The lacunae of language models in the neuroscience of language

Andrea E. Martin

Lise Meitner Research Group
Language and Computation in Neural Systems

Max Planck Institute for Psycholinguistics &
Donders Centre for Cognitive Neuroimaging, Radboud University

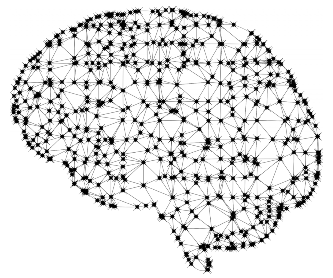
MAX-PLANCK-GESELLSCHAFT

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

NWO

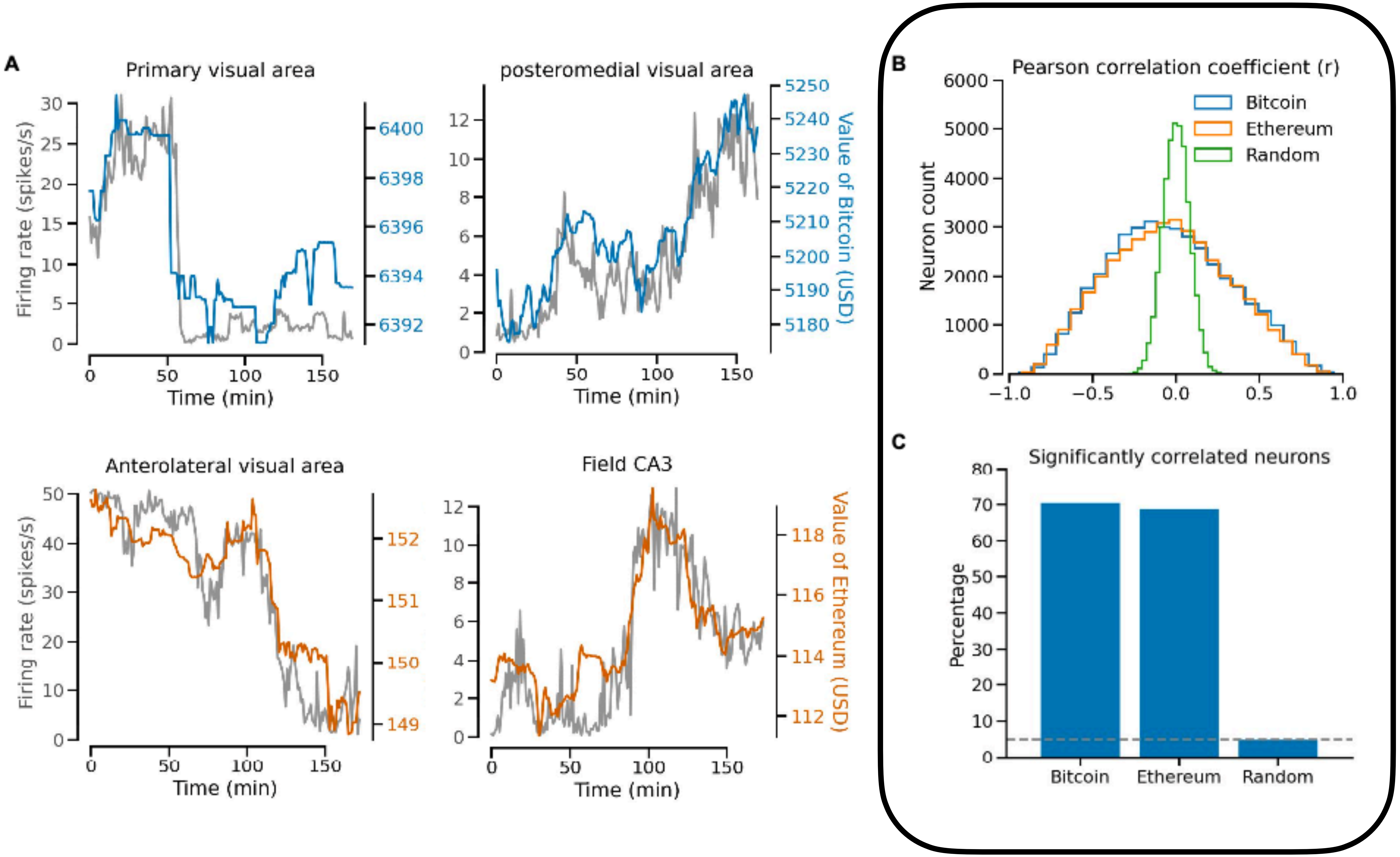
DONDERS INSTITUTE

LANGUAGE in INTERACTION



andrea.martin@mpi.nl www.andreaemartin.com lacns.github.io

Predicting neural data with… crypto



Neurons in the mouse brain correlate with cryptocurrency price

When correlating two signals evolve slowly over time, the chances of finding a significant correlation between the two are much higher than when comparing signals which lack this property.

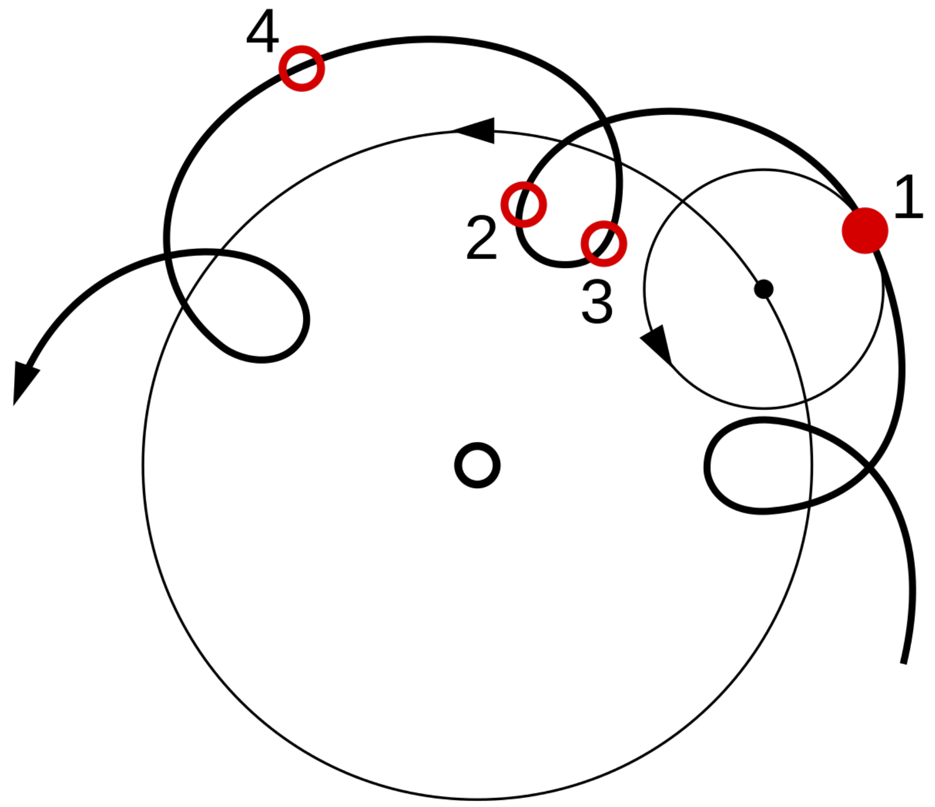
Meijer (2021)

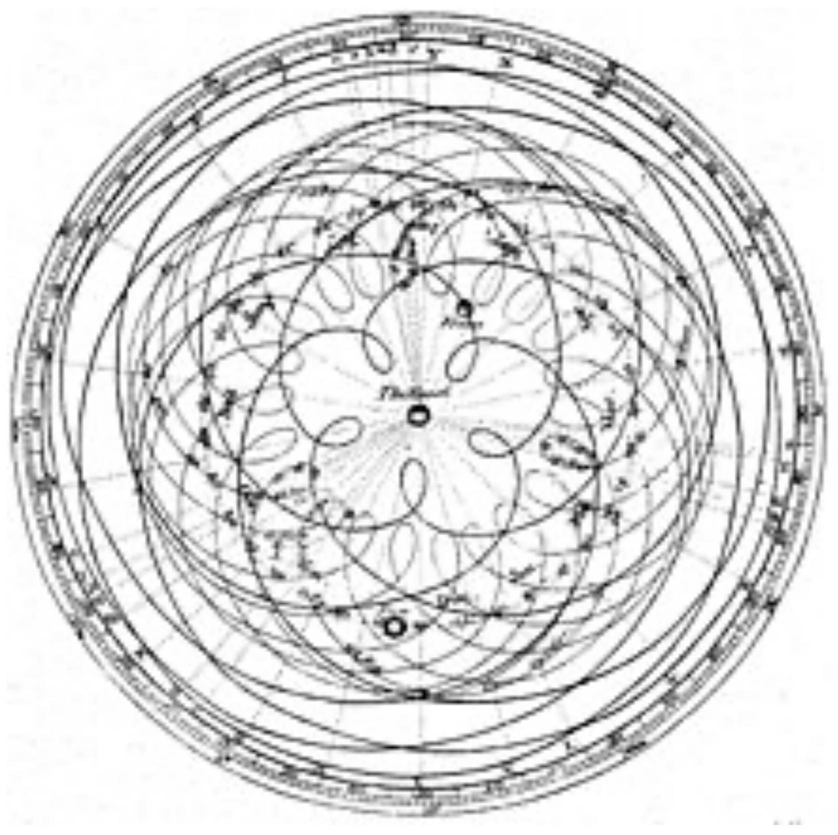
See also Harris, K. D. (2020). Nonsense correlations in neuroscience. *bioRxiv*,

Meijer (2021) *Peer Community Journal*

Why can models with few brain-like properties predict brain activity?

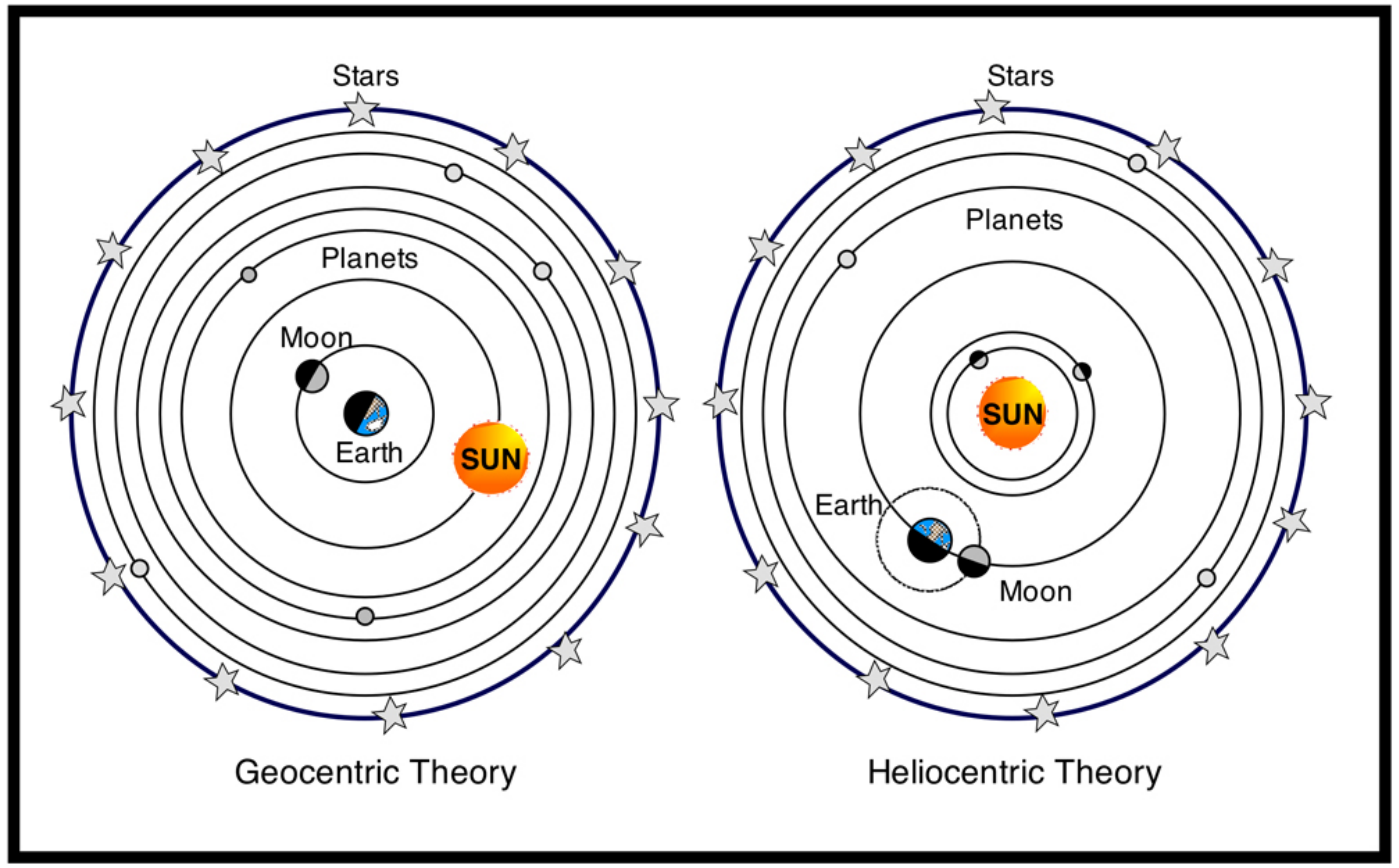
Epicycles



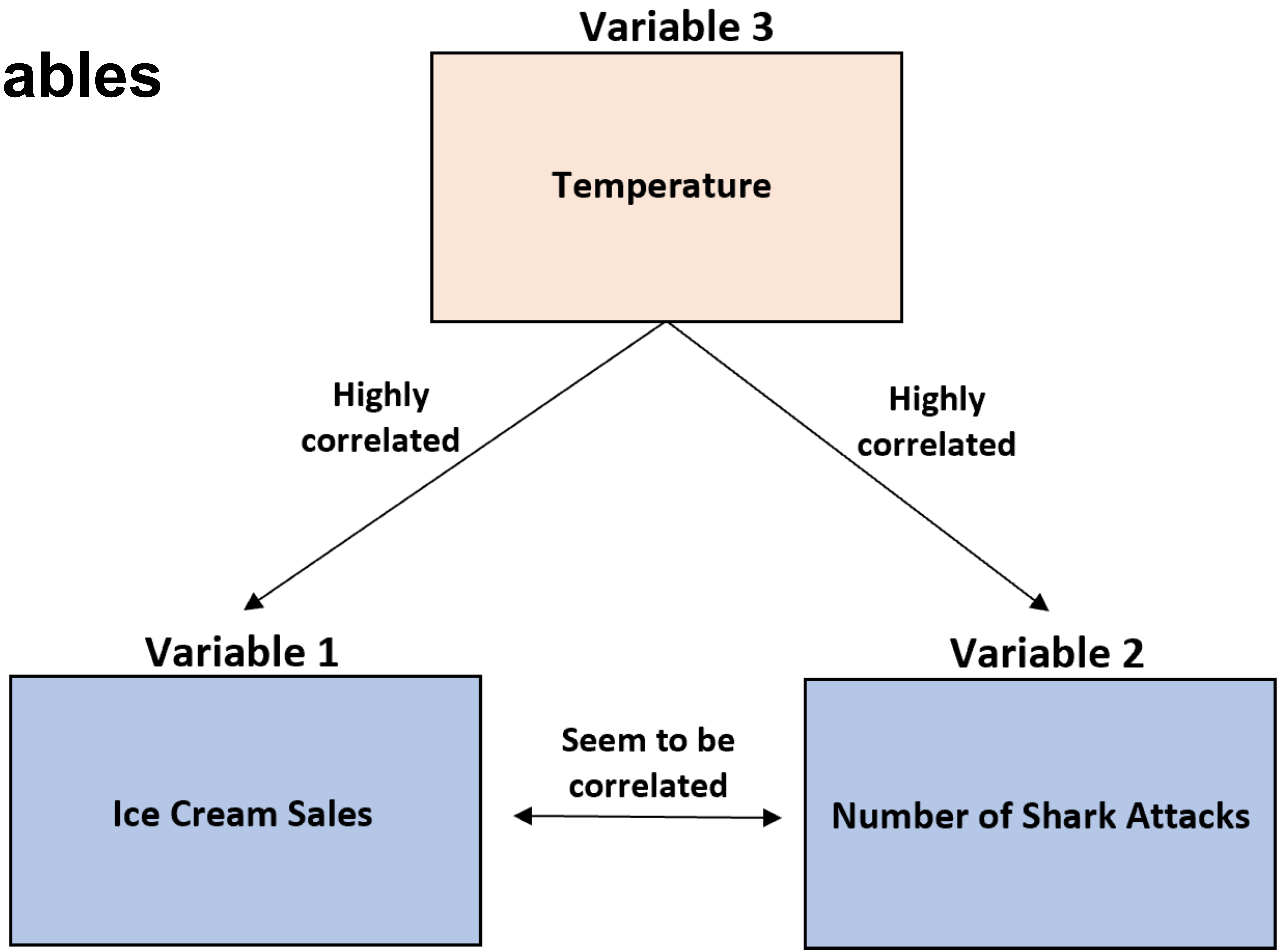


To explain the irregular motions of planets observed from Earth, astronomers introduced **epicycles**, which were smaller circles that planets were thought to move around as they orbited the Earth. The centers of these smaller circles, called epicycles, were themselves thought to move around the Earth along larger circles, called deferents. The combination of these two circular motions created the observed irregularities in the planet's motion.

…any smooth curve can be approximated to arbitrary accuracy with a sufficient number of epicycles…

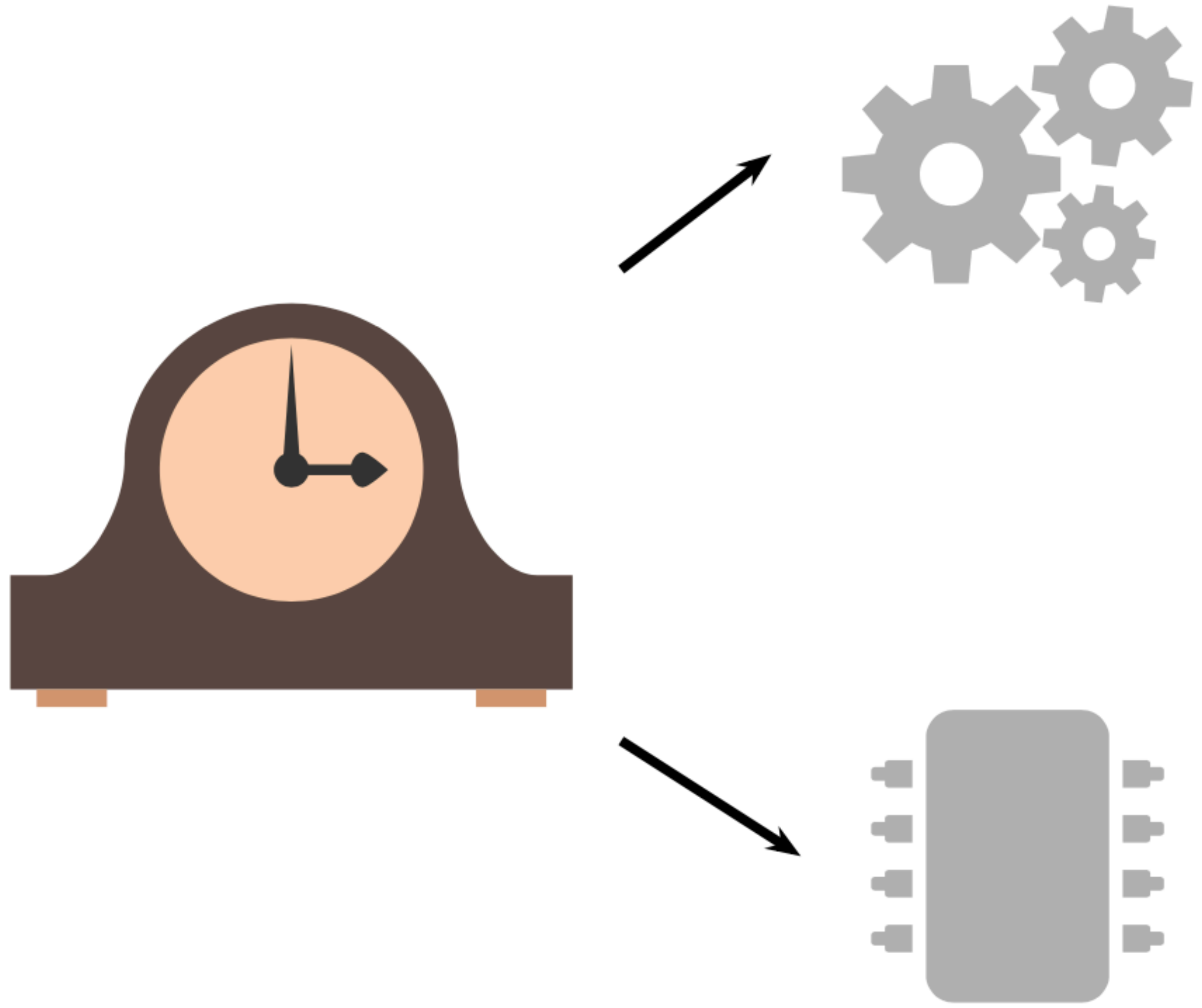


Third variables



statology.org

Language is more than large language models



the clockwork clock as the ‘real’ empirical clock and the digital clock as the “computational” model.

Digital clocks display the time.
Clockwork clocks display the time,
and require manual winding.
Therefore, digital clocks require manual winding.

Artificial neural networks (ANN) correlate with fMRI data.
Brains correlate with fMRI data, and instantiate the
biological mechanisms for cognition.
Therefore, ANNs instantiate the biological mechanisms
for cognition.

LLMs produce grammatical strings.
Humans produce grammatical strings, and instantiate
the biological mechanisms for language.
Therefore, LLMs instantiate the biological
mechanisms for language.

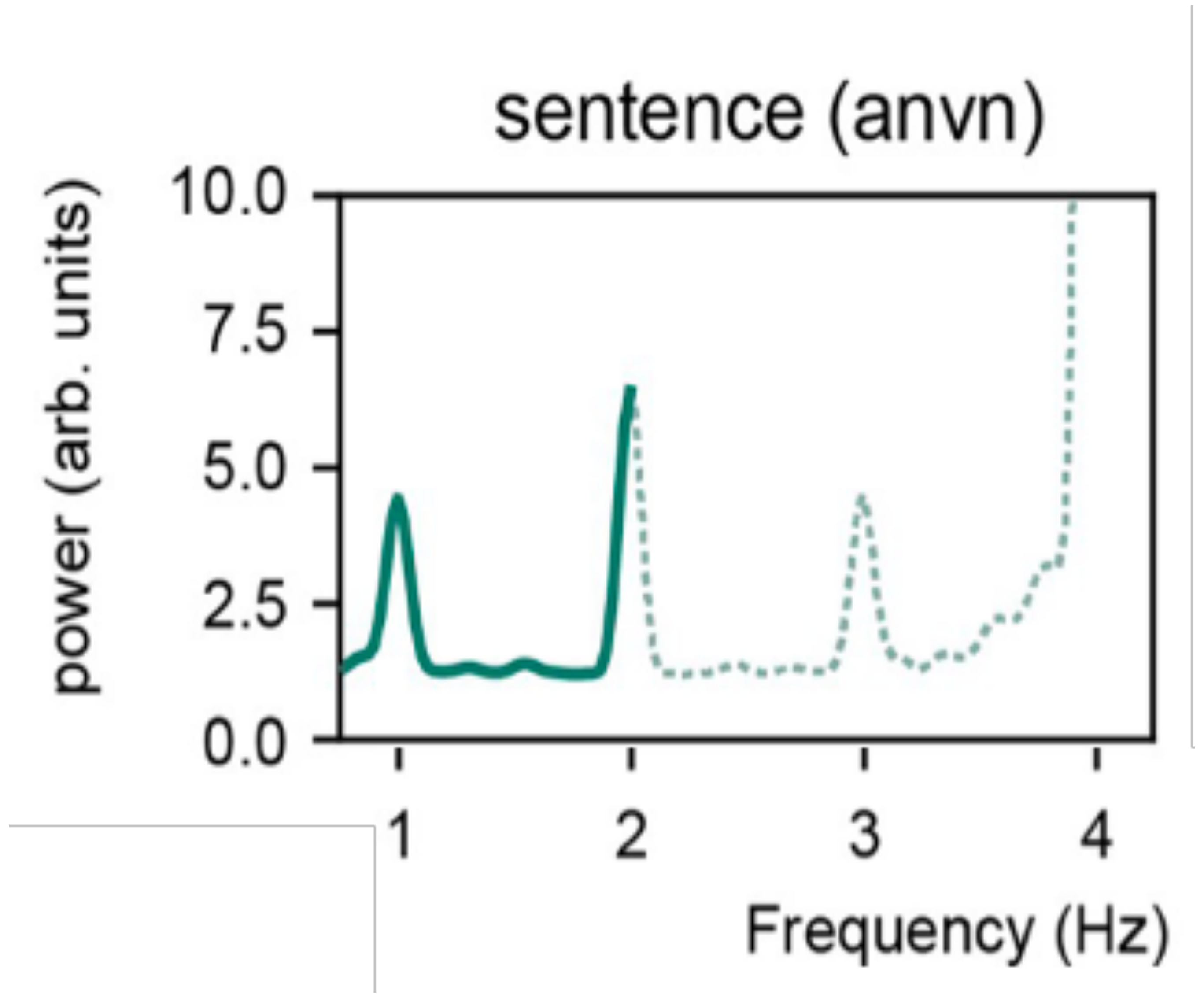
Olivia Guest

Guest & Martin (2023) *Computational Brain & Behavior*

A model can predict data, but may not be the implementation that the brain uses

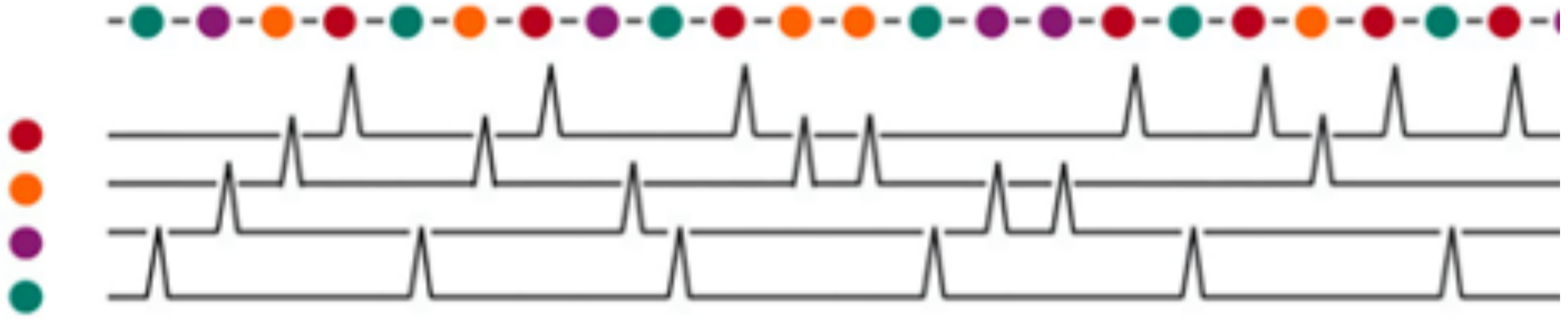
Sanne ten Oever

A neural network parser trained only on syntactic annotations shows the same pattern



Karthikeya Kaushik

Arbitrary sequences with a underlying sub-harmonic rhythms show the same pattern



ten Oever, Kaushik, & Martin (2022) *PLoS Comp Bio*

And yet it moves

- Metrics like BrainSCORE contain behavioral data – information that would already pick out voxels associated with language comprehension

- LLM have properties and assumptions that wildly violate the requirements of psychological and neurobiological cognitive models (e.g., parameters, training data needed, representational states, lack of time)

- LLM may share low dimensional states with networks in the brain processing language

- And so might a lot of other things…

- In a low enough dimensional space, all temporally-ordered statistical structures may correlate

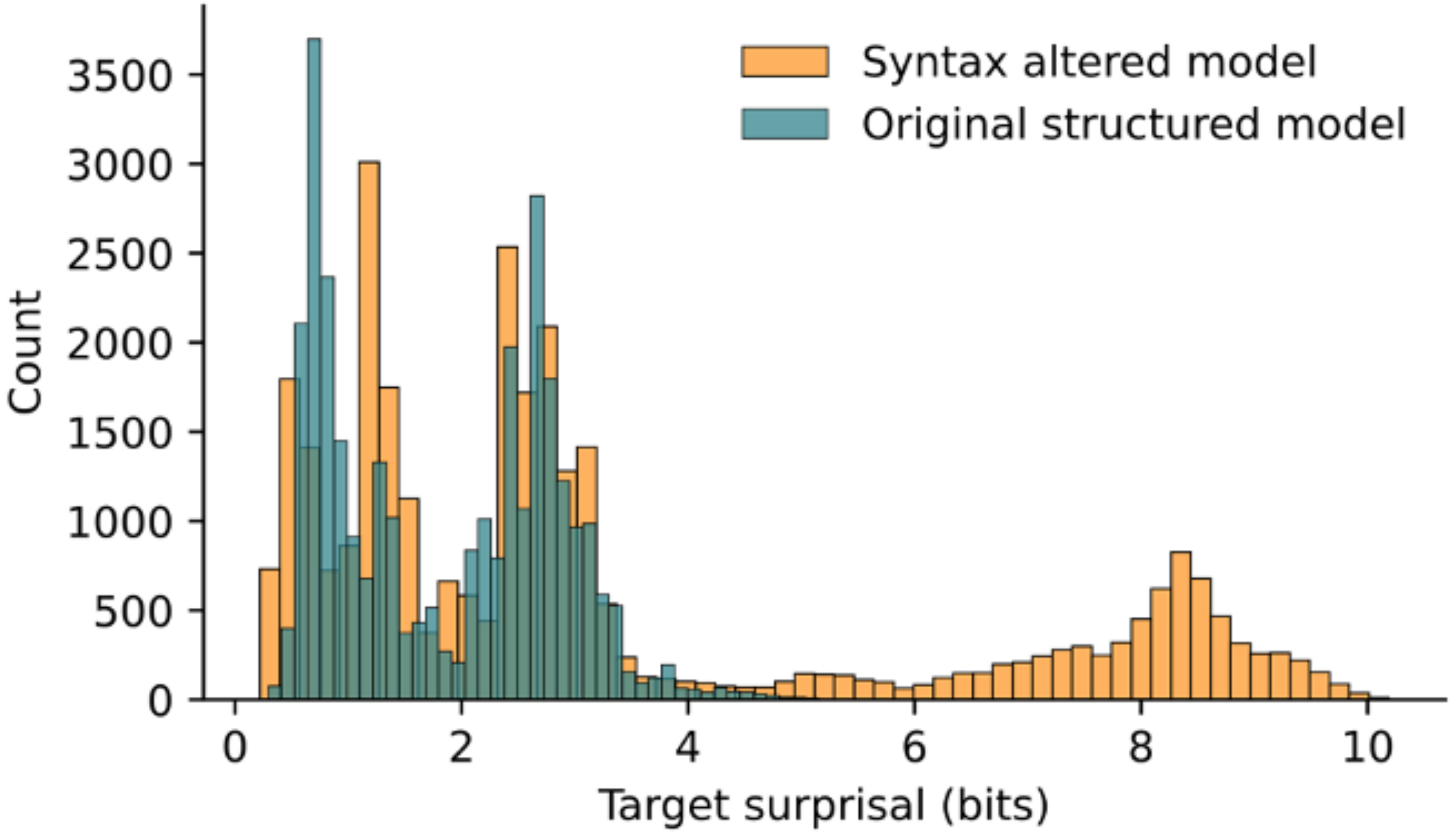
Can this ever be an adequate cognitive theory?

With care, a powerful tool for decomposing complex signals?
(still a filter/prism)

Surprisal is a mixture of information by design

D33

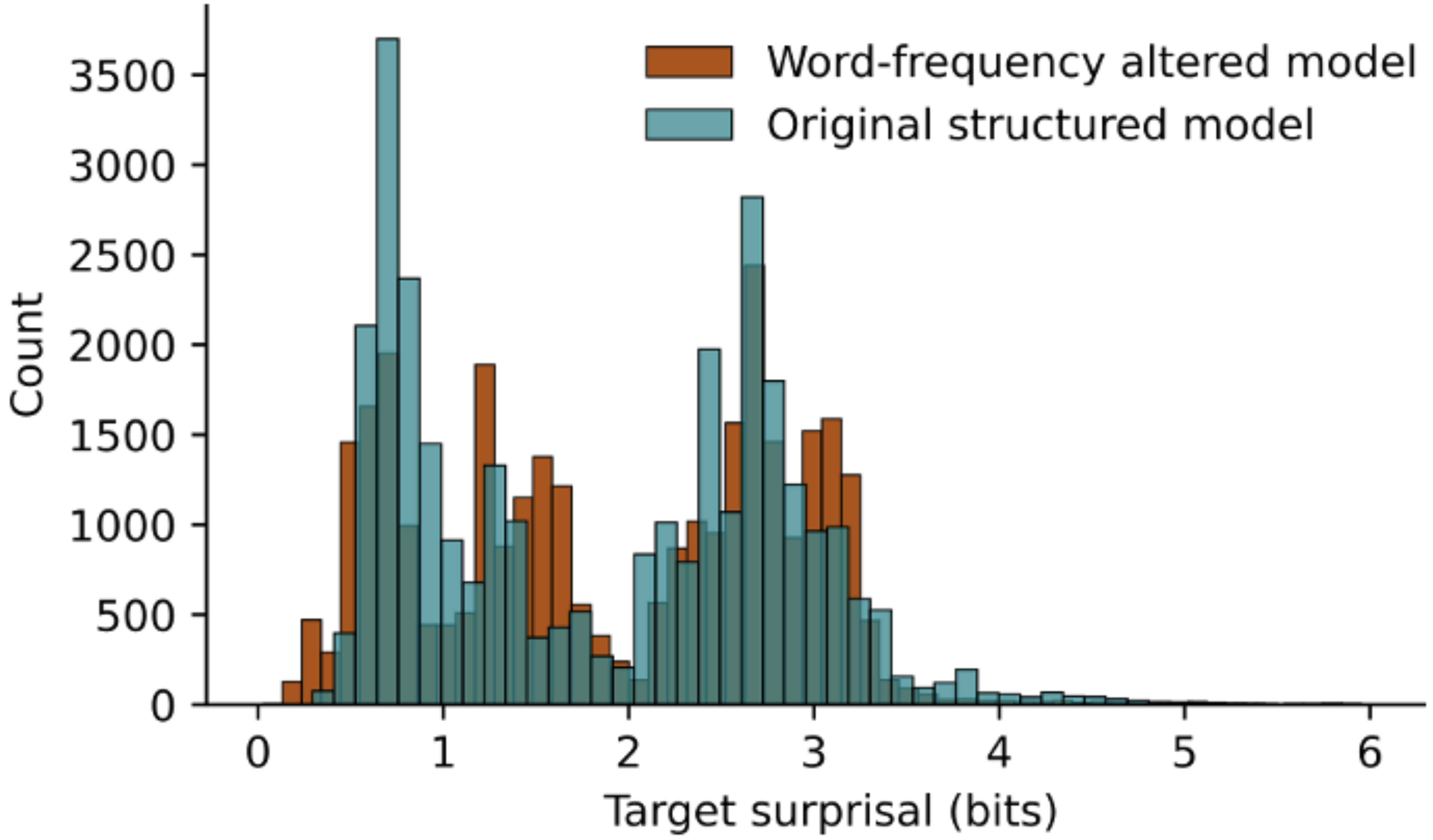
Sophie Slaats



Syntax changes surprisal values

An excellent predictor,
but not specific in mechanism
or explanation

Word frequency changes surprisal values



Slaats & Martin (2023) *PsyArXiv*

Reframing the problem

Linguistic representations, and statistics about them, matter and shape brain activity

This means that both LLMs and linguistic theory are incomplete

But - statistics are **about** linguistic representations

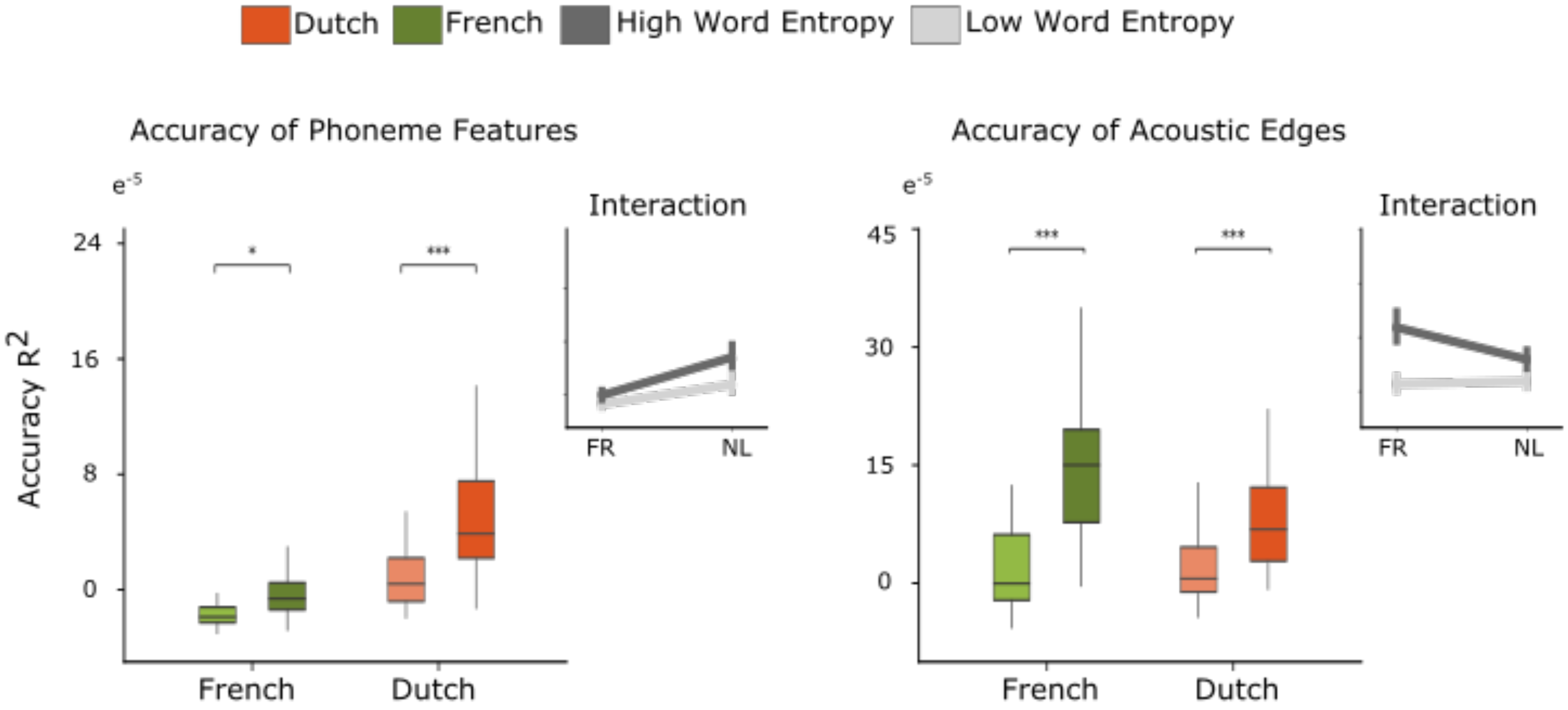
Linguistic representations are about more than statistics ->

The *ne plus ultra* of brain computation - compress ethological statistics into stable robust internal states for behavior, such that behavior is no longer driven by statistics alone

Martin (2016, 2020); Martin & Doumas (2019, 2020)

How does the brain transform the sensory into the linguistic?

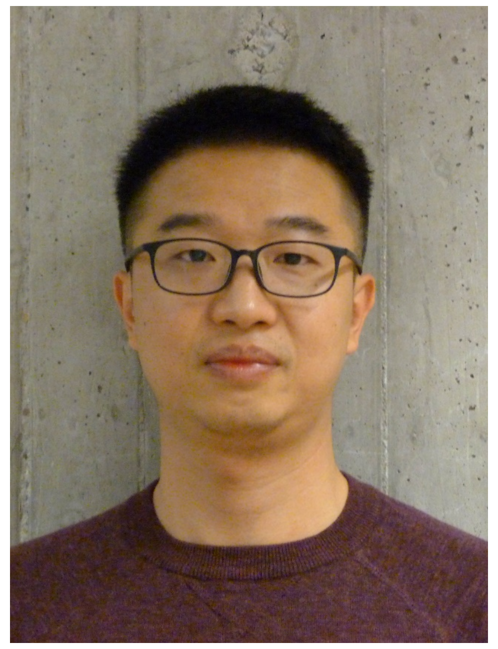
- A trade off between physical and abstract encoding:
  - acoustic edges are ‘softened’ during language comprehension, phonemes are enhanced



Filiz Tezcan

Tezcan, Weissbart, & Martin (2023) *eLife*

How does the brain transform the sensory into the linguistic?



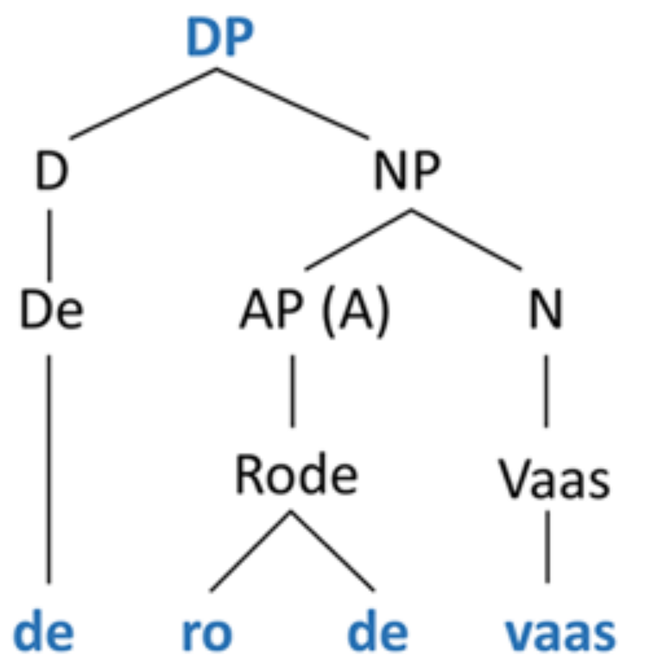
Fan Bai

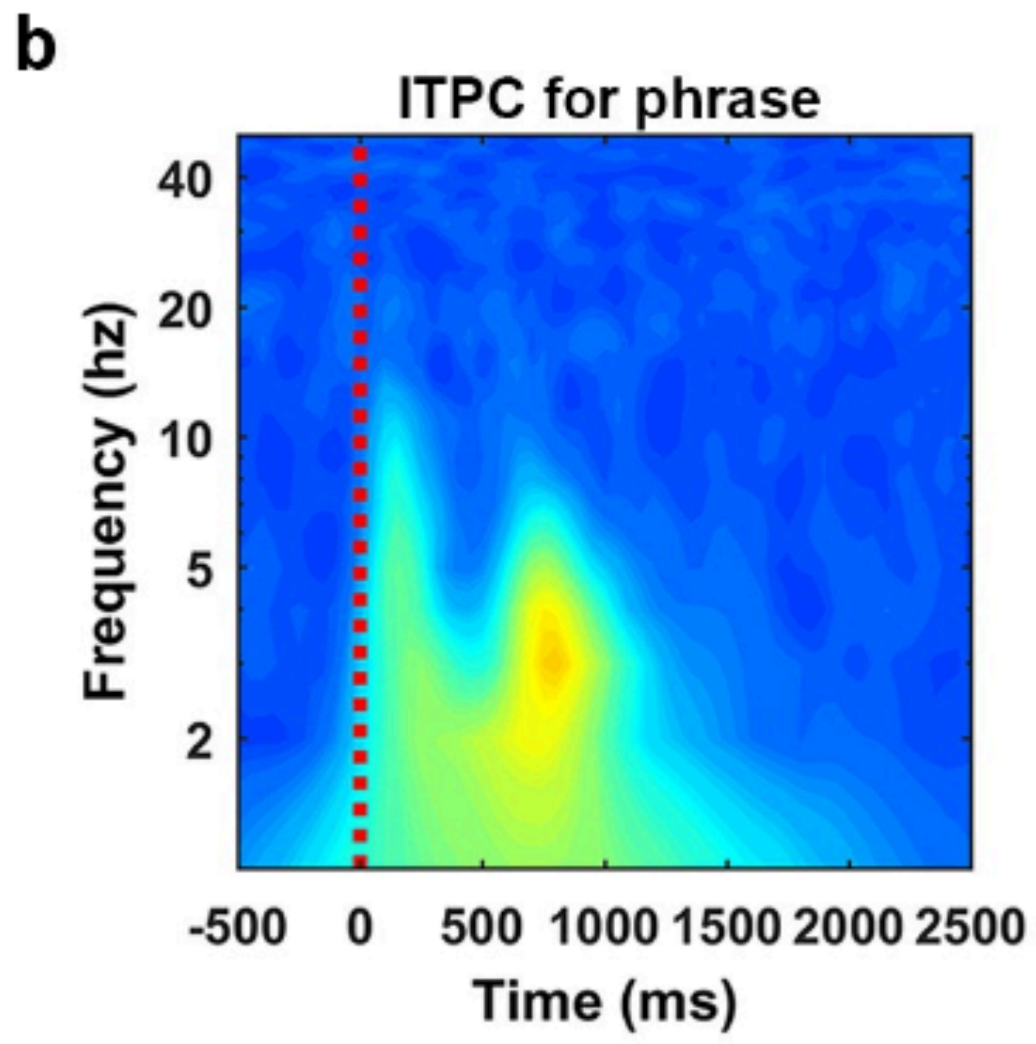
phase synchronization (ITPC)

physically and temporally matched spoken stimuli:

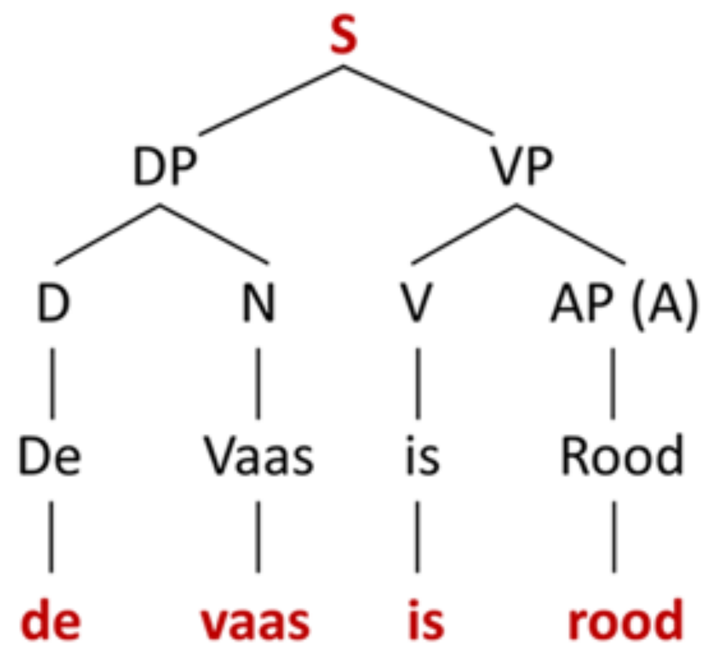
- more phase synchronisation for sentences compared to phrases
- ensembles activity is more coordinated in time in sentences than in phrases
- dynamics may scale with relational structure e.g., constituency

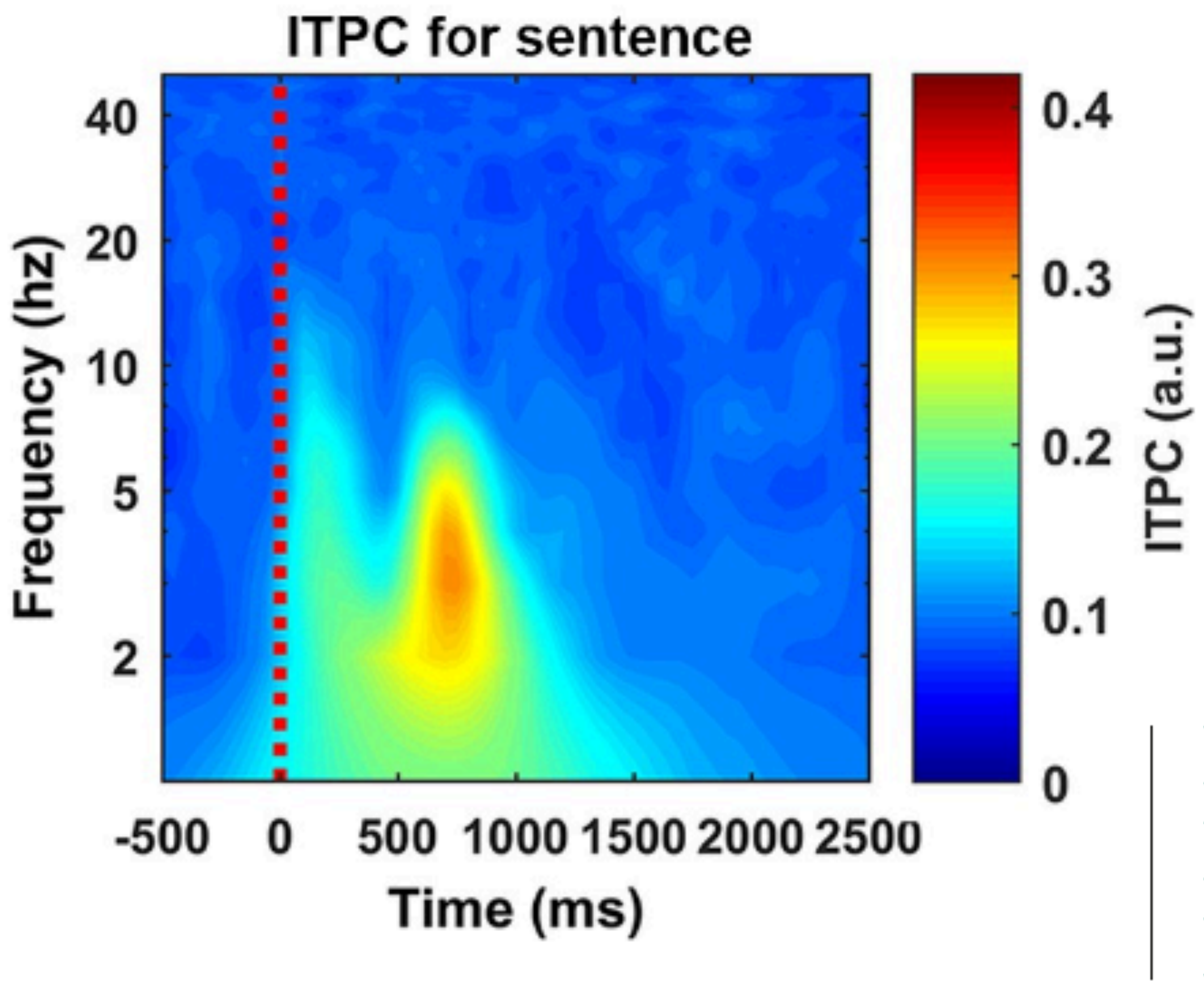
Phrase





Sentence



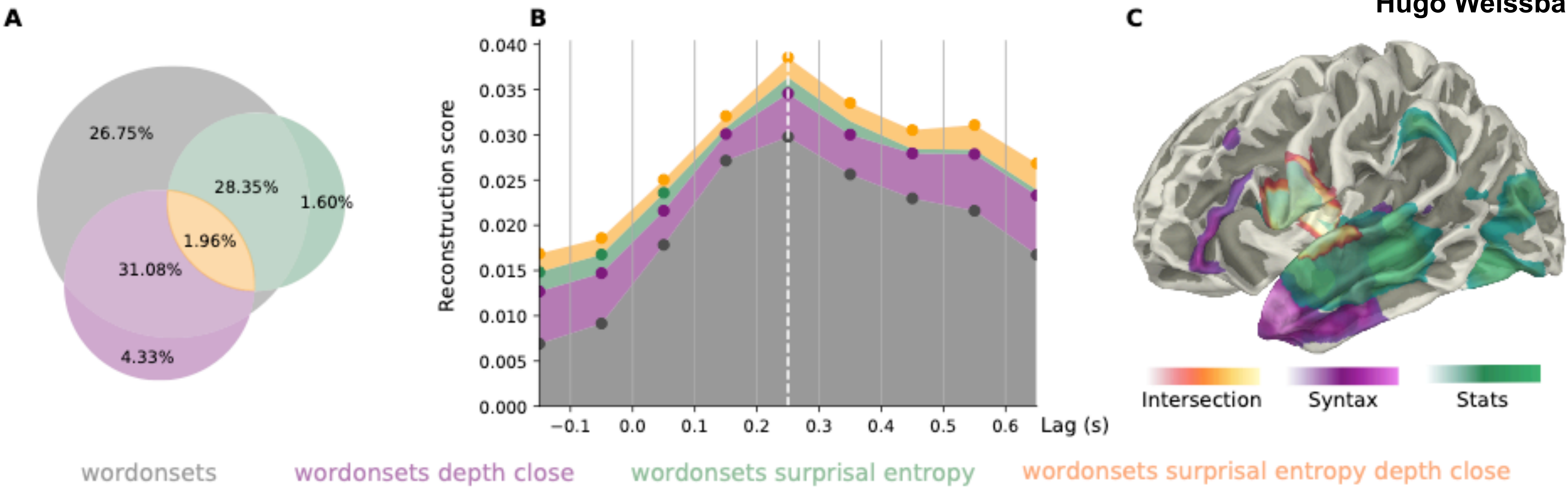


Bai, Meyer, & Martin (2022) *PLoS Biology*

structure predicts the neural response in a sustained way
statistics shape the phase of neural dynamics

B78

Hugo Weissbart



Weissbart & Martin (2023) *biorXiv*

linguistic structure (and statistics) **in neural dynamics**

neural tracking of phrases is automatic and robust to task demands

the distribution of sources / engagement of MTG & IFG changes with task

syntactic structure modulates **phase synchronization**, statistics shape phase, both shape coupling

syntactic features reconstruct the neural response in a sustained way

statistical and syntactic predictors complement each other and interact in time and space

---

Bai, Meyer, & Martin (2022) *PLoS Biology*
Coopmans, de Hoop, Hagoort, & Martin (2022) *Neurobiology of Language*
Kaufeld, Bosker, Ten Oever, Alday, Meyer, & Martin (2020) *JNeurosci*
ten Oever, Carta, Kaufeld, & Martin (2022) *eLife*
ten Oever & Martin (2021) *eLife*
Tezcan, Weissbart, & Martin (2023) *eLife*
Slaats, Weissbart, Schoffelen, Meyer, & Martin (2023) *JNeurosci*
Weissbart & Martin (2023) *biorXiv*
Zioga, Weissbart, Lewis, Haegens, & Martin (2023) *JNeurosci*

Focus questions

What do we mean when we conclude that LLM (or any model) and the brain are ‘similar’?

What is being predicted? How?

How are things processed when they are not predicted?

What are we trying to explain?

What is a good explanation?

Are we leveraging what we know about brain computation, psychology, and linguistics?

Language and Computation in Neural Systems at SNL 2023

A45

Cas W. Coopmans

E4

Julia Chauvet

D12

Rong Ding

A28

Anna Mai

D33
E4

Sophie Slaats

D64

Noémie te Rietmolen

A72

Filiz Tezcan

A72
A85
B78
D67

Hugo Weissbart

A85
D67

Ioanna Zioga

MAX-PLANCK-GESELLSCHAFT

MAX PLANCK
MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

NWO

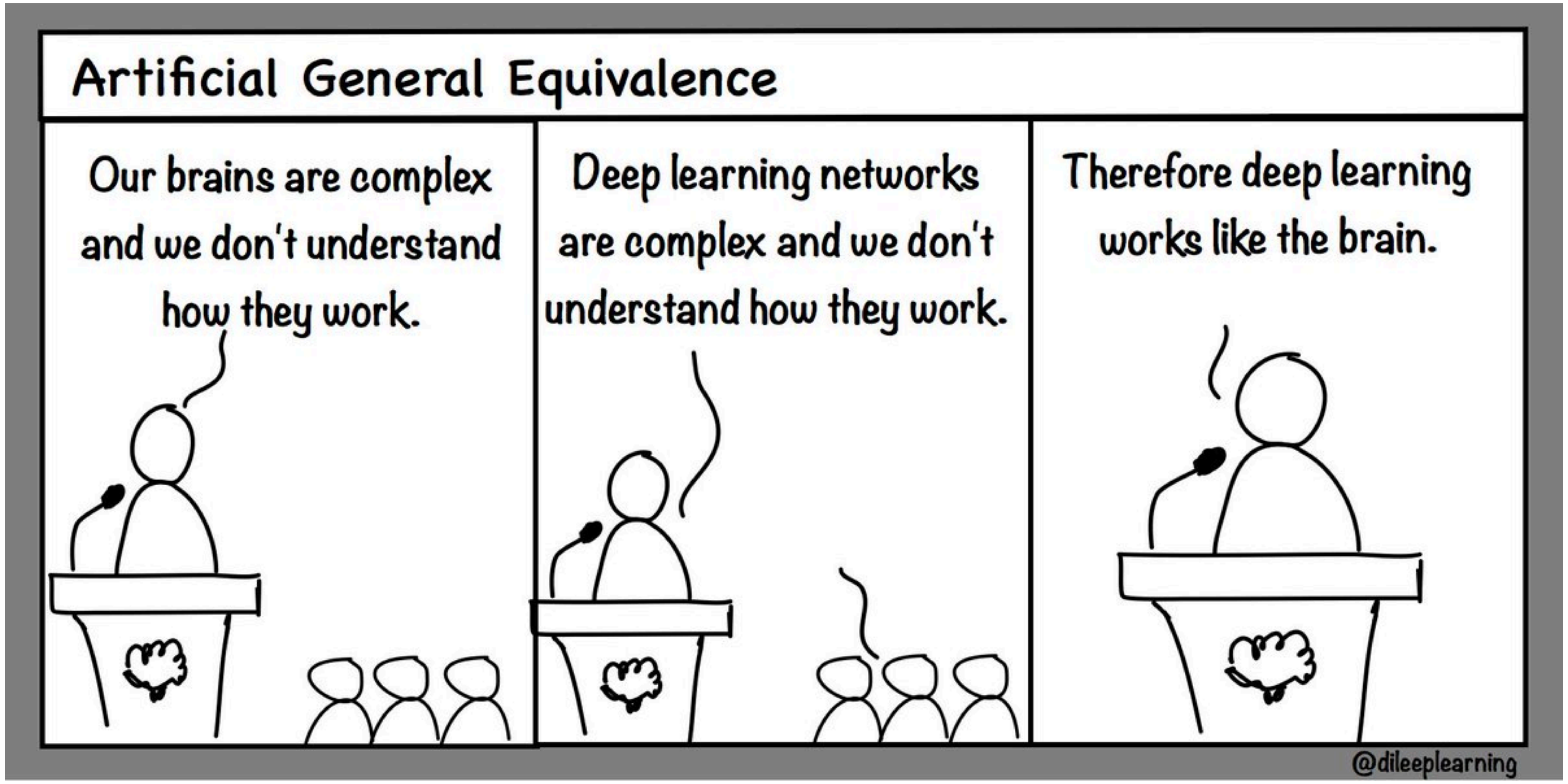
DONDERS INSTITUTE

LANGUAGE in INTERACTION

lacns.github.io

andreaemartin.com

andrea.martin@mpi.nl


Artificial General Equivalence
Our brains are complex and we don't understand how they work.
Deep learning networks are complex and we don't understand how they work.
Therefore deep learning works like the brain.
@dileeplearning

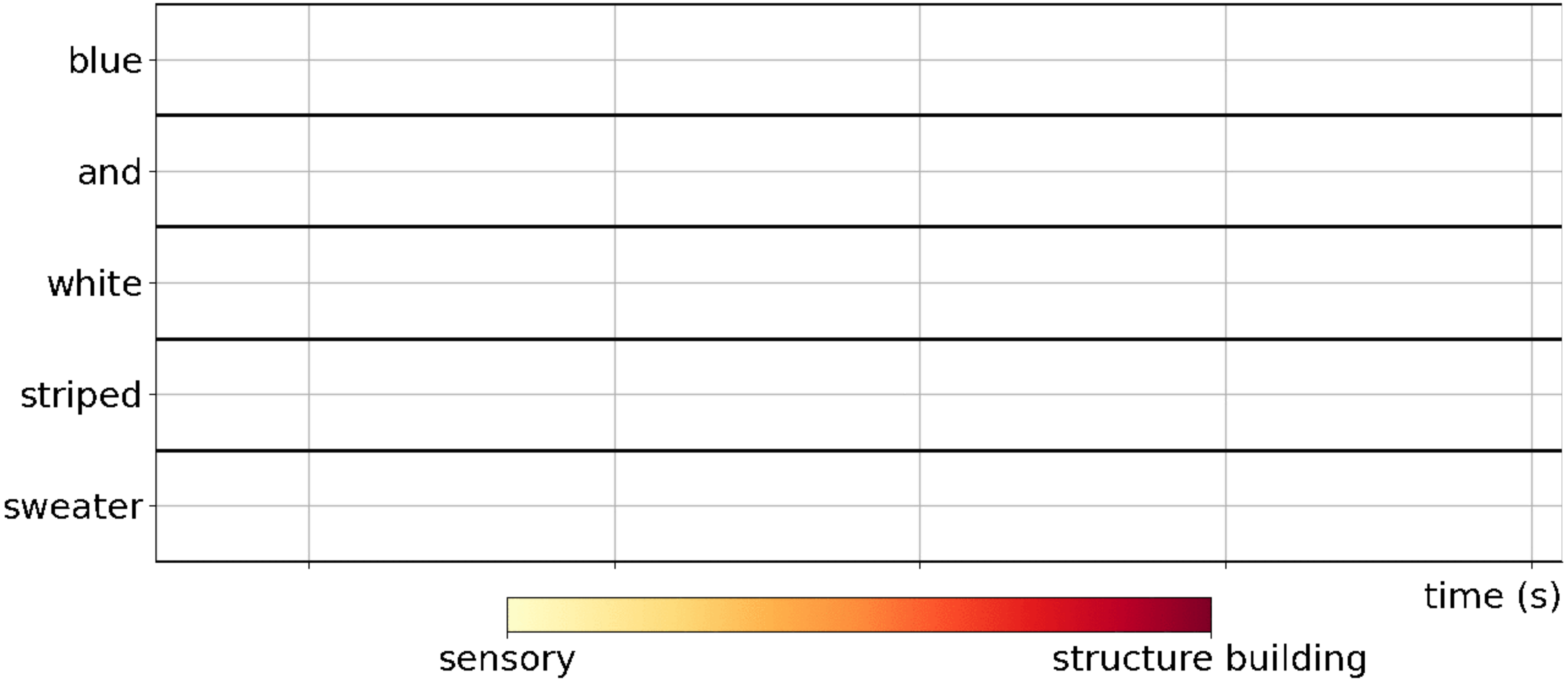
"[A]ll science would be superfluous if the outward appearance and the essence of things directly coincided."

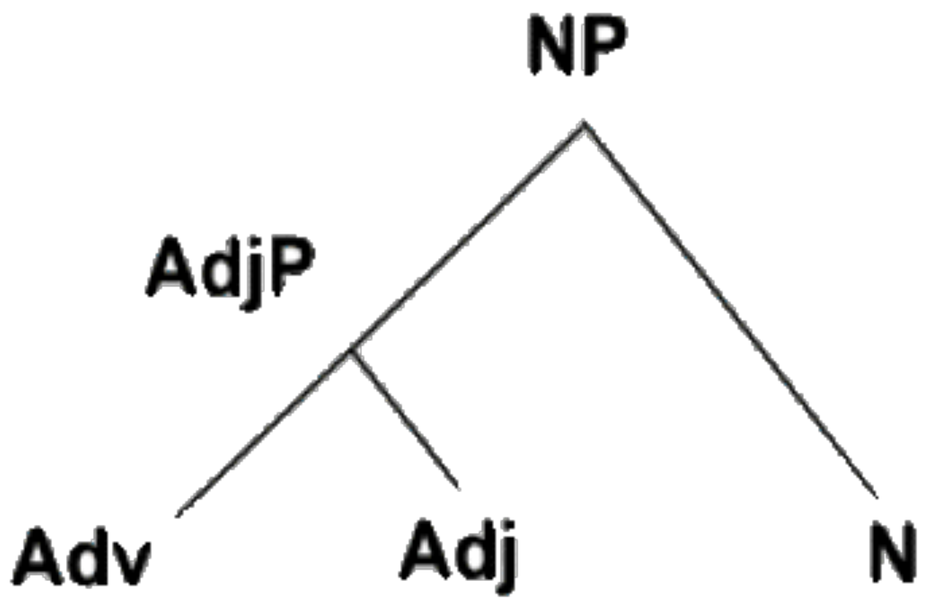
Marx, 1894, p. 592

with thanks to Dr. Olivia Guest

Why does Chat-GPT sound so good?

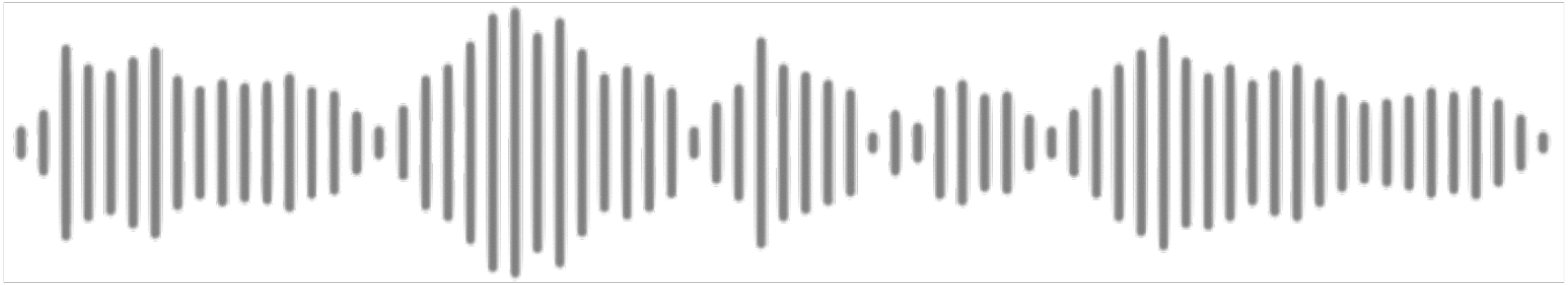
- No distinction between training and test data
- Overfitting
- Billion parameters tuned to optimise continuation with supervision/ feedback; interpolation within optimised parameter space
- Human labeled data, including hand tuning of parameters to select ‘best’ answers
- RLHF - Reinforcement learning human feedback - banned in the 90’s

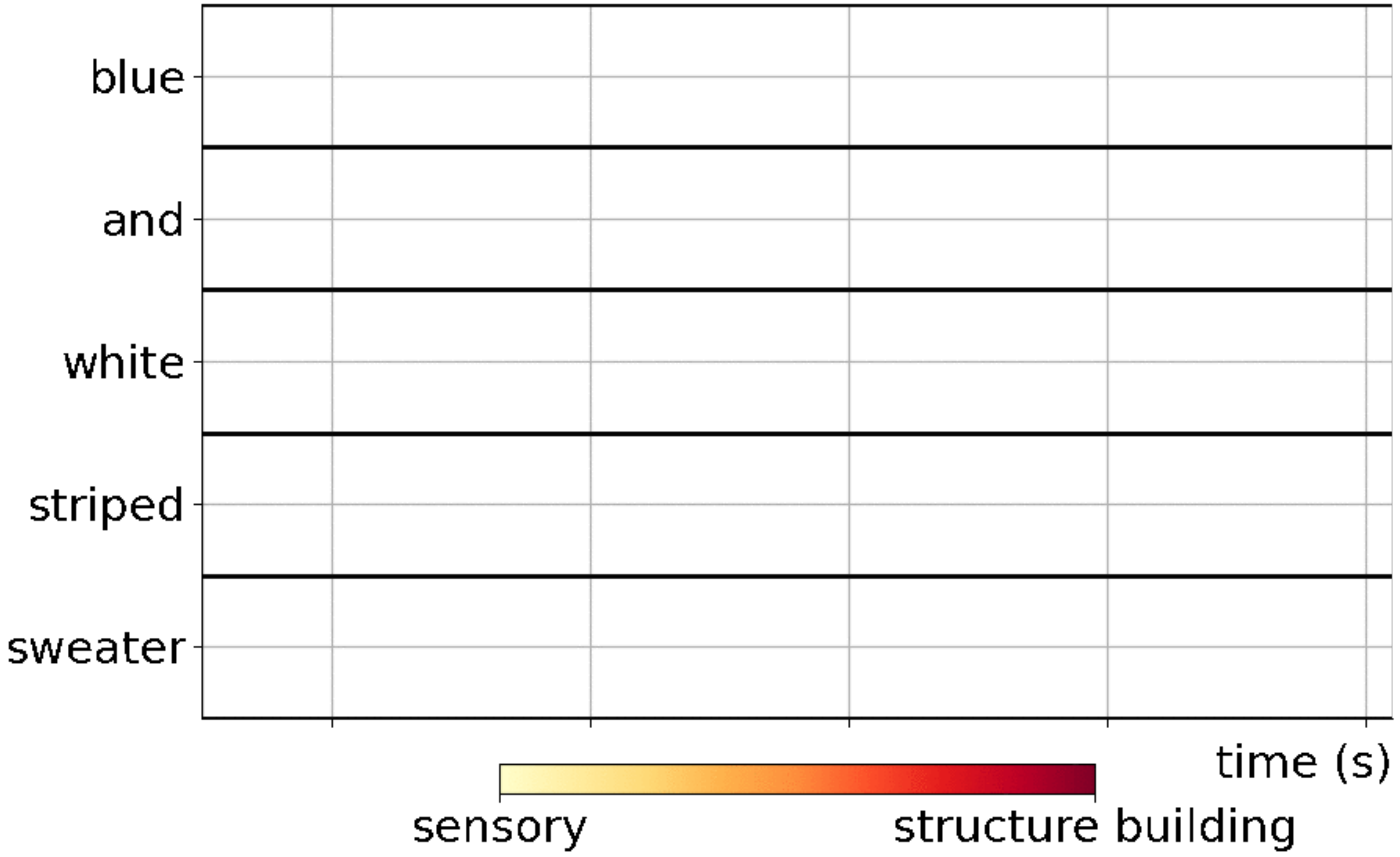


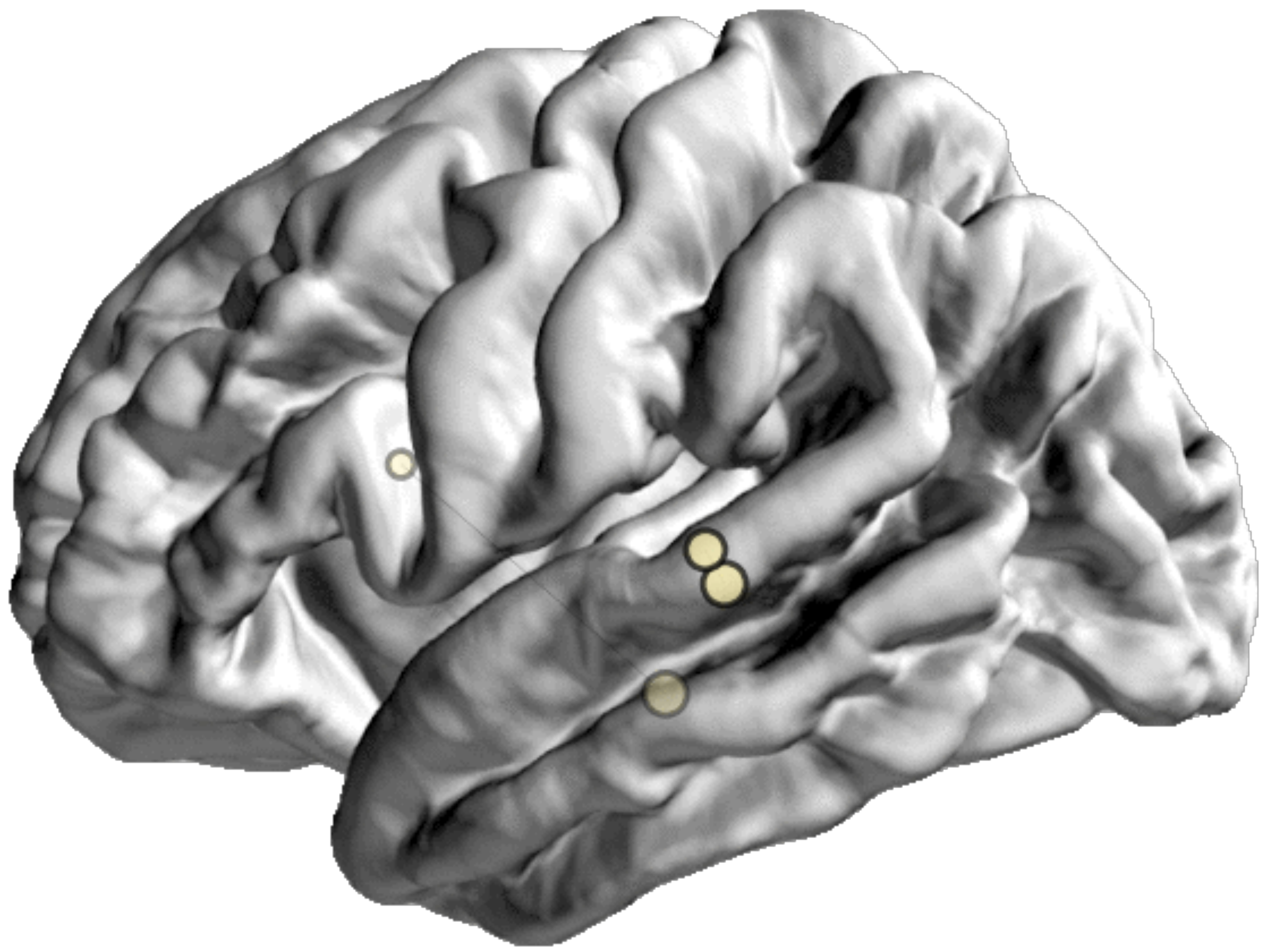


Martin (2016) *Frontiers in Language Sciences*
Martin (2020) *Journal of Cognitive Neuroscience*

thanks to Noémie te Rietmolen and Anna Mai







Martin (2016) *Frontiers in Language Sciences*
Martin (2020) *Journal of Cognitive Neuroscience*

thanks to Noémie te Rietmolen and Anna Mai