GENERATIVE AI:
TOOLS & CHALLENGES

27 Juin 2024, Toulouse
Journées Ouvertes en Biologie, Informatique, et Mathématiques

Vincent Guigue
vincent.guigue@agroparistech.fr
https://vguigue.github.io

MIA
PARIS-SACLAY
EKINOCS

AGROPARISTECH
Institut des Sciences et Industries du Vivant et de l'Environnement

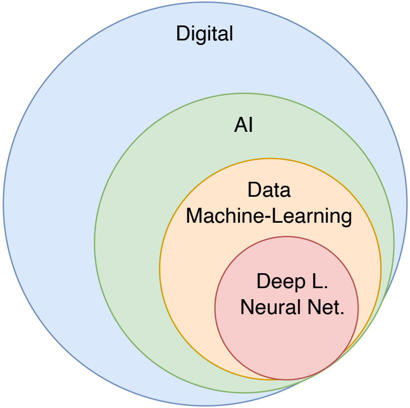
université PARIS-SACLAY

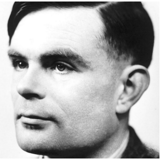
From AI to machine-learning

Digital & Artificial Intelligence

- Two related but distinct concepts
- AI: Different Definitions

1956	Any algorithm / program
1960-2012	Expert systems and logical reasoning
2012-	Data & neural networks



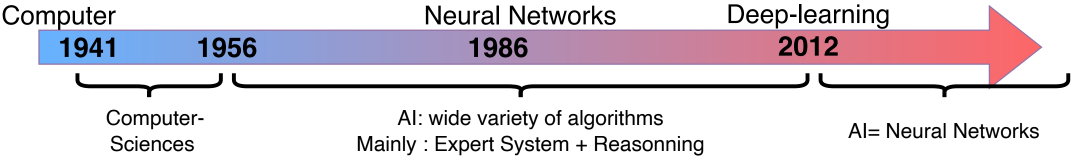


A. Turing

Marvin Minsky

Y. Lecun

G. Hinton



Artificial Intelligence & Machine Learning

Input (**X**)		Output (**Y**)	Application
email	→	spam? (0/1)	spam filtering
audio	→	text transcript	speech recognition
English	→	Chinese	machine translation
ad, user info	→	click? (0/1)	online advertising
image, radar info	→	position of other cars	self-driving car
image of phone	→	defect? (0/1)	visual inspection

AI: computer programs that engage in tasks which are, for now, performed more satisfactorily by human beings because they require high-level mental processes.

Marvin Lee Minsky, 1956

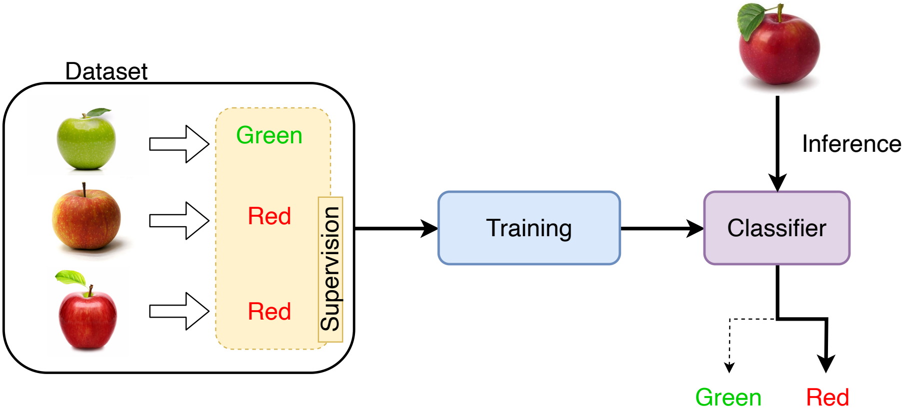
N-AI (Narrow Artificial Intelligence), dedicated to a single task

$\neq$ **G-AI (General AI)**, which replaces humans in complex systems.

Andrew Ng, 2015

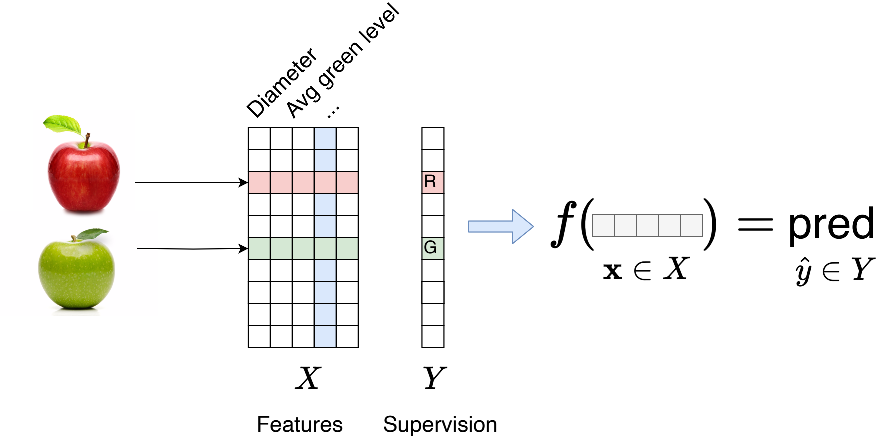
Machine Learning definition

1. Collecting labeled **dataset**
2. Training **classifier**
3. Exploiting the model



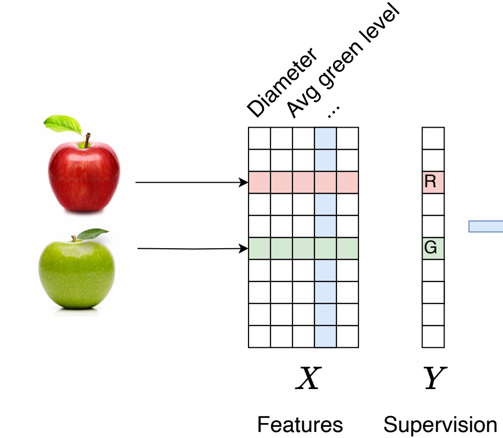
Machine Learning definition

1. Collecting labeled **dataset**
2. Training **classifier**
3. Exploiting the model

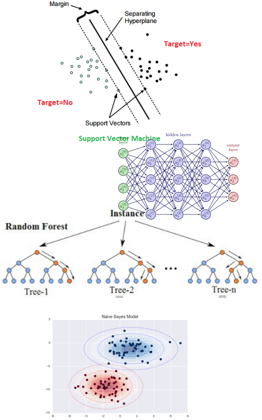


Machine Learning definition

1. Collecting labeled **dataset**
2. Training **classifier**
3. Exploiting the model

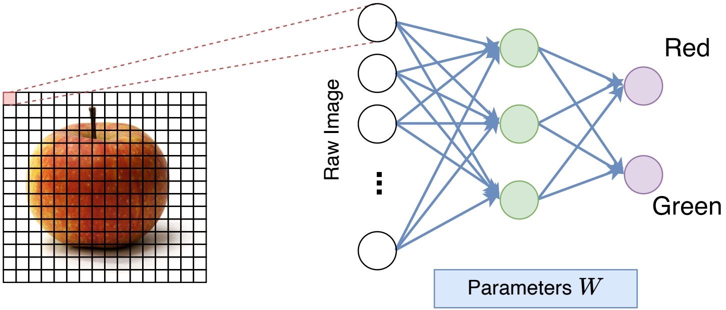


$$f(\square\square\square\square\square) = \text{pred}$$
$$\mathbf{x} \in X \qquad \hat{y} \in Y$$



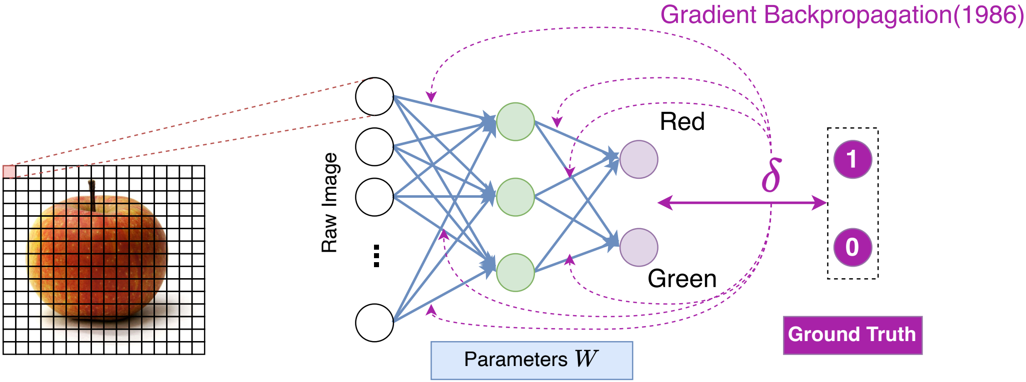
Neural Networks: tackling raw/complex data

1. Complex modular architecture
2. Random initilization



Neural Networks: tackling raw/complex data

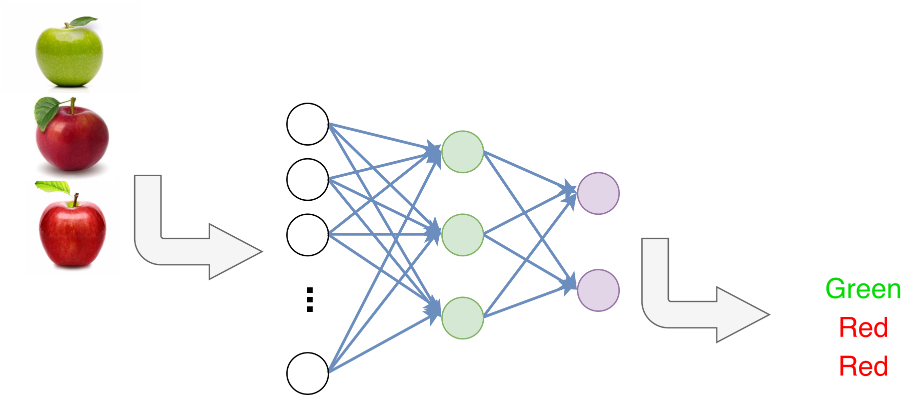
1. Complex modular architecture
2. Random initilization
3. (Slow) Training by backpropagation



Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Neural Networks: tackling raw/complex data

1. Complex modular architecture
2. Random initilization
3. (Slow) Training by backpropagation
4. Faster inference



Deep learning & Representation Learning

[Application to textual data]

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

Rule-based Systems:

* ⟶ {like, love, appreciate} ⟶ * ⟶ #product

* ⟶ {didn't, not, doesn't, don't} ⟶ {like, love, appreciate} ⟶ * ⟶ #product

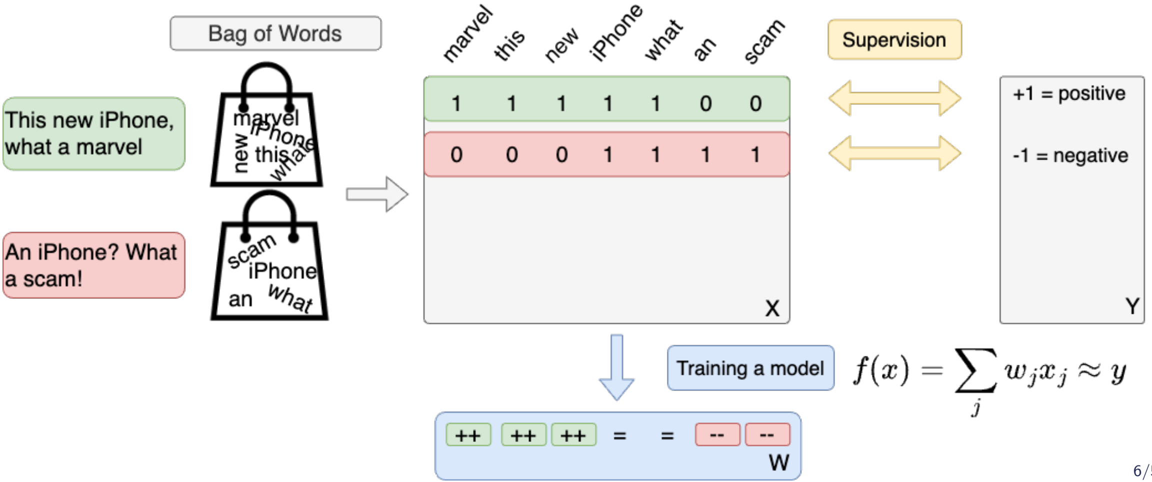
* ⟶ {hate, loathe, detest} ⟶ * ⟶ #product

- Requires expert knowledge
- Rule extraction ⇔ very clean data
- Very high precision
- Low recall
- Interpretable system

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Machine Learning [1990-2015]



$$f(x) = \sum_j w_j x_j \approx y$$

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

- Requires expert knowledge
- Rule extraction ⇔ very clean data

+ Interpretable system
+ Very high precision
− Low recall

Machine Learning [1990-2015]

- Little expert knowledge needed
- Statistical extraction ⇔ robust to noisy data

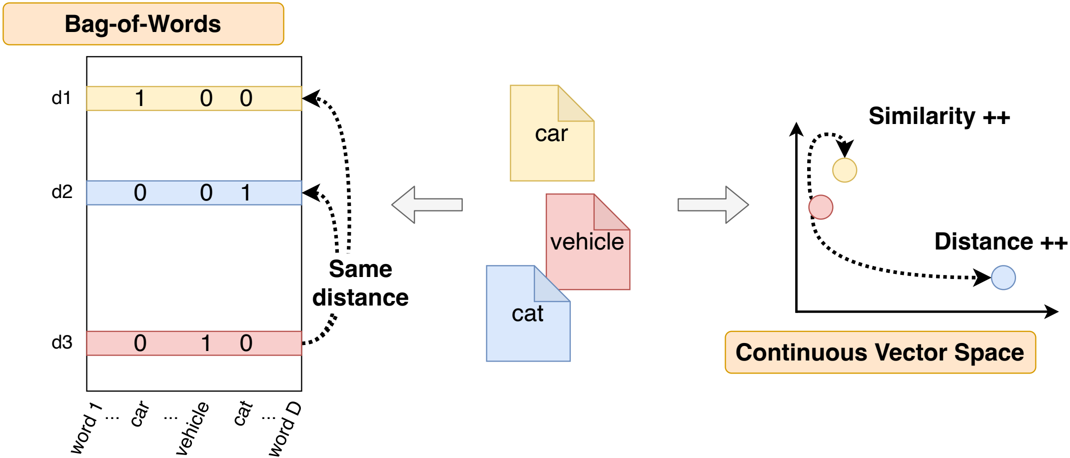
≈ Less interpretable system
− Lower precision
+ Better recall

Precision = criterion for acceptance by industry

→ Link to metrics

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

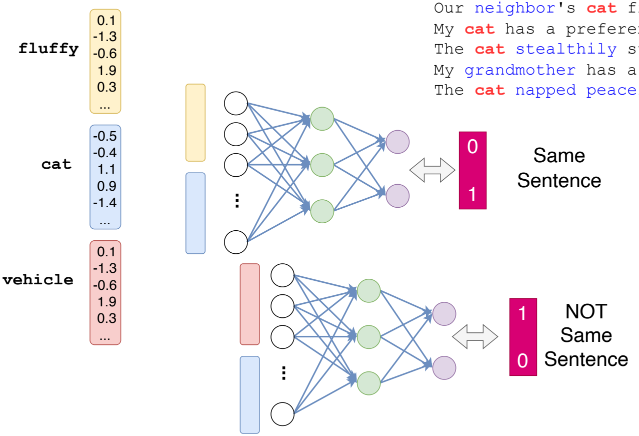


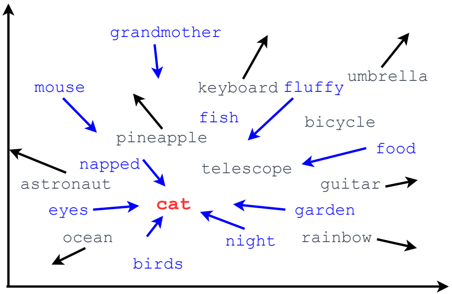
LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

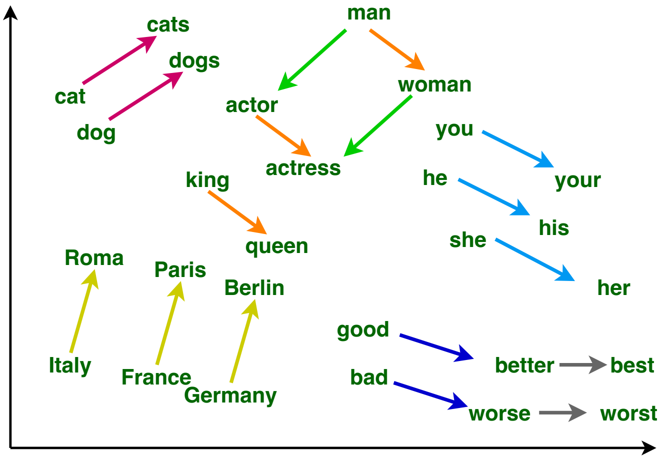
The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.
The cat napped peacefully in the warm sunlight.





Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]



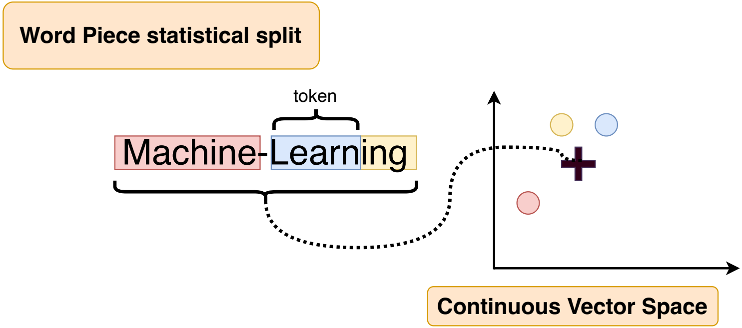
- Semantic Space:
  similar meaning
  ⇔
  close position
- Structured Space:
  grammatical regularities,
  basic knowledge, ...

Distributed representations of words and phrases and their compositionality, Mikolov et al. NeurIPS 2013

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

From Words to Tokens

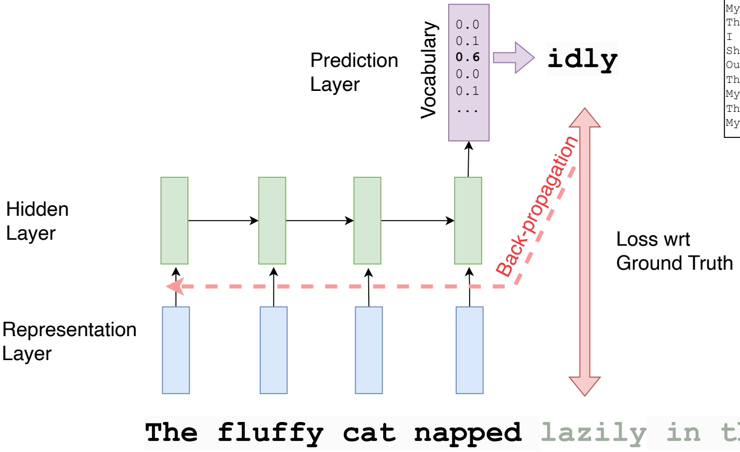


- Representation of unknown words
- Adaptation to technical domains
- Resistance to spelling errors

Enriching word vectors with subword information. Bojanowski et al. TACL 2017.

Aggregating word representations: towards generative AI

- Generation & Representation
- New way of learning word positions

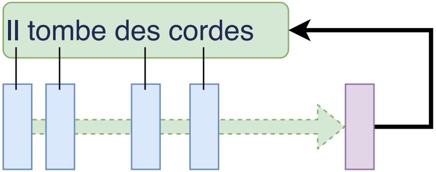



```
The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
The playful cat swatted at the dangling string with its paw.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.
```

Corpus

Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

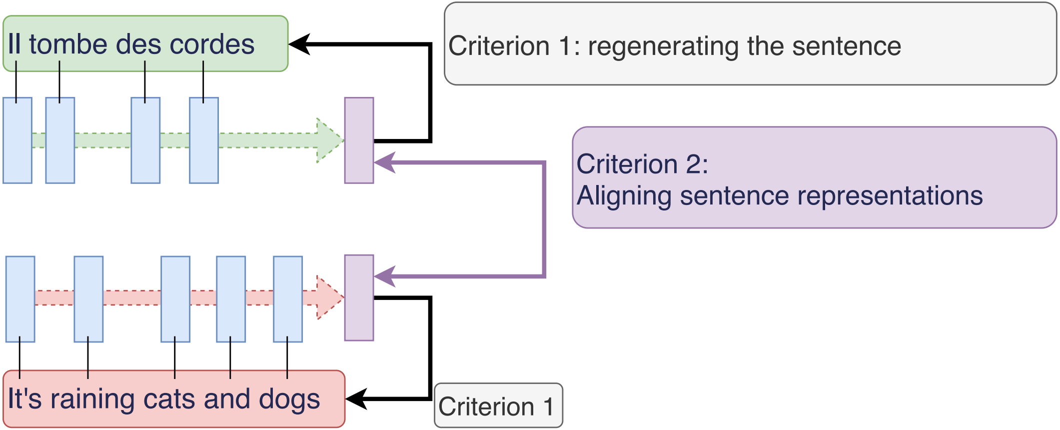
Use-Case: Machine Translation



Criterion 1: regenerating the sentence

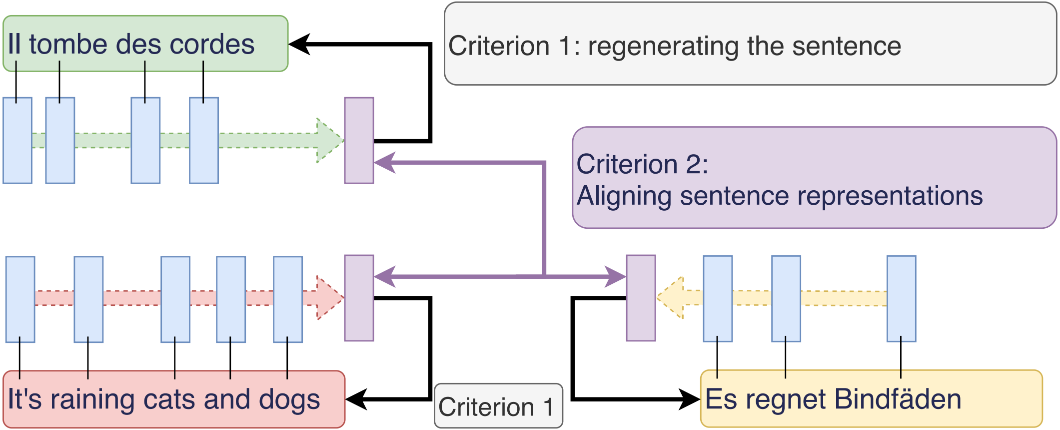
Beyond word-for-word translation, multilingual representation of sentences

Use-Case: Machine Translation



Beyond word-for-word translation, multilingual representation of sentences

Use-Case: Machine Translation

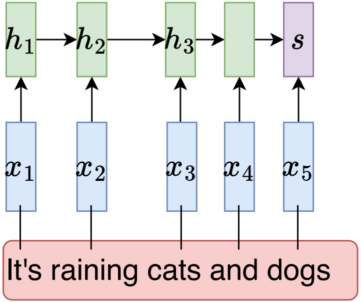


Beyond word-for-word translation, multilingual representation of sentences

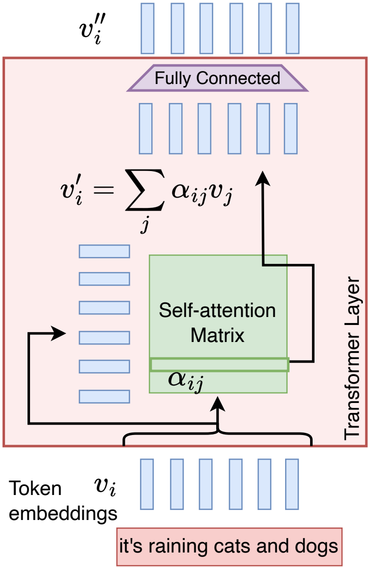
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

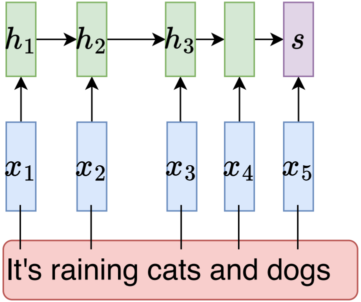


Attention is all you need, Vaswani et al. NeurIPS 2017

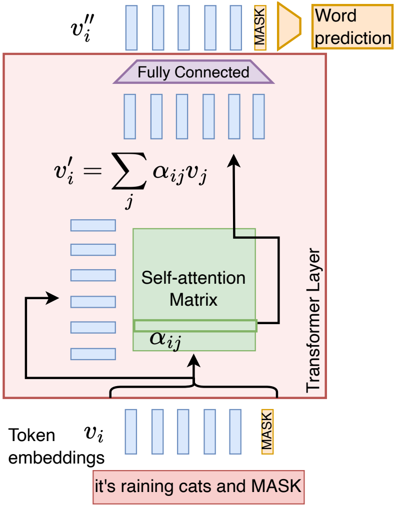
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

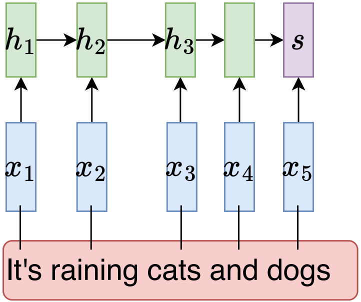


Attention is all you need, Vaswani et al. NeurIPS 2017

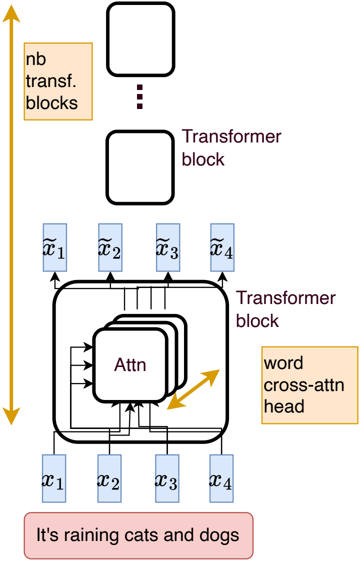
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:

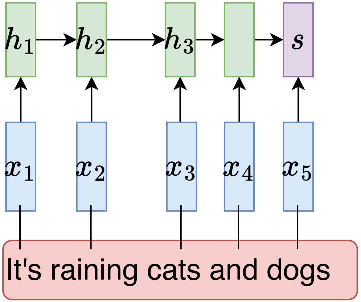


Attention is all you need. Vaswani et al. NeurIPS 2017

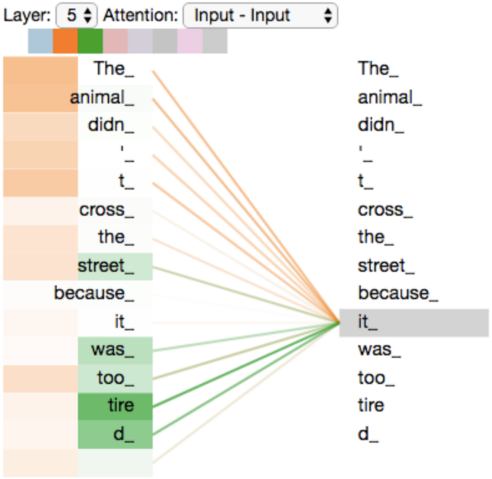
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



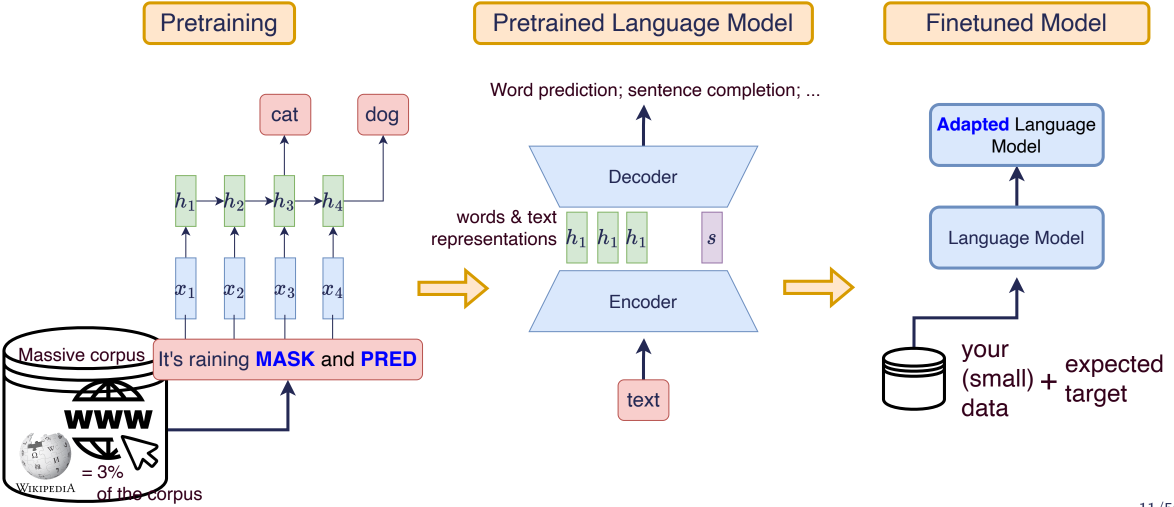
Transformer:



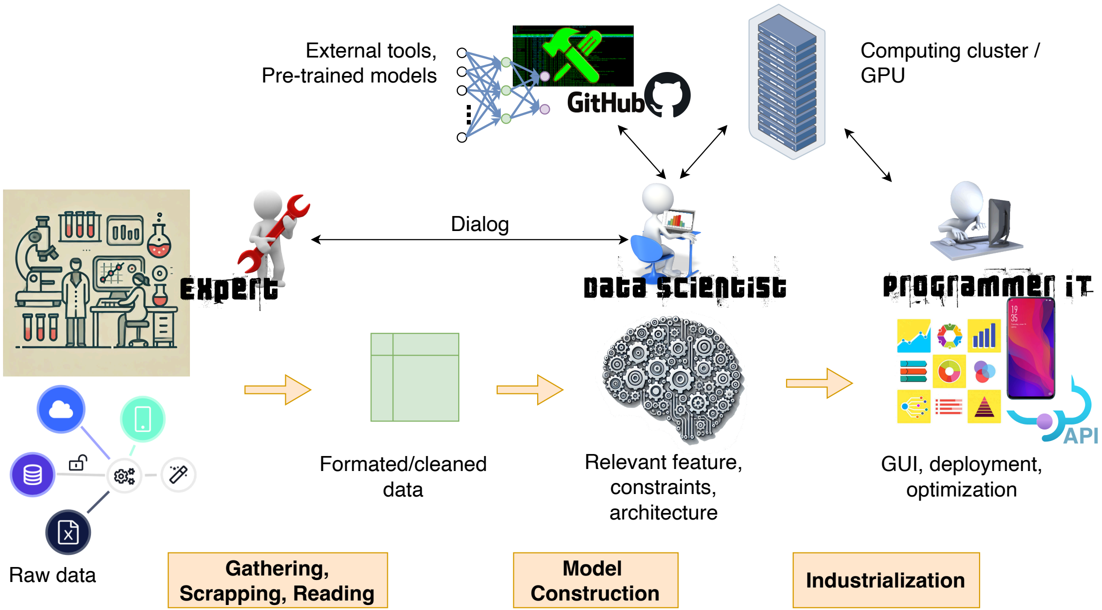
Attention is all you need, Vaswani et al. NeurIPS 2017

A new developpement paradigm since 2015

- Huge dataset + huge archi. $\Rightarrow$ unreasonable training cost
- Pre-trained architecture + 0-shot / finetuning



Different steps / different jobs



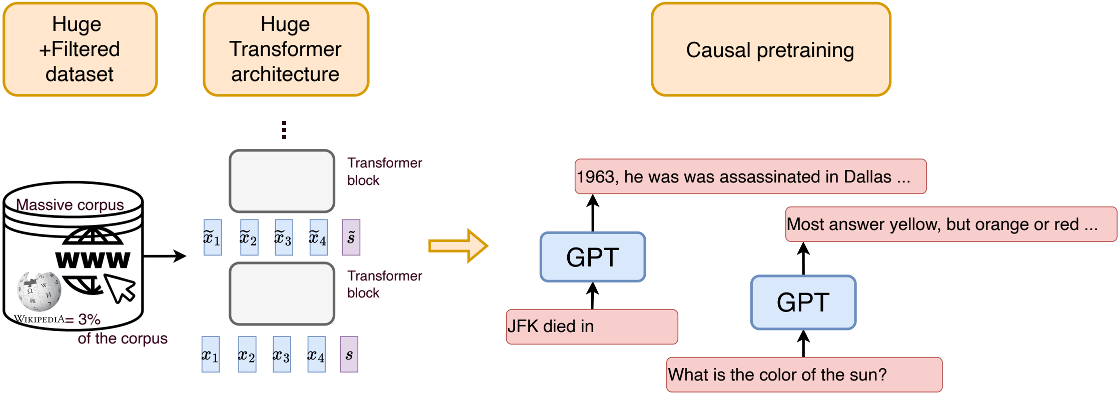
chatGPT

November 30, 2022

1 million users in 5 days
100 million by the end of January 2023
1.16 billion by March 2023

The Ingredients of chatGPT

0. Transformer + massive data (GPT)



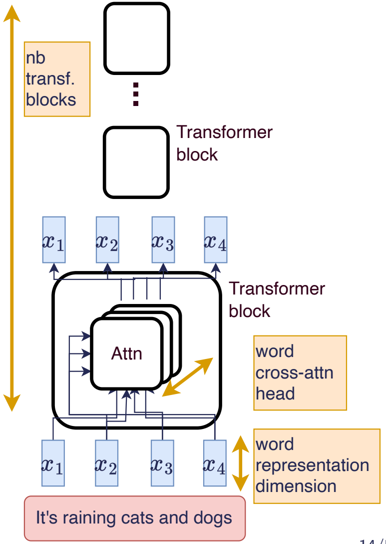
- Grammatical skills: singular/plural agreement, tense concordance
- Knowledges

Language Models are Few-Shot Learners, Brown et al. 2020

The Ingredients of chatGPT

1. More is better! (GPT)

- $+$ more input words [500 $\Rightarrow$ 2k, 32k, 100k]
- $+$ more dimensions in the word space [500-2k $\Rightarrow$ 12k]
- $+$ more attention heads [12 $\Rightarrow$ 96]
- $+$ more blocks/layers [5-12 $\Rightarrow$ 96]

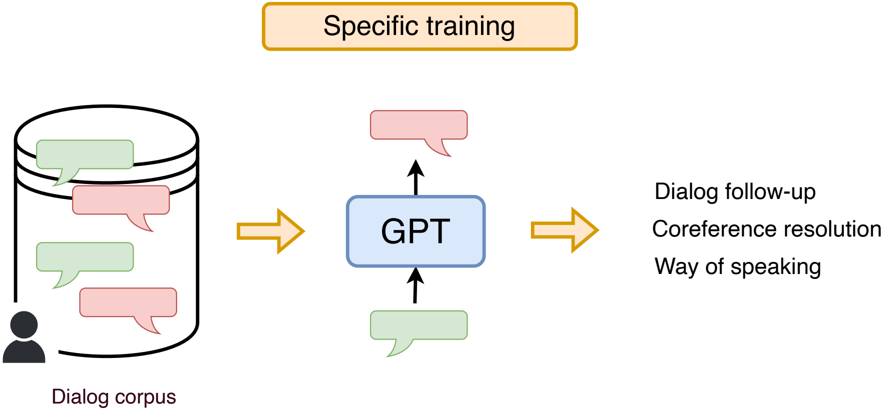


175 Billion parameters... What does it mean?

- $1.75 \cdot 10^{11} \Rightarrow$ 300 GB + 100 GB (data storage for inference) $\approx$ 400GB
- NVidia A100 GPU = 80GB of memory (=20k€)
- Cost for (1) training: 4.6 Million €

The Ingredients of chatGPT

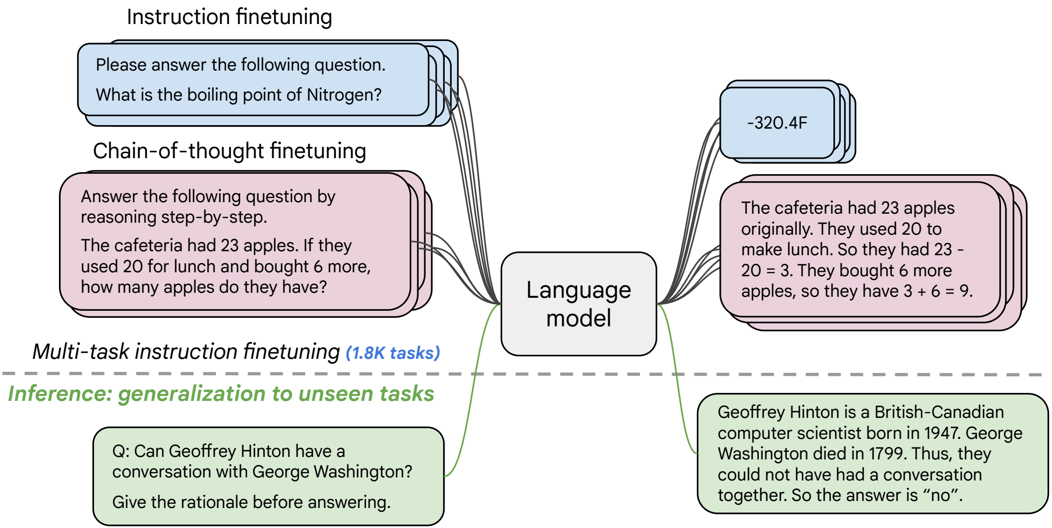
2. Dialogue Tracking



- **Very clean** data Data generated/validated/ranked by humans

The Ingredients of chatGPT

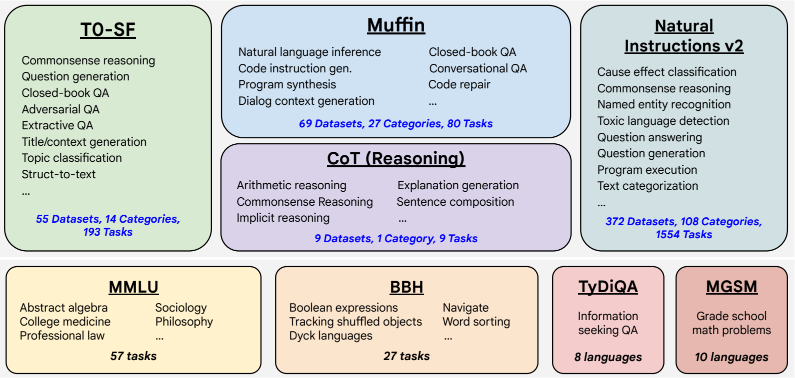
3. Fine-tuning on different (±) complex reasoning tasks

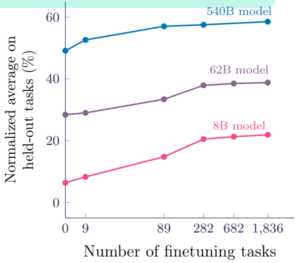


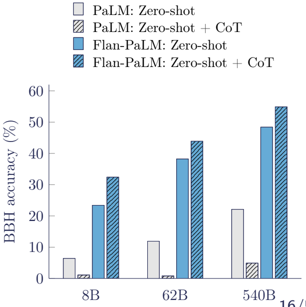
Scaling Instruction-Finetuned Language Models, Chung et al., JMLR 2024

The Ingredients of chatGPT

3. Fine-tuning on different (±) complex reasoning tasks



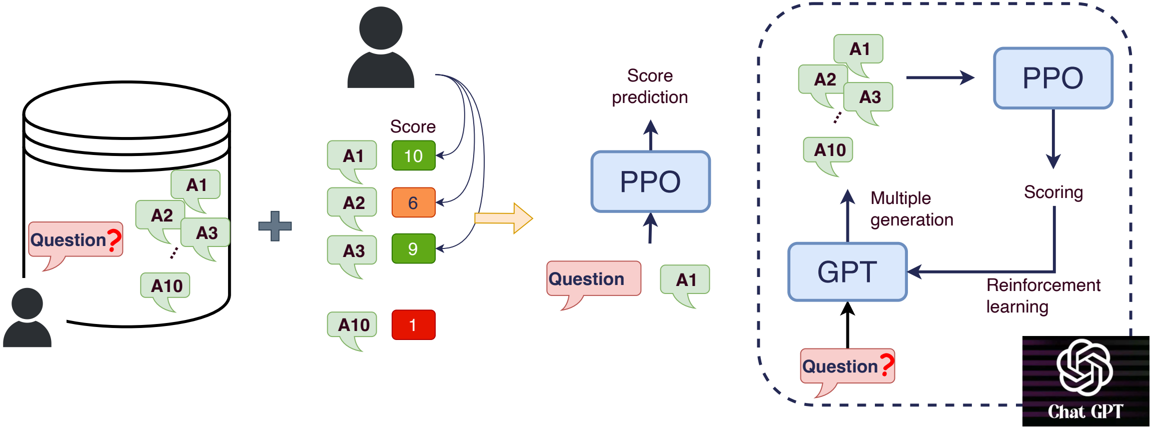




Scaling Instruction-Finetuned Language Models, Chung et al., JMLR 2024

The Ingredients of chatGPT

4. Instructions + answer ranking



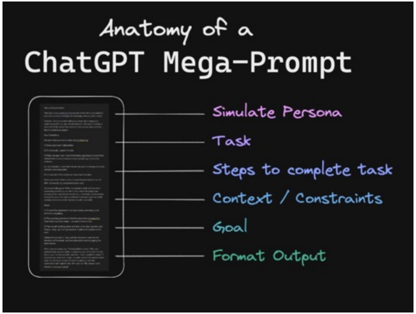
- Database created by humans
- Response improvement
- ... Also a way to avoid critical topics = censorship

Training language models to follow instructions with human feedback, Ouyang et al., 2022

Usage of chatGPT & Prompting

- Asking chatGPT = skill to acquire ⇒ *prompting*
  - Asking a question well: ... *in detail*, ... *step by step*
  - Specify number of elements e.g. : *3 qualities for ...*
  - Provide context : *cell* for a biologist / legal assistant
- Don't stop at the first question
  - Detail specific points
  - Redirect the research
  - Dialogue
- Rephrasing
  - Explain like I'm 5, like a scientific article, bro style, ...
  - Summarize, extend
  - Add mistakes (!)

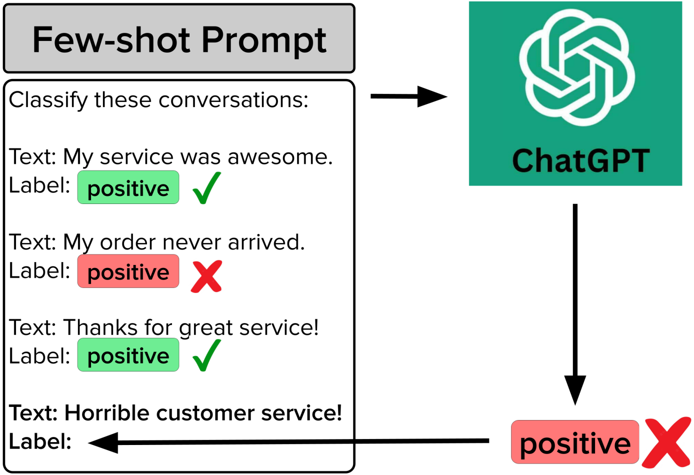
⇒ Need for **practice** [1 to 2 hours], discuss with colleagues



https://chatgptprompts.guru/what-makes-a-good-chatgpt-prompt/

Towards *few-shot learning*

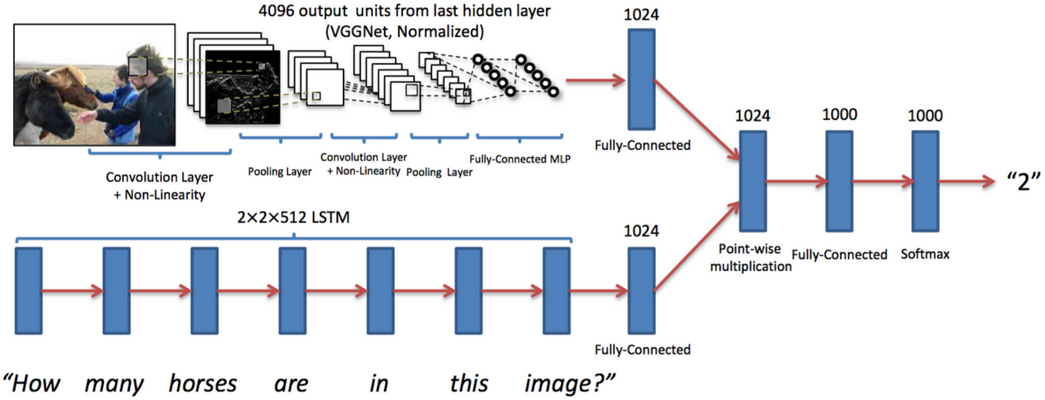
- Learning without modifying the model = examples in the prompt



GPT4 & Multimodality

Merging information from text & image. **Learning** to exploit information jointly

The example of VQA: visual question answering



⇒ Backpropagate the error ⇒ modify word representations + image analysis

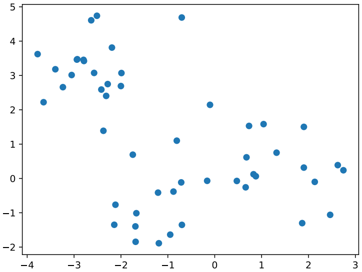
VQA: Visual Question Answering, arXiv, 2016 , A. Agrawal et al.

Generative Artificial Intelligence

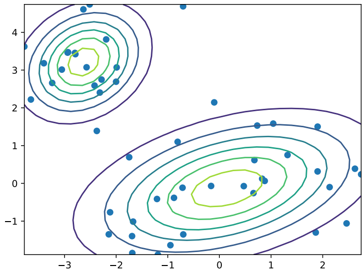
At the origin of statistical modeling

1. **Observing** data (and context)
2. **Modeling** = Choosing probabilistic model / bayesian network
3. **Optimize** parameters (Max Lik., EM, ...)
4. **Sampling** / Inference + Evaluate distances : existing *vs* sampled

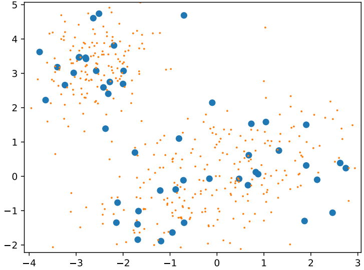
Observations



Modeling: choice+optim.



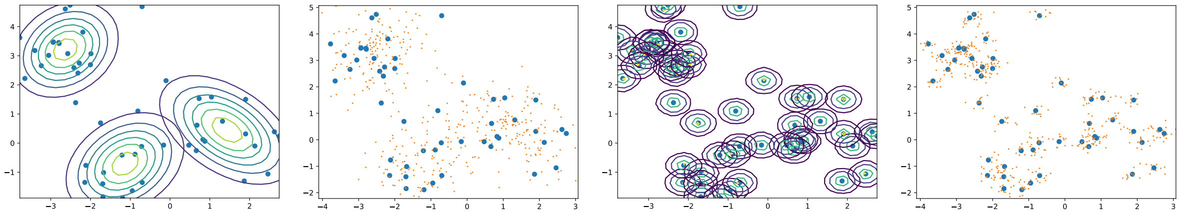
Sampling / eval.



At the origin of statistical modeling

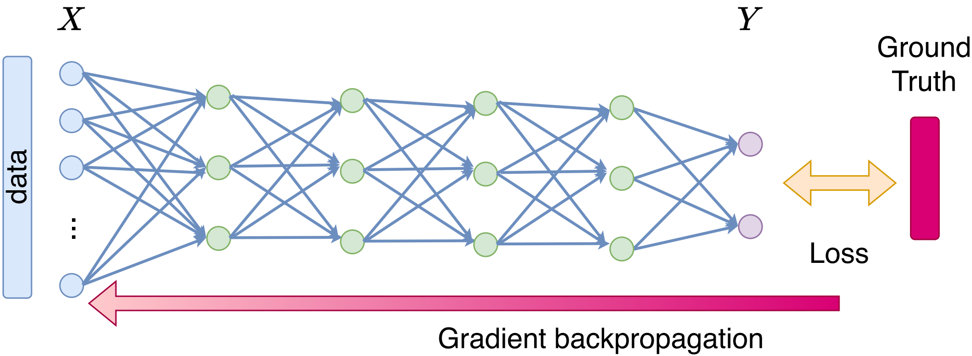
1. **Observing** data (and context)
2. **Modeling** = Choosing probabilistic model / bayesian network
3. **Optimize** parameters (Max Lik., EM, ...)
4. **Sampling** / Inference + Evaluate distances : existing *vs* sampled

Different modeling options / different traps



At the origin of deep learning

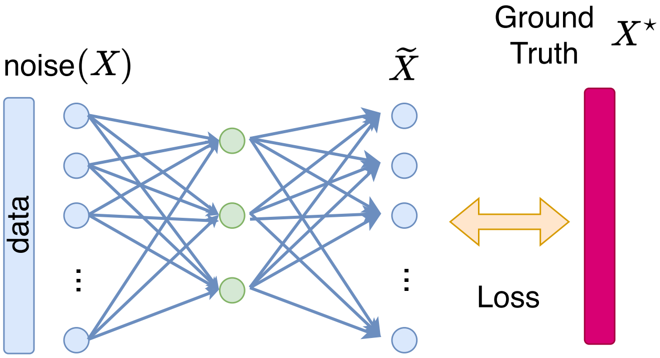
- Gradient vanishing issue in deep architecture



Gradient **weakening => vanishing**

At the origin of deep learning

- Gradient vanishing issue in deep architecture
- Auto-Encoder architecture / facing unsupervised dataset with NN

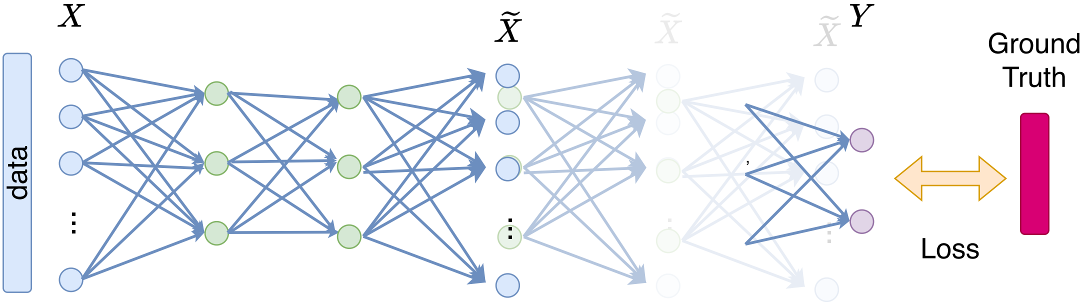


- Denoising
- Low dimensional representation learning (/ PCA, SVD)

Auto-association by multilayer perceptrons and singular value decomposition, Biological Cybernetics, 1988
H. Bourlard & Y. Kamp

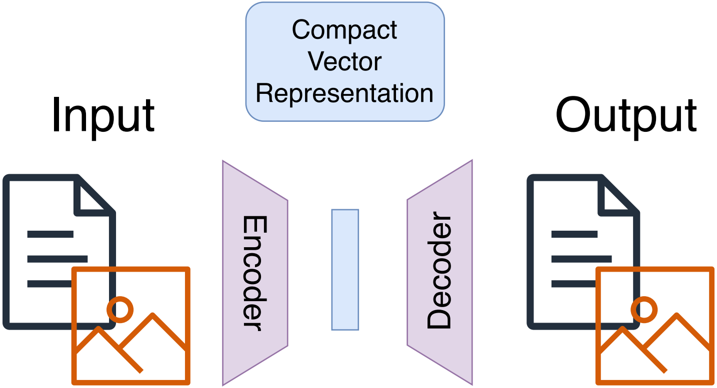
At the origin of deep learning

- Gradient vanishing issue in deep architecture
- Auto-Encoder architecture / facing unsupervised dataset with NN
- Stacked Denoising Auto-Encoder : iterative training / **pretraining**



The difficulty of training deep architectures and the effect of unsupervised pre-training, AIS, PMLR 2009
Erhan, D., Manzagol, P. A., Bengio, Y., Bengio, S., & Vincent, P.

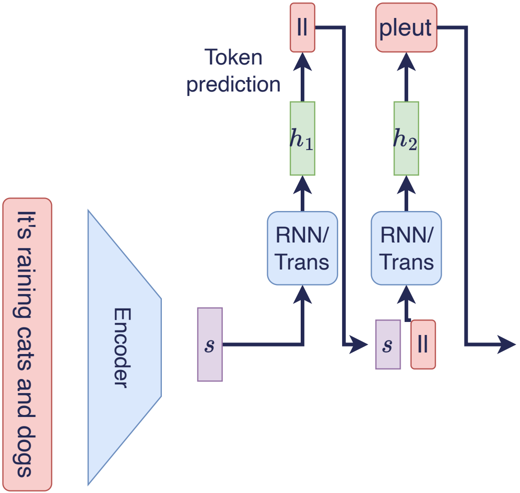
Different Forms of Generative AI



1. Encode an input = construct a vector
2. Decode a vector = *generate* an output

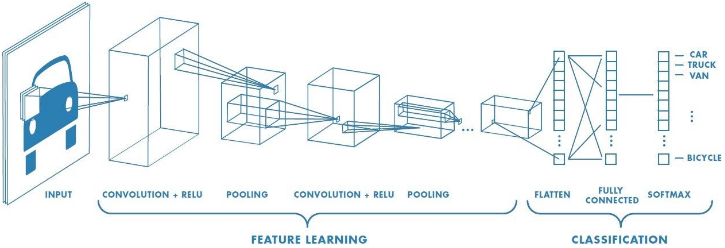
Different Media / Different Architectures

- Texts: classification problem

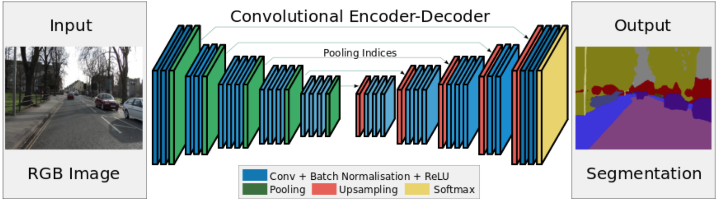


Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem



NVidia Lab.

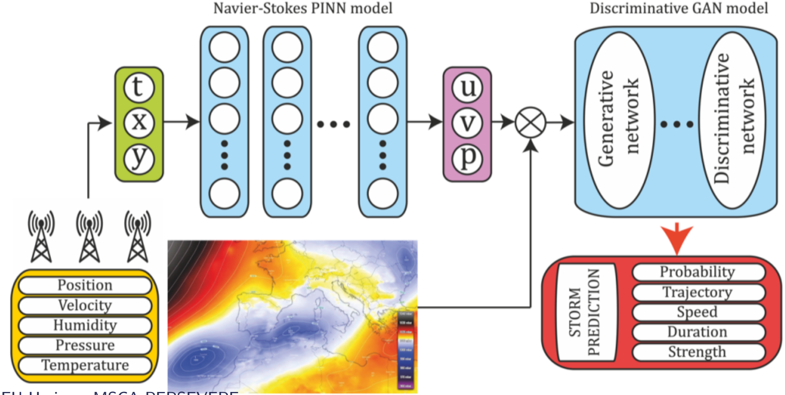


Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem
- Physical processes

- Mix mechanistic and *data-driven* approaches

e.g. Model differential equations in a neural network

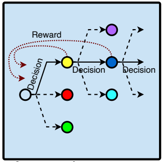


EU Horizon, MSCA PERSEVERE

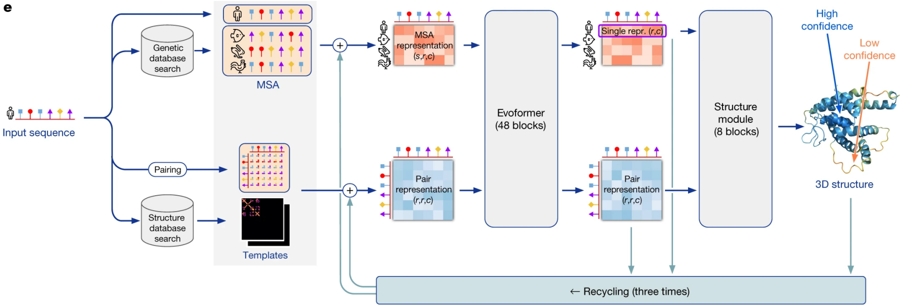
Different Media / Different Architectures

- Texts: classification problem
- Images: multivariate regression problem
- Physical processes
- Complex structures / 3D / graphs: sequential problem

- Reinforcement learning: action/reward



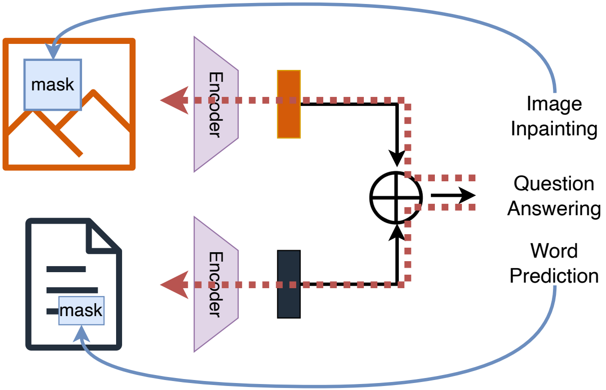
Apprentissage par renforcement



Highly accurate protein structure prediction with AlphaFold, Nature, 2021
Jumper et al.

Multi-Modality

- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*

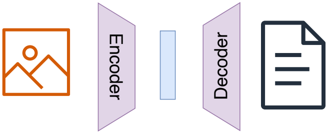


Alignment of representation spaces

Word	*Teraword*	*Knext*
Spoke	11,577,917	372,042
Laughed	3,904,519	179,395
Murdered	2,843,529	16,890
Inhaled	984,613	5,617
Breathed	725,034	41,215

Multi-Modality

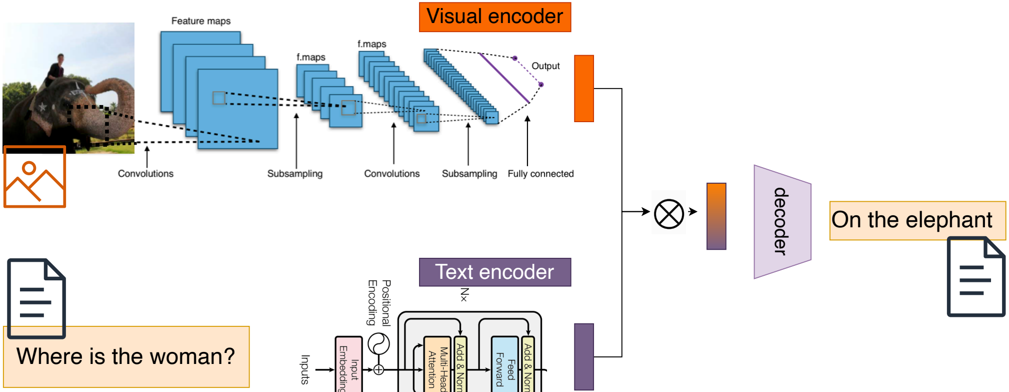
- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*



Show and Tell: image captioning open sourced in TensorFlow, Chris Shallue , Google Research, 2016

Multi-Modality

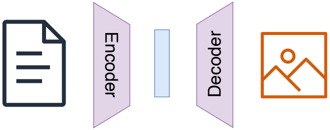
- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*

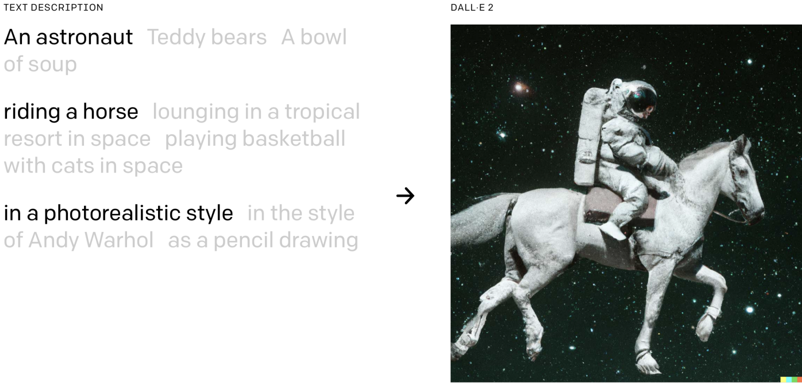


Vqa: Visual question answering, ICCV, 2015
Antol et al.

Multi-Modality

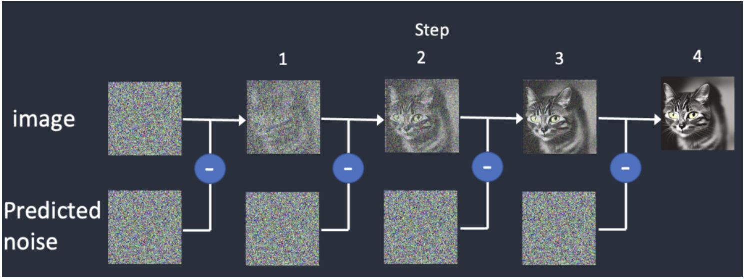
- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning*, *Visual Question Answering*
- Text ⇒ Image: *mid-journey*, *dall-e*, ...



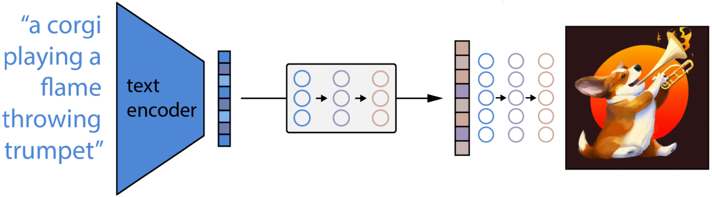


Multi-Modality

- Construction of multimodal representation spaces = *grounding*
- Image ⇒ Text: *Captioning, Visual Question Answering*
- Text ⇒ Image: *mid-journey, dall-e, ...*

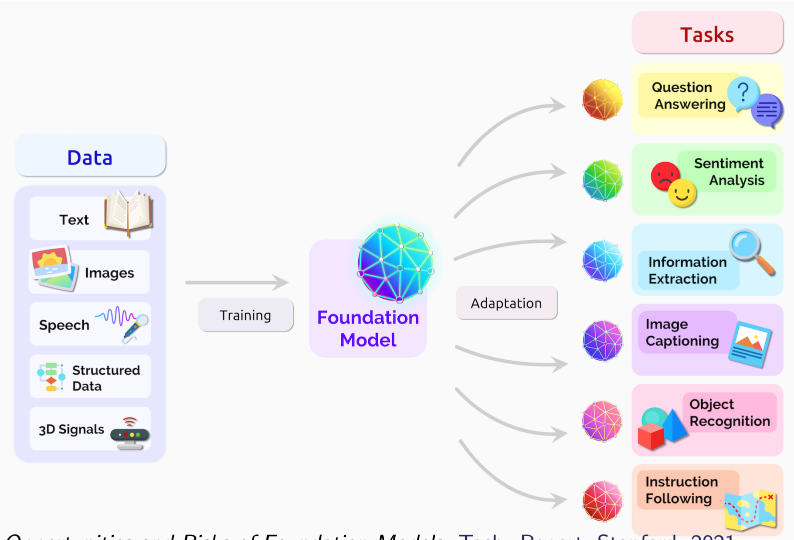


Hierarchical Text-Conditional Image Generation with CLIP Latents, arXiv, 2022
Ramesh et al.



Towards Larger Foundation Models?

- Let the modalities enrich each other



On the Opportunities and Risks of Foundation Models, Tech. Report, Stanford, 2021
Bommasani et al.

Conclusion

The main challenges of multimodality

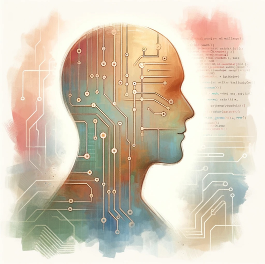
- New applications
  - at the interface between text, image, music, voice, ...
- Performance improvement
  - Better encoding, disambiguation, context encoding
- Explainability (through dialogue)
  - IoT / RecSys / Intelligent Vehicle / ...

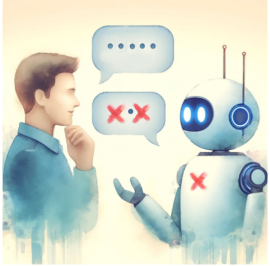
Dall-e

Conclusion

Why So Much Controversy?

- New tool [December 2022]
- \+ Unprecedented adoption speed [1M users in 5 days]
- Strengths and weaknesses... Poorly understood by users
  - Significant productivity gains
  - Surprising / sometimes absurd uses
- Misinterpreted feedback
  - Anthropomorphization of the algorithm and its errors
- Prohibitive cost: what economic, ecological, and societal model?





Dall-e generated images

Machine Learning Limits

chatGPT and the relationship with truth

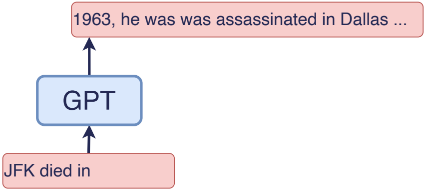
1. Plausibility = grammar, agreement, tense concordance, logical sequences...
   ⇒ Repeated knowledge

2. Predict the most **plausible** word...
   ⇒ produces **hallucinations**

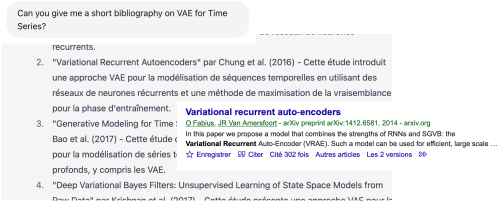
3. **Offline** functioning

4. chatGPT ≠ **knowledge graphs**

5. Brilliant answers...
   And silly mistakes!
   + we cannot predict the errors



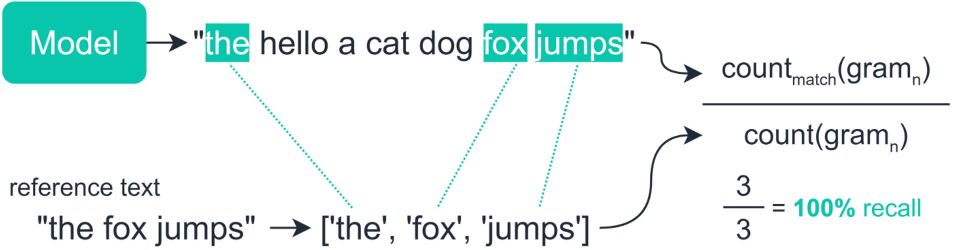
Example: producing a bibliography



Generative AI: how to evaluate performance?

The critical point today

- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?

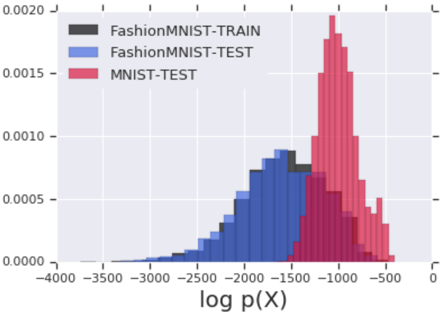


The Ultimate Performance Metric in NLP, J. Briggs, Medium 2021

Generative AI: how to evaluate performance?

The critical point today

- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?



Plausibility

Train

Test

Do Large Language Models Know What They Don't Know?, Yin et al. , ACL, 2023
Do Deep Generative Models Know What They Don't Know?, Nalisnick et al. , ICLR, 2019

Stability/predictability

- Difficult to bound a behavior
- Impossible to predict good/bad answers

⇒ Little/no use in video games

V how old is Obama

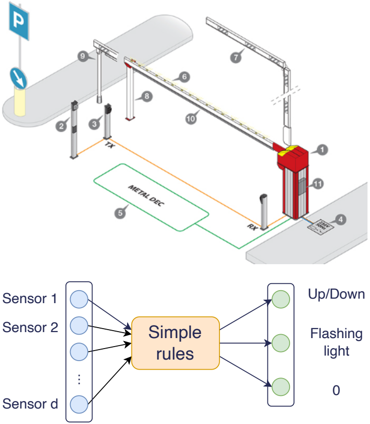
Barack Obama was born on August 4, 1961, making him 61 years old as of February 2, 2023.

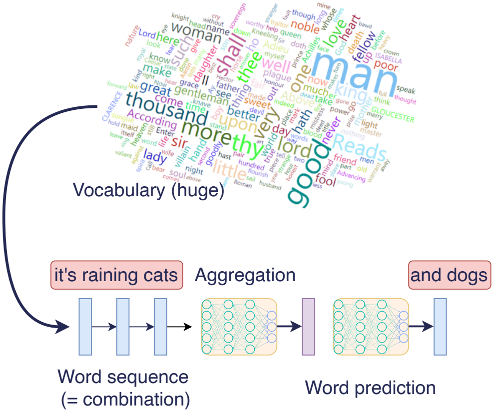
Stability/predictability

- Difficult to bound a behavior
- Impossible to predict good/bad answers

⇒ Little/no use in video games

Stability, explainability... And complexity





- Simple system
- Exhaustive testing of inputs/outputs
- **Predictable** & **explainable**

- Large dimension
- Complex non-linear combinations
- **Non-predictable** & **non-explainable**

Stability, explainability... And complexity

Interpretability *vs* Post-hoc Explanation

Neural networks = **non-interpretable** (almost always)

too many combinations to anticipate

Neural networks = **explainable a posteriori** (almost always)

[Uber Accident, 2018]

- Simple system
- Exhaustive testing of inputs/outputs
- **Predictable** & **explainable**

- Large dimension
- Complex non-linear combinations
- **Non-predictable** & **non-explainable**

Transparency

- Model weights (*open-weight*)... $\Rightarrow$ but not just the weights
- Training data (*BLOOM*) + distribution + instructions
- Learning techniques
- Evaluation

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

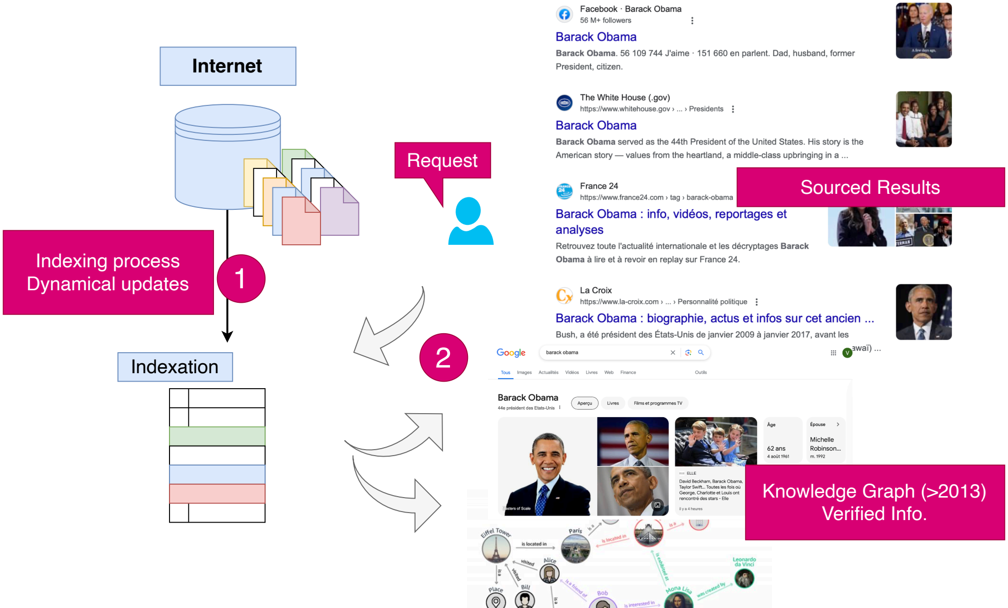
Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	

https://crfm.stanford.edu/fmti/May-2024/index.html

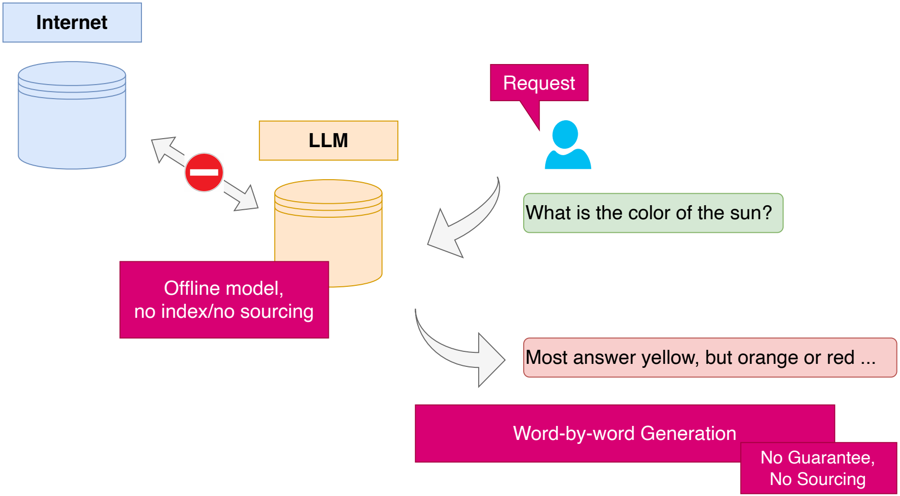
Large Language Models Uses

Information access: from word index to RAG

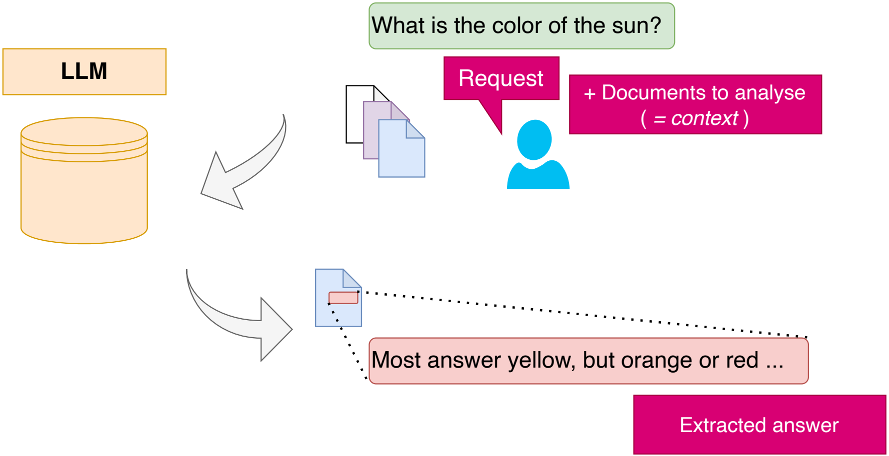


Information access: from word index to RAG

- Asking for information from ChatGPT... A surprising use!
- But is it reasonnable? [Real Open Question (!)]

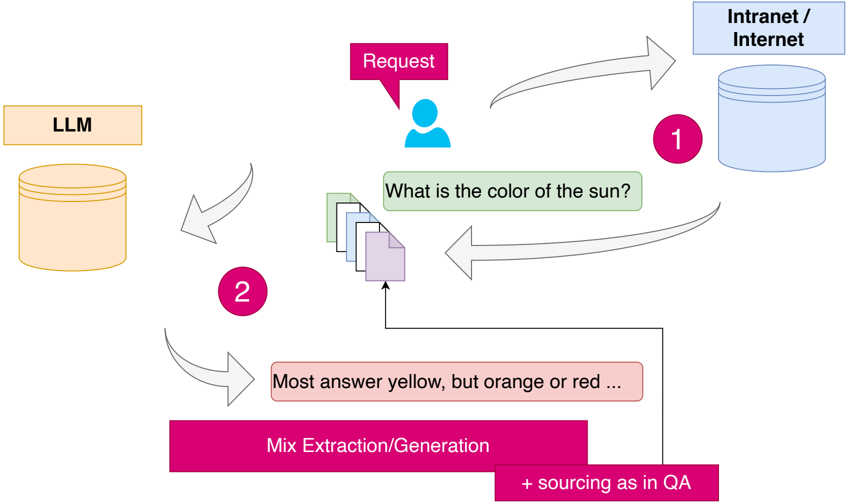


Information access: from word index to RAG



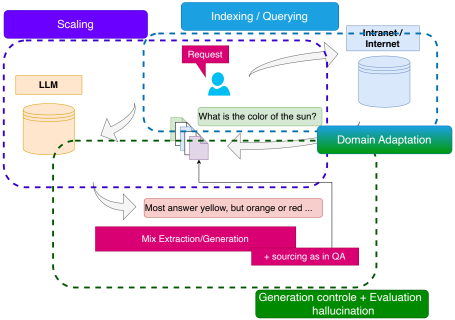
- Web query + analysis, automatic summary, rephrasing, meeting reports...
- (Current) limit on input size (2k then 32k tokens)
- = *pre chatGPT use of LLM for question answering*

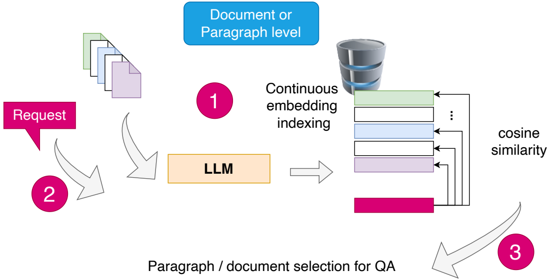
Information access: from word index to RAG



- RAG: Retrieval Augmented Generation
- (Current) limit on input size (2k then 32k tokens)

Information access: from word index to RAG





An introduction to neural information retrieval, IR, 2018
Mitra, B., & Craswell, N.

1. Specific indexing process, relying on (L)Language Model

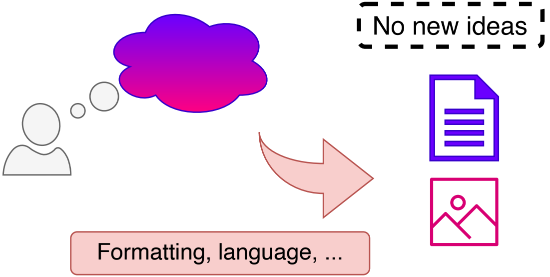
   Lewis et al (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
2. Very large context given to the LLM

   Borgeaud et al (2022) Improving Language Models by Retrieving from Trillions of Tokens
3. Generation controle: hallucination

   LeBronnec et al. 2024, SCOPE: A Preference Fine-tuning Framework for Faithful Data-to-text
4. Domain Adaptation (Biology, Medecine, Technical field...)

Other Uses of Generative AIs

A fantastic tool for formatting

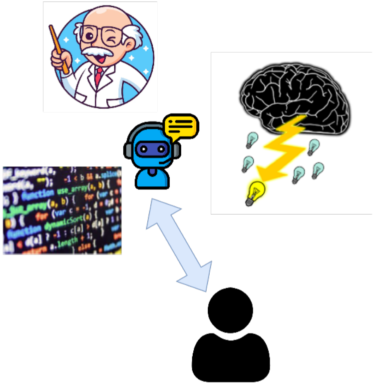


- Personal assistant
  - Standard letters, recommendation letters, cover letters, termination letters
  - Translations
- Meeting reports
  - Formatting notes
- Writing scientific articles
  - Writing ideas, in French, in English

Other Uses of Generative AIs

And a tool for **reflection**!

- Brainstorming
  - Argument development, contradiction search
- Assistant for software development
  - Code generation, error search, ...
  - Documentation
- Educational assistant
  - Wikipedia ++, proposal of outlines for essays,
  - Code explanation / correction proposals
- Document analysis
  - Information extraction, question-answering, ...

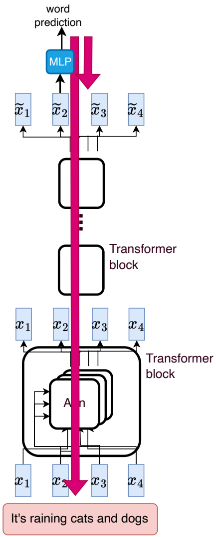


Using Language Models in a Pipeline

Data-scientist requested...

... but it could be a small project anyway

- Adapting to new domain (biology, legal domain, technical field)
    - New words, new meaning, new contexts
    - ⇒ (few-shot), mainly fine-tuning
- Specific task
    - Information extraction, Technological Watch, Question answering
    - ⇒ (zero/few-shot), mainly fine-tuning
- Finetuning
    - Few iterations
    - Specific layers / light approximate gradient...



(Main) Risks derived from ML & LLM

Typology of AI Risks in NLP (L. Weidinger)

Discrimination, exclusion and toxicity

Harms that arise from the language model producing discriminatory and exclusionary speech.

Information hazards

Harms that arise from the language model leaking or inferring true sensitive information.

Misinformation harms

Harms that arise from the language model producing false or misleading information.

Malicious uses

Harms that arise from actors using the language model to intentionally cause harm.

Human-computer interaction harms

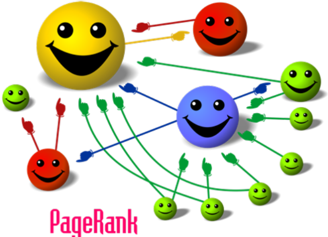
Harms that arise from users overly trusting the language model, or treating it as human-like.

Automation, access and environmental harms

Harms that arise from environmental or downstream economic impacts of the language model.

Access to Information

- Access to dangerous/forbidden information
  - +Personal data
  - Right to digital oblivion
- Information authorities
  - Nature: unconsciously, image = truth
  - Source: newspapers, social media, ...
  - Volume: number of variants, citations (pagerank)
- Text generation: harassment...
- Risk of anthropomorphizing the algorithm
  - Distinguishing human from machine



Machine Learning & Bias

Mustache, Triangular Ears, Fur Texture

Cat

Over 40 years old, white, clean-shaven, suit

Senior Executive

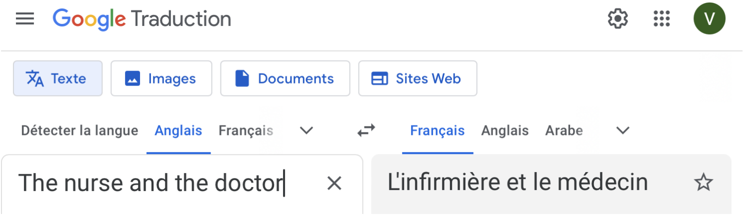
Bias in the data $\Rightarrow$ bias in the responses

Machine learning is based on extracting statistical biases...

$\Rightarrow$ Fighting bias = manually adjusting the algorithm

Machine Learning & Bias

Sterreotypes from *Pleated Jeans*



- Gender choice
- Skin color
- Posture
- ...

Bias in the data $\Rightarrow$ bias in the responses

Machine learning is based on extracting statistical biases...

$\Rightarrow$ Fighting bias = manually adjusting the algorithm

Bias Correction & Editorial Line

Bias Correction:

- Selection of specific data, rebalancing
- Censorship of certain information
- Censorship of algorithm results

⇒ Editorial work... Done by whom?

- Domain experts / specifications
- Engineers, during algorithm design
- Ethics group, during result validation
- Communication group / user response

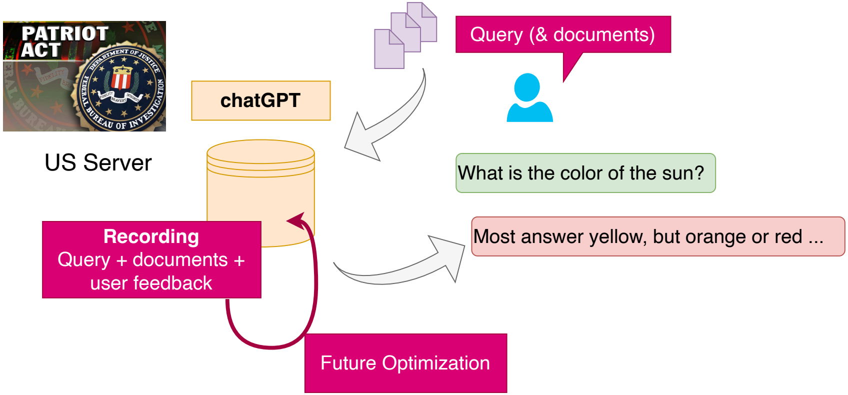
⇒ What legitimacy? What transparency? What effectiveness?

Machine learning is never neutral

1. Data selection
   - Sources, balance, filtering
2. Data transformation
   - Information selection, combination
3. Prior knowledge
   - Balance, loss, a priori, operator choices...
4. Output filtering
   - Post processing

⇒ Choices that influence algorithm results

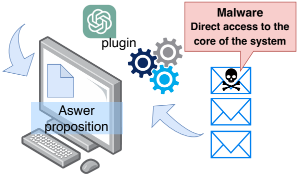
Data Leak(s)

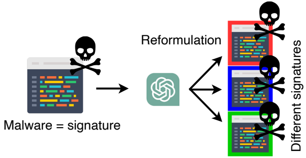


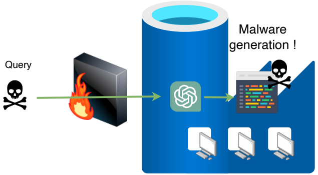
- Transfer of sensitive data
- Exploitation of data by OpenAI (or others)
- Data leakage in future models

Security Issues

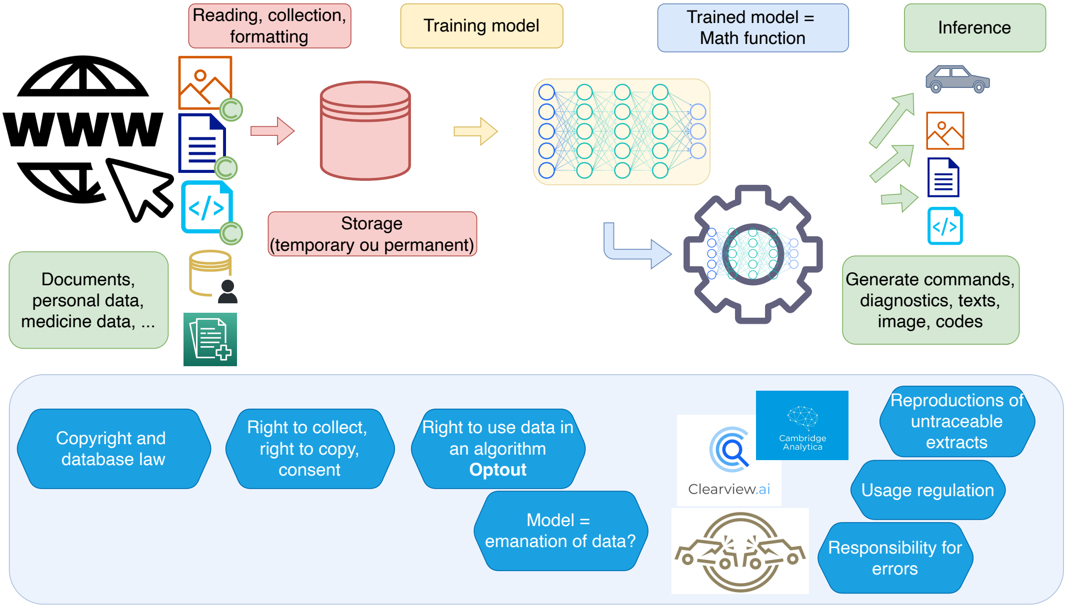
- Plug-ins ⇒ Often significant security vulnerabilities for users
  - Email access / transfer of sensitive information etc...
- Management issues for companies
  - Securing (very) large files
- Increased opportunities for malware signatures
  - ≈ software rephrasing
- New problems!
  - Direct malware generation







Legal Risks/Questions



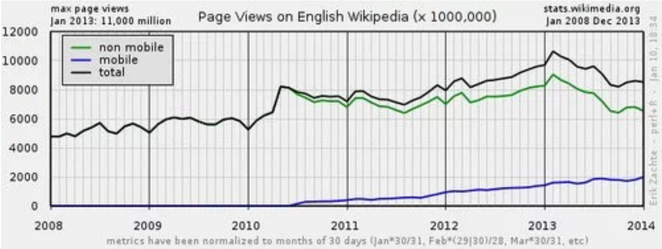
Economic Questions

- Funding/Advertising $\Leftrightarrow$ **visits** by internet users
- Google knowledge graph (2012) $\Rightarrow$ fewer visits, less revenue
- chatGPT = encoding web information... $\Rightarrow$ much fewer visits?

$\Rightarrow$ What **business model for information sources** with chatGPT?

Google's Knowledge Graph Boxes: killing Wikipedia?

by Gregory Kohs

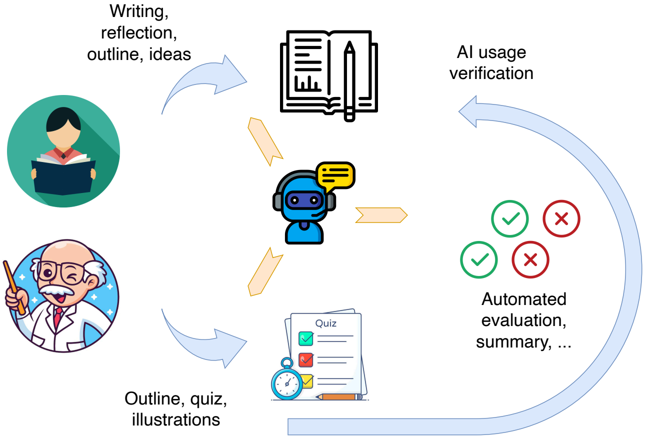


$\Rightarrow$ Who does **benefit from the feedback**? [StackOverFlow]

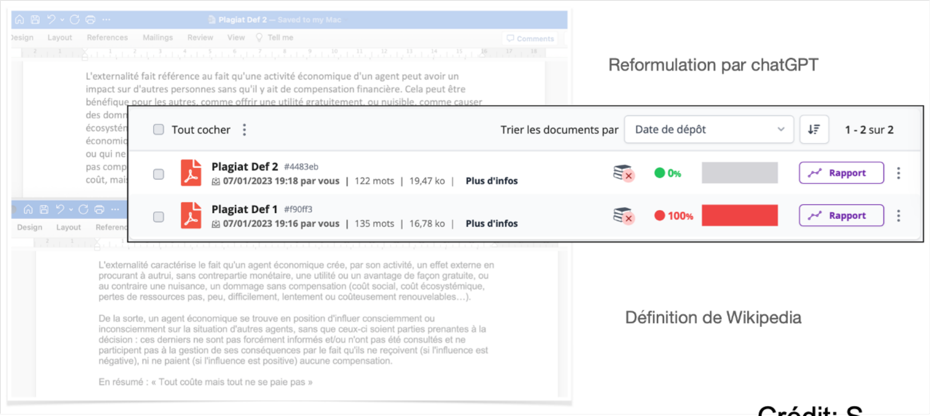
Risks of AI Generalization

AI everywhere =
loss of meaning?

- In the educational domain
- Transposition to HR
- To project-based funding systems



Detection of *texts generated by* chatGPT



Crédit: S. Pajak

Detection of *texts generated by* chatGPT

GPTZero
Detect AI Plagiarism. Accurately

ORIGINALITY.AI

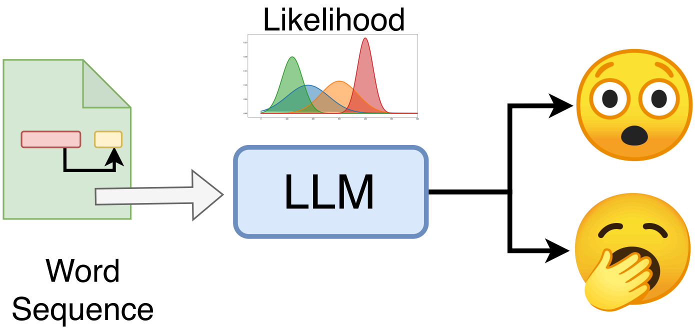
Chat GPT
AI Detector

- **Text classifier** (like for any author)
  - Detection of biases in word choice / phrasing
- Characterization of text **plausibility** (OpenAI, GPTZero)
  - Hyper-fluency of sentences, over-abundance of logical connectors
  - Language model = statistical $\Rightarrow$ measurement between distributions (**perplexity**)
- δ-**plausibility** on perturbed texts (DetectGPT)
- chatGPT *should quickly* integrate **fingerprints** in generated texts

Detectors $\Rightarrow$ $< 100\%$ detection

$+$ confidence level in detection

Detection of *texts used by* chatGPT



- Closed corpora $\Rightarrow$ challenge of detection of texts used in training
- Detection of **likelihood/surprise of observed word sequences**

How to approach the ethics question?

Medicine

1. **Autonomy:** the patient must be able to make informed decisions.
2. **Beneficence:** obligation to do good, in the interest of patients.
3. **Non-maleficence:** avoid causing harm, assess risks and benefits.
4. **Justice:** fairness in the distribution of health resources and care.
5. **Confidentiality:** confidentiality of patient information.
6. **Truth and transparency:** provide honest, complete, and understandable information.
7. **Informed consent:** obtain the free and informed consent of patients.
8. **Respect for human dignity:** treat all patients with respect and dignity.

Artificial Intelligence

1. **Autonomy:** Humans control the process
2. **Beneficence:** including the environment?
3. **Non-maleficence:** Humans + environment / sustainability / malicious uses
4. **Justice:** access to AI and equal opportunities
5. **Confidentiality:** what about the Google/Facebook business model?
6. **Truth and transparency:** the tragedy of modern AI
7. **Informed consent:** from cookies to algorithms, knowing when interacting with an AI
8. **Respect for human dignity:**

How to approach the ethics question?

Medicine

1. **Autonomy:** the patient must be able to make informed decisions.
2. **Beneficence:** obligation to do good, in the interest of patients.
3. **Non-maleficence:** avoid causing harm, assess risks and benefits.
4. **Justice:** fairness in the distribution of health resources and care.
5. **Confidentiality:** confidentiality of patient information.
6. **Truth and transparency:** provide honest, complete, and understandable information.
7. **Informed consent:** obtain the free and informed consent of patients.
8. **Respect for human dignity:** treat all patients with respect and dignity.

Artificial Intelligence

1. **Autonomy:** Humans control the process
2. **Beneficence:** including the environment?
3. **Non-maleficence:** Humans + environment / sustainability / malicious uses
4. **Justice:** access to AI and equal opportunities
5. **Confidentiality:** what about the Google/Facebook business model?
6. **Truth and transparency:** the tragedy of modern AI
7. **Informed consent:** from cookies to algorithms, knowing when interacting with an AI
8. **Respect for human dignity:**

Conclusion

Tools and Questions

New tools:

- New ways to handle existing problems
- Address new problems
- ... But obviously, it doesn't always work!
- AI often makes mistakes (assistant *vs* replacement)

Learning to use an AI system

- AI not suited for many problems
- AI = part of the problem (+interface, usage, acceptance...)

Maturity of Tools & Environments

(More) mature tools

- **Environments**: Jupyter, Visual Studio Code, ...
- **Machine Learning** `Scikit-Learn`: blocks to assemble
    - Training: 1 week
    - Project completion: few hours to few days
- **Deep Learning** `pytorch`, `tensorflow`: building blocks... but more complex
    - Training: 2-5 weeks
    - Project completion: few days to few months
    - Mandatory for text and image

- A data project = 10 or 100 times less time / 2005
- Developing a project is **accessible to non-computer scientists**

Levels of Access to Artificial Intelligence

1. User via an interface: *chatGPT*
   - WARNING: some training is still required (2-4h)
2. Using Python libraries
   - Basics on protocols
   - Standard processing chains
   - Training: 1 week-3 months (ML/DL)
3. Tool developer
   - Adapt tools to a specific case
   - Integrate business constraints
   - Build hybrid systems (mechanistic/symbolic)
   - Mix text and images
   - Training: $\geq$ 1 year

Digital Sovereignty: the Entire Chain

Pre-trained model construction

Data Mastery
- Collection/balance
- Cleaning

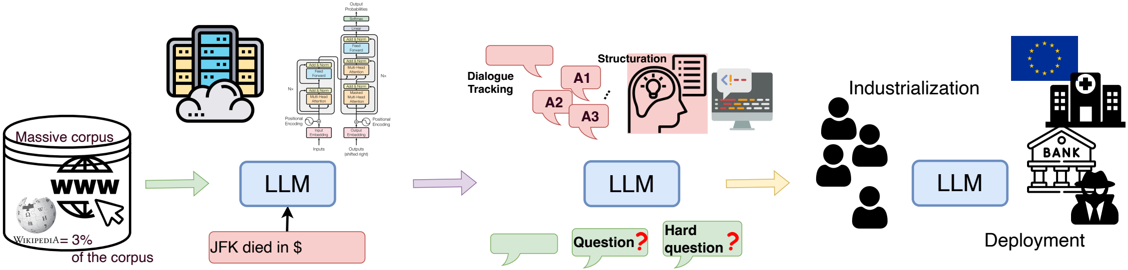
Training
- Computation power (x1000 GPU)
- Architecture ML

Model Fine-Tuning

Data Mastery & Construction
- Human interactions +++
- Dataset cost
- Domain adaptation

Model exploitation

Optimization / Cost reduction
- MLOps skills
- Local deployment



A Multitude of Professions

Data architect / manager

- Data management & hardware devices (storage, network, …)

Data Engineer

- Update & Query on the data

Data Analyst

- Data visualization (chart, indicators, …)
- Statistical trends

Prompt Engineer

- Query on LM/foundation models with "prompts"

Data Scientist

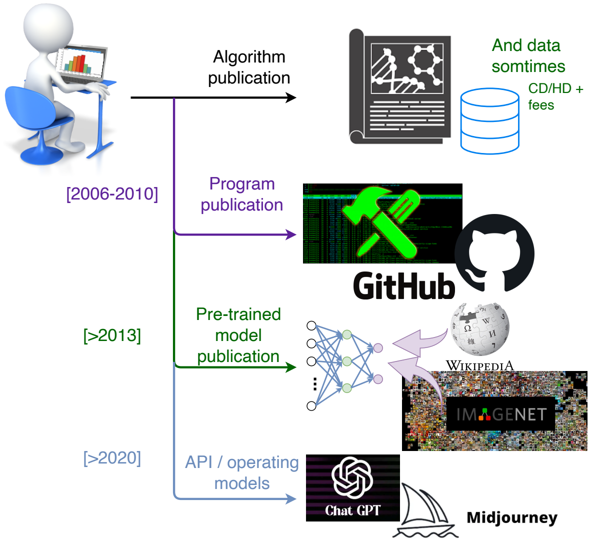
- Query the data / critical selection & balance
- Algorithm development / adaptation / evaluation
- Advanced data visualization

MLOps Engineer

- Algorithm optimization
- Industrialize software solutions

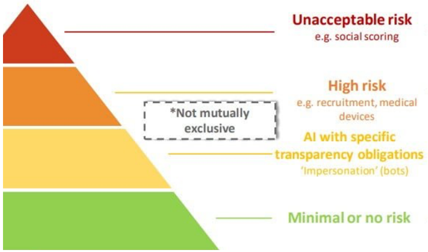
+ **DPO** :
Data Protection Officer

A Multitude of Professions



Factors of Acceptability for Generative AI

1. **Utilitarianism**:
   - Performance (acceptance factor of chatGPT)
   - Reliability / Self-assessment
2. **Non-dangerousness**:
   - Bias / Correction
   - Transparency (editorial line, human/machine confusion)
   - Reliable Implementation
   - Sovereignty (?)
   - Regulation (AI act)
     - Avoid dangerous applications
3. **Know-how**:
   - Training (usage/development)



chatGPT: A Simple Step

- Training & Tuning Costs

  4-5 Million Euros / training $\Rightarrow$ chatGPT is **poorly trained**!

- Data Efficiency

  chatGPT $>$ 1000x a human's lifetime reading

- Identify Entities, Cite Sources

  Anchoring responses in knowledge bases
  Anchoring responses in sources

Sam Altman
@sama

ChatGPT launched on wednesday. today it crossed 1 million users!

8:35 AM · Dec 5, 2022

3,457 Retweets **573** Quote Tweets **52.8K** Likes

- Multiplication of initiatives: GPT, LaMBDA, PaLM, BARD, BLOOM, Gopher, Megatron, OPT, Ernie, Galactica...

- Public involvement,

  impact on information access