

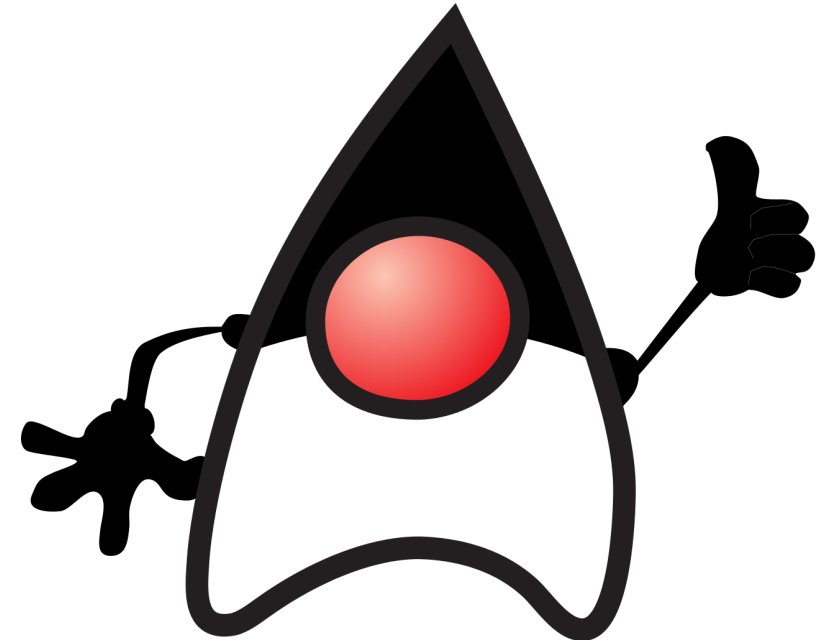
A Practical-ish Introduction to Data Science

Who Am I?

 @markawest

Who Am I?

- **Previously Java Developer and Architect.**



Who Am I?

- Previously Java Developer and Architect.
- **Currently building and managing a team of Data Scientists at Bouvet Oslo.**

bouvet

Who Am I?

- Previously Java Developer and Architect.
- Currently building and managing a team of Data Scientists at Bouvet Oslo.
- **Leader javaBin (Norwegian Java User Group).**



Agenda

**What is Data
Science?**

**Machine
Learning
Algorithms**

**Practical
Example**

Agenda

**What is Data
Science?**

**Machine
Learning
Algorithms**

**Practical
Example**

Agenda

**What is Data
Science?**

**Machine
Learning
Algorithms**

**Practical
Example**

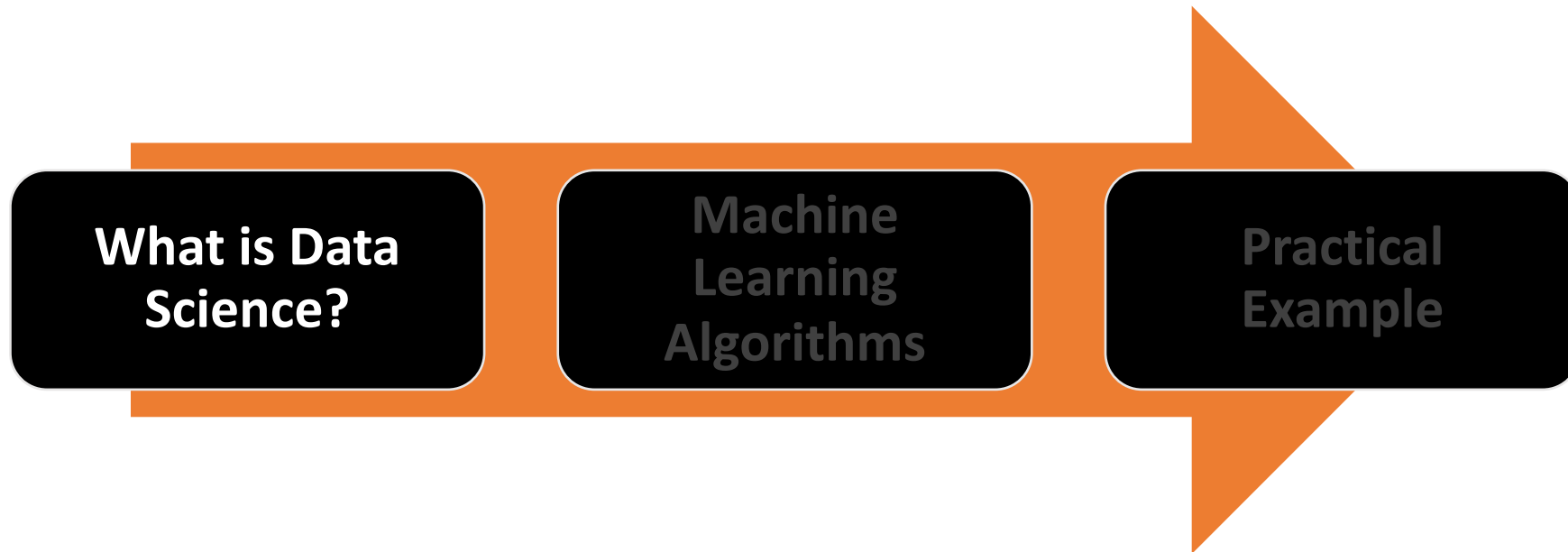
Agenda

**What is Data
Science?**

**Machine
Learning
Algorithms**

**Practical
Example**

What is Data Science?



*“Data Science... is an **interdisciplinary** field of **scientific** methods, processes, and systems to **extract knowledge** or insight from data...”*

Wikipedia

 @markawest

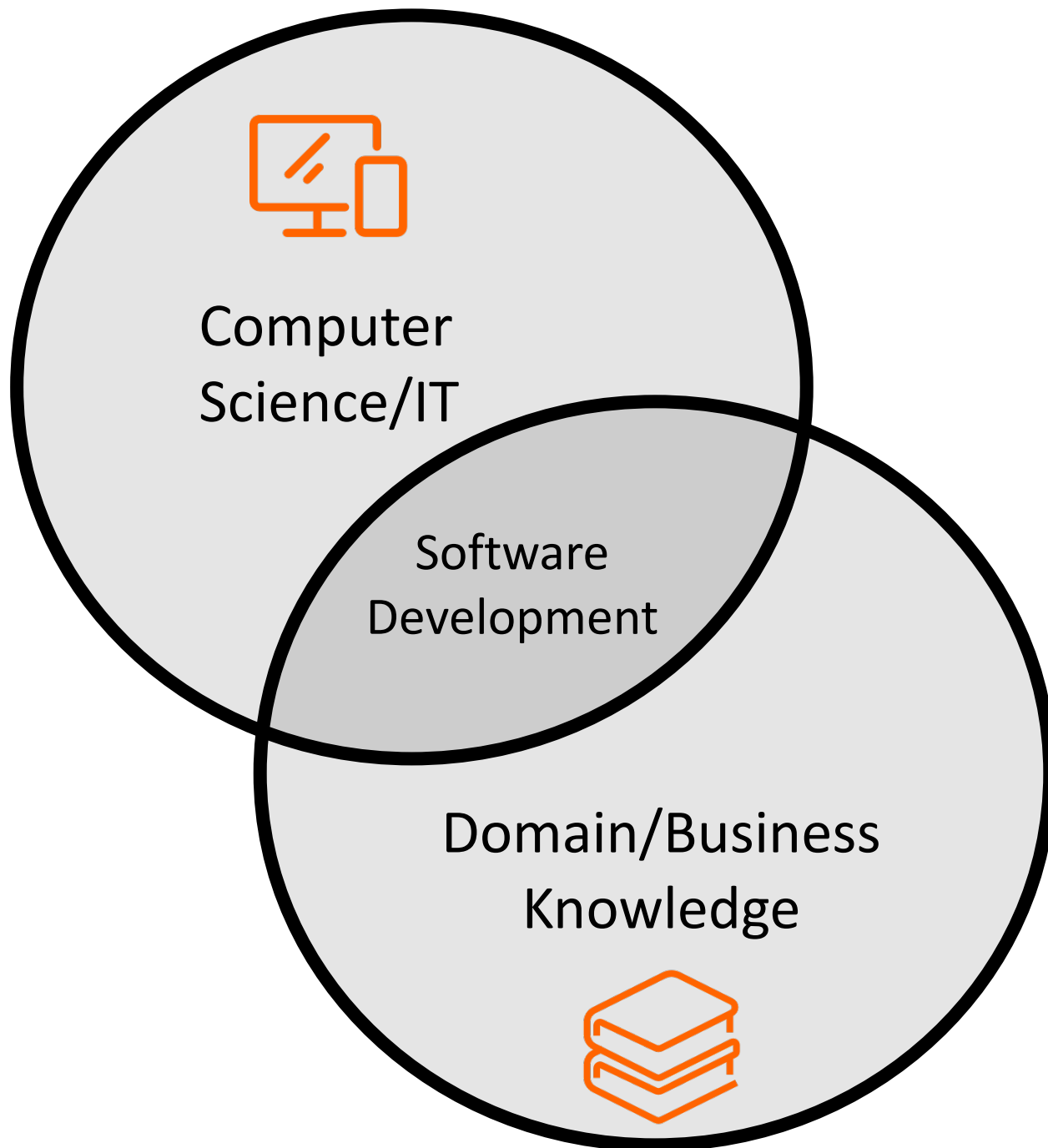
*“Data Science... is an **interdisciplinary** field of **scientific** methods, processes, and systems to **extract knowledge** or insight from data...”*

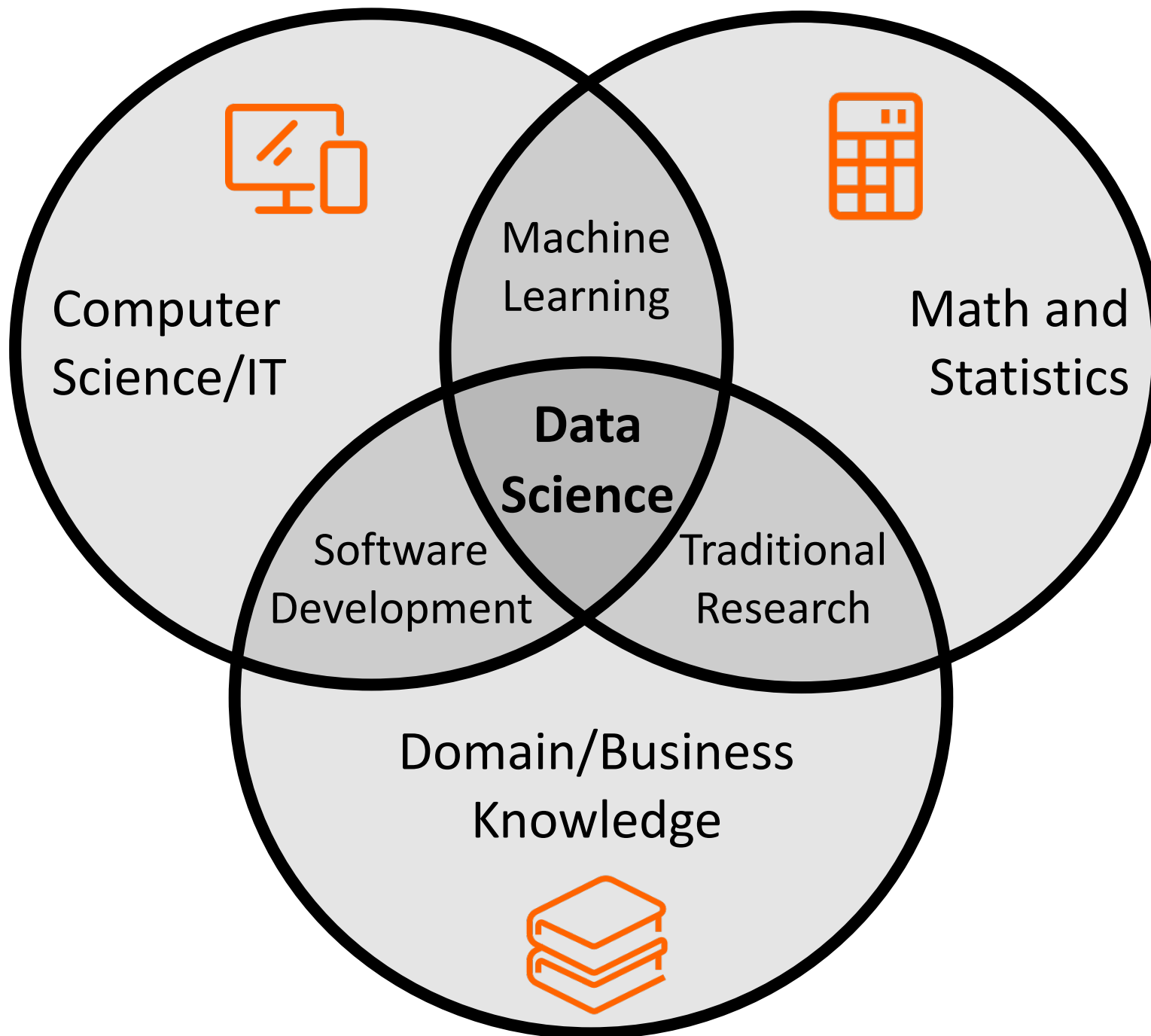
Wikipedia

 @markawest



Computer
Science/IT





*“Data Science... is an **interdisciplinary** field of **scientific** methods, processes, and systems to **extract knowledge** or insight from data...”*

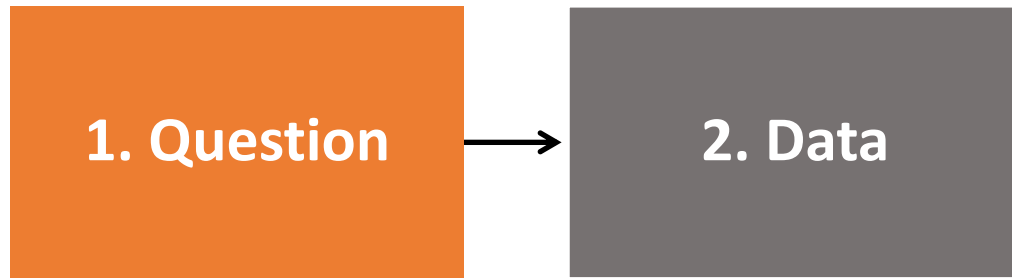
Wikipedia

 @markawest

Data Science Process : Hypothesis Driven

1. Question

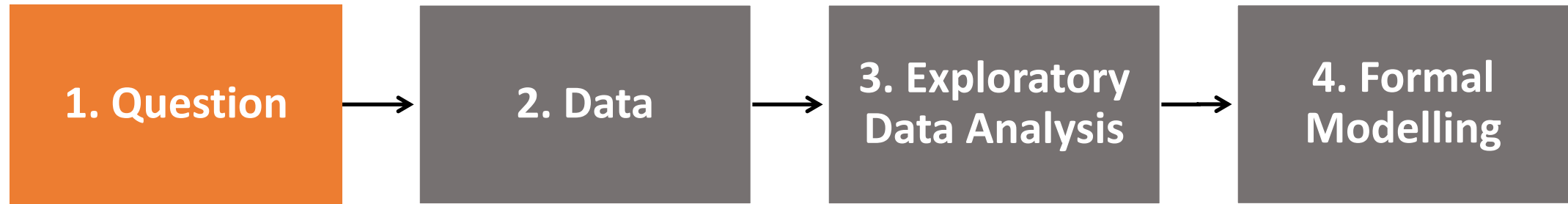
Data Science Process : Hypothesis Driven



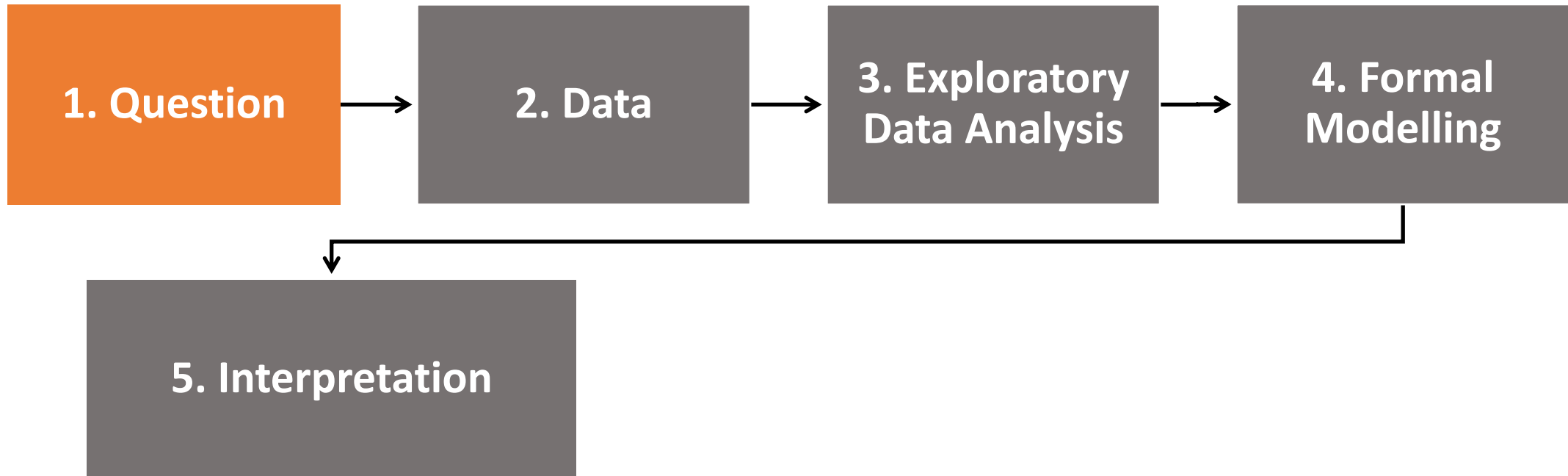
Data Science Process : Hypothesis Driven



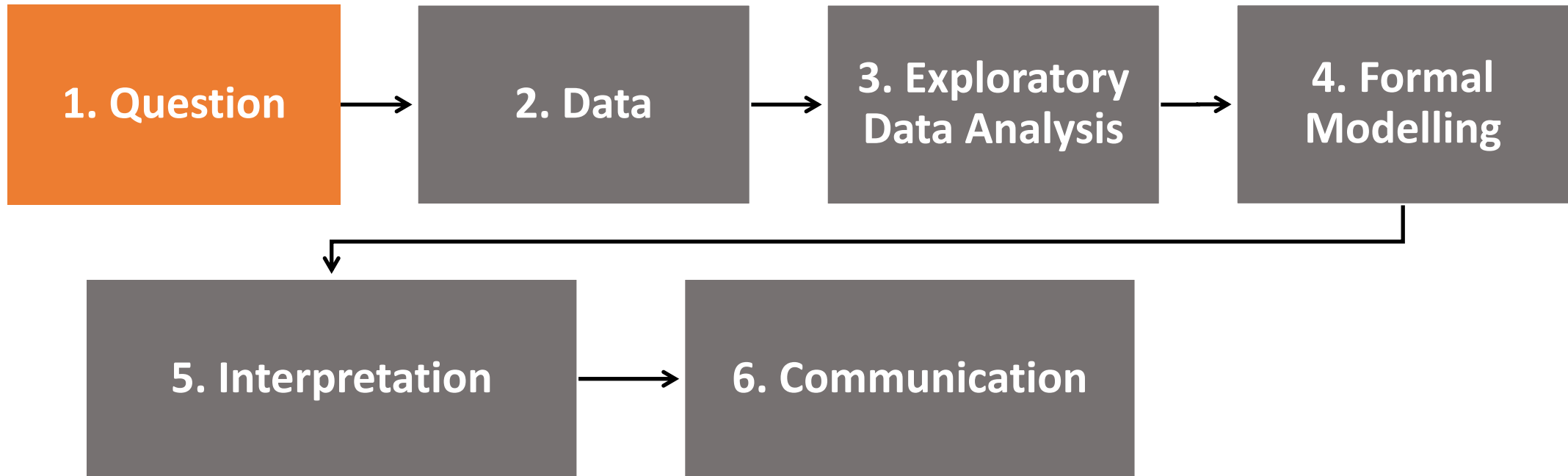
Data Science Process : Hypothesis Driven



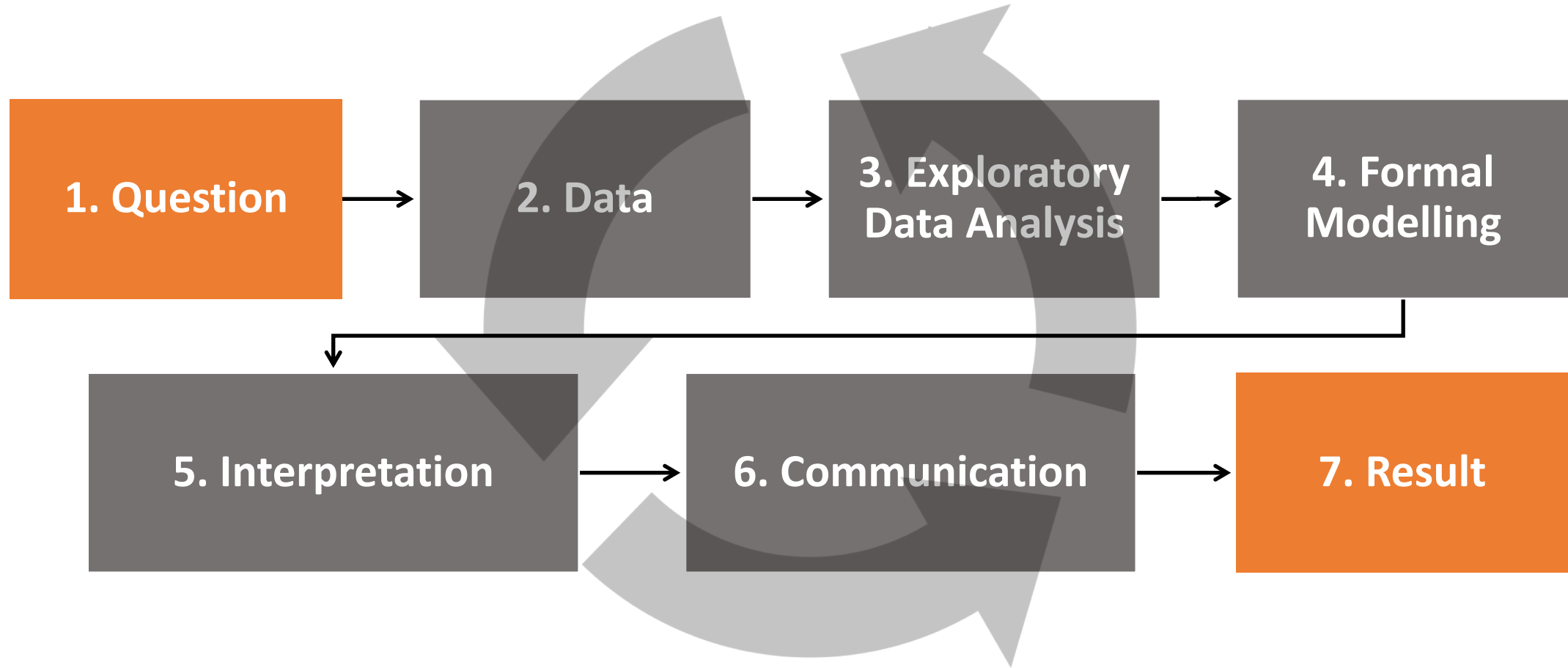
Data Science Process : Hypothesis Driven



Data Science Process : Hypothesis Driven



Data Science Process : Hypothesis Driven



Roles Required in a Data Science Project

Roles Required in a Data Science Project

- Prove / disprove hypotheses.
- Information and Data gathering.
- Data wrangling.
- Algorithm and ML models.
- Communication.

**Data
Scientist**



Roles Required in a Data Science Project

- Prove / disprove hypotheses.
- Information and Data gathering.
- Data wrangling.
- Algorithm and ML models.
- Communication.

**Data
Scientist**



- Build Data Driven Platforms.
- Operationalize Algorithms and Machine Learning models.
- Data Integration.
- Monitoring.

**Data
Engineer**



Roles Required in a Data Science Project

- Prove / disprove hypotheses.
- Information and Data gathering.
- Data wrangling.
- Algorithm and ML models.
- Communication.

**Data
Scientist**



- Build Data Driven Platforms.
- Operationalize Algorithms and Machine Learning models.
- Data Integration.
- Monitoring.

**Data
Engineer**



- Storytelling.
- Build Dashboards and other Data visualizations.
- Provide insight through visual means.

**Data
Visualization**



Roles Required in a Data Science Project

- Prove / disprove hypotheses.
- Information and Data gathering.
- Data wrangling.
- Algorithm and ML models.
- Communication.

Data Scientist



- Build Data Driven Platforms.
- Operationalize Algorithms and Machine Learning models.
- Data Integration.
- Monitoring.

Data Engineer



- Storytelling.
- Build Dashboards and other Data visualizations.
- Provide insight through visual means.

Data Visualization



- Project Management.
- Manage stakeholder expectations.
- Maintain a Vision.
- Facilitate.
- Evangelize.

Process Owner



*“Data Science... is an **interdisciplinary** field of **scientific** methods, processes, and systems to **extract knowledge** or insight from data...”*

Wikipedia

 @markawest



Data Science as an Evolution of BI

	Business Intelligence	Data Science Adds..
Data Sources	Structured Data, most often from Relational Database Management Systems (RDBMS).	Unstructured Data (log files, audio, images, emails, tweets, raw text, documents).
Available Tools	Data Visualization, Statistics.	Machine Learning.
Goals	Provide support to strategic decision making, based on historical data.	Provide business value through advanced functionality.

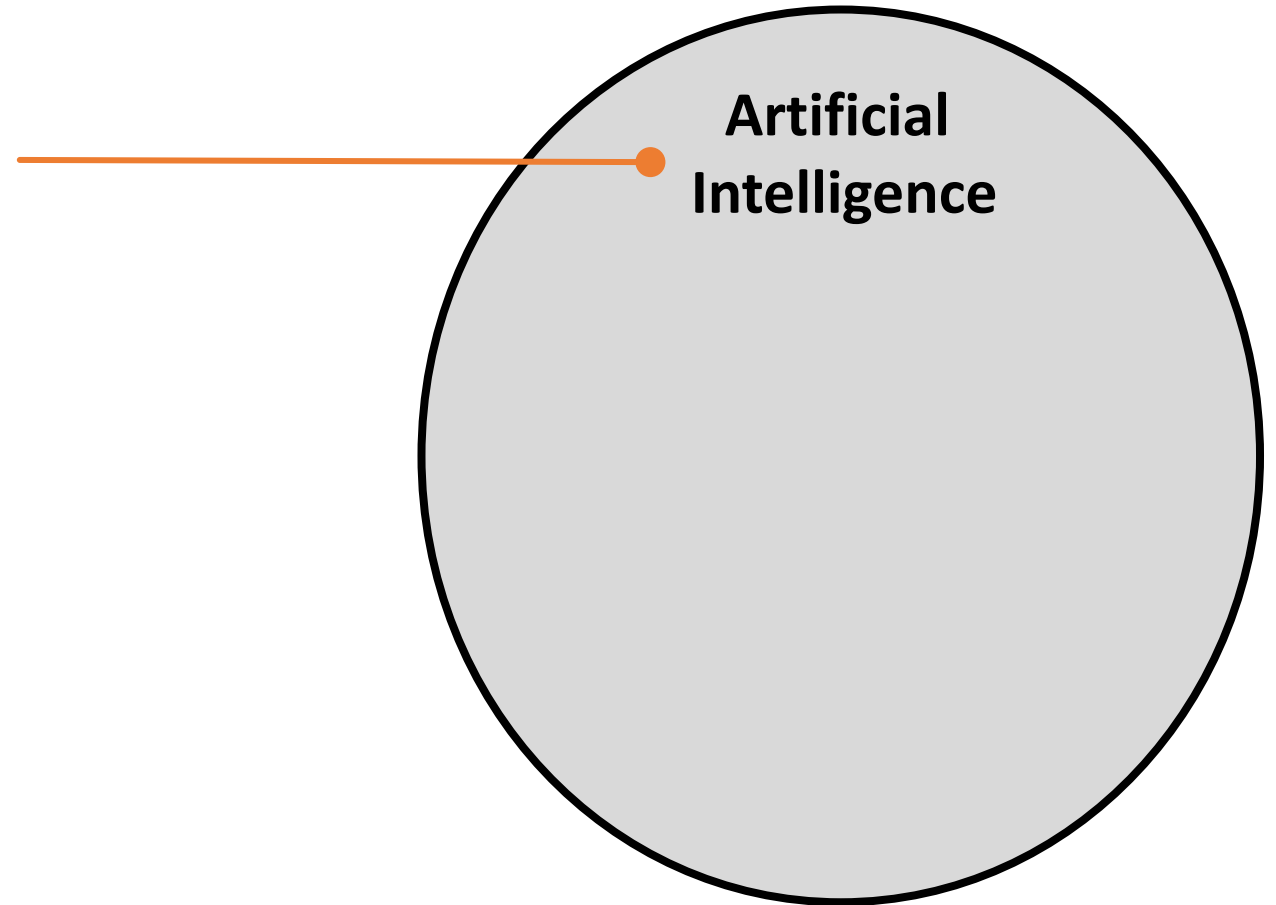
Source: <https://www.linkedin.com/pulse/data-science-business-intelligence-whats-difference-david-rostcheck>

Machine Learning: A Tool for Data Science

Machine Learning: A Tool for Data Science

Artificial Intelligence

Enabling computers to mimic human intelligence and behavior.



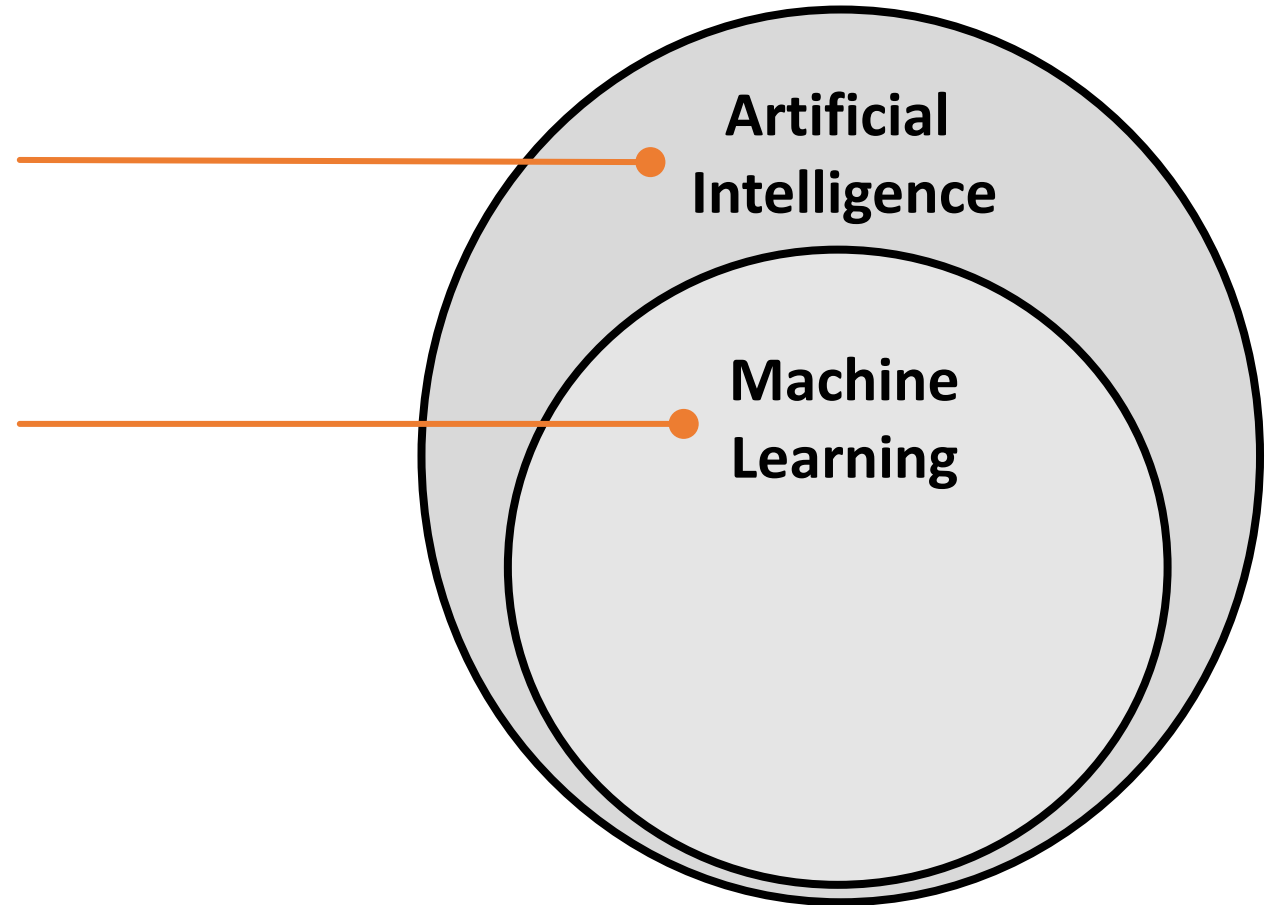
Machine Learning: A Tool for Data Science

Artificial Intelligence

Enabling computers to mimic human intelligence and behavior.

Machine Learning

Algorithms allowing computers to learn, make predictions and describe data without being explicitly programmed.



Machine Learning: A Tool for Data Science

Artificial Intelligence

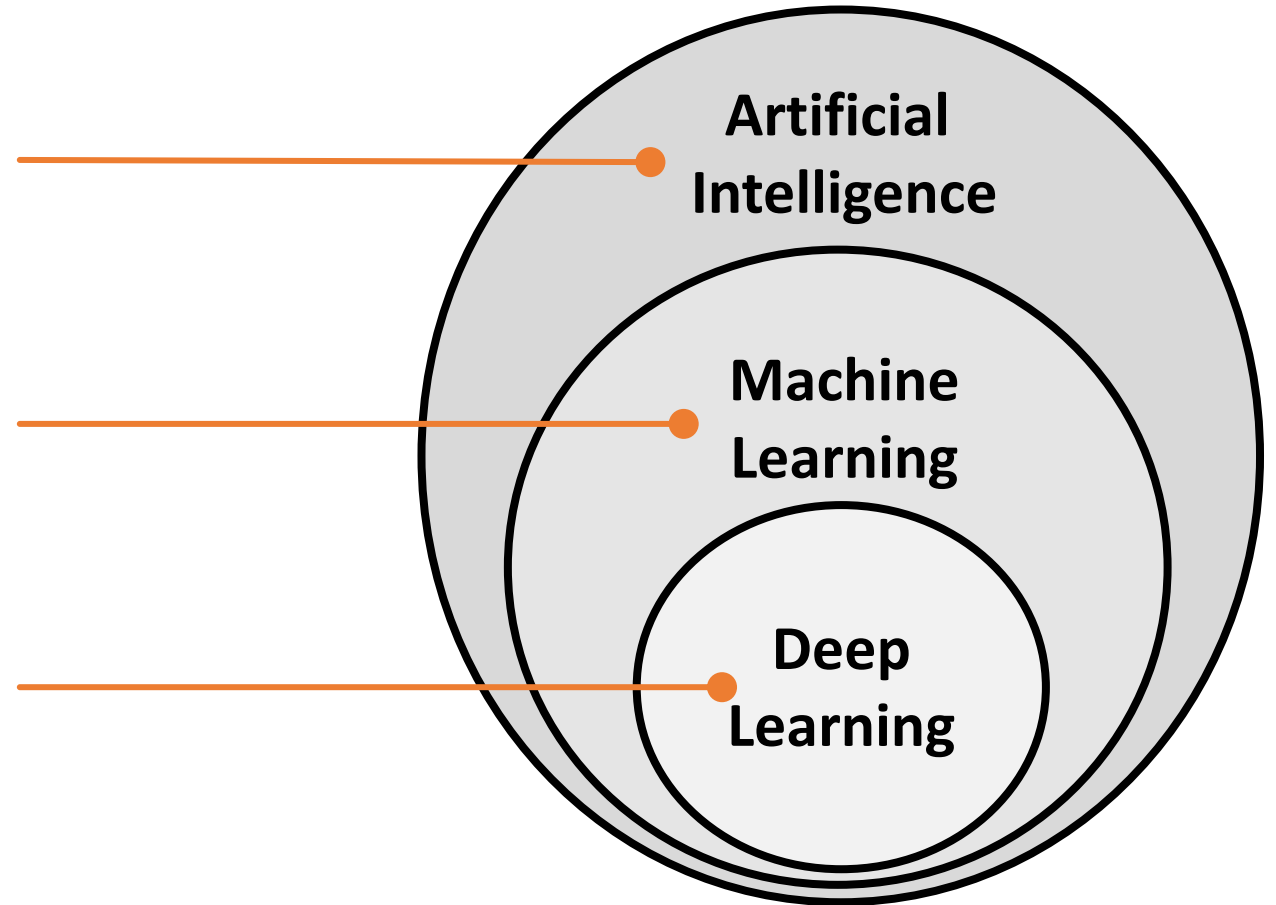
Enabling computers to mimic human intelligence and behavior.

Machine Learning

Algorithms allowing computers to learn, make predictions and describe data without being explicitly programmed.

Deep Learning

Black box learning with multi-layered Neural Networks.



What is Data Science: Key Takeaways

- Data Scientists require **Math and Statistics** skills in addition to traditional Software Development.
- Data Science is **Hypothesis Driven**.
- Data Science projects **require a range of competencies/roles**.
- Data Science can be seen as an **evolution of Business Intelligence**, providing additional capabilities through the application of **cutting edge technologies** and **unstructured data**.

Machine Learning Algorithms

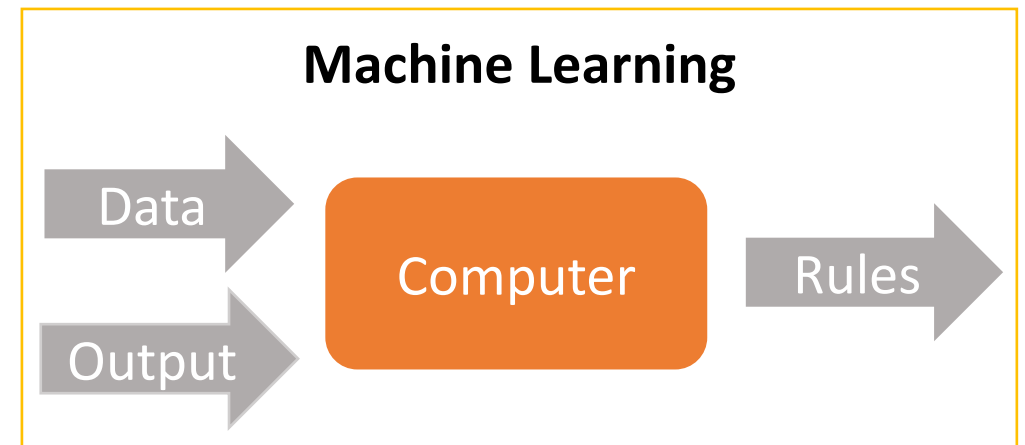
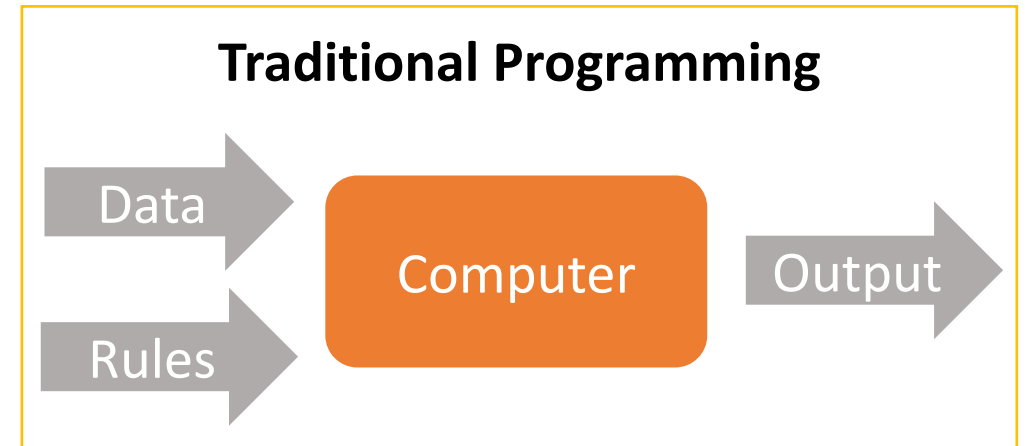
What is Data Science?

Machine Learning Algorithms

Practical Example

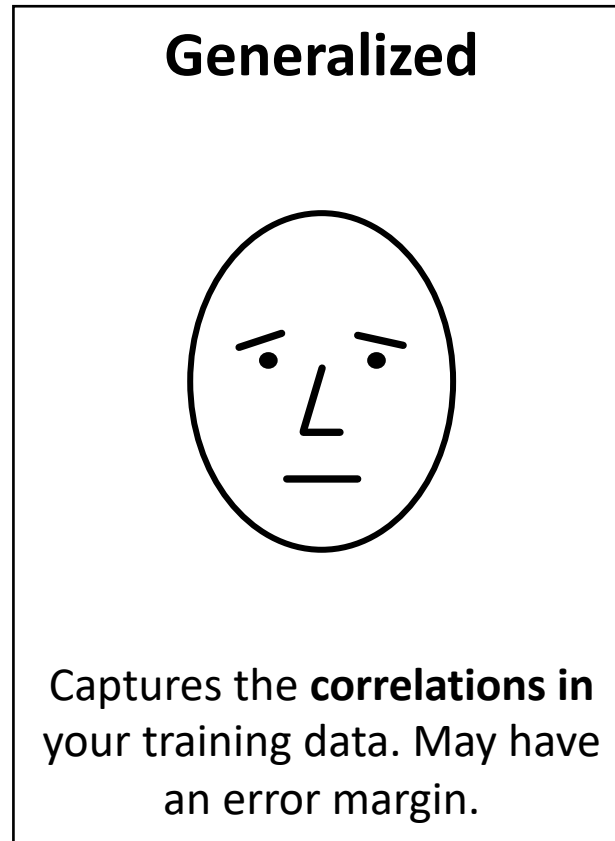
*“Machine Learning:
Field of study that gives
computers the ability to
learn without being
explicitly programmed.”*

Arthur L. Samuel
IBM Journal of Research and Development, 1959



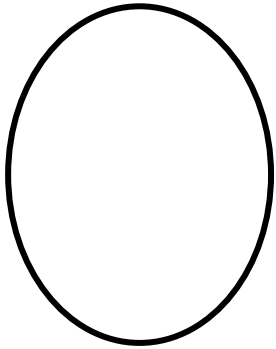
The Art of The Generalized Model

The Art of The Generalized Model



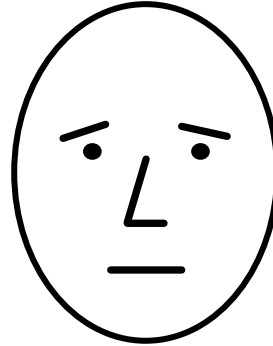
The Art of The Generalized Model

Underfitted



Model overlooks **underlying patterns in your training data.**

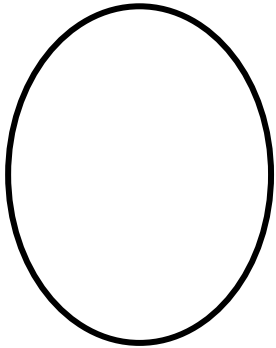
Generalized



Captures the **correlations** in your training data. May have an error margin.

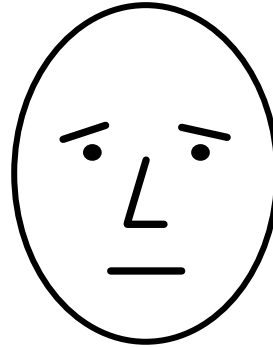
The Art of The Generalized Model

Underfitted



Model overlooks **underlying patterns in your training data**.

Generalized



Captures the **correlations in your training data**. May have an error margin.

Overfitted



Model **memorizes** the training data rather than finding underlying patterns.

Machine Learning Types

Supervised Learning

Model trained on **historical** data. Resulting model can be used to make predictions on new data.

Use Case: **Predicting** a value based on patterns discovered in previous data.

Unsupervised Learning

Algorithm finds **trends and patterns** in data, without prior training on historical data.

Use Case: **Describing** your data based on statistical analysis.

Reinforcement Learning

Model uses a **feedback loop** to iteratively improve its performance.

Use Case: **Learning** how to best solve a problem based on trial and error.

Common Machine Learning Algorithm Types

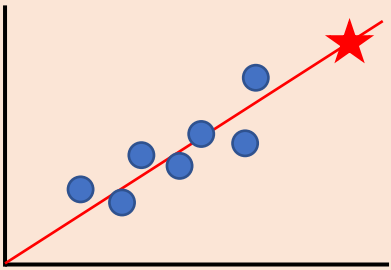
Supervised Learning

Unsupervised Learning

Common Machine Learning Algorithm Types

Supervised Learning

Regression



Unsupervised Learning

Example Machine Learning Algorithms

Supervised Learning

Regression

Linear
Regression

Classification

Decision Trees

Unsupervised Learning

Clustering

K-Means

Example Machine Learning Algorithms

Supervised Learning

Regression

**Linear
Regression**

Classification

Decision Trees

Unsupervised Learning

Clustering

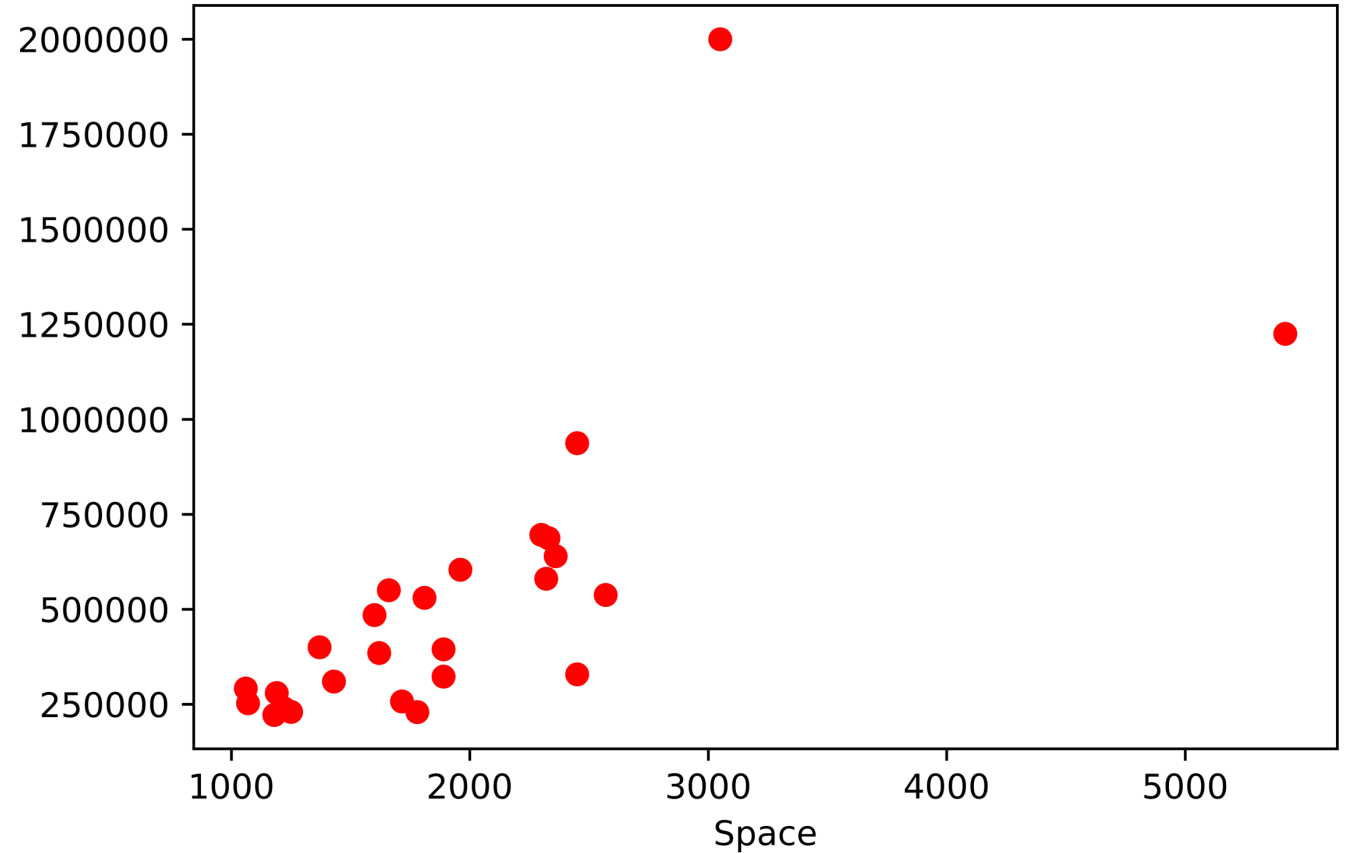
K-Means

Linear Regression

Floor Space	House Price
1 180	221 900
2 570	538 000
770	180 000
1 960	604 000
1 680	510 000
...	...
...	...
5 240	1 225 000

Feature

Label

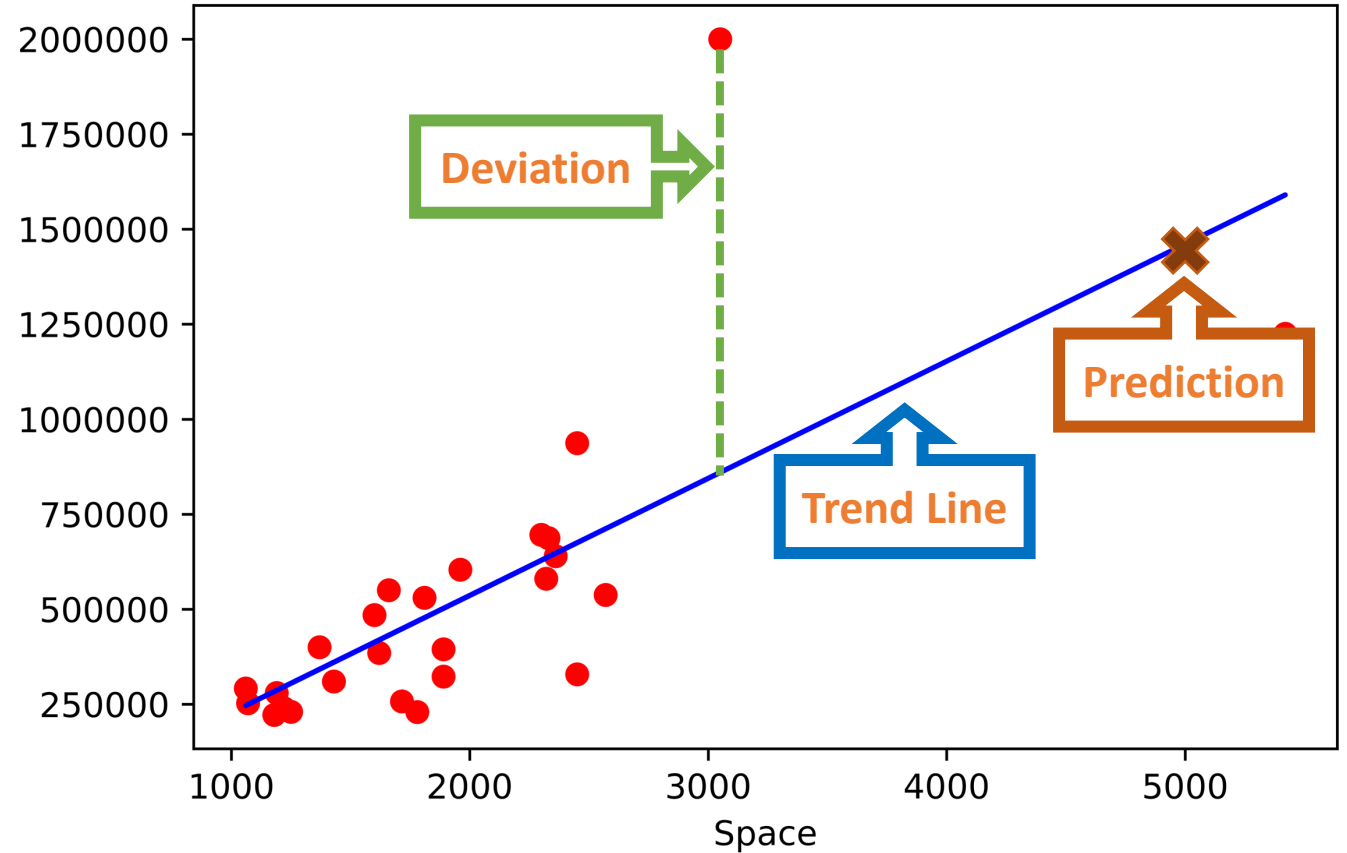


Linear Regression

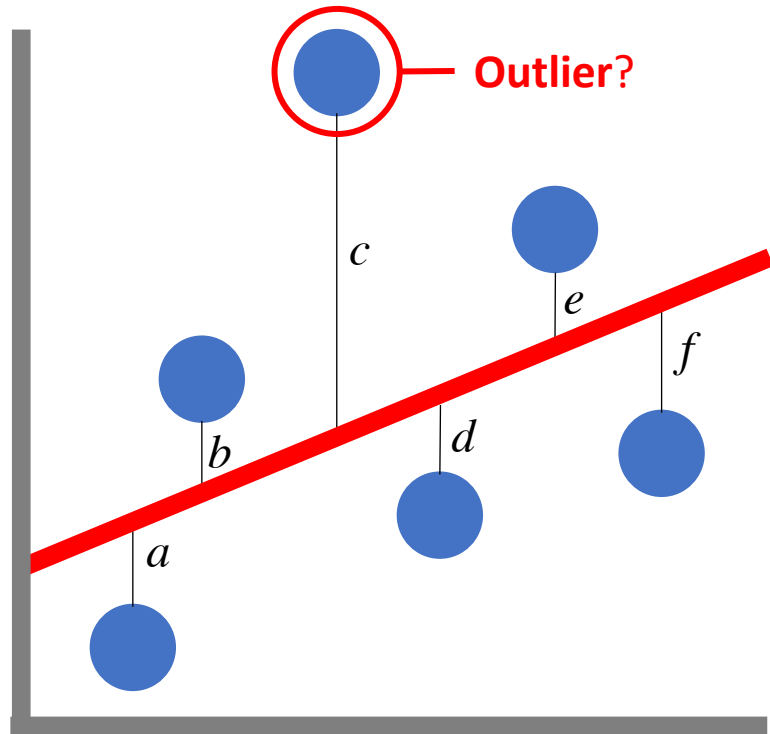
Floor Space	House Price
1 180	221 900
2 570	538 000
770	180 000
1 960	604 000
1 680	510 000
...	...
...	...
5 240	1 225 000

Feature

Label



Fitting a trend line: Ordinary Least Squares



$$a^2 + b^2 + c^2 + d^2 + e^2 + f^2 = \text{sum of squared error}$$

Linear Regression Notes

Benefits

- Simple to understand.
- Transparent.

Limitations

- Outliers skew trend line.
- Doesn't work with non-linear relationships.

Some Alternatives

- Non-linear Least Squares.
- Tree algorithms.

Example Machine Learning Algorithms

Supervised Learning

Regression

Linear
Regression

Classification

Decision Trees

Unsupervised Learning

Clustering

K-Means

Decision Tree: Calculating the Best Split

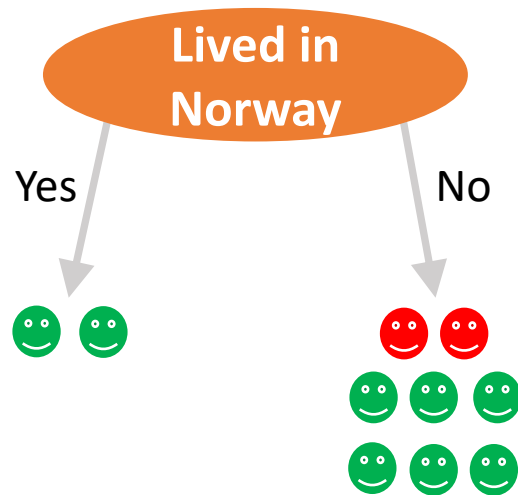
Goal: Build a Decision Tree for deciding who gets a payrise this year, based on historical payrise data.

Name	Placements	Complaints	Lived in Norway	Payrise
Don	Yes	Yes	Yes	Yes
Lewis	Yes	Yes	No	Yes
Mike	Yes	No	Yes	Yes
Danny	Yes	Yes	No	Yes
Dan	No	No	Yes	No
Elliot	Yes	No	No	Yes
Luke	Yes	No	No	Yes
Tom	Yes	Yes	No	Yes
Nathan	No	Yes	Yes	No
Owen	Yes	No	No	Yes

Features

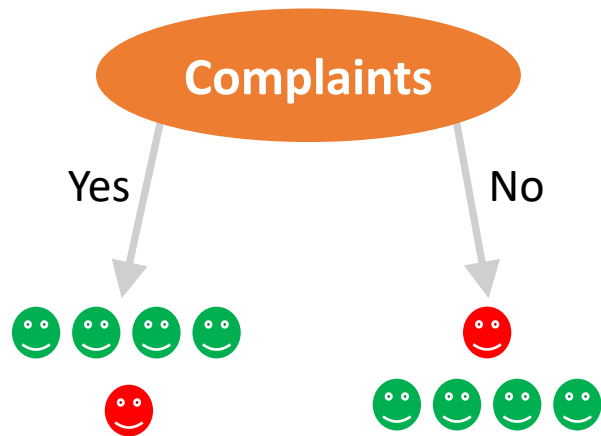
Labels

Decision Tree: Calculating the Best Split



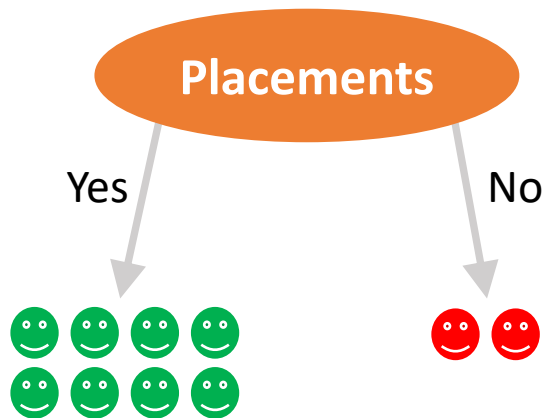
Name	Placements	Complaints	Lived in Norway	Payrise
Don	Yes	Yes	Yes	Yes
Lewis	Yes	Yes	No	Yes
Mike	Yes	No	Yes	Yes
Danny	Yes	Yes	No	Yes
Dan	No	No	Yes	No
Elliot	Yes	No	No	Yes
Luke	Yes	No	No	Yes
Tom	Yes	Yes	No	Yes
Nathan	No	Yes	Yes	No
Owen	Yes	No	No	Yes

Decision Tree: Calculating the Best Split



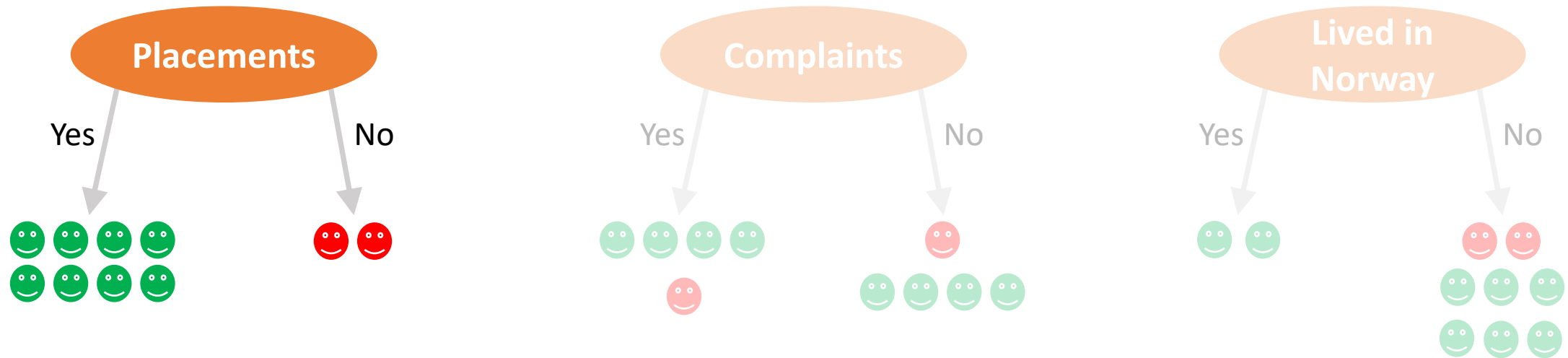
Name	Placements	Complaints	Lived in Norway	Payrise
Don	Yes	Yes	Yes	Yes
Lewis	Yes	Yes	No	Yes
Mike	Yes	No	Yes	Yes
Danny	Yes	Yes	No	Yes
Dan	No	No	Yes	No
Elliot	Yes	No	No	Yes
Luke	Yes	No	No	Yes
Tom	Yes	Yes	No	Yes
Nathan	No	Yes	Yes	No
Owen	Yes	No	No	Yes

Decision Tree: Calculating the Best Split



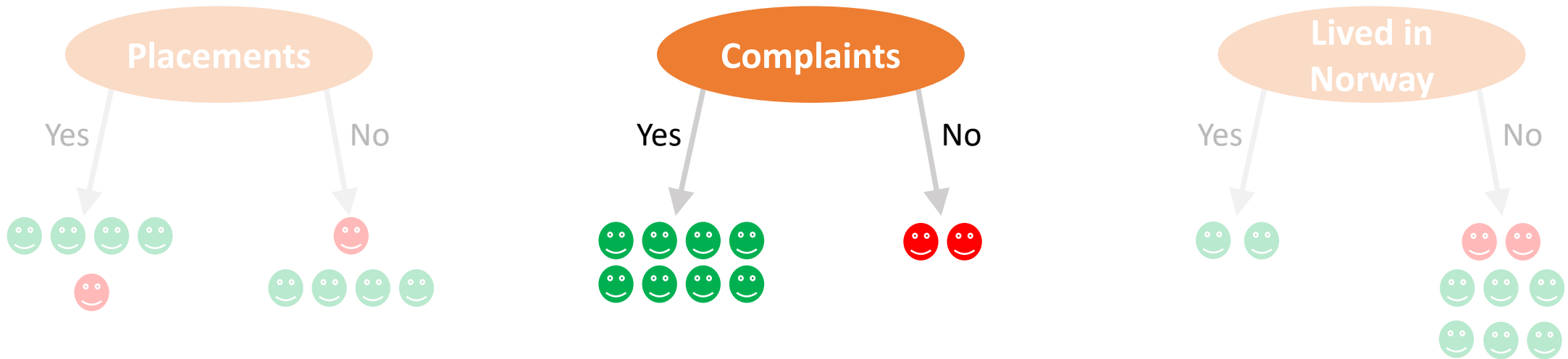
Name	Placements	Complaints	Lived in Norway	Payrise
Don	Yes	Yes	Yes	Yes
Lewis	Yes	Yes	No	Yes
Mike	Yes	No	Yes	Yes
Danny	Yes	Yes	No	Yes
Dan	No	No	Yes	No
Elliot	Yes	No	No	Yes
Luke	Yes	No	No	Yes
Tom	Yes	Yes	No	Yes
Nathan	No	Yes	Yes	No
Owen	Yes	No	No	Yes

Decision Tree: Calculating the Best Split



Recruiters	Placements	Complaints	Lived in Norway	Payrise
8	8	4	2	Yes
2	0	1	2	No

Bad Data Leads to a Bad Model



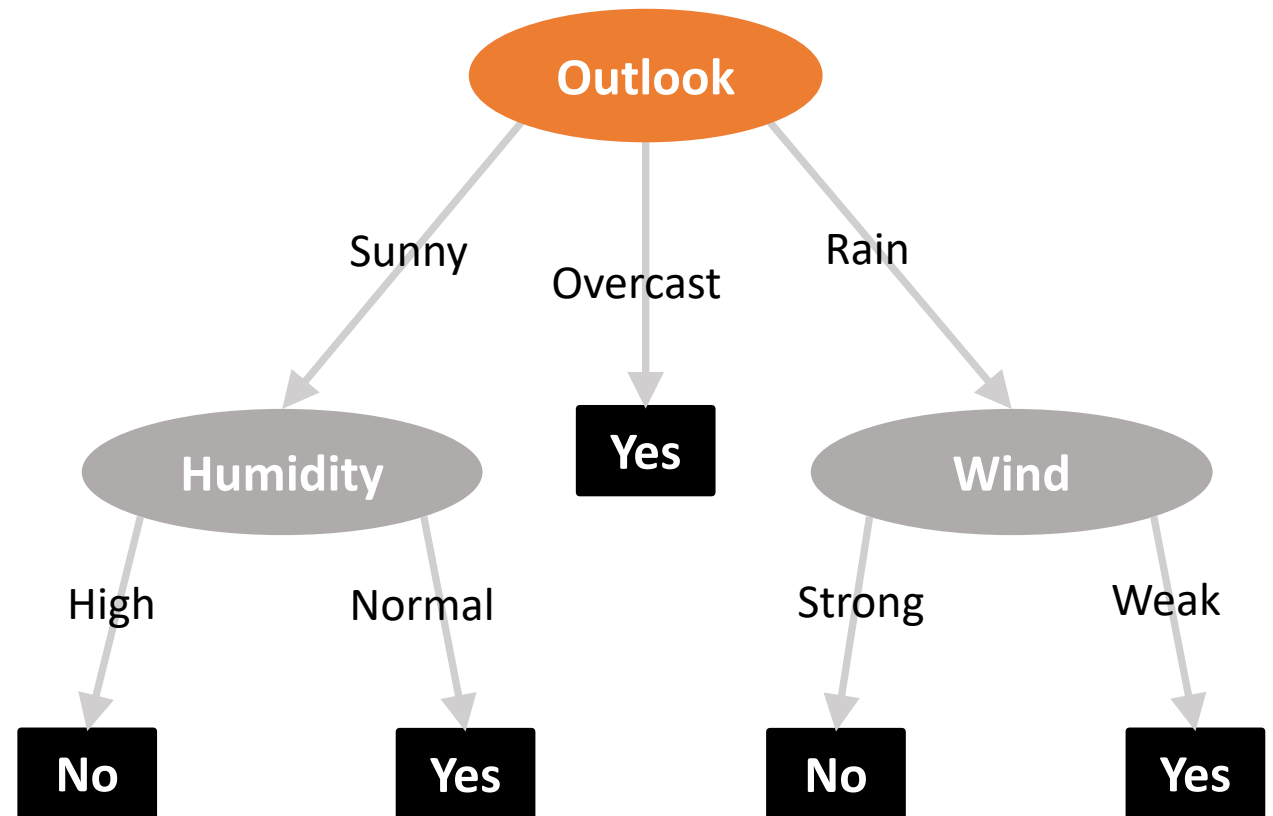
Recruiters	Placements	Complaints	Lived in Norway	Payrise
8	7	8	2	Yes
2	1	0	2	No

Decision Tree: Recursive Partitioning

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
...	...	...	...	...
...	...	...	...	...
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Features

Labels



Decision Tree Notes

Benefits	Limitations	Some Alternatives
• White Box. • Flexible (use for both regression and classification). • Robust to outliers. • Handle non-linear boundaries.	• Susceptible to overfitting. • Changes to where the Data is sliced can produce different results.	• Support Vector Machine. • Logistic Regression. • Random Forests.

Example Machine Learning Algorithms

Supervised Learning

Regression

Linear
Regression

Classification

Decision Trees

Unsupervised Learning

Clustering

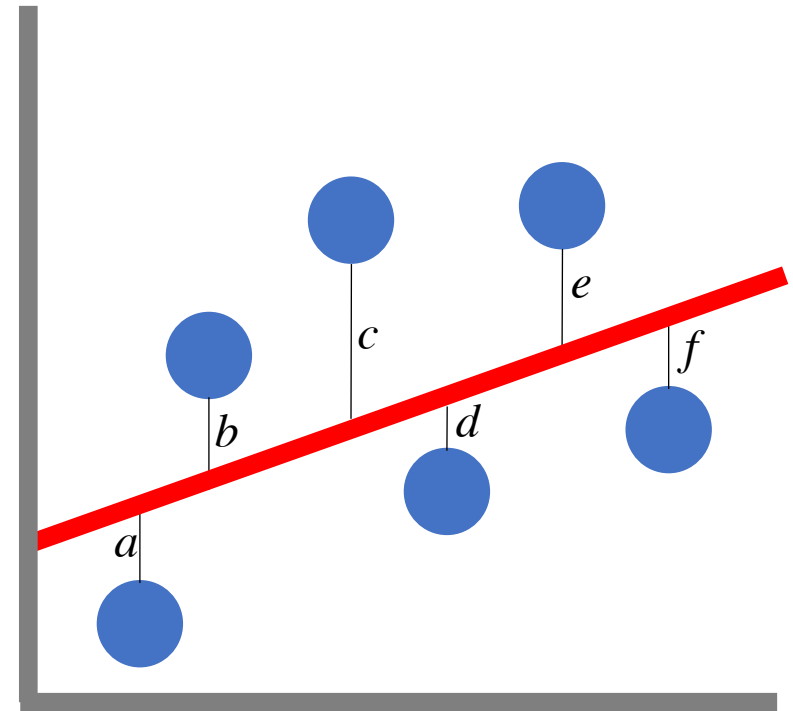
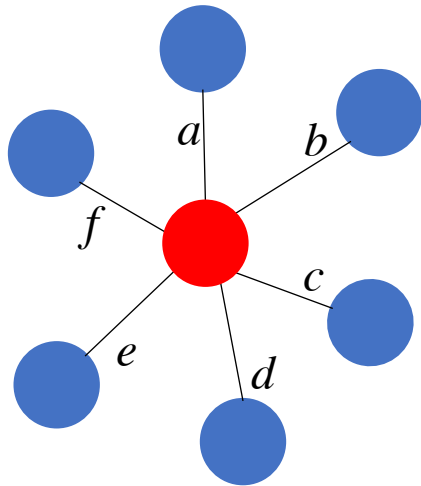
K-Means

K-Means Clustering



- K = The amount of clusters the algorithm will try to find.
- K = Should be large enough to extract meaningful patterns but small enough that clusters remain clearly distinct.
- **So how do we calculate K?**

Sum of Squared Errors



$$a^2 + b^2 + c^2 + d^2 + e^2 + f^2 = \text{sum of squared error}$$

K-Means: Calculating the K value



- **Scree Plots** allow us to find optimal number of clusters.
- Shows the **Sum of Squared Errors** for different numbers of clusters.
- The optimal K value is at the “Elbow” of the plot.

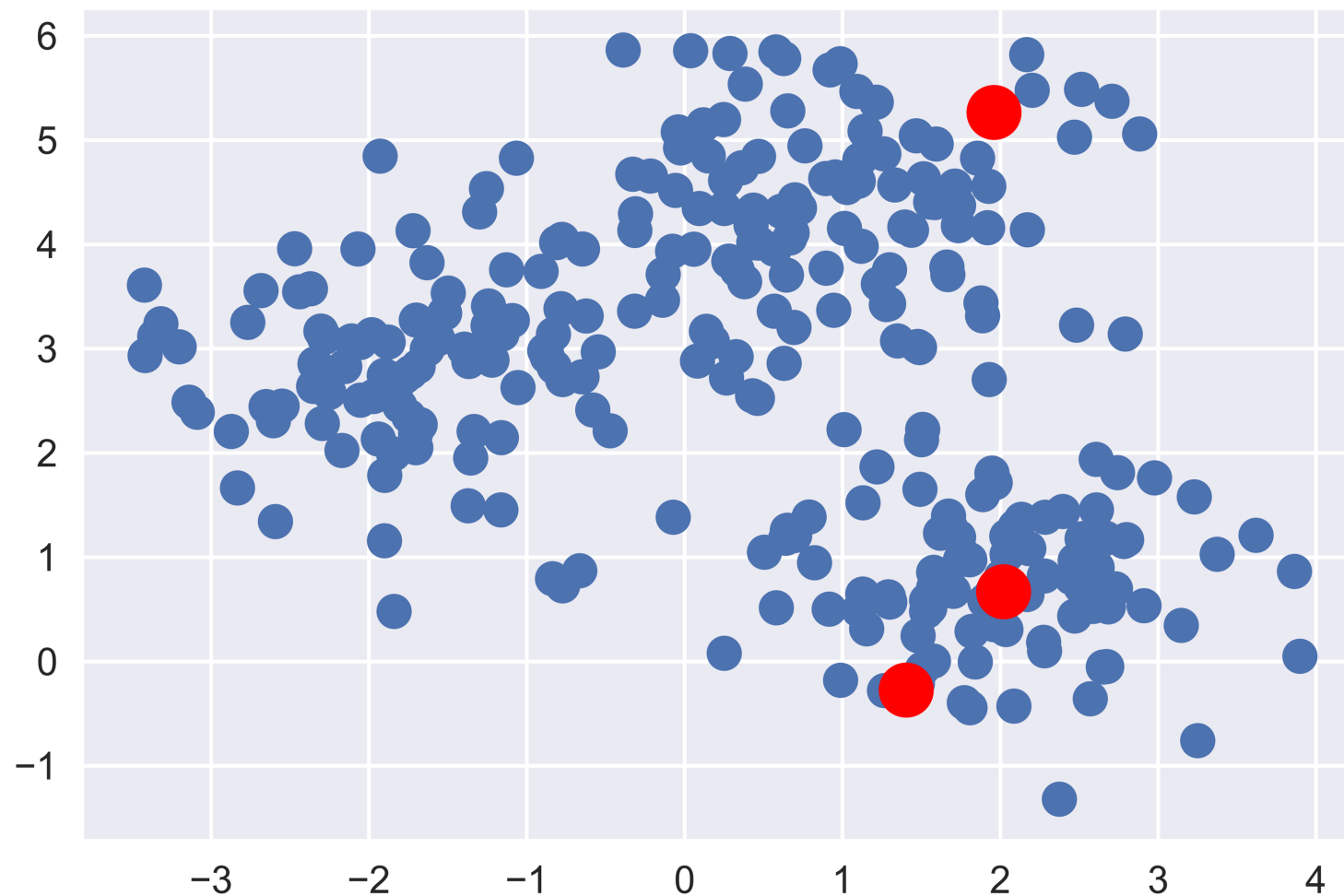
K-Means Demo

Randomly allocate centroids



K-Means Demo

Randomly allocate centroids



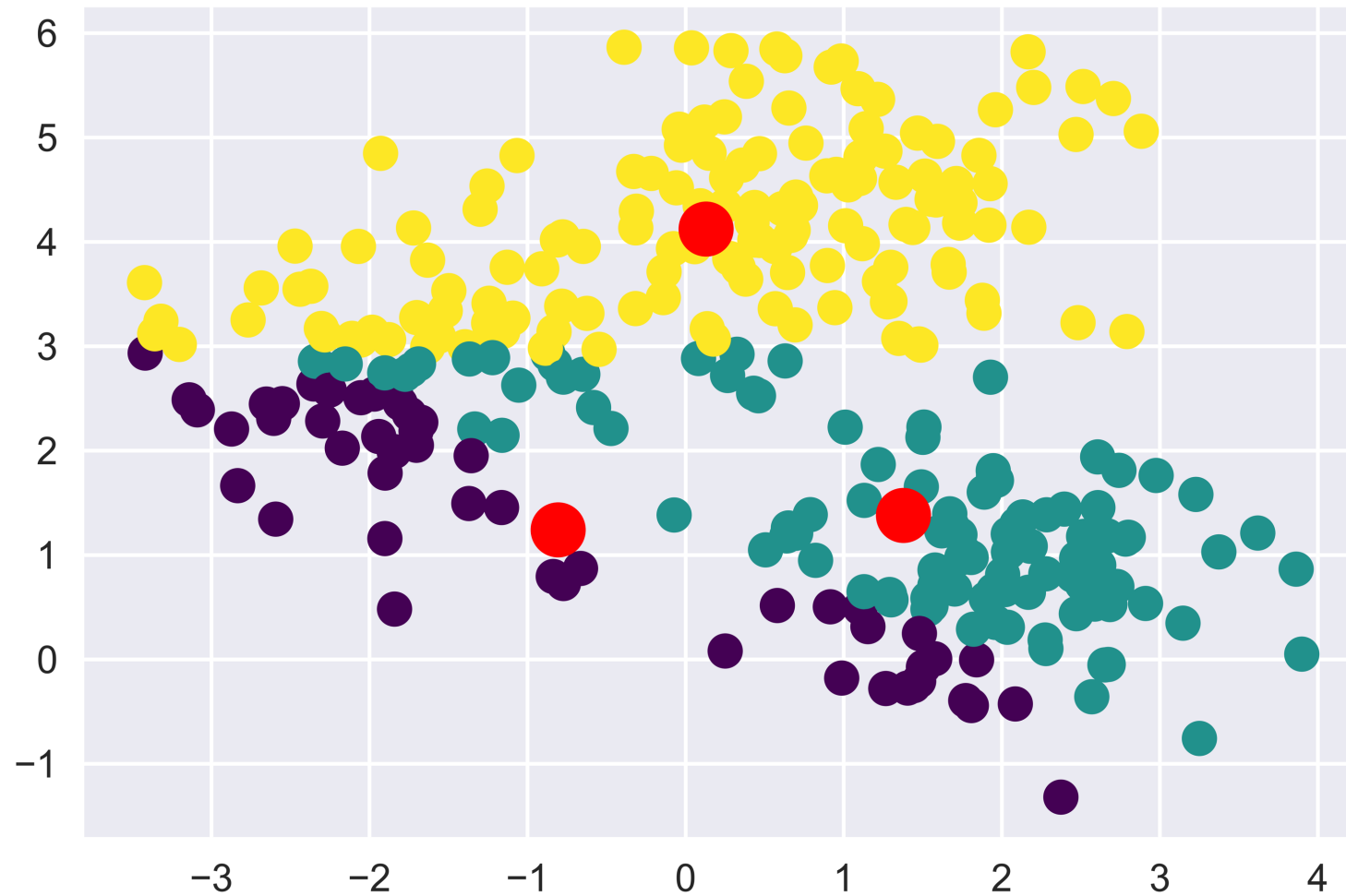
K-Means Demo

Iteration 1: Calculate cluster membership based on nearest centroid



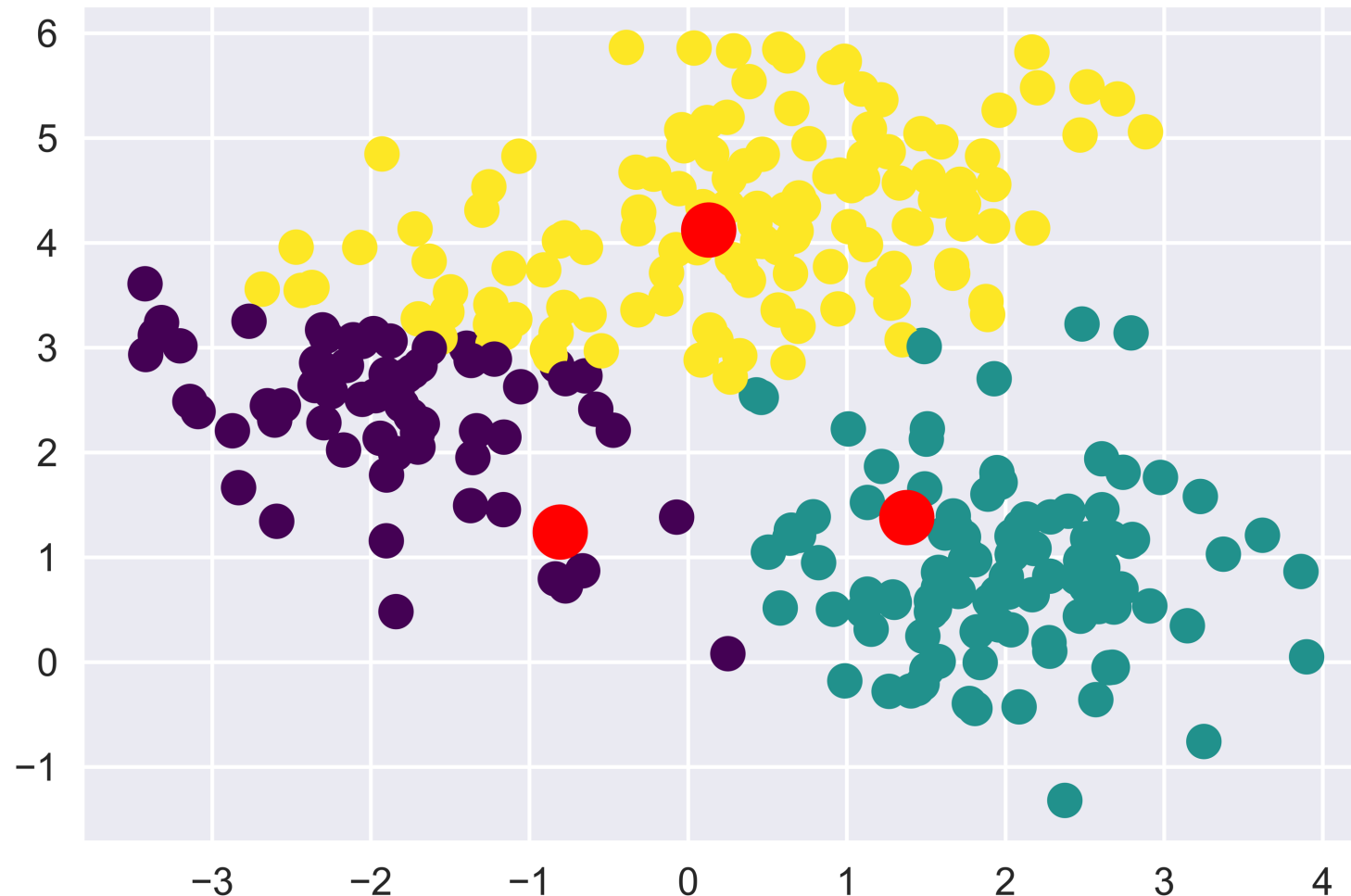
K-Means Demo

Iteration 1: Move centroids to the center of their cluster



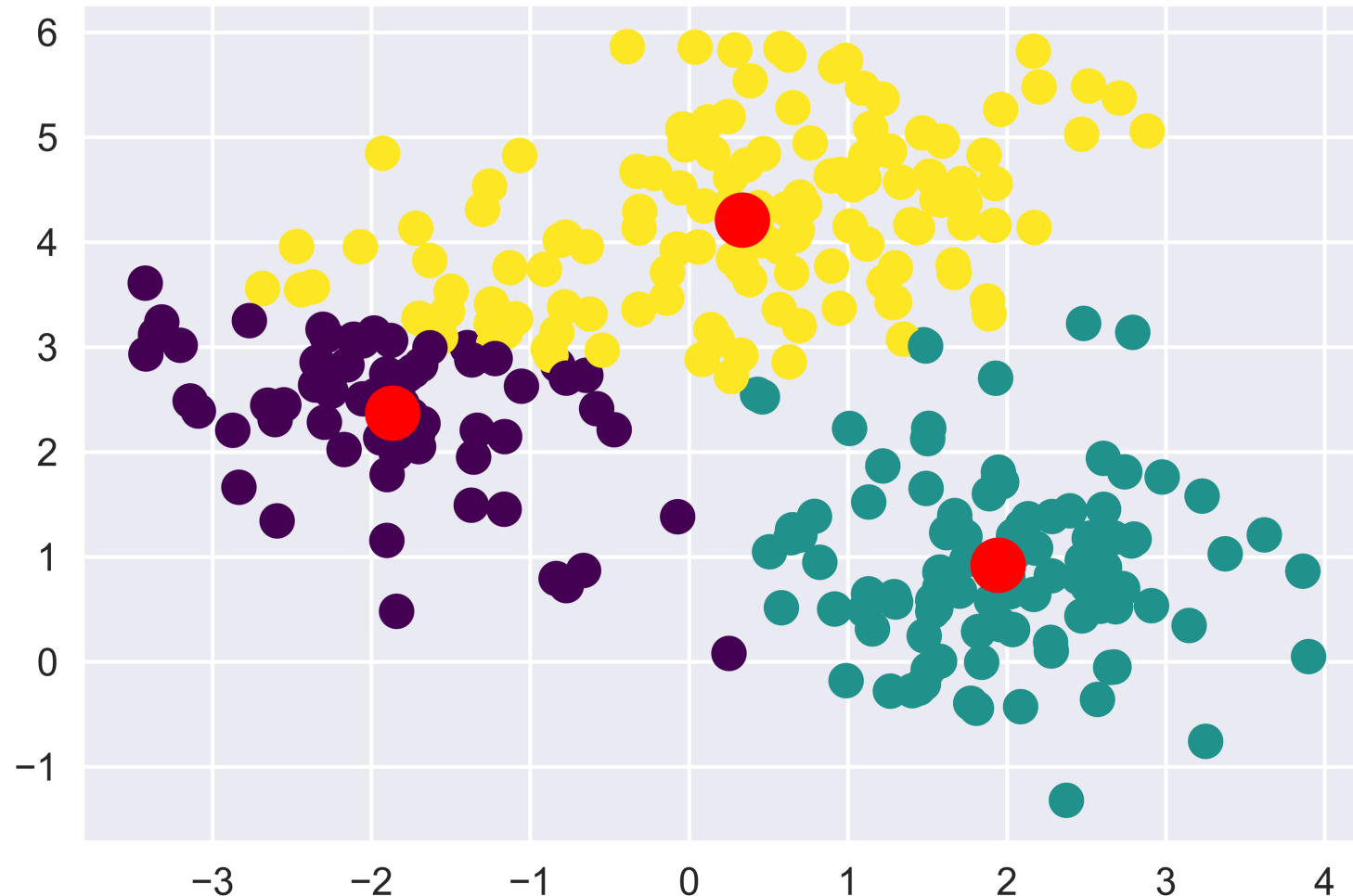
K-Means Demo

Iteration 2: Recalculate cluster membership based on nearest centroid



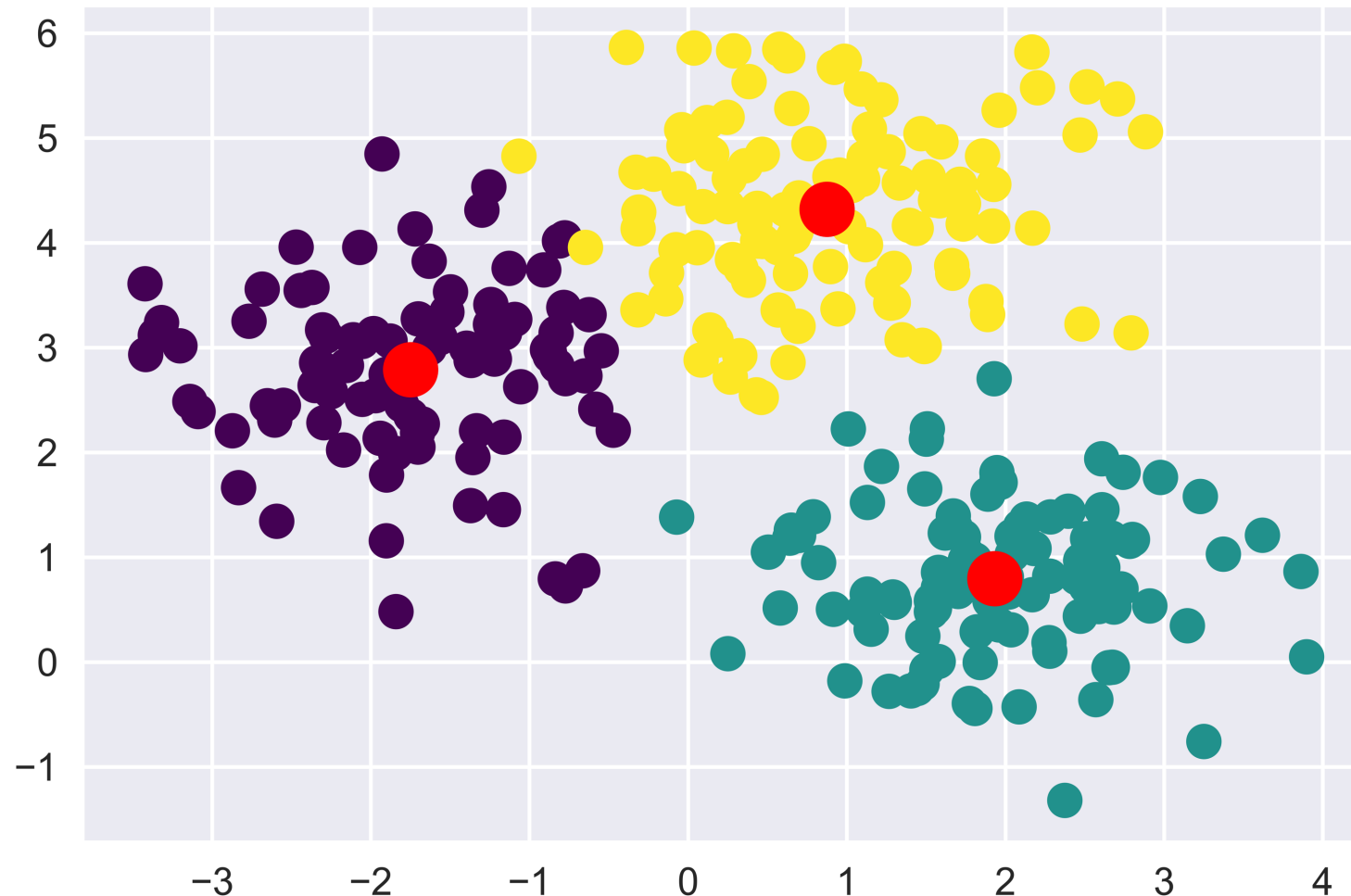
K-Means Demo

Iteration 2: Move centroids to the center of their cluster



K-Means Demo

After 6 iterations: Clusters and centroids stabilise, algorithm stops



K-Means Clustering Notes

Benefits

- Fast and highly effective at uncovering basic data patterns.
- Works best for spherical, non-overlapping clusters.

Limitations

- Each data point can only be assigned to one cluster.
- Clusters are assumed to be spherical.

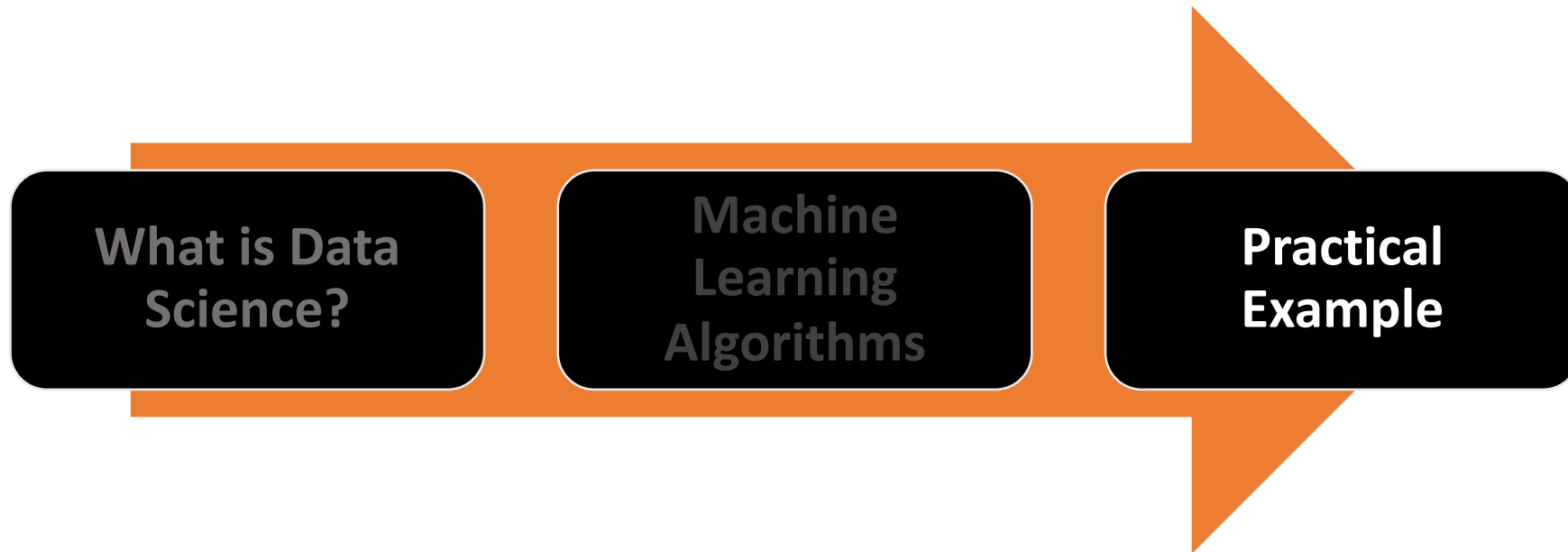
Some Alternatives

- Gaussian mixtures.
- Fuzzy K-Means.

Machine Learning Algorithms: Key Takeaways

- The three main types of Machine Learning are **Supervised**, **Unsupervised** and **Reinforcement** Learning.
- **Machine Learning is more than Neural Networks** and Deep Learning.
- A successful Machine Learning Model needs to find the balance between **Overfitting** and **Underfitting**.
- **Machine Learning Algorithms are merely tools**. Good results come from understanding how they work and tuning them correctly.

Practical Example



Use Case: Titanic Passenger Survival



Goal: Build a classification model for predicting Titanic survivability.

Hypothesis

That it is possible
to predict Titanic
survivability based
on **Age, Gender**
and **Ticket Class**.



Variable	Description
PassengerId	Unique Identifier
Survival	Survived = 1, Died = 0
Pclass	Ticket class (1, 2 or 3)
Sex	Gender ('male' or 'female')
Age	Age in years
Sibsp	Number siblings / spouses aboard the Titanic
Parch	Number parents / children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation
Name	Passenger name, including honorific.

Titanic Dataset

Tools



Practical Example: Key Takeaways

- **Scikit-learn** and **Jupyter Notebooks** provide a free and flexible basis for starting with Data Science. Use the **Anaconda** distribution to save time on installation!
- **Feature Engineering** is a vital skill for Data Scientists.
- **Domain Knowledge** is key to succeed!
- **Split your data** into **Test** and **Training** sets.
- **Tweaking Hyperparameters** may give better results (but you should be able to explain how your tweak improved model performance).

Tips for Getting Started with Data Science

- Become a Data Engineer!
- Learn Python or R (SQL is also useful)!
- Learn some statistical methods!
- Understand the Data Science process!
- Practice with Kaggle!

Jupyter

Titanic Demo

Last Checkpoint: 27/11/2019 (autosaved)

File

Edit

View

Insert

Cell

Kernel

Widgets

Help

Trusted

Python 3

+

Run

Code

Slide Type

Slide

Feature Engineering 2 : Fix "Sex" Categorical Variable

According to our hypothesis, Sex is also an indicator of survivability. However in our Dataset it is represented as string values, which scikit-learn doesn't like. So we'll need to convert it to a numerical value.

In [10]:

Slide Type

Fragment

```
gendermap = {"male": 0, "female": 1}
titanic_df["Sex"] = titanic_df["Sex"].map(gendermap)
titanic_df.head(5)
```

Out[10]:

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

Slide Type

Slide

Does the Data Reflect Our Hypothesis?

Here we'll use Matplotlib to get visual insight into our Data.

First, how does survivability relate to Class?

In [11]:

Slide Type

Fragment

```
sns.countplot(x='Pclass', hue='Survived', data=titanic_df)
```

Out[11]:

<matplotlib.axes._subplots.AxesSubplot at 0x1a22c1a940>

Pclass	Survived 0	Survived 1
1	80	96
2	97	80
3	365	120

[https://github.com/
markwest1972/titanic](https://github.com/markwest1972/titanic)

 @markawest

Thanks for listening!