A Quest to understand the neural mechanisms of human visual intelligence

James DiCarlo MD, PhD

Peter de Florez Professor of Neuroscience
Director, MIT Quest for Intelligence
Investigator, McGovern Institute for Brain Research
Investigator, Center for Brains, Minds and Machines
Massachusetts Institute of Technology

A Quest to understand the neural mechanisms of human visual intelligence

= "Reverse engineering human visual intelligence"

James DiCarlo MD, PhD

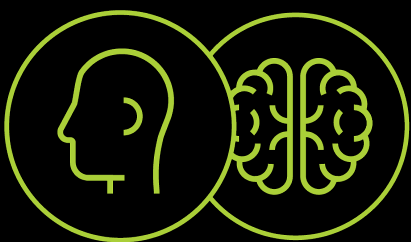
Peter de Florez Professor of Neuroscience
Director, MIT Quest for Intelligence
Investigator, McGovern Institute for Brain Research
Investigator, Center for Brains, Minds and Machines
Massachusetts Institute of Technology

The neuroscientific goal of reverse engineering:

Account for human visual intelligence …
(behavioral capabilities)

…using mechanisms of the brain…
(networks of simulated neurons)

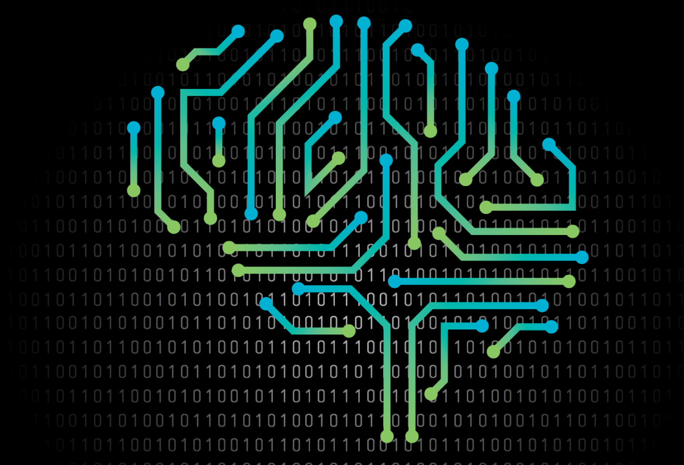
…in the language of engineering
(predictive, built systems).



SCIENCE

Mind & Brain

Measurements
& Discoveries



***Specific* artificial neural networks as implemented (and leading) scientific hypotheses**

ENGINEERING

Software, Hardware, Robotics

Synthesis & Creation

My talk today: Ongoing progress on a foundational piece of visual intelligence

Local reverse engineering team:

Current group:

Robert Ajemian
Yoon Bai
Joel Dapello
Kohitij Kar
Micheal Lee
Tiago Marques
Ratan Murty
Alina Peter
Jon Prescott-Roy
Sachi Sanghavi
Martin Schrimpf
Christopher Shay
Chris Stawarz

- *Office of Naval Research*
- *National Science Foundation (CBMM)*
- *Simons Global Brain*
- *IBM/MIT Watson Lab*
- *Semiconductor Research Corporation (SRC)/DARPA*

Key collaborators:

Ed Boyden (MIT)
SueYeon Chung (Columbia)
David Cox (IBM)
Danny Gutfreund (IBM)
Nancy Kanwisher (MIT)
Lynne Kiorpes (NYU)
Fei-Fei Li (Stanford)
Jitendra Malik (UC Berkeley)
J. Anthony Movshon (NYU)
Tomaso Poggio (MIT)
Kaushik Roy (Purdue)
Josh Tenenbaum (MIT)
Andreas Tolias (Baylor CM)
Dan Yamins (Stanford)

Alumni scientists:

Arash Afraz (=> Prof., NIH)
Paul Aparicio (=> NIH)
Pouya Bashivan (=>Prof., McGill)
Charles Cadieu (=> Caption Health)
David Cox (=> IBM, VP AI research)
Ha Hong (=> Caption Health)
Chou Hung (=> Army research)
Elias Issa (=> Prof., Columbia)
Xioaxuan Jia (=>Allen Institute)
Kamila Jozwik (=> U. Cambridge)
Gabriel Kreiman (=> Prof., Harvard)
Hyodong Lee (=> Google)
Nuo Li (=> Prof., Baylor College Med)
Najib Majaj (=> NYU)
Shay Ohayon (=> Google)
Nicolas Pinto (=> Apple => Cygni)
Rishi Rajalingham (=>MIT)
Nicole Rust (=> Prof., U Penn)
Dan Yamins (=> Prof., Stanford)
Davide Zoccolan (=> Prof., SISSA)

Human visual intelligence…

Guidance from brain and cognitive sciences: ~10 deg at center of gaze, ~200 msec snapshots

Guidance from brain and cognitive sciences: ~10 deg at center of gaze, ~200 msec snapshots

Image adapted from MIT Street Scenes Database (Courtesy of Tommy Poggio)

Guidance from brain and cognitive sciences: **~10 deg at center of gaze, ~200 msec snapshots**

Foundational component of visual intelligence:

Core object recognition

Is a car here?

Is a person here?

What is the pose of the car?

...

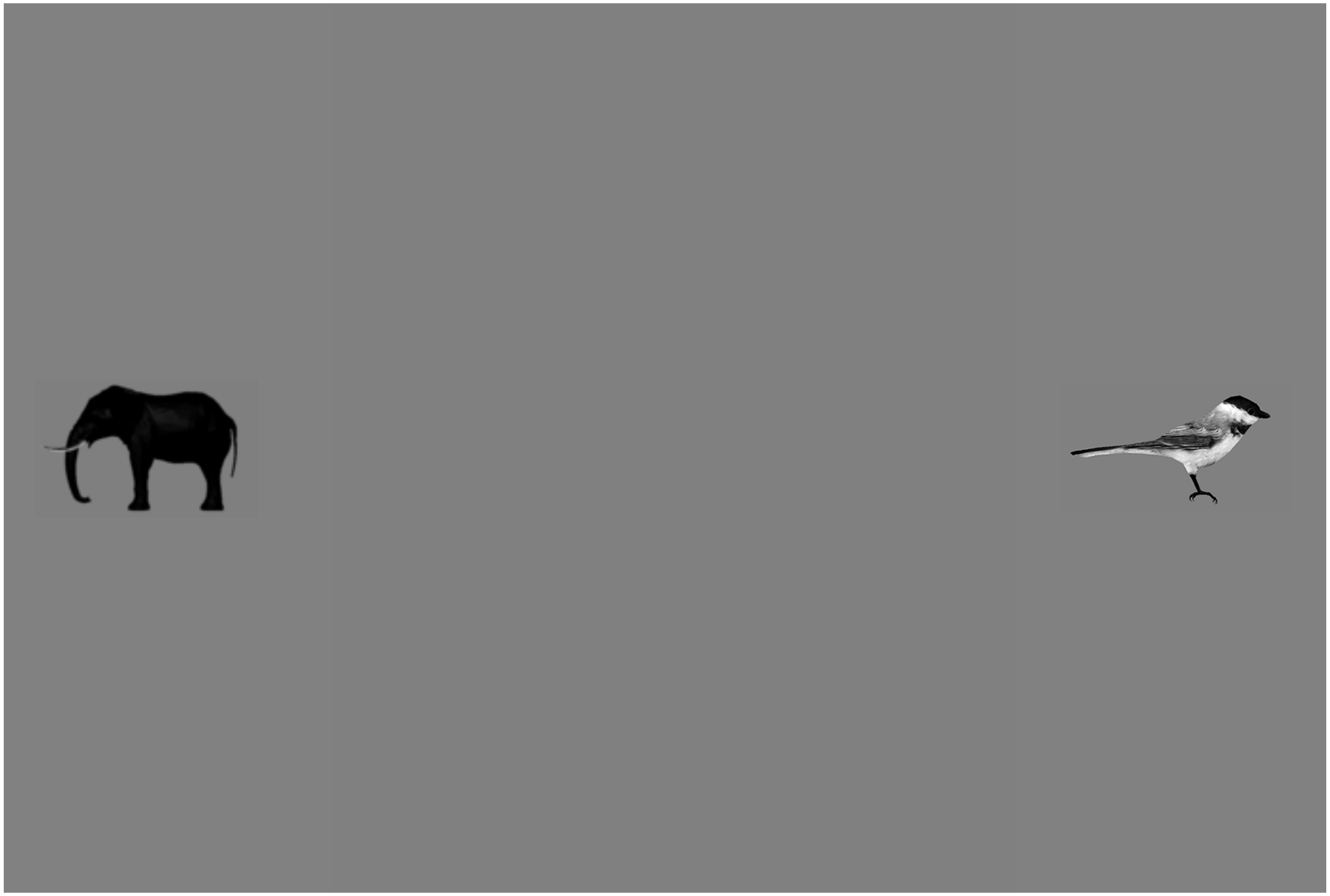
~200 msec

~200 msec

...

~200 msec

Example behavioral test trials

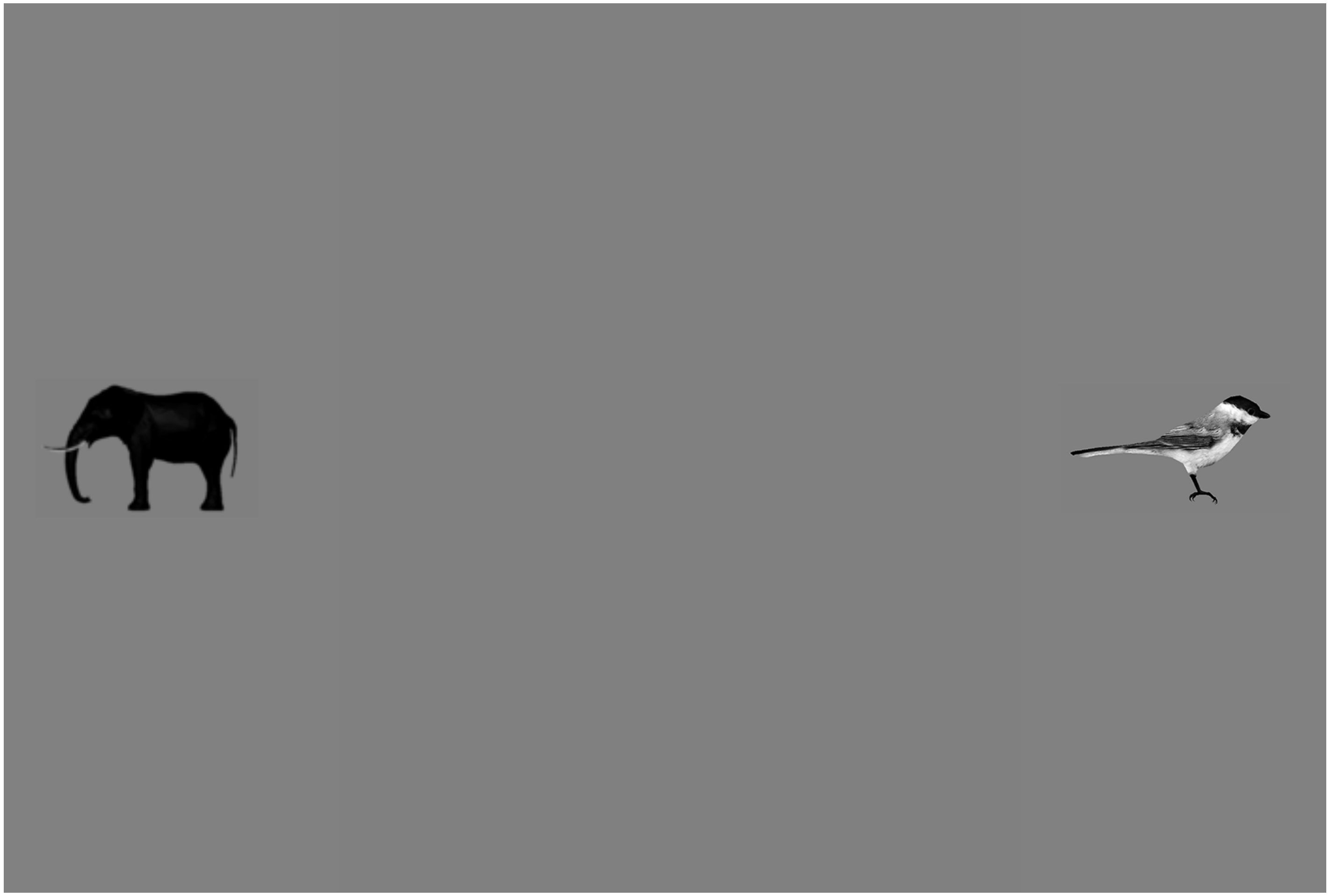


8 deg image at center of gaze, ~100 msec viewing time

Example behavioral test trials

8 deg image at center of gaze, ~100 msec viewing time

Example behavioral test trials



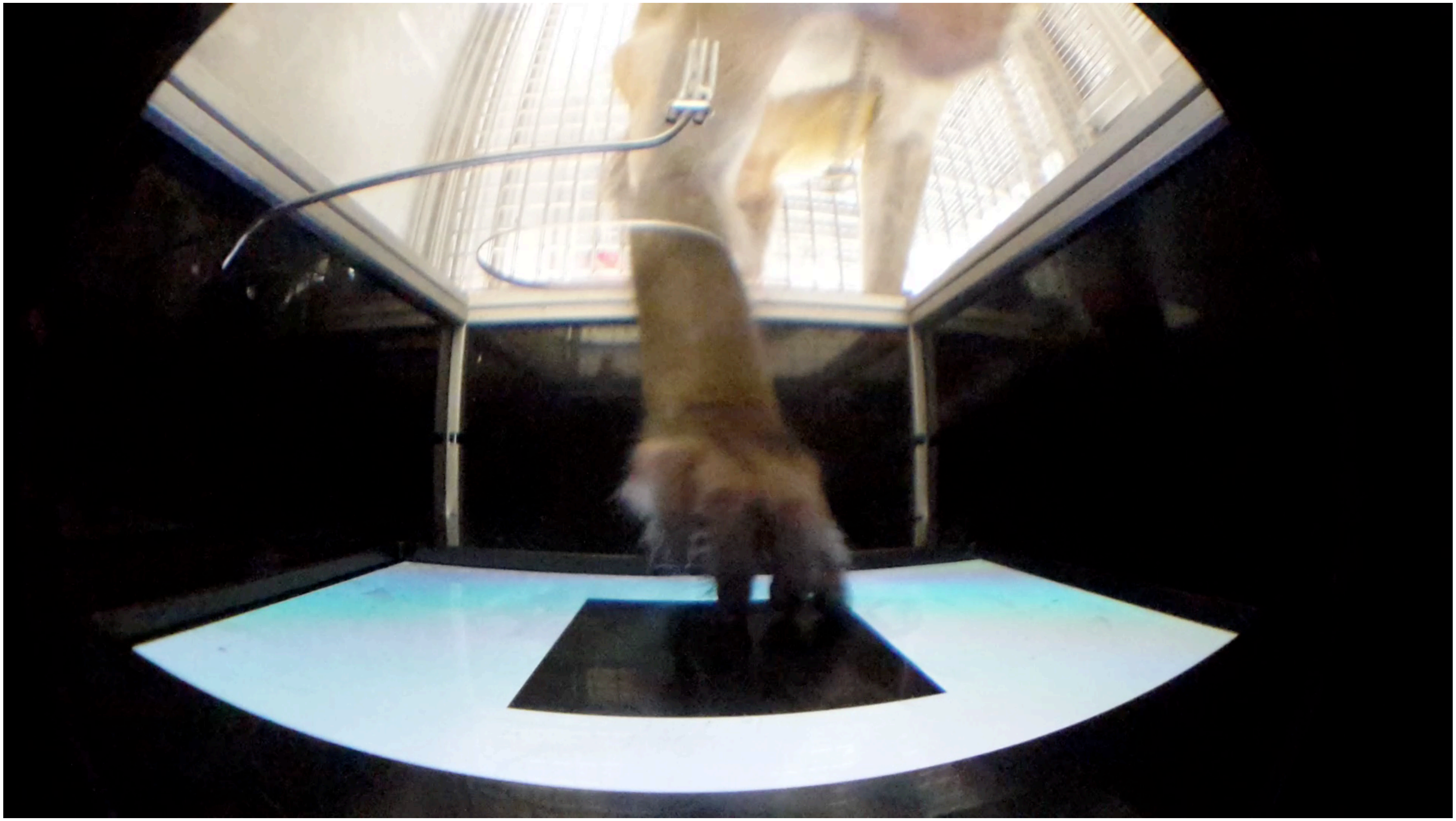
8 deg image at center of gaze, ~100 msec viewing time

Species A

Species B

Computer vision systems **<** **primate, Homo sapien**

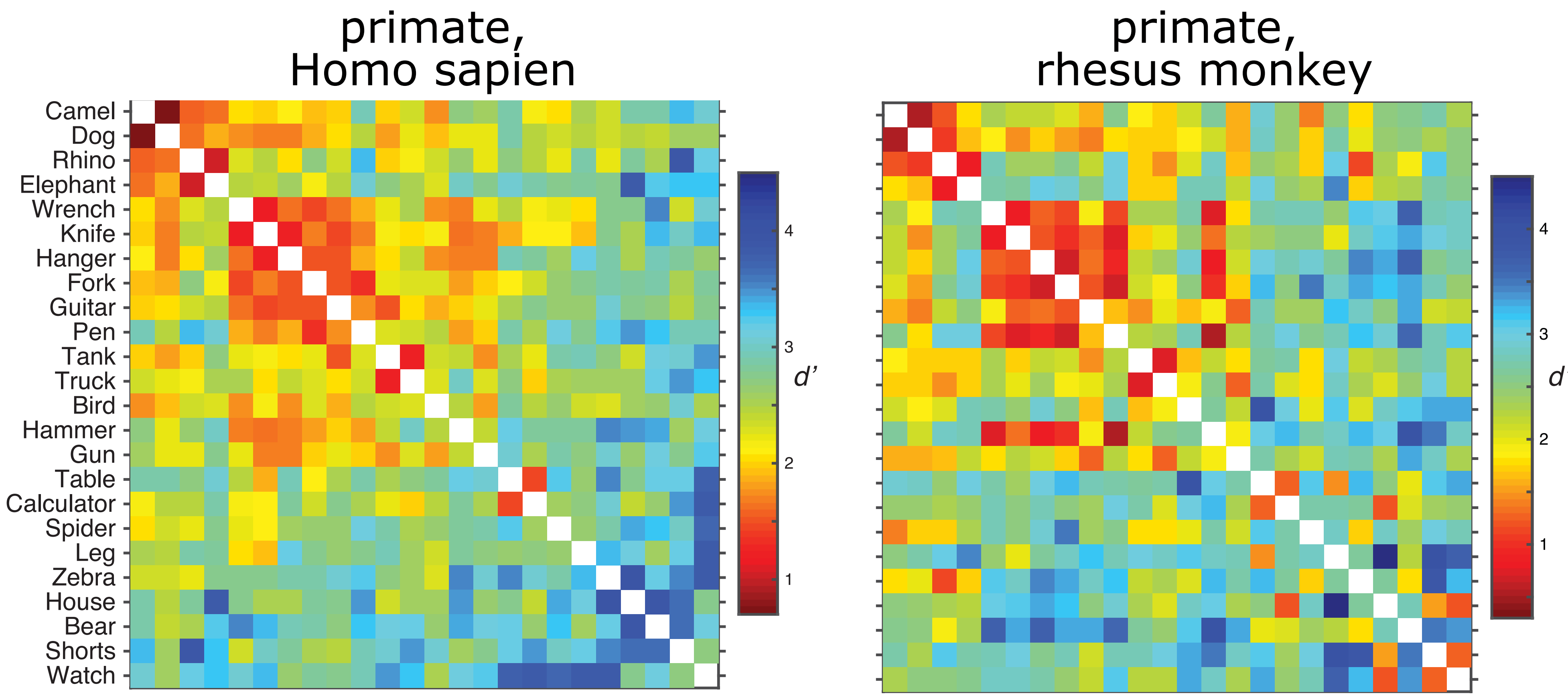
Intelligence test domain: Core Visual Object Perception



Intelligence test domain: Core Visual Object Perception

Behavior test performance for 276 core object recognition tasks

(note: many images in each task)

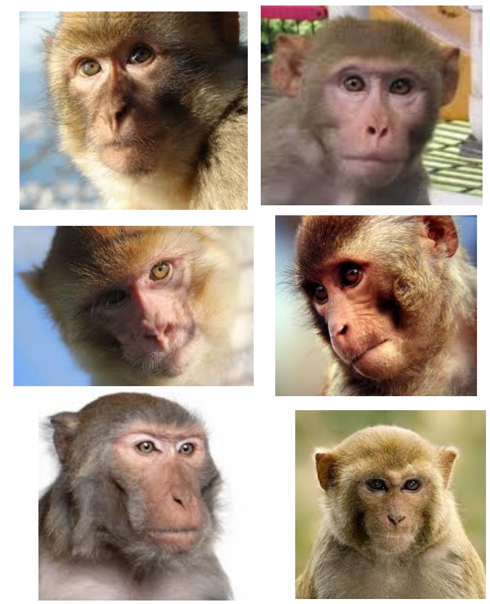


Rajalingham, Schmidt, & DiCarlo, ***Vision Sciences Society*** *(2014)*
Rajalingham, Schmidt, & DiCarlo, ***J. Neuroscience*** *(2015)*
Rajalingham, Issa, Kar, Schmidt, & DiCarlo, ***CCN*** *(2017)*

Intelligence test domain: Core Visual Object Perception

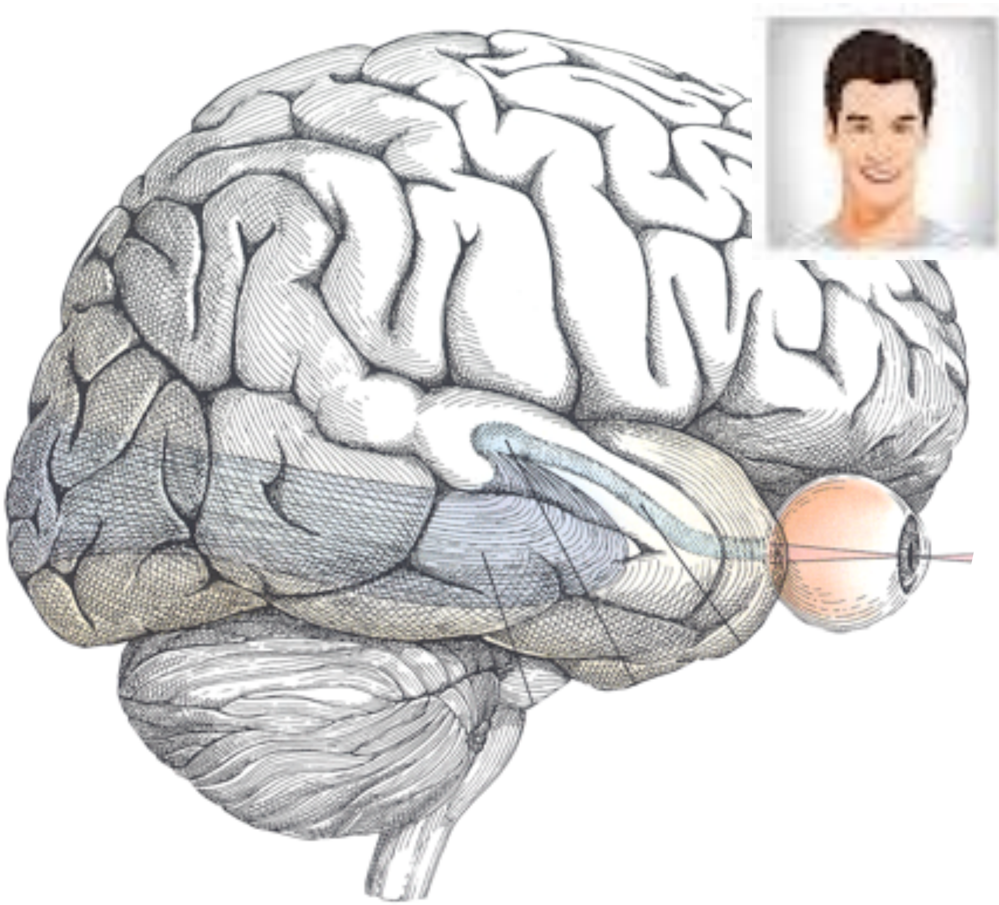
Primates

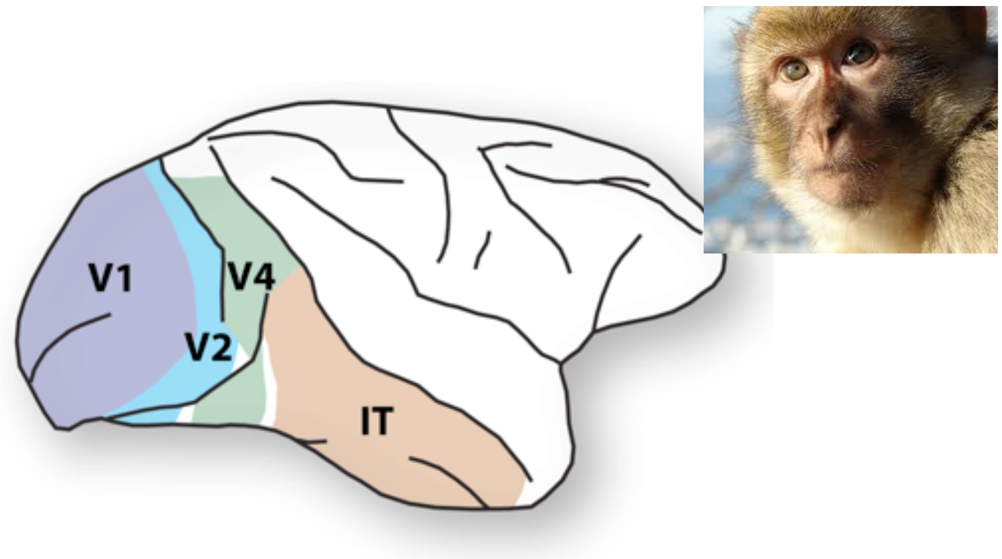
Want-to-be
primates

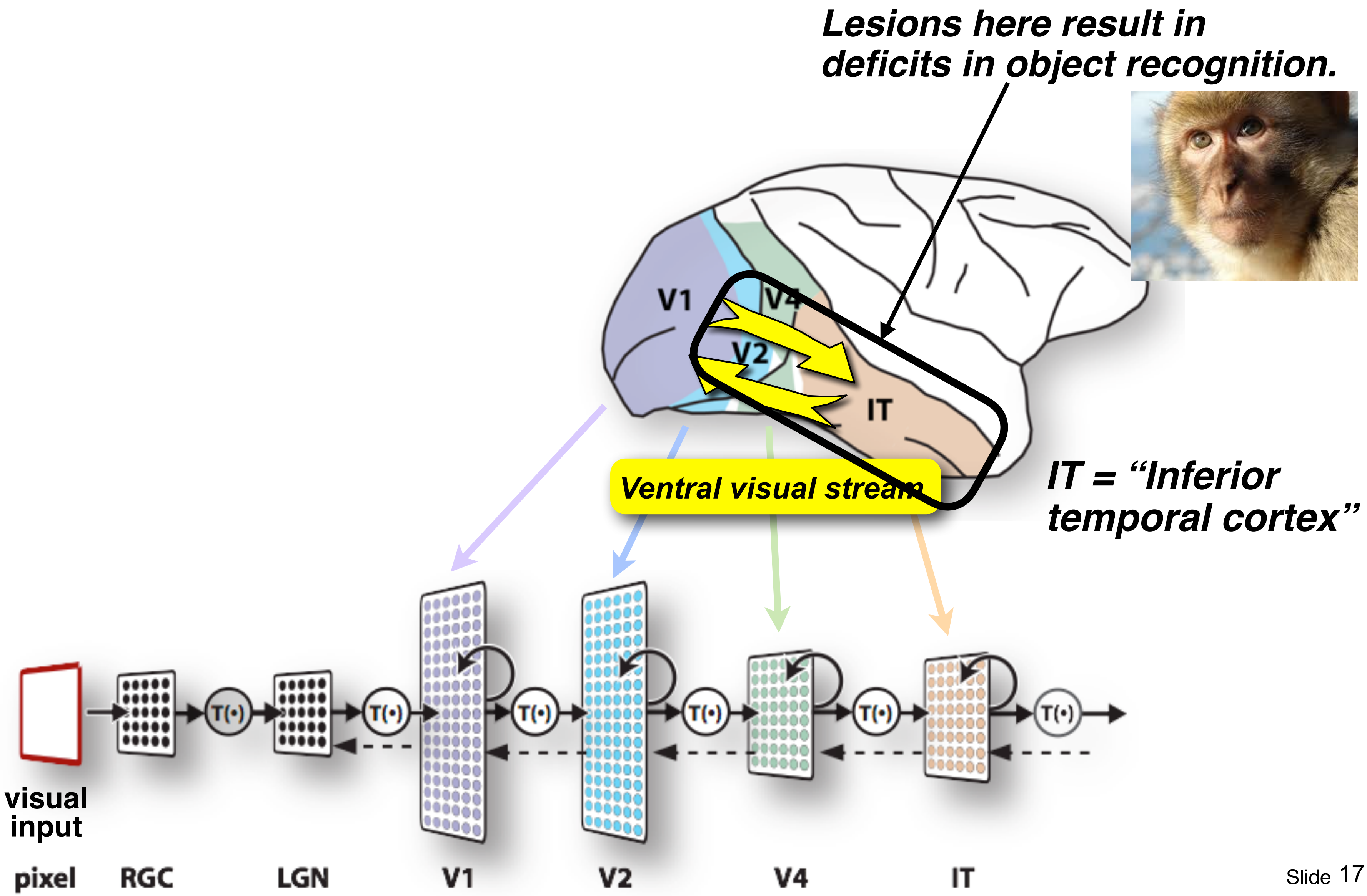


Computer vision systems < [primate, Homo sapien ≅ primate, rhesus monkey]

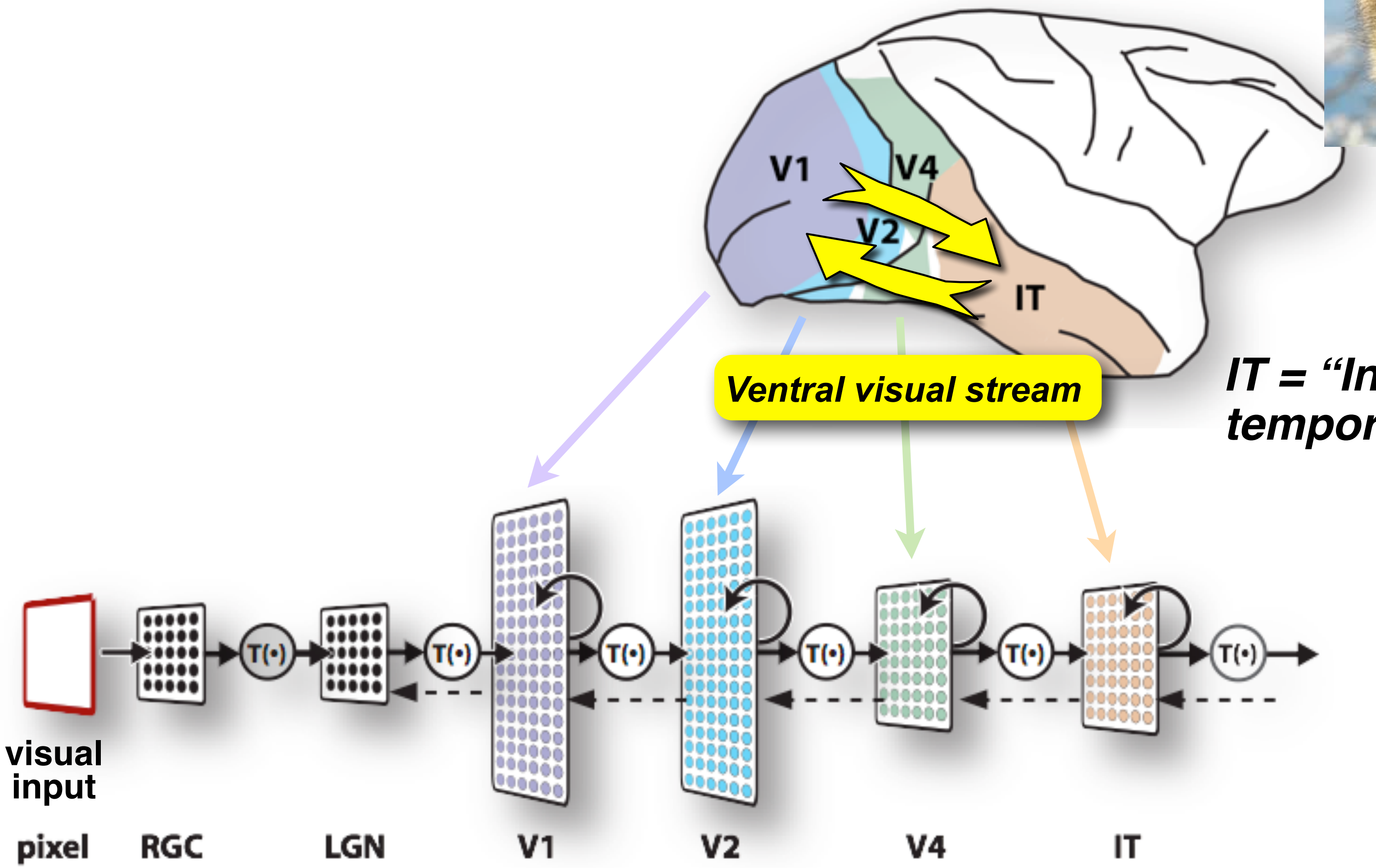
Intelligence test domain: Core Visual Object Perception



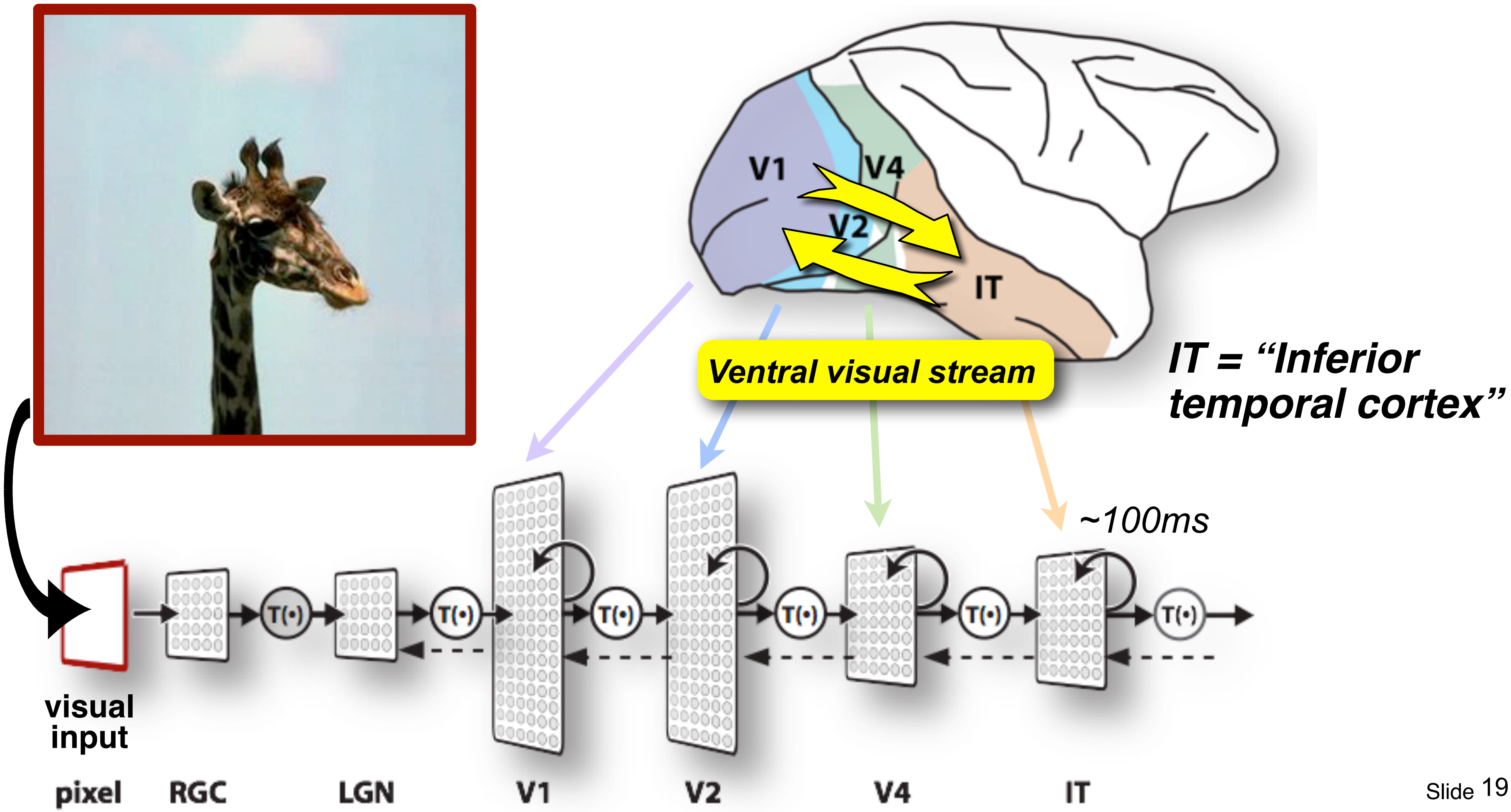

V1
V4
V2
IT

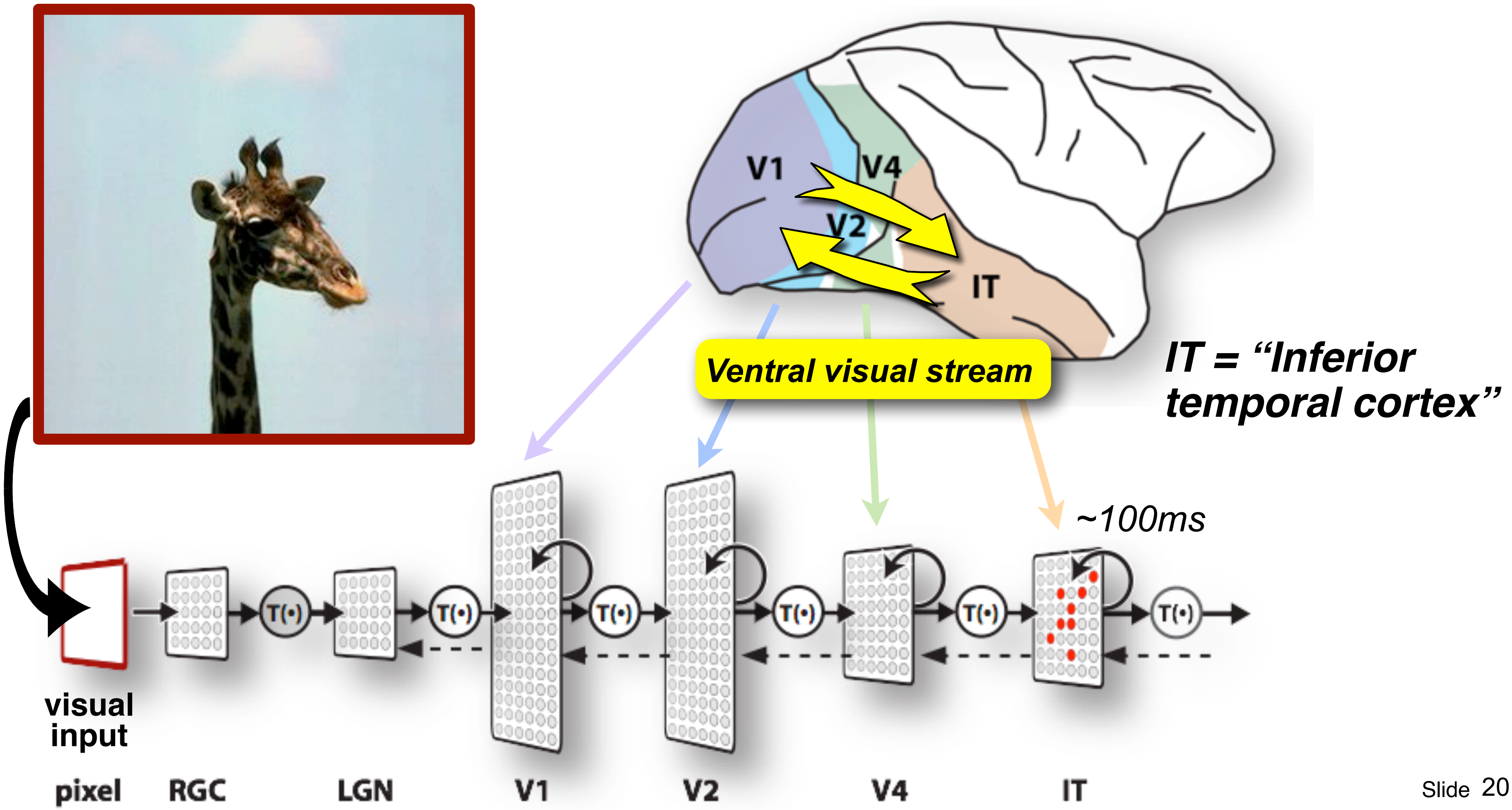


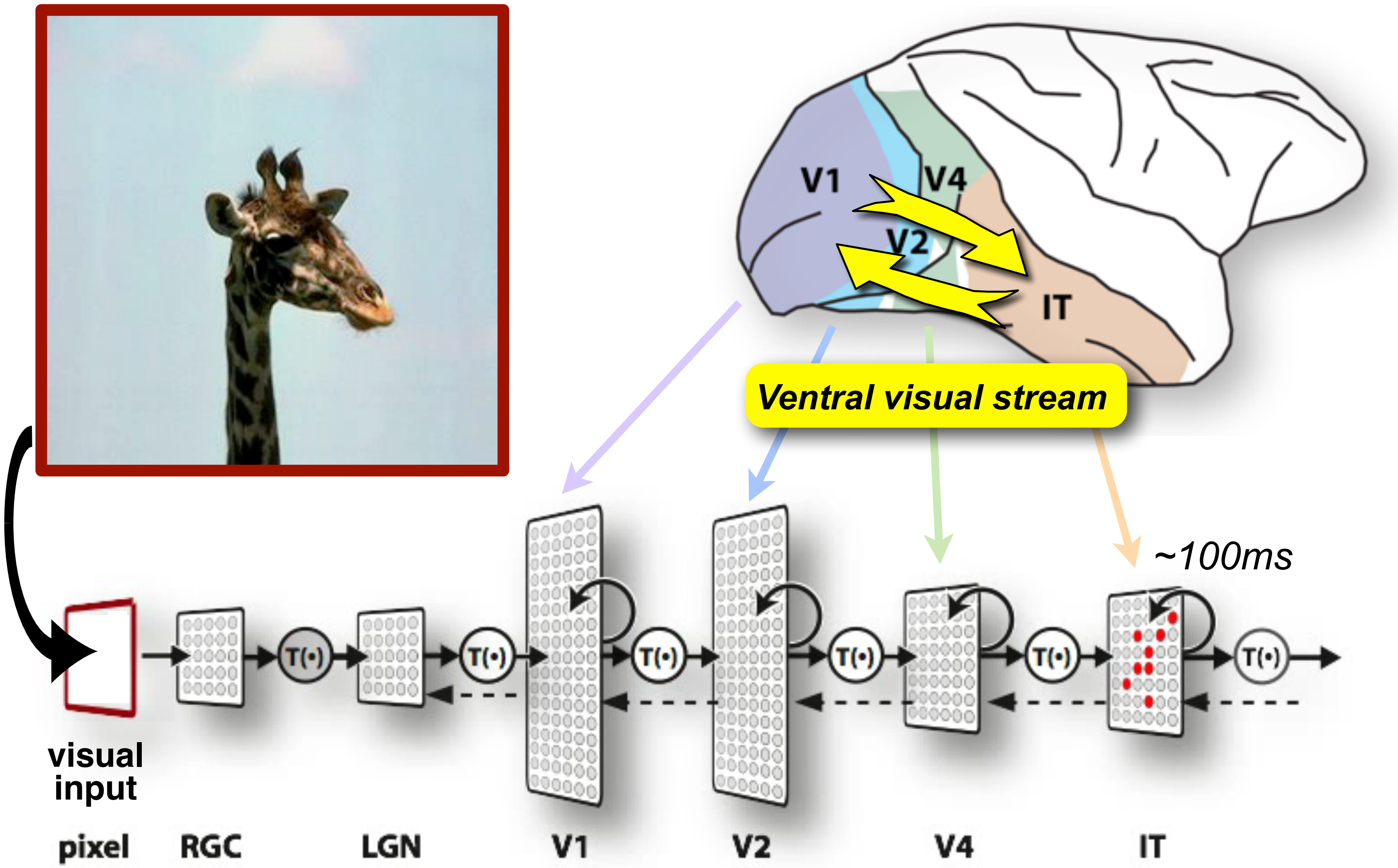
Lesions here result in deficits in object recognition.
V1
V4
V2
IT
Ventral visual stream
IT = "Inferior temporal cortex"
T(•)
visual input
pixel
RGC
LGN
V1
V2
V4
IT

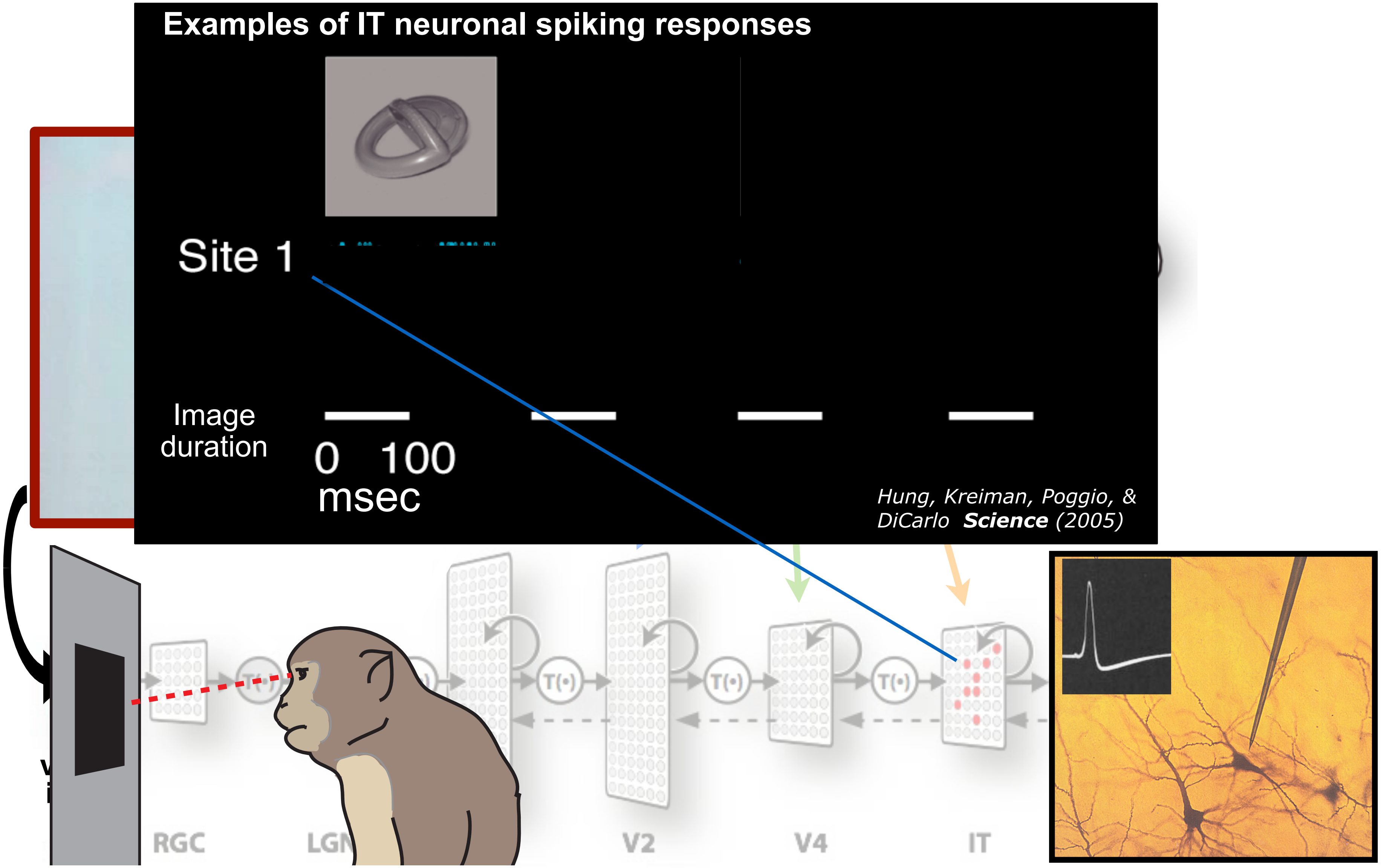


IT = "Inferior temporal cortex"

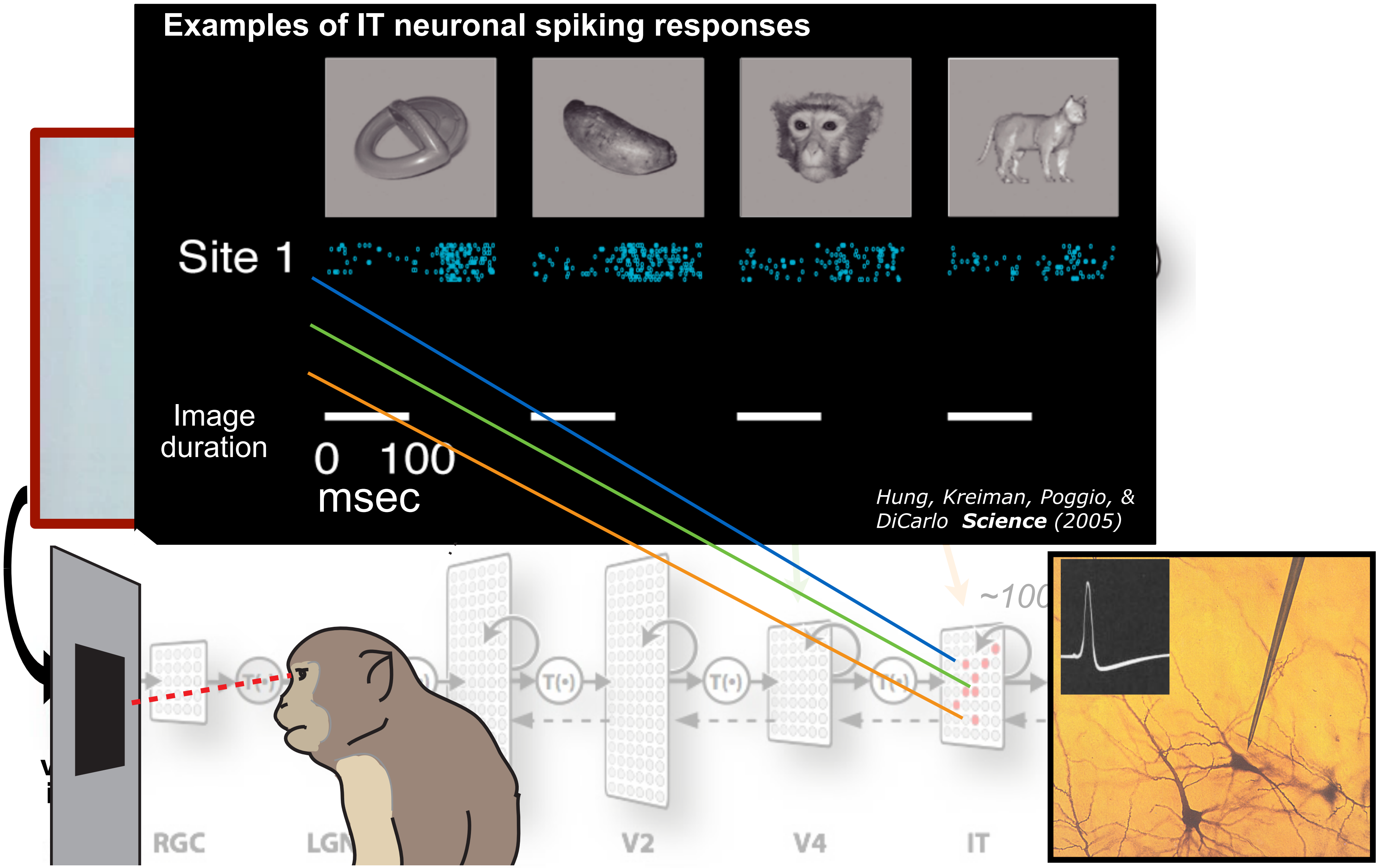

V1
V4
V2
IT
Ventral visual stream
IT = “Inferior temporal cortex”
~100ms
T(•)
visual input
pixel
RGC
LGN
V1
V2
V4
IT

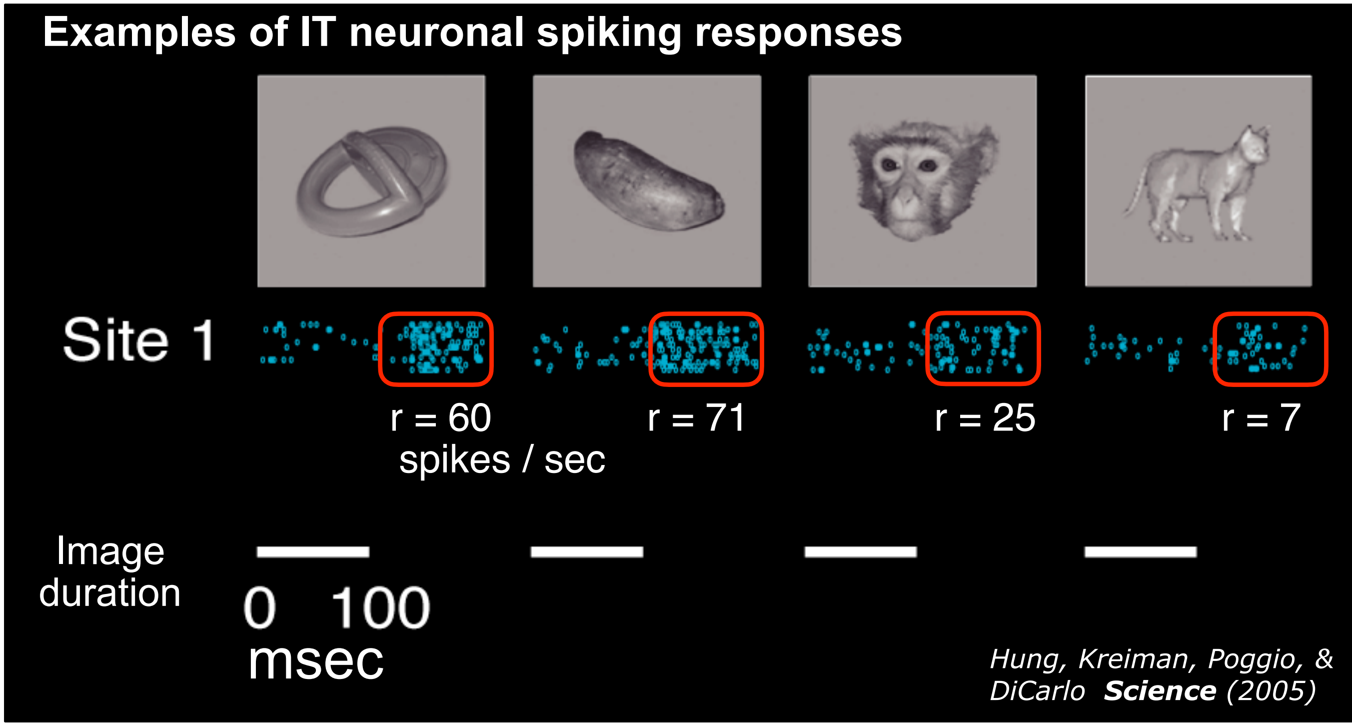

V1
V4
V2
IT
Ventral visual stream
IT = "Inferior temporal cortex"
~100ms
T(•)
visual input
pixel
RGC
LGN
V1
V2
V4
IT

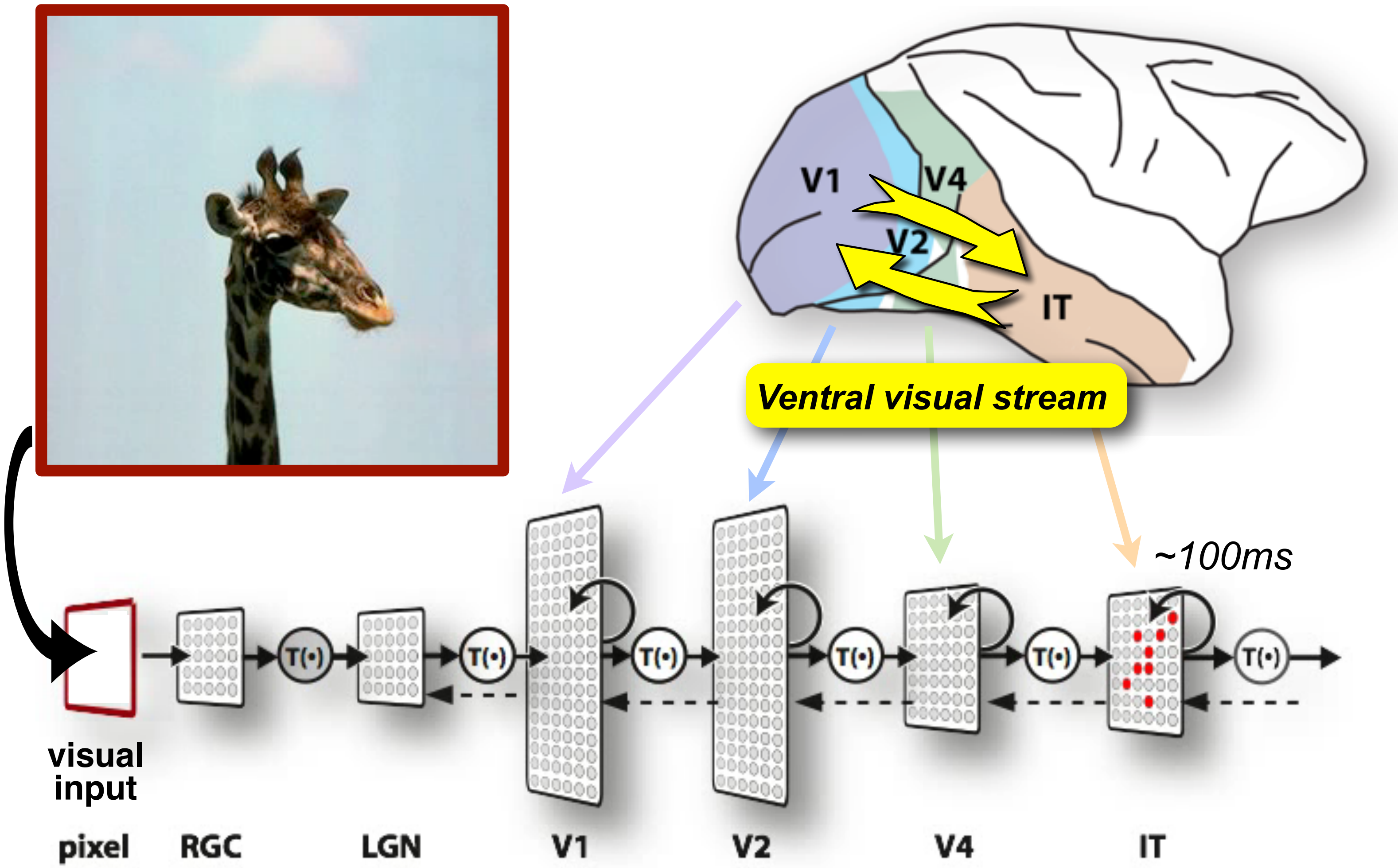

V1
V4
V2
IT
Ventral visual stream
~100ms
T(•)
visual
input
pixel
RGC
LGN
V1
V2
V4
IT

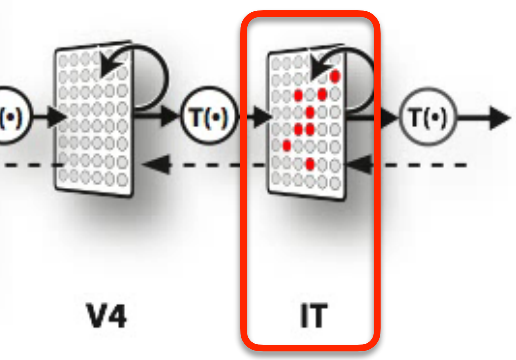


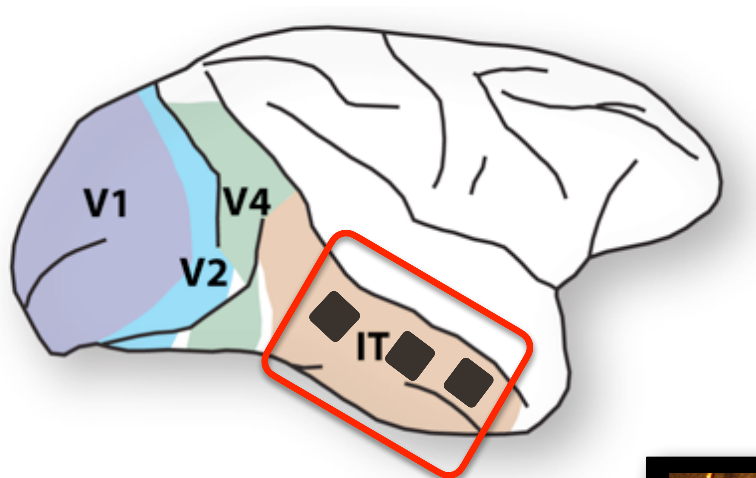
Examples of IT neuronal spiking responses
Site 1
Image duration
0 100
msec
Hung, Kreiman, Poggio, & DiCarlo **Science** (2005)
T(•)
T(•)
T(•)
RGC
LGN
V2
V4
IT

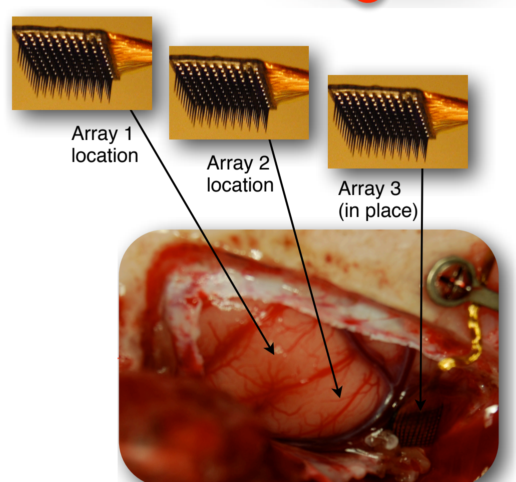

Examples of IT neuronal spiking responses
Site 1
Image
duration
0 100
msec
Hung, Kreiman, Poggio, &
DiCarlo Science (2005)
~100
RGC
LGN
V2
V4
IT

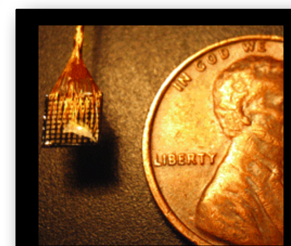

Examples of IT neuronal spiking responses
Site 1
r = 60
spikes / sec
r = 71
r = 25
r = 7
Image
duration
0 100
msec
Hung, Kreiman, Poggio, &
DiCarlo Science (2005)

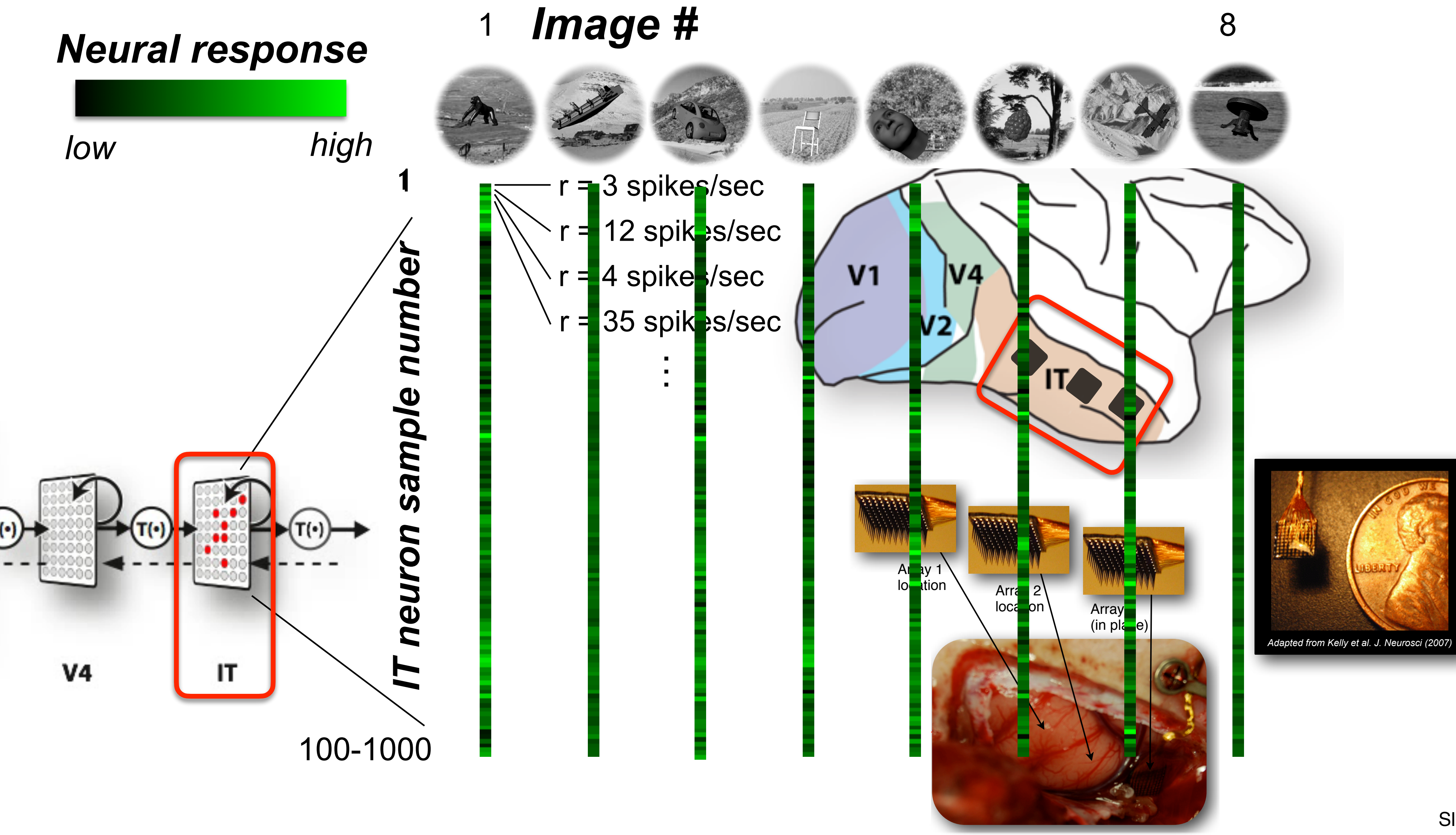

V1
V4
V2
IT
Ventral visual stream
~100ms
T(•)
visual
input
pixel
RGC
LGN
V1
V2
V4
IT



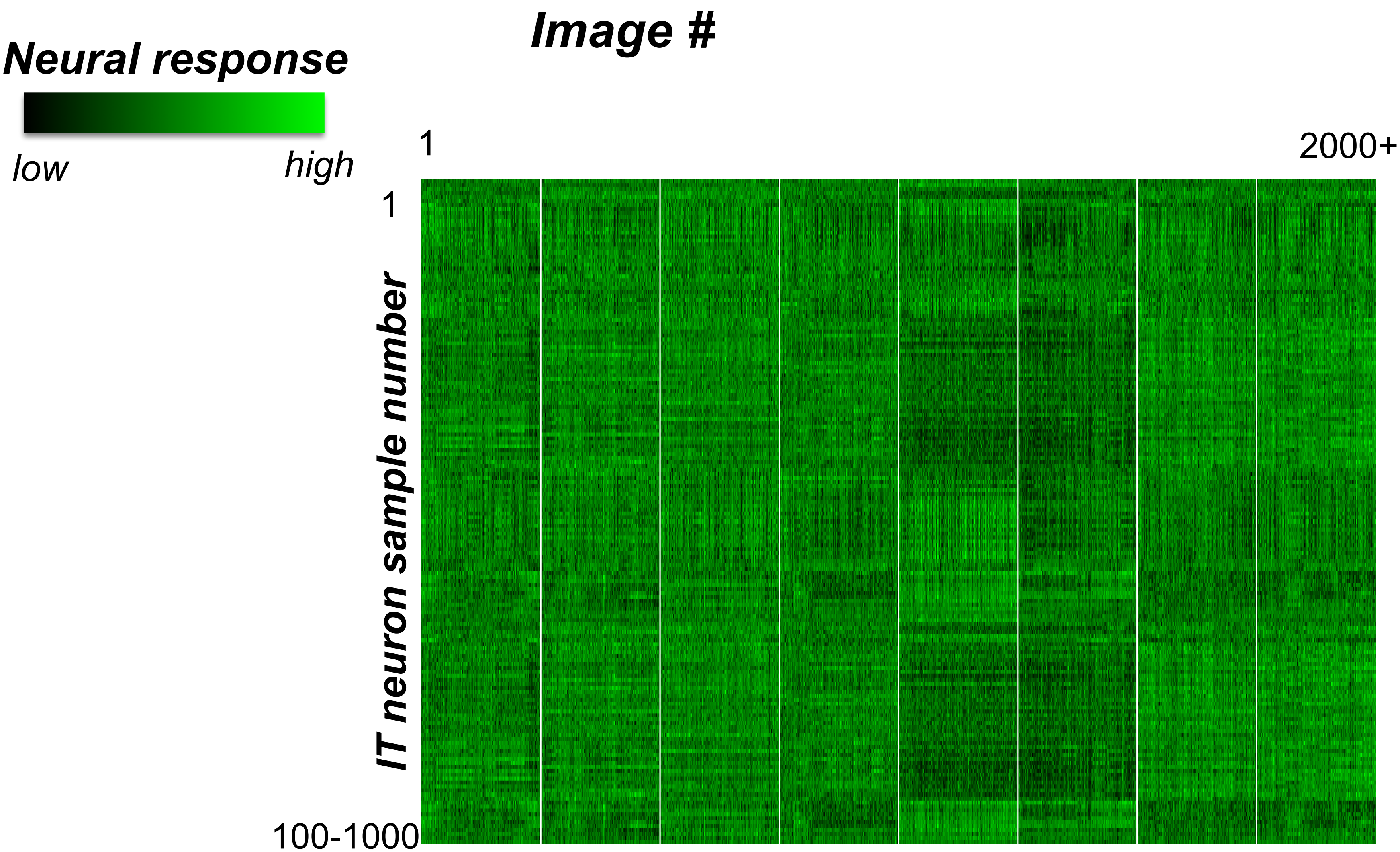





Adapted from Kelly et al. J. Neurosci (2007)



Neural response
low
high
1
Image #
8
1
r = 3 spikes/sec
r = 12 spikes/sec
r = 4 spikes/sec
r = 35 spikes/sec
IT neuron sample number
100-1000
T(•)
T(•)
V4
IT
V1
V4
V2
IT
Adapted from Kelly et al. J. Neurosci (2007)


Neural response
low
high
Image #
1
2000+
1
IT neuron sample number
100-1000

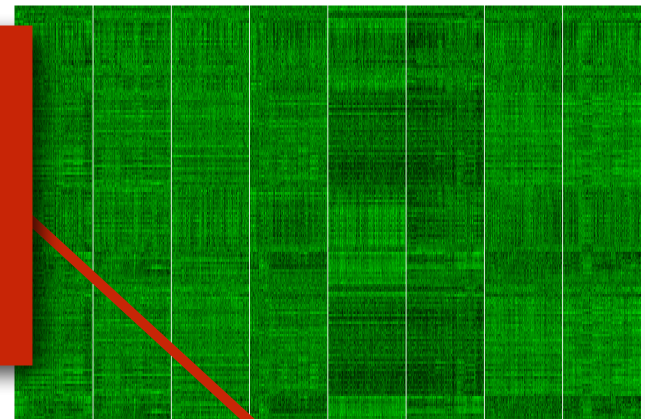
Hung*, Kreiman*, Poggio and DiCarlo, **Science** (2005);
Rust & DiCarlo, **J Neuroscience** (2010);
Majaj et al. **J Neuroscience** (2015)

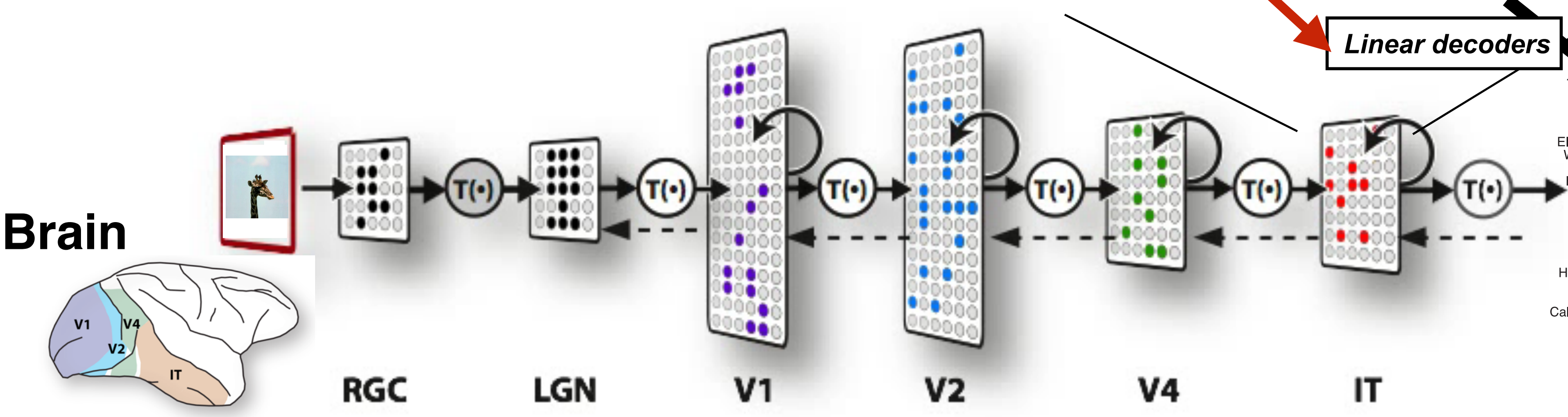
The IT neural population representation explains & predicts object recognition behavior !

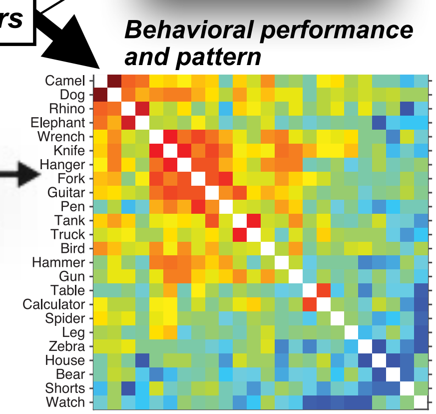
The parameters of this model of object perception tell us how we should manipulate IT neural responses to predictably modify object percepts.

(Afraz et al. ***PNAS*** *2015 ; Rajalingham,* ***Neuron*** *2018)*

Towards brain machine interfaces …



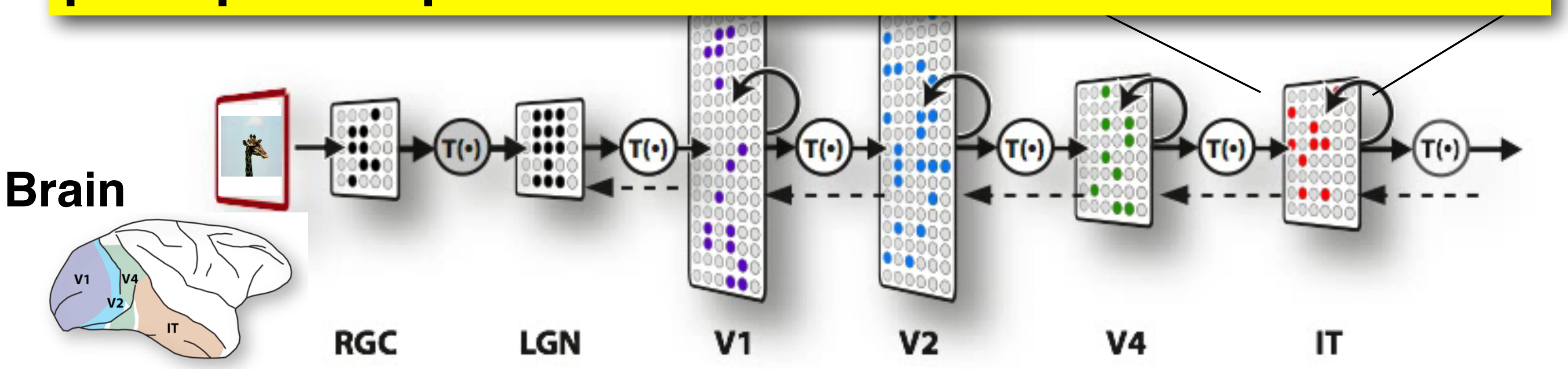




The IT neural population representation explains & predicts object recognition behavior

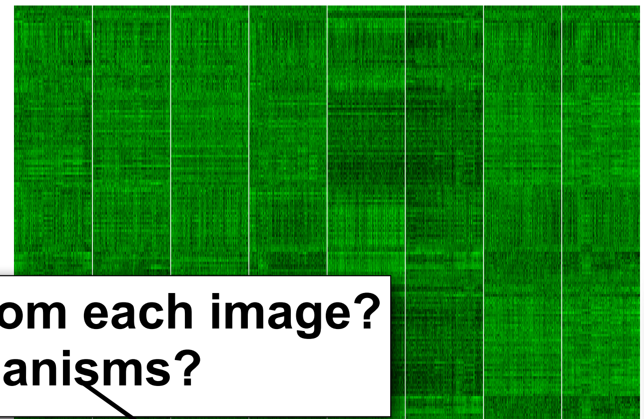
AI relevance: Primates are behaviorally higher performing than computer vision systems because their brain can compute this IT neural representation !

One key take-away: explaining the mean IT firing rates is ~sufficient to (computationally) explain behavior & perceptual report



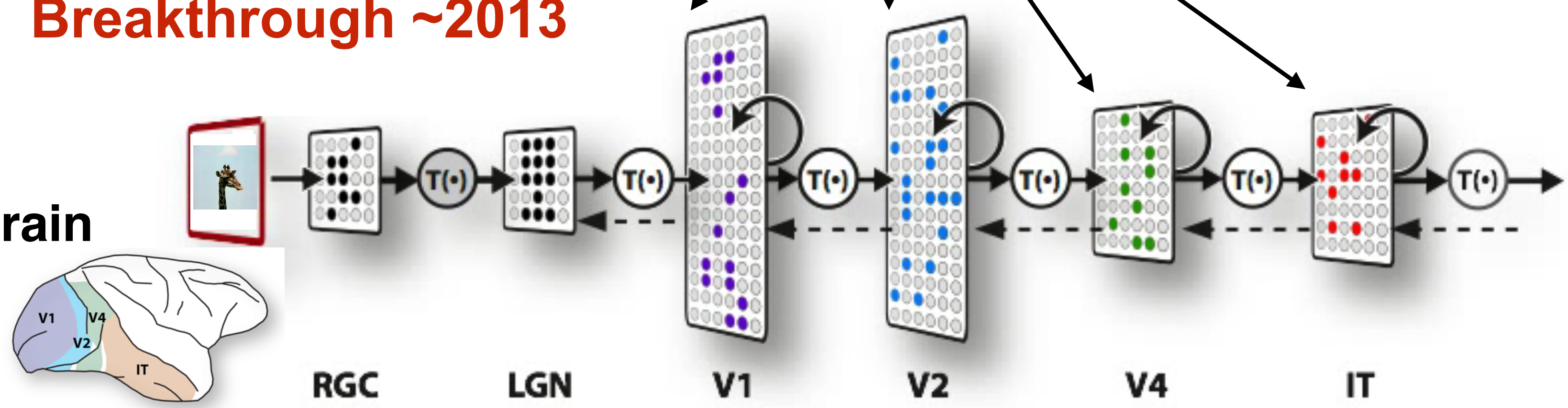
The IT neural population representation explains & predicts object recognition behavior

AI relevance: Primates are behaviorally higher performing than computer vision systems because their brain can compute this IT neural representation !



But <u>how</u> is the IT representation computed from each image? That is, what are the intervening neural mechanisms?

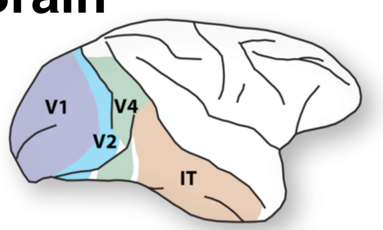
Breakthrough ~2013

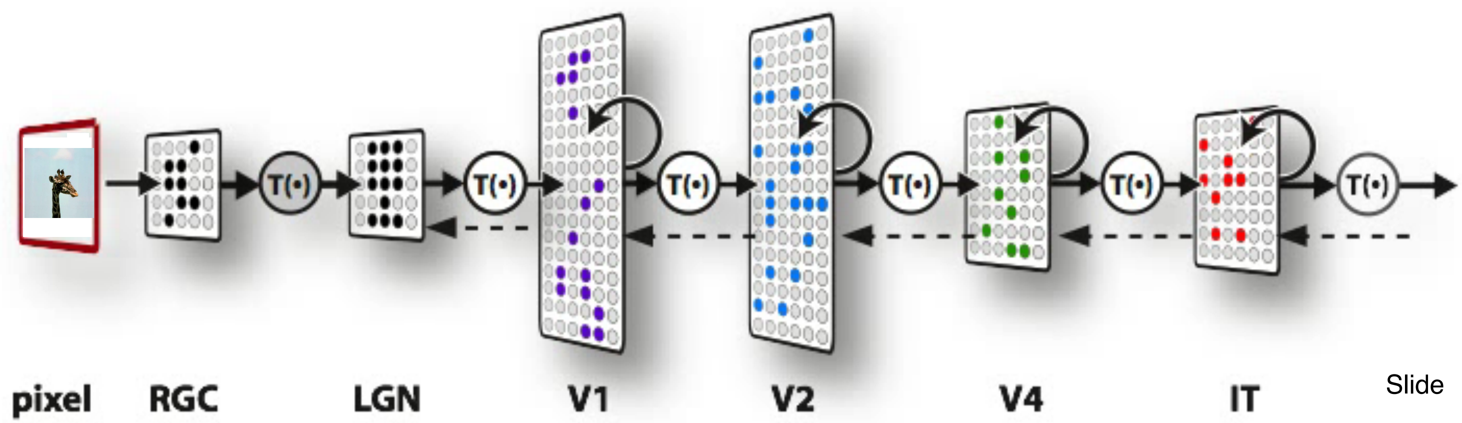


Background

GUIDANCE FROM NEUROSCIENCE (many labs):

Brain



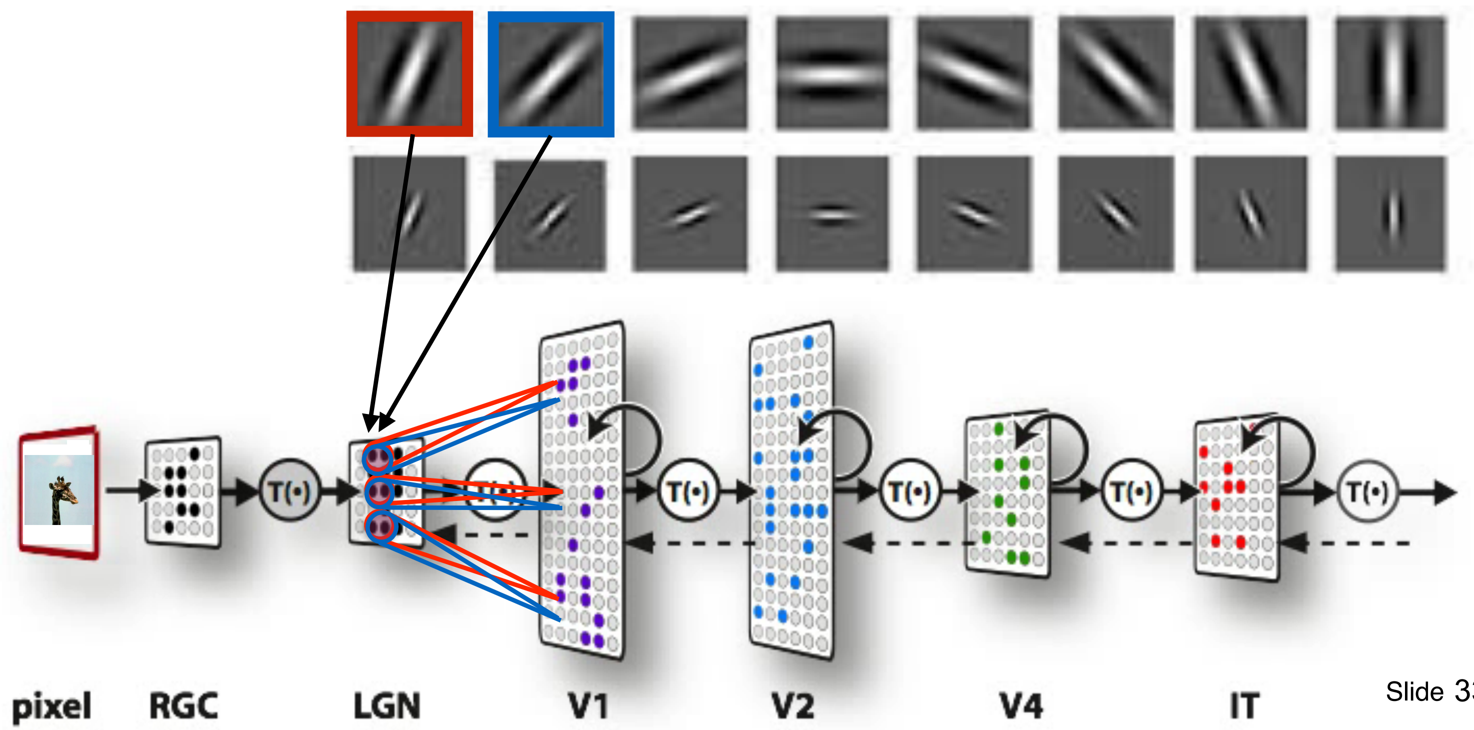


Background

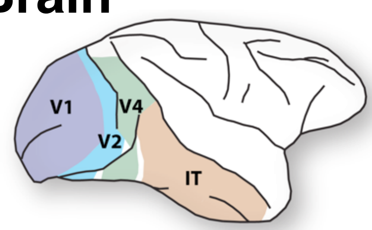
GUIDANCE FROM NEUROSCIENCE (many labs):

Each "area" of processing:

- ***Spatially local, ~linear filters***
- ***Different types of such filters***
- ***Each repeated spatially over the input (~convolution)***



Brain

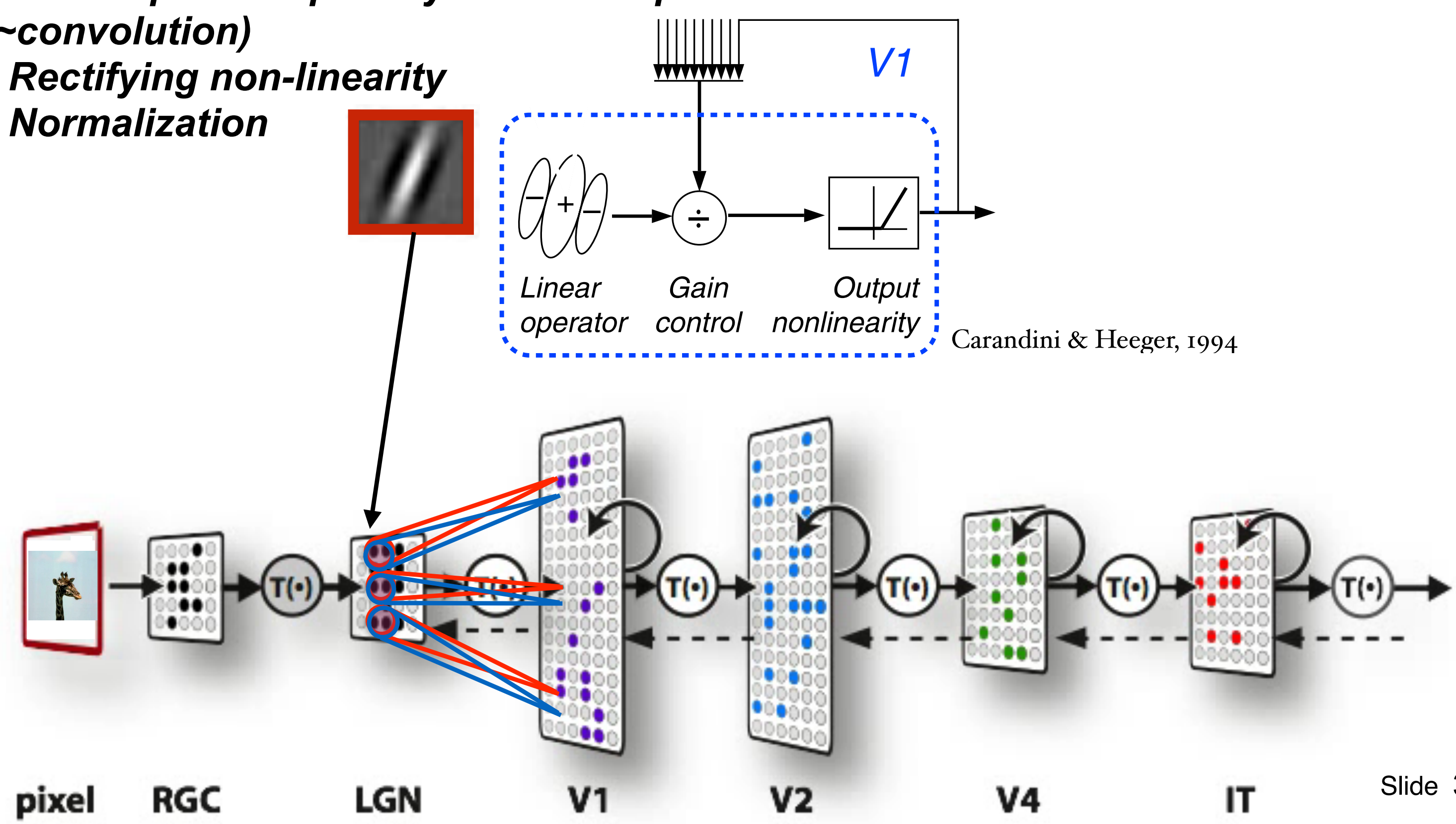


Background

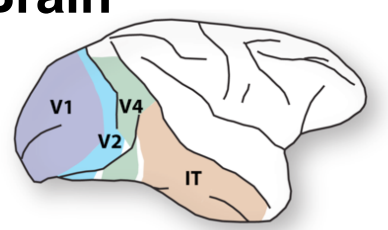
GUIDANCE FROM NEUROSCIENCE (many labs):

Each "area" of processing:

- ***Spatially local, ~linear filters***
- ***Different types of such filters***
- ***Each repeated spatially over the input (~convolution)***
- ***Rectifying non-linearity***
- ***Normalization***



Brain



Background

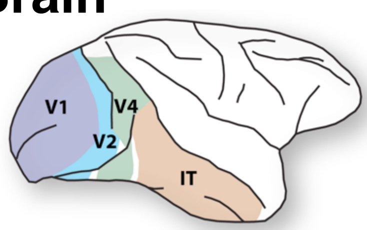
GUIDANCE FROM NEUROSCIENCE (many labs):

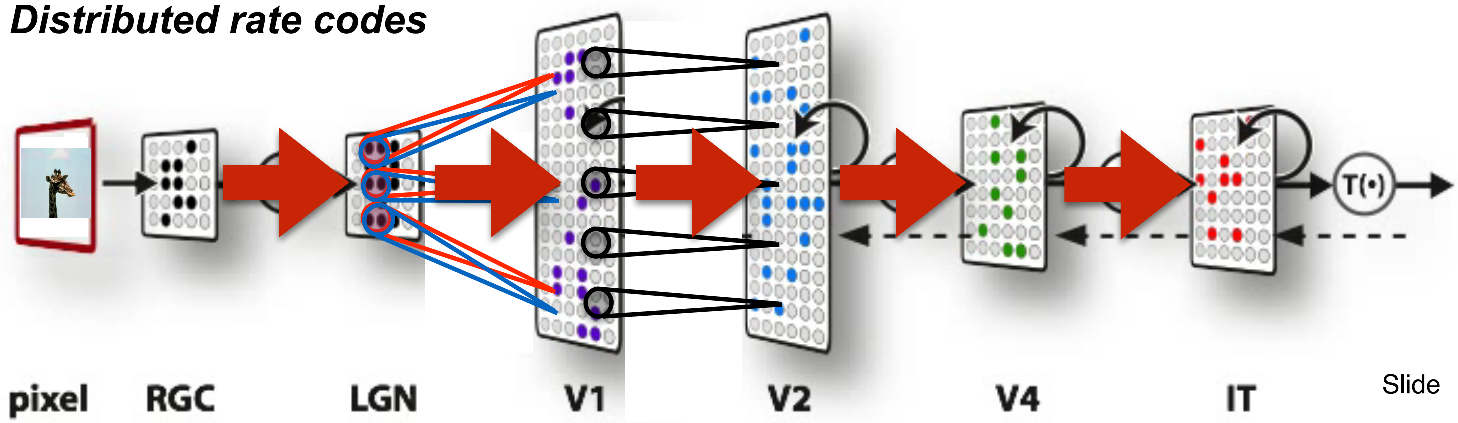
Each "area" of processing:
- ***Spatially local, ~linear filters***
- ***Different types of such filters***
- ***Each repeated spatially over the input (~convolution)***
- ***Rectifying non-linearity***
- ***Normalization***

- ***"Deep" <u>series of areas</u>***
- ***Similar "style" operations at each successive area***

- ***Fast ~feedforward does a lot!***
- ***Distributed rate codes***

Brain





Background

GUIDANCE FROM NEUROSCIENCE (many labs):

Each "area" of processing:
- ***Spatially local, ~linear filters***
- ***Different types of such filters***
- ***Each repeated spatially over the i (~convolution)***
- ***Rectifying non-linearity***
- ***Normalization***

- ***"Deep" <u>series of areas</u>***
- ***Similar "style" operations at each successive area***

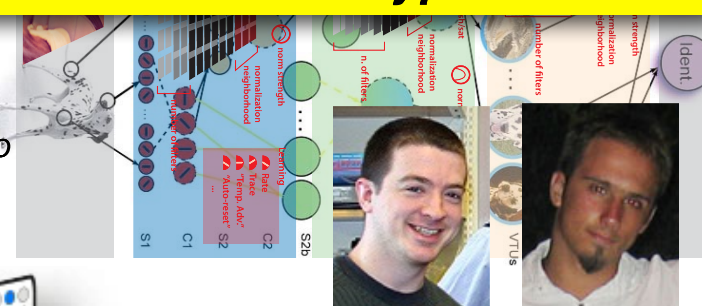
- ***Fast ~feedforward does a lot!***
- ***Distributed rate codes***

Resulted in proposed feedforward artificial neural networks (ANNs):

Pinto and Cox 2008, 2010

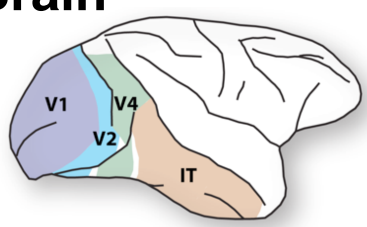
The building of such models is critically important to basic science research: Each is a testable mechanistic hypothesis!

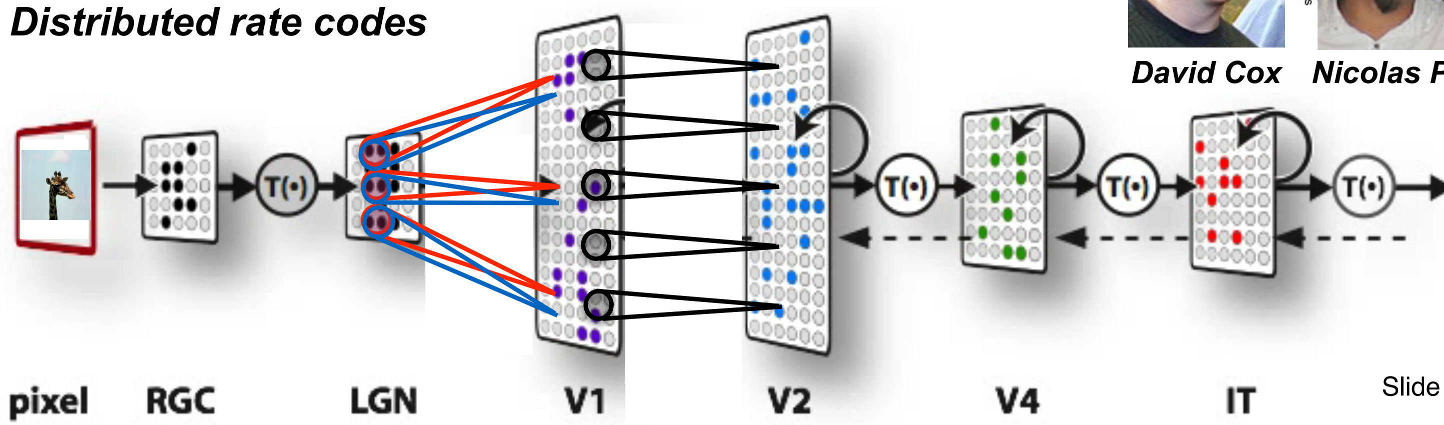
Tomaso Poggio
BCS/MIT



David Cox ***Nicolas Pinto***

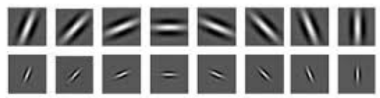
Brain





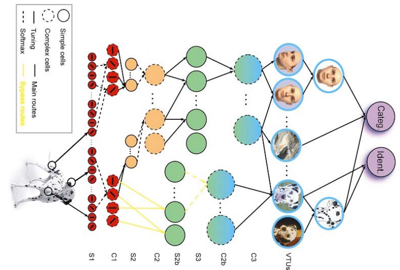
Hypotheses (specific ANN models):

~1980-2010

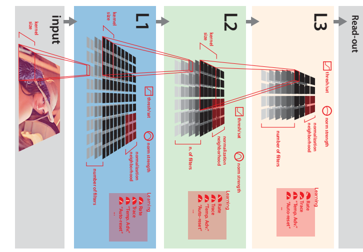


"V1-like" models

"V2-like" models



HMAX models



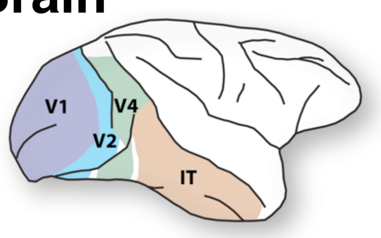
PLoS09 models

...

GUIDANCE FROM NEUROSCIENCE:

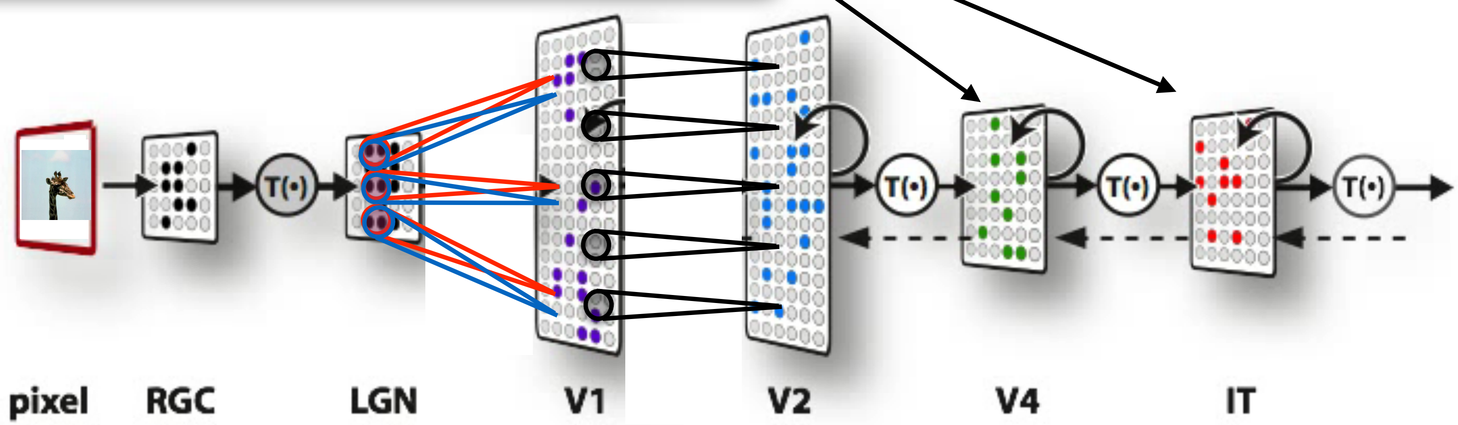
SCIENCE

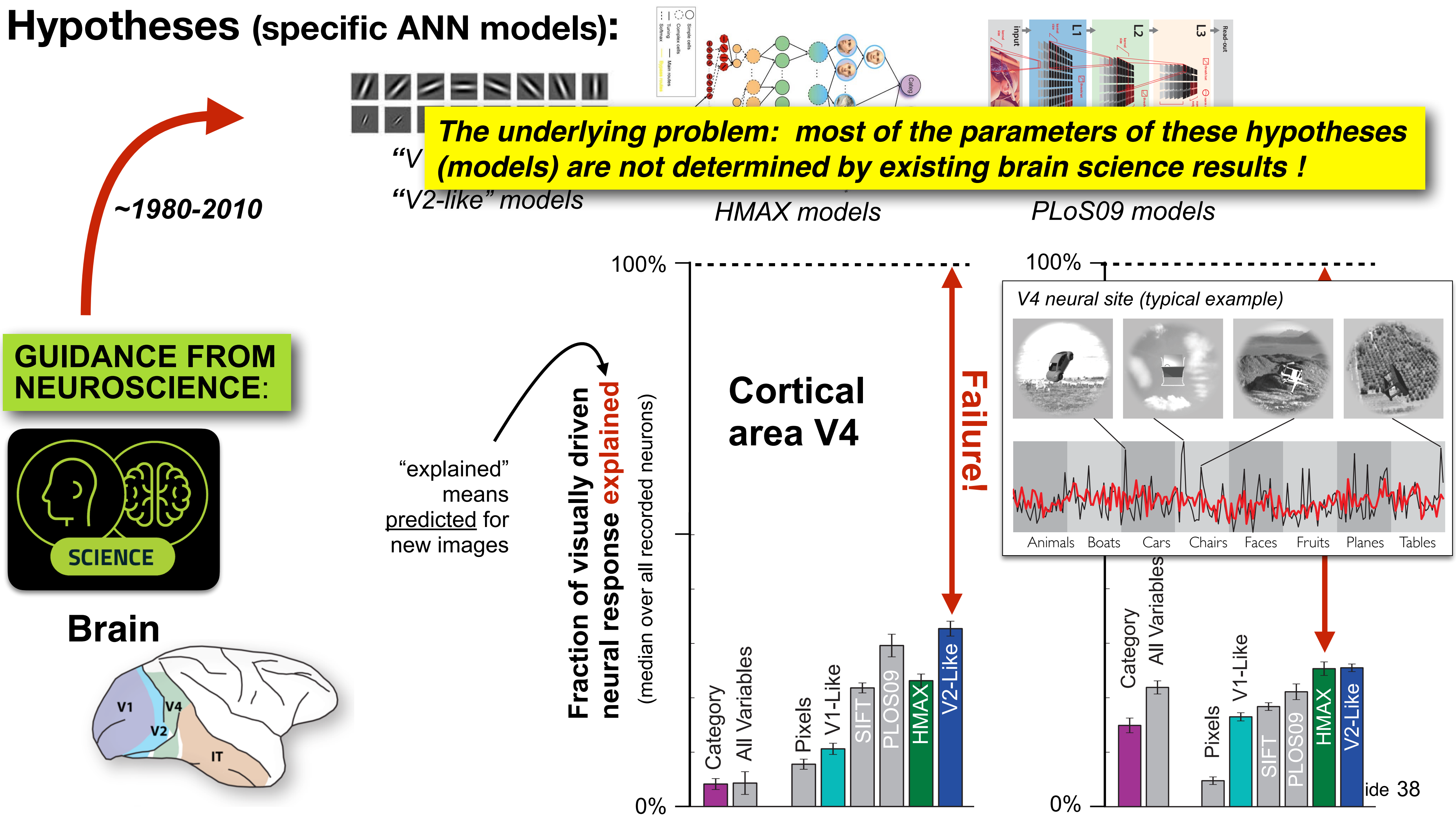
Brain



Unfortunately, <u>all</u> of these specific mechanistic hypotheses were inadequate.

E.g. they each failed to accurately predict the internal neural responses to new sets of test images.

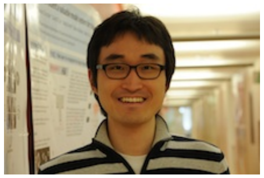




Hypotheses (specific ANN models):
~1980-2010
GUIDANCE FROM NEUROSCIENCE:
SCIENCE
Brain
V1
V2
V4
IT
"V
"V2-like" models
HMAX models
PLoS09 models
The underlying problem: most of the parameters of these hypotheses (models) are not determined by existing brain science results !
"explained" means predicted for new images
Fraction of visually driven neural response explained
(median over all recorded neurons)
100%
0%
Cortical area V4
Failure!
Category
All Variables
Pixels
V1-Like
SIFT
PLOS09
HMAX
V2-Like
V4 neural site (typical example)
Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables
ide 38

~2013: Collaborative breakthrough

Dan Yamins



Ha Hong

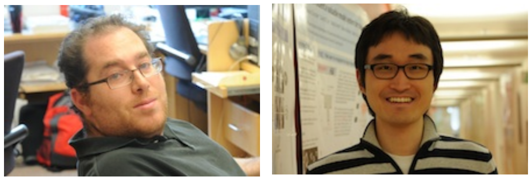
Yamins, Hong, Solomon, Seibert and DiCarlo ***NIPS (2013), PNAS (2014)***

The underlying problem: most of the parameters of these hypotheses (models) are not determined by existing brain science results !

~2013: Collaborative breakthrough

Cognitive science guided the task

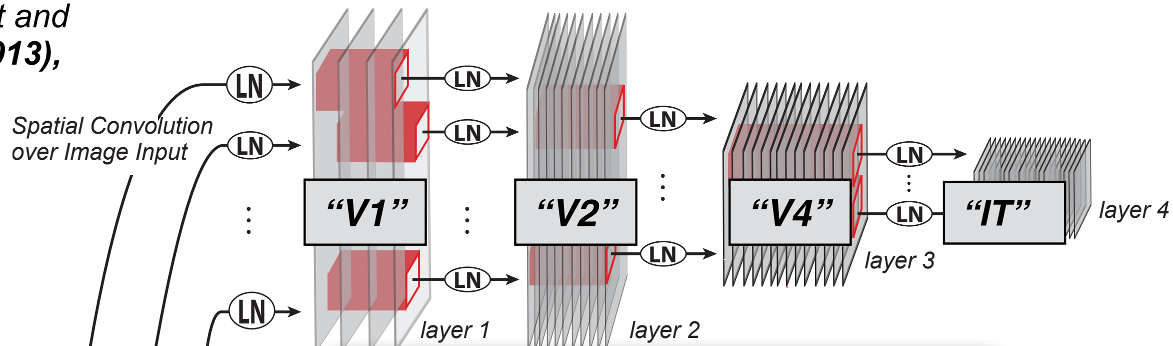
The underlying problem: most of the parameters of these hypotheses (models) are not determined by existing brain science results !



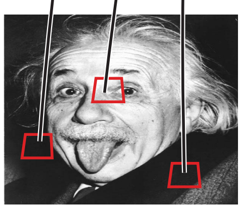
Dan Yamins ***Ha Hong***

Yamins, Hong, Solomon, Seibert and DiCarlo ***NIPS (2013), PNAS (2014****)*

Artificial neural network (ANN)

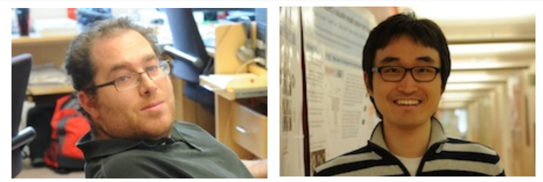


Neuroscience guided the parameters of system macro- and meso- architecture



Example test image (one of many)

~2013: Collaborative breakthrough

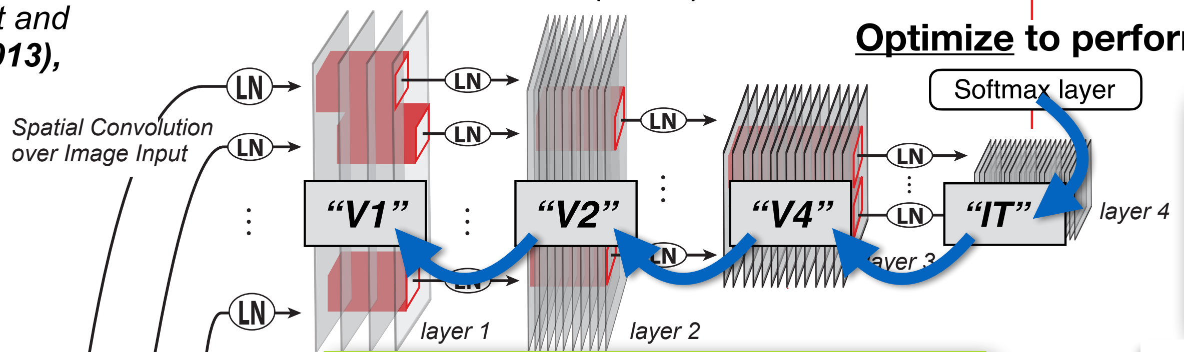

Dan Yamins *Ha Hong*

*Yamins, Hong, Solomon, Seibert and DiCarlo **NIPS (2013), PNAS (2014)***

Cognitive science guided the task

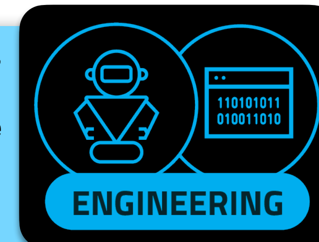
Core object recognition

Artificial neural network (ANN)



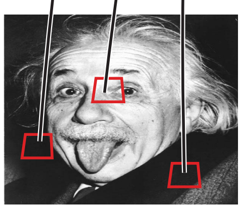
<u>Optimize</u> to perform this task!

Engineering tools to tune the microarchitecture parameters to perform well on this task !



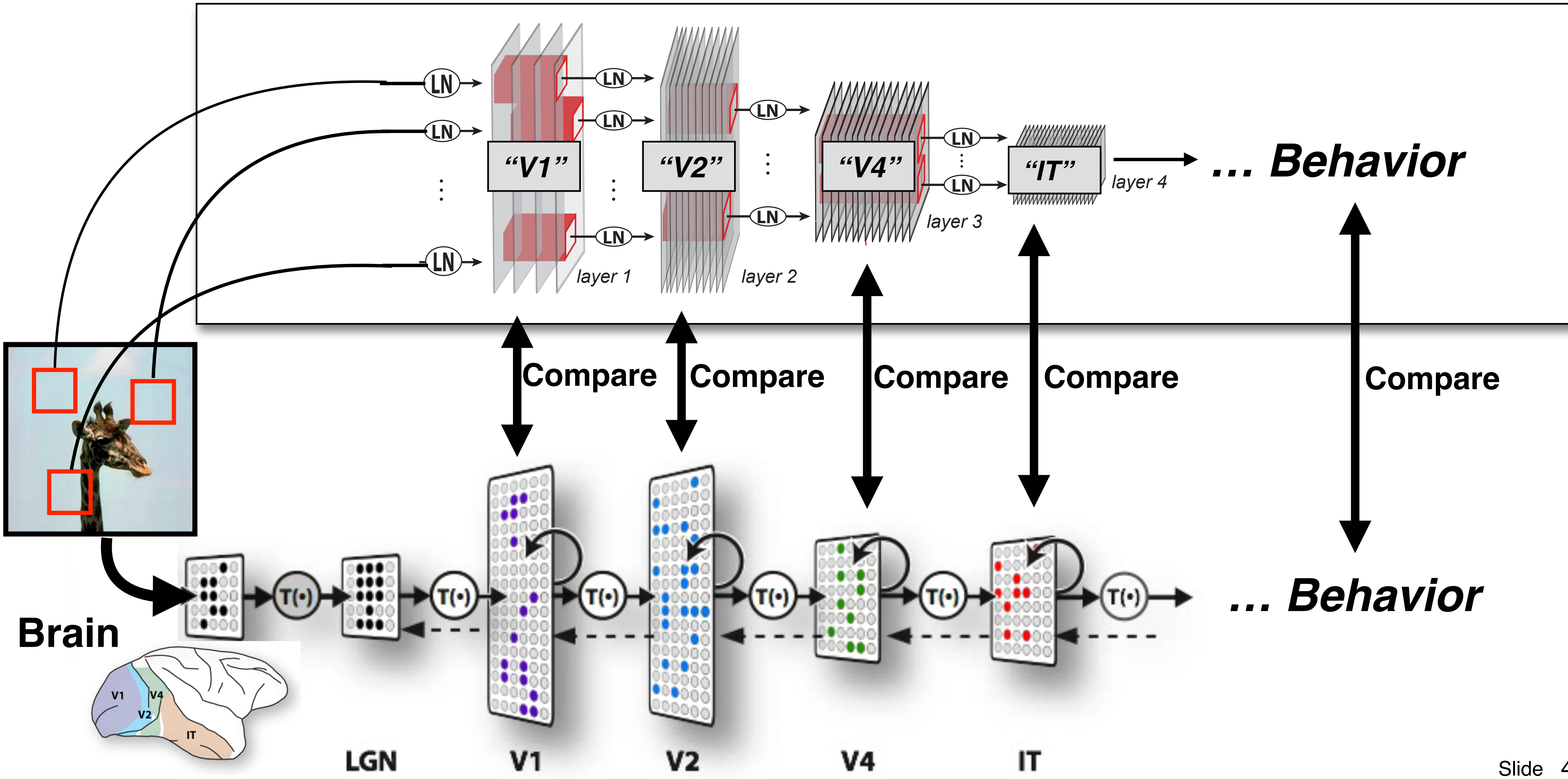
Neuroscience guided the parameters of system macro- and meso- architecture

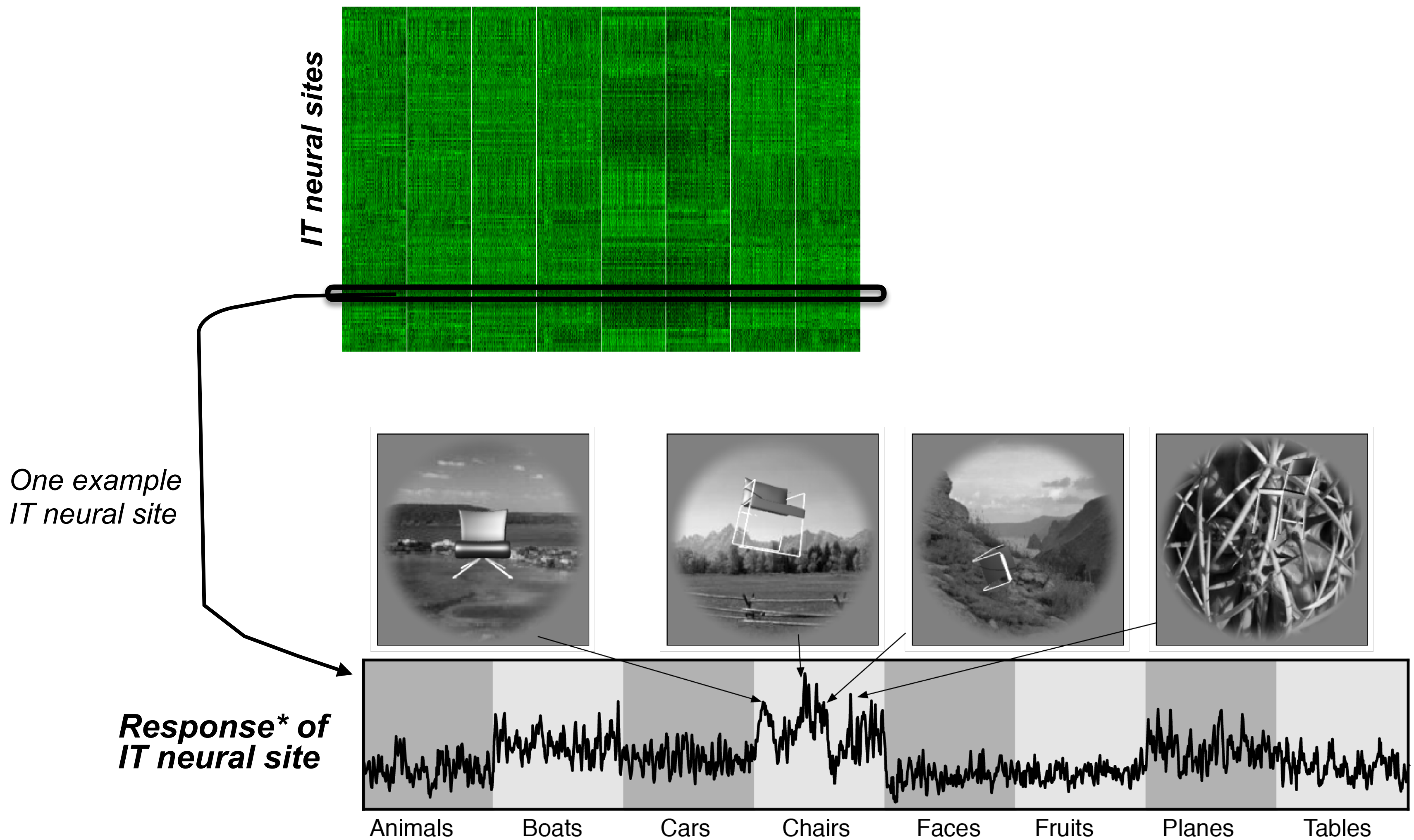
- Each choice of all system parameters is an entire new artificial ventral stream!


Example test image (one of many)

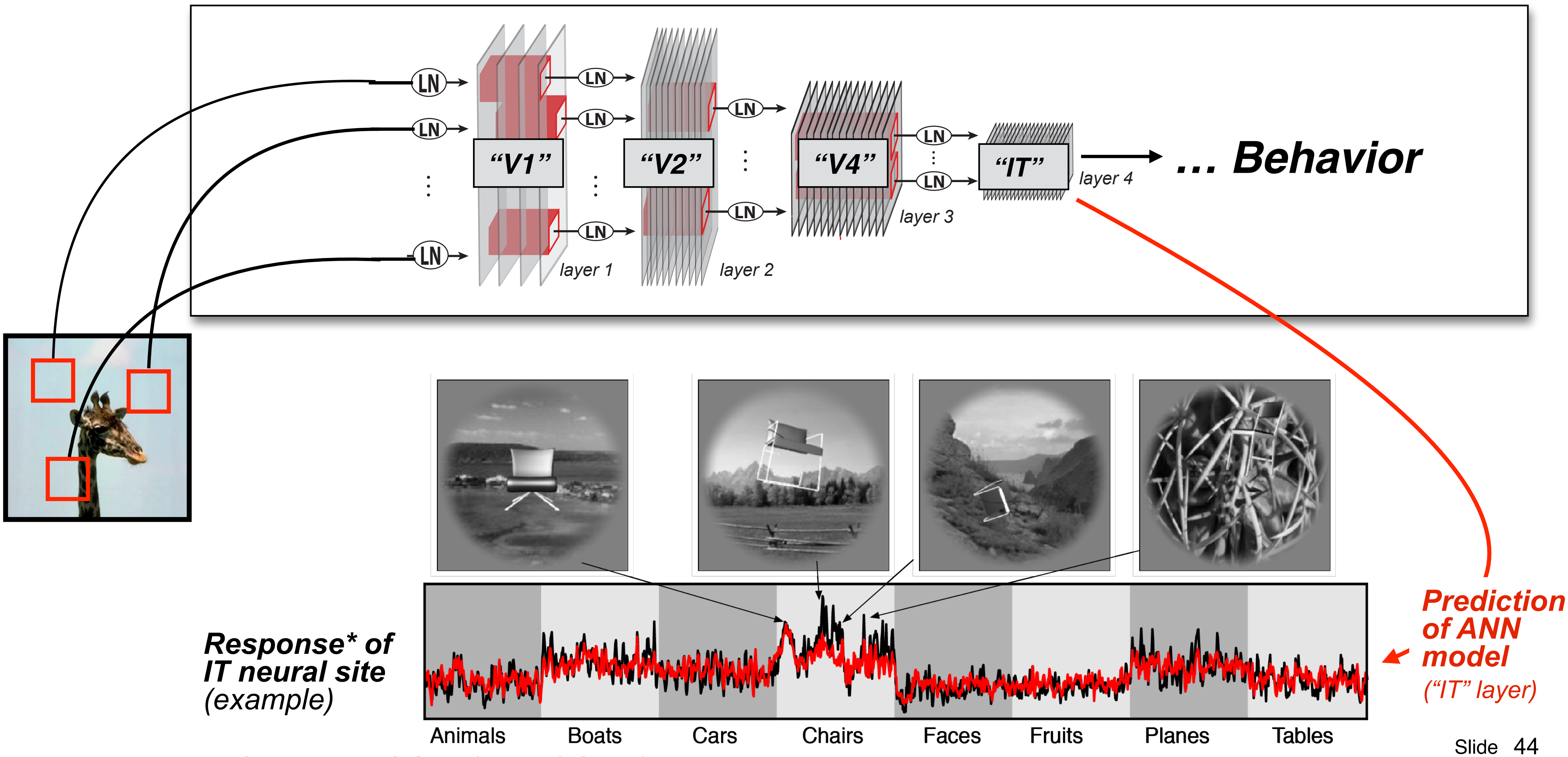
Discovery: the individual "neurons" inside some of these models behave very much like the individual neurons that we and others were recording in the monkey brain !

A specific deep ANN (a neurally-mechanistic scientific hypothesis!)



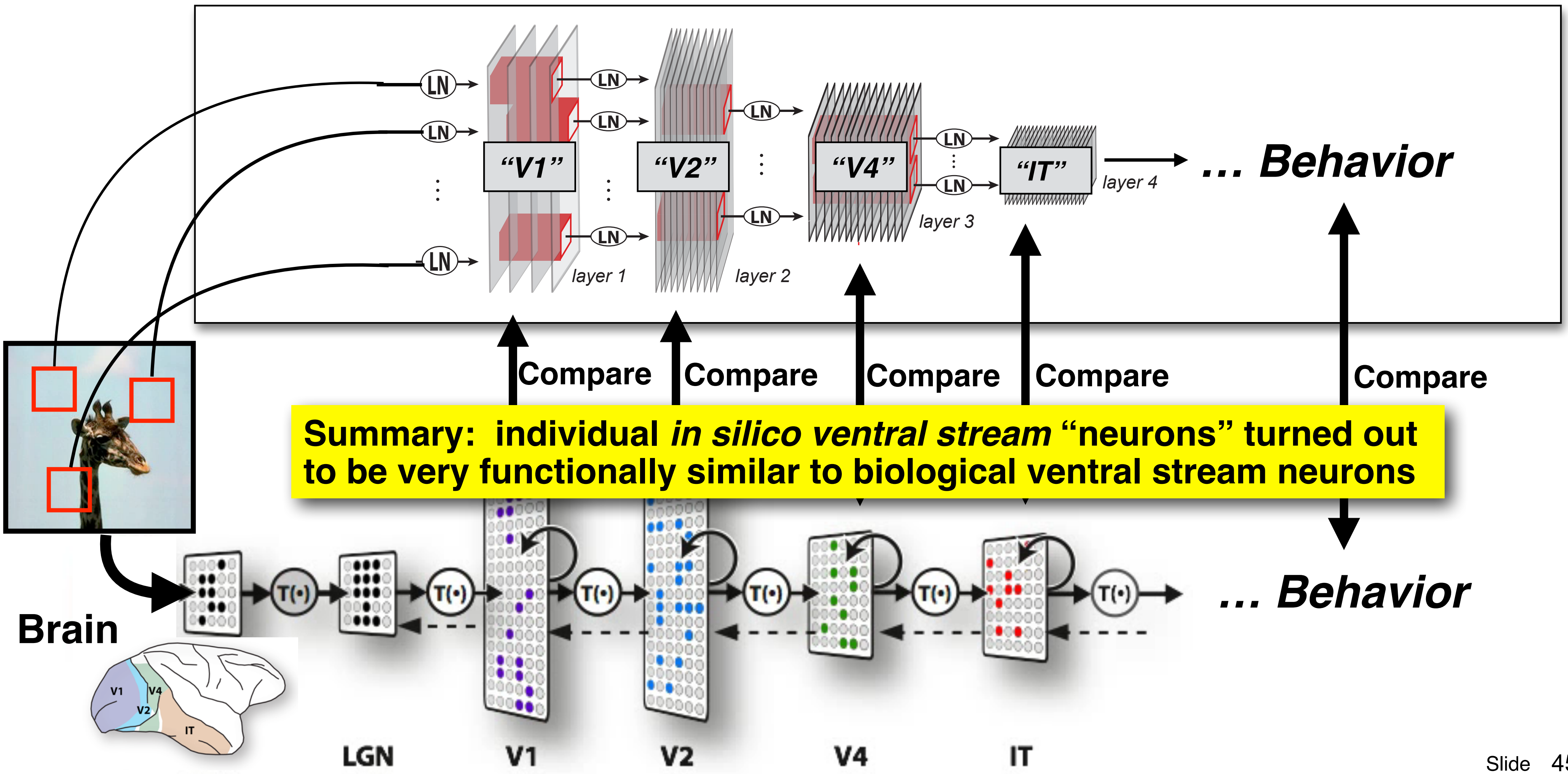

IT neural sites
One example
IT neural site
Response* of
IT neural site
Animals
Boats
Cars
Chairs
Faces
Fruits
Planes
Tables

A specific deep ANN (a neurally-mechanistic scientific hypothesis!)



Yamins, Hong, … DiCarlo ***NeurIPS (2013)***, ***PNAS (2014)***

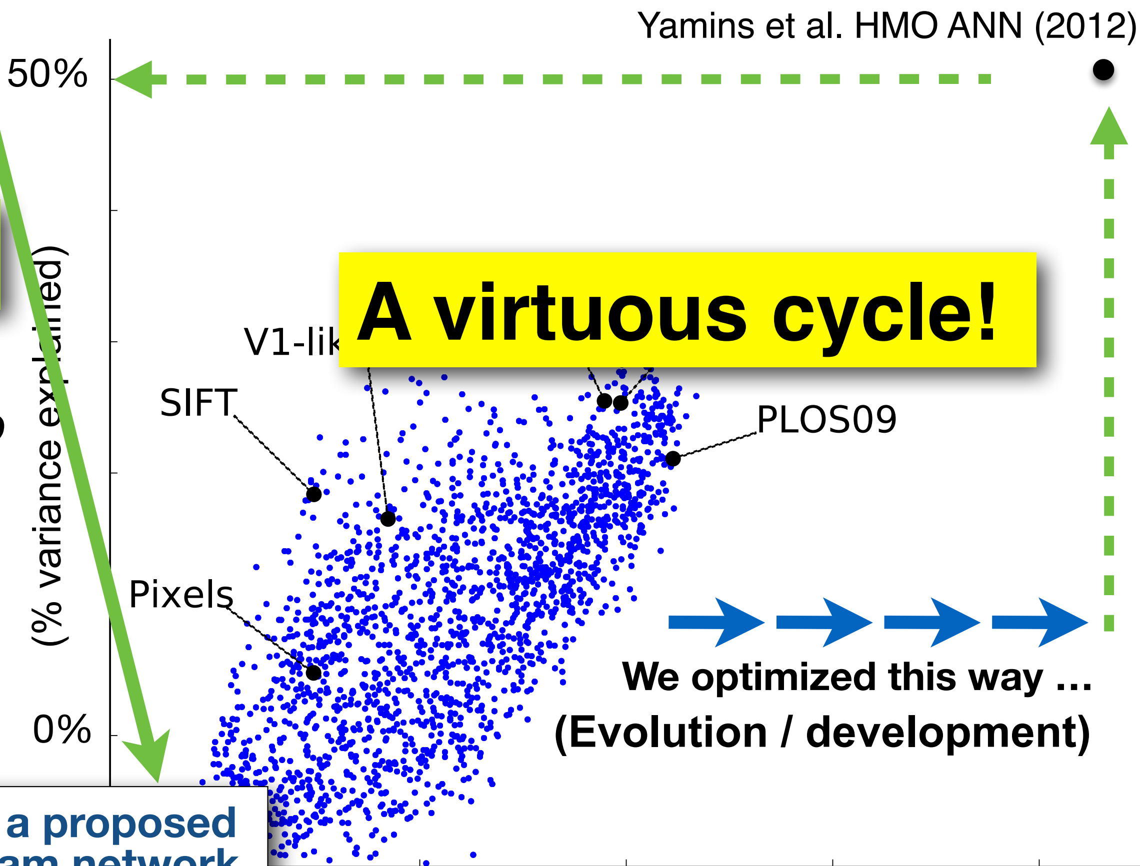
A specific deep ANN (a neurally-mechanistic scientific hypothesis!)



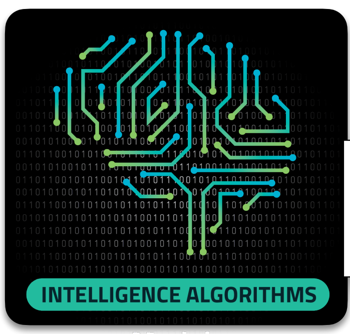
Summary: individual *in silico ventral stream* "neurons" turned out to be very functionally similar to biological ventral stream neurons

A neuroscience goal

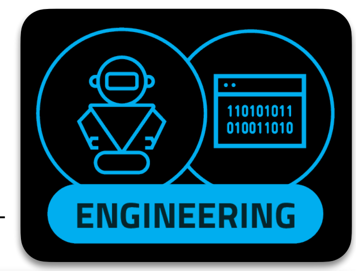
Match score of the *in silico* "IT" neurons with primate IT neurons



Accuracy on core object recognition tasks

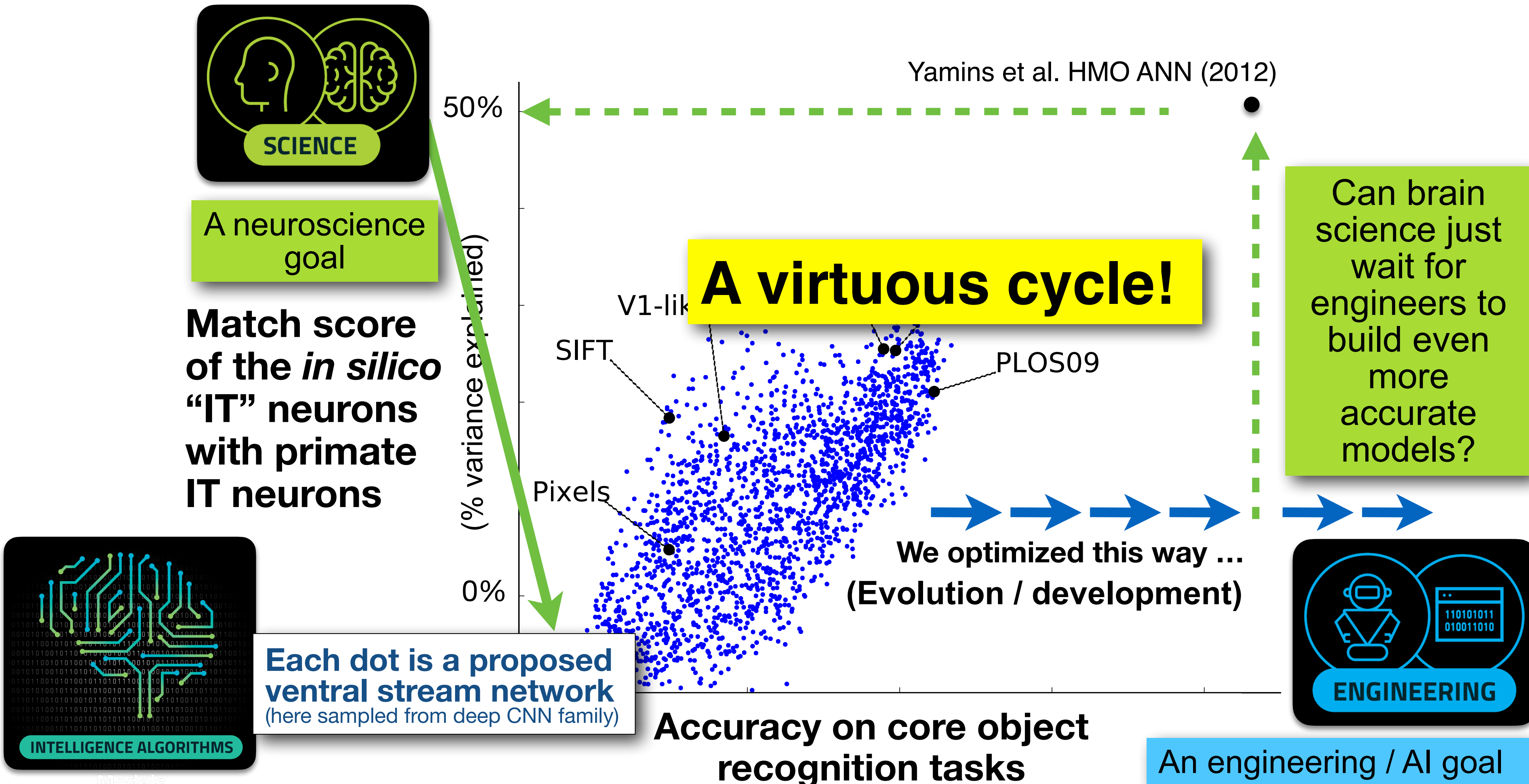


Each dot is a proposed ventral stream network
(here sampled from deep CNN family)



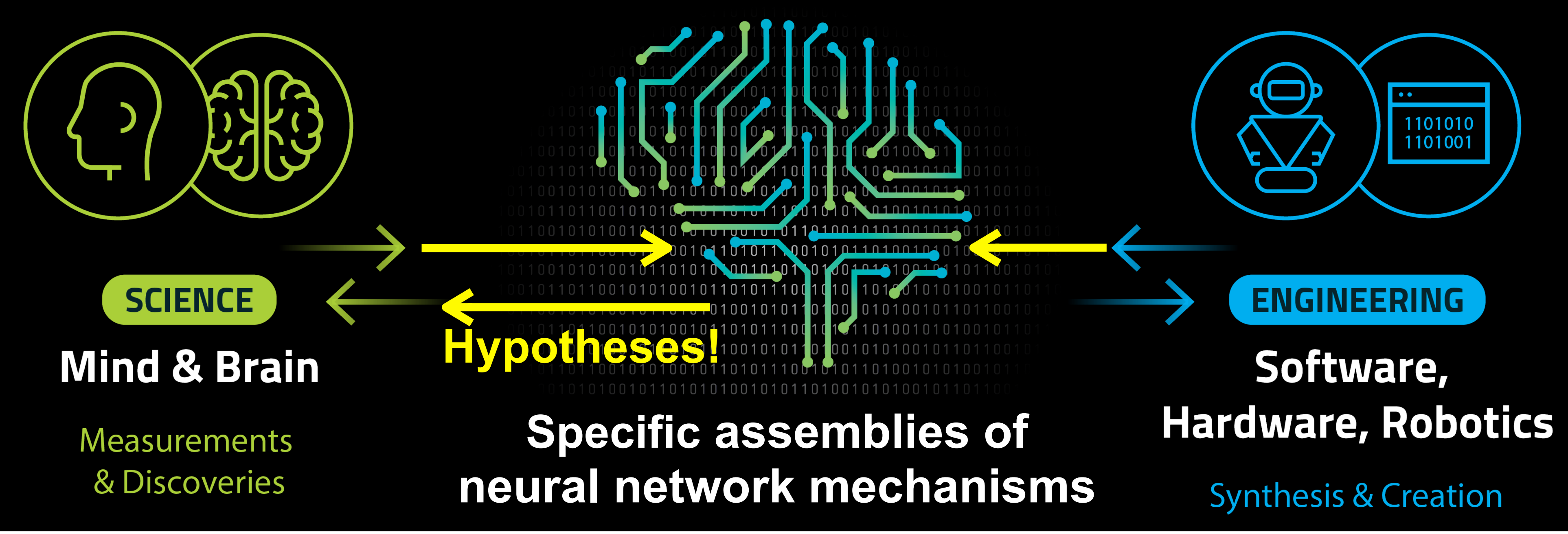
An engineering / AI goal

Adapted from Yamins, Hong, Solomon, Seibert and DiCarlo ***PNAS (2014)***



Adapted from Yamins, Hong, Solomon, Seibert and DiCarlo ***PNAS (2014)***

An implicit collaboration!



This approach is leading to rapid progress in other areas of brain science:

Vision (Kriegeskorte, Oliva, Konkle, Ganguli, Kanwisher, Tsao, …)
Audition (McDermott & Yamins)
Somatosensation (Hartmann & Yamins)
Decision making (Sussillo & Newsome, Freedman, …)
Motor planning and control (Jazayeri, Batista, Churchland, …)
Navigation (Fiete, …)

An implicit collaboration!
To the benefit of both fields

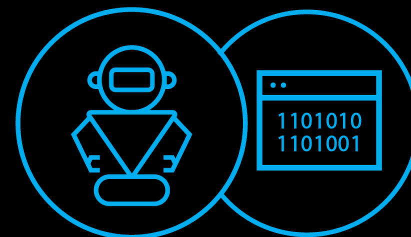
The New York Times

Scientists See Promise in Deep-Learning Programs

Using an artificial intelligence technique inspired by theories about how the brain recognizes patterns, technology companies are reporting startling gains in fields as diverse as computer vision, speech recognition and the identification of promising new molecules for designing drugs.

“AI”

es of
hanisms

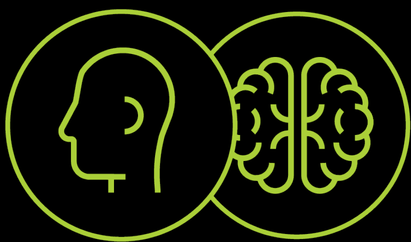


ENGINEERING

Software, Hardware, Robotics

Synthesis & Creation

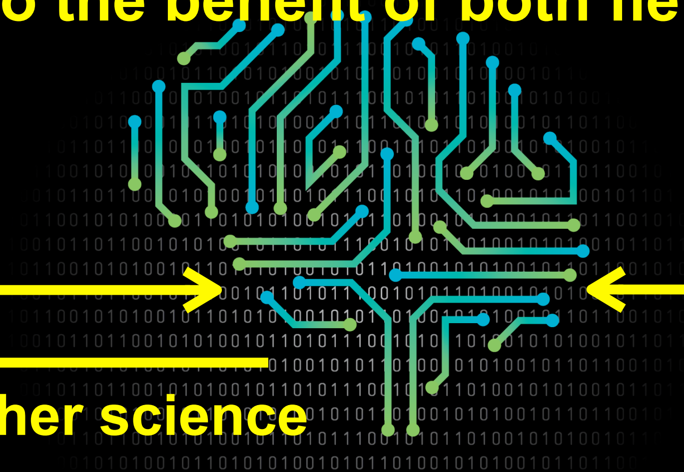
An implicit collaboration!
To the benefit of both fields



SCIENCE

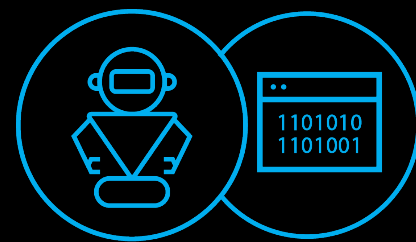
Mind & Brain

Measurements
& Discoveries



Further science

Specific assemblies of neural network mechanisms

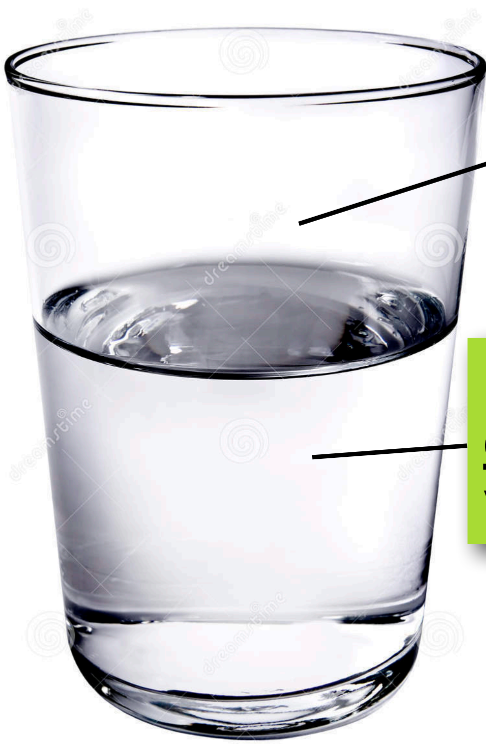


ENGINEERING

Software, Hardware, Robotics

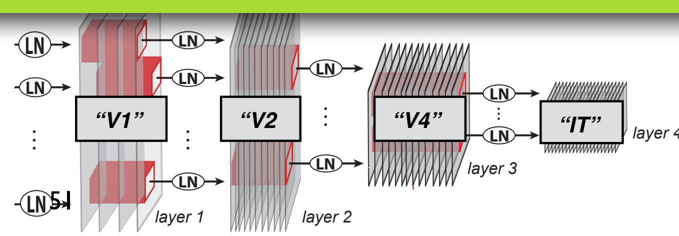
Synthesis & Creation

Current leading mechanistic models of the ventral visual stream…



But, no ANN model aces all of our brain and behavioral tests.

Particular ANNs can now reasonably accurately **<u>explain</u>** / **<u>predict</u>** the workings of the ventral visual stream (single-neuron-level & behavioral-level)



Slide

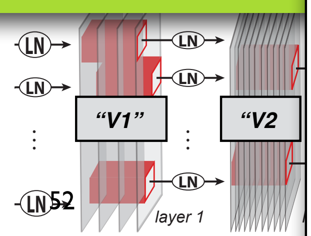
Current leading mechanistic models of the ventral visual stream…

Brain-Score

Summary: clear progress, but the current scientific hypotheses (models) are still incomplete (i.e. demonstrably inaccurate in important ways)

model aces all of our brain and behavioral tests.

Particular ANNs can now **explain** / **predict** the wo
visual stream (single-neuron



Rank ▼	Model	average	V1: FreemanZi	V2: FreemanZi	V4: Majaj2015	IT: Majaj2015	IT-temporal: Ka	behavior: Raja	ImageNet
1	CORnet-S *Kubilius et al., 2018*	.539	.429	.430	.781	.730	.316	.545	.747
2	resnet-50_v2 *He et al., 2015*	.509	.478	.532	.780	.710	X	.553	.756
2	resnet-101_v1 *He et al., 2015*	.509	.455	.524	.775	.738	X	.561	.764
4	densenet-169 *Huang et al., 2016*	.508	.445	.537	.786	.736	X	.543	.759
4	resnet-101_v2 *He et al., 2015*	.508	.466	.527	.784	.713	X	.555	.774
6	resnet-152_v1 *He et al., 2015*	.507	.459	.528	.779	.740	X	.533	.768
6	densenet-201 *Huang et al., 2016*	.507	.454	.533	.777	.738	X	.537	.772
8	resnet-50_v1 *He et al., 2015*	.506	.456	.528	.782	.747	X	.526	.752

Current leading mechanistic models of the ventral visual stream…

Brain-Score

Summary: clear progress, but the current scientific hypotheses (models) are still incomplete **(i.e. demonstrably inaccurate in important ways)**

What can brain scientists do with the current best mechanistic hypotheses?

How should we improve these mechanistic hypotheses?

Pouya Bashivan

Kohitij Kar

RESEARCH ARTICLE SUMMARY

SCIENCE

NEUROSCIENCE

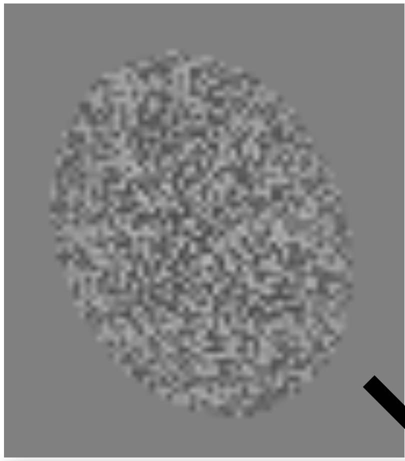
Neural population control via deep image synthesis

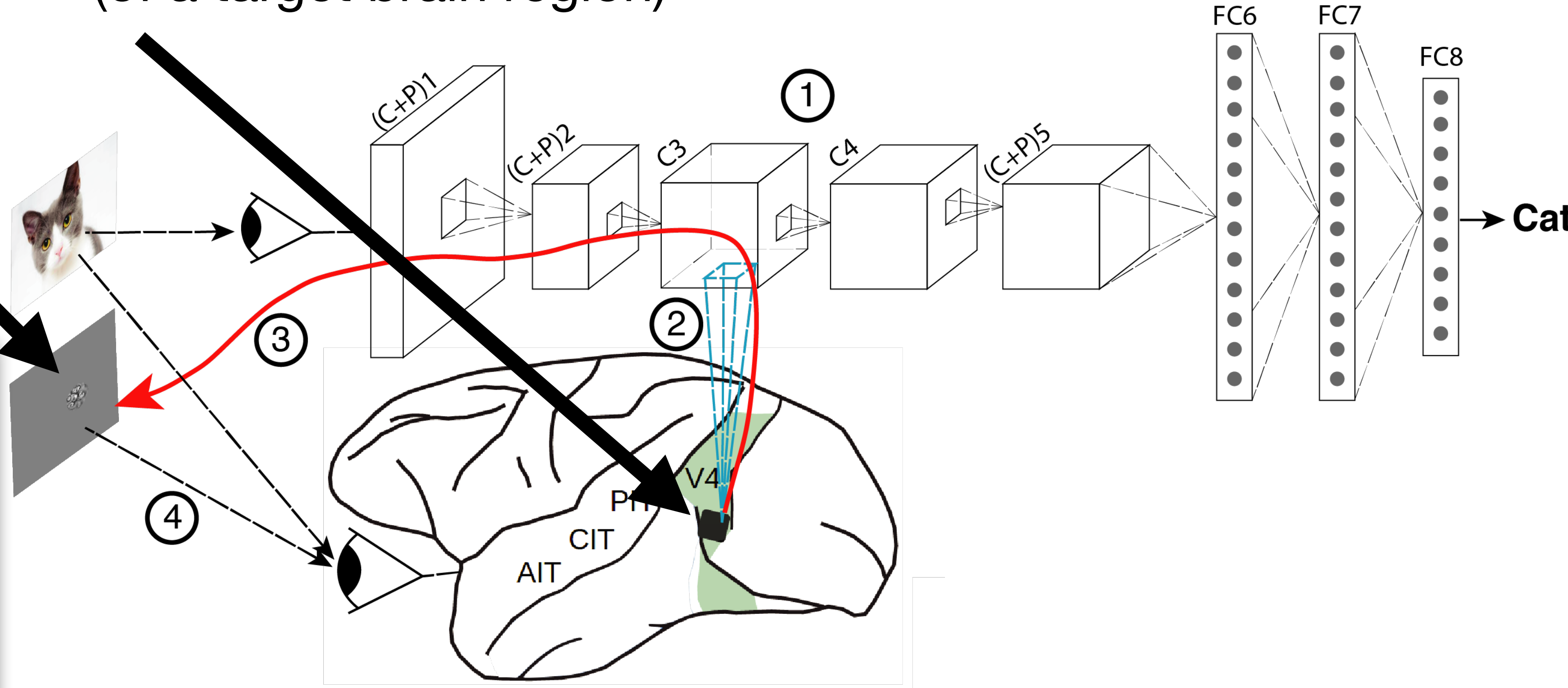
Pouya Bashivan*, Kohitij Kar, James J. DiCarlo

What can brain scientists <u>do</u> with the current best mechanistic hypotheses?

Bashivan, Kar and DiCarlo ***CCN*** *(2018),* ***Science*** *364 (2019)*

A control goal: set a desired neural state
(of a target brain region)





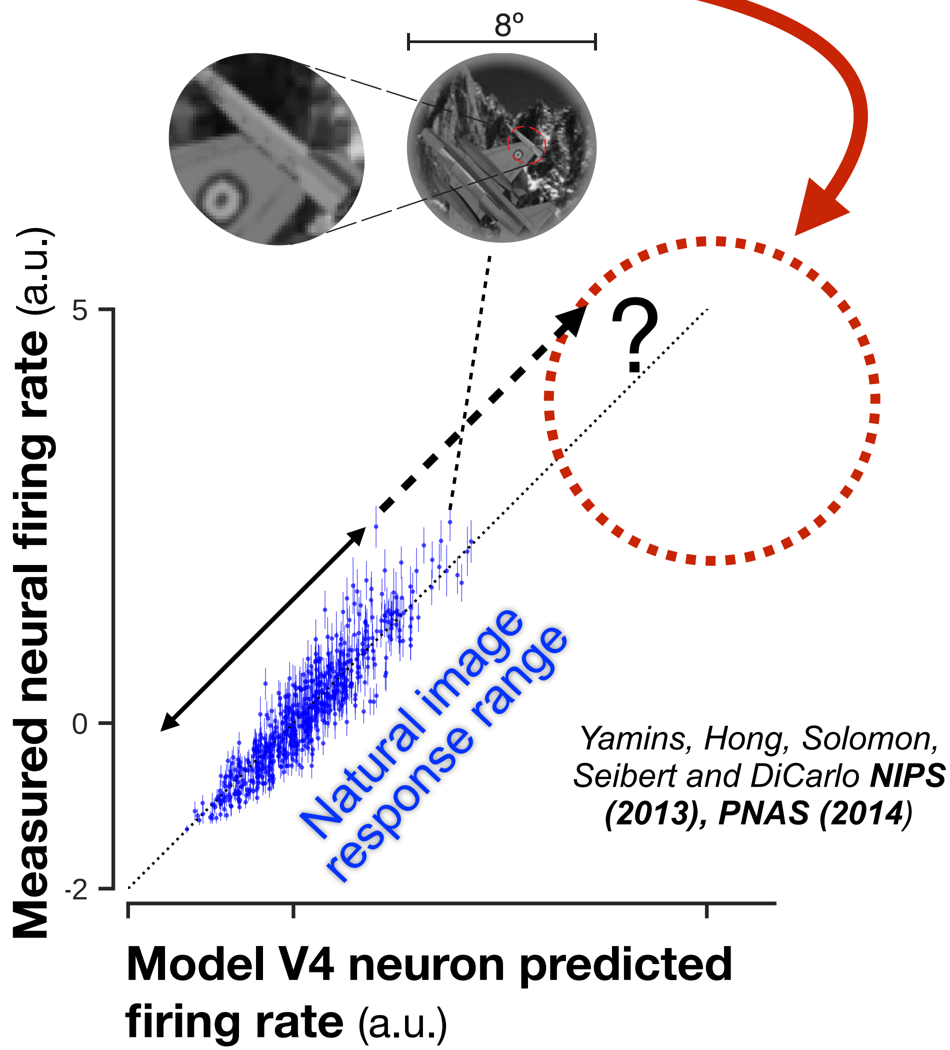
We used our in silico ventral stream models to design precise patterns of light energy on the eyes to try to "set" different internal states of the brain.

We found that we could use these model-designed images to successfully set neural activity states deep in the brain !

Bashivan, Kar and DiCarlo ***CCN*** *(2018),* ***Science*** *364 (2019)*

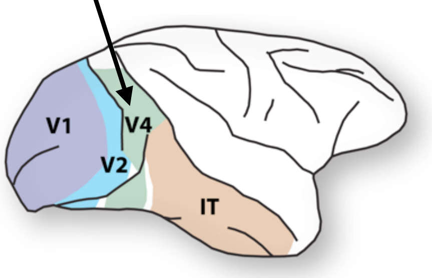
Control goal 1: "super-activate"

Drive any single neural site's activity beyond the maximum response observed thus far.



Responses of an example V4 neural site:

(a mid-level visual area)



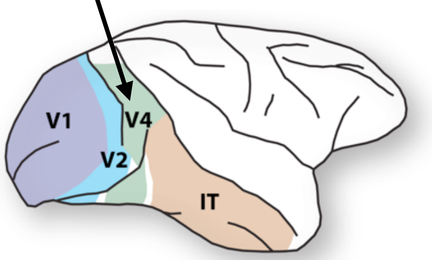
*Bashivan, Kar and DiCarlo **CCN** (2018), **Science** 364 (2019)*

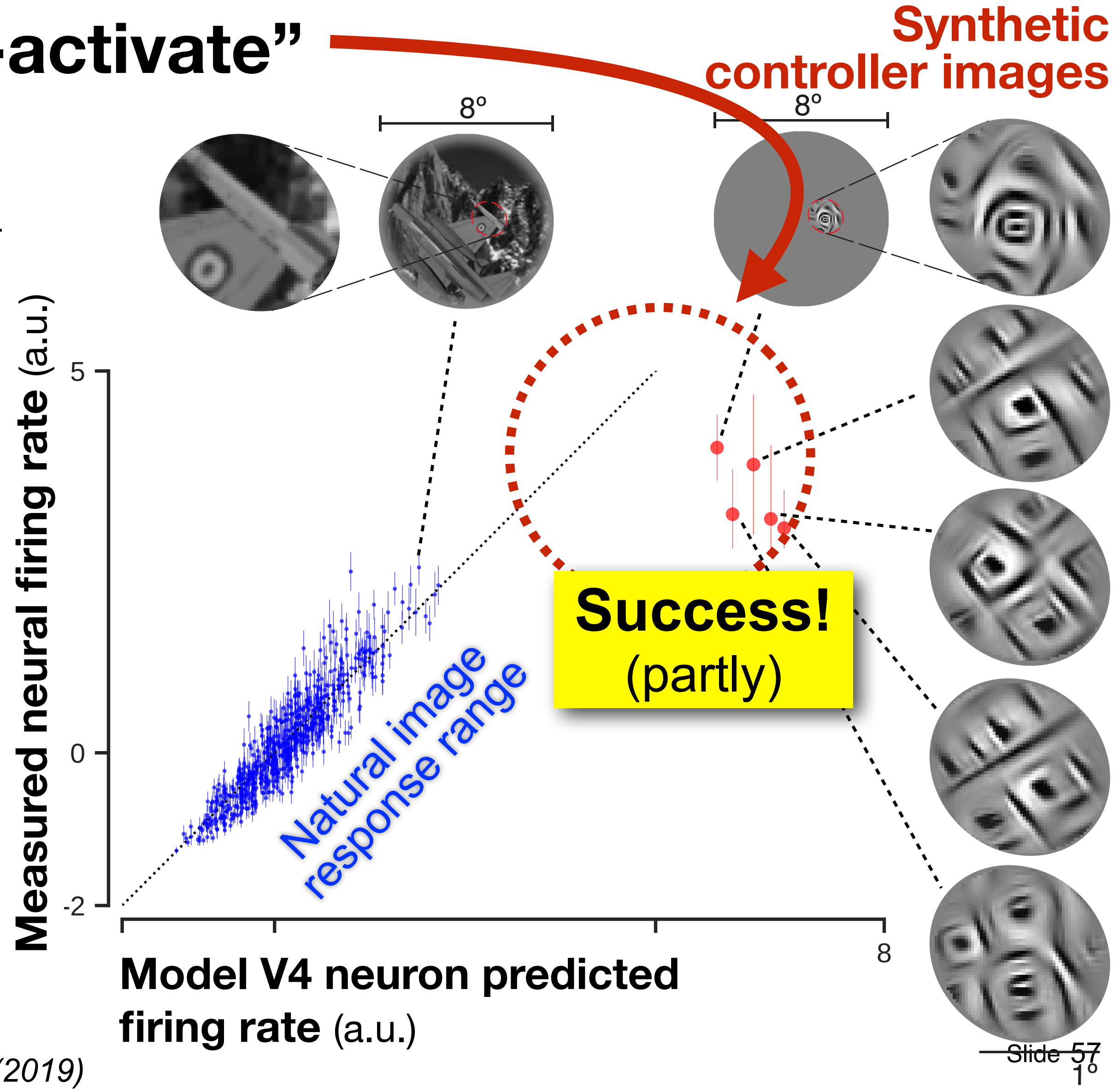
Control goal 1: "super-activate"

Drive any single neural site's activity beyond the maximum response observed thus far.

Responses of an example V4 neural site:

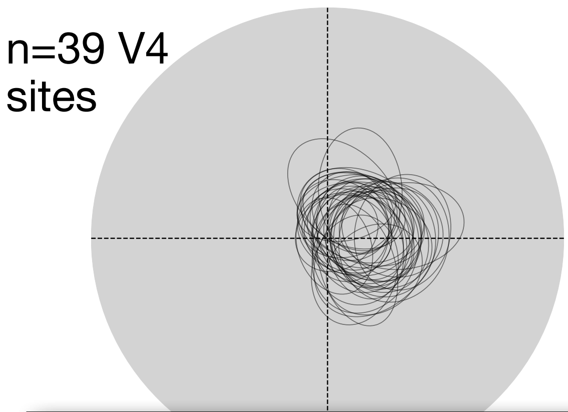
(a mid-level visual area)





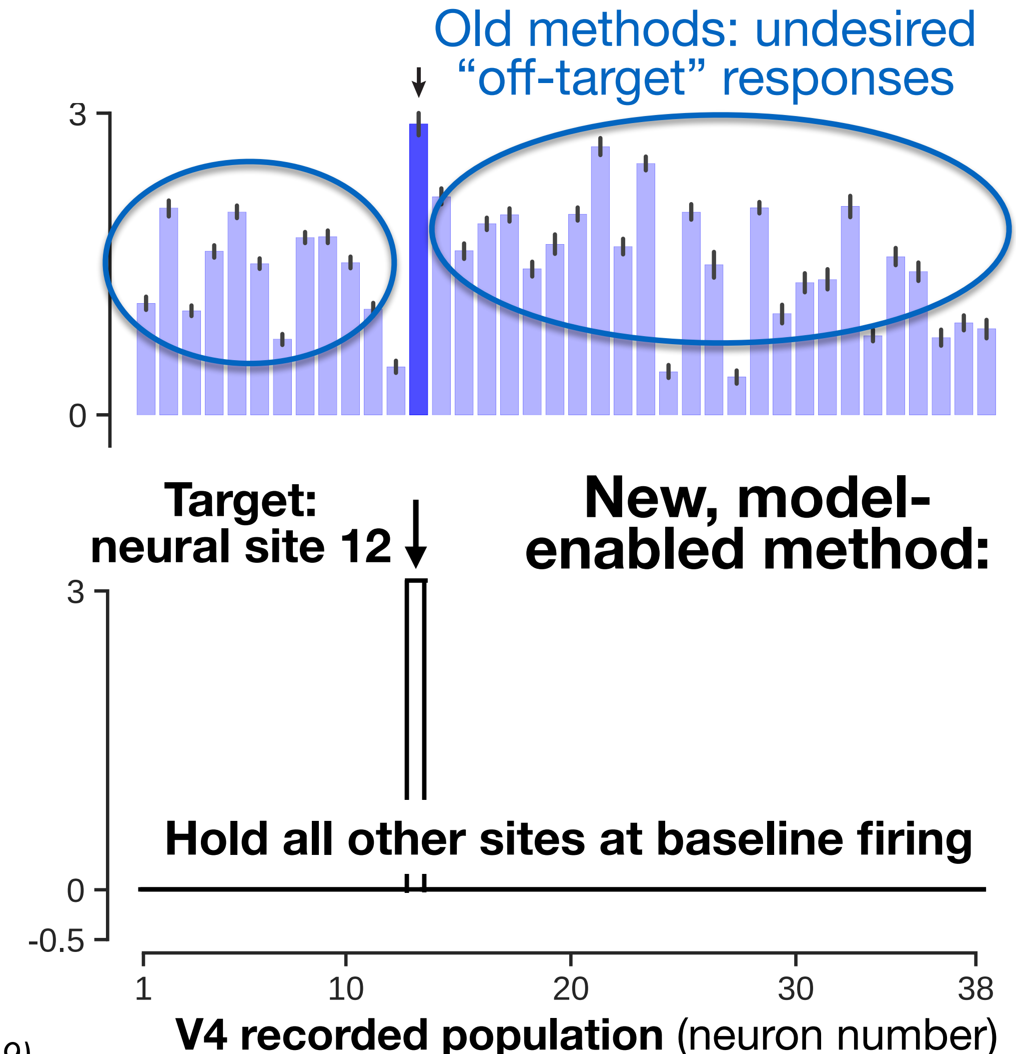
Bashivan, Kar and DiCarlo ***CCN*** *(2018),* ***Science*** *364 (2019)*

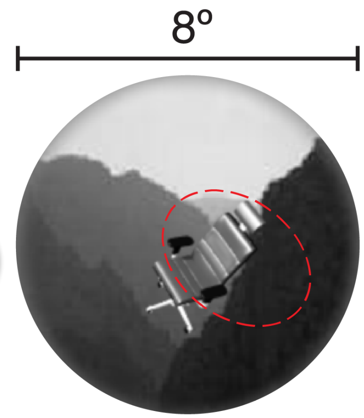
Application: control the neural population state deep in the brain

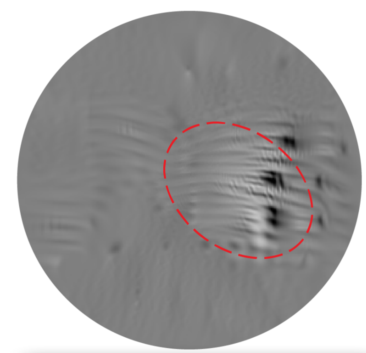


All respond to stimuli in the same part of the visual field

Example goal:
drive one target neural site to high activity, while driving all other neural sites low.







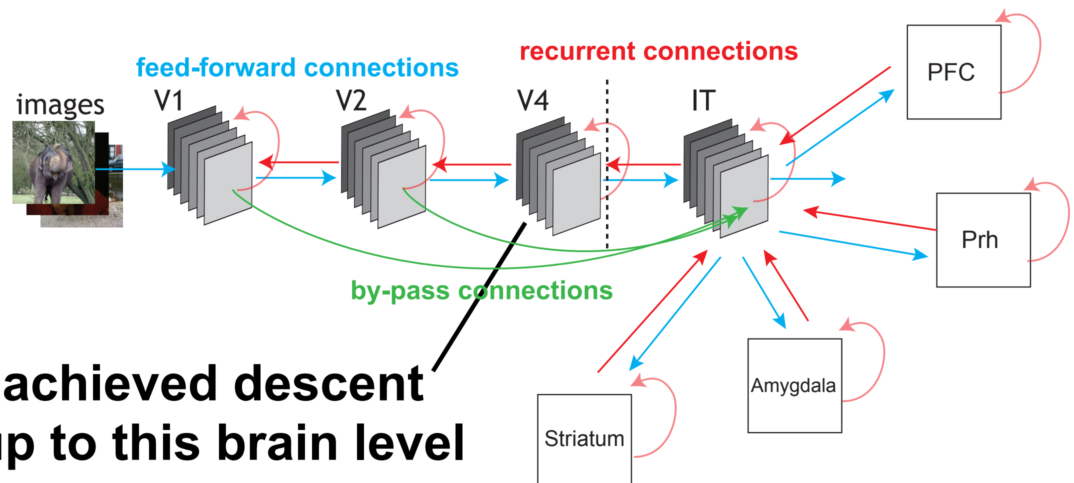
synthetic "controller image"

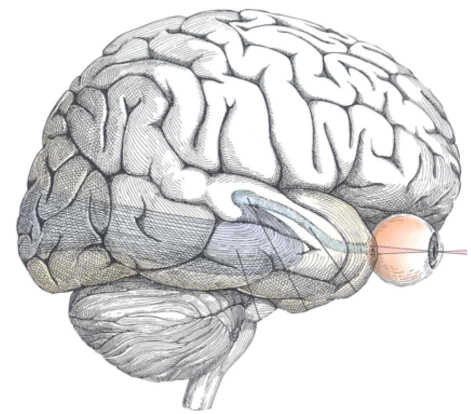
*Bashivan, Kar and DiCarlo **Science** 364 (2019)*

A new application superpower for scientists?:

The ability to control patterns of neural activity deep in the brain by precisely designing the visual input (images, movies)

Images, movies, or video games





So far, we have achieved descent neural control up to this brain level

*Bashivan, Kar and DiCarlo **Science** 364 (2019)*

As we further improve our models, this superpower will further improve!

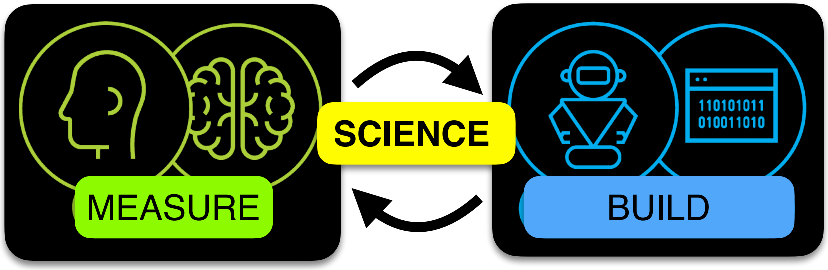
Current leading mechanistic models of the ventral visual stream…

Brain-Score

Summary: clear progress, but the current scientific hypotheses (models) are still incomplete (i.e. demonstrably inaccurate in important ways)

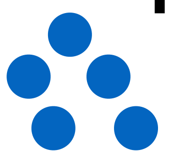
What can brain scientists do with the current best mechanistic hypotheses?

How should we <u>improve</u> these mechanistic hypotheses?



Test/break current hypotheses

Use those breaks to build better hypotheses



Current best ventral stream models

Two major lines of ongoing experimentally-driven work:

1. **Study & improve the early components of these ventral stream hypotheses**

White box adversarial attack

ANN: "Terrier" + Perturbation ⟶ ANN: "Church"

in

Amygdala

Striatum

How shoul

mechani

Dapello, Marques*, Schrimpf, Geiger, Cox & DiCarlo **NeurIPS** (2020)*

Summary take home messages

1. **Background: The ventral visual stream produces an IT neural population representation that carries linearly decodable, image generalizable solutions for all (tested) core object recognition tasks.**

2. **Optimizing deep artificial neural network (ANN) architectures for core recognition tasks leads to internal neural representations in those ANNs that are remarkably similar to the internal neural representations of the ventral visual stream.** (*NIPS* 2013, *PNAS* 2014)

   - **This result (above) includes IT "face neurons".**
   - **This result (above) is consistent with, but does <u>not</u> imply, that the brain learns by classical backpropagation.**

3. **These same (optimized) ANN models can be used to guide the construction of novel synthetic images to super-activate ventral stream neurons and control sub-populations of neurons** (*CCN*, 2018; *Science*, 2019)

Summary take home messages

4. **Nevertheless, these same (optimized) ANNs are <u>not</u> yet functionally <u>identical</u> to the ventral visual stream.** (*e.g. J Neuroscience*, 2018; “Brain-Score” *bioRxiv* 2018)

5. **One difference is the lack of recurrent circuits, and recent IT neurophysiology suggests that fast-acting, automatically-evoked recurrent circuits enable the ventral stream’s superior performance on many images** (*Nature Neuro.*, 2019).

6. **We and our collaborators are building a series of new models that incorporate more biological constraints — thus far, these model show computer vision gains in efficiency (depth) and gains in robustness to image perturbation.**

dicarlo@mit.edu