Exploring AI
The RSE way

Date : 10 May 2024 | Speaker : Ayon Roy | Venue : Online for GL Bajaj ITM

Event : FDP on Confluence of Deep Learning Generative AI and Data Science

Visit - https://ayon-roy.netlify.app

Responsible AI
Sustainable AI
Ethical AI

Visit - https://ayon-roy.netlify.app

Hello **World!**

I am **Ayon Roy**

Executive Data Scientist @ NielsenIQ

Z by HP Global Data Science Ambassador

Mentored/Judged **100+** Hackathons

Delivered **100+** Technical Talks

Brought **Kaggle Days Meetup** Community in India for the 1st time

If you haven't heard about me yet, you might have been living under the rocks. Wake up !!

Agenda

- Growth of AI
- New Challenges with AI
- Responsible & Ethical AI
- Sustainable AI
- What’s the Future

Why Data & AI matters in 2024 ?

To analyze, extract information from huge datasets which maybe beyond the ability of general tools to manage, process data.

Volume : Scale of Data

Variety : Different types of Data

Velocity : Speedy Ingestion of new Data

Veracity : Uncertainty in the Data

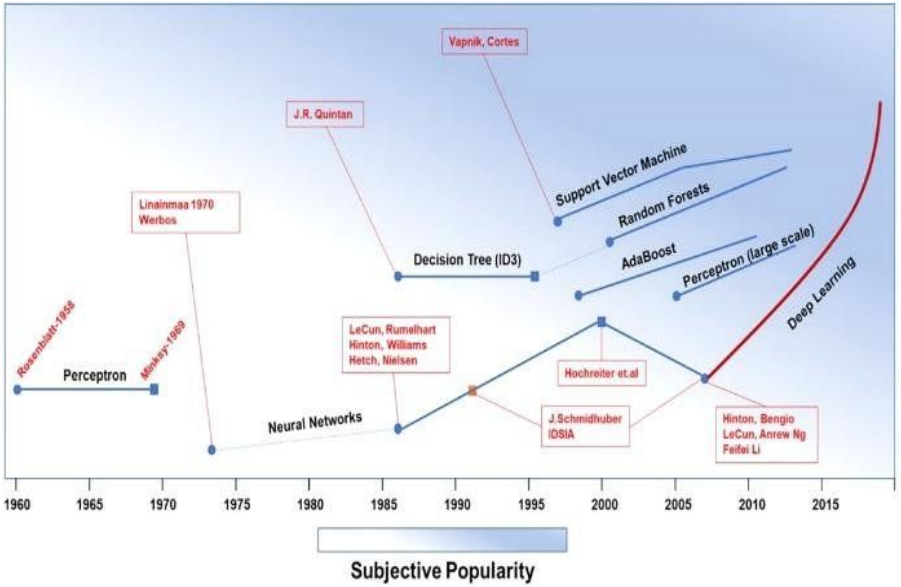
AI Systems

Artificial Intelligence Systems are projects which are undertaken with the long-term goal of simulating the human brain in real time, complete with artificial consciousness and artificial general intelligence.

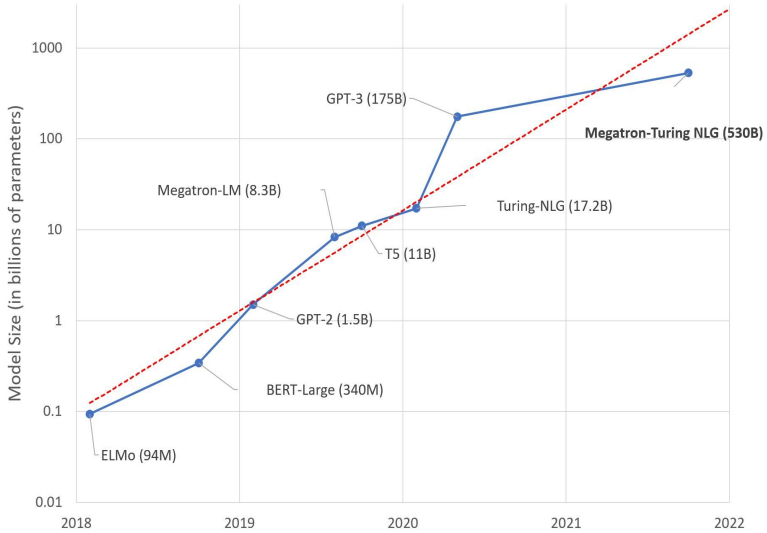
How do we simulate the human brain in real time & bring artificial consciousness ?

Data + Model + Compute

Evolution of AI



https://www.researchgate.net/publication/349864030_Review_of_machine_learning_and_deep_learning_application_in_mine_microseismic_event_classification



https://huggingface.co/blog/large-language-models

Understanding growth in AI's support systems

AI is driven by 3 primary factors - Data, Model (Algorithms) & Compute

While model development have taken the spotlight for a few years now & have transformed the way AI advancements are happening with significant improvements in efficiency.

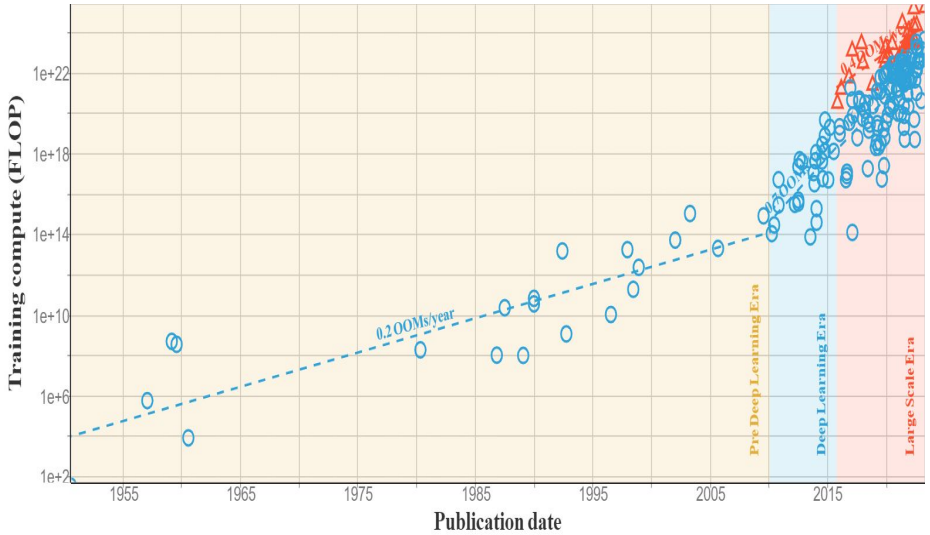
Data has been becoming increasingly available, particularly with the advent of **"big data"** in recent years.

At the same time, progress in computing hardware has been rapid, with increasingly powerful and specialised AI hardware.

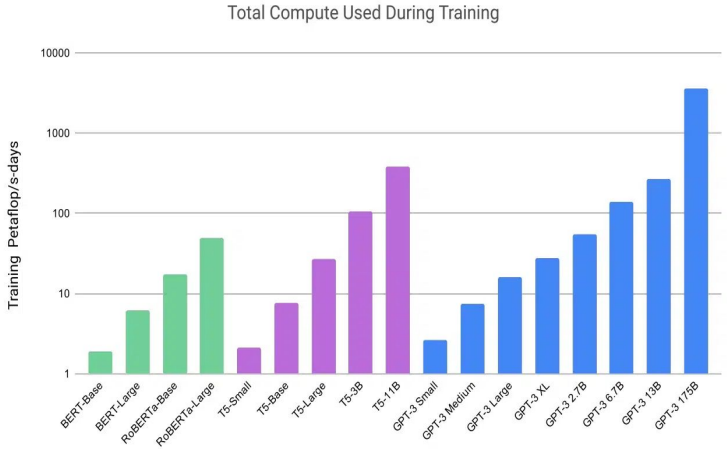
What is less obvious is the *relative* importance of these factors, and what this implies for the future of AI. A paper on "Scaling Laws for Neural Language Models" Kaplan *et al.* (2020) studied these developments through the lens of **scaling laws**, identifying three key variables:

- Number of parameters of a machine learning model
- Training dataset size
- Compute required for the final training run of a machine learning model (henceforth referred to as **training compute**)

Growth in Computing power requirement



https://epochai.org/blog/compute-trends



https://blogs.nvidia.com/blog/2022/10/10/llms-ai-horizon/

Visit - https://ayon-roy.netlify.app

Growth in Data usage

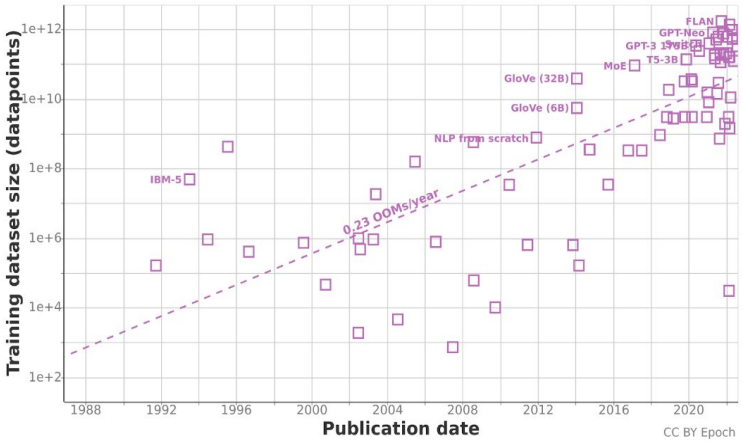


Figure 3: Evolution of language datasets

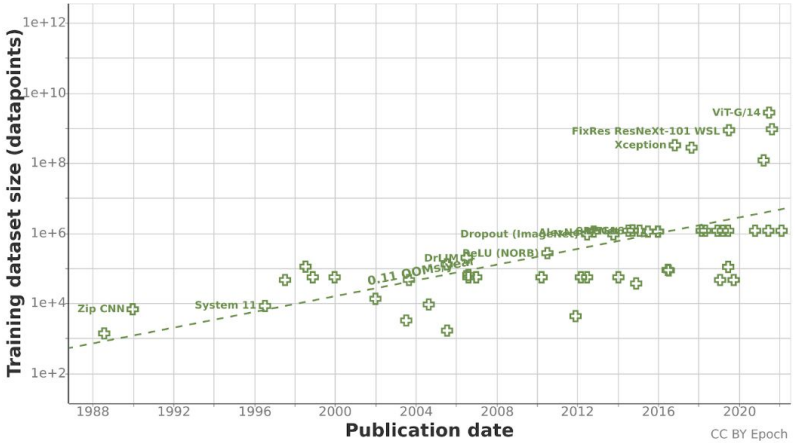
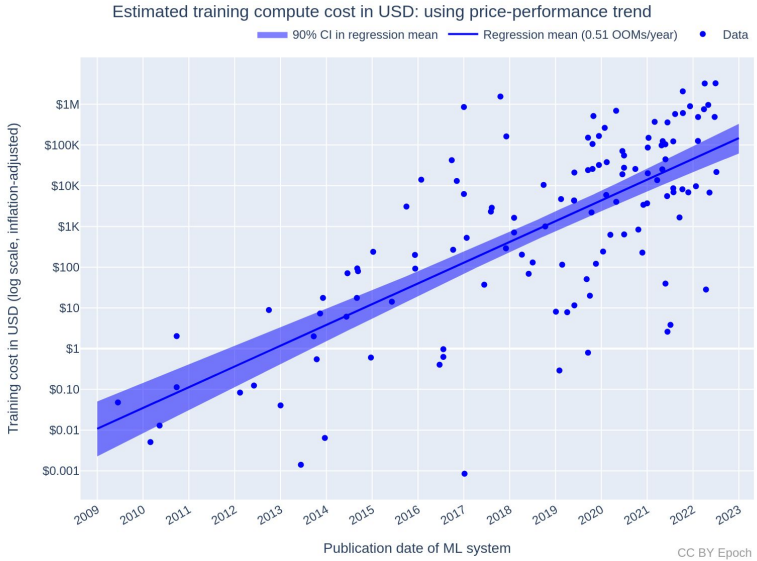


Figure 2: Evolution of vision datasets. A significant number of models is concentrated near 6e4 and 1e6, which are the sizes of MNIST and ImageNet, respectively.

https://epochai.org/blog/trends-in-training-dataset-sizes

Visit - https://ayon-roy.netlify.app

Growth in Compute cost ($)



https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems

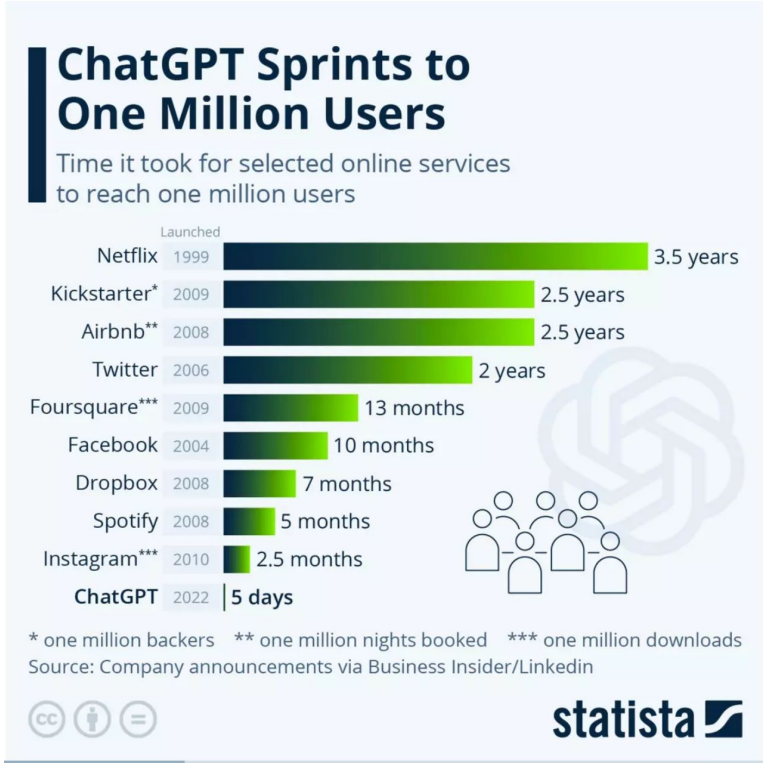
Visit - https://ayon-roy.netlify.app

The talk of the town in 2024

Generative AI

Visit - https://ayon-roy.netlify.app

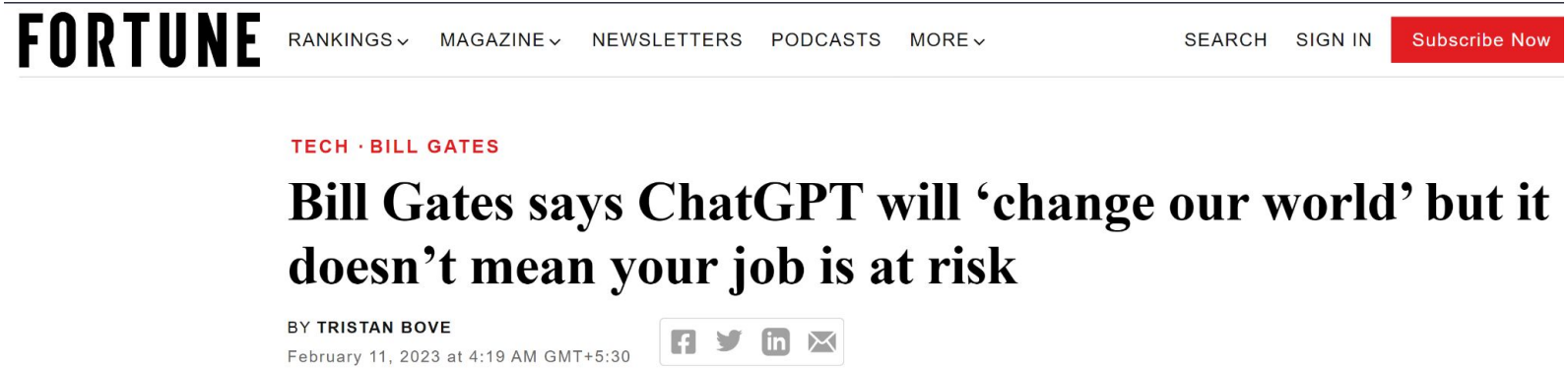
The Big Bang Moment of Generative AI



30% of outbound messages from business will be written by AI by 2025 (Gartner)

$110bn Generative AI market size by the end of the decade

Visit - https://ayon-roy.netlify.app



Visit - https://ayon-roy.netlify.app

62% of consumers would prefer to use a customer service bot rather than waiting for a human agent to answer their queries.

Source : https://www.tidio.com/blog/chatbot-statistics/

A few popular Generative AI tools

ChatGPT by Open AI

Bard by Google

BLOOM by BigScience

Galactica AI by META

DALL-E2 by OpenAI

Imagen by Google

Stable Diffusion by Stability AI

MidJourney v4 by MidJourney

Vall-Eby Microsoft for speech

PointEby OpenAI for 3D objects

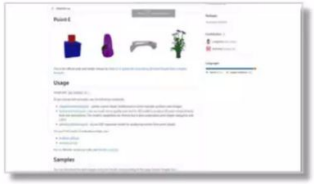


Imagen Video by Google

Chinchilla by Deep Mind for text

Visit - https://ayon-roy.netlify.app

Introduction to Generative AI

- Generative AI is a subset of artificial intelligence (AI) that involves creating models capable of generating new data or content.
- This is different from discriminative AI models that are trained to classify or predict existing data.
- Generative AI models can create new images, music, text, or other types of content that resemble or expand on the data they were trained on.

- Some popular generative AI models include Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and autoregressive models.

- The use of generative AI is growing rapidly across a variety of industries, including art, music, fashion, gaming, and healthcare.

- Generative AI has the potential to revolutionize the way we create and consume content, and can enable new forms of human-machine interaction.

How Generative AI really works ?

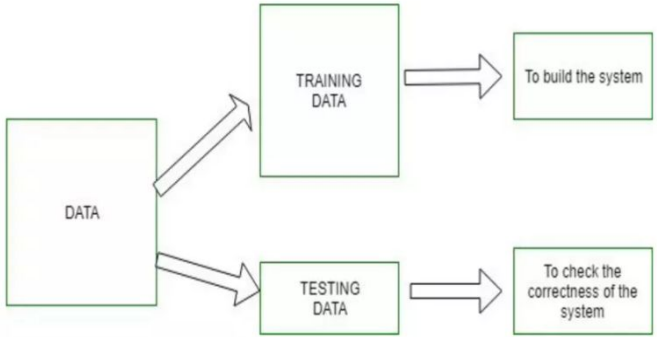
- Generative AI involves training models to learn patterns and capture the underlying structure of a given dataset.
- These trained models are then capable of generating new and original content that is similar to the training data but not an exact replica.
- Examples : Generative AI can be applied to various domains, such as image synthesis, text generation, music composition, and more, to create content that is realistic, diverse, and creative.

When you hear **“AI,”** think **“statistical pattern-matching”**

- Oracle describes AI this way:

  AI has become a catchall term for applications that perform complex tasks that once required human input, such as communicating with customers online or playing chess.

  *The term is often used interchangeably with … **machine learning (ML)** and deep learning.*



The data is **“tokenized”** *(= made into “chunks” of words, punctuation marks, pixels, etc.)* during this process - remember this for later

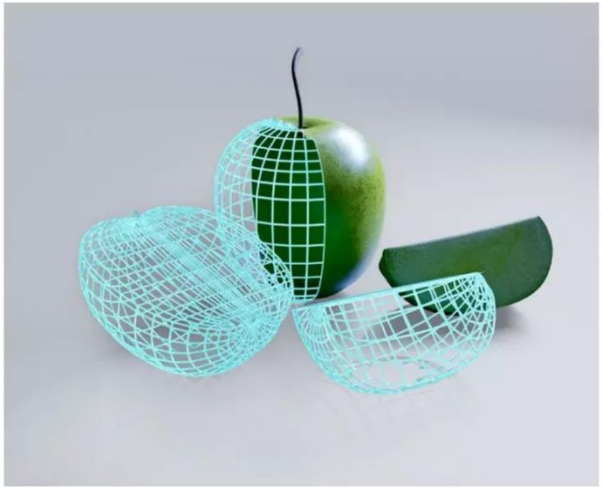
Text from What is Artificial Intelligence (AI)? Oracle, n.d. Retrieved May 16, 2023 from https://www.oracle.com/artificial-intelligence/what-is-ai/
Image from Pattern Recognition. GeeksforGeeks. Retrieved May 16, 2023 from https://www.geeksforgeeks.org/pattern-recognition-introduction/

Now, AI can synthesize *part* or *all* of a creative work

- McKinsey defines generative AI as:

  … Algorithms (such as ChatGPT) that can be used to create new content, including audio, code, images, text, simulations, and videos.

  Recent breakthroughs in the field have the potential to drastically change the way we approach content creation.



Text and image from What is generative AI? McKinsey. Retrieved May 16, 2023 from https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai

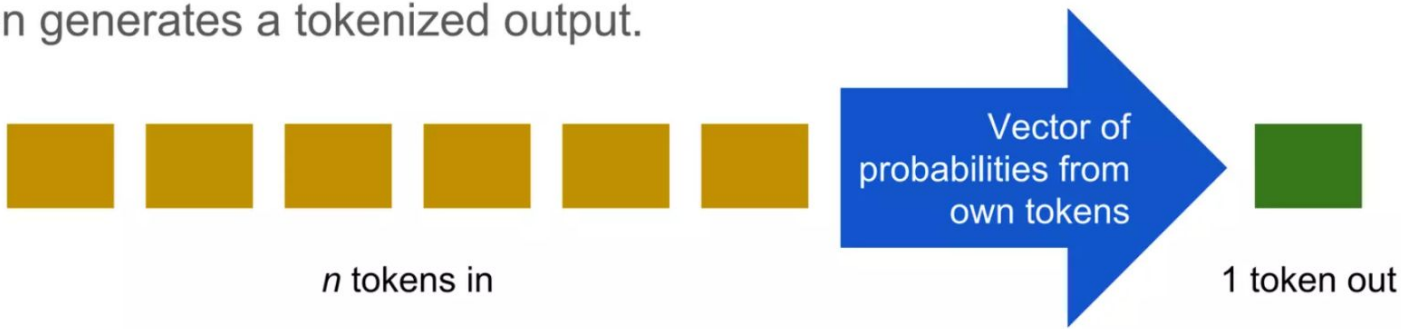
Simple probabilities (but trillions of them) are at the root of all large language models (LLMs)

Is it "aardvark"?

Or "birthday"?

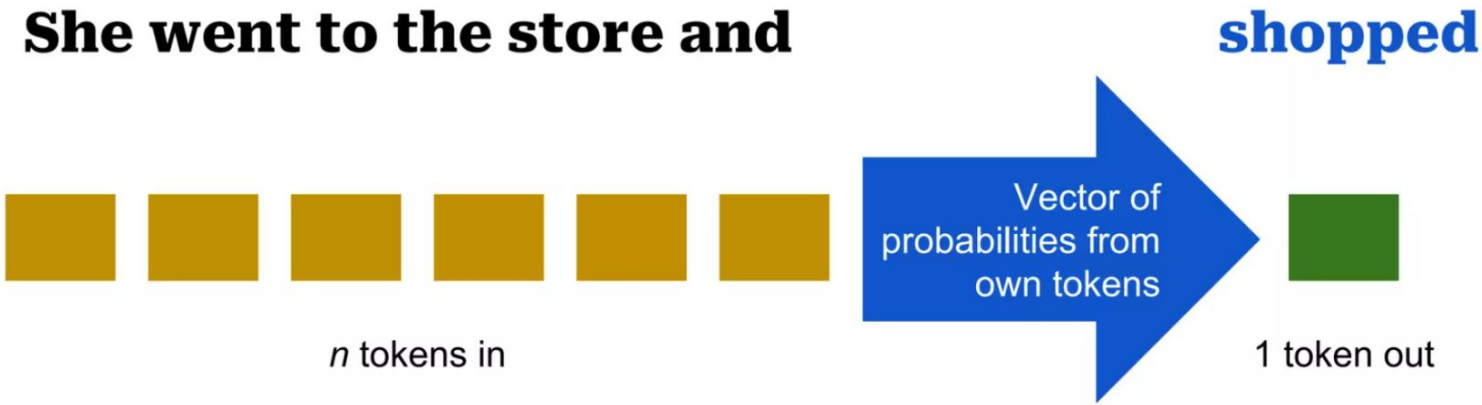
The prompts are converted into tokens *(= "chunks" of words, punctuation marks, pixels, etc.)*, then the system analyzes what is likely to come next, based on the tokens in its own dataset (as many as 32,000 in GPT-4!).

It then generates a tokenized output.



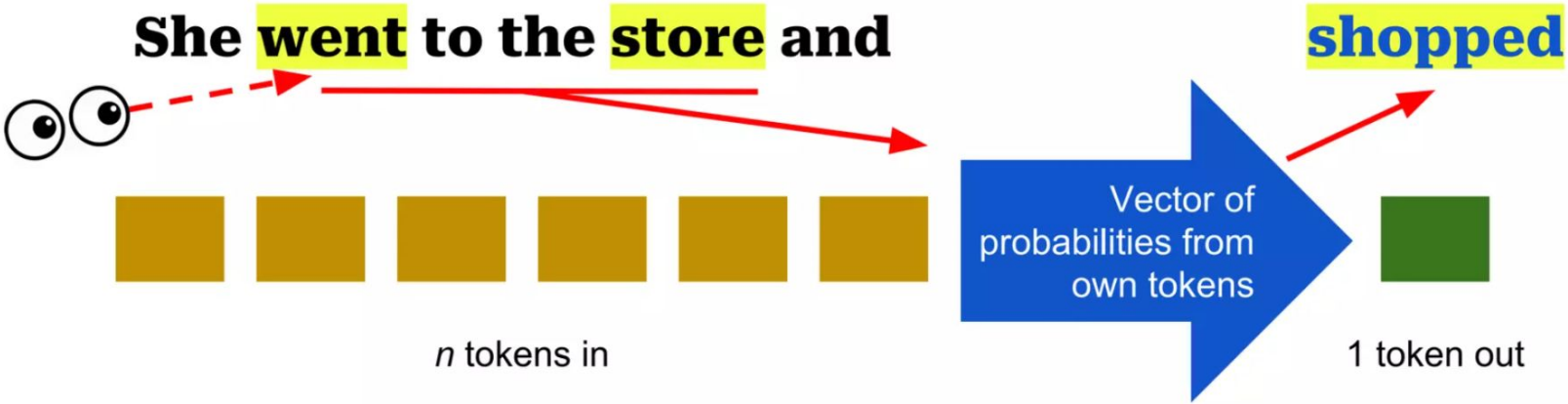
Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

With each output, it keeps re-analyzing the probabilities to decide next tokens.



Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

Transformers (the “T in “GPT”) know how to **direct attention to specific parts of the input** to guide their selection of the output - such as verb tenses, objects.



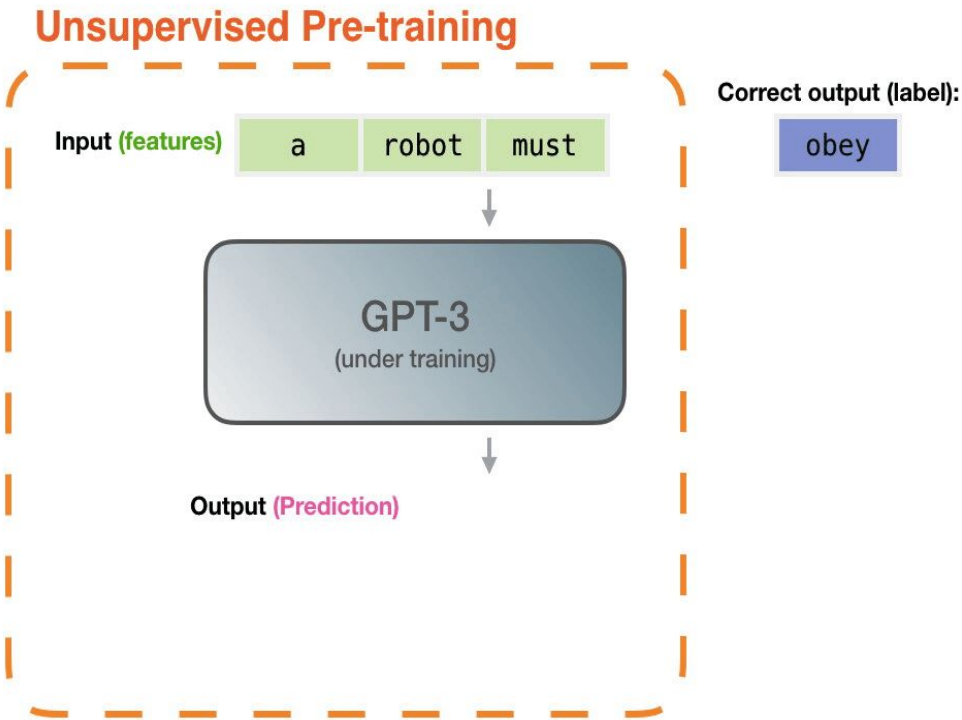
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/1706.03762
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

The system can give you different answers to the same inputs:

Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/

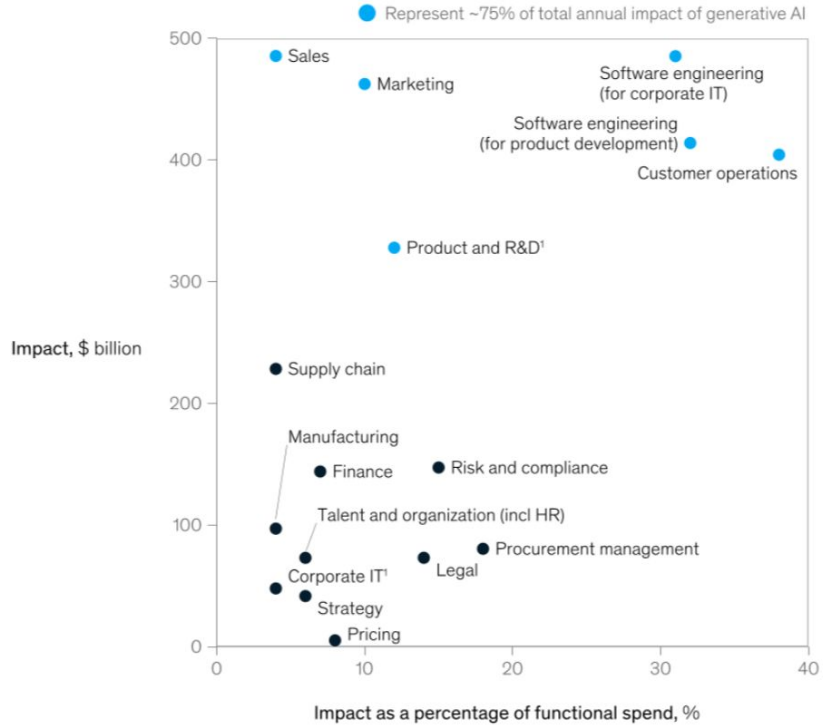
“Hallucinations” - when the output doesn’t seem to make sense - are why it is important not to accept everything it outputs at face value.

Murray Shanahan. 2022. Talking About Large Language Models. arXiv [cs.CL]. Retrieved from http://arxiv.org/abs/2212.03551
Bea Stollnitz. How generative language models work. Retrieved May 10, 2023 from https://bea.stollnitz.com/blog/how-gpt-works/


Unsupervised Pre-training
Input (features)
a
robot
must
GPT-3
(under training)
Output (Prediction)
Correct output (label):
obey

Impact of Generative AI

Using generative AI in just a few functions could drive most of the technology's impact across potential corporate use cases.



Note: Impact is averaged.
[1]Excluding software engineering.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

Visit - https://ayon-roy.netlify.app

Domain wise Generative AI impact

Generative AI use cases will have different impacts on business functions across industries.

Generative AI productivity impact by business functions[1]

Low impact — High impact

	Total, % of industry revenue	**Total,** $ billion	Marketing and sales	Customer operations	Product and R&D	Software engineering	Supply chain and operations	Risk and legal	Strategy and finance	Corporate IT[2]	Talent and organization
			760–1,200	340–470	230–420	580–1,200	280–530	180–260	120–260	40–50	60–90
Administrative and professional services	**0.9–1.4**	150–250									
Advanced electronics and semiconductors	**1.3–2.3**	100–170									
Advanced manufacturing[3]	**1.4–2.4**	170–290									
Agriculture	**0.6–1.0**	40–70									
Banking	**2.8–4.7**	200–340									
Basic materials	**0.7– 1.2**	120–200									
Chemical	**0.8–1.3**	80–140									
Construction	**0.7–1.2**	90–150									
Consumer packaged goods	**1.4–2.3**	160–270									
Education	**2.2–4.0**	120–230									
Energy	**1.0– 1.6**	150–240									
Healthcare	**1.8–3.2**	150–260									
High tech	**4.8–9.3**	240–460									
Insurance	**1.8– 2.8**	50–70									
Media and entertainment	**1.5– 2.6**	60–110									
Pharmaceuticals and medical products	**2.6–4.5**	60–110									
Public and social sector	**0.5–0.9**	70–110									
Real estate	**1.0–1.7**	110–180									
Retail[4]	**1.2–1.9**	240–390									
Telecommunications	**2.3–3.7**	60–100									
Travel, transport, and logistics	**1.2–2.0**	180–300									
		2,600–4,400									

Note: Figures may not sum to 100%, because of rounding.
[1]Excludes implementation costs (eg, training, licenses).
[2]Excluding software engineering.
[3]Includes aerospace, defense, and auto manufacturing.
[4]Including auto retail.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

Visit - https://ayon-roy.netlify.app

What to do now ?

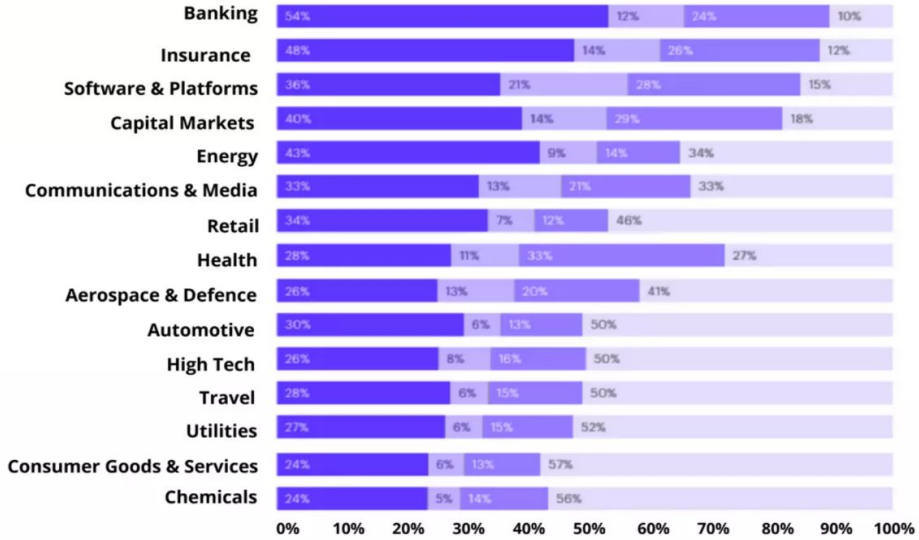
Companies must reinvent work to find a path to generative AI value. Business leaders must lead the change, starting now, in job redesign, task redesign and reskilling people.

Nearly 6 in 10 organisations plan to use ChatGPT for learning purposes and over half are planning pilot cases in 2023. Over 4 in 10 want to make a large investment.

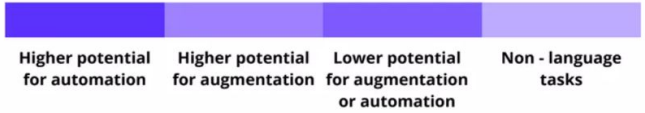
40% of working hours across industries can be impacted by Large Language Models (LLMs)

Why is this the case? Language tasks account for 62% of total worked time in the US. Of the overall share of language tasks, 65% have high potential to be automated or augmented by LLMs.

katonic.ai



Work time distribution by industry and potential AI impact:
Based on their employment levels in the US in 2021



Visit - https://ayon-roy.netlify.app

How to use Generative AI models effectively ?

Select a foundation model

AI21labs ANTHROPIC
co:here Google
Hugging Face OpenAI
Meta PowerML
... and more. Check model license

Define the task

- Classification
- Entry Extraction
- Translation
- Other

Prepare Data

- Gather, pre-process, label
- Split new data:
- Training set
- Validation set

Choose a fine-tuning strategy

- Fine tune full model or modify only last layers;
- Other layers are frozen
- Or use other strategy

Configure the model

Set Up
- Model architecture
- Hyperparameters
- Optimizer

Fine-tune the model

Train model on your specific tasks/data
- Run through multiple epochs
- Adjust hyperparameters

Monitor loss/accuracy on validation set.

Evaluate

Use the validation data set to evaluate the performance of the model

Iterate and improve

Repeat previous steps until you achieve satisfactory results.

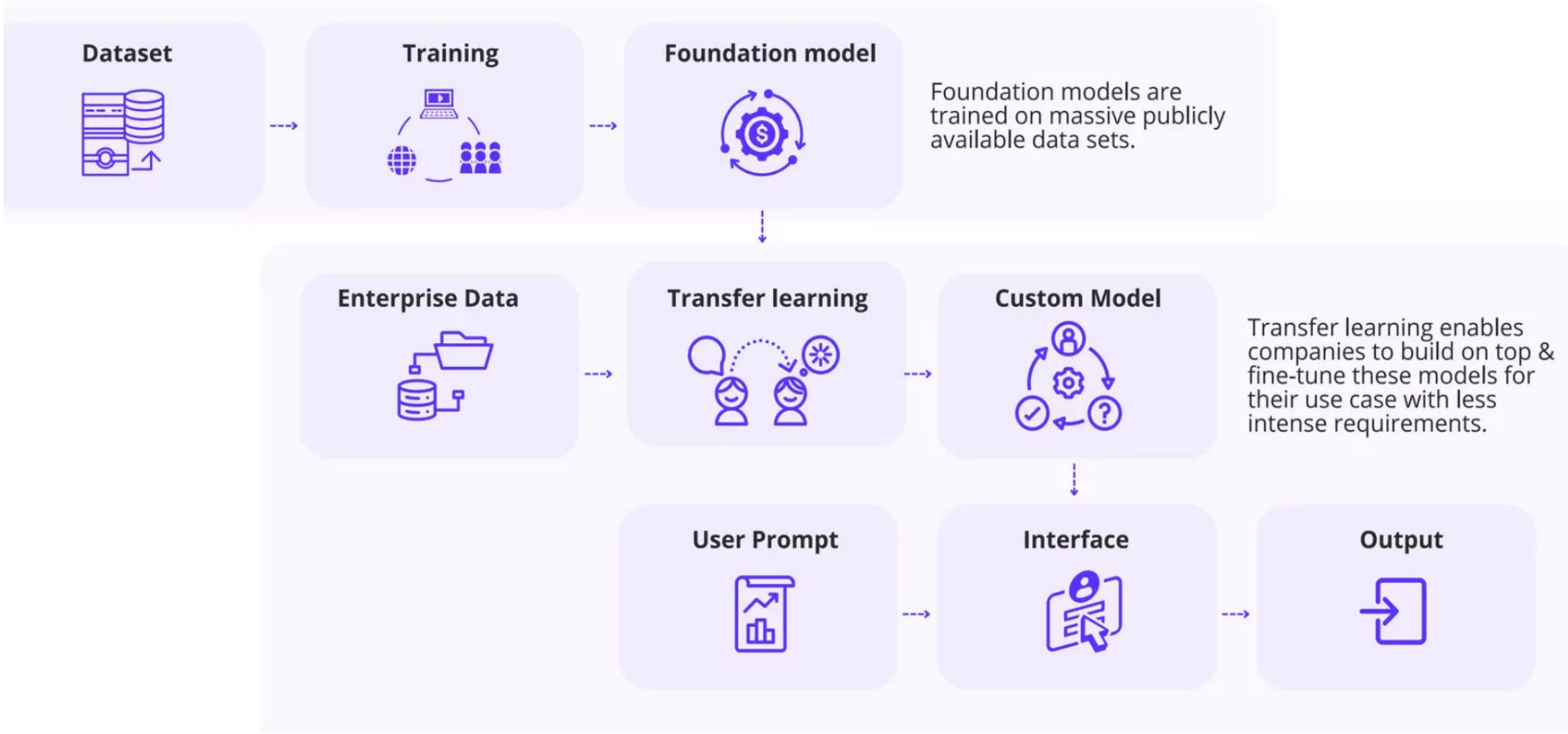
Test the model

Use model on held-out test set to confirm its performance on unseen data

Deploy or use the model

Visit - https://ayon-roy.netlify.app

Fine Tuning Generative AI models

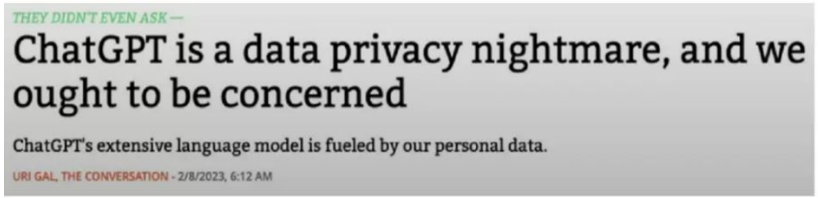


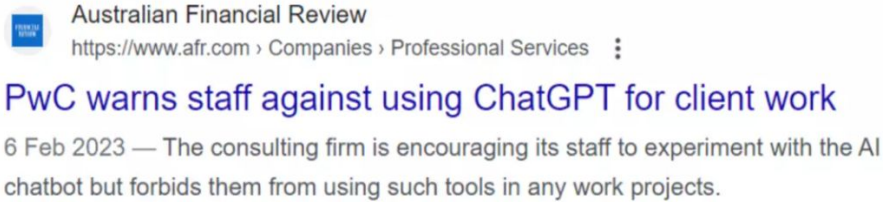
Visit - https://ayon-roy.netlify.app

Challenges in Generative AI

- Faces a challenge where the AI model may produce limited and repetitive output, lacking diversity and creativity.
- The training process of generative models can be sensitive to settings and may require careful adjustments. It can sometimes be challenging to achieve stable and consistent results.
- Generative AI models can unintentionally reflect biases present in the training data, potentially resulting in biased or unfair output. Addressing these biases and ensuring ethical use of AI is an important challenge.
- Training and utilizing generative models often require significant computational power and resources, which can pose challenges for individuals or organizations with limited access to such resources.
- Understanding how and why generative AI models generate specific content can be difficult. These models often lack transparency and explanations, making it challenging to trust and interpret their output.

Data Security, Privacy & Cost aspect of Generative AI





Visit - https://ayon-roy.netlify.app

- Violations of intellectual property
  - Check the Terms of Service - will your inputs or prompts be used as training data?
- Violations of academic integrity
  - Do a spot check of outputs, using a search engine, to see if any are wholly from another work
  - Analyze submitted work using Open AI's AI Text Classifier or the multi-service GPTZero

- Generative AI tools are great for PRODUCTIVITY - they can be nifty shortcuts to dispose of low-value tasks and / or to jumpstart creativity
- Generative AI tools should always be used - *and taught to be used* - with a critical mind, because they are prone to mistakes and "hallucinations"

And that’s where, is the biggest opportunity for RSE

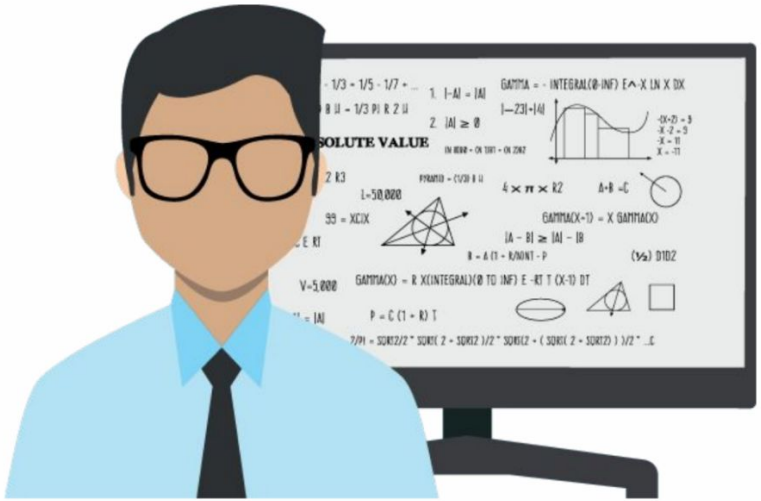
Visit - https://ayon-roy.netlify.app

Responsible AI

Sustainable AI

Ethical AI

Visit - https://ayon-roy.netlify.app



HUMANS AND MACHINES HAVE COEXISTED FOR CENTURIES
IF HARNESSED CORRECTLY, IT HAS CHANGED THE WORLD

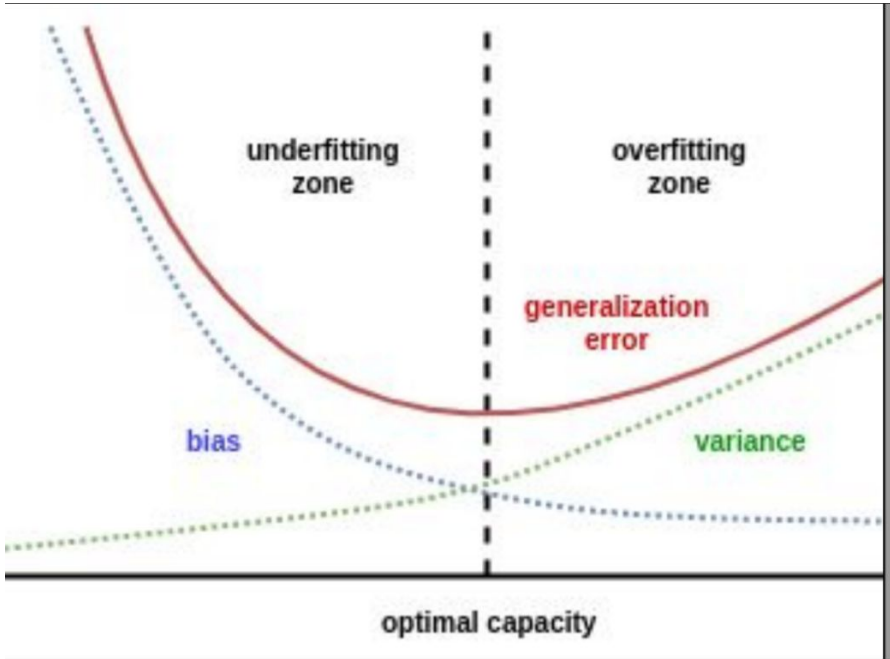
Visit - https://ayon-roy.netlify.app

RISK

- a possibility of something dangerous or unpleasant happening; a situation that could be dangerous or have a bad result

RISK MANAGEMENT

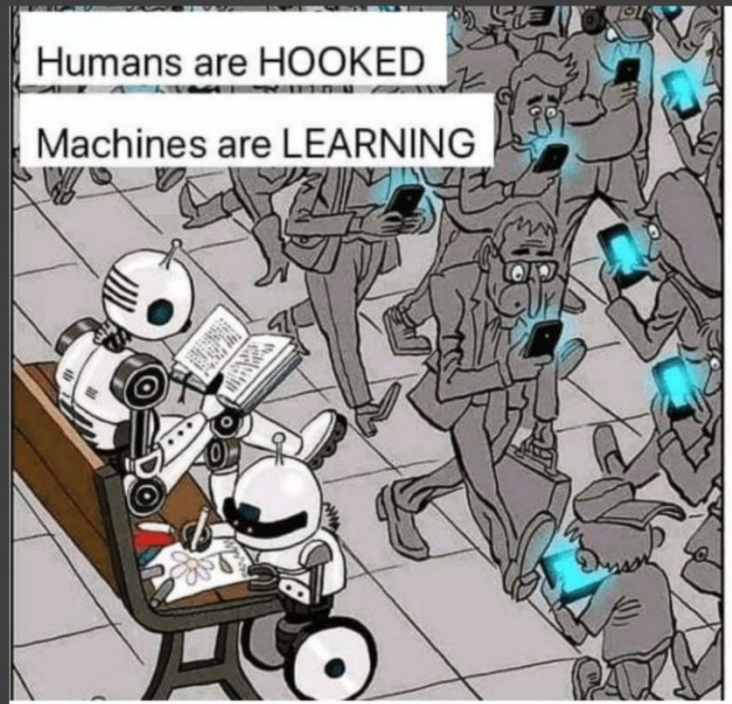
- Risk management is the process of identifying, assessing and controlling threats to an organization's capital and earnings.



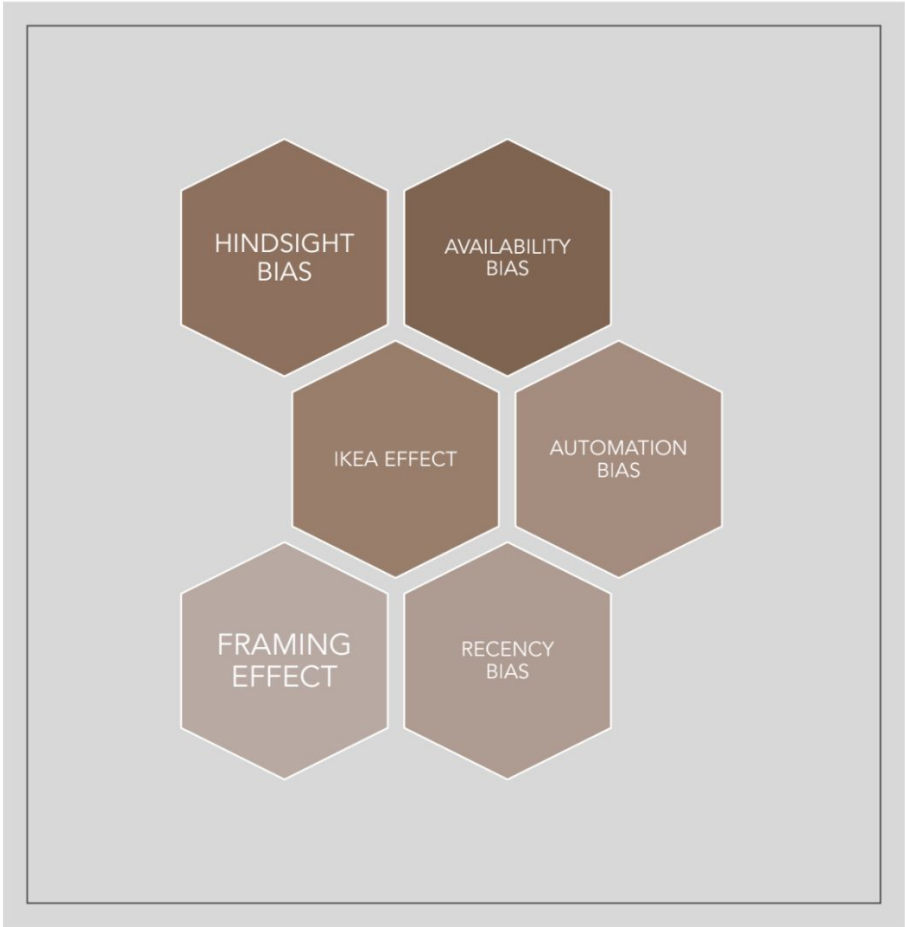
	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Negative Type II
Actual Negative	False Positive	True Negative

ERRORS IN MACHINE LEARNING MODELS

HUMANS AREN'T PERFECT, BUT TECHNOLOGY IS DESIGNED TO BE NEAR PERFECT



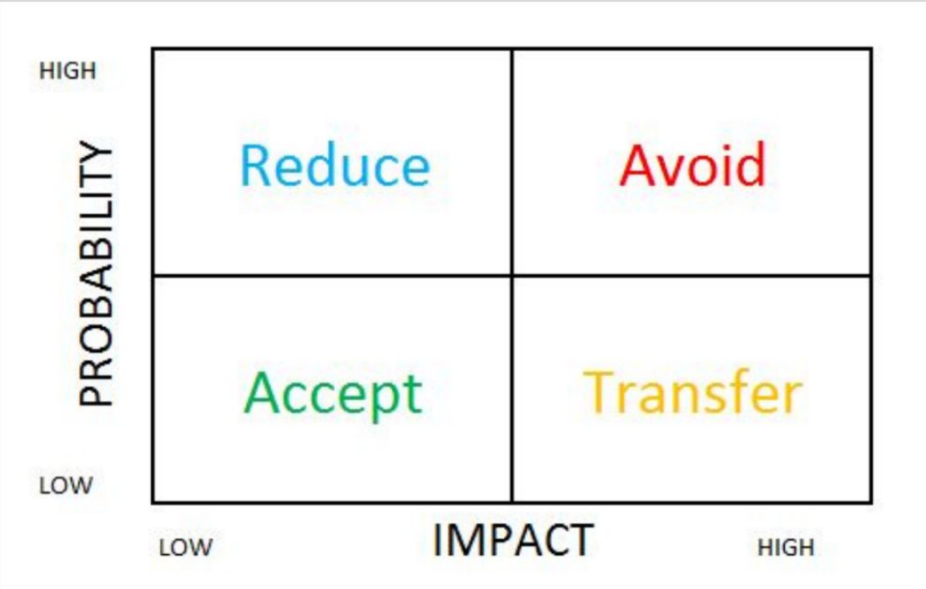
HUMANS ARENT AS RATIONAL AS THEY BELIEVE THEMSELVES TO BE



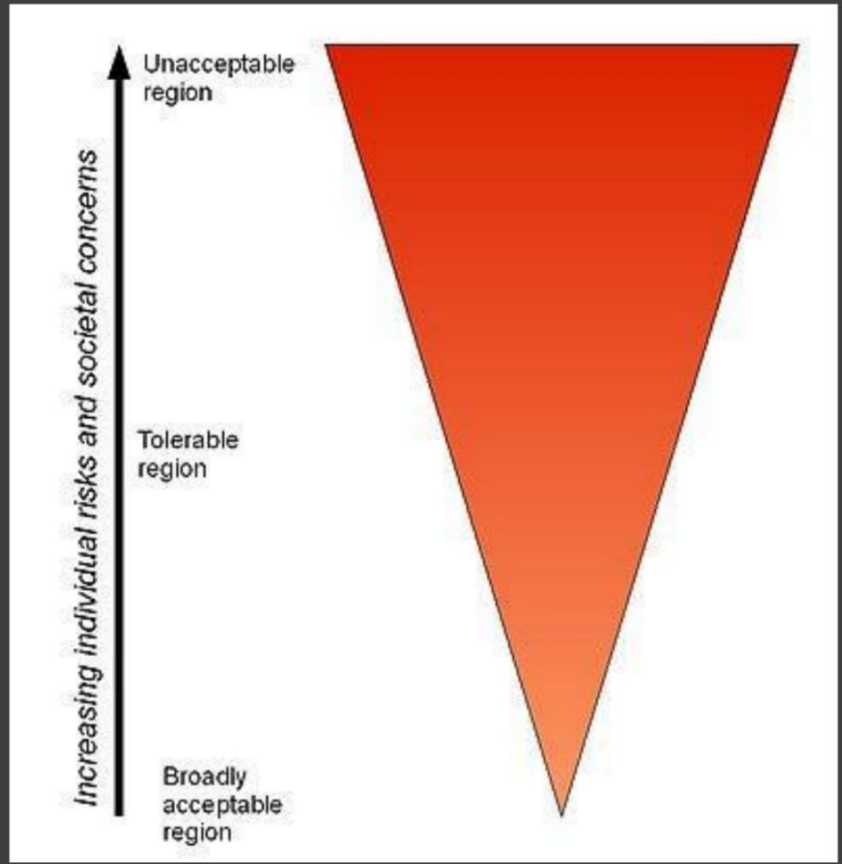
Visit - https://ayon-roy.netlify.app

THE RISK MANAGEMENT PROCESS

STANDARD RISK MANAGEMENT MODELS 1 - TARA



AS LOW AS REASONABLY POSSIBLE



REGULATION AND COMPLIANCE IN PLACE

- The Comprehensive Capital Analysis and Review (CCAR)
- Dodd-Frank Act stress testing (DFAST)
- FAIR Lending compliance standards
- GDPR (Right to information)
- EU AI regulation 2021
- Personal Data Protection bill (India)

What can be done to mitigate these risks? -1

Data audits

1. Is the training data a true reflection of the population?
2. Is the data collection methodology reliable?
3. Will this data drift in near future?
4. What was the motivation for collecting the data?
5. Did they follow a scientific experiment designing process, or just went with what's convenient?
6. What can go wrong?

What can be done to mitigate these risks? -2

MODEL AUDITS

1. Is the model using correct metric for evaluation?
2. Is the model fair to protected classes?
3. Is it suitable for use in high stake situations?
4. What would be the ideal retraining and refresh frequency?
5. Does the model comply with the local/industry specific laws?

What can be done to mitigate these risks? -3

HUMAN CENTRIC DISCUSSIONS

1. Critical Thinking
2. Being mindful of prejudices vs Intuition
3. Involving domain experts
4. Formulating AI and Data Strategies and implementing it
5. Ensuring stakeholder diversity

Responsible AI

Sustainable AI

Ethical AI

Visit - https://ayon-roy.netlify.app

Major approaches used while doing AI

Model Centric & Data Centric Approach

Model-Centric Approach

This involves designing empirical tests around the model to improve the performance. This consists of finding the right model architecture and training procedure among a huge space of possibilities.

Data-centric approach

This consists of systematically changing/enhancing the datasets to improve the accuracy of your AI system. This is usually overlooked and data collection is treated as a one off task.

https://towardsdatascience.com/from-model-centric-to-data-centric-artificial-intelligence-77e423f3f593#:~:text=Data%2Dcentric%20approach,as%20a%20one%20off%20task

Community's Bias towards Model Centric Approach

The steel sheets defect detection was one of the examples brought during the session — assuming a series of images from steel sheets we want to develop the best model to detect these defects that can happen during the process of steel sheets manufacturing. There are 39 different defects that we want to be able to identify. By developing a computer vision model with well-tuned hyperparameters, it was able to reach a **76.2% accuracy baseline system**, but the goal is to achieve **90% accuracy**. *How can this be done*?

Steel Sheets Detection Challenge

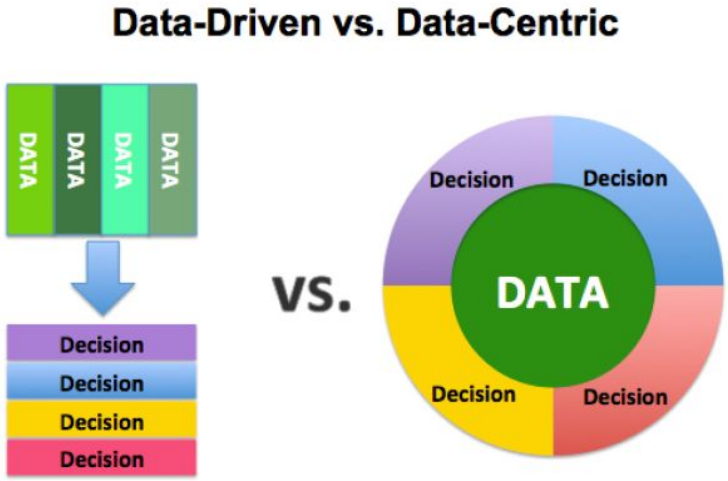
https://www.youtube.com/watch?v=06-AZXmwHjo&t=148s

Difference in Results

Knowing that the baseline model was already good, the task to have it improved to achieve 90% accuracy sound almost impossible — for the model-centric, the improvements based on Network Architecture search and using the state-of-the-art architectures, whereas, for the data-driven, the approach taken was to identify inconsistencies and clean noisy labels. The results were mind-blowing:

Steel sheets defects detection	**Baseline**	**Model-centric**	**Data-centric**
Accuracy	76.2%	+0% (76.2%)	+16.9% (93.1%)

Beware of the Trade Off



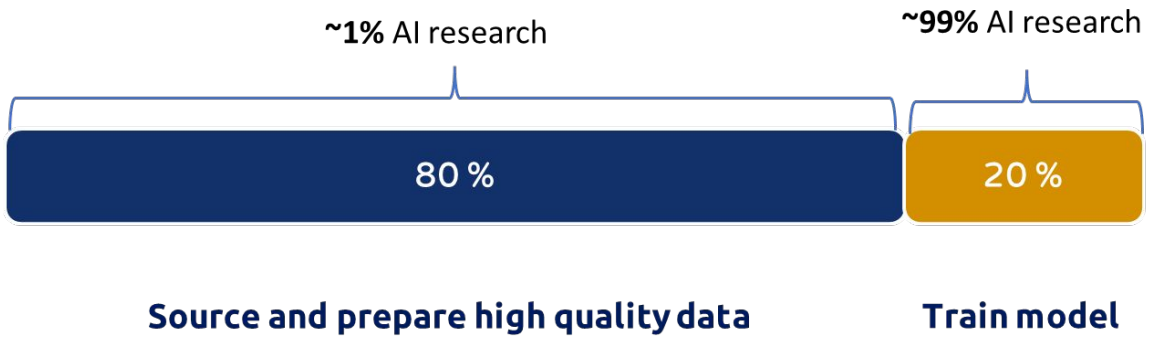
https://neptune.ai/blog/data-centric-vs-model-centric-machine-learning

But..

Check for Data Quality

https://www.lotame.com/why-is-data-quality-important/

Need for Data Centric Approaches



https://www.youtube.com/watch?v=06-AZXmwHjo&t=1835s

Use of Data effectively for AI, ML

Our projections predict that we will have exhausted the stock of low-quality language data by 2030 to 2050, high-quality language data before 2026, and vision data by 2030 to 2060. This might slow down ML progress.

All of our conclusions rely on the unrealistic assumptions that current trends in ML data usage and production will continue and that there will be no major innovations in data efficiency. Relaxing these and other assumptions would be promising future work.

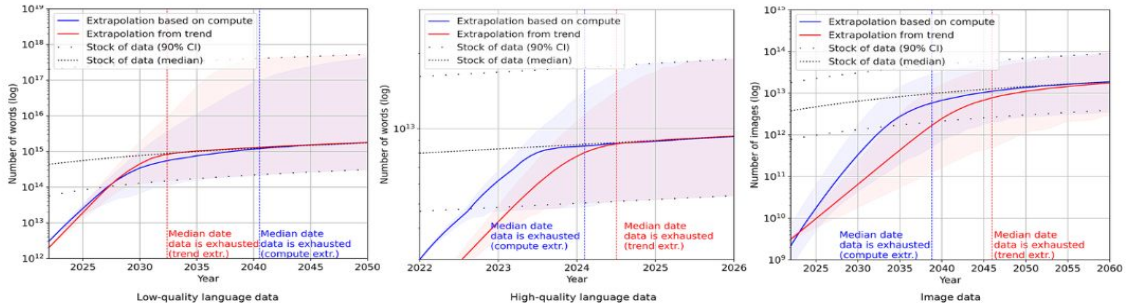
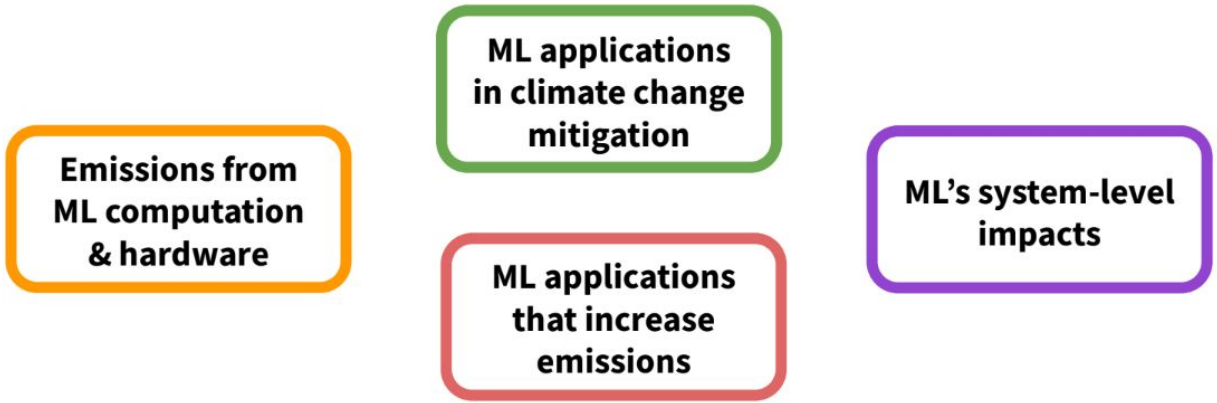


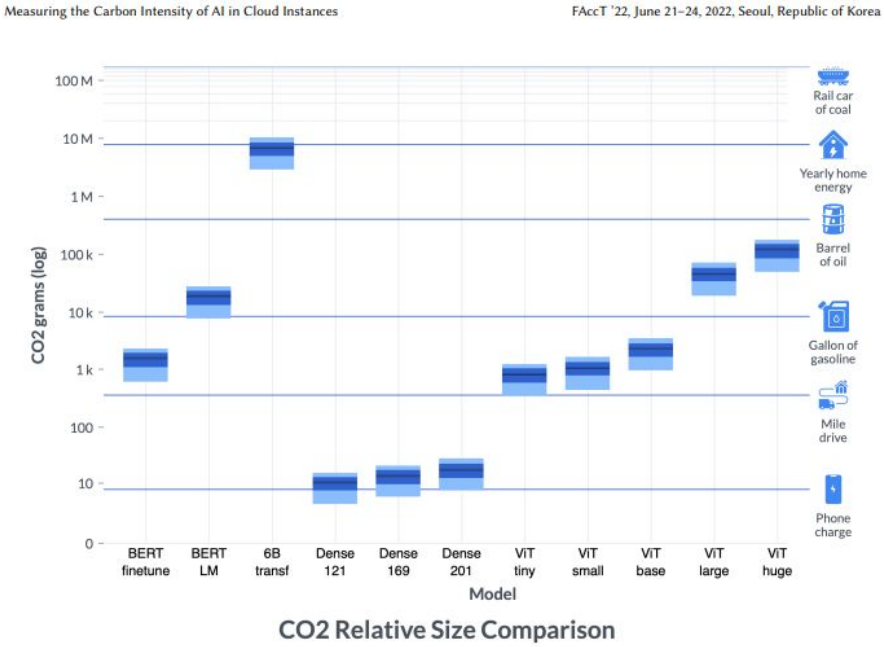
Figure 1: ML data consumption and data production trends for low quality text, high quality text and images.

https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset

AI-ML Carbon Footprint



Emissions from AI Cloud Instances



Measuring the Carbon Intensity of AI in Cloud Instances : https://arxiv.org/pdf/2206.05229.pdf

Emissions for the research paper's 11 experiments. For each model they show a vertical blue bar, where the top of the bar is the max, the bottom is the min, and the black line represents the average emissions (across regions and time of year).

First and fourth quartiles are represented by the light blue at the top and bottom of each vertical blue bar. The largest training runs (e.g., 6 billion parameter LM) releases a significant amount of emissions, no matter the region (and recall the 6 billion parameter LM is only trained for 13% of a full run, so a full run would emit about an order of magnitude more emissions than reported here).

The smallest experiments emit very little. Presented on a log scale, with references on the right indicating equivalent sources of emissions per the United States Environmental Protection Agency.

Necessity for exploring SustAInability

Visit - https://ayon-roy.netlify.app

Approaching SustAInability as you build AI Systems

One consequence of this increase in computing is the heavy environmental impact of training machine learning models. A recent research paper — Energy and Policy Considerations for Deep Learning in NLP — notes that an inefficiently trained NLP model using Neural Architecture Search can emit **more than 626,000 pounds of CO_2.** That's about **five times the lifetime emissions of an average American car!**

https://wandb.ai/amanarora/codecarbon/reports/Tracking-CO2-Emissions-of-Your-Deep-Learning-Models-with-CodeCarbon-and-Weights-Biases--VmlldzoxMzM1NDg3

Visit - https://ayon-roy.netlify.app

Comparison of Certain NLP Models

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	$41–$140
Transformer_{big}	P100x8	1515.43	84	201	192	$289–$981
ELMo	P100x3	517.66	336	275	262	$433–$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	$3751–$12,571
BERT_{base}	TPUv2x16	—	96	—	—	$2074–$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	$942,973–$3,201,722
NAS	TPUv2x1	—	32,623	—	—	$44,055–$146,848
GPT-2	TPUv3x32	—	168	—	—	$12,902–$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Relevant Research Paper : https://arxiv.org/pdf/1906.02243.pdf

SustAInable Development

UN Brundtland Commission in its report "Our Common Future", published in 1987 defines sustainable development as "**development that meets the needs of the present without compromising the ability of future generations to meet their own needs**.

It contains within it two key concepts:

- the concept of 'needs', in particular the essential needs of the world's poor, to which overriding priority should be given;
- the idea of limitations imposed by the state of technology and social organization on the environment's ability to meet present and future needs.

SustAInable Future , is it ?

AI has the potential to assist in the achievement of each of the SDGs. This is illustrated by a study by the McKinsey Global Institute, which by November 2018 had identified as many as 135 cases of AI use worldwide that support the SDGs. [Whether concrete, partial or simply potential, these cases demonstrate real opportunities for AI applications]

The proposed hypothesis is that

> By becoming aware of the potential precariousness of the resources, room for maneuver and other properties intrinsic to the nature of the actor, the business and industry sector, nongovernmental organizations, the scientific and technological community, and local authorities, i.e. the State, are the actors best placed not only to impact generally on the progress towards the SDGs, but also to be the channels through which AI can most ideally, until then, impact beneficially on Sustainable Development and reach the SDGs.

A use case for SustAInable future

What do farmers want to know?

- ★ When to plant?
- ★ Crop performance
- ★ Potential threats to production (e.g. climate change)
- ★ Actual threats to production (e.g.,nearby pest/disease outbreak or weather forecasts)
- ★ Soil moisture, rainfall, temperature, etc.
- ★ Productivity potential (yield gap)
- ★ Suitability of crops (would a different crop or variety grow better?)

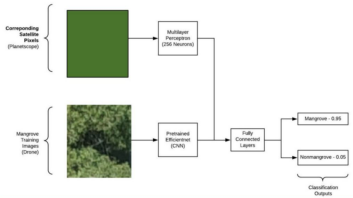
What do policymakers want to know?

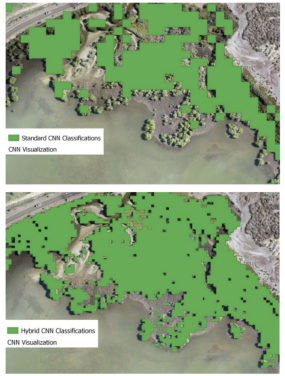
- ★ Crop performance
- ★ Potential threats to production
- ★ Actual threats to production
- ★ When to intervene
- ★ How to intervene
- ★ Productivity potential
- ★ Suitability of crops
- ★ How suitability will change
- ★ Measure impacts of policies

Visit - https://ayon-roy.netlify.app

A few other use-cases

Mangrove classification

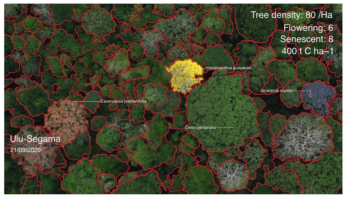




Fusing different data resolutions improves classification accuracy
Challenges: Labeled Data

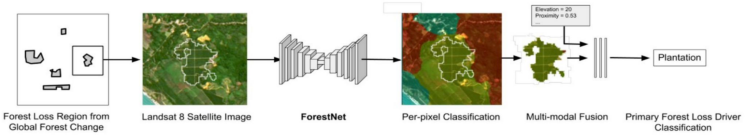
Mangrove Ecosystem Detection using Mixed-Resolution Imagery with a Hybrid-Convolutional Neural Network
Hicks *et al.*, CCAI ICML20

Drone-Based Biomass Estimation



Challenges:
Limited labeled data for model training

Deforestation driver classification



Model	Predictors	Val		Test	
		Acc	F1	Acc	F1
RF	Visible	0.56	0.49	0.49	0.44
RF	Visible + Aux	0.72	0.67	0.67	0.62
CNN	Visible	0.80	0.75	0.78	0.70
CNN + SDA	Visible	0.82	0.79	0.78	0.73
CNN + SDA + PT	Visible	0.83	0.80	0.80	0.74
CNN + SDA + PT	Visible + Aux	**0.84**	**0.81**	**0.80**	**0.75**

Fusing multi-modal data sources improve classification accuracy
Challenges: Labeled Data

ForestNet: Classifying Drivers of Deforestation in Indonesia using Deep Learning on Satellite Imagery
Irvin *, Sheng * *et al.* CCAI NeurIPS20

https://ai-for-sdgs.academy/
https://omdena.com/projects/

Visit - https://ayon-roy.netlify.app

AI used for Energy Saving

DATA CENTER DESIGN

Meta Previews New Data Center Design for an AI-Powered Future

Meta is retooling its massive digital infrastructure for AI, fine-tuning everything from tiny chips to giant data centers. As part of that shift, Meta confirmed that its new design will make extensive use of liquid cooling.

Rich Miller

May 18, 2023

https://www.datacenterfrontier.com/data-center-design/article/33005296/meta-previews-new-data-center-design-for-an-aipowered-future

DeepMind's AI cuts energy costs for cooling buildings

Research firm DeepMind has built an AI to optimise cooling systems in buildings. In tests, it reduced energy usage by around 10 per cent

TECHNOLOGY 20 December 2022

By Jeremy Hsu

https://www.newscientist.com/article/2352075-deepminds-ai-cuts-energy-costs-for-cooling-buildings/

How can you approach SustAInability?

Visit - https://ayon-roy.netlify.app

Focus on your day to day activities

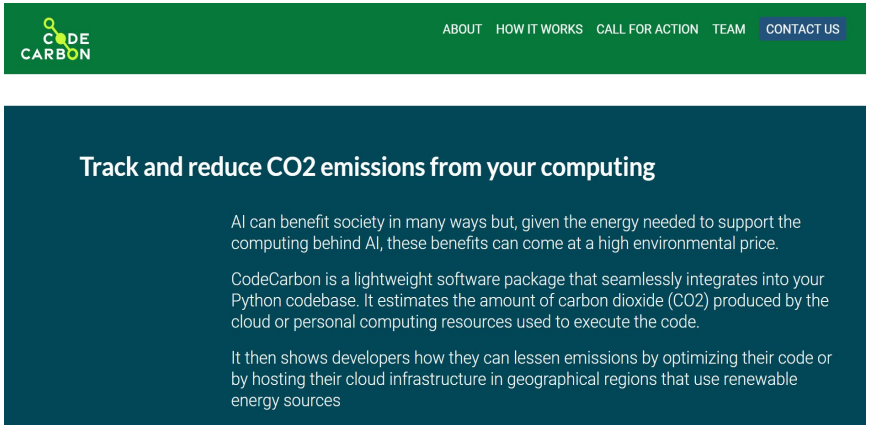
Computing-related

- Measure your footprint with tools such as ML CO2 Impact, CodeCarbon, Carbontracker, or tools specifically for Azure or Hugging Face
- Reduce your impacts by choosing more efficient models, and reducing wasteful model retraining & execution

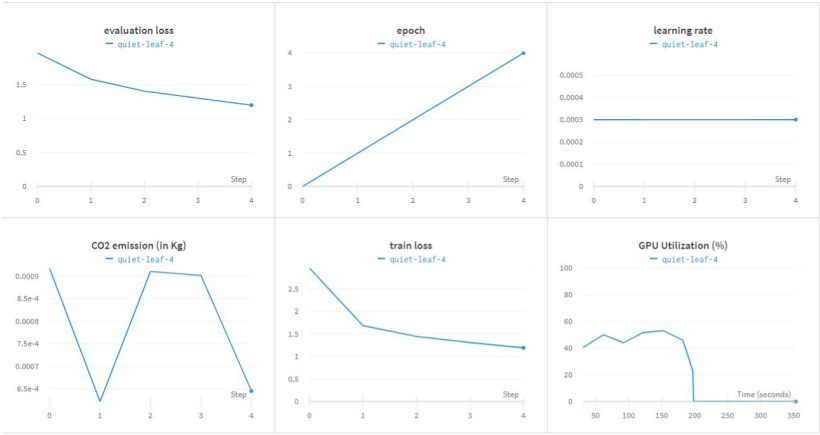
Application-related

- Quantify and evaluate the application impacts where possible
- Be transparent about impacts in publications and with stakeholders (quantitatively and qualitatively)
- Choose what you (or the ML community) works on

Tracking CO2 Emissions of Your Deep Learning Models with CodeCarbon + Weights & Biases



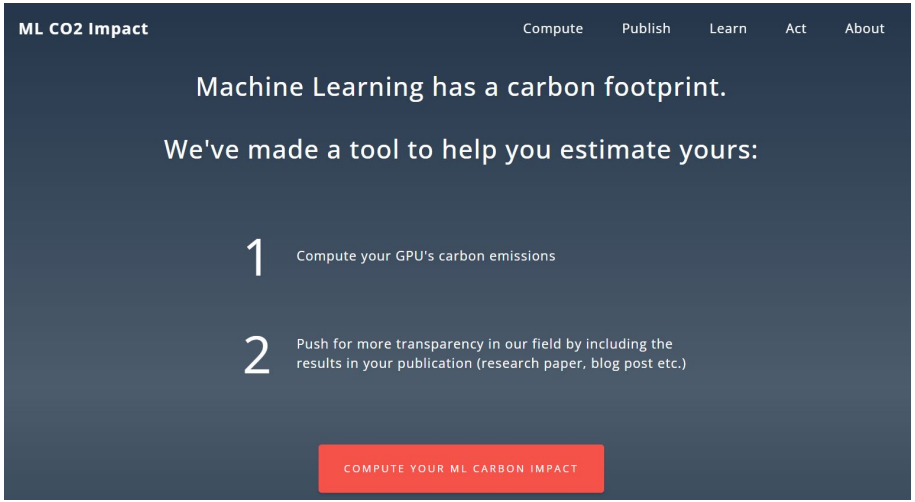
https://codecarbon.io/



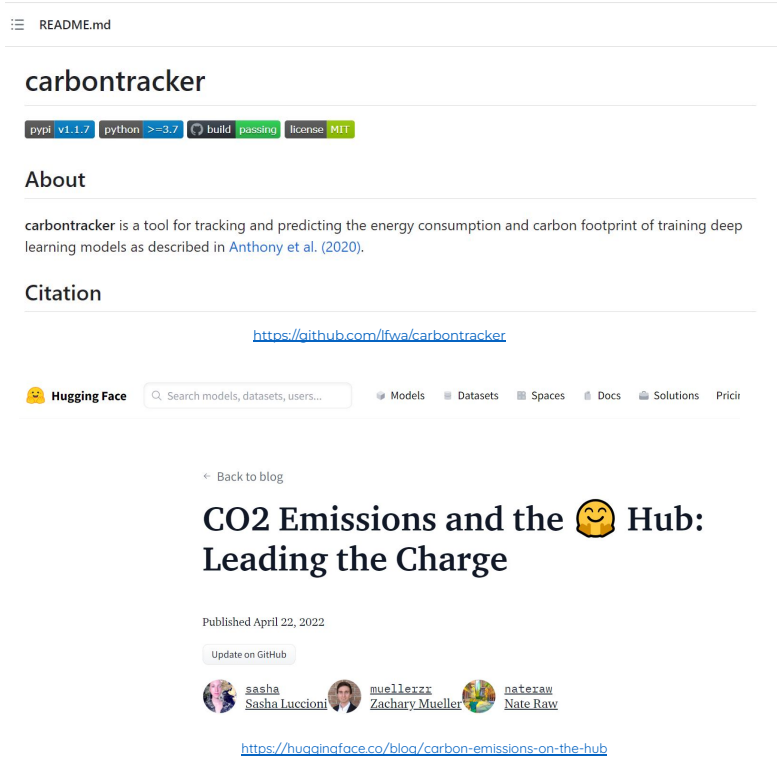
https://wandb.ai/amanarora/codecarbon/reports/Tracking-CO2-Emissions-of-Your-Deep-Learning-Models-with-CodeCarbon-and-Weights-Biases--VmlldzoxMzM1NDg3

Visit - https://ayon-roy.netlify.app

A few other interesting ideas to explore



https://mlco2.github.io/impact/



https://huggingface.co/blog/carbon-emissions-on-the-hub

Visit - https://ayon-roy.netlify.app

What’s the future ?

Visit - https://ayon-roy.netlify.app

WORLD ECONOMIC FORUM

Top 10 skills of 2025

Type of skill

- Problem-solving
- Self-management
- Working with people
- Technology use and development

Source: Future of Jobs Report 2020, World Economic Forum.

Lets focus on the RSE while doing AI

Danke Schoen

Questions ? Any Feedbacks ? Did you like the talk?
Tell me about it.

If you think I can help you,
connect with me via

Email : ayon-roy@outlook.com

LinkedIn : https://www.linkedin.com/in/ayon-roy

Website : https://AYON-ROY.NETLIFY.APP/

Scan to Connect

Thank You