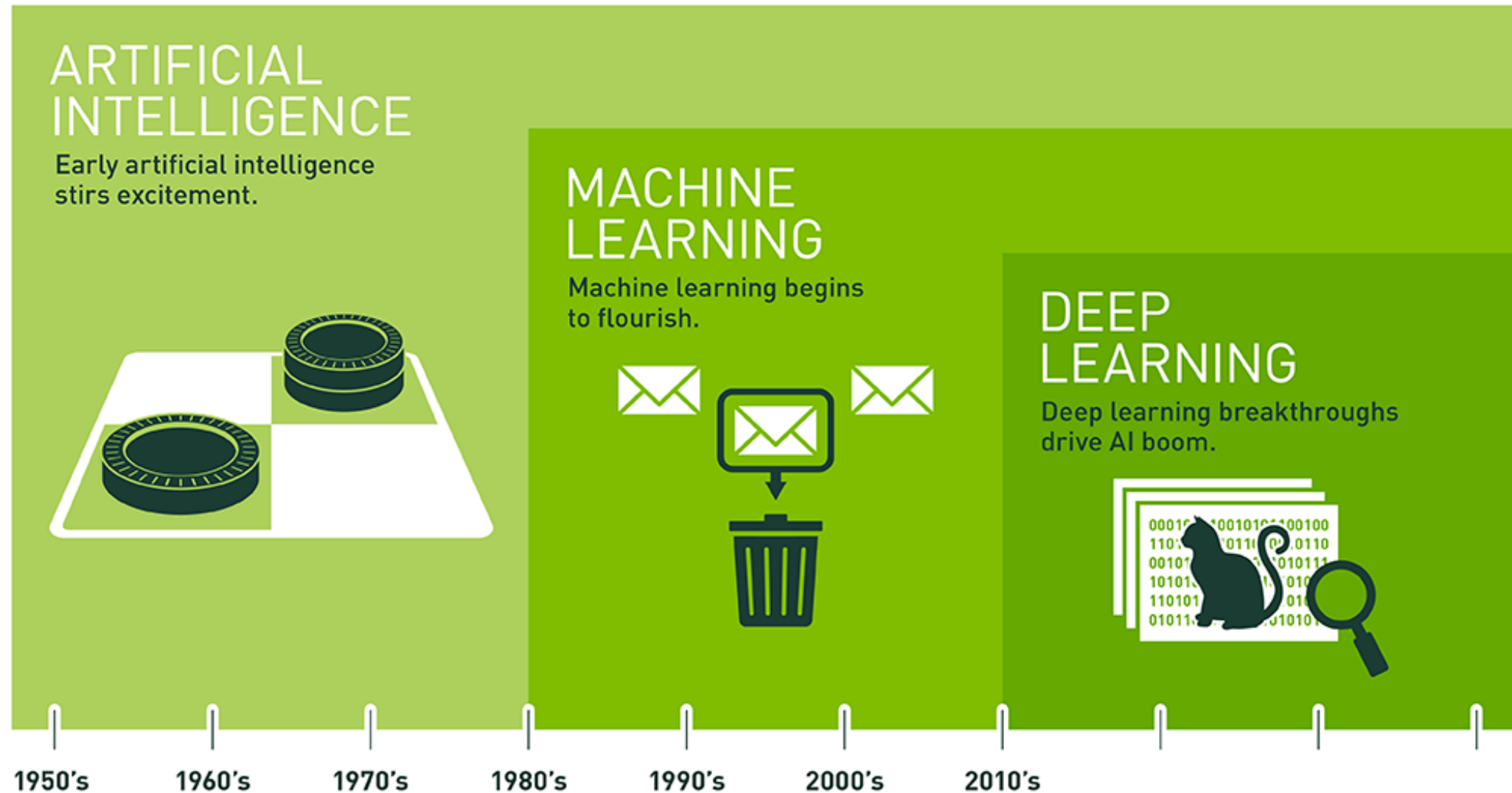
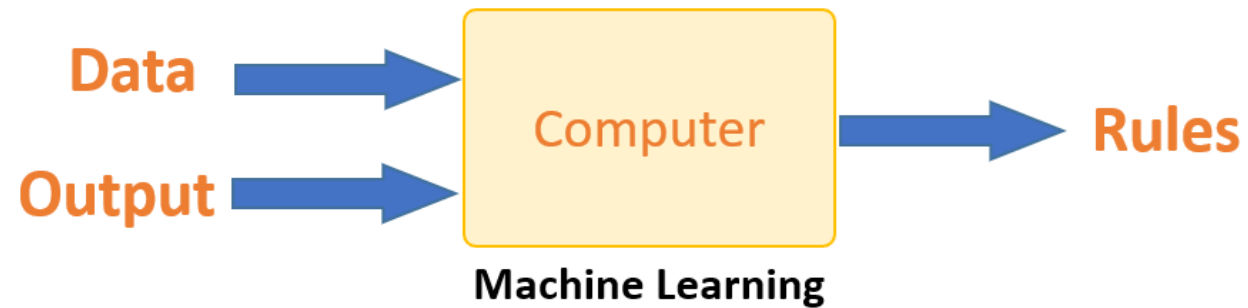
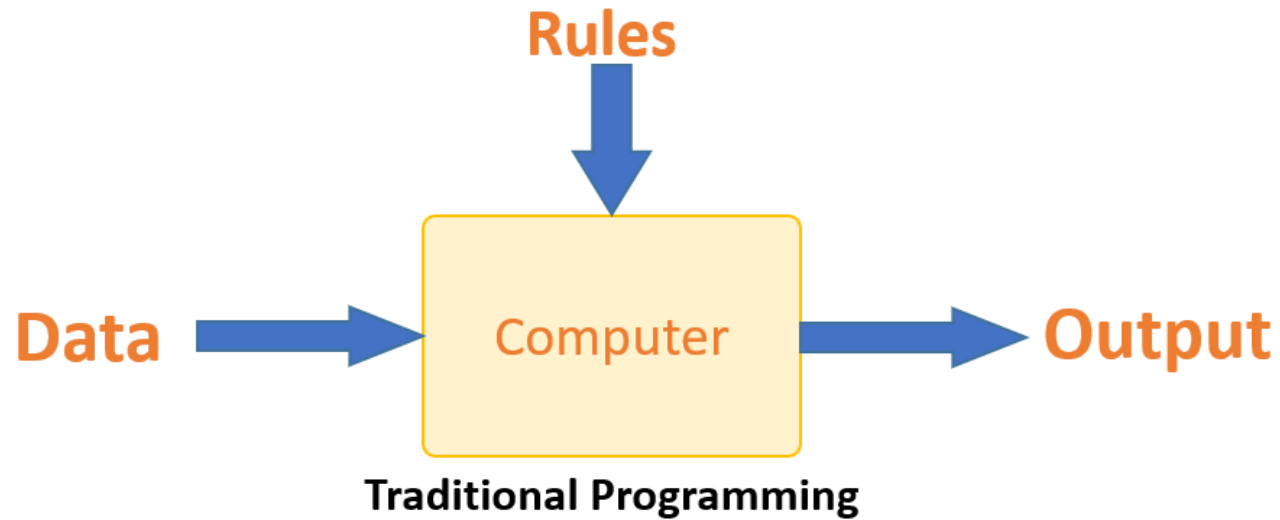


AI, ML and DL

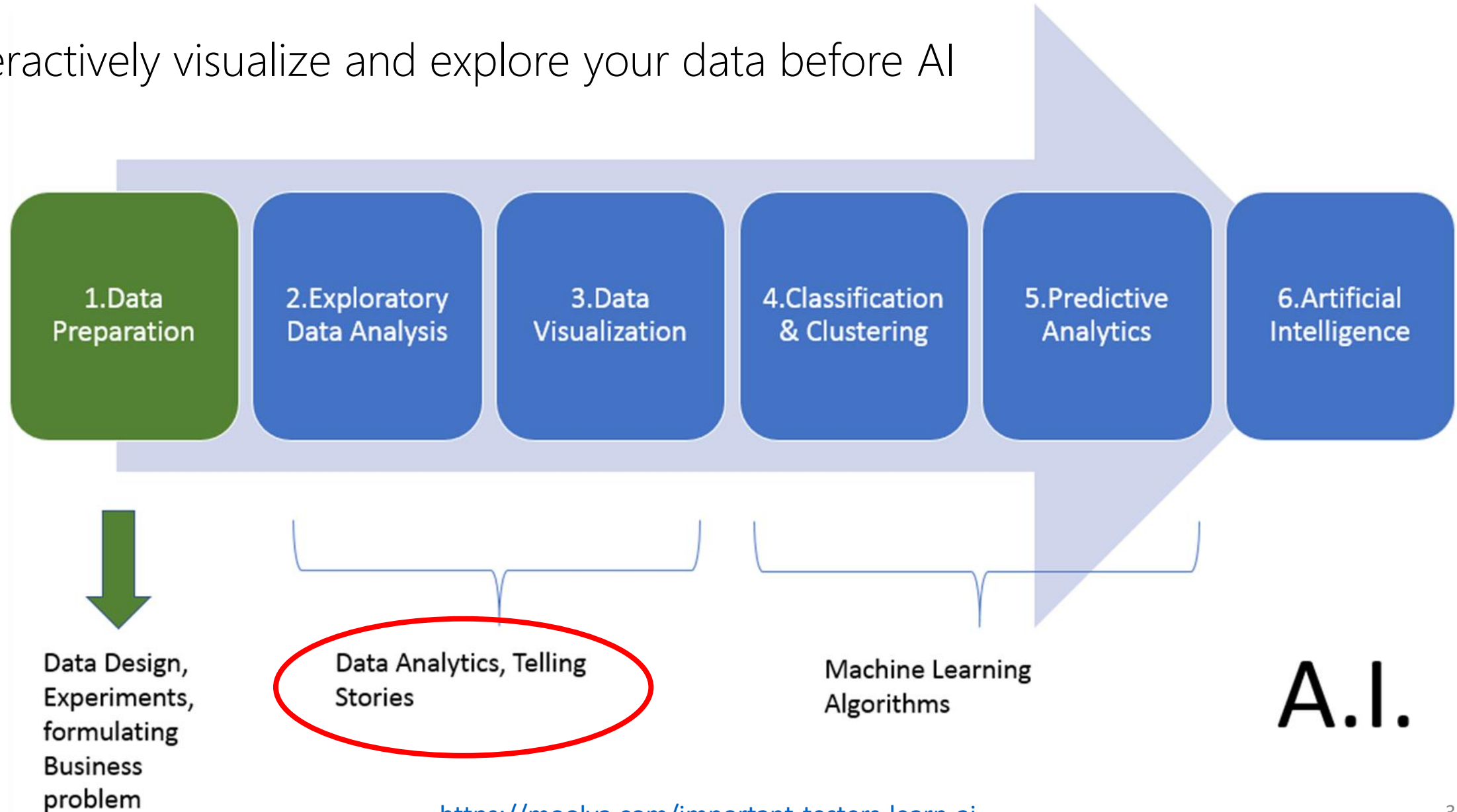


Traditional programming vs machine learning

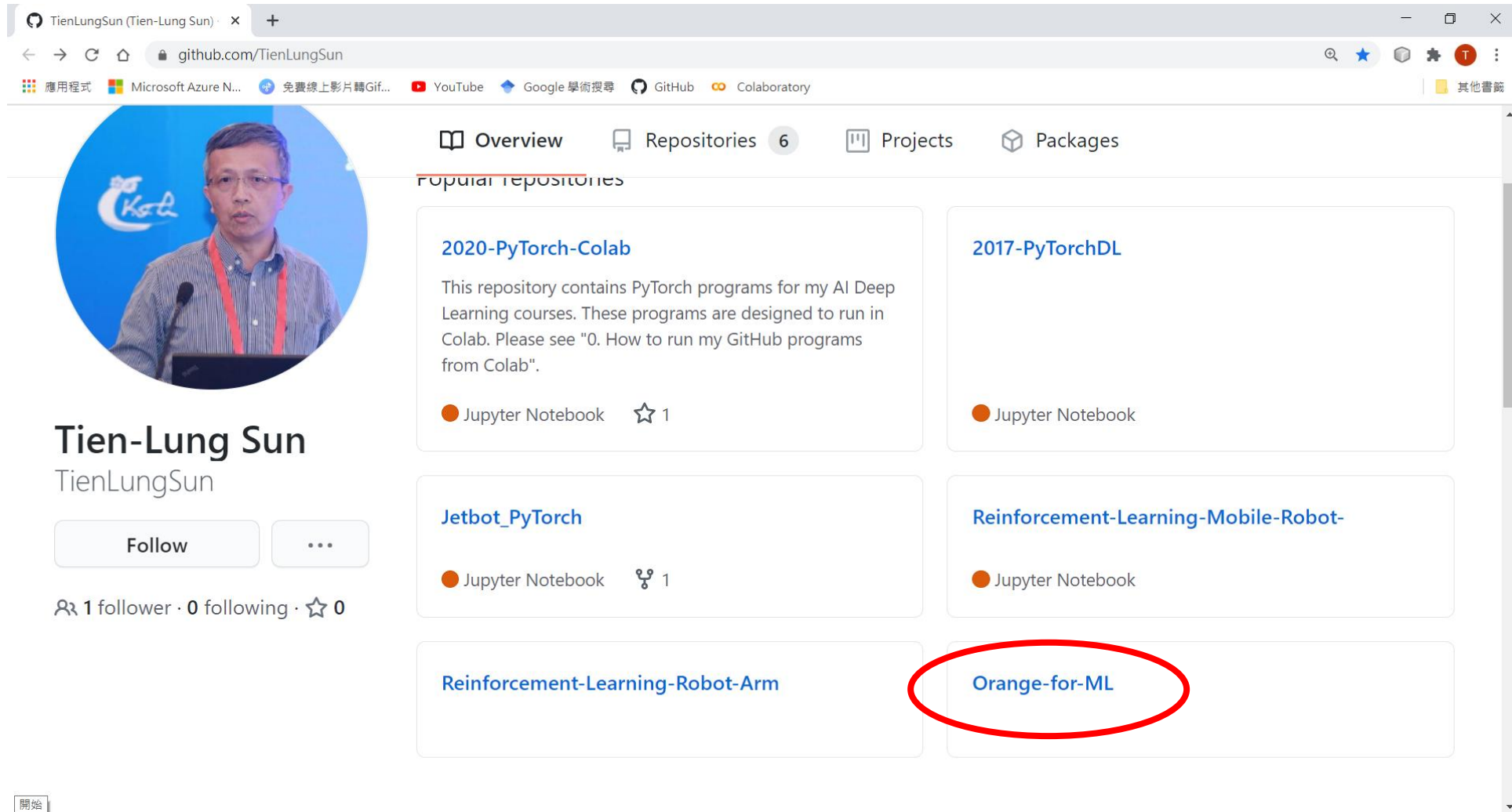


HI before AI

Interactively visualize and explore your data before AI



Download data files from my GitHub



The screenshot shows a web browser window displaying the GitHub profile of TienLungSun. The browser's address bar shows the URL `github.com/TienLungSun`. The profile page includes a circular profile picture of a man with glasses, the name **Tien-Lung Sun**, and the username `TienLungSun`. Below the name is a `Follow` button and a menu icon. The statistics show `1 follower · 0 following · 0 stars`. The `Repositories` tab is selected, showing a list of six repositories. The repository `Orange-for-ML` is circled in red. The other repositories are `2020-PyTorch-Colab`, `2017-PyTorchDL`, `Jetbot_PyTorch`, `Reinforcement-Learning-Mobile-Robot-`, and `Reinforcement-Learning-Robot-Arm`. Each repository card includes a description, a file type icon (Jupyter Notebook), and a star count.

Repository Name	Description	File Type	Stars
<code>2020-PyTorch-Colab</code>	This repository contains PyTorch programs for my AI Deep Learning courses. These programs are designed to run in Colab. Please see "0. How to run my GitHub programs from Colab".	Jupyter Notebook	1
<code>2017-PyTorchDL</code>		Jupyter Notebook	
<code>Jetbot_PyTorch</code>		Jupyter Notebook	1
<code>Reinforcement-Learning-Mobile-Robot-</code>		Jupyter Notebook	
<code>Reinforcement-Learning-Robot-Arm</code>			
<code>Orange-for-ML</code>			

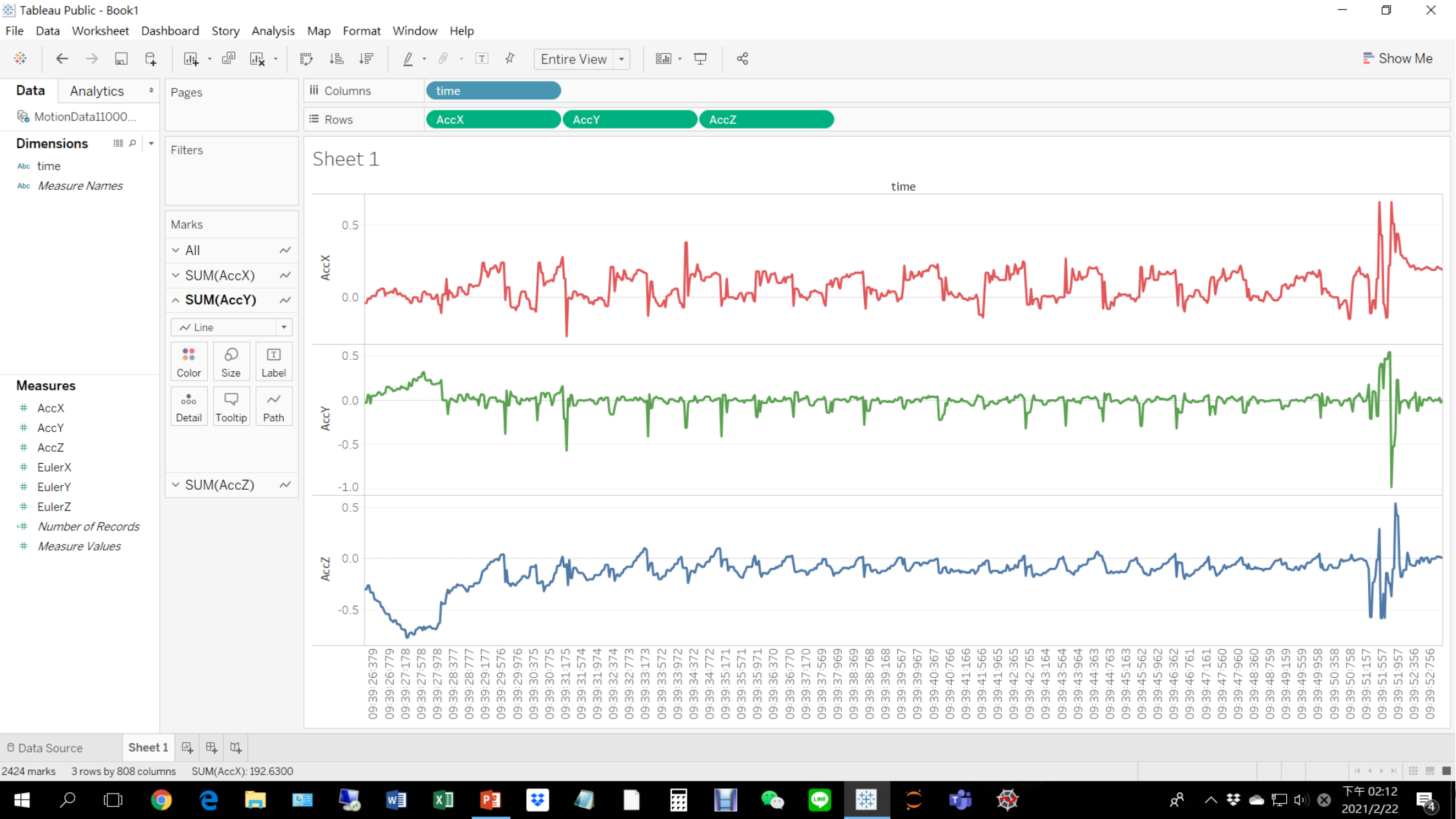
<https://github.com/TienLungSun>

Interactively visual exploration – (1) Tableau Public



<https://public.tableau.com/s/>

Tableau Public



Practice – Tableau public

1. Download and install Tableau Public
2. Visualize the motion data file
3. Search Kaggle (<https://www.kaggle.com/>) to find a sensor data file (see next slide)
4. Use Tableau public to visualize the data file

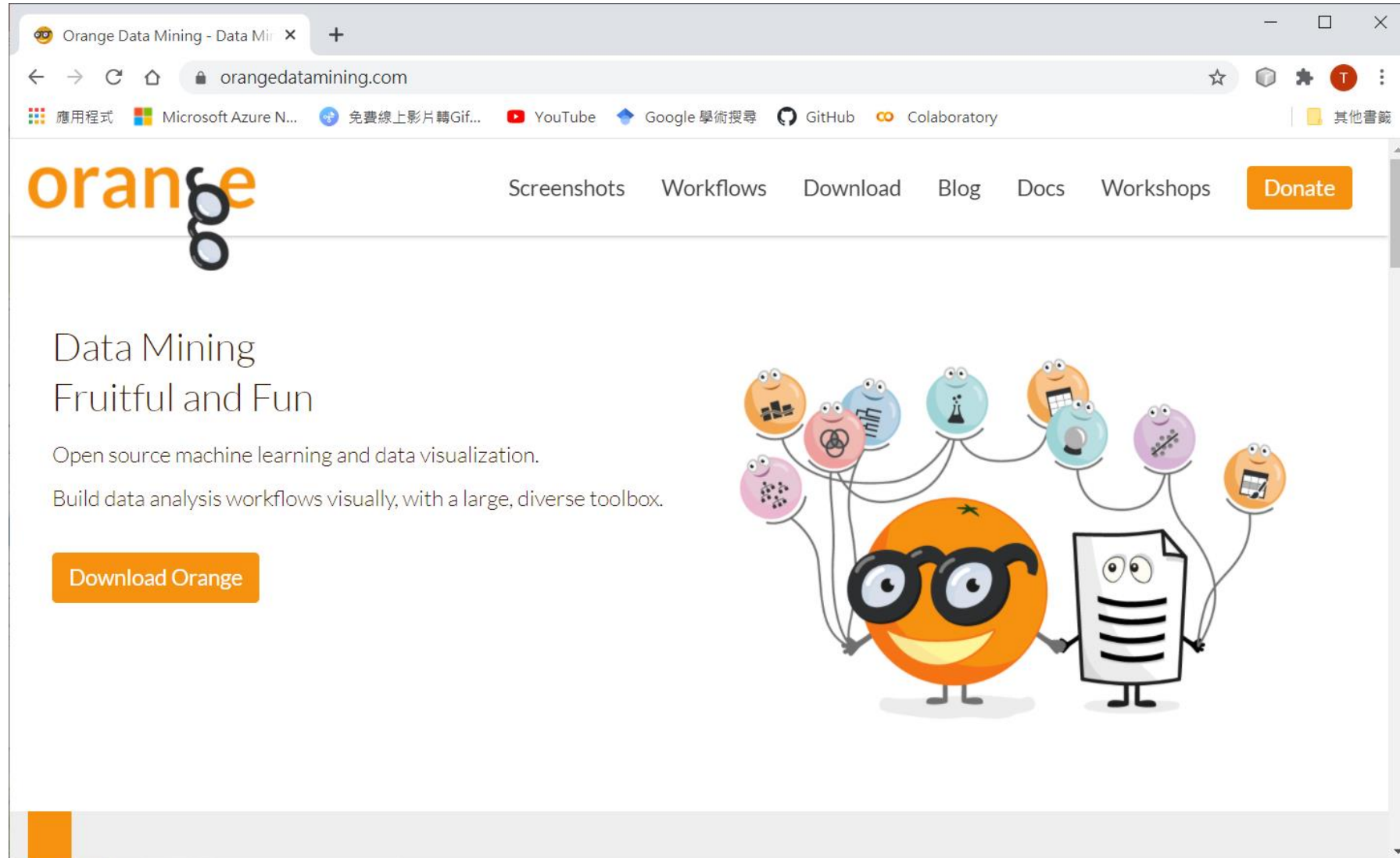
Kaggle

The screenshot shows the Kaggle website interface. The browser address bar displays `kaggle.com/datasets?search=motion+sensor`. The left sidebar contains navigation links: Home, Compete, Data (highlighted with a red circle), Code, Communities, Courses, and More. The main content area features a search bar with the text "motion sensor" (also highlighted with a red circle) and a "Filters" button. Below the search bar, there are tabs for "Datasets", "Tasks", "Computer Science", "Education", "Classification", "Computer Vision", "NLP", and "Data Visualization". The "Datasets" tab is selected, showing a list of 70 datasets. The top three datasets are:

- MotionSense Dataset : Smartphone Sensor Data - HAR** by Mohammad Malekzadeh, updated 3 years ago. Usability 7.6, 361 Files (CSV), 72 MB. 171 votes, Silver medal.
- Emotions Sensor Data Set** by jon.bill, updated 2 years ago. Usability 8.2, 1 File (CSV), 27 KB. 122 votes, Bronze medal.
- Hand Gesture Recognition Database** by GTI, updated 3 years ago. Usability 7.5, 20000 Files (other), 2 GB. 397 votes, Silver medal.

A cookie notice at the bottom states: "We use cookies on Kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies." with "Got it" and "Learn more" buttons. The Windows taskbar at the bottom shows the time as 下午 03:22 on 2021/2/22.

Interactively visual exploration – (2) Orange



<https://orangedatamining.com/>

Orange

The screenshot displays the Orange3 data mining software interface. The main workflow window shows a sequence of three widgets: CSV File Import, Select Columns, and t-SNE, connected by data links. The Select Columns widget is currently active, showing a list of available variables on the left and a list of selected features on the right. The features list includes t1, t2, t3, t4, t5, t6, n1, and n2, which are circled in red. The target variable is set to y1, also circled in red. The t-SNE widget settings are visible on the right, showing a perplexity of 30, PCA components of 15, and a color mapping to y1. The resulting t-SNE plot shows two distinct clusters of data points, colored blue (0) and red (1).

Workflow: CSV File Import → Data → Select Columns → Data → t-SNE

Select Columns Widget:

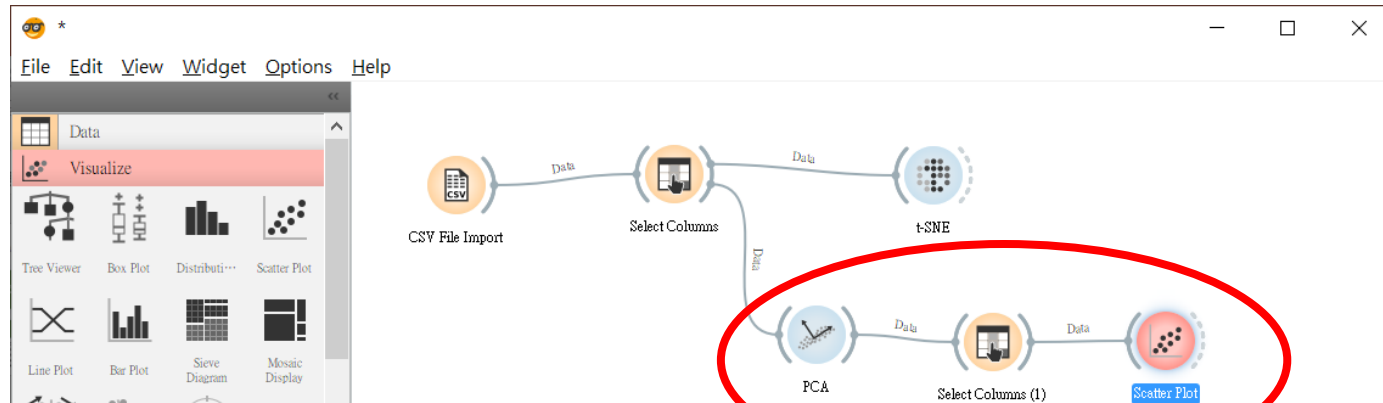
- Available Variables: Gender, TUG, BBS, y2, No, age, time, Step frequency, Steps
- Features: t1, t2, t3, t4, t5, t6, n1, n2
- Target Variable: y1

t-SNE Widget Settings:

- Perplexity: 30
- Exaggeration: 1
- PCA components: 15
- Normalize data: ☒
- Color: y1
- Shape: (Same shape)
- Size: (Same size)
- Label: (No labels)
- Symbol size: [Slider]
- Opacity: [Slider]
- Jittering: [Slider]
- Show color regions: ☐
- Show legend: ☒

t-SNE Plot: A scatter plot showing two clusters of data points. The legend indicates that blue points represent class 0 and red points represent class 1.

Orange



Scatter Plot

Interactive scatter plot visualization with intelligent data visualization enhancements.

[more...](#)

Select Columns (1)

Available Variables

Filter

- t1
- t2
- t3
- t4
- t5
- t6
- n1
- n2
- n3
- n4
- n5
- n6
- px
- py
- pz

Features

Filter

- PC1
- PC2

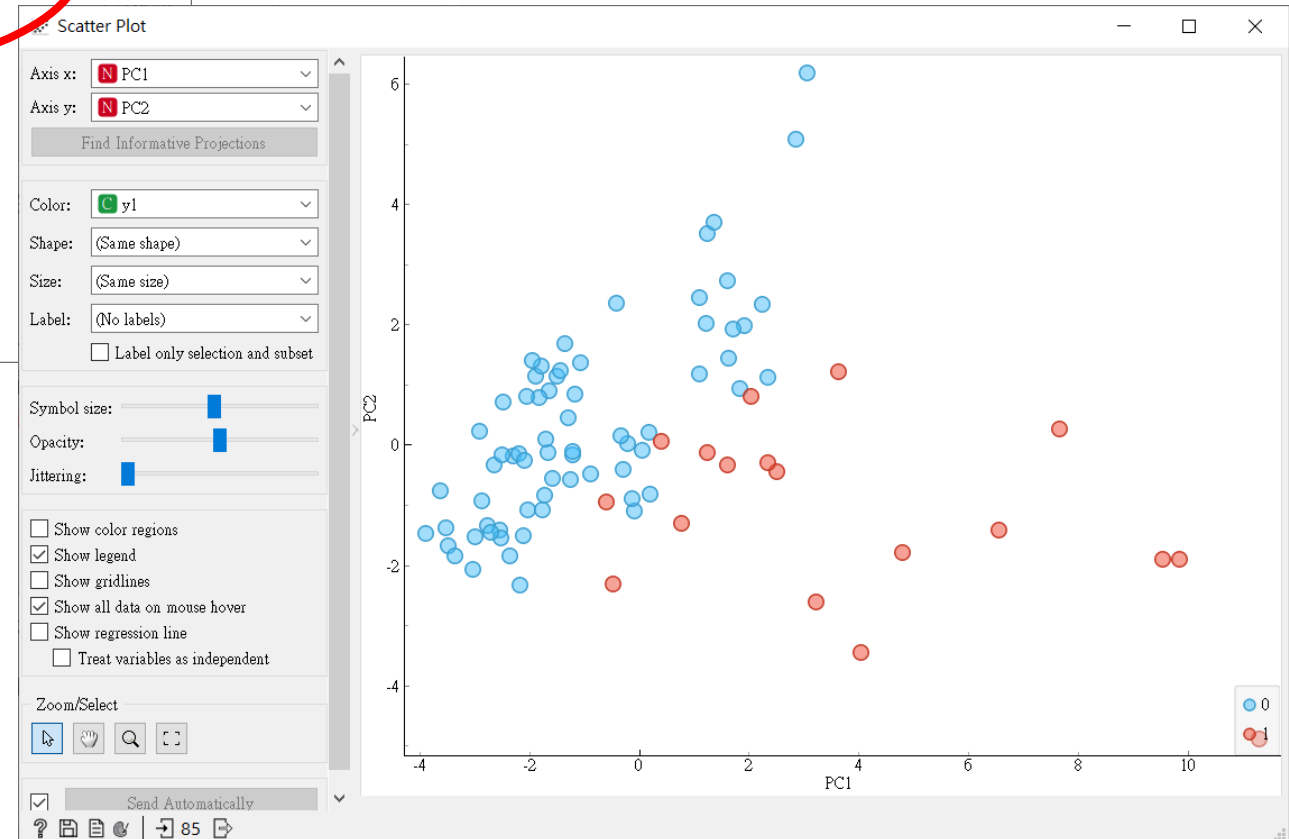
Target Variable

y1

Meta Attributes

Reset

Send Automatically



Practice – Orange

1. Download and install Orange
2. Visualize the 3M TUG data file
3. Search Kaggle to find a classification data file
4. Use Orange to visualize the high dimensional data

Kaggle

Find Open Datasets and Machi x +

kaggle.com/datasets?search=classification

應用程式 Google 學術搜尋 YouTube Colaboratory GitHub TienLungS... 李弘毅 ML

kaggle

Home

Compete

Data

Code

Communities

Courses

More

Search

Sign In Register

+ New Dataset Your Work

Datasets

classification

Filters

Datasets Tasks Computer Science Education Classification Computer Vision NLP Data Visualization

209 Tasks See All

Species Classification
Sachin Sharma · 48 Submissions

Classification Problem
Pratham Tripathi · 25 Submissions

Mushroom Classification using Random Forest Classifier
Sanchita Karmakar · 34 Submissions

Fetal Health Classification
Larxel · 26 Submissions

Iris Species
UCI Machine Learning · Usa...

Drug Classification
Pratham Tripathi · Usability 1...

Mushroom Classifica...
UCI Machine Learning · Usa...

Fetal Health Classific...
Larxel · Usability 10.0

3,010 Datasets

Hottest

We use cookies on Kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies. Got it Learn more

Windows Search File Explorer Edge Mail Calendar Photos OneDrive Word Excel PowerPoint Teams Zoom Slack Visual Studio Code Docker Desktop

下午 03:26
2021/2/22

Data visualization – (3) Python coding

The screenshot shows the Google Colaboratory web interface. A modal window is open for connecting to a GitHub repository. The modal has a header with tabs: 範例 (Examples), 最近 (Recent), Google 雲端硬碟 (Google Drive), **GitHub** (highlighted with a red circle), and 上傳 (Upload). Below the tabs, there is a search bar with the text "輸入 GitHub 網址或依機構或使用者搜尋" (Enter GitHub URL or search by organization or user). The search results show "TienLungSun" (highlighted with a red circle) as the selected user. Below the user name, there is a dropdown menu for the repository, currently showing "TienLungSun/Orange-for-ML". To the right of the repository name is a dropdown for the branch, currently showing "main". Below these, there is a list of notebooks available in the repository, including "Read_and_visualize_high_dimensional_data.ipynb" and "Read_and_visualize_sensor_data_file.ipynb". At the bottom right of the modal, there are buttons for "新增筆記本" (Add notebook) and "取消" (Cancel). The background shows the Colaboratory interface with a sidebar on the left containing a directory tree and a main area on the right for code execution. The Windows taskbar is visible at the bottom of the screen.

歡迎使用 Colaboratory

檔案 編輯 檢視畫面 插入

目錄

- 開始使用
- 數據資料學
- 機器學習
- 其他資源
- 機器學習範例
- 區段

範例 最近 Google 雲端硬碟 **GitHub** 上傳

輸入 GitHub 網址或依機構或使用者搜尋 ☐ 包括私人存放區

TienLungSun 🔍

存放區: TienLungSun/Orange-for-ML 分支版本: main

路徑

- Read_and_visualize_high_dimensional_data.ipynb
- Read_and_visualize_sensor_data_file.ipynb

新增筆記本 取消

如要執行上方儲存格中的程式碼，請按一下進行選取，再按一下程式碼左側的播放鍵，或是使用鍵盤快速鍵「Command/Ctrl + Enter 鍵」。按

下午 03:11 2021/2/22

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

Data visualization with Python

Read and visualize sensor data file

檔案 編輯 檢視畫面 插入 執行階段 工具 說明

+ 程式碼 + 文字 複製到雲端硬碟

連線 編輯

```
from google.colab import files
uploaded = files.upload()
```

選擇檔案 未選擇任何檔案

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving MotionData110000_001001_2018_09_14_09_39_26.csv to MotionData110000_001001_2018_09_14_09_39_26.csv

```
[ ] fnameLst = list(uploaded.keys())
    fname = fnameLst[0]
    print(fname)
```

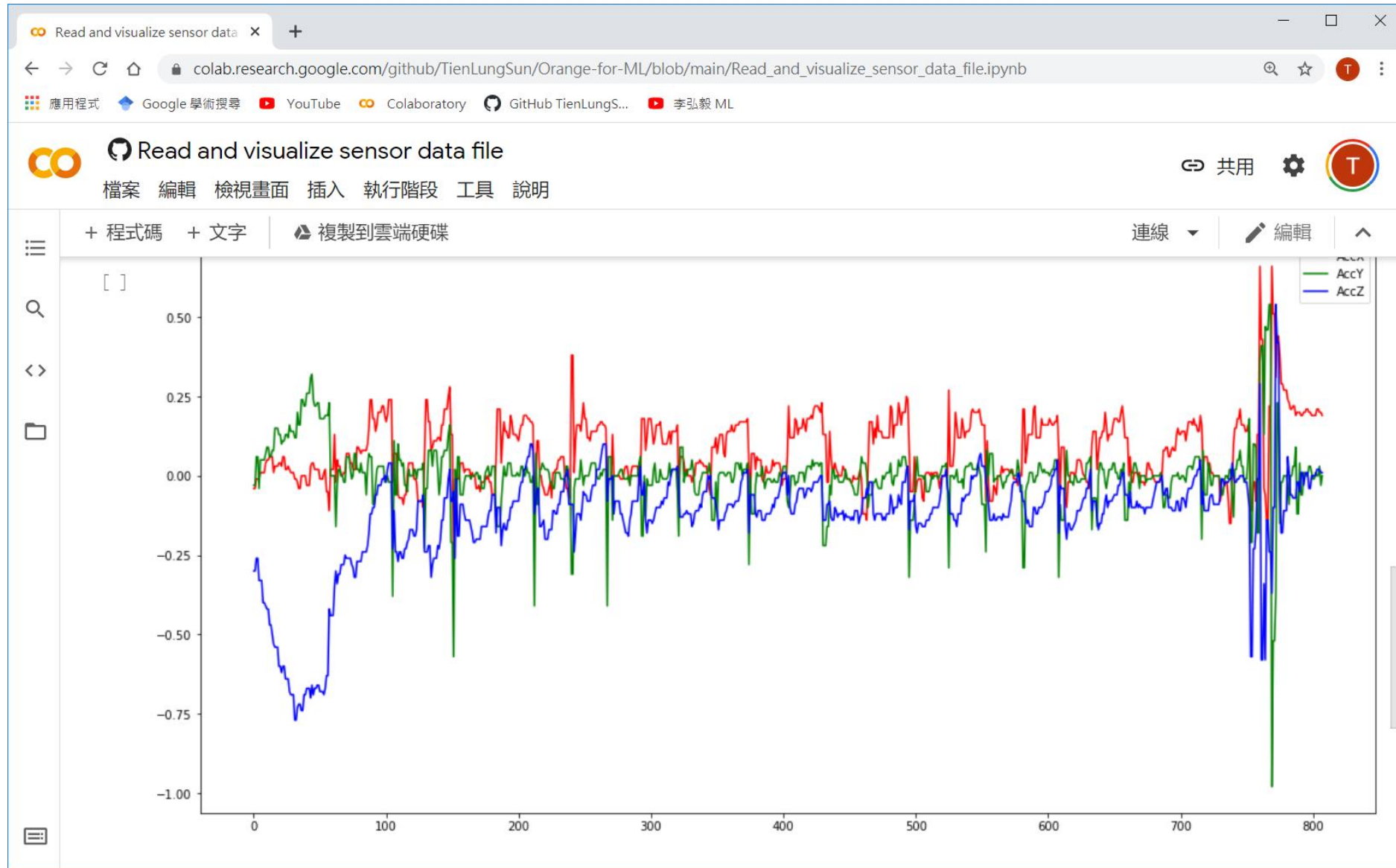
MotionData110000_001001_2018_09_14_09_39_26.csv

```
[ ] import pandas as pd
    df=pd.read_csv(fname)
```

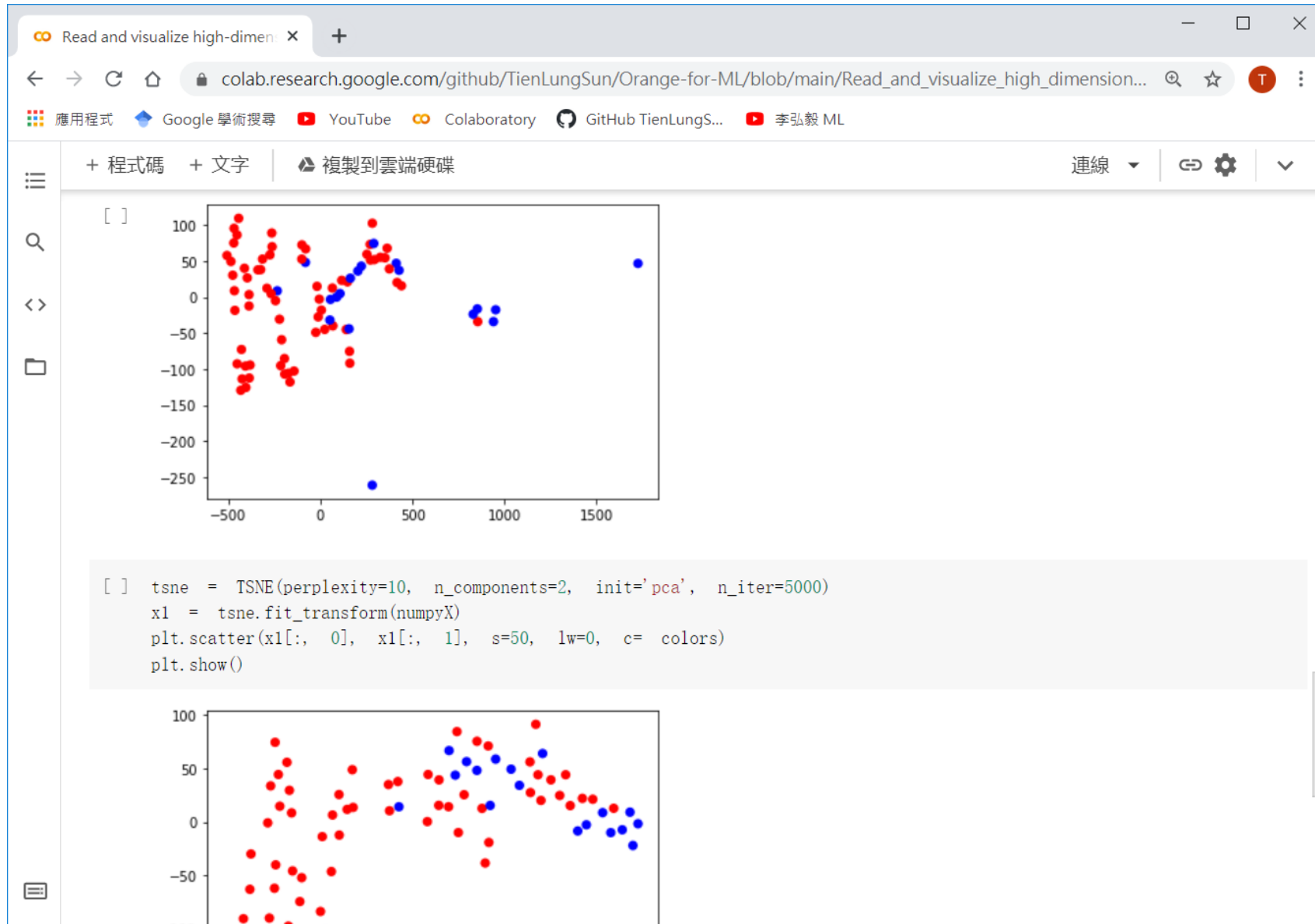
```
[ ] df.head(5)
```

	time	AccX	AccY	AccZ	EulerX	EulerY	EulerZ
0	09:39:26:180	-0.04	-0.03	-0.30	-106.05	-66.92	2.16

Data visualization with Python



Data visualization with Python

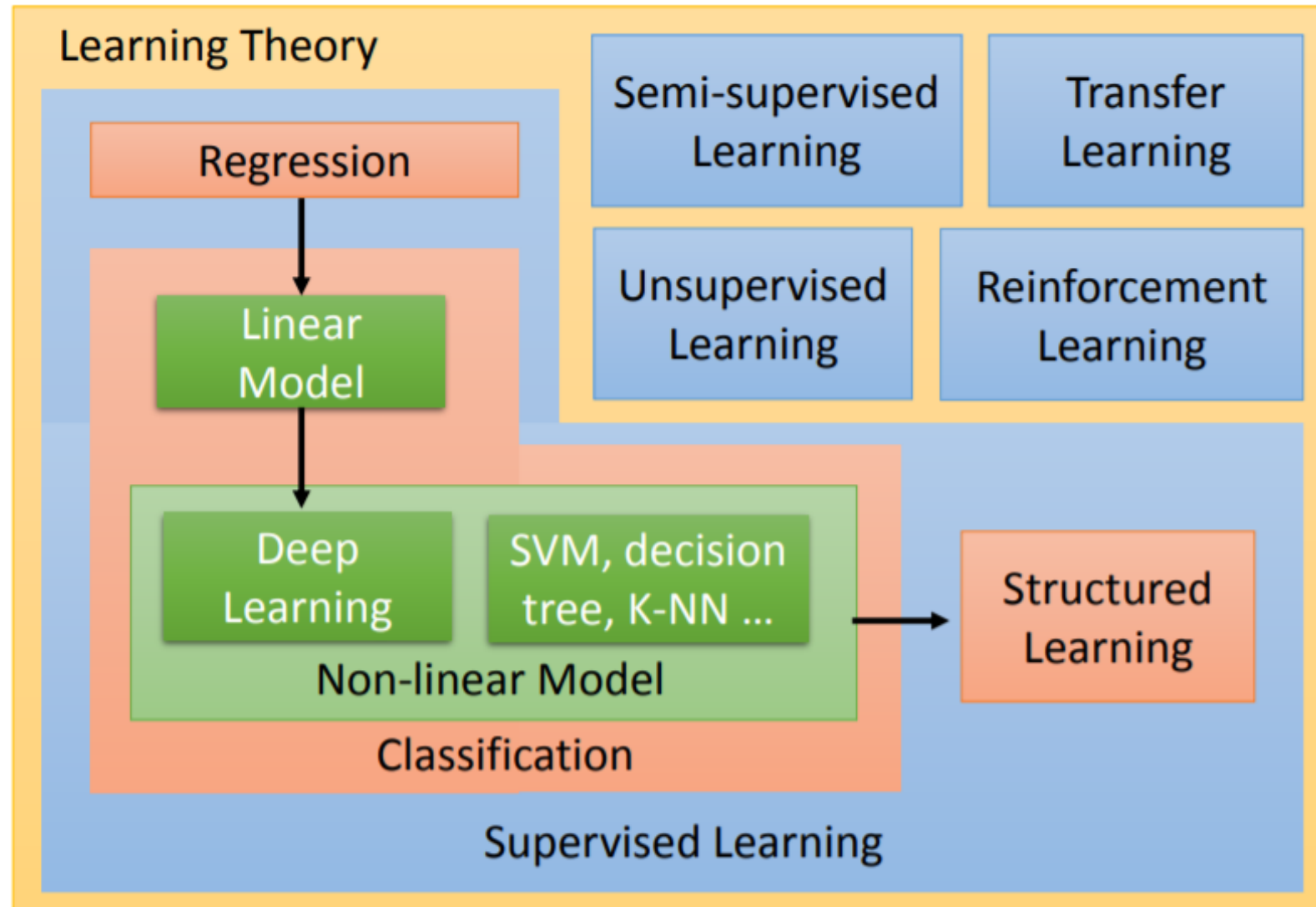


Practice – Python coding

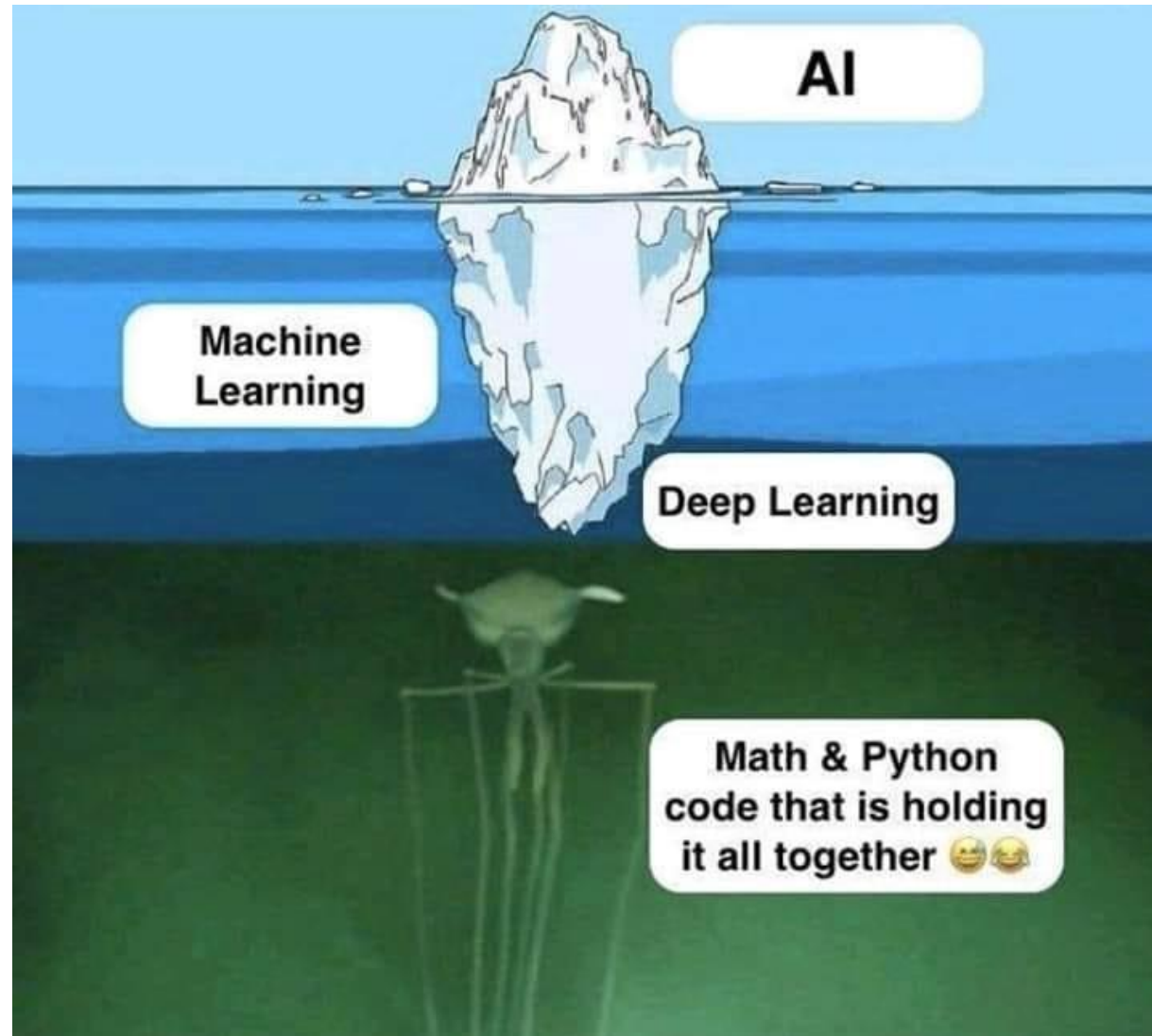
1. Log in to Colab
2. Run python code to visualize the motion sensor data file
3. Run python code to visualize the 3M TUG data file
4. Run python code to visualize the two data files you download from Kaggle

Employ AI (ML/DL) to learn from high-dimensional data
to assist human visual data exploration

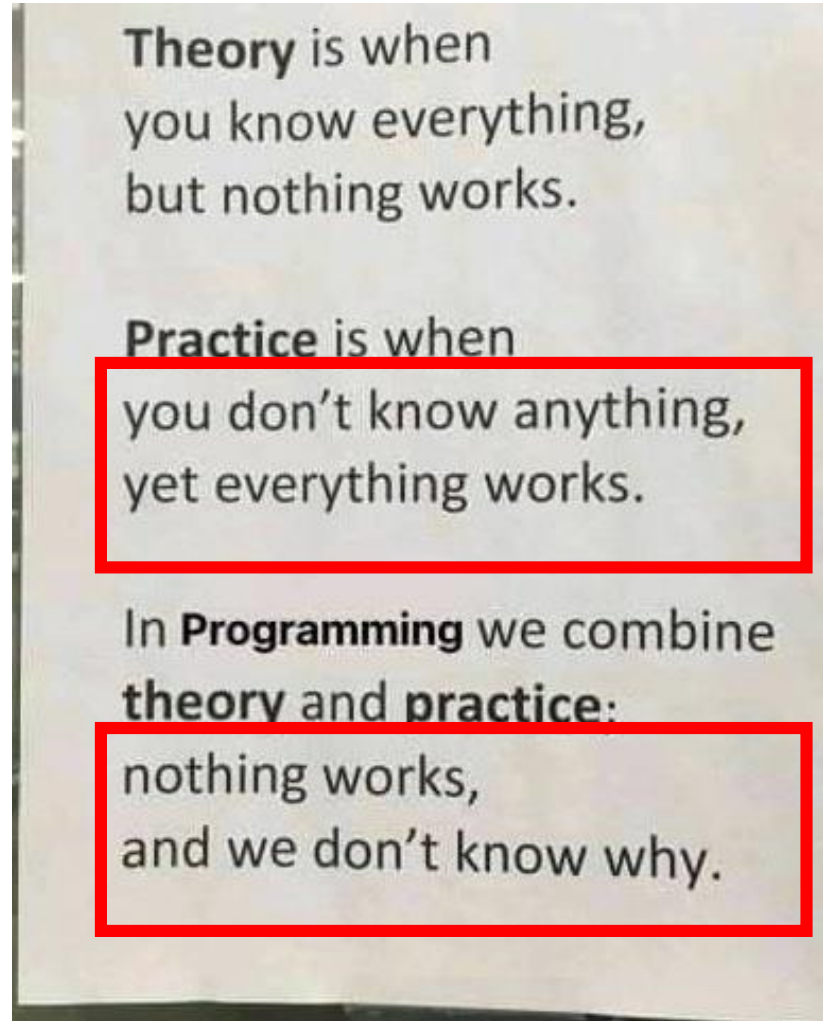
How computer learns from data ?



Math + Python coding



Why we need to study math + python coding together ?



Run exemplar PyTorch code from GitHub

Starts to train your own NN for your thesis

Python development tools

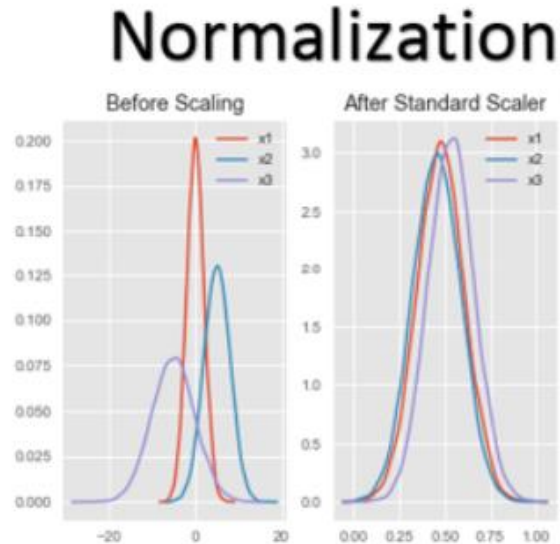


Before we start, two more issues not covered in this class but
are important in AI development practice

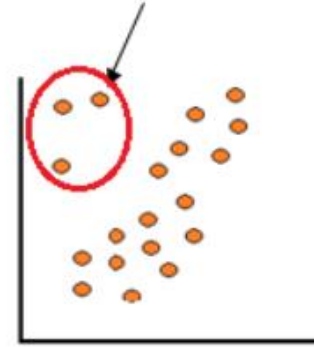
Data pre-processing

Topics to be covered:

1. Standardization
2. Scaling with sparse data and outliers
3. Normalization
4. Categorical Encoding
5. Imputation



Outliers



Imputation

	First	Second	Third
0	100.0	30.0	NaN
1	90.0	45.0	40.0
2	NaN	56.0	80.0
3	95.0	NaN	98.0

Encoding

Food Name	Apple	Chicken	Broccoli
Apple	1	0	0
Chicken	0	1	0
Broccoli	0	0	1

Data preprocessing in Python

Steps involved in data preprocessing :

1. Importing the required Libraries
2. Importing the data set
3. Handling the Missing Data.
4. Encoding Categorical Data.
5. Splitting the data set into test set and training set.
6. Feature Scaling.

<https://aaaanchakure.medium.com/data-preprocessing-3cd01eefd438>

Steps in Data Preprocessing in Machine Learning

1. Acquire the dataset
2. Import all the crucial libraries
3. Import the dataset
4. Identifying and handling the missing values
5. Encoding the categorical data
6. Splitting the dataset
7. Feature scaling

<https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/>

Why Data Preprocessing?

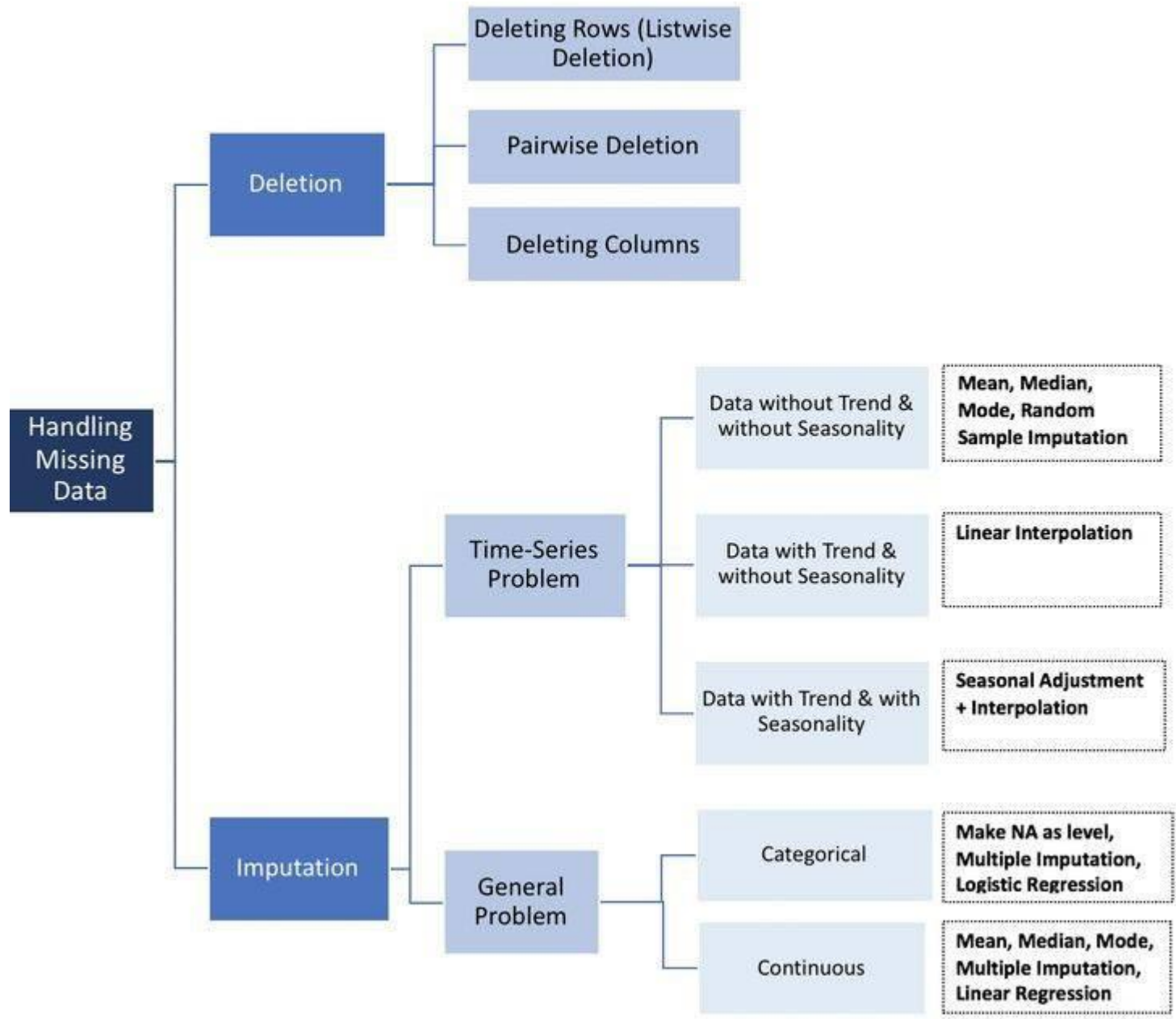
- **Data in the real world is not clean**
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data, e.g., occupation=" "
 - **noisy**: containing errors or outliers, e.g., Salary="-10"
 - **inconsistent**: containing discrepancies in codes or names, e.g., Age="42" Birthday="03/07/1997"
- **No quality data, no quality mining results!**
- **Quality decisions must be based on quality data**
 - Duplicates or missing data may cause incorrect or misleading analyses.

Data Preprocessing: Major Tasks

- **Data cleaning**
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
 - Integration of multiple databases, data cubes, or files
- **Data transformation**
 - Normalization and aggregation
- **Data reduction**
 - Obtains reduced representation in volume but produces the same or similar analytical results
- **Data discretization**
 - Part of data reduction but with particular importance, especially for numerical data

Data Cleaning Tasks

- **Fill in missing values**
- **Identify outliers and smooth out noisy data**
- **Correct inconsistent data**
- **Resolve redundancy caused by data integration**

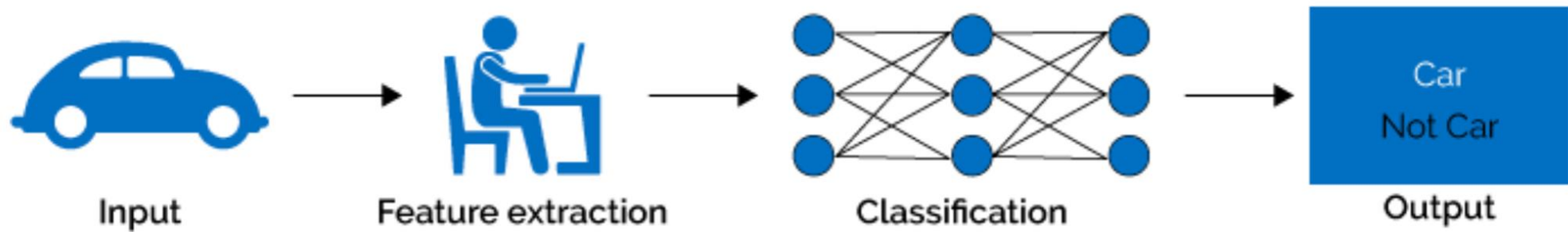


Data Transformation

- **Smoothing: remove noise from data**
- **Aggregation: summarization, data cube construction**
- **Generalization: concept hierarchy climbing**
- **Normalization: scaled to fall within a small, specified range**
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
- **Attribute/feature construction**
 - New attributes constructed from the given ones

Feature engineering

Machine Learning



Deep Learning

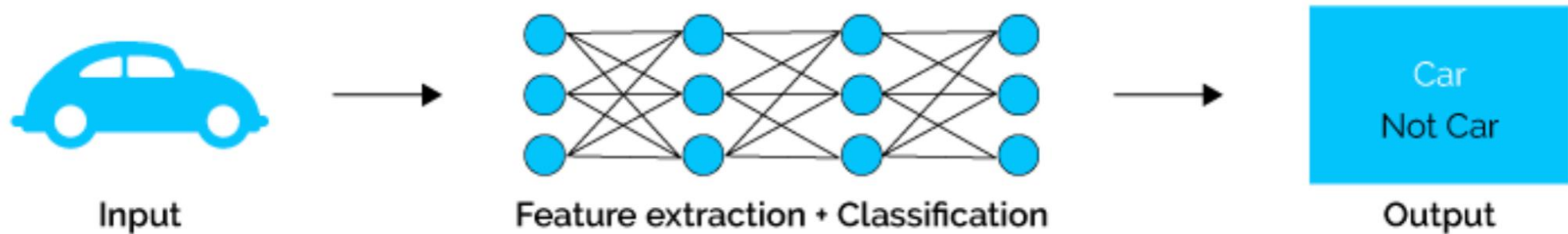
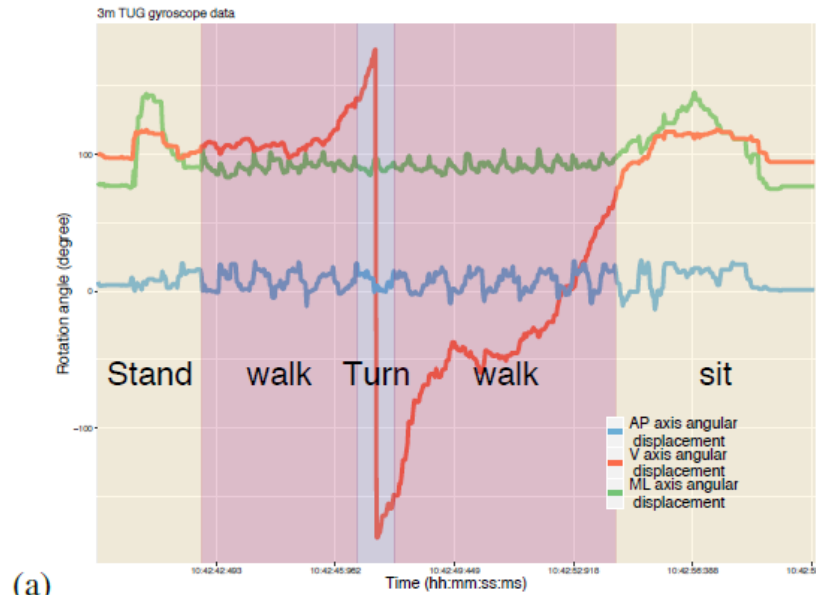
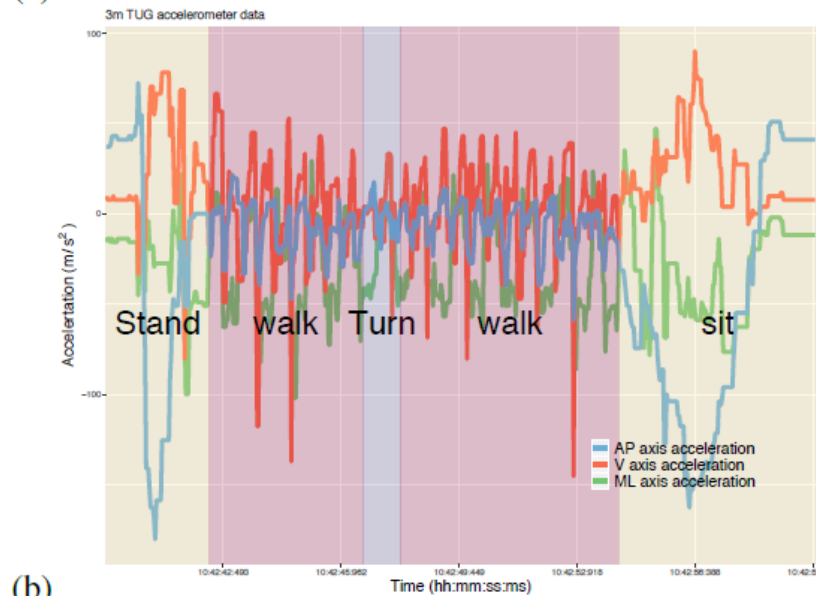


TABLE II
SUMMARY OF FEATURES

TUG sensor-based features		
Feature name	Direction	Unit
Sit to stand		
Angle range	ML	deg
Acceleration range	V, AP	m/s^2
Completion time		s
Walking		
Acceleration CV	V, AP, ML	
Acceleration range	V, AP, ML	m/s^2
Acceleration median	V, AP, ML	m/s^2
Acceleration RMS	V, AP, ML	m/s^2
Angular velocity CV	V, AP, ML	
Angular velocity range	V, AP, ML	deg/s
Angular velocity median	V, AP, ML	deg/s
Angular velocity RMS	V, AP, ML	deg/s
Speed		m/s
Turning		
Acceleration CV	V, AP, ML	
Acceleration range	V, AP, ML	m/s^2
Acceleration median	V, AP, ML	m/s^2
Acceleration RMS	V, AP, ML	m/s^2
Angular velocity CV	V, AP, ML	



(a)



(b)

Feature engineering

- › Supervised Feature Selection Methods
 - › Wrapper Feature Selection Methods
 - › Filter Feature Selection Methods
 - › Embedded or Intrinsic Feature Selection Methods
- › Feature Selection with Statistical Measures
 - › Univariate Feature Selection
- › Feature Selection Strategies
 - › Selection Method
 - › Transform Variables
- › Which Feature Selection Method is the Best?
- › Feature Selection Implementations
 - › Feature Selection For Regression models
 - › Classification Feature Selection