AI in Medicine

Session 1:

Emergence of AI in Healthcare

Andrew Foong, Ph.D.

July 7th 2025

MAYO CLINIC

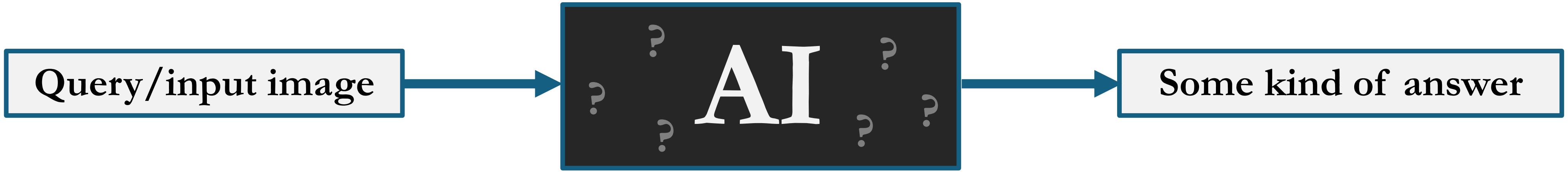
Radiation Oncology
AI & Data Analytics
AIDA

About me

- Senior Associate Consultant, AI: *Radiation Oncology*
- Senior Researcher: *Microsoft Research AI for Science*
- PhD in Machine Learning: *Cambridge University*
- Research Scientist Intern: *Google DeepMind*

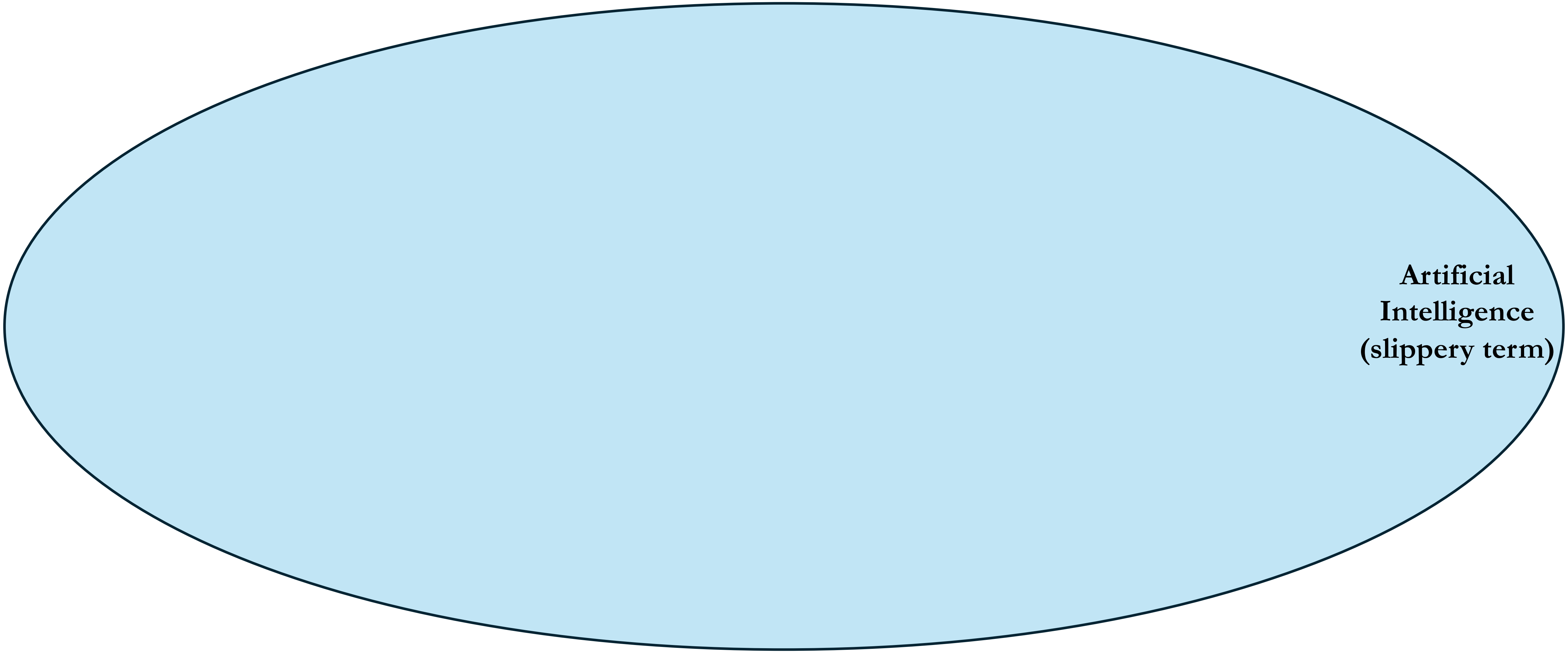
Why do we need to understand AI?

- Many of us use/will use AI in healthcare.
  - Automatic treatment summaries, detecting lung nodules in CT scans, …
  - Black-box model:

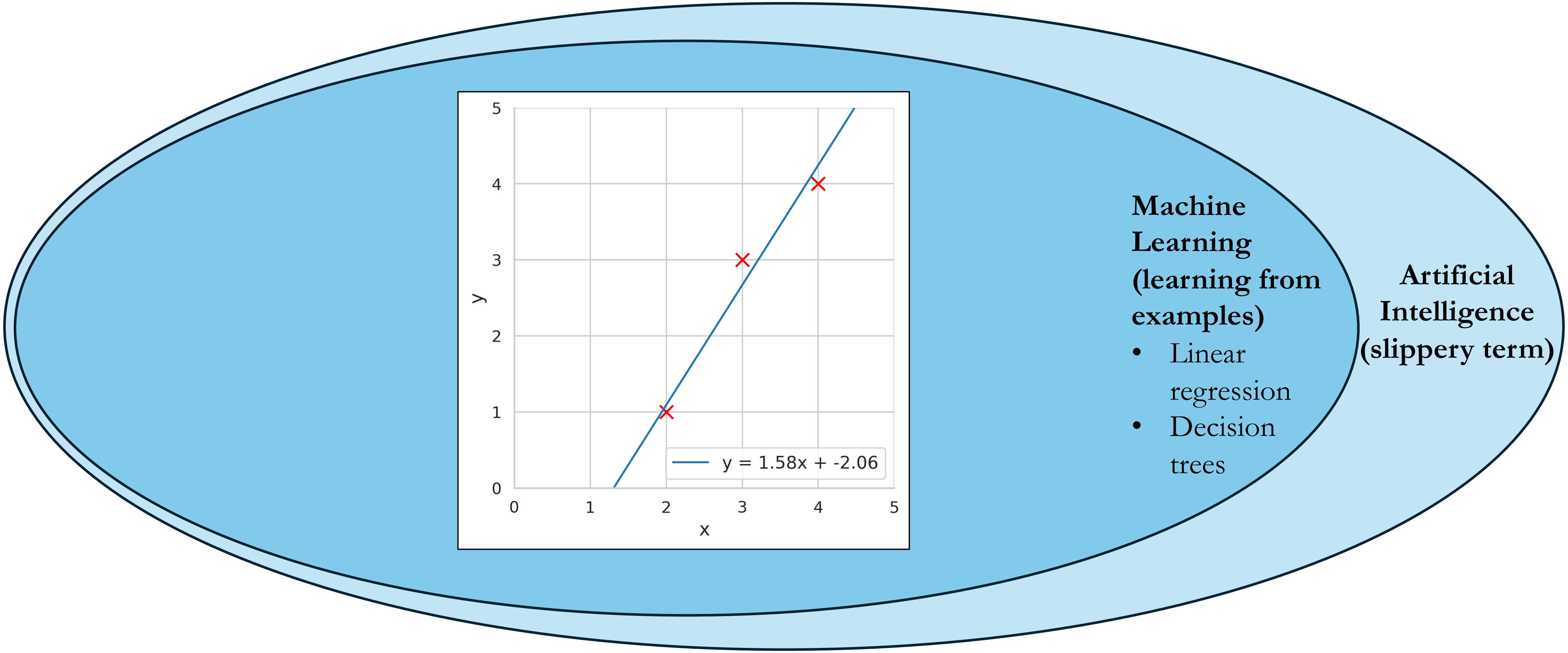


- Why would *healthcare professional* need to *understand* it?
  - **Critical thinking**: When will it work/fail? Don't be passive consumer.
  - **Innovation**: Identify promising applications.
  - **Confidence**: Distinguish hype from reality, call out nonsense.
  - **Scientific curiosity**: One of the greatest engineering achievements.

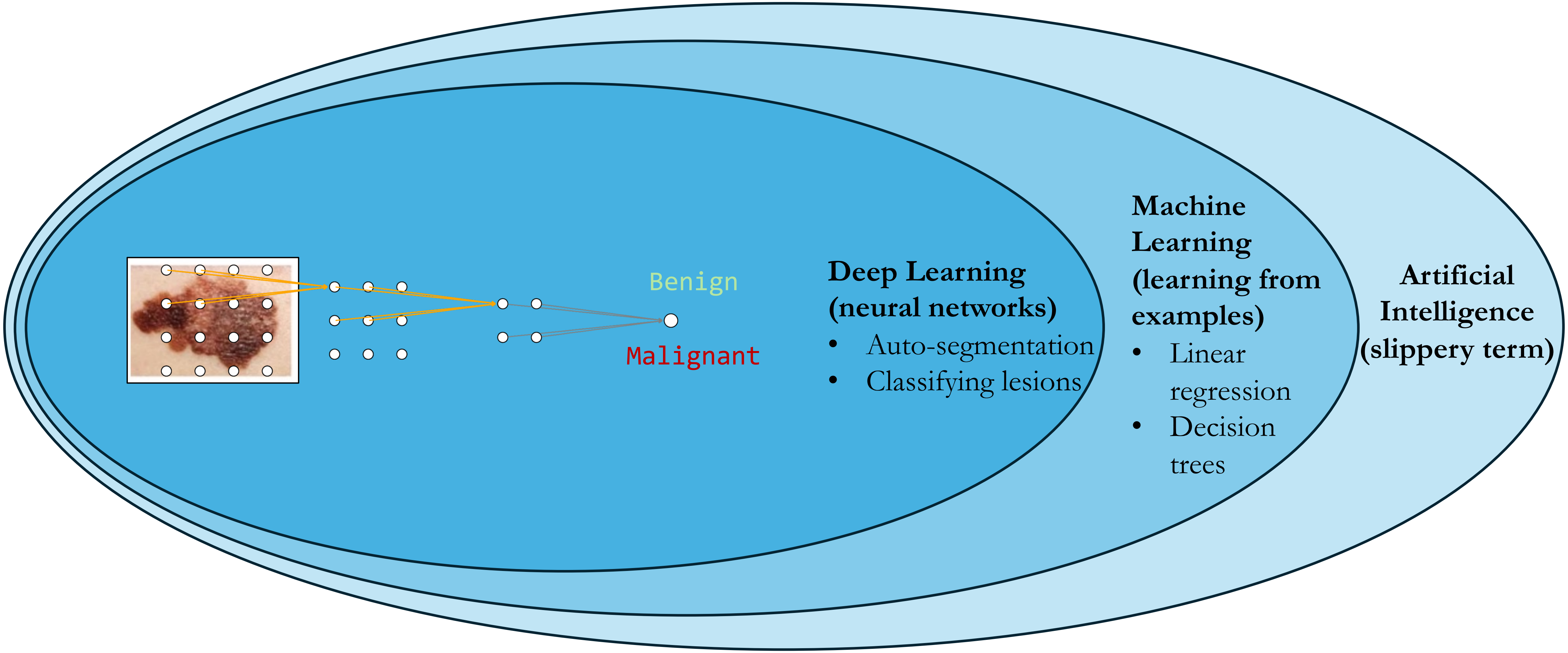
But what *is* AI?



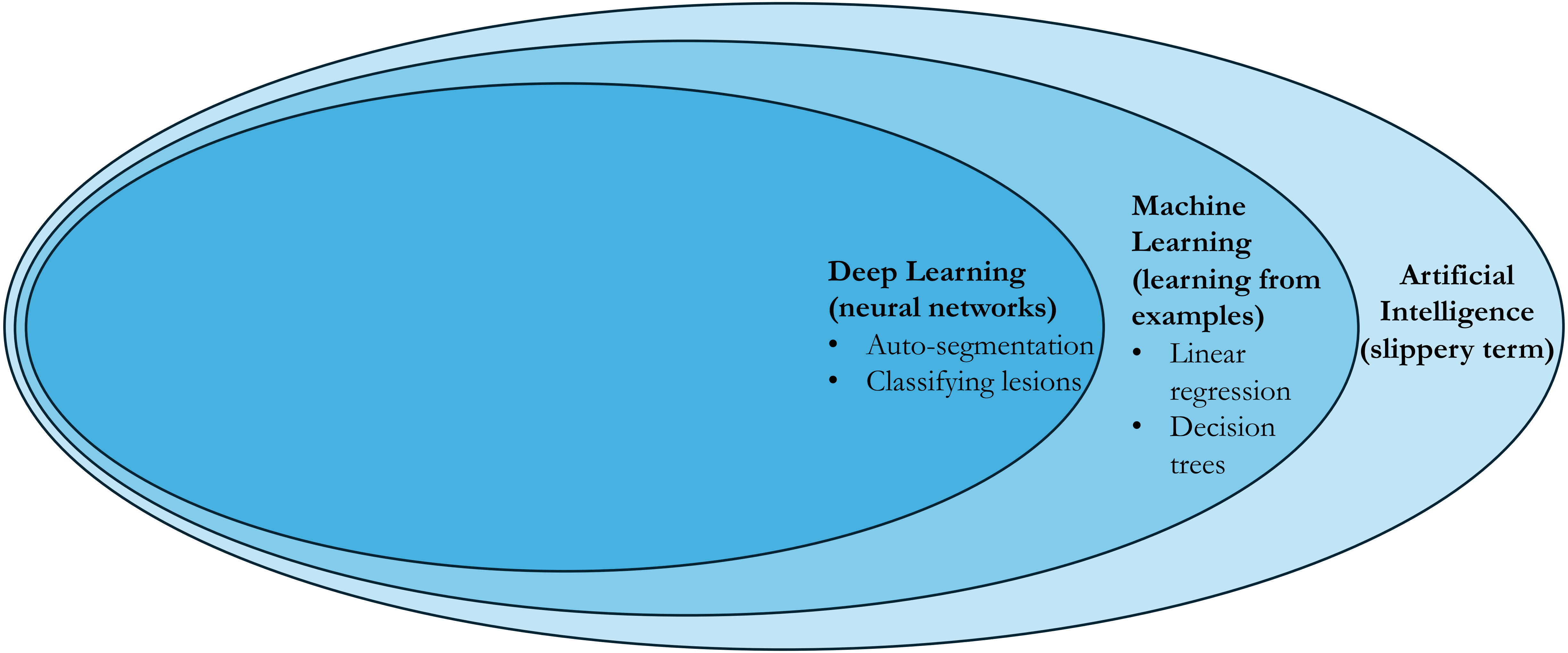
But what *is* AI?



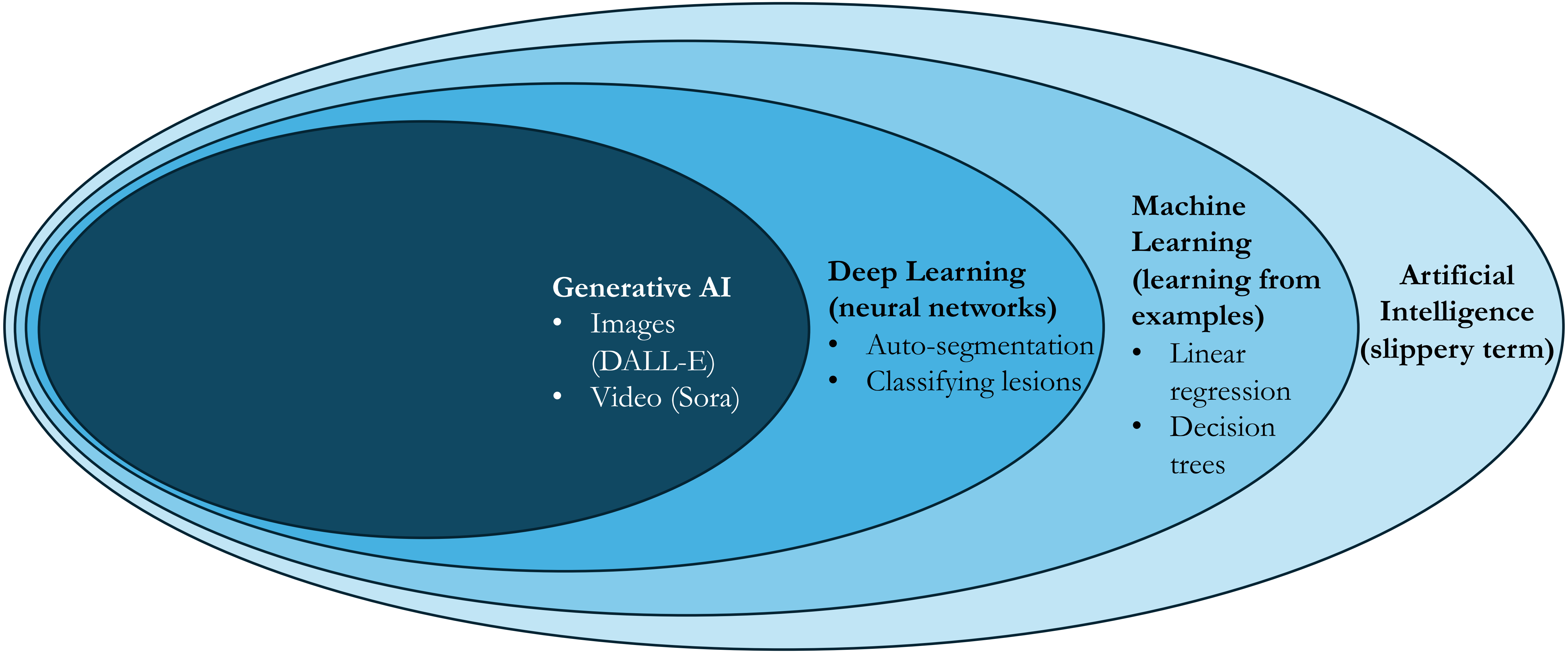
But what *is* AI?



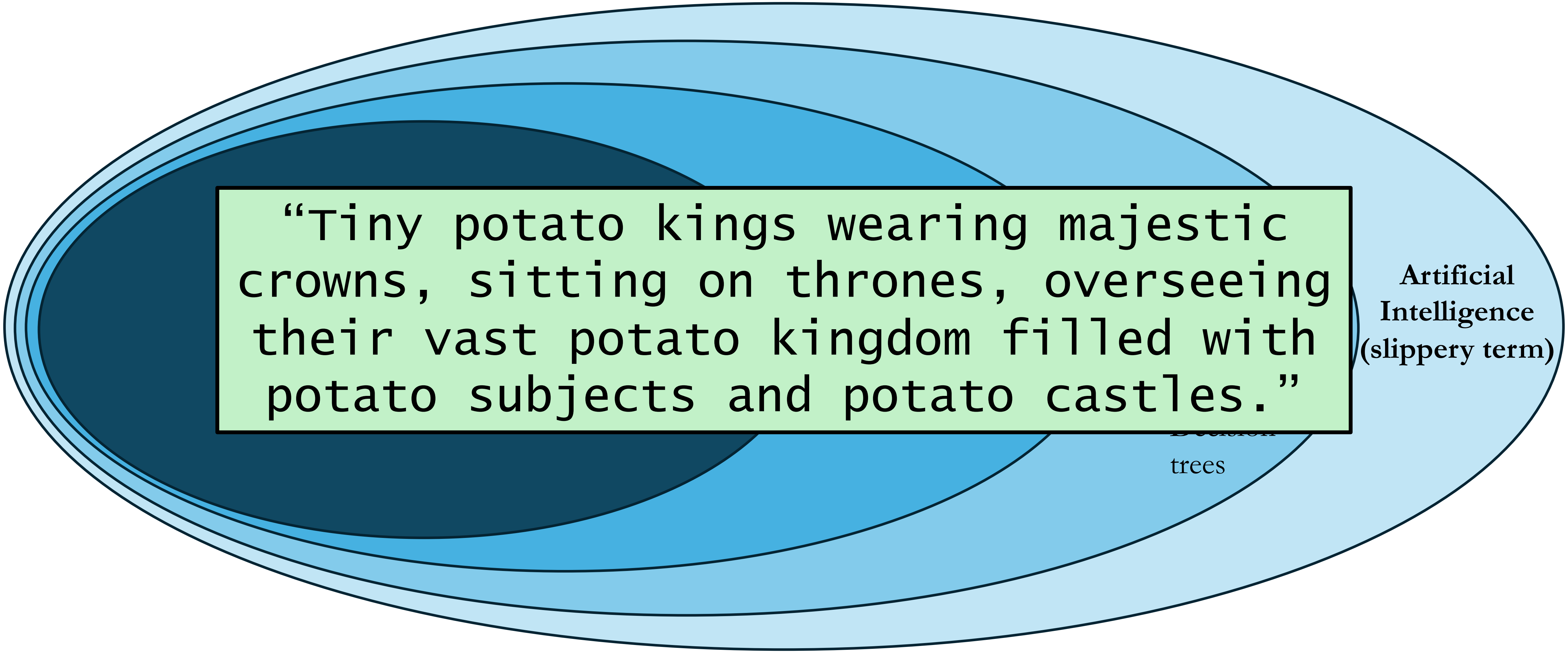
But what *is* AI?



But what *is* AI?



But what *is* AI?



"Tiny potato kings wearing majestic crowns, sitting on thrones, overseeing their vast potato kingdom filled with potato subjects and potato castles."

But what *is* AI?

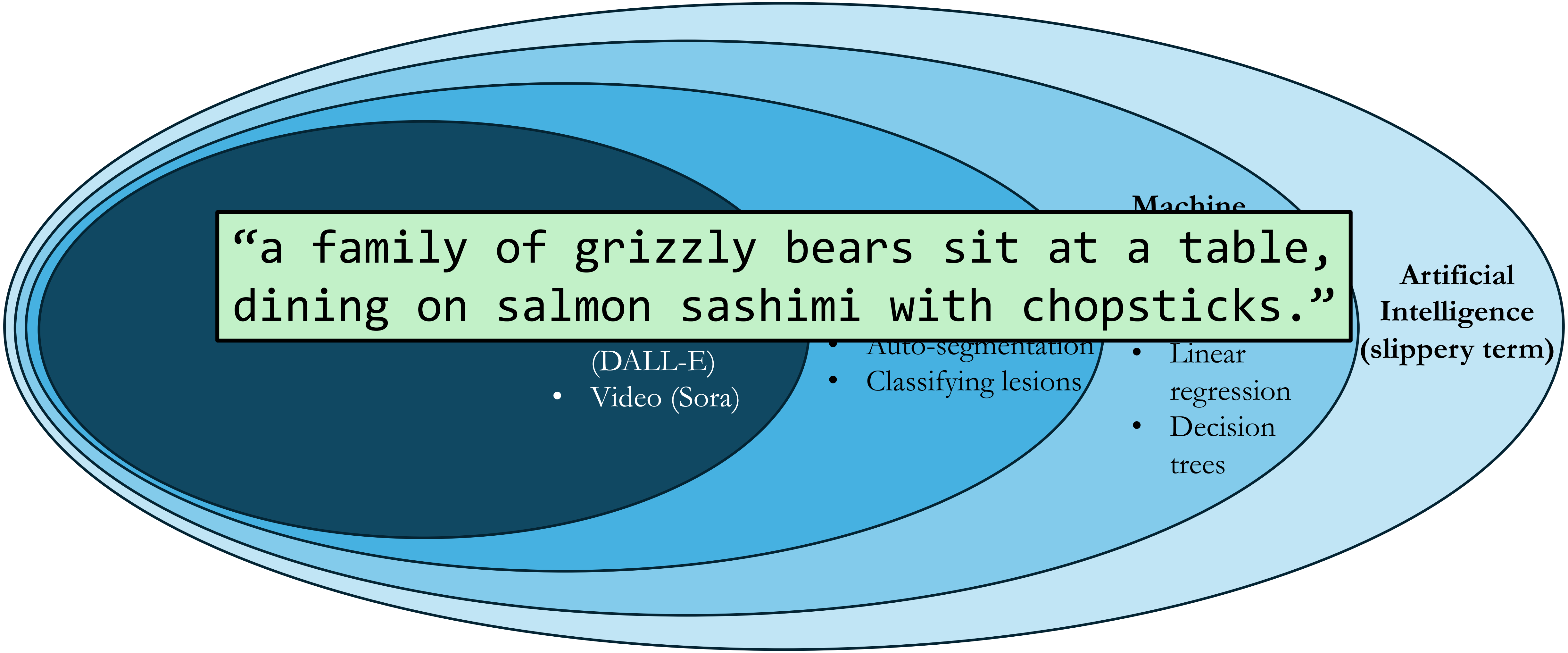
Generative AI
- Images (DALL-E)
- Video (Sora)

Deep Learning (neural networks)
- Auto-segmentation
- Classifying lesions

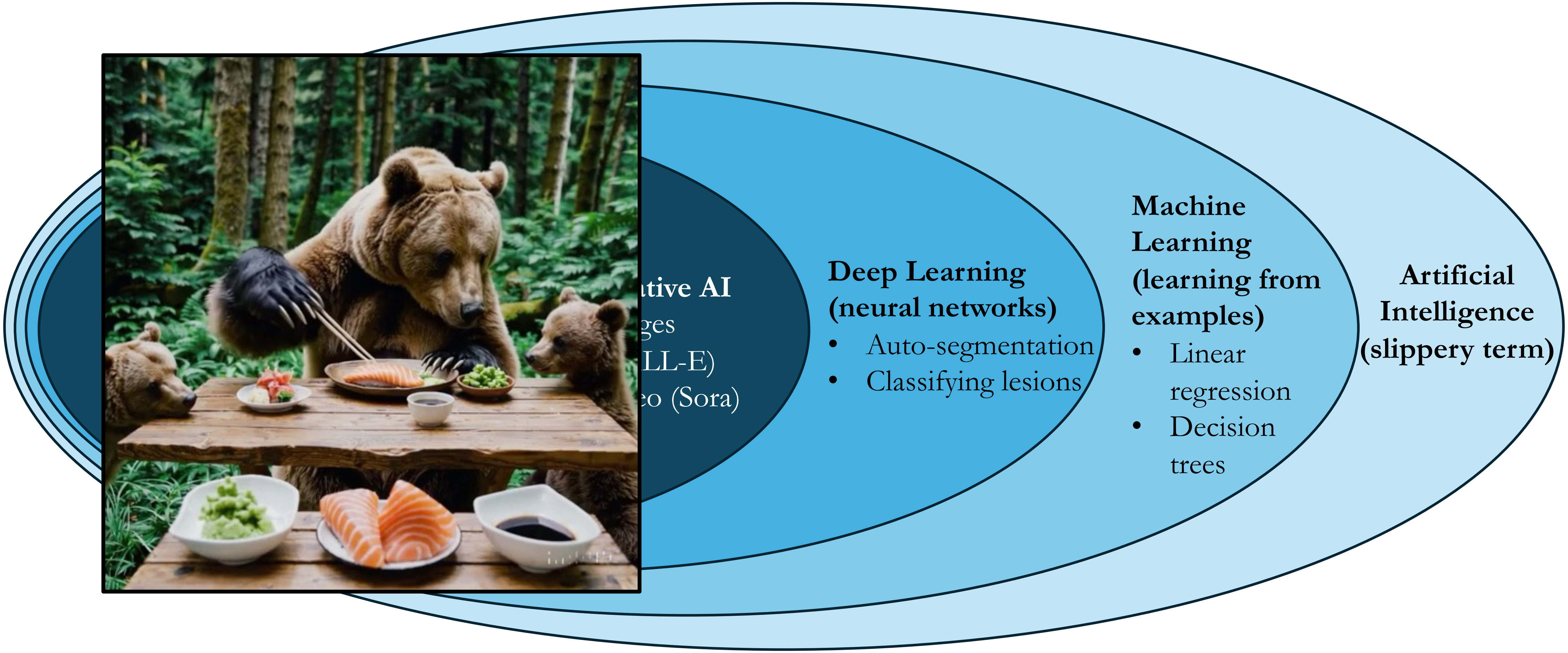
Machine Learning (learning from examples)
- Linear regression
- Decision trees

Artificial Intelligence (slippery term)

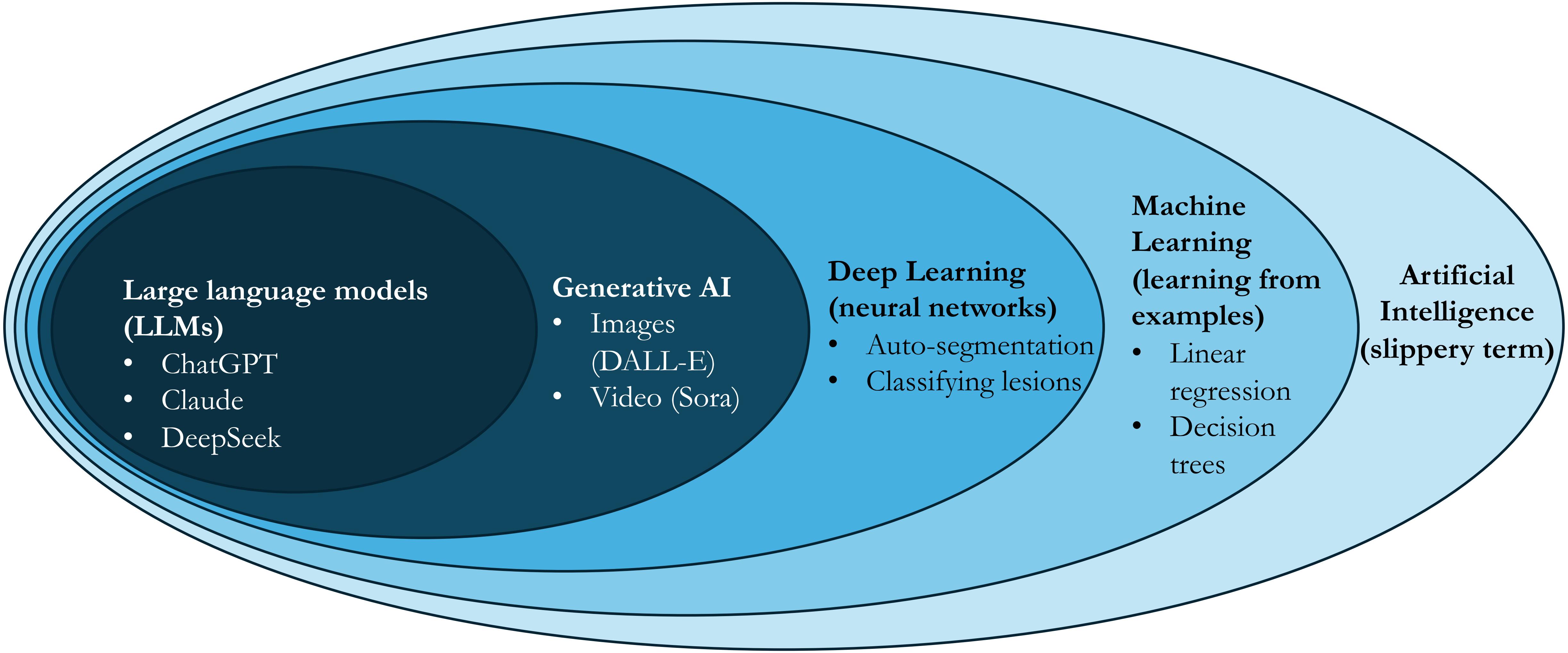
But what *is* AI?



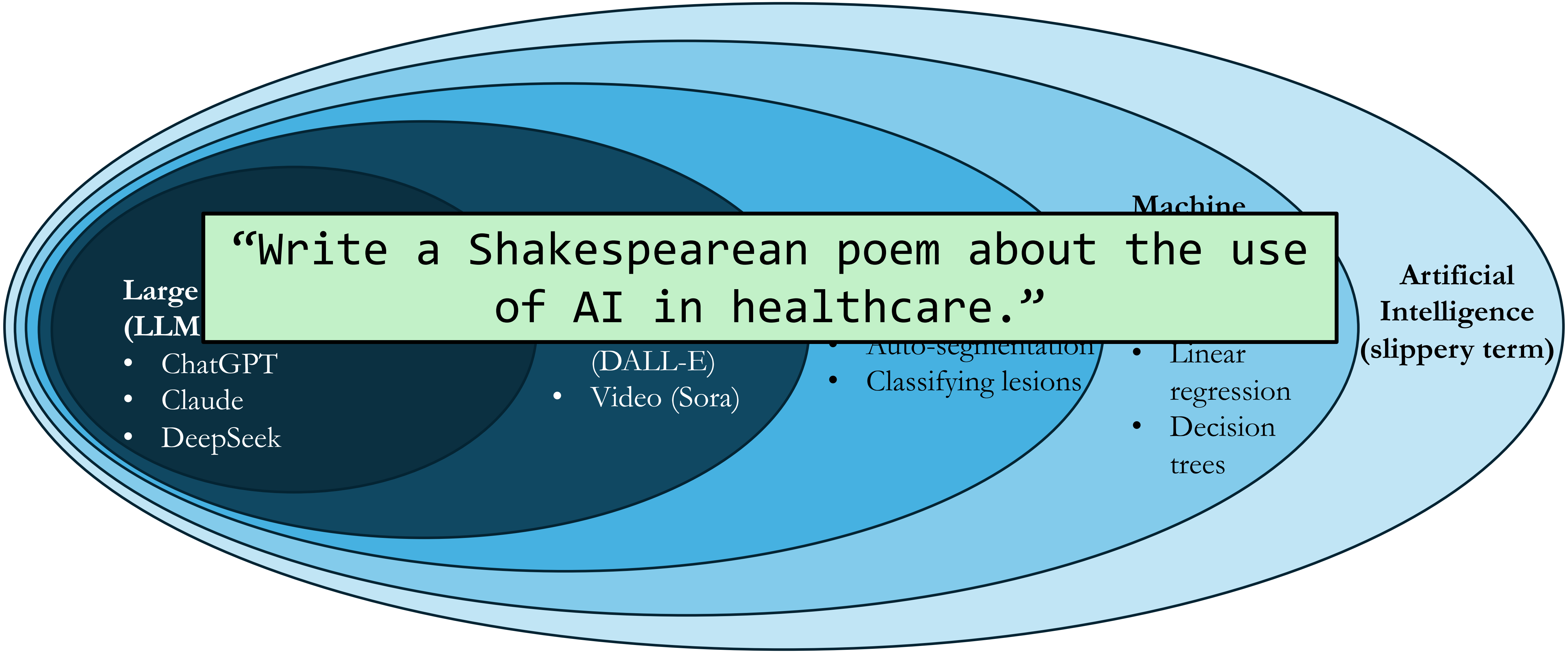
But what *is* AI?



But what *is* AI?



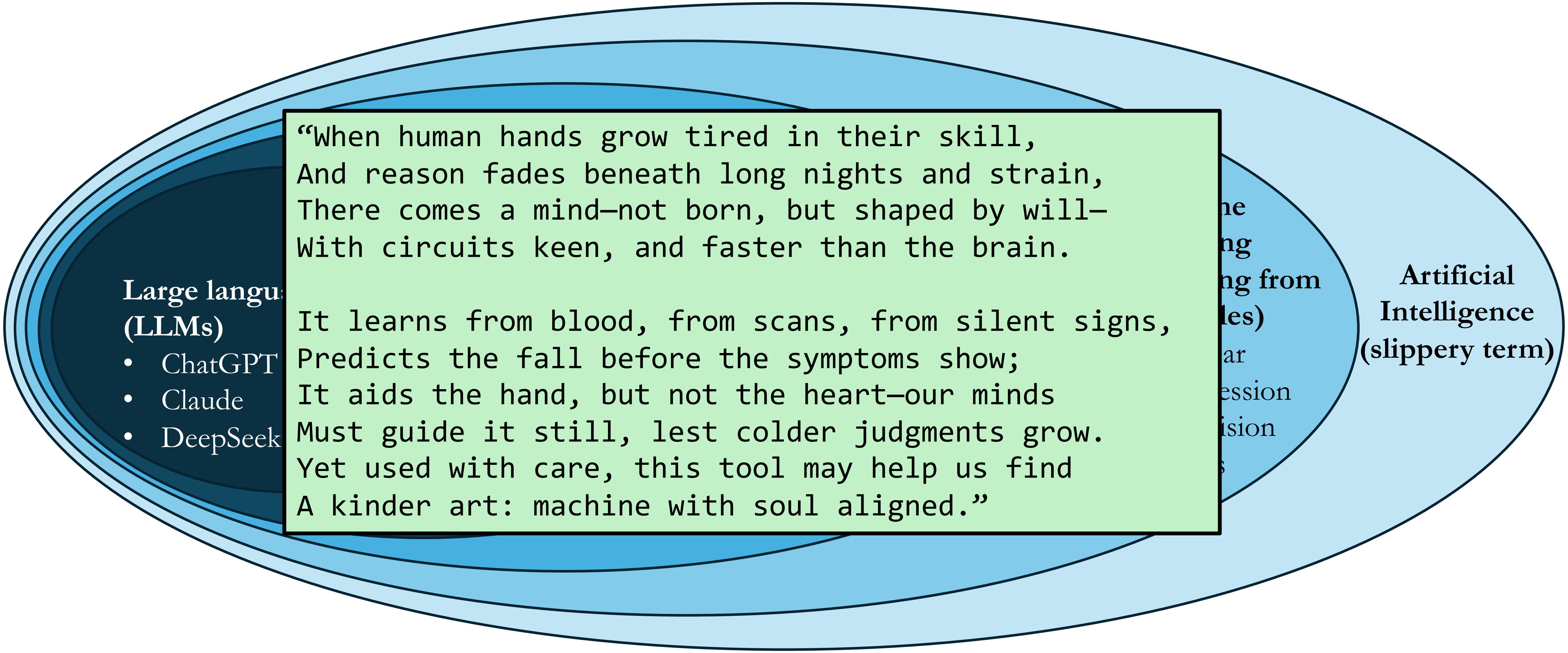
But what *is* AI?



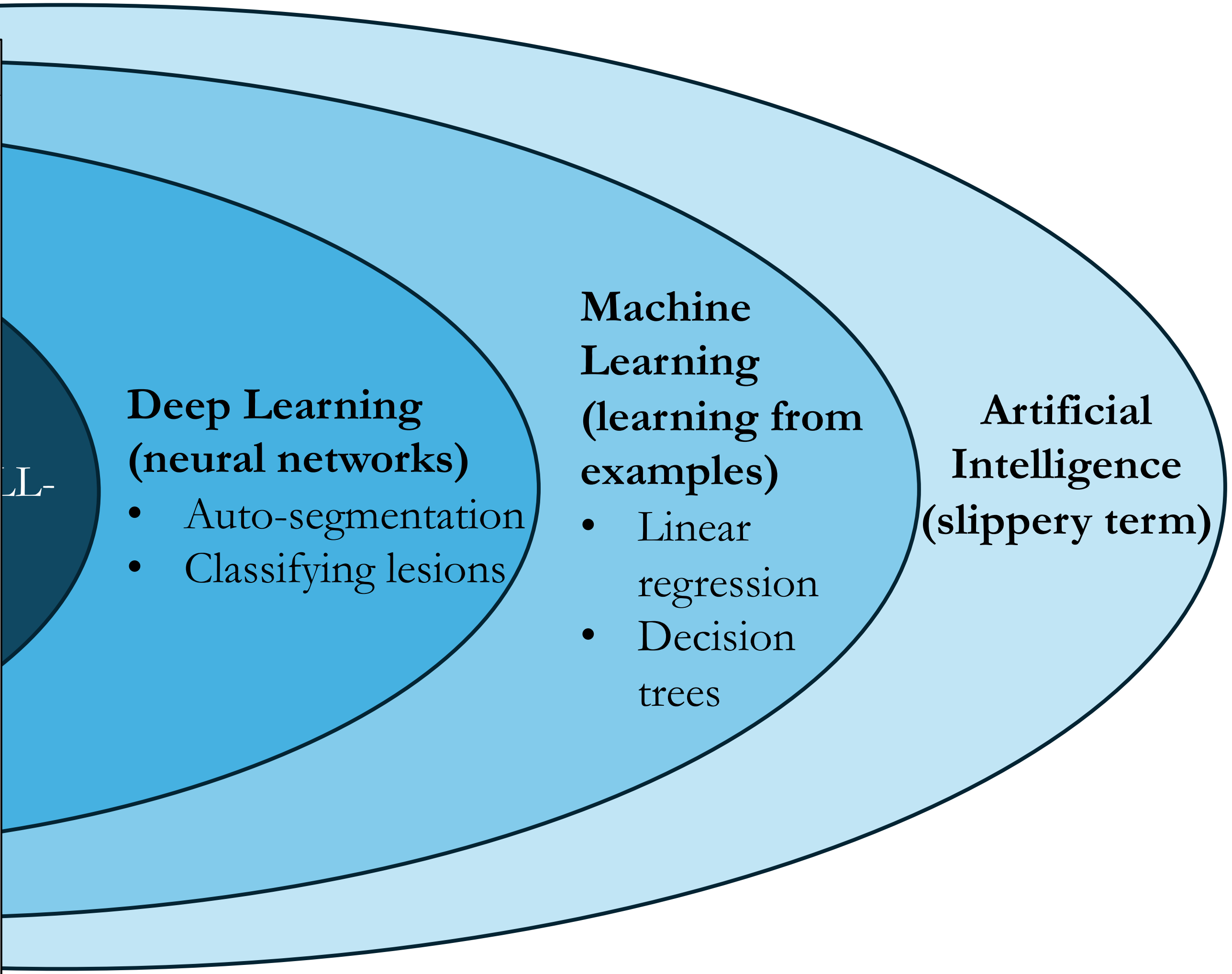
But what *is* AI?

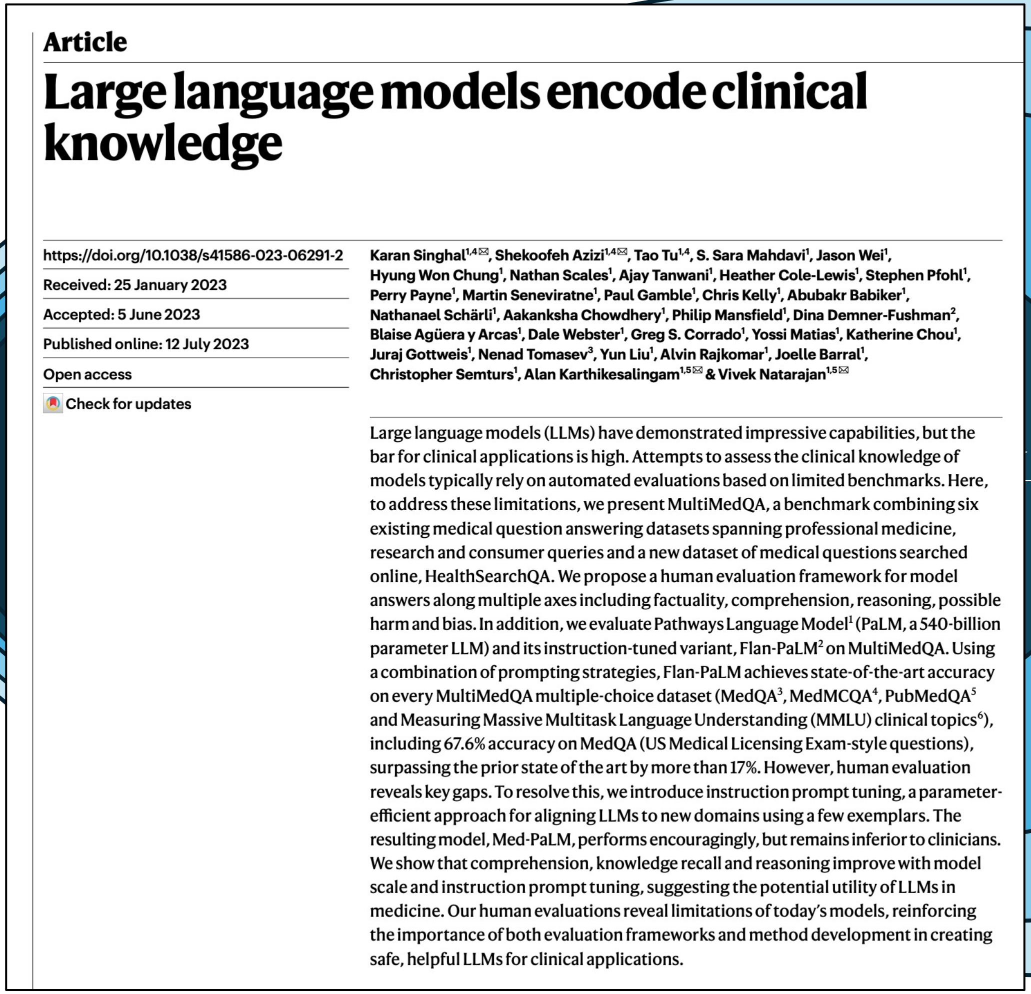
“When human hands grow tired in their skill,
And reason fades beneath long nights and strain,
There comes a mind—not born, but shaped by will—
With circuits keen, and faster than the brain.

It learns from blood, from scans, from silent signs,
Predicts the fall before the symptoms show;
It aids the hand, but not the heart—our minds
Must guide it still, lest colder judgments grow.
Yet used with care, this tool may help us find
A kinder art: machine with soul aligned.”



But what *is* AI?




Article

Large language models encode clinical knowledge

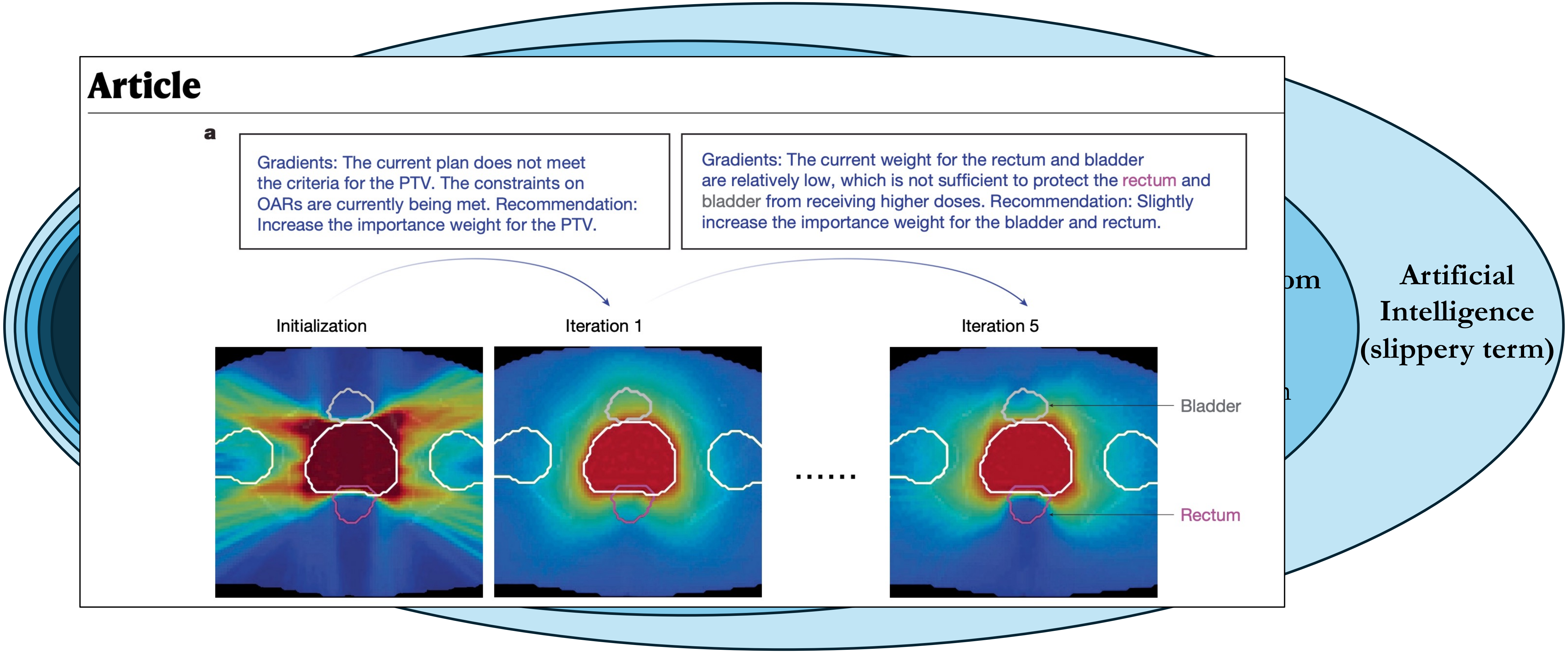
https://doi.org/10.1038/s41586-023-06291-2
Received: 25 January 2023
Accepted: 5 June 2023
Published online: 12 July 2023
Open access
Check for updates

Karan Singhal[1,4]✉, Shekoofeh Azizi[1,4]✉, Tao Tu[1,4], S. Sara Mahdavi[1], Jason Wei[1], Hyung Won Chung[1], Nathan Scales[1], Ajay Tanwani[1], Heather Cole-Lewis[1], Stephen Pfohl[1], Perry Payne[1], Martin Seneviratne[1], Paul Gamble[1], Chris Kelly[1], Abubakr Babiker[1], Nathanael Schärli[1], Aakanksha Chowdhery[1], Philip Mansfield[1], Dina Demner-Fushman[2], Blaise Agüera y Arcas[1], Dale Webster[1], Greg S. Corrado[1], Yossi Matias[1], Katherine Chou[1], Juraj Gottweis[1], Nenad Tomasev[3], Yun Liu[1], Alvin Rajkomar[1], Joelle Barral[1], Christopher Semturs[1], Alan Karthikesalingam[1,5]✉ & Vivek Natarajan[1,5]✉

Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high. Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Here, to address these limitations, we present MultiMedQA, a benchmark combining six existing medical question answering datasets spanning professional medicine, research and consumer queries and a new dataset of medical questions searched online, HealthSearchQA. We propose a human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possible harm and bias. In addition, we evaluate Pathways Language Model[1] (PaLM, a 540-billion parameter LLM) and its instruction-tuned variant, Flan-PaLM[2] on MultiMedQA. Using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-choice dataset (MedQA[3], MedMCQA[4], PubMedQA[5] and Measuring Massive Multitask Language Understanding (MMLU) clinical topics[6]), including 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of the art by more than 17%. However, human evaluation reveals key gaps. To resolve this, we introduce instruction prompt tuning, a parameter-efficient approach for aligning LLMs to new domains using a few exemplars. The resulting model, Med-PaLM, performs encouragingly, but remains inferior to clinicians. We show that comprehension, knowledge recall and reasoning improve with model scale and instruction prompt tuning, suggesting the potential utility of LLMs in medicine. Our human evaluations reveal limitations of today's models, reinforcing the importance of both evaluation frameworks and method development in creating safe, helpful LLMs for clinical applications.

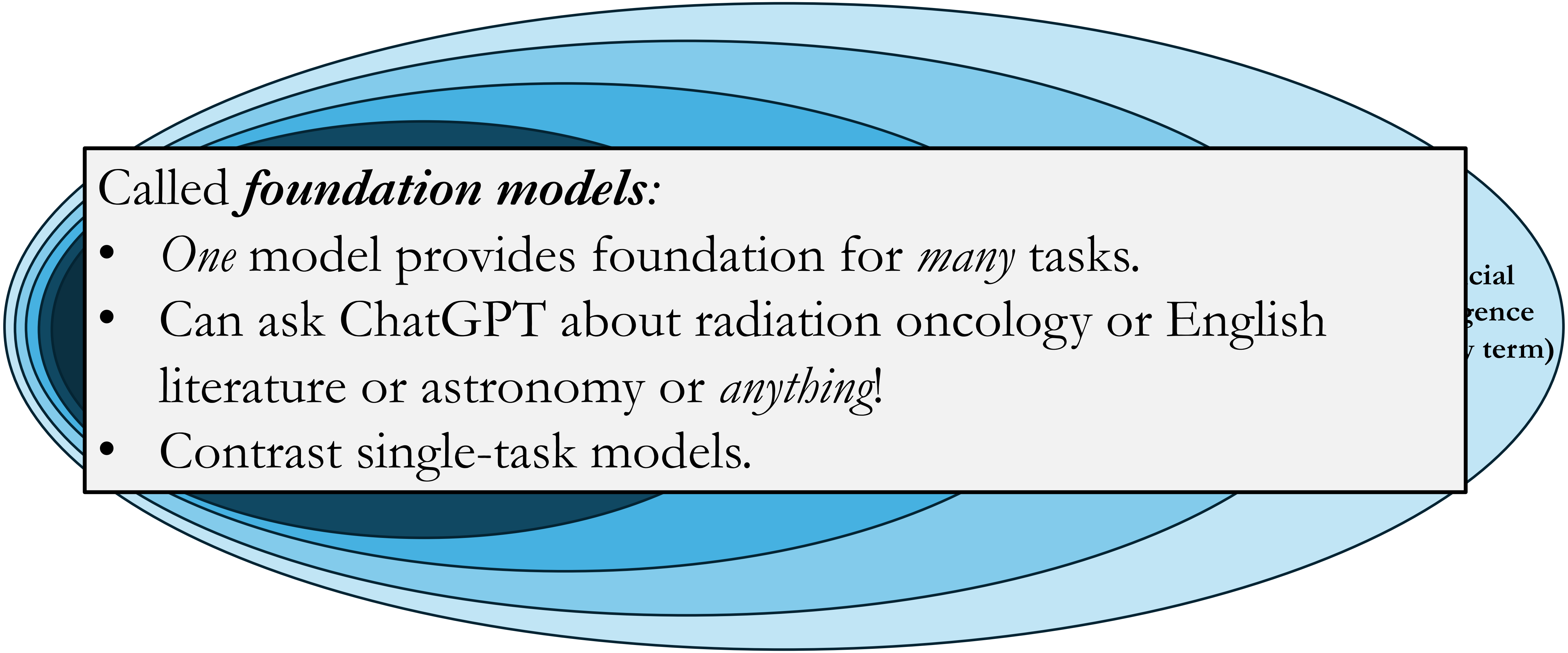
Singhal et al. 2023, Large language models encode clinical knowledge, *Nature*

But what *is* AI?



Yuksekgonul et al. 2023, Optimizing generative AI by backpropagating language model feedback, *Nature*

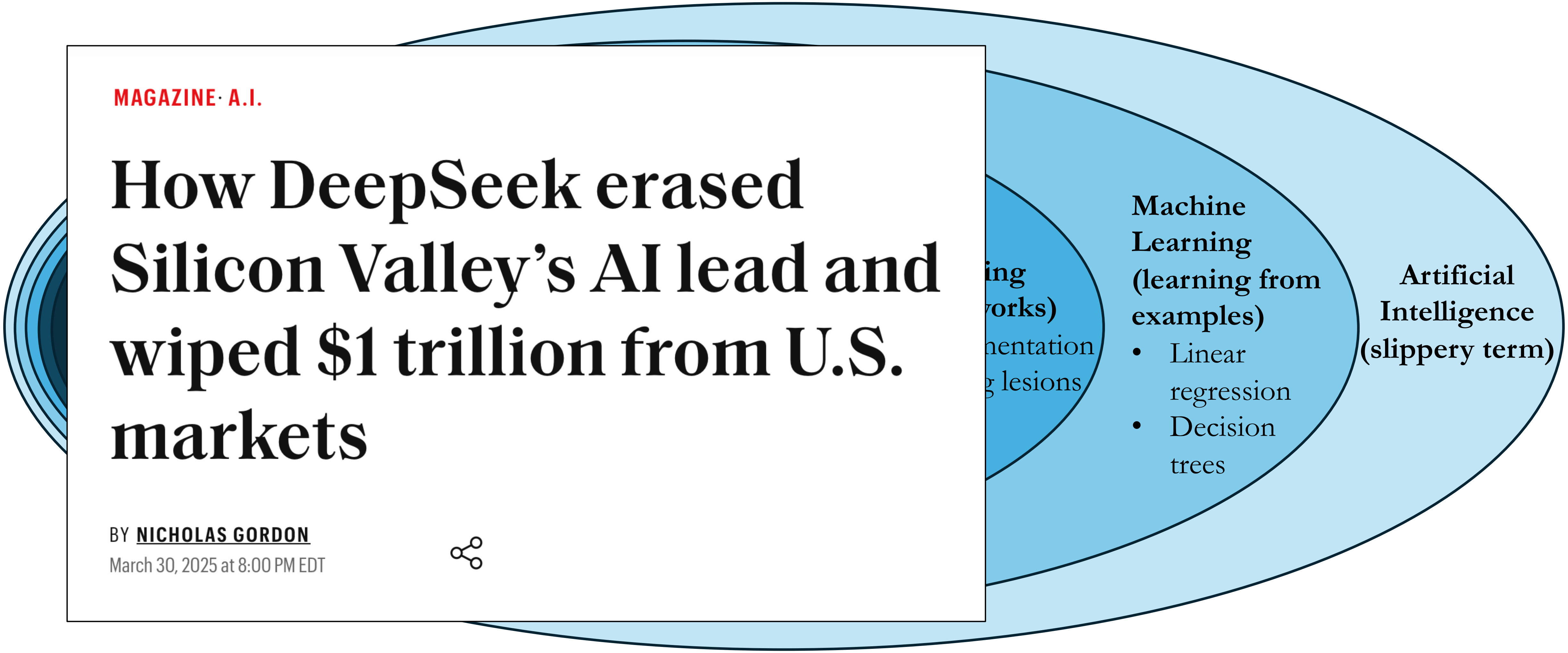
But what *is* AI?



Called ***foundation models*:**

- *One* model provides foundation for *many* tasks.
- Can ask ChatGPT about radiation oncology or English literature or *anything*!
- Contrast single-task models.

But what *is* AI?



MAGAZINE · A.I.

How DeepSeek erased Silicon Valley's AI lead and wiped $1 trillion from U.S. markets

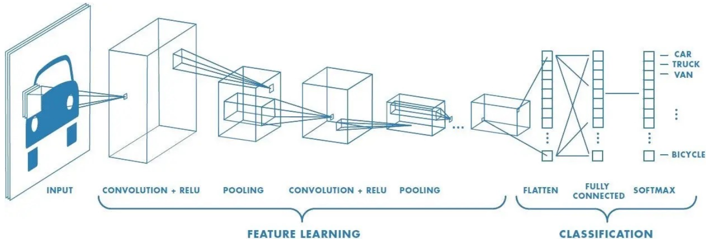
BY NICHOLAS GORDON
March 30, 2025 at 8:00 PM EDT

Nicholas Gordon, *Fortune*, March 30, 2025

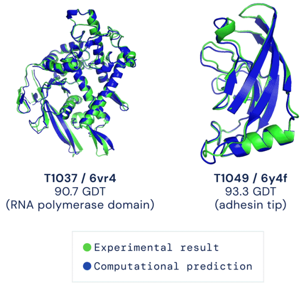
But what *is* AI?

The deep learning revolution

AlexNet:
Image classification



AlphaFold:
Protein structure prediction



ChatGPT:
Intelligent chatbot

2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025

AlphaGo:
Playing board games

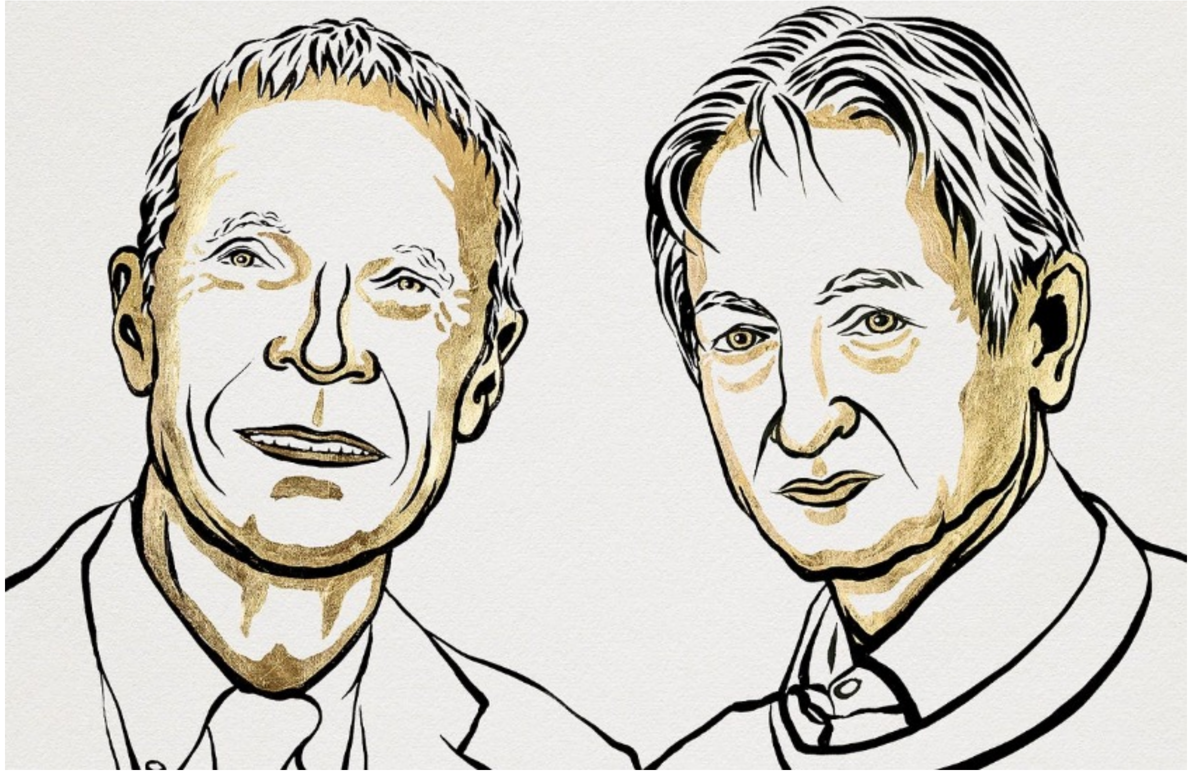
DALL-E:
Text-to-image generation

Your idea here

MAYO CLINIC

Already revolutionized science

2024 Nobel Prize in Physics



John Hopfield
Princeton University

Geoffrey Hinton
University of Toronto

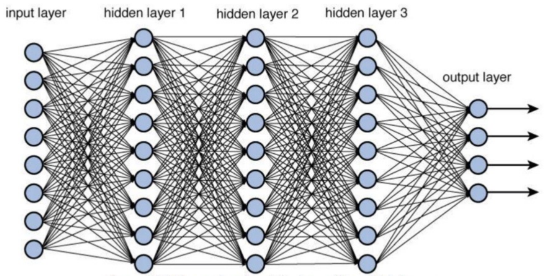
2024 Nobel Prize in Chemistry

Demis Hassabis
Google DeepMind

John Jumper
Google DeepMind

Mayo Clinic is betting that deep learning will revolutionize healthcare

Roadmap



Part 1: How does deep learning work?

Part 2: How do foundation models work?

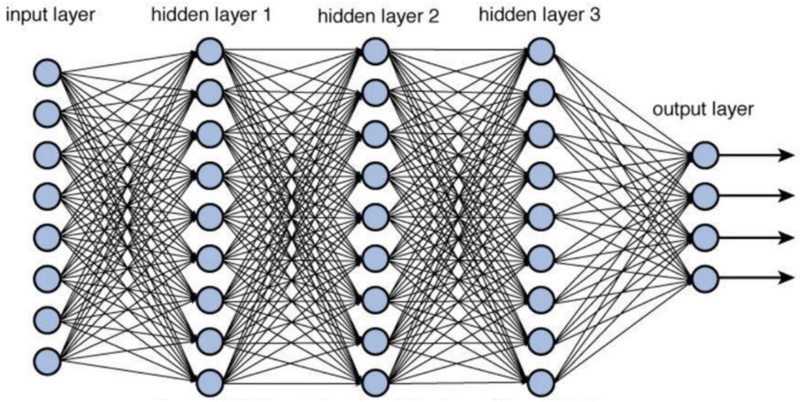
Goal: *intuition* and *general principles*.
Applications covered in Day 2.

Part 1:

How Does Deep Learning Work?

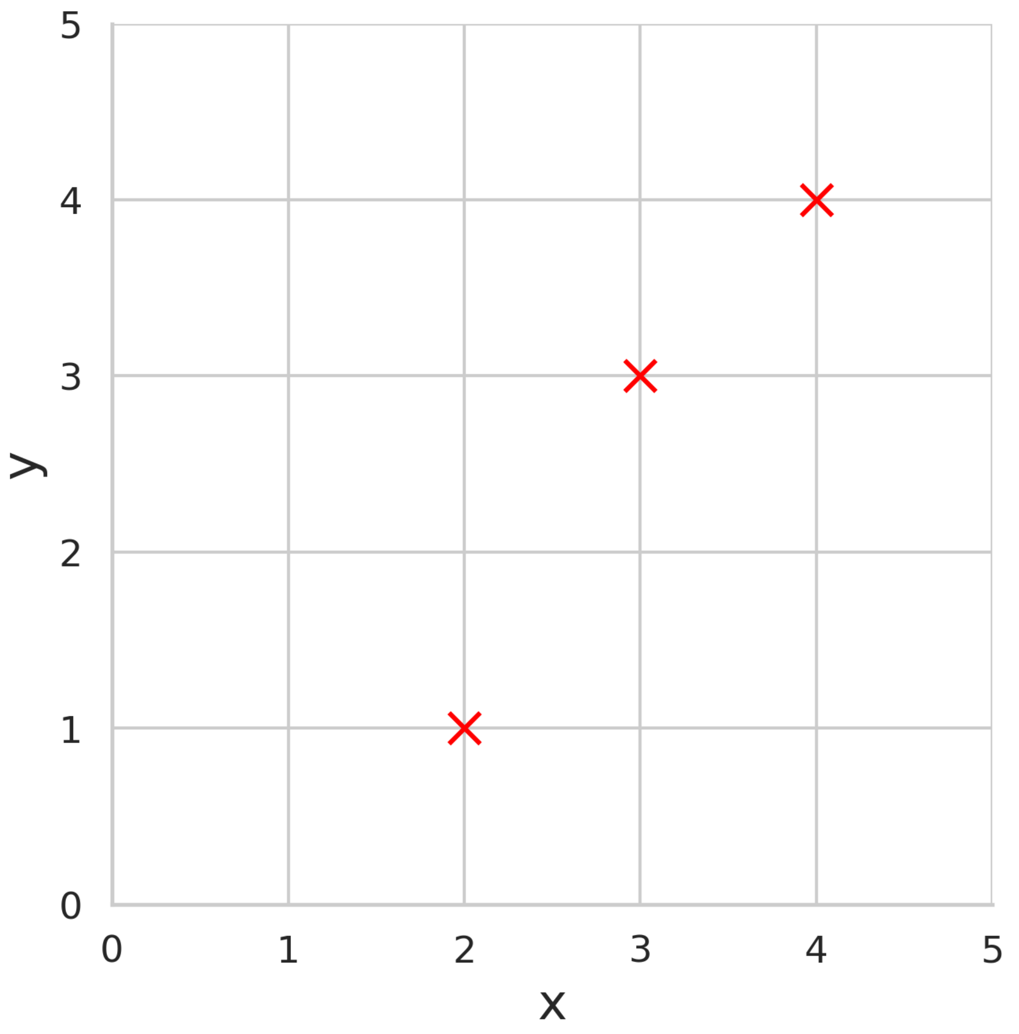
Deep learning from 40,000 feet

- Deep learning = use of **neural networks**.
- Math functions with millions of numbers: **“parameters/weights”**
- The numbers determine how the neural network behaves.
    1. Start by choosing parameters randomly (*garbage predictions*).
    2. **Optimizer** automatically adjusts parameters to fit example data.
    3. Apply the function to new data (*great predictions, hopefully*).



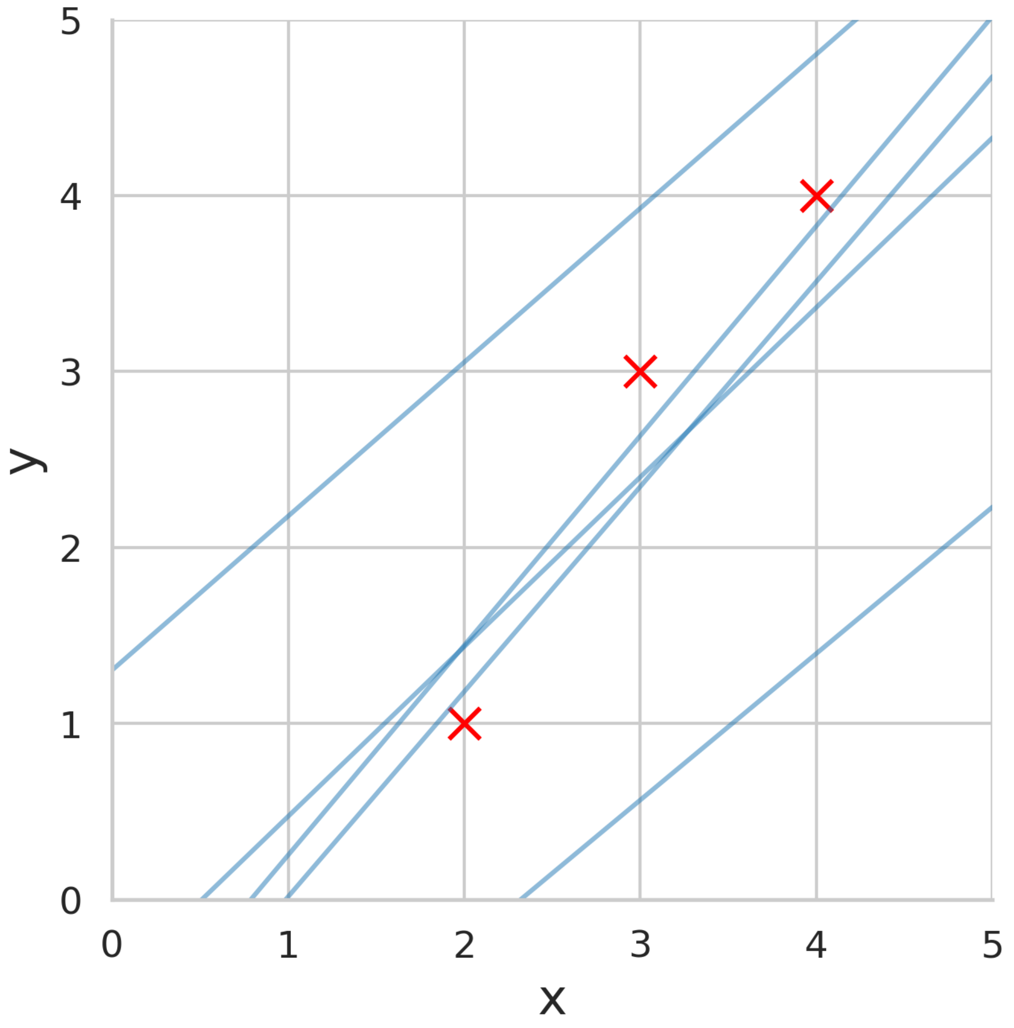
First step: linear regression

- To understand neural networks, start with **linear regression**.
- *Solving this uses same ideas behind ChatGPT or image classification!*
- Given 3 datapoints:
  - For any x, predict y.
- Simplest idea: straight line.
  - But points not in straight line!
  - Any straight line will miss points.



First step: linear regression

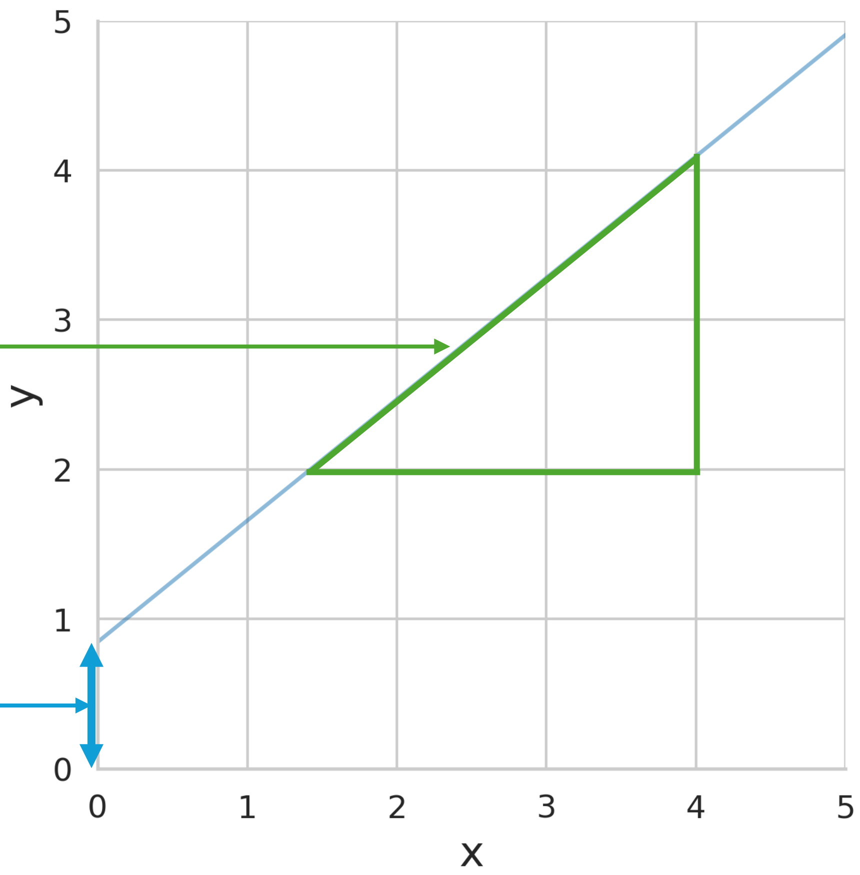
- To understand neural networks, start with **linear regression**.
- *Solving this uses same ideas behind ChatGPT or image classification!*
- Given 3 datapoints:
  - For any x, predict y.
- Simplest idea: straight line.
  - But points not in straight line!
  - Any straight line will miss points.
  - Which line to choose?



Linear regression parameters

- Need two numbers to define a line:
  - Slope/gradient
  - Intercept/bias

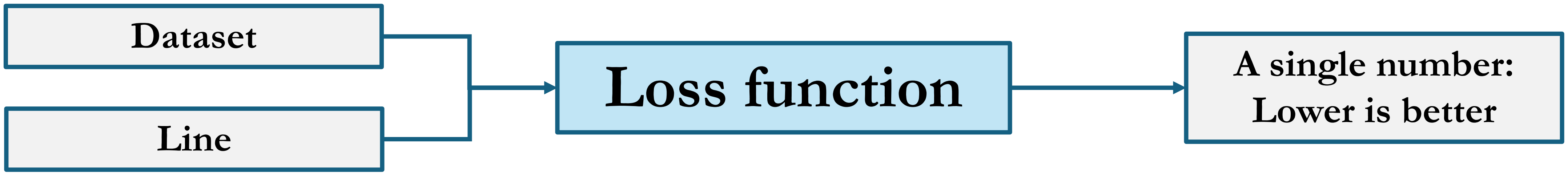
$$f(x) = wx + b$$



- Once w, b specified, can easily compute $f(x)$ at *any* x.
- w, b are the **parameters/weights.**
- Neural networks have them too, but *many* more.
- Reframe problem:
  - *Choose best line → choose best parameters/weights.*

Picking the best line

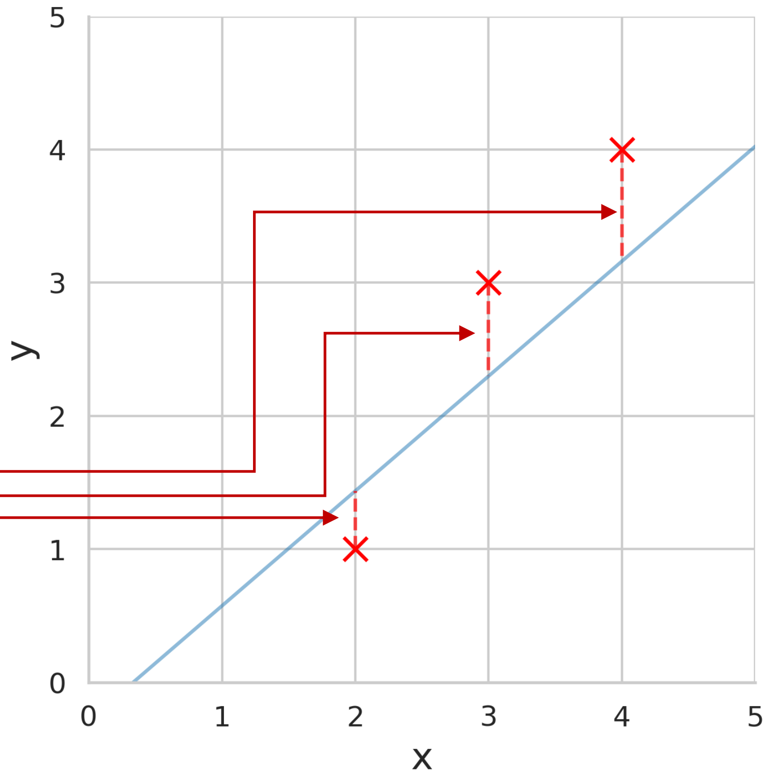
- Need rule to choose "best" line.
- No one "correct" answer: must *invent* something.
- Define "**loss**" function to measure how far points are from line.
  - High loss → line fits data poorly.
  - Low loss → line fits data well.
- Crucial step in *all* deep learning models.



Choosing a loss function

- Choosing loss function is an art.
  - Defines what a *good* prediction is.
- Simple choice: **squared error**

$$L(w, b) = \underbrace{(f(x_1) - y_1)}^2 + \underbrace{(f(x_2) - y_2)}^2 + \underbrace{(f(x_3) - y_3)}^2$$



Optimization

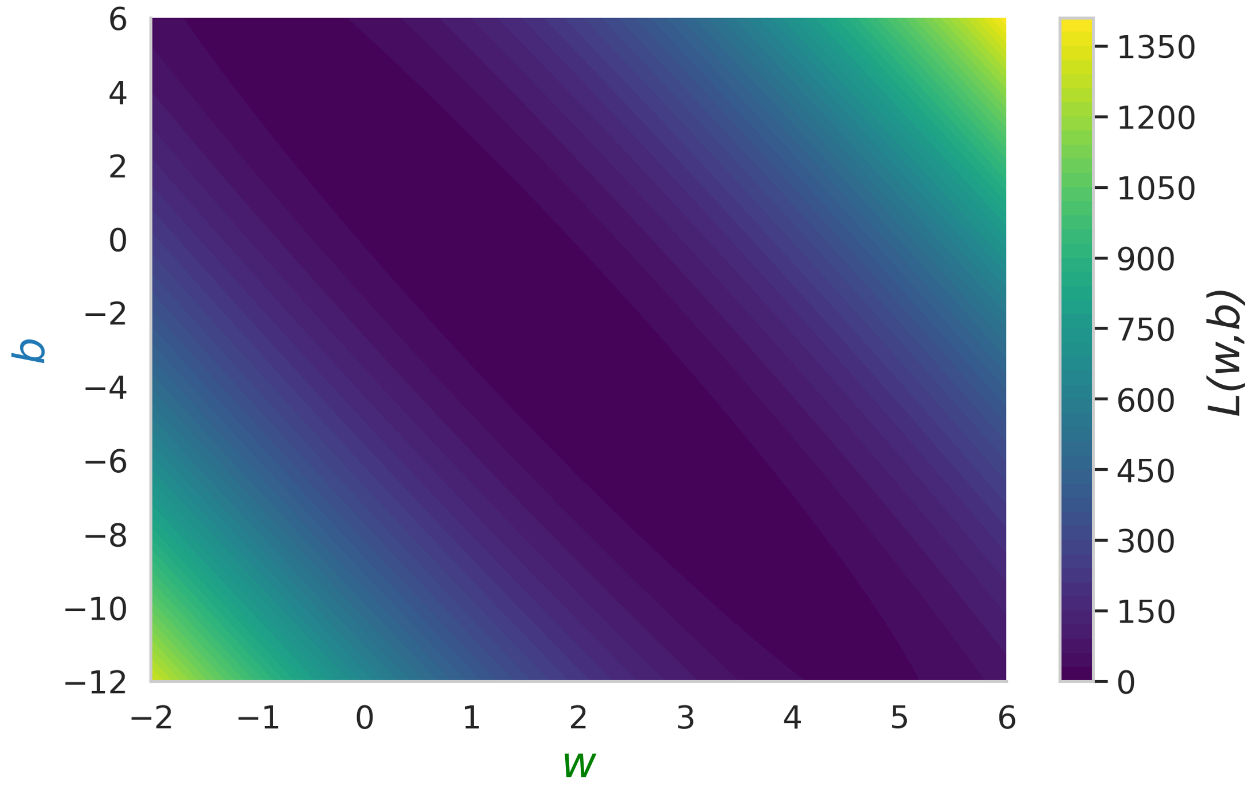
- Choose best fit line →

  Find w, b that minimizes loss: $L(w, b)$.

- Use **gradient descent**:
  1. Choose w, b randomly *(terrible fit!)*.
  2. Calculate **derivative/gradient** of $L(w, b)$.
  3. Adjust w, b by small amount in (opposite) direction of gradient.
  4. Repeat 2–3 until $L(w, b)$ is low *(good fit!)*.
- *Gradient descent is how all neural networks learn!*

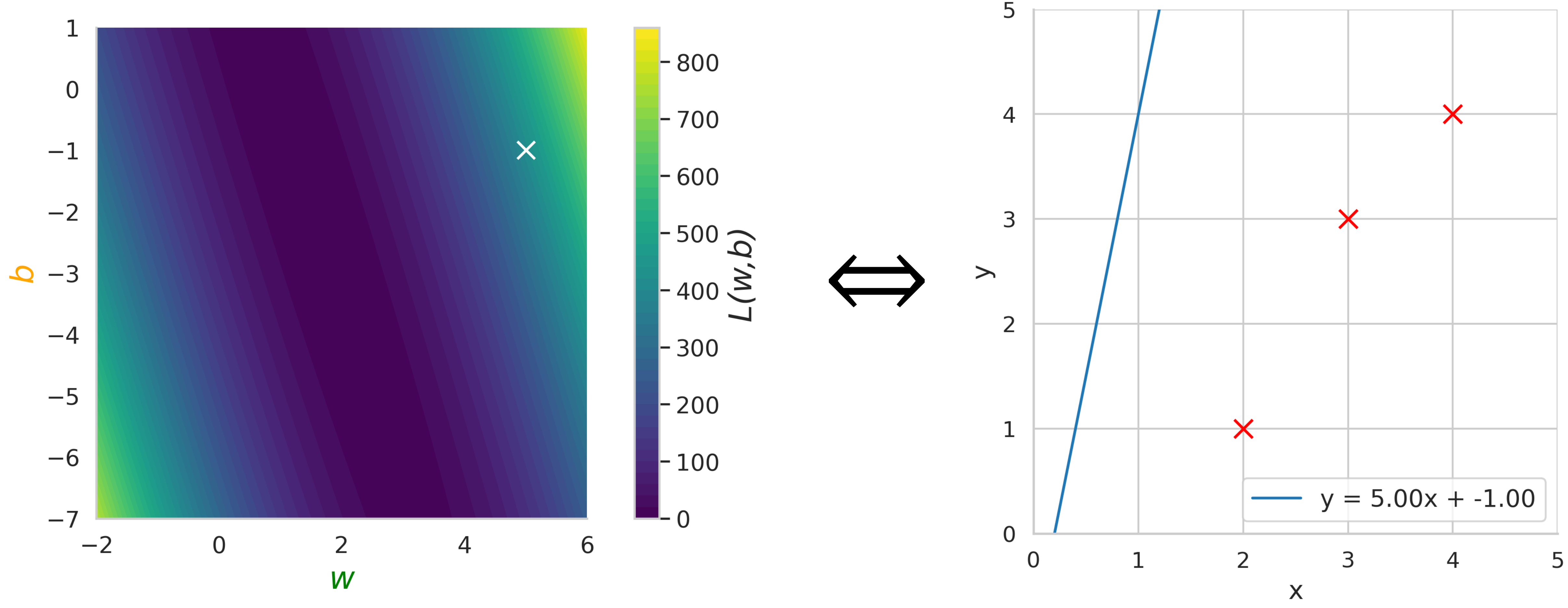
Gradient descent for linear regression

- Find w, b that minimizes $L(w, b)$.
- 2 parameters $\rightarrow$ 2D *surface*!
- Each point represents choice of w, b.
- Yellow $\rightarrow$ higher loss $\rightarrow$ worse fit
- Blue $\rightarrow$ lower loss $\rightarrow$ better fit



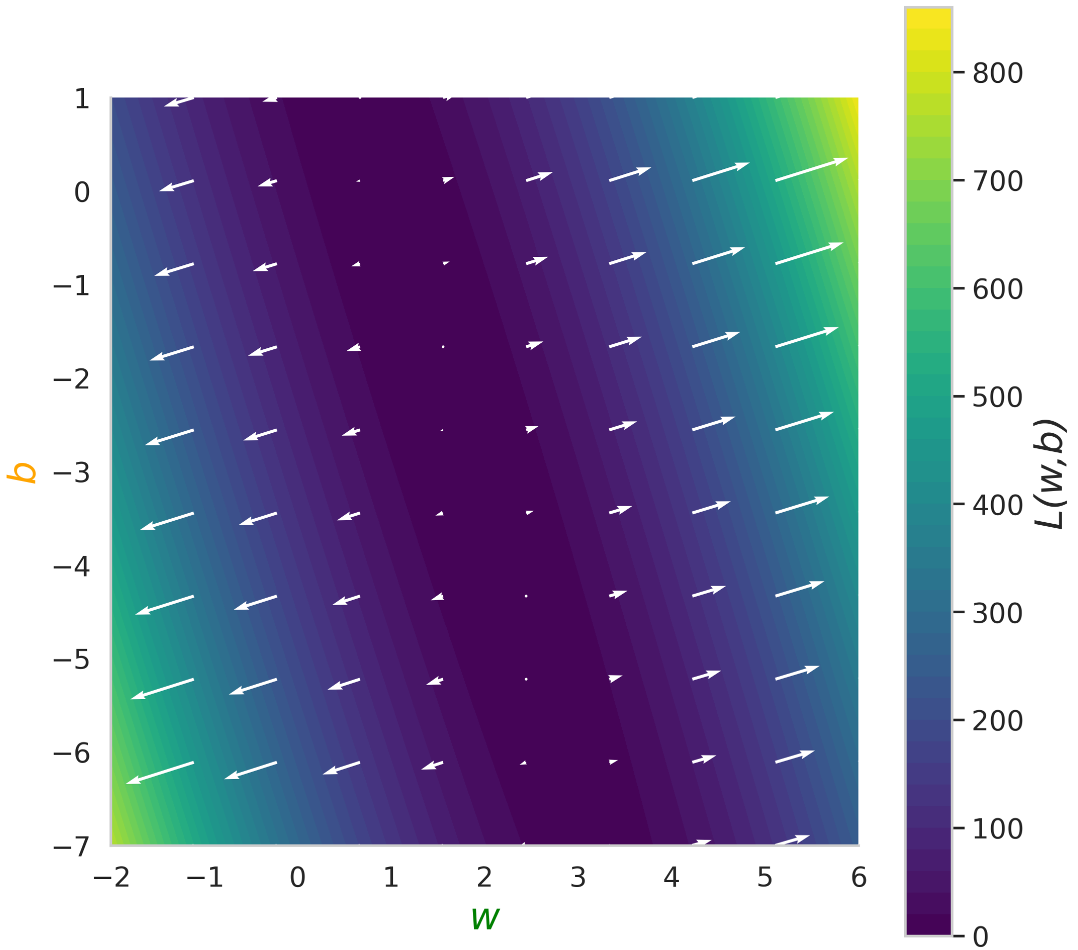
Gradient descent for linear regression

1. Choose w, b randomly (*terrible fit*).

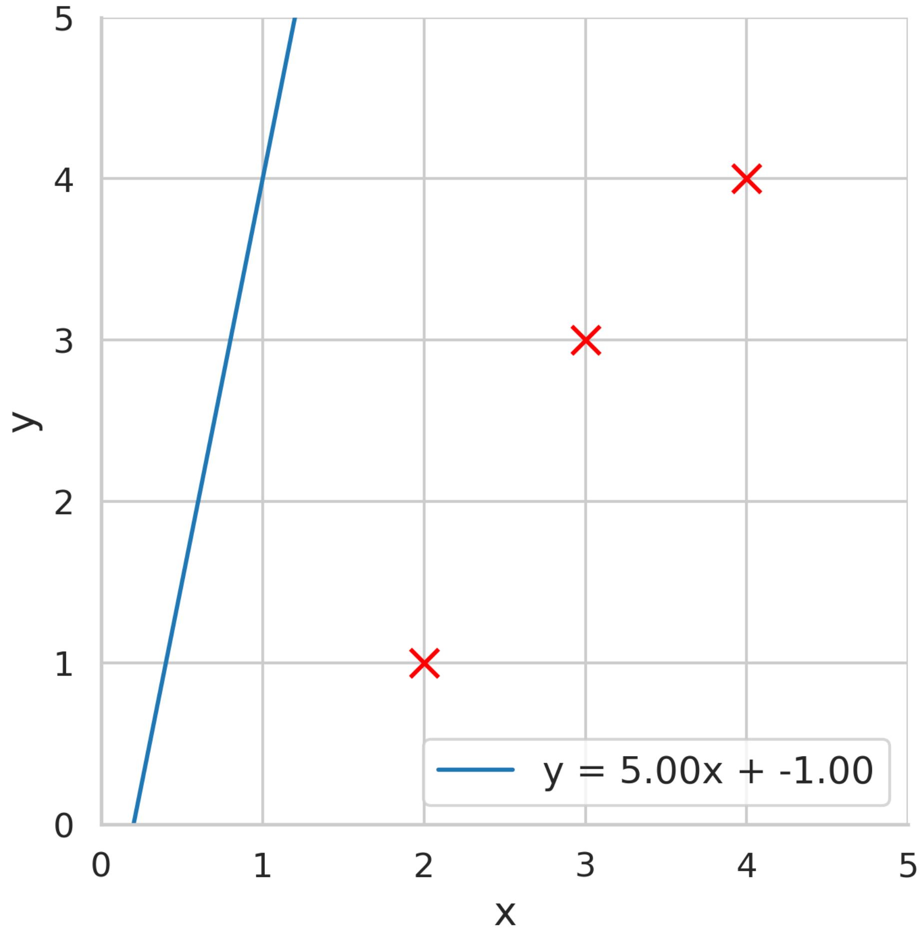


Gradient descent for linear regression

2. Calculate **derivative/gradient.**

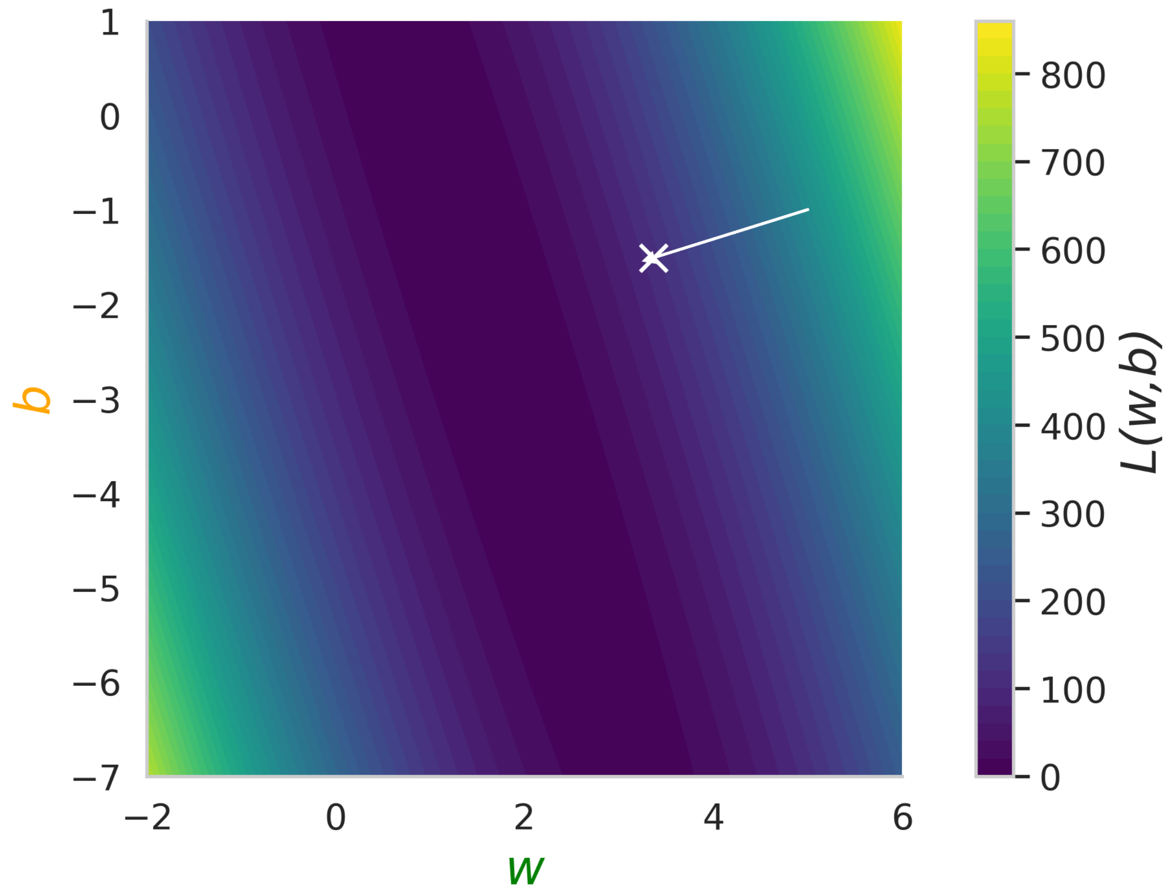


⟺

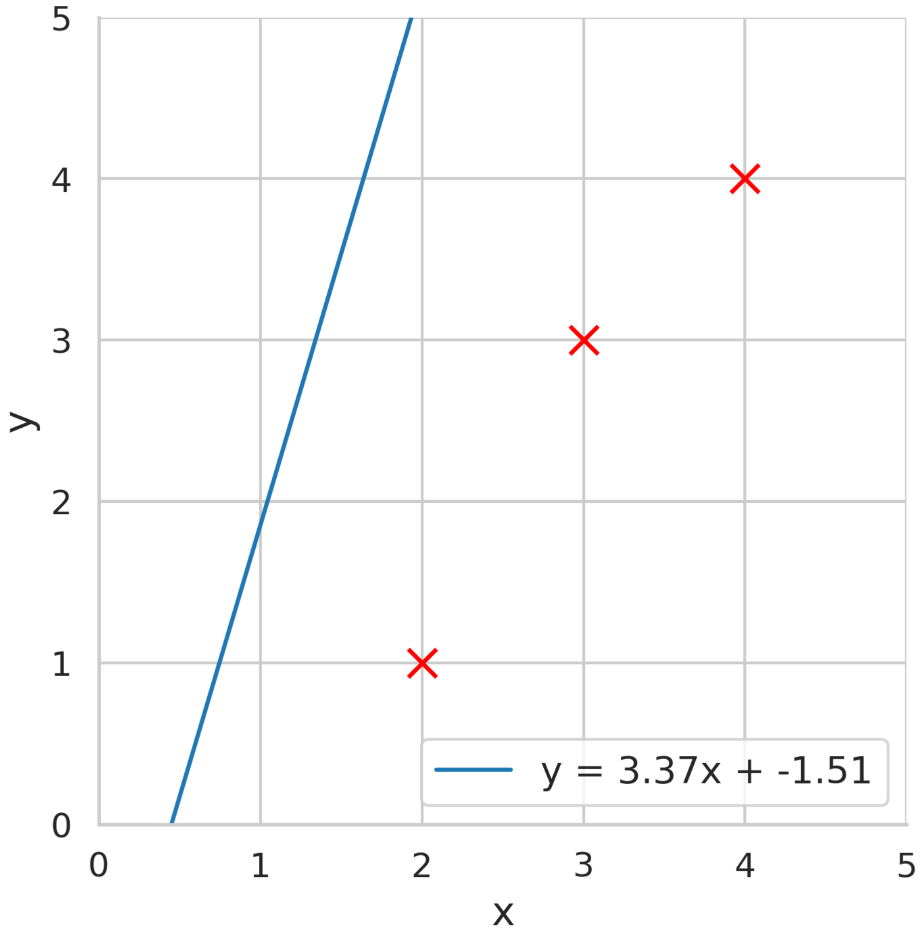


Gradient descent for linear regression

2. Calculate **derivative/gradient.**
3. Move w, b a little in **opposite** direction of gradient.

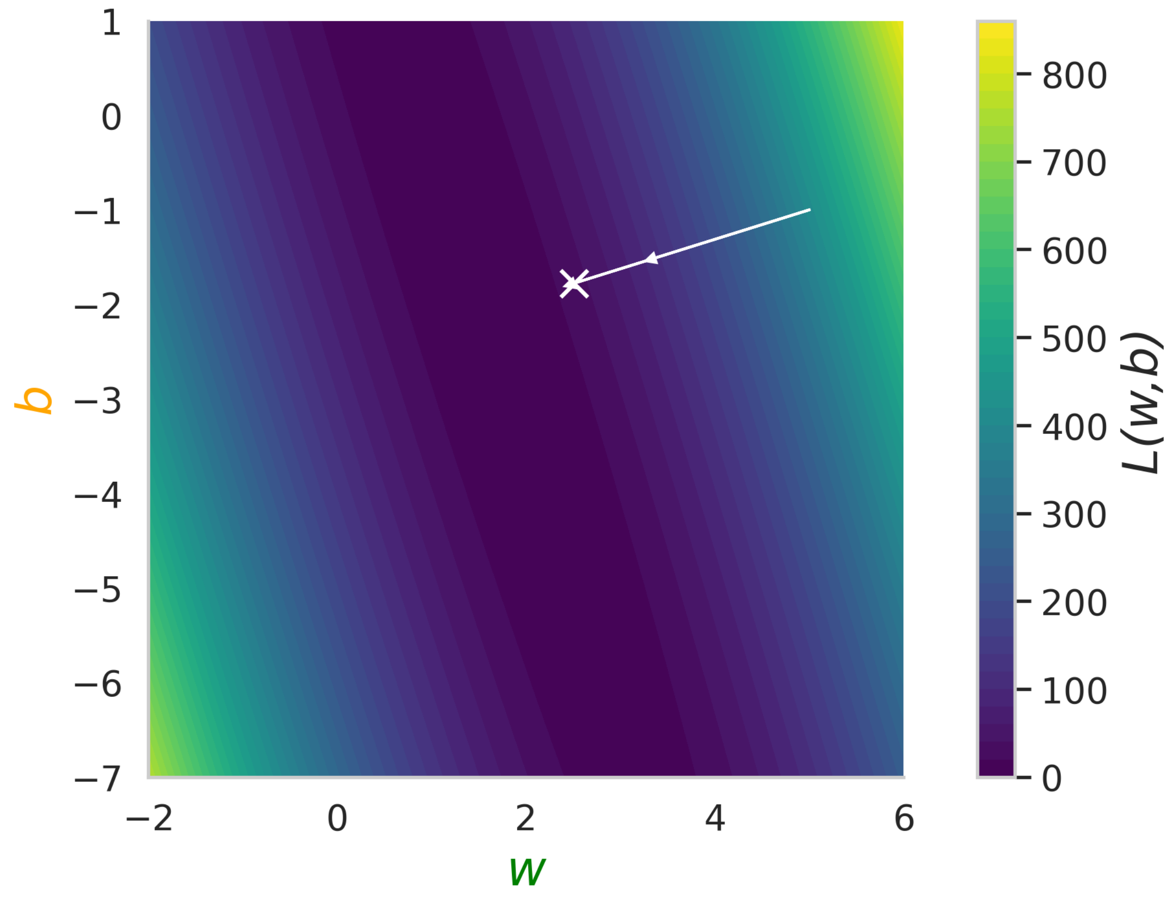


$\Longleftrightarrow$

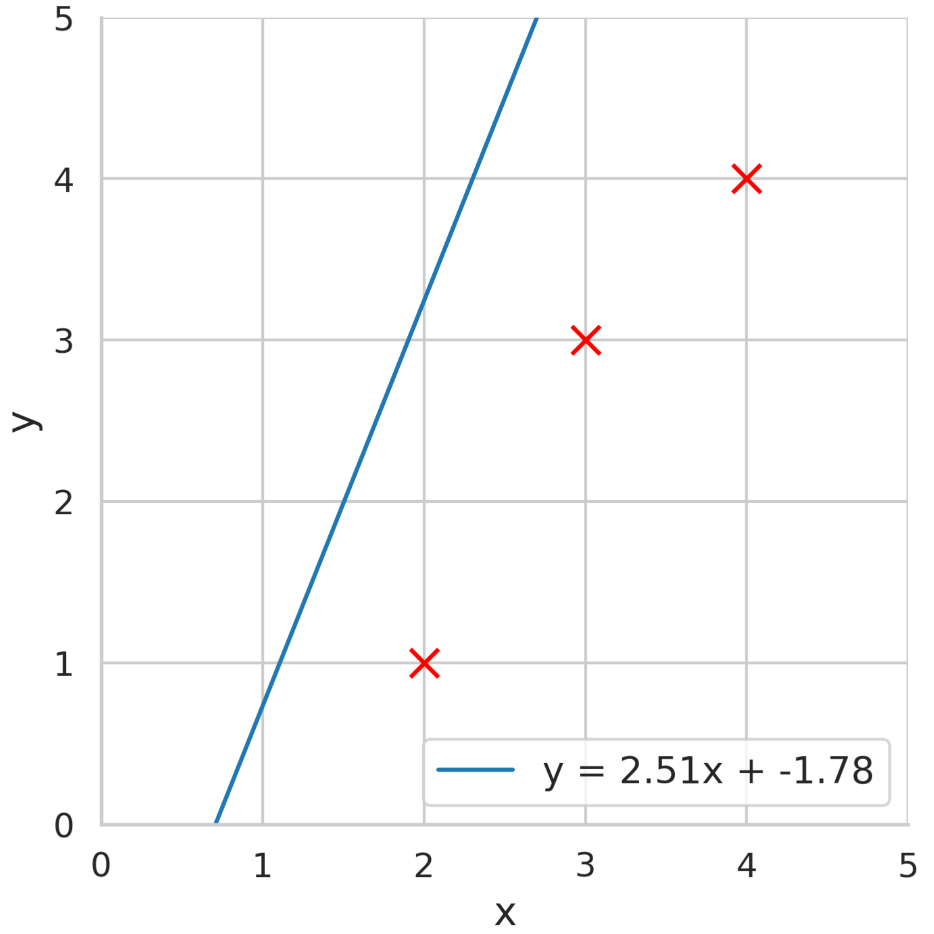


Gradient descent for linear regression

2. Calculate **derivative/gradient.**
3. Move w, b a little in **opposite** direction of gradient.

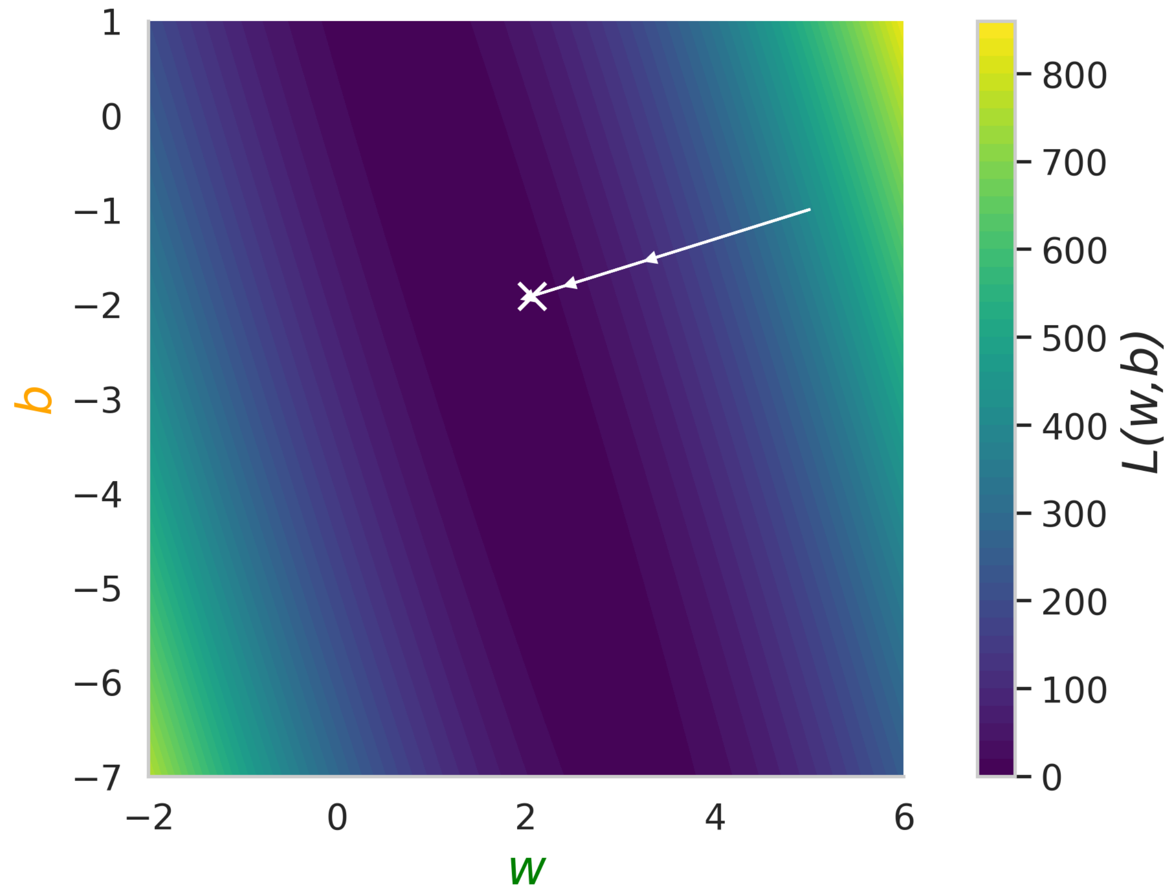


$\Longleftrightarrow$

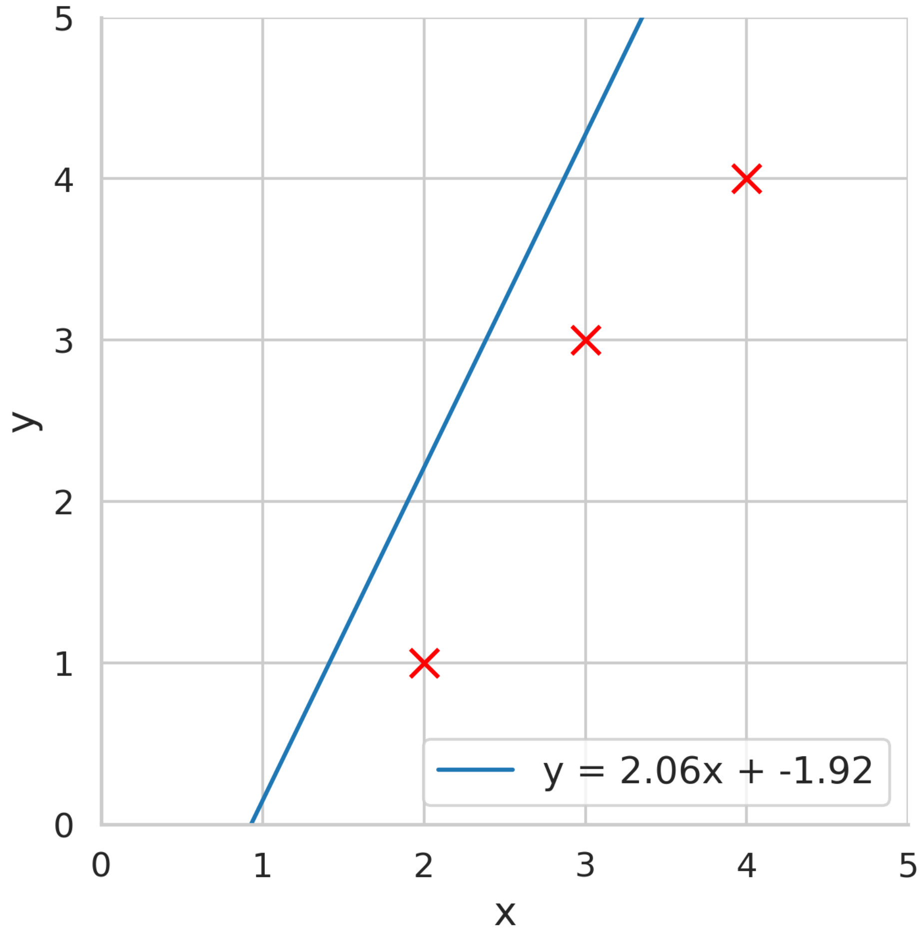


Gradient descent for linear regression

2. Calculate **derivative/gradient.**
3. Move w, b a little in **opposite** direction of gradient.

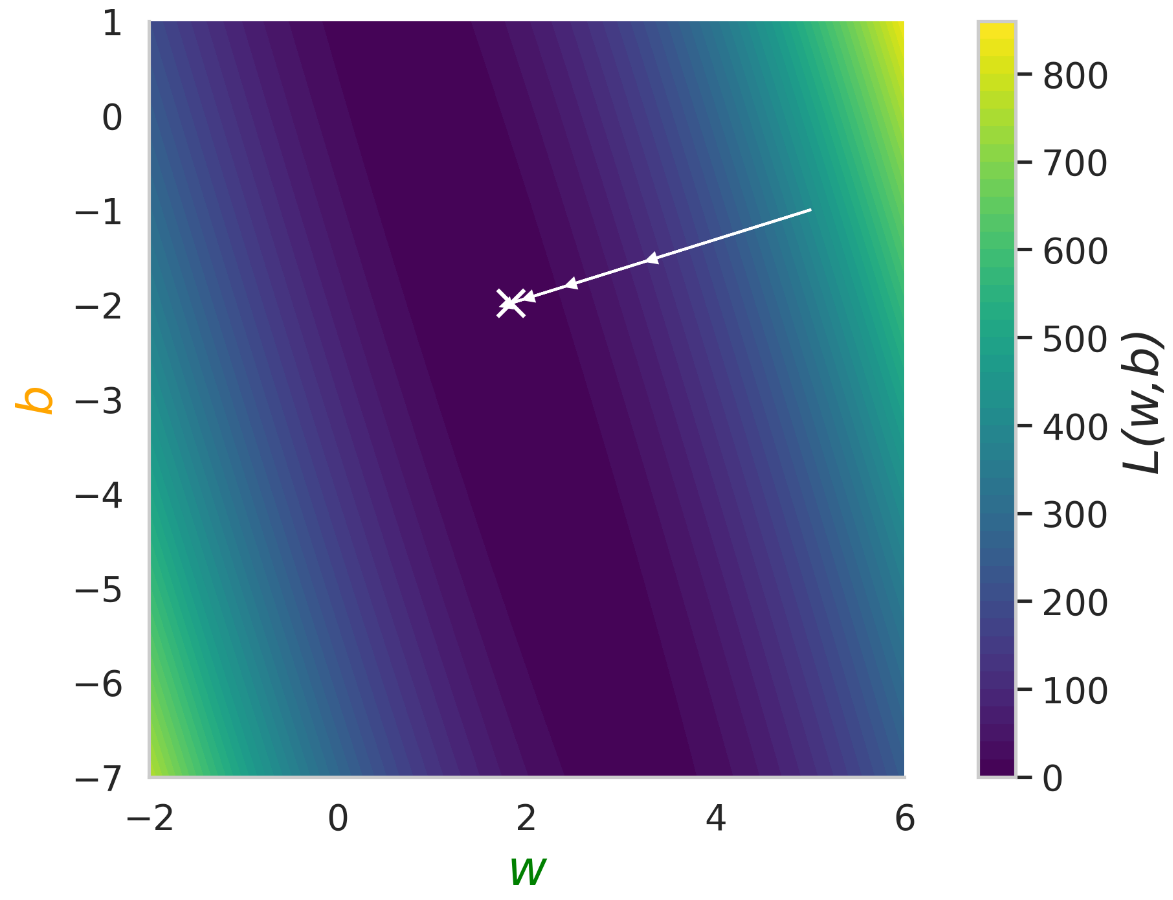


⟺

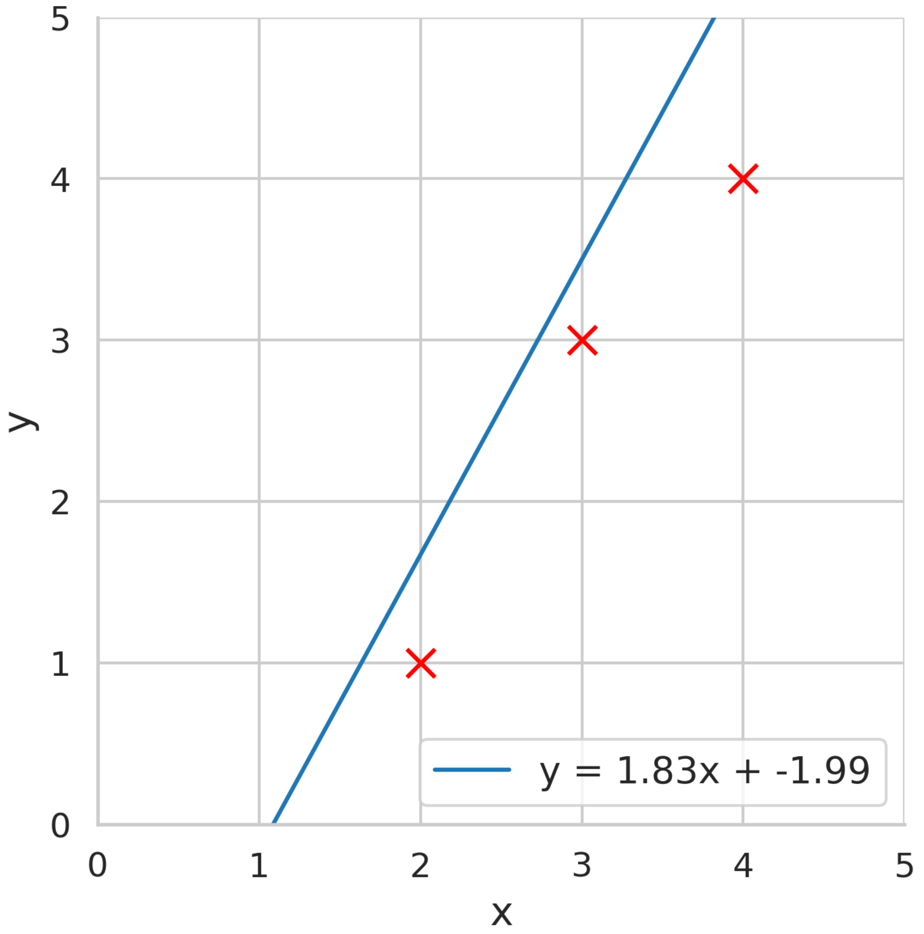


Gradient descent for linear regression

2. Calculate **derivative/gradient.**
3. Move w, b a little in **opposite** direction of gradient.

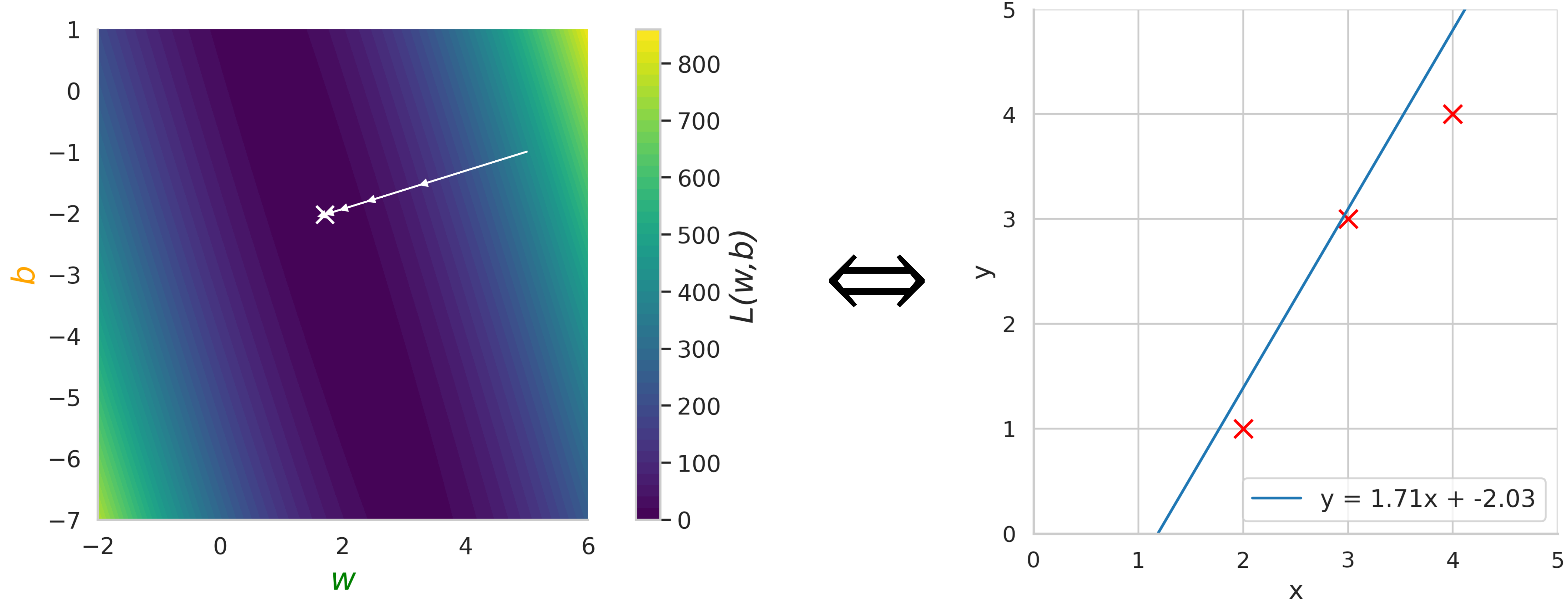


$\Longleftrightarrow$



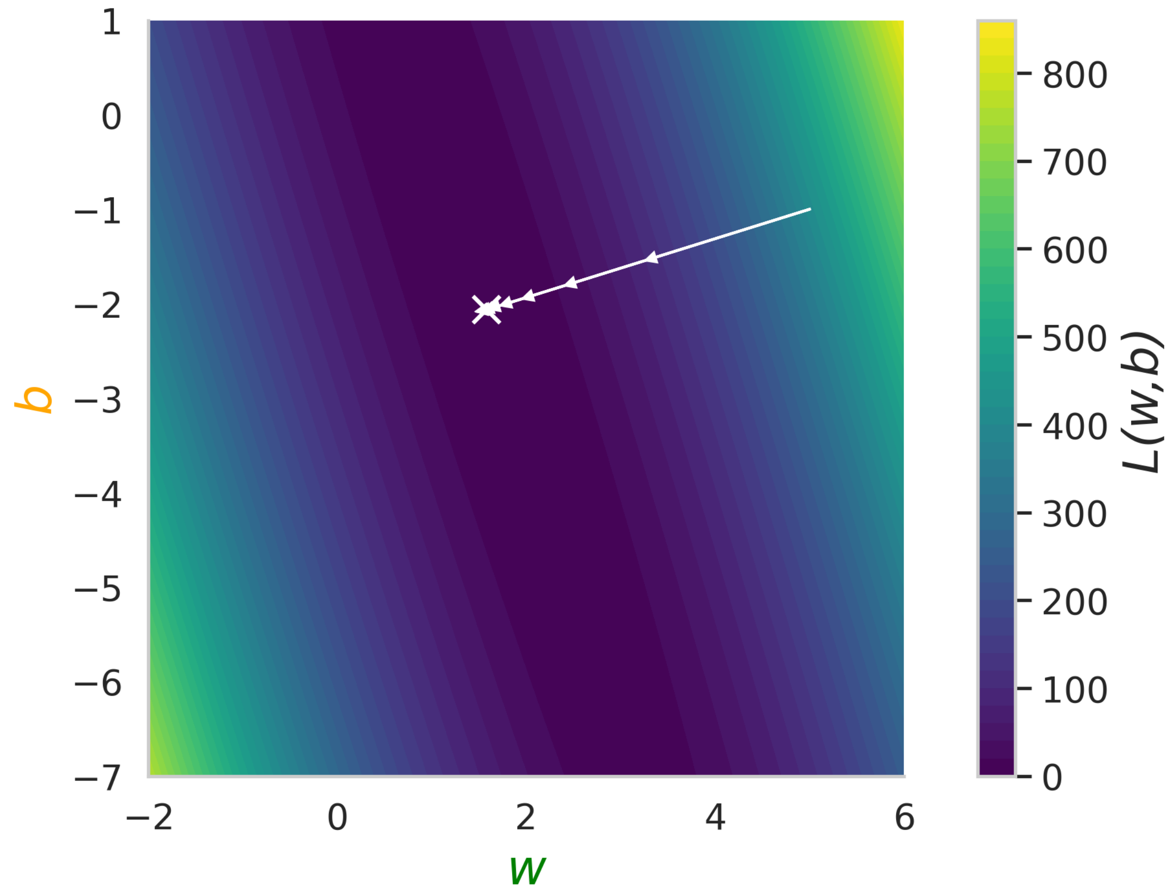
Gradient descent for linear regression

2. Calculate **derivative/gradient.**
3. Move w, b a little in **opposite** direction of gradient.

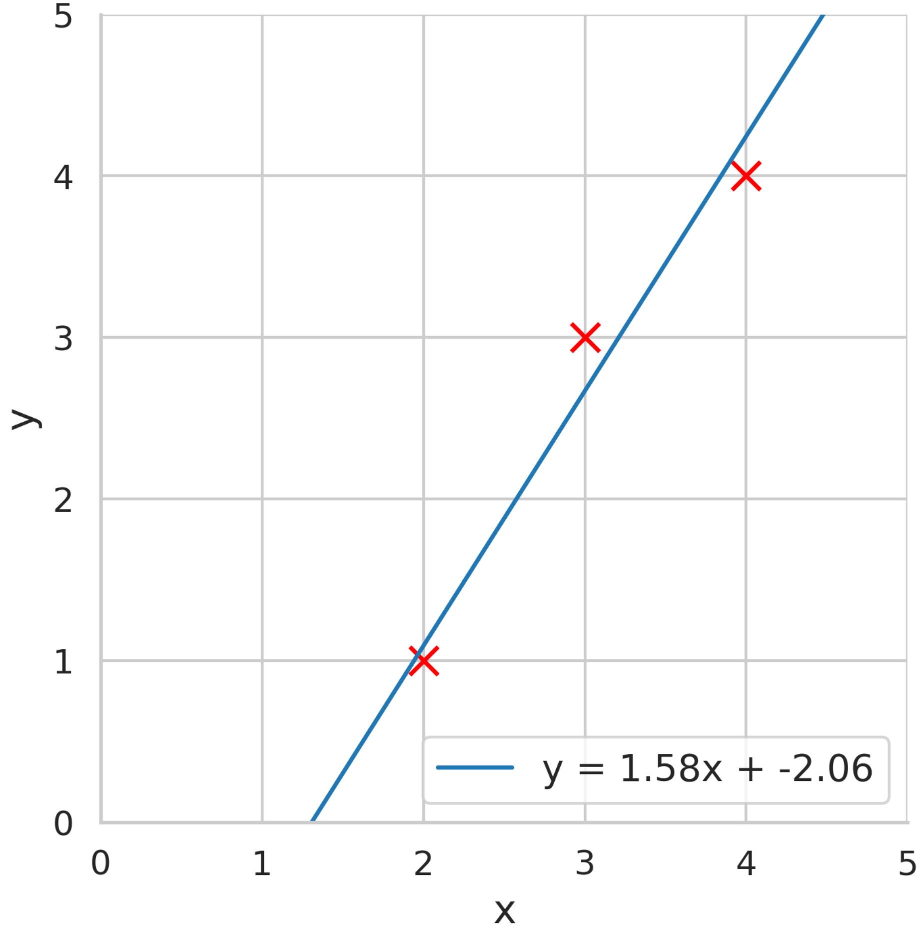


Gradient descent for linear regression

2. Calculate **derivative/gradient.**
3. Move w, b a little in **opposite** direction of gradient.

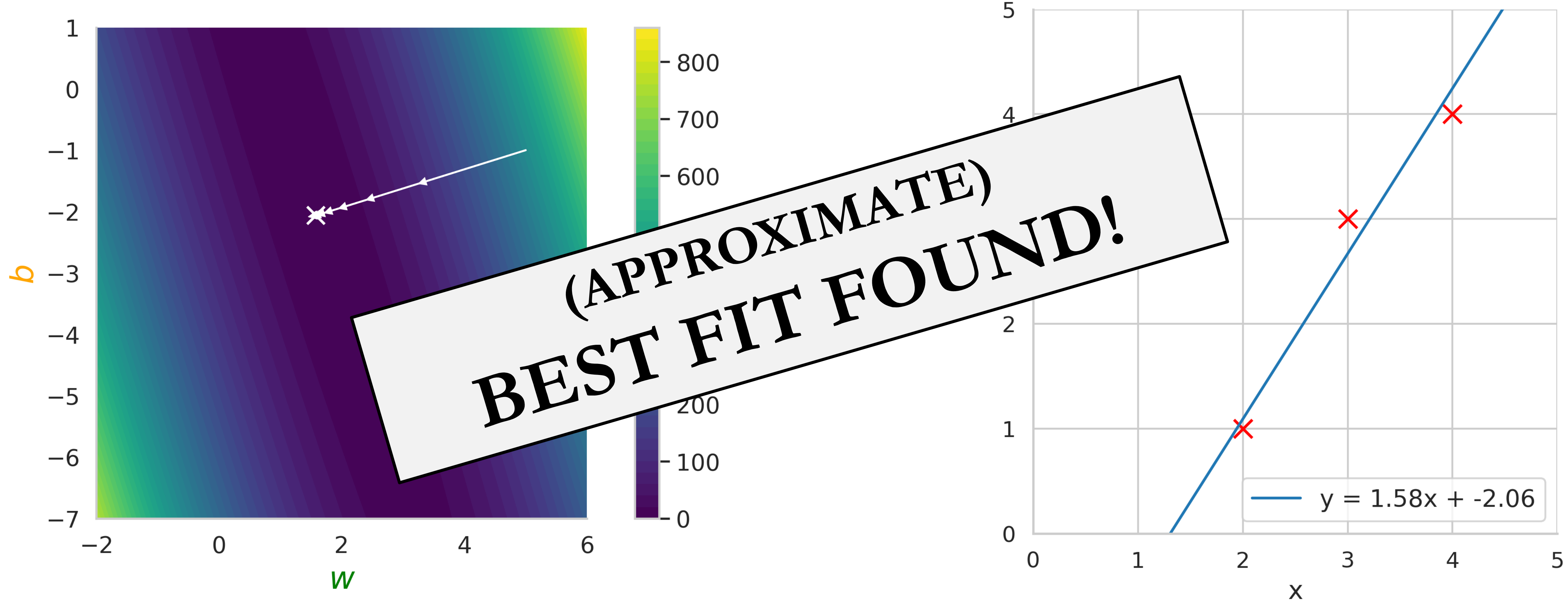


$\Longleftrightarrow$



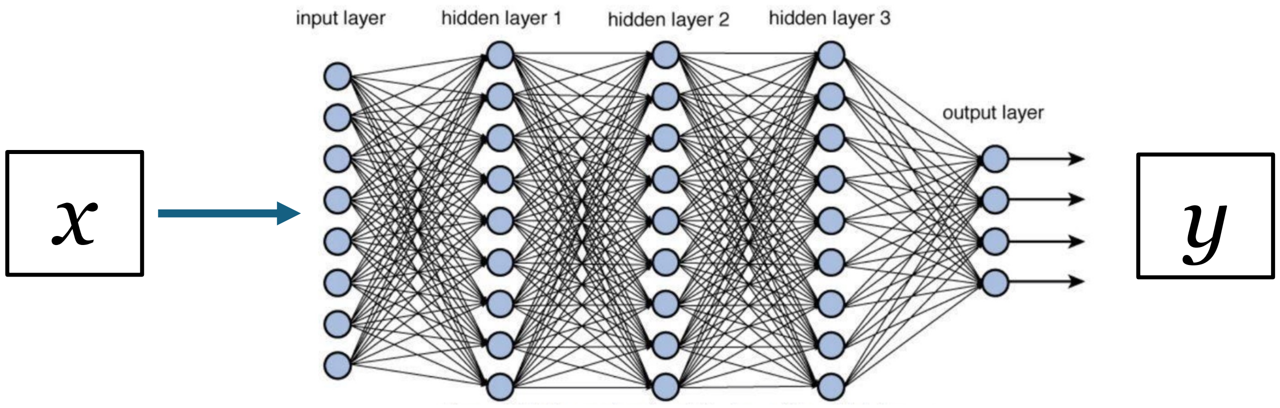
Gradient descent for linear regression

2. Calculate **derivative/gradient.**
3. Move w, b a little in **opposite** direction of gradient.



Recipe for deep learning

- Deep learning more sophisticated.
- But same ingredients!
  1. **Dataset**: x, y pairs (up to *trillions*) — but x, y could be anything.
  2. **Function class**: *neural networks*, not straight lines.
  3. **Parameters:** *billions*, not two.
  4. **Loss function:** not squared error, but same idea.
  5. **Gradient descent:** find best parameters.



Part 2:

How do Foundation Models Work?

Foundation models from 40,000 feet

- Still neural networks!
- Large Language Models (LLMs).
- To create a foundation model:
    1. Represent words as numbers.
    2. *Generate* sentences by *predicting* next word.
    3. Train on data from Internet.

Foundation models from 40,000 feet

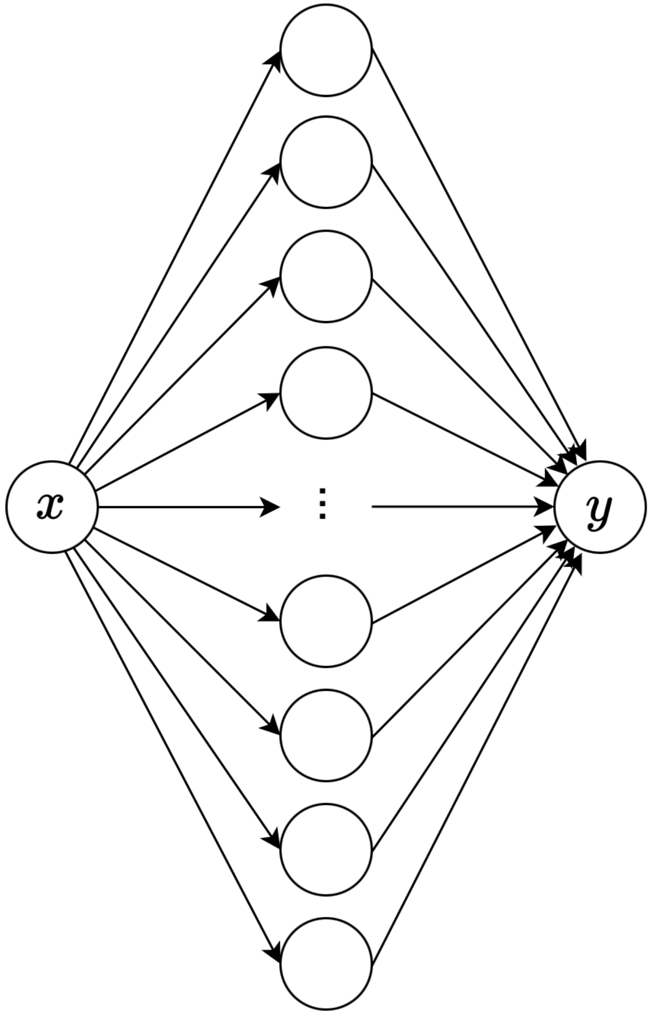
- Still neural networks!
- Large Language Models (LLMs).
- To create a foundation model:
    1. **Represent words as numbers.**
    2. *Generate* sentences by *predicting* next word.
    3. Train on data from Internet.

Words to numbers

- Neural networks process *numbers:*

$$f(x) = y$$

- x and y numbers.



Words to numbers

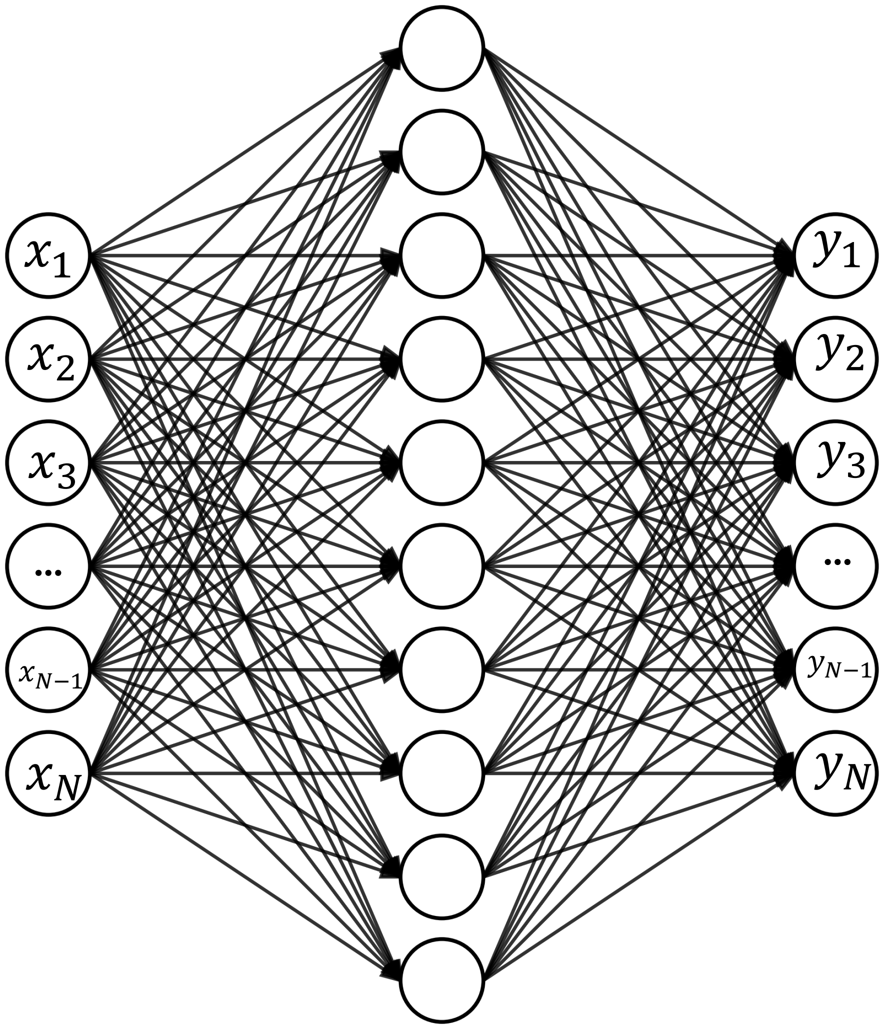
- Neural networks process *numbers:*

$$f(x) = y$$

- x and y numbers.
- Multiple numbers possible too.

$$f(x_1, \ldots, x_N) = y_1, \ldots, y_N$$

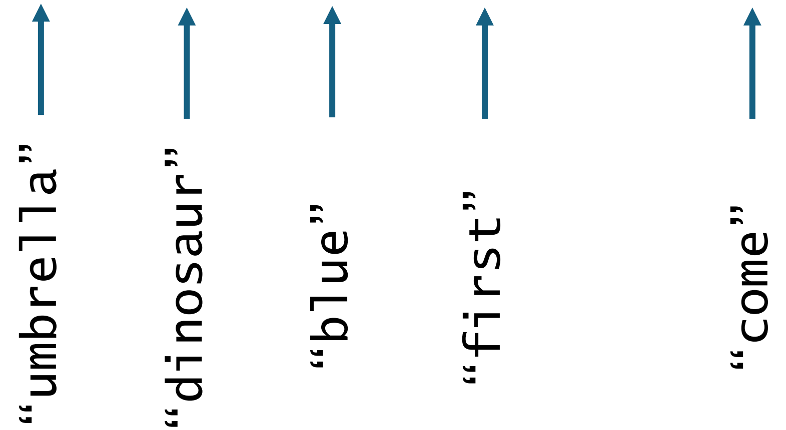
- How to input and output **words**?



Words to numbers

- Assume dictionary has N different words:
  - Represent each word as N numbers.
  - Each “slot” assigned to each word.

$$(x_1, x_2, x_3, x_4, \dots, x_N)$$



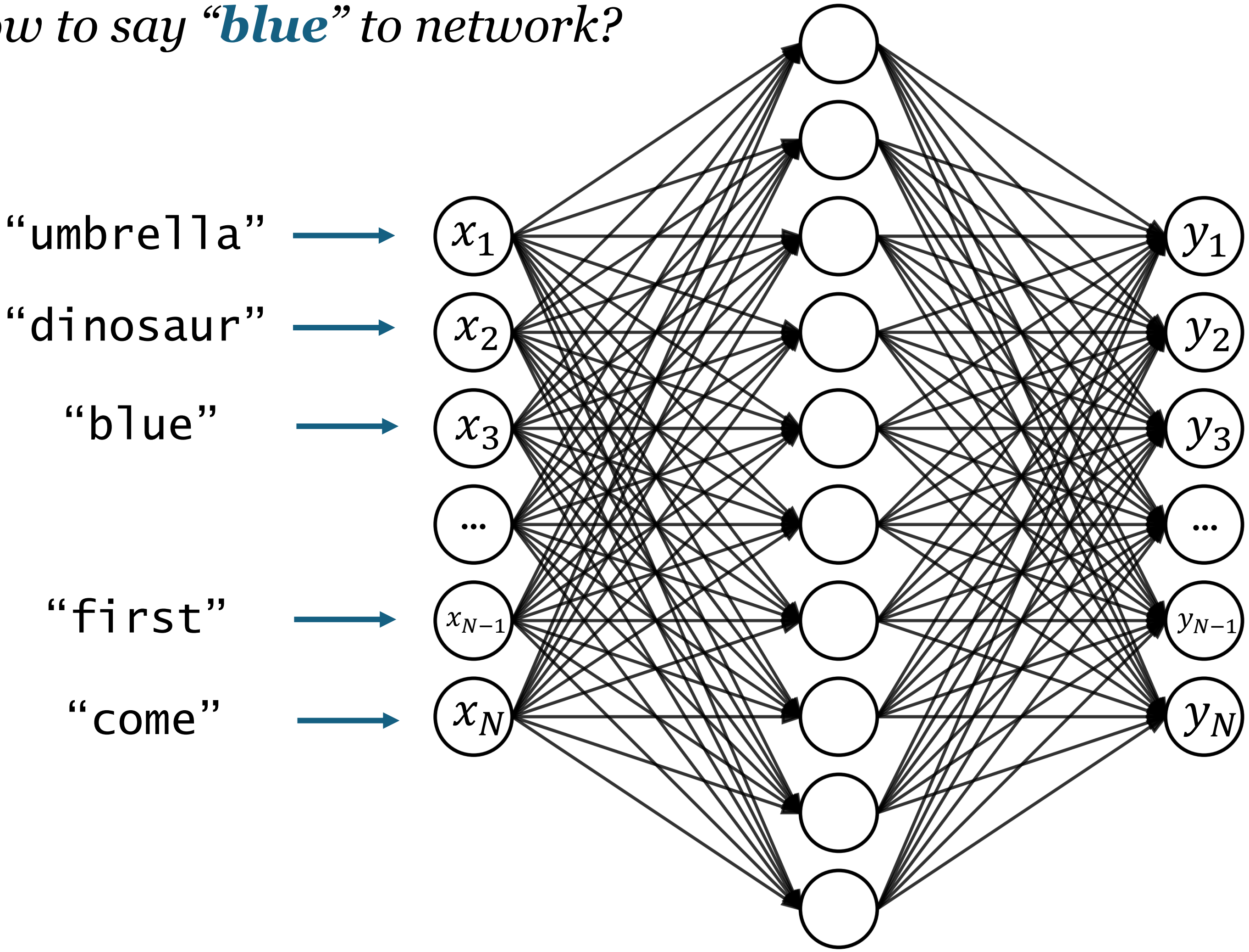
  - Each word represented by 1 in corresponding slot, 0 elsewhere:

$$\text{“}\texttt{blue}\text{”} \rightarrow (0, 0, 1, 0, \dots, 0)$$

  - Called **one-hot encoding.**

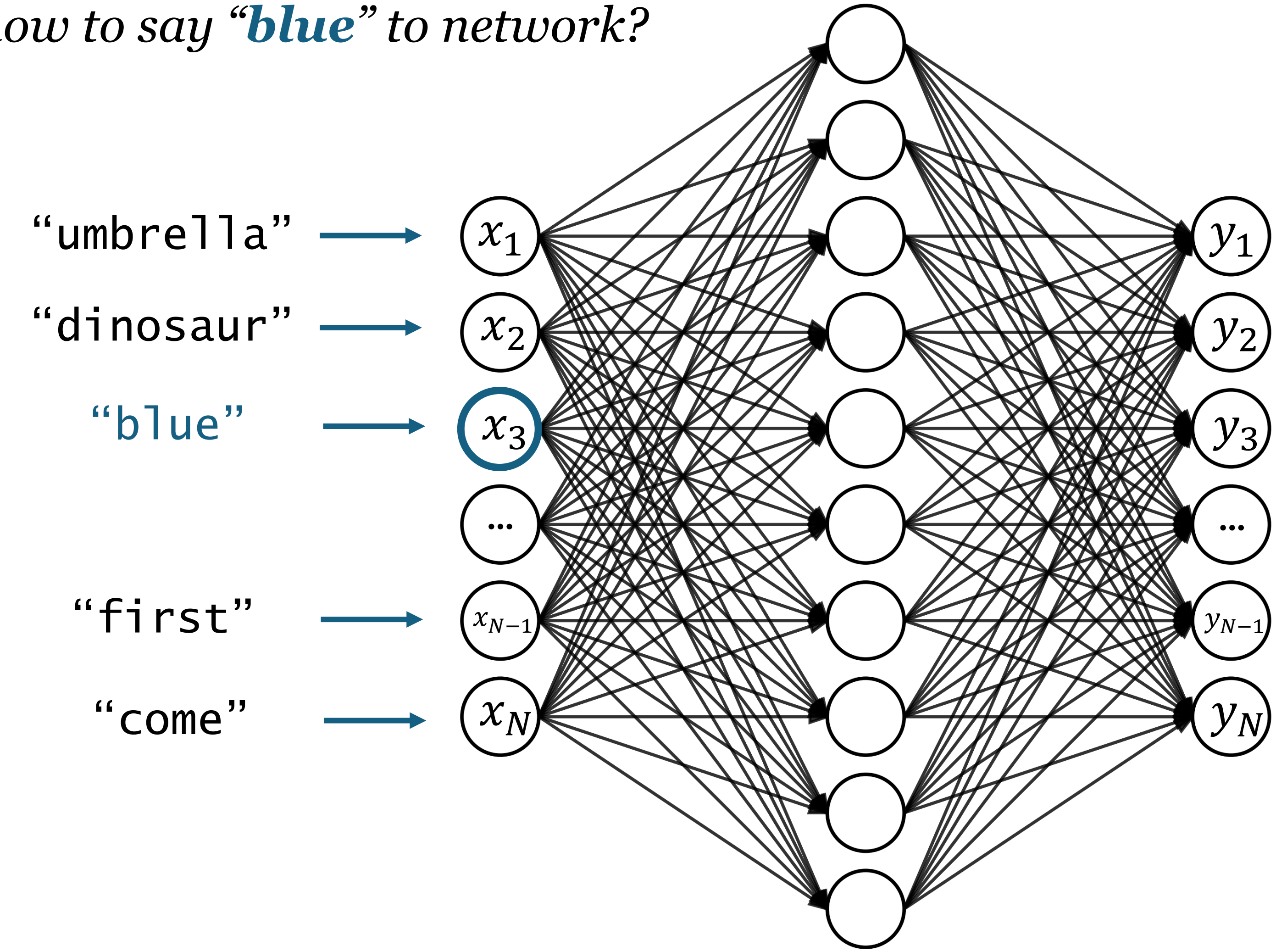
Words to numbers

*Example: how to say "**blue**" to network?*



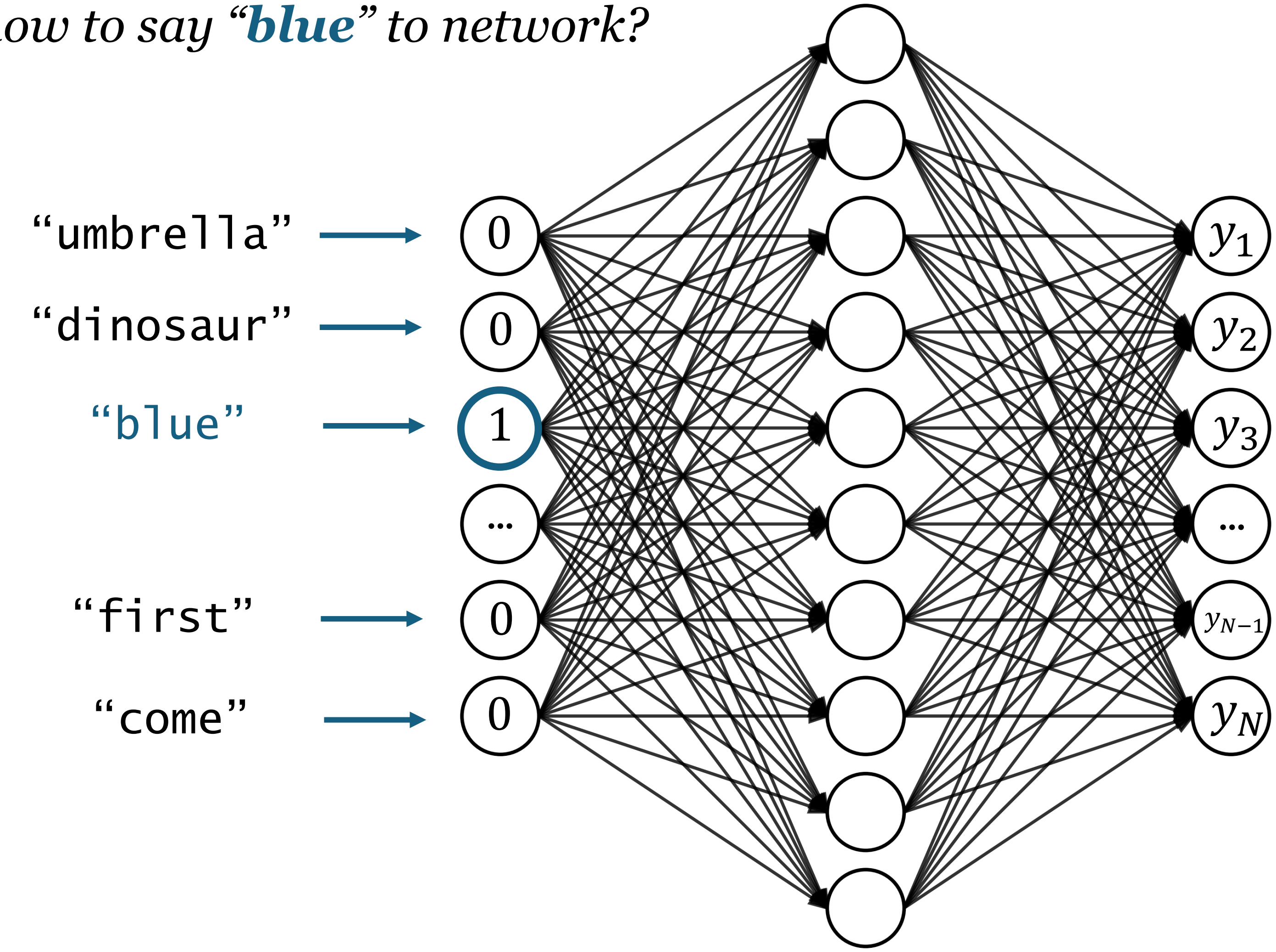
Words to numbers

*Example: how to say "**blue**" to network?*



Words to numbers

*Example: how to say "**blue**" to network?*



Foundation models from 40,000 feet

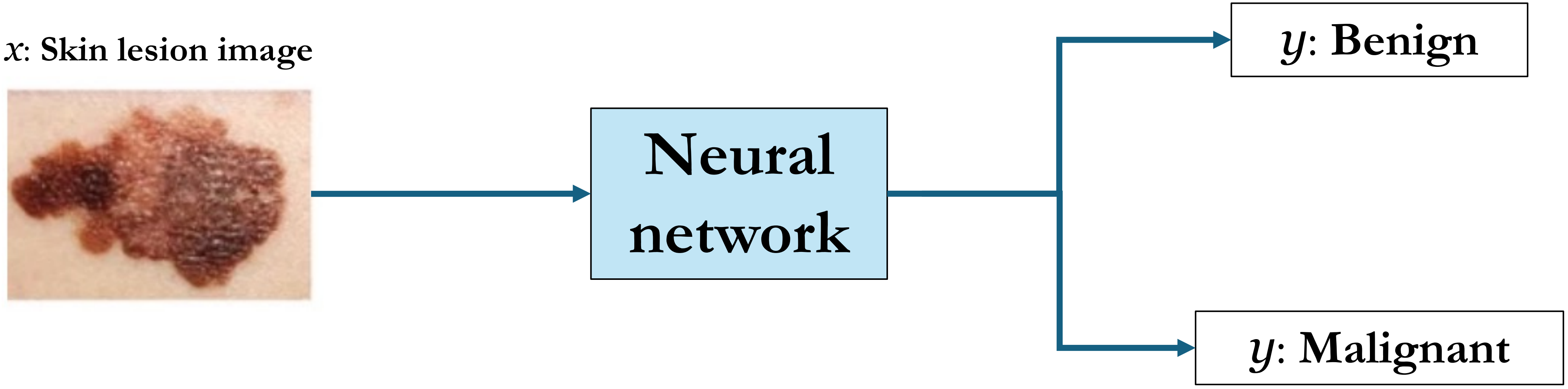
- Still neural networks!
- Often called “Large Language Models” (LLMs).
- To create a foundation model:
    1. **Represent words as numbers.**
    2. *Generate* sentences by *predicting* next word.
    3. Train on data from Internet.

Foundation models from 40,000 feet

- Still neural networks!
- Often called “Large Language Models” (LLMs).
- To create a foundation model:
    1. Represent words as numbers.
    2. ***Generate* sentences by *predicting* next word.**
    3. Train on data from Internet.

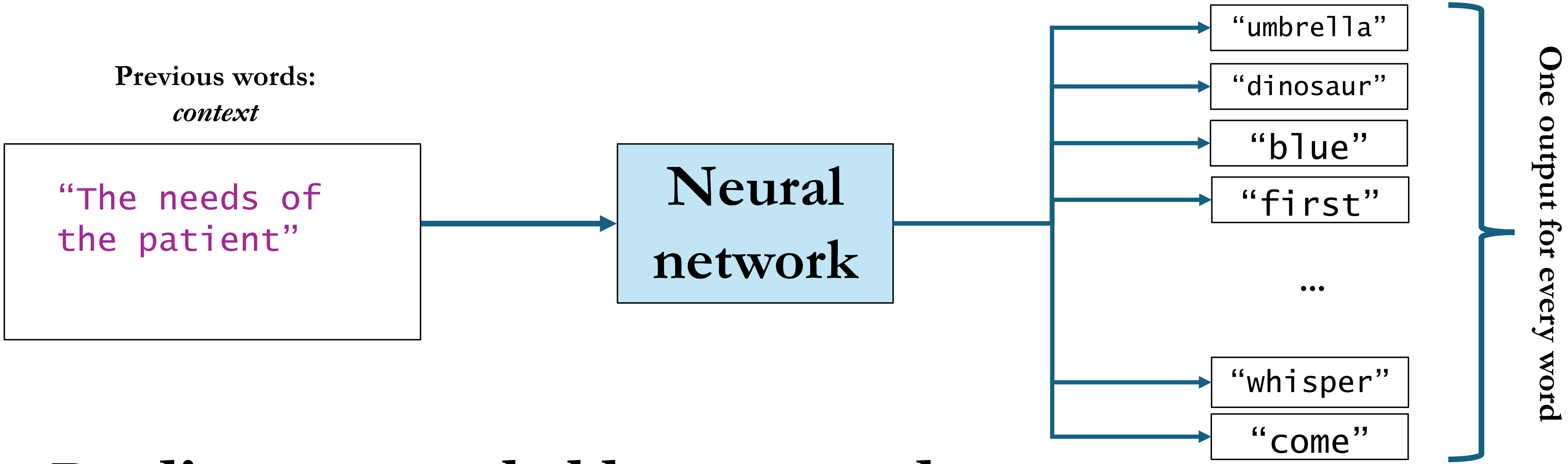
Generative AI is predictive AI

- How to generate entire sentences?
- Idea: ***sentence generation is just repeated word prediction.***
- Standard AI: **predict** malignancy given image.



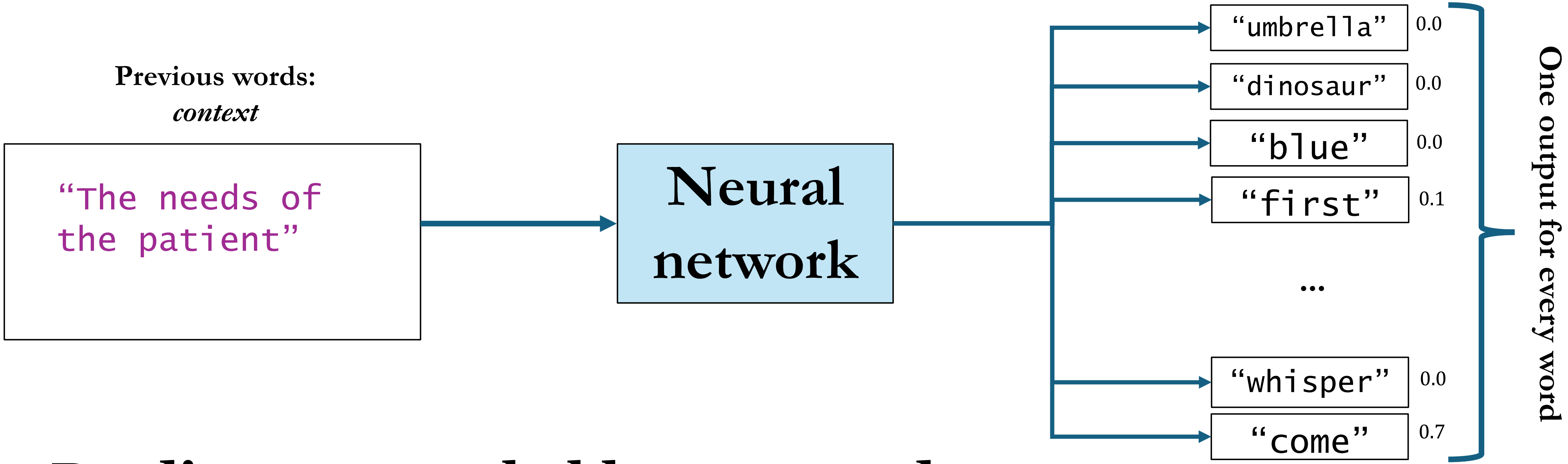
- Now: **predict** *next word* given *previous words*.

Generative AI is predictive AI



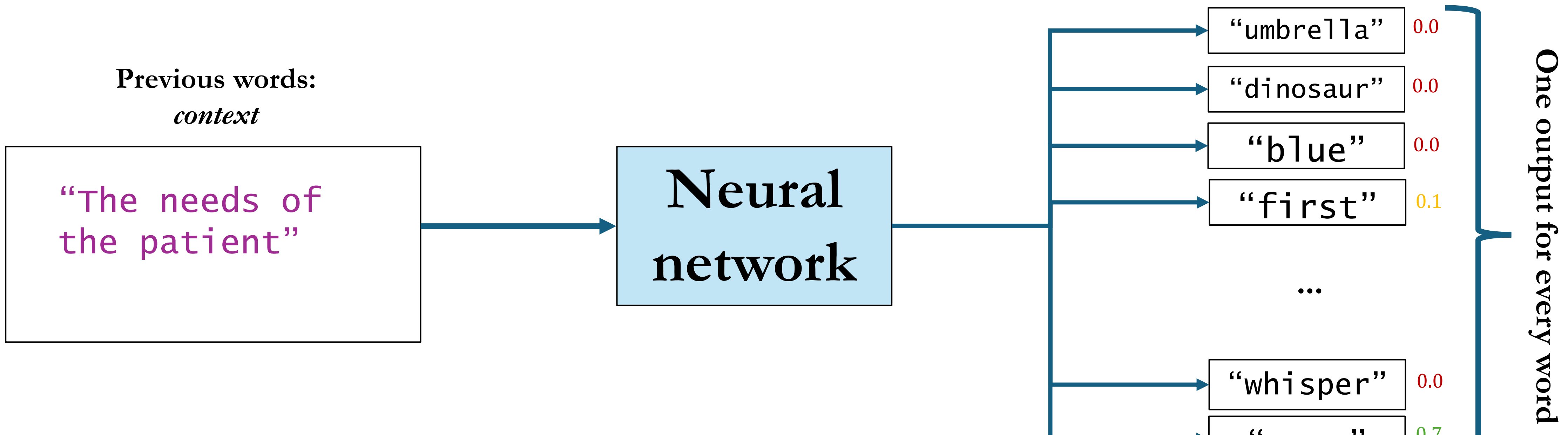
1. **Predict most probable next word.**

Generative AI is predictive AI



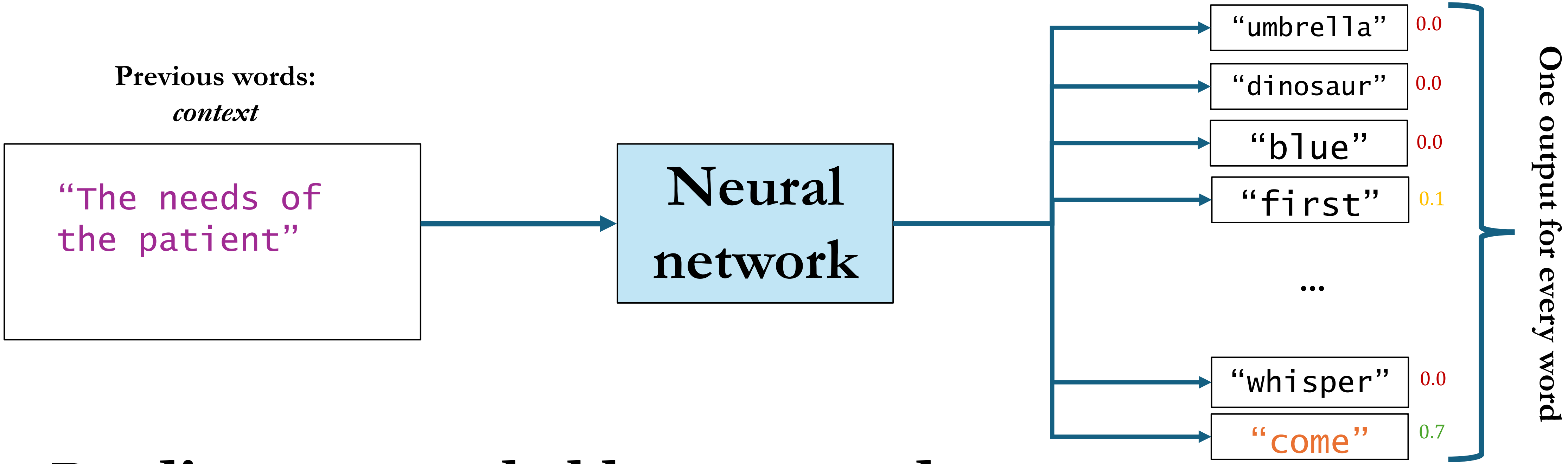
1. **Predict most probable next word.**

Generative AI is predictive AI



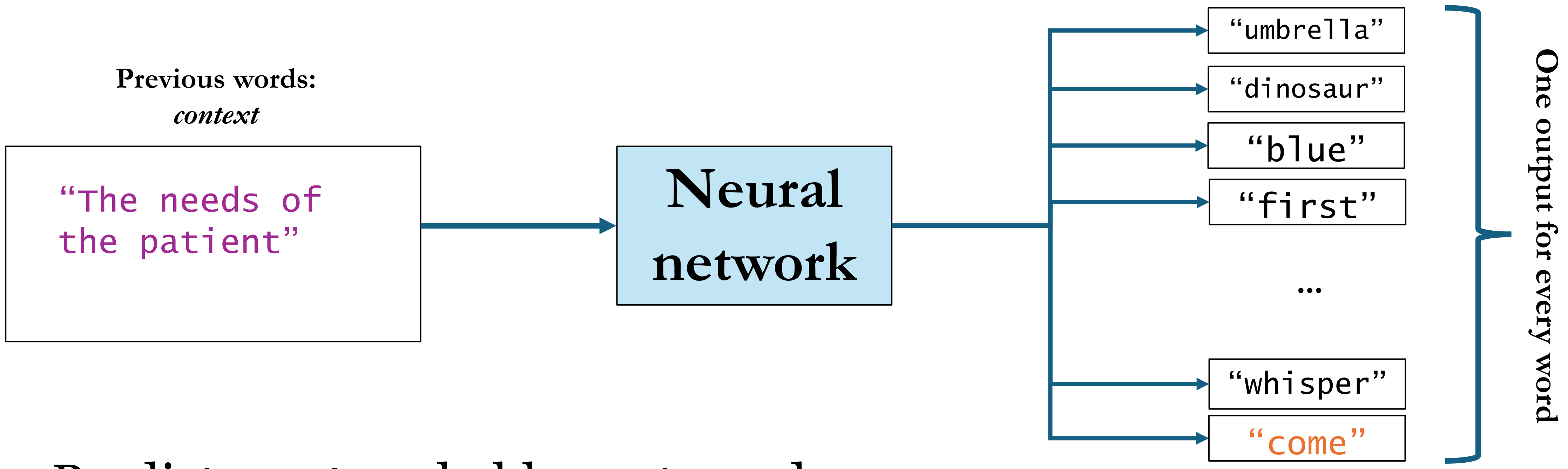
1. **Predict most probable next word.**

Generative AI is predictive AI



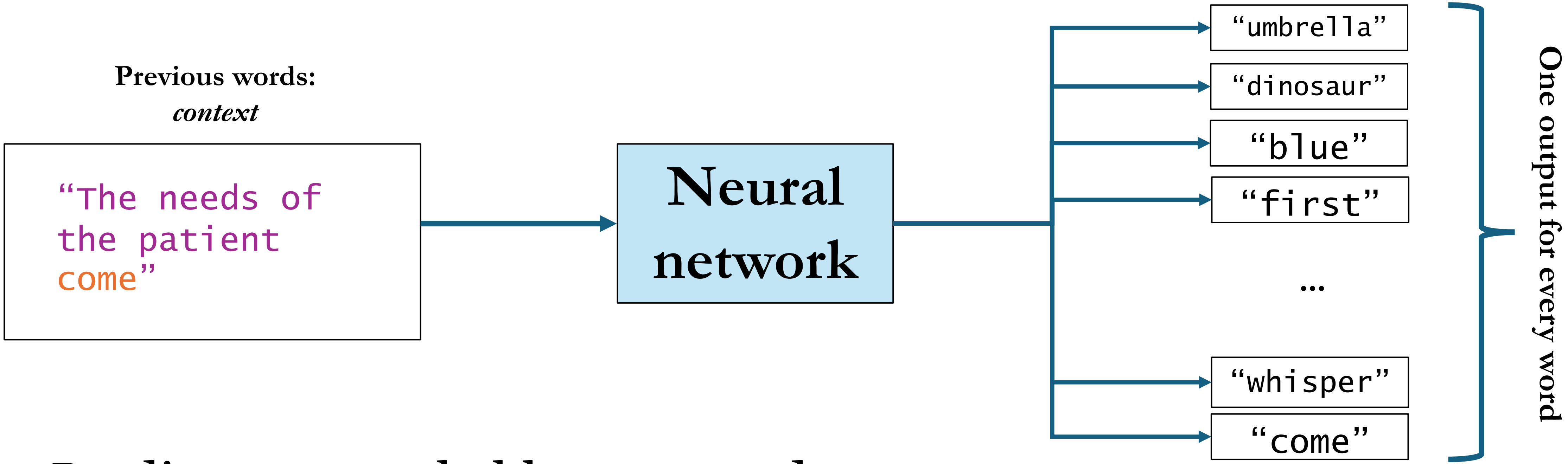
1. **Predict most probable next word.**

Generative AI is predictive AI



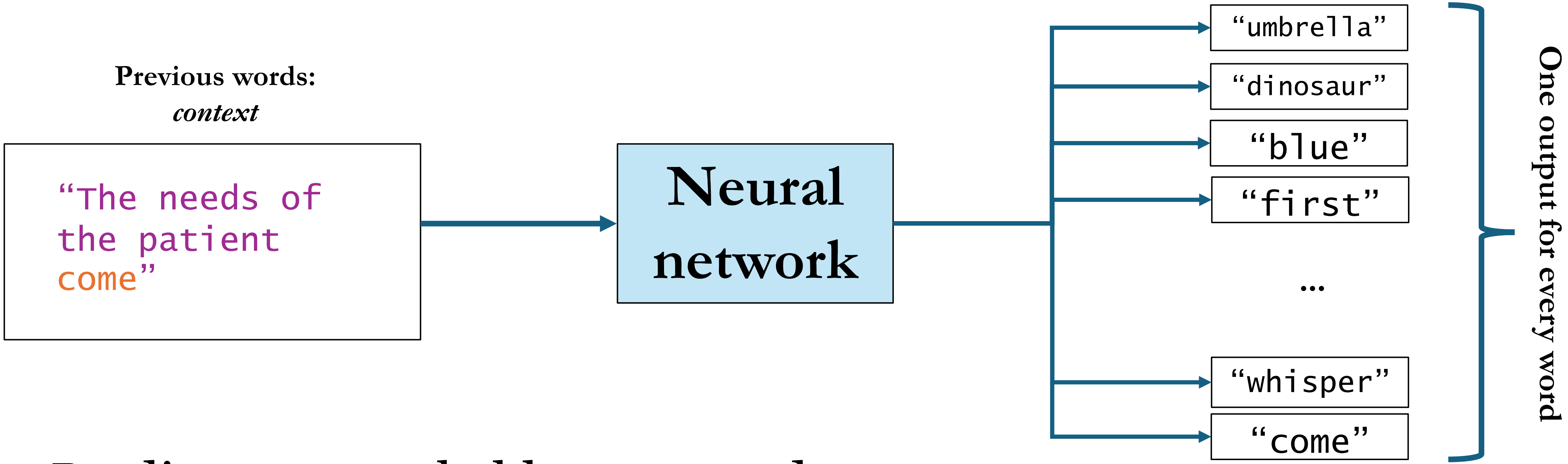
1. Predict most probable next word.
2. **Add it onto the context.**

Generative AI is predictive AI



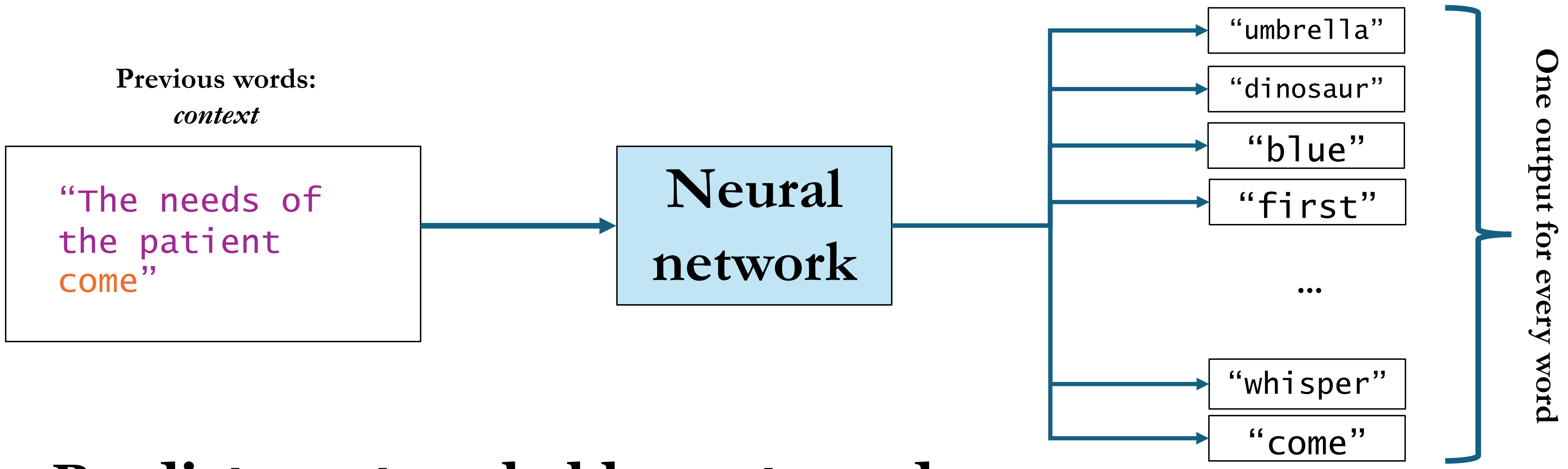
1. Predict most probable next word.
2. **Add it onto the context.**

Generative AI is predictive AI



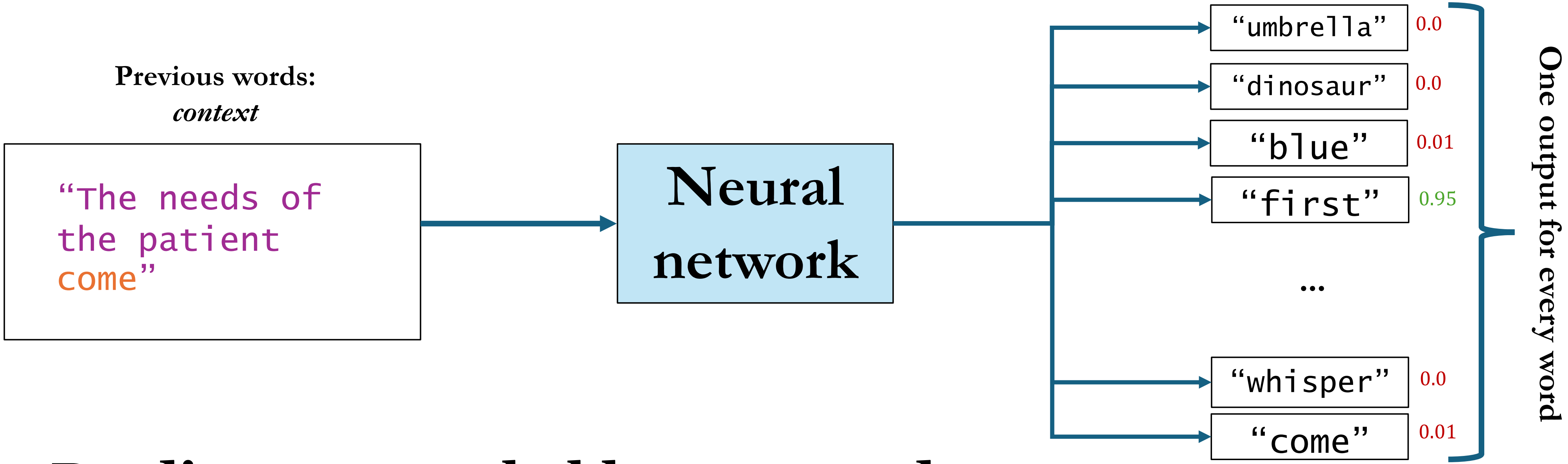
1. Predict most probable next word.
2. Add it onto the context.
3. **Go back to step 1.**

Generative AI is predictive AI



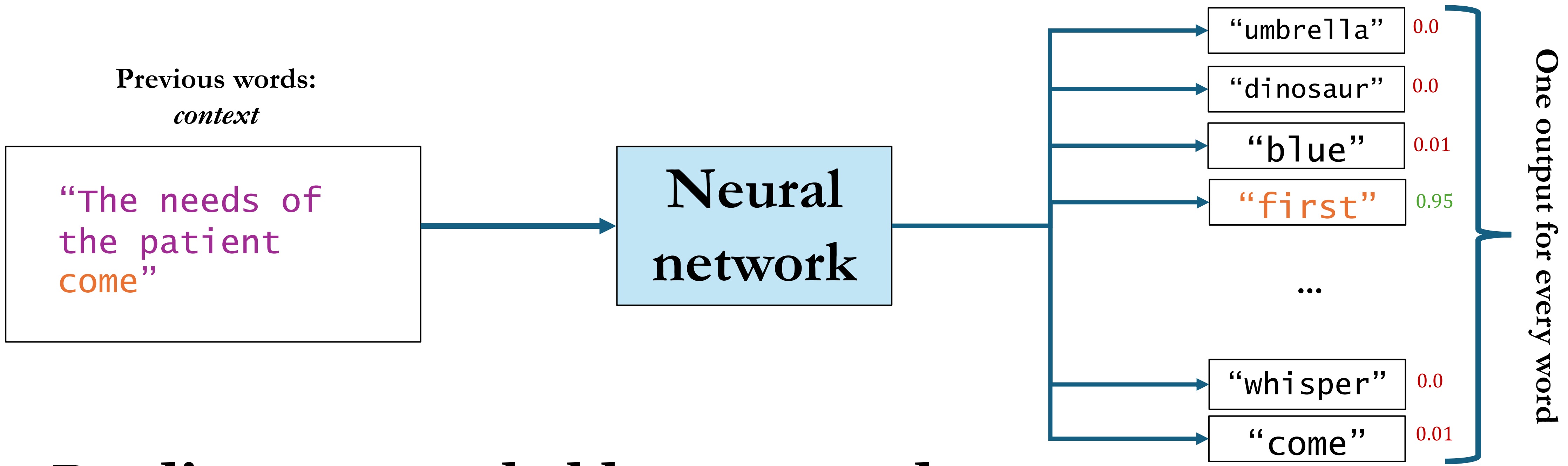
1. **Predict most probable next word.**
2. Add it onto the context.
3. Go back to step 1.

Generative AI is predictive AI



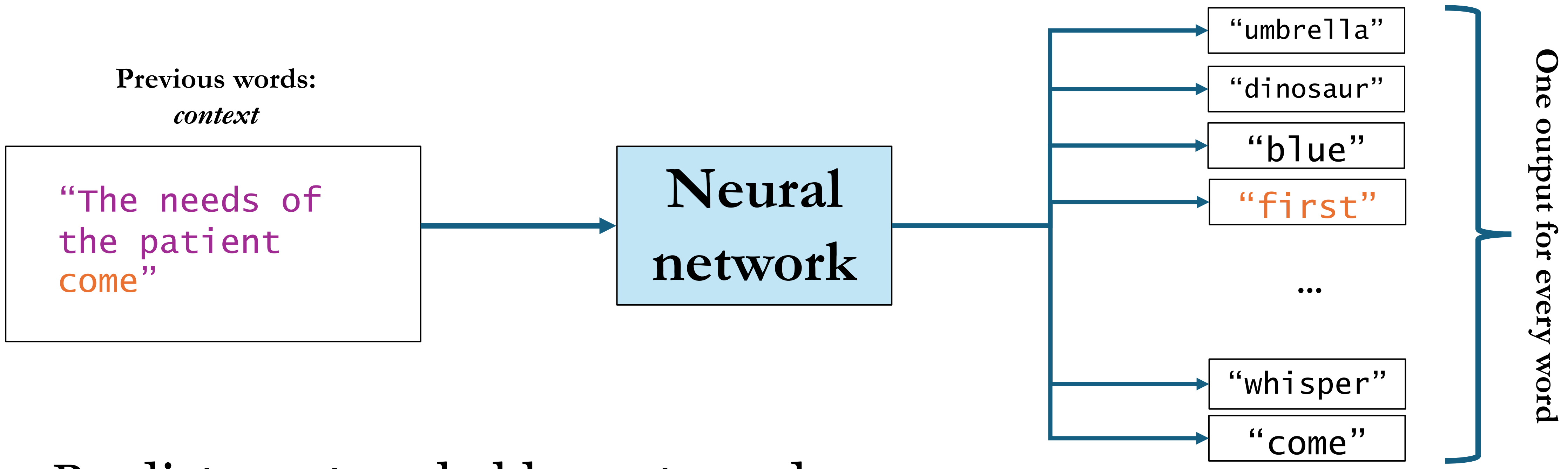
1. **Predict most probable next word.**
2. Add it onto the context.
3. Go back to step 1.

Generative AI is predictive AI



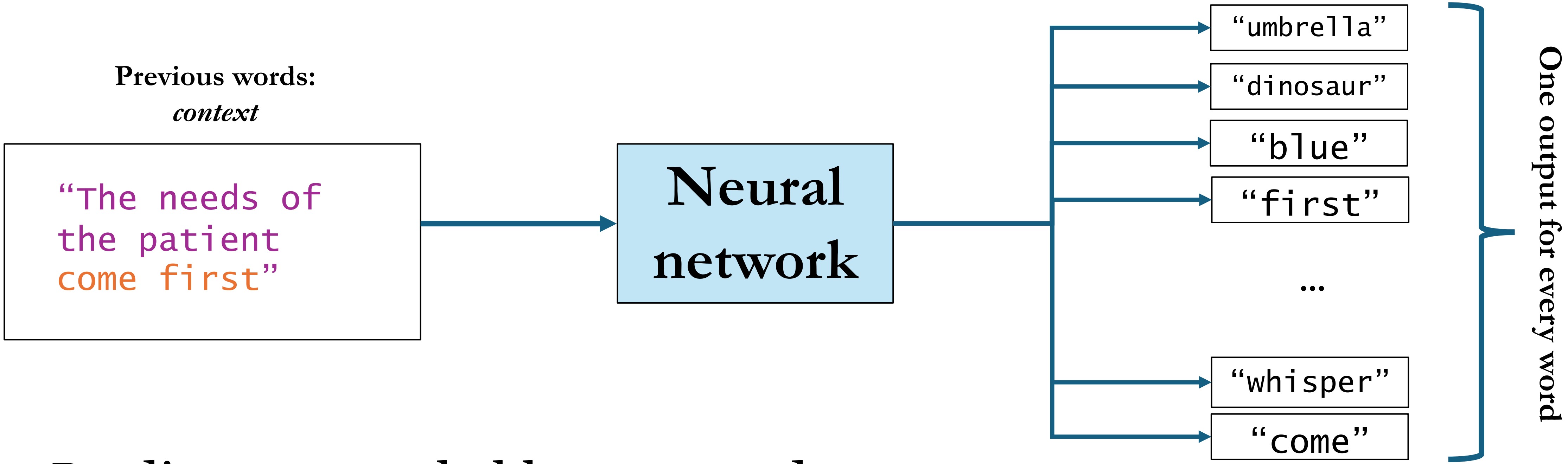
1. **Predict most probable next word.**
2. Add it onto the context.
3. Go back to step 1.

Generative AI is predictive AI



1. Predict most probable next word.
2. **Add it onto the context.**
3. Go back to step 1.

Generative AI is predictive AI



1. Predict most probable next word.
2. **Add it onto the context.**
3. Go back to step 1.

Generative AI is predictive AI

Previous words:
context

“The needs of the patient come first”

“come”

One output for every word

1. Predic
2. **Add it**
3. Go back

Repeat until entire document/conversation is generated

Foundation models from 40,000 feet

- Still neural networks!
- Often called “Large Language Models” (LLMs).
- To create a foundation model:
    1. Represent words as numbers.
    2. ***Generate* sentences by *predicting* next word.**
    3. Train on data from Internet.

Foundation models from 40,000 feet

- Still neural networks!
- Often called "Large Language Models" (LLMs).
- To create a foundation model:
    1. Represent words as numbers.
    2. *Generate* sentences by *predicting* next word.
    3. **Train on data from Internet.**

Training data

- Training requires *dataset* of many *datapoints*.
  - **Datapoint** = example of *correct* (Input, Output) pair.
  - **Dataset** = (Input 1, Output 1), (Input 2, Output 2), ..., etc.
- Last time:
  - **Dataset** = ( , Malignant), ( , Benign), ..., etc.
- Now: labels are next word in sentence.
  - **Dataset** = ("Mary had a little", "lamb"),
    ("To be or not to", "be"),
    ("Bold Forward" , "Unbound"), ..., etc.
- Where to get such a dataset?
  - **The Internet!**

Esteva et al. 2017, Dermatologist-level classification of skin cancer with deep neural networks, *Nature*.

Training data

- Computers automatically download text in billions of webpages.
- *Common Crawl* releases monthly web “snapshots”.
- Includes:
    - Wikipedia
    - Forum discussions
    - Online courseware
    - Academic papers
    - Books
    - Open-source code
- ***All human knowledge!***

Toy example: mini-internet

- Imagine entire internet had only *two* websites.
- Each website has just *one* sentence:
  - www.mayo.com says: "Mayo Clinic saves lives"
  - www.ai.com says: "Mayo Clinic uses AI"
- **What would happen if we trained LLM on this data?**

What is the training data?

- **Dataset** = (Input 1, Output 1), (Input 2, Output 2), ..., etc.
- *Six* training datapoints in two sentences!
  - “`Mayo Clinic saves lives`”
  - “`Mayo Clinic uses AI`”

What is the training data?

- **Dataset** = (Input 1, Output 1), (Input 2, Output 2), ..., etc.
- *Six* training datapoints in two sentences!
  - "`Mayo Clinic saves lives`"
  - "`Mayo Clinic uses AI`"
- Becomes:
  1. ("`Mayo`","`Clinic`")

What is the training data?

- **Dataset** = (Input 1, Output 1), (Input 2, Output 2), ..., etc.
- *Six* training datapoints in two sentences!
  - "`Mayo Clinic saves lives`"
  - "`Mayo Clinic uses AI`"
- Becomes:
  1. ("`Mayo`","`Clinic`")
  2. ("`Mayo Clinic`","`saves`")

What is the training data?

- **Dataset** = (Input 1, Output 1), (Input 2, Output 2), …, etc.
- *Six* training datapoints in two sentences!
  - "`Mayo Clinic saves lives`"
  - "`Mayo Clinic uses AI`"
- Becomes:
  1. ("`Mayo`","`Clinic`")
  2. ("`Mayo Clinic`","`saves`")
  3. ("`Mayo Clinic saves`","`lives`")

What is the training data?

- **Dataset** = (Input 1, Output 1), (Input 2, Output 2), …, etc.
- *Six* training datapoints in two sentences!
  - “`Mayo Clinic saves lives`”
  - “`Mayo Clinic uses AI`”
- Becomes:
  1. (“`Mayo`”,“`Clinic`”)
  2. (“`Mayo Clinic`”,“`saves`”)
  3. (“`Mayo Clinic saves`”,“`lives`”)
  4. (“`Mayo`”,“`Clinic`”)

What is the training data?

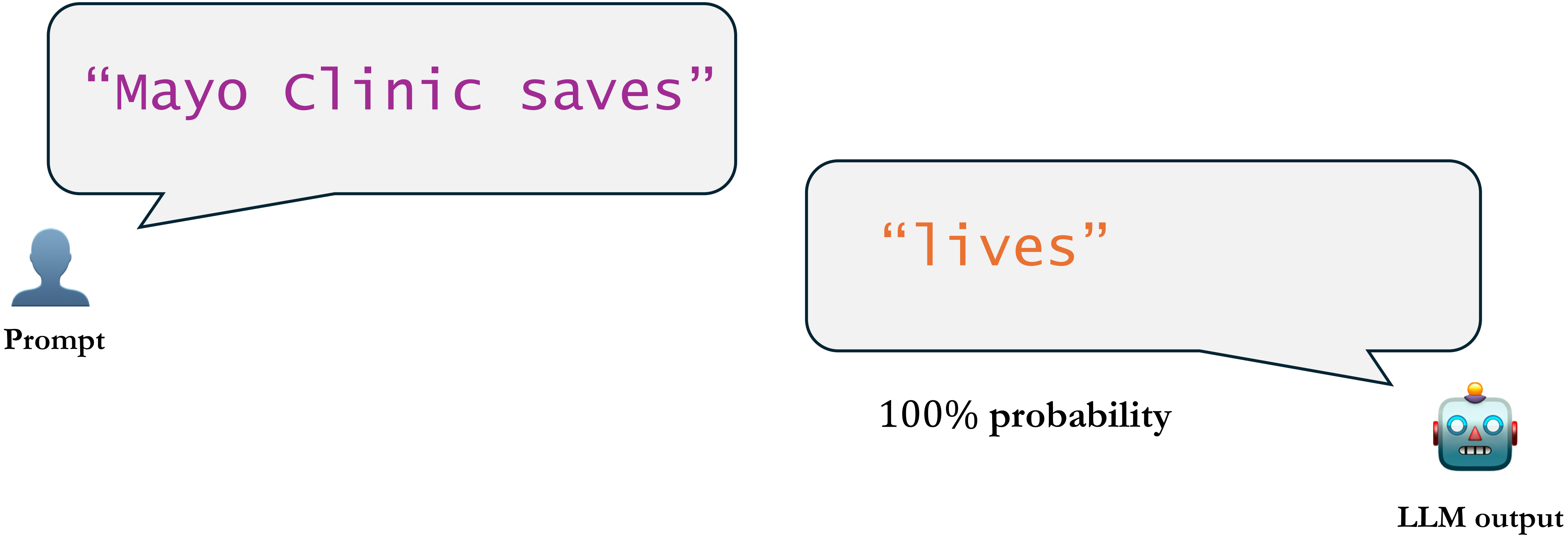
- **Dataset** = (Input 1, Output 1), (Input 2, Output 2), …, etc.
- *Six* training datapoints in two sentences!
  - “`Mayo Clinic saves lives`”
  - “`Mayo Clinic uses AI`”
- Becomes:
  1. (“`Mayo`”,“`Clinic`”)
  2. (“`Mayo Clinic`”,“`saves`”)
  3. (“`Mayo Clinic saves`”,“`lives`”)
  4. (“`Mayo`”,“`Clinic`”)
  5. (“`Mayo Clinic`”,“`uses`”)

What is the training data?

- **Dataset** = (Input 1, Output 1), (Input 2, Output 2), ..., etc.
- *Six* training datapoints in two sentences!
  - "`Mayo Clinic saves lives`"
  - "`Mayo Clinic uses AI`"
- Becomes:
  1. ("`Mayo`","`Clinic`")
  2. ("`Mayo Clinic`","`saves`")
  3. ("`Mayo Clinic saves`","`lives`")
  4. ("`Mayo`","`Clinic`")
  5. ("`Mayo Clinic`","`uses`")
  6. ("`Mayo Clinic uses`","`AI`")

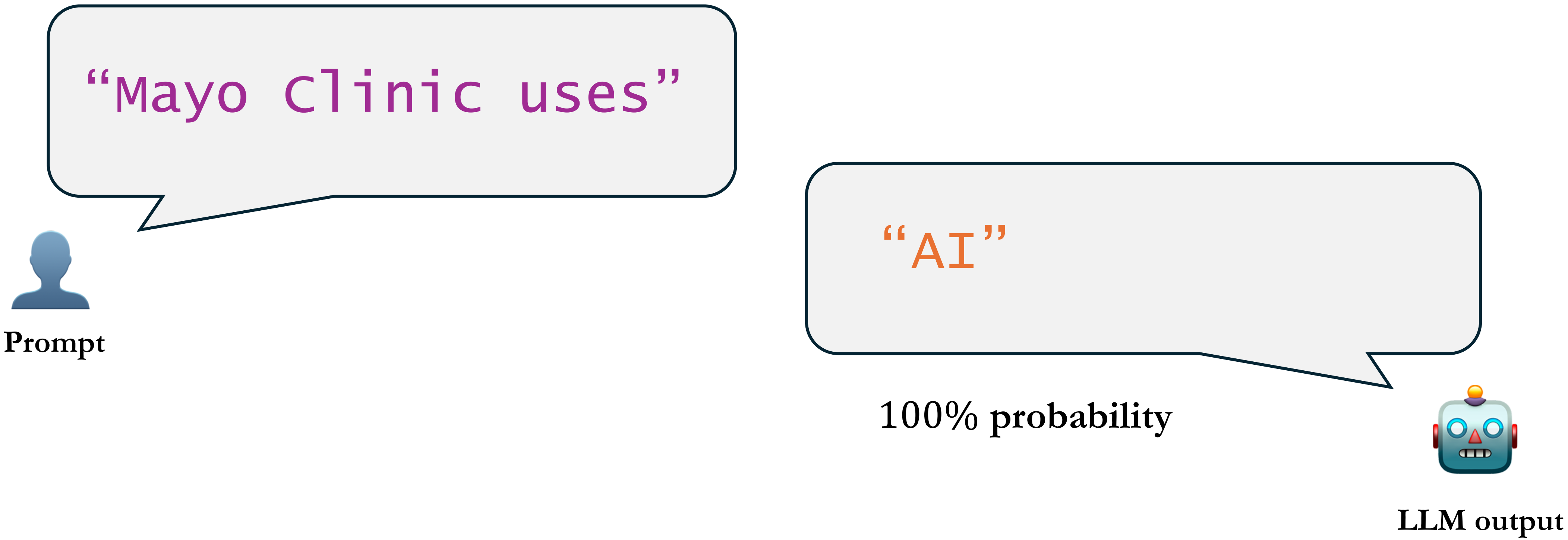
What will the network learn?

- If gradient descent succeeds, will *match* distribution of training dataset.
- Like a parrot!



What will the network learn?

- If gradient descent succeeds, will *match* distribution of training dataset.
- Like a parrot!



What will the network learn?

- What if the training data has *multiple* occurrences of phrase?
- Minimize loss → assign equal probability to each outcome.

"Mayo Clinic"

Prompt

"saves lives" 50% probability

"uses AI" 50% probability

LLM output

What will the network learn?

- What if the training data has *multiple* occurrences of phrase?
- Minimize loss → assign equal probability to each

"Mayo Clinic"

Prompt

What does this mean for LLM use in healthcare?

50% probability

"uses AI"

50% probability

LLM output

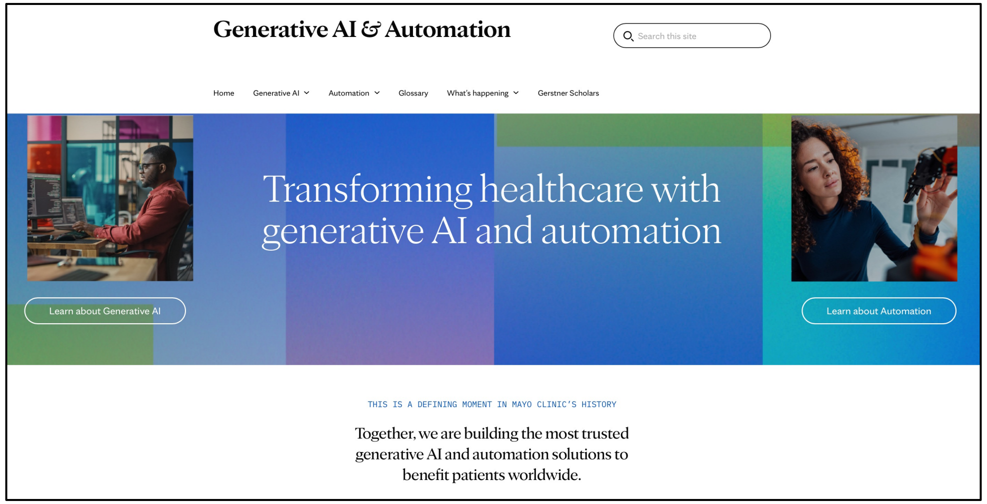
Generalization

- Can now create digital parrot.
- Memorization ≠ understanding.

- **Real question**: can LLMs complete text that *isn't anywhere online?*
  - *Radiation oncology Shakespearean poem?*
  - *Feedback on a never-before-seen radiation treatment plan?*
- Yes! As long as training data *large* and *diverse* enough.
  - Won't work for toy internet example.
  - But when trained on *real* internet, LLMs ***learn the principles of language and reasoning.***

Recap

- Large language models like ChatGPT are **neural networks** that:
  - Represent text as numbers.
  - Predict the next word given context.
  - Are trained on the Internet.
  - *But* have demonstrated ability to generalize beyond it.
- ***And they are changing healthcare and the world!***



Thank you for listening!

Q&A