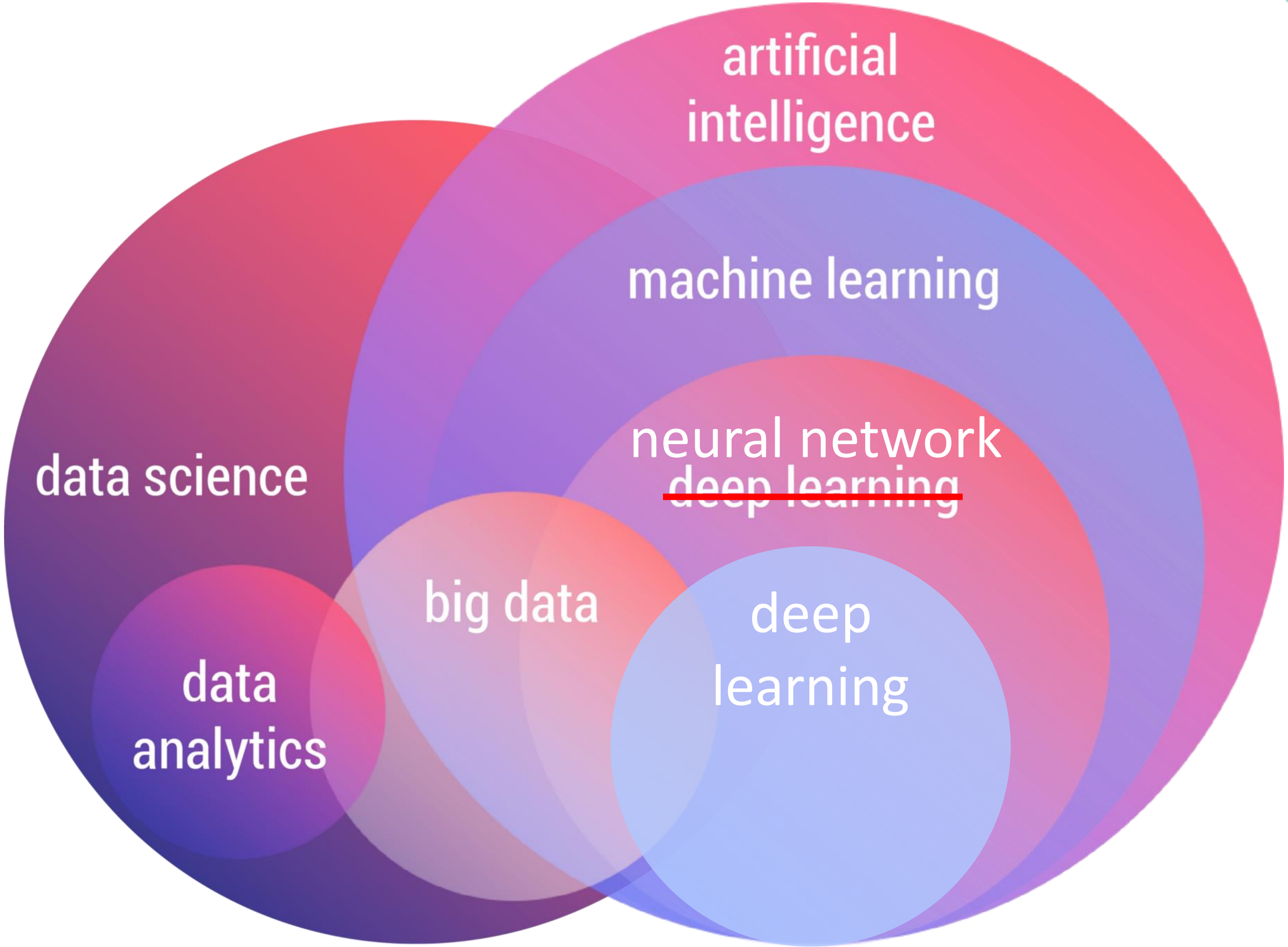
Introduction to **Big Data** in Business

Dr.Pongrapee Kaewsaiha

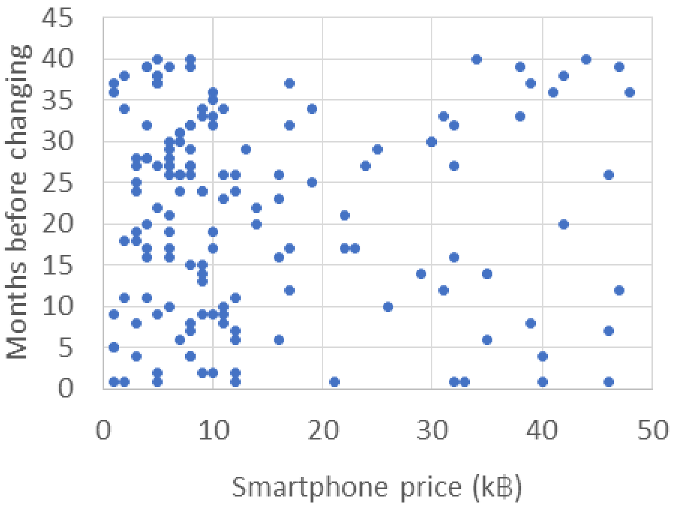
College of Hospitality
Industry Management
Suan Sunandha Rajabhat University

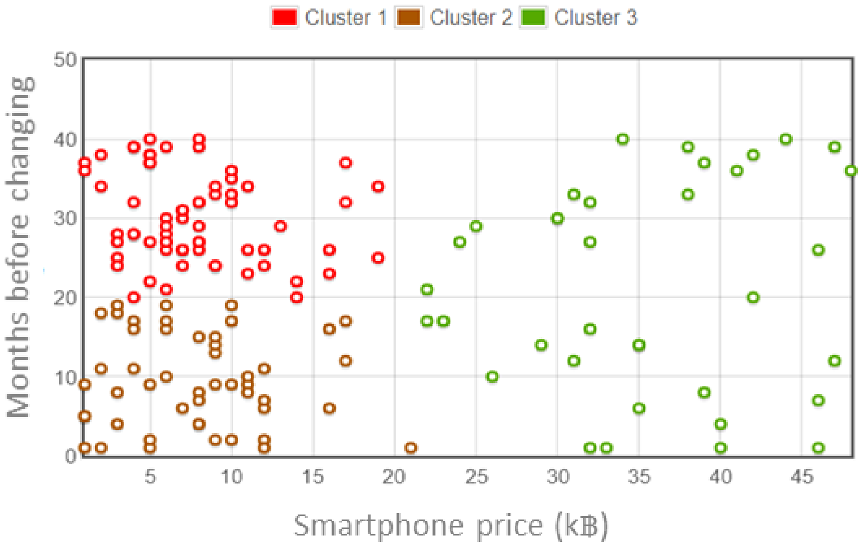


Source: towardsdatascience.com

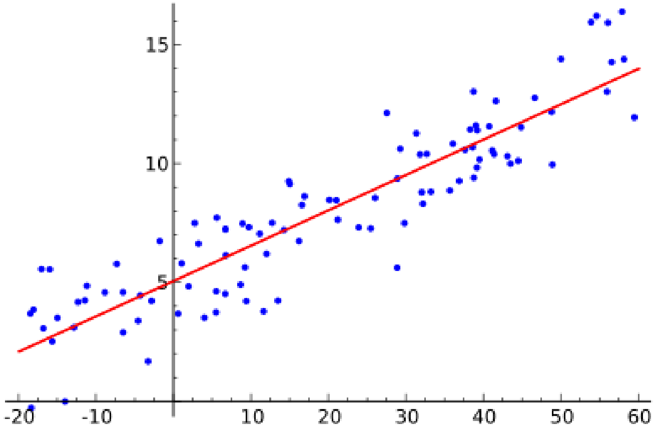
Purposes of data science

Find hidden pattern within the dataset



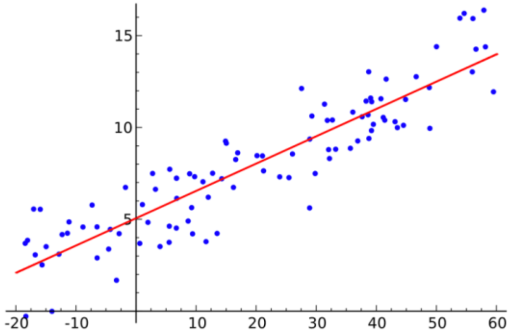


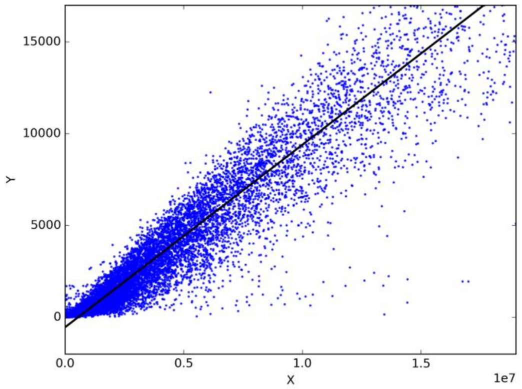
Predict trends from the existing dataset



Big data and its challenges

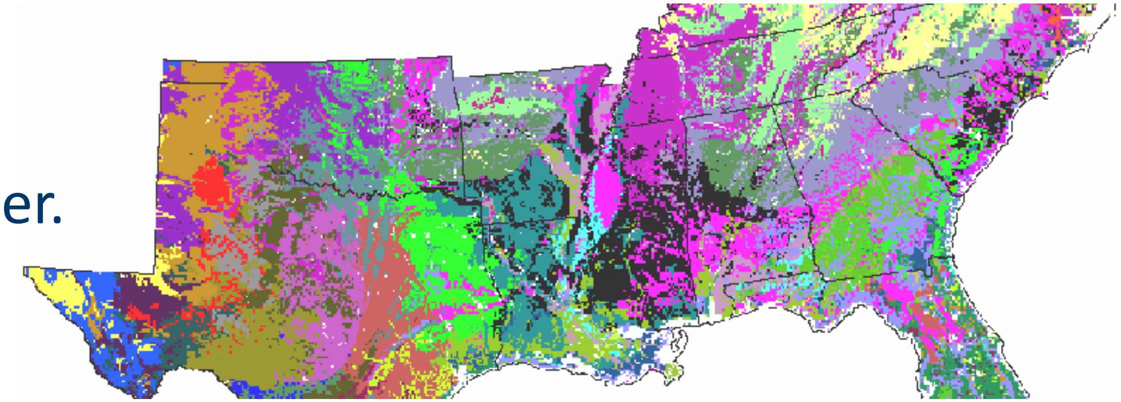
The bigger the dataset is, the more accurate the prediction would be. But big data is prone to outlier, noise, and bias.





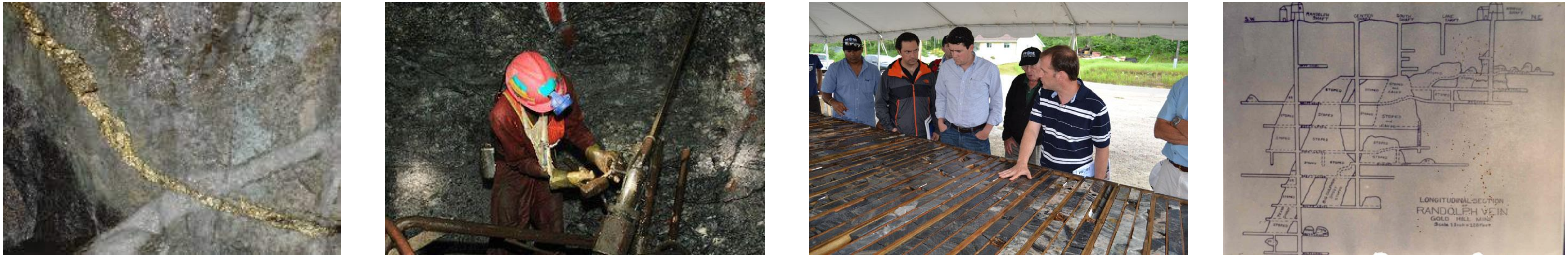
Not every single piece of data is useful. There is less than 1% gold in the rock formation.

Huge dataset can freeze the computer.



Big data and its challenges

Extract or locate useful/valid/valuable data. → Data mining



Gold vein probe

Use programming languages with lower complexity. → Python, R

Data types

Binary – Only two possible options *(True/False), (Yes, No), (Positive/Negative)*

Categorical – Two or more options

- Customer type: Eat-in, Take-away, Delivery
- Level of education: Below high-school, High-school, Diploma, …
- …

Integer – Must be a whole number

- Number of customers: There cannot be 1.25 customers.
- Boxes: Although it is possible to fill only half a box, we do not accept that in the inventory record.
- …

Data types

Continuous – Numerical data that can have decimal places

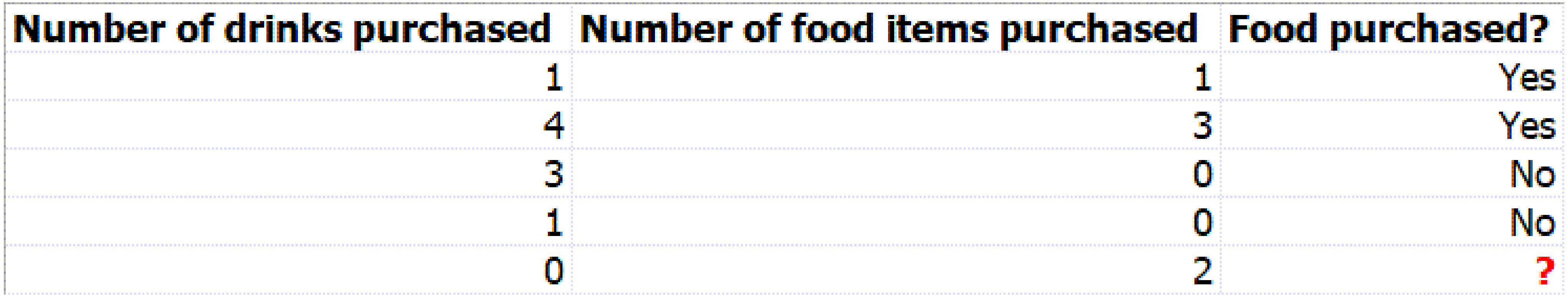
- Price: $10.99, 20.25 THB, …
- Length: 24.3 cm, 1’ 3.66”, …
- Weight: 12.65 kg, 3.24 pounds, …
- …

Formatted data – Must be encoded before it can be analyzed

- Image → Pattern recognition
- Voice → Voice commands
- Digital file → Digital file conversion
- …

Some data needs to be engineered

If we want to group customers based on whether they buy food or not, we need to translate the number of food purchased into binary.



Number of drinks purchased	Number of food items purchased	Food purchased?
1	1	Yes
4	3	Yes
3	0	No
1	0	No
0	2	?

Number of food items purchased > 0 : **Yes**
Number of food items purchased = 0 : **No**

Sometimes we need to interpret the data

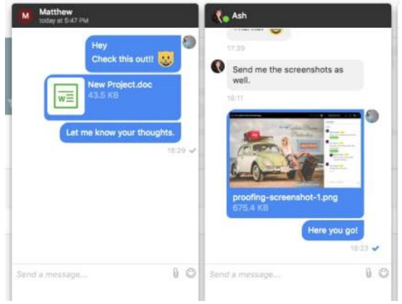
Document submissions

Group 1

- Submit in person
- Submit via postal mail
- Submit via messenger service

Group 2

- Submit by email
- Share via cloud drive
- Submit via instant messaging



Matthew
today at 5:47 PM

Hey
Check this out!!

New Project.doc
43.5 KB

Let me know your thoughts.

18:29

Send a message...

Ash

17:39

Send me the screenshots as well.

18:11

proofing-screenshot-1.png
675.4 KB

Here you go!

18:23

Send a message...

What's the difference between Group 1 and Group 2?

Sometimes we need to interpret the data

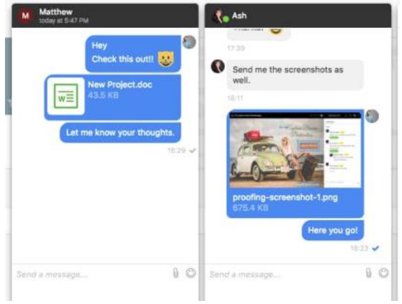
Document submissions

Group 1

- Submit in person
- Submit via postal mail
- Submit via messenger service

Group 2

- Submit by email
- Share via cloud drive
- Submit via instant messaging



Matthew
today at 5:47 PM

Hey
Check this out!!

New Project.doc
43.5 KB

Let me know your thoughts.

18:29

Ash

17:39

Send me the screenshots as well.

18:11

proofing-screenshot-1.png
675.4 KB

Here you go!

18:22

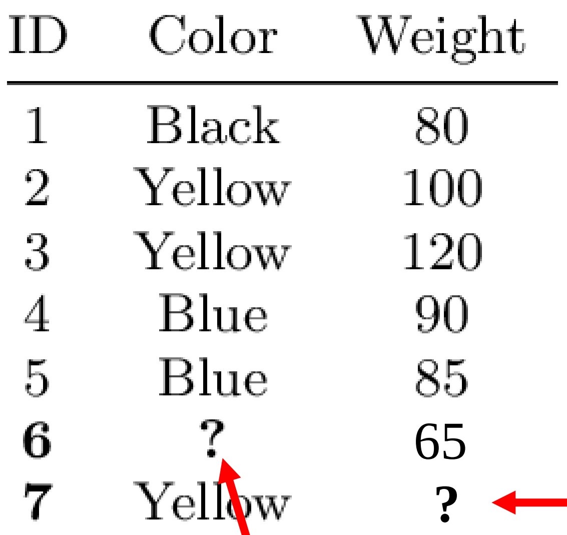
Send a message...

Send a message...

What's the difference between Group 1 and Group 2?

The algorithm groups the data for you, but you must interpret the result by yourselves.

Dealing with missing data



ID	Color	Weight
1	Black	80
2	Yellow	100
3	Yellow	120
4	Blue	90
5	Blue	85
6	**?**	65
7	Yellow	**?**

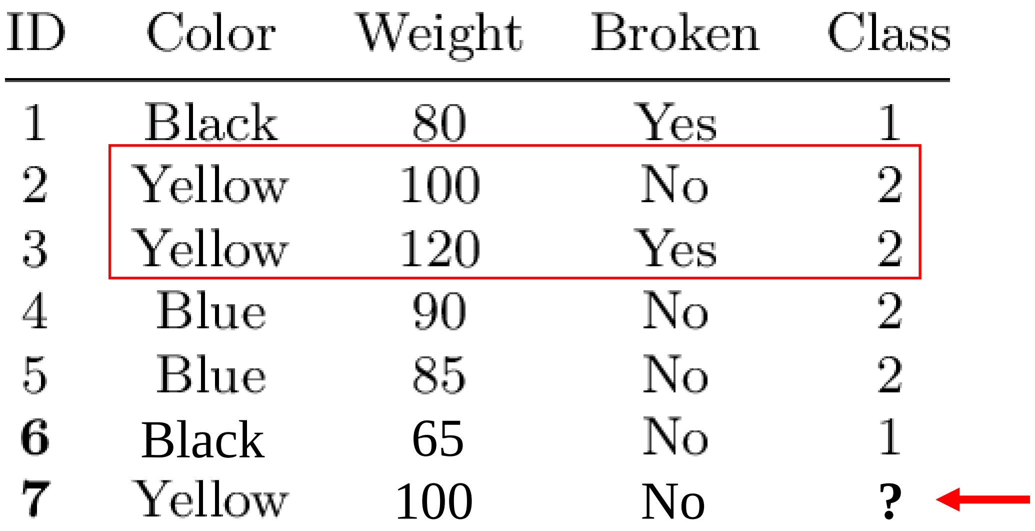
We should deal with missing data in a way that minimize the effect.

Method 1: Replace with central tendency

This should be “**100**” (the group average).

This should be “**Black**” since there are two Blue, three Yellow, but only one Black.

Dealing with missing data

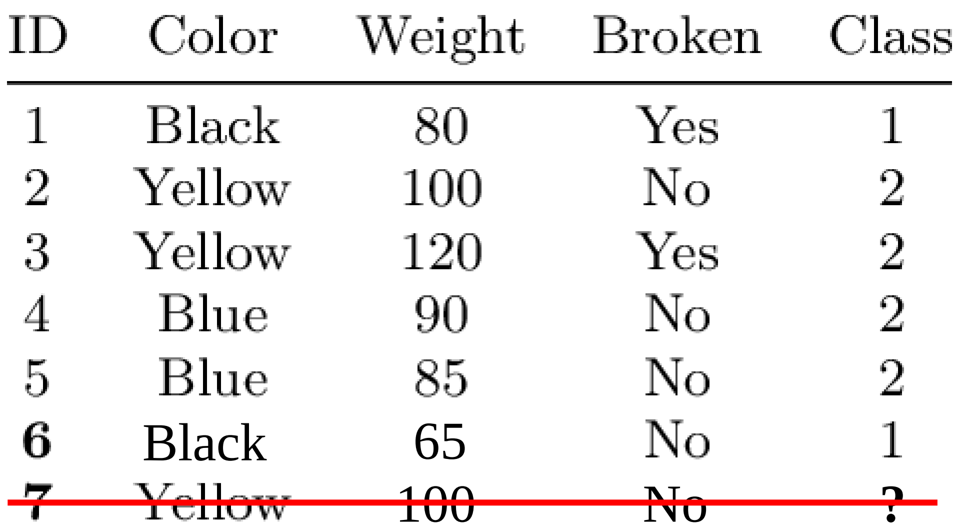


ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	Black	65	No	1
7	Yellow	100	No	**?**

Method 2: Find the reference

This should be “**2**”
following the above “yellow” records.

Dealing with missing data



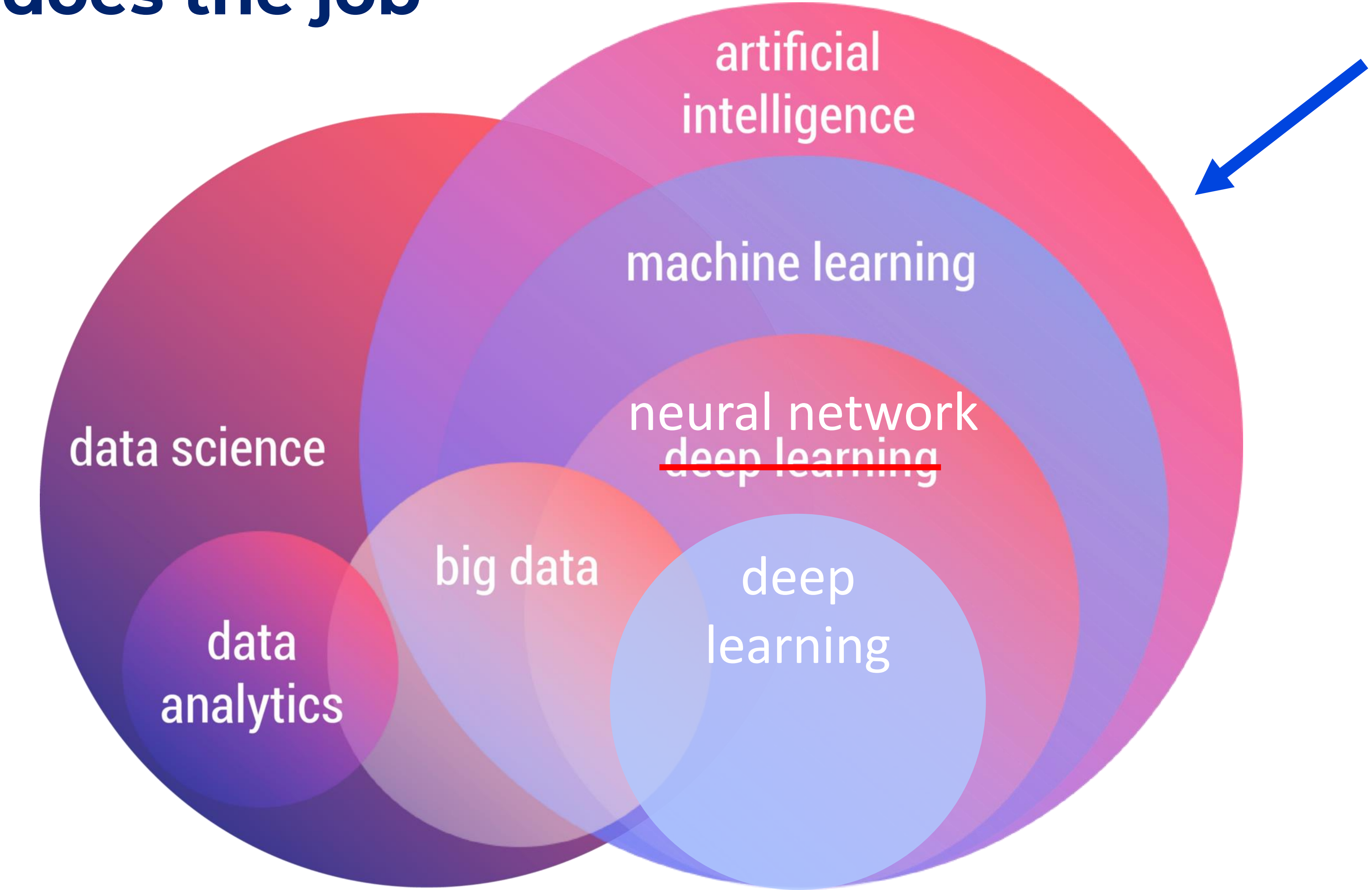
ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	Black	65	No	1
~~**7**~~	~~Yellow~~	~~100~~	~~No~~	~~?~~

Method 3: Remove the entire record

Do this as a last choice because

- It lowers the data size → Lower accuracy
- It can be considered as concealing the transactions (illegal)

Let AI does the job



Supervised learning

We give AI the pattern, called “**training**” dataset.

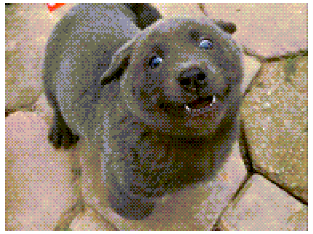
In this case, we have a database of cat and dog photos.

This is formatted data.
(Do you remember that?)
So, the AI needs to extract the pattern first.

Cats

Dogs

Then we give the input, called “**test**” dataset.



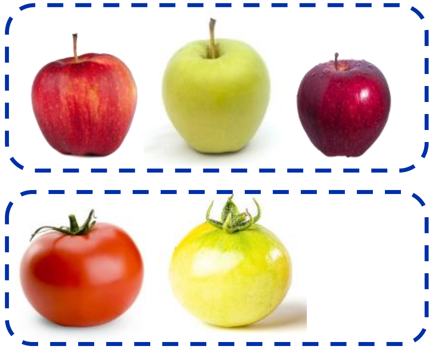
Cat or dog?

AI will compare it with the database and pick the closest answer.

Unsupervised learning

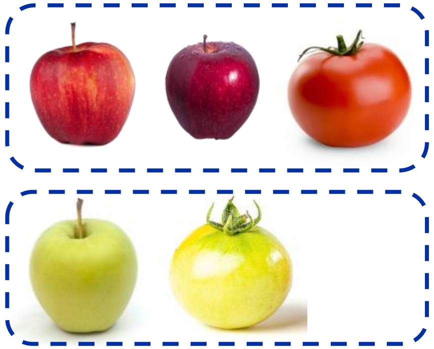
We give a mixture of apple and tomato photos without telling which one is an apple and which one is a tomato.

The algorithm will try to extract their features (e.g., skin color and stem) and try to group them based on similarity.



If it tries to group these photos based on "**stem**," it should come up with two groups like this. But it cannot tell which are apples and which are tomatoes because we did not provide this information.

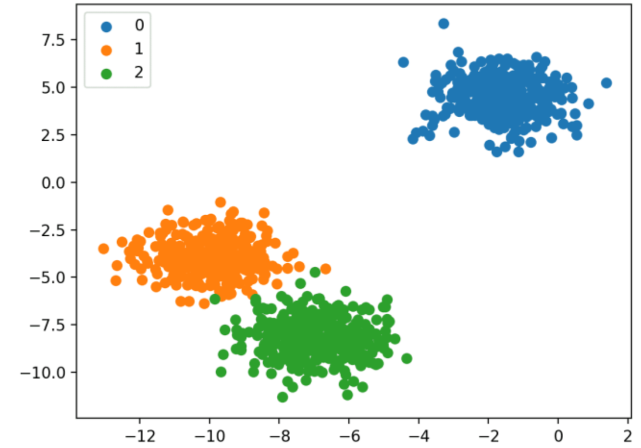
Unsupervised learning



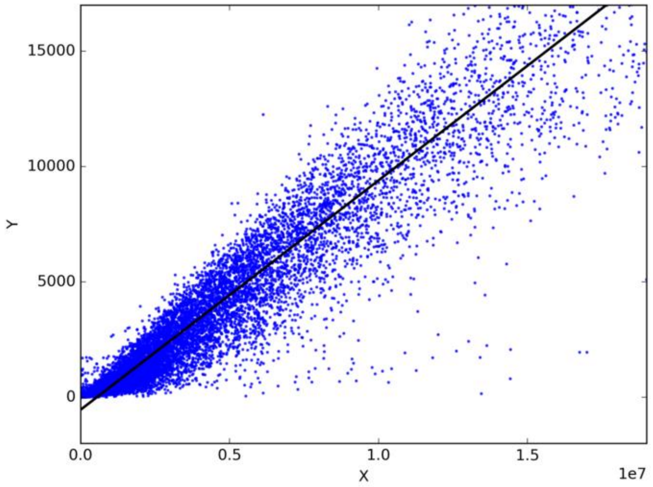
If it tries to group these photos based on "**skin color**," it should come up with two groups like this. The user must interpret the result manually and see if the grouping is reasonable or give any useful information or not.

Problem types

For binary and categorical output, we often do the "**classification**" problem.

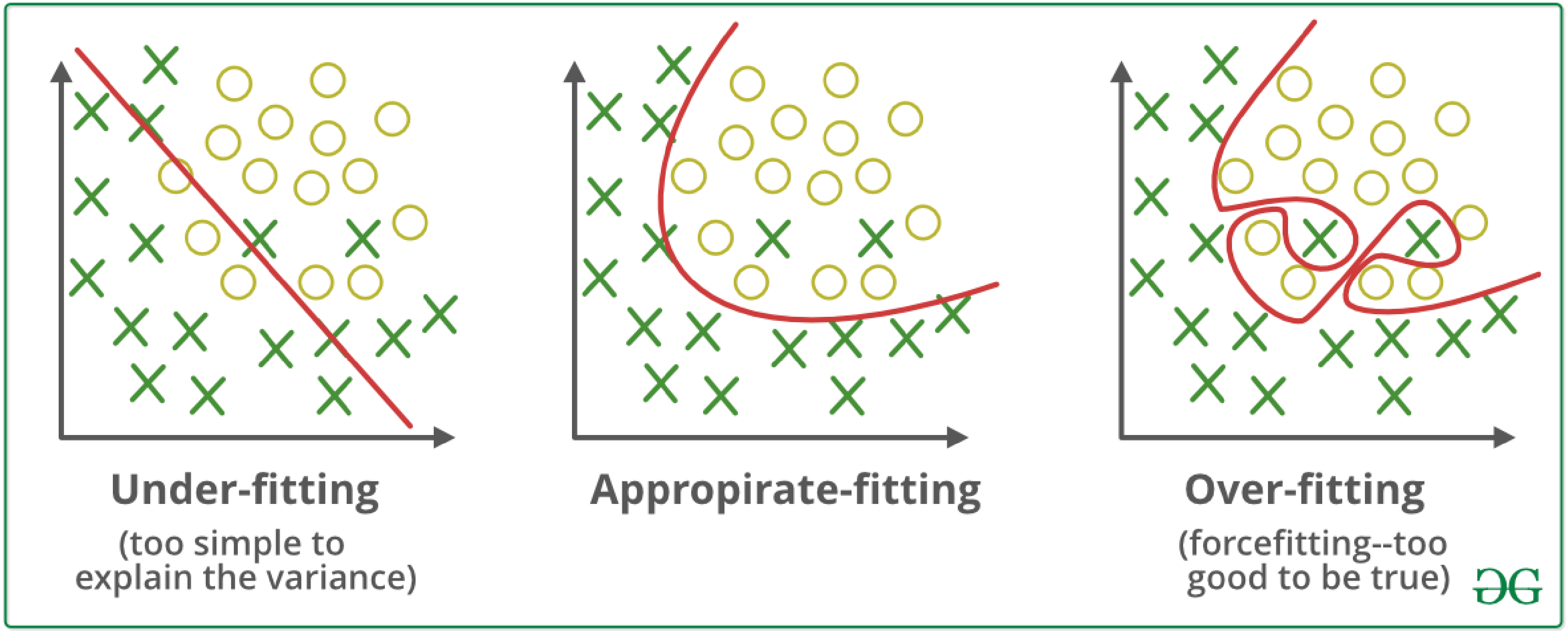


For integer and continuous output, we often do the "**regression**" problem.



Parameter adjustment

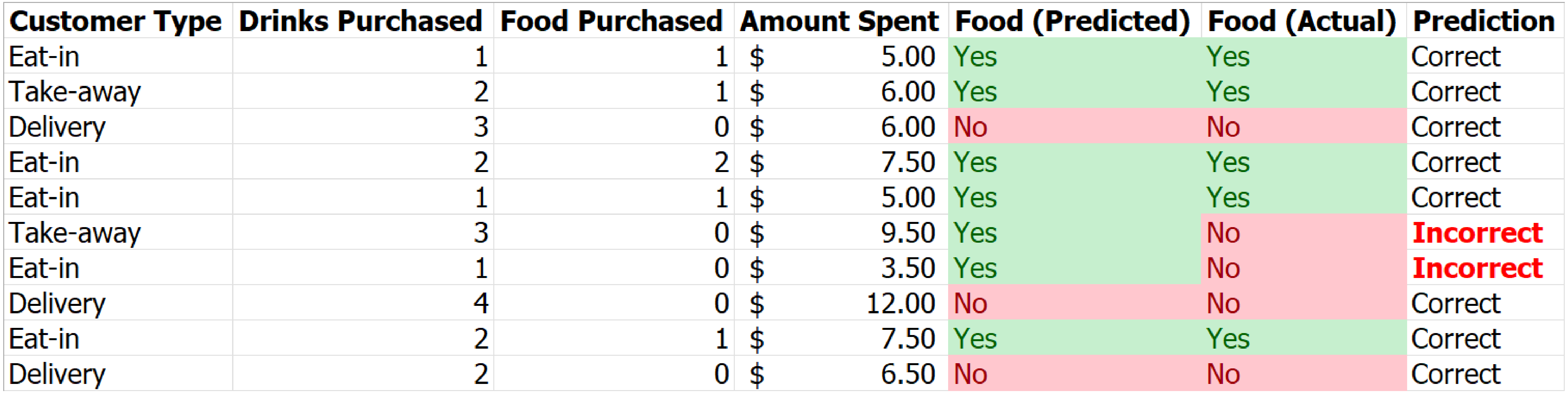
Although we apply the same algorithm to the same dataset, the results can be different depending on the parameter adjustment.


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Prediction accuracy

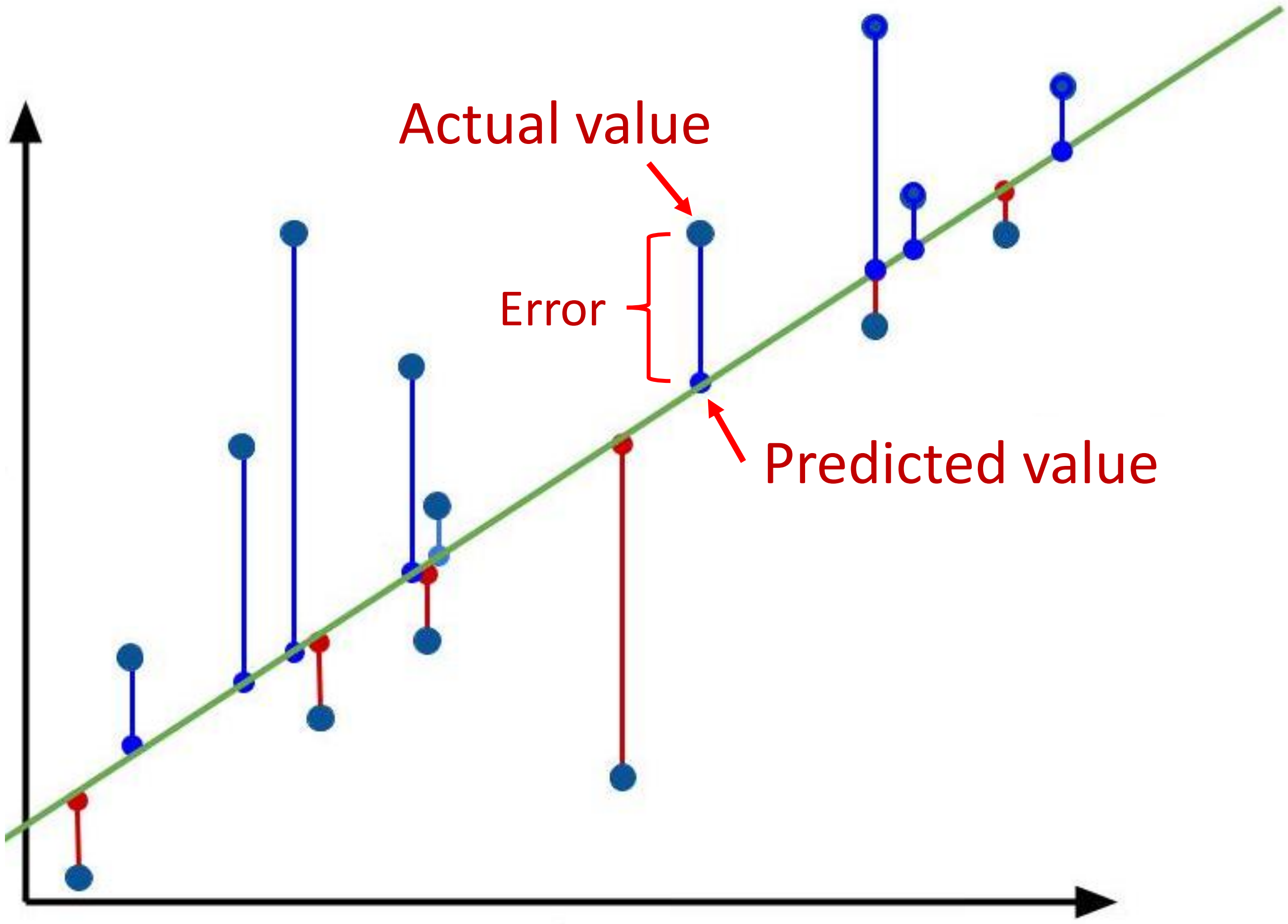
Compare predicted and actual output data to see if the prediction is correct.

Classification metric



Customer Type	Drinks Purchased	Food Purchased	Amount Spent	Food (Predicted)	Food (Actual)	Prediction
Eat-in	1	1	$ 5.00	Yes	Yes	Correct
Take-away	2	1	$ 6.00	Yes	Yes	Correct
Delivery	3	0	$ 6.00	No	No	Correct
Eat-in	2	2	$ 7.50	Yes	Yes	Correct
Eat-in	1	1	$ 5.00	Yes	Yes	Correct
Take-away	3	0	$ 9.50	Yes	No	**Incorrect**
Eat-in	1	0	$ 3.50	Yes	No	**Incorrect**
Delivery	4	0	$ 12.00	No	No	Correct
Eat-in	2	1	$ 7.50	Yes	Yes	Correct
Delivery	2	0	$ 6.50	No	No	Correct

Regression metric



Confusion metric

		Prediction	
		Will buy	Will not buy
Actual	Buy	**1** (True positive)	**5** (False negative)
	Don’t buy	**5** (False positive)	**89** (True negative)

90 true predictions out of 100 transactions

→ Accuracy = 90%

This system predicts **negative** results more accurate.

True positive

We predict that customers will buy, and they actually buy it.

True negative

We predict that customers won’t buy, and they don’t.

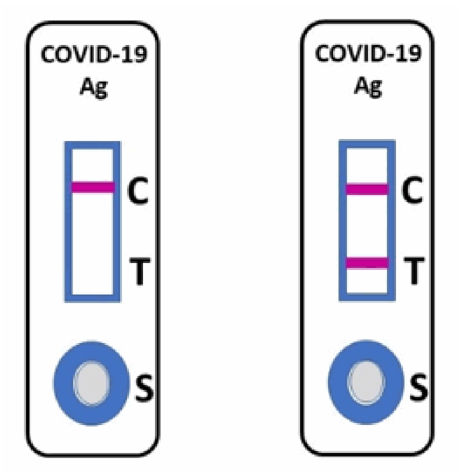
False positive

We predict that customers will buy but they don’t.

False negative

We predict that customers won’t buy but they buy it.

COVID-19 test kits are developed based on data science concept.



Negative **Positive**

True positive: Infected and the test kit gives a positive result.

- The patient is sent to isolation and receive treatment.

True negative: Not infected and the test kit gives a negative result.

- The patient can continue working.

False positive: Not infected but the test kit gives a positive result.

- The healthy person is sent to isolation and receive treatment.

False negative: Infected but the test kit gives a negative result.

- The patient can continue working and spread virus further.

Earthquake warning system also relies on proper parameter adjustment.

True positive: Earthquake is coming, and the system sounds the alarm.

- People evacuate to the safe place.

True negative: No earthquake and no alarm.

- People continue what they are doing.

False positive: No earthquake but the alarm sounds.

- People evacuate to the safe place and return as nothing happens.

False negative: Earthquake is coming but no alarm.

- People continue what they are doing and the earthquake hits.

Example: Earthquake prediction

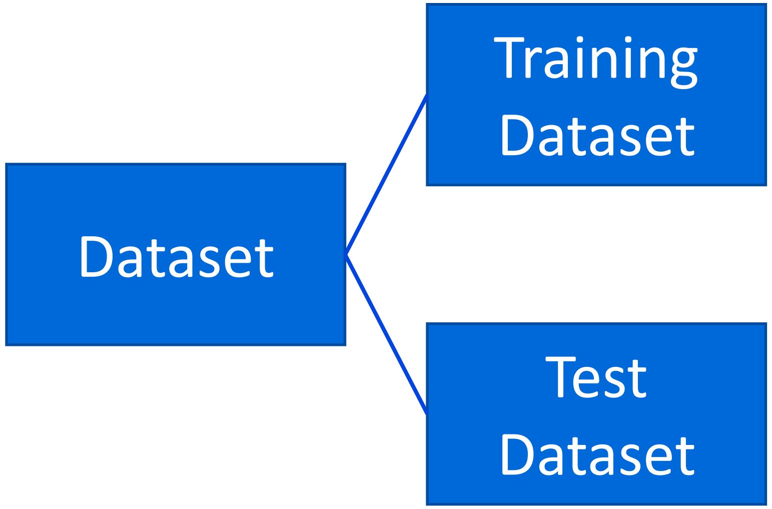
False negatives = Too many loss

(forecasting that no earthquake will occur, but it happens)

False positive = No loss

(anticipating an earthquake but not happening)

Validation

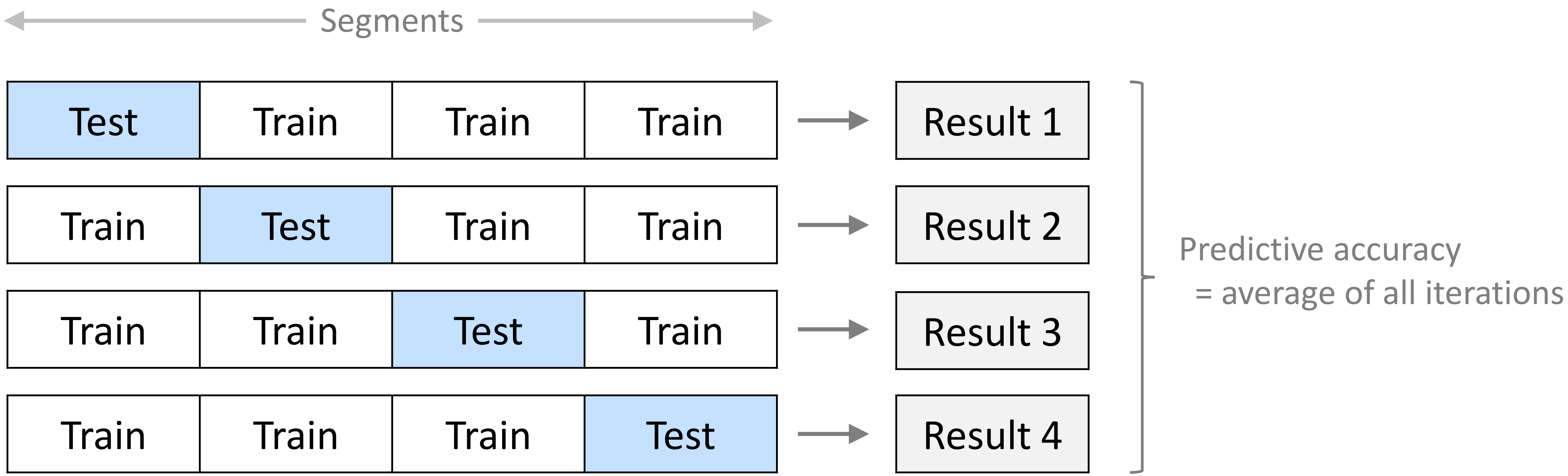


Create and refine forecast models

Assess the model's predictive accuracy

- We split the dataset into two groups, one for training and one for test.
- If the result doesn’t change, the algorithm is strong enough.
- But doing so will decrease the data size → Lower the accuracy.

Cross-validation



Test	Train	Train	Train	→	Result 1
Train	Test	Train	Train	→	Result 2
Train	Train	Test	Train	→	Result 3
Train	Train	Train	Test	→	Result 4

Predictive accuracy
= average of all iterations

If the data size is not big enough, we can divide it into segments.