INTRODUCTION TO DATA SCIENCE AND MACHINE LEARNING

TAN PEI SENG

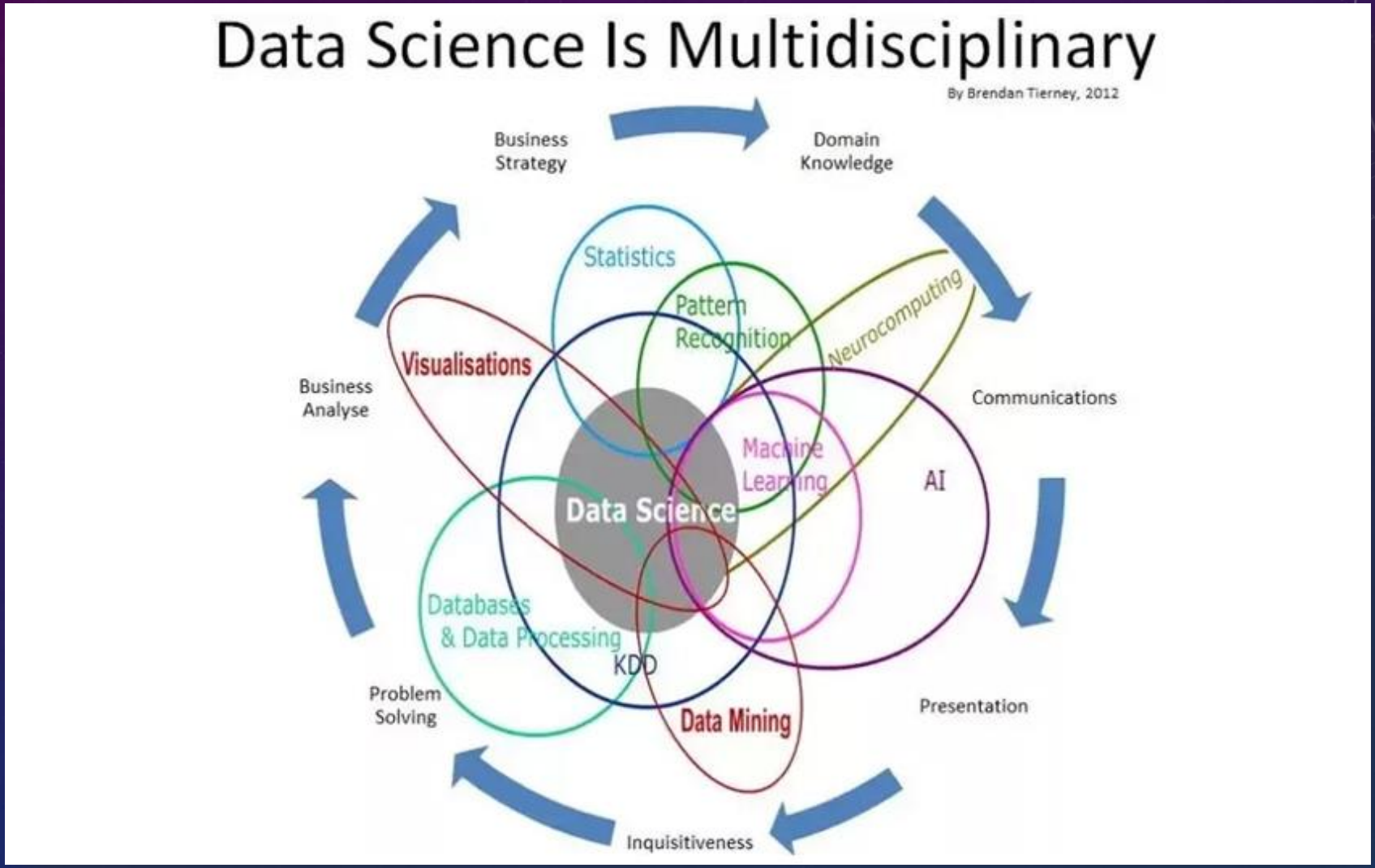
CONTENT COVERED

- Data Science
  - The Definition of Data Science
  - The Role of a Data Scientist in a Project
  - The Jobs Available, Salary and Skills Required for Data Scientist
  - The Flow of a Data Science Project
- Machine Learning (ML)
  - The Definition of ML
  - The Differences between Artificial Intelligence (AI), ML and Deep Learning (DL)
  - The Types of ML and its Applications and Algorithm
  - The Introduction of Instance Space, Label Space and Hypothesis Space of ML model
  - The flow of creating ML model
  - The metric evaluation of ML classification model
  - Generalization, Underfitting and Overfitting


ARE YOU
READY?!

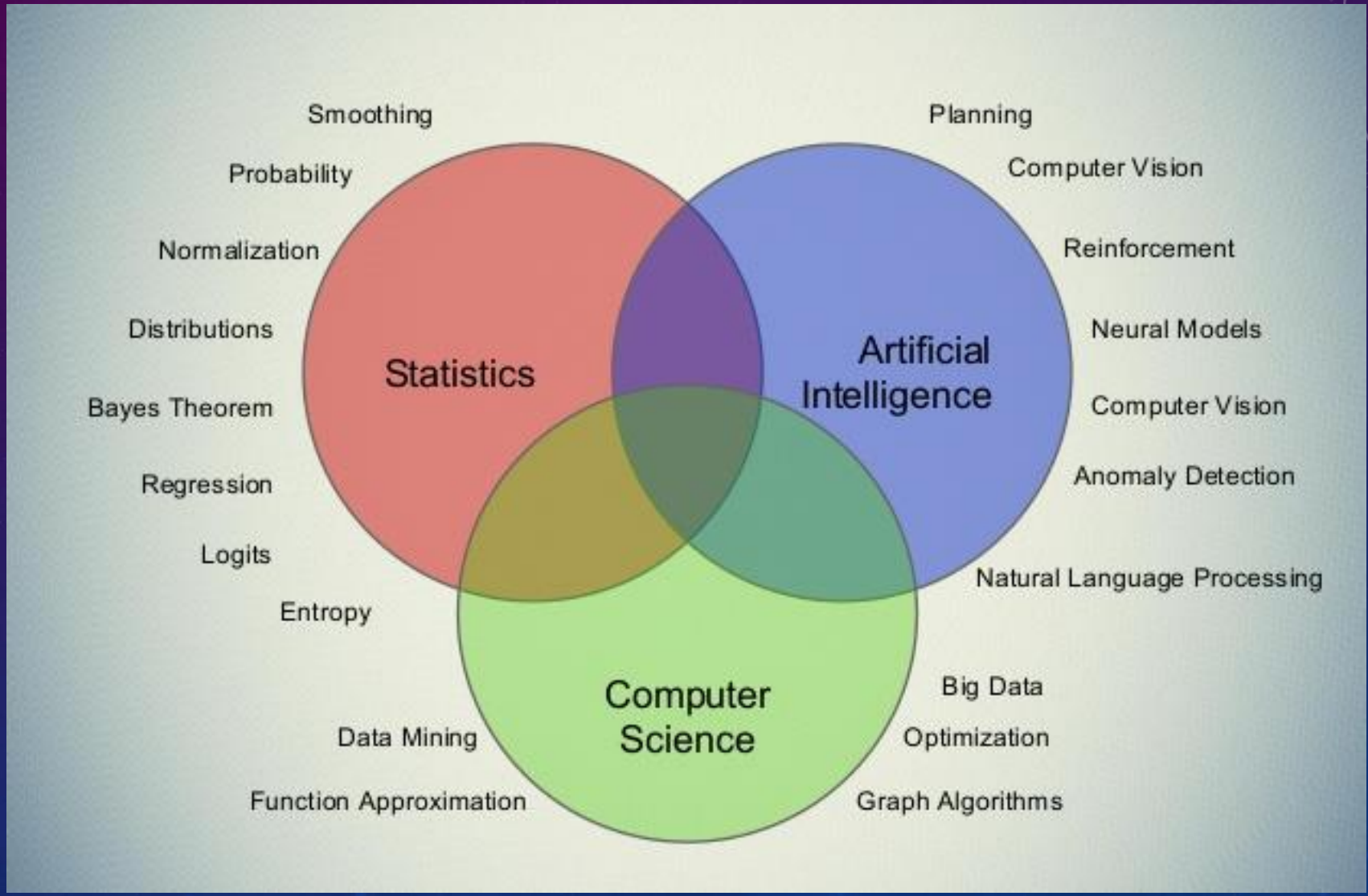

DETECTION
WWW
SOCIAL MEDIA
SERVICES
PROJECTS
BIG DATA
COMPUTING
INFORMATION
TECHNOLOGY
PROMOTION
CONTENT
PROCESSING
MODELS
E-MARKETING
COMMUNICATION
MULTIMEDIA
NETWORK
CONSUMER
ORGANIZATION
PLANNING
CODING
SOFTWARE
BRANDING
CONSUMER
COMPUTER
DEMAND
MARKETS
WEB MARKETING
PREDICTIVE
PROGRAM
ANALYTICS
EVENTS
PROGRAMMING
DATA MINING
DATA SCIENCE
INFORMATION
MACHINE LEARNING
VISION
ENGINEERING
RESEARCH
PROBABILITY
WEB SERVICES
PROJECTS
BIG
DATA
SERVICE
WEB DEV
KDD
COMPUTING
MATHS
PATTERN
STRATEGY
WORLDWIDE
VISUALIZATION
BIG
DATA
ENGINEERING
SOLUTIONS
PLANNING
MEDIA
MOBILE
PRO
PRICING
CODING
SEGMENTATION
SOCIAL
NETWORK
INFORMATION
SOCIAL
NETWORKS
STATISTICS
DIGITAL
TARGET

THE DEFINITION OF DATA SCIENCE



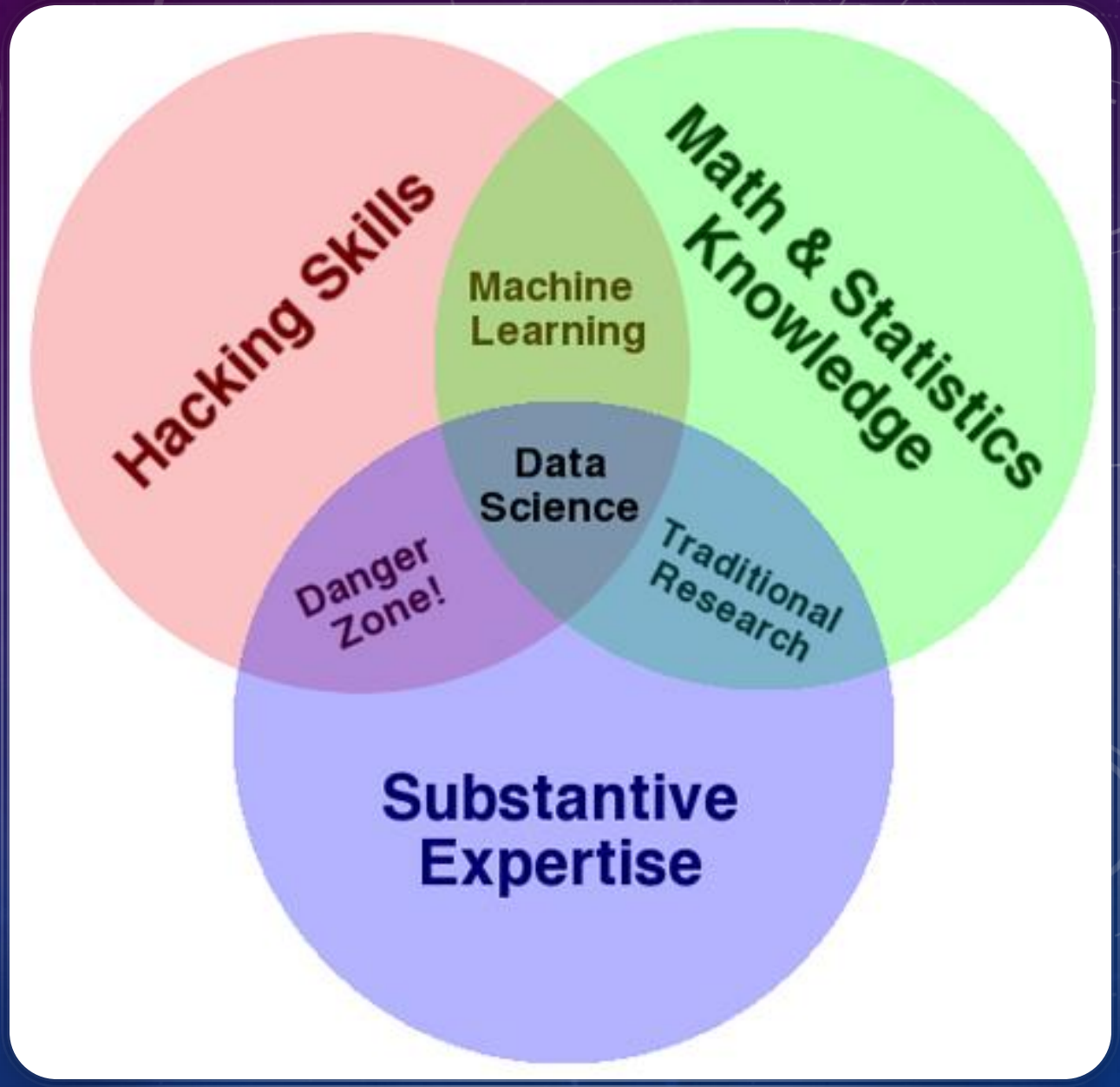
https://www.simplilearn.com/ice9/free_resources_article_thumb/Data-Science-vs.-Data-Analytics-vs.-Machine-Learning1.jpg

THE DEFINITION OF DATA SCIENCE



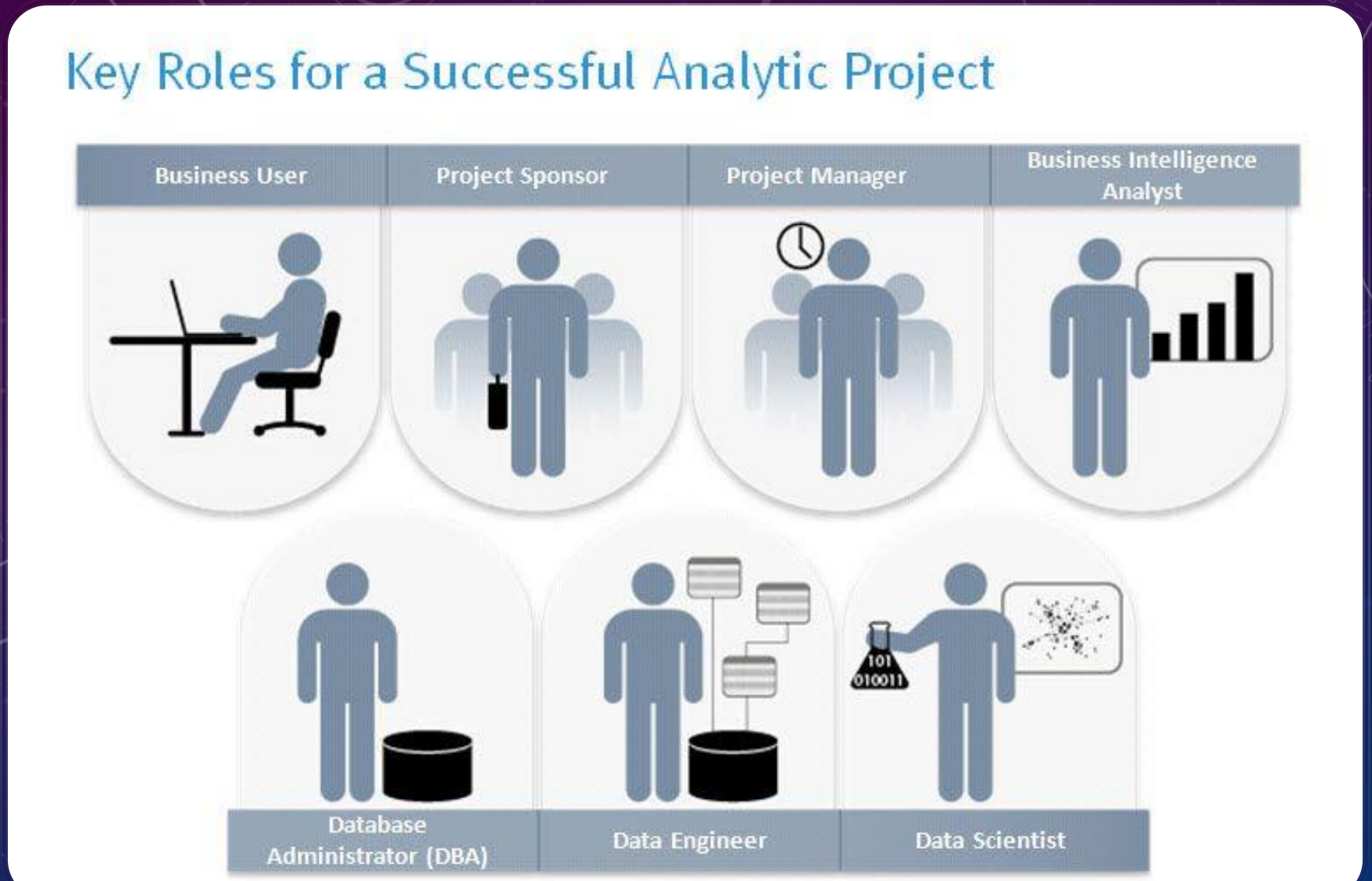
https://www.stoodnt.com/blog/wp-content/uploads/2017/12/20170722123543_introduction-to-machine-learning-300x188.jpg

THE DEFINITION OF DATA SCIENCE



https://www.simplilearn.com/ice9/free_resources_article_thumb/Data-Science-vs.-Data-Analytics-vs.-Machine-Learning.jpg

http://3rd-i.info/iii/cht.jpg



THE ROLE OF A DATA SCIENTIST IN A PROJECT

THE JOBS AVAILABLE, SALARY AND SKILLS REQUIRED FOR DATA SCIENTIST

- https://www.facebook.com/drhanlau/photos/a.2097466600312043/2097466646978705/?type=3&theater

THE JOBS AVAILABLE, SALARY AND SKILLS REQUIRED FOR DATA SCIENTIST

Quora Home Answer Spaces Notifications 3 Search Qu

Debasis Nayak, lives in Malaysia (2017-present)
Updated Oct 15

It depends from which background and what kind of experience you have.

If you are a fresher then it's very hard to find a job in Data scientist role.

If you are experienced like 1–3 year then min salary will be 6000–9000 RM.

Then if you have more experience, then you can understand. (10000+)

It also depends from which country you are from , if you are from Japan/India then you will get handsome salary, also nearly same for local Malaysian but you have to be very talented, because you will get very tough competition from foreigner Indians.

JobStreet.com
No. 1 Job Site in Malaysia

Search Jobs By Title, Skills or Keywor

5 data scientist jobs Save as email job alert

Search Criteria

data scientist

All Locations

All Specializations

Minimum Salary (MYR)

All Jobs Direct Employers

Senior Data Scientist

Macdonald + Company

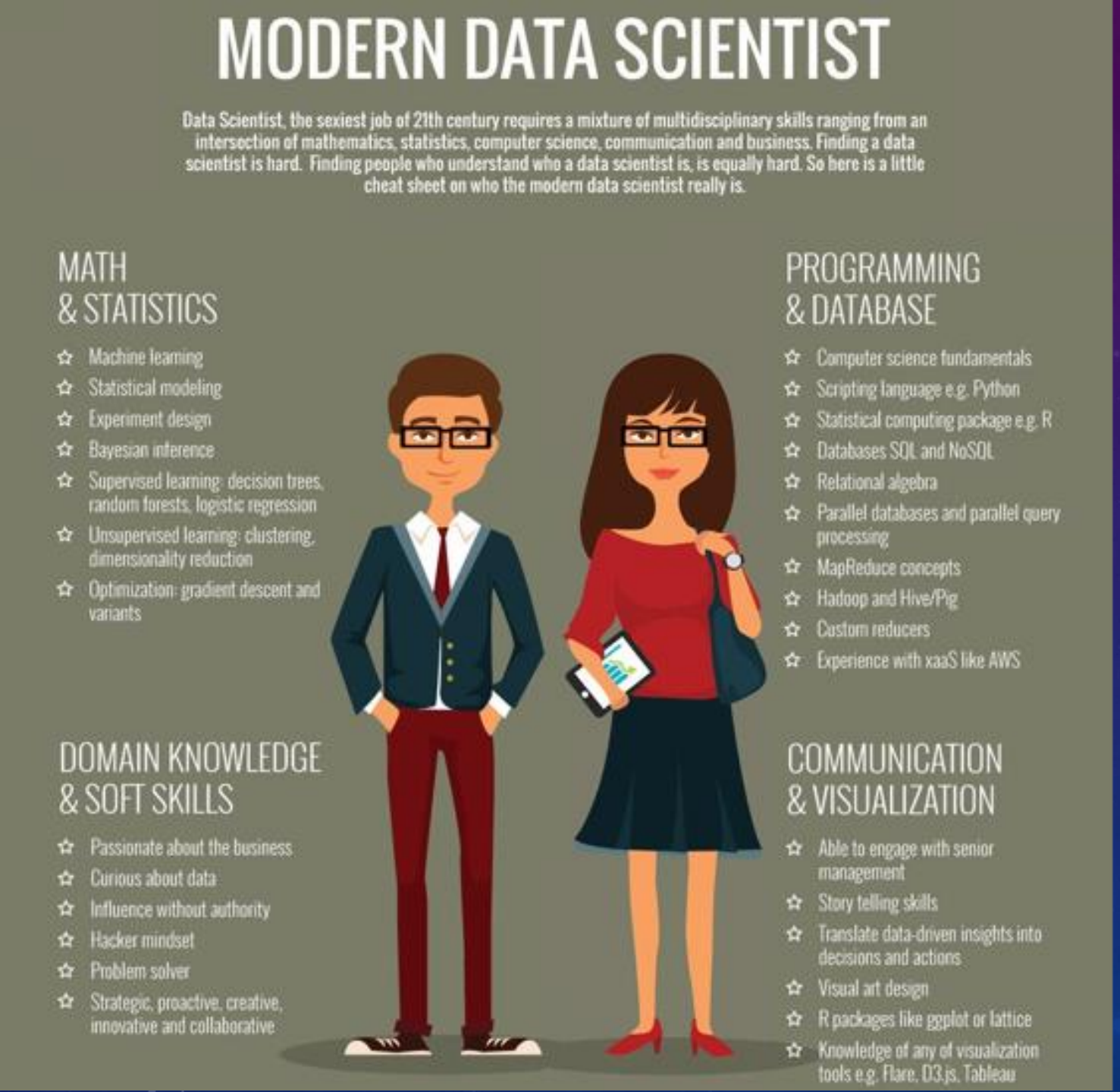
Central (Singapore)

$ MYR 15,000 - 25,000

A global financial institution has newly established an Analytics Centre of Excellence in Kuala Lumpur, Malaysia. They are looking for a senior...

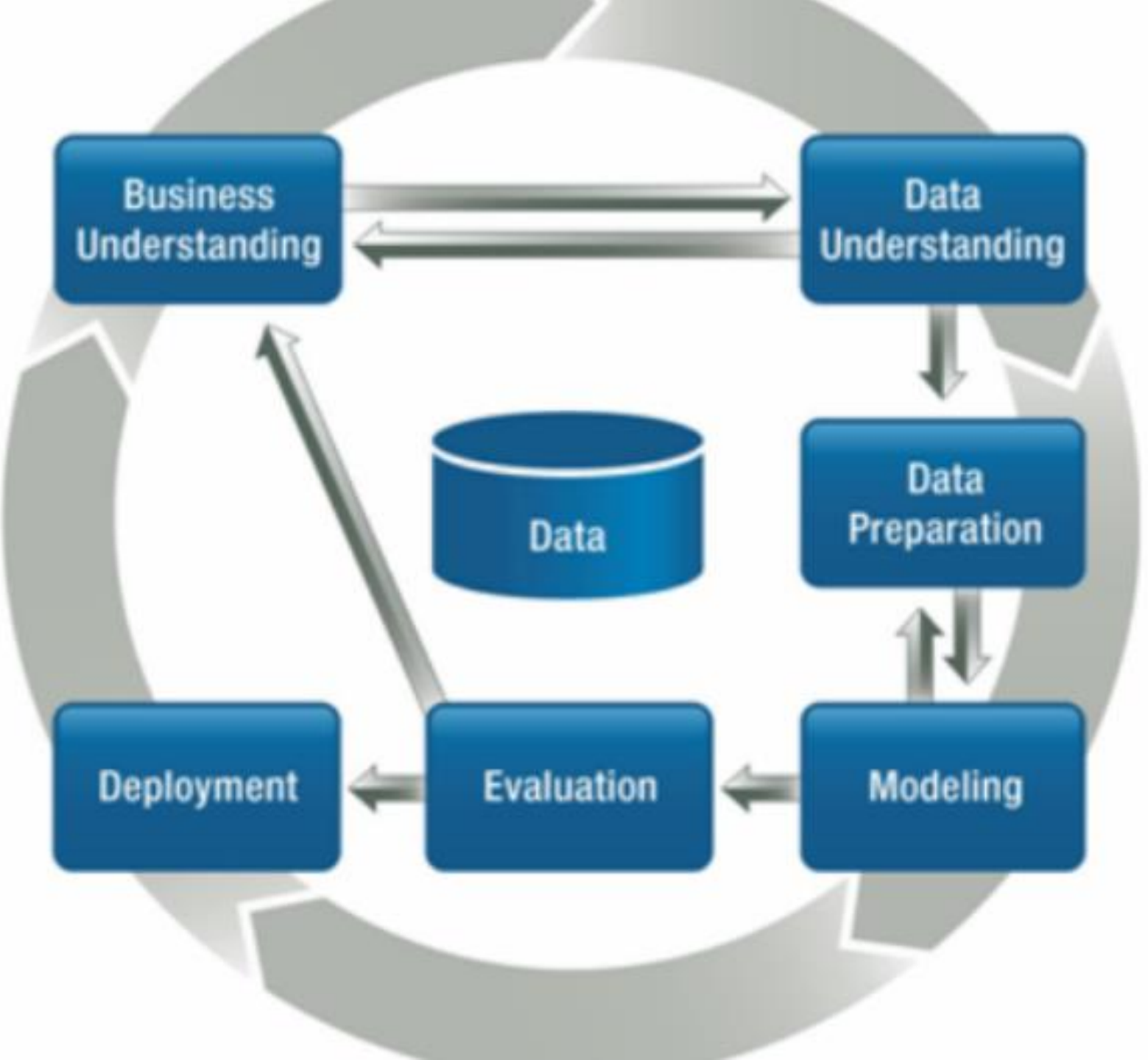
16 hours ago • more

Macdonald & Company



THE JOBS AVAILABLE, SALARY AND SKILLS REQUIRED FOR DATA SCIENTIST

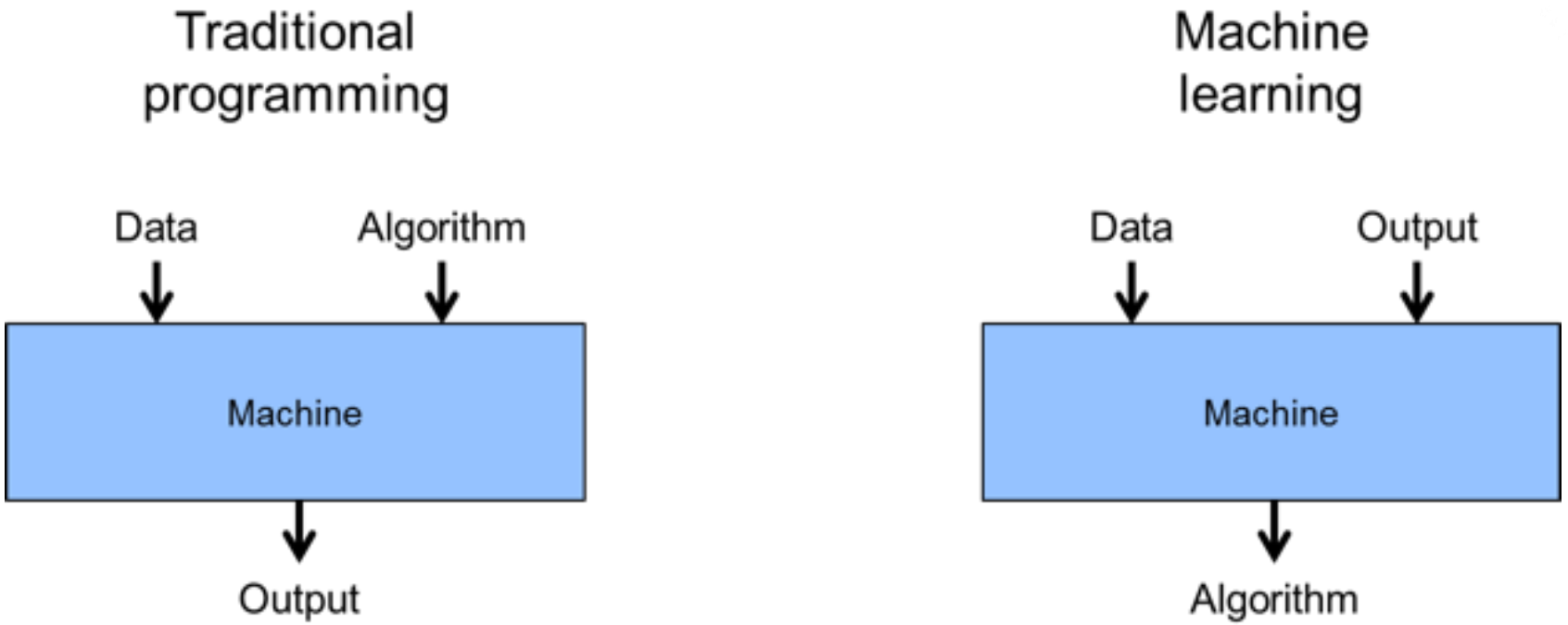
https://pbs.twimg.com/media/DVfEQJHWsAA_vst.jpg



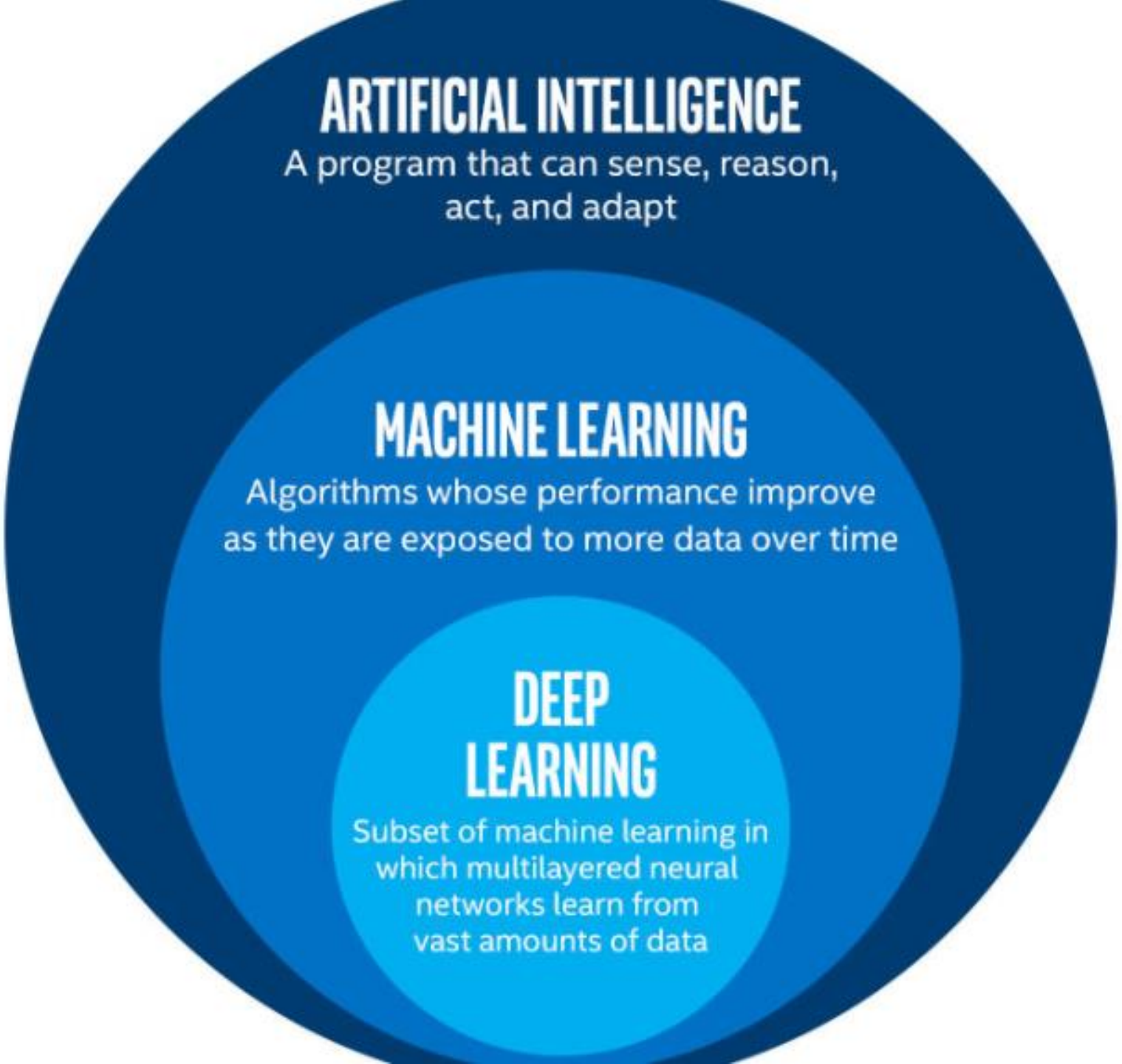
THE FLOW OF A DATA SCIENCE PROJECT

http://www.rosebt.com/uploads/8/1/8/1/8181762/142692.png?1344373798

THE DEFINITION OF MACHINE LEARNING



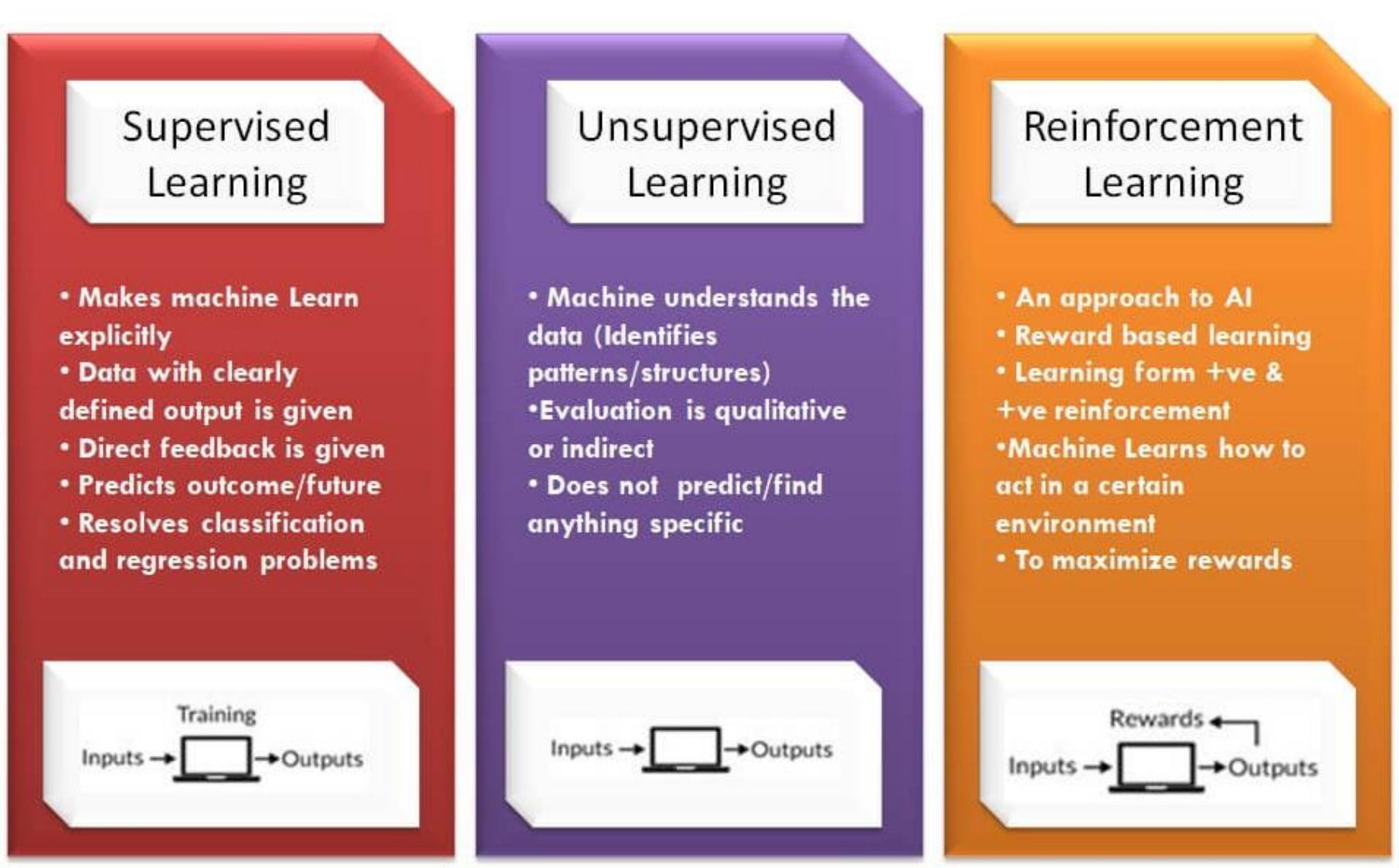
https://cdn.datafloq.com/cms/2017/08/05/machine-learning.png



THE DIFFERENCES BETWEEN ARTIFICIAL INTELLIGENCE (AI), ML AND DEEP LEARNING (DL)

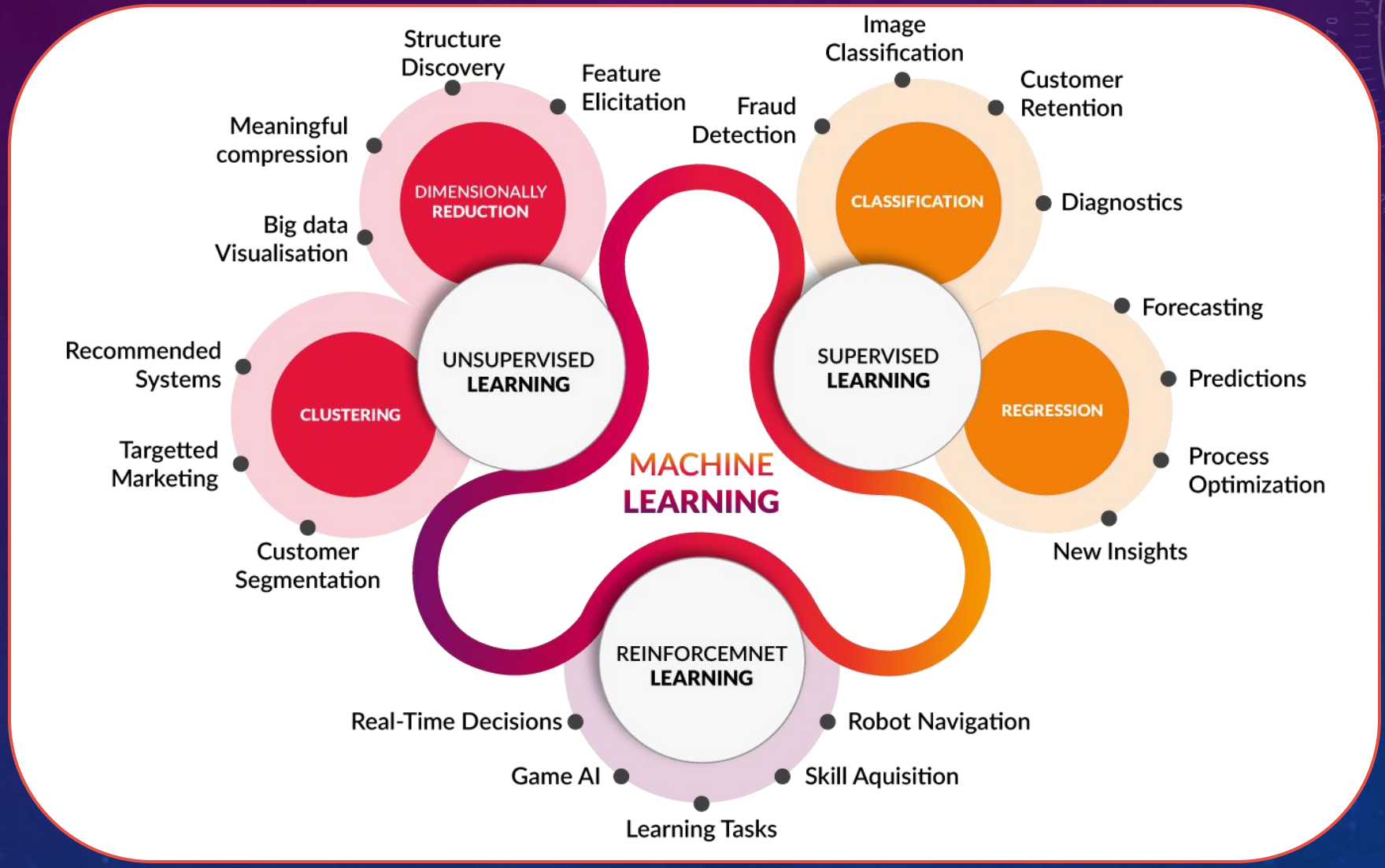
https://cdn-images-1.medium.com/max/1200/1*TiORvHgrJPme_lEiX3olVA.png

THE TYPES OF ML AND ITS APPLICATIONS AND ALGORITHMS



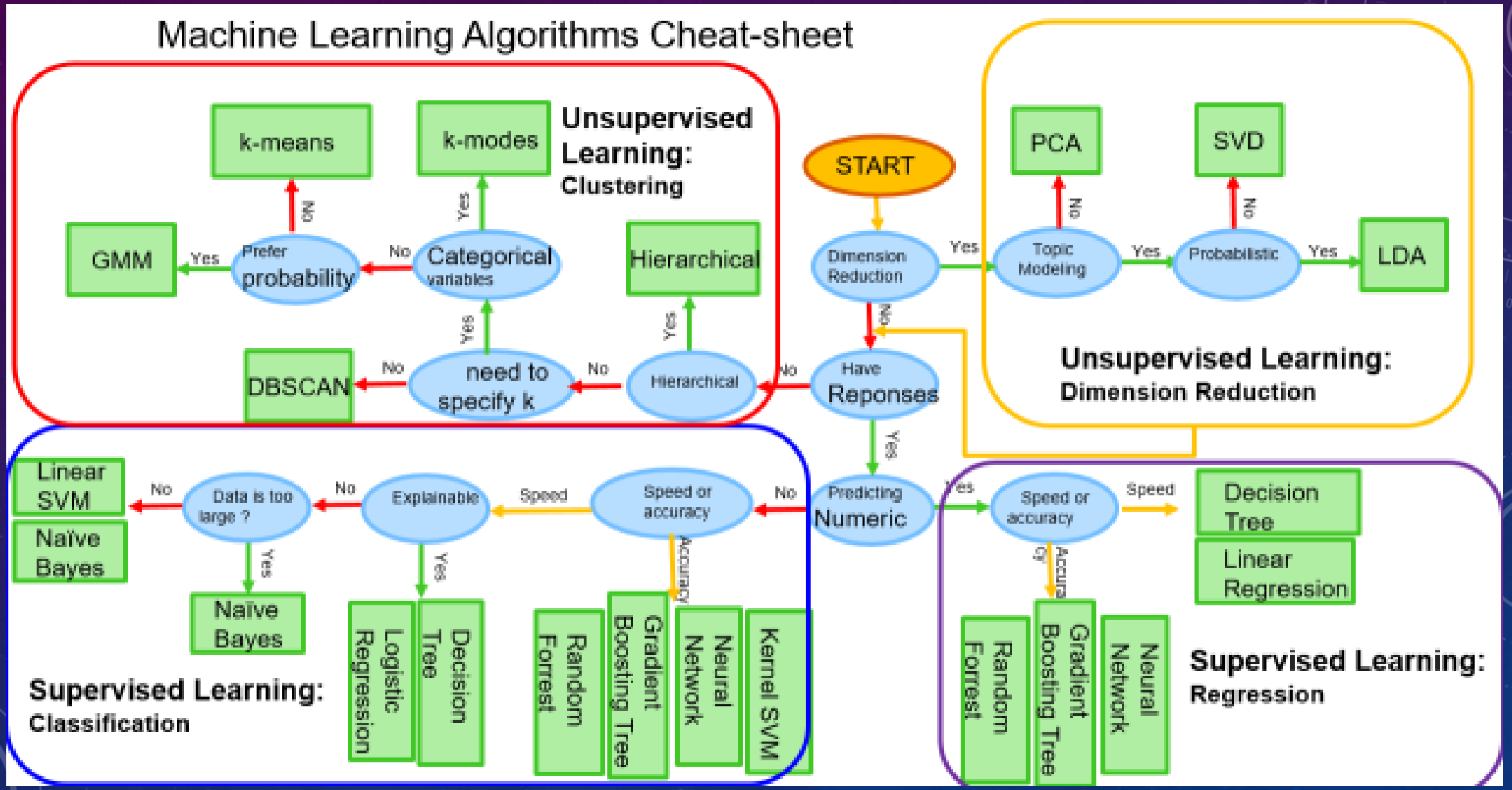
https://www.newtechdojo.com/wp-content/uploads/2018/03/Types-of-Machine-Learning-%E2%80%93-At-a-Glance-optimized.jpg

THE TYPES OF ML AND ITS APPLICATIONS AND ALGORITHMS



https://cdn-images-1.medium.com/max/2000/1*8wU0hfUY3UK_D8Y7tbIyFQ.png

THE TYPES OF ML AND ITS APPLICATIONS AND ALGORITHMS



https://whatsthebigdata.files.wordpress.com/2017/04/machinelearning_cheatsheet.png?w=640

THE INTRODUCTION OF INSTANCE SPACE, LABEL SPACE AND HYPOTHESIS SPACE OF ML MODEL

The Car Searching

from Examples

Name	Label
Ferrari	-
Mazda 8	+
Mazda CX5	-
Buggati Chiron	-
Honda City	-
Toyota Vios	-
Toyota Avanza	+
Toyota Vellfire	+
Honda Odyssey	+
Mini Cooper R53	-
Kia Carnival	+

Searching for a family car

How are the labels generated?

24/09/2018

3

THE INTRODUCTION OF INSTANCE SPACE, LABEL SPACE AND HYPOTHESIS SPACE OF ML MODEL

The Car Searching from Examples

Name	Label
Ferrari	-
Mazda 8	+
Mazda CX5	-
Buggati Chiron	-
Honda City	-
Toyota Vios	-
Toyota Avanza	+
Toyota Vellfire	+
Honda Odyssey	+
Mini Cooper R53	-
Kia Carnival	+

What is the label for "Hyundai Startex"?

What about the label for "Honda Accord"?

THE INTRODUCTION OF INSTANCE SPACE, LABEL SPACE AND HYPOTHESIS SPACE OF ML MODEL

The Car Searching from Exampl

Name	Label
Ferrari	-
Mazda 8	+
Mazda CX5	-
Buggati Chiron	-
Honda City	-
Toyota Vios	-
Toyota Avanza	+
Toyota Vellfire	+
Honda Odyssey	+
Mini Cooper R53	-
Kia Carnival	+

How are the labels generated?

Is it depends on the price and engine power?

x_1: price, x_2 : engine power

THE INTRODUCTION OF INSTANCE SPACE, LABEL SPACE AND HYPOTHESIS SPACE OF ML MODEL

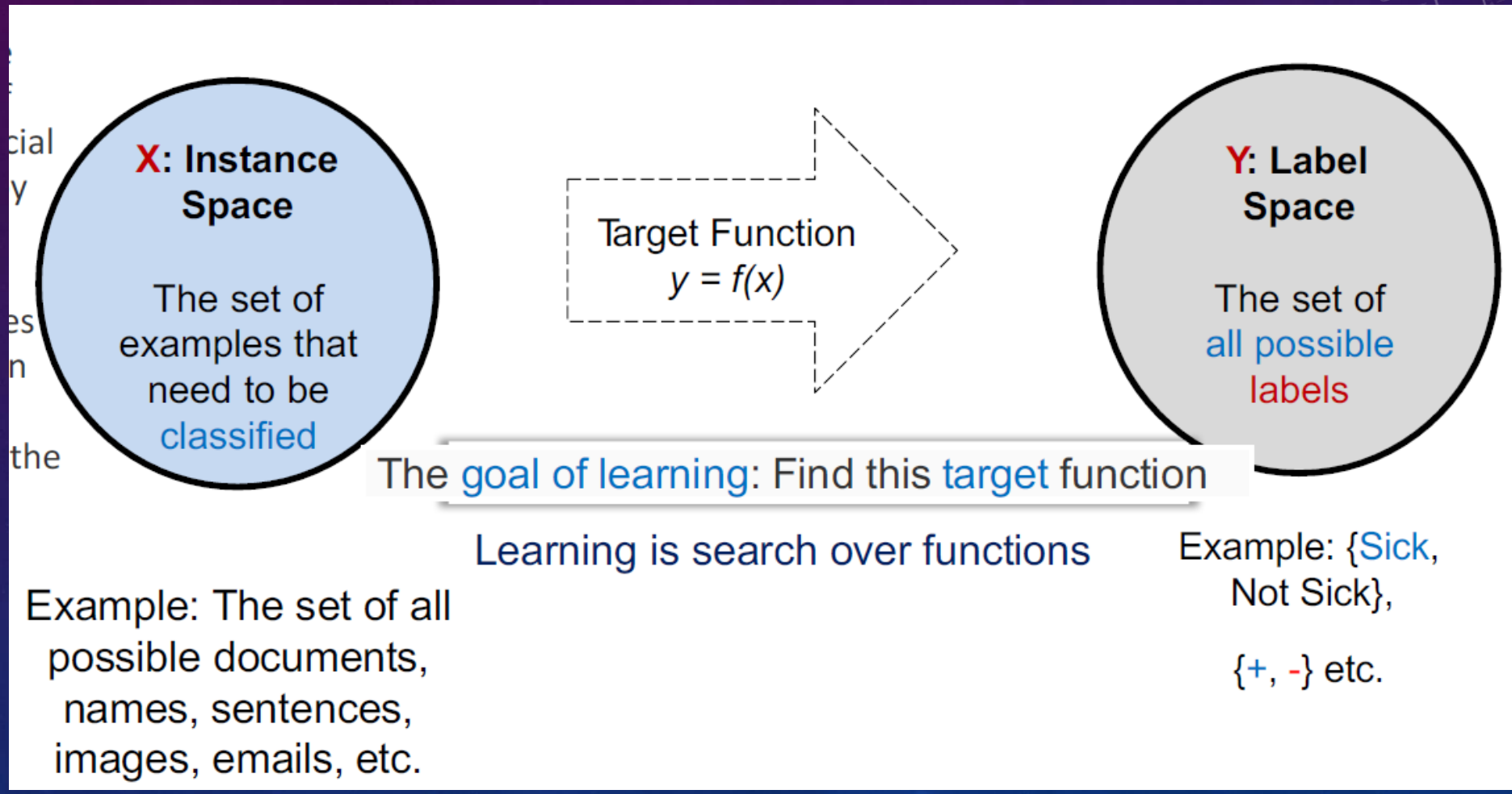
The Car Searching

- *Class C* of a "family car"
  - Prediction: Is car *x* a family car?
  - Knowledge extraction:
    - What do people expect from a family car?
- Output:

  Positive (+) and negative (–) examples
- Input representation:

  x_1: price, x_2 : engine power

THE INTRODUCTION OF INSTANCE SPACE, LABEL SPACE AND HYPOTHESIS SPACE OF ML MODEL



INSTANCE SPACE – INPUTS

Instances $x \in X$ are defined by features/attributes

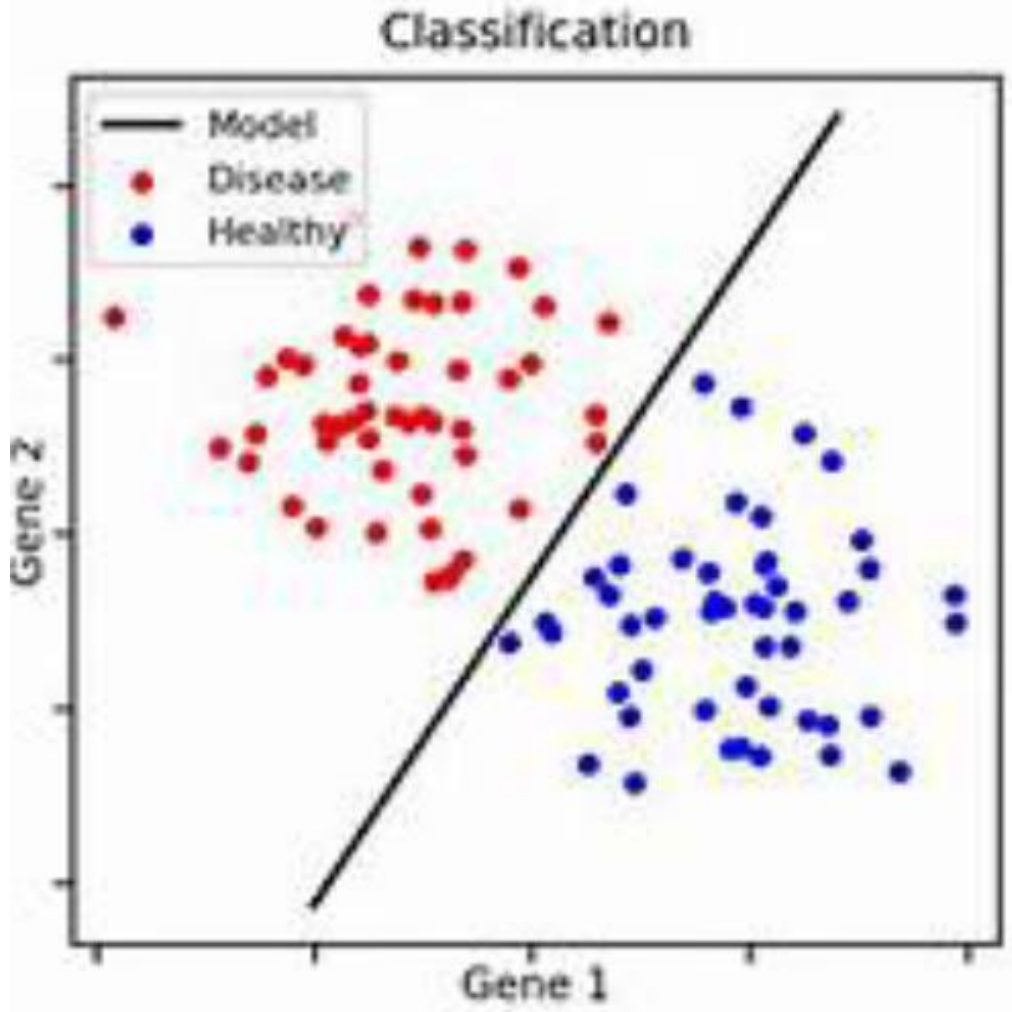
The choice of features is crucial to how well a task can be learned.

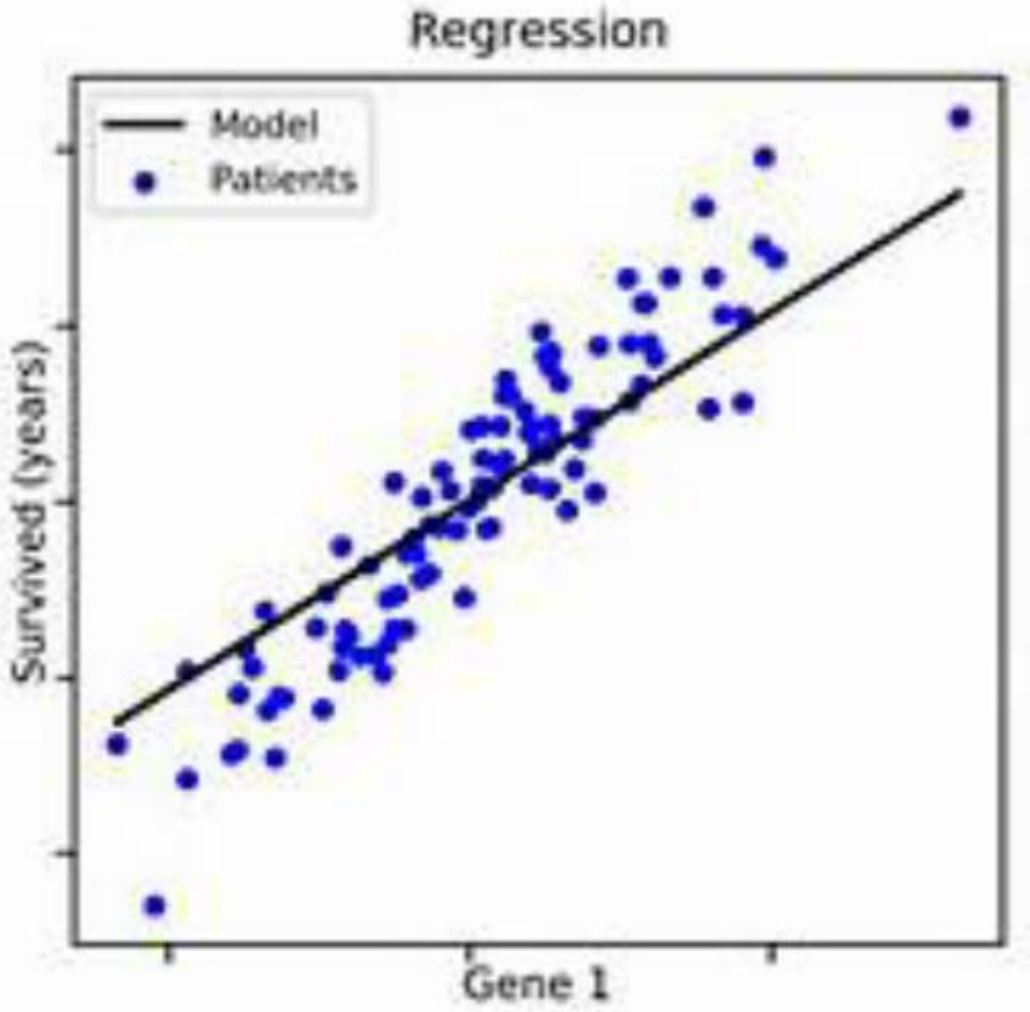
What are other possible features in the Car Searching?

- *Number of seats?*
- *The look?*
- *The comfortable level?*
- *??*

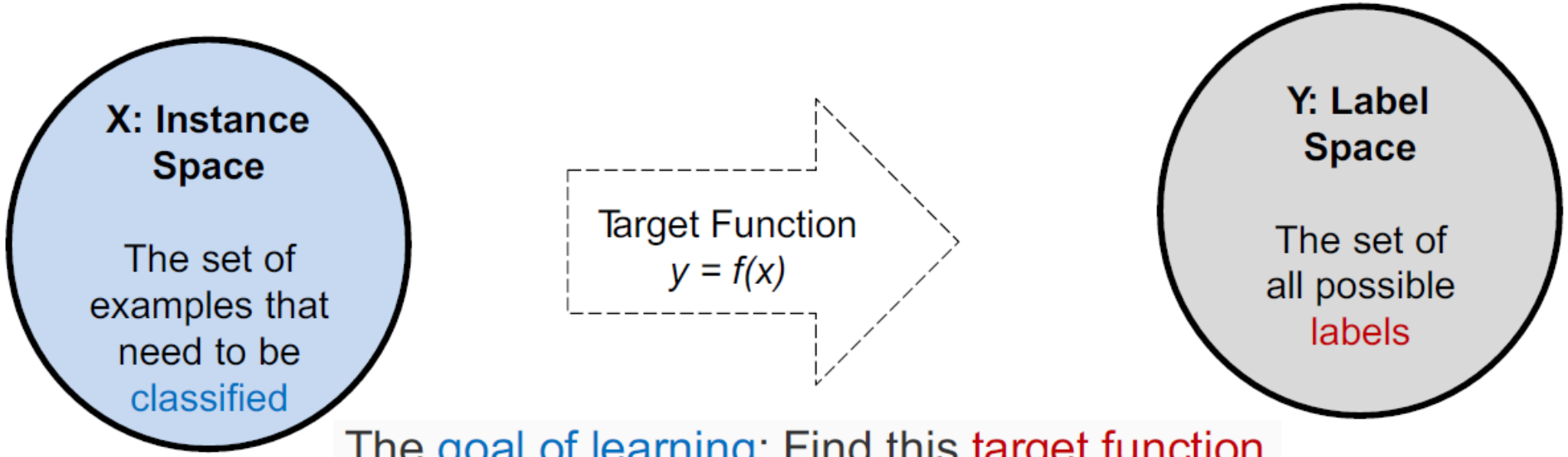
LABEL SPACE – OUTPUTS

- Determines what kind of supervised learning task we are dealing with
- **Classification**: Output is categorical
  - **Binary** classification: Two possible labels, $y \in \{-1, 1\}$ (e.g., [Yes, No], [Sick, Not Sick])
  - **Multi-class** classification: More than two possible labels [Like, Dislike, Neutral]
- **Regression**: Output is numerical (real numbers)


Classification
Model
Disease
Healthy
Gene 2
Gene 1


Regression
Model
Patients
Survived (years)
Gene 1

HYPOTHESIS SPACE – ALGORITHM

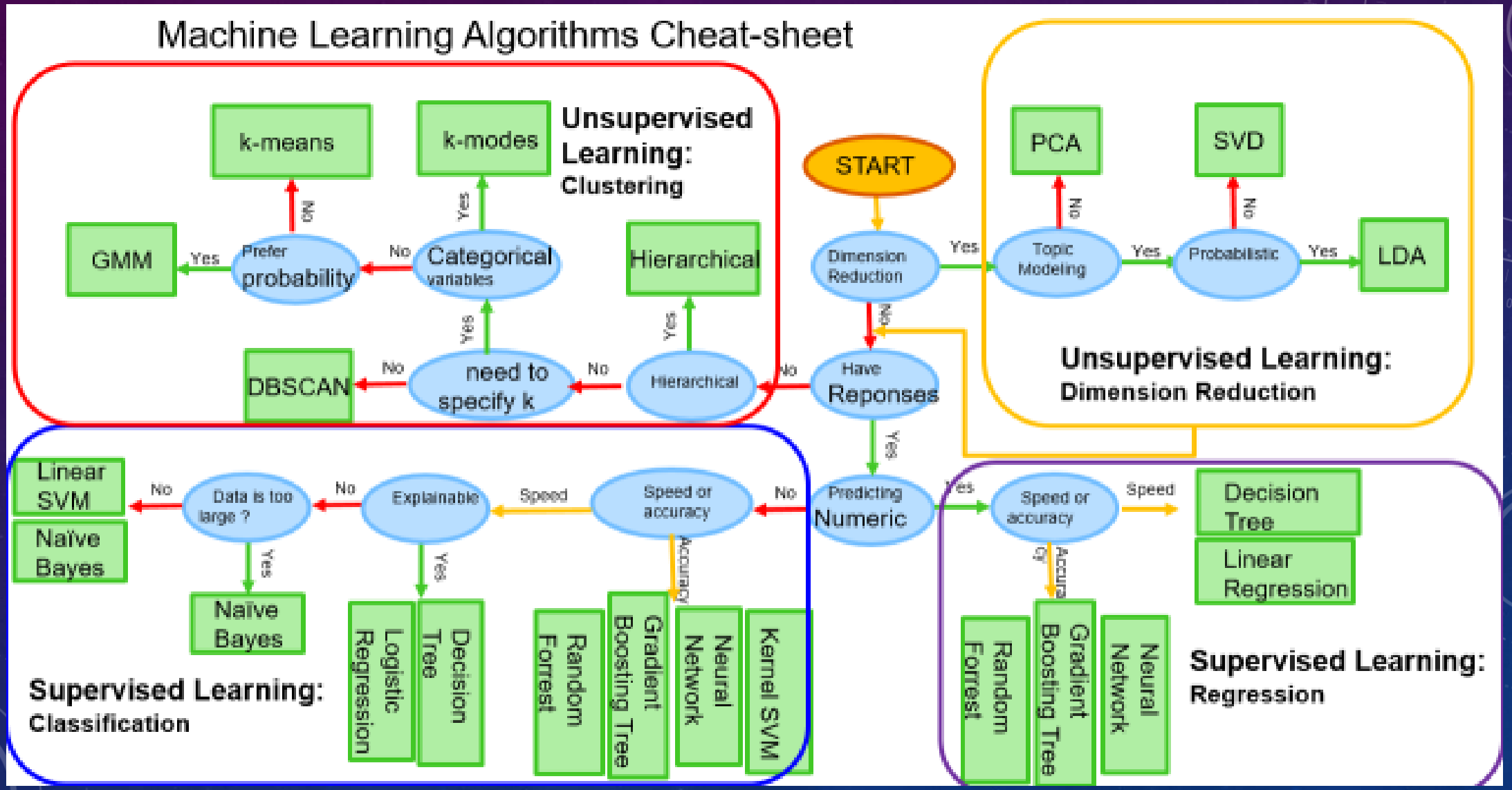


The goal of learning: Find this target function

Learning is search over functions

The **hypothesis space** is the set of functions we consider for this search

HYPOTHESIS SPACE – ALGORITHM

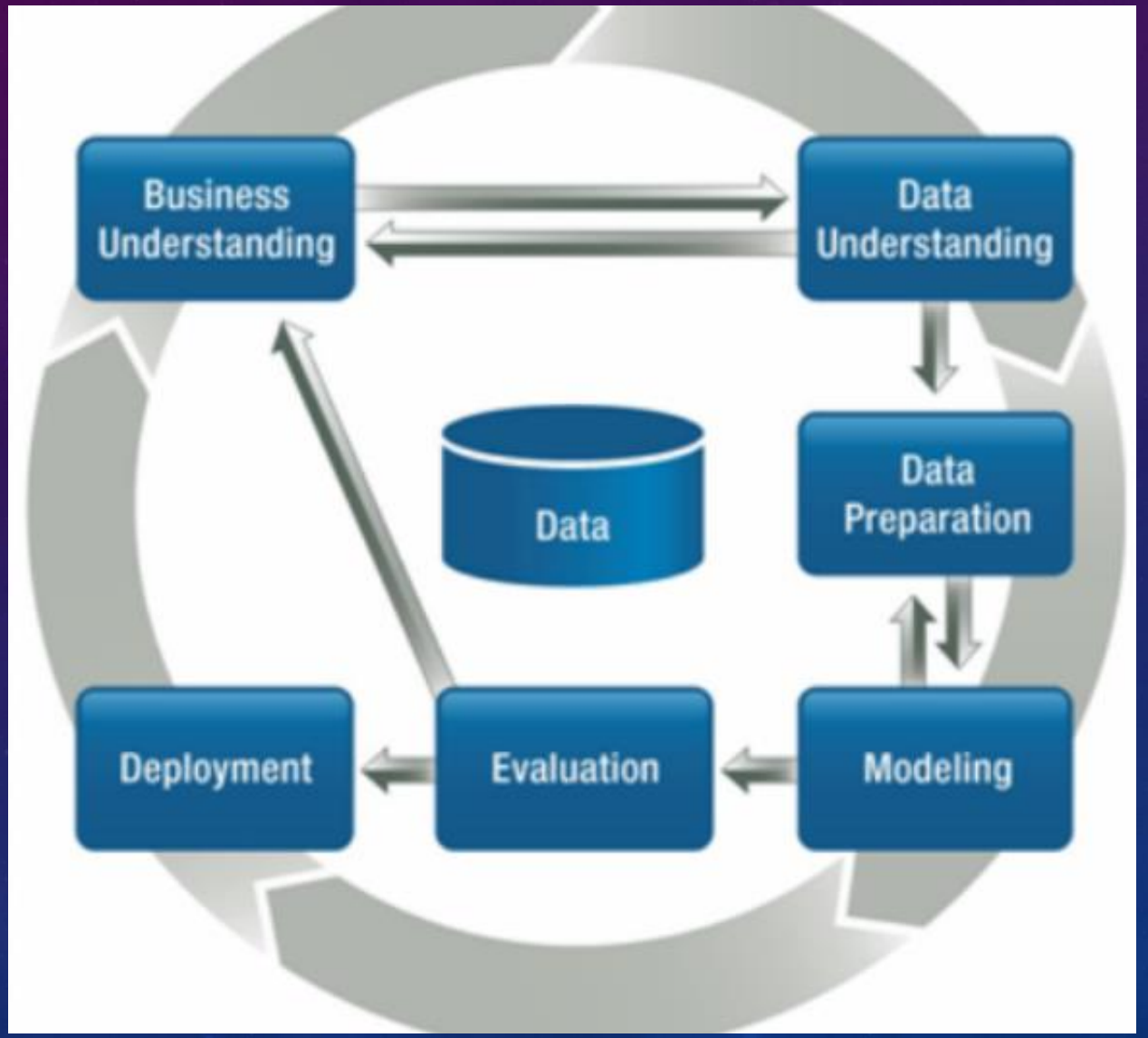


https://whatsthebigdata.files.wordpress.com/2017/04/machinelearning_cheatsheet.png?w=640

INSTANCES, LABELS AND ALGORITHMS

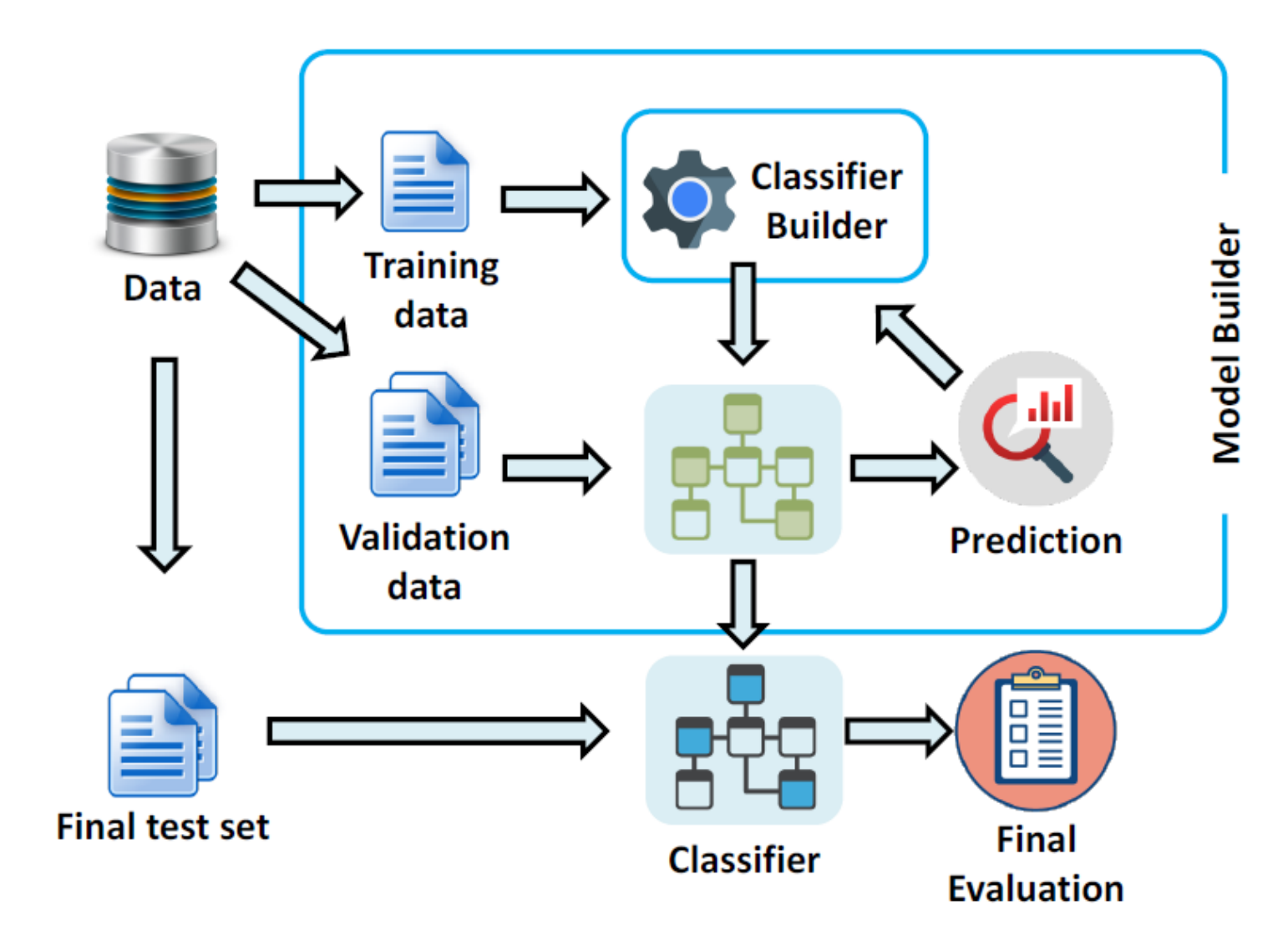
- What is our **instance space**?
  - What are the inputs to the problem? What are the features?
- What is our **label space**?
  - What is the predictive task?
- What is our **hypothesis space**?
  - What functions should the learning algorithm search over?

THE FLOW OF CREATING ML MODEL



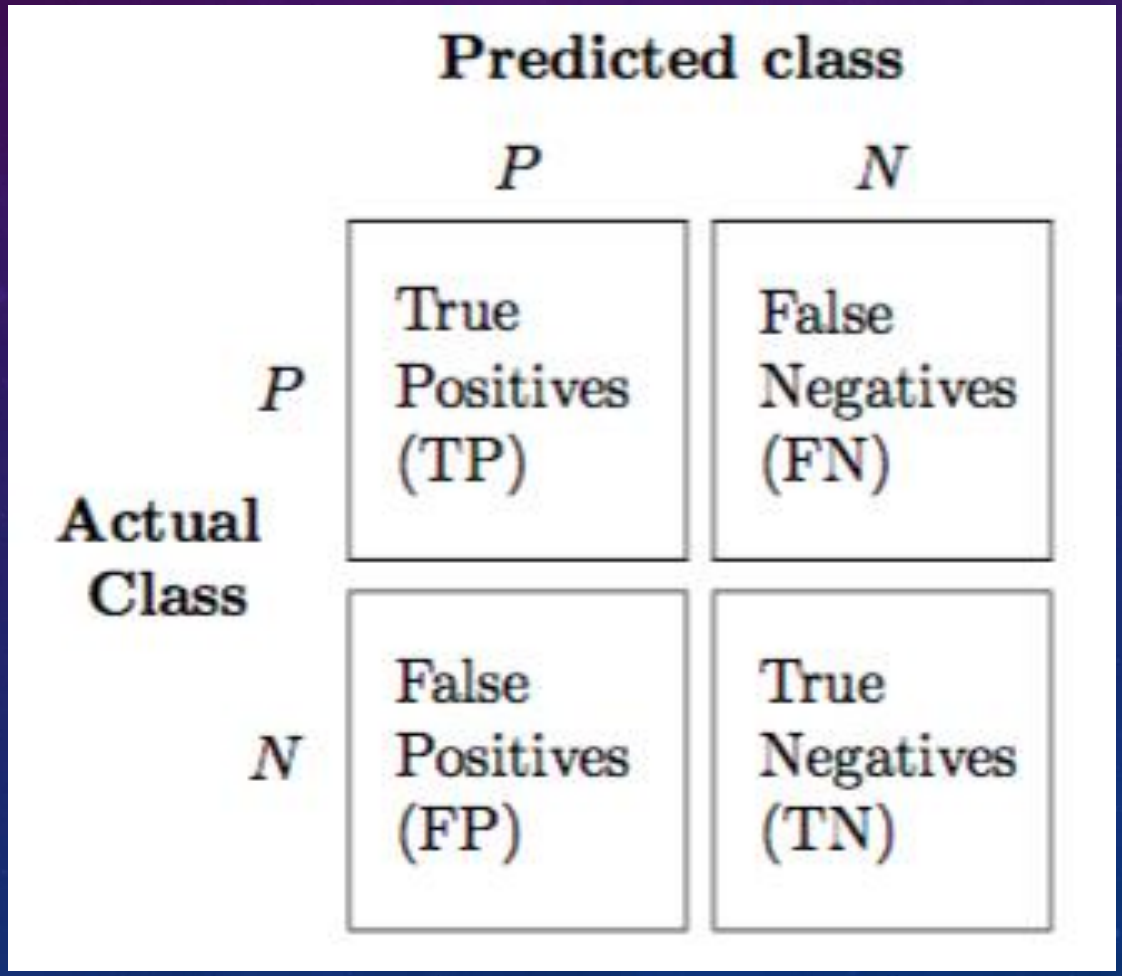
http://www.rosebt.com/uploads/8/1/8/1/8181762/142692.png?1344373798

THE FLOW OF CREATING ML MODEL



THE METRIC EVALUATION OF CLASSIFICATION MODEL

1. Confusion Matrix



http://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix_files/confusion_matrix_1.png

THE METRIC EVALUATION OF CLASSIFICATION MODEL

2. Precision, Recall and F1 Score

Predictive Model: Evaluation

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

predictive result / classification		actual result / classification: yes	actual result / classification: no
	yes	**tp** (true positive)	**fp** (false positive) — Type 1 error
	no	**fn** (false negative)	**tn** (true negative)

$$\text{Precision} = \frac{tp}{tp + fp}$$

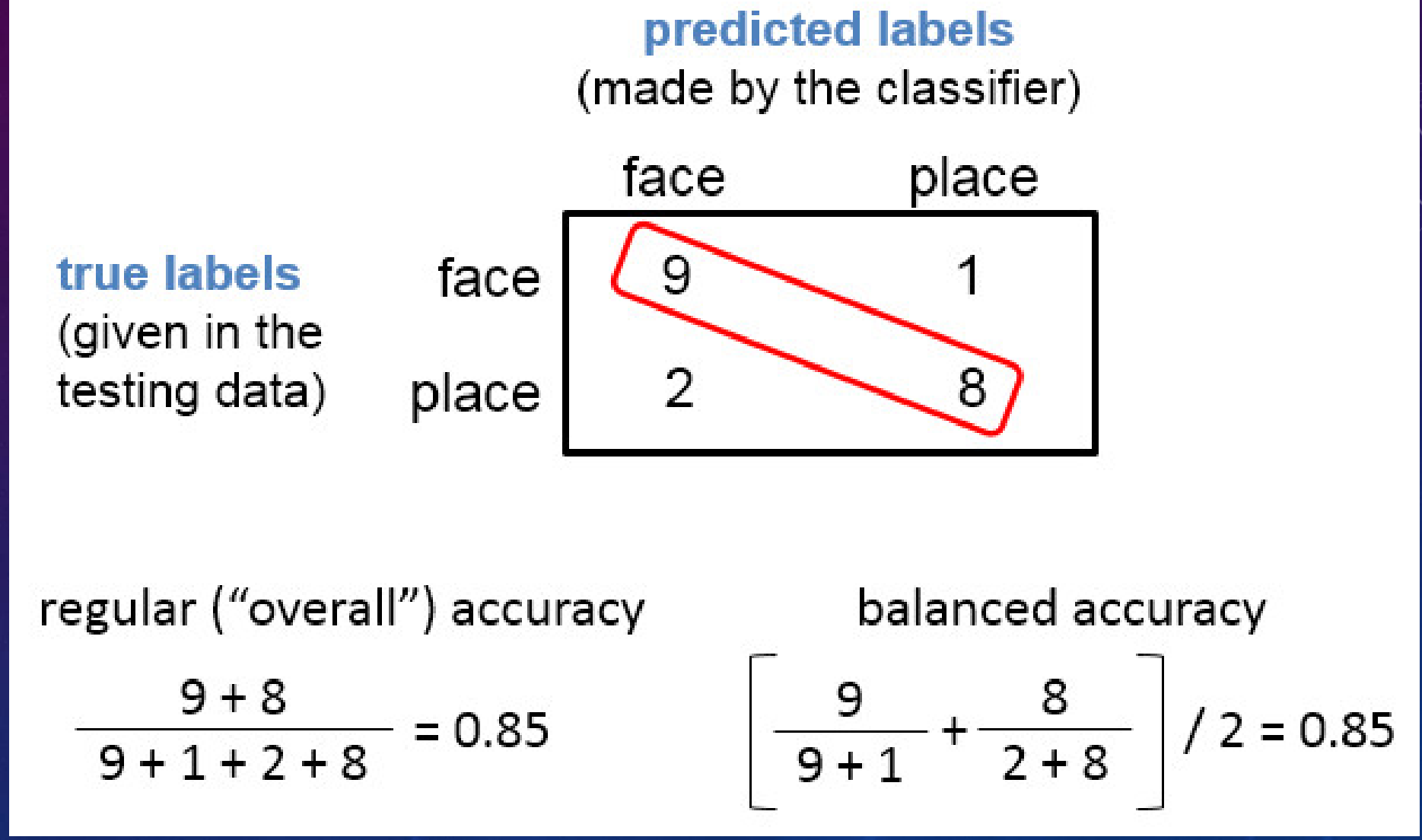
$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\text{True Negative Rate} = \frac{tn}{tn + fp}$$

https://i.stack.imgur.com/OW5Lt.jpg

EXAMPLE



regular ("overall") accuracy

$$\frac{9+8}{9+1+2+8} = 0.85$$

balanced accuracy

$$\left[\frac{9}{9+1} + \frac{8}{2+8}\right] / 2 = 0.85$$

http://1.bp.blogspot.com/-SRXhc6xrnM4/VnQ4-uCMzbI/AAAAAAAAA4U/Cb7kpPchhik/s1600/cmatrix.jpg

EXAMPLE

Classifier	Precision	Recall	F1 Score	Accuracy
GaussianNB	0.35556	0.80000	0.49321	0.76429
DecisionTree	0.60000	0.60000	0.60000	0.88571
SVC (kernel='linear')	0.71429	0.25000	0.37037	0.87857
KMeans (n_clusters=2)	0.12500	0.25000	0.16667	0.64286

https://cdn-images-1.medium.com/max/1600/1*d1vefJgFI5_hBtdhs6wLCA.png

EXAMPLE

Methods	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbours	0.952	0.074	0.268	0.116
Linear SVM	0.968	0.721	0.385	0.502
Decision Tree	0.951	0.250	0.385	0.303
Random Forests	0.958	0.320	0.315	0.317
Adaboost	0.960	0.230	0.567	0.327
Naive Bayesian	0.801	0.527	0.183	0.111
Variance of Laplacian	0.958	0.113	0.161	0.133
NIQE [9]	0.958	0.210	0.248	0.227
CNN-no augmentation [14]	0.968	0.700	0.466	0.560
CNN-translational augmentation	0.974	0.750	0.600	0.667
CNN-k-space augmentation	0.977	0.779	0.642	0.704
CNN with k-space+translational augmentation	**0.982**	**0.809**	**0.652**	**0.722**

https://www.researchgate.net/profile/Ilkay_Oksuz/publication/327050195/figure/tbl1/AS:667871336226818@1536244343080/Mean-accuracy-precision-recall-and-F1-score-results-of-image-classification-for-motion.png

GENERALIZATION

- How well a model trained on the training set predicts the right output for new instances.

UNDERFITTING AND OVERFITTING

Overfitting

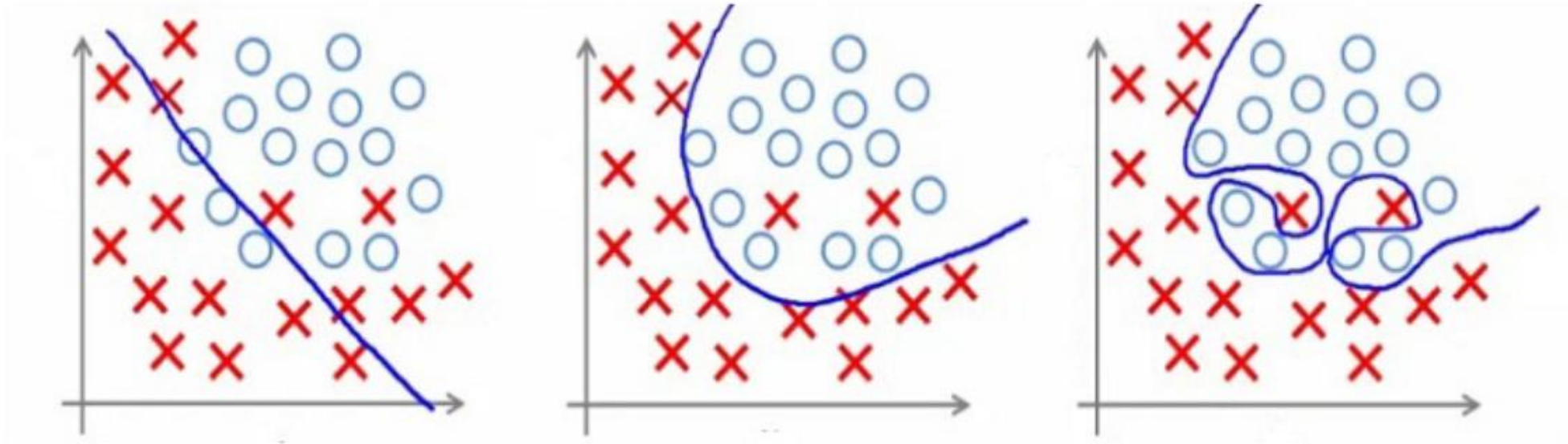
- Refers to a model that models the training data **too well**.
  - when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
- Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function.

Underfitting

- Refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
- Often not discussed as it is easy to detect given a good performance metric.

UNDERFITTING AND OVERFITTING

Example: Classification



Under-fitting

(too simple to explain the variance)

Appropriate-fitting

Over-fitting

(forcefitting -- too good to be true)

May be caused by noise (unwanted anomaly in the data)

SUMMARY

- Data Science
  - The Definition of Data Science
  - The Role of a Data Scientist in a Project
  - The Jobs Available, Salary and Skills Required for Data Scientist
  - The Flow of a Data Science Project
- Machine Learning (ML)
  - The Definition of ML
  - The Differences between Artificial Intelligence (AI), ML and Deep Learning (DL)
  - The Types of ML and its Applications and Algorithm
  - The Introduction of Instance Space, Label Space and Hypothesis Space of ML model
  - The flow of creating ML model
  - The metric evaluation of ML classification model
  - Generalization, Underfitting and Overfitting

THANK YOU