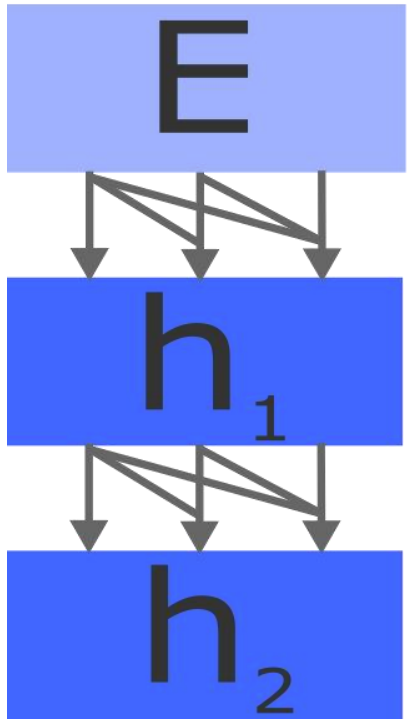
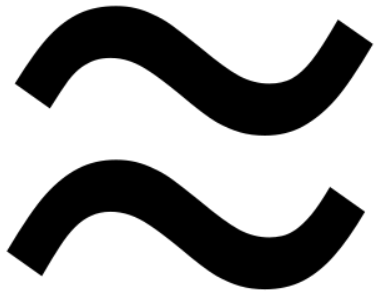
The neural architecture of language: Integrative modeling converges on predictive processing

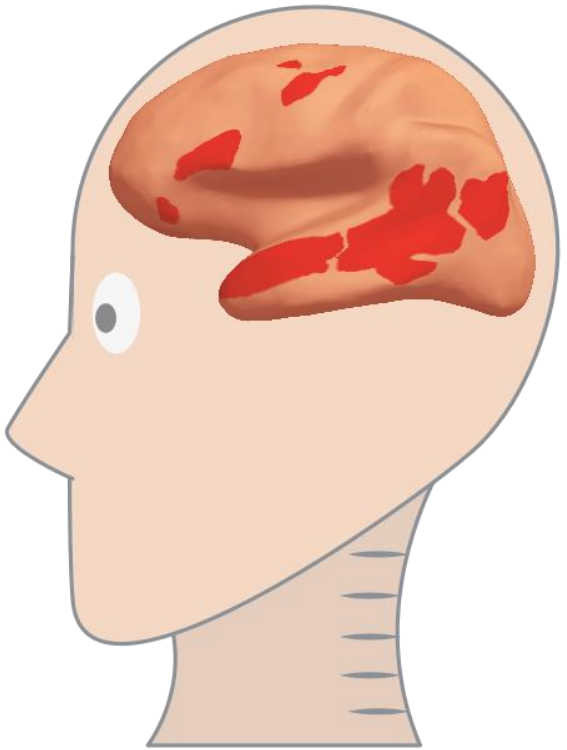
Martin Schrimpf

msch@mit.edu

@martin_schrimpf







MIT Quest for Intelligence

brain+cognitive sciences

McGOVERN INSTITUTE FOR BRAIN RESEARCH

CENTER FOR Brains Minds+ Machines

The neural architecture of language: Integrative modeling converges on predictive processing

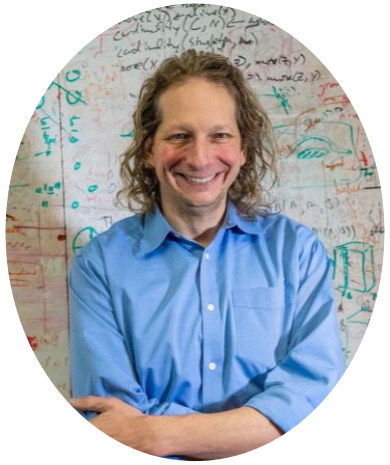
Idan Blank

Greta Tuckute

Carina Kauf

Eghbal Hosseini

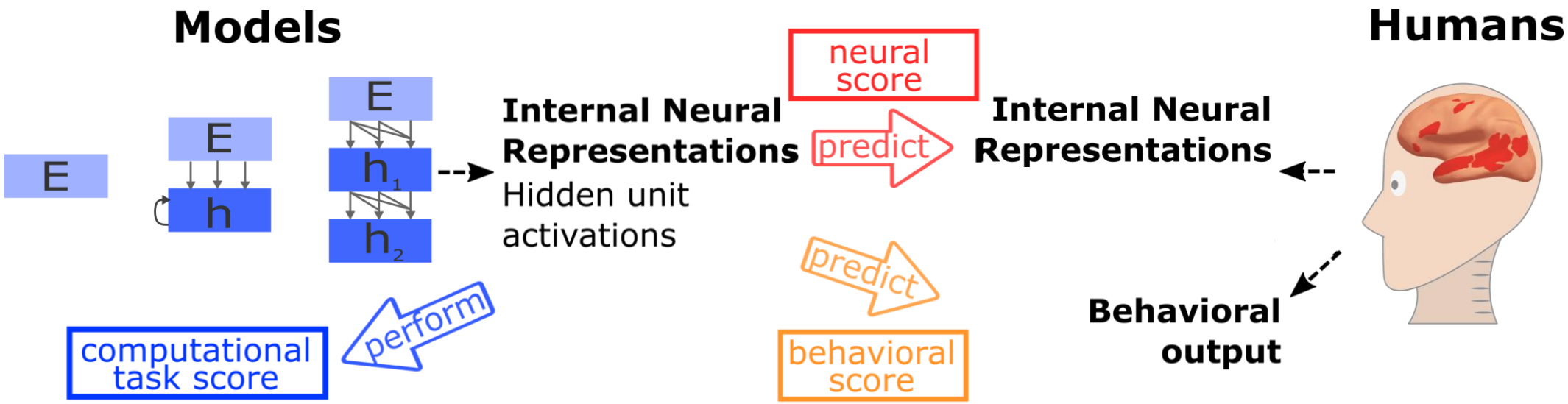
Nancy Kanwisher



Josh Tenenbaum

Ev Fedorenko

Integrative Modeling:
link neural mechanisms, behavior, and computation



Schrimpf et al. Neuron 2020

Modeling Sensory Systems: Vision

V1

edges

Hubel & Wiesel 1962

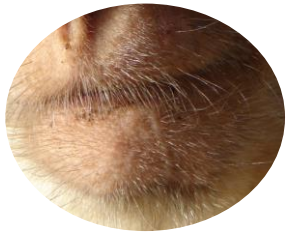
Necessary first steps! But …

V2

textures?

Freeman, Ziemba*, Heeger, Simoncelli, Movshon 2015*

V4



curvatures??

Pasupathy & Connor 1999

Bashivan, Kar*, DiCarlo 2019*

IT

objects???

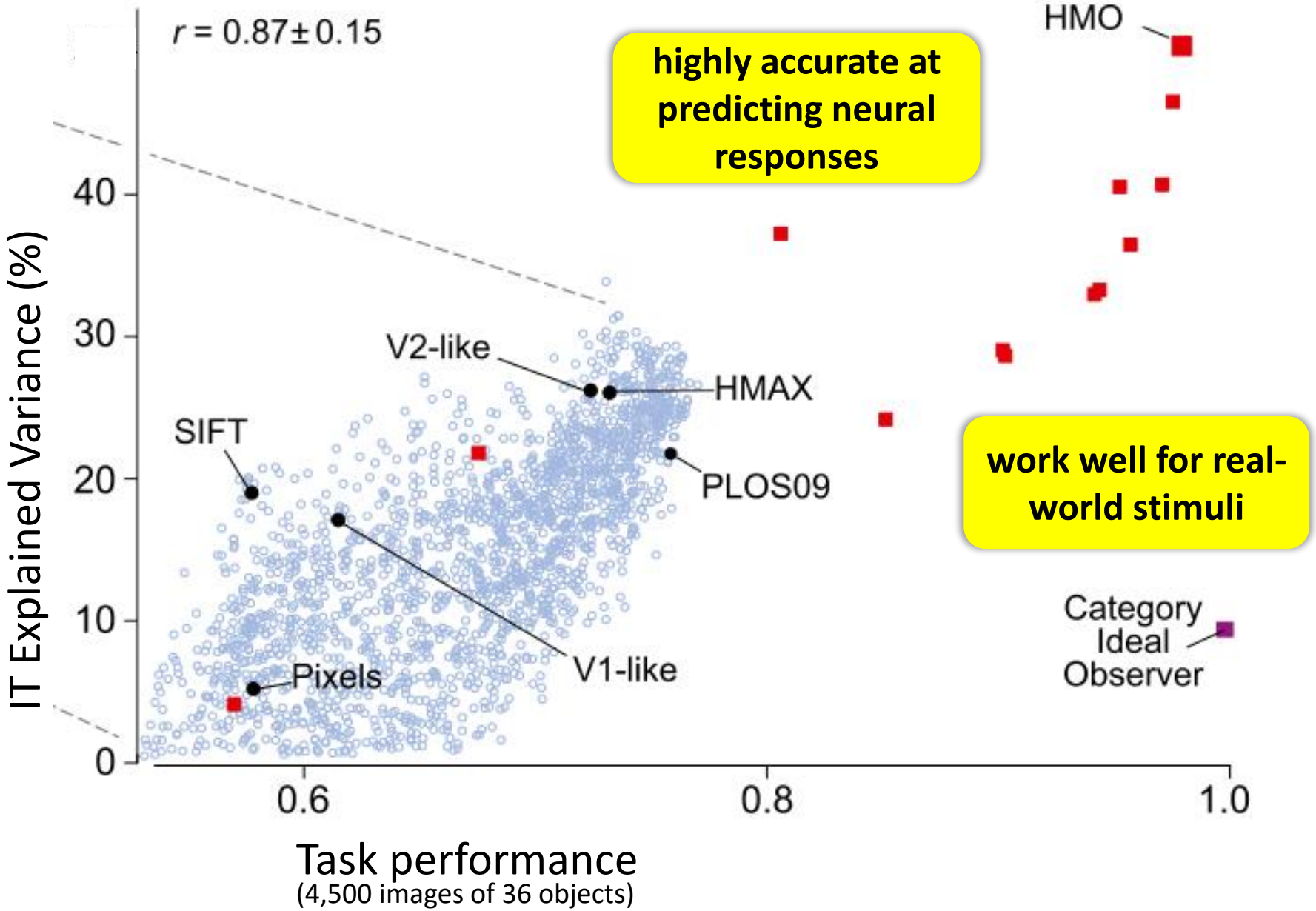
Hung, Kreiman*, Poggio, DiCarlo 2005*

purely descriptive

very little predictive power

often fall short for real-world stimuli

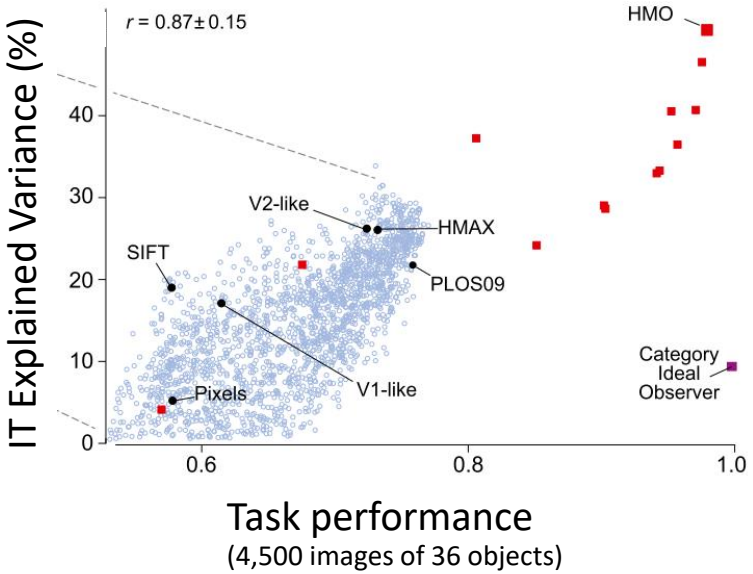
Enter artificial neural networks

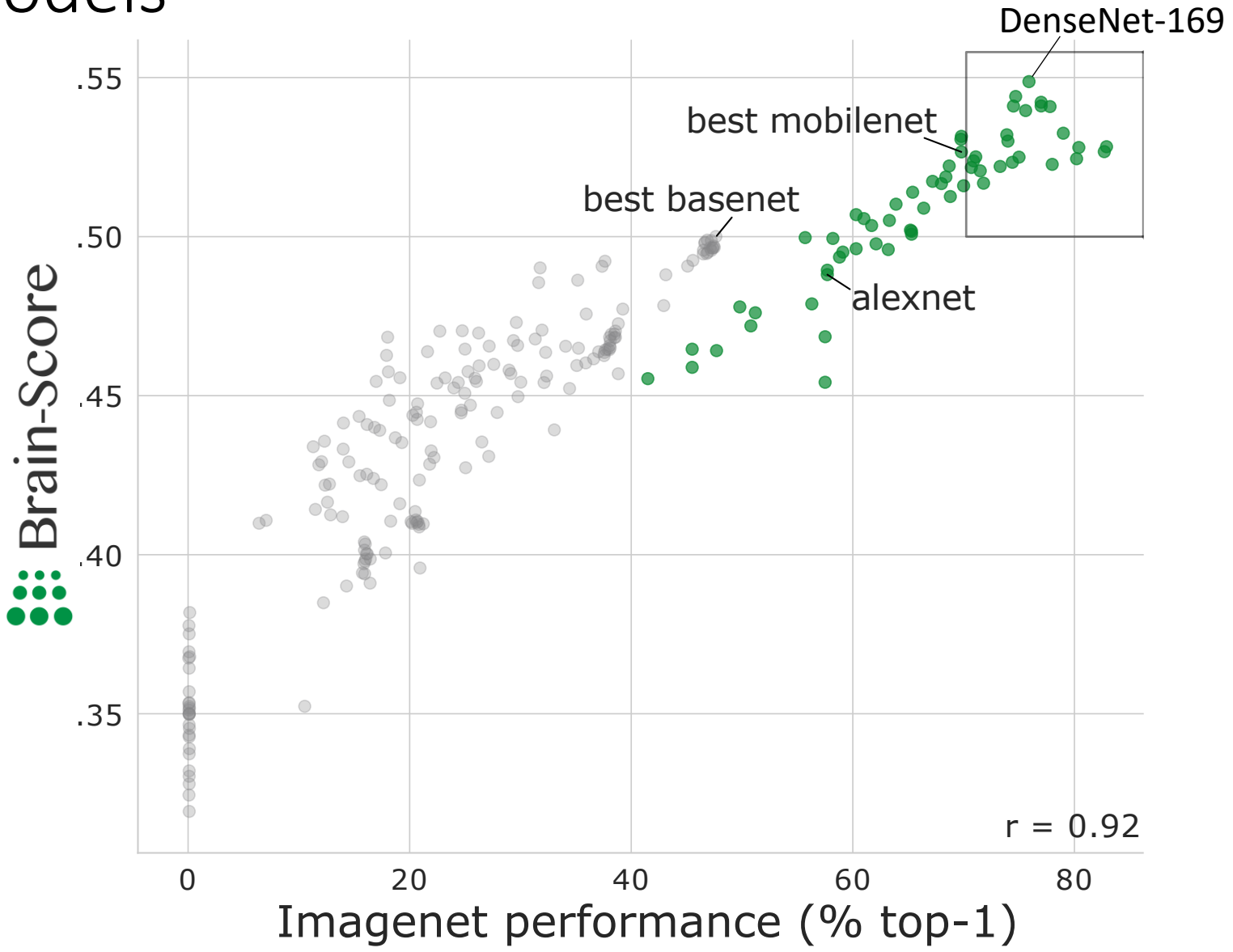


make predictions on any new input

Yamins, Hong*, et al. 2013 | Khaligh-Razavi & Kriegeskorte 2014 | Yamins*, Hong*, et al. 2014 | Yamins & DiCarlo 2016 | Schrimpf*, Kubilius*, et al. 2018*

Recent progress in sensory systems: increasingly accurate neurally-mechanistic models





Yamins, Hong*, et al. 2013 | Khaligh-Razavi & Kriegeskorte 2014 | Yamins*, Hong*, et al. 2014 | Zhuang et al. 2017 | Kell et al. 2018*

www.brain-score.org *Schrimpf*, Kubilius*, et al. 2018 | Kubilius*, Schrimpf*, et al. 2019*

Modeling **higher cognition**

Perception

Language

High-level reasoning

Artificial neural networks have worked well in modeling sensory cortex – could they also predict higher cognition?

The human language network

working definition (Fedorenko et al.):
a set of **left-lateralized** regions on the lateral surfaces of **frontal** and **temporal** cortex that support **high-level** language processing.

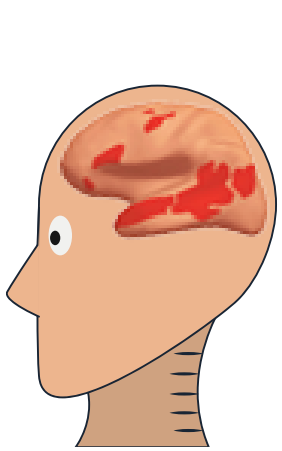
Language > Perceptually matched control

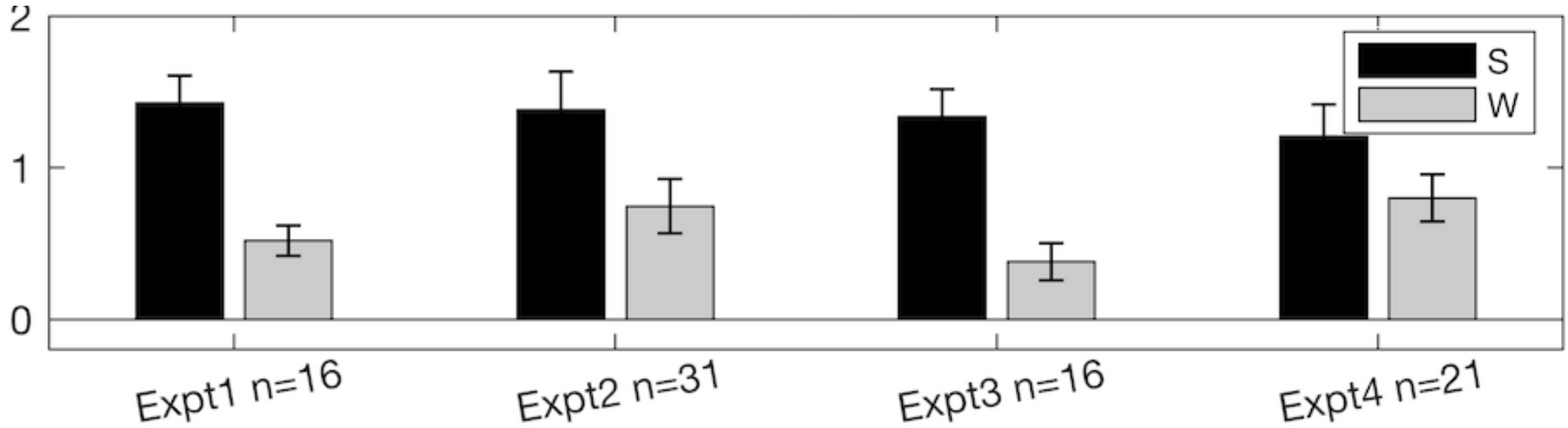
Sentences > Lists of nonwords

Fedorenko and Thompson-Schill 2014 | Braga, DiNicola and Buckner 2019

The human language network

the dog is taking a bath	>	blanket the green in ran giraffe a sad the





Key signature: stronger response to sentences than lists of unconnected words

Fedorenko, Behr and Kanwisher 2011 | Fedorenko et al. 2020

What are the mechanisms underlying human language comprehension?

the dog is taking a bath

"meaning"

What are the mechanisms underlying human language comprehension?

stimulus

present

record

observed data

Want to find the neural mechanisms that predict our data

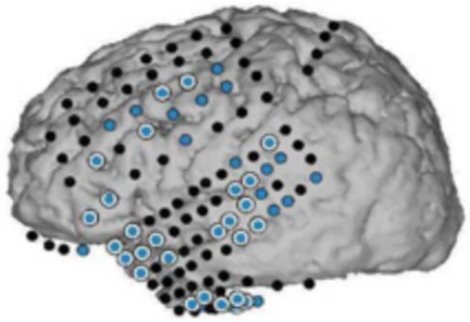
Computational understanding of human language processing

Guide and constraint next-generation NLP models

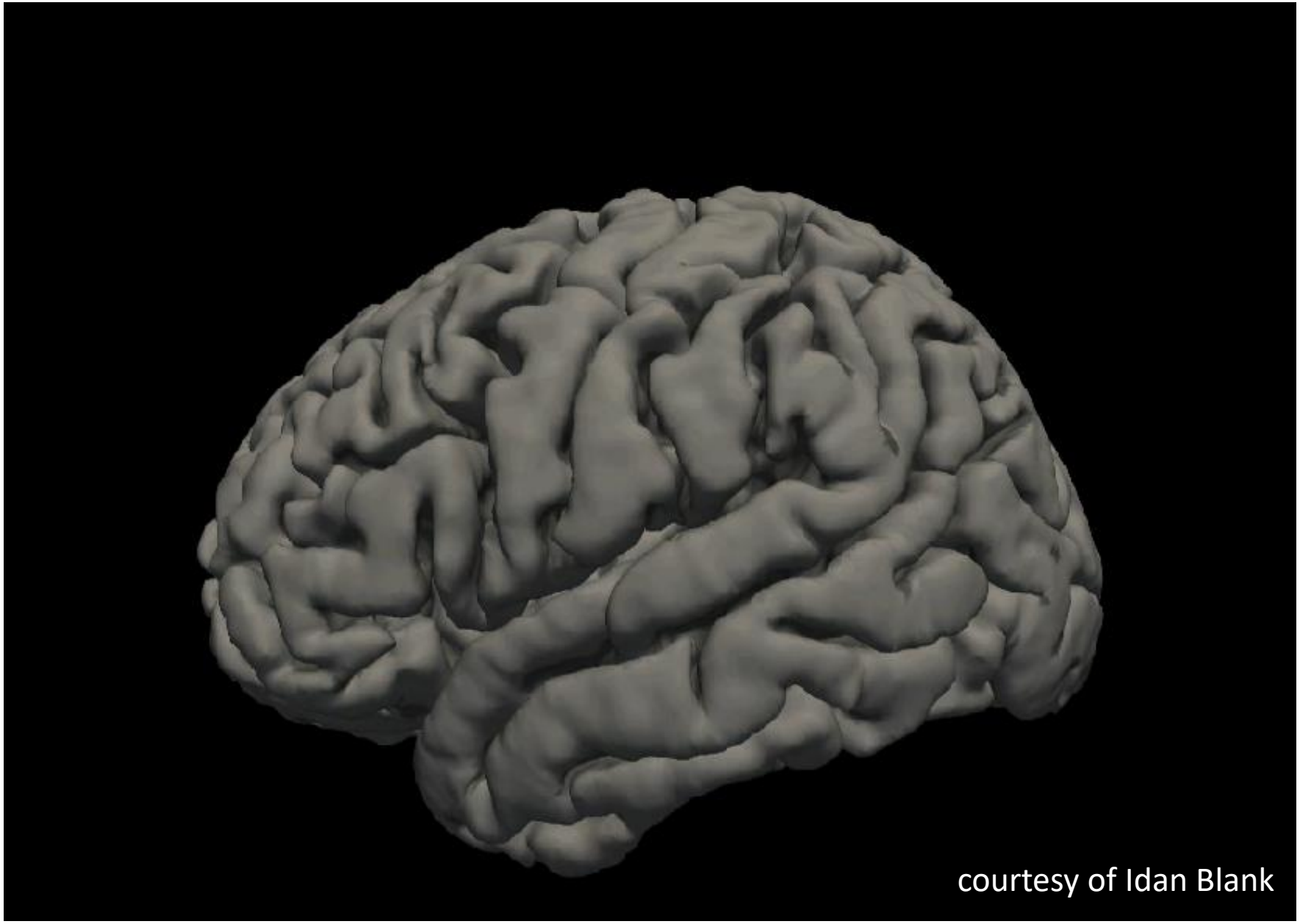
Future clinical applications (BMI)

The data target: human neural recordings

fMRI



ECoG



The data target: human neural recordings

Pereira et al. 2018 fMRI

627 sentences x 13,517 voxels in 10 subjects

Beekeeping encourages the conservation of local habitats. | It is in every beekeeper's interest …

Fedorenko et al. 2016 ECoG

416 words x 97 electrodes in 5 subjects

ALEX | WAS | TIRED | SO | HE | TOOK | A | NAP

Blank et al. 2014 fMRI

1,317 story fragments x 60 fROIs in 5 subjects

If you were to journey to the | North of England, you would come to a valley | that is surrounded by moors as high as | mountains. It is in this | valley where you would find the city of Bradford, | …

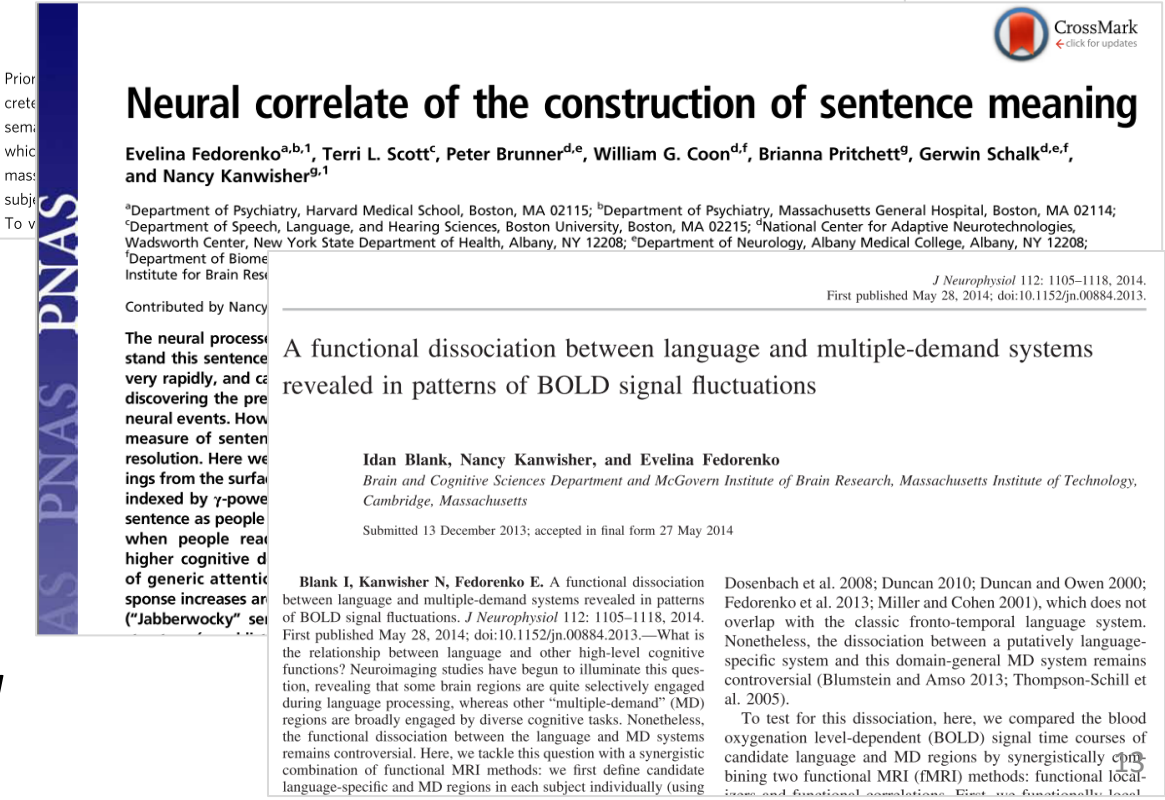
nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-03068-4 OPEN

Toward a universal decoder of linguistic meaning from brain activation

Francisco Pereira[1], Bin Lou[1], Brianna Pritchett[2], Samuel Ritter[3], Samuel J. Gershman[4], Nancy Kanwisher[2,5], Matthew Botvinick[3,6] & Evelina Fedorenko[5,7,8]



CrossMark click for updates

PNAS

Neural correlate of the construction of sentence meaning

Evelina Fedorenko[a,b,1], Terri L. Scott[c], Peter Brunner[d,e], William G. Coon[d,f], Brianna Pritchett[g], Gerwin Schalk[d,e,f], and Nancy Kanwisher[g,1]

[a]Department of Psychiatry, Harvard Medical School, Boston, MA 02115; [b]Department of Psychiatry, Massachusetts General Hospital, Boston, MA 02114; [c]Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA 02215; [d]National Center for Adaptive Neurotechnologies, Wadsworth Center, New York State Department of Health, Albany, NY 12208; [e]Department of Neurology, Albany Medical College, Albany, NY 12208; [f]Department of Biome… Institute for Brain Res…

Contributed by Nancy…

The neural process… stand this sentence… very rapidly, and ca… discovering the pre… neural events. How… measure of senten… resolution. Here we… ings from the surfa… indexed by γ-powe… sentence as people… when people rea… higher cognitive d… of generic attentio… sponse increases ar… ("Jabberwocky" se…

J Neurophysiol 112: 1105–1118, 2014.
First published May 28, 2014; doi:10.1152/jn.00884.2013.

A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations

Idan Blank, Nancy Kanwisher, and Evelina Fedorenko

Brain and Cognitive Sciences Department and McGovern Institute of Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts

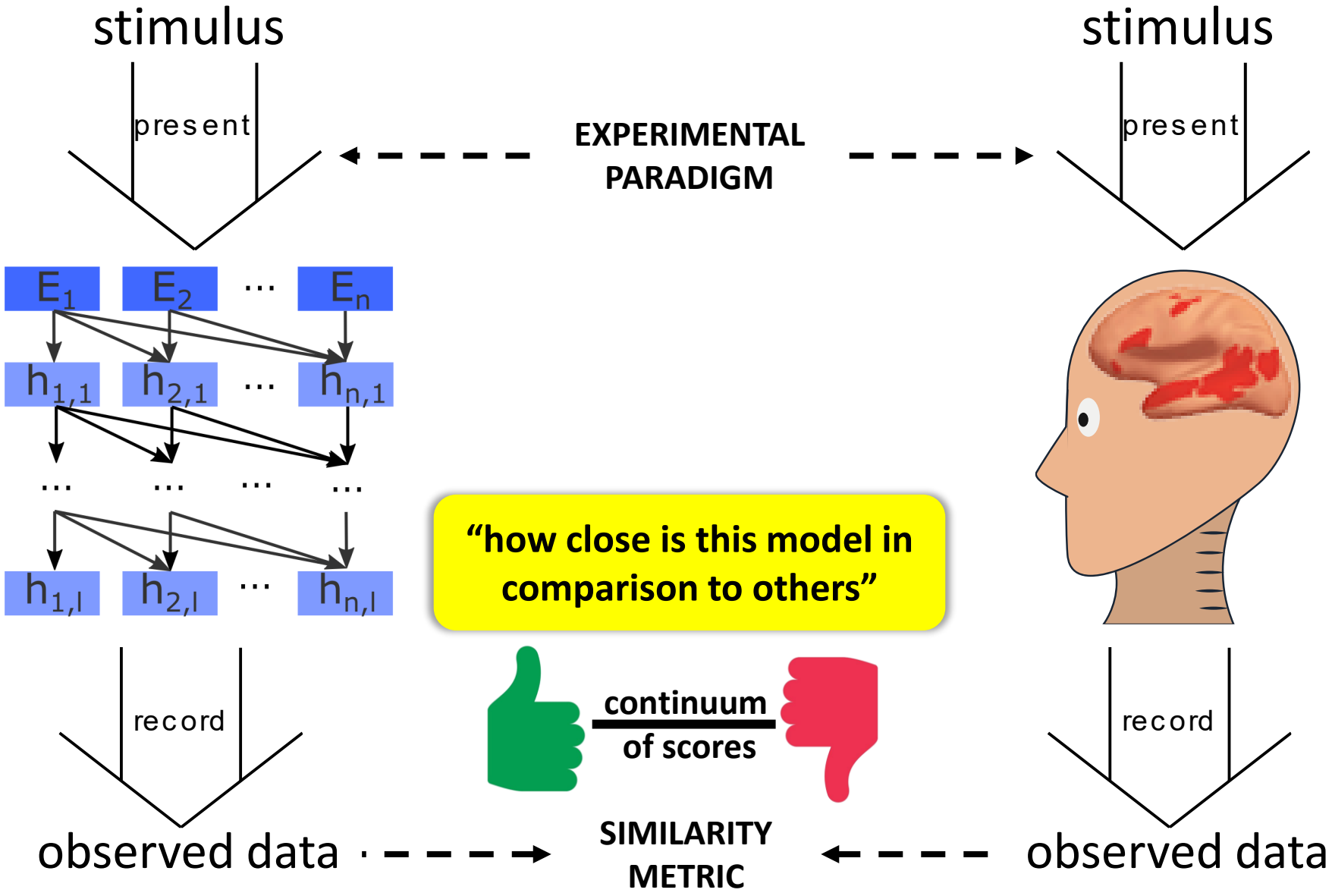
Submitted 13 December 2013; accepted in final form 27 May 2014

Blank I, Kanwisher N, Fedorenko E. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J Neurophysiol* 112: 1105–1118, 2014. First published May 28, 2014; doi:10.1152/jn.00884.2013.—What is the relationship between language and other high-level cognitive functions? Neuroimaging studies have begun to illuminate this question, revealing that some brain regions are quite selectively engaged during language processing, whereas other "multiple-demand" (MD) regions are broadly engaged by diverse cognitive tasks. Nonetheless, the functional dissociation between the language and MD systems remains controversial. Here, we tackle this question with a synergistic combination of functional MRI methods: we first define candidate language-specific and MD regions in each subject individually (using

Dosenbach et al. 2008; Duncan 2010; Duncan and Owen 2000; Fedorenko et al. 2013; Miller and Cohen 2001), which does not overlap with the classic fronto-temporal language system. Nonetheless, the dissociation between a putatively language-specific system and this domain-general MD system remains controversial (Blumstein and Amso 2013; Thompson-Schill et al. 2005).

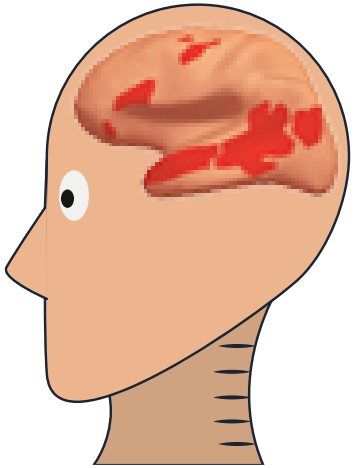
To test for this dissociation, here, we compared the blood oxygenation level-dependent (BOLD) signal time courses of candidate language and MD regions by synergistically combining two functional MRI (fMRI) methods: functional local-

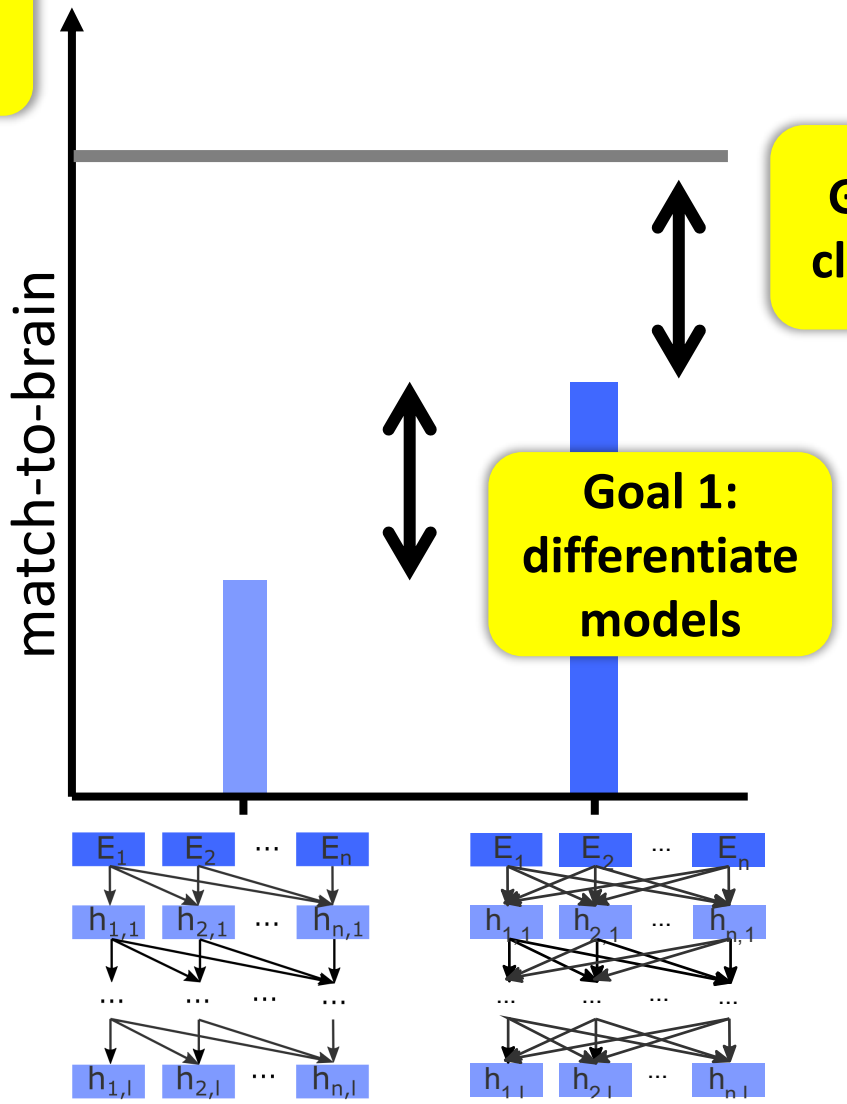
Quantifying match-to-brain: Benchmarking



Quantifying match-to-brain: Benchmarking

We only care about best-matching model (for now)

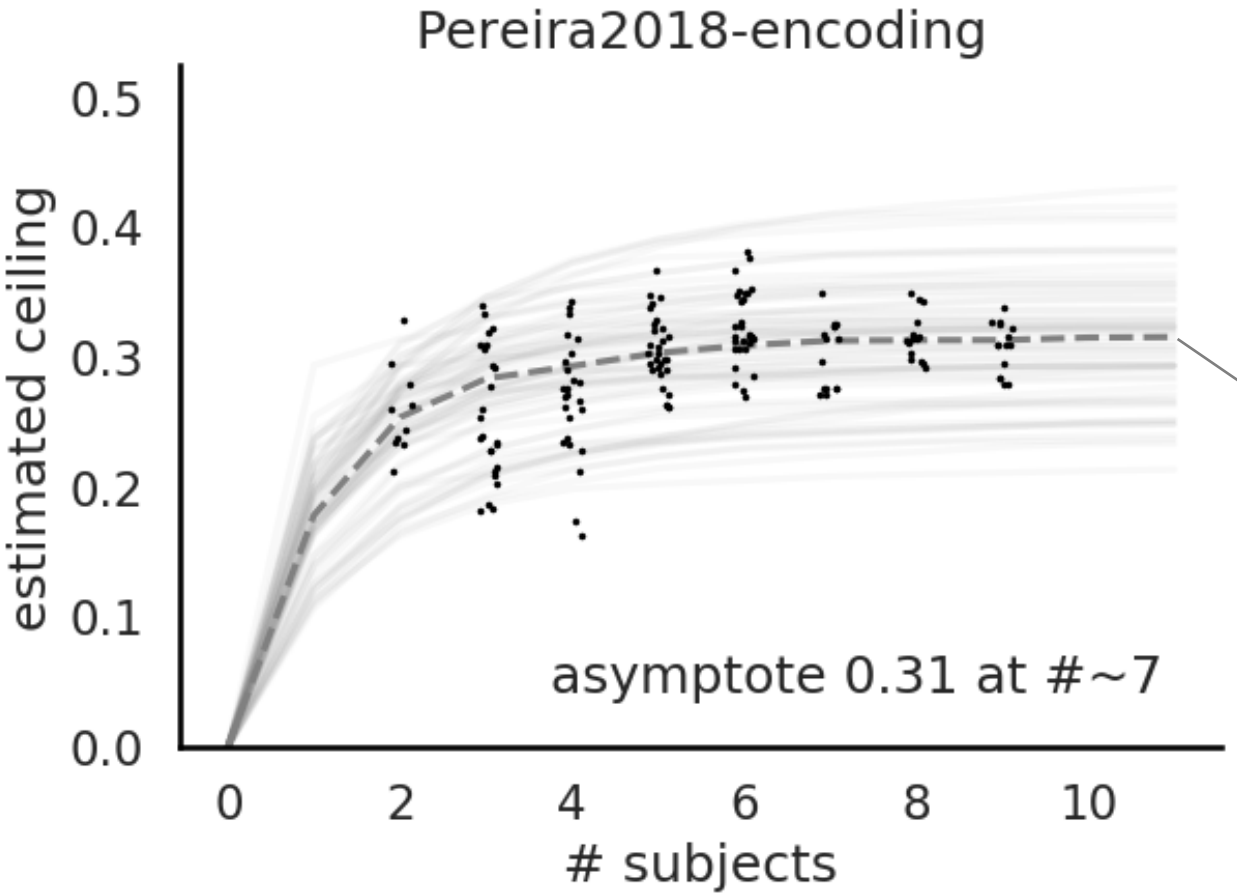


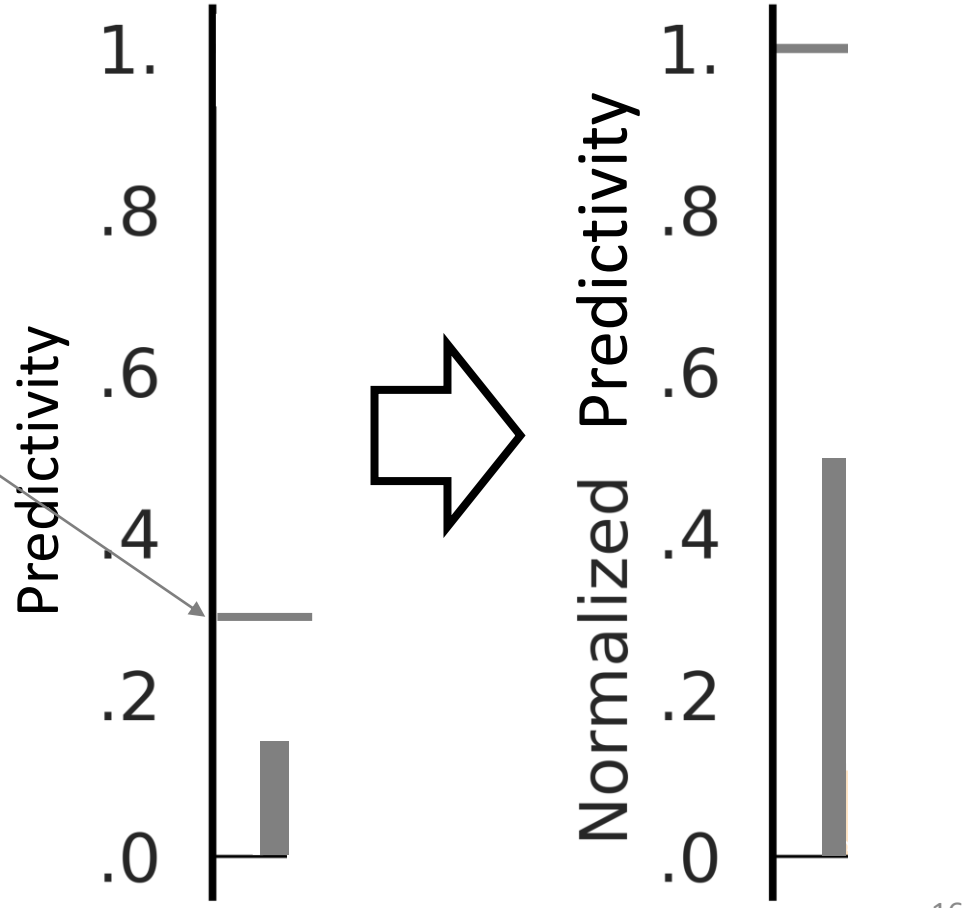


Goal 2 “how close are we” – **ceiling**

compute how well a pool of subjects predicts a held-out subject

$$predictivity_{normalized} = \frac{predictivity}{ceiling}$$





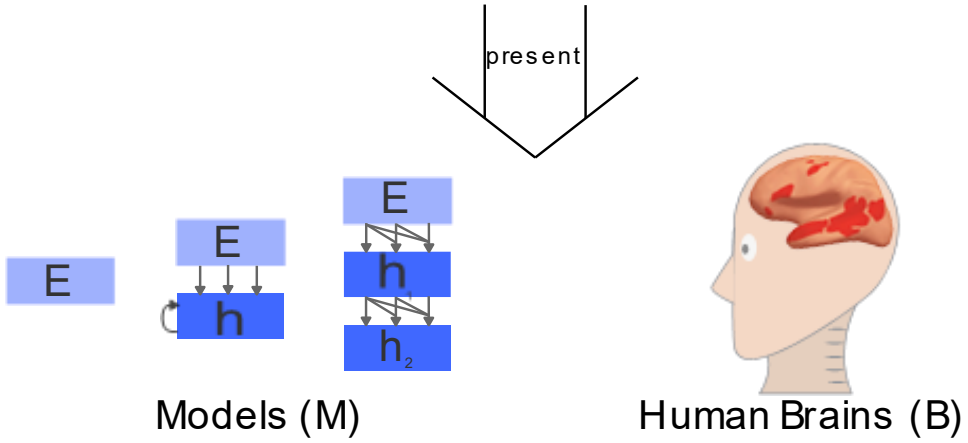
Stimuli

Pereira2018 "Beekeeping encourages the conservation of local habitats. It is in every beekeeper's interest..."

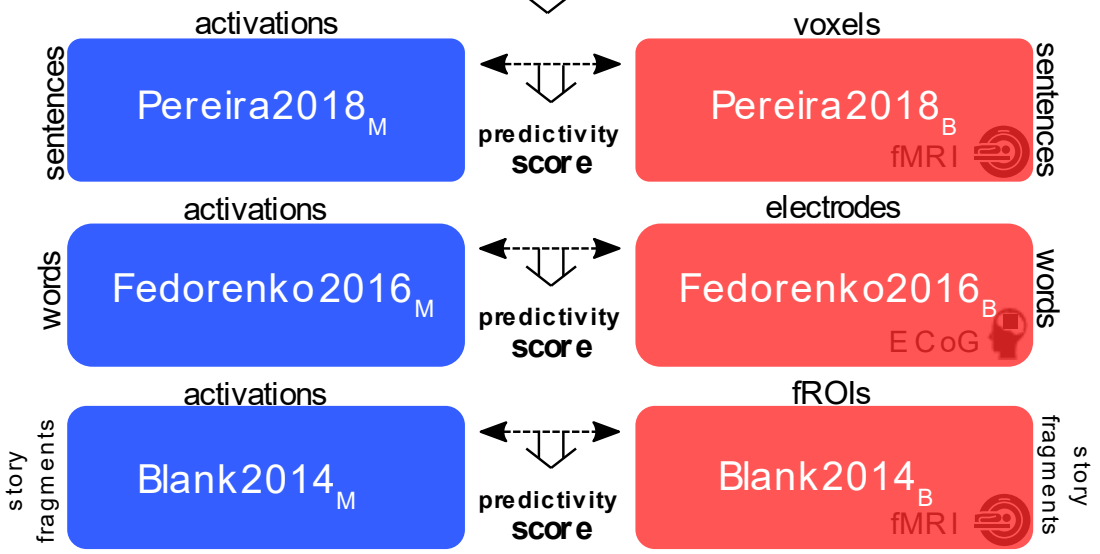
Fedorenko2016 "Alex was tired so he took a nap."

Blank2014 "If you were to journey to the North of England, you would come to a valley that is surrounded by moors as high as mountains. It is in this valley where you…"

Experimental Participants



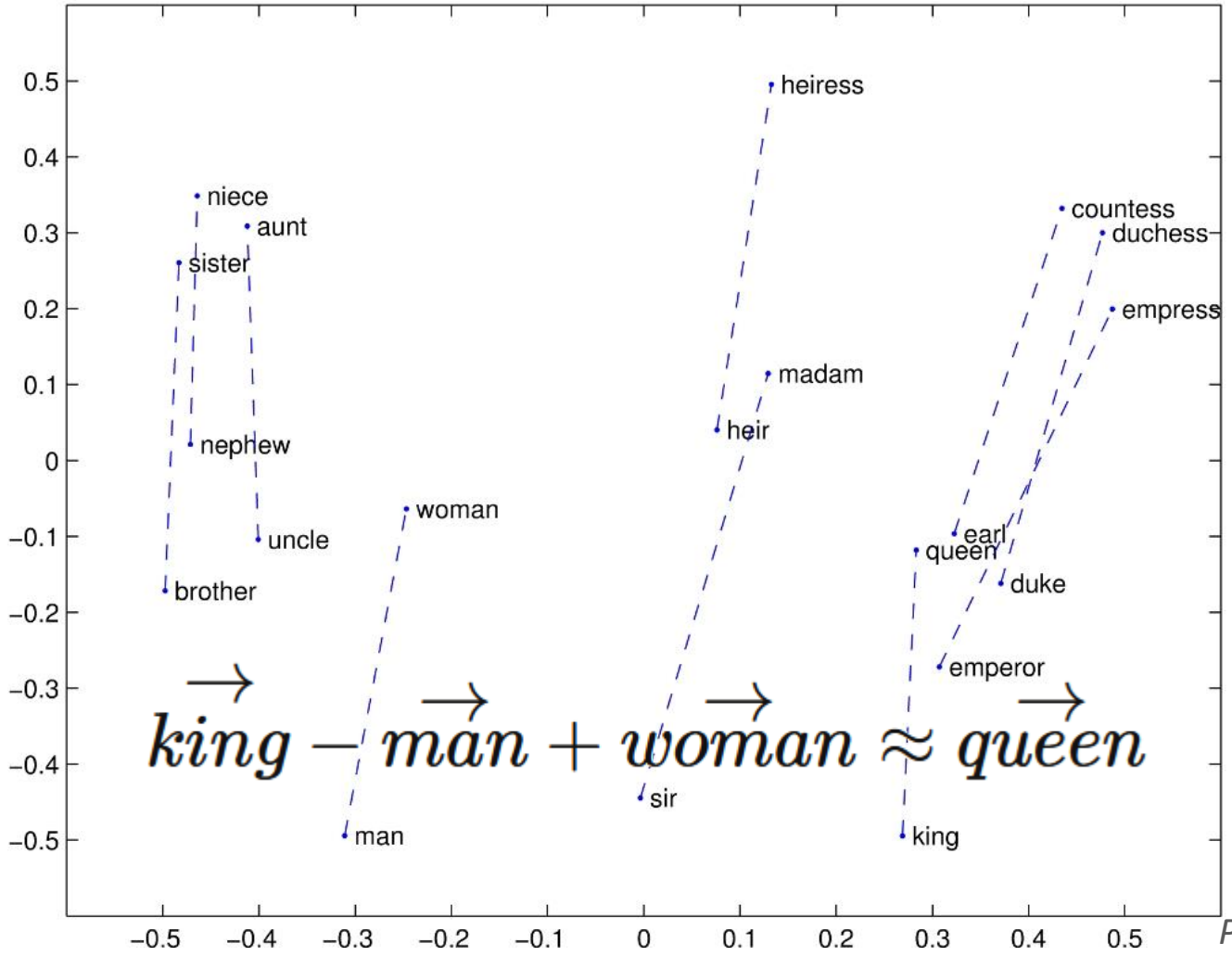
Comparative Measurements



Integrative Modeling: one model to predict them all

Models considered (n=43)

Embedding type models: GloVe, word2vec, topicETM

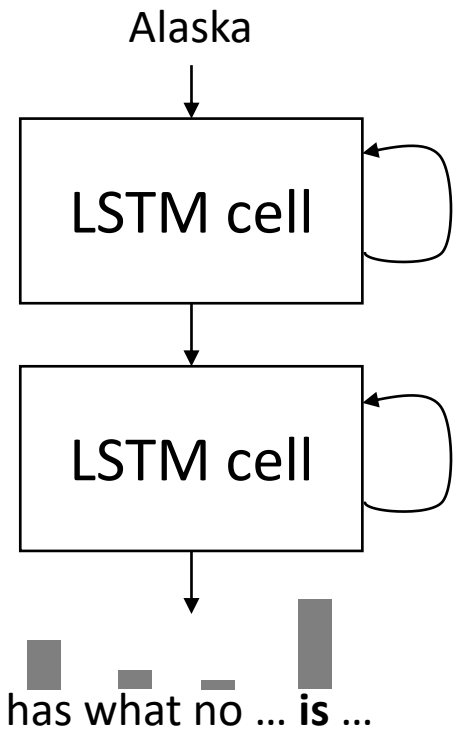


Pennington, Socher, Manning 2014

Models considered (n=43)

Embedding type models: GloVe, word2vec, topicETM

Recurrent networks: skip-thoughts, LSTM lm_1b



Language Modeling

Alaska is

Alaska is about

Alaska is about twelve

Alaska is about twelve times

Alaska is about twelve times larger

Alaska is about twelve times larger than

Alaska is about twelve times larger than New

Alaska is about twelve times larger than New York

Jozefowicz, vinyals, Schuster, Shazeer, Wu 2016

Image from https://www.quora.com/What-is-a-masked-language-model-and-how-is-it-related-to-BERT

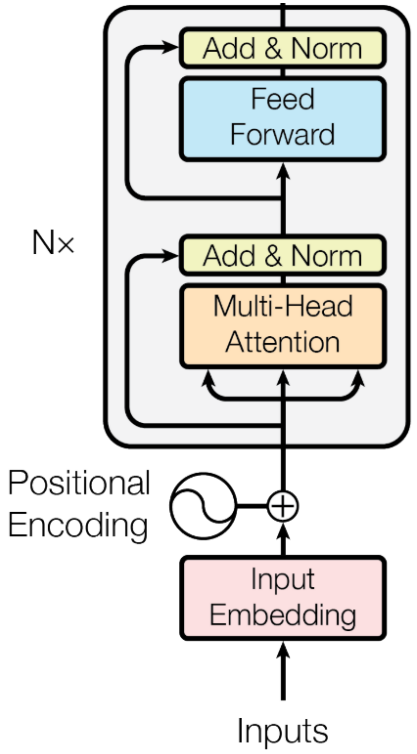
Models considered (n=43)

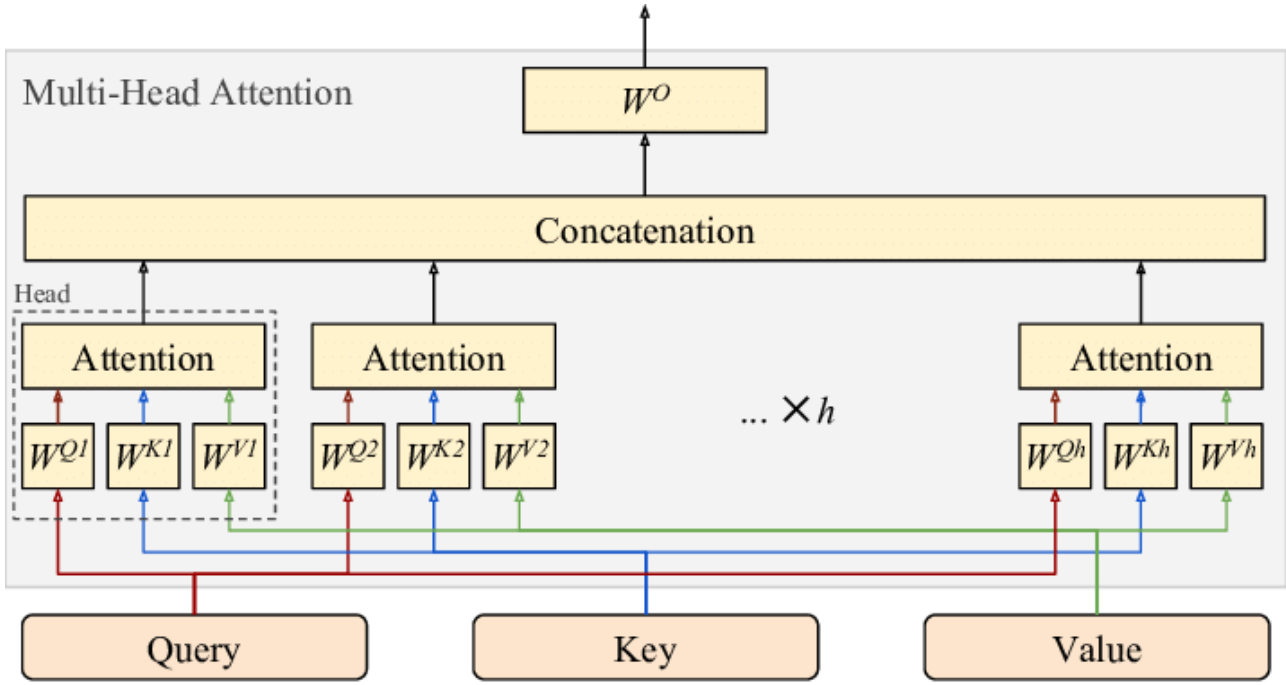
Embedding type models: GloVe, word2vec, topicETM

Recurrent networks: skip-thoughts, LSTM lm_1b

Transformers

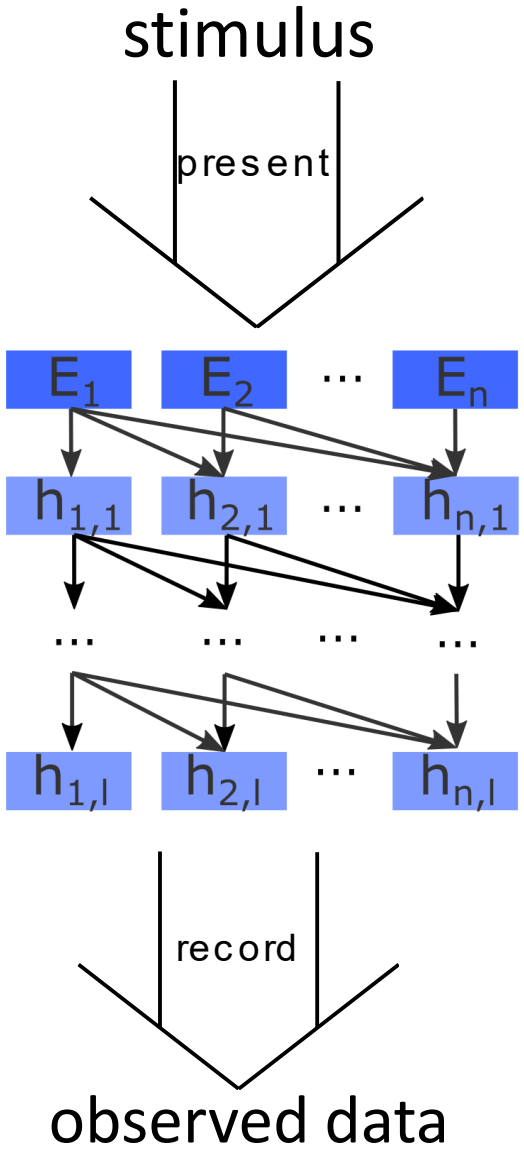
- BERTs
- RoBERTas
- XLMs
- Transformer-XLs
- XLNets
- CTRL
- T5s
- AlBERTs
- GPTs



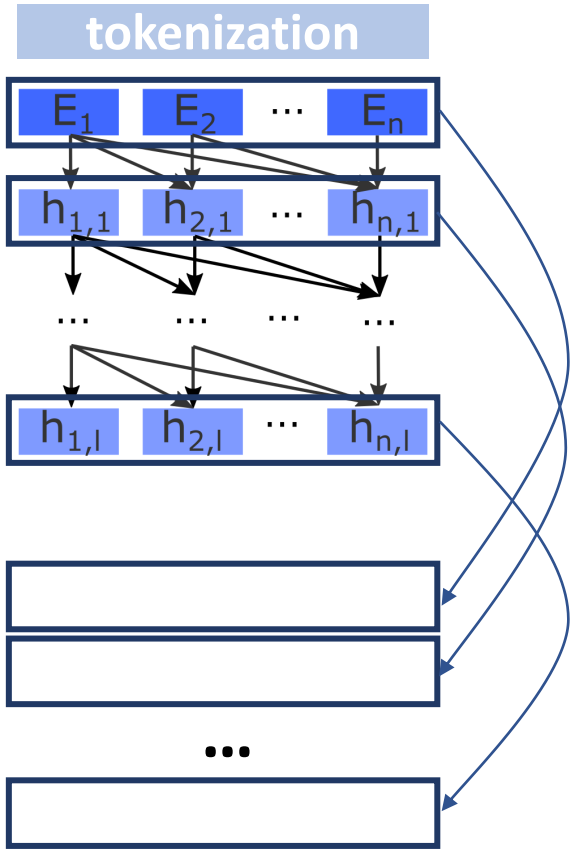


e.g. Pennington et al. 2014 | Jozefowicz et al. 2016 | Vaswani, Shazeer*, Parmar*, Uszkoreit*, Jones*, Gomez*, Kaiser*, Polosukhin* 2017 | Devlin et al. 2018*

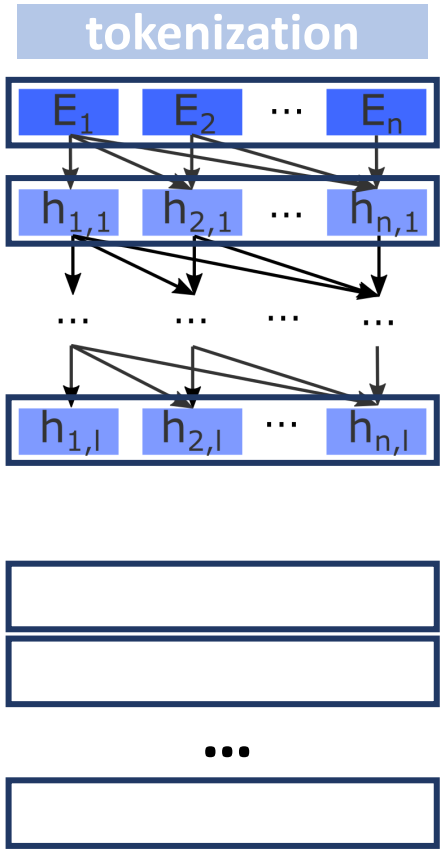
Treating models as experimental subjects



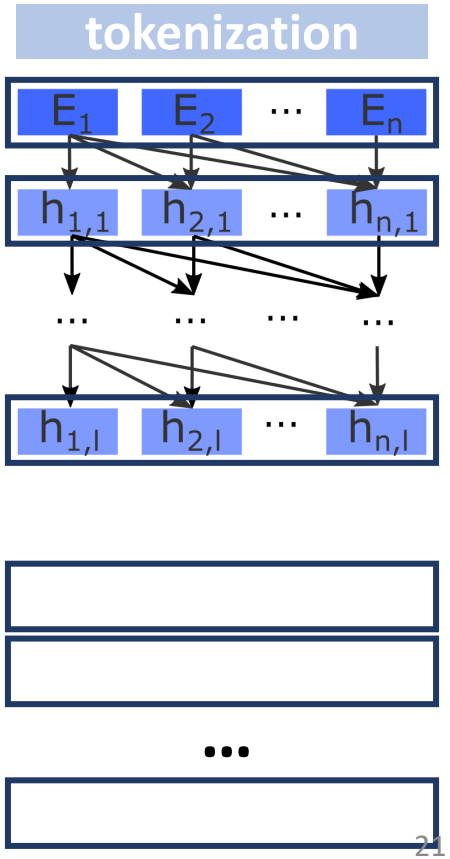
Beekeeping encourages the conservation of local habitats.



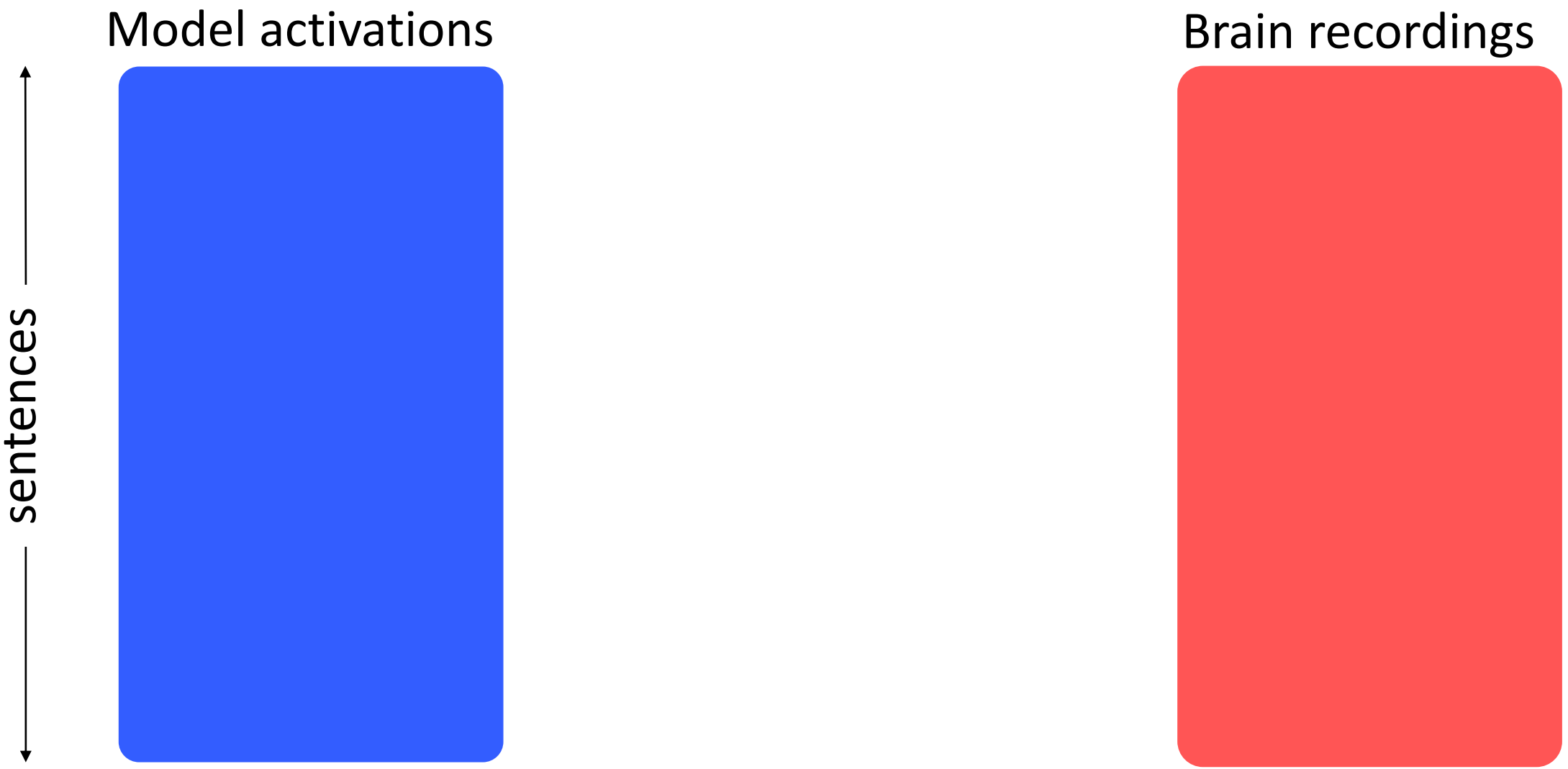
It is in every beekeeper's interest to conserve local plants that produce pollen.



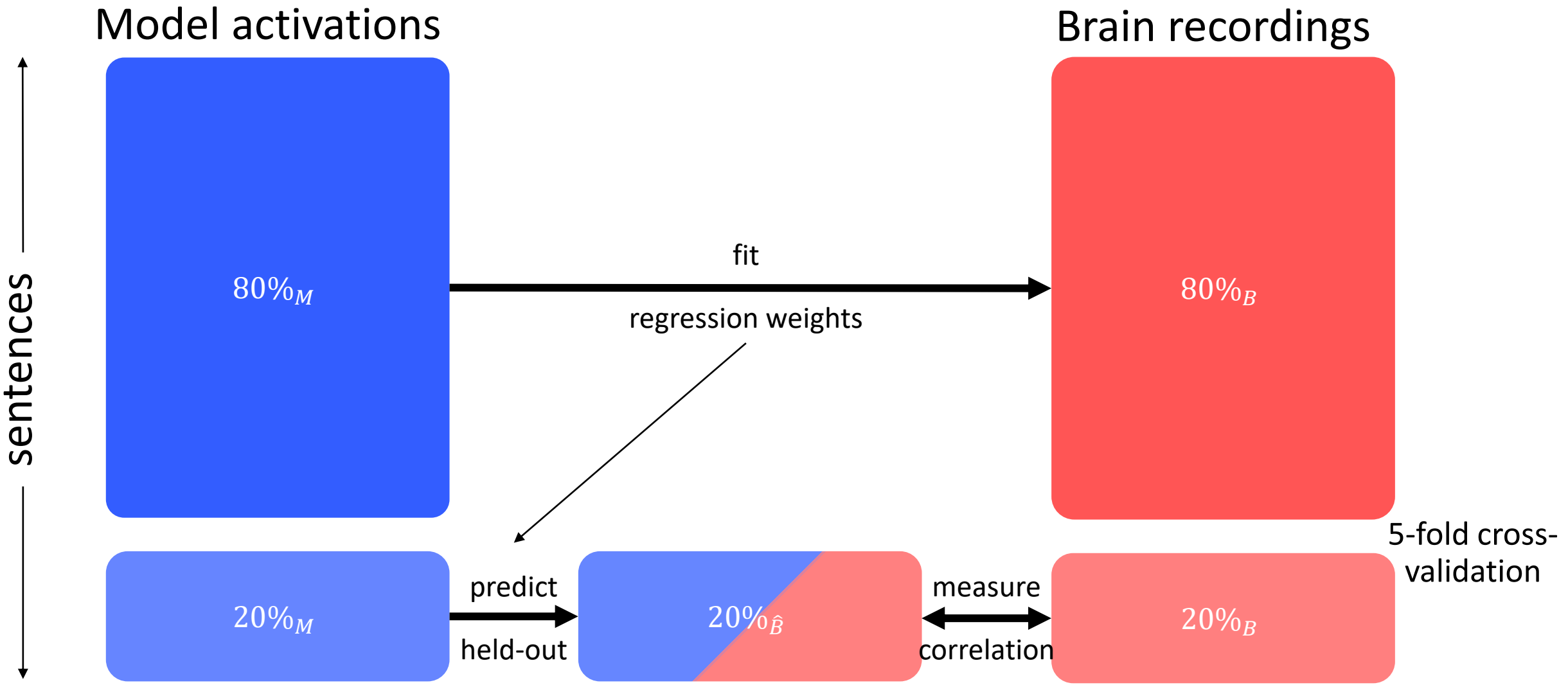
…



Similarity metric: Neural Predictivity / Encoding

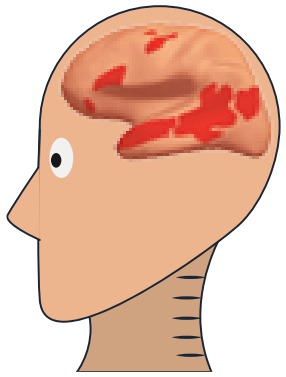


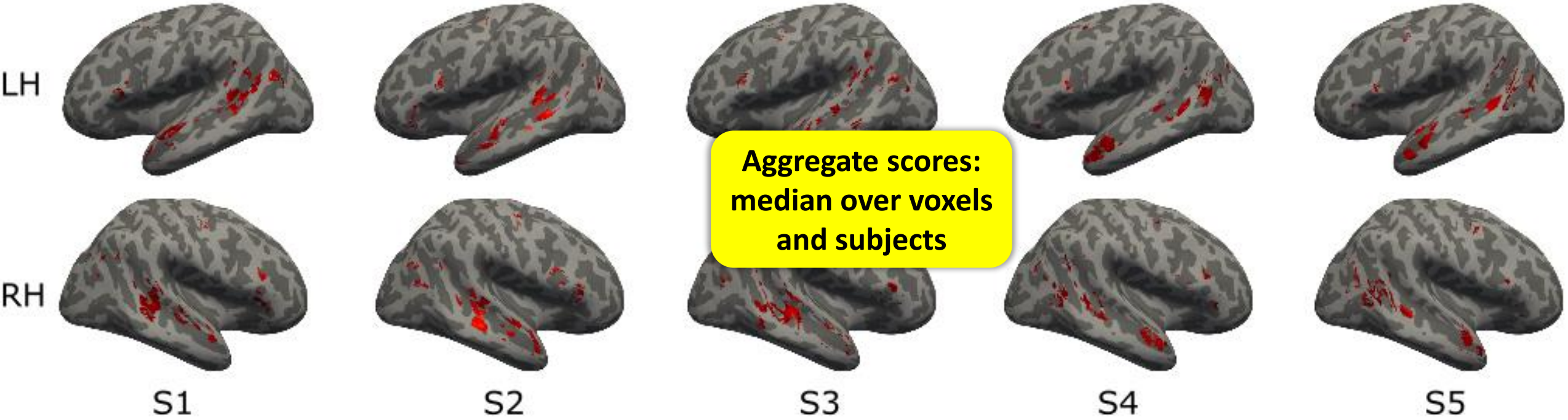
Similarity metric: Neural Predictivity / Encoding



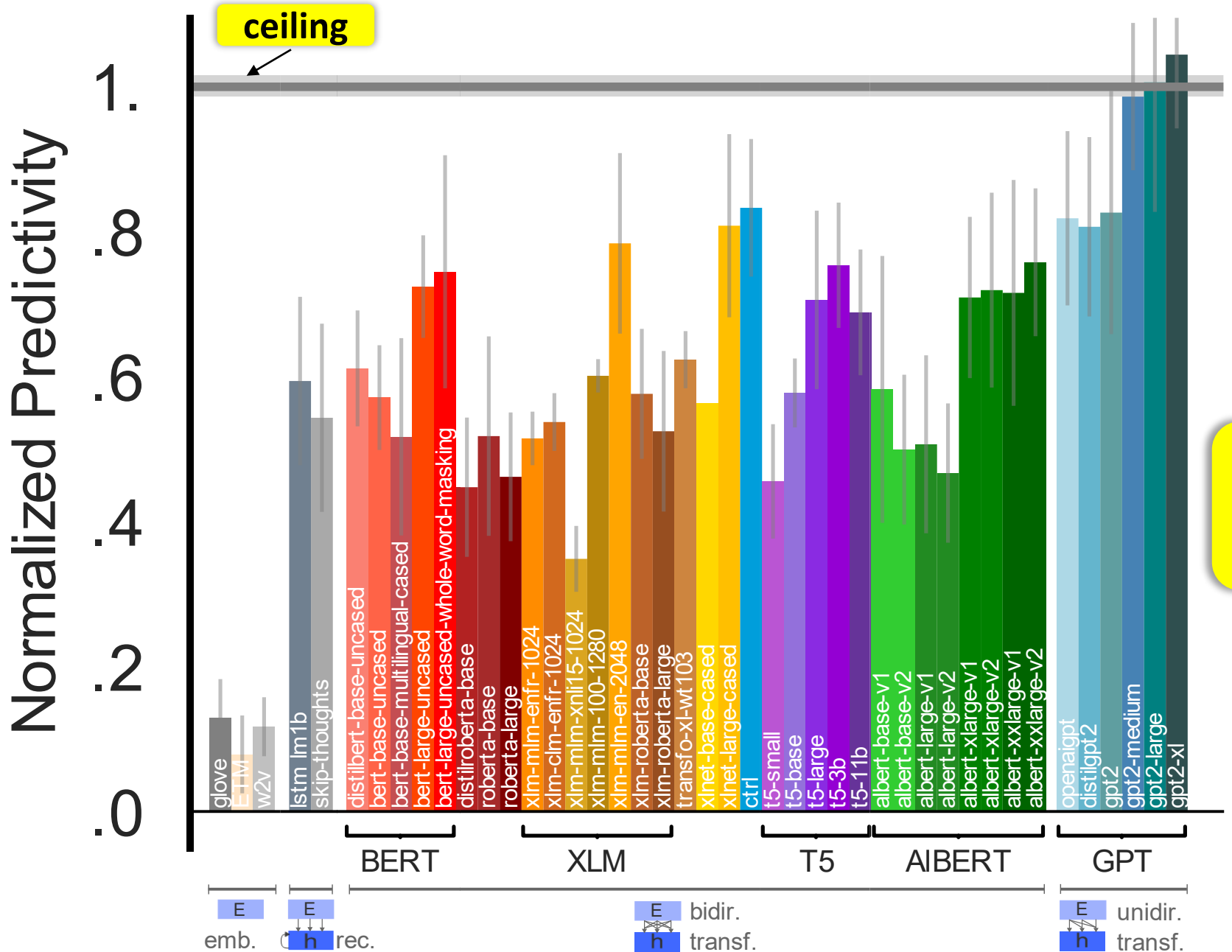
GloVe voxel-wise predictivity scores

GloVe →





Certain language models predict human language recordings

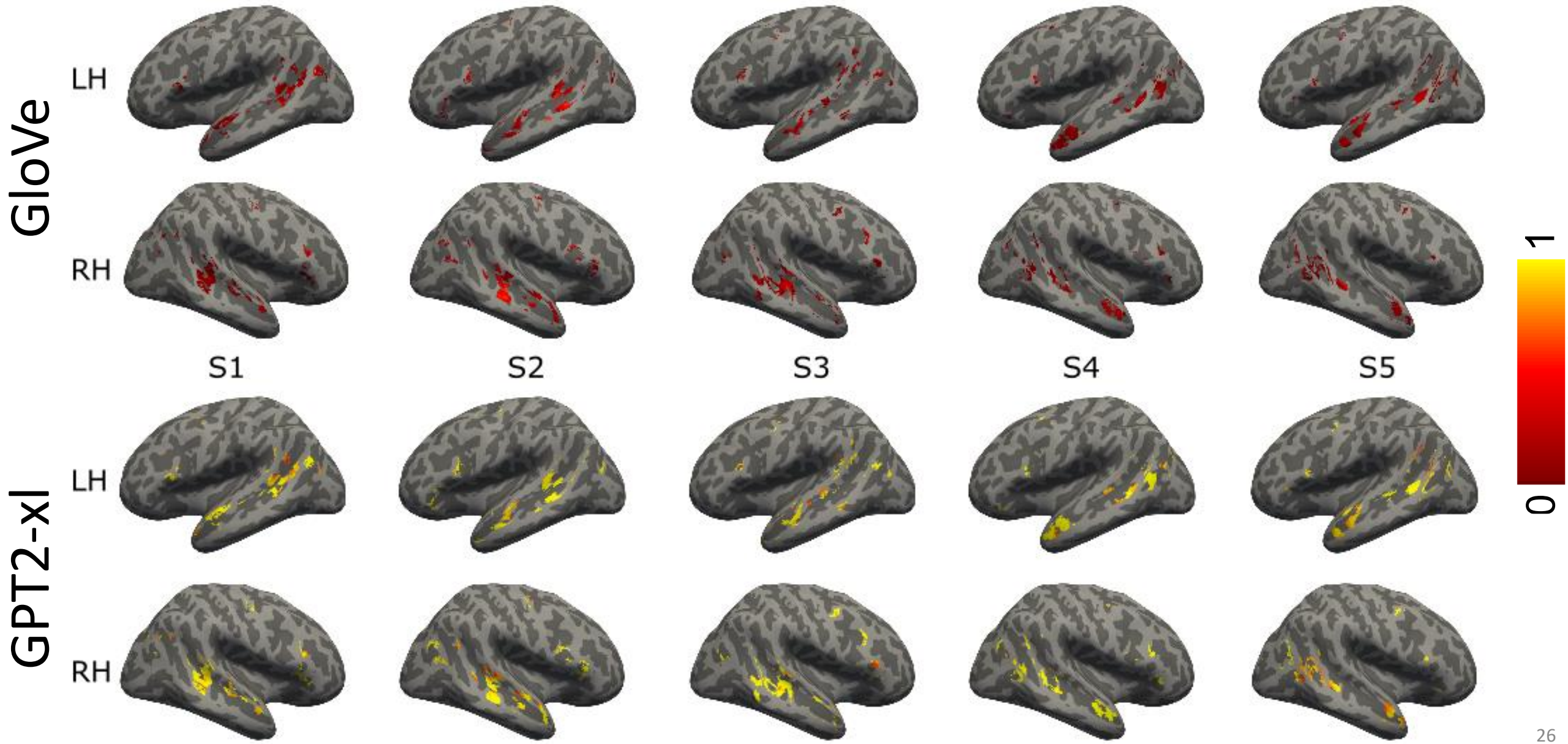


gpt2-xl hits our estimated ceiling for this benchmark

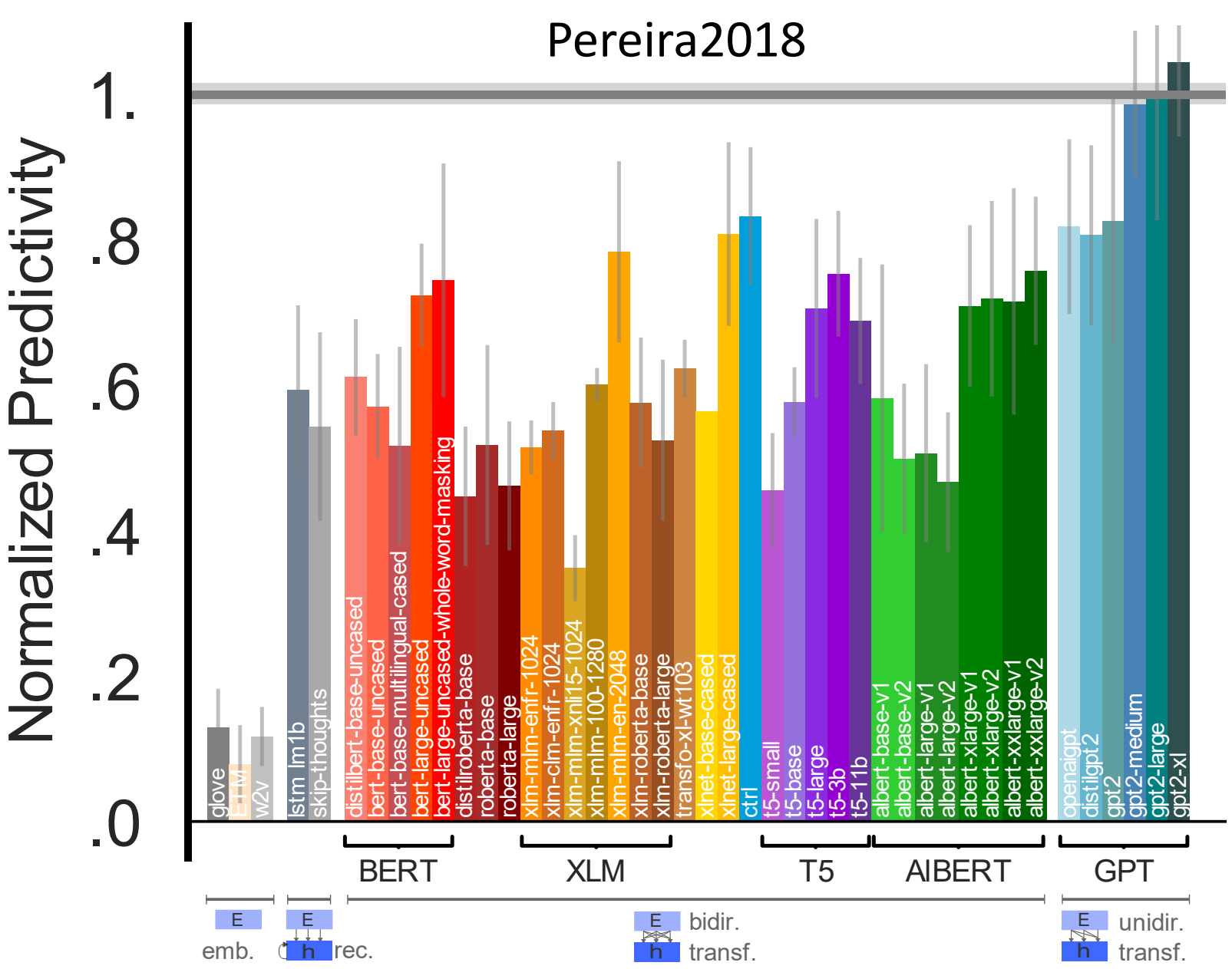
Small differences can lead to very different brain predictivities, warranting a full survey

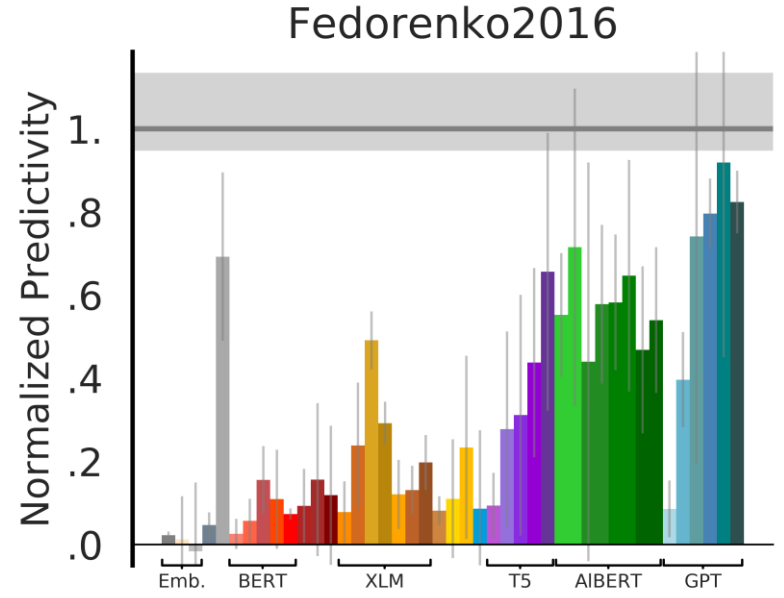
Jain & Huth 2018
Gauthier & Ivanova 2018
Jat et al. 2019
Toneva & Wehbe 2019
Gauthier & Levy 2020
Wang et al. 2020

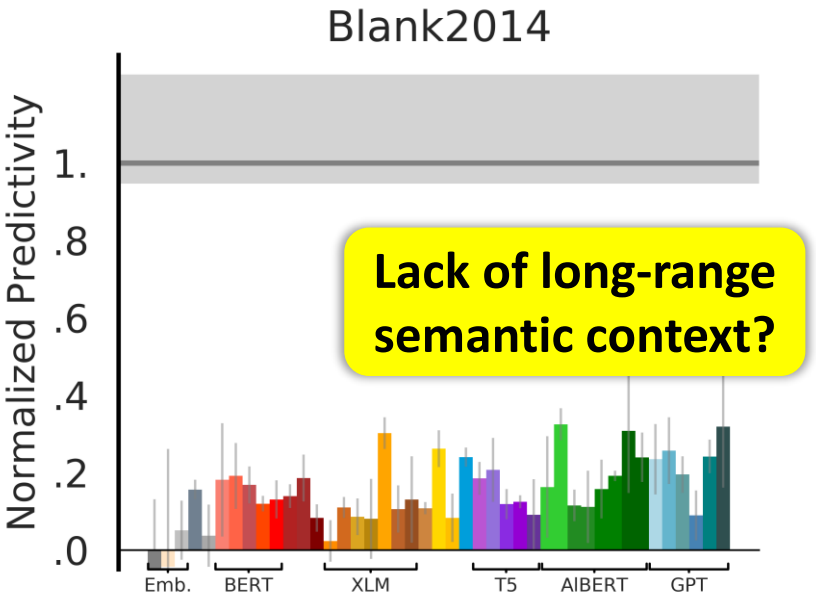
GPT2-xl accurately predicts a large portion of voxels



Language Models predict human language recordings



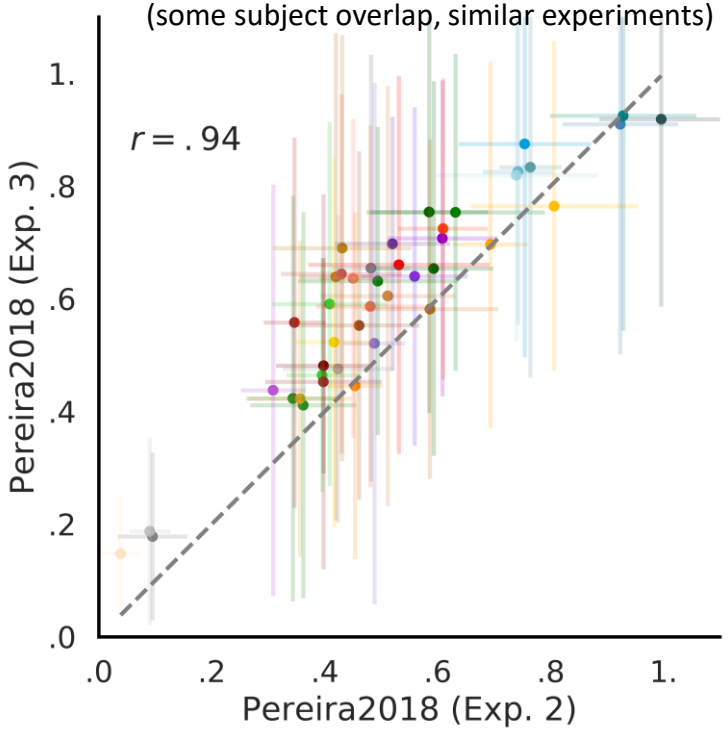




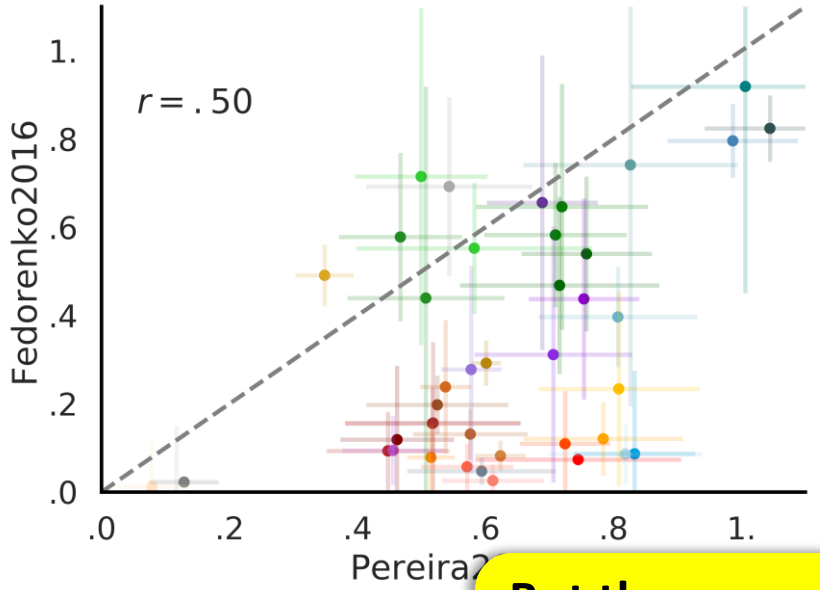
Model scores across benchmarks are correlated, although differences exist

Scores generalize to a good extent

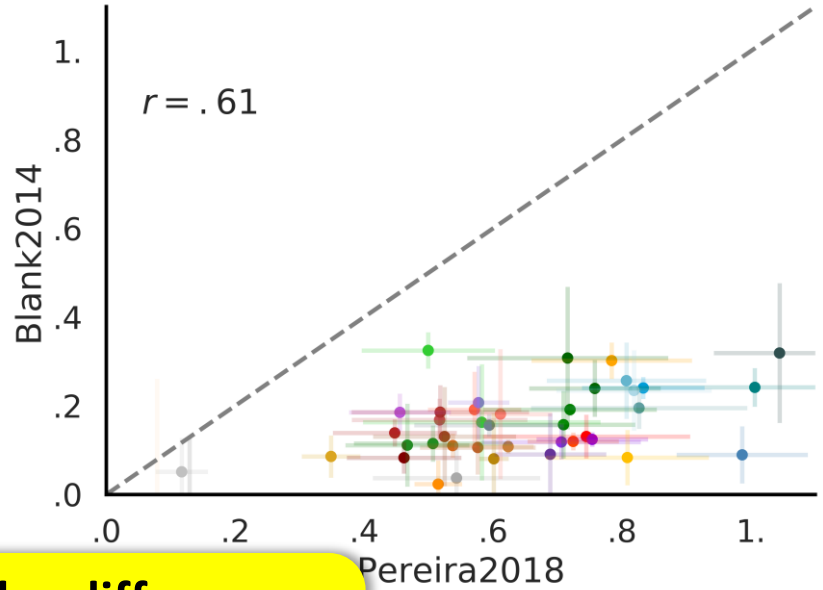
within Pereira2018
(some subject overlap, similar experiments)



Pereira2018 → Fedorenko2016

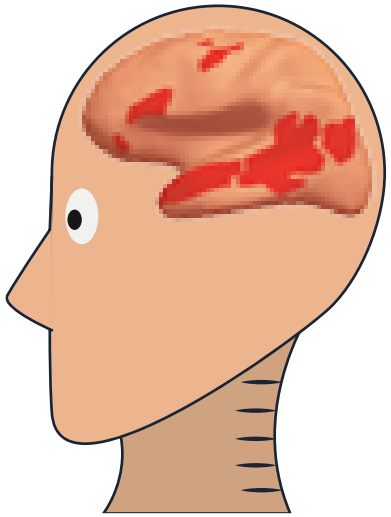


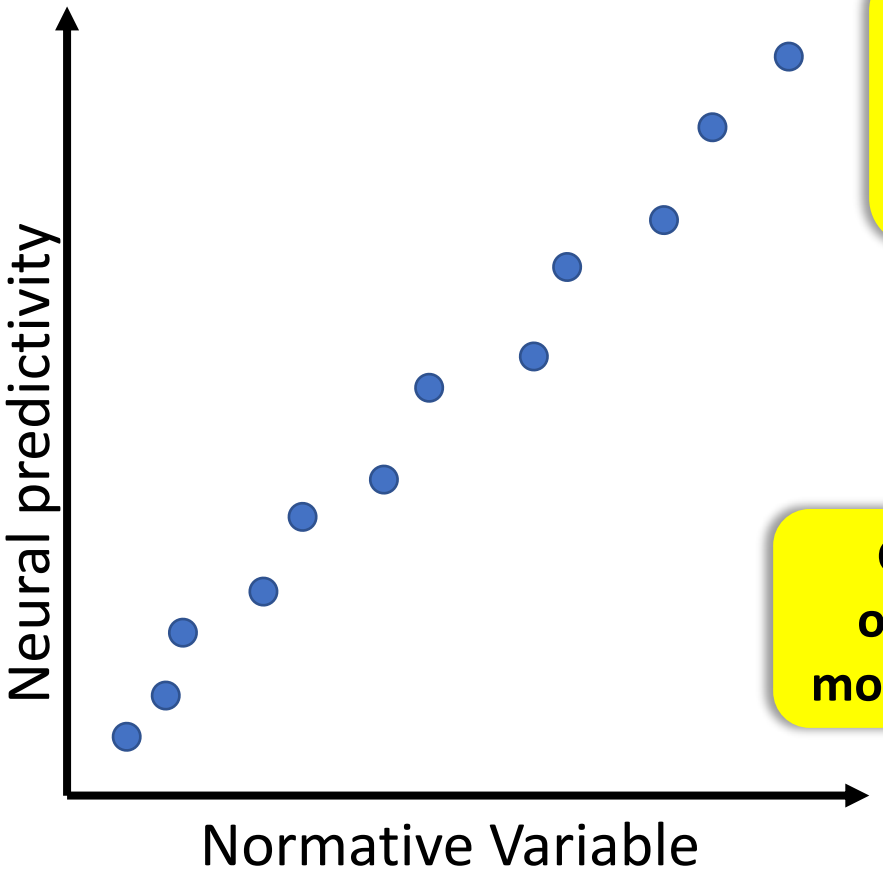
Pereira2018 → Blank2014



But there are also differences, making each individual benchmark valuable

What explains the model differences?





Goal 1: possible explanation *why* some models are better than others (hinting at optimization in the brain)

Goal 2: if x-axis is easier to optimize than y-axis, we can more *efficient*ly improve models

Next-Word Prediction on WikiText-2

= Gold dollar =
The gold dollar or gold one @-@ dollar piece was a coin struck as a regular issue by the United States Bureau of the Mint from 1849 to 1889 . The coin had three types over its lifetime , all designed by Mint Chief Engraver James B. Longacre . The Type 1 issue had
…

	WikiText-2		
	Train	**Valid**	**Test**
Articles	600	60	60
Tokens	2,088,628	217,646	245,569
Vocab	33,278		
OoV	2.6%		

Alaska

Alaska is

Alaska is about

Alaska is about twelve

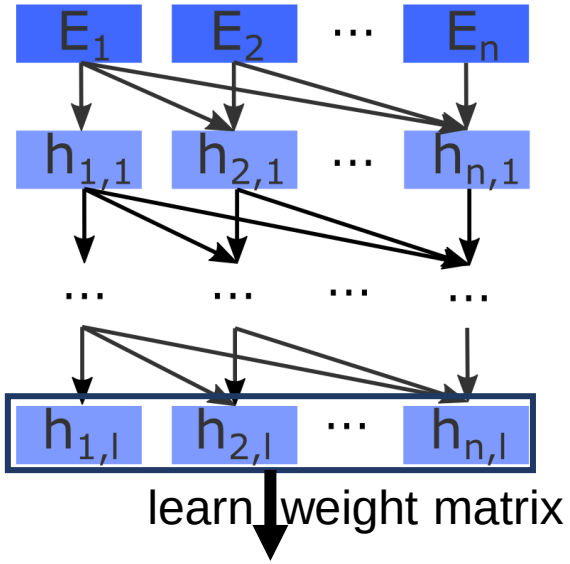
Alaska is about twelve times

Alaska is about twelve times larger

Alaska is about twelve times larger than

Alaska is about twelve times larger than New

Alaska is about twelve times larger than New York

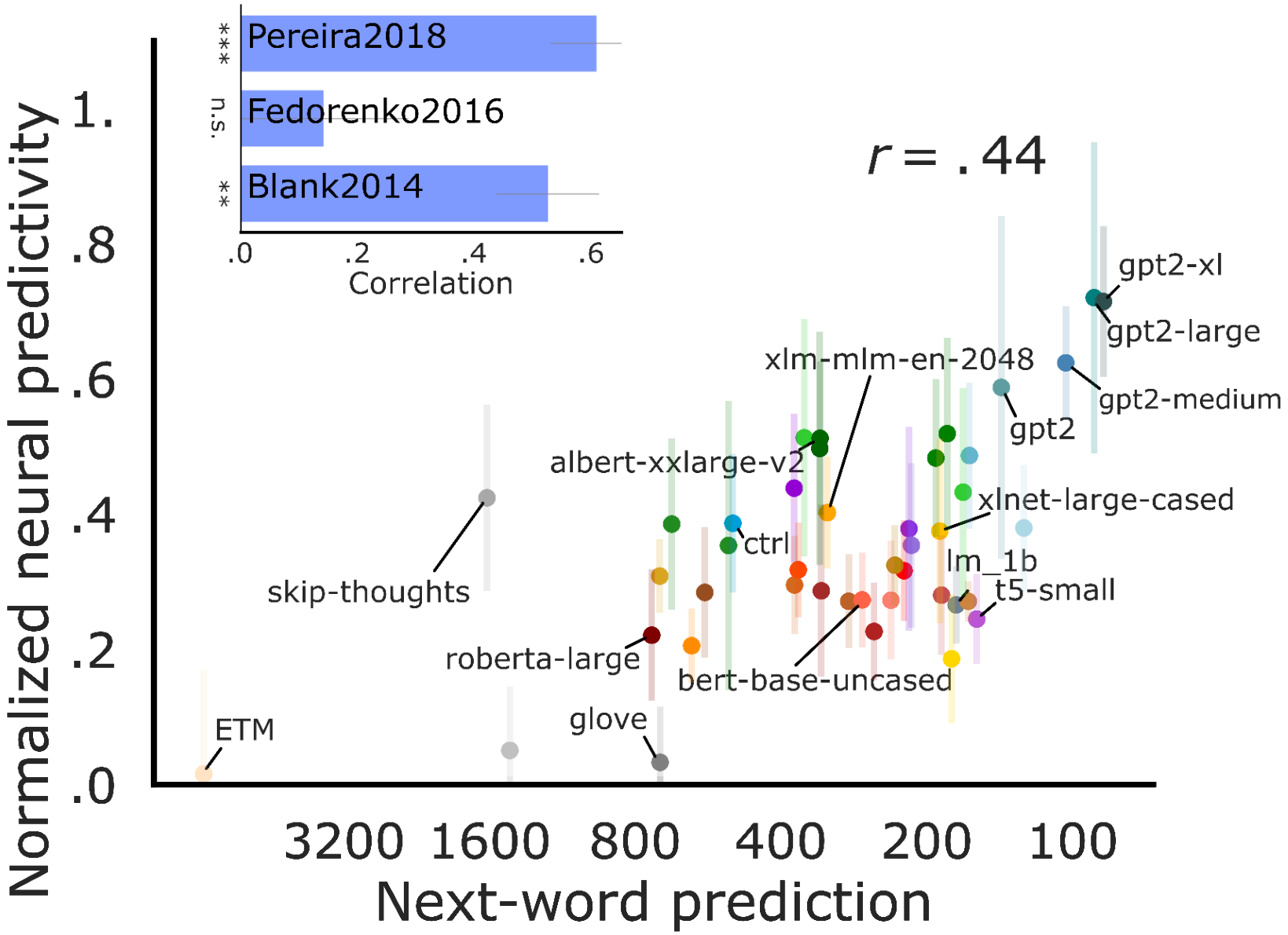


afternoon | alaska | animation | article | …

Surprisal of seeing actual next word:
perplexity = exp(NLL Loss)

Merity et al. 2016

Next-Word Prediction performance correlates with neural predictivity



What about other language tasks?

9 “General Language Understanding Evaluation” tasks:

Sentence grammaticality (CoLa)

Sentence sentiment (SST-2)

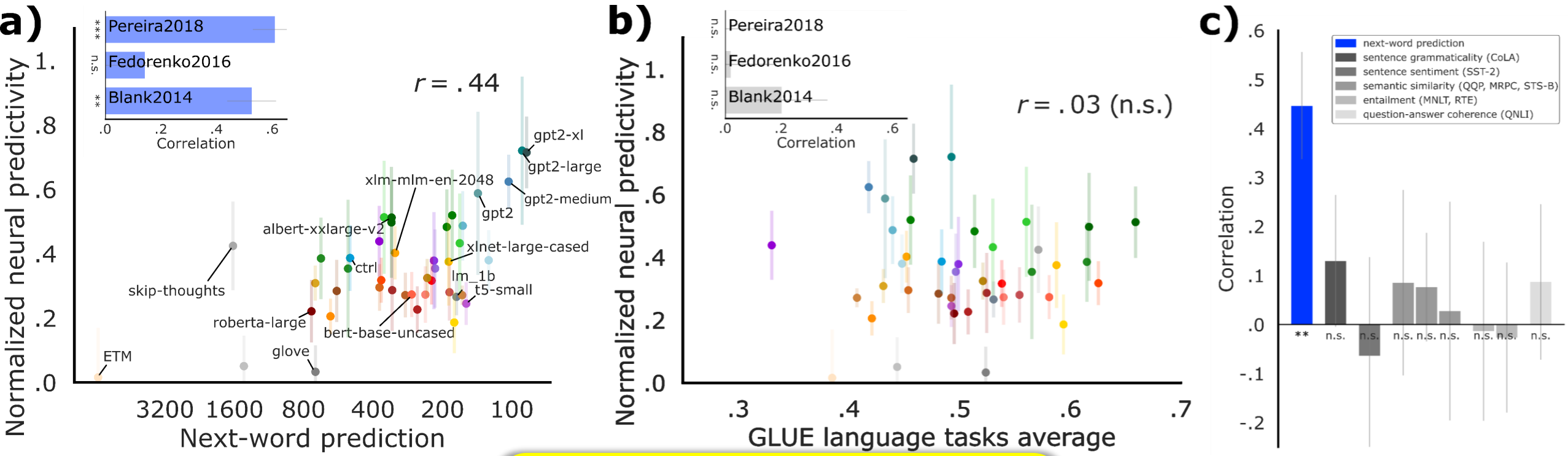
Semantic similarity (QQP, MRPC, STS-B)

Entailment (MNLT, RTE)

Question-answer coherence (QNLI)

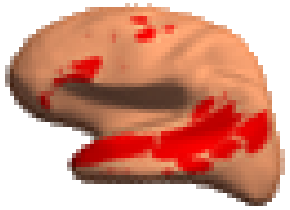
Winograd (WNLI; ignored due to known issues)

Next-Word Prediction performance **selectively** correlates with neural predictivity

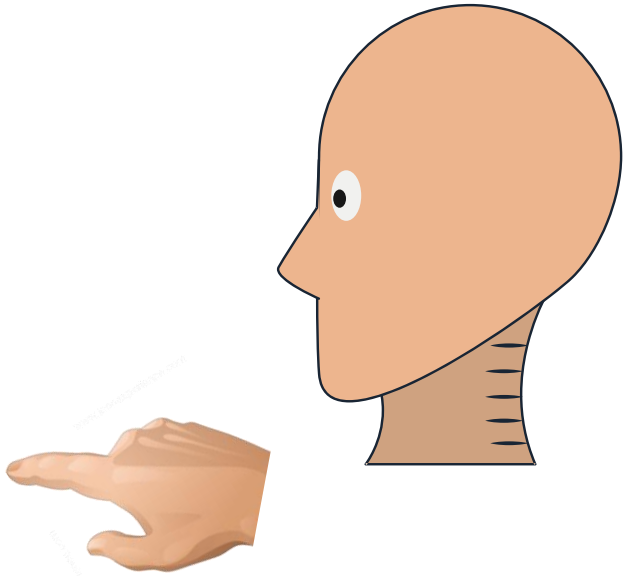


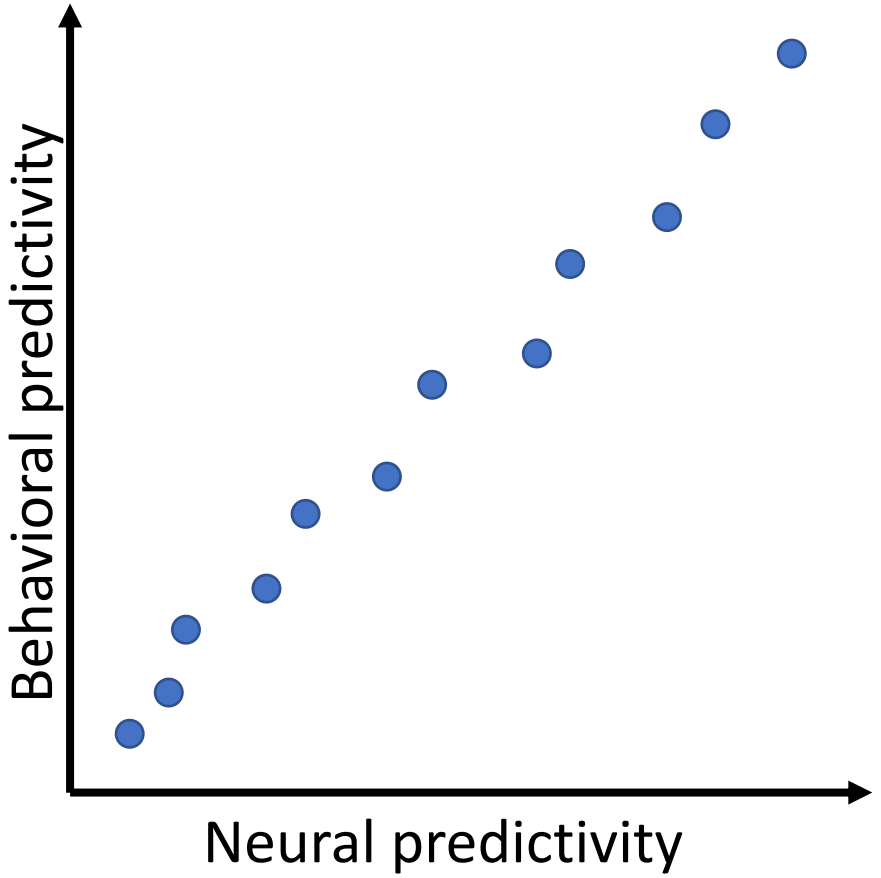
Online prediction may fundamentally shape language processing in the brain

Is any of this behaviorally relevant?



Is any of this behaviorally relevant?



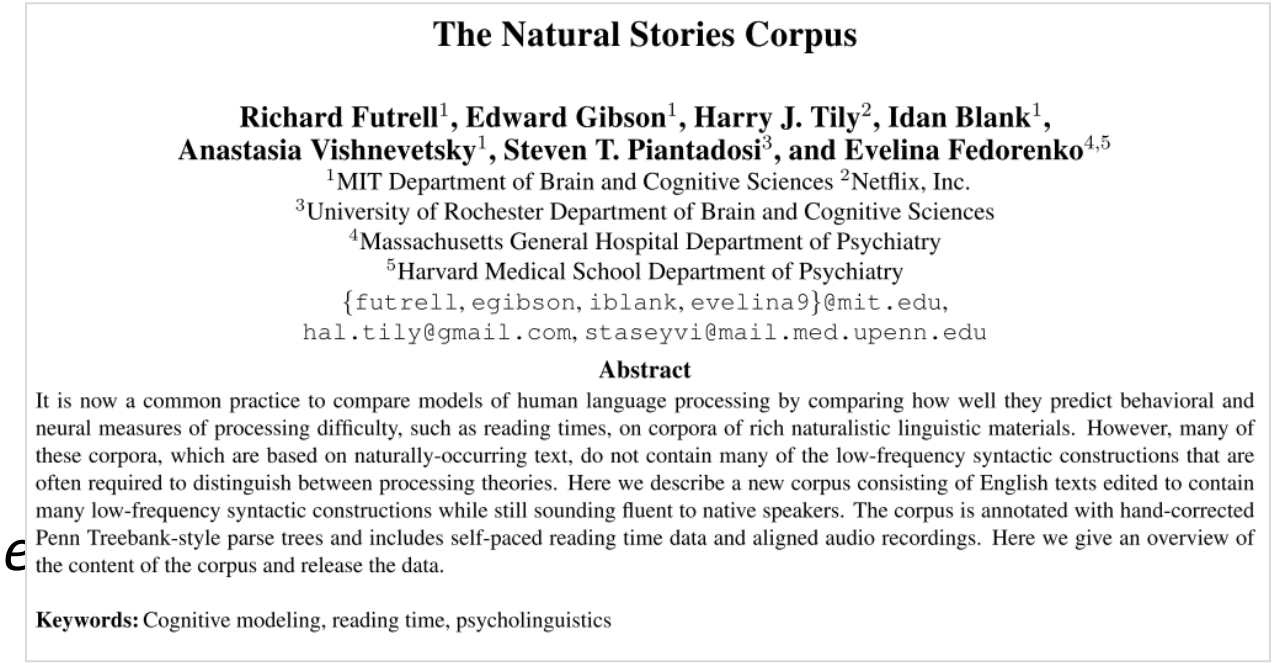


Behavioral target: human reading times

Futrell et al. 2018

10256 words x 179 subjects

If | you | were | to | journey | to | the | North | of | England, | you | would | come | to | a | valley | that | is | surrounded | by | moors | as | high | as | mountains. | It | is | in | this | valley | where | you | would | find | the | city | of | Bradford, | where | once | a | thousand | spinning | …



The Natural Stories Corpus

Richard Futrell[1], **Edward Gibson**[1], **Harry J. Tily**[2], **Idan Blank**[1], **Anastasia Vishnevetsky**[1], **Steven T. Piantadosi**[3], **and Evelina Fedorenko**[4,5]
[1]MIT Department of Brain and Cognitive Sciences [2]Netflix, Inc.
[3]University of Rochester Department of Brain and Cognitive Sciences
[4]Massachusetts General Hospital Department of Psychiatry
[5]Harvard Medical School Department of Psychiatry
{futrell, egibson, iblank, evelina9}@mit.edu,
hal.tily@gmail.com, staseyvi@mail.med.upenn.edu

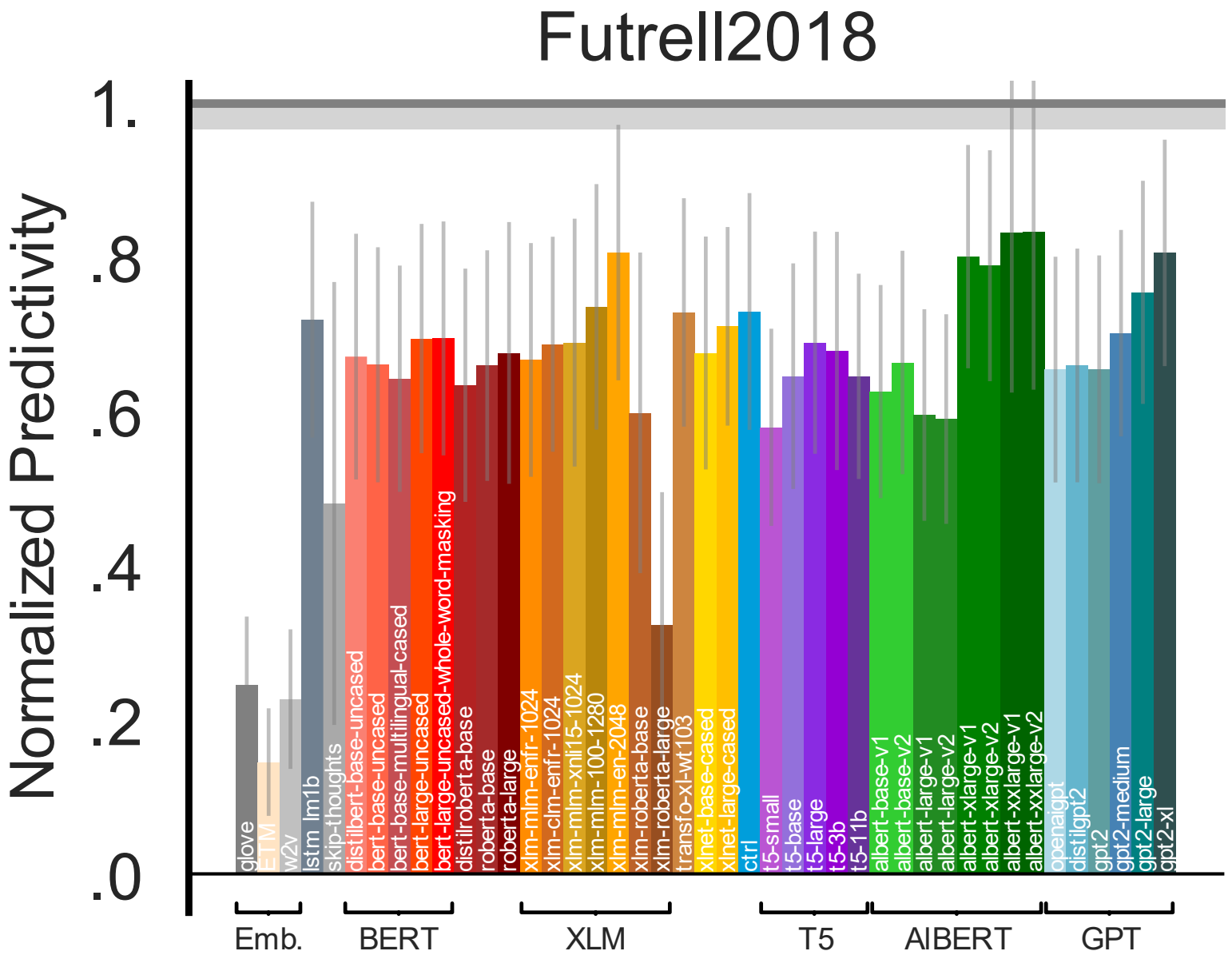
Abstract

It is now a common practice to compare models of human language processing by comparing how well they predict behavioral and neural measures of processing difficulty, such as reading times, on corpora of rich naturalistic linguistic materials. However, many of these corpora, which are based on naturally-occurring text, do not contain many of the low-frequency syntactic constructions that are often required to distinguish between processing theories. Here we describe a new corpus consisting of English texts edited to contain many low-frequency syntactic constructions while still sounding fluent to native speakers. The corpus is annotated with hand-corrected Penn Treebank-style parse trees and includes self-paced reading time data and aligned audio recordings. Here we give an overview of the content of the corpus and release the data.

Keywords: Cognitive modeling, reading time, psycholinguistics

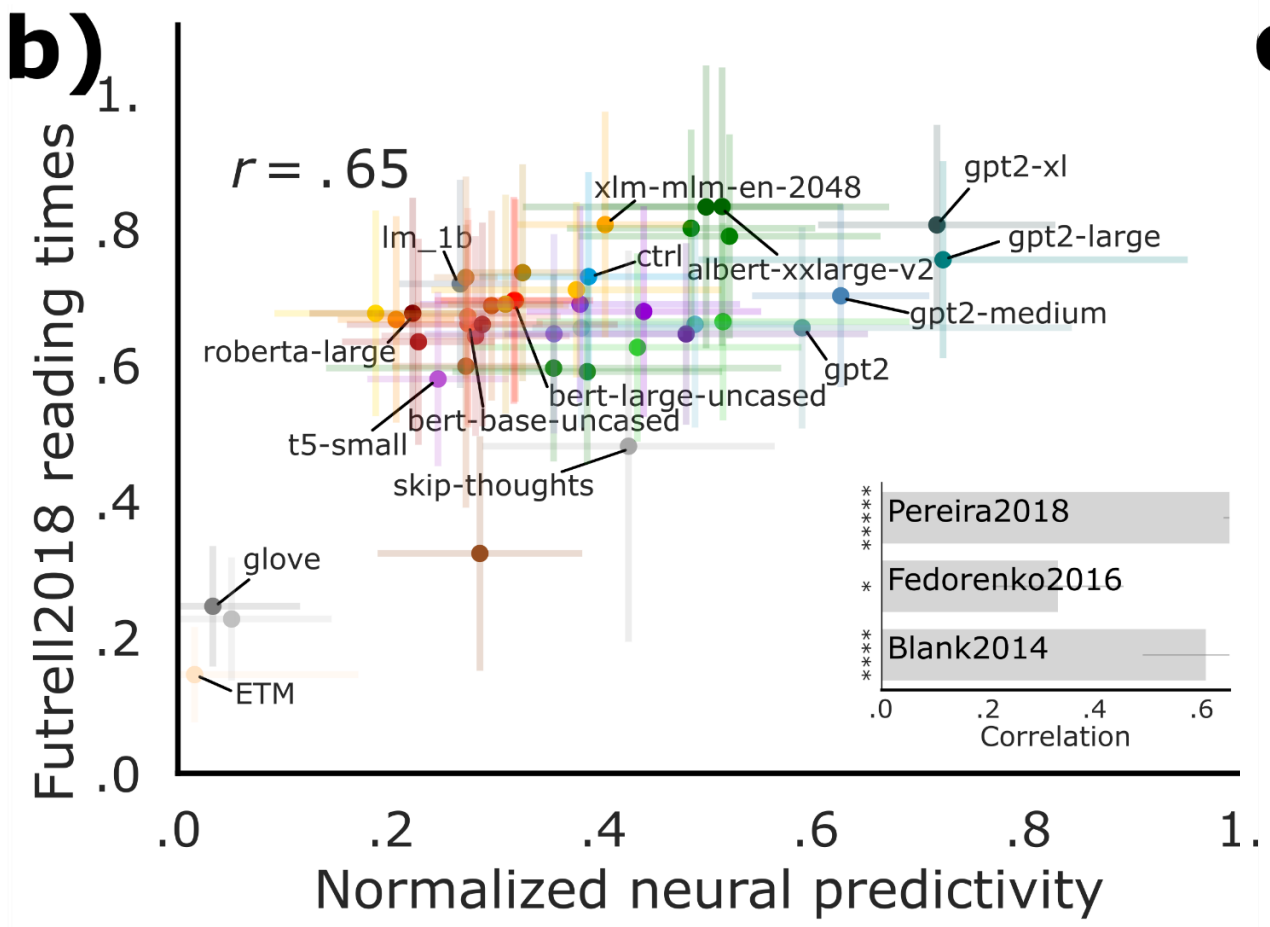
Treat reading times as representation target

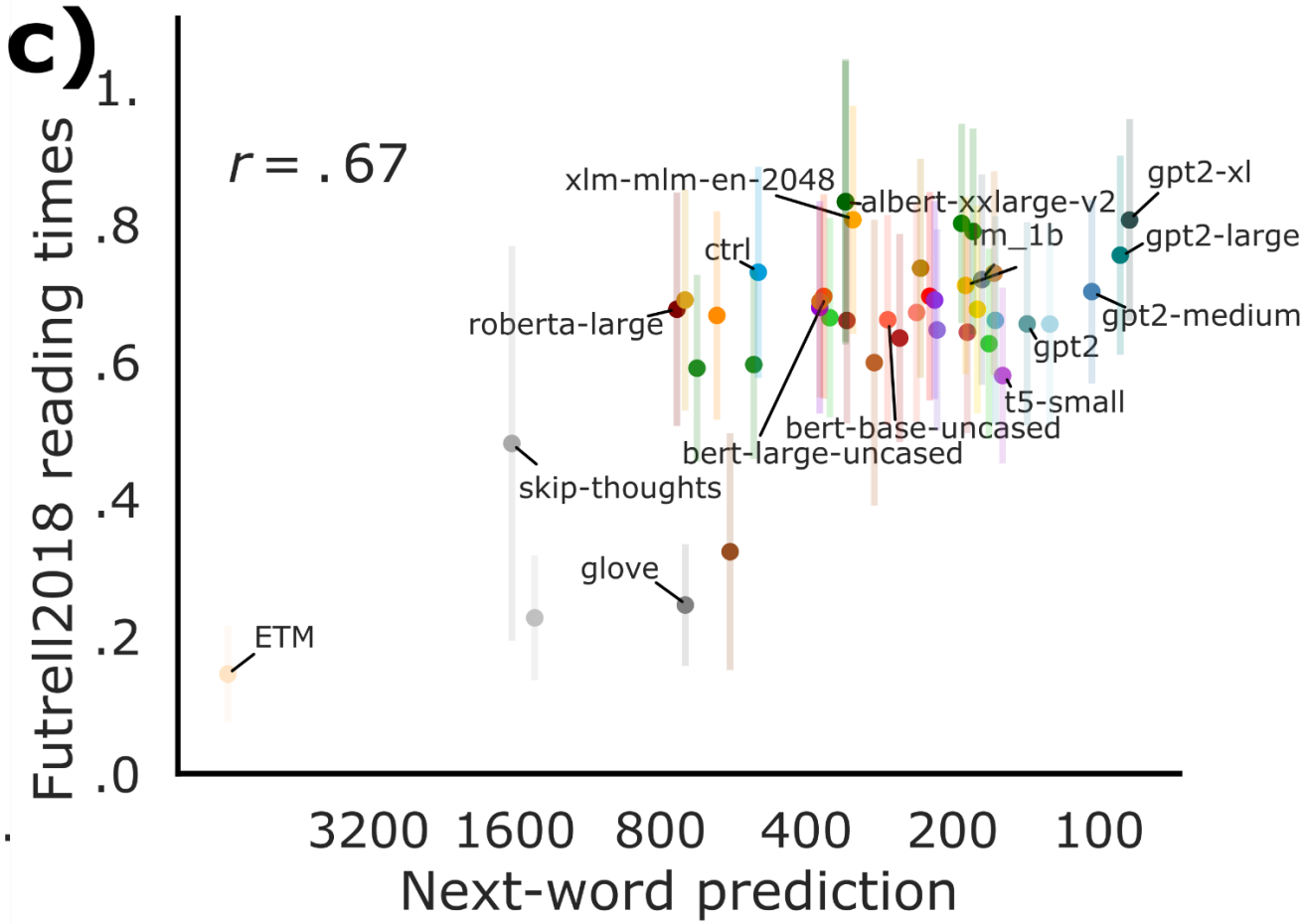
Behavioral scores



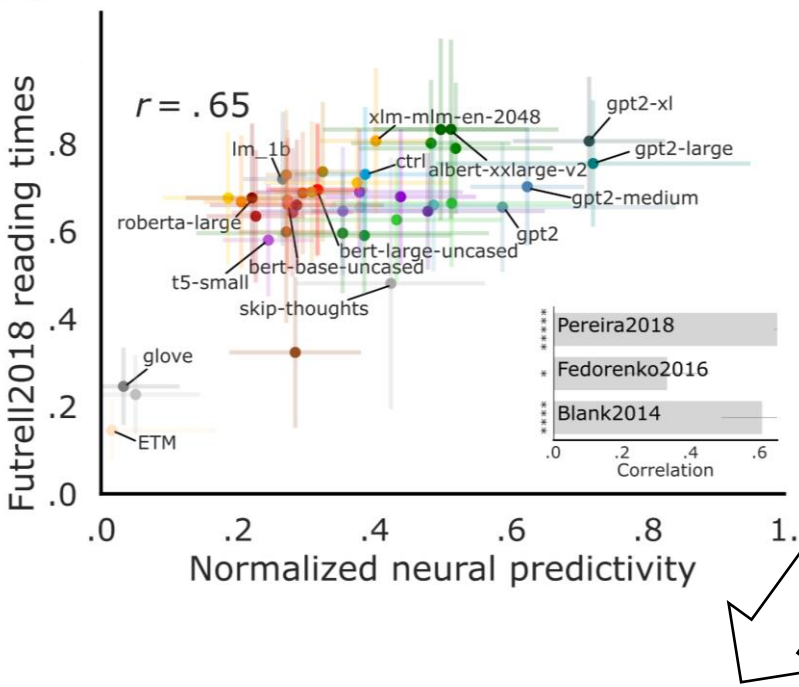
Neural scores correlate with Behavioral scores

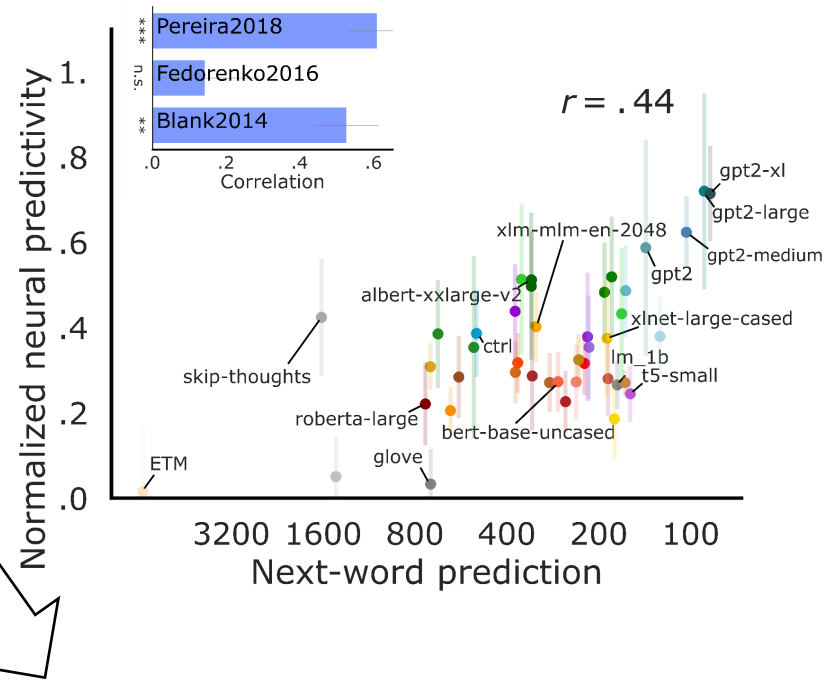
Task scores correlate with Behavioral scores





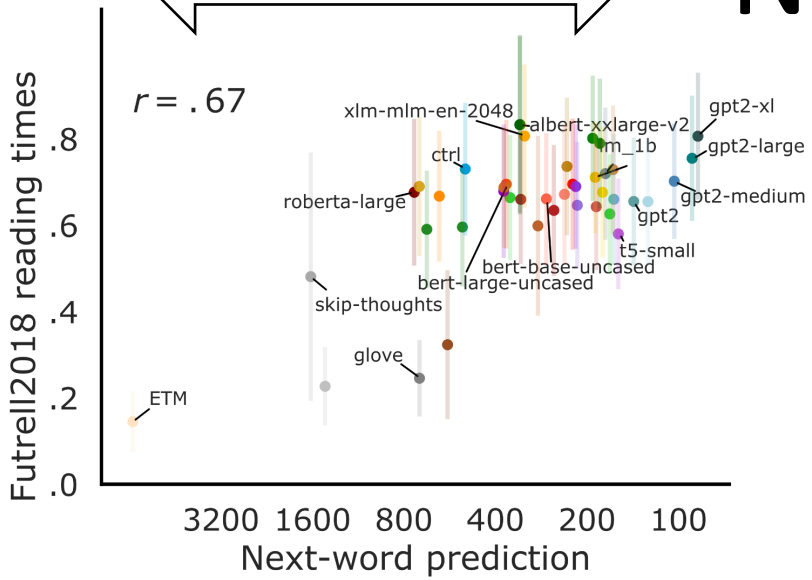
Neural



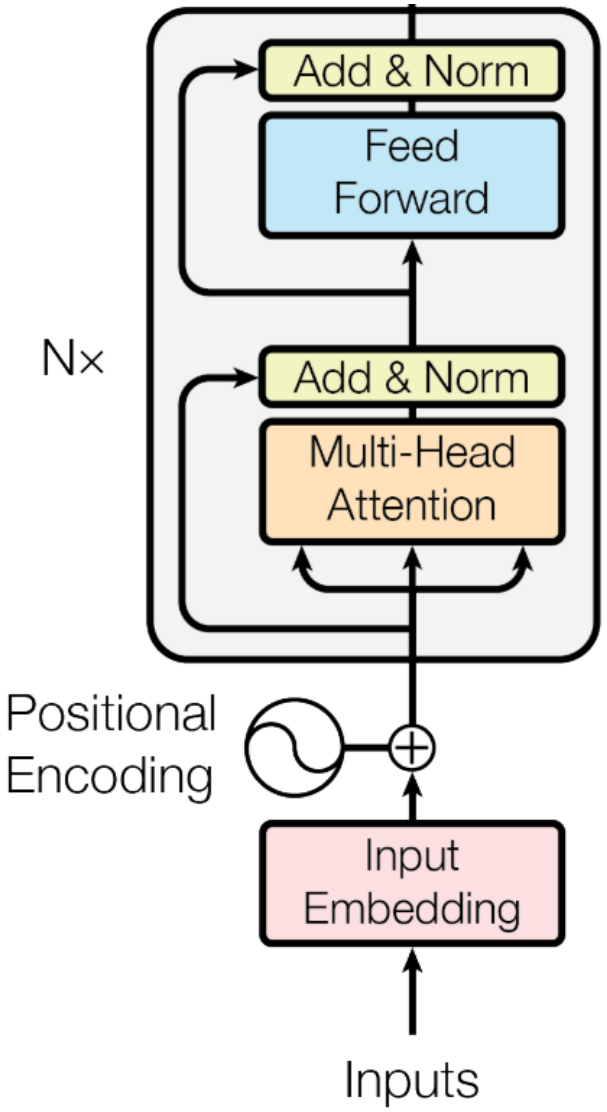


Behavioral

Normative Task



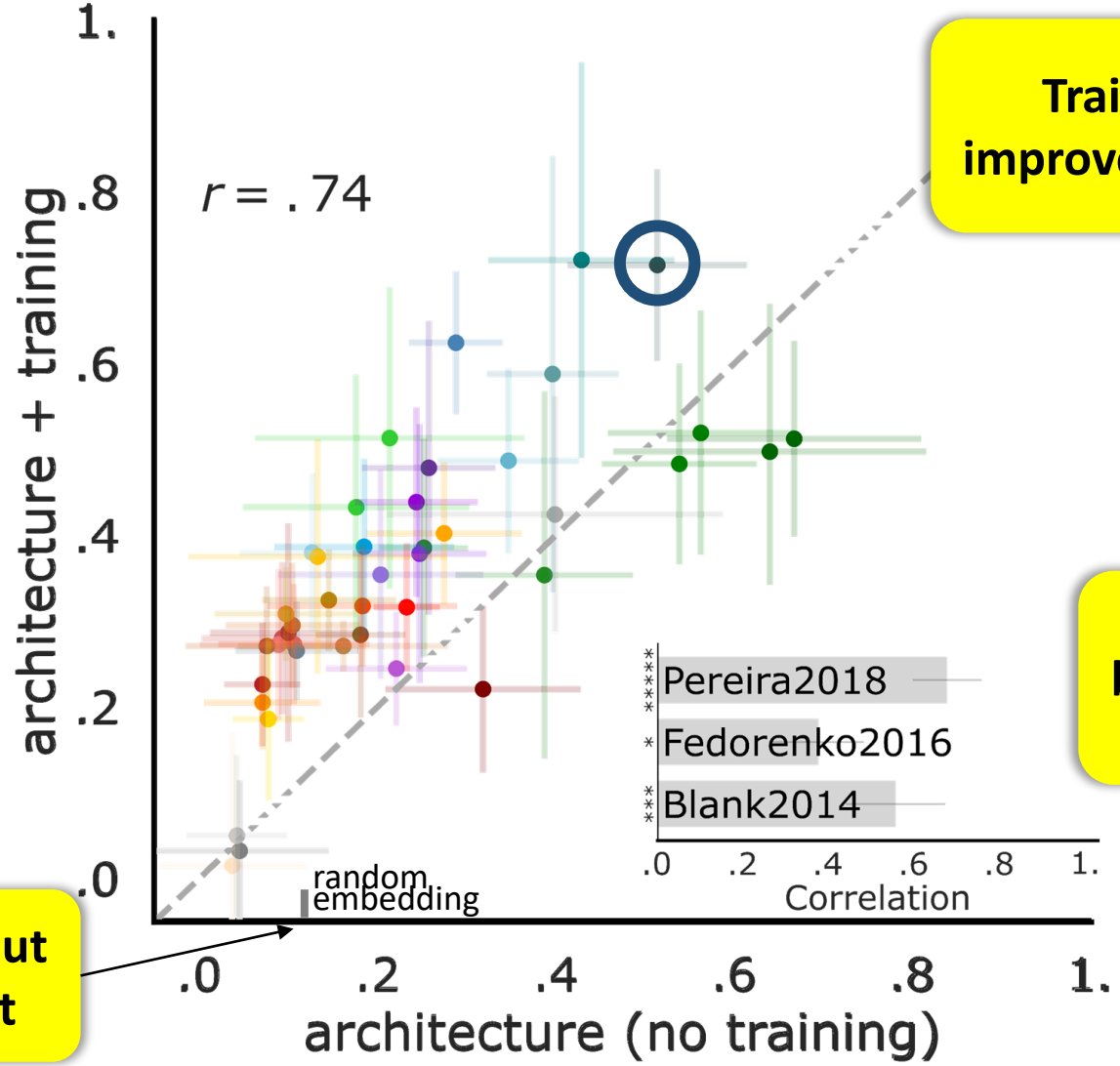
What is the relative importance of evolutionary and learning-based optimization?



Evolution ≃ community optimization over architectural properties

Experience-dependent learning ≃ updating of weights over training

Architecture substantially contributes to models' brain predictivity



Training generally improves scores by ~53%

Inherent structure might be a key driver of brain-like language representations

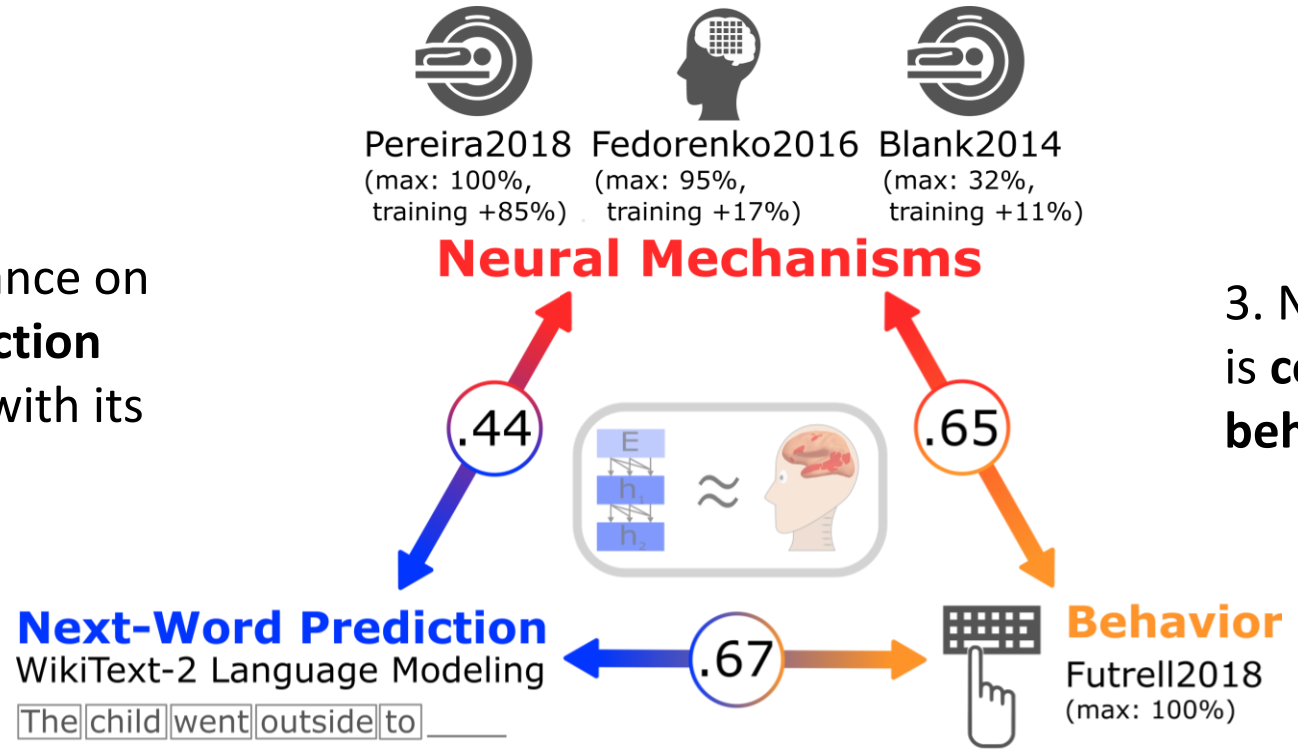
Large feature size without structure is insufficient

Summary

Certain artificial neural networks are viable **mechanistic hypotheses** for how **predictive language processing** is implemented in human neural tissue.

1. Across three datasets, specific models such as gpt2-xl **consistently predict human recordings**

2. Model performance on a **next-word prediction task is correlated** with its brain-predictivity



3. Neural-predictivity is **correlated with behavior-predictivity**

4. **Architecture** substantially contributes to brain-predictive representations