LINKING ARTIFICIAL AND BIOLOGICAL NEURAL NETWORKS II

Prof. Alexander Huth

4.9.2020

RECAP

* **System construction** to **system identification**

  * First, build a system that solves a task

  * Second, collect data from brain that is solving the same task

  * Third, use representations from the system you built to predict responses in the

RECAP

* In the context of vision
  * Use convolutional neural network (e.g. AlexNet, OverFeat) as **linearizing transform**

* i.e. use **activations** from different layers of convolutional neural network as **features** of the input images for system identification

RECAP

* From fMRI paper (Eickenberg et al.)
  * Features/representations extracted from **different layers** of the artificial network are best for predicting **different areas** in visual cortex
* The **hierarchy** matches: early layers in network are best for “early” visual cortex

TODAY

* Continuation of vision w/ another example (Yamins et al.)

* Discussion of “task” / “computational goal”

* Lead-in to language (Jain & Huth, 2018)

CONSTRUCTION TO IDENTIFICATION

* Key questions about Eickenberg et al. scheme
  * Does this work at all? ***YES***
    * If so, is this trivial? ***HMM…***

* Do activations from different layers predict different parts of the brain? ***YES***
  * Does this correspond to known computational hierarchies? ***YES***
  * If so, is *this* trivial? ***HMM…***

NEURAL EXAMPLE

* Yamins, Hong, Cadieu, Solomon, Seibert, & DiCarlo (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*

NEURAL EXAMPLE

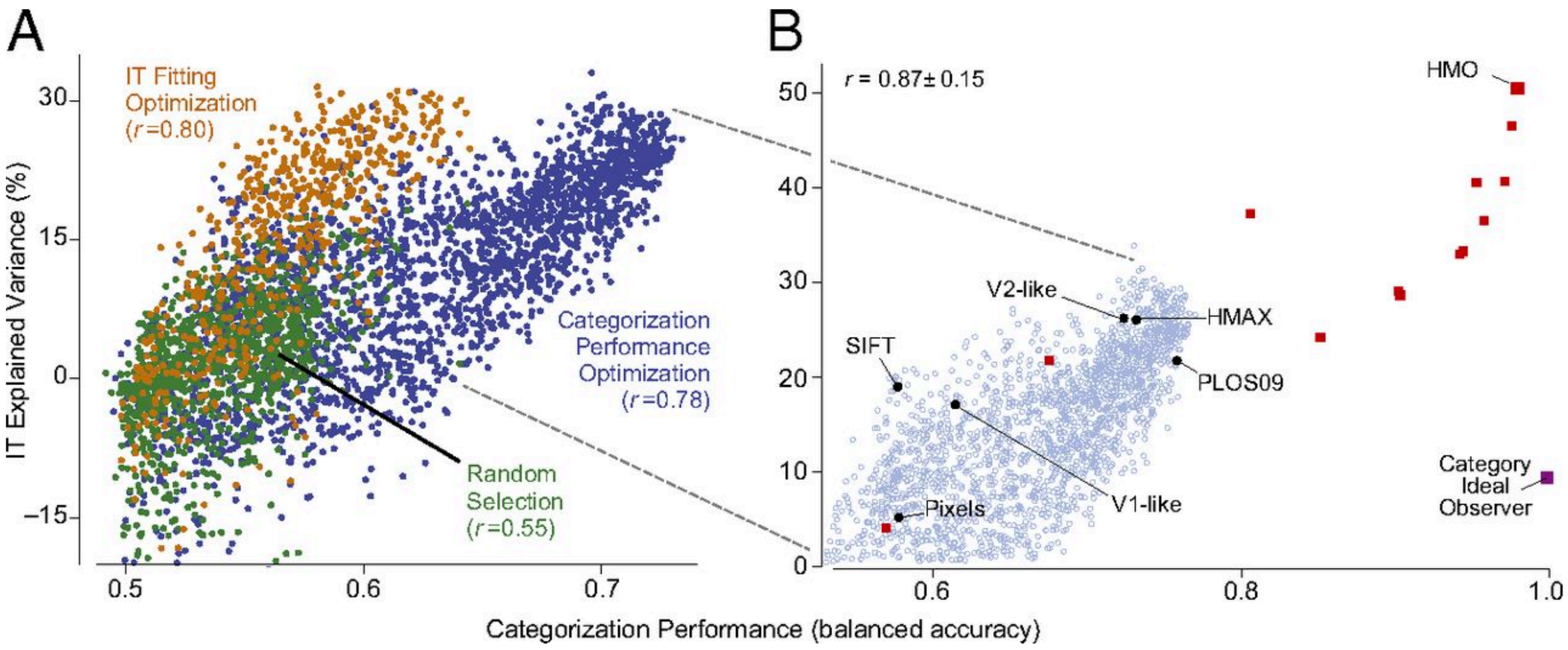
* Neural recordings from areas V4 and IT (inferotemporal cortex) in macaques while they view naturalistic images

* Hierarchical modular optimization (HMO) models (CNN, ~equivalent to AlexNet) used to extract features

NEURAL EXAMPLE

* **Question 1:** Are model performance at object recognition & model performance at predicting brain data correlated?

NEURAL EXAMPLE



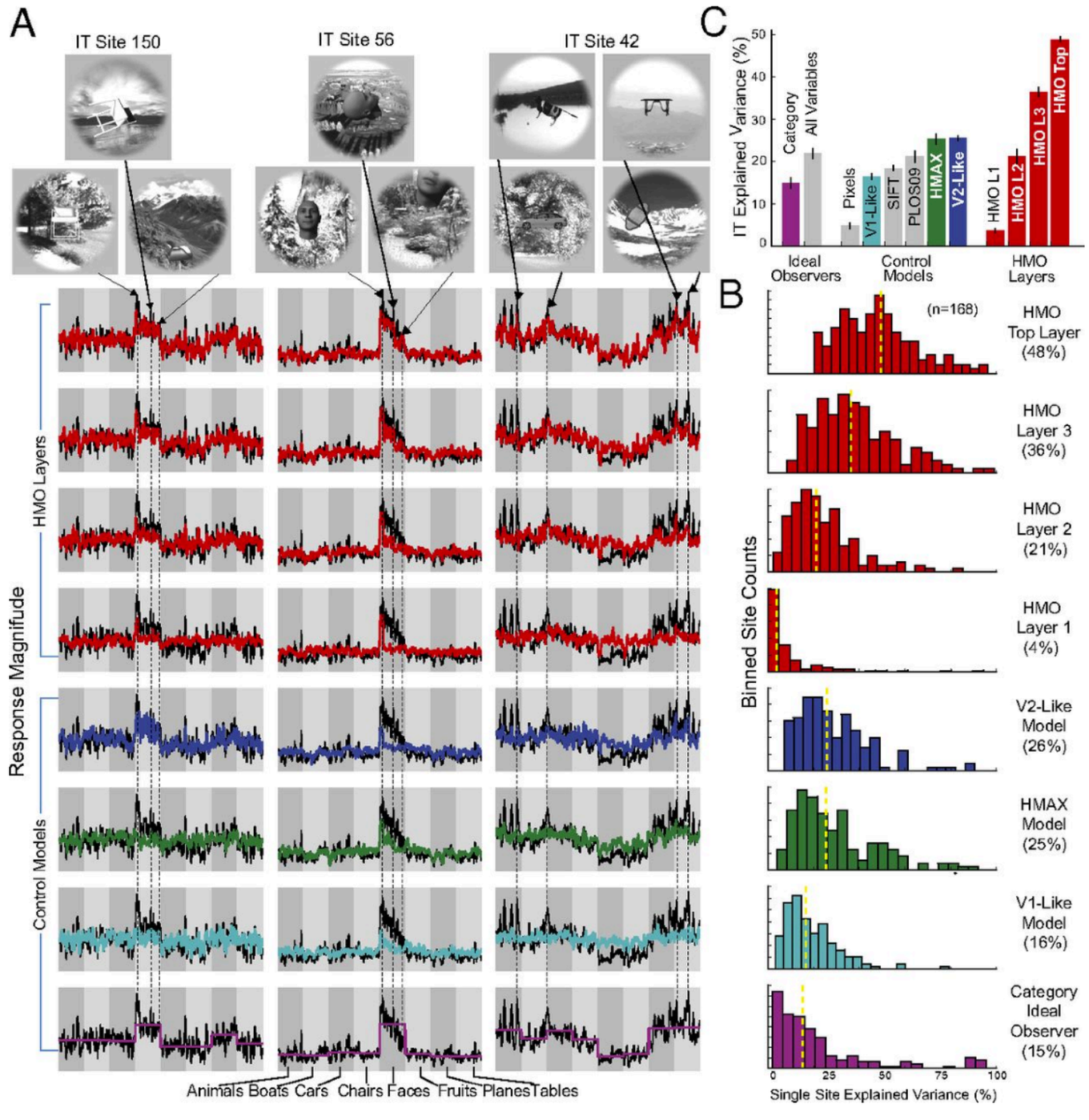
* **Answer 1:** Yes, broadly!

NEURAL EXAMPLE

* **Question 2:** Are different layers of the model good at predicting different regions in cortex?

NEURAL EXAMPLE

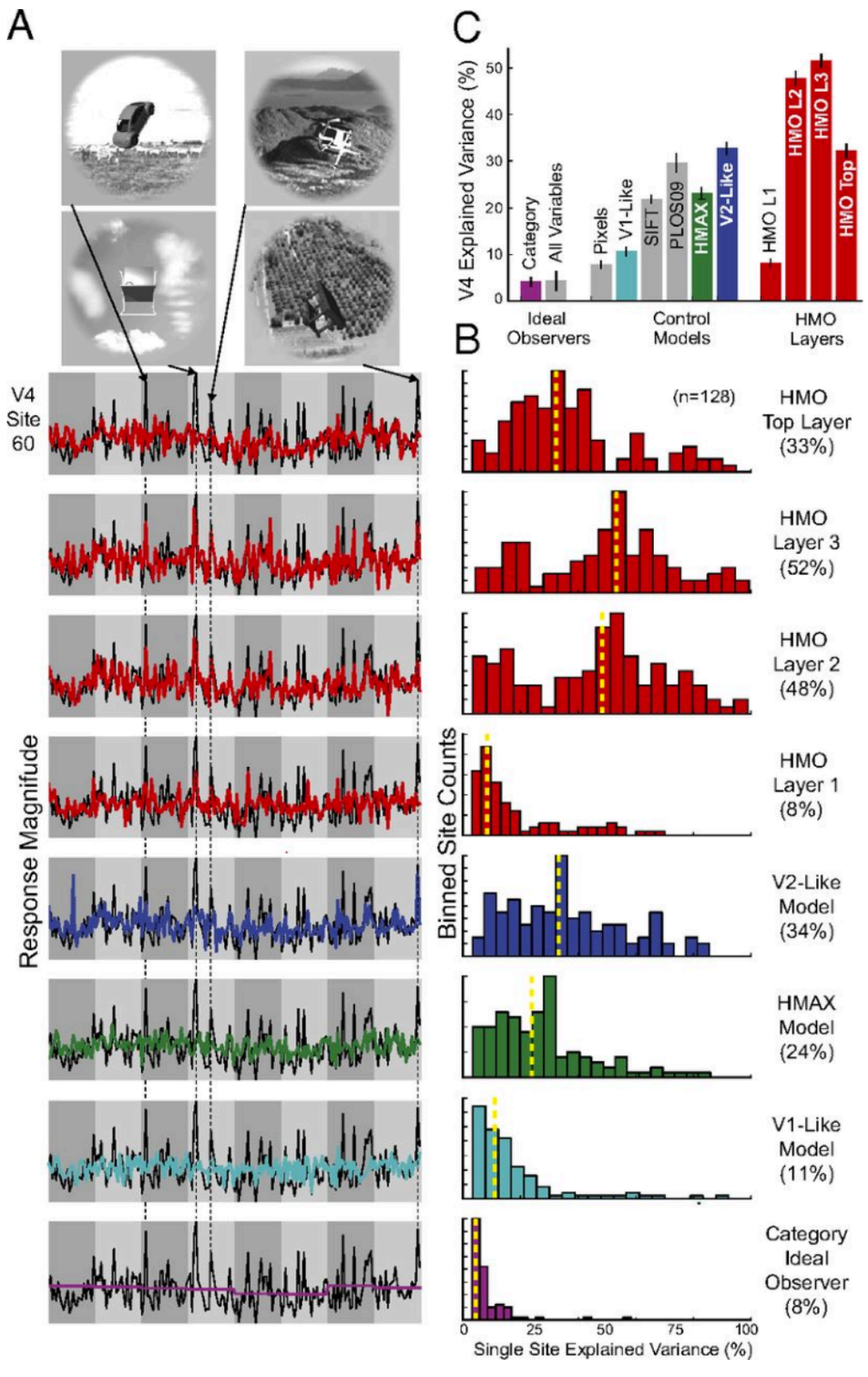
* IT neural responses are best explained by the “Top Layer” of the model



NEURAL EXAMPLE

* V4 neural responses are best explained by Layer 3 of the model!

* **Answer 2:** YES



CONSTRUCTION TO IDENTIFICATION

* Key questions about this scheme overall:

  * Does this work at all? ***YES***

    * If so, is this trivial? ***HMM…***

  * Do activations from different layers predict different parts of the brain? ***YES***

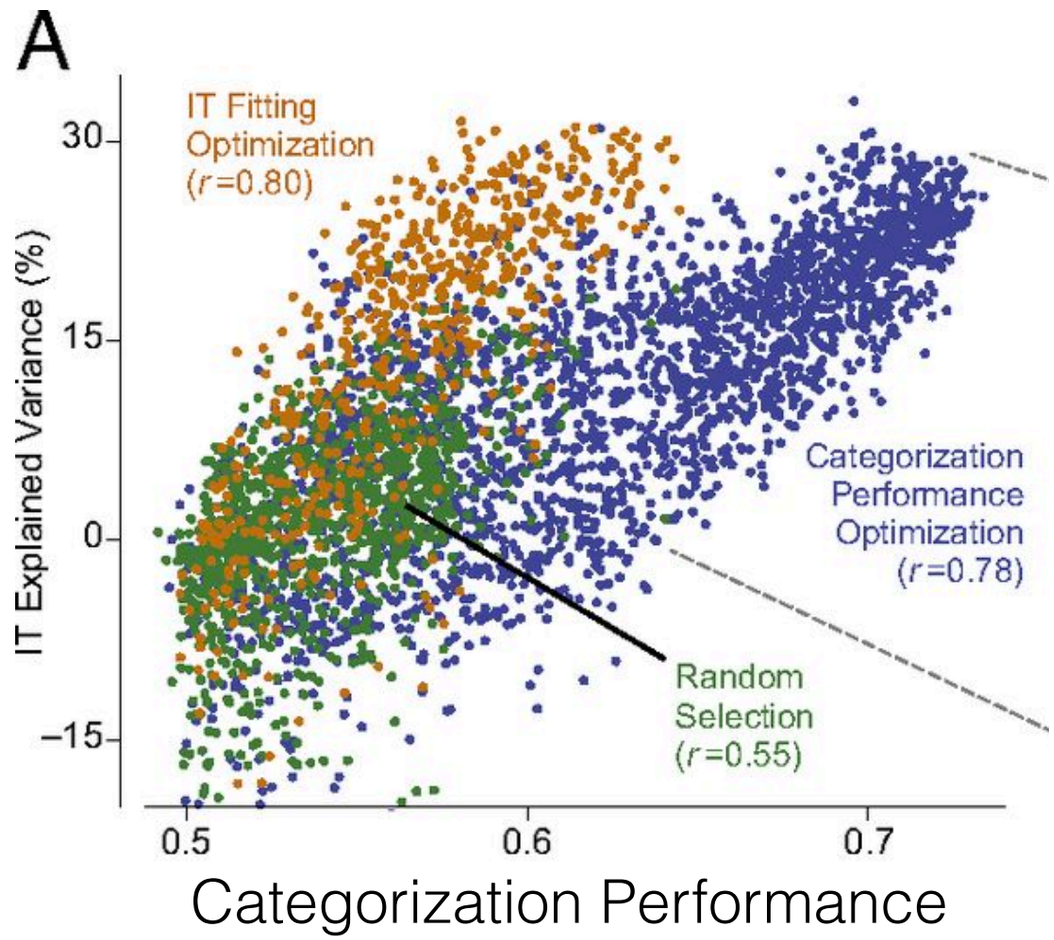
    * Does this correspond to known computational hierarchies? ***YES***

    * If so, is *this* trivial? ***HMM…***

HOW MUCH DOES TASK MATTER?

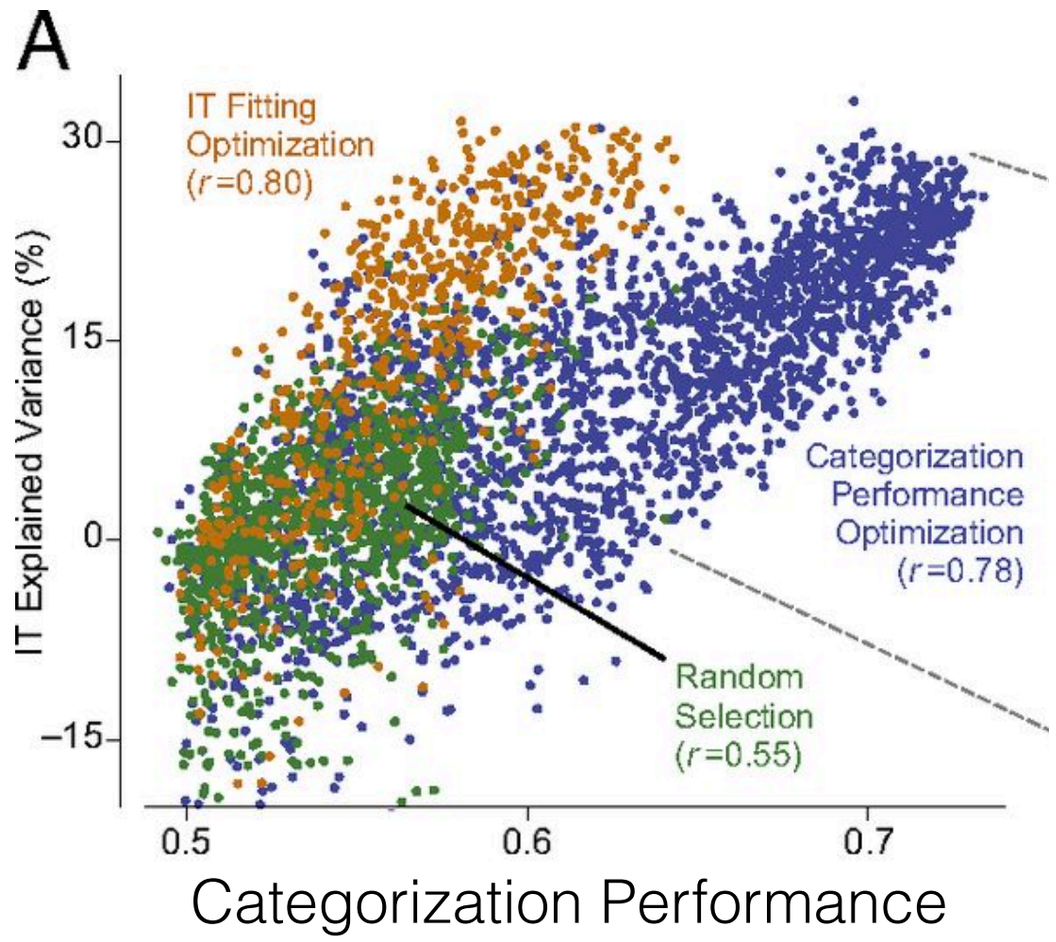
* Yamins et al. show that **untrained** neural networks do **reasonably well** at predicting brain data (& doing image categorization)

* i.e. building a supervised model on top of a **library of random non-linear features** works pretty well



HOW MUCH DOES TASK MATTER?

* ***Why?*** Let's take 5 minutes to discuss in small groups, then share thoughts w/ the rest of the class!



LANGUAGE

* Experimental setting:
  * A subject listens to language (e.g. a narrative story) while BOLD responses are recorded from cortex using fMRI

* We want to do **system identification**: build some model that predicts BOLD responses from the language stimuli

LANGUAGE

* Suppose we want to apply the “**system construction**” approach:
  * build an artificial neural network that solves **some task**
  * then use its representations (as a **linearizing transform**) to model the brain
* ***What task?***

LANGUAGE

* Suppose we want to apply the “**system construction**” approach:
  * build an artificial neural network that solves **some task**
  * then use its representations (as a **linearizing transform**) to model the brain
* ***What task?***

LANGUAGE

* In vision: **object categorization** seems like a good task (modulo earlier discussion), because it requires the network to learn lots of things

* What is an equivalent for language?

  * Ideally something related to language **meaning**

LANGUAGE MODELS

* One solution may be **language models**, which have taken on a similar role in the NLP field to image classification models (e.g. AlexNet) in computer vision

* The task of a language model is to predict a word from its context

  * e.g. $P(w_i|w_1,...,w_{i-1})$

LANGUAGE MODELS

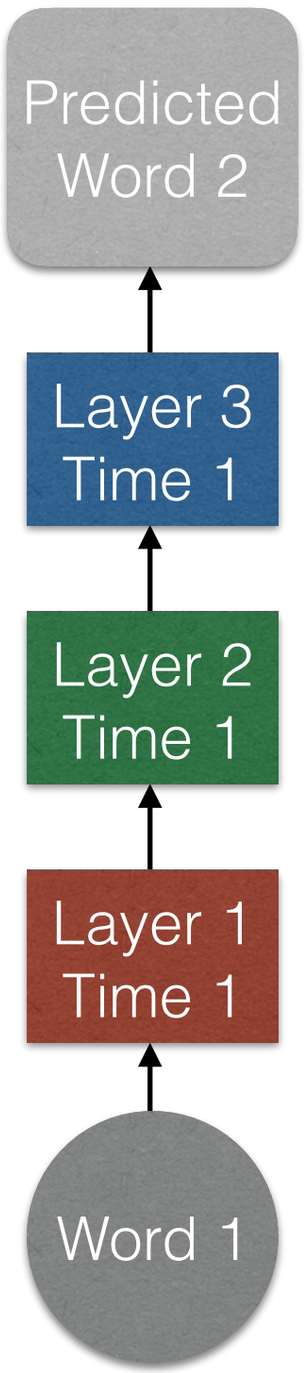
* Language models can use many different architectures

* One is a **recurrent neural network (RNN)**

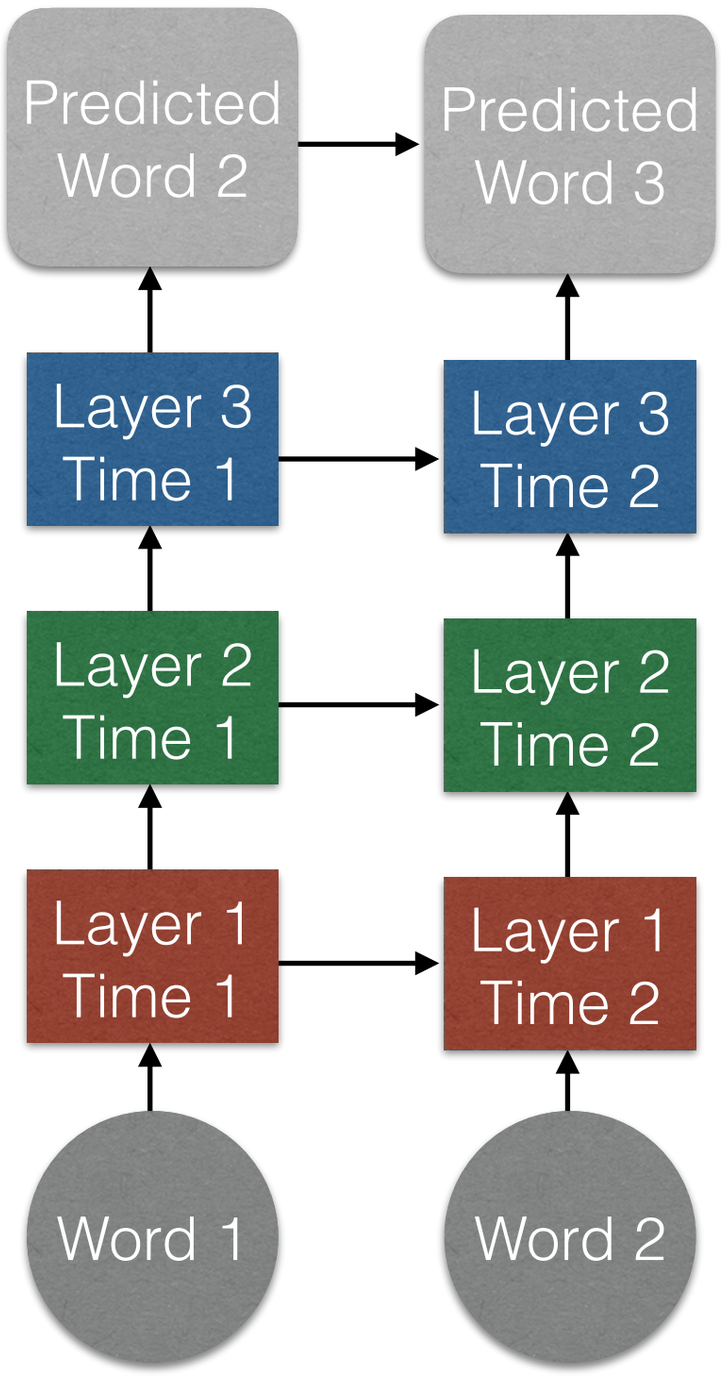
* In particular, a variant RNN called a **long short-term memory (LSTM) network**

  * (We'll talk talk about this network & how it works in detail on Tuesday)

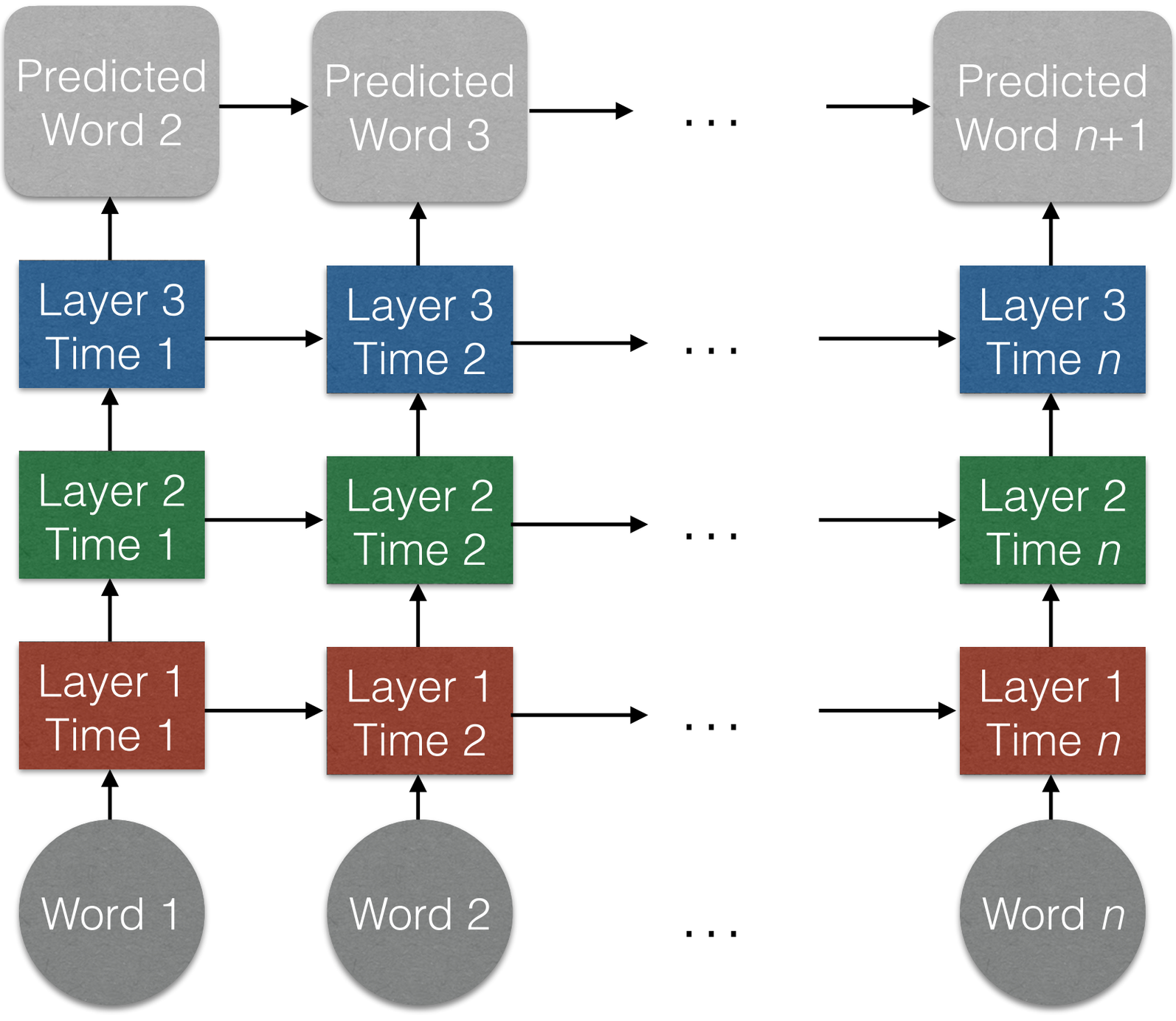
LSTM LANGUAGE MODEL



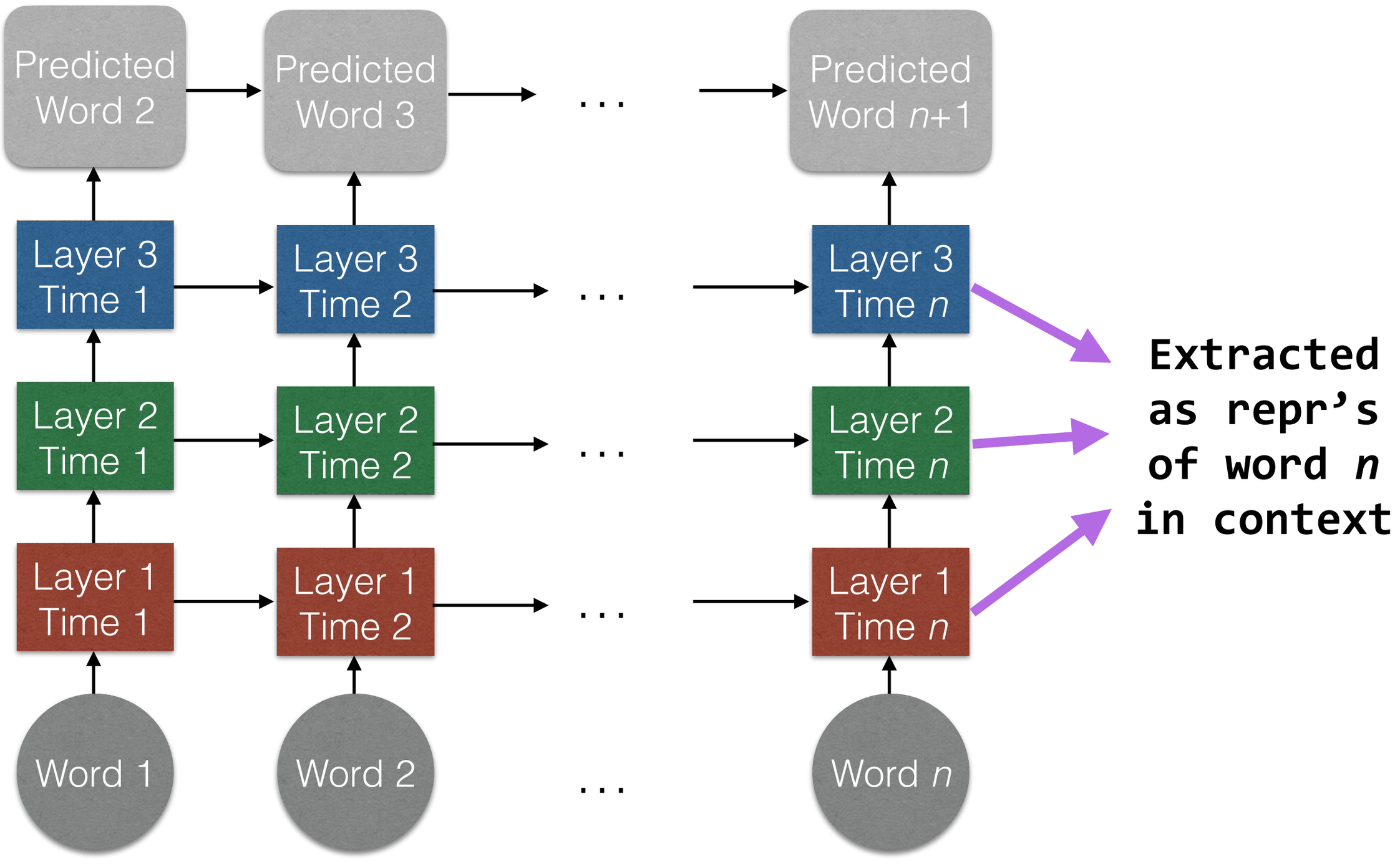
LSTM LANGUAGE MODEL



LSTM LANGUAGE MODEL



LSTM LANGUAGE MODEL



FMRI EXAMPLE USING LSTM LANGUAGE MODEL

* Jain & Huth (2018) Incorporating context into language encoding models for fMRI. *NeurIPS*

FMRI EXAMPLE USING LSTM LANGUAGE MODEL

* Approach:
  * Train LSTM language model on text
  * Do fMRI experiment
  * Use LSTM language model to extract features from fMRI stimuli
  * Build linearized system identification model using LSTM-derived features

FMRI EXAMPLE USING LSTM LANGUAGE MODEL

* In visual models different features can be extracted from different **layers**

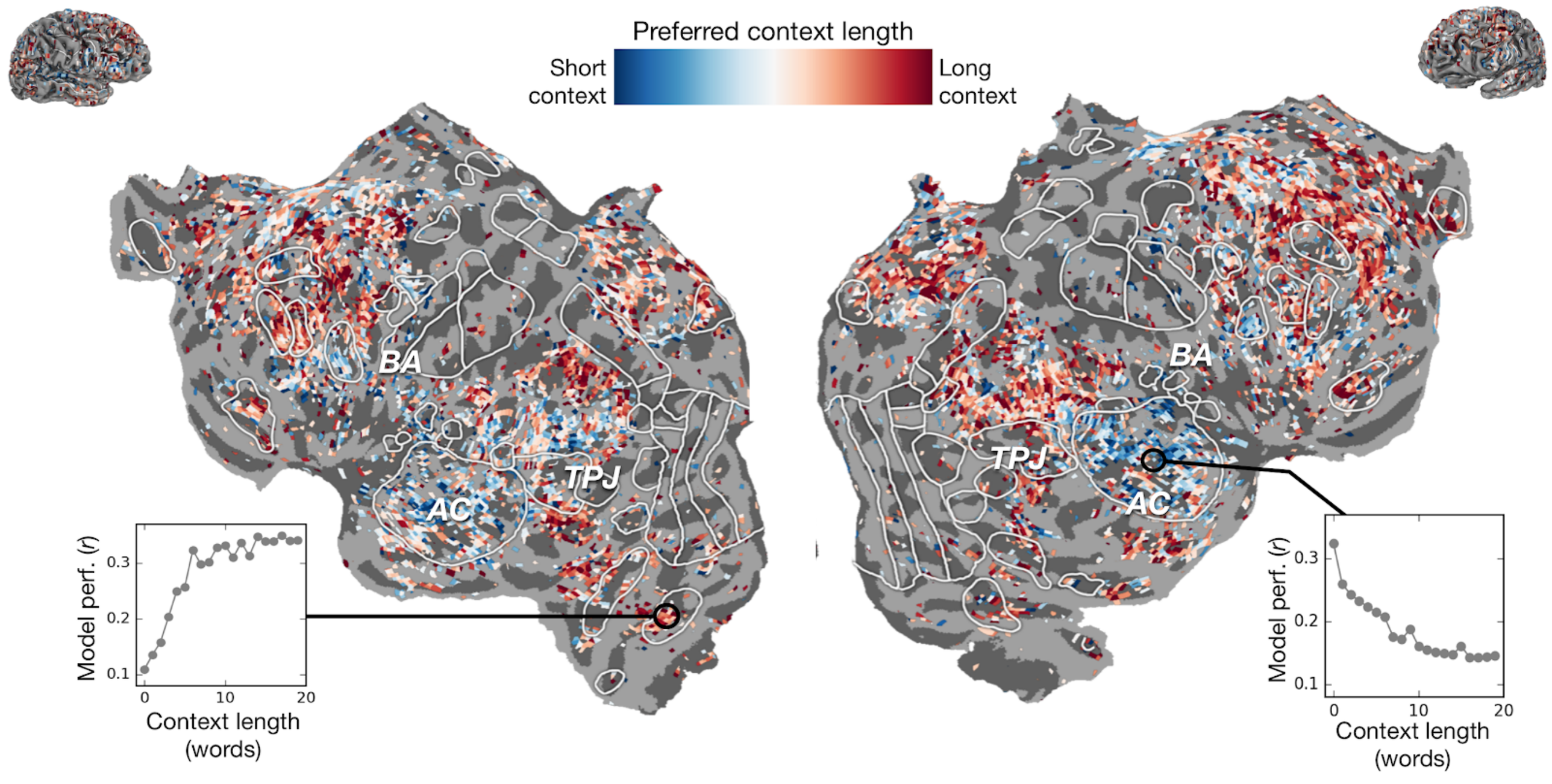
* In language models different features can be extracted from different **layers** *and* with different **amounts of context**

  * e.g. $P(w_i|w_{i-c},\ldots,w_{i-1})$

“context length”

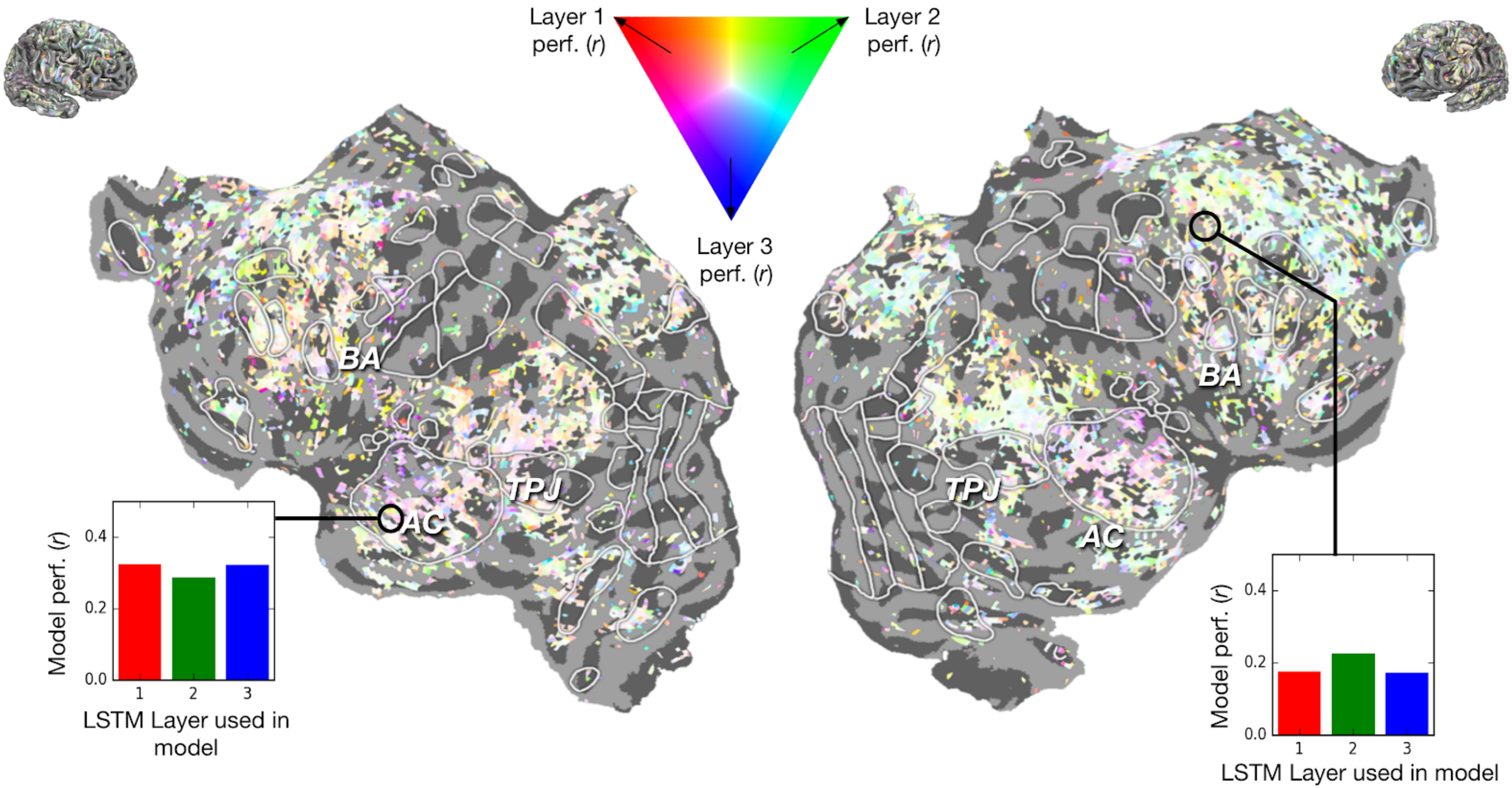
FMRI LSTM EXAMPLE

* Different brain areas prefer different amounts of context



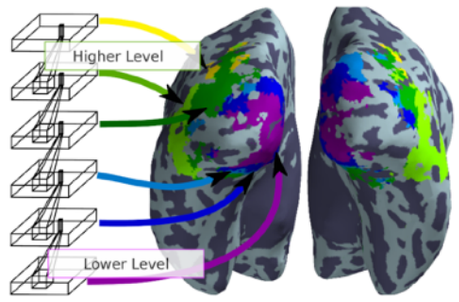
FMRI LSTM EXAMPLE

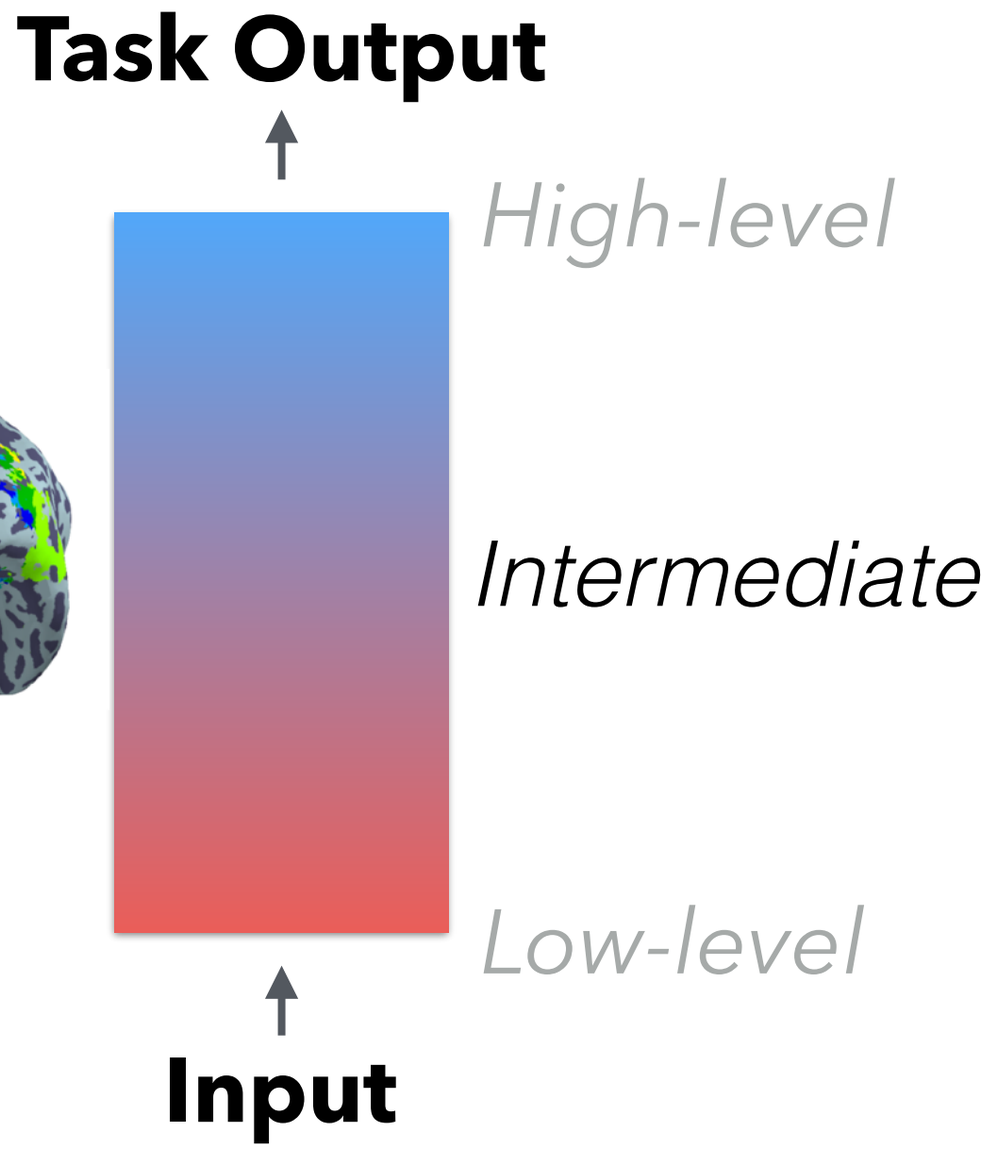
* But “layer preference” does not recapitulate known hierarchies, unlike Eickenberg, etc.



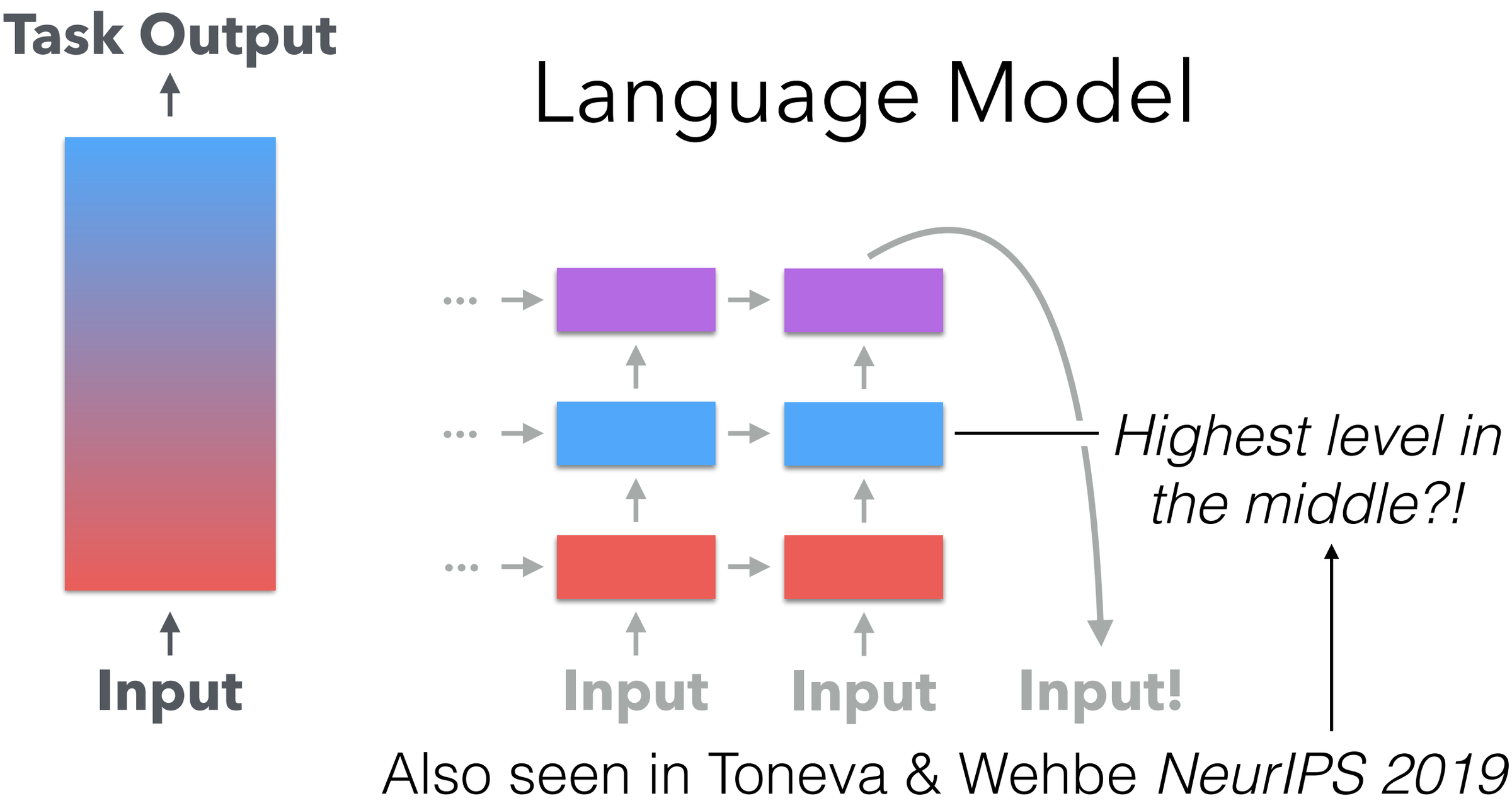
FMRI LSTM EXAMPLE

* In visual models, there is a clear “progression” of representations from low- to high-level





FMRI LSTM EXAMPLE



Also seen in Toneva & Wehbe *NeurIPS 2019*

NEXT TIME

* Recurrent neural networks in more detail