What Just Happened?

Learning from our Frabjous experiment

What just happened?

We taught a computer to classify shapes as *Frabjous* or *Not Frabjous*.

- Did it just memorize our answers?

What just happened?

We taught a computer to classify shapes as *Frabjous* or *Not Frabjous*.

- Did it just memorize our answers?
- Why could it make predictions about *new* shapes?
  - soon...

What just happened?

We taught a computer to classify shapes as *Frabjous* or *Not Frabjous*.

- Did it just memorize our answers?
- Why could it make predictions about *new* shapes?
  - soon...
- What, exactly, did it learn?
  - the deep question ...

One thing we did *not* do

We never gave the computer a definition of *frabjous*.

We gave it **examples** and our **judgments**.

examples, not a rule

- We said: “This one is Frabjous.” “That one is not.”
- We did *not* say: “A polygon is Frabjous exactly when . . . ”
- The network had to use the examples to form its own way of classifying new shapes.

Question: Could two groups teach the same computer two different meanings?

examples, not a rule

- We said: “This one is Frabjous.” “That one is not.”
- We did *not* say: “A polygon is Frabjous exactly when . . . ”
- The network had to use the examples to form its own way of classifying new shapes.

Question: Could two groups teach the same computer two different meanings?

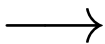
Sure

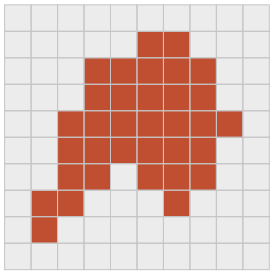
under the hood

How does the computer see a shape?



A picture of a polygon

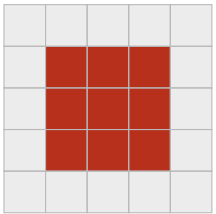




A grid of colored pixels

Each pixel is represented by **numbers** describing its color.

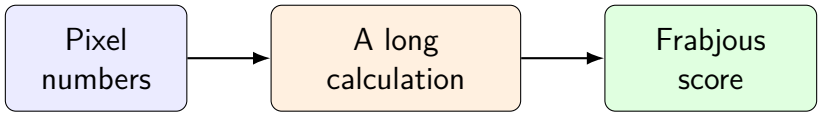
More detail: what are pixel numbers?



Each little square has three numbers:

- How much red
- How much green
- How much blue
- etc ...

From scanning numbers to a logical (?) prediction



The computer produces a number between 0 and 1.

How does it make a predication ?

- The calculation contains many adjustable numbers.
- These numbers are called **weights**.
- Changing the weights changes the computer's predictions.
- (This is the math we are skipping :-)

What happens when we teach it?

1. The computer makes a prediction.

What happens when we teach it?

1. The computer makes a prediction.

2. We give our judgment.

What happens when we teach it?

1. The computer makes a prediction.
2. We give our judgment.
3. The computer measures how far its prediction was from our judgment.

What happens when we teach it?

1. The computer makes a prediction.
2. We give our judgment.
3. The computer measures how far its prediction was from our judgment.
4. It **adjusts the weights** to try to reduce that difference.

What happens when we teach it?

1. The computer makes a prediction.
2. We give our judgment.
3. The computer measures how far its prediction was from our judgment.
4. It **adjusts the weights** to try to reduce that difference.
5. THIS IS THE CRUCIAL IDEA

What happens when we teach it?

1. The computer makes a prediction.
2. We give our judgment.
3. The computer measures how far its prediction was from our judgment.
4. It **adjusts the weights** to try to reduce that difference.
5. THIS IS THE CRUCIAL IDEA

This is how the AI "learns" from user input.

Based on how successful it was,

It adjusts the "weights"

Other predictions change!

- The *same* weights are used for every shape.
- Teaching it about shape 7 can change its prediction about shape 42.
- That is more than keeping a list of answers.

summarizing

careful...

- Our classroom network is **tiny**; modern AI systems are vastly larger and more complex.

careful...

- Our classroom network is **tiny**; modern AI systems are vastly larger and more complex.
- But both learn by adjusting numerical weights using examples and feedback.

careful...

- Our classroom network is **tiny**; modern AI systems are vastly larger and more complex.
- But both learn by adjusting numerical weights using examples and feedback.
- Learning a pattern is not the same as being guaranteed to give the right answer.

More detail: what this little AI cannot do

- Twenty or even a hundred polygons are *very few* training examples.
- Color may be easy to learn; concavity and counting corners can be harder.
- The network can be confidently wrong.
- Unlike a proof in logic, its prediction comes with no guarantee.

More detail: the difference from traditional logic

Logic

- State precise premises and rules.
- Apply valid reasoning.
- A valid inference preserves truth.

Our neural network

- Supply labeled examples.
- Adjust weights to fit examples.
- Predictions may be mistaken.

Both involve mathematics; they offer different kinds of guarantees.

More detail: did we create a meaning?

- At first, *Frabjous* had no agreed definition.
- Our judgments influenced the computer's classifications.
- Its predictions may have influenced our judgments in return.

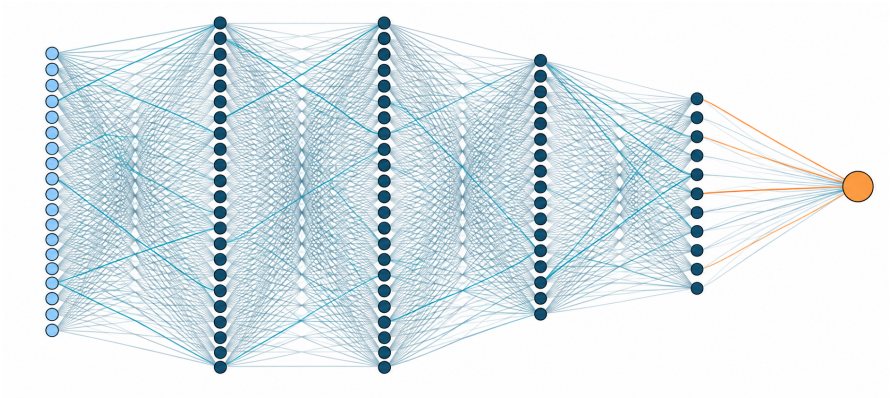
Perhaps the meaning was emerging through the interaction itself.

One last question

Who decided what *frabjous* means?

Us? The computer? Our interaction?

A neural network



Inputs | Layers of calculations | Output

Illustrative network; the app's exact architecture may differ.

A large language model

An LLM is a neural network trained to predict **the next word** from the words that came before it.

Our Frabjous network

- Input: numbers describing a shape.
- Output: a score for Frabjous.

A language model

- Input: numbers representing words in context.
- Output: probabilities for possible next words.

Both use learned weights to make predictions.

Predicting the next word

The cat sat on the _____

Possible next words: mat, floor, sofa, ...

- The model assigns a probability to each possible next word.
- Earlier words affect those probabilities.
- A word is selected using those probabilities. It need not be the most likely one.

How does it learn to predict words?

1. During training, the model sees a passage with the next word held back.
2. It predicts which word comes next.
3. The actual next word in the passage supplies the feedback.
4. Training **adjusts the weights** to make that word more likely in that context.

The same learning idea, repeated over enormous amounts of text.

A response grows one word at a time

Words so far	Next word selected
The cat sat on the	mat
The cat sat on the mat	and
The cat sat on the mat and	waited

Each selected word joins the context for the next prediction. In a chat, the context also includes your question and available conversation history.

Using context normally leaves the model's weights unchanged. **Training** is when the weights change.

A plausible continuation can still be factually wrong.