Machine Learning

Fields around the Data

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

"Data is the new oil" By Clive Humby

- BUT!
- Data by itself is **messy** and must be refined and analyzed to be useful!
  - Oil also is not so useful unless refined and changed into usable gaz, chemicals, etc
  - **Messy data**: heterogeneous values, missing entries, and large errors

Data is among the world's **most valuable resources**

(Why)?

- Because of how much potential **revenue and business value** it can provide
- Companies **collect, understand, use, sell and buy data**
  - Data Monetization
- “If you’re **not paying** for the product, then **you’re the product**”
  - Netflix documentary, The Social Dilemma

Importance of data

- If there is no data, there is will be no data-centered fields
- Services compete on **collecting data** about you
  - Your location, visits, habits, what you buy, what you watch, for how long
- They **use** these to learn how to change your behaviour
  - Grap more your attention such as reels/shorts, netflix recommendations, products
  - More interesting social media posts to stay more on the platform
    - Then collect more data
    - Show more **ads** that you most probably click!

Question!

- Nowadays, there are many services that offer very short clips for the users
  - Facebook reels, Youtube shorts, Tiktok shorts
- Assume 2 users uploaded the following videos:
  - Video 1: A woman shows how to use a new brand for Women's Electric Shaver
  - Video 2: A woman in a bikini behaving in a sexual way
- Ali and Sarah are **twins** at the age 13. Today is their first existence on the internet and they created accounts on a social media providing basic information like age, address and education
  - Which video will Ali most probably see? What about Sarah?

Question!

- While your mum walking in the home, she noticed 2 Facebook reels with almost naked women (or violence)
- Your mum said these things are on your page because you keep looking for such things!
- Is this a correct judgment?
- If not, what to do to build more correct judgement?

2.5 **quintillion** bytes of data were created every day. (SG Analytics, 2020) ... so?

- There is an extremely large amount of data uploaded every minute/day
  - People sent *500 million tweets* daily. (TechJury, 2020)
- This large scale data (**big data**) makes many old **analysis/modeling/storage** techniques useless!
- However, still many domains generate small-to-medium data

Data Types

- Structured data is **quantitative** such as integer values (age), real values (weight) , category (male), dates, and names
  - It can be stored in databases easily (rows/cols in relational database)
- Unstructured data is **qualitative** data and includes text (e.g. email's **body**) , image, video (set of frames/images), audio, etc.
  - Sometimes, we just store them as they are
    - Or transform to a format that is fast to read/store
  - Sometimes stored in NoSQL database
- Many data in the world are messy unstructured data
- Deep learning revolutionizes problems based on text, image and audio

Fields around The data

- Due to its importance, there are many fields around the data (*overlapping*)
- Data mining
  - Find **information**/patterns in **structured** data
- Data analysis / Data analytics
- Data Science
  - Find **insights**, answer questions in (un)**structured** data and do **forecasting**
  - It involves data scraping, cleaning, visualization, statistics, etc
- Machine Learning
  - Build wide variety of models based on the data and facts (gt) available about it
- *The last 2 fields are major in the data domain nowadays. Why?*

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."