Introduction to Big Data

Pooya Jamshidi

pooya.jamshidi@ut.ac.ir

Ilam University

School of Engineering,
Computer Group

February 15, 2025

Class Info

- **When:**
  - Saturdays 8:00am – 9:30am
  - Mondays 9:30am – 11:00am
- **Where:** Class 301

Main References

- Course handouts.
- Main textbooks:
  - J. Leskovec, A. Rajaraman, and J. Ullman, *Mining of Massive Data Sets*, Cambridge University Press, **2020**. (mmds.org)
  - M. Kleppmann, *Designing Data-Intensive Applications*, O'Reilly Media, Inc., **2015**.
  - J. Densmore, *Data Pipelines Pocket Reference*, O'Reilly Media, Inc., **2021**.

Prerequisites

- Generally, this course do not have any **official prerequisites**, but you're expected to be familiar with the following subjects:
  - **Python programming language** & **Linux operating system**
  - **Database** or a working knowledge of using **SQL**
  - Familiarity Computer networks or distributed systems
  - Familiarity with machine learning and deep learning concepts

Course Content

- **Data Mining**
  - Crawling and targeted crawling
- **Docker Essentials**
- **Batch Processing, Map-Reduce and the New Software Stack**
  - Unix tools
  - MapReduce / Hadoop
  - Hive and Presto
  - Apache Spark
  - Elasticsearch / Solr

Course Content (Cont.)

- **Storage and Retrieval**
  - Data structures
  - Technologies (NFS, GFS & HDFS)
- **Encoding and Evolution**
  - Encoding data formats
  - Dataflow modes
- **Mining Data Streams**
  - Apache Kafka & Apache Flink

Course Content (Cont.)

- **Data pipelines**
  - Patterns
  - Apache Airflow
- **Link Analysis**
  - Google PageRank
- **Advertising on the Web**
- **Mining Social-Network Graphs**
- **Data Visualization**
  - Apache Superset
- **Finding Similar Items (if time allows)**
- **Clustering (if time allows)**
- **Data Privacy and Ethics**

**I may not get a chance to go over all these topics. These subjects may be presented out of order.*

Course Audience

- This course will be useful for AI and software students and more broadly for anyone involved in machine learning.
- More broadly, anyone interested in data science/engineering and data-driven decision making will benefit from this course.
- I aim to maintain a balance between practical and theoretical subjects.
- Be open to new subjects, and you'll enjoy this course

Grading

- **Assignments:**
  - Analytical homeworks (HW): $2 \times$ 1pt
  - Computer assignments (CA): $3 \times$ 2pts
  - Final project: 2pts
- **Exams:**
  - Pop quizzes: Up to 2pts
  - Mid-term: 5pts
  - Final exam: 5pts
- **Late assignments policy:** Late assignments will be penalized at the rate of **10% per day** or fraction thereof for the first **two days**. After that (two days), no late assignment will be accepted.
- **Exam dates:**
  - Mid-term exam time: ?
  - Final exam time: **1404/03/29 8:00am**

A Famous Tweet

Big Data Borat (@BigDataBorat)

"In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data."

6:17 AM · Feb 27, 2013 · Twitter Web Client
544 Retweets · 24 Quote Tweets · 394 Likes

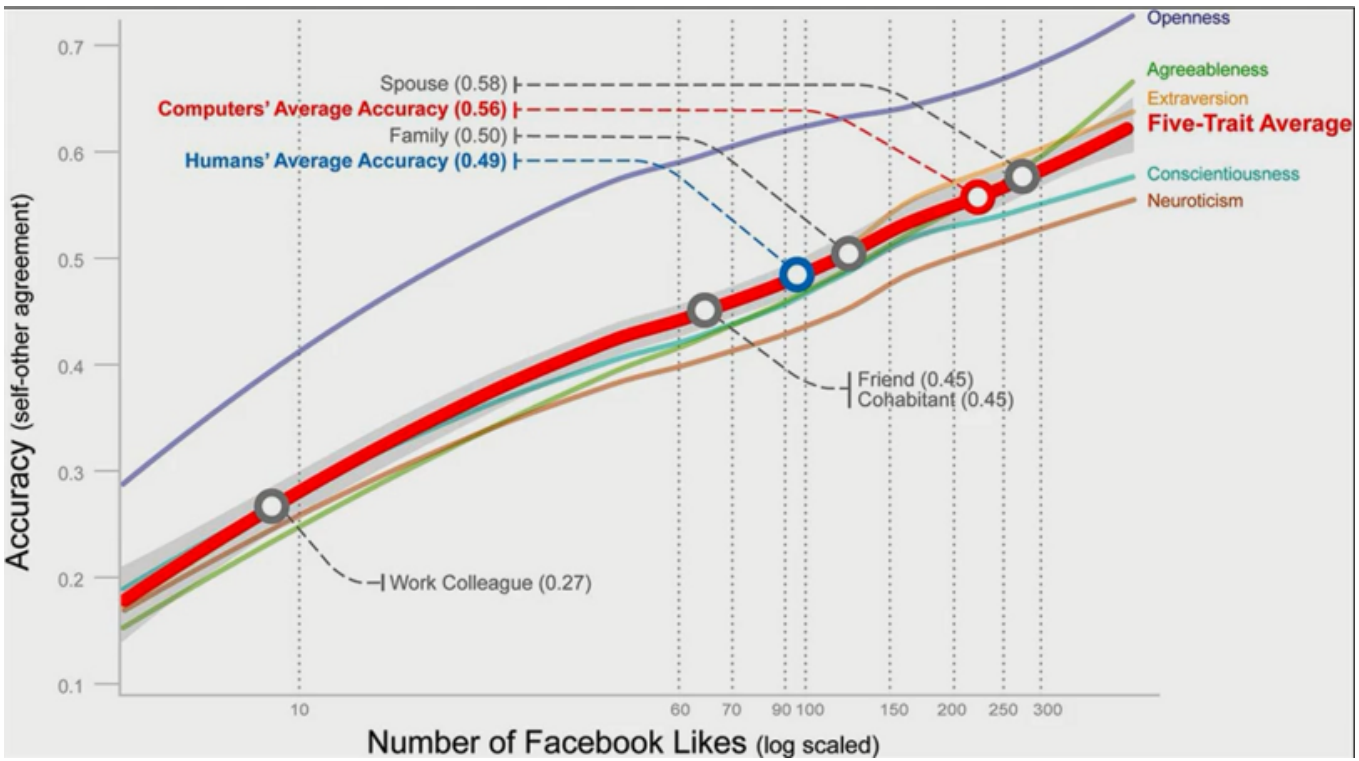
Experienced vs. Novice Machine Learning Engineer/Scientist

- Generally, an easy-to-spot difference between an **experienced** and a **novice** (yet knowledgeable) machine learning engineer/scientist boils down to understanding the following:
  1. **Data preparation:** Includes tasks like *guideline creation*, *working with data annotators*, *data cleanup*, etc.
  2. **Model deployment to production:** Understanding how to transition models from development to production environments.

What is big data?

- **How much data is actually considered big?**
  - 1 GB, 10 GB, 100 GB, 1 TB, etc.
- **Big data means your memory is small!**
- **How to handle big data?**
  - Sampling
  - Distributing
  - Streaming

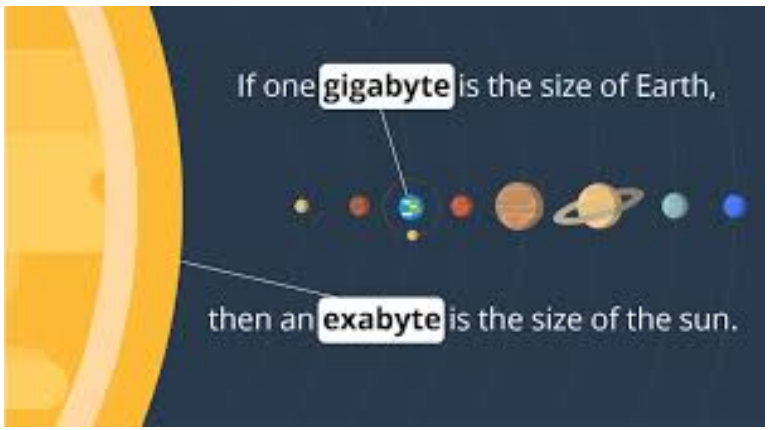
Why is data important?

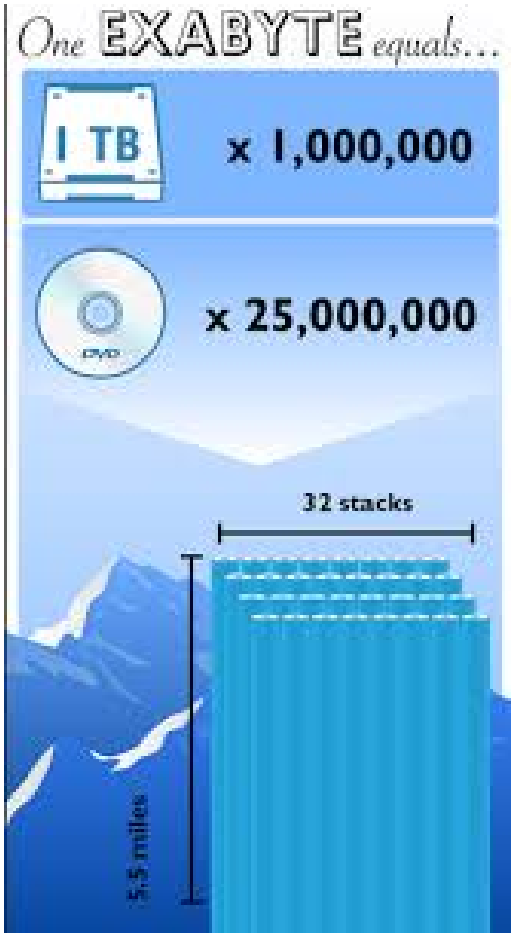


How much data does Facebook have?

- It contains an extremely heterogeneous set of data:
  - **Binary blobs:** e.g., photos & videos
  - **Textual data:** e.g., post contents
  - **Metadata:** e.g., impressions & metadata
- Facebook stores several **exabytes of data**, and the size grows exponentially.

Source 1 — Source 2


If one gigabyte is the size of Earth,
then an exabyte is the size of the sun.


One EXABYTE equals...
1 TB
x 1,000,000
DVD
x 25,000,000
32 stacks
5.5 miles


Some have speculated that **5 exabytes** likely equals all of the words ever spoken by humans.
Websites
Family Photos
Data Storage
Ransomware
Medical Files
Streaming Videos
Music
To have recorded 1 exabyte of data, you would have to have started a video call **237,823 years ago.**
That's about the time modern homo sapiens emerged on the planet.

Where is our Big Data?

Consider scenarios in Iran!