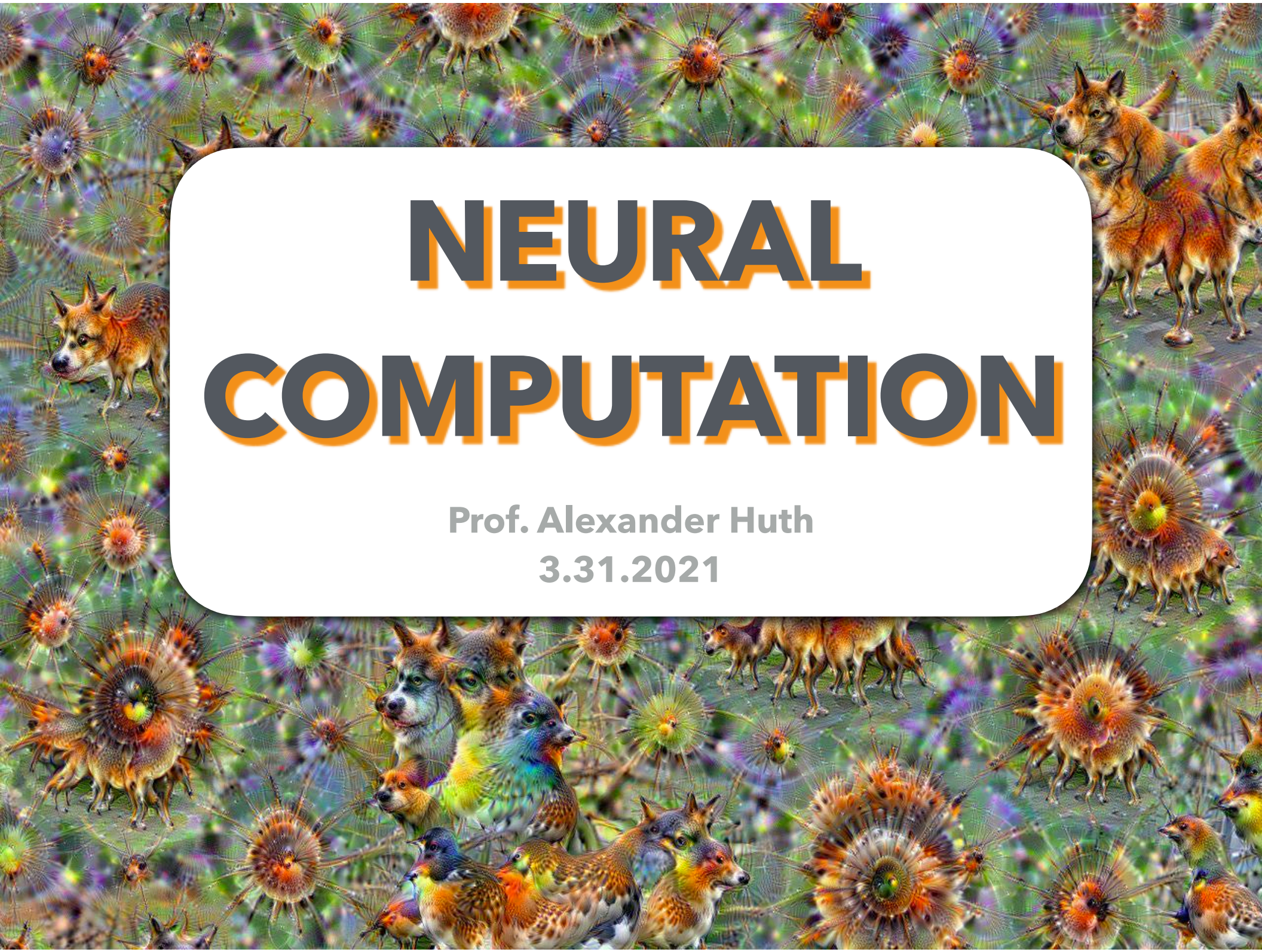

NEURAL
COMPUTATION
Prof. Alexander Huth
3.31.2021

RECAP

* **Deep** neural networks (i.e. with many hidden layers)
  * Trained using gradient **backpropagation**
  * Can do impressive things like **image classification** (e.g. **AlexNet**)
  * Using tricks like:
    * **localized** receptive fields
    * **convolution**
    * **downsampling** (i.e. max pooling)

RECAP

* Subtle switch from **system identification** to **system construction**

* *System identification*: building functions to explicitly approximate **representations** used by a given computational system

* *System construction*: building functions with (ideally) the same **behavior** or **goal** as a given computational system

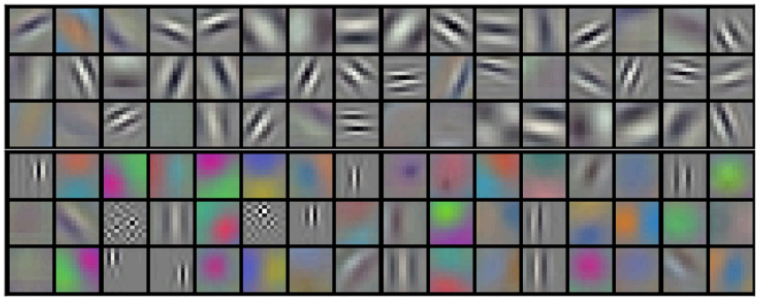
RECAP: ALEXNET

* Does it solve the problem the same way that our brains do?

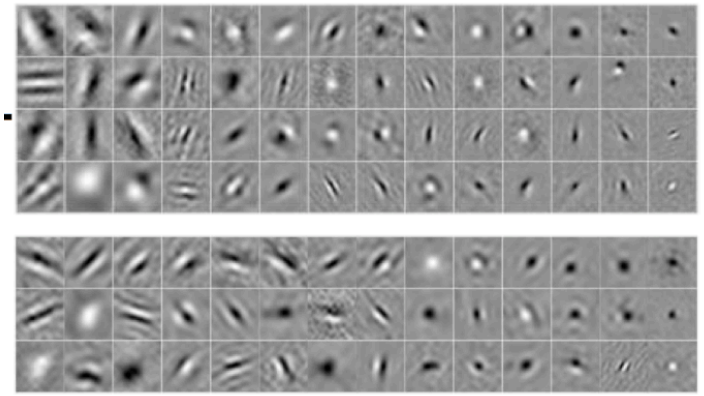
* How would we know?

RECAP: ALEXNET

* Features learned by 1st layer of AlexNet



* Receptive fields of V1 neurons from a macaque



CONSTRUCTION TO IDENTIFICATION

* What would it mean for a constructed system (e.g. AlexNet) to solve a problem **similarly** to how an existing system (e.g. a brain) does?

CONSTRUCTION TO IDENTIFICATION

* Recall Marr's levels of description:
  * **Computational**: both systems solve the same task ✅
  * **Algorithmic/Representational**: both systems use the same algorithm or representations ? ? ?
  * **Implementation**: both systems use the same implementation (neurons, etc.) ⛔

CONSTRUCTION TO IDENTIFICATION

* Recall Marr's levels of description:
  * **Computational**: both systems solve the same task ✅
  * **Algorithmic/Representational**: both systems use the same algorithm or representations ? ? ?
  * **Implementation**: both systems use the same implementation (neurons, etc.) ⛔

CONSTRUCTION TO IDENTIFICATION

* What would it mean for a constructed system (e.g. AlexNet) to solve a problem **similarly** to how an existing system (e.g. a brain) does?

  * *It would mean that both use the same algorithm or representations to solve the problem*

  * How do we test whether a system uses some representation?

CONSTRUCTION TO IDENTIFICATION

* Key idea: *use the constructed system as a linearizing transform for system identification*

EXTRACTING REPRESENTATIONS FROM A NEURAL NETWORK

* The **activations** at each layer of a deep neural network can serve as “**features**”

* The **functions** that compute the activations (from the input) thus serve as **linearizing transformations**

EXTRACTING REPRESENTATIONS FROM A NEURAL NETWORK

* (“whiteboard” notes)

CONSTRUCTION TO IDENTIFICATION

1. Collect dataset of brain responses Y using stimuli X

2. Train/find artificial system

3. Extract activations $a(X)$ for the stimuli in the neuroscience experiment

4. Use activations as features in linearized encoding model: $Y = a(X) \cdot b$

CONSTRUCTION TO IDENTIFICATION

* Key questions:
  * Does this work at all?
    * If so, is this trivial?
  * Do activations from different layers predict different parts of the brain?
    * Does this correspond to known computational hierarchies?
    * If so, is *this* trivial?

FMRI EXAMPLE

* Eickenberg, Gramfort, Varoquaux, & Thirion (2016) Seeing it all: convolutional network layers map the function of the human visual system. *Neuroimage*

FMRI EXAMPLE

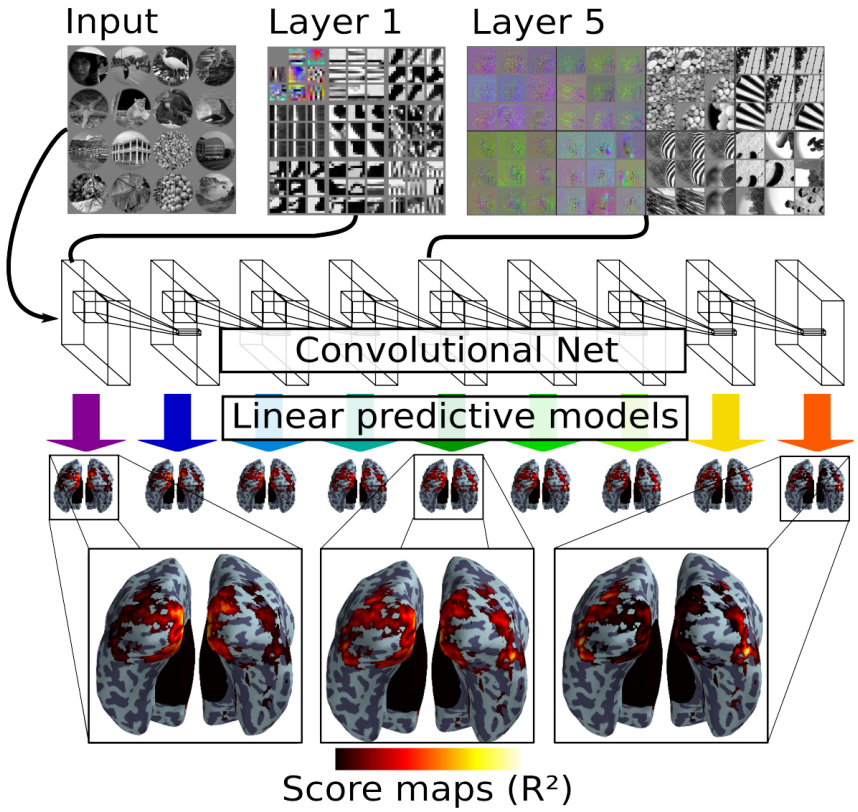
* fMRI data come from experiment we discussed previously (see *lecture 10*, Kay et al., Nature 2008)

* Two human subjects viewed 1000s of images while BOLD signals were recorded from visual cortex using fMRI

FMRI EXAMPLE

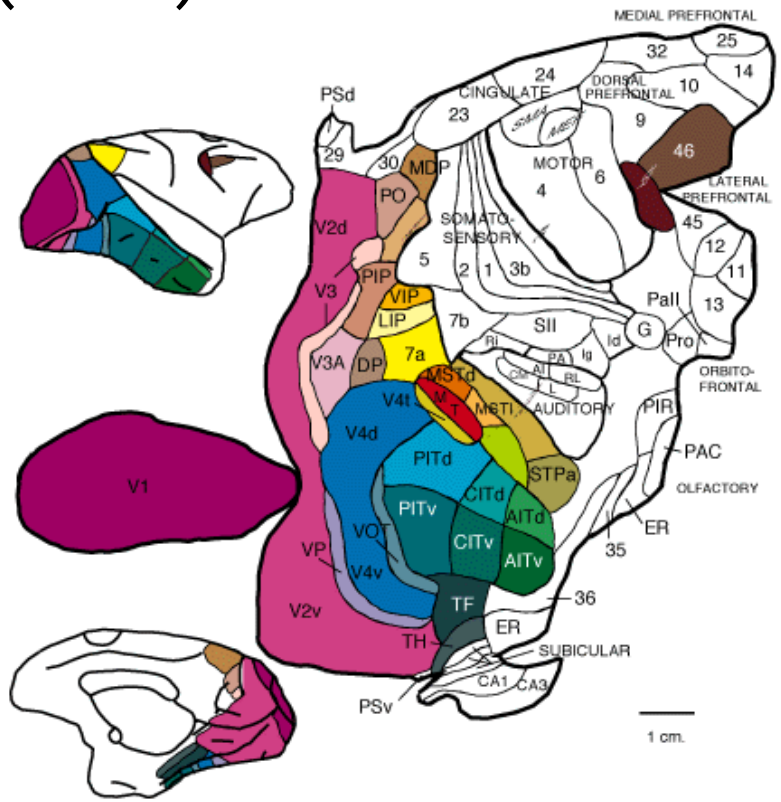
* Used “Overfeat” convolutional neural network (similar architecture to AlexNet)

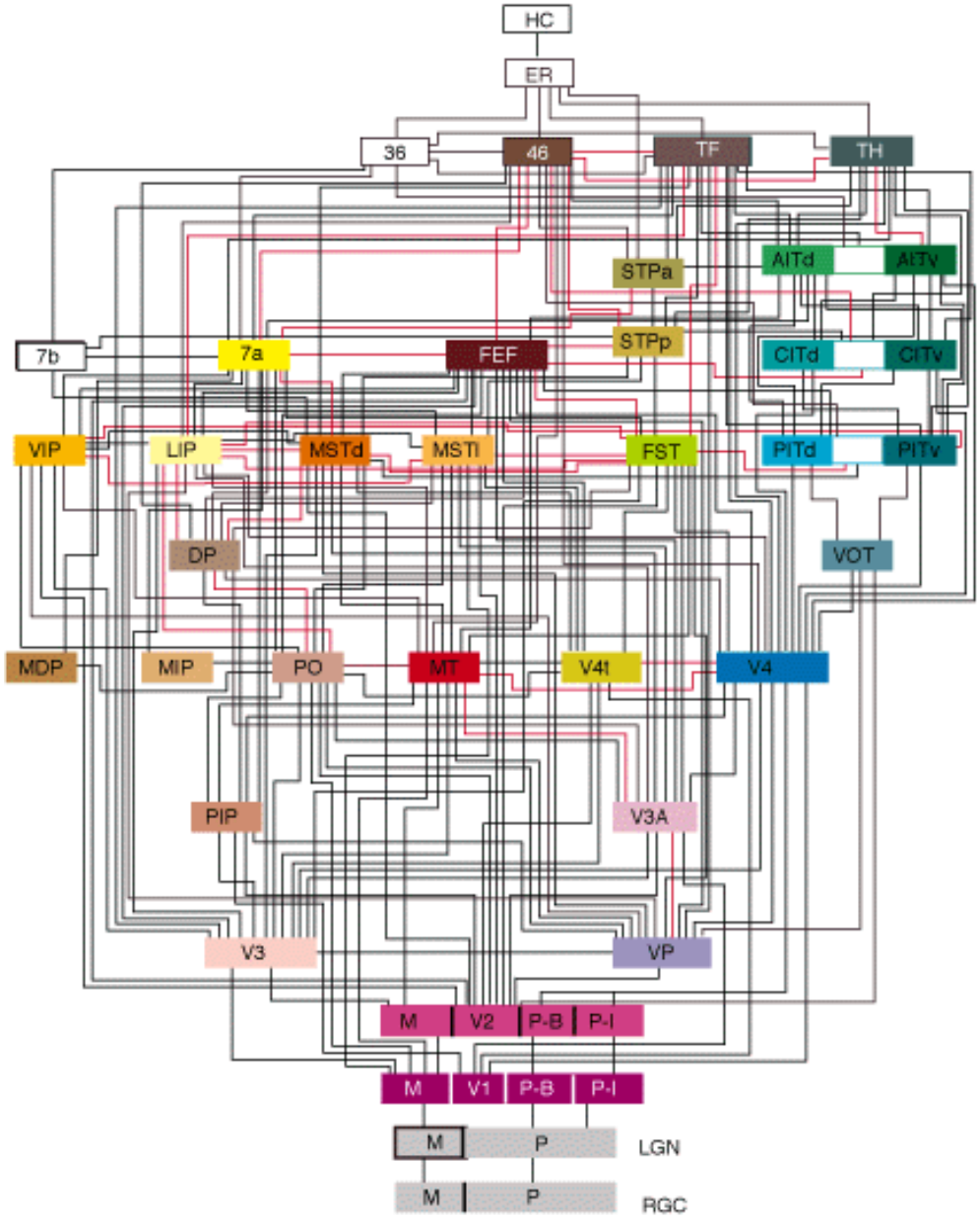
* Built model for each voxel using activations from each layer, then tested on held-out data



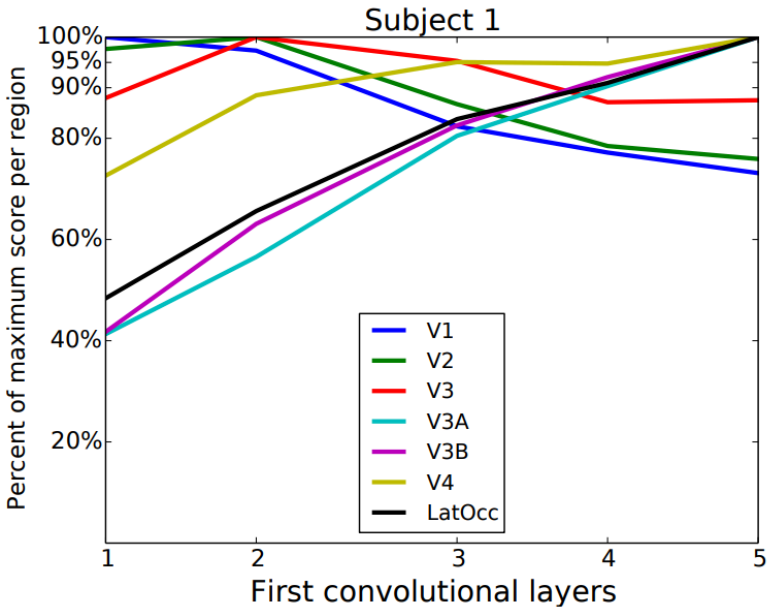
VISUAL CORTEX

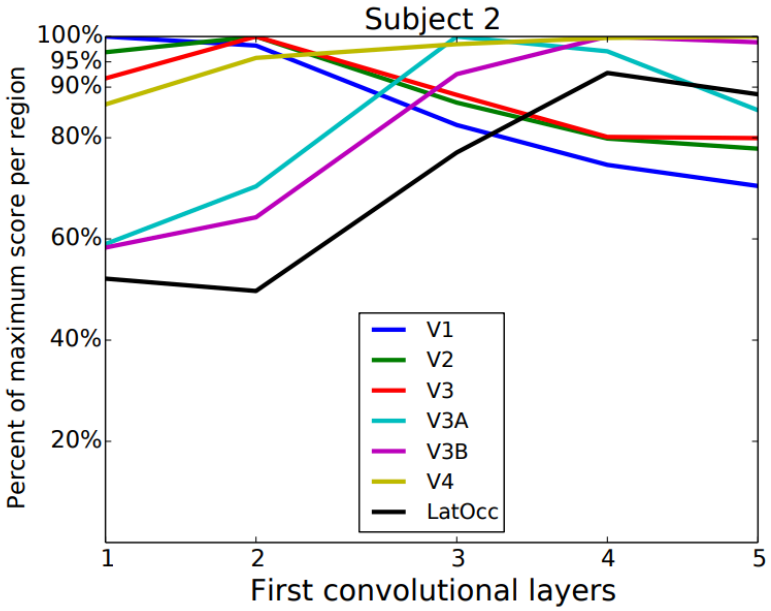
* **Connectivity** of areas in macaque visual cortex
* Felleman & Van Essen (1991)





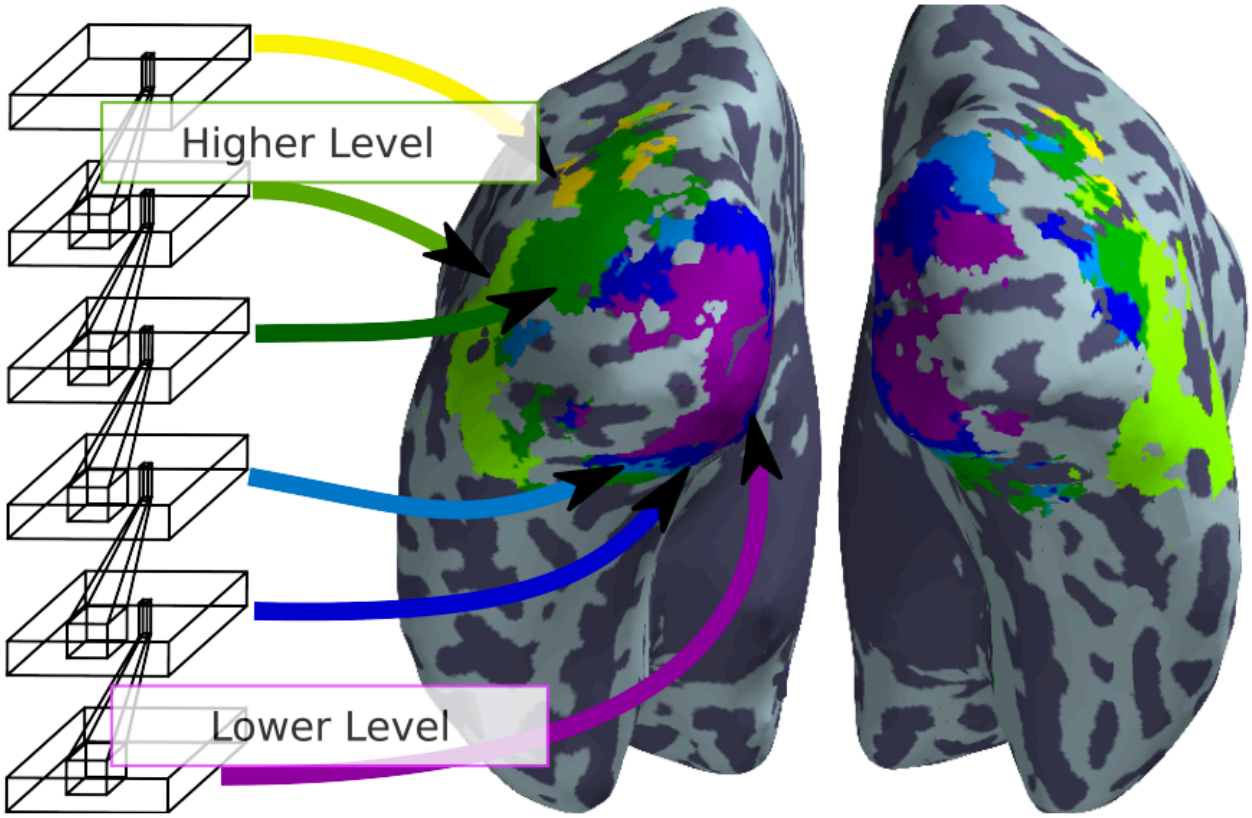
FMRI EXAMPLE





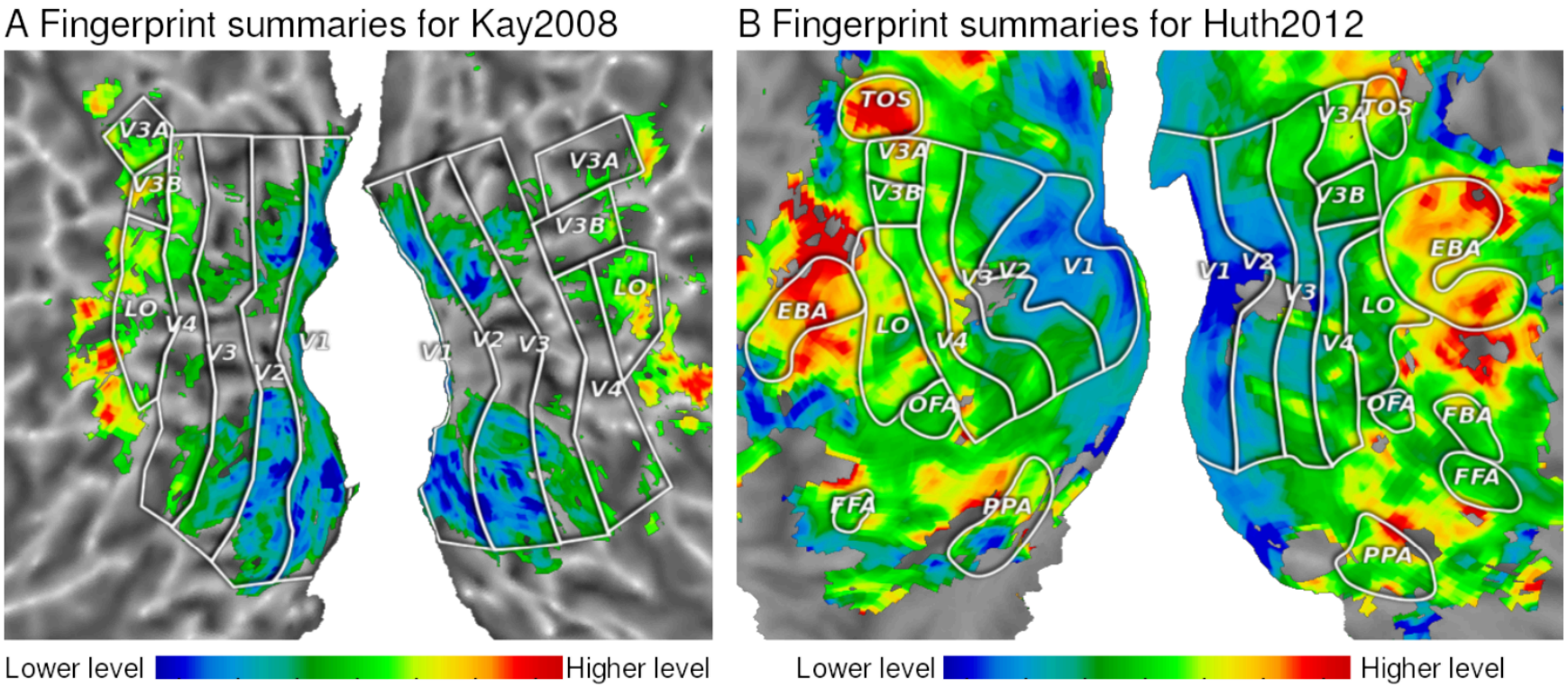
* Lower layers of the network predict “earlier” visual areas, higher layers predict “later” visual areas!

FMRI EXAMPLE



FMRI EXAMPLE

* ***Fingerprint:*** summary model performance metric giving something like “best layer”



CONSTRUCTION TO IDENTIFICATION

* Key questions:
  * 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…***

NEXT TIME

* Similar analysis for neurophysiology data