HACKWAGON
• ACADEMY •

DATA SCIENCE 102:
DATA ANALYTICS WITH PYTHON

AGENDA

- Data Analytics Vs Data Science
- Data Landscape Today
- What DS102 Covers

WHAT IS DATA ANALYTICS

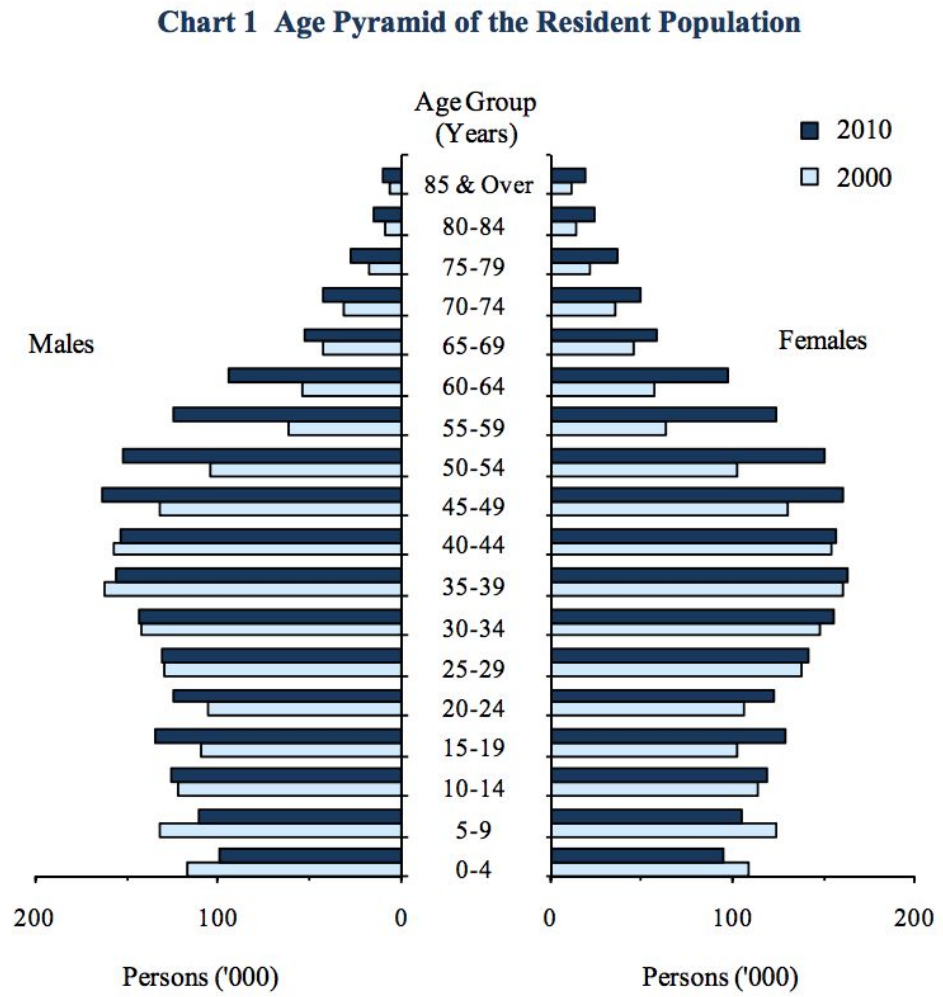
- Data Analytics is the process of **collecting** & **extracting** meaningful **insights** from data

- These insights aid in business decision making, that help to **support a business strategy**

- Business Strategy includes but is not limited to:
  - Private Entities
  - Government Bodies
  - Social Sector Enterprises & Organisations
  - Non Profit Oragnisations

AGE PYRAMID

- The Agy Pyramid (Singapore Census, 2010) is a useful visualisation tool to observe aging trends. This will support discussions on healthcare & mobility needs in the future

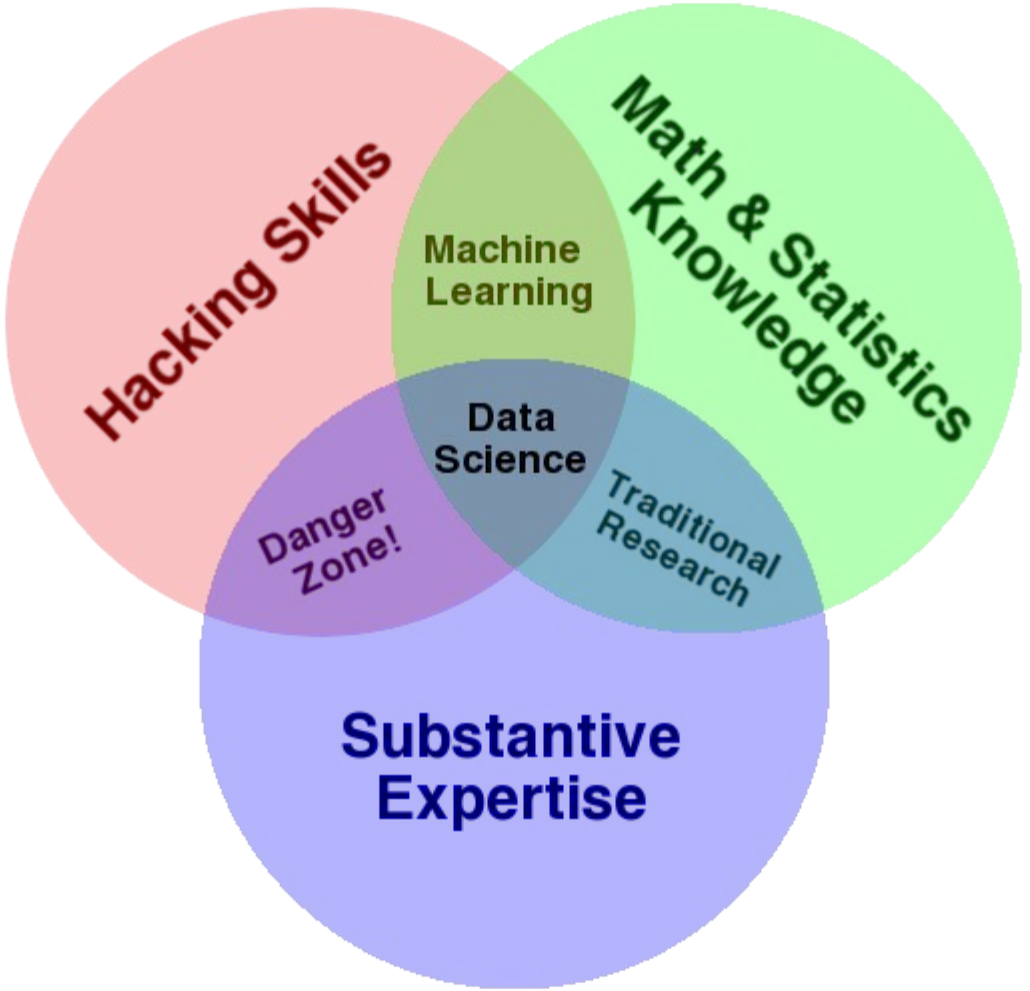


THE DATA SCIENCE VENN DIAGRAM

- **Hacking Skills**: Ability to code to develop algorithms to efficiently store, process and visualise data. *(What we teach you in DS101 & DS102)*

- **Maths & Statistics Knowledge:** Ability to form and apply mathematical models, and summarize these to any given datasets. (What we cover in DS102)

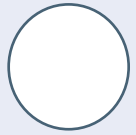
- **Substantive Expertise**: Domain expertise to form the right business scenarios and questions



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

WHAT IS DATA ANALYTICS & DATA SCIENCE

- **Descriptive Analytics**: summarize raw data and make it something that is interpretable by humans. It describe the past and answer: “What has happened?” (Example: Annual Sales for Shoes in a Adidas)
- **Predictive Analytics**: Using statistical models and forecast techniques to understand the future and answer: “What could happen?” (Example: Predicted sales for shoes in the next quarter)
- **Prescriptive Analytics**: Using optimization and simulation algorithms to advice on possible outcomes and answer: “What should we do?”
(Example: How to allocate stocks in the right stores to increase sales)
- **Machine Learning**: The science of getting computers to act without being explicitly programmed. Basically Machine Learning is concerned with the methods and models

	Data Analytics	Data Science
Descriptive		
Predictive		
Prescriptive		
Machine Learning		

Source: https://datascientistinsights.com/2013/09/09/data-analytics-vs-data-science-two-separate-but-interconnected-disciplines/

DATA SCIENCE PROJECT STAGES

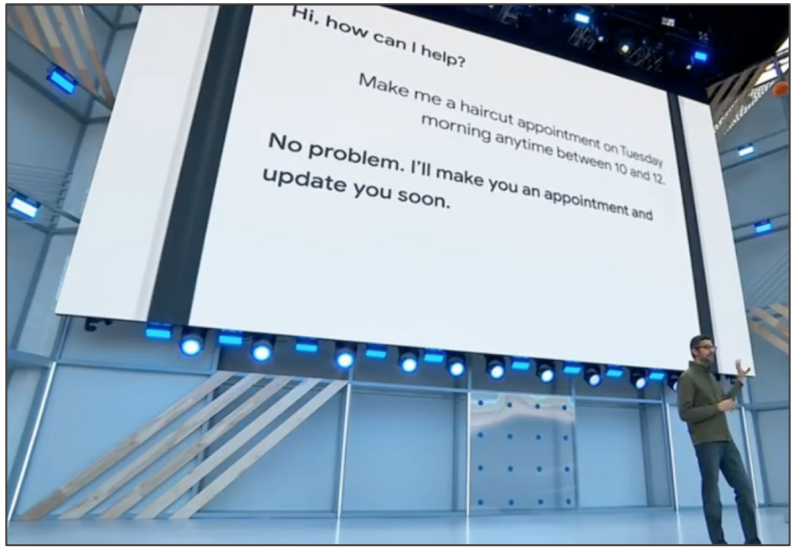
Problem Specification	Data Gathering & Preprocessing	Descriptive Analytics	Machine Learning	Deployment
• Understand business scenario • Define the project problem and scope • Define limitations of the project	• Develop a system to gather data • Clean and prepare raw data for processing • Usually the most time consuming stage in a data science project	• Exploratory Data Analysis • Basic understanding of the dataset • Answer the initial assumptions you may have of the data	• Depending on the scope/nature of your project, apply necessary machine learning models to tackle the problem • Train the machine learning model and assess its' performance	• Consult with project stakeholders on suitability of model • Deploy model for live usage

DATA LANDSCAPE

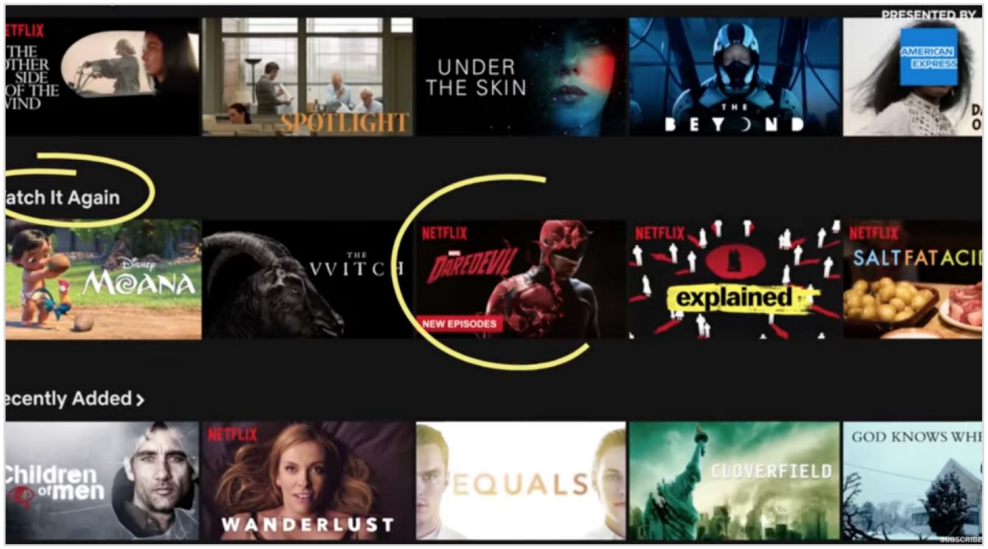
- Tech Companies
- Fortune 500s Take Notice
- Public Sector
- Data in the News

TECH COMPANIES LEAD THE WAY

Internet companies demonstrate how machine learning can be used in cutting-edge technologies in data, including the domains of voice and imagery



Google Duplex making a phone call



Netflix Recommender

Sources:
https://www.youtube.com/watch?v=axCBA3VD5dQ (Netflix A/B Testing with Thumbnails)
https://www.youtube.com/watch?v=D5VN56jQMWM (Google Duplex)

FORTUNE 500 TAKES NOTICE

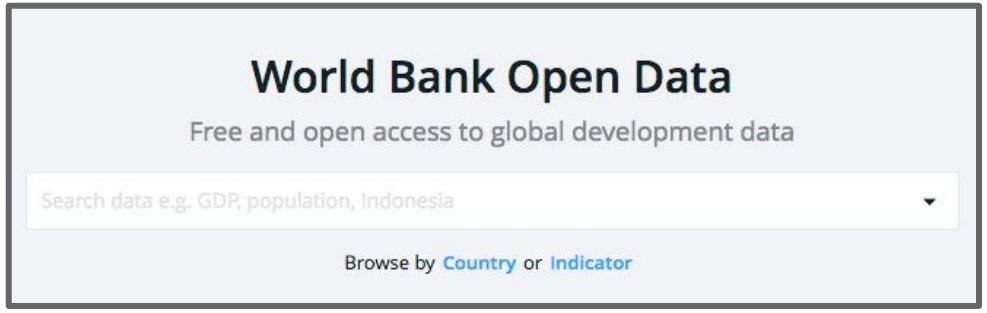
- Toyota invests US$1B in Grab to help develop connected services
- Disney uses AI to predict quality of short story narratives

Sources:
https://www.straitstimes.com/business/companies-markets/toyota-to-invest-us1b-in-grab
https://www.financialexpress.com/industry/toyota-invests-in-uber-rival-grab-to-extend-ride-share-foray/832901/
https://emerj.com/ai-sector-overviews/ai-at-disney-viacom-and-other-entertainment-giants/

PUBLIC SECTOR

Local government & international are providing publicly-available datasets for research in policy-making

Data.gov.sg, Singapore



World Bank Open Data, World Bank

Sources:
https://www.straitstimes.com/business/companies-markets/toyota-to-invest-us1b-in-grab
https://www.financialexpress.com/industry/toyota-invests-in-uber-rival-grab-to-extend-ride-share-foray/832901/
https://emerj.com/ai-sector-overviews/ai-at-disney-viacom-and-other-entertainment-giants/

DATA IN THE NEWS - Amazon Scraps AI Recruiting Tool

- **Amazon scraps AI recruiting tool -** Amazon realised the algorithm did not rate candidates in a gender neutral way as it was learning from resumes mostly submitted by men
- **Google urged to drop 'Dragonfly Project' -** 'Dragonfly' would be a censored version of the search engine developed with the aid of the Chinese government, enable censorship of certain topics

Source:
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
https://www.bbc.com/news/technology-46357008
https://www.bbc.com/news/technology-46604085

WHAT WILL BE COVERED IN DS102

- Course Content
- Modes of Assessment

COURSE CONTENT

- Exploratory Data Analysis with Pandas
  - Data Wrangling
  - Numerical
  - Categorical
- Web Scraping
- Data Visualisation
- Text Mining
- Machine Learning

COURSE OUTLINE

MODES OF ASSESSMENT

- Problem Sets (60%)
  - 3 Problems Sets
- Project (40%)
- Grading
  - 70% or above - Certificate of Distinction
  - 60% or above - Certificate of Merit
  - 0% to 60% - Certificate of Pass

PROBLEM SETS

- Problem Sets (P-Sets) are used to apply the concepts learnt during class
- Each P-Set score is merely **a form of feedback** to give you a measure of your understanding of the material.
- We encourage you to form study groups to discuss the problems and to approach the TAs and instructors when you are stuck

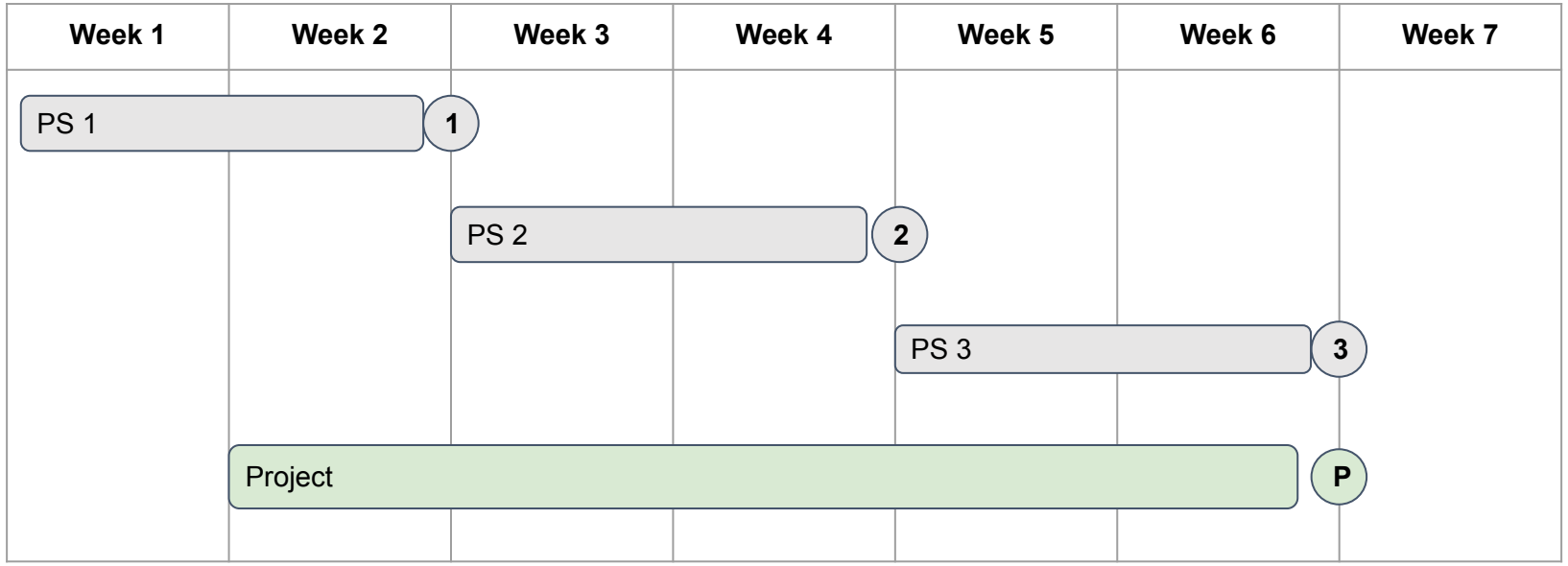
PROBLEM SETS

- 3 Problem sets, every 2 lessons
- Submission before next class starts (2359hrs)
  - Submissions after the class scores are halved
  - Submissions 7 days late and beyond are zero
  - Do as much as you can
  - Individual submission

PROJECT

- Exploratory Data Analysis on an Open Data Source
- Submission: **pairs or up to a team of 4**, **within the class**
- Project Timeline
  - Lesson 2 - Project Release
  - Before Lesson 5 - Proposal Submission (Optional)
  - Before Lesson 7 - Final Submission

DS102 WORKLOAD



Duration to Complete Problem Set	3	Problem Set Submission
Duration to Complete Project	P	Project Submission