COMP309/AIML421 — ML Tools and Techniques

Week 1-Lecture 2

Machine Learning Tasks

Dr Qi Chen

School of Engineering and Computer Science

Victoria University of Wellington

Qi.Chen@ecs.vuw.ac.nz

Week Overview

★ AI and Machine Learning

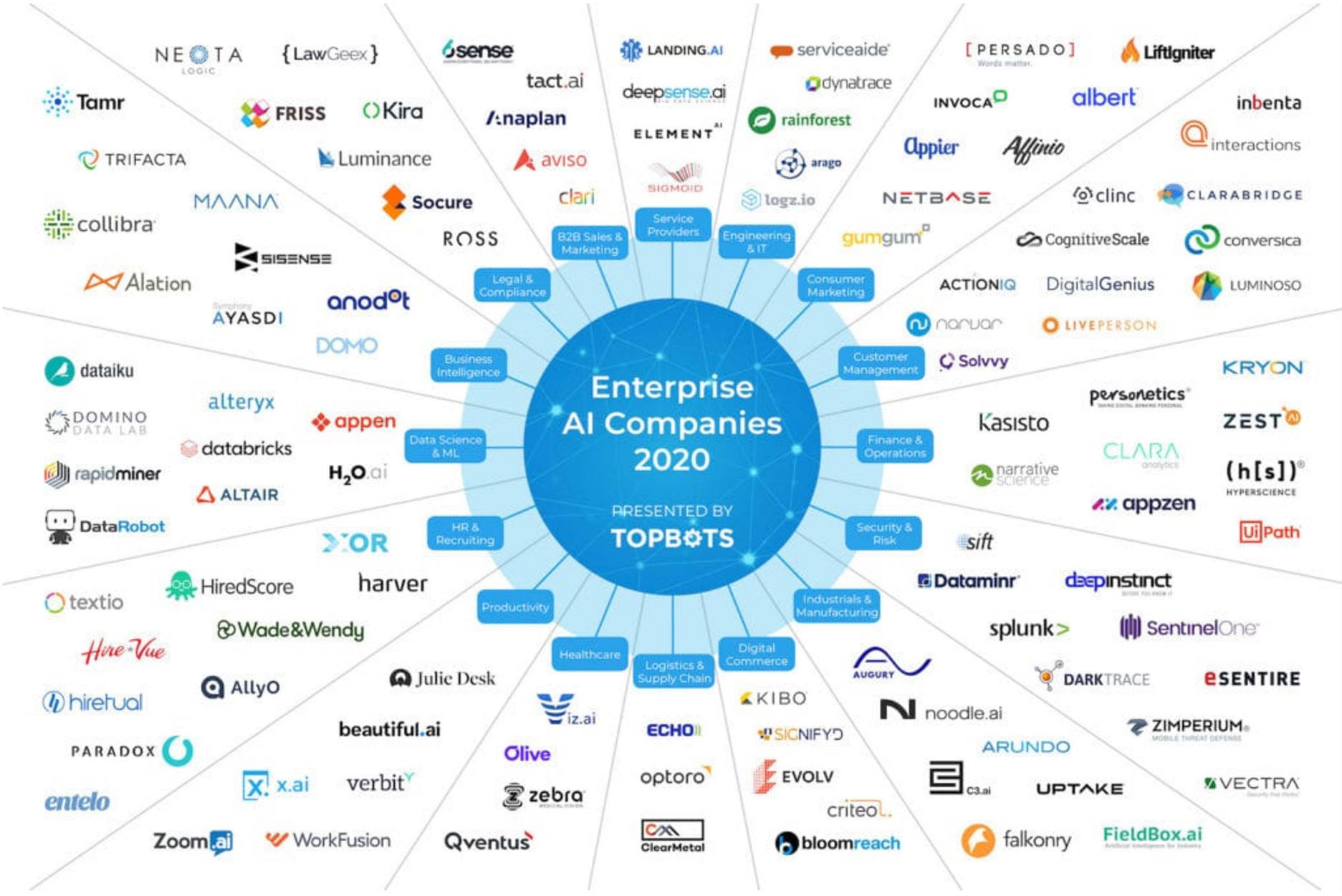
★ Machine Learning Scope: Data, Task, Model, and Algorithm

★ Data handle by Machine Learning

★ Machine Learning Tasks

Artificial Intelligence Companies

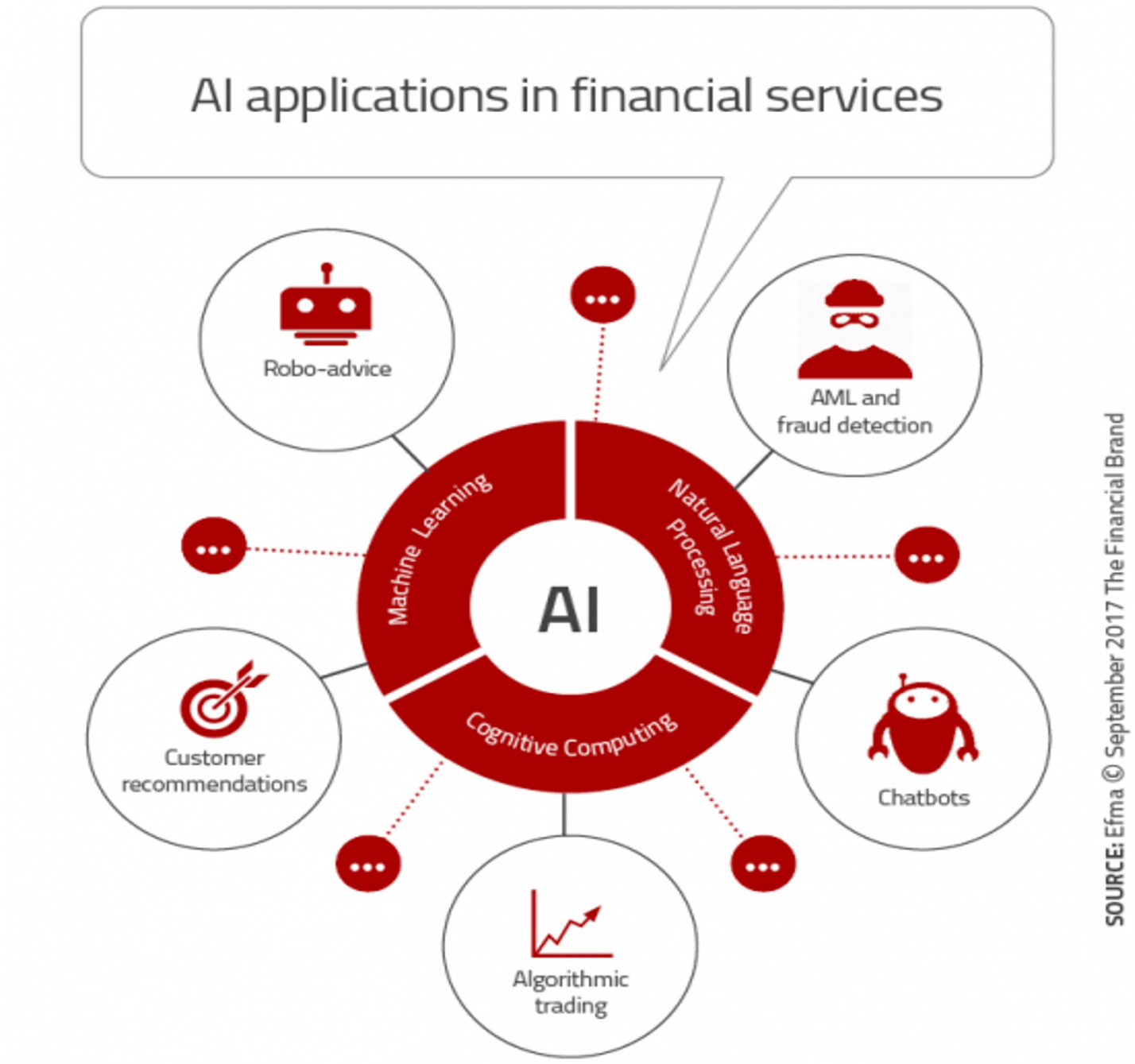
Artificial Intelligence Companies



cf in Aotearoa :

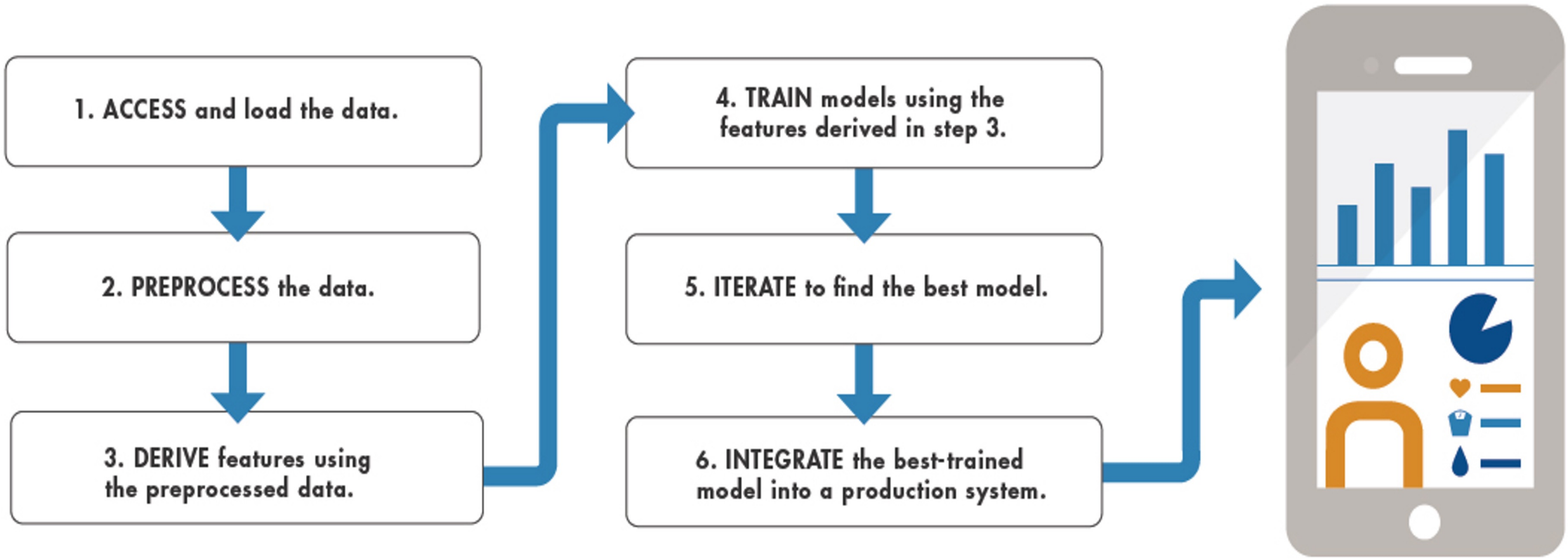
https://newzealand.ai/nz-ai-companies

Not This - This is Application



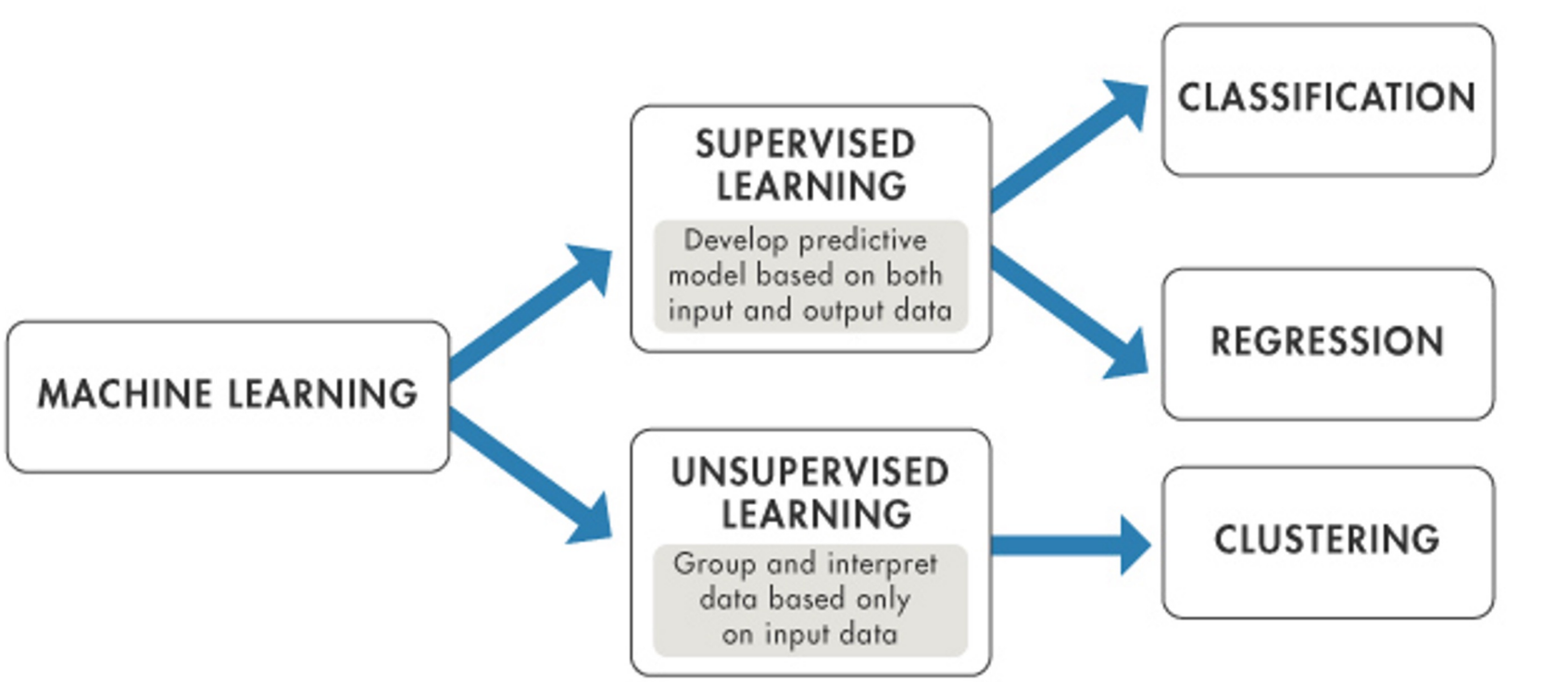
Not this either - this is the process of ML

This is the process of ML



https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

MATLAB gives the following ontology of ML



But it’s a lot richer than this

https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

Supervision available to the learner

lots:

Supervised learning: The environment contains a teacher that provides the correct response for certain environmental states. The goal is for the learner to output the correct response: “do what the teacher would do”.

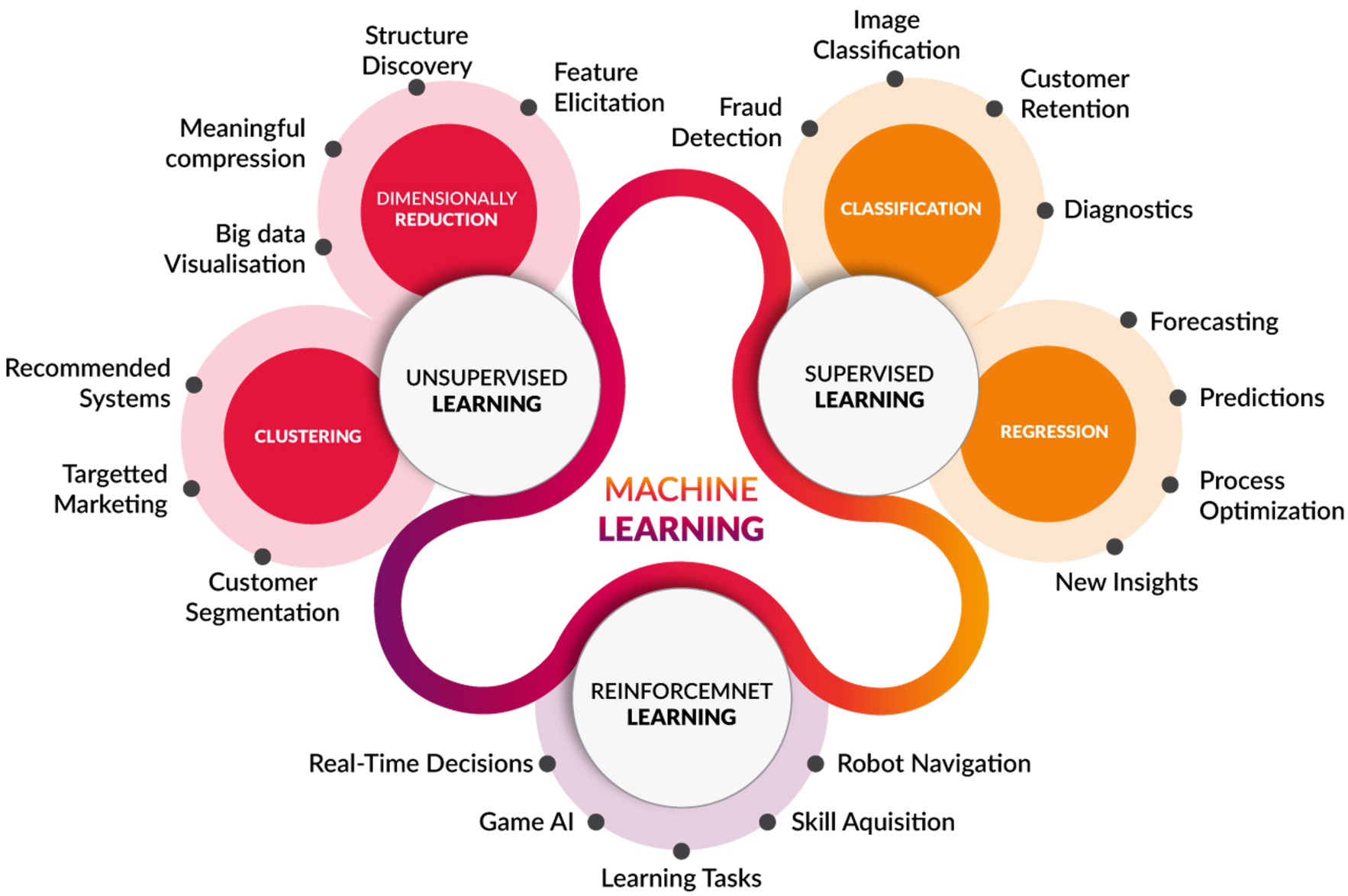
none:

Unsupervised learning: No such teacher to say whether the learner’s output is correct. Instead the learning system has an internally defined teacher with a prescribed goal that does not need utility feedback of any kind.

some:

Reinforcement learning: Again, no such teacher to say whether the learner’s output is correct. Instead of a label, the environment provides reward or punishment to indicate the utility of actions that were actually taken by the system.

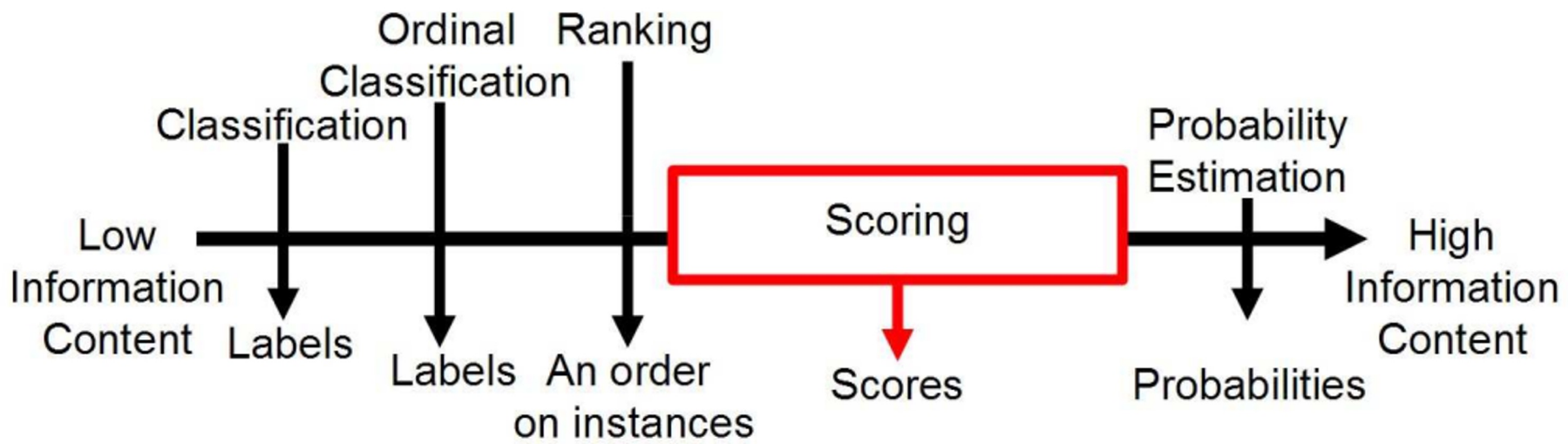
3-way split



http://www.cognub.com/index.php/cognitive-platform/

Outcomes (outputs) – can be simple… or more informative

- Notice that different outcomes can have different amounts of information content
- *e.g. here, a classifier classifies☺ some novel input pattern*

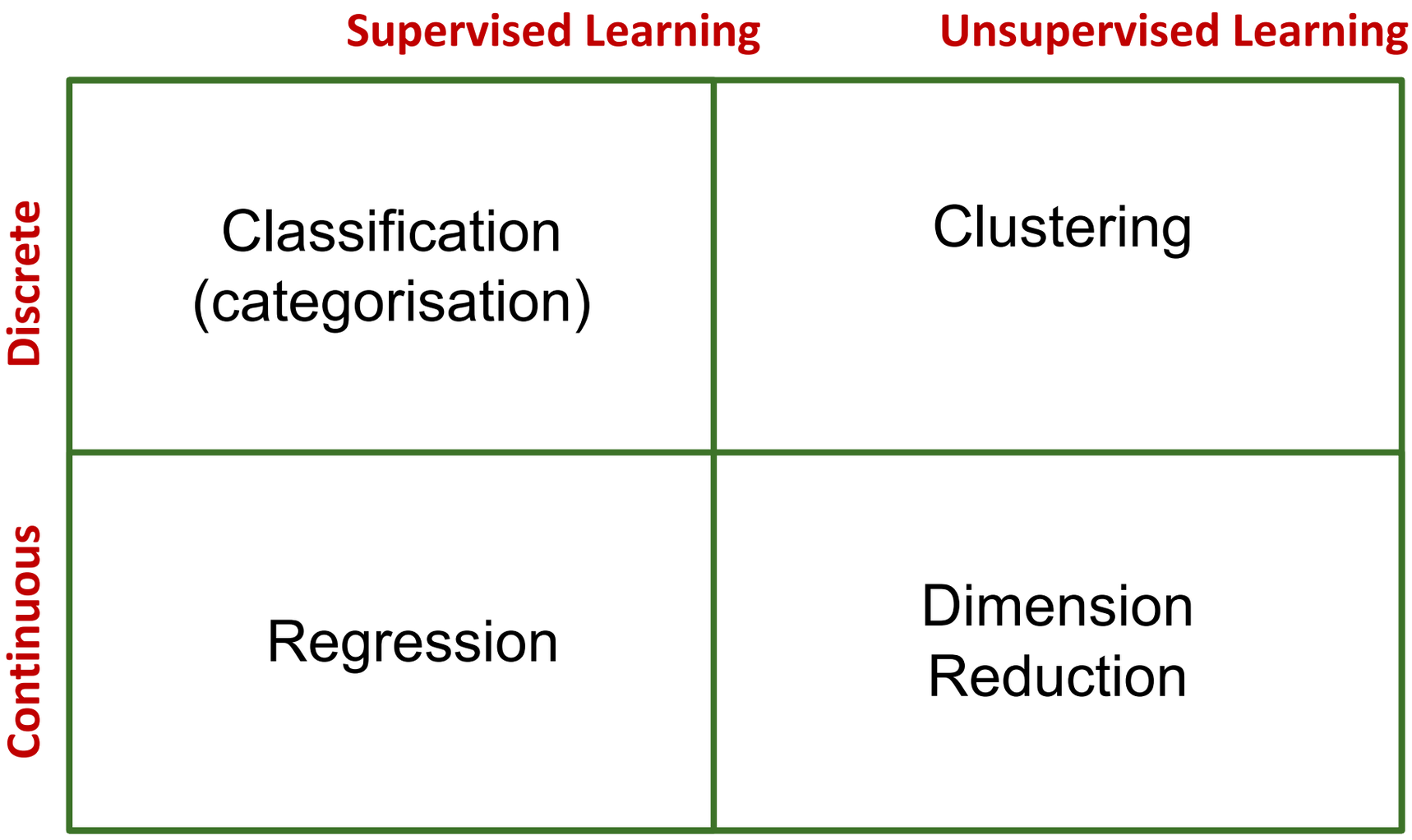


Prediction Outcomes

notice we can think of the classifier's response a "prediction", of a kind

https://www.researchgate.net/publication/220699539_Smooth_Receiver_Operating_Characteristics_smROC_Curves

The main divisions within learners



(leaving Reinforcement learning aside)

The main divisions within learners

	Supervised Learning	Unsupervised Learning
Discrete	Classification (categorisation)	Clustering
Continuous	Regression	Dimension Reduction

Semi-supervised Learning?

Adapted from: https://www.slideshare.net/GirishKhanzode/supervised-learning-52218215

Based on Types of Response

Basic/Immediate: The learning system responds immediately to the input with an output.

(classification tasks, e.g. disease prognosis...)

Delayed: The learning system may require some history of inputs before any output can be generated.

(e.g. maze navigation, text classification... anything where "history" matters)

Sequence of outputs: The learning system responds with an output to each input, perhaps to reach some "ultimate goal".

(temporal tasks, e.g. missile avoidance...)

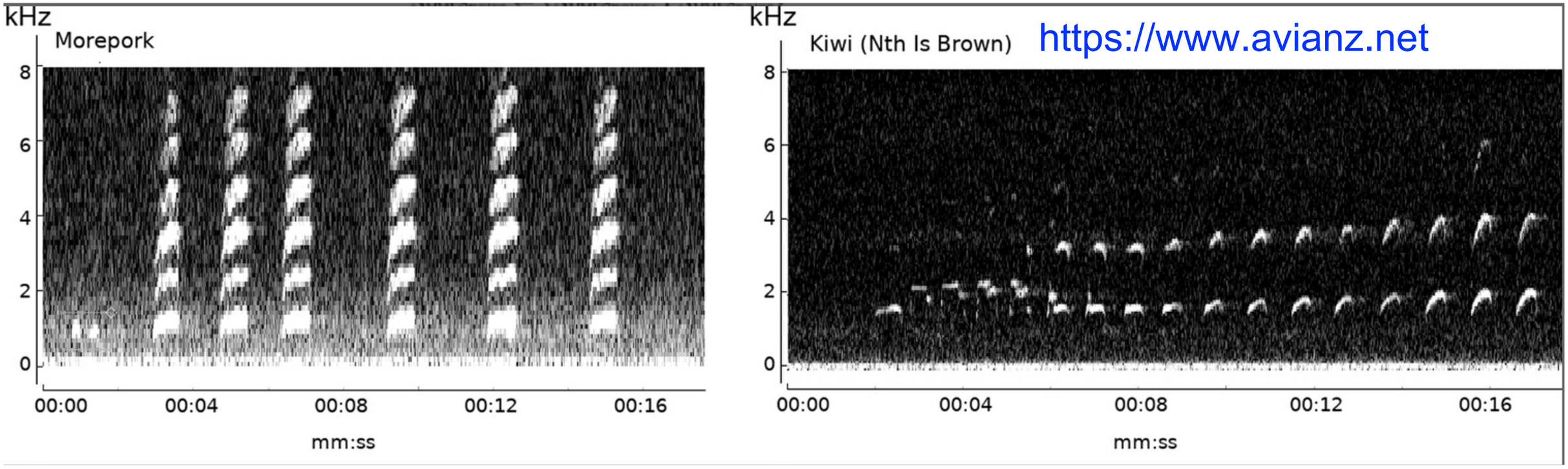
Types of Input

Could be…

i.i.d. (**i**ndependent, **i**dentically **d**istributed)

or a sequence, like this:

V1	V2	V3	V4	Class
3.6216	8.6661	-2.8073	-0.44699	1
4.5459	8.1674	-2.4586	-1.4621	1
3.866	-2.6383	1.9242	0.10645	1
3.4566	9.5228	-4.0112	-3.5944	1
0.32924	-4.4552	4.5718	-0.9888	1
4.3684	9.6718	-3.9606	-3.1625	1
3.5912	3.0129	0.72888	0.56421	1
2.0922	-6.81	8.4636	-0.60216	1
3.2032	5.7588	-0.75345	-0.61251	1
1.5356	9.1772	-2.2718	-0.73535	1
1.2247	8.7779	-2.2135	-0.80647	1



or…?

Classification

Predict a category(class)

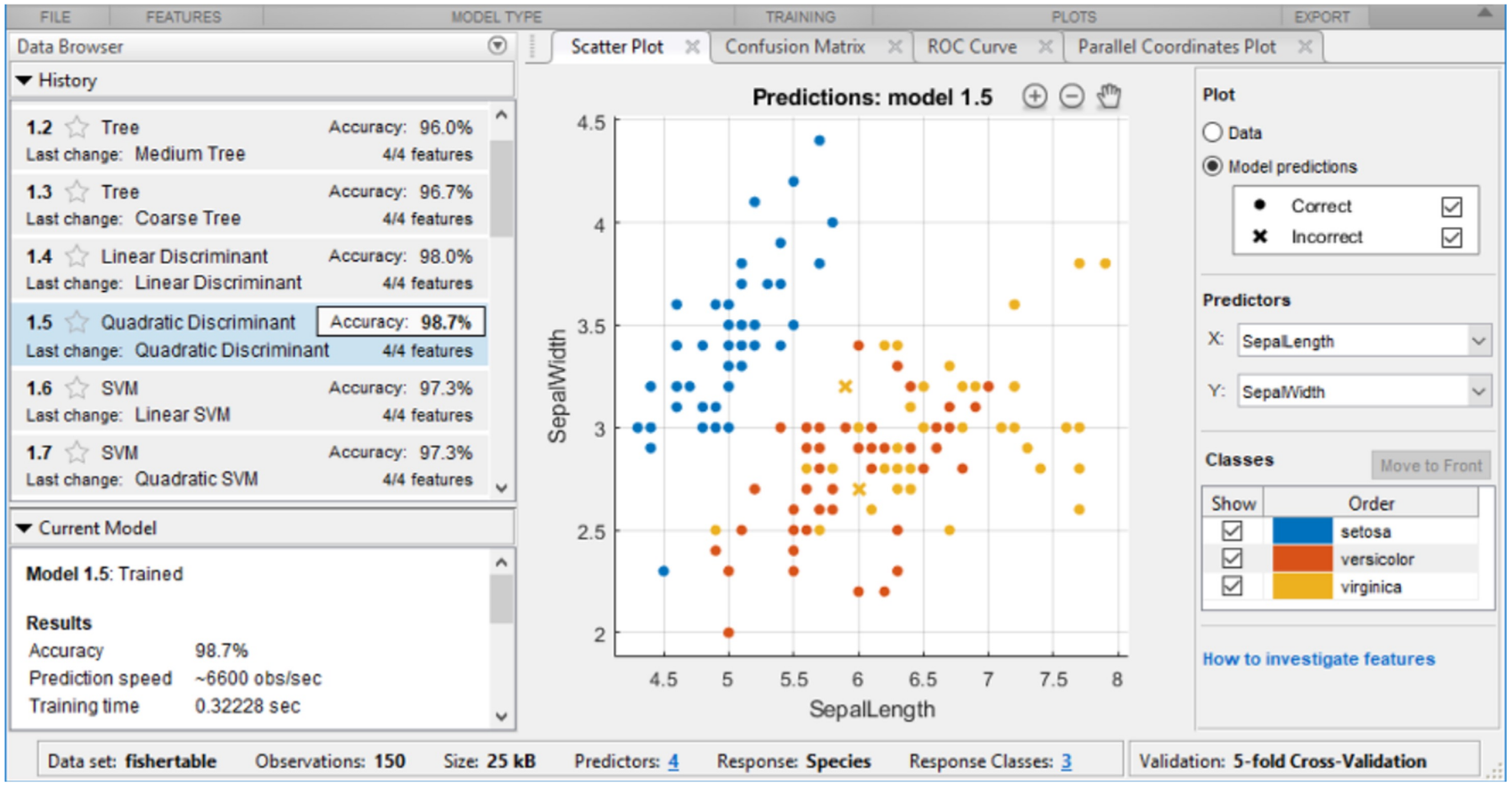
V1	V2	V3	V4	Class
3.6216	8.6661	-2.8073	-0.44699	1
4.5459	8.1674	-2.4586	-1.4621	1
3.866	-2.6383	1.9242	0.10645	1
3.4566	9.5228	-4.0112	-3.5944	1
0.32924	-4.4552	4.5718	-0.9888	1
4.3684	9.6718	-3.9606	-3.1625	1
3.5912	3.0129	0.72888	0.56421	1
2.0922	-6.81	8.4636	-0.60216	1
3.2032	5.7588	-0.75345	-0.61251	1
1.5356	9.1772	-2.2718	-0.73535	1
1.2247	8.7779	-2.2135	-0.80647	1

species of iris: (a 3-way classification)

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
…					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
…					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
…					

Classification

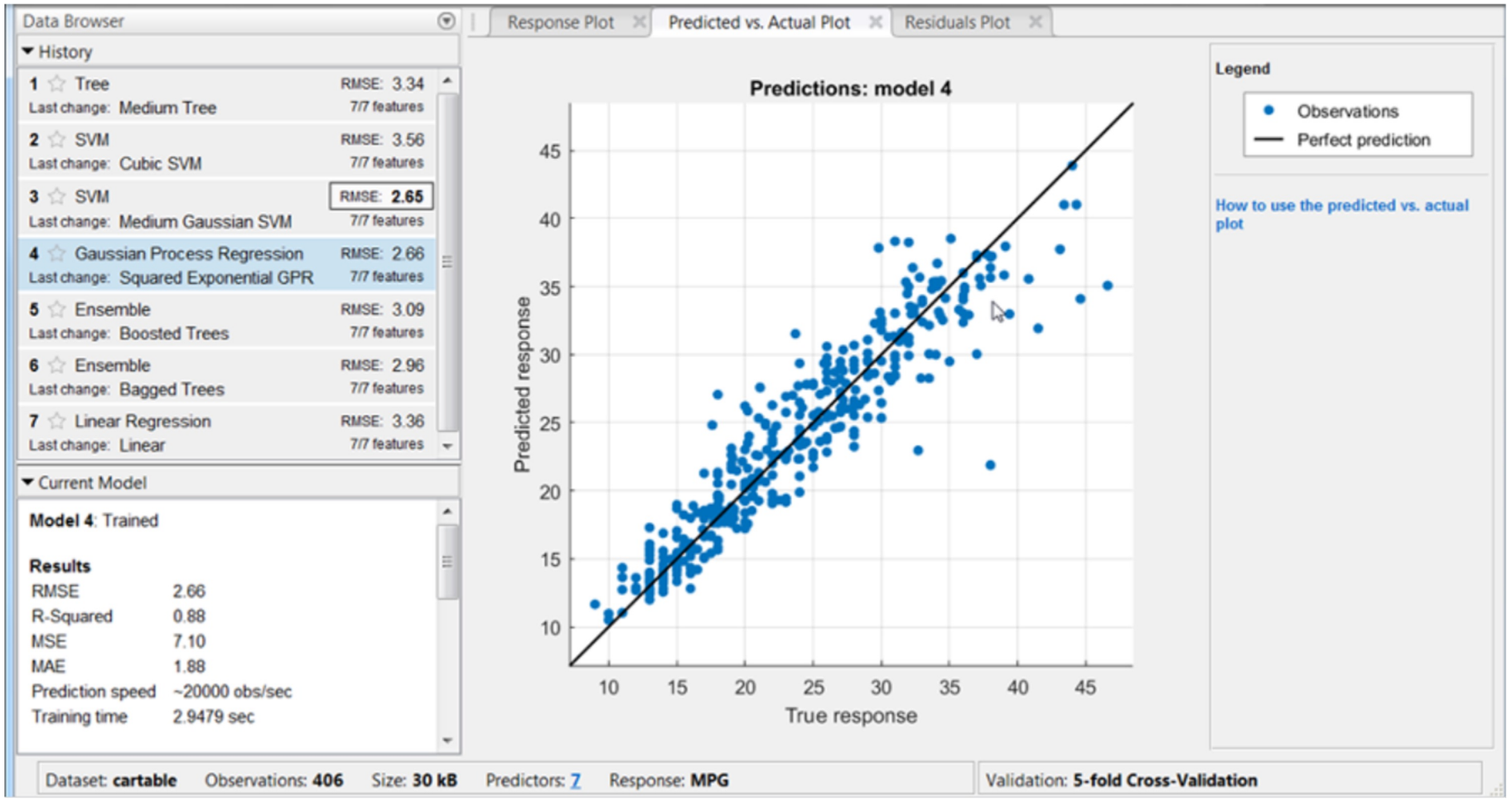
species of iris:
(a 3-way
classification)



https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

Regression

Predict one or more floats



Clustering

Finding groups of items that are "similar"

Clustering is *unsupervised:* class of an example is not known

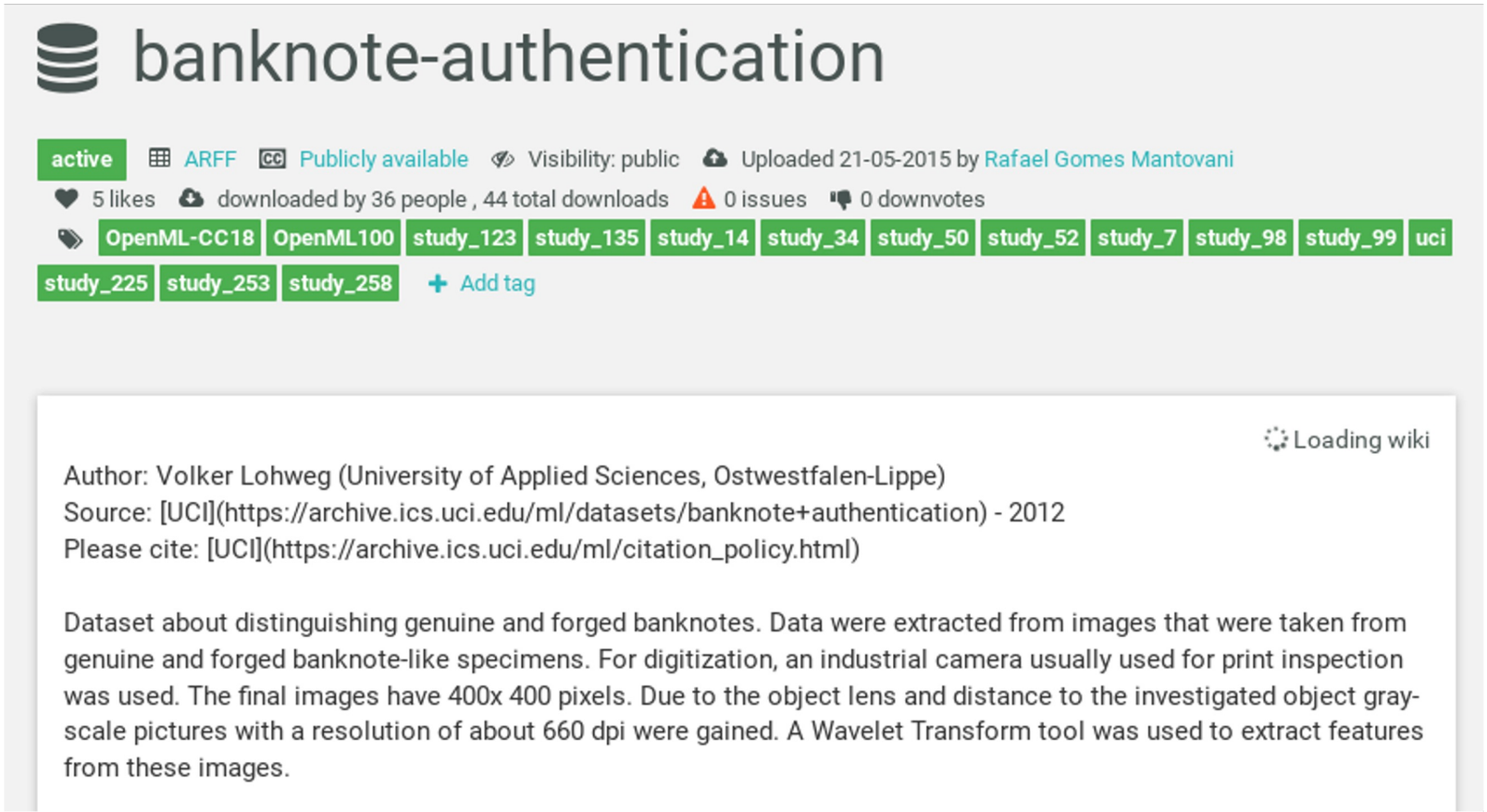
Success often measured subjectively – it is fundamentally ill-posed!

	Sepal length	Sepal width	Petal length	Petal width	Type
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3.0	1.4	0.2	Iris setosa
...					
51	7.0	3.2	4.7	1.4	Iris versicolor
52	6.4	3.2	4.5	1.5	Iris versicolor
...					
101	6.3	3.3	6.0	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

Dimension Reduction

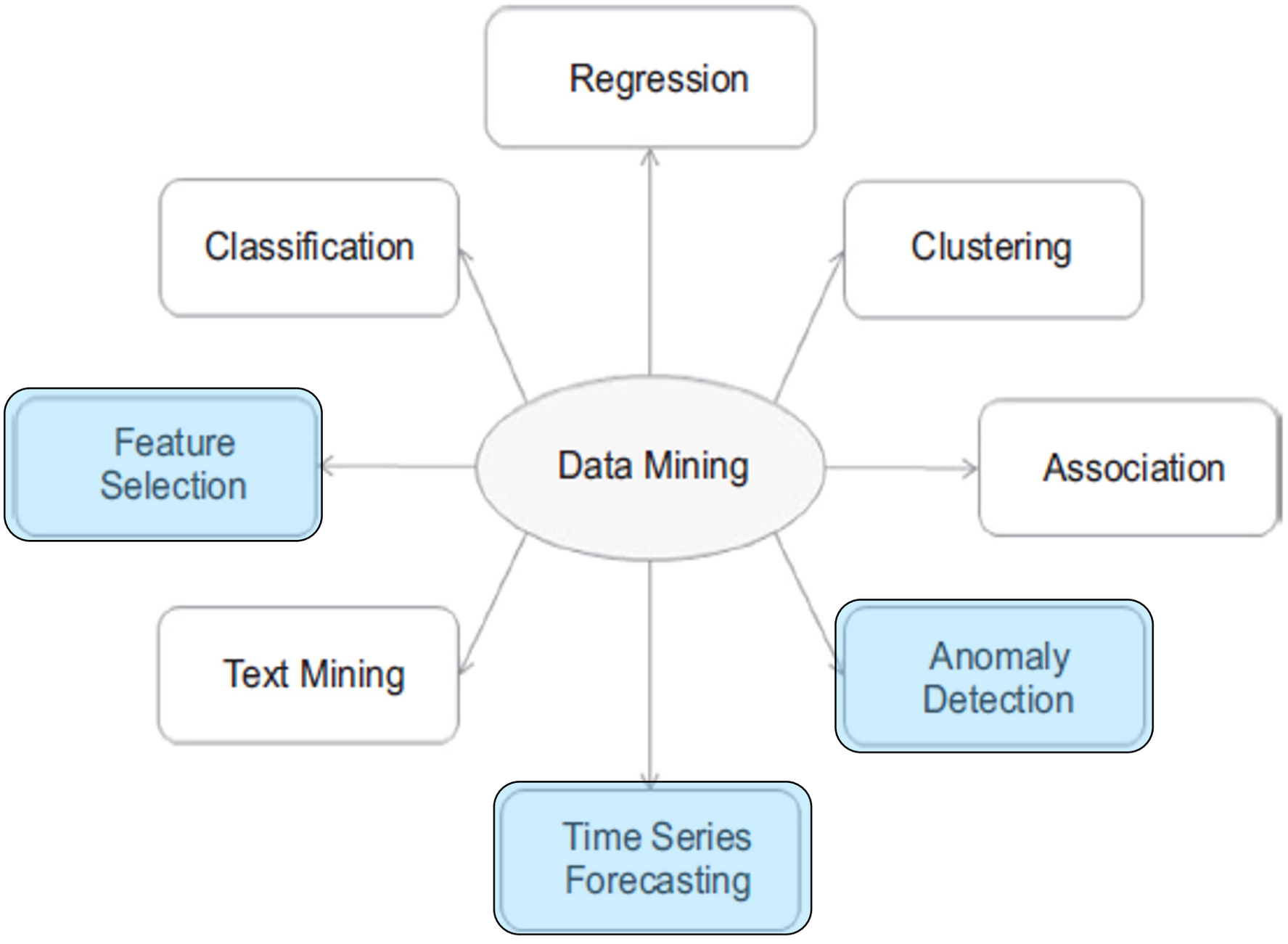
Raw data is often high-dimensional, which is a problem:

1. data with >3 dimensions is hard for humans to interpret / work with!
2. ML algorithms also struggle with high-dimensional data (ultimately, due to the curse of dimensionality)



Hence, interest in ML methods that can identify the main directions of interest in data, for example (e.g. PCA: Principle Components Analysis, and others)

Some Others


Regression
Classification
Clustering
Feature Selection
Data Mining
Association
Text Mining
Anomaly Detection
Time Series Forecasting

Some Others

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known data set.	Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbors	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known data set.	Linear regression, logistic regression	Predicting unemployment rate for next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, density based, local outlier factor (LOF)	Fraud transaction detection in credit cards Network intrusion detection
Time series	Predict the value of the target variable for a future time frame based on historical values.	Exponential smoothing, autoregressive integrated moving average (ARIMA), regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	k-means, density-based clustering (e.g., density-based spatial clustering of applications with noise [DBSCAN])	Finding customer segments in a company based on transaction, web, and customer call data
Association analysis	Identify relationships within an item set based on transaction data.	Frequent Pattern Growth (FP-Growth) algorithm, Apriori algorithm	Find cross-selling opportunities for a retailer based on transaction purchase history

https://tanthiamhuat.wordpress.com/data-mining/