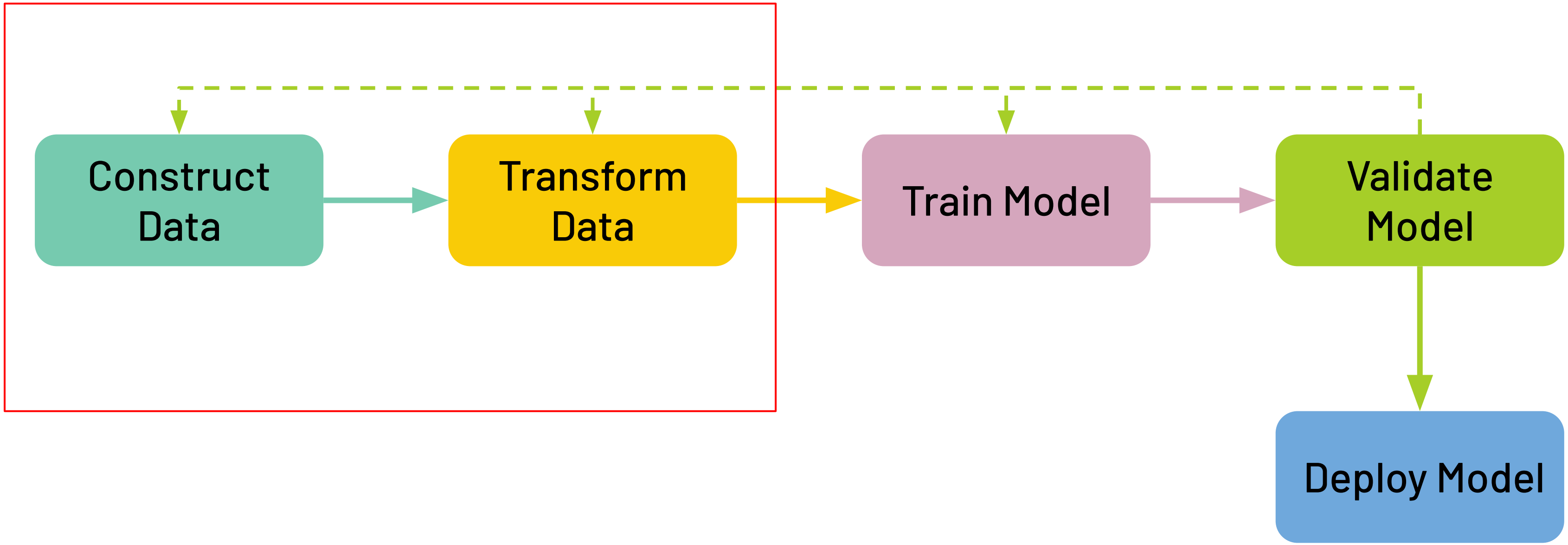
Data Preprocessing

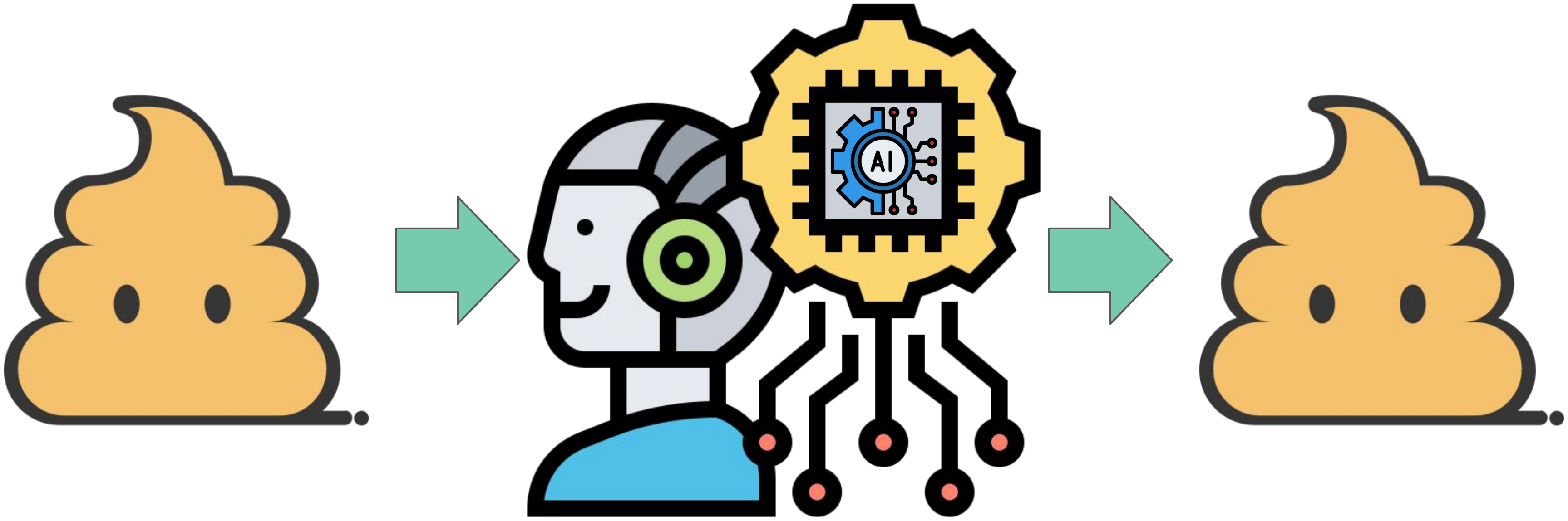
MACHINE LEARNING

Pakarat Musikawan

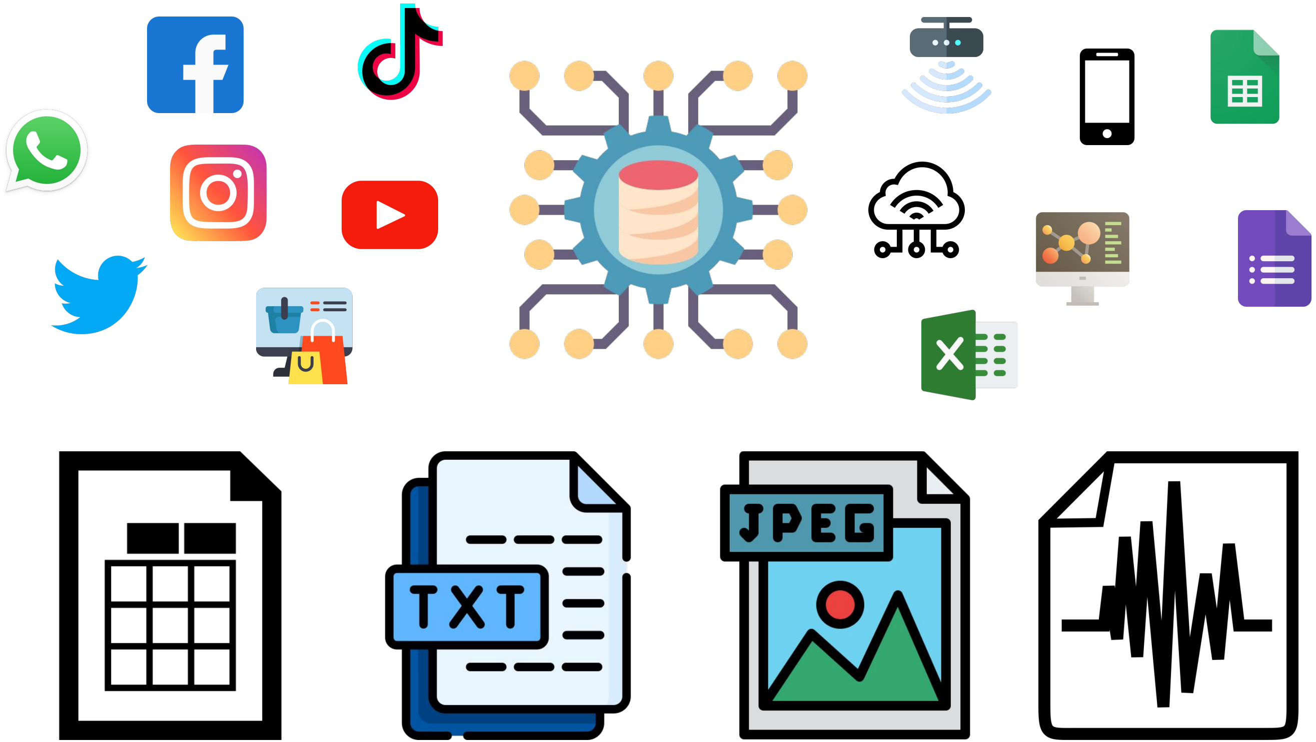
ML Pipeline


Construct Data
Transform Data
Train Model
Validate Model
Deploy Model

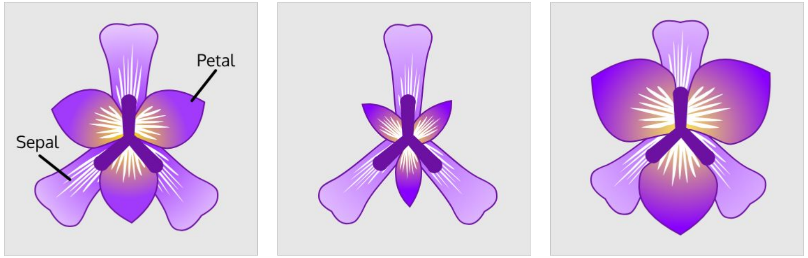
Garbage in, garbage out



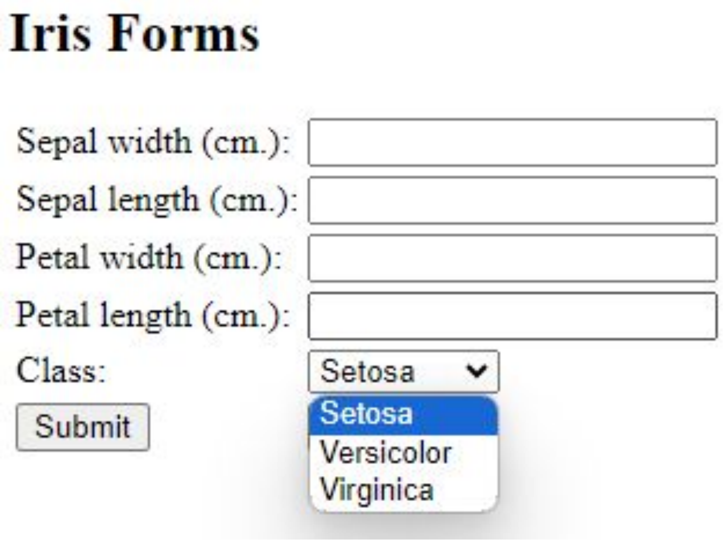
Data Collection



Data Collection



Iris Versicolor Iris Setosa Iris Virginica

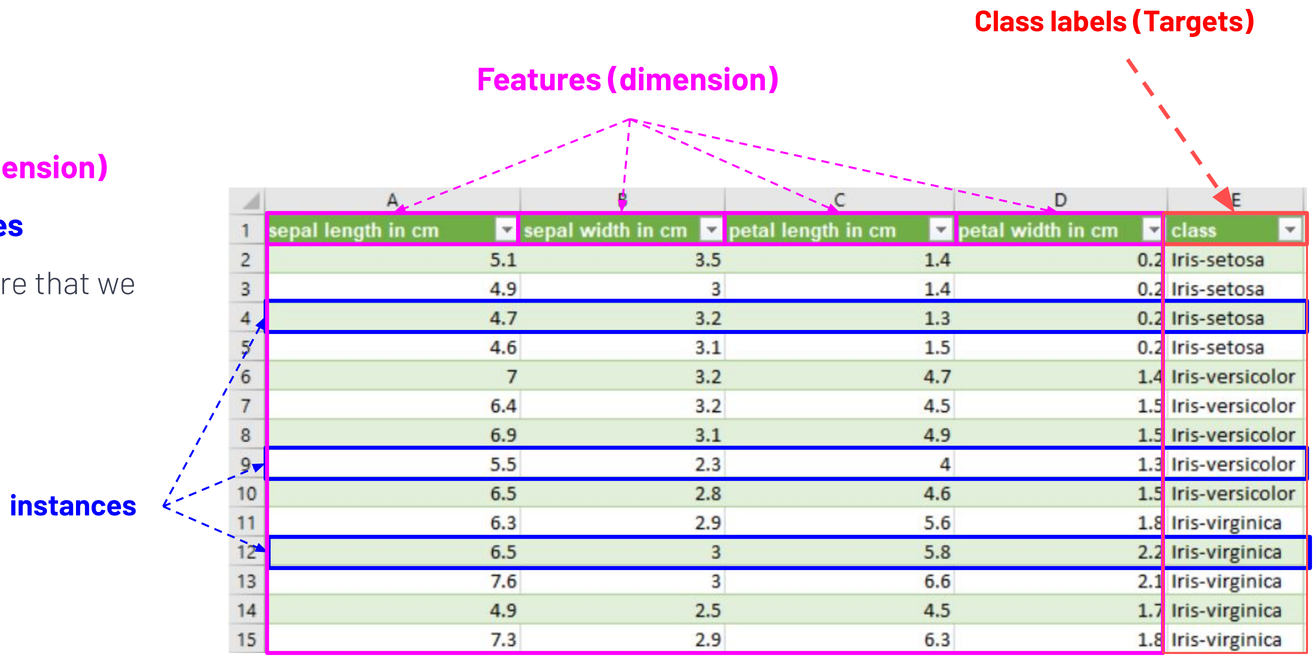


Data Collection: Tabular

	A	B	C	D	E
1	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	7	3.2	4.7	1.4	Iris-versicolor
7	6.4	3.2	4.5	1.5	Iris-versicolor
8	6.9	3.1	4.9	1.5	Iris-versicolor
9	5.5	2.3	4	1.3	Iris-versicolor
10	6.5	2.8	4.6	1.5	Iris-versicolor
11	6.3	2.9	5.6	1.8	Iris-virginica
12	6.5	3	5.8	2.2	Iris-virginica
13	7.6	3	6.6	2.1	Iris-virginica
14	4.9	2.5	4.5	1.7	Iris-virginica
15	7.3	2.9	6.3	1.8	Iris-virginica

Data Collection: Tabular

- Columns denote **feature (dimension)**
- Rows denote labeled **instances**
- **Class label (Target)** is a feature that we want to predict



	A	B	C	D	E
1	sepal length in cm	sepal width in cm	petal length in cm	petal width in cm	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	7	3.2	4.7	1.4	Iris-versicolor
7	6.4	3.2	4.5	1.5	Iris-versicolor
8	6.9	3.1	4.9	1.5	Iris-versicolor
9	5.5	2.3	4	1.3	Iris-versicolor
10	6.5	2.8	4.6	1.5	Iris-versicolor
11	6.3	2.9	5.6	1.8	Iris-virginica
12	6.5	3	5.8	2.2	Iris-virginica
13	7.6	3	6.6	2.1	Iris-virginica
14	4.9	2.5	4.5	1.7	Iris-virginica
15	7.3	2.9	6.3	1.8	Iris-virginica

Data Collection: JSON

```
{
    {
        “Id”: 1,
        “Name”: “Pizza Hut”,
        “Type”: “R”
    },
{
        “Id”: 2,
        “Name”: “Sears Tower”,
        “Type”: “A”
    },
}
```

Id	Name	Type
1	Pizza Hut	R
2	Sears Tower	A

Data Collection: CSV

"sepal.length","sepal.width","petal.length","petal.width","variety"

5.1,3.5,1.4,.2,"Setosa"

4.9,3,1.4,.2,"Setosa"

4.7,3.2,1.3,.2,"Setosa"

4.6,3.1,1.5,.2,"Setosa"

5,3.6,1.4,.2,"Setosa"

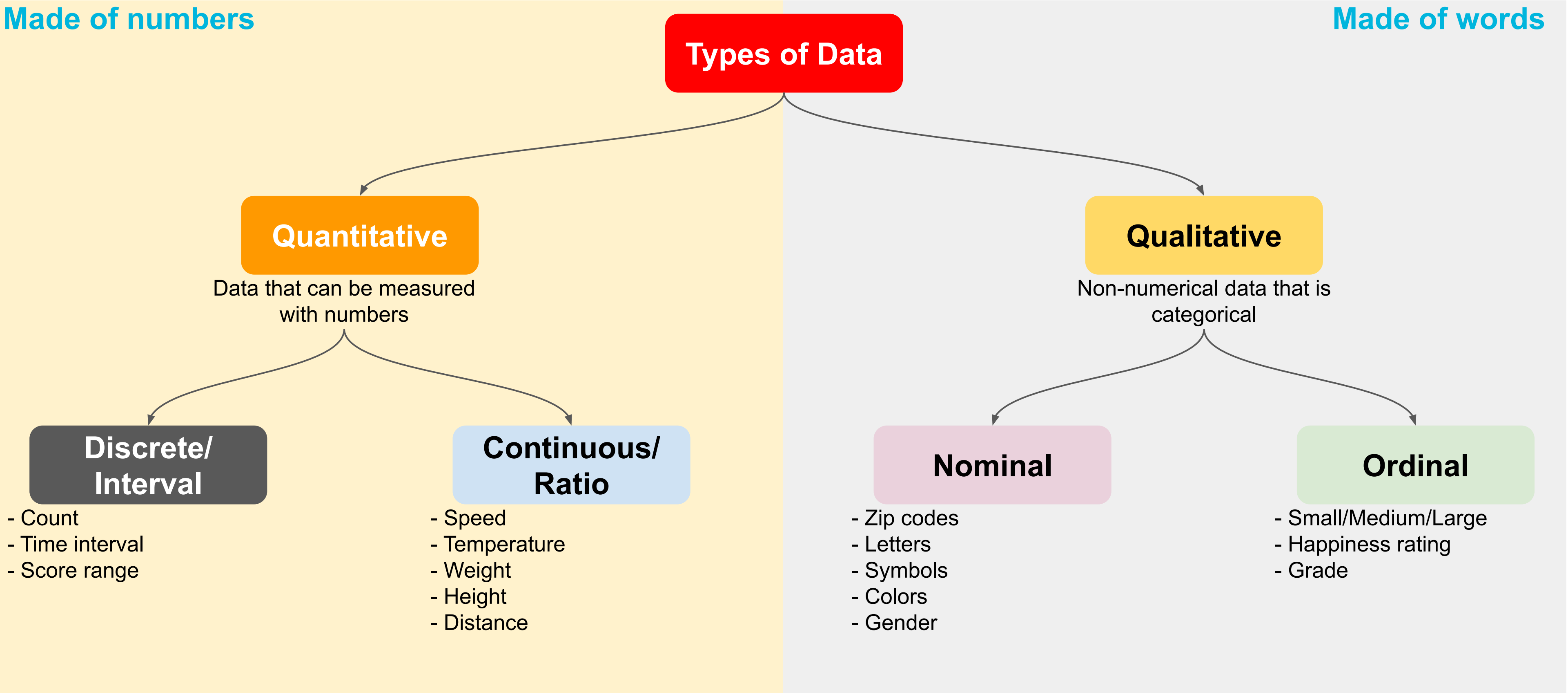
5.4,3.9,1.7,.4,"Setosa"

4.6,3.4,1.4,.3,"Setosa"

5,3.4,1.5,.2,"Setosa"

sepal.length	sepal.width	petal.length	petal.width	variety
5.1	3.5	1.4	.2	Setosa
4.9	3	1.4	.2	Setosa
4.7	3.2	1.3	.2	Setosa
4.6	3.1	1.5	.2	Setosa
5	3.6	1.4	.2	Setosa
5.4	3.9	1.7	.4	Setosa
4.6	3.4	1.4	.3	Setosa
5	3.4	1.5	.2	Setosa

Types of Data



Data Quality

Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data

Tid	Redund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	NULL	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	100000K	Yes
6	No	Divorced	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	75K	No
9	No	Married	75K	No

Data Quality

Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data

Tid	Redund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	NULL	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	100000K	Yes
6	No	Divorced	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	75K	No
9	No	Married	75K	No

Data Quality

Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data

Tid	Redund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	NULL	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	100000K	Yes
6	No	Divorced	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	75K	No
9	No	Married	75K	No

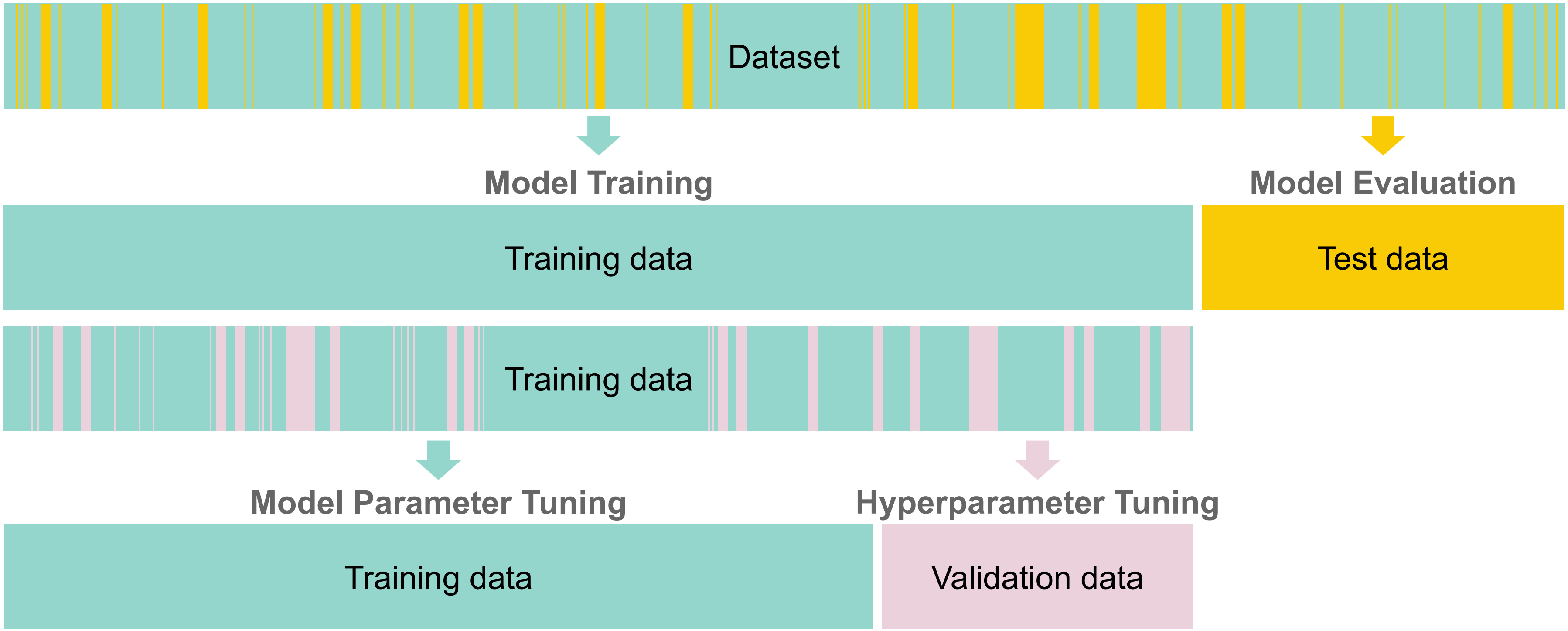
Data Quality

Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data

Tid	Redund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	NULL	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	100000K	Yes
6	No	Divorced	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	75K	No
9	No	Married	75K	No

Splitting Data



Splitting Data

	A	B	C	D	E	F	G	H
1	filename	width	height	class	xmin	ymin	xmax	ymax
2	00000022.	600	450	ak47	142	197	567	300
3	00000028.	600	439	ak47	251	11	388	392
4	00000030.	600	900	ak47	90	221	467	374
5	00000034.	500	389	ak47	56	42	444	322
6	00000038.	600	450	ak47	19	9	597	402
7	00000039.	600	600	ak47	160	240	380	382
8	00000039.	600	600	ak47	245	288	400	434
9	00000039.	600	600	ak47	6	160	367	381
10	00000052.	600	438	ak47	325	11	388	101
11	00000052.	600	438	ak47	383	1	435	191
12	00000055.	482	200	ak47	263	147	318	180
13	00000079.	480	480	ak47	2	332	480	436
14	00000079.	480	480	ak47	1	198	478	310
15	00000098.	240	240	ak47	5	94	235	147
16	00000099.	600	427	ak47	259	73	417	206
17	00000112.	600	800	ak47	175	258	494	503
18	00000112.	600	800	ak47	1	200	293	323
19	00000112.	600	800	ak47	379	293	545	587
20	00000121.	600	376	ak47	1	46	599	259
21	00000122.	300	257	ak47	119	54	200	103
22	00000127.	380	570	ak47	195	218	372	570
23	00000130.	480	480	ak47	21	246	176	295
24	00000130.	480	480	ak47	11	12	194	70
25	00000130.	480	480	ak47	13	87	158	165
26	00000144.	600	450	ak47	21	19	597	359
27	00000147.	360	170	ak47	8	59	344	163
28	00000151.	600	337	ak47	1	43	305	301
29	00000163.	600	963	ak47	238	423	419	869
30	00000169.	480	480	ak47	4	132	478	330

Dataset

2	00000022.	600	450	ak47	142	197	567	300
3	00000028.	600	439	ak47	251	11	388	392
4	00000030.	600	900	ak47	90	221	467	374
5	00000034.	500	389	ak47	56	42	444	322
6	00000038.	600	450	ak47	19	9	597	402
7	00000039.	600	600	ak47	160	240	380	382
8	00000039.	600	600	ak47	245	288	400	434
9	00000039.	600	600	ak47	6	160	367	381
10	00000052.	600	438	ak47	325	11	388	101
11	00000052.	600	438	ak47	383	1	435	191
12	00000055.	482	200	ak47	263	147	318	180
13	00000079.	480	480	ak47	2	332	480	436
14	00000079.	480	480	ak47	1	198	478	310
15	00000098.	240	240	ak47	5	94	235	147
16	00000099.	600	427	ak47	259	73	417	206
17	00000112.	600	800	ak47	175	258	494	503
18	00000112.	600	800	ak47	1	200	293	323
19	00000112.	600	800	ak47	379	293	545	587
20	00000121.	600	376	ak47	1	46	599	259
21	00000122.	300	257	ak47	119	54	200	103
22	00000127.	380	570	ak47	195	218	372	570
23	00000130.	480	480	ak47	21	246	176	295
24	00000130.	480	480	ak47	11	12	194	70
25	00000130.	480	480	ak47	13	87	158	165

Training data

26	00000144.	600	450	ak47	21	19	597	359
27	00000147.	360	170	ak47	8	59	344	163
28	00000151.	600	337	ak47	1	43	305	301
29	00000163.	600	963	ak47	238	423	419	869
30	00000169.	480	480	ak47	4	132	478	330

Test data

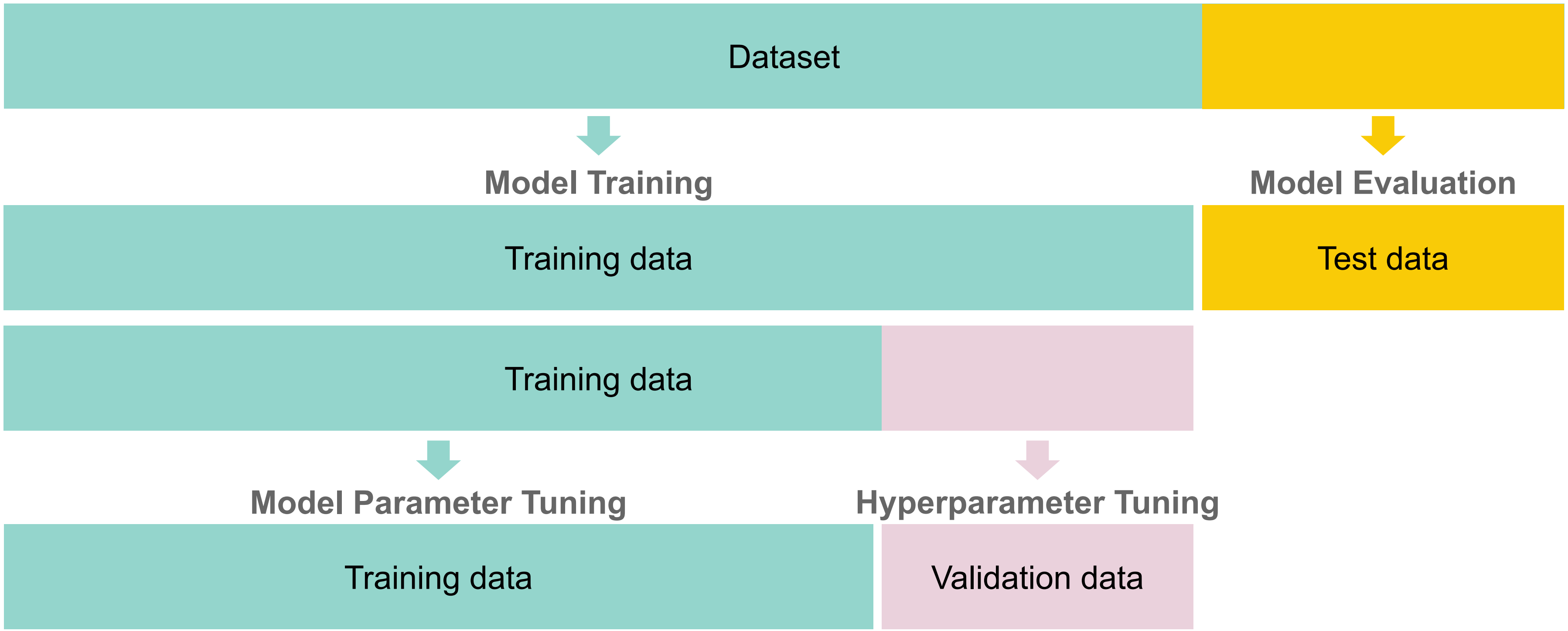
2	00000022.	600	450	ak47	142	197	567	300
3	00000028.	600	439	ak47	251	11	388	392
4	00000030.	600	900	ak47	90	221	467	374
5	00000034.	500	389	ak47	56	42	444	322
6	00000038.	600	450	ak47	19	9	597	402
7	00000039.	600	600	ak47	160	240	380	382
8	00000039.	600	600	ak47	245	288	400	434
9	00000039.	600	600	ak47	6	160	367	381
10	00000052.	600	438	ak47	325	11	388	101
11	00000052.	600	438	ak47	383	1	435	191
12	00000055.	482	200	ak47	263	147	318	180
13	00000079.	480	480	ak47	2	332	480	436
14	00000079.	480	480	ak47	1	198	478	310
15	00000098.	240	240	ak47	5	94	235	147
16	00000099.	600	427	ak47	259	73	417	206
17	00000112.	600	800	ak47	175	258	494	503
18	00000112.	600	800	ak47	1	200	293	323
19	00000112.	600	800	ak47	379	293	545	587
20	00000121.	600	376	ak47	1	46	599	259

Training data

21	00000122.	300	257	ak47	119	54	200	103
22	00000127.	380	570	ak47	195	218	372	570
23	00000130.	480	480	ak47	21	246	176	295
24	00000130.	480	480	ak47	11	12	194	70
25	00000130.	480	480	ak47	13	87	158	165

Validation data

Splitting Data - Time Series



Data Transformation

Label encoding

iris-setosa	0
iris-versicolor	1
iris-setosa	0
iris-virginica	2
iris-versicolor	1
iris-versicolor	1
iris-virginica	2

Data Transformation

One-Hot encoding

iris-setosa
iris-versicolor
iris-setosa
iris-virginica
iris-versicolor
iris-versicolor
iris-virginica

iris-setosa	iris-versicolor	iris-virginica
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	1

Data Transformation

One-Hot encoding

Patient_ID	Gender	BP(S)	BP(D)	Heart Rate	Temperature
001	Female	120	80	75	98.5
002	Female	125	82	70	98.7
003	Male	145	90	89	98.6
004	Male	140	87	92	98.5

Patient_ID	Female	Male	BP(S)	BP(D)	Heart Rate	Temperature
001	1	0	120	80	75	98.5
002	1	0	125	82	70	98.7
003	0	1	145	90	89	98.6
004	0	1	140	87	92	98.5

Data Transformation

One-Hot encoding

Patient_ID	Gender	BP(S)	BP(D)	Heart Rate	Temperature
001	Female	120	80	75	98.5
002	Female	125	82	70	98.7
003	Male	145	90	89	98.6
004	Male	140	87	92	98.5

Patient_ID	Gender	BP(S)	BP(D)	Heart Rate	Temperature
001	0	120	80	75	98.5
002	0	125	82	70	98.7
003	1	145	90	89	98.6
004	1	140	87	92	98.5

0: Female
1: Male

Data Transformation

Scaling

$$\boldsymbol{X}_i' = L + \frac{\boldsymbol{X}_i - \min(\boldsymbol{X}_i)}{\max(\boldsymbol{X}_i) - \min(\boldsymbol{X}_i)} \times (U - L)$$

$\boldsymbol{X}_i$	the i-th original column data
$\boldsymbol{X}_i'$	the i-th scaled column data
$\max(\boldsymbol{X}_i)$	the maximum value of the i-th column of the training data
$\min(\boldsymbol{X}_i)$	the minimum value of the i-th column of the training data
L	the lower bound of the desired range
U	the upper bound of the desired range

Data Transformation

Scaling

$$X'_i = L + \frac{X_i - \min(X_i)}{\max(X_i) - \min(X_i)} \times (U - L)$$

$$L = 0, U = 1$$

Age	Loan
25	40000
35	60000
45	80000
20	20000
35	120000
52	18000
23	95000
40	62000
60	100000
48	220000
33	150000
48	142000

$$0 + \frac{25 - 20}{60 - 20} \times (1 - 0) = 0.125$$

$$0 + \frac{142000 - 18000}{220000 - 18000} \times (1 - 0) = 0.614$$

Age	Loan
0.125	0.109
0.375	0.208
0.625	0.307
0.000	0.010
0.375	0.505
0.800	0.000
0.075	0.381
0.500	0.218
1.000	0.406
0.700	1.000
0.325	0.653
0.700	0.614

Data Transformation

Normalization

$$\boldsymbol{X}'_i = \frac{\boldsymbol{X}_i - \mu_i}{\sigma_i}$$

$\boldsymbol{X}_i$ the i-th original column data

$\boldsymbol{X}'_i$ the i-th normalized column data

μ_i the average of the i-th column of the training data

σ_i the standard deviation of the i-th column of the training data

$$\mu_i = \frac{\sum_{j=1}^{N} x_{j,i}}{N} = \frac{x_{1,i} + x_{2,i} + \ldots + x_{N,i}}{N}$$

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N} (x_{j,i} - \mu_i)^2}{N}} = \sqrt{\frac{(x_{1,i} - \mu_i)^2 + \ldots + (x_{N,i} - \mu_i)^2}{N}}$$

Data Transformation

Normalization

$$\boldsymbol{X}'_i = \frac{\boldsymbol{X}_i - \mu_i}{\sigma_i}$$

Age	Loan
25	40000
35	60000
45	80000
20	20000
35	120000
52	18000
23	95000
40	62000
60	100000
48	220000
33	150000
48	142000

$$\mu_1 = \frac{25 + 35 + \ldots + 33 + 48}{12} = 38.66$$

$$\sigma_1 = \sqrt{\frac{(25 - 38.66)^2 + \ldots + (48 - 38.66)^2}{12}} = 12.39$$

$$\mu_2 = \frac{40000 + 6000 + \ldots + 150000 + 142000}{12} = 92250$$

$$\sigma_2 = \sqrt{\frac{(40000 - 92250)^2 + \ldots + (40000 - 92250)^2}{12}} = 59188.64$$

Data Transformation

Normalization

$$X'_i = \frac{X_i - \mu_i}{\sigma_i}$$

Age	Loan
25	40000
35	60000
45	80000
20	20000
35	120000
52	18000
23	95000
40	62000
60	100000
48	220000
33	150000
48	142000

$$\frac{25 - 38.66}{12.39} = -1.103$$

$$\frac{142000 - 92250}{59188.64} = 0.841$$

Age	Loan
-1.103	-0.883
-0.296	-0.545
0.511	-0.207
-1.507	-1.221
-0.296	0.469
1.076	-1.254
-1.264	0.046
0.108	-0.511
1.722	0.131
0.753	2.158
-0.457	0.976
0.753	0.841

Workshop

1 ทำการ Scaling ข้อมูล Training set และ Test set ให้อยู่ในช่วง [-1, 1]

แสดงค่า Min / Max ของทุกคอลัมน์

2 ทำการ Normalization ข้อมูล Training set และ Test set ที่กำหนดให้

แสดงค่า Average / SD ของทุกคอลัมน์

Training set

X1	X2	X3
189	19500	21
280	25000	34
159	19000	28
177	26000	25

Test set

X1	X2	X3
299	15000	35
90	27000	22
150	22000	15

แสดงการคำนวณวิธีละ 1 แถวข้อมูล