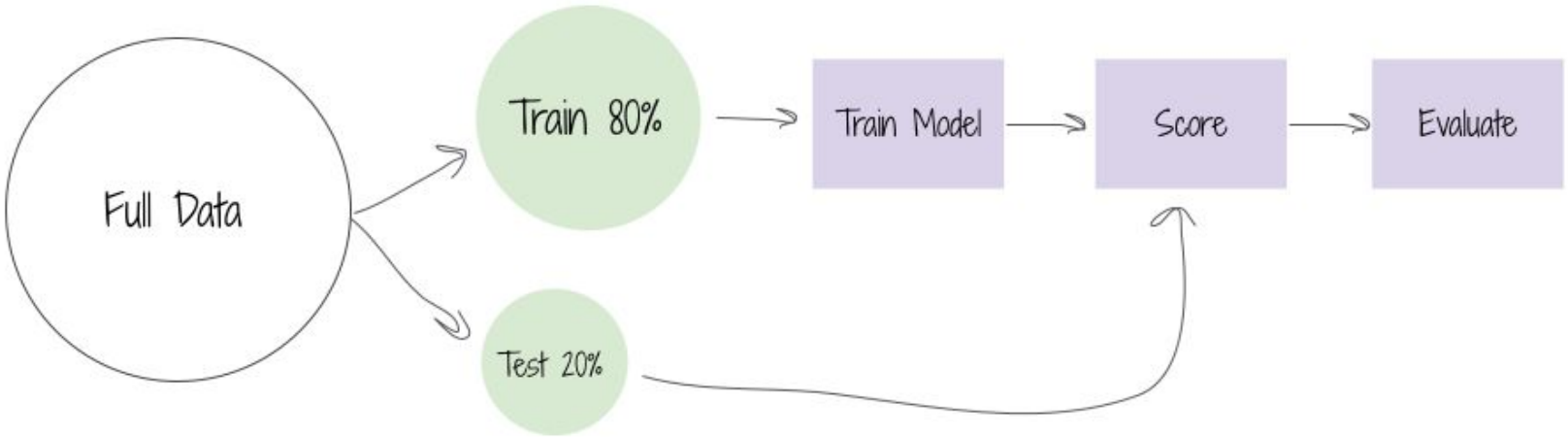
Bootcamp Live 07
Introduction to Machine Learning

Instructor: Kasidis Satangmongkol

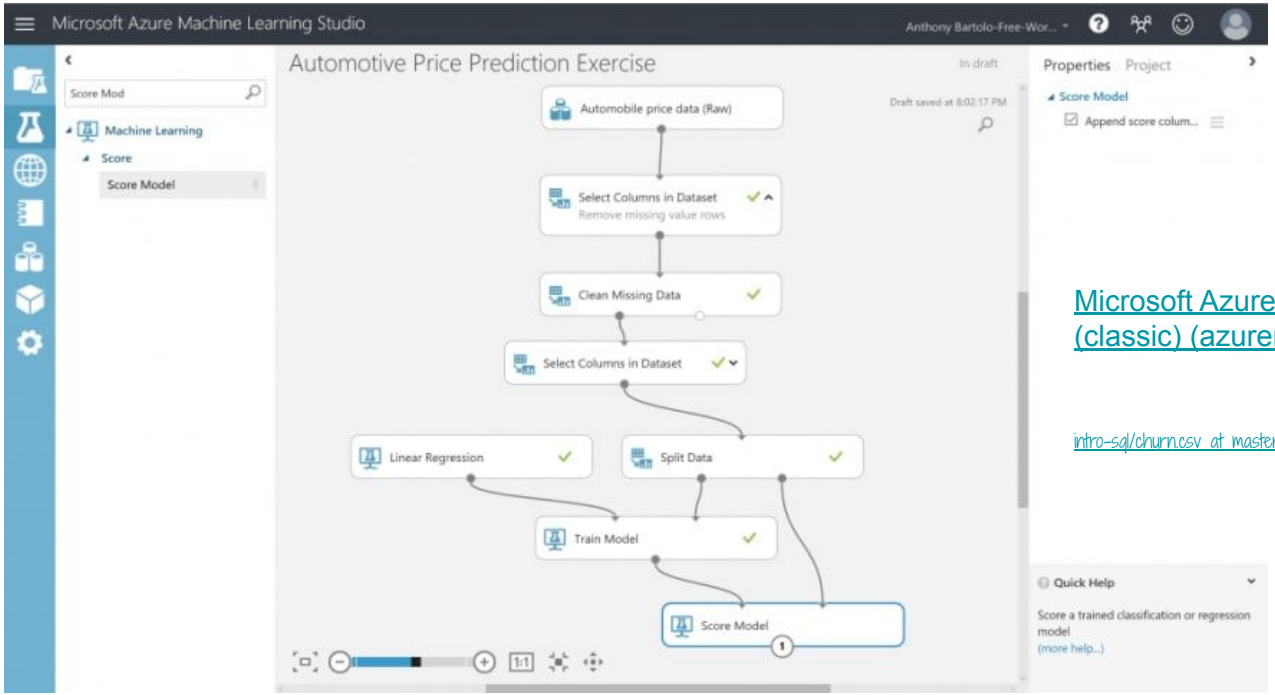
You can download full ML slide in Notion

Slide and data in our school Google Drive

Simple pipeline to build ML models



Build your first model with Azure ML Studio



[Microsoft Azure Machine Learning Studio (classic) (azureml.net)]

[intro-sql/churn.csv at master · toyeiei/intro-sql (github.com)]

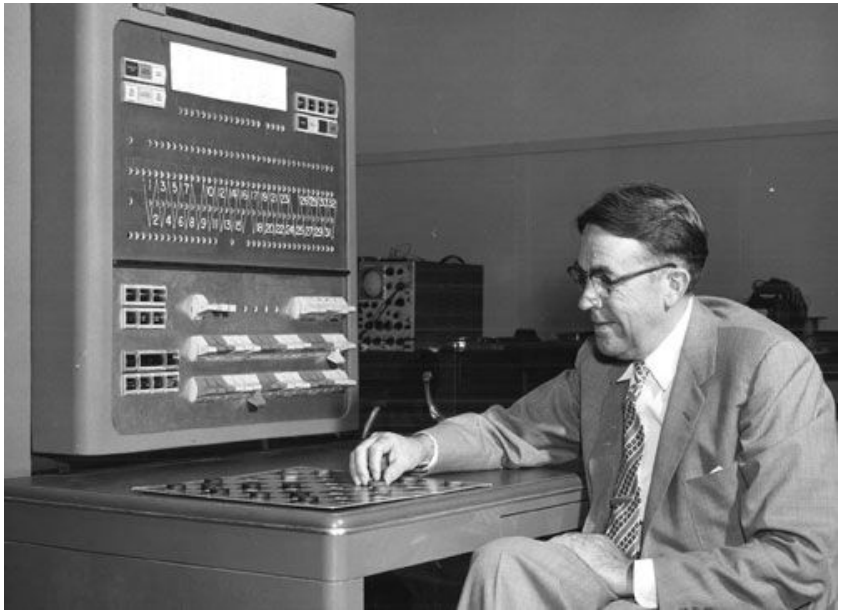
Example dataset


Machine Learning
When a computer can learn to recognize pattern

Essential ML

- what exactly is machine learning
- supervised vs. unsupervised
- regression vs. classification
- train test split vs. cross validation
- model selection + hyperparameter
- model evaluation

What is Machine Learning



Arthur Samuel (1959)

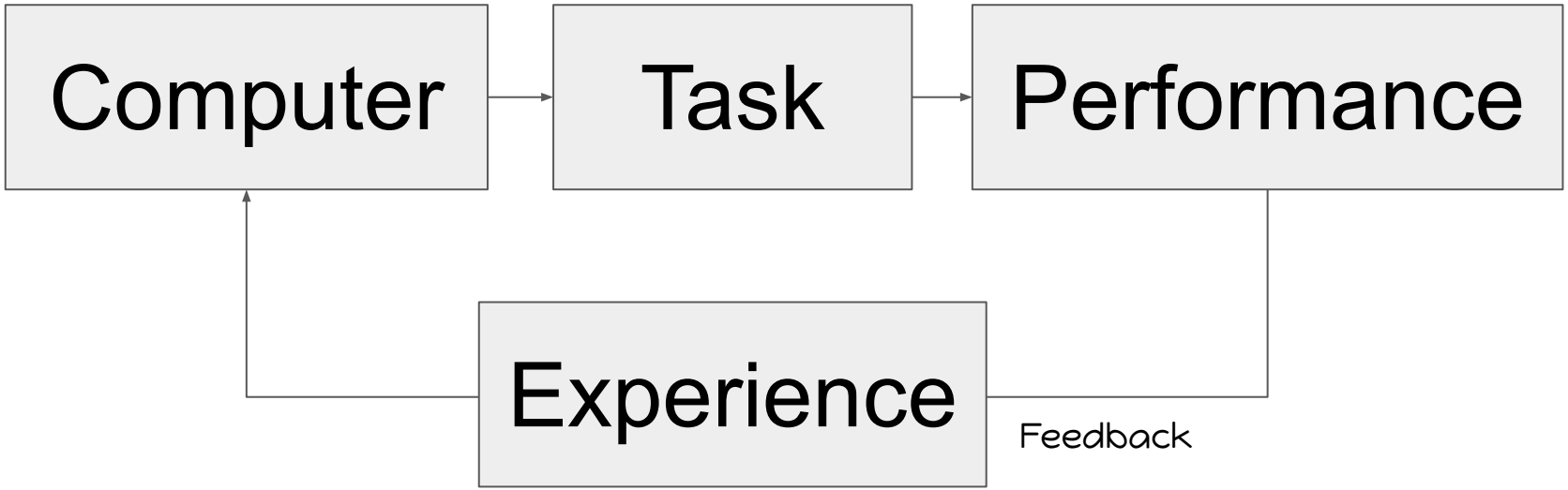
Field of study that gives computers **the ability to learn without being explicitly programmed**.

How do we learn?

Human learn from **experience**.

Computer learn from **data**.

Simple Idea



ML Glossary #1

- dataset
- data points
- features
- label or target

Data Point

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2

Dataset: BostonHousing

Features (X)

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2

Dataset: BostonHousing

Label / Target (Y)

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2

Dataset: BostonHousing

Features (X)

Label / Target (Y)

Supervised Learning

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985												5.21	28.7
7	0.08829												12.43	22.9
8	0.14455												19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2

Dataset: BostonHousing

Mapping

Features (X)

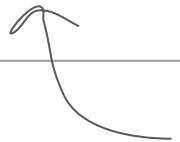
	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33
6													
7													
8													
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26

Unsupervised Learning

Dataset: BostonHousing

Quick summary

Supervised Learning	Unsupervised Learning
Has features (x) and labels (y)	Has features (x) without labels (y)
The goal is **PREDICT**	The goal is to **SUMMARISE**
Example algorithms - Regression - Classification	Example algorithms - Clustering - Association Rules - Principal Component Analysis



คอร์สเราโฟกัสที่ supervised learning

Problem 01

AIS อยากจะทำ market survey กับลูกค้า (ทุกค่าย) ทั้งหมด 3000 คน เพื่อจะดูว่าตลาดคนไทยมีลูกค้าอยู่กี่ประเภท? i.e. customer segmentation

Problem 02

Gmail มีตัวกรอง email ว่าอันไหนคือ spam อันไหนคือ ham (อีเมล์ดี)

Problem 03

อิ้งเขียนโค้ดทำ web scraping จากเว็บไซต์ขายรถยนต์มือสอง เพื่อจะดูว่ารถยนต์ Toyota รุ่น 2015 เครื่อง 1.5 ลิตร ขับมาแล้ว 20000 โล ควรจะซื้อราคาเท่าไรดี?

น้องอิ้ง!

Types of Supervised Learning

1. Regression	2. Classification
Predict **numeric** labels	Predict **categorical** labels
Examples - house price - customer satisfaction - personal income - how much a customer will spend	Examples - yes/ no question - churn prediction - conversion - weather forecast - default prediction
100, 200, 250, 190, 300, 500, etc.	0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, etc.

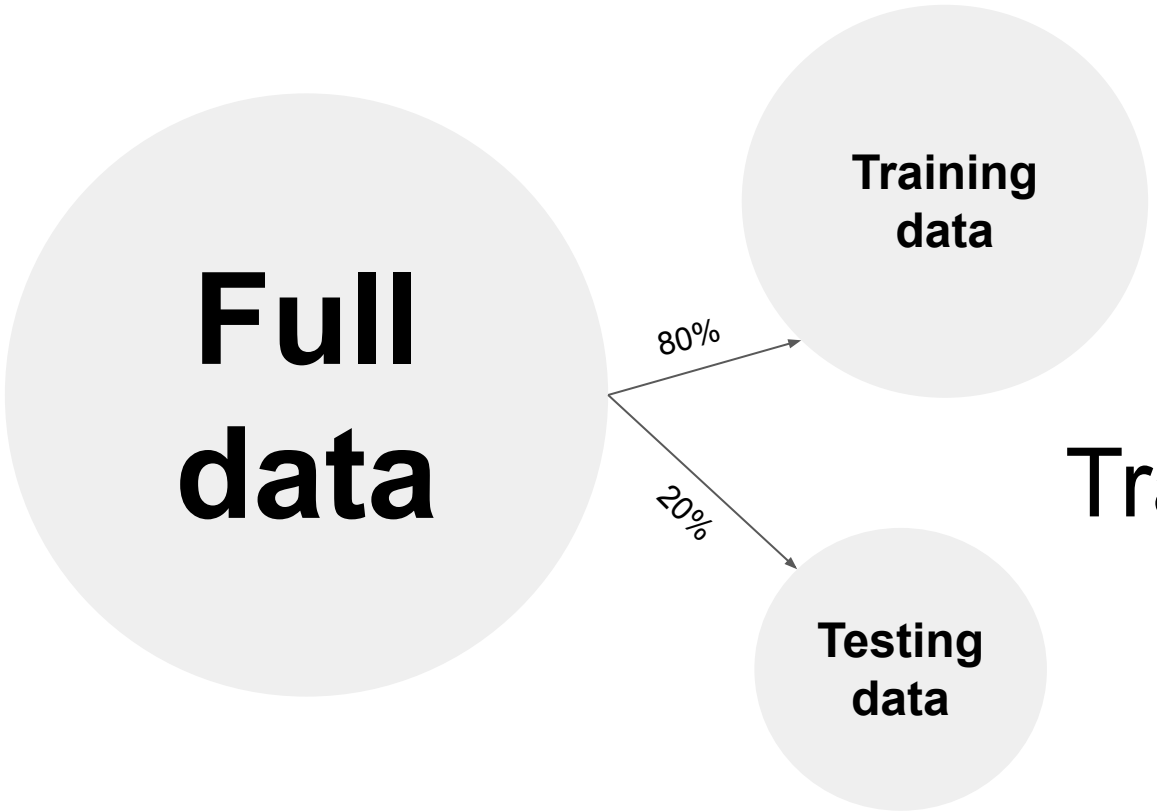
Now let's get into the details :)

3 steps to build ML

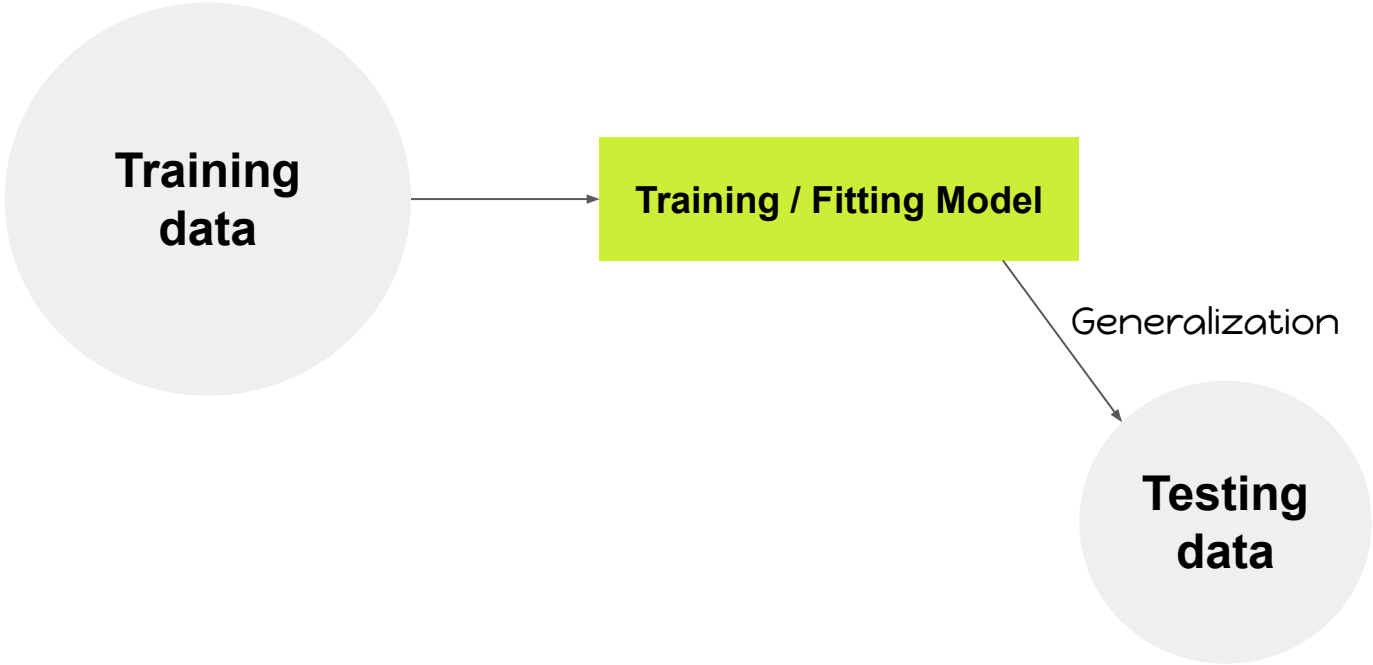
- prepare data
- train algorithm
- test/ evaluate algorithm

ML Glossary 02

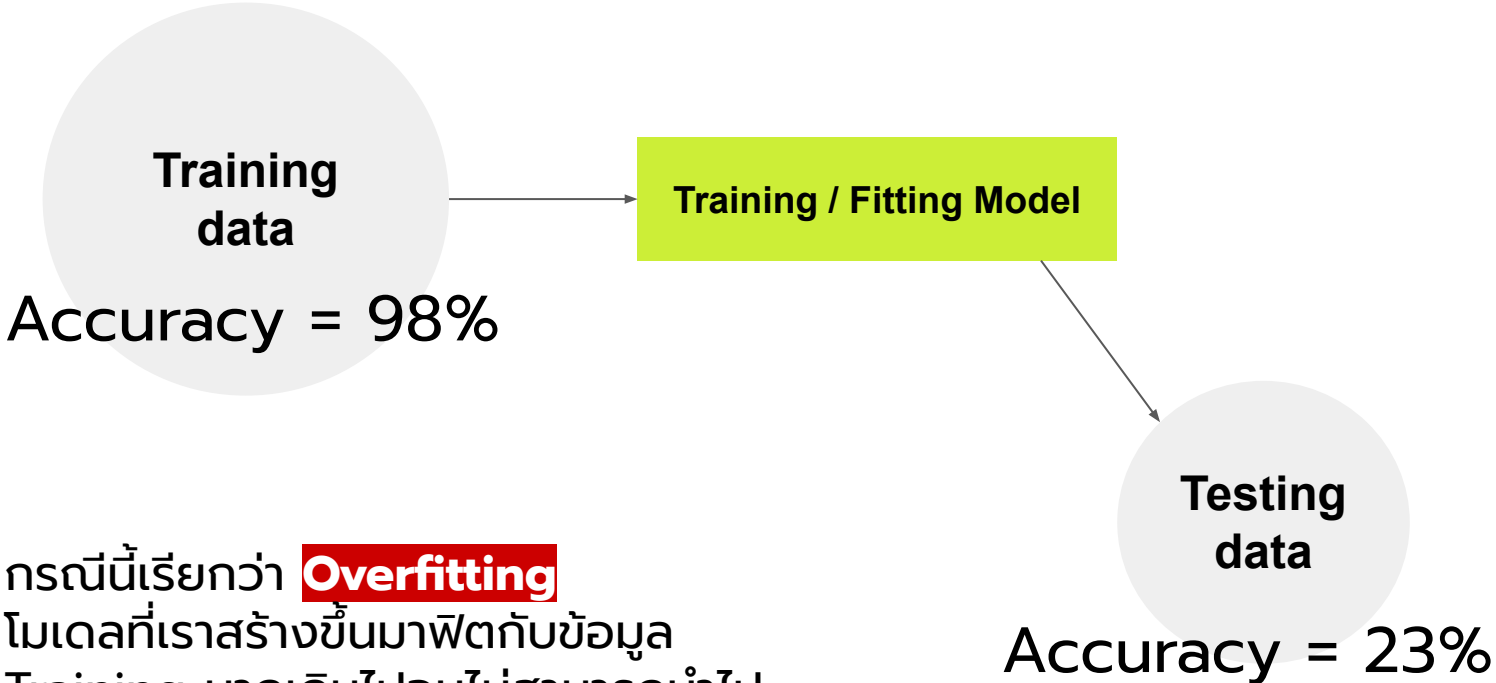
- train test split
- training set
- testing/ validation set
- overfitting



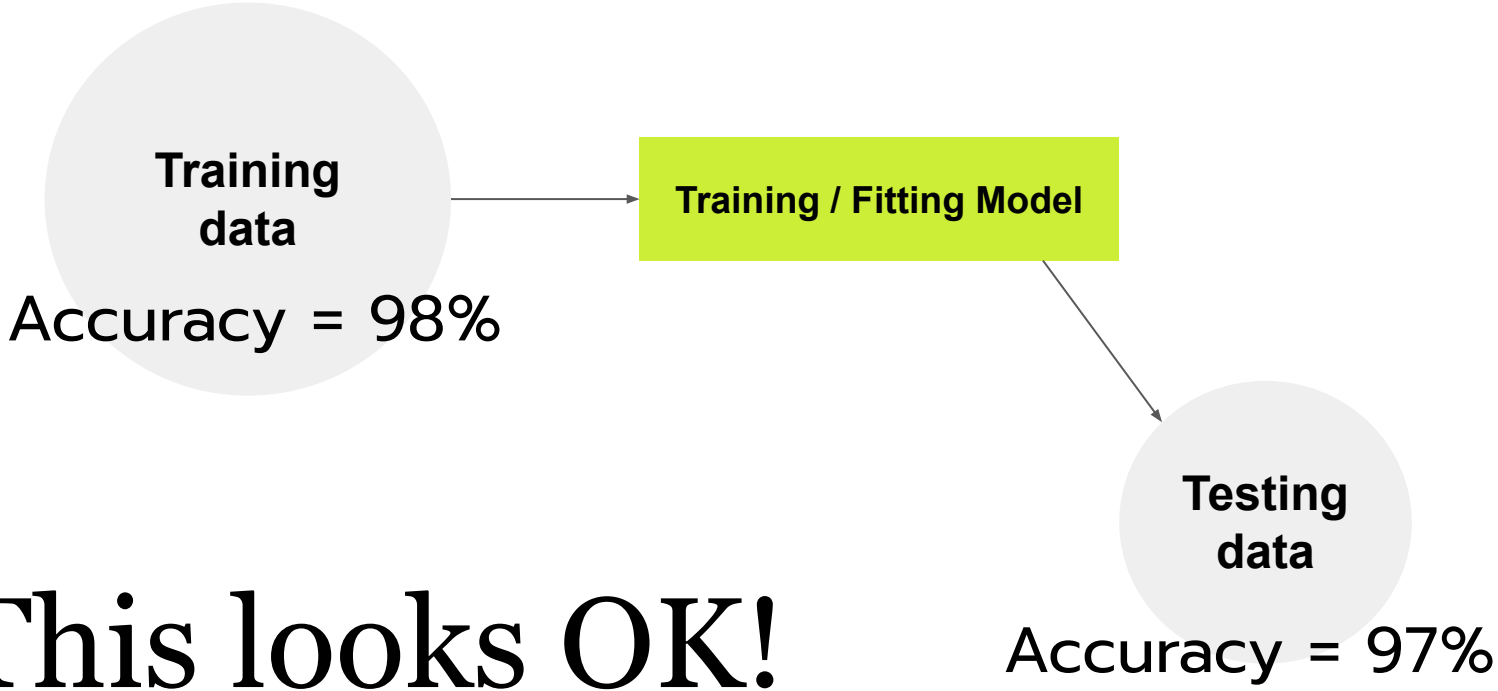
Train test split



เราต้องถามคำถามนี้เสมอ
โมเดลที่เราสร้างขึ้นมาเอาไปใช้จริงได้หรือเปล่า? i.e.
ความถูกต้องของโมเดลกับ test data เป็นเท่าไร



กรณีนี้เรียกว่า **Overfitting**
โมเดลที่เราสร้างขึ้นมาฟิตกับข้อมูล Training มากเกินไปจนไม่สามารถนำไปใช้กับ Testing/ Unseen data ได้



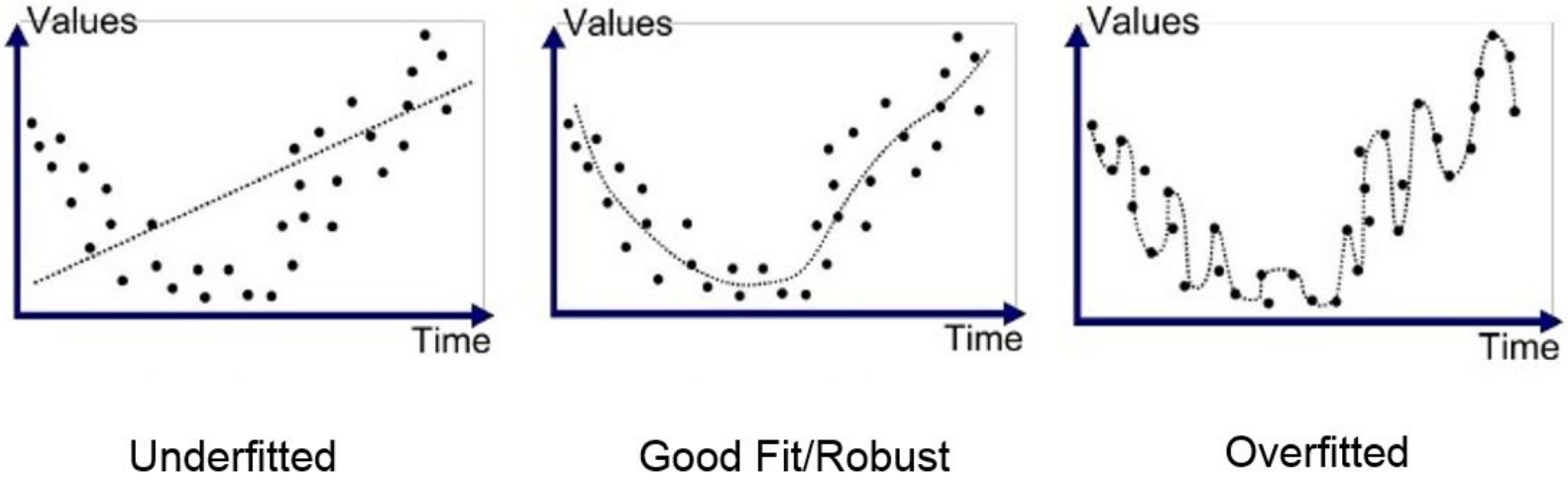
This looks OK!

Golden Rule

เราจะไม่ทดสอบโมเดลด้วยข้อมูลชุดเดิมที่ใช้เทรนโมเดล

i.e. เราจะไม่ใช้ training data วัดผลว่าโมเดลของเราทำงานดีไหม? แต่ต้องเป็น unseen data ที่โมเดลไม่เคยเห็นมาก่อน

Our goal is in the middle -> Just Right



https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Discuss: Overfitting คืออะไร?

เขียนคำตอบได้ที่นี่

Discuss: แล้วถ้า Underfitting ล่ะ?

เขียนคำตอบได้ที่นี่

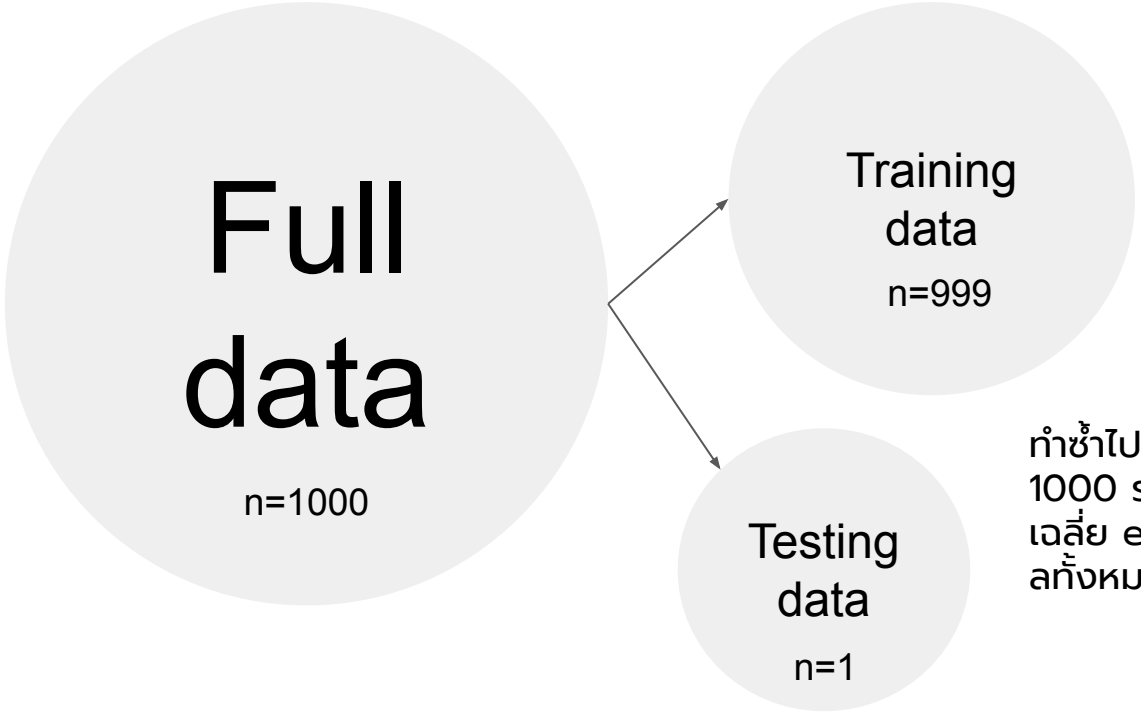
ในทางปฏิบัติ Train Test Split (ส่วนมาก) จะไม่ใช่วิธีที่ดีที่สุดในการสร้างโมเดล ML

เราใช้เทคนิคที่เรียกว่า **Resampling** สำหรับเทรนโมเดลเพื่อผลลัพธ์ที่ดีกว่า

ML Glossary 03

- resampling
  - leave one out CV
  - bootstrap
  - k-fold cross validation

Leave One Out CV



ทำซ้ำไปเรื่อยๆจนกว่าจะเทรนโมเดลครบ 1000 รอบ (ตามจำนวน n) แล้วหาค่าเฉลี่ย error หรือ accuracy ของโมเดลทั้งหมด

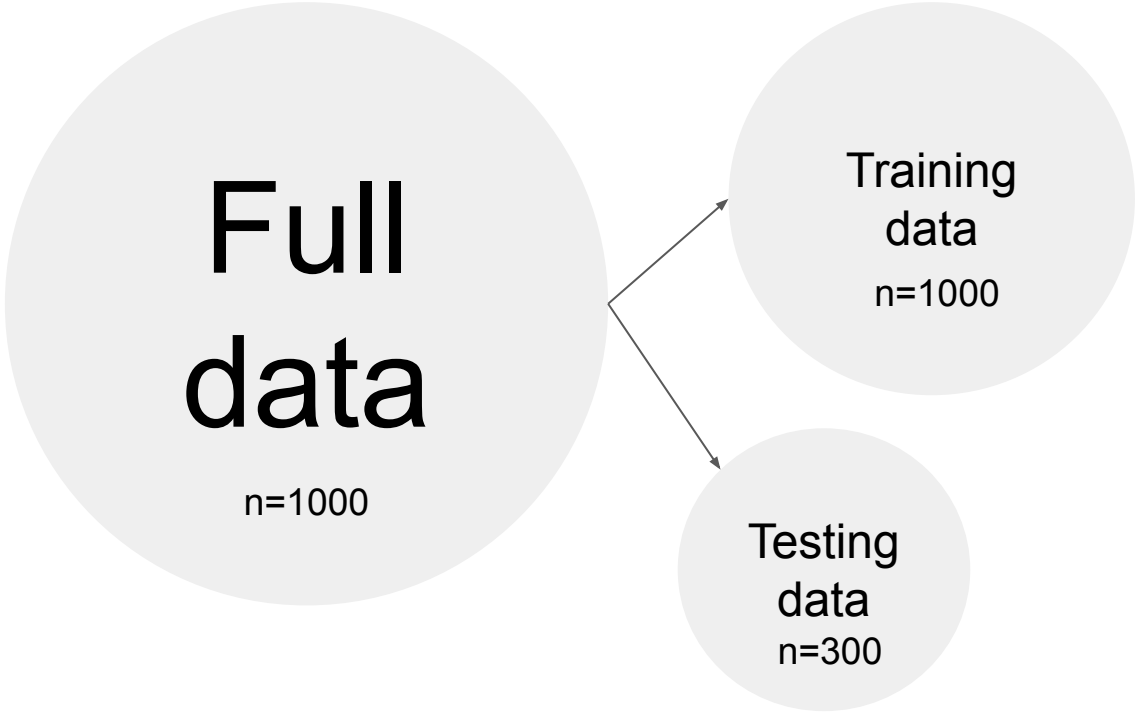
Leave One Out CV

1	2	3	4	...	...	997	998	999	1000	iteration 1
1	2	3	4	...	...	997	998	999	1000	iteration 2
1	2	3	4	...	...	997	998	999	1000	iteration 3
1	2	3	4	...	...	997	998	999	1000	iteration 4
1	2	3	4	...	...	997	998	999	1000	iteration 5

■ ■ ■

1	2	3	4	...	...	997	998	999	1000	iteration 999
1	2	3	4	...	...	997	998	999	1000	iteration 1000

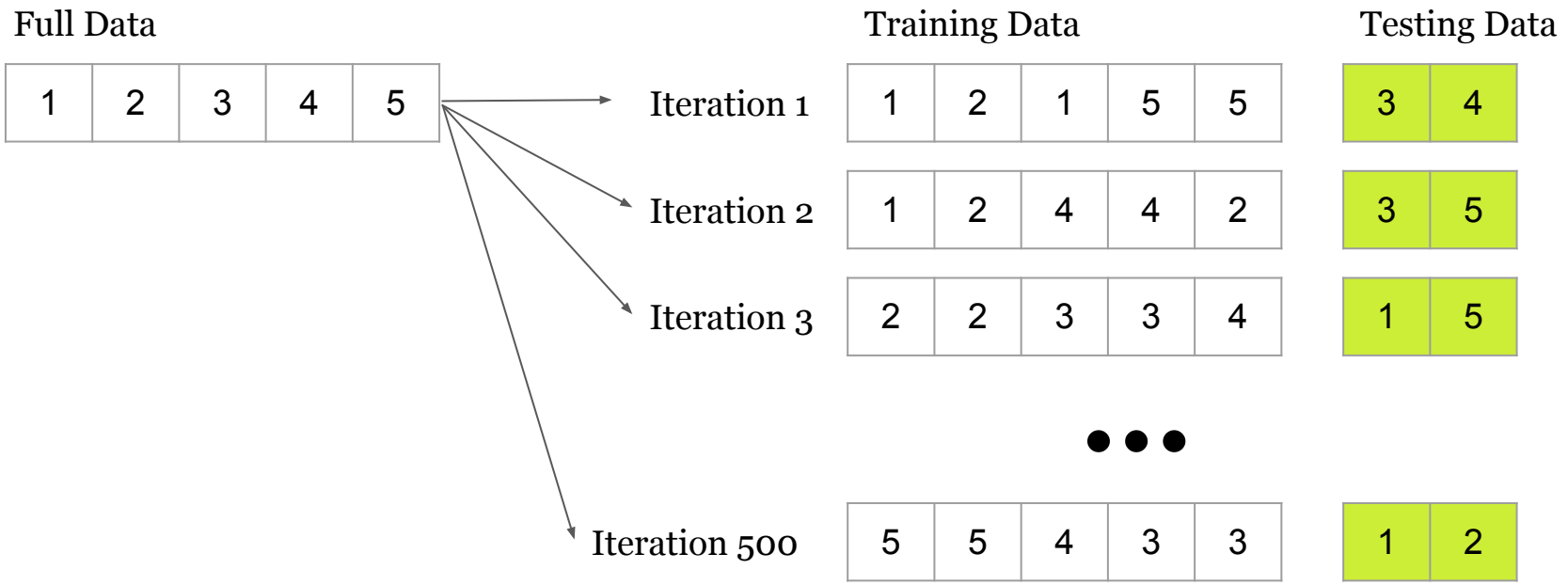
Bootstrap



Sampling with replacement

ใช้การสุ่มซ้ำ n=1000 เหมือน full dataset

Bootstrap



The error will be averaged over 500 training iterations

K-Fold Cross Validation

1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5
1	2	3	4	5

iteration 1: train {2,3,4,5} test {1} -> error 18%

iteration 2: train {1,3,4,5} test {2} -> error 20%

iteration 3: train {1,2,4,5} test {3} -> error 30%

iteration 4: train {1,2,3,5} test {4} -> error 15%

iteration 5: train {1,2,3,4} test {5} -> error 19%

Average error = (18+20+30+15+19) / 5 = 20.4%

ปกติเรานิยมใช้ค่า **K=5** หรือ **K=10**

Discuss: LOOCV, Bootstrap, K-Fold ทั้งสามวิธีแตกต่างกันอย่างไร?

เขียนคำตอบได้ที่นี่

Essential ML

- OK what exactly is machine learning
- OK supervised vs. unsupervised
- OK regression vs. classification
- OK train test split vs. cross validation
- model selection + hyperparameter
- model evaluation