AI in InfoSec? Lets talk about it…

Authors: Joff Thyer, and Derek Banks

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]

https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs

Historical Notes

- The field grew out of Neural Network research circa 1959
- Bernard Widrow and Marcian Hoff at Stanford
  - Developed ADALINE, and MADALINE
  - **M**ultiple **ADA**ptive **LIN**ear **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]

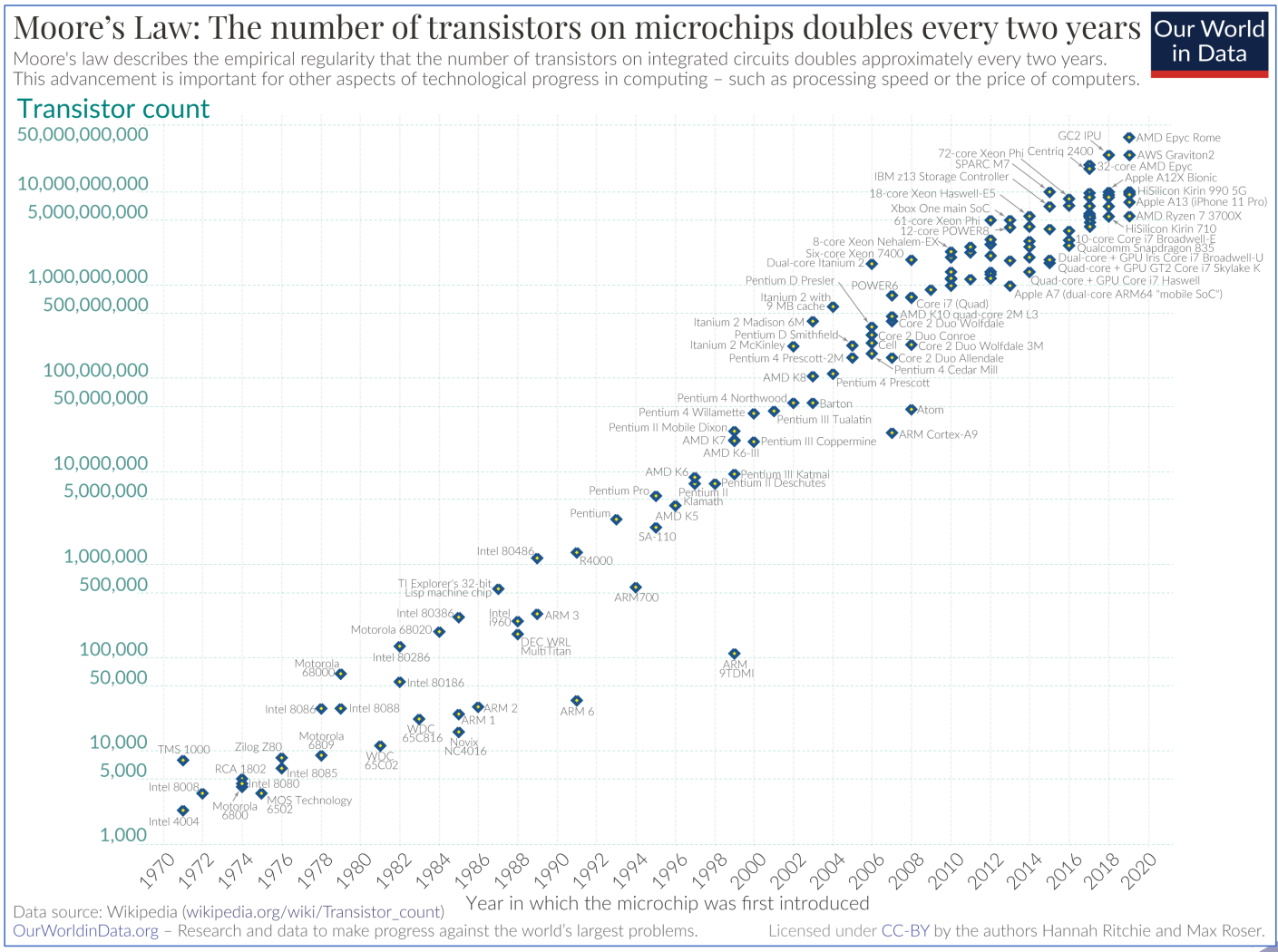
https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs



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

https://github.com/RiverGumSecurity/IntroAILabs

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://synthedia.substack.com/p/a-timeline-of-large-language-model

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

https://github.com/RiverGumSecurity/IntroAILabs

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://synthedia.substack.com/p/a-timeline-of-large-language-model

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

https://github.com/RiverGumSecurity/IntroAILabs

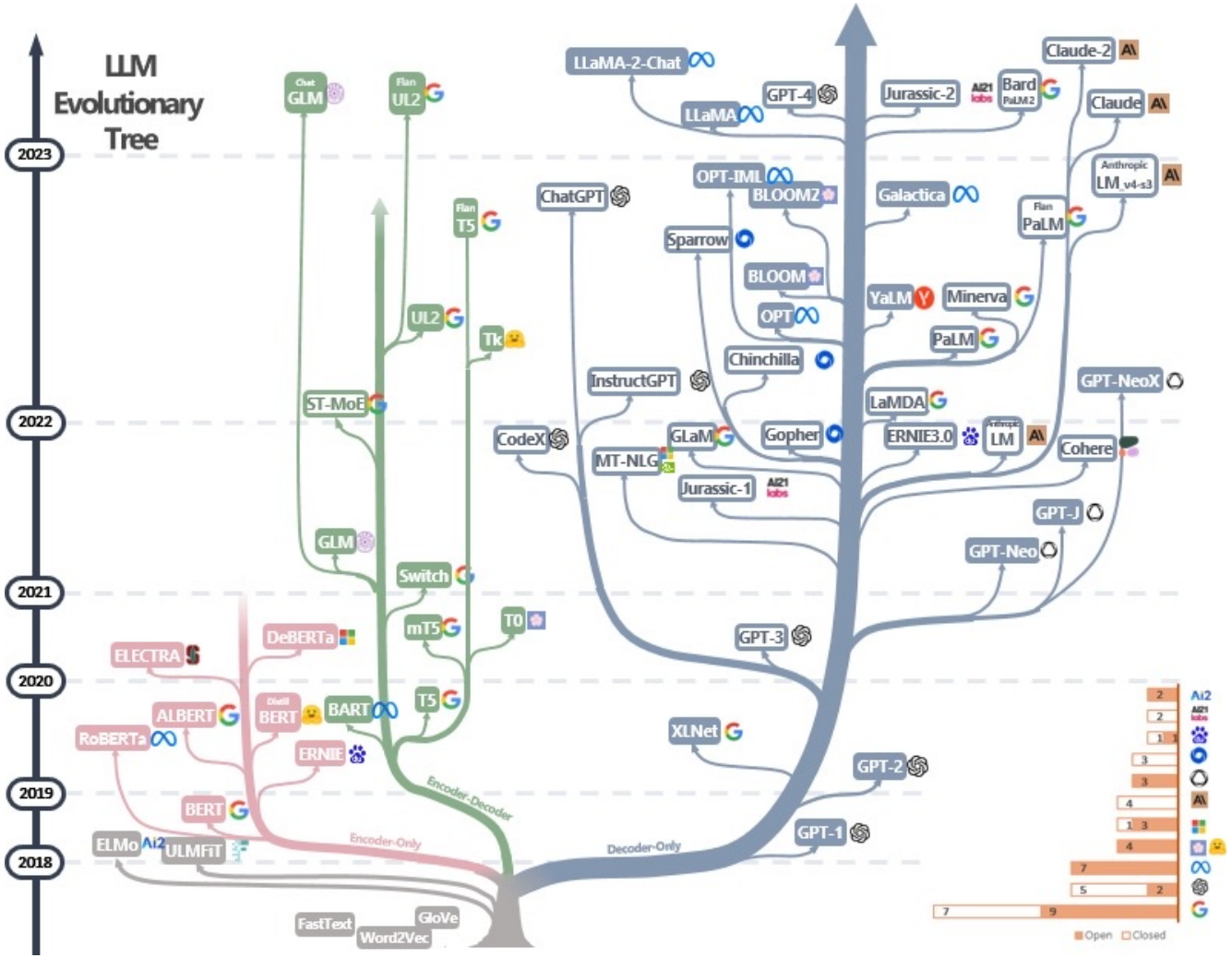
Historical Timeline…

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

Ref: https://synthedia.substack.com/p/a-timeline-of-large-language-model

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

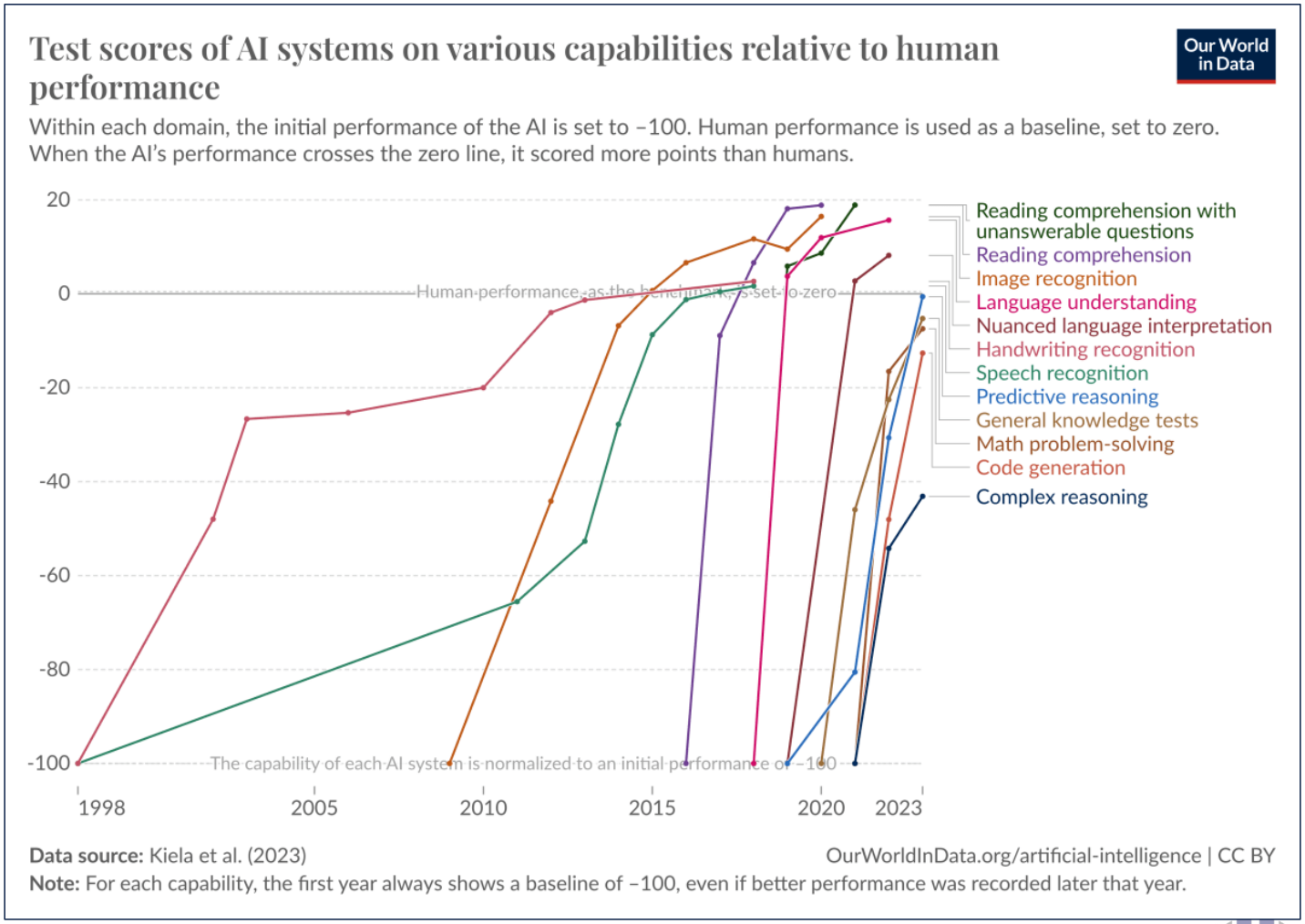
https://github.com/RiverGumSecurity/IntroAILabs



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

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

https://github.com/RiverGumSecurity/IntroAILabs



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

https://github.com/RiverGumSecurity/IntroAILabs

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

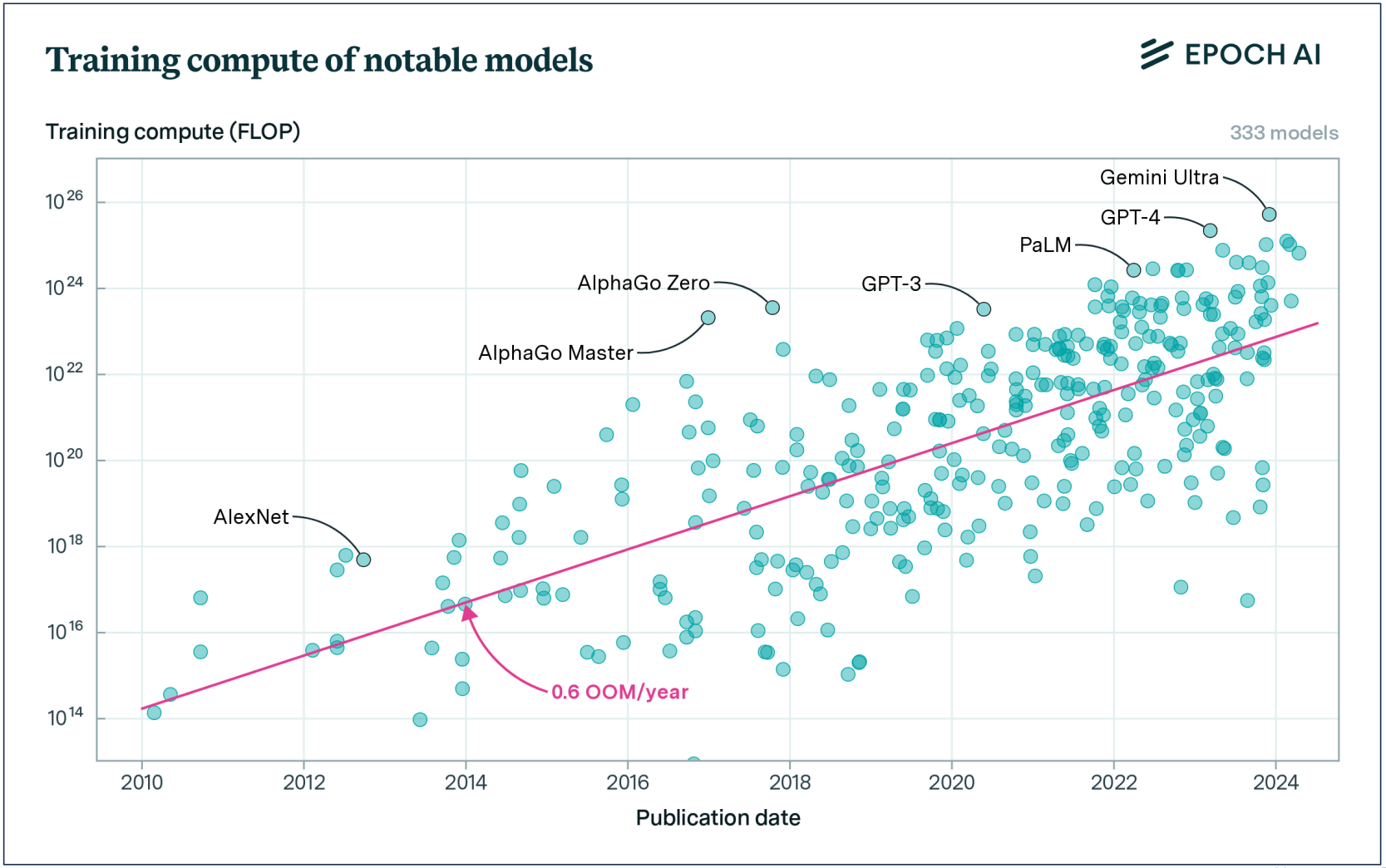
	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

https://situational-awareness.ai/

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

https://github.com/RiverGumSecurity/IntroAILabs



https://situational-awareness.ai/

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

https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs

2023-2027 (Projection)

Compute	Algorithmic Efficiency	Unhobbling
2-3 OOMs	1-3 OOMs	? 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

https://situational-awareness.ai/

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

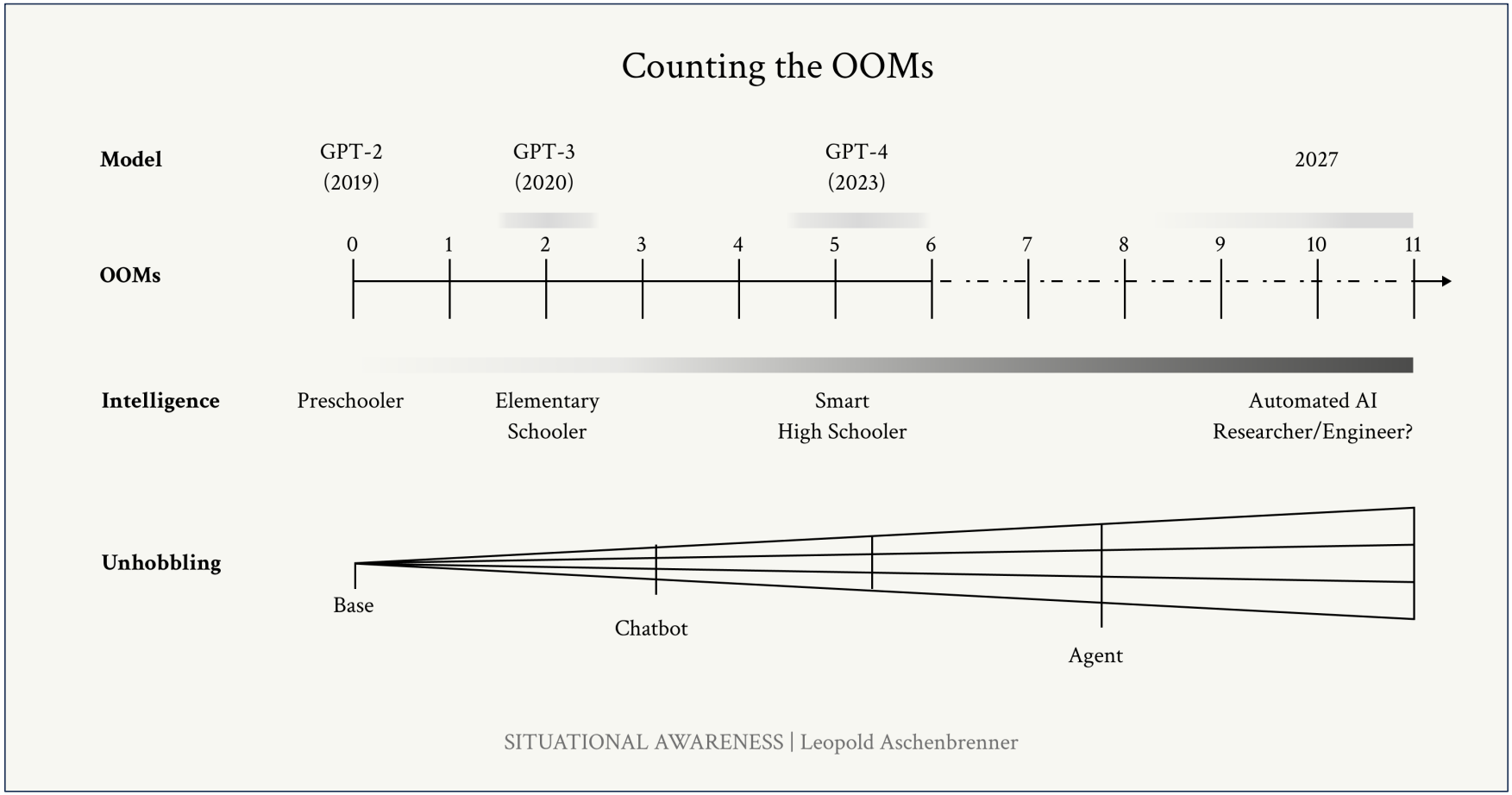
https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs



https://situational-awareness.ai/

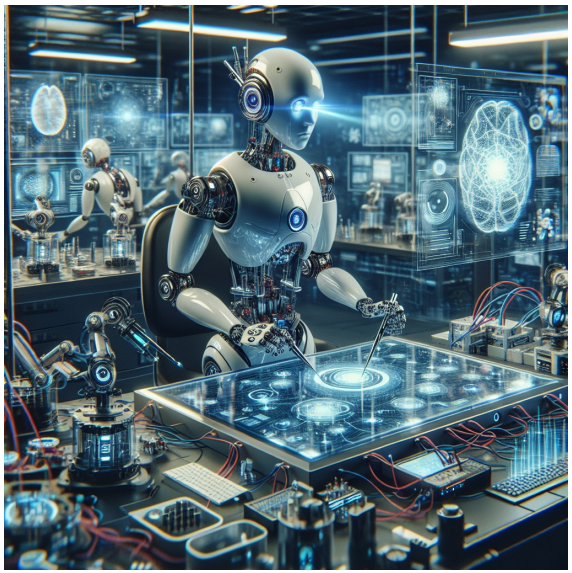
[Some images created with the assistance of DALL·E·3]

https://github.com/RiverGumSecurity/IntroAILabs

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"

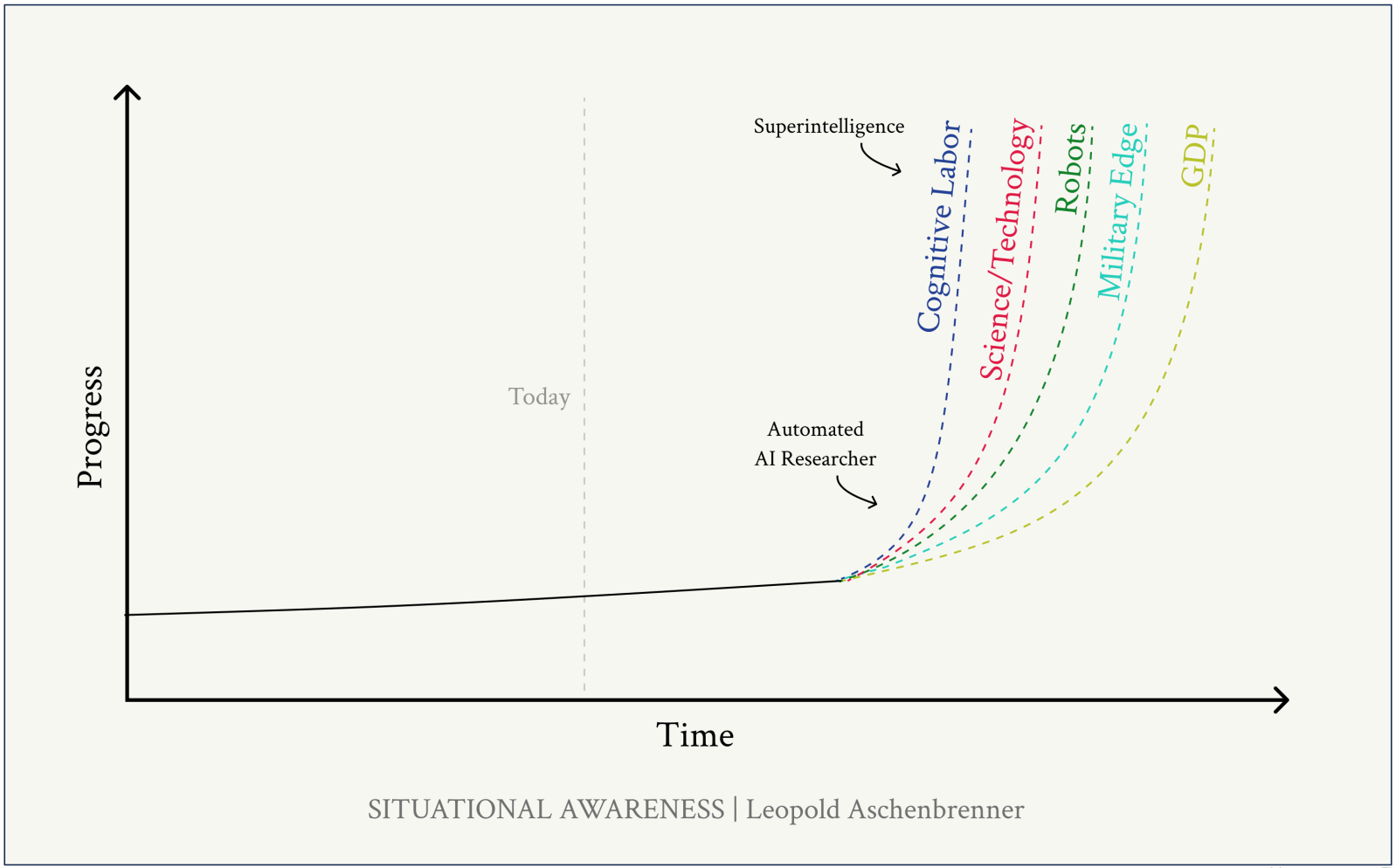
"It's strikingly plausible we'd go from AGI to superintelligence very quickly, perhaps in less than one year."



https://situational-awareness.ai/

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

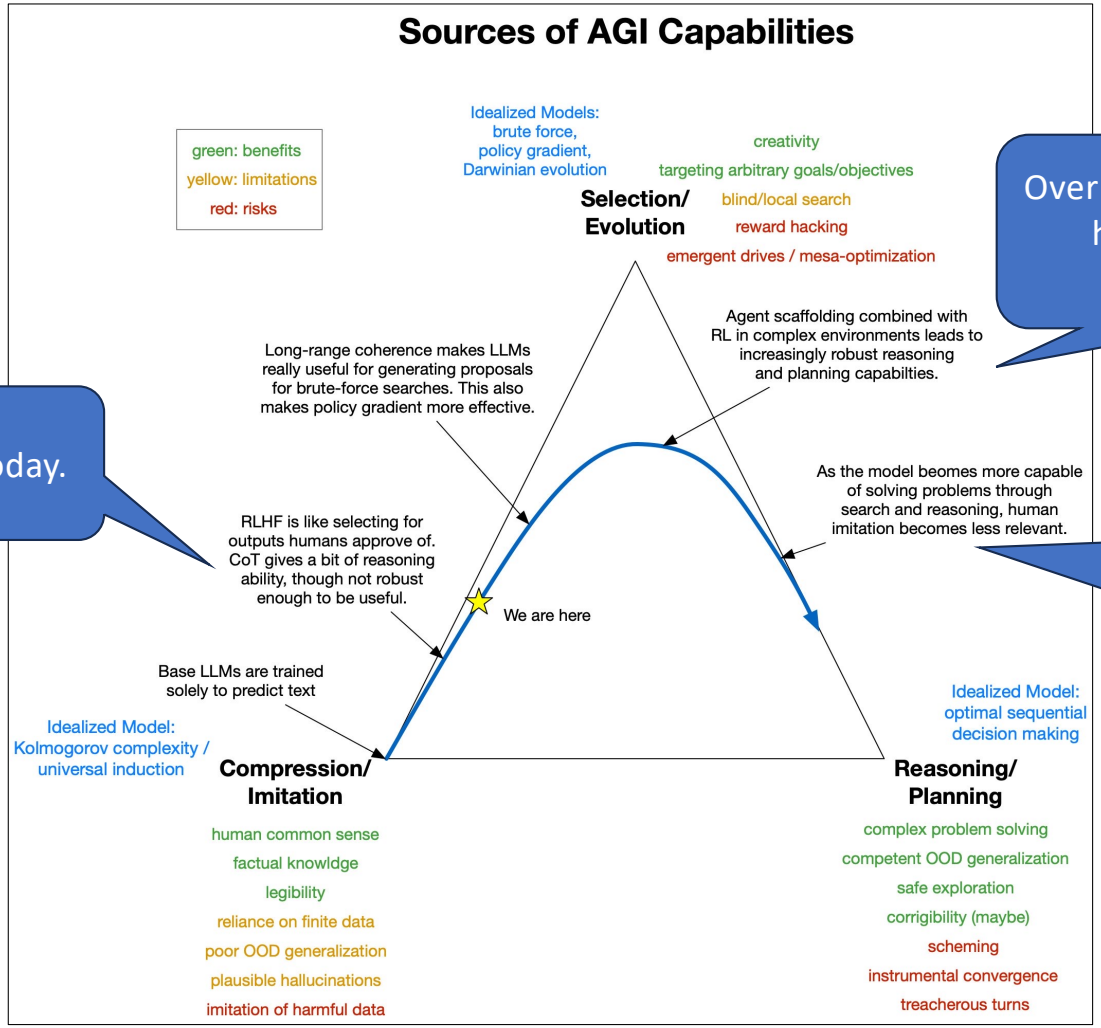
https://github.com/RiverGumSecurity/IntroAILabs



https://situational-awareness.ai/

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

https://github.com/RiverGumSecurity/IntroAILabs



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

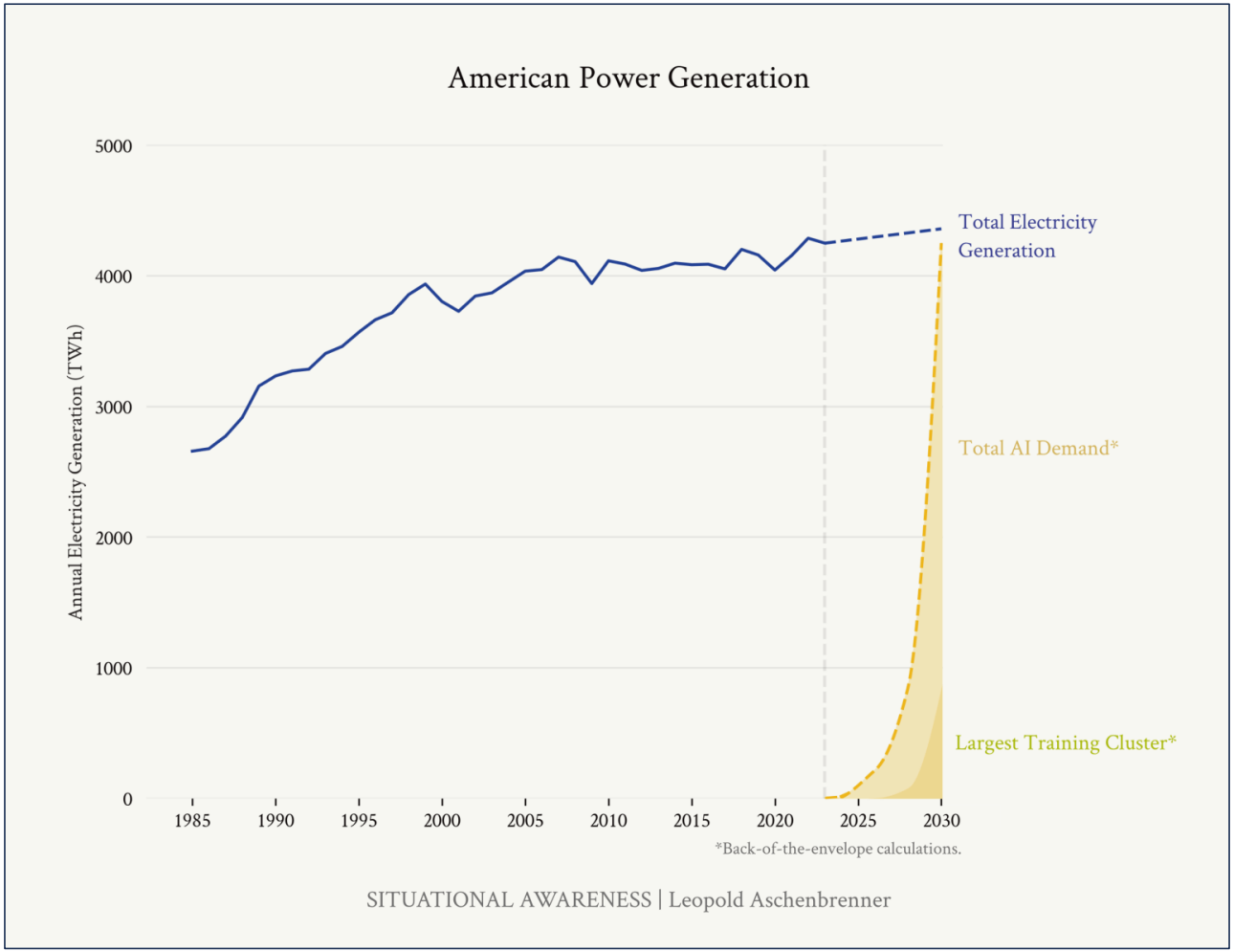
We are focused here today.

Will humans matter at all when we get here?

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

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

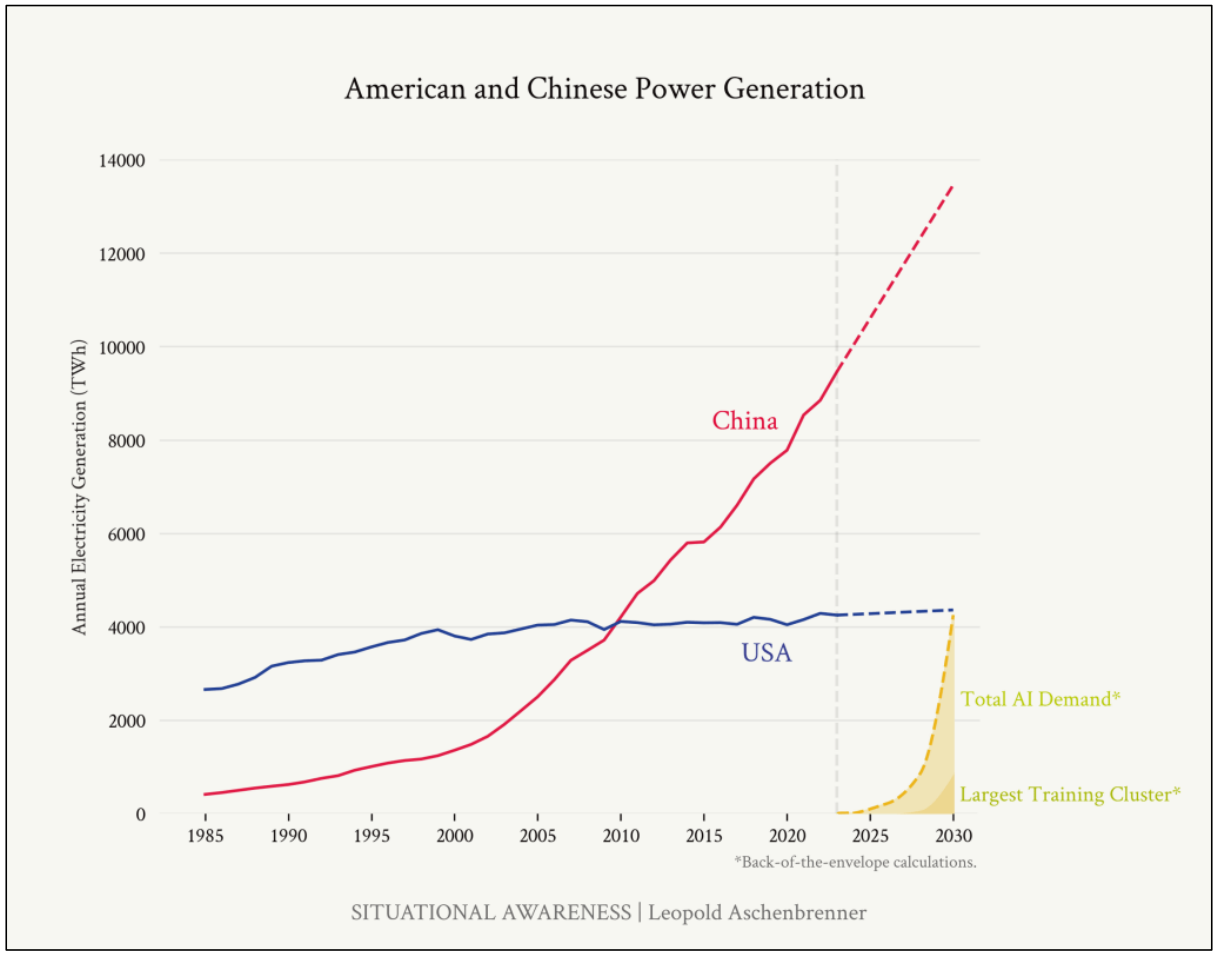
https://github.com/RiverGumSecurity/IntroAILabs



https://situational-awareness.ai/

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

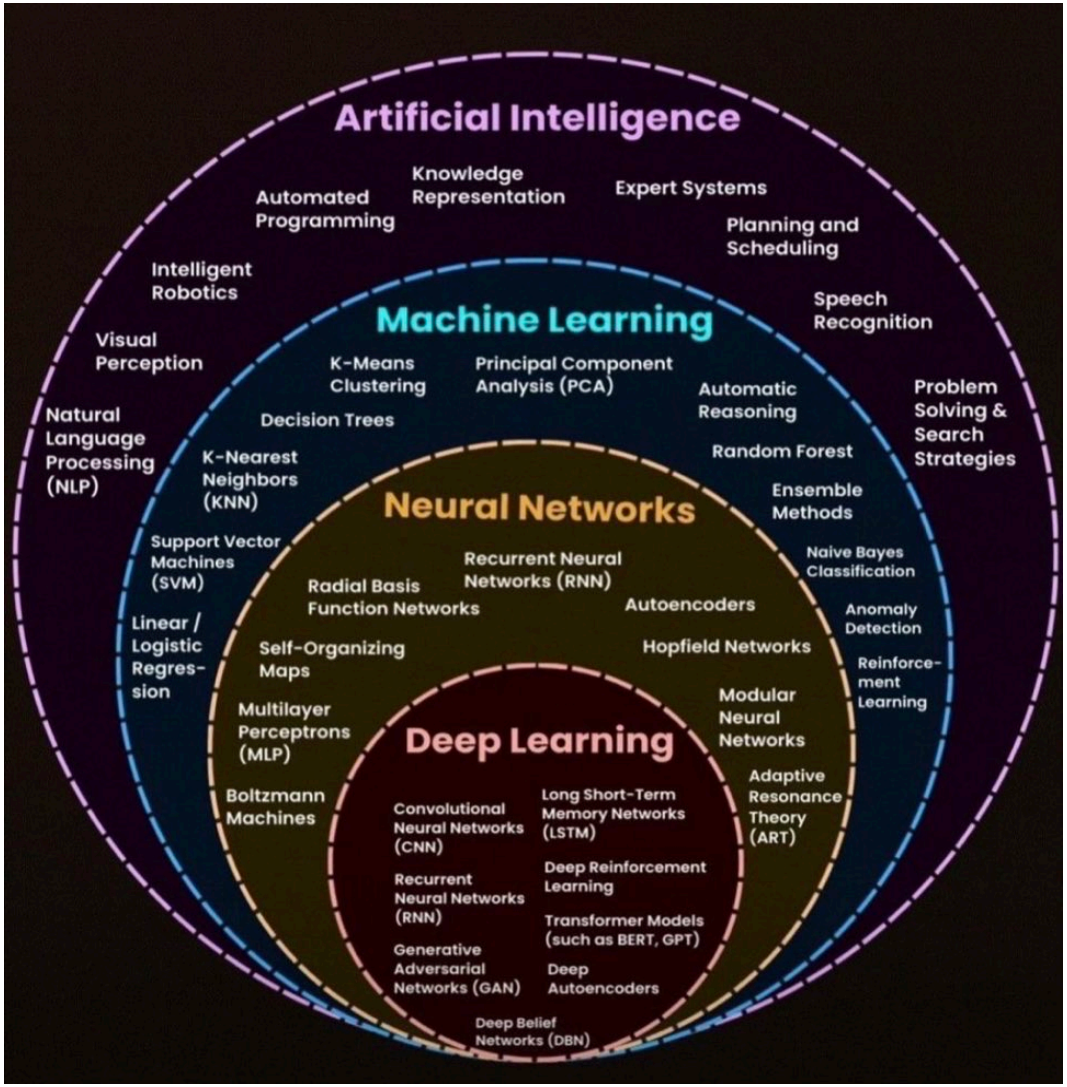
https://github.com/RiverGumSecurity/IntroAILabs



https://situational-awareness.ai/

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

https://github.com/RiverGumSecurity/IntroAILabs



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

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

https://github.com/RiverGumSecurity/IntroAILabs

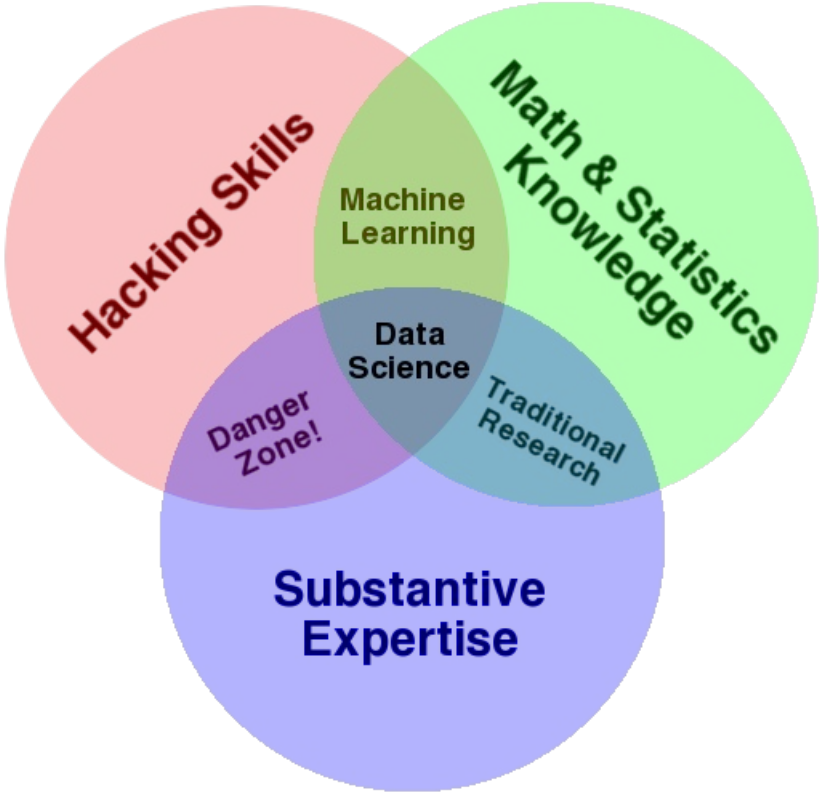
Data Science in Cyber Security

- 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]

https://github.com/RiverGumSecurity/IntroAILabs

AI/ML/Data Science and Hackers



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

https://github.com/RiverGumSecurity/IntroAILabs

Machine Learning

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

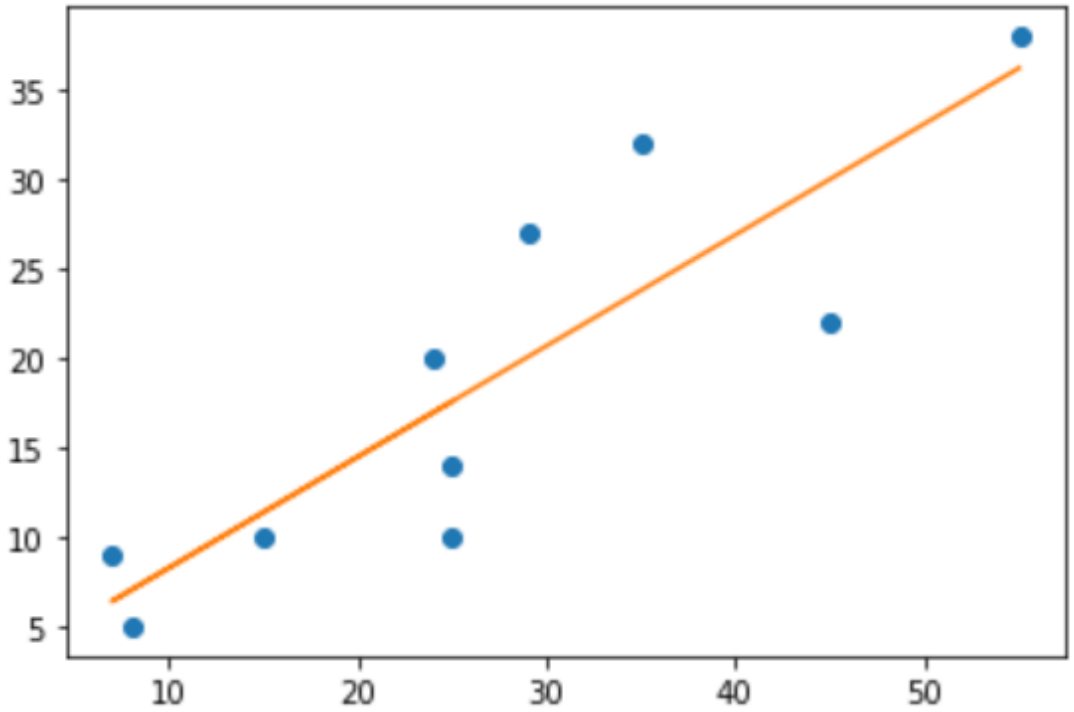
- Burkov, Andriy. The Hundred-Page Machine Learning Book (p. xvii). Andriy Burkov

$$y = mx + b$$

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

https://github.com/RiverGumSecurity/IntroAILabs

Machine Learning



Dependent Variable (Response Variable)

Independent Variables (Predictors)

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

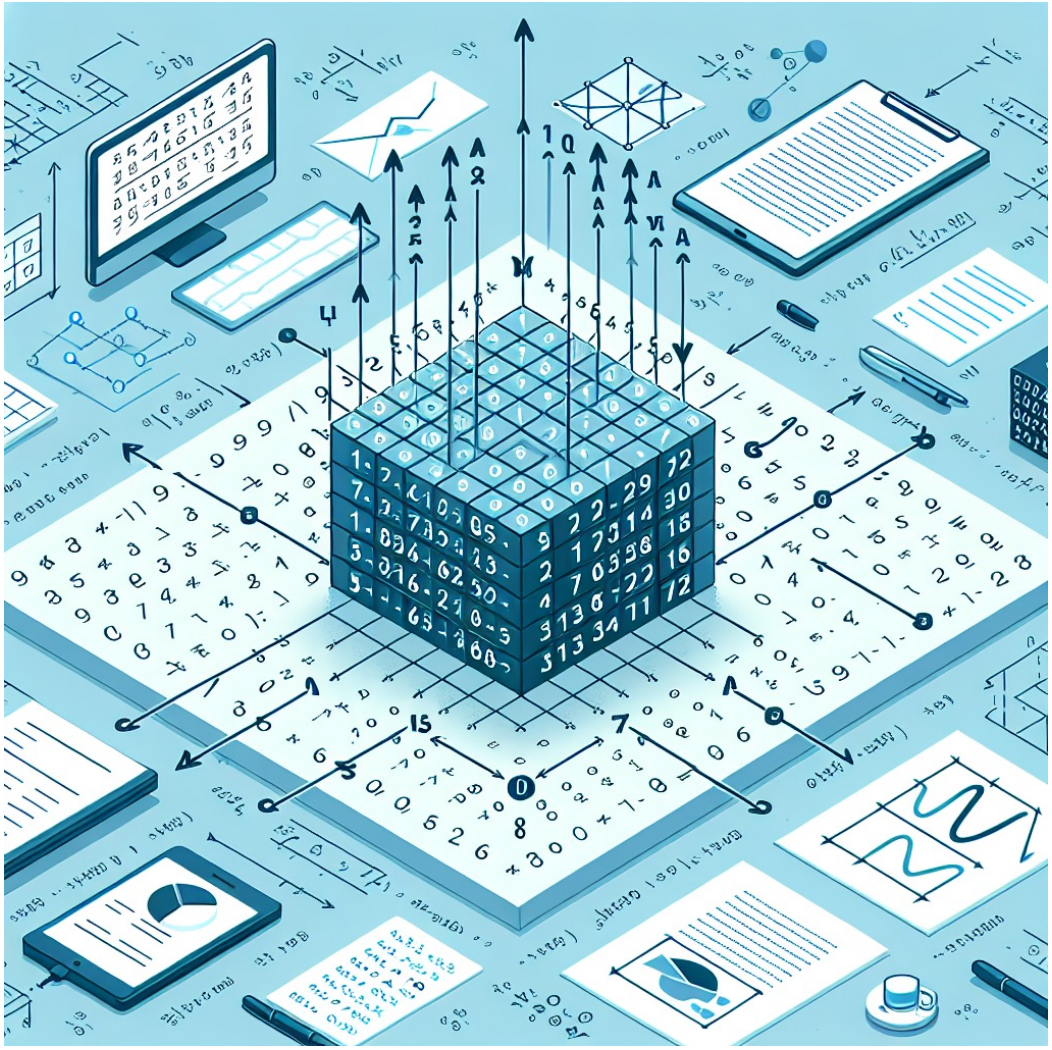
Y intercept

Slope Coefficient

Error Term

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

https://github.com/RiverGumSecurity/IntroAILabs



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

https://github.com/RiverGumSecurity/IntroAILabs

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»

UNSUPERVISED

Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»

ASSOCIATION

«Find what clothes I often wear together»

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

[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]

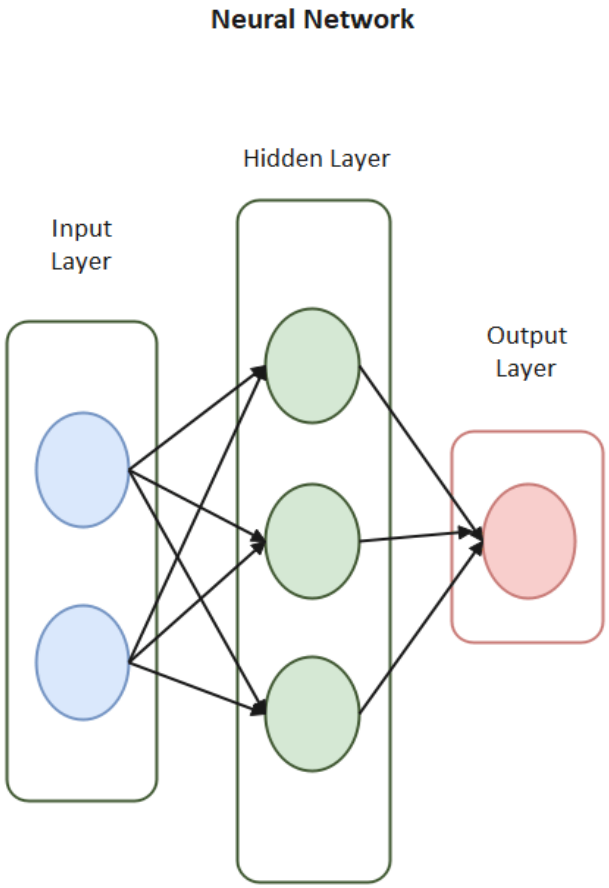
https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs



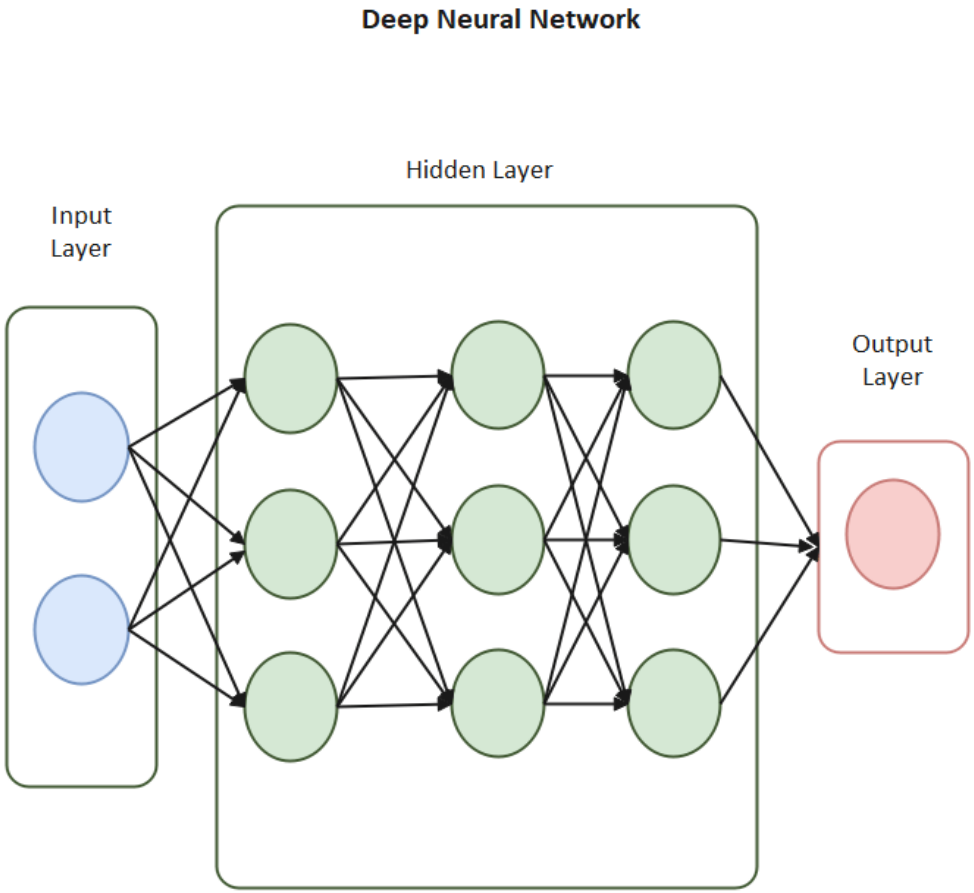


Image credit: https://images.edrawsoft.com/articles/neural-network-diagram/example1.png

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

https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs

Large Language Models

- AI system trained on vast amounts of text data on 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]

https://github.com/RiverGumSecurity/IntroAILabs

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

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

https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs

“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?

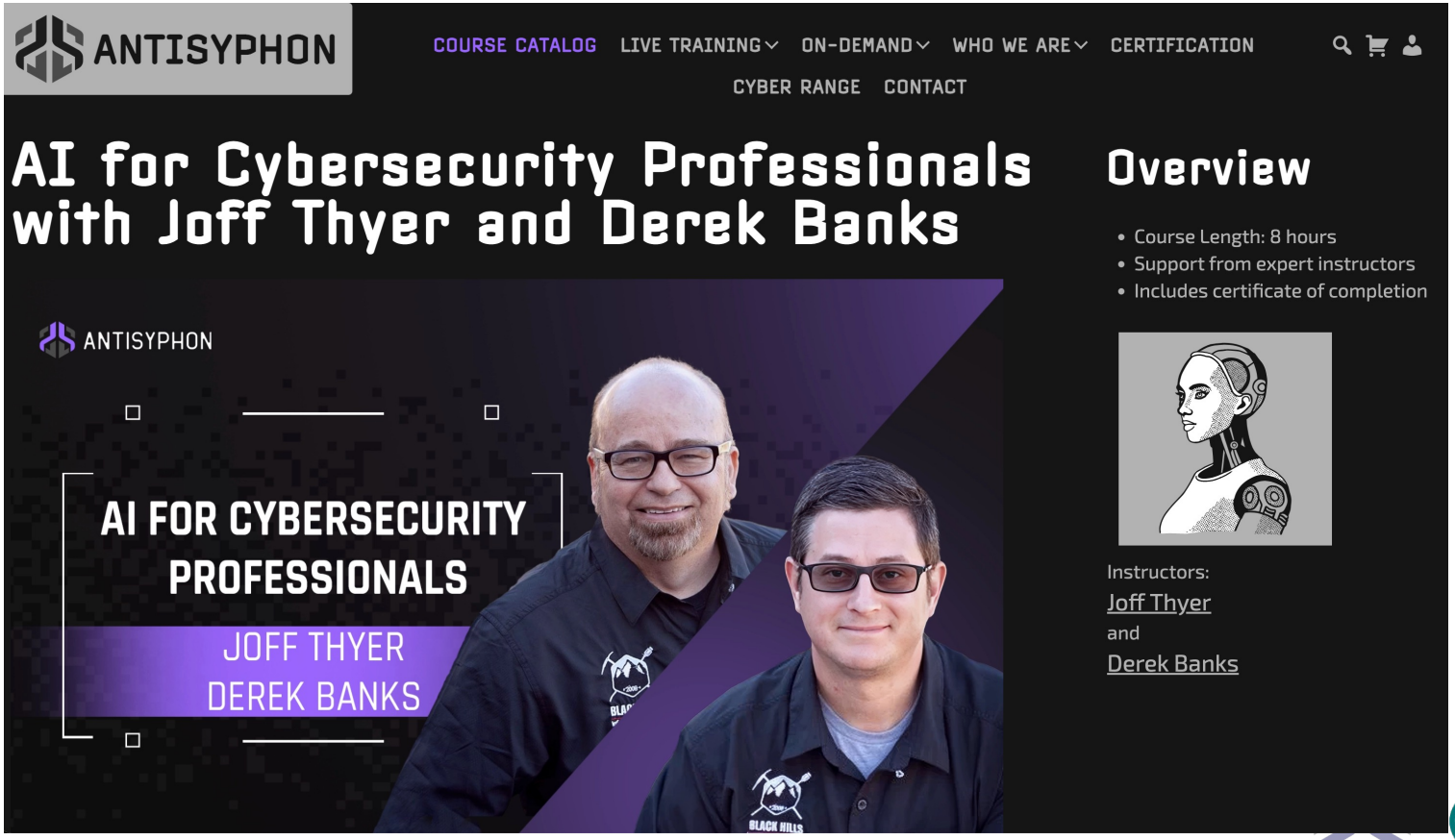
[Some images created with the assistance of DALL·E·3]

https://github.com/RiverGumSecurity/IntroAILabs

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

https://github.com/RiverGumSecurity/IntroAILabs

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



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

https://github.com/RiverGumSecurity/IntroAILabs

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]

https://github.com/RiverGumSecurity/IntroAILabs