kuala lumpur

school of ai

Data Analytics 101

Importance of Data

- Data is information in a raw and unorganized form that can be digitally manipulated to represent conditions, objects, or ideas.

- Common types of data include sales figures, marketing research results, readings from weather sensors, or a list of cities and their populations.

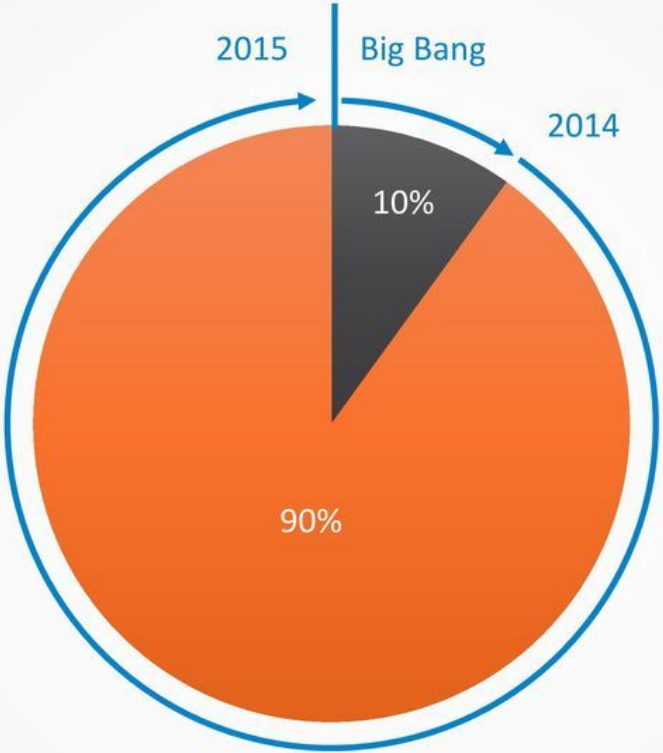
- We now generate an estimated 2.5 quintillion bytes of data each day.

- 90% of world's data comes from the past few years

- A 1% data-driven productivity improvement in aviation could save $30 billion in fuel savings worldwide over 15 years

https://data.bsa.org/

90% of the World's Data



was created in the last 2 years

http://www.sintef.no/home/Press-Room/Research-News/Big-Data--for-better-or-worse/

Data Analytics

- Application of software as a way of transforming and modeling data in order to derive useful information, insights or meaning from data.

- It is often used to uncover hidden patterns or unknown correlations, and aid in decision-making.

https://data.bsa.org/

Components of Data Analytics

- Collecting
- Processing
- Statistical Data analysis

Tools used by a Data Analyst

- Database systems
  - SQL, NoSQL, Hadoop

- Analyzing data
  - Python
  - R
  - Excel

Why do I need Python, when I have Excel ?

- Excel is good for viewing data, performing basic analysis, and drawing simple graphs
- Audit trail of what was done is not known in Excel
- Would know exactly what has been done in python
- Repeatability
- Transparency and sharing
- Easier to clean data (missing data, dates in different formats, no headers)
- Analysis on big amount of data is easier
- Python is free and open source

Tools used in this class

- Anaconda
  - Jupyter Notebook
- We can run python in Jupyter Notebook
  - Execution of code can be done line by line

- If you do not want to use Jupyter Notebook you can use
  - Google Colabatory (online)
  - Colaboratory is a research tool for machine learning education and research.
  - It’s a Jupyter notebook environment that requires no setup to use.(Chrome /Firefox)
  - Access to Google GPU (limited)

Resources

- https://data.bsa.org/
- https://github.com/jakevdp/PythonDataScienceHandbook
- http://pandas.pydata.org/pandas-docs/stable/10min.html
- https://www.kaggle.com/
- http://www.data.gov.my/
- https://www.anaconda.com/
- https://colab.research.google.com/