Full Stack AI Projects

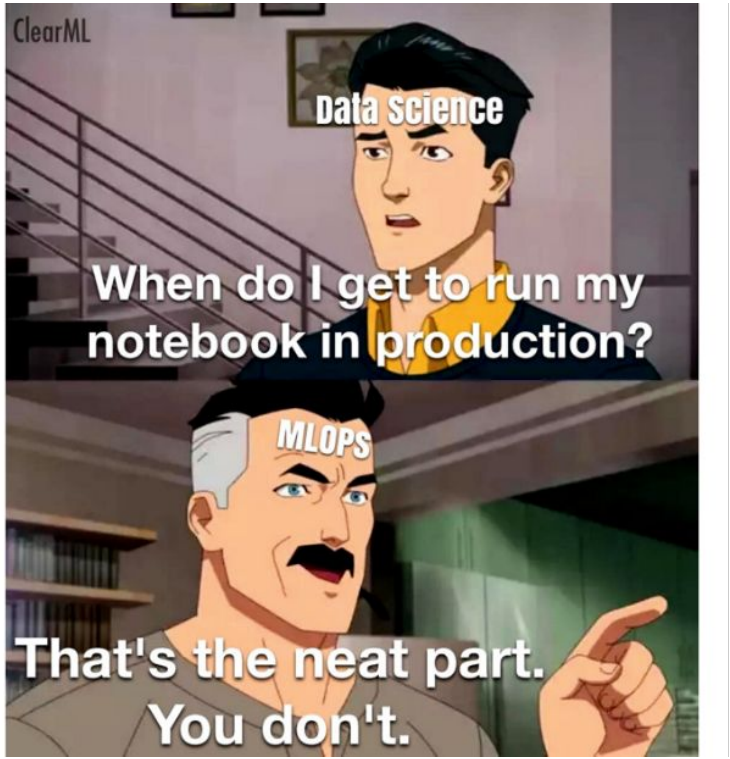
Data Management Tools

Henry Ruiz
GDE ML
@devharuiz
https://haruiz.github.io/

What we have discussed so far?

What is machine learning system design?

The process of **defining the <u>interface, algorithms, data, infrastructure, and hardware</u>** for a machine learning system to satisfy specified requirements.

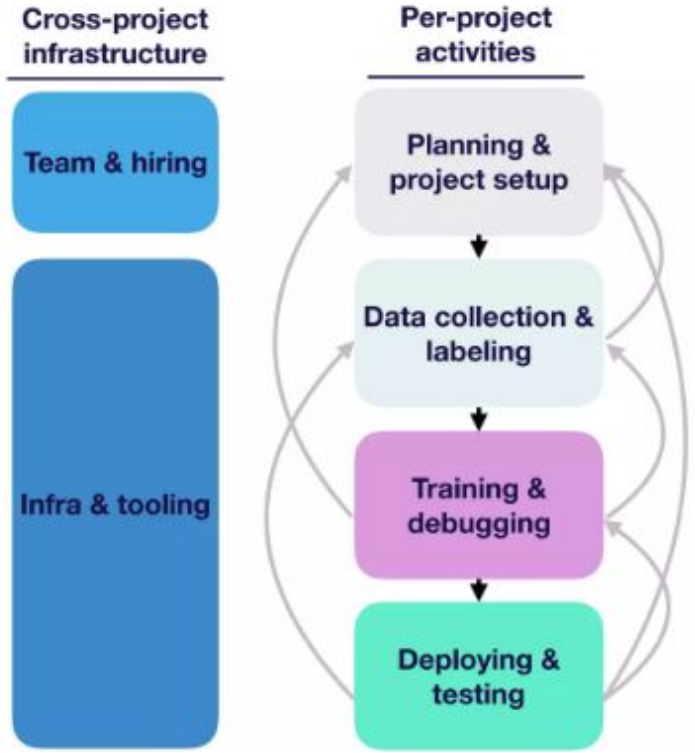


So, What is required to design and build an ML system?

What we have discussed so far?

Define a methodology

Like any other software solution, ML systems require a **well-structured methodology** to maximize the success rate of the implementation.



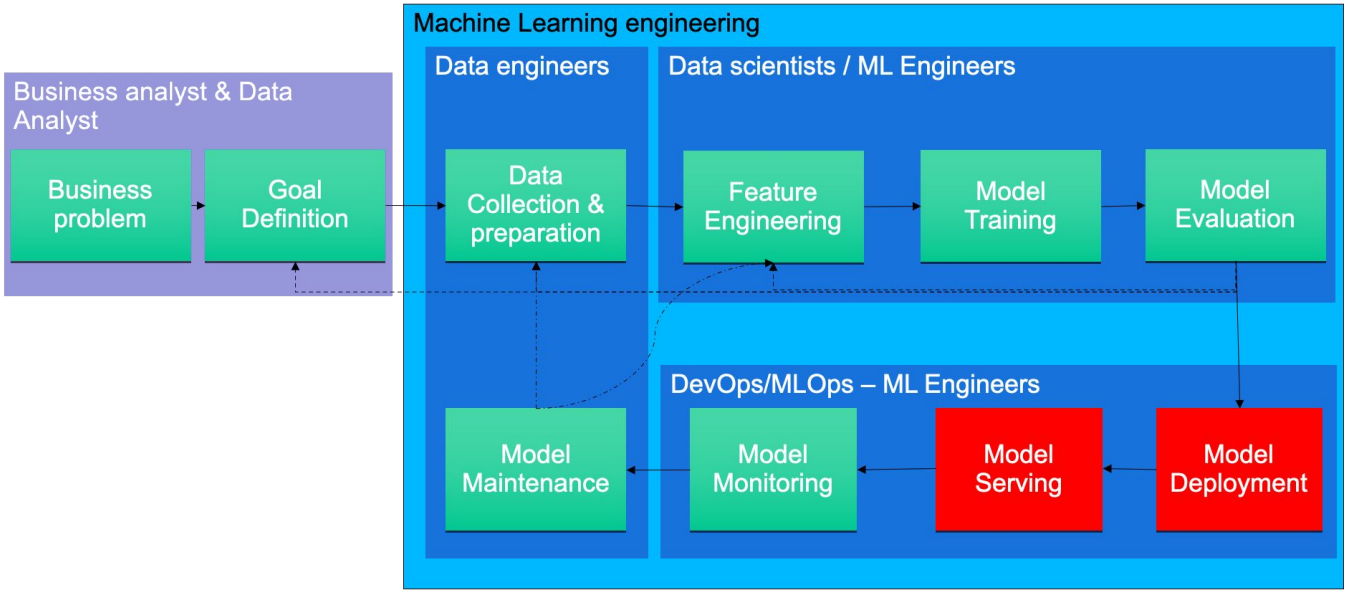
What we have discussed so far?

So, What is required to design and build an ML system?

Set up a team

Managing and leading ML and Data Science teams require unique skills.

ML teams have diverse roles.

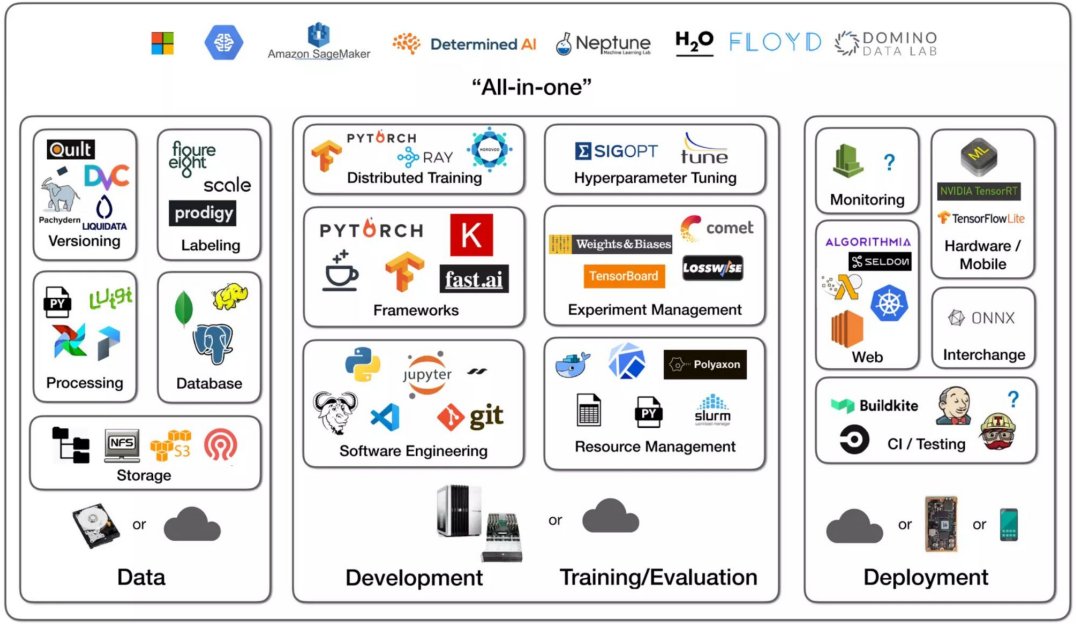


What we have discussed so far?

Define Development infrastructure & Tooling

The number of **available tools to work with ML seems endless.**

Selecting the appropriate tools depends on:
the kind of problem, **type of solution, deployment scenario, capacity building**, team experience, cost, hardware and software infrastructure, etc.



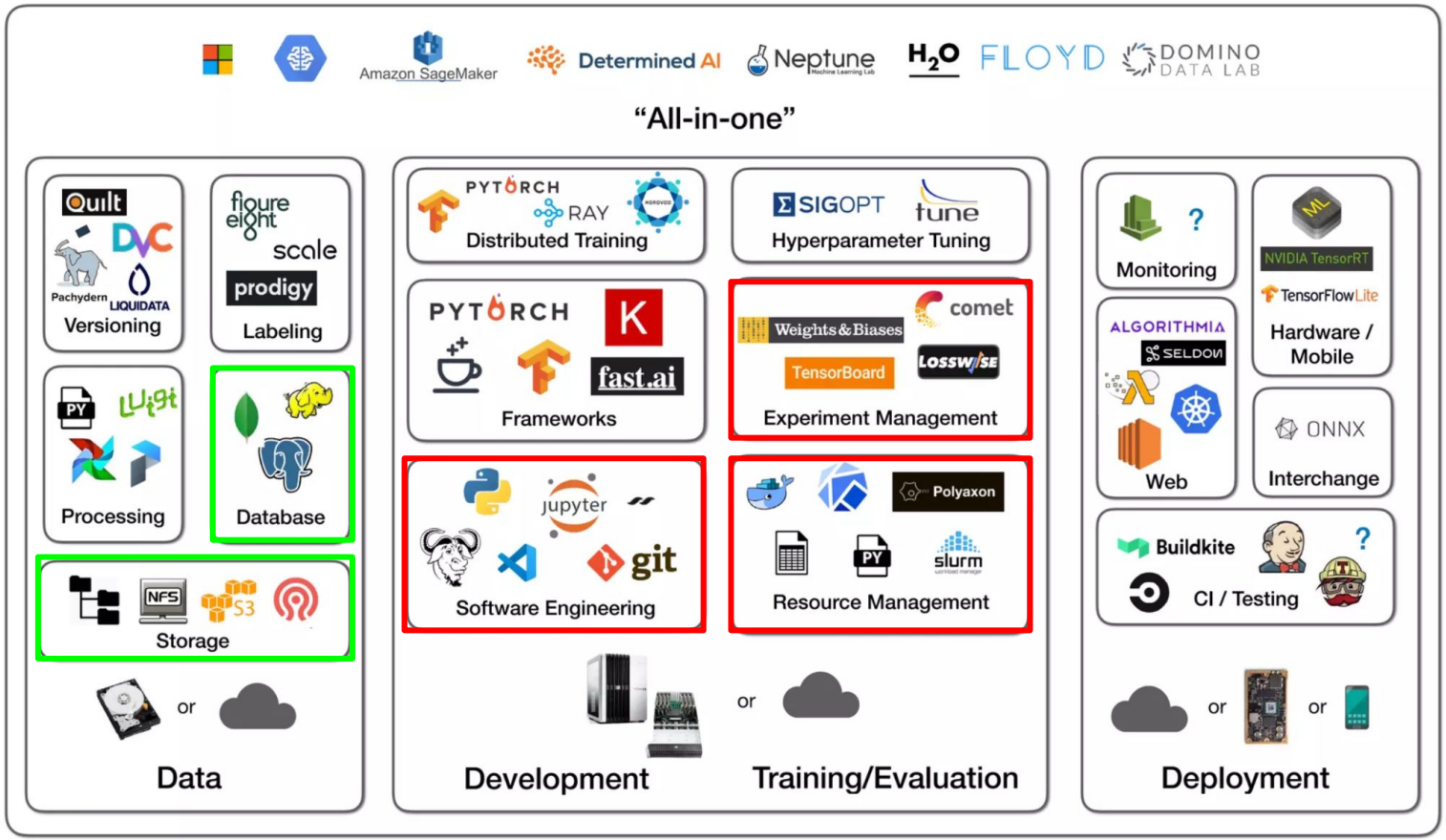
Last class…

Experiment and Model Management

Experiment management refers to **tools and processes that help us keep track of code, model parameters, and data sets that are iterated on during the model development lifecycle**. Such tools are essential to effective model development. There are several solutions here:

- TensorBoard: A non-exclusive Google solution effective at one-off experiment tracking. It is difficult to manage many experiments.
- MLflow: A non-exclusive Databricks project that includes model packaging and more, in addition to experiment management. It must be self-hosted.
- Weights and Biases: An easy-to-use solution that is free for personal and academic projects! Logging starts simply with an "experiment config" command.
- Other options include Neptune AI, Comet ML, and Determined AI, all of which have solid experiment tracking options.

Software Engineering Tools



Data Management Tools

Data management is the process of collecting, storing, organizing, maintaining, and utilizing data effectively and efficiently.In ML, it is an important component, since data is the fuel of the algorithms.

“Data is the new currency”
“Data is the new oil”
“Data is the new economy”

What is data?

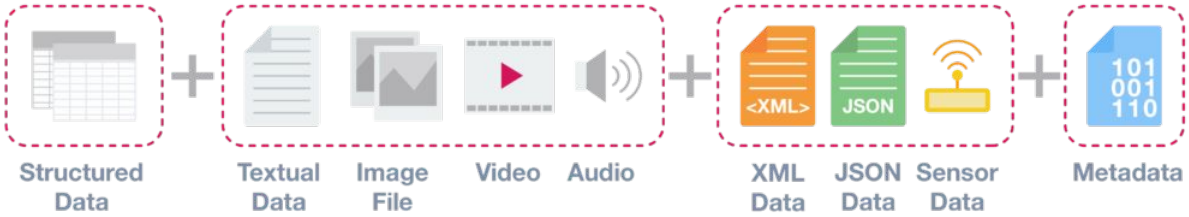
Data refers *to* ***a collection of facts, information, or statistics that are often represented in a numerical or digital form.*** Data can be in various forms, such as text, images, audio, video, or any other format ***that can be stored and analyzed using digital technologies.***

Data is an important resource for decision-making in many fields, including business, science, healthcare, education, and government. With the advent of big data and advanced analytics technologies, the analysis and interpretation of data have become increasingly important for organizations ***to gain insights into patterns,*** **trends, and behaviors that can inform decision-making and drive innovation.**

Structure vs Unstructured Data

Structured data refers to **data that is organized and formatted in a specific way, typically within a fixed schema**. Structured data **can be easily analyzed, searched, and processed using algorithms or software tools. Examples of structured data include spreadsheets, databases, and tables.**

On the other hand, unstructured data refers to **data that is not organized in a specific way and does not have a fixed schema.** Unstructured data can be more difficult to analyze and process compared to structured data because it lacks a consistent format. **Examples of unstructured data include text documents, social media posts, audio and video recordings, and images.**



How data is generated?

Structured data is generated through intentional efforts to organize and format data in a specific way. For example, a company might use a database management system to store and manage data in a structured format, or an individual might use a spreadsheet to organize and analyze data.

Unstructured data, on the other hand, is generated through natural and unorganized means. For example, unstructured data can be generated through social media posts, emails, audio and video recordings, and images. These types of data are often generated by individuals or groups without any specific intention to format or organize the data. As a result, unstructured data is typically less consistent in format and can be more difficult to analyze and process.

The age of generative AI

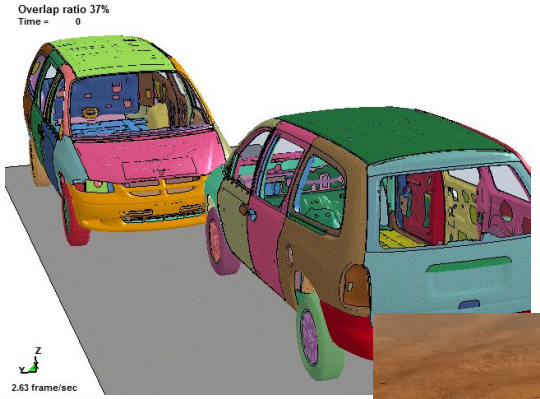
Synthetic data is artificially generated **data that mimics the statistical properties and structure of real-world data or our dataset distribution.**

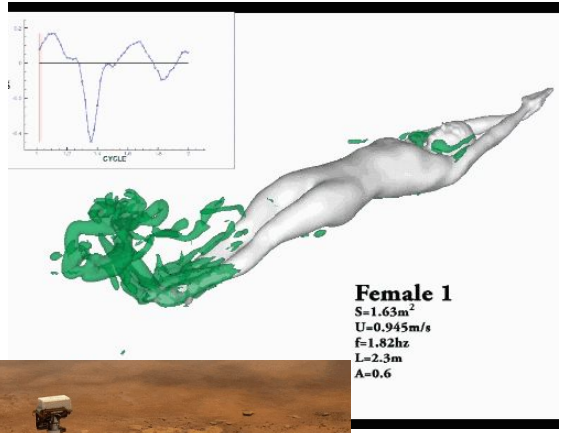
Synthetic data **can be used in situations where real data is scarce, sensitive, or proprietary, or where there is a need to test or validate algorithms or models.**

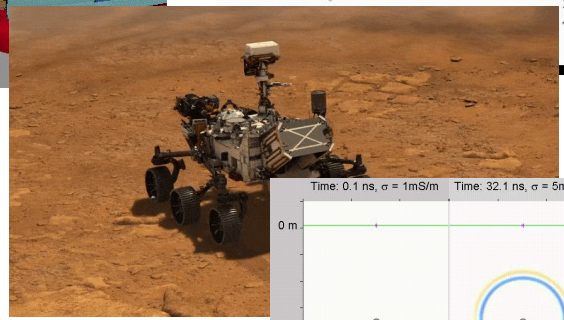
Fig. 1: Synthetic images generated using recent text-to-image models: DALL·E 2 [3], stable diffusion [4] and GLIDE [5].

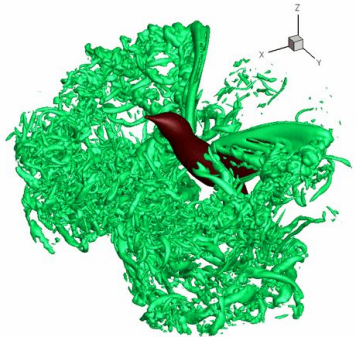
The age of generative AI - How the data is generated

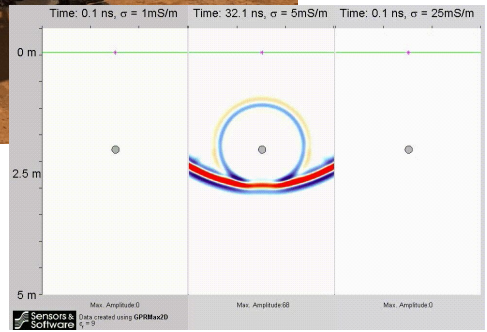
Simulation: Simulation involves **modeling real-world phenomena and generating data based on those models.** For example, a simulation could be used to generate synthetic weather data based on historical patterns and statistical distributions











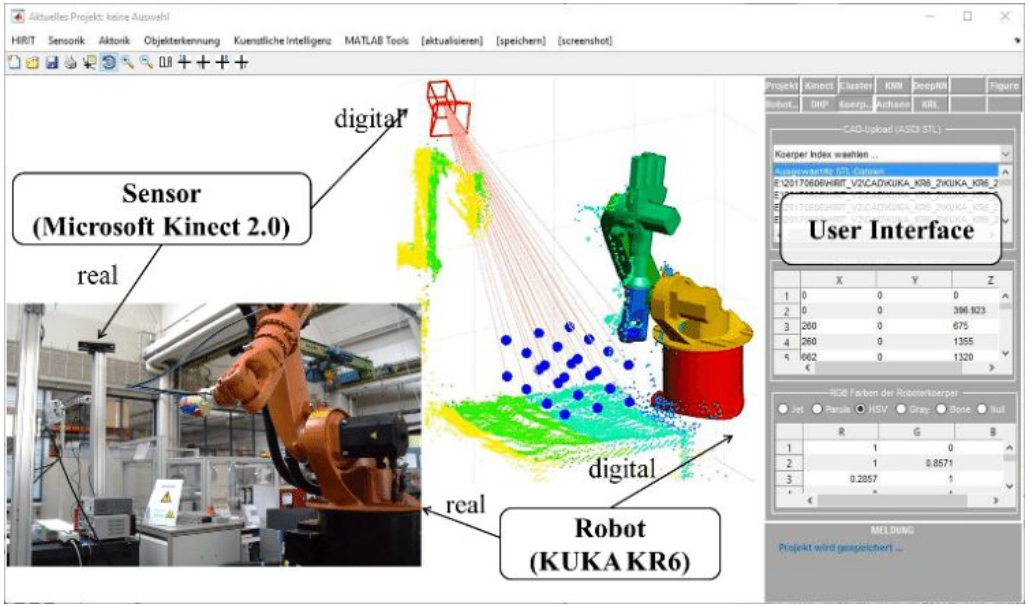
https://www.nvidia.com/en-us/deep-learning-ai/industries/robotics/

Digital Twin

A digital twin **is a virtual replica of a physical system or process that is created using data from sensors, machines, or other sources that** can be used to simulate, monitor, and optimize the performance of the physical system in real-time.

Digital twins typically **consist of three main components: the physical asset**, **the digital model, and the data that links the two**. The physical asset can be a machine, a building, or even an entire city, while the digital model is a virtual representation of that asset.

The data that links the two includes information about the physical asset's design, performance, and maintenance history, as well as real-time data from sensors and other sources.

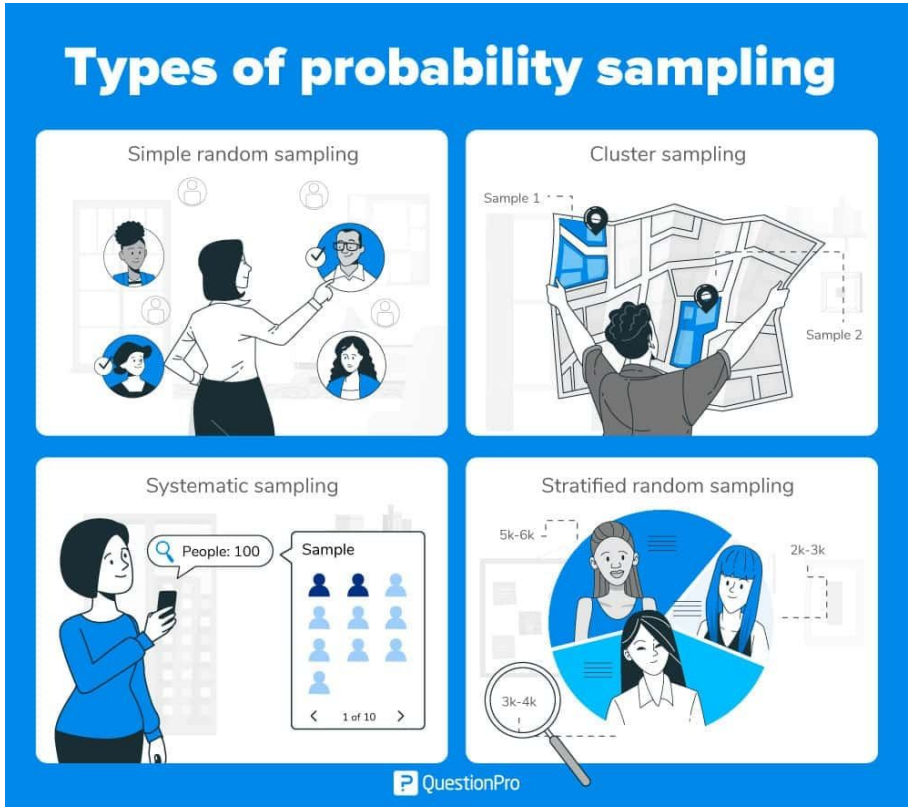


https://www.nvidia.com/en-sg/omniverse/solutions/digital-twins/

https://www.youtube.com/watch?v=6-DaWgg4zF8&ab_channel=NVIDIA

The age of generative AI - How the data is generated

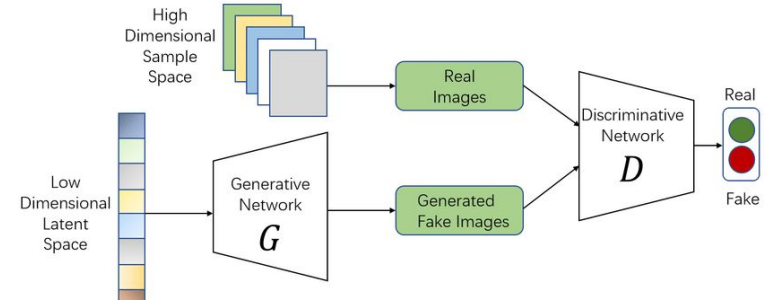
Random sampling: Random sampling involves **generating data by randomly selecting values from a known distribution**. For example, synthetic data could be generated by randomly selecting values from a Gaussian distribution to mimic the statistical properties of a real dataset.

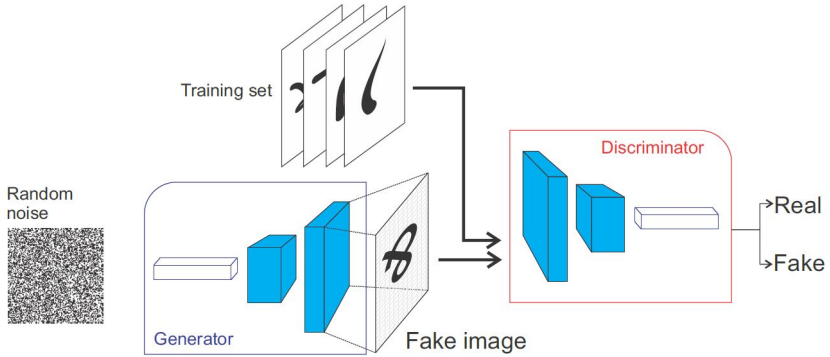


The age of generative AI - How the data is generated

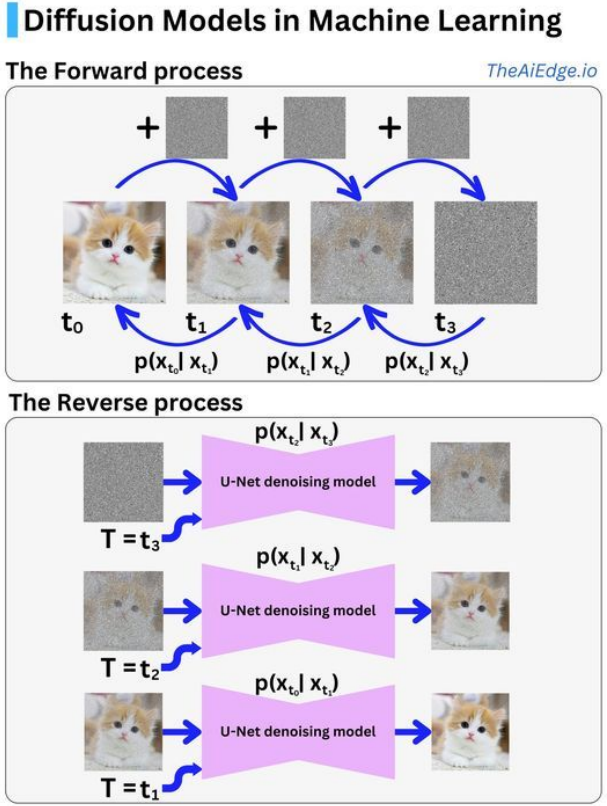
Generative models: Generative models **involve training machine learning algorithms to generate new data based on patterns in existing data.** For example, a generative model could be trained on a set of real images to generate new, synthetic images that have similar visual characteristics.

GANS and Diffusion Models





Learn the distribution of the data to generate new samples

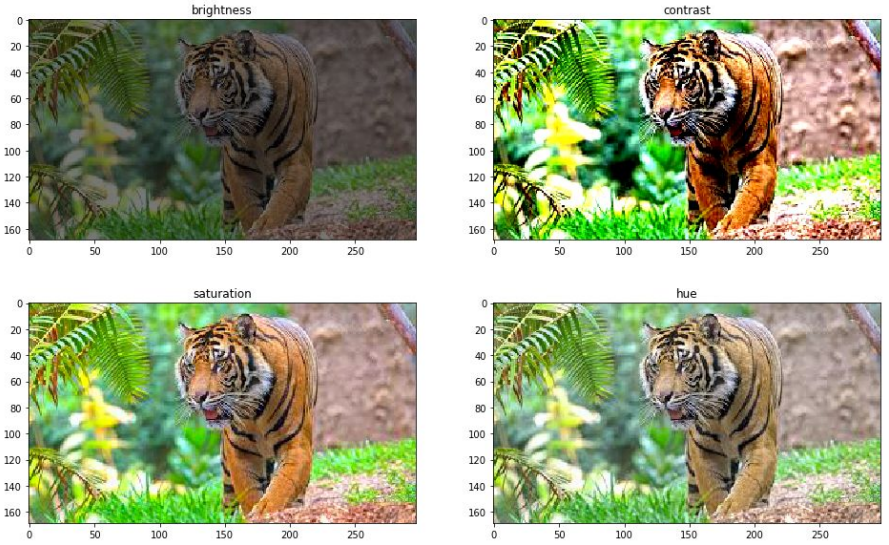


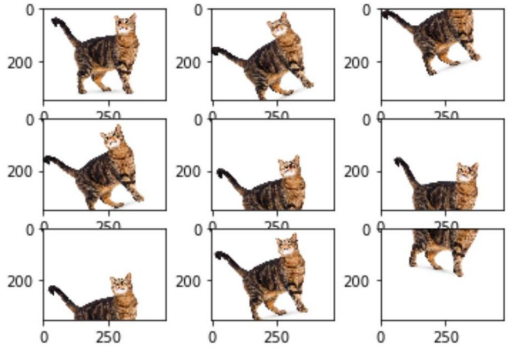
Iteratively and gradually make an image noisier and noisier, and then train a model to learn how to denoise it to get back to the original image.In Other words, Diffusion models can generate coherent images from noise.

Science behind diffusion models

Data Augmentation

Data augmentation: Data augmentation **involves manipulating existing data to create new, synthetic data.** For example, synthetic data could be generated by **applying random rotations, translations, or scaling to existing images.**



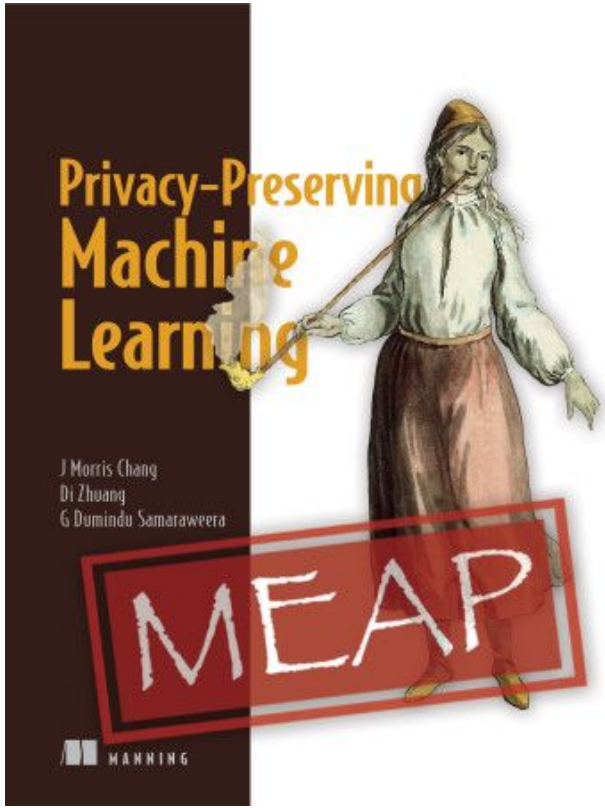


What about privacy?

Synthetic Data refers to **data that is generated artificially, typically through computer algorithms or simulations.** This type of data is used to augment training data sets, improve machine learning models and increase the size of the data set, among other things.

Data Augmentation, on the other hand, refers to transforming the existing data set to increase its size and diversity. This can be achieved through rotation, scaling, flipping, cropping images, or adding random noise to the data.

Data Anonymization refers to transforming personal data into a form that does not identify individuals. This is important to protect the privacy of individuals whose data is being used for machine learning models or other data science applications.



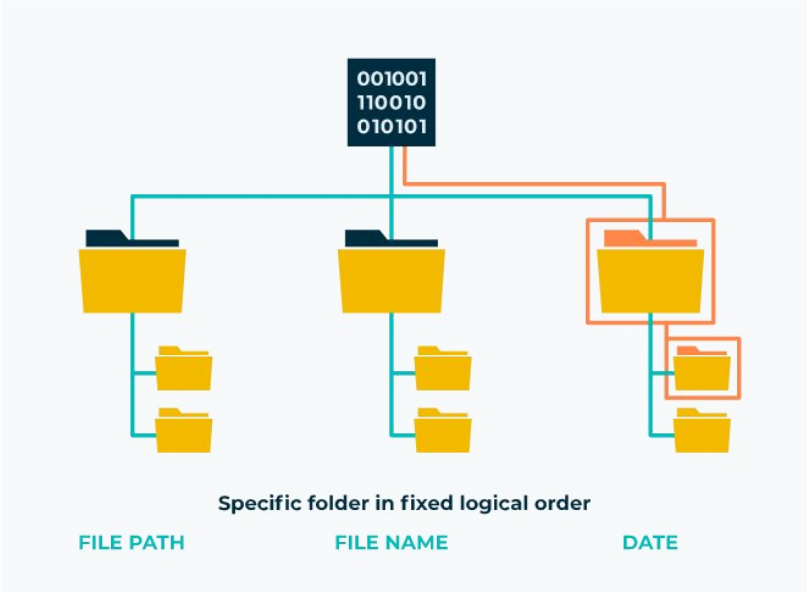
https://github.com/tensorflow/privacy

How is data stored?

Filesystem storage: data is organized and managed on a storage device such as a hard drive, solid-state drive (SSD), or flash drive. It provides a hierarchical structure of directories (also known as folders) and files, which can be accessed and manipulated by the operating system.

The most common file systems used in modern operating systems are NTFS (New Technology File System) on Windows and APFS (Apple File System) on macOS, but there are many other file systems, including ext4, FAT32, and HFS+.

The choice of file system depends on factors such as the operating system, the type of storage device, and the intended use of the storage device.system and applications.



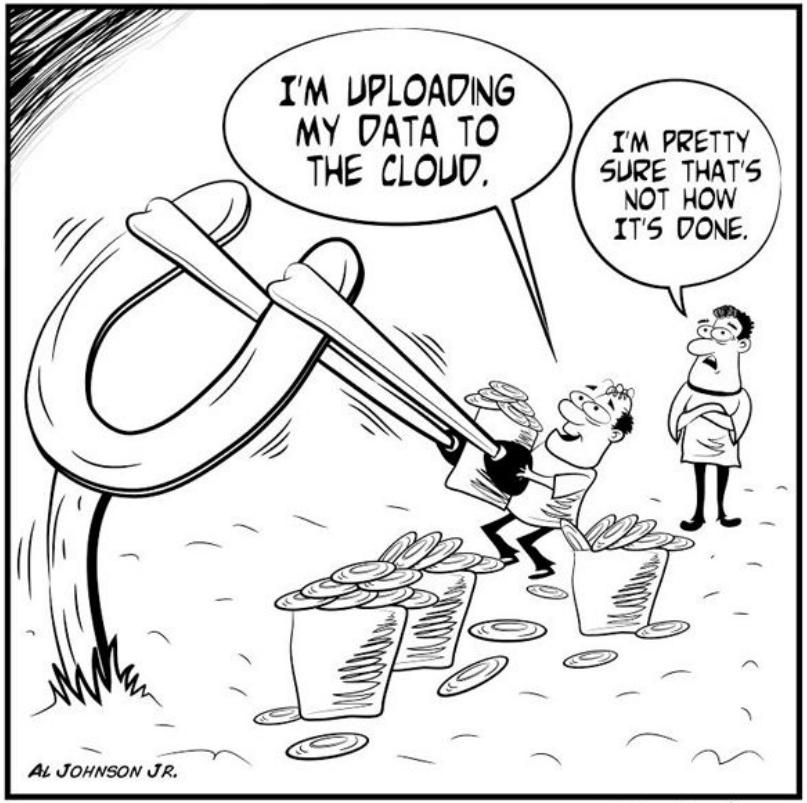
How is data stored?

Object Storage: The object storage is **an API over the filesystem.** Its fundamental unit is an object. **Data is usually store in binary format (an image, a sound file, a text file, etc.).** In object storage, **each object is given a unique identifier**, known as an object ID, which can be used to retrieve and access the object. The object ID can be a simple string or a complex hash, depending on the object storage system.

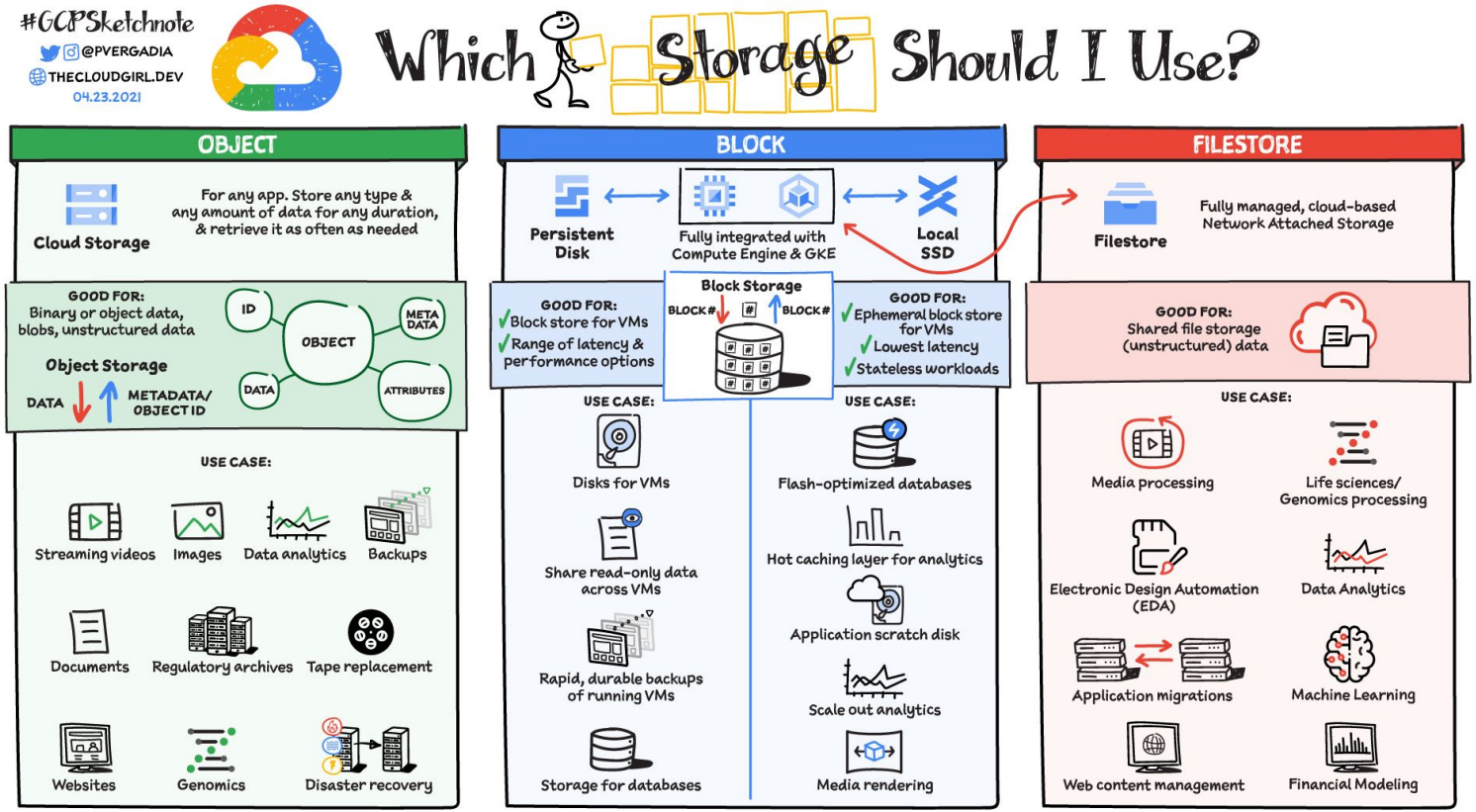
Object storage systems **are designed for massive scalability and can store large amounts of unstructured data, such as multimedia files, documents, and social media data.** It is the predefined storage model adopted by cloud platforms.

Some popular object storage systems include Amazon S3, Google Cloud Storage, and Microsoft Azure Blob Storage.

Data is generally accessed by using code API's



Which Storage Should I Use?



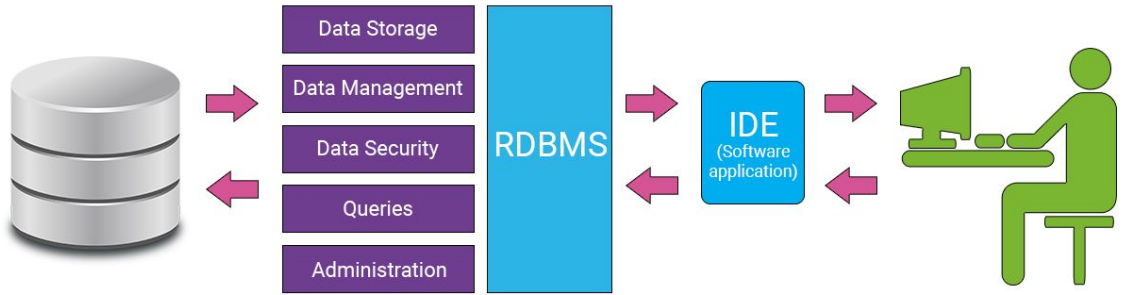
All credits to https://thecloudgirl.dev/

How is data stored?

Databases: organized collections of data that are designed to be easily accessed, managed, and updated. It allow users to store and retrieve data efficiently and securely.

A database typically consists of one or more tables, which are collections of data organized into rows and columns. Each table represents a specific type of data, such as customer information or product inventory. Within a table, each row represents a unique record or instance of the data, and each column represents a specific attribute or characteristic of the data.

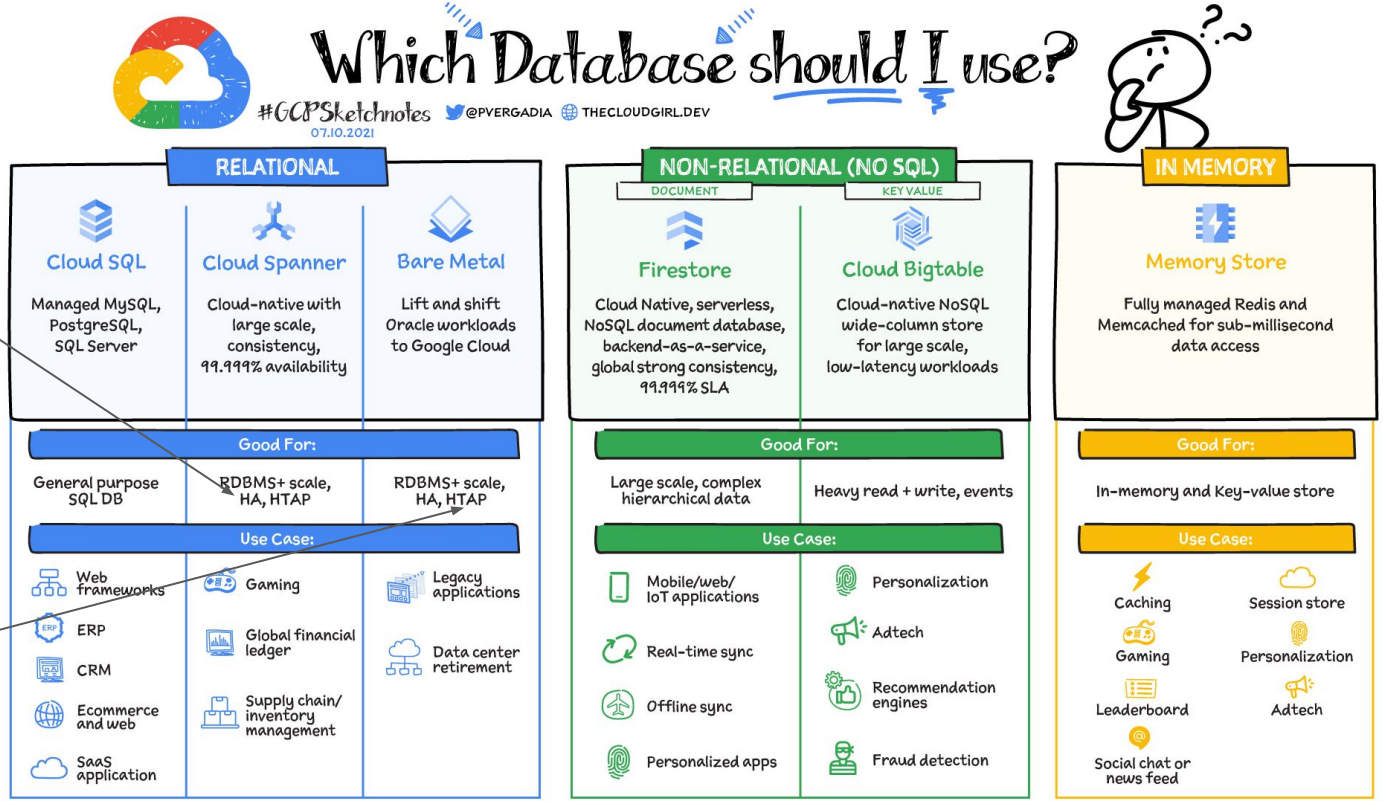
Databases are typically managed by a database management system (DBMS), which provides tools and interfaces for creating, modifying, and querying the data. The most commonly used DBMSs are relational database management systems (RDBMS), such as MySQL, Oracle, and Microsoft SQL Server, which use the SQL (Structured Query Language) programming language for managing and querying data.



Which database use?

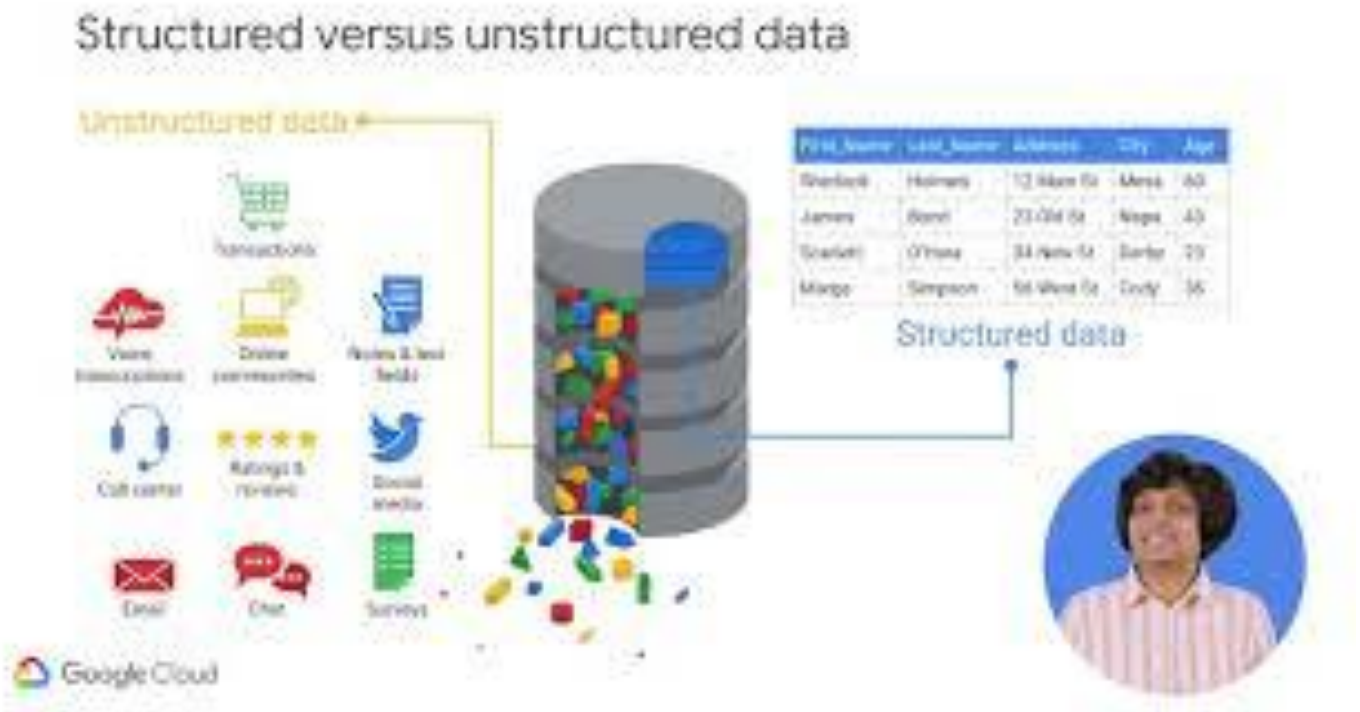
High-availability storage (HA storage)

Hybrid Transactional/Analytical Processing

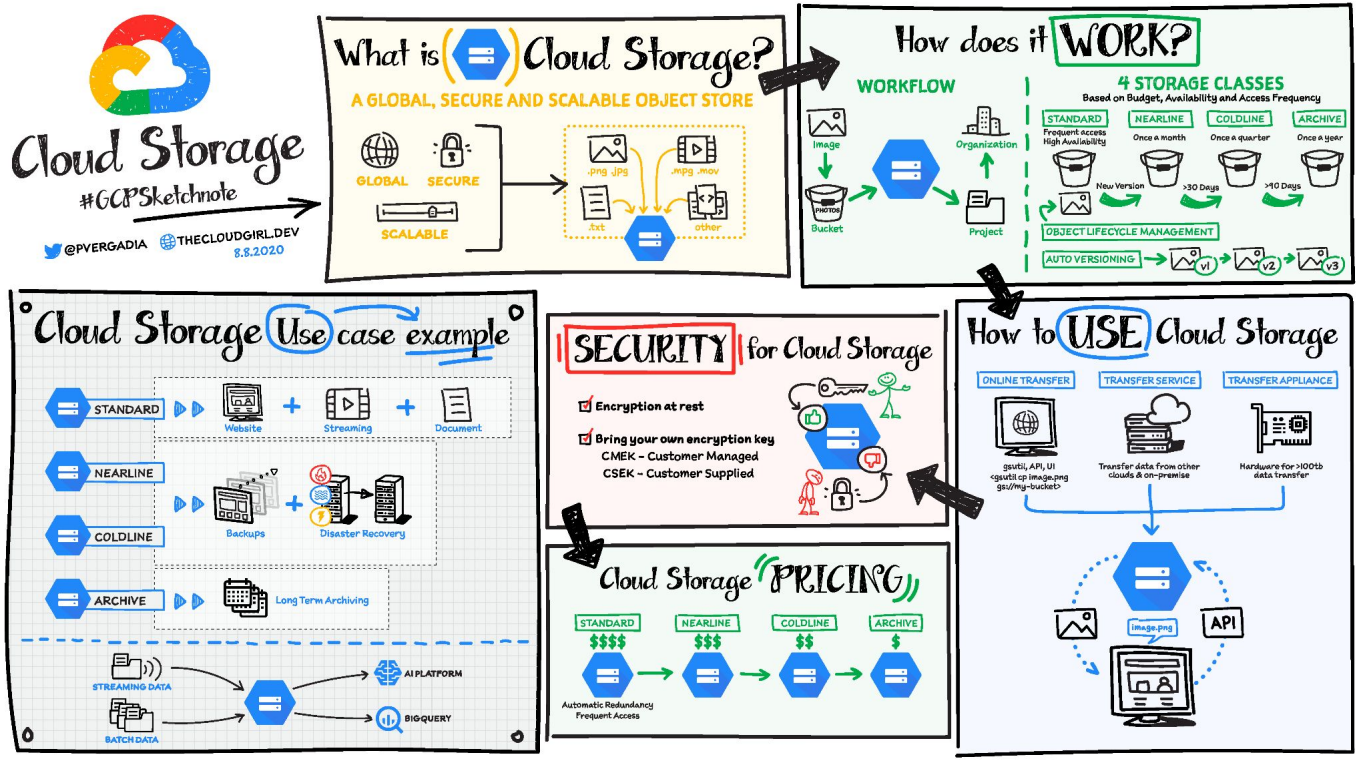


All credits to https://thecloudgirl.dev/

GCP(Google Cloud Platform)



Cloud storage - big picture



All credits to https://thecloudgirl.dev/

Big data solutions

Data warehouse:

- A data warehouse is a centralized repository of data that is designed for reporting and analysis.
- It typically contains structured data from multiple sources, such as transactional systems, and is organized to support specific business processes or analytical use cases.
- Data warehouses use a schema-on-write approach, which means that data is structured and formatted before it is loaded into the warehouse.
- Data warehouses typically use ETL (extract, transform, load) processes to move data from source systems into the warehouse and transform it into a format that is optimized for reporting and analysis.

Data lake:

- A data lake is a large, centralized repository of raw data that is stored in its native format and can be accessed and analyzed by a variety of tools and users.
- Data lakes are designed to support a wide range of use cases, from exploratory data analysis to machine learning and AI.
- Data lakes use a schema-on-read approach, which means that data is stored in its native format and is structured and formatted as it is accessed or queried.
- Data lakes can contain both structured and unstructured data from a variety of sources, including social media, IoT devices, and other unstructured sources.
- Data lakes use tools and interfaces such as data catalogs and metadata management to help users discover, understand, and use the data in the lake.

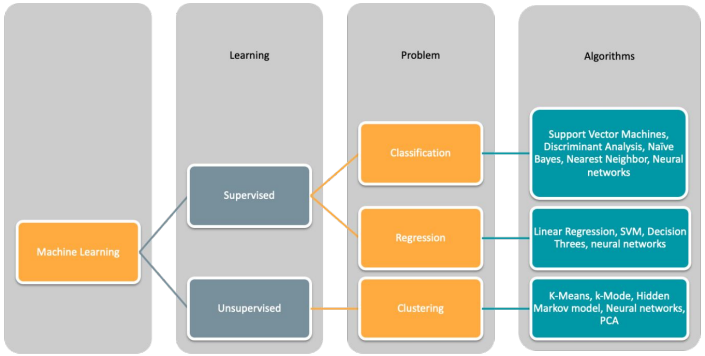
How data is presented to the model?

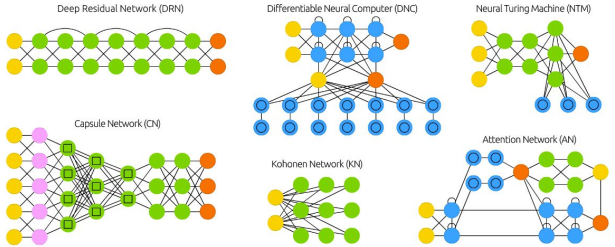
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

Structured data

Non-Structured data

Fit data



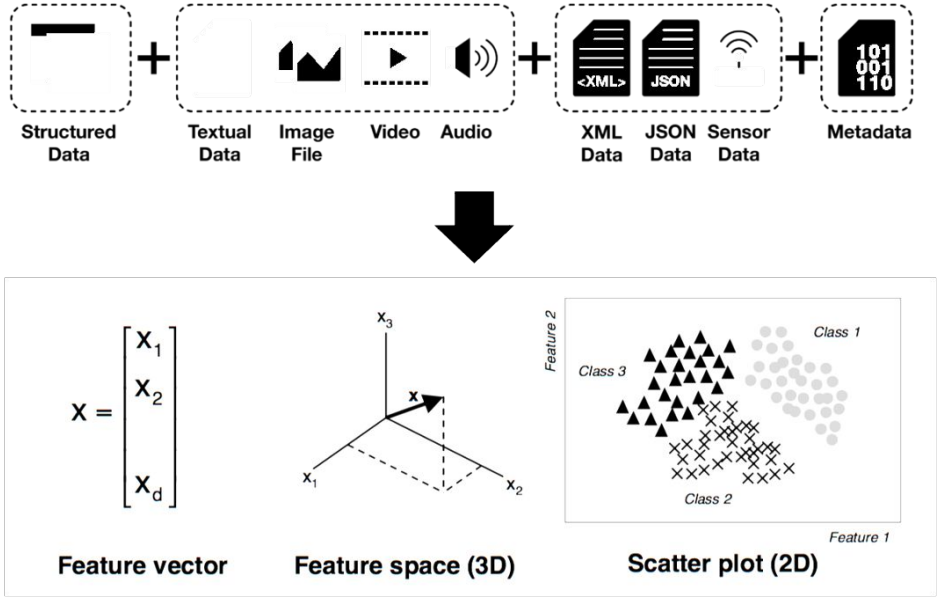


https://www.asimovinstitute.org/neural-network-zoo/

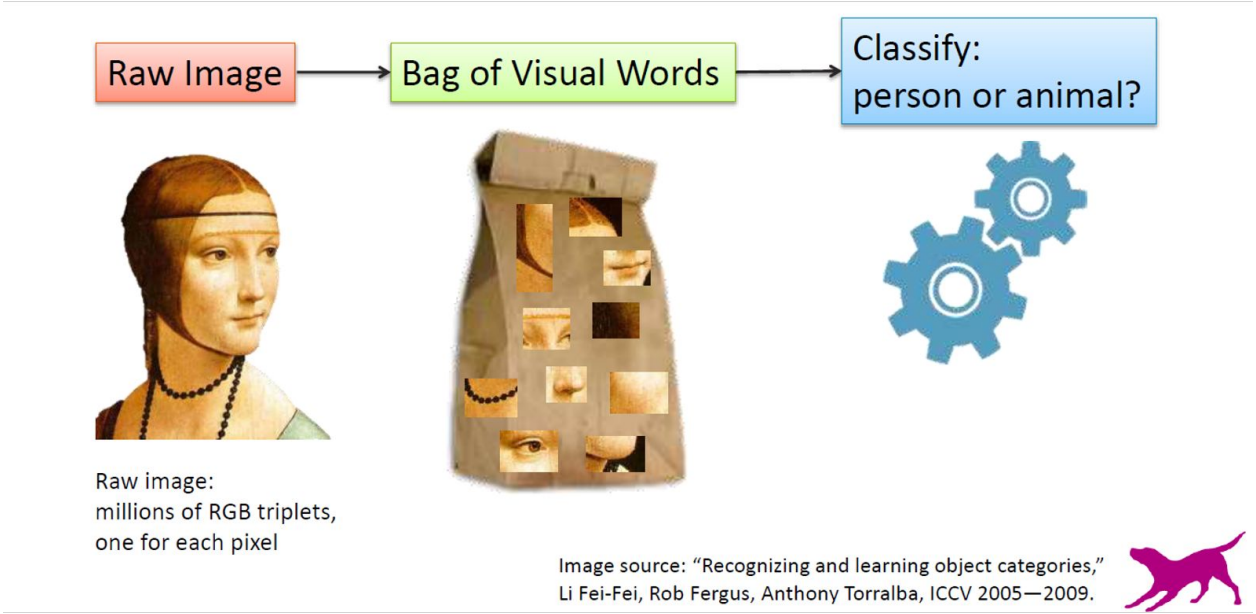
Model

Algorithm/Architecture Selection

How does my model process and work with data during training?



Feature space



How data is visualized ?

https://datavizcatalogue.com/

https://clauswilke.com/dataviz/