DeUmbra

A Path to Strong AI

Jonathan Mugan
@jmugan

Keynote
Data Day Texas
June 2022

Why we want AI

David (Deddy) Dayag, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

We need strong artificial intelligence (AI) so it can help us

- understand the nature of the universe to satiate our curiosity,
- devise cures for diseases to ease our suffering,
- and expand to other star systems to ensure our survival

Why we want AI

David (Deddy) Dayag, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

- What happened before the big bang? Why does the universe exist at all?
- Why is there something rather than nothing?
- Why doesn't my cloud setup work?

Teach me things I don't know by pinpointing exactly where I don't understand and explain it using concepts I do.

Why we want AI

Two desires for AI:

- Teaching: requires **meaning** for language and world understanding with dynamic coordination
- Discovery: requires **meaning** and **search** with access to experimental domain

This talk is about where we are and what we need to do.

Outline

- Why we want AI
- Recent big-compute methods have been surprisingly good
- We still need meaning
- How to get there
- A pseudocode of consciousness

Outline

- Why we want AI
- Recent big-compute methods have been surprisingly good
- We still need meaning
- How to get there
- A pseudocode of consciousness

Illustration of Progress

mid 2010s

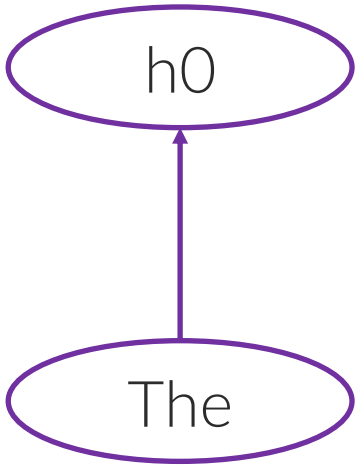
Foundational Models

Term from Stanford paper
https://fsi.stanford.edu/publication/opportunities-and-risks-foundation-models

Foundational models are big neural networks trained on massive amounts of data, such as news articles and images with captions.

Encoding sentence meaning into a vector

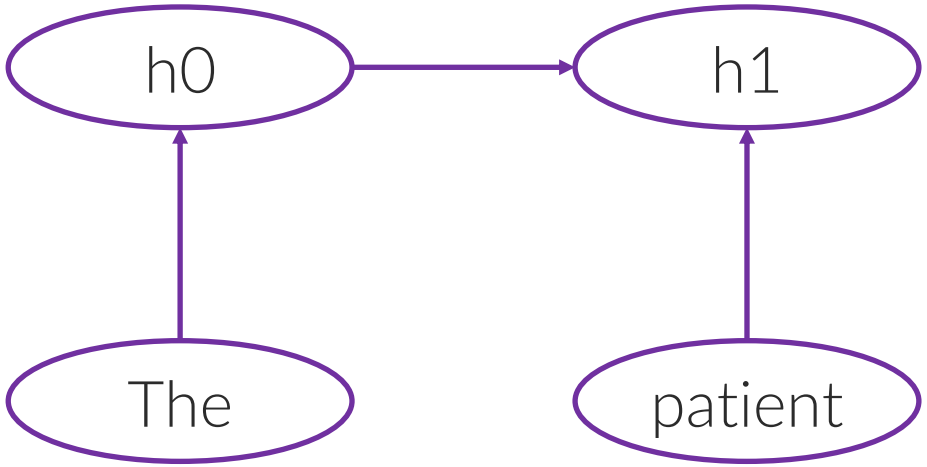
"The patient fell."



Using a recurrent neural network (RNN).

Encoding sentence meaning into a vector

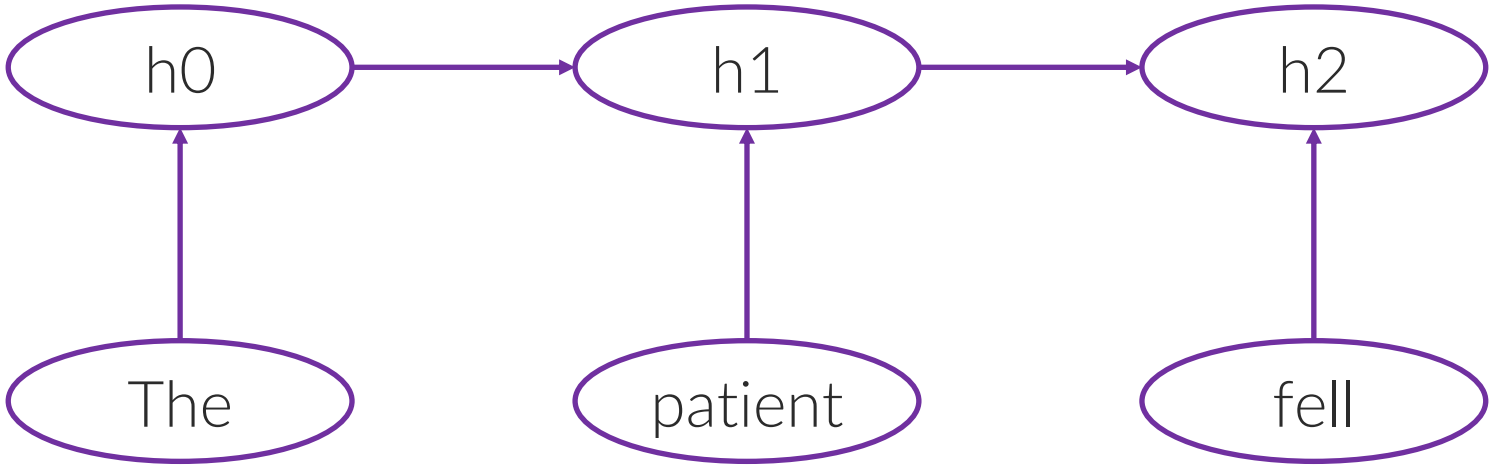
"The patient fell."



Using a recurrent neural network (RNN).

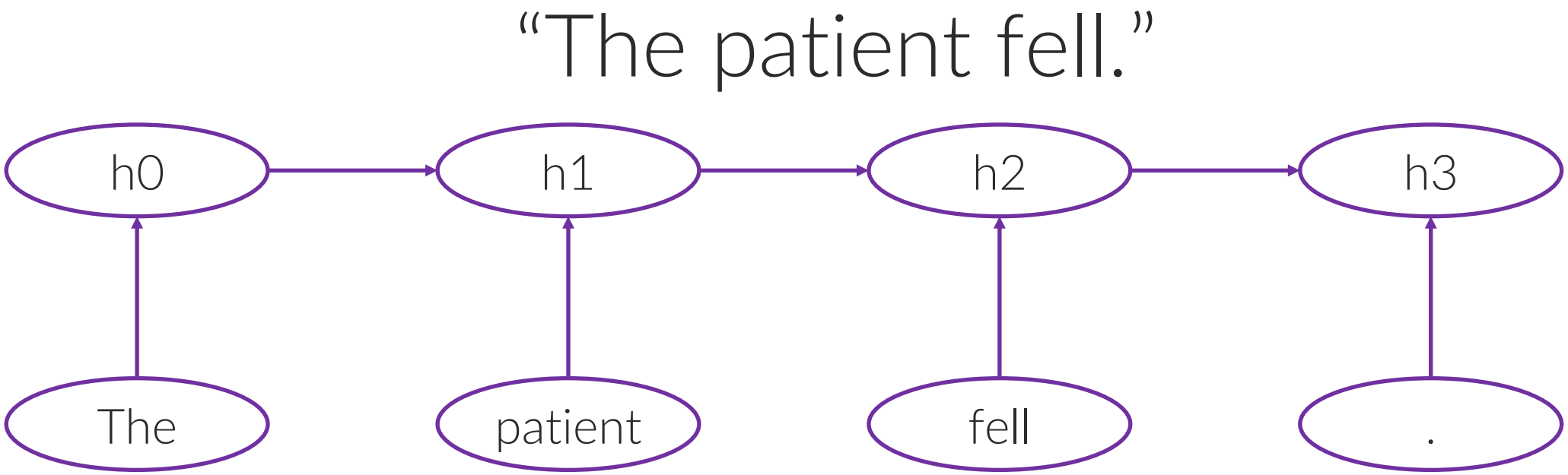
Encoding sentence meaning into a vector

"The patient fell."



Using a recurrent neural network (RNN).

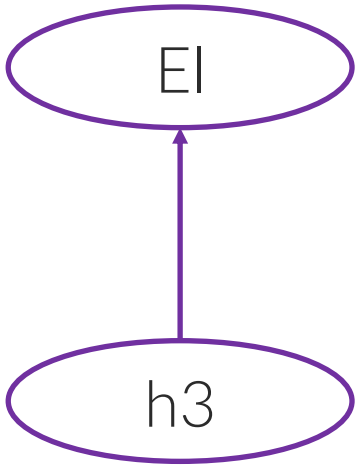
Encoding sentence meaning into a vector



RNN is like a hidden Markov model but doesn't make the Markov assumption and benefits from a vector representation.

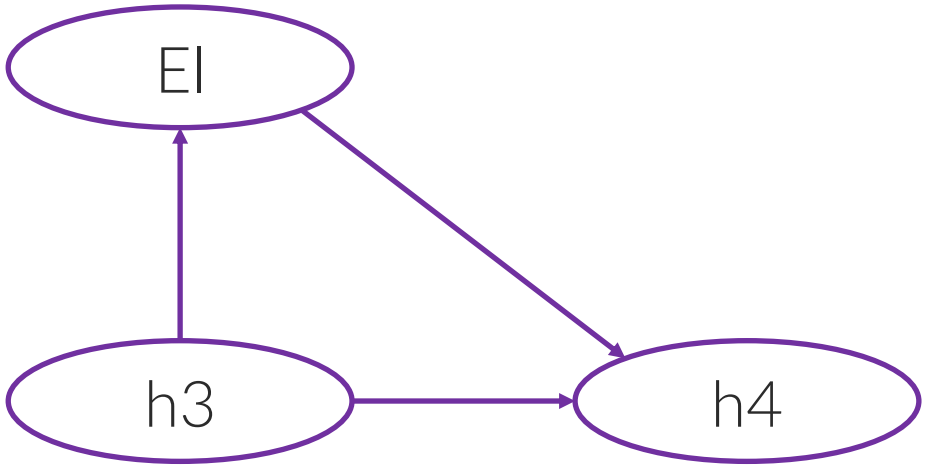
Decoding sentence meaning

Machine translation.



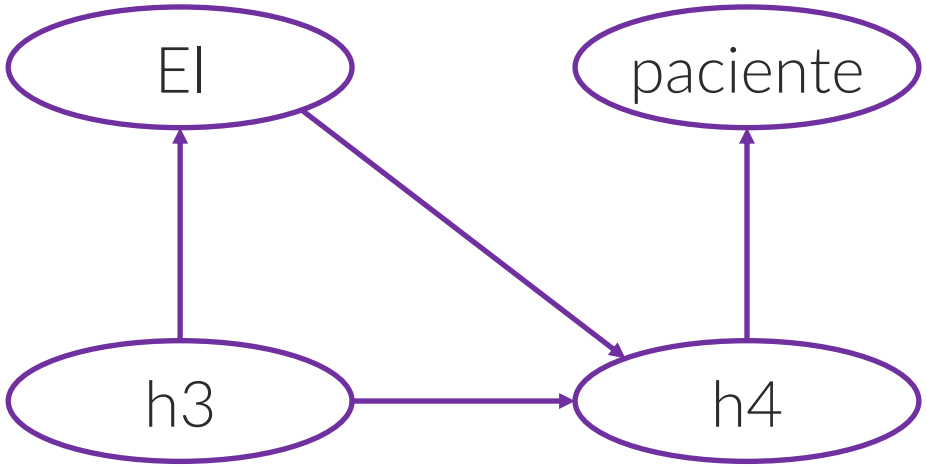
Decoding sentence meaning

Machine translation.



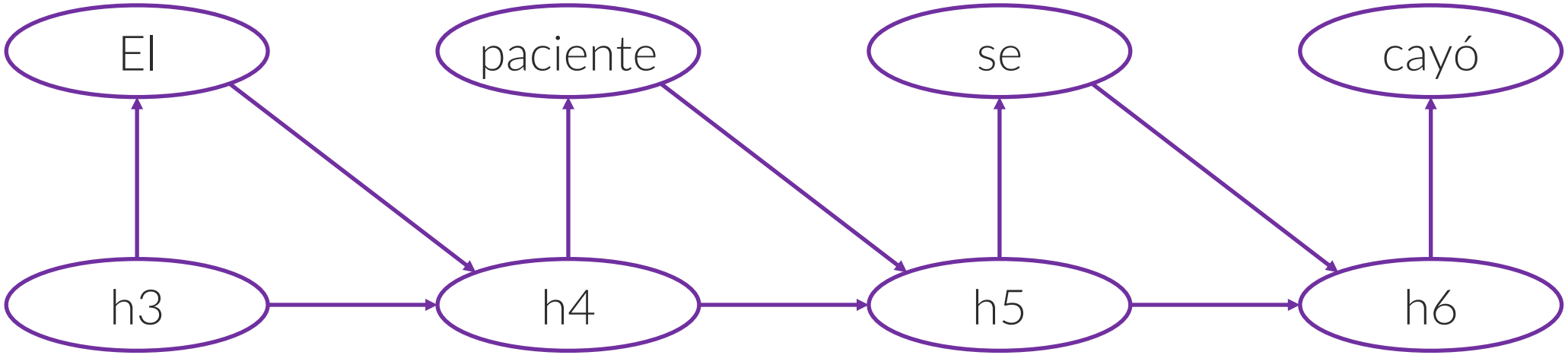
Decoding sentence meaning

Machine translation.



Decoding sentence meaning

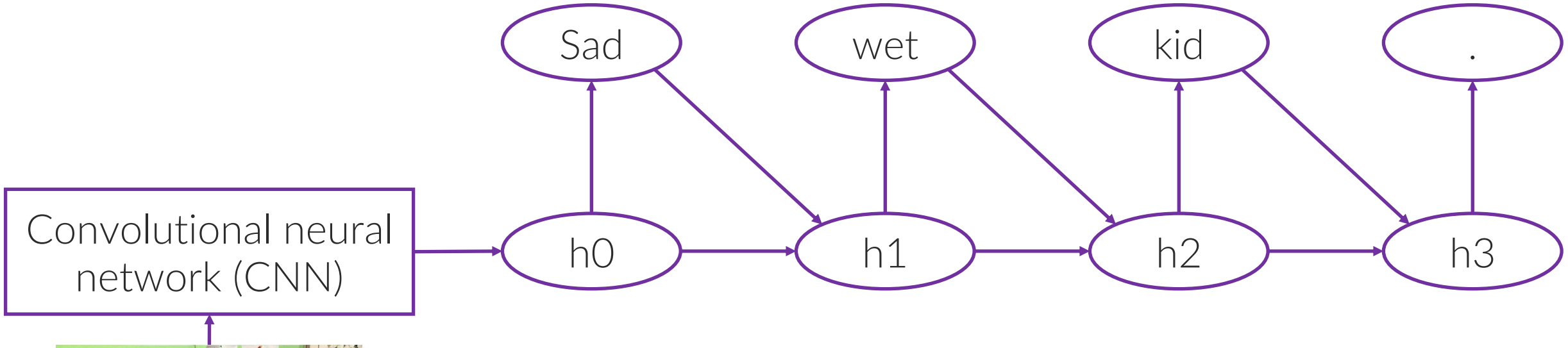
Machine translation.



[Cho et al., 2014]

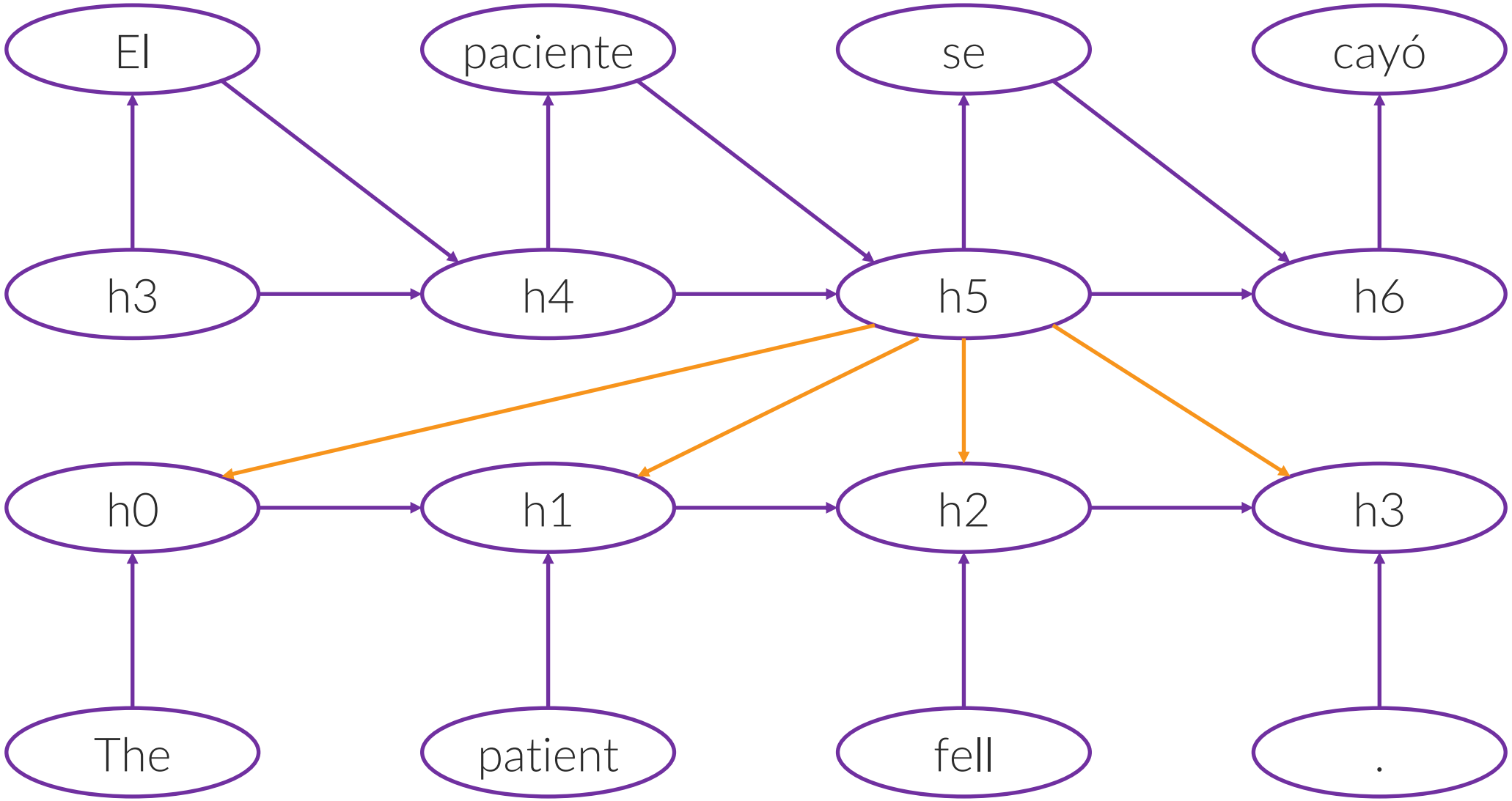
- It keeps generating until it generates a stop symbol.
- It is using a kind of interpolation from a huge set of training data.

Can also generate image captions



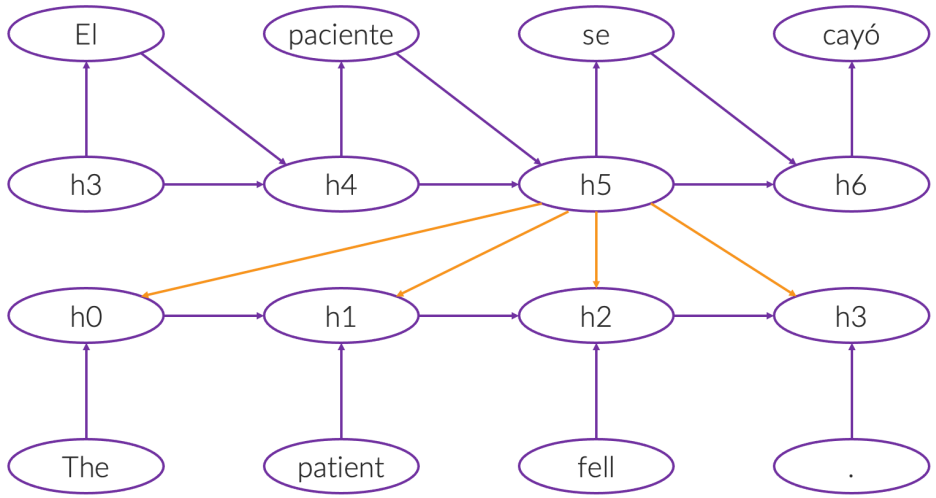
[Karpathy and Fei-Fei, 2015]
[Vinyals et al., 2015]

Attention [Bahdanau et al., 2014]



Transformers: Attention is all you need

https://arxiv.org/abs/1706.03762



And GANs and diffusion models and other cool and amazing methods.

Illustration of Progress

mid 2010s

Foundational Models

Foundational models do a great job of text and image encoding and decoding

- prompt engineering
- model integration

Now

Model Systems

Language models are given prompts

Prompt

```
from transformers import pipeline

text_generator = pipeline("text-generation")
print(text_generator("The space aliens came to me to talk about",
                     max_length=100))
```

```
Downloading: 100%|██████████| 230/230 [00:00<00:00, 73.4kB/s]
Setting `pad_token_id` to 50256 (first `eos_token_id`) to generate
sequence
[{'generated_text': 'The space aliens came to me to talk about thei
r own race to whom they had come to enslave. Because they were alie
ns, they were more like friends, the only beings of intelligence th
at humans came to know. I knew that for much but it just so happene
d that at least one of them was a human. And not even that one coul
d live in my life, only by way of the Force, at least not to have a
ny sort of sense of morality or morality or self. I was also'}]
```

HuggingFace https://huggingface.co/

Prompts can also encode what kind of answer you want, and the model "learns" just from that

Helsinki is the capital and largest city of Finland. It is in the region of … The Helsinki metropolitan area includes the urban core of Helsinki, Espoo, …

Q: what is the most populous municipality in Finland?
A: Helsinki
Q: how many people live there?
A: 1.4 million in the metropolitan area
Q: what percent of the foreign companies that operate in Finland are in Helsinki?
A: 75%
Q: what towns are part of the metropolitan area?
A:

Before, you had to fine-tune on a particular task, but with prompt engineering, you teach it in real time.

Example from "Language Models are Few-Shot Learners" by Brown and friends at OpenAI
https://arxiv.org/pdf/2005.14165.pdf

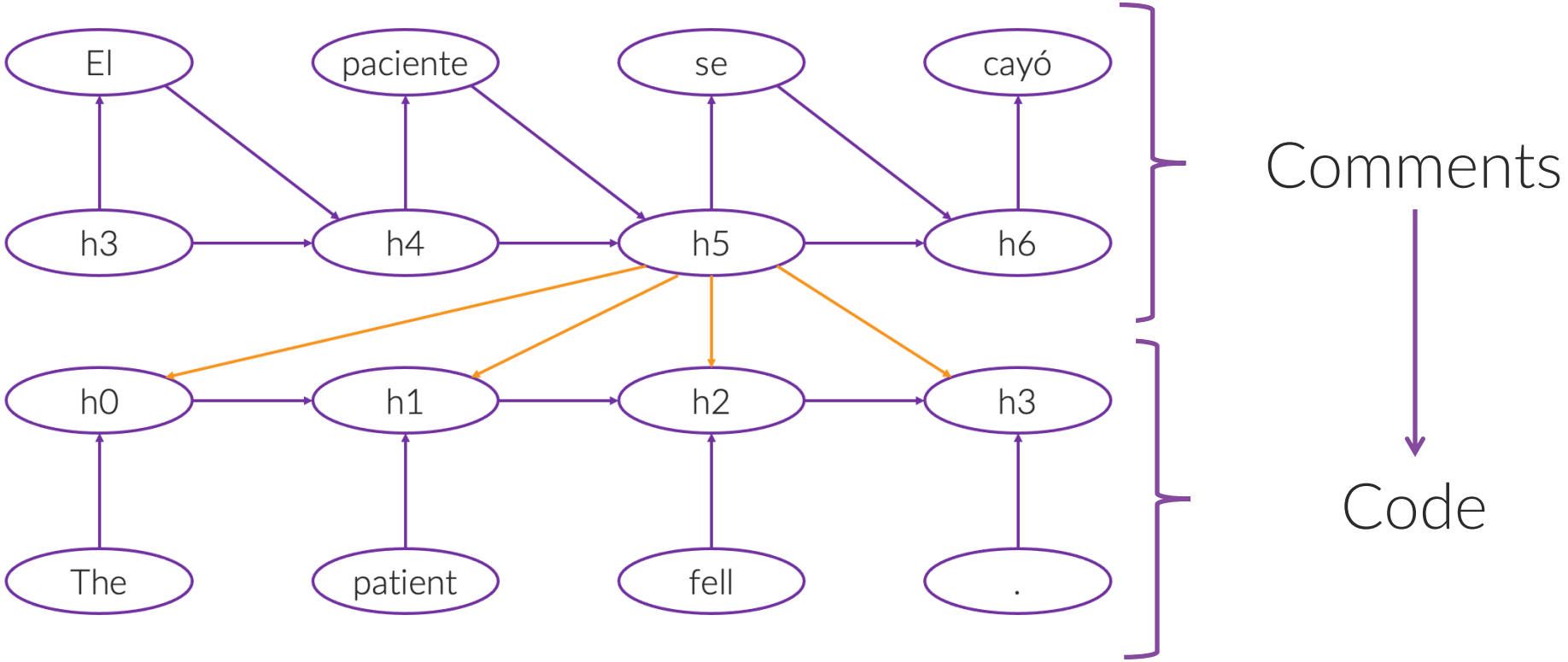
This whole thing is the prompt

Model output: *Helsinki, Espoo, Vantaa, Kuaniainen, and surrounding commuter towns*

OpenAI Codex and GitHub Copilot

https://openai.com/blog/openai-codex/
https://copilot.github.com/

Train on comment and code combinations.



Combining prompt engineering and code generation

Drori et al. make changes to college math questions, and CODEX automatically writes code to solve them using Python libraries.
https://arxiv.org/pdf/2112.15594.pdf

Example from paper:

Question: Find the differential $\delta w. w = \ln(x^2 + y^2 + z^2)$

Engineered prompt: In differential equations, write a function using sympy to find the differential of w=ln(x^2 + y^2 + z^2)

```
import sympy as sp
x, y, z = sp.symbols('x y z')
w = sp.log(x**2 + y**2 + z**2)
print(sp.diff(w,x))
print(sp.diff(w,y))
print(sp.diff(w,z))
```

Generated output

Extending Prompt Engineering

Also see “Chain of Thought Prompting Elicits Reasoning in Large Language Models” by Wei and friends at Google Brain. https://arxiv.org/abs/2201.11903

Extends prompt engineering by explaining why the previous answers are correct.

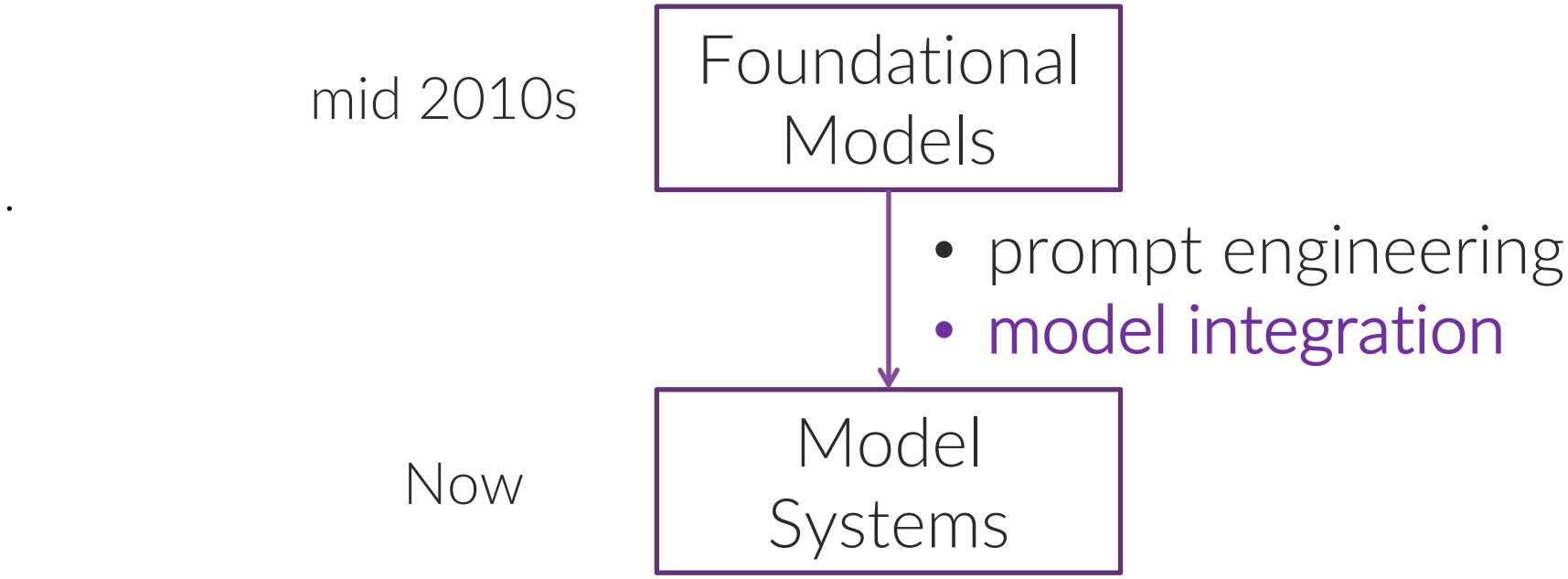
If you don’t just give it the answers like in the Helsinki example but say why those answers were computed, prompt engineering works even better.

Helsinki is the capital and largest city of Finland. It is in the region of … The Helsinki metropolitan area includes the urban core of Helsinki, Espoo, …

Q: what is the most populous municipality in Finland?
A: Helsinki
Q: how many people live there?
A: 1.4 million in the metropolitan area
Q: what percent of the foreign companies that operate in Finland are in Helsinki?
A: 75%
Q: what towns are part of the metropolitan area?
A:

Extend from “Language Models are Few-Shot Learners” by Brown and friends at OpenAI to say why the answers are what they are, and reasoning improves.

Illustration of Progress



OpenAI Dall-e 2

- Trained on combinations of images and text; CLIP and diffusion models
- Can create images from whatever you type

https://openai.com/dall-e-2/

Thanks to you @hardmaru for the images!

Prompt:
"Photograph of two robot cats going on a date in Manhattan at night"

OpenAI Dall-e 2

- Trained on combinations of images and text; CLIP and diffusion models
- Can create images from whatever you type

https://openai.com/dall-e-2/

Thanks to you @hardmaru for the images!

Prompt:
"Darth Vader on the cover of Vogue magazine"

OpenAI Dall-e 2

- Trained on combinations of images and text; CLIP and diffusion models
- Can create images from whatever you type

https://openai.com/dall-e-2/

Thanks to you @hardmaru for the images!

Prompt:
"Photograph of Apes attending the World Economic Forum in Davos"

Using language models to guide robot action

Have a single-armed mobile robot in a kitchen setting.

Do As I Can, Not As I Say:
Grounding Language in Robotic Affordances
https://say-can.github.io/

Robotics at Google and Everyday Robots

Human: I spilled my drink, can you help?

Robot: tries the text associated of each action and sees which one most likely in the language model.

Action 1: get sponge
Action 2: get vacuum

→ Language Model

I spilled my drink, can you help? I'll get a sponge. score = .062
I spilled my drink, can you help? I'll get a vacuum. score = .013

Since sponge has a higher likelihood in the language model, the sponge is the best action.

They are extracting world knowledge from language, very cool, but still not enough, as we will see.

DeepMind Flamingo: Conversations about pictures

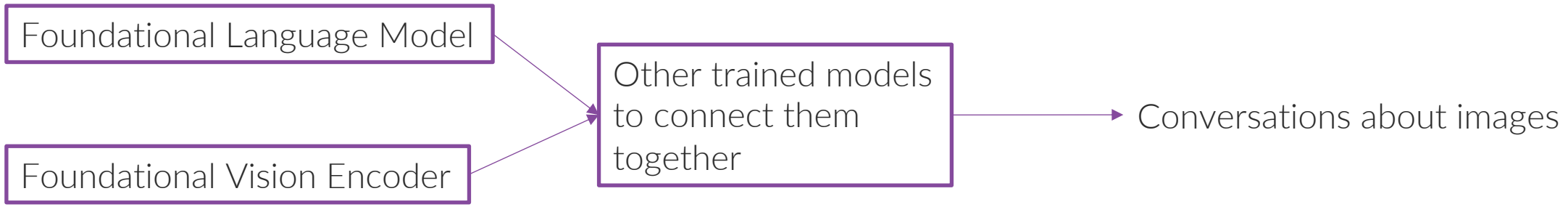


Image of weird cloth monster in soup

Example from DeepMind
https://www.deepmind.com/blog/tackling-multiple-tasks-with-a-single-visual-language-model

Human: What is in this picture

Flamingo: It's a bowl of soup with a monster face on it

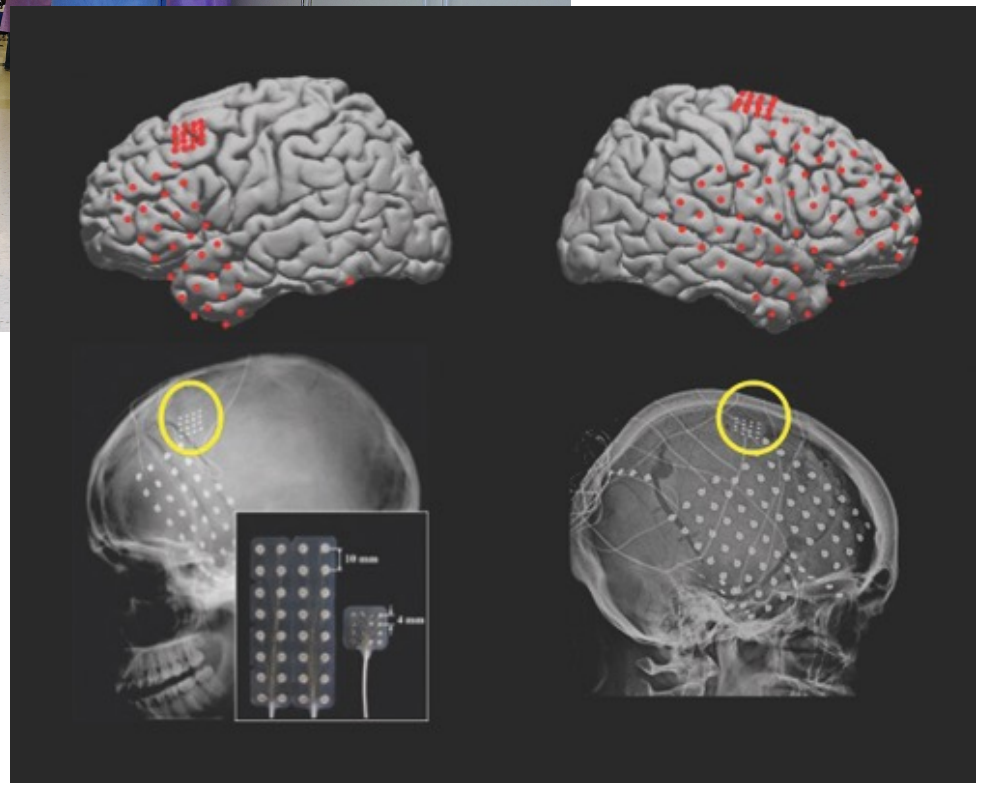
Human: What is the monster made out of?

Flamingo: It is made out of vegetables

Human: No, it's made out of a kind of fabric, can you see what kind?

Flamingo: It's made out of a woolen fabric.

ThoughtLog: The Automatic Diary by DeUmbra



- Researchers at the University of Buda (top)
- ThoughtLog implants (right)

Released on
April 1, 2022

ThoughtLog reads and automatically stores your thoughts in the cloud, even while you sleep.

Trained with an encoder-decoder model of brain waves to thoughts.

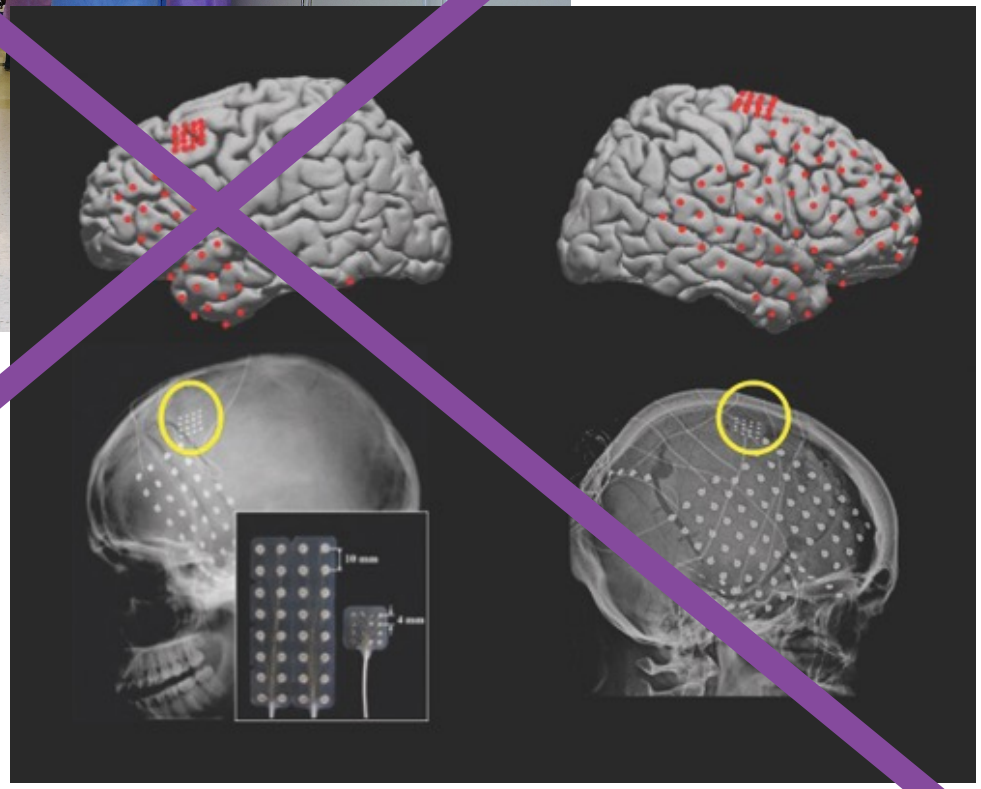
- Share your thoughts with loved ones!
- Become a better person!
- Help make our community safer!

ThoughtLog is completely free!

- Supported by advertising
- Encoder-decoder can go both ways
- (Not to worry. Most trial participants could distinguish sponsored thoughts from organic ones.)

For more details, see https://deumbra.com/2022/04/introducing-thoughtlog-the-free-automatic-diary/

ThoughtLog: The Automatic Diary by DeUmbra



- Researchers at the University of Buda (top)
- ThoughtLog implants (right)

Released on April 1, 2022

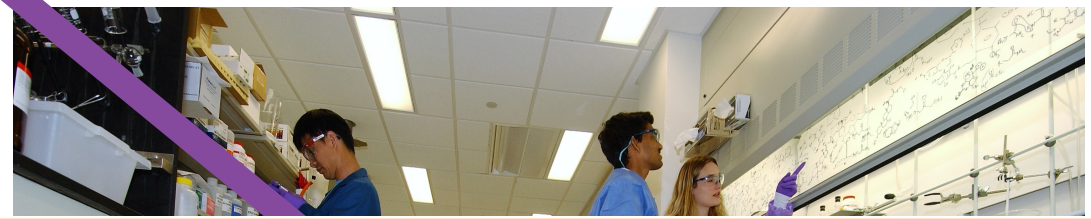
Not real.

Was an April Fool's Day joke, but as we go through this, we need to keep watch for unintended consequences of new technology.

Everything else in this talk is real.

For more details, see https://deumbra.com/2022/04/introducing-thoughtlog-the-free-automatic-diary/

ThoughtLog: The Automatic Diary by DeUmbra



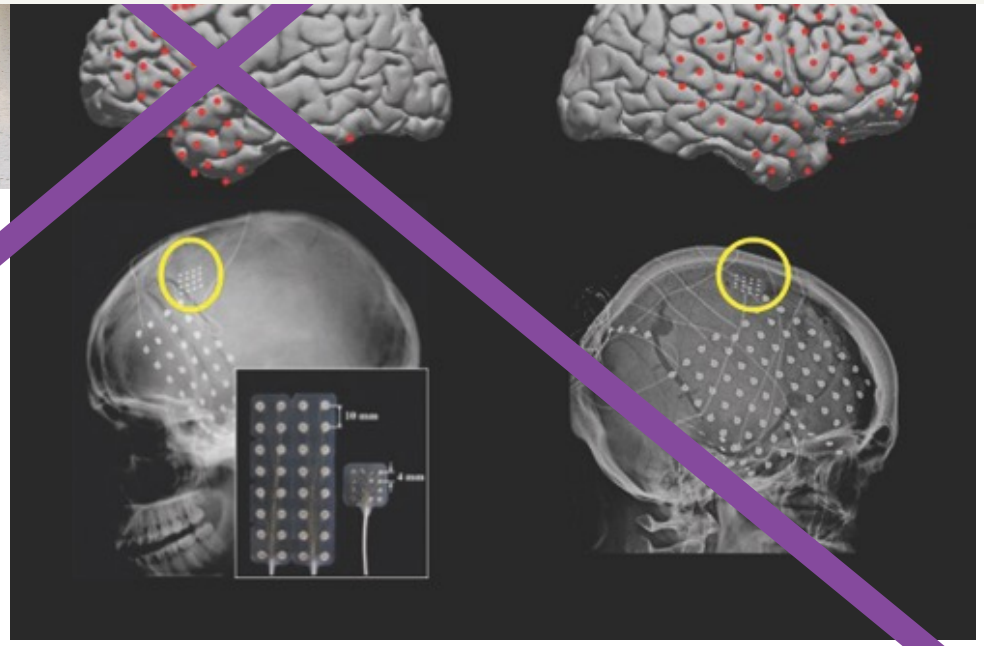
Not real.

Hacker News new | threads | past | comments | ask | show | jobs | submit

Was an April Fools Day

6. ▲ Reconstructing images a person sees via non-invasive brain scans (mind-vis.github.io)
65 points by yboris 4 hours ago | flag | hide | 49 comments

this, we need to keep watch for unintended consequences of new technology.



- Researchers at the University of Buda (top)
- ThoughtLog implants (right)

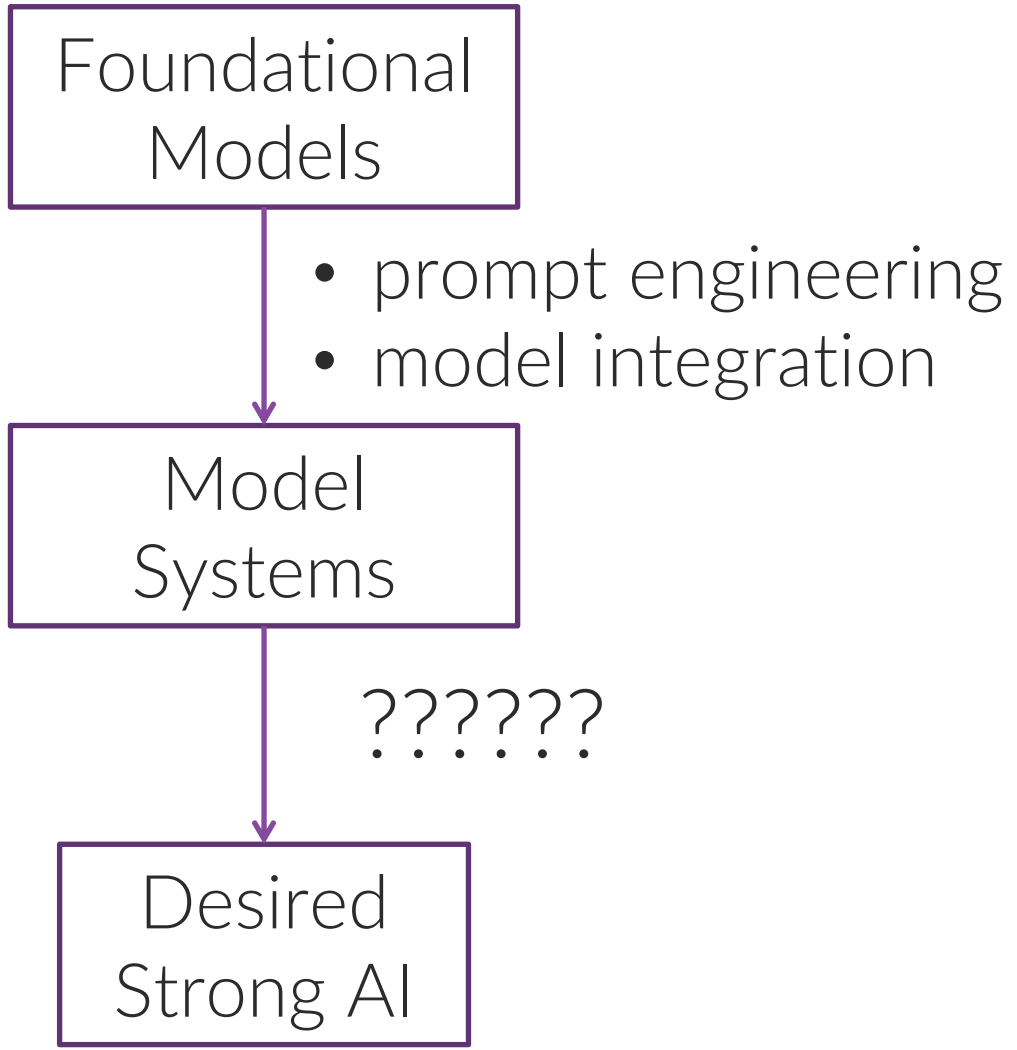
Released on April 1, 2022

Everything else in this talk is real.

For more details, see https://deumbra.com/2022/04/introducing-thoughtlog-the-free-automatic-diary/

Outline

- Why we want AI
- Recent big-compute methods have been surprisingly good
- We still need meaning
- How to get there
- A pseudocode of consciousness



- I view what we have seen so far as interpolation.
- Interpolation will get better, but it seems it must have limits, and I don't see how it can teach us to explore star systems.

We aren't there yet

These are amazing results, but I still can't consistently have a satisfying conversation with a personal assistant.

If the assistant doesn't understand, we are stuck.

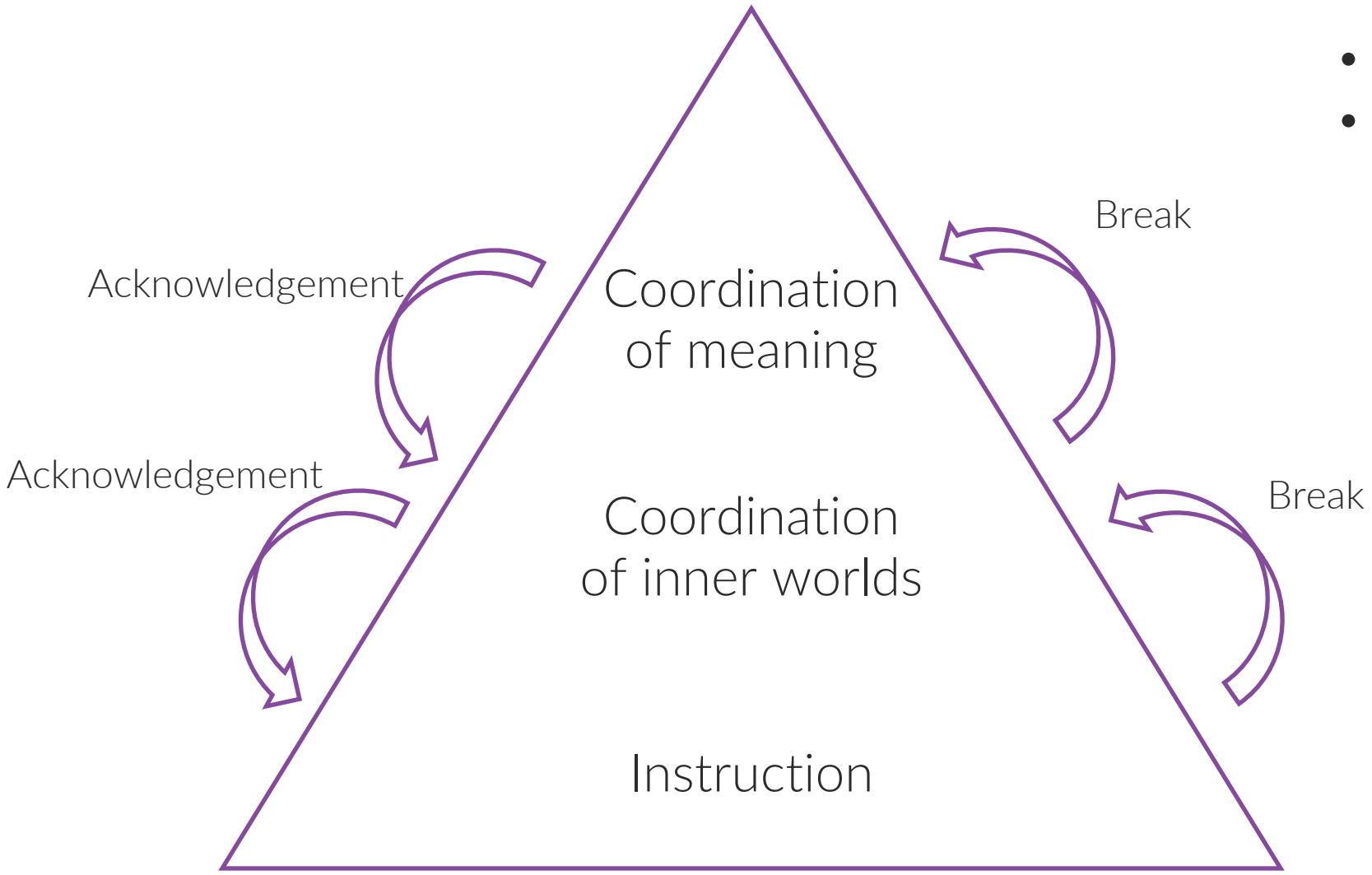
You can't have a bot walk you through how to set your cloud infrastructure.

You can't have a tutor identify what exactly you don't understand and explain it.

Maybe it's just that most modern bots haven't made it into products yet, but is an open question whether these methods will take us far enough

Let's dig a little deeper into what we need

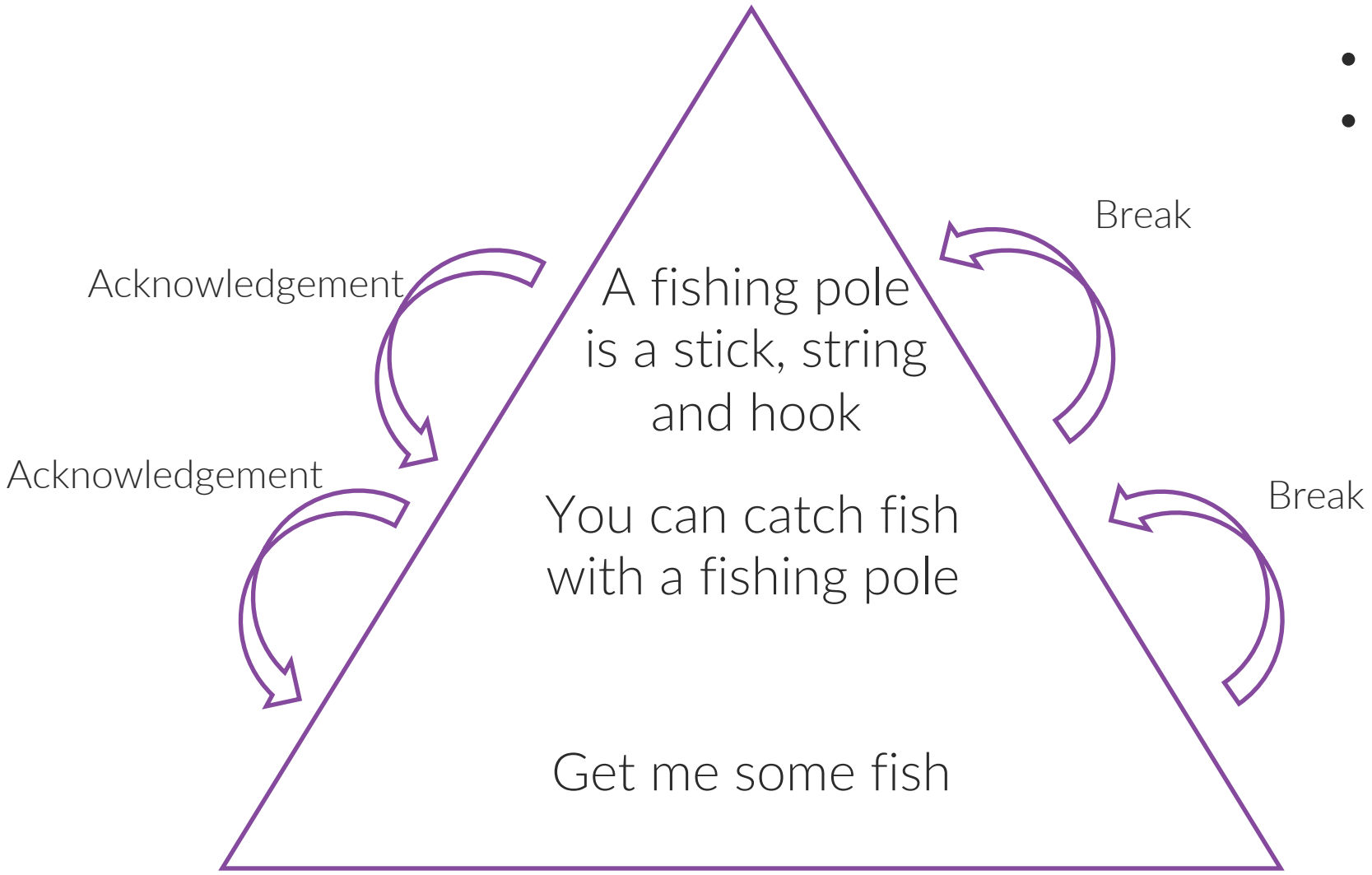
We negotiate meaning as we go



- Levels of discourse
- Complicated to go up and down the pyramid

Modified from Gärdenfors (2014), which was based on Winter (1998)

We negotiate meaning as we go



- Levels of discourse
- Complicated to go up and down the pyramid

Modified from Gärdenfors (2014), which was based on Winter (1998)

What does “meaning” mean?

How can we encode this in a computer?

Overview of "meaning"

Reminiscent of Robert Brandom through Richard Evans
https://www.doc.ic.ac.uk/~re14/Evans-R-2020-PhD-Thesis.pdf

"The table is on the table." → *What?*

Overview of "meaning"

Reminiscent of Robert Brandom through Richard Evans
https://www.doc.ic.ac.uk/~re14/Evans-R-2020-PhD-Thesis.pdf

"The table is on the table."

- If you want to pick up top table, you must first walk to the table
- If you push the bottom table, the top table will fall
- Party guests will think this is weird looking.

Overview of "meaning"

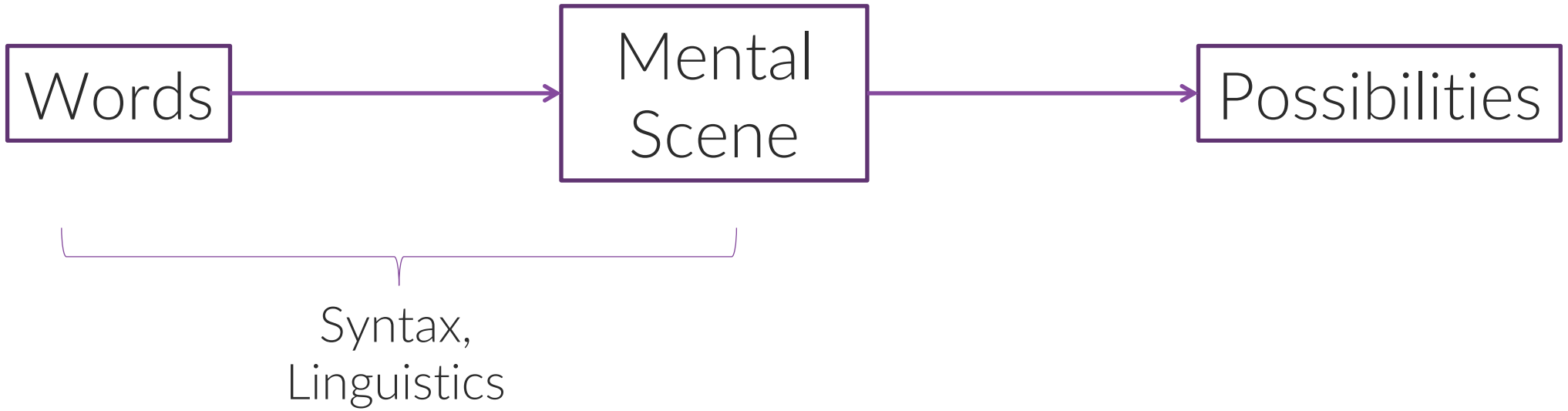
Reminiscent of Robert Brandom through Richard Evans
https://www.doc.ic.ac.uk/~re14/Evans-R-2020-PhD-Thesis.pdf

Two ways to not understand in a conversation:

1. Wrong mental scene
2. Not knowing the possibilities

Overview of "meaning"

Reminiscent of Robert Brandom through Richard Evans
https://www.doc.ic.ac.uk/~re14/Evans-R-2020-PhD-Thesis.pdf



Tip-of-the-linguistics iceberg for making a mental scene

Bob goes to the store.

Linguistics: tense, aspect, mood
https://en.wikipedia.org/wiki/Tense-aspect-mood

- Tense: action in time, past or future
- Aspect: duration of action
- Mood: whether it has happened or not, subjunctive

Tense
Bob went to the store
Bob is going to the store
Bob will go to the store

Aspect
Bob was going to go to the store … when it happened
Bob used to go to the store

Mood
Bob can go to the store
If Bob were to go the store … he would see the apples

Even more
Bob has gone to the store
Bob had gone to the store
I believe Bob is going to the store
I'm angry that Bob is going to the store

Tip-of-the-linguistics iceberg for making a mental scene

Bob goes to the store.

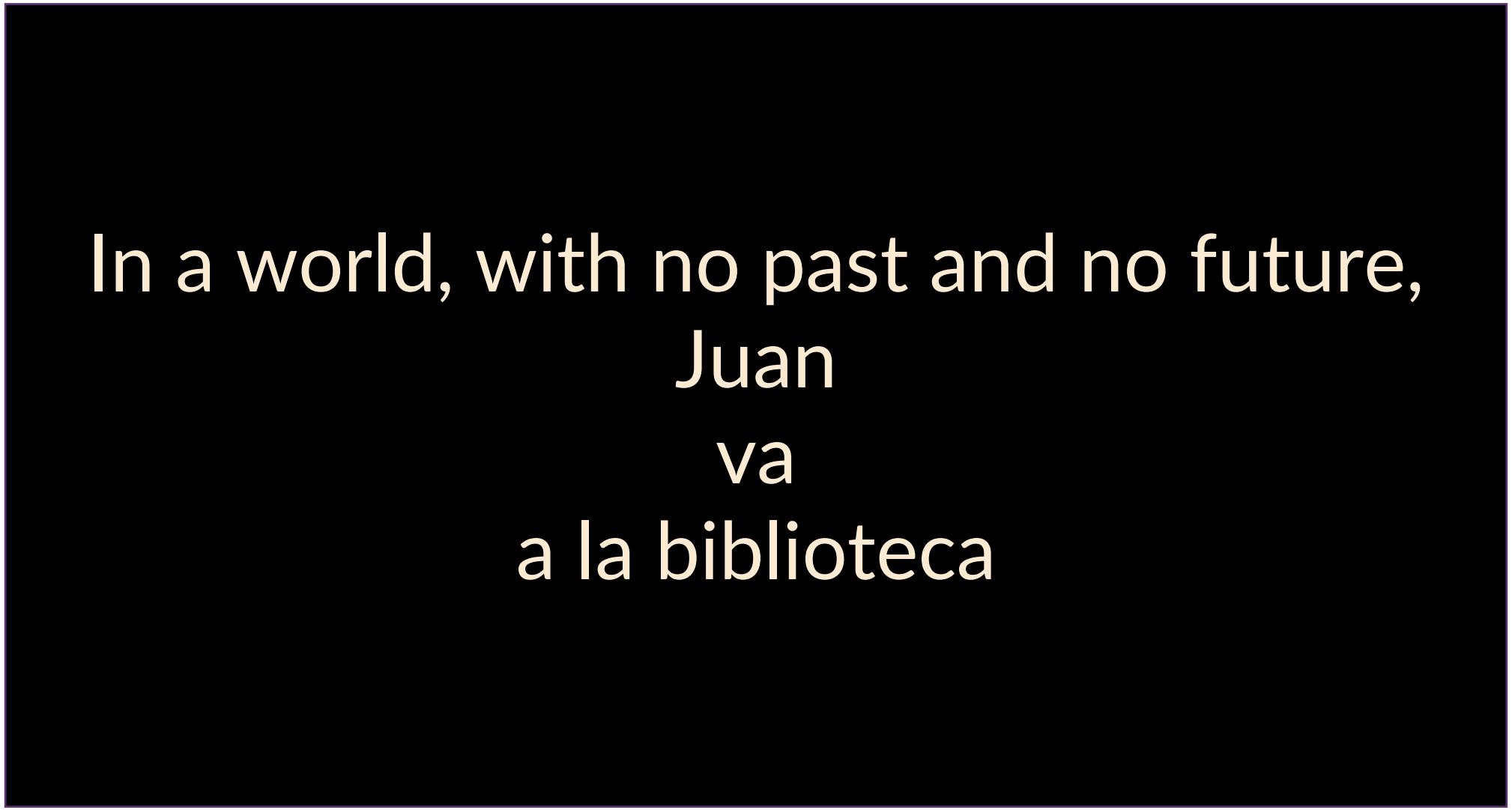
Linguistics: tense, aspect, mood
https://en.wikipedia.org/wiki/Tense-aspect-mood

- Tense: action in time, past or future
- Aspect: duration of action
- Mood: whether it has happened or not, subjunctive

Each of these versions evokes a slightly different mental scene.

You don't realize this until you study a foreign language, usually in year 2.

An idea that always makes me chuckle: if high school Spanish 1 had a movie trailer, it might go like this



But it is more than words that evokes the mental scene: conversation has its own rules (pragmatics)

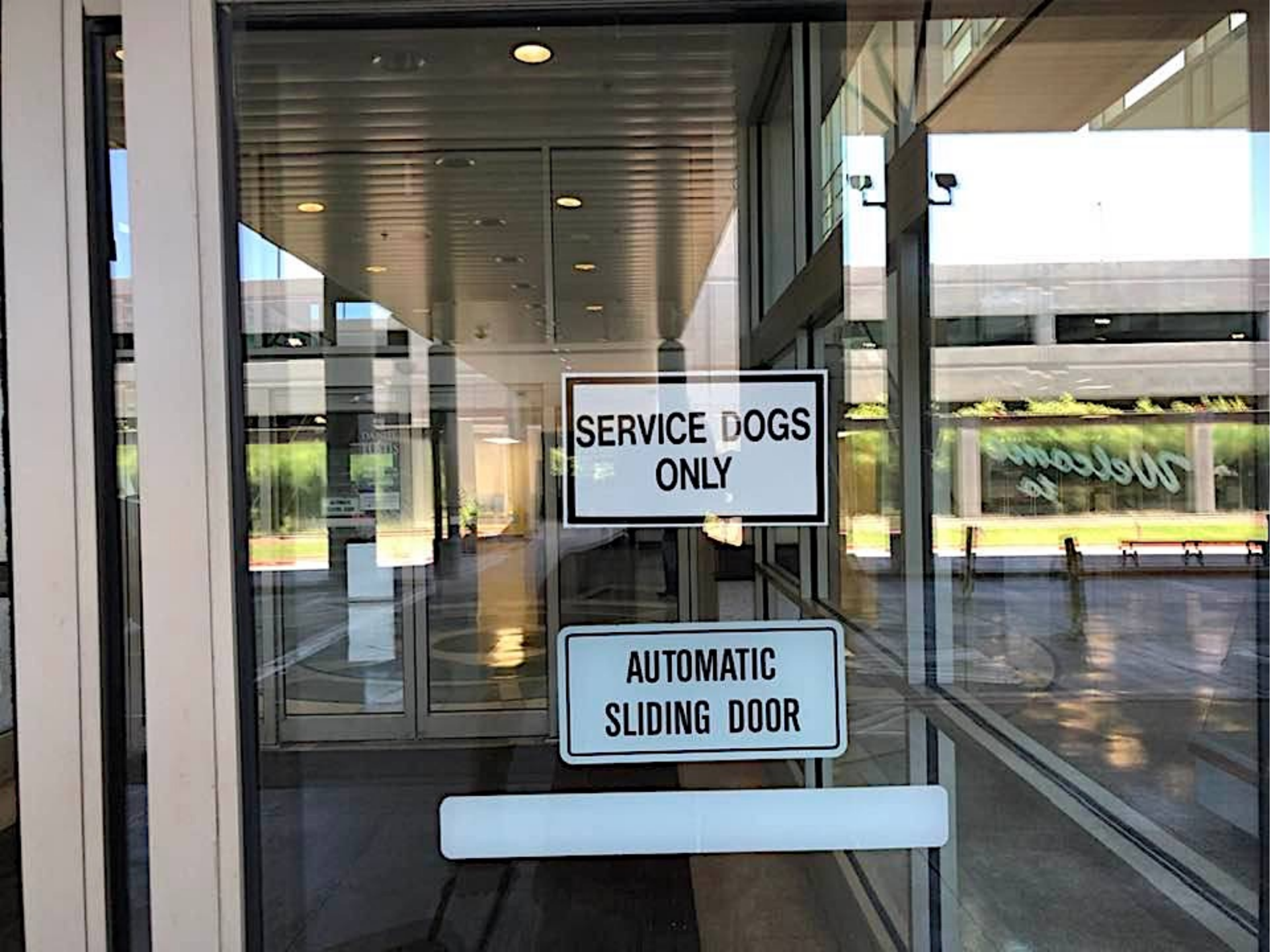
- Conversational maxims: Grice (1975, 1978)
- Breaking these rules is a way to communicate more than the meaning of the words.

Maxim of Quantity: Say only what is not implied.
Yes: "Bring me the table."
No: "Bring me the table by transporting it to my location."
What did she mean by that?

Maxim of Quality: Say only things that are true.
Yes: "I hate carrying tables."
No: "I love carrying tables, especially when they are covered in fire ants."
She must be being sarcastic.

Maxim of Relevance: Say only things that matter.
Yes: "Bring me the table."
No: "Bring me the table and birds sing."
What did she mean by that?

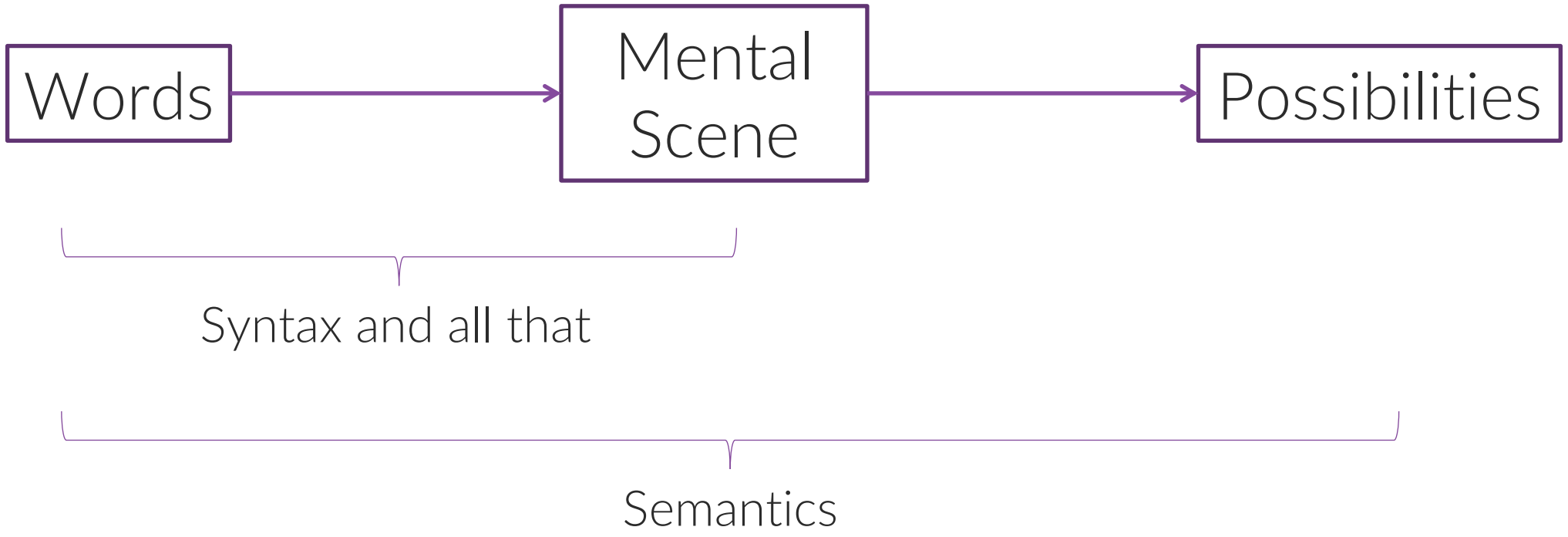
Maxim of Manner: Speak in a way that can be understood.
Yes: "Bring me the table."
No: "Use personal physical force to levitate the table and transport it to me."
What did she mean by that?



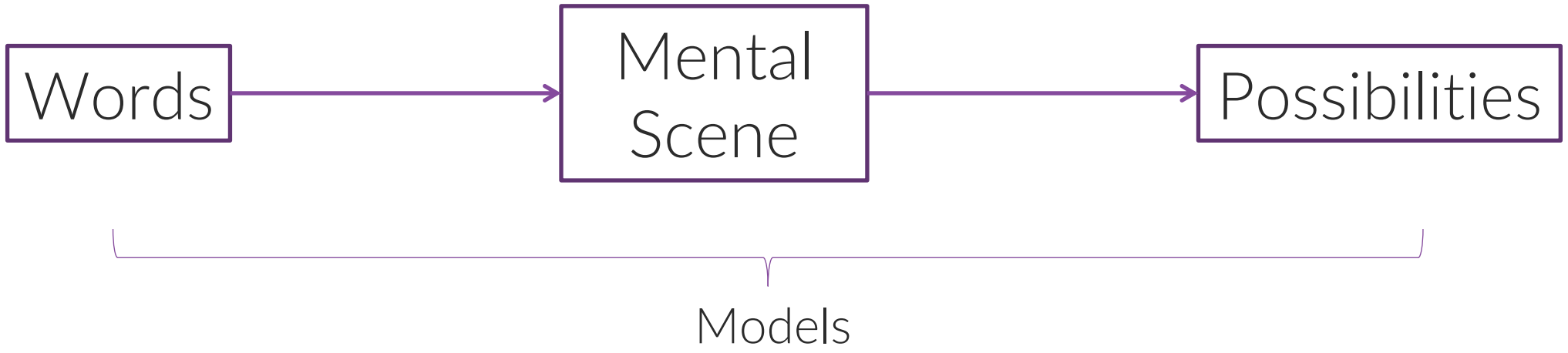
Words are only hints at possible meanings

When I saw this, my first thought was, “Where do people enter?”

Overview of “meaning”

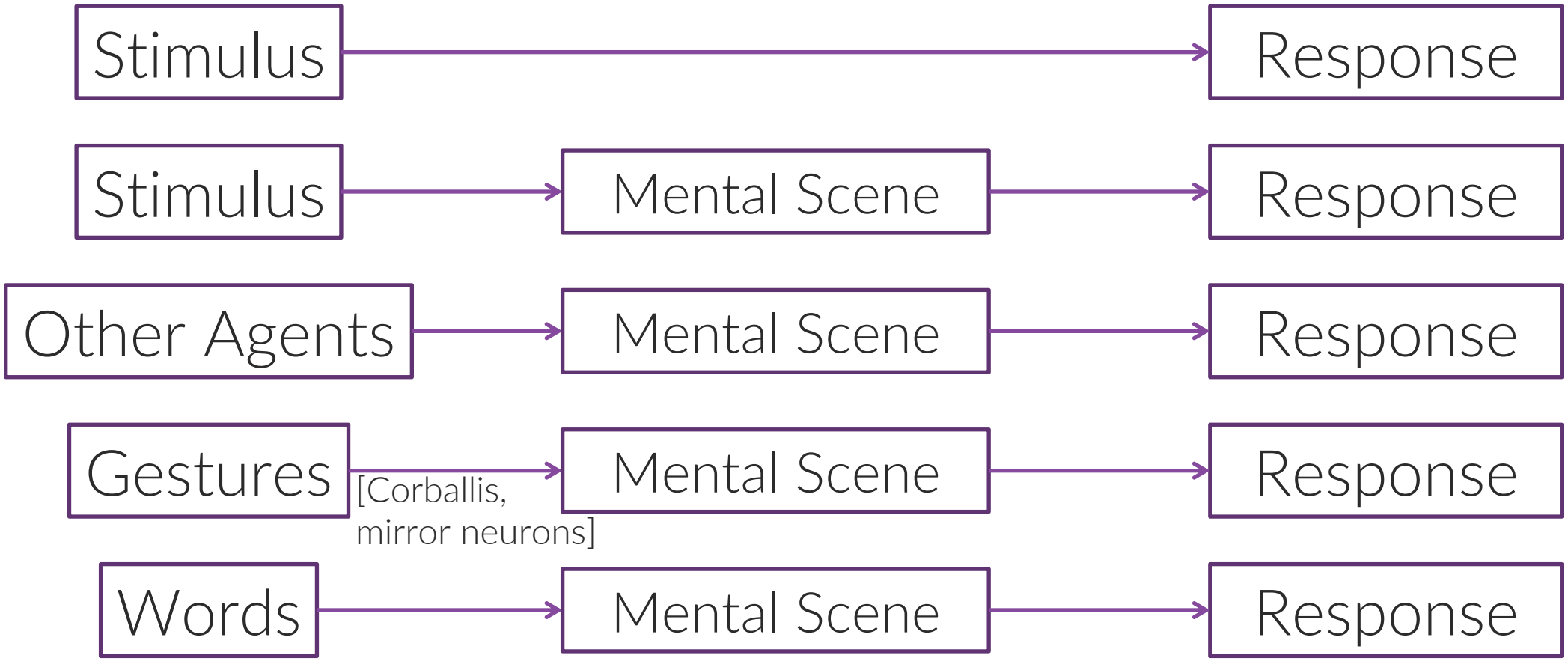


Overview of "meaning"



- Models enable meaning by building the mental scene and creating possibilities.
- Current deep learning methods interpolate from large datasets. Seems like sophisticated stimulus-response; is that sufficient?

How language may have come about: stimulus-response to flexible intelligence through models



Words are another kind of stimulus—they are only hints at possible meanings.

Outline

- Why we want AI
- Recent big-compute methods have been surprisingly good
- We still need meaning
- How to get there
- A pseudocode of consciousness

Illustration of Progress

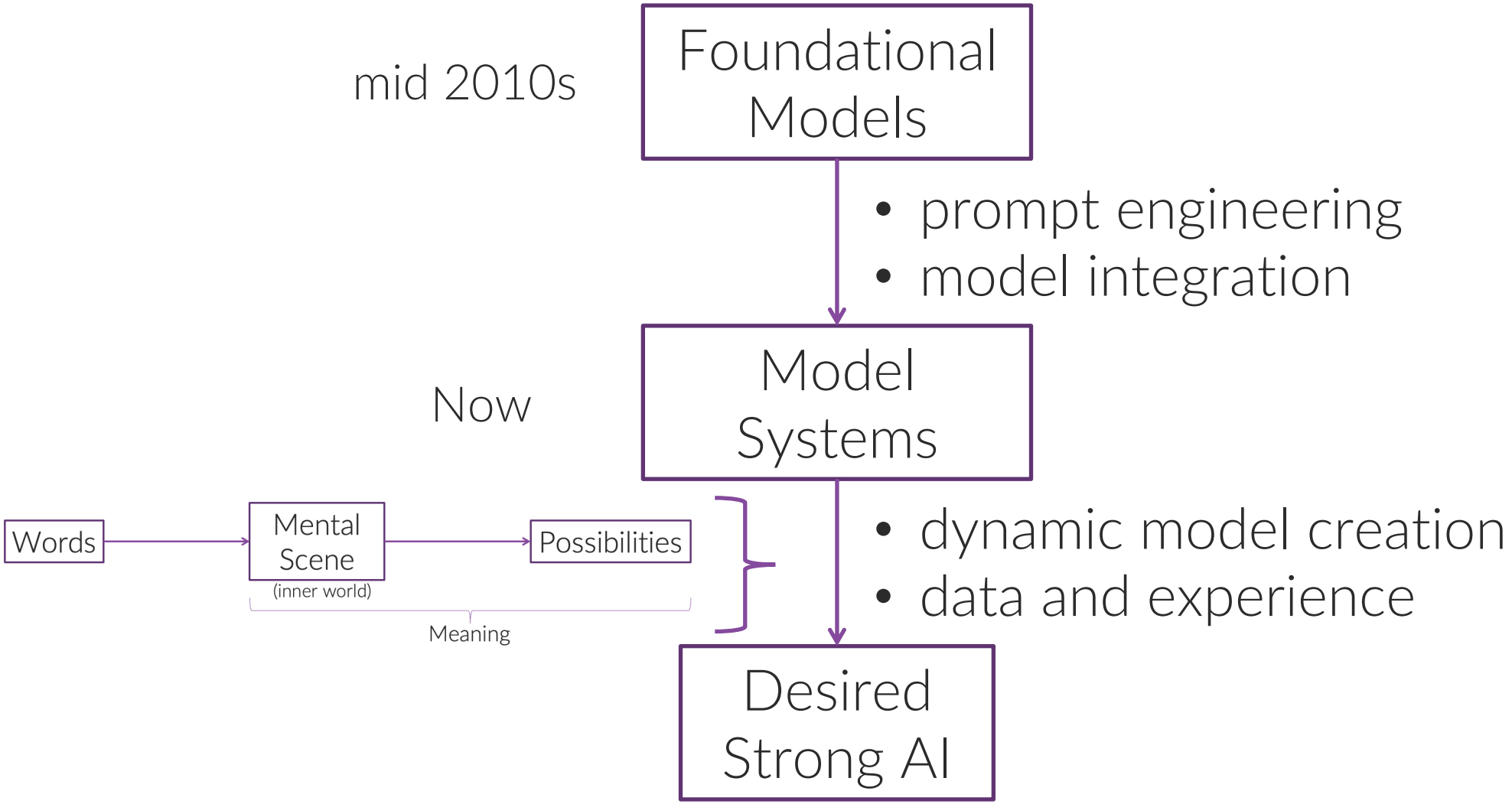
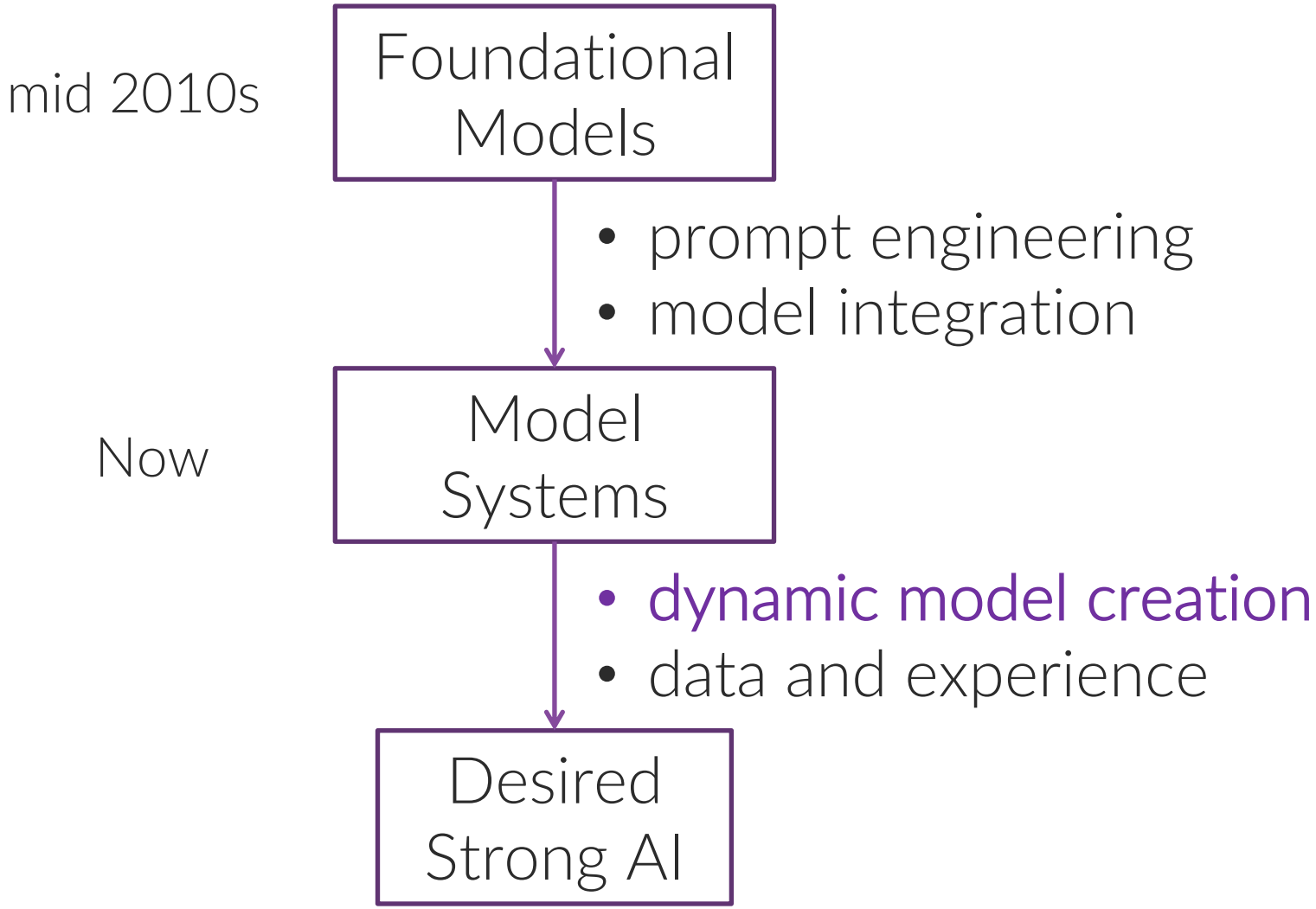


Illustration of Progress



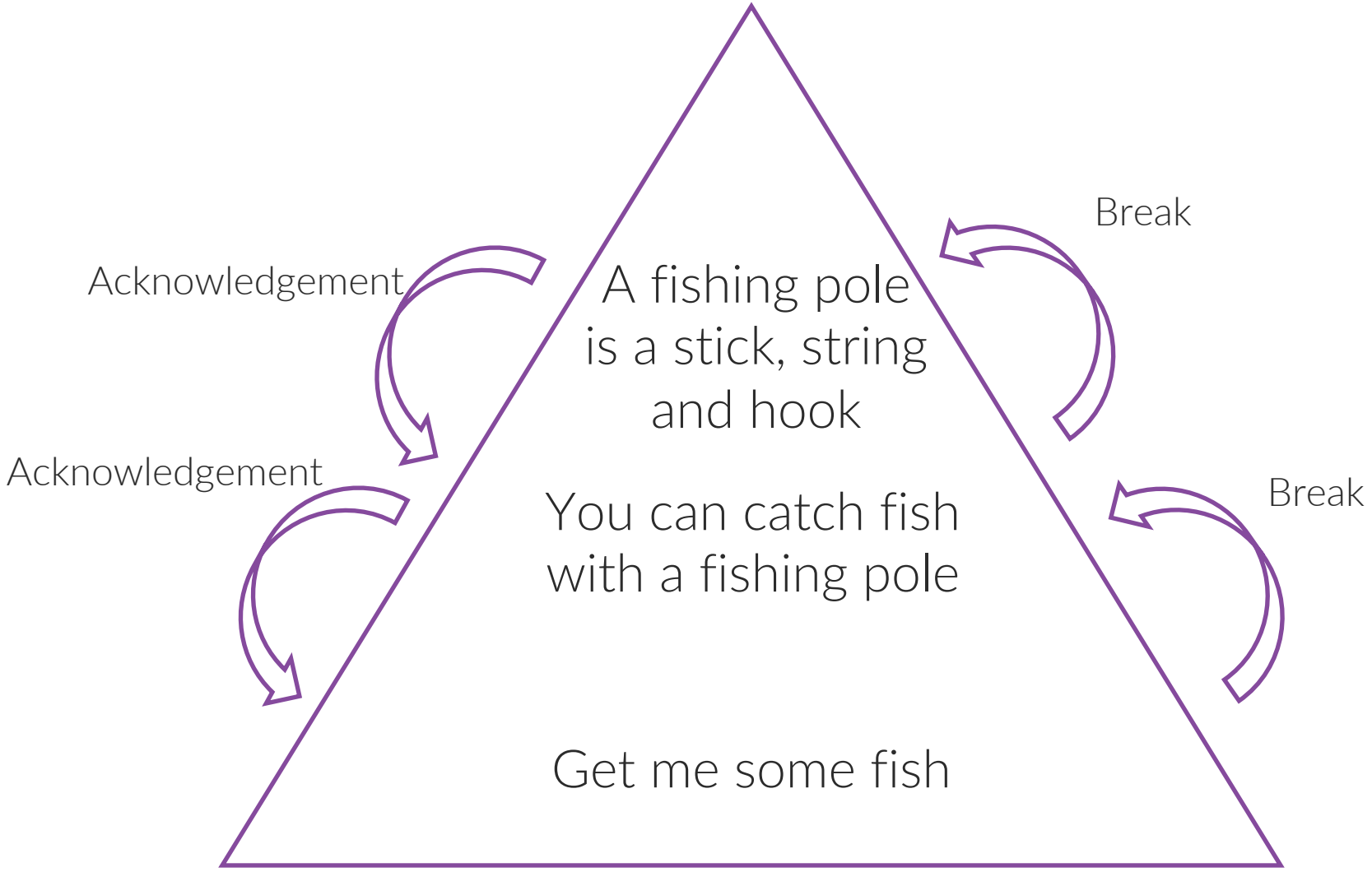
Models provide possibilities and set up the mental stage

A model says what a table is + the possibilities

"The table is on the table."

- If you want to pick up top table, you must first walk to the table
- If you push the bottom table, the top table will fall
- Party guests will think this is weird looking.

Models must be constructed dynamically to enable conversations and discovery



Modified from Gärdenfors (2014), which was based on Winter (1998)

Models must be constructed dynamically to enable conversations and discovery

Consider this derivative of an old joke*:

Your ego is so big that Thanos had to snap twice.

In the movies, Thanos is a Malthusian concerned with sustainability. He works to secure technology to reduce by half the number of persons in the universe, which is triggered by him snapping his fingers.

We think of the technology as removing souls, but this joke focuses on the mass of souls removed, and the derivative extends it with ego consisting of mass. To understand the joke, you must dynamically change your model.

* Best references I could find to the original https://www.youtube.com/watch?v=ztkB__b6zBM&ab_channel=PiyushPatel

In addition to dynamic creation, to bring models closer to covering the complexity of the world, models must be

Causal so they generalize better

- Imagine you know that you must water your peach trees in the summer.
- What if you have a spring with no rain?
- A causal model enables you to infer that you must water then too, even if you have no experience with a rainless spring.

Your model says that ~~Brawndo~~ water is what plants crave.

Rich and deep

- Rich models scale out horizontally to include more context. Leaves fall in autumn. When they fall, what happens to them? Do they accrue on the ground? Do they stay there forever? Why rake them?
- Deep models contain long chains of causal structure. Why do leaves fall in autumn? Why does it get cold? If it doesn't get cold one year, what will the leaves do? Why is cold related to leaves falling?

There is always a point below which we take it on faith. But at that point we no longer have the power of reasoning to expand our knowledge of world state.

Our causal, rich, and deep models need to compose for maximum coverage

Foundational Metaphors (See Mark Johnson and others. Steven Pinker talks about two main ones)

- Force
  - An offer or a person can be *attractive*
  - A broken air conditioner can *force* you to move a meeting
- Location in space
  - AI has come *a long way* in the last 15 years

Conceptual Blending (*The Way We Think* by Gilles Fauconnier and Mark Turner)

- That running back is a truck
- Dall-e 2 could generate a good picture of this, but it couldn't imagine what it is like to tackle a vehicle

Analogies (Melanie Mitchell and Douglas Hofstadter)

- We can broadly apply the story of sour grapes (Hofstadter in *Surfaces and Essences*)

Note that these often require bodily experience, as we will discuss soon.

What do these models look like, and what can they do?

Consider logic:

Can be simple like `taco(x) -> good(x)`

Powerful. You can infer that an item is good without being told it. You also know that if it's bad, it ain't a taco. Inference is a lever for information.

Can be more sophisticated

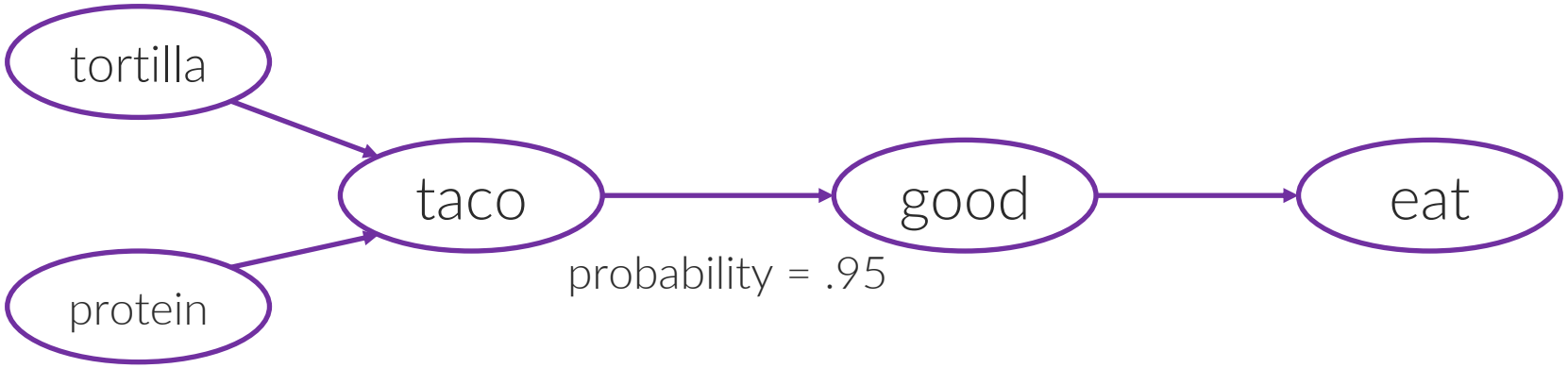
```
(=>
    (instance ?MAKING Making)
    (exits (?ARTIFACT)
        (and
            (instance ?ARTIFACT Artifact)
            (result ?MAKING ?ARTIFACT))))
```

If you make something, the result is an artifact, and that artifact is the result of the making process.

From SUMO: Suggested Upper Merged Ontology
http://www.adampease.org/OP/

Gets really sophisticated with answer-set programming and blazing fast SAT solvers.

Bayesian networks (and other graphical models)



We may have to consider the possibility of a bad taco.

In graphical models like this, your set of variables is fixed.

Probabilistic Programming

Probabilistic programing can create probabilistic models, such as Bayesian networks, dynamically.

Like setting up a computation graph in deep learning using Python.

Best resource: https://probmods.org/
Also, Pyro: https://pyro.ai/

Sets up model creation as a search over programs.

We do have a way to automatically write programs (Open AI Codex and GitHub Copilot) besides brute-force search, keep an eye out for clever ways to tie foundational models into model systems that create models.

```python
import torch
import pyro
import pyro.distributions as dist
import pyro.poutine as poutine
from pyro.infer import MCMC, NUTS

def model(prior_elves):
    num_elves = pyro.sample("num_elves",
                dist.Normal(prior_elves, torch.tensor(2.0)))
    num_rocks = num_elves * 4
    num_logs = num_elves * 6
    _rocks_observed = pyro.sample("rocks_observed",
                        dist.Normal(num_rocks, torch.tensor(3.0)))
    _logs_observed = pyro.sample("logs_observed",
                        dist.Normal(num_logs, torch.tensor(3.0)))
    return num_elves
def conditioned_model(model, data, prior_elves):
    return poutine.condition(model, data=data)(prior_elves)
data = {"rocks_observed": torch.tensor(4),
    "logs_observed": torch.tensor(6),
    }
prior_elves = torch.tensor([.2,.2,.2,.2,.2])
nuts_kernel = NUTS(conditioned_model, jit_compile=False)
mcmc = MCMC(nuts_kernel,
                num_samples=10,
                warmup_steps=5,
                num_chains=1)
mcmc.run(model, data, prior_elves)
mcmc.summary(prob=.5)
```

Probabilistic programming example in Pyro

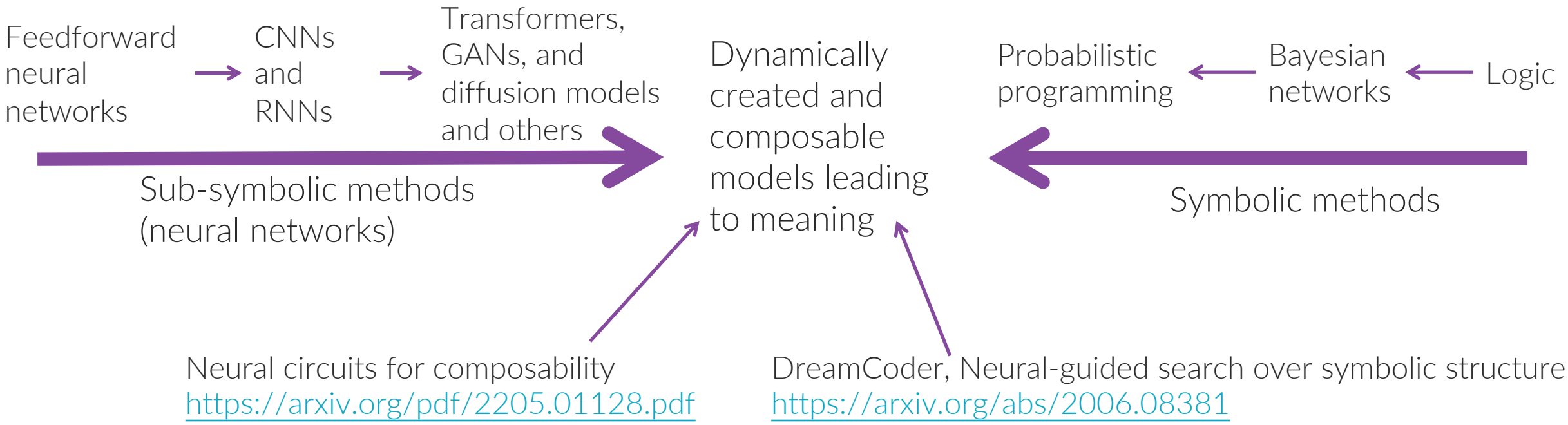
Imagine we want to know how many elves there are in a forest and how many modifications to the forest they have made.

There are two modifications an elf can make: painting a rock or carving a log. Each day, an elf can paint rocks or carve logs. We can only partially observe the forest.

- If we see a lot of elves, we can infer that there will be a lot painted rocks or carved logs or both.
- If we see a lot of painted rocks, we can infer there are lots of elves.
- If we are certain of the number of elves, then seeing a lot of painted rocks means there are fewer carved logs.

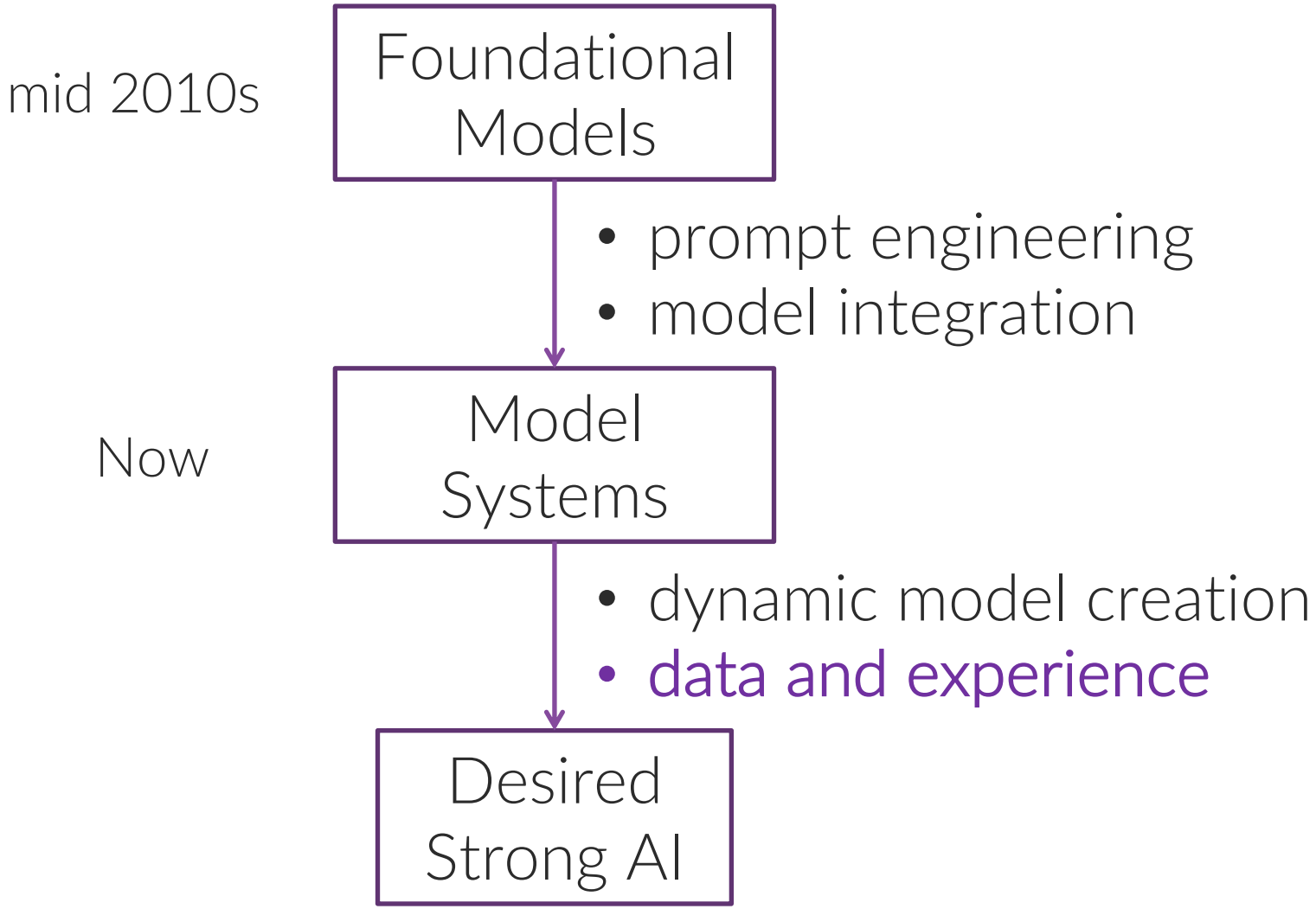
Closing the Gap on Dynamic Model Creation

- Neural networks are good at starting from the "pixel level" and fitting data
- Symbolic methods provide powerful inference, but they are largely built by hand, and we don't have a good way to automatically do metaphors, blending, and analogies



Still an open question how we will get here, but we need the data we get as children.

Illustration of Progress



AI needs the experience children get

Foundational models learn from text, but too many basic things aren't written down, because everyone knows them

Some of you may recall the romance novel I'm writing:

He pushed the table between them aside to embrace her, and all the objects on the table moved as well, and the liquid within the glasses spilled, and some things fell to the ground, but because it was carpet it wasn't as loud as it would be if it was a hard floor, and the table had friction with the floor, so it didn't fly through the window and into the street, and …

Words are only hints at possible meanings; to understand those hints we need experience in addition to text, even videos may be insufficient.

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You start off in a place and see the cup

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You pick up the cup

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You look up

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You rotate left

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You move left

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You move forward

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You open the microwave

One environment is AI2Thor by Allen AI

https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You place the cup in

One environment is AI2Thor by Allen AI

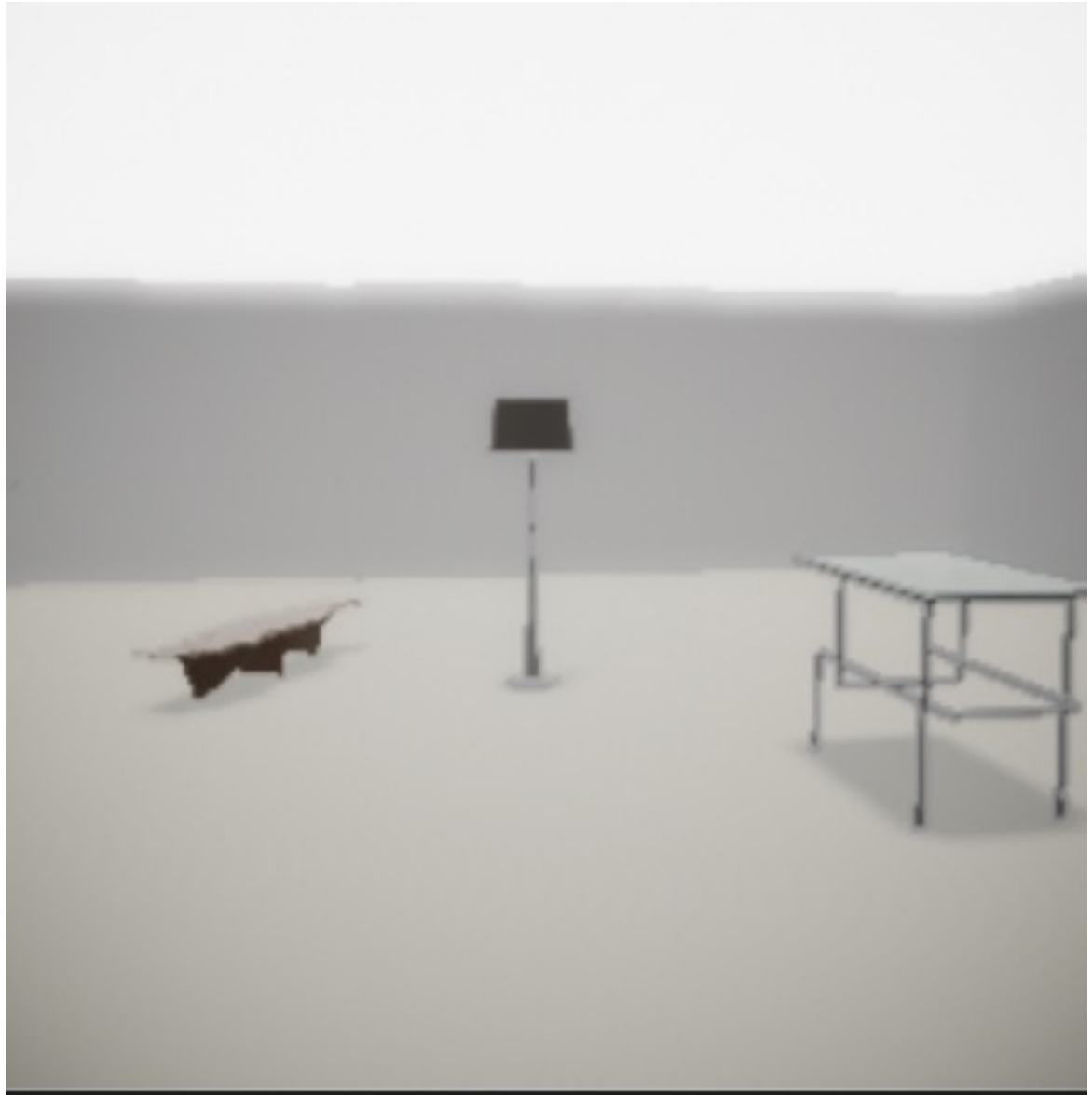
https://ai2thor.allenai.org/

Thor visible room rearrangement challenge
https://ai2thor.allenai.org/rearrangement/

This example from (they have since killed the link)
https://ai2thor.allenai.org/ithor/documentation/overview/examples/

You close the microwave

ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation

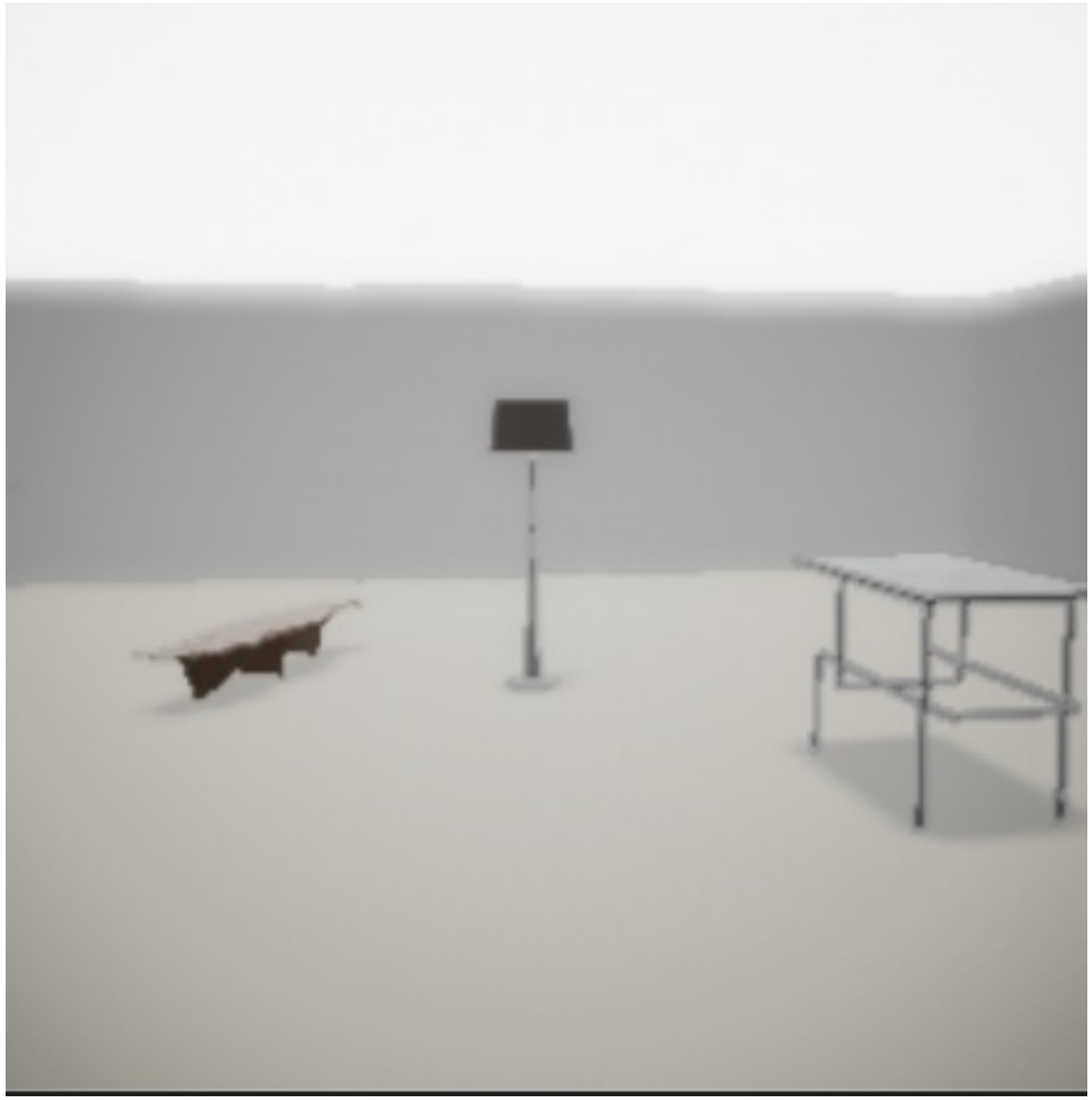


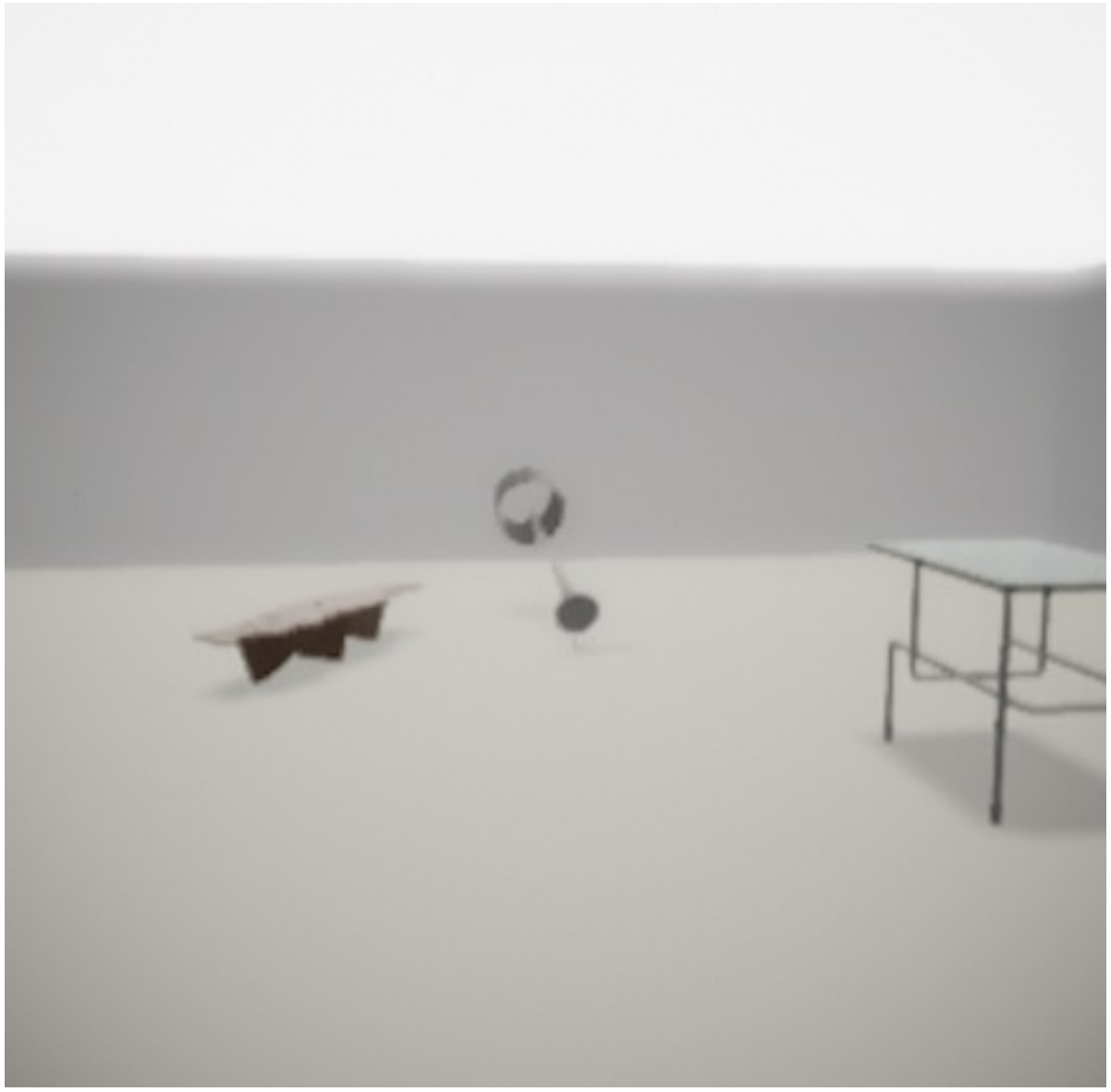
An engine where you can do 3d simulations.

https://arxiv.org/pdf/2007.04954.pdf
http://www.threedworld.org/
https://github.com/threedworld-mit/tdw

Image of a lamp falling over is generated when you run this (they have since killed the link)
https://github.com/threedworld-mit/tdw/blob/master/Python/example_controllers/objects_and_images.py

ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation





Toward simulations of our world

- I'm excited for these worlds to become our personal world
- The beginnings of a mirror world (*Mirror Worlds*, David Gelernter, 1993)

Microsoft Flight Simulator
https://www.flightsimulator.com/

Imagine you could fly. You could approach downtown Austin from the east from the airport.

On my son's computer he got in 5th grade ("the potato")

Also see https://deumbra.com/2020/09/microsoft-flight-simulator-2020-is-an-inflection-point-for-virtual-worlds-and-our-own/

A digital assistant in your own personal mirror world

An assistant like Siri could live in your own personal mirror world.

Then it could have the context to understand your meaning by creating the correct mental scene.

If you go to the pharmacy, it goes with you, into a virtual pharmacy.

From digital assistant to conversation partners

What would it take to have a digital assistant become a conversational partner and maybe a friend?

If it were your friend, what would it say?

To make decent conversation, it would also have to have goals of its own. What would it want?

For life, the value function comes from homeostasis (Antonio Damasio).

Would it need to be conscious?

Outline

- Why we want AI
- Recent big-compute methods have been surprisingly good
- We still need meaning
- How to get there
- A pseudocode of consciousness

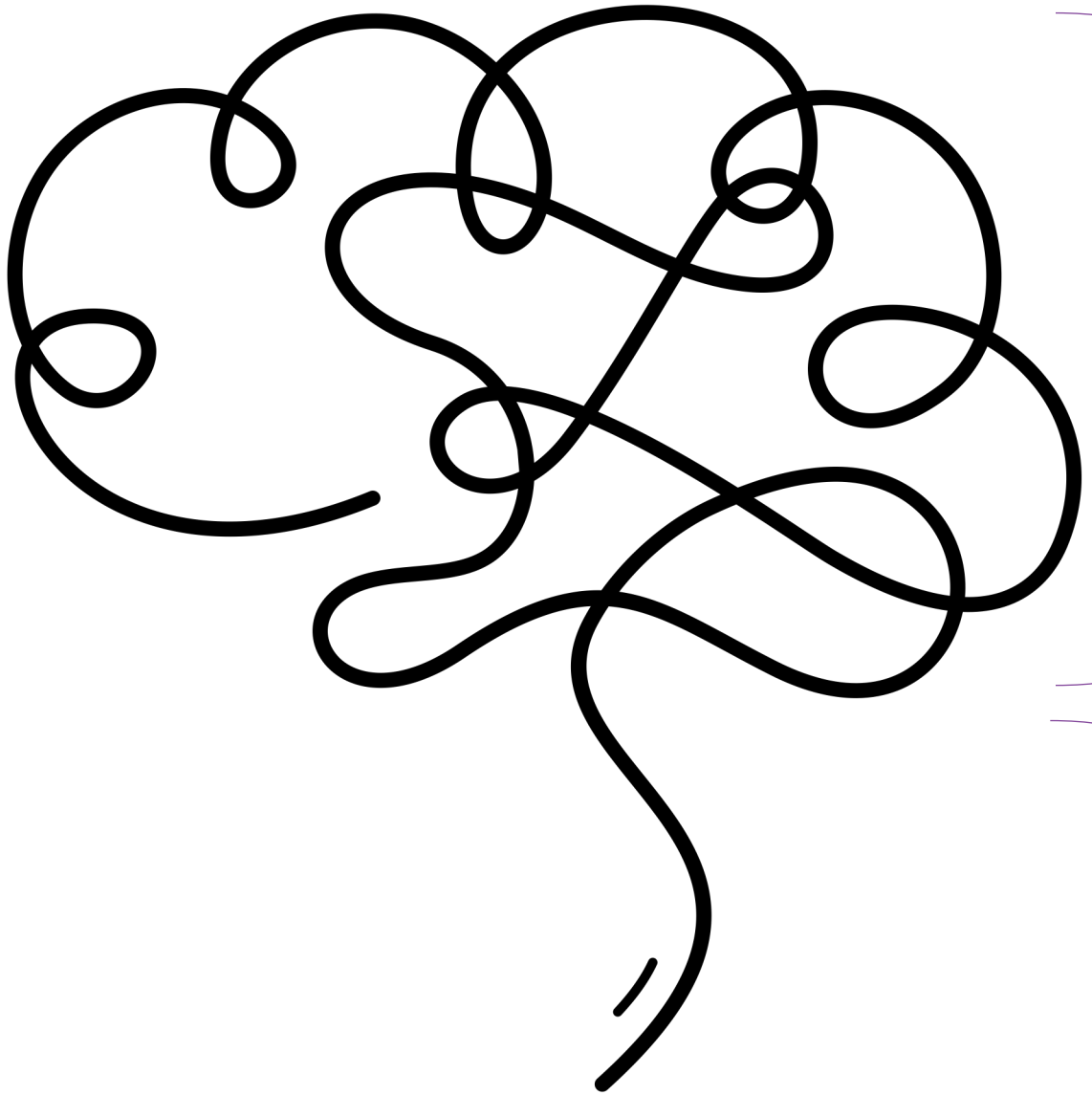
Lots of theories of consciousness

Global workspace theory of Baars, 1988. Consciousness is "fame in the brain" (Dennett). Important representations of the external environment need to be available to all parts of the brain.

Attention to internal states. *Rethinking Consciousness*, Graziano. It's more than having a global workspace—it is the internal focus of attention into that workspace (reminiscent of Hofstadter's strange loops). You are sitting there thinking about that time in third grade and …

Consciousness happens when there is a "breakdown," Heidegger. When you are hammering the hammer does not exist, unless there is a problem. Consciousness is debug mode, when you must stop and think, Ballard.
→ Developmental psychologist Allison Gopnik says that because adults spend so much time on autopilot that children are more conscious than adults. This is why it takes them so long to put their socks on.
→ Breakdowns may slow down perceived time, which is maybe why life speeds up as you get older but can slow down on vacation.

You feel your way through, Solms. Consciousness must "feel like" something (qualia) because you can't add a reduction in hunger to a reduction in cold; as a being maintaining homeostasis, you must handle each need one at a time because either of those would kill you if not sufficiently addressed.



Cortex

$T(s, a) \rightarrow s'$

Thinking, simulating (Hawkins)

drives and emotions that give problems to be solved; hunger, thirst

ideas and imagined states to be evaluated

Brain Stem

$V(s)$

Consciousness is here;
we feel our way through (Solms)

A pseudocode of consciousness

```
while alive:
  if there is a breakdown:
    1. the cortex populates mental state
    2. internal attention says where to focus
    3. non-interchangeable value function
       says what is good
  else:
    continue in auto zombie mode
```

Reinforcement learning notation

receive state s
if expected state $s_e \neq s$
$T(s, a) \rightarrow s'$
Attention$(s') \rightarrow \hat{s}$
$V(\hat{s})$ s.t. $V(s_1) + V(s_2) \neq V(s_1 + s_2)$
and choose policy π

$\pi(s) \rightarrow a$

According to this organization of ideas, if we can avoid building a non-interchangeable value function in a computer we can keep it from being conscious. Will save us a lot of trouble.

An idea that makes me chuckle: What if by some fluke virus, printers are conscious? Would explain why they are so recalcitrant.

Conclusion

Foundational Models

Model Systems

1. Current deep learning models appear to be interpolation, which allows for sophisticated stimulus-response agents
2. Moving beyond stimulus and response requires meaning
3. Meaning is mental scene + possibilities
4. Open question 1: Symbolic models can build mental scenes and produce possibilities; can they be made dynamically and robustly?
5. Open question 2: Can deep neural networks display sufficient composability?

To get here, we need meaning

Conclusion

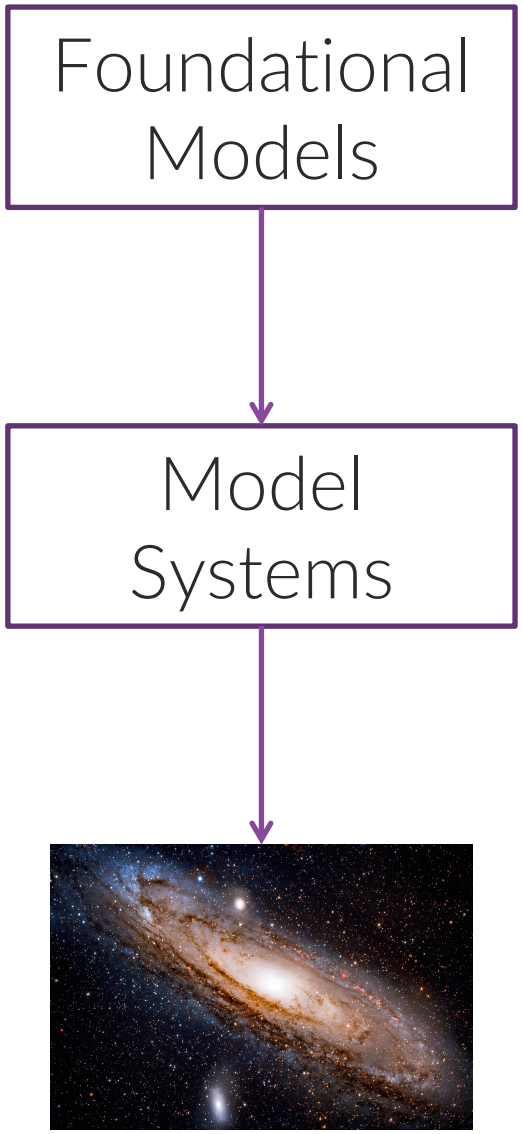
Foundational Models

Model Systems

We know we need world experience; will meaning come from neural networks or symbolic models?

neural networks ? meaning ? symbolic models

Conclusion



You'll know you are starting to get strong AI when

- you have conversations with computers where new meanings are established, and
- computers can create new knowledge and explain it to us.

See my article in The Gradient for more details
https://thegradient.pub/strong-ai-requires-autonomous-building-of-composable-models/

6500 River Place Blvd.
Bldg. 3, Suite 120
Austin, TX. 78730

Extra slides

Can AI tell us why this picture is funny?

Volleyball locker room at UT Austin.

https://www.flickr.com/photos/obamawhitehouse/4921383047/

Karpathy, 2012
http://karpathy.github.io/2012/10/22/state-of-computer-vision/

Yannic Kilcher discusses how Flamingo gets close
https://www.youtube.com/watch?v=smUHQndcmOY&t=152s&ab_channel=YannicKilcher

- You can't ask why it is funny, but you can ask "Where is Obama's foot positioned?"

Tweet by Florin Bulgarov
https://twitter.com/florinzf/status/1522633003511517187/photo/1

Robotic lab