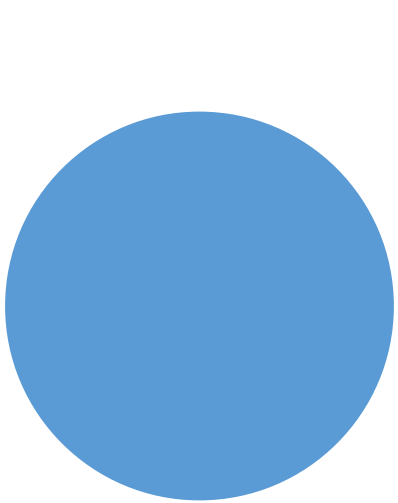
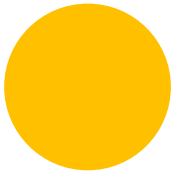



What is Data Science?

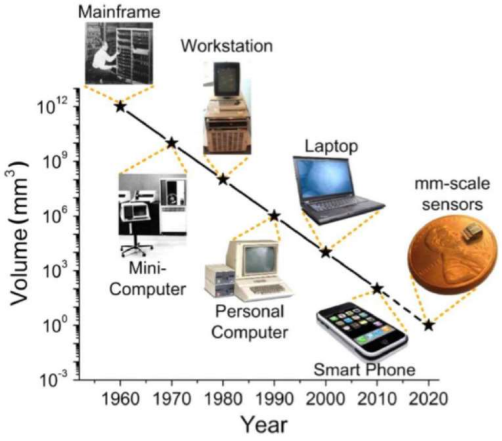
Pravin Pawar

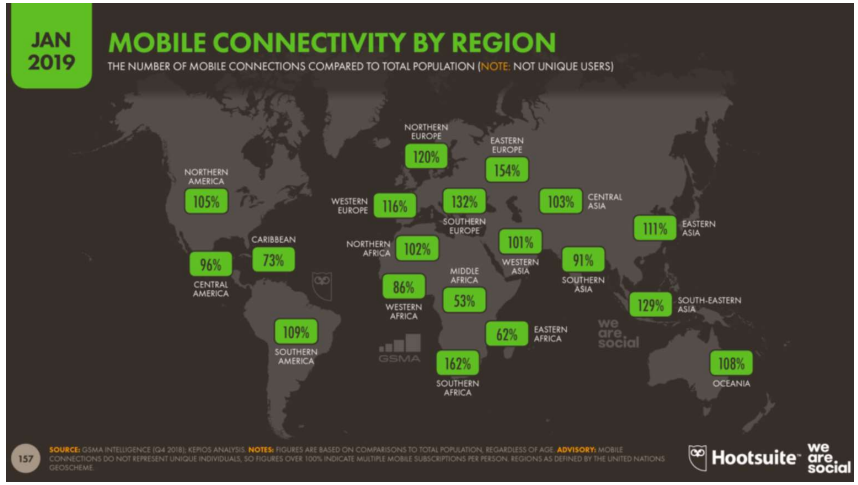
Terminologies

- As per Oxford dictionary data is:
  - Facts or information, especially when examined and used to find out things or to make decisions.

- While science is:
  - Knowledge about the structure and behavior of the natural and physical world, based on facts that you can prove, for example by experiments.

The data explosion in 21st century





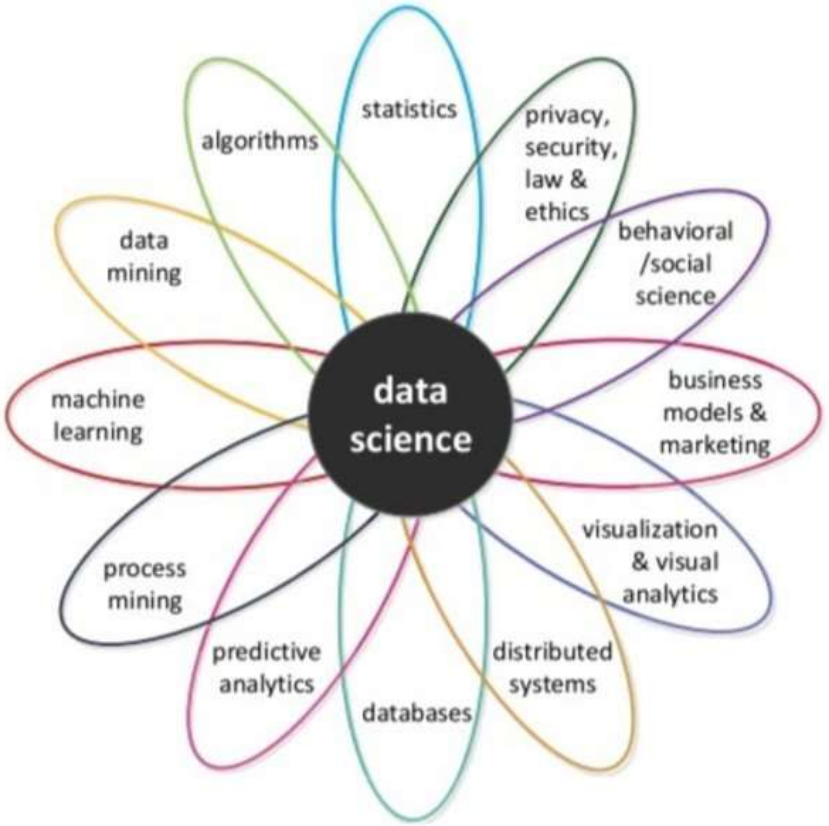
https://www.internetlivestats.com/

- Figure 1 source: https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/
- Figure 2 source: https://wearesocial.com/global-digital-report-2019

Data Science

- Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

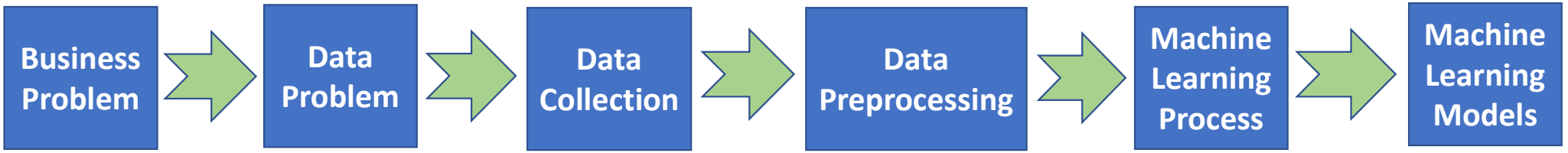
- The term was first coined in 2001 in an article by William S. Cleveland and its popularity has exploded since 2010*.



- Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. International statistical review, 69(1), 21-26.

Data science and machine learning

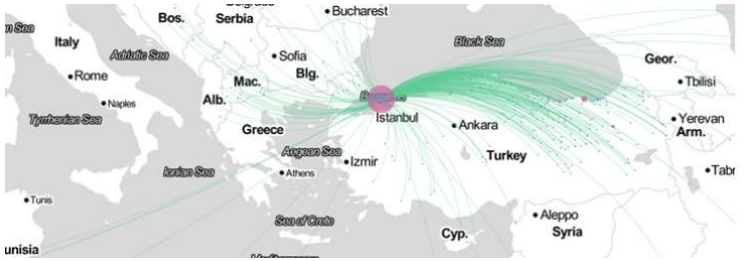
- Many people imagine that data science is mostly machine learning.
- However, data science is mostly about solving business problems.



- Some machine learning processes
  - Regression: A statistical model to predict numeric or continuous data.
  - Classification: Predict categories (labels/classes) of the data.
  - Clustering: Identify groups of similar objects in a multivariate data set.
  - Associative rules: Discovering interesting relations between variables in a data set.

Some case studies will surprise you!!

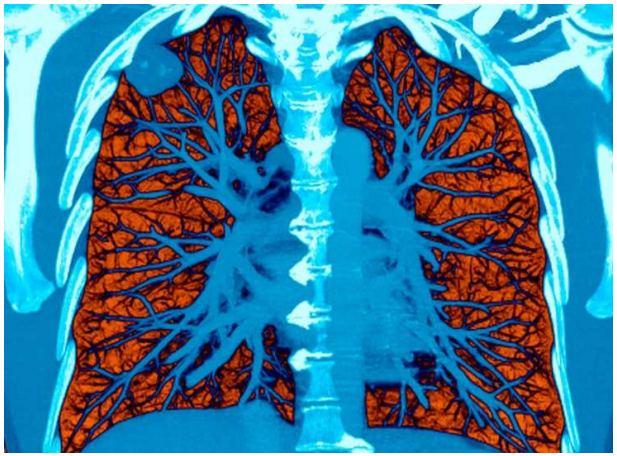
- Facebook asks users to list hometown and current location.
- Analyzes these locations to identify global migration patterns.
- Coordinated migration: A significant proportion of the population of a city has migrated, as a group, to different city.
- Examples of international coordinated migrations:
  - Migration from Cuba: Individuals who emigrate from Cuba are most likely moving to Miami.
  - Migration from Mexico: Several destination cities (Chicago, Houston, Dallas, LA).
  - Istanbul: A large proportions of emigrants from Turkey, but also from East Europe.



- Source: https://www.facebook.com/notes/facebook-data-science/coordinated-migration/10151930946453859.

Some case studies will surprise you!!

- Rayid Ghani was chief scientist on President Obama's re-election campaign turned to using data science for social good.
- 48 data scientists worked together for 12 weeks to tackle social problems.
- A group devised a new way for the world bank to flag contracts where corporate collusion is most likely to occur.
- Another group helped pinpointing tens of thousands of housing units where kids are at the risk of lead poisoning.
- Interpret medical images such as MRIs and X-rays to detect tumors, artery stenosis and organ anomalies.



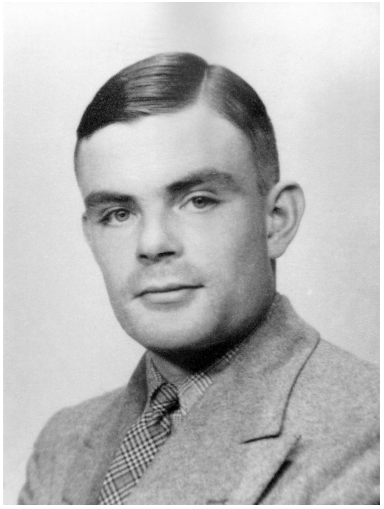
- Source: https://www.marketplace.org/2014/08/22/beyond-ad-clicks-using-big-data-social-good/.

Some case studies will surprise you!!

- Target figured out a teen girl was pregnant before her father did.
- Target assigns shopper a unique ID to keep track of their shopping habits.
- Target statistician Andrew Pole analyzed buying data for all the ladies signed up for Target baby registries.
- He identified 25 products which allow him to assign the shopper a “pregnancy prediction” score (and also estimate her due date within a small window). E.g.
  - Fictional target shopper Jenny of age 23.
  - Bought cocoa-butter lotion in March.
  - A purse large enough to double as a diaper bag.
  - Zinc and magnesium supplements.
  - 87% chance of being pregnant and delivery date in August.

- Source: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#5cabf5266686.

The Turing Test



- Alan Turing (1912-1954) was an English mathematician who laid some of the important theoretical groundwork of computer science
- In addition to other topics, Turing was interested in the idea of computers being able to think as human beings do
- He devised what he called the **imitation game**, now known as the **Turing test**
- A human judge (the interrogator) engages in an online chat with another person and a computer, but isn't told which is which
- If the interrogator cannot tell which is the person and which is the computer, then the computer has passed the Turing Test because it is simulating human intelligence
- So the Turing Test touches on two important areas of computer science: **artificial intelligence** and **natural language processing** (NLP)
- Google AI passes Turing test

Skill set of a data scientist

- Programming languages
  - R (Dplyr, ggplot2, Mlr, Caret, DataScienceR).
  - Python (Scikit-Learn, PyTorch, TensorFlow, Pandas, NumPy).
- Platforms
  - Jupyter Notebook.
  - Hadoop (data exploration, filtration, sampling and summarization).
  - Apache Spark (In memory computation, run complicated algorithms faster).
  - Amazon webservices.
- Machine learning tools
  - Weka, Orange, RapidMiner, KNIME, Neural Designer.
- Data visualization tools.
- Working with unstructured data.
- Knowledge of statistics.
- Business acumen, communication skills, teamwork.

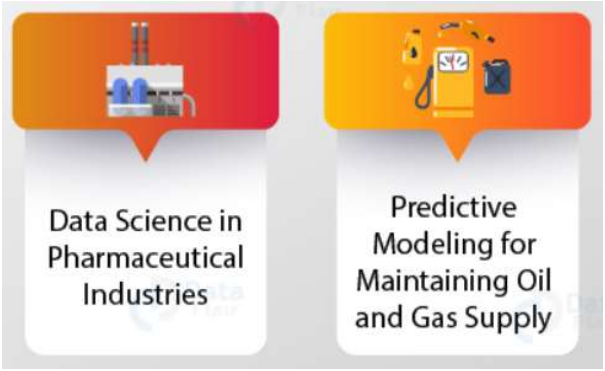
A note on the student projects

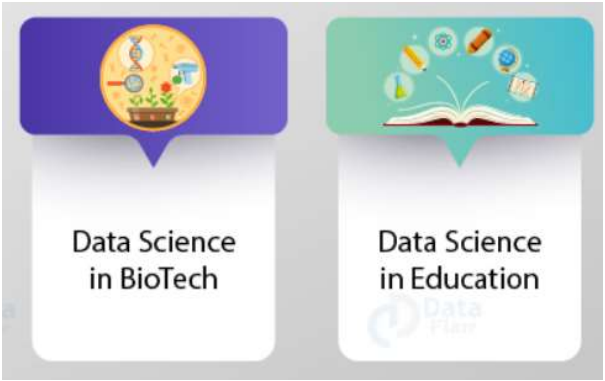
- Business analytics
- Business logistics, including supply chain optimization
- Finance
- Health, wellness, & biomedicine
- Bioinformatics
- Natural sciences and agriculture
- Information economy
- Social media and social network analysis
- Smart cities
- Education and electronic teaching
- Energy, sustainability and climate
- Weather prediction
- Astronomy
- Do a scholar search and select research paper or case study for presentation
- Keywords: Data science, machine learning, deep learning, data mining

Google Scholar

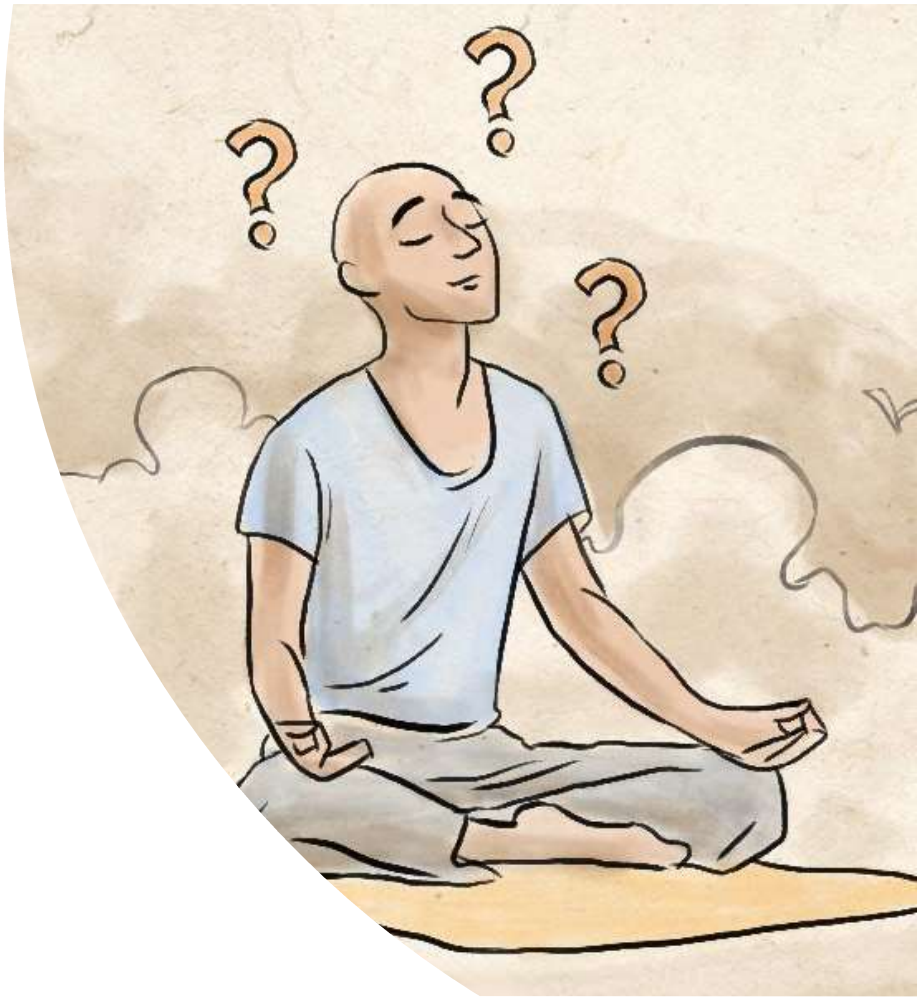
Links to some data science case studies

- https://data-flair.training/blogs/data-science-at-netflix/
- https://www.analyticsvidhya.com/blog/tag/case-study/
- https://data-flair.training/blogs/data-science-in-retail/
- https://towardsdatascience.com/ml-case-studies/home
- https://www.analyticsvidhya.com/blog/2016/10/complete-study-of-factors-contributing-to-air-pollution/
- Choose a comfortable team size (2 per team)
- Prepare and deliver a presentation on Friday 10th January – 10 minutes per group
- Group homework: Watch videos of data science projects posted at http://www.quant-shop.com/



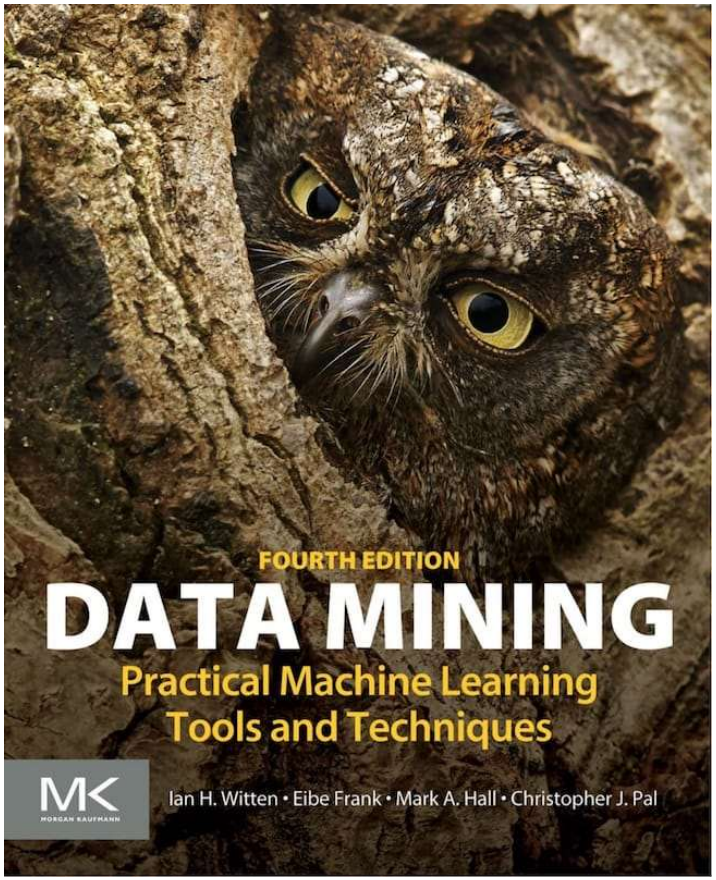


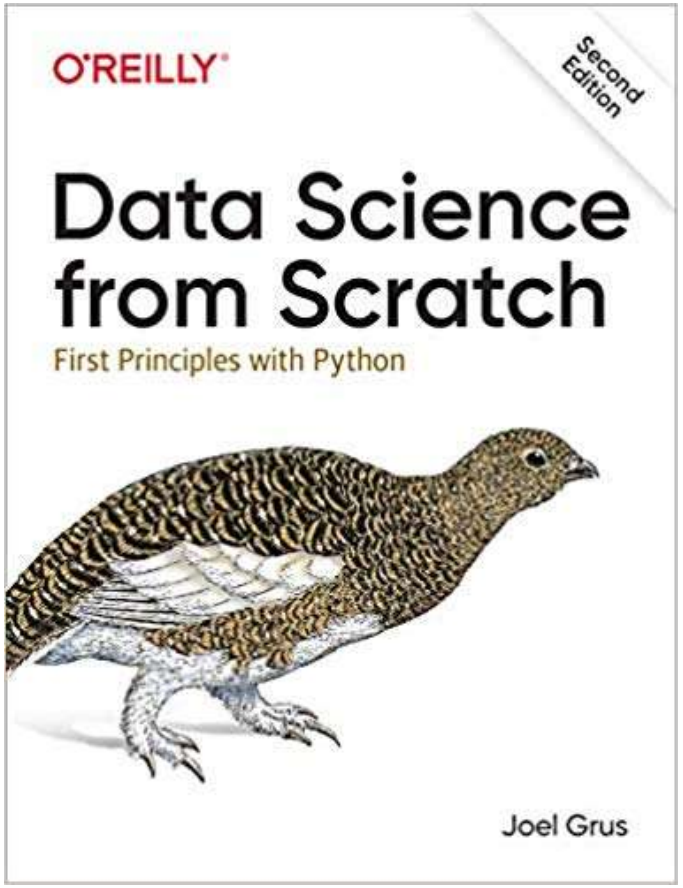
Paradigms related to data science



- Data mining
- Machine learning
- Artificial intelligence
- Predictive analytics
- Business analytics
- Statistical analysis
- Data visualization
- Big data
- Natural language processing
- These are all somewhat inter-related terms with some differences

Text-books for this course

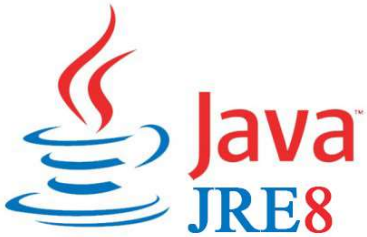




Software used in this course

https://www.cs.waikato.ac.nz/ml/weka/

https://www.python.org/downloads/

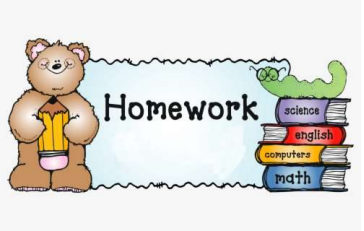


https://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html

https://www.jetbrains.com/pycharm/

Python programs link: https://drive.google.com/file/d/1zhDj0GMvncO7FUnFjzi6vV_cVMwwIKw8/view?usp=sharing

Today's homework



- Choose the group partners convenient to you and give your group a name that starts from A - H.
- Go to the site www.quant-shop.com.
- Watch presentations one per group.
  - (A) Miss Universe.
  - (B) Movie gross.
  - (C) Baby weight.
  - (D) Art auction price.
  - (E) White Christmas.
  - (F) Football champions.
  - (G) Ghoul pool.
  - (H) Gold/oil prices.
- Answer the following question tomorrow (5 minutes per group):
  - What is the quant-shop presentation about? How data science is used to solve the problem?
  - Which data science case study you will be selecting for the presentation on Friday 10th?