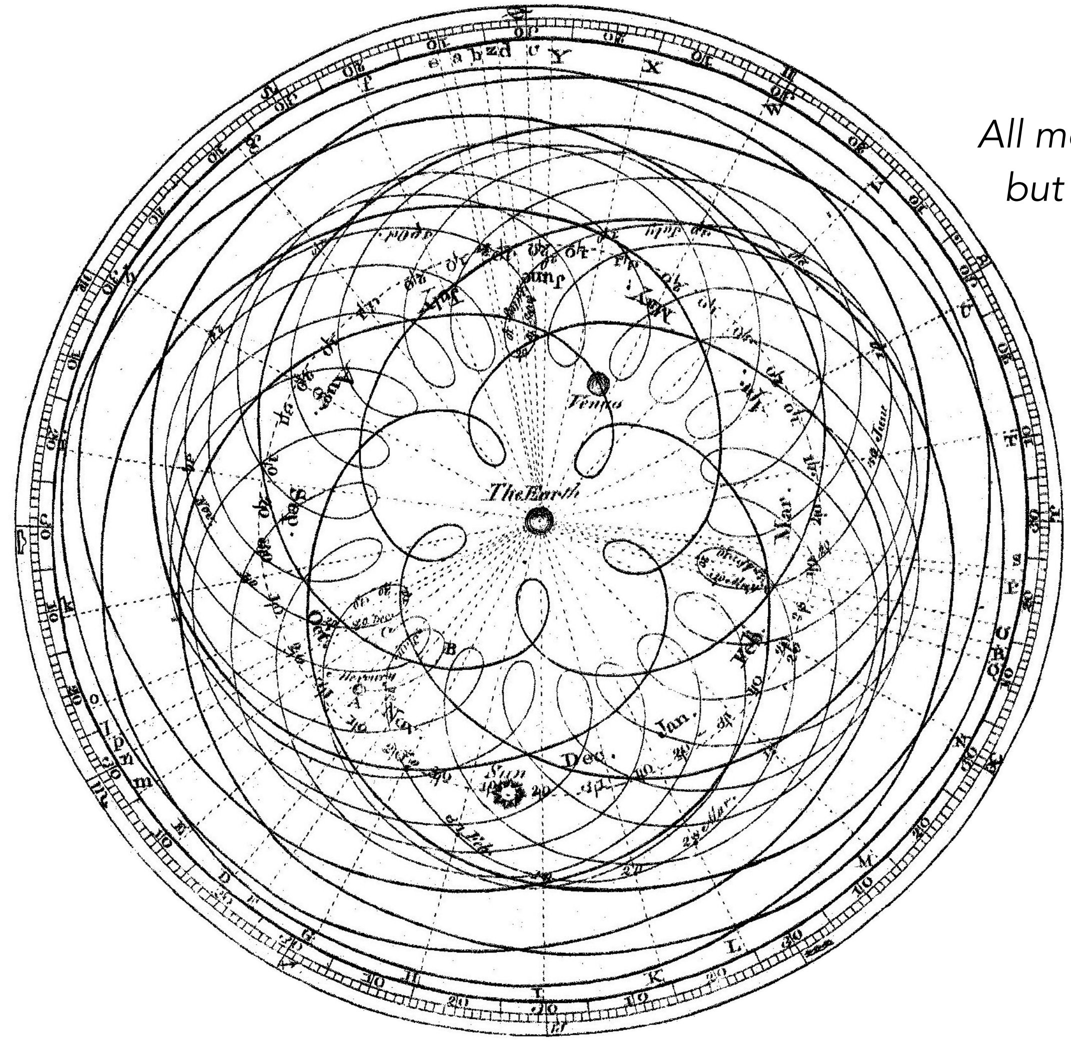


All models are wrong,
but some are useful.

– George E. P. Box

Ilya Kuzovkin

Understanding Information Processing in Human Brain by Interpreting Machine Learning Models

Supervised by Raul Vicente

UNIVERSITY OF TARTU
1632

Modeling is a well-proven way of obtaining knowledge

MANUAL

AUTOMATED

https://towardsdatascience.com/black-boxes-and-their-intrusion-620aa3c4c56b https://cs.hse.ru/en/bayesgroup/education/moml http://ircamera.as.arizona.edu/NatSci102/NatSci102/lectures/kepler.htm
https://www.forbes.com/sites/andrewbusby/2019/03/04/new-pwc-survey-reveals-consumer-data-is-the-most-highly-valued/#6a084d3d640c https://www.helioseducore.com/keplers-laws-of-planetary-motion

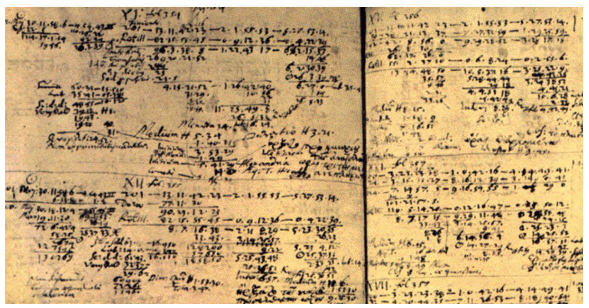
Modeling is a well-proven way of obtaining knowledge

DATA

ALGORITHM

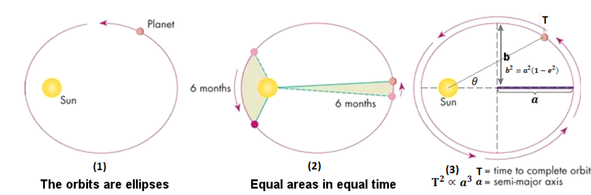
MODEL

MANUAL



Tycho Brahe's observations

Johannes Kepler



Three Laws of Planetary Motion

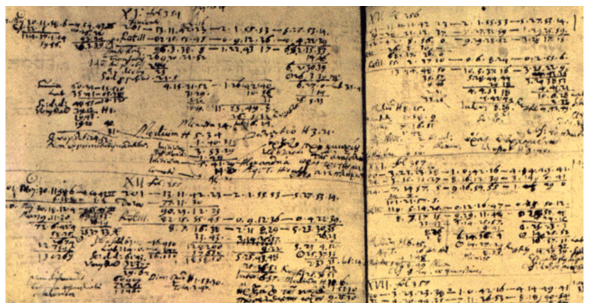
AUTOMATED

https://towardsdatascience.com/black-boxes-and-their-intrusion-620aa3c4c56b https://cs.hse.ru/en/bayesgroup/education/moml http://ircamera.as.arizona.edu/NatSci102/NatSci102/lectures/kepler.htm
https://www.forbes.com/sites/andrewbusby/2019/03/04/new-pwc-survey-reveals-consumer-data-is-the-most-highly-valued/#6a084d3d640c https://www.helioseducore.com/keplers-laws-of-planetary-motion

Modeling is a well-proven way of obtaining knowledge

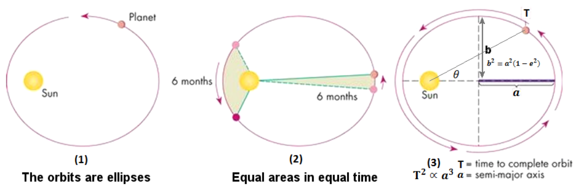
DATA | ALGORITHM | MODEL

MANUAL



Tycho Brahe's observations

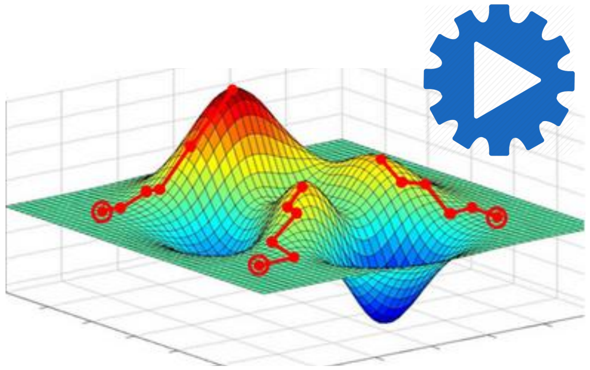
Johannes Kepler



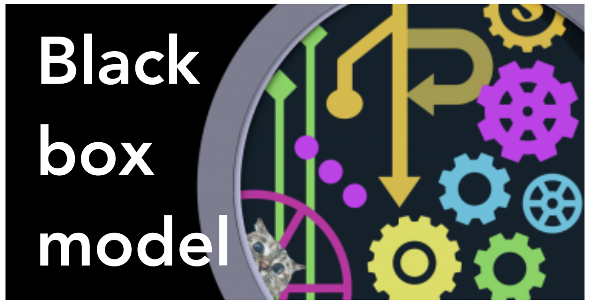
Three Laws of Planetary Motion

AUTOMATED

Some data



A machine learning algorithm



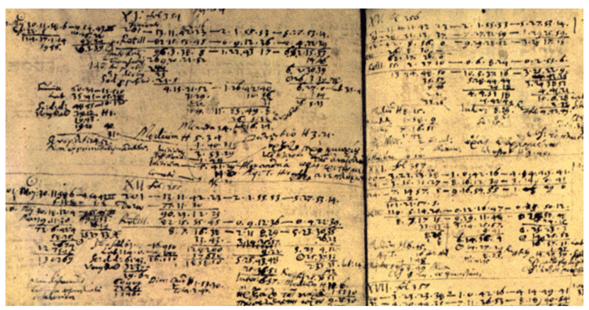
The model

https://towardsdatascience.com/black-boxes-and-their-intrusion-620aa3c4c56b https://cs.hse.ru/en/bayesgroup/education/moml http://ircamera.as.arizona.edu/NatSci102/NatSci102/lectures/kepler.htm
https://www.forbes.com/sites/andrewbusby/2019/03/04/new-pwc-survey-reveals-consumer-data-is-the-most-highly-valued/#6a084d3d640c https://www.helioseducore.com/keplers-laws-of-planetary-motion

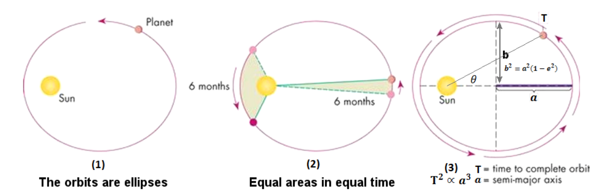
Modeling is a well-proven way of obtaining knowledge

DATA | ALGORITHM | MODEL | KNOWLEDGE

MANUAL


Tycho Brahe's observations

Johannes Kepler

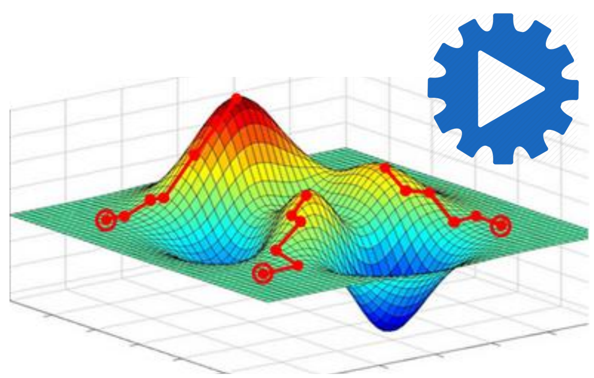


Three Laws of Planetary Motion

=
MODEL

AUTOMATED


Some data


A machine learning algorithm

The model

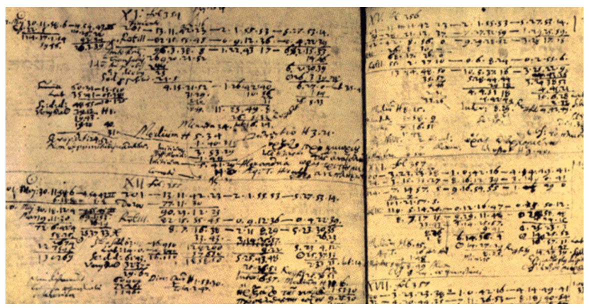
IT'S THERE
BUT
HIDDEN

https://towardsdatascience.com/black-boxes-and-their-intrusion-620aa3c4c56b https://cs.hse.ru/en/bayesgroup/education/moml http://ircamera.as.arizona.edu/NatSci102/NatSci102/lectures/kepler.htm
https://www.forbes.com/sites/andrewbusby/2019/03/04/new-pwc-survey-reveals-consumer-data-is-the-most-highly-valued/#6a084d3d640c https://www.helioseducore.com/keplers-laws-of-planetary-motion

Modeling is a well-proven way of obtaining knowledge

DATA | ALGORITHM | MODEL | KNOWLEDGE

MANUAL

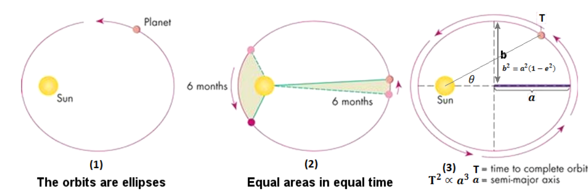


Tycho Brahe's observations

→

Johannes Kepler

→



Three Laws of Planetary Motion

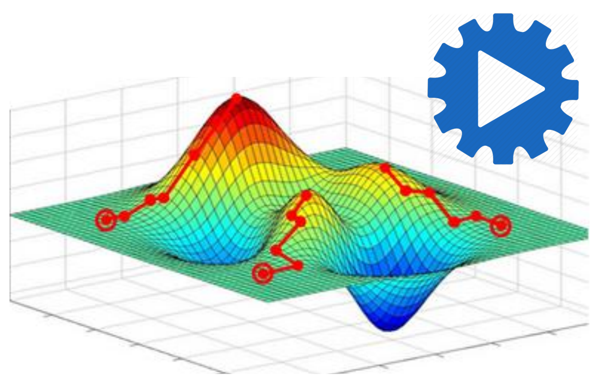
=

MODEL

AUTOMATED

Some data

→



A machine learning algorithm

→

The model

IT'S THERE
BUT
HIDDEN

Machine learning can uncover knowledge

Model interpretation is required to articulate it

https://towardsdatascience.com/black-boxes-and-their-intrusion-620aa3c4c56b https://cs.hse.ru/en/bayesgroup/education/moml http://ircamera.as.arizona.edu/NatSci102/NatSci102/lectures/kepler.htm
https://www.forbes.com/sites/andrewbusby/2019/03/04/new-pwc-survey-reveals-consumer-data-is-the-most-highly-valued/#6a084d3d640c https://www.helioseducore.com/keplers-laws-of-planetary-motion

In life sciences model interpretation has a special significance

https://www.wired.com/story/the-rise-of-dna-data-storage/ https://www3.mpifr-bonn.mpg.de/staff/rbeck/MKSP/pictures.html https://www.youtube.com/watch?v=Evsqy0a_Lrk https://www.sciencemag.org/news/2018/08/scientists-tweak-dna-viable-human-embryos
http://andysbrainblog.blogspot.com/2015/04/slice-analysis-of-fmri-data-with-spm.html https://astronomy.com/news/2019/01/maybe-you-really-can-use-black-holes-to-travel-the-universe https://www.sciencenewsforstudents.org/article/life-earth-mostly-green

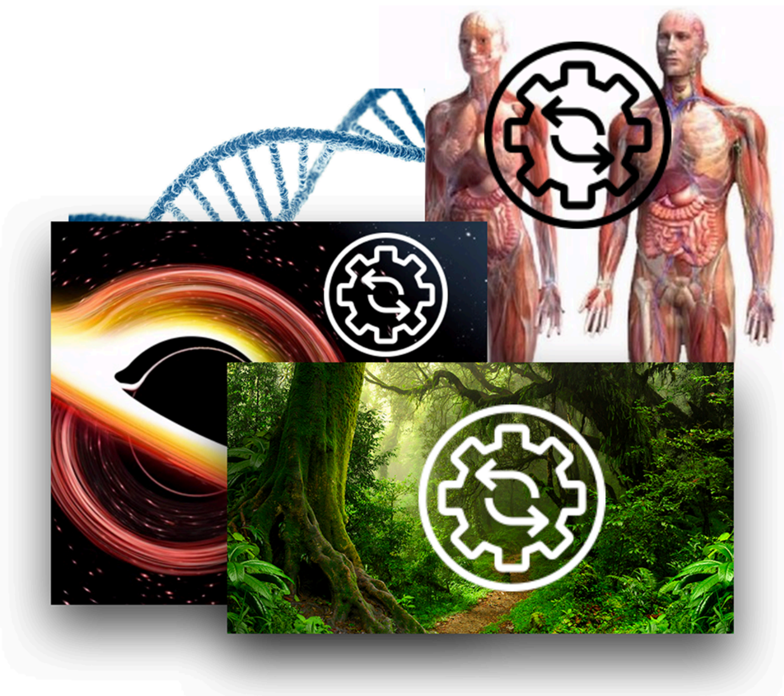
In life sciences model interpretation has a special significance

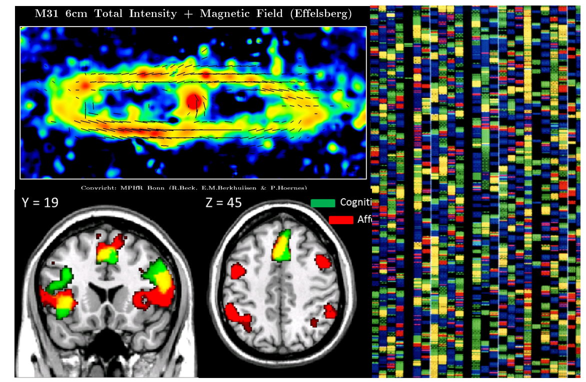
Interpretability in ML

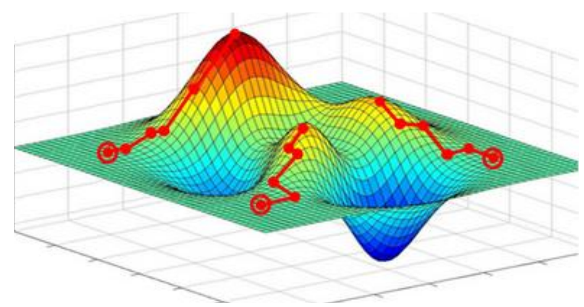
- Trust in model's decision
- Legal transparency
- Debugging

https://www.wired.com/story/the-rise-of-dna-data-storage/ https://www3.mpifr-bonn.mpg.de/staff/rbeck/MKSP/pictures.html https://www.youtube.com/watch?v=Evsqy0a_Lrk https://www.sciencemag.org/news/2018/08/scientists-tweak-dna-viable-human-embryos
http://andysbrainblog.blogspot.com/2015/04/slice-analysis-of-fmri-data-with-spm.html https://astronomy.com/news/2019/01/maybe-you-really-can-use-black-holes-to-travel-the-universe https://www.sciencenewsforstudents.org/article/life-earth-mostly-green

In life sciences model interpretation has a special significance







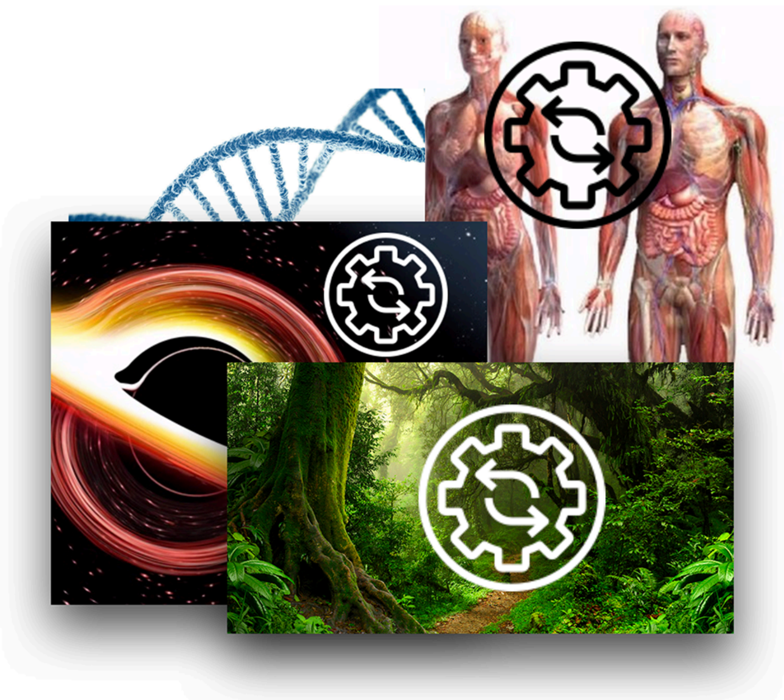
A model that can make correct predictions about reality

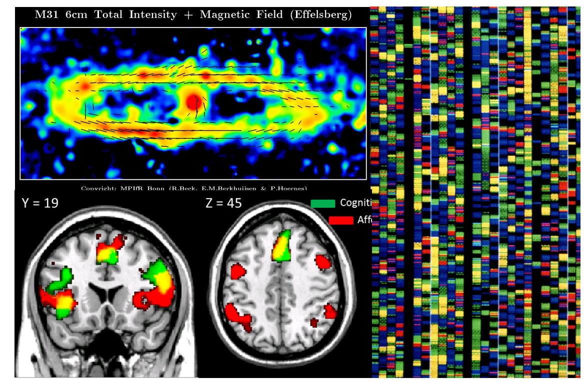
Interpretability in ML

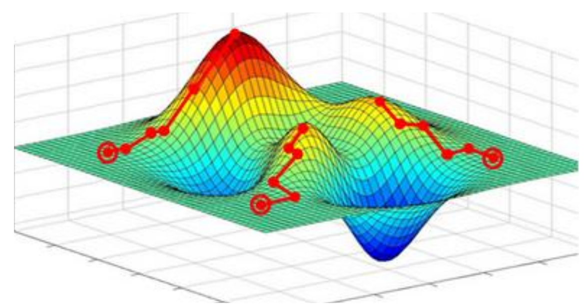
- Trust in model's decision
- Legal transparency
- Debugging

https://www.wired.com/story/the-rise-of-dna-data-storage/ https://www3.mpifr-bonn.mpg.de/staff/rbeck/MKSP/pictures.html https://www.youtube.com/watch?v=Evsqy0a_Lrk https://www.sciencemag.org/news/2018/08/scientists-tweak-dna-viable-human-embryos
http://andysbrainblog.blogspot.com/2015/04/slice-analysis-of-fmri-data-with-spm.html https://astronomy.com/news/2019/01/maybe-you-really-can-use-black-holes-to-travel-the-universe https://www.sciencenewsforstudents.org/article/life-earth-mostly-green

In life sciences model interpretation has a special significance







A model that can make correct predictions about reality

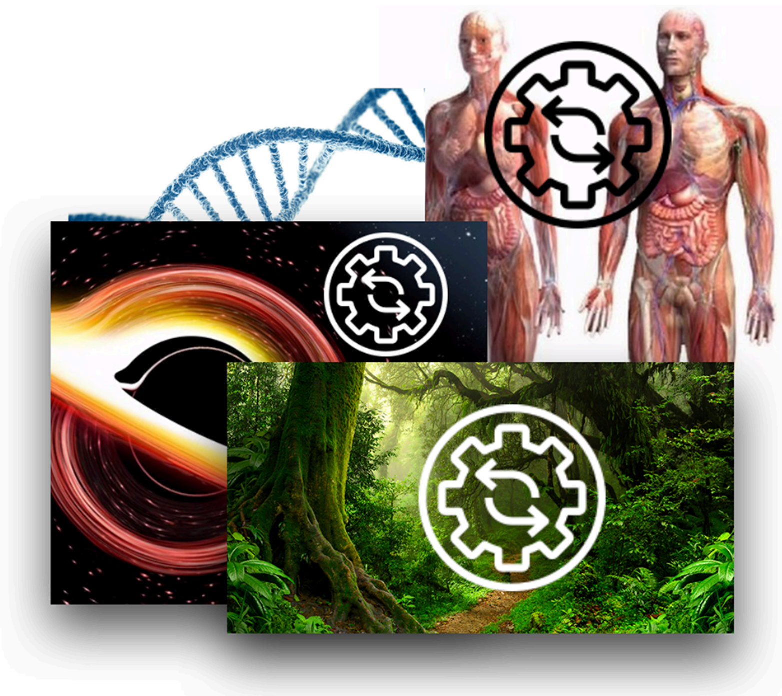
Interpretability in ML

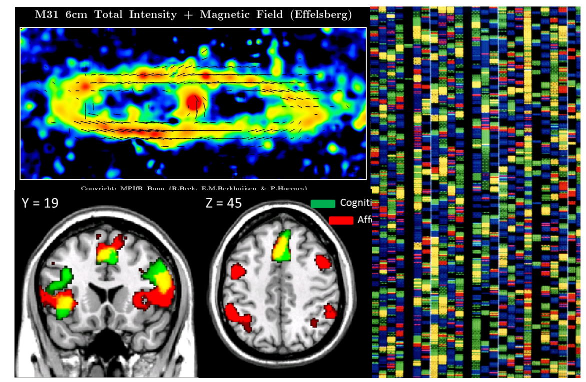
- Trust in model's decision
- Legal transparency
- Debugging

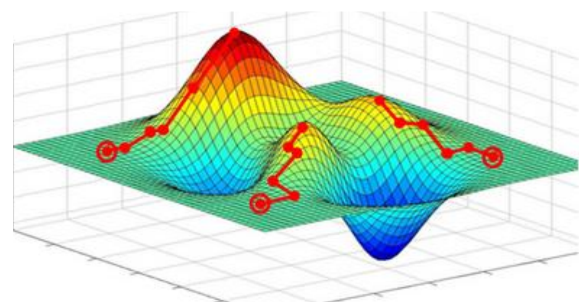
At this point the model knows something about the world that the scientist does not

https://www.wired.com/story/the-rise-of-dna-data-storage/ https://www3.mpifr-bonn.mpg.de/staff/rbeck/MKSP/pictures.html https://www.youtube.com/watch?v=Evsqy0a_Lrk https://www.sciencemag.org/news/2018/08/scientists-tweak-dna-viable-human-embryos
http://andysbrainblog.blogspot.com/2015/04/slice-analysis-of-fmri-data-with-spm.html https://astronomy.com/news/2019/01/maybe-you-really-can-use-black-holes-to-travel-the-universe https://www.sciencenewsforstudents.org/article/life-earth-mostly-green

In life sciences model interpretation has a special significance







A model that can make correct predictions about reality

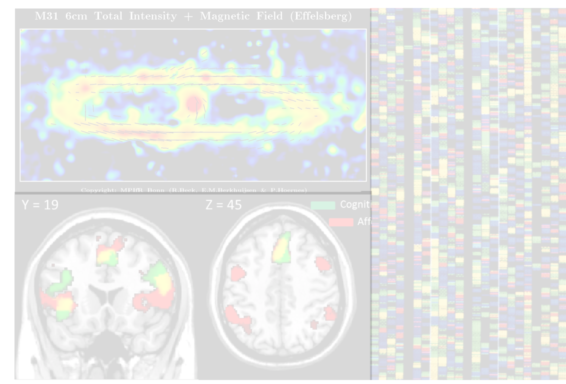
At this point the model knows something about the world that the scientist does not

Interpretability in ML

- Trust in model's decision
- Legal transparency
- Debugging
- Articulating the knowledge about the underlying process that the model has captured

https://www.wired.com/story/the-rise-of-dna-data-storage/ https://www3.mpifr-bonn.mpg.de/staff/rbeck/MKSP/pictures.html https://www.youtube.com/watch?v=Evsqy0a_Lrk https://www.sciencemag.org/news/2018/08/scientists-tweak-dna-viable-human-embryos
http://andysbrainblog.blogspot.com/2015/04/slice-analysis-of-fmri-data-with-spm.html https://astronomy.com/news/2019/01/maybe-you-really-can-use-black-holes-to-travel-the-universe https://www.sciencenewsforstudents.org/article/life-earth-mostly-green

In life sciences model interpretation has a special significance



A model that can make correct predictions about reality

Interpretability in ML

- Trust in model's decision
- Legal transparency
- Debugging
- Articulating the knowledge about the underlying process that the model has captured

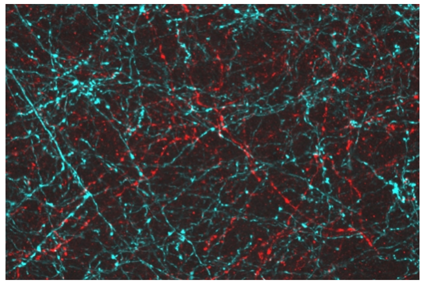
At this point the model knows something about the world that the scientist does not

Machine-learned models as scientific theories

- Both capture and model observations
- Both make correct predictions on new observations
- Computers can generate and test theories faster than humans
- In the big data regime there is not enough scientists to sift through all potential explanations of the data
- Algorithms have different bias than humans, hence find different solutions

https://www.wired.com/story/the-rise-of-dna-data-storage/ https://www3.mpifr-bonn.mpg.de/staff/rbeck/MKSP/pictures.html https://www.youtube.com/watch?v=Evsqy0a_Lrk https://www.sciencemag.org/news/2018/08/scientists-tweak-dna-viable-human-embryos
http://andysbrainblog.blogspot.com/2015/04/slice-analysis-of-fmri-data-with-spm.html https://astronomy.com/news/2019/01/maybe-you-really-can-use-black-holes-to-travel-the-universe https://www.sciencenewsforstudents.org/article/life-earth-mostly-green

Applying this principle in Neuroscience



"Identifying task-relevant spectral signatures of perceptual categorization in the human cortex"
Scientific Reports, 2020

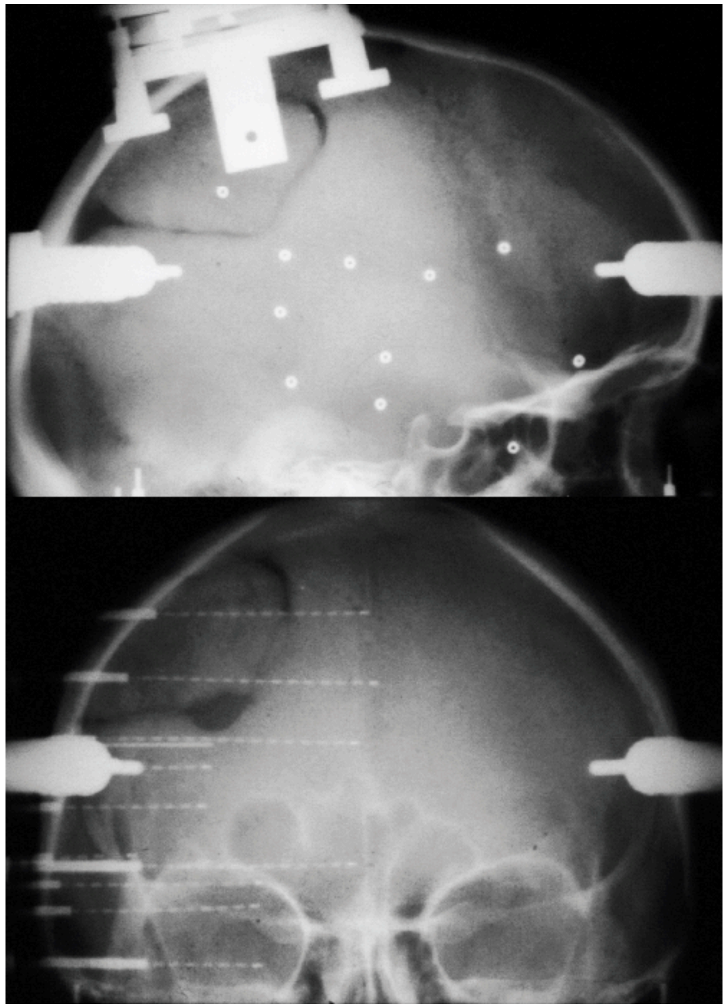
"Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex"
Communications Biology, 2018

"Mental state space visualization for interactive modeling of personalized BCI control strategies"
Journal of Neural Engineering, 2020

Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex

CRNL
Centre de Recherche Neurosciences Lyon

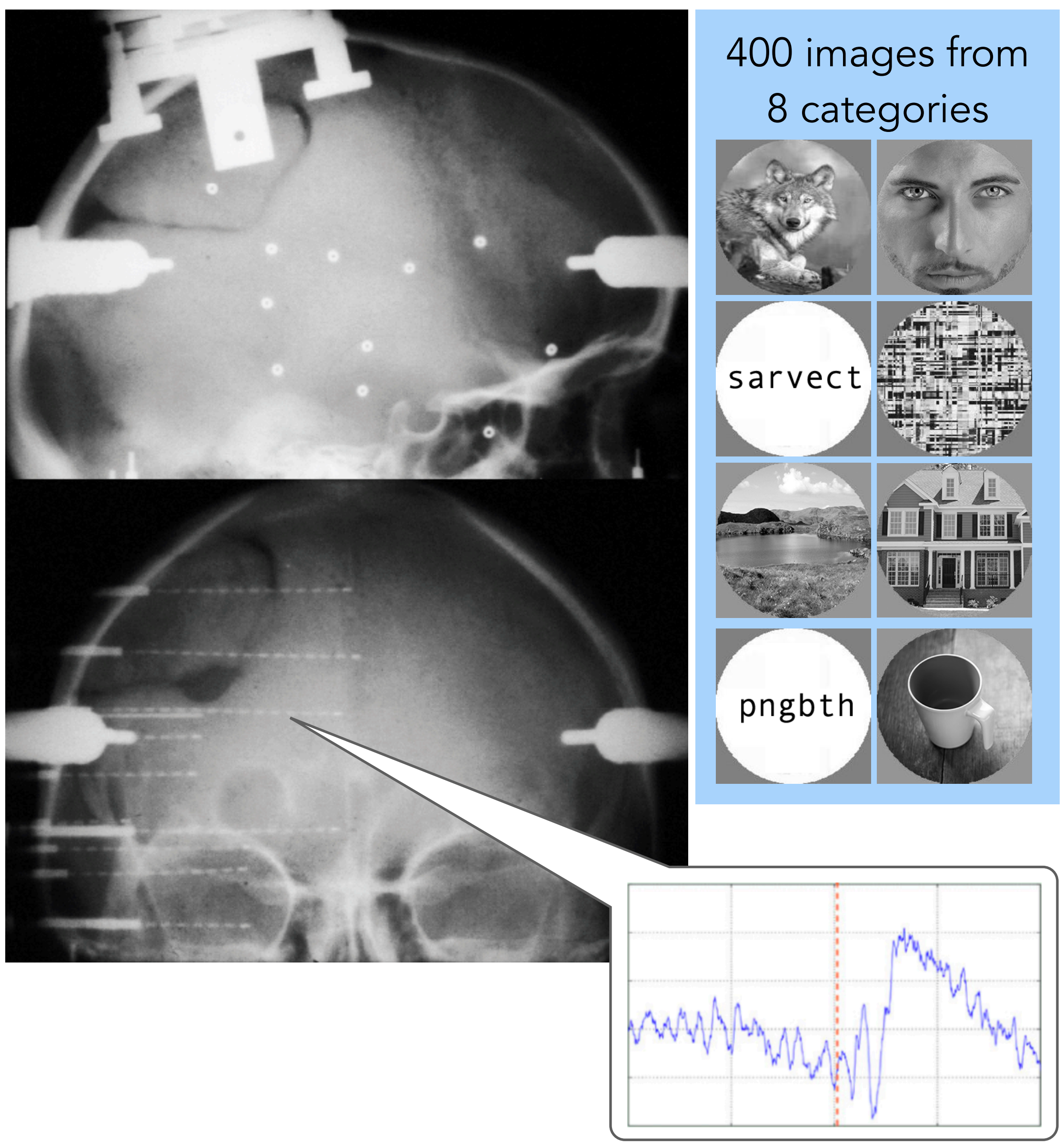
100 patients
12,000 electrodes



Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex

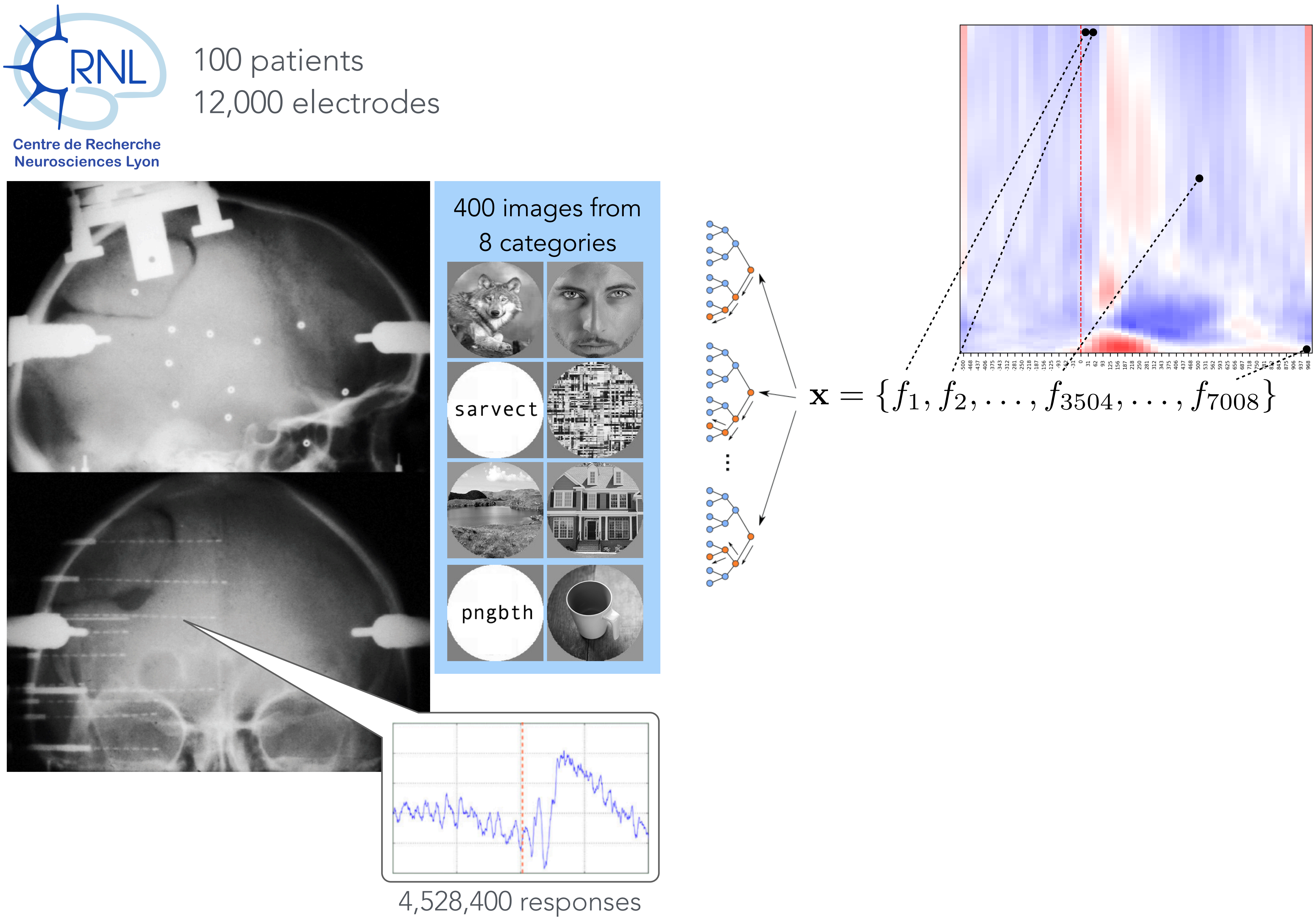
CRNL
Centre de Recherche
Neurosciences Lyon

100 patients
12,000 electrodes



4,528,400 responses

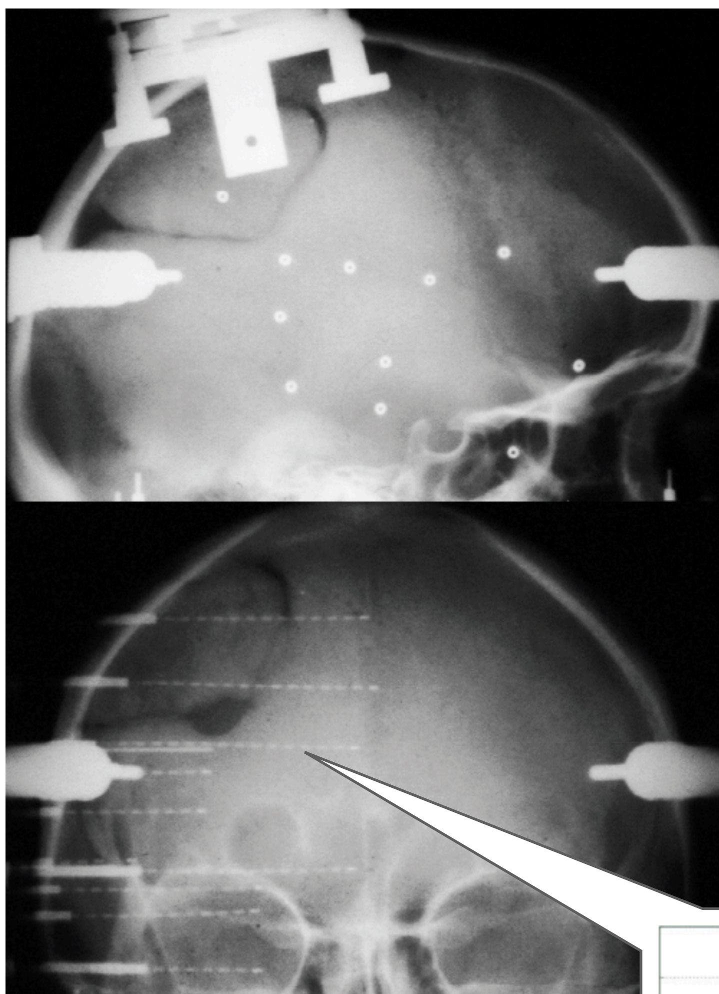
Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex

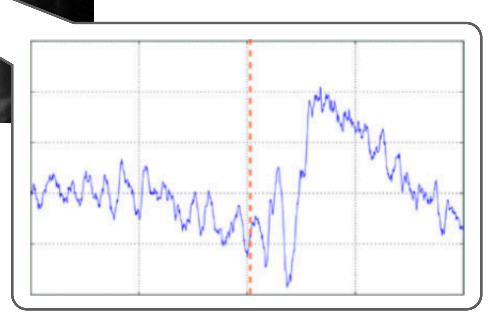


Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex

CRNL
Centre de Recherche Neurosciences Lyon

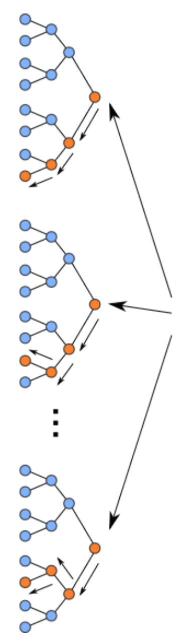
100 patients
12,000 electrodes



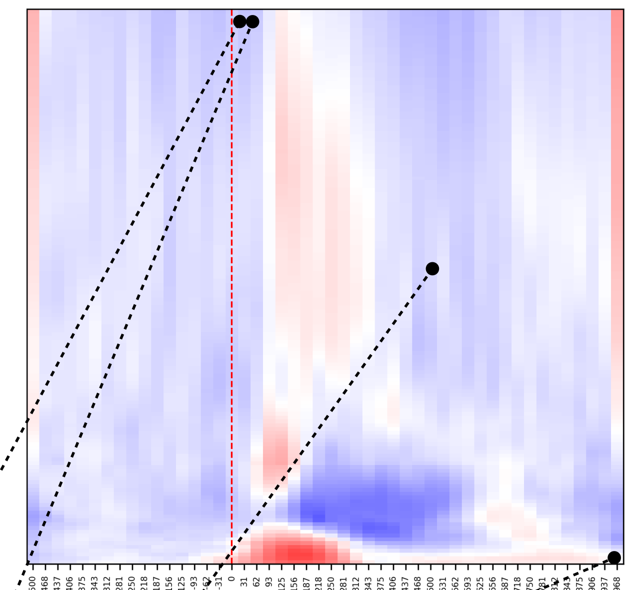


4,528,400 responses

400 images from 8 categories



$$\mathbf{x} = \{f_1, f_2, \ldots, f_{3504}, \ldots, f_{7008}\}$$

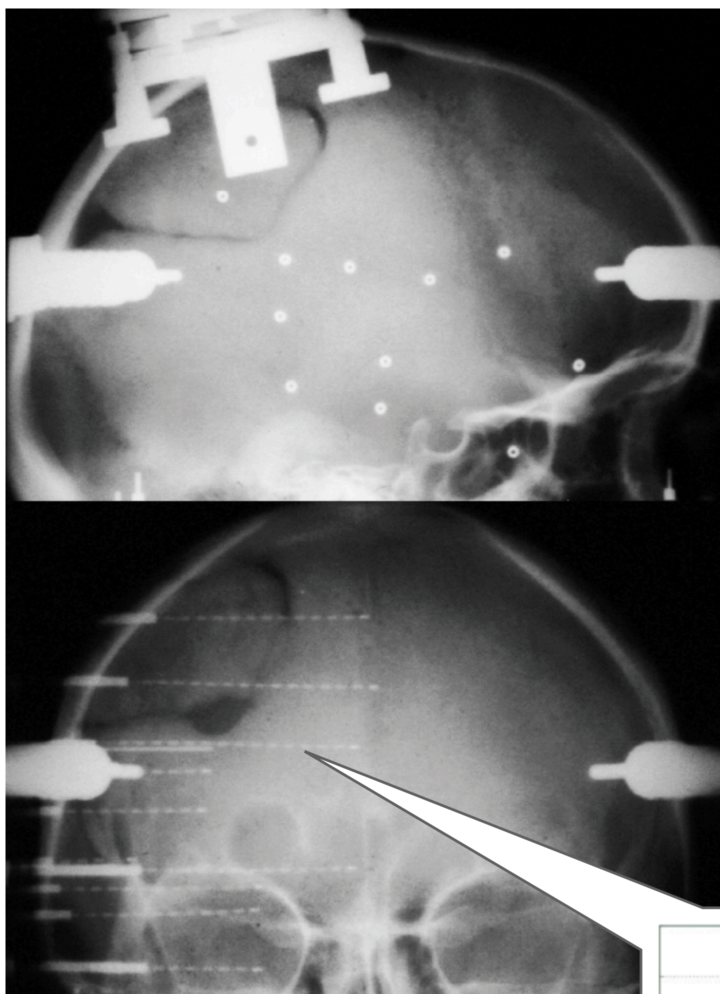


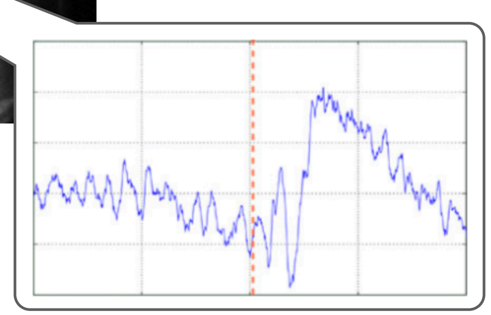
Which parts of this brain activity are generated by the **underlying process** of mental image categorization?

Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex

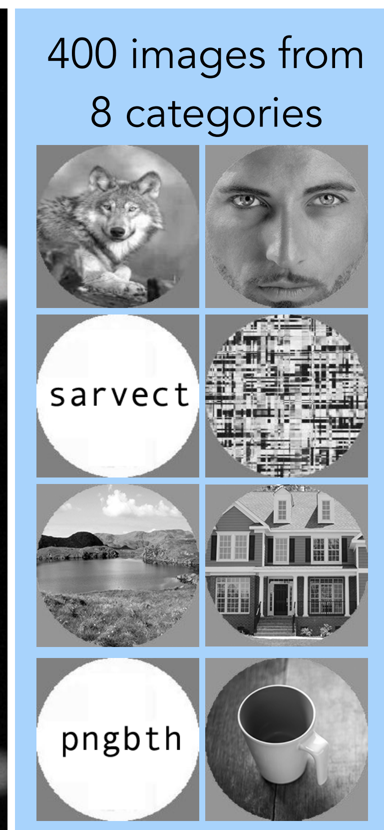
Centre de Recherche Neurosciences Lyon

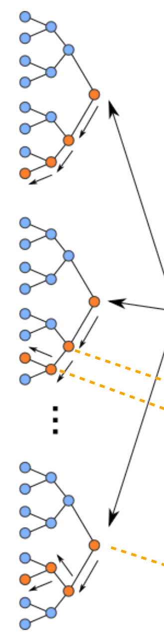
100 patients
12,000 electrodes

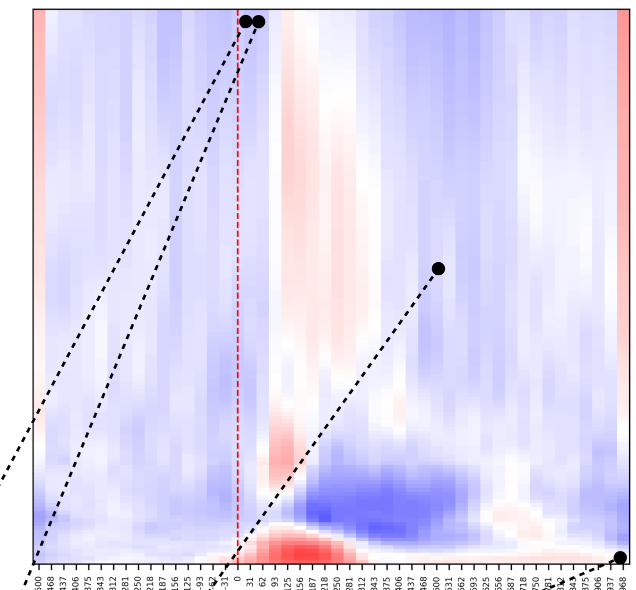




4,528,400 responses

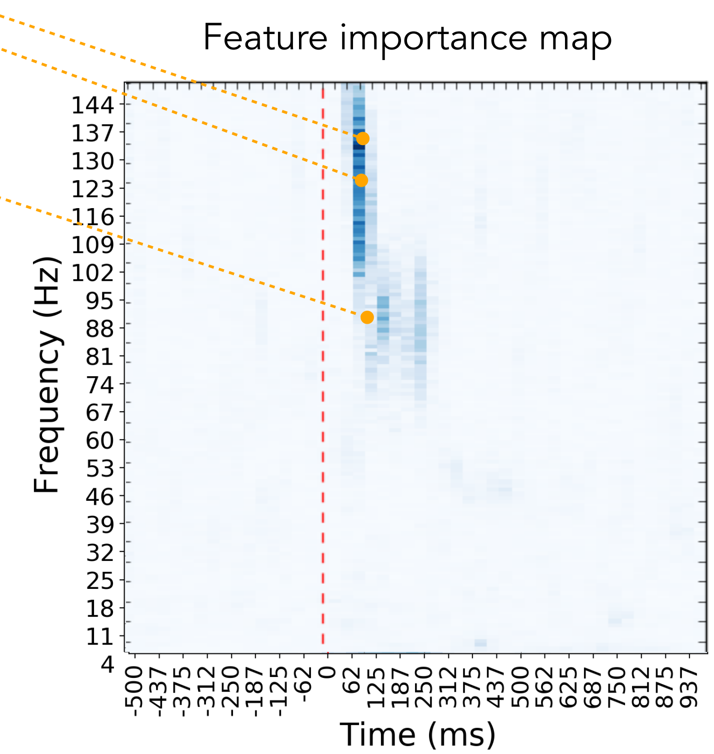




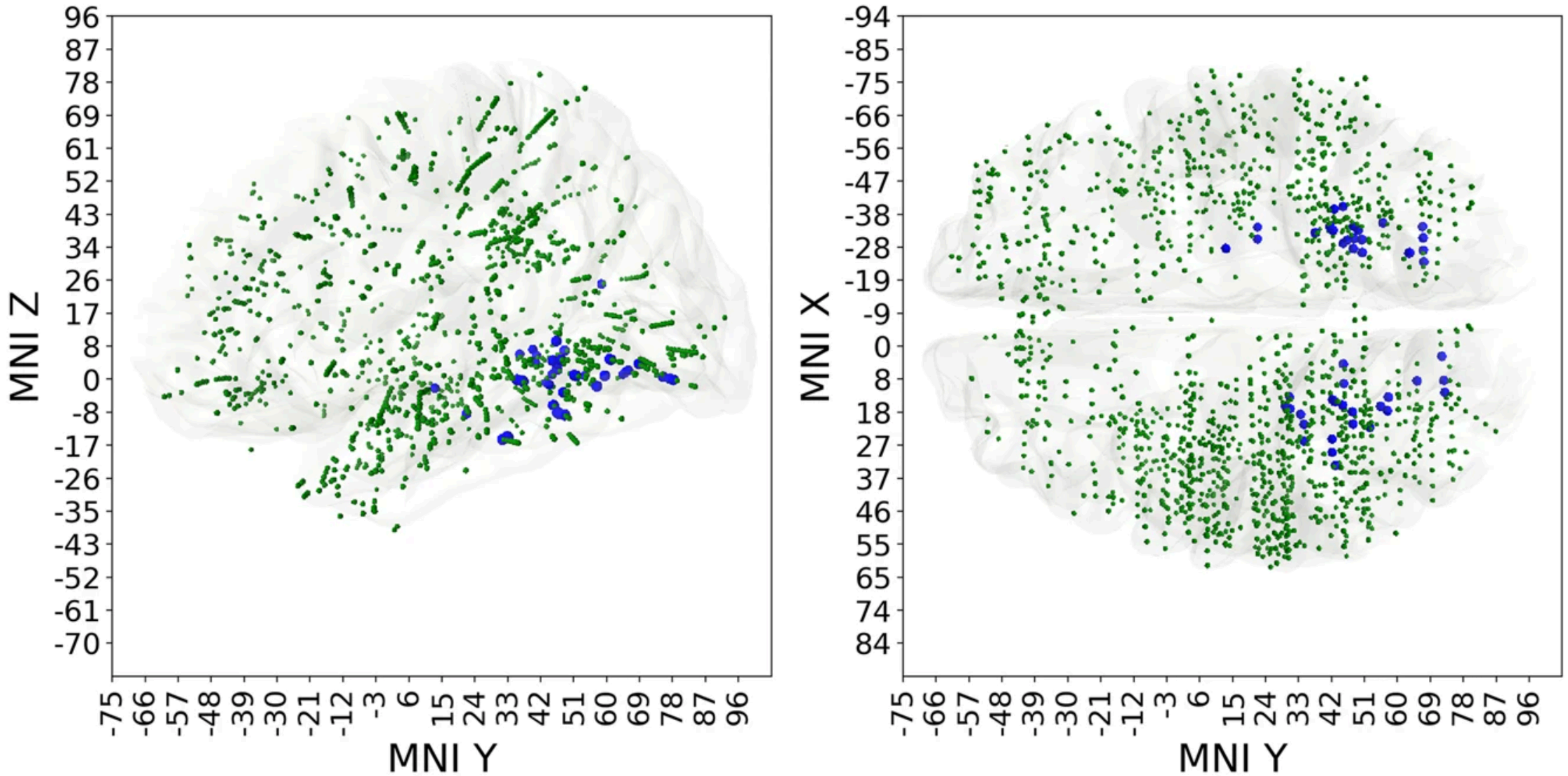


$$\mathbf{x} = \{f_1, f_2, \ldots, f_{3504}, \ldots, f_{7008}\}$$

Which parts of this brain activity are generated by the **underlying process** of mental image categorization?

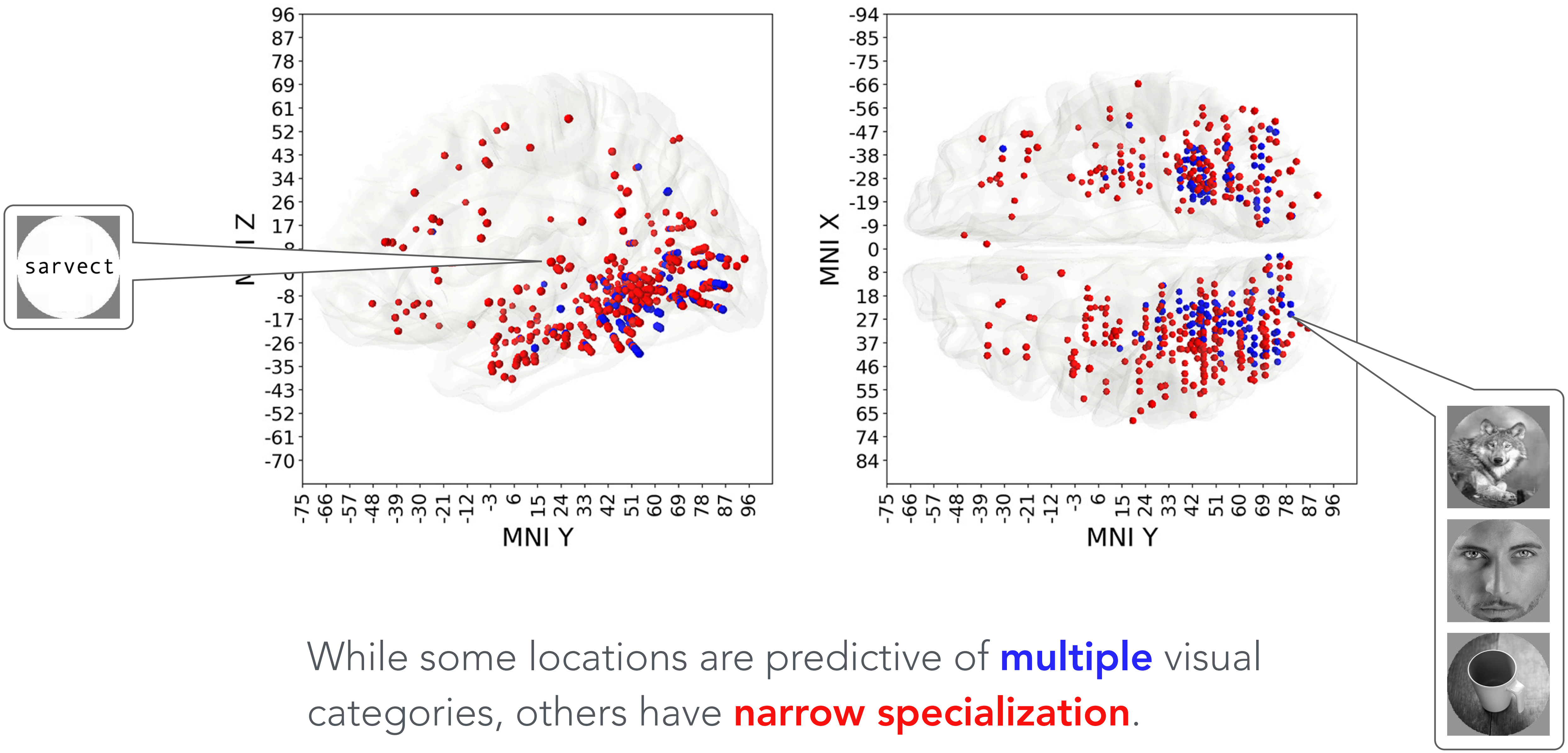


Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex



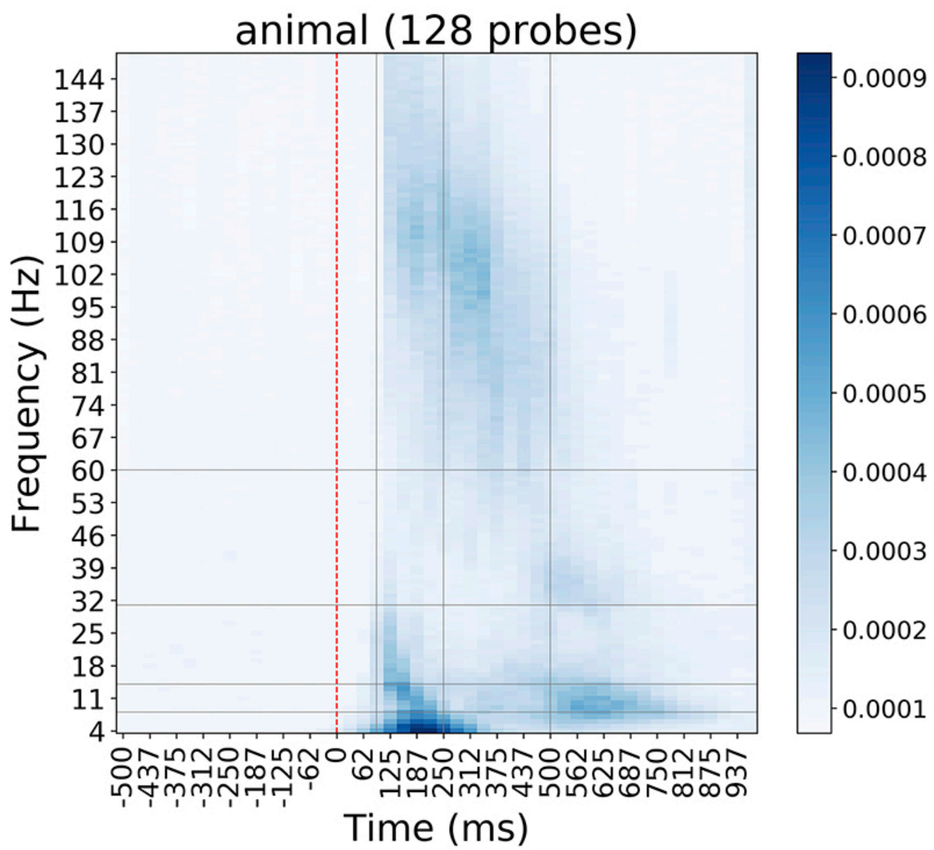
Multiple brain regions **respond** when a stimulus is shown, but only a small fraction (5%) are **predictive** of a category.

Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex

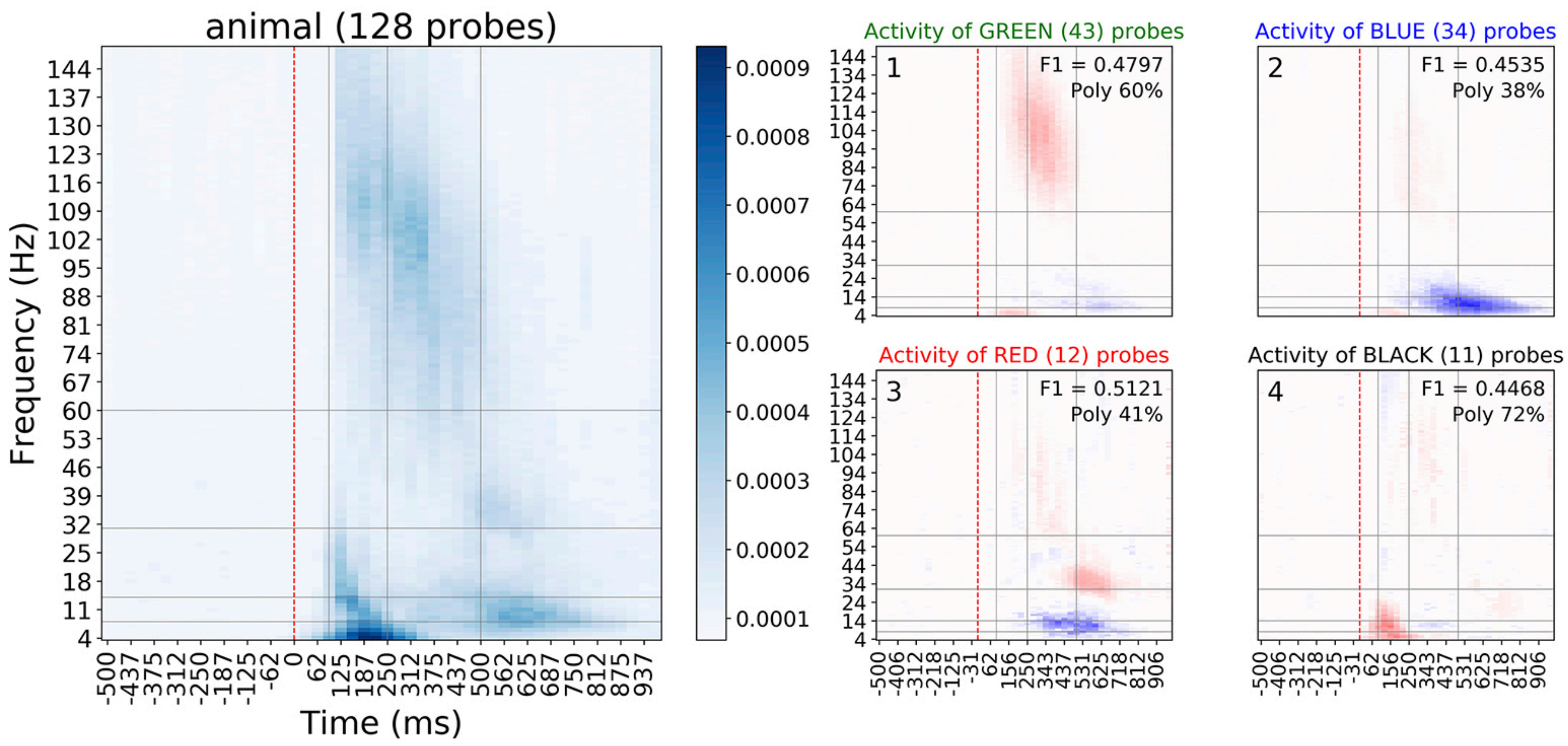


While some locations are predictive of **multiple** visual categories, others have **narrow specialization**.

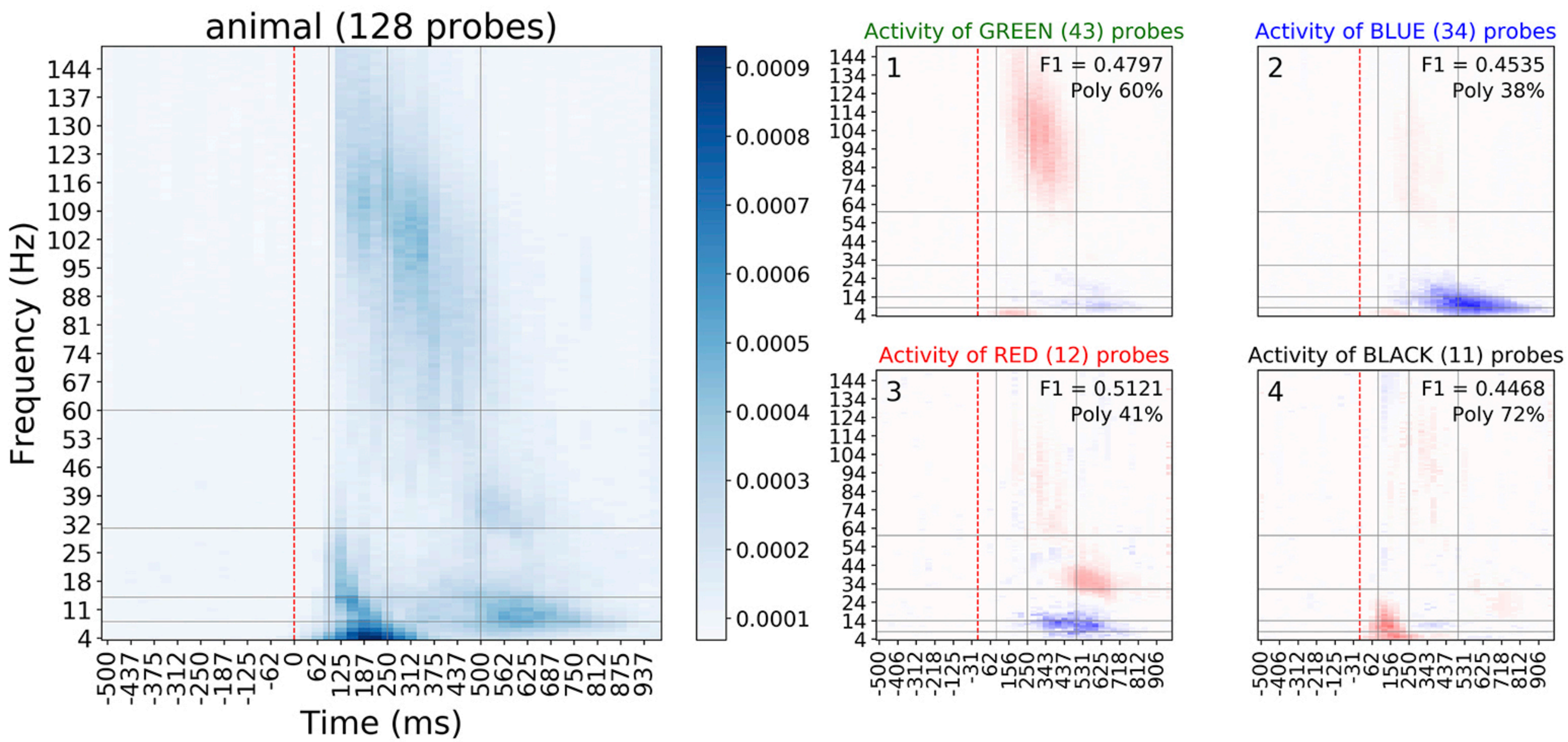
Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex



Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex



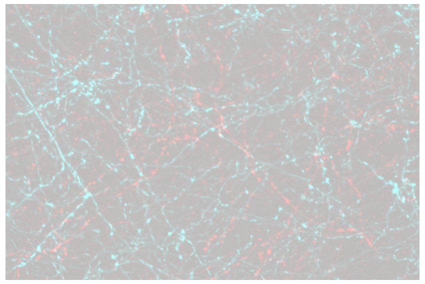
Identifying task-relevant spectral signatures of perceptual categorization in the human visual cortex



It is believed that high-frequency activity reflects the information processing during high cognitive tasks. We show that low-frequency activity is almost as important for the task at hand.

Across all categories the classifier relied on power decreases in different brain networks, not only on the increases to perform the classification

Applying this principle in Neuroscience

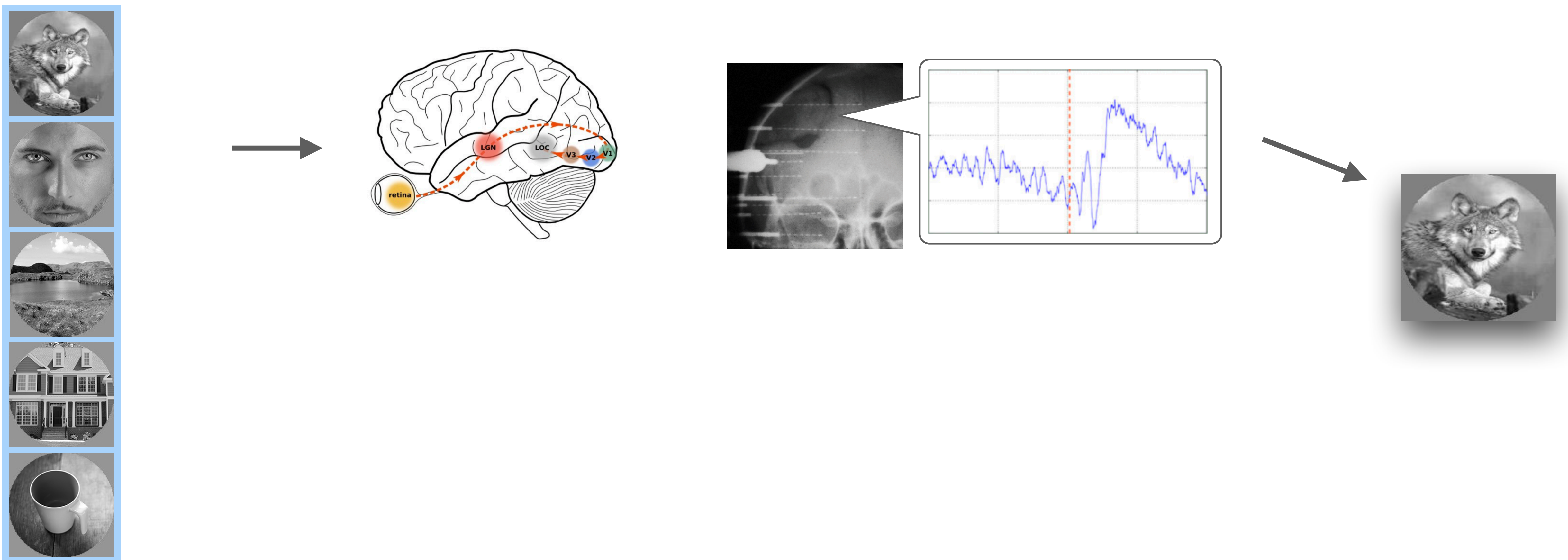


"Identifying task-relevant spectral signatures of perceptual categorization in the human cortex"
Scientific Reports, 2020

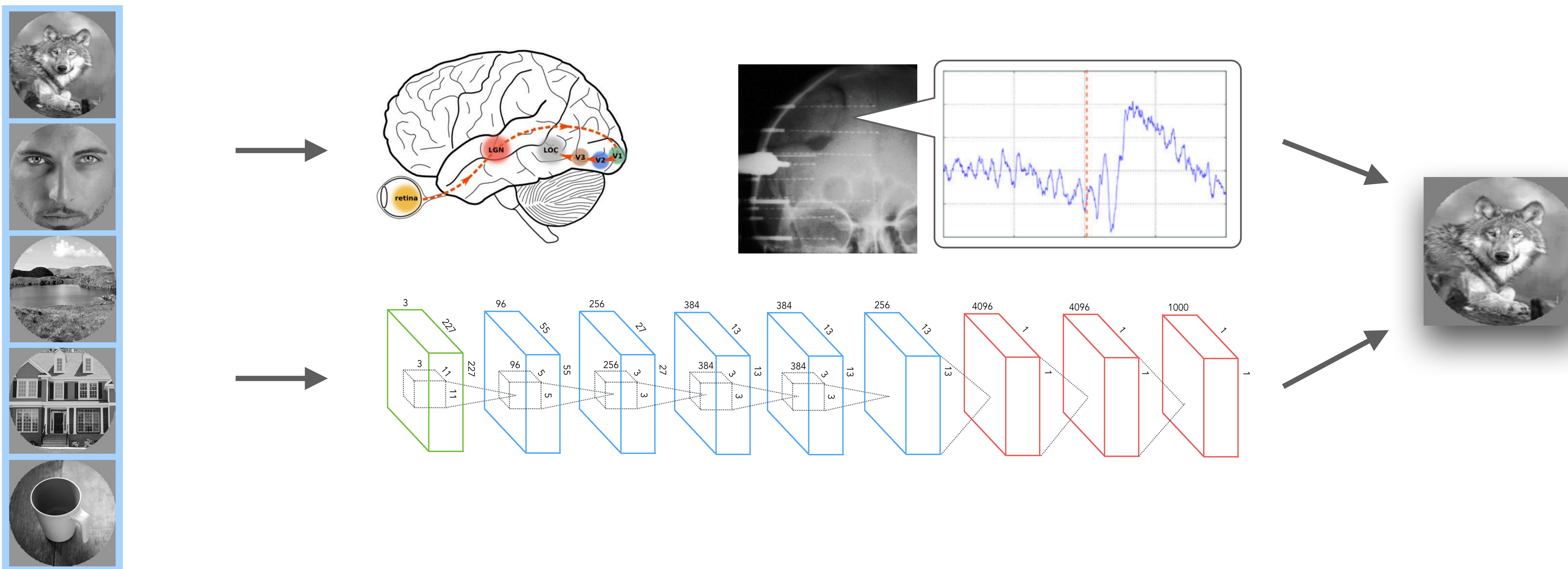
"Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex"
Communications Biology, 2018

"Mental state space visualization for interactive modeling of personalized BCI control strategies"
Journal of Neural Engineering, 2020

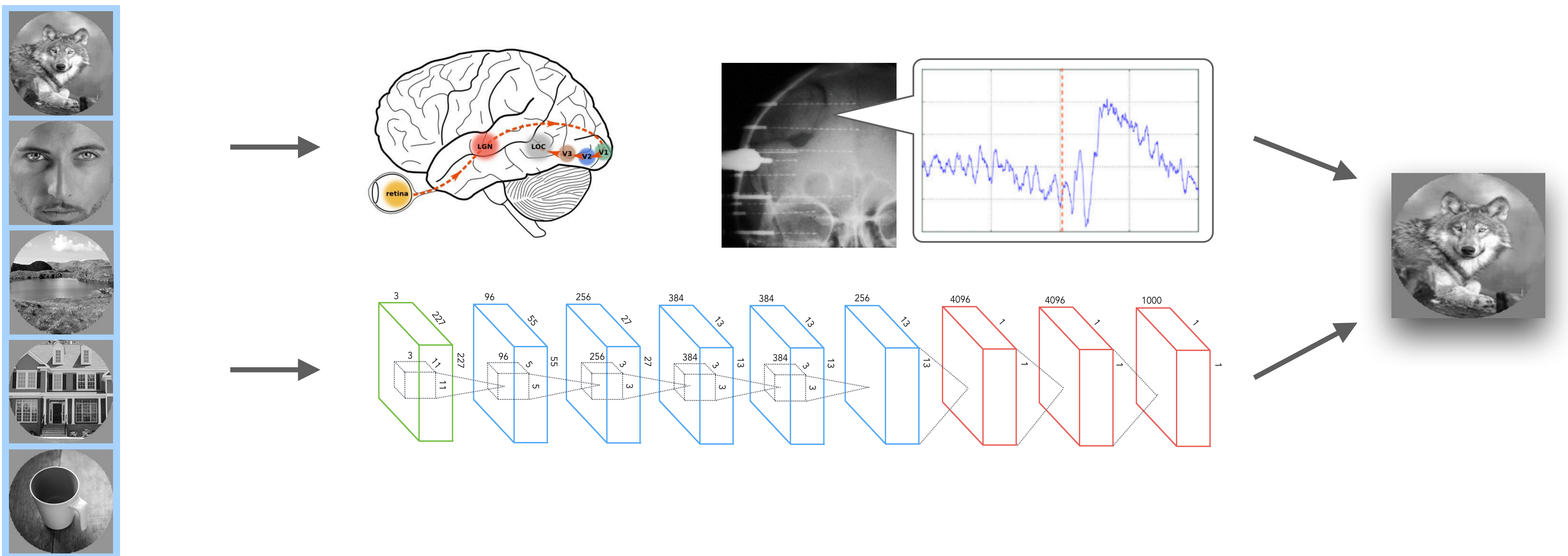
Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



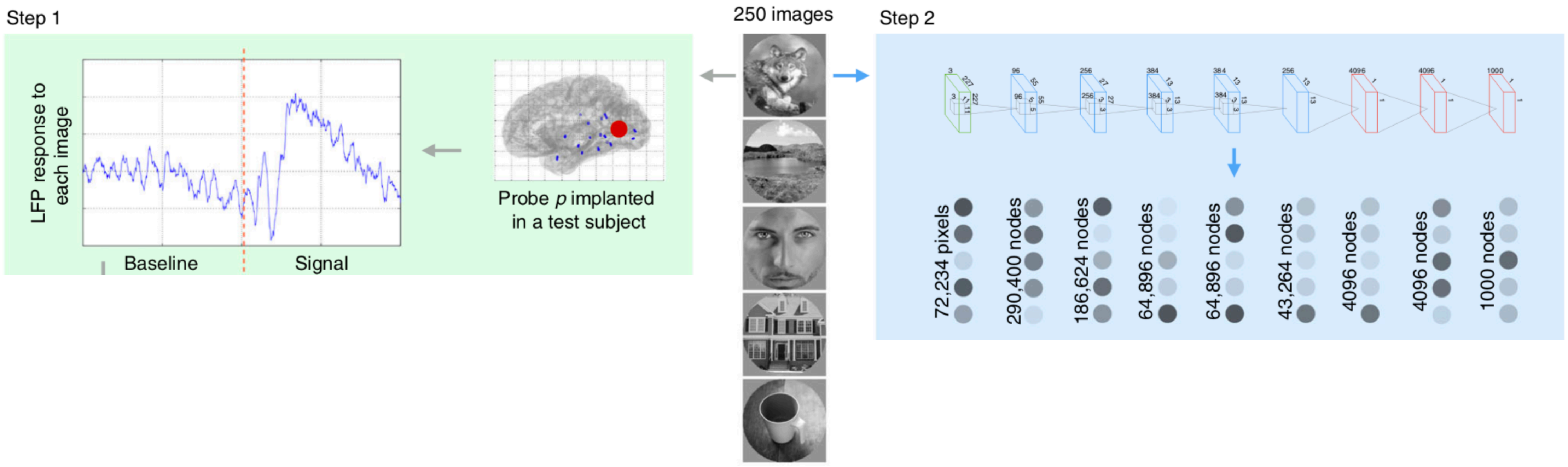
Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



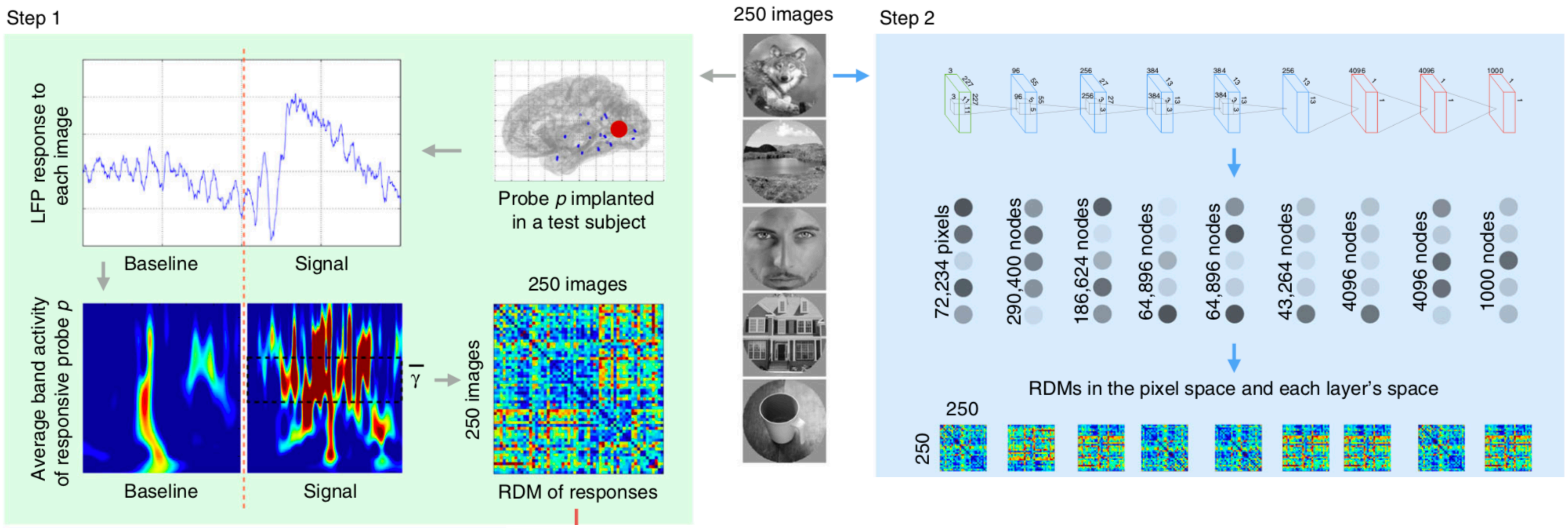
From previous fMRI research we knew that there is a mapping between the hierarchies

Intracranial electrophysiological data allowed us to learn when and at which frequencies

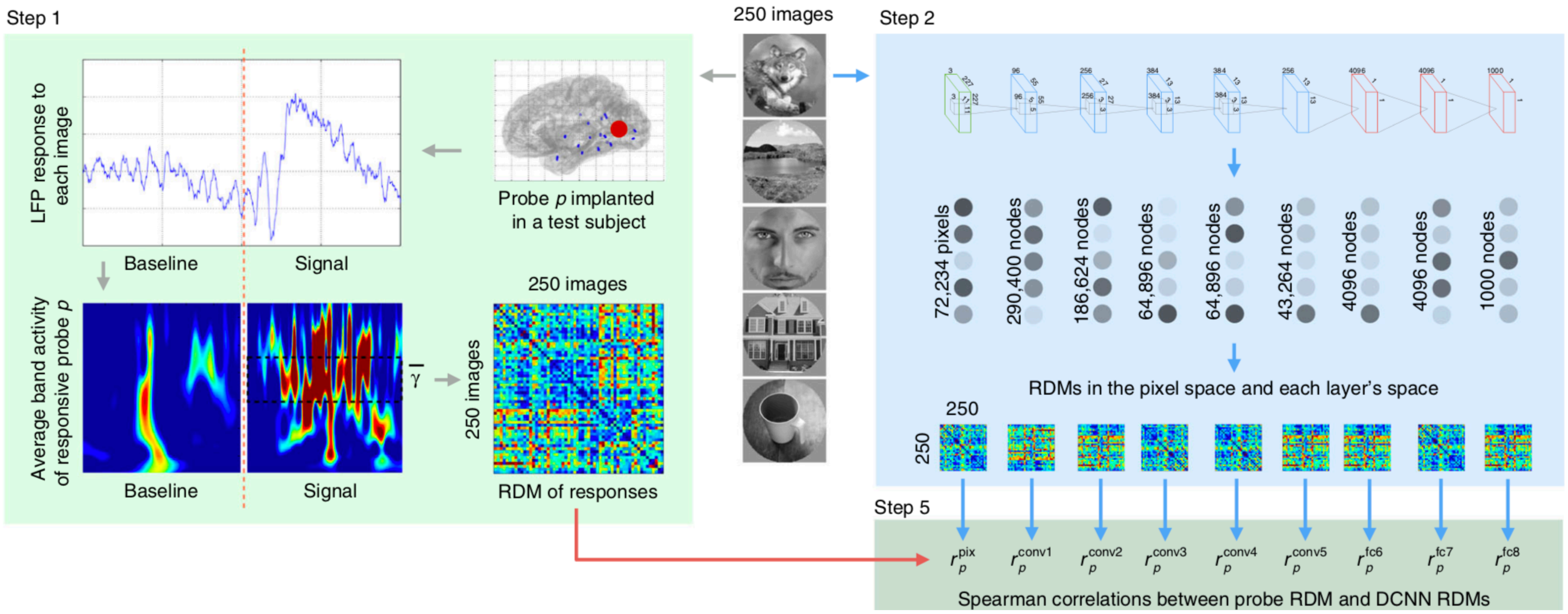
Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



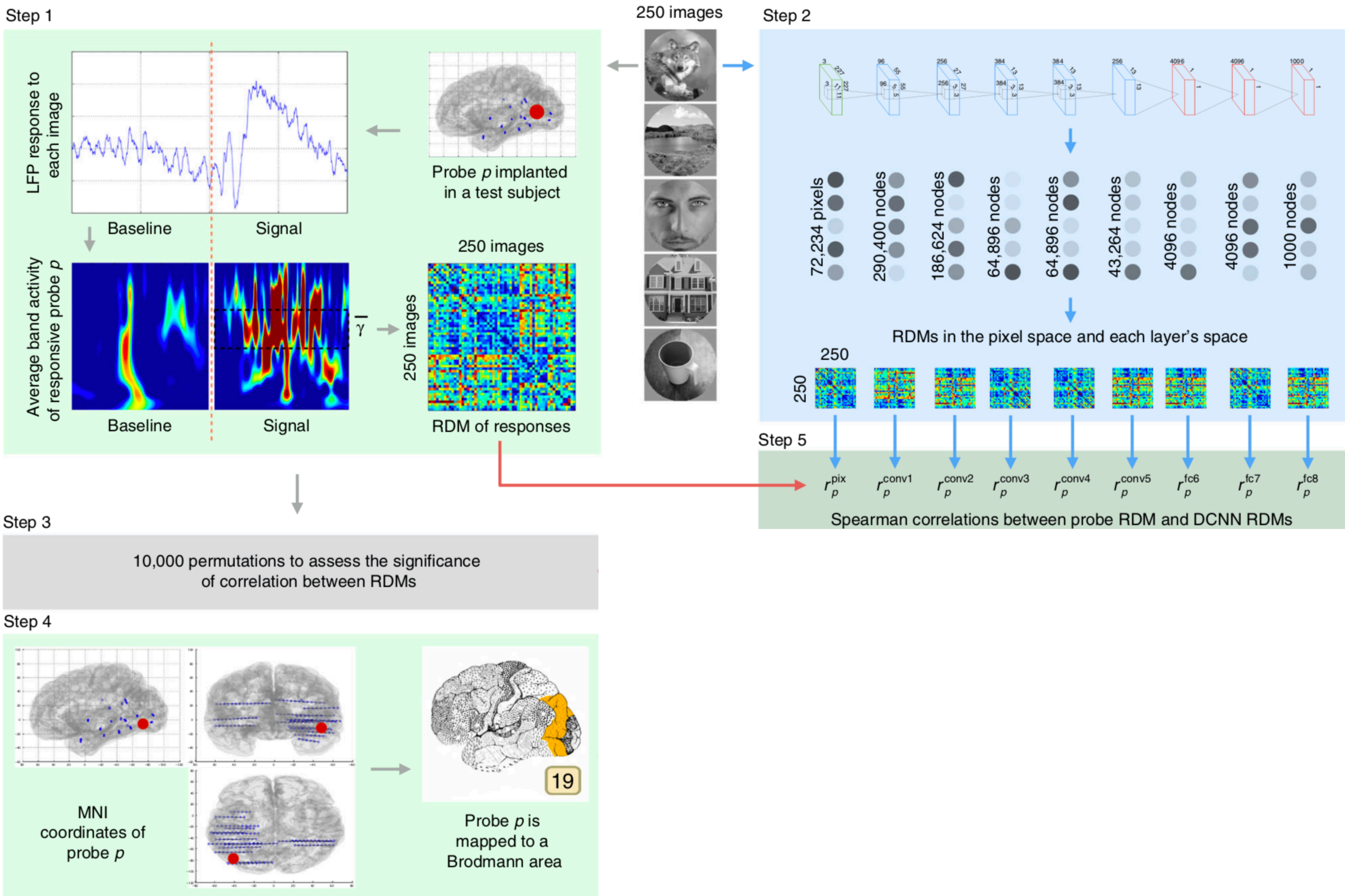
Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



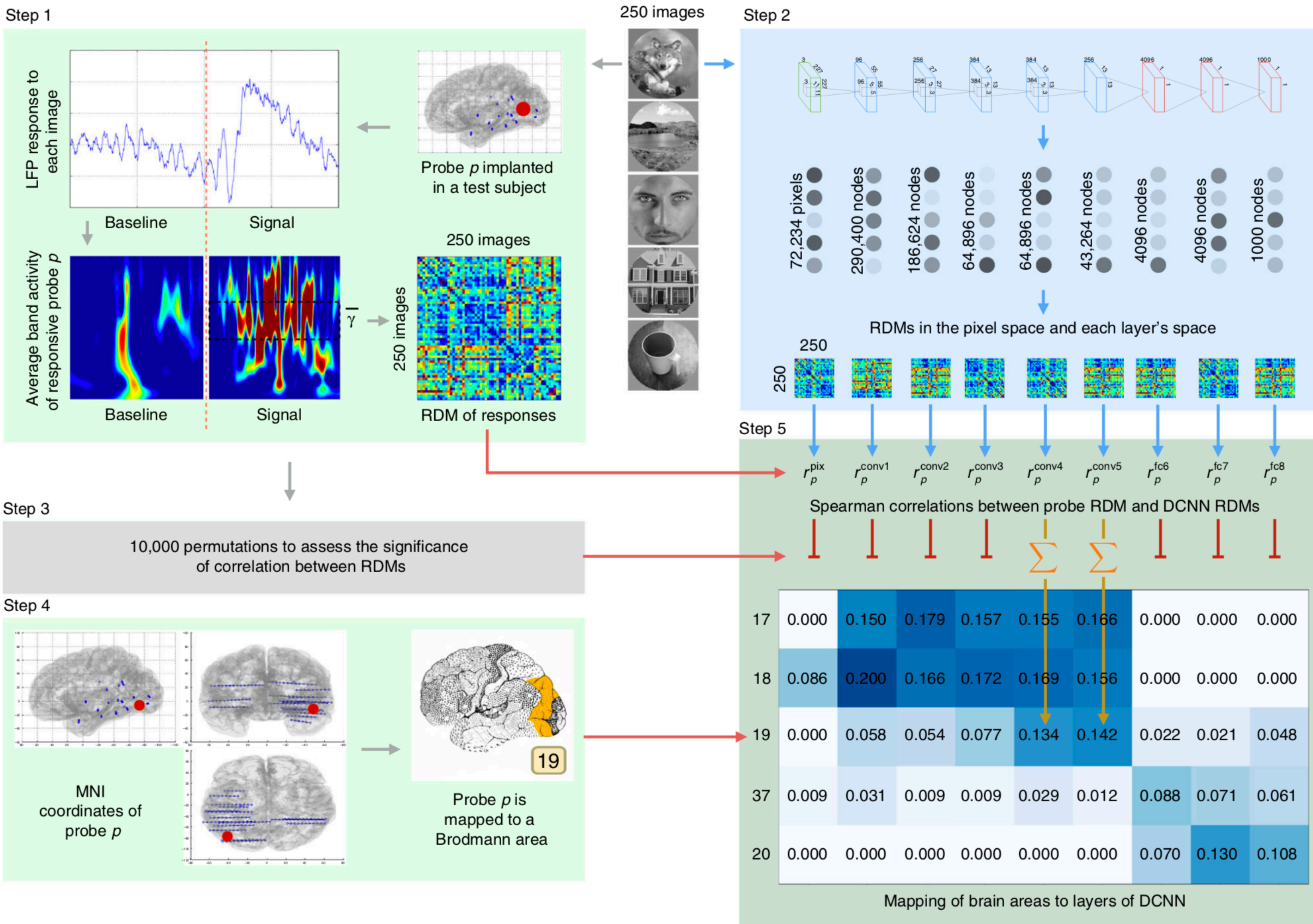
Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex

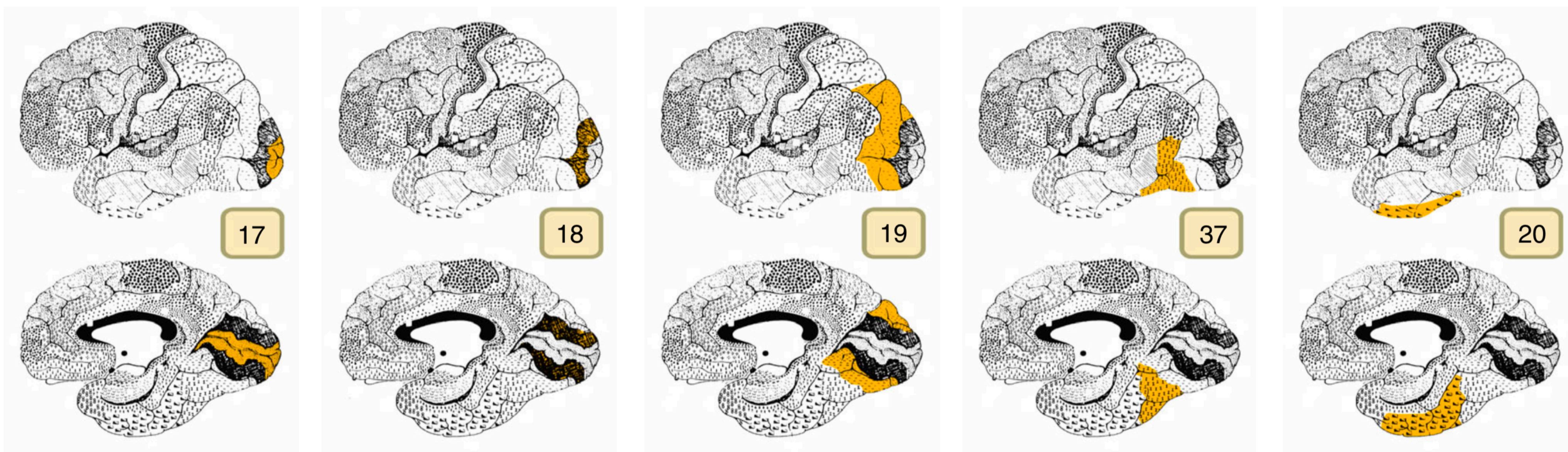


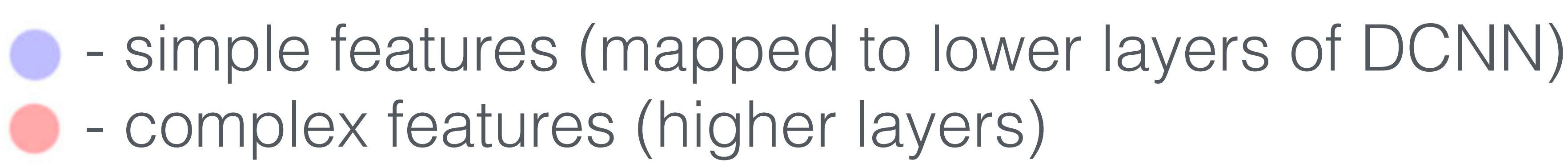
Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex

17	0.000	0.150	0.179	0.157	0.155	0.166	0.000	0.000	0.000
18	0.086	0.200	0.166	0.172	0.169	0.156	0.000	0.000	0.000
19	0.000	0.058	0.054	0.077	0.134	0.142	0.022	0.021	0.048
37	0.009	0.031	0.009	0.009	0.029	0.012	0.088	0.071	0.061
20	0.000	0.000	0.000	0.000	0.000	0.000	0.070	0.130	0.108

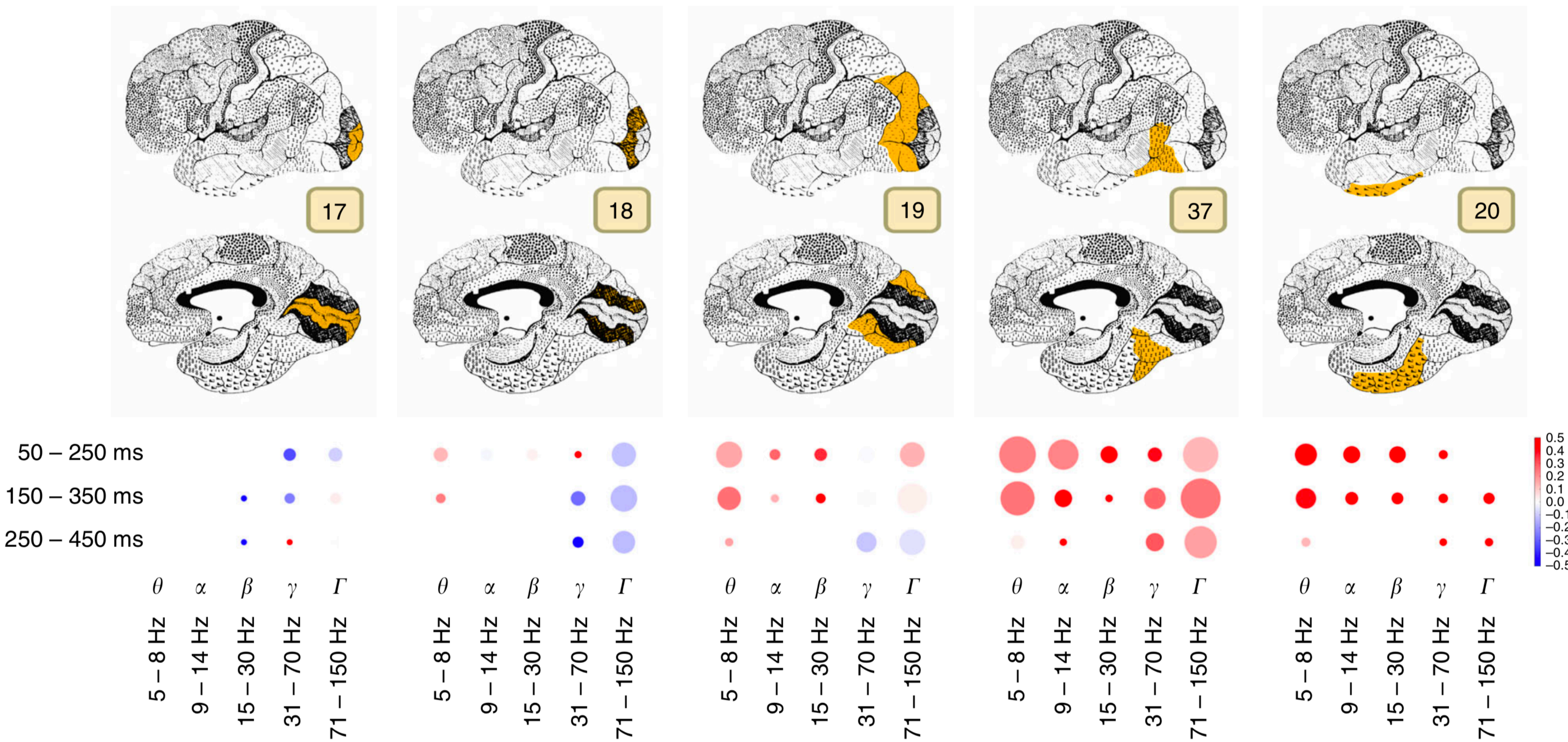
Mapping of brain areas to layers of DCNN

Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex



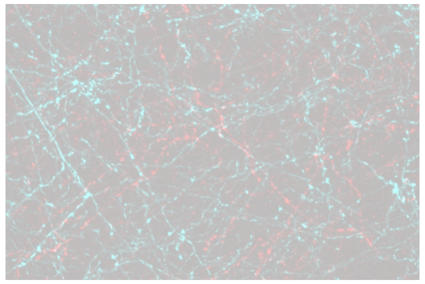


Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex

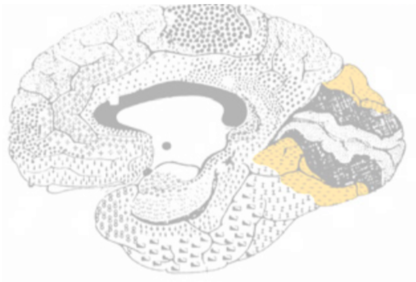


- simple features (mapped to lower layers of DCNN)
- complex features (higher layers)

Applying this principle in Neuroscience



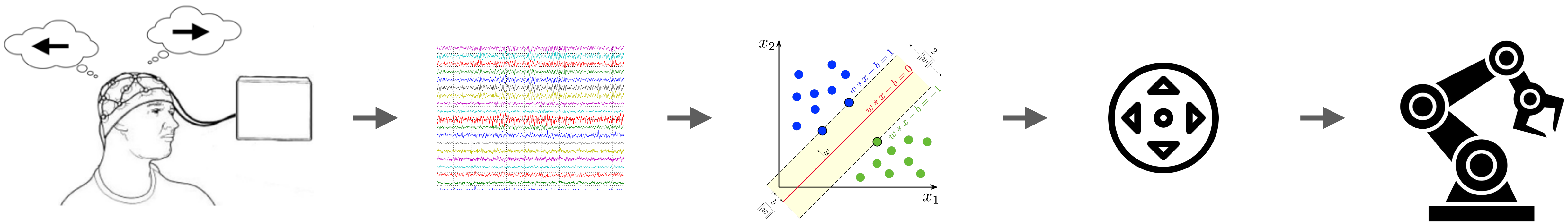
"Identifying task-relevant spectral signatures of perceptual categorization in the human cortex"
Scientific Reports, 2020



"Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex"
Communications Biology, 2018

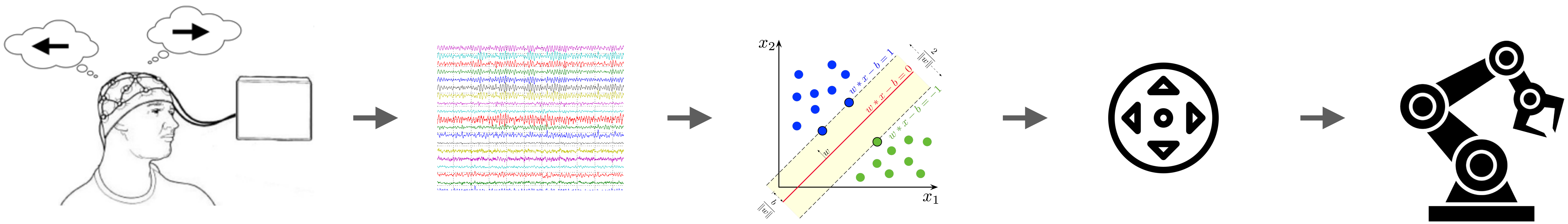
"Mental state space visualization for interactive modeling of personalized BCI control strategies"
Journal of Neural Engineering, 2020

Mental state space visualization for interactive modeling of personalized BCI control strategies



Mental actions → Brain signal → Machine learning algorithm → Control commands → An external device

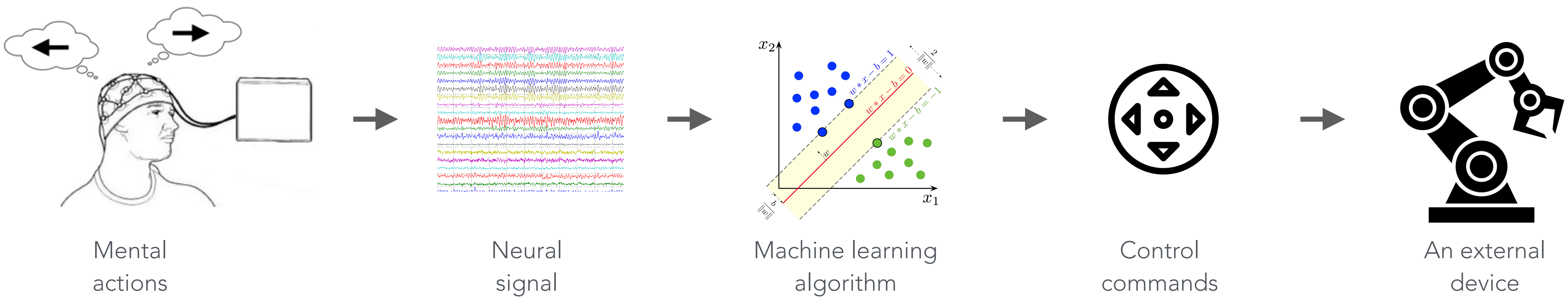
Mental state space visualization for interactive modeling of personalized BCI control strategies



The user must produce *mental actions* that are distinguishable by the machine and do that consistently

The algorithm must distinguish between different *mental actions*

Mental state space visualization for interactive modeling of personalized BCI control strategies

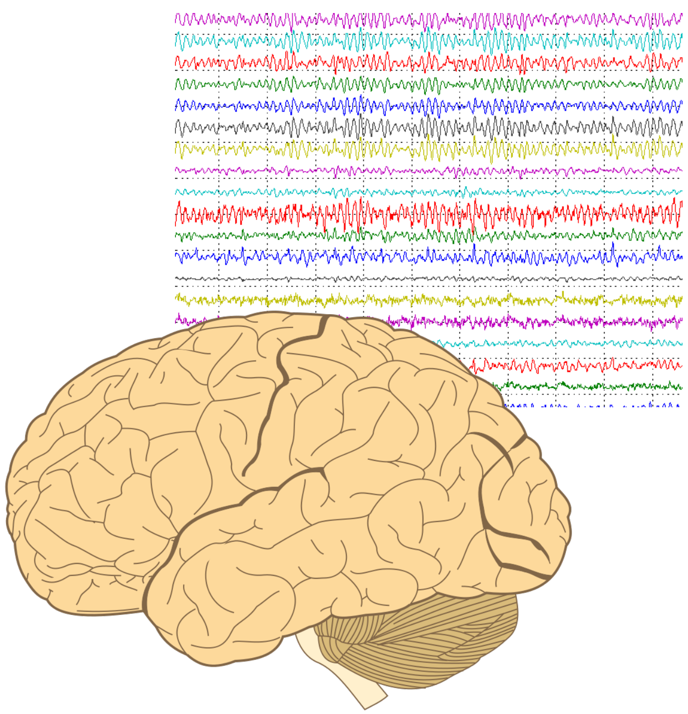


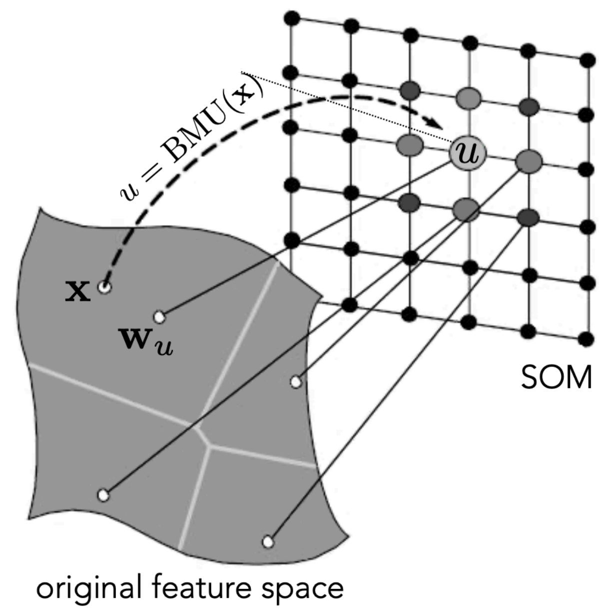
The user must produce *mental actions* that are distinguishable by the machine and do that consistently

The algorithm must distinguish between different *mental actions*

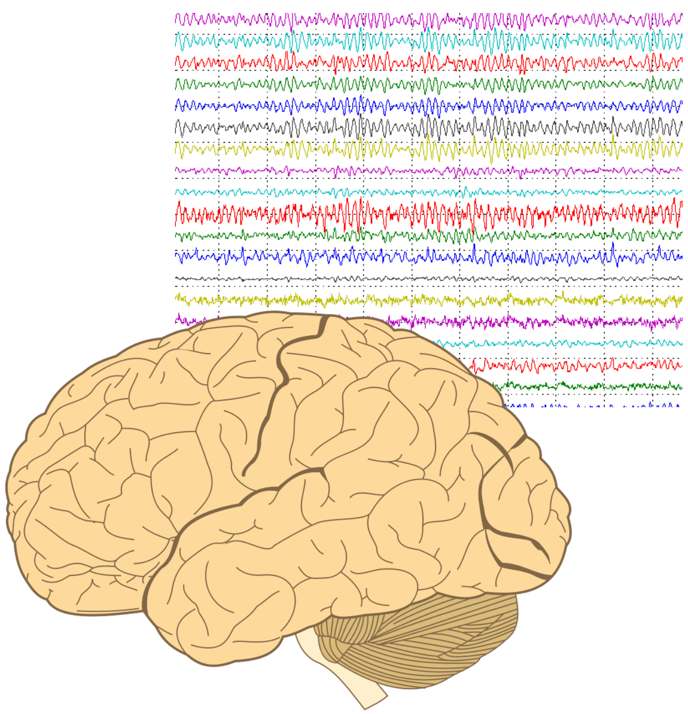
If the user could see machine's representation of his mental actions he could find out which ones are suitable

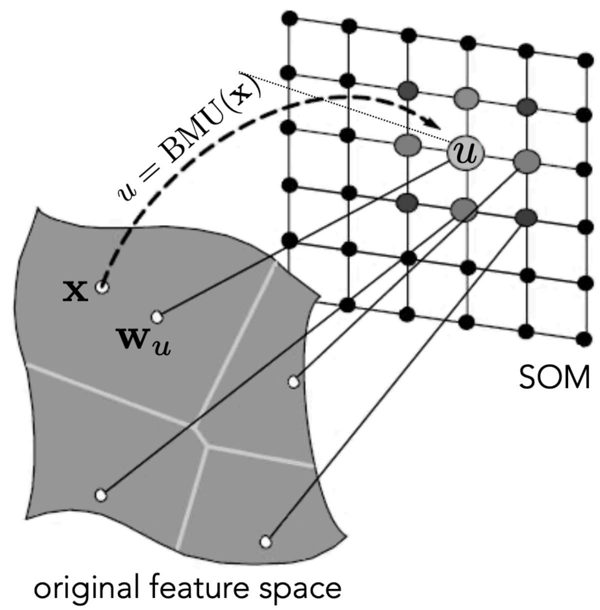
Mental state space visualization for interactive modeling of personalized BCI control strategies

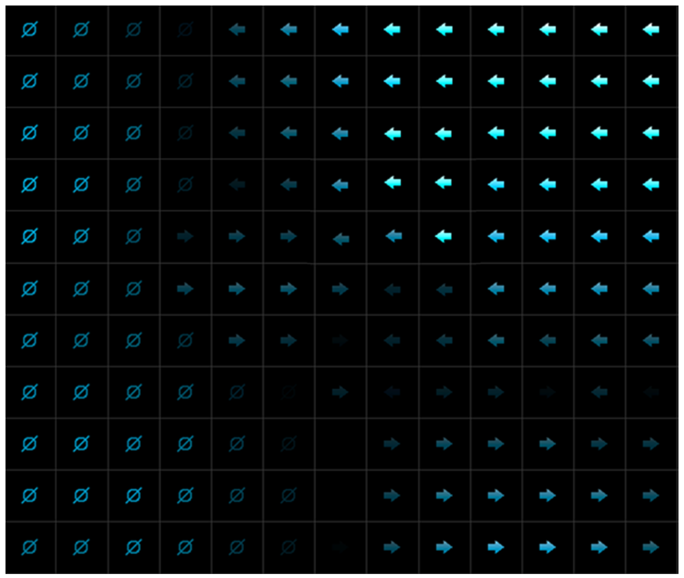




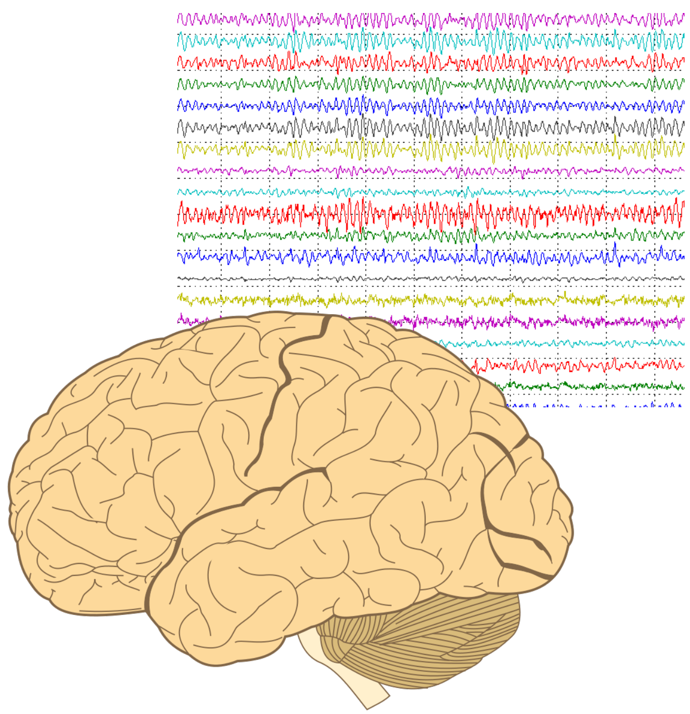
Mental state space visualization for interactive modeling of personalized BCI control strategies

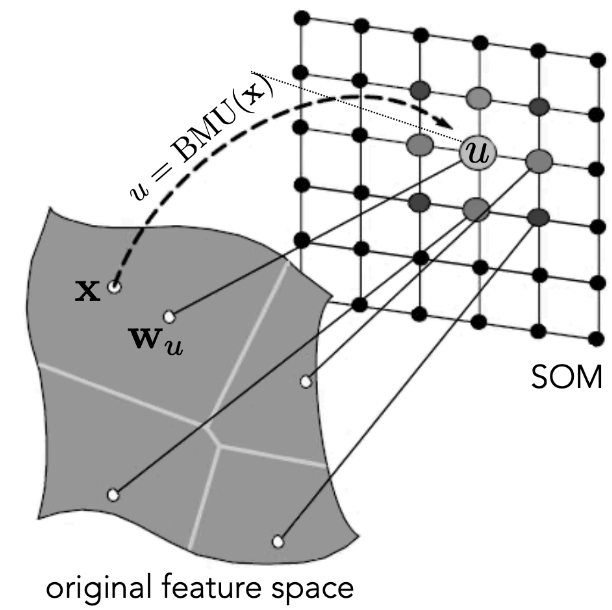


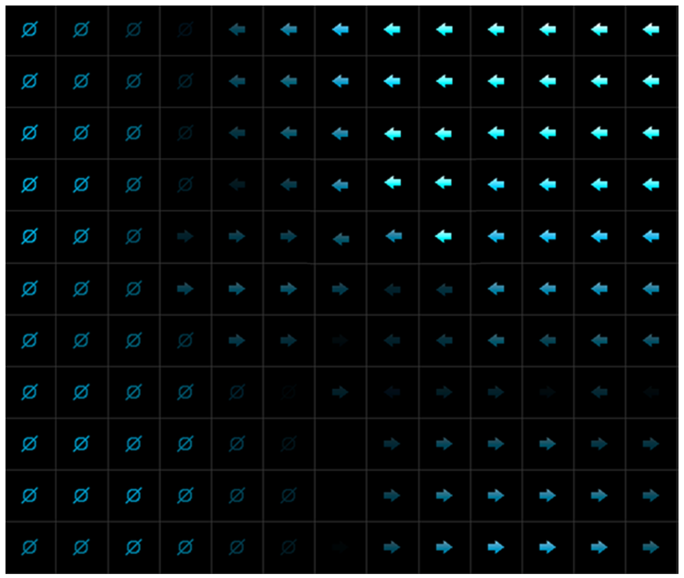


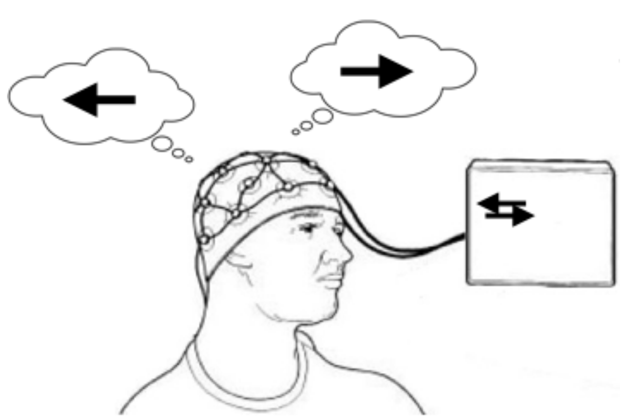


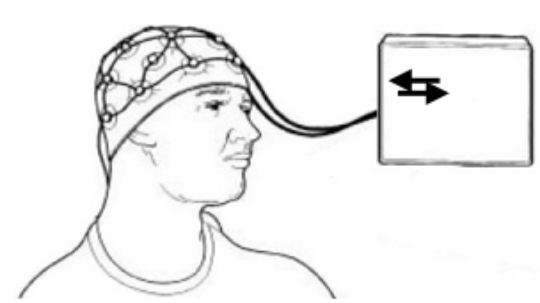
Mental state space visualization for interactive modeling of personalized BCI control strategies

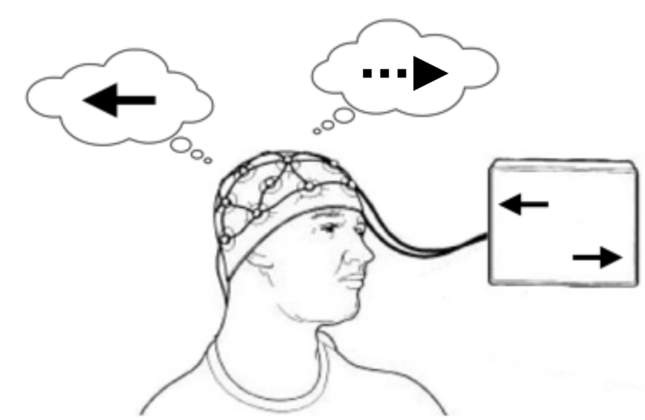










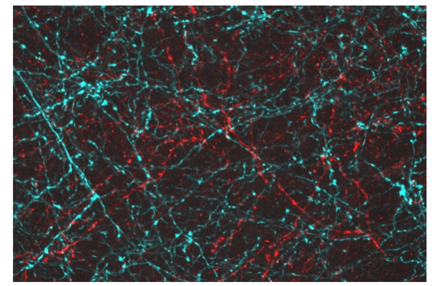


I see that the signal for ← is very similar to the signal for →

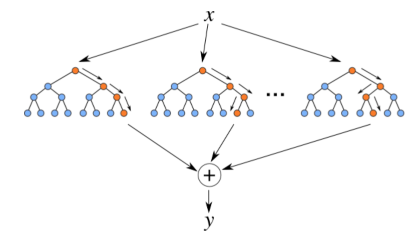
I should try to think → differently

RESEARCH PROJECT | LEVEL OF NEURAL ORGANIZATION | KNOWLEDGE REPRESENTAION IN THE MODEL

Kuzovkin et al.,
"Identifying task-relevant spectral signatures of perceptual categorization in the human cortex"
Scientific Reports, 2020 *(in review)*



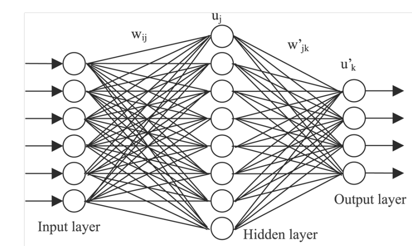
Local responses of a neural population



Random Forests: feature-based **rules**

Kuzovkin et al.,
"Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex"
Communications Biology, 2018

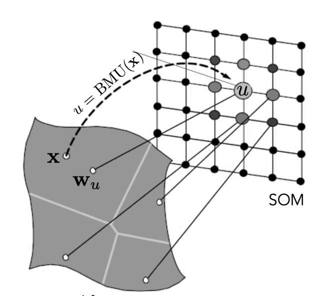
Hierarchy of visual areas



DNNs: **distributed representations** over features (input or latent)

Kuzovkin et al.,
"Mental state space visualization for interactive modeling of personalized BCI control strategies"
Journal of Neural Engineering, 2020

Correlates of mental states ("thoughts")

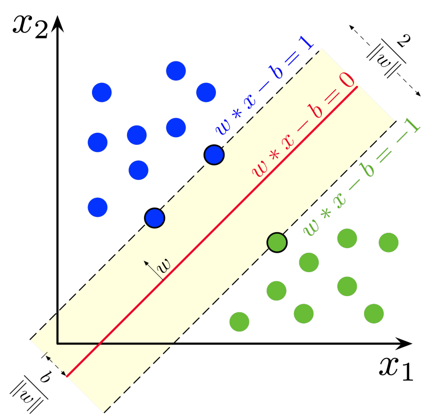


Self-Organizing Maps: **topology** of the samples

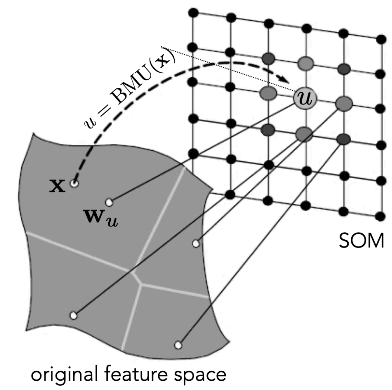
http://news.mit.edu/2014/optogenetic-toolkit-goes-multicolor-0209

Interpretability adds a new axis for algorithm selection

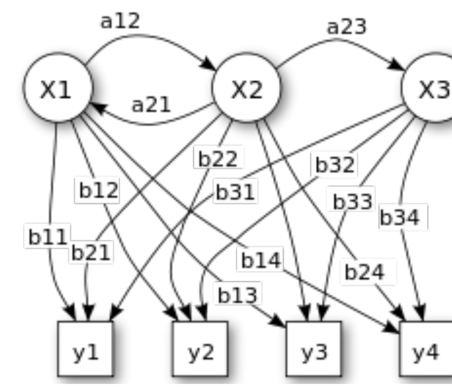
Different machine learning algorithms capture knowledge
into different representations



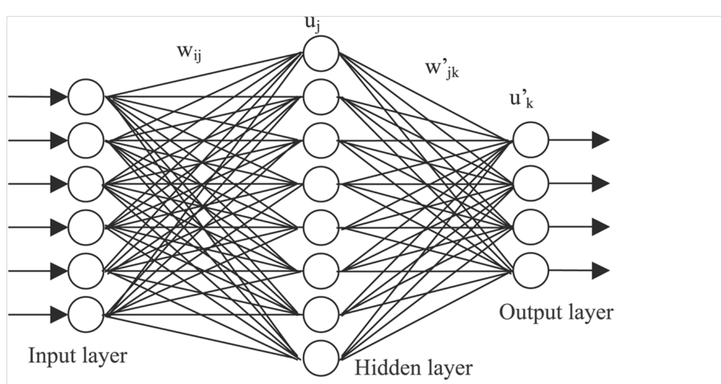
Support Vector Machines:
points in the feature space



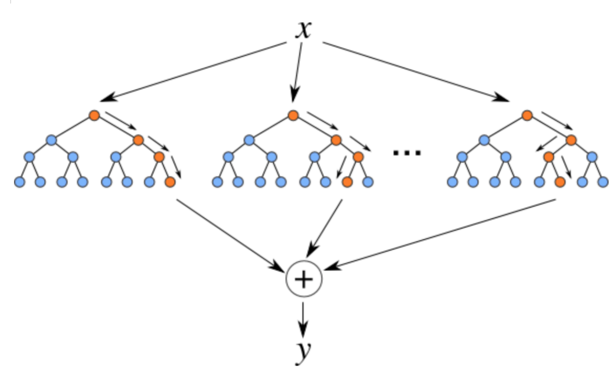
Self-Organizing Maps:
topology of the samples



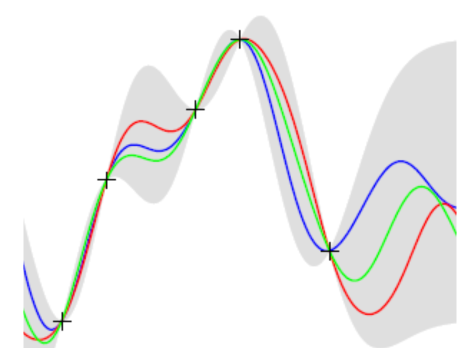
Hidden Markov Models:
states and transitions



DNNs: **distributed representations**
over features (input or latent)



Random Forests:
feature-based **rules**

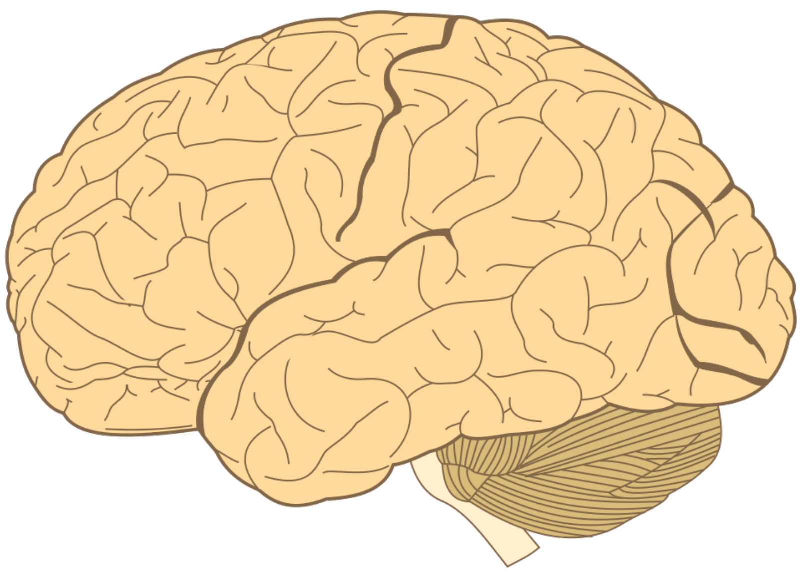


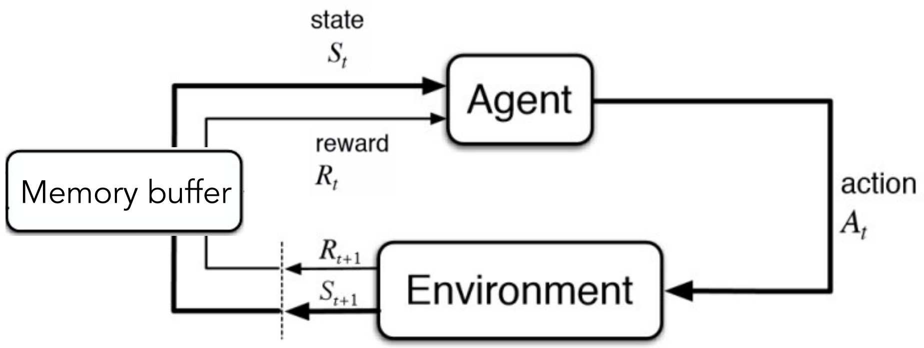
Gaussian Processes:
functions

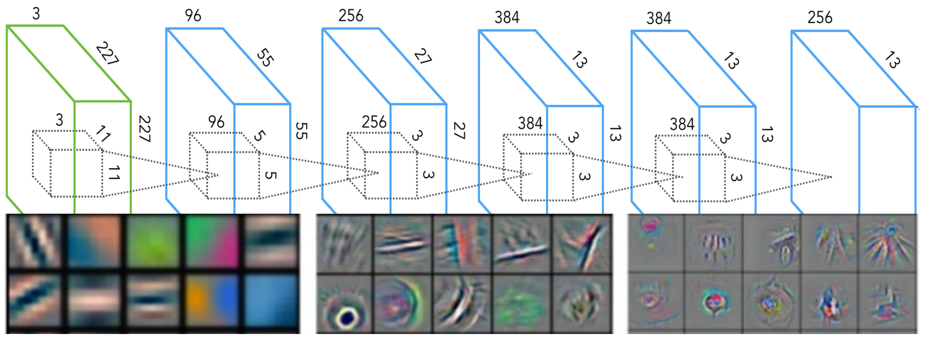
Pick the one that will reveal the knowledge you are after,
not the one that just gives the best performance on a metric.

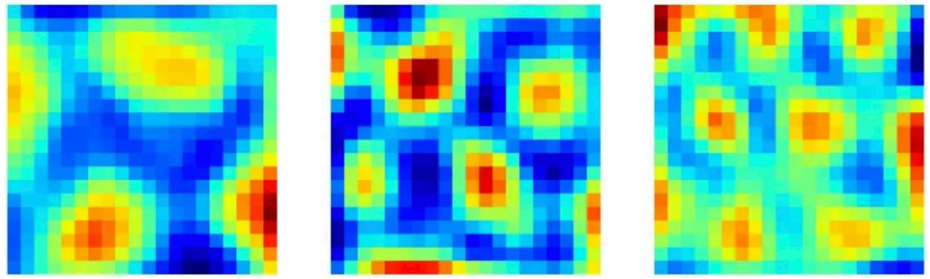
https://www.economind.org/single-post/2017/11/23/Is-Economic-Modelling-a-futile-attempt-before-the-everlasting-Uncertainty https://blog.toadworld.com/2018/08/31/random-forest-machine-learning-in-r-python-and-sql-part-1
https://math.stackexchange.com/tags/neural-networks/info https://www.superdatascience.com/blogs/the-ultimate-guide-to-self-organizing-maps-soms

Curiously similar mechanisms in biological and artificial systems

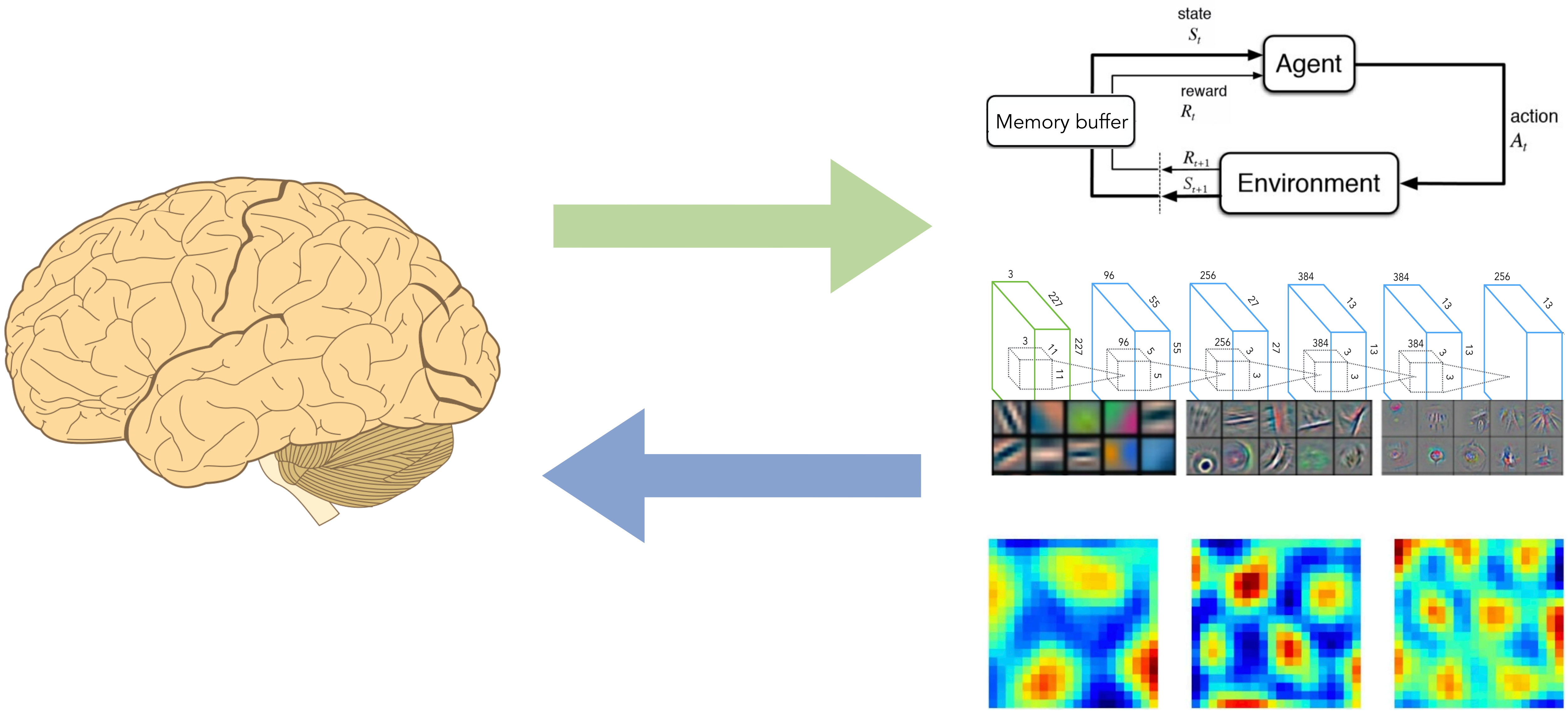








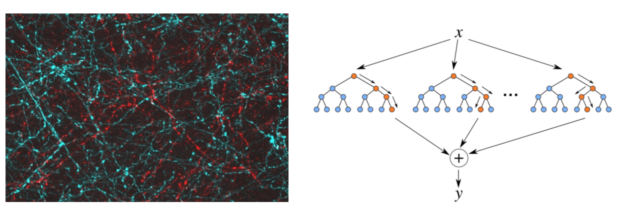
Curiously similar mechanisms in biological and artificial systems



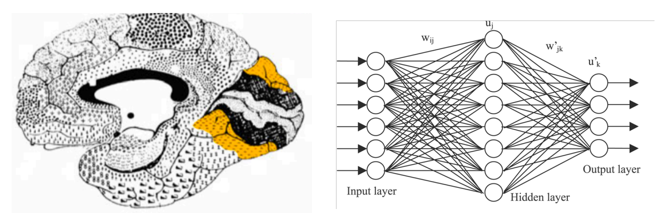
Interpreting the mechanisms of machine learning models
can shed light on the mechanisms of the brain

Discussion

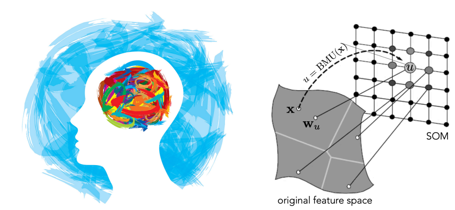
- Modeling is a well-proven way of obtaining knowledge
- Machine-learned models do capture the knowledge, but an additional step of interpretation is required
- In life sciences model interpretation has a special significance
- Three examples of applying this principle in Neuroscience



"Identifying task-relevant spectral signatures of perceptual categorization in the human cortex"
Scientific Reports, 2020



"Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex"
Communications Biology, 2018



"Mental state space visualization for interactive modeling of personalized BCI control strategies"
Journal of Neural Engineering, 2020

- Interpretability adds a new axis for algorithm selection
- Interpreting the mechanisms of machine learning models can shed light on the mechanisms of the brain

This would not be possible without

the support and knowledge given by the **University of Tartu** *and the* **Institute of Computer Science**,

Raul Vicente *establishing the Computational Neuroscience Lab and supervising my PhD studies*,

Konstantin Tretyakov *introducing me to the field of machine learning,*

Sven Laur *further developing my understanding of machine learning,*

the work done by **Juan R. Vidal** *and the co-authors from* **Lyon Neuroscience Research Center**,

the constant flow of ideas born at the seminars, lunch breaks, discussions with **Anna Leontjeva, Tambet Matiisen, Jaan Aru, Ardi Tampuu**, **Kristjan Korjus** and all lab members and alumni, students, and co-authors,

my **parents** and family, *who showed me the value of knowledge and provided with the opportunity to pursue it.*

Thank you!

Neuroscience

Aims to understand learning systems and intelligence by analyzing and reverse engineering the existing example

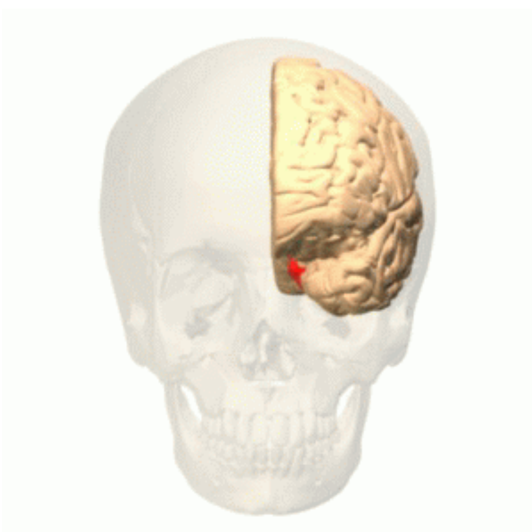
a special kind of synergy that leads to **curious similarities**

Aims to build an intelligent system ground-up by figuring out the building blocks and rules of interactions between them

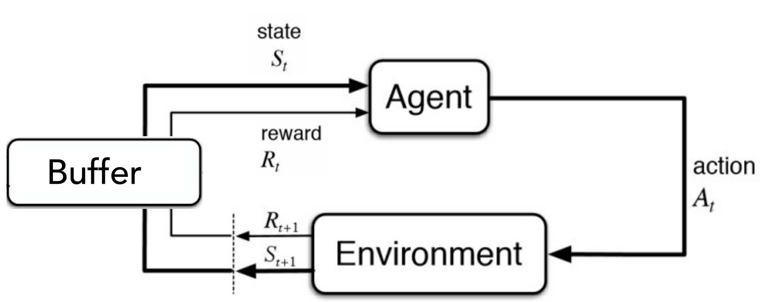
Machine Learning & AI

Curiously similar mechanisms in biological and artificial systems

Experience replay



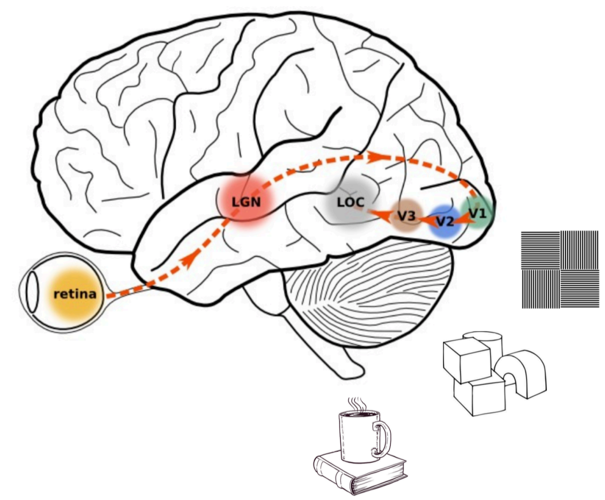
Hippocampus



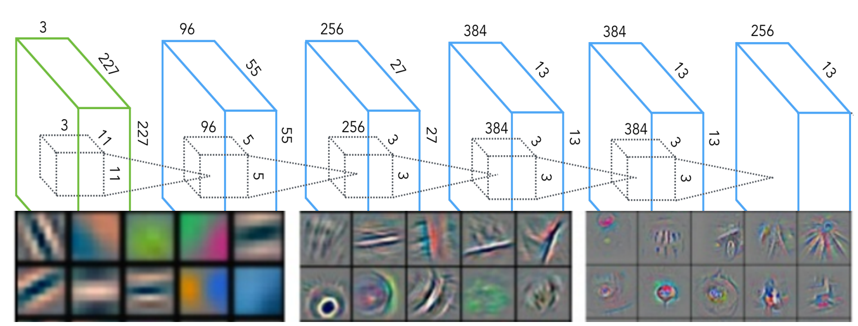
Deep Q-Learning

Memory consolidation ⇄ Experience replay

Hierarchy of visual layers



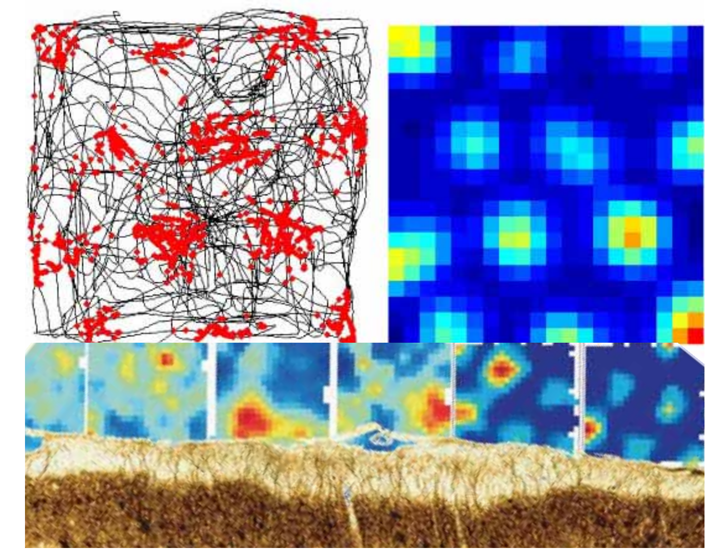
Layers of visual cortex



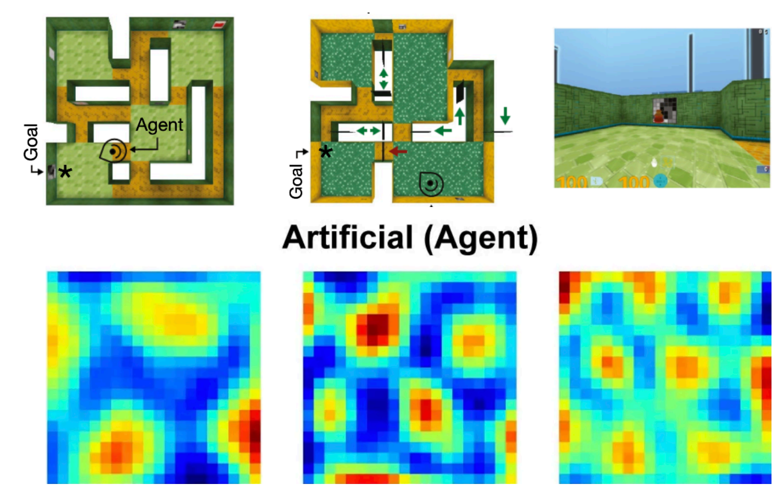
Deep convolutional neural network

- Layered structure
- Hierarchy of representational complexity
- Receptive fields ⇄ convolutional filters

An efficient spacial code



Grid cells in entorhinal cortex



Self-emergent spacial code

Grid-based code for location