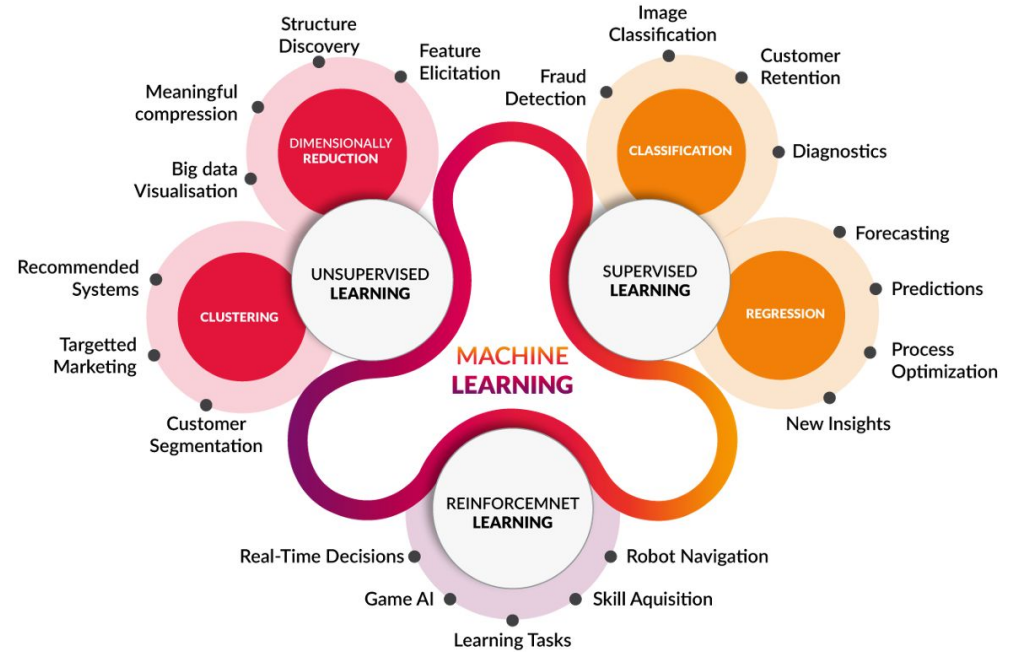
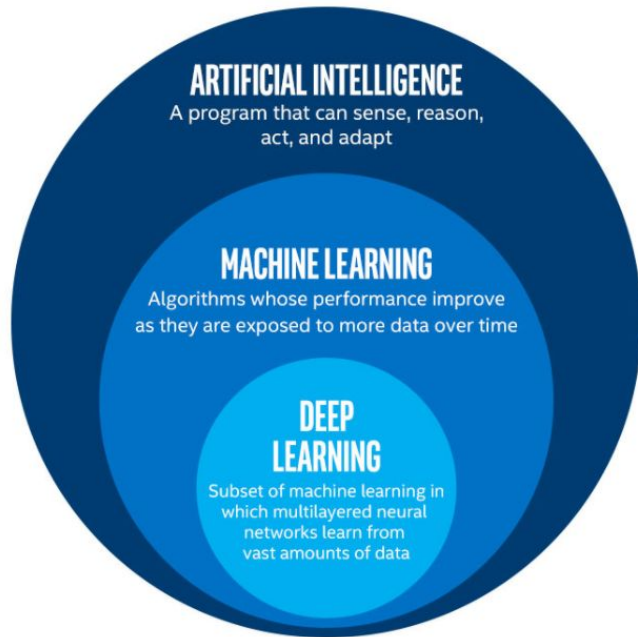


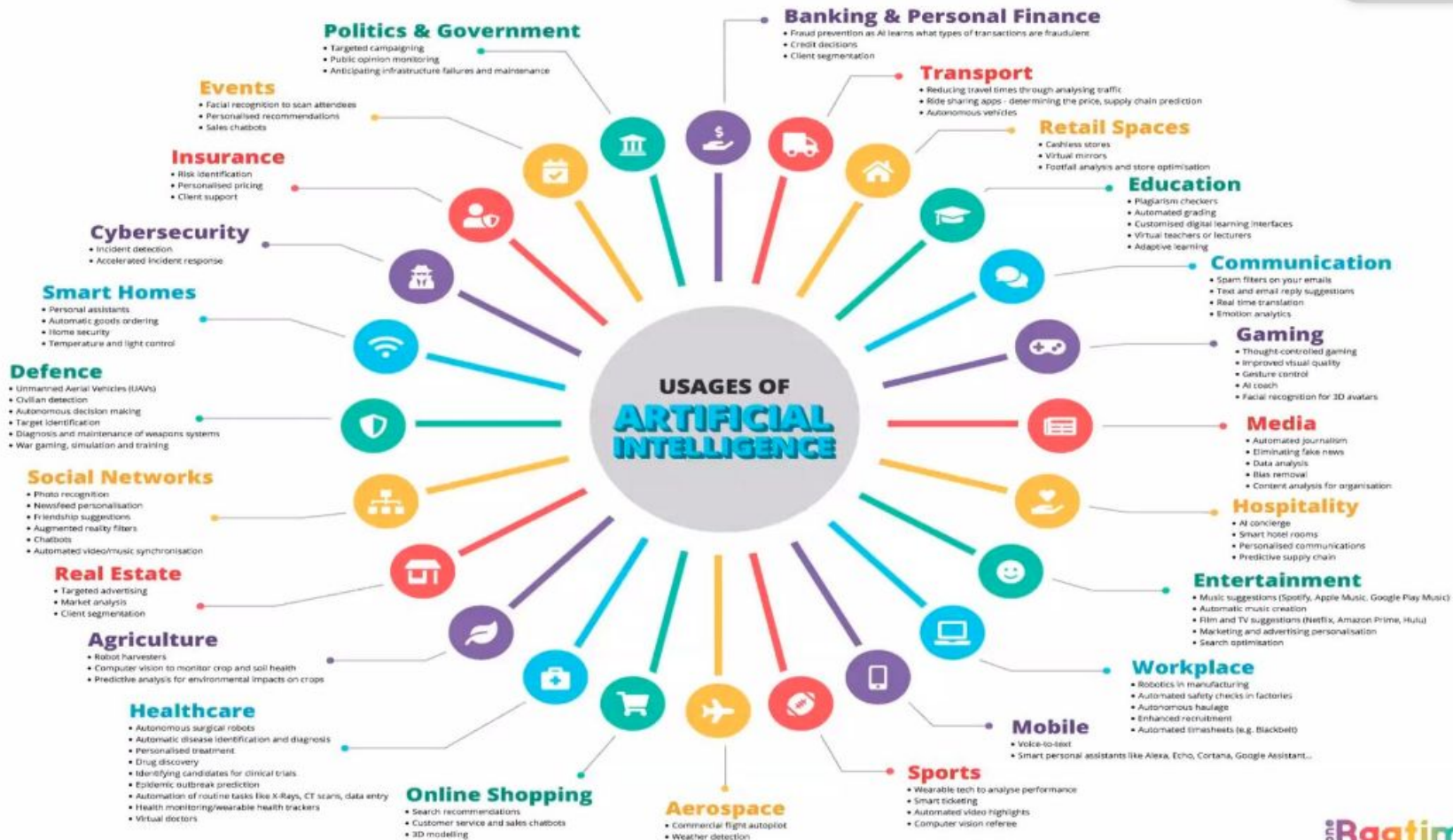
AI Data and Model Landscape

Giang Nguyen

2023-11-15 [AI4EOSC](#) Platform: User's Workshop

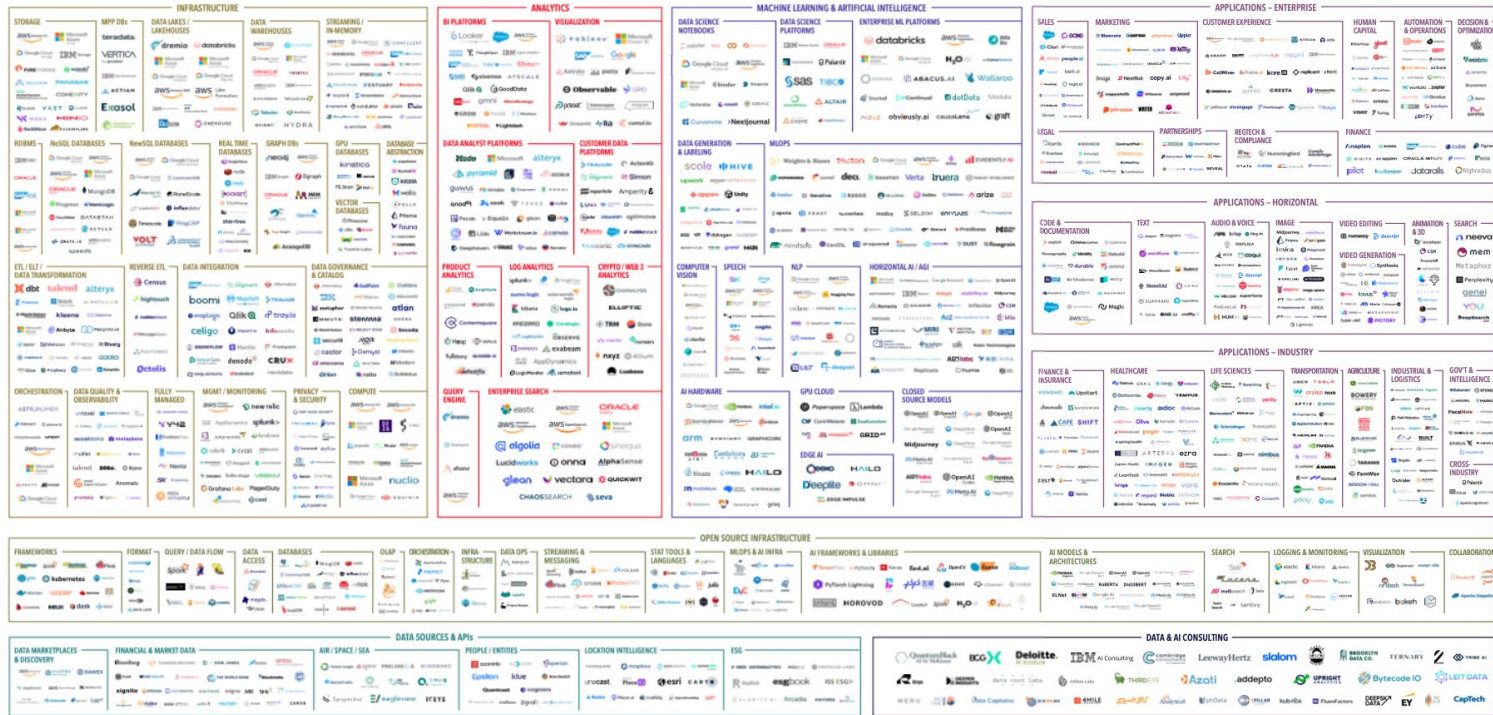
Artificial Intelligence (AI) and Machine Learning





The 2023 MAD: Machine Learning, AI and Data

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



Version 1.0 - Feb 2023 © Matt Turck (@matturck), Kevin Zhang (@ykevinzhang) & FirstMark (@firstmarkcap)

Blog post: matturck.com/MAD2023

Interactive version: MAD.firstmarkcap.com

Comments? Email MAD2023@firstmarkcap.com

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Exponential acceleration of Generative AI

ChatGPT - a kind of general purpose intelligence

AI generated content, AI bias, AI hallucination
→ **Responsible AI**

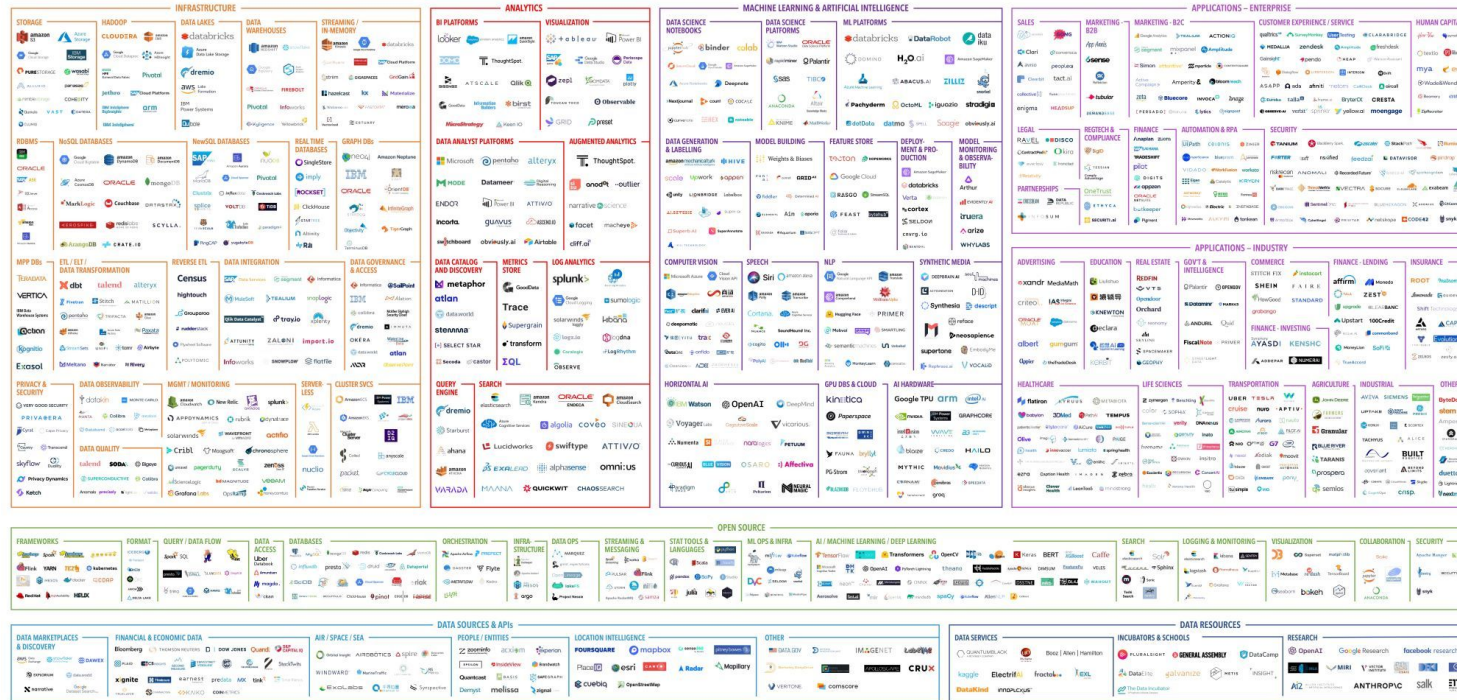
Hadoop - gradual **disappearance** of the Big Data technology

Crypto/web3 analytics

Data mesh, products, contracts: dealing with software complexity

The 2021 MAD: Machine Learning, AI and Data

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



Version 3.0 - November 2021

© Matt Turck (@mtturturk), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)

mttturck.com/data2021

FIRSTMARK
SMALL-SCALE VENTURE CAPITAL

Data mesh

Busy year for **DataOps**

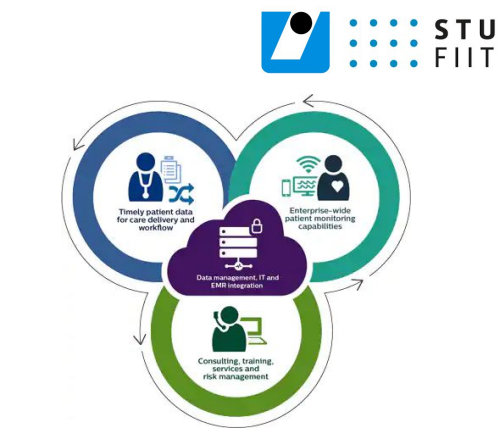
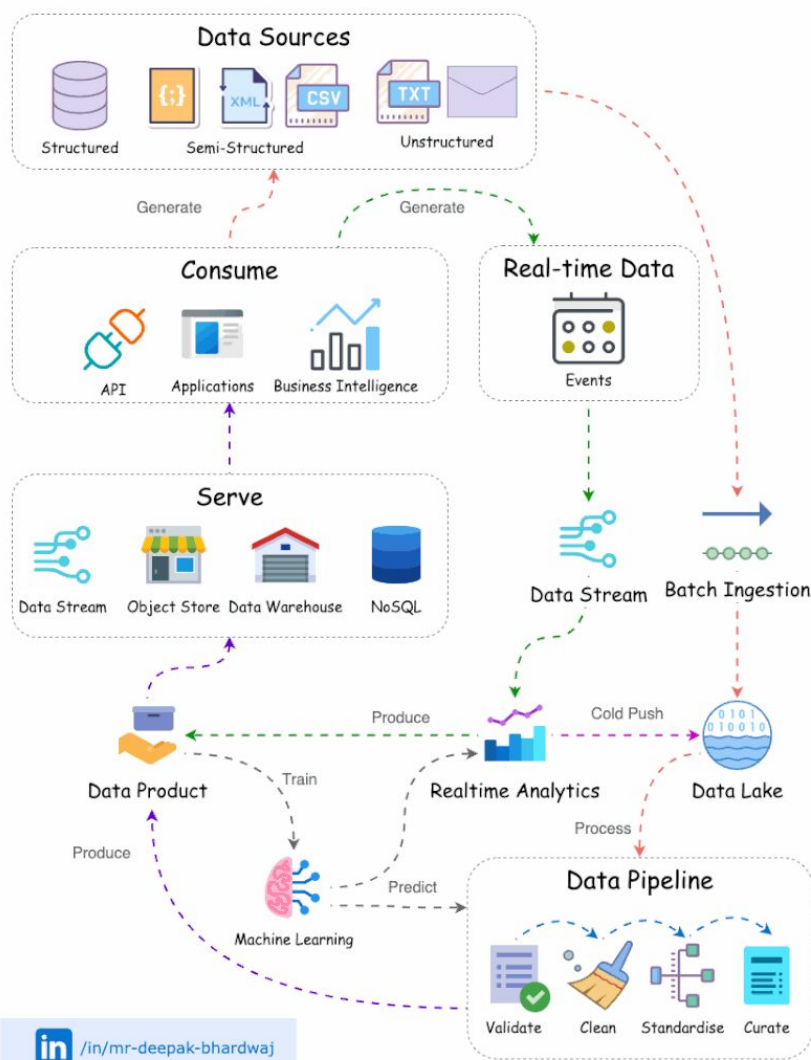
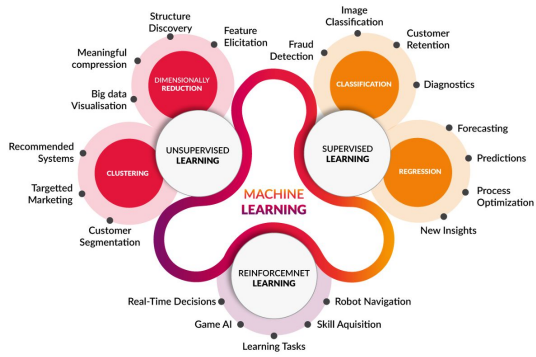
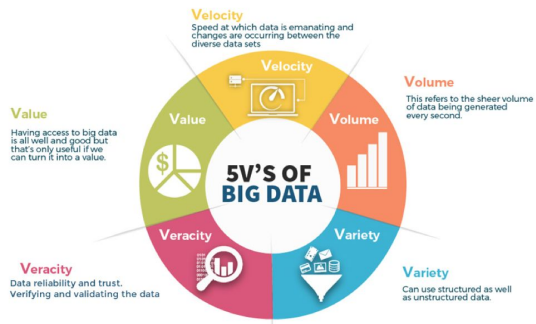
It's time for real time

Metrics stores

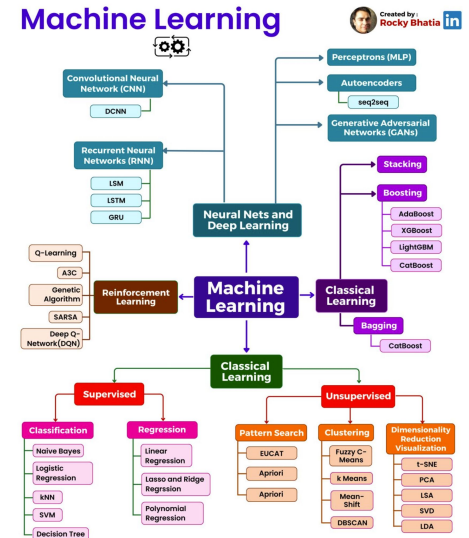
Reverse ETL

Cross-org data sharing

Privacy and security



Machine Learning



Imbalanced Datasets

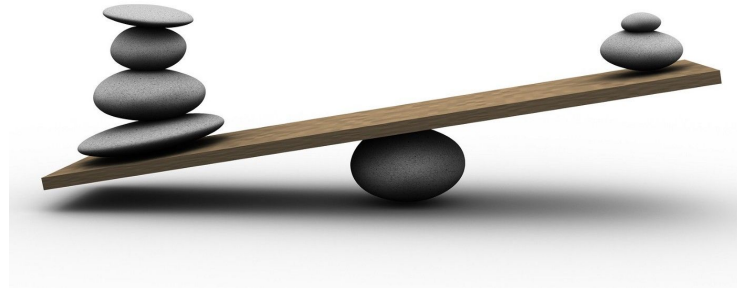
The best way to approach Machine Learning solution is to start by analyzing and exploring the dataset in **EDA**

One of the common issues found in datasets is imbalanced classes issue

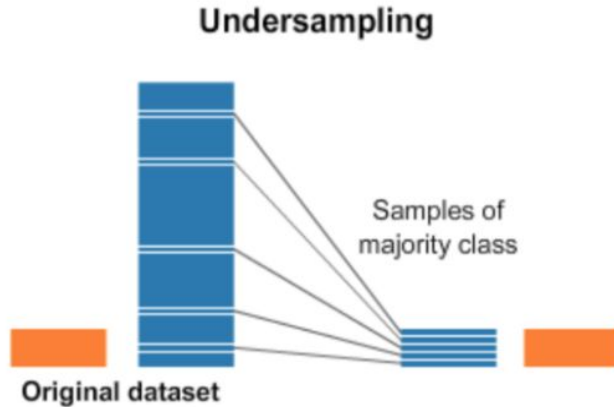
Data imbalance usually reflects an unequal distribution of classes within a dataset

Binary classification models without fixing this problem will be very biased

Feature correlation need to be improved

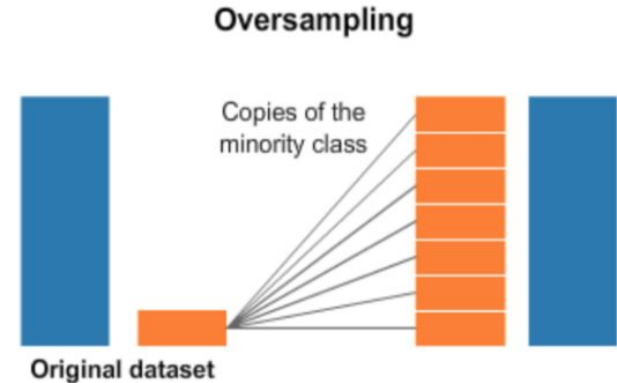


Undersampling



Randomly deleting some of the observations from the majority class in order to match the numbers with the minority class.

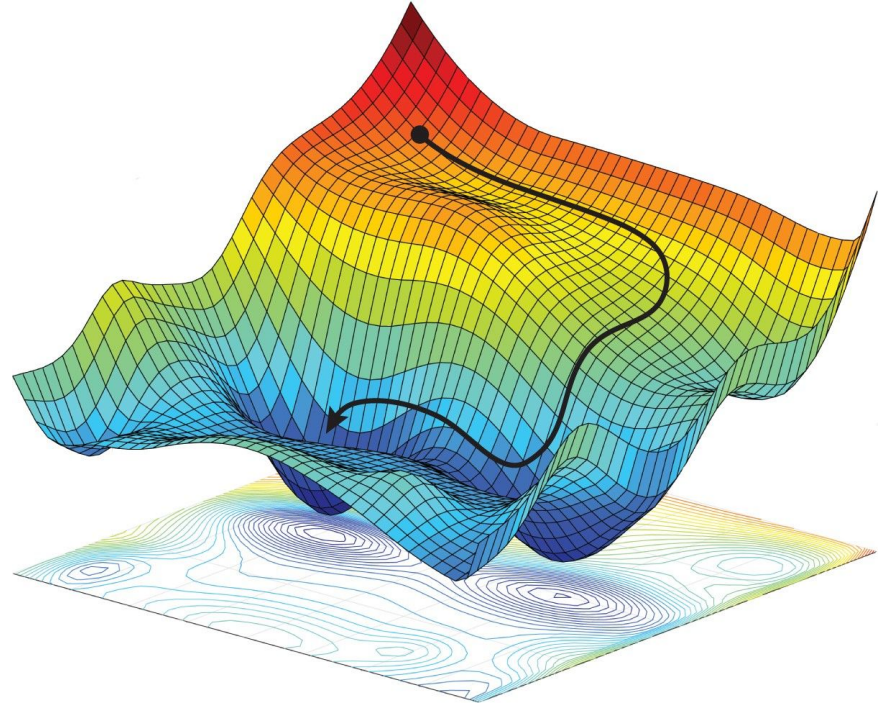
Oversampling



Synthetic Minority Over-sampling Technique (SMOTE) – looks at the feature space for the minority class data points and considers its k nearest neighbours.

Optimization for Machine Learning

- **Feature selection**
- **Model training**
- **Hyper-parameter tuning**
- Others



Making Better Decisions in an Uncertain World

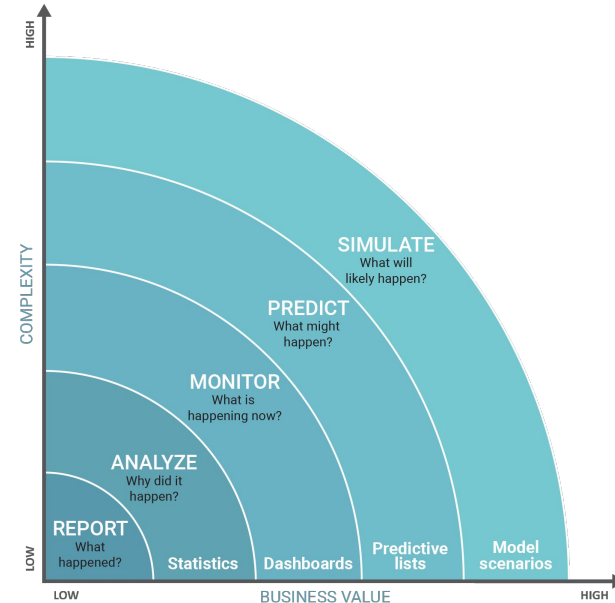
Optimization techniques and applications

- Exhaustive search
- Local search
- Gradient descent
- Nature-inspired methods

- Agents
- Games
- Simulations

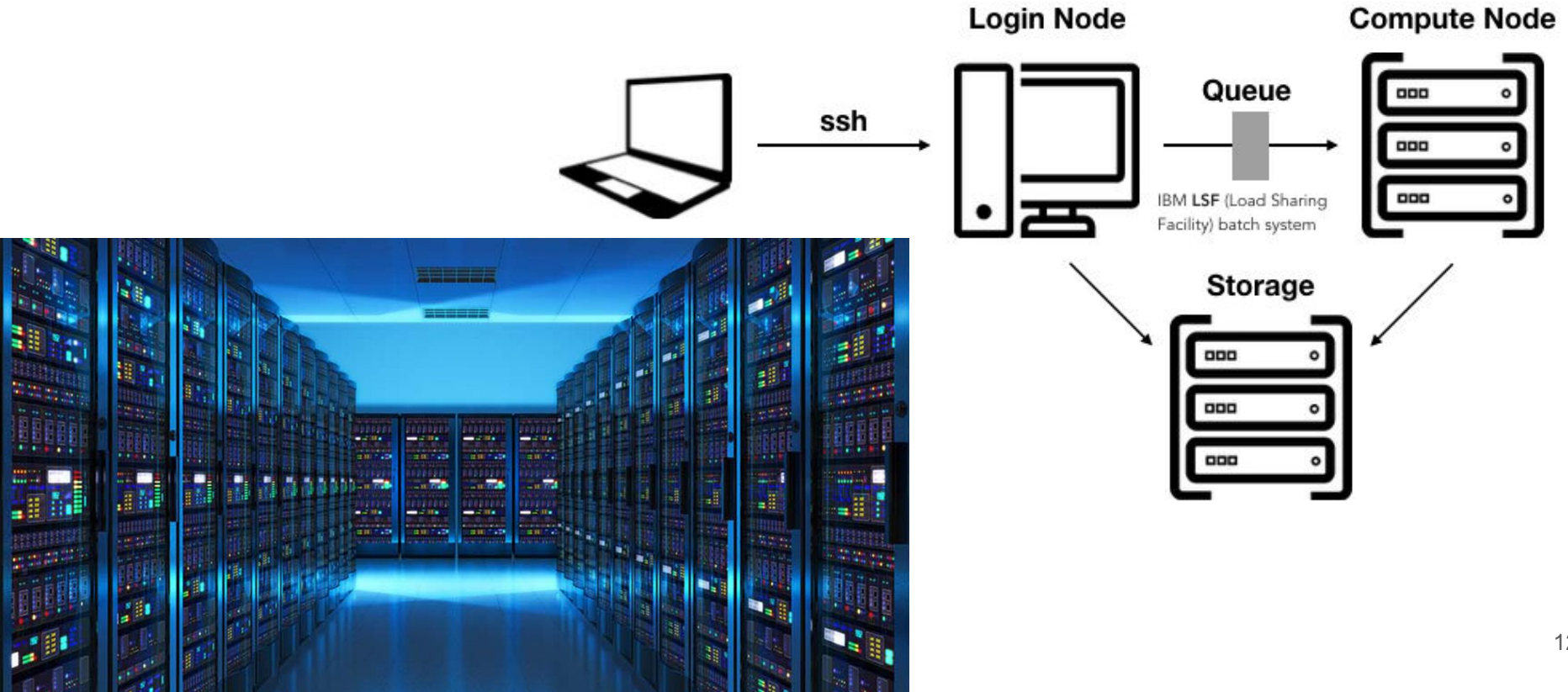
Imperfect but real-time decisions

Approximately optimal solutions in an acceptable time

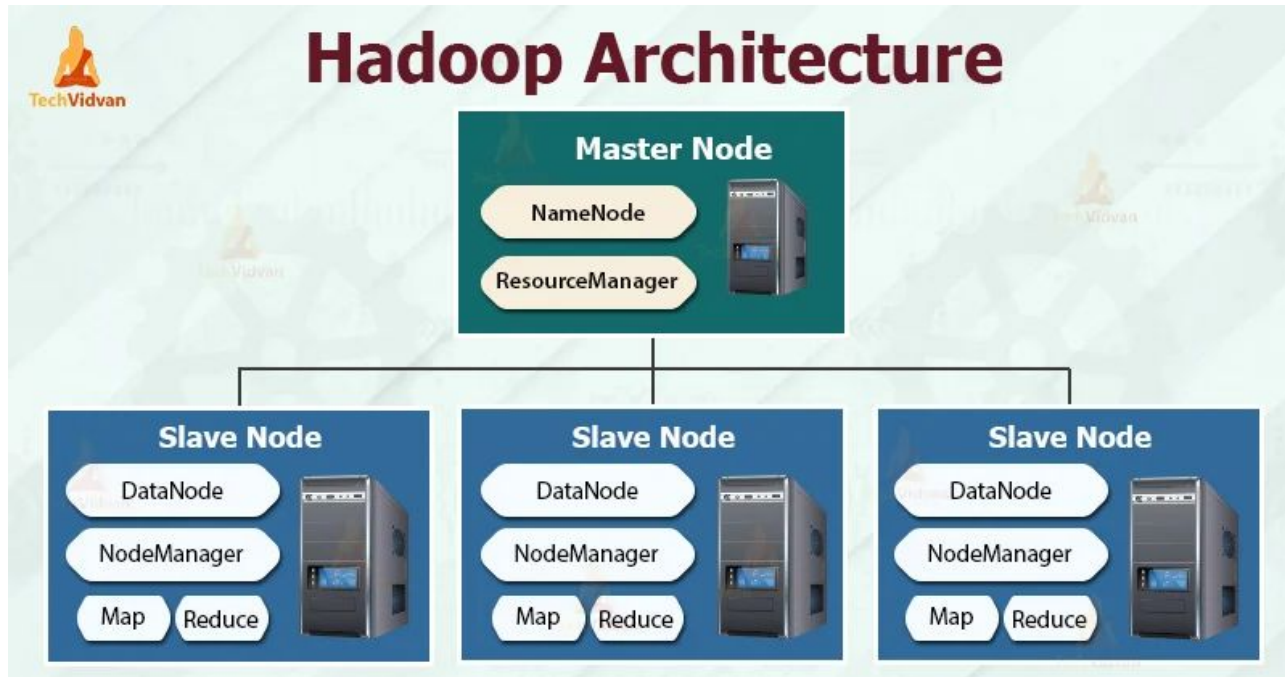


2003–2023: A Brief History of Big Data

HPC cluster and compute-intensive applications



Hadoop cluster and data-intensive applications



Summing up 20 years of Hadoop

2008–2012 Rise of the Hadoop vendors

- Cloudera, Hortonworks + MapR, Datastax
- Column-oriented storage format:
Apache Parquet, Apache ORC

2010–2014 Hadoop 2.0 and the Spark

- Hadoop resource manager: **YARN**
(Yet Another Resource Manager)
- **Apache Spark** faster, better, and great replacement for MapReduce
ML into the ecosystem → Apache Mahout,
Apache MLlib
- LinkedIn Apache Kafka
- Twitter Apache Storm
- Amazon (cloud) AWS with Netflix
- Microsoft Azure (cloud)
- **Google Cloud Platform**

2014–2016 Reaching the top

- Apache Spark, Apache Flink, Google's DataFlow
- Airbnb Apache **Airflow** (open-source schedulers)
- Spotify **Luigi** (open-source schedulers)
- Google's BigQuery
- Amazon's Redshift
- Snowflake

2016–2020 Containerisation and DL → the downfall of Hadoop

- Massive migration of data infrastructures to the cloud
- Amazon S3, Google Storage or Azure Blob Storage are used instead of HDFS
- Docker containerisation framework
- **Kubernetes** (container orchestration system)
- **DL** → **Tensorflow, Keras, PyTorch**

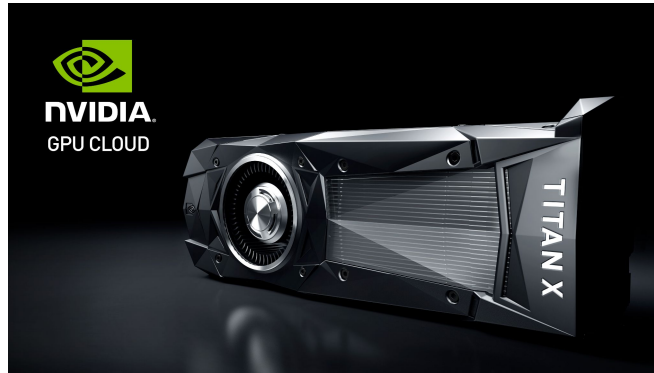
The need of GPUs – they are not in Hadoop clusters

Current shift from Hadoop ecosystems to
Cloud services like Kubernetes

The rise of Deep Learning
The downfall of Hadoop



Hardware and software available for DL



Deep Learning (DL) evolution timeline

1940s NN were proposed
1960s DNN were proposed
1990 **the first DL application** →
handwritten digit recognition – **LeNet**

Since 2010 **DL success** is due to the 3 factors

1. new algorithmic advances → better performance
2. availability of huge amount of data to train NNs
3. increase of computing power

Stanford Vision Lab, Stanford University,
Princeton University

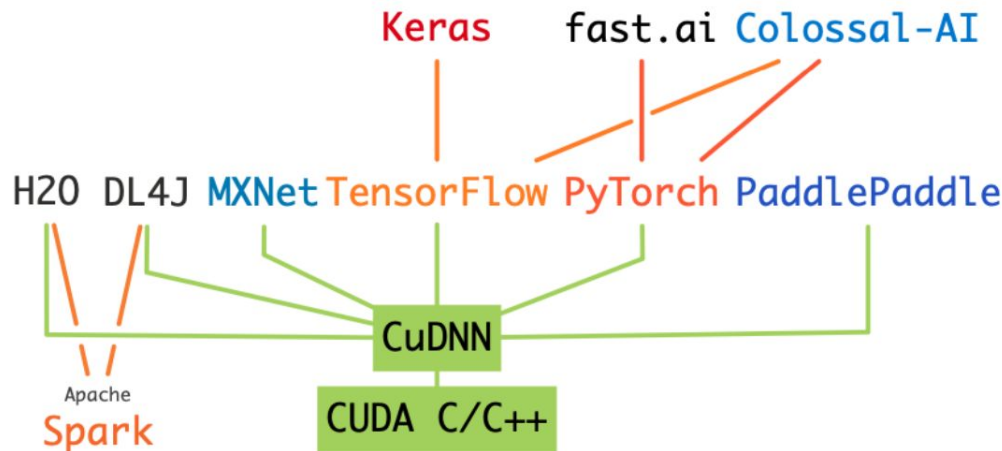
ImageNet Large-Scale Visual Recognition Challenge ILSVRC (2010–2017)

- Computer vision challenge
- Evaluates algorithms for object detection and image classification at large scale

Motivations

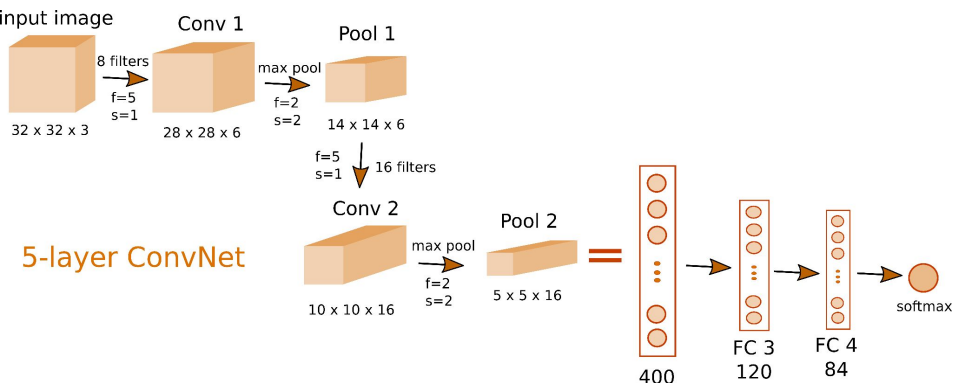
- to allow researchers to compare progress in detection across a wider variety of objects – taking advantage of the quite **expensive labeling effort**
- to measure the progress of **computer vision** for large scale image indexing for **retrieval and annotation**

Neural Networks and Deep Learning frameworks



Convolutional Neural Networks (CNN)

Input - Conv - MaxPooling - Conv - MaxPooling - FC - FC(Output)



Iconic CNN architectures

- **1990 LeNet** deployed in ATMs to recognize digits for check deposits
- 2012 AlexNet the first CNN wonned ILSVRC
- 2013 ZFNet one of the top at ILSVRC 2013
- **2014 VGG** classified as the second in ILSVRC
- 2015 GoogLeNet ILSVRC winner
 inception module composed of parallel connections
 which drastically reduced the number of parameters
- **2016 ResNet** Residual Net - ILSVRC winner
- 2016 SqueezeNet focuses in heavily reducing model size
 using deep compression
- UNet, **DenseNet**, MobileNet, ShuffleNet, ...

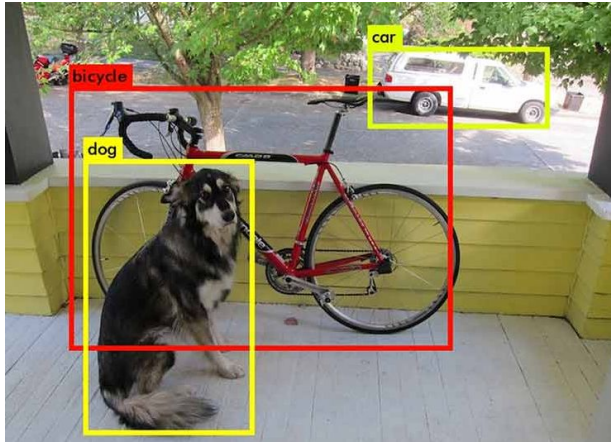
Object detection

Image classification: CNNs

- Classifies a picture
- Predicts probability of object

Classification ws. Localization: YOLO, R-CNN

- Detects an object in a picture
- Predicts probability of object and where it is located



Face recognition

Face verification (one-to-one lookup)

- Is this the correct person?

Face recognition (one-to-one lookup)

- Is this one of the k people in the database?



2020: contradictory year for facial recognition

Jan 2020: America's first confirmed wrongful arrest by facial recognition technology (Robert Williams, **Black man**)

Jan 2020: **ClearviewAI** – small facial recognition company that ran its algorithm on a DB of **billions of pictures** grabbed **from social media**

Facial recognition is known to be less accurate for **darker-skinned** people

Facial recognition was/is widely used by **police departments** in the United States

The use of biometrics → Organic part of digital life

- Apple Face ID system to unlock iPhones

Image segmentation

Image segmentation

- segmenting an image into fragments
- assigning a label to each of those

Categories

- **Semantic segmentation**
- **Instance segmentation**
- Panoptic segmentation: combination of both semantic and instance segmentation

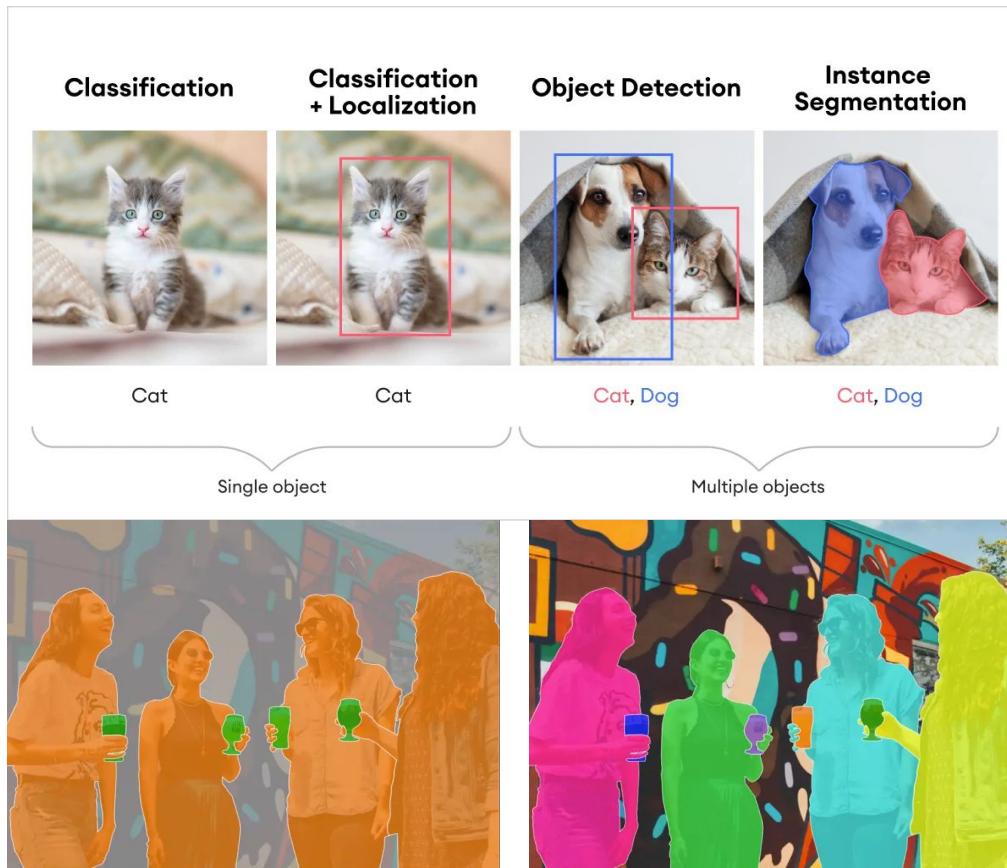


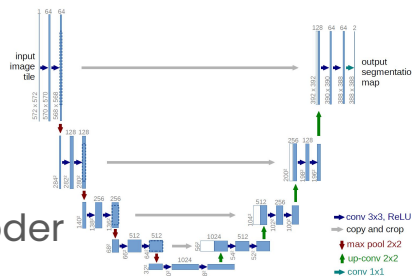
Image segmentation techniques

Classic

- Region-based segmentation
- Edge detection segmentation
- Thresholding
- Clustering

Deep learning

- U-Net: encoder-decoder
- [Mask R-CNN](#)
- DeepLab Versions
- Interactive segmentation
- **Meta's SAM (Segment Anything Model)** → privacy-preserving images



Benchmark datasets

- [Pascal VOC dataset](#)
- [MS COCO dataset](#)

Metrics

- [IOU \(Intersection Over Union\)](#)

Applications

Medical imaging, Video surveillance, Autonomous vehicles, Agriculture, Satellite imagery, Robotics, Art and design, Gaming, Fashion and retail

R-CNN and YOLO → Real-time object detection

Object detection: YOLO, R-CNN

- Detects up to several objects in a picture
- Predicts probabilities of objects and where they are located

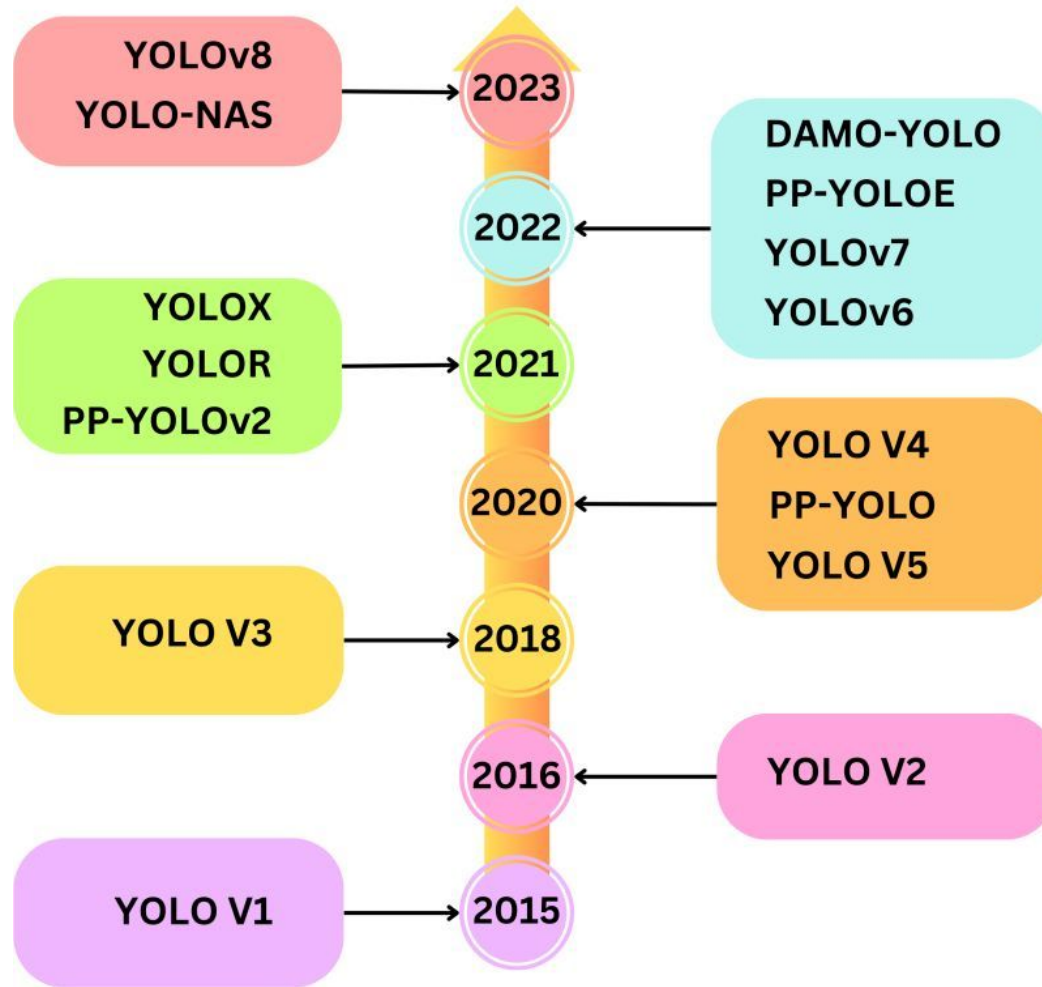


You only look once (YOLO)

- **YOLO v3**: Joseph Redmon, 2018
 - announced his Computer Vision research stop due to ethical concerns ([Twitter 2.2020](#))
- **YOLO v4**: A. Bochkovskiy, fork, 4.2020
- **YOLO v5**: G. Jocher, PyTorch, 6.2020
- **PP-YOLO**: X. Long, Baidu, 7.2020
- **YOLOv7** in 2022
- **YOLOv8** in 2023

The YOLO timeline

TheAiEdge.io



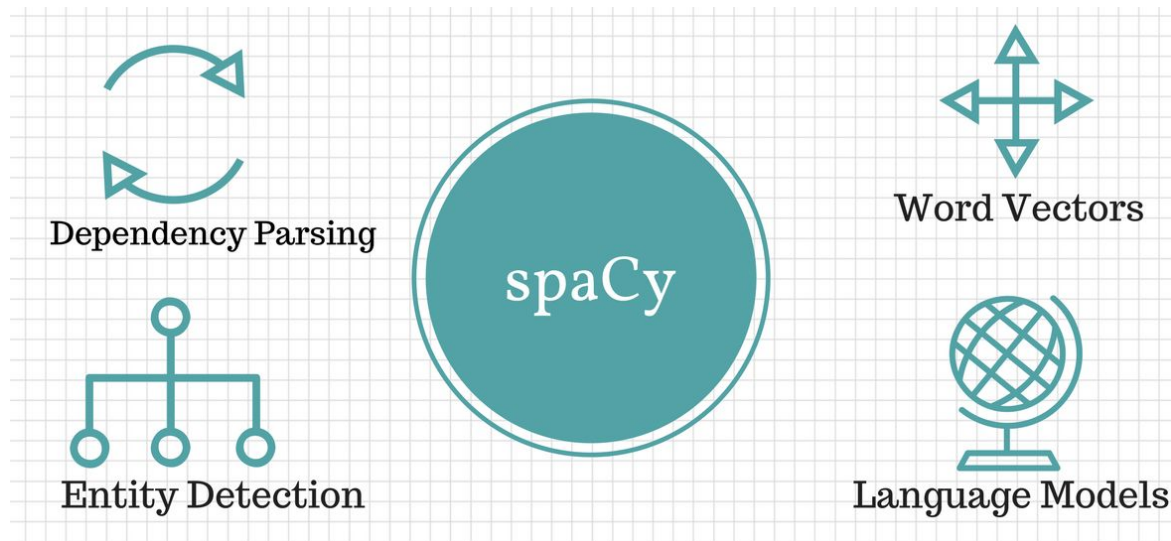
Recurrent Neural Networks

Encoder-Decoder

Natural Language Processing (NLP)

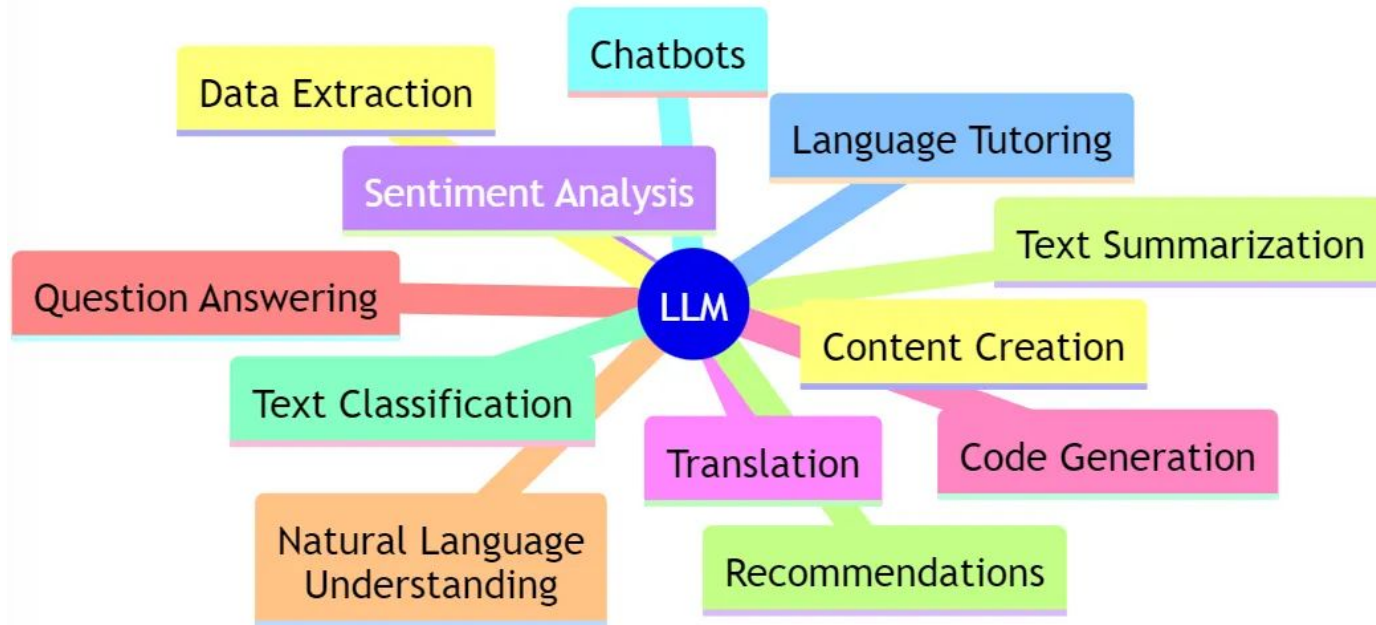


2019-2020 Big Years of NLP and DL



Natural Language Processing (NLP)

- Feature Extraction for ML – The Art of the Science
- NLP in the context with ML/DL/AI
 - **NLP Pipelines**, [NLTK](#), Stanza, Spacy, scikit-learn, ...
 - Basic NLP task for text processing: stop-words removal, tokenization, ...
 - Basic text processing: regular expression, normalization, extraction, ...
 - Basic language modeling: probabilistic N-grams
- Text Modeling with Machine Learning
 - **Sentiment Analysis in Text**
- Text Preprocessing as Feature Extraction for ML
 - Text representation: **hand-crafted** and automatic
 - Feature Extraction: **Bag-of-Words, IF-IDF**, Word Embedding
 - Distance metrics, Document similarity
 - Encoding Data, Feature Selection, Dimensionality Reduction
- Distributional Semantics and Contextual Representation
 - NLP Transformer family: Encoder-Decoder
 - **Pretrained Large Language Models (LLM) with Deep Learning (DL)**



Model Deployment and Curation

Model Deployment – 87% have never been

4 years ago

it was unusual to find a company already deploying models in production.

2 years ago

it was unusual to find a company with 50 models in production.

Today

2 independent surveys estimate that

1/3 of companies that use AI have more than 50 models in production

Algorithmia (2021) said **40%**,

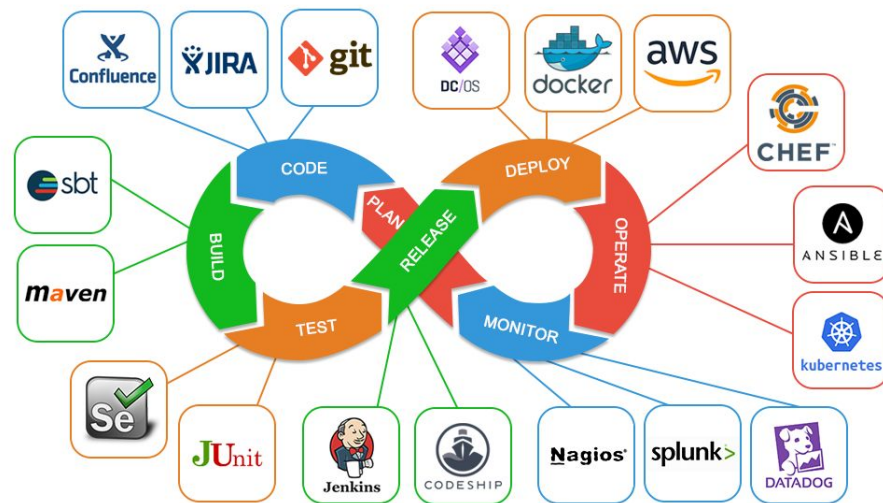
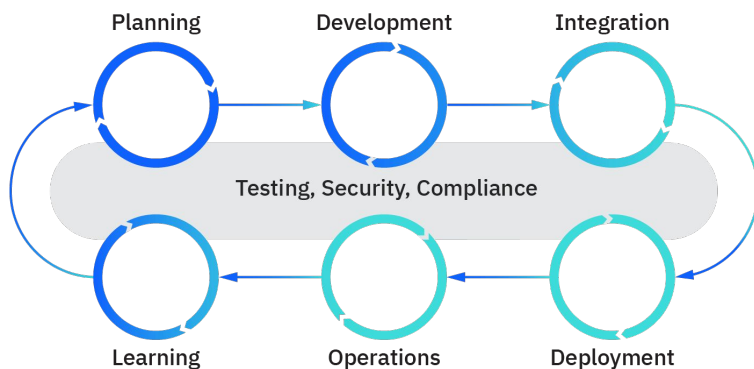
Arize (2022) said **36%**

Deploying your ML model is just the Beginning

Data and model curation in production!

Deployment → CD/CI → DevOps

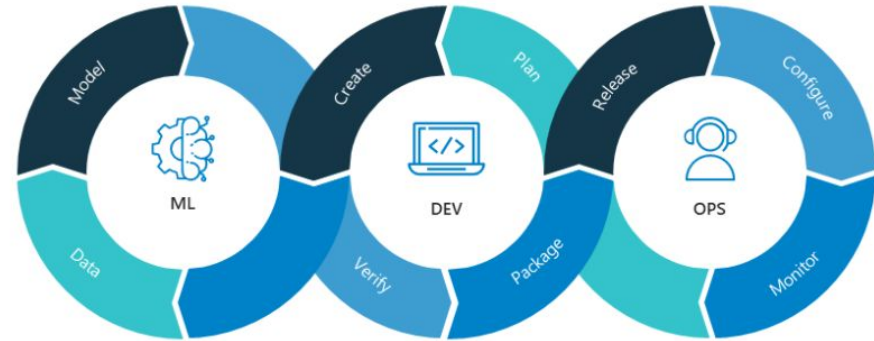
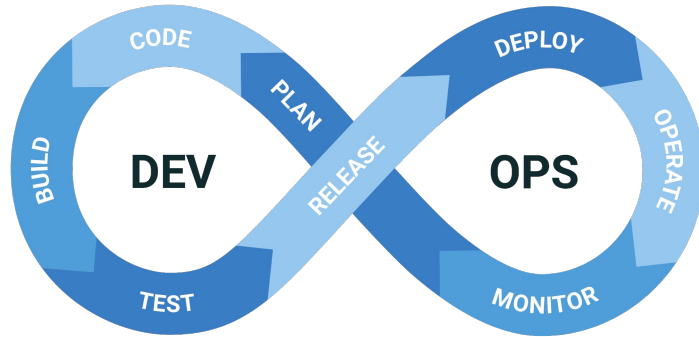
DevOps speeds delivery of higher quality software by combining and automating the work of software development and IT operations teams



DevOps → MLOps and DataOps

DevOps

MLOps



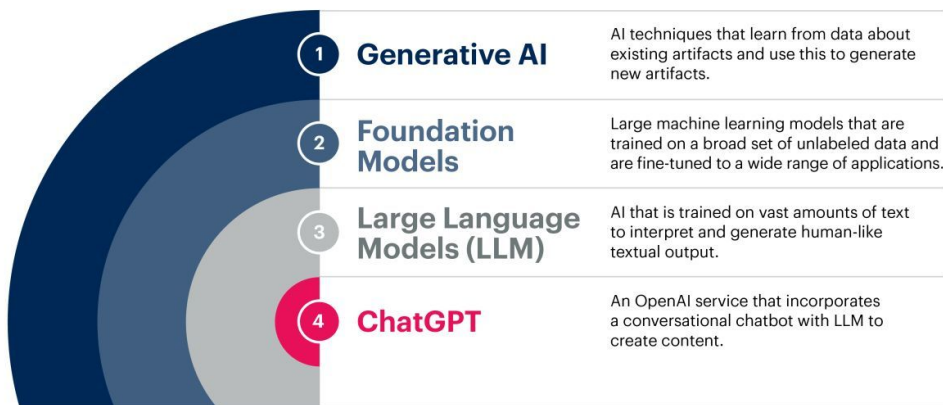
DevOps = Faster software delivery + software quality assurance (QA)
 MLOps = DevOps applied to Machine Learning (ML) operations

Data distributions: DevOps and MLOps

In real world scenario of Machine Learning (ML) deployment

- The change in data distribution is called **data drift**
 - Data drift affects the performance of ML model used in deployment
- Data drift leads to **concept drift** (another name **model drift**)
 - Degradation of ML model performance in production
- Heavy work of ML engineering in **DevOps** and **MLOps**

What Is Generative AI?



Source: Gartner
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. 2421958

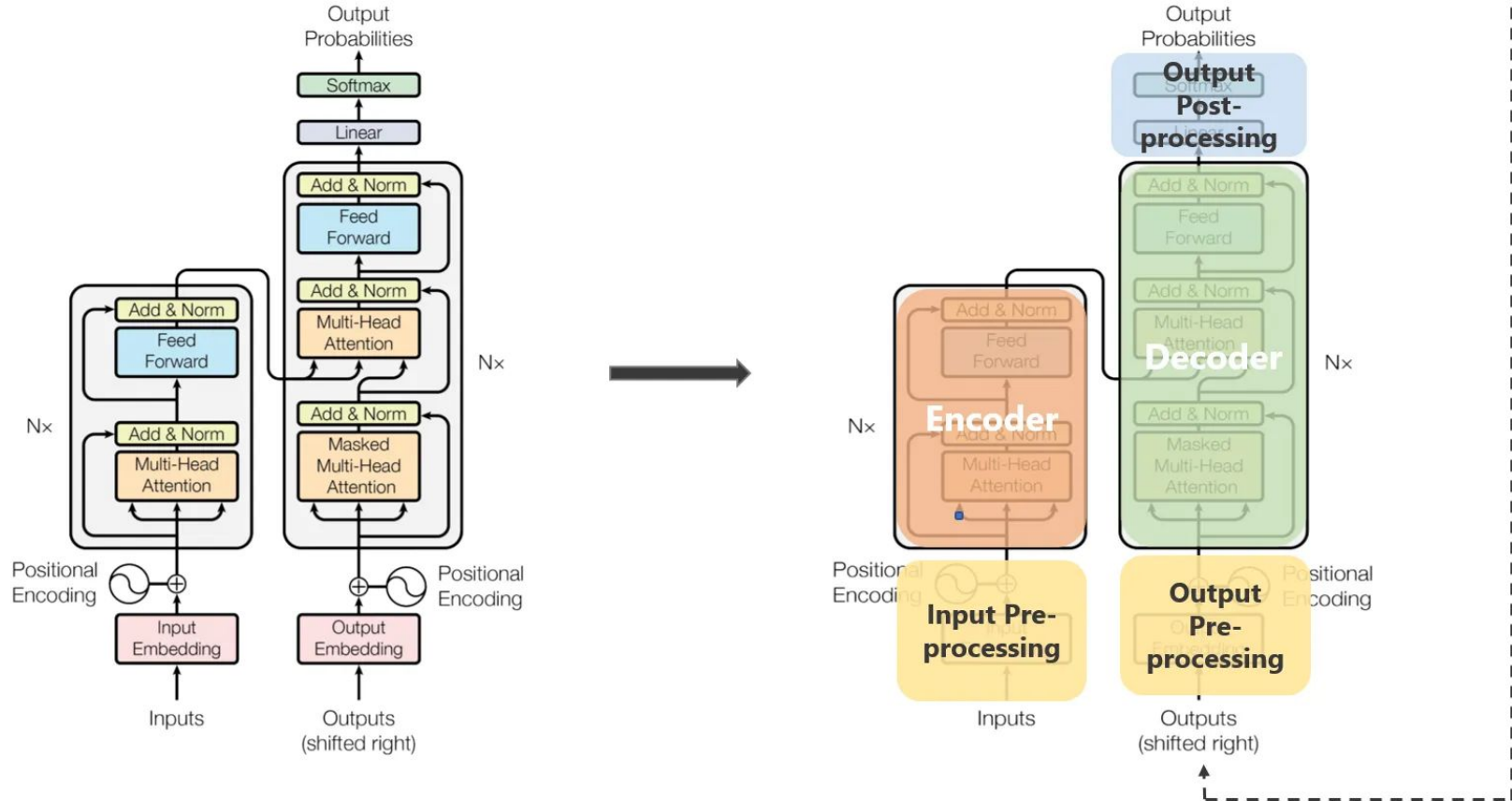
Gartner.



(NLP) Transformers

Transformer (2017)

"Attention is all you need"



NLP Transformers: Benefits and Drawbacks

- **highly parallelizable**, meaning that they can process multiple parts of a sequence at the same time → significantly speeds up training and inference
- capture **long-term dependencies** in text → better understand the overall context and generate more coherent text
- **high computational demand**
- **sensitive to the quality and quantity of the training data**



2022 Leading Large Languages Models (LLMs)

- BERT** Pre-training of Deep Bidirectional Transformers for Language Understanding
- GPT-2** Language Models Are Unsupervised Multitask Learners
- XLNet** Generalized Autoregressive Pretraining for Language Understanding
- RoBERTa** A Robustly Optimized BERT Pretraining Approach
- ALBERT** A Lite BERT for Self-supervised Learning of Language Representations
- T5** Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
- GPT-3** Language Models Are Few-Shot Learners
- GPT-3.5, GPT-4, ...**

Vision Transformer (ViT)

2017, Transformers introduced for NLP

2020, adapted for computer vision → ViT

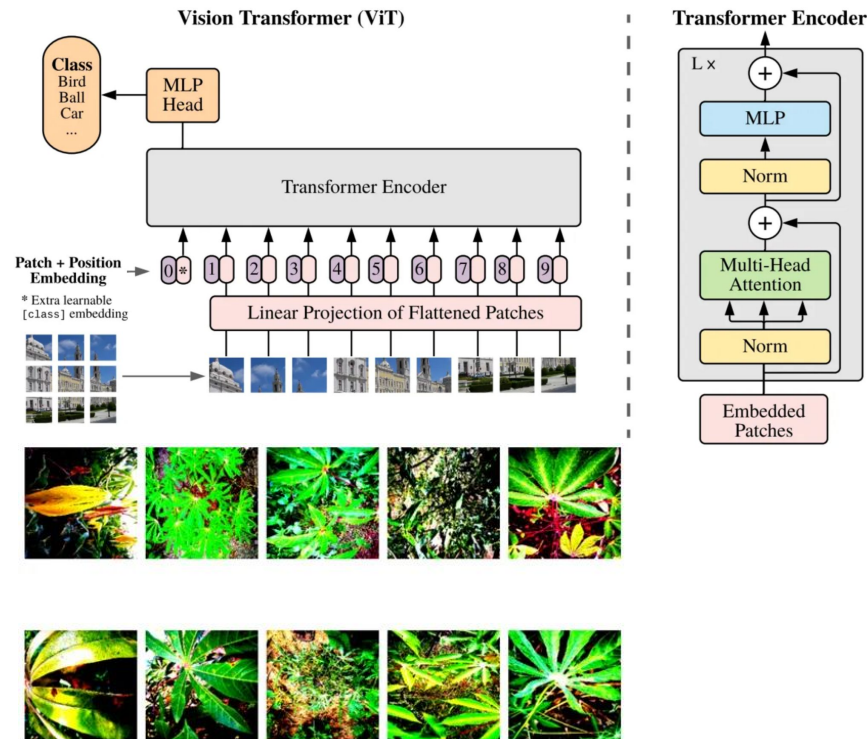
- The attention mechanism in a ViT repeatedly transforms representation vectors of **image patches**, incorporating more and more **semantic relations** between image patches in an image (like NLP way)
- CNNs achieve excellent results even with training based on data volumes that are **not as large** as those required by Vision Transformers



Vision Transformer vs. CNN

- ViT possess a different kind of bias toward exploring topological relationships between patches, which leads them to be able to capture also global and wider range relations but at the cost of a more onerous training in terms of data
- ViT more robust to input image distortions

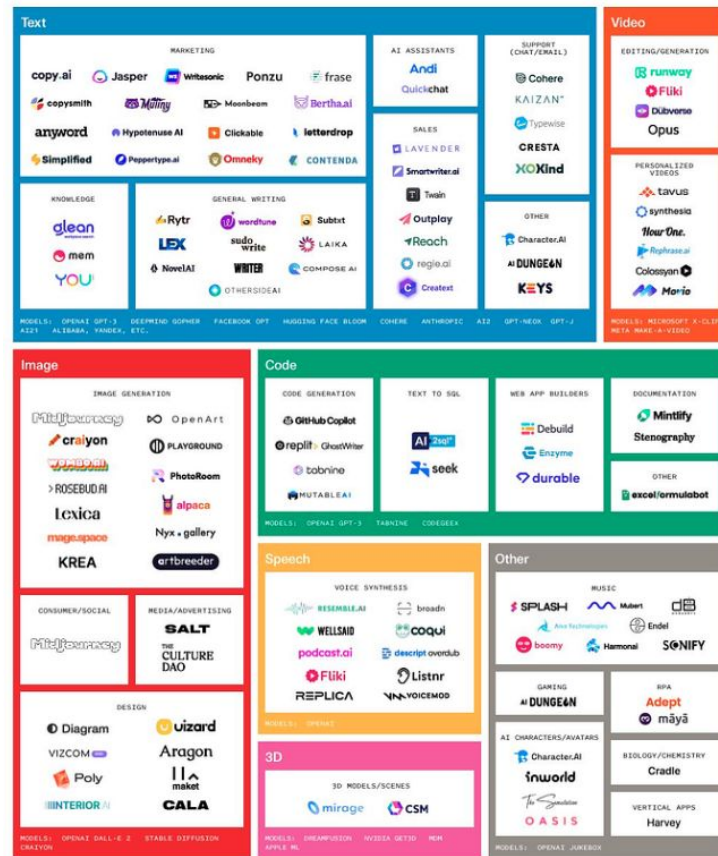
Not so clear winner between CNN and ViT
→ combination of both



Large Multimodal Models (LMMs)

Generative AI applied to other modalities

- **Image Generation:**
Dall-E | MidJourney | Stable Diffusion | DreamStudio
- **Audio Generation:**
Whisper | AudioGen | AudioLM
- **Search Engines:** Neeva | You
- **Code Generation:** Copilot | Codex
- **Text Generation:** Jasper



Everything is still **NOT** working

2022: Why Meta's latest large language model survived only **three days** online

- Meta = Facebook's parent company
- **Galactica** – new large language model designed to **generate scientific text**
- Demo online **2022 November 15**, take down after **3 days**
- Model trained on **48 million** examples of
 - scientific articles,
 - websites,
 - textbooks, lecture notes, and
 - encyclopedias
- blind spot → severe limitations of large language models
- not able to distinguish truth from falsehood



2016: Microsoft launched a **chatbot** called **Tay on Twitter** — then shut it down **16h** later when Twitter users turned it into a racist, homophobic sexbot

GALACTICA

Language Models that Cite

GALACTICA models are trained on a large corpus comprising more than 360 millions in-context citations and over 50 millions of unique references normalized across a diverse set of sources. This enables GALACTICA to suggest citations and help discover related papers.

Machine Learning Math Computer Science Biology Physics

Input:

The paper that presented a new computing block given by the formula:

$$f(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

GALACTICA Suggestions:

Attention is All you Need

Vaswani et al., 2017

[View full prompt](#)

Examples: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#)

It is amazing, threatening, challenging. The technology WILL BE incredible.
AI and Ethics?

Something is working :) sometime

Language model: more truthful and **less toxic**, trained with **human-in-loop**

OpenAI just released GPT-3 text-davinci-003,

The results are impressive in comparison with with 002!





Brief summary on (LLM) GPT-3 hype

Even with 175B parameters and 450 gigabytes of input data, it's **not a reliable** interpreter of the world

Models presents **security and uncontrollability problems**, including false content and biased information during content generation

Unreliable doesn't mean useless. This is not to say that GPT-3-related is devoid of practical applications

Can be considered to use as **intelligent auxiliary tasks**

Cannot directly interface with the end-users

GPT-3 is a microcosm of the best and worst example in AI today (W. D. Heaven, 2021)

LLMs and LMMs demands

- Significant step in technological evolution
- Resource demands
 - LLMs are usually trained in **high performance infrastructures** with **multiple GPUs** for several months, e.g., Microsoft Azure infrastructure
 - The energy cost is **several hundreds thousands EUR/USD per month**
 - Each LLM/LMM: a “**piece**” made from long-time high energy consumption
- Unrepairable logical **inconsistencies** and inaccuracies in their outputs, which come from the design as the most probably output – not the correct output
- **Ethical concerns: explainability, content control, bias mitigation**

Responsible AI

Principles of Artificial Intelligence, Privacy, Fairness and Data Fusion



Responsible AI – the only way to mitigate AI risks

1. **Privacy**
2. **Security and Safety**
3. **Ethics**
4. **Fairness**
5. **Accountability**
6. **Transparency**

Responsible AI is a standard for ensuring that AI is safe, trustworthy and unbiased

Responsible AI

As creators of artificial intelligence systems, we have a duty to guide the development and application of AI in ways that fit our social values



Accountability
Fairness
Public Trust

Trustworthy AI

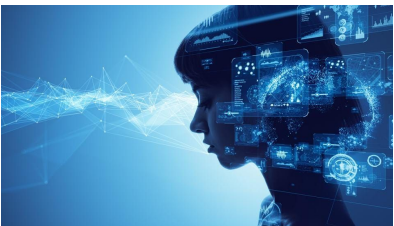
- AI opens up virtually limitless benefit potential
- Potential risks from AI, by addressing any areas where controls or processes are found to be lacking or inadequate. The risks include
 - biased decision-making
 - the interpretability of AI decisions
 - a lack of explainability
- **Trustworthy AI is a methodology** for the implementation of AI methods in real organizations with
 - fairness
 - model explainability, and
 - accountabilityin its core

The purpose of Artificial Intelligence

is to augment

Human Intelligence

(not replacement)



Human-Centric AI (HCAI)

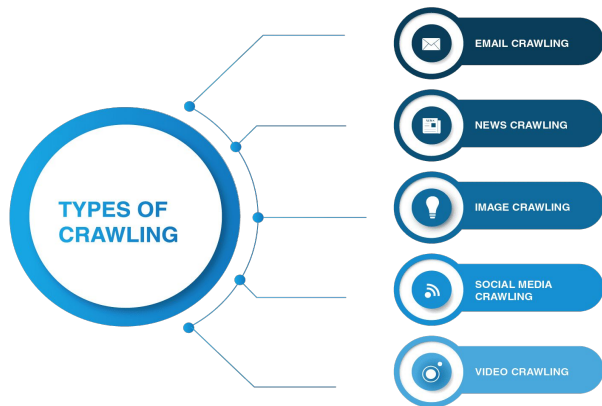
“The **human-centric** approach to AI strives to ensure that **human values** are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for **fundamental rights**, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for **human dignity**, in which the human being enjoys a unique and inalienable **moral** status.

This also entails consideration of the natural environment and of other living beings that are part of the **human ecosystem**, as well as a **sustainable** approach enabling the flourishing of **future generations** to come.”



GDPR (2016, EU)

General Data Protection Regulation (GDPR)



DSA (2022, EU)

Regulation on Digital Services Act (DSA)



Privacy – It is all about personal data – Personal Identifiable Information (PII)

Privacy-Preserving Data Mining (PPDM)

Data Transformations

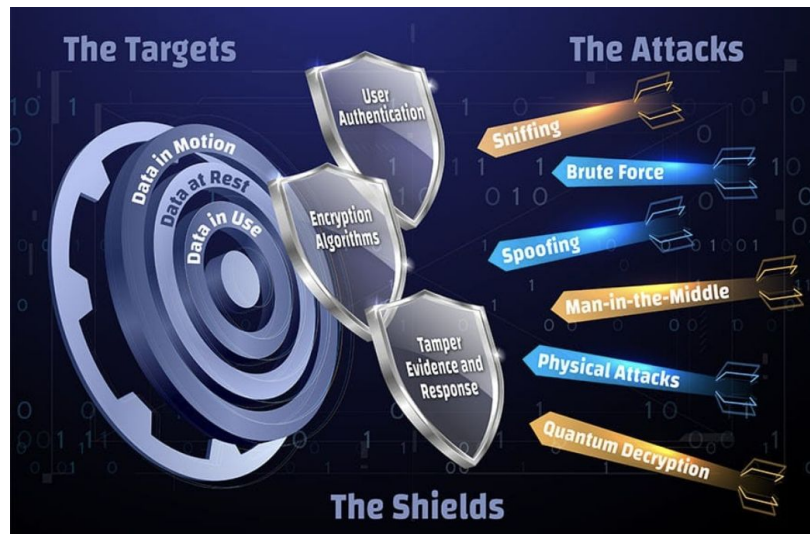
- Pseudonymization
- Anonymization: k -anonymity, l -diversity
- Differential Privacy

Secure Computations

- Encryption: symmetric, asymmetric, hashing
- Homomorphic Encryption: fully, partially

Trusted Environments

- Specialized Hardware
- Secure Execution Environment



Data Protection and Federated Learning

Centralized learning

- Problems: access right, owner right

Distributed learning

- Problems: sharing agreements, retention and disposal, minimization



Federated Learning

- Multiple clients train a model under central server orchestration, while keeping data decentralized
- Sharing and aggregating the model updates as transferred artifact



Responsible AI

Responsible development of human-centric and trustworthy AI systems

Thank you for your attention!
AI Data and Model Landscape

Giang Nguyen

