The Algonauts Project:

Tutorial Day 1

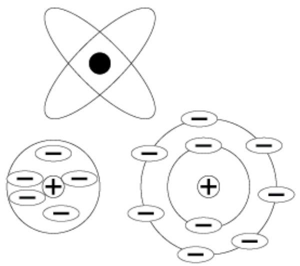
Comparing Brains and DNNs: Theory of Science

Radoslaw Martin Cichy

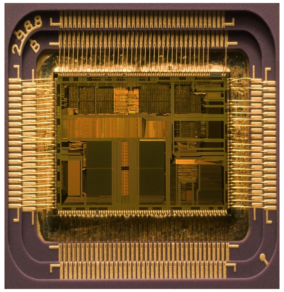
Heated debate

	Critique	Endorsement
Overall potential	Limitations; divergence what a DNN and humans can do; different approach needed	Unprecedented opportunity, new convergence of cognitive science & AI; new framework
Explanation	DNNs may predict, but do not explain phenomena	Explanations of different kinds than usual; post-hoc explanations
Interpretation	DNNs are black boxes – opaque how each part contributes	Concede opaqueness; but in-silico experimentation
Biological realism	While inspired by the brain, in infinite ways DNN differ	Abstraction & idealization essential for modelling; today‘s DNNs starting point for increasing realism
Scientific validity	Current use of DNNs is unscientific because untheoretical	The origin of a model is irrelevant, other factors (e.g. predictive or explanatory power) cound

A bird's eye view from philosophy of science



Model nature
Plurality, diversity & origin



Prediction
Akin to technology: tool and benchmark

$$
\begin{aligned}
\frac{\partial z_k}{\partial w_{ij}} &= \frac{\partial z_k}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} \\
&= \frac{\partial}{\partial a_j} a_j w_{jk} \frac{\partial a_j}{\partial w_{ij}} \\
&= w_{jk}\frac{\partial a_j}{\partial w_{ij}} \\
&= w_{jk}\frac{\partial g_j(z_j)}{\partial w_{ij}}
\end{aligned}
$$

Explanation
Akin to theory: kinds of explanation

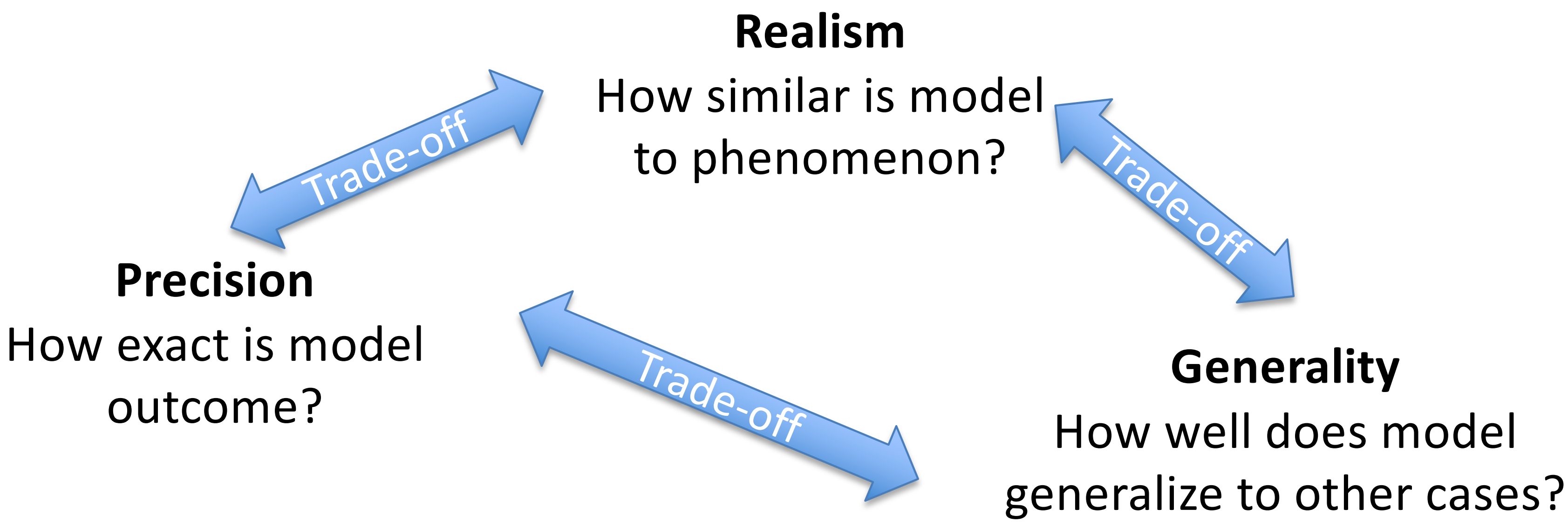
$\left.\begin{array}{l} \\ \end{array}\right\}$ The two major goals of science

Exploration
Starting point for new theories

$\left.\begin{array}{l} \\ \end{array}\right\}$ Overlooked, yet fundamental & ubiquitous

Claim 1: We need many models; theoretical desiderata

Theoretical desiderata = what we want a model to be for theoretical reasons

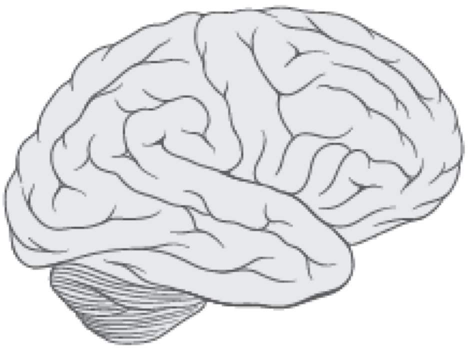


If target class is inhomogenous, no model fulfills all desiderata
Cognitive phenomena are inhomogenous (evolution/experience).

$\Rightarrow$ There is no one perfect model. We need many models.

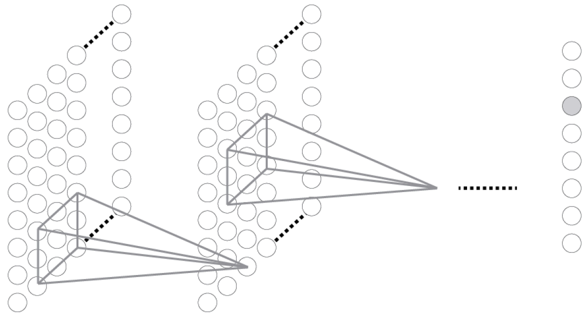
Claim 1: We need many models; non-theoretical desiderata

Non-theoretical desiderata = what we want a model to be for practical reasons



A perfect brain model that is incredibly slow to evaluate, hard to manipulate, ethically restricted

Trade - off



An inexact model that is very fast, easy to manipulate, and ethically unproblematic

⇒ Non-theoretical desiderata often take precedence

⇒ DNNs appear attractive on many non-theoretical desiderata

Claim 2: Best models are diverse

Question:

Given many models for many desiderata – will they all be of the same kind (e.g. all DNNs) or all different?

Plausibility argument:

In any branch of science…
… at any degree of maturity…
… there are models of different kinds.

$\Rightarrow$ DNNs have a place in the diverse set of models in cognitive science

Claim 3: The origin of models is irrelevant

Challenge:

Scientific models are derived from theory to instantiate or test it

⇒ DNNs are not derived from theory, so they are not proper models

Reality check from scientific practise:

- Rarely deduced straight-forwardly from theory
- More art than logic
- No predefined set of rules
- Process involves creativity, chance and transfer
- Again: non-theoretical desiderata relevant

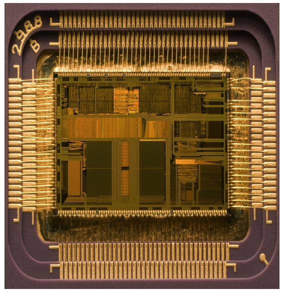
⇒ **Origin of a model is irrelevant**

⇒ **DNN being hijacked by cognitive science akin to ready-mades is OK**

(Duchamp 1917)

A bird's eye view from philosophy of science

Model nature
Plurality, diversity & origin



Prediction
Akin to technology: tool and benchmark

Explanation
Akin to theory: kinds of explanation

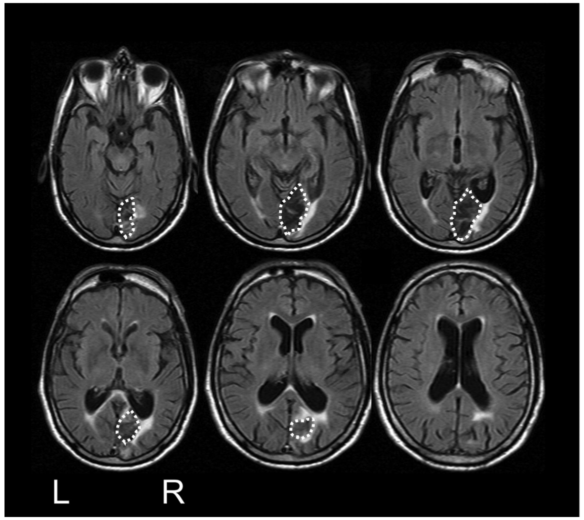
Exploration
Starting point for new theories

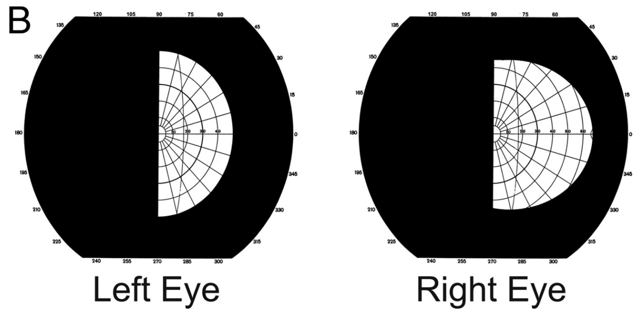
Claim 1: Use DNNs as a tool for practical aim

Without recurrence to explanation

Examples

- Medical application
  => neural prothesis



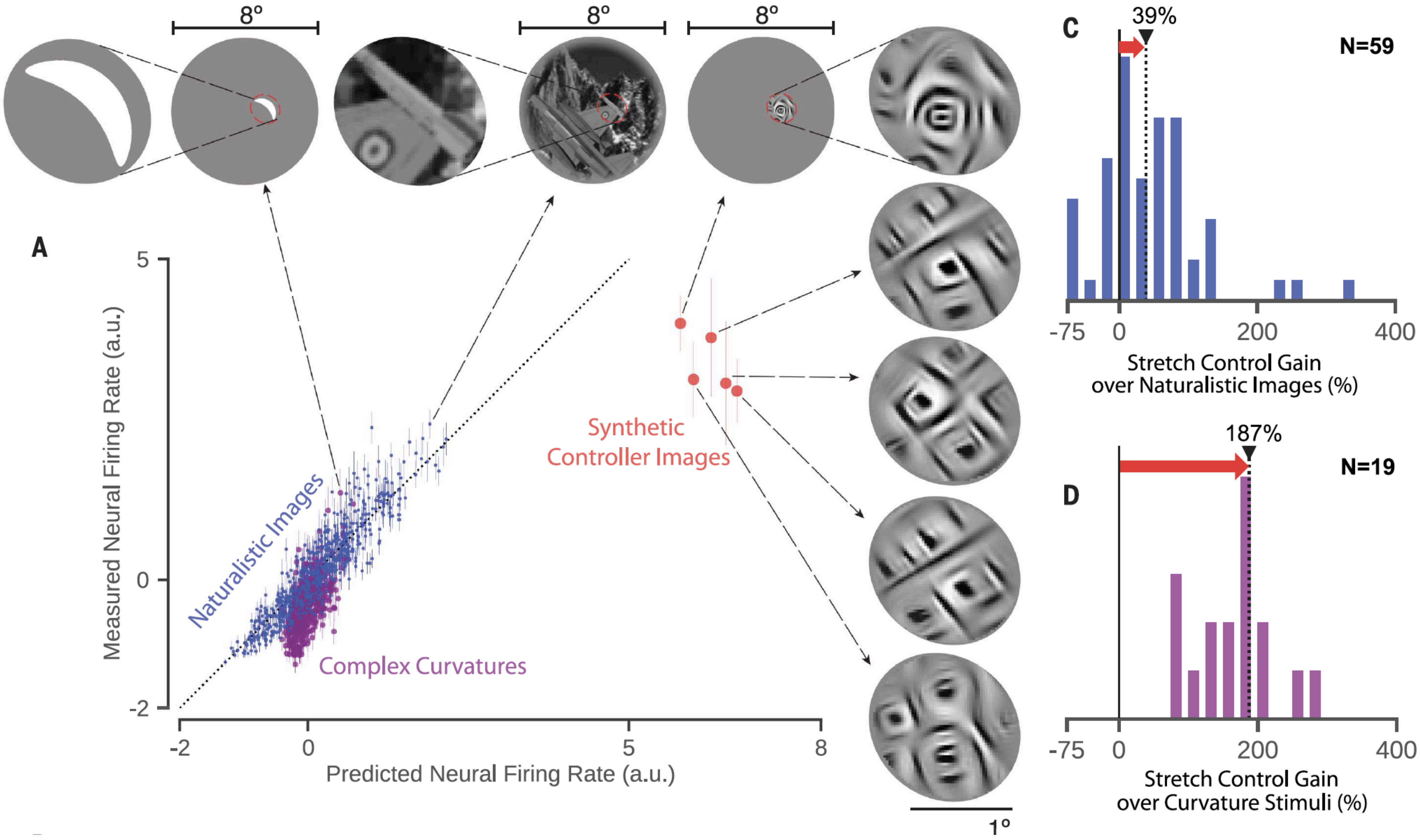


Striemer et al., 2009

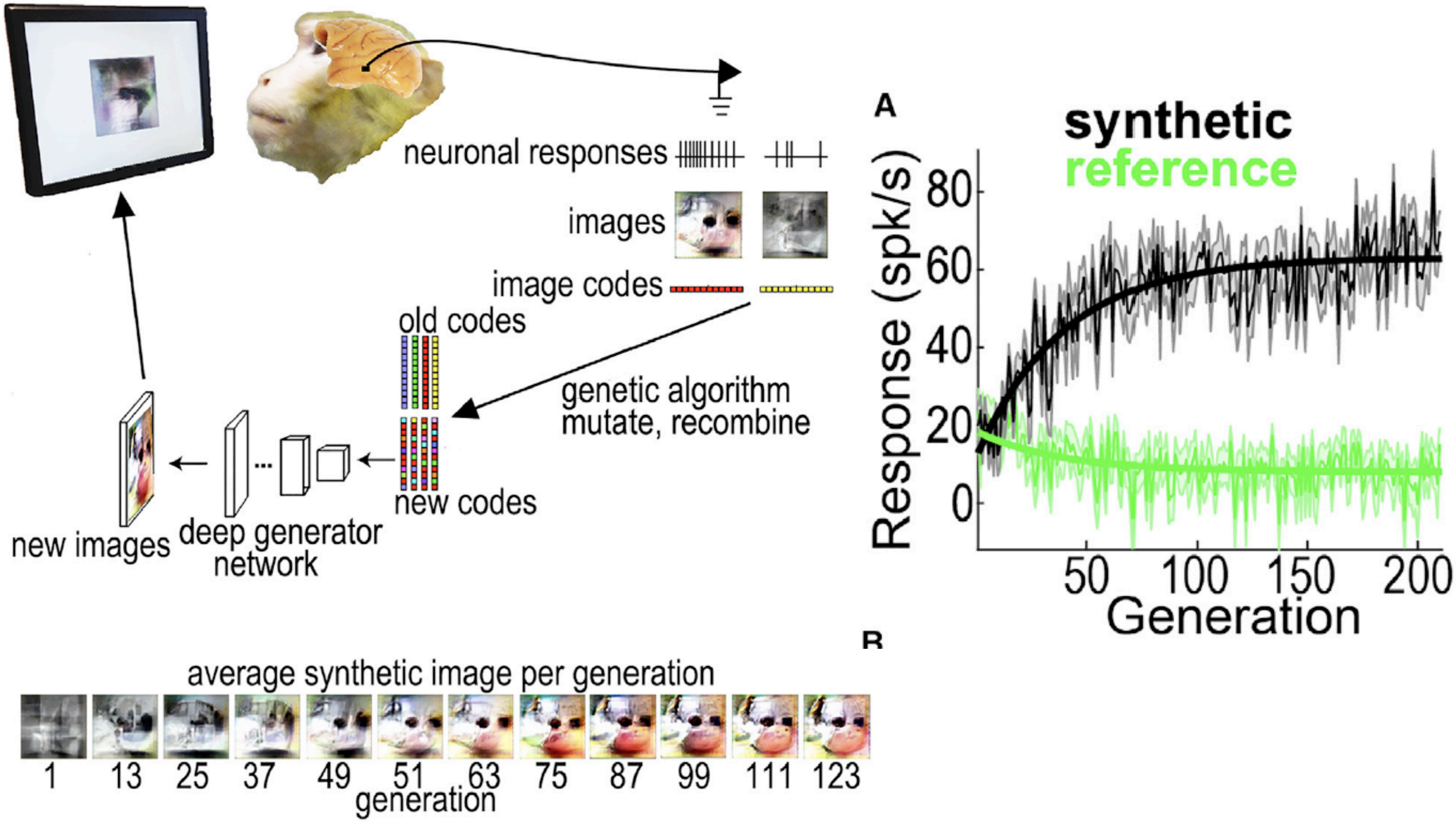
- Experimental design optimization => experimental control

Example: Neural population control via deep image synthesis

Pouya Bashivan*, Kohitij Kar*, James J. DiCarlo†



Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences



Claim 2: Benchmarking as stepping stone for explanation

		Score
Rank	Team Name	Average Noise Normalized R^2 (%)
	Noise Ceiling	*100*
1	agustin	26.91
2	Aakash	24.89
3	rmldj	24.56
…	…	…
24	AlexNet-OrganizerBaseline	7.41

⇒ Pre-select models by performance for further inquiry

⇒ Comparison of models can reveal factors relevant for success

⇒ Good prediction baseline for explanation of complex functions

A bird’s eye view from philosophy of science

Model nature
Plurality, diversity & origin

Prediction
Akin to technology: tool and benchmark

$$\begin{aligned}\frac{\partial z_k}{\partial w_{ij}} &= \frac{\partial z_k}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} \\ &= \frac{\partial}{\partial a_j} a_j w_{jk} \frac{\partial a_j}{\partial w_{ij}} \\ &= w_{jk}\frac{\partial a_j}{\partial w_{ij}} \\ &= w_{jk}\frac{\partial g_j(z_j)}{\partial w_{ij}}\end{aligned}$$

Explanation
Akin to theory: kinds of explanation

Exploration
Starting point for new theories

Exploratory power of DNNs – the challenge

The received view: mathematical-theoretical modelling

- Identify a few relevant variables
- Each variable identified a priori with part of phenomenon modelled
- Use math to model variables & their interaction

$\Rightarrow$ Changes in model variable directly interpretable as changes in the world

DNNs

- ~ millions of parameters
- Parameters learned rather than set a priori
- Relationship of variables to the world is opaque

$\Rightarrow$ DNNs are a black box. One cannot explain one black box (e.g. brain) by another one (DNN). Thus DNNs lack explanatory power.

Claim 1: DNNs provide teleological explanations

Teleological: From Greek telos (end, goal, purpose), related to a goal, aim or purpose

DNN

Question

Why does a unit behave such and such?

Answer

Because it fulfill its function in enabling a particular objective

Rather than

Because it represents this or that feature of the world

Brain

Analogous exchanging “unit” for “neuron”

Claim 2: Appearance nonwithstanding DNNs offer standard vanilla explanations

DNNs defined by handful of parameters set a priori, e.g.

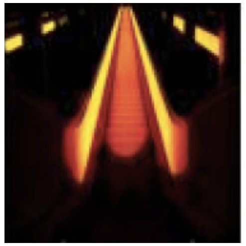
- architecture
- training material
- training procedure
- objective

Variables directly refer to phenomena in the world.

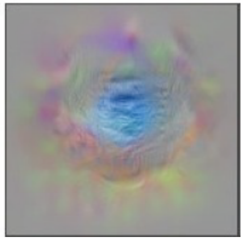
$\Rightarrow$ The model is thus transparent, and not a black box.

Claim 3: Strong potential for post-hoc explanations

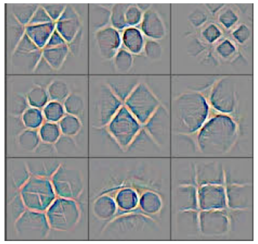
Idea: Making DNNs transparent will enable explanatory power



Zhou et al., 2015



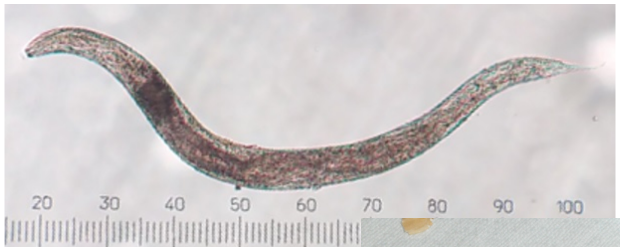
Yosinski et al., 2015



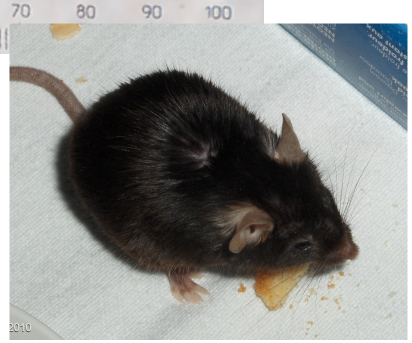
Zeiler & Fergus 2013

Zhou et al., 2018

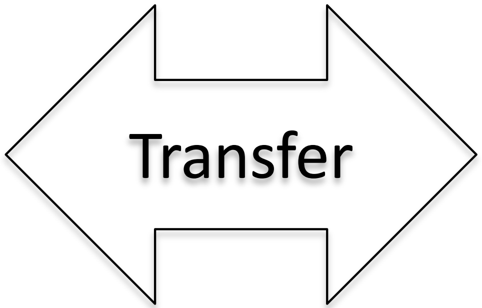
Analogy: model organisms in biology

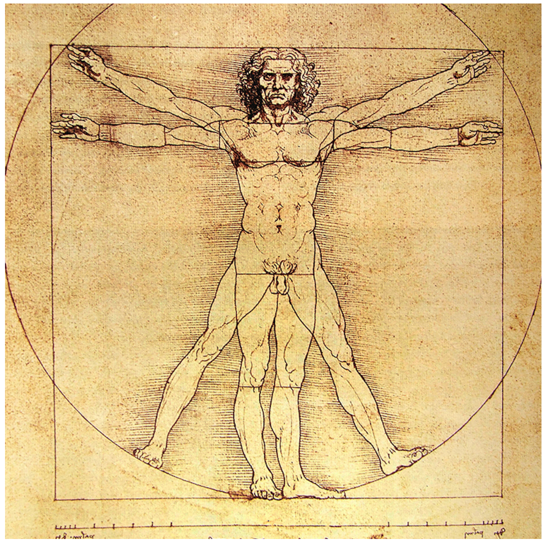


C. elegans



Mus musculus





Homo sapiens

A bird's eye view from philosophy of science

Model nature
Plurality, diversity & origin

Prediction
Akin to technology: tool and benchmark

Explanation
Akin to theory: kinds of explanation

Exploration
Starting point for new theories

Exploration: DNNs as starting point for new theories

With a fully-fledged theory, deriving hypotheses and testing them in experiments is the rule.

But what do you do when there is no fully-fledged theory?

$\Rightarrow$ **Exploration**

Claim 1: Exploration generates new hypotheses

Analogies (Mary Hesse)

Positive: characteristics we know model and target **do** share

Negative: characteristics we know model and target ***do not share***

Neutral: characteristics of which we ***do not know*** whether they are shared

Brain – DNN example

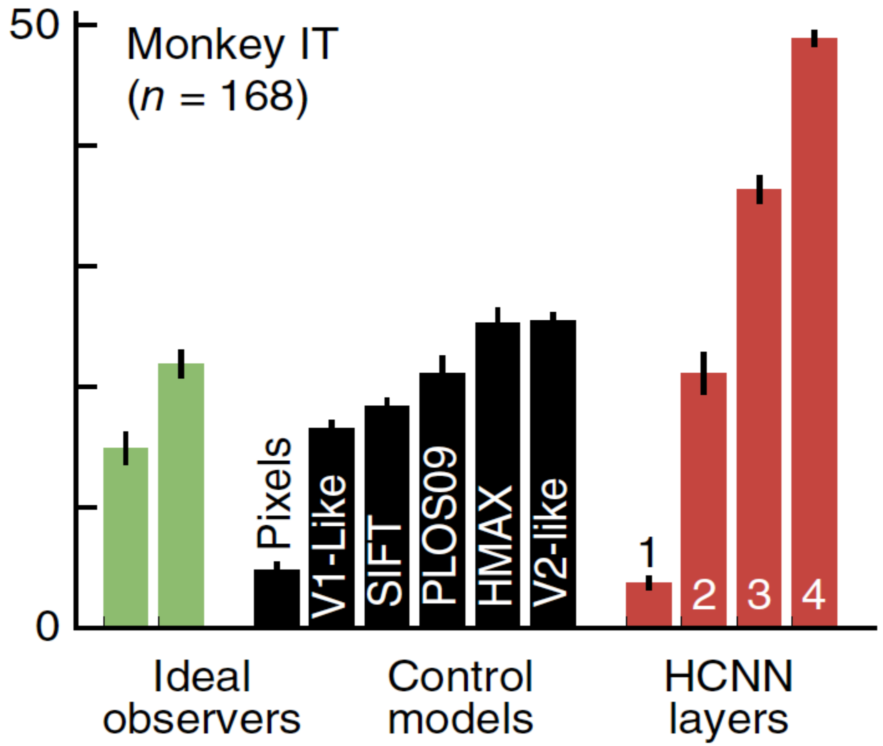
Brains and DNNs have simple discrete entities (neurons/ units) as computational building blocks

Brains are made of sugars, lipids, proteins and water, DNNs not

Potential for learning new facts about the target

Claim 2: DNNs offer proof-of-principle demonstrations

Proof-of-principle demonstration
Demonstration that it works in theory by showing that it works in practise

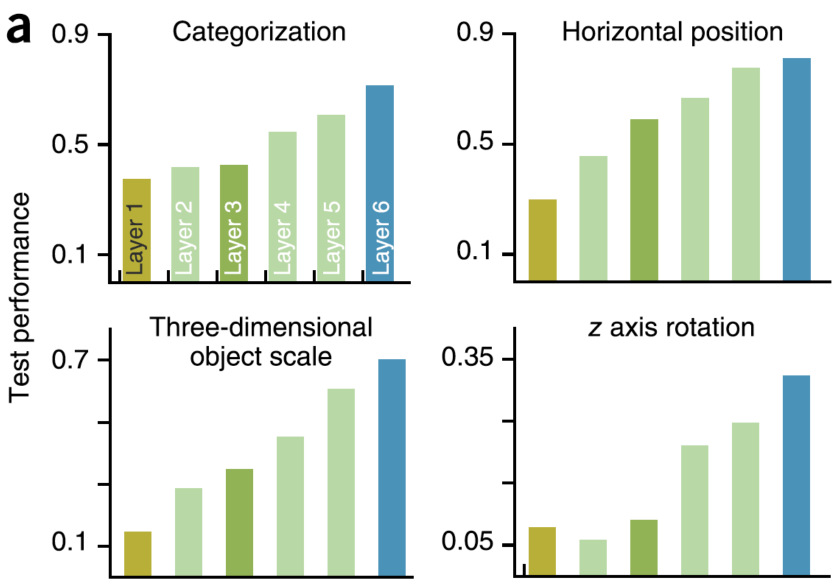


Example
A purely feed-forward DNN predicts neural activity in IT well.

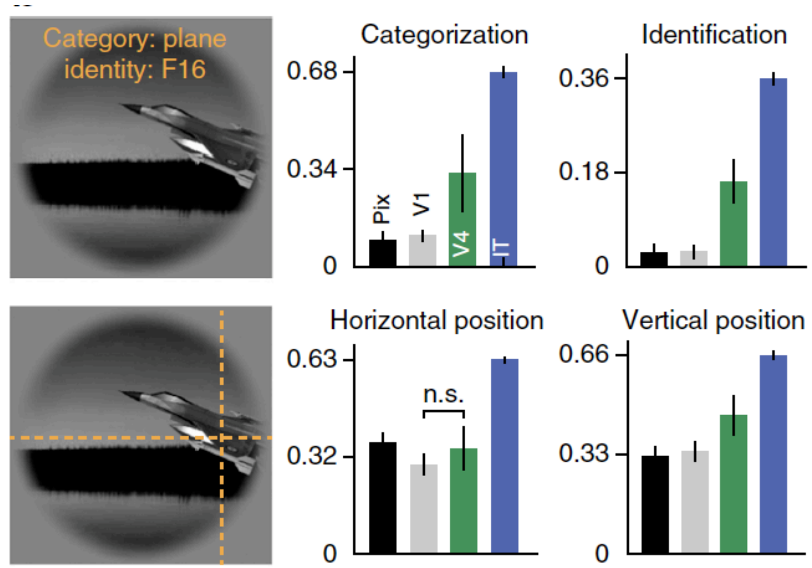
Upshot
⇒Feasibility invites further investigation of feed-forward solutions

Claim 3: Assessment of the suitability of the target

Example: Category – orthogonal properties (Hong et al., 2016)



DNN



Monkey IT

Caveats and limitations of DNN exploration

1) Standards for judging quality/success are less developed & implicit

$\Rightarrow$ Give DNNs benefit of the doubt to avoid curbing development prematurely

2) Same model: exporative in one context, explanatory in another

$\Rightarrow$ Clearly indicate how the model is used

3**) Danger of mistaking the model for the world**

$\Rightarrow$ Modelling must always be checked by experimentation

Summary

Model nature

Plurality

Trade-offs between desiderata (theoretical and non-theoretical)

Diversity

Co-existence and continuous success of diverse models anywhere in science

Origin

Irrelevant to scientific relevance of a model

Model use

Prediction

DNNs as tools to reach a practical aim

- Neural prothesis
- Experimental design development & optimization

Benchmarking as stepping stone to explanation

- Model selection for further inquiry
- Model comparison

Explanation

DNNs as means to test hypotheses

- Teleological type of explanation
- Mechanistic-theoretical explanation: DNNs defined by few interpretable parameters
- Opaqueness of DNNs an interim stage to be overcome by post-hoc explanations

Exploration

DNNs as starting points for new theories

- Generation of new hypotheses via neutral analogies
- Proof-of-principle demonstrations motivate further inquiry
- Assessment of the suitability of the target

Cichy & Kaiser, TICS 2019