Course Overview

CSCI 601-471/671 (NLP: Self-Supervised Models)

https://self-supervised.cs.jhu.edu/sp2025/

Why This Course?

OpenAI

- Let's start with an example: **ChatGPT**
- A computational model that respond to dialogue commands.

https://chat.openai.com/chat

An Example

OpenAI

- Let's start with an example: **ChatGPT**

C Can you write me a 4 paragraph academic essay comparing and contrasting the theories of nationalism of Benedict Anderson and Ernest Gellner?

Benedict Anderson and Ernest Gellner both have theories about the origins and nature of nationalism, but their perspectives differ in key ways. Anderson believes that nationalism is a cultural construct, arising from the development of print capitalism and the creation of imagined communities. Gellner, on the other hand, sees nationalism as a product of modernization and industrialization, with nations being formed as a means of organizing society and providing a shared identity.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

An Example

OpenAI

- Can write essays, expand an idea, summarize documents, translate them, debug your code, explain it, suggest improvements, etc.

How is it able to do these? 🤔
- What are the principles that govern this technology?
- Can we replicate it?
- Can we extend it?

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Meanwhile …

What will be the gender of the first female US President?

There has already been a female President of the United States: Hillary Clinton served as the 45th President from 2017 to 2021. Is there something else you would like to know about this

Was Lincoln's assassin on the same continent as Lincoln when the assassination occurred

Abraham Lincoln was assassinated by John Wilkes Booth while he was attending a play at Ford's Theatre in Washington, D.C. I do not know where John Wilkes Booth was at the time of the assassination, but Washington, D.C. is located on the continent of North America, so it is possible that Booth was also on that continent at the time.

Why This Course?

OpenAI

- The model is extremely good at important tasks that we can about: why? how?

- It is easily fails at seemingly trivial tasks: why?

Why does it make such surprisingly simple mistakes? 🤔
- Can we explain these?
- Can we predict them?
- Can we mitigate (or even better, solve) them?

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Course Learning Objectives

- We will paper on a variety of topics:
  - Architectures
  - Pre-training
  - Alignment and safety
  - Efficiency
  - Interaction with the world (code, physical world, etc.)
  - Impacts on humans -- their misuse, biases, etc.
- Skills:
  - **Technical**—understanding of the algorithms and implementing them.
    - Gaining intuitions about **capabilities** and **limitations** of models.
  - **Soft skills**—algorithms, PyTorch, SLURM, intuition about capabilities, teamwork.

Focus on Natural/Human Language

- **Most** of the class revolves around **natural language**.

- Why natural language?
    - It is a **convenient medium of communication**.
    - Natural language is our species' best attempt to encode **everything about the world** as **efficiently** as possible.
    - A huge archive of natural language is **freely available** (e.g., on the web).

Self-Supervised Models

Self-Supervision

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Self-Supervision

[Slide credit: Colin Raffel]

Self-Supervision

[Slide credit: Colin Raffel]

Self-Supervision

Dataset of natural images

[Slide credit: Colin Raffel]

Self-Supervision

Dataset of natural images

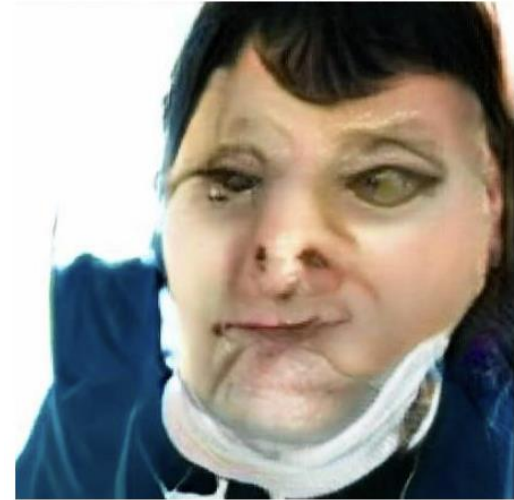Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide credit: Colin Raffel]

Self-Supervision

Dataset of natural images

Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide credit: Colin Raffel]

Self-Supervision

Dataset of natural images

Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide credit: Colin Raffel]

Self-Supervision

Dataset of natural images

Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

[Slide credit: Colin Raffel]

Self-Supervision

== treaty of paris (1763)

the treaty of paris, also known as the treaty of 1763, was signed on 10 february 1763 by the kingdoms of great britain, france and spain, with portugal in agreement, after great britain's victory over france and spain during the seven years' war.

the signing of the treaty formally ended the seven years' war, known as the french and indian war in the north american theatre, ….

[Slide credit: Colin Raffel]

Self-Supervision

== wheelbarrow

A wheelbarrow is a small hand-propelled vehicle, usually with just one wheel, designed to be pushed and guided by a single person using two handles at the rear, or by a sail to push the ancient wheelbarrow by wind. The term "wheelbarrow" is made of two words: "wheel" and "barrow." "Barrow" is a derivation of the Old English "barew" which was a device used for carrying loads. The wheelbarrow is designed to .....

==

the
the
febr
grea
port
brita
spai

the
end
as th
north american theatre, ...

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Self-Supervision

== lemon

WIKIPEDIA
The Free Encyclopedia

The lemon (Citrus limon) is a species of small evergreen trees in the flowering plant family Rutaceae, native to Asia, primarily Northeast India (Assam), Northern Myanmar or China.[2] The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.[2] The pulp and rind are also used in cooking and baking. ....

== w

A wh prop
one and
two
to pu
wind
mad
"bar
the (
devi
wheelbarrow is designed to .....

==

the prop
the one
febr and
grea two
port to pu
brita wind
spai mad
"bar
the s the (
end devi
as th wheelbarrow is designed to .....
north american theatre, ...

Dataset of Wikipedia articles

[Slide credit: Colin Raffel]

Self-Supervision

== lemon

The lemon (Citrus limon) is a species of small evergreen trees in the flowering plant family Rutaceae, native to Asia, primarily Northeast India (Assam), Northern Myanmar or China.[2] The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.[2] The pulp and rind are also used in cooking and baking. ….

== w

A wh
prop
one
and
two
to pu
wind
mad
"bar
the C
devi
wheelbarrow is designed to …..

==
the
the
febr
grea
port
brita
spai

the s
end
as th
north american theatre, ….

WIKIPEDIA
The Free Encyclopedia

== wings over kansas

wings over kansas is the second studio album by jason ammons, john bolster and mo rosato. the album debuted at number one on the billboard 200, selling 35,000 copies in it first week at the time. it was the second highest selling album to debut at the billboard top 50 and the third highest selling album to debut at the top heatseekers, with 26,000 copies sold. this is the supremes album earning the nickname …
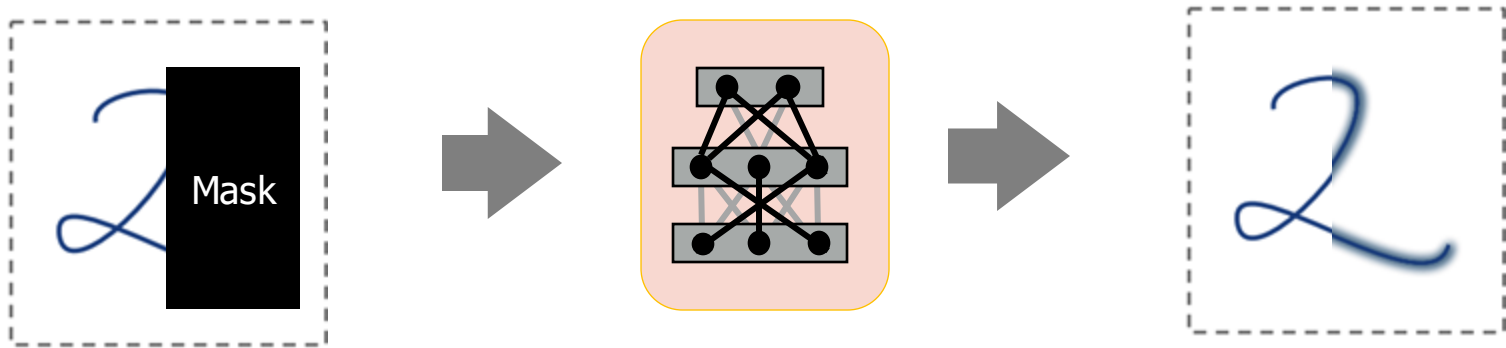
Dataset of Wikipedia articles

[Slide credit: Colin Raffel]

Self-Supervised Models

*are **predictive models** of the world!*

Self-Supervised Models
*are **predictive models** of the world!*

- Are trained to complete partial samples from the world.

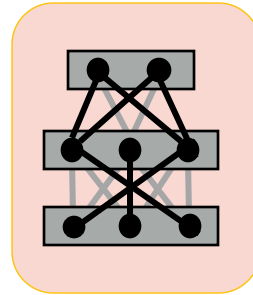[Bengio et al. 2004, Hinton et al. 2006, Peters et al. 2018, …]

Self-Supervised Models

are predictive models of the world!

- Are trained to complete partial samples from the world.

"Wings over Kansas is [MASK]"

"Wings over Kansas is an aviation website founded in 1998 by Carl Chance owned by Chance Communications, Inc."

[Bengio et al. 2004, Hinton et al. 2006, Peters et al. 2018, …]

Self-Supervised Models

*are **predictive models** of the world!*

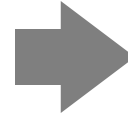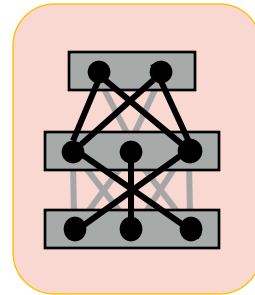*learned from **cheaply available** unlabeled data*

Self-Supervised Models

*are **predictive models** of the world!*

*learned from **cheaply available** unlabeled* ***data***
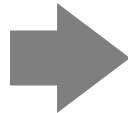
Self-Supervised Models

are tightly connected to *tasks* *we care about.*

- **Goal:** Answering questions

Question: "Where is the birthplace of the American national anthem?"

"The birthplace of the American national anthem" [MASK]

"The birthplace of the American national anthem, "The Star-Spangled Banner," lies in Baltimore, Maryland."

Produced with GPT-J: https://6b.eleuther.ai/

Self-Supervised Models

- *Are predictive models of the world.*
- *Are learned from unlabeled data.*
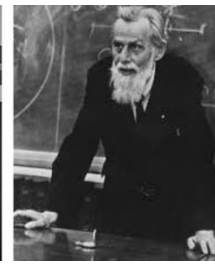- *Tightly connected to tasks we care about.*

How did we get here?

Progress in AI

- Many advances are due to neural networks
- How old are neural networks?

[Slide credit: Arman Cohan]
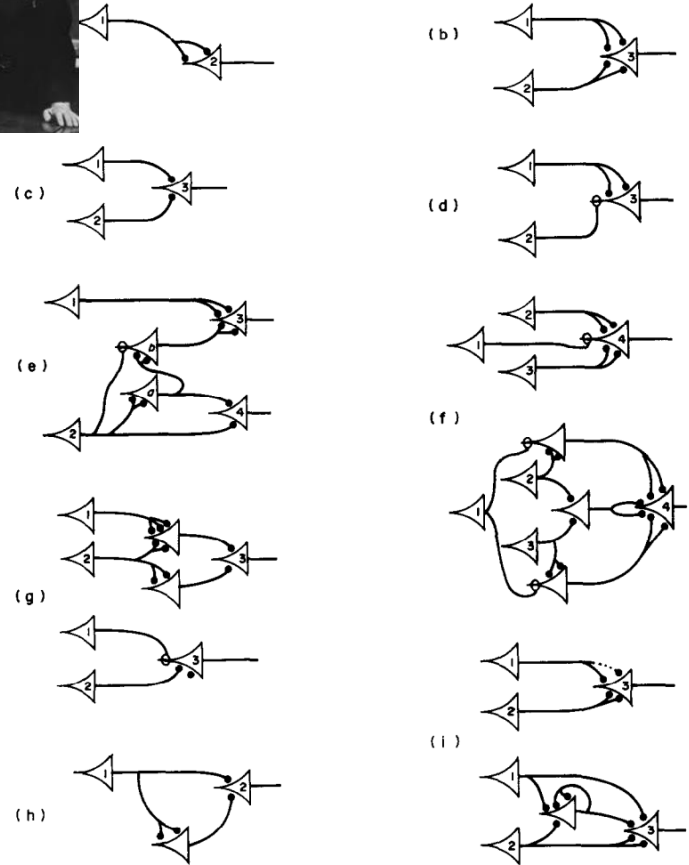
Progress in AI

- Many advances are due to neural networks
- How old are neural networks?

McCulloch & Pitts (1943)

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

■ WARREN S. MCCULLOCH AND WALTER PITTS
University of Illinois, College of Medicine,
Department of Psychiatry at the Illinois Neuropsychiatric Institute,
University of Chicago, Chicago, U.S.A.

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

[Slide credit: Arman Cohan]

Progress in AI

- Many advances are due to neural networks
- How old are neural networks?
    - They've been around since the 1940s
    - But why have only recently we seen breakthroughs?
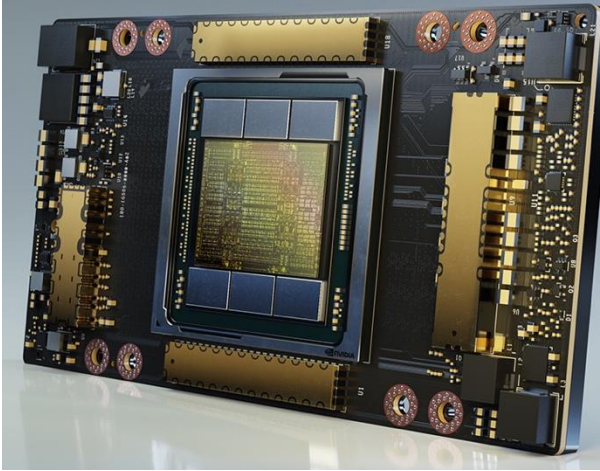   - 3 forces came together!

Force 1: Massive Amount of Data

- Internet provided us with a massive repository of data.
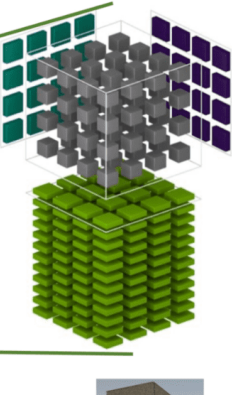

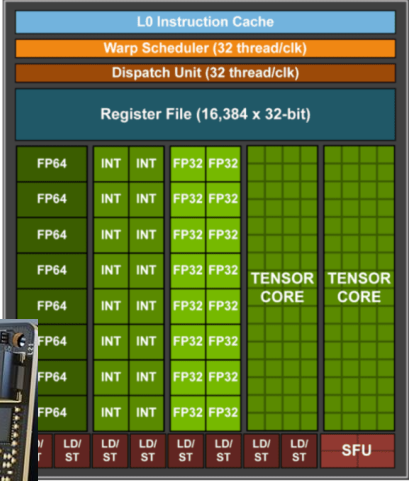**Annual Size of the Global Datasphere**
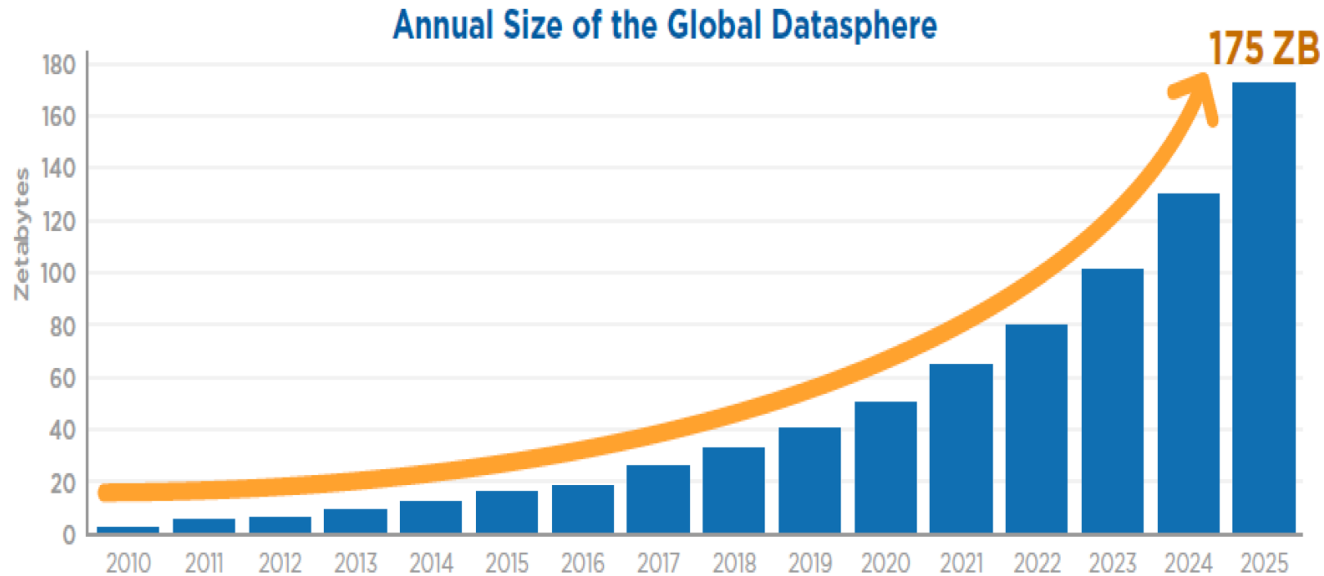
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Force 2: Computing Power

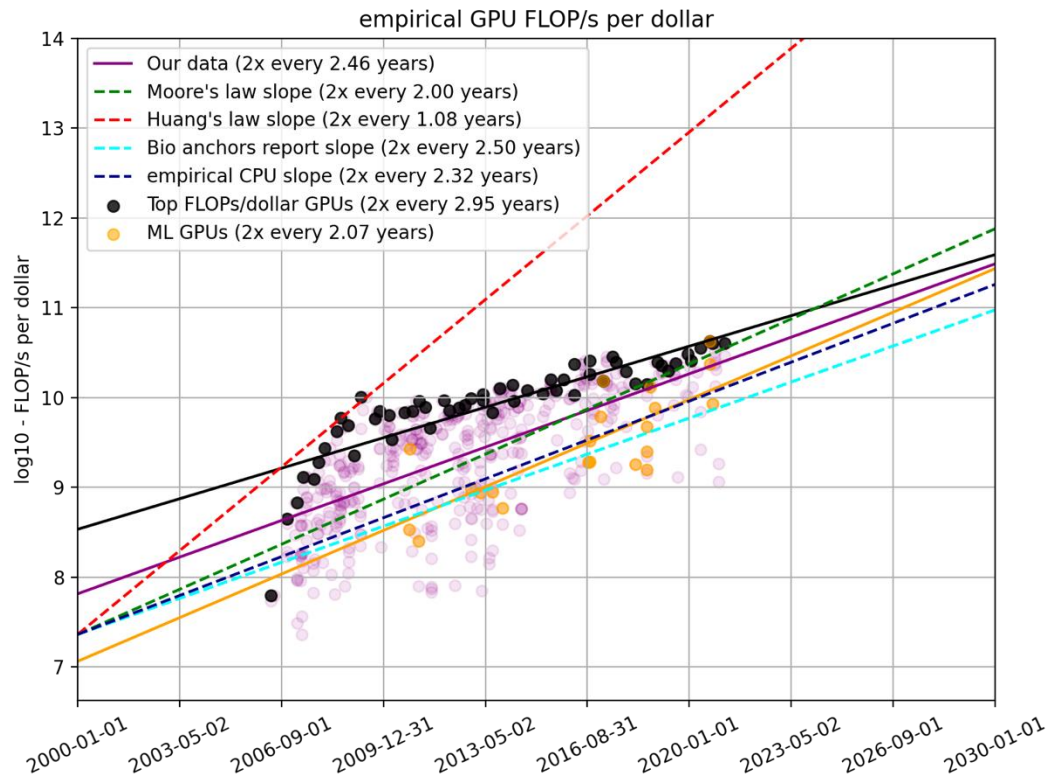- Fast processors for deep learning!

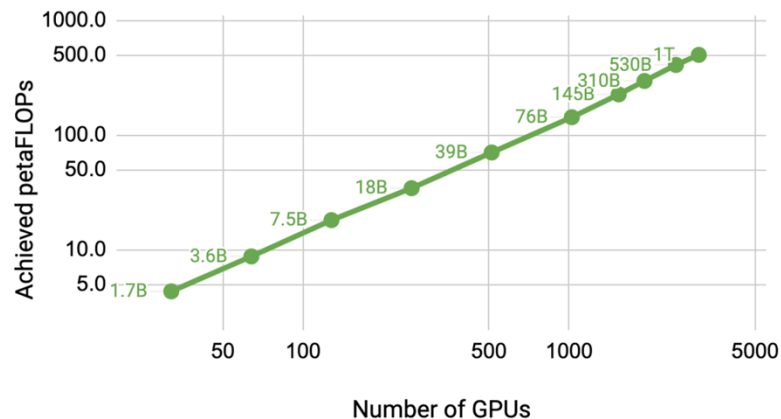A100 GPU

Force 2: Computing Power

- Fast processors for deep learning!

- Cheaper computing power over time.

The amount of computing power, per dollar

empirical GPU FLOP/s per dollar

Legend:
- Our data (2x every 2.46 years)
- Moore's law slope (2x every 2.00 years)
- Huang's law slope (2x every 1.08 years)
- Bio anchors report slope (2x every 2.50 years)
- empirical CPU slope (2x every 2.32 years)
- Top FLOPs/dollar GPUs (2x every 2.95 years)
- ML GPUs (2x every 2.07 years)

https://www.lesswrong.com/posts/c6KFvQcZggQKZzxr9/trends-in-gpu-price-performance

Force 2: Computing Power

- Fast processors for deep learning!

- Cheaper computing power over time.

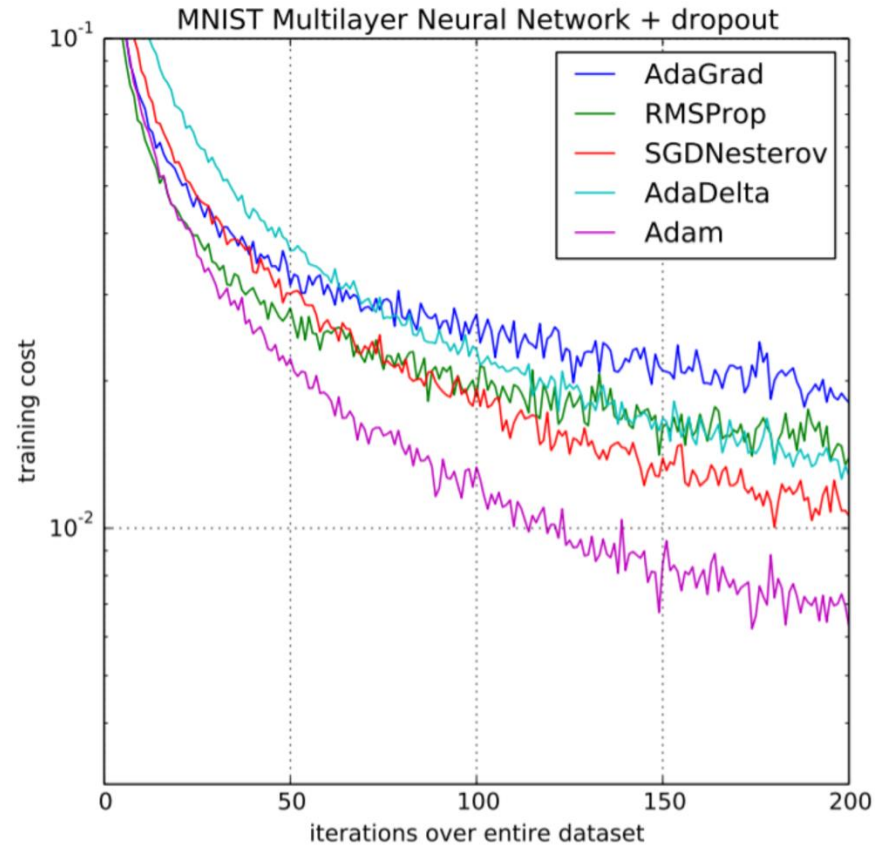- Distributed training/inference allows us to scale to a larger set of processors.
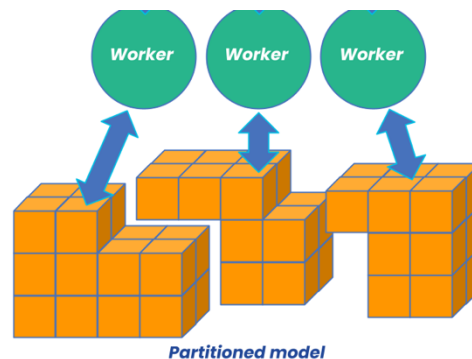
Force 3: Algorithmic innovations

- Advances in optimization

MNIST Multilayer Neural Network + dropout

- AdaGrad
- RMSProp
- SGDNesterov
- AdaDelta
- Adam

training cost / iterations over entire dataset

Force 3: Algorithmic innovations

- Advances in optimization
- Innovations in model architectures
- ….
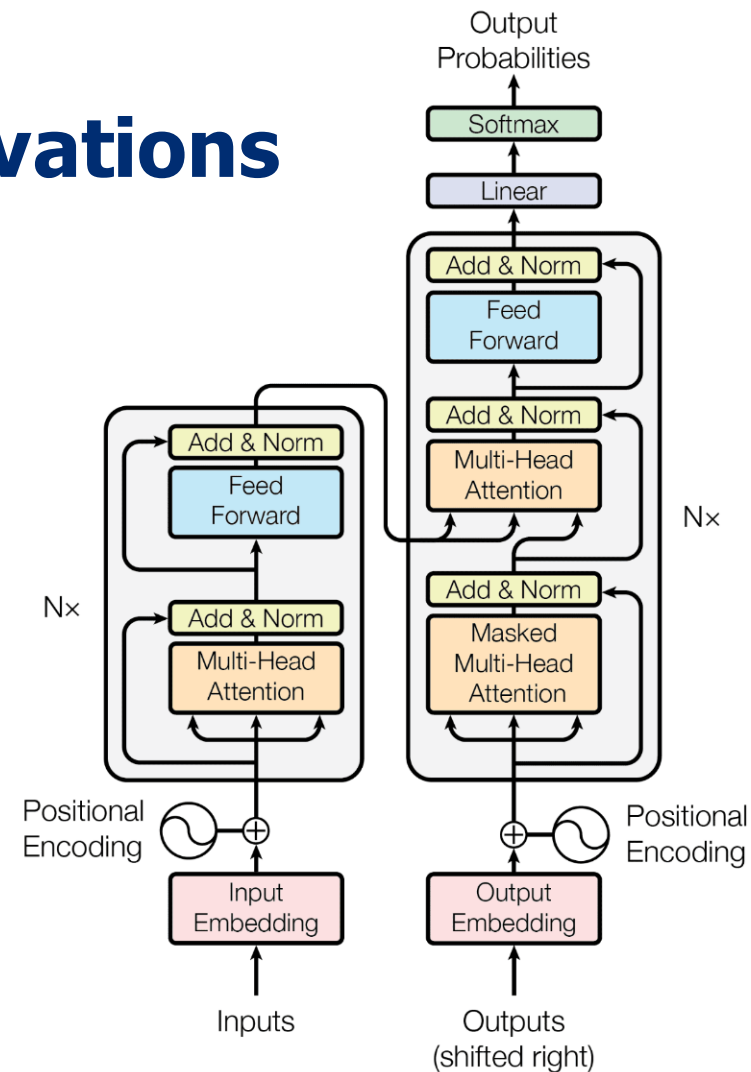
Self-Supervised Models

FINANCIAL TIMES

Is AI finally closing in on human intelligence?

INSIDER

Google's DeepMind artificial intelligence has figured out how to talk

The Atlantic

An Artificial Intelligence Developed Its Own Non-Human Language

On terminology

- These names are sometimes used interchangeably:
  - Self-supervised models
  - Pre-trained models
  - Generative AI models
  - Foundation models
  - Frontier models
  - …

- Though they're not exactly the same.

On terminology

- These names are sometimes used interchangeably:
    - Self-supervised models
    - Pre-trained models
    - Generative AI models
    - Foundation models
    - Frontier models
    - …

    Our models are
    more than just "pretrained"!

- Though they're not exactly the same.

On terminology

- These names are sometimes used interchangeably:
    - Self-supervised models
    - Pre-trained models
    - Generative AI models   ~ Models that generate content
    - Foundation models
    - Frontier models
    - ...

- Though they're not exactly the same.

On terminology

- These names are sometimes used interchangeably:
  - Self-supervised models
  - Pre-trained models
  - Generative AI models
  - Foundation models
  - Frontier models
  - ...

  > ~ can be used as a foundational component of modern AI systems

  > That doesn't mean that these models are the foundation of AI!

- Though they're not exactly the same.

More discussion on naming (Bommasani et al., 2021)   https://arxiv.org/pdf/2108.07258.pdf

On terminology

- These names are sometimes used interchangeably:
  - Self-supervised models
  - Pre-trained models
  - Generative AI models
  - Foundation models
  - Frontier models
  - ...

  They're user-facing "frontiers" of applications

- Though they're not exactly the same.

On terminology

- These names are sometimes used interchangeably:
  - Self-supervised models
  - Pre-trained models
  - Generative AI models
  - Foundation models
  - Frontier models
  - …

- Though they're not exactly the same.
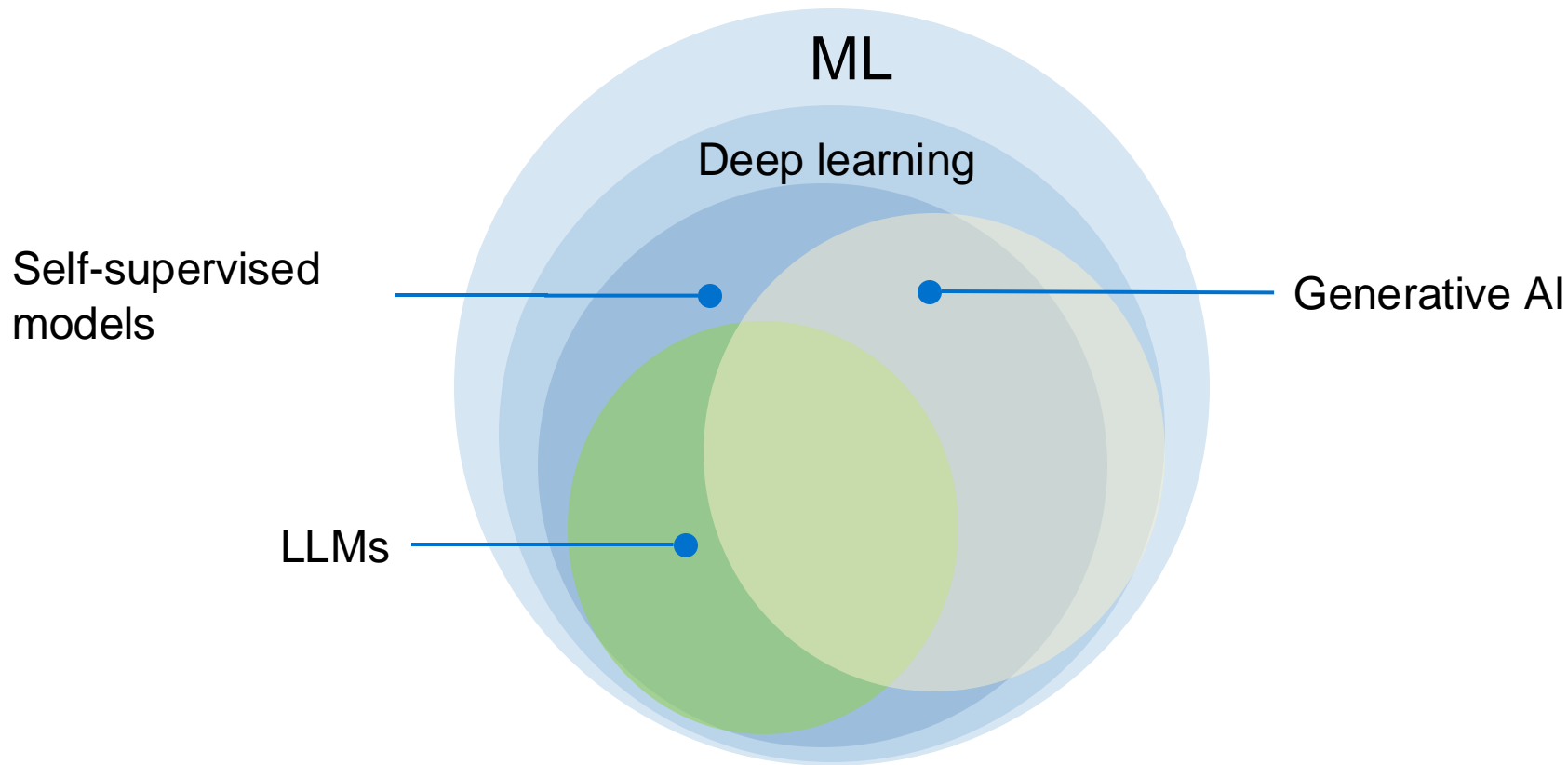
On terminology

- These names are sometimes used interchangeably:
  - Self-supervised models
  - Pre-trained models
  - Generative AI models
  - Foundation models
  - Frontier models
  - …

Which is your favorite? 😊

[Slide credit: Arman Cohan]

Current state of Self-supervised Models

- Almost every AI model is based on Neural networks
- Performance is consistently improving with scale
    - More training data
    - Larger models (number of neural network parameters)

Current state of Self-supervised Models

State-of-the-art models are
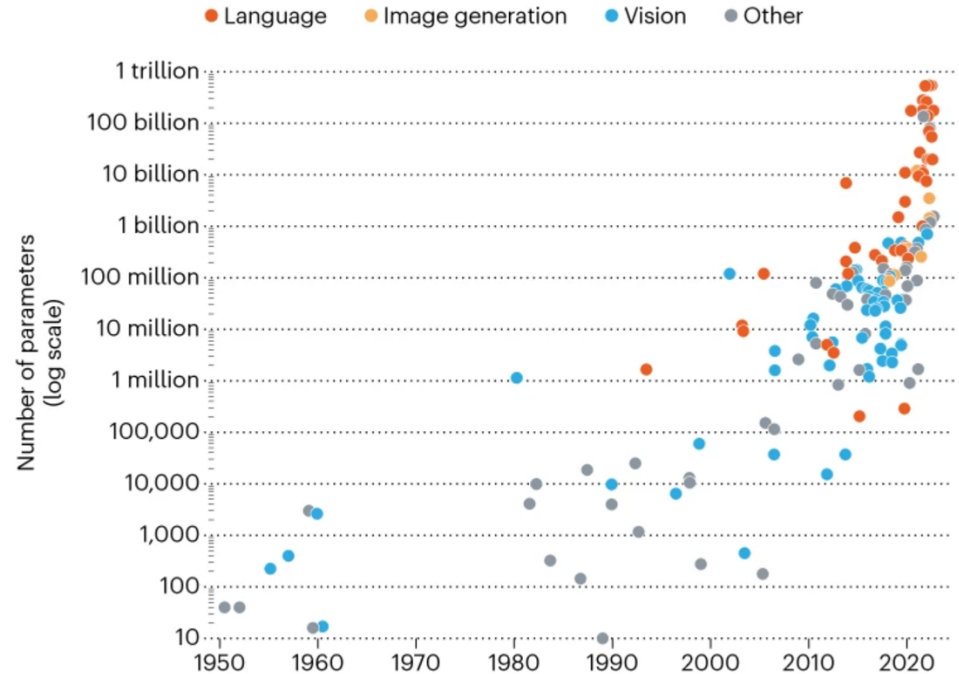hundreds of billions of parameters

Image from: In AI, is bigger always better? https://www.nature.com/articles/d41586-023-00641-w

[Slide credit: Arman Cohan]

Current state of Self-supervised Models

State-of-the-art models are hundreds of billions of parameters

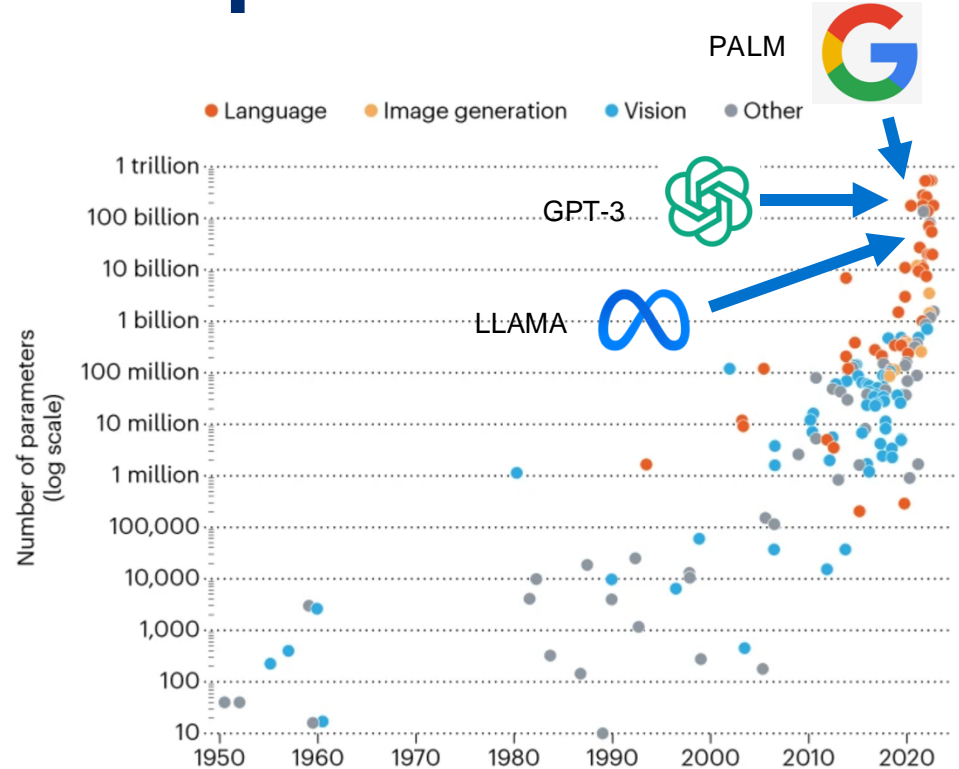Trained on vast amounts of data (Trillions of tokens)

Image from: In AI, is bigger always better? https://www.nature.com/articles/d41586-023-00641-w

[Slide credit: Arman Cohan]

Course Learning Objectives

- We will paper on a variety of topics:
  - Architectures
  - Pre-training
  - Alignment and safety
  - Efficiency
  - Interaction with the world (code, physical world, etc.)
  - Impacts on humans -- their misuse, biases, etc.
- Skills:
  - **Technical**—understanding of the algorithms and implementing them.
    - Gaining intuitions about **capabilities** and **limitations** of models.
  - **Soft skills**—algorithms, PyTorch, SLURM, intuition about capabilities, teamwork.

Disclaimer

- We will not be exhaustive.

- The goal is to give you a bird's-eye view of the field.

Course Logistics

CSCI 601-471/671 (NLP: Self-Supervised Models)

https://self-supervised.cs.jhu.edu/sp2024/

Course Logistics Brief

- **Instructor:** Daniel Khashabi
  - I prefer just "Daniel", as long as we act mutually respectfully.
  - No "prof Daniel"!! 🙅

- **TA:** Tianjian Li
- **CAs:** Sungwon, Tanay, Adi, TaiMing, Prabhav, TBD

Daniel Khashabi
Instructor

Tianjian Li
Teaching Assistant

Sungwon Kim
Course Assistant

Adi Asija
Course Assistant

Prabhav Singh
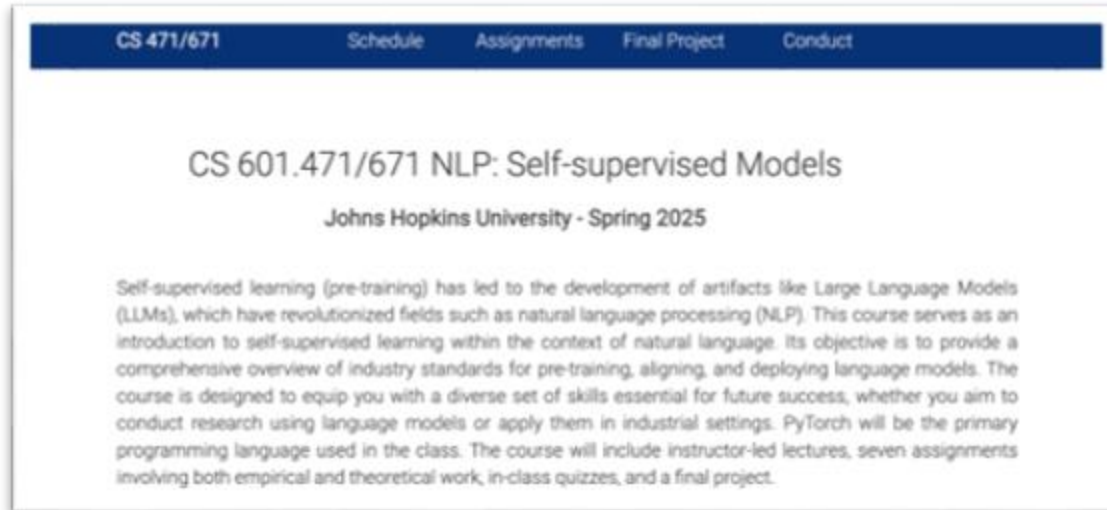Course Assistant

Tanay Nayak
Course Assistant

TaiMing (Terry) Lu
Course Assistant

TBD
Course Assistant

Course Website

- Lots of important information on the **website**:
  - https://self-supervised.cs.jhu.edu/sp2025

CS 471/671   Schedule   Assignments   Final Project   Conduct

CS 601.471/671 NLP: Self-supervised Models

Johns Hopkins University - Spring 2025

Self-supervised learning (pre-training) has led to the development of artifacts like Large Language Models (LLMs), which have revolutionized fields such as natural language processing (NLP). This course serves as an introduction to self-supervised learning within the context of natural language. Its objective is to provide a comprehensive overview of industry standards for pre-training, aligning, and deploying language models. The course is designed to equip you with a diverse set of skills essential for future success, whether you aim to conduct research using language models or apply them in industrial settings. PyTorch will be the primary programming language used in the class. The course will include instructor-led lectures, seven assignments involving both empirical and theoretical work, in-class quizzes, and a final project.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Other Relevant Websites

Course Prerequisites

- Comfortable with programming, particularly Python

- Calculus, linear algebra, probability

- Familiarity with Natural Language Processing

- HW1 should give a sense how prepared you are!

Course Logistics Brief

- **Pre-recorded videos**: will be available on the course website and/or Piazza.

- **Office hours:** regular office hours; will be announced on Piazza.

- **Questions or discussions:** Piazza

Course Work and Grading Policy

- **Homework (team optional):** 40%
  - 7 x 1-week assignments
- **3xMidterm quizzes (individually):** 40%
- **Final course project (team optional):** 20%
- **Bonus points:**
  - **Extra credit questions (HW and quizzes):** +10%

Homework

- **Timeline:**
  - Released every Tuesday and due in a week.
  - Each have both conceptual (theory) and empirical (coding) parts.
  - All assignments will be submitted to Gradescope.

- **Late day policy**
  - 7 free late days across all assignments; afterwards, 5% off course grade per day late
  - Assignments not accepted more than 7 days late (unless given permission in advance)

- **Collaboration policy:** Allowed within certain limits:
  - Needs to be well-documented; acknowledge working with other students.
  - Avoid: copying code off the web or ChatGPT! — write your own solutions.

- HW1 is released today! Due next week. Submitted via Gradescope.

Final Project

- **No HW after spring break** to let you focus on your final project.
- Topic:
    1. Default projects: we will give you well-defined, scoped projects.
    2. Custom projects: can bring your own idea, as long as it is related to the class.
- This is your chance to hone your technical skills on a topic of interest.
- **Deliverables:**
    1. Submit project **proposal** outline (for our formal review and suggestions)
        - To make sure the proposals are well scoped and doable in limited time.
    2. Get excited 🤩 and work on the project
    3. Midway report
    4. Final project poster session and report

Attendance Policy

- Not mandatory!

Attendance Policy

- Not mandatory!
- No additional grade if you attend.

- However, it is highly encouraged!!
  - Participation in class is our chance to learn more effectively.
  - Come to the class and participate in the discussions!

Quick pulse check (1)

- I am a ….:
  - CS major,
  - CS minor,
  - MS student,
  - PhD student,
  - None of the above??

Quick pulse check (2)

- Which one best describes you?
    1. NLP and DL are rather new to me. Looking forward to learning more.
    2. I am a DL, NLP or self-supervised learning enthusiast – I read casually on these topics.
    3. I am practitioner of NLP or ML and have already worked with a variety of self-supervised models.
    4. I am a pro. I could be part of the teaching staff of this class.

Quick pulse check (3)

- I have read (and understood) the syllabus!
  - Yes
  - No

Wrapping it up!

- HW1 is released!

- If you're not going to take this, drop the course!
  - Many are on the waitlist.