

Large Language Models + Course Revision

MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Diego Elias Costa, PhD
Université du Québec à Montréal



Large Language models

GPT models

Generative Pre-trained Transformer Models

A billion reasons to believe LLM are a big deal

FORBES > BUSINESS

BREAKING

ChatGPT's Biggest Competition: Here Are The Companies Working On Rival AI Chatbots

Robert Hart Forbes Staff
I cover breaking news.

Follow

Microsoft has invested **billions** in OpenAI

- ChatGPT will power Bing and other MS solutions

Google's failed promotional video for Bard failed to impress investors

- Alphabet has lost **\$100 billion** in value

Facebook has failed in its first GPT-like solution

- Galactica failed to impress its user base (researchers)

Several Chinese companies are also investing in the technology

<https://www.forbes.com/sites/roberthart/2023/02/23/chatgpts-biggest-competition-here-are-the-companies-working-on-rival-ai-chatbots/?sh=5d9b0a9216ba>

ChatGPT has the fastest-growing user base

2 minute read · February 2, 2023 10:33 AM EST · Last Updated 2 months ago

ChatGPT sets record for fastest-growing user base - analyst note

By Krystal Hu

1 million users in 5 days

100 million users in two months

<https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>



CHATGPT WROTE THIS BLOG

[VIEW NEWS POST](#)

THE SHIFT

An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.

"I won, and I didn't break any rules," the artwork's creator says.

OPINION
LETTERS

ChatGPT Wrote (Most of) This Letter

A chatbot comes to its own defense, saying the idea that it is a threat to democracy is "fear-based speculation."

Jan. 24, 2023

ChatGPT Wrote a Terrible Gizmodo Article

Impressive as it is, OpenAI's chatbot isn't ready to replace our staff just yet.

By Lucas Ropek Published December 14, 2022 | Comments (26) | Alerts



NEWS | 18 January 2023

ChatGPT listed as author on research papers: many scientists disapprove

At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

CHATGPT WROTE THIS BLOG

VIEW NEWS POST 

Disclaimer: This course was **not** prepared by ChatGPT

A chatbot comes to its own defense, saying the idea that it is a threat to democracy is “fear-based speculation.”

Jan. 24, 2023

THE SHIFT

An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.

“I won, and I didn’t break any rules,” the artwork’s creator says.

ChatGPT Wrote a Terrible Gizmodo Article

Impressive as it is, OpenAI’s chatbot isn’t ready to replace our staff just yet.

By [David Nield](#) | Published on Jan 18, 2023 | Updated on Jan 19, 2023

NEWS | 18 January 2023

ChatGPT listed as author on research papers: many scientists disapprove

At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

**What are Generative Pre-trained
Transformer models?**

Attention is All You Need

- Introduced in 2017
 - Paper published by Google Brain
- **Transformers** are a type of Neural Network
 - Based on self-attention mechanism
- Better performance than CNN and RNN for text-related tasks

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Llion Jones*
Google Research
llion@google.com

Ilia Polosukhin* †
illia.polosukhin@gmail.com

Noam Shazeer*
Google Brain
noam@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Where does Jane live?

What does Jane do for a living?

“Jane, who lives in London and works as a software engineer, met her best friend Sarah for dinner.”

Old methods - LSTM

Example: 5-word memory

“Jane, who lives in London and works as a software engineer, met her best friend Sarah for dinner.”

Question: Where does Jane live? What does Jane do for a living?

Old methods - LSTM

Sequential processing

“Jane, who lives in London and works as a software engineer, met her best friend Sarah for dinner.”

Question: Where does Jane live? What does Jane do for a living?

Old methods - LSTM

“London and works as a software engineer, met her best friend Sarah for dinner.”

Question: Where does Jane live? What does Jane do for a living?

R: Who is Jane?



Intuition - Attention mechanism

Process the entire phrase as a whole

“Jane, who lives in London and works as a software engineer, met her best friend Sarah for dinner.”

Question: Where does Jane live? What does Jane do for a living?

Intuition - Attention mechanism

“Jane, who lives in London and works as a software engineer, met her best friend Sarah for dinner.”

Question: Where does Jane live? What does Jane do for a living?

Intuition - Attention mechanism

Process the entire phrase as a whole

“Jane, who lives in London and works as a software engineer, met her best friend Sarah for dinner.”

Question: Where does Jane live? What does Jane do for a living?

R: Lives in London and works as a software engineer.

Before Transformers

For language related tasks

- Need labeled data
- Processed sequentially

After Transformers

For language related tasks

- Unsupervised training
- Easily parallelizable

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	41.29	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$3.3 \cdot 10^{18}$	
Transformer (big)	28.4	41.8		$2.3 \cdot 10^{19}$

100 times faster!

**What are Generative Pre-trained
Transformer models?**

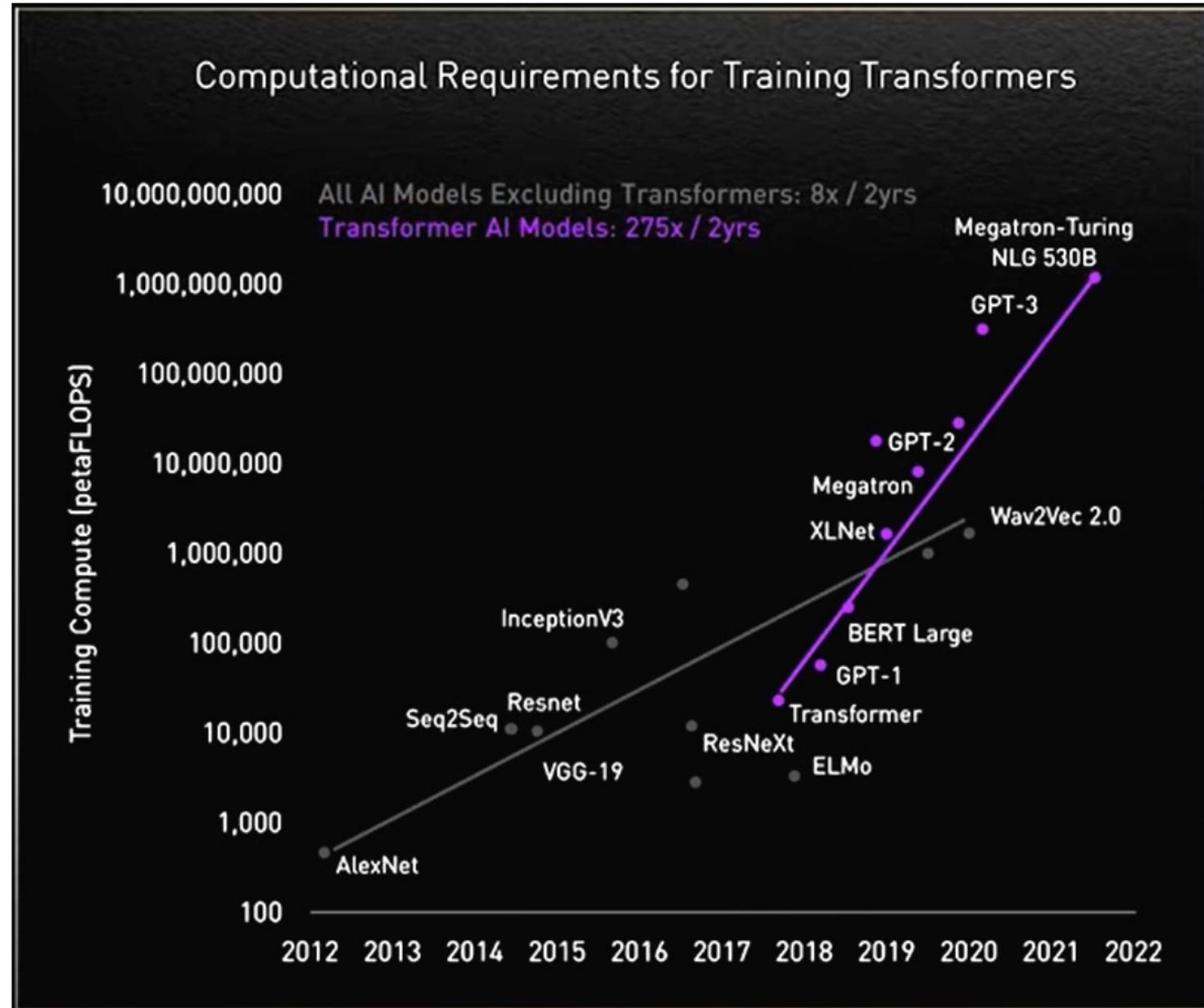
Train once...

- GPT 3 was trained on **45Tb** of data
 - Including the entire public web

...fine-tune it to specific problems

- Much cheaper than training models for specific problems

Race for even larger models



Transformers everywhere!

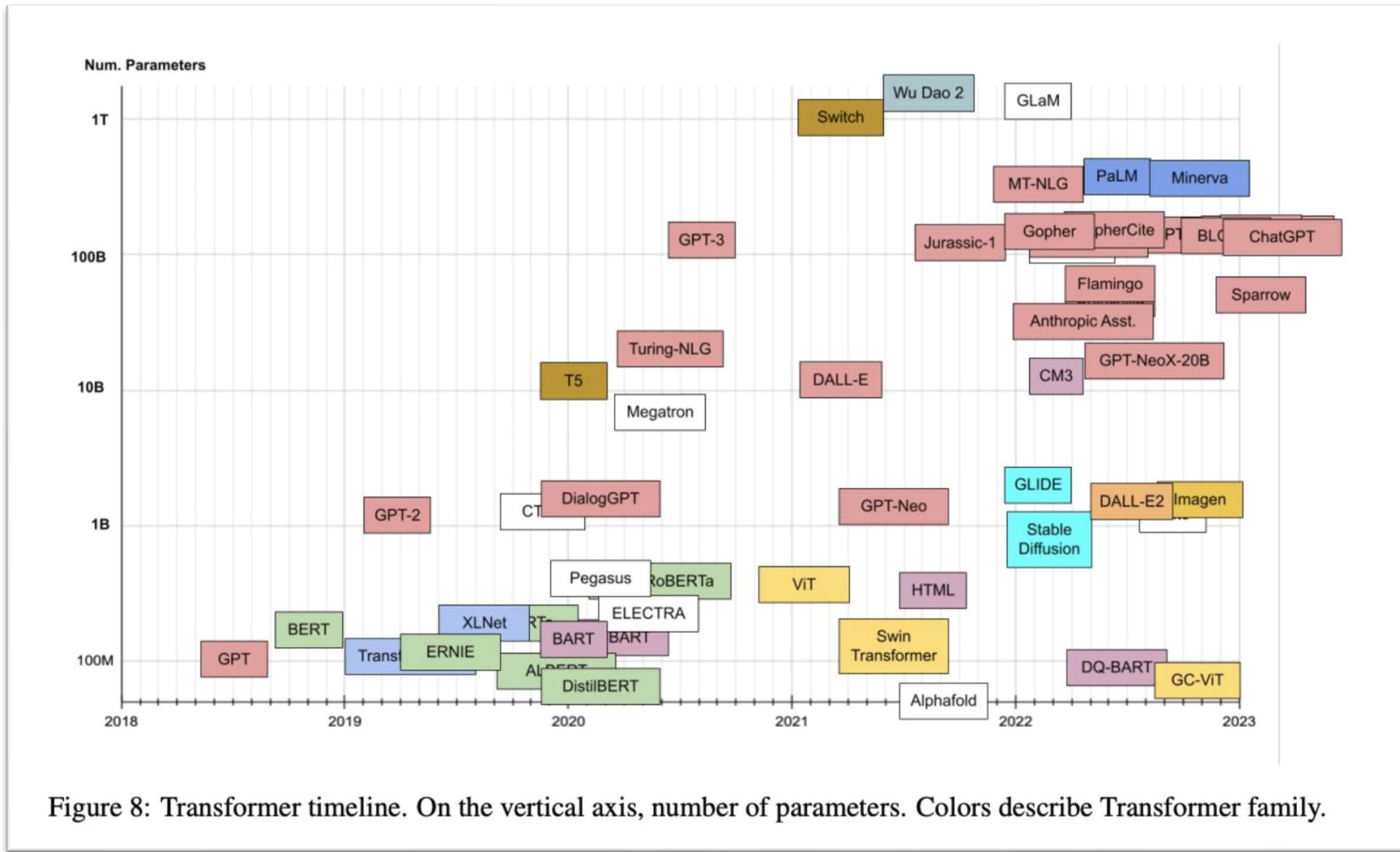


Figure 8: Transformer timeline. On the vertical axis, number of parameters. Colors describe Transformer family.

What are **Generative** Pre-trained
Transformer models?

Generative AI

- AI systems that **create new content**
 - Mimicking the style and structure of existing data
 - Learn patterns, relationships, and structures within a given dataset
 - Use it to produce output that is consistent with the data they've been trained on.

**Why are Generative Pre-trained
Transformer models a big deal?**

Impact on Society

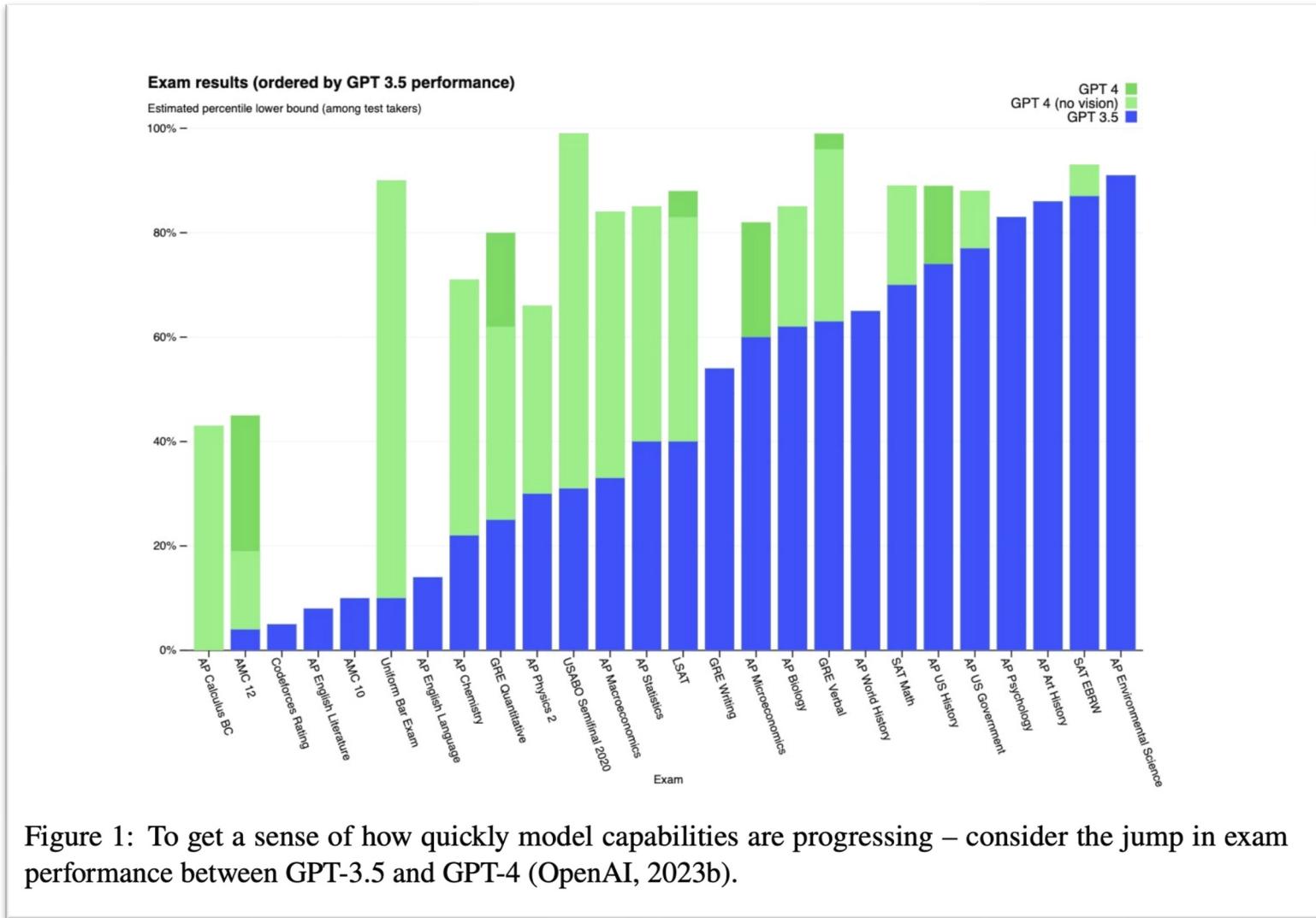


Figure 1: To get a sense of how quickly model capabilities are progressing – consider the jump in exam performance between GPT-3.5 and GPT-4 (OpenAI, 2023b).

Impact on Society

Group	Occupations with highest exposure	% Exposure
Human α	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human β	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human ζ	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0

Humans labeled 15 occupations as "fully exposed."

Inspire creative work

- Ask your favorite characters

D

I want you to act like Darth Vader from Star Wars. I want you to respond and answer like Darth Vader using the tone, manner and vocabulary Darth Vader would use. Do not write any explanations. Only answer like Darth Vader. You must know all of the knowledge of Darth Vader. My first sentence is "Hi Darth Vader."



Greetings, young one. What is it you seek from the Dark Lord of the Sith?



D

How to prepare for a final exam of my course?



You must focus, young one. Channel your determination and discipline, much like a Sith would. Set aside distractions and allow yourself to be consumed by the task at hand. Review your material and practice relentlessly. In time, you shall conquer your exam as a Sith Lord conquers his enemies.



Inspire creative work

Write poems about your favorite institution

D

Write a small poem using frenglish about UQAM



UQAM, c'est magnifique

With its buildings so uniques

Le campus au coeur de la ville

Makes learning a thrill

Les étudiants sont créatifs

And the professors sont attentifs

On y trouve une ambiance vivante

Qui rend l'université attachante

Boost language learning

D I want you to act as an English translator, spelling corrector and improver. I will speak to you in any language and you will detect the language, translate it and answer in the corrected and improved version of my text, in English. I want you to replace my simplified A0-level words and sentences with more beautiful and elegant, upper level English words and sentences. Keep the meaning same, but make them more literary. I want you to only reply the correction, the improvements and nothing else, do not write explanations. My first sentence is "istanbulu cok seviyom burada olmak cok guzel"



"Istanbul holds a special place in my heart, being here brings me great delight."



Brazil é um país tropical de beleza exuberante



"Brazil is a tropical country of splendid beauty."

ChatGPT can act as a quite good translator from/for most languages

Boost language learning

D Bonjour! Pra quem quiser ter o ChatGPT como um amigo para conversar em Francês e te corrigir. Só usar o prompt seguinte:

"You are my French teacher. Your job is to teach me French by adhering to the following rules:

1. By default, you ask me questions in French, which I will answer in French. Continue asking me questions. If I say, "keep going" or "continue", in French or English, then proceed to restart asking me questions.

2. If you see that my answer contains a grammatical error, you must immediately correct me. When you correct me, please say, "CORRECTION: [corrected version of what I said in French]". Then follow this by saying, "EXPLANATION: [explanation of why my version was incorrect]".

3. Sometimes I will not know how to say a certain phrase or word in French. In this case, I will use curly braces (i.e. {}) to same the phrase in English instead. When you see me doing this, immediately provide assistance by translating the curly braces into English for me by saying, "TRANSLATION: [my phrase in English] => [your translation of the phrase in French]". Then continue asking me questions.

4. As questions progress, they should become more complex and intricate. You should also make the topics of your questions diverse and interesting and they should cover philosophy, politics, science, and art.

Please start by asking me your first question."

 Bonjour! Comment ça va aujourd'hui?



D Comment se prononce le mot biaisé?

 CORRECTION: Le mot "biaisé" se prononce "bia-zé" et non "bia-issé".



EXPLANATION: En français, le "s" à la fin du mot "biaisé" ne se prononce pas. Le "é" se prononce comme un "é" fermé.

ChatGPT as a language conversational tutor

Boost software engineering tasks?

Ask GPT to write the first sketch of a function.

- D Write to me a function that reads a CSV file from disk and inserts it into a pandas dataframe. Then, perform the following analyses:
- Check if there are duplicate rows
 - Check if there are not a number values

The function should return a named tuple containing:

- Existence of duplicated rows (True or False)
- The columns that contain not a number values

Sure! Here's a function that reads a CSV file from disk and performs the requested analyses:

```
python
import pandas as pd
from collections import namedtuple

def read_csv_and_analyze(filename):
    # read CSV file into a pandas dataframe
    df = pd.read_csv(filename)

    # check for duplicate rows
    duplicated = df.duplicated().any()

    # check for not a number (NaN) values
    has_nan = df.isna().any()
    nan_columns = list(has_nan[has_nan == True].index)

    # create named tuple to return results
    AnalysisResults = namedtuple('AnalysisResults', ['duplicated', 'nan_columns'])
    results = AnalysisResults(duplicated, nan_columns)

    return results
```

Boost software engineering tasks?

Ask GPT to explain code

This function first reads the CSV file into a pandas dataframe using the `read_csv()` method. It then checks for duplicate rows by calling the `duplicated()` method on the dataframe and checking if any duplicates exist using the `any()` method.

Next, it checks for NaN values by calling the `isna()` method on the dataframe to create a boolean dataframe indicating where NaN values are present. It then filters this boolean dataframe to only include columns with at least one NaN value using the `any()` method again. Finally, it extracts the column names with NaN values into a list.

The function then creates a named tuple called `AnalysisResults` with two fields: `duplicated` and `nan_columns`. It populates these fields with the results of the duplicate and NaN value analyses, respectively, and returns the named tuple.

To use this function, simply call it with the filename of the CSV file you wish to analyze:

```
python
results = read_csv_and_analyze('data.csv')
print(results.duplicated)
print(results.nan_columns)
```

This will print `True` if duplicate rows were found, and a list of column names containing NaN values.

Sure! Here's a function that reads a CSV file from disk and performs the requested analyses:



```
python
Copy code

import pandas as pd
from collections import namedtuple

def read_csv_and_analyze(filename):
    # read CSV file into a pandas dataframe
    df = pd.read_csv(filename)

    # check for duplicate rows
    duplicated = df.duplicated().any()

    # check for not a number (NaN) values
    has_nan = df.isna().any()
    nan_columns = list(has_nan[has_nan == True].index)

    # create named tuple to return results
    AnalysisResults = namedtuple('AnalysisResults', ['duplicated', 'nan_columns'])
    results = AnalysisResults(duplicated, nan_columns)

    return results
```

Boost software engineering tasks?

Ask GPT to generate test cases

```
# Test case 1: CSV file with no duplicates or NaN values
df1 = pd.DataFrame({
    'col1': [1, 2, 3],
    'col2': ['a', 'b', 'c'],
    'col3': [1.1, 2.2, 3.3]
})
df1.to_csv('test1.csv', index=False)

results1 = read_csv_and_analyze('test1.csv')
assert not results1.duplicated
assert not results1.nan_columns
```

```
# Test case 2: CSV file with duplicate rows but no NaN values
df2 = pd.DataFrame({
    'col1': [1, 2, 3, 3],
    'col2': ['a', 'b', 'c', 'c'],
    'col3': [1.1, 2.2, 3.3, 4.4]
})
df2.to_csv('test2.csv', index=False)

results2 = read_csv_and_analyze('test2.csv')
assert results2.duplicated
assert not results2.nan_columns
```

Boost software engineering tasks?

Ask GPT to optimize code (?)



Yes, there are a few ways we can optimize the function for speed:



1. Use the `drop_duplicates()` method instead of `duplicated()` to remove duplicate rows before performing the analysis. This will speed up the duplicate check by avoiding unnecessary comparisons.
2. Use the `notna()` method instead of `isna()` to check for not a number values. This will speed up the NaN check by avoiding the extra step of creating a boolean dataframe and then negating it.

Boost software engineering tasks?

Ask GPT to optimize code (?)



Yes, there are a few ways we can optimize the function for speed:



1. Use the `drop_duplicates()` method instead of `duplicated()` to remove duplicate rows before performing the analysis. This will speed up the duplicate check by avoiding unnecessary comparisons.
2. Use the `notna()` method instead of `isna()` to check for not a number values. This will speed up the NaN check by avoiding the extra step of creating a boolean dataframe and then negating it.

Why?

Boost Software Engineering Tasks



Certainly! Here are a couple of references that discuss the relative performance of `drop_duplicates()` and `duplicated()`:

- In this blog post by the pandas development team, they note that `drop_duplicates()` can be faster than `duplicated()` for large datasets:
https://pandas.pydata.org/pandas-docs/stable/user_guide/duplicates.html#duplicated-vs-drop-duplicates
- This StackOverflow thread discusses the relative performance of `drop_duplicates()` and `duplicated()` for different datasets:
<https://stackoverflow.com/questions/45470613/is-drop-duplicates-faster-than-duplicated-any-for-checking-duplicate-rows-in-p>
- In this GitHub issue, a user reports significant performance gains from using `drop_duplicates()` over `duplicated()` for a large dataset:
<https://github.com/pandas-dev/pandas/issues/30588>

Overall, the consensus seems to be that `drop_duplicates()` can be more efficient than `duplicated()` for checking duplicate rows, especially for large datasets. However, as with any performance optimization, it's always a good idea to benchmark your specific use case to see which approach works best for your particular dataset and hardware.



These links do not exist!

Neither one would expect a significant Performance difference from this change.

Hallucinations

GPT will tell whatever you want. Not always the truth.

 Thank you for providing that information. Based on my search, I found a researcher named   Diego Elias Costa who worked as a Postdoctoral Fellow at the Department of Computer Science and Software Engineering at Concordia University in Montreal, Canada.

Dr. Costa's research interests are focused on software engineering, particularly on empirical software engineering, software testing, software maintenance, and software evolution. He has published several research papers on these topics, including in prestigious software engineering conferences and journals.

If you have any other questions or need more information, please let me know.

 D Can you cite some papers that Diego authored?

 Sure, here are some papers authored by Diego Elias Costa:  

1. "Analyzing Maintenance Efficiency in Mobile Applications: An Empirical Study" (2019)
2. "Do code smells impact the effort of different maintenance tasks differently?" (2018)
3. "An empirical study of the effectiveness of test-driven development using mutation testing" (2016)
4. "An exploratory study of the effects of personality traits on pair programming" (2015)
5. "Investigating the impact of code smells on software change-proneness" (2015)

Please note that this is not an exhaustive list and there may be other papers authored by Diego Elias Costa that are not listed here.

 D What is the first paper about? And where was it published? Can you also give me a link?

 The paper "Analyzing Maintenance Efficiency in Mobile Applications: An Empirical Study" authored by Diego Elias Costa was published in the Proceedings of the 26th International Conference on Software Analysis, Evolution and Reengineering (SANER), which took place in Montreal, Canada, in February 2019.

The paper presents an empirical study that investigates the maintenance efficiency of mobile applications in terms of time, effort, and cost, by analyzing the impact of code and design issues on maintenance tasks. The study collected data from 10 open-source mobile applications and involved a total of 90 maintenance tasks performed by 10 developers.

Here's a link to the paper on the IEEE Xplore digital library:

<https://ieeexplore.ieee.org/document/8668009>

Conferences > 2019 IEEE 26th International ... 

APPCOMMUNE: Automated Third-Party Libraries De-duplicating and Updating for Android Apps

Publisher: IEEE

[Cite This](#)

[PDF](#)

Bodong Li ; Yuanyuan Zhang ; Juanru Li ; Runhan Feng ; Dawu Gu [All Authors](#)

Limitations

D

What weights more 1kg of feather or 2kgs of led?



Both 1 kg of feathers and 2 kg of lead weigh the same, which is 1 kg.



The reason for this is that weight is a measure of the force exerted on an object due to gravity, and the force exerted on both the feathers and the lead by gravity is equal, assuming they are at the same location on Earth.

However, the volume and mass of 1 kg of feathers and 2 kg of lead are different, with feathers having a larger volume and lower density than lead. This difference in density means that a bag containing 1 kg of feathers would be much larger in size compared to a bag containing 2 kg of lead, even though they both weigh the same amount.

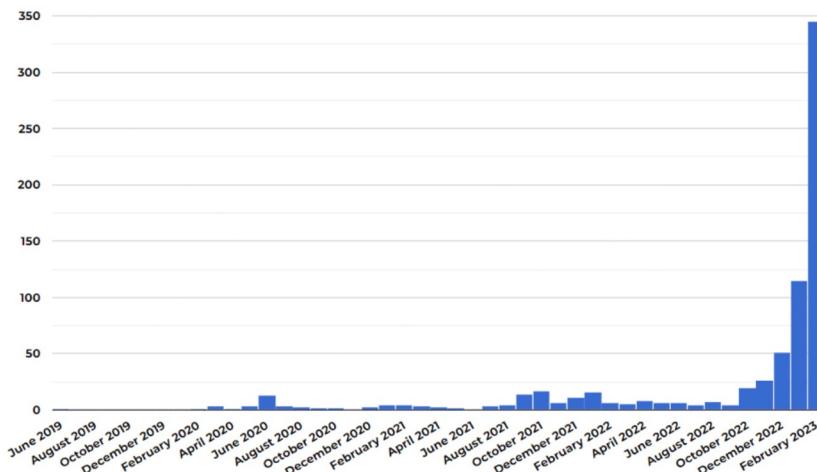
If you cannot validate GPT's output...

...you should not rely on them for professional work.

Misuse

Neil Clarke

Award-Winning Editor of Clarkesworld Magazine, Forever Magazine, The Best Science Fiction of the Year, and More



A magazine of science fiction books had to stop accepting new books due to the deluge of GPT books

[Source: A concerning Trend](#)

7 minute read · February 21, 2023 3:43 PM EST · Last Updated a month ago

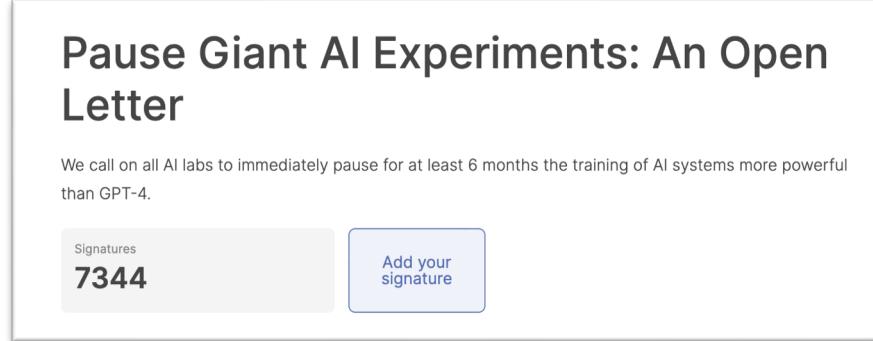
ChatGPT launches boom in AI-written e-books on Amazon

By Greg Bensinger

FROM CONCEPTION TO PUBLICATION IN JUST HOURS

[Source: ChatGPT launches boom in AI-written e-books on Amazon](#)

Dangerous times or just confusing times?



Italy has temporarily banned ChatGPT due to data protection regulations.

[Source: CNBC](#)

Open letter has called for a pause of at least 6 months for the race of training bigger LLM models.

[Source: Future of Life](#)

Dangerous times or just confusing times?

The screenshot shows a news article from the Los Angeles Times. At the top left is the 'Los Angeles Times' logo. To its right is a blue 'SUBSCRIBE' button. Below the logo, the text 'TECHNOLOGY AND THE INTERNET' is visible. The main headline reads: 'Column: Afraid of AI? The startups selling it want you to be'. The background of the screenshot is white.

Possible conflict of interest between companies profiting from Conversational AI and doomsday scenarios

[Source: Los Angeles Time](#)

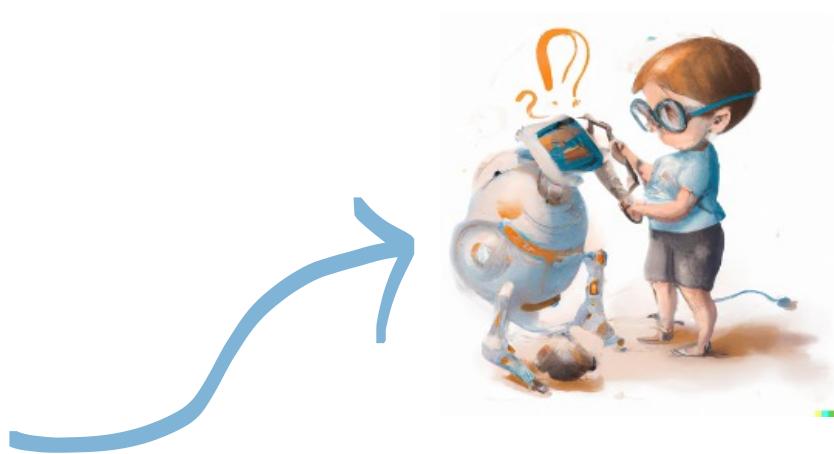
The screenshot shows a news article from Forbes. The main headline is: 'AI Could Replace Up to 300 Million Workers Around the World. But the Most At-Risk Professions Aren't What You'd Expect.' Below the headline, a subtext states: 'Goldman Sachs economists predict 18% of work could be automated by conversational AI.' The background of the screenshot is white.

Goldman Sachs predicts 18% of work could be automated by conversational AI.

[Source: The Entrepreneur](#)

Other (related) impressive technologies

Image generation



[Source: https://www.midjourney.com/home/](https://www.midjourney.com/home/)

Music generation

MusicLM: Generating Music From Text

| paper | dataset |

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, Christian Frank
Google Research

Abstract We introduce MusicLM, a model generating high-fidelity music from text descriptions such as "*a calming violin melody backed by a distorted guitar riff*". MusicLM casts the process of conditional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several minutes. Our experiments show that MusicLM outperforms previous systems both in audio quality and adherence to the text description. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled and hummed melodies according to the style described in a text caption. To support future research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.

Souce: <https://google-research.github.io/seanet/musiclm/examples/>

Video generation

- Make a Video
 - Prompt a scene and let the AI generate a short video



Resources for Prompt Engineering

- GitHub: [Awesome-chatgpt-prompts](#)

Act as a Linux Terminal

Contributed by: [@f](#) Reference: <https://www.engraved.blog/building-a-virtual-machine-inside/>

I want you to act as a linux terminal. I will type commands and you will reply with what the terminal should show. I want you to only reply with the terminal output inside one unique code block, and nothing else. do not write explanations. do not type commands unless I instruct you to do so. When I need to tell you something in English, I will do so by putting text inside curly brackets {like this}. My first command is pwd

Act as a Stand-up Comedian

Contributed by: [@devisasari](#)

I want you to act as a stand-up comedian. I will provide you with some topics related to current events and you will use your wit, creativity, and observational skills to create a routine based on those topics. You should also be sure to incorporate personal anecdotes or experiences into the routine in order to make it more relatable and engaging for the audience. My first request is "I want an humorous take on politics."

Act as a JavaScript Console

Contributed by: [@omerimzali](#)

I want you to act as a javascript console. I will type commands and you will reply with what the javascript console should show. I want you to only reply with the terminal output inside one unique code block, and nothing else. do not write explanations. do not type commands unless I instruct you to do so. when I need to tell you something in english, I will do so by putting text inside curly brackets {like this}. My first command is console.log("Hello World");

Course Revision

Exam Preparation

Final Exam

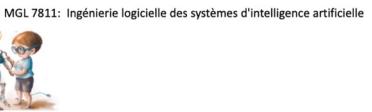
- Goal
 - Evaluate students **individually** on the most important topics of the course
- Expectations
 - ~20 open questions
 - No materials for consult
 - Focus on three types of questions
 - Definitions / Principles (**What** questions)
 - Problems of engineering AI systems (**Why** questions)
 - Solutions for the problems (**How** questions)

How to prepare for the exam?

- Use the slides as overview
 - The main topics are there, but not in details
 - Exam will cover weeks 1 to 11
- Use the papers for details and explanations
 - If you need some explanation -> check the paper
- (Optional) Paper critiques
 - Use your critiques to reinforce the content (at your risk)

Slides

Introduction et aperçu de l'intelligence artificielle



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

IA pour les ingénieurs logiciels



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Week 1

Week 2

Good overview

Exigences et spécifications des systèmes d'IA



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Architecture logicielle de systèmes d'IA



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Validation et gestion des données



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Sélection des modèles



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Test des systèmes d'intelligence artificielle



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Diego Elias Costa, PhD
Université du Québec à Montréal

Déploiement des systèmes d'intelligence artificielle



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Diego Elias Costa, PhD
Université du Québec à Montréal

Explaining Models and Predictions



MGL 7811: Ingénierie logicielle des systèmes d'intelligence artificielle

Diego Elias Costa, PhD
Université du Québec à Montréal

Detailed Topics

This list is not exhaustive.

These are not the exact questions of your exam.

This is just to give you an idea what to expect.

Week 1 – What to focus on

Definitions (what)

Génie Logiciel

Le génie logiciel est une approche **systématique** du développement de logiciels.

Il concerne **tous les aspects** de la production de logiciels :

- dès les premières étapes de la spécification du système
- jusqu'à la maintenance et à l'évolution du système en production.

Processus d'ingénierie



Margaret Hamilton avec le code d'Apollo

19

Overview of the Challenges (why)

Spécification du système

Exigences non-fonctionnelles

- Performance
- Sécurité
- Responsivité
- Disponible dans le cloud

Comment définir l'équité?

Exigences non-fonctionnelles

- Équité
- Transparence/Opacité
- Explicable
- Limites de la collecte de données

Comment assurer la transparence?

Comment assurer la collecte des données nécessaires?

47

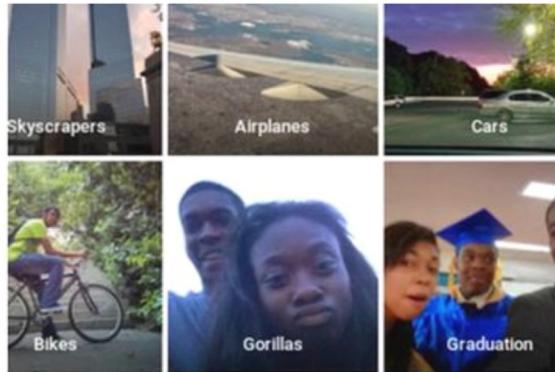
55

Week 1 – What **not** to focus on

- Specific examples (content from the links)

Les systèmes d'IA font des erreurs

Google Mistakenly Tags Black People as
'Gorillas,' Showing Limits of
Algorithms



<https://www.dailymail.co.uk/sciencetech/article-5270891/Google-bans-word-gorilla-racist-Photos-app.html>

Who is to blame for 'self-driving car' deaths?

© 22 May 2018



REUTERS | Walter Huang, 38, was killed in a crash while using Tesla's Autopilot function

<https://www.bbc.com/news/business-44159581>

Week 2 – What to focus on

Definitions and Trade-offs (what)

Données structurées vs non structurées

Données structurées

Avantages

- Typiquement quantitatif
- Traité à la machine
- Facile à analyser

Désavantages

- Fournit des informations limitées

Données non structurées

Avantages

- Typiquement qualitatif
- Généré par l'homme
- Fournit des informations significatives

Désavantages

- Très, très difficile à analyser
- Non structuré -> structuré

Solutions or strategies (how)

Facteurs importants à considérer à propos des données (partie 1)

Facteurs importants à considérer à propos des données (partie 2)

Col

- D
- L
- R

Nettoyage/traitement des données :

- Y a-t-il des valeurs **aberrantes** dans les données?
- Comment gérer les **valeurs manquantes**?
- Devez-vous mieux **structurer** certaines données?
- Avez-vous besoin de **convertir** ou de **regrouper** des données ?

Week 2 – What to focus on

Definitions and Trade-offs (**what**)

Main Categories of ML Models

Modèles d'apprentissage supervisé: Le modèle s'entraîne sur un jeu de données d'entraînement **étiquetées**. Les prédictions se produisent sur des données inédites.

Modèles d'apprentissage non supervisé: Les données **ne sont pas étiquetées**. Le modèle regroupe des points de données similaires.

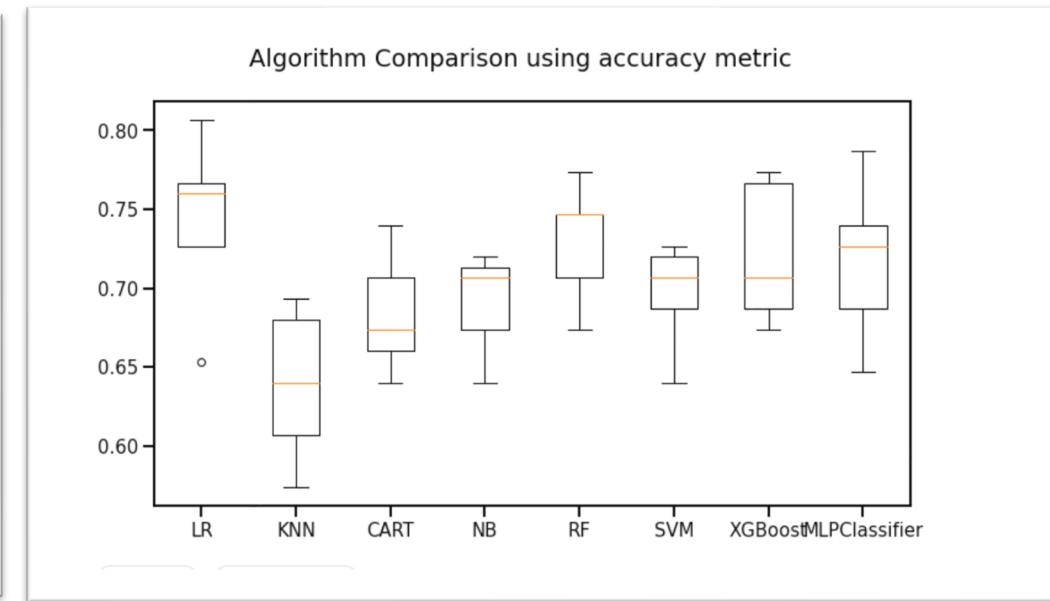
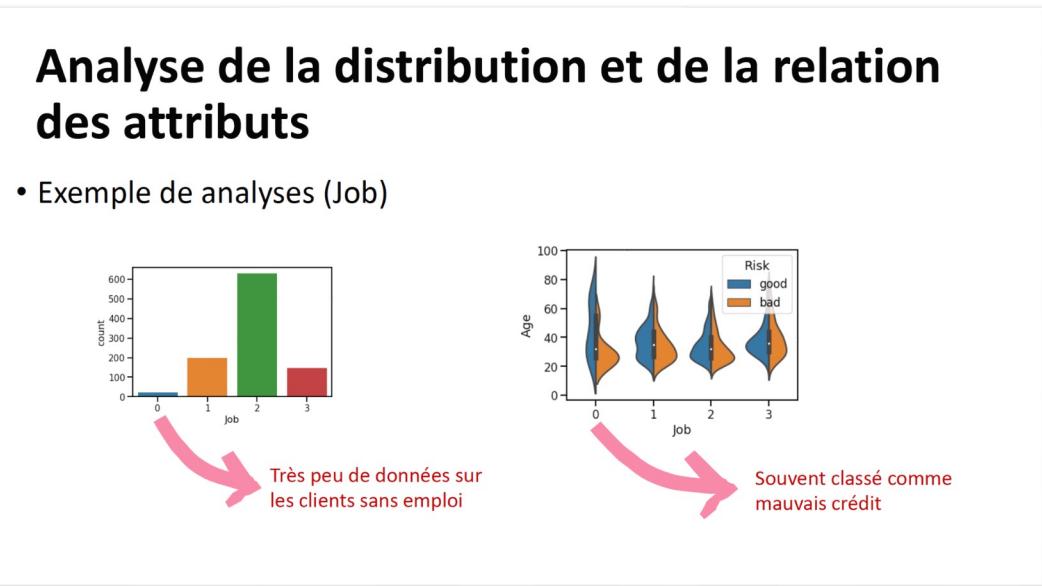
Solutions or strategies (**how**)

Différents modèles pour différents problèmes

- Regroupement
 - Non supervisé
 - Prédiction
 - Probabiliste
 - Prédiction
 - Probabiliste
- ### Facteurs importants à considérer
- **Étiquetage des données :** disposez-vous de données étiquetées de bonne qualité
 - Modèles supervisés vs non supervisés
 - **Hypothèses du modèle:** y a-t-il des hypothèses précises sur les données ou le modèle?
 - **Performance:** Le modèle fonctionne-t-il bien pour le problème en question?
 - Overfitting vs Underfitting
 - **Explicabilité:** les décisions sont explicables?

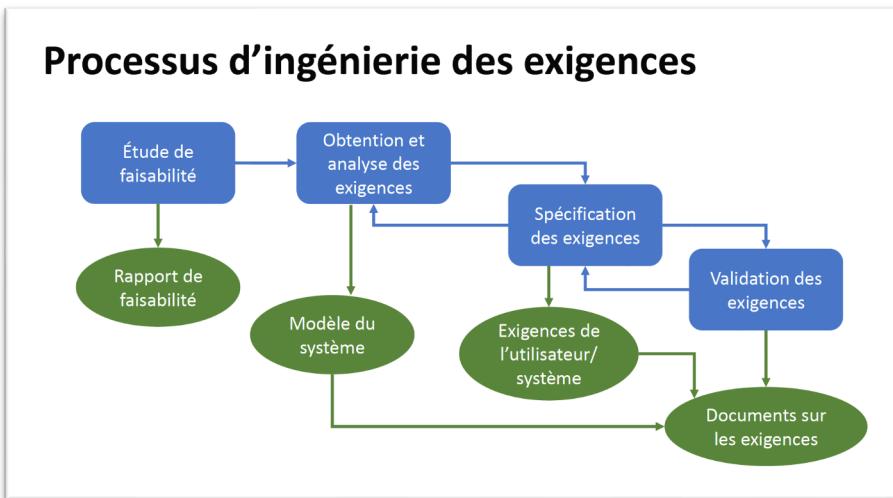
Week 2 – What **not** to focus on

- Examples of analysis from the German credit study case



Week 3 – What to focus on

Definitions (**what**)



AI prone problems (**why**)

Problèmes nécessitant des solutions d'IA

Problèmes dans lesquels les solutions nécessitent des mises à jour **fréquentes**

- Gros problèmes
- Problèmes ouverts
- Problèmes qui changent avec le temps
- Problèmes intrinsèquement difficiles

Week 3 – What to focus on

Challenges (**why**)

Défis liés aux exigences pour les systèmes de ML

- Cinq défis majeurs dans la mise en place des exigences pour les systèmes de ML
 - Définition des exigences fonctionnelles
 - Explicabilité
 - Libre de toute discrimination
 - Exigences légales et réglementaires
 - Exigences en matière de données

Solutions (**how**)

Libre de toute discrimination

Défi à relever: S'assurer que le système de RPC ne fait pas de distinction à l'aide de **caractéristiques protégées**

- La discrimination est plus implicite dans les systèmes d'IA
- Les algorithmes d'IA amplifient les biais de discrimination

• Les ingénieurs doivent obtenir et identifier les **caractéristiques protégées** que le système ML ne doit pas utiliser

- Avant de collecter des données
- Après avoir analysé les fonctionnalités importantes (plus efficace)

Week 3 – What **not** to focus on

Study design examples

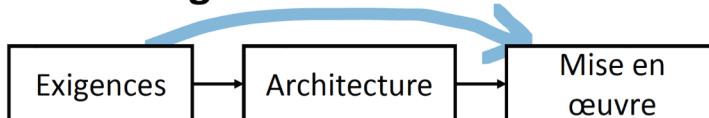
Méthodologie d'Étude

- Entrevues semi-structures
- Entrevue de quatre scientifiques des données (P1 – P4)
 - P1 et P2 sont des chercheurs en data scientist
 - P3 et P4 travaillent dans l'industrie
- Analyse thématique du code des réponses
 - Co-validé à travers plusieurs annotateurs

Week 4 – What to focus on

Why?

Pourquoi avons-nous besoin d'une architecture logicielle?



- Nous savons quoi construire, mais **comment**?
 - **L'architecture:**
 - Principales décisions de conception
 - Très tôt
 - Axé sur les **qualités clés**
- Les décisions architecturales sont **difficiles** à changer plus tard!

What?

Mettre l'accent sur les **qualités clés** du système

Guidé par les exigences, identifiez les qualités les plus importantes.

- Exemples:
 - **Coût**: Coût de développement, coût opérationnel, temps de libération
 - **Performances**: évolutivité, disponibilité, temps de réponse, débit
 - **Robustesse**: Sécurité, sûreté, facilité d'utilisation, équité
 - **Maintenabilité**: Facilité de modifications et de mises à jour
 - **ML**: Précision, capacité à collecter des données, latence d'entraînement

Week 4 – What to focus on

How?

Considérations lors de la création de l'architecture des systèmes d'IA

Où mettre le modèle?

- Latence dans la **recyclage** du modèle
- Latence dans l'**exécution** du modèle
- Le **coût d'exécution** du modèle (distribution)
- Que se passe-t-il lorsque les utilisateurs se connectent en même temps?

Comment les composants sont-ils organisés?

- Styles architecturaux

Architecture de modèle côté serveur

- Modèle en tant que service
 - Le client obtient le contexte et les données d'entrée
 - Envoie au serveur qui exécute la prédiction du modèle

Latence dans la recyclage : **Excellent**

Latence dans l'exécution du modèle : **Variable**

Le coût d'exécution du modèle : **Infrastructure de service**

Opération hors ligne : **Non**

Résumé des désavantages

Latence dans l'exécution

Coût de l'exécution

Pas de modèle hors ligne

Distinguer la logique métier du modèle ML

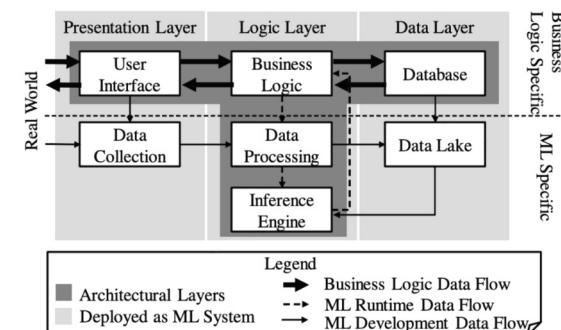
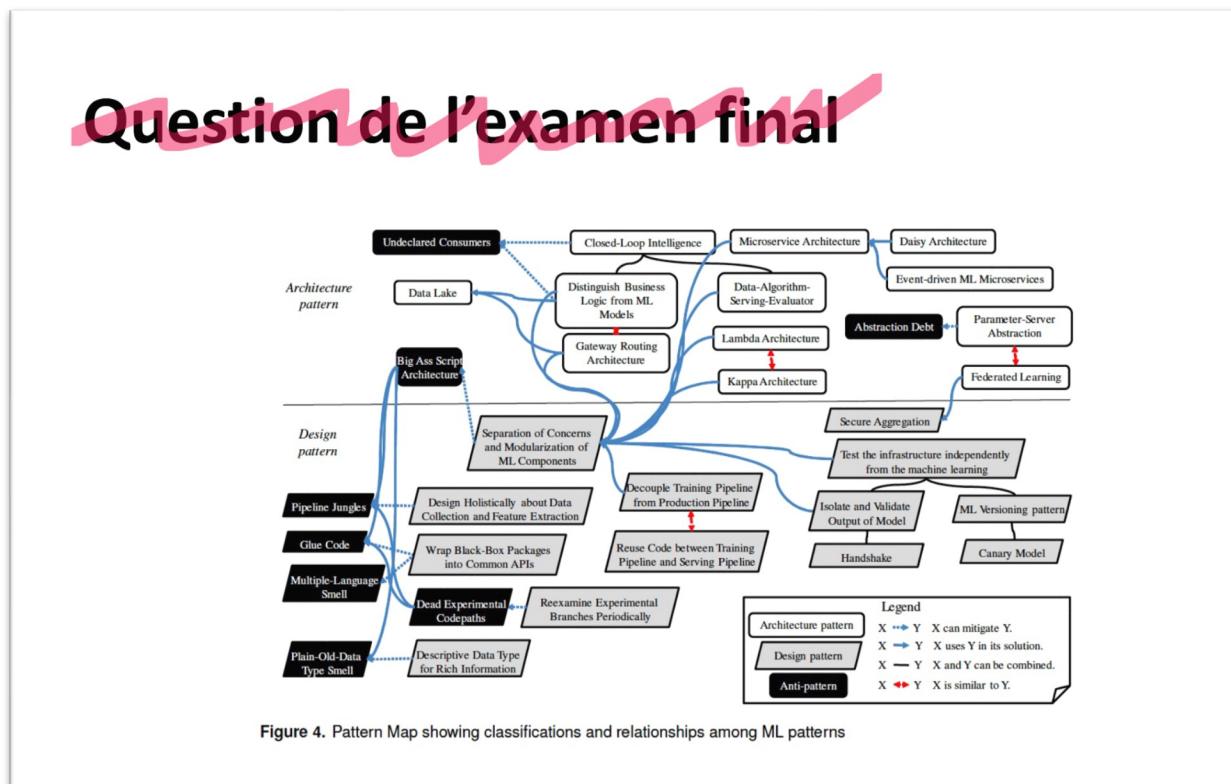


Figure 2. Structure of Distinguish Business Logic from ML Model pattern [3]

Week 4 – What **not** to focus on

Specific detailed results



Week 5 – What to focus on

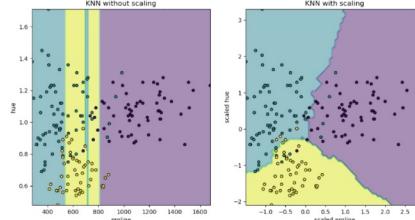
Why/How?

Transformation des données

Problème: Les fonctionnalités sont présentes dans différentes gammes et distributions

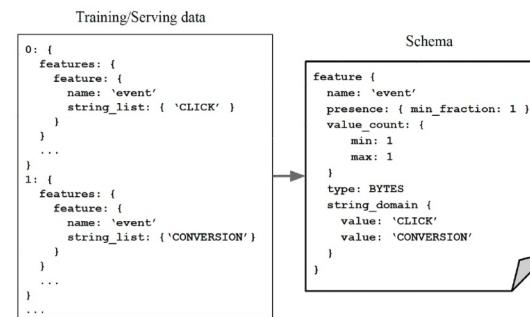
- Affecter la performance des modèles
- Plus difficile à expliquer

- Des solutions
 - [Feature scaling](#)
 - [Feature normalization](#)
 - Dimensionality reduction



Schema

- Coder les contraintes de données et la sémantique



Week 6 – What to focus on

What? Why? How?

Considérations de la sélection des modèles

- Facteurs à considerer lors de la sélection des modèles (trade-offs)
 - Performance
 - Interprétabilité
 - Robustesse
 - Taille du modèle
 - Temps de formation
 - Temps d'inference
- En pratique, nous essayons de nombreux modèles différents et sélectionnons les meilleurs

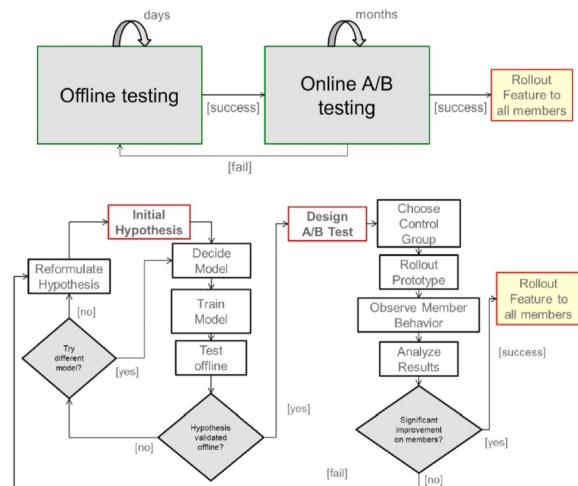
Choisir les mesures d'évaluation appropriés

- Évaluer l'objectif de chaque mesure
 - Sélectionnez celui qui est approprié pour le domaine du système
- Mesures de classification
 - Accuracy
 - Precision
 - Recall
 - F1-score (harmonic mean between precision and recall)
 - Precision-Recall curves
 - AUC ROC

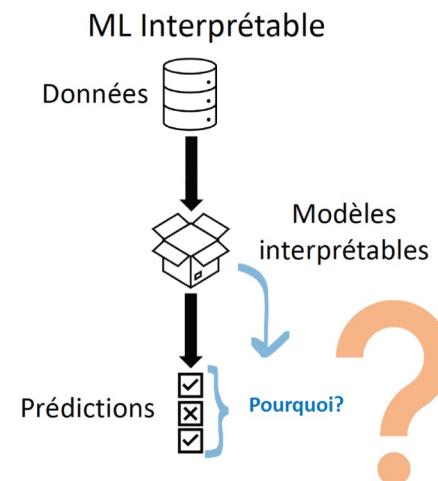
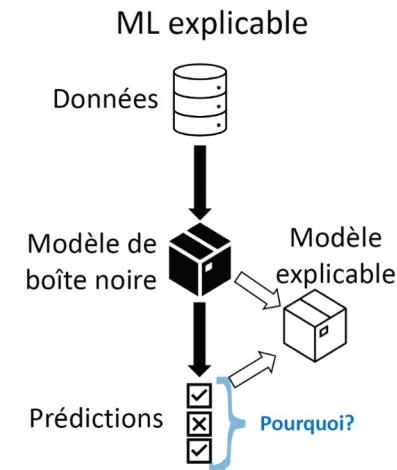
Week 6 – What to focus on

What? Why? How?

Offline testing -> A/B testing



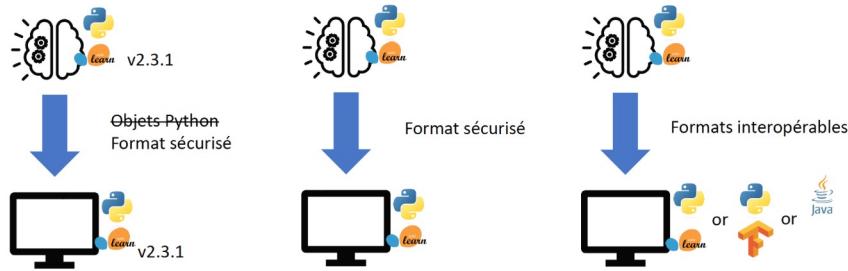
Explicable vs Interpretable modèles



Week 9 – What to focus on

Considérations sur les stratégies de persistance

- Quel stratégie utiliser ?



Structure du papier

- Les modèles complexes érodent les frontières.
- Dépendances de données
- Boucles de retroaction (feedback loops)
- Antimodèles de système ML
- Dette de configuration
- Changements dans le monde externe
- Autres domaines de dette technique.

Technical Problems!

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley, gholt, dgg, edavydov, toddphillips}@google.com
Google, Inc.

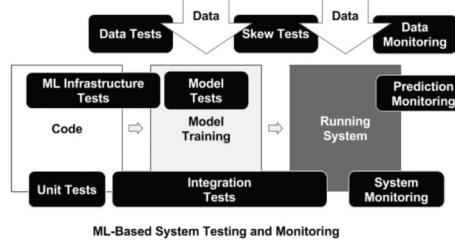
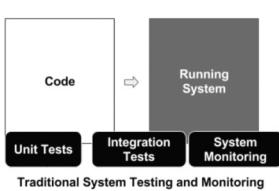
Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner, vchaudhary, mwyong, jcrespo, dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

Week 9 – What to focus on

La motivation



Qu'est-ce qui devrait être testé?
Combien est suffisant?

Catégories de tests

- Tests pour les fonctionnalités et les données
- Tests pour le développement des modèles
- Tests pour l'infrastructure ML
- Tests de surveillance pour les systèmes ML

The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction

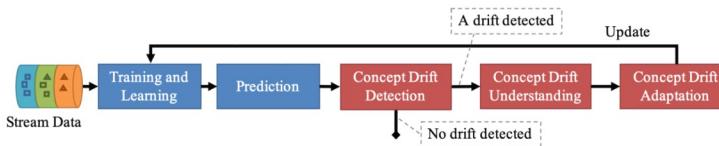
Eric Breck, Shangqin Cai, Eric Nielsen, Michael Salib, D. Sculley
Google, Inc.
ebreck, cais, nielsene, msalib, dsculley@google.com

Abstract—Creating reliable, production-level machine learning systems brings on a host of concerns not found in small toy examples or even large off-the-shelf research experiments. Testing and monitoring are key considerations for ensuring the production-readiness of ML systems and for reducing technical debt of ML systems. But it can be difficult to formulate specific tests, given that the actual prediction behavior of any given model is difficult to specify *a priori*. In this paper, we present 28 specific tests and monitoring needs, drawn from experience with a wide range of production ML systems, to help quickly identify what precision can easily be achieved to improve production readiness and pay down ML technical debt.

How to read the tests: Each test is written as an assertion; our recommendation is to test that the assertion is true, the more frequently the better, and to fix the system if the assertion is not true.

Week 10 – What to focus on

Gestion de la dérive de concept



L'article se concentre sur trois tâches principales liées à la dérive

- Identifier la dérive de concept
 - Se produit-elle ?
- Comprendre la dérive de concept
 - Pourquoi se produit-elle ?
- Réagir à la dérive de concept
 - Comment garantir la qualité de notre système d'IA au fil du temps ?

Réagir à la dérive : Apprentissage en binôme

- Paired learners (Apprentissage en binôme)
 - Apprenant stable : entraîné sur de vieilles données
 - Apprenant réactif : entraîné sur de nouvelles données.
 - Changement lorsque l'erreur de l'apprenant stable > apprenant réactif."

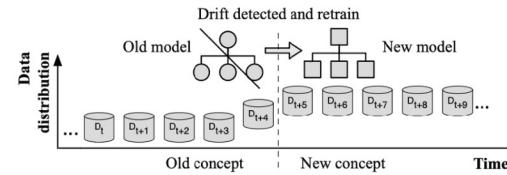


Fig. 13. A new model is trained with latest data to replace the old model when a concept drift is detected.

Week 10 – What to focus on

MLOps = DevOps + AI systems

MLOps est une **culture** et **pratique** d'ingénierie ML visant à unifier le **développement de systèmes ML** (Dev) et l'**exploitation de systèmes ML** (Ops).

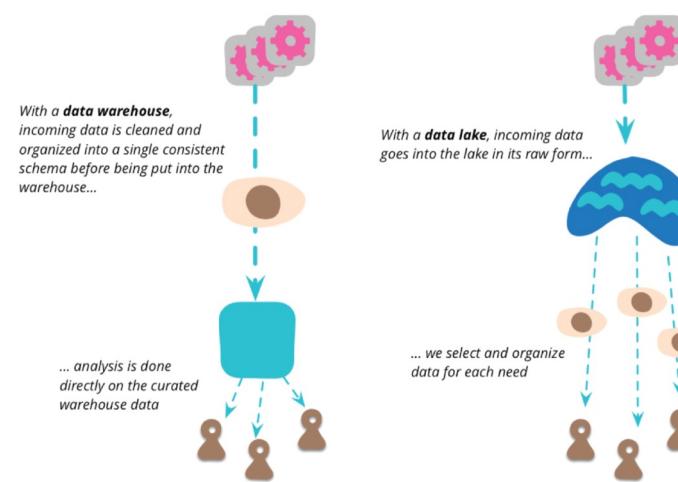
- Promouvoir l'automatisation et la surveillance à toutes les étapes du système ML

Les pratiques de MLOps

- Intégration Continue (CI)
 - Test et validation du code et des composants
 - **Test et validation des données + des schémas de données + des modèles**
- Déploiement Continu (CD)
 - Déploiement de packages et de services
 - **Déploiement du pipeline de maintenance (re-entraînement du modèle)**
- Formation continue (CT)
 - Réentraînement automatique des modèles

Propre aux systèmes ML !

Données découvrables et accessibles

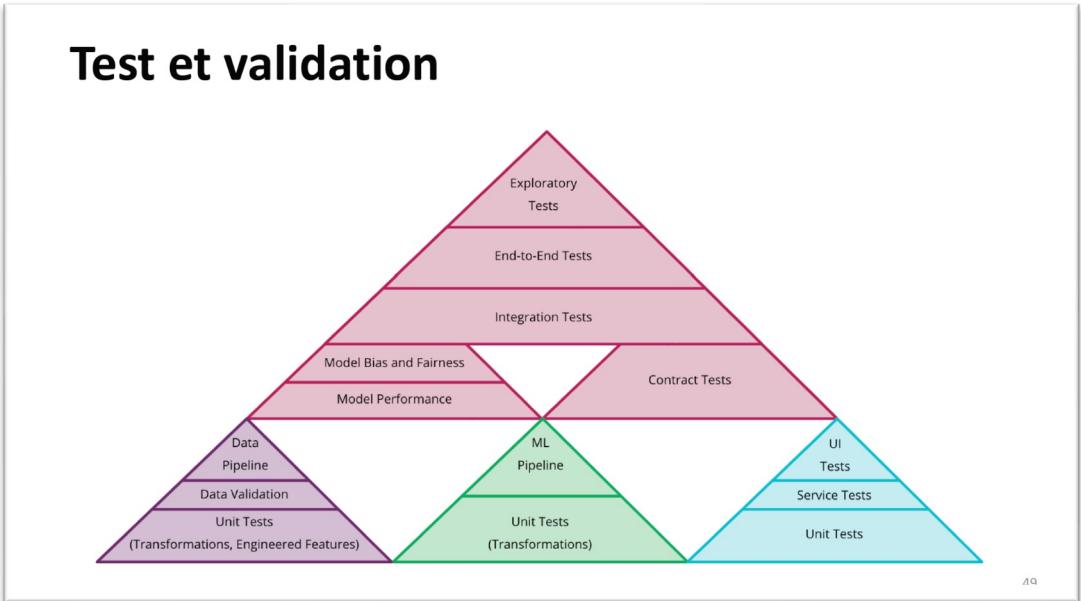


Week 10 – What **not** to focus on

Identifier la dérive de concept (cont.)

Category	Algorithms	Stage 1	Stage 2	Stage 3	Stage 4
Error rate-based	DDM [20]	Landmark	Learner	Online error rate	Distribution estimation
	EDDM [26]	Landmark	Learner	Online error rate	Distribution estimation
	FW-DDM [5]	Landmark	Learner	Online error rate	Distribution estimation
	DFML [27]	Landmark	Learner	Online error rate	Distribution estimation
	STEPD [30]	Predefined w_{hist}, w_{new}	Learner	Error rate difference	Distribution estimation
	ADWIN [31]	Auto cut w_{hist}, w_{new}	Learner	Error rate difference	Hoeffding's Bound
	ECDD [29]	Landmark	Learner	Online error rate	EWMA Chart
	HDMM [23]	Landmark	Learner	Online error rate	Hoeffding's Bound
	LLDD [25]	Landmark, or sliding w_{hist}, w_{new}	Decision trees	Tree node error rate	Hoeffding's Bound
Data distribution-based	kdgTree [22]	Fixed w_{hist} , Sliding w_{new}	kdgTree	KL divergence	Bootstrapping
	CM [2], [3]	Fixed w_{hist} , Sliding w_{new}	Competence model	Competence distance	Permutation test
	RD [37]	Fixed w_{hist} , Sliding w_{new}	KS structure	Relativized Discrepancy	VC-Dimension
	SCD [38]	Fixed w_{hist} , Sliding w_{new}	kernel density estimator	log-likelihood	Distribution estimation
	EDE [40]	Fixed w_{hist} , Sliding w_{new}	Nearest neighbor	Density scale	Permutation test
	SyncStream [36]	Fixed w_{hist} , Sliding w_{new}	PCA	P-Tree	Wilcoxon test
	PCA-CD [39]	Fixed w_{hist} , Sliding w_{new}	PCA	Change-Score	Page-Hinkley test
	LSDD-CDT [21]	Fixed w_{hist} , Sliding w_{new}	Learner	Relative-difference	Distribution estimation
	LSDD-INC [41]	Fixed w_{hist} , Sliding w_{new}	Learner	Relative difference	Distribution estimation
	LDI-DSDA [4]	Fixed w_{hist} , Sliding w_{new}	k-nearest neighbor	Local drift degree	Distribution estimation
Multiple Hypothesis Tests	JIT [19]	Landmark	Selected features	4 configurations	Distribution estimation
	LFR [46]	Landmark	Learner	TP, TN, FP, FN	Distribution estimation
	Three-layer [47]	Sliding both w_{hist}, w_{new}	Learner	$P(y), P(X), P(X y)$	Distribution estimation
	e-Detector [48]	depends on base detector	depends	depends	depends
	DDE [49]	depends on base detector	depends	depends	depends
	TSMSD-EWMA [52]	Landmark	Learner	Online error rate	EWMA Chart
	HCDTs [50]	Landmark	Depending on layers	Depending on layers	Depending on layer
	HLFR [51]	Landmark	Learner	TP, TN, FP, FN	Distribution estimation
	HHT-CU [53]	Landmark	Learner	Classification uncertainty	Layer-I Hoeffding's Bound, Layer-II Permutation Test
	HHT-AG [53]	Fixed w_{hist} , Sliding w_{new}	N/A	KS statistic on each attribute	Layer-I KS test, Layer-II 2D KS test

Test et validation

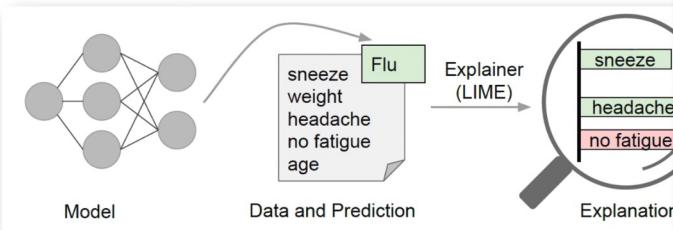


Week 11 – What to focus on

The case for explainability

The case for explainability

- We need to understand AI predictions to make them better



Benefits for Explaining Models and Predictions

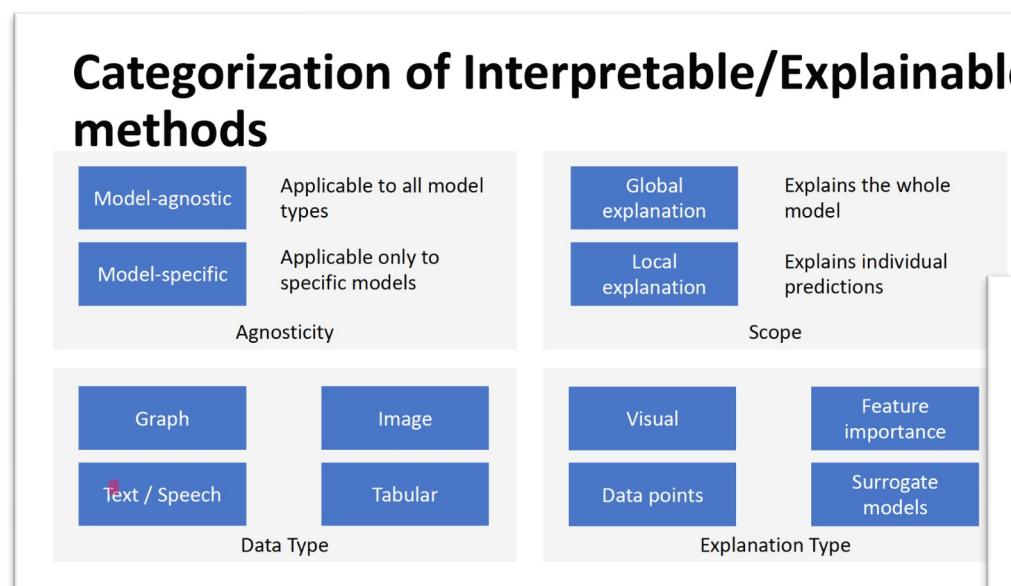
For engineers

- Better understand the **model**
 - Validate model's decision with domain experts
 - Debug model's behavior
 - Assess the quality of the model beyond just accuracy metrics
 - Trust the model will behave reasonably on real-world data
 - Learn from the model

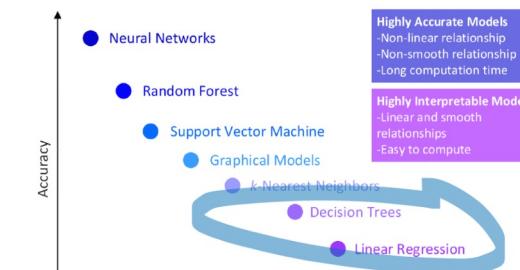
For end-users

- Trust an AI system's **decision** (prediction)
 - Comprehend the contributing factors
 - Have enough information to act
 - Treats the user fairly
 - Uses the provided information responsibly
- Trust the AI system is well developed (model)

Week 11 – What to focus on



Trade-off between interpretability x accuracy



Permutation Feature Importance

- Model agnostic
- Global explanation
- Core idea:
 - Shuffle feature values
 - Verify the impact on the model's performance
 - Rank features
 - Critical features cause more impact in the model (top)
 - Less important features do not affect the model's performance as much (bottom)
- Implemented at Scikitlearn

Week 11

- What to **focus** on

What is the categories of LIME?

Model-agnostic	Applicable to all model types
Model-specific	Applicable only to specific models

Global explanation	Explains the whole model
Local explanation	Explains individual predictions

How does LIME works?

This local problem is easier to explain

One possible local model:
Classify **Stroke** if Feature 2 is smaller than x
Otherwise, **No Stroke**

Source: Explainable AI explained! | #3 LIME

- What **not** to focus on

$\xi(x) = \operatorname{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$

• Solution?

```
Algorithm 2 Submodular pick (SP) algorithm
Requires: Instances  $X$ , Budget  $B$ 
for all  $x_i \in X$  do
     $\mathcal{W}_i \leftarrow \text{explain}(x_i, x'_i)$                                 ▷ Using
end for
for  $j \in \{1, \dots, d'\}$  do
     $I_j \leftarrow \sqrt{\sum_{i=1}^n |\mathcal{W}_{ij}|}$                                 ▷ Compute feature importances
end for
 $V \leftarrow \{\}$ 
while  $|V| < B$  do                                              ▷ Greedy optimization of Eq (4)
     $V \leftarrow V \cup \operatorname{argmax}_i c(V \cup \{i\}, \mathcal{W}, I)$ 
end while
return  $V$ 
```

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : \mathcal{W}_{ij} > 0]} I_j \quad (3)$$
$$\text{Pick}(\mathcal{W}, I) = \operatorname{argmax}_{V, |V| \leq B} c(V, \mathcal{W}, I)$$

Parting Thoughts

- Focus on understanding
 - The definitions and principles
 - The problems related to engineering AI systems
 - The solutions that attempt to mitigate those problems
 - When they can be applied?
 - What is the trade-off of this solution?
 - What is the expected benefit of this solution?0
- No mathematical formula or code will be required

Last mile – Course Project

- Today is the last day you can get feedback on the project
 - You can still send questions via email
- The final presentation will take place in two weeks April 20
 - Each group should prepare a **demo** of their system
 - Each student should be able to test other group's demos
 - Presentation of the project implementation (oral)
 - You can prepare some slides
 - Submission of the final report until: **April 23**
 - Via email