A complex network graph composed of numerous small, glowing white nodes connected by thin white lines, creating a dense web-like pattern.

goML Part 1

[Class Materials]

https://bit.ly/goML_Jan2021

DAY 1

Mr Seow Khee Wei / Mr Shubham Khare

Before we start...

- **Mute** your microphone when not speaking
- **Unmute** when you are answering questions / or asking questions in class
- Give me **feedback** as I need to know how you are doing so that I can adjust my pace or explain any concepts again.



[source](#)



Warm up!

Step 1: Go to the following url

https://bit.ly/goML_warmup



Step 2: facilitator will walk you through the following 2 questions

- 1) Write down what you know about or prior experience with Artificial Intelligence and machine learning**

- 2) What do you hope to get out of this class.**



5 mins



Introduction



Name
SEOW Khee Wei

Telegram
@kwseow

Email
seow_khee_wei@rp.edu.sg

LinkedIn
www.linkedin.com/in/kwseow

Education

Nvidia Certified Instructor
Advanced Certificate in Learning and Performance
NanoDegree (AI)
MBA
BEng

Experience

Assistant Director (Capability and Industry)
Entrepreneur
Programme Chair (Interactive & Digital Media)
Technology Centre Manager (IT Security)
Solution Manager
R&D Engineer

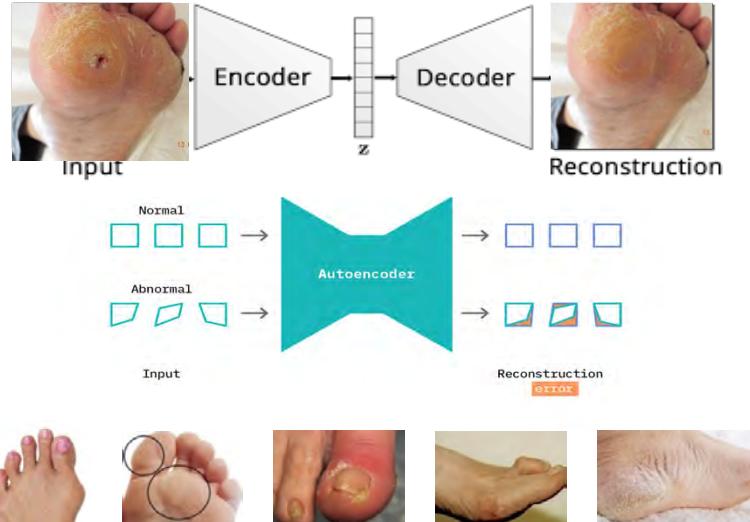


Projects

Scene Understanding



Diabetic Foot At Risk Reconstruction



NER model for Medical Certificates





Programmes

Day 1	<p>What is Data?</p> <p>Types of Machine learning - Supervised and unsupervised learning Understand the common learning methods and how they are used to solve problems.</p> <p>What they can do and what they cannot do</p> <p>Use case sharing</p> <p>Activity – Numpy</p>	<p>Machine learning Workflow Develop a reusable pipeline for project workflow</p> <p>Data Visualization, Preparation and Cleaning Preparing your data for machine learning including feature engineering</p> <p>Activity – Pandas, Matplotlib, Seaborn, data prep</p>
Day 2	<p>Regression techniques</p> <ul style="list-style-type: none"> - Training a Regression Model using Linear Regression - Training a Regression Model using Neural Network - HDB resale price predictor <p>Classification techniques</p> <p>Activity –</p> <ul style="list-style-type: none"> - Classification with Logistic Regression - Classification with Neural Network - Multi Class Classification 	<p>Model improvement Improve the performance of any model using simple hyperparameter tuning</p> <p>Activity: HyperParameter</p> <p>Create successful projects that matters Brainstorming to find your ML use case</p> <p>ML Project Checklist</p> <p>Quiz</p>



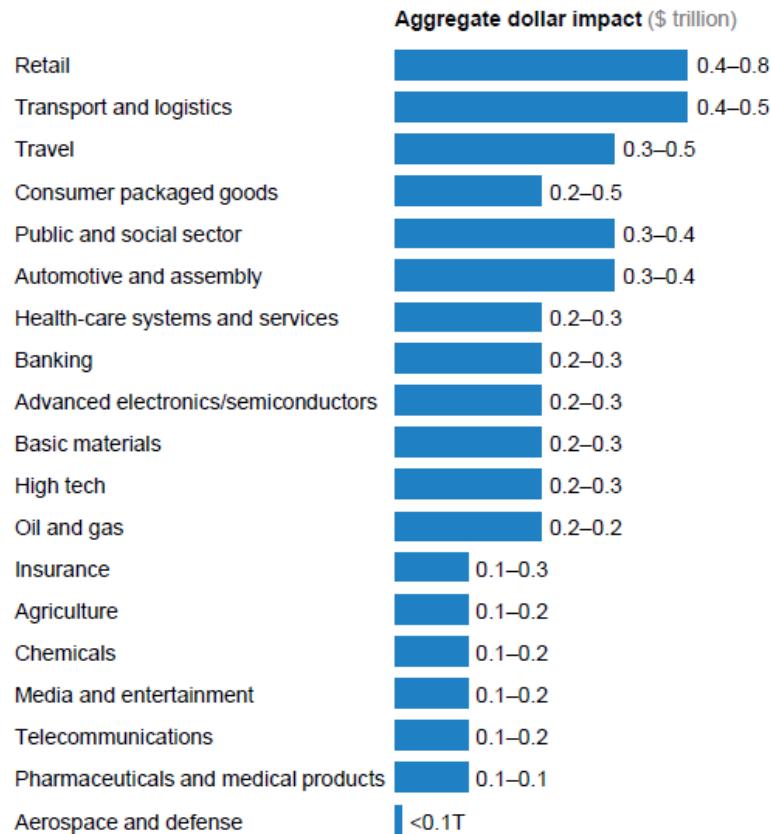
Intro to AI and ML

Which industry will not be affected by AI?

AI value creation
By 2030

\$13
trillion

The potential value of AI by sector

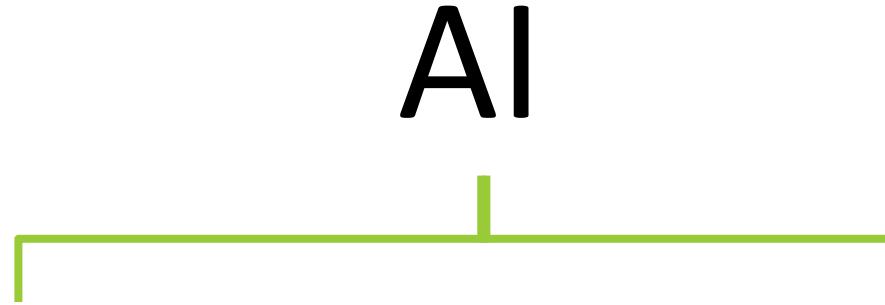


NOTE: Artificial Intelligence here includes neural networks only. Numbers may not sum due to rounding.

[Source: McKinsey Global Institute]



ANI vs AGI

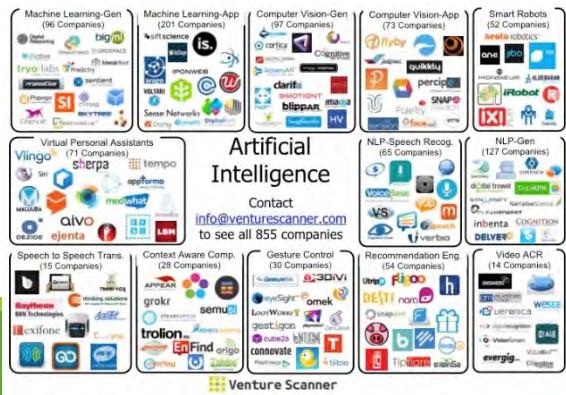


ANI

(Artificial Narrow Intelligence)

E.g. Smart Speaker, self-driving car, web search, AI in farming, route planning for logistic, AI in healthcare, finance

Just using some of the capabilities of AI for a narrow function



AGI

(Artificial General Intelligence)

Do anything human can do

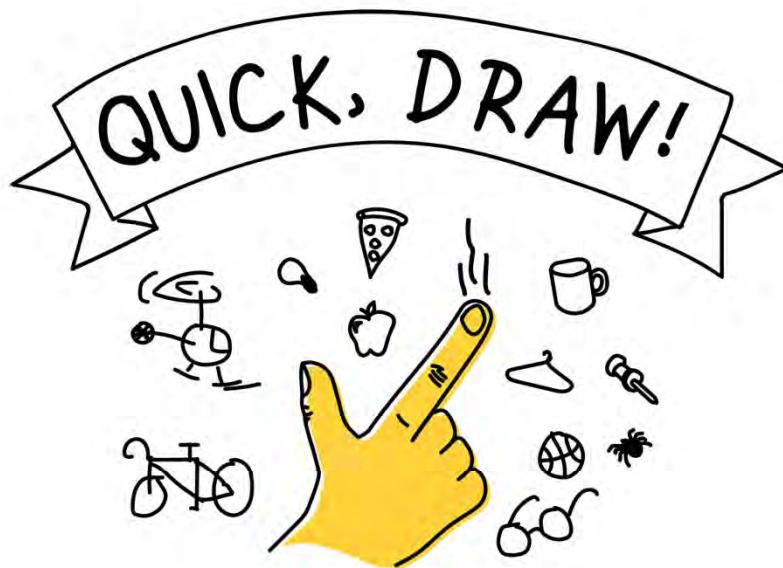
Capable of adapting to any situation
Able to respond as a human would





Activity: Quickdraw Game

<https://g.co/quickdraw>



Can a neural network learn to recognize doodling?

Help teach it by adding your drawings to the [world's largest doodling data set](#), shared publicly to help with machine learning research.

Let's Draw!



5 mins



Activity: Discussion

Step 1: Go to the following url

https://bit.ly/activity_21



Step 2: Pen down your thots on the following 3 questions

- 1) How does the game work?**

- 2) How is it recognizing your drawing?**

- 3) How could we program this?**

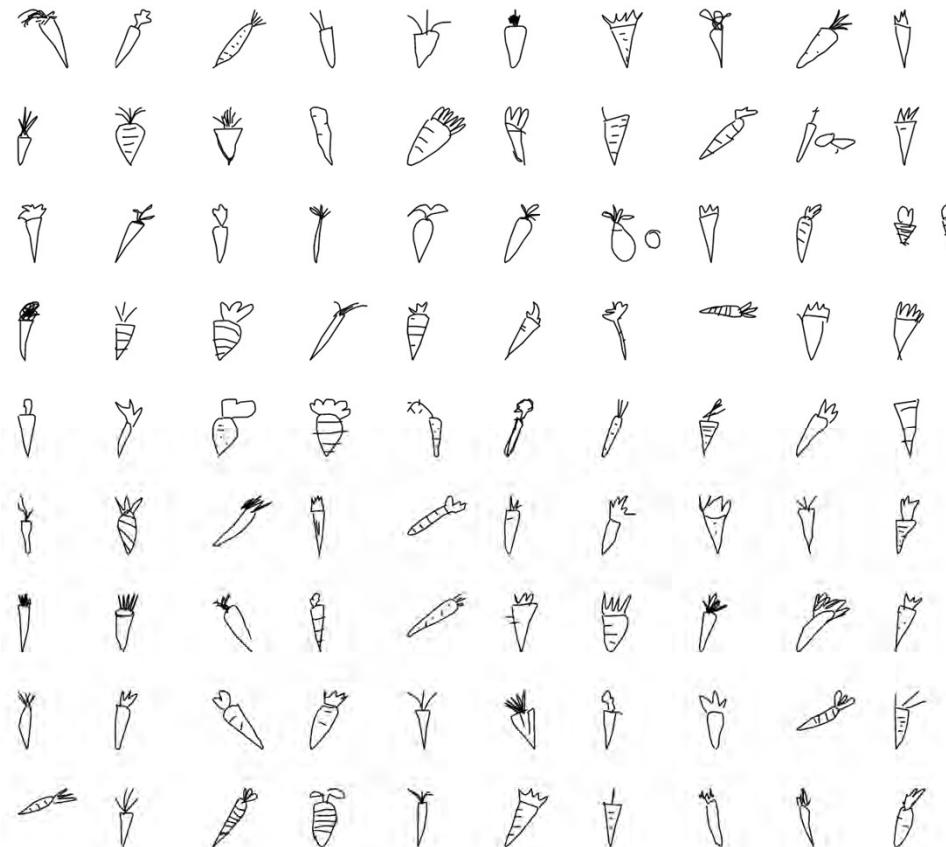


10 mins



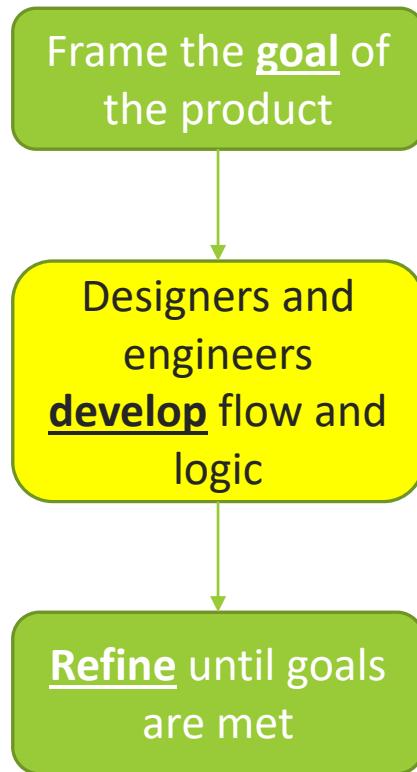
How does ML work in Quickdraw?

g.co/quickdrawdata





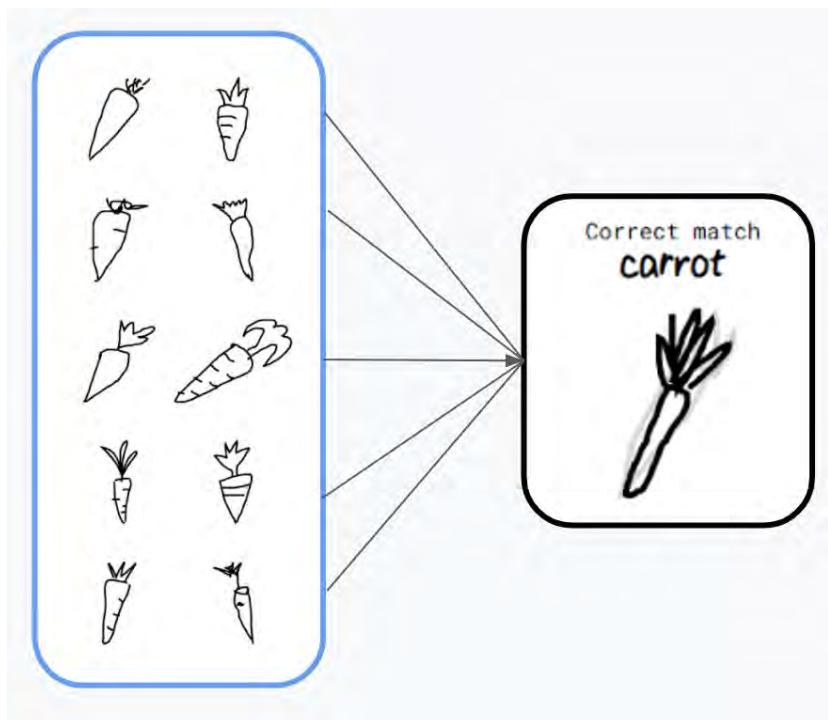
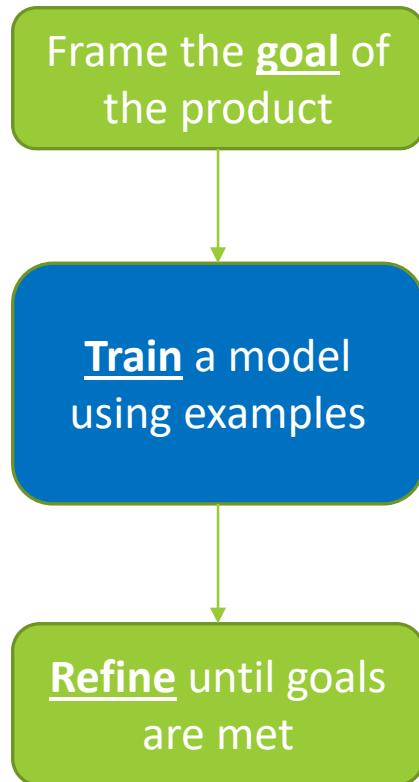
Rule Based



```
if object.height > 10 :  
    do x  
if object.color is blue :  
    do y  
if object.numberOfLegs > 2:  
    do z  
...
```



Machine Learning





Which approach to use?

Alphabetising
a list of song titles

Rule-based

Machine Learning



Which approach to use?

Classifying an
object in a photo

Rule-based

Machine Learning



Which approach to use?

Ranking web
search results

Rule-based

Machine Learning



Download notebooks

- Go to https://bit.ly/goML_Jan2021 to download notebooks.zip
- Unzip and copy the whole directory to your home directory or where your jupyter notebook's startup directory.
- You should have the following (in my case, jupyter notebook starts in c:\Users\seow_khee_wei)

	Name	Date modified	Type	Size
seow_khee_wei				
Documents	.ipynb_checkpoints	12/31/2021 10:03 ...	File folder	
Downloads	data	12/30/2021 4:58 PM	File folder	
Pictures	models	12/30/2021 4:58 PM	File folder	
Dec 2021 (Micron)	test	12/31/2021 10:19 ...	File folder	
Nov 2021	1_1_numpy.ipynb	12/30/2021 5:01 PM	IPYNB File	167 KB
SOI	1_2_Pandas.ipynb	12/30/2021 5:05 PM	IPYNB File	162 KB
AI Guild	1_3_Matplotlib.ipynb	12/30/2021 5:06 PM	IPYNB File	208 KB
SLMP-1	1_4_Seaborn.ipynb	12/30/2021 5:07 PM	IPYNB File	822 KB
SkinAI-1A	1_5_Data_Preparation.ipynb	12/30/2021 5:09 PM	IPYNB File	53 KB
Anaconda Setup	2_1_Train_a_Regression_Model_Using_Lin...	12/30/2021 5:11 PM	IPYNB File	25 KB
AY2122-S2-C300	2_2_Train_a_Regression_Model_Using_Ne...	12/30/2021 5:19 PM	IPYNB File	37 KB
goML_Micron_notebook	2_3_HDB_Retale.ipynb	12/30/2021 5:21 PM	IPYNB File	35 KB
Jan 2022 (Micron)	2_3_HDB_Retale_Solutions.ipynb	12/29/2021 6:24 PM	IPYNB File	148 KB
dropbox-NamespaceExtensionRole.Pe	2_4_Train_a_Classifier_with_Logistic_Regr...	12/30/2021 5:22 PM	IPYNB File	12 KB
OneDrive - Personal	2_5_Train_a_Classifier_with_Neural_Netwo...	12/30/2021 5:25 PM	IPYNB File	5 KB
Attachments	2_6_Classifier_Iris.ipynb	12/30/2021 5:26 PM	IPYNB File	5 KB
COURSES	2_7_Classifier_Use_Case.ipynb	12/30/2021 5:28 PM	IPYNB File	18 KB
DATASETS	2_8_Regression_Model_Hyperparameter_...	12/30/2021 5:51 PM	IPYNB File	26 KB
Documents	2_9_Classifier_Hyper_parameter_tuning.ip...	12/30/2021 6:41 PM	IPYNB File	22 KB
PERSONAL	2_10_Neural_Network_Hyper_parameter_...	12/30/2021 7:05 PM	IPYNB File	11 KB
Pictures	environment.yml	12/29/2021 7:15 PM	YML File	10 KB
	notebooks.zip	12/31/2021 8:25 AM	Compressed (zipp...	2,583 KB
	test_import.ipynb	12/31/2021 10:20 ...	IPYNB File	4 KB

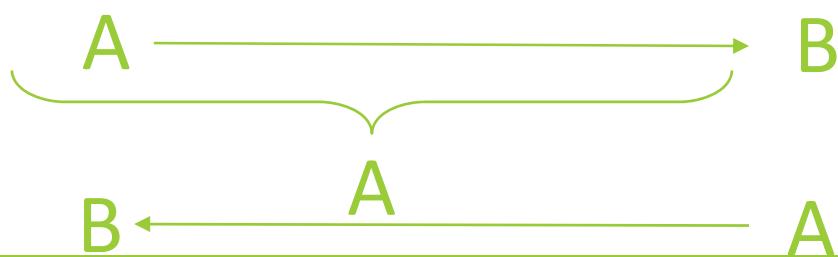


What is Data

Floor area sqm (Sqm)	Storey range	Street name	Remaining lease	Resale price (\$)
131	01 TO 03	TAMPINES ST 45	74 years 01 month	540,000
120	10 TO 12	TAMPINES ST 71	74 years 11 months	578,888
119	04 TO 06	TAMPINES ST 71	74 years 10 months	545,000
121	01 TO 03	TAMPINES ST 83	66 years 10 months	520,000
146	04 TO 06	TAMPINES AVE 5	65 years 06 months	768,000
146	07 TO 09	TAMPINES AVE 5	66 years 05 months	755,000
148	04 TO 06	TAMPINES ST 11	63 years 09 months	655,000

image	label
	cat
	not cat
	cat
	not cat

A → B





Acquiring data

- Manual labeling



- From observing behaviours

user ID	time	price (\$)	purchased
4783	Jan 21 08:15:20	7.95	yes
3893	March 3 11:30:15	10.00	yes
8384	June 11 14:15:05	9.50	no
0931	Aug 2 20:30:55	12.90	yes

- Collecting from machines' log

machine	temperature (°C)	pressure (psi)	machine fault
17987	60	7.65	N
34672	100	25.50	N
08542	140	75.50	Y
98536	165	125.00	Y

- Download from websites/partner ships

Datasets

Explore, analyze, and share quality data. Learn more about data types, creating, and collaborating.

[New Dataset](#)

Search datasets

Datasets Tasks Computer Science Education Classification Computer Vision NLP Data Visualization

Trending Datasets

Cleaned data for the chatbot with model weights | Indian Male baby names | US Personal Expenditures by State 1997-2019

Registry of Open Data on AWS

This registry exists to help people discover and share datasets that are available via AWS resources. Learn more about sharing data on AWS. See all usage examples for datasets listed in this registry.

See datasets from Allen Institute for Artificial Intelligence (AI2), Digital Earth Africa, Facebook Data for Good, NASA Space Act Agreement, NIH STRIDES, NOAA Big Data Program, Space Telescope Science Institute, and Amazon Sustainability Data Initiative.

[View Registry](#)

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA), a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive molecular profiles of the key genomic, epigenetic, and other alterations of cancer. TCGA has analyzed more than 10,000 tumor and normal tissue samples, allowing for the comprehensive characterization of 33 cancer types and subtypes, including 10 rare cancers. The dataset contains open Clinical Supplement, Biospecimen Supplement, RNA-Seq Gene Expression Quantification, miRNA-Seq Isoform Expression Quantification.



Dataset

- Some sites where you can download open dataset

GitHub

<https://kwseow.github.io/>

kaggle

<https://www.kaggle.com/datasets>

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

<https://www.re3data.org/>

Google

Dataset Search Beta

<https://datasetsearch.research.google.com/>



<https://www.kdnuggets.com/datasets/index.html>



Mendeley Data

<https://data.mendeley.com/>



Open Manufacturing Dataset

UCI SECOM Dataset

Semiconductor manufacturing process dataset

<https://www.kaggle.com/paresh2047/uci-semcom>

Superconductivity Data Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Two files contain data on 21263 superconductors and their relevant features.

<http://archive.ics.uci.edu/ml/datasets/Superconductivity+Data>

Gas sensor arrays in open sampling settings Data Set

Download: [Data Folder](#), [Data Set Description](#)

<https://archive.ics.uci.edu/ml/datasets/Gas+sensor+arrays+in+open+sampling+settings>

CNC Mill Tool Wear

Variational CNC machining data

<https://www.kaggle.com/shasun/tool-wear-detection-in-cnc-mill/data>



PHM DATA Challenge 18: Etching tool fault detection (PdM)

<https://phmsociety.org/conference/annual-conference-of-the-phm-society/annual-conference-of-the-prognostics-and-health-management-society-2018-b/phm-data-challenge-6/>



Data is messy

- Garbage in, garbage out

- Data problems
 - Incorrect labels
 - Missing values

- Multiple types of data

Images, audio, text
unstructured

size of house (square feet)	# of bedrooms	price (1000\$)
523	1	115
645	1	0.001
708	unknown	210
1034	3	unknown
unknown	4	355
2545	unknown	440

structured





AI Terminology

- Machine Learning vs Data Science

Floor area sqm (Sqm)	Storey range	Street name	Remaining lease	Resale price (\$)
131	01 TO 03	TAMPINES ST 45	74 years 01 month	540,000
120	10 TO 12	TAMPINES ST 71	74 years 11 months	578,888
119	04 TO 06	TAMPINES ST 71	74 years 10 months	545,000
121	01 TO 03	TAMPINES ST 83	66 years 10 months	520,000
146	04 TO 06	TAMPINES AVE 5	65 years 06 months	768,000
146	07 TO 09	TAMPINES AVE 5	66 years 05 months	755,000
148	04 TO 06	TAMPINES ST 11	63 years 09 months	655,000

ML: A → B

Running AI system

(e.g. websites / mobile app)

A

DS: Homes with 3 bedrooms are more expensive than homes with 2 bedrooms of a similar size.

Hindsight/Insight

Foresight

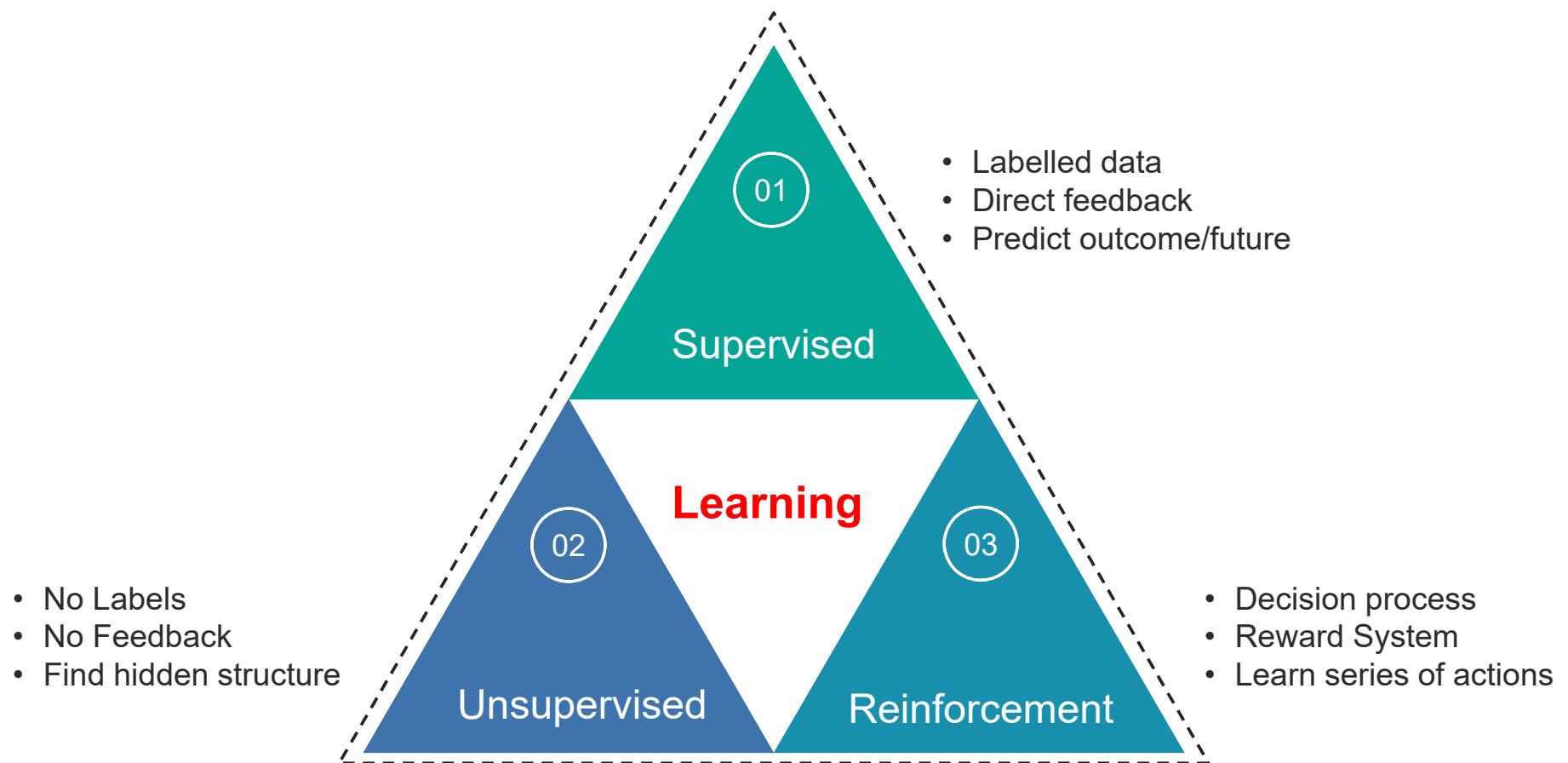


ML vs DS

- Machine Learning
 - Field of study that gives computers the ability to learn without being explicitly programmed – Arthur Samuel (1959)
 - Often a piece of software
- **Data Science**
 - Science of extracting knowledge and insights from data
 - Often a slide deck



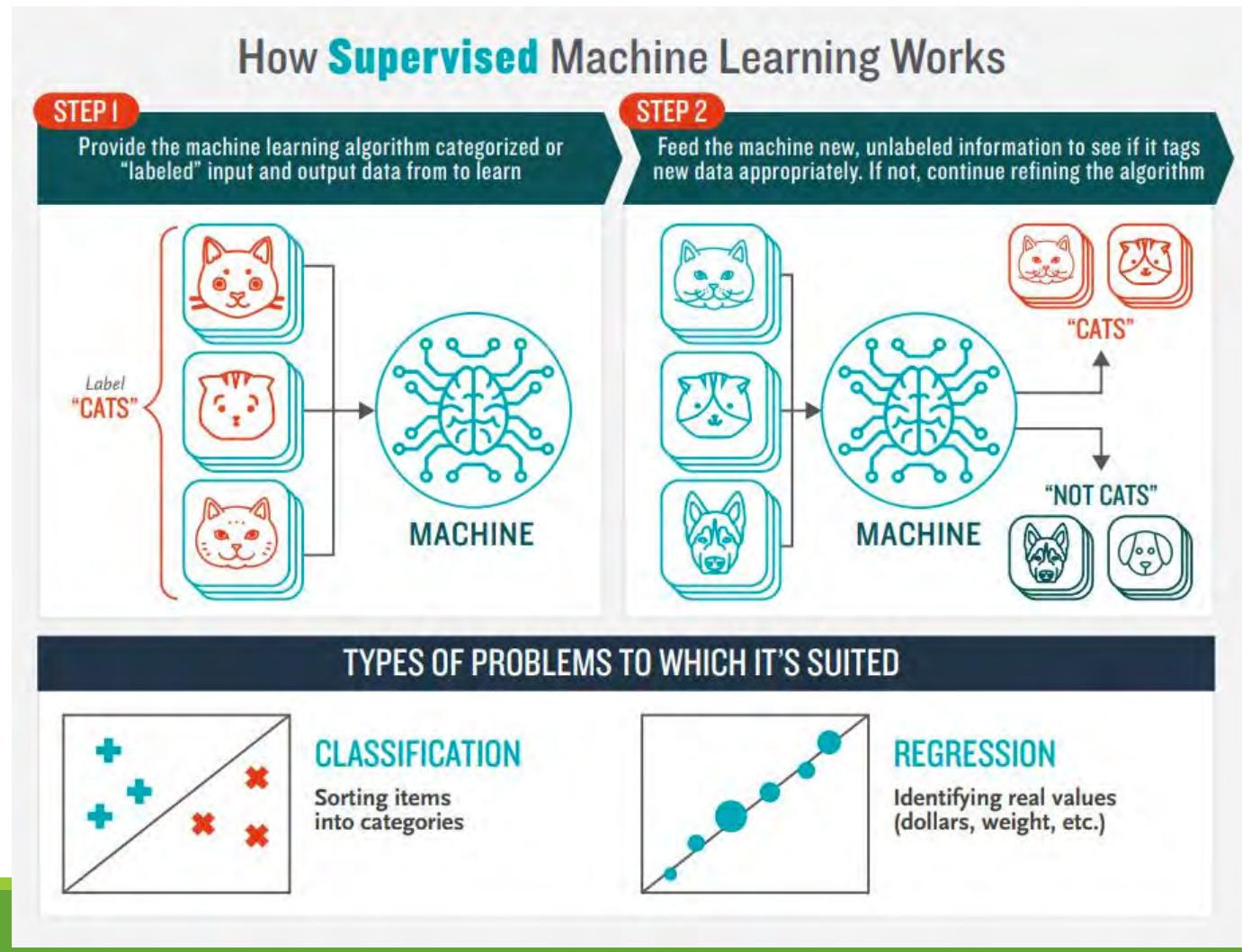
Three kind of machine learning





Supervised Learning

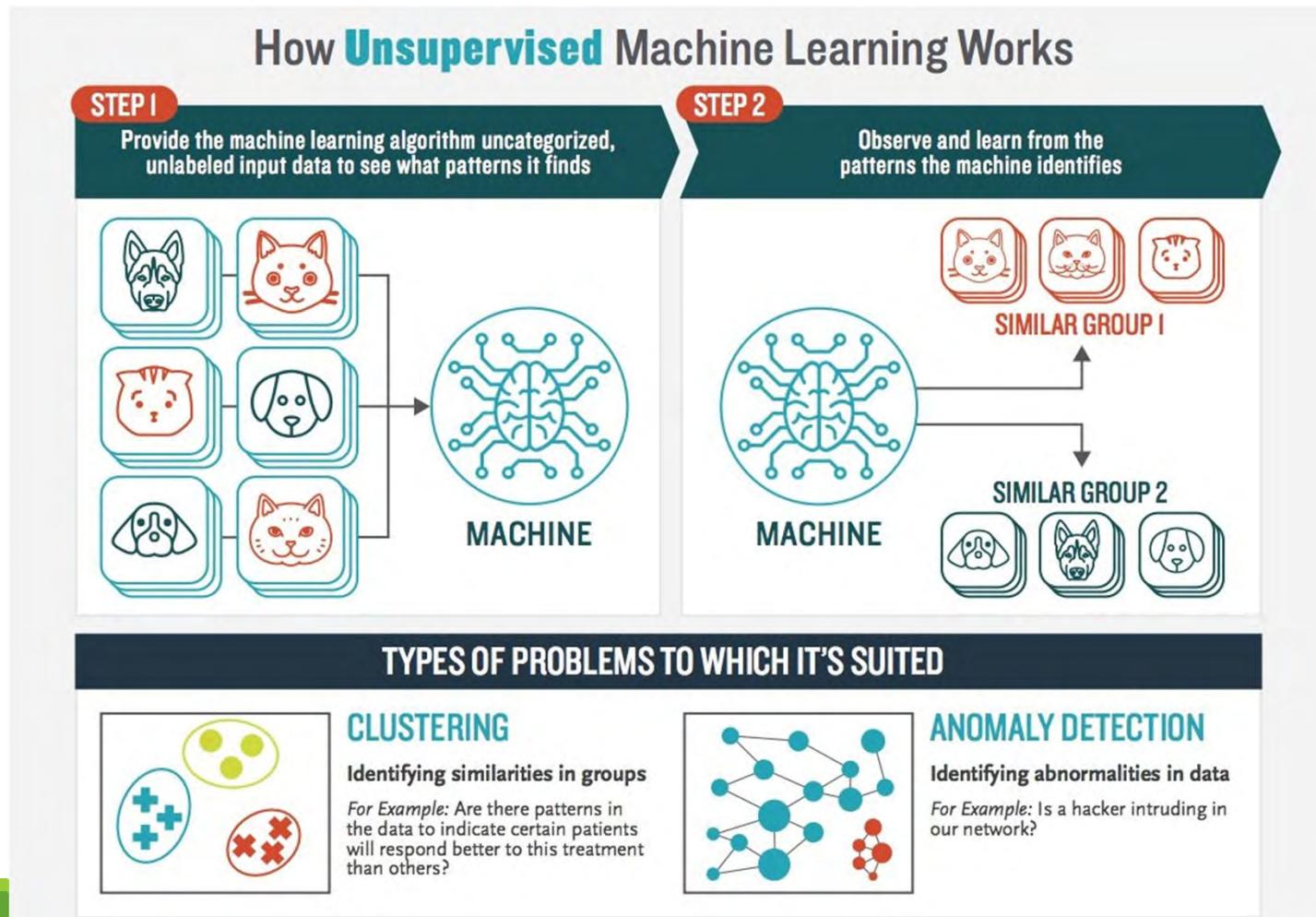
Data points have **known** outcome





Unsupervised Learning

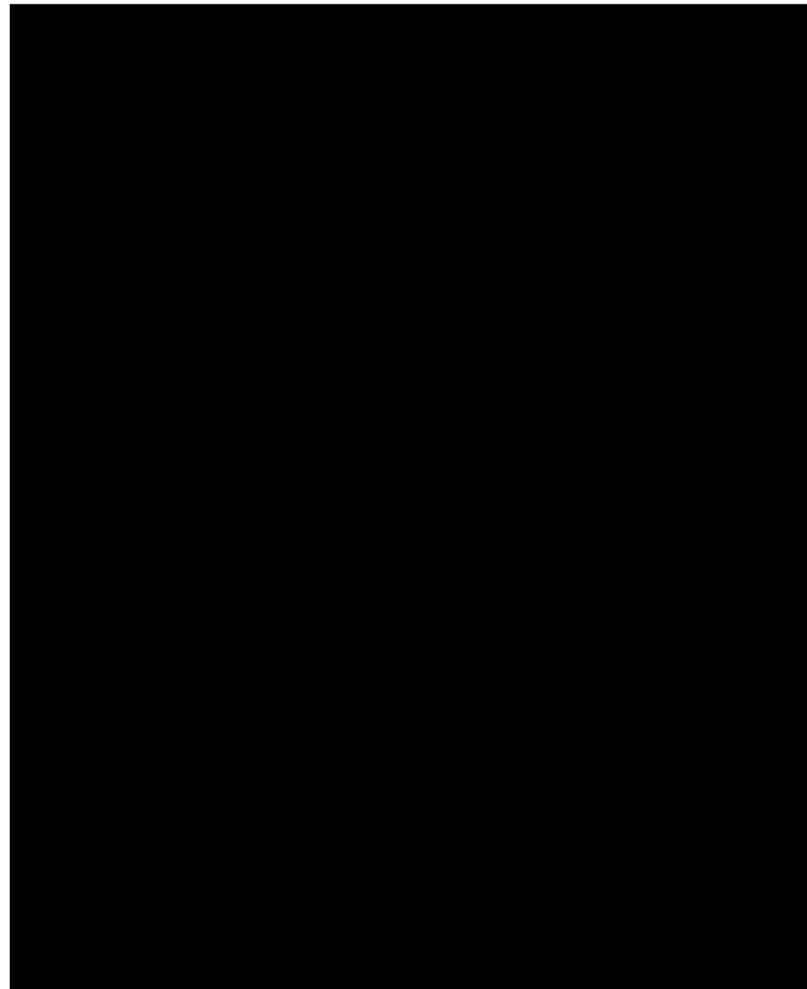
Data points have **unknown** outcome





Reinforcement Learning

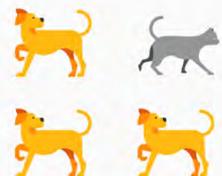
- Reinforcement learning





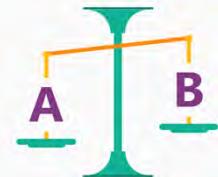
5 fundamental questions

**Is this weird?
(Anomaly detection)**



Is this pressure gauge reading normal?
Is this message from the internet typical?

**Is this A or B?
(Classification)
(discrete values)**



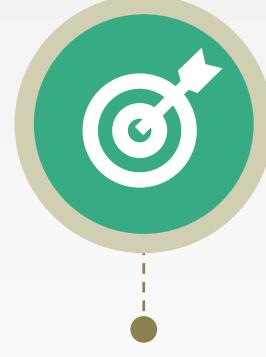
Will this tire fail in the next 1,000 miles: Yes or no?
Which brings in more customers: a \$5 coupon or a 25% discount?

**How many?
How Much?
(Regression)
(Continuous)**



What will the temperature be next Tuesday?
What will my fourth quarter sales be?

**How is this organized?
(Clustering)**



Which viewers like the same types of movies?
Which printer models fail the same way?

**What should I do?
(Reinforce Learning)**



If I'm a self-driving car: At a yellow light, brake or accelerate?
For a robot vacuum: Keep vacuuming, or go back to the charging station?



Type of machine learning

Labeling email as
spam or not-spam

Anomaly Detection

Classification

Regression

Clustering

Reinforcement learning



Type of machine learning

Identifying trends amongst a group of people who have bought a new music release

Anomaly Detection

Classification

Regression

Clustering

Reinforcement learning



Type of machine learning

Predicting the strength of a password

Anomaly Detection

Classification

Regression

Clustering

Reinforcement learning



Type of machine learning

Does this foot
look unusual?

Anomaly Detection

Classification

Regression

Clustering

Reinforcement learning



Machine Learning Example

- Suppose you wanted to identify fraudulent credit card transactions.
- You could define features to be:
 - Transaction time
 - Transaction amount
 - Transaction location
 - Category of purchase
- The algorithm could learn what feature combinations suggest unusual activity.





Machine Learning Limitations

- Suppose you wanted to determine if an image is of a cat or a dog.
- What features would you use?
- This is where **Deep Learning** can come in.



Dog and cat recognition

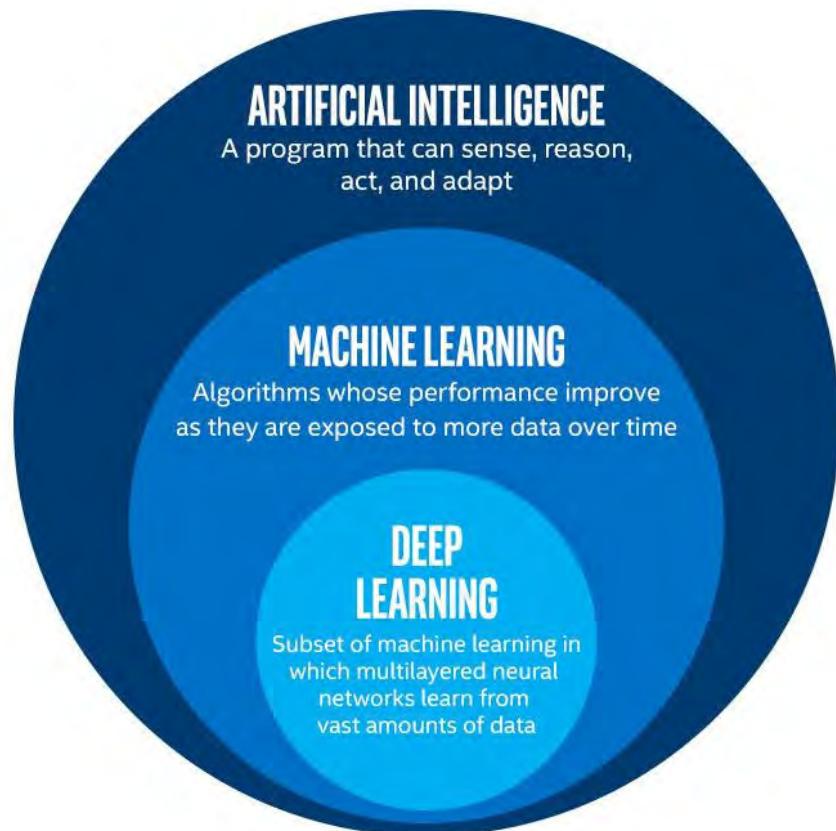


What is deep learning?

Deep Learning

“Machine learning that involves using very complicated models called “deep neural networks”.”
(Intel)

Models determine best representation of original data; in classic machine learning, humans must do this.

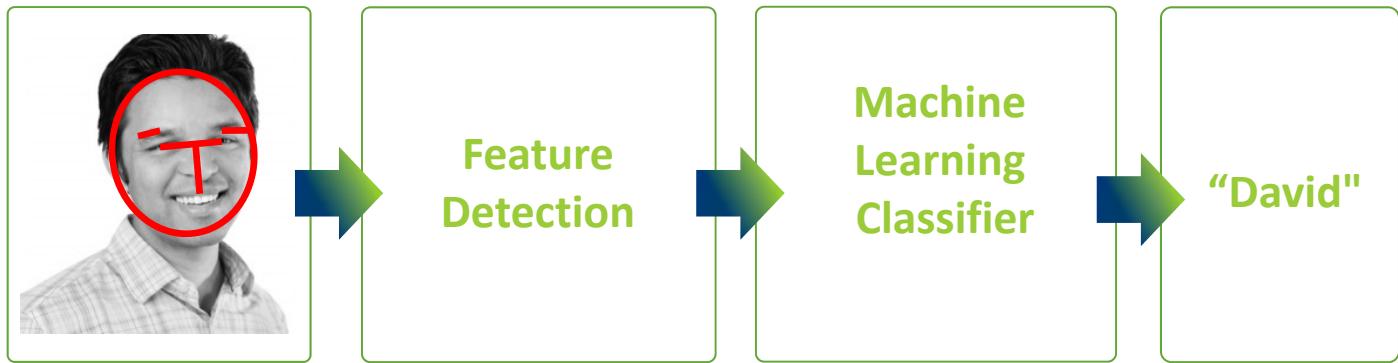




Deep Learning Example

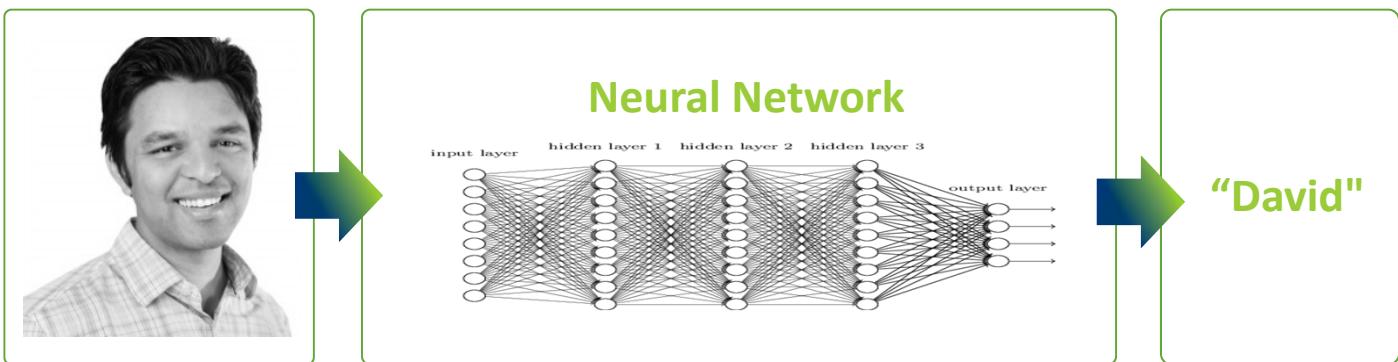
Classic Machine Learning

Step 1: Determine features.
Step 2: Feed them through model.



Deep Learning

Steps 1 and 2 are combined into 1 step.

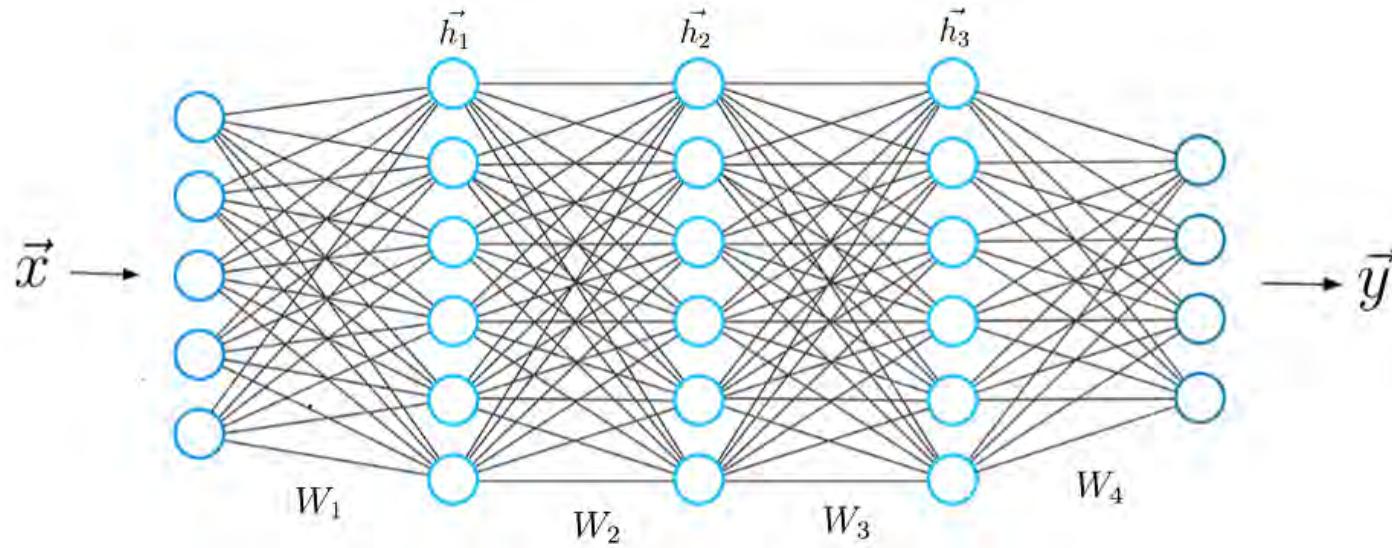




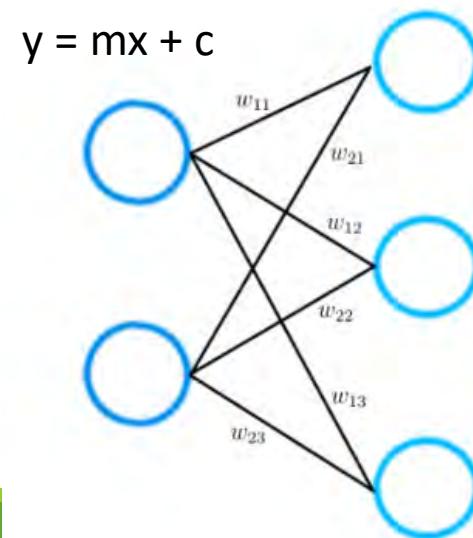
What is a Neural Network?



Neural Networks

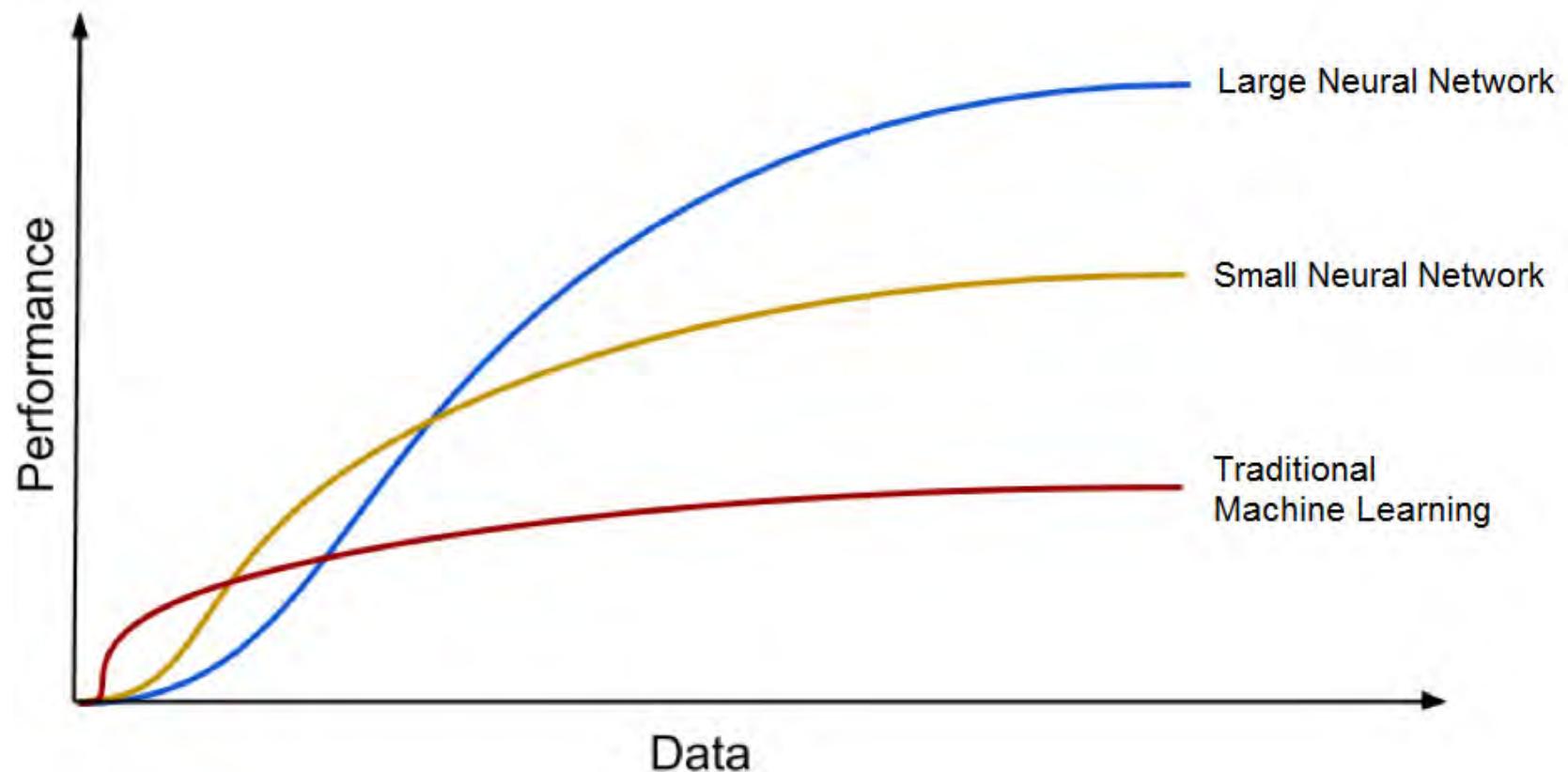


The challenge in training a neural network is finding a set of weights that give the most accurate output.





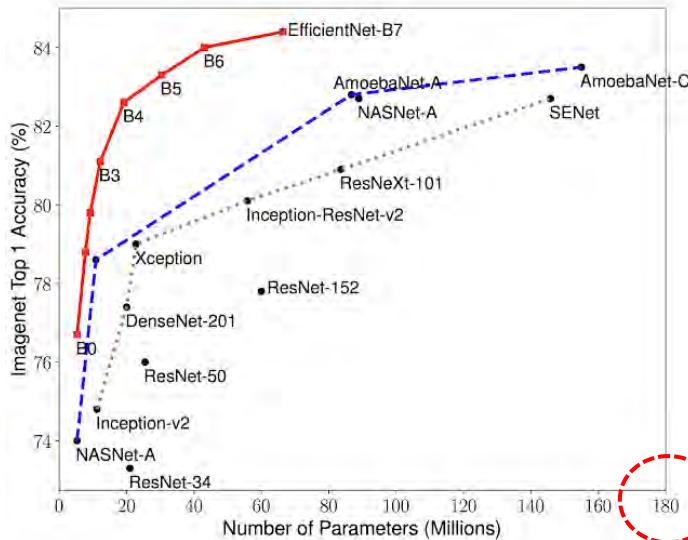
Performance



Deep Learning Algorithms get better with the increasing amount of data.



Size

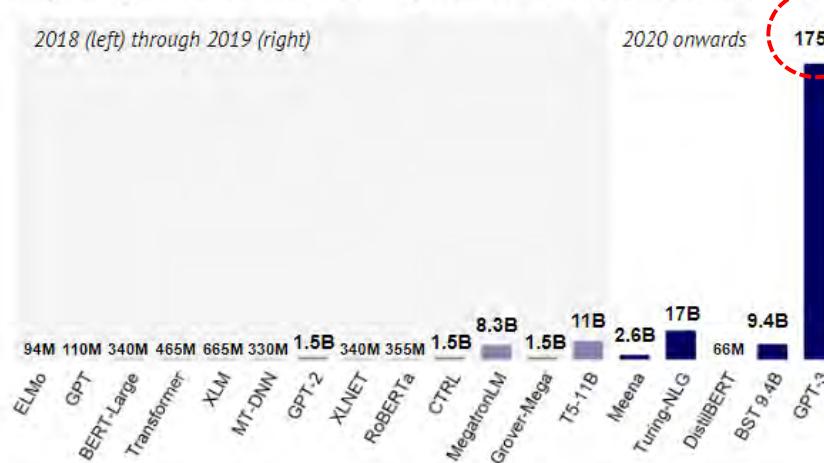


Huge models, large companies and massive training costs dominate the hottest area of AI today, NLP.

2018 (left) through 2019 (right)

2020 onwards

175B



Note: The number of parameters indicates how many different coefficients the algorithm optimizes during the training process.



Tools

Frameworks



Edge On-device Inference



Visualisation and experiment tracking



TensorBoard



Weights & Biases



neptune.ai



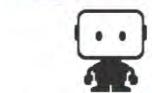
mlflow

Low Code / Code Free



PELTARION

AutoML



DataRobot

ML infrastructure and
compute



Paperspace



Lambda



FLOYDHUB

The fastest way to build, train, and deploy deep learning models



Cloud inference & ML as a service



Amazon SageMaker



Azure Machine Learning



Google Cloud





ML Technical Tools

- Research Publications
 - Arxiv
- Open source repositories
 - Github

Cornell University

arXiv.org

arXiv is a free distribution service and an open-access archive for 1,302,153 scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. Materials on this site are not peer-reviewed by arXiv.

Subject search and browse: Physics [Search](#) [Form interface](#) [Catchup](#)

News
Read about recent news and updates on arXiv's [blog](#). (View the former "what's new" pages here). Read [robots beware](#) before attempting any automated download.

Physics

- Astrophysics ([astro-ph](#) new, recent, search)
 - Includes: Astrophysics of Galaxies; Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics;
- Condensed Matter ([cond-mat](#) new, recent, search)
 - Includes: Disordered Systems and Neural Networks; Materials Science; Mesoscale and Nanoscale Physics; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons
- General Relativity and Quantum Cosmology ([gr-qc](#) new, recent, search)
 - Includes: Black Holes; Cosmology; Gravitation; Relativity and Statistical Mechanics
- High Energy Physics - Experiment ([hep-ex](#) new, recent, search)
 - Includes: Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
- High Energy Physics - Phenomenology ([hep-ph](#) new, recent, search)
 - Includes: High Energy Physics - Phenomenology
- High Energy Physics - Theory ([hep-th](#) new, recent, search)
 - Includes: Adelphi Model; AdS/CFT Correspondence; Black Holes; Brane World Scenario; Causal Dynamical Triangulation; Conformal Field Theory; D-Branes; Duality; Extra Dimensions; Gauge Theories; Geometrodynamics; General Relativity; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Plasma Physics; Popular Physics; Quantum Gravity; Quantum Physics; String Theory; Symmetries and Group Representations; Theoretical High Energy Physics
- Mathematical Physics ([math-ph](#) new, recent, search)
 - Includes: Analysis; Mathematical Physics
- Nonlinear Sciences ([nlin](#) new, recent, search)
 - Includes: Adaptation and Self-Organizing Systems; Cellular Automata and Lattice Gases; Chaotic Dynamics; Exactly Solvable and Integrable Systems; Pattern Formation and Solitons
- Nuclear Theory ([nucl-th](#) new, recent, search)
 - Includes: Nuclear Theory
- Physics ([physics](#) new, recent, search)
 - Includes: Acoustics; Atmospheric and Oceanic Physics; Atomic and Molecular Clusters; Atomic Physics; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Plasma Physics; Popular Physics; Quantum Physics
- Quantum Physics ([quant-ph](#) new, recent, search)

We gratefully acknowledge support from the Simons Foundation and member institutions.

Login

Search... All fields [Search](#)

[Help](#) | [Advanced Search](#)

COVID-19 Quick Links

See COVID-19 SARS-CoV-2 preprints from

- arXiv
- medRxiv and bioRxiv

Important: Preprints posted on arXiv are not peer-reviewed by arXiv. They are posted by authors without peer review to guide clinical practice or health-related behavior and should not be used without consulting multiple experts in the field.

Why GitHub? [Team](#) [Enterprise](#) [Explore](#) [Marketplace](#) [Pricing](#)

[Search GitHub](#) [Sign in](#) [Sign up](#)

Where the world builds software

Millions of developers and companies build, ship, and maintain their software on GitHub—the largest and most advanced development platform in the world.

Email address [Sign up for GitHub](#)

65+ million Developers 3+ million Organizations 200+ million Repositories 72% Fortune 50



CPU vs GPU

CPU – Central Processing Unit - Computer processor



GPU – Graphical Processing Unit



Cloud vs On-premises

A graphic element consisting of a white circle with a blue double-line border. Inside the circle is a black silhouette of a teacup with steam rising from it. A small white tag with the word "Tea" is visible inside the cup. The circle is positioned in the center of the slide, partially overlapping a large blue downward-pointing arrow.

10 Mins Break

Back by 11:05



What AI can and cannot do

What happens if you try?

Input (A)

User email

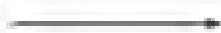


Output (B)

2-3 paragraph response

1000 examples

"My box was damaged."



Thank you for your email.

"Where do I write a review?"



Thank you for your email.

"What's the return policy?"



Thank you for your email.



Customer service

<https://www.youtube.com/watch?v=g89RUEWywPQ>



What ML can and cannot do

Input (A)	Output (B)	Application
Email	Spam? (0/1)	Spam filtering
Audio	Text transcripts	Speech Recognition
English	Chinese	Machine Translation
Image, radar info	Position of other cars	Self-driving car
Image of phone	Detect? (0/1)	Visual inspection

Anything you can do with 1 second of thought, we can probably now or soon automate



What ML can and cannot do

- The toy arrived two days late, so I wasn't able to give it to my niece for her birthday. Can I return it??

Refund Request

Input text (A) →
Refund/Shipping/Other (B)

Oh, sorry to hear that. I hope your niece had a good birthday. Yes, we can help with

Diagnose pneumonia from 10,000 labelled images



Diagnose pneumonia from 10 images of a medical textbook chapter explaining pneumonia



Strengths & Weaknesses of ML

Learning a "simple" concept (<= 1 sec)

You have lots of data available

Learning complex concepts from small amount of data

Asked to perform on new types of data



Why use AI?

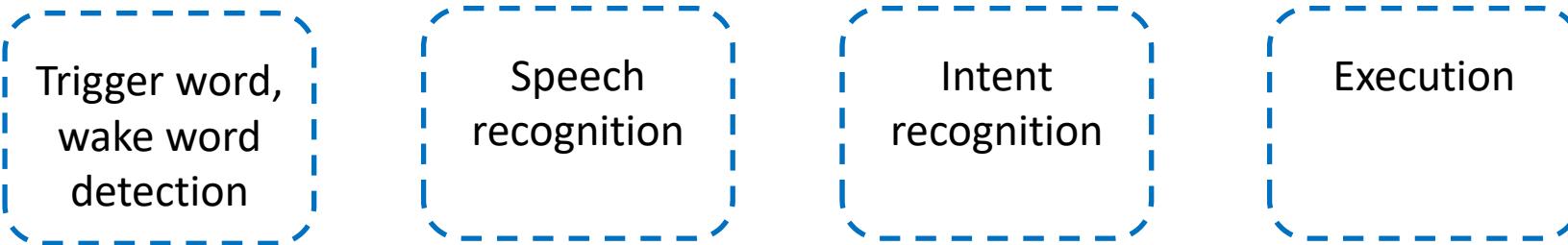
Why use AI	When not to use AI
<i>Solve problems that can't be solved with unintelligent, automated systems</i>	Repetitive, Deterministic Automation Tasks o Just code it or record it and be done with it
Simple Rules Don't Work o If you can use simple rules processing or workflow, you should o Machine learning is probabilistic not deterministic	Formulaic Analytics o This is what big data Business Intelligence (BI) platforms are for
Requires Object Identification or Classification o You can't easily program image recognition. Let the neural network do it.	Systems that require 100% accuracy o If it's trained, then it can't be always right
Pattern matching across large quantities of data o Could you build advanced programs without AI to do it? Maybe. ML is easier	Situations with very little training data o If you can't train it, then machine learning won't work
Probabilistic vs. Deterministic Patterns o If behavior doesn't happen the same way every single time, consider AI.	Situations where hiring a person is easier, cheaper, and faster o Sometimes the human brain just wins out
Advanced Statistics or Analytics is Too Complex o Machine learning does better than non-learning formulas.	A need to do "AI" without understanding what it is or what it's for o There's lots of value in AI without being vague

Use case sharing



Case Study : Smart Speaker

"Hey device, tell me a joke"



Trigger word,
wake word
detection

Audio → "hey
device" ? (0/1)

Speech
recognition

Audio →
"Tell me a joke"

Intent
recognition

"Tell me a joke" →
Joke?
Time?
Music?
Call?
Weather?

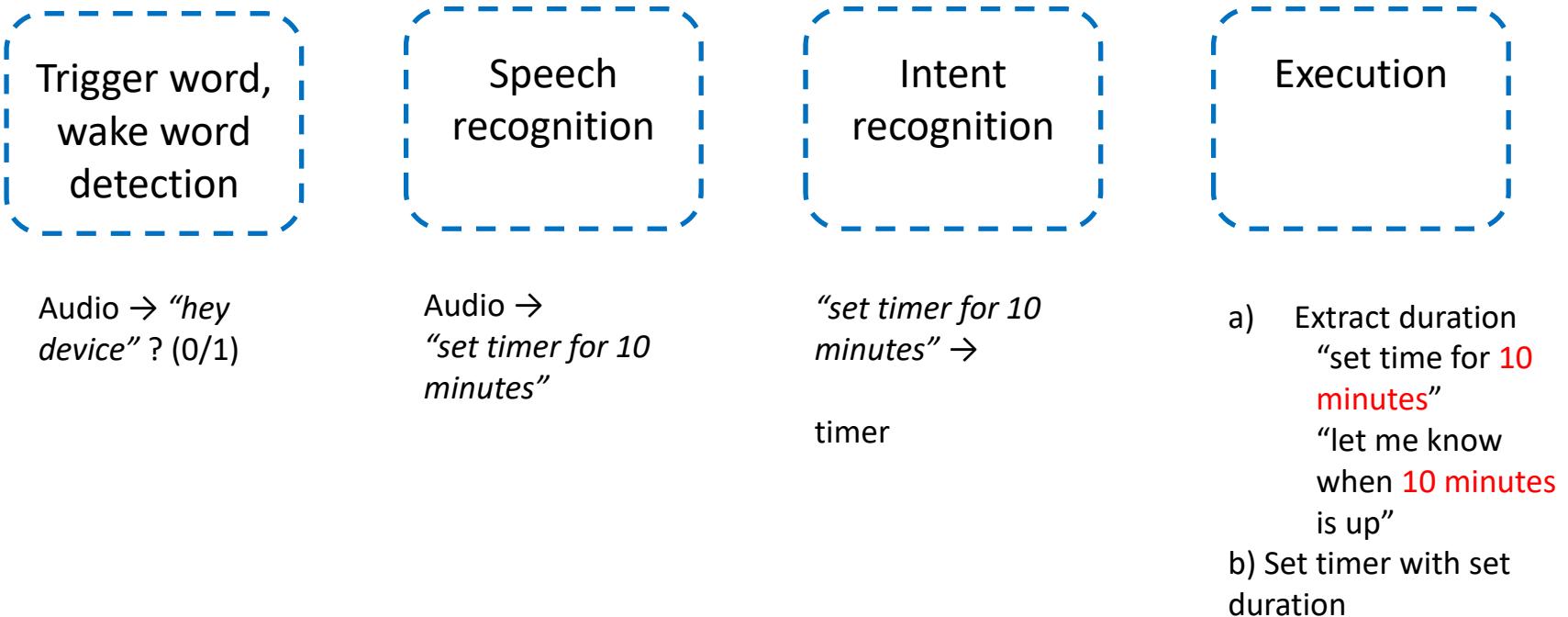
Execution

A piece of software
that can randomly
get a joke and tell
the user



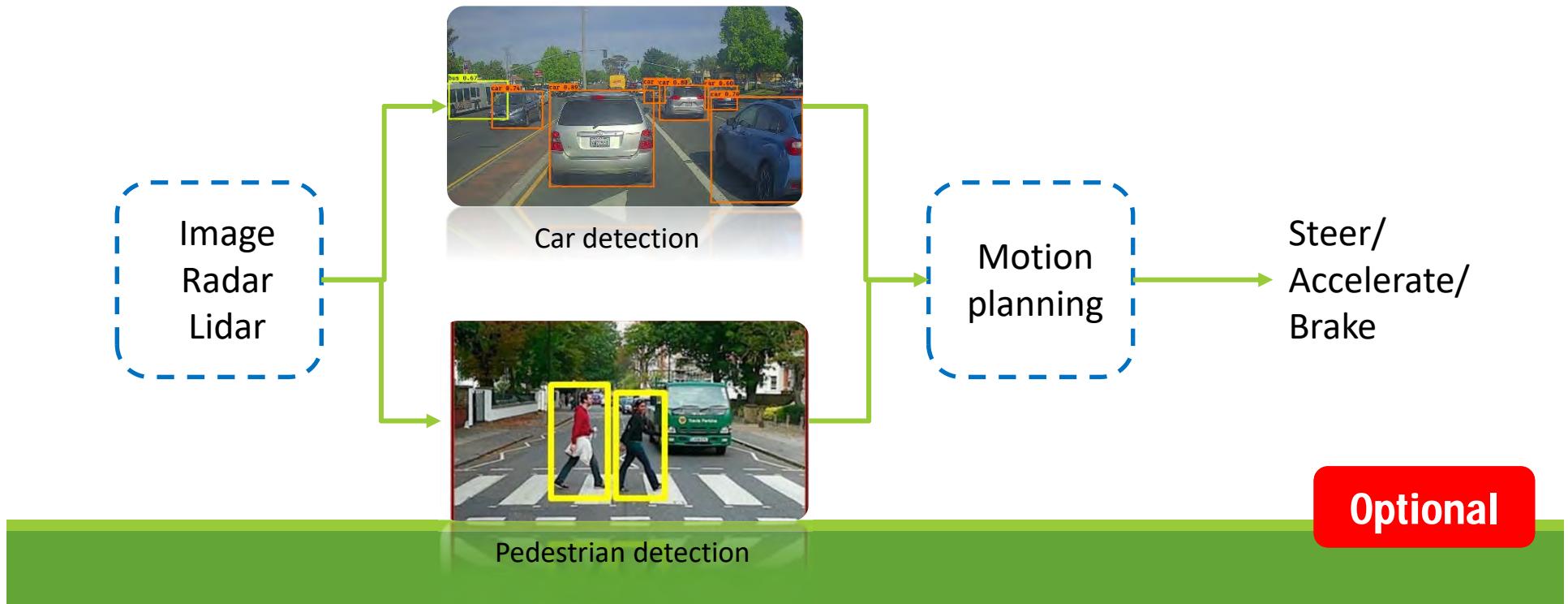
Case Study : Smart Speaker

"Hey device, set timer for 10 minutes"



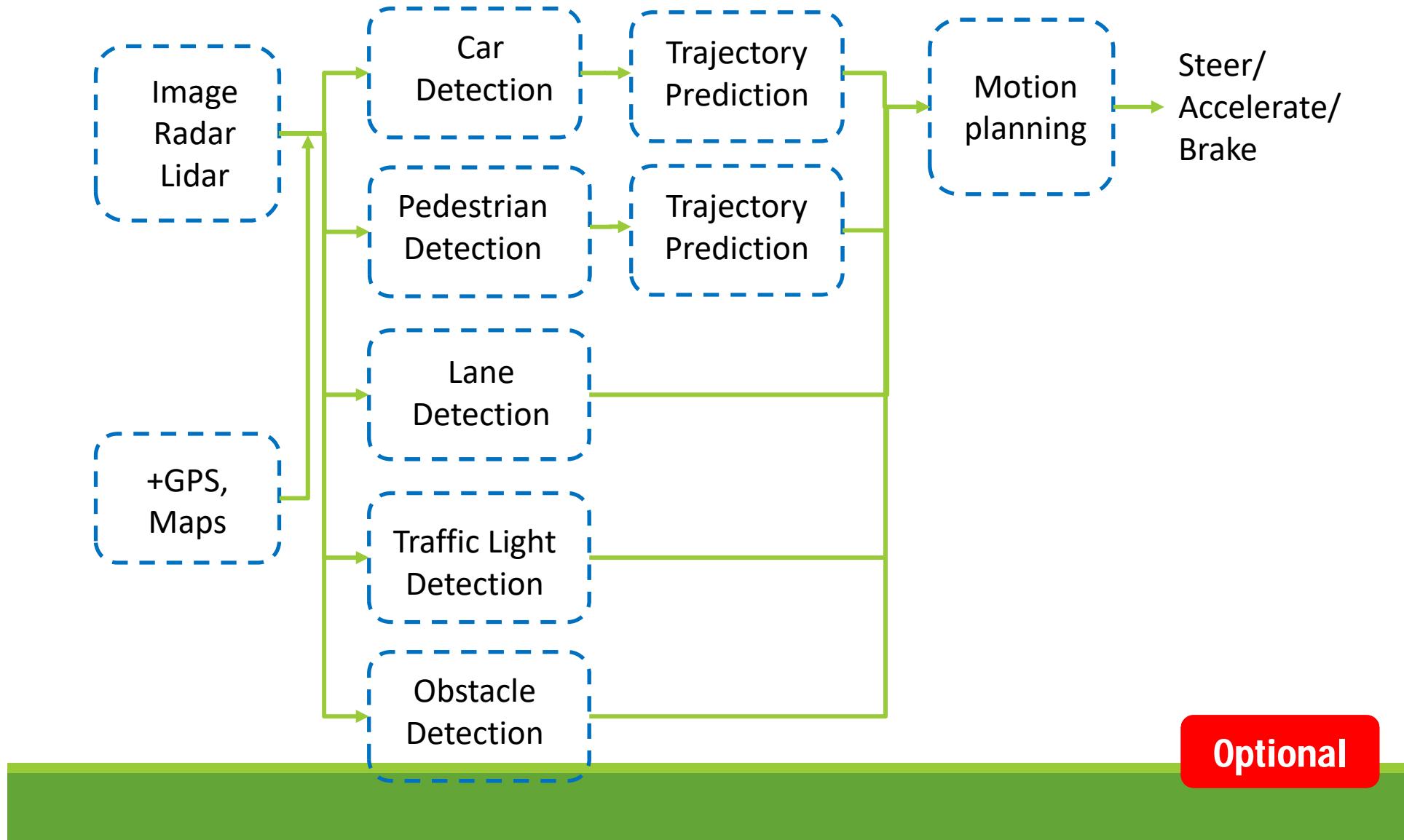


Case Study : Self-driving Car





Case Study : Self-driving Car





Create successful AI use cases and projects

-  01 *Save time → Process automation*
-  02 *Reduce cost → Predicting the future*
-  03 *Improve people's lives → at work or at home*
-  04 *Increase revenue → with predictive marketing*



How?

- Achieving the greater levels of ***productivity***
- Improving ***accuracy***
- Improving ***reliability***
- Eliminating ***drudgery and dull tasks***
- Improving ***relevance***
- Achieving ***greater scale***
- Enabling ***continuous monitoring and auditing***
- Extracting ***more value from data***
- Providing ***assistance to those in need***





AI team composition

- Example Roles
 - Software Engineer -> Joke execution, ensure self-driving reliability
 - Machine Learning Engineer -> generate A->B mapping,
 - Machine Learning Researcher -> extend sota in ML
 - Applied ML Scientist -> between the two above.
 - Data Scientist -> still evolving, examine data and provide insights, make presentation to team/executive
 - Data Engineer
 - Organizing data,
 - Make sure data is saved in an easily accessible, secure and cost effective way.
 - AI Product Manager
 - Help decide what to build, what's feasible and valueable

Skills frameworks : <https://www.skillsfuture.gov.sg/skills-framework/ict>

Data and Artificial Intelligence

<https://www.imda.gov.sg/-/media/IMTalent-Portal-Revamp/3-Guidances/Skills-Planning/SF-ICT/ICT-Navigation-Tool-2020.pdf>

Optional



AI team composition

- Getting started with a small team
 - 1 software engineer, or
 - 1 machine learning engineer/data scientist
 - Nobody but yourself

Optional



Major AI Application

Computer Vision

Image classification/Object Recognition



Dog



- Object Detection



- Image Segmentation



- Style Colorization





Major AI Application

- Natural Language Processing

- Text Classification

email → Spam/Non-Spam

Product description → Product category

- Sentiment recognition

"The food was good" → ❤️ ❤️ ❤️ ❤️

"Service was horrible" → ❤️

- Information retrieval

Web search

- Name entity recognition

organization

person

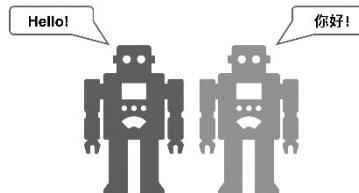
Ousted WeWork founder Adam Neumann listed his

Manhattan penthouse for \$37.5 million

location

monetary value

- Machine translation

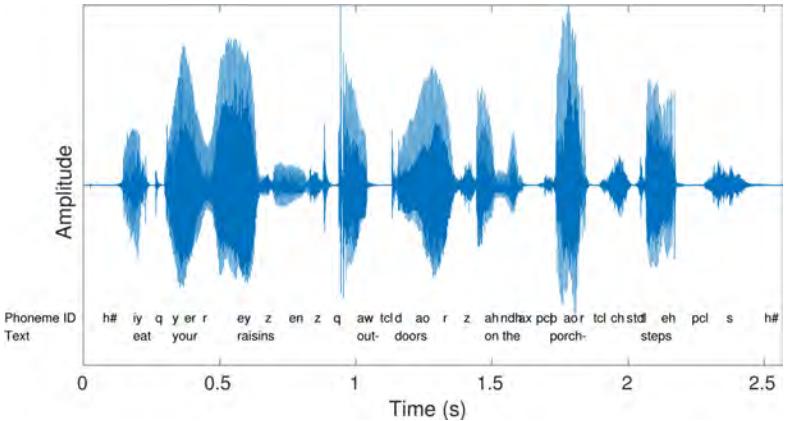


NLP task website: <https://nlpprogress.com/>



Major AI Application

- Speech
 - Speech recognition (speech-to-text)
 - Trigger word/wakeword detection
 - Speaker ID
 - Speech synthesis (text-to-speech, TTS)





Survey of Major AI Application Area

Speech



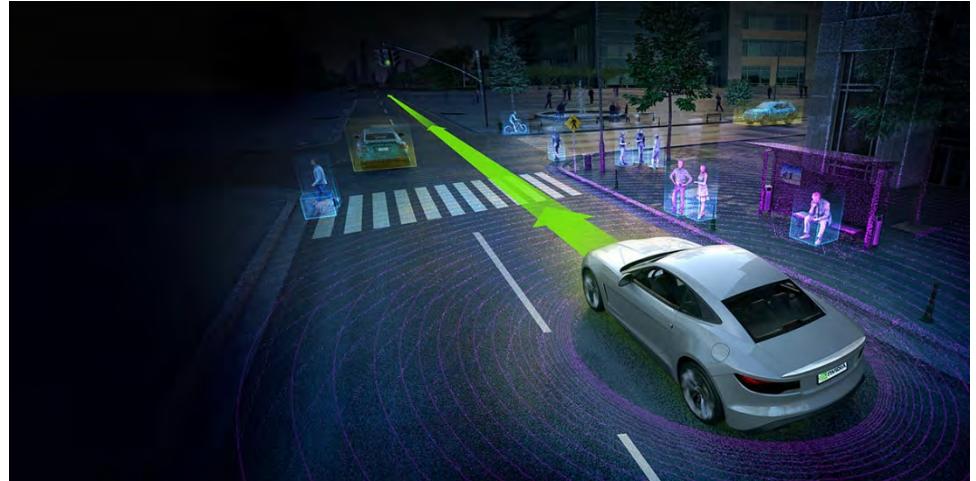
<https://www.youtube.com/watch?v=VDLs4u7l-Aw>

Optional



Major AI Application

- Robotics
 - Perception: Figuring out what's in the world around you
 - Motion planning: finding a path for the robot to follow
 - Control: sending commands to the motors to follow a path





Major AI Application

- General Machine Learning
 - Unstructured data (images, audio, text)
 - Structured data





Examples Of AI And ML In Practice



<https://www.youtube.com/watch?v=AXgjEW2nA9I>



AI/ML workflow



Amazon
Echo / Alexa



Google
Home



Apple
Siri



Baidu
DuerOS

“Hey device, tell me a joke”



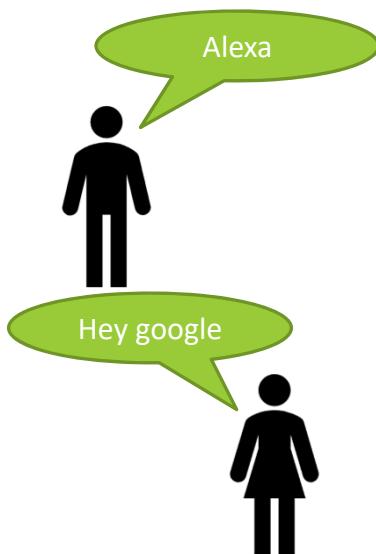


Smart Speaker

Collect
data

Train
model

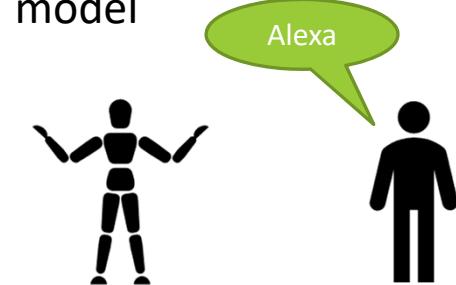
Deploy model



Iterate many times
until good enough

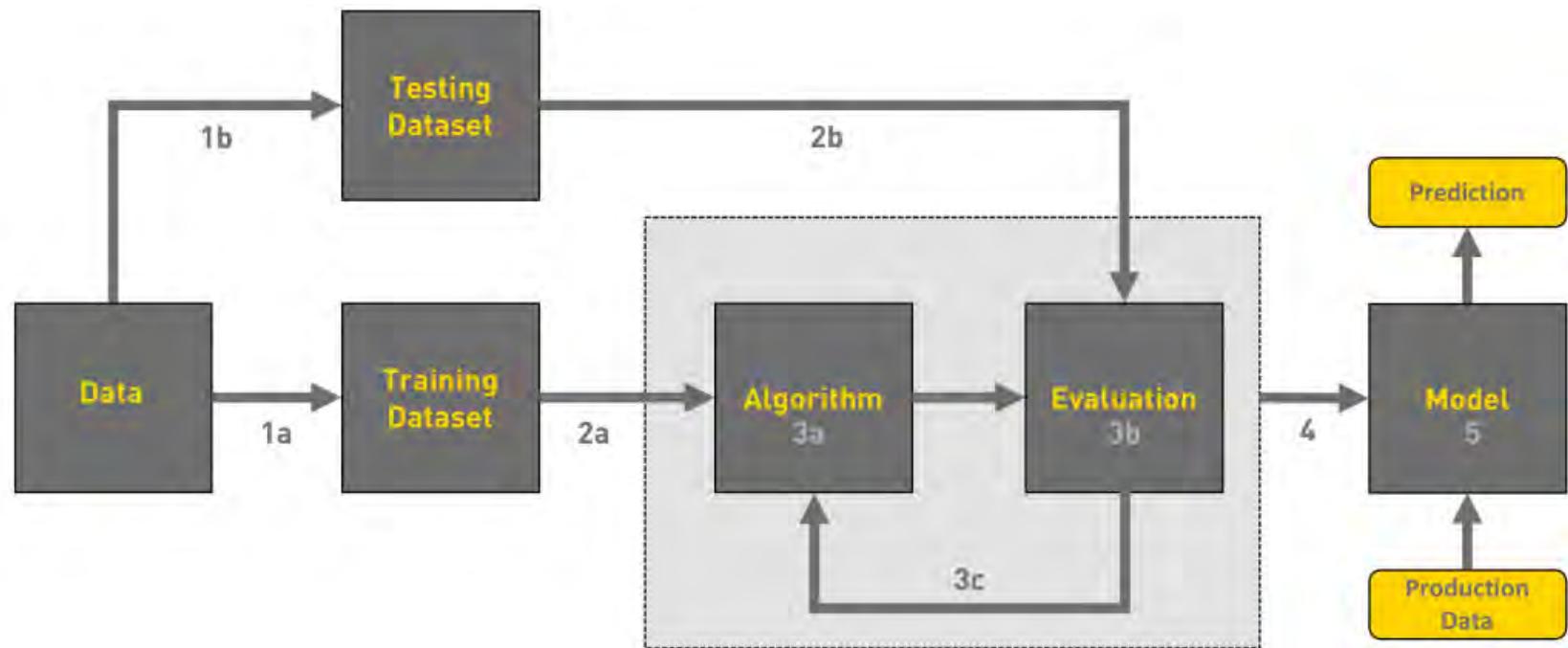
$A \rightarrow B$
audio #1 – “Alexa”
audio #2 – “hey google”

Get data back
Maintain/update
model



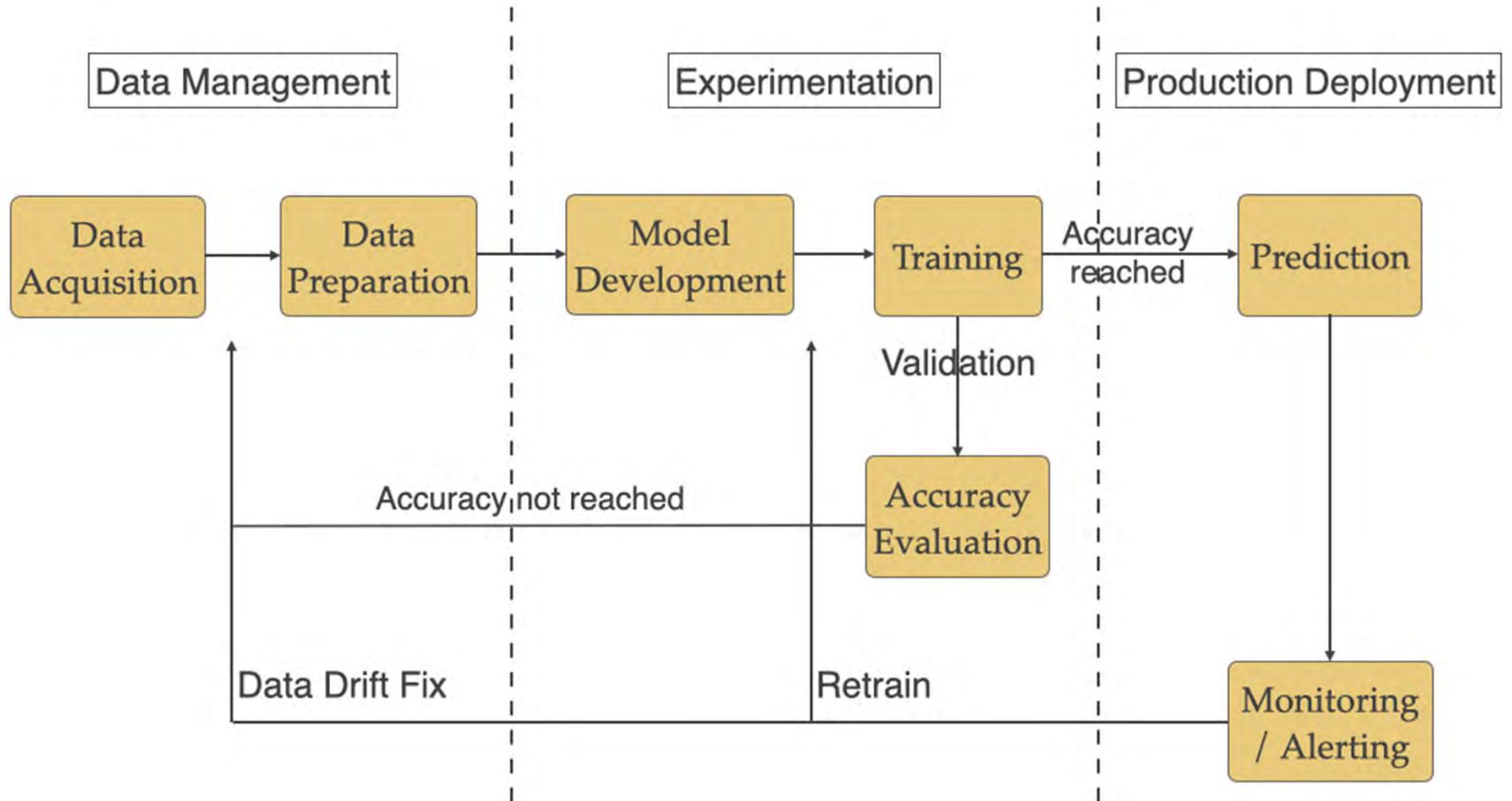


Machine Learning Workflow



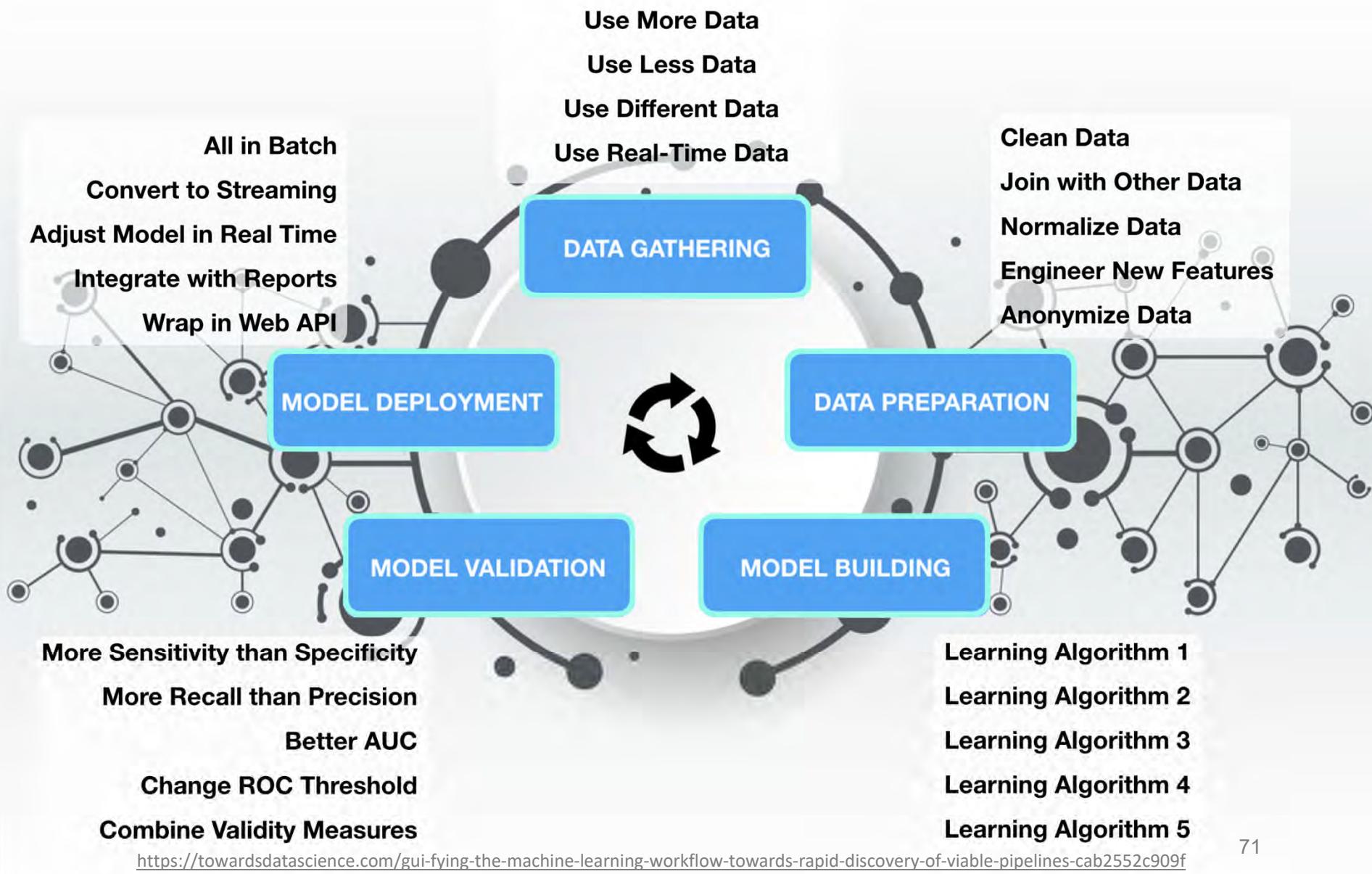


Machine Learning workflow





AI/ML Workflow





Tools for this workshop

- For this course, we will need the following tools:
 - Numpy
 - Pandas
- We will be focusing on the usage of the tools in this LU and how it can be used.
- The tools are used to input, manipulate and format the data.
- The tools can be used to perform mathematical functions on the data



Relationship between NumPy and Pandas

- Pandas like SciPy and Matplotlib are high level manipulation tools built on top of NumPy
- NumPy is a low level manipulation tool

Pandas

SciPy

Matplotlib

NumPy



What is NumPy?

- NumPy is the fundamental package for scientific computing
- Numpy is a Linear Algebra Library for Python
- It is very important to Data Science as almost all of the libraries in the python ecosystem rely on NumPy as one of the key development block
- NumPy is very fast as it has binding to C Libraries



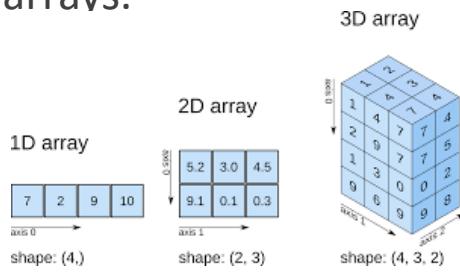
What is included in NumPy?

- Includes functionalities for
 - multidimensional arrays
 - high-level mathematical functions
 - linear algebra operations
 - Fourier transforms
 - pseudorandom number generator



NumPy

- N-dimensional array (**ndarray**) is an important object define in NumPy
- Collection of items of the **same type**
- Items in the collection accessed by **zero-based index**
- Ndarrays comes in **vectors** and **matrices**
 - Vectors: 1d arrays
 - Matrices 2d arrays.



Example of Vectors

[1, 2, 3]

Example of Matrices

$\begin{bmatrix} 3, & 4 \\ 5, & 6 \end{bmatrix}$

- Each item in an ndarray has the **same size memory block**



Python List Vs NumPy Array

- The advantage of NumPy (compared to Python List) are as follows:
 - **requires less memory**
 - **faster to process**
 - **requires less coding**



Usage of NumPy

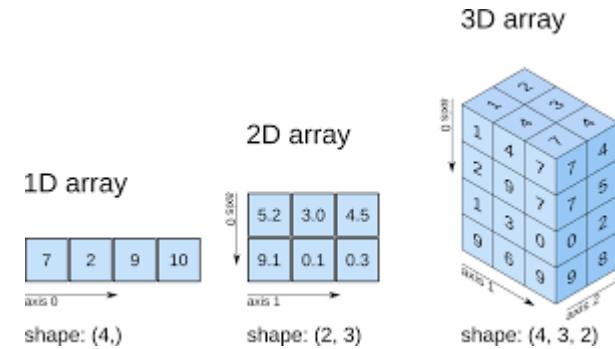
- NumPy arrays are the main way of storing data
- It comes in Vectors and Matrices
- Vectors are 1-D arrays, matrices are 2-D arrays

Example of Vectors

$$[1, 2, 3]$$

Example of Matrices

$$\begin{bmatrix} 3, & 4 \\ 5, & 6 \end{bmatrix}$$





Why we need Structured Data Types

- Define the fields(columns) needed
- Define the data type of the fields(columns) expected from the data
- Needed for reading text files and formatting to the required structure



Built-In Data Type

- NumPy has the following Built-In Data Types:

Data Type	Prefix	Numpy as np
Boolean	'b'	np.bool_
Integer(signed)	'l'	np.int
Integer(unsigned)	'u'	np.uint
Float	'f'	np.float
Date Time	'M'	np.datetime64
Unicode String	'U'	np.Unicode
String	'S'	np.str



NumPy - Structured data

1. Define the fields (columns) needed
2. Define the data type of the fields (columns) expected from the data

```
student=np.dtype([('name','S20'), ('age', 'i1'), ('marks', 'f4')])
```

3. Adding data to array using the student data type

```
s_array = np.array([('Mary', 21, 50),('John', 18, 75)], dtype=student)
```



NumPy – Accessing Structured data

1. Example:

```
student=np.dtype([('name','S20'), ('age', 'i1'), ('marks', 'f4')])
```

```
s_array = np.array([('Mary', 21, 50),('John', 18, 75)], dtype=student)
```

```
s_array['name'] = [b'Mary' b'John']
```

```
s_array[0]['name'] = b'Mary'
```

```
s_array[0]['age'] = 21
```



Activity 1.1 - Numpy

- Activity - Numpy 1.1 numpy
- Creating Numpy arrays
- Methods in Numpy
- Attributes of Array
- Array with Structured data
- Indexing, Slicing and Concatenating Array
- Reading a text file
- Operations
- Statistical Functions



Exercises:

- Create an 1-Dimension array of 36 numbers using arange(36) and store in narr1
- Change narr1 to 2-Dimensional Array (4 X 9) and store in narr2
- Change narr1 to 3-Dimensional array (3 X 3 X 4) and store in narr3
- Change narr1 array to 3 Dimensional array (2 X 3 X 3)?
- Advanced Indexing

Step 1:
Watch and listen to the
instructor's demonstration



15 mins

Step 2:
Work through the activities

Target to finish by 12:30



30 mins

Individual Activity



LUNCH BREAK



60 mins Lunch Break

Lunch break 12:30 - 13:30



Pandas

- A set of Python libraries for data wrangling and analysis
- Useful for early stages of data inspection, preprocessing and data cleaning
- It can work with data from a variety of sources
- It has excellent performance and built-in visualisation features



Save the Panda



Pandas and NumPy

- As mentioned Pandas is built on top of NumPy, with many features that are similar to NumPy
- Likewise, dtype defines the data type that are used in the various Pandas' various Data structure

This is the reason why NumPy is emphasized first



Pandas Data Structure and Dimensions

- Pandas has the following data structures
 - Series
 - Data Frames
 - Panel
- The data structure are built on NumPy
- Pandas reduces the complexities of handling two or higher-dimensional arrays



Overview of Pandas Data Structure and Dimensions

Data Structure	Dimension	Description
Series	1	1D labelled homogeneous array, size immutable. Data is mutable
Data Frames	2	General 2D labelled, data and size mutable tabular structure with potentially heterogeneously typed columns.
Panels	3	General 3D labelled, data and size mutable array.

- All data structures except Series are mutable (can be changed)
- Data Frame is the most common data structure used



Key Points of Series Data Structure

- One-dimensional labelled array
- Data needs to be **homogenous**
- The size is **immutable**
- Values of the data are **mutable**



Creation of Series Data Structures

- Series data structures can be created from the following:
 - ndarray
 - dict
 - Scalar



Key Points of Data Frame Data Structure

- Two Dimensional labelled respectively
- Data can be **heterogeneous**
- **Size and data is mutable**
- Able to perform Arithmetic operations on rows and columns

Columns

Student ID	Student Name	Student Marks	Student Grade
S001	Tom Hardy	78	B+
S002	Florence Chen	68	C+
S003	S W Koh	83	A
S004	June Loh	43	F

Rows



Creation of Data Frame Data Structures

- Data Frame can be created from the following inputs
 - Lists
 - Dict
 - Series
 - Ndarray
 - Another Data Frame



Row Selection and Slicing

- The row from the Dataframe can be retrieved by label or integer
- Use *loc* to retrieve by label
- Use *iloc* to retrieve by integer



Explanation of Adding Column Code

- A new column can easily added by assigning it to a label as shown in the code

```
df['three']=pd.Series([10,20,30],index=['a','b','c'])
```



The advantage of using the index (label) is to indicate the row of where the elements should be added. The index here is an optional parameter



Deletion of Column from Data Frame

- Using the drop() function
 - Delete a row / column
 - `d.drop([column1, column2], axis=1)` - delete columns
 - Eg. `d = d.drop(['Name', 'Age'], axis = 1)`
 - `d.drop([row1, row2], axis=0)` - delete rows
 - Example: `d = d.drop([1, 3, 5], axis = 0)`



Sorting the array

- There maybe instances where there is a need to sort the data
 - E.g. Analysing and training certain range of data.
- Two sorting options are available in Pandas:
 - By Label
 - By Actual Value



Sorting by Values

- The array can be sorted by the label(index) using `sort_value()`. It accept a 'by' argument to determine which column to sort
- The array can be sorted by multiple columns. For example:
 - `sort_value(by=['col1', 'col2'])`
 - the array will be sorted by col1 followed by col2



Pandas Descriptive Statistics

- Pandas like NumPy has statistical functions

Function	Description
count()	Number of Non Null observations(records)
sum()	Sum of Values
mean()	Mean of Values
median()	Median of values
mode	Mode of Values
std()	Standard Deviation of Values
max()/min()	Maximum and Minimum of values respectively
abs()	Absolute Value
prod()	Product of values
cumsum()/cumprod()	Cumulative sum and product of values respectively



Activity 1.2 – Pandas

Note: You need to start the Jupyter to perform ALL the exercises

- 1 - Creating a Series Data Structure
- 2 - Accessing Elements in a Series Data Structure
- 3 - Creating Data Frames
- 4 - Loading Data from csv, txt files
- 5 - Row Selection and Slicing
- 6 - Deleting Column from Data Frame
- 7 - Sorting
- 8 - Merging of Data Frames
- 9 - Descriptive Statistics

(1.2-Pandas)

Exercises:
- Complete exercises in colab

Target to finish by 14:35

Step 1:

Watch and listen to the instructor's demonstration



20 mins

Step 2:

Work through the activities

Individual Activity



20 mins



What is Data Visualisation?

Data visualisation is the presentation of **data** in a pictorial or graphical format.

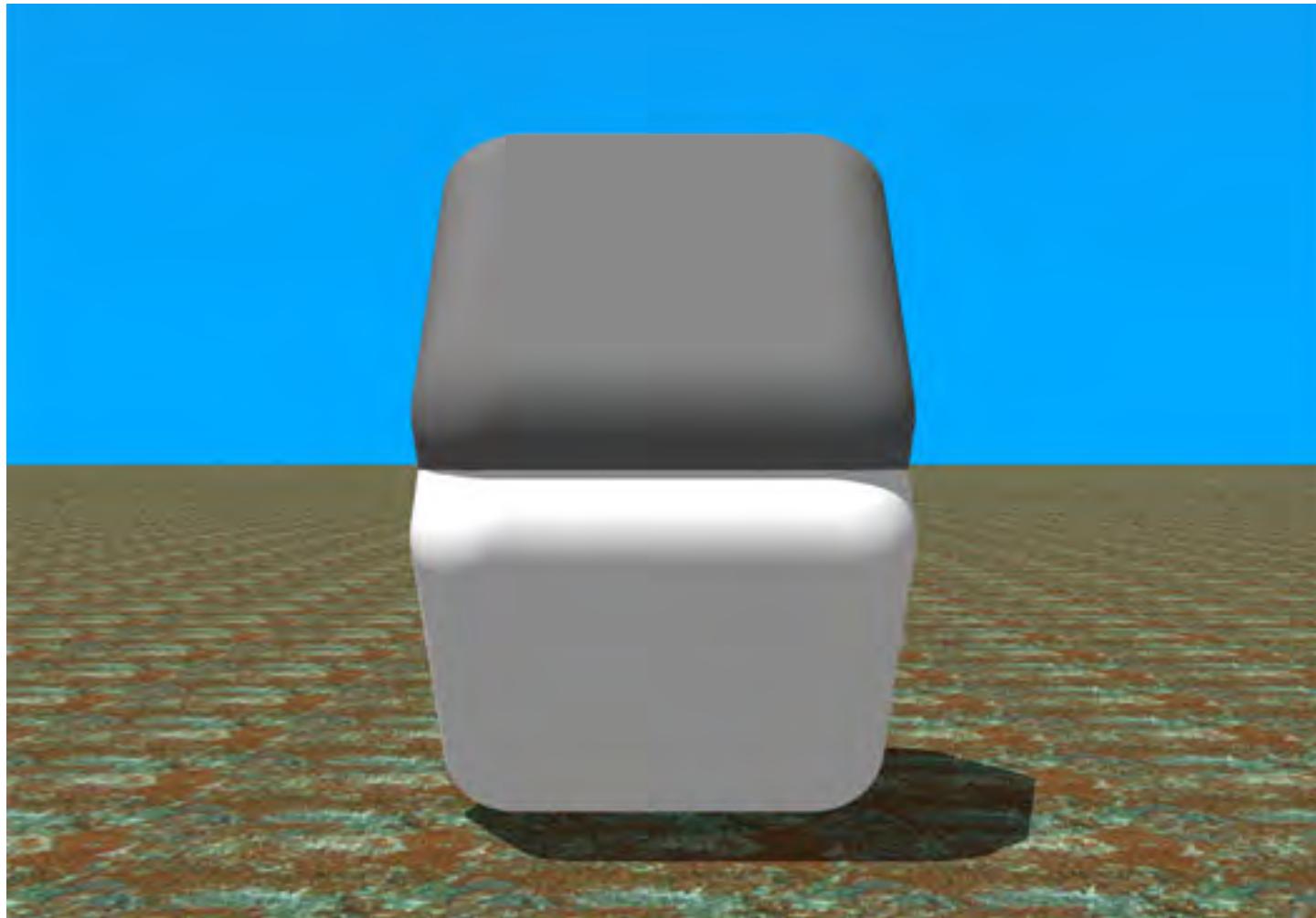
Anytime information is presented in visual format or concept, like a bar graph or pie chart, it is creating data visualisations.

But Data Visualisation is more than just charts and tables! It is an ability to tell a story, connecting one dataset to another!





Seeing is believing... ?



Source: Dr Beau Lotto (www.LottoLab.org) and Jeffrey A. Shaffer, University of Cincinnati



Goals of Data Visualization

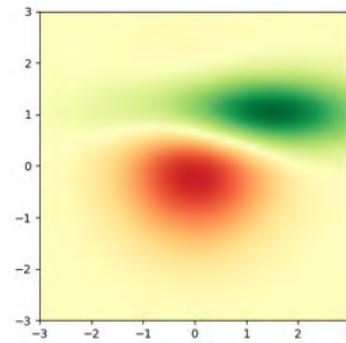
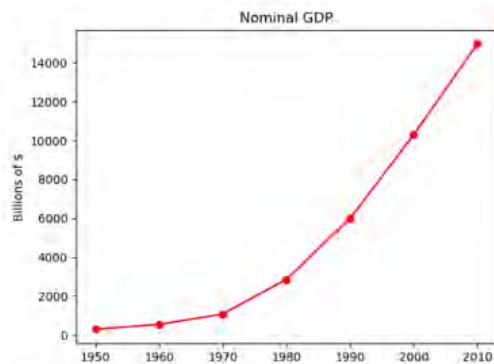
- **Explanatory** : Explain data to solve problems.
- **Exploratory** : Explore large data sets for better understanding.

- Data is growing at astronomical rate !
 - How to **manage and present** the data , without losing information, becomes a big challenge.
 - **Too much data** at our disposal could result in the mismanagement in the processing of the data for analysis (especially with inappropriate tools or methods)
- Human Minds work well with Visuals !
 - **Reading charts is easier** than going through pages of rows and columns of numbers



Matplotlib

- In Data Science, very often the data needs to be visualized in order to draw insights from the datasets.
- Basic examples (<http://matplotlib.org/examples/>)





Installation

- sudo apt-get install libpng12-dev
- sudo apt-get install python-dev
- sudo pip install matplotlib



Matplotlib – A simple plot

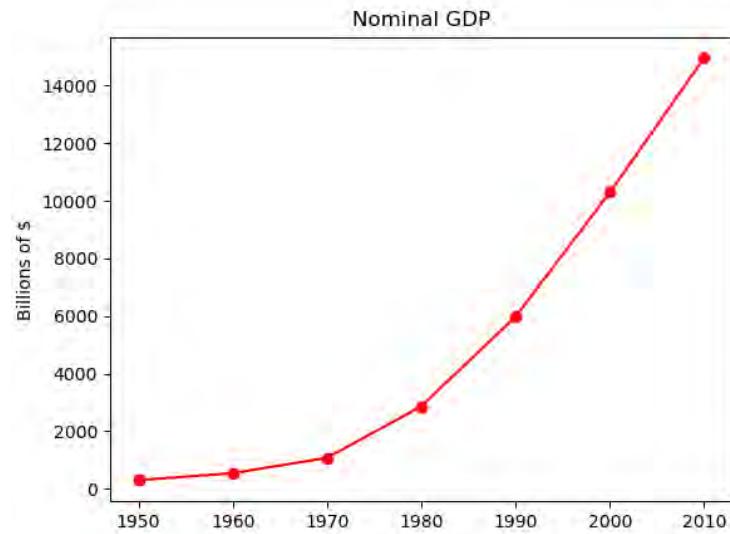
- Refer to 1.3-Matplotlib

```
from matplotlib import pyplot as plt
years = [1950, 1960, 1970, 1980, 1990, 2000, 2010]
gdp = [300.2, 543.3, 1075.9, 2862.5, 5979.6, 10289.7, 14958.3]

#create a line chart, years on x-axis, gdp on y-axis
plt.plot(years, gdp, color='red', marker='o', linestyle='solid')

#add a title
plt.title("Nominal GDP")

# add a label to the y-axis
plt.ylabel("Billions of $")
plt.show()
```





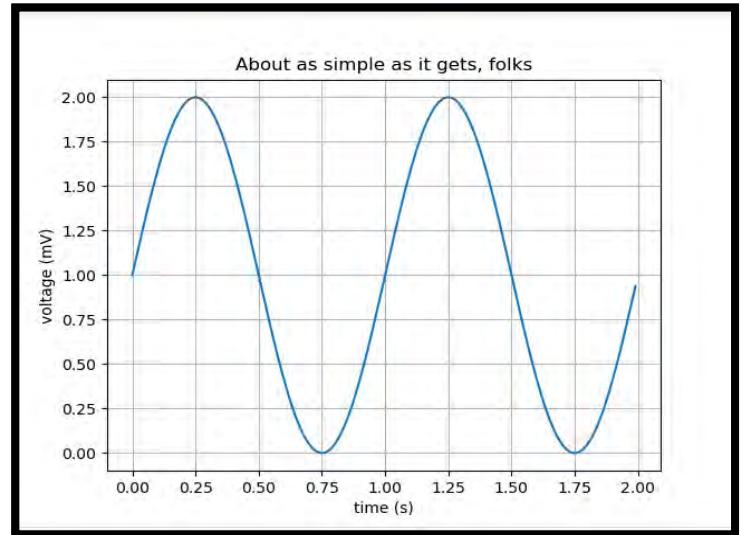
Matplotlib – A voltage plot

```
import matplotlib.pyplot as plt  
import numpy as np
```

```
# Data for plotting  
t = np.arange(0.0, 2.0, 0.01)  
s = 1 + np.sin(2 * np.pi * t)
```

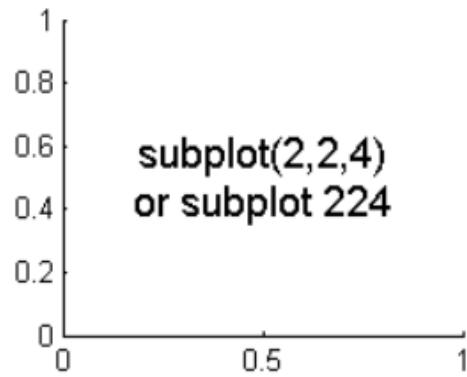
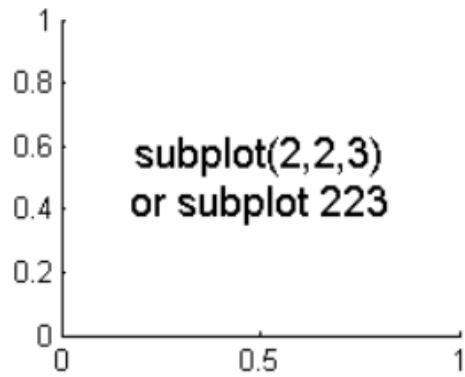
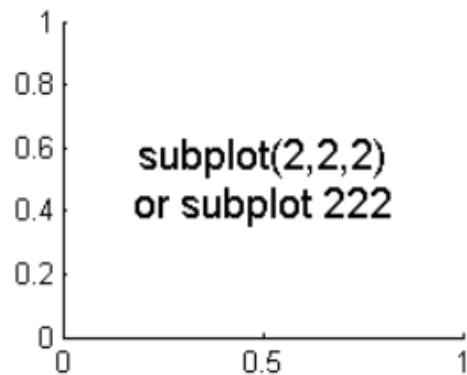
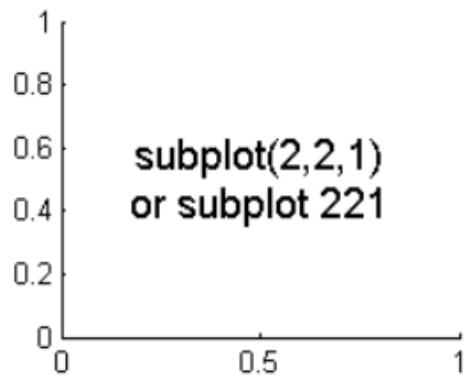
```
# Note that using plt.subplots below is equivalent to using  
# fig = plt.figure and then ax = fig.add_subplot(111)  
fig, ax = plt.subplots()  
ax.plot(t, s)
```

```
ax.set(xlabel='time (s)', ylabel='voltage (mV)',  
       title='About as simple as it gets, folks')  
ax.grid()  
  
fig.savefig("test.png")  
plt.show()
```





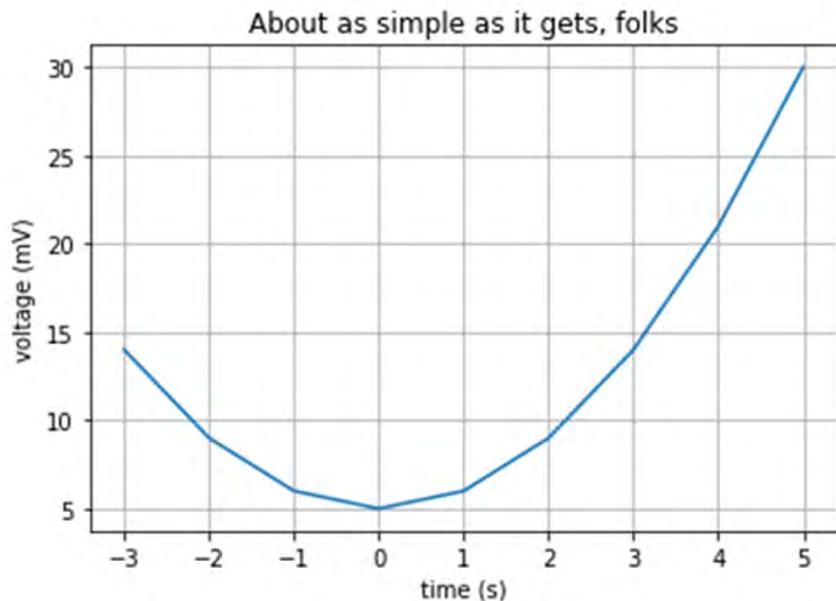
About...subplot(row, column, position)



```
import matplotlib.pyplot as plt  
  
fig = plt.figure()  
  
fig.add_subplot(221) #top left  
fig.add_subplot(222) #top right  
fig.add_subplot(223) #bottom left  
fig.add_subplot(224) #bottom right  
  
plt.show()  
  
# (111) = only one subplot or graph
```



Activity 1.3 – Matplotlib

**Step 1:**

Watch and listen to the instructor's demonstration



15 mins

Step 2:

Work through the activities

Individual Activity

15 mins

A graphic element consisting of a white circle with a blue double-line border. Inside the circle is a black silhouette of a teacup with steam rising from it. A small white rectangle containing the word "Tea" is positioned at the bottom left of the cup. The circle is set against a dark blue background that slopes upwards from left to right.

15 Mins Break

Back by 15:20



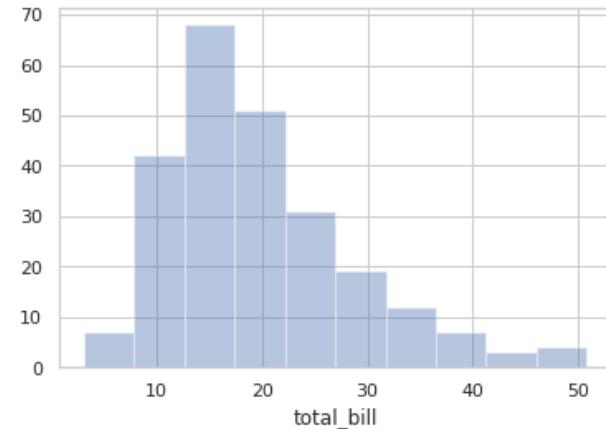
Seaborn

- Python data visualisation library built on top of matplotlib.
- Provides high-level commands, convenient way to create a variety of plot types
- Used typically for statistical data exploration and statistical model fitting.
- Useful and easy options for univariate and bivariate visualization and for comparing data
- Ref: <https://github.com/mwaskom/seaborn> - documentation and examples

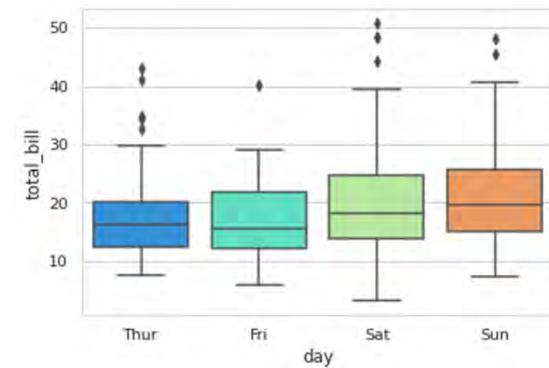


Seaborn

- **Distribution Plots** - to visualise distribution of a data set
 - distplot(...)
 - jointplot(...)
 - pairplot(...)



- **Categorical Plots** - to visualise data where one or more variables is “categorical” (divided into discrete groups).
 - countplot(...)
 - barplot(...)
 - boxplot(...)

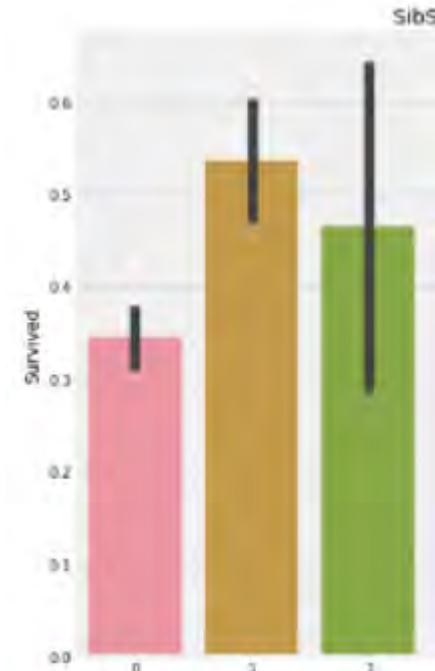




Error bar in barplot

- **Error bar**

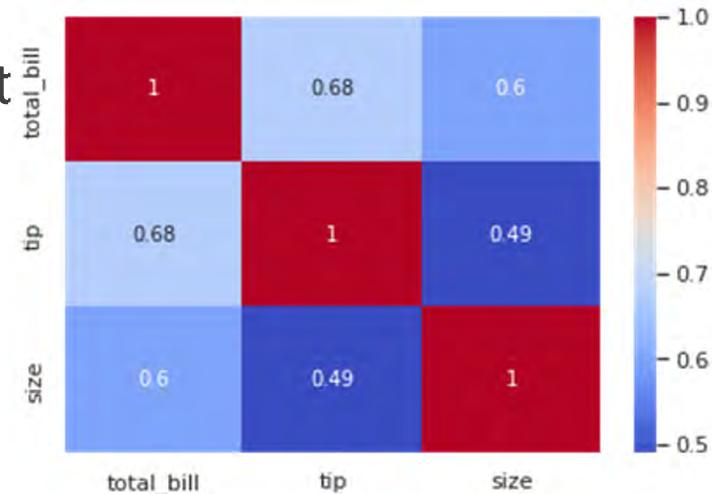
- The error bar is the confidence interval for the variable
- It is the interval where the 95 % of your variable lies in.
- How spread the data are around the mean value (small SD bar = low spread, data are clumped around the mean; larger SD bar = larger spread, data are more variable from the mean).



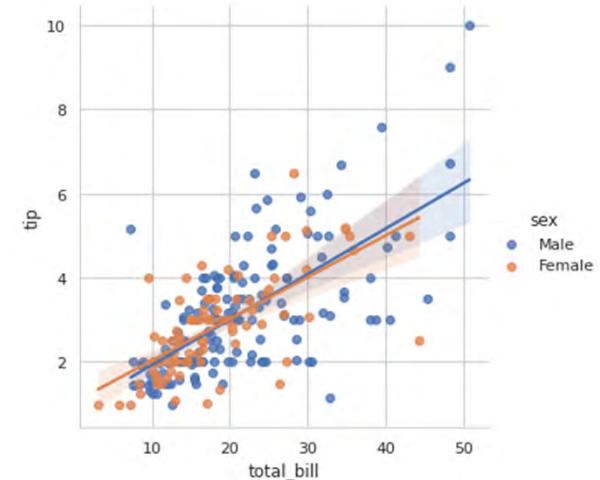


Seaborn

- **Matrix Plots** - to visualise distribution of a matrix data set
 - heatmap(...)
 - clustermap(...)



- **Regression Plots** - to visualise linear regression model or their data and features.
 - lmplot(...)





Activity 1.4 – Seaborn

1 to 5 of 5 entries								Filter	?
index	total_bill	tip	sex	smoker	day	time	size		
0	16.99	1.01	Female	No	Sun	Dinner	2		
1	10.34	1.66	Male	No	Sun	Dinner	3		
2	21.01	3.5	Male	No	Sun	Dinner	3		
3	23.68	3.31	Male	No	Sun	Dinner	2		
4	24.59	3.61	Female	No	Sun	Dinner	4		

Show 25 ▾ per page

Like what you see? Visit the [data table notebook](#) to learn more about interactive tables.

Step 1:

Watch and listen to the instructor's demonstration



15 mins

Step 2:

Work through the activities

Individual Activity

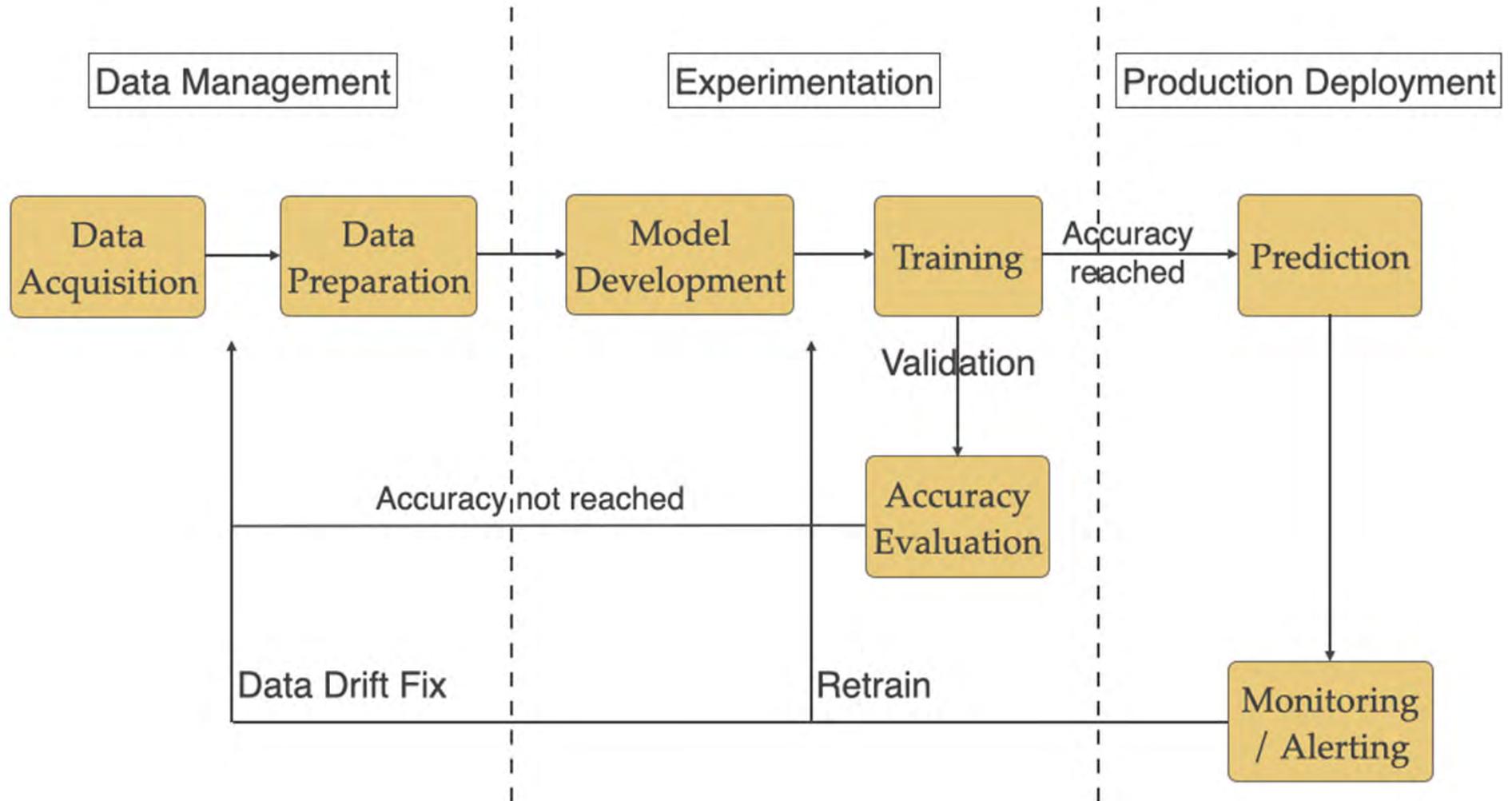


15 mins

Data Preparation



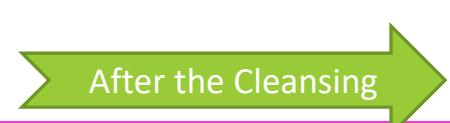
Machine Learning workflow





Data Preparation

- Data can be gathered from many sources and the quality of the data may not be 'clean'
- Data preparation involves the following
 - Data Cleaning
 - Segmentation of Data for Training, Validation and Testing
- Data scientist will be spending most of the time in data preparation

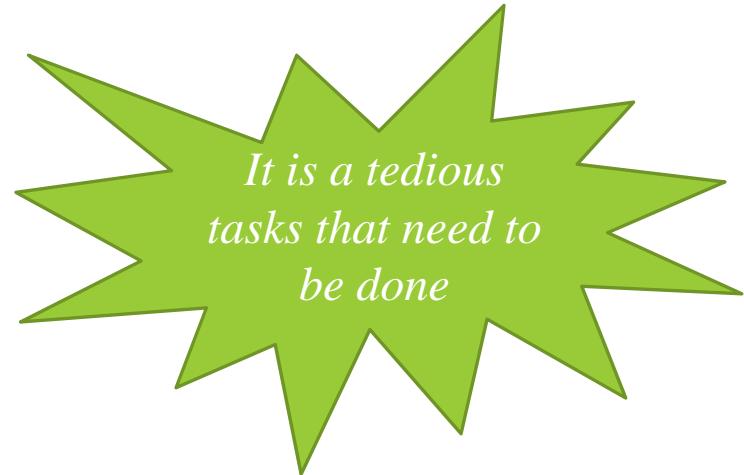


Similarity to data cleansing, we need to clean up the data as the situation when it is gathered may not ideal



Data Cleaning

- Data that are gathered are usually far from ideal as there may be:
 - Incorrect Data Format (e.g. Date should be ivy/mmm/did)
 - Missing Data
 - Incorrect Data Type (e.g. Data is numeric but a string value is given)
 - Data Duplication
 - Etc.





WHY IS IT IMPORTANT!

- Poor data could lead to wrong classification
- When there is wrong classification of data, it will lead to wrong predication and outcome.
- Wrong predication and outcome will be 'costly' to the organization

For example, the Land Transport Authority may like to investigate the public transport usage. If the data is not clean, LTA may not have the wrong conclusion and prediction.



Data Cleaning

- In Data Science (e.g. Data Analytics, AI), it will require large amount of data and manually cleaning of these data will be **impossible**
- Tools like Python will be able to **automate** the process of cleaning the data based on 'rules'
- The data will be manipulated based the **rules** set in the programmed



Explanation to the Data Cleaning Programme

```

# Scenario 1
# This programme is to clean the data from the input file
import pandas as pd
import numpy as np

#Example
df = pd.read_csv('Data/dirty_data.csv')

#Create a duplicate for df
df_Ex1= df.copy()

# Rule 1
# This sets of routines will null all non-numeric values from the individual columns
df_Ex1['floor_size'] = pd.to_numeric(df_Ex1['floor_size'], errors="coerce")

# Scenario 1 Exercise 1: Add the codes required to null all non-numeric values
# ..... COMPLETE CODE
#-----


# Computer the means
df_mean=df_Ex1.mean()
df.loc(row, column)

# Rule 2
# This set of routines are to convert null value to the mean value of the columns
df_Ex1.loc[df_Ex1['floor_size'].isnull(),'floor_size'] = df_mean['floor_size']

# Scenario 1 Exercise 2: Add the codes required to handle missing values
# ..... COMPLETE CODE
#-----


#Save to CSV file
df_Ex1.to_csv('Data/clean_data_scenario1.csv')

```

Creating a duplicate of the dataframe df
Changes made to df_Ex1 will not change the values of df1

This step will nullify ALL non-numeric or blank values for a particular column

Remember the loc function

Boolean function that will return the index of row that meet the condition

Save DataFrame to a csv file



Explanation to the Data Cleaning Programme

```
# Scenario 2
# This programme is to clean the data from the input file
import pandas as pd
import numpy as np

#Example
df = pd.read_csv('Data/dirty_data.csv')

#Create a duplicate for dataframe df
df_Ex2 = df.copy()

df_Ex2a = df_Ex2.drop(df_Ex2[df_Ex2['median_income'].str.isalpha()== True].index)
#Complete the code
# Add the required codes for the rest of the columns
```

Drop the rows from the
DataFrame

Function that check
whether the value is
characters



Writing the Data Cleaning Programme

Scenario 1

- Rules for cleaning the data
 - Ensure that all non-numeric characters are nullified
 - Ensure that all nullified values are given the mean values of the respective columns

Scenario 2

- Rules for cleaning the data
 - Remove rows where there are non-numeric value
 - Ensure that all nullified values are given the mean values of the respective columns





Data Preparation for building the model

- Machine Learning will require a model to perform prediction
- The model used are usually statistical algorithms that will perform the ‘prediction’
- Prediction are based on the probability of the data point that matches certain criterion





The last step in Data Preparation

- After the data are ‘cleaned’, it is important that the data split into the following datasets
 - Training
 - Validation
 - Test

Note in some case, validation is known as the test and test is known as the holding datasets

- It is important to note that the data are randomly split into the respective datasets (Training, Validation and Test)



Purpose of the Respective Datasets

- **Training**
 - Training datasets are used for training the models. For each data that passed through the model, it will 'refine' the model.
- **Validation**
 - Validation datasets are used for the checking the model.
- **Testing**
 - Test datasets plays the similar role as the validation datasets but it is to ensure that the model best fit the hypothesis

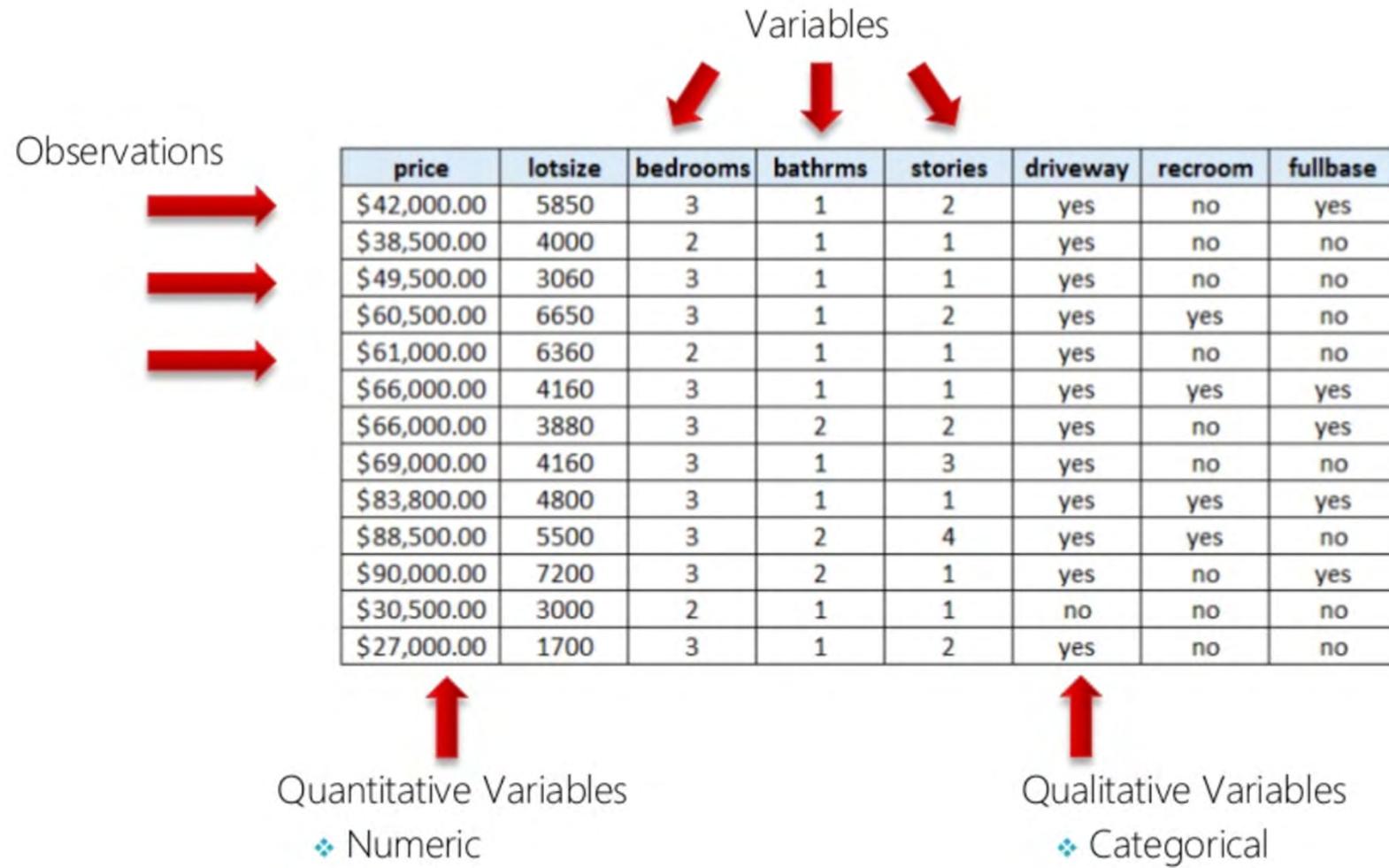
Normal Ratio of splitting the data:

Training :70% Validation 20% Testing : 10%

Note this ratio is able to change based on requirements



Features of a Data Set



Ref: Derek Kane, Data Science – EDA & Model Selection



Features of a Data Set

Dependent Variable

Independent Variables



price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
\$42,000.00	5850	3	1	2	yes	no	yes
\$38,500.00	4000	2	1	1	yes	no	no
\$49,500.00	3060	3	1	1	yes	no	no
\$60,500.00	6650	3	1	2	yes	yes	no
\$61,000.00	6360	2	1	1	yes	no	no
\$66,000.00	4160	3	1	1	yes	yes	yes
\$66,000.00	3880	3	2	2	yes	no	yes
\$69,000.00	4160	3	1	3	yes	no	no
\$83,800.00	4800	3	1	1	yes	yes	yes
\$88,500.00	5500	3	2	4	yes	yes	no
\$90,000.00	7200	3	2	1	yes	no	yes
\$30,500.00	3000	2	1	1	no	no	no
\$27,000.00	1700	3	1	2	yes	no	no

The dependent variable is the variable that we are interested in predicting and the independent variables are the variables which may or may not help to predict the dependent variable.



Splitting the Dataset

- An important step in predictive model building involves splitting the dataset into a training and testing dataset. The training dataset is used for model building while the testing dataset is for validating the model.
- The purpose is to ensure that the model are validated against data that are not part of model construction.
- To do this, we randomly select observations for training dataset (60%) and the remaining into the test dataset(40%)

Training
Dataset

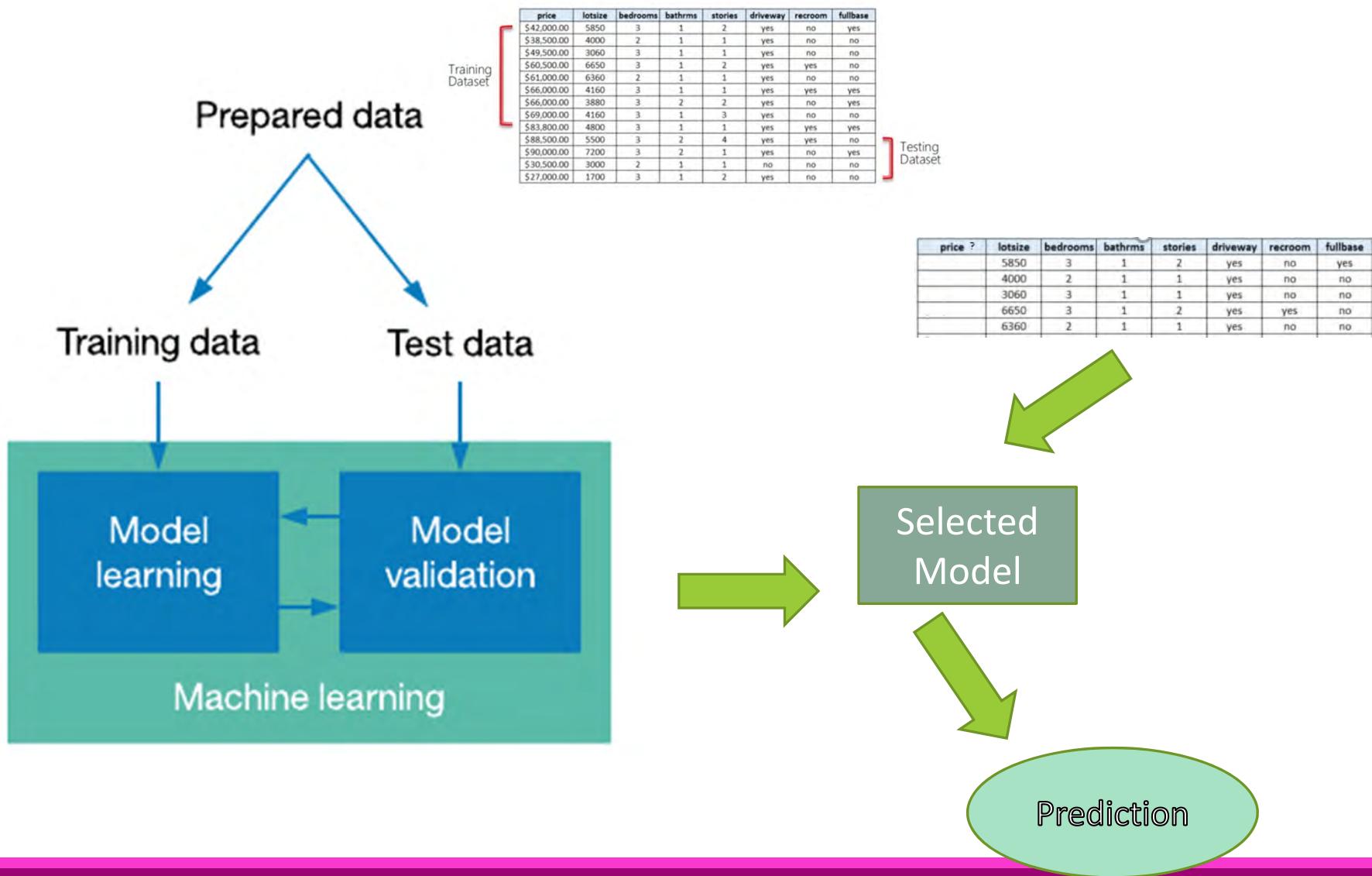
price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase
\$42,000.00	5850	3	1	2	yes	no	yes
\$38,500.00	4000	2	1	1	yes	no	no
\$49,500.00	3060	3	1	1	yes	no	no
\$60,500.00	6650	3	1	2	yes	yes	no
\$61,000.00	6360	2	1	1	yes	no	no
\$66,000.00	4160	3	1	1	yes	yes	yes
\$66,000.00	3880	3	2	2	yes	no	yes
\$69,000.00	4160	3	1	3	yes	no	no
\$83,800.00	4800	3	1	1	yes	yes	yes
\$88,500.00	5500	3	2	4	yes	yes	no
\$90,000.00	7200	3	2	1	yes	no	yes
\$30,500.00	3000	2	1	1	no	no	no
\$27,000.00	1700	3	1	2	yes	no	no

Testing
Dataset

Ref: Derek Kane, Data Science – EDA & Model Selection



Training, Testing and Prediction





Example of Data Splitting Programme

- Ratio of Training : Validation: Test is 70:20:10

```
df = pd.read_csv('..../Data/out/clean_data_scenario1.csv')

#Create a duplicate for dataframe df
df_Example= df.copy()

train, validate, test = np.split(df_Example.sample(frac=1), [int(.7*len(df_Example)), int(.9*len(df_Example))])
```

Split location in the array



Activity 1.5 – Data Preparation

- 3 cleaning exercises
 - non-numeric characters are set to null
 - non-numeric characters are dropped
 - Splitting the data for training and testing.



Step 1:
Watch and listen to the
instructor's demonstration



15 mins

Target to finish by 16:50am

Step 2:
Work through the activities



30 mins

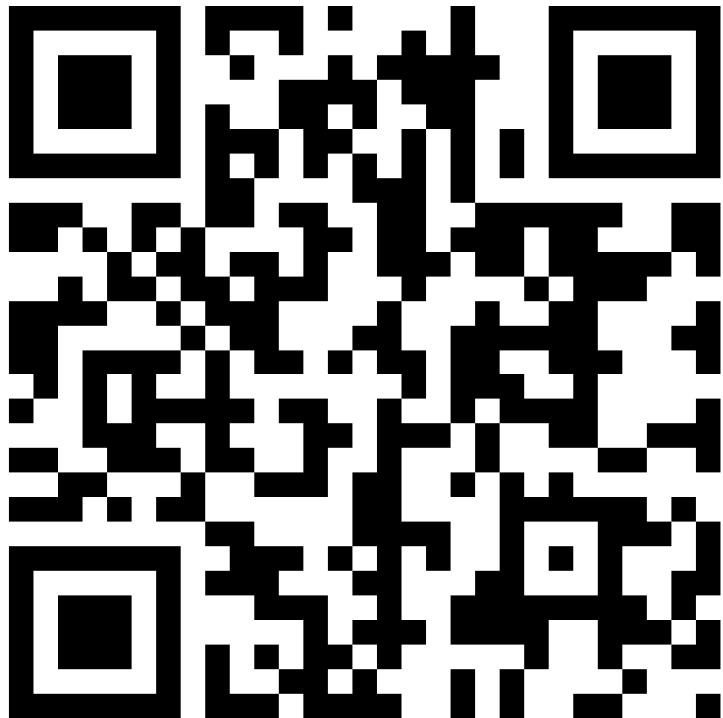
Individual Activity



Debrief

Step 1: Go to the following url

http://bit.ly/dlwp_debrief



Step 2: facilitator will walk you through the following



5 mins



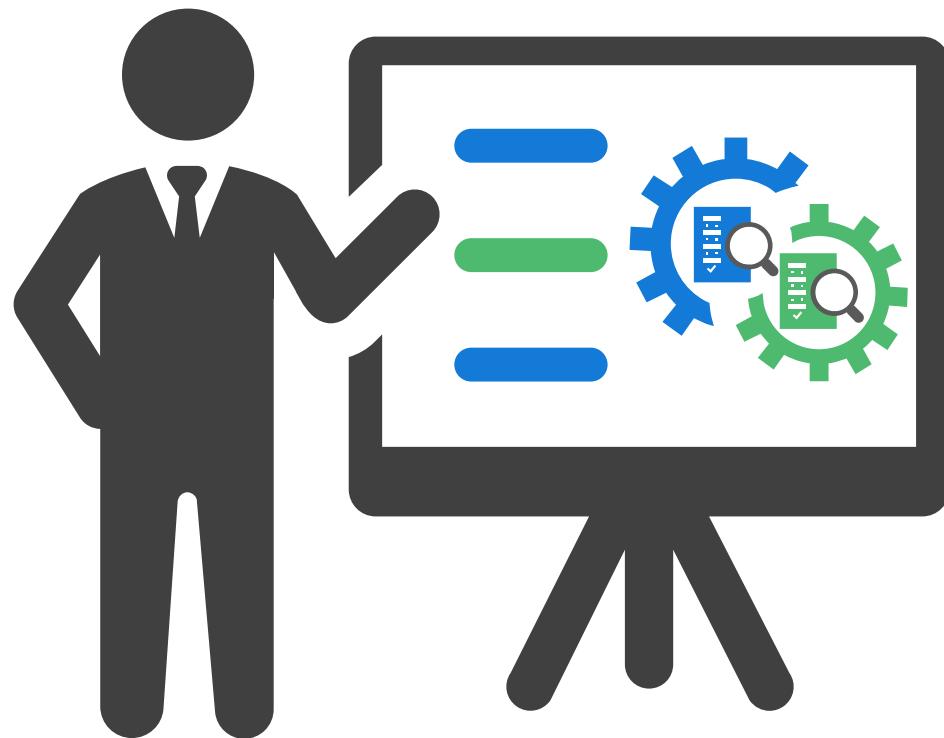
Quiz

https://bit.ly/kw_poll





Q&A



Email
seow_khee_wei@rp.edu.sg

Telegram
[@kwseow](https://t.me/kwseow)

Source code:

135



Thank you