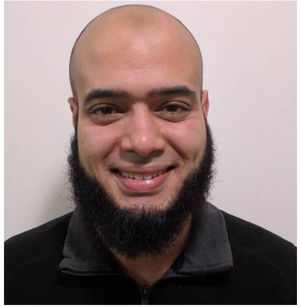
# Machine Learning

# Data Storage

**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 Storage

- WAYMO will derive a car for data collection for 100 hours storing **8 cameras**.
- How much data storage do we need to store?

# Data Storage

- First, we need to get all relevant information
  - How many frames per camera? All camera same rates? Say yes, 30 FPS
  - What is the resolution of the image? E.g. 1920 x 1080 pixels
  - What is the image format? Standard RGB
  - Do we need to store compressed or uncompressed? Uncompressed

- For a standard RGB color image (24 bits per pixel: 8 bits per R/G/B)
  - Pixels: 1920 × 1080 = 2,073,600 pixels
    - Each pixel 24 bits (for uncompressed)
  - Total bits: 2,073,600 x 24 = 49,766,400 bits = 6,220,800 bytes = **5.93 megabyte**
    - Recall 1 byte = 8 bits. 1 Megabyte = 1024 * 1024 byte

# Data Storage

- Total number of frames (30 FPS)
- 8 camera x 100 hours x 60 minutes-per-second x 60 seconds-per-minutes x 30  = 86400000 frames
- Then total disk space: 86400000 * 5.93 = 512352000 MB = 512.352 terabyte
    - Imagine all such data from only 100 hours driving!
    - We also may store metadata
    - Other sensor information: Radar and Ridar
    - Later storage for frames annotation (e.g. labels, boxes, semantic segmentation, etc)

# Data Storage

- On low level, storage typically is HDD and SDD
- In large scale data, we need to carefully decide
    - What will be the storage type? Compute how much data do u need
- File Systems
    - Local, Network File System (NFS), Distributed File Systems (e.g. Hadoop DFS)
- Databases
    - Relational Databases (e.g. MySQL) vs NoSQL Databases: (e.g. MongoDB, Cassandra)
- Cloud Storage: AWS S3, Google Cloud Storage, Azure Blob Storage
- Data Warehouses (processed data): Snowflake, BigQuery, Redshift
- Data Lakes (unprocessed data): AWS Lake Formation, Azure, Databricks
- Ad-hoc: TFRecords, Parquet, Avro, ORC, Feather

# Popular Data File Formats

- **CSV** is used for tabular data (manipulate by pandas)
- **JSON** is used for hierarchical or nested data (e.g. metadata)
- **YAML** is used in configuration files (supports Comments)
- **Parquet**: Columnar storage format optimized for use with big data processing frameworks like Hadoop and Spark. Efficient for read-heavy workloads.
- **HDF5** (Hierarchical Data Format version 5): Used for storing large datasets including *multidimensional arrays*
- **SQLite**: Lightweight disk-based database that can be a good fit for small to medium-sized datasets.

# Privacy Tip

- Don’t collect something that requires privacy (e.g. car plate license / people)
- Don't record in a private property
- Be careful about subjects
    - Minimize Data: Only collect data that is absolutely necessary
    - If possible, **anonymize** the data so that the identity of subjects cannot be easily linked
    - **Informed Consent**: Always obtain informed **consent** from the individuals whose data is being collected. Make sure they understand what data is being collected, for what **purpose**, and how it will be used.
    - **Consider**: Data Encryption, Access Controls, Secure Storage, Retention Policies, Regular Audits, Data Breach Plans, Data Sharing, Training and Awareness, Compliance

“Acquire knowledge and impart it to the people.”

“Seek knowledge from the Cradle to the Grave.”