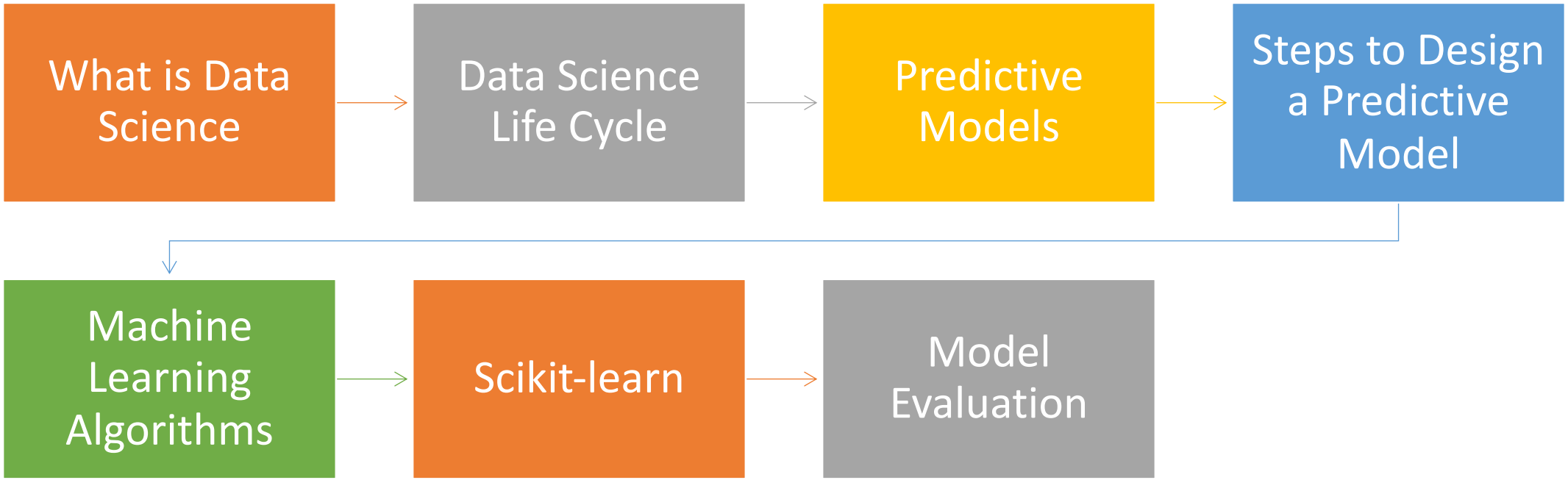
Introduction to Machine Learning

The Center of Applied Data Science

Outline


What is Data Science
Data Science Life Cycle
Predictive Models
Steps to Design a Predictive Model
Machine Learning Algorithms
Scikit-learn
Model Evaluation

Data

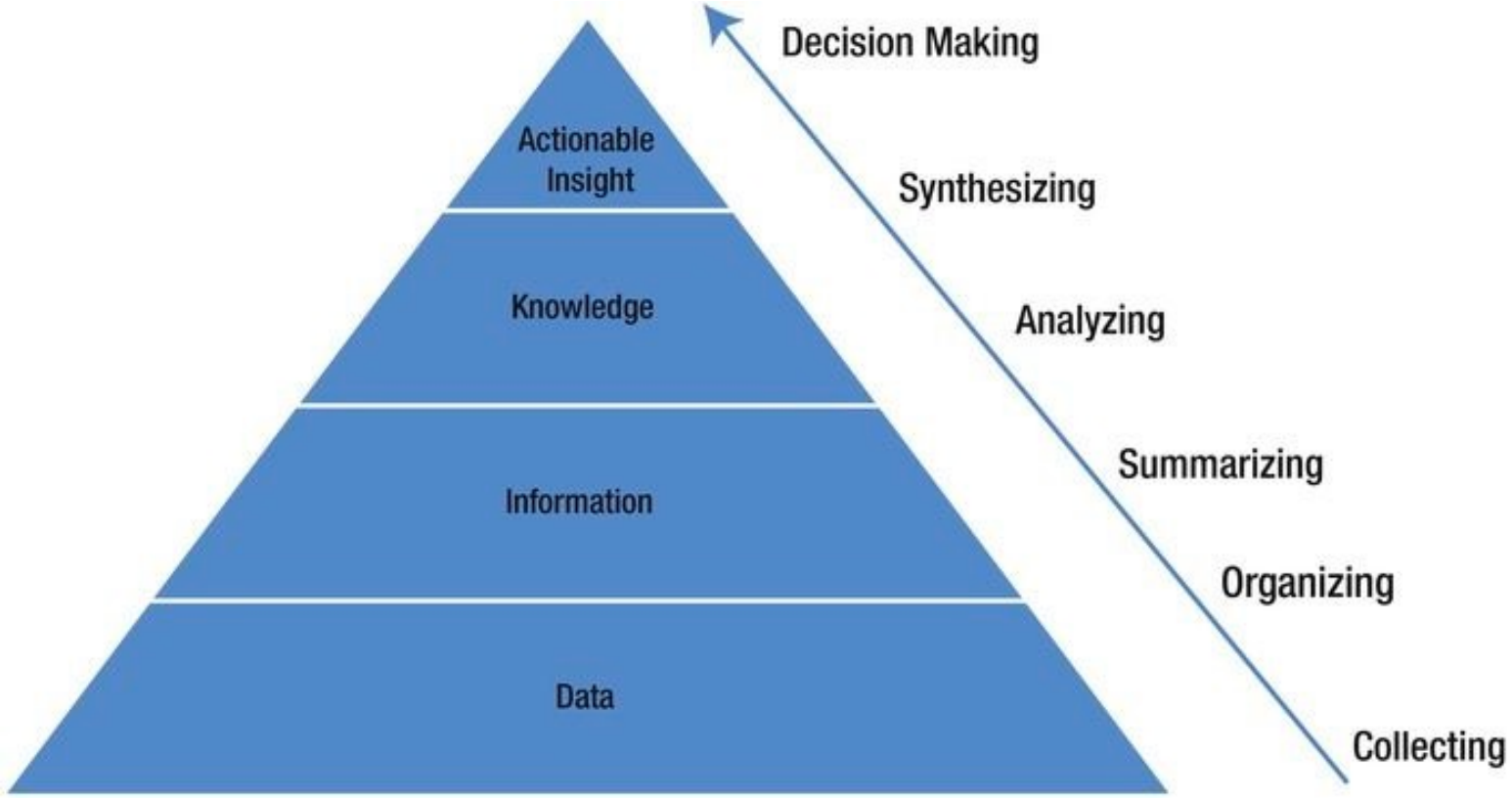
Facts and statistics collected together for reference or analysis.

Science

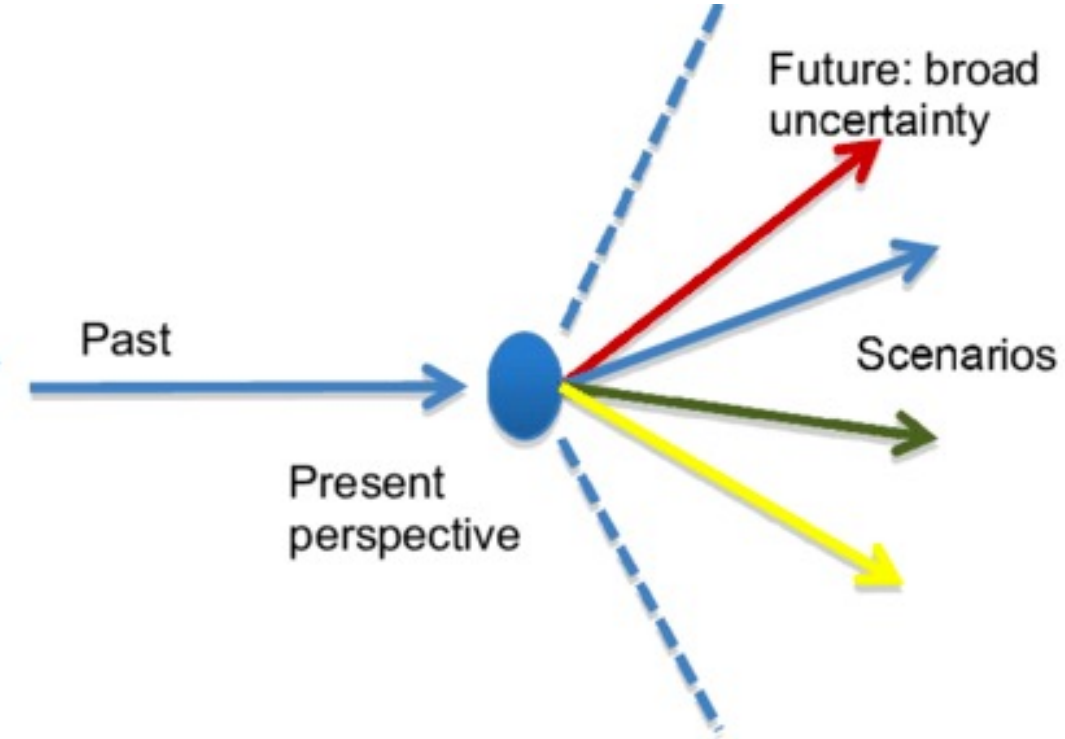
Systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.

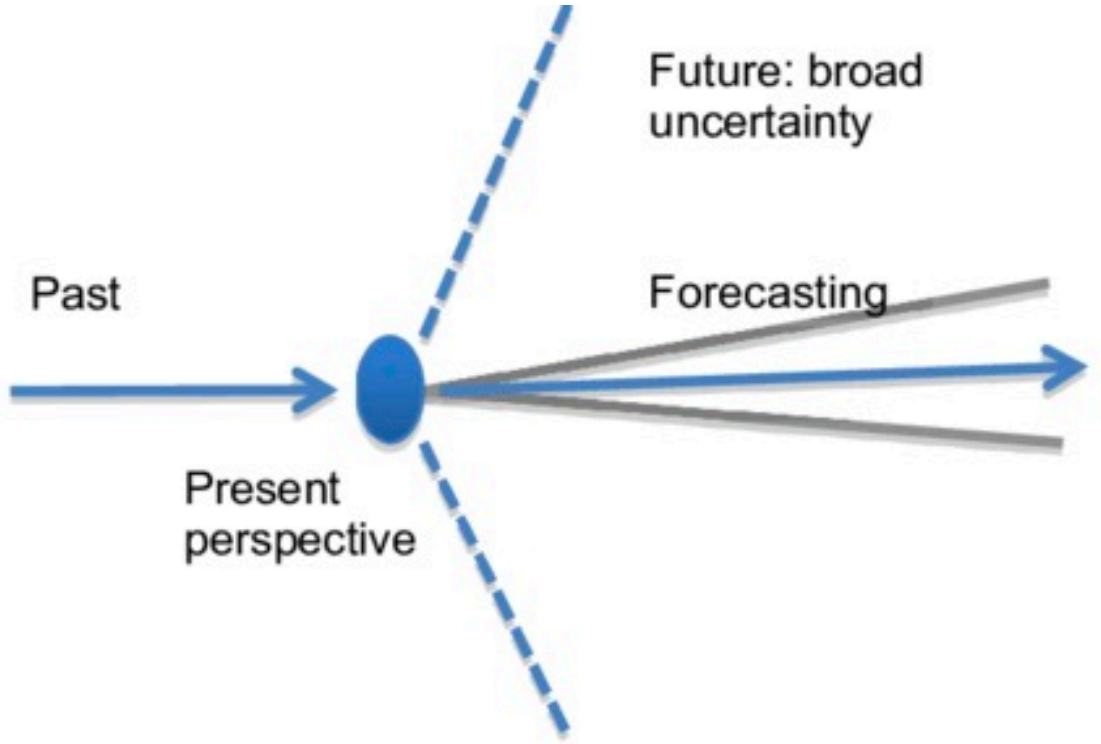
The Center of Applied Data Science

Data Science



The Center of Applied Data Science





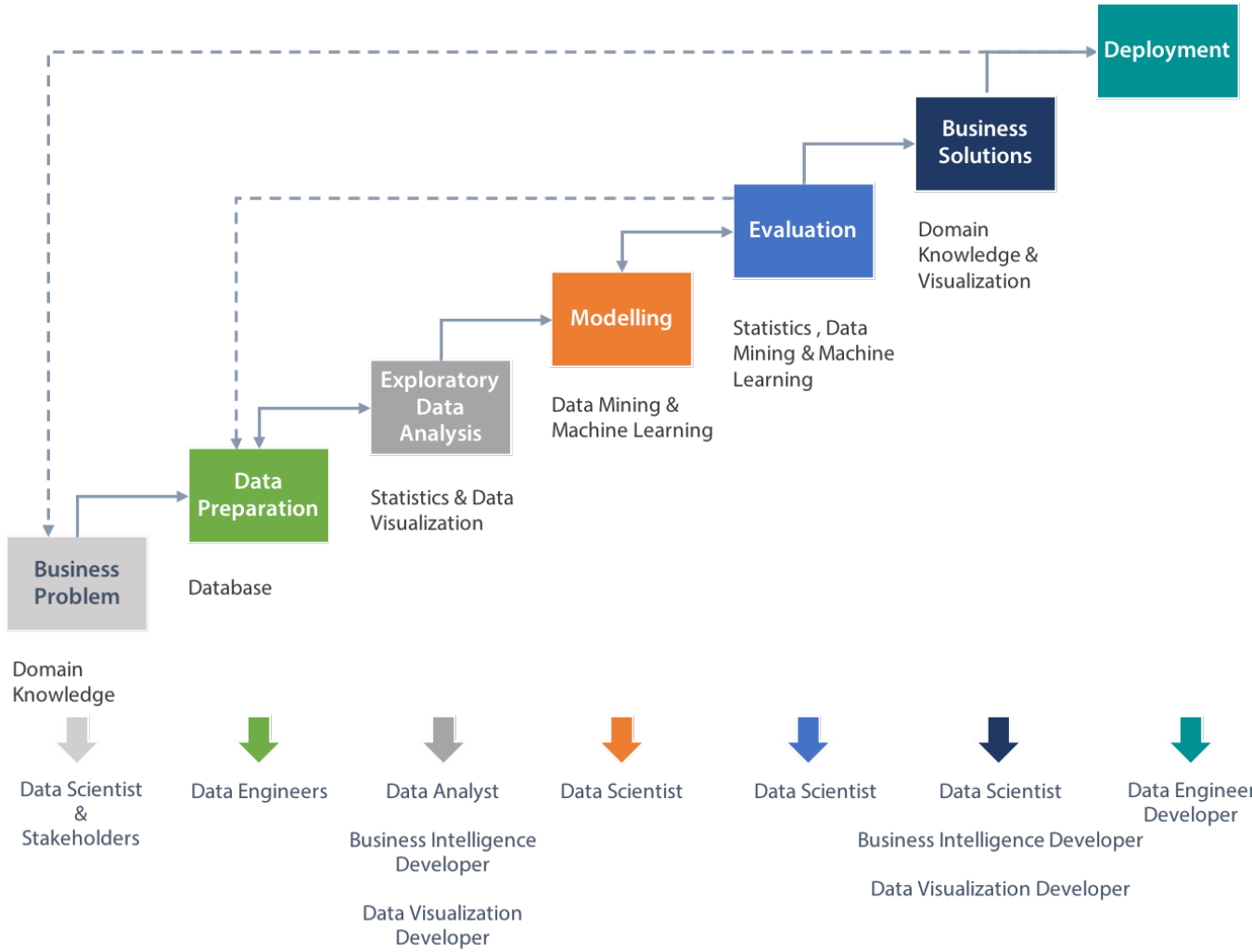
Data Science: Uncertainty

The Center of Applied Data Science

Data Science: Questions

What happened?	Descriptive Analytics
Why did something happen?	Diagnostic Analytics
What will happen next?	Predictive Analytics
What's the best decision now?	Prescriptive Analytics

Data Science Life Cycle

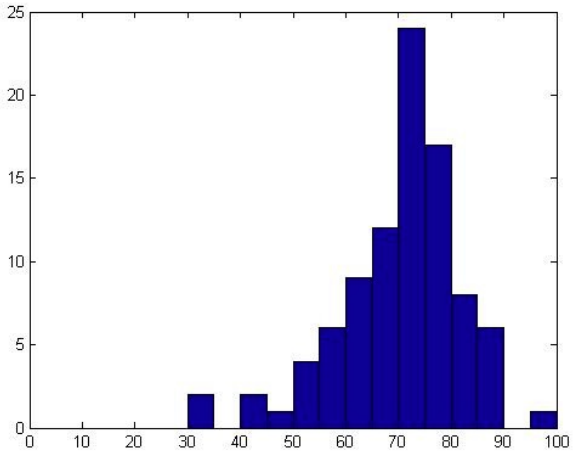


Steps to Design a Predictive Model

The Center of Applied Data Science

Step1: Business Problem
Who, What, When, Where, Why

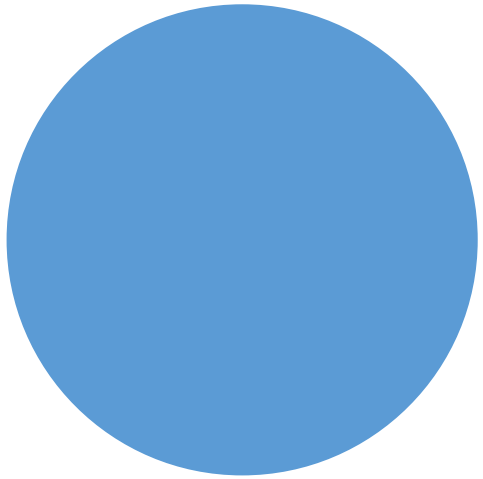
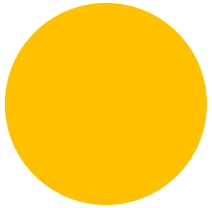
- Who has the best performance in our company?
- What products or services sell better?
- When should the company launch its next product?
- Where does the organization have big wins?
- Why are sales down despite many advertisement channels?



	A	B	C	D	E
1	StudentID	Homework	Midterm	Project	Final
2	4560	100	97	100	95
3	5540	85	90	88	90
4	6889	92	85	88	87
5	6817	65	85	87	89

Step2: Data Preparation
Import, Clean, Explore, and Visualize Data

The Center of Applied Data Science

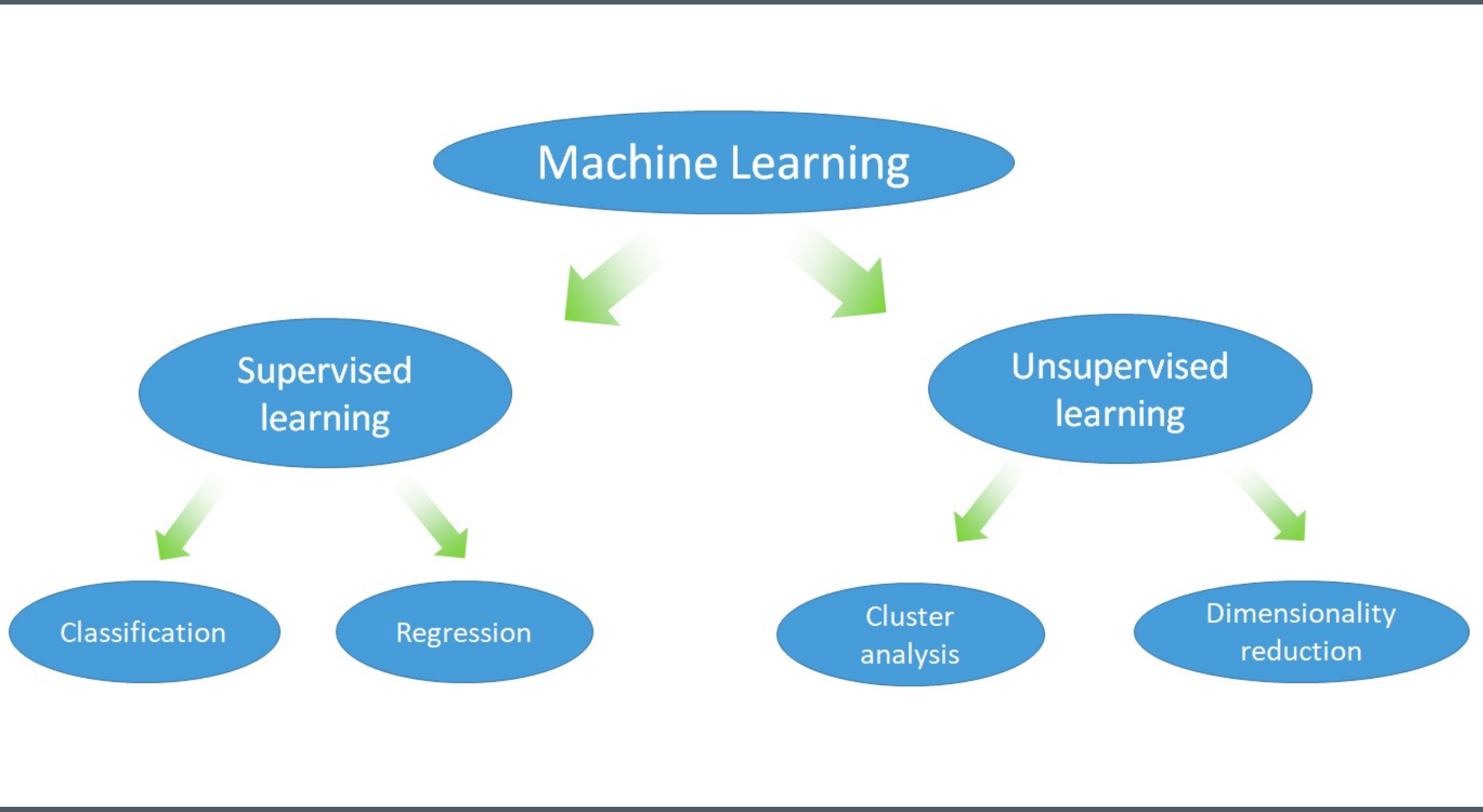
Step3: Build and Evaluate Machine Learning Model

Step4: Interpret the results regarding the question

Machine Learning

Machine learning is the study of using a computer to automatically find patterns in data

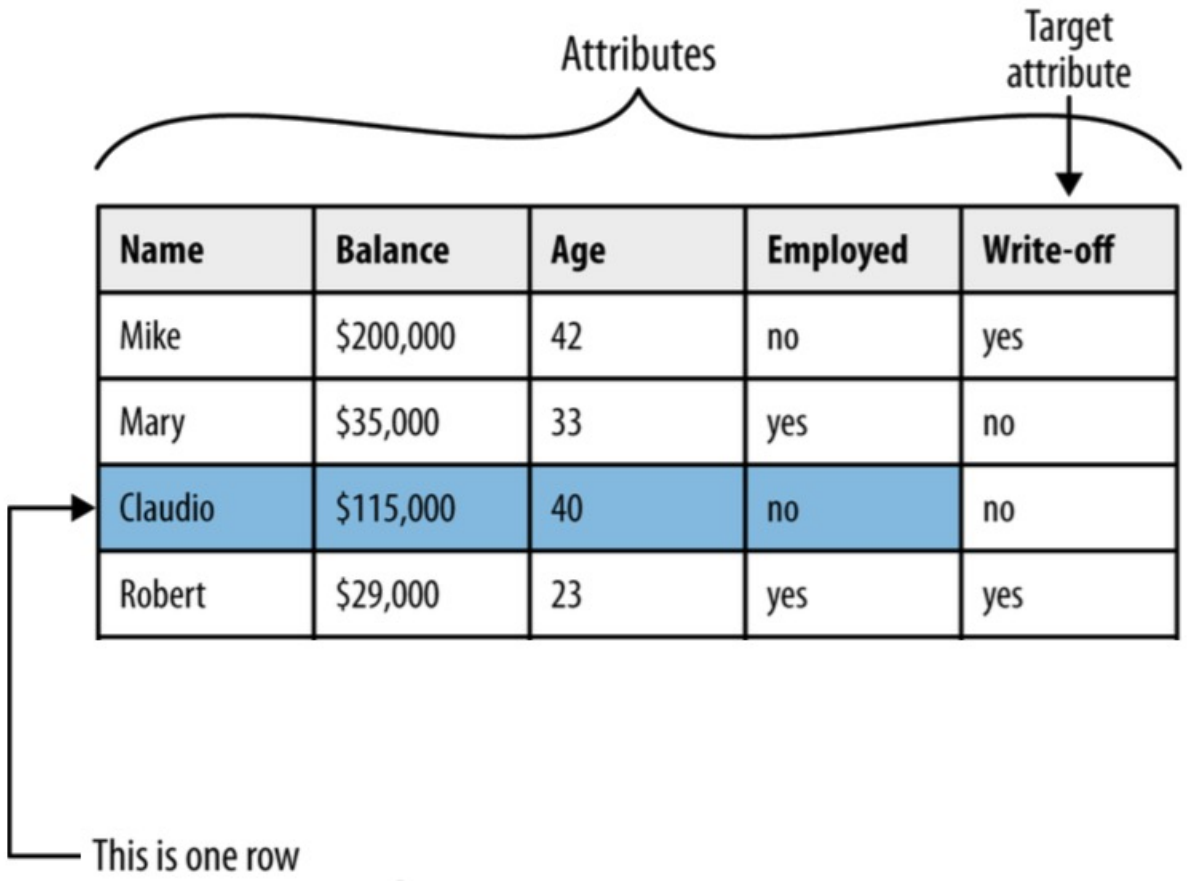
Obvious patterns are already identified by domain experts, so the real advantage of machine learning is to extract non-intuitive patterns



http://www.diegocalvo.es/en/machine-learning-supervised-unsupervised/

Terminology

- The following terms are used interchangeably:
  - **Row**, observation, datapoint
  - Column, **attribute**, feature
  - **Target attribute**, target, label
- A target attribute is one we are interested in predicting
- To make this prediction we build a model
  - Models are the fundamental building blocks of data science



Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes

Two stages of modeling
Fit and predict

- Fit
  - Using available data (including targets), a set of rules is created
- Predict
  - For a new, unobserved datapoint (target unknown), the set of rules is applied

Returning to example

Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	?

New row

We would like for a model to predict this?

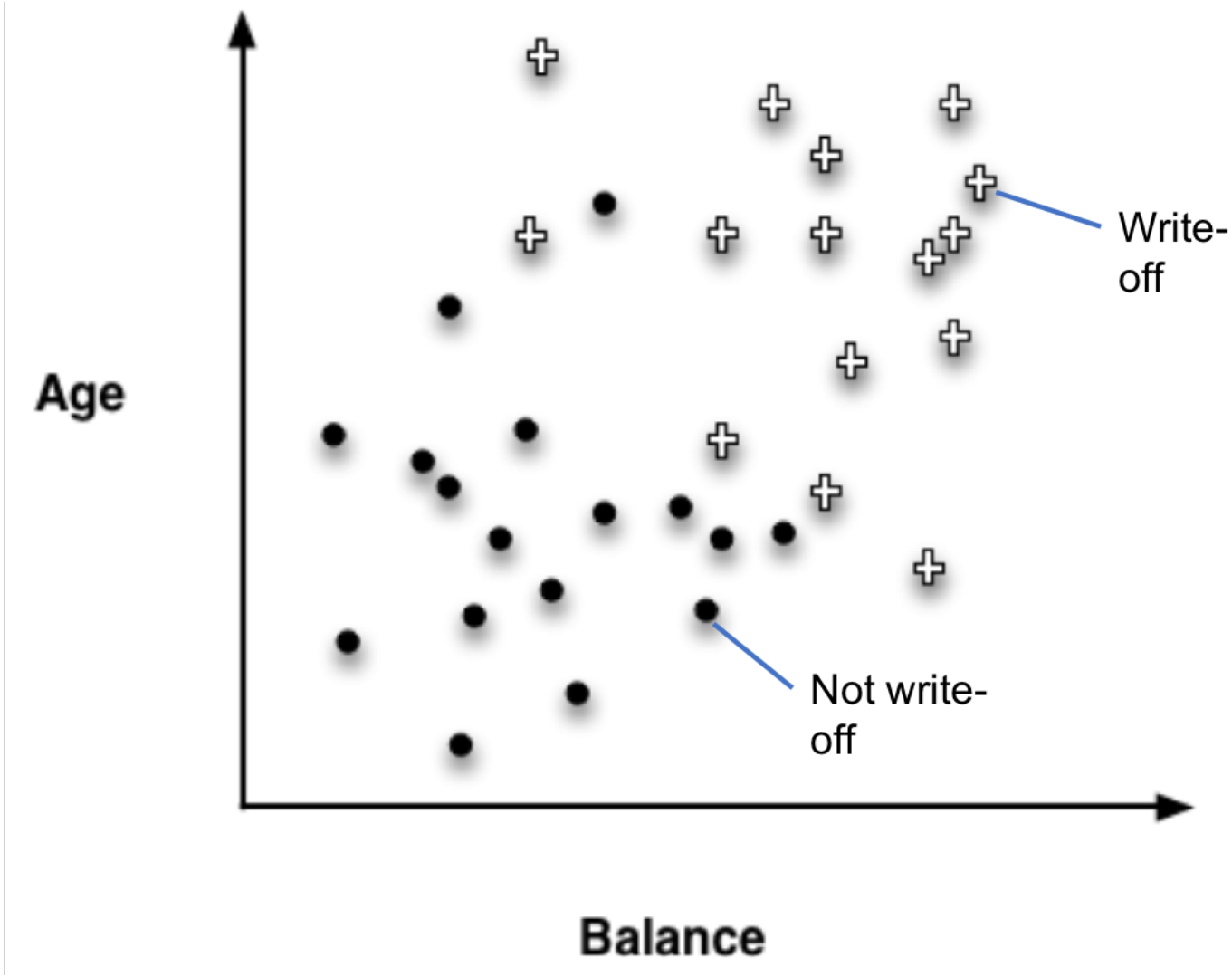
The Center of Applied Data Science

Classification

- Classification is used when the target attribute is a category
- Model is built based on target attributes of historical data
- Examples:
  - Predicting churn
  - Predicting write-off

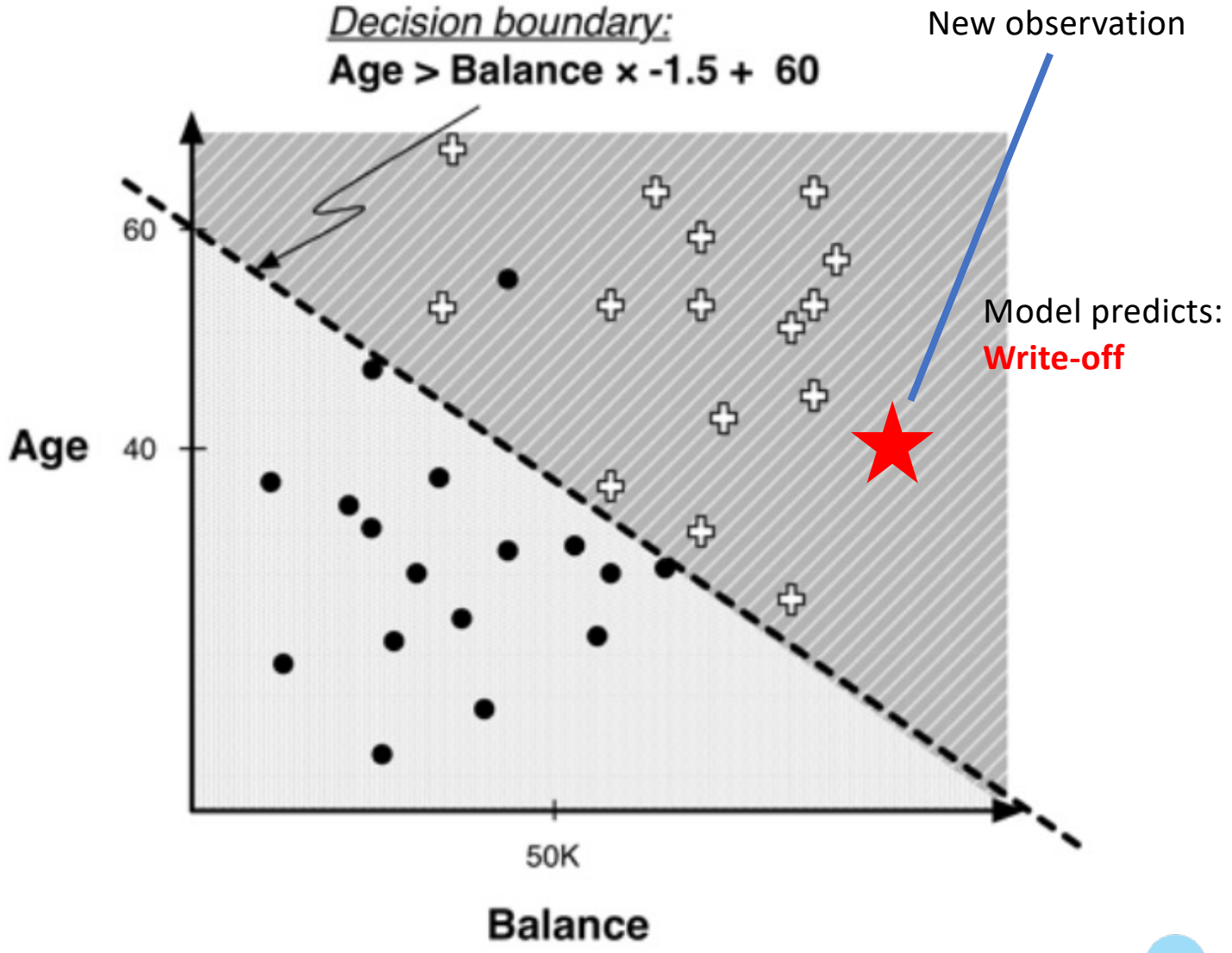
The Center of Applied Data Science

Visualization of classification algorithms



The Center of Applied Data Science

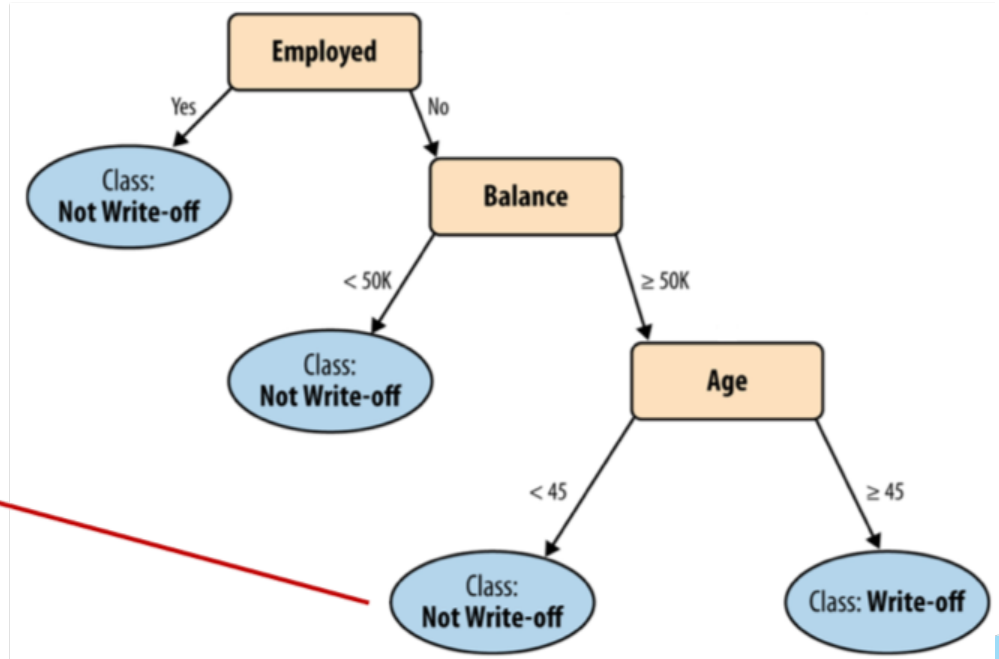
Visualization of classification algorithms: Linear classifier



Decision Tree
One type of model

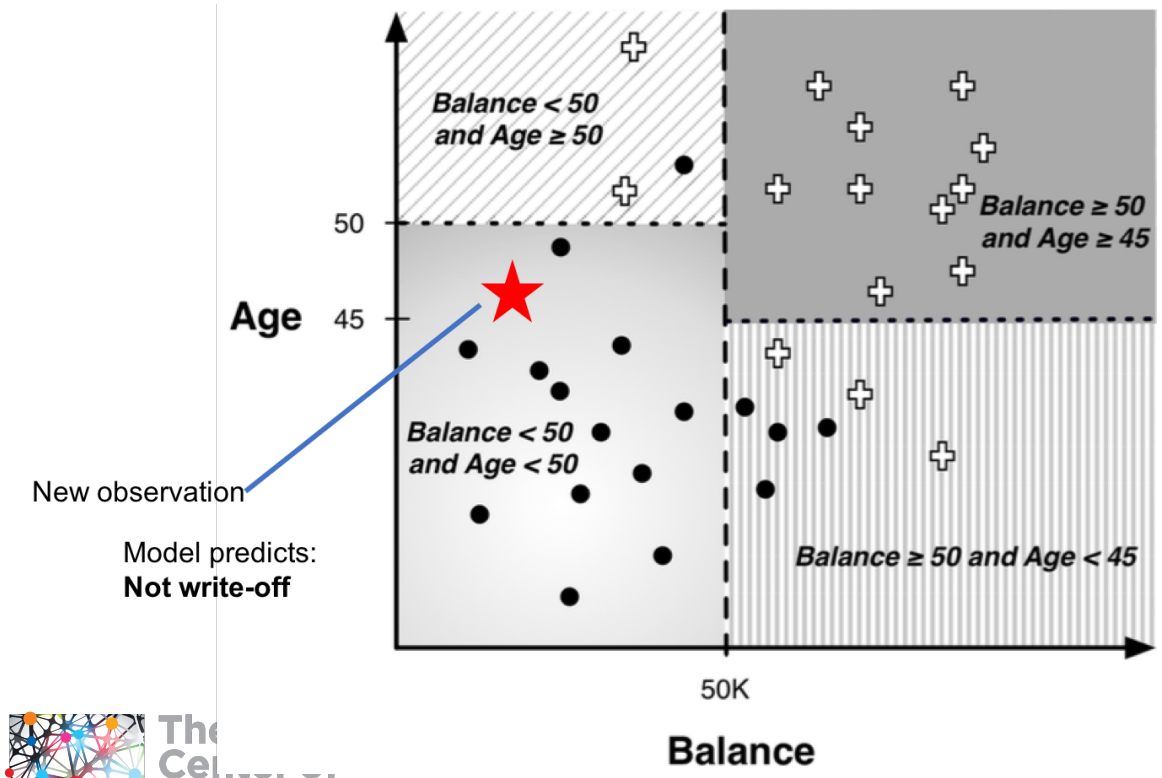
Name	Balance	Age	Employed	Write-off
Mike	$200,000	46	No	Yes
Mary	$335000	33	Yes	Yes
Claudio	$115,000	40	No	No
Robert	$29,000	23	Yes	Yes
Dora	$72,000	31	No	?

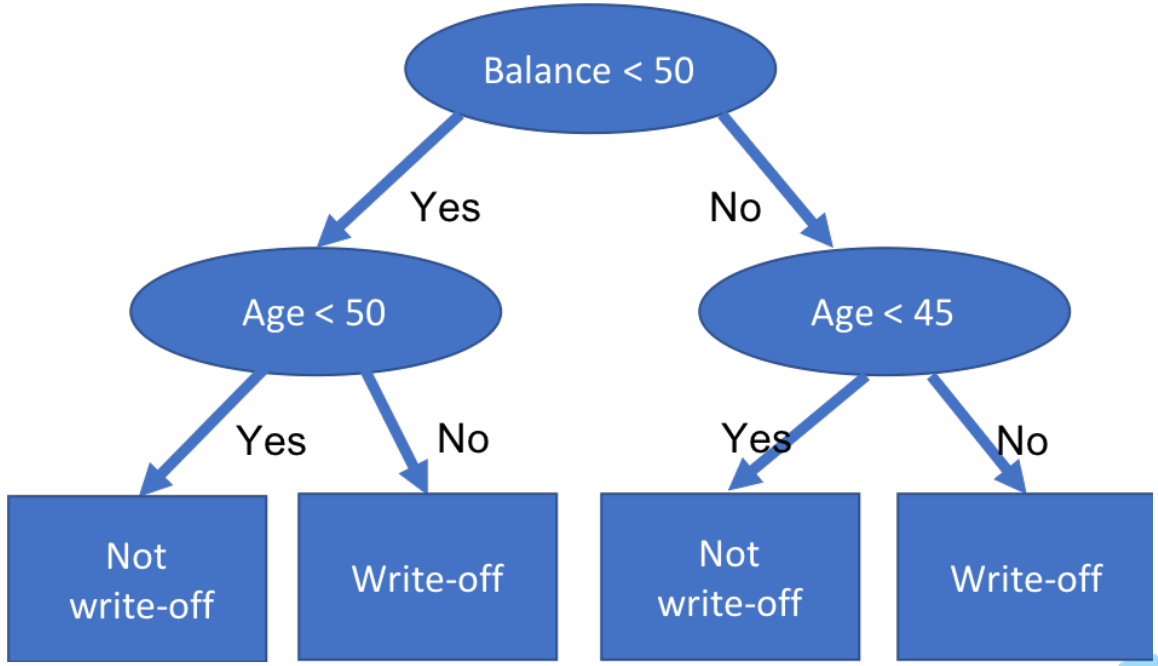
No



The Center of Applied Data Science

Visualization of classification algorithms: Decision tree

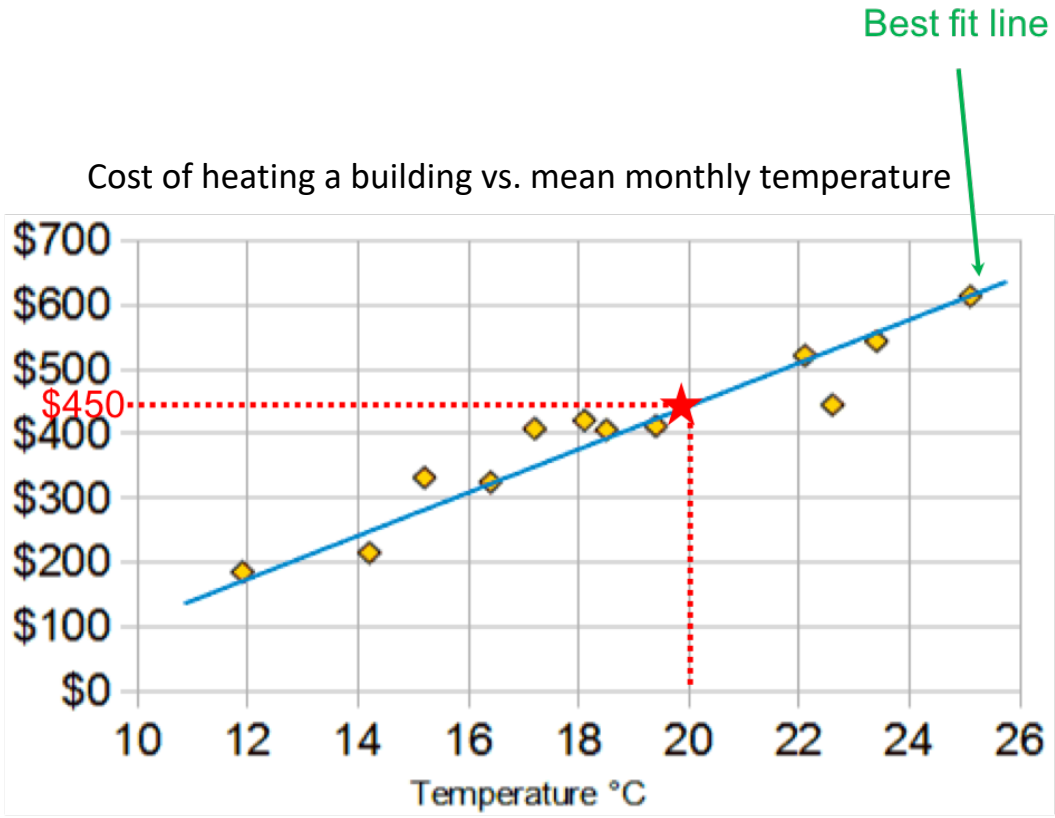




The Center for Applied Data Science

Regression

Temperature (C)	Cost
12	192
14.5	205
15.2	325
16.7	320
…	…
20	?

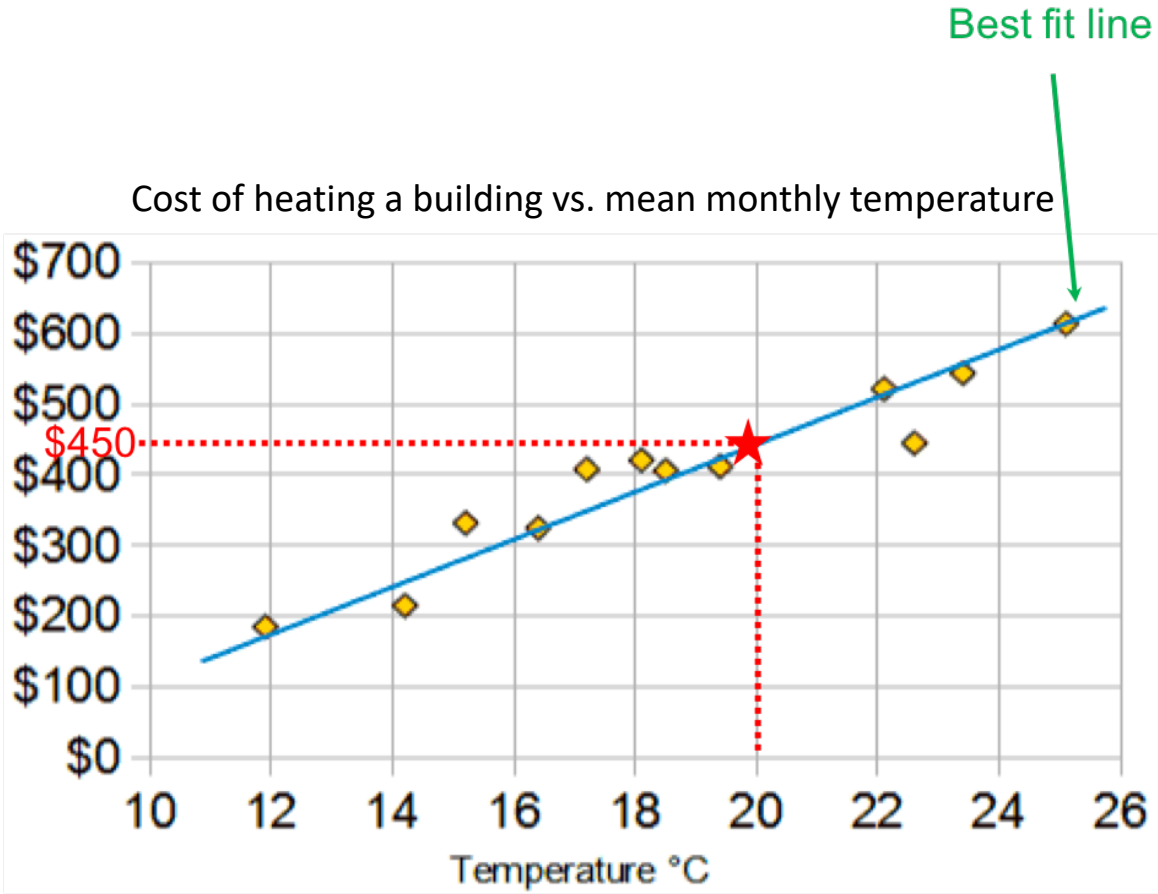


Question: What would be the cost for a temperature of 20°C?

The Center of Applied Data Science

Regression

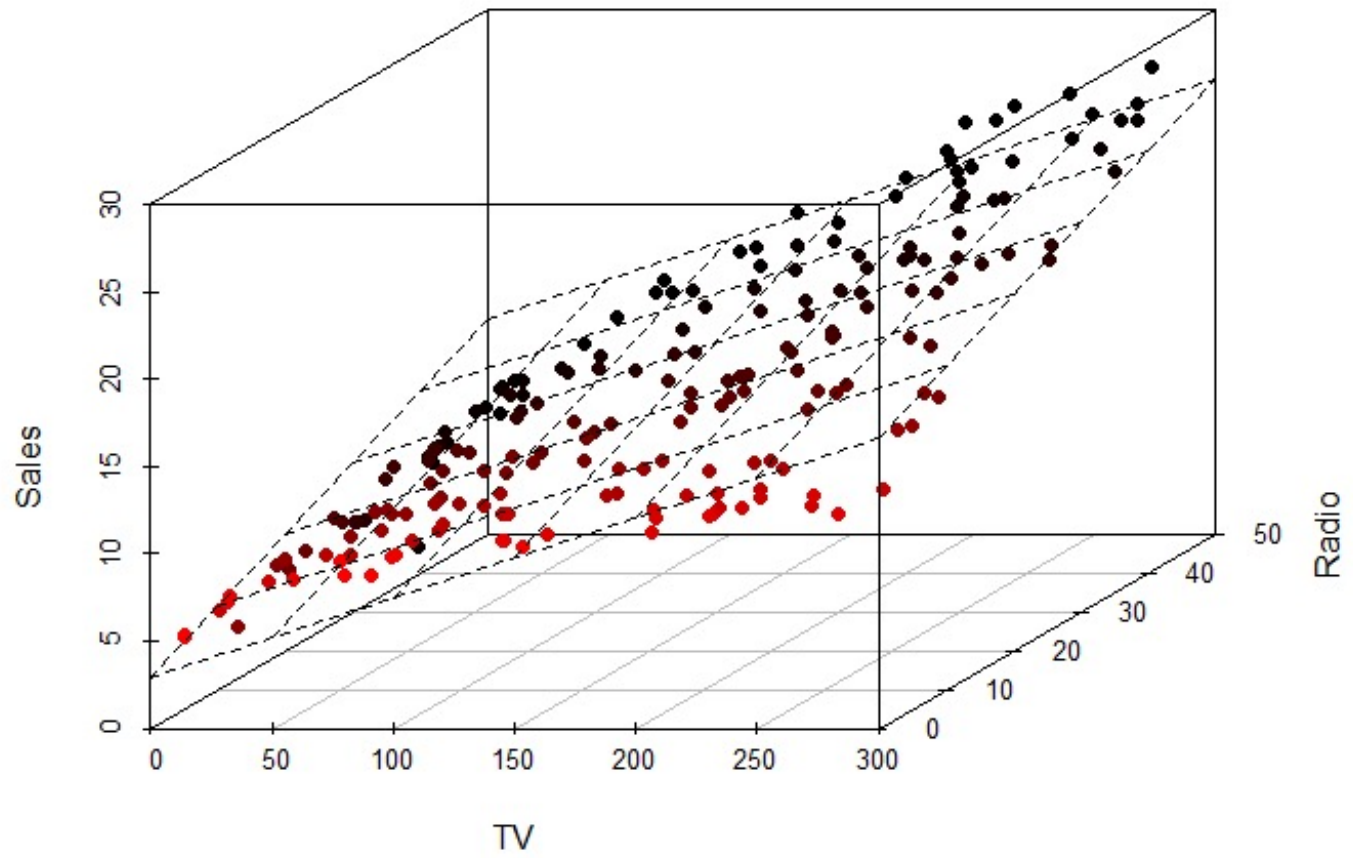
Regression predicts a target attribute which is a real number



Question: What would be the cost for a temperature of 20°C?

The Center of Applied Data Science

Regression

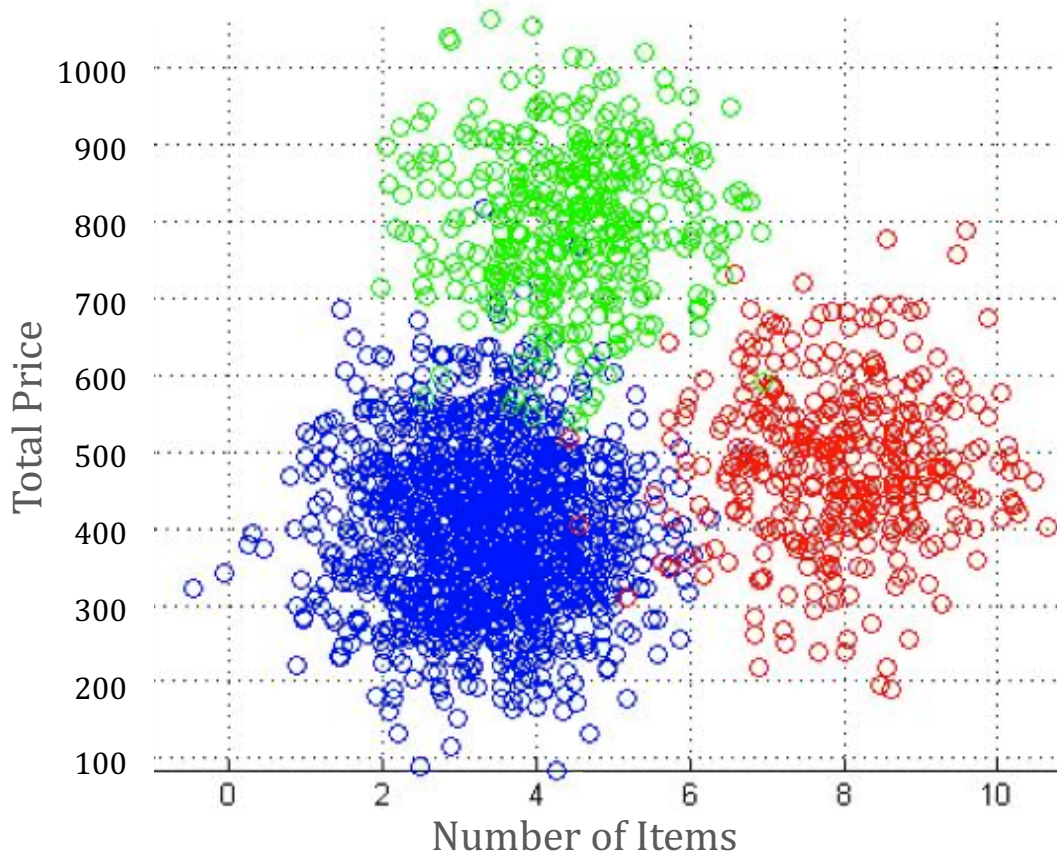


The Center of Applied Data Science

Clustering

Number of Items	Total Price
2	50
4	1500
9	199
8	550
…	…

Question: How many clusters of buying behavior you observe in the data?

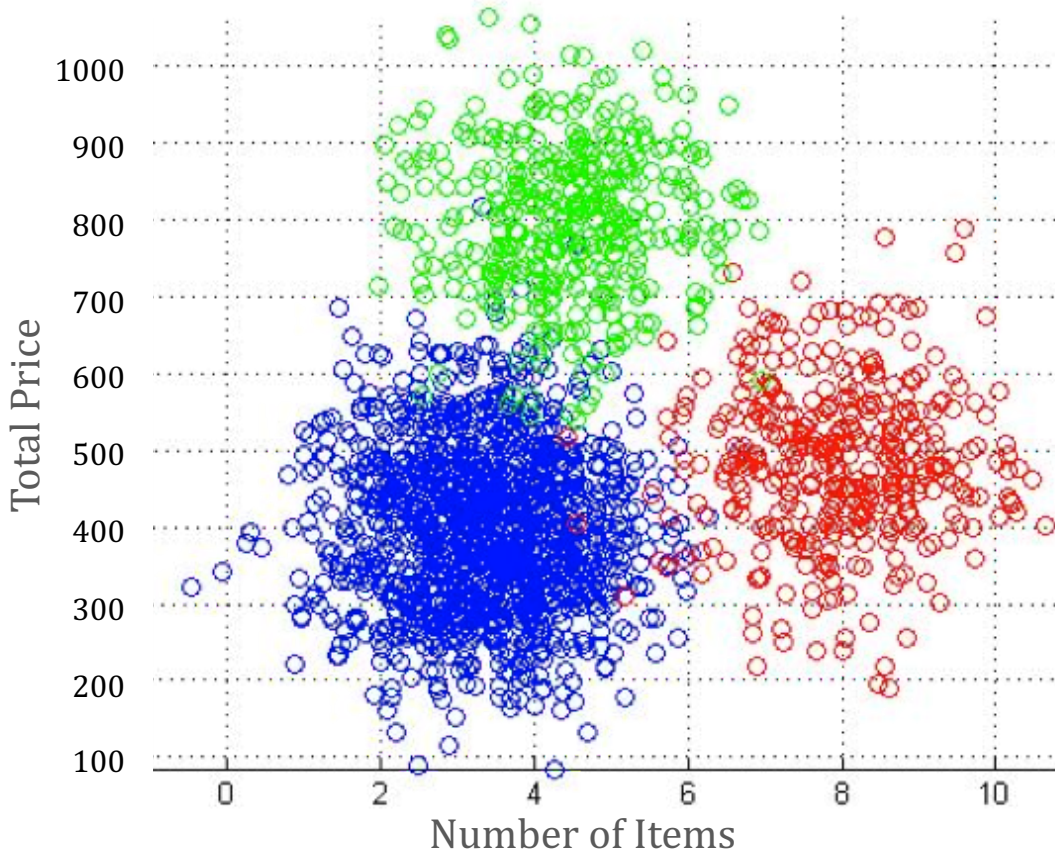


The Center of Applied Data Science

Clustering

- Clustering is similar to classification except that there are no category labels in the dataset

- Instead, we are looking for a natural grouping of the datapoints



The Center of Applied Data Science

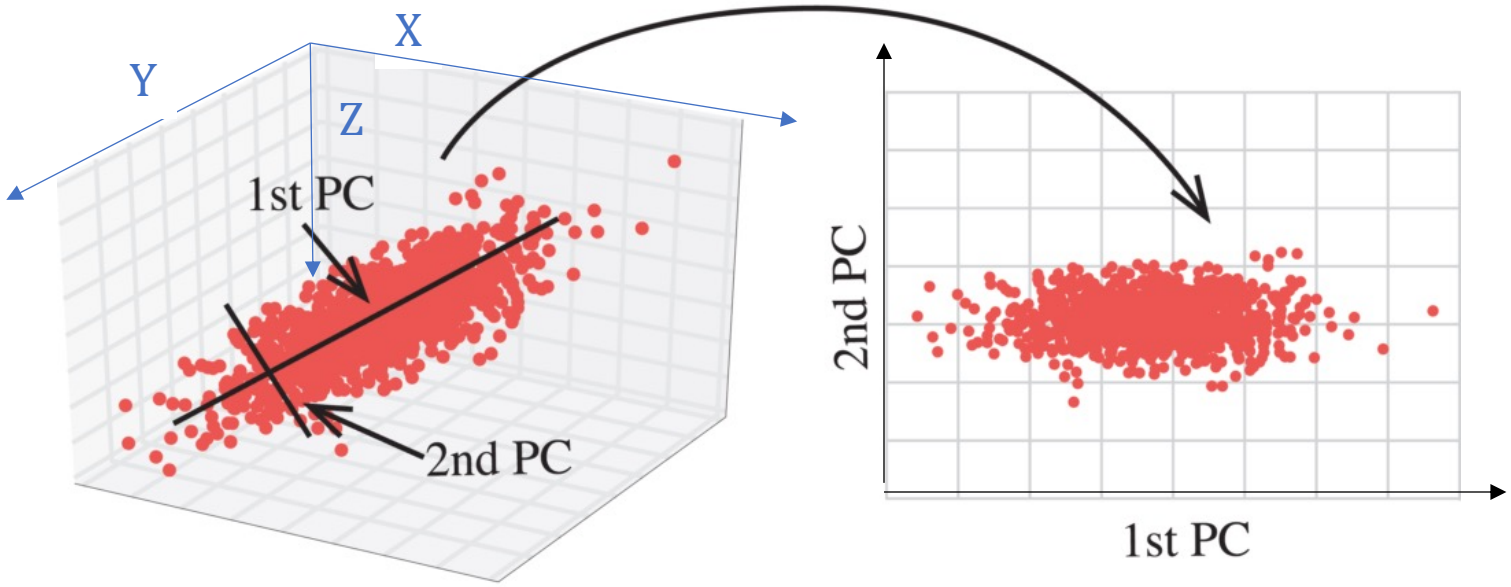
Dimensionality reduction

3 Dimension

X	Y	Z
0.5	-1	2.3
1	2.7	-3.3
0.2	1.2	3.8
…	…	…

2 Dimension

PC1	PC2
?	-?
?	?
?	?
…	…



3 Dimension → 2 Dimension

The Center of Applied Data Science

Dimensionality reduction

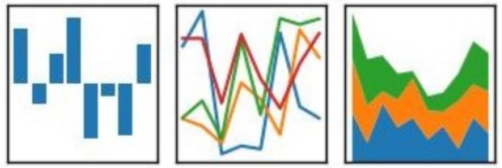
- In some datasets, there are hundreds or thousands of attributes (columns)
  - **Example**: Netflix gives predictions for each user next based on their ratings of previous movies. In this case, every single movie in the database is an attribute
- A technique for reducing the number of attributes in a dataset
- Commonly used as a preprocessing step
- Collapses the attributes into new "latent dimensions"
- No guarantee that the reduced dimensions are comprehensible

The Center of Applied Data Science

Open source project containing a good number of state-of-the-art ML algorithms

Good documentation, active user community

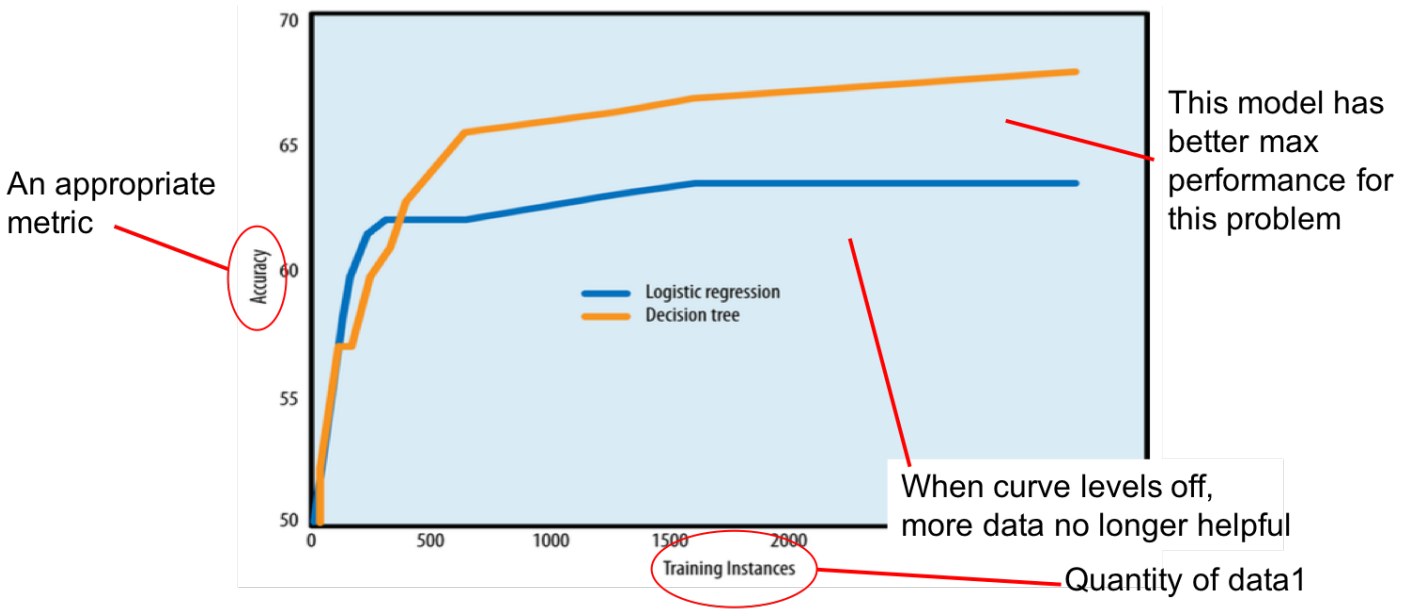
Works well with a number of other libraries (numpy, pandas)



Other libraries and tools

Evaluating models

Example:
The learning curve



The learning curve helps answer two questions

- Which model is better for this problem?
- Do we have enough data?

The Center of Applied Data Science

Thank You

The Center of Applied Data Science