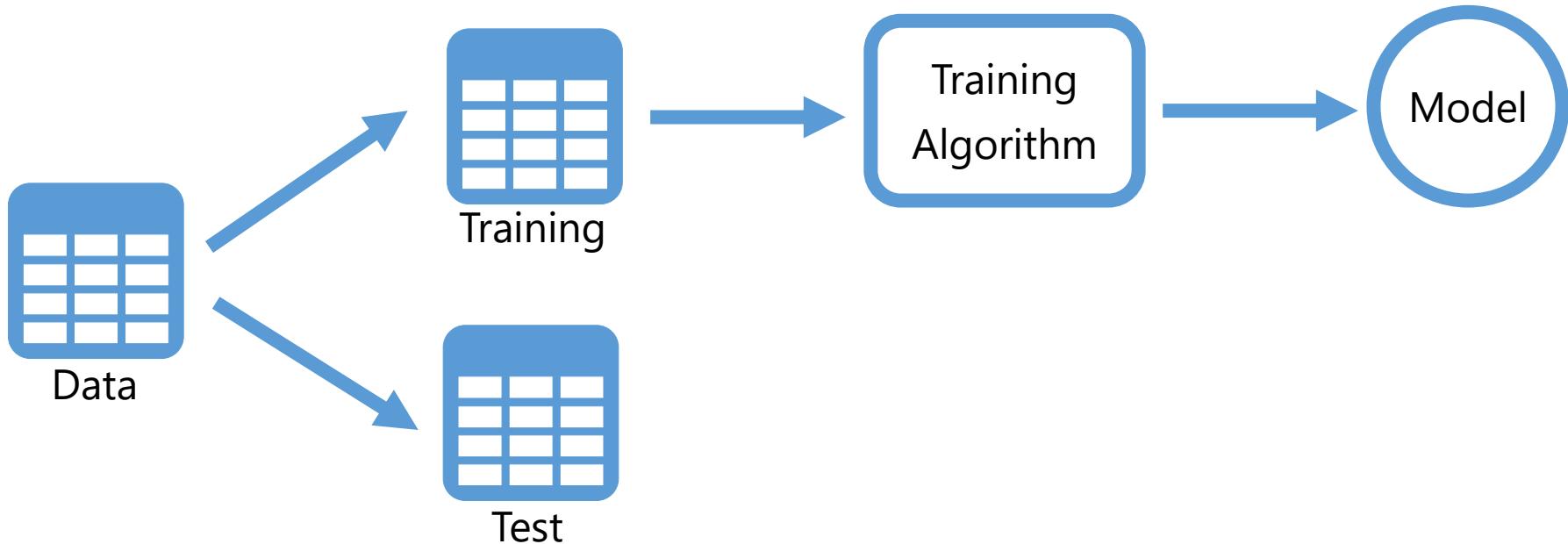
How does machine learning work?

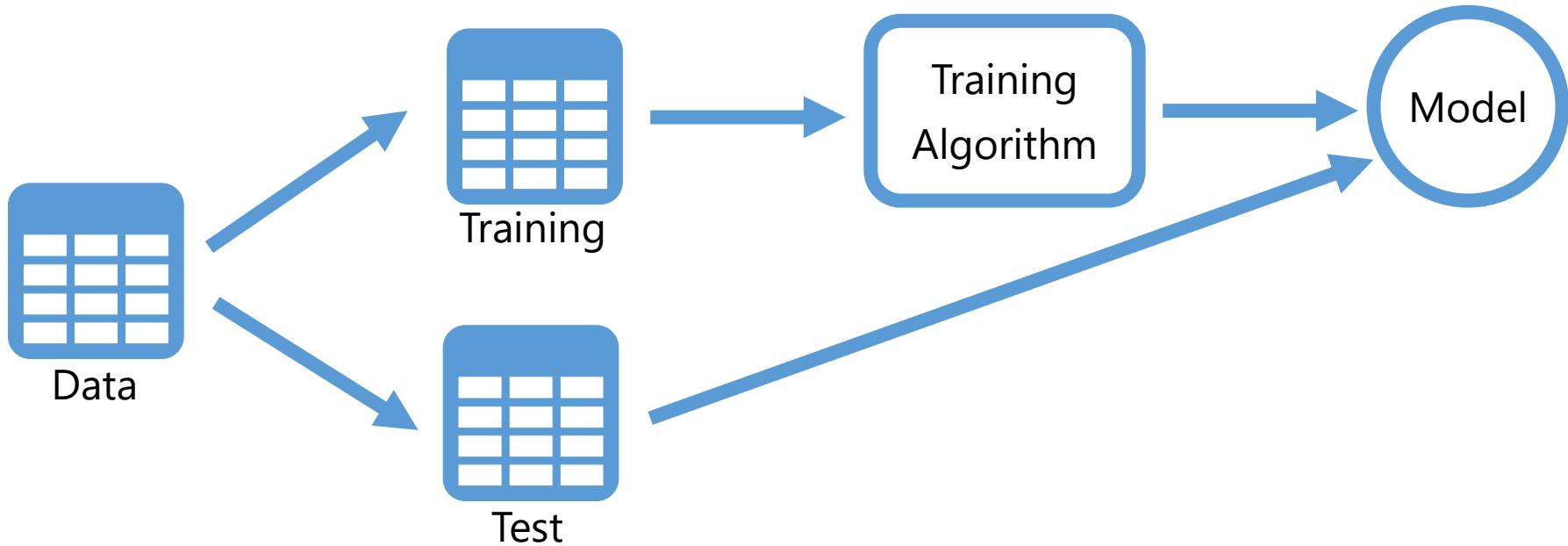


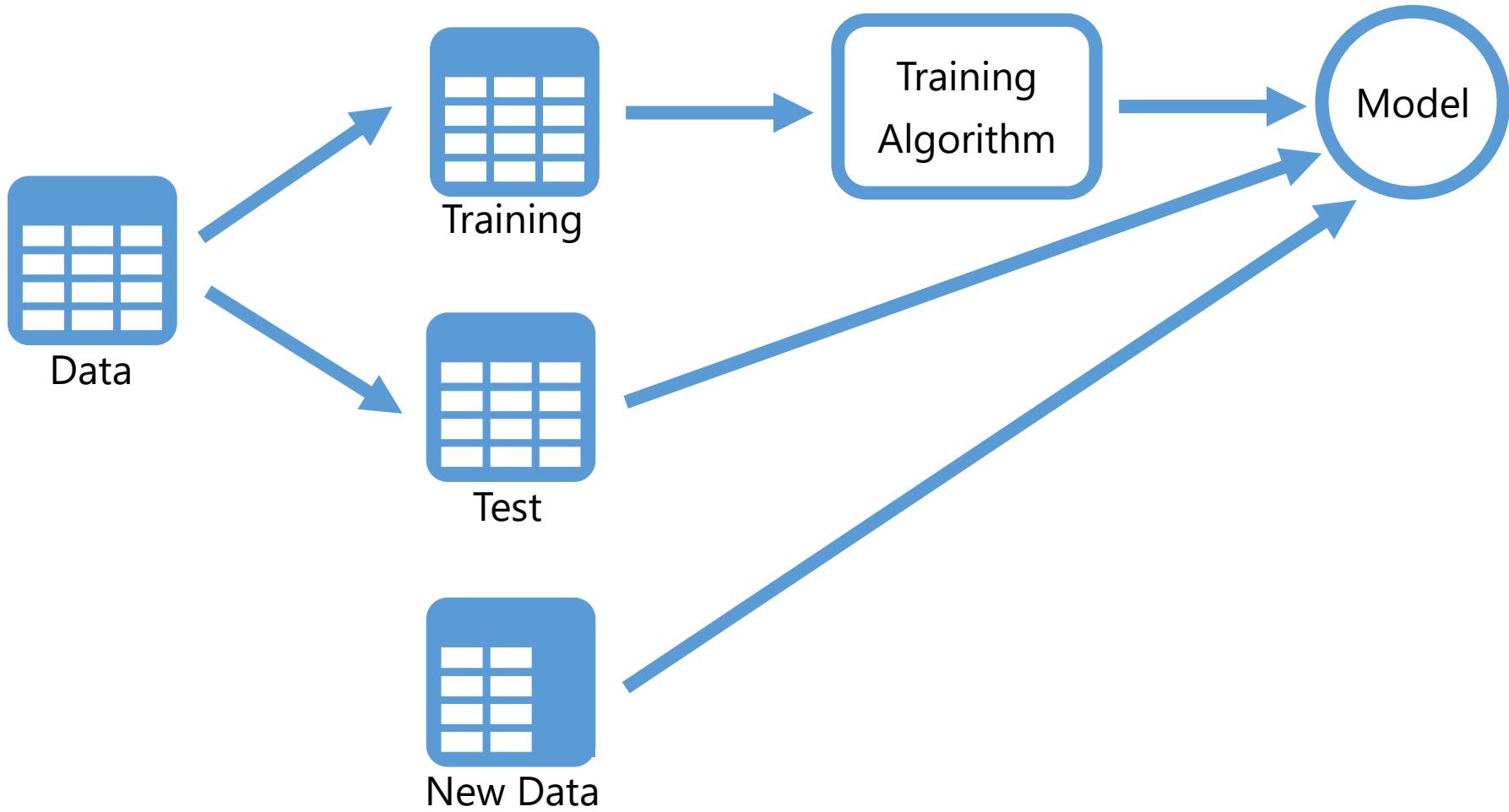
Data

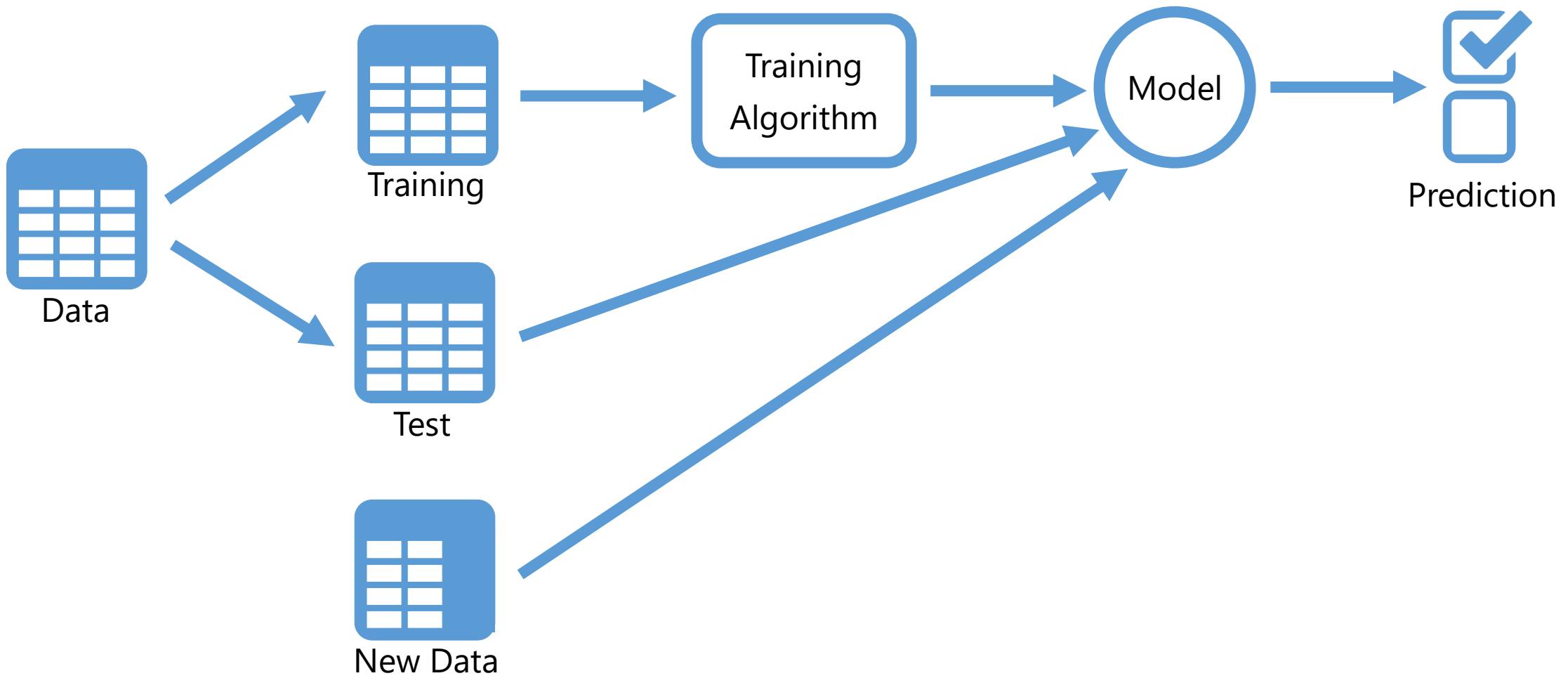


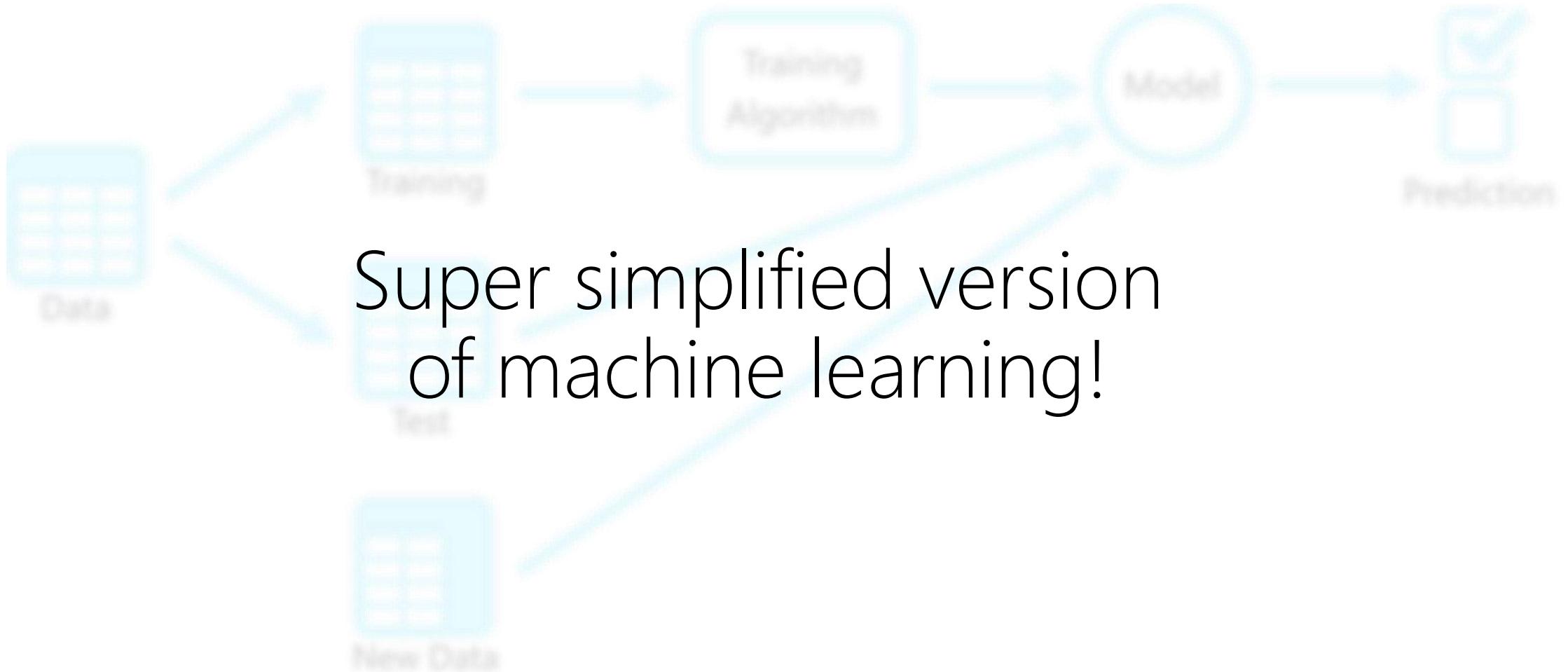




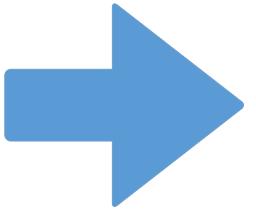
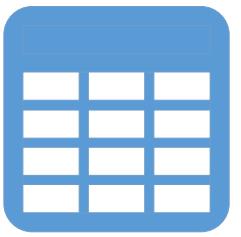
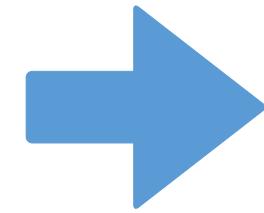






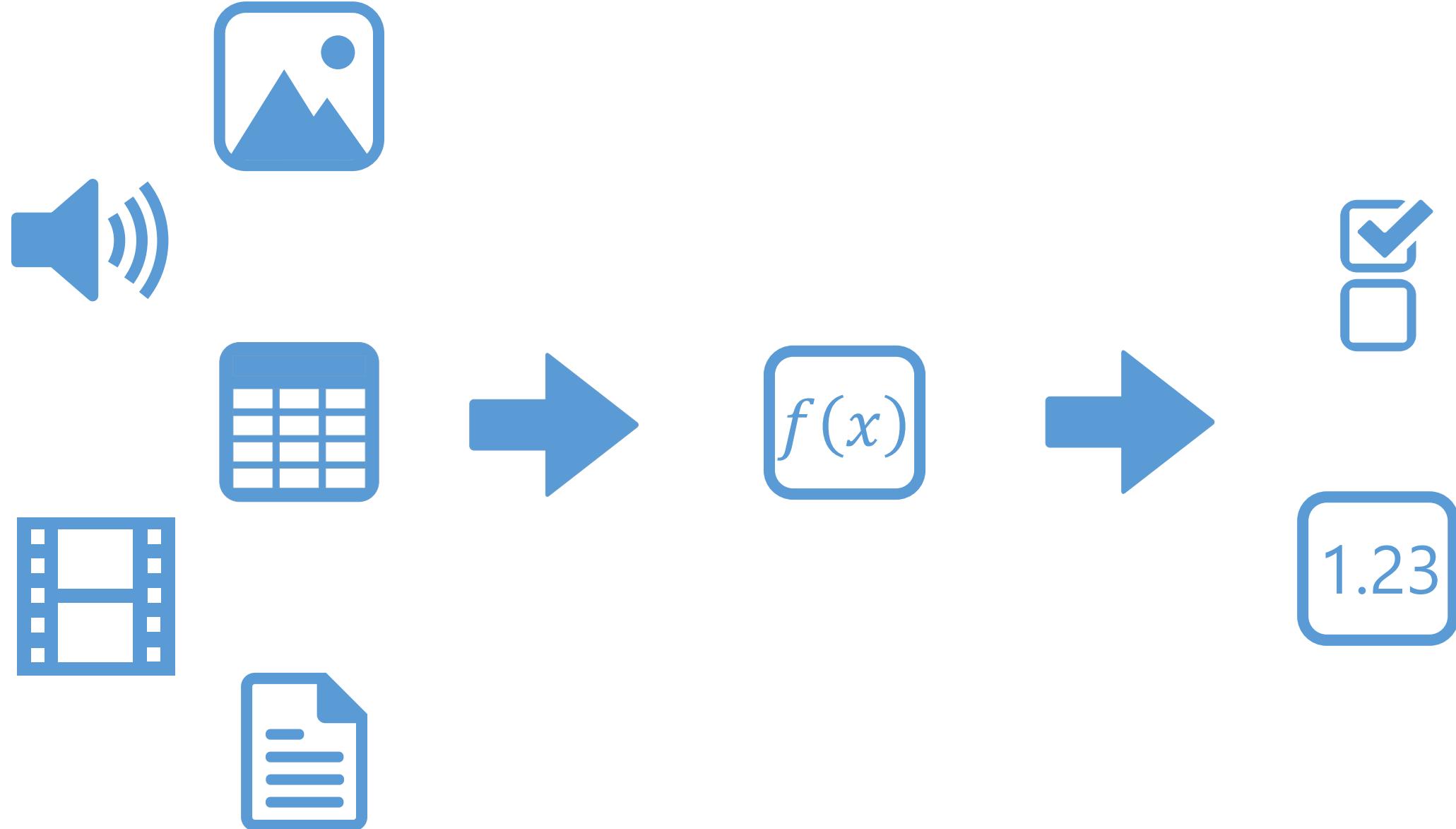


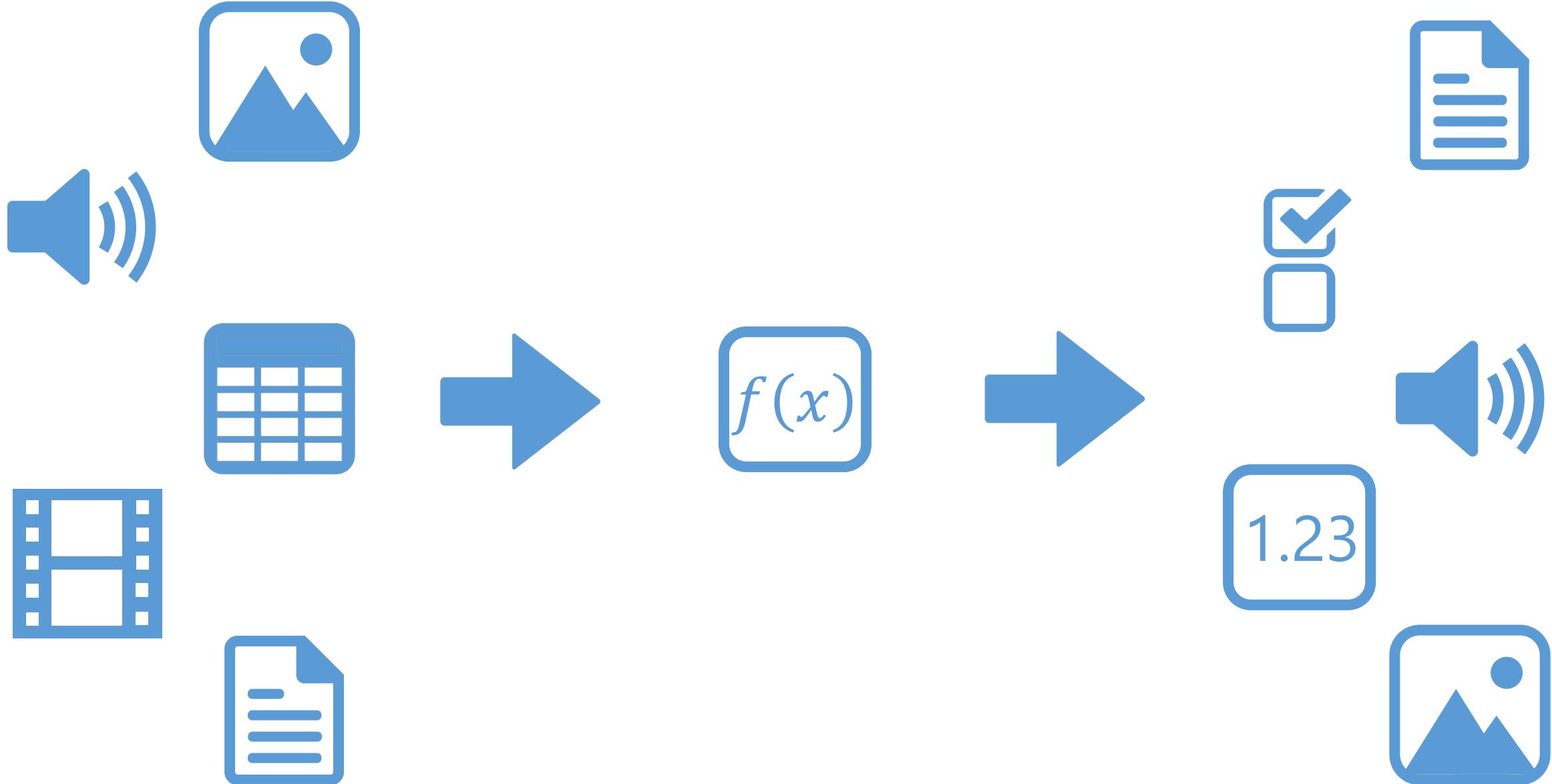
What can machine learning do?

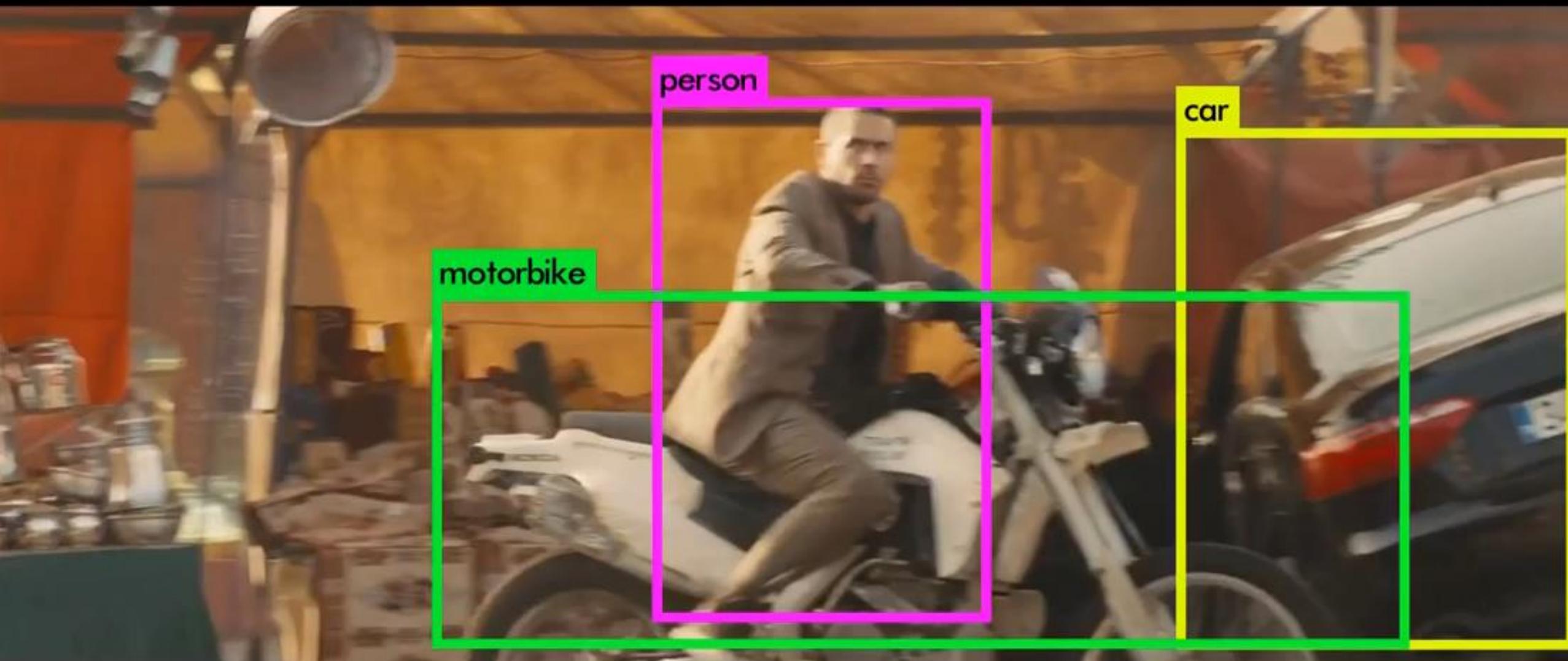
 $f(x)$ 

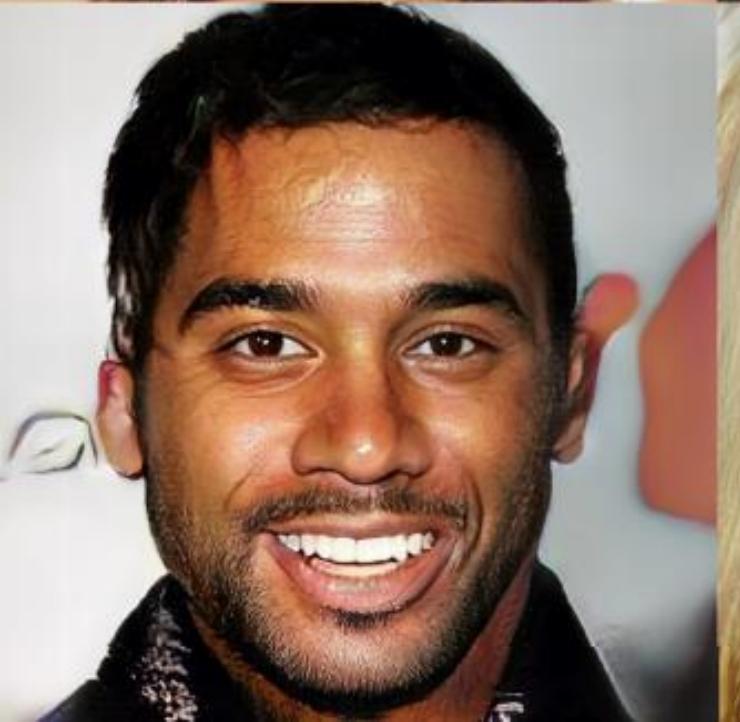
1.23









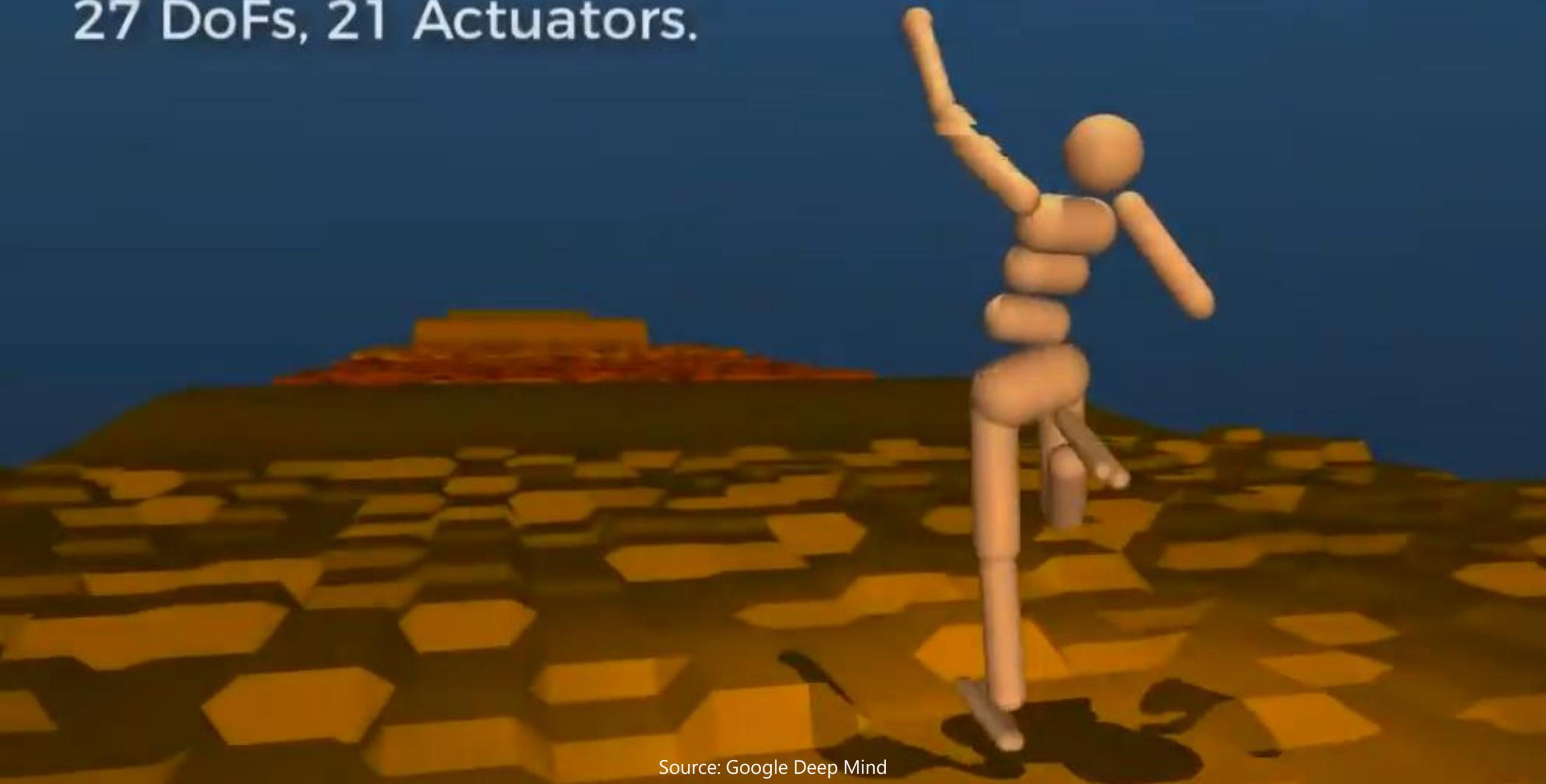


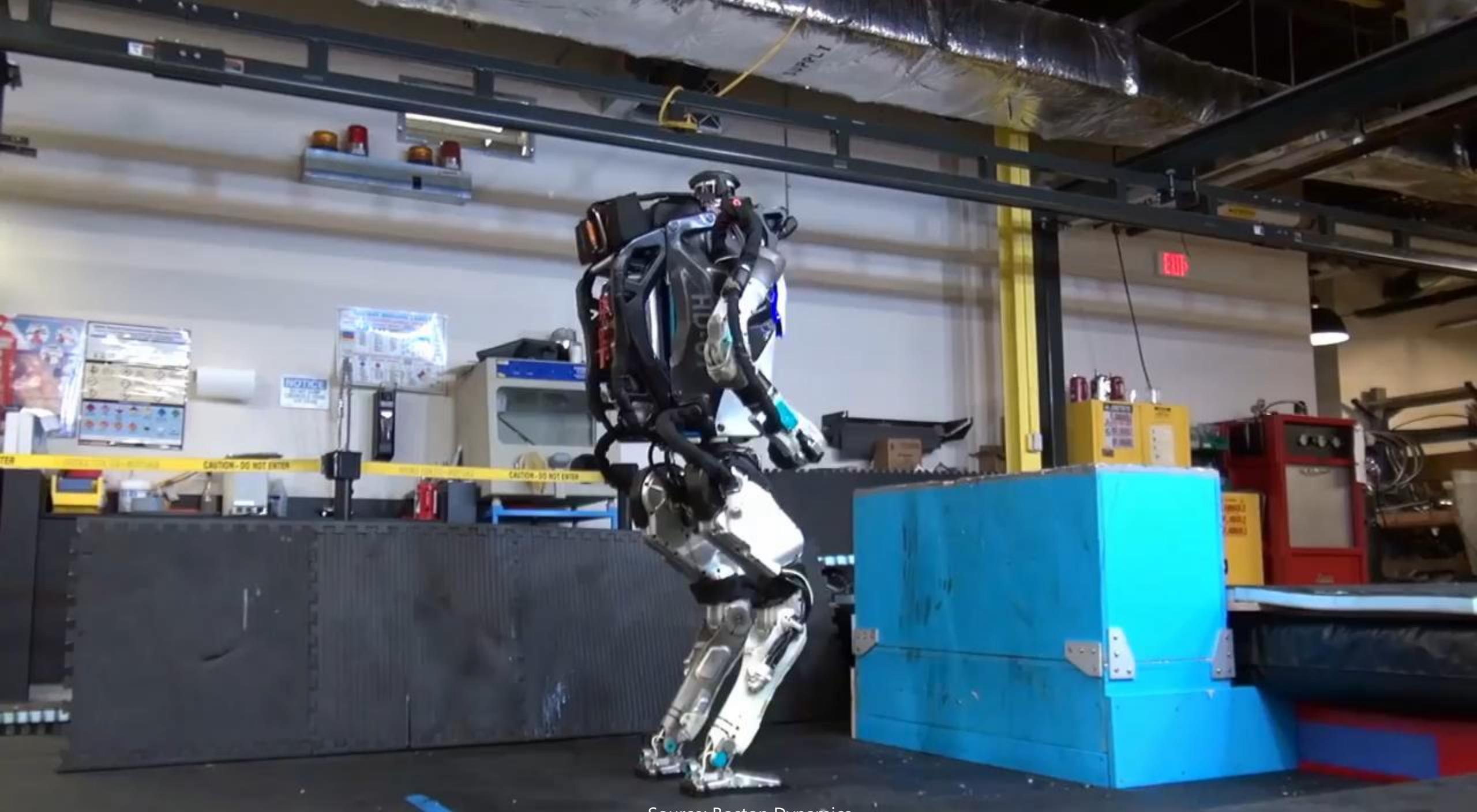
Source: Nvidia



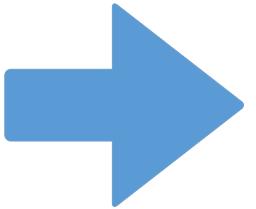
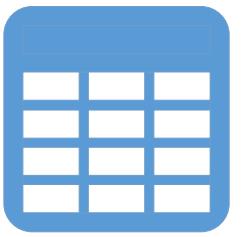
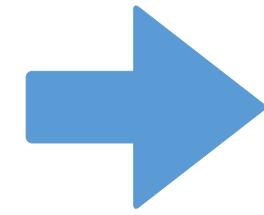
Source: <http://grail.cs.washington.edu/projects/AudioToObama/>

Humanoid:
27 DoFs, 21 Actuators.





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
10,000+ extensions



FREE

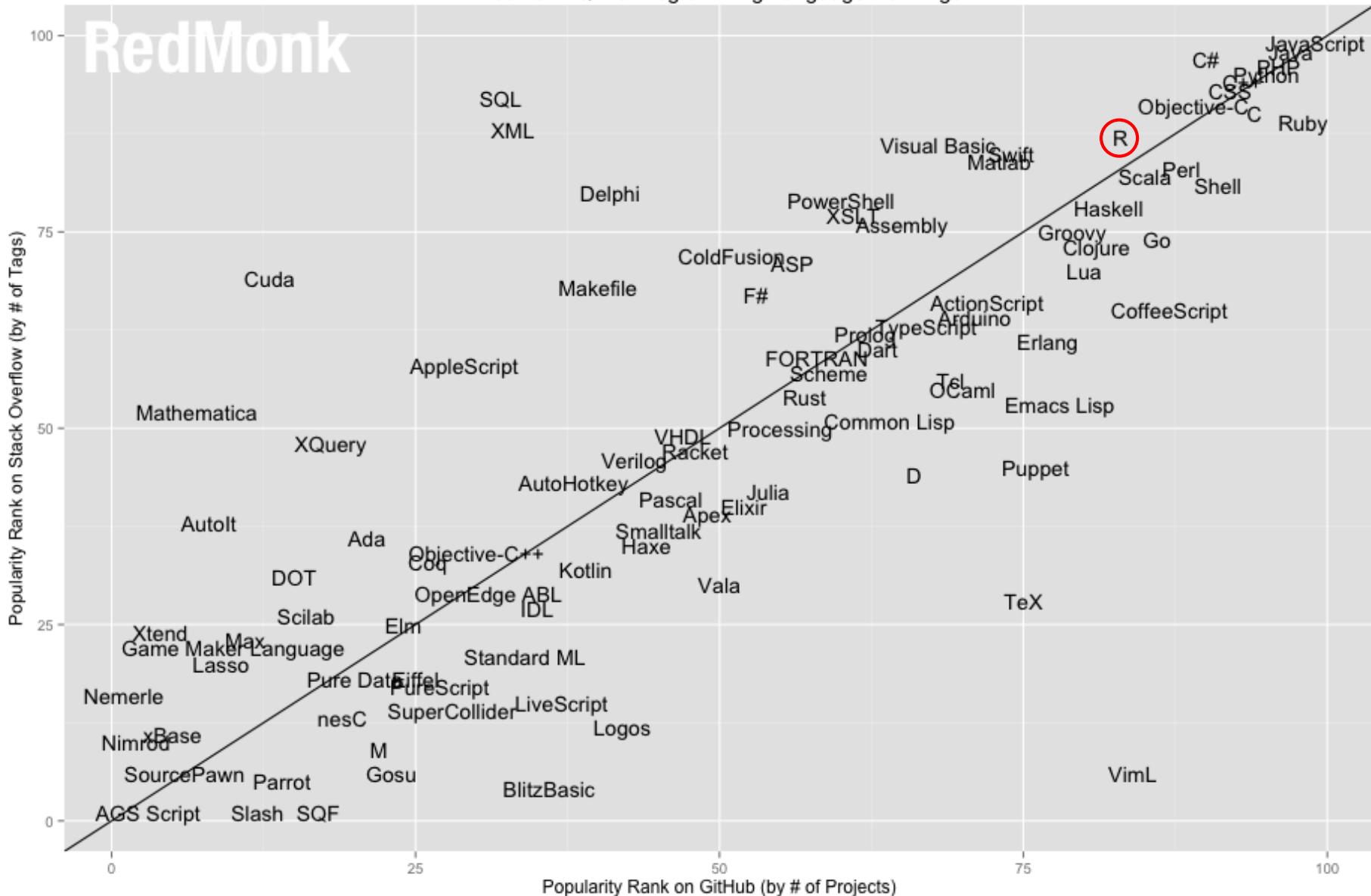


A low-angle photograph of the Statue of Liberty against a clear blue sky. Her right arm is raised high, holding the torch aloft. Her left arm is bent, holding a tablet or smartphone that displays the word "FREE".

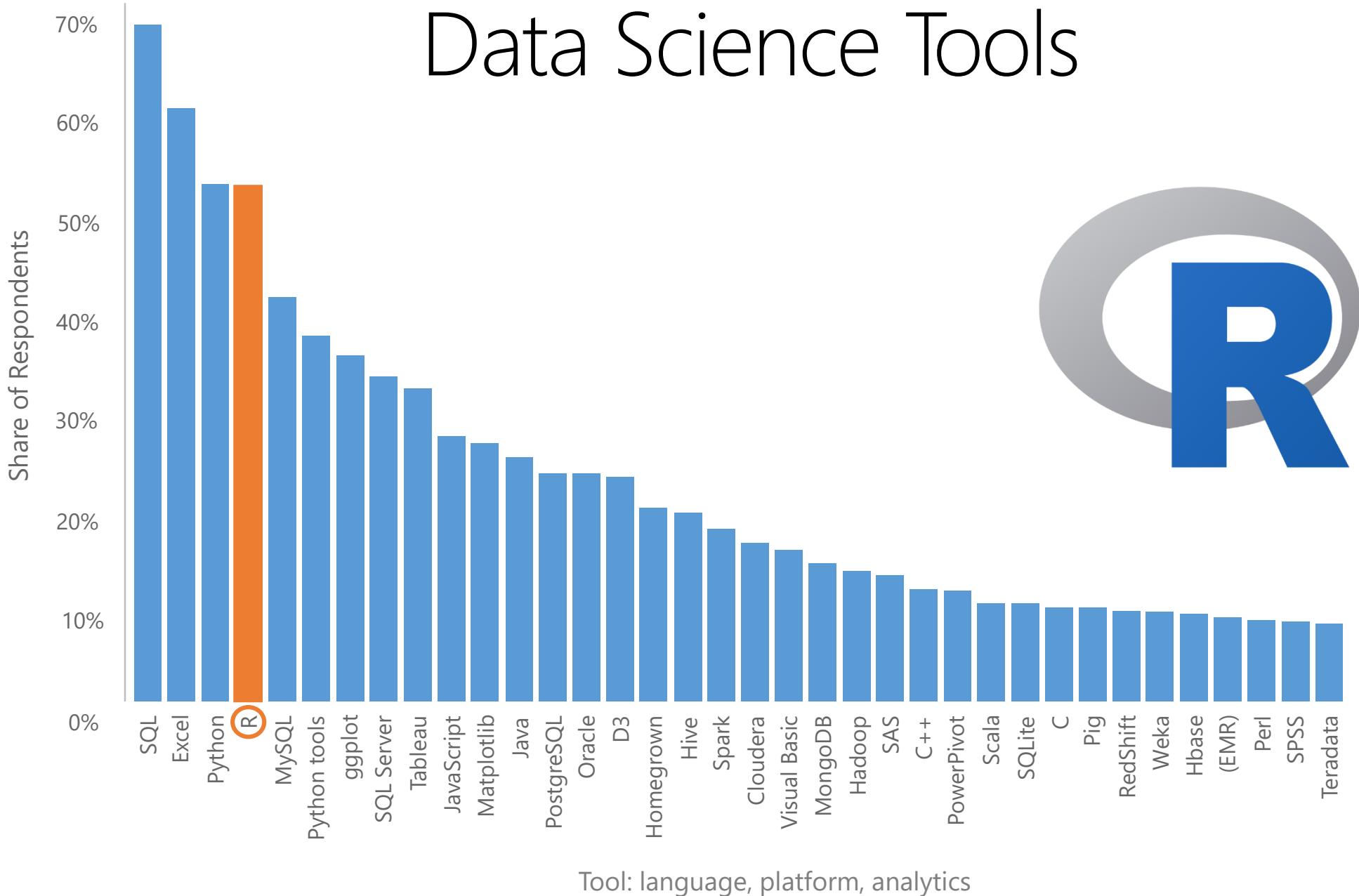
FREE

RedMonk

RedMonk Q116 Programming Language Rankings



Data Science Tools



Source: O'Reilly 2015 Data Science Salary Survey

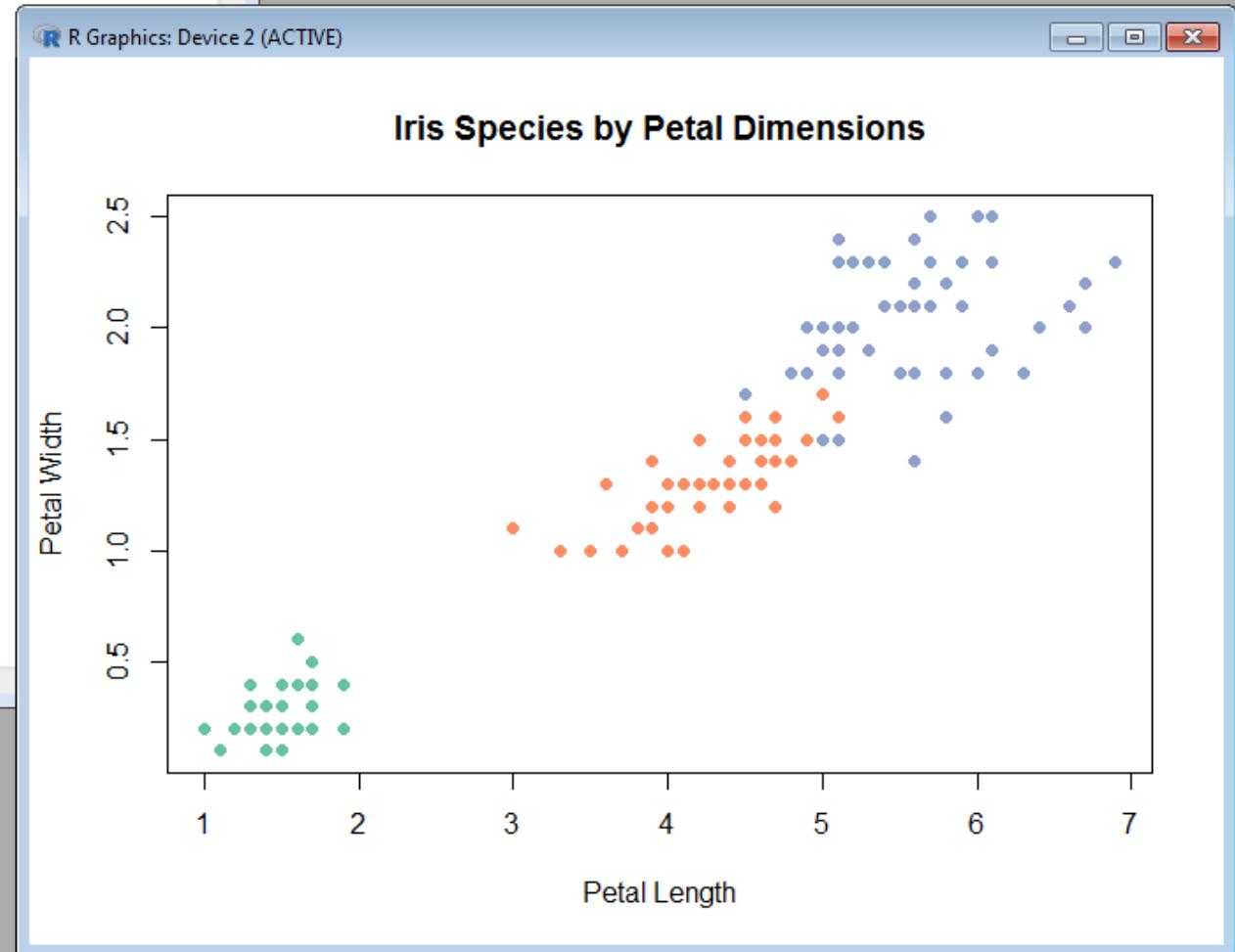


R Console

```
> # 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
```



RStudio

File Edit Code View Plots Session Build Debug Tools Help

Script.R * Go to file/function Addins Project: (None)

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 Global Environment Data Values palette 150 obs. of 5 variables chr [1:3] "#66C2A5" "#FC8D62" "#8DA0C... Files Plots Packages Help Viewer Publish Iris Species by Petal Dimensions Petal Width Petal Length

The figure is a scatter plot titled "Iris Species by Petal Dimensions". The x-axis is labeled "Petal Length" and ranges from 1 to 7. The y-axis is labeled "Petal Width" and ranges from 0.5 to 2.5. There are three distinct clusters of data points representing different iris species: setosa (green dots), versicolor (orange dots), and virginica (blue dots). The setosa species has the lowest petal lengths and widths, ranging approximately from 1.0 to 2.0. The versicolor species has intermediate values, ranging approximately from 3.0 to 5.5. The virginica species has the highest values, ranging approximately from 5.0 to 7.0.

Script.R - Microsoft Visual Studio

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

Matthew Renze

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)
```

R Interactive

```
> # 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
>
```

Variable Explorer

Name	Value	Class	Type
iris	150 obs. of 5 variables	data.frame	list
palette	chr [1:3] "#6C2A5" "#FC8D62" "#8DA0CF	character	character

R Plot

Iris Species by Petal Dimensions

A scatter plot titled "Iris Species by Petal Dimensions". The x-axis is labeled "Petal Length" and ranges from 1 to 7. The y-axis is labeled "Petal Width" and ranges from 0.5 to 2.5. The plot shows three distinct clusters of data points corresponding to the Iris species: Setosa (green), Versicolor (orange), and Virginica (blue). The data points are scattered across the plot area, with a general trend where Petal Length increases as Petal Width increases.

Solution Explorer R Plot R Package Manager R Help

Error List Output Azure App Service Activity

Ready Ln 30 Col1 Ch1 INS ↑ 7 ⌂ 0 ⌂ Root ⌂ master ⌂

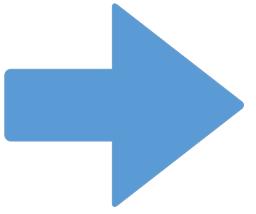
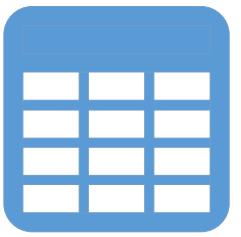
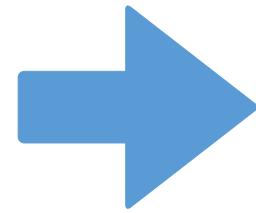
Demo 1

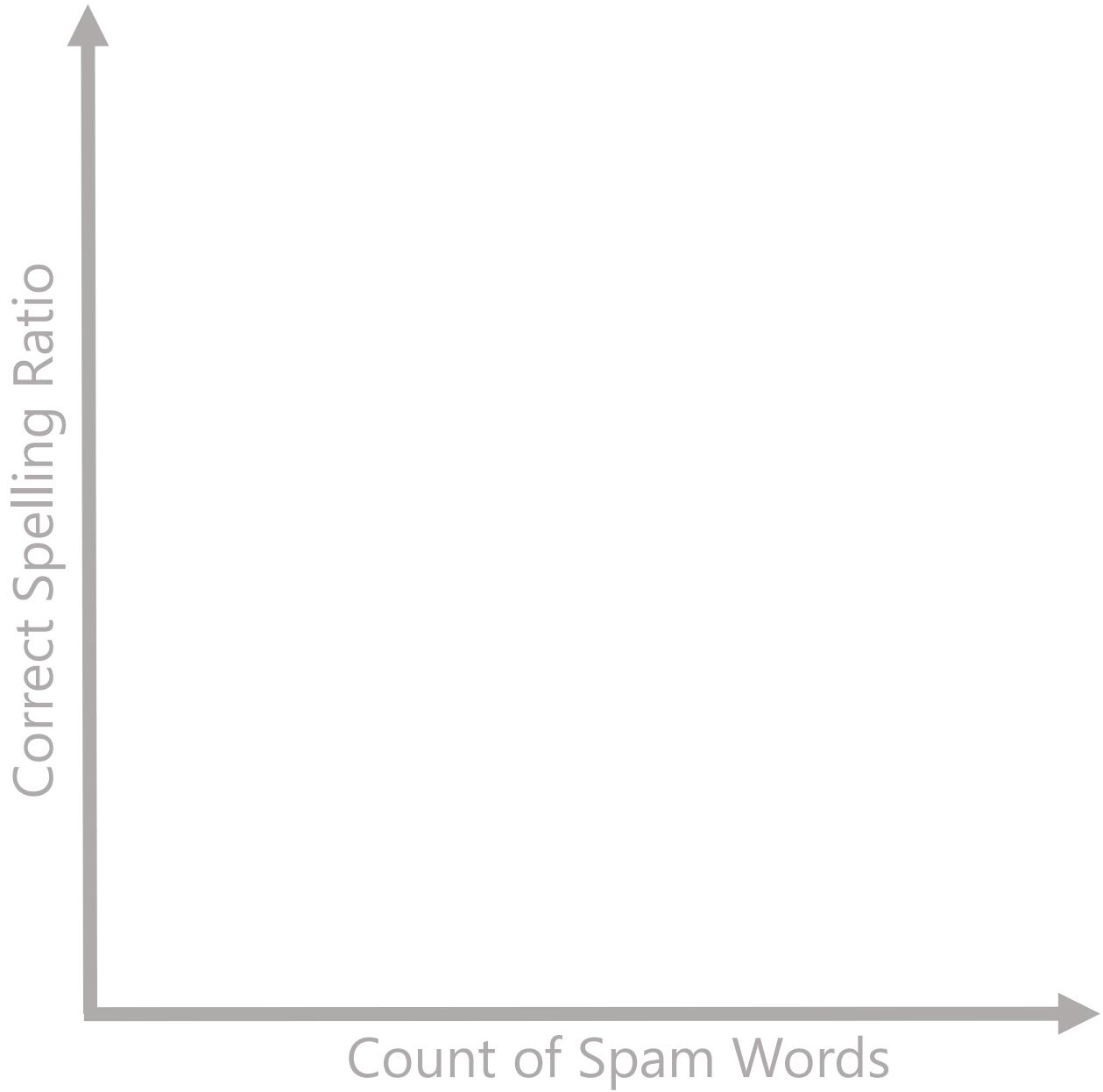
R Language Basics

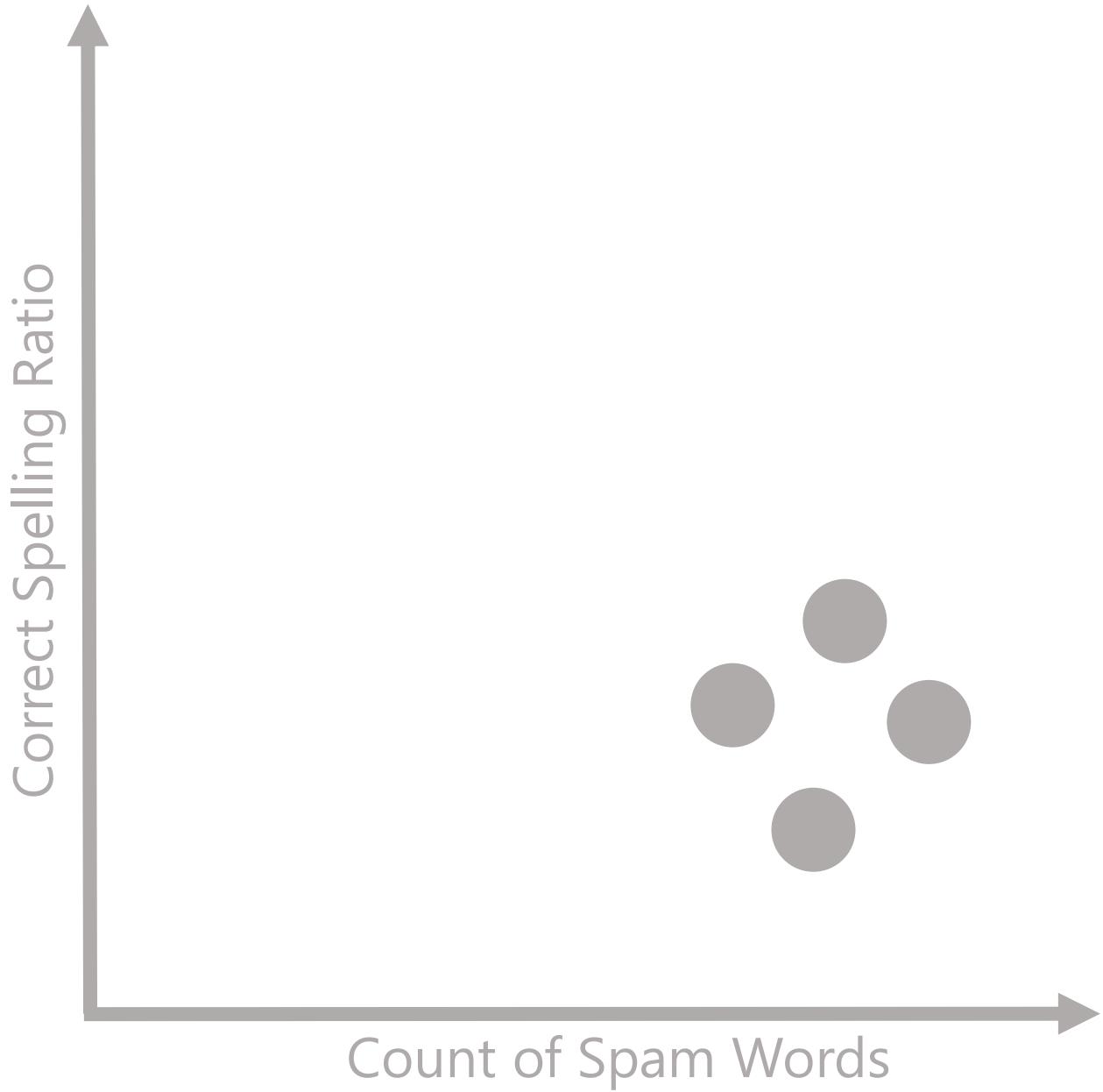
Lab 1

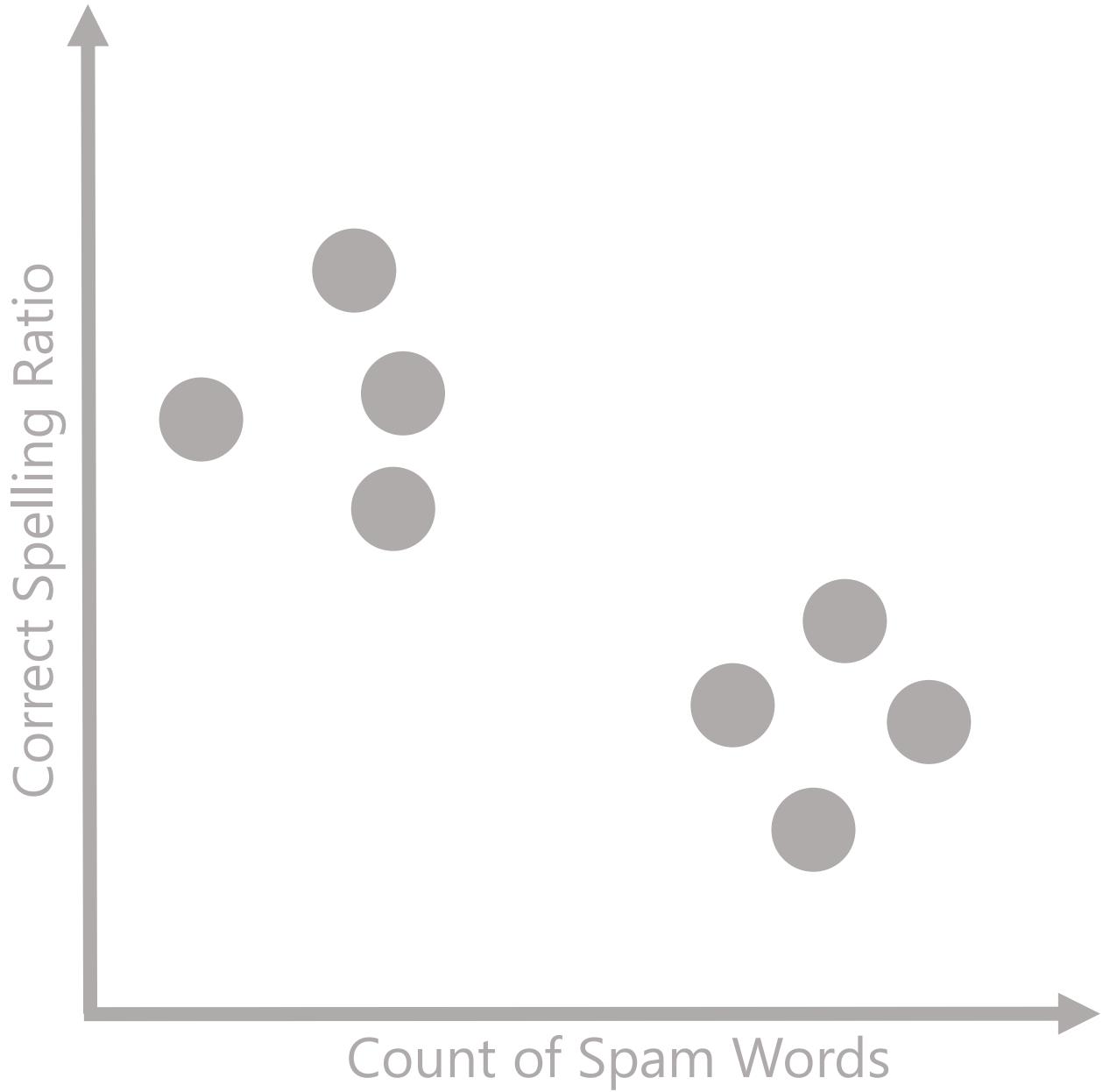
R Language Basics

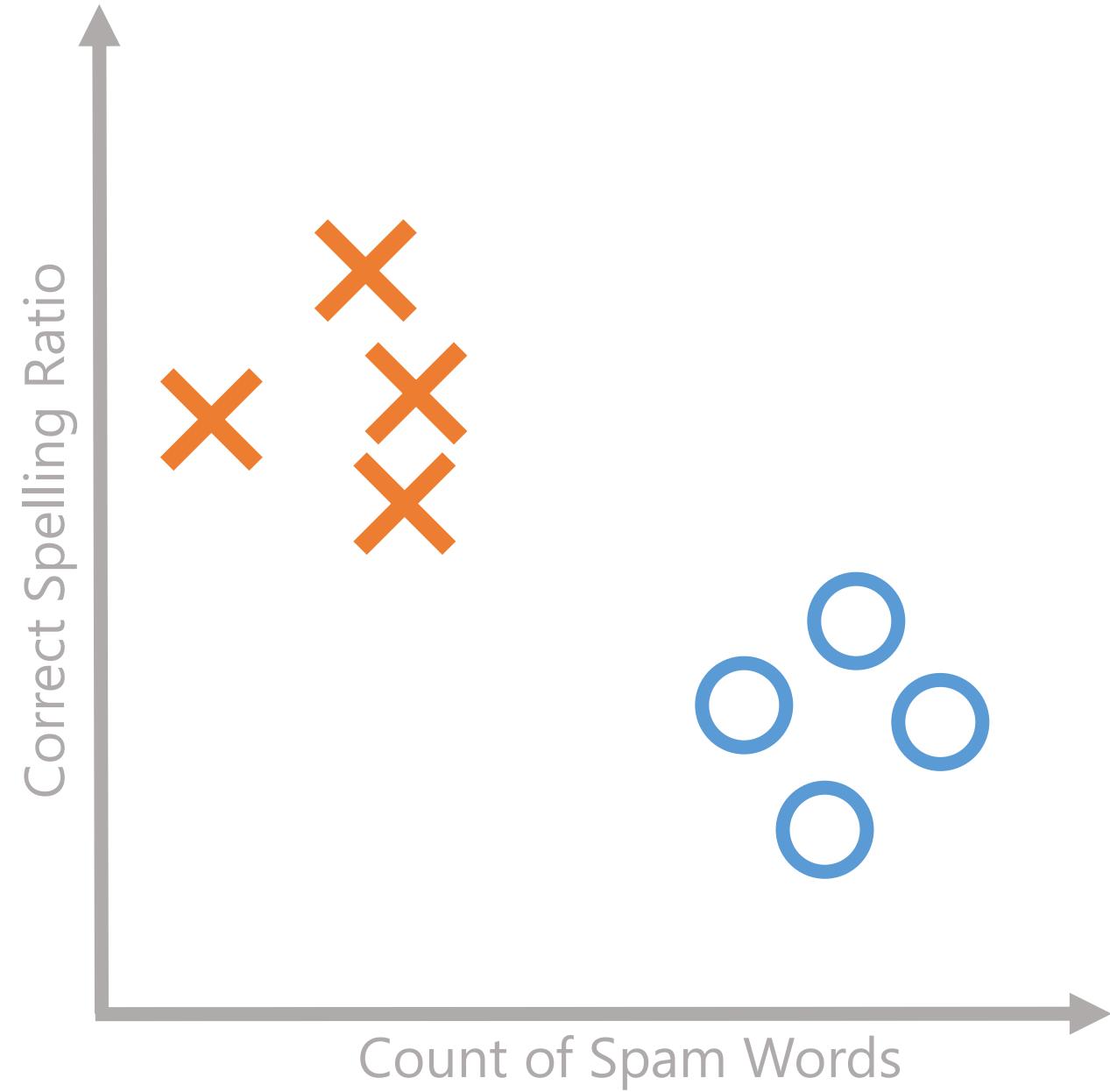
Classification

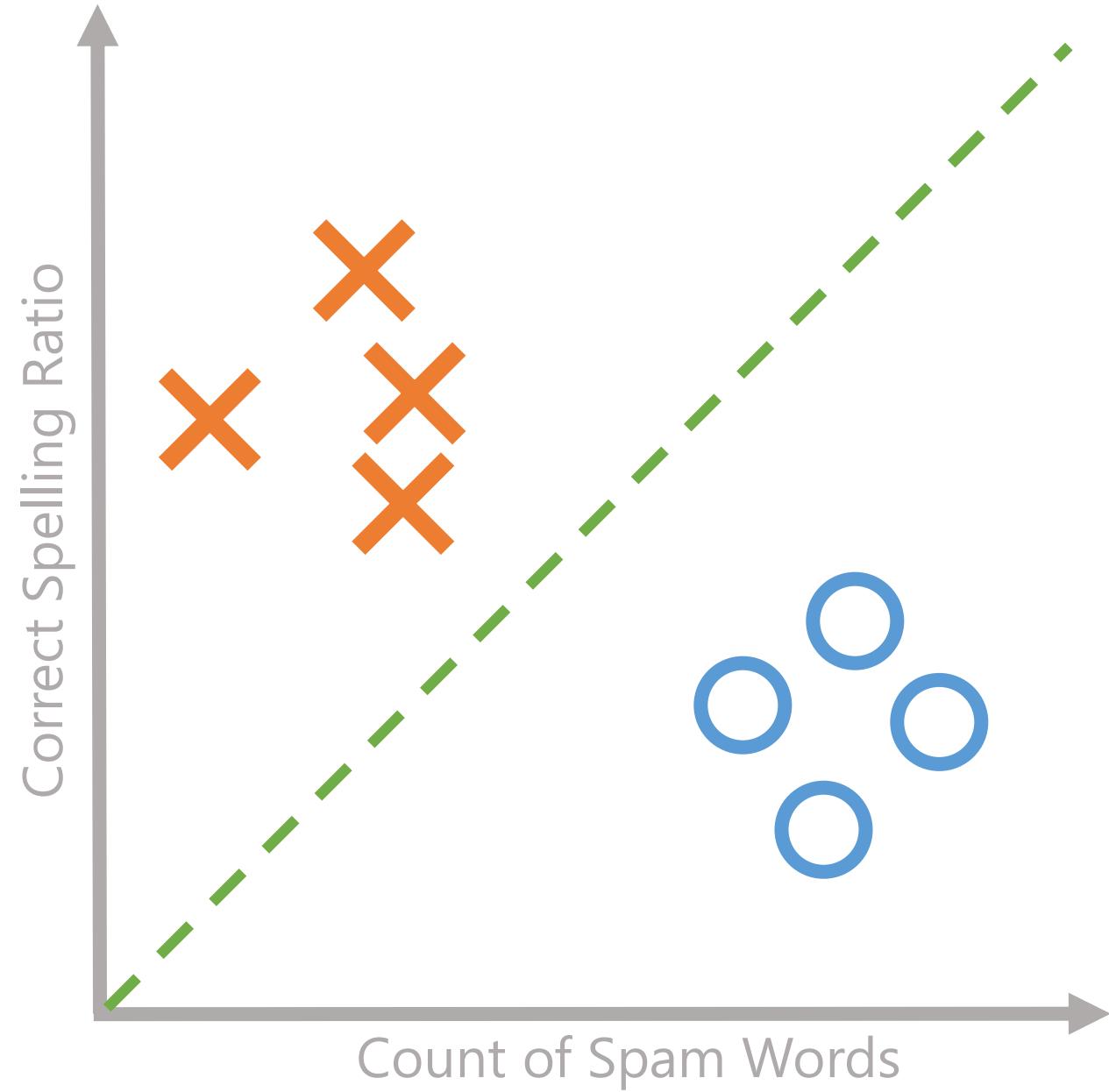
 $f(x)$ 

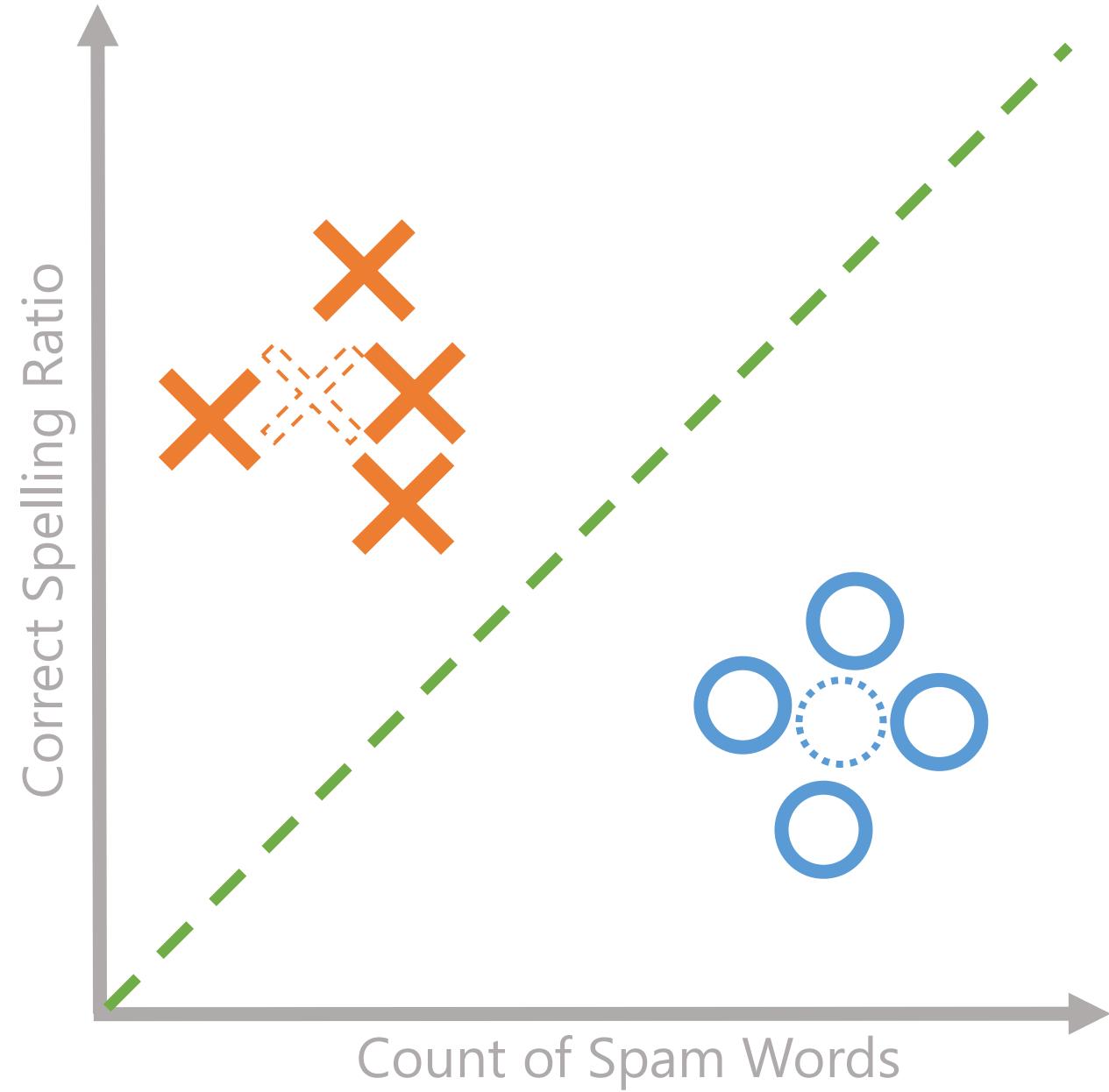






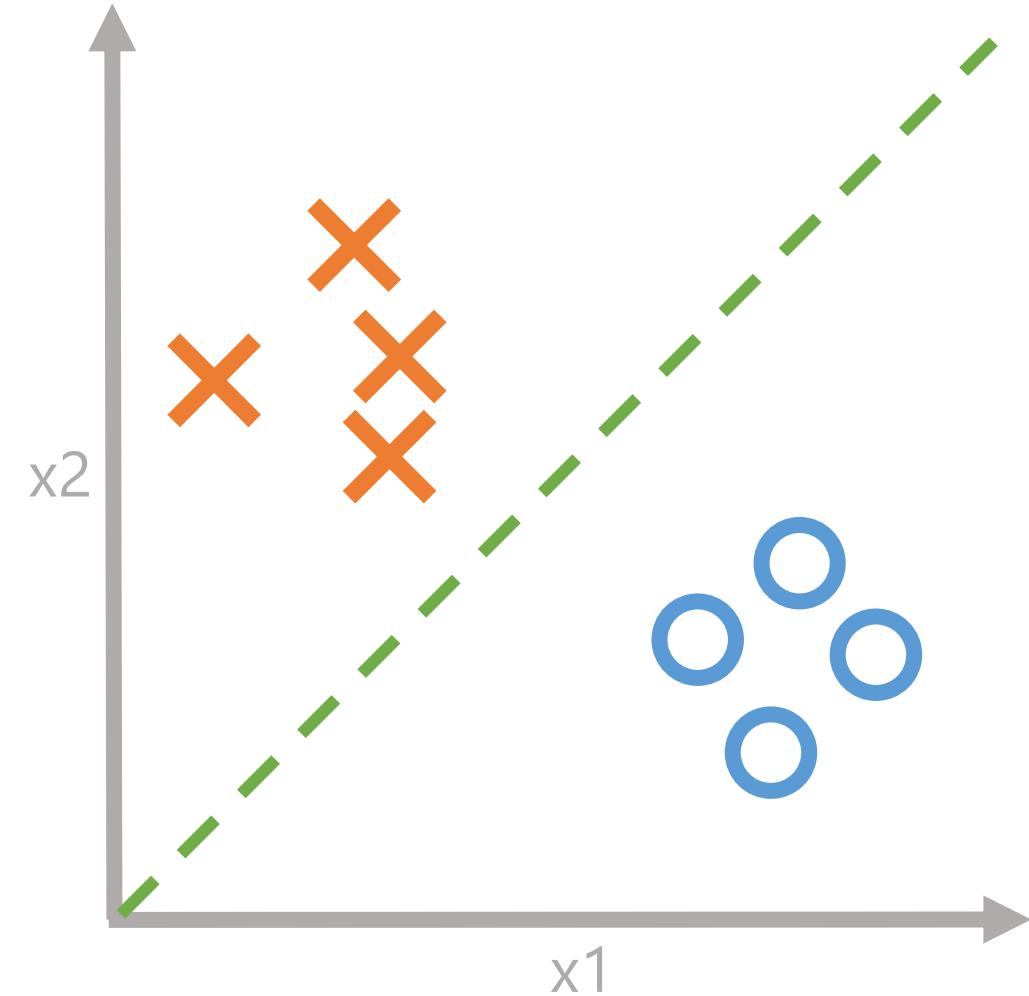




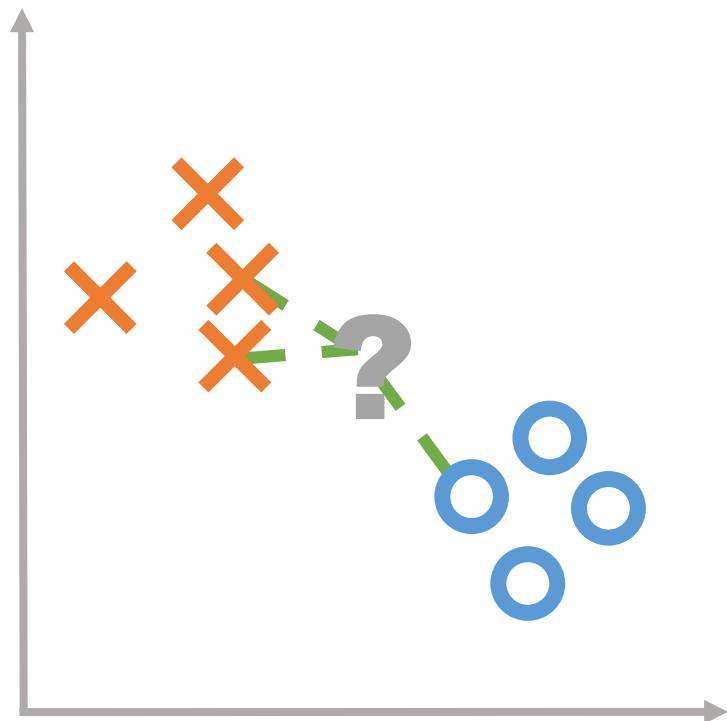


Classification Algorithms

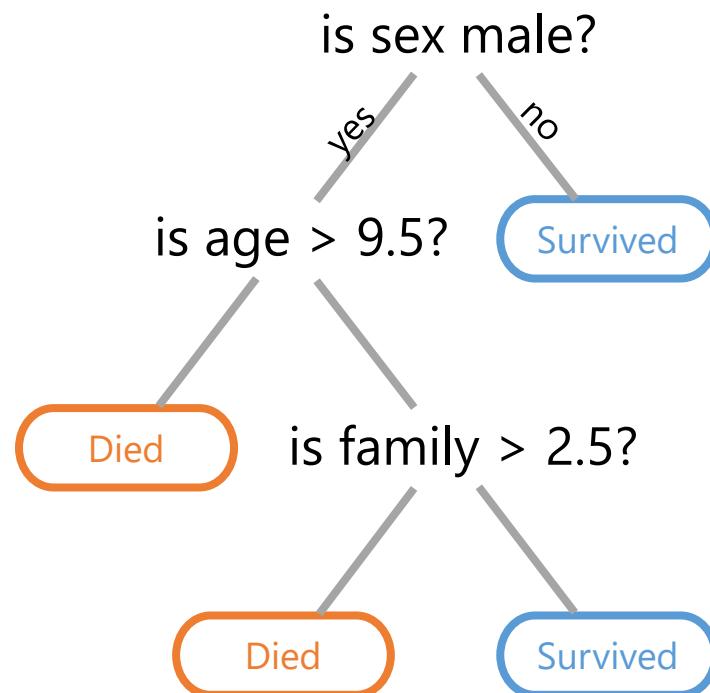
k-Nearest Neighbors
Decision Tree Classifier
Naïve Bayes Classifier
Support Vector Machine
Neural Network Classifier



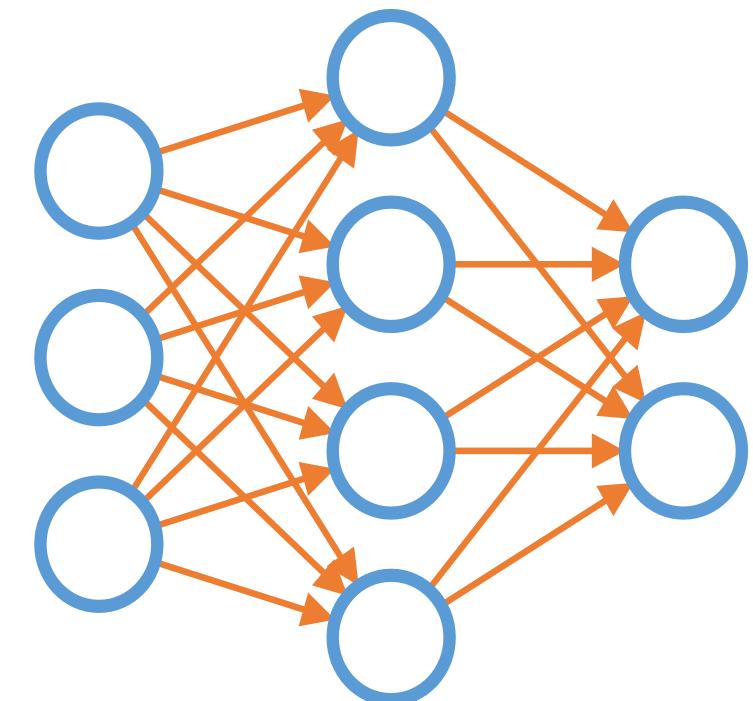
Classification Algorithms



k-Nearest Neighbors



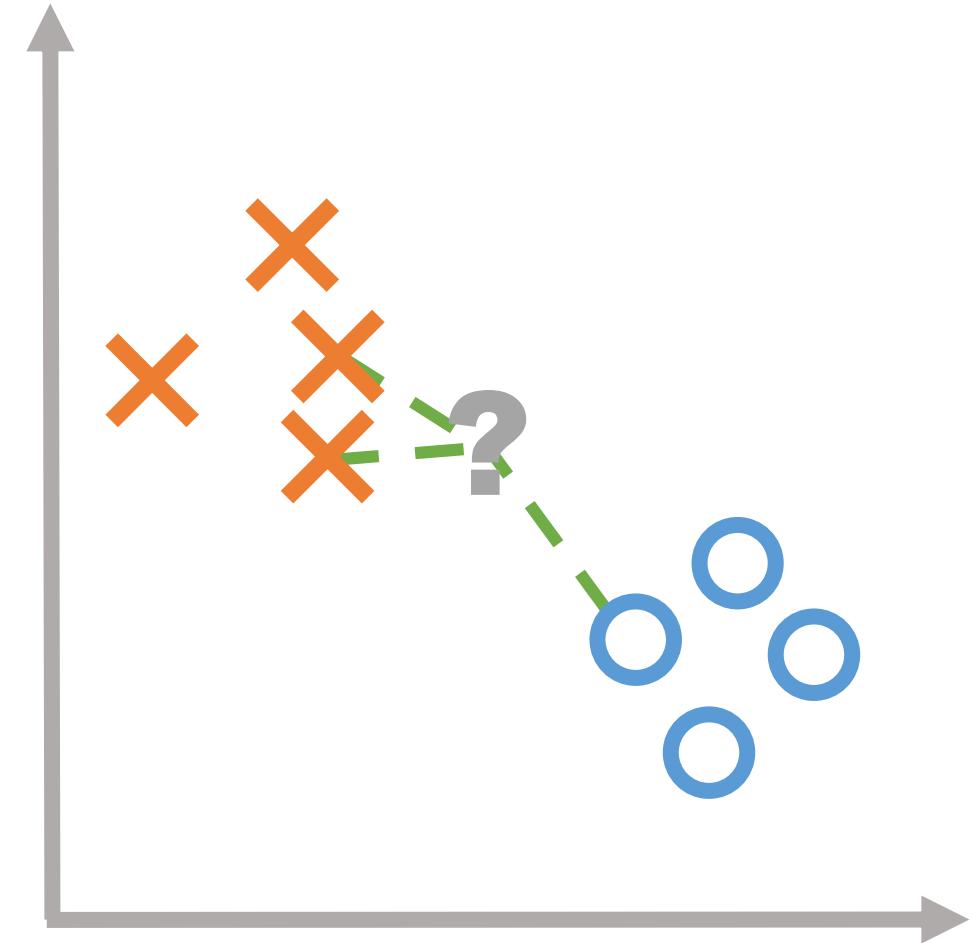
Decision Tree



Neural Network

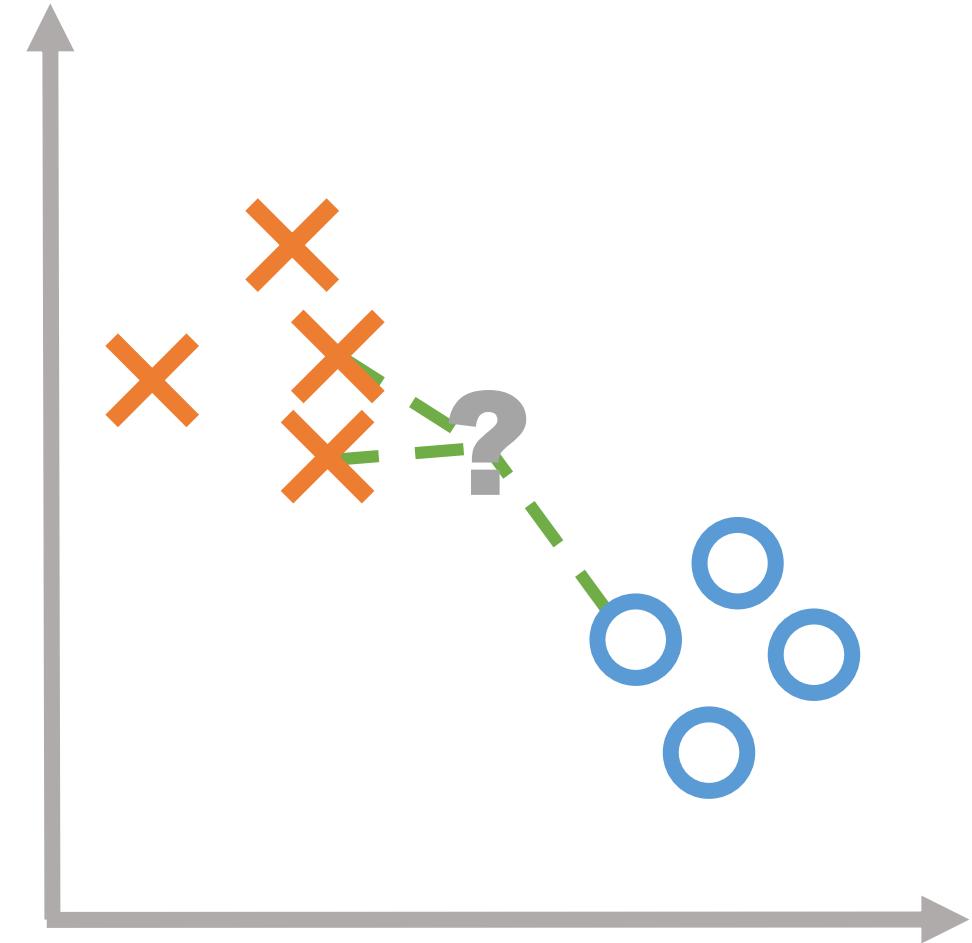
K-Nearest Neighbors Classifier

Supervised learning



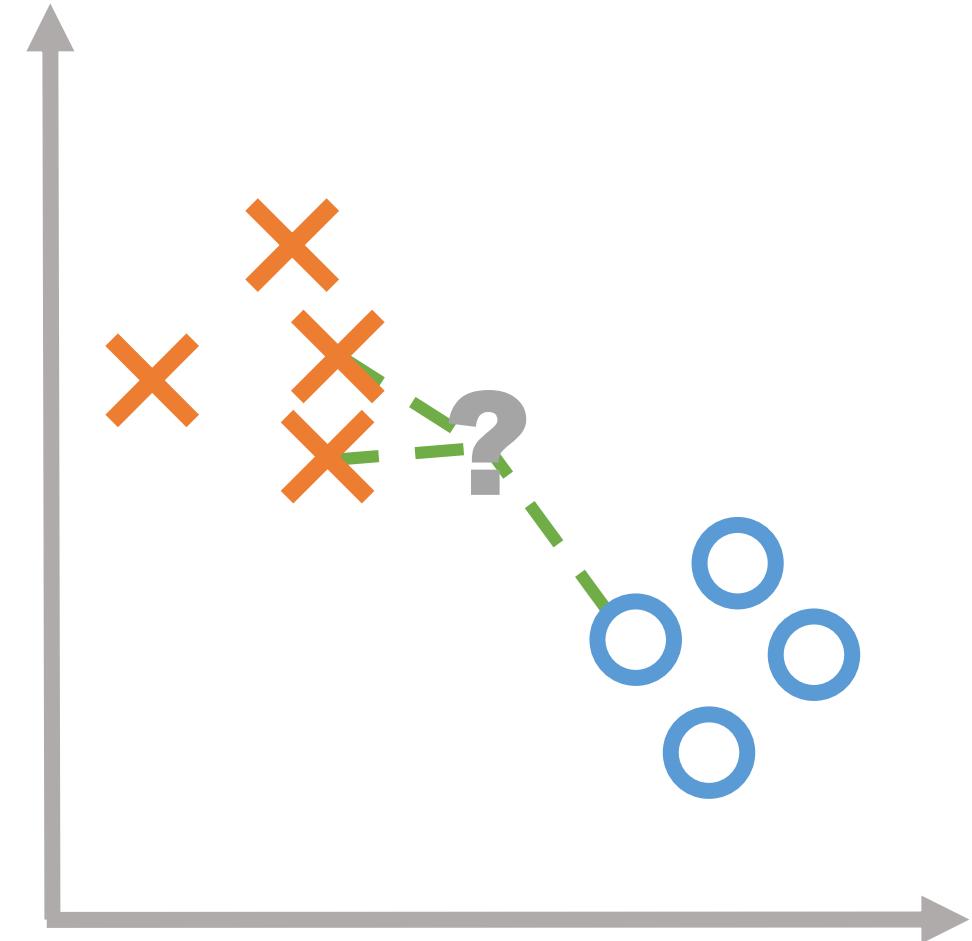
K-Nearest Neighbors Classifier

Supervised learning
Uses class of neighbors



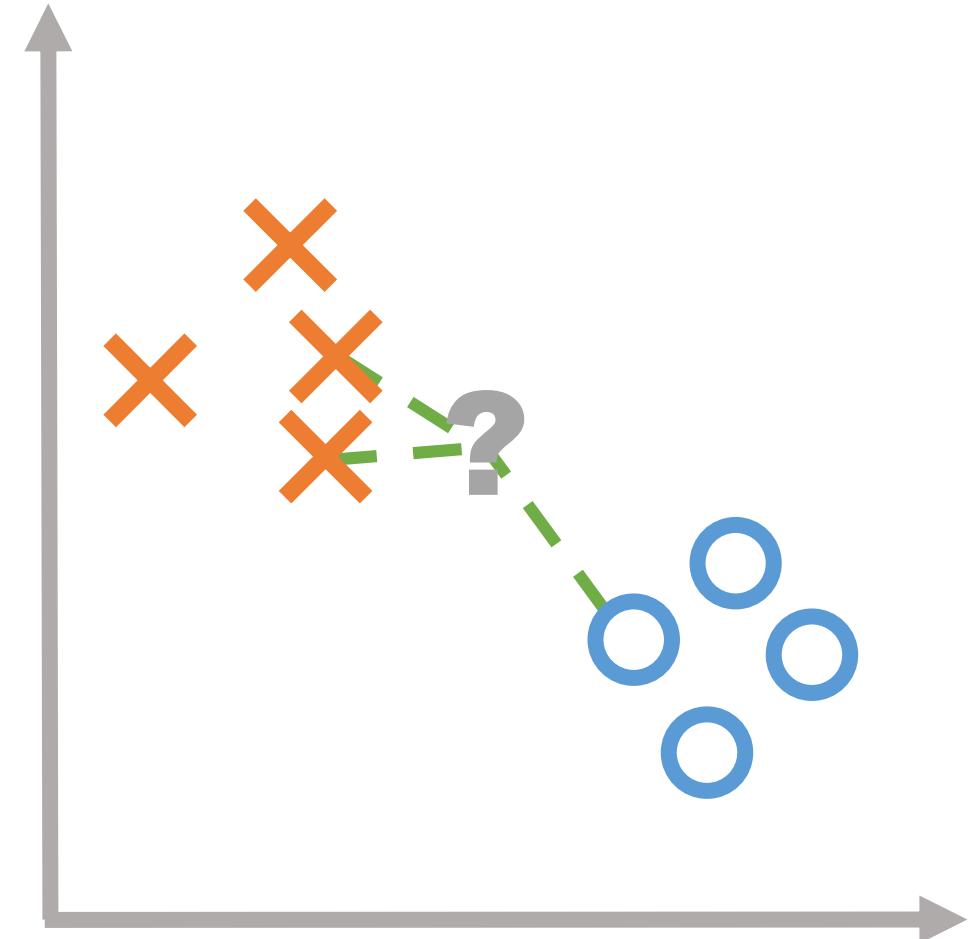
K-Nearest Neighbors Classifier

Supervised learning
Uses class of neighbors
 k specifies how many



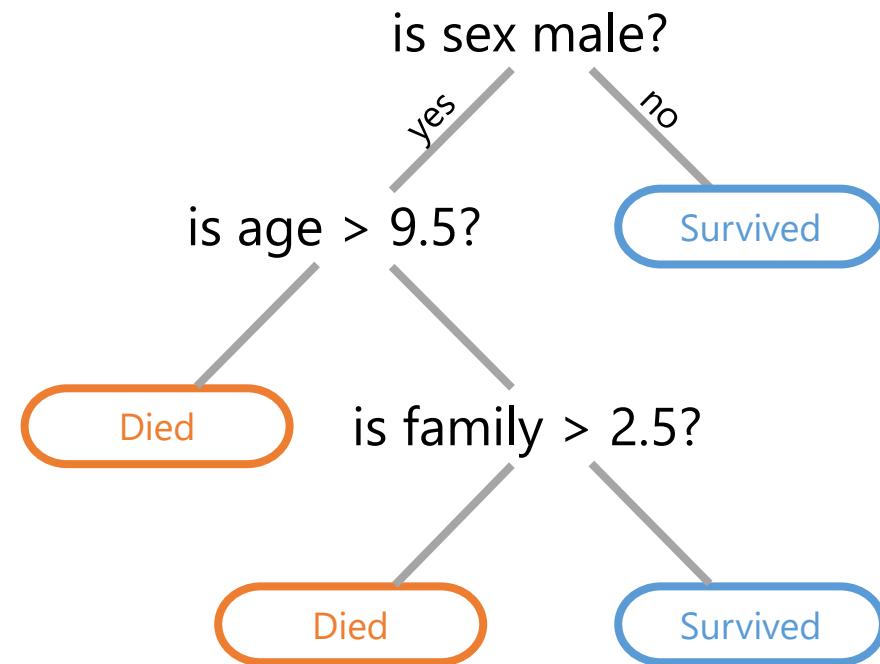
K-Nearest Neighbors Classifier

- Supervised learning
- Uses class of neighbors
- k specifies how many
- Simple and easy



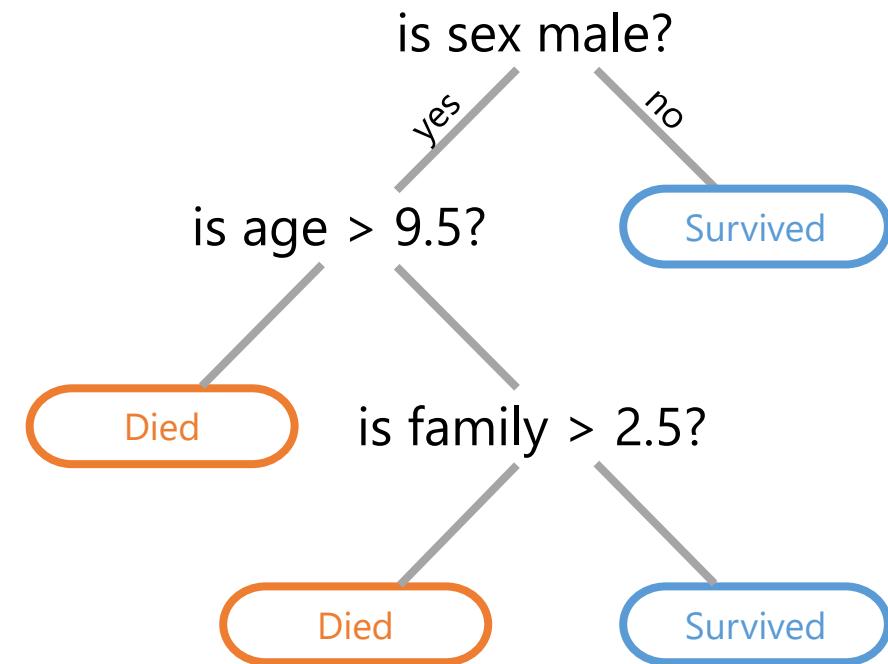
Decision Tree Classifier

Supervised learning



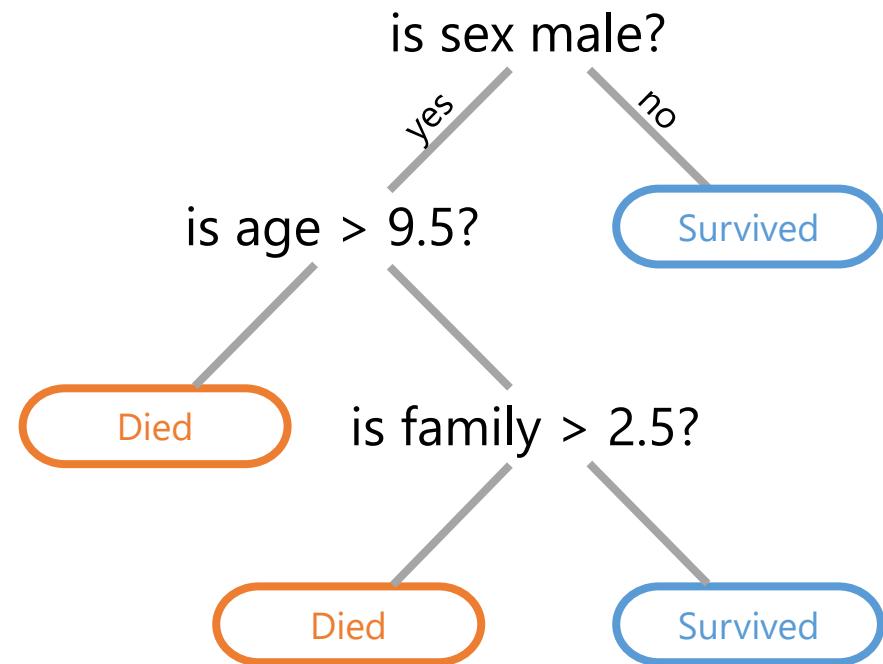
Decision Tree Classifier

Supervised learning
Tree of decisions



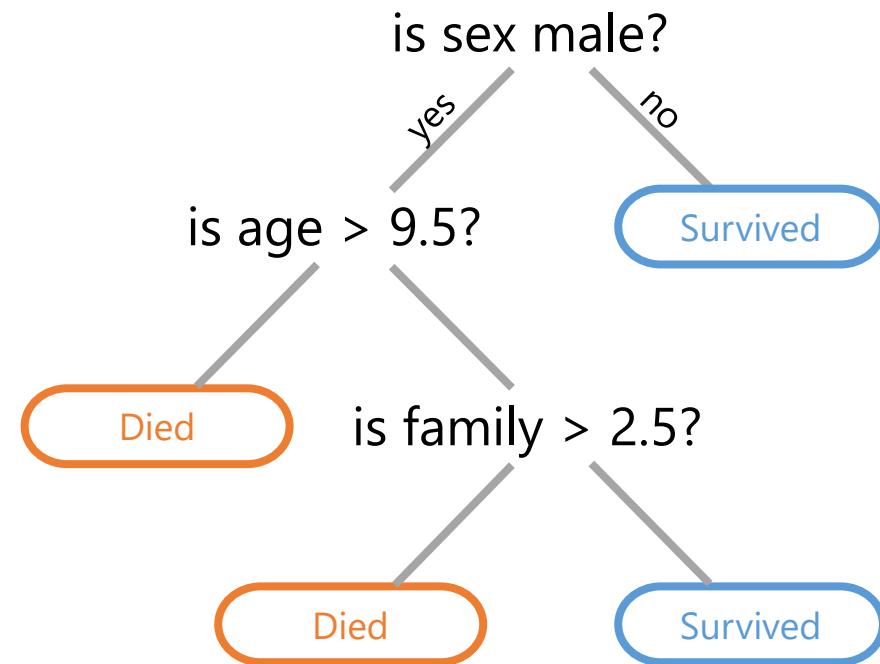
Decision Tree Classifier

Supervised learning
Tree of decisions
Information gain



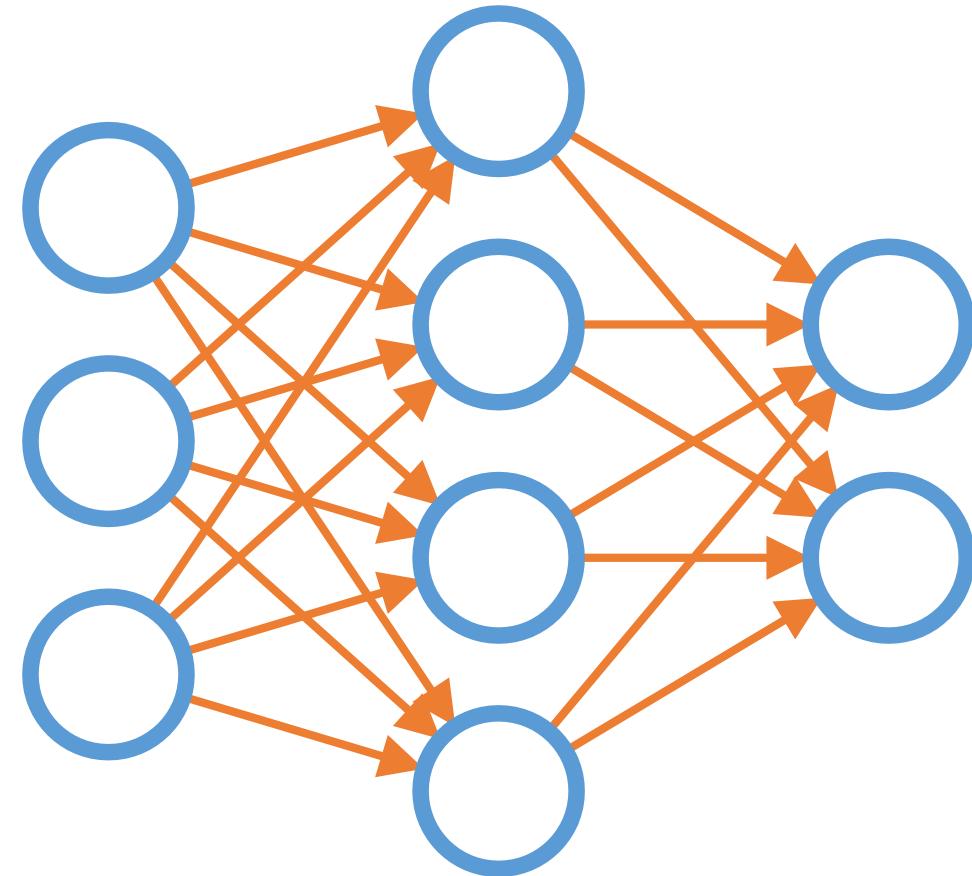
Decision Tree Classifier

Supervised learning
Tree of decisions
Information gain
Simple and easy



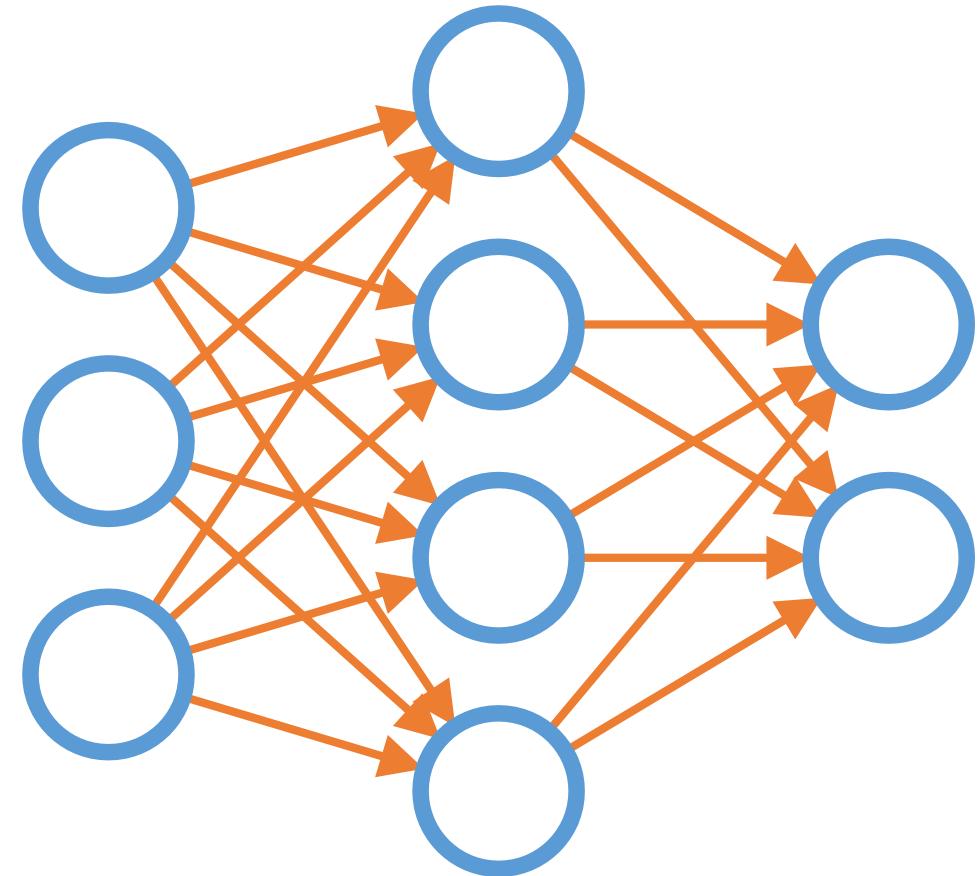
Neural Network Classifier

Supervised learning



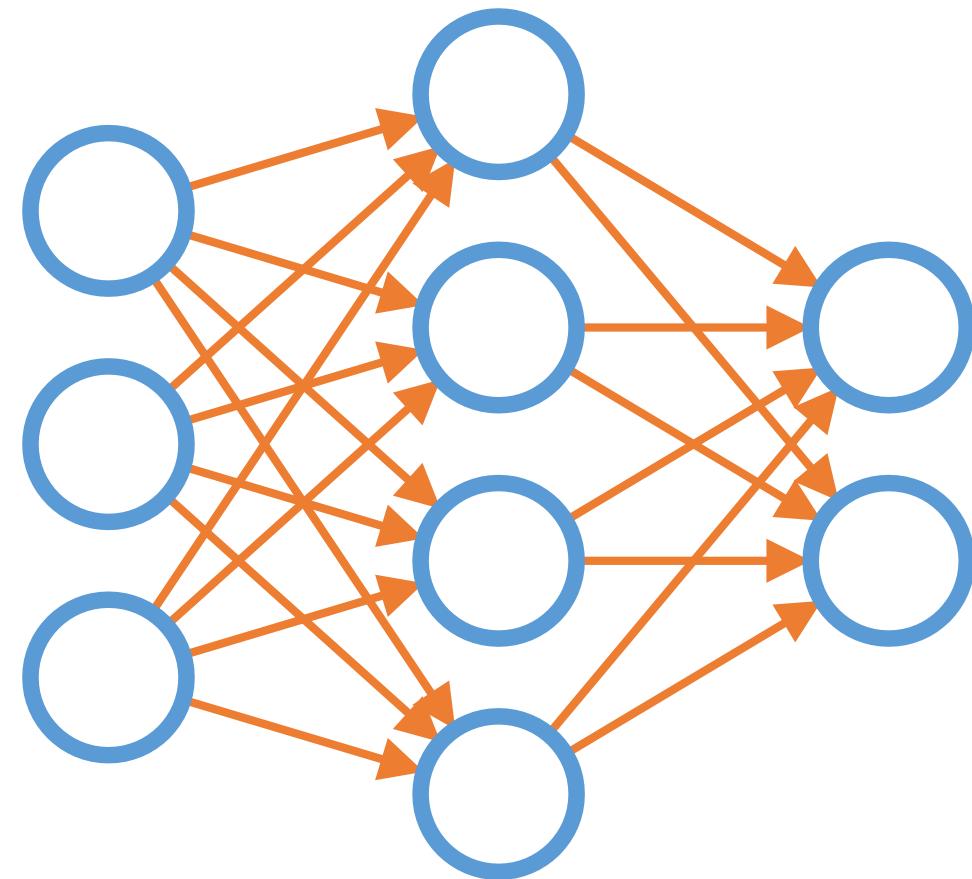
Neural Network Classifier

Supervised learning
Neurons in a brain



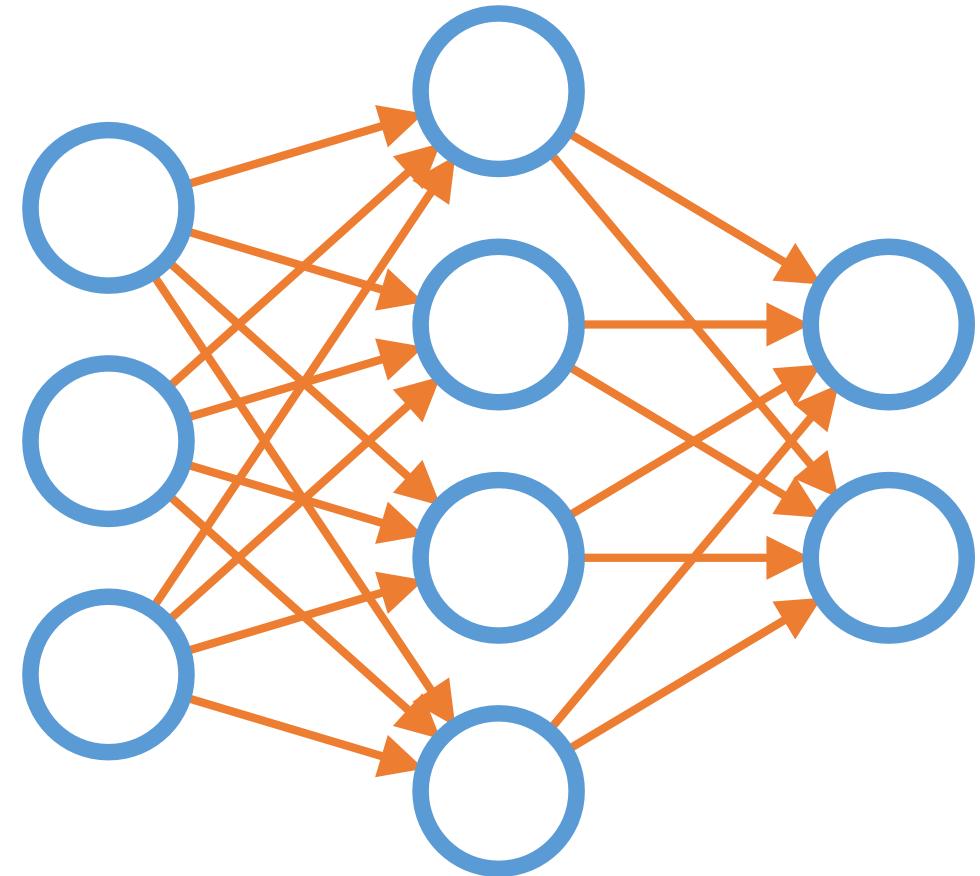
Neural Network Classifier

Supervised learning
Neurons in a brain
Weighted connections



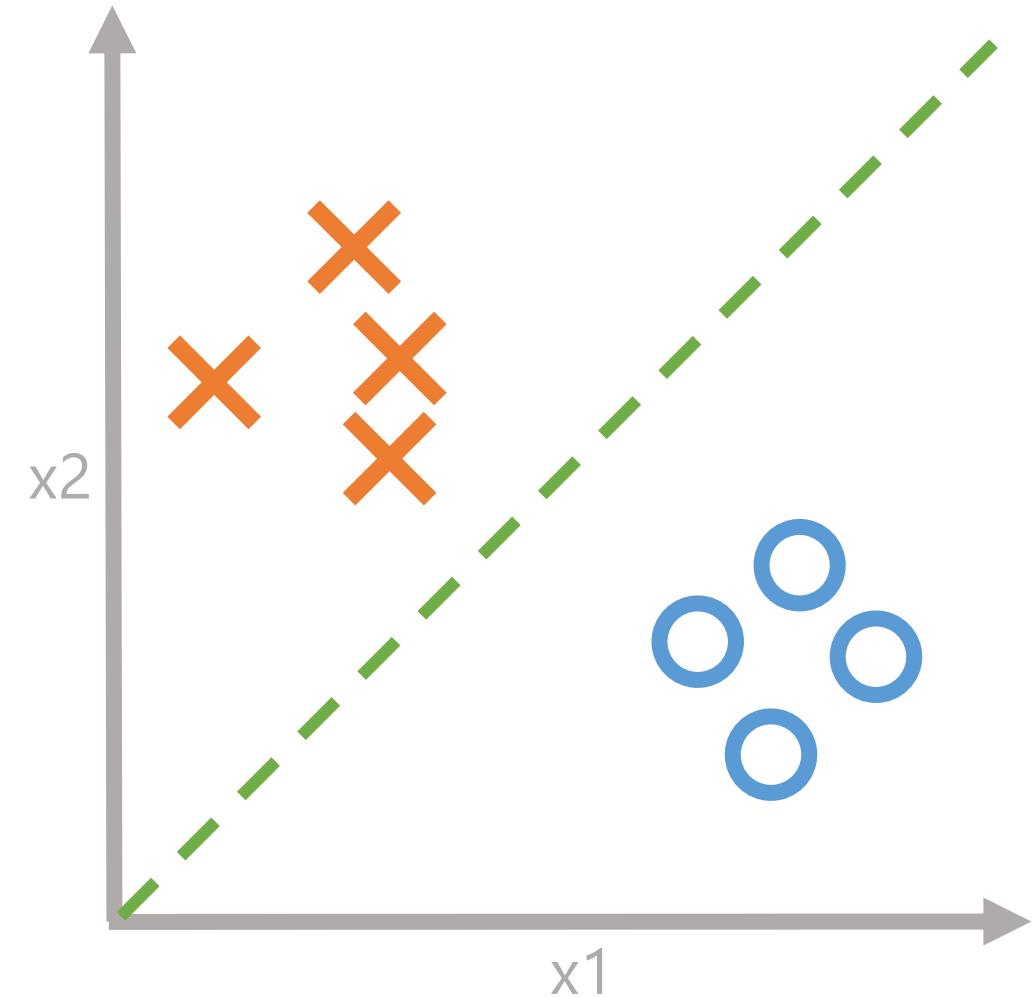
Neural Network Classifier

Supervised learning
Neurons in a brain
Weighted connections
Complex



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

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 Data Set

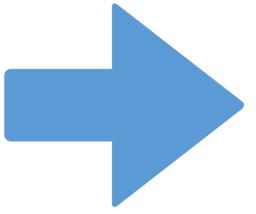
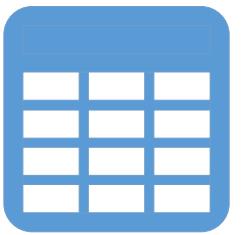
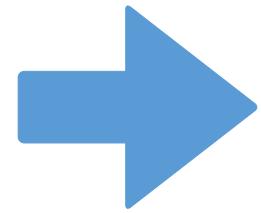
Lab 2A – Classification (Easy)

Goal: Predict species based on
petal and sepal measurements

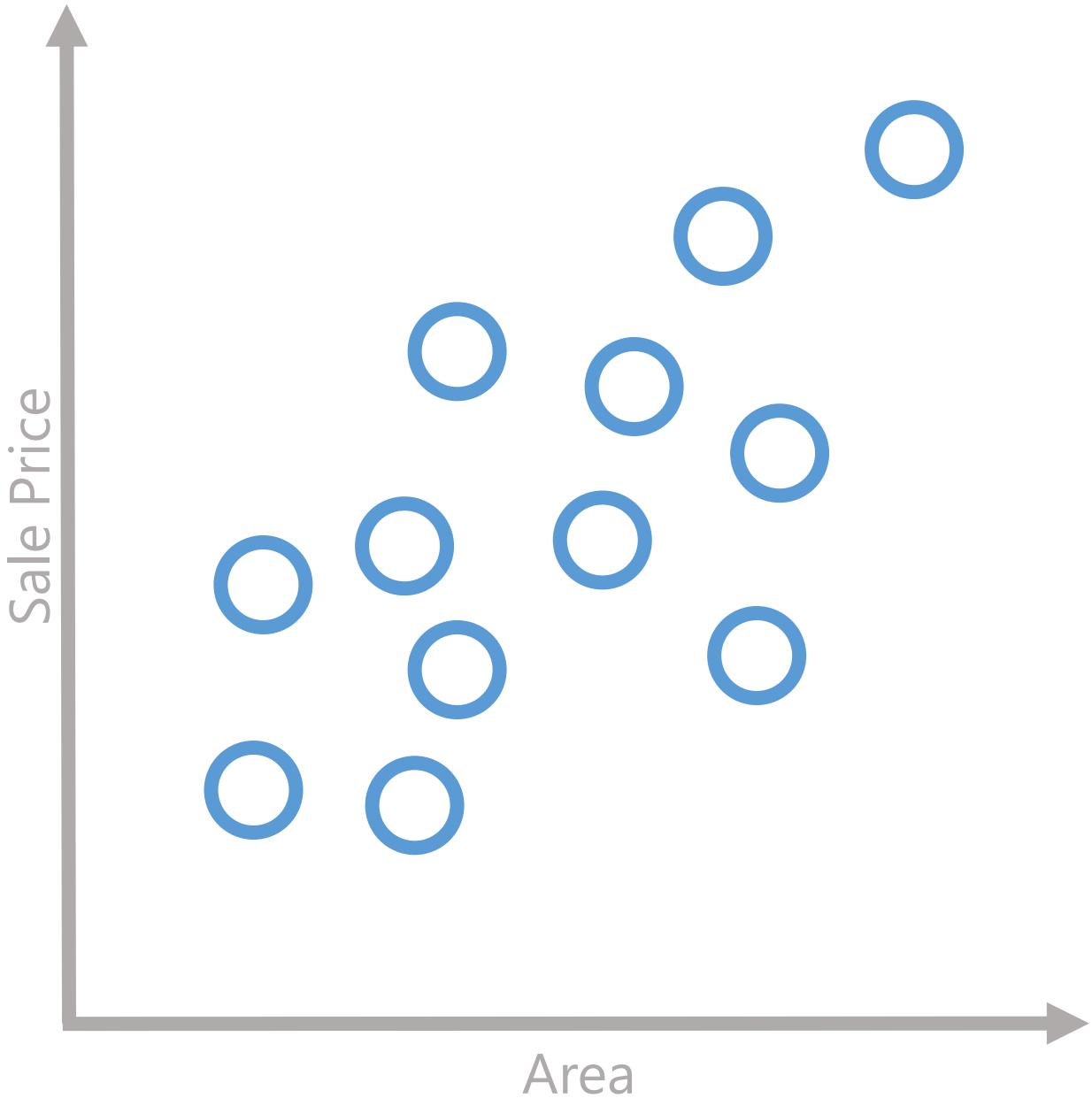
Lab 2B – Classification (Hard)

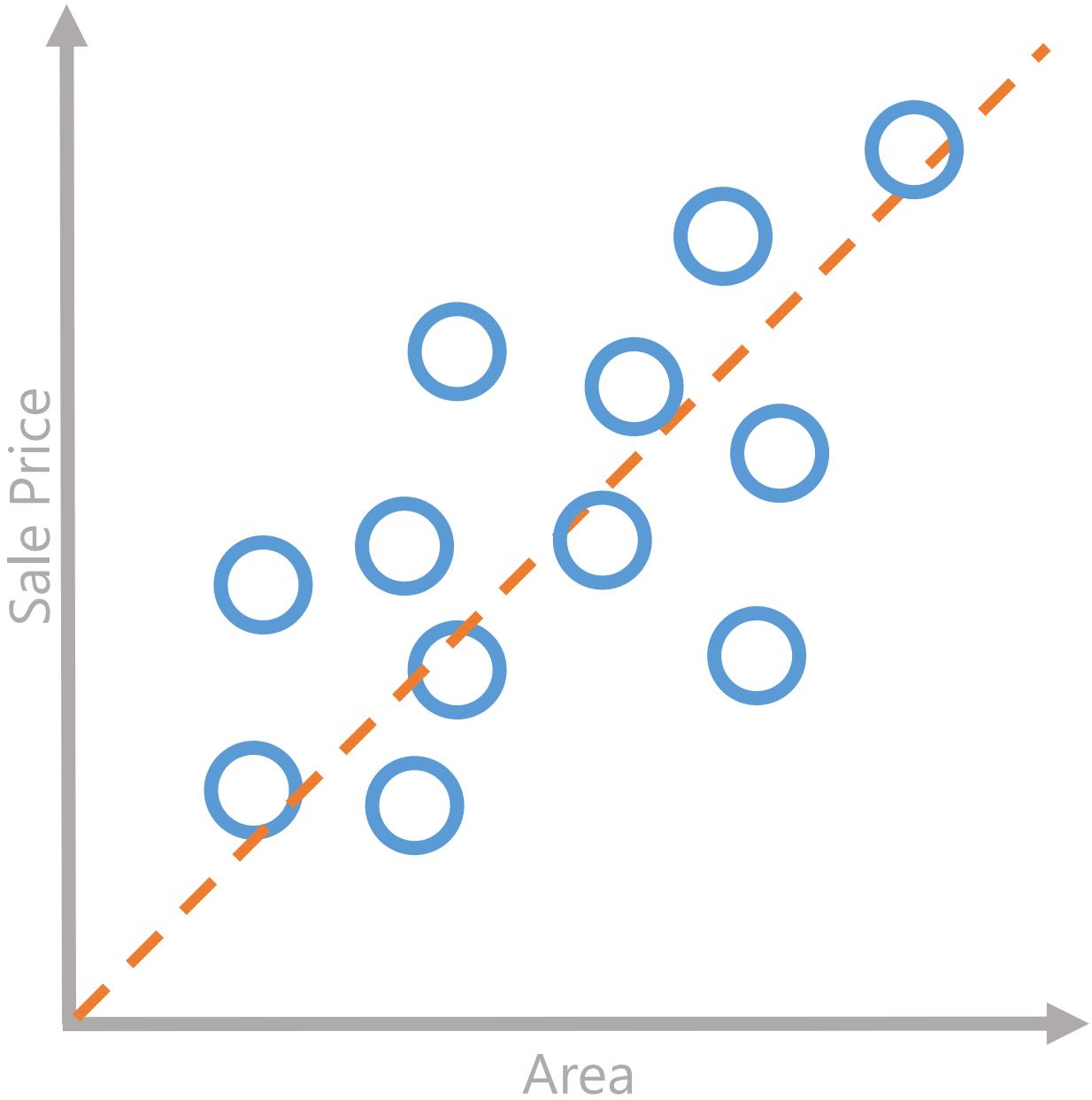
Goal: Predict the risk of
an insurance policy

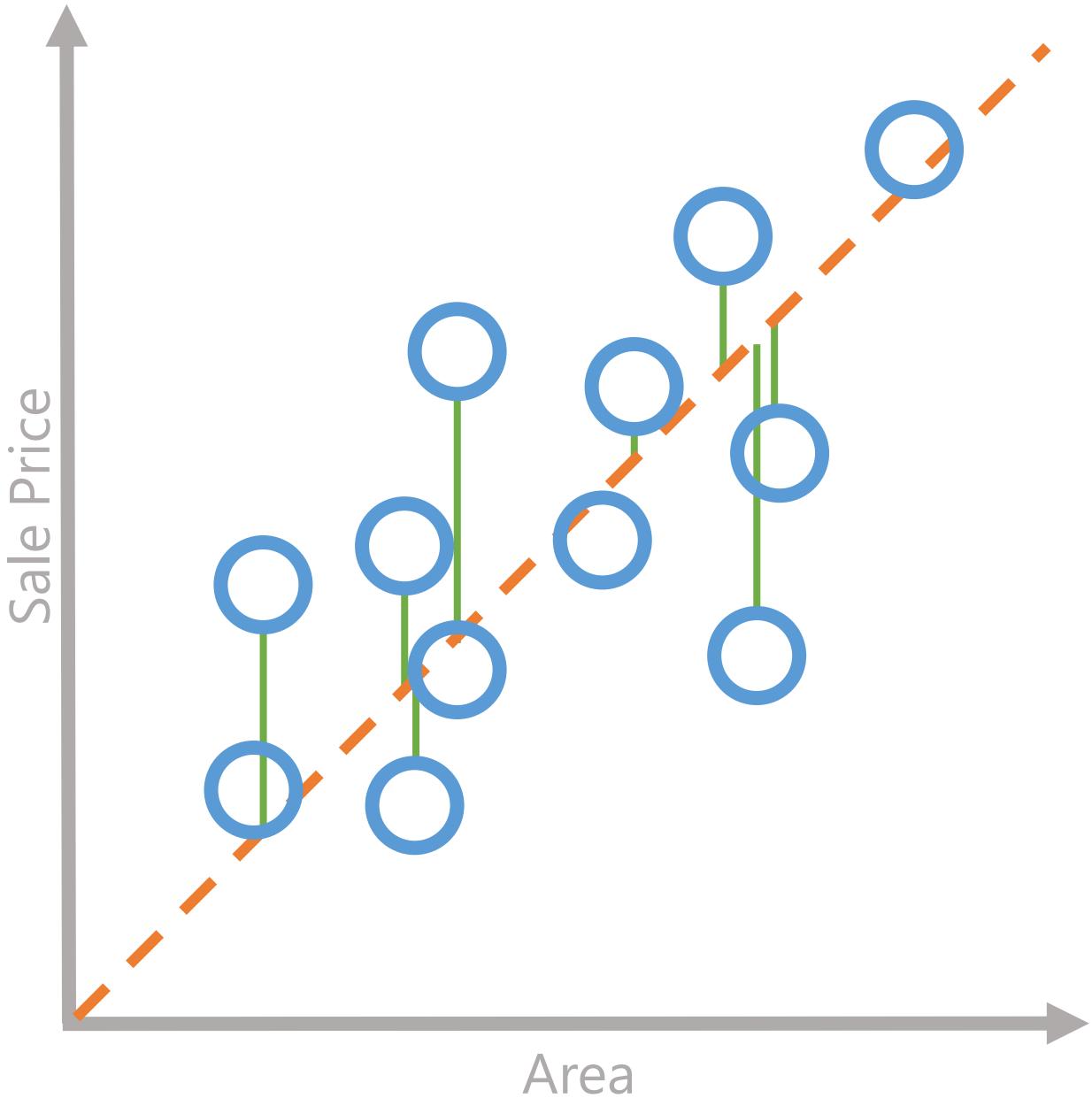
Regression

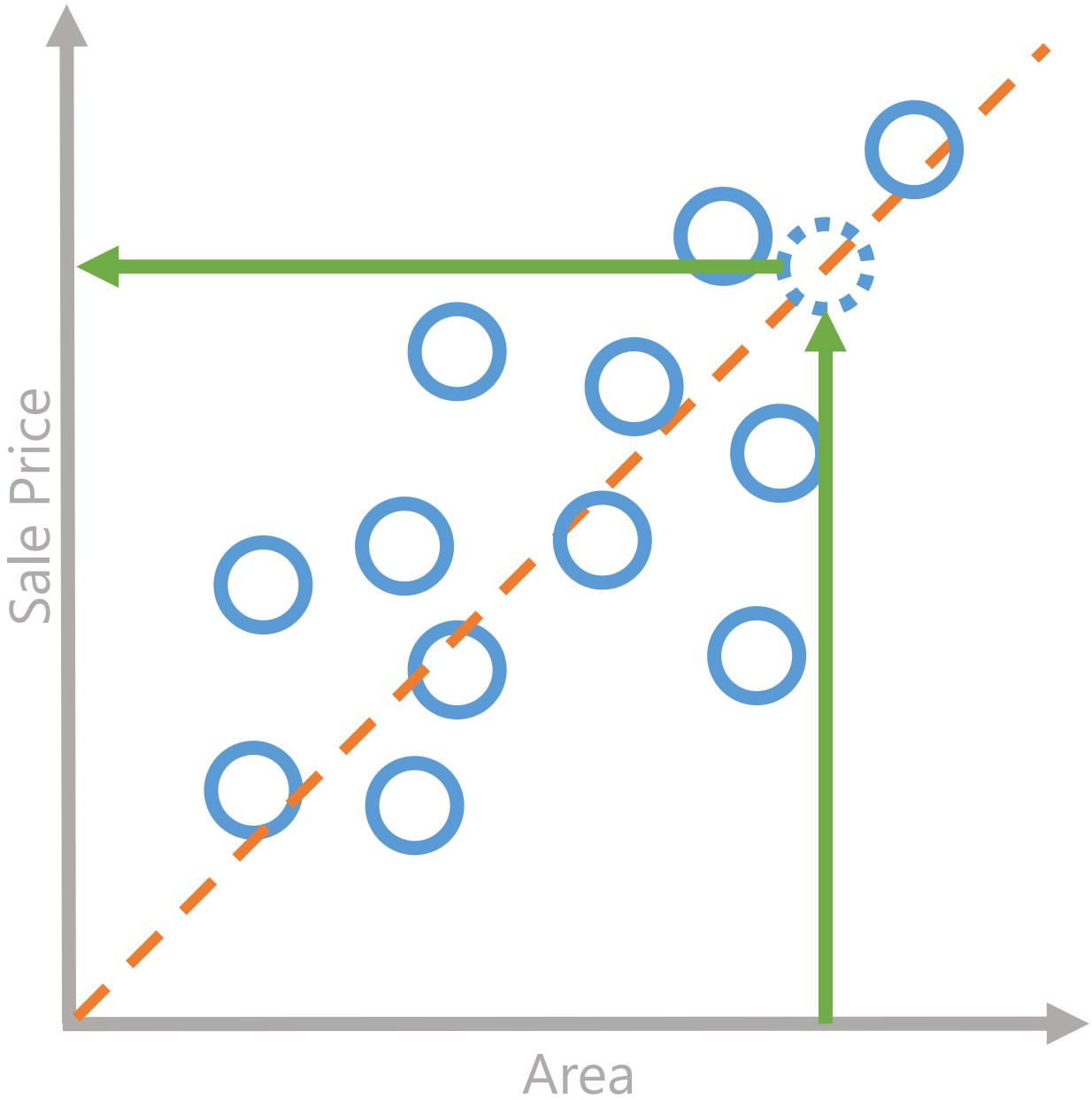
 $f(x)$ 

1.23









Regression Algorithms

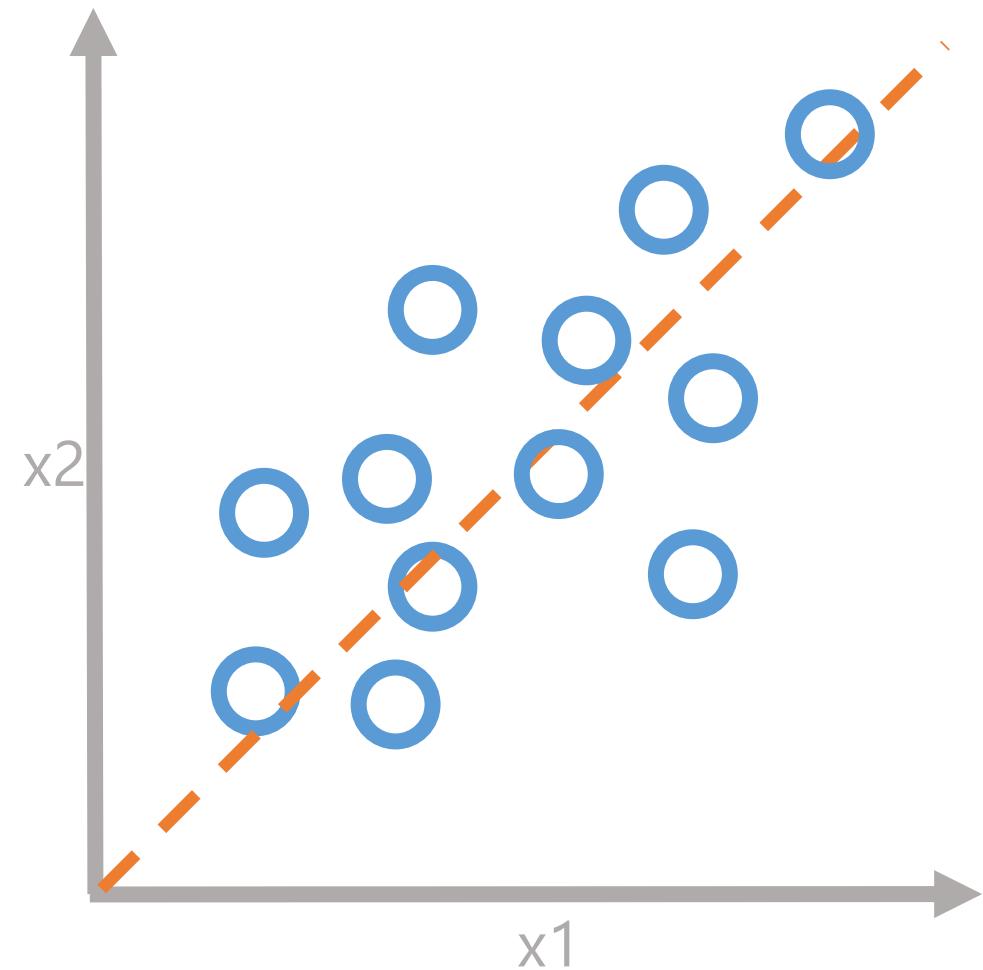
Linear Regression

Polynomial Regression

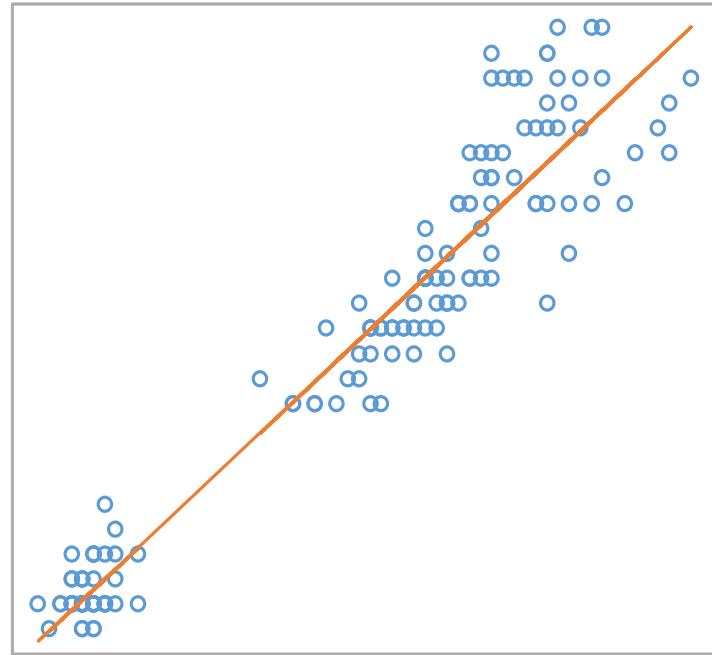
Lasso Regression

ElasticNet Regression

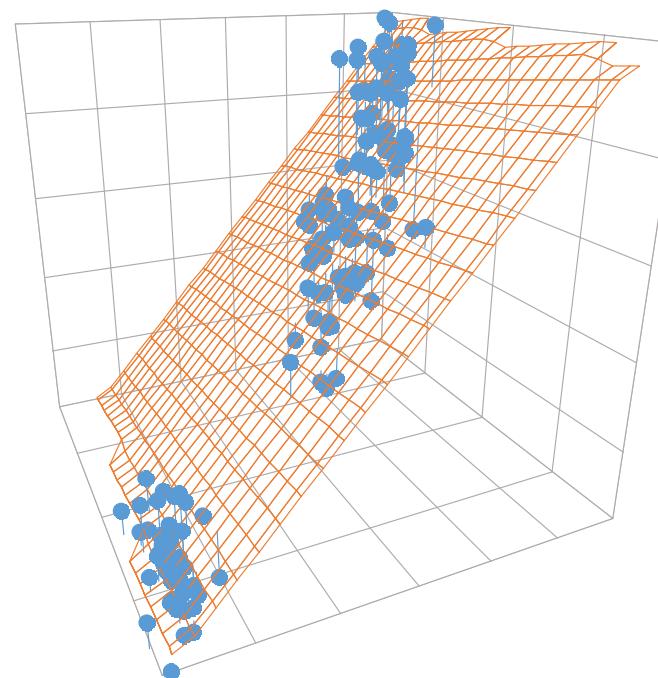
Neural Network Regression



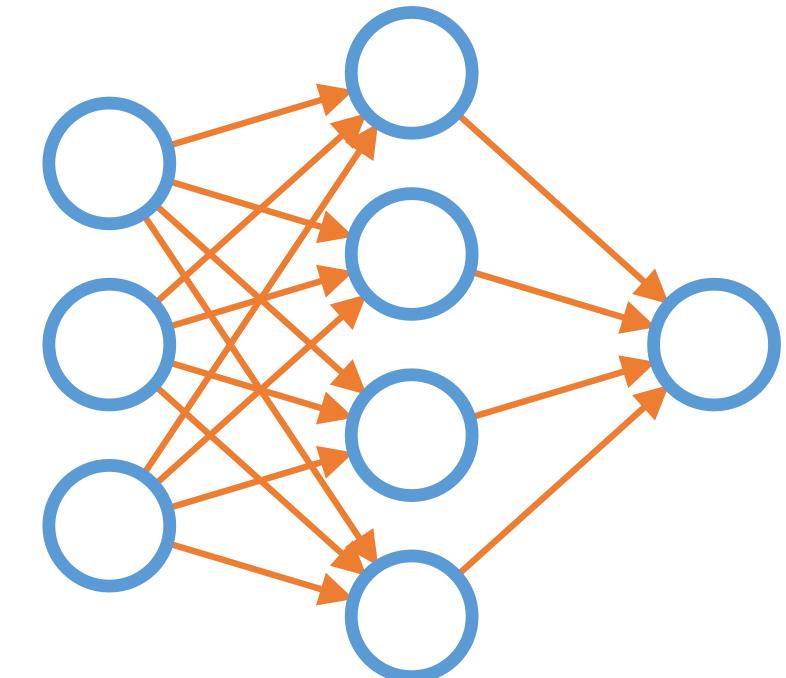
Regression Algorithms



Simple Linear



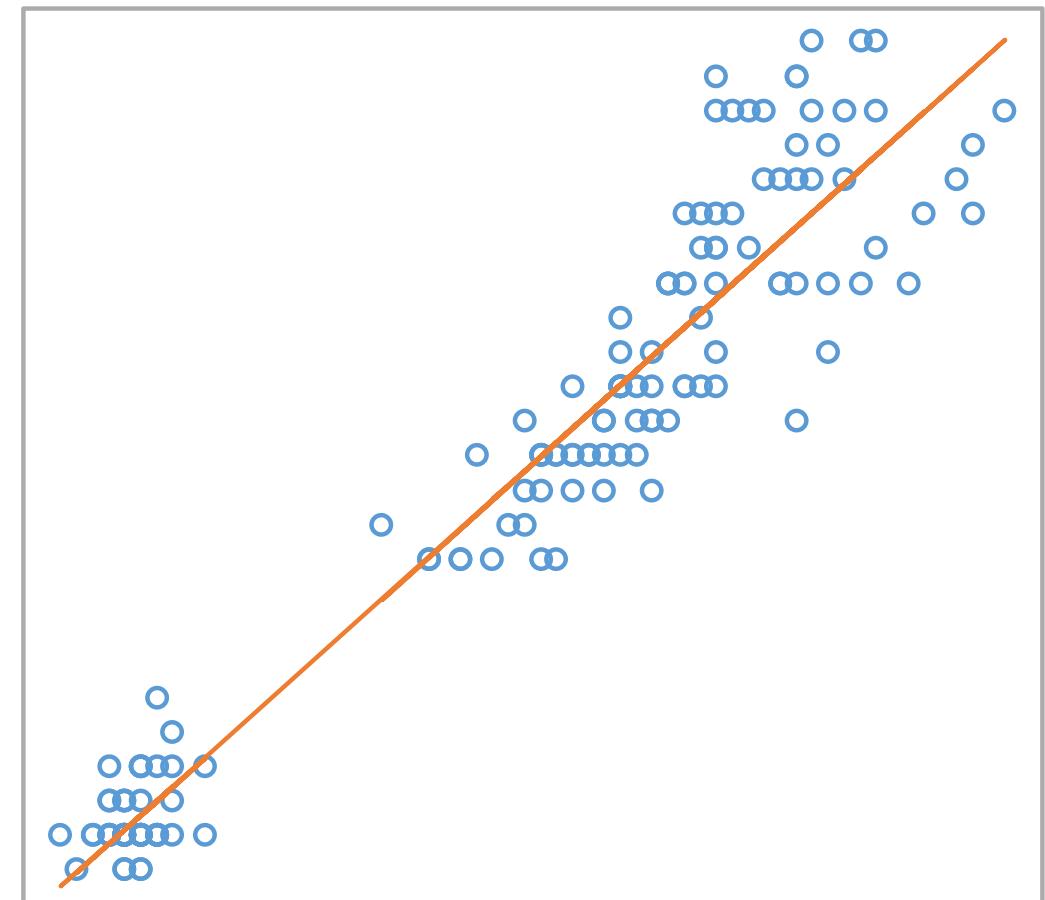
Multiple Linear



Neural Network

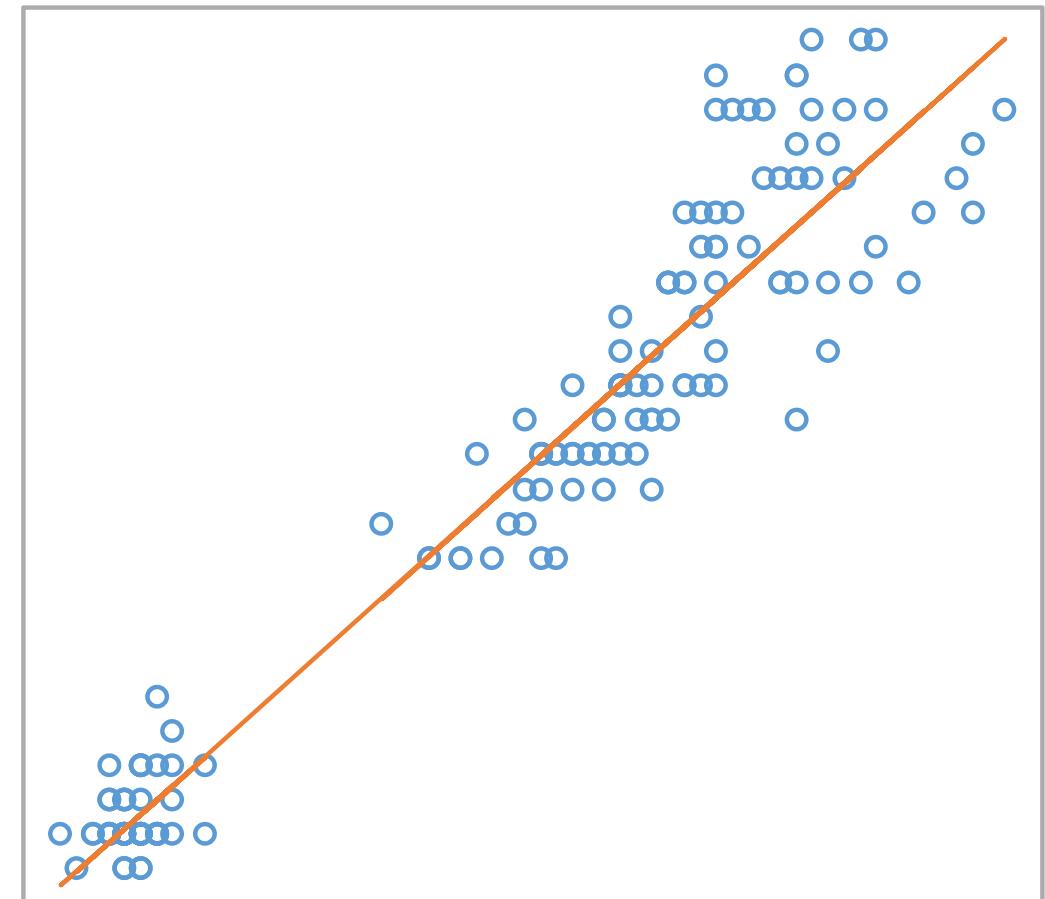
Simple Linear Regression

Relationship



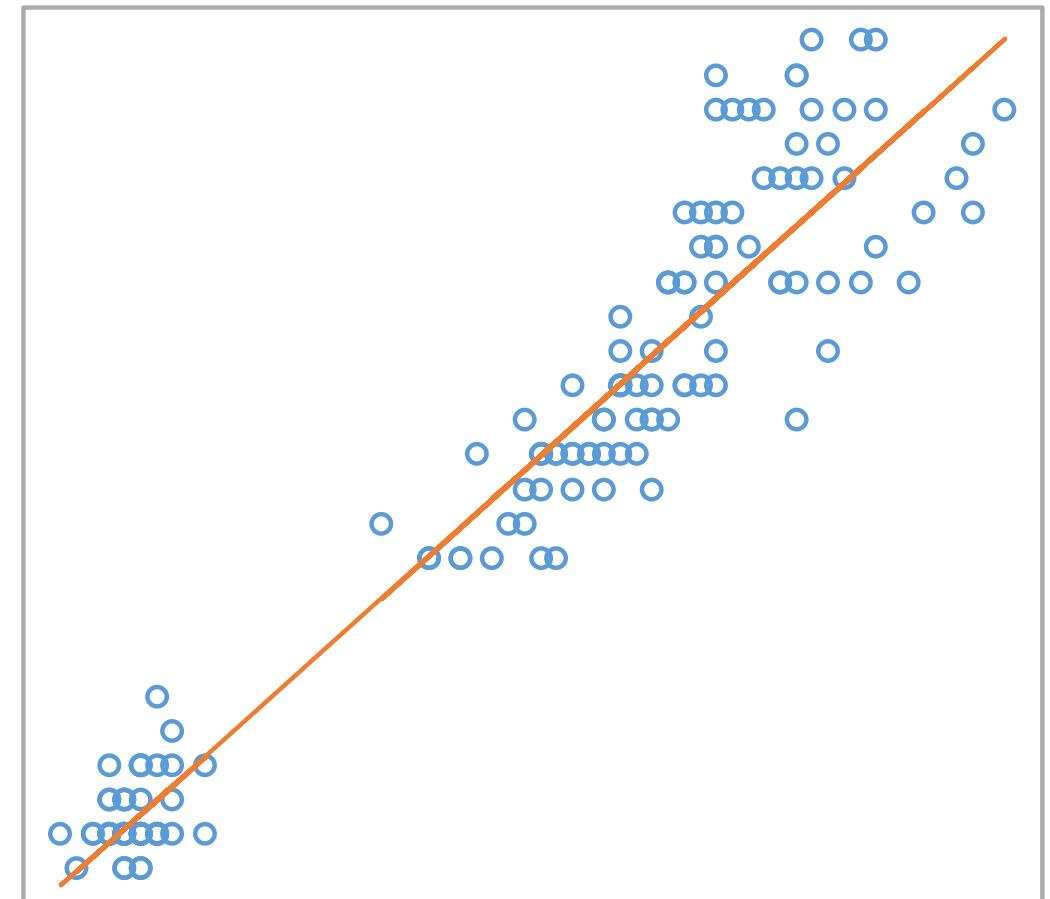
Simple Linear Regression

Relationship
Linear model



Simple Linear Regression

Relationship
Linear model
 $y = m \cdot x + b$



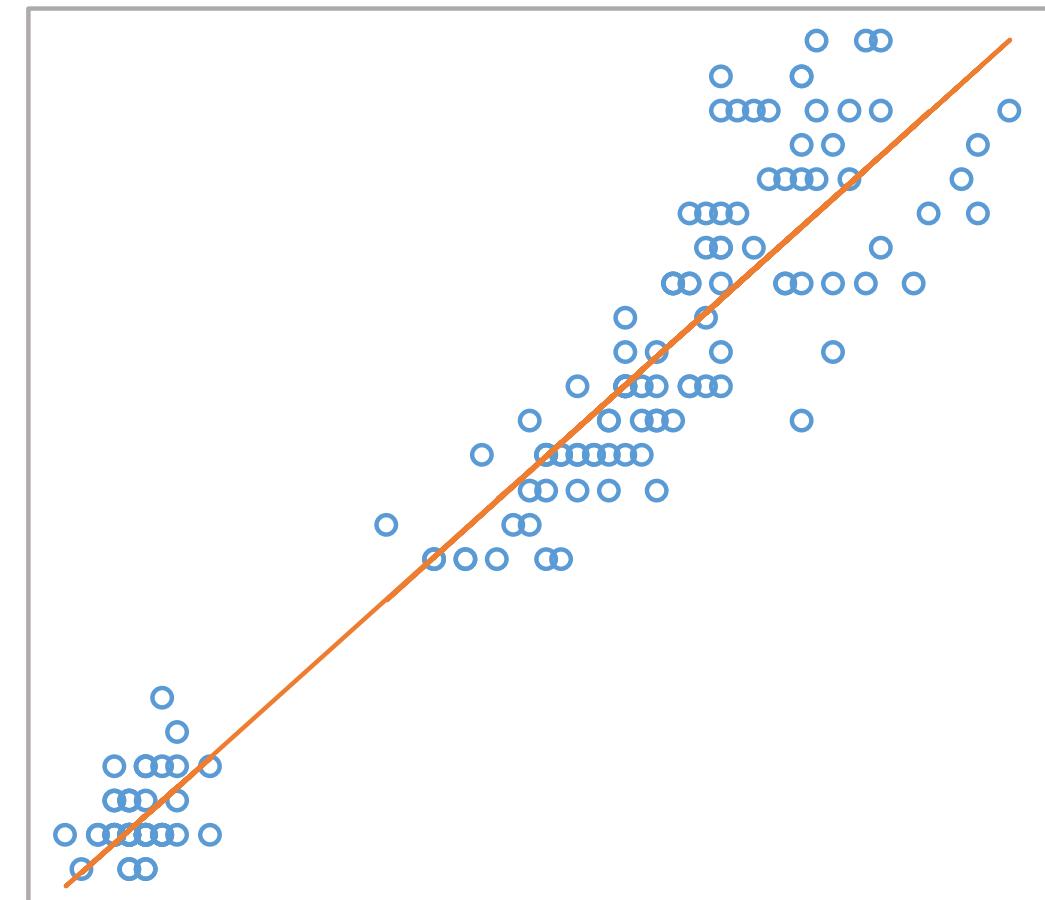
Simple Linear Regression

Relationship

Linear model

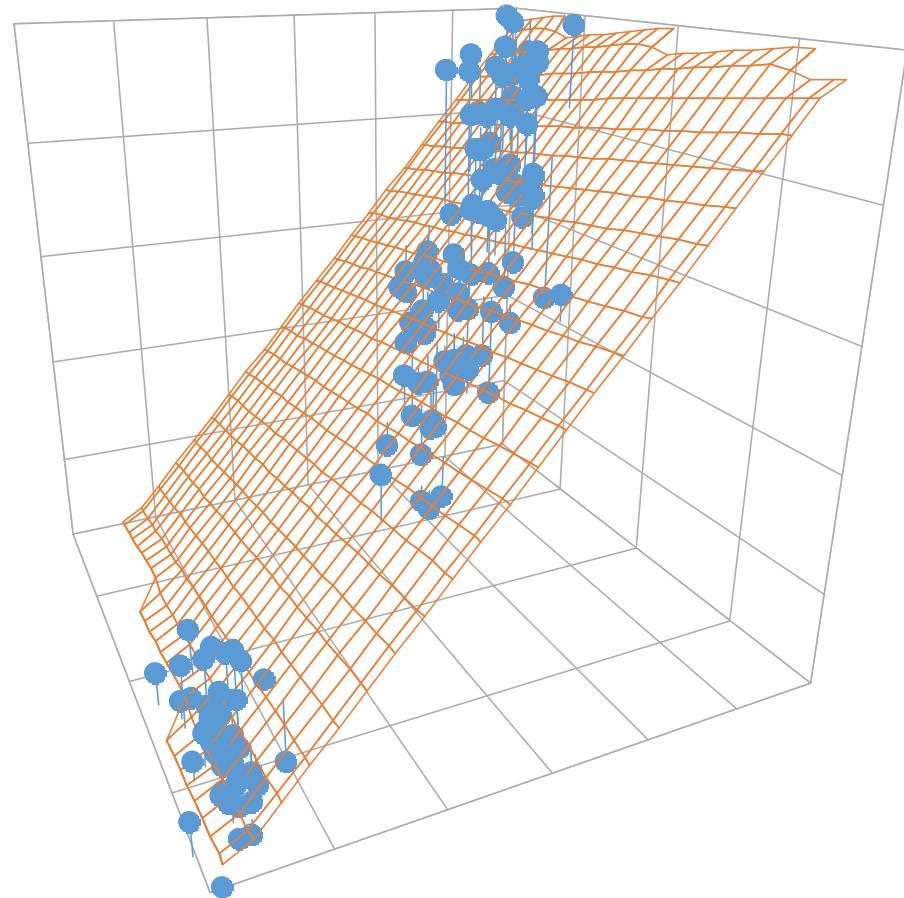
$$y = m \cdot x + b$$

Parameters estimated



Multiple Linear Regression

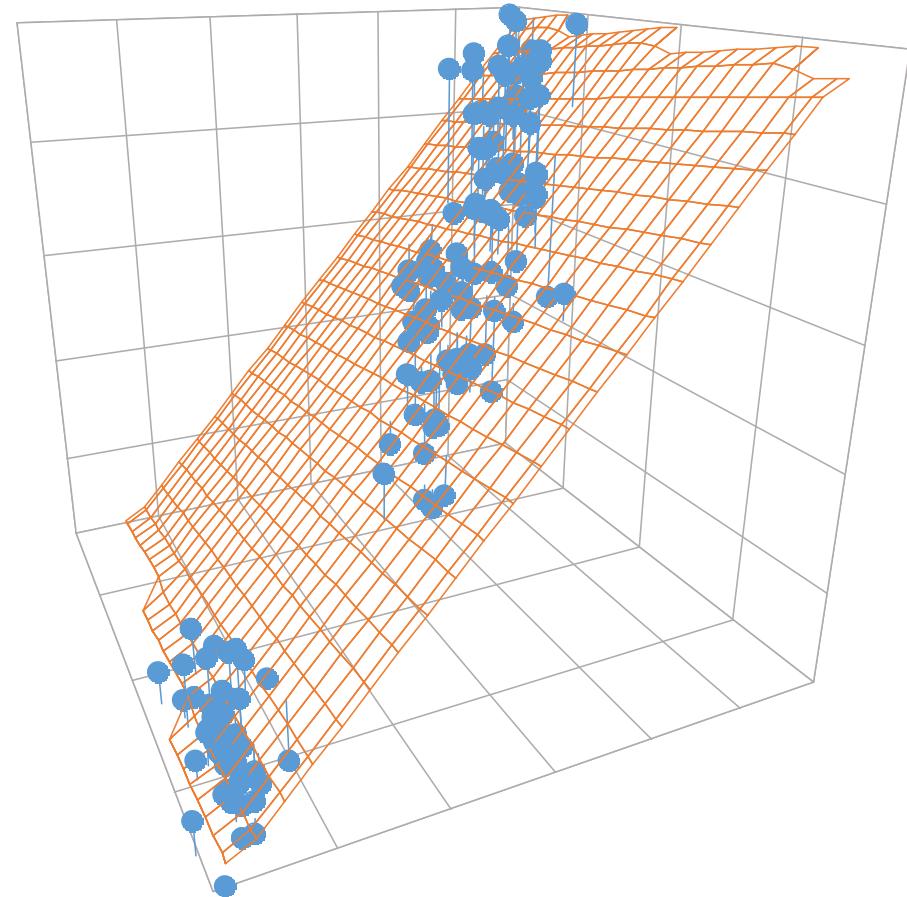
Similar to SLR



Multiple Linear Regression

Similar to SLR

Multiple variables

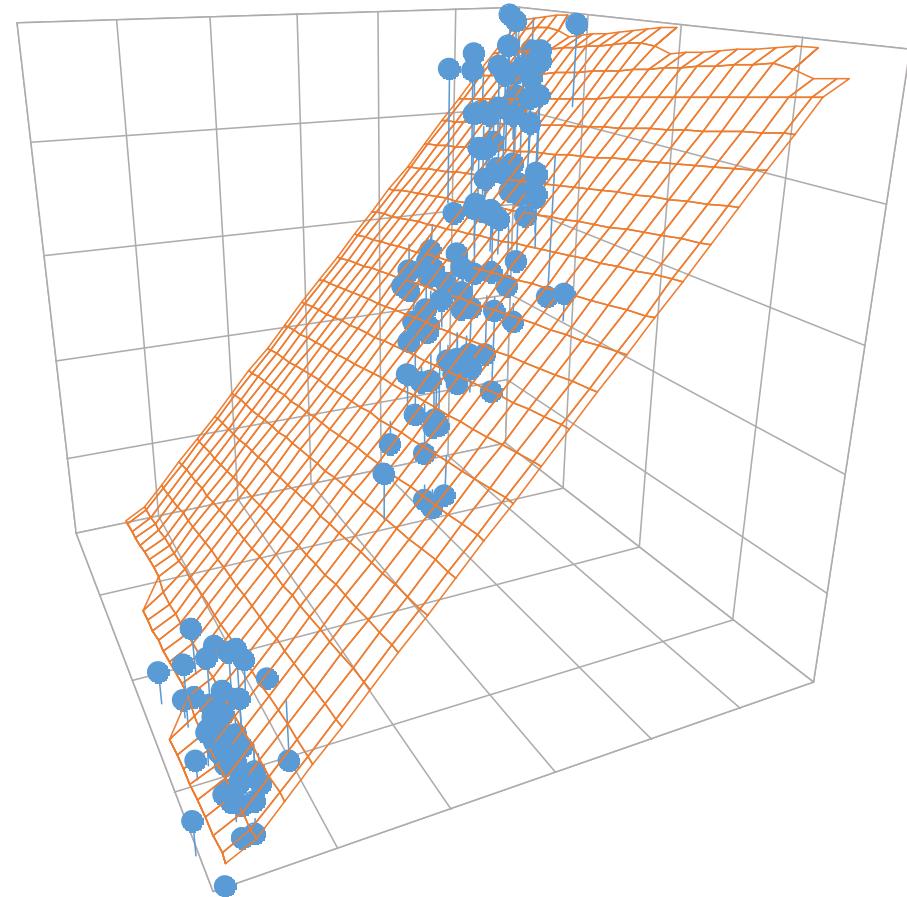


Multiple Linear Regression

Similar to SLR

Multiple variables

Multiple slopes



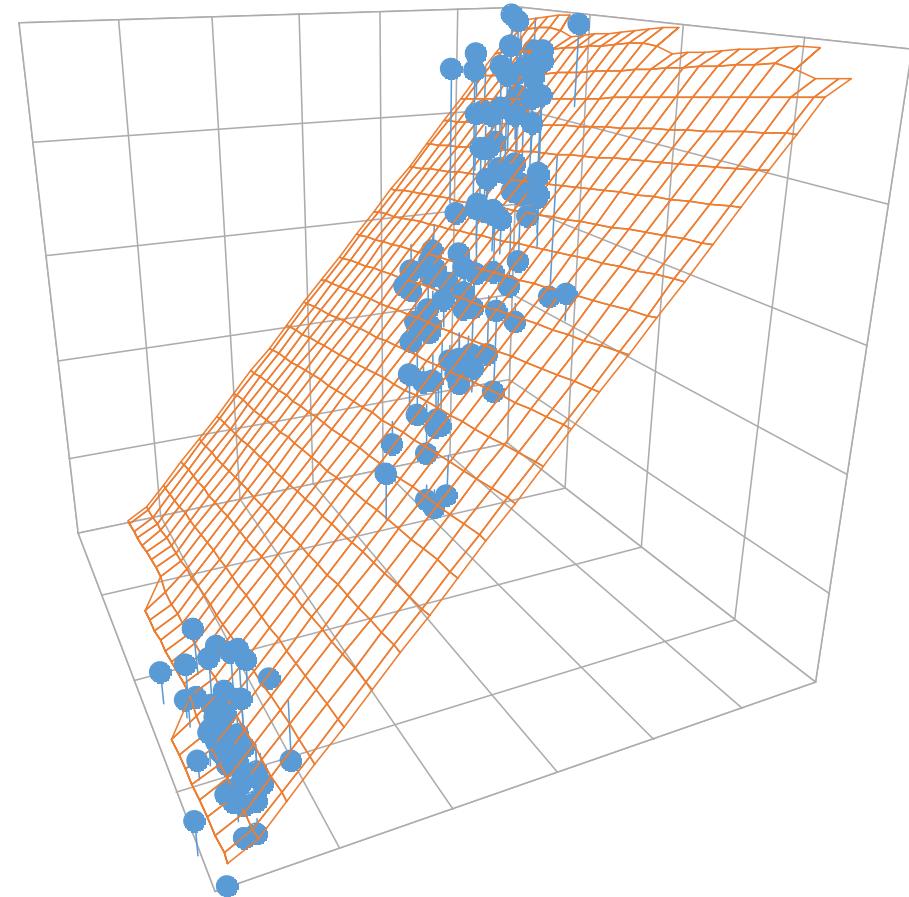
Multiple Linear Regression

Similar to SLR

Multiple variables

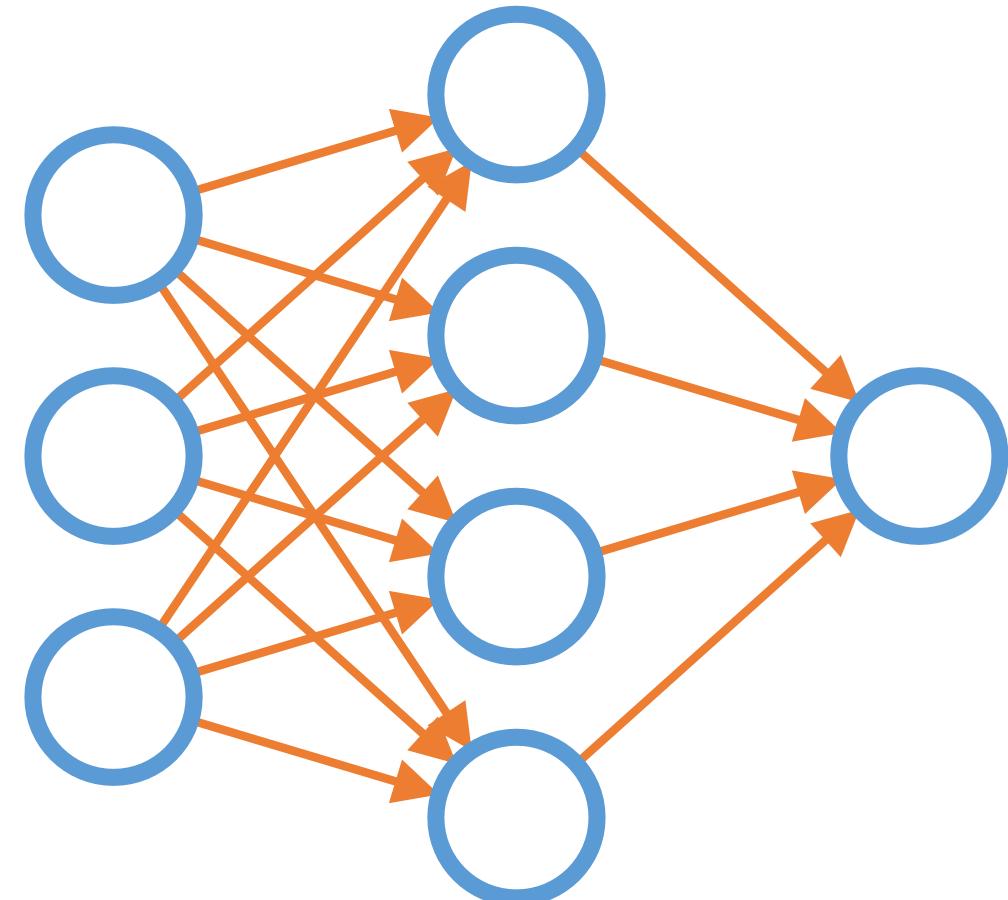
Multiple slopes

Categorical variables



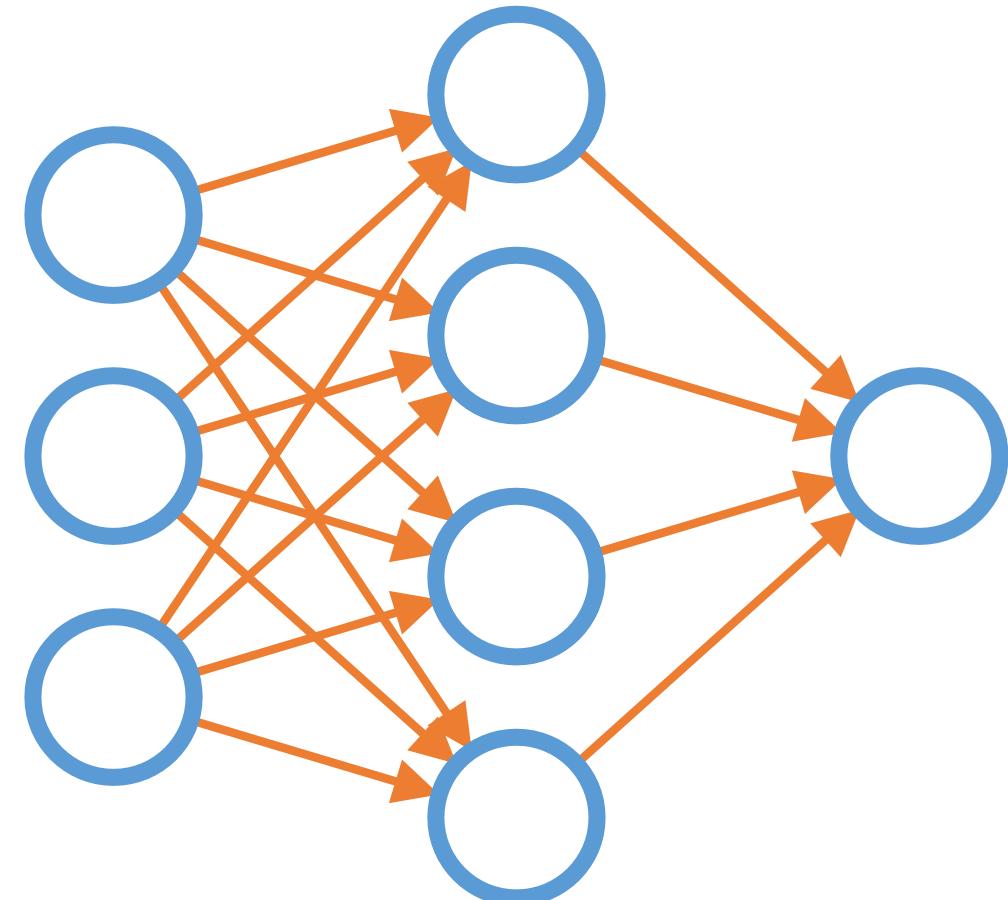
Neural Network Regression

Similar to NN classifier



Neural Network Regression

Similar to NN classifier
Numeric output



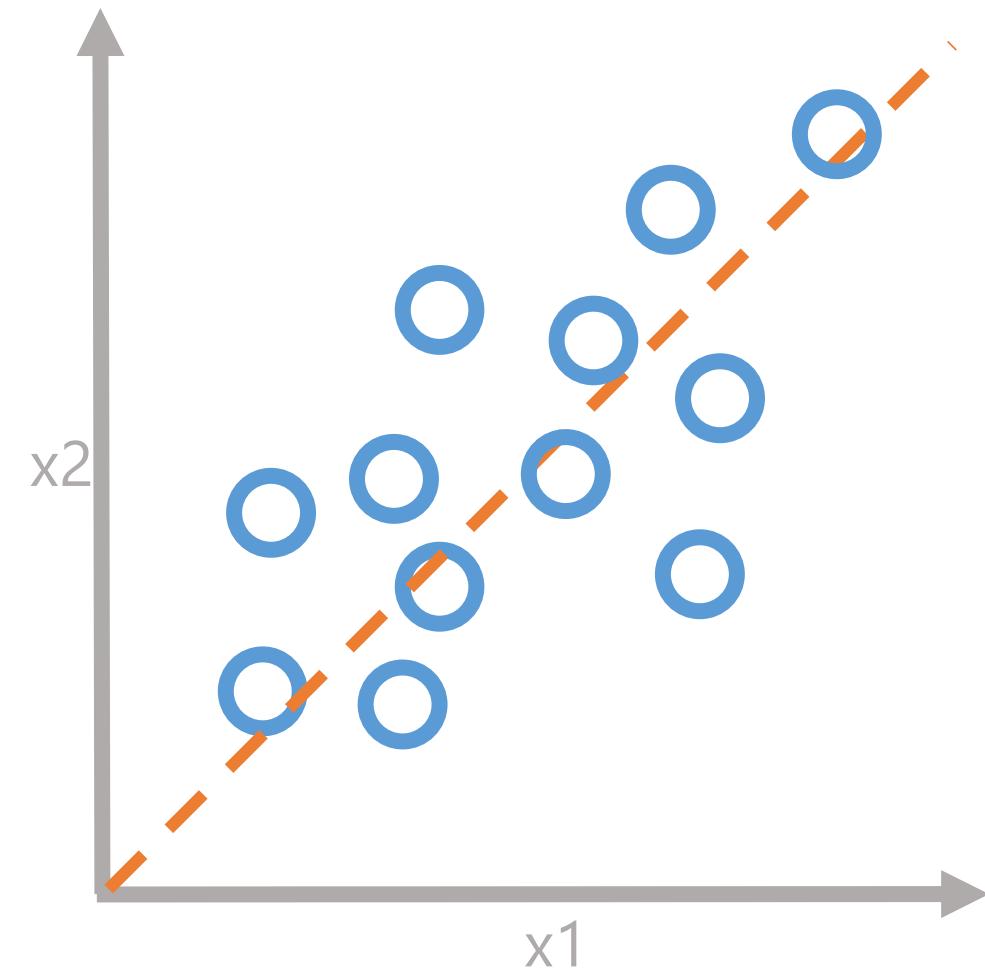
Real-World Examples

How much profit will we make?

What will the price be tomorrow?

How many units will they 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?

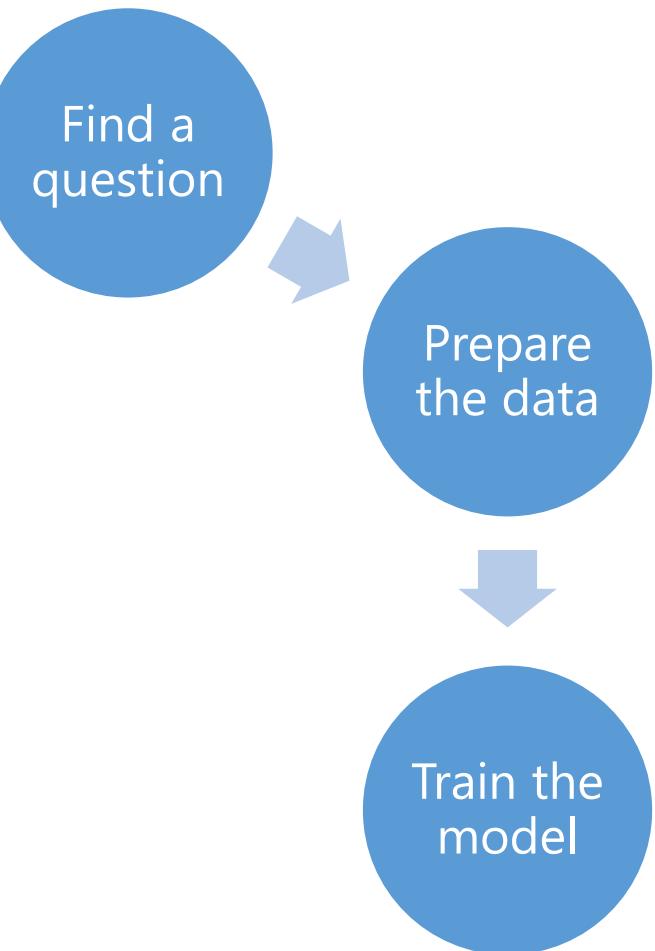


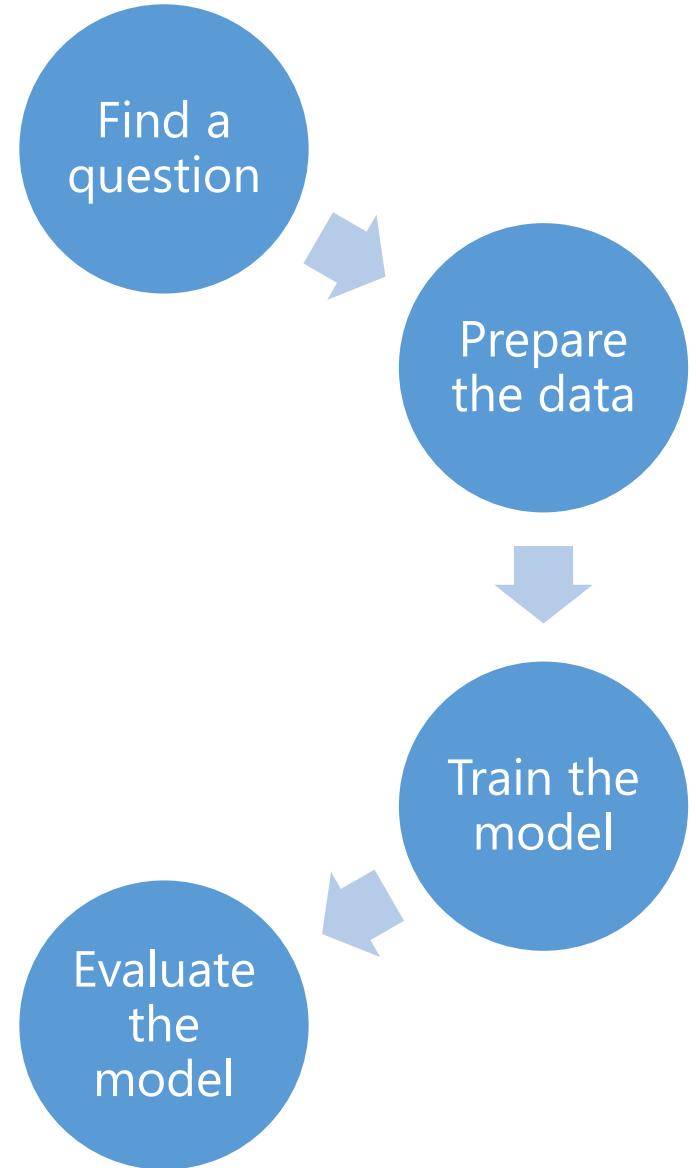
Find a
question

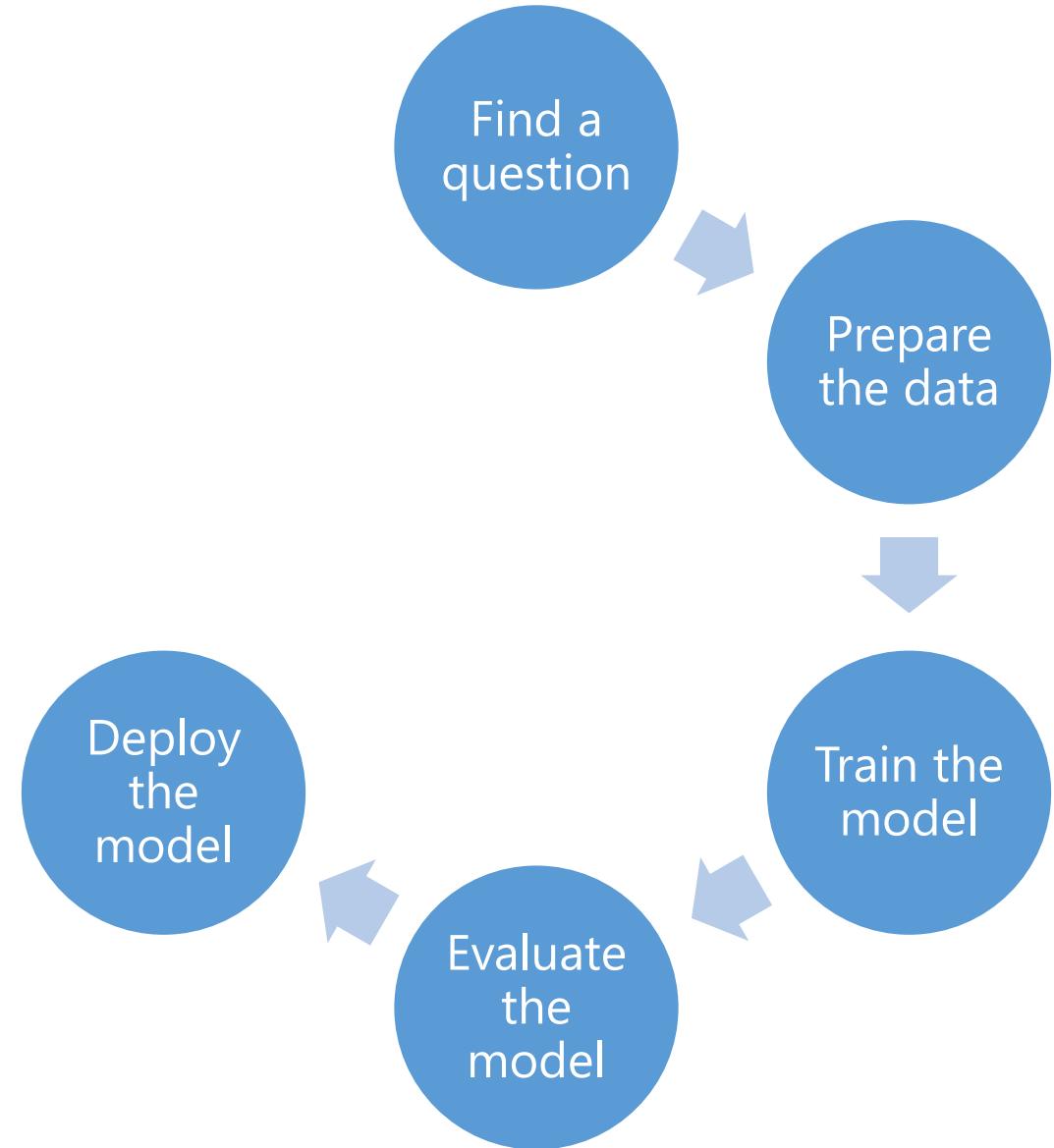
```
graph TD; A((Find a question)) --> B((Prepare the data))
```

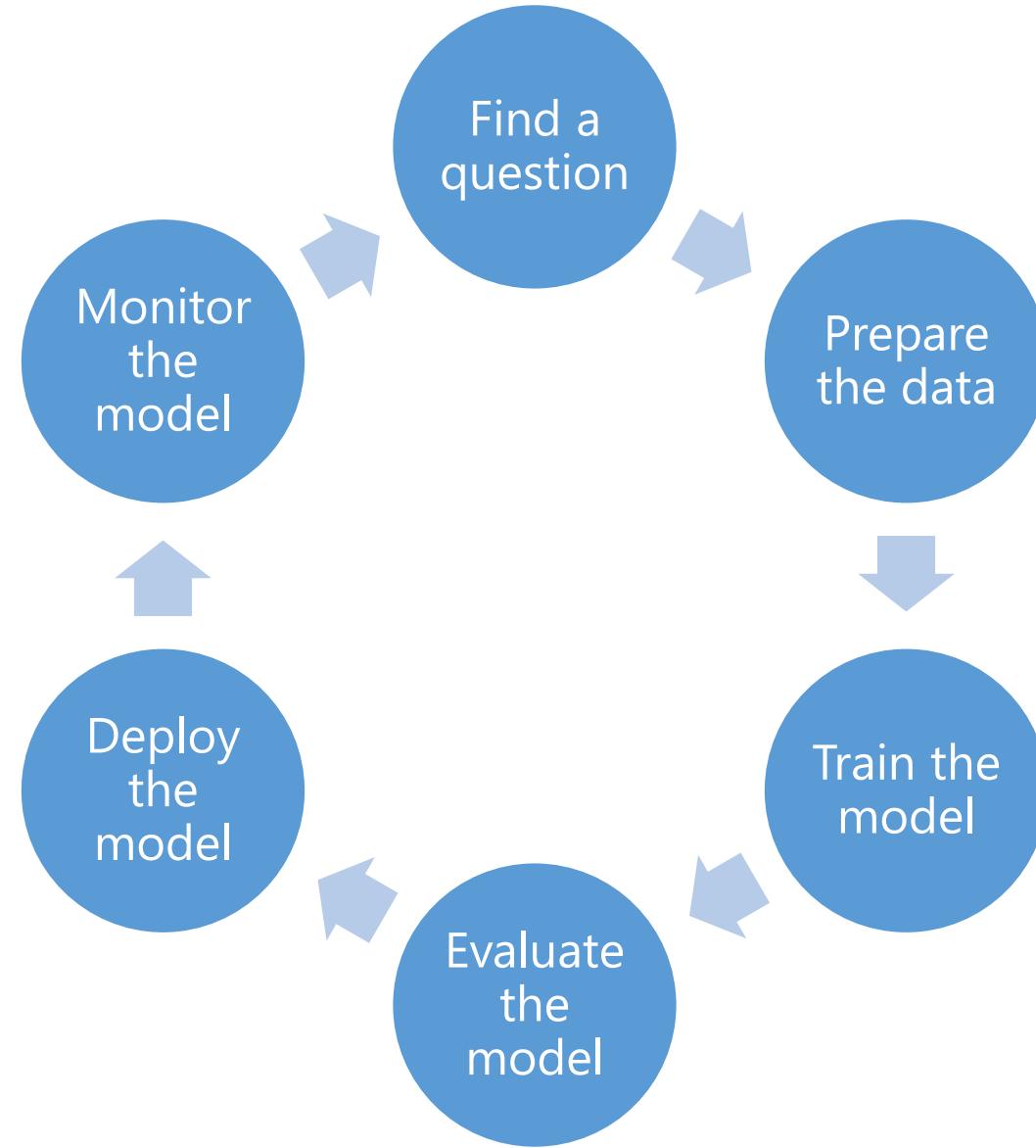
Find a
question

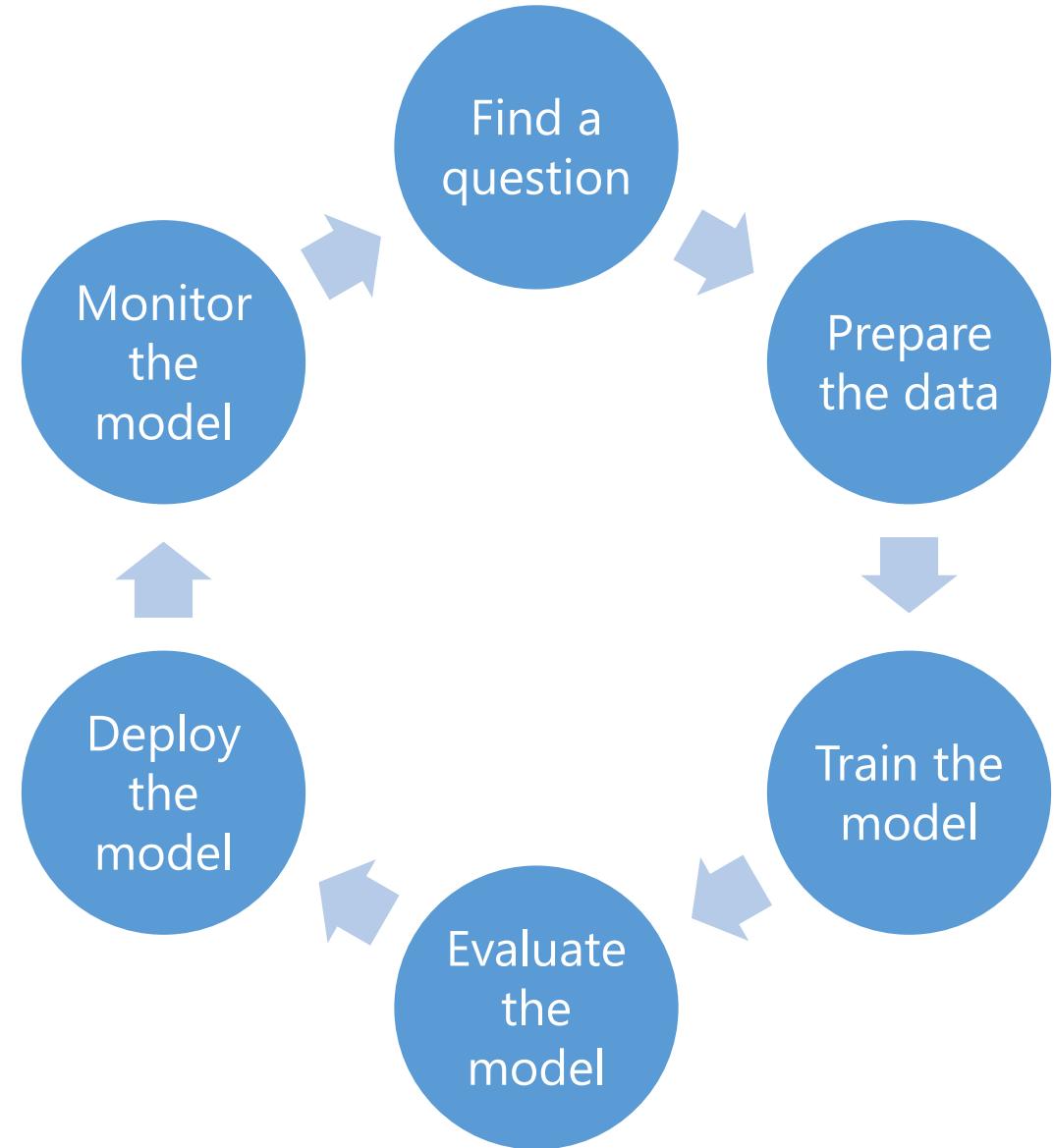
Prepare
the data











Creating accurate and robust
models is not easy

Find a
question

Monitor
the
model

Prepare
the data

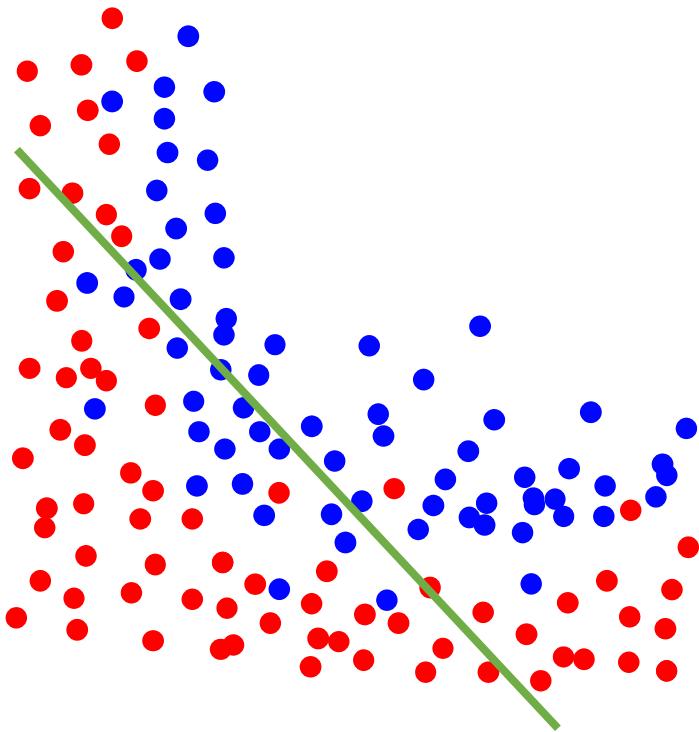
Deploy
the
model

Train the
model

Evaluate
the
model

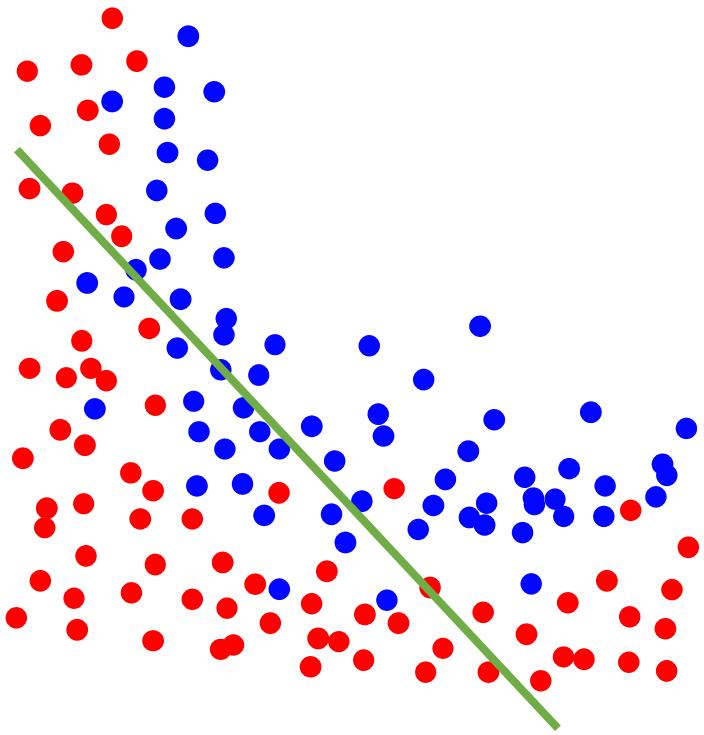
Goodness of Fit

Goodness of Fit

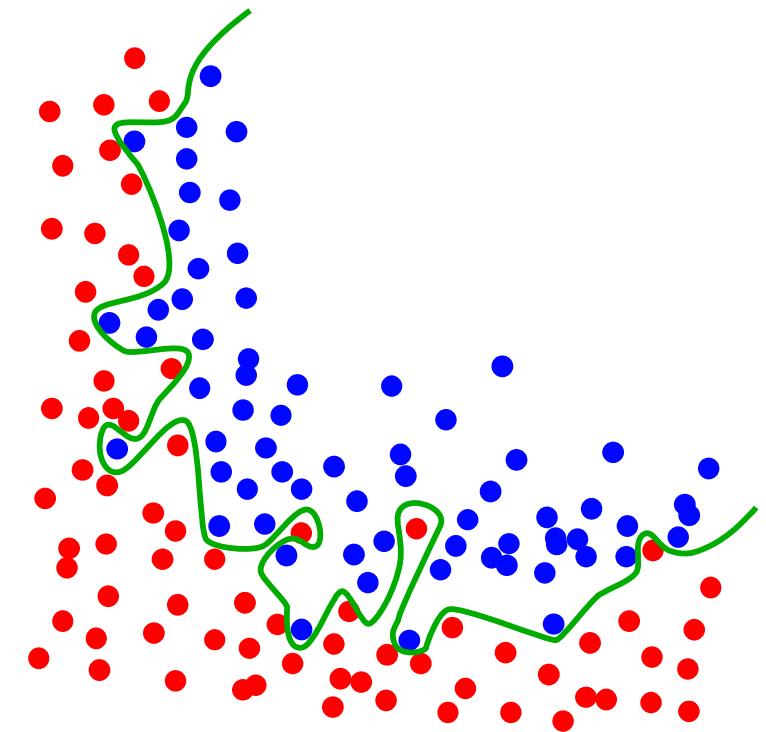


Underfit

Goodness of Fit

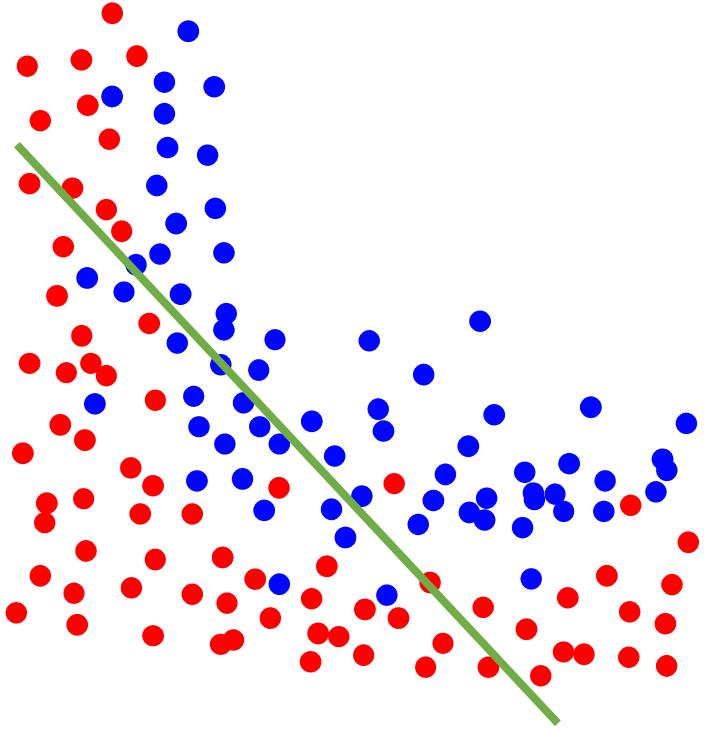


Underfit

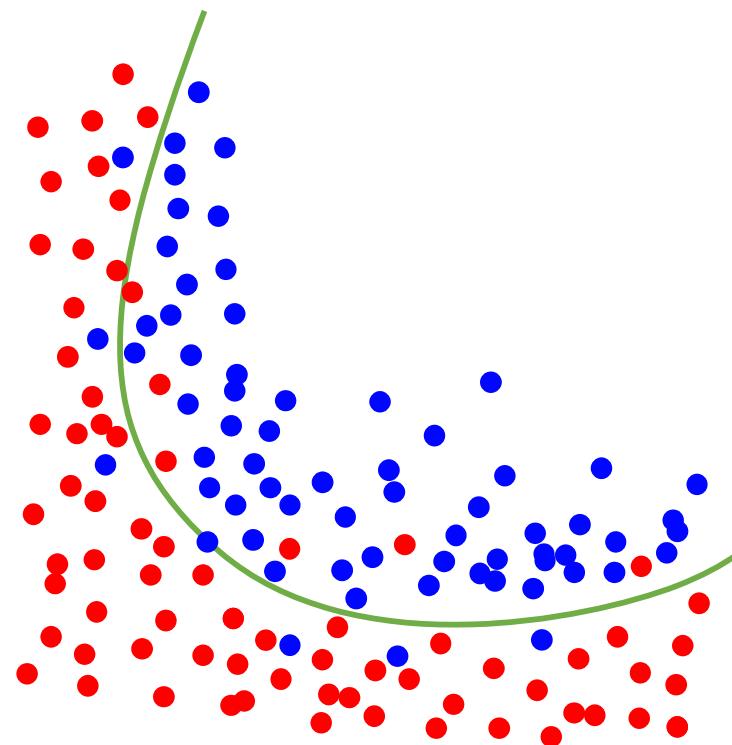


Overfit

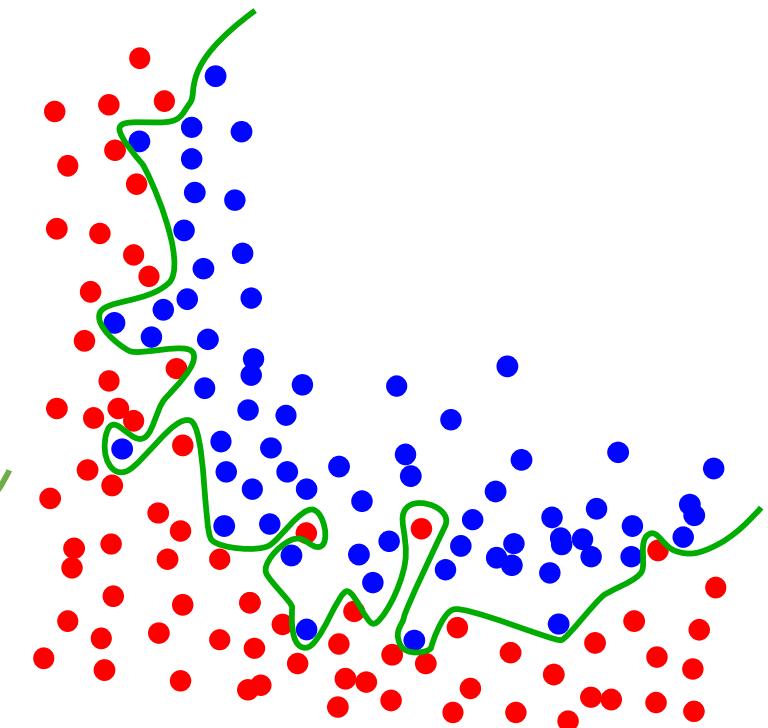
Goodness of Fit



Underfit



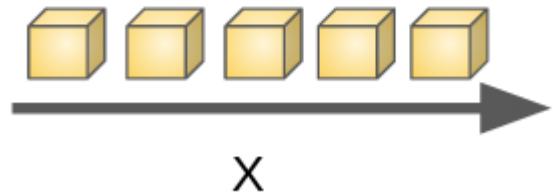
Good fit



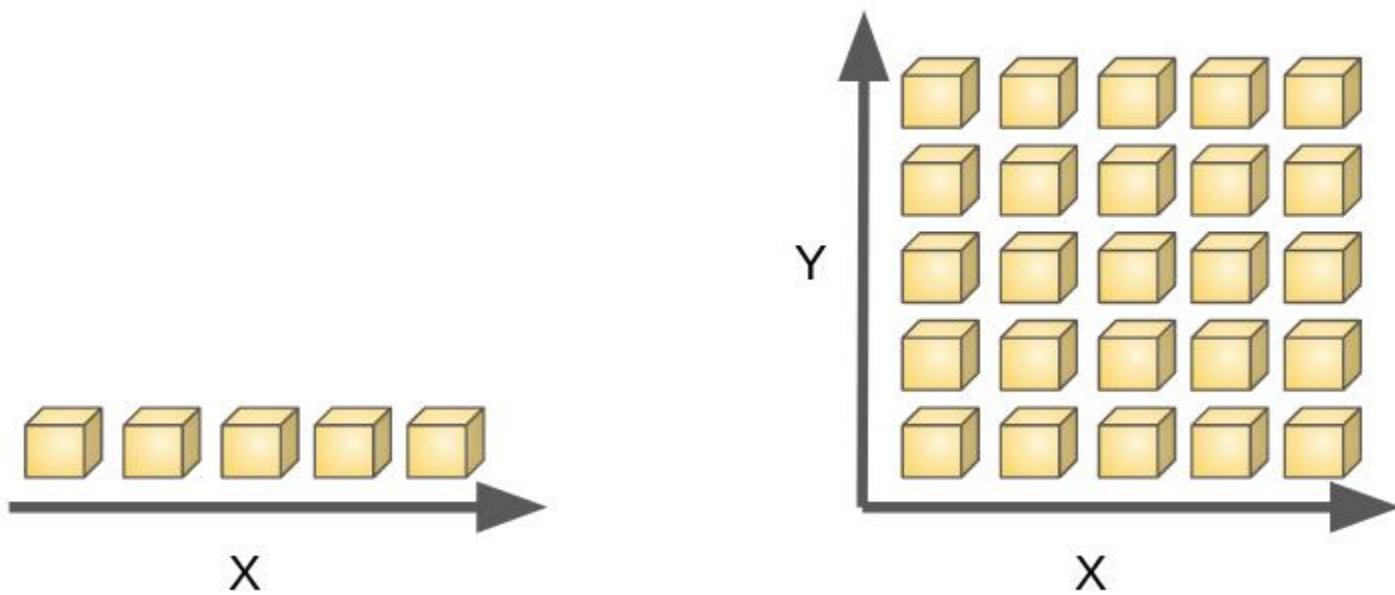
Overfit

Curse of Dimensionality

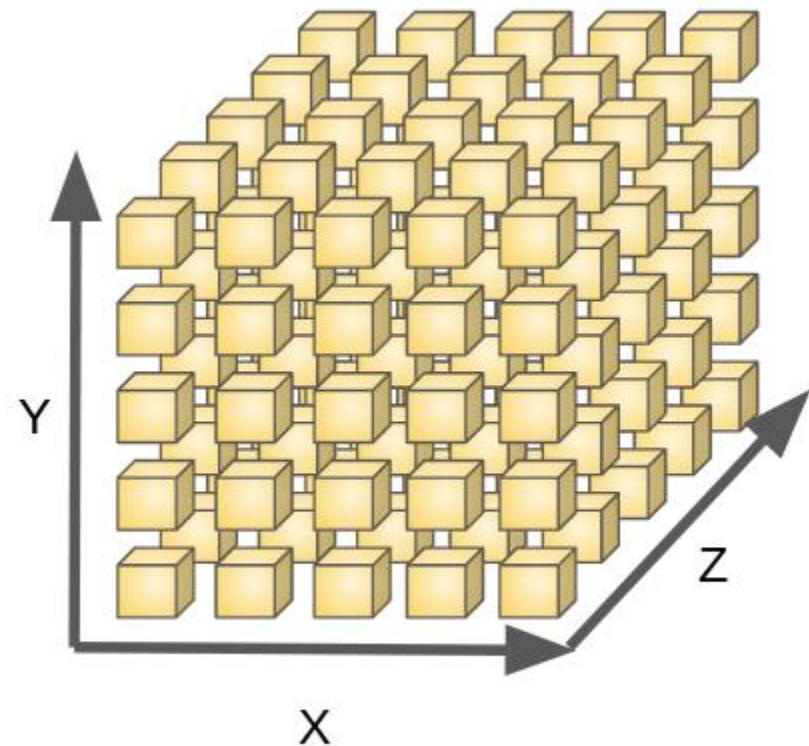
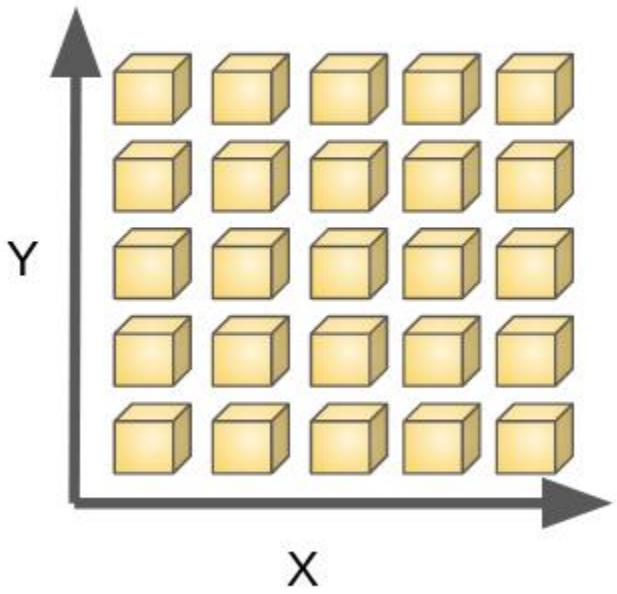
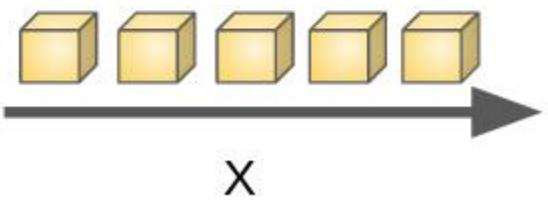
Curse of Dimensionality



Curse of Dimensionality



Curse of Dimensionality



A police officer in uniform stands in a dark environment, holding a flashlight that illuminates the scene. Another officer is partially visible in the background. The scene is dimly lit, with the primary light source being the flashlight.

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)

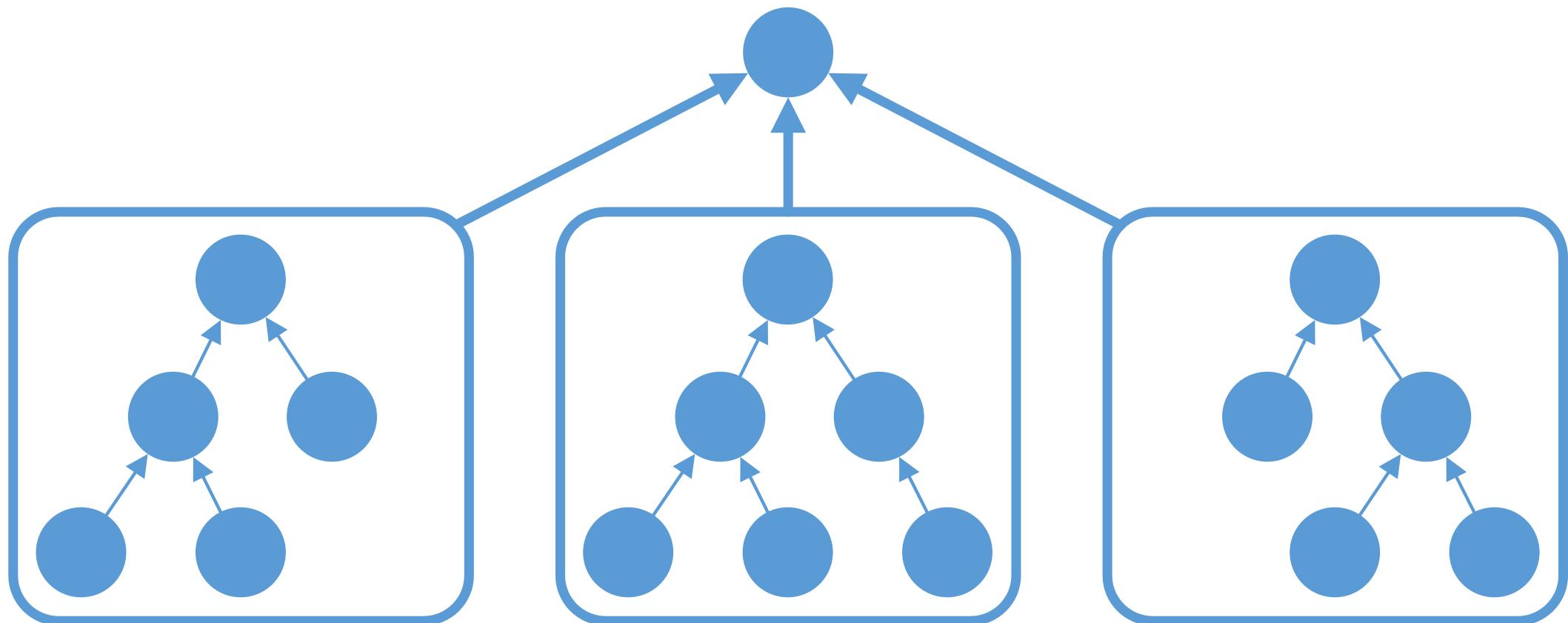
Goal: Predict survivors
of the Titanic

Conclusion

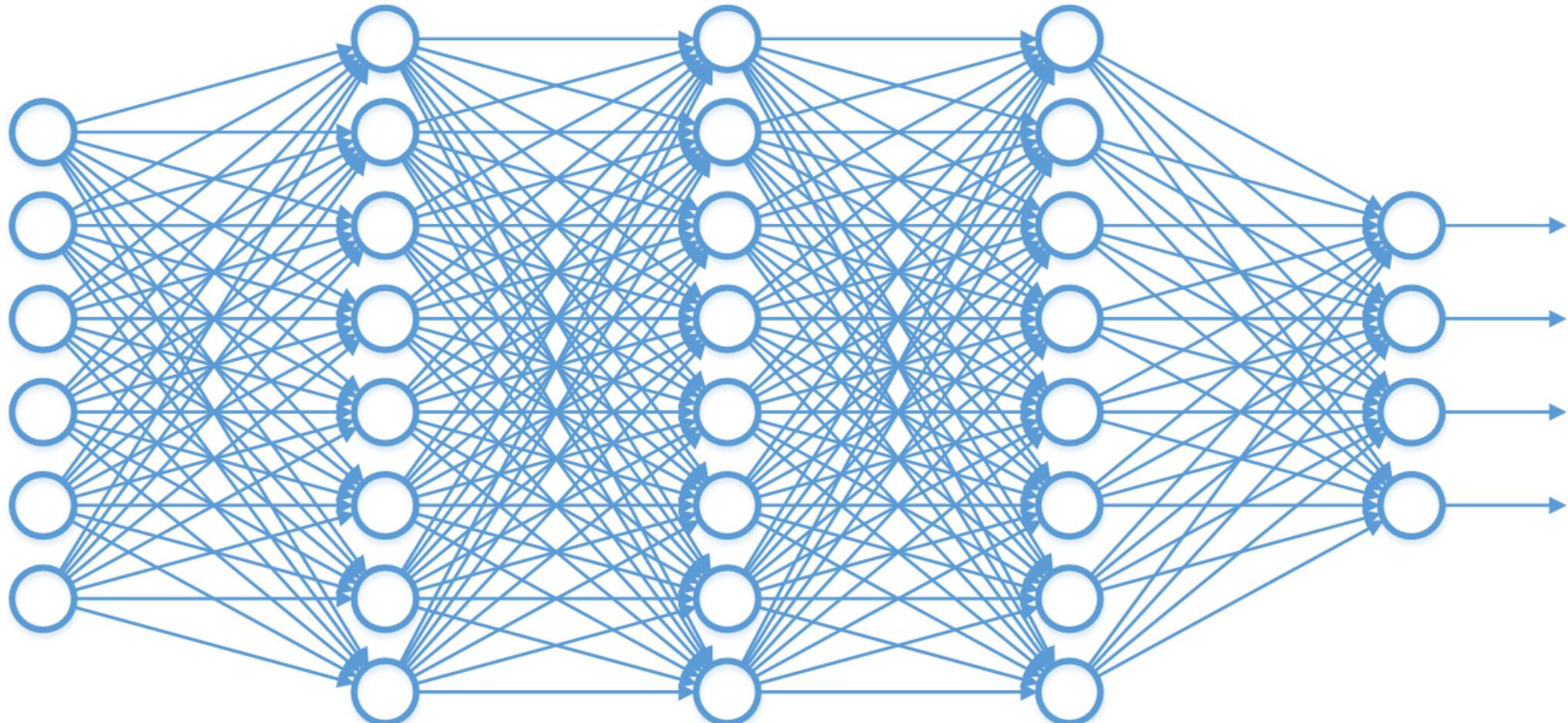


This is just the tip of the iceberg!

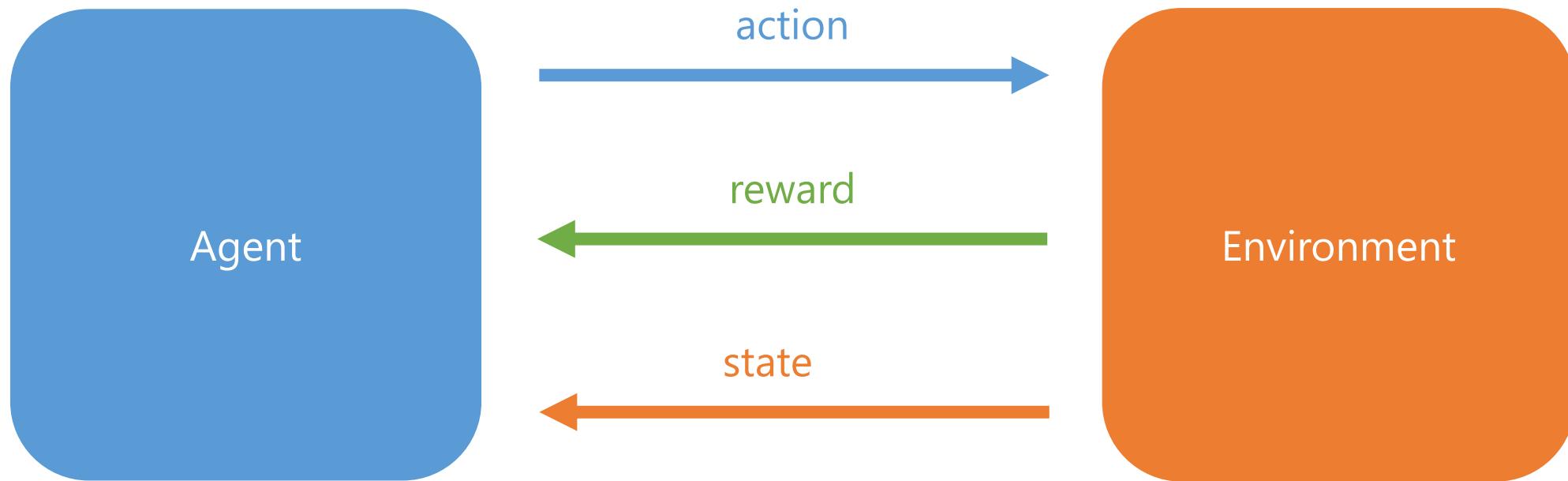
Ensemble Learning



Deep Learning



Reinforcement Learning



Where do we go from here?

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>

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.



Matthew is a data science consultant, author for [Pluralsight](#), international public speaker, a [Microsoft MVP](#), [ASPIInsider](#), and open-source software contributor.

2017-08-14 - Dev on Fire Interview

Feedback

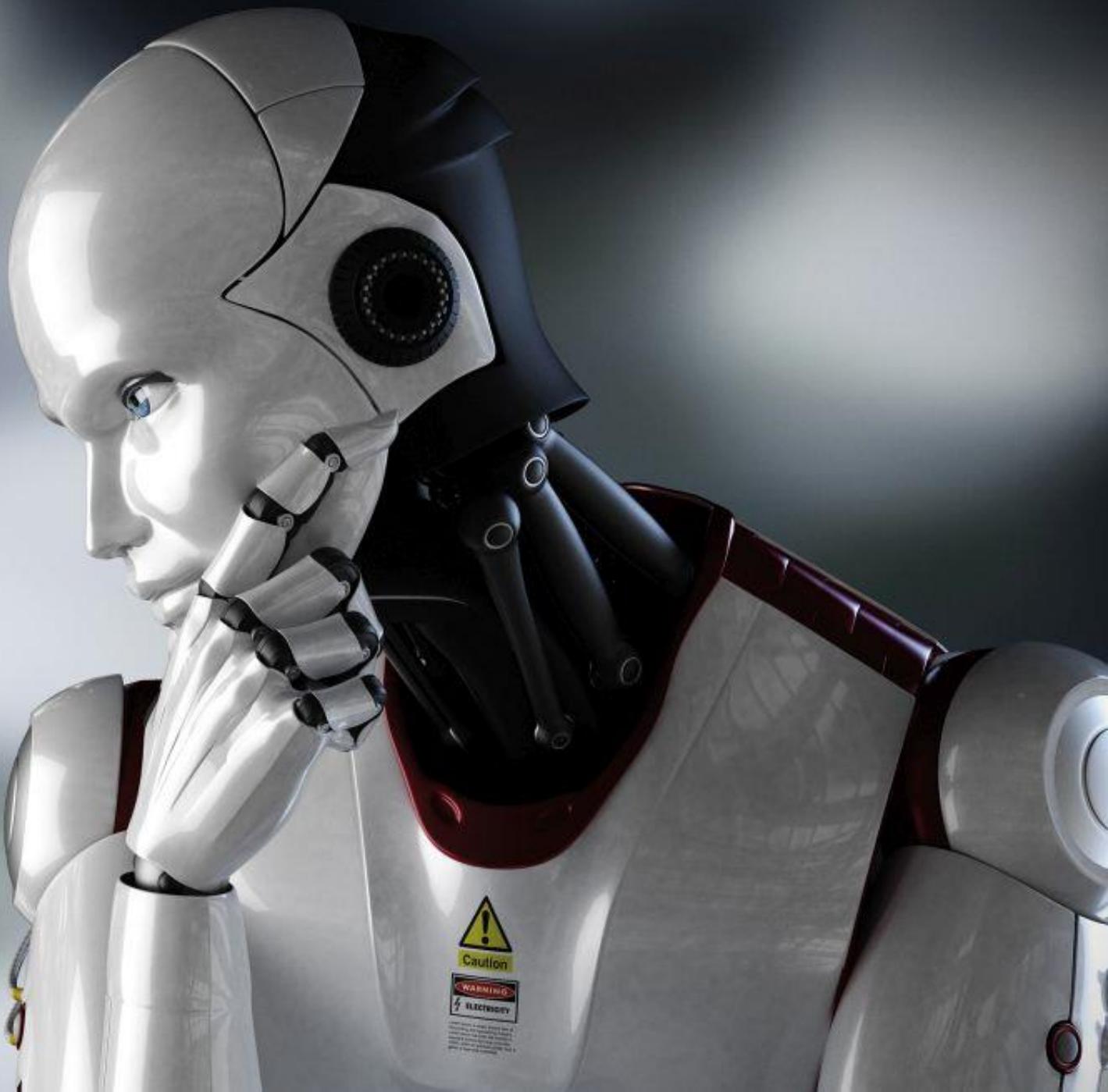
Very important to me!
What did you like?
What could I improve?

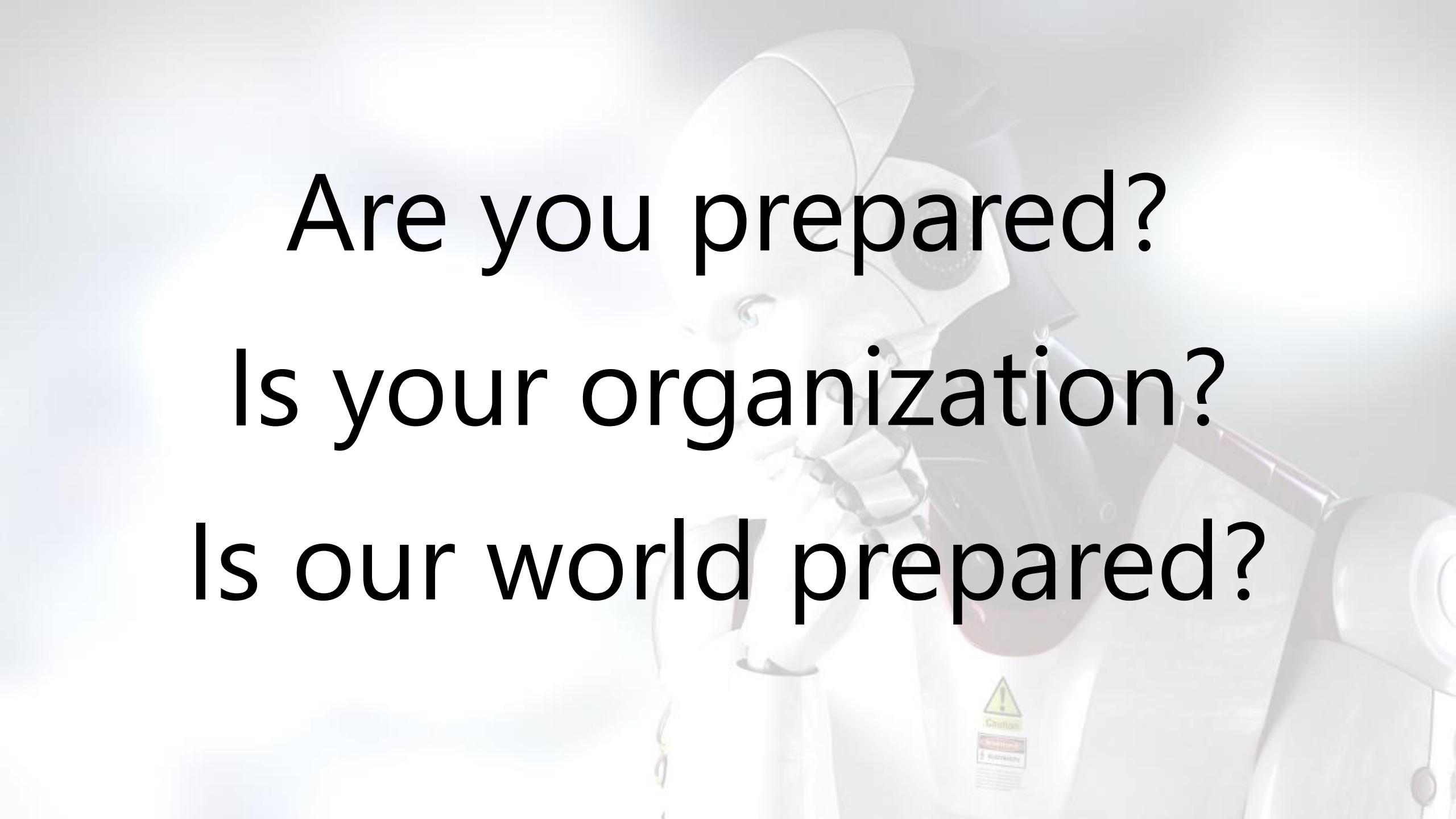


Conclusion

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



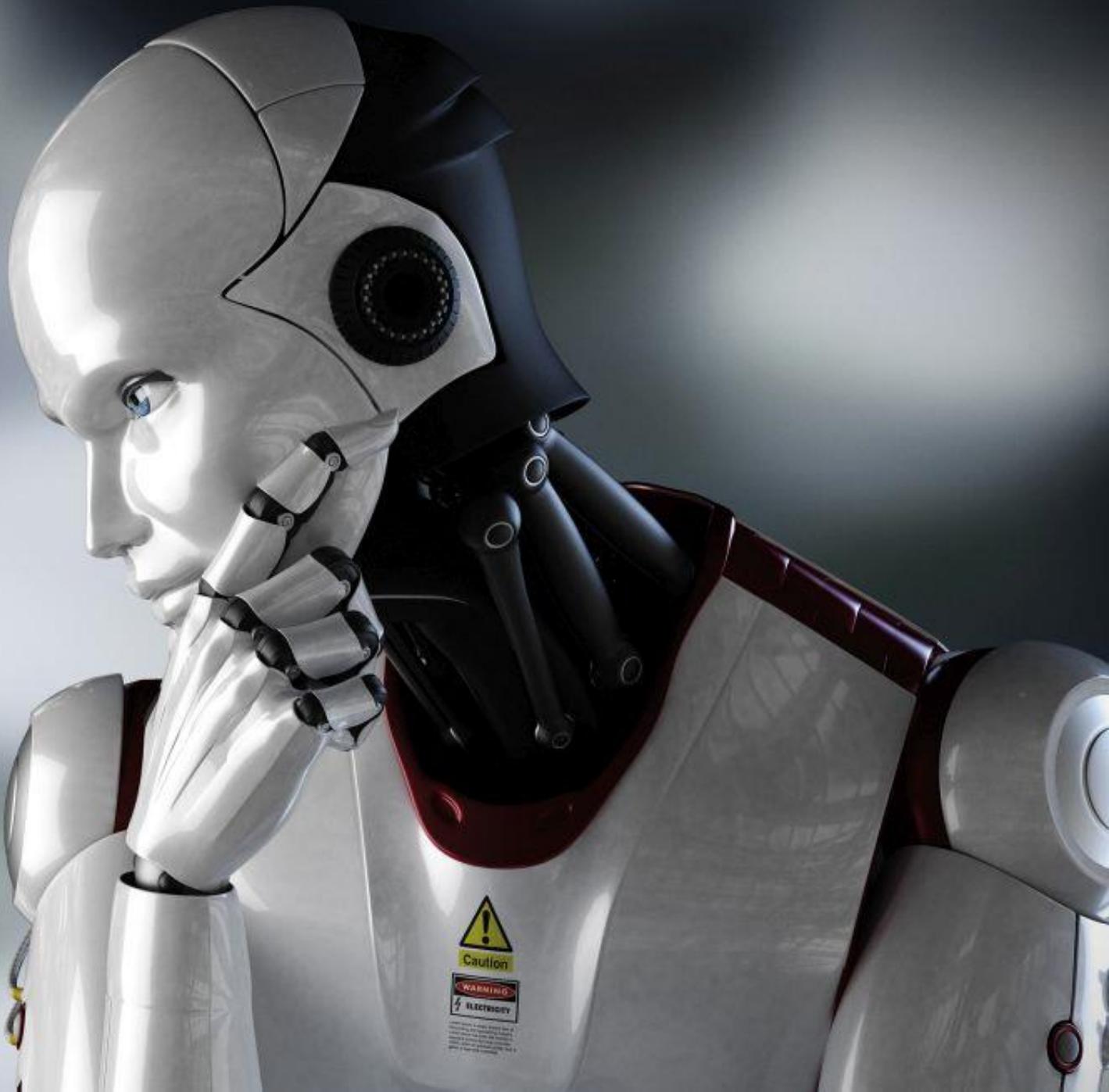


A man in a white protective suit and mask, holding a clipboard and a pen, looking down at something off-camera.

Are you prepared?

Is your organization?

Is our world prepared?



Contact Info

Matthew Renze

Data Science Consultant
Renze Consulting

Twitter: [@matthewrenze](https://twitter.com/matthewrenze)

Email: info@matthewrenze.com

Website: www.matthewrenze.com



Thank You! :)