

AI in InfoSec? Lets talk about it...

Authors: Joff Thyer, and Derek Banks

Copyright © 2024



Importance to Information Security

- Increasing Demand
- Skills and training deficit
- Human resource deficit
- A need to accelerate task completion
- Cost reduction

[Some images created with the assistance of DALL·E ·3]

Key Challenges and Opportunities

- We are in the hype cycle phase with AI and Natural Language Processing (NLP)
- Multiple opportunities exist to apply data science/AI/ML to information security problems
 - Time series analysis
 - Text classification
 - Anomaly Detection
 - Analyst "6th sense"
- Attacks against AI technology

[Some images created with the assistance of DALL·E ·3]

What is Artificial Intelligence?

- **“The science and engineering of making intelligent machines”**
 - The term was coined by Stanford Professor John McCarthy in 1955
- **Autonomous Systems** can independently plan and execute a sequence of steps to reach a specific goal.
- **Machine Learning (ML)** is a part of AI studying how computers improve perception, knowledge, and thinking by leveraging data.
- **Deep Learning** is the use of large multi-layer neural networks organized in a human brain like configuration.



[Some images created with the assistance of DALL·E ·3]

Historical Notes

- The field grew out of Neural Network research circa 1959
- Bernard Widrow and Marcian Hoff at Stanford
 - Developed ADALINE, and MADALINE
 - Multiple **ADA**ptive **LI**near **E**lements
- MADALINE solved a real-world problem of “phone audio echo”
 - “Predicting the next bit value in a sequence”



[Some images created with the assistance of DALL·E ·3]

Von Neumann architecture impact

- Neural network research and application was viable
- Von Neumann's computer architecture dominated
 - Serial architecture not well suited for neural nets
 - Neural network research stagnated
- In 1982, research was reinvigorated by John Hopfield at Caltech.

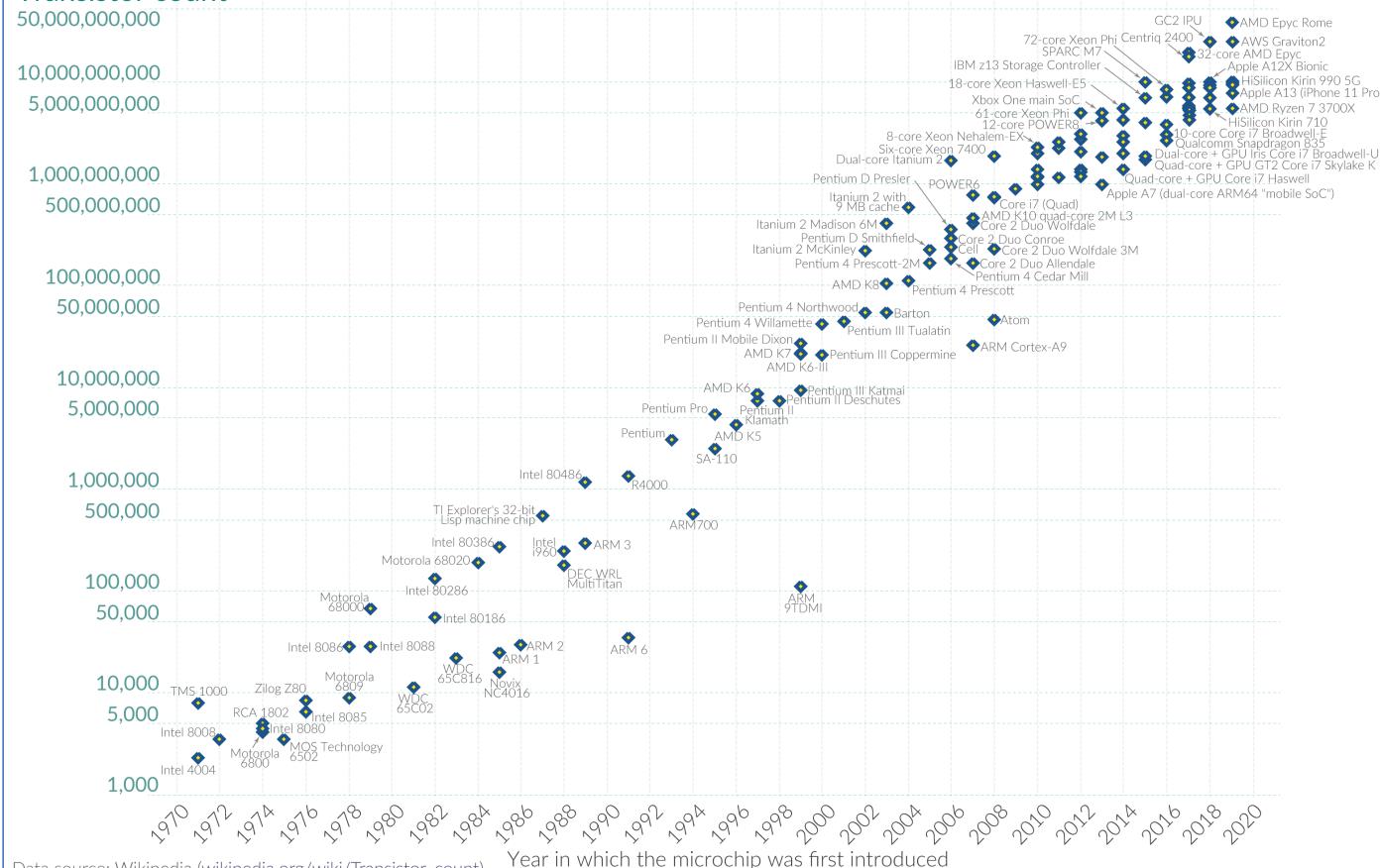
[Some images created with the assistance of DALL·E ·3]

Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World in Data

Transistor count



Data source: Wikipedia ([wikipedia.org/wiki/Transistor_count](https://en.wikipedia.org/w/index.php?title=Transistor_count&oldid=1000000000))

OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

Historical AI Tech Timeline

- 1966: Eliza developed at MIT by Joseph Weizenbaum
- 1972: Statistically Trained Natural Language Processor at MIT
- 1997: Long Term Short Term (LSTM) model developed Hochreiter and Schmidhuber
- 1999: Nvidia introduces their GPU
- 2000 - 2016: Multiple new AI models released
 - IBM, Facebook/Meta, Google
 - OpenAI funded in 2015

Ref: <https://synthetica.substack.com/p/a-timeline-of-large-language-model>

[Some images created with the assistance of DALL·E ·3]

8

<https://github.com/RiverGumSecurity/IntroAIJobs>



Historical Timeline...

- 2017: Introduction of Transformer models at Google
- 2018: Google makes Tensor Flow Processors (TPUs) available
 - Open AI publishes paper on GPT
 - Google introduces BERT – an NLP trained by Google
- 2019: OpenAI releases GPT-2
 - Baidu, Microsoft, Facebook, and AWS all competing

Ref: <https://synthetica.substack.com/p/a-timeline-of-large-language-model>

[Some images created with the assistance of DALL·E ·3]

9

<https://github.com/RiverGumSecurity/IntroAIJobs>



Historical Timeline...

- 2022 - 2024: An explosive growth of AI-LLM deployments



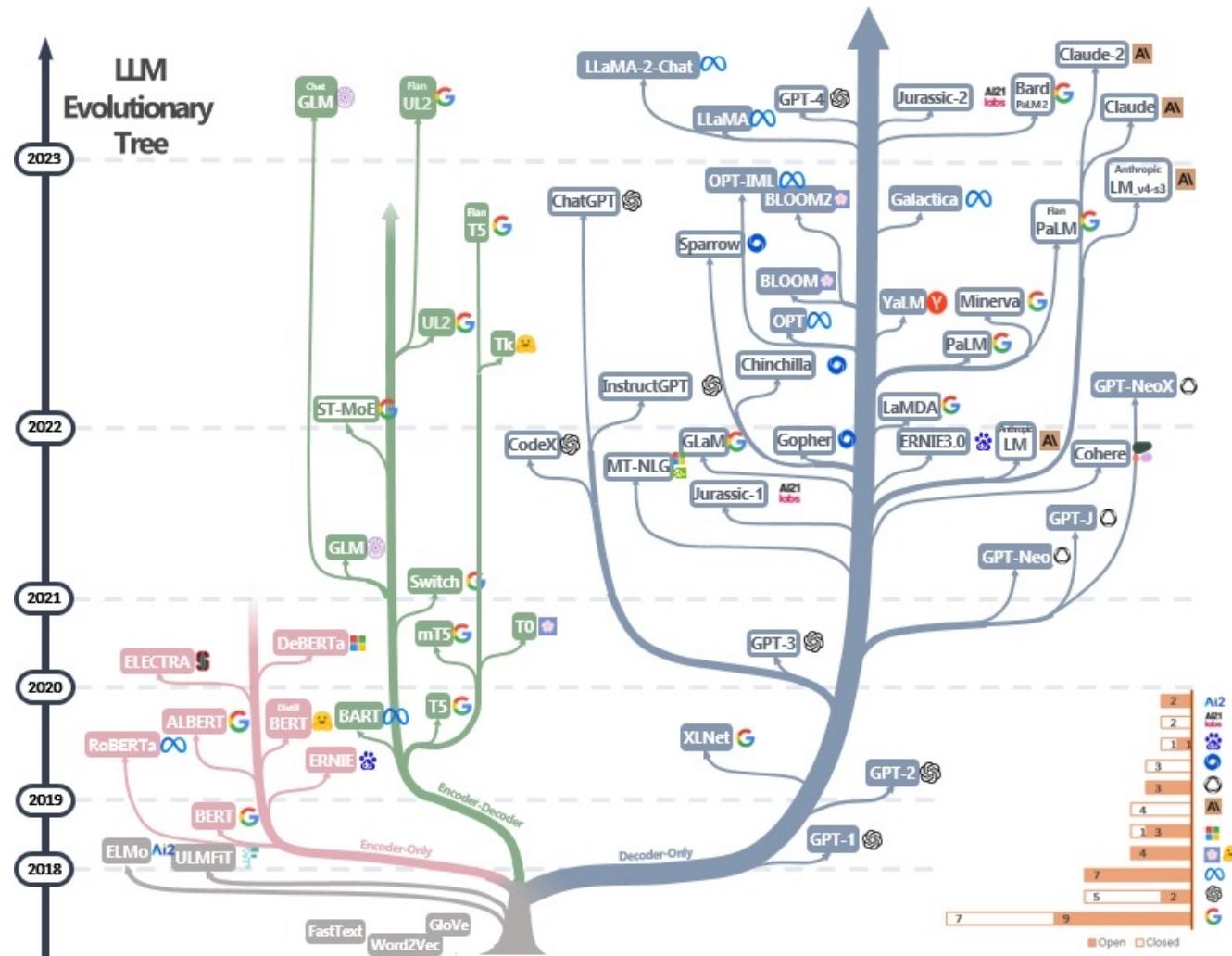
Ref: <https://synthetica.substack.com/p/a-timeline-of-large-language-model>

[Some images created with the assistance of DALL·E ·3]

10

<https://github.com/RiverGumSecurity/IntroAI Labs>





Ref: <https://github.com/Mooler0410/LLMsPracticalGuide>

[Some images created with the assistance of DALL·E 3]

13

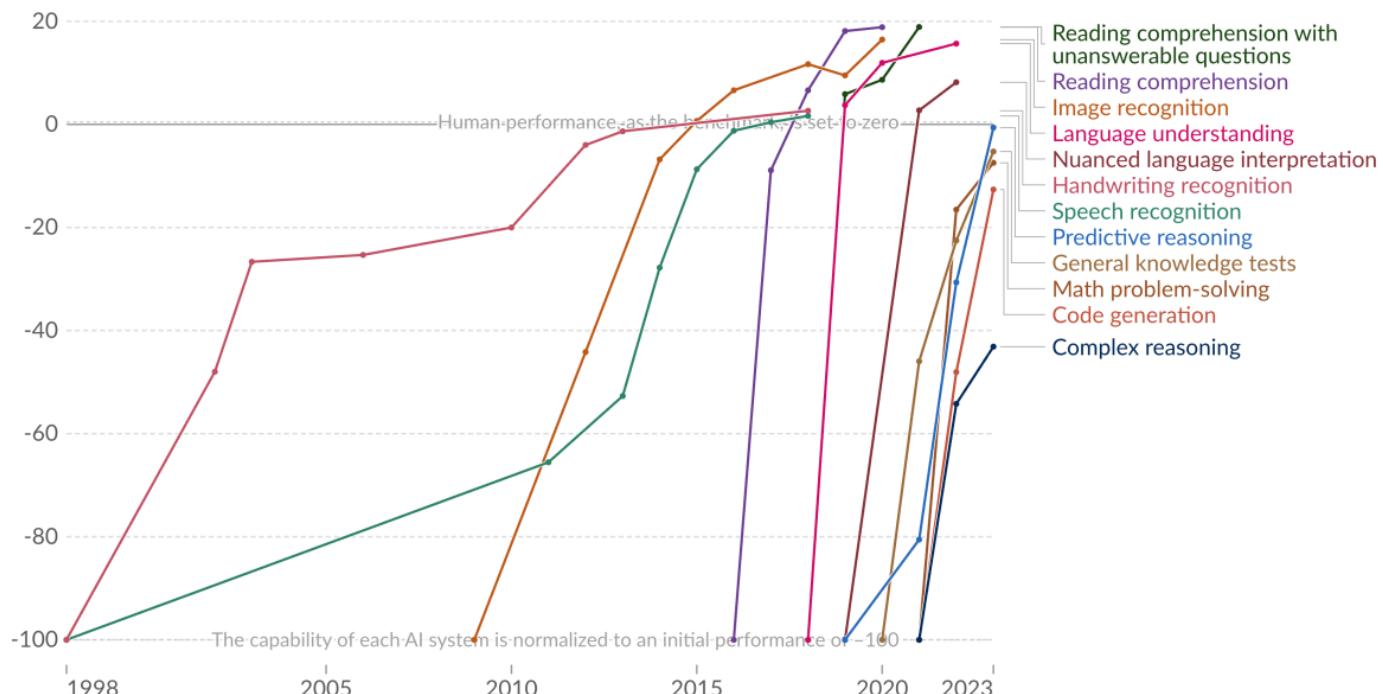
<https://github.com/RiverGumSecurity/IntroAILabs>



Test scores of AI systems on various capabilities relative to human performance

Our World in Data

Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

OurWorldInData.org/artificial-intelligence | CC BY

[Some images created with the assistance of DALL·E 3]

Performance on common exams
(percentile compared to human test-takers)

	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

<https://situational-awareness.ai/>

13

<https://github.com/RiverGumSecurity/IntroAI Labs>

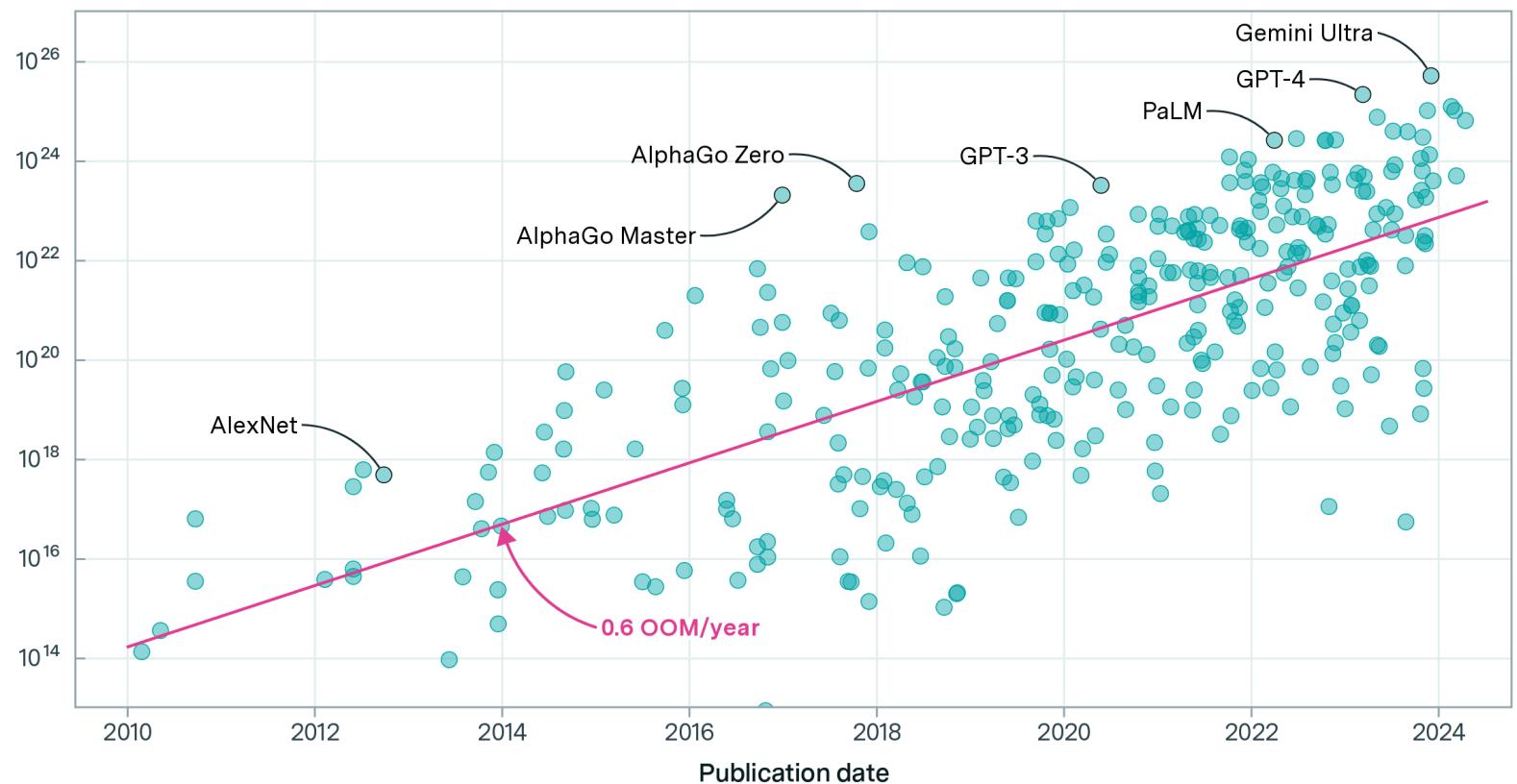
[Some images created with the assistance of DALL·E 3]



Training compute of notable models



Training compute (FLOP)



<https://situational-awareness.ai/>

[Some images created with the assistance of DALL·E 3]

14

<https://github.com/RiverGumSecurity/IntroAI Labs>



The State of AI-LLMs

- “*Current frontier models like Llama 3 are trained on the internet—and the internet is mostly crap, like e-commerce or SEO or whatever.*”
- “*Many LLMs spend the vast majority of their training compute on this crap, rather than on really high-quality data (e.g. reasoning chains of people working through difficult science problems).*”
- “*Imagine if you could spend GPT-4-level compute on entirely extremely high-quality data—it could be a much, much more capable model.*”

<https://situational-awareness.ai/>

[Some images created with the assistance of DALL·E ·3]

15

<https://github.com/RiverGumSecurity/IntroAIJobs>



Accelerating Future Concepts

- Orders of Magnitude (OOMs) in pursuit of acceleration
 - Best guesses as to upcoming changes that increase OOMs
 - Estimates between 2023 - 2027
- Compute
 - Likely a 2 – 3 OOM improvement in processing power, and system scaling
- Algorithmic Efficiency
 - Continuous improvement in algorithm design with likely 1 – 2 OOM improvement
- Unhobbling
 - Creative ways of removing model limitations and improving operational integration. Hard to quantify.

<https://situational-awareness.ai/>

[Some images created with the assistance of DALL·E ·3]

16

<https://github.com/RiverGumSecurity/IntroAI Labs>



2023-2027 (Projection)

Compute

2-3 OOMs

Algorithmic
Efficiency

1-3 OOMs

Unhobbling

? OOMs

Onboarding problem
Test-time compute/
System II
Using a computer

3-6 OOMs (best guess: ~5 OOMs) of base scaleup

Chatbot to Agent

Based on public estimates.

SITUATIONAL AWARENESS | Leopold Aschenbrenner

[Some images created with the assistance of DALL·E 3]

<https://situational-awareness.ai/>

17

<https://github.com/RiverGumSecurity/IntroAIJobs>

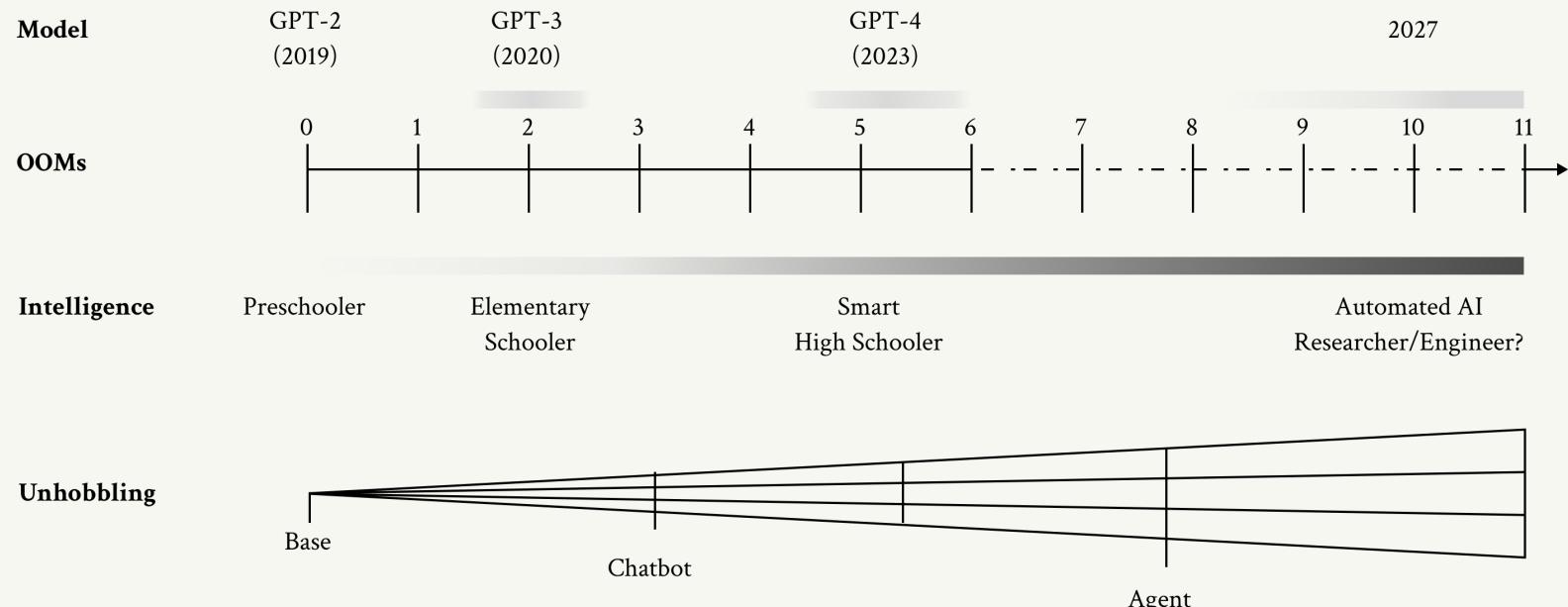


Acceleration Concepts Combined

- Today we have LLM's that are “smart high schoolers”
 - They ace the ACT, solve college level math, and accelerate tasks
- Adding the predicted Order of Magnitude improvements together
- Predicted intelligence might transition over time:
 - Smart High Schooler -> Knowledge driven PhD level researcher

[Some images created with the assistance of DALL·E ·3]

Counting the OOMs



SITUATIONAL AWARENESS | Leopold Aschenbrenner

<https://situational-awareness.ai/>

[Some images created with the assistance of DALL·E 3]

19

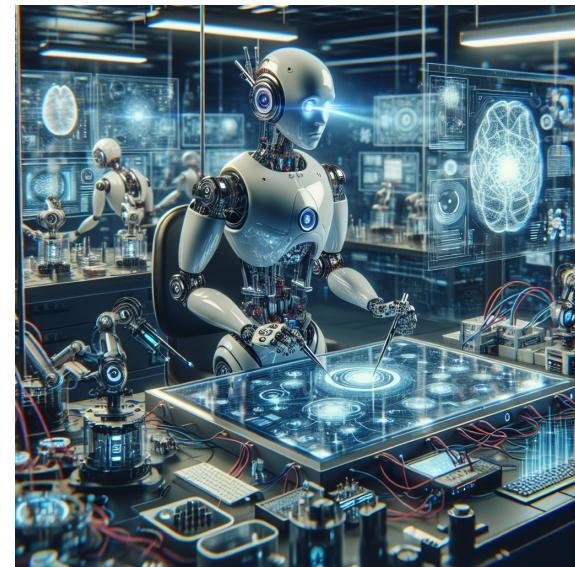
<https://github.com/RiverGumSecurity/IntroAIJobs>



The Future AGI on Researching AI!

“expect 100 million automated researchers each working at 100x human speed not long after we begin to be able to automate AI research”

“It’s strikingly plausible we’d go from AGI to superintelligence very quickly, perhaps in less than one year.”



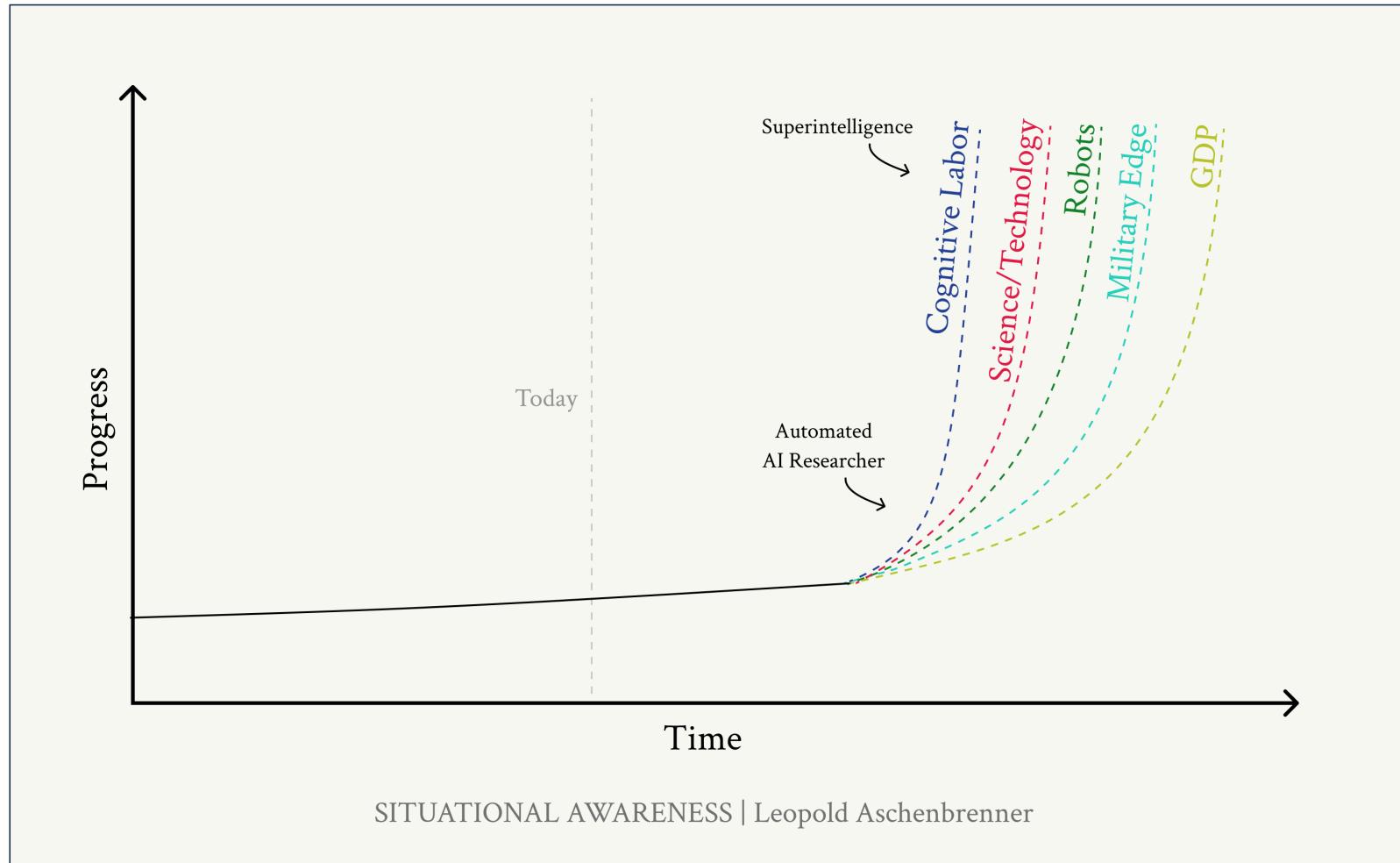
<https://situational-awareness.ai/>

[Some images created with the assistance of DALL·E ·3]

20

<https://github.com/RiverGumSecurity/IntroAI Labs>





<https://situational-awareness.ai/>

[Some images created with the assistance of DALL·E ·3]

21

<https://github.com/RiverGumSecurity/IntroAI Labs>



Sources of AGI Capabilities

green: benefits
yellow: limitations
red: risks

Idealized Models:
brute force,
policy gradient,
Darwinian evolution

creativity
targeting arbitrary goals/objectives
blind/local search
reward hacking

Selection/
Evolution

emergent drives / mesa-optimization

Long-range coherence makes LLMs
really useful for generating proposals
for brute-force searches. This also
makes policy gradient more effective.

Agent scaffolding combined with
RL in complex environments leads to
increasingly robust reasoning
and planning capabilities.

As the model becomes more capable
of solving problems through
search and reasoning, human
imitation becomes less relevant.

Idealized Model:
Kolmogorov complexity /
universal induction

Compression/
Imitation

human common sense
factual knowledge
legibility
reliance on finite data
poor OOD generalization
plausible hallucinations
imitation of harmful data

Idealized Model:
optimal sequential
decision making

Reasoning/
Planning

complex problem solving
competent OOD generalization
safe exploration
corrigibility (maybe)
scheming
instrumental convergence
treacherous turns

We are focused here today.

Over here is where we really
have alignment/risk
challenges.

Will humans matter at
all when we get here?

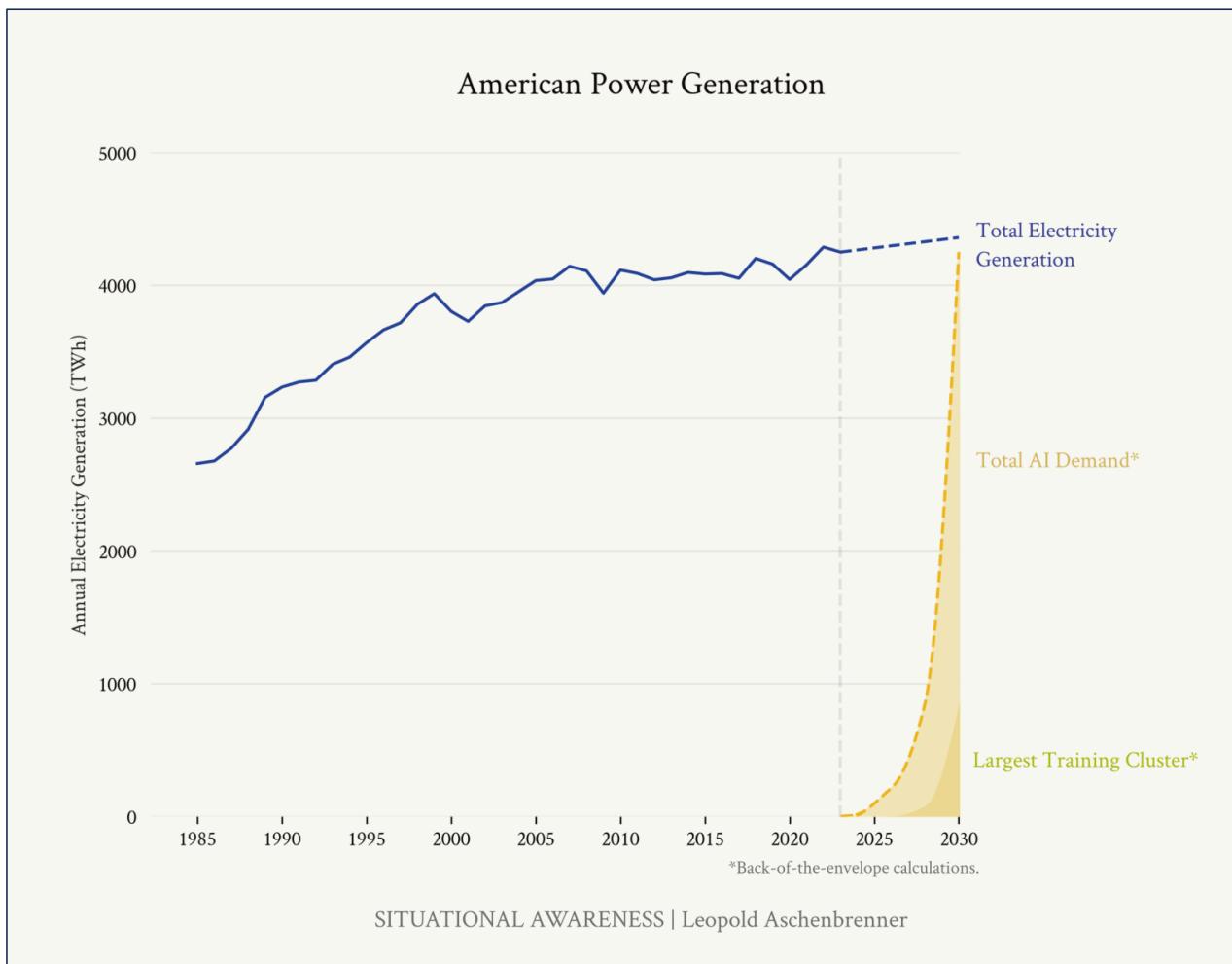
<https://x.com/RogerGrosse/status/1758506017791279440>

22

<https://github.com/RiverGumSecurity/IntroAI Labs>

[Some images created with the assistance of DALL·E 3]





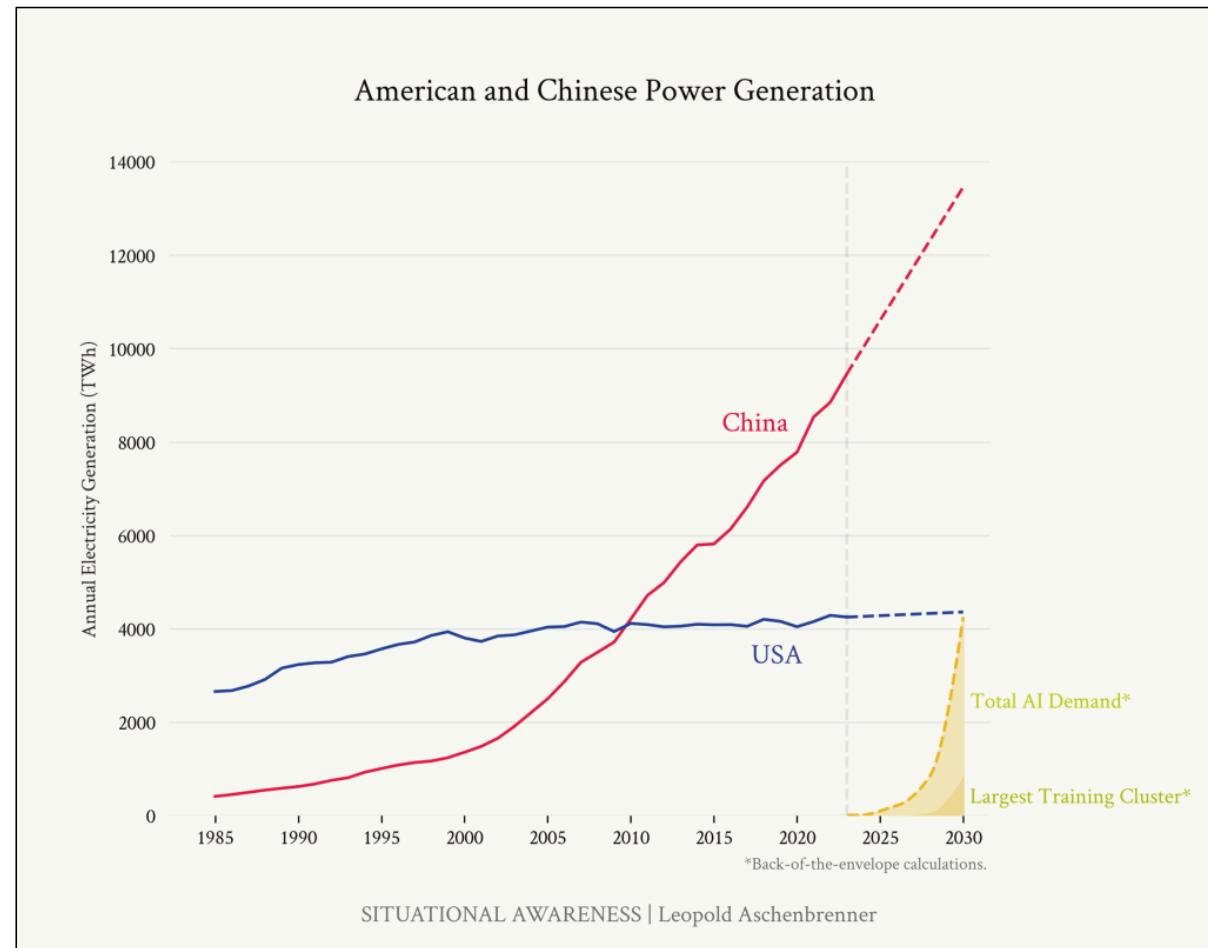
<https://situational-awareness.ai/>

23

[Some images created with the assistance of DALL·E 3]

<https://github.com/RiverGumSecurity/IntroAIJobs>





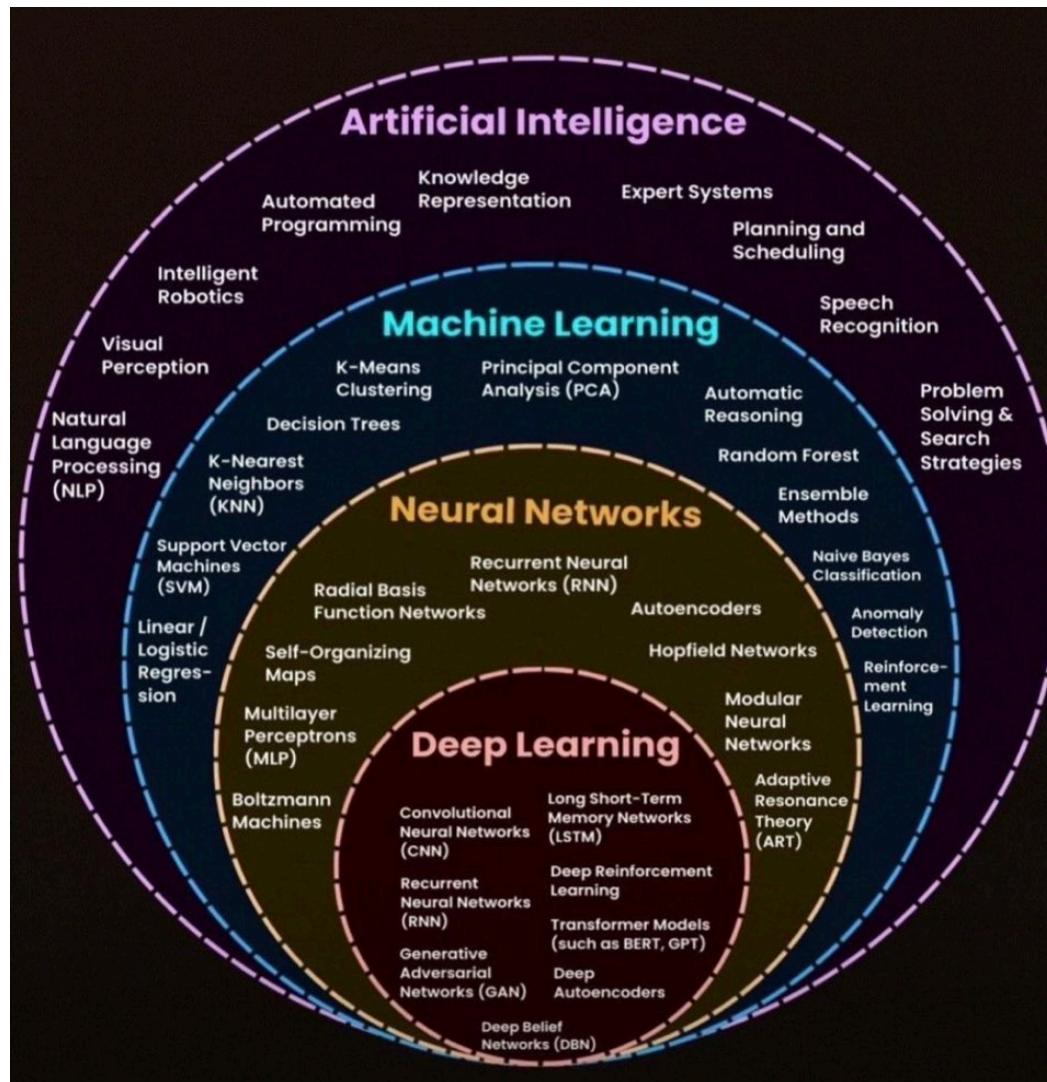
<https://situational-awareness.ai/>

24

[Some images created with the assistance of DALL·E 3]

<https://github.com/RiverGumSecurity/IntroAI Labs>





[Some images created with the assistance of DALL·E ·3]

25

<https://github.com/RiverGumSecurity/IntroAllabs>

<https://x.com/InterestingSTEM/status/1820019685875761252>

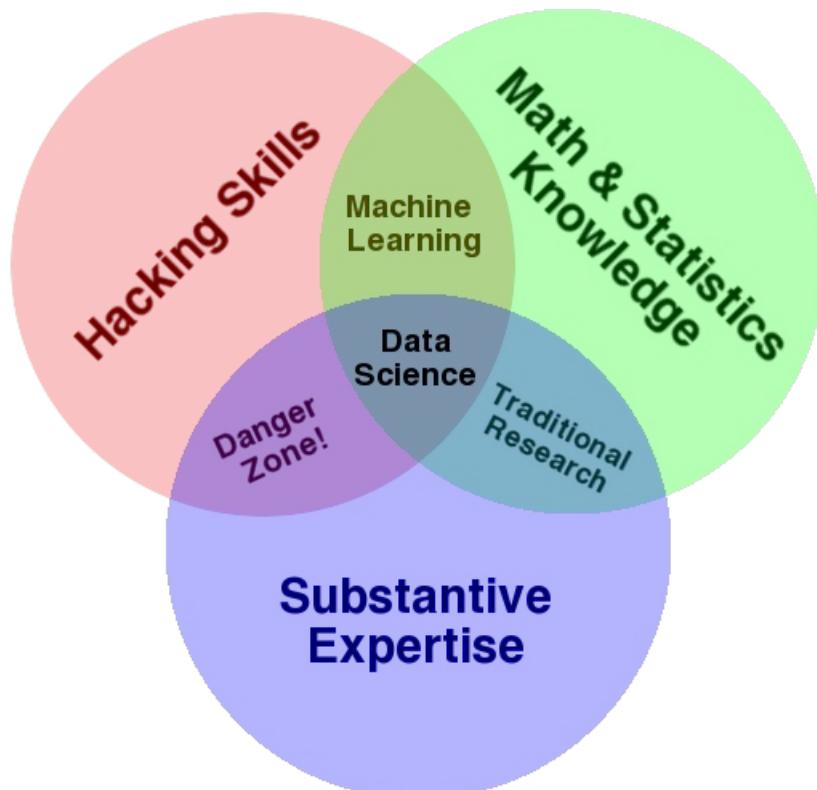


Data Science in Cyber Security

- Enhanced threat detection through ML/AI and statistical analysis
- Efficient detection, monitoring, and incident response
 - Ex: Risk Based Alerting (RBA) using aggregate risk scores to reduce alert fatigue
- Increased understanding of attack vectors
 - Data science tools and techniques provide enhanced views of attack patterns improving analyst understanding
- Improved Security Policies
 - Understanding the underlying data informs better decision making

[Some images created with the assistance of DALL·E ·3]

AI/ML/Data Science and Hackers



[Some images created with the assistance of DALL·E 3]

27

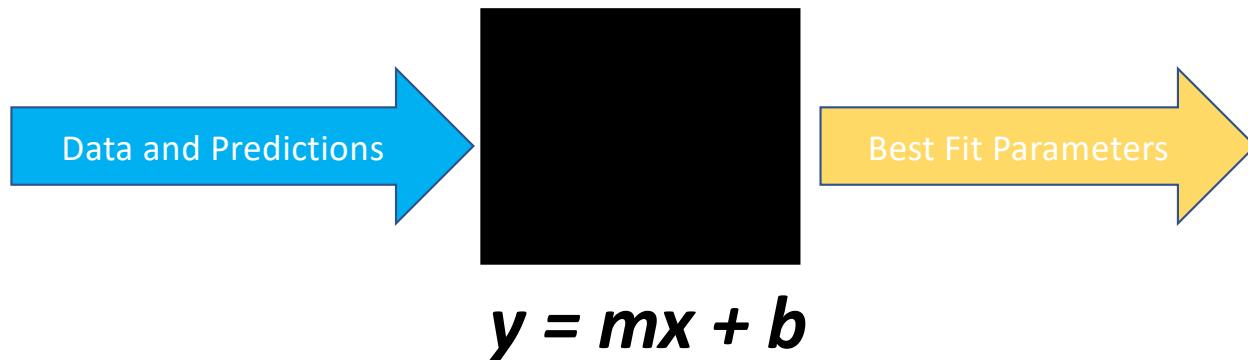
<https://github.com/RiverGumSecurity/IntroAI Labs>



Machine Learning

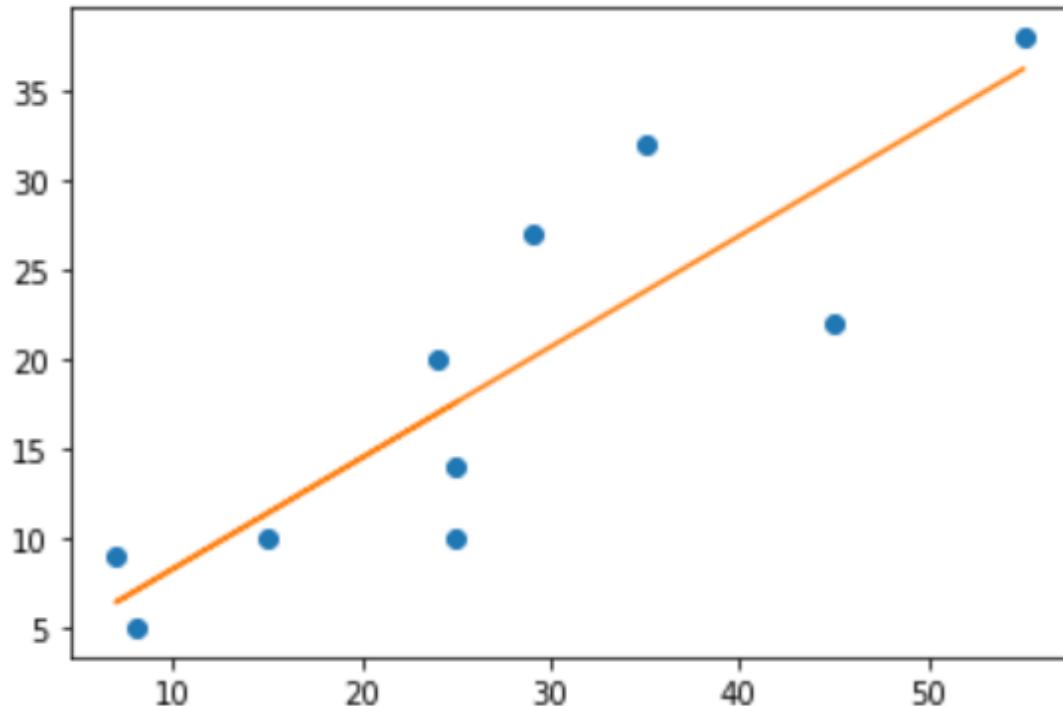
Let's start by telling the truth: machines don't learn."

- Burkov, Andriy. *The Hundred-Page Machine Learning Book* (p. xvii). Andriy Burkov



[Some images created with the assistance of DALL·E ·3]

Machine Learning

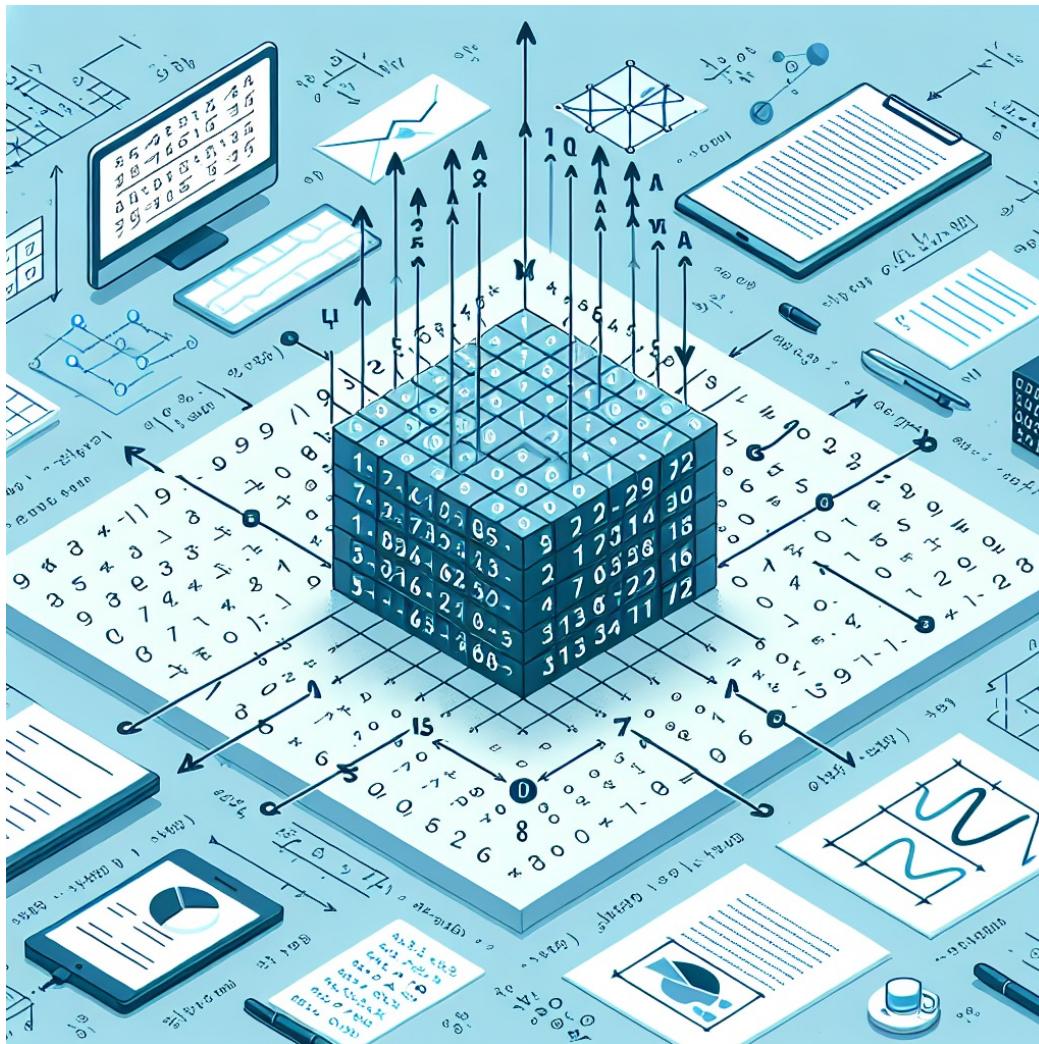


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$

Diagram illustrating the components of a linear regression equation:

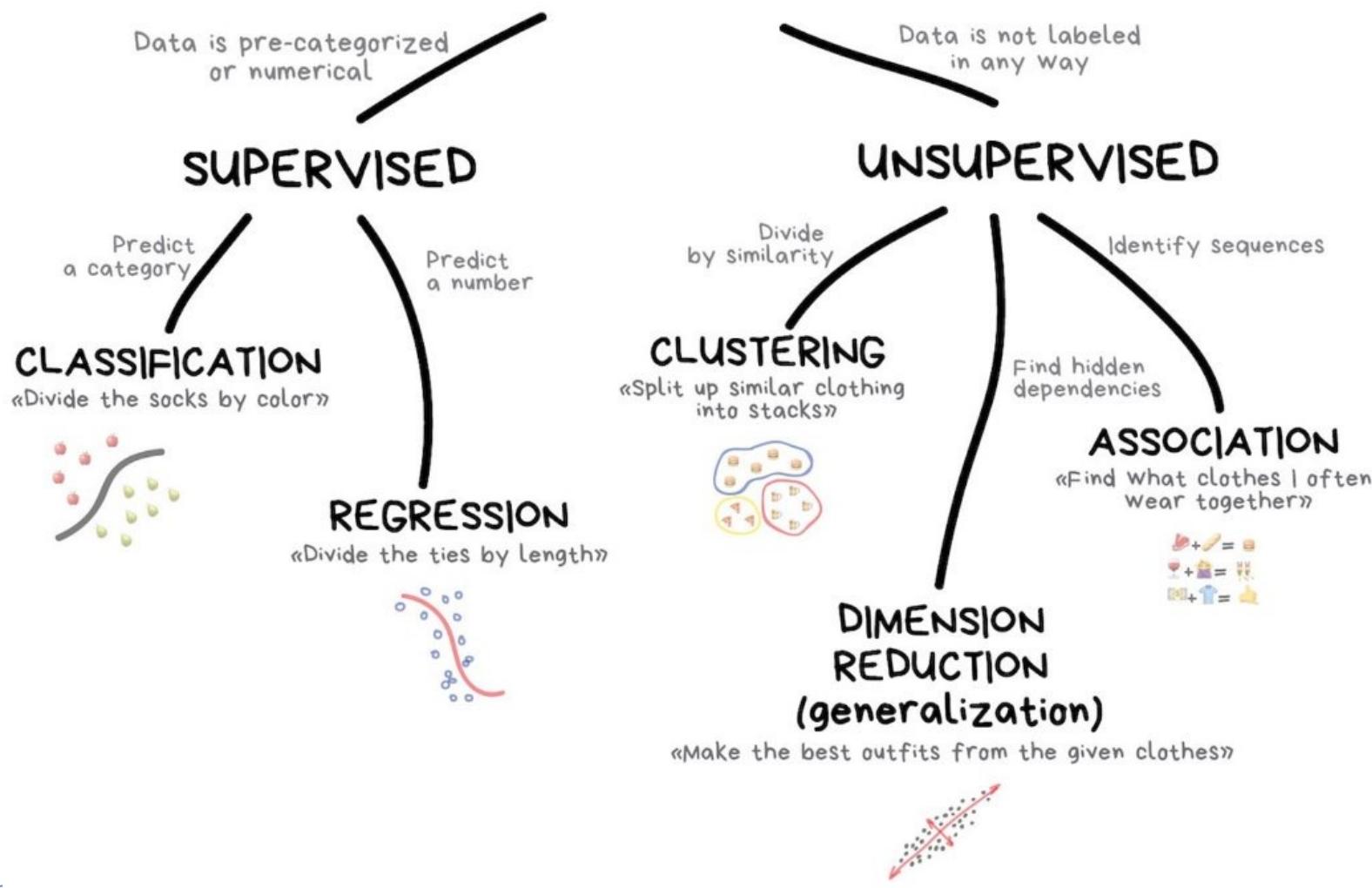
- Dependent Variable (Response Variable) points to Y
- Independent Variables (Predictors) points to X_1, X_2, \dots
- β_0 is labeled Y intercept
- β_1 is labeled Slope Coefficient
- ε is labeled Error Term

[Some images created with the assistance of DALL·E 3]



[Some images created with the assistance of DALL·E 3]

CLASSICAL MACHINE LEARNING



[Sc

Supervised Machine Learning

- Uses labeled dataset to where each input data point has the corresponding correct output (also known as a label)
- Training process is where the algorithm “learns” to map inputs to the most efficient outputs by adjusting parameter values
- Training process evaluated by a validation data set scoring metrics used determine model performance
- Trained model then used to make predictions on unseen data
 - Data has to match expected parameters
- Cybersecurity examples:
 - Phishing vs. non-phishing emails
 - Malware detection

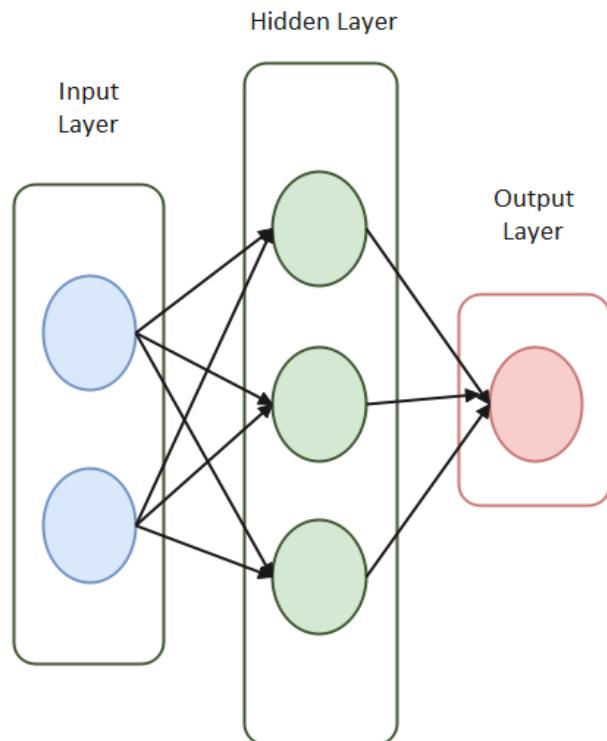
[Some images created with the assistance of DALL·E ·3]

Unsupervised Machine Learning

- Primary goal is to discover hidden patterns and structures in the dataset
- Uses unlabeled data with no predefined outcomes
- Useful for clustering, dimensionality reduction, and anomaly detection
- Cyber security examples:
 - Network security analysis
 - Fraud detection

[Some images created with the assistance of DALL·E ·3]

Neural Network



Deep Neural Network

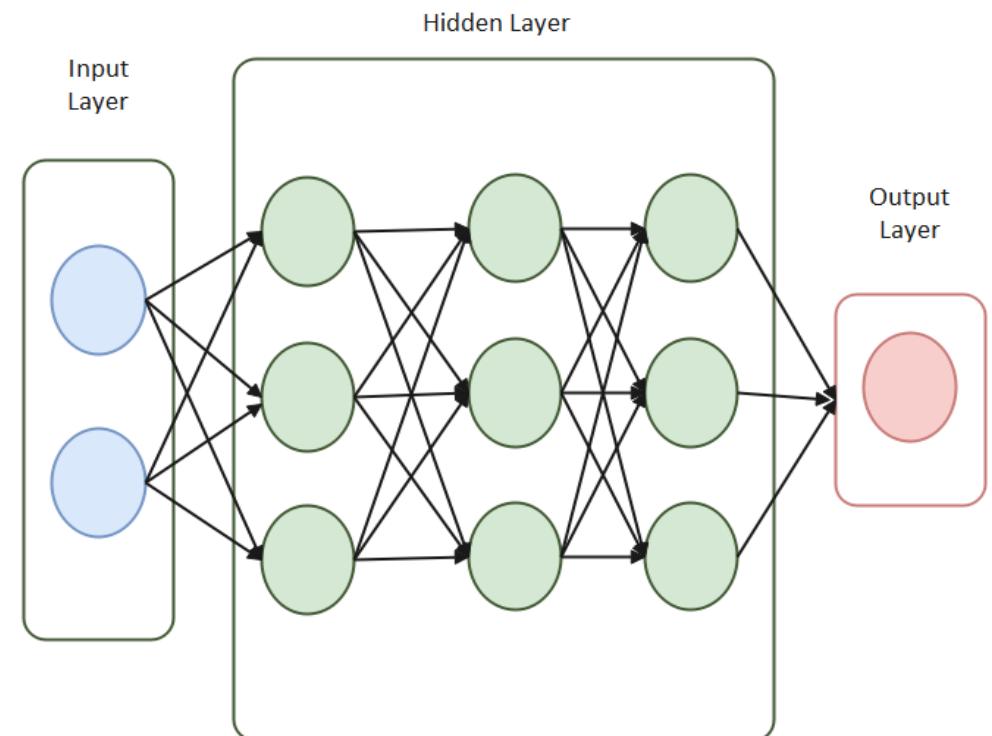


Image credit: <https://images.edrawsoft.com/articles/neural-network-diagram/example1.png>

[Some images created with the assistance of DALL·E ·3]

Neural Networks

- Inspired by the structure of the human brain
- Classic structure of a neural network includes an input layer, hidden layers, and an output layer
- Training involves adjusting the weights of connections between neurons
- Deep learning is a type of neural network that typically involves three or more layers of neurons

[Some images created with the assistance of DALL·E ·3]

Large Language Models

- AI system trained on vast amounts of text data on huge computing clusters
- Typically based on transformer models with billions of parameters
- Can perform various tasks like translation, summarization, question-answering, and text generation
- Uses “self-supervised” learning on diverse text sources to capture language patterns and context

[Some images created with the assistance of DALL·E ·3]

Prompt Engineering

- Practice of designing effective prompts to guide large language models in generating relevant and useful outputs
- Properly crafted prompts can significantly improve the quality of a model's response
- Techniques include iterative refinement, balancing prompt detail, and using instructional, contextual, or open-ended prompts

LLM Terminology

- **Token:** Unit of text that the model uses to process and generate language
 - Words, word-stems, characters, punctuation
- **Context Window:** maximum amount of text (measured in tokens) that the model can consider at one time when processing or generating language
- **Agent:** system or entity that is capable of performing tasks autonomously

[Some images created with the assistance of DALL·E ·3]

“fabric” – Augmenting Humans using AI

- Daniel Miessler -> helping us integrate AI
 - <https://github.com/danielmiessler/fabric>
- Mission statement: “Human Flourishing via AI Augmentation”
 - Think of “fabric” as a front-end helper to LLM interaction
 - Magnifying / enhancing the human elements NOT replacing!
- A human driven approach in his workflow
 - What problems are we trying to solve?
 - Some of these human challenges:
 - Too much content to ingest
 - Forgetting content watched
 - How do I focus on the right take home points from content?

[Some images created with the assistance of DALL·E ·3]



[Some images created with the assistance of DALL·E ·3]

<https://www.antisyphontraining.com/course/ai-for-cybersecurity-professionals-with-joff-thyer-and-derek-banks/>

The screenshot shows the Antisyphon website with the following details:

- Header:** Antisyphon logo, COURSE CATALOG, LIVE TRAINING, ON-DEMAND, WHO WE ARE, CERTIFICATION, CYBER RANGE, CONTACT, search icon, shopping cart, and user profile.
- Section:** AI for Cybersecurity Professionals with Joff Thyer and Derek Banks.
- Image:** A photo of two men, Joff Thyer and Derek Banks, wearing dark shirts with a "BLACK HILLS" logo.
- Text:** AI FOR CYBERSECURITY PROFESSIONALS, JOFF THYER, DEREK BANKS.
- Overview:** Course Length: 8 hours, Support from expert instructors, Includes certificate of completion.
- Image:** A stylized illustration of a robot head.
- Text:** Instructors: [Joff Thyer](#) and [Derek Banks](#).
- Logos:** River Gum Security logo (tree icon) and Black Hills Information Security logo (mountain icon).

[Some images created with the assistance of DALL·E 3]

Thanks!

- Twitter/X:
 - @joff_Thyer – Malware Pit Boss / ~~Chief Intern~~
 - @0xderuke – SOC Technology Wrangler
- LinkedIn
 - <https://www.linkedin.com/in/joffthyer/>
 - <https://www.linkedin.com/in/derek-banks-117b0012/>



[Some images created with the assistance of DALL·E ·3]

