K6312 Information Mining & Analysis

Chen Zhenghua & Zhao Rui

Course Information

Course Webpage

Chen Zhenghua
(Instructor)

zhenghua.chen@ntu.edu.sg

Zhao Rui
(Instructor)

rui.zhao@ntu.edu.sg

https://k6312.github.io/

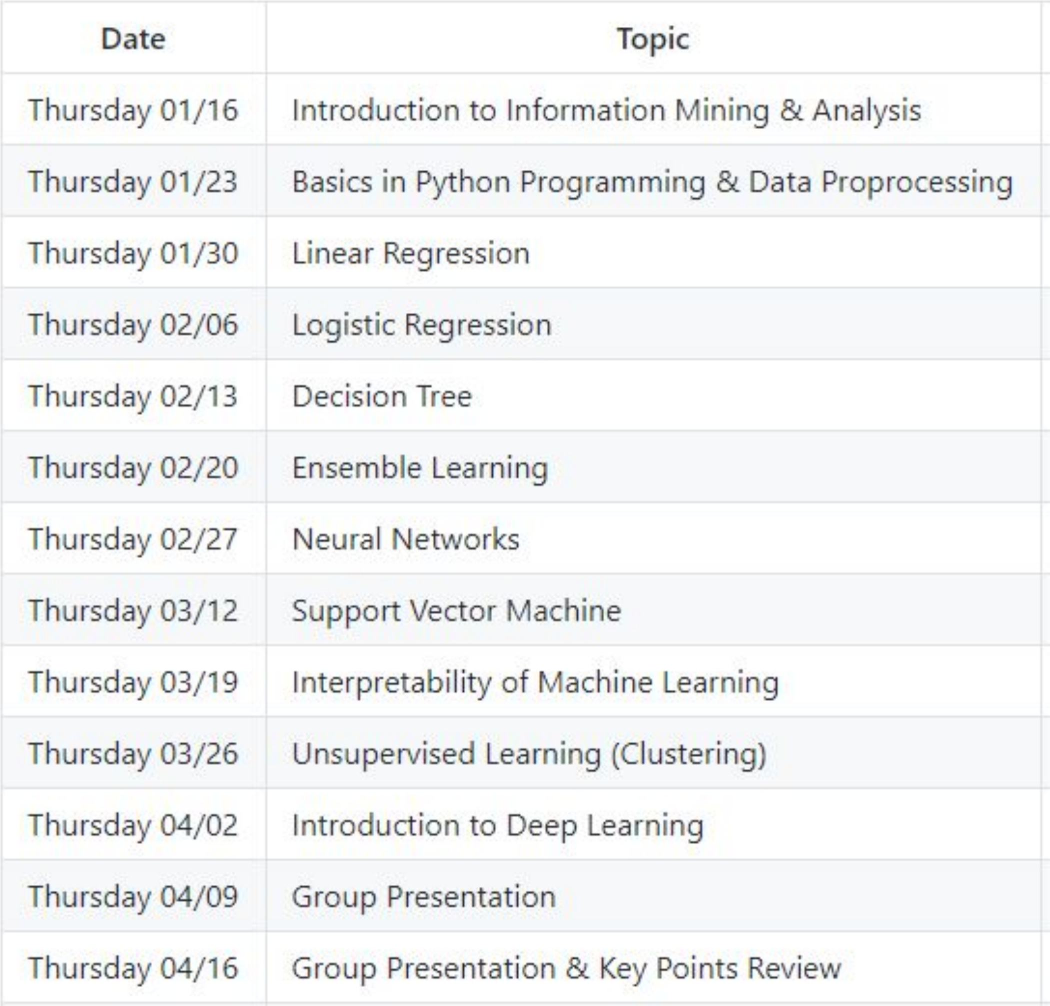
Goals of this Course

Learn how to analyse various data

- Know the basics for data mining and analysis
- Various data mining algorithms
  - Pre-processing, visualization, supervised learning, unsupervised learning, etc.
- Practice on real applications
  - Classification, regression, clustering, etc.

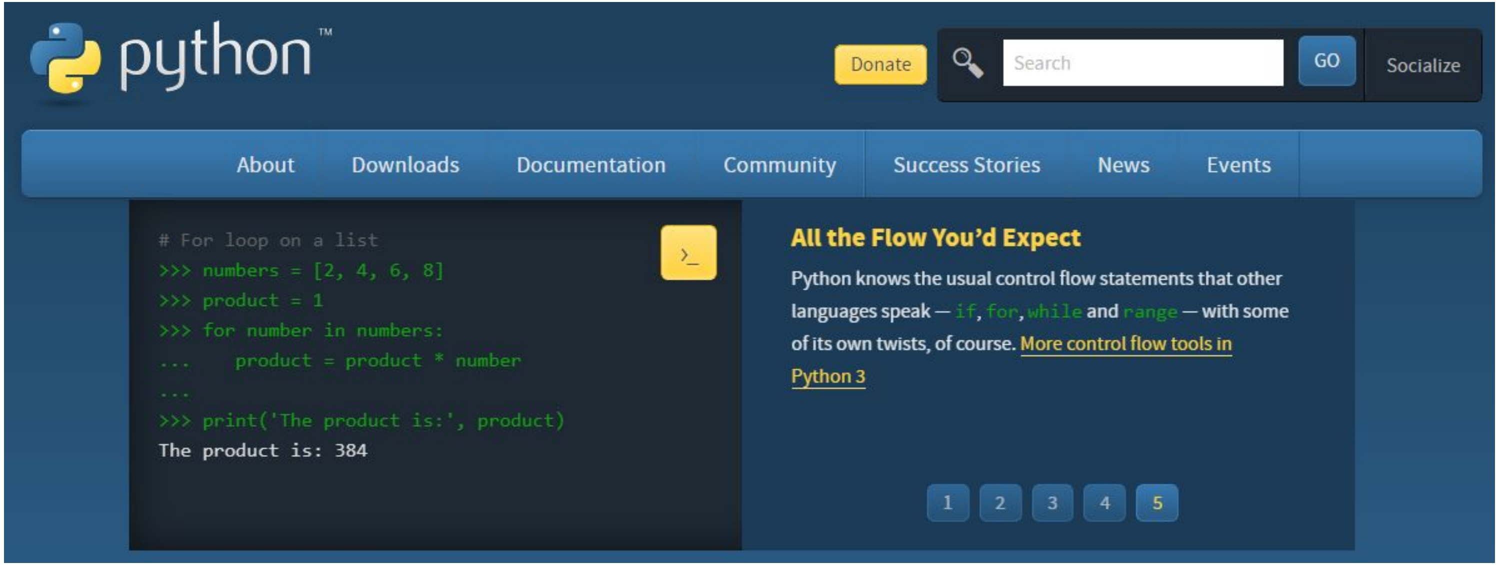
Course Assessment

- Class Participation (10%)
- Assignments (10%)
  - Two 90-minutes in-class assignments (5% each)
- Group Project (30%)
  - Project Report (15%)
  - Final Presentation (15%)
- Final Exam(50%)
  - Closed book



Date	Topic	
Thursday 01/16	Introduction to Information Mining & Analysis	Chen Zhenghua
Thursday 01/23	Basics in Python Programming & Data Proprocessing	Chen Zhenghua
Thursday 01/30	Linear Regression	Chen Zhenghua
Thursday 02/06	Logistic Regression	Zhao Rui
Thursday 02/13	Decision Tree	Zhao Rui
Thursday 02/20	Ensemble Learning	Chen Zhenghua
Thursday 02/27	Neural Networks	Chen Zhenghua
Thursday 03/12	Support Vector Machine	Chen Zhenghua
Thursday 03/19	Interpretability of Machine Learning	Zhao Rui
Thursday 03/26	Unsupervised Learning (Clustering)	Chen Zhenghua
Thursday 04/02	Introduction to Deep Learning	Chen Zhenghua
Thursday 04/09	Group Presentation	Chen Zhenghua
Thursday 04/16	Group Presentation & Key Points Review	Chen Zhenghua

Programming Language for Hands-on



```
# For loop on a list
>>> numbers = [2, 4, 6, 8]
>>> product = 1
>>> for number in numbers:
...     product = product * number
...
>>> print('The product is:', product)
The product is: 384
```

All the Flow You'd Expect

Python knows the usual control flow statements that other languages speak — `if`, `for`, `while` and `range` — with some of its own twists, of course. More control flow tools in Python 3

Let us Start

Information Mining & Analysis on *Data*

- Smartphone sensor data to detect human activities
- Infer health status based on activity levels

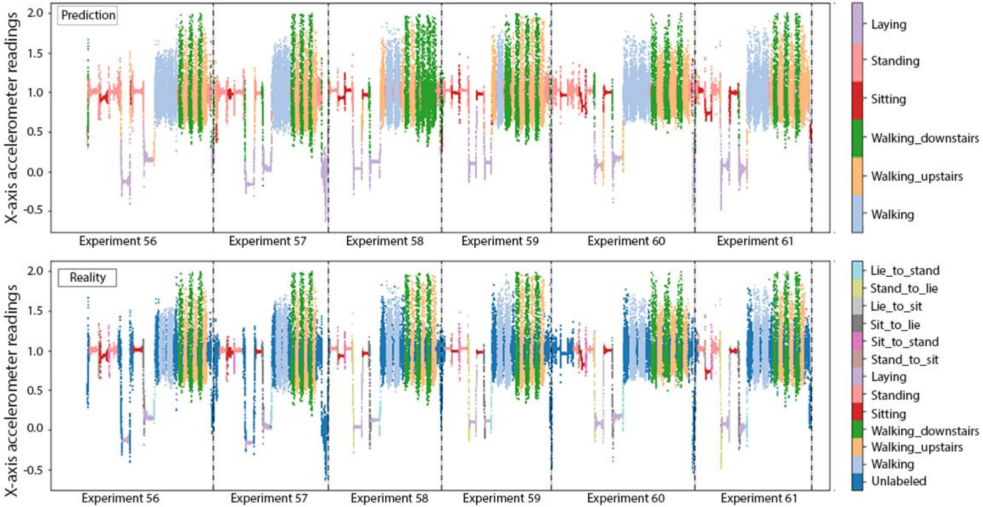
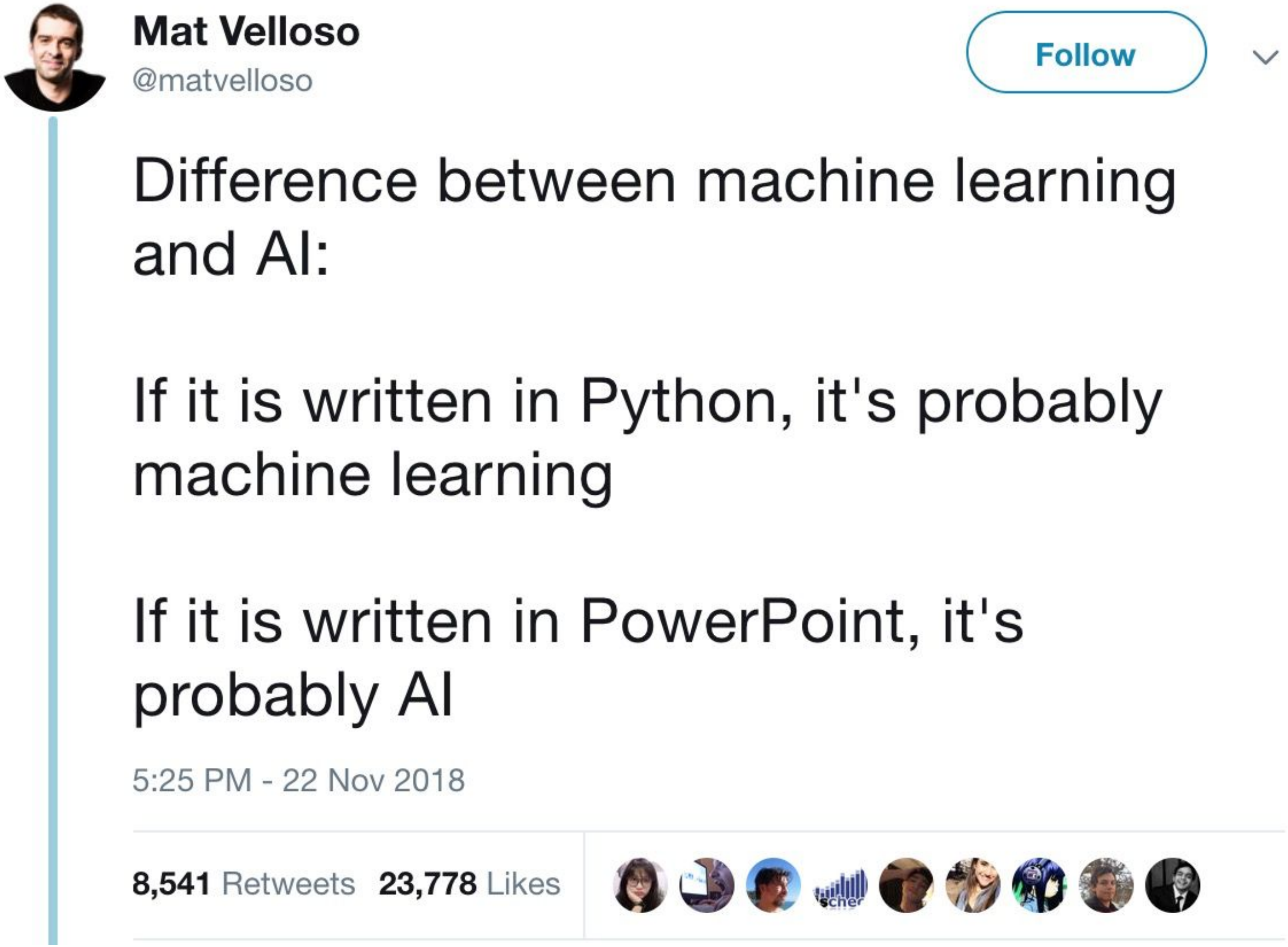


Image source: https://mhealth.jmir.org/2019/2/e11201/, https://securityboulevard.com/2019/08/healthcare-is-in-cybercriminals-crosshairs/

What is Machine Learning



Mat Velloso
@matvelloso

Difference between machine learning and AI:

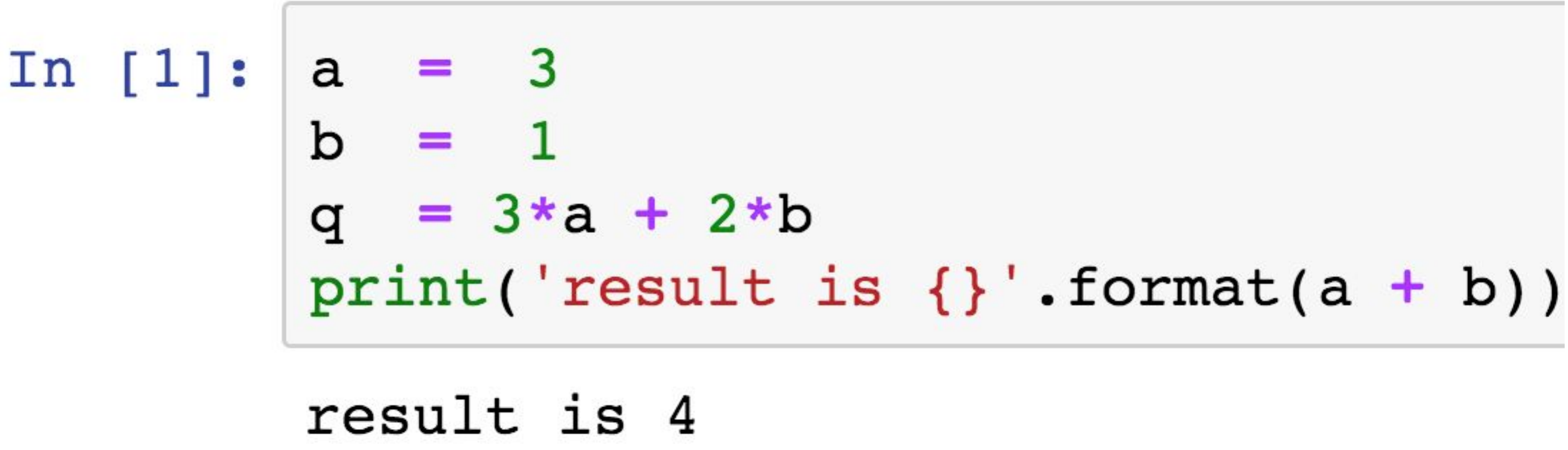
If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI

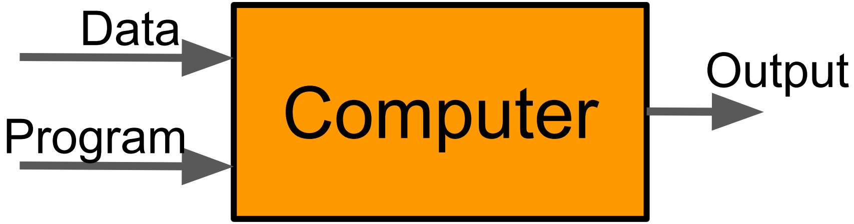
5:25 PM - 22 Nov 2018

8,541 Retweets 23,778 Likes

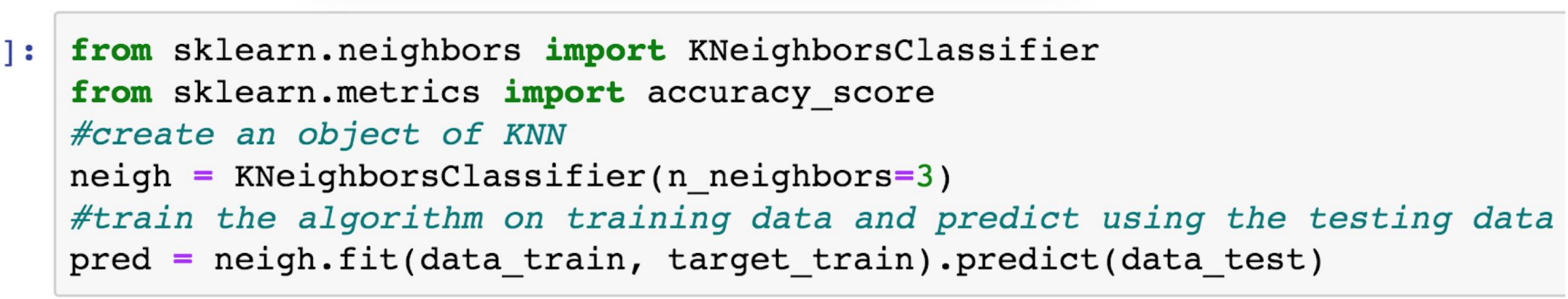
Python Programming



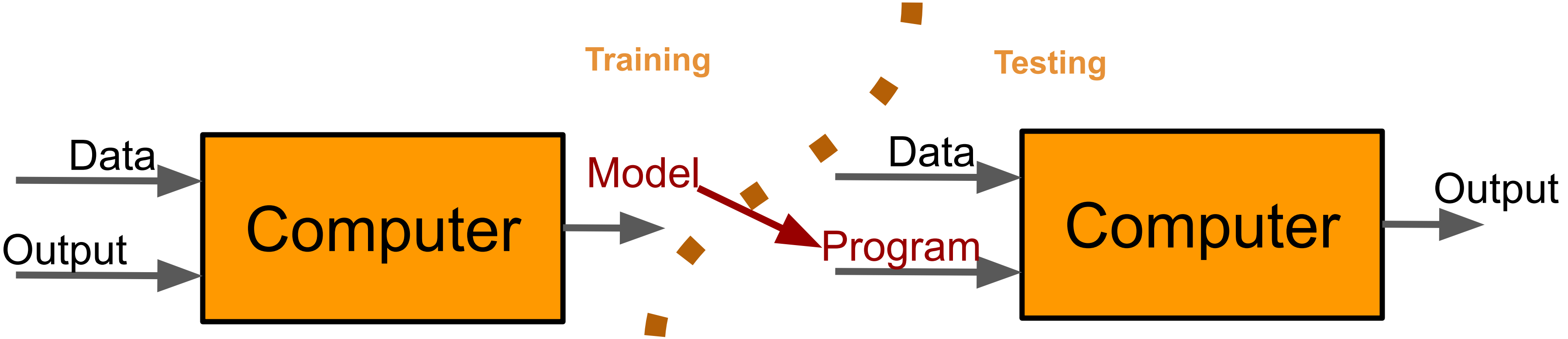
```
In [1]: a  =  3
        b  =  1
        q  = 3*a + 2*b
        print('result is {}'.format(a + b))

        result is 4
```



Machine Learning



```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
#create an object of KNN
neigh = KNeighborsClassifier(n_neighbors=3)
#train the algorithm on training data and predict using the testing data
pred = neigh.fit(data_train, target_train).predict(data_test)
```



Definition of Machine Learning

“A computer program is said to learn from experience **E**
with respect to some class of tasks **T**
and performance measure **P**,
if its performance at tasks in **T**, as measured by **P**,
improves with experience **E**”

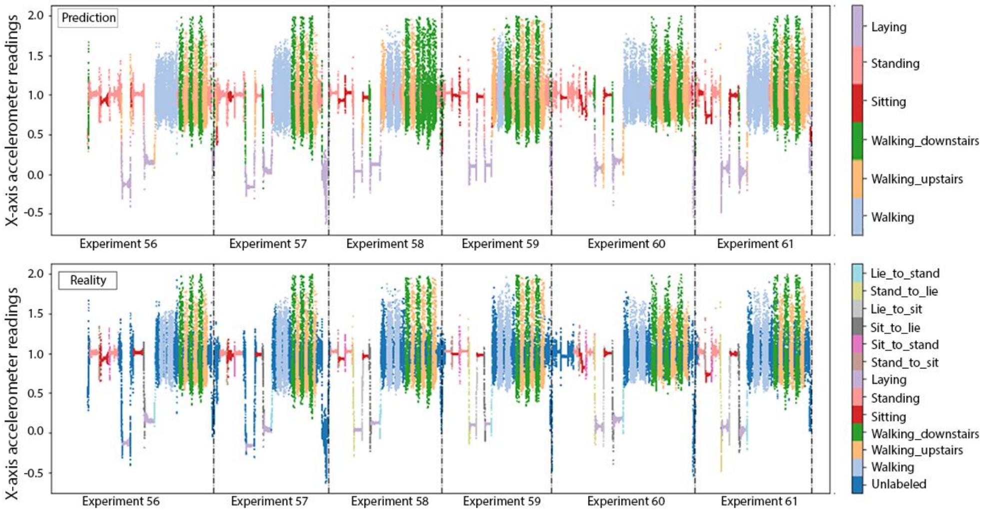
Tom Mitchell

T, **P**, **E** are three basic elements to define a complete machine learning tasks

AlphaGo

T : **Play Go Games**
P : **Win rates of all matches**
E : **Match Experiences with many go players or itself**

Activity Recognition



T : Identify different activities

P : Accuracy that human activities are identified

E : Dataset of labelled sensory data

Data Mining

What is Data Mining

- Definition: The process of discovery of interesting, meaningful and actionable patterns hidden in large amounts of data
- Data mining tasks
  - ***Classification*** maps data into predefined groups or classes
  - ***Regression*** is used to map a data item to a real valued prediction variable
  - ***Clustering*** groups similar data together into clusters
- Preferential Questions
  - Which technique to choose?
    - Classification/regression/clustering
    - Answer: Depends on what you want to do with data?

Some Definitions ...

- Concepts: kinds of things that can be learned
    - Example: the relation between patient characteristics and the probability to be diabetic
- Instances/Samples: the individual, independent examples of a concept
    - Example: a patient, candidate drug etc.
- Attributes/Features: measuring aspects of an instance
    - Example: age, weight, lab tests, microarray data etc.

Data Mining Process

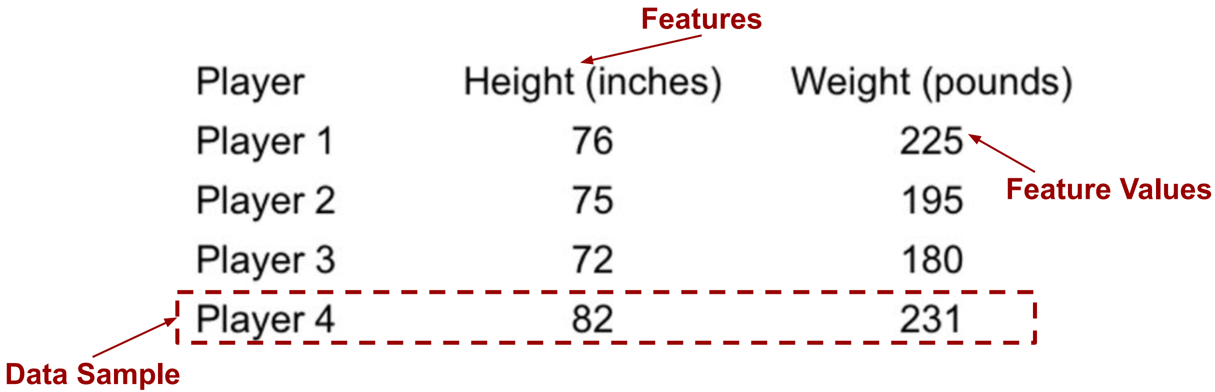
- Problem formulation
- Data collection
- Pre-processing: cleaning
  - Interpretation for missing data
  - Outlier detection
  - Normalization
- Feature Engineering
  - Find relevant representations of data samples
- Choose machine learning algorithms
- Result evaluation and visualization

Structured Data

- Structured Data
  - Machine learning/predictive algorithms need fixed-length vectors as inputs
  - Structured data is easily to be handled/prepared by our humans
  - Can be represented by columns and rows.
  - Each row is a data sample. Each column is attribute/feature.
- A toy task: predict the position of the basketball player

Structured Data for Toy Example

- Structured: just like the excel file or csv



Player	Height (inches)	Weight (pounds)	Position
Player 1	76	225	C
Player 2	75	195	PG
Player 3	72	180	SF
Player 4	82	231	PF

Features

Feature Values

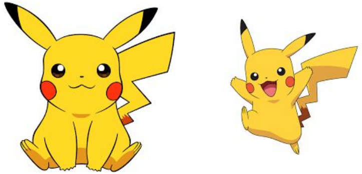
Data Sample

Labels

Unstructured

- The original data can not be stored in an “table”
- More abstract, more fuzzy, and more high-dimensionality

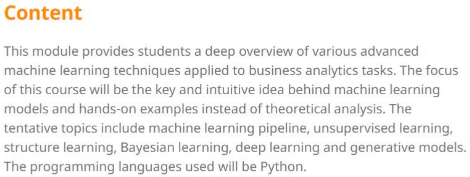
Images



Audio

Video

Text

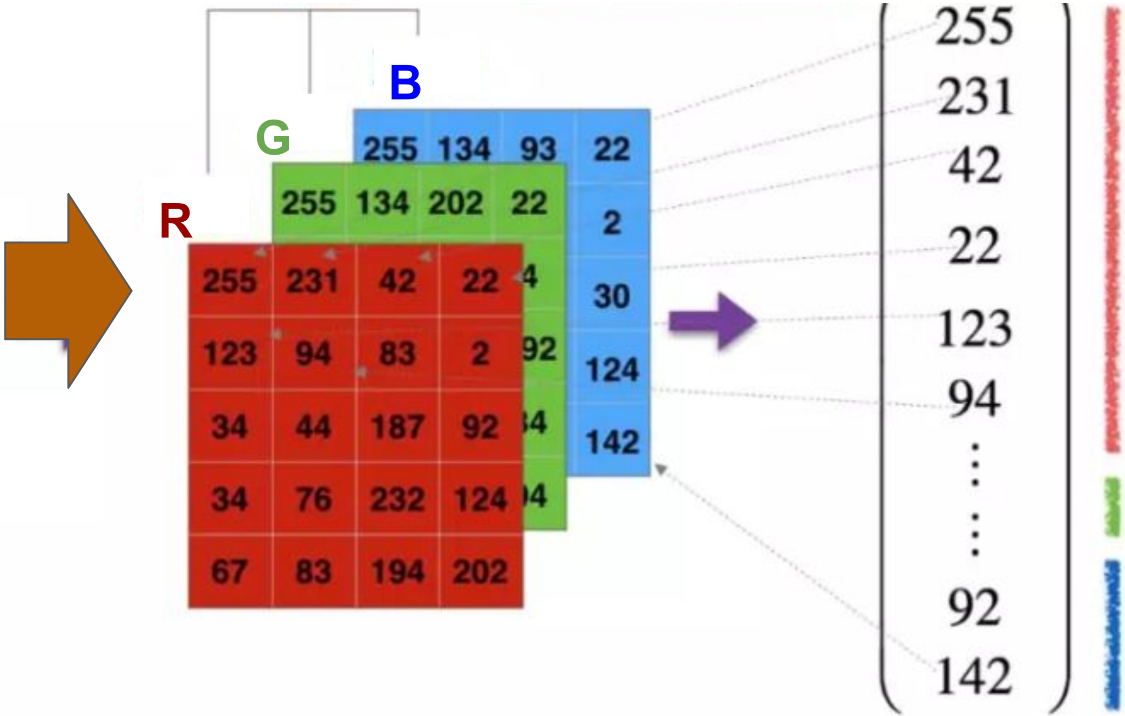


Content

This module provides students a deep overview of various advanced machine learning techniques applied to business analytics tasks. The focus of this course will be the key and intuitive idea behind machine learning models and hands-on examples instead of theoretical analysis. The tentative topics include machine learning pipeline, unsupervised learning, structure learning, Bayesian learning, deep learning and generative models. The programming languages used will be Python.

Environment around agent

For Images



255	231	42	22
123	94	83	2
34	44	187	92
34	76	232	124
67	83	194	202

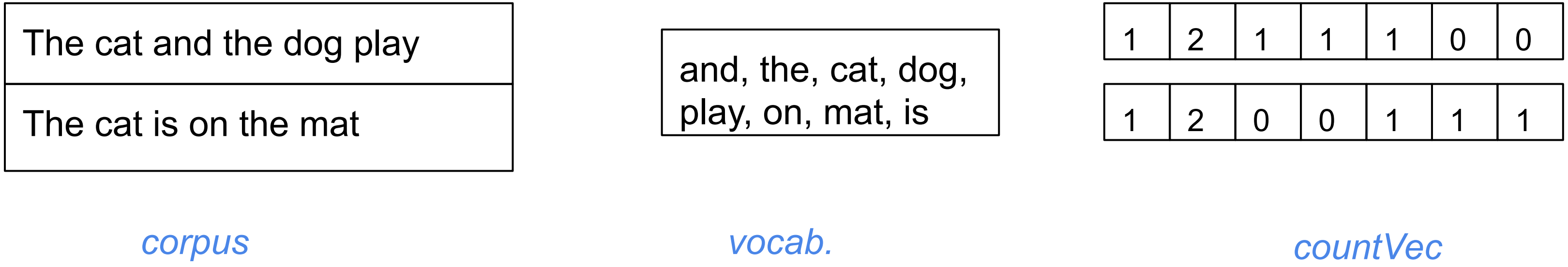
Flatten of red matrix

Flatten of green matrix

Flatten of blue matrix

For Text

- One of the main themes supporting text mining is **the transformation of text into numerical data**.
- Although the initial presentation is document format, the data move into a classical data-mining encoding (from unstructured to structured).
  - Each data is a vector
  - The length of the vector should be fixed
- Each row represents a document and each column a word.



The cat and the dog play
The cat is on the mat

corpus

and, the, cat, dog,
play, on, mat, is

vocab.

1	2	1	1	1	0	0
1	2	0	0	1	1	1

countVec

Data Mining Applications

Classification: Applications 1

- **Direct Marketing**
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decide to buy and which decided otherwise. This *{buy, don’t buy}* decision forms the class labels.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Applications 2

- **Fraud Detection**
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc.
    - Label past transactions as fraud or fair transactions. This forms the class label.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Regression: Applications 1

- **Stock Price**
    - Goal: Predict stock price of a company
    - Approach:
        - To identify key attributes related to stock price
            - Profit, Financing, GDP of the country, political environment, etc.
            - Past prices and trends.
        - Select a proper model to predict

Regression: Applications 2

- **Crowd Counting Using Computer Vision**
  - Goal: Count the number of subjects for a specific area.
  - Approach:
    - To identify key attributes in images
      - Color Gradient Histogram, Grayscale Pixel Values, etc.
    - Select a proper model to link the attributes with true labels

ShanghaiTech Dataset

Clustering: Applications 1

- **Document Clustering:**
    - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
    - Approach:
        - To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
    - Gain: Information Retrieval can utilize the clusters to relate a new document or search terms to clustered documents.

Clustering: Applications 2

- **Market Segmentation:**
  - Goal: Subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.