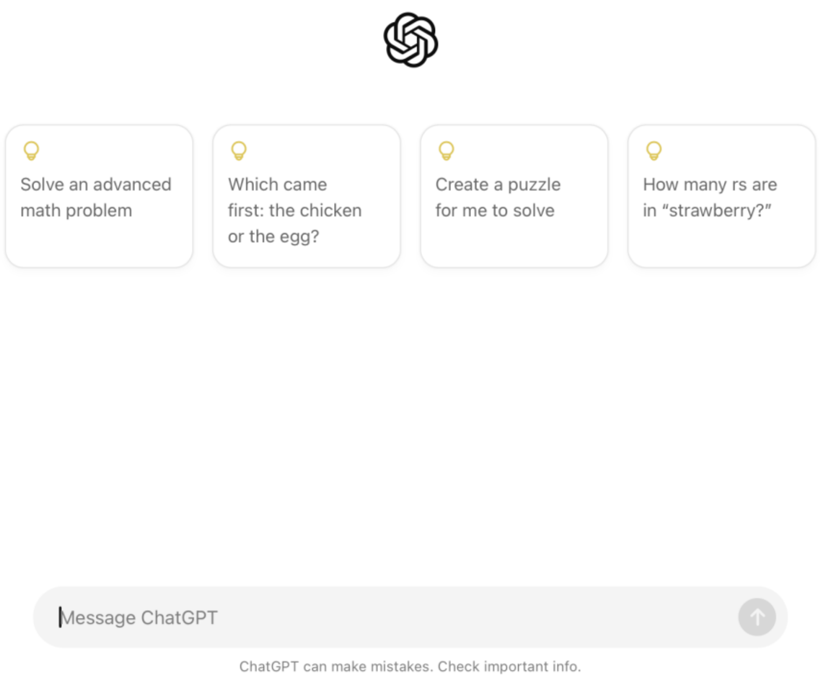
Data and Artificial Intelligence: Empowering The Future

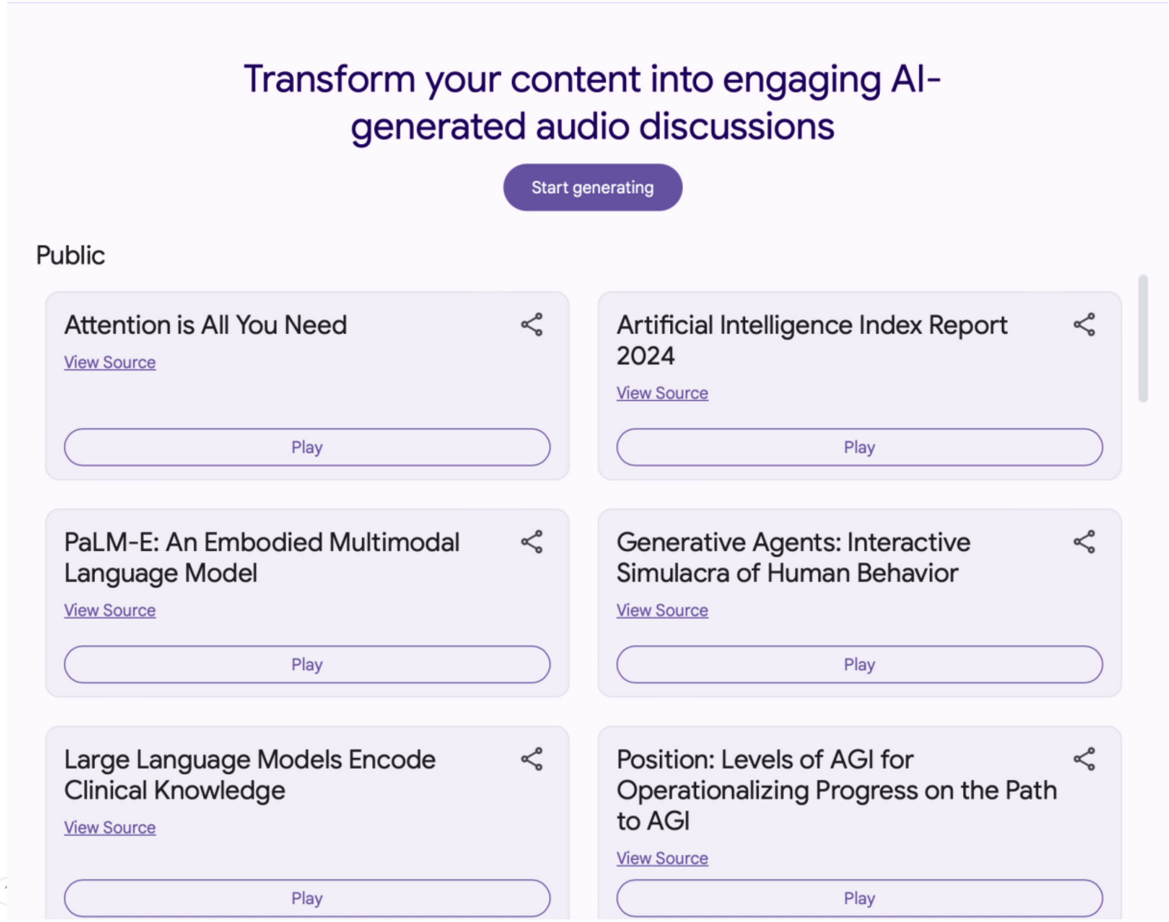
Sanjay Sarma, Asia School of Business & MIT

Capabilities of Gen AI

1. **Generative:** Create new text, videos, audio
2. **Interaction:** LLM’s are the future of interfaces
3. **Knowledge:** Terabytes of notes, manuals weaponized
4. **Reasoning:** Just getting started; GPT 5 will stun you

Impact is evolving…rapidly

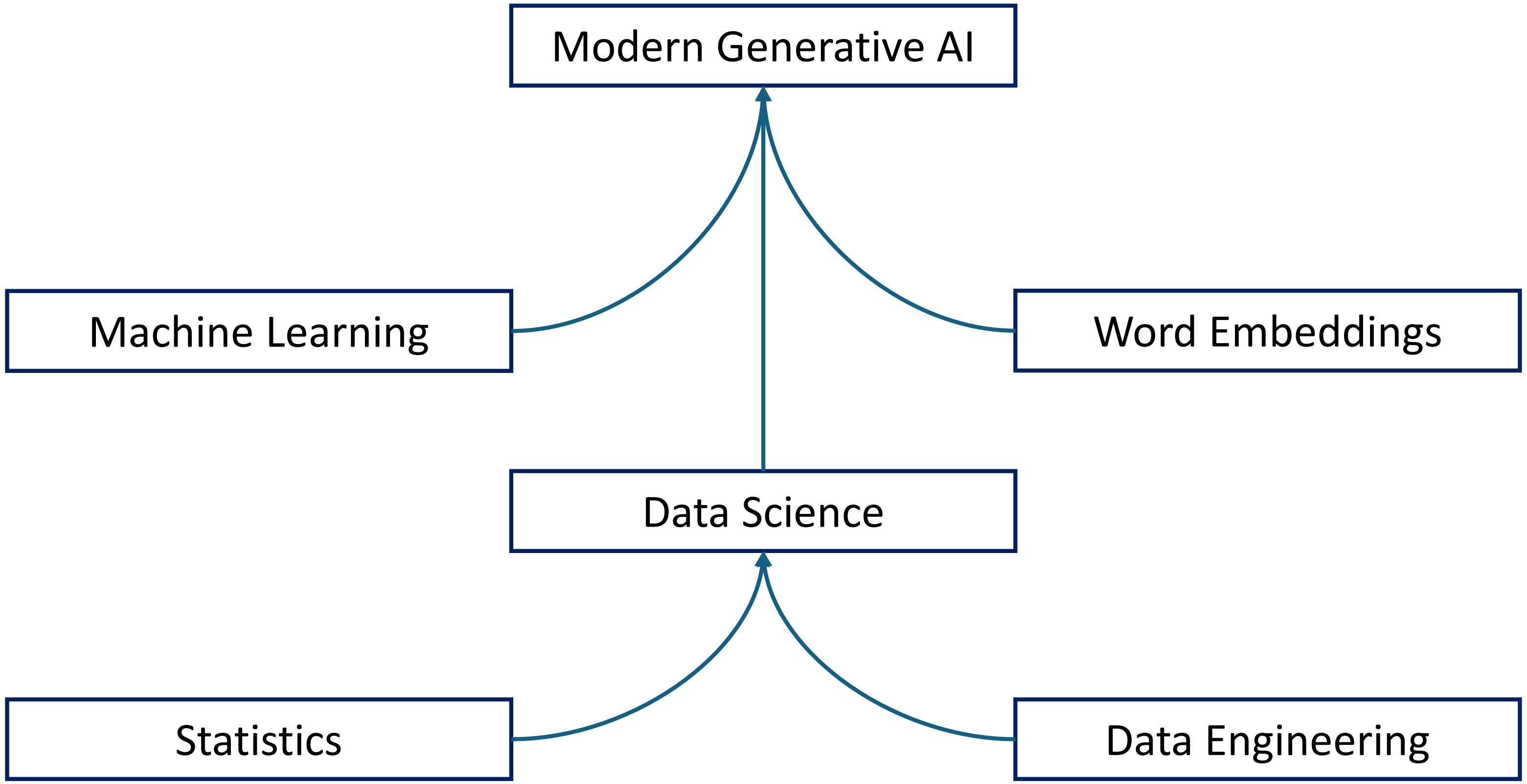




History of AI

= Big Data + Big Stats + Big Vectors

A Historical Confluence



Where does the word “data: come from?

- Data: 1640s, "a fact given or granted," classical plural of datum.
- From Latin datum. Proto-Indo-European (Sanskrit: Data, Dana)

- 1897 as "numerical facts collected for future reference.”
- Transmission: 1946.
- ***Data-processing*** 1954;
- ***database*** 1962;
- ***data-entry*** is by 1970.

www.etymologyonline.com

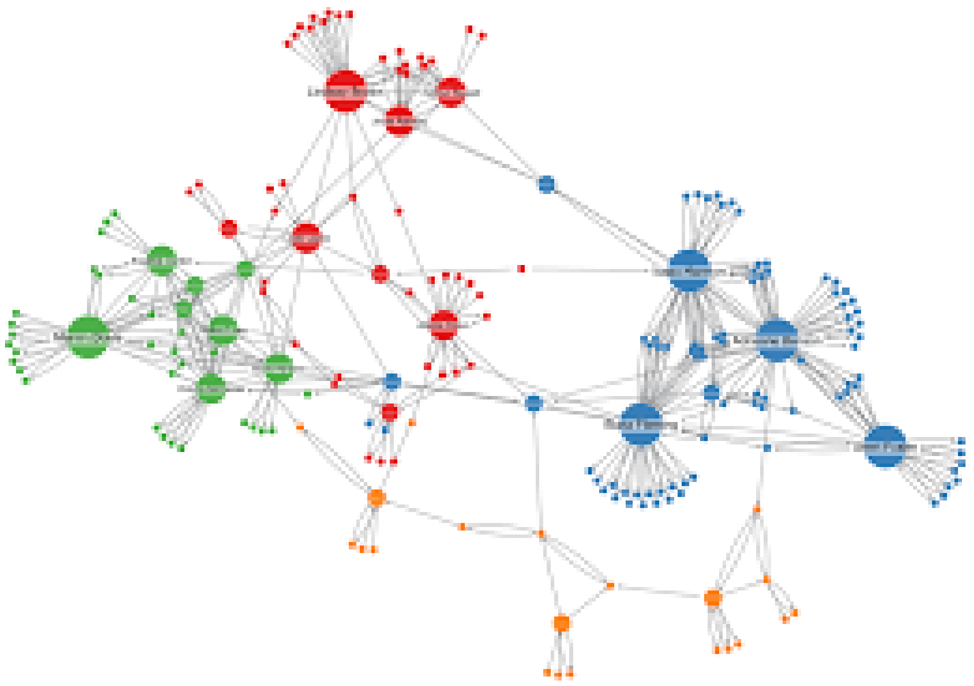
History of “Statistics”

- Comes from ancient civilizations such as Babylon
    - Counting was the beginning: hexagesimal, decimal
- The term comes from the word for “state” (like statecraft)
    - Counts of goods, estimation of taxes
    - Census taking, mortality (John Graunt1662)
- Probability Theory
    - Pascal, Fermat, Bernoulli, de Moivre
- Pre-modern stats: 1850-1945
    - Gauss, Nightingale, Pearson, Fisher, (Egon) Pearson, Bayes, von Neumann, Tukey
- Modern stats
    - Pearson, Bayesian, Ulam, Bradley, Efrom

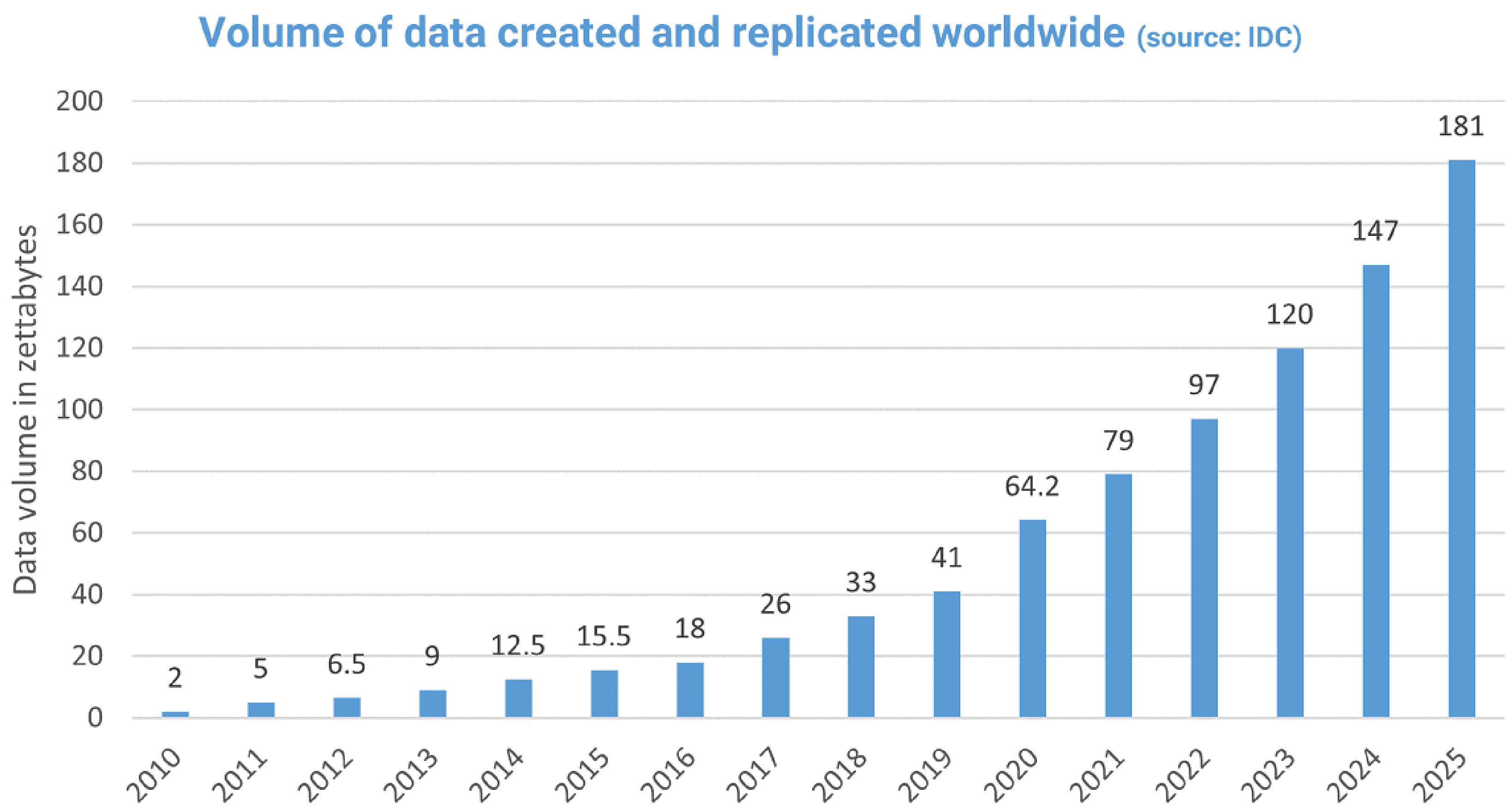
Modern statistics

- Basic statistics (descriptive, exploratory)
- Data pre-processing and cleaning
- Unsupervised learning and clustering

- Supervised learning:
  - Regression, linear and logistic
  - Decision trees
  - Random forests
  - Support vector machines
  - K-Nearest Neighbors
  - *etc.*

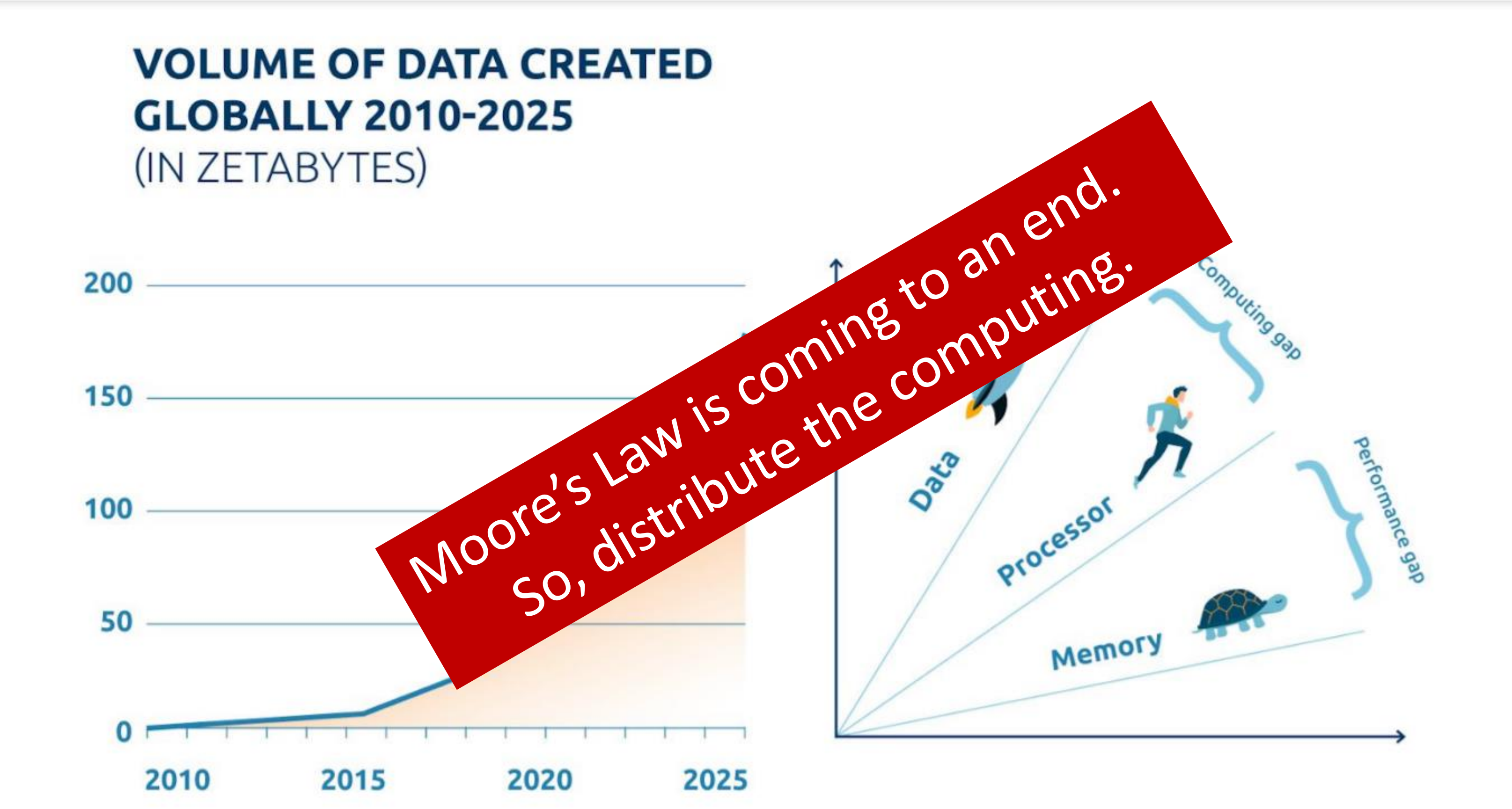


Enter Big Data



https://www.red-gate.com/blog/database-development/whats-the-real-story-behind-the-explosive-growth-of-data

Data outcompetes processing

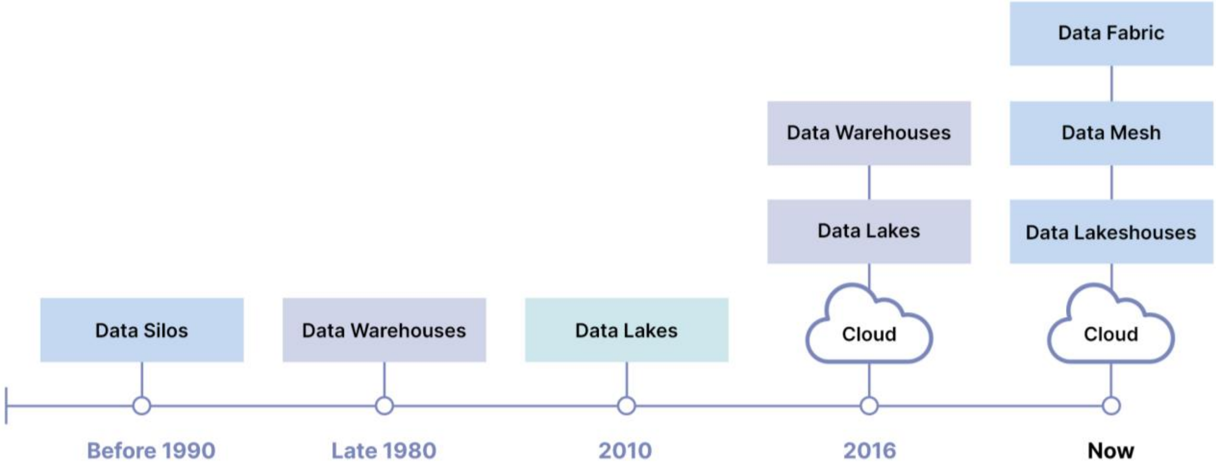


Moore's Law is coming to an end.
So, distribute the computing.

https://medium.com/grovf/embracing-the-exponential-growth-of-data-towards-the-breakthrough-of-memory-scaling-d29a94f2e45c

Data Engineering

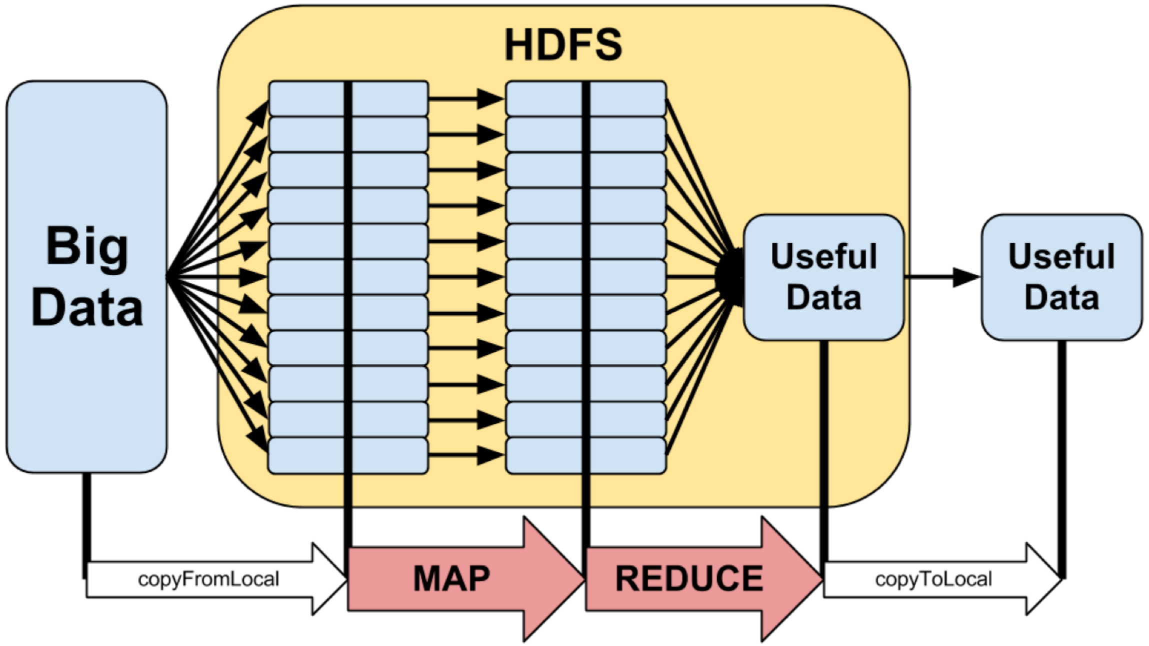
New distributed data architectures



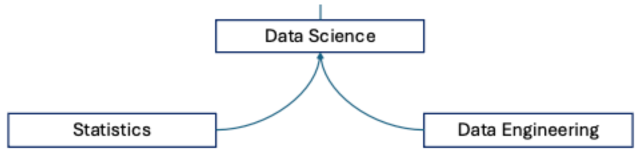
https://www.zdnet.com/article/cloud-computing-will-virtually-replace-traditional-data-centers-within-three-years/

https://www.infocepts.ai/blog/dont-decide-on-a-data-architecture-until-you-read-this/

New Distributed Computation: Hadoop



https://www.glennklockwood.com/data-intensive/hadoop/overview.html

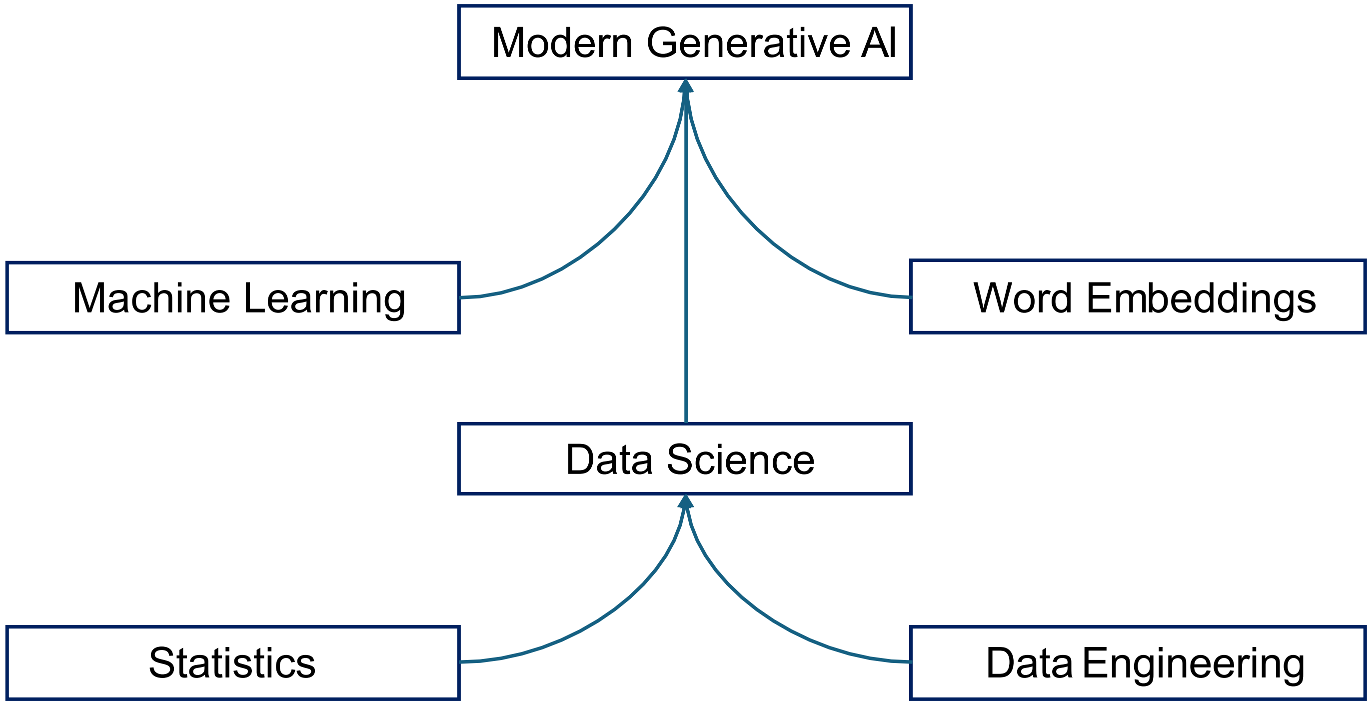


And that is
Data Science

C. F. Jeff Wu suggested rebranding statistics as data science.

“Statistics = Data Science?”

Back to our map



Machine Learning

Start with the data

Labeled (supervised)

Unlabeled (unsupervised)

3 Waves of AI

GPU's, Large Datasets

Wave 1: Neural Networks
1940s -1990s

- Neural Networks
- Expert Systems

Deep Blue

Wave 2: Deep Learning*:*
2011 - present

- Deep Neural Networks
- Pattern Recognition
- Matching, Prediction

Siri

Google Lens

Wave 3: Generative AI
2017 - present

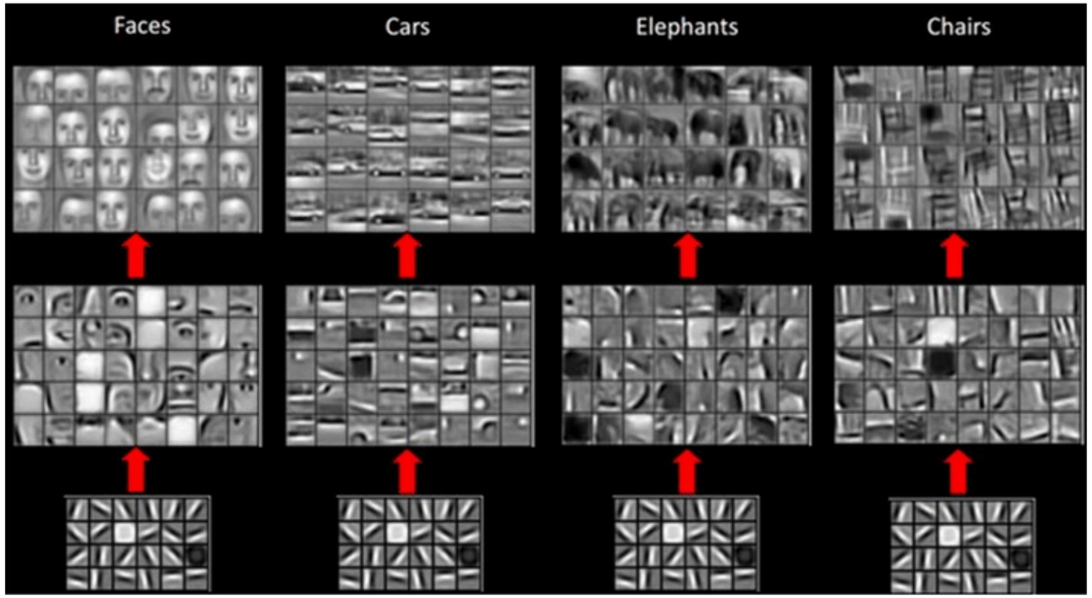
- Large Language Models ("LLMs")
- Text, Image, Video, Science Generation

ChatGPT

MS Copilot

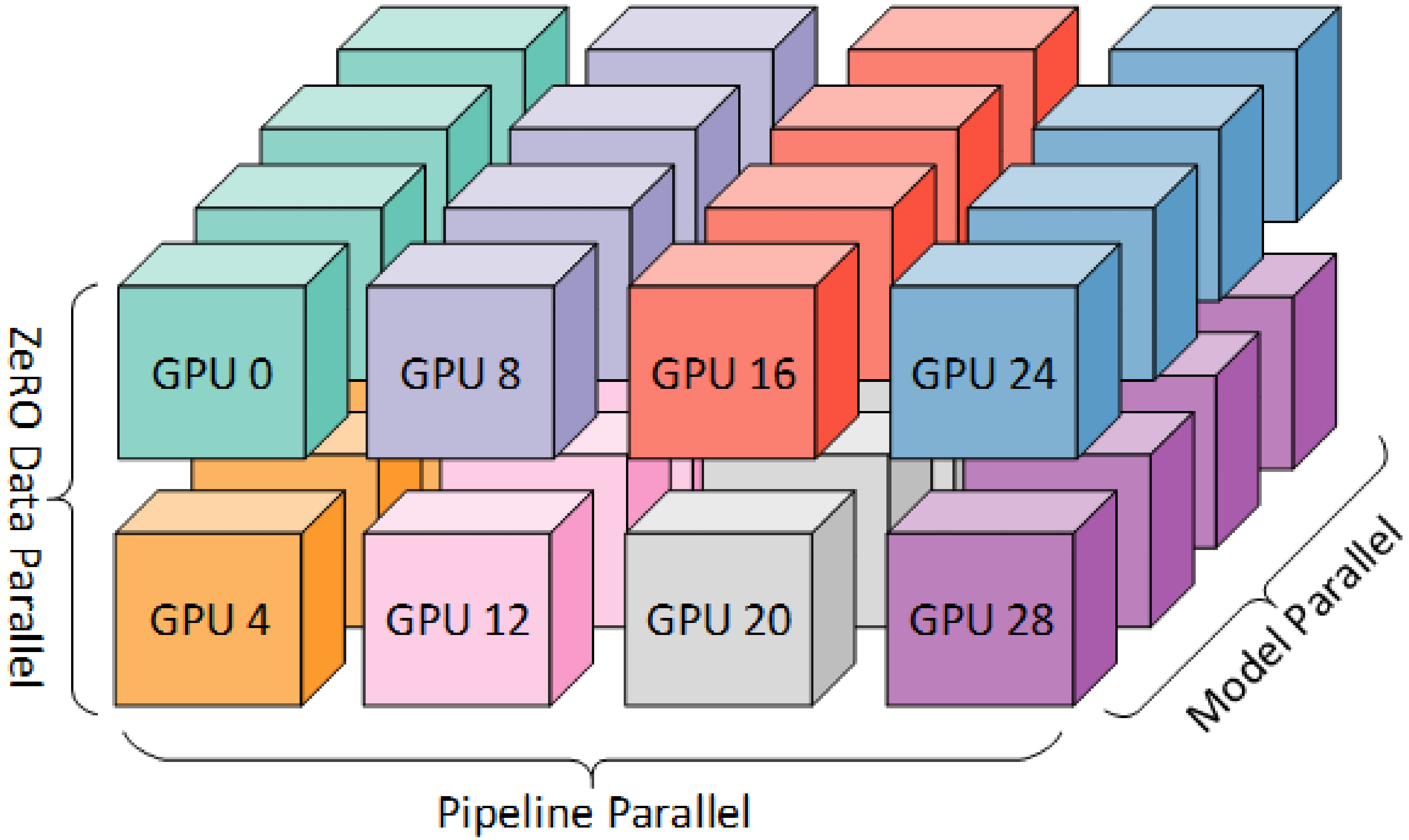
Neural Networks are not Regression

1. Non-Linearity and Complexity
   - Non-linear activation functions: ReLU, sigmoid, or tanh
2. Representation Learning
   - No need to handcraft features
3. Universal Approximation
   - Hornik, K., Stinchcombe, M., & White, H. (1989). *Multilayer feedforward networks are universal approximators*. Neural Networks, 2(5), 359-366.
4. Architecture Variability
   - CNN, RNN's, Transformers
5. Scalability through parallelization, big data
   - Cloud, Hadoop, GPU's, TPU's, transfer earning,



https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac

More Engineering How GPU's parallelize AI



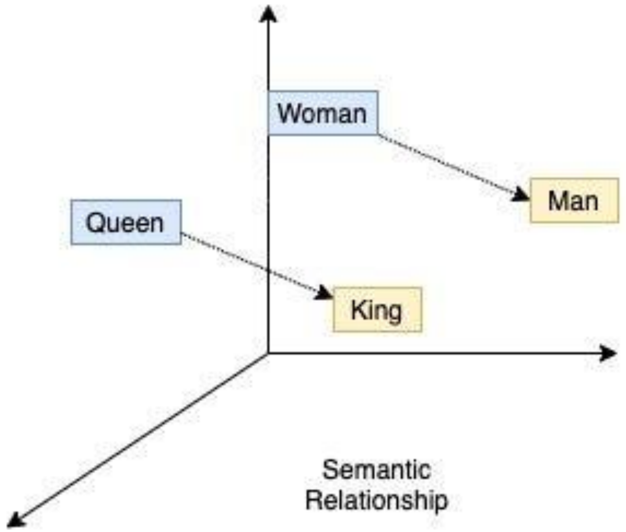
https://www.linkedin.com/pulse/accelerating-ai-art-parallelization-model-training-kirubasagar-v-ase5c/

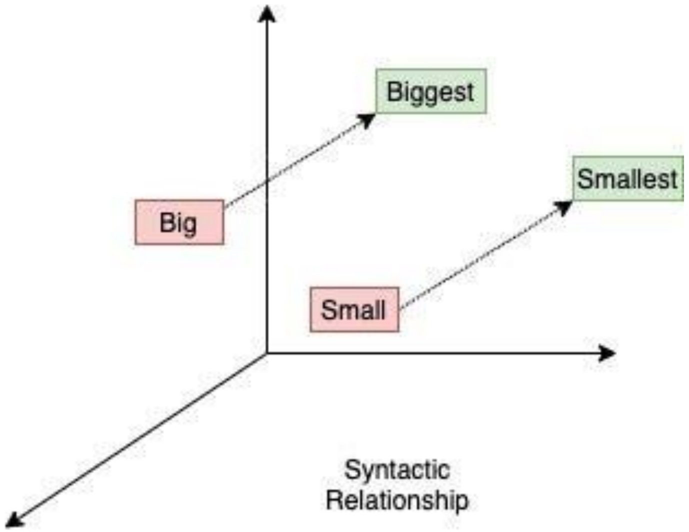
Large Language Models *are based on* Word Embeddings

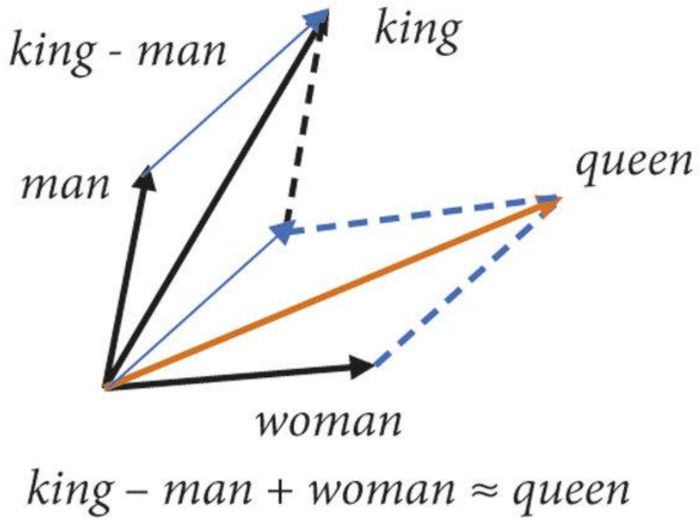
Reducing words to numbers

Words as Vectors

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Proceedings of International Conference on Learning Representations (ICLR)*.
- Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computa- tional Linguistics.







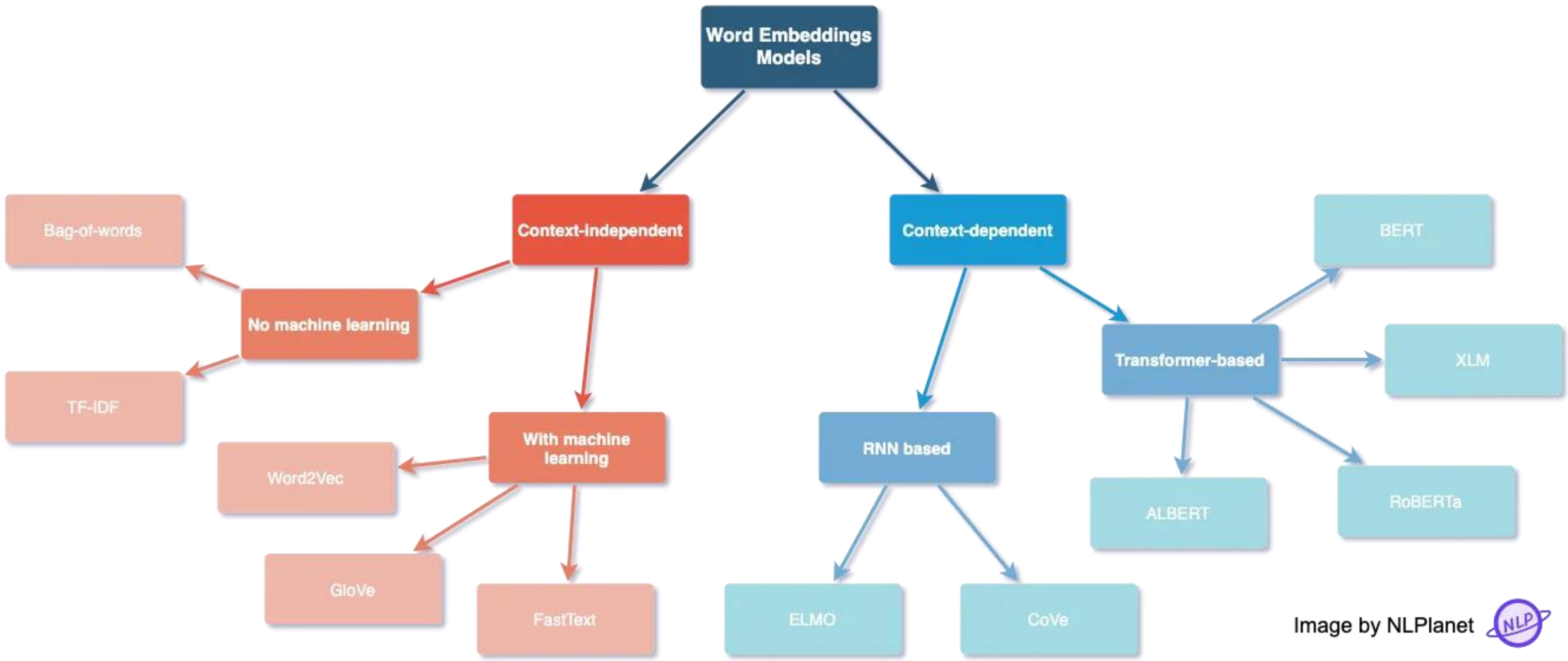
Analogical Reasoning

https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30

Liang, Wentao, Lu Wang, Jialuo She, and Yuqing Liu. "Detecting Resource Release Bugs with Analogical Reasoning." *Scientific Programming* 2022, no. 1 (2022): 3518673.

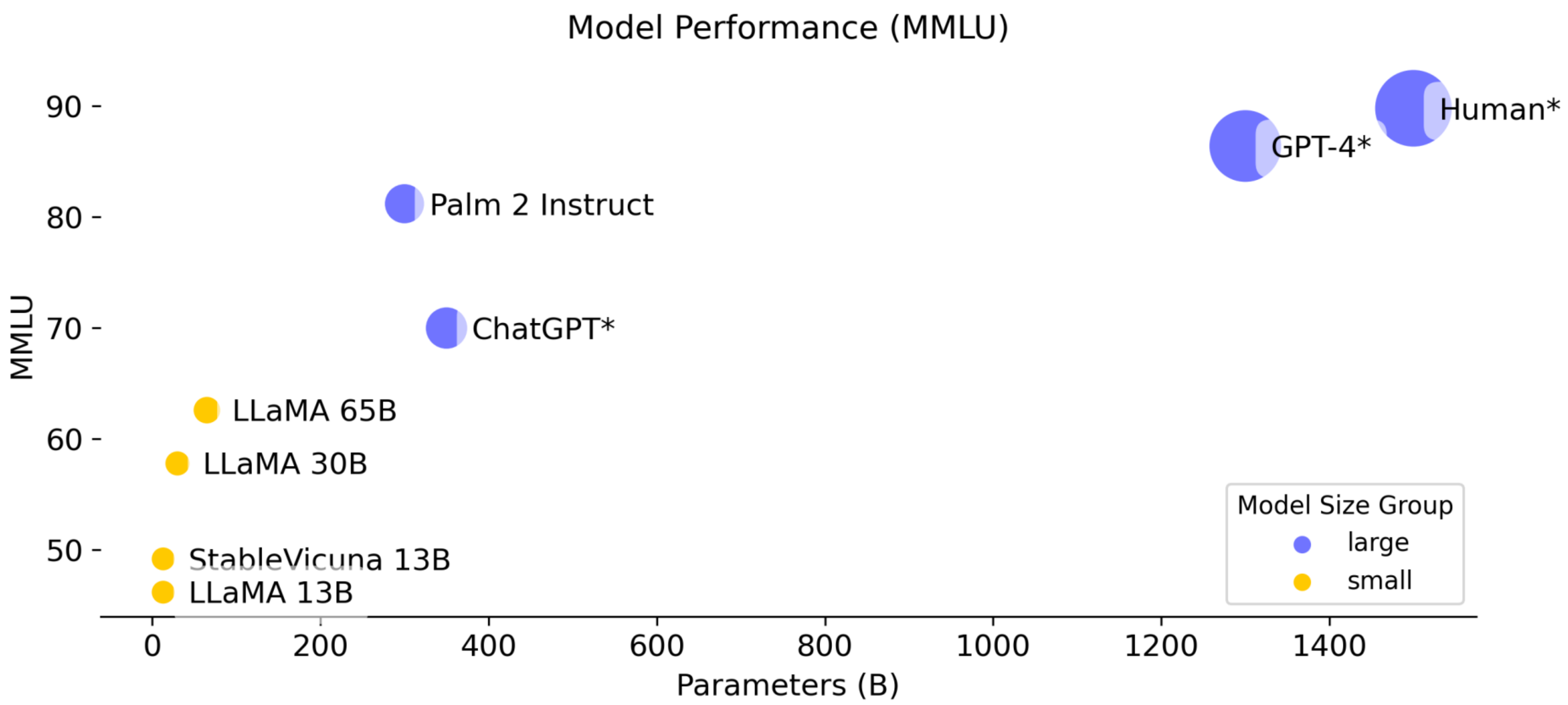
Words become numbers

https://medium.com/nlplanet/two-minutes-nlp-11-word-embeddings-models-you-should-know-a0581763b9a9



Massive Multitask Language Understanding

MMLU tests models on 57 different subjects



*Exact model size is unknown. | Data from InstructEval GitHub.

https://newsletter.victordibia.com/p/understanding-size-tradeoffs-with

Conclusion

Barriers Being Breached by Gen AI

Content	Structured vs Unstructured	80% of data is unstructured
Old AI	Supervised vs Unsupervised	Passive information unleashed

Robots learn to perform chores by watching YouTube

Brian Heater / 9:09 AM PDT • June 22, 2023 Comment

Impact on jobs, economy