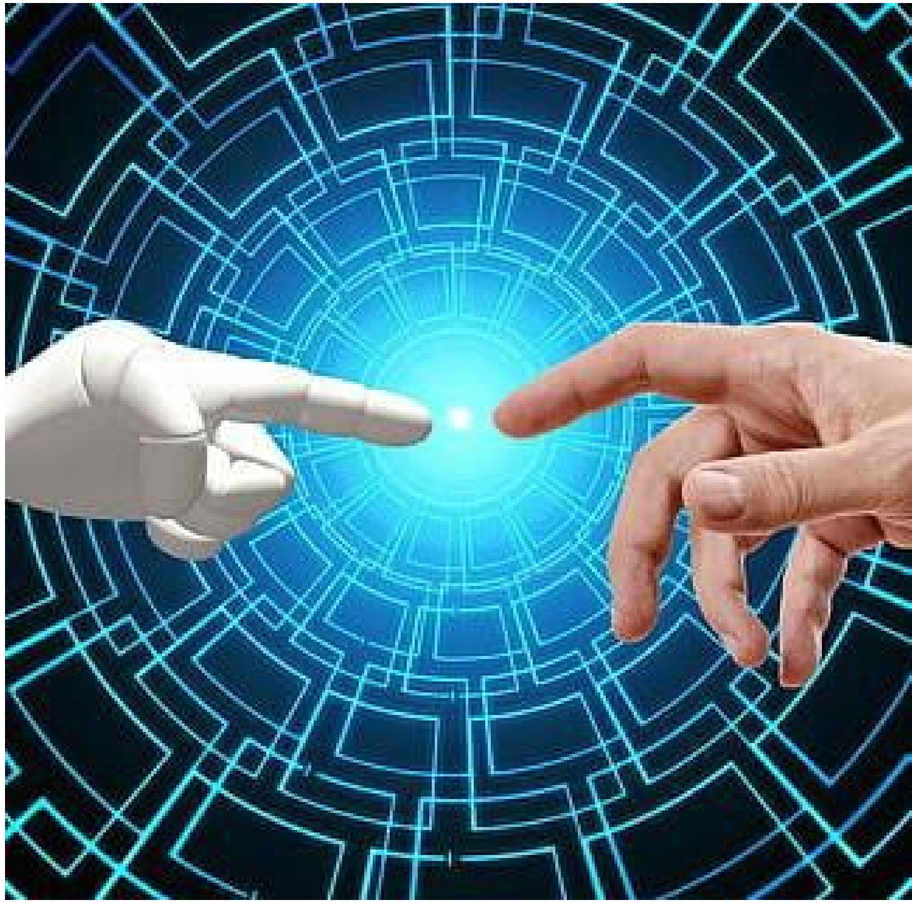
Artificial Intelligence (AI) for Engineering

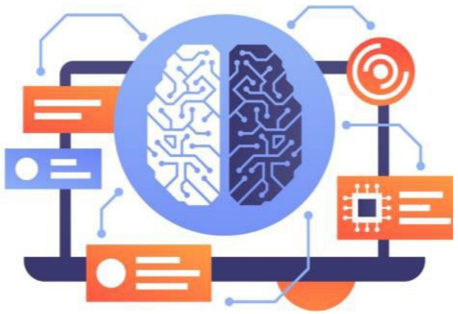
COS40007

Dr. Afzal Azeem Chowdhary

Lecturer, SoCET, Swinburne University of Technology

Seminar 2: 11th March 2025

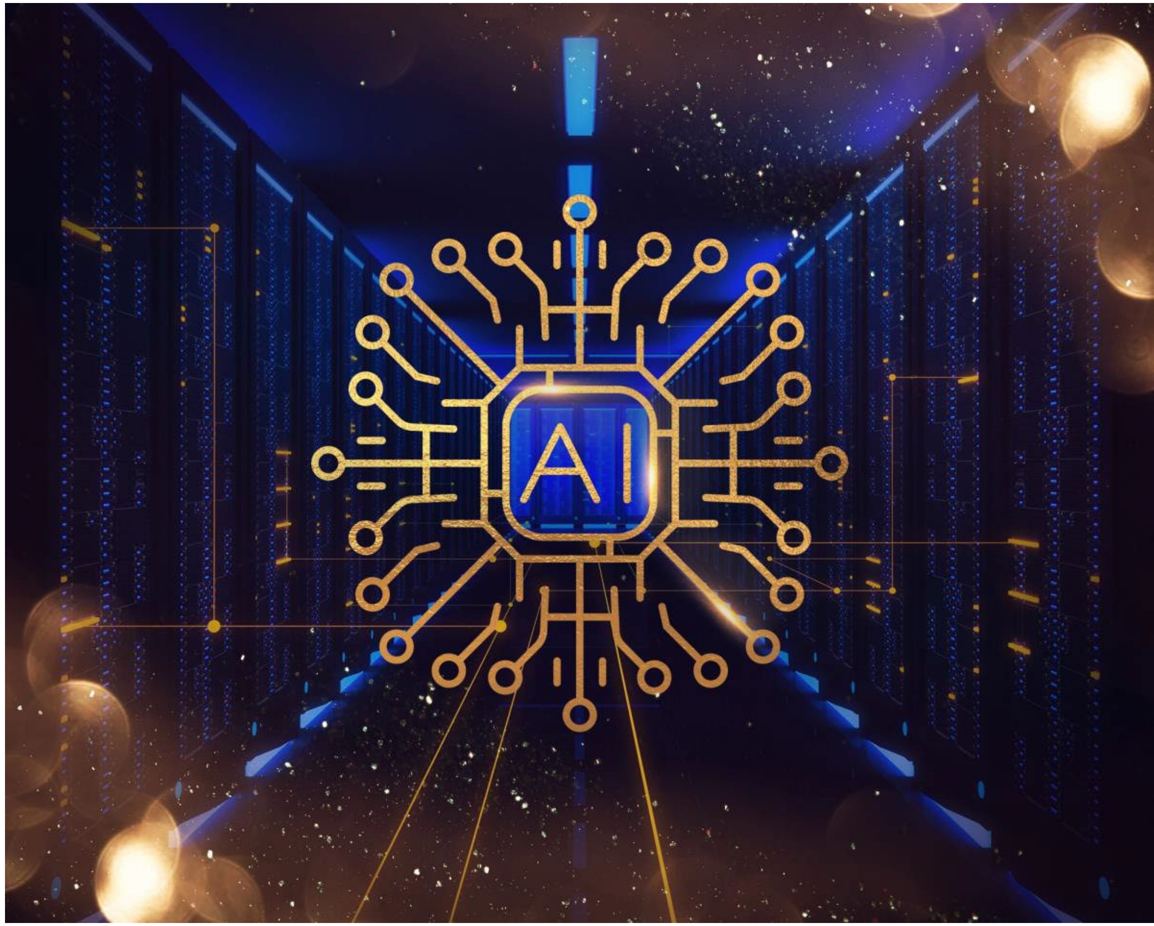




SWINBURNE UNIVERSITY OF TECHNOLOGY

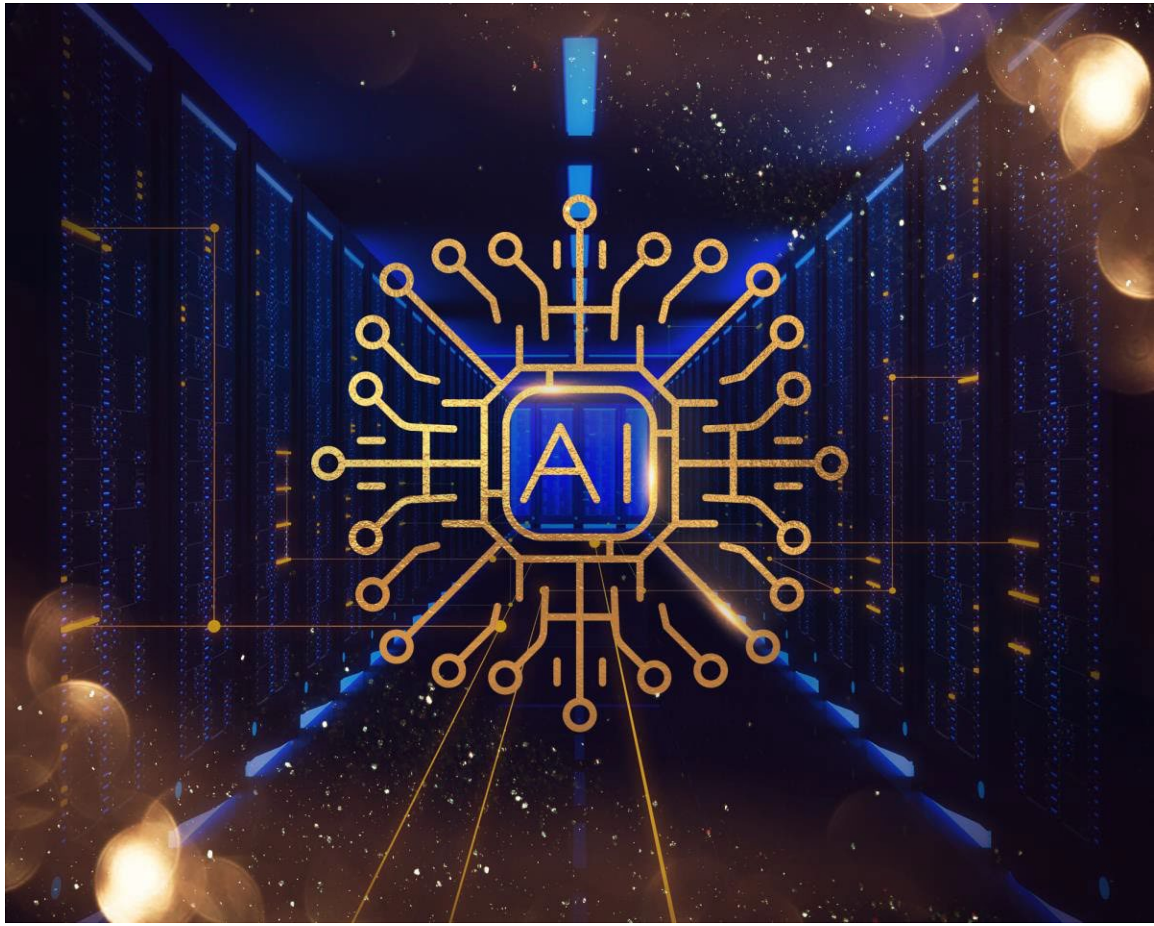
Overview

- ❑ Steps of Machine Learning
- ❑ Data collection
- ❑ Data cleaning and Feature Engineering
- ❑ Model Training
- ❑ Testing and Evaluation
- ❑ Model Improvement



Required Reading

- Chapter 1 and Chapter 4 of "Machine Learning with Pytorch and Scikit-Learn"
- A Reference Guide to Feature Engineering Methods

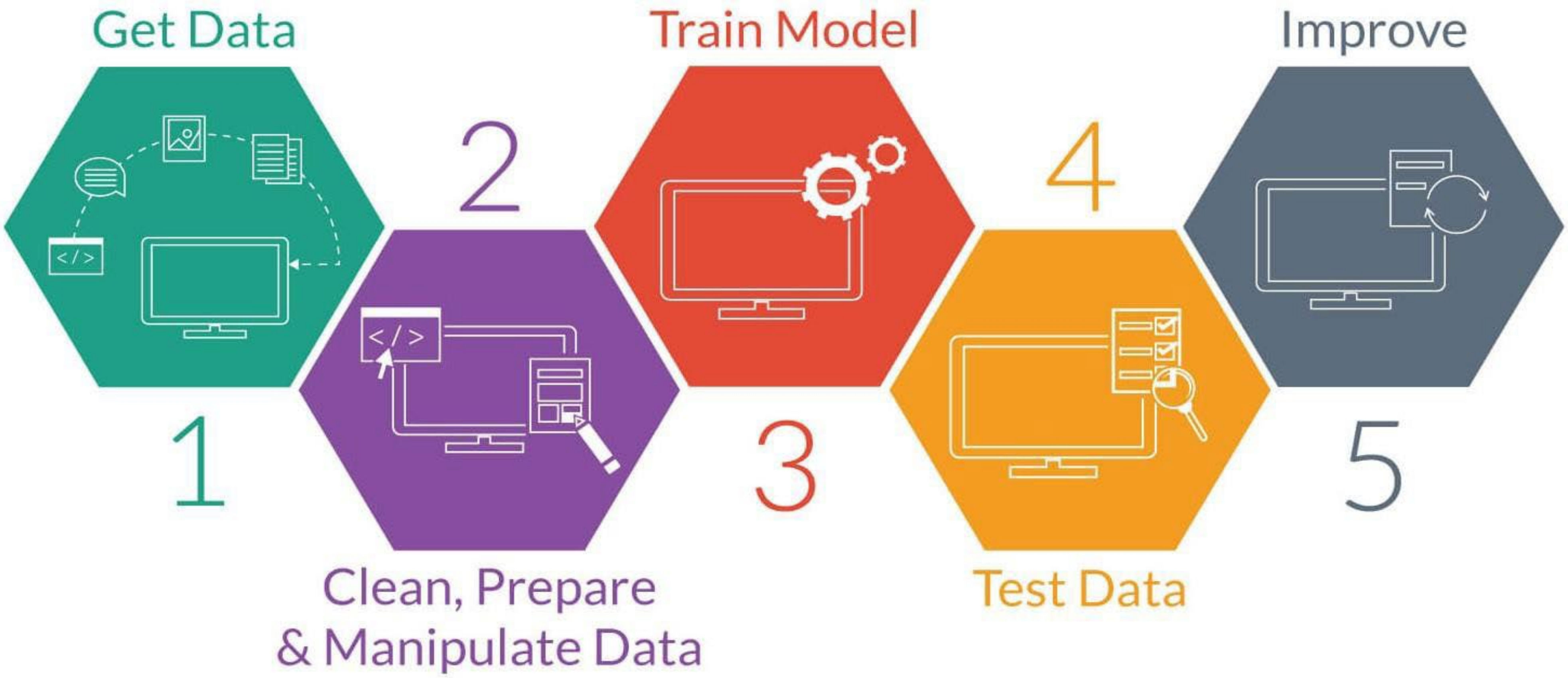


SWINBURNE UNIVERSITY OF TECHNOLOGY

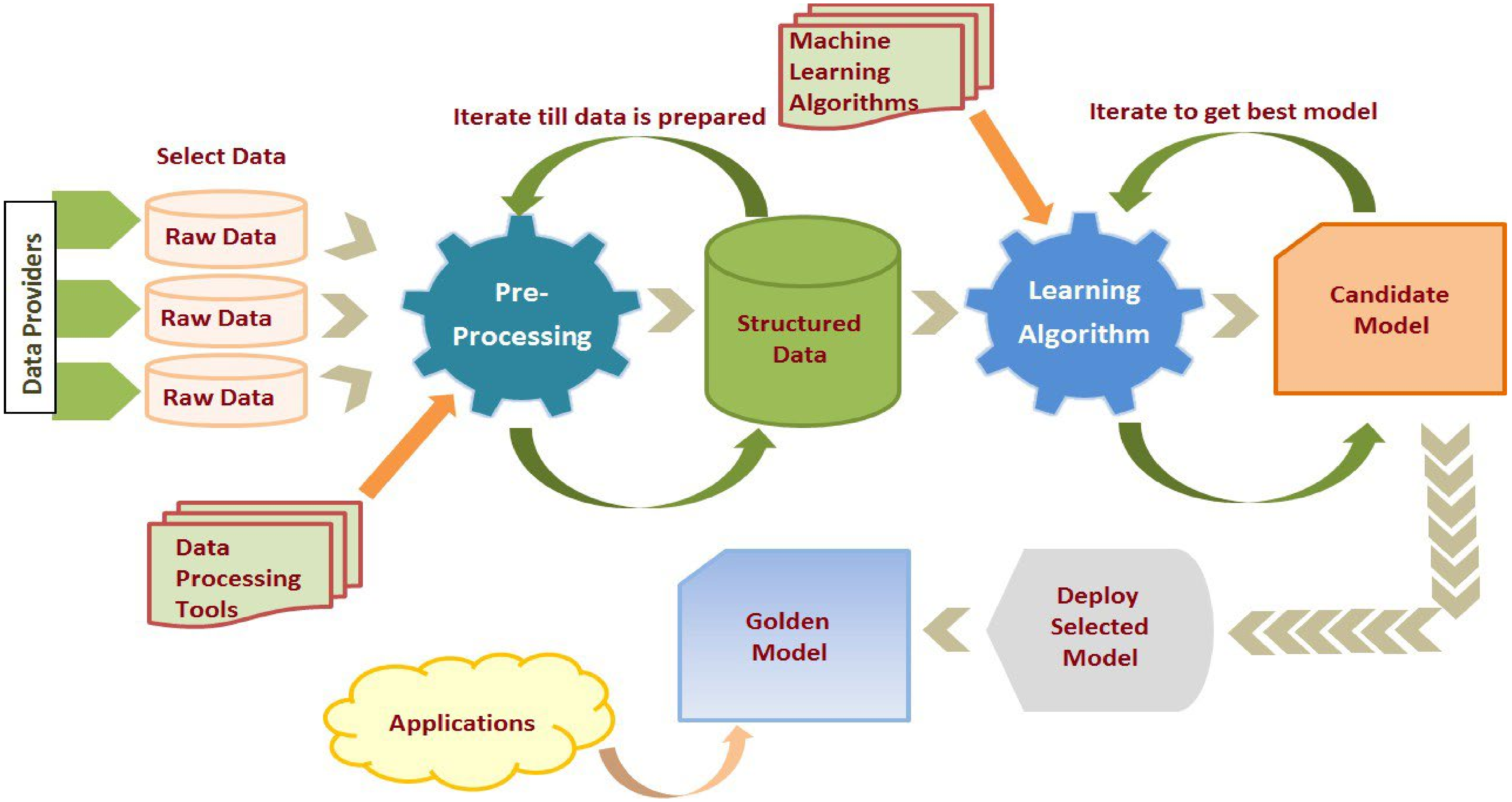
At the end of this you should be able to

- Understand the steps you need to complete to develop machine learning models
- Understand how to perform data pre-processing
- Understand model training and development process

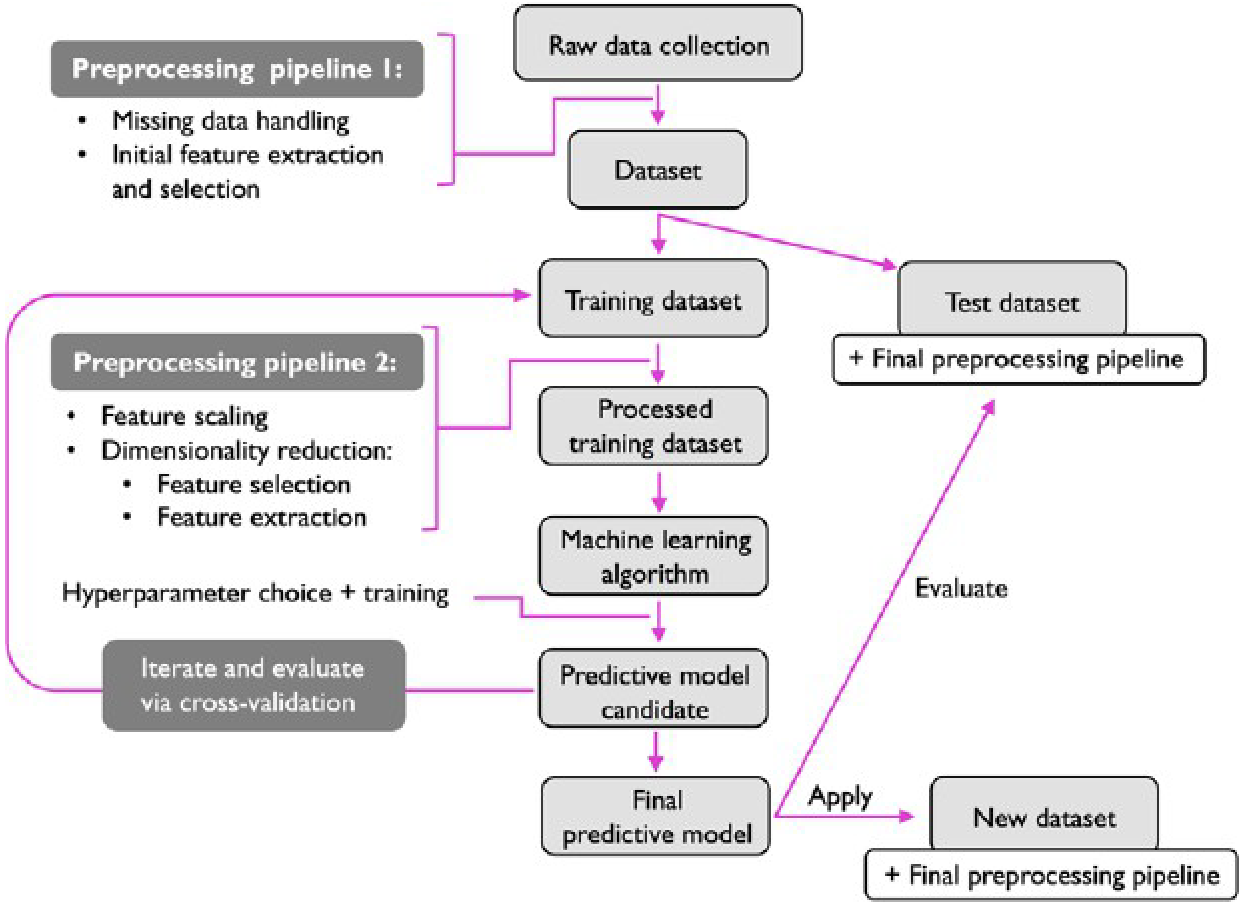
Steps of Machine Learning



How to use Machine Learning in applications ?



Workflow for predictive modelling



Data collection

Raw data

- Database tables
- Data file dumps from machine, processes
- Continuous time series data from sensors
- Images
- Videos
- Text

Raw data rarely comes in the form and shape that is necessary for the optimal performance of a learning algorithm. Thus, the preprocessing of the data is one of the most crucial steps in any machine learning application

Data preparation

- It is essential to have quality data that you can use to train your models
- If the data has minor discrepancies or missing information, it can greatly impact your model's accuracy.
- Data preparation takes 80% of the total data engineering effort
- Real-world data may be noisy or impure. Data preparation produces a narrower dataset than the source, which can boost data collection performance dramatically.

Example of Raw data

Given the strength of sonar, the prediction of whether or not an object is a mine or a rock returns at different angles.

```
0.0200,0.0371,0.0428,0.0207,0.0954,0.0986,0.1539,0.1601,0.3109,0.2111,0.1609,0.1582,0.2238,0.0645,0.0660,0.227
3,0.3100,0.2999,0.5078,0.4797,0.5783,0.5071,0.4328,0.5550,0.6711,0.6415,0.7104,0.8080,0.6791,0.3857,0.1307,0.2
604,0.5121,0.7547,0.8537,0.8507,0.6692,0.6097,0.4943,0.2744,0.0510,0.2834,0.2825,0.4256,0.2641,0.1386,0.1051,0
.1343,0.0383,0.0324,0.0232,0.0027,0.0065,0.0159,0.0072,0.0167,0.0180,0.0084,0.0090,0.0032,R
0.0453,0.0523,0.0843,0.0689,0.1183,0.2583,0.2156,0.3481,0.3337,0.2872,0.4918,0.6552,0.6919,0.7797,0.7464,0.944
4,1.0000,0.8874,0.8024,0.7818,0.5212,0.4052,0.3957,0.3914,0.3250,0.3200,0.3271,0.2767,0.4423,0.2028,0.3788,0.2
947,0.1984,0.2341,0.1306,0.4182,0.3835,0.1057,0.1840,0.1970,0.1674,0.0583,0.1401,0.1628,0.0621,0.0203,0.0530,0
.0742,0.0409,0.0061,0.0125,0.0084,0.0089,0.0048,0.0094,0.0191,0.0140,0.0049,0.0052,0.0044,R
0.0262,0.0582,0.1099,0.1083,0.0974,0.2280,0.2431,0.3771,0.5598,0.6194,0.6333,0.7060,0.5544,0.5320,0.6479,0.693
1,0.6759,0.7551,0.8929,0.8619,0.7974,0.6737,0.4293,0.3648,0.5331,0.2413,0.5070,0.8533,0.6036,0.8514,0.8512,0.5
045,0.1862,0.2709,0.4232,0.3043,0.6116,0.6756,0.5375,0.4719,0.4647,0.2587,0.2129,0.2222,0.2111,0.0176,0.1348,0
.0744,0.0130,0.0106,0.0033,0.0232,0.0166,0.0095,0.0180,0.0244,0.0316,0.0164,0.0095,0.0078,M
0.0100,0.0171,0.0623,0.0205,0.0205,0.0368,0.1098,0.1276,0.0598,0.1264,0.0881,0.1992,0.0184,0.2261,0.1729,0.213
1,0.0693,0.2281,0.4060,0.3973,0.2741,0.3690,0.5556,0.4846,0.3140,0.5334,0.5256,0.2520,0.2090,0.3559,0.6260,0.7
340,0.6120,0.3497,0.3953,0.3012,0.5408,0.8814,0.9857,0.9167,0.6121,0.5006,0.3210,0.3202,0.4295,0.3654,0.2655,0
.1576,0.0681,0.0294,0.0241,0.0121,0.0036,0.0150,0.0085,0.0073,0.0050,0.0044,0.0040,0.0117,R
0.0762,0.0666,0.0481,0.0394,0.0590,0.0649,0.1209,0.2467,0.3564,0.4459,0.4152,0.3952,0.4256,0.4135,0.4528,0.532
6,0.7306,0.6193,0.2032,0.4636,0.4148,0.4292,0.5730,0.5399,0.3161,0.2285,0.6995,1.0000,0.7262,0.4724,0.5103,0.5
459,0.2881,0.0981,0.1951,0.4181,0.4604,0.3217,0.2828,0.2430,0.1979,0.2444,0.1847,0.0841,0.0692,0.0528,0.0357,0
.0085,0.0230,0.0046,0.0156,0.0031,0.0054,0.0105,0.0110,0.0015,0.0072,0.0048,0.0107,0.0094,R
```

Example of Raw data

Time-series data for x,y,z acceleration from accelerometer sensor

timestamp	x	y	z
1502851906	0.371338	0.575684	0.69751
1502851906	0.21875	0.470215	0.672607
1502851906	0.161377	0.360107	0.707764
1502851906	0.164307	0.302734	0.666504
1502851906	0.243652	0.258545	0.632813
1502851906	0.326172	0.226074	0.577637
1502851906	0.358643	0.196045	0.577393
1502851906	0.428223	0.176025	0.656738
1502851906	0.460205	0.172363	0.603516
1502851906	0.47876	0.134521	0.559326
1502851906	0.411865	0.112793	0.531738
1502851906	0.384033	0.085938	0.487061
1502851906	0.404053	0.042236	0.435059
1502851906	0.428955	-0.01685	0.387207
1502851906	0.175293	-0.08179	0.121094
1502851906	-0.12915	-0.03638	-0.00732
1502851906	-0.39819	-0.02417	-0.04663

SWINBURNE UNIVERSITY OF TECHNOLOGY

Example of Raw data

Server time	Laptop time	Refrac date	Refrac time	Measured solids	Sample temperature	Certainty	Part	Scale	
27/08/2019 3:46	27/08/2019 3:46	27/08/2011	00:53:30	65.5	20	26	brd	bx	no
27/08/2019 4:11	27/08/2019 4:11	27/08/2011	01:18:10	65	20	25	brd	bx	no
27/08/2019 4:40	27/08/2019 4:40	27/08/2011	01:46:55	66.2	20	25	brd	bx	no
27/08/2019 5:00	27/08/2019 5:00	27/08/2011	02:07:46	69	20	28	brd	bx	no
27/08/2019 5:01	27/08/2019 5:01	27/08/2011	02:08:47	68.6	20	27	brd	bx	no
27/08/2019 5:04	27/08/2019 5:04	27/08/2011	02:11:37	68.1	20	25	brd	bx	no
27/08/2019 5:21	27/08/2019 5:21	27/08/2011	02:27:58	67.4	20	22	brd	bx	no
27/08/2019 5:47	27/08/2019 5:47	27/08/2011	02:54:07	67.2	20	21	brd	bx	no
27/08/2019 5:51	27/08/2019 5:51	27/08/2011	02:58:20	41.3	20	28	vegmite	bx	no
27/08/2019 6:08	27/08/2019 6:08	27/08/2011	03:15:44	62.6	20	21	vegmite	bx	no
27/08/2019 6:12	27/08/2019 6:12	27/08/2011	03:19:41	62.3	20	21	vegmite	bx	no
27/08/2019 8:06	27/08/2019 8:06	27/08/2011	05:13:21	0	20	101	vegmite	bx	no
27/08/2019 8:45	27/08/2019 8:45	27/08/2011	05:51:53	66.3	20	22	vegmite	bx	no
27/08/2019 8:49	27/08/2019 8:49	27/08/2011	05:56:39	65.1	20	21	vegmite	bx	no
27/08/2019 9:42	27/08/2019 9:42	27/08/2011	06:49:34	63.9	20	23	vegmite	bx	no

SWINBURNE UNIVERSITY OF TECHNOLOGY

Example of Raw data

RunId	RunDateTime	WorkOrderId	ProductId	RunType	QtyPlanned	QtyActual	QtyUofm	UserId	Notes	IsCompleted	IsCurrent
425	32:42.2	WO451	2		725.4	726	KG	1		1	0
426	36:04.2	WO451	3		6.25	7.5	KG	1		1	0
427	50:39.4	WO451	4		81.25	81.5	KG	1		1	0
428	01:43.9	WO451	24		600	600	KG	1		1	0
429	04:53.1	WO451	16		18	18	KG	1		1	0
430	08:54.7	WO451	24		600	240	KG	1		1	0
431	14:30.8	WO451	7		100	101.5	KG	1		1	0
432	25:35.1	WO451	24		600	600	KG	1		1	0
433	31:25.6	WO451	19		1	2.5	KG	1		1	0
434	31:41.6	WO451	24		600	16	KG	1		1	0
435	31:45.1	WO451	24		600	8	KG	1		1	0
436	39:50.7	WO451	18		0.12	2.5	KG	1		1	0
437	39:57.6	WO451	24		600	7	KG	1		1	0
438	40:01.5	WO451	20		1	0	KG	1		1	0
439	40:05.7	WO451	24		600	5	KG	1		1	0
440	43:30.5	WO451	21		40	40.5	KG	1		1	0
441	46:46.9	WO451	24		600	195	KG	1		1	0
442	49:21.0	WO451	2		80.6	84	KG	1		1	0
443		WO461	2		805.5		KG	1		0	0
444		WO461	24		600		KG	1		0	0
445		WO461	3		11		KG	1	Increase stirrer	0	0
446		WO461	24		600		KG	1		0	0
447		WO461	4		25		KG	1		0	0
448		WO461	24		600		KG	1		0	0
449		WO461	8		23.2		KG	1	Reduce stirrer	0	0
450		WO461	24		600		KG	1		0	0
451		WO461	7		90		KG	1		0	0

A	B	C	D	E	
ProductId	ProductCode	ProductName	ProductType	BatchSize	Note
1	1	Assembly Stock	Manufactured	1000	
2	Water	Water	Raw Material		
3	PC0041	Pearl Caustic	Raw Material		
4	PN0045	Sodim Tripol	Raw Material		
5	PD0011	Fluorsceine	Raw Material		
6	PD0032	Tartrazine Yel	Raw Material		
7	LC0040	Product description	Raw Material		
8	LN0114	Product description	Raw Material		
9	LF0019	Lemon Fragrance	Raw Material		
10	PC0042	Sodium Meta	Raw Material		
11	PC0002	Product description	Raw Material		
12	LN0010	Butyl Glycoether	Raw Material		
13	LN0070	SLES 70%	Raw Material		
14	LF0019	Lemon Fragrance	Raw Material		
15	LN0120	Surfactant	Raw Material		
16	LC0057	Triethanolamine	Raw Material		
17	LC0021	Labs Acid	Raw Material		
18	LD0163	Product description	Raw Material		
19	LC0013	Acticide	Raw Material		
20	LF0099	Toasted Coconut	Raw Material		
21	LN0005	CDE 80	Raw Material		
22	2	T2000	Manufactured	1000	
23	3	Bulldog Blue	Manufactured	1000	
24	Mixing	Manual Action	Recipe Action		
25	4	Wash & Shine	Manufactured	1000	
26	PD0067	CARMOISINE	Raw Material		
27	LN0088	Silicone Emul	Raw Material		

LogId	WorkOrde	LogDateTime	SensorId	LogDescription	LogData
19582	WO451	09:33.8	2	Current Value	0
19583	WO451	09:33.9	1	Weight Value	1
19584	WO451	09:39.6	1	Weight Value	0.5
19585	WO451	09:40.9	1	Weight Value	0
19586	WO451	09:41.3	1	Weight Value	0.5
19587	WO451	09:42.0	1	Weight Value	0
19588	WO451	09:42.4	1	Weight Value	0.5
19589	WO451	09:43.6	1	Weight Value	8.5
19590	WO451	09:44.0	1	Weight Value	25
19591	WO451	09:44.4	1	Weight Value	24
19592	WO451	09:44.8	1	Weight Value	23.5
19593	WO451	09:45.2	1	Weight Value	0.5
19594	WO451	09:46.8	1	Weight Value	0
19595	WO451	09:47.2	1	Weight Value	5.5
19596	WO451	09:47.6	1	Weight Value	26.5
19597	WO451	09:48.0	1	Weight Value	28
19598	WO451	09:48.4	1	Weight Value	27
19599	WO451	09:49.2	1	Weight Value	27.5

SWINBURNE UNIVERSITY OF TECHNOLOGY

Data Pre-processing

Data cleaning

- Remove constant feature – does not have an impact on the outcome
- Remove irrelevant feature – id values
- Remove duplicate features (across columns) and samples (across rows) – because this causes data imbalance and over-fitting during training
- Identify and remove outliers – as they fall well outside decision boundary and can skew your data
- Identify and remove highly correlated features - Some features may be highly correlated and therefore redundant to a certain degree (because they same information about the target variable)

Data Imputation

- Detect missing features / incorrect or missing values
- Detecting missing features can be done by plotting the histogram of each feature
  - unusual outlier spikes indicate the use of special values,
  - a spike in the middle of the distribution is a sign that mean/median imputation has already been performed.
- To fix missing features
  - Sometimes, use the entire feature's mean/median/mode for imputation.
  - For time-series data, impute using value repetition or interpolation is good.

Data Imputation

- Categorical Imputation: Missing categorical variables are generally replaced by the most commonly occurring value in other records
- Numerical Imputation: Missing numerical values are generally replaced by the mean of the corresponding value in other records

Discretization

- Discretization involves taking a set of data values and grouping sets of them together logically into bins
- Binning can apply to numerical values as well as to categorical data values.
- Grouping of equal intervals (e.g., from seconds to minute)
- Grouping based on equal frequencies
- Grouping based on sorting

Feature Encoding

- Ordinal features (such as age) may have integer values, but they differ from numeric features
- Tree-based models can use label-encoding (i.e. fixed strings or integers denoting class membership) and don't need further preprocessing.
- Non-tree methods require that categorical features be one-hot encoded (each category is converted to variable with value 0/1)

Normalisation

- Scaling or normalisation is suitable for achieving low training loss particularly for non-tree-based methods.
- Numerical features can often benefit from transformations. Log transformation, np.log(1 + x), is a powerful transformation that is particularly helpful when a feature observes a power-law relationship.

Dimensionality Reduction

- Dimensionality Reduction techniques are useful for compressing the features onto a lower-dimensional subspace.
- Reducing the Dimensionality of our feature space has the advantage of requiring less storage space, and the learning algorithm can run much faster.
- Improve the predictive performance of a model if the dataset contains a large number of irrelevant features (or noise).

Data Shuffling

- During preprocessing, it’s important to shuffle your dataset before splitting it into train/validation/test subsets.
- Utilize the stratify feature of sklearn.model_selection.train_test_split() to ensure a consistent distribution of your minority targets across all your subsets.
- Help our machine learning algorithm not only perform well on the training dataset but also generalise well to new data.

Feature Generation

- Mapping existing features into a new space, for example, the date -> day of the week.
- Combining multiple features into a composite. Example: sum of 2 columns.
- Aggregating data to find patterns: Example: mean values of each minute per second time-series data.
- Merging auxiliary data.

Train Model

- It is essential to compare at least a handful of different learning algorithms to train and select the best-performing model
- Different techniques summarised as “cross-validation” can be used for validation during the training process
- In cross-validation, the dataset is further divided into training and validation subsets to estimate the model’s generalisation performance.

Train Model : Parameter Tuning

- We cannot expect that the default parameters of the different learning algorithms provided by software libraries are optimal for our specific problem task
- Frequent use of hyperparameter optimisation techniques that help us to fine-tune the performance of our model
- We can think of those hyperparameters as parameters that are not learned from the data but represent the knobs of a model that we can turn to improve its performance

Train Model : Parameter Tuning

- We cannot expect that the default parameters of the different learning algorithms provided by software libraries are optimal for our specific problem task.
- Frequent use of hyperparameter optimisation techniques that help us to fine-tune the performance of our model.
- We can think of those hyperparameters as parameters that are not learned from the data but represent the knobs of a model that we can turn to improve its performance.

Evaluating models

- After we have selected a model that has been fitted on the training dataset, we can use the test dataset to estimate how well it performs on this unseen data
- If we are satisfied with its performance, we can now use this model to predict new, future data.
- Data must also be in a pre-processed format for the test dataset.
- One commonly used metric for evaluation is accuracy, which is defined as the proportion of correctly classified instances

Learn, Practice and Enjoy the AI journey