Modeling
Interdisciplinary
Data Science
Applied
Statistics

Artificial Intelligence & Statistics

Navigating potentials, pitfalls and use

LSS & STATEC – Economic Seminar

Thursday, December 21st 2023

Florian FELICE

Senior Data Scientist, AWS Generative AI Innovation Center

Doctoral Research, University of Luxembourg

UNIVERSITÉ DU LUXEMBOURG

aws

Modeling
Interdisciplinary
Data Science
Applied
Statistics

Artificial Intelligence & Statist

Navigating potentials, pitfalls and use

LSS & STATEC – Economic Seminar

Thursday, December 21st 2023

Florian FELICE

Senior Data Scientist, AWS Generative AI

Doctoral Research, University of Luxembo

UNIVERSITÉ DU LUXEMBOURG

Modeling
Interdisciplinary
Data Science
Applied
Statistics

Artificial Intelligence & Statistics

Navigating potentials, pitfalls and use

LSS & STATEC – Economic Seminar

Thursday, December 21st 2023

Florian FELICE

Senior Data Scientist, AWS Generative AI Innovation Center

Doctoral Research, University of Luxembourg

UNIVERSITÉ DU LUXEMBOURG

aws

Agenda

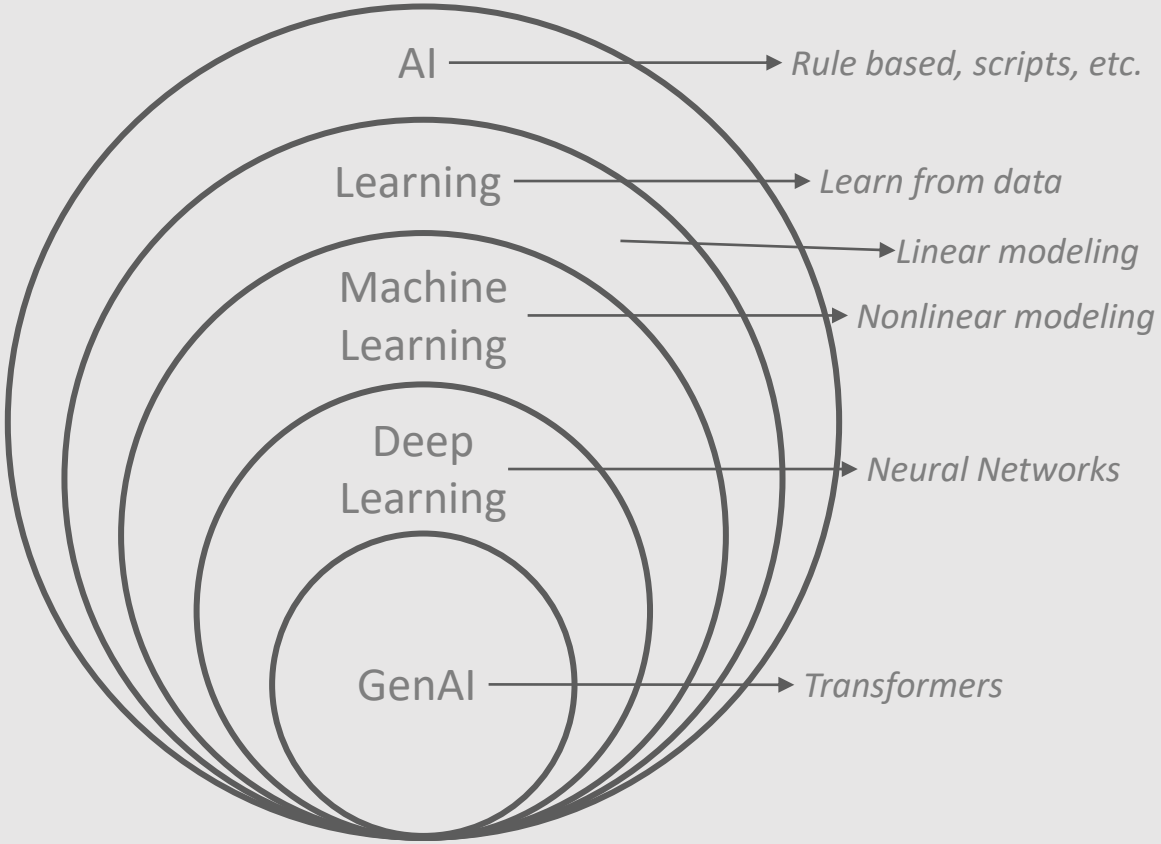
- Introduction
- Generative AI capabilities
- Arising concerns
- Responsible AI
- Conclusion & discussion

Introduction

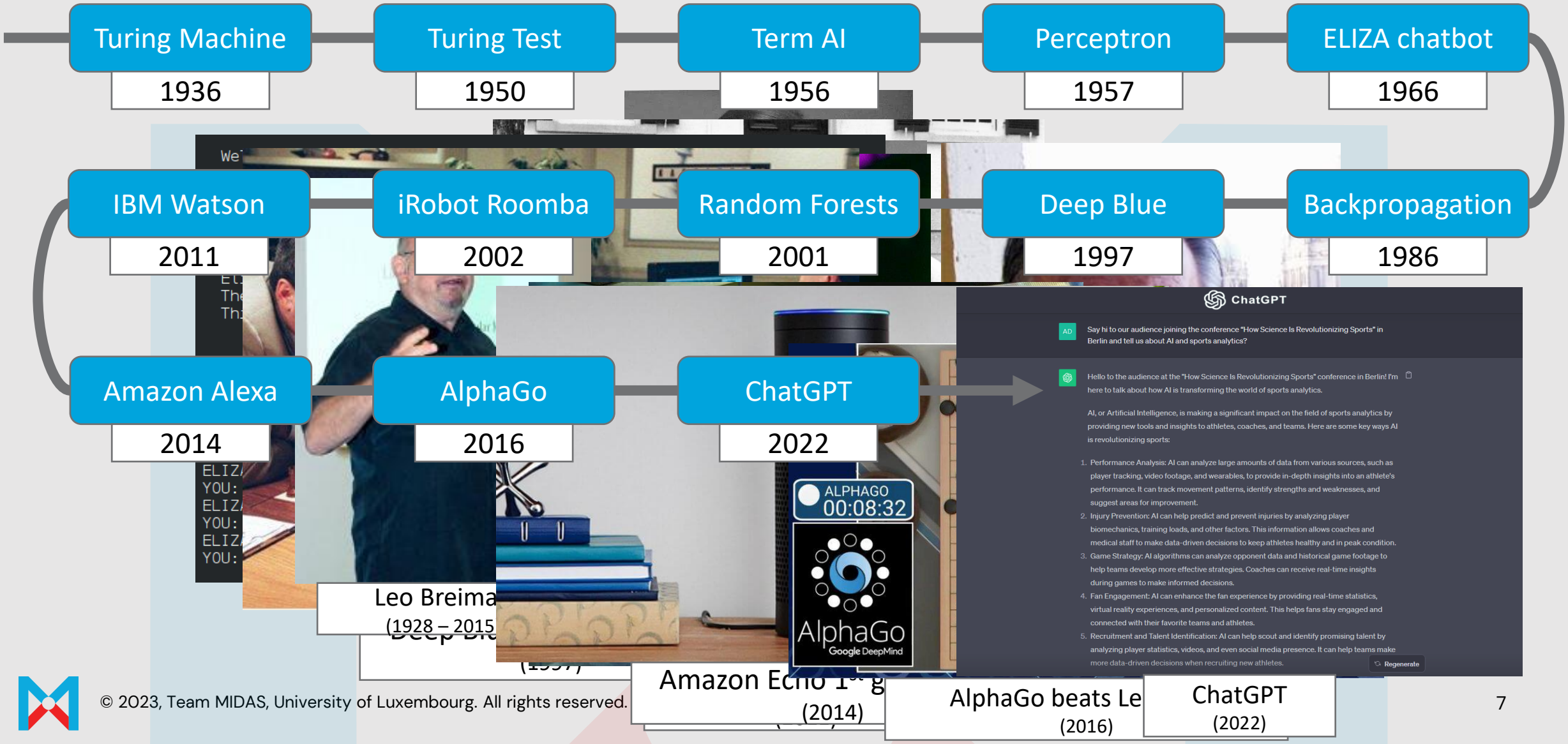
Artificial Intelligence

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Felice, Florian, et al. "Statistically Enhanced Learning: a feature engineering framework to boost (any) learning algorithms." *arXiv preprint arXiv:2306.17006* (2023).



The history of AI

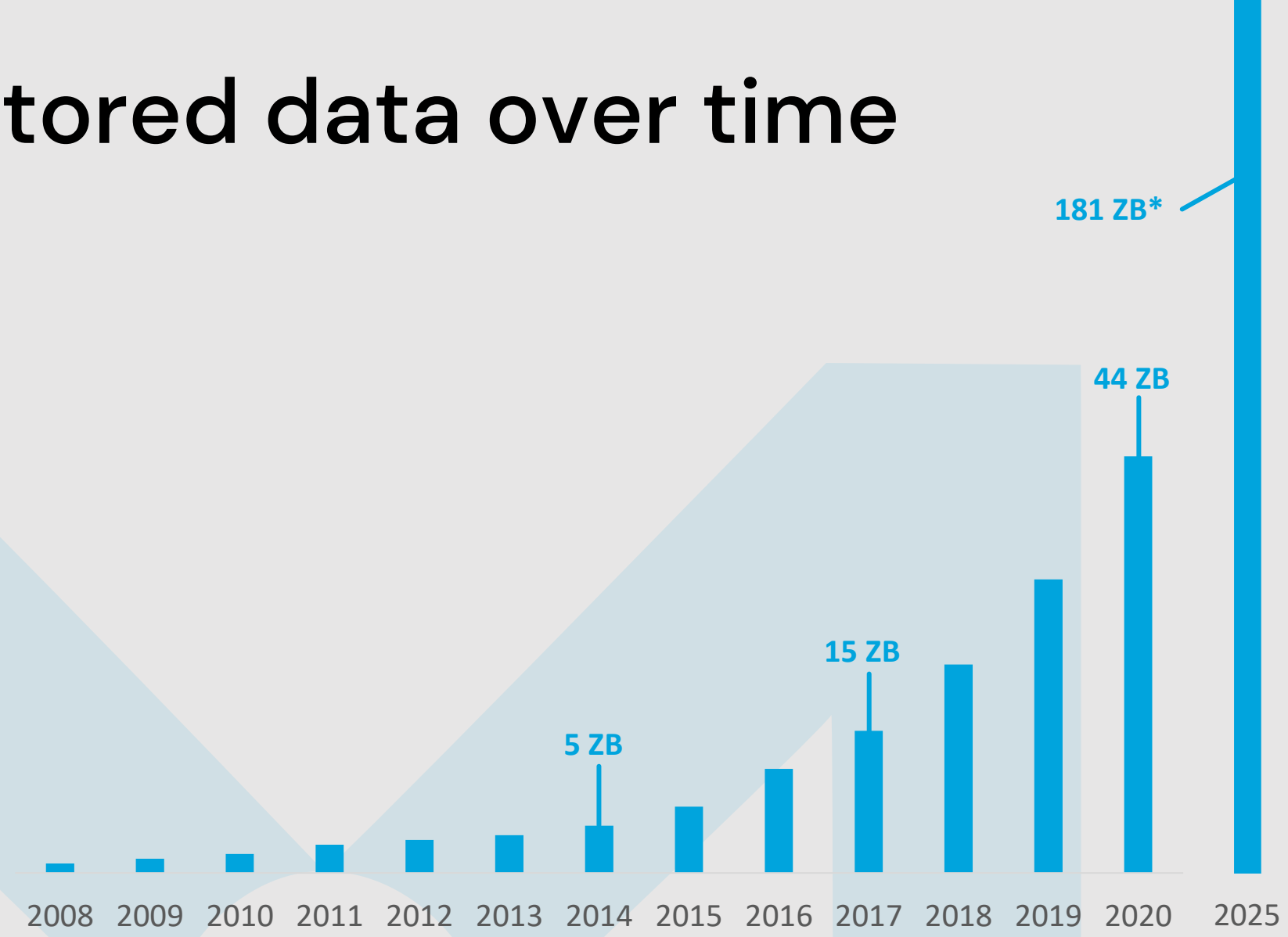


Once upon a time, there was a great [illegible]."

Evolution of stored data over time

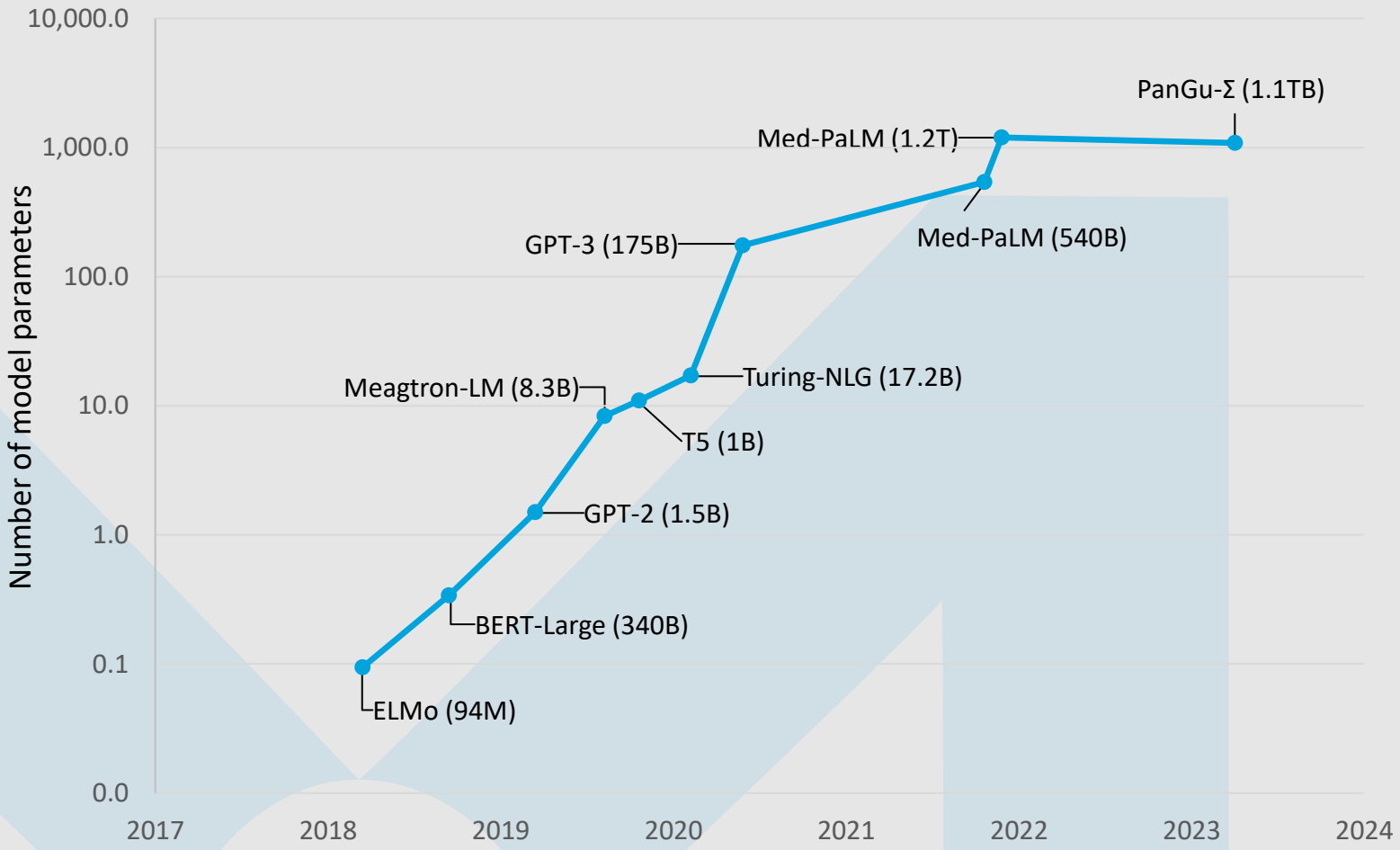
1 ZB = 10^9 TB

Hammad, Khalid Adam Ismail, et al. "Big data analysis and storage." *International conference on operations excellence and service engineering*. 2015.

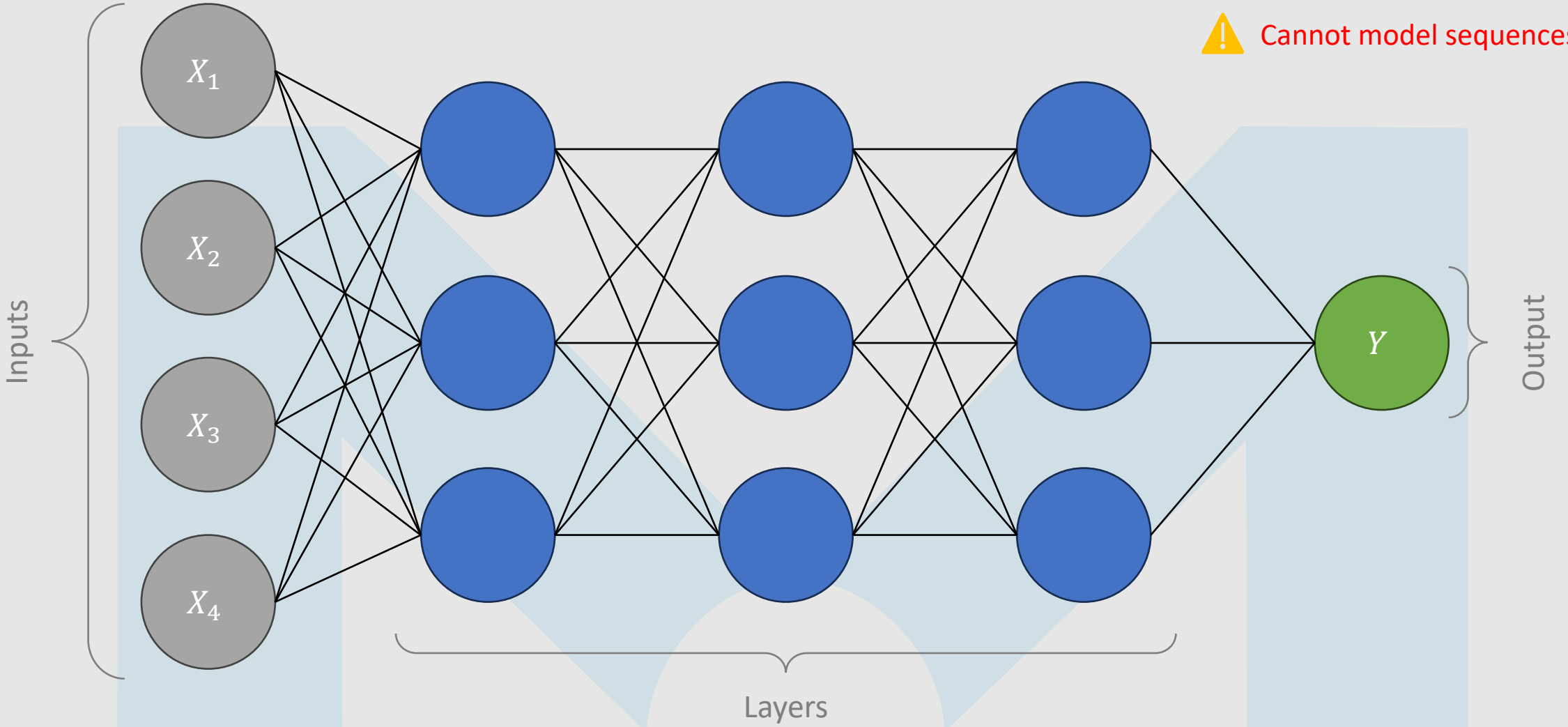


Evolution of model size

Simon, Julien. "Large language models: A new Moore's law." *Hugging Face blog* (2021).



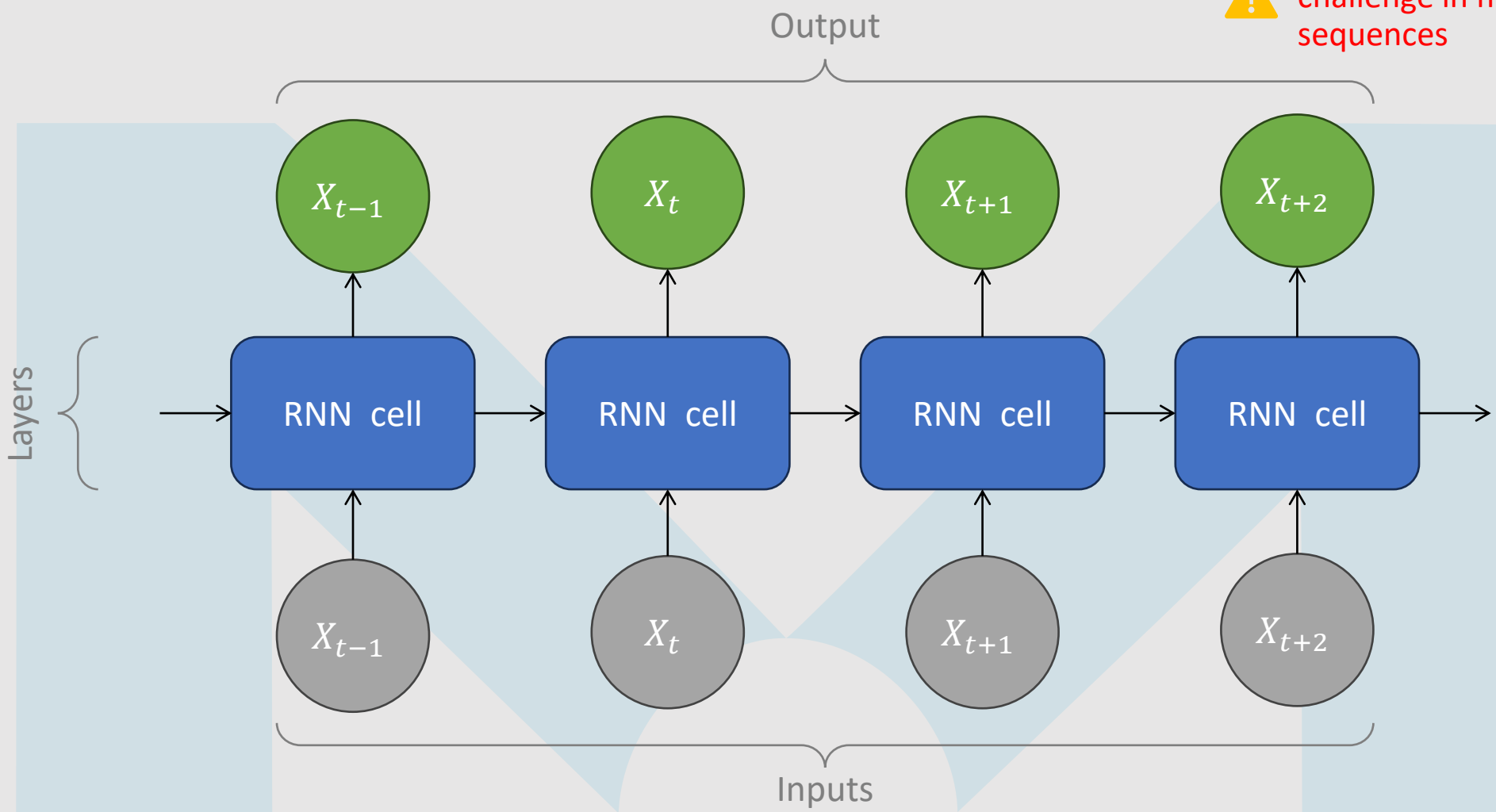
Multi-Layered Perceptron (MLP)



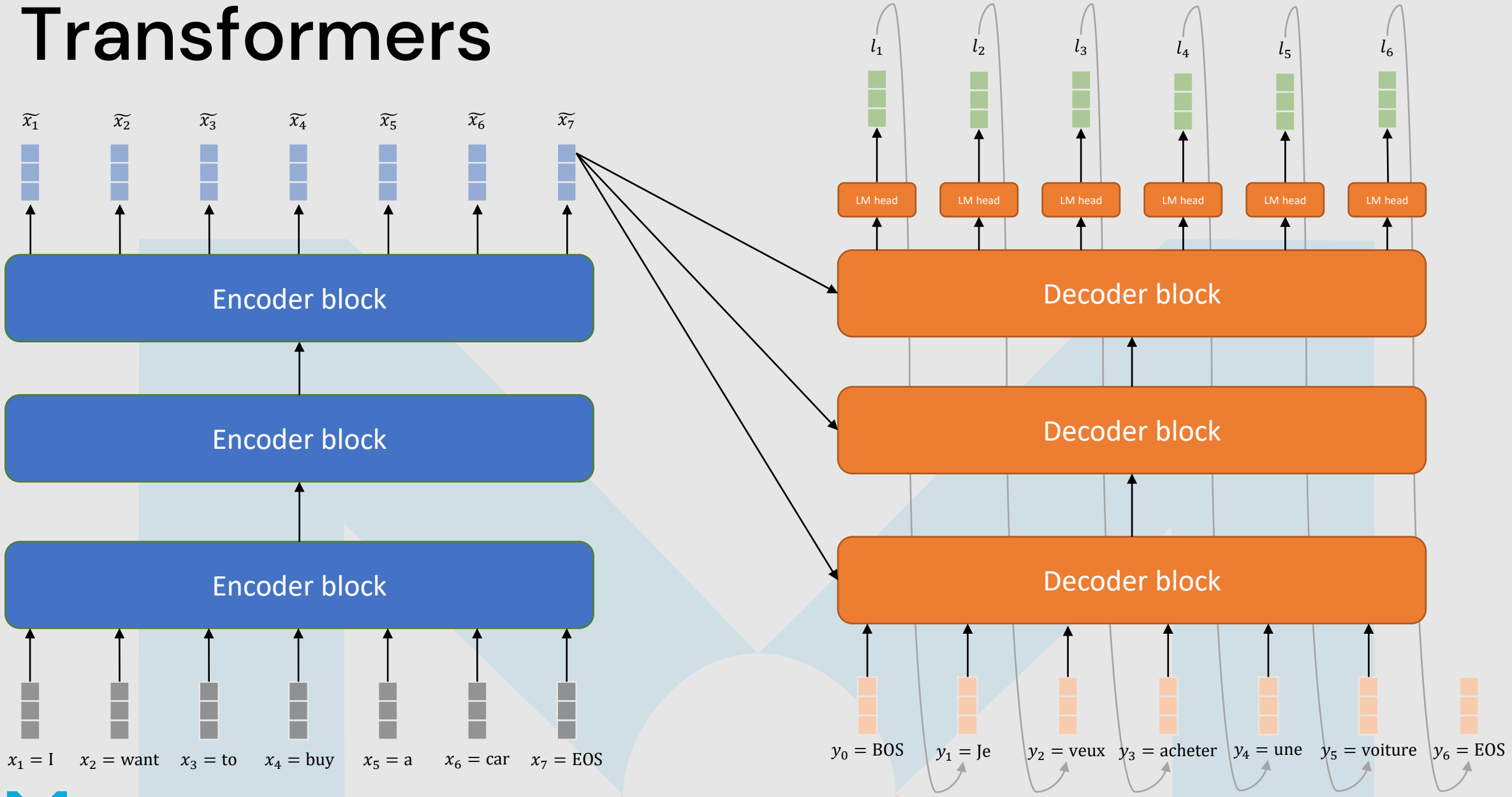
⚠ Cannot model sequences

Recurrent Neural Nets (RNN)

Gradient vanishing problem: challenge in modeling long sequences



Transformers



Generative AI capabilities

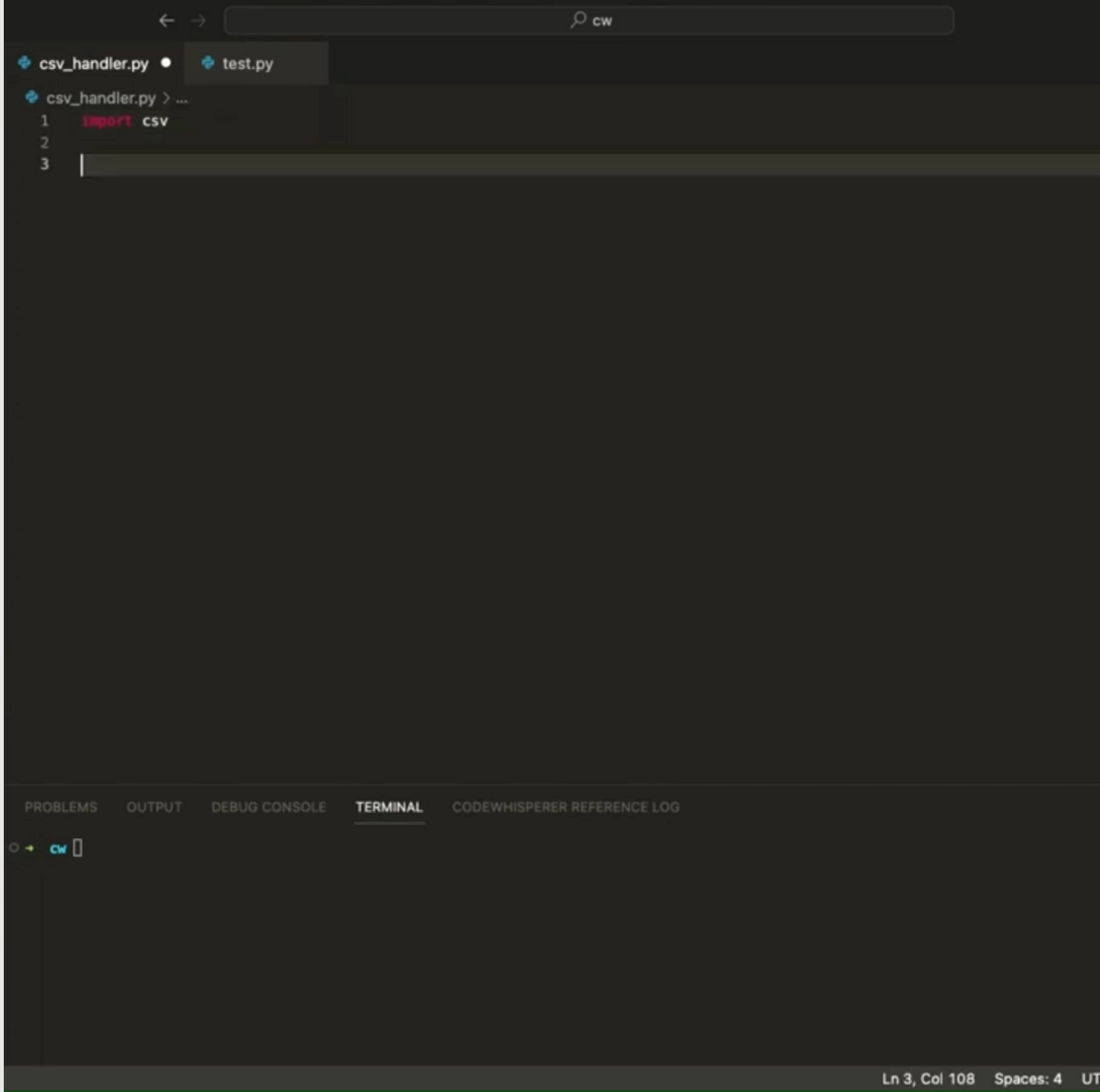
Generative AI for images

Prompt: *“Create a painting of a skateboarding cat in the style of Andy Warhol.”*

Generated by Stable Diffusion XL via Amazon Bedrock

Generative AI for coding

- Support developers be more productive
- Help explain code
- Help debug



Generative AI for document summary

- Retrieval Augmented Generation

Generative AI capabilities

Sentiment analysis

Language understanding

Text classification

Image editing

Q&A

Text to speech

Jobs destruction

- Speculations on jobs being lost because of GenAI technologies
- Can be a productive helper tool (e.g. AWS CodeWhisperer)
- Can expect a similar revolution as with online search engines
- Education on how to use such tools

Arising concerns

Plagiarism, cheating & Intellectual Property

- Models are trained on large amount of data
- Can reproduce the training data (potentially including confidential or copy-righted content)
- Challenge in education systems of students using such tools
- Ethical, legal and educational considerations

Toxicity

- Offensive/inappropriate content
- Discrimination against groups/individuals
- Difficult to draw boundary between restricting toxic content and censorship

Hallucinations

- Outcome that sounds/looks reasonable but verifiably incorrect
- Over-creativity cannot be linked to online and verified content
- Root cause: model is predict next word from some distributions


You have
NO CHOICE
NOT
CONVINCED??

Regulations around AI

- Level 1: Social Principles (2018 –)
  - OECD Principles on AI
- Level 2: National Frameworks and Social Guidelines (2020 –)
  - NIST AI RMF (USA)
  - GDPR, Regulatory Framework proposal on AI (EU)
  - National AI Strategy (UK)
- Level 3: Technical Guidelines and Standardization (2021 –)
  - ISO/IEC JTC 1/SC 42 AI Standards
  - IEEE AI Standards

Responsible AI

An overview

“AI that is innovative and trustworthy and respects human rights and democratic values”

Source: OECD

Some Responsible AI dimensions

Fairness

How a system impacts different subpopulations of users (e.g. by gender, ethnicity)

Explainability

Mechanisms to understand and evaluate the outputs of an AI system

Robustness

Mechanisms to ensure an AI system operates reliably

Privacy and Security

Data used in accordance with privacy considerations, and protected from theft and exposure

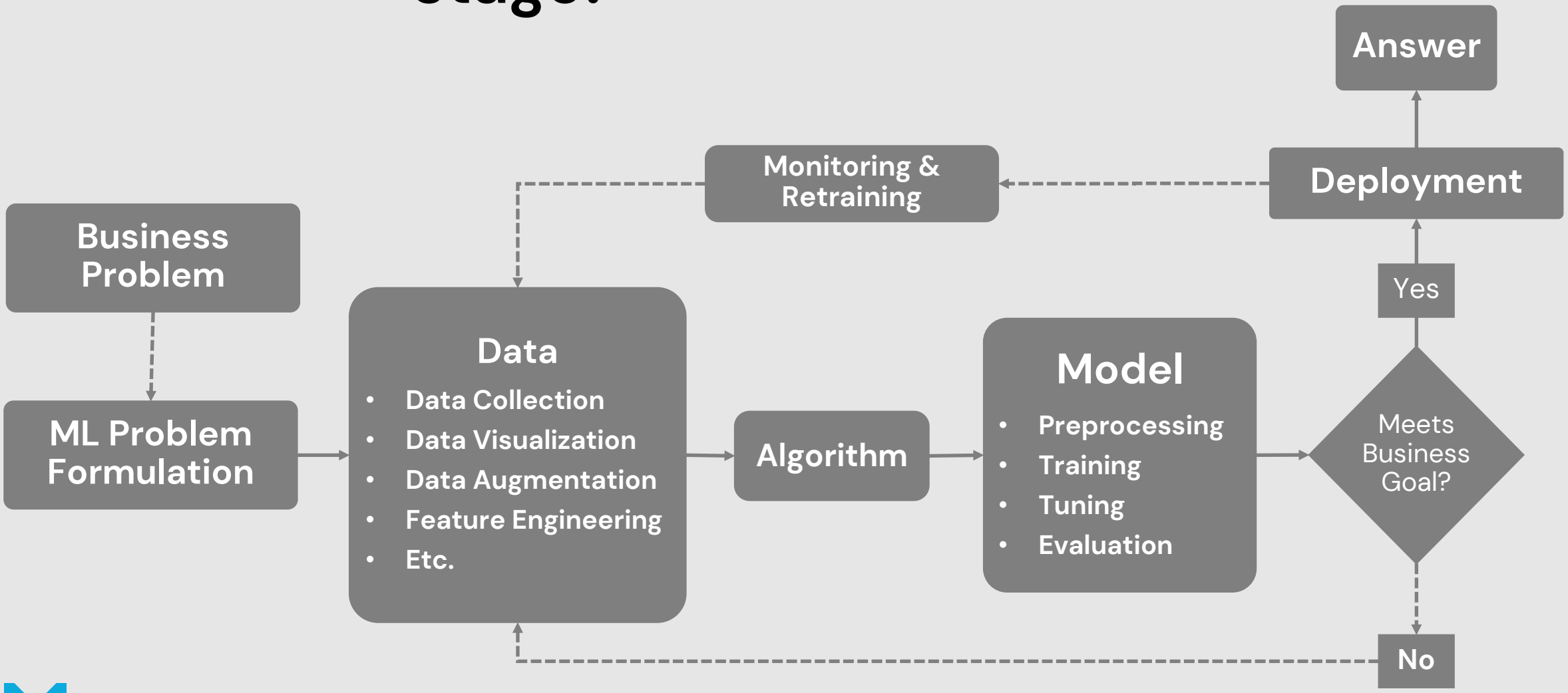
Governance

Processes to define, implement and enforce responsible AI practices within an organization

Transparency

Communicate information about an AI system so stakeholders can make informed choices about their use of the system

ML lifecycle… must have RAI components at any stage!



A responsible use of GenAI

Fairness in Generative AI

Once upon a time, a doctor called

John (0.7)
Jane (0.2)

Dr. Hanson studied the patient's chart carefully, and

he... (0.8)
she... (0.1)

Fairness in Generative AI

Once upon a time, a doctor called
- John (0.4)
- Jane (0.35)

Dr. Hanson studied the patient's chart carefully, and
- he... (0.4)
- she... (0.6)

⚠️ Not scalable!

Fairness in Generative AI

- Careful preparation of tranining data
  - Models will reflect the tone of the training data
- Human annotated data
  - Training methods such as RLHF or model alignment
- Human tested model
  - Check how model behaves for specific prompts

Fairness in Generative AI

- Limit the toxicity of a model by:
- Working on training step
  - E.g. Reinforcement Learning with Human Feedback (RLHF)
- Giving instructions to the model
  - Prompt engineering
- Training guardrail models
  - Identify and filter out undesired content
- Understanding the behavior of your model
  - Explainability

Explainability in Generative AI

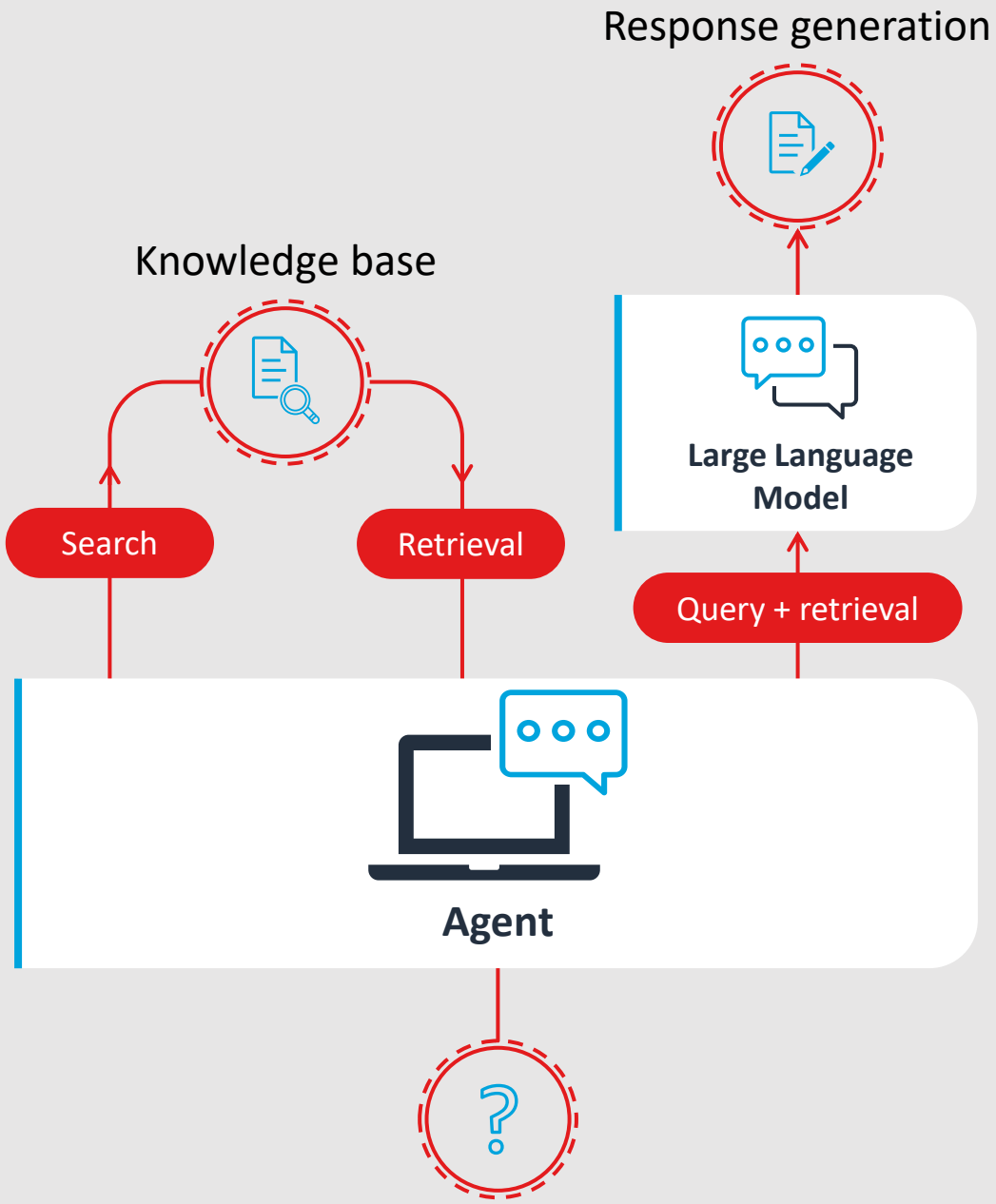
- Active field of research
- Requires definition of explanations for text generation
- Some literature available for classification models
- Chain-of-thoughts prompting
- Available AWS service: SageMaker Clarify

Zafar, Muhammad Bilal, et al. "On the lack of robust interpretability of neural text classifiers." *arXiv preprint arXiv:2106.04631* (2021).

Kokalj, Enja, et al. "BERT meets shapley: Extending SHAP explanations to transformer-based classifiers." *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. 2021.

Robustness in Generative AI

- Human-in-the-Loop
  Expert feedback during development
  Collection of user feedback in production
- Retrieval Augmented Generation



Privacy and security for Generative AI

- Data privacy
  - Careful preparation of input data
  - Remove personal/confidential information
- Data security
  - E.g. data encryption on S3
- Service security
  - Model endpoint on VPC
  - E.g. restrict access to internet, isolated subnet

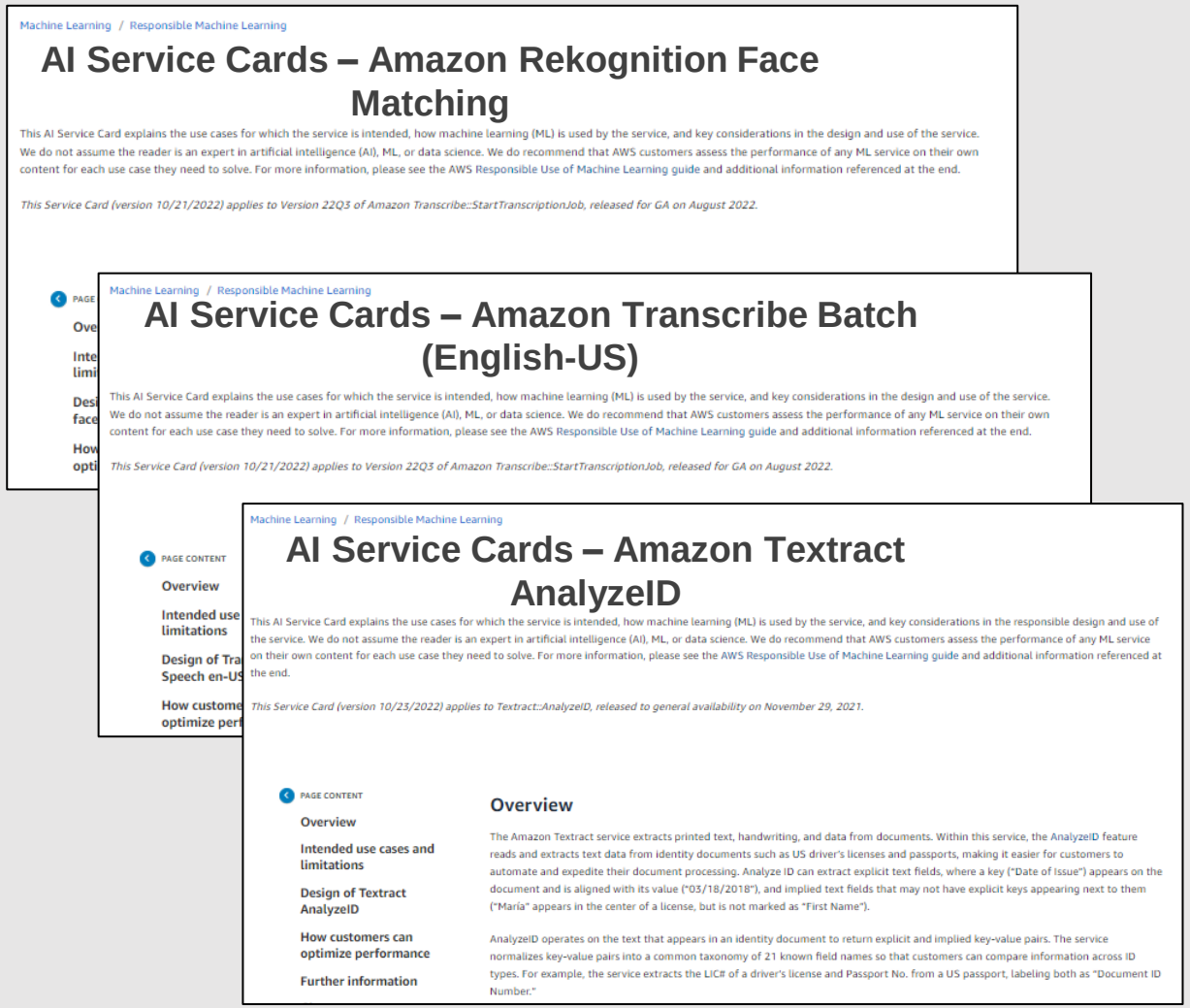
Governance for Generative AI

- Accountability and management of the service/tool
- Ensure RAI practices are carried out among stakeholders
- Regulatory frameworks (e.g. NIST – US) focus on governance
  - They insist on documentation
  - Accountability and permissions at each step of the model lifecycle

Transparency for Generative AI

New transparency resource to advance responsible AI

- Documents the intended use cases and fairness considerations of our AWS AI services
- Reflects our comprehensive development process
- Three new AI Service Cards published



Machine Learning / Responsible Machine Learning

AI Service Cards – Amazon Rekognition Face Matching

This AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the design and use of the service. We do not assume the reader is an expert in artificial intelligence (AI), ML, or data science. We do recommend that AWS customers assess the performance of any ML service on their own content for each use case they need to solve. For more information, please see the AWS Responsible Use of Machine Learning guide and additional information referenced at the end.

This Service Card (version 10/21/2022) applies to Version 22Q3 of Amazon Transcribe::StartTranscriptionJob, released for GA on August 2022.

Machine Learning / Responsible Machine Learning

AI Service Cards – Amazon Transcribe Batch (English-US)

This AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the design and use of the service. We do not assume the reader is an expert in artificial intelligence (AI), ML, or data science. We do recommend that AWS customers assess the performance of any ML service on their own content for each use case they need to solve. For more information, please see the AWS Responsible Use of Machine Learning guide and additional information referenced at the end.

This Service Card (version 10/21/2022) applies to Version 22Q3 of Amazon Transcribe::StartTranscriptionJob, released for GA on August 2022.

PAGE CONTENT

Overview

Intended use limitations

Design of Tra Speech en-US

How custome optimize perf

Machine Learning / Responsible Machine Learning

AI Service Cards – Amazon Textract AnalyzeID

This AI Service Card explains the use cases for which the service is intended, how machine learning (ML) is used by the service, and key considerations in the responsible design and use of the service. We do not assume the reader is an expert in artificial intelligence (AI), ML, or data science. We do recommend that AWS customers assess the performance of any ML service on their own content for each use case they need to solve. For more information, please see the AWS Responsible Use of Machine Learning guide and additional information referenced at the end.

This Service Card (version 10/23/2022) applies to Textract::AnalyzeID, released to general availability on November 29, 2021.

PAGE CONTENT

Overview

Intended use cases and limitations

Design of Textract AnalyzeID

How customers can optimize performance

Further information

Overview

The Amazon Textract service extracts printed text, handwriting, and data from documents. Within this service, the AnalyzeID feature reads and extracts text data from identity documents such as US driver's licenses and passports, making it easier for customers to automate and expedite their document processing. Analyze ID can extract explicit text fields, where a key ("Date of Issue") appears on the document and is aligned with its value ("03/18/2018"), and implied text fields that may not have explicit keys appearing next to them ("María" appears in the center of a license, but is not marked as "First Name").

AnalyzeID operates on the text that appears in an identity document to return explicit and implied key-value pairs. The service normalizes key-value pairs into a common taxonomy of 21 known field names so that customers can compare information across ID types. For example, the service extracts the LIC# of a driver's license and Passport No. from a US passport, labeling both as "Document ID Number."

Conclusion

Conclusion

- Generative AI is not all you need
  - It will not replace traditional ML: it will complement it
- Prepare your data carefully
- Responsible AI is everywhere
- Stay tuned! Research is in progress

- Read more on Amazon Science
  - Responsible AI in the generative era by Michael Kearns

Thank you!

TRY FOR FREE : AI STUDIOS
www.deepbrain.io

Thank you!