Building LLM Solutions

Module 1: What on Earth is going on!?

Hamza Farooq
Founder & CEO
Traversaal.ai

01

Introduction

Nice to meet you all!

- My name is Hamza Farooq
- Founder @traversaal.ai
- Ex- Google, Walmart Labs
- 15+ of experience in Machine Learning
- Adjunct Professor at Stanford & UCLA

Nice to meet you, from the TA!

- My name is Nikka Mofid
- Machine Learning Engineer @ Google
- Stanford Masters in Electrical Engineering focused on Machine Learning
- Ex-Nvidia, Adobe AI, Microsoft, and Amazon

Class Etiquette

- This is a virtual class, please keep your video on – as much as possible
- Please drop your questions in the chat, we will take breaks to answer questions
- Use Discussions forum for any questions

Caution

- This is an extremely technical course in terms of coding
- You will be coding for each assignment and in each class
- You will work in teams to submit the assignments

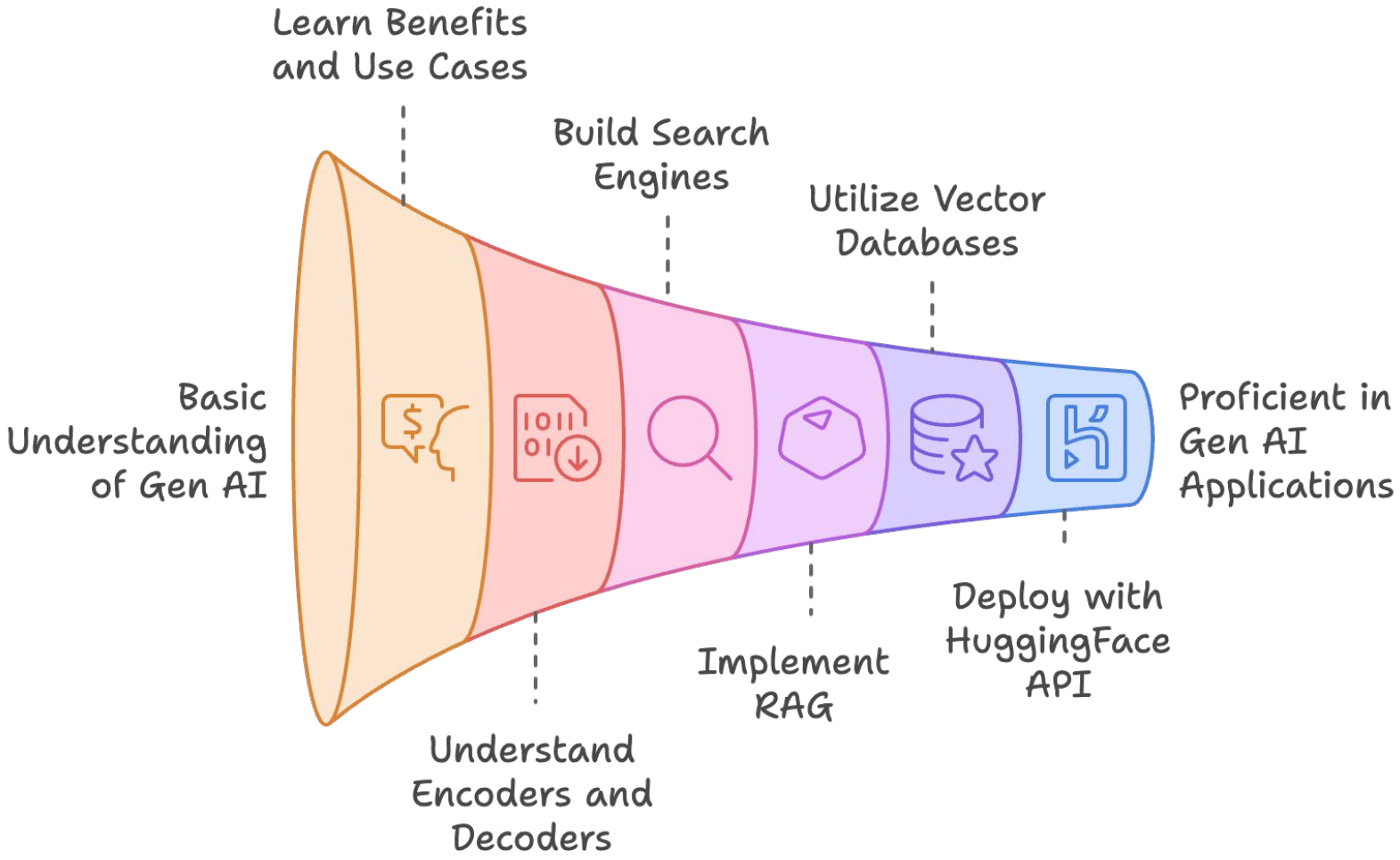
Course Overview

- Module 1: Intro to Everything
- Module 2: Overview of Transformer with overview of Encoders and Decoders Models, Self Attention
- Module 3: Building basic search using fundamental blocks
- Module 4: Full Scale Search
- Module 5: Generation using Decoder Models and API
- Module 6: Fine-tuning

Guideline for success in this course

- Able to answer the benefits and Use Cases of Gen AI
- Understand the following concepts and being able to work with:
  - Encoders
  - Decoders
  - Search Engines from scratch (keyword + Semantic)
  - RAG
  - Vector Databases
  - HuggingFace API Deployments (Gradio)
  - Fine Tuning

Rubric


Learn Benefits and Use Cases
Build Search Engines
Utilize Vector Databases
Basic Understanding of Gen AI
Proficient in Gen AI Applications
Understand Encoders and Decoders
Implement RAG
Deploy with HuggingFace API

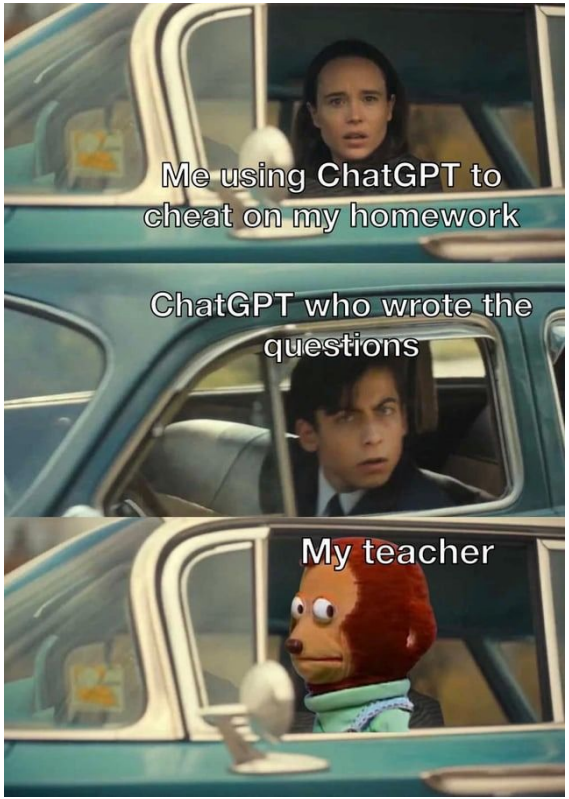
Learning outcomes for Module – 1

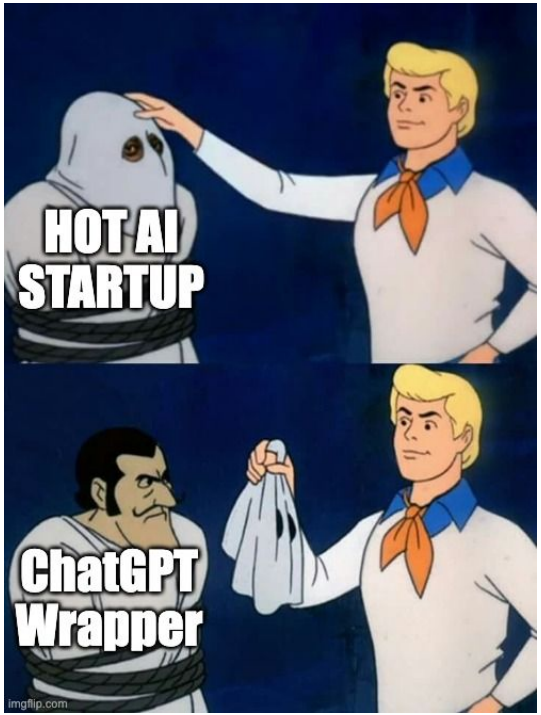
- Why on Earth do we need to change?
- Basic intro to Machine Learning
- What are foundational models?
- NLP Overview
- A larger overview of application of LLMs
- Working Prototypes

02

Here we go!

Why are we here?





The main question: What can Gen AI do for you?

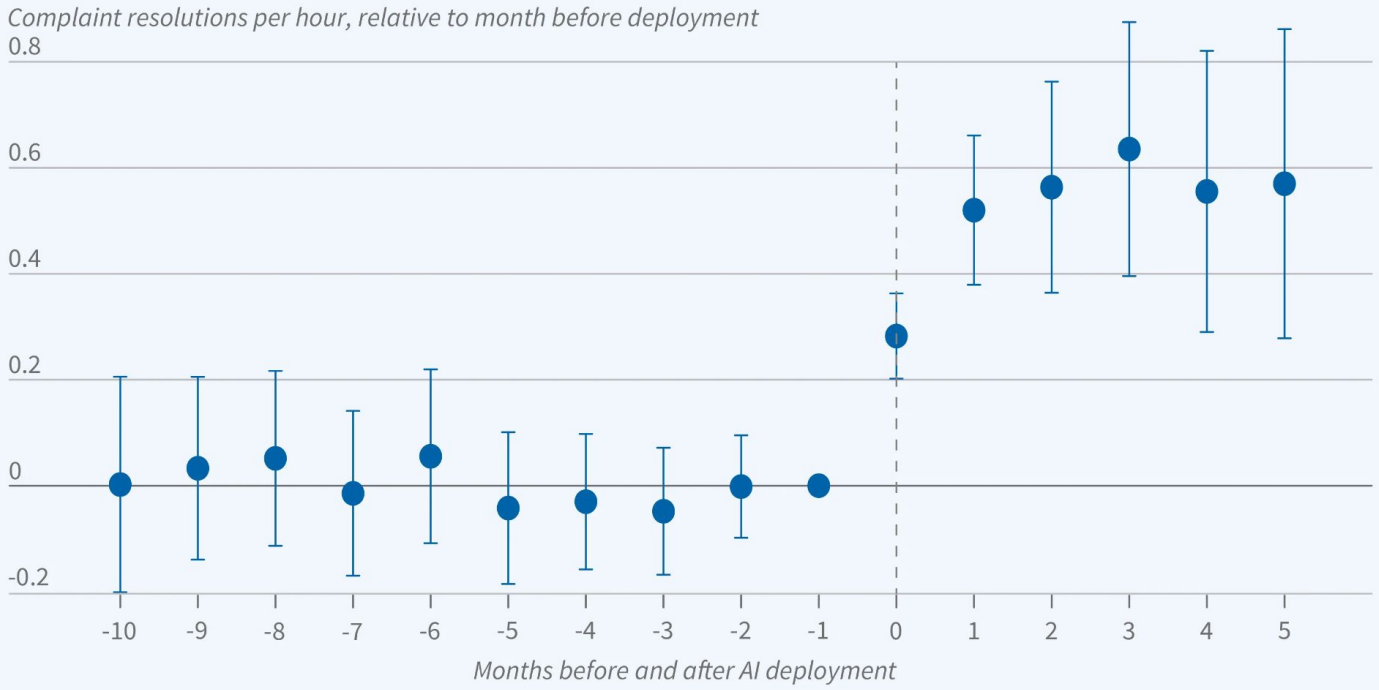

Why
Uber ?

Products 1970s → **Product + Services** 1990s → **Customer Centric** 2000s → **Experience Centric** Today

Source

One word: Productivity

AI Assistance and Customer Complaint Resolutions



Thin bars represent 95% confidence intervals
Source: Researchers' calculations using data from customer support agents provided by a Fortune 500 enterprise software company


Increased Innovation
Drive innovation with data-driven insights and predictive analytics, enhancing your user experience
Faster Time-to-Market
With AI, you can now automate tasks, analyze data and accelerate decision-making, reducing build times
Top 5 Benefits of AI to Propel Your Next-Gen Software Ahead
Cost Savings
Automate your tasks with AI and reduce the need for manual labor to gain improved operational efficiency
Improved Product Experience
Resolving issues is now faster with improved product usability and reduced human errors with AI
Enhanced UX Personalisation
Use AI-generated recommendations and improve the usability of interfaces based on user behavior

Source

Let’s explore our first example

Find me a stylish Nike blue t-shirt specifically designed for golf, with a comfortable fit and moisture-wicking fabric, in size medium and at a reasonable price point

Search

- Query 1: Find me a Nike t-shirt.

- Query 2: Find me a blue t-shirt.

- Query 3: Find me a t-shirt specifically designed for golf.

- Query 4: Find me a stylish t-shirt.

- Query 5: Find me a t-shirt with a comfortable fit.

- Query 6: Find me a t-shirt with moisture-wicking fabric.

- Query 7: Find me a t-shirt in size medium.

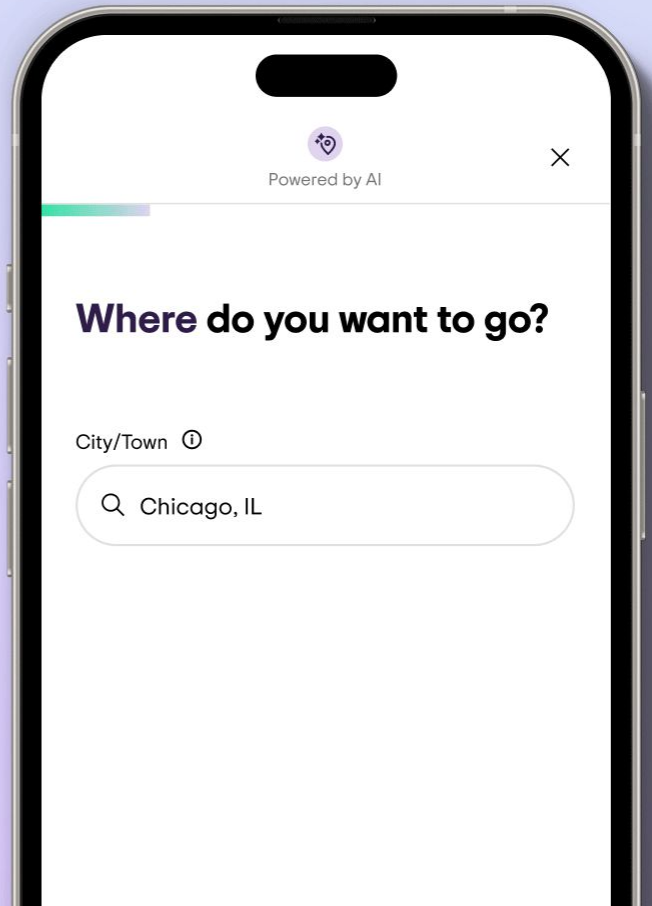
- Query 8: Find me a t-shirt at a reasonable price point.

- Combine Results:

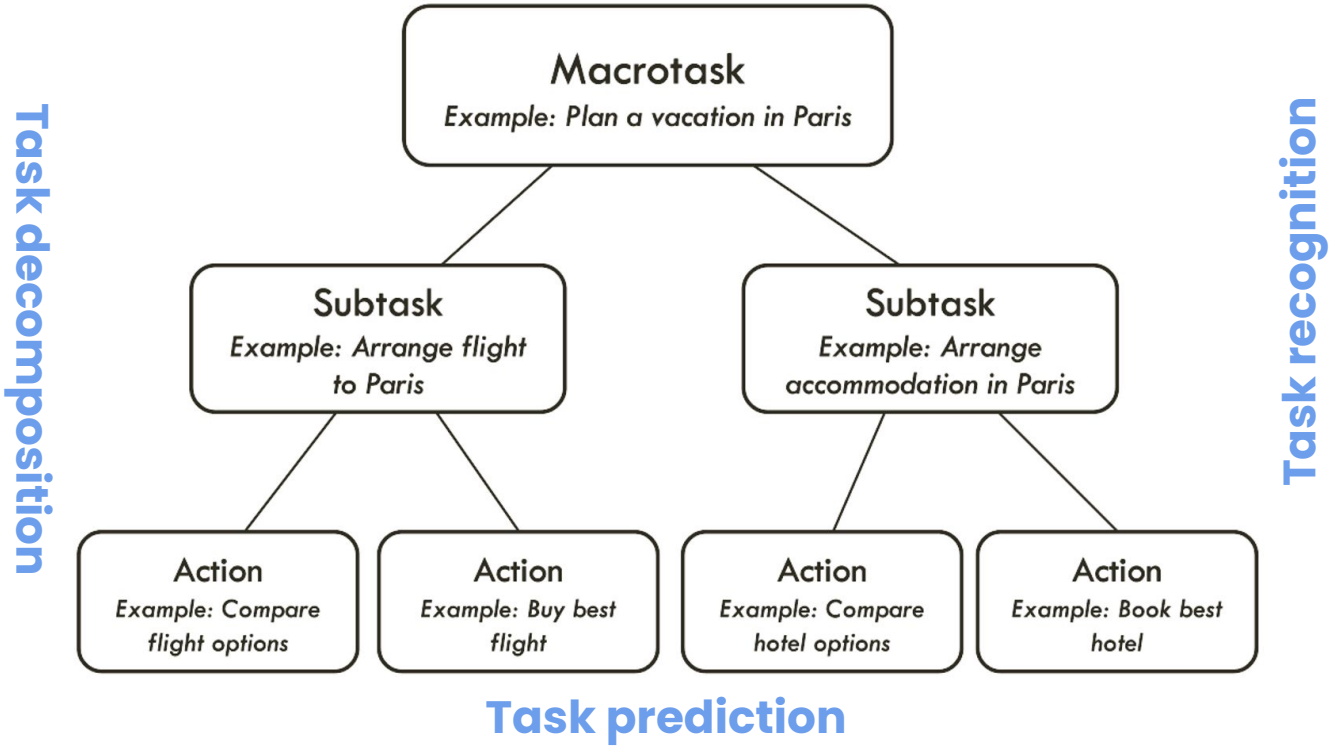
- "Find me a stylish Nike blue t-shirt specifically designed for golf, with a comfortable fit and moisture-wicking fabric, in size medium and at a reasonable price point."

Tripadvisor

Kick-start your travel planning



Dividing a complex search into subtasks, to then retrieve the results from them, and combine them into one answer, by searching the internet is quite a hard task



Source

And...

Complex search is difficult, multi-faceted and requires deeper engagement

- There are 10 billion search queries a day, **an estimated half of them go unanswered.**
- That's because, **people are using search to do things it wasn't originally designed to do.**
- It's great for finding a website, but **for more complex questions or tasks, too often it falls short.**
- Complex search tasks involve **multiple queries, multiple sessions or requires deep engagement with search**

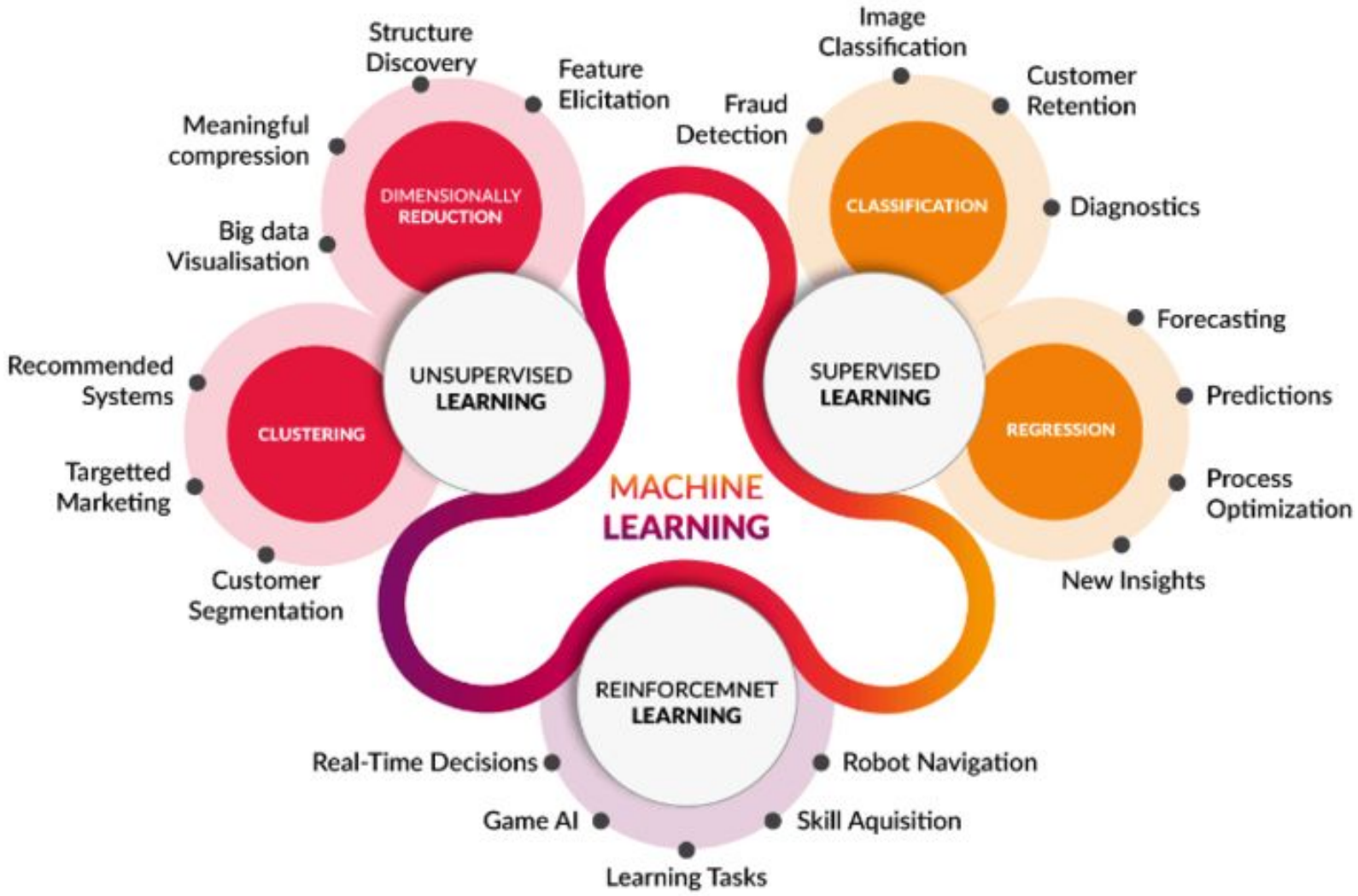
Let’s revisit some basic terminologies

Machine Learning Models

Machine learning is a discipline that focuses on creating algorithms and software capable of “learning” from the information we provide, and effectively perform specific types of tasks, as a result of this training.

Some examples:

- Netflix Recommendation Models
- Sales Prediction
- Self Driving Cars


MACHINE LEARNING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REINFORCEMNET LEARNING
DIMENSIONALLY REDUCTION
CLUSTERING
CLASSIFICATION
REGRESSION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
Recommended Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Fraud Detection
Diagnostics
Forecasting
Predictions
Process Optimization
New Insights
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Generative AI

Generative artificial intelligence (generative AI) is a type of AI that can create new content and ideas, including conversations, stories, images, videos, and music.

AI technologies attempt to mimic human intelligence in nontraditional computing tasks like image recognition, natural language processing (NLP), and translation.

Generative AI market in focus

Use case	Marketing	Creative writing	Sales
Text	ChatGPT, Jasper, WordAI, anyword, copy.ai, frase, Writesonic, copysmith, Hypotenuse AI, articleforge, Rytr, QuillBot, SE Ranking, byword_	wordtune, neuraltext, AI-WRITER, ShortlyAI, Scrivener, AI DUNGEON	Smartwriter.ai, LAVENDER, DRIFT, Outplay, Twain, exceed, people.ai, regie.ai
Video	Generation & editing: runway, Fliki, PICTORY, descript, invideo, lumen5, OPUS	Personalized video: tavus, synthesia, Colossyan	
Image	Generation & editing: Midjourney, plask, alpaca, Picsart, Playground, DALL·E, Jasper Art, Photosonic	Design: maket, INTERIOR AI, Looka, uizard, SWAPP	
Code	Code generation: GitHub Copilot, tabnine, seek, mutable.ai, Enzyme, Moderne	App building: appypie, mendix, Debuild, DURABLE	

Use case

Applications

Music

Generation & editing

splash, amper, Aiva, ecrett music, soundful, beatoven.ai, boomy

Music customization

SOUNDRAW, moises, LALAL.AI, audioshake, LANDR

Speech

Text-speech / speech-text

WELLSAID, coqui, RESEMBLE.AI, Speechify, LISTNR, REPLICA, PODCASTLE, LOVO, MURFAI

3D

Art

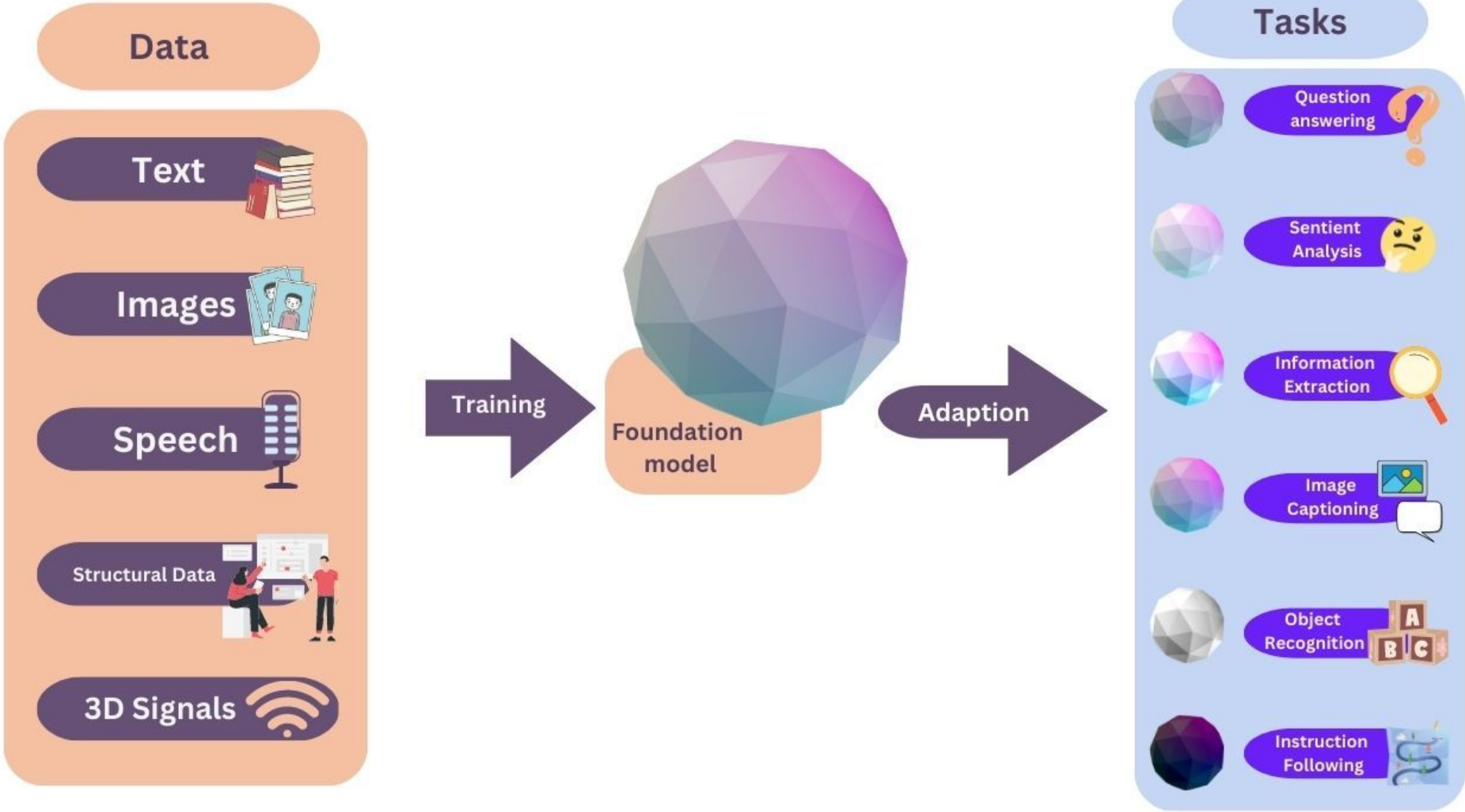
mirage, Spline, LUMA AI, DEEPMOTION

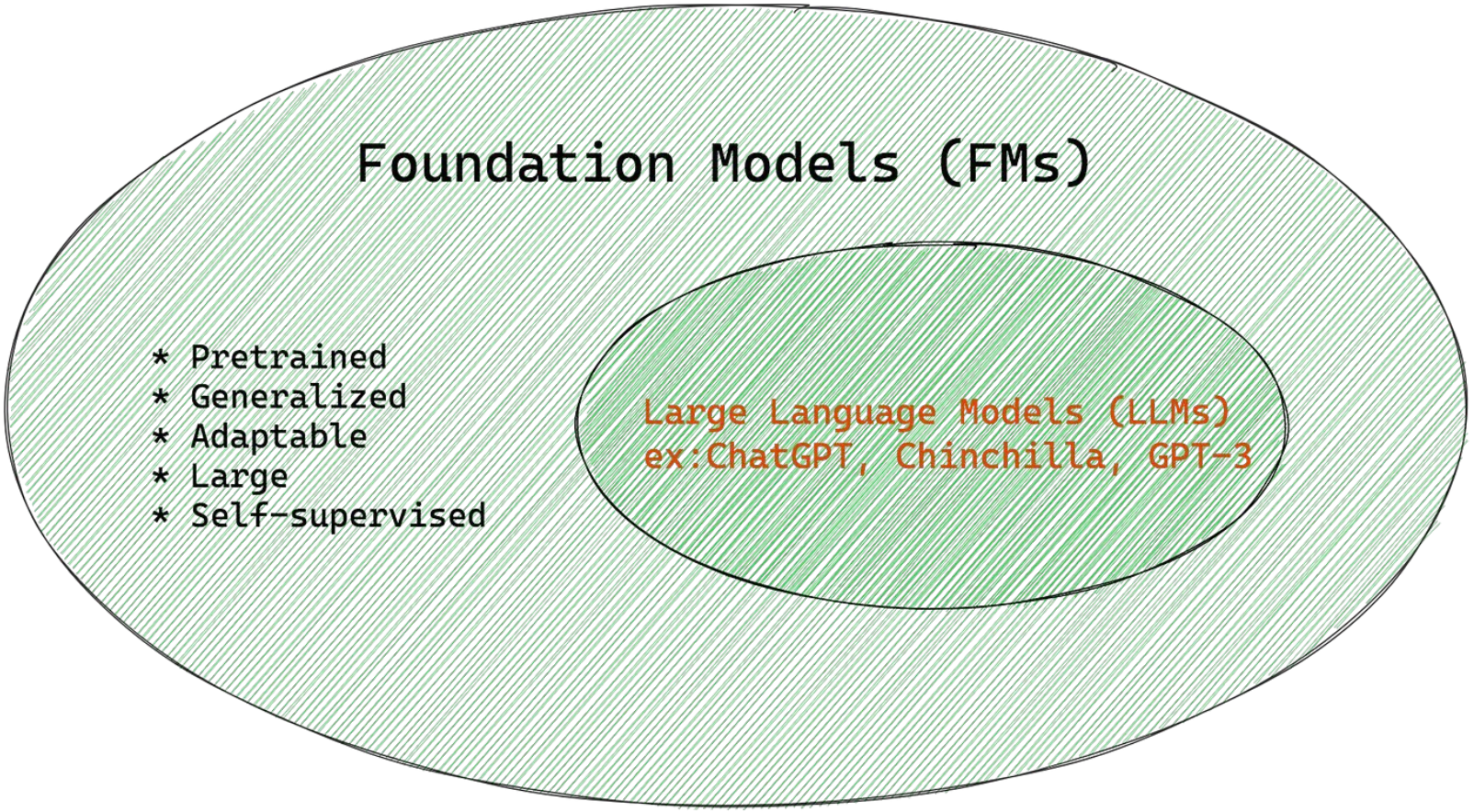
Gaming

Leonardo.AI, ROSEBUD.AI, Move AI

Foundational Models

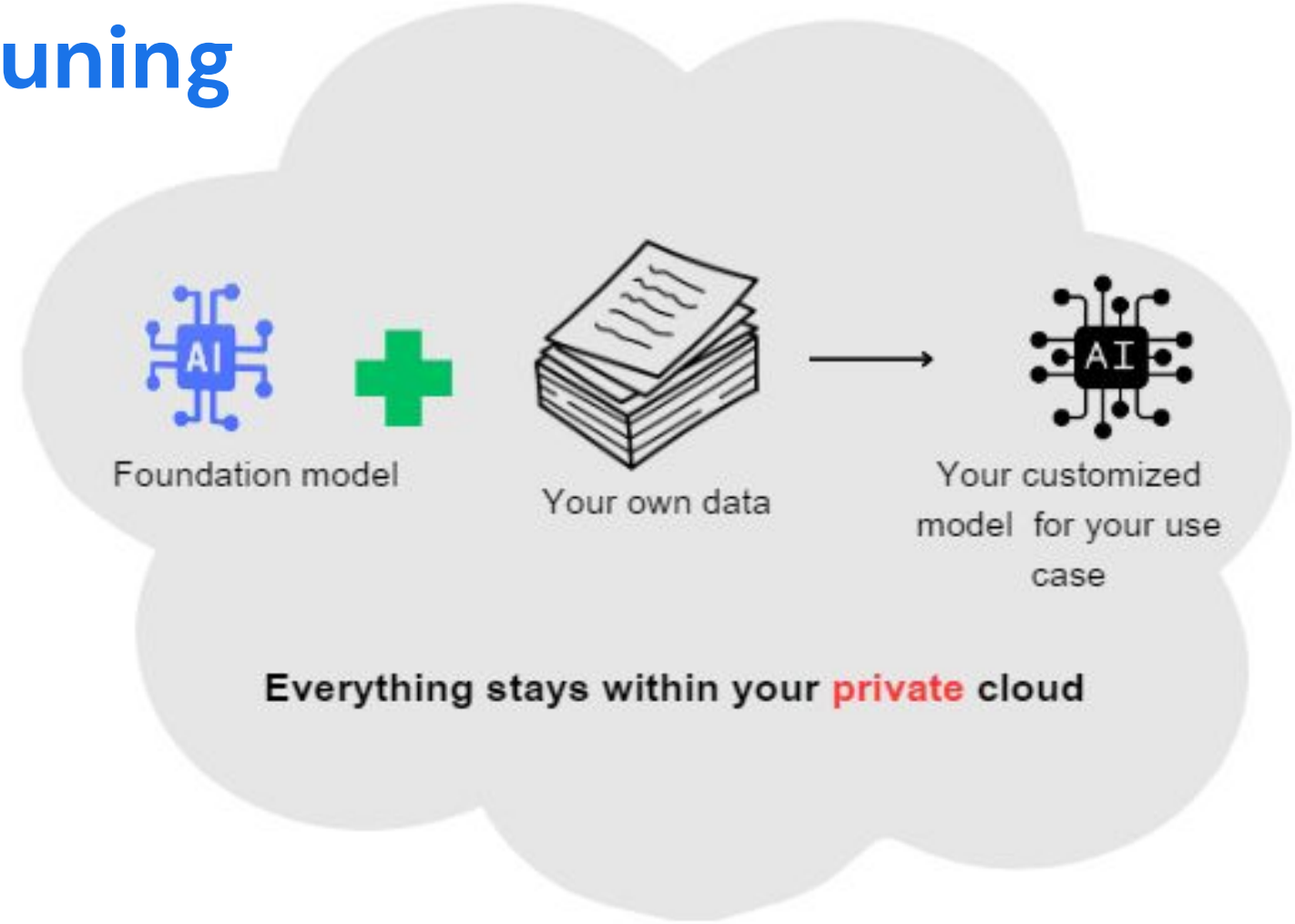
A foundation model (also called base model) is a large machine learning (ML) model trained on a vast quantity of data at scale (often by self-supervised learning or semi-supervised learning), such that it can be adapted to a wide range of downstream tasks.


Data
Text
Images
Speech
Structural Data
3D Signals
Training
Foundation model
Adaption
Tasks
Question answering
Sentient Analysis
Information Extraction
Image Captioning
Object Recognition
Instruction Following



FMs are models trained on broad data (using self-supervision at scale)
that can be adapted to a wide range of downstream tasks.
https://hai.stanford.edu/news/reflections-foundation-models

fine-tuning


AI
Foundation model
Your own data
AI
Your customized model for your use case
Everything stays within your private cloud

Foundational Models vs. LLMs

Large Language Models (LLMs) differ from Foundational Models in their scope of language understanding.

LLMs are specifically focused on language-related systems, while Foundational Models are attempting to stake out a broader function-based concept, which could accommodate new types of systems in the future

What is a language model?

A language model refers to a type of model specifically designed to generate human-like text or predict the probability of a sequence of words. Language models learn patterns and statistics from large amounts of text data, enabling them to generate sensible and contextually appropriate sentences.

In short…

The cat

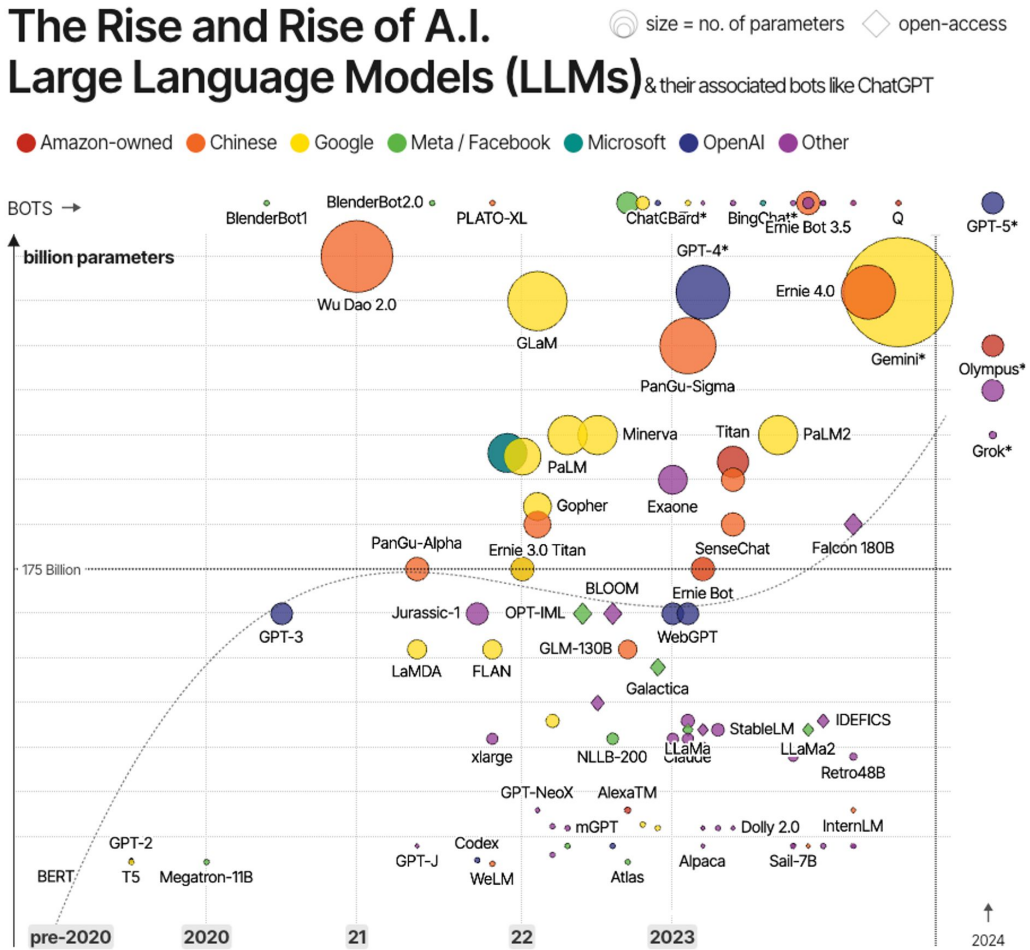
Enter, Large Language Models

Extensive Training Data: Large language models are trained on vast amounts of text data, often comprising billions or even trillions of words. This extensive training data helps the models learn patterns, grammar, context, and a wide range of language nuances, enabling them to generate coherent and contextually appropriate responses.

Complex Architectures: Large language models employ complex architectures, such as transformer networks, that contain numerous layers and millions or even billions of parameters. These architectures enable the models to capture intricate language structures, understand semantics, and generate high-quality text by leveraging the vast amount of training data they have been exposed to. The large number of parameters allows the models to learn fine-grained details and provide nuanced responses.

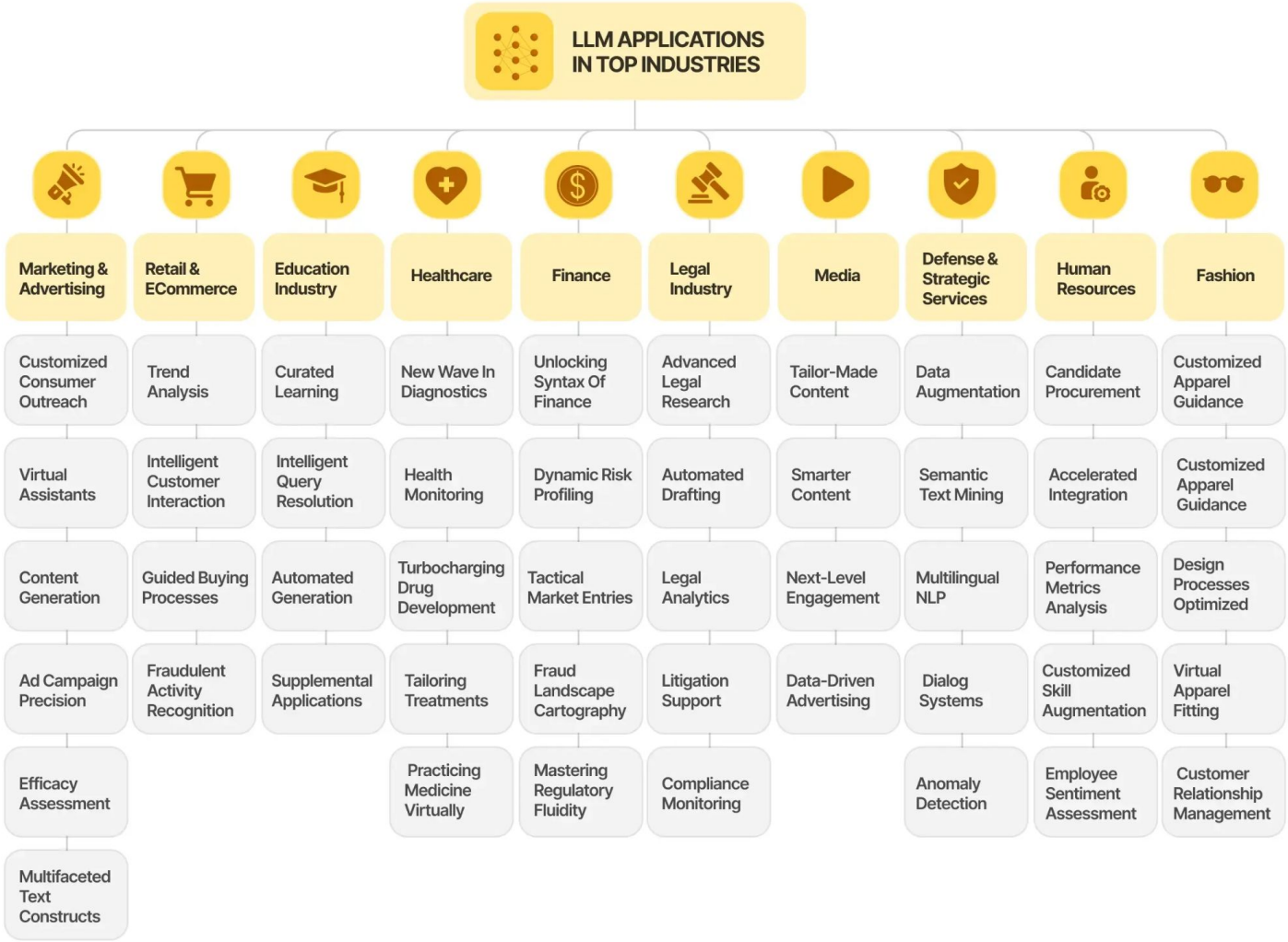
.

The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT

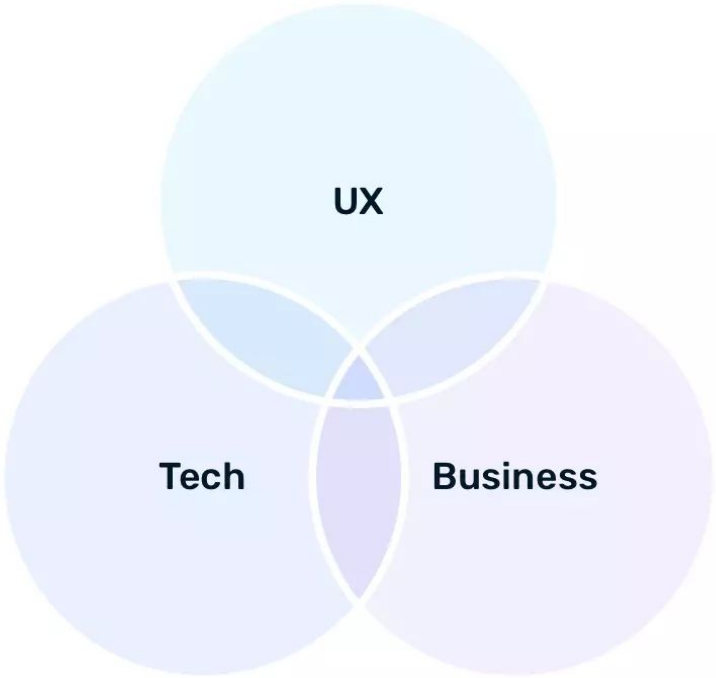


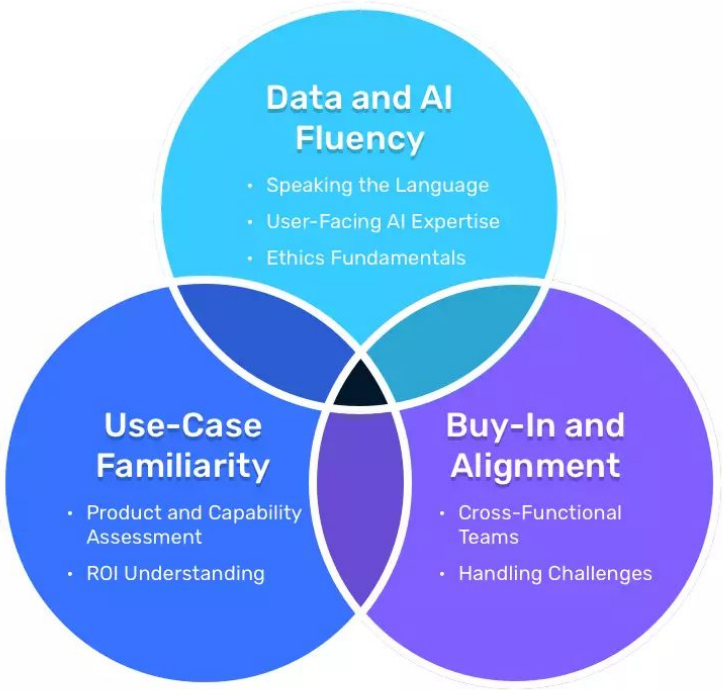
David McCandless, Tom Evans, Paul Barton
Information is Beautiful // UPDATED 6th Dec 23

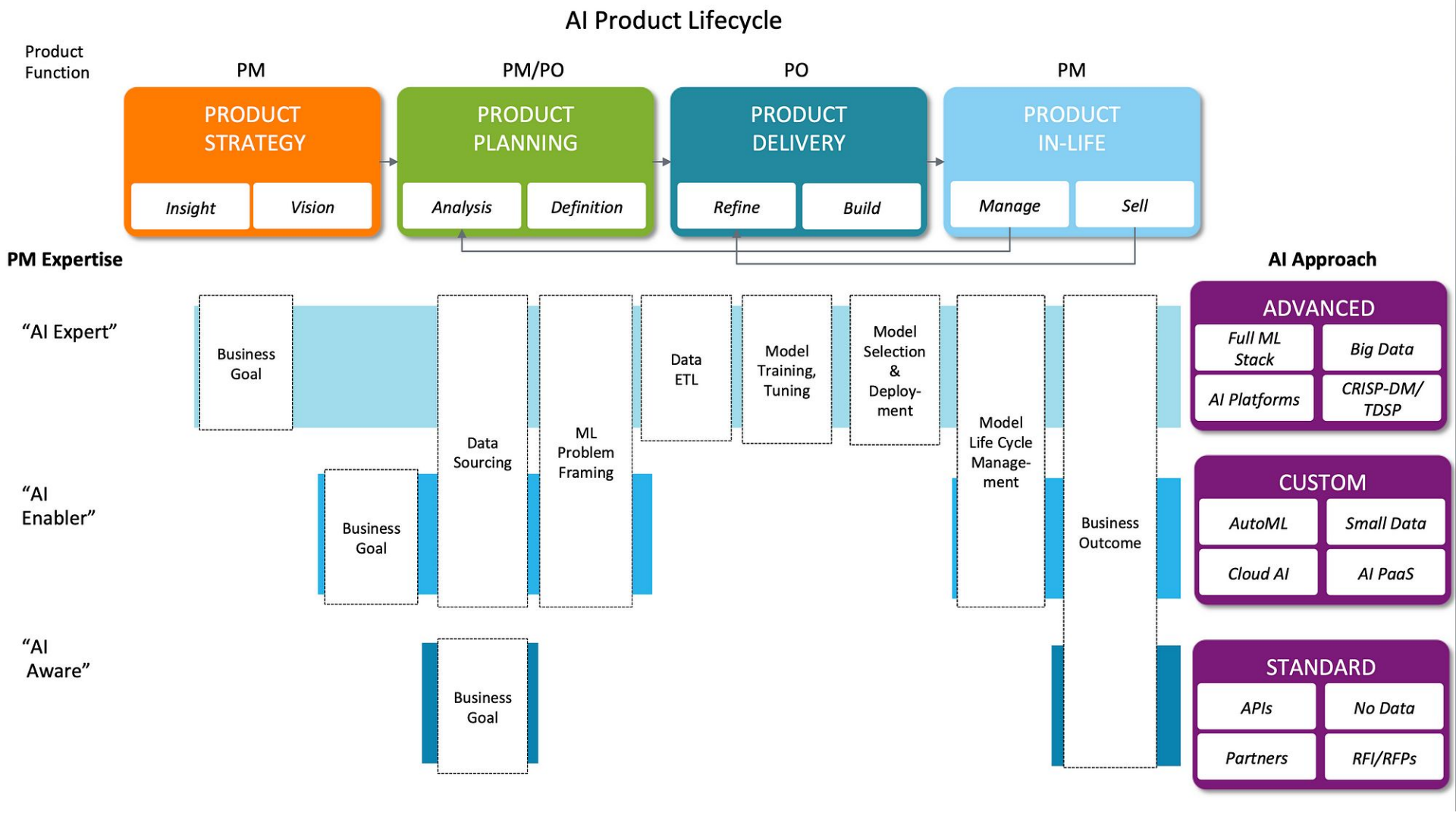
source: news reports, LifeArchitect.ai
* = parameters undisclosed // see the data


LLM APPLICATIONS IN TOP INDUSTRIES
Marketing & Advertising
Customized Consumer Outreach
Virtual Assistants
Content Generation
Ad Campaign Precision
Efficacy Assessment
Multifaceted Text Constructs
Retail & ECommerce
Trend Analysis
Intelligent Customer Interaction
Guided Buying Processes
Fraudulent Activity Recognition
Education Industry
Curated Learning
Intelligent Query Resolution
Automated Generation
Supplemental Applications
Healthcare
New Wave In Diagnostics
Health Monitoring
Turbocharging Drug Development
Tailoring Treatments
Practicing Medicine Virtually
Finance
Unlocking Syntax Of Finance
Dynamic Risk Profiling
Tactical Market Entries
Fraud Landscape Cartography
Mastering Regulatory Fluidity
Legal Industry
Advanced Legal Research
Automated Drafting
Legal Analytics
Litigation Support
Compliance Monitoring
Media
Tailor-Made Content
Smarter Content
Next-Level Engagement
Data-Driven Advertising
Defense & Strategic Services
Data Augmentation
Semantic Text Mining
Multilingual NLP
Dialog Systems
Anomaly Detection
Human Resources
Candidate Procurement
Accelerated Integration
Performance Metrics Analysis
Customized Skill Augmentation
Employee Sentiment Assessment
Fashion
Customized Apparel Guidance
Customized Apparel Guidance
Design Processes Optimized
Virtual Apparel Fitting
Customer Relationship Management

How can things change for you?


UX
Tech
Business

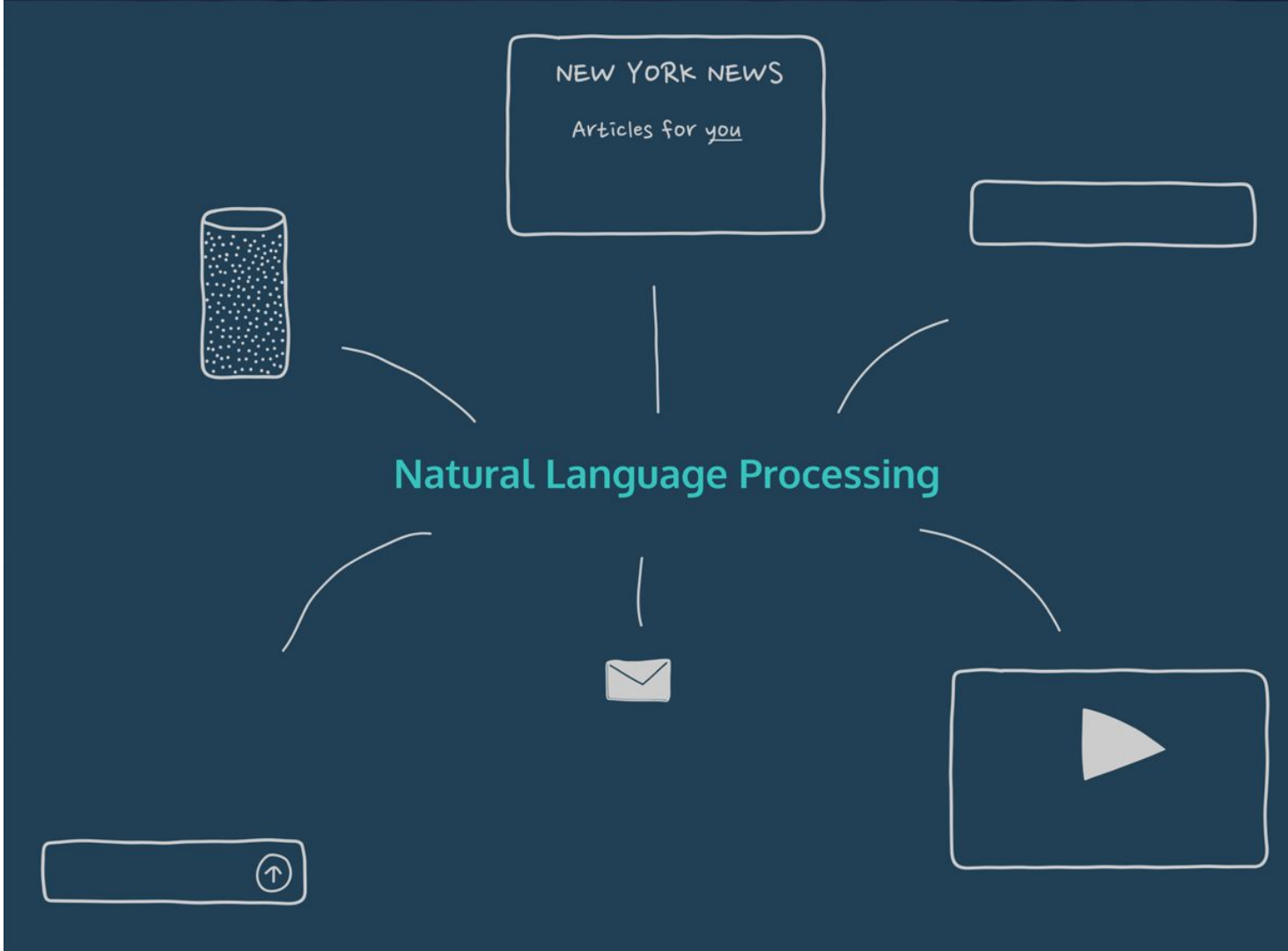

Data and AI Fluency
• Speaking the Language
• User-Facing AI Expertise
• Ethics Fundamentals
Use-Case Familiarity
• Product and Capability Assessment
• ROI Understanding
Buy-In and Alignment
• Cross-Functional Teams
• Handling Challenges


AI Product Lifecycle
Product Function
PM
PM/PO
PO
PM
PRODUCT STRATEGY
Insight
Vision
PRODUCT PLANNING
Analysis
Definition
PRODUCT DELIVERY
Refine
Build
PRODUCT IN-LIFE
Manage
Sell
PM Expertise
AI Approach
"AI Expert"
"AI Enabler"
"AI Aware"
Business Goal
Data Sourcing
ML Problem Framing
Data ETL
Model Training, Tuning
Model Selection & Deploy-ment
Model Life Cycle Manage-ment
Business Outcome
ADVANCED
Full ML Stack
Big Data
AI Platforms
CRISP-DM/ TDSP
CUSTOM
AutoML
Small Data
Cloud AI
AI PaaS
STANDARD
APIs
No Data
Partners
RFI/RFPs

03

Let’s go back in time

We live in a world of NLP



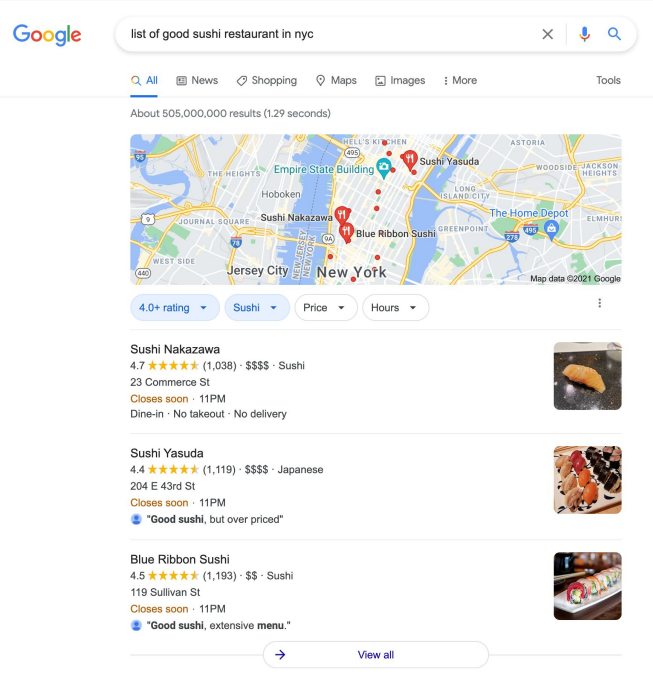
What is NLP anyways?

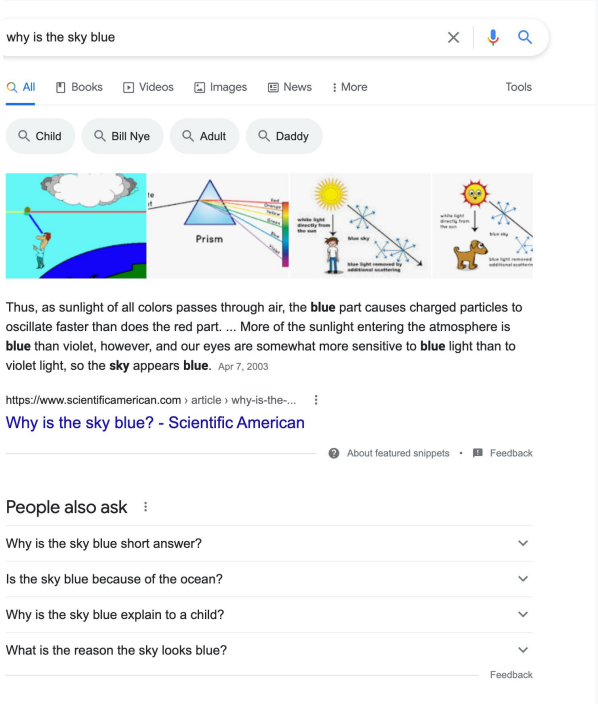
Natural Language Processing (NLP) is defined as the branch of Artificial Intelligence that provides computers with the capability of understanding text and spoken words, in the same way a human being can.

It incorporates machine learning models, statistics, and deep learning models into computational linguistics i.e. rule-based modeling of human language to allow computers to understand text, spoken words and understands human language, intent, and sentiment.

Applications – 1

- Information retrieval
- Information extraction
- Question answering

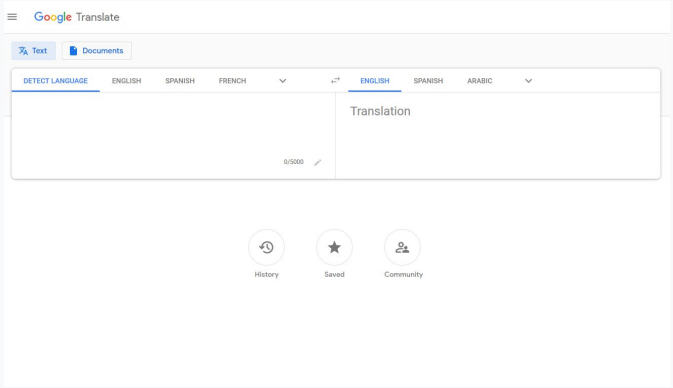


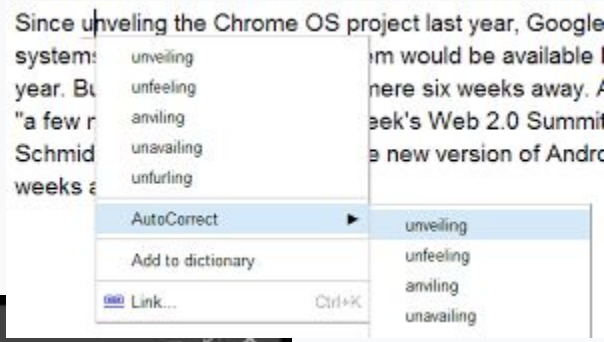


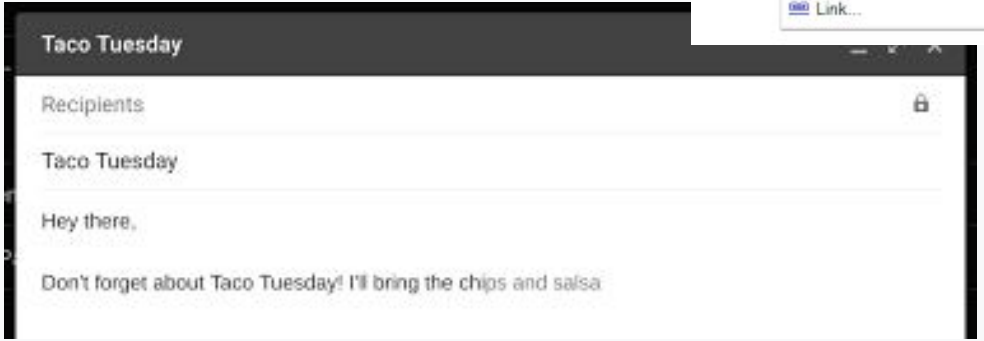
Applications -2

- Machine Translation
- Summarization
- Auto Completion
- Spell Correction

Many More...







NLP Ambiguities

There are different types of ambiguities present in natural language:

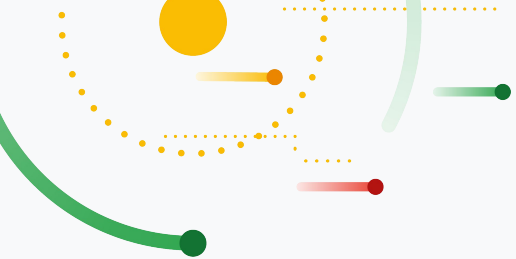
1. Lexical Ambiguity: It is defined as the ambiguity associated with the meaning of a single word. A single word can have different meanings. Also, a single word can be a noun, adjective, or verb. For example, The word “bank” can have different meanings. It can be a financial bank or a riverbank. Similarly, the word “clean” can be a noun, adverb, adjective, or verb.

NLP Ambiguities

2. Syntactic Ambiguity: It is defined as the ambiguity associated with the way the words are parsed. For example, The sentence “Visiting relatives can be boring.” This sentence can have two different meanings. One is that visiting a relative’s house can be boring. The second is that visiting relatives at your place can be boring.

NLP Ambiguities

3. Semantic Ambiguity: It is defined as ambiguity when the meaning of the words themselves can be ambiguous. For example, The sentence “Mary knows a little french.” In this sentence the word “little french” is ambiguous. As we don’t know whether it is about the language french or a person.



Common NLP tasks

Common NLP tasks

NLP systems

- Natural language understanding
- Natural language generation and summarization
- Natural language translation

Natural language understanding

- Extract information (e.g. about entities or events) from text
- Translate raw text into a meaning representation
- Reason about information given in text
- Execute NL instructions

Natural language generation and summarization

- Translate database entries or meaning representations to raw natural language text
- Produce (appropriate) utterances/responses in a dialog
- Summarize (newspaper or scientific) articles, describe images

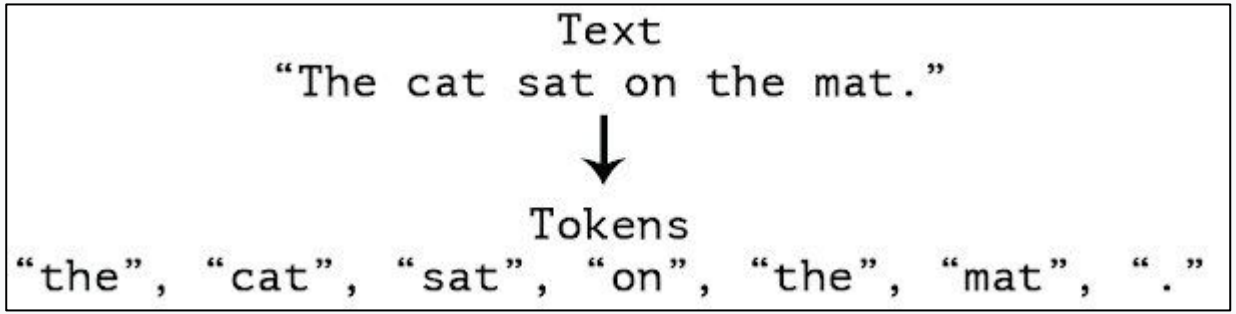
Natural language translation

- Translate one natural language to another

How does a machine understand what we are saying?

- **Tokenization**
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

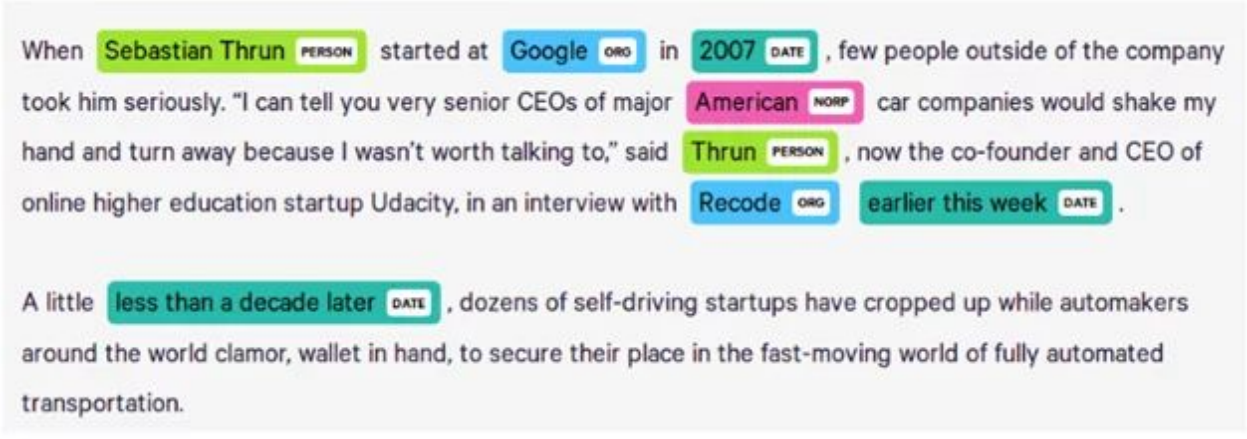
- Tokenization is the process of breaking down a text into individual units called tokens.
- Tokens are typically words, but can also be phrases or even individual characters, depending on the application.
- Tokenization is a crucial step in natural language processing tasks such as machine translation, sentiment analysis, and named entity recognition.



Common NLP tasks

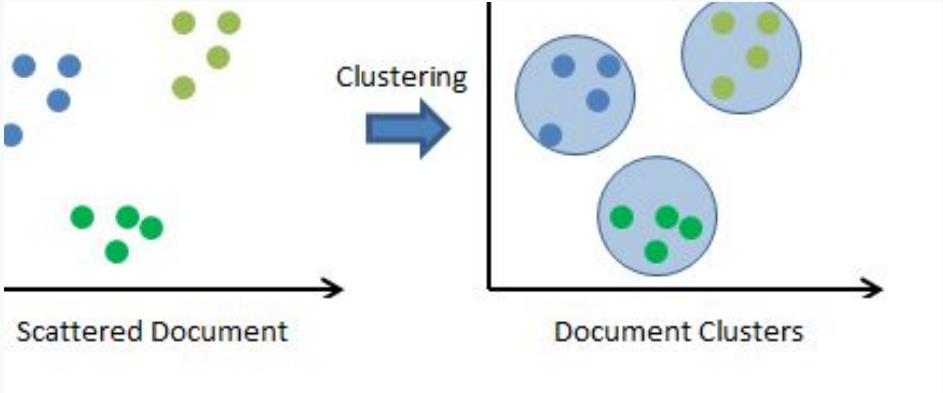
- Tokenization
- **Named Entity Recognition (NER)**
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

Named entity recognition (NER) is the process of identifying and categorizing named entities in a text, such as people, organizations, locations, and dates.



Common NLP tasks

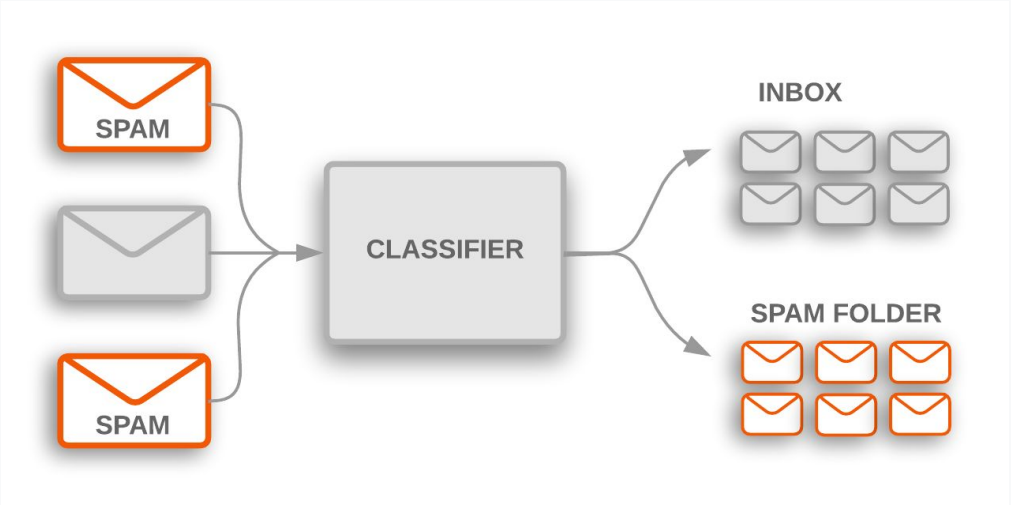
- Tokenization
- Named Entity Recognition (NER)
- **Text representation**
- Text classification
- Natural language generation
- Multimodal NLP

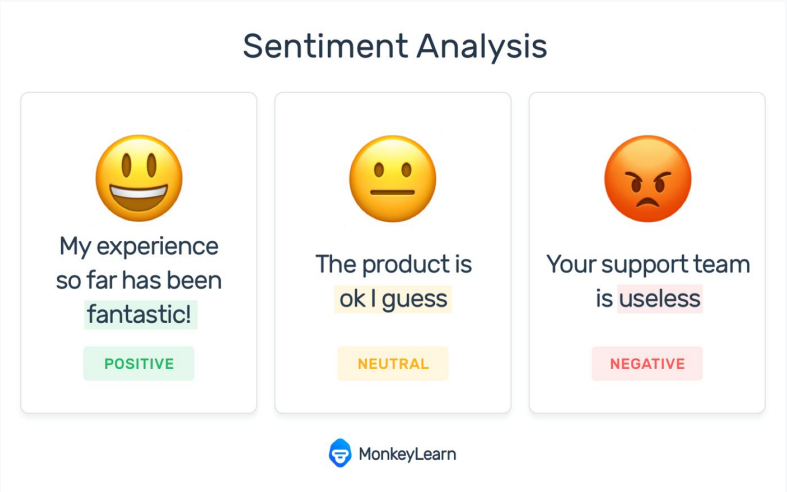


- Text representation is the process of converting unstructured text data into a structured format that can be used for natural language processing tasks.
- It involves selecting a suitable representation scheme, such as bag-of-words, word embeddings, or topic models, to capture the key features and characteristics of the text data in a numerical form that can be processed by machine learning algorithms.

Common NLP tasks

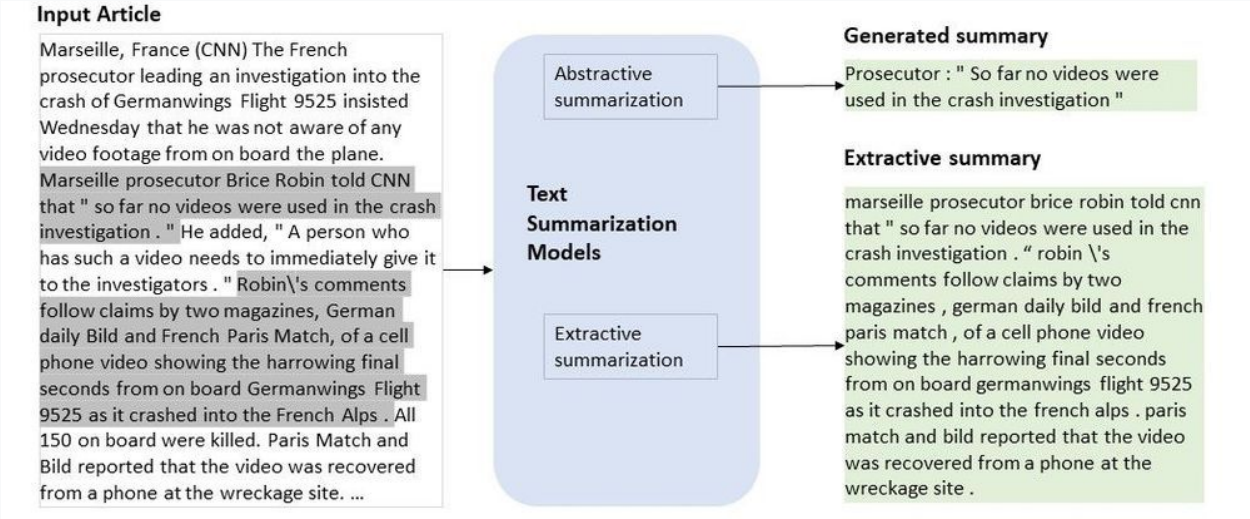
- Tokenization
- Named Entity Recognition (NER)
- Text representation
- **Text classification**
- Natural language generation
- Multimodal NLP

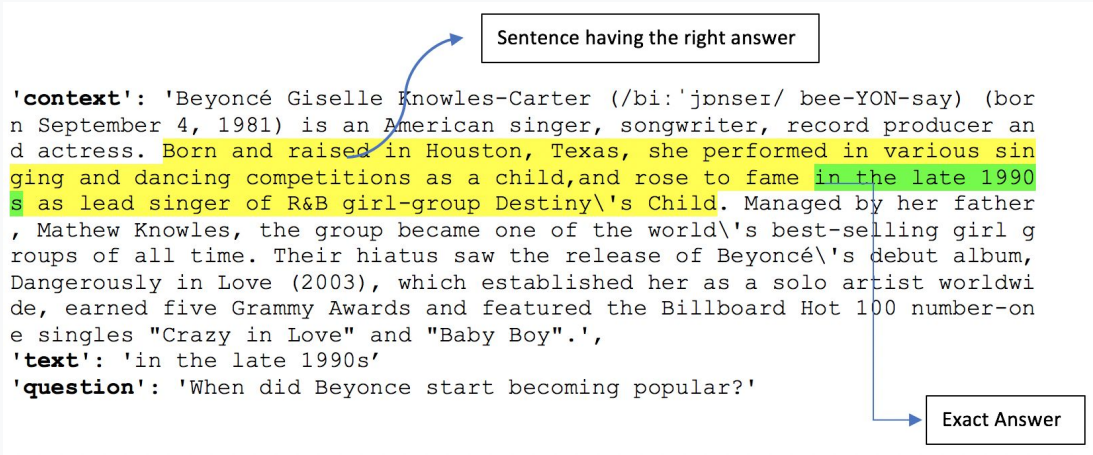




Common NLP tasks

- Tokenization
- Named Entity Recognition (NER)
- Text representation
- Text classification
- **Natural language generation**
- Multimodal NLP

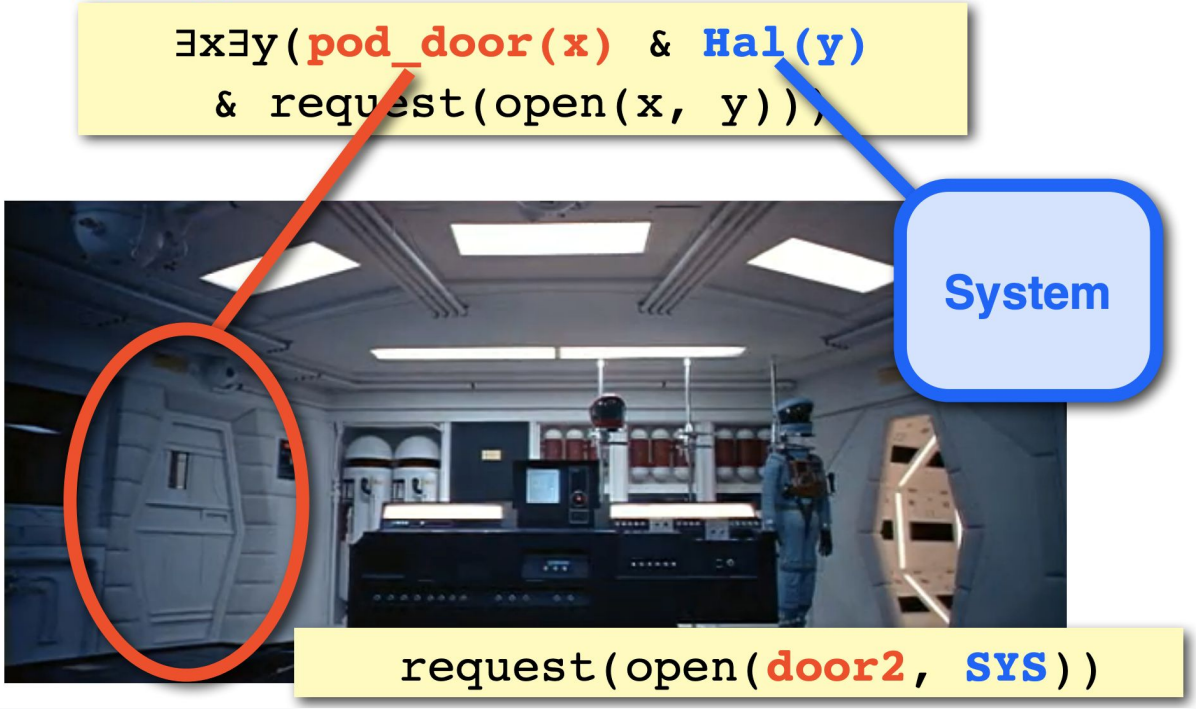




Common NLP tasks

- Tokenization
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- **Multimodal NLP**

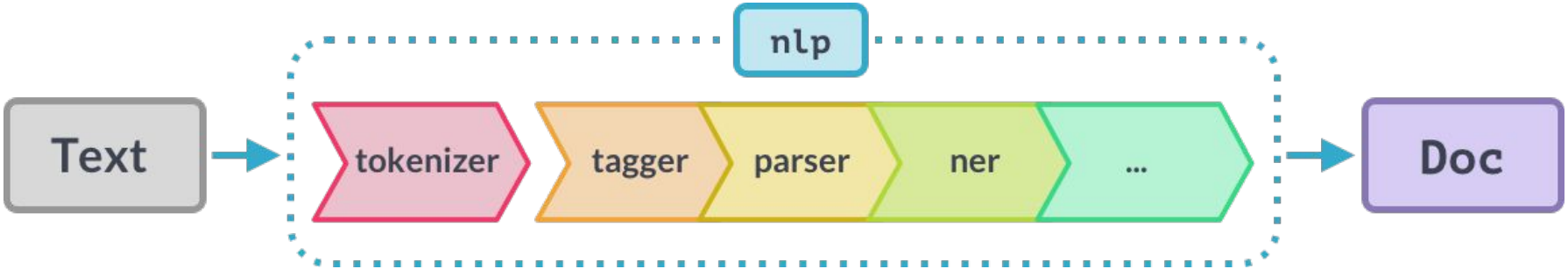
Multimodal NLP: mapping from language to the world



spaCy Package

spaCy is an open-source library used for natural language processing in python. It is extremely popular for processing a large amount of unstructured data generated at a vast scale in the industry and generate useful and meaningful insights from the data.

spaCy NLP Pipeline



Let’s code

Colab

Appendix