EECS 4980/5980 Special Topics: Generative AI and LLMs

Introduction to Generative AI

Monday, August 26, 2024

Liang (Leon) Cheng, Ph.D., Professor

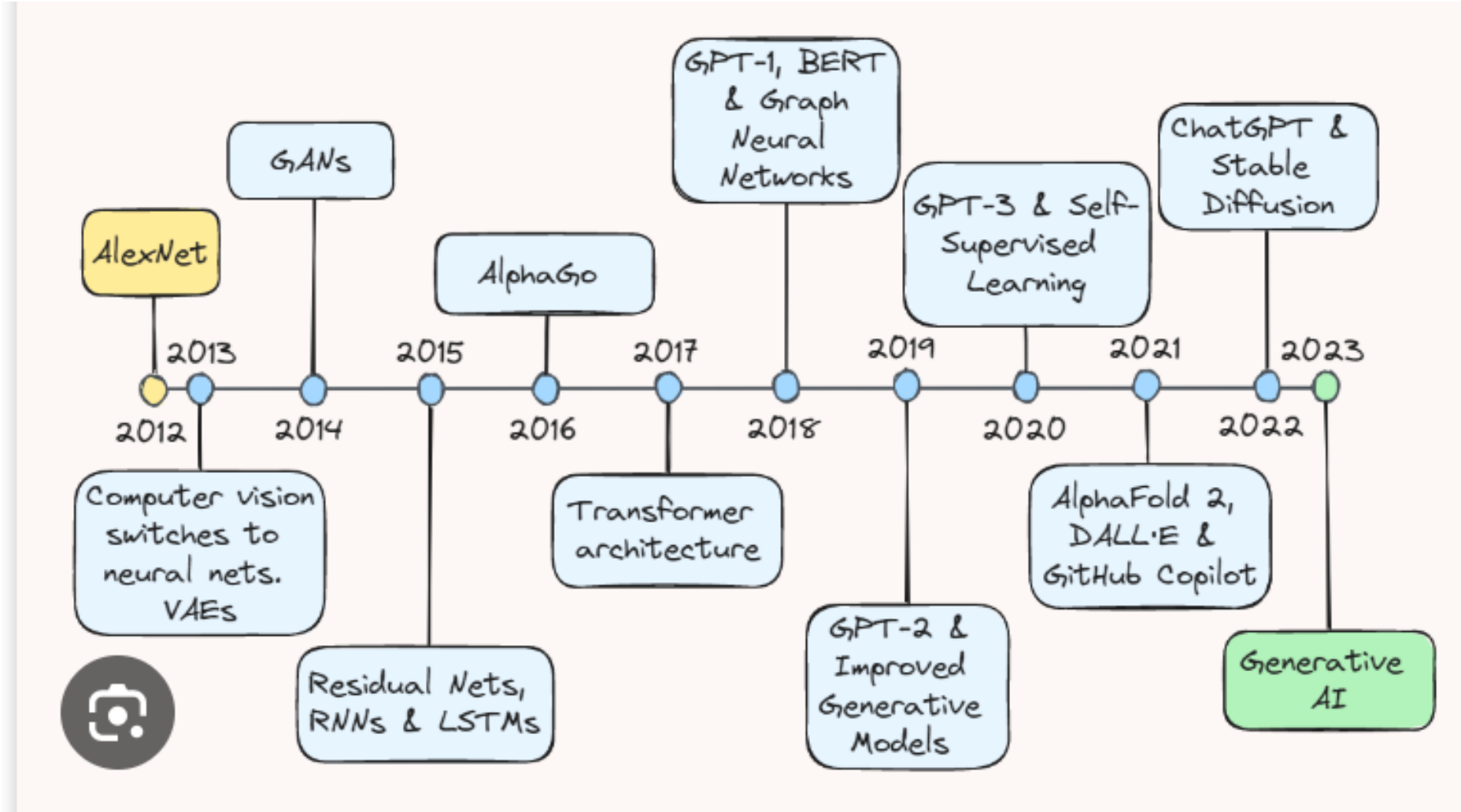
THE UNIVERSITY OF TOLEDO

AI Terms

- Artificial Intelligence: computer systems that emulate human capabilities.
- Machine Learning: computer systems that makes decisions by learning from data (supervised learning or unsupervised learning).
- Deep Learning: multi-layer learning through neural networks.
- Generative AI: AI capable of generating text, images, videos, or other data using generative models, often in response to prompts.
- Large Language Models (LLM's): ML's that use Deep Learning to process NLP (Natural Language Processing).

AI History

- Alan Turing – machine intelligence
- John McCarthy (1956) coins term “AI” at Dartmouth
  - Focus on encoding logic reasoning and rule-based predictive systems
  - Statistical models used to develop encoding systems
- AI Winter (1970’s) – results not being realized – grants down
- AI Revival (1980’s-1990’s) – expert systems & advanced algorithms
- AI Booming with computing power and data storage increase
  - from symbolic AI and rule-based systems to statistical and probabilistic approaches



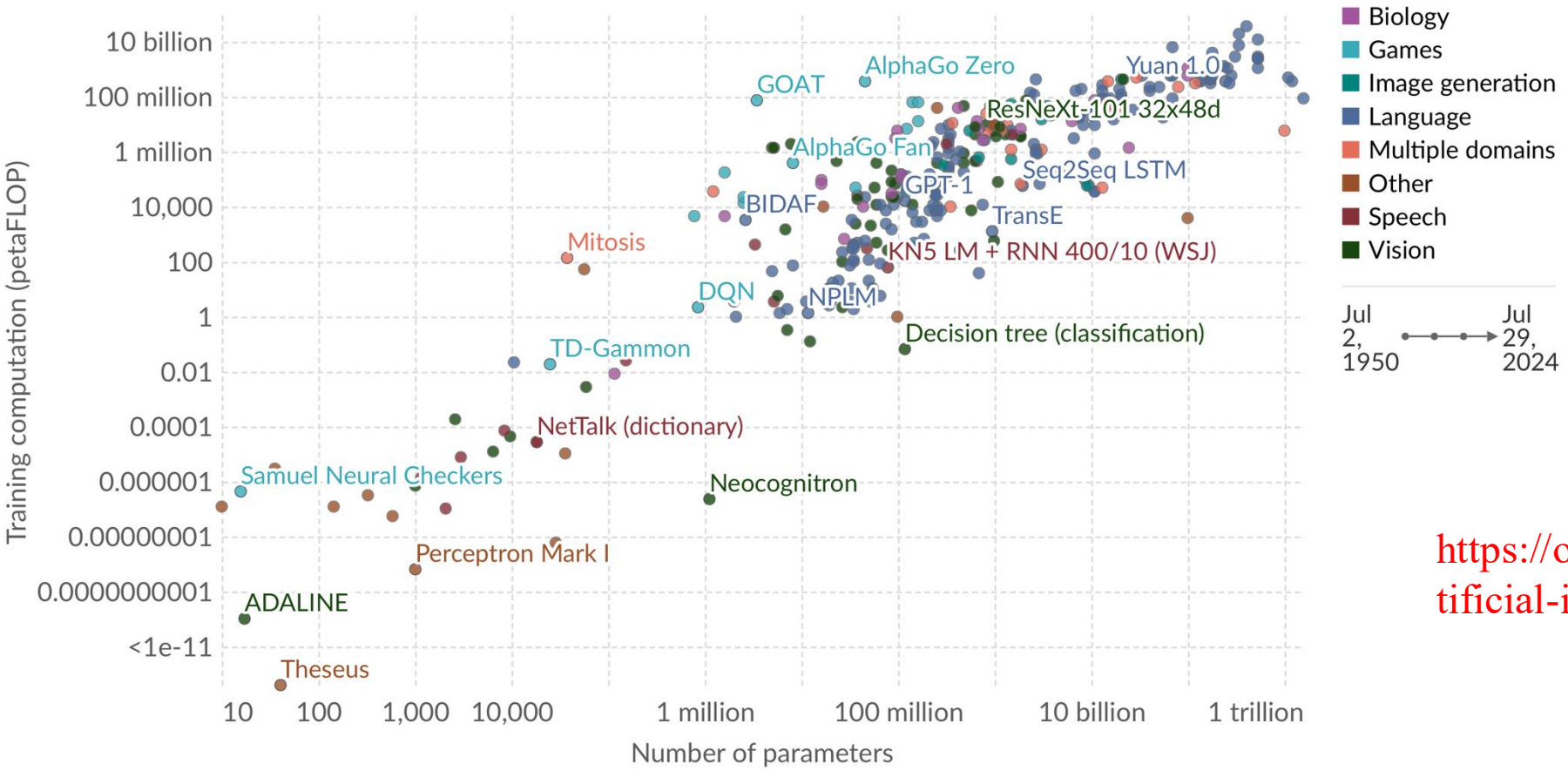
https://www.linkedin.com/pulse/decade-ai-review-muhammad-talha-waseem

THE UNIVERSITY OF TOLEDO

Training computation vs. parameters in notable AI systems, by domain

Our World in Data

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations[1] estimated from AI literature, albeit with some uncertainty. Parameters are variables in an AI system whose values are adjusted during training to establish how input data gets transformed into the desired output.



https://ourworldindata.org/grapher/artificial-intelligence-parameter-count

Data source: Epoch (2024)

OurWorldInData.org/artificial-intelligence | CC BY

Note: Parameters are estimated based on published results in the AI literature and come with some uncertainty. The authors expect the estimates to be correct within a factor of 10.

1. Floating-point operation: A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

THE UNIVERSITY OF TOLEDO

Semiconductor Development for AI

- Graphics Processing Units
- AI Accelerators / ASICs
- Neuromorphic Chips
- High-Bandwidth Memory
- Chiplets and Interposers

Generative AI Use Cases

- Personalized marketing and ads
- Detecting toxic or harmful content
- Text summarization
- Rewriting
- Information extraction
- Q&A

Generative AI on AWS: Building Context-Aware Multimodal Reasoning Applications by Chris Fregly, Antje Barth, Shelbee Eigenbrode

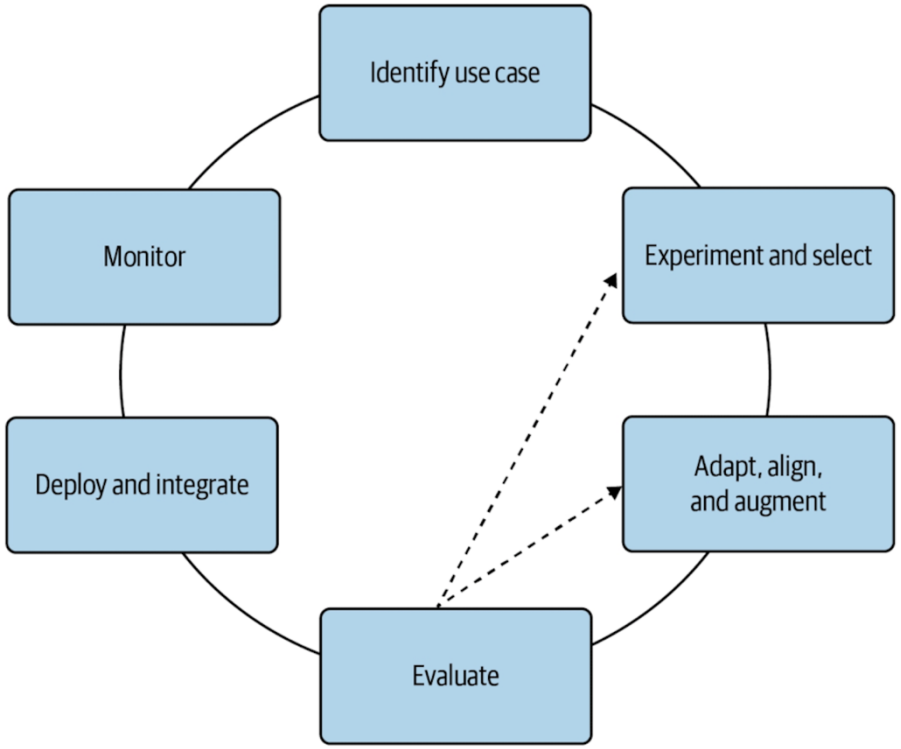
THE UNIVERSITY OF TOLEDO

AI-Driven Semiconductor Design & Fabrication

- Algorithmic chip design, e.g. DeepMind's AlphaChip
- Generative AI and digital twins for semiconductor design
- AI in fabrication
  - Automated defect detection
  - Predictive maintenance

Generative AI Project Life Cycle

- Foundation models
  - Very large and complex neural network models with billions of parameters, which are learned during the training phase (pretraining)
  - BERT (Bidirectional Encoder Representations from Transformers, 340 million in 2018 using 16 GB data by Google)
  - GPT-4 (170 trillion in 2023 using 45 GB data, OpenAI)
  - Claude 2, 3, … (Anthropic)
  - Llama 2, 3, … (Meta)
  - Stable Diffusion (Stability AI)



THE UNIVERSITY OF TOLEDO

Summary and HW#1

- Introduction to Generative AI
- Syllabus