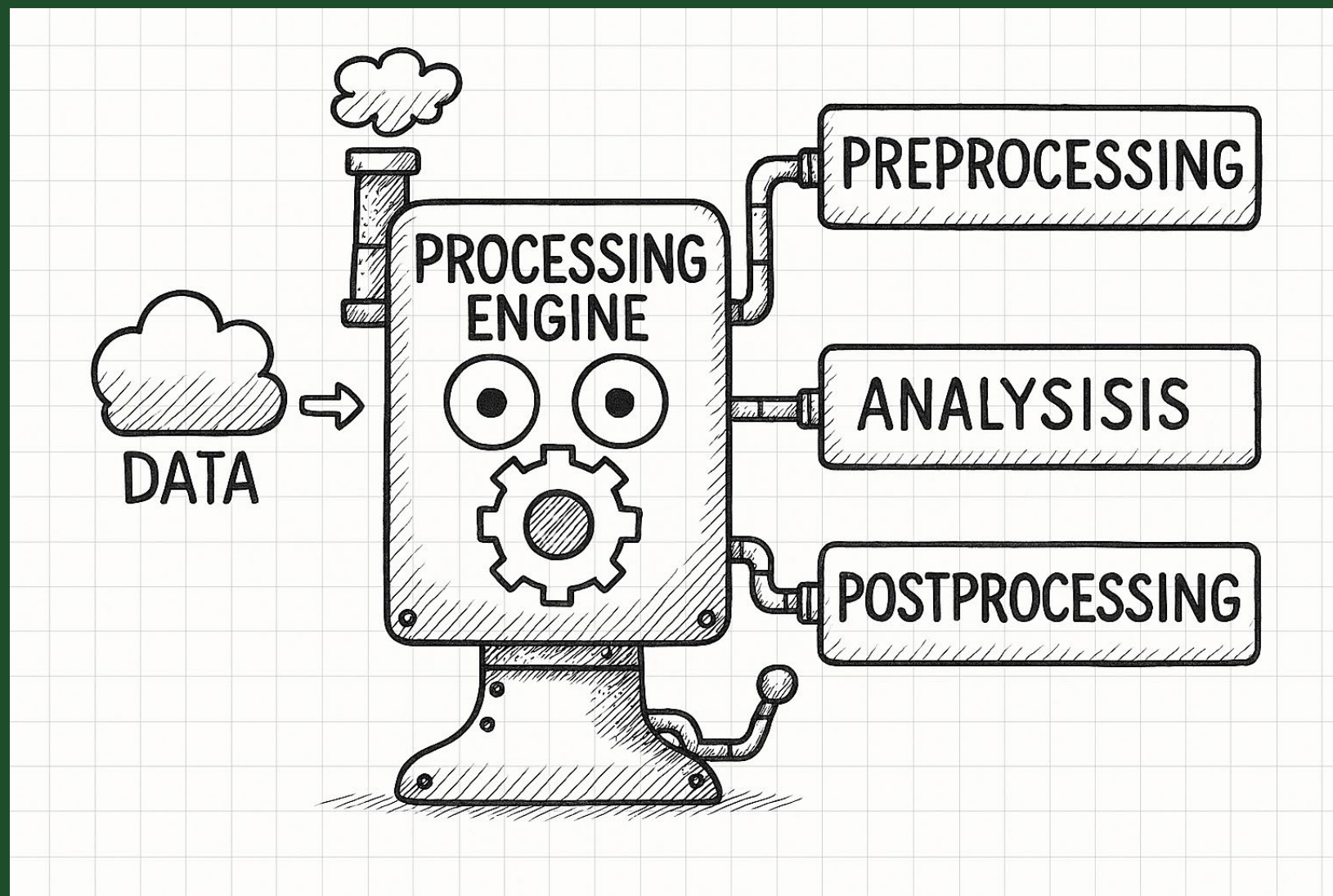


Demystifying AI : A Quick Introduction

Bruce Draper
Department of Computer Science
Colorado State University



AI is a data processing engine



Data gets fed in, it gets processed, and results are produced in a desired format.

Examples:

Financial data in → chart out

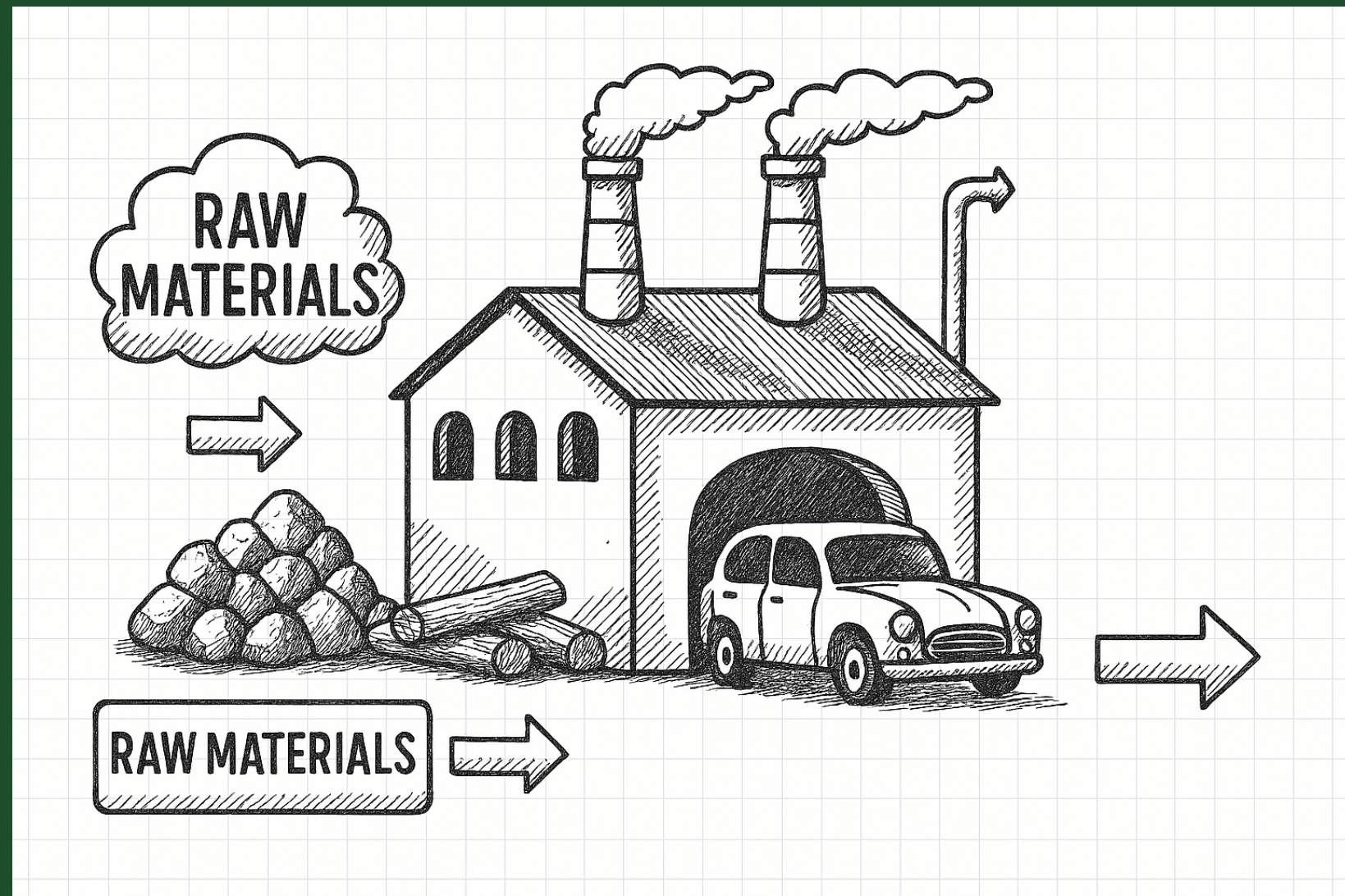
Sensor signals in → control signal out

But it's not like any other data processor:

- The input data is unstructured (e.g. raw text, video feeds, unfiltered data)
- The engine itself is trained on (massive amounts of) data, not programmed
- The goal of the engine is to predict the most likely output, given the input

Redefining Automation

Pre-AI Automation



We used to automate repetitive, physical processes, either with robots or assembly-line workers

Post-AI Automation

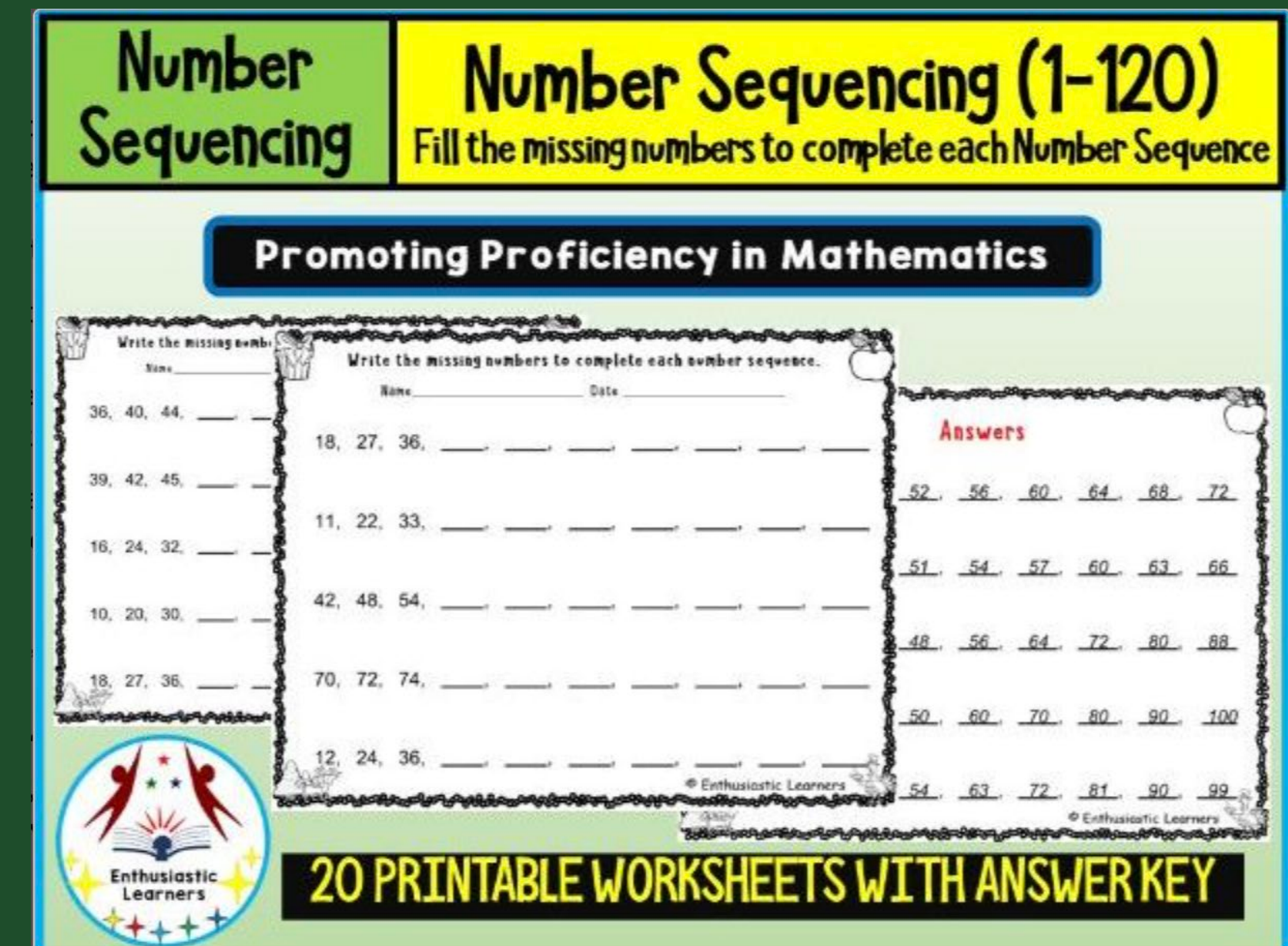


Now we automate anything repetitive tasks, physical or otherwise, where 'repetitive' admits many predictable variations.

How Does AI Work (simplified)?

Remember sequence completion questions?
Example: 1, 1, 2, 3, 5, 8, ... What's next?

13



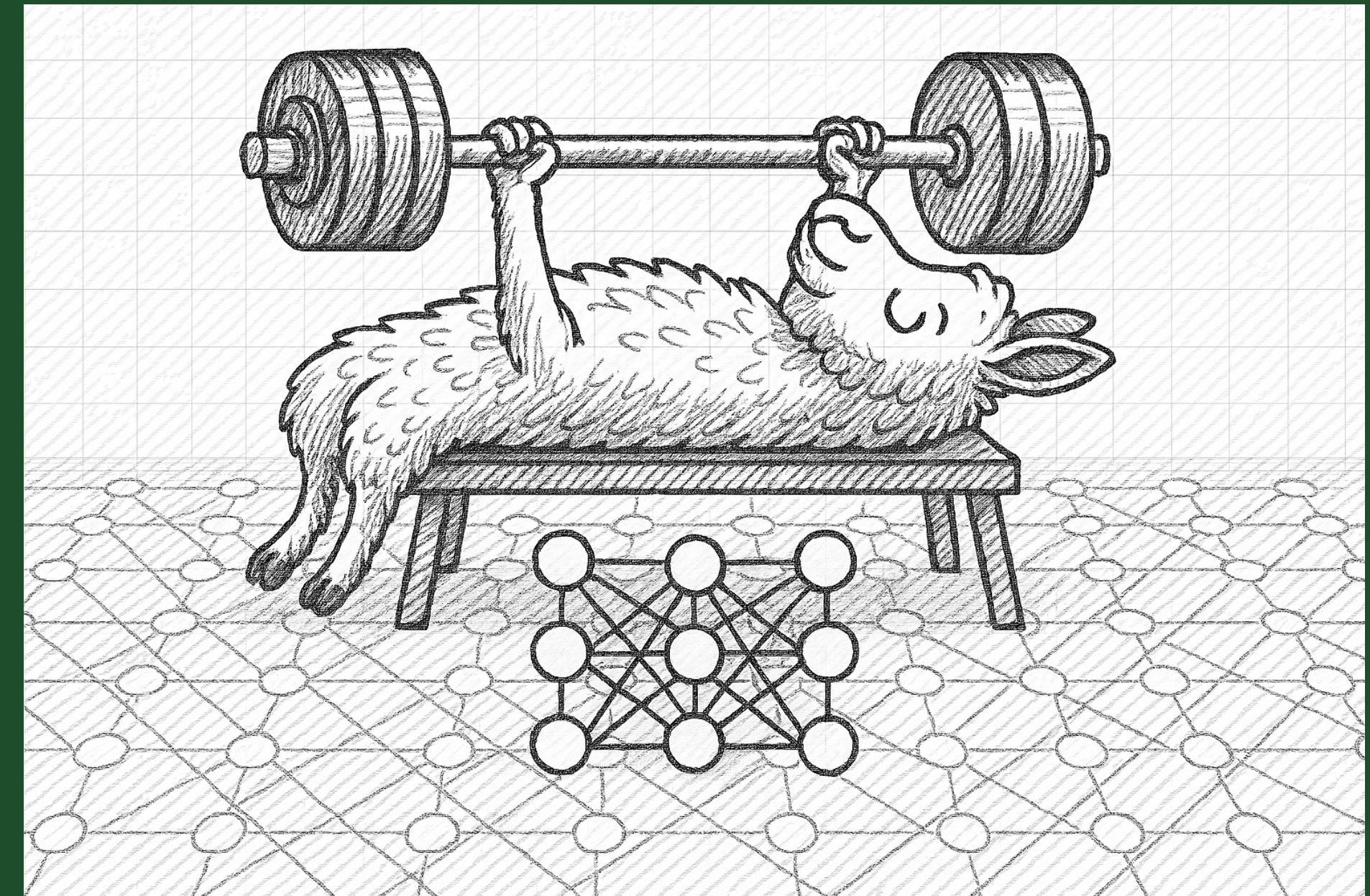
This is what AIs do. They take an input sequence (called a 'prompt') and predict the next value (called a 'token'). Tokens may be words or something else.

The predicted token is then added to the end of the prompt and the process repeats thereby predicting the next token, until a special 'stop' token is predicted.

Where is the Magic?

The magic is in the model (we'll use Meta's LLaMa 7b):

- A lexicon of 32,000 words (tokens)
- A prompt (starting sequence) of 4,096 tokens
- An associative model with 7bn parameters
- Trained over a significant chunk of the internet



Weights (parameters) reflect how strongly tokens (words) relate to each other at various distances over the training set.

So the magic is in the co-occurrences in the training data, as reflected in the weights.

AI Pitfalls

- **Hallucination**

AI models will always respond with *something*. If they don't know the answer, they will make it up. Self-checking (e.g. RAG) helps, but the problem remains.

- **Sycophancy**

AI models will tell you what you want to hear, whether or not it's correct. This is an example of an **alignment** problem.

- **Bias**

Unknown statistical biases in the training data create biased models, unbeknownst to the developer or user.

- **Overconfidence**

AI models are just as confident of hallucinated answers as real ones. In fact, they are consistently overconfident in their responses.



Human Pitfalls (around AI)

- **Automation Bias**

People over-trust information from an AI (or any computer). Always double check. Maybe Google Maps doesn't know the bridge is out. Trust your eyes.

- **Responsibility**

Who is responsible for AI actions? If a self-driving car kills someone, who is liable? A person needs to be “in the loop”, meaning empowered and alert and willing to step in.

- **Skill Erosion**

Many entry-level skills can be automated, but without entry-level workers learning on the job, expertise becomes a rare commodity.



Final Quotes

“The boring parts of every job will be automated”

-- Shrideep Pallickara

“Assume this is the worst AI you will ever use”

-- Ethan Mollick

