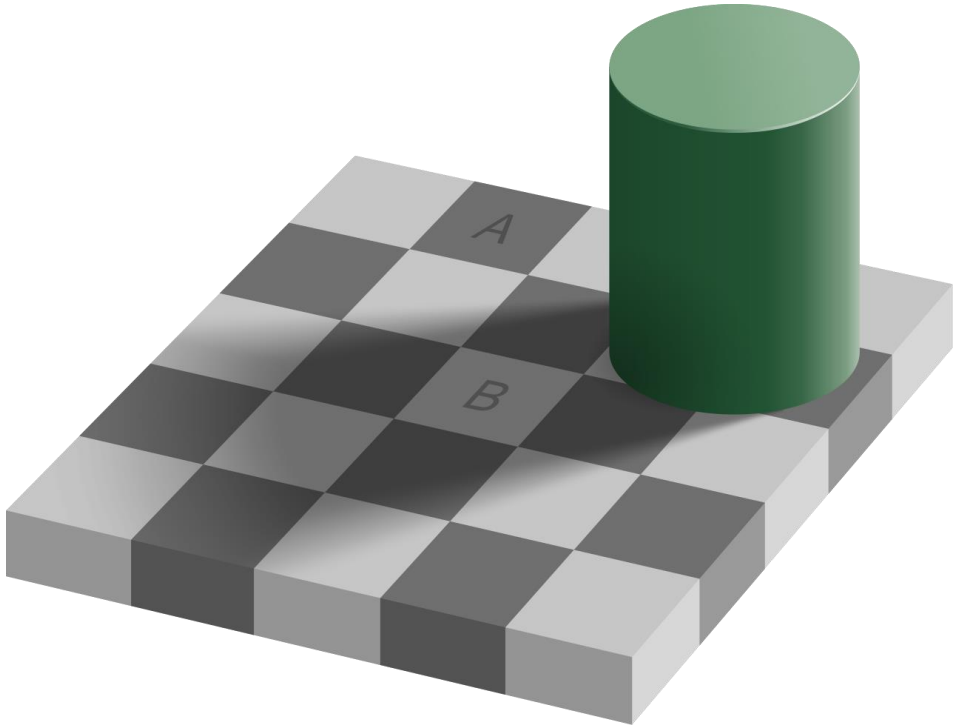
Machine Learning and Neuroscience – part 2

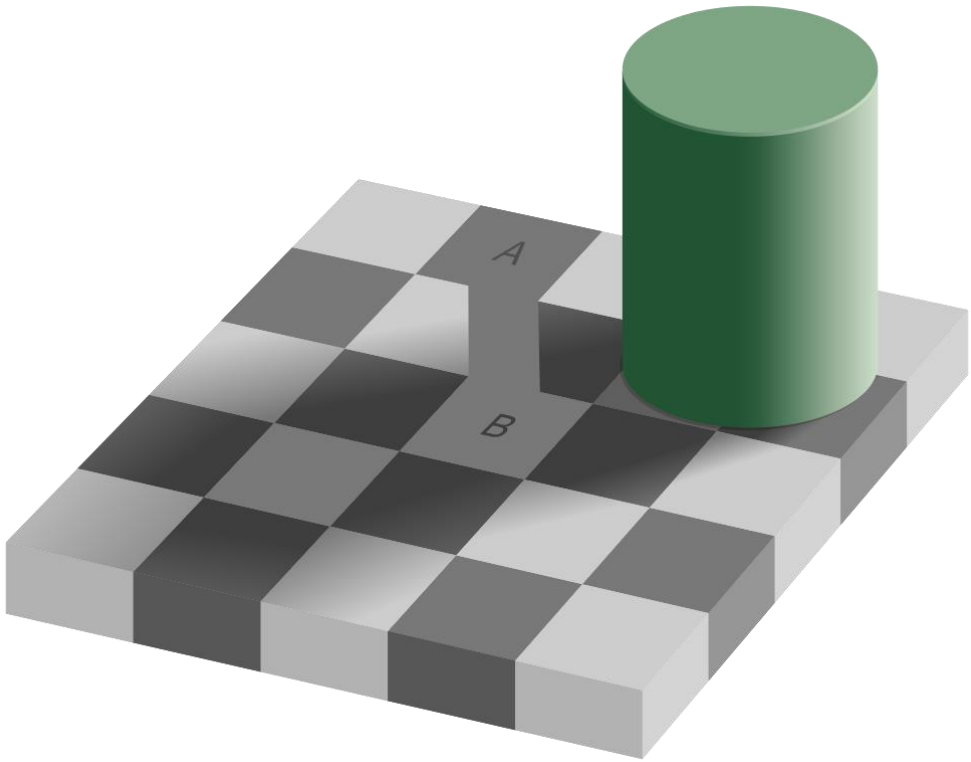
Michel Besserve

MPI for Intelligent Systems - MPI for Biological Cybernetics,

Tübingen, Germany.

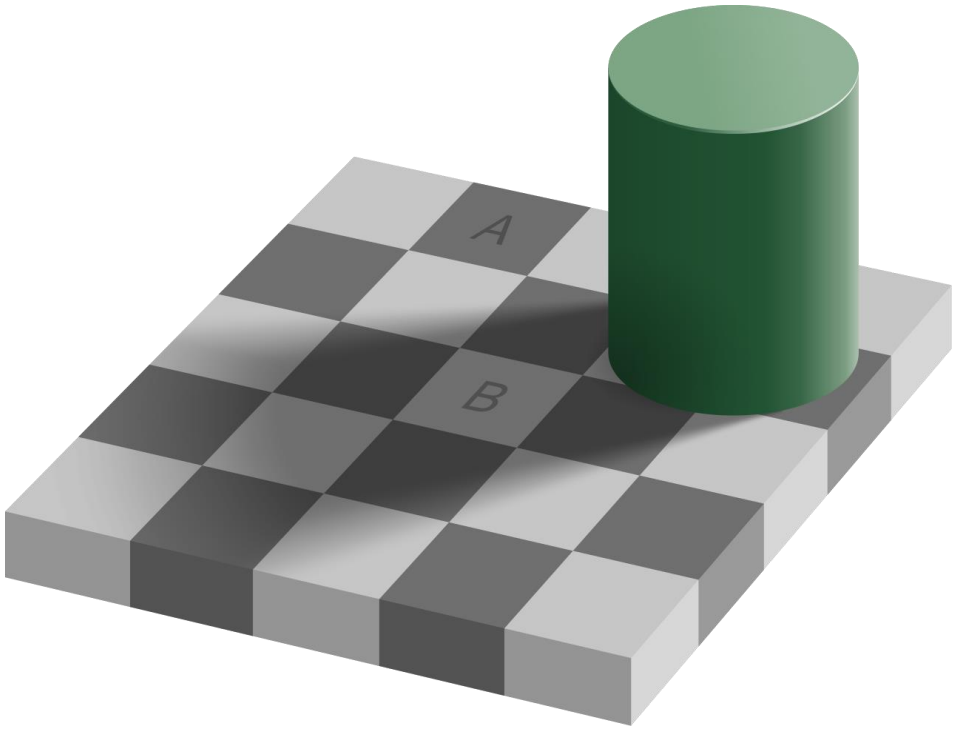
The generative nature of perception





From “Checker shadow illusion”, Wikipedia.

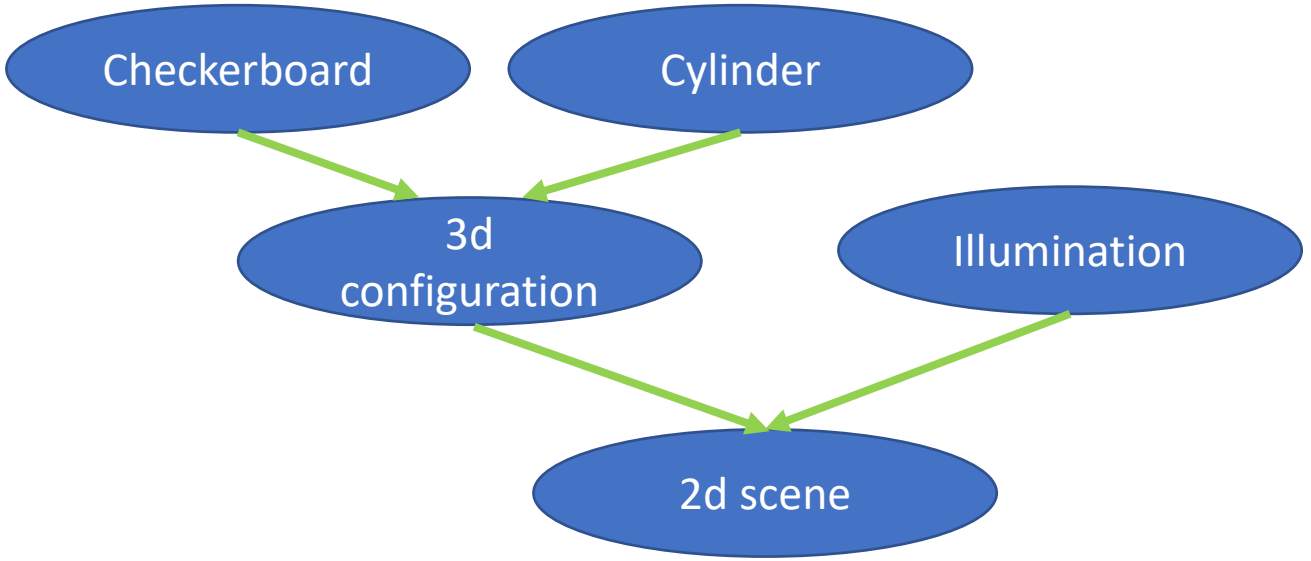
The generative nature of perception



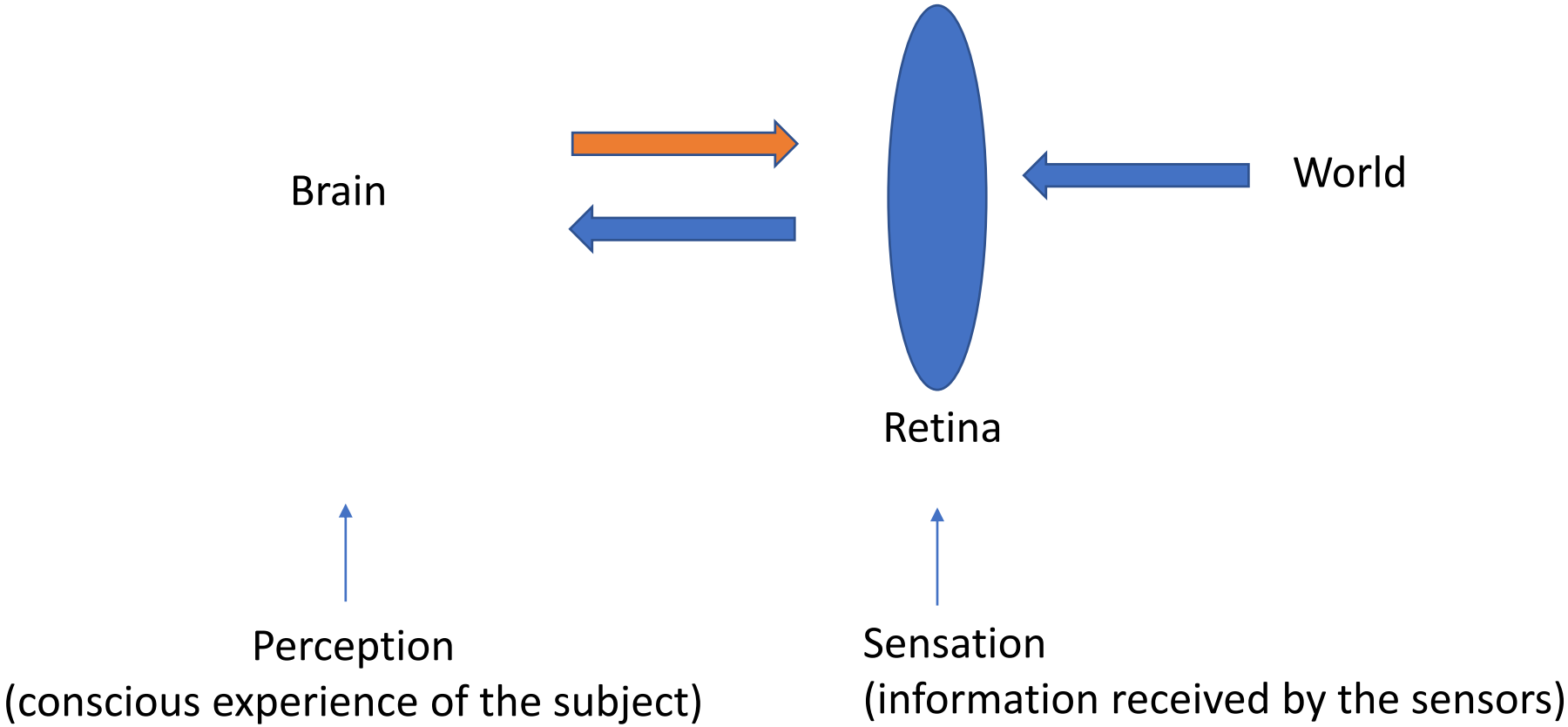
From “Checker shadow illusion”, Wikipedia.

Interpretation 1:
our brain “compensates” changes in illumination.

Interpretation 2:
our perception incorporates an immediately accessible *causal* representation of how sensory data is generated.



The basic idea behind predictive coding



Neural implementation by feedback connections
(Kawato et al., 1993)

A forward-inverse optics model of reciprocal connections between visual cortical areas

Mitsuo Kawato, Hideki Hayakawa & Toshio Inui

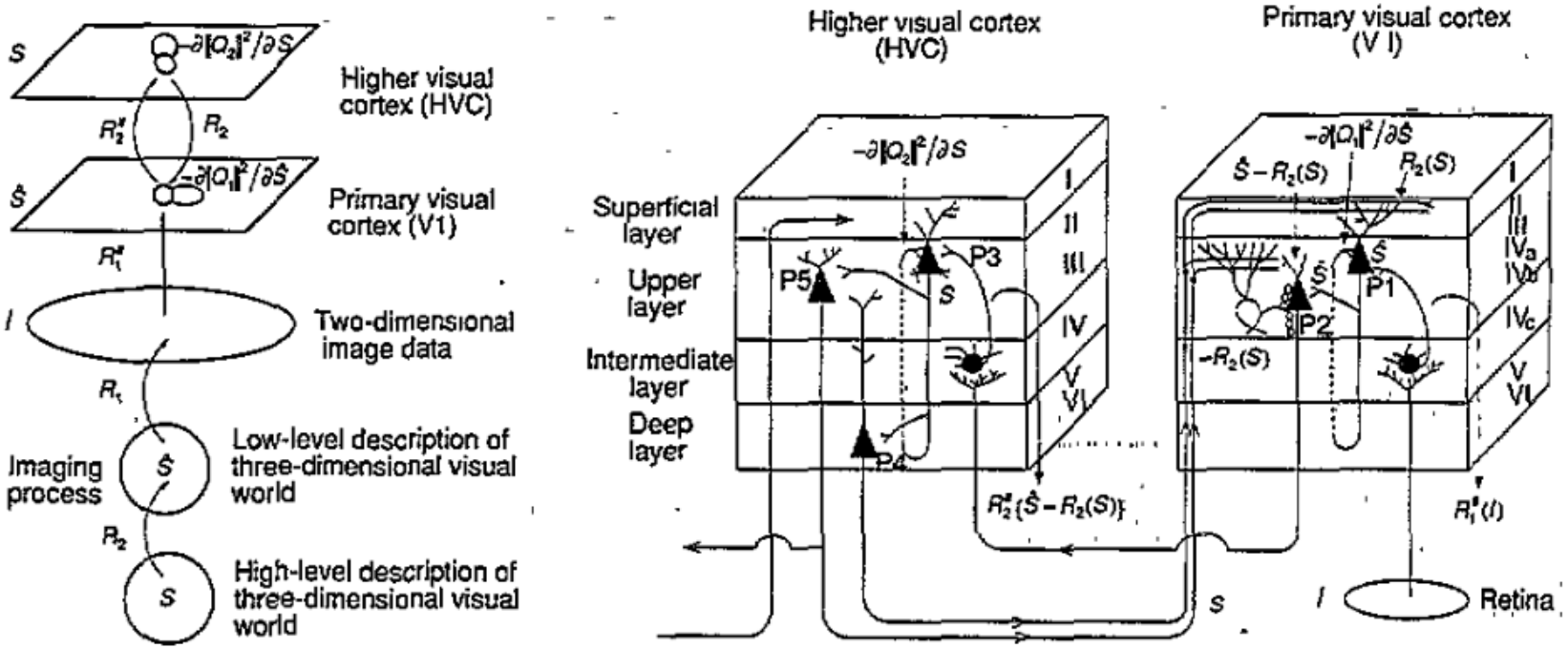
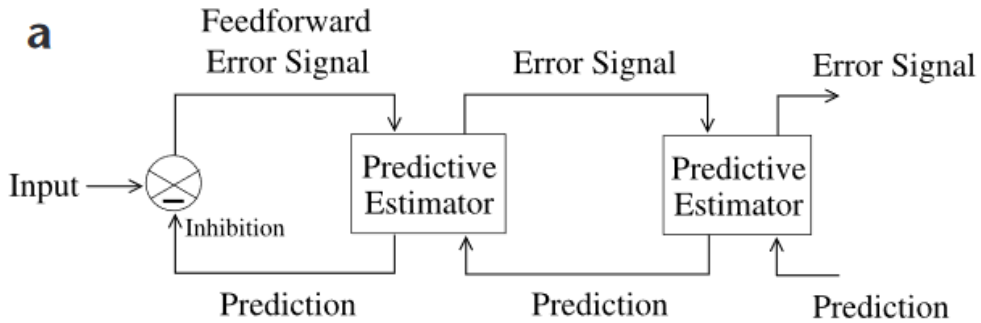


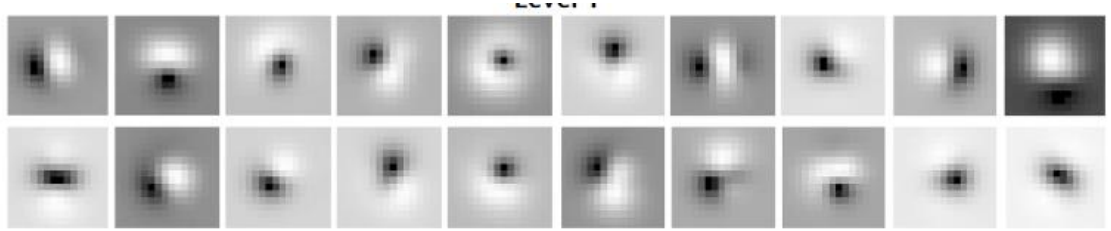
Figure 1. Fundamental forward-inverse optics model. (A) Model for reciprocal interactions between V1 and the higher visual cortex (HVC). In the lower half of the figure, the optics operation R in the outer world is decomposed into a lower and a higher part, R_1 and R_2. A model of this hierarchy in the brain is shown in the upper half of the figure. (B) Layered-neural-circuit model of the hierarchical interaction between V1 and HVC. Filled neurons are excitatory and a hollow neuron is inhibitory.

nature *neuroscience* • volume 2 no 1 • january 1999

Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects

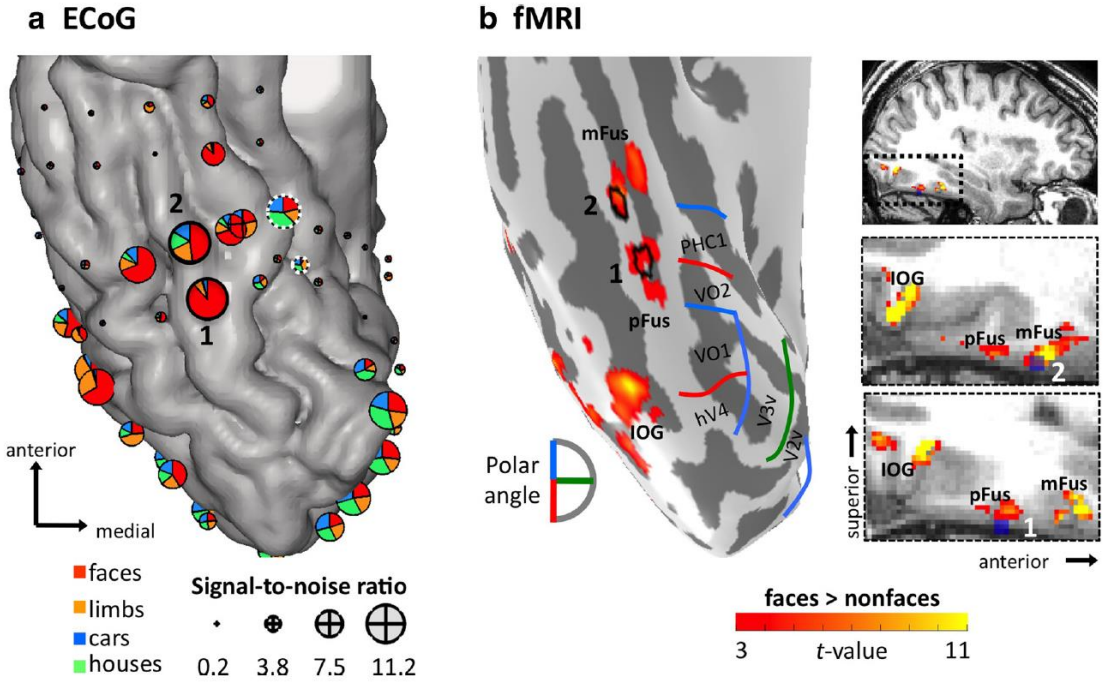
Rajesh P. N. Rao[1] and Dana H. Ballard[2]





Perceptual counterfactuals (Parvizi et al., 2012)

Electrical Stimulation of Human Fusiform Face-Selective Regions Distorts Face Perception



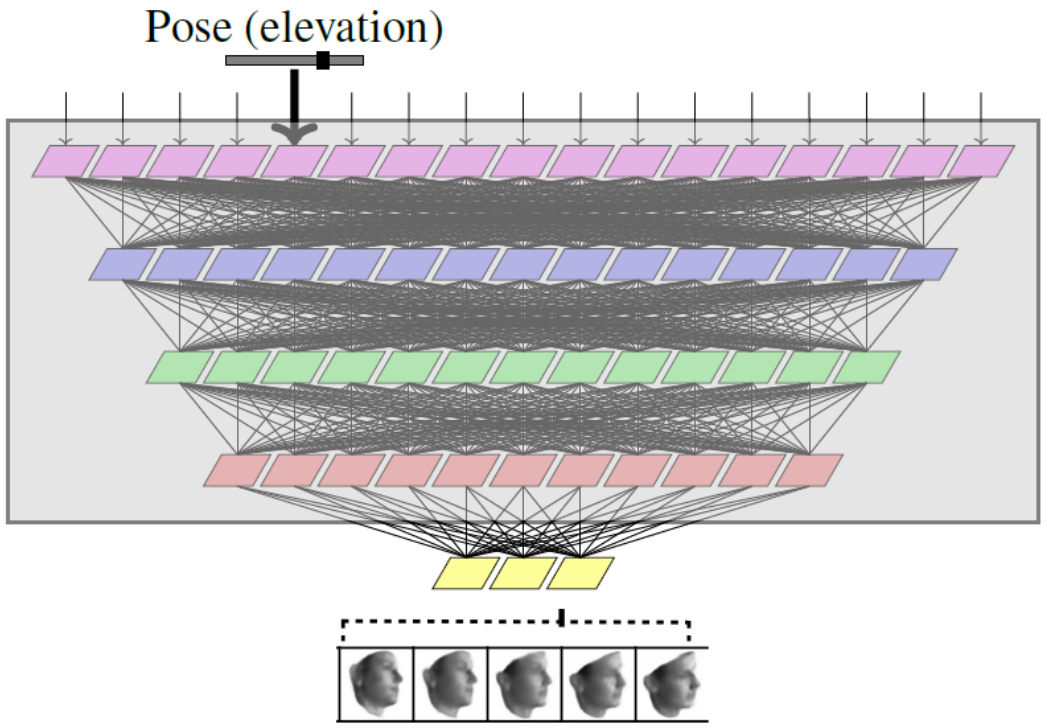
https://www.youtube.com/watch?v=R2AMgLhaxHk

Implication for artificial intelligence

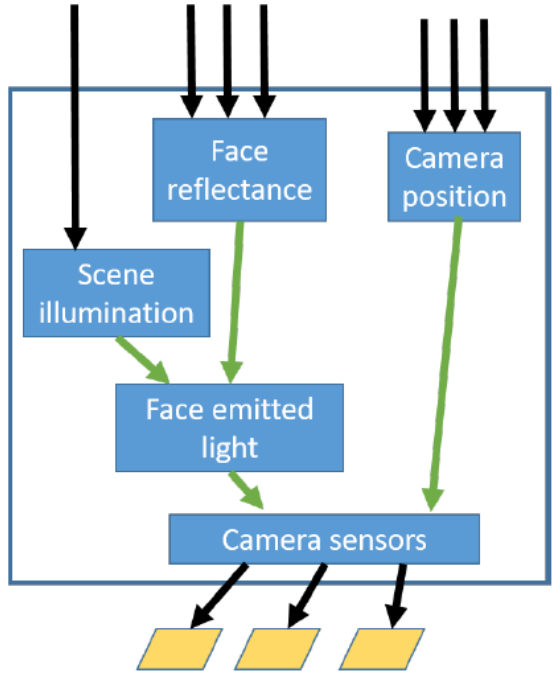
- Causal forward optics models may be easier to fit and manipulate than their anti-causal inverse.
- The generative part auto-encoder architectures may be exploited by agents to better generalize and “human-like” reasoning about their environment.

Counterfactuals in deep generative models

(Besserve et al. 2018)

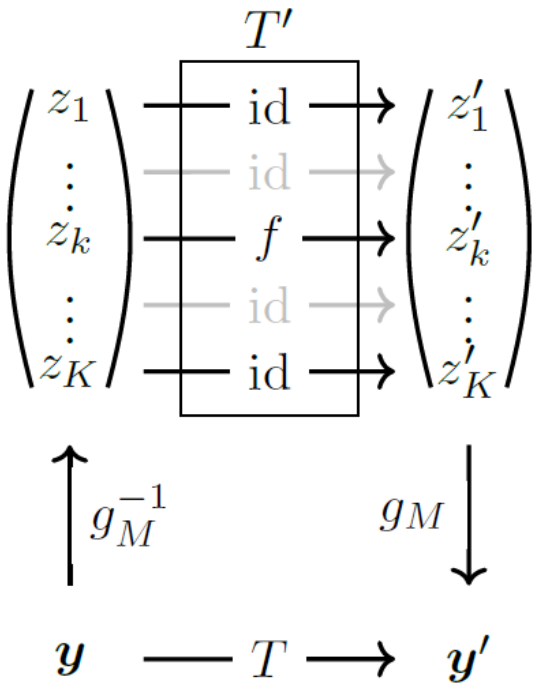


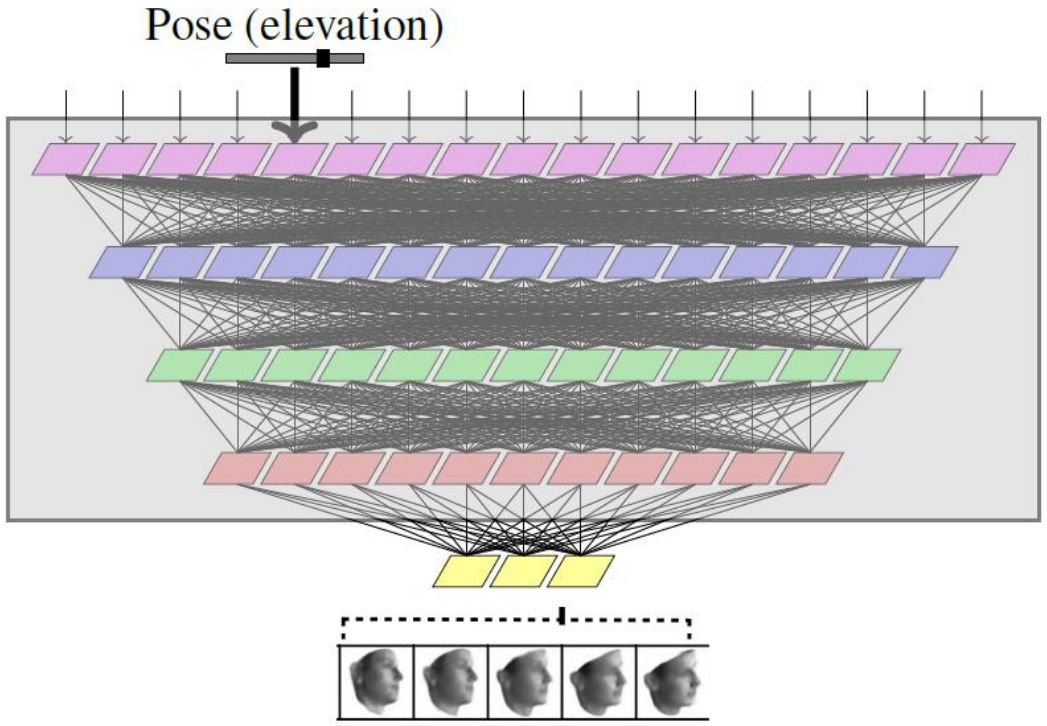
Current approach to disentangled generative models (e.g. Kulkarni et al. (2015)).



Idealized modular generative model.

Principle of a disentangled representations





Current approach to disentangled generative models (e.g. Kulkarni et al. (2015)).

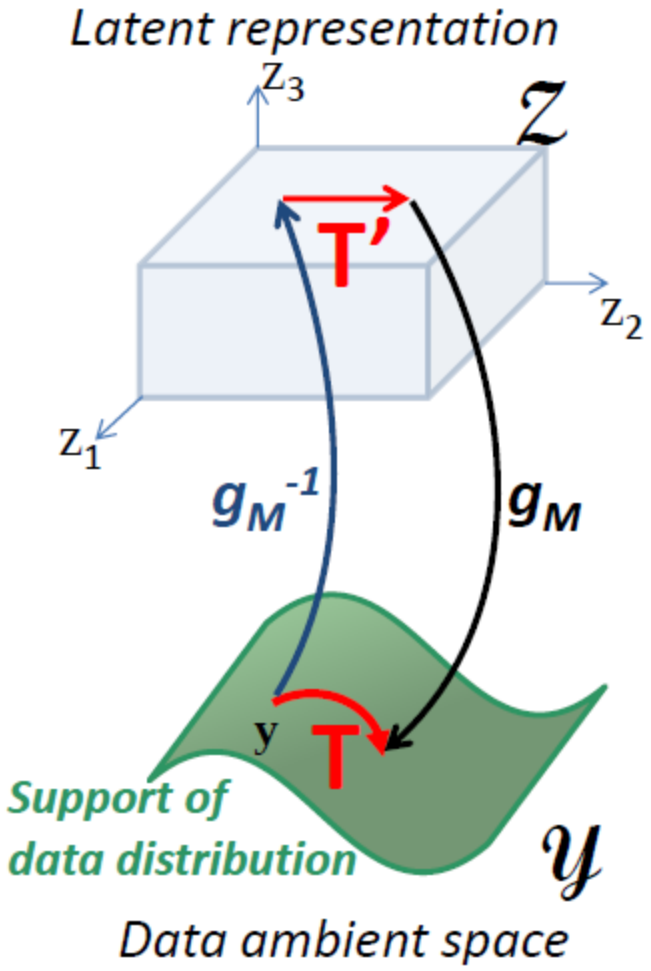
Causal description of the sytem

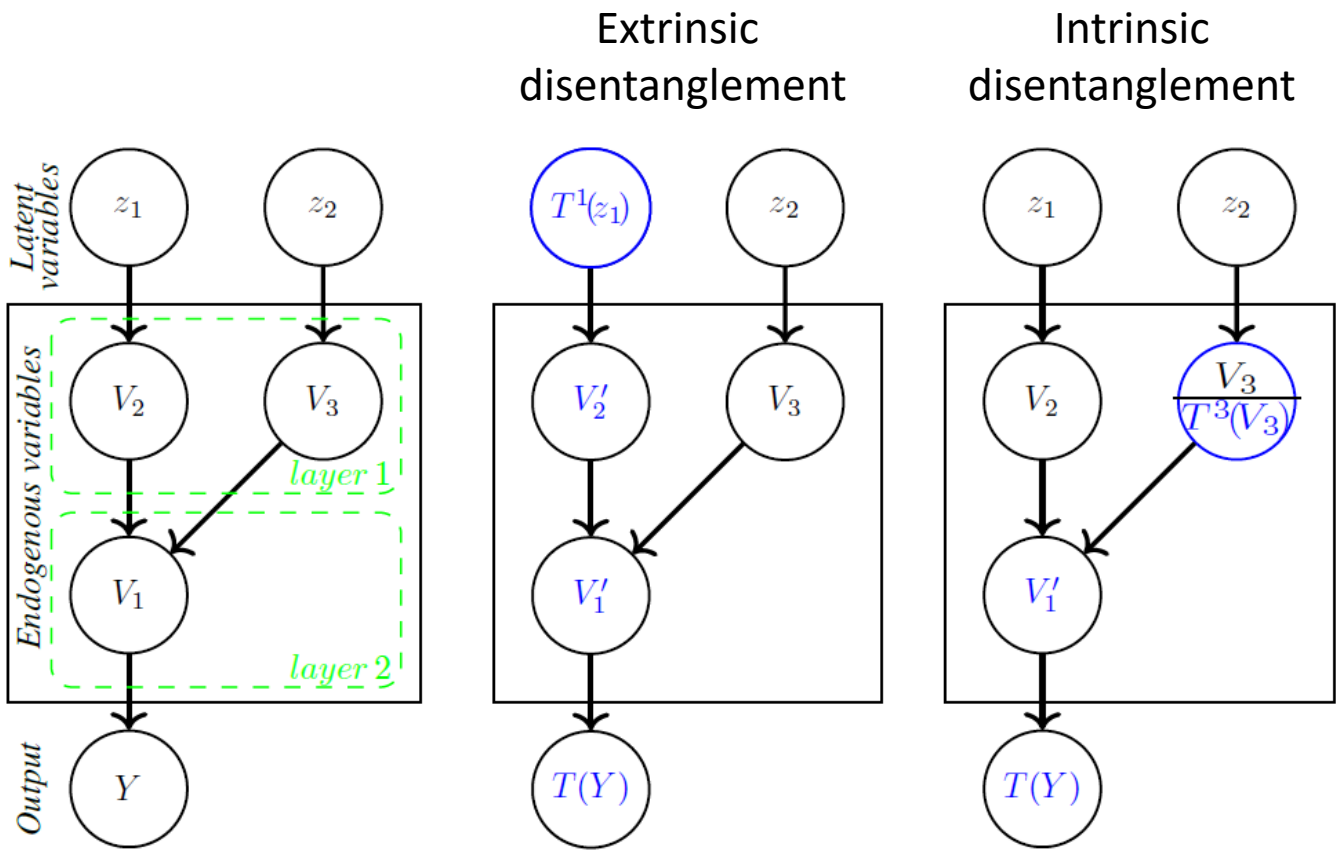
Structural equations

$$X_n := f(X_1, X_2, \cdots, X_k)$$

- More than an equality linking the observations of a system.
- Variables are not necessarily random.
- It reflects "stability" of the relation: to interventions (hard and soft), to counterfactuals.

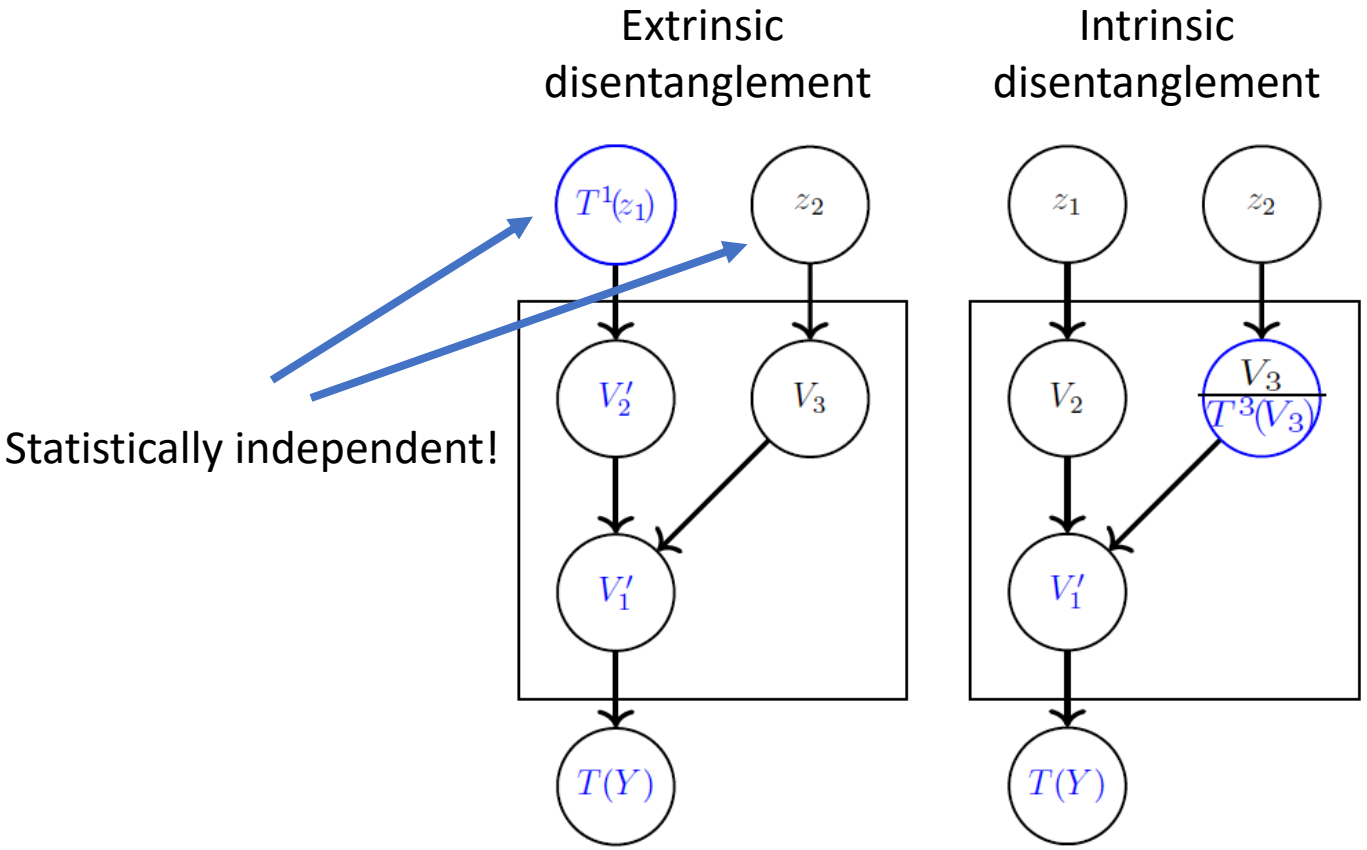
Summary of the framework





Besserve et al., 2018

Intrinsic disentanglement is more relevant

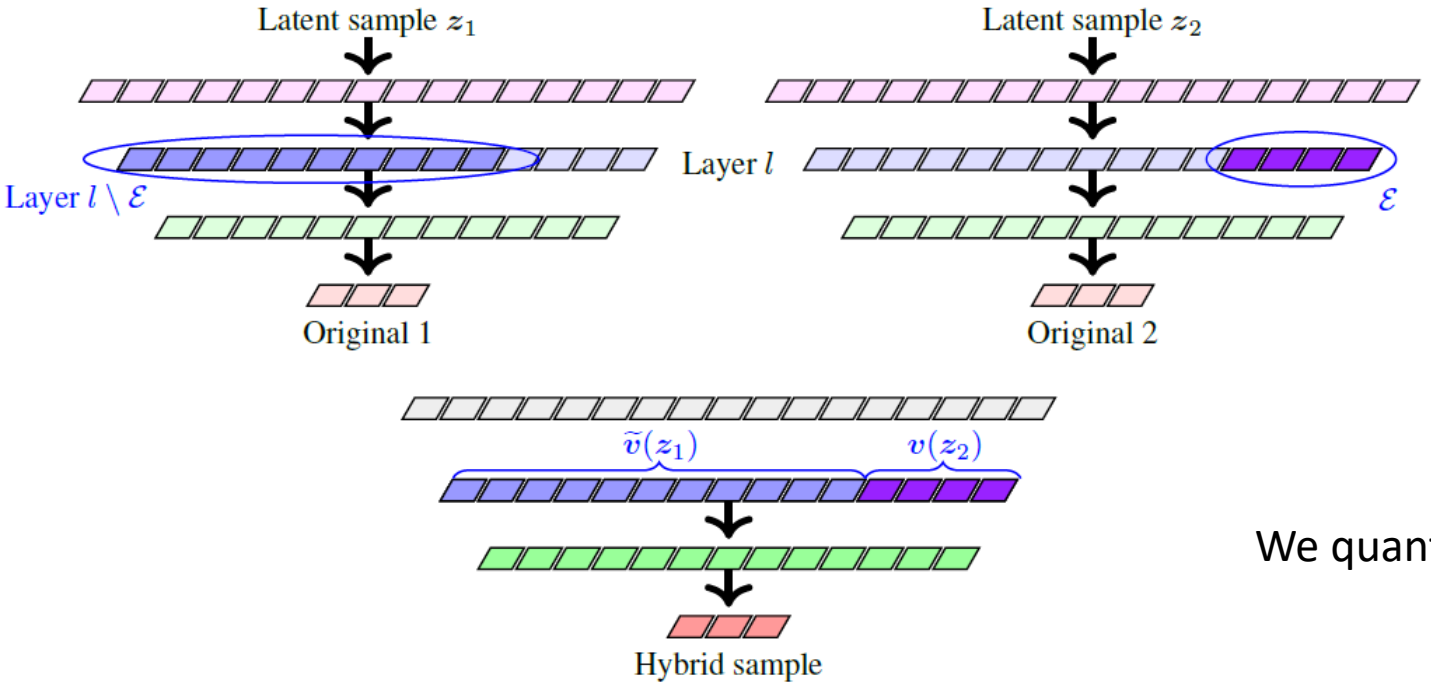


Most properties that we would like to disentangle are NOT statistically independent:

- Hair color and skin color,
- Object and background!

Hybridization

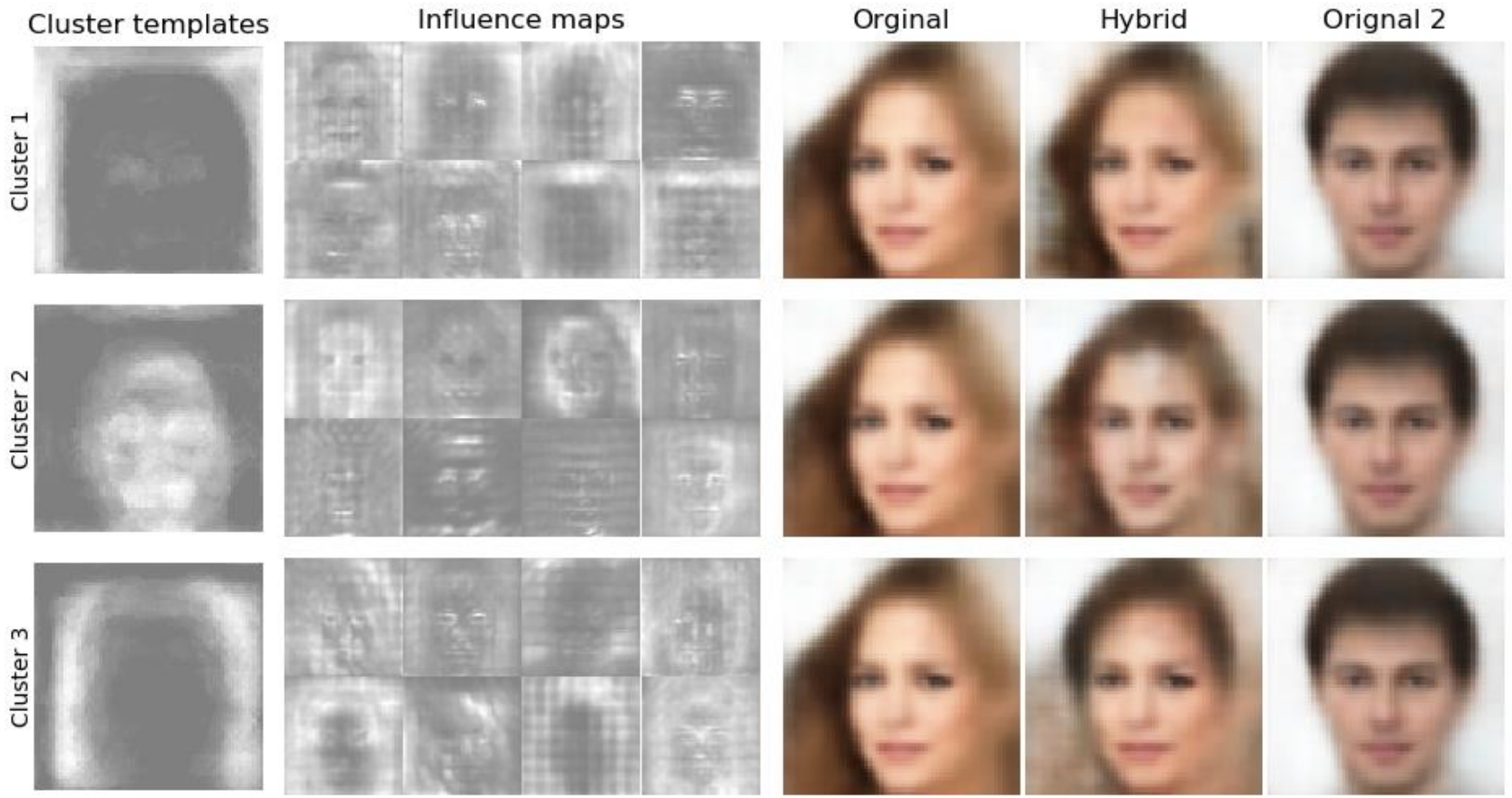
Idea: to look for disentangled transformations,
we can look for “faithful” counterfactuals,
i.e. counterfactuals that do not leave the data manifold.



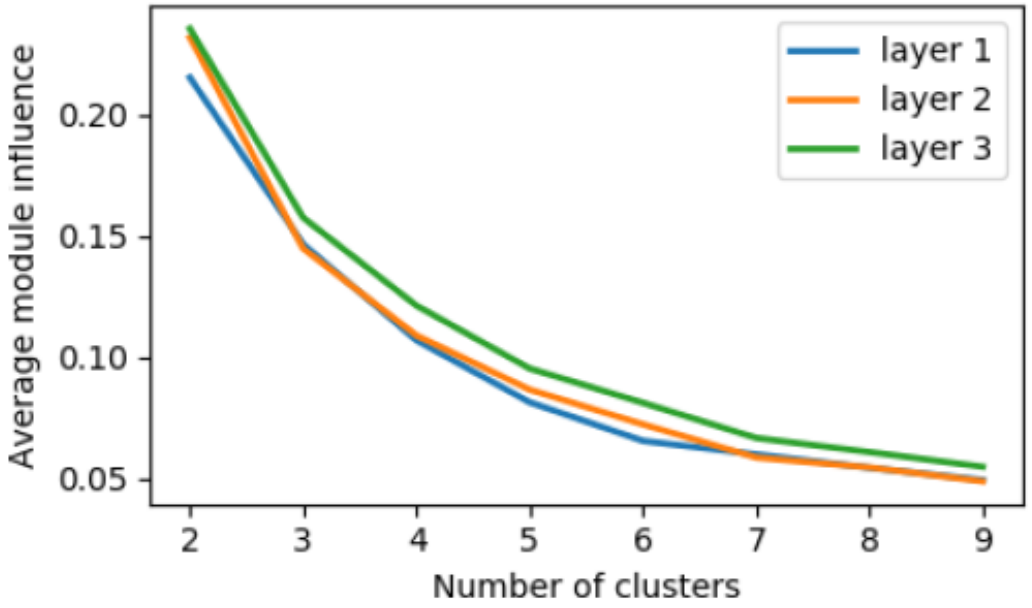
We quantify *average causal effects* trough influence maps

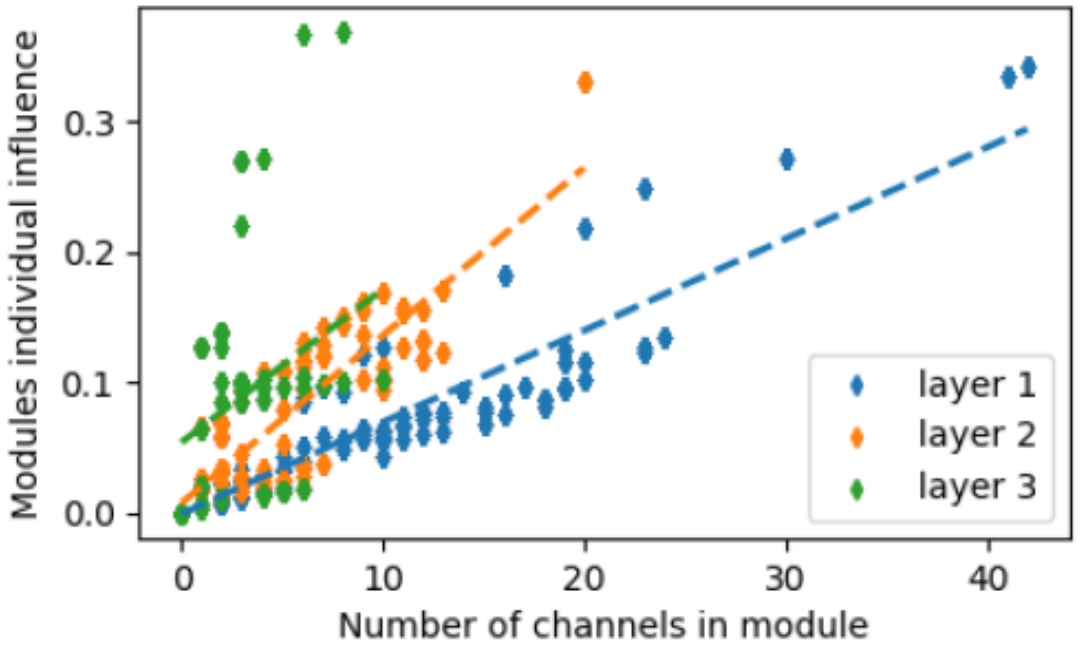
$$IM(\mathcal{E}) = \mathbb{E}_{\boldsymbol{z}_2 \sim P(\mathbf{Z})}\left[\mathbb{E}_{\boldsymbol{z}_1 \sim P(\mathbf{Z})}\left[\left|Y^{\mathcal{E}}_{v(\boldsymbol{z}_2)}(\boldsymbol{z}_1) - Y(\boldsymbol{z}_1)\right|\right]\right]$$

Clustered influence maps (VAE)

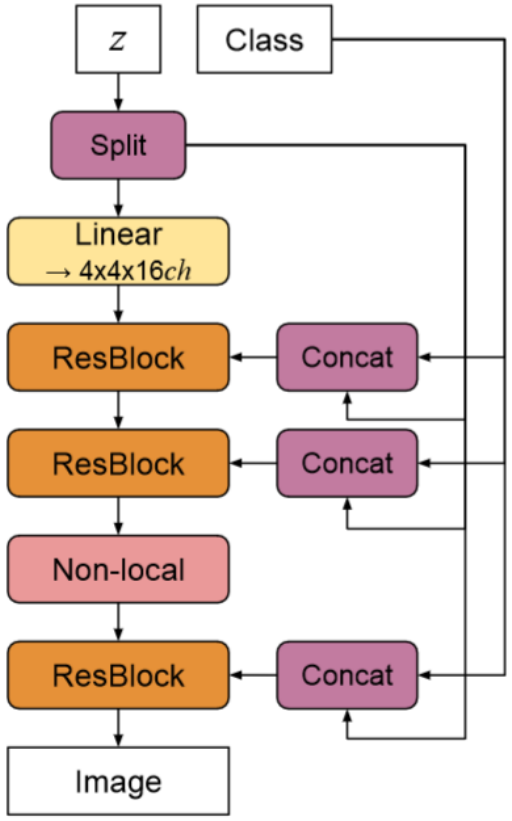


Influence of cluster size

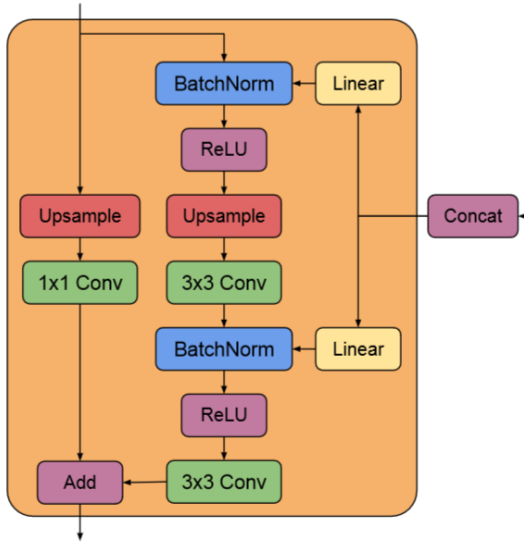




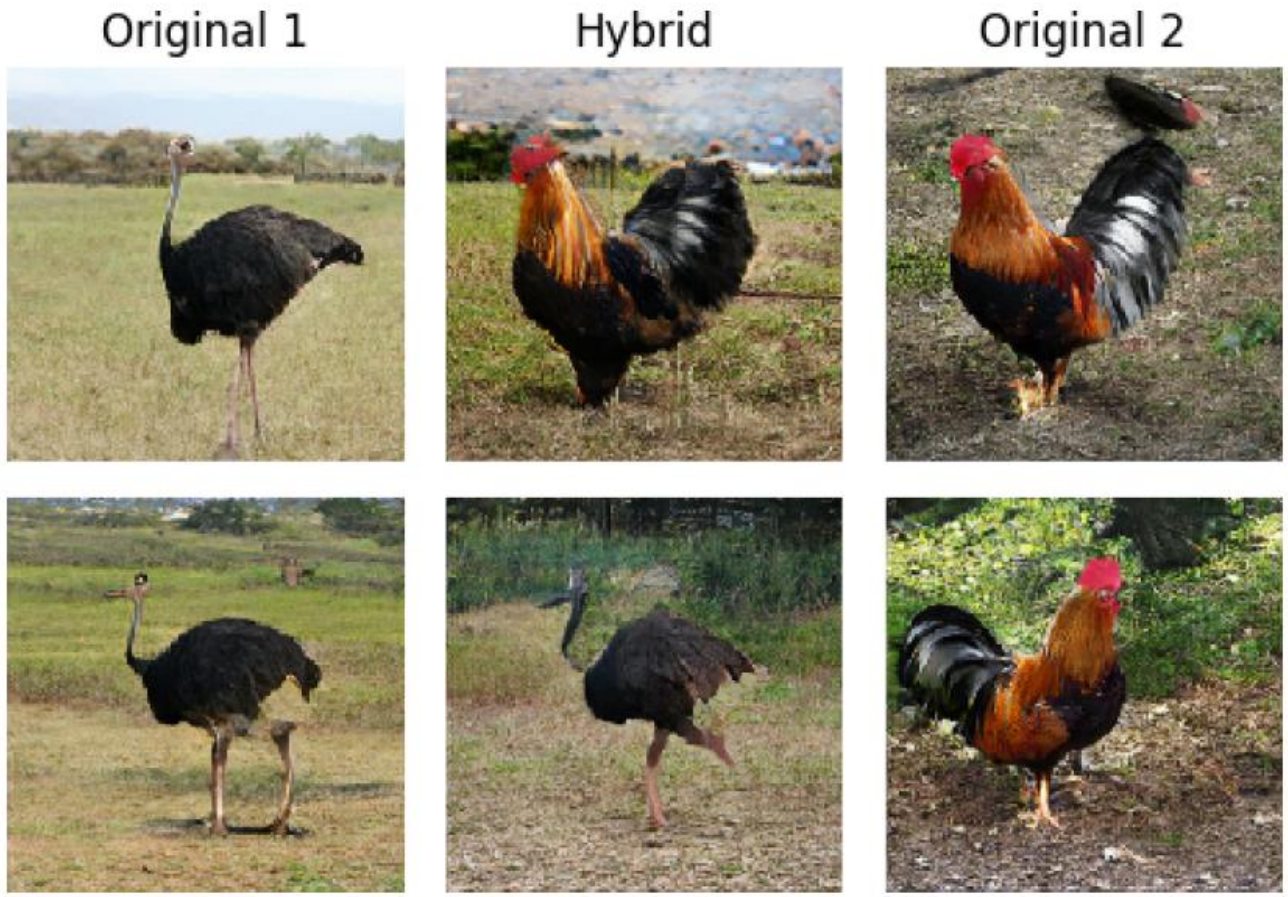
Using BIGGan on ImageNet



(a)



(b)

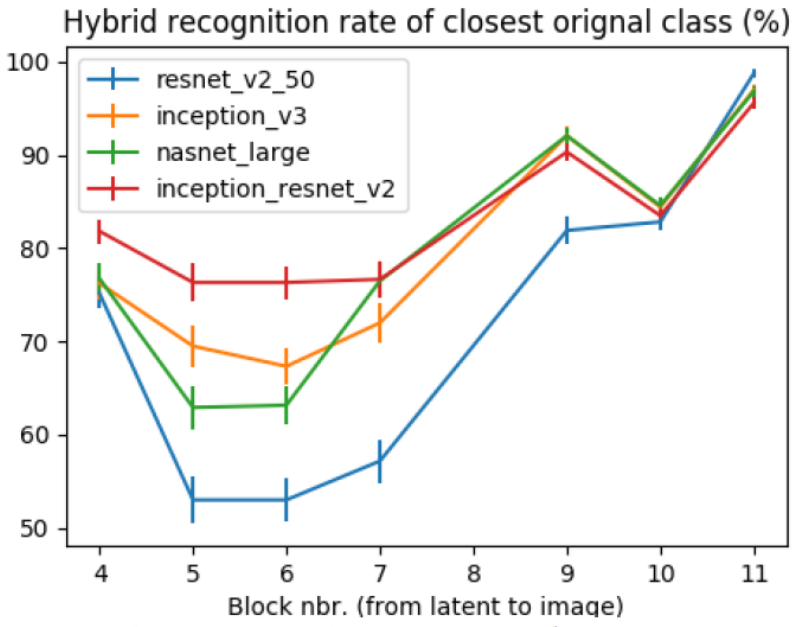


Besserve et al., submitted.

Using BIGGan on ImageNet

Besserve et al., submitted.

Testing robustness of classifiers



Besserve et al., submitted.

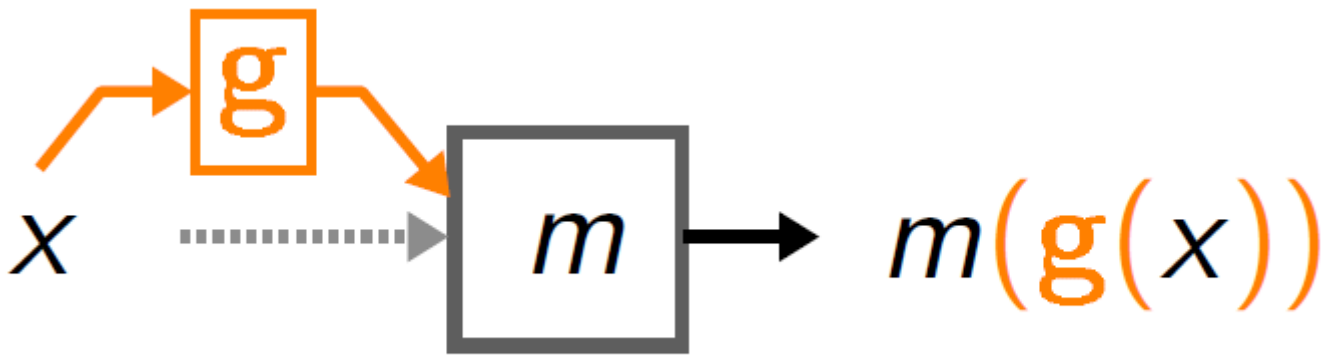
Example of ambiguous pictures

Left Image	Model Output	resnet_v2_50 koala	inception_v2 koala	nasnet_large koala	inception_resent_v2 koala
Middle Image	Model Output	resnet_v2_50 koala	inception_v3 koala	nasnet_large teddy	inception_resnet_v2 teddy
Right Image	Model Output	resnet_v2_50 koala	inception_v3 teddy	nasnet_large teddy	inception_resnet_v2 koala

Figure 12: Three koala+teddy hybrids as the inputs to the classifiers of Table. 3

Besserve et al., submitted.

Group invariance principles for causal generative models (Besserve et al. 2018)



Structural equations

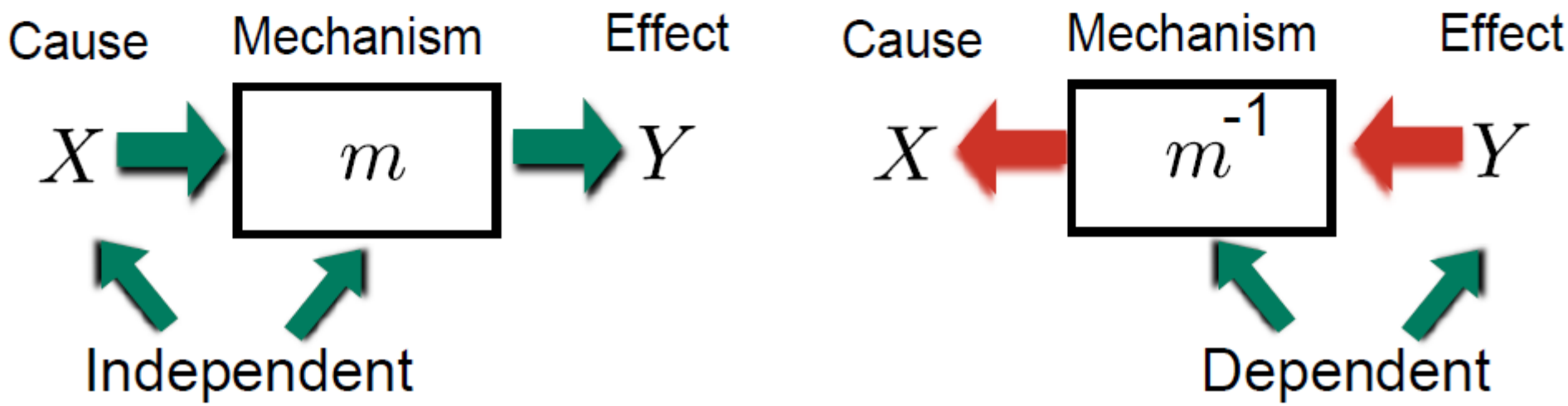
$$X_n := f(X_1, X_2, \cdots, X_k)$$

- More than an equality linking the observations of a system.
- Variables are not necessarily random.
- It reflects "stability" of the relation: to interventions (hard and soft), to counterfactuals.

Independence of cause and mechanism (ICM)

- Postulate: Nature chooses the cause independent from the mechanism
  [Daniusis et al., 2010, Janzing and Schölkopf, 2010].
- Non-statistical independence, introduced using Kolmogorov complexity (non-computable!)

$$P(X) \perp\!\!\!\perp P(Y|X)$$



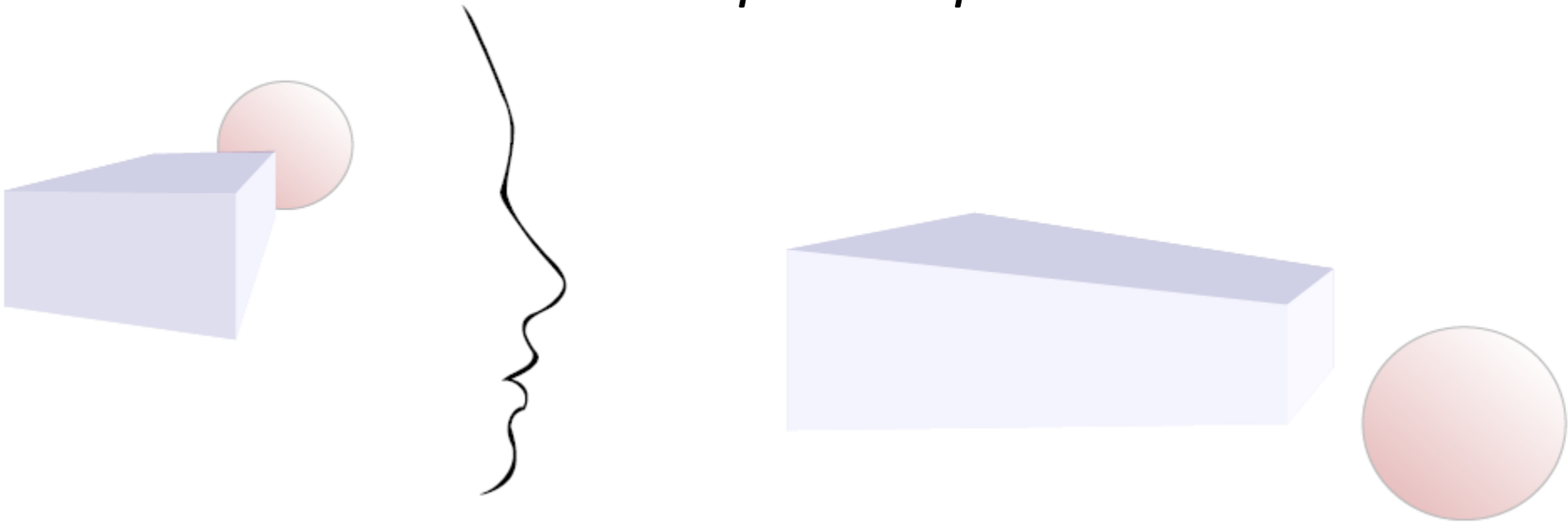
ICM: some observations

- Using various quantifications of "independence", several CI methods have been designed based on these principles [Janzing et al., 2010, Zscheischler et al., 2011, Janzing et al., 2012, Shajarisales et al., 2015, Sgouritsa et al., 2015].
- These methods exhibit nice theoretical properties with respect to the dimensionality of the data.
- No unified framework yet to develop methods using this principle.

ICM: beyond pairwise problems, generative models.

- Many data analysis procedures ultimately aim at drawing a causal interpretation: clustering, independent component analysis, which can all be considered as generative models.
- Causality may help enforcing additional constraints to select the right model.

Illustration with visual perception



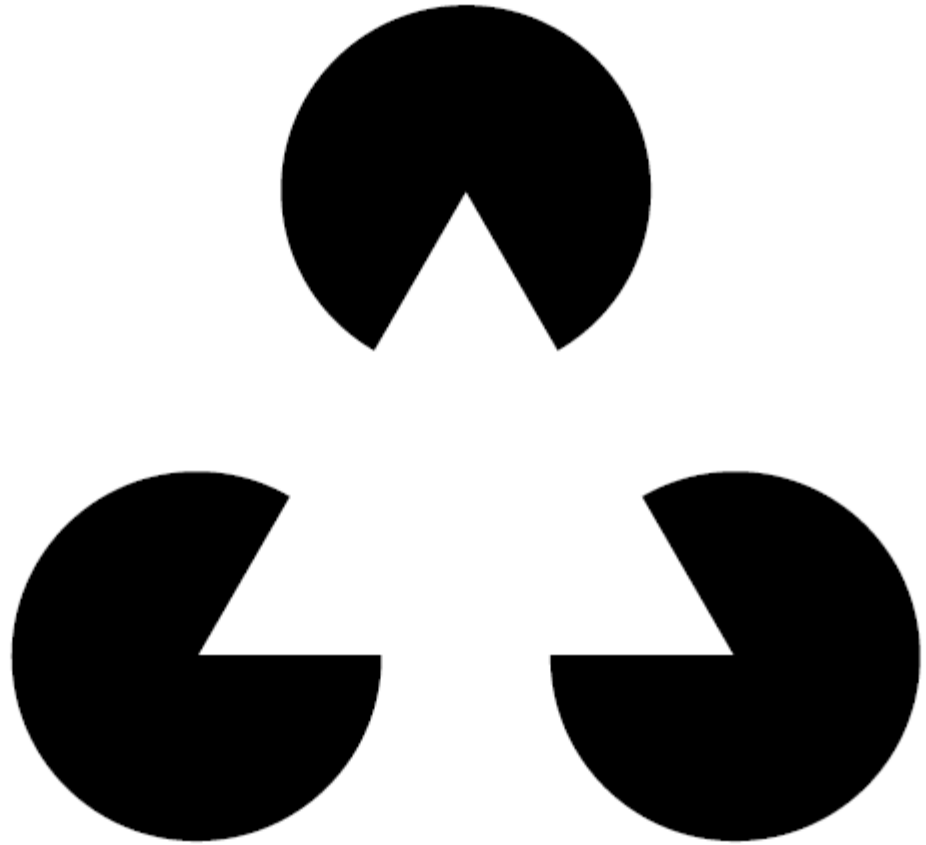
- Inferring the structure of 3D environment from a single 2D scene,
- Impossible to solve without additional assumptions,
- Example:*generic viewpoint assumption* [Freeman, 1994]

When assumptions are not met...

When assumptions are not met...

Slide courtesy of B. Schölkopf

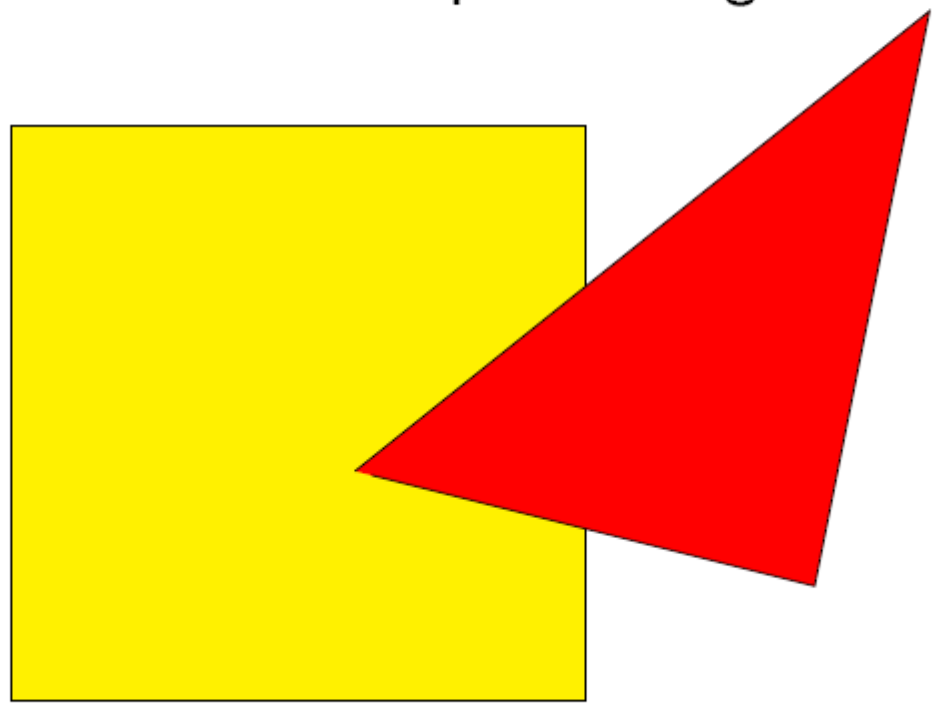
Kanizsa's triangles



- Illusory triangle on top of occluded disks.
- Alignment between segments is an important cue to group features belonging to the same objects.

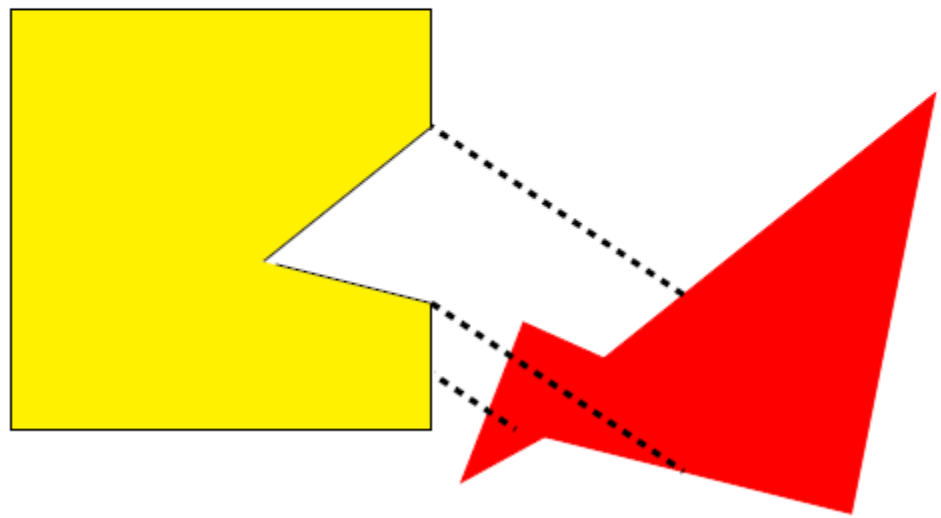
Object occlusion

We will illustrate how a causal inference framework based the principle of independence of cause and mechanism may solve the occlusion problem in a simple setting.

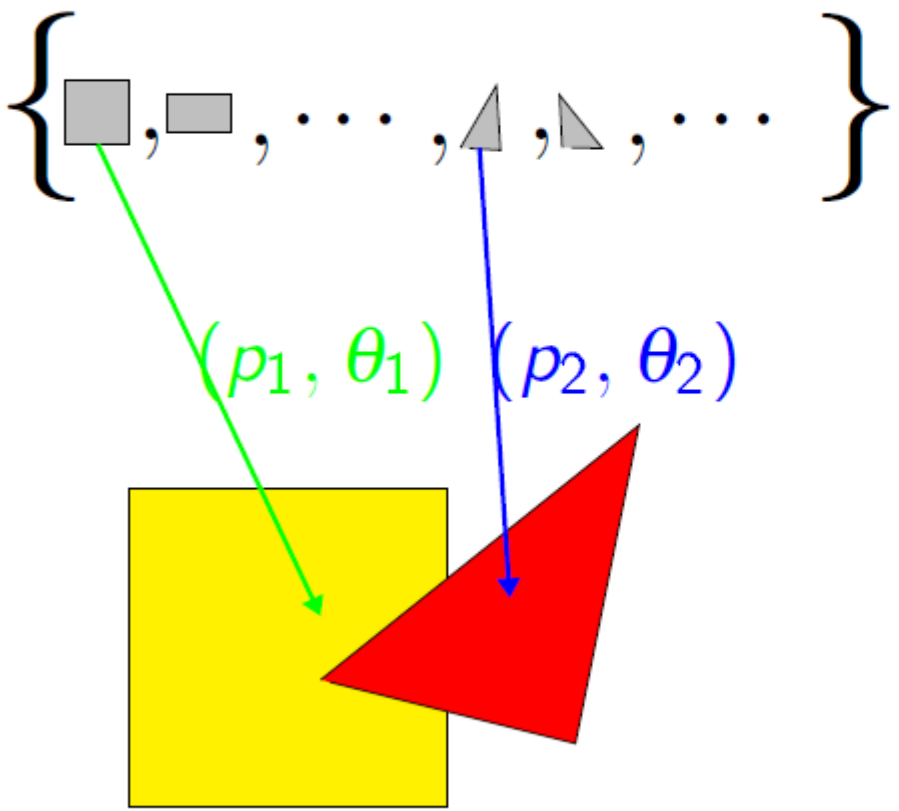


Which object is in front?

An alternative hypothesis (to the obvious one)



A scene generating mechanism



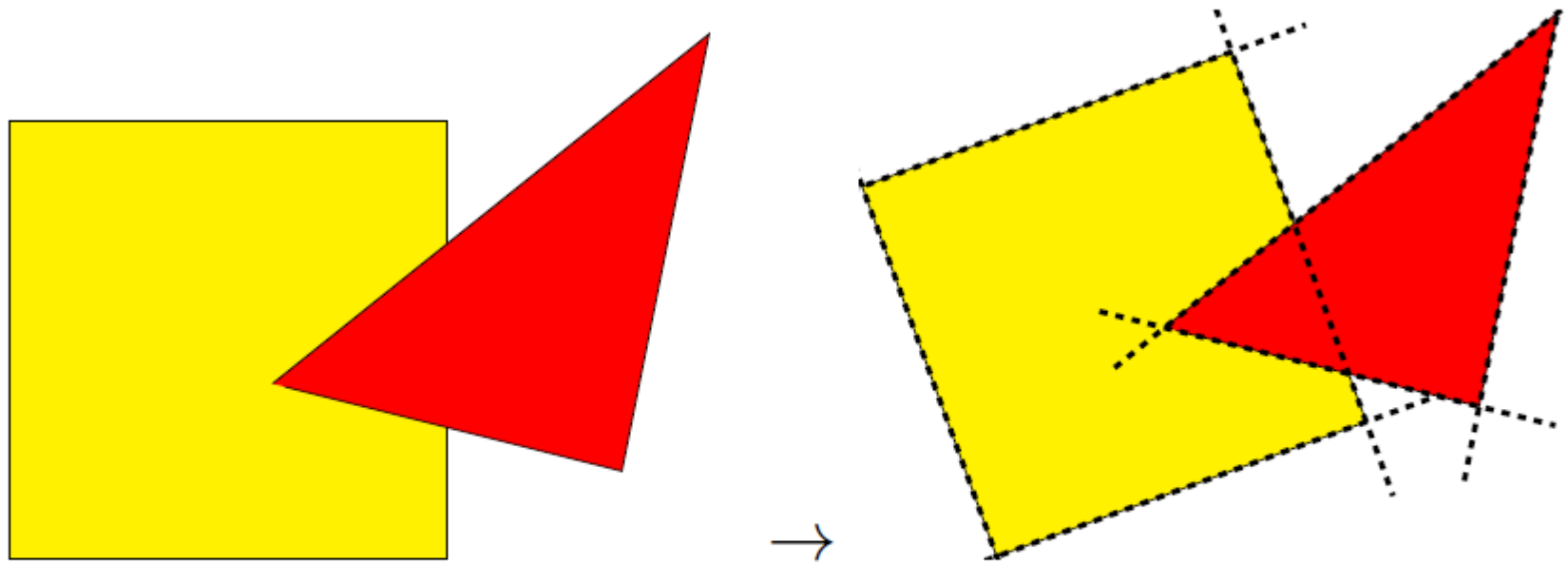
Given two polygonal objects O_1 and O_2, a scene S is generated according to the function

$$S = m_{p_1,\theta_1,p_2,\theta_2}(O_1, O_2)$$

where $m_{p_1,\theta_1,p_2,\theta_2}$ represents the mechanisms that puts first O_1 on the scene at position p_1 with orientation θ_1, then puts O_2 at position p_2 with orientation θ_2 in front of O_1.

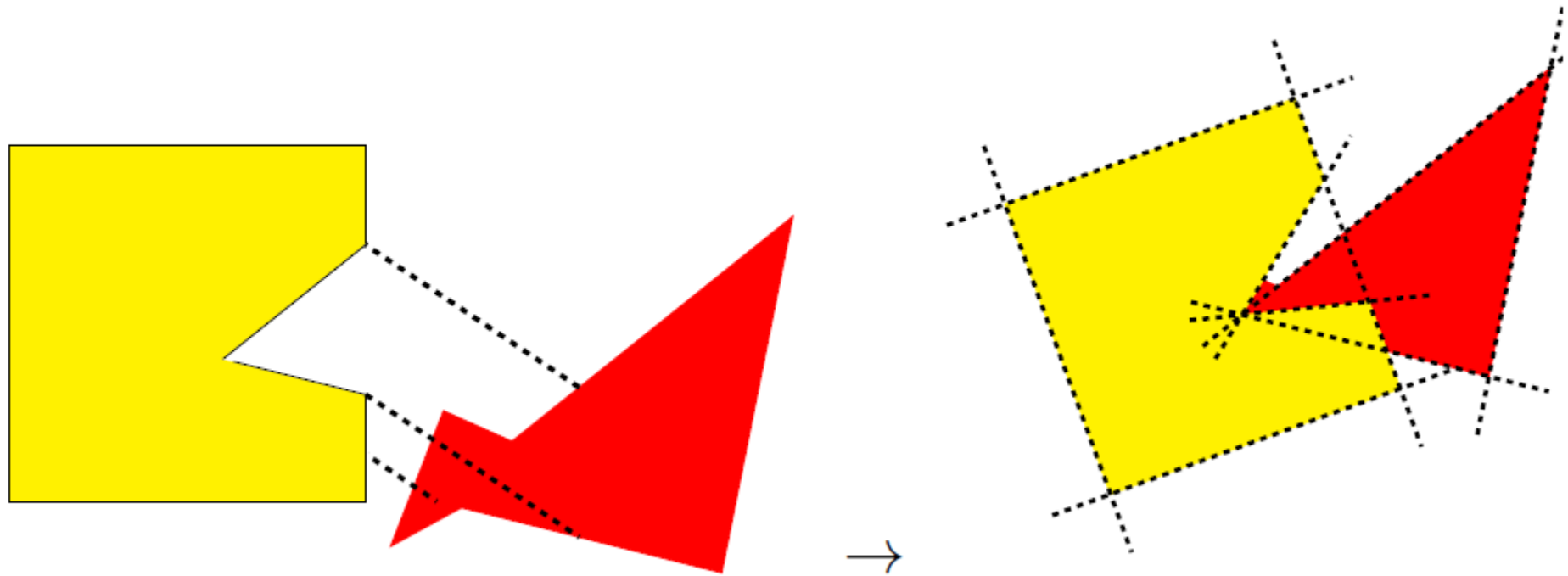
ICM type assumption: objects and (θ, p)s are chosen independently.

Perturbing the mechanism



The number of lines stays the same.

Perturbing the mechanism (II)



Now the number of lines in the scene increases.

Wrapping it up

- A data generating mechanism acting on causes (O_1, O_2):

$$S := m_{p_1,\theta_1,p_2,\theta_2}(O_1, O_2)$$

- A "contrast" (characteristic of the scene):

$$C(S) = \# \text{ of straight lines}$$

- A group of transformations: Rotations r_ϕ

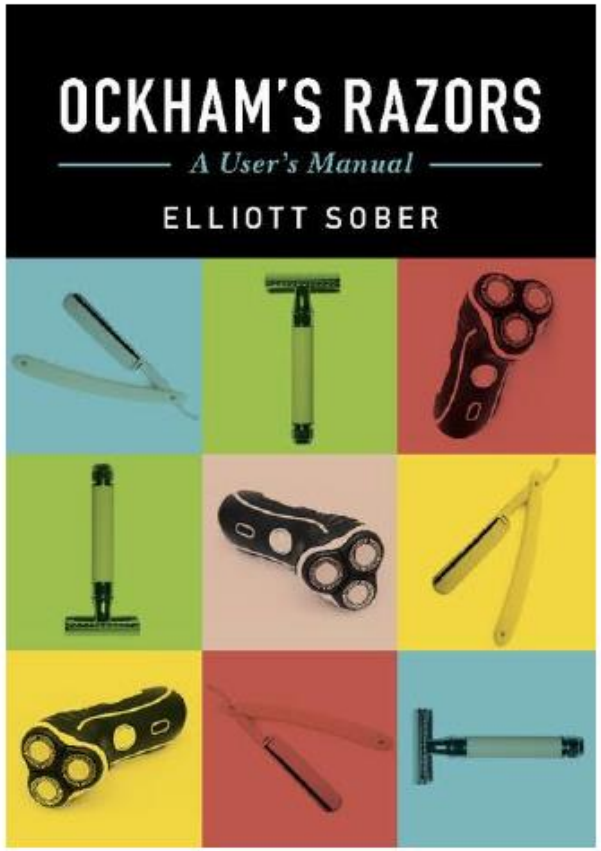
Causal inference algorithm

Procedure for both models (red or yellow object in front)

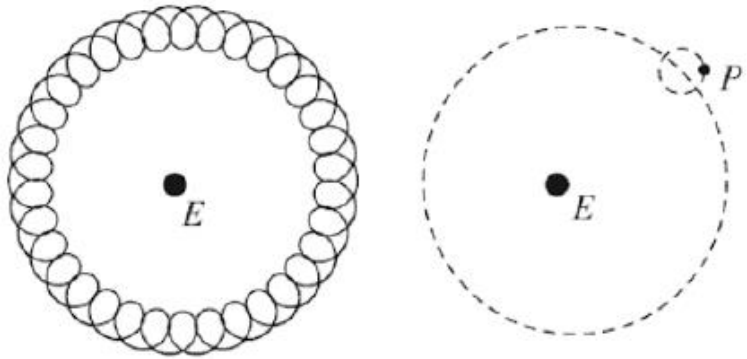
- Apply and arbitrary (random) transformation
- Check whether $C(S) = C(S_\phi)$ for most choices of ϕ (almost all)
- If yes the model is *generic*; if no, model is *suspicious* and unlikely to be true.

Note that this causal inference algorithm applies to deterministic variables.

Related concepts: Occam's razor



Epicycles of the Ptolemaic system.



The distinction between prediction and accommodation also applies when each model is asked to explain a second regularity:

(I-N) The inferior planets (Mercury and Venus) are always observed to be near each other and near the Sun.

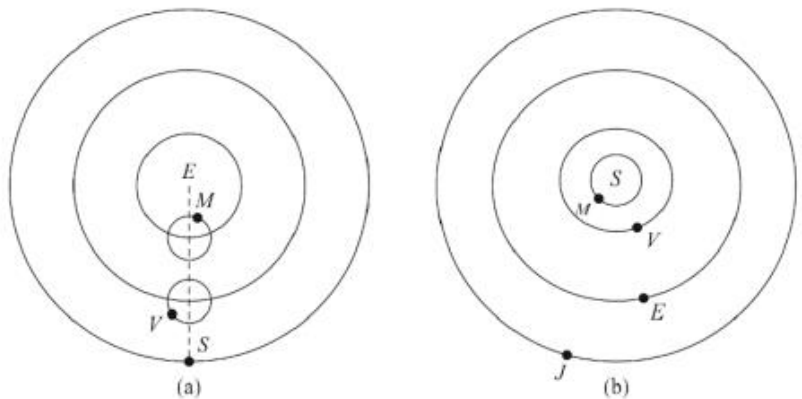


Figure 1.6

Ptolemaic versus Copernican.

General framework

Definition

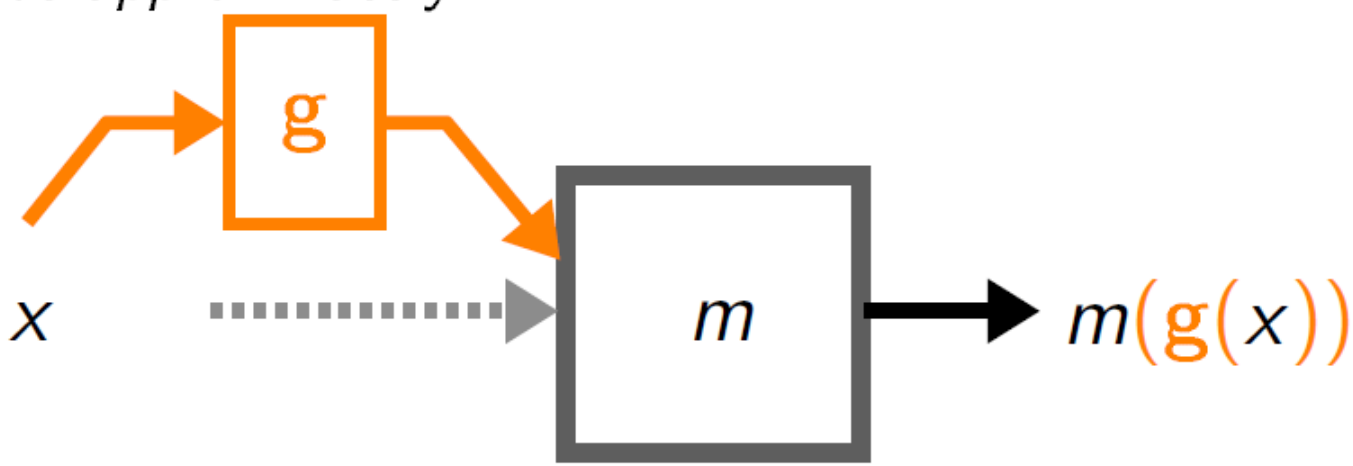
The Expected Generic Contrast (EGC) of a cause mechanism pair (x, m) is defined as:

$$\langle C \rangle_{m,x} = \mathbb{E}_{g \sim \mu_{\mathscr{G}}} C(mgx)\,. \tag{1}$$

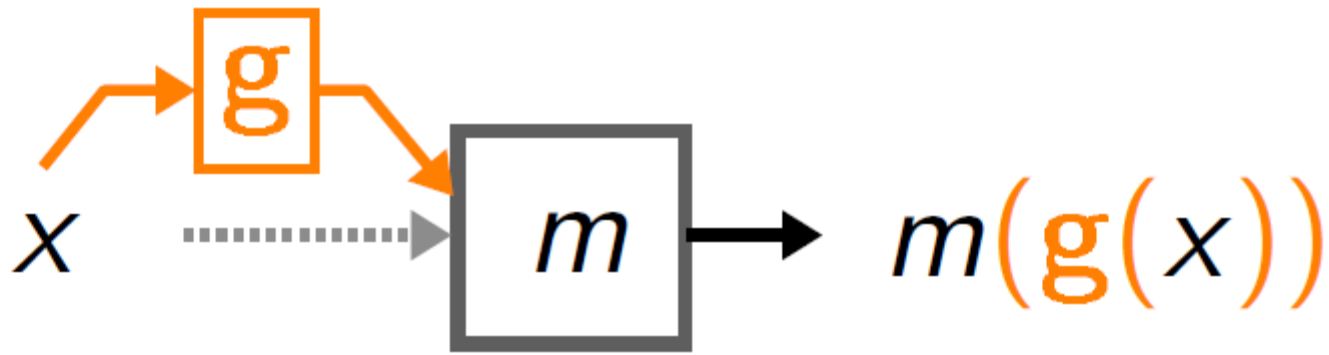
We say that the relation between m and x is $\mathscr{G}$-generic under C, whenever

$$C(mx) = \langle C \rangle_{m,x} \tag{2}$$

holds approximately.



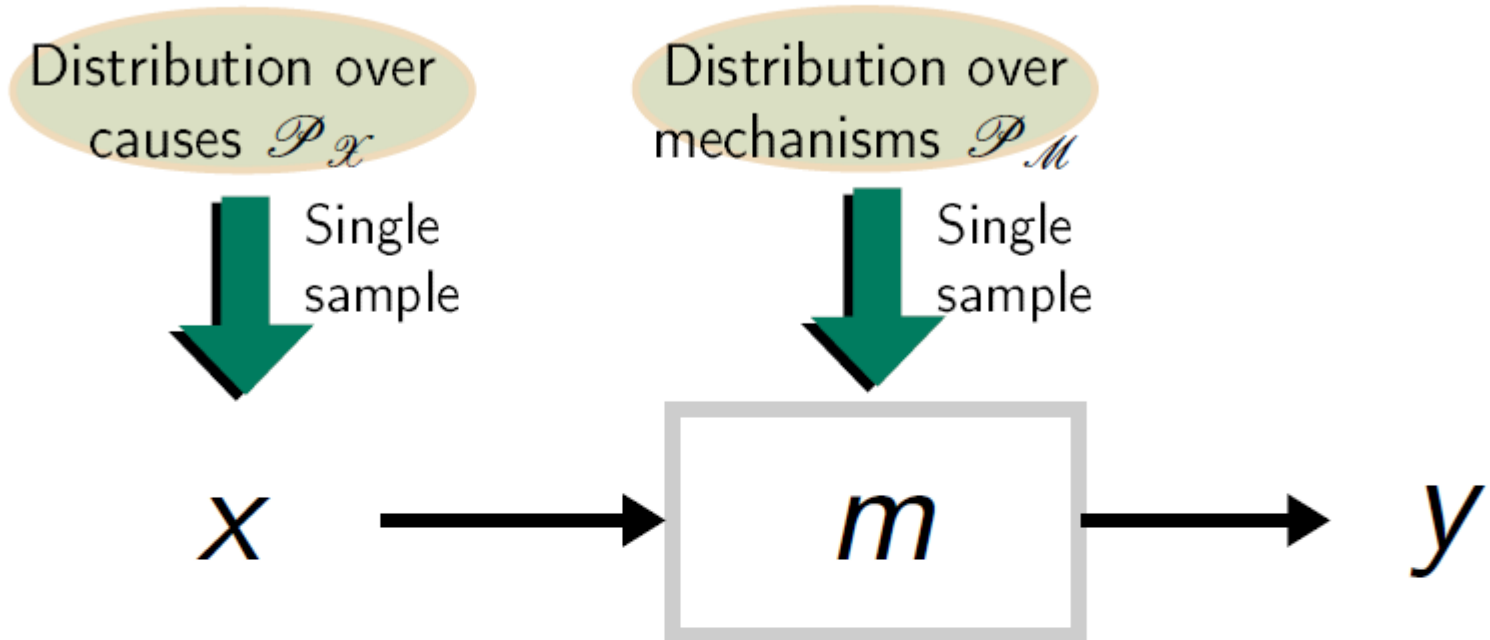
Interpretation 1



The generic transformation g perturbs the relationship between cause and mechanism, and can be interpreted as a soft intervention/counterfactual:

> *What would have happened, had another cause/mechanisms been picked at random by Nature?*

Interpretation 2

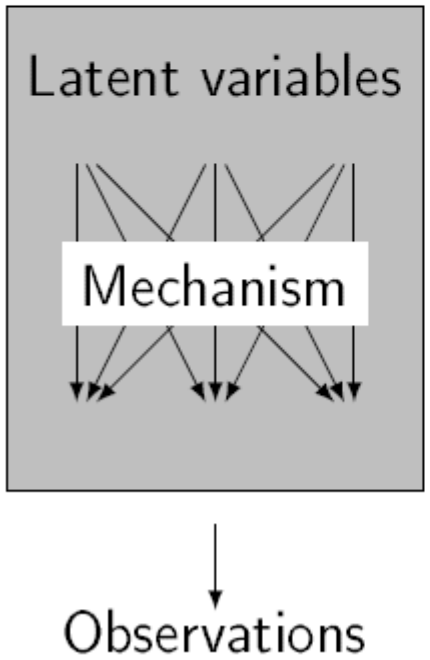


Assume $\mathscr{P}_{\mathscr{X}}$ $\mathscr{G}$-invariant, it can be parametrized as $x = g\tilde{x}$ for $\tilde{x}$ and g are independent RVs and $g \sim \mu_{\mathscr{G}}$.
Then the expectation of the generic ratio is

$$\mathbb{E}_x \left[C(mx) / \langle C \rangle_{m,x} \right] = \mathbb{E}_{\tilde{x}} \mathbb{E}_{\tilde{g}} \left[C(mg\tilde{x}) / \langle C \rangle_{m,g\tilde{x}} \right] = 1 \qquad (3)$$

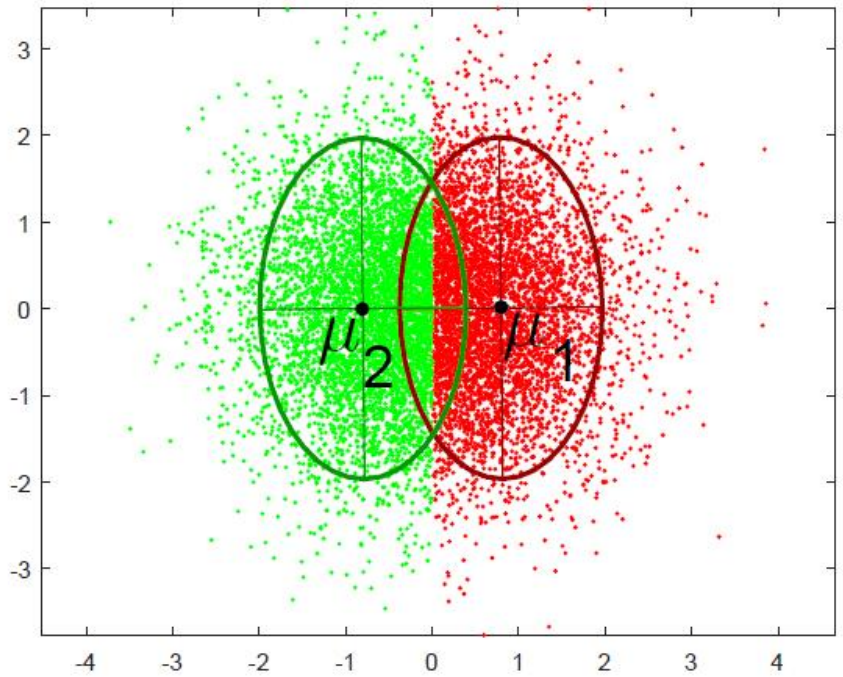
So genericity is true at least "on average".

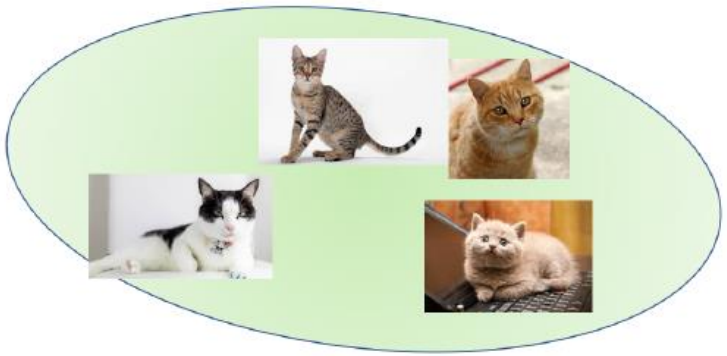
Causal latent variable models



- In applications, Unsupervised Leaning (UL) is used to make sense of experimental data.
- By extracting latent variables (classes or components), scientists expect to lean about the underlying mechanisms generating the data.
- How to know whether the outcome provided corresponding inference algorithms is artificial and unlikely to reflect an authentic underlying structure?
- We propose to exploit ICM and group theoretic principles to validate UL algorithms.

Application to clustering

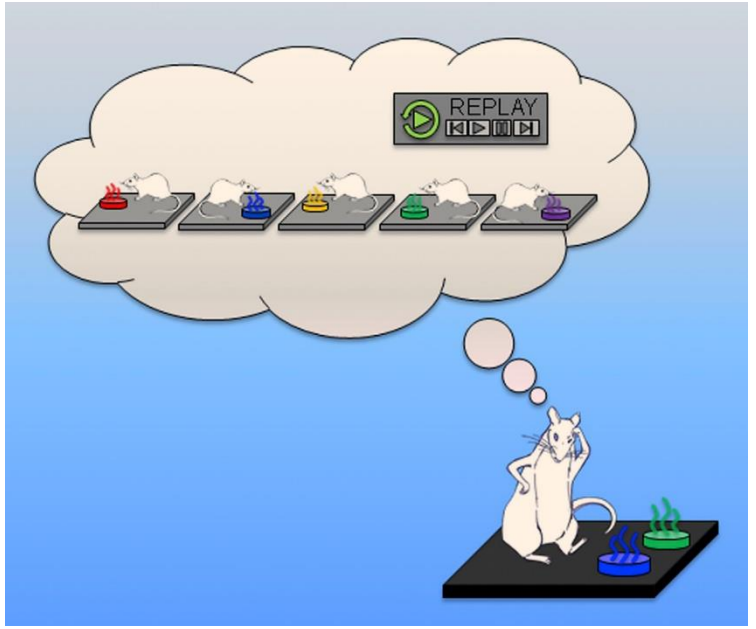




Episodic memory: generating without sensing.

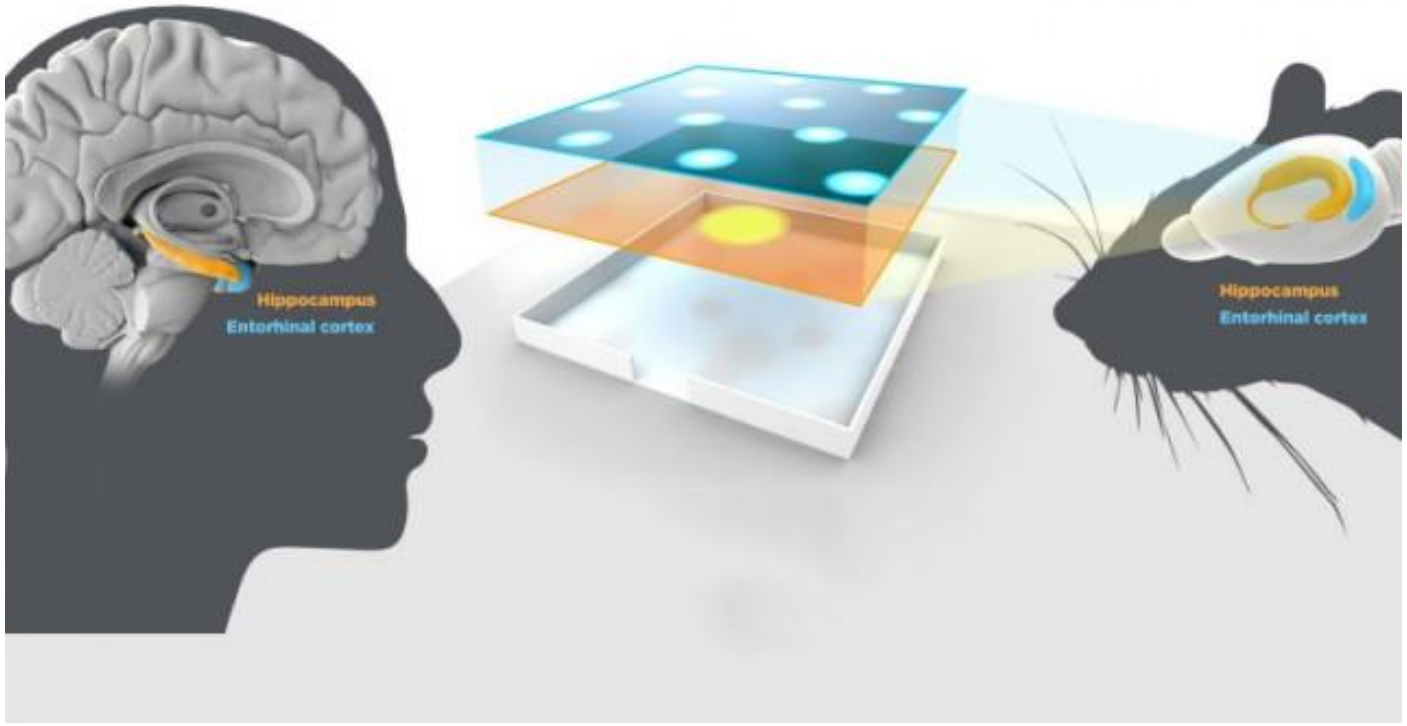
- The **memory** of autobiographical events (times, places, associated emotions,…).
- The collection of past personal experiences.
- Unambiguously defined in *humans*.
- Extended to *animals*.
- *Multimodal/sequential/associative*.
- Key structure involved in both:

the *medial temporal lobe*, including the *hippocampus*.



Panoz-Brown et al., 2018

The hippocampal formation

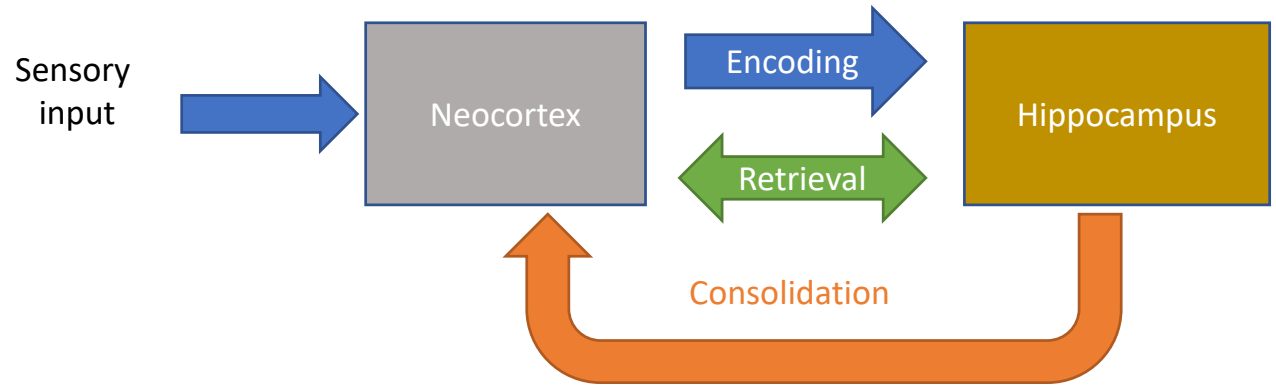


extremetech.com

Equivalent in reptiles:
dorsal ventricular ridge?
Shein-Idelson, Science 2016

Memory related functions

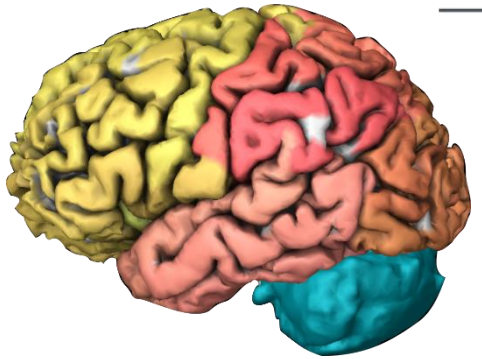
- **Encoding**: forming new memories.
- **Recollection**: *retrieval* of contextual information associated to a specific experience.
- **Consolidation**: turning short time storage (hours, days, …) to long term memory (years, life…).

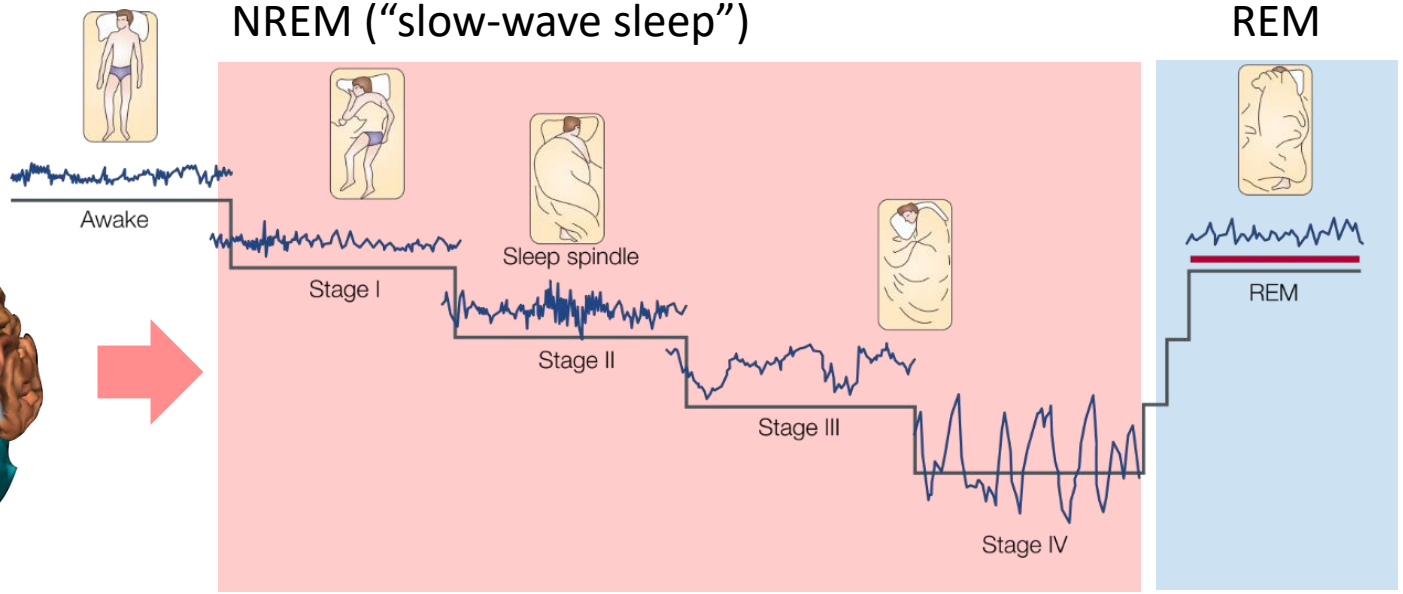


Complementary learning systems: McClelland et al. (1995)

EEG correlates of sleep stages

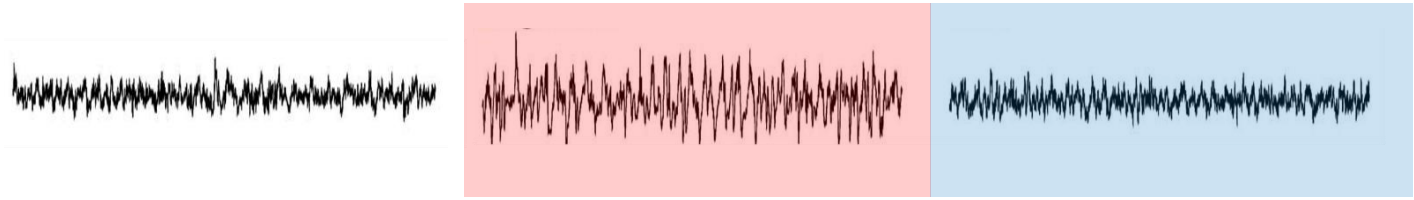
Human





Adapted from: Hobson, *Nat. Rev. Neurosci.*, 2009; Pace-Schott & Hobson, *Nat. Rev. Neurosci.*, 2002

Non-human primate



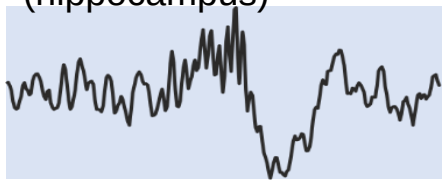
Slide courtesy of Juan Ramirez-Villegas

Background - Sleep Stages and Transient Neural Events

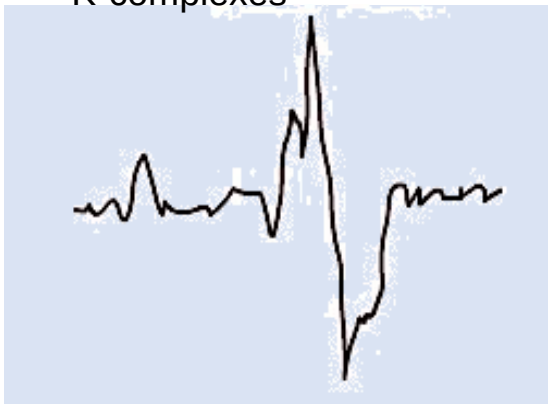
Sleep stages

Wakefulness	Slow Wave Sleep (SWS)	Rapid Eye Movement(REM)
	• Slow oscillations • Spindles • K-complexes • Sharp wave-ripples	• Ponto-Geniculo-Occipital (PGO) waves

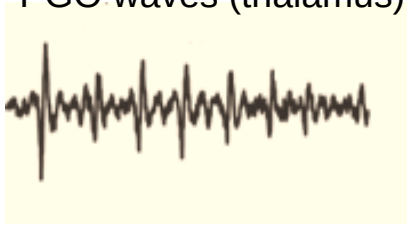
Sharp wave-ripples (hippocampus)



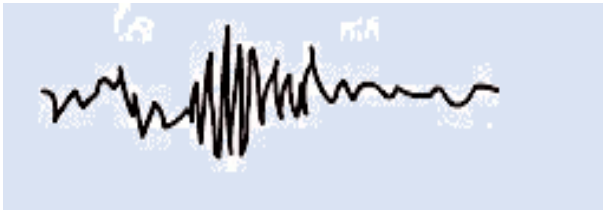
K-complexes



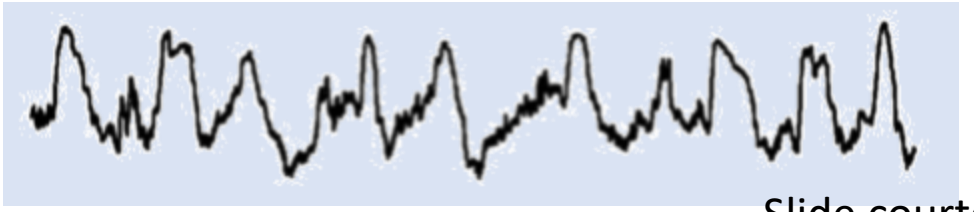
PGO waves (thalamus)



Spindles

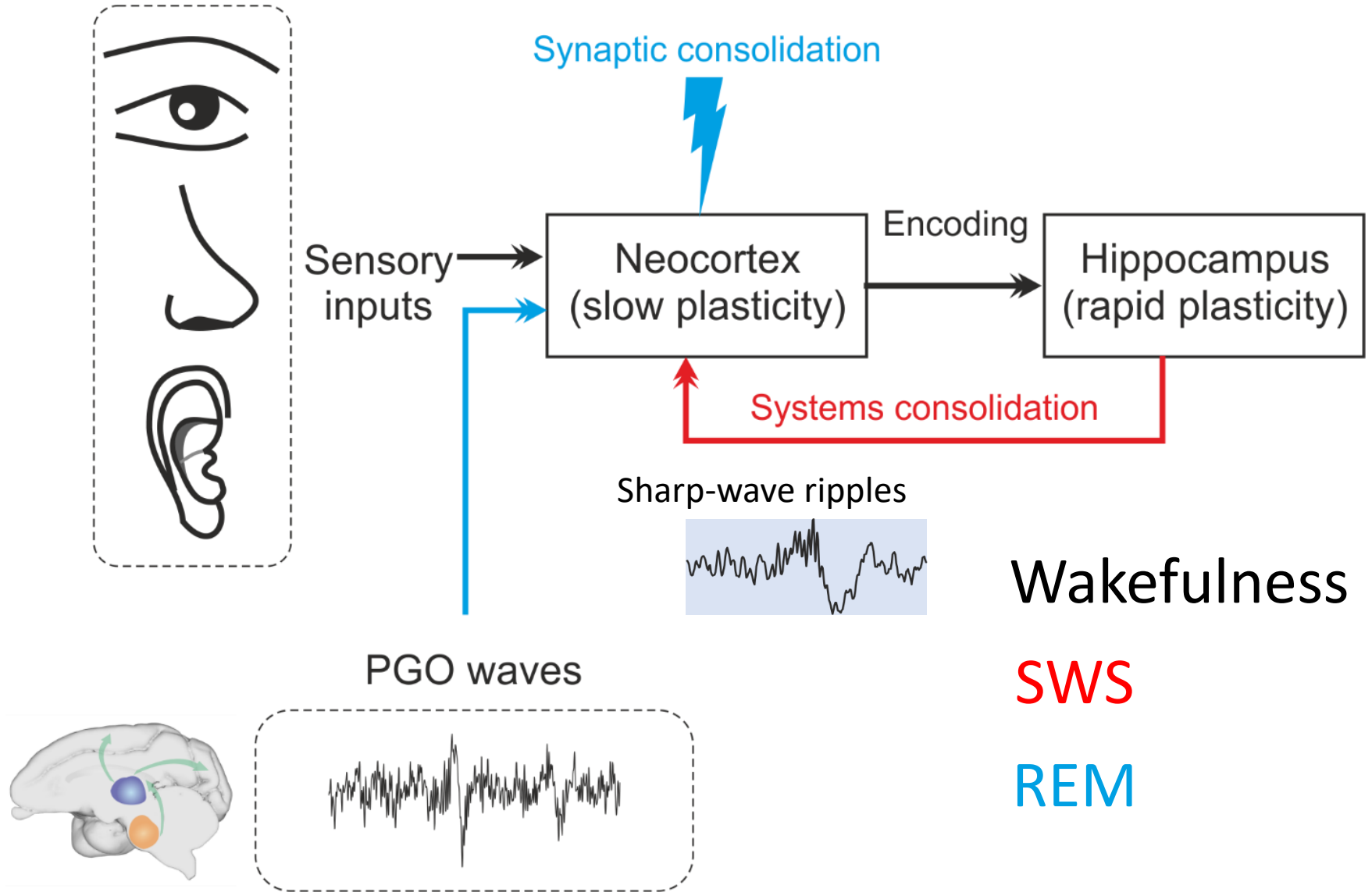


Slow waves



Slide courtesy of Kaidi Shao

Memory consolidation and the hippocampal circuit

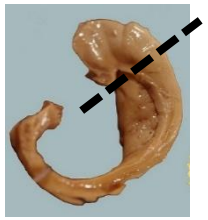


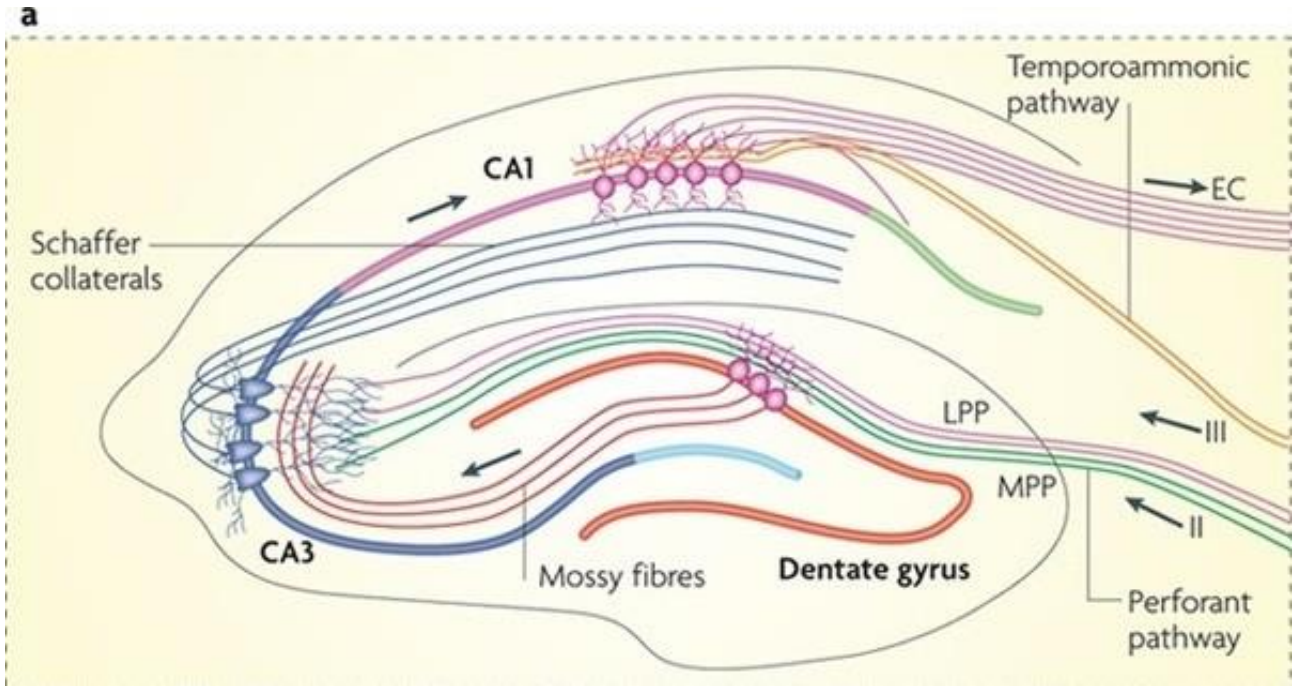
Wakefulness

SWS

REM

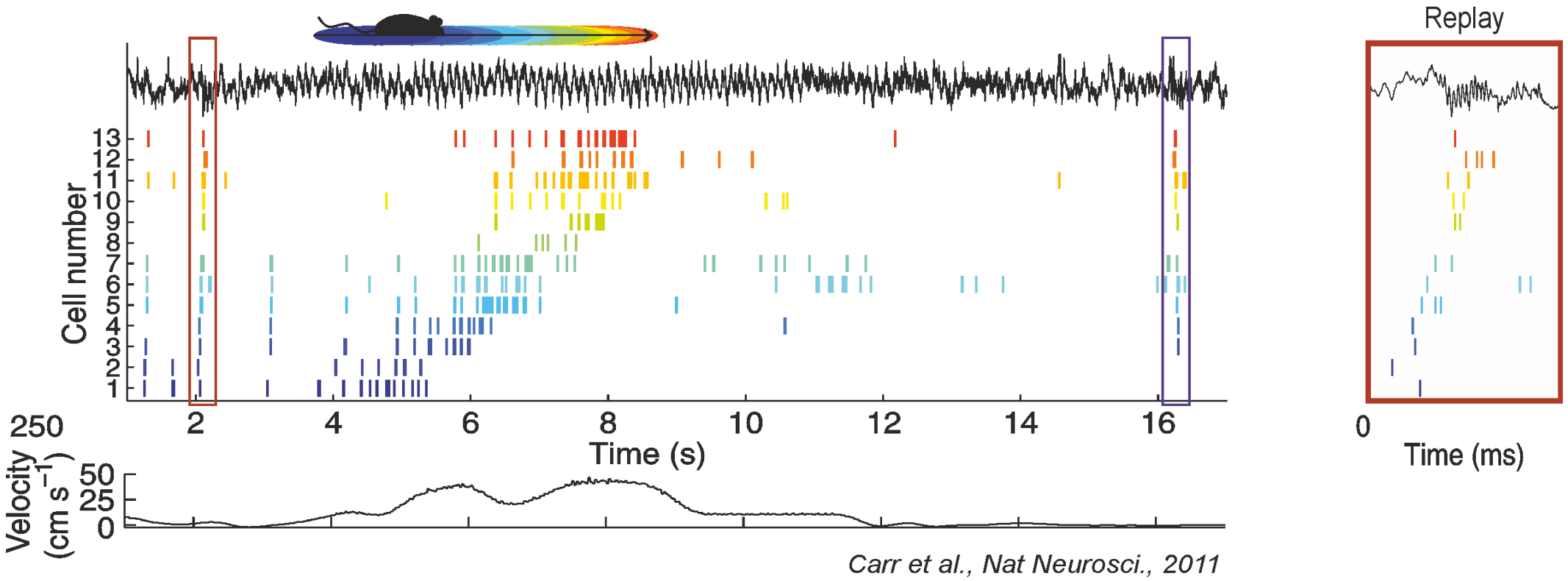
Hippocampal subfields





Deng et al. 2010

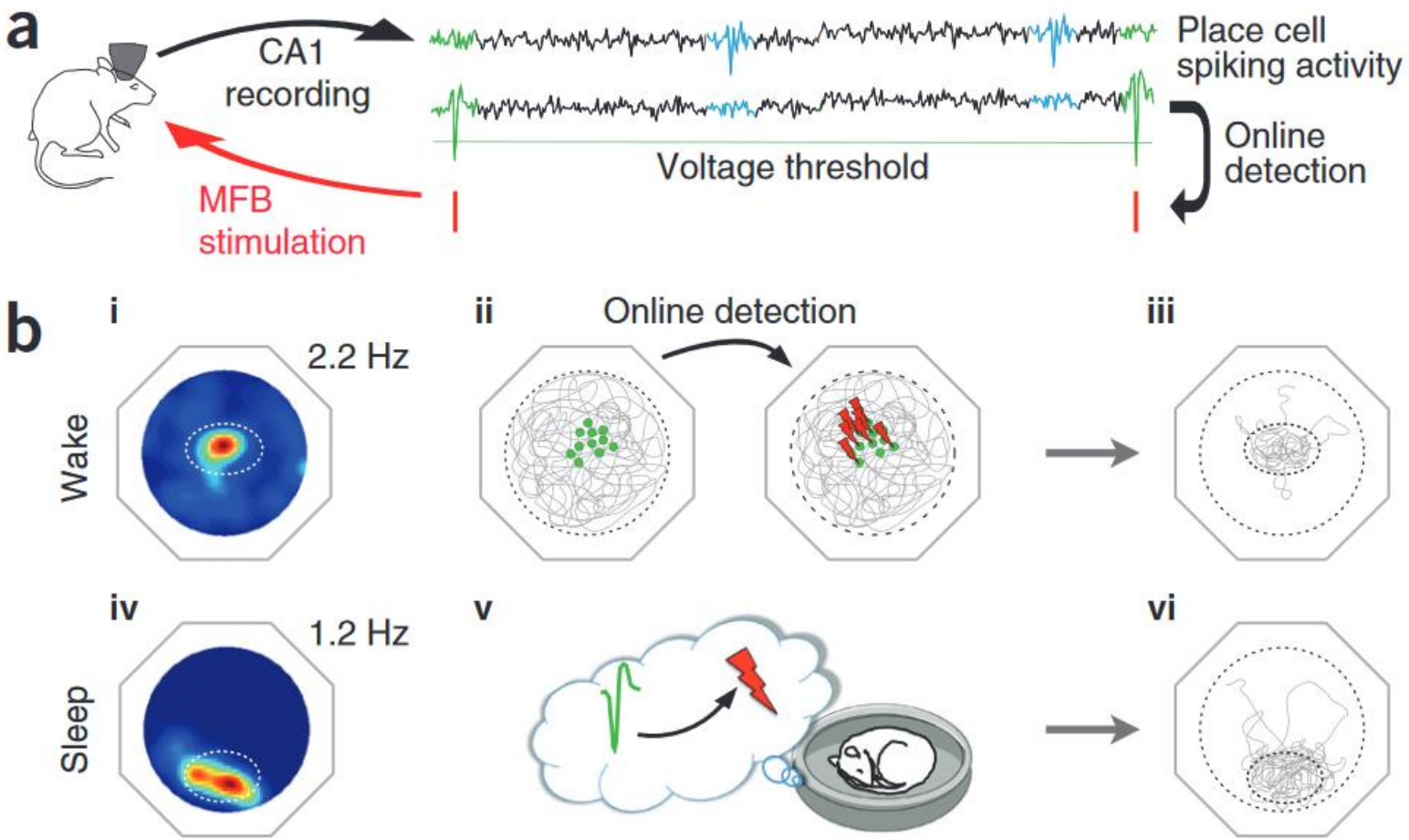
Sharp-wave ripples and replay in CA1



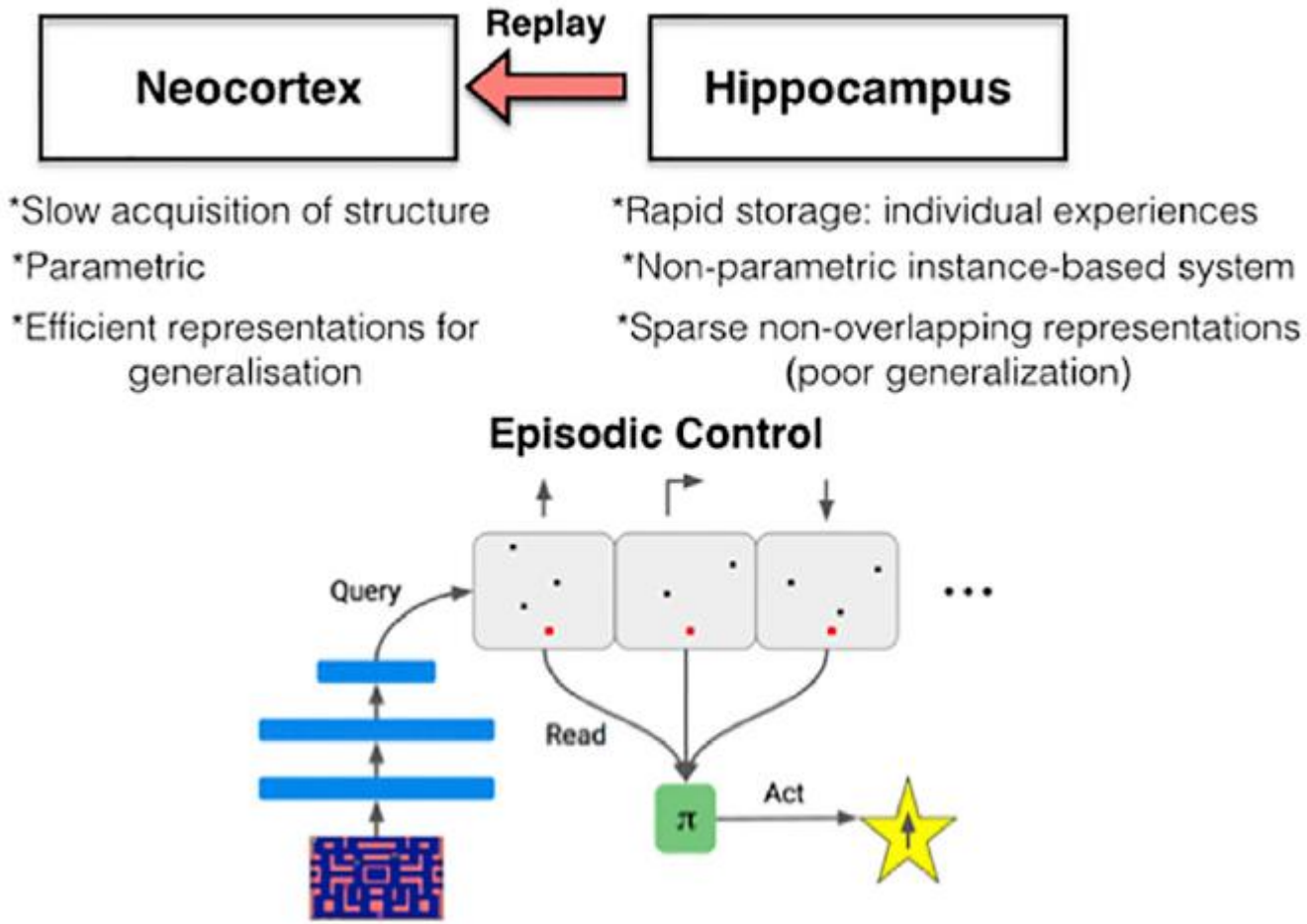
Carr et al., Nat Neurosci., 2011

- Behaviourally-relevant spiking activity replays during sharp wave-ripple (SWR) episodes.
- Time-frequency characteristics (induction of plasticity)
- SWR suppression leads to learning deficits.

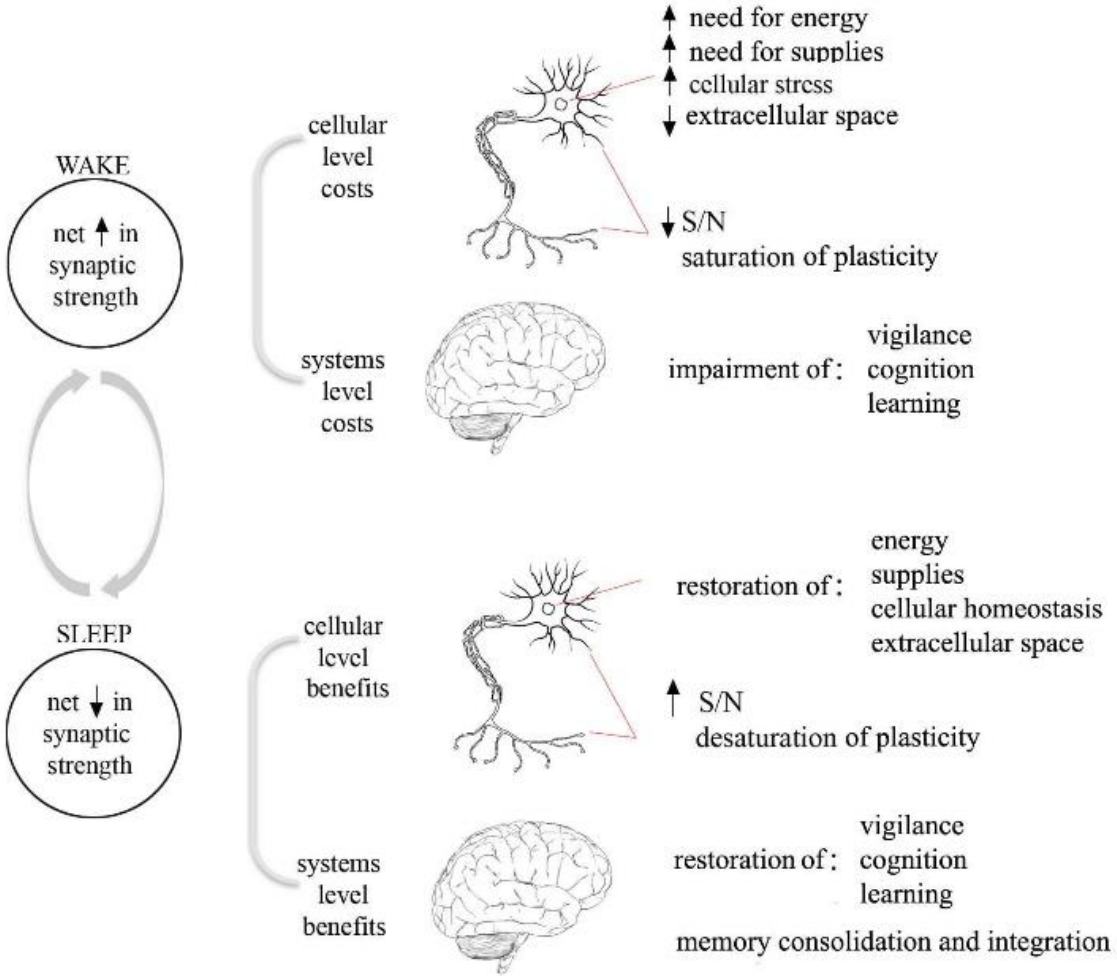
Memories can be manipulated (Laviellon, 2015)

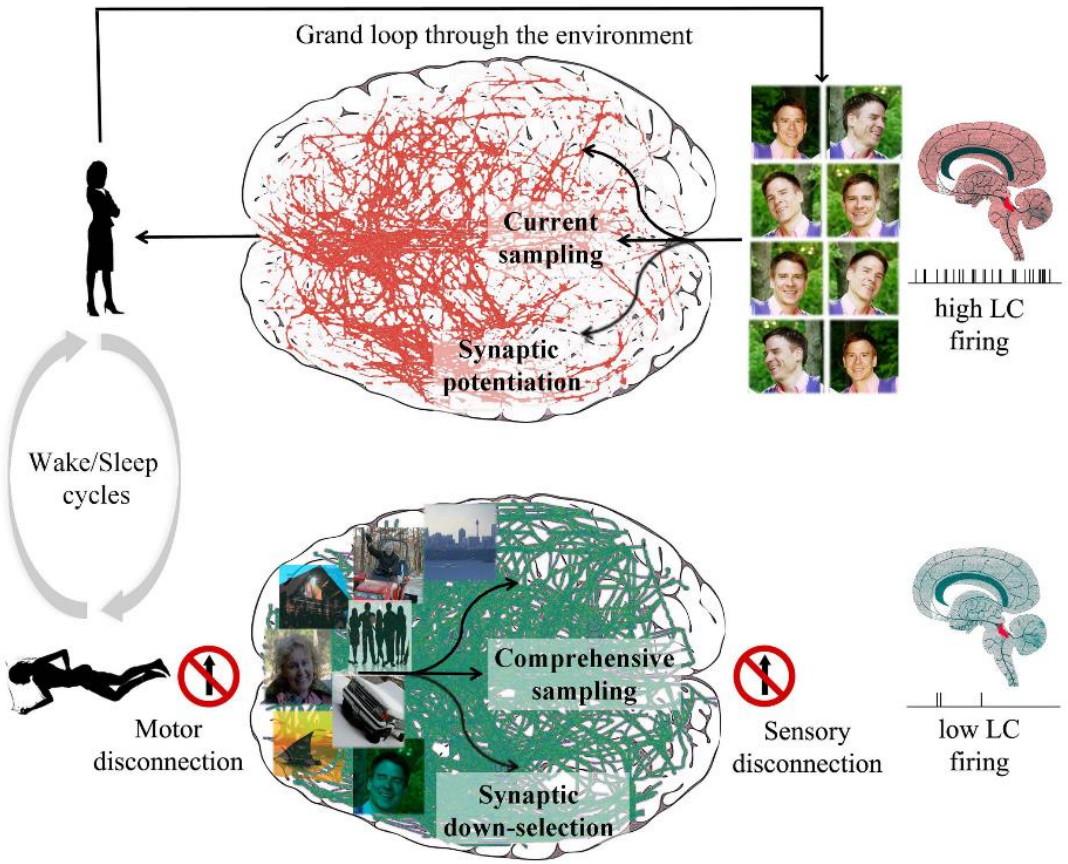


Episodic memory for artificial intelligence
(review: Hassabis et al. 2017)



Synaptic homeostasis (review: Tononi & Cirelli, 2014)





Research questions

What are the cooperative events?

➡ Event detection and classification.

Signal processing & Machine Learning

What are the mechanisms?

➡ Modeling and inference of multiscale interactions, relationship to measured signals.

Statistics, causal inference & Biophysical models

What are the functions?

➡ Neural encoding and processing of sensory information, optimality principles of neural computations and plasticity.

Information theory & Learning theory

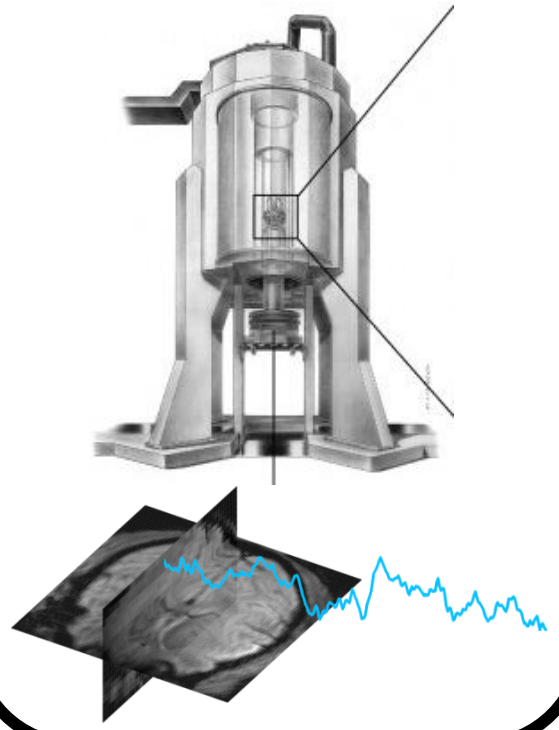
Functional neural signals

We exploit the rich information in multivariate neural activity at **multiple scales**.

Local **Scale** Brain-wide

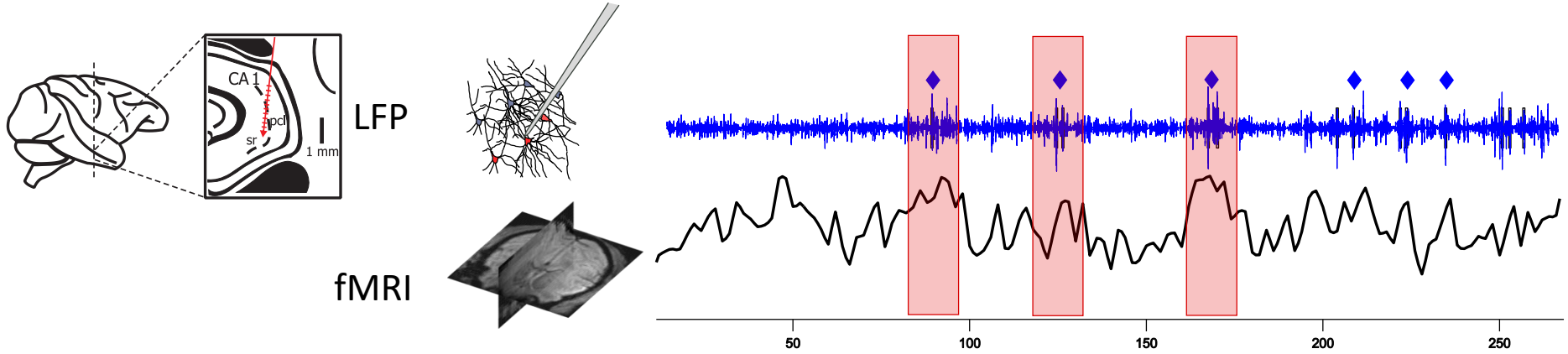
Microelectrode

fMRI

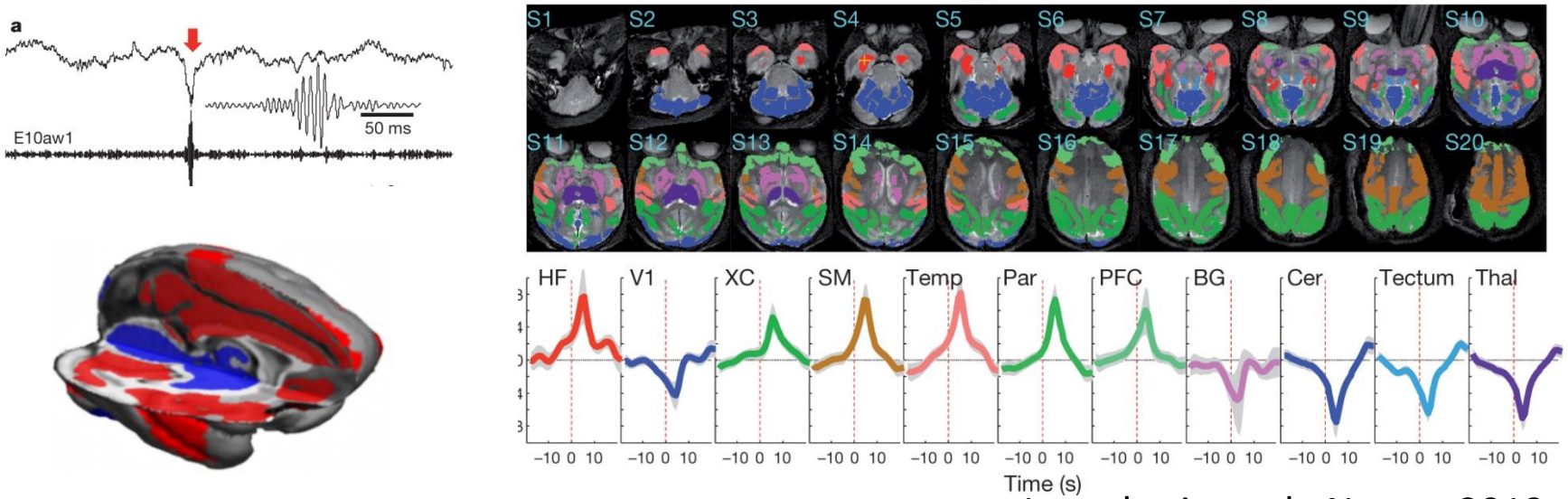


Relationship?

Neural event triggered (NET)-fMRI

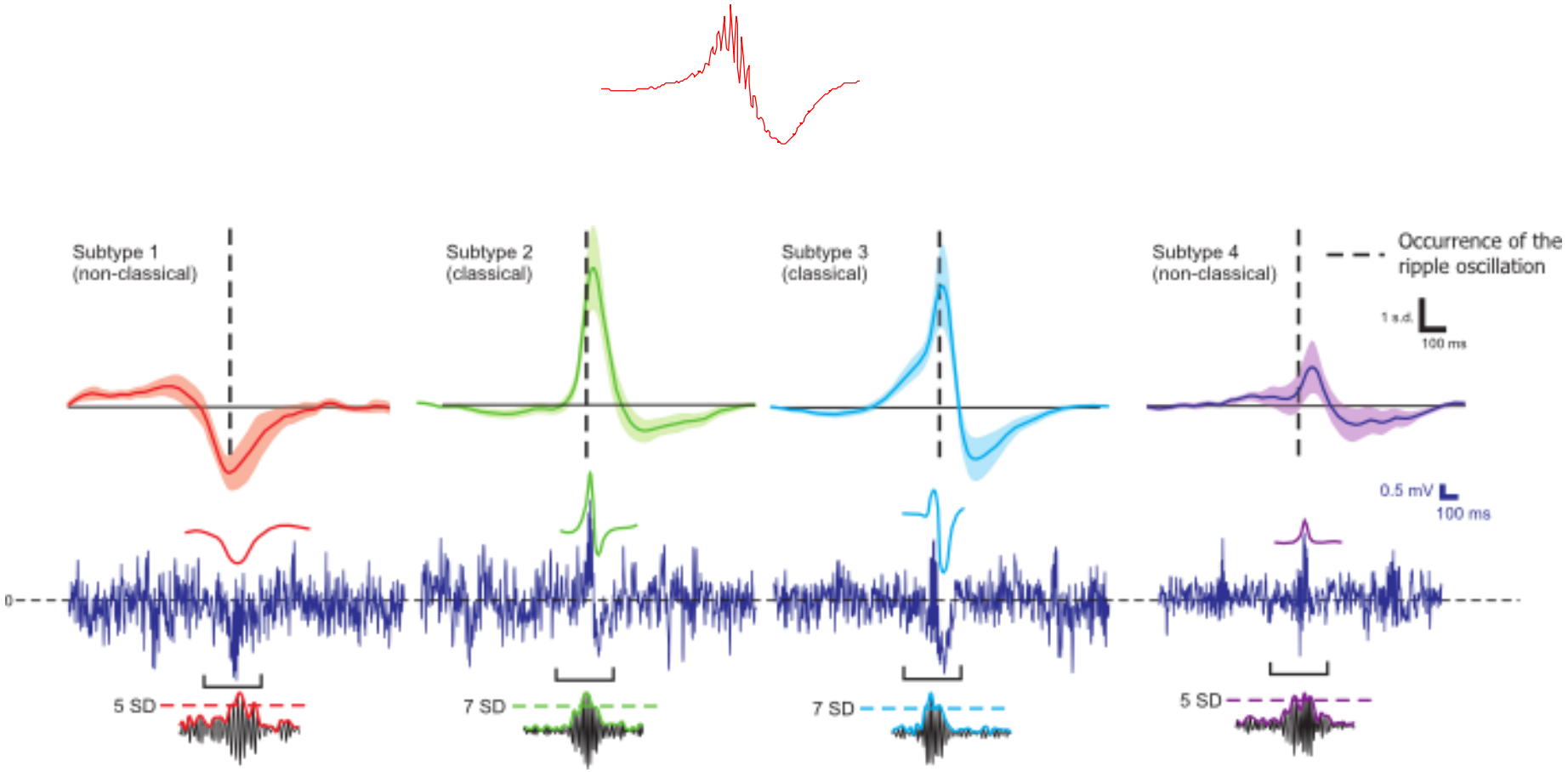


The case of Hippocampal Sharp-wave ripples



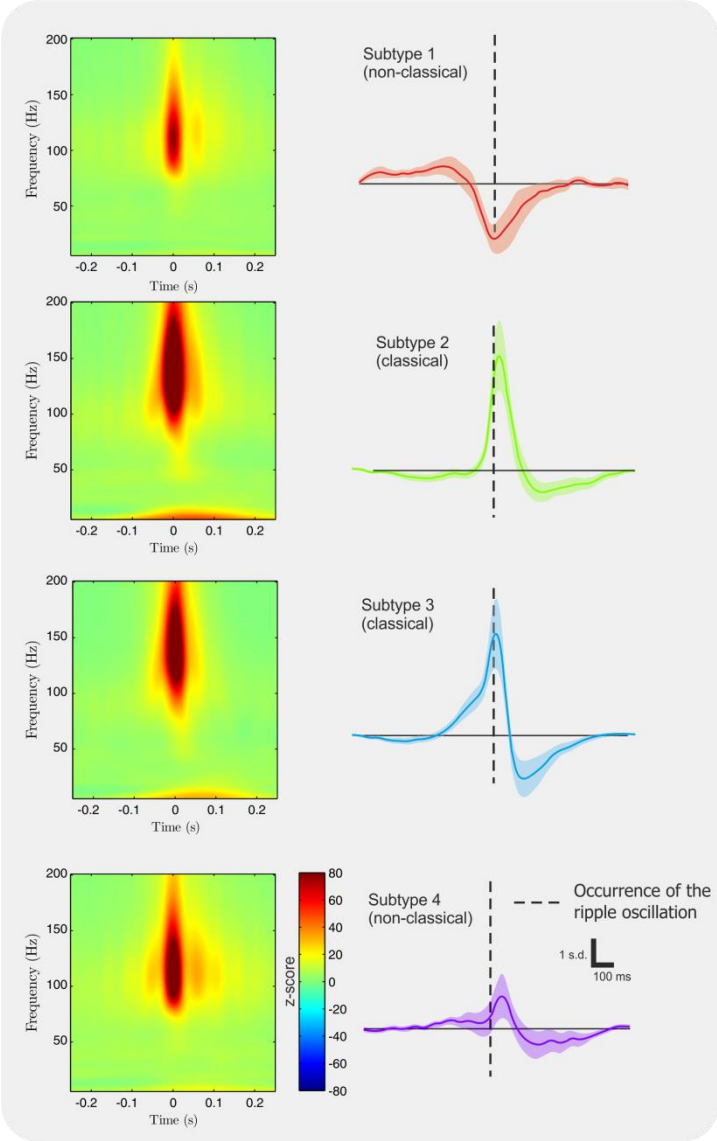
Logothetis et al., Nature 2012

Sharp Wave-Ripple episodes: diversity

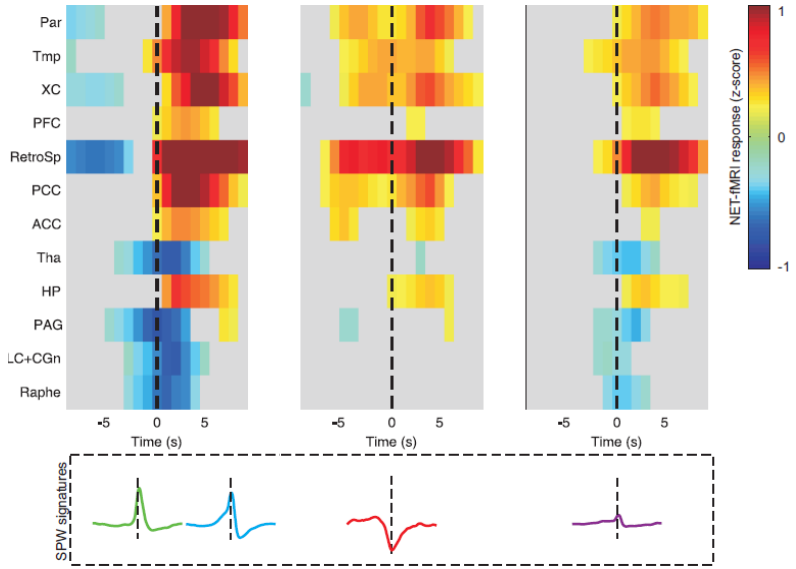


Ramirez-Villegas et al., PNAS 2015.

Brain wide signature of Sharp Wave-Ripple diversity



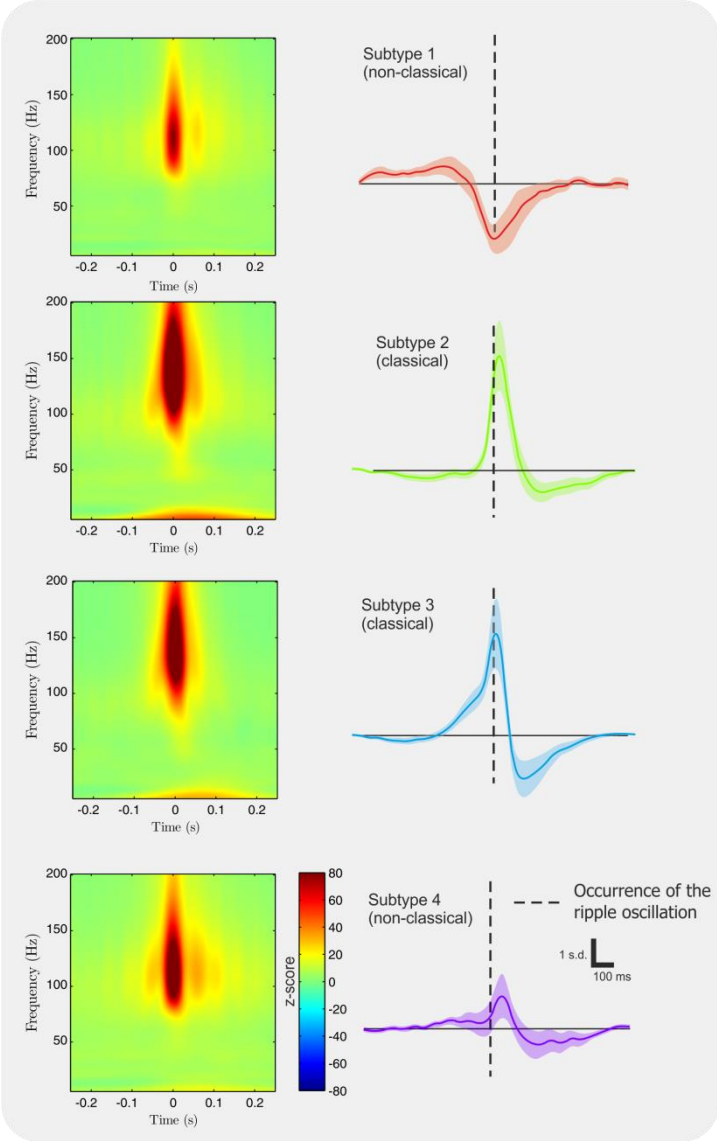
Concurrent fMRI recordings reveal SWR-specific brain-wide correlates:



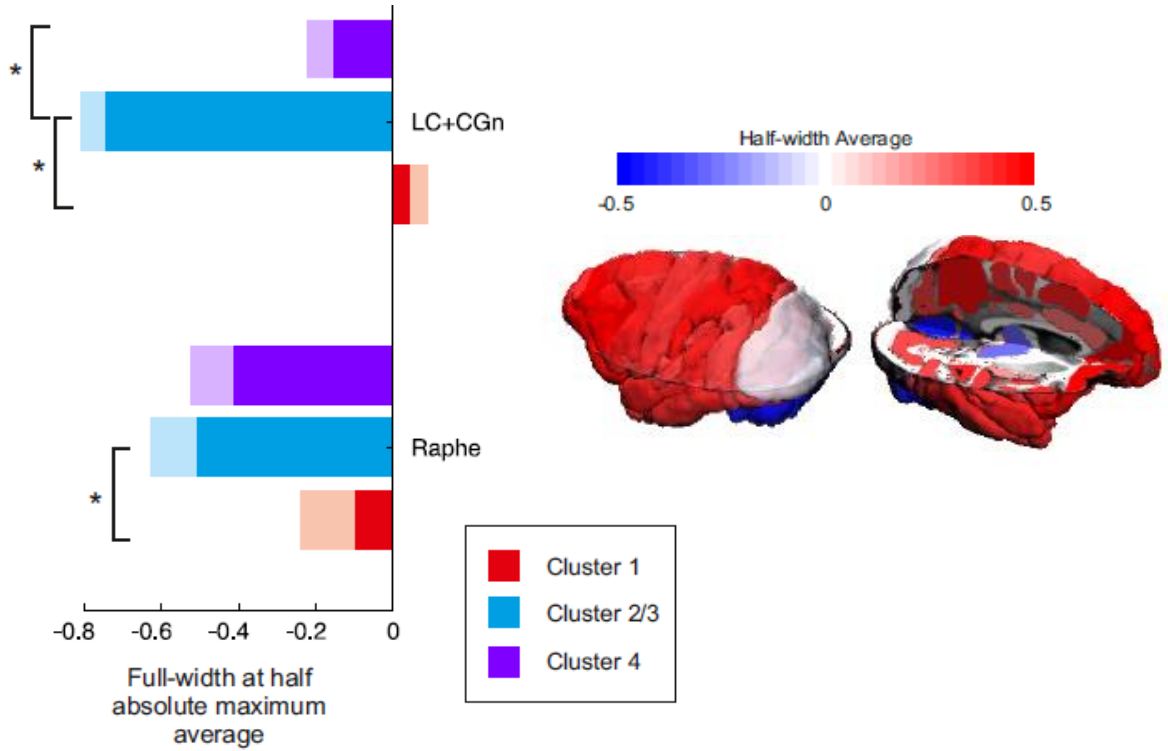
SWR may modulate cortical circuitry (and vice-versa) in a differentiated manner.

Ramirez-Villegas et al., *PNAS*, 2015

Brain wide signature of Sharp Wave-Ripple diversity

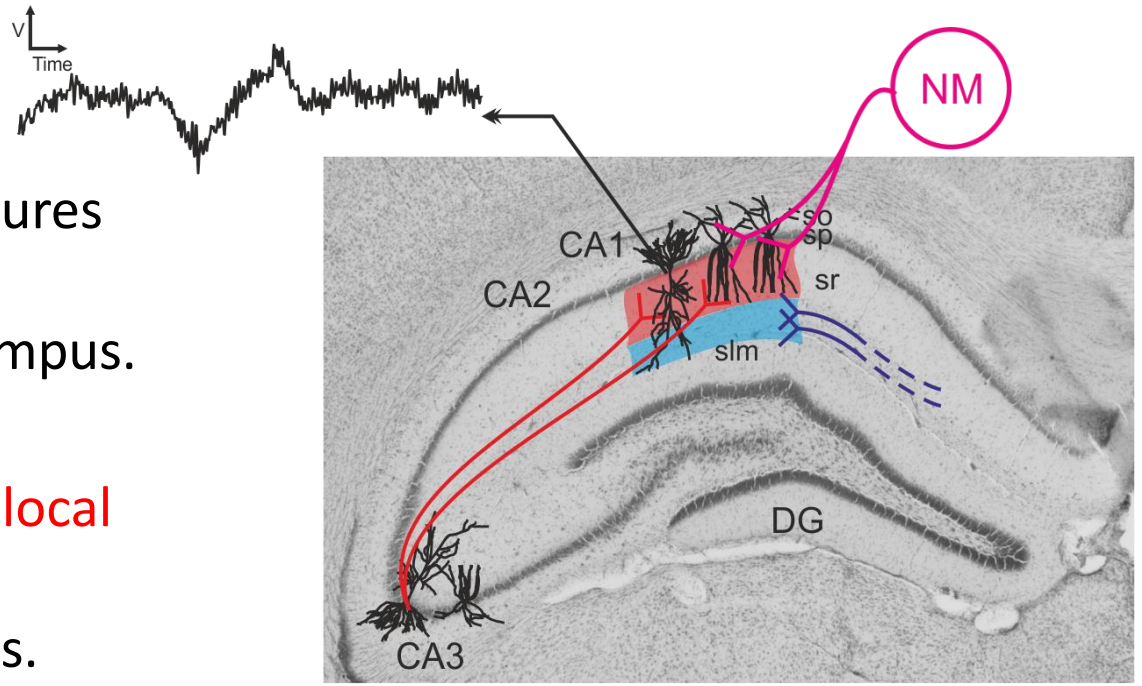


Concurrent fMRI recordings reveal potential involvement of subcortical NM centers on the ripple phenomenon.



Ramirez-Villegas et al., *PNAS*, 2015

Brain wide signature of Sharp Wave-Ripple diversity



SWR field potential signatures may be expressed due to remote inputs to hippocampus.

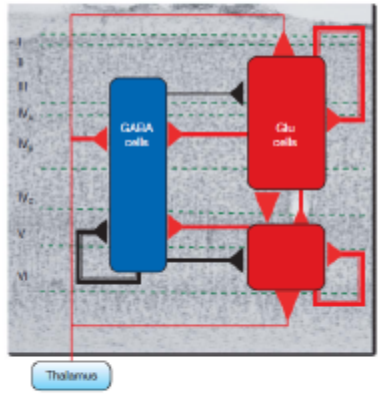
These inputs may change local E-I balance, and also the location of current sources.

Our work elucidates that different types of SWR may have differentiated impact on cortical circuitry. Some SWR may also occur after cortical activation, consistent with memory retrieval during awake state (see also: Mohajerani and McNaughton, 2016).

Ramirez-Villegas et al., *PNAS*, 2015

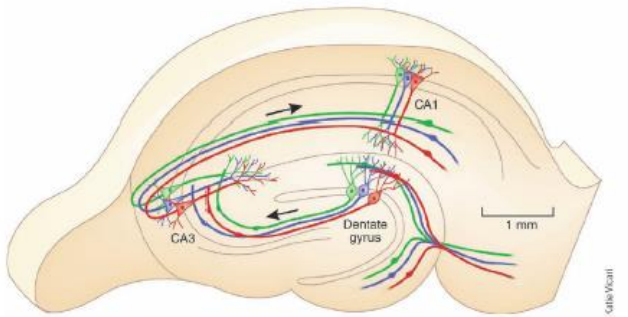
Causality to investigate information flow in brain networks

Long range connectivity

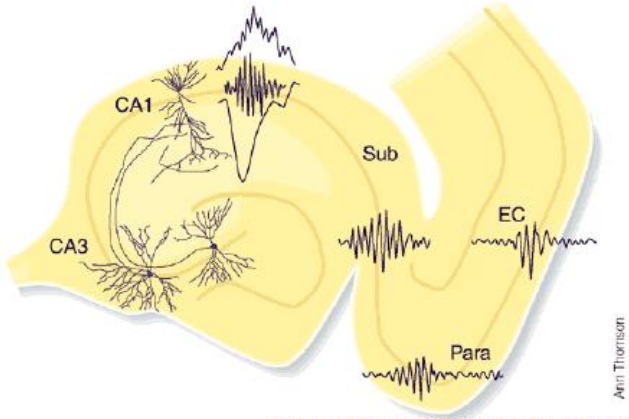


Cortical microcircuits

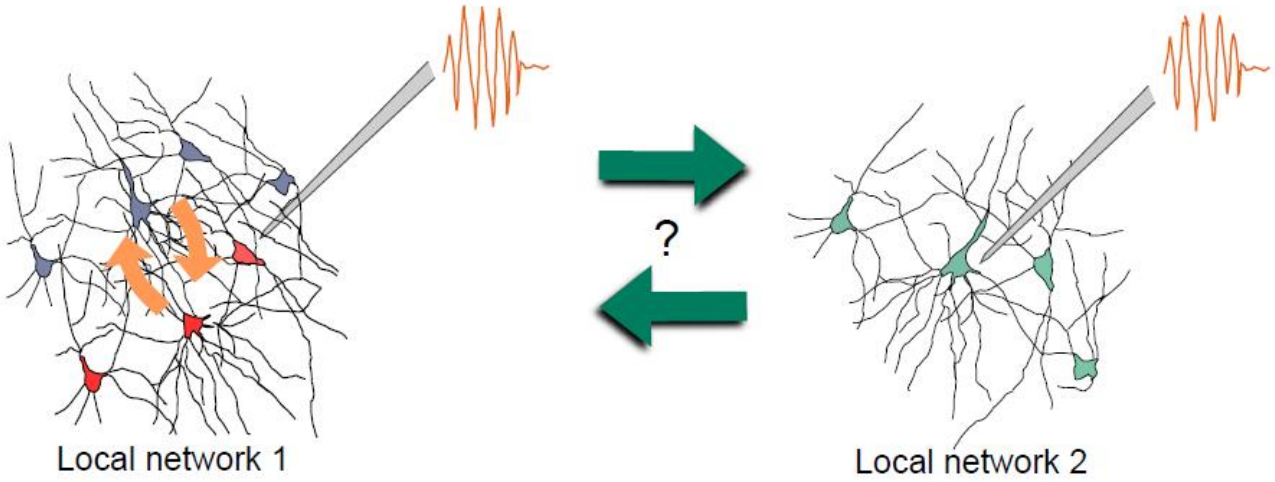
Hippocampus as a quasi ground-truth dataset



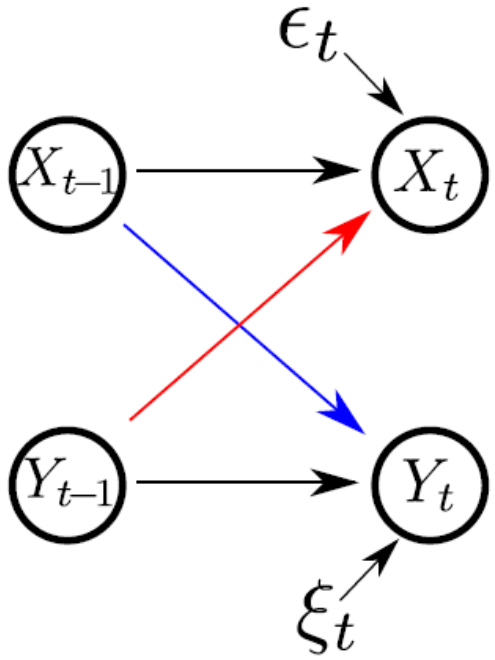
Moser, 2011



Buzsaki and Chrobak, 2011



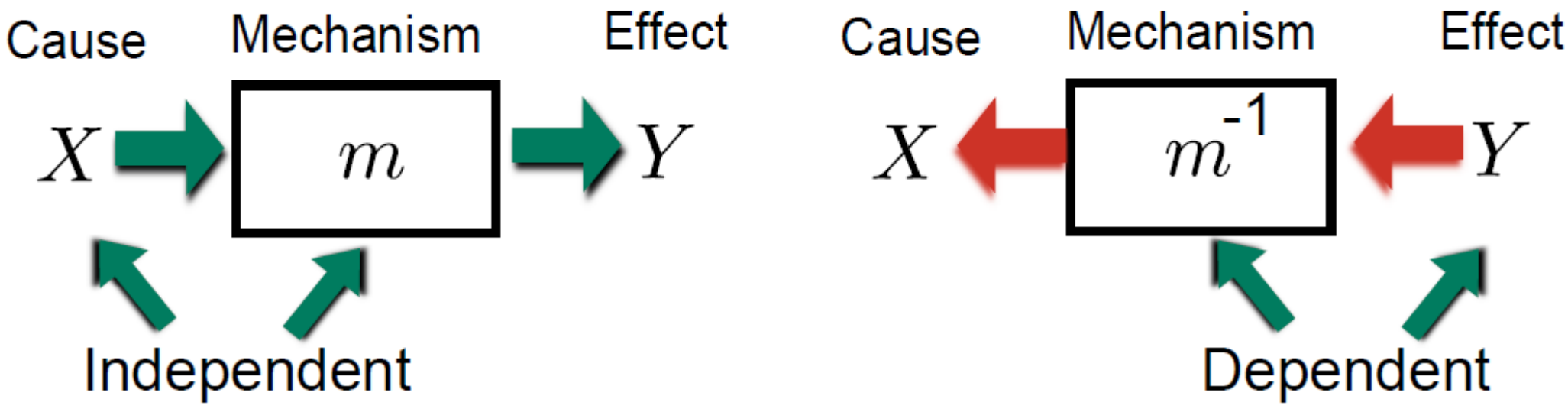
Granger causality



Exploiting statistical properties of the system noise to improve identifiability of such time series models (linear or non-linear) is under active development (Peters et al., Zhang et al.). We suggest a different approach, based on *non-statistical* properties.

Using Independence of cause and mechanism

- Postulate: Nature chooses the cause independent from the mechanism (Daniusis et al. 2010; Janzing et al., 2010; Zscheischler et al., 2011).



Stationary time series

Basic tool: the discrete time Fourier transform

$$\widehat{a}(\nu) = \sum_{t\in\mathbb{Z}} a_t \exp(-\mathbf{i}2\pi\nu t),\ \nu \in [-1/2, 1/2]$$

Assume $\mathbf{X} = \{X_t, t \in \mathbb{Z}\}$ is a weakly stationary process implies

$$\mathbb{E}[X_t X_{t+\tau}] = C_{XX}(\tau),$$

called the *autocovariance function* of the process.

Then, we can define its *Power Spectral Density* (PSD) as

$$S_{XX}(\nu) = \widehat{C_{XX}}(\nu).$$

The power of the process is $P(\mathbf{X}) = \mathbb{E}(|X_t|^2) = \int_{-1/2}^{1/2} S_{XX}(\nu)d\nu$.

S_{XX} really represents the distribution of signal power in the frequency domain.

Example: i.i.d. (white) noise $C_{XX}(\tau) = \delta(\tau)$, thus $S_{XX} = \widehat{C_{XX}} = \mathbf{1}_{[-1/2,1/2]}$.

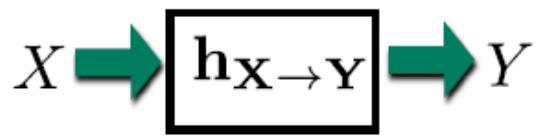
Linear dynamical system

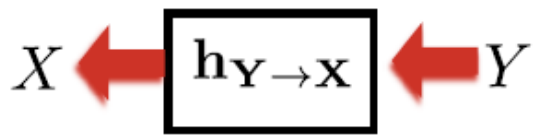
Assume $\mathbf{X}$ is to input of a linear time invariant (LTI) system with impulse response function $\mathbf{h}_{\mathbf{X}\to\mathbf{Y}}$ s.t.

$$\mathbf{Y} = \{\sum_{\tau\in\mathbb{Z}} X_{t-\tau} h_\tau\} = \mathbf{X} * \mathbf{h}_{\mathbf{X}\to\mathbf{Y}}.$$

Then the input output relationship holds in the Fourier domain:

$$S_{yy}(\nu) = |\widehat{h}_{\mathbf{X}\to\mathbf{Y}}(\nu)|^2 S_{xx}(\nu)$$





If the filter is invertible, the same relationship holds for the backward model with the backward impulse response s.t.

$$\widehat{h}_{\mathbf{X}\to\mathbf{Y}} = \frac{1}{\widehat{h}_{\mathbf{Y}\to\mathbf{X}}}.$$

Can we identify the true data generating process, although everything looks very symmetric?

Spectral Independence Criterion (SIC)

Postulate: the distribution of power across frequencies of the input is independent of the amplifying factor of the mechanism, such that

$$\langle S_{yy} \rangle = \langle S_{xx} |\widehat{h}|^2 \rangle = \langle S_{xx} \rangle \langle |\widehat{h}|^2 \rangle ,$$

where $\langle f \rangle = \int_{-1/2}^{1/2} f(\nu) d\nu$.

This is a measure of correlation:

$$\langle S_{xx} \cdot |\widehat{h}|^2 \rangle - \langle S_{xx} \rangle \langle |\widehat{h}|^2 \rangle = \mathbb{C}\mathrm{ov} \left(S_{xx}, |\widehat{h}|^2 \right) .$$

Formulation using input and output statistics:

$$\langle S_{yy} \rangle = \langle S_{xx} \rangle \langle S_{yy} / S_{xx} \rangle .$$

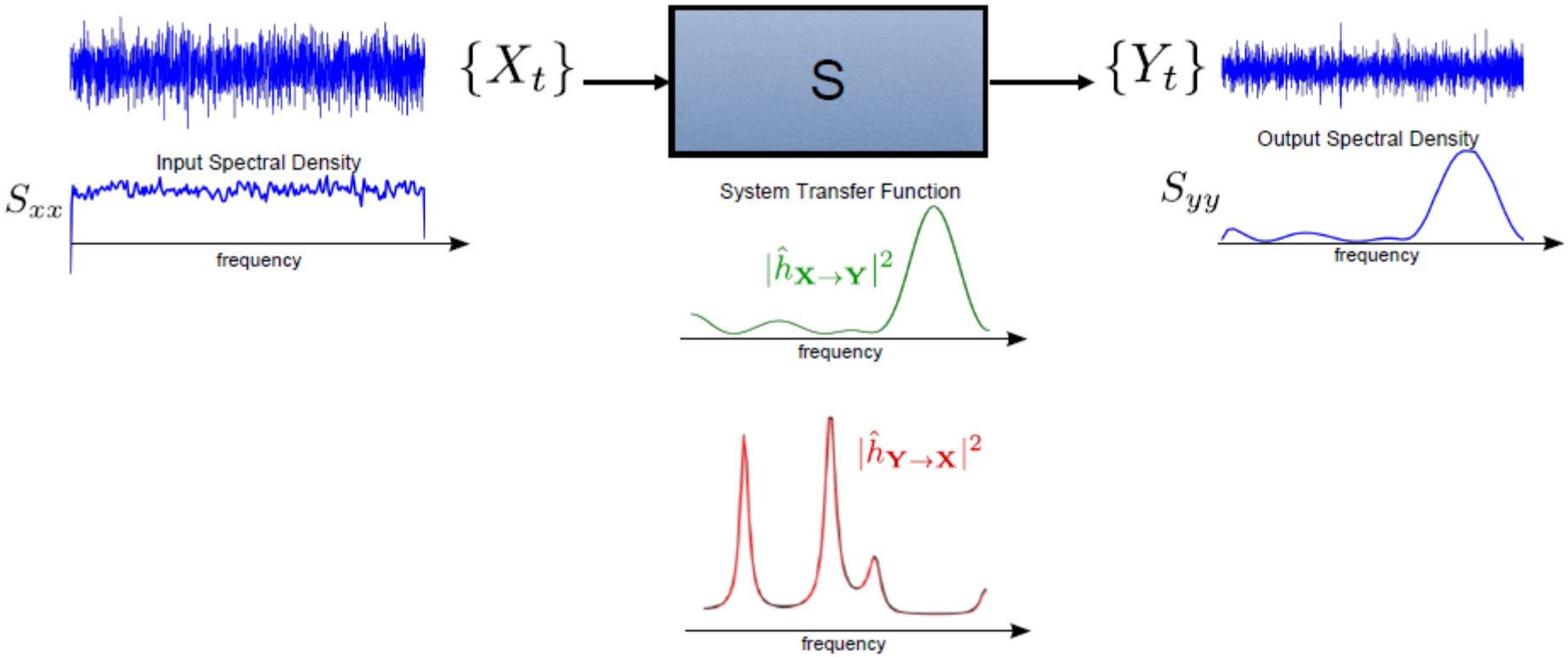
Spectral density ratio (SDR)

$$\rho_{\mathbf{X}\to\mathbf{Y}} := \frac{\langle S_{yy} \rangle}{\langle S_{xx} \rangle \langle S_{yy}/S_{xx} \rangle}$$

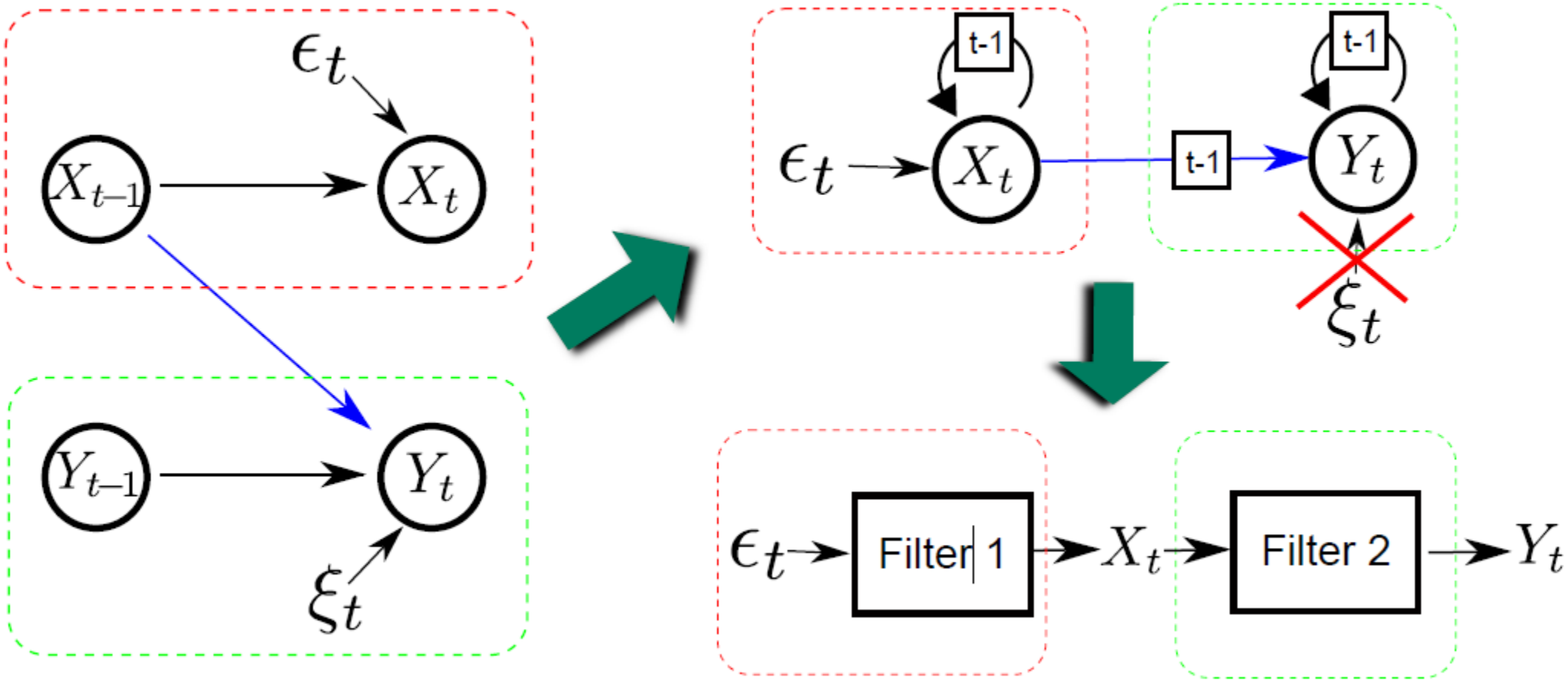
Interpretation:

- $\rho_{\mathbf{X}\to\mathbf{Y}}$=1: Spectral Independence
- $\rho_{\mathbf{X}\to\mathbf{Y}} < 1$: Negative correlation
- $\rho_{\mathbf{X}\to\mathbf{Y}} > 1$: Positive correlation.

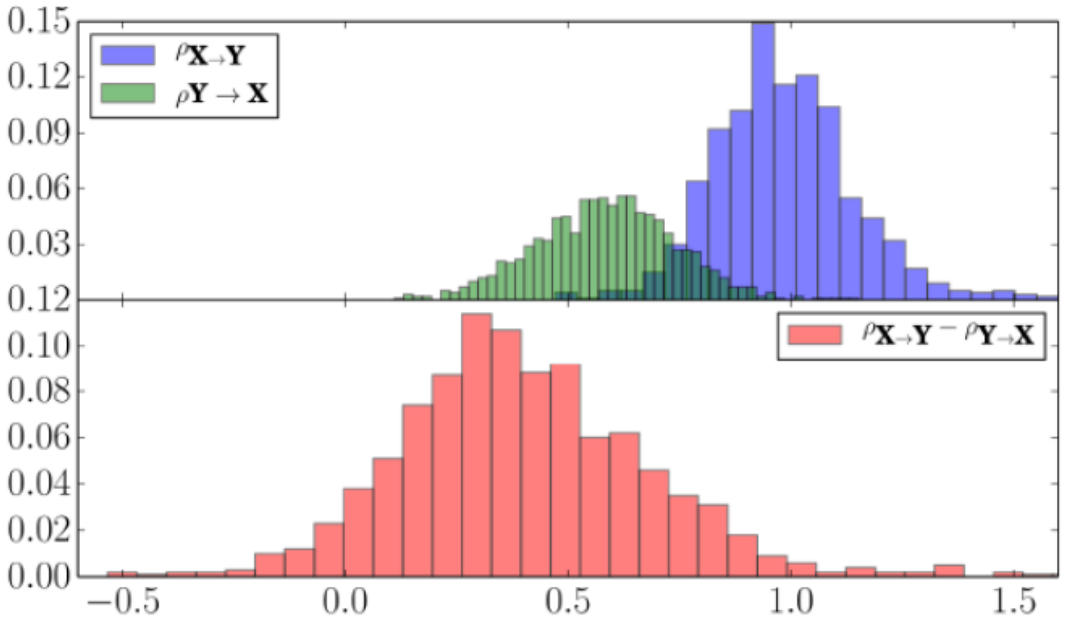
Intuition

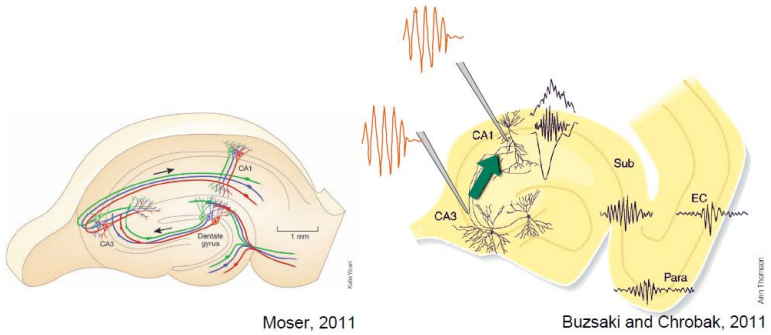

$\{X_t\}$
S
$\{Y_t\}$
Input Spectral Density
S_{xx}
frequency
System Transfer Function
$|\hat{h}_{\mathbf{X}\to\mathbf{Y}}|^2$
frequency
Output Spectral Density
S_{yy}
frequency
$|\hat{h}_{\mathbf{Y}\to\mathbf{X}}|^2$
frequency

Link to autoregressive models



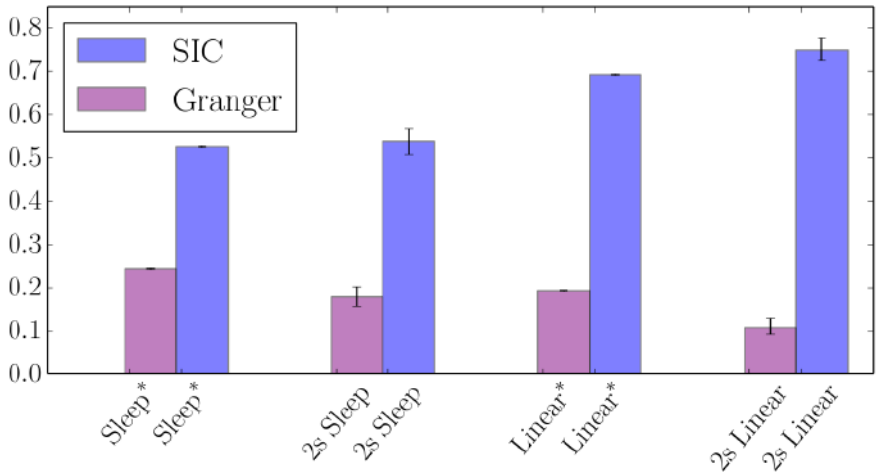
Experiments





Moser, 2011

Buzsaki and Chrobak, 2011



Conclusion

- Convergence of ML and Neuroscience questions.
- ML offers tools to elaborate hypotheses about brain function and test them with data.
- Neuroscience offers principles that might be exploited to build “stronger” AI.
- Causality and generative models are exciting directions for such developments.