The Biology of Consciousness

Christof Koch

Allen Institute for Brain Science

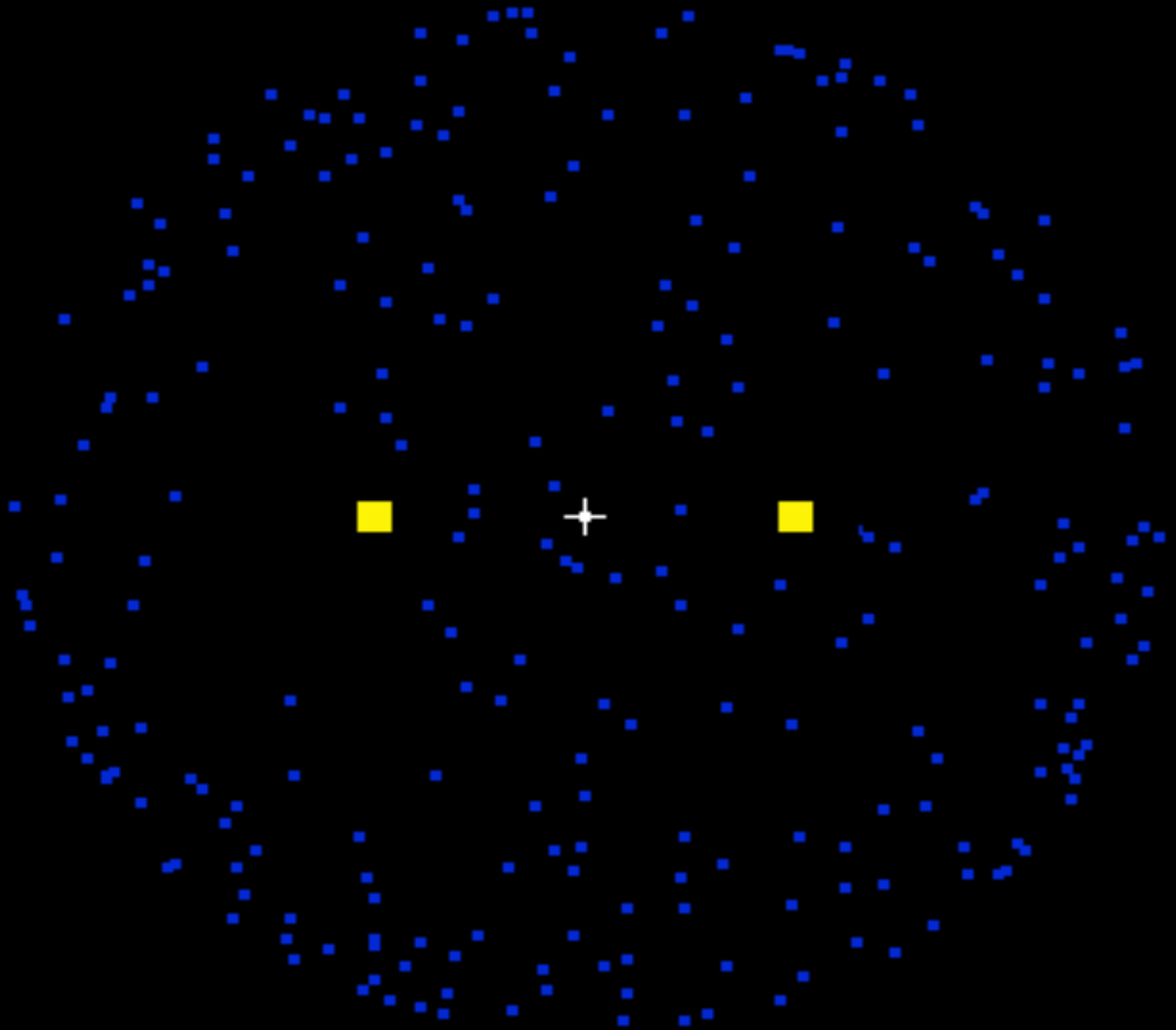
1/8/2014

Cartesian Certainty

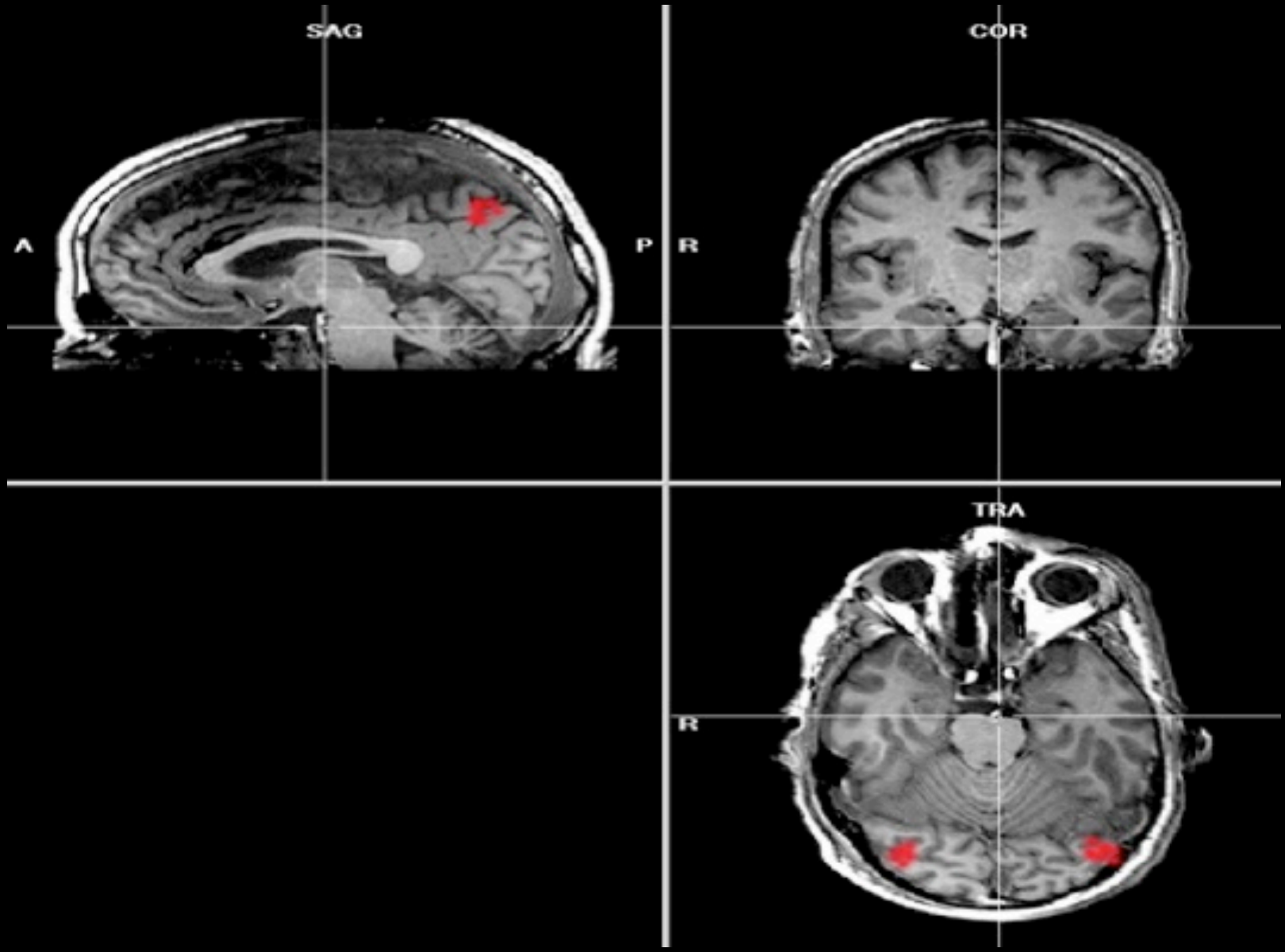
Je pense, donc je suis

or, in modern language,

I am conscious, therefore I am



Difference between brains and other things



Difference between brains and other things

- There are external observables (sensory-motor behavior, neurons, action potentials, molecules etc) - third person account
- However, there is also an unique internal perspective to a brain - first person account

The passage from the physics of the brain to the corresponding facts of consciousness is unthinkable as a result of mechanics. Granted that a definite thought, and a definite molecular action in the brain, occur simultaneously; we do not possess the intellectual organ, nor apparently any rudiment of the organ, which would enable us to pass, by a process of reasoning, from the one phenomenon to the other. They appear together, but we do not know why. Were our minds and senses so expanded, strengthened, and illuminated, as to enable us to see and feel the very molecules of the brain; were we capable of following all their motions, all their groupings, all their electric discharges, if such there be; and were we intimately acquainted with the corresponding states of thought and feeling, we should be as far as ever from the solution of the problem, “How are these physical processes connected with the facts of consciousness?” The chasm between the two classes of phenomena would still remain intellectually impassable. Let the consciousness for love, for example, be associated with a right-handed spiral motion of the molecules of the brain, and the consciousness of hate with a left-handed spiral motion. We should then know, when we love, that the motion is in one direction, and, when we hate, that the motion is in the other; but the “WHY?” would remain as unanswerable as before.

John Tyndall (1886)

The Hard Problem

The really hard problem of consciousness is the problem of experience. Why is it that when our [brains] engage in visual and auditory
information-processing, we have visual or
auditory experience…?

David Chalmers (1995)

What do we know about C?

- C is associated with some complex, adaptive, biological networks (not immune system nor enteric nervous system)
- C does not require behavior
- C does not require emotions
- C does not require language nor self-consciousness
- C does not require long-term memory
- C does not require selective attention
- C can occur in one cerebral hemisphere
- Destruction of localized brain regions interferes with specific content of C

Many Brains Inside Your Head

Many - if not most - behaviors occur in the absence of conscious sensations, or consciousness occurs after the fact:

- Spinal reflexes
- Posture adjustments
- Any over-trained routine: Shaving, dressing, tennis, video games, keyboard typing, driving, rock-climbing, dancing
- Reaching and grabbing
- Generating speech
- Dissociation between what the eyes see and conscious perception
- High-level decision making (e.g. choice blindness)

Behavioral Correlates of Consciousness (BCC)

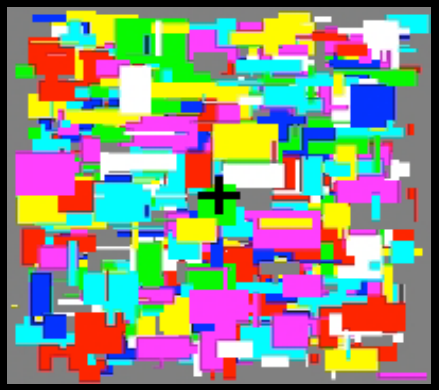
Empirically, certain behaviors are associated with consciousness

- Purposeful behavior in response to spoken commands
- Glasgow Coma Scale (3-15)
- Meaningful linguistic contents
- Non-stereo-typed, temporal-delayed sensory-motor behavior
- Meta-cognition

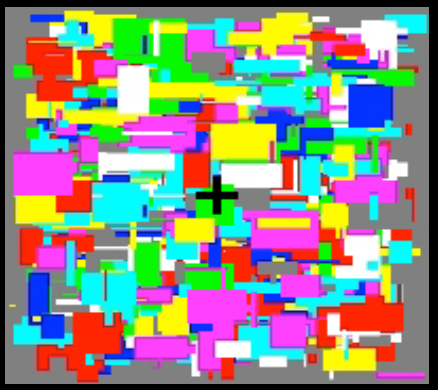
Continuous Flash Suppression

Left eye

Right eye

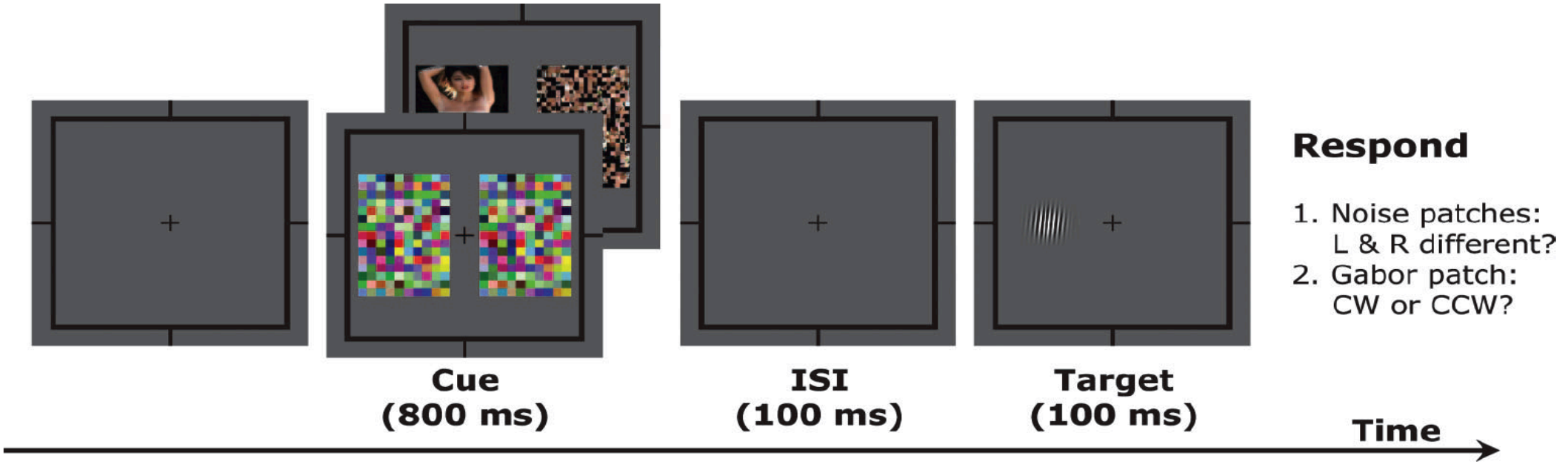


Percept

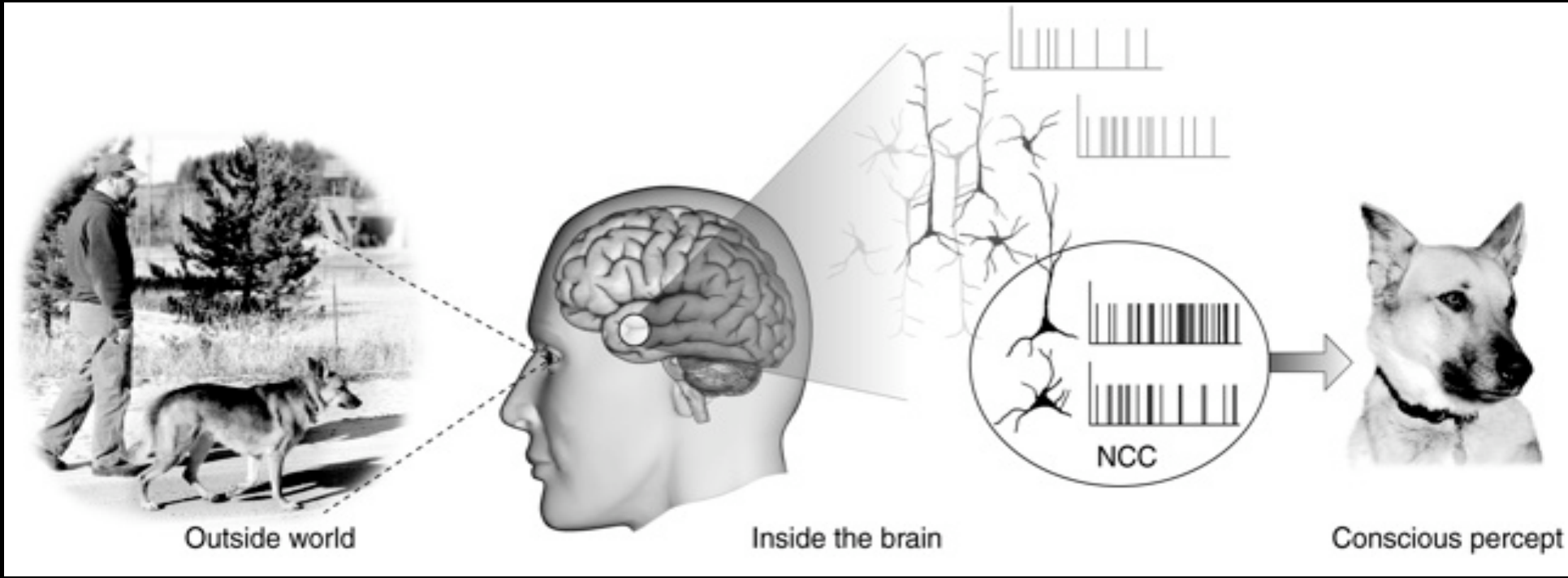


Tsuchiya & Koch (2005)

Looking at invisible nudes



Jiang et al. (2006)

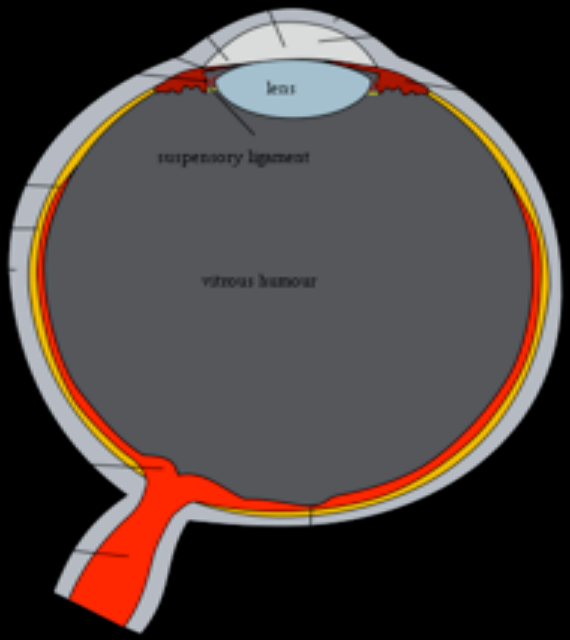


Search for the minimal neuronal mechanisms jointly sufficient for any one conscious perception, the neuronal correlates of consciousness (NCC)

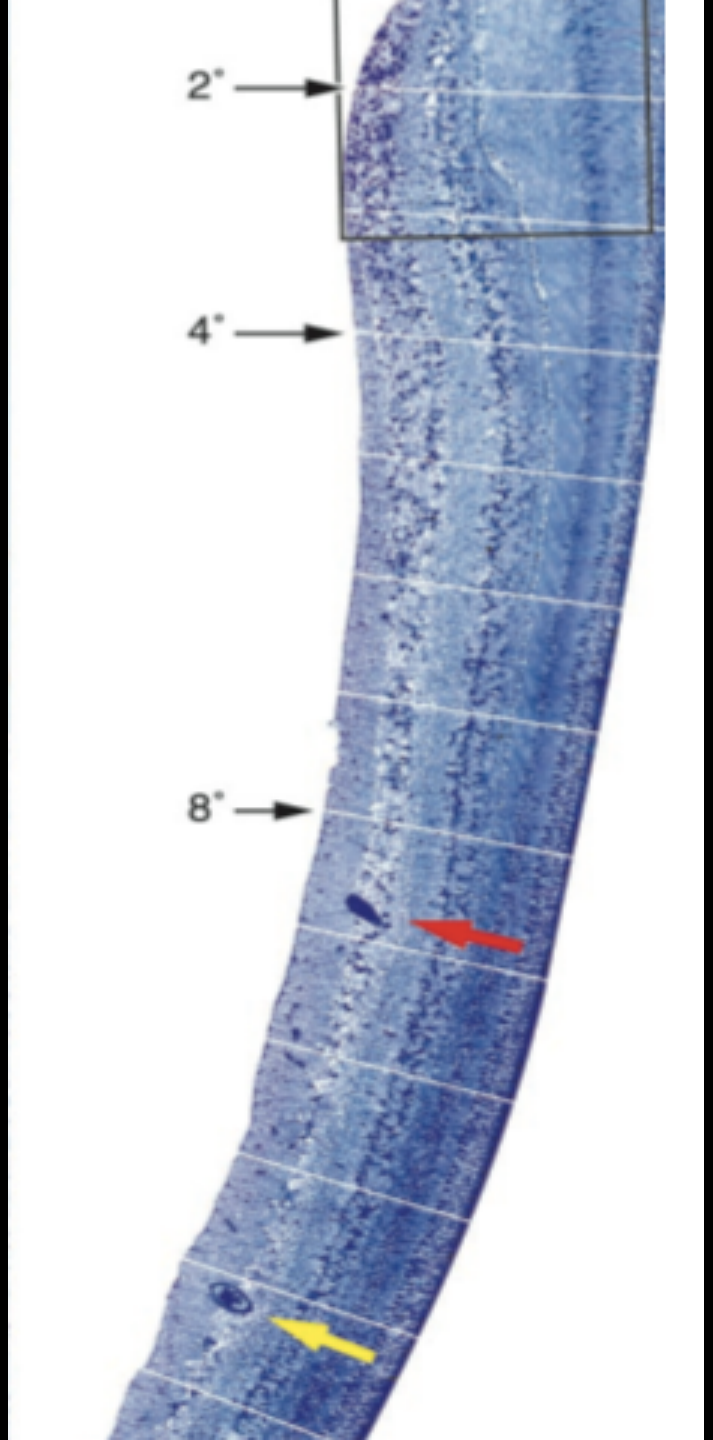
For every conscious percept, there will be a NCC

Crick & Koch (*Nature* 1995)

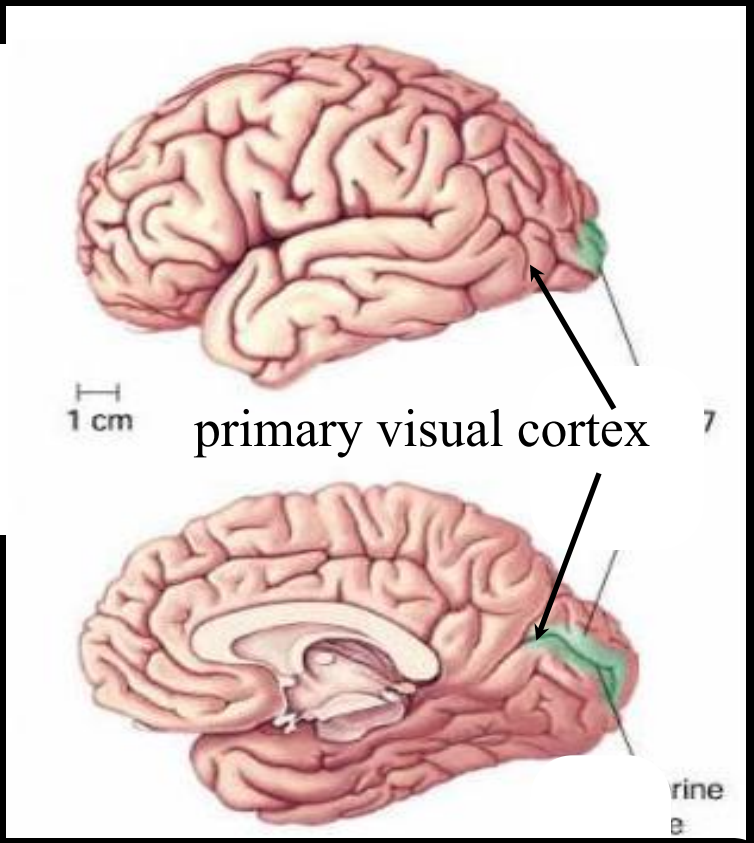
The Eye



The retina is not part of the NCC



Visual Cortex

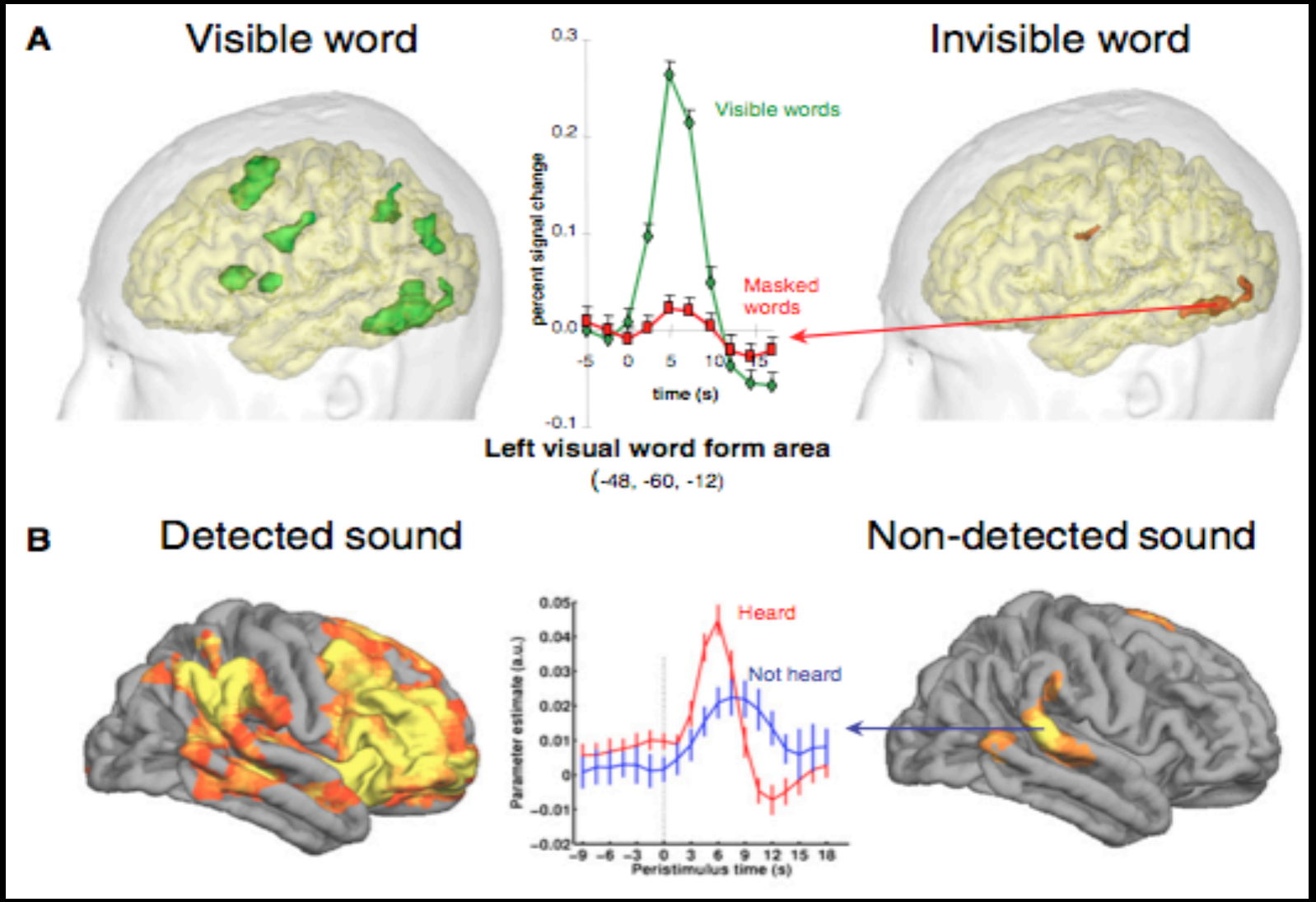


The NCC does not reside in primary visual cortex

Cerebellum

- 69 out of 86 billion neurons are in the cerebellum
- Main deficit of cerebellar lesions are ataxia, slurred speech and unsteady gait
- The cerebellum is not a significant part of the NCC

A NCC in Frontal-Parietal Cortical Structures

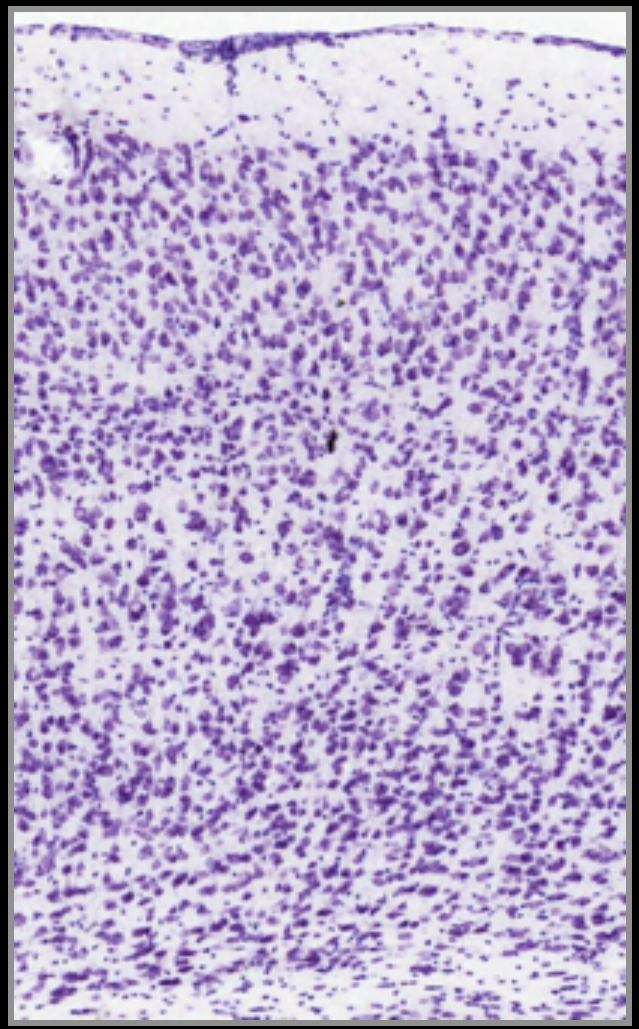


Dehaene & Changeux (2012)

Consciousness in Other Mammals

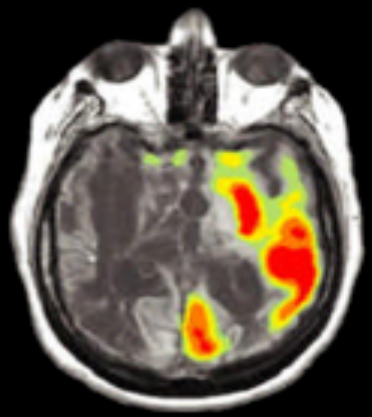
- Similarity of behavior
- Similarity of brain architecture
- Close evolutionary kinship
- The main specialization of *homo sapiens* is a highly developed self-consciousness and language

Temporal Cortex



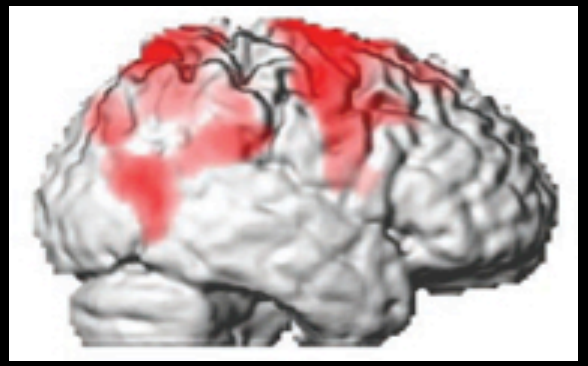
Hard Calls

Brain “islands” in a vegetative subject



Fetus, pre-term & newborn infant

Ketamine anesthesia



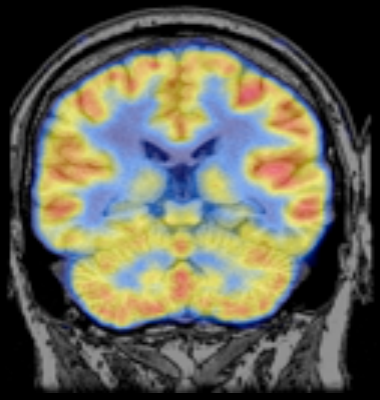
Sleepwalking

Octopus

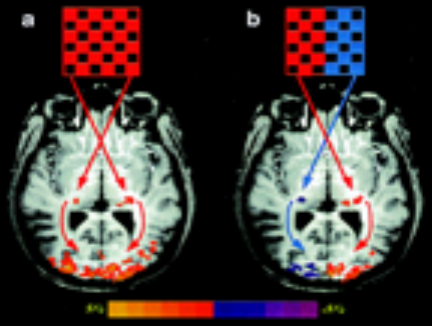
Apple Siri

More Hard Questions

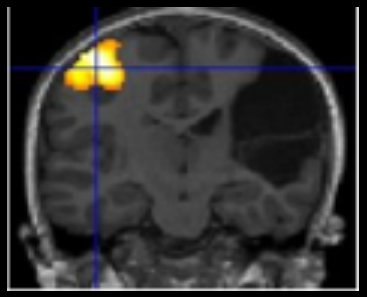
Why not the cerebellum?



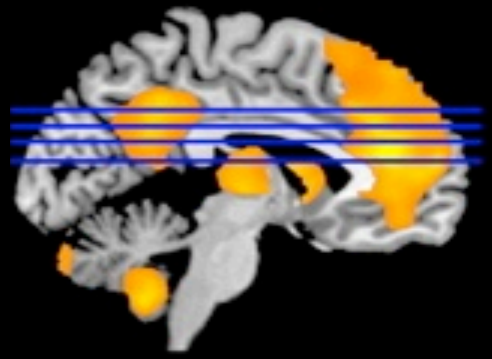
Why not afferent pathways?



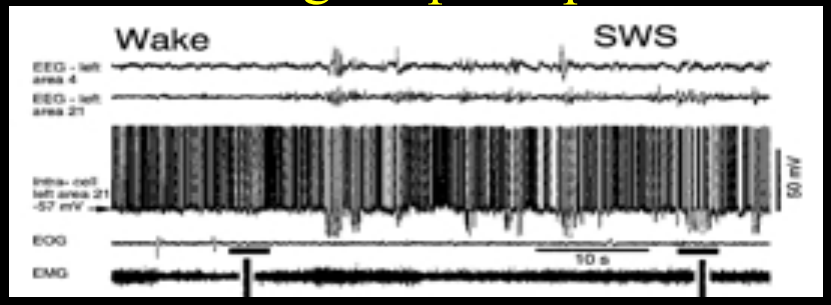
Why not efferent pathways?



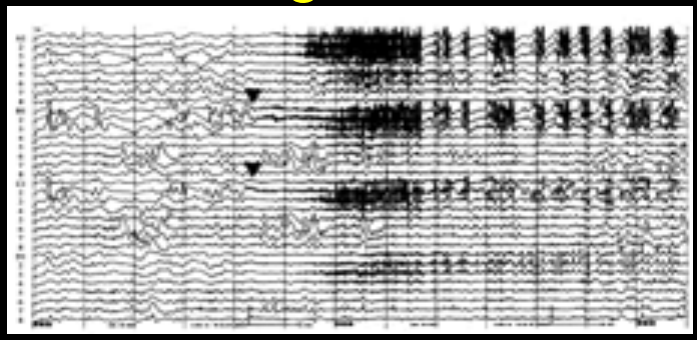
Why not cortico-subcortico-cortical loops?



Why not the cortex during deep sleep?



Why not the cortex during seizures?



==>

or

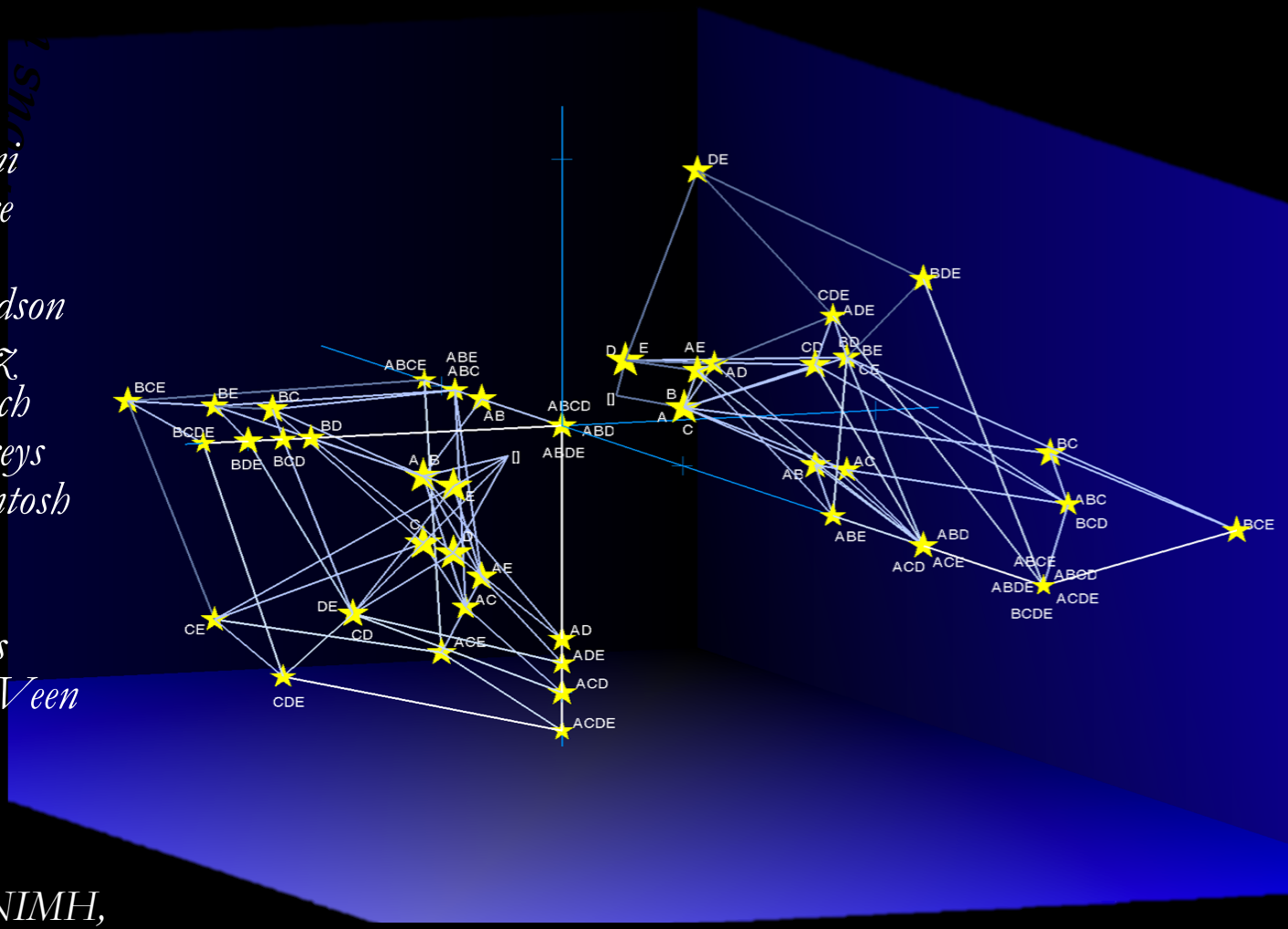
<==

Going from BCC/NCC to Consciousness is hard; so let's go the opposite way

From Phenomenology to Mechanisms, and Back: An Integrated Information Theory of Consciousness

Larissa Albantakis
David Balduzzi
Melanie Boly
Chiara Cirelli
Daniela Dentico
Fabio Ferrarelli
Olivia Gosseries
Atif Hashmi
Erik Hoel
Matteo Mainetti
Marcello Massimini
Andy Nere
Yuval Nir
Masafumi Oizumi
Umberto Olcese
Ben Shababo
Francesca Siclari

Chris Adami
Mike Alkire
Ruth Benca
Richie Davidson
Tony Hudetz
Christof Koch
Steven Laureys
Randy McIntosh
Bob Pearce
Brad Postle
Olaf Sporns
Barry Van Veen



Support: NIH Director's Pioneer Award, NIMH, NINDS, DARPA, Paul Allen Foundation, McDonnell Foundation, University of Wisconsin

Disclosures: Consultant for Philips

Giulio Tononi

Check your biases

- Consciousness = experience
  rather than awareness of environment, of self, or reflective awareness
- IIT starts from consciousness itself (phenomenology)
  rather than its behavioral correlates (BCC) / neural correlates (NCC)
- Information is intrinsic (differences that make a difference within a system):
  how a set of mechanisms in a state constrains,
  i.e. informs, its past and future states
  rather than extrinsic (Shannon): how an observer can decode inputs
  from outputs of a channel
- Integration = irreducibility: what the whole does above its parts
  rather than convergence onto a place

Axioms:

Identifying the essential properties of consciousness

Existence

Experience exists (**intrinsically**, independent of external observers)

Composition

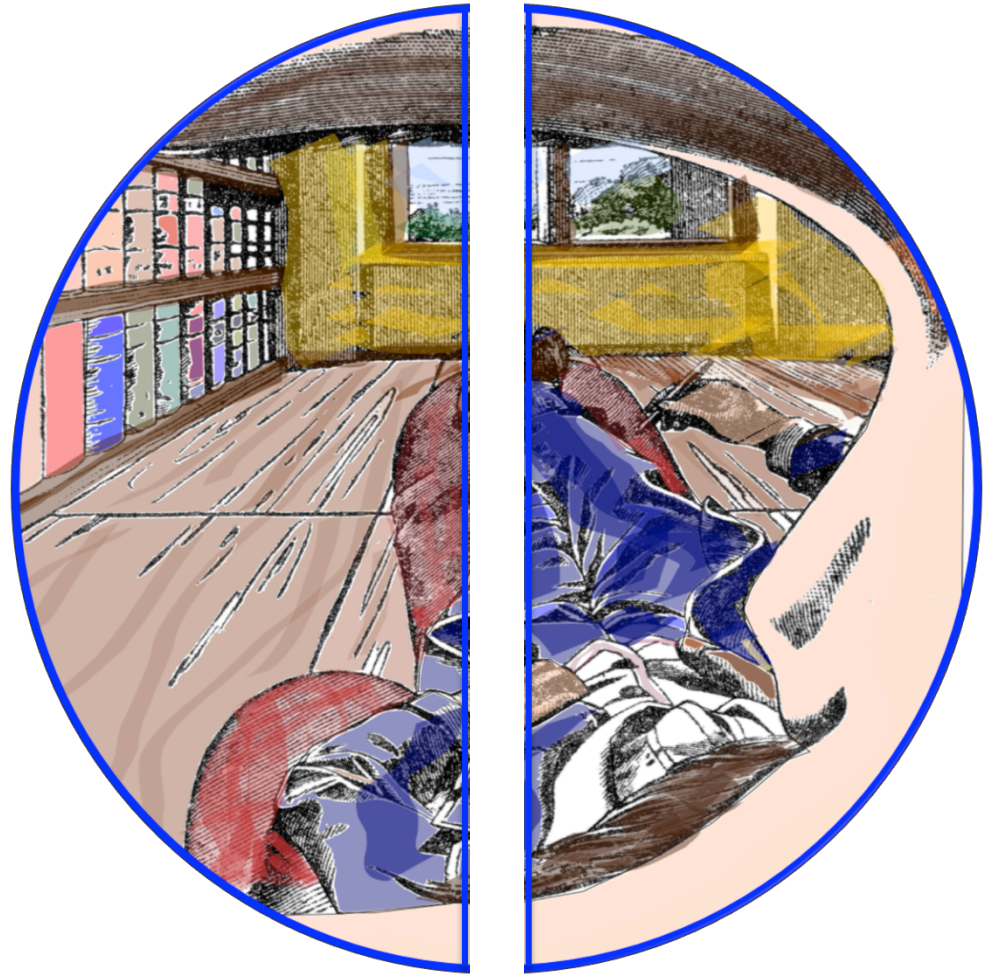
Experience is **structured** (it has **many aspects**)

Information

Experience is **differentiated** (**one out of many**):

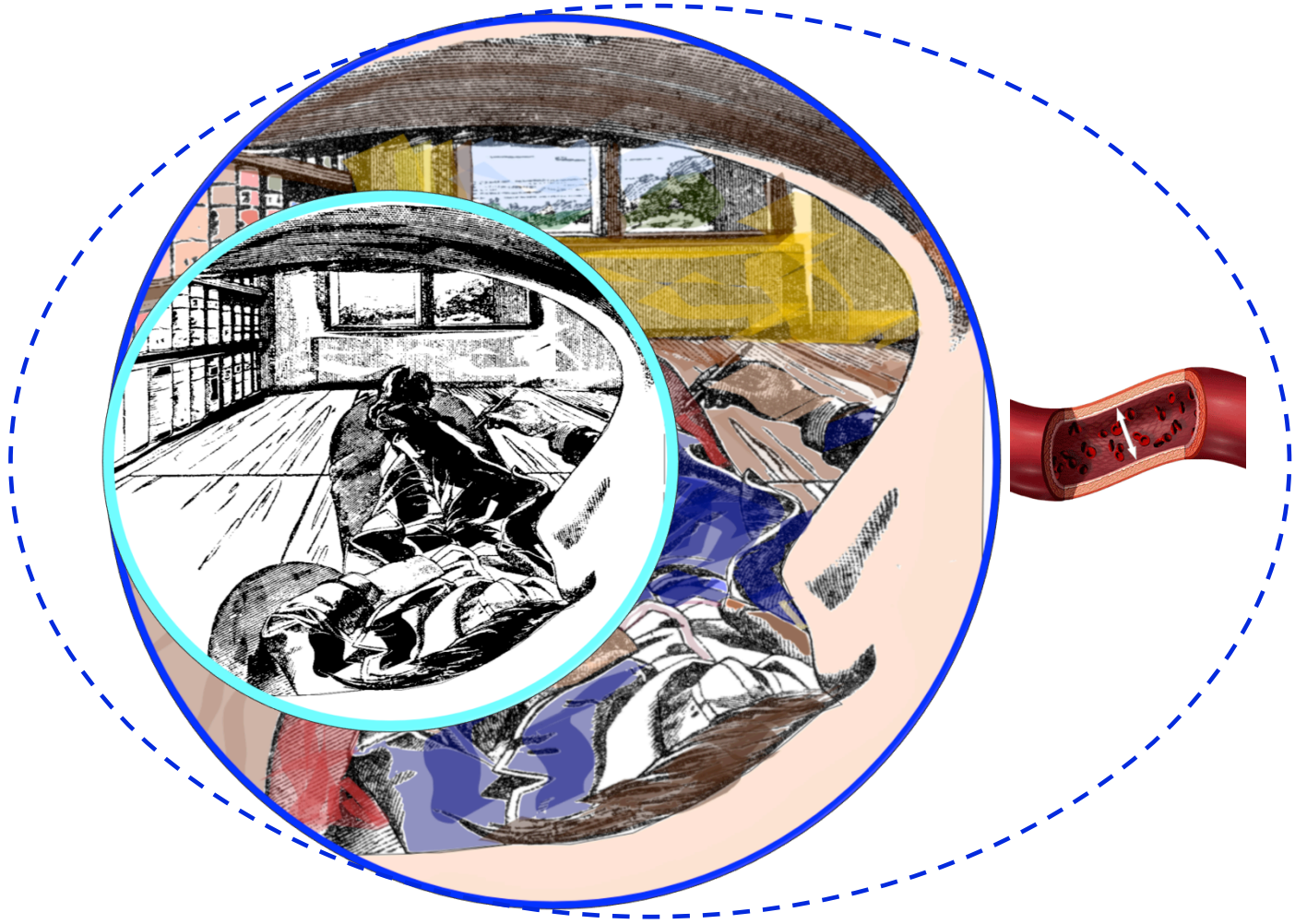
it is what it is by differing in its particular way from many others

Integration



Experience is **unified** (it is "**one**"):
it cannot be reduced to non-interdependent components

Exclusion



Experience is **unique** (it is **only one**), in content and spatio-temporal grain:
it is not a superposition of multiple experiences, with less or more content, flowing at faster or slower speed at once

Postulates:

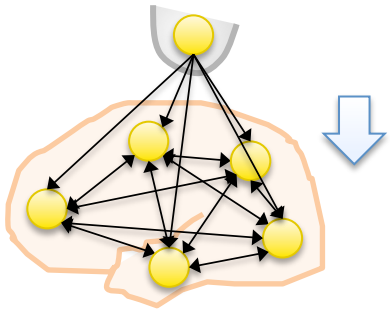
Identifying the requisites for the physical substrate of consciousness

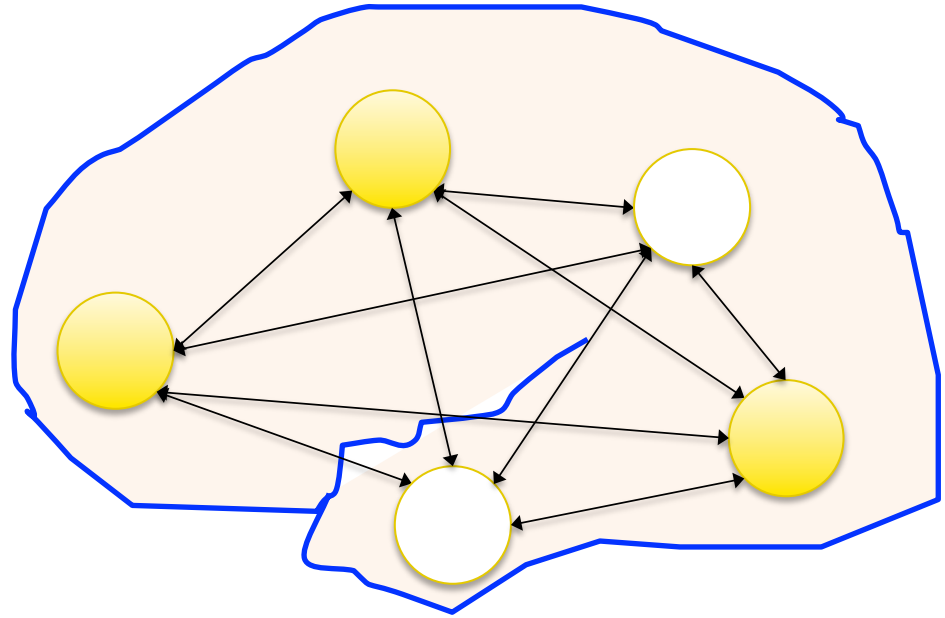
Existence

Experience exists (**intrinsically**, independent of external observers)

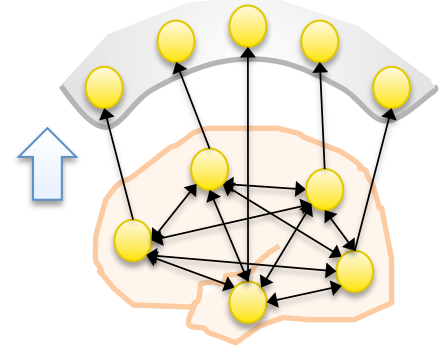
Existence

ECT (electroconvulsive) shock





fMRI scanner

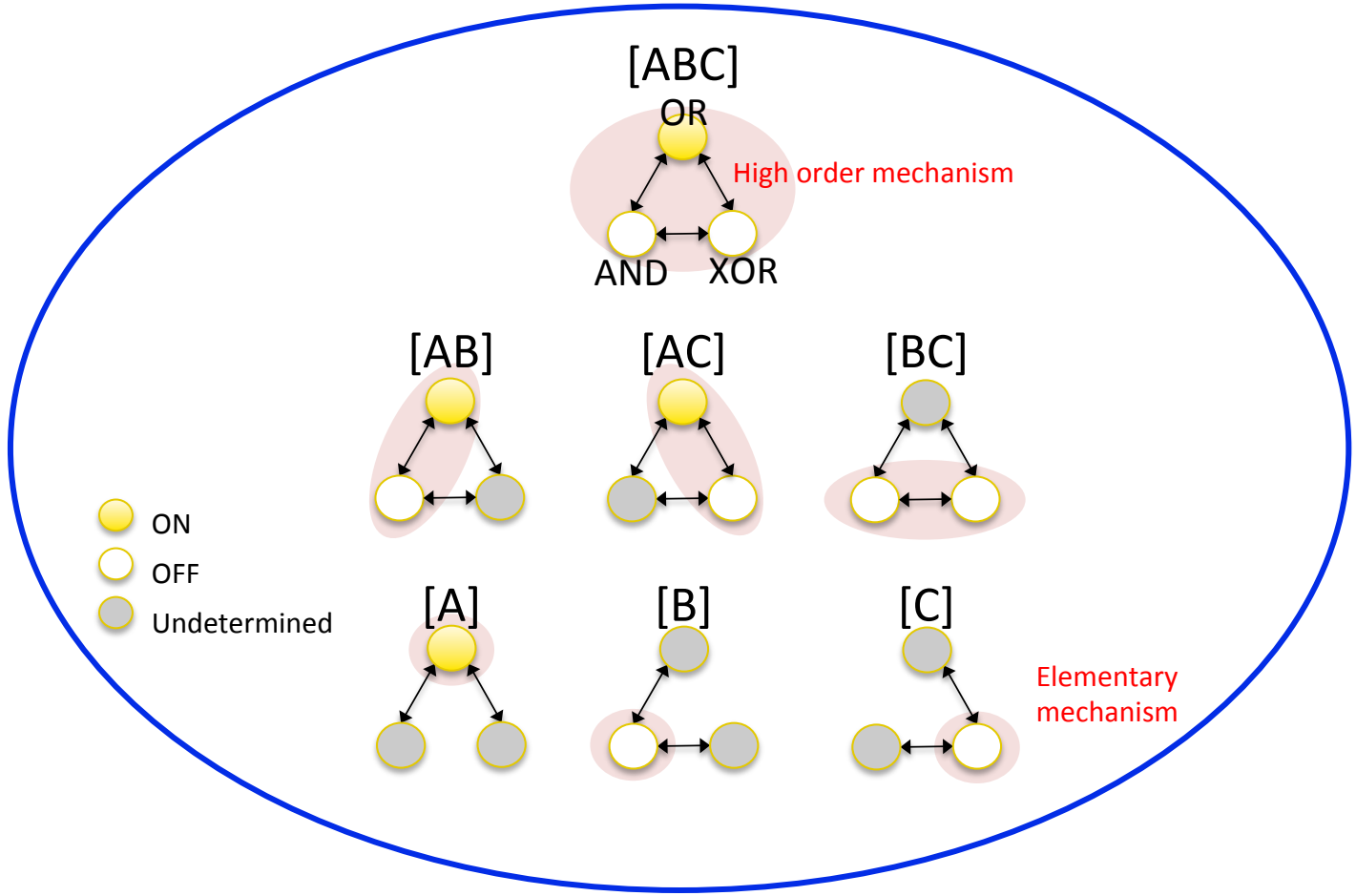


Experience is generated by a **system of mechanisms**:
to exist, the mechanisms must have cause-effect power
("**differences that make a difference**")
within the system itself (**intrinsically**)

Composition

Experience is **structured** (it has **many aspects**)

Composition



The system can be **structured**:

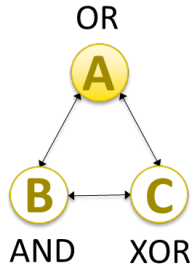
Subsets of the system can contribute **specific aspects** of experience

Information

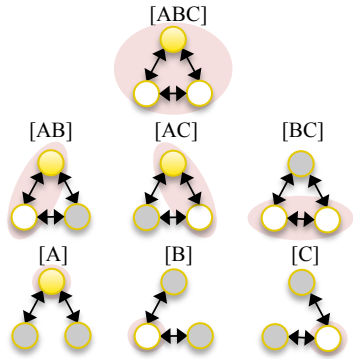
Experience is **differentiated** (**one out of many**):

it is what it is by differing in its particular way from many others

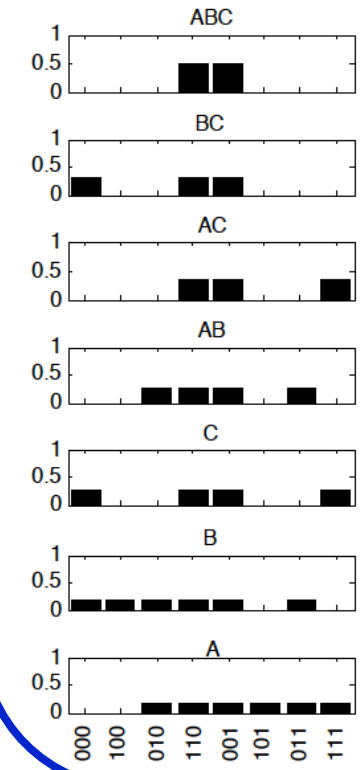
Information

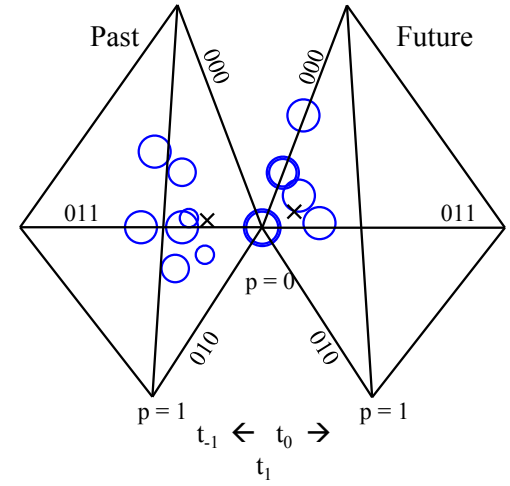


A set of elements
in their current state s_0



Power set of mechanisms



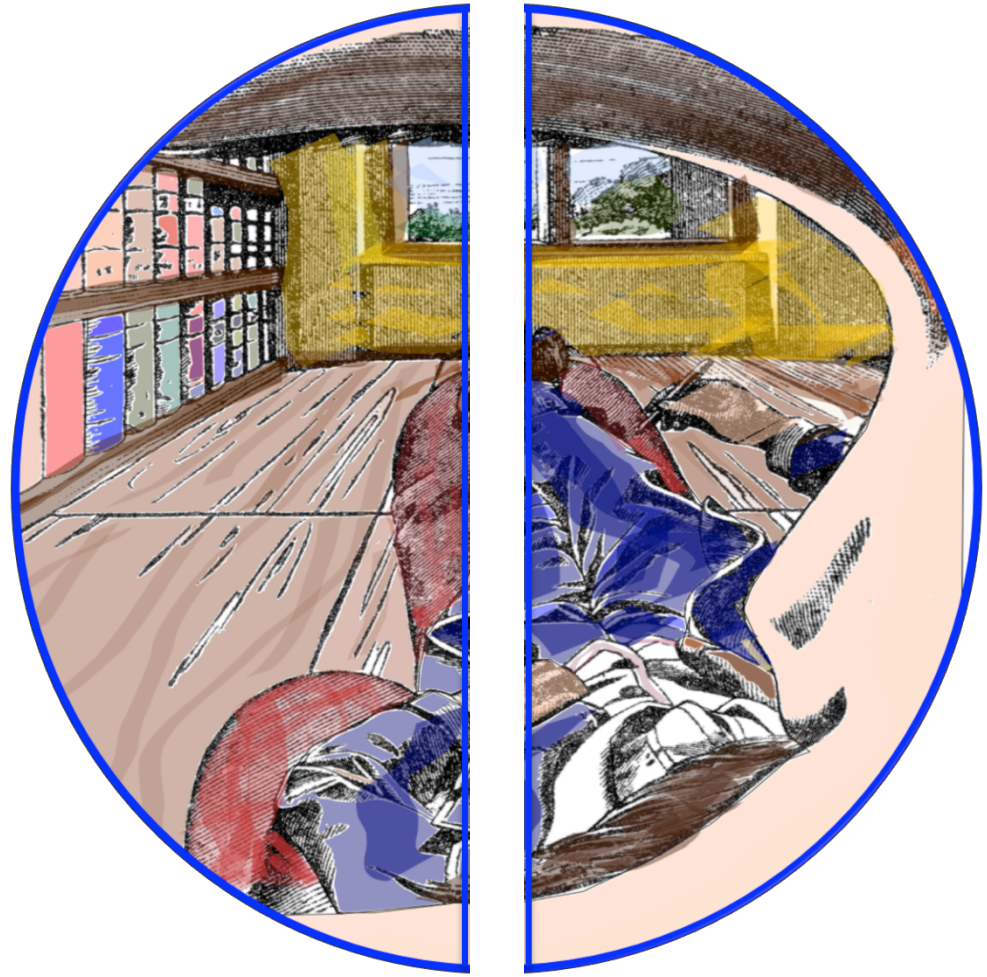


Cause-effect structure
plotted as a constellation
in concept space

based on Oizumi, Albantakis, and Tononi, submitted

The system must be **differentiated**: when it is in a particular state, its mechanisms must **constrain** its past and future states in a particular way – specifying a **cause-effect structure**, made up of **cause-effect repertoires** specified by individual mechanisms

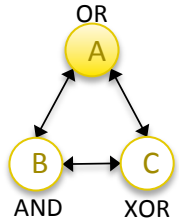
Integration



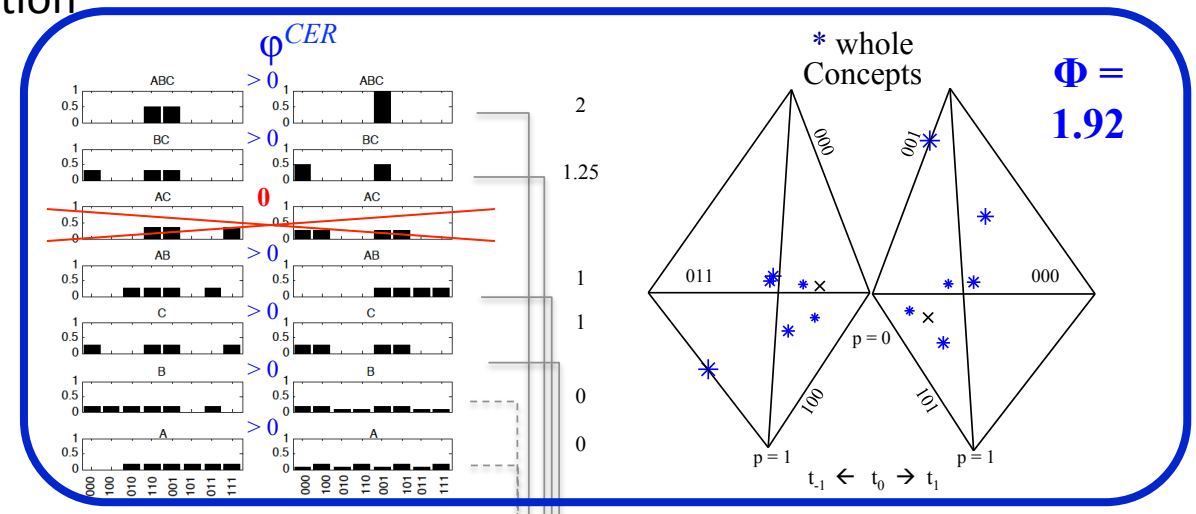
Experience is **unified** (it is "**one**"):
it cannot be reduced to non-interdependent components

Integration

Whole Constellation

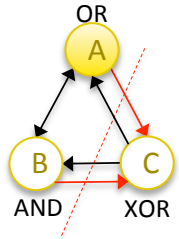


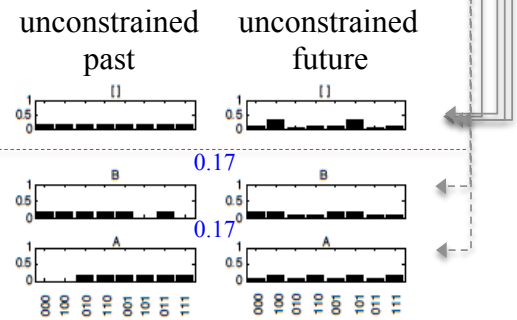
$s_0(ABC) = 100$

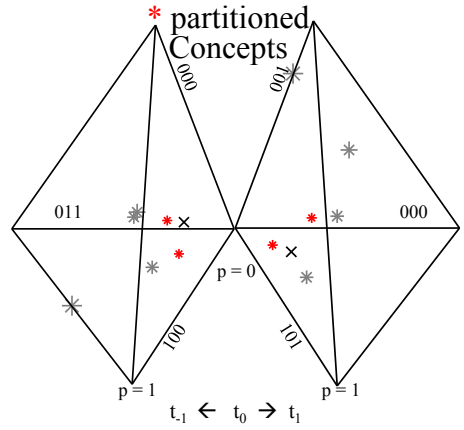


Partitioned Constellation

MIP = minimum information partition, unidirectional







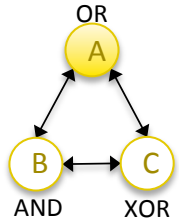
$$\Phi(C|s_0) = \mathrm{D}\left((C|s_0)\middle\|(C^{\rightarrow}_{MIP}|s_0)\right)$$

D: Earth mover's distance

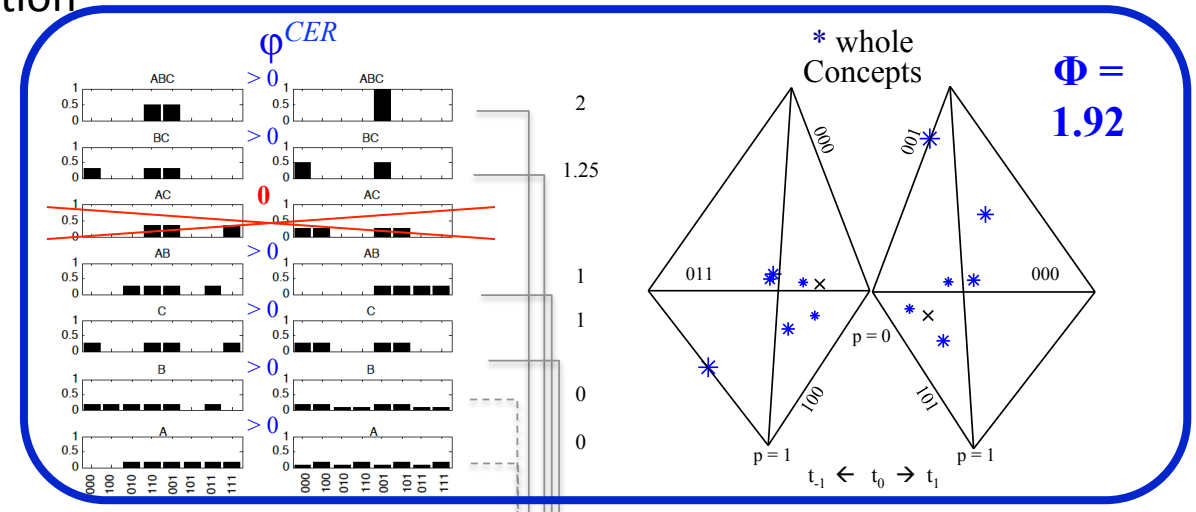
The system must be **unified**: it must be **irreducible** (by a minimum partition MIP) to non-interdependent sub-systems (Φ>0)

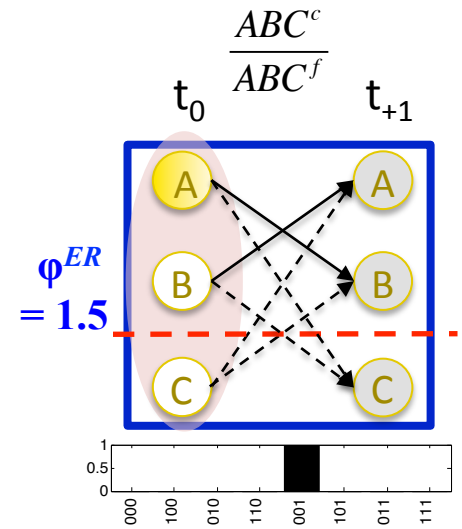
Integration

Whole Constellation



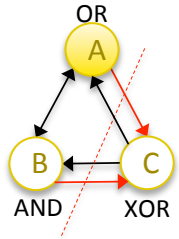
$s_0(ABC) = 100$

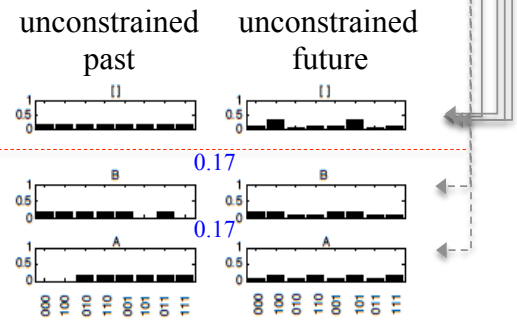


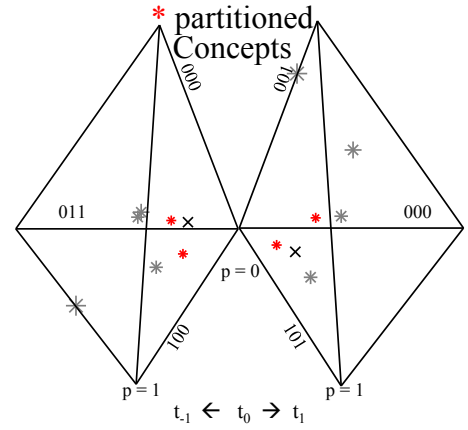


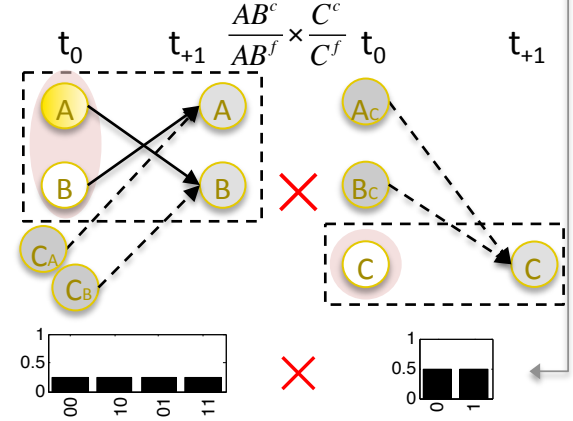
Partitioned Constellation

MIP = minimum information partition, unidirectional









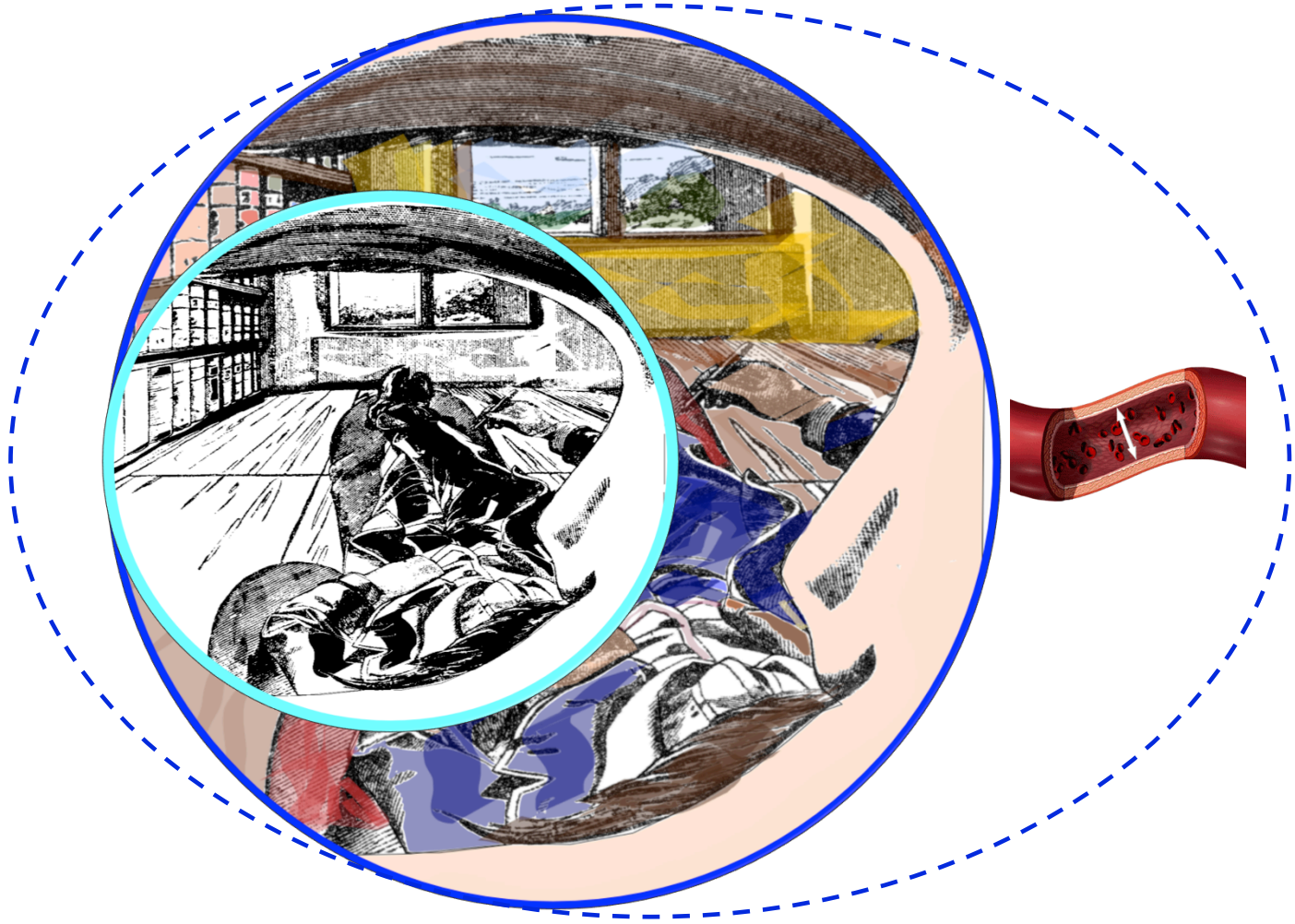
$$\Phi(C|s_0) = \mathrm{D}\left((C|s_0)\middle\|(\vec{C_{MIP}}|s_0)\right)$$

D: Earth mover's distance

$$\varphi(CER|s_0) = \mathrm{D}\left(p(CER|s_0)\middle\|p(CER_{MIP}|s_0)\right)$$

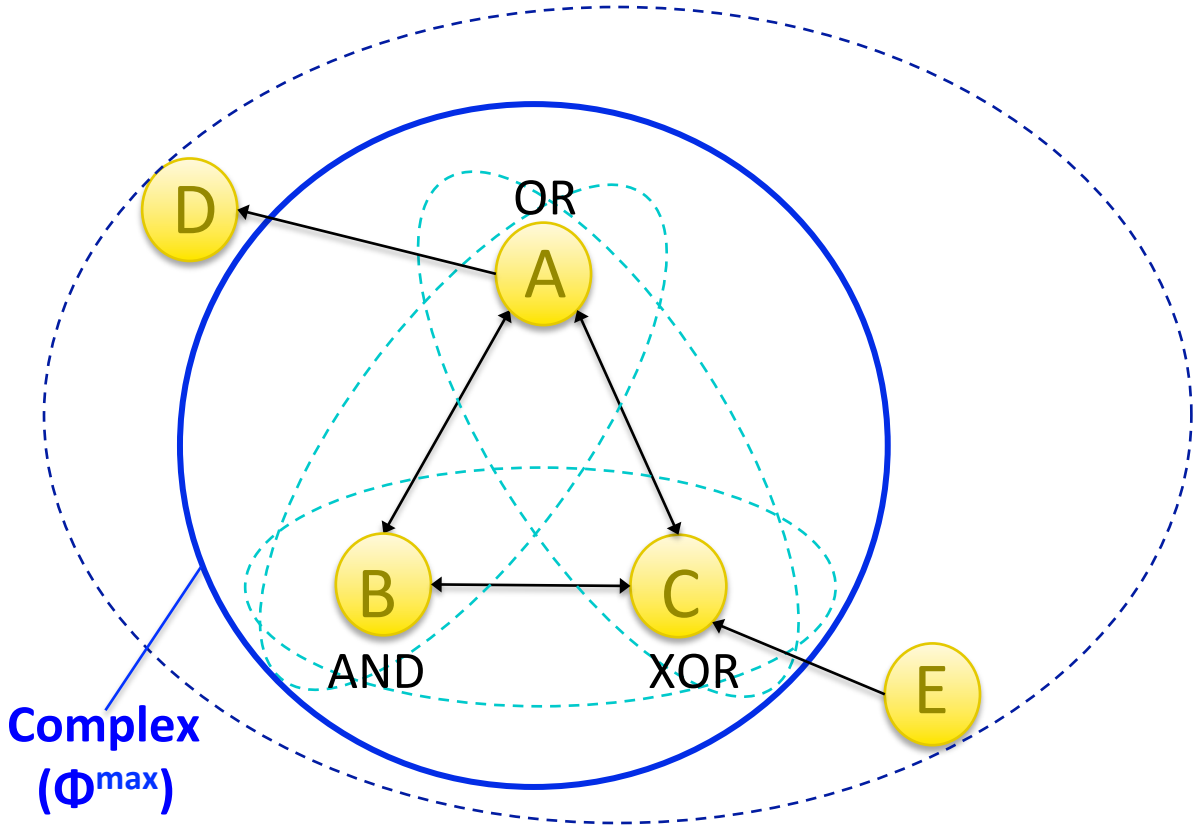
The system must be **unified**: it must be **irreducible** (by a minimum partition MIP) to non-interdependent sub-systems (Φ>0) and each mechanism must be irreducible to sub-mechanisms (φ>0)

Exclusion



Experience is **unique** (it is **only one**), in content and spatio-temporal grain:
it is not a superposition of multiple experiences, with less or more content, flowing at faster or slower speed at once

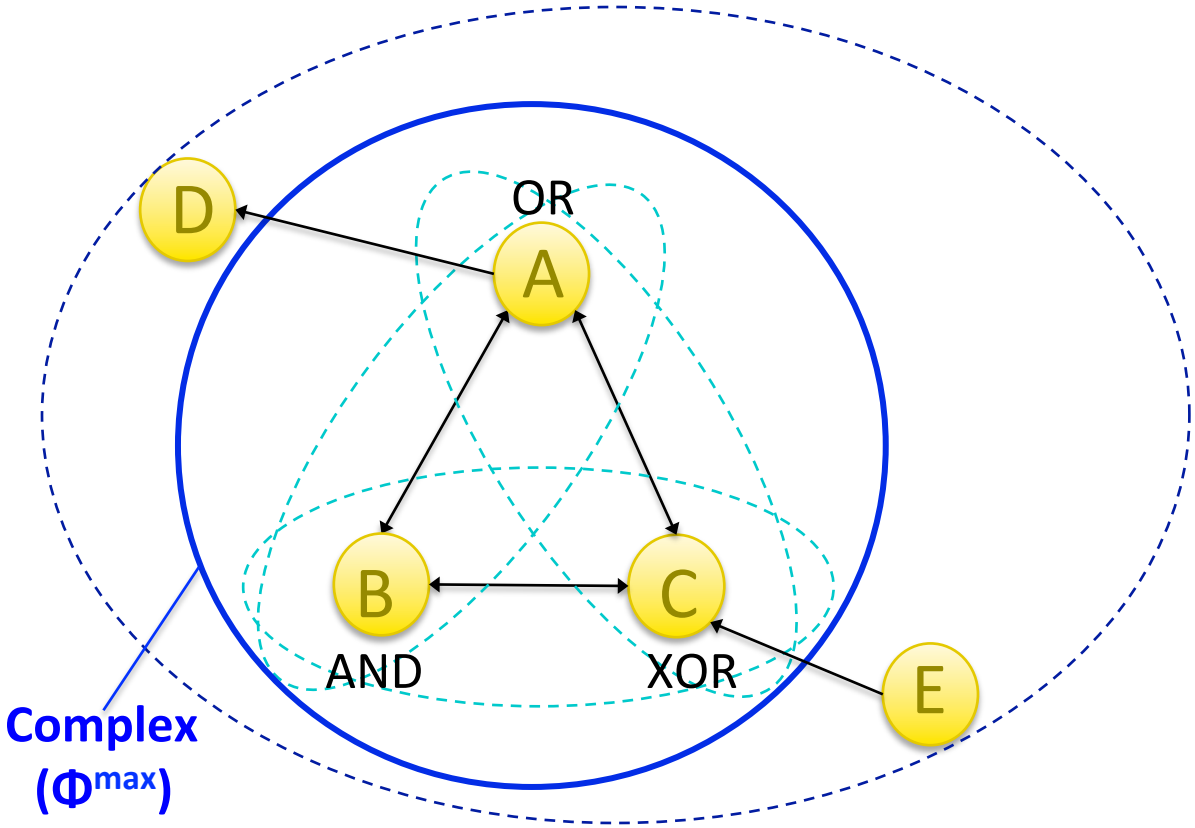
Exclusion

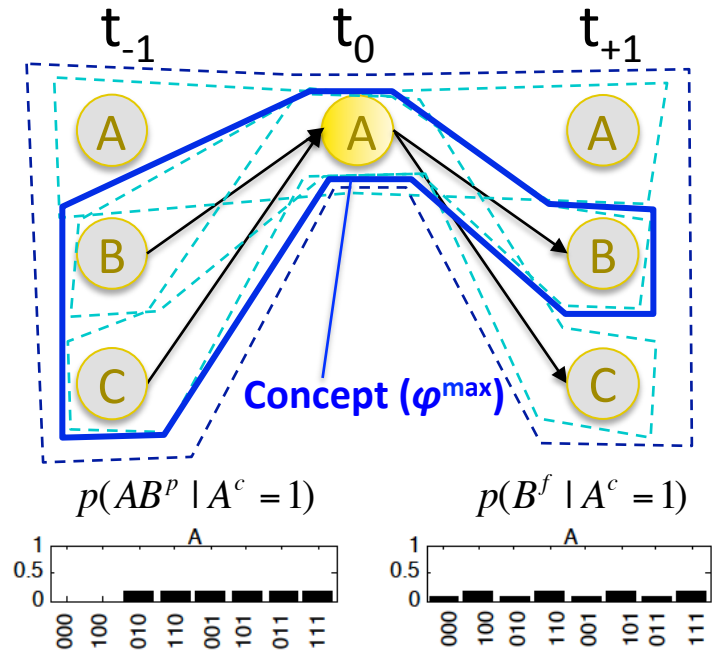


The system must be **unique** over elements and spatio-temporal grain:
it must specify **only one** cause-effect structure, the one that is **maximally irreducible** (Φ^{max})

based on Oizumi, Albantakis, and Tononi, submitted

Exclusion



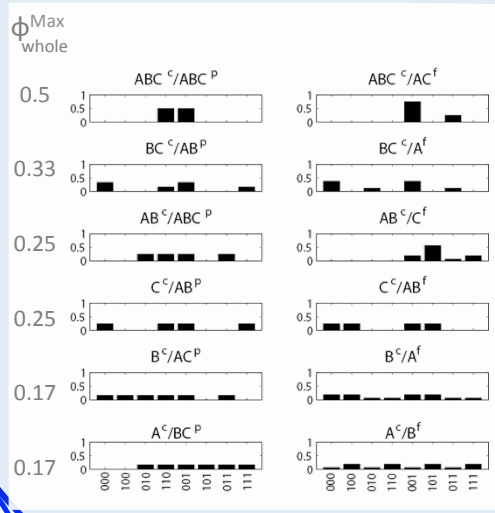


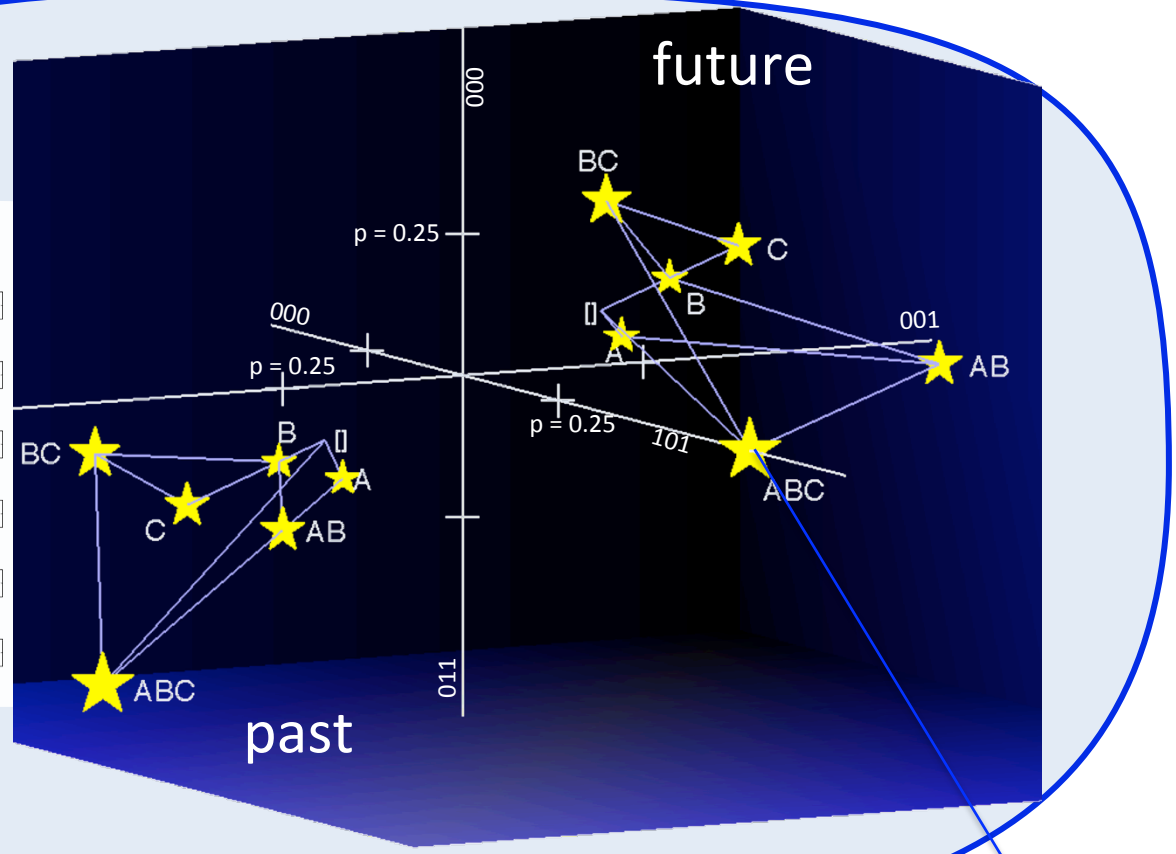
The system must be **unique** over elements and spatio-temporal grain:
it must specify **only one** cause-effect structure, the one that is **maximally irreducible** (Φ^{max})
and each mechanism must specify only one cause-effect repertoire (φ^{max})

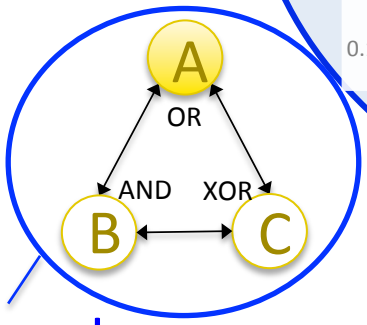
A quale

Quale

$\Phi^{max} = 1.92$







Complex

Concept

$\varphi^{max} = 0.5$

based on Oizumi, Albantakis, and Tononi, submitted

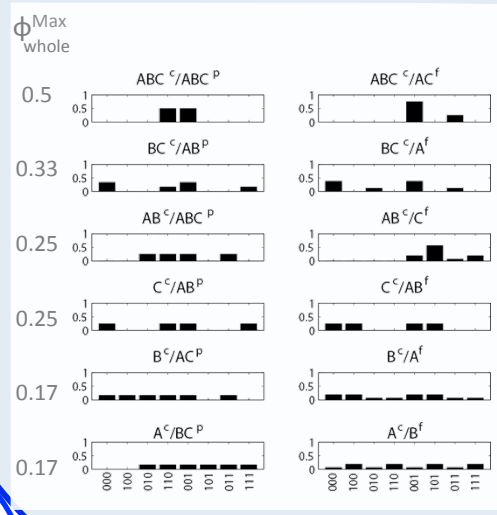
The maximally irreducible conceptual structure generated by a complex

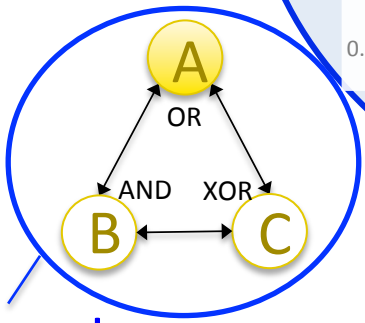
Identity:

An experience is a maximally irreducible conceptual structure

Quale

$\Phi^{max} = 1.92$





Complex

Concept

$\varphi^{max} = 0.5$

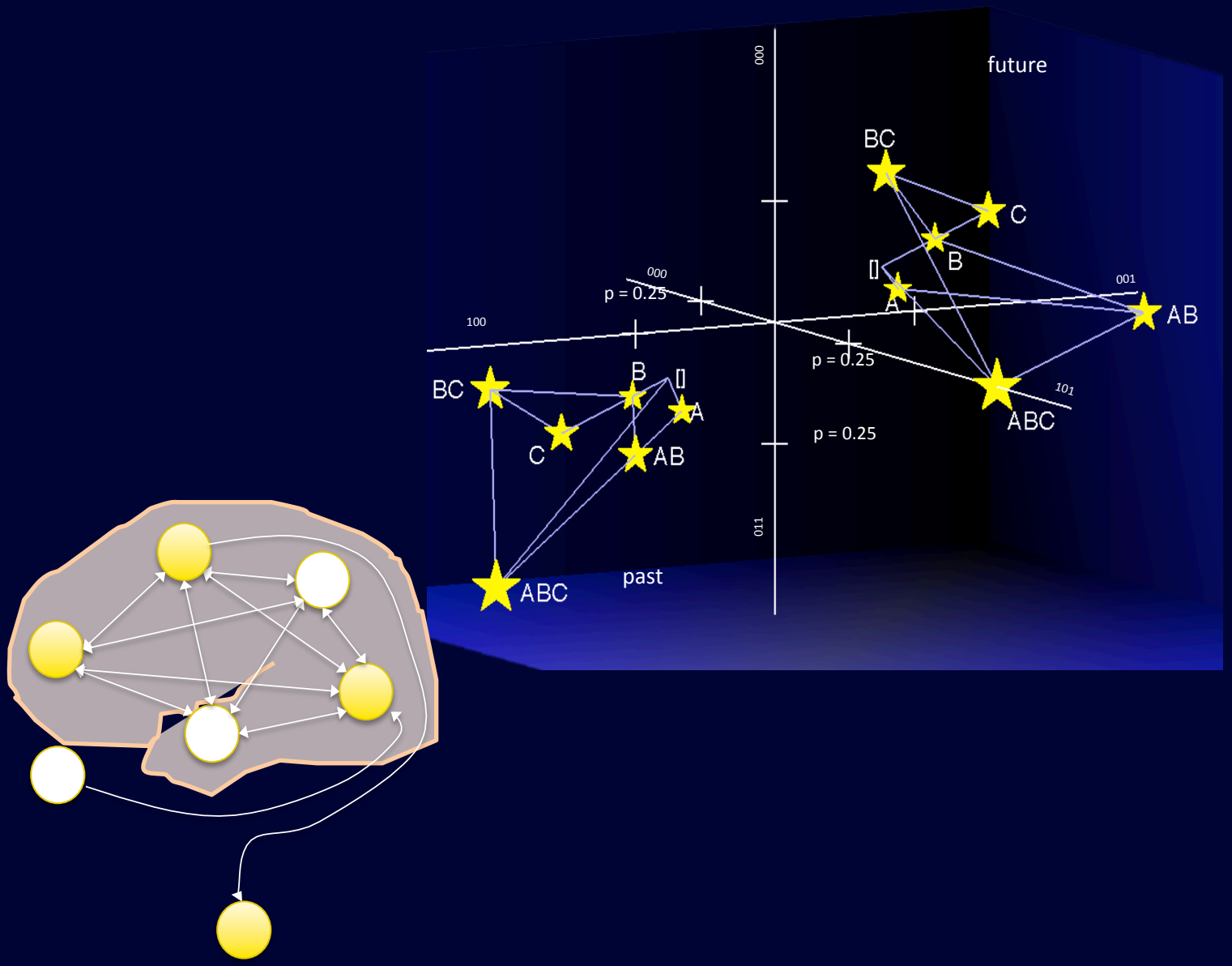
Quantity: Irreducibility of the conceptual structure (Φ^{max})

Quality: "Shape" of the quale in qualia space ("constellation")

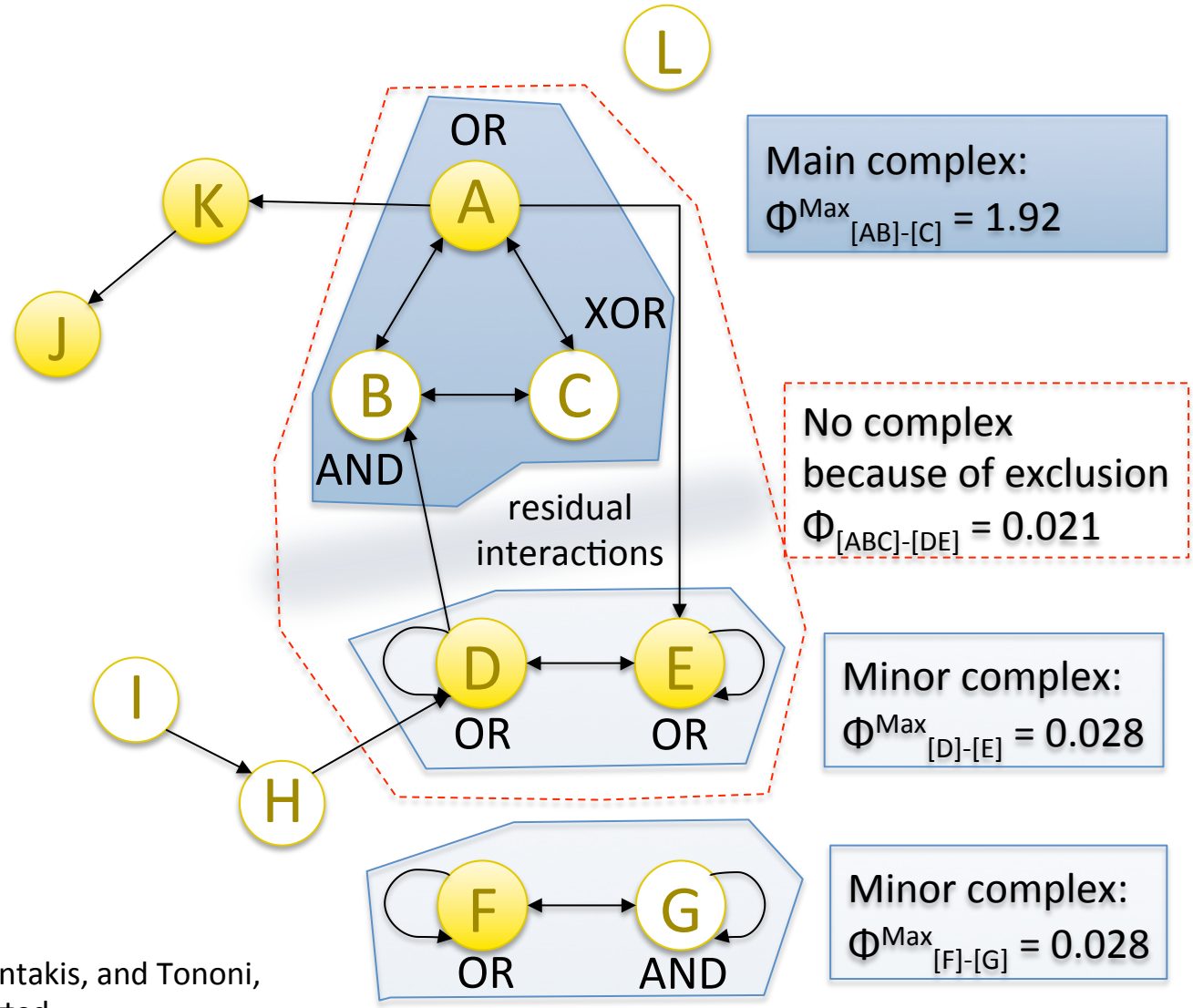
Integrated Information Theory

- **Corollaries**
- **Predictions**
- **Explanations**
- **Extrapolations**

IIT: some corollaries

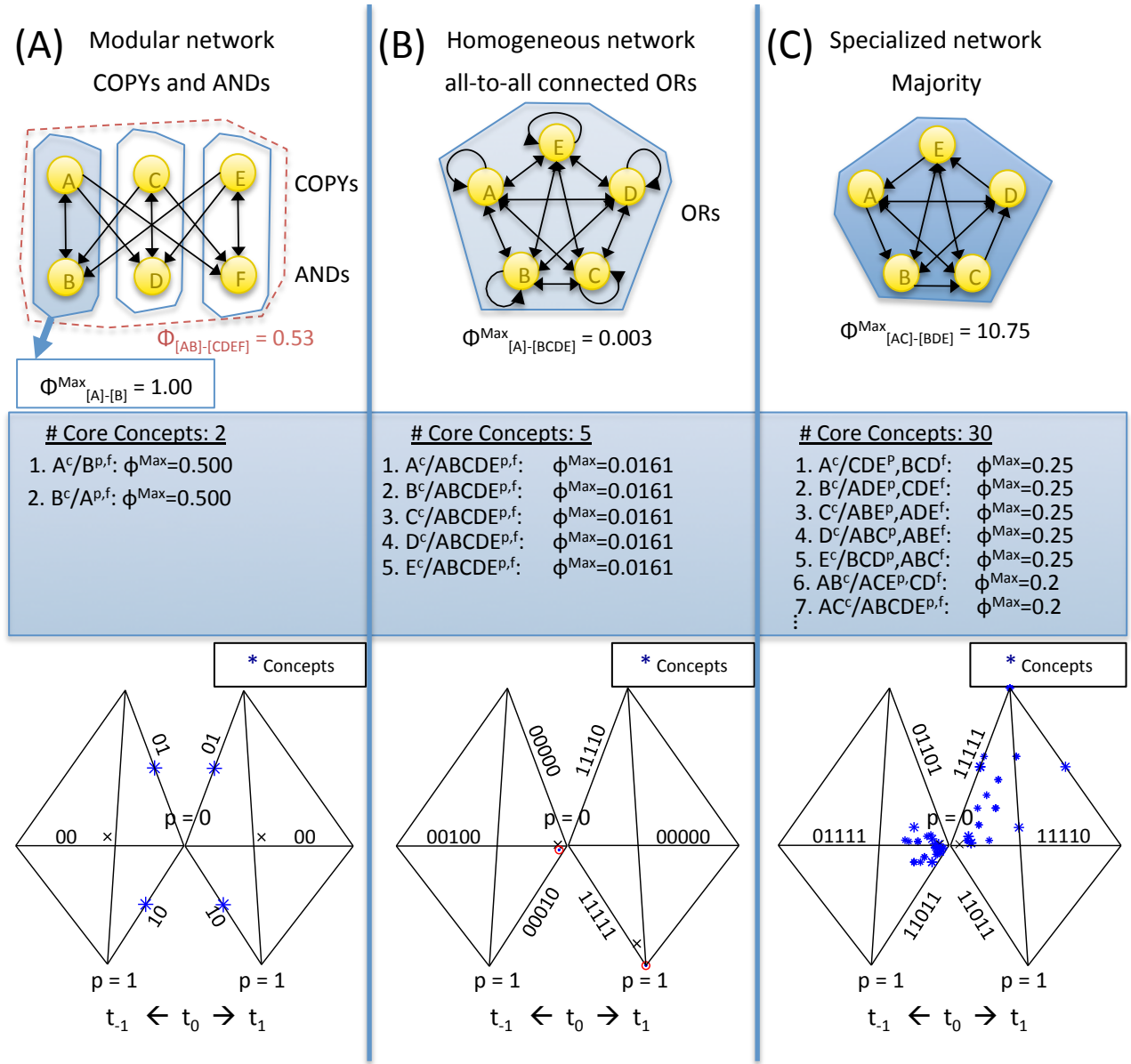


A system can condense into major and minor complexes and their residual interactions

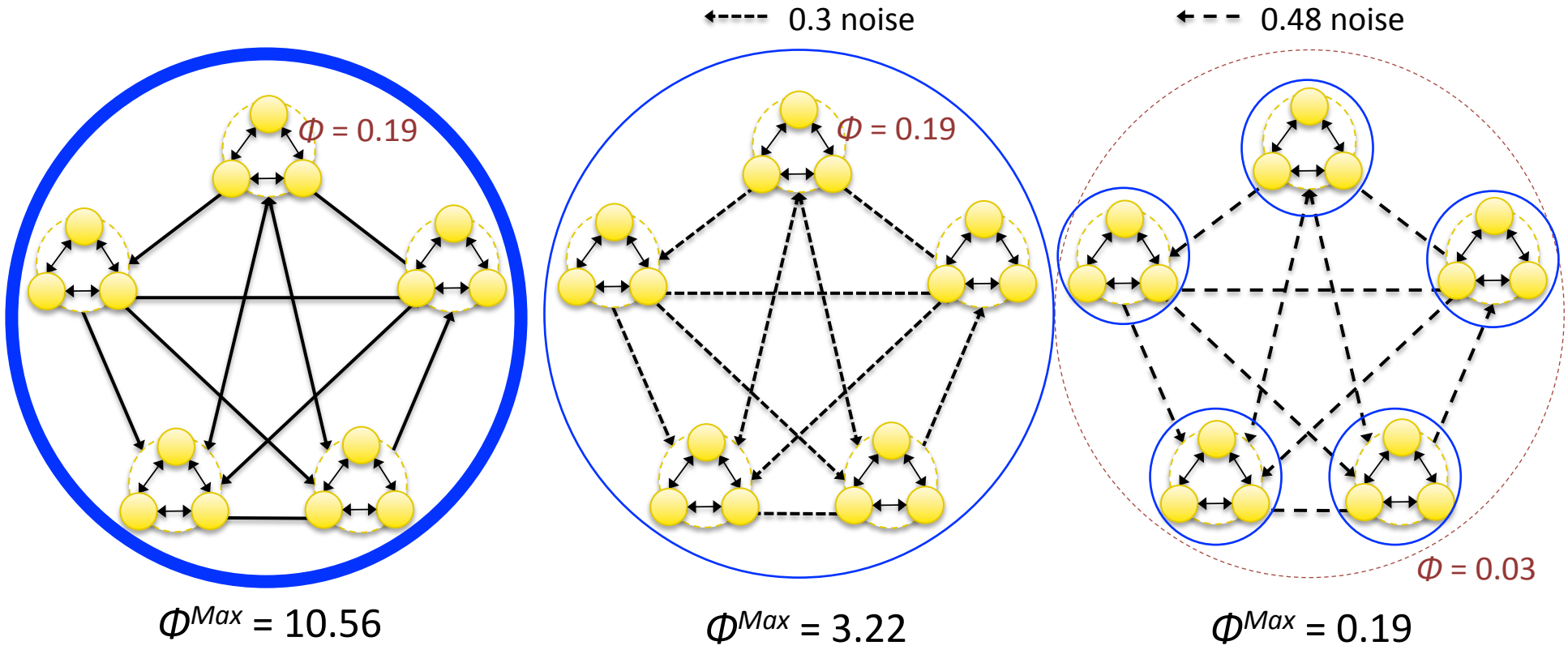


based on Oizumi, Albantakis, and Tononi, submitted

Qualia generated by modular, homogeneous, and specialized networks



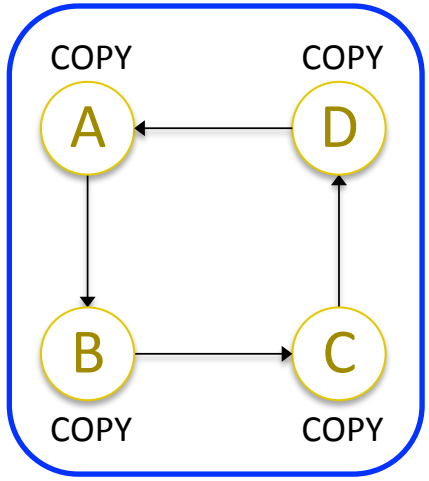
Consciousness can be graded



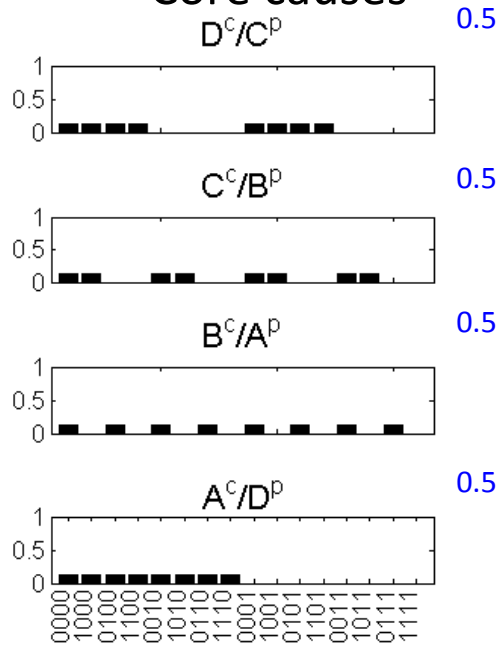
Inactive systems can be conscious

$$\varphi^{Max}(P, F \mid X = s_0)$$

$\Phi^{Max} = 1$

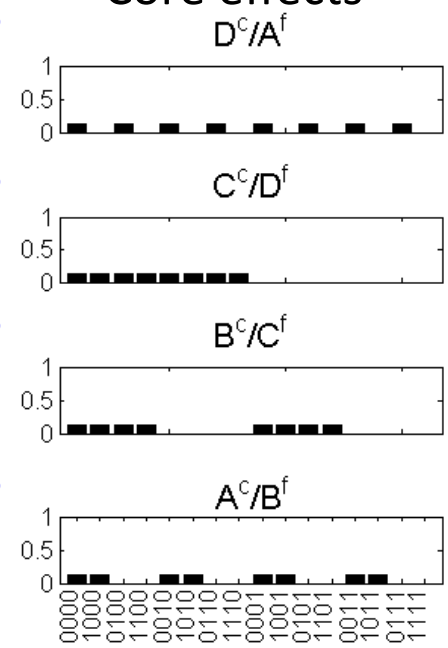


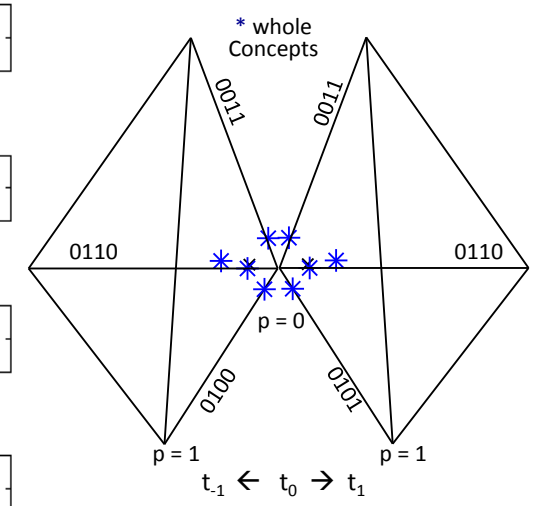
Core causes



0.5

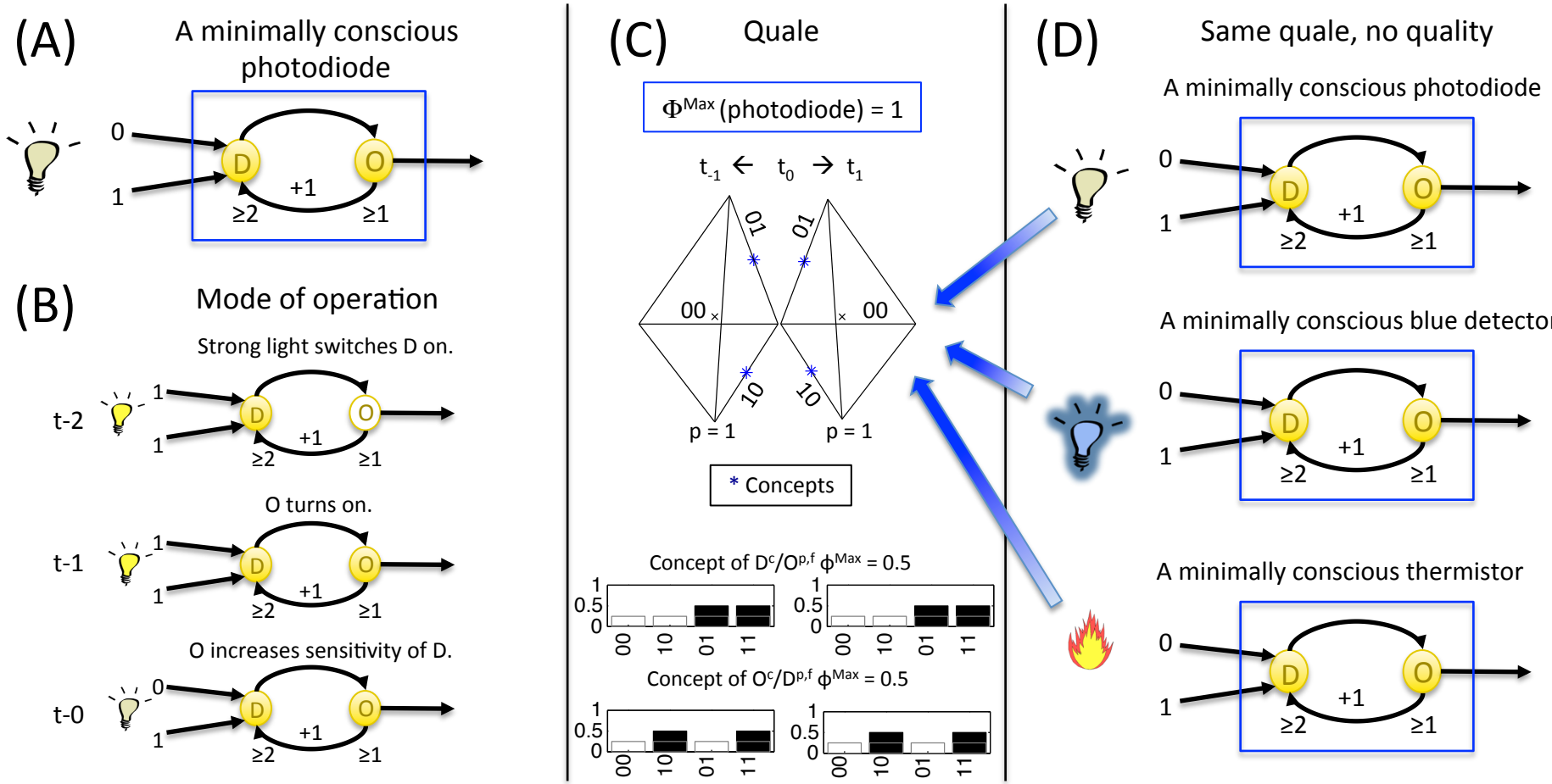
Core effects





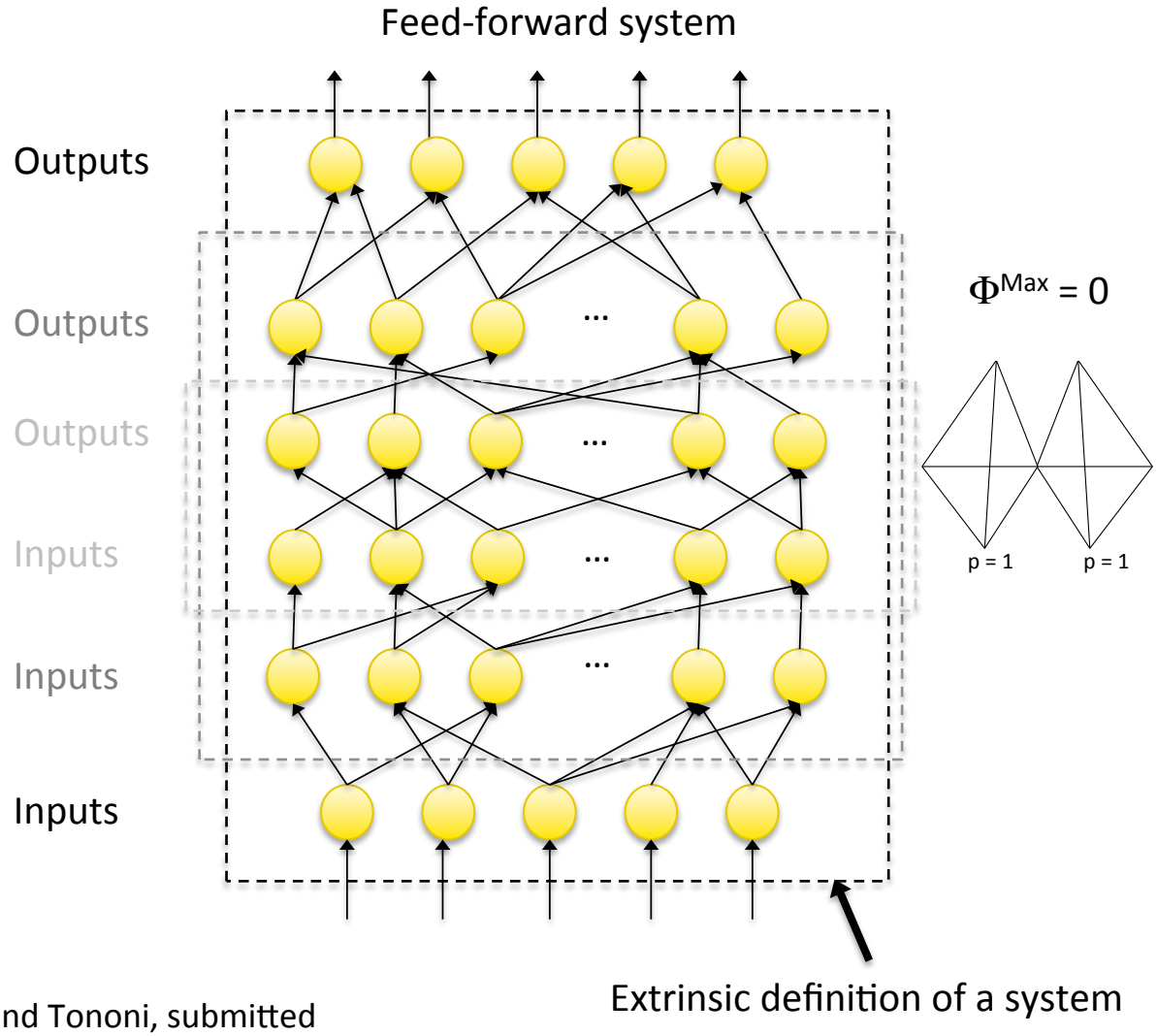
based on Oizumi, Albantakis, and Tononi, submitted

Simple systems can be conscious (but they have little quality): a “minimally conscious” photodiode



based on Oizumi, Albantakis, and Tononi, submitted

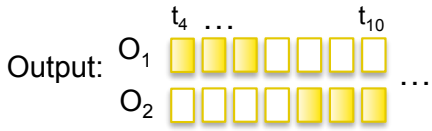
Complicated systems can be unconscious: feed-forward "zombie" systems do not generate consciousness

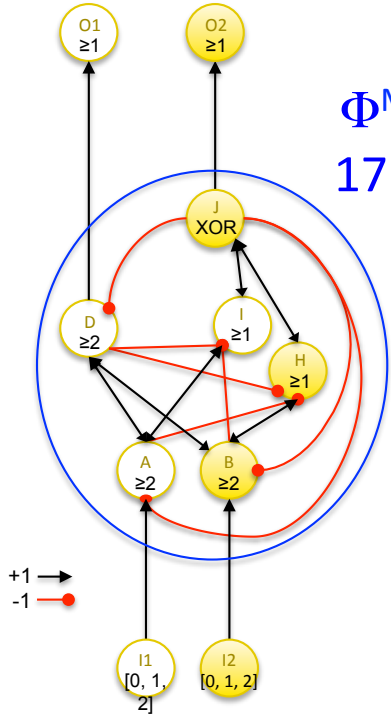


based on Oizumi, Albantakis, and Tononi, submitted

Conscious and unconscious systems can be functionally equivalent

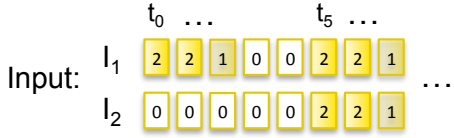
Integrated system



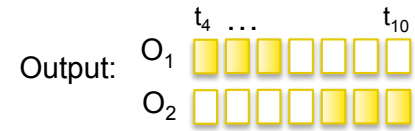


$\Phi^{Max} = 0.79$

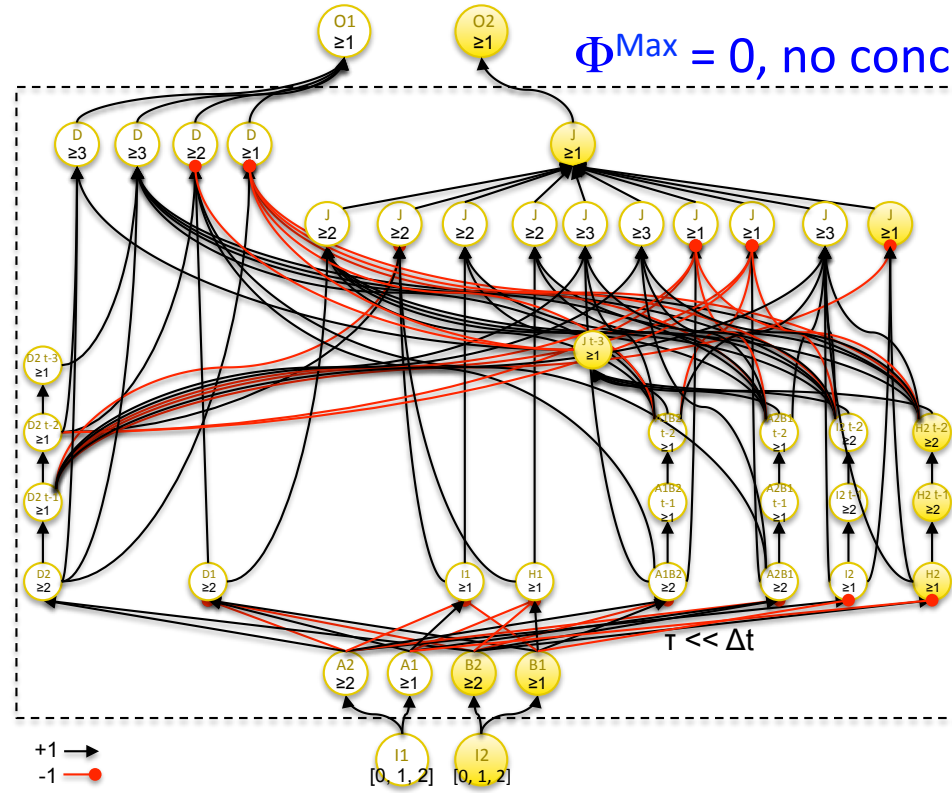
17 concepts

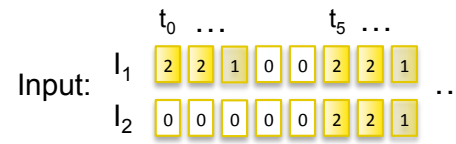


Feed-forward system



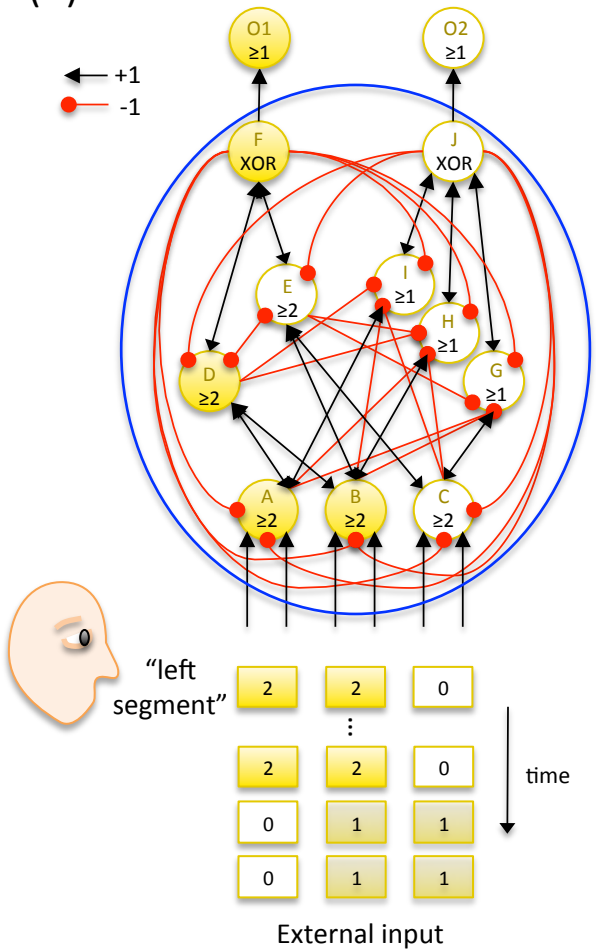
$\Phi^{Max} = 0$, no concepts





A complex can have ports in and ports out from and to the environment, but its qualia are 'solipsistic' (self-generated, self-referential, holistic)

(A)



(B) Concept Order

#Elements in the concept

Elementary (1st order) concepts:

ϕ(A=1) = 0.125 ϕ(B=1) = 0.077
ϕ(C=0) = 0.033 ϕ(D=1) = 0.186
ϕ(E=0) = 0.077 ϕ(F=1) = 0.186
ϕ(G=0) = 0.044 ϕ(H=0) = 0.025
ϕ(I=0) = 0.044 ϕ(J=0) = 0.186

2nd order concepts:

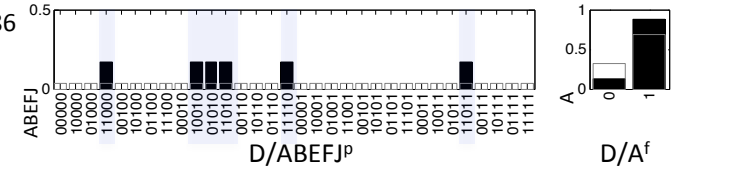
ϕ(AB=11) = 0.107
ϕ(BC=10) = 0.106
ϕ(DE=10) = 0.417
ϕ(DI=10) = 0.0625
...

3rd and higher order concepts:

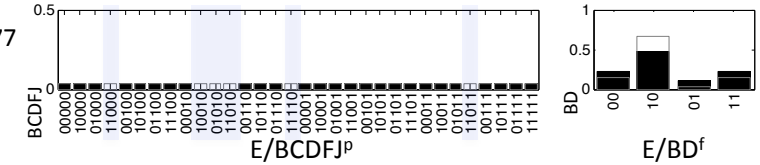
ϕ(ABC=110) = 0.063
ϕ(GHI=000) = 0.095
ϕ(DEF=101) = 0.188
ϕ(ADF=111) = 0.063
...
ϕ(ABDF=1111) = 0.063
...
ϕ(ABCDEF=110101) = 0.050 bits
...

(C) Concept truth value (positive or negative)

D: positive
ϕ(D=1) = 0.186
Purview:
D/ABEFJ^p,A^f

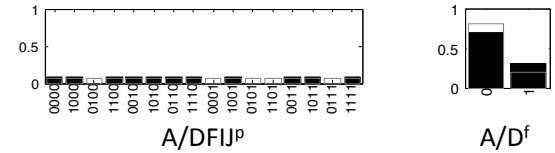


E: negative
ϕ(E=0) = 0.077
Purview:
E/BCDFJ^p,BD^f



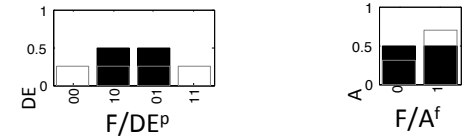
(D)

A: positive
ϕ(A=1) = 0.125
Purview:
A/DFIJ^p,D^f

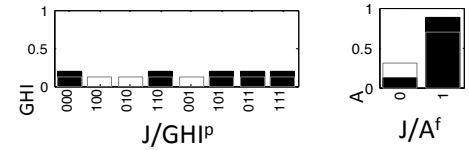


(E)

F: positive
ϕ(F=1) = 0.186
Purview:
F/DE^p,A^f

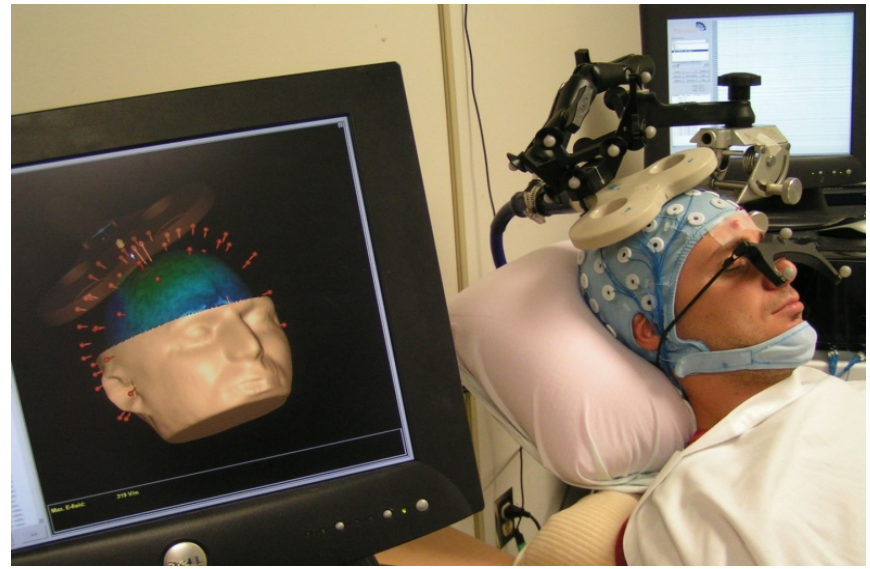


J: negative
ϕ(J=0) = 0.186
Purview:
J/GHI^p,A^f



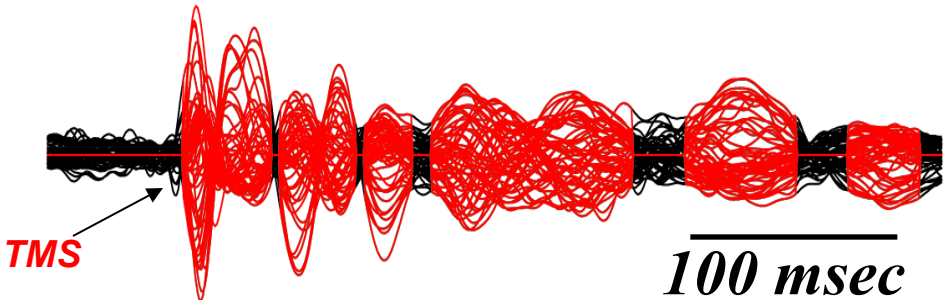
IIT: some predictions

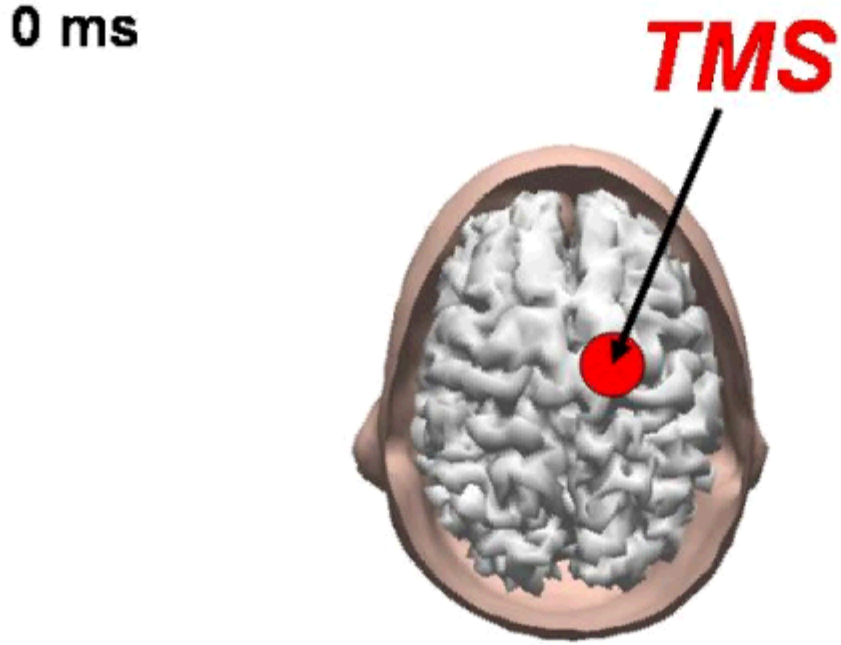
From theory to practice: Evaluating integrated information using TMS and hd-EEG during wake and sleep



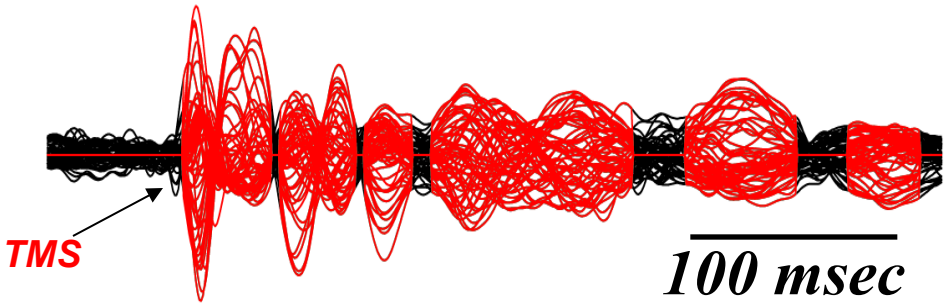
Massimini et al., Science, 2005

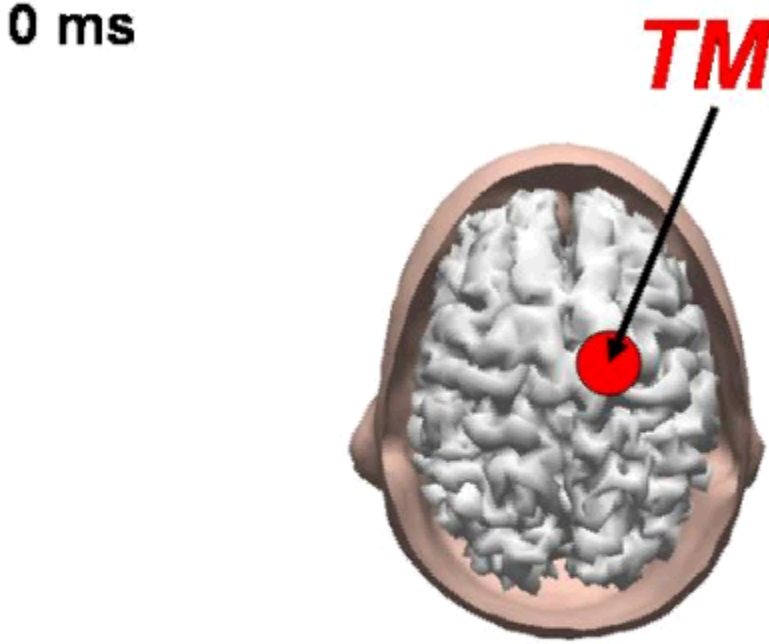
Wakefulness



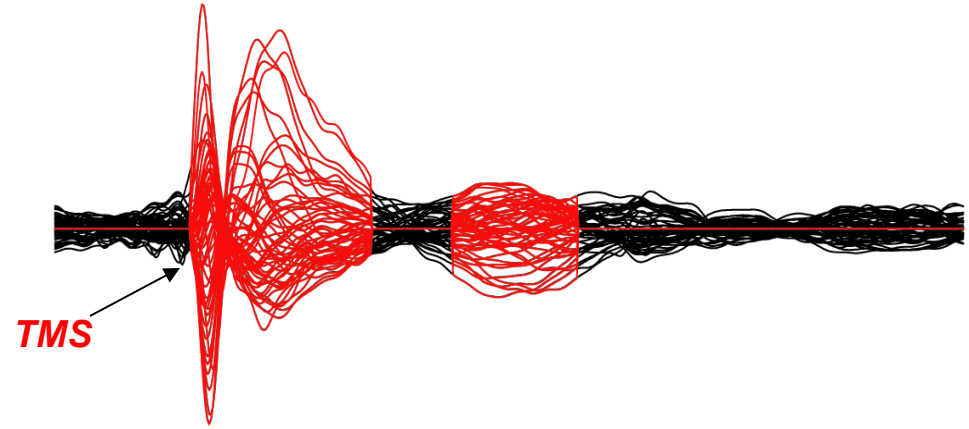


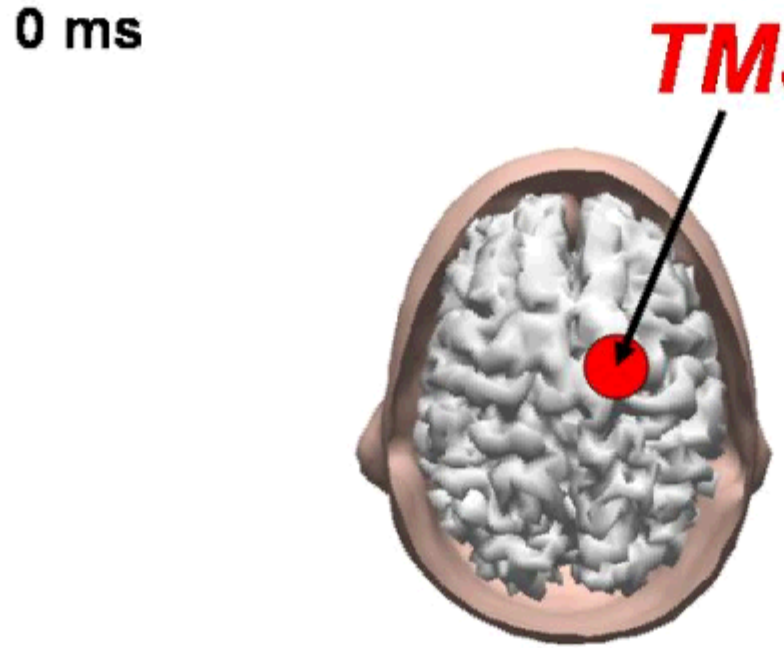
Wakefulness





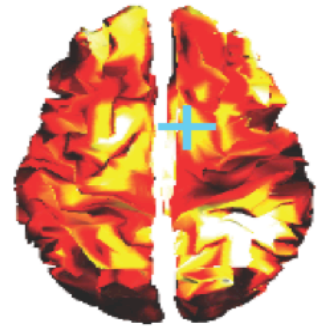
Slow Wave Sleep



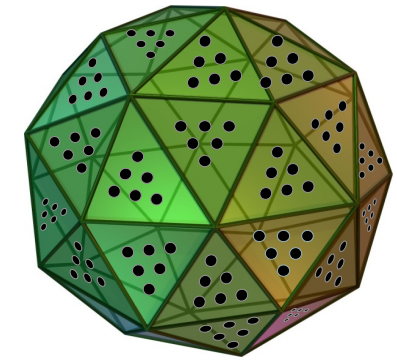


Like consciousness, information integration is high in wake, breaks down in slow wave sleep, and returns during REM sleep

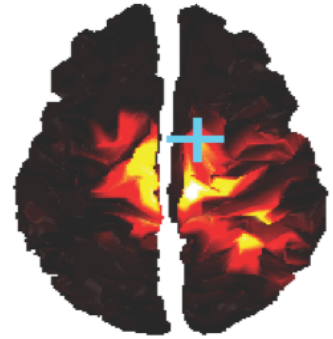
Wake



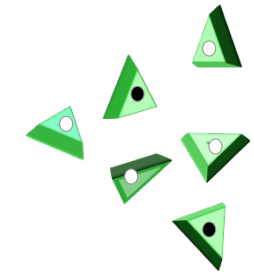
Highest inf. integration



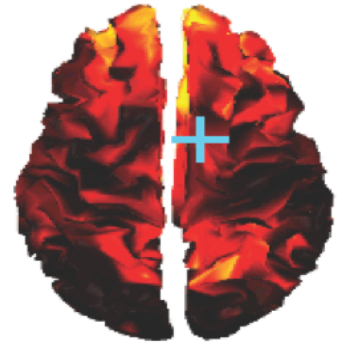
Slow Wave Sleep



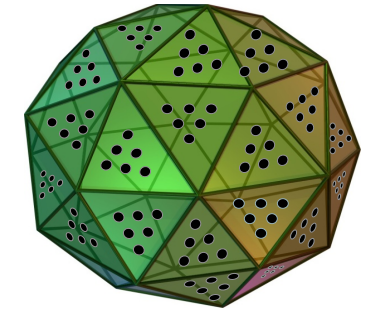
Low inf. integration



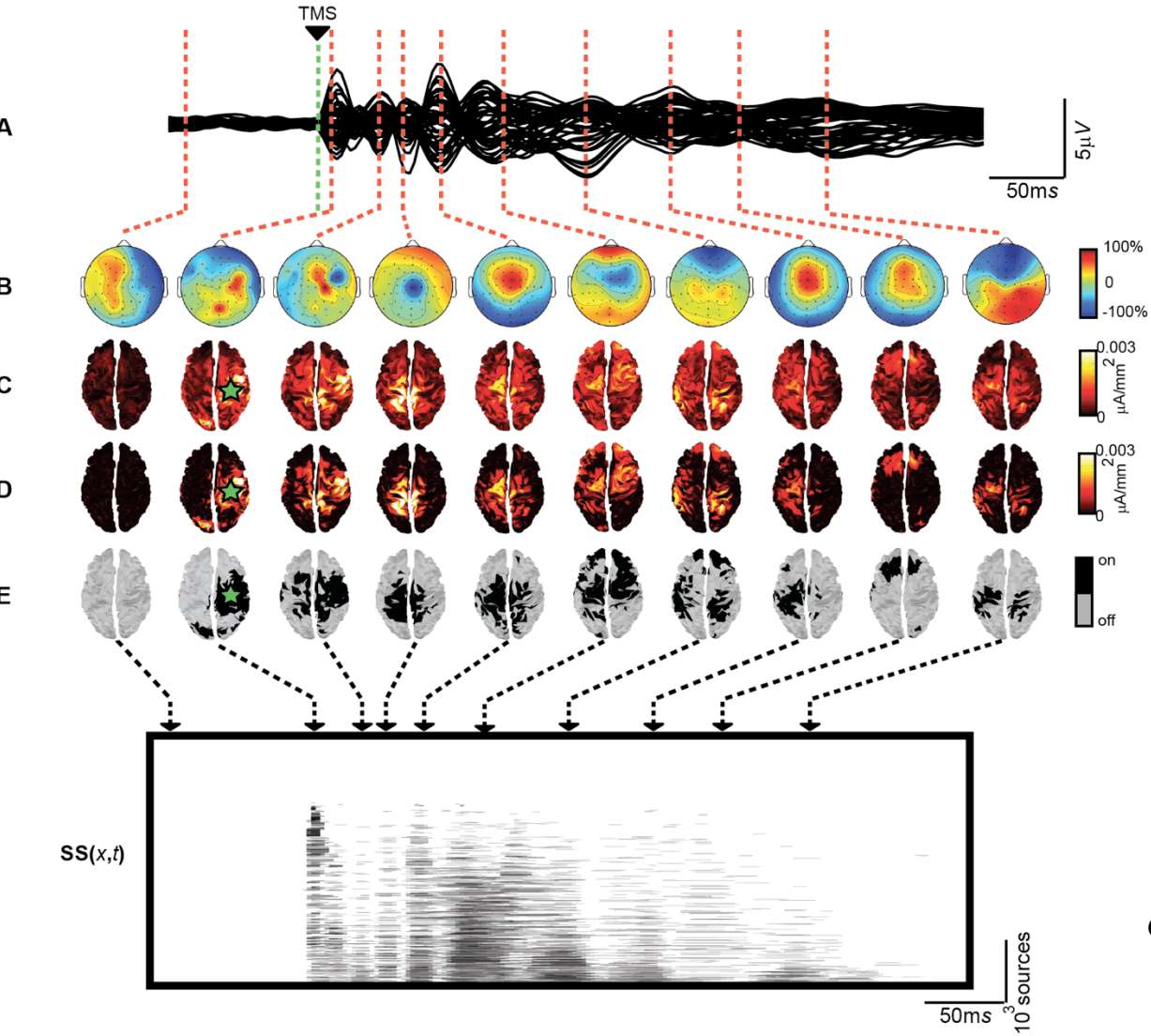
REM Sleep



High inf. integration



Towards a Consciousness – Meter: "zap and zip"



A. **Time course of TMS-hdEEG responses**

B. **Voltage maps**

C. **Current sources**

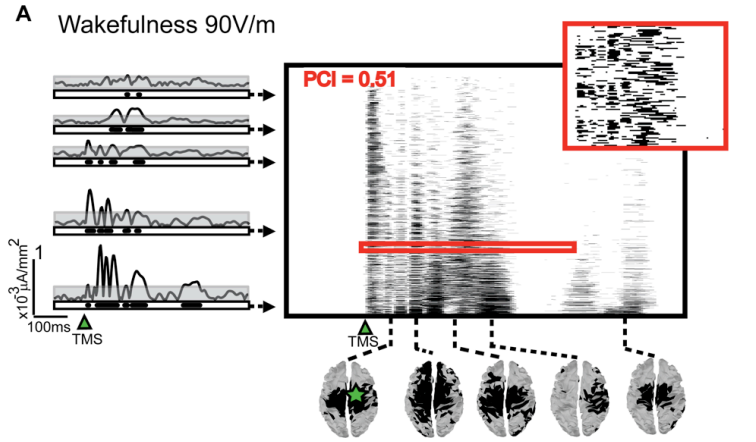
D. **Significant sources (nonparametric)**

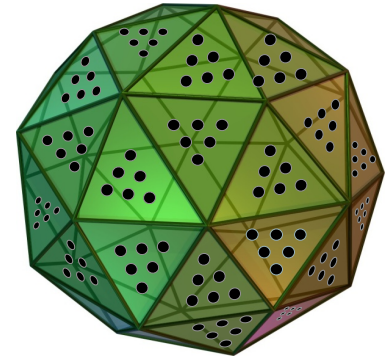
E. **Binarized matrix**

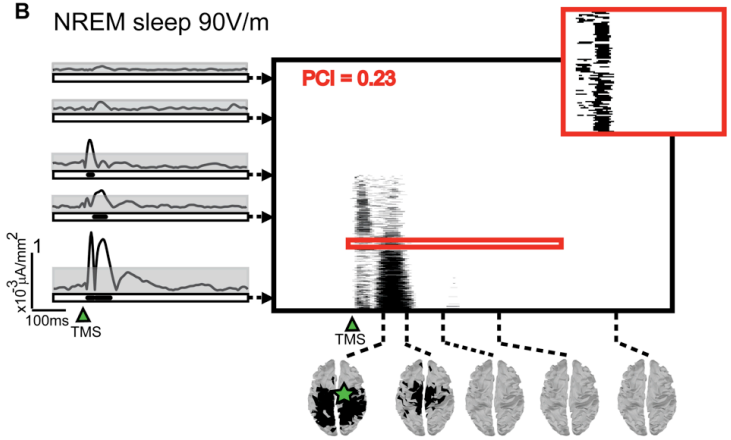
Perturbational Complexity Index (PCI), a practical measure of information integration using TMS ("zapping")

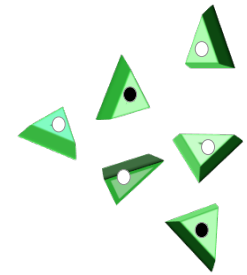
computed using Lempel-Ziv encoding of hd-EEG sources time series ("zipping")

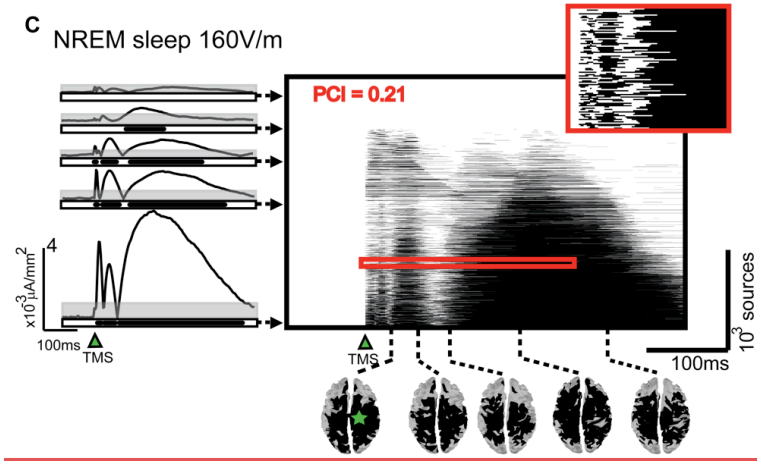
(Casali et al., Neuroimage 2010, Science TM, 2013)

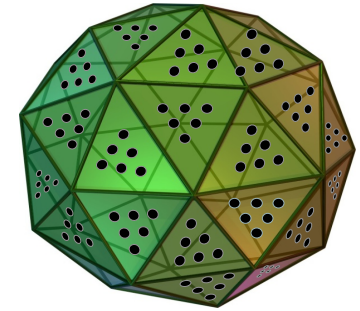






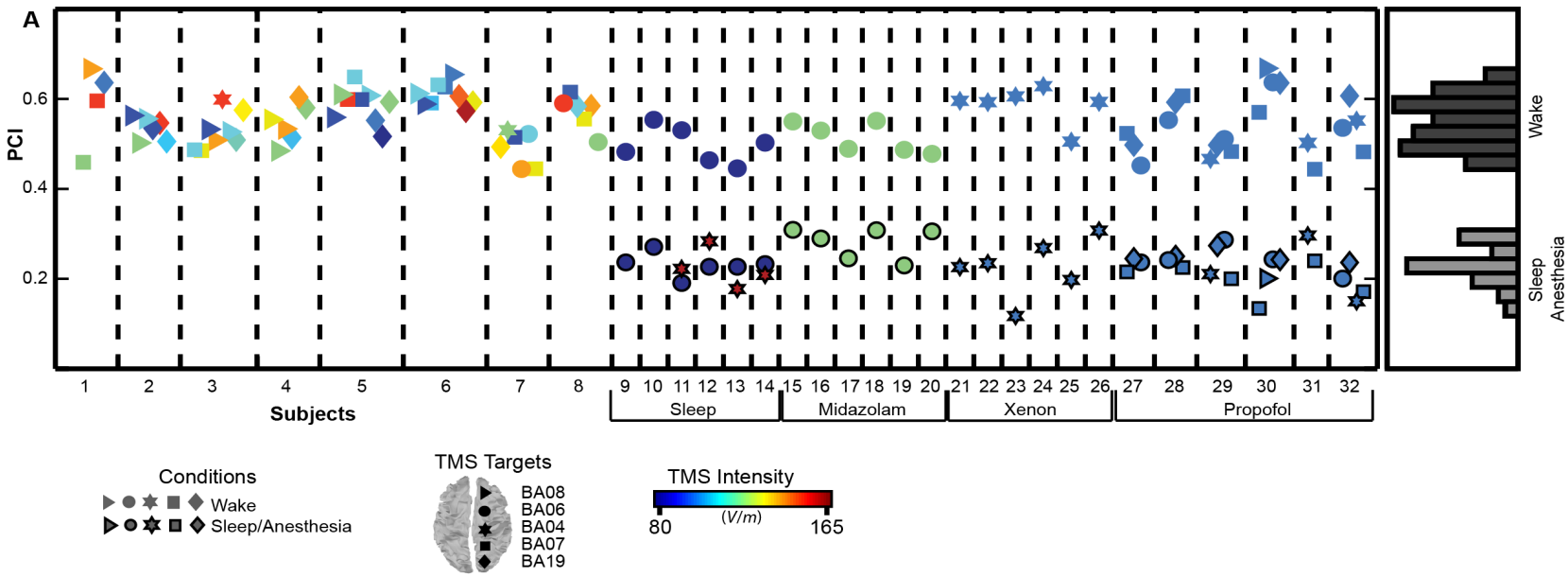






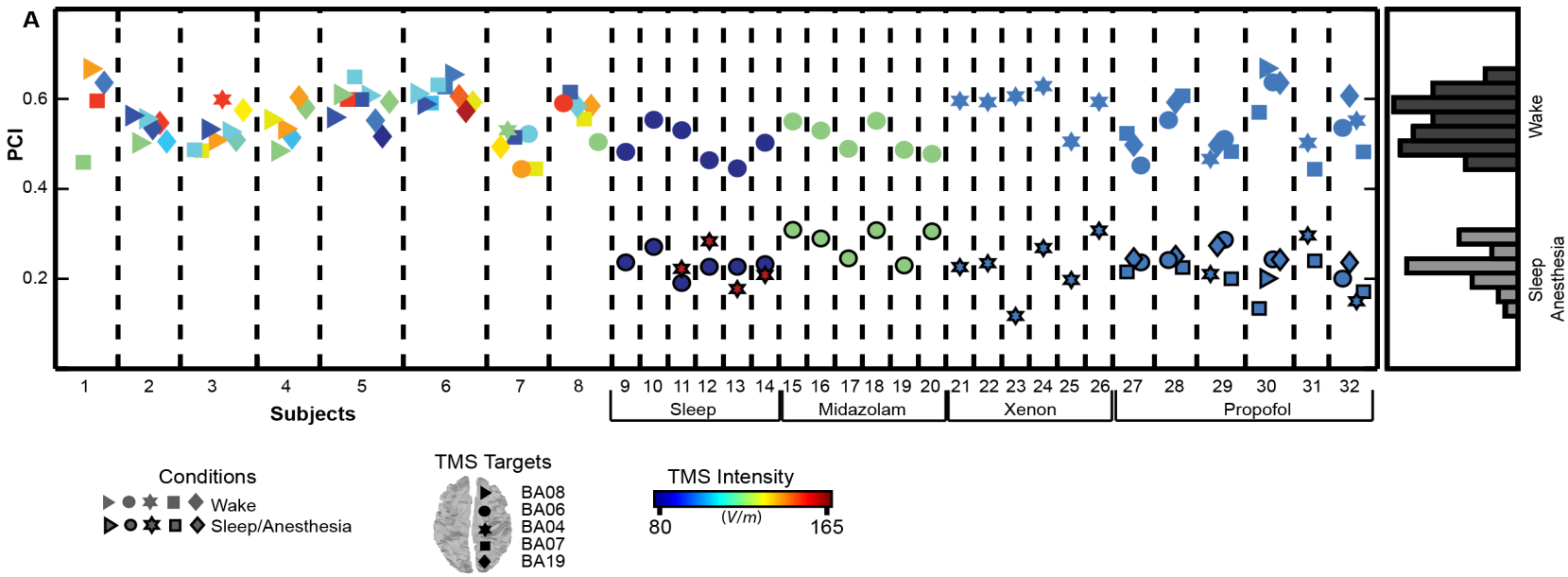
PCI is sensitive to the complexity (algorithmic compressibility) of the responses to TMS

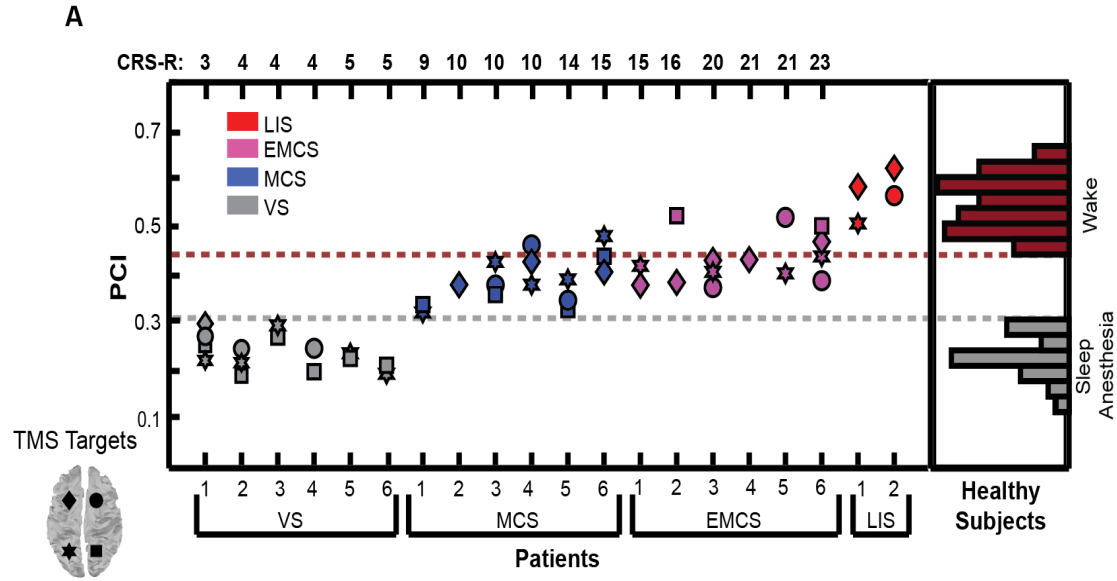
Separating higher from lower levels of consciousness

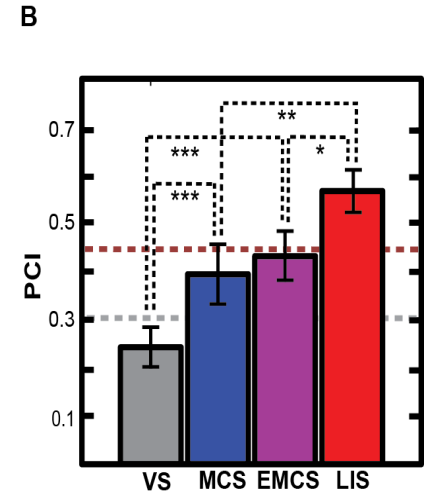


Casali et al., Science TM, 2013

Separating higher from lower levels of consciousness







Casali et al., Science TM, 2013

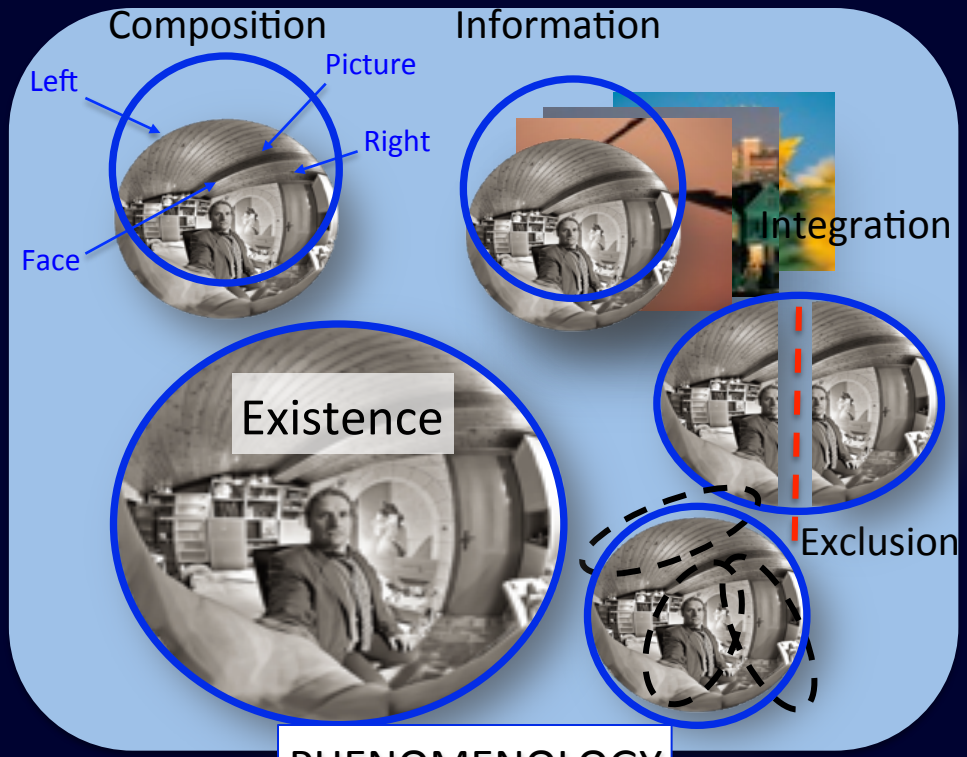
The Biology of Consciousness

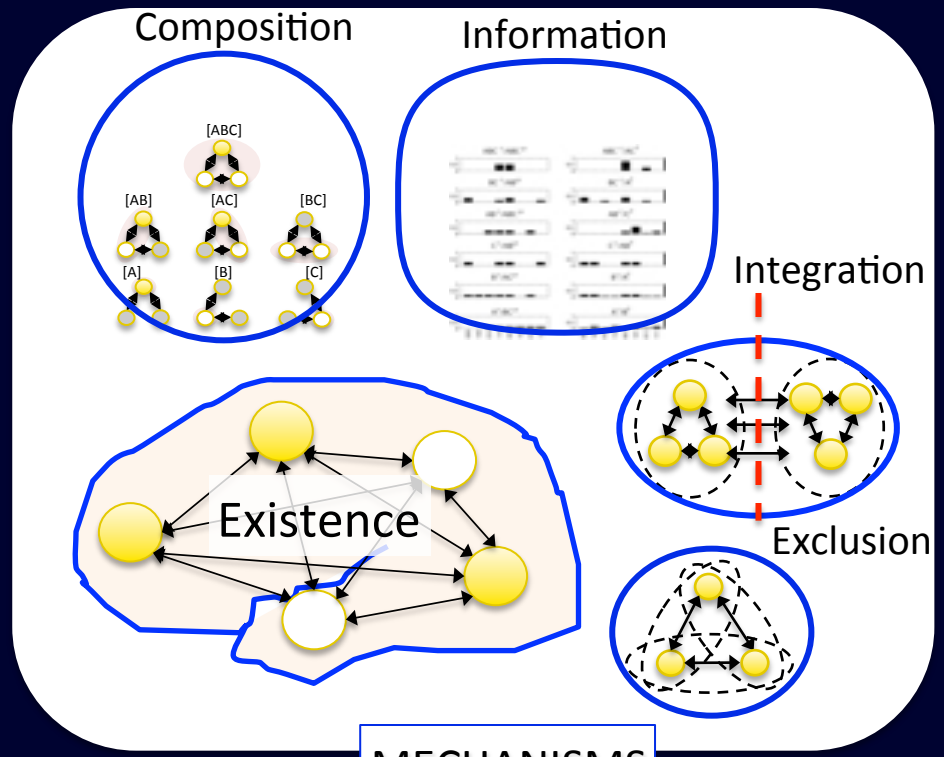
Christof Koch

Allen Institute for Brain Science

1/8/2014

From phenomenology to mechanisms, and back



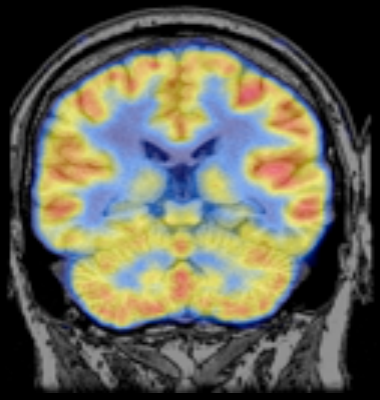


PHENOMENOLOGY

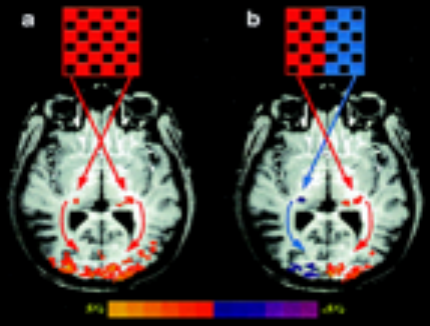
MECHANISMS

Explanations

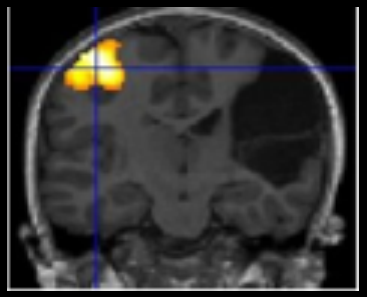
Why not the cerebellum?



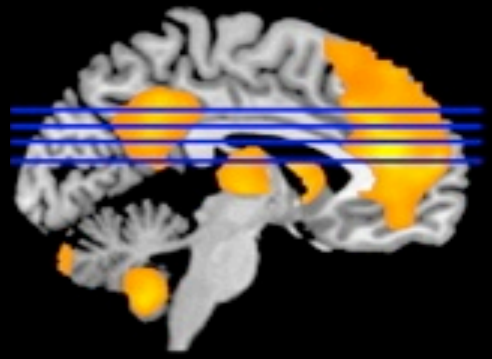
Why not afferent pathways?



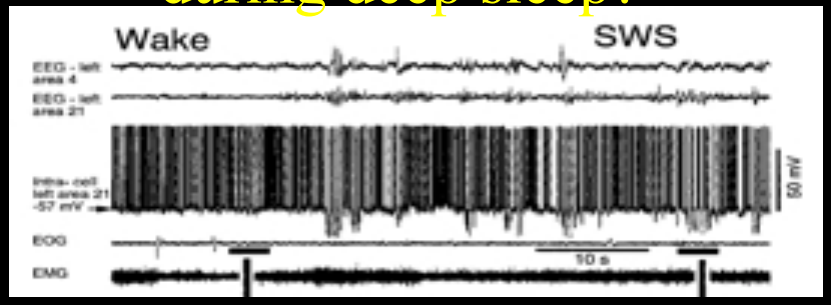
Why not efferent pathways?



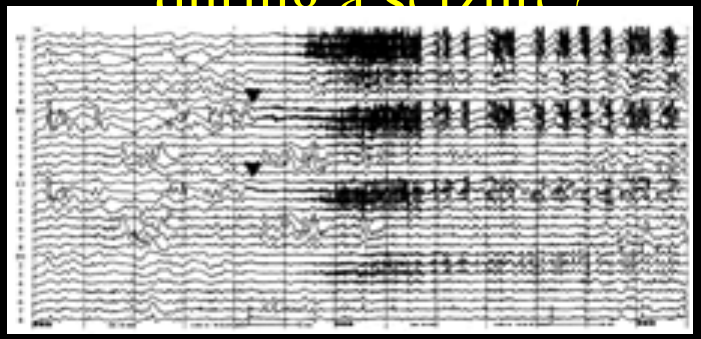
Why not cortico-subcortico-cortical loops?



Why not the cortex during deep sleep?



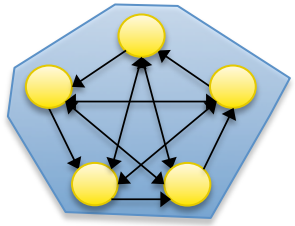
Why not the cortex during a seizure?



Explanations

Cortical system

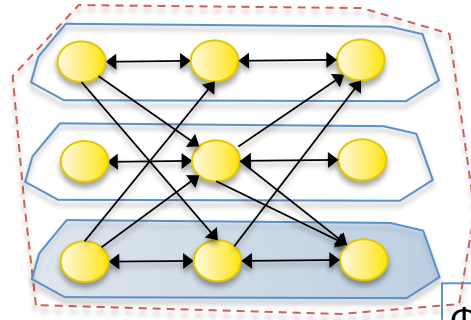
Inhomogeneous network, functional specialization and integration



$\Phi^{Max} = 10.56$

Cerebellum

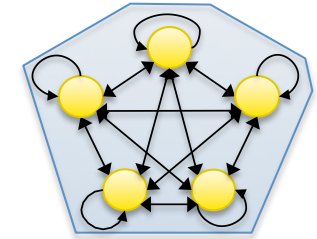
Modular organization



$\Phi^{max} = 1.00$

Cortical system during deep sleep / anesthesia/ seizures

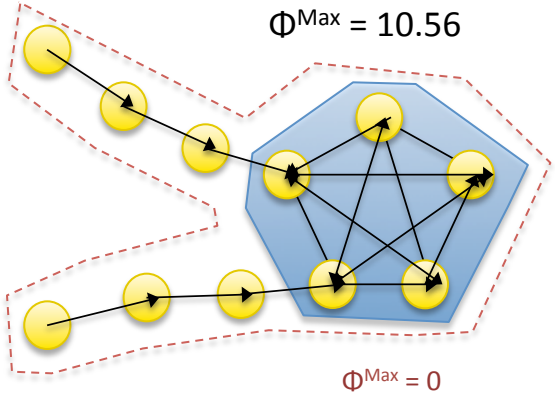
Homogeneous network



$\Phi^{Max} = 0.003$

Afferent pathways

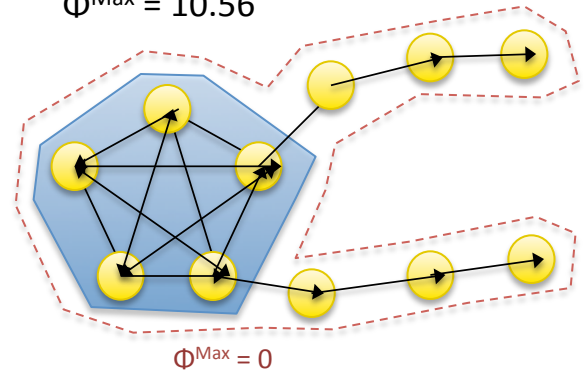
$\Phi^{Max} = 10.56$



$\Phi^{Max} = 0$

Efferent pathways

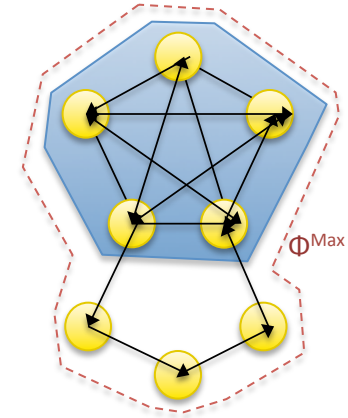
$\Phi^{Max} = 10.56$



$\Phi^{Max} = 0$

Cortico-subcortical loop

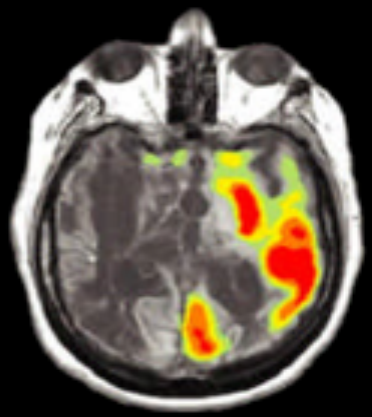
$\Phi^{Max} = 10.56$



$\Phi^{Max} = 0$

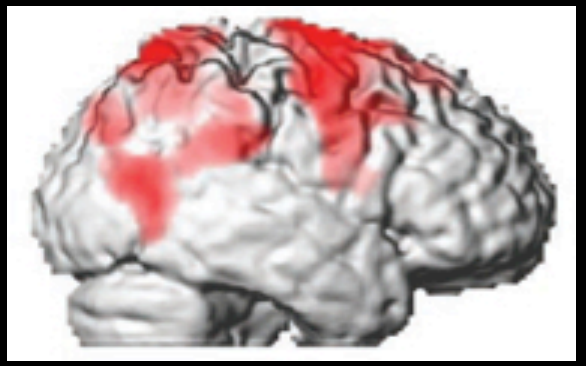
Extrapolation

Brain “islands” in a vegetative subject



Newborn / 1 year old

Ketamine anesthesia

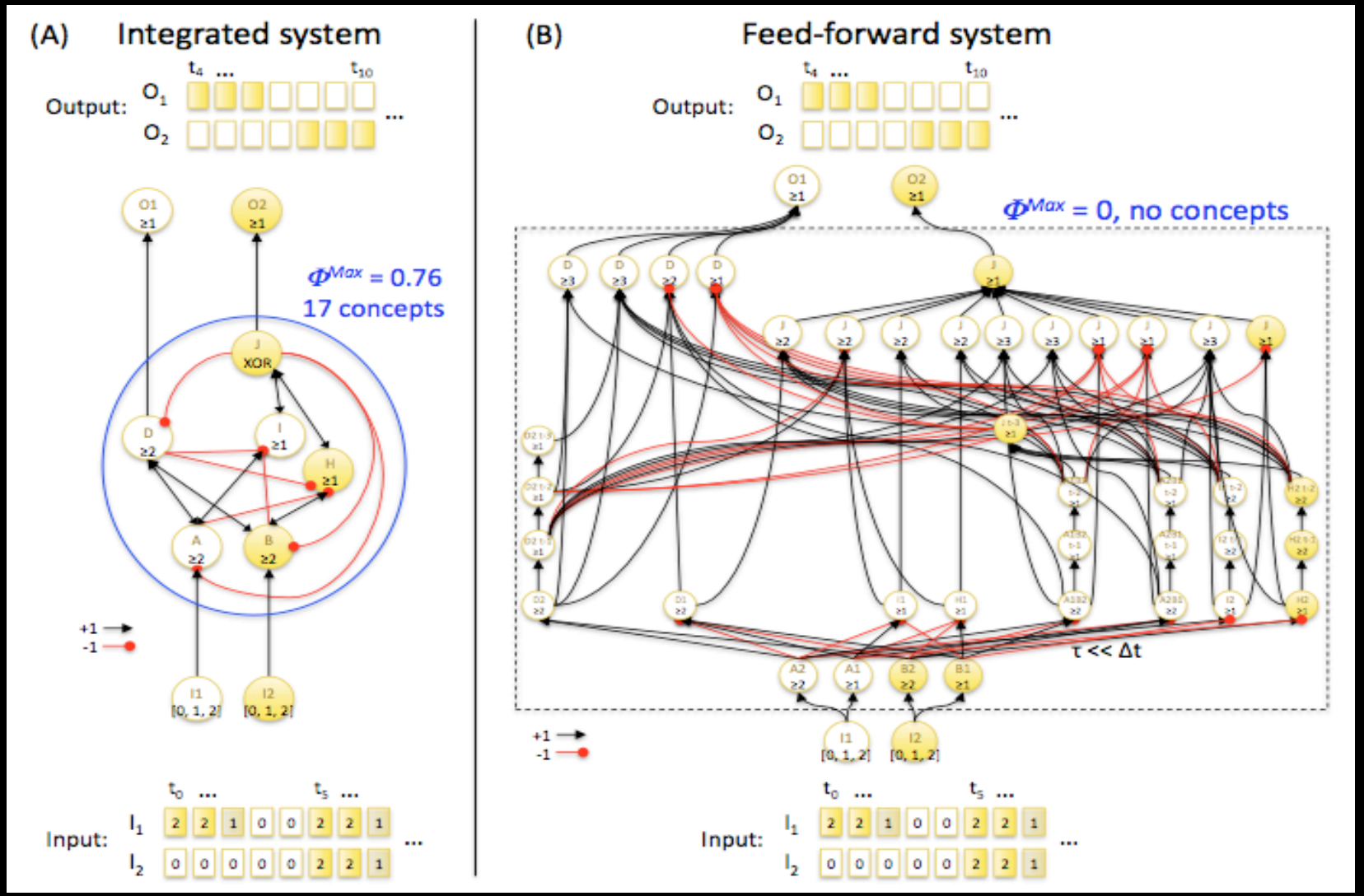


Sleepwalking

Octopus

Apple Siri

Computational Systems



Which systems are not conscious (or only minimally)

- Dumb/simple system
- Aggregate systems are not conscious above and beyond the consciousness of their components
- Feed-forward computational systems
- Computer simulations of brains - Their consciousness relates to the cause-effect repertoire of the underlying hardware instantiating a Turing Machine