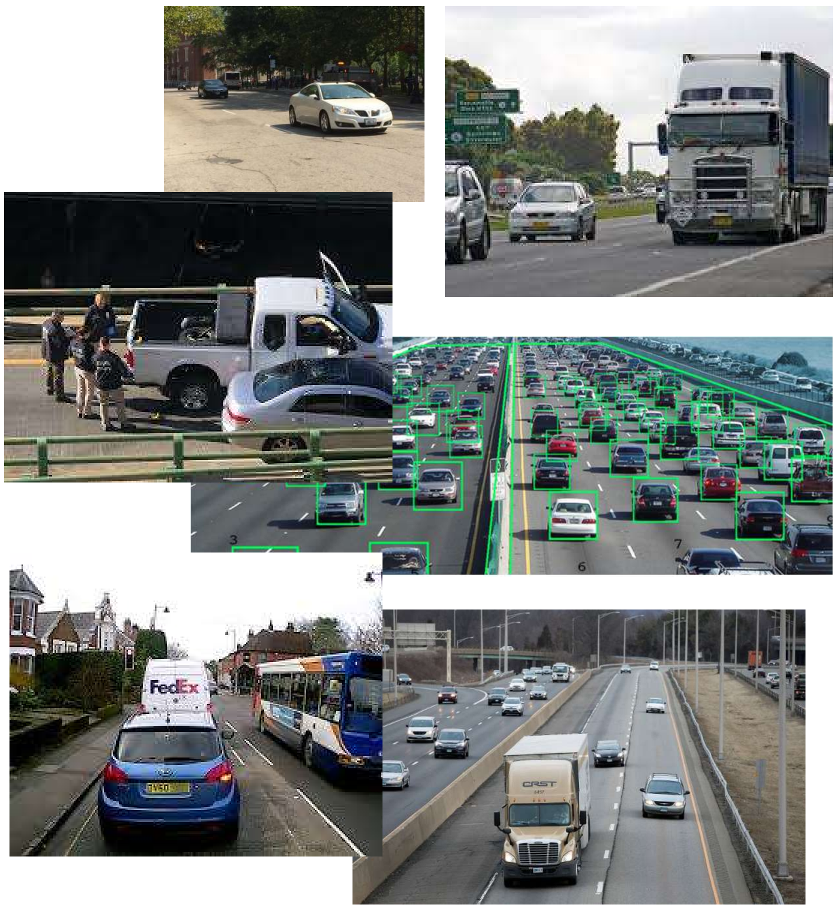
MACHINE LEARNING

ML WORKFLOW – STEPS AND PROCEDURES

WHY?

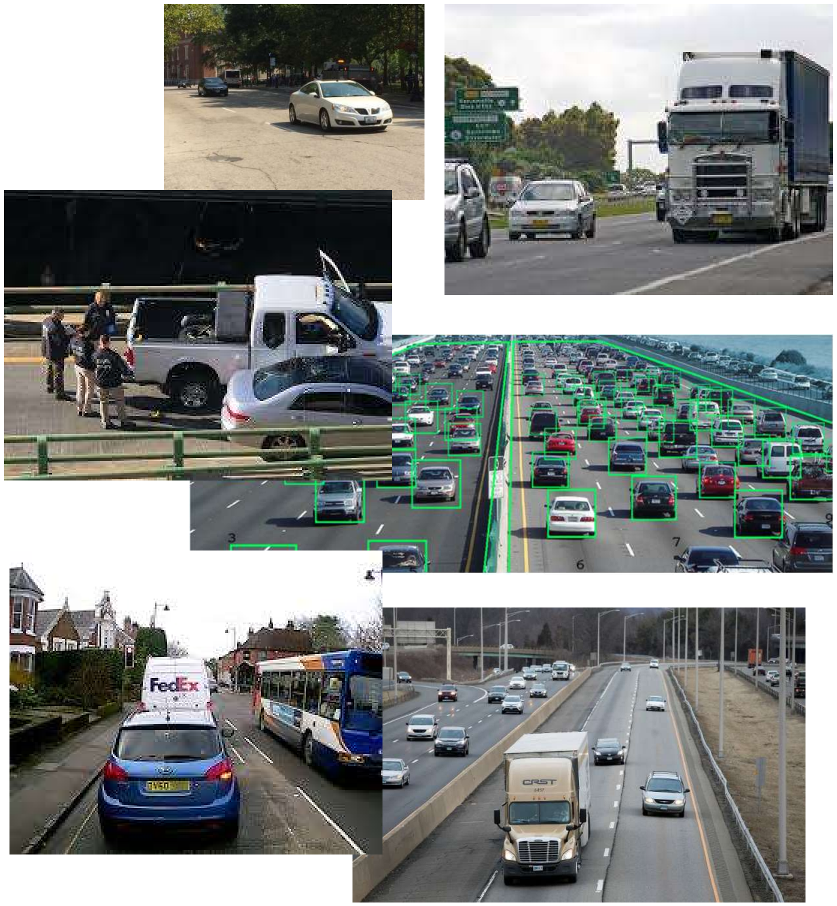
REAL LIFE EXAMPLE

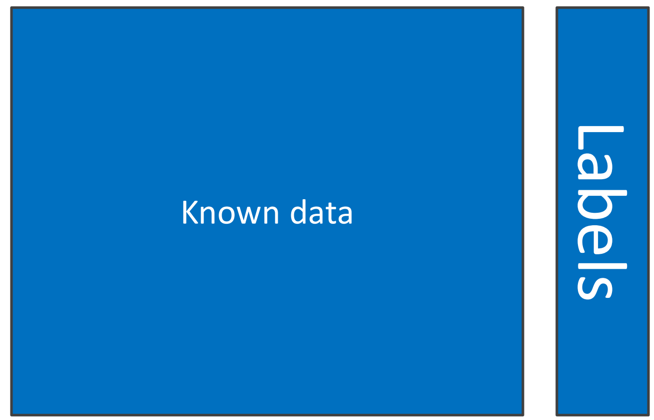
REAL LIFE EXAMPLE – TRUCK OR NO TRUCK



Data

REAL LIFE EXAMPLE – TRUCK OR NO TRUCK





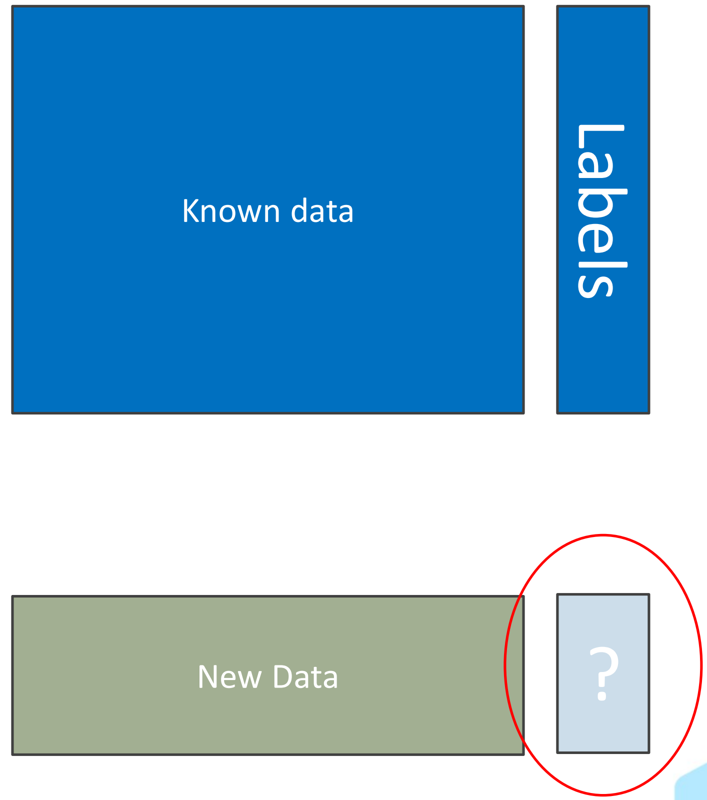
REAL LIFE EXAMPLE – TRUCK OR NO TRUCK

Known data

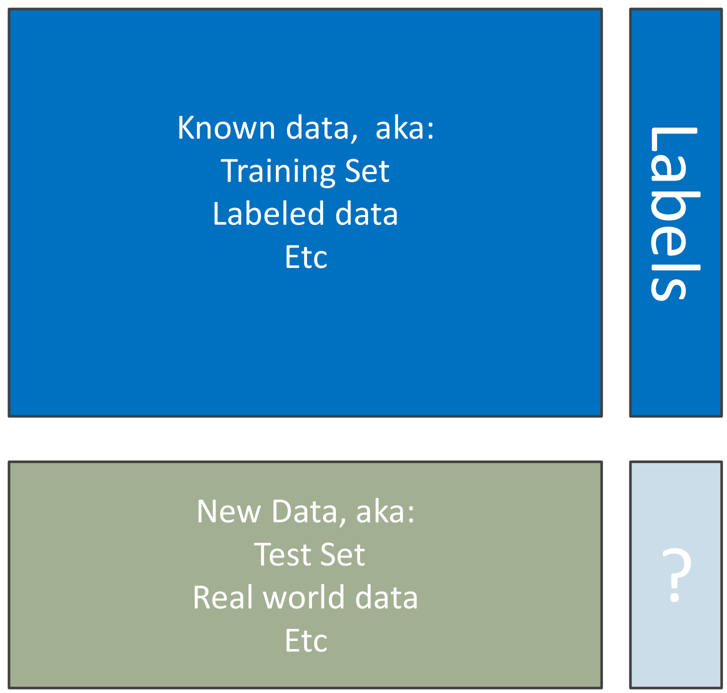
Labels

New Data

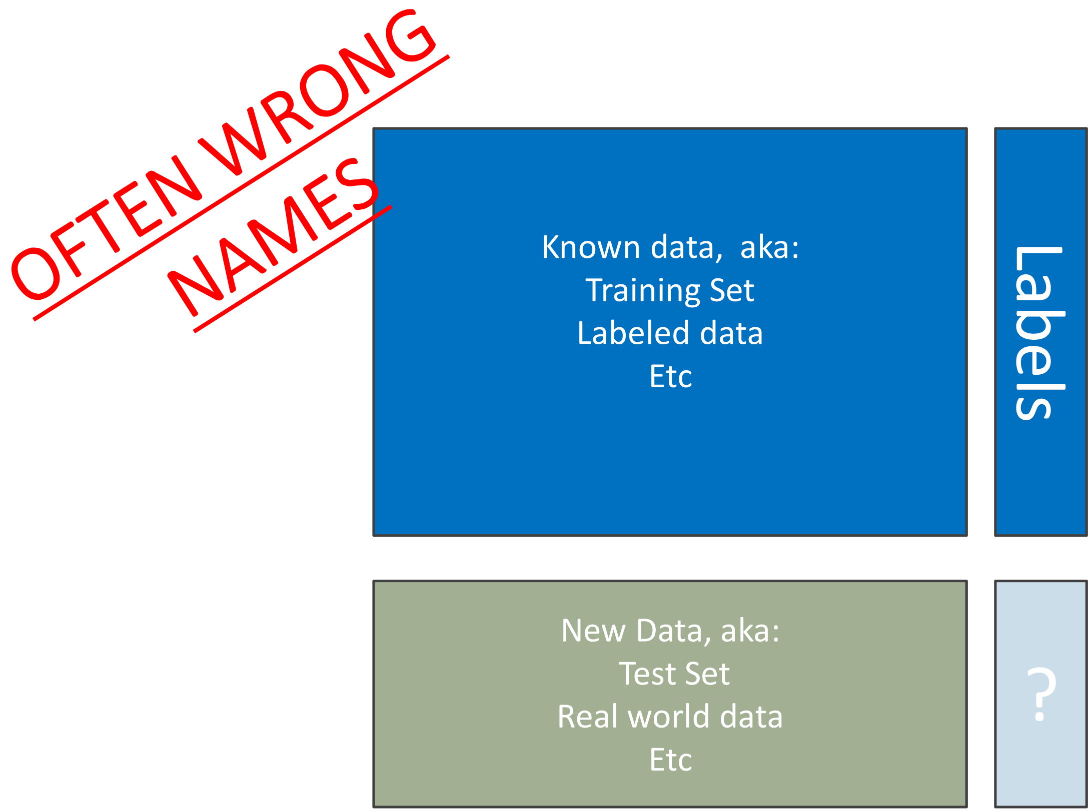
REAL LIFE EXAMPLE – TRUCK OR NO TRUCK



REAL LIFE EXAMPLE – TRUCK OR NO TRUCK

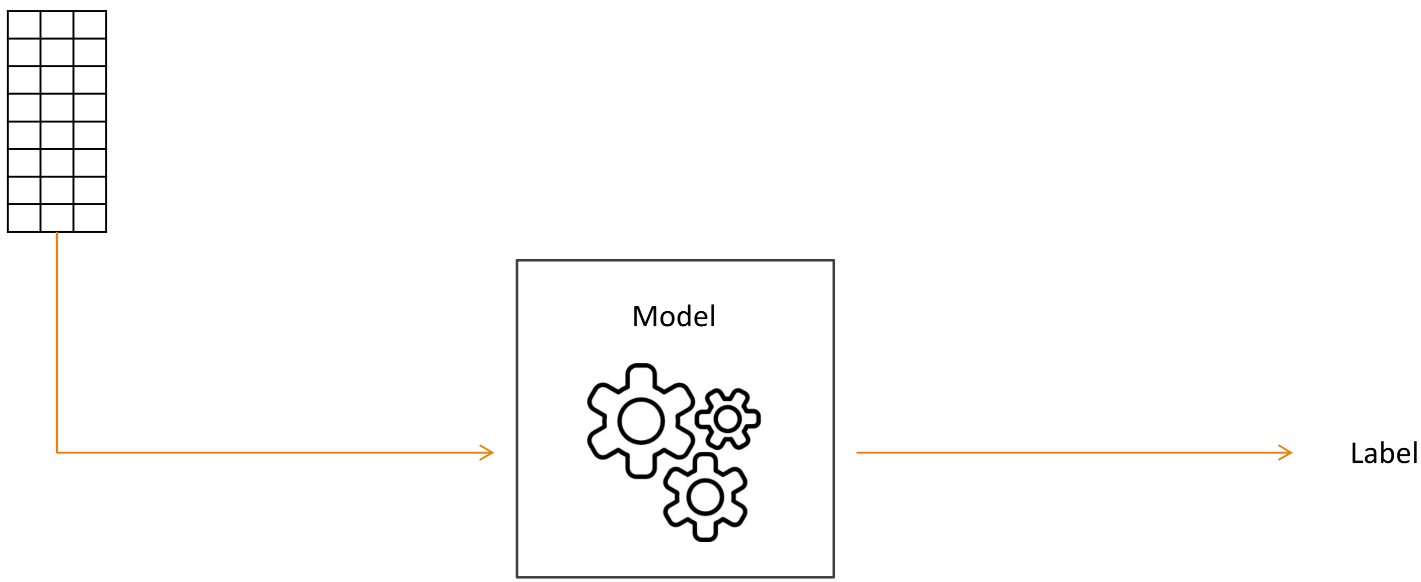


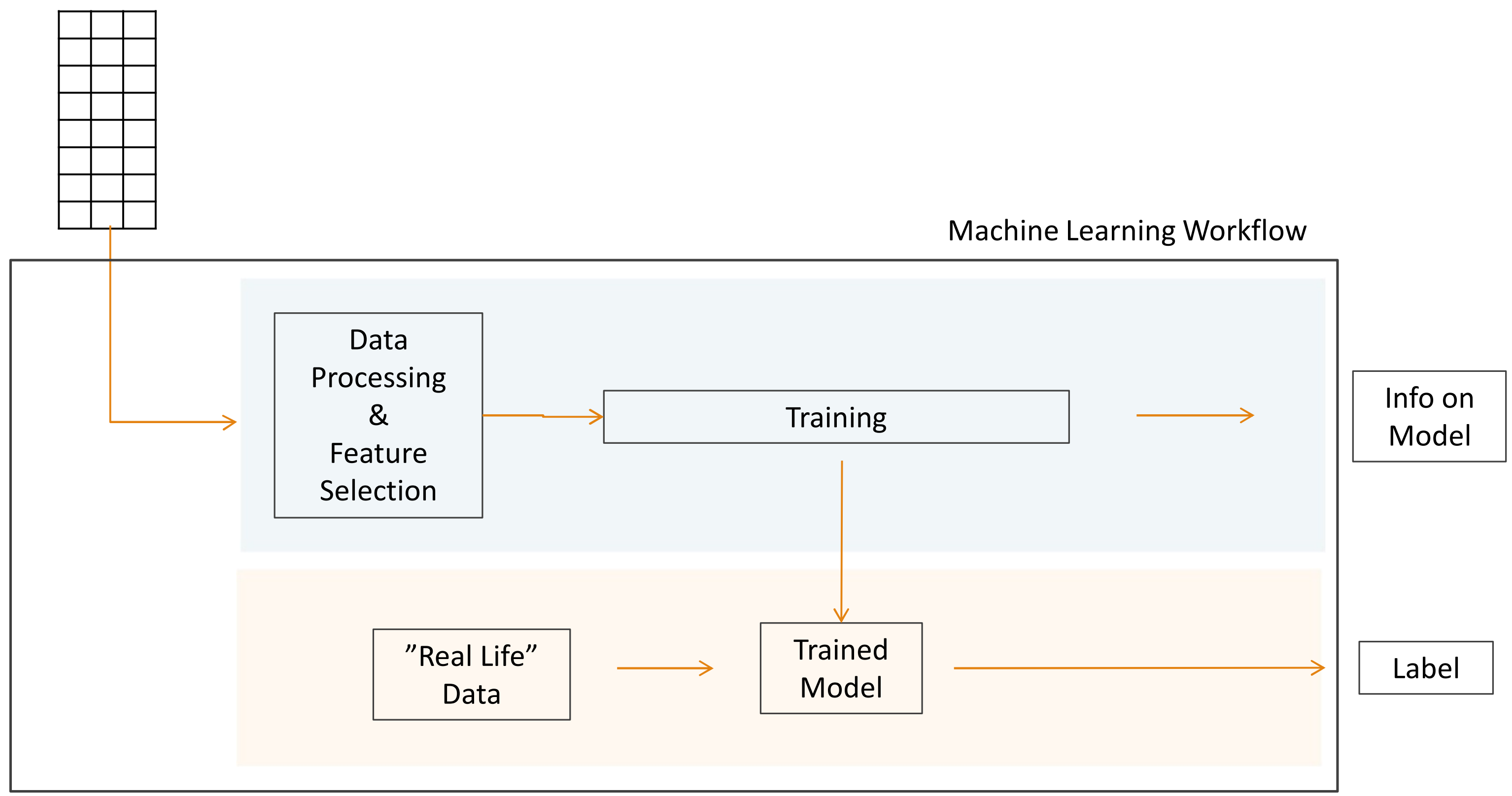
REAL LIFE EXAMPLE – TRUCK OR NO TRUCK

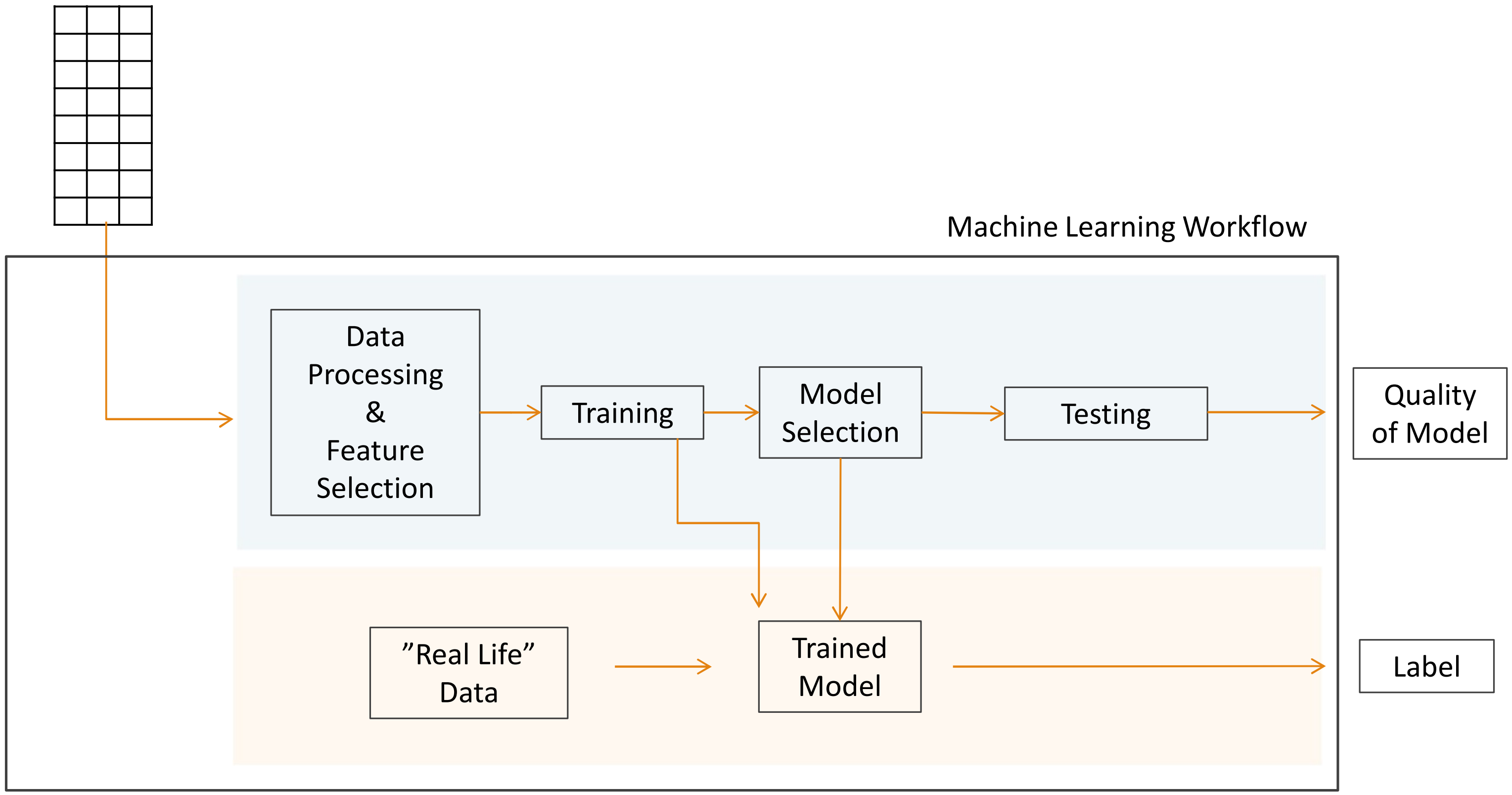


<u>HOW?</u>

MACHINE LEARNING PROCESS


Model
Label

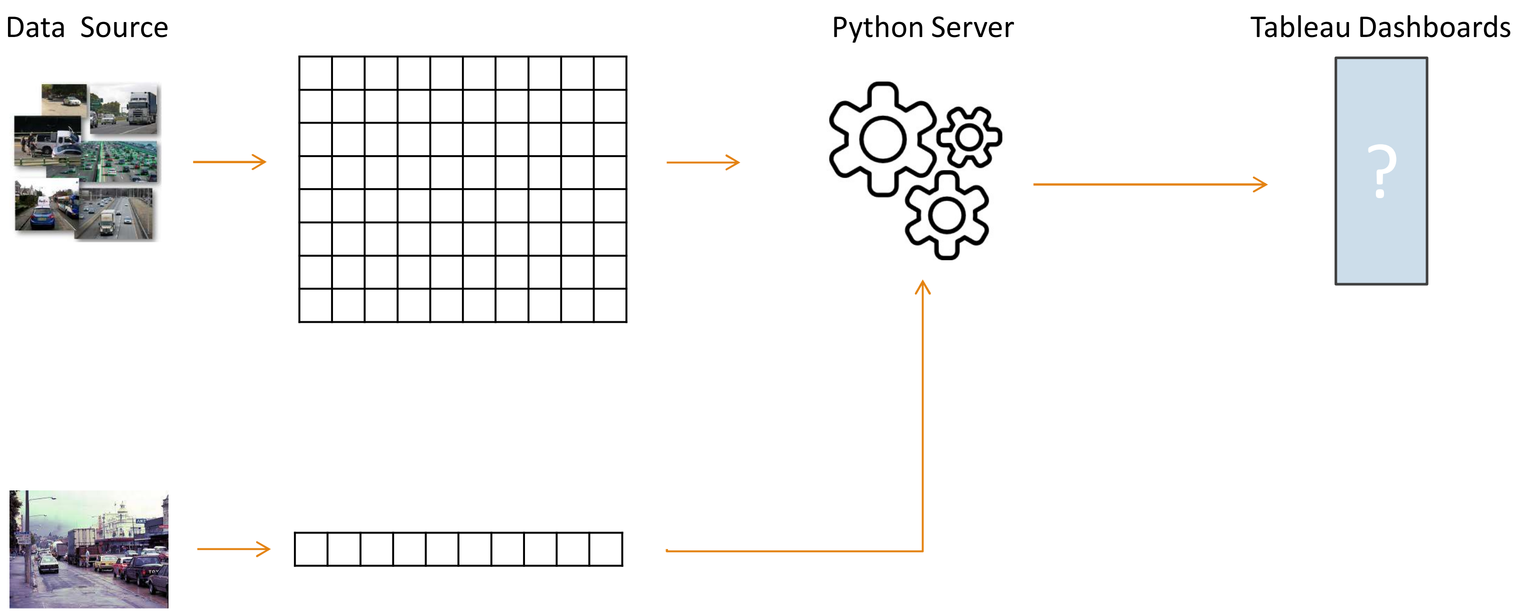

Machine Learning Workflow
Data Processing & Feature Selection
Training
Info on Model
"Real Life" Data
Trained Model
Label



Machine Learning Workflow
Data Processing & Feature Selection
Training
Model Selection
Testing
Quality of Model
"Real Life" Data
Trained Model
Label

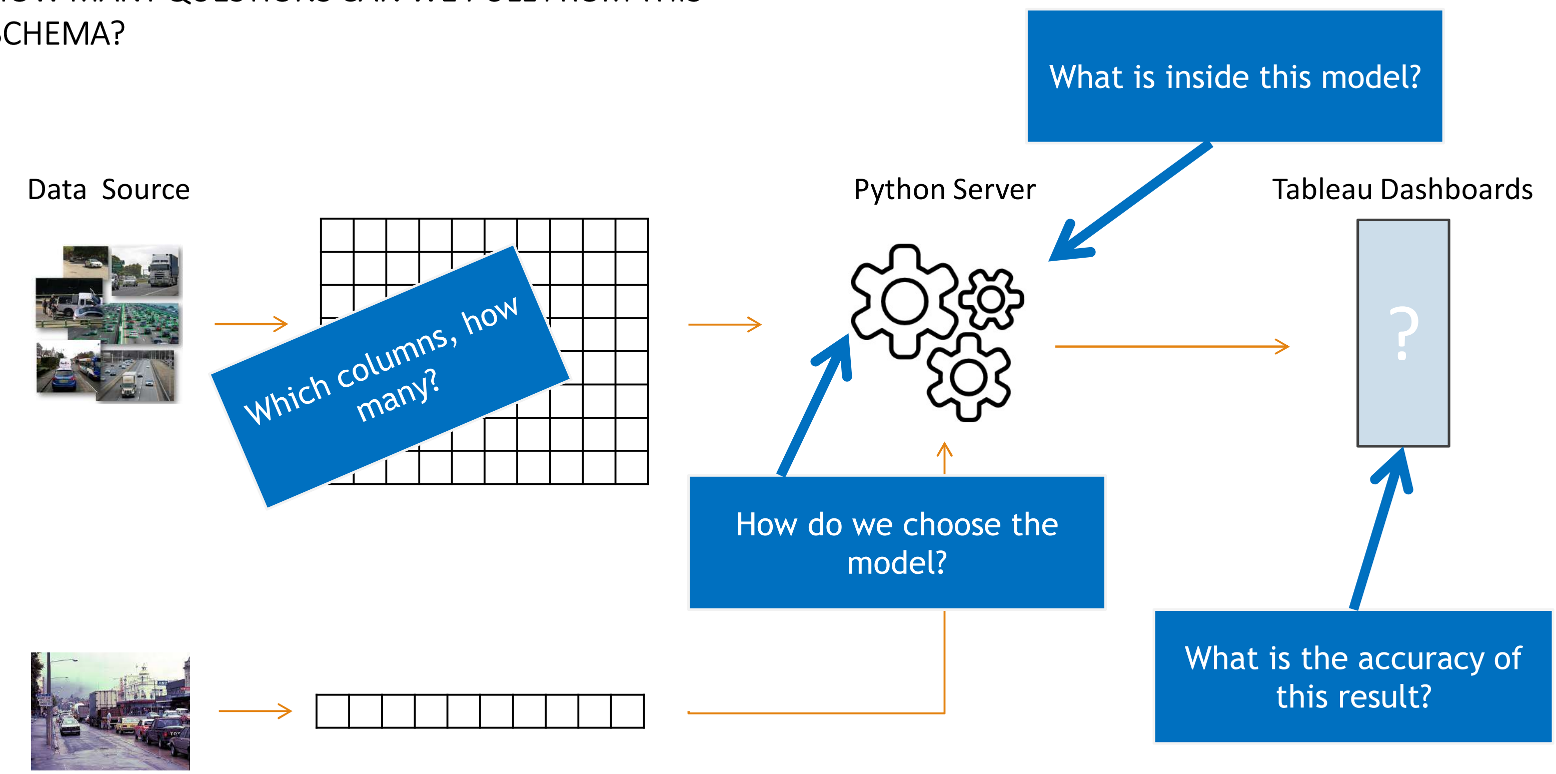
CLASSIFICATION IN MACHINE LEARNING

HOW MANY QUESTIONS CAN WE PULL FROM THIS SCHEMA?



CLASSIFICATION IN MACHINE LEARNING

HOW MANY QUESTIONS CAN WE PULL FROM THIS SCHEMA?



<u>WHAT?</u>

MACHINE LEARNING PROCESS

MACHINE LEARNING WORKFLOW OPERATIONS

K-fold cross validation

Mutual Information

Training Data

Cross validation

Test Set

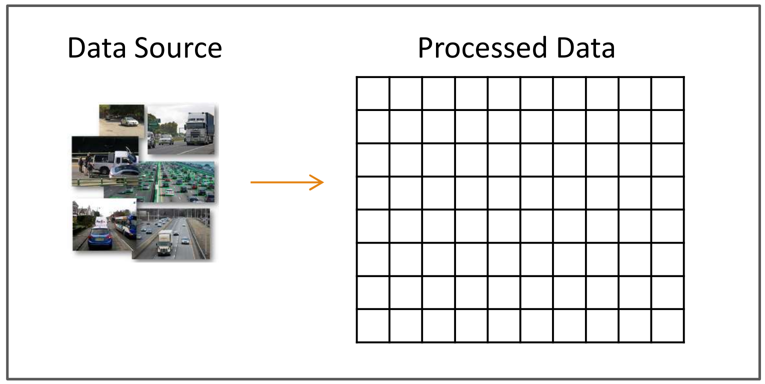
Feature Selection

Mutual Information

MACHINE LEARNING PROCESS

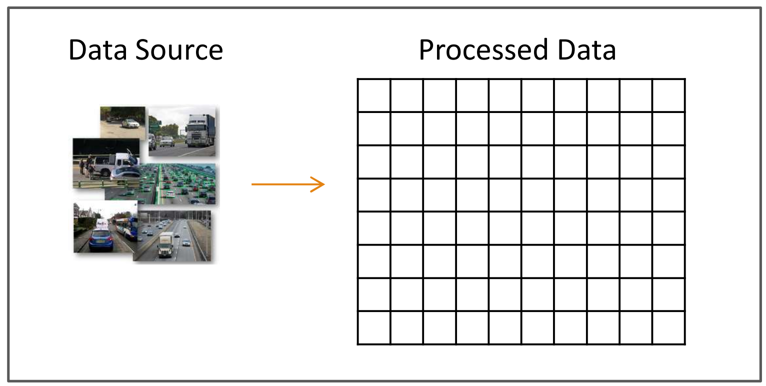
FEATURE SELECTION

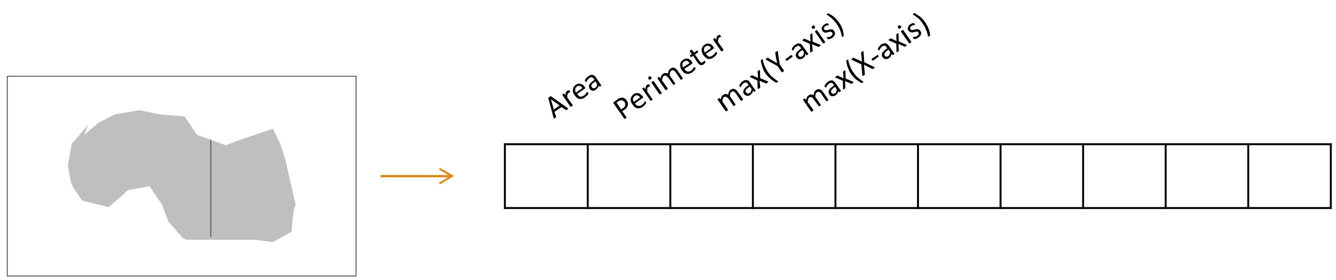
MACHINE LEARNING WORKFLOW OPERATIONS – FEATURE ANALYSIS



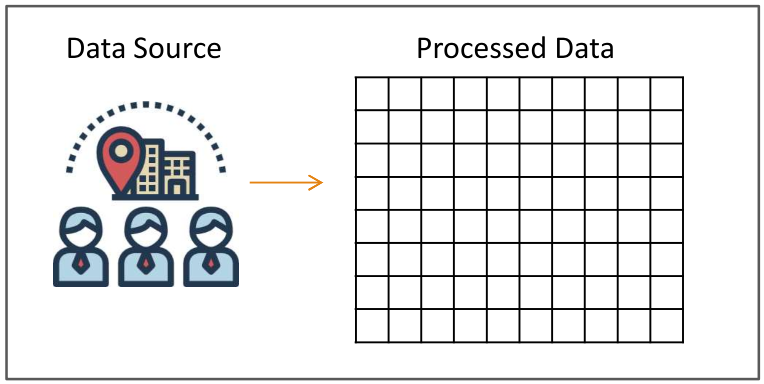
WHAT IS EACH COLUMN?

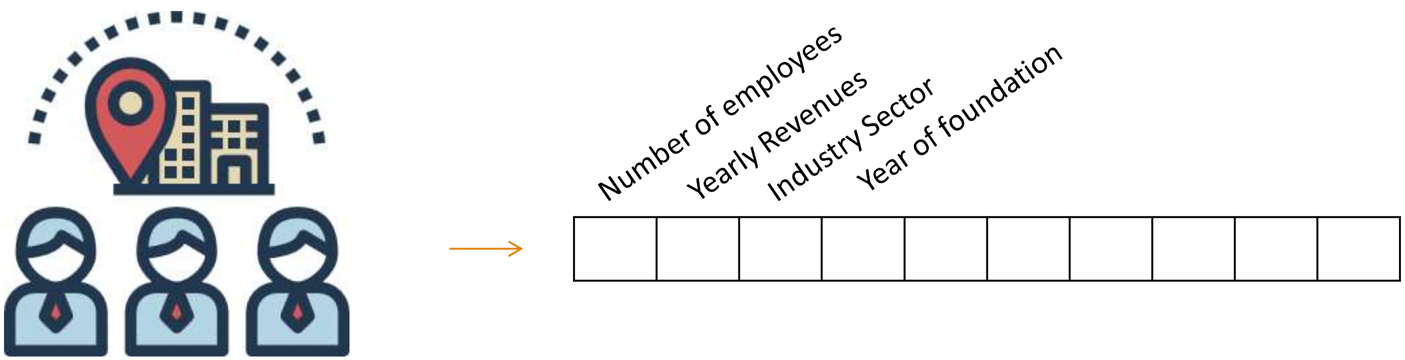
MACHINE LEARNING WORKFLOW OPERATIONS – FEATURE ANALYSIS





MACHINE LEARNING WORKFLOW OPERATIONS – FEATURE ANALYSIS





FEATURE SELECTION…

… MORE ON THIS LATER ON

MACHINE LEARNING PROCESS

TRAINING

MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING

KNN

Decision Trees

Random Forest

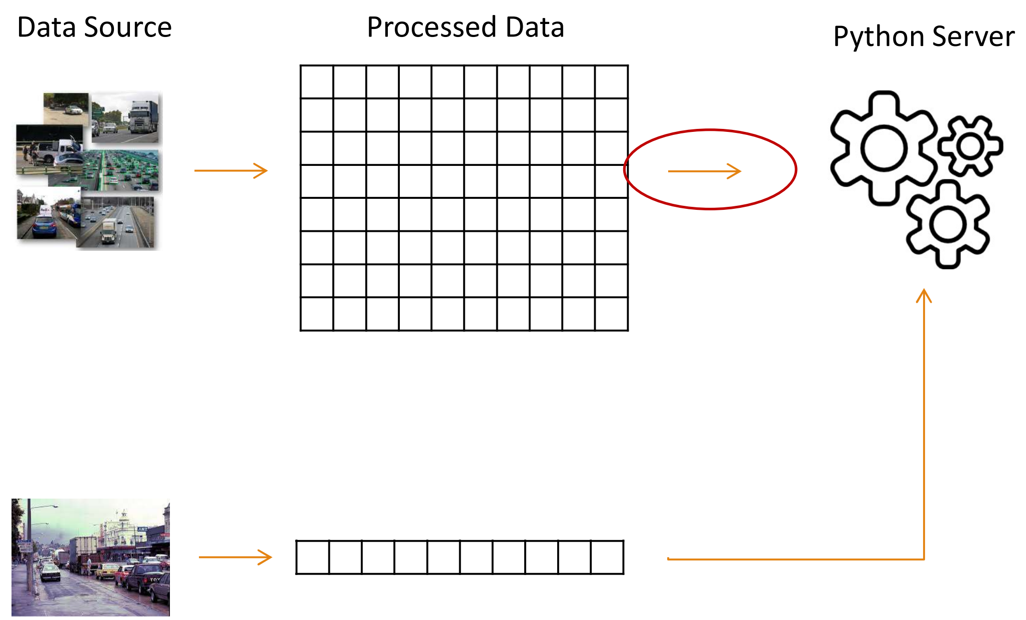
M.L.P.

Expectation-Maximization

Neural Networks

S.V.M.

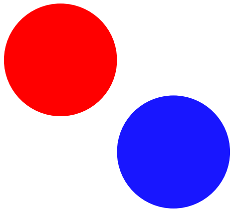
MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING



MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING

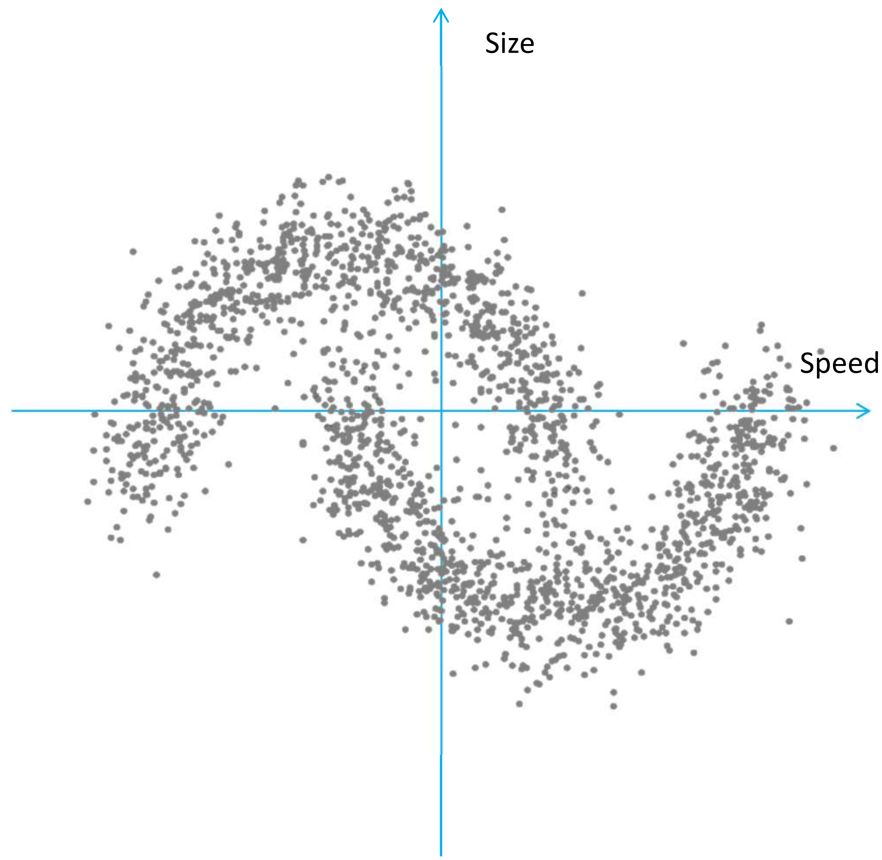
Known data

Labels

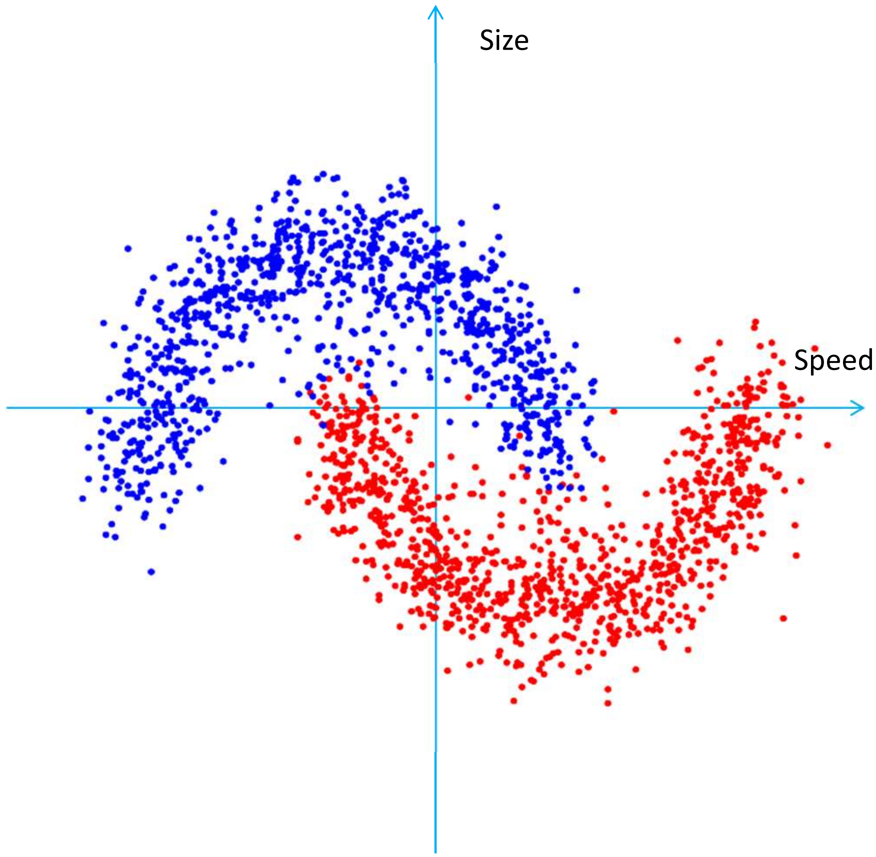


MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING

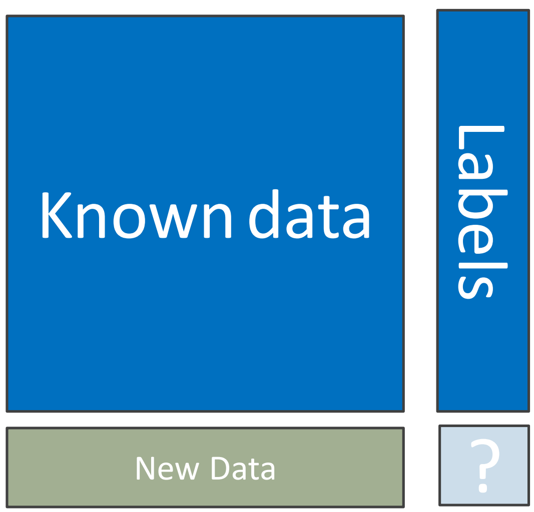
Known data

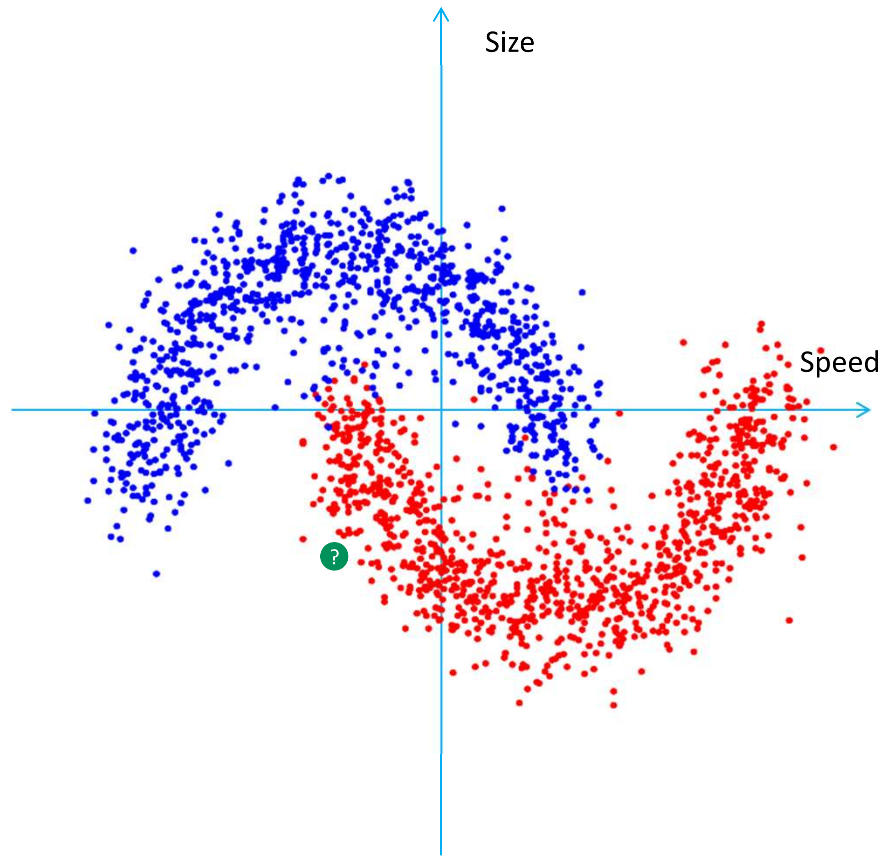


MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING

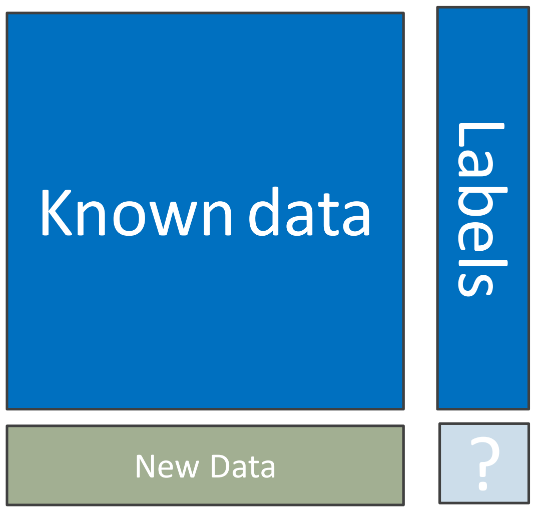


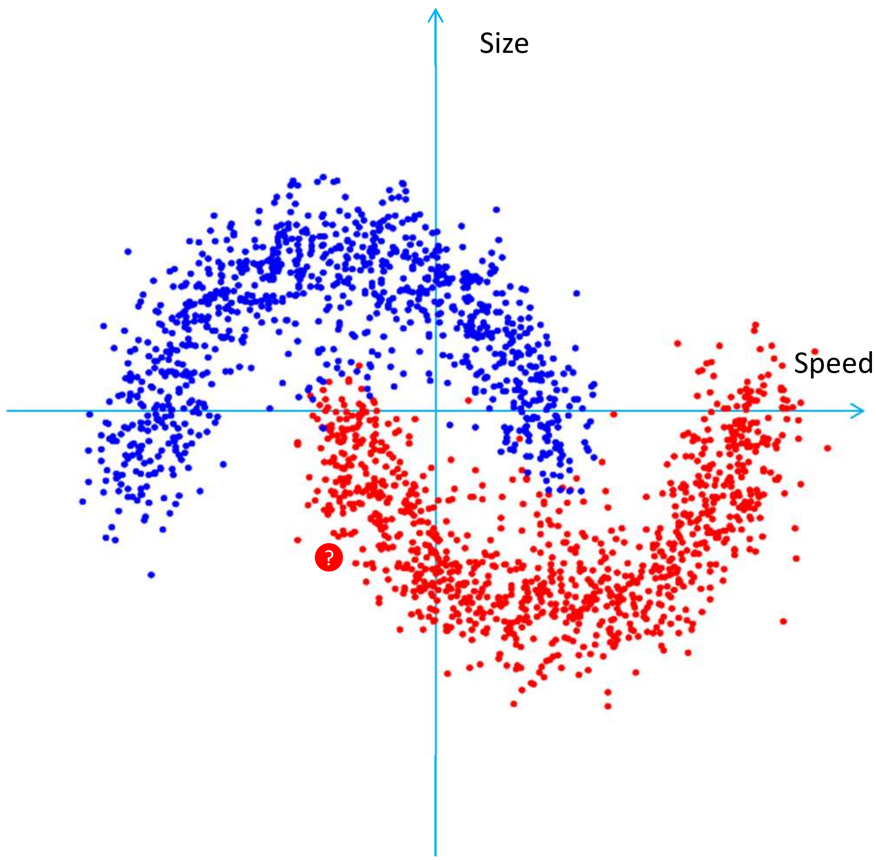
MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING



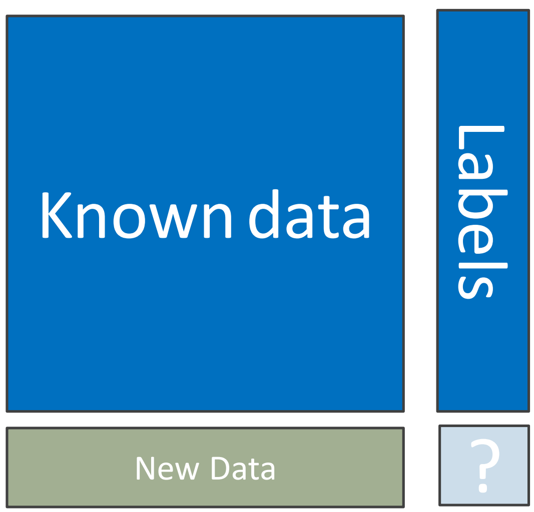


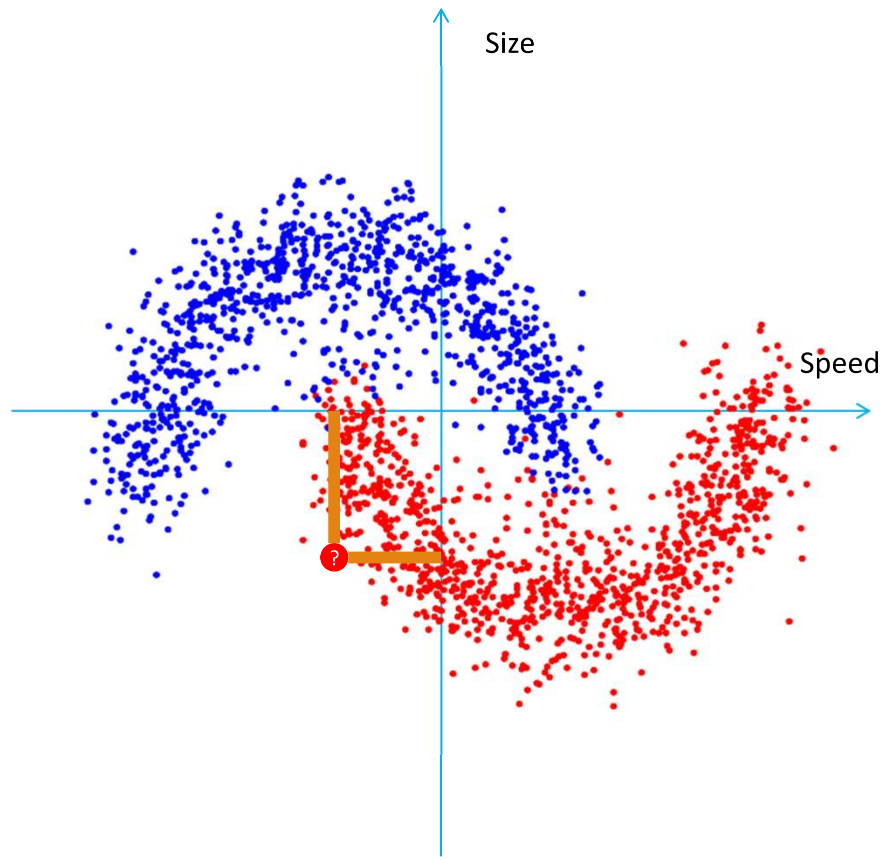
MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING





MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING





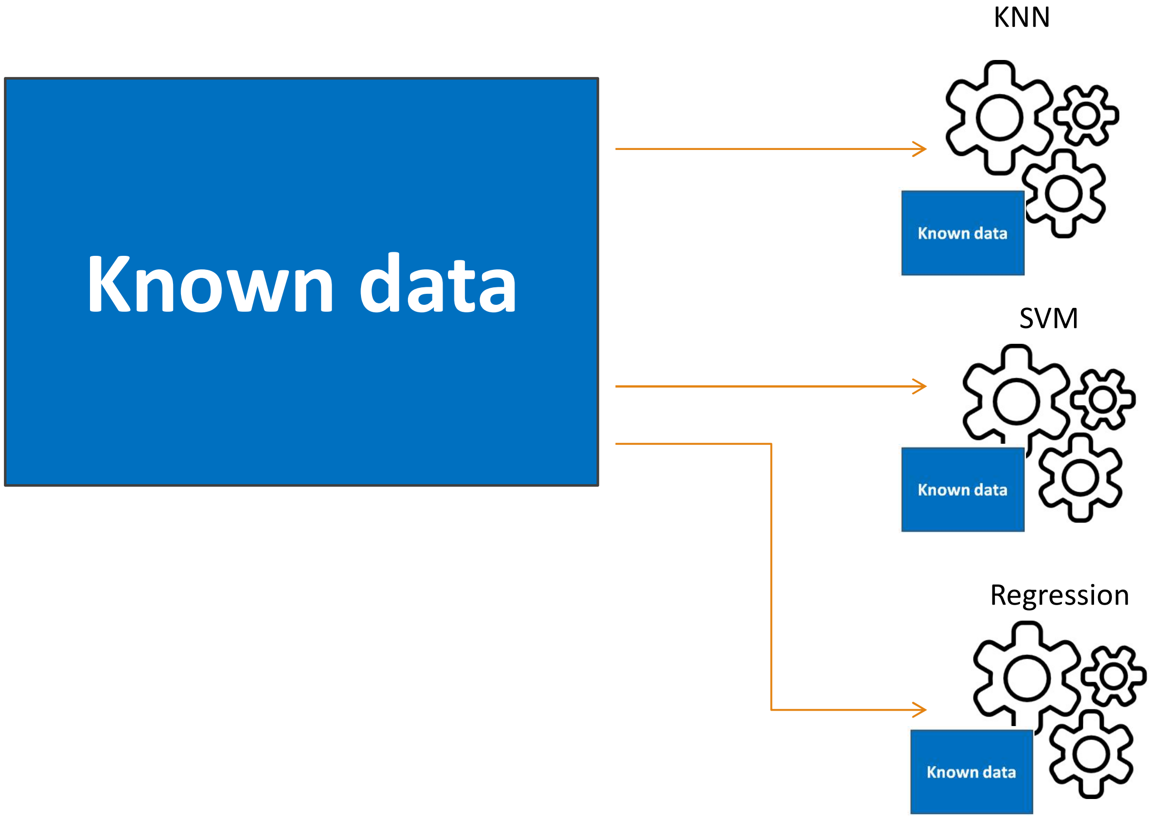
MACHINE LEARNING WORKFLOW OPERATIONS – TRAINING

DONT THINK JUST ABOUT KNN

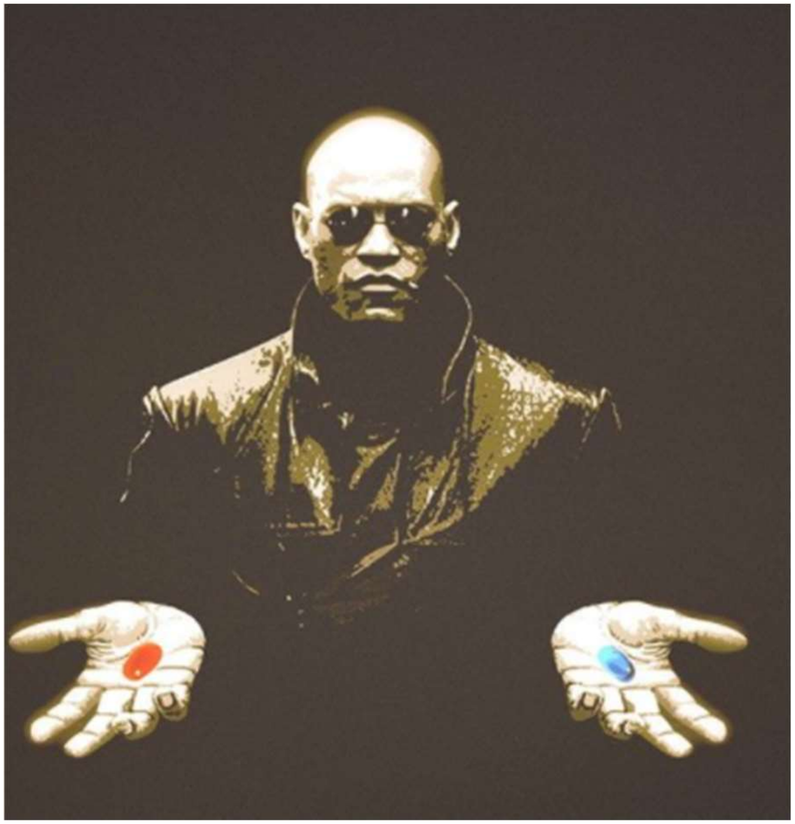
MACHINE LEARNING PROCESS

MODEL SELECTION

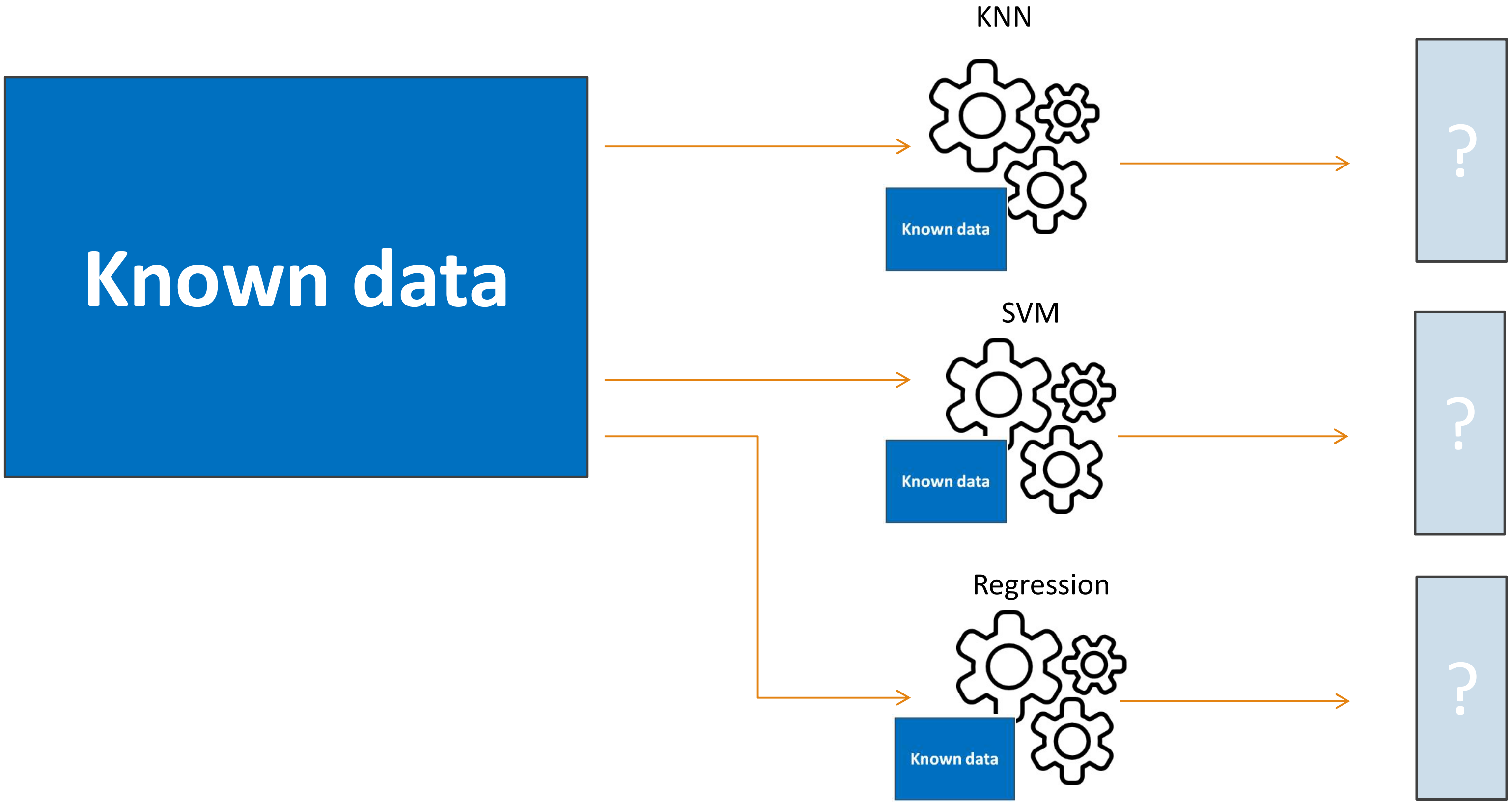
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



MODEL SELECTION – WE CAN ONLY PICK ONE

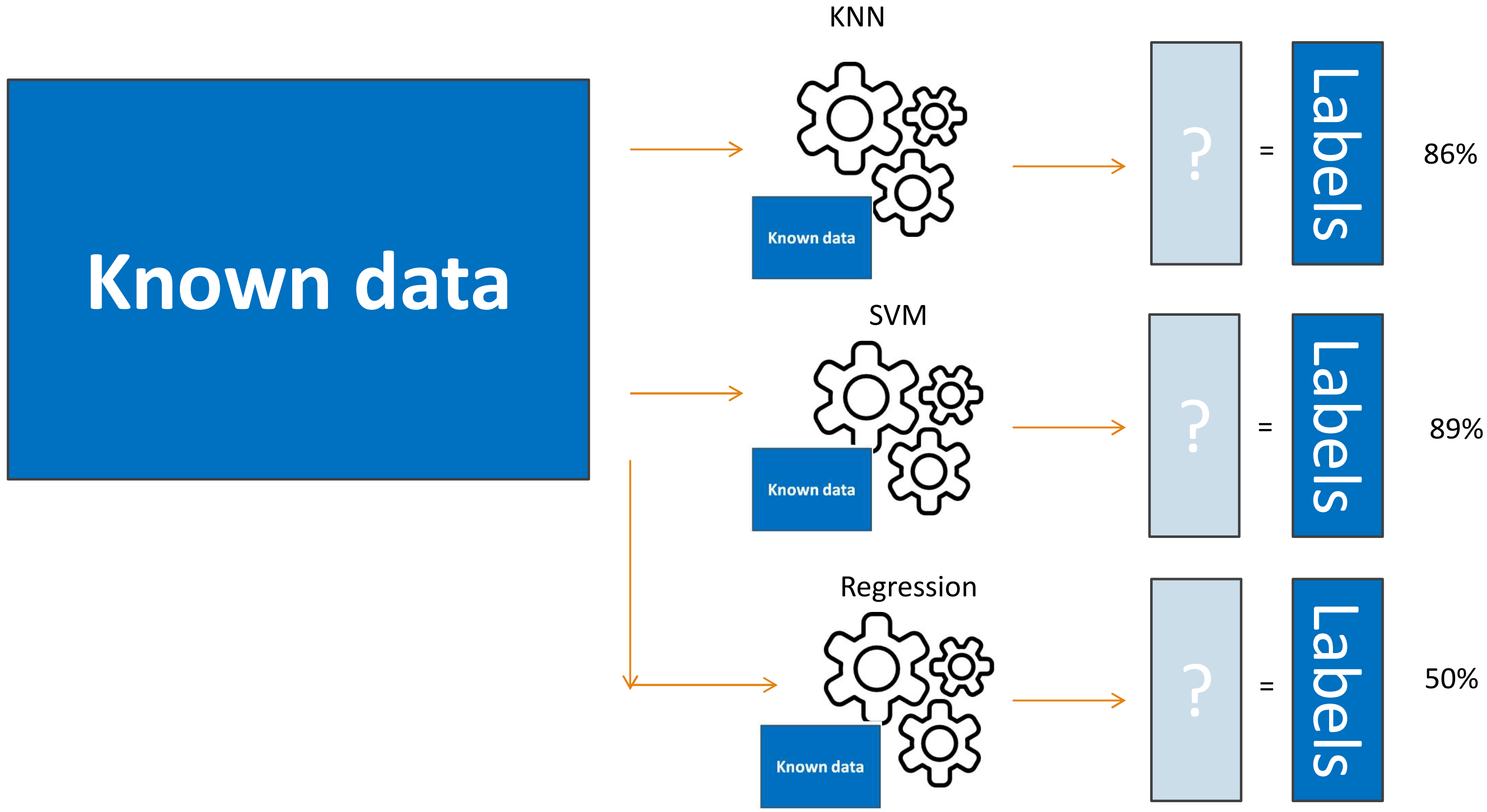


MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION

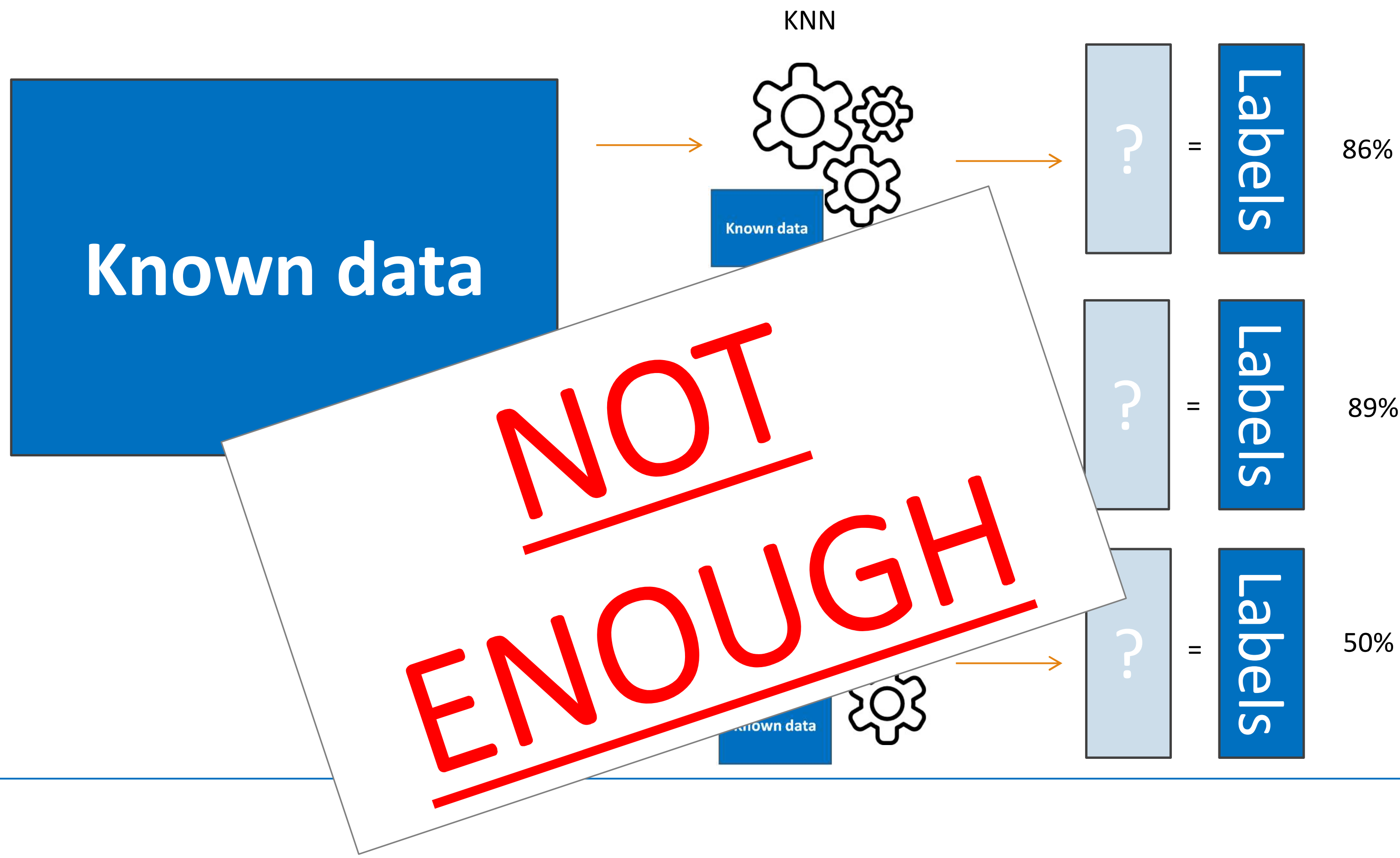


MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION

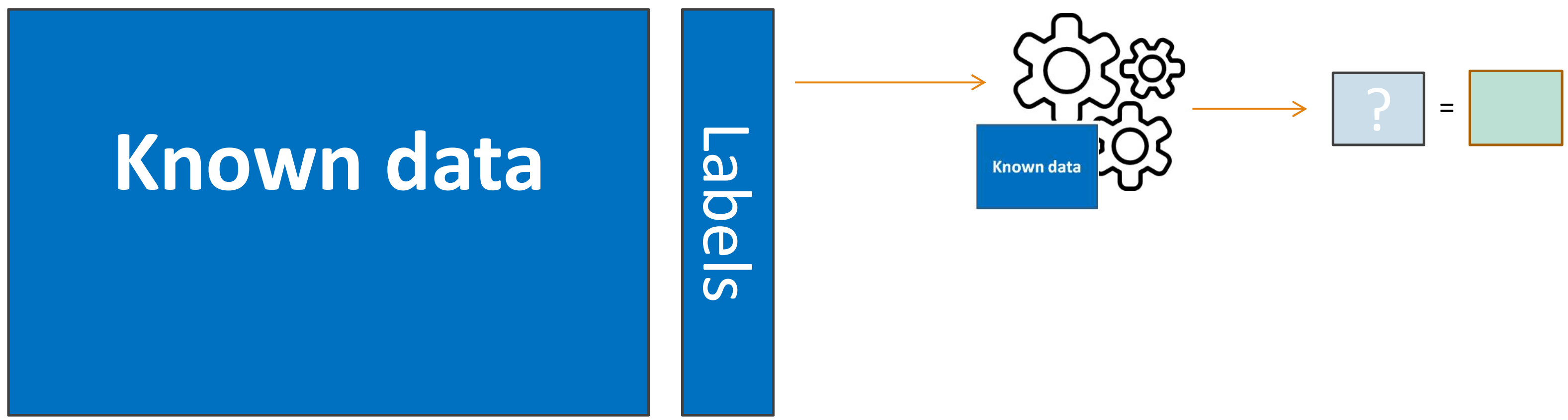
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



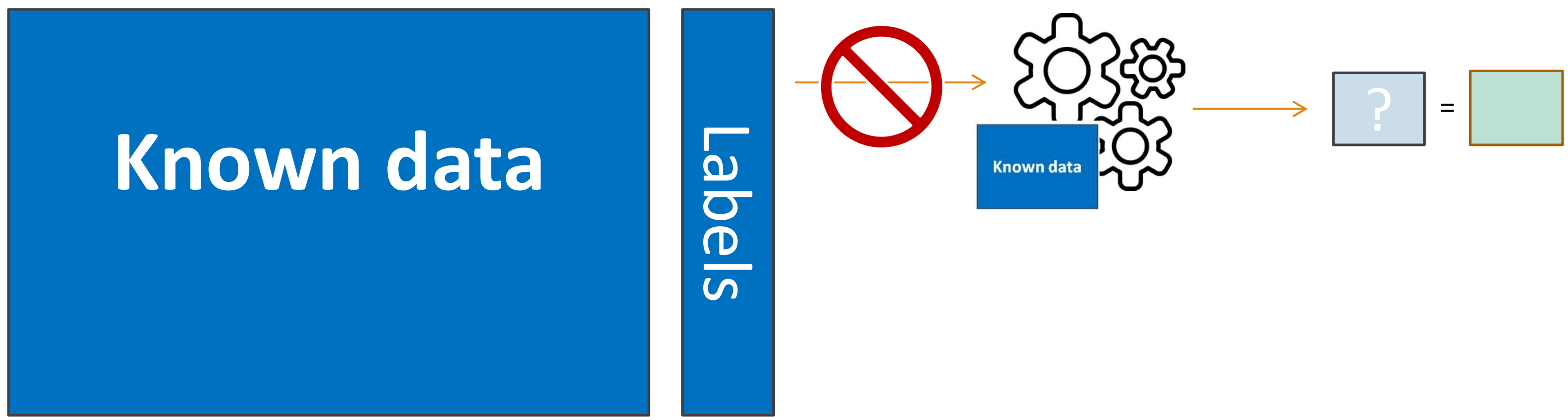
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



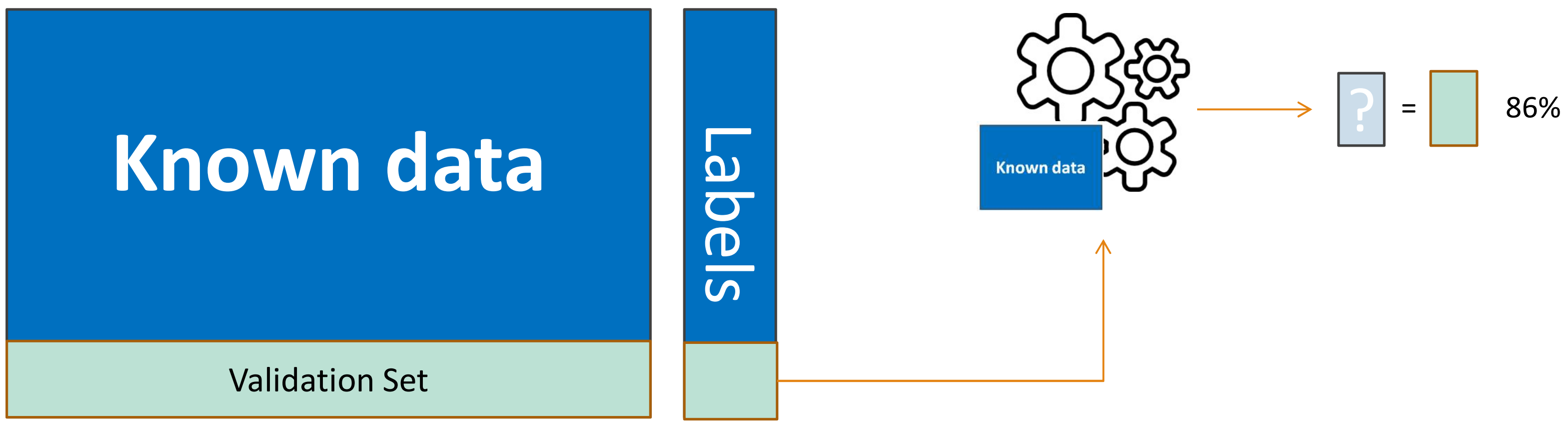
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



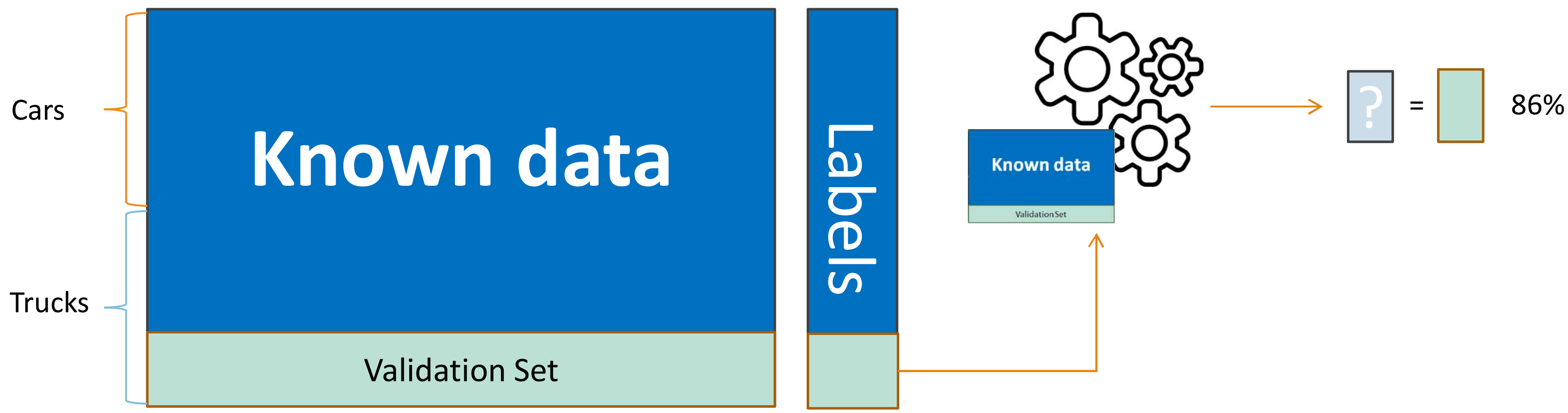
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



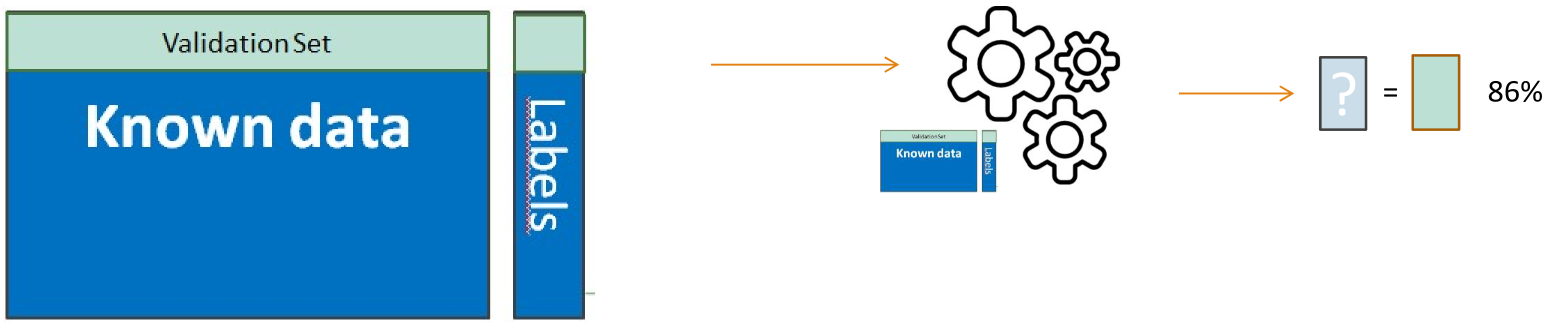
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



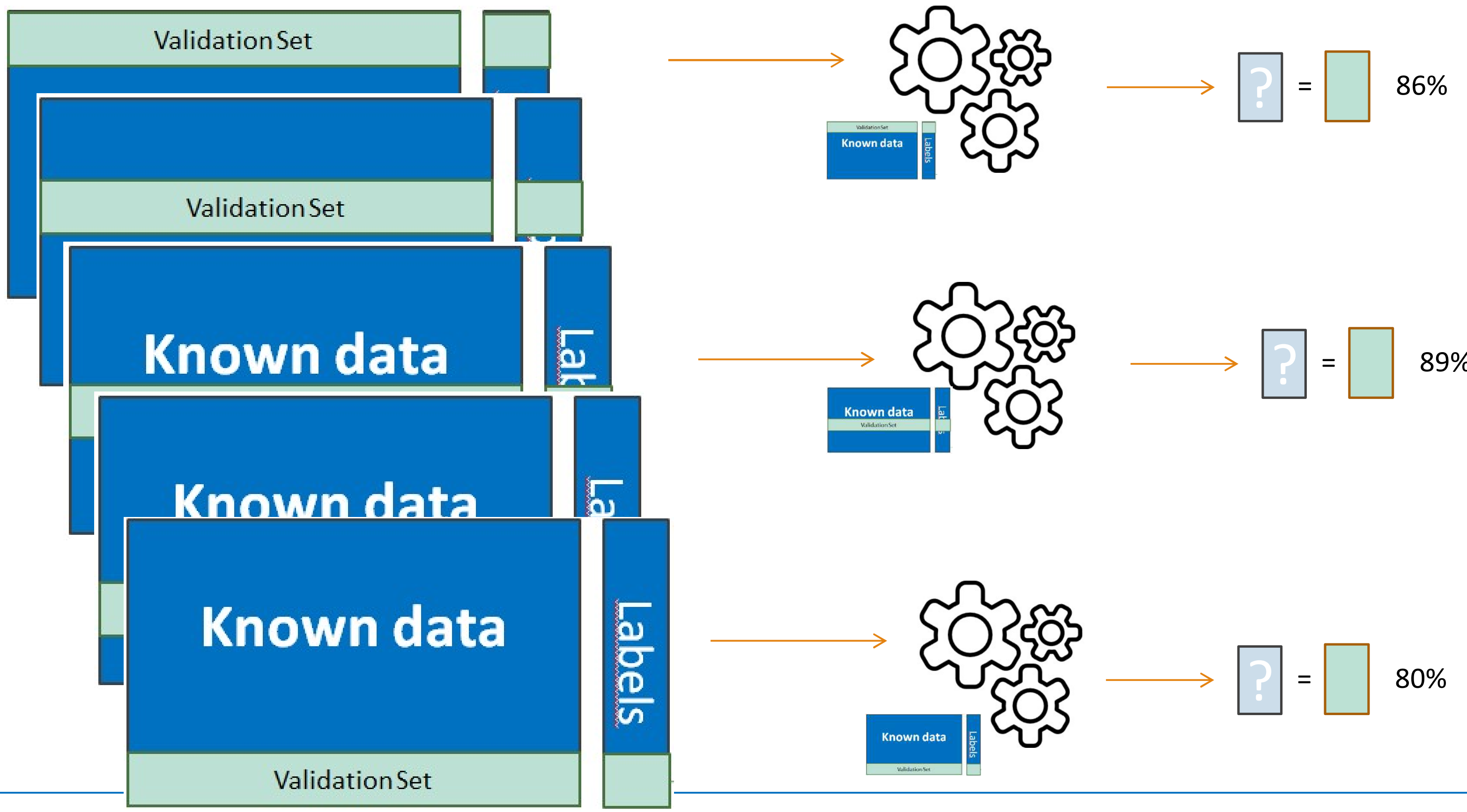
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



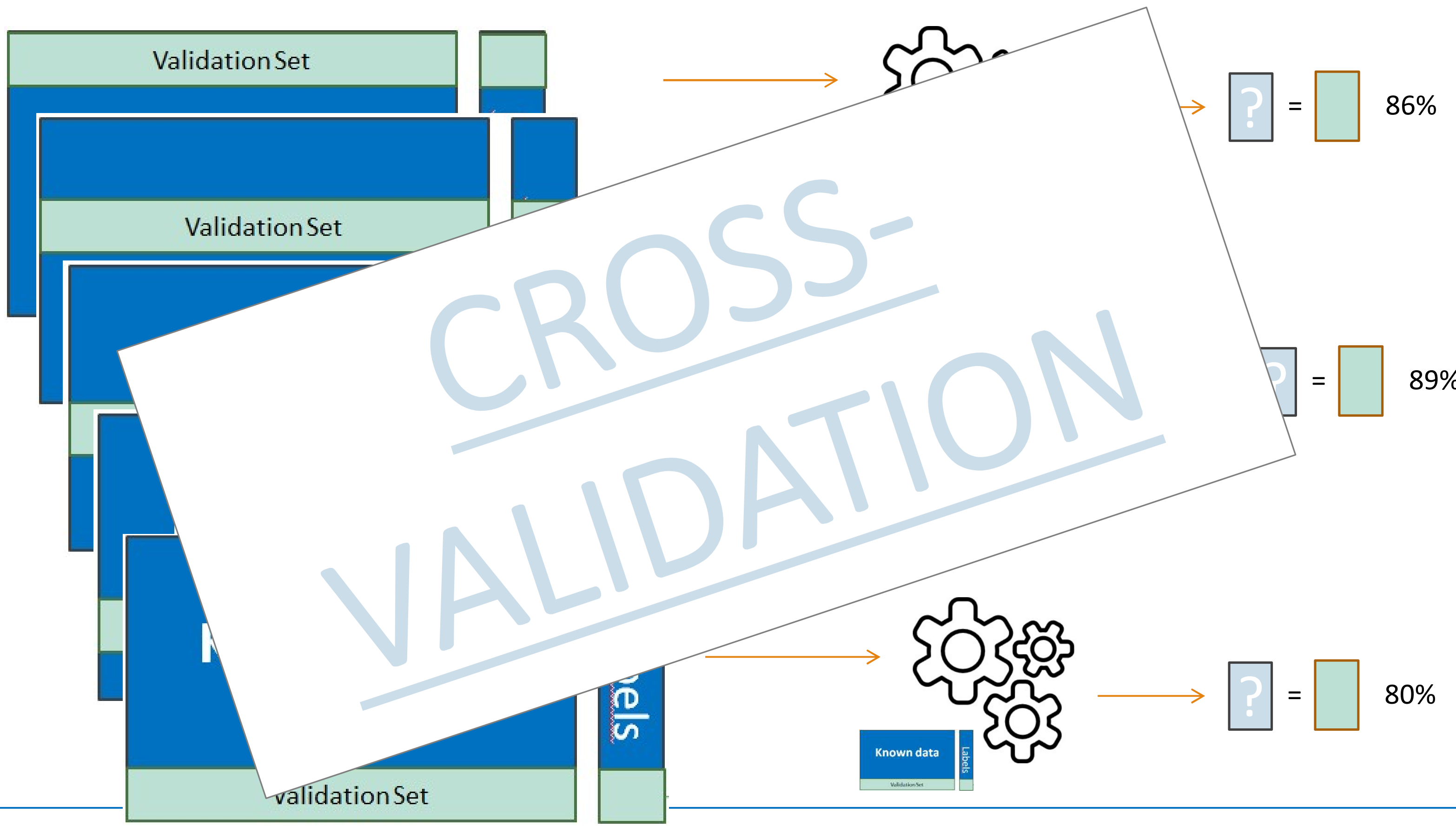
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



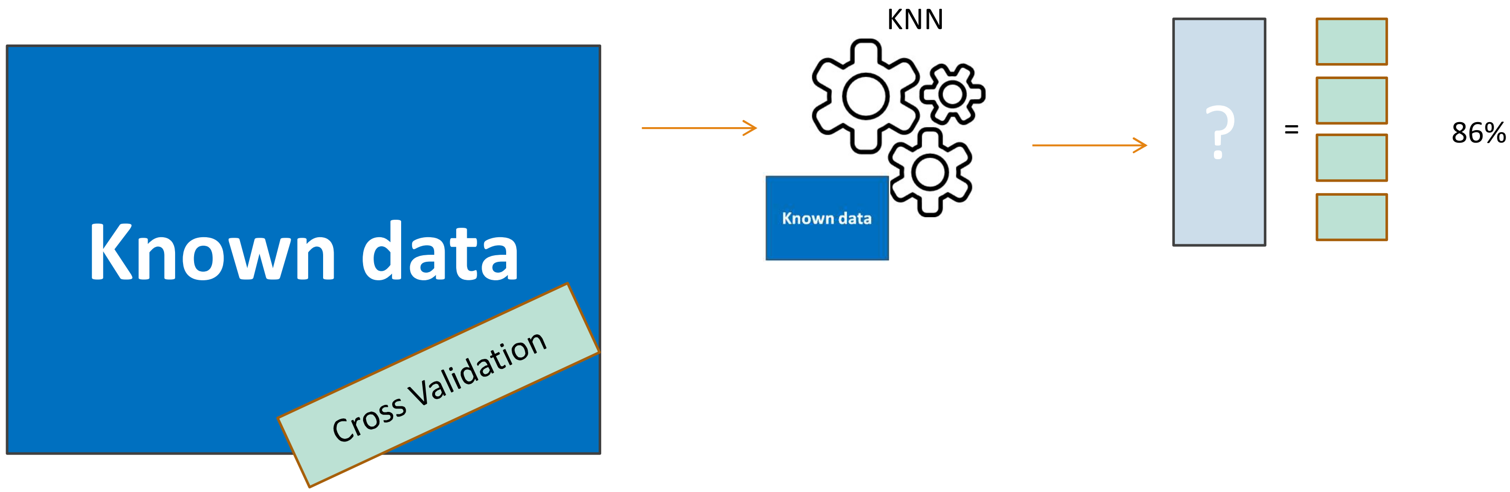
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



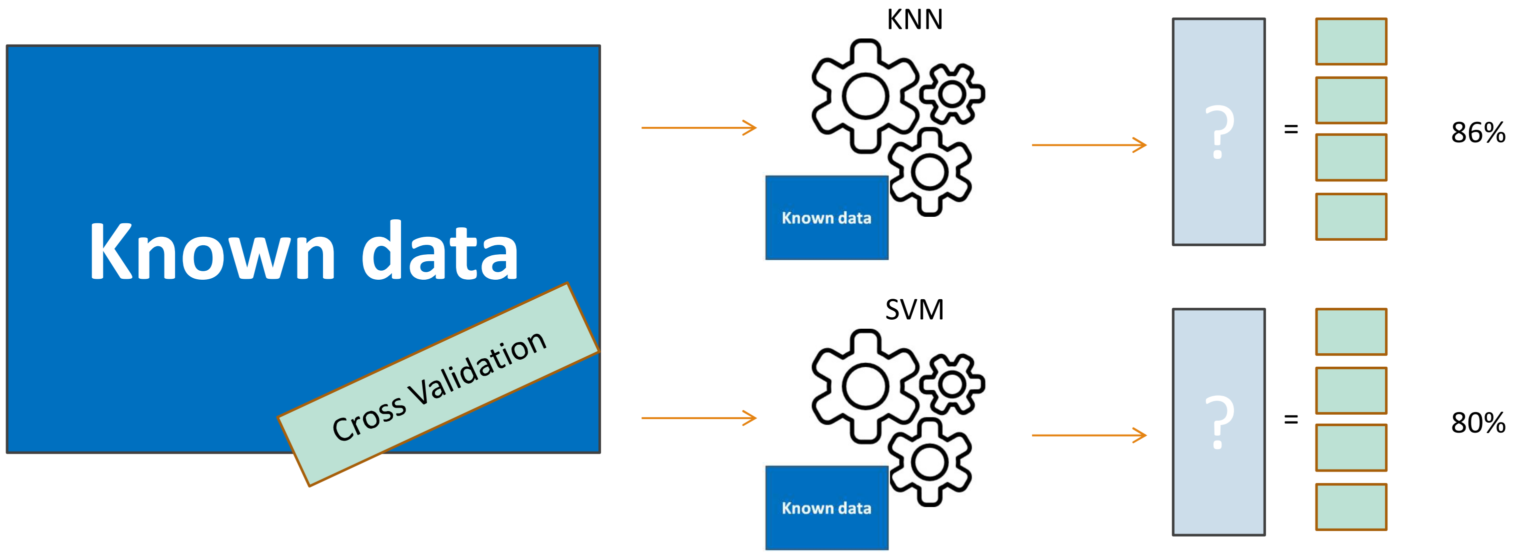
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



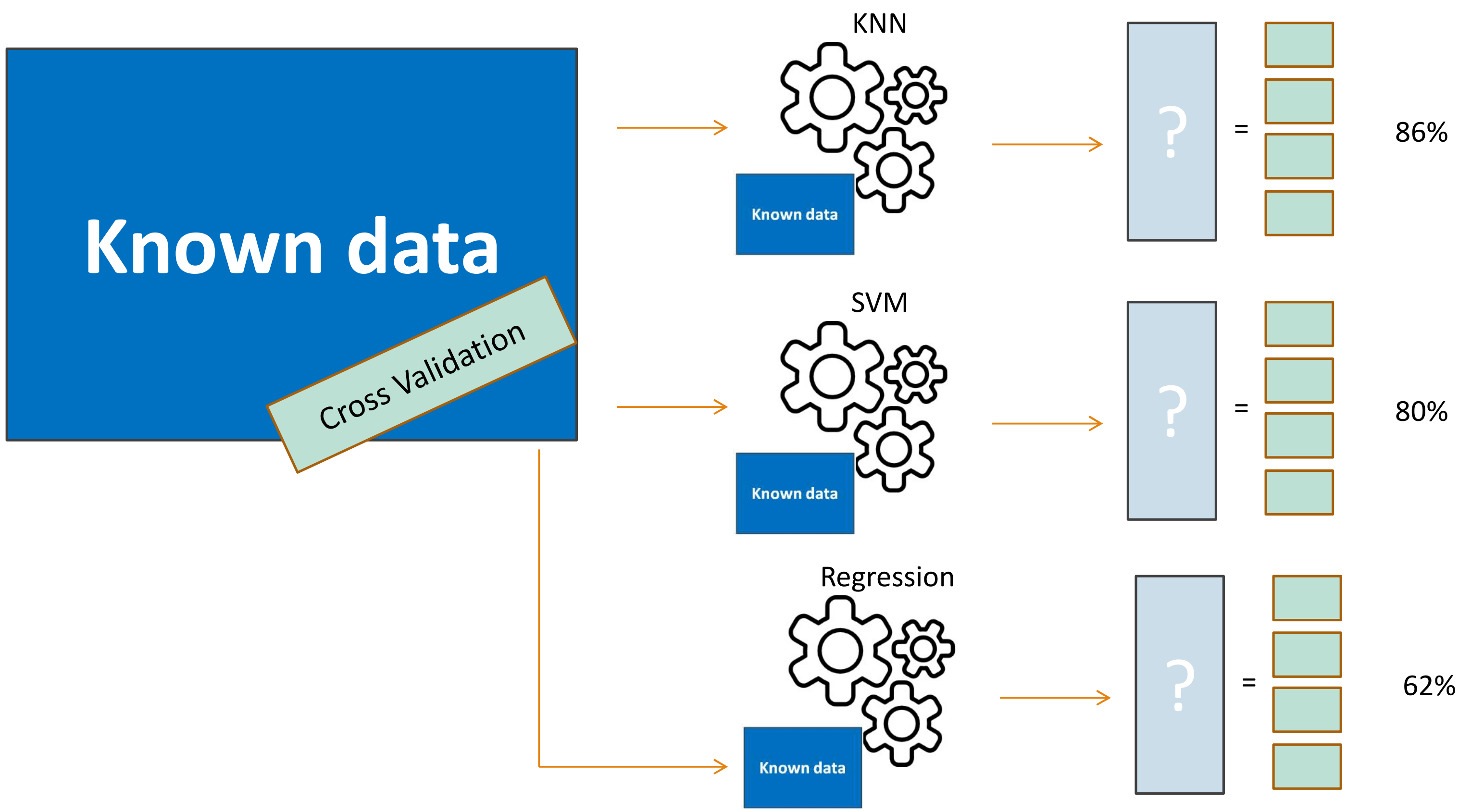
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION

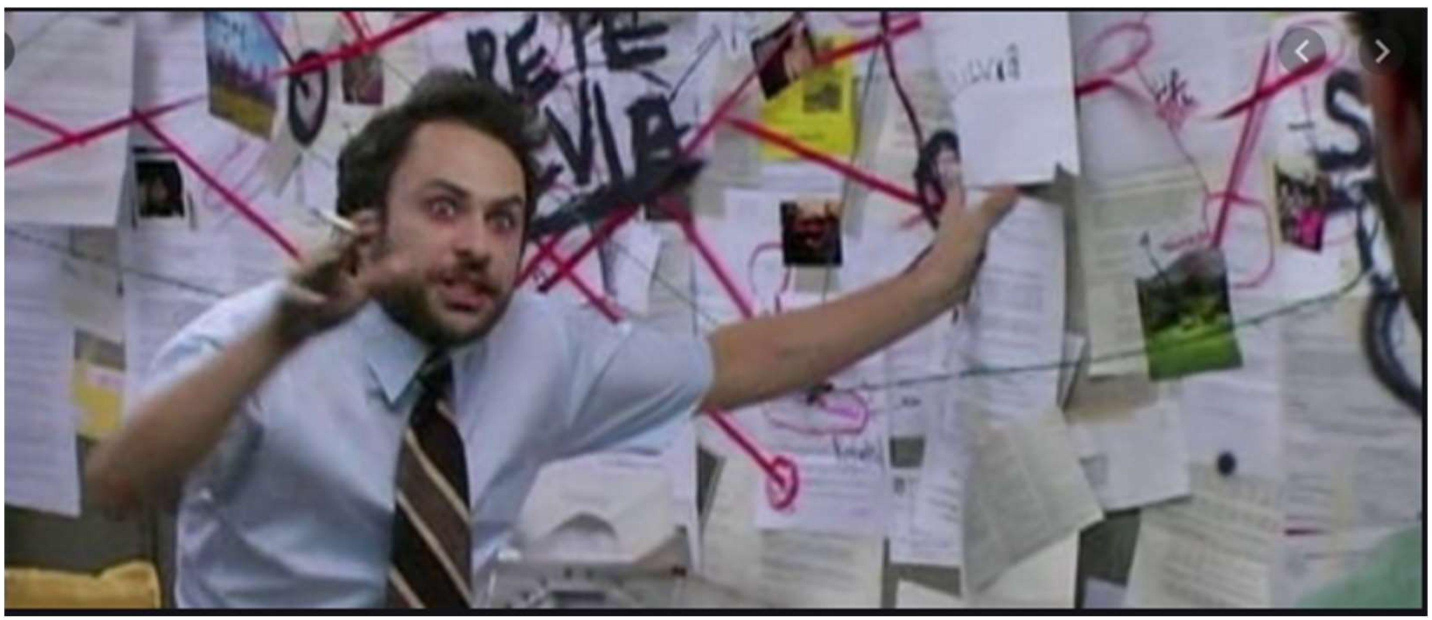


MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



MODEL SELECTION – ~~WE CAN ONLY PICK ONE~~

ACTUALLY NO… BUT THAT IS A FAIRLY ADVANCED TOPIC WE WILL LEAVE FOR LATER



NOW, OUR MODEL IS
TRAINED AND PICKED
AND
READY TO CLASSIFY NEW DATA. TRUE

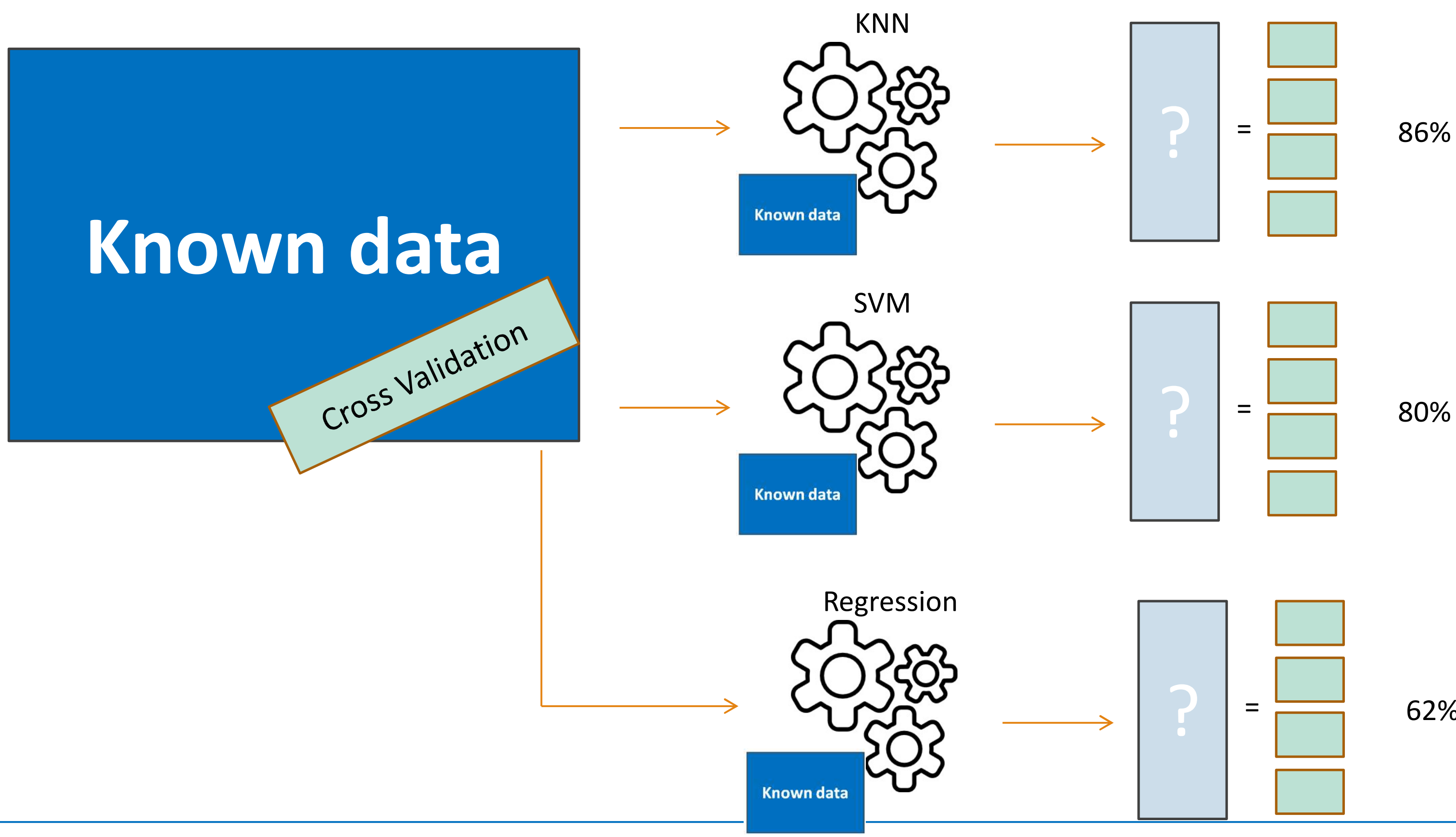
NOW, OUR MODEL IS
TRAINED AND PICKED
AND
READY TO CLASSIFY NEW DATA. TRUE

BUT CAN WE SAY HOW GOOD IT IS?

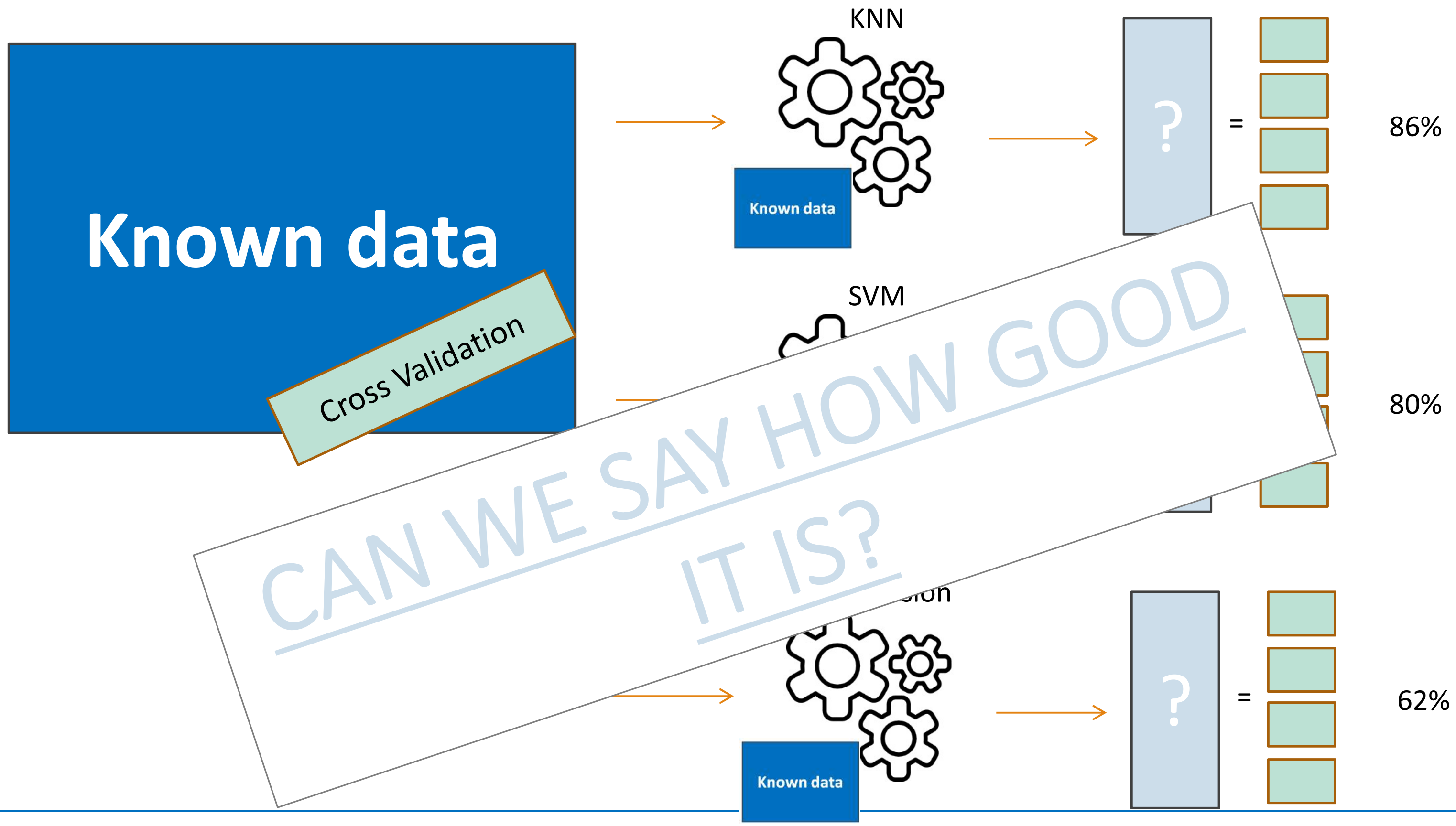
MACHINE LEARNING PROCESS

TEST SET

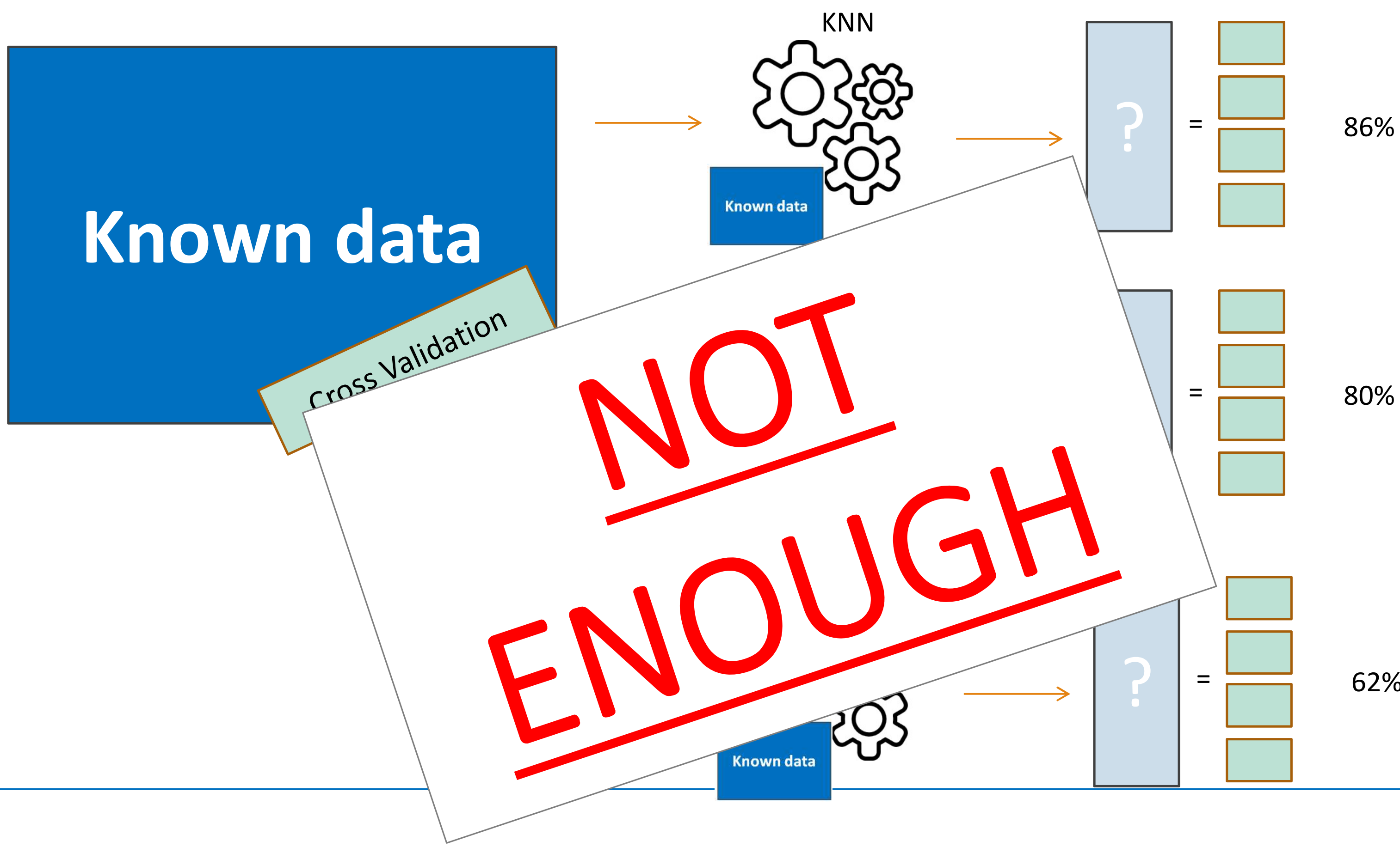
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



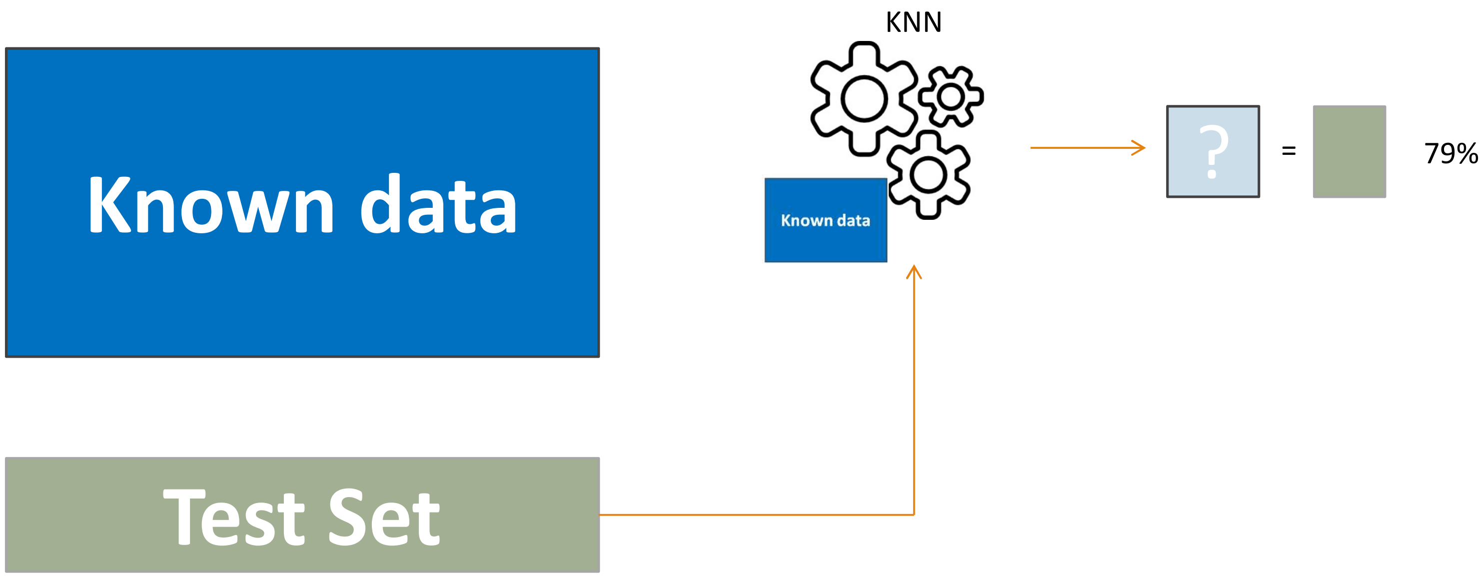
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



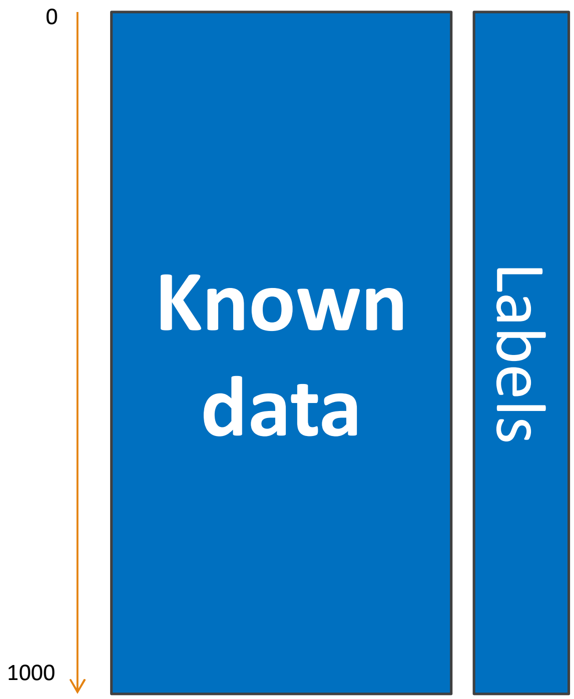
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



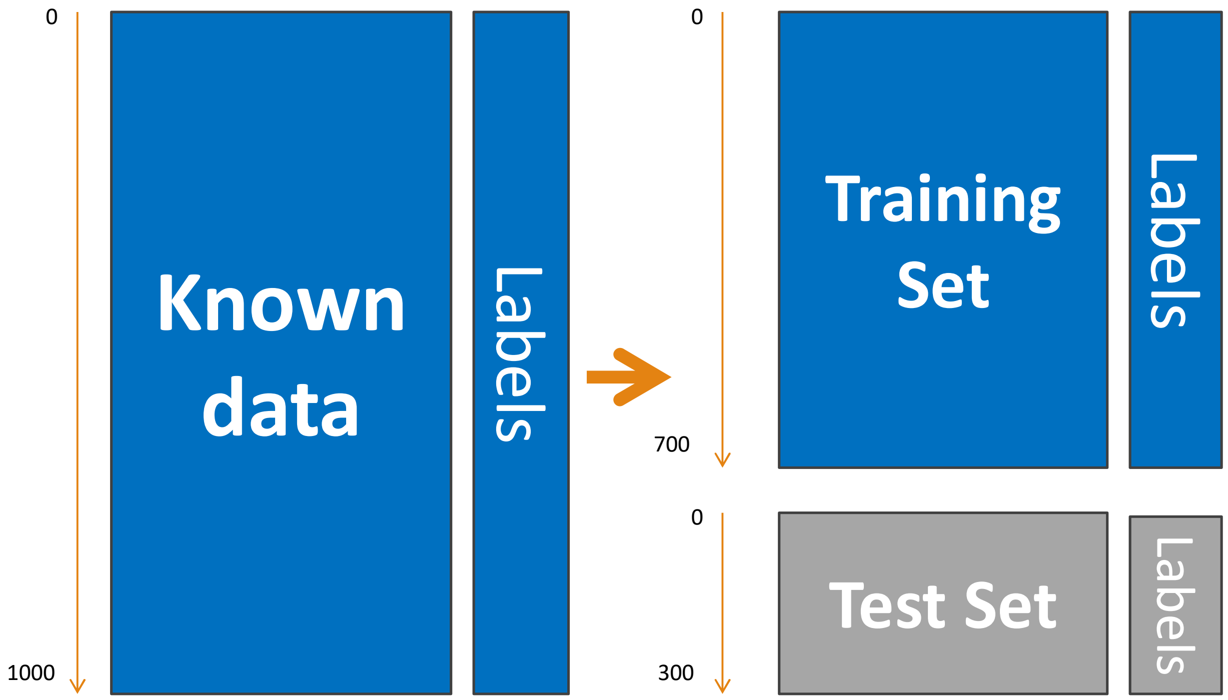
MACHINE LEARNING WORKFLOW OPERATIONS – TEST SET



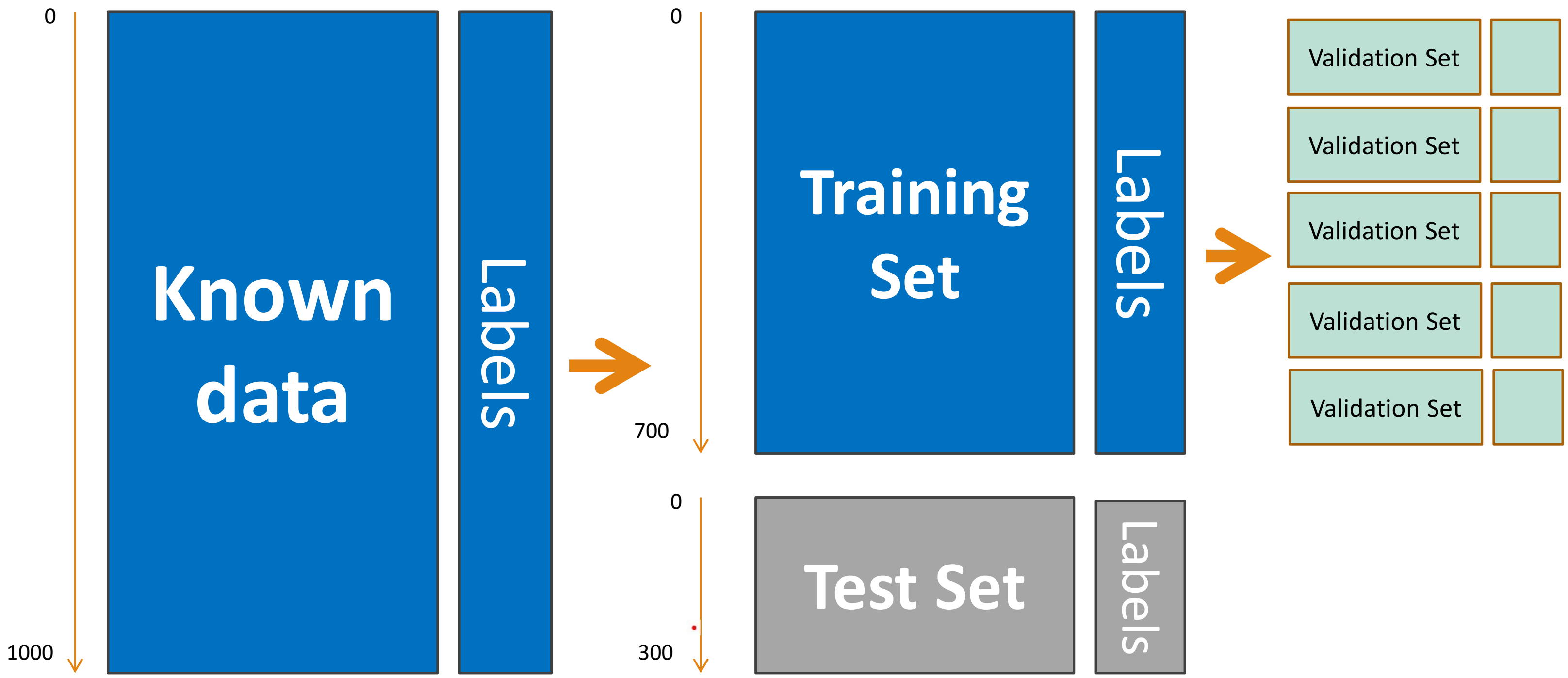
MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW

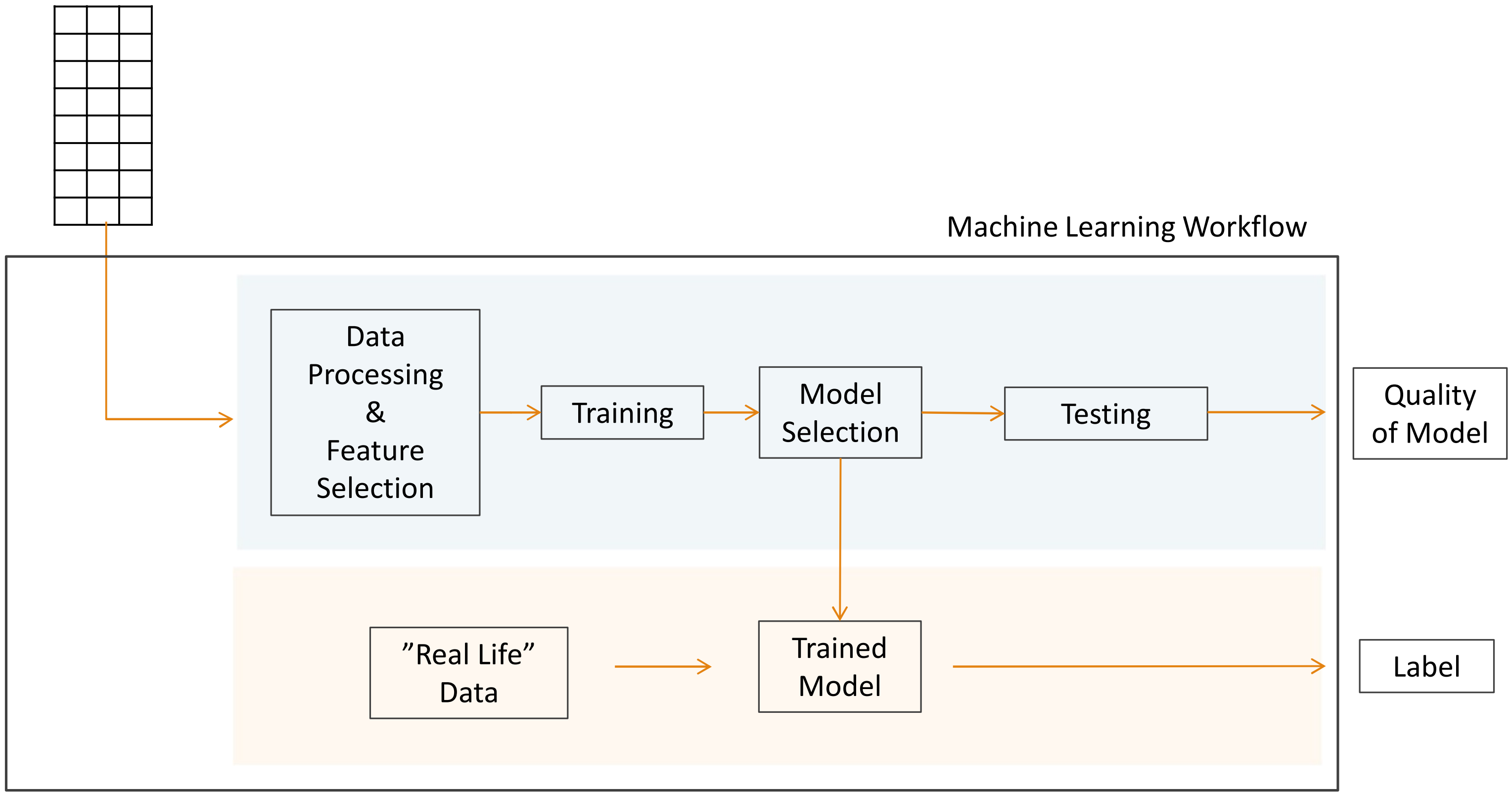


MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW



MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW




Machine Learning Workflow
Data Processing & Feature Selection
Training
Model Selection
Testing
Quality of Model
"Real Life" Data
Trained Model
Label

PERFORMANCE METRICS

LETS EVALUATE OUR MODEL

MACHINE LEARNING WORKFLOW– PERFORMANCE METRICS

True Positive

Classes

Confusion Matrix

False Positive Rate

Specificity

F1 score

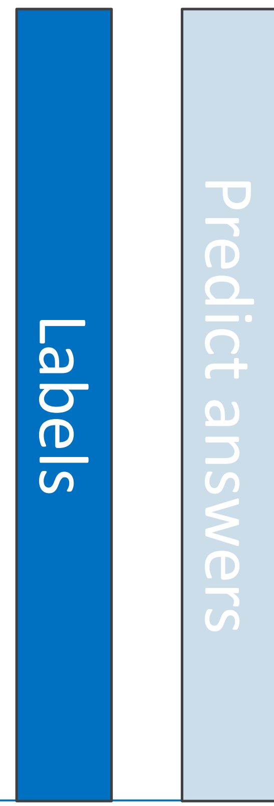
Precision

MACHINE LEARNING PROCESS

ACCURACY

MACHINE LEARNING WORKFLOW– PERFORMANCE METRICS

ACCURACY:



MACHINE LEARNING WORKFLOW– PERFORMANCE METRICS

ACCURACY:

Labels

Predict answers

MACHINE LEARNING WORKFLOW– PERFORMANCE METRICS

ACCURACY:

Labels	0	1	1	1	0	0	1	0	1	0
Predicted Answers	0	1	0	1	0	0	1	0	1	0

Accuracy: (# grey cells / # total) - 90%

MACHINE LEARNING WORKFLOW– PERFORMANCE METRICS

ACCURACY (FOR SEVERAL CLASSES):

Labels	0	1	2	1	2	0	1	2	1	0
Predicted Answers	0	1	0	1	2	0	0	2	1	2

MACHINE LEARNING WORKFLOW– PERFORMANCE METRICS

ACCURACY (FOR SEVERAL CLASSES):

Labels	0	1	2	1	2	0	1	2	1	0
Predicted Answers	0	1	0	1	2	0	0	2	1	2

Accuracy: (# grey cells / # total) - 70%

IRON HACK

MACHINE LEARNING PROCESS

CONFUSION MATRIX

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES…

MACHINE LEARNING WORKFLOW– CONFUSION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES...

Predicted Labels (output of the model)

Correct Labels (provided in the data)

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES…

Predicted Labels (output of the model)

Correct Labels (provided in the data)

A			

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES…

Predicted Labels (output of the model)

Correct Labels (provided in the data)

	A		
A			

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES…

Predicted Labels (output of the model)

Correct Labels (provided in the data)

	A		
A	266		

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES…

Predicted Labels (output of the model)

Correct Labels (provided in the data)

	A	B	C
A	266 ✓	21 ✗	30 ✗

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES...

Predicted Labels (output of the model)

Correct Labels (provided in the data)

	A	B	C
A	266	21	30
B		289	

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES...

Predicted Labels (output of the model)

Correct Labels (provided in the data)		A	B	C
	A	266	21	30
	B		289	
	C			300

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES…

Predicted Labels (output of the model)

Correct Labels (provided in the data)

	A	B	C
A	266 ✓	21 ✗	30 ✗
B	18 ✗	289 ✓	50 ✗
C	16 ✗	10 ✗	300 ✓

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES...

Predicted Labels (output of the model)

Correct Labels (provided in the data)

	A	B	C	
A	266	21	30	317
B	18	289	50	357
C	16	10	300	326
	300	320	380	TOTAL: 1000

CLASSIFICATION IN MACHINE LEARNING – TYPES OF ANSWERS

Features	Correct Label	Predicted Label
Entry 1	0	0
Entry 2	1	1
Entry 3	1	0
Entry 4	0	1

CLASSIFICATION IN MACHINE LEARNING – TYPES OF ANSWERS

Features	Correct Label	Predicted Label
Entry 1	0	0
Entry 2	1	1
Entry 3	1	0
Entry 4	0	1

TN (True Negative)

Classification was **correct**: the entry was initially from class “0” or “Negative” and our classification model identified it as “Negative”

CLASSIFICATION IN MACHINE LEARNING – TYPES OF ANSWERS

Features	Correct Label	Predicted Label
Entry 1	0	0
Entry 2	1	1
Entry 3	1	0
Entry 4	0	1

TP (True Positive)

Classification was **correct**: the entry was initially from class “1” or “Positive” and our classification model identified it as “Positive”

CLASSIFICATION IN MACHINE LEARNING – TYPES OF ANSWERS

Features	Correct Label	Predicted Label
Entry 1	0	0
Entry 2	1	1
Entry 3	1	0
Entry 4	0	1

FN (False Negative)

Classification was **incorrect**: The entry was initially from class 1, but our model missclassified this entry, outputing a label of class 0.

This entry was incorrectly classified as a Negative entry, hence the name "False Negative"

CLASSIFICATION IN MACHINE LEARNING – TYPES OF ANSWERS

Features	Correct Label	Predicted Label
Entry 1	0	0
Entry 2	1	1
Entry 3	1	0
Entry 4	0	1

FP (False Positive)

Classification was **incorrect**: The entry was initially from class 0, but our model missclassified this entry, outputing a label of class 1.

This entry was considered as a Positive by our model. This classification is incorrect, hence the name "False Positive"

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES…

Predicted Labels
(output of the model)

Correct Labels
(provided in the data)

	A
A	266
B	18
C	16
	300

For Class A:

Out of 300 entries classified as class A, 266 were in fact originally of class A. 34 (18+16) were of classes B and C, dispite the model's classification.

This is a rate of 266/300 = 88%

This rate is called $Precision_A = \frac{TP}{TP+FP}$

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS USE THIS TO UNDERSTAND HOW GOOD THE MODEL IS FOR EACH CLASS

Predicted Labels (output of the model)

Correct Labels (provided in the data)

	A	B	C	
A	266	21	30	317

For Class A:

Out of 317 entries of class A, the model classified 266 correctly and 51 (21+30) incorrectly.
This is a rate of 266/317 = 83,9%

This rate is called $Recall_A = \frac{TP}{TP+FN}$

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS SAY OUR TEST SET HAD 1000 ENTRIES...

Predicted Labels (output of the model)

Correct Labels (provided in the data)	A	B	C	
A	266 ✓	21 ✗	30 ✗	317
B	18 ✗	289 ✓	50 ✗	357
C	16 ✗	10 ✗	300 ✓	326
	300	320	380	TOTAL: 1000

IRON HACK

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS IMAGINE THE EXAMPLE OF FRAUD DETECTION

WHICH DO YOU THINK THE CLIENT WOULD PREFER?

A – Fraud
B – Not Fraud

Predicted Labels (output of the model)

Correct Labels (provided in the data)

TOTAL 1000	A	B
A	21	1
B	51	927

Predicted Labels (output of the model)

Correct Labels (provided in the data)

TOTAL 1000	A	B
A	2	20
B	1	976

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS IMAGINE THE EXAMPLE OF FRAUD DETECTION

A – Fraud
B – Not Fraud

WHICH DO YOU THINK THE CLIENT WOULD PREFER?

Predicted Labels
(output of the model)

Correct Labels
(provided in the data)

TOTAL 1000	A	B
A	21 ✓	1 ✗
B	51 ✗	927 ✓

Predicted Labels
(output of the model)

Correct Labels
(provided in the data)

TOTAL 1000	A	B
A	2 ✓	20 ✗
B	1 ✗	976 ✓

Trade off between
TP<->FN and even FP...

MACHINE LEARNING WORKFLOW– CLASSIFICATION MATRIX

LETS IMAGINE THE EXAMPLE OF FRAUD DETECTION

A – Fraud
B – Not Fraud

WE ALSO CANT HAVE TOO HIGH FP…

Predicted Labels (output of the model)

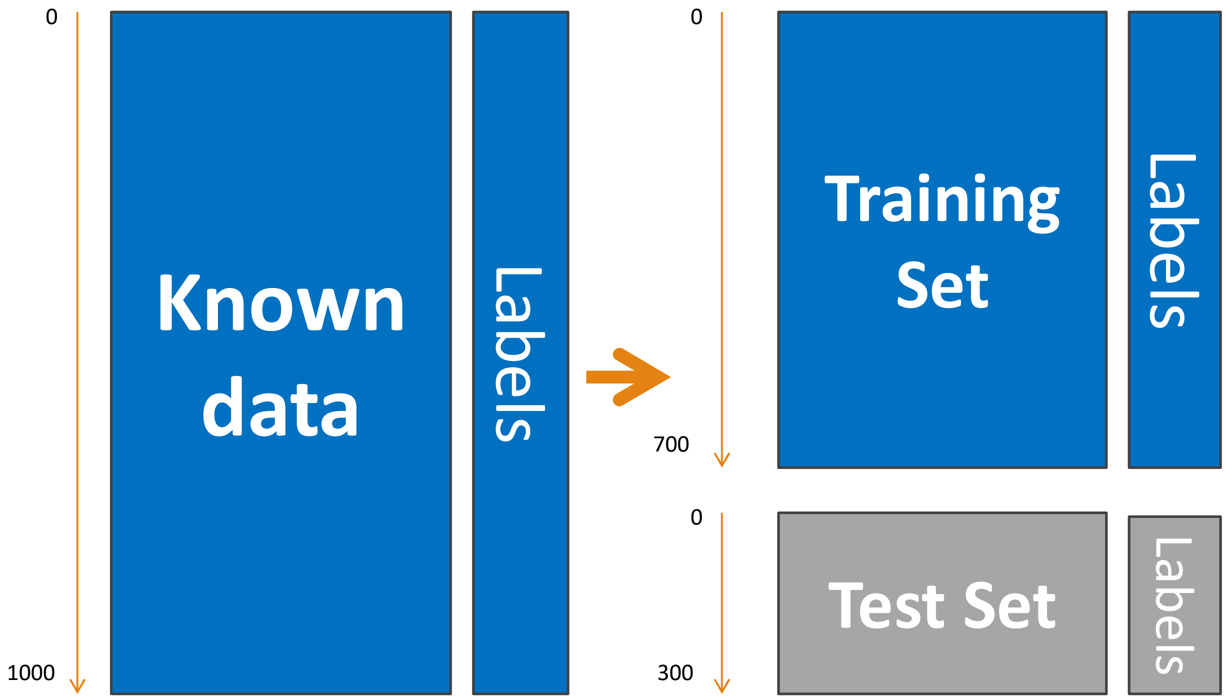
Correct Labels (provided in the data)

TOTAL 1000	A	B
A	21	1
B	927	51

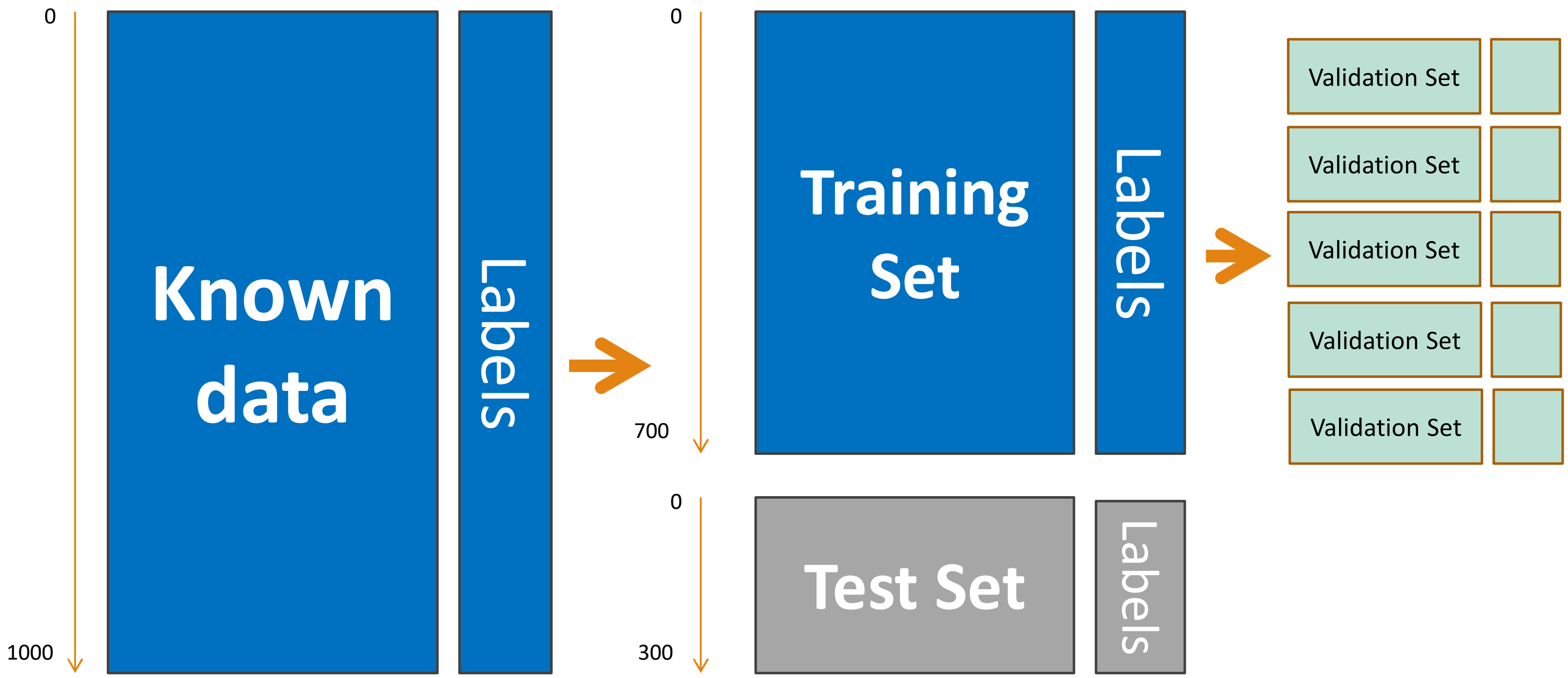
MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW

MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW

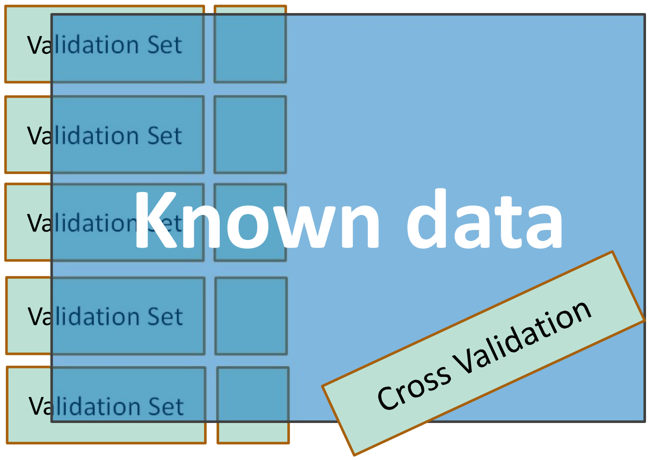
MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW



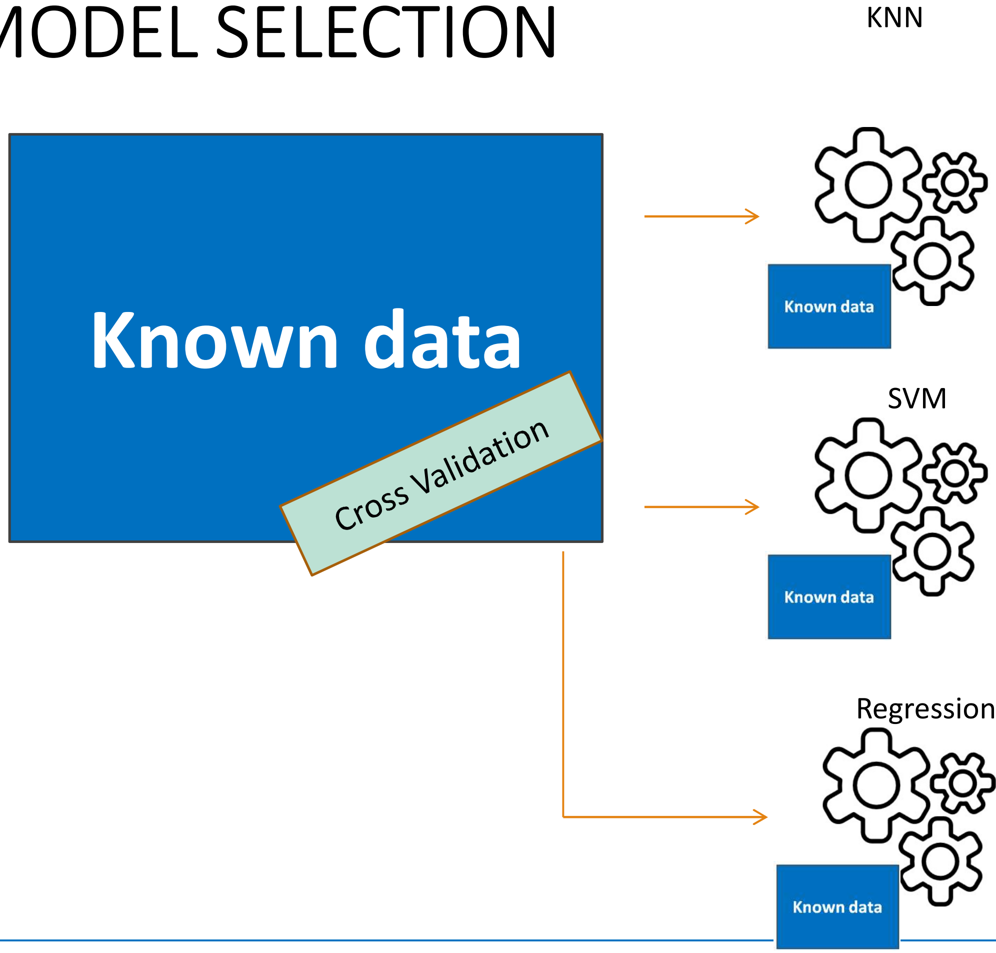
MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW



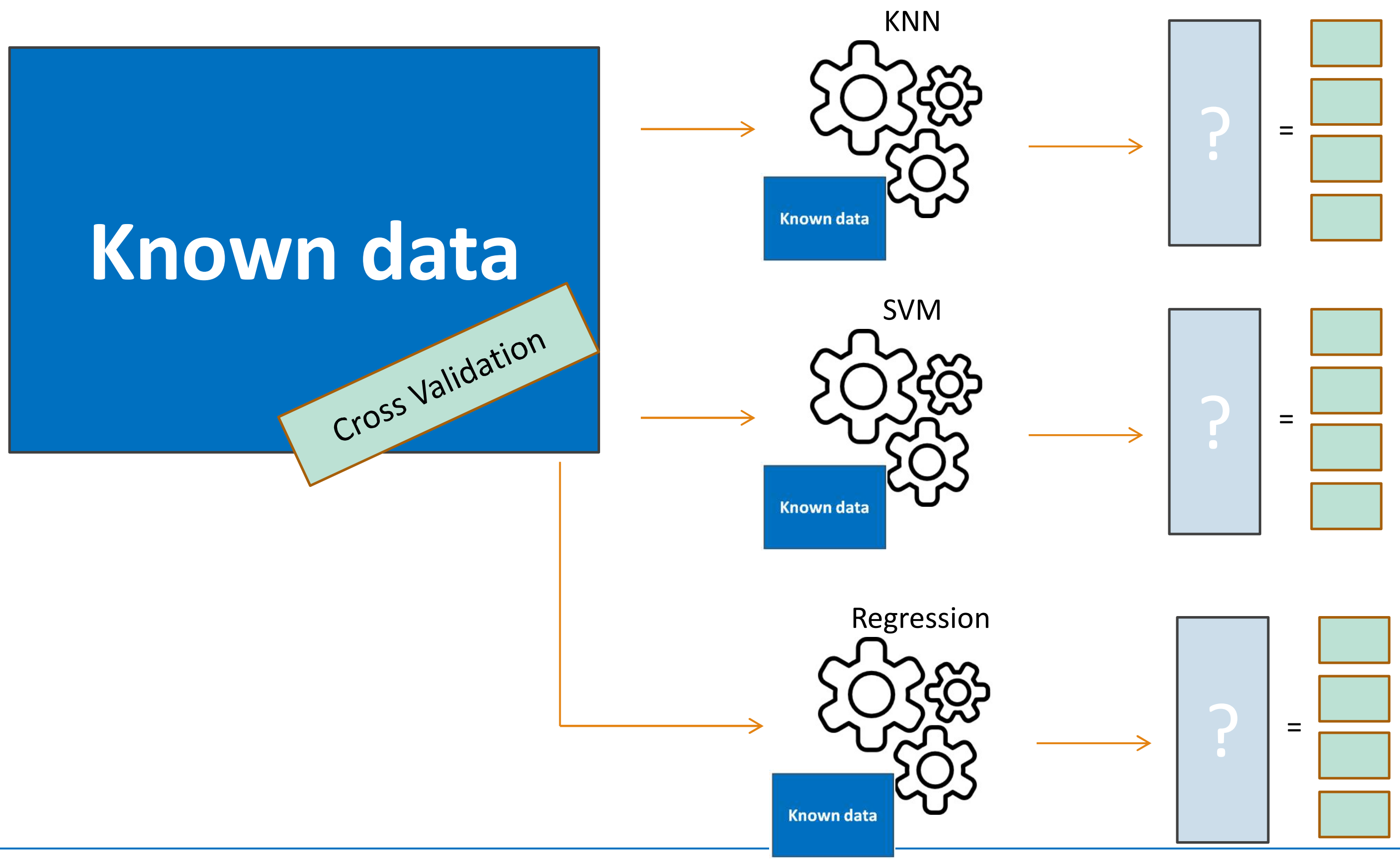
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



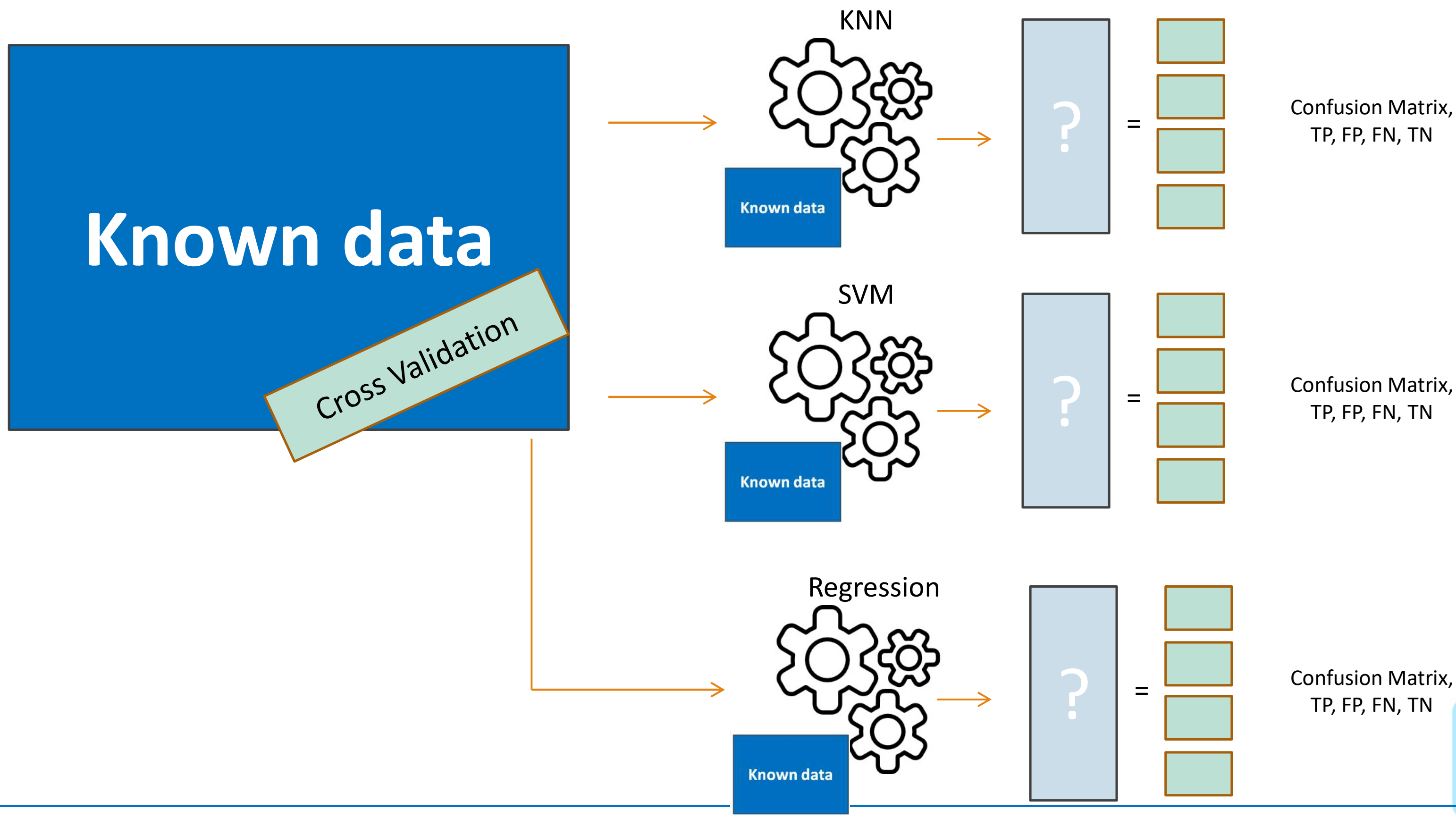
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



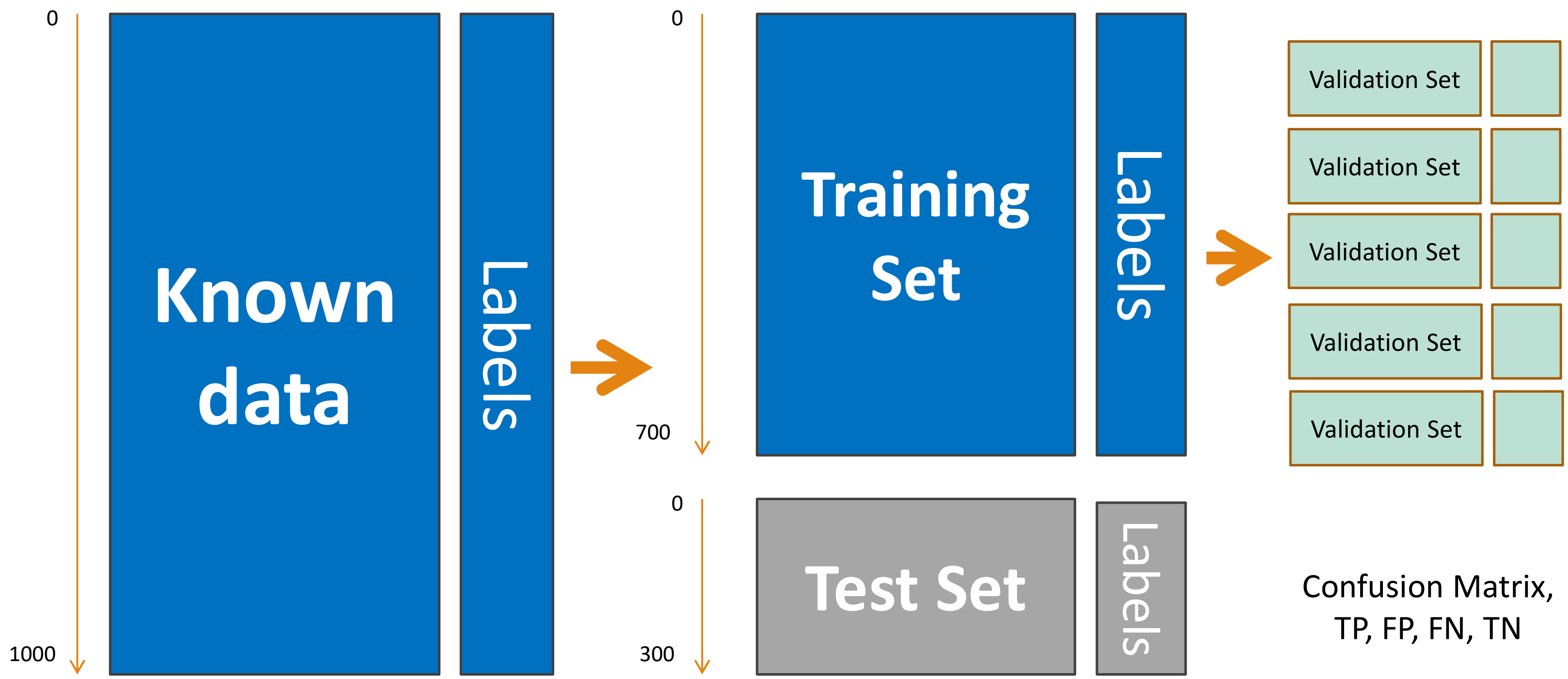
MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



MACHINE LEARNING WORKFLOW OPERATIONS – MODEL SELECTION



MACHINE LEARNING WORKFLOW OPERATIONS – LETS REVIEW



MACHINE LEARNING WORKFLOW

POINTS OF ATTENTION

OVERFITTING

OVERFITTING

In statistics, overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably"

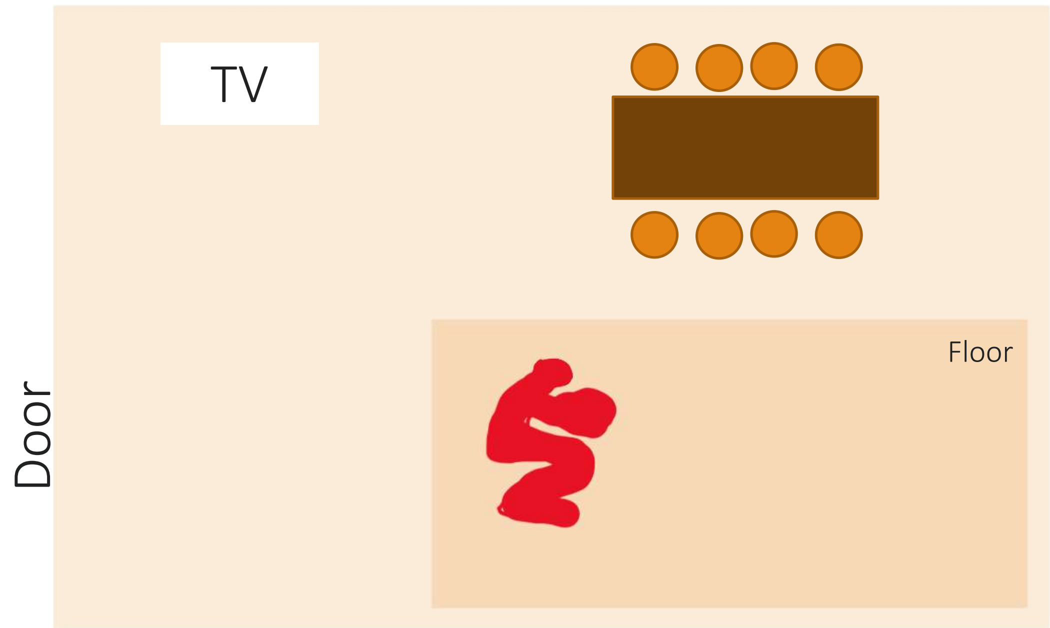
OVERFITTING

In statistics, ov … f an analysis
that corresp … particular
set of data, a … ional data
or predict futu …

WHAT?

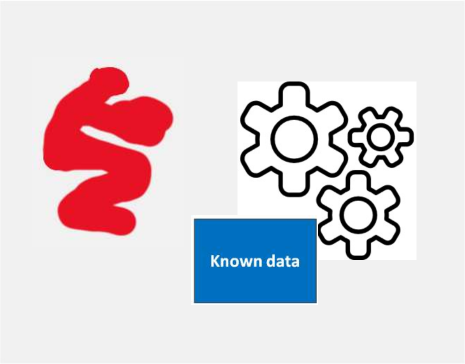
OVERFITTING

Let's imagine that you need to build a place for a person to sleep…



OVERFITTING

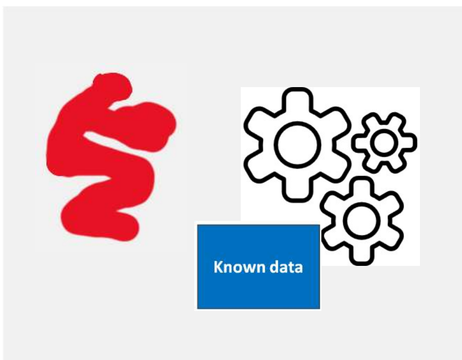
Please use Machine Learning to come up with the best furniture shape for a human to sleep BASED ON THE DATA THAT YOU HAVE



OVERFITTING

Please use Machine Learning to come up with the best furniture shape for a human to sleep BASED ON THE DATA THAT YOU HAVE

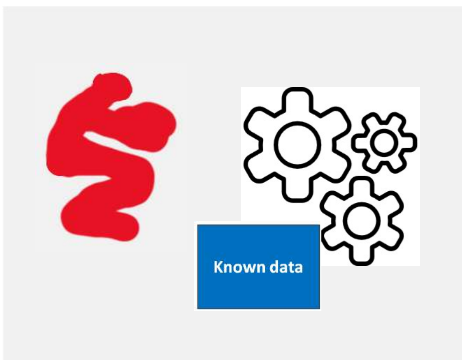
Intuitively, which is best?



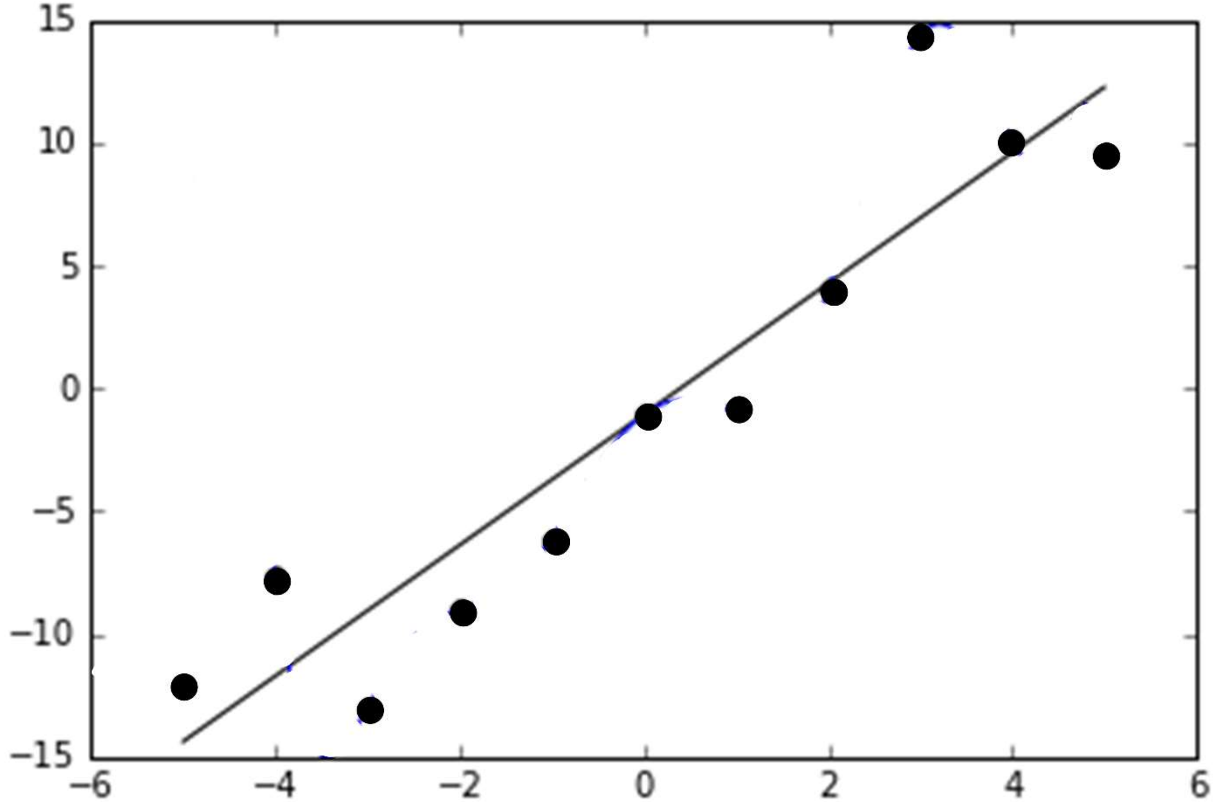
OVERFITTING

Please use Machine Learning to come up with the best furniture shape for a human to sleep BASED ON THE DATA THAT YOU HAVE

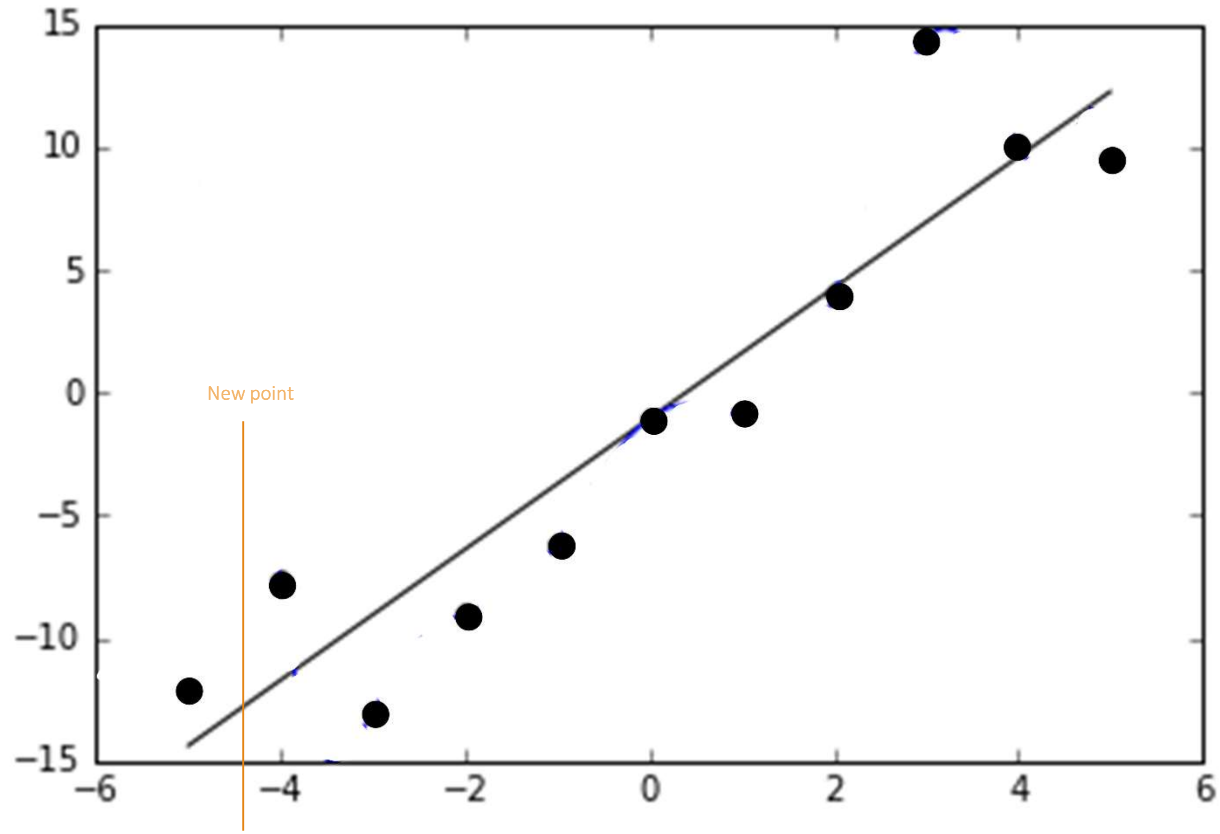
Intuitively, which is best?



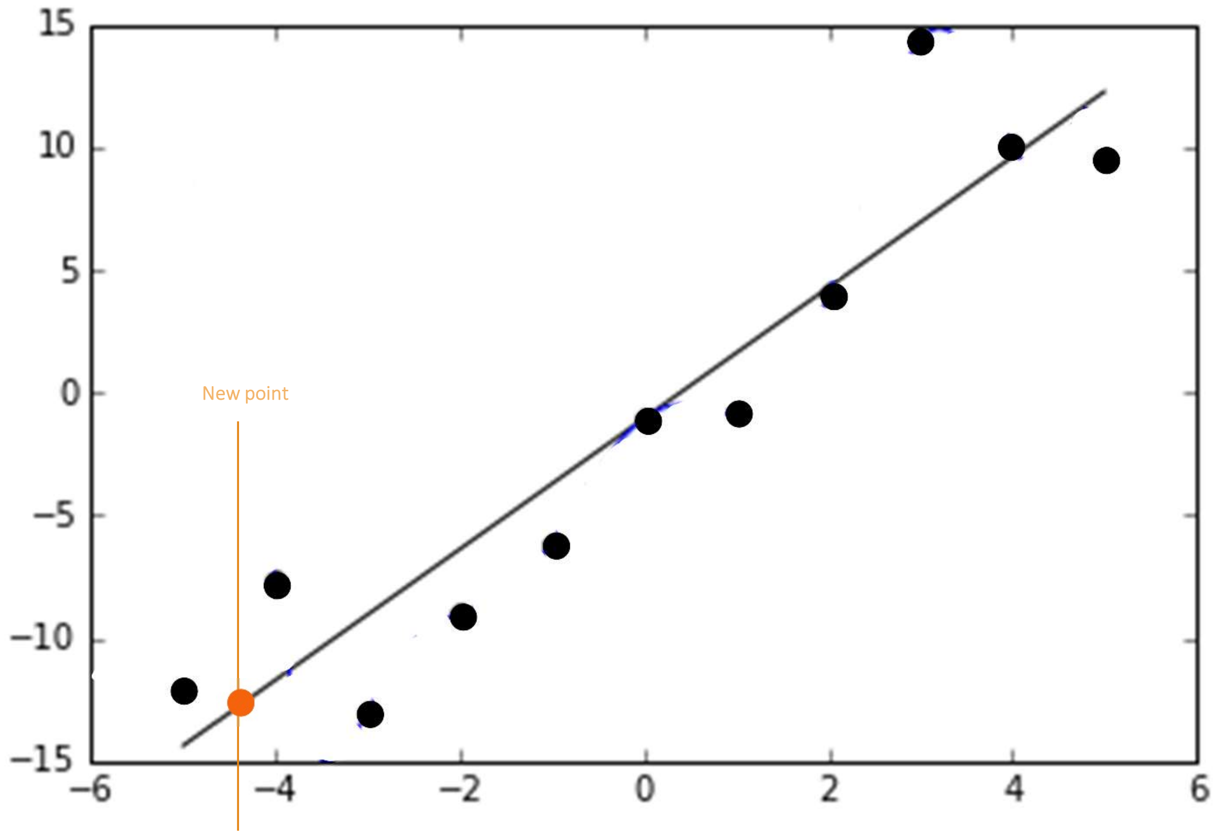
OVERFITTING



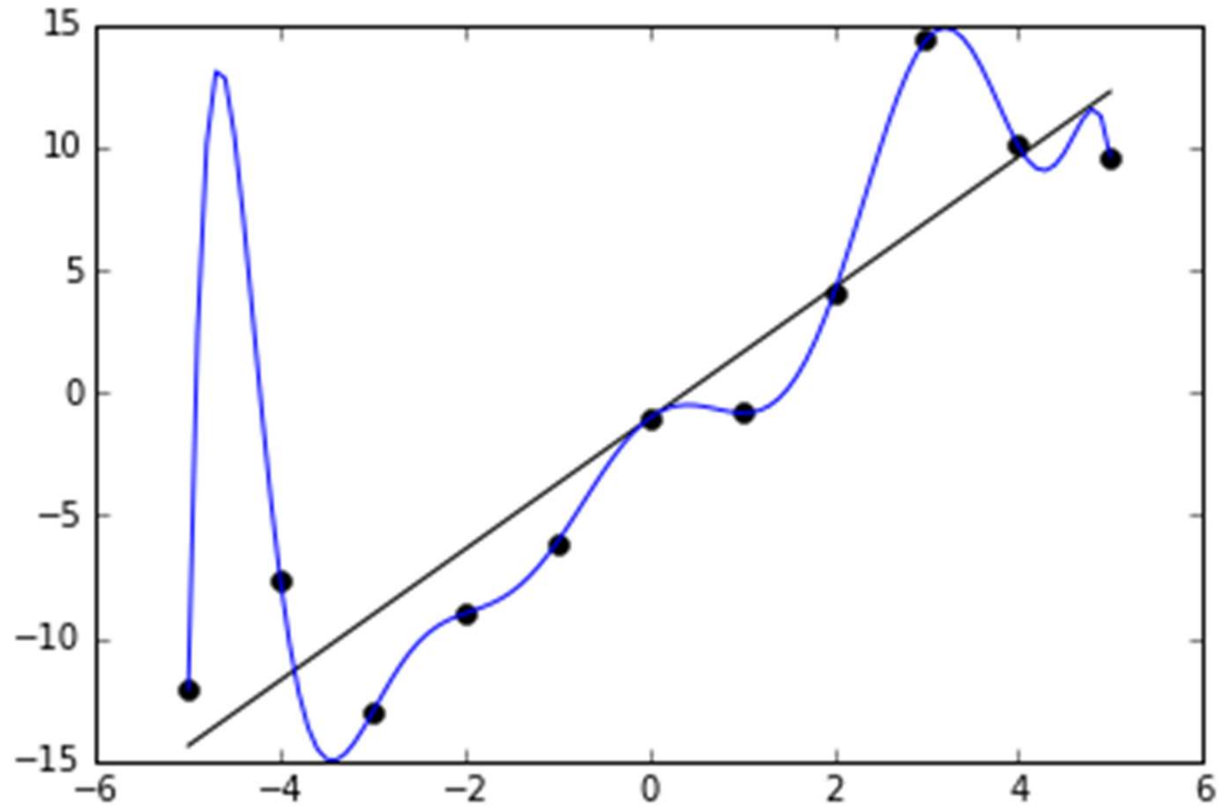
OVERFITTING



OVERFITTING

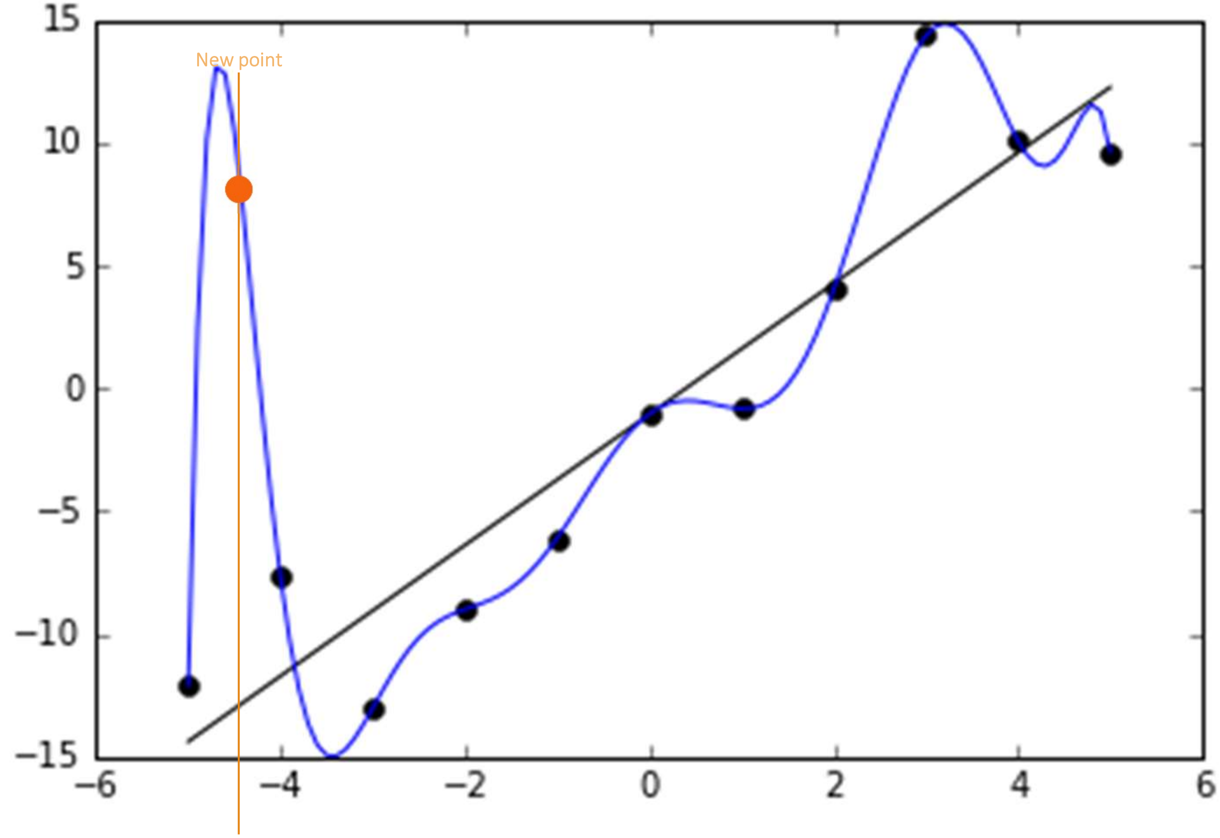


OVERFITTING

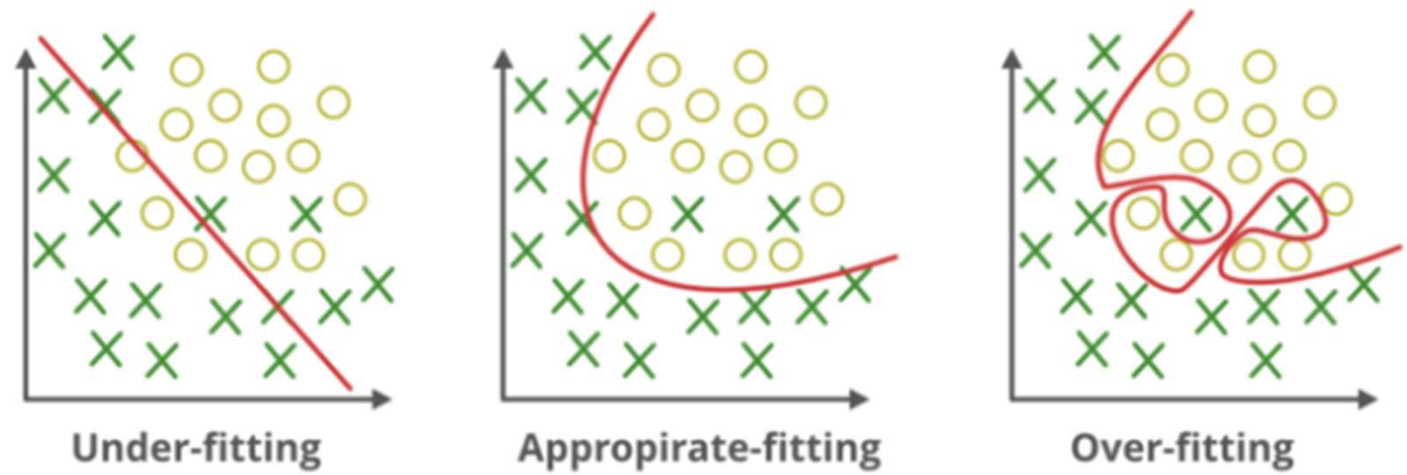


IRON HACK

OVERFITTING



OVERFITTING



OVERFITTING

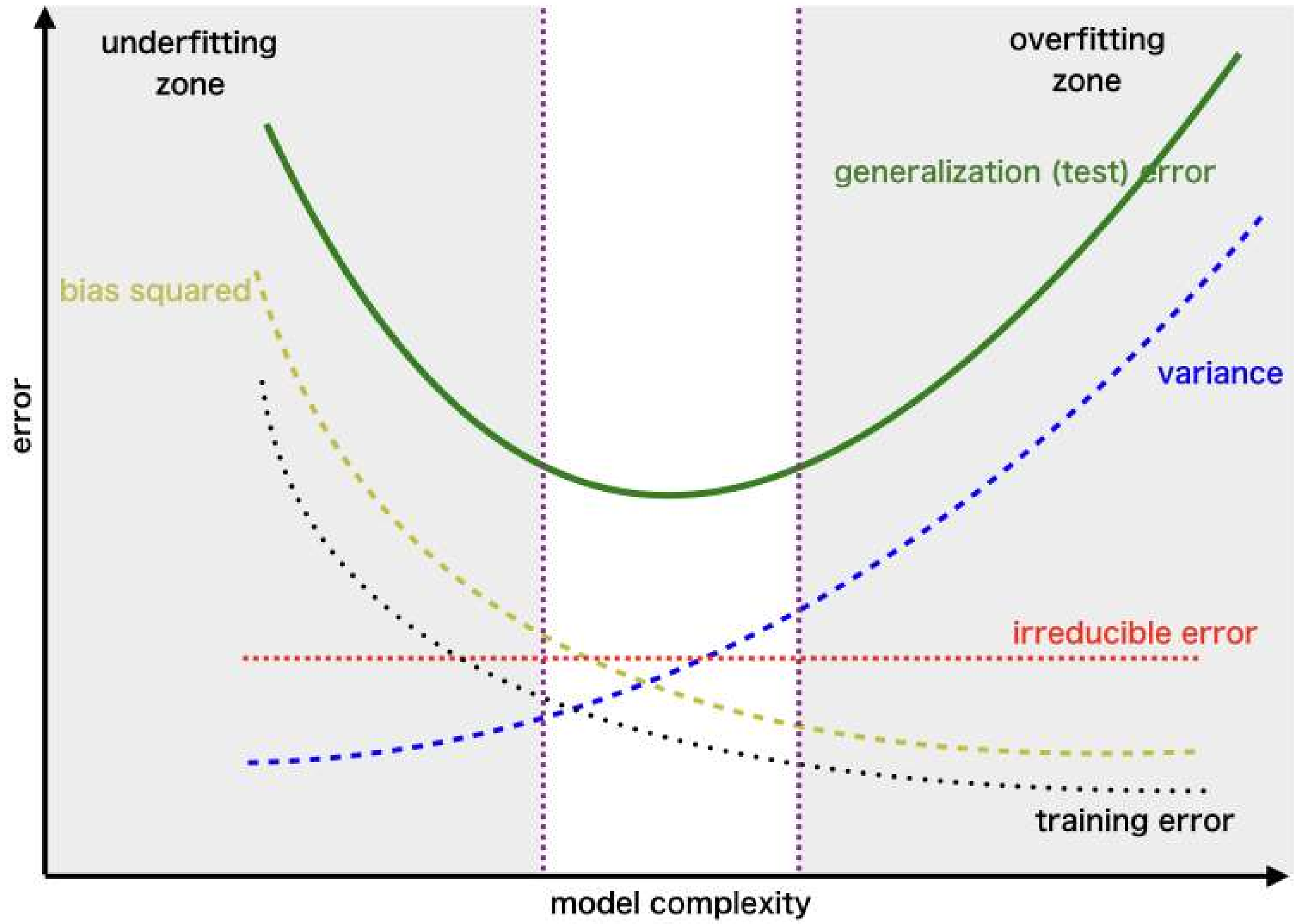
How do we cause overfitting?

Well, it depends a lot on the model!!!
Hence your need to understand it and not just using "model".fit(X_train, y_train)

How do we identify overfitting

Comparing comparing the evaluation of your model between the training labels and the test labels

OVERFITTING



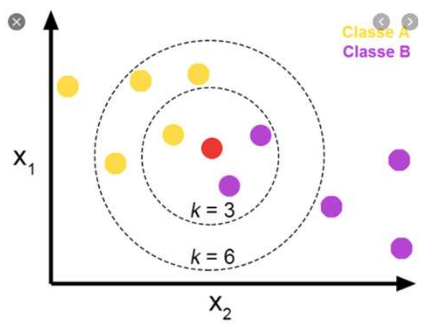
THANK YOU

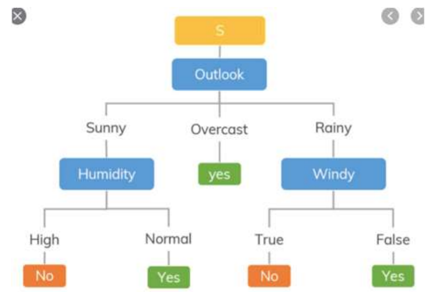
SUPERVISED LEARNING

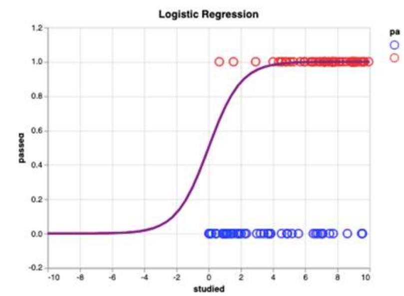
REVIEW AND APPLICATIONS

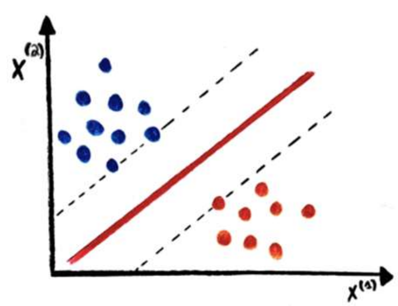
MACHINE LEARNING MODELS

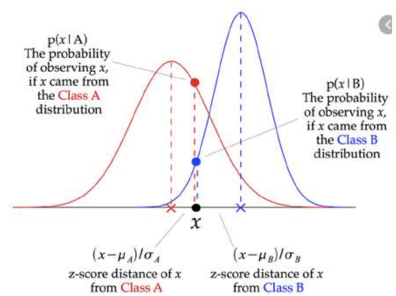
THERE ARE MANY MACHINE LEARNING MODELS…











MACHINE LEARNING MODELS

THERE ARE MANY MACHINE LEARNING MODELS...

You DONT NEED TO KNOW
how to code/build them.

!BUT!

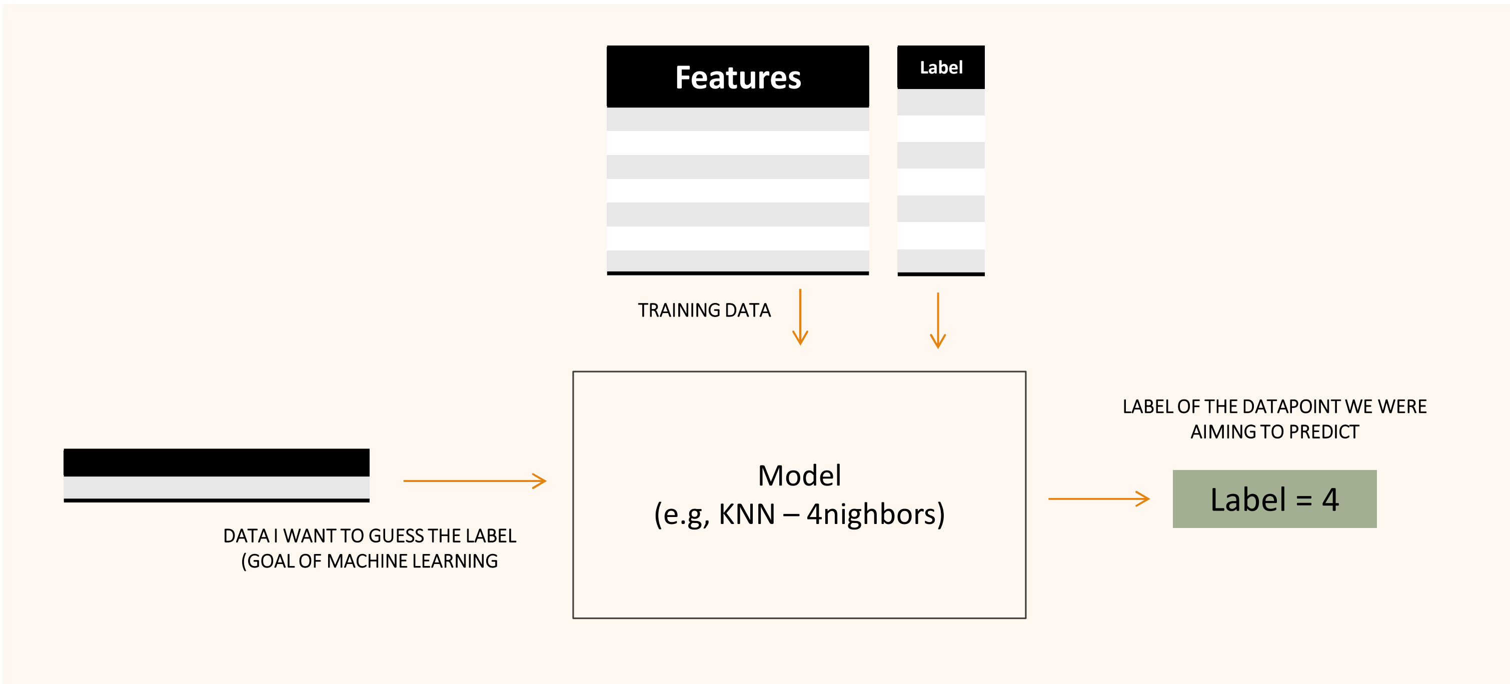
The more UNDERSTAND THEM,
the better you will know

what “CHOICES”
make sense

(e.g, Normalize or not, take care with overfit, etc)

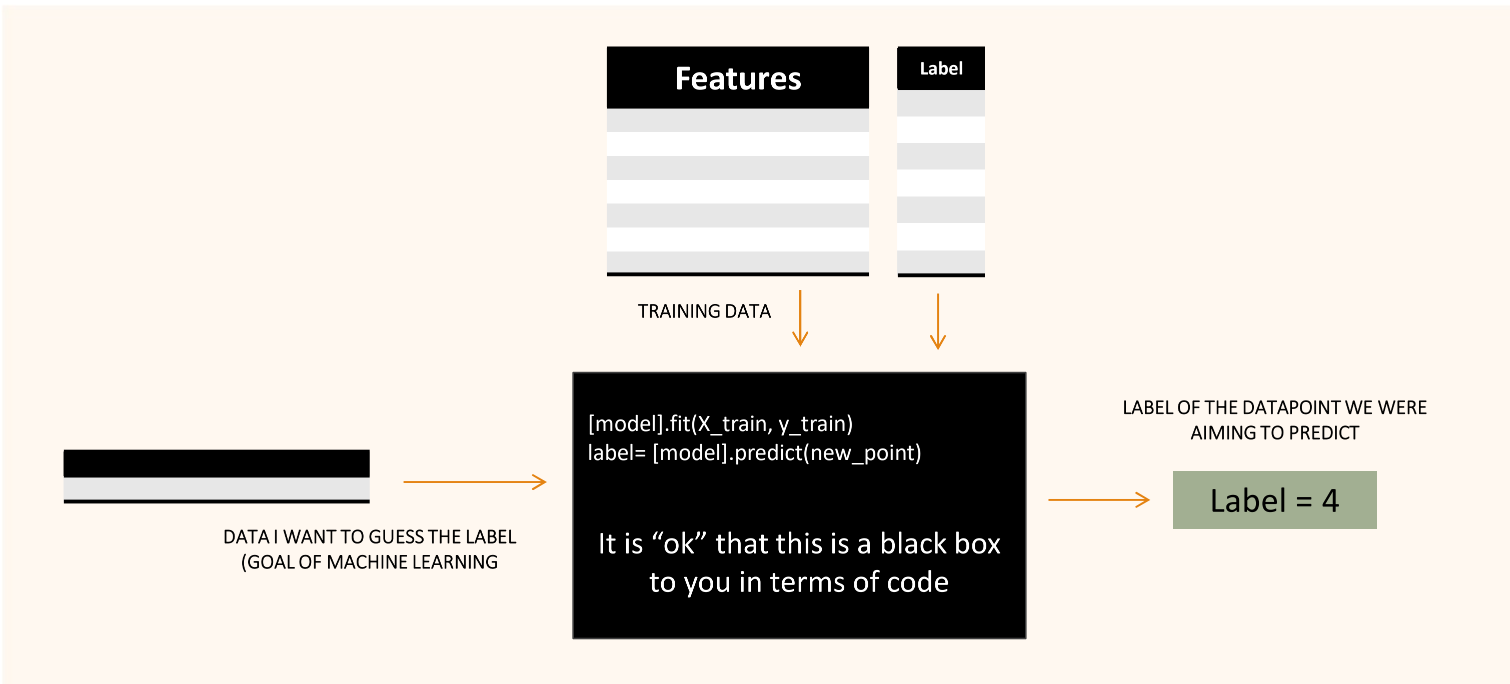
MACHINE LEARNING MODELS

LET'S SIMPLIFY

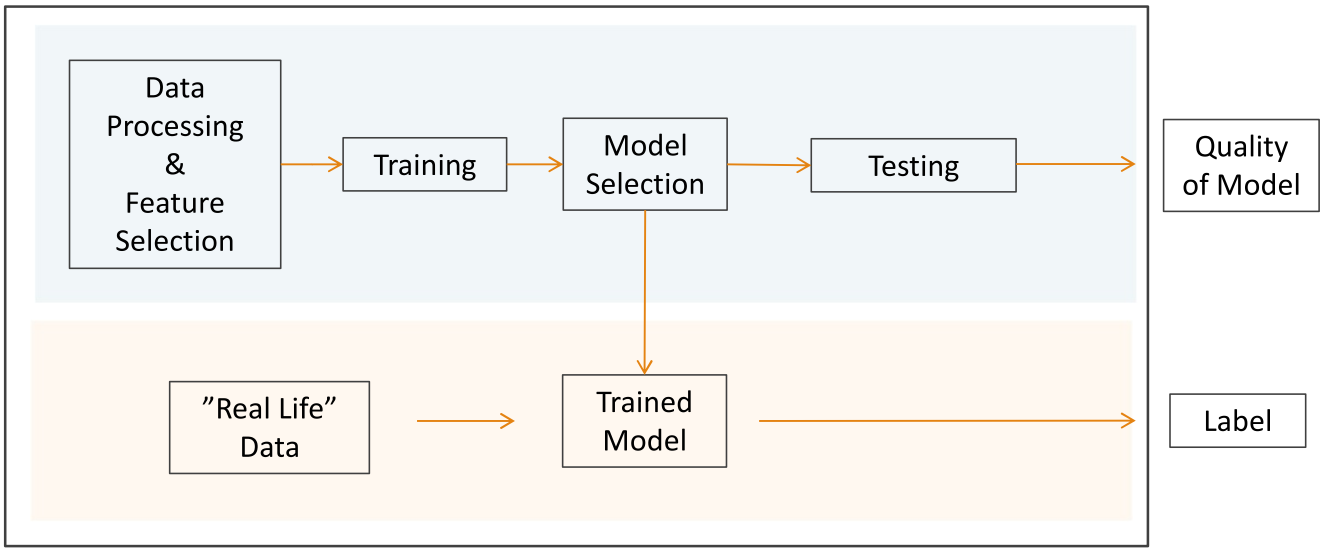


MACHINE LEARNING MODELS

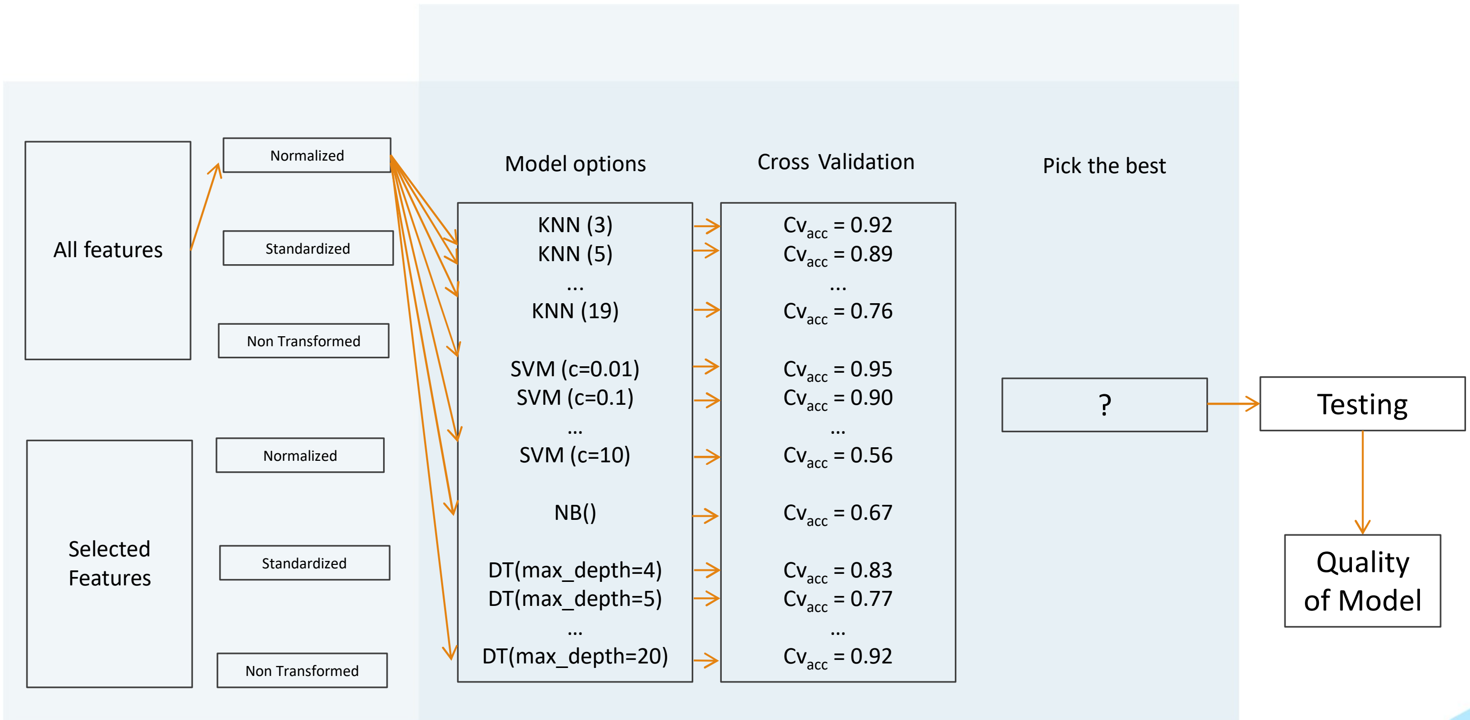
LET'S SIMPLIFY



MACHINE LEARNING WORKFLOW - CHOICES



MACHINE LEARNING WORKFLOW - CHOICES



MACHINE LEARNING WORKFLOW - CHOICES