COM6002 Big Data Management

Introduction to big data management

Big Data

- What is big data?
  - **Big data** primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software.
  - https://en.wikipedia.org/wiki/Big_data
- When some people talk about **Big Data**, they simply mean the data they are working on
  - And the data is not large in volume

Characteristics of Big Data

- 4 V's of Big Data
  - Volume
  - Variety
  - Veracity
  - Velocity

Volume: Too much data to handle

- Example task:
  - Study the social sentiment about bitcoin since launch (2009)
- Data source: Twitter data archive
  - https://archive.org/search.php?query=collection%3Atwitterstream&sort=-publicdate
  - 5M text messages per day. Around 2GB per day. (Twitter stats: around 500M messages per day)

Variety: Multiple types / formats

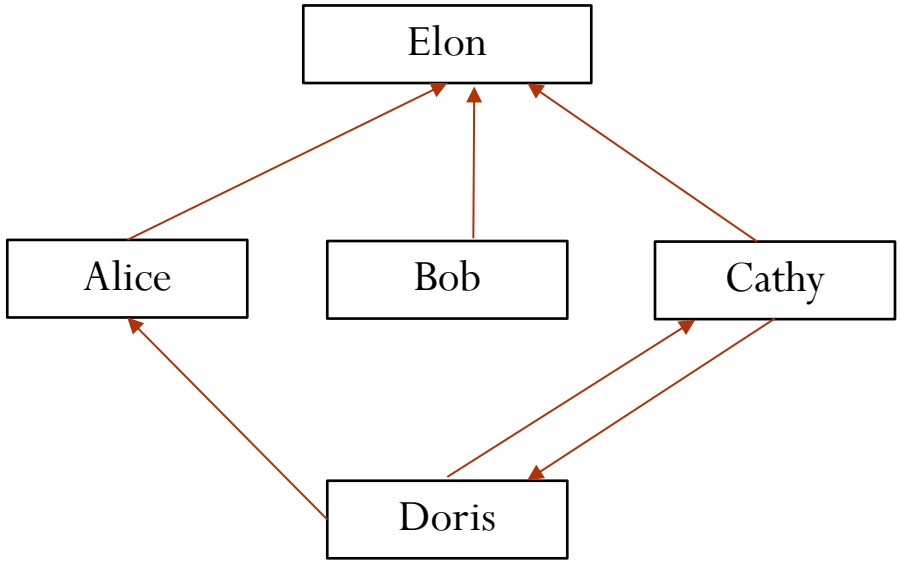
- What kind of data can you find on Twitter?

A sample tweet
(Twitter message)

- Multi-lingual text
- Images
- Videos
- …

Graph data

- Alice is following Elon, Bob is following Elon, …
  - Can be represented in a **graph**
- Graph analysis:
  - Who is a KOL (Key Opinion Leader)?



Veracity: data quality

- Is your data always correct?

Mechanism?

Terrorist?

GIGO - Garbage in, garbage out

- Data quality is very important!

- Other data quality concerns
  - Missing data
  - Duplicated data
  - Ambiguous data
  - Inconsistency

"There is a bird in a cage that can talk."
Who can talk?

Velocity: speed of data

- Big Data is often **real-time** data
  - Data is generated continuously
  - The data source does not always keep all the data for you to download later
- Example scenario: let's keep information about bitcoin on exchange platforms
  - Data source:
    - https://www.binance.com/en/futures/BTCUSDT
    - What data should we keep?
    - The trade queue every update?

Another example scenario

- Home security camera app
  - Will you keep the video data?

Factors Affecting Data Size

1. Resolution:
   - Common resolutions include 720p (HD), 1080p (Full HD), and 4K.
   - Higher resolutions generate larger files.
2. Frame Rate:
   - Common frame rates are 15, 30, or 60 frames per second (fps).
   - Higher frame rates increase the amount of data collected.
3. Compression:
   - Video compression formats like H.264 or H.265 can reduce file sizes significantly.
   - H.265 is more efficient than H.264.
4. Duration:
   - Continuous recording vs. motion-activated recording affects total data.

Estimation Example

Let's calculate an approximate data size for a 1080p camera recording at 30 fps with H.264 compression:

- **Resolution:** 1920x1080 pixels
- **Bitrate:** A typical bitrate for 1080p at 30 fps is about 4 Mbps (megabits per second).

Daily Data Calculation:
- **Data per second:** 4 Mbps = 0.5 MB/s (1 byte = 8 bits)
- **Data per minute:** 0.5 MB/s × 60 seconds = 30 MB/min
- **Data per hour:** 30 MB/min × 60 minutes = 1800 MB/hour = 1.8 GB/hour
- **Data per day:** 1.8 GB/hour × 24 hours = **43.2 GB/day**

https://www.quora.com/How-much-data-of-CCTV-camera-is-collected-one-day-in-GBs-and-where-it-store-its-heavy-data

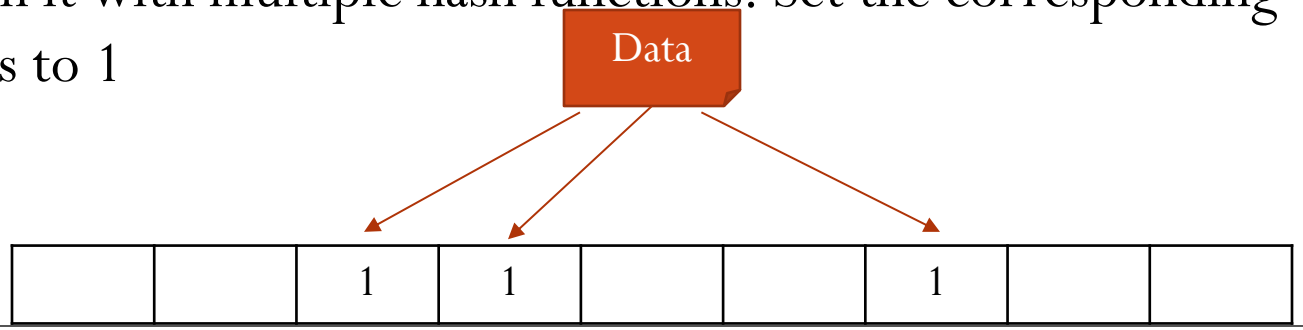
What data should we keep?

Streaming algorithms

- What if the data generation speed is far more than our storage capability?
- Streaming algorithms
  - https://en.wikipedia.org/wiki/Streaming_algorithm
- Key design issues:
  - How to extract key summaries and what to keep?
  - Can we find the exact answers to queries?
  - If it is an approximation, how close will it be? Any bounds on the performance?

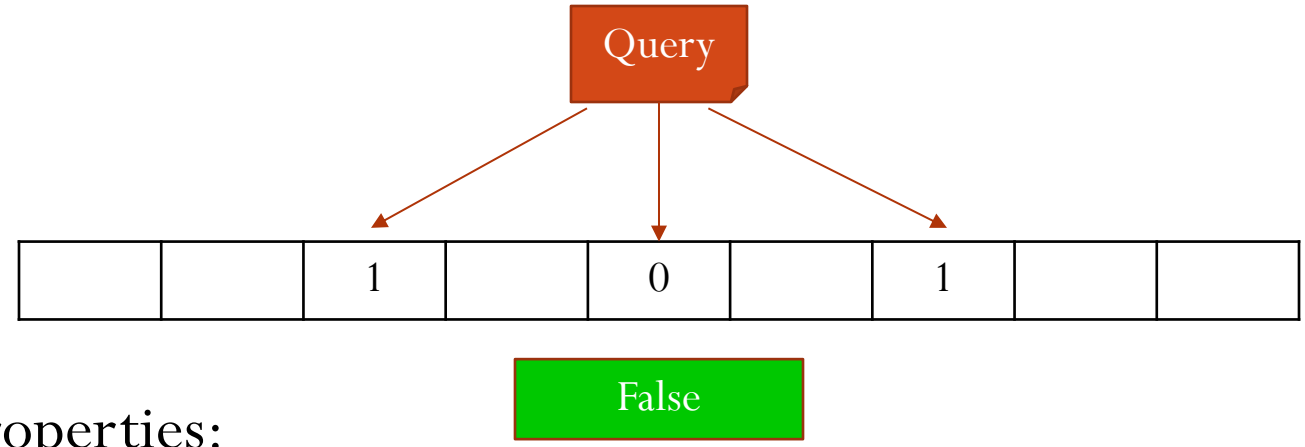
Example: Bloom filter

- Note:
  - The details of this algorithm will NOT be asked in the exam
  - It just gives you a feeling about how a streaming algorithm works
- Usage:
  - To check if a data item has appeared before
- A bloom filter is simply an array of 0 or 1
- Processing a data item
  - Hash it with multiple hash functions. Set the corresponding slots to 1



Example: Bloom filter

- Query (is the data item seen before?):
  - Hash the query with the same set of hash functions, return true if all slots are 1



- Properties:
  - There is no false negative
  - Bit array is space-efficient
- Q: Is bloom filter better than a simple hash table?

Why do we have Big Data?

- Internet!
  - Many data sources are available to the public
    - Example:
      - Social network: Twitter
      - Finance: Cryptocurrency market data
- Increased capability of machines, e.g., increased storage
  - We tend to store more data

1.44MB
1990's

256 GB
~ HKD $100

Google

15 GB
Personal
Free

Kingston DataTraveler Exodia M USB 256GB (DTXM/256GB)

容量: 256GB

HK$ 110 - 120 行

比較報價

The habit has changed!

- All of us are generating more data
  - Personal: Photos / videos
  - Business: more online services

History of data management

- Computer files
  - Baseline option for a digitalized copy of data
- Relational DataBase Management System (RDBMS)
  - 1970-now
  - Has a well-defined **table** structure
  - Many advantages over file storage, e.g., search efficiency, better backup and recovery
  - Main query language: SQL (Structured Query Language)

Example relational data

Date	Open	High	Low	Close	Volume
1/1	3.2	3.5	3.1	3.2	1.6M
2/1	3.2	3.3	2.9	3	1M
3/1	3	3.6	2.9	3.5	2M
4/1	3.5	3.9	3.1	3.6	2.3M

History of data management

- NoSQL database
  - NoSQL stands for Not-only SQL
  - Data is not well structured as a table
- There are different types of NoSQL databases
  - Graph database
  - Key-value store
  - Document store
- Example: MongoDB is a document store

Example MongoDB data

- Q: What happens if we put the data in a table?

Symbol: ETH
Date: 1/1
Close: 11.3
BTC Correlation: 0.77

Symbol: BTC
Date: 1/1
Open: 3.1
High: 3.3
Low: 2.9
Close: 3.1
Sentiment: 0.39
Basis: 0.0003
Whale transactions: 1056

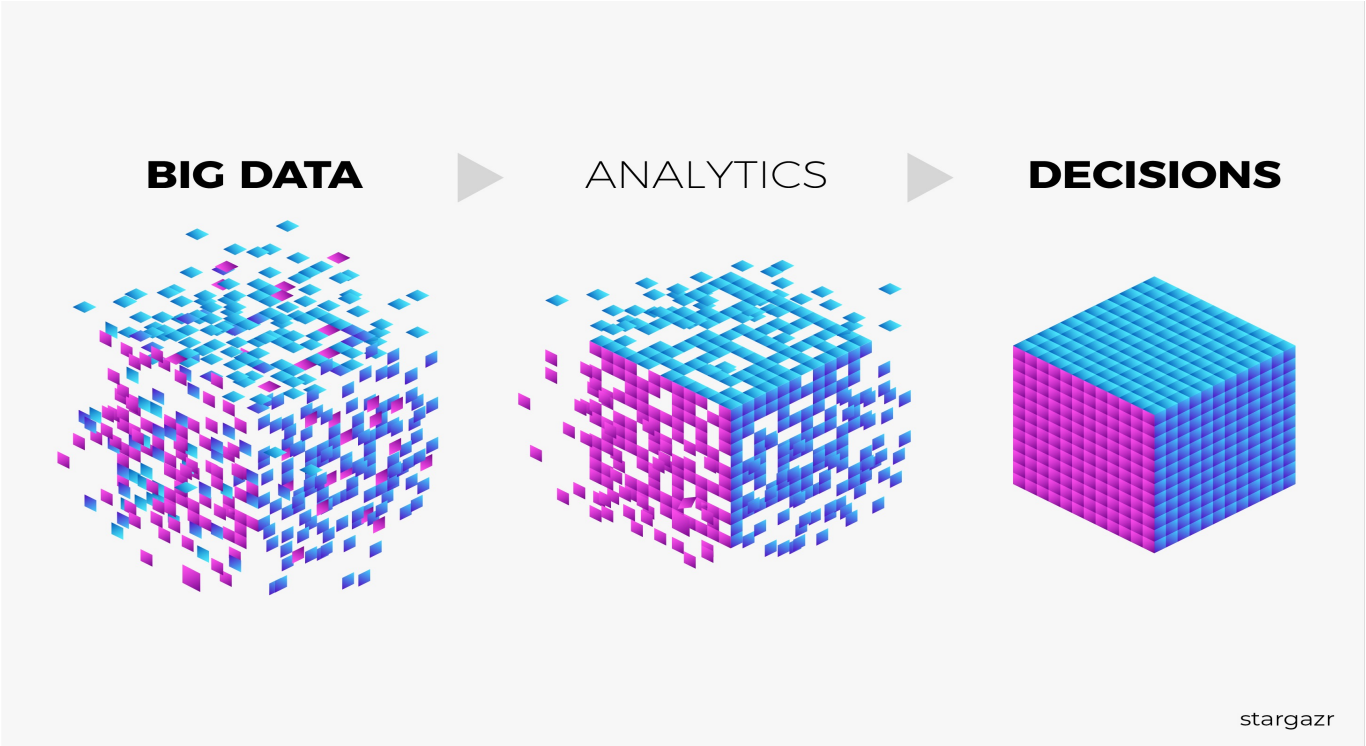
Symbol: DOGE
Date: 1/1
Close: 0.000065
Twitter mention: 25,694

Which type of database should we use?

- **<u>This is the main theme of this module!</u>**
- More details about RDBMS vs NoSQL will be discussed after we learn more about RDBMS and NoSQL in the coming lectures
  - Actually, we need to consider other options like HDFS as well…
- Pre-requisite for using RDBMS
  - Can your data fit in a tabular format?

Use of Big Data

- Big Data is almost understood as Big Data **Analytics**



Data analytics has a long history…

- Story of "Beer and Diapers" in early 1990's
  - Market basket analysis in a supermarket
    - Analyze what products customers purchase together
  - Common expected result:
    - Bread and butter
  - Something interesting:
    - Beer and diapers have been purchased together frequently

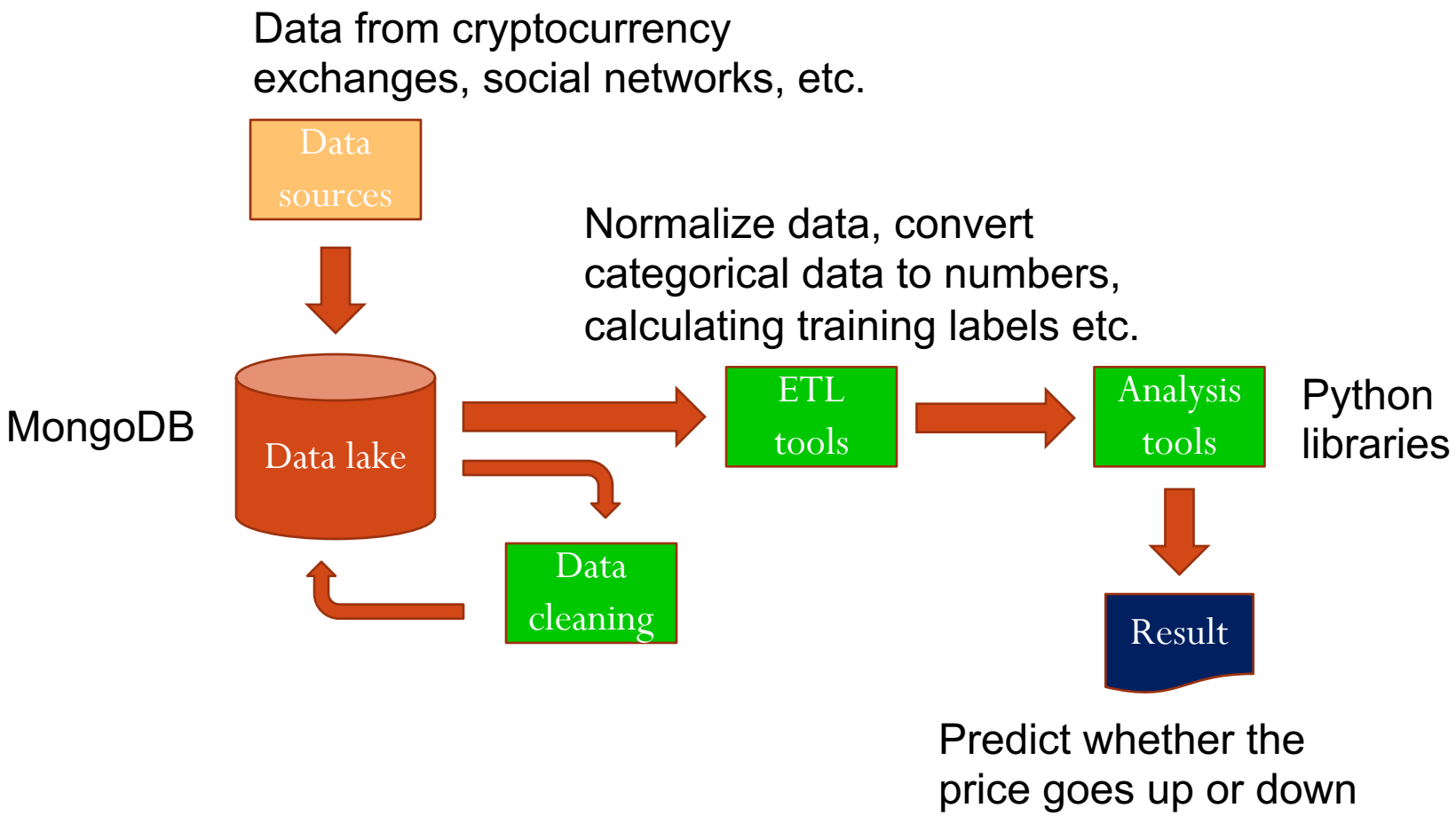
Diapers

- Nowadays, any data analytics tasks may be referred as Big Data Analytics
  - Is this a good sign?

Big Data Analytics tools

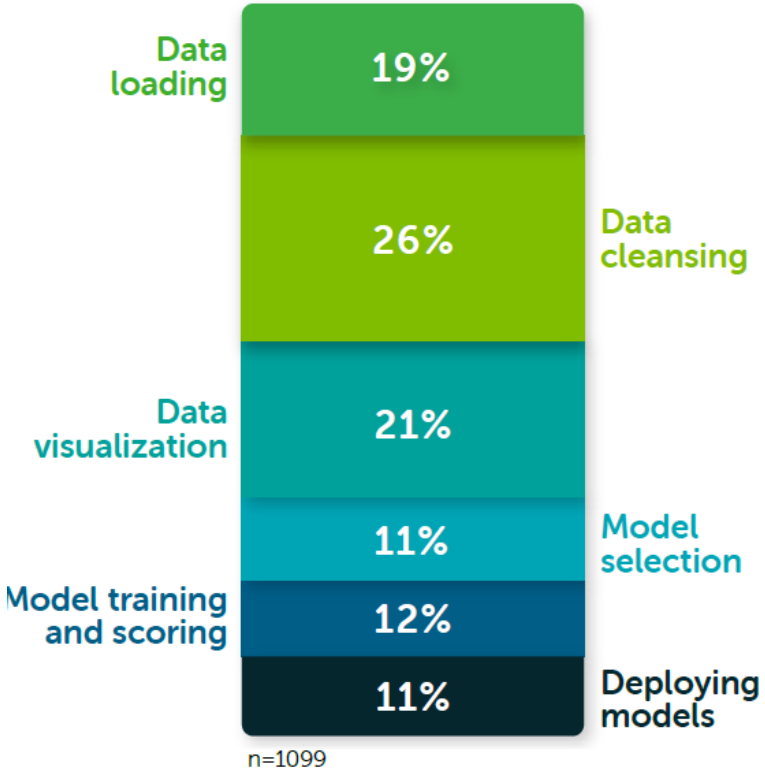
- Data cleaning
  - Improve the data quality of your data
- ETL
  - Extract-transform-load pipeline to preprocess your data for your analysis
- Analytics tools
  - Machine learning / Data mining / Artificial Intelligence

Big Data Analytics pipeline and example



Which part is more important?

- Go to google and find the results from similar surveys.



https://blog.ldodds.com › 2020/01/31 › do-data-scientis...

Do data scientists spend 80% of their time cleaning ... - Lost Boy

31 Jan 2020 — **Data scientists** spend 80% of their **time** cleaning **data** rather than creating

Why is data management important?

- https://towardsdatascience.com/your-ai-is-only-as-good-as-your-data-quality-42be9ab533b9

Praneeth Vasarla

May 16, 2021 · 9 min read · Listen

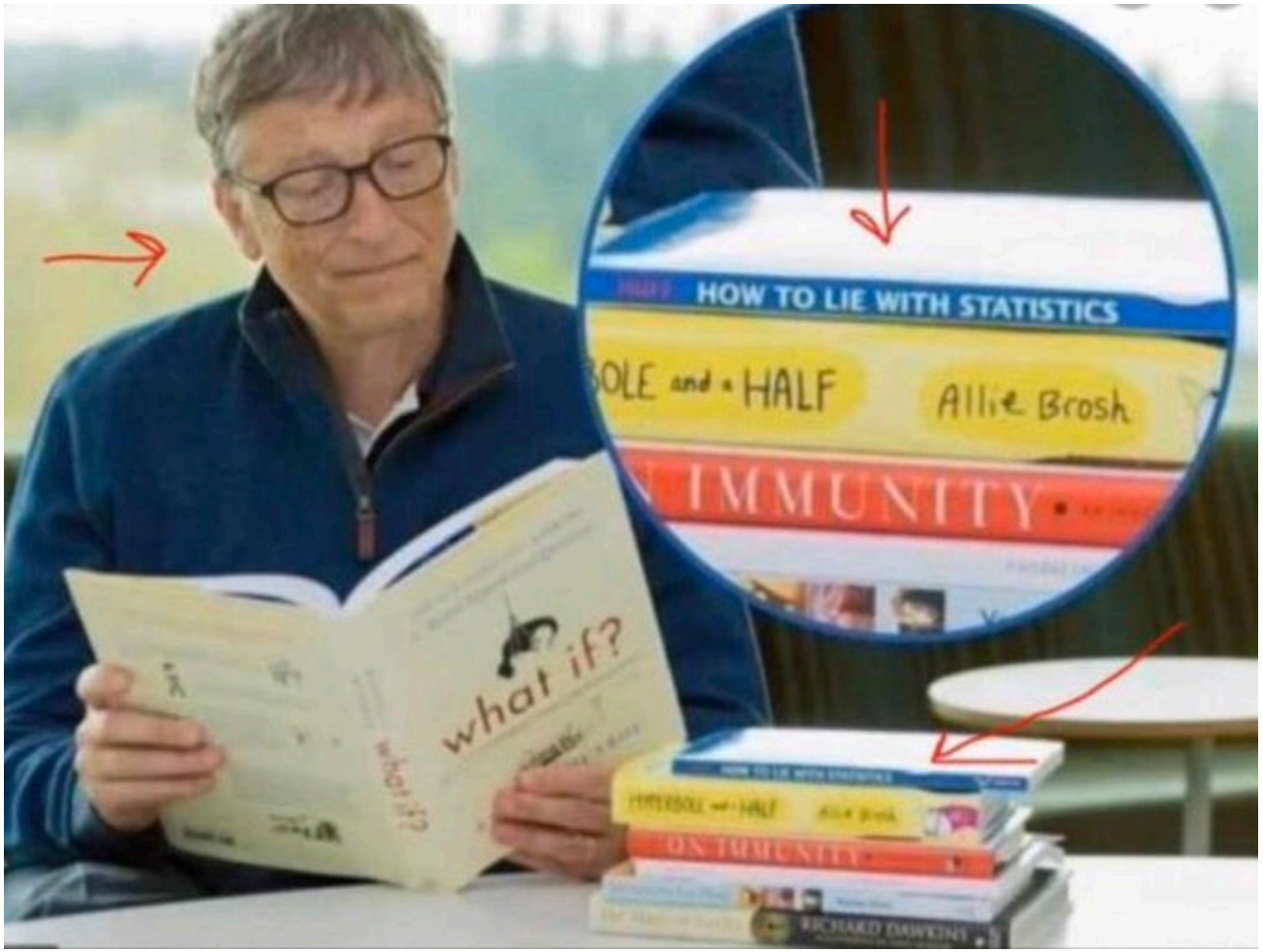
Your AI Is Only As Good As Your Data Quality!

The importance of data preparation in the process of model creation

University education vs work

- You will learn a lot of theories, maths, and so-called hands-on skills at the University
  - Most of time you spent at work is not about the above knowledge
  - NOTE:
    - Don't get confused. These are still very important! They make you a real professional
    - Always remember, a key to success of your data science / AI project is the data

Working with data



Good luck

- Big Data is not always a successful story

Google Flu Trends is dead – long live Google Trends?

By rmjlmcd, on 23 January 2018

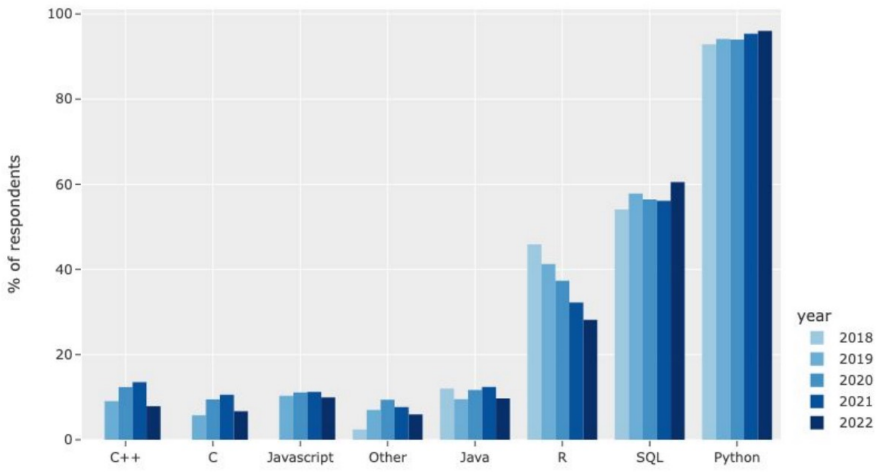
- The initial Google paper stated that the Google Flu Trends predictions were 97% accurate comparing with CDC data.
- In the 2012/2013 season, it predicted twice as many doctors' visits as the US Centers for Disease Control and Prevention (CDC) eventually recorded

Current trends in data science

- Kaggle Report 2022
  - No data collection exercise in Kaggle Report 2023
- Programming language

Kaggle DS & ML Survey 2022

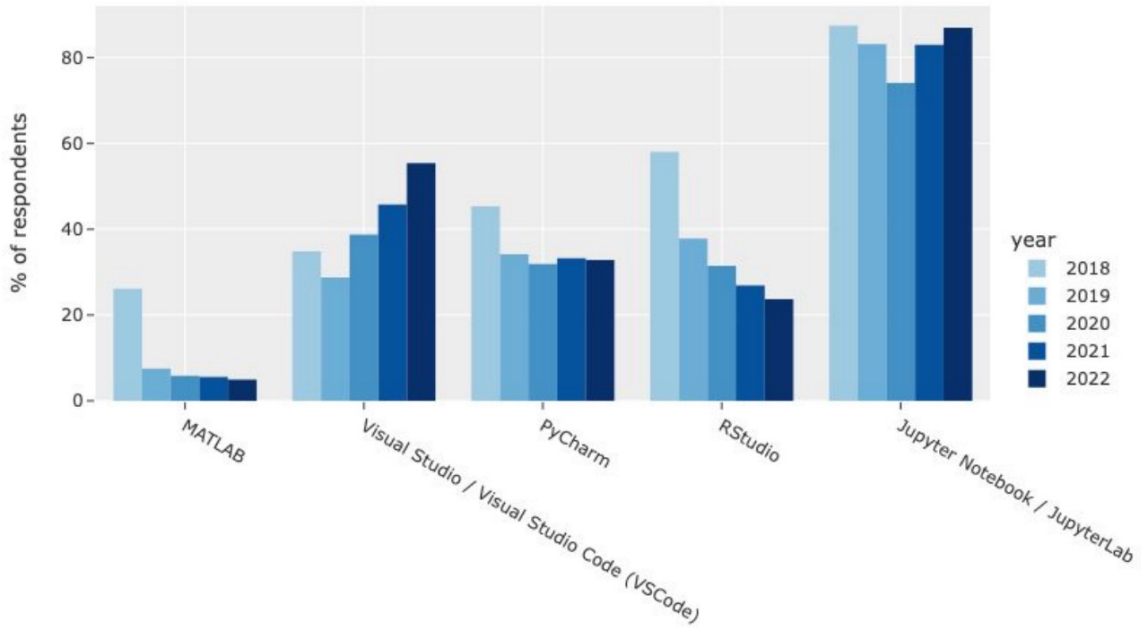
Python and SQL remain the two most common programming skills for data scientists



IDE (Integrated Development Environment)

Kaggle DS & ML Survey 2022

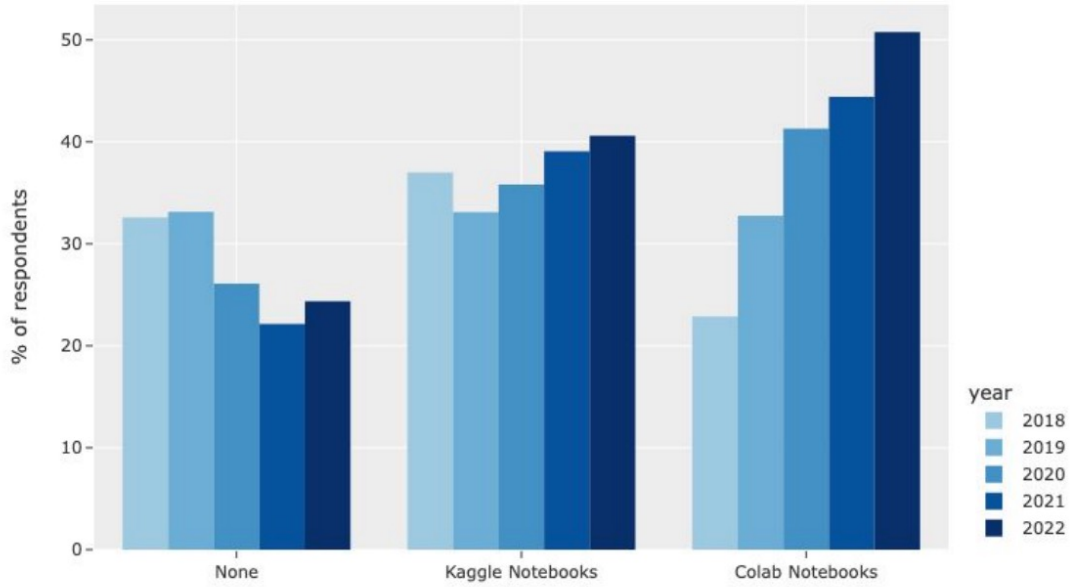
VSCode is now used by over 50% of working data scientists



Cloud notebook environment

Kaggle DS & ML Survey 2022

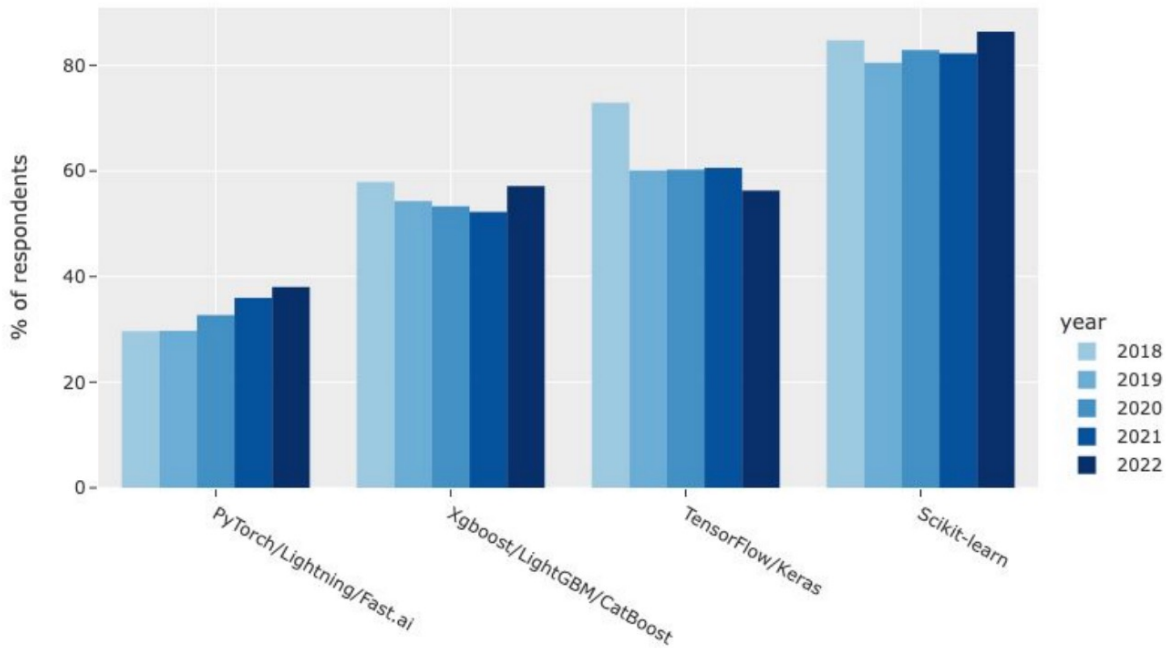
Colab notebooks are the most popular cloud-based Jupyter notebook environment



Machine learning framework

Kaggle DS & ML Survey 2022

Scikit-learn is the most popular ML framework while PyTorch has been growing steadily year-over-year

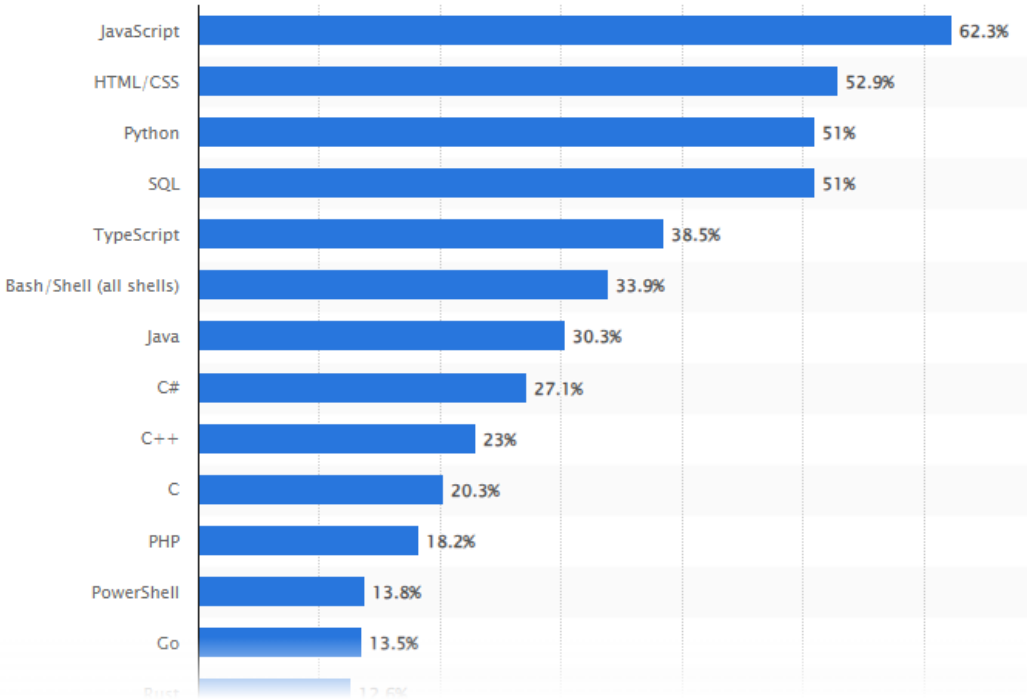


Self-study:
More results on Kaggle Report

- https://www.kaggle.com/kaggle-survey-2022

Statistics from other sources (2024)

- Programming language
  - https://www.statista.com/statistics/793628/worldwide-developer-survey-most-used-languages/
  - Not just for data science / AI



What is your career plan?

- Understanding the market needs
  - Example: Jobs DB
  - https://hk.jobsdb.com/

Note: I just pick the first 3 jobs that state the salary when I search on 25 Aug 2024 around 3pm. No cherry picking.

Lead Data Scientist

Sun Cupid Industries Ltd View all jobs

Cheung Sha Wan, Sham Shui Po District

Business/Systems Analysts (Information & Communi

Full time

$35,000 – $45,000 per month

Posted 2d ago

Key Qualifications

- **Education:** Bachelor's degree in Data Science, Computer Science, Statistics, Engineering, or a related discipline.
- **Experience:** A minimum of 5 years of experience in a relevant field, with a demonstrated history of leveraging data to drive significant business improvements.
- **Technical Expertise:**
  - Proficiency in programming languages such as Python or R.
  - Expertise in data manipulation and query languages like SQL.
  - Familiarity with machine learning frameworks (e.g., TensorFlow, PyTorch).
  - Knowledge of big data technologies (e.g., Hadoop, Spark) is highly desirable.
  - Experience in Android application development is a big advantage.
  - Strong analytical and problem-solving capabilities.
- **Leadership and Soft Skills:**
  - Exceptional communication and collaboration skills.
  - Proven leadership and project management abilities.
  - Strategic thinking with a keen eye for identifying business opportunities through data.

Adecco

Senior AI Specialist (NLP | Modern Fashion Enterprise | Perm)

ADECCO Personnel Limited View all jobs

Wan Chai, Wan Chai District

Engineering - Software (Information & Communication Technology)

Full time

$50,000 – $60,000 per month

Posted 1d ago

Requirements:

- Bachelor's or master's degree in a quantitative field, such as Mathematics, Statistics, or a related discipline.
- 3-4 years of hands-on experience in executing data-driven projects and working with large, diverse data sets.
- Mastery of natural language processing (NLP), cloud infrastructure (AWS), computer vision, programming languages (Python/R), database querying (SQL), and data visualization tools (e.g., Power BI).
- Proven track record of leveraging cloud computing platforms (AWS, Azure, Google Cloud) to drive data-powered initiatives.
- Innate curiosity, a research-oriented mindset, and the agility to juggle multiple priorities effectively.
- Extensive experience in working with large-scale data sets and leading complex, impactful analytical projects.
- Comprehensive knowledge of a diverse array of machine learning techniques, including clustering, decision trees, neural networks, and a deep understanding of their real-world applications and limitations.
- Expertise in utilizing state-of-the-art data visualization tools and frameworks to drive data-driven decision-making.

Robert—
—Walters

Lead Data Scientist - FS (8 yrs up)

Robert Walters (HK) Ltd View all jobs

Central and Western District

Other (Information & Communication Technology)

Full time

$90k - $100k p.a.

Requirements:
- Bachelor's degree in Computer Science or related field, advanced degree preferred.
- Minimum 8-10 years in cloud-based Data or AI/ML applications management.
- Experience leading data teams and driving business impact with data.
- Proficiency in Python, SQL, and data science libraries.
- Familiarity with NLP algorithms, deep learning frameworks, and Azure/Databricks is a plus.