



What is Data Science?

Pravin Pawar
CSE351/519 – Introduction to Data Science
SUNY Korea.

1

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.

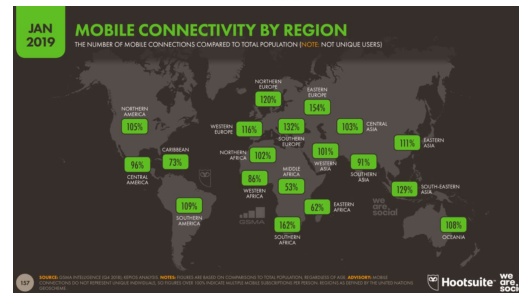
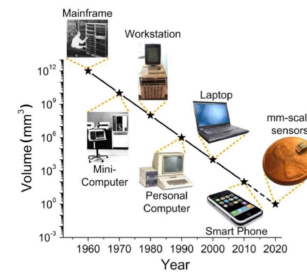



2

The data explosion in 21st century



<https://www.internetlivestats.com/>

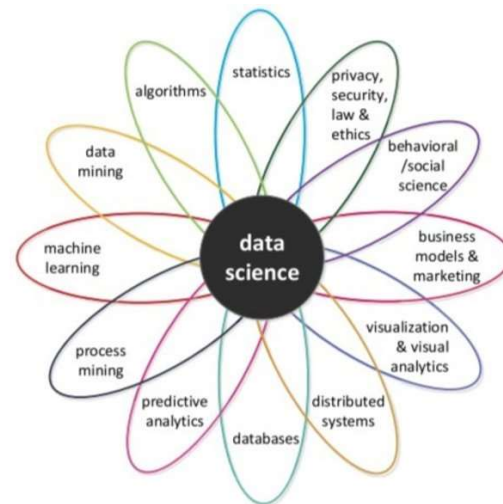


- Figure 1 source: <https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/>
- Figure 2 source: <https://wearesocial.com/global-digital-report-2019>

3

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.

4

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.

5

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.



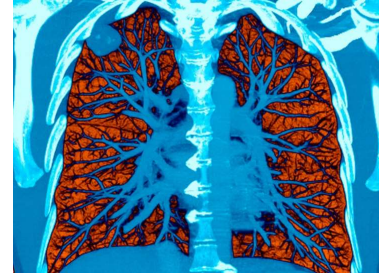
- Source: <https://www.facebook.com/notes/facebook-data-science/coordinated-migration/10151930946453859>.

6

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/>.

7

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>.

8

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](#)

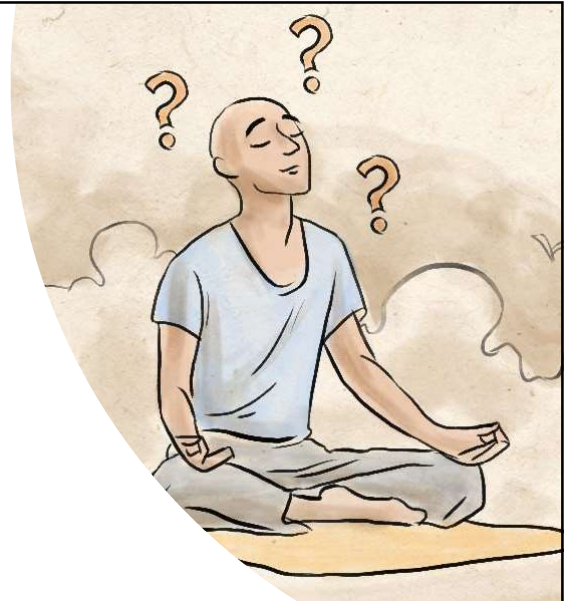


9

9

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



10

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/>



13

Homework

(To be presented on Wednesday 11 March)



- Choose the group partners convenient to you and give your group a name that starts from A - E.
- Go to the site www.quant-shop.com.
- Watch presentations one per group (No overlapping of topics).
 - (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.
- Present your report on Wednesday 11 March (6 mins per group + 2 mins QA):
 - What is the quant-shop presentation about?
 - How data science is used to solve the problem?

14