The bumpy road to strong AI

Paul Cisek

Groupe de recherche sur le système nerveux central
GRSNC

Montréal Artificial Intelligence & Neuroscience
November 19, 2017

“Strong AI”

- “Strong AI”, a.k.a. “artificial general intelligence”
  - Able to solve complex problems across a wide range of domains
  - Able to represent commonsense knowledge
  - Able to learn new tasks
  - Able to communicate in a natural language
  - In short: human-like intelligence
- “Weak AI”
  - Every type of AI that is not “strong”
    - Domain specific
    - Hard-wired
    - Examples:
      - Expert systems, machine vision
      - ELIZA, Eugene Goostman, SIRI
  - In short: all AI that has been developed so far

The promise of AI

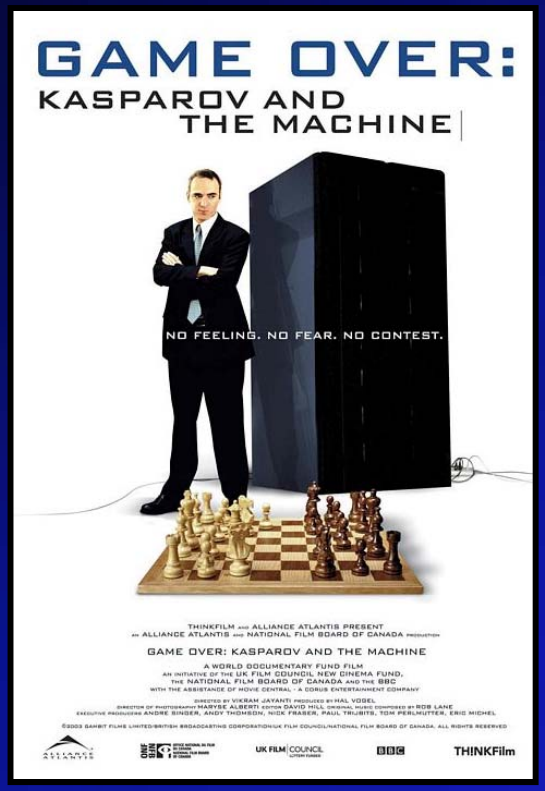
- *"every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it"*
  McCarthy, Minsky, Rochester, Shannon, 1955
- "*machines will be capable, within twenty years, of doing any work a man can do*"
  Herbert Simon, 1965
- "*Within a generation ... the problem of creating 'artificial intelligence' will substantially be solved*"
  Marvin Minsky, 1967

The threat of AI

GOFAI

- Good Old-Fashioned Artificial Intelligence
  - Central idea: “Intelligence” is the ability to reason
    - Make logical inferences from a set of facts
      - E.g. “see dark clouds”, “hear pitter-patter” => “it is raining”
    - Make smart decisions
      - E.g. “it is raining” => “bring an umbrella”
  - Symbolic systems
    - Symbols stand for atoms of knowledge (facts, beliefs, etc.)
    - Thinking: Rule-based manipulation that generates new knowledge
      - E.g. “**if** A and B **then** C”

Successes and failures

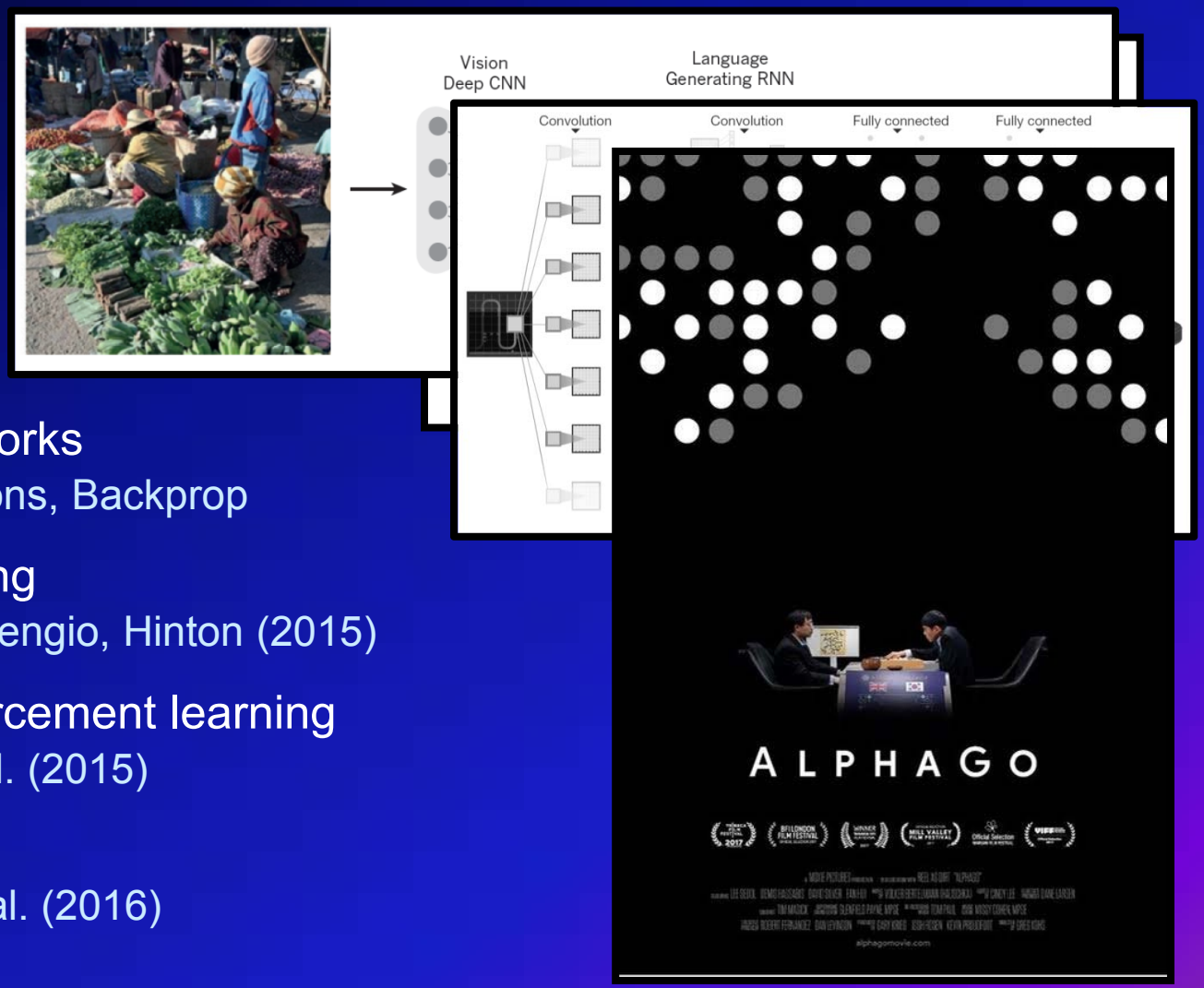


- "Fifth generation computer"
  - Begun in 1982
  - Funded by Japan's Ministry of International Trade and Industry
  - A ten-year project to build massive inference engines capable of analyzing huge data sets
  - Massively parallel architecture
- "CYC project"
  - Begun in 1984
  - Funded by the CIA, in response to above…
  - Symbolic inference engine written in LISP and CycL
  - Human coded knowledge rules
- Both failed to achieve strong AI

Why did GOFAI fail?

- No learning
  - Rules were hard-wired by the programmers
  - Not possible for the programmers to list all relevant contingencies
    - "Frame problem"
  - Not possible for the system to *generalize* to new context
    - Stuck in a specific domain
    - Not "strong"
  - Has it been solved?

Successes



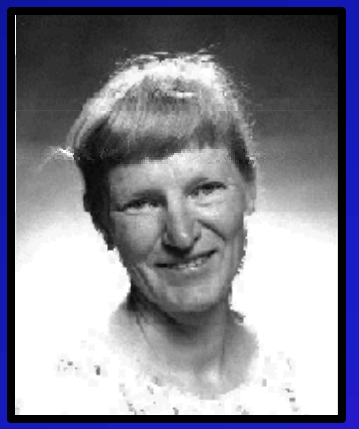
- Neural networks
  - Perceptrons, Backprop
- Deep learning
  - LeCun, Bengio, Hinton (2015)
- Deep reinforcement learning
  - Mnih et al. (2015)
- AlphaGo
  - Silver et al. (2016)

Why did GOFAI fail?

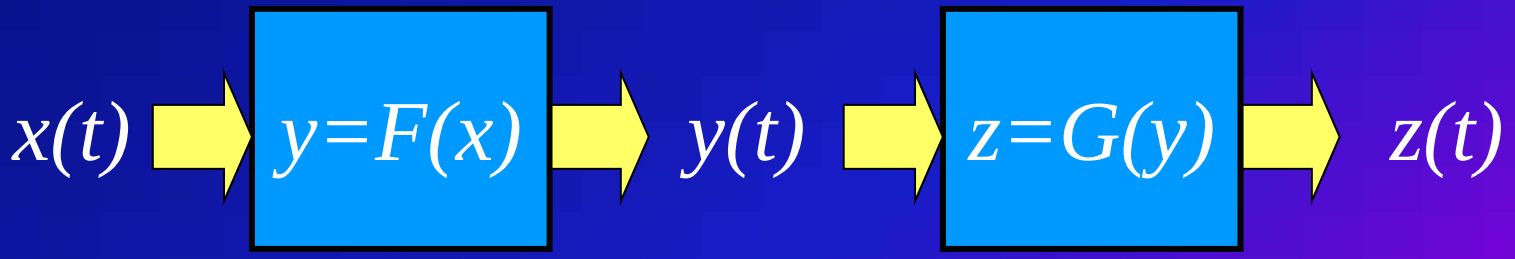
- No learning
  - Rules were hard-wired by the programmers
  - Not possible for the programmers to list all relevant contingencies
    - "Frame problem"
  - Not possible for the system to *generalize* to new context
    - Stuck in a specific domain
    - Not "strong"
  - Has it been solved?
    - In large part, yes
- No semantics
  - The symbols are meaningful only to the programmers
  - Semantics are external to the system
  - "Chinese room argument"
    - John Searle (1980)
  - "Symbol grounding problem"
    - Stevan Harnad (1990)
  - Has *it* been solved?

What is meaning?

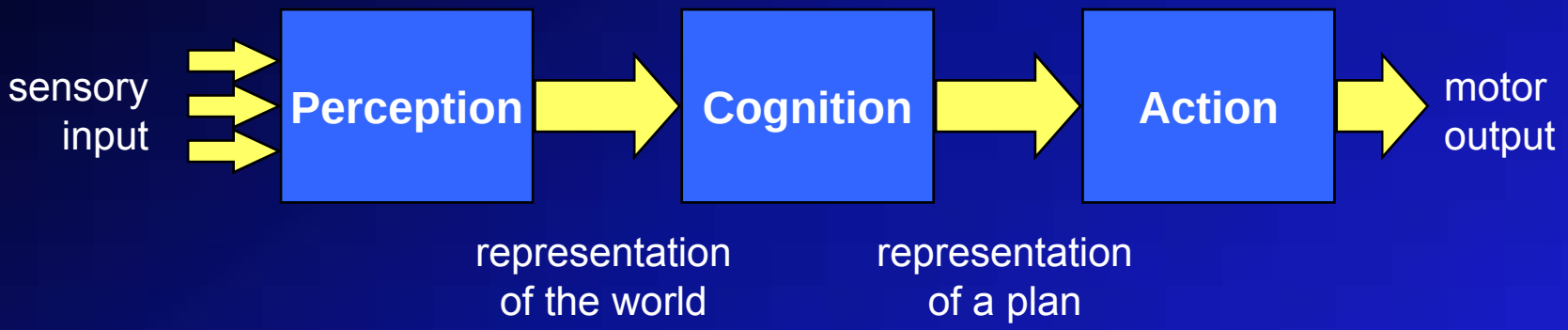
- Semantics are about function
  - Millikan (1989): The meaning of a representation lies in how it is ***used***
    - "Let us view the system, then, as divided into two parts or two aspects, one of which produces representations for the other to consume. What we need to look at is the consumer part, at what it is to use a thing *as* a representation."
  - Also Dretske (1981), Mingers (1996), etc.



Ruth Millikan



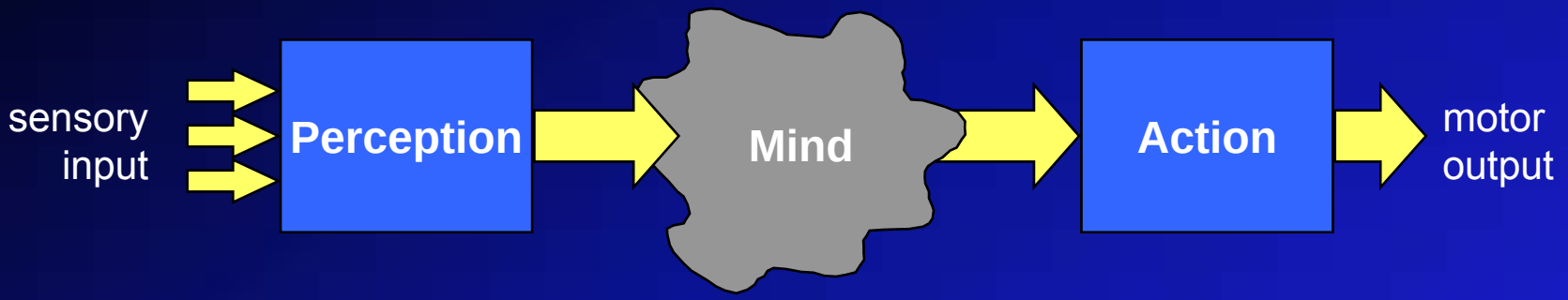
What is behavior?



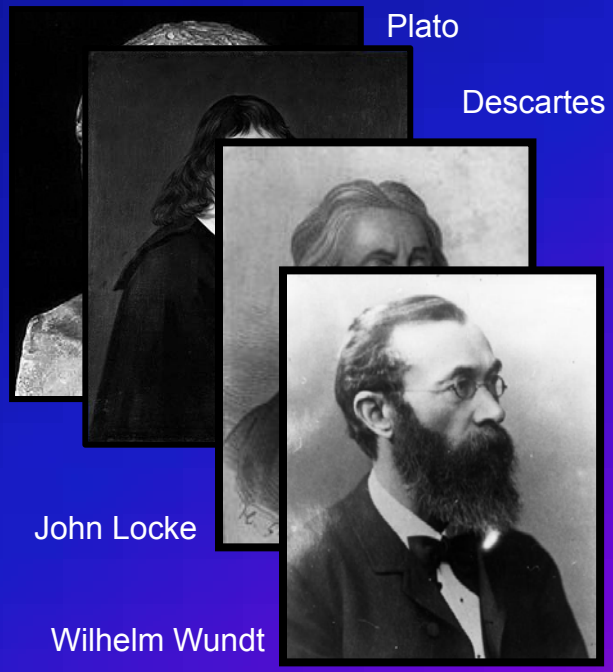
- University courses
- Textbooks
- Journals
- Conferences
- Academic departments
- Grant review committees
- Scientists
- Questions we ask
- Theories we propose

Q: From where does this view originate?

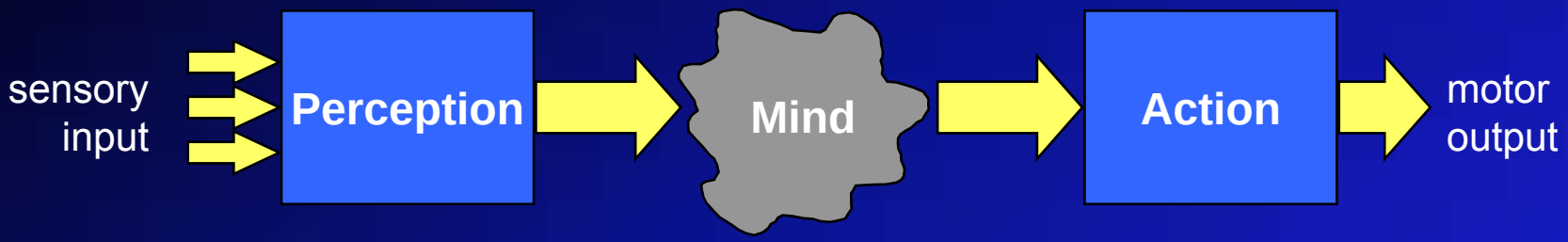
"Dualism"



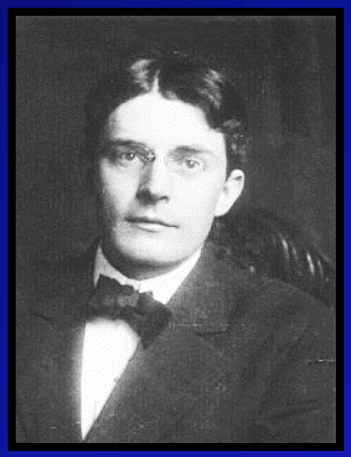
- The mind is non-physical
  - There must exist interfaces between the non-physical mind and the physical world
- Psychology: Study of the psyche
  - The mind is studied through careful disciplined introspection



“Behaviorism”

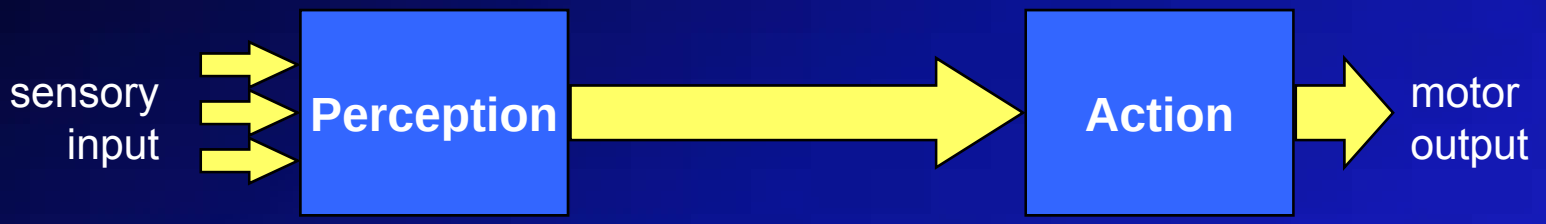


– Stop all of this metaphysical nonsense…

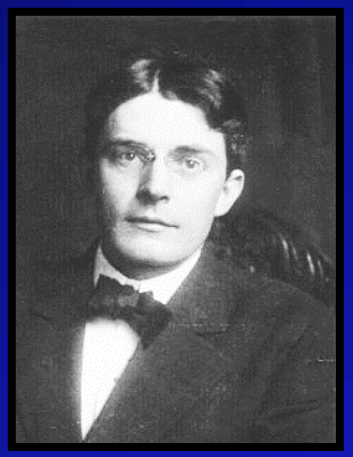


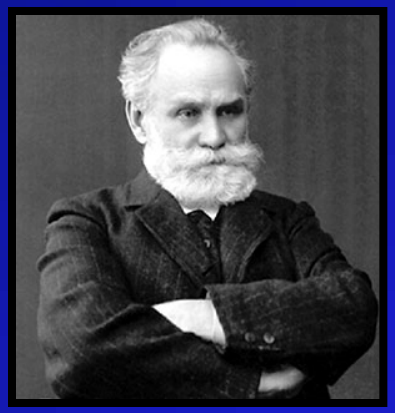
John Watson

"Behaviorism"

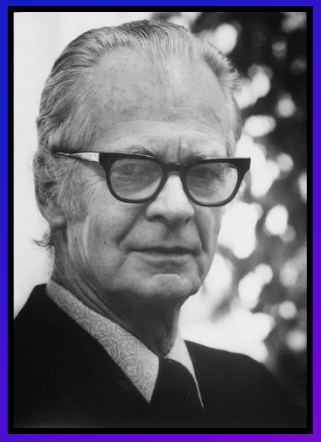


- Stop all of this metaphysical nonsense…
- Don't need a "*mind*"; Perception and Action can be directly linked
- Subject matter: Learning laws which establish the linkage

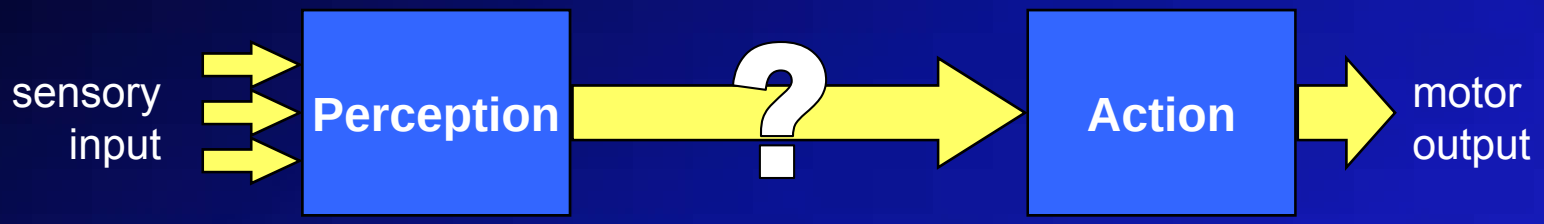

John Watson


Ivan Pavlov


Edward Thorndike


B.F. Skinner

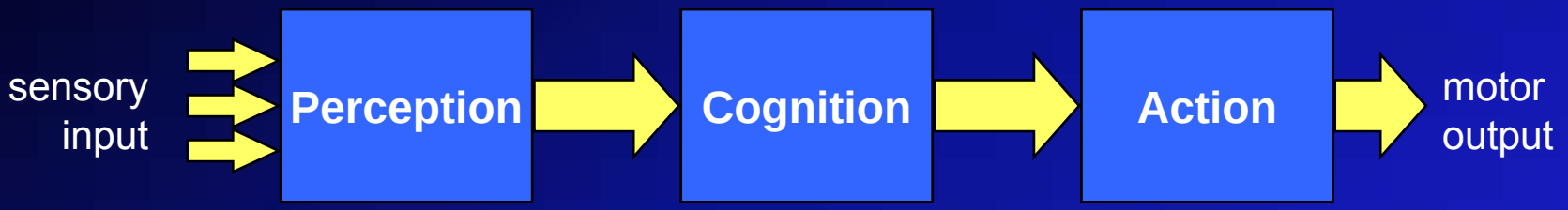
“Behaviorism”



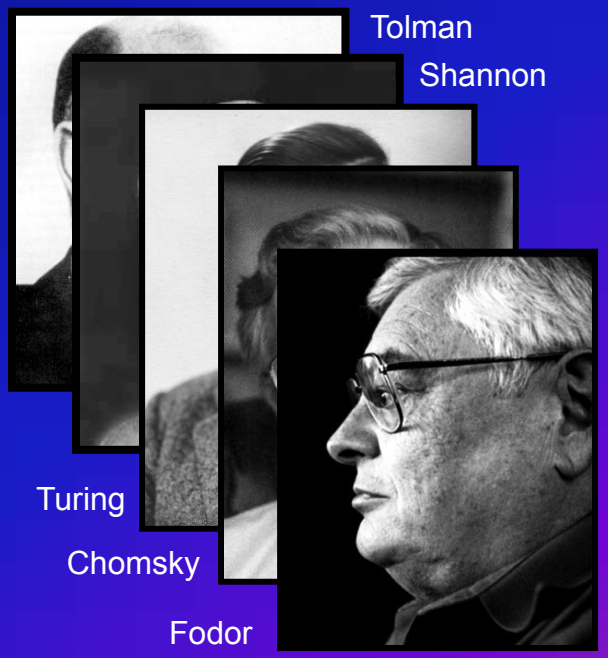
- Internal processes are indispensable
  - We can infer beyond our experience

Tolman

“Cognitivism”



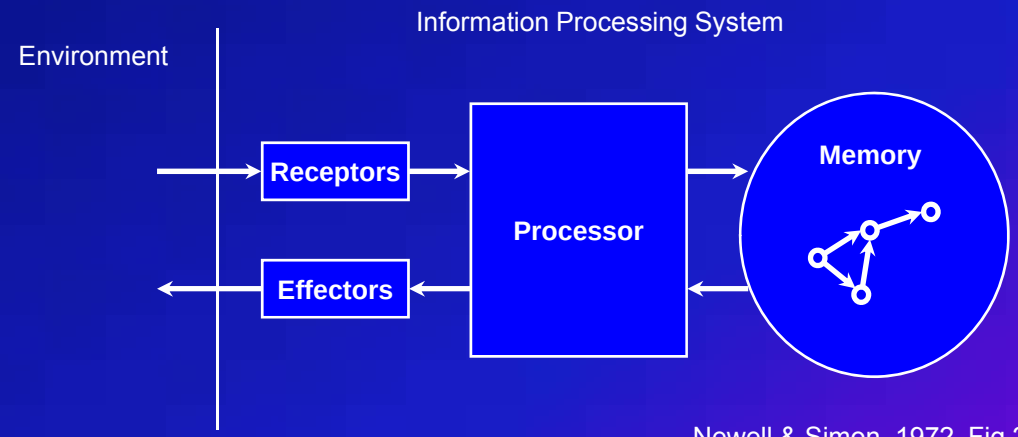
- Internal processes are indispensable
  - We can infer beyond our experience
- “Cognition” takes the mind’s place
- A fully physical process – but what?
- “Information processing”
  - Definition of “information”
  - Definition of “processing”
- Cognition is a computational process
  - Language
  - Thought



"Cognitivism"

- The "computer metaphor"
  - Perception is like input processing
  - Action is like output processing
  - Cognition is like *computation*:
    Manipulation of representations
    (Newell & Simon, Pylyshyn)
  - The mind is the software (Block)
  - The hardware is separate

Newell & Simon



Newell & Simon, 1972, Fig 2.1

What is thinking?

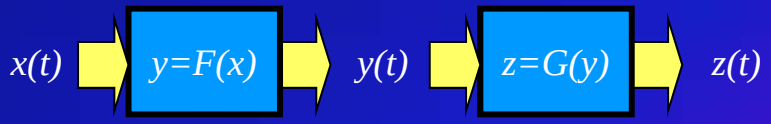
- *"[T]hinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures"* Thagard, *The Stanford Encyclopedia of Philosophy* 2008
- *"The principle function of the central nervous system is to represent and transform information and thereby mediate appropriate decisions and behaviors."* deCharms & Zador, *Ann. Rev. Neurosci.* 2000

Representations

- "Descriptive" representations
  - Capture ***knowledge*** about the world and the organism
  - Explicit
  - Objective, accurate to external reality, uncontaminated by internal states
  - Examples:
    - Reconstructed visual image
    - 3-D map of the world
    - Desired path of the hand in space
    - Expected value of a choice

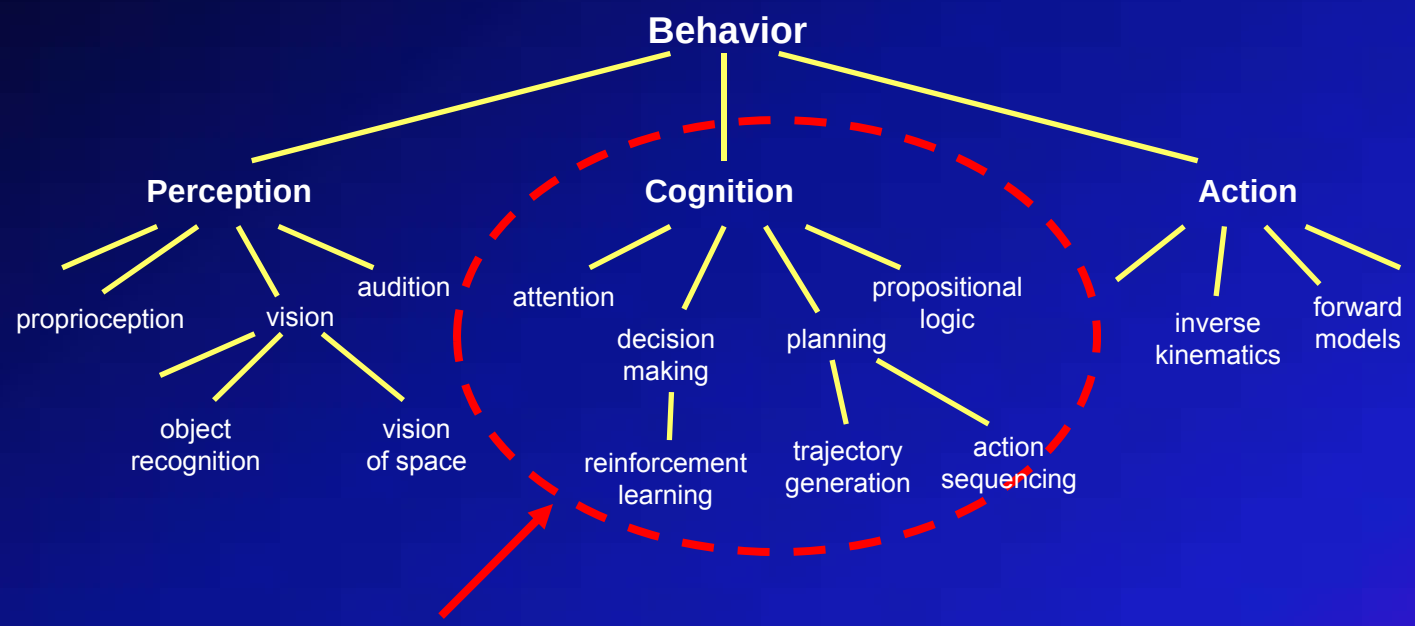
David Marr

Descriptive representations delineate the conceptual borders between the processes that produce them and the processes that consume them.



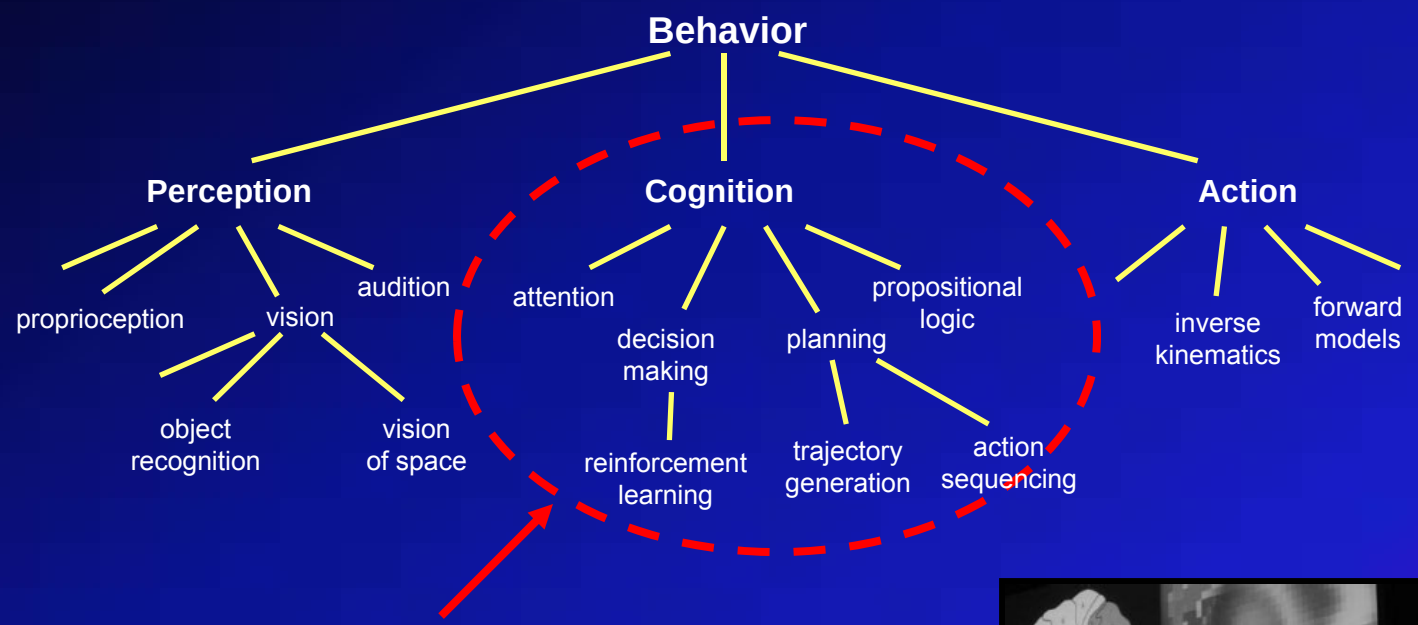
This provides a method for a functional decomposition of the large problem of behavior into smaller (and thus presumably more tractable) problems

Functional decomposition

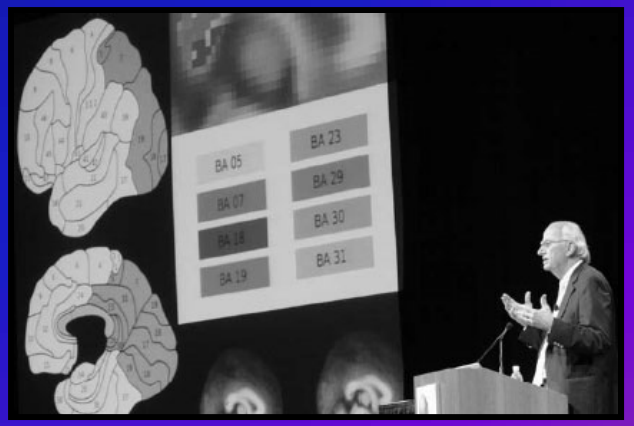


- Cognitive Science

Functional decomposition



- Cognitive **Neuro**science
  - How are psychological / cognitive functions produced by the brain?
  - Based on the concepts of cognitivism
    - Computation, descriptive representations, working memory, attentional filters, motor programs, etc.



Michael Gazzaniga

What is thinking?

- *"[T]hinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures"* Thagard, *The Stanford Encyclopedia of Philosophy* 2008

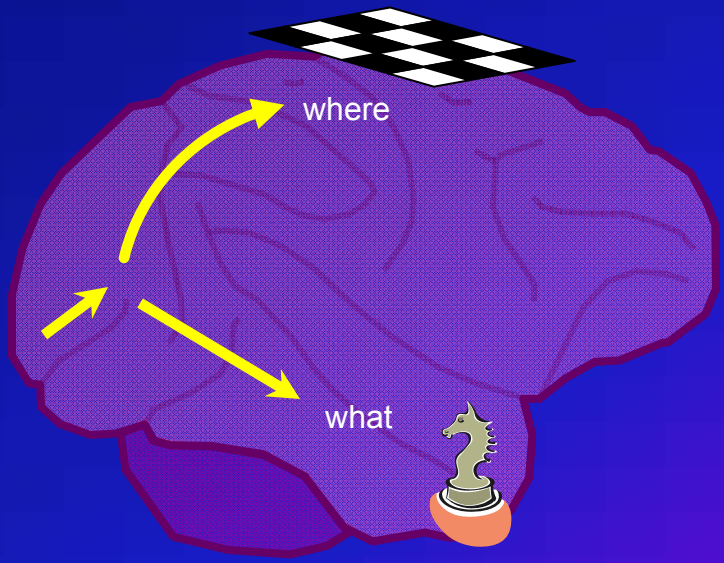
- *"The principle function of the central nervous system is to represent and transform information and thereby mediate appropriate decisions and behaviors."* deCharms & Zador, *Ann. Rev. Neurosci.* 2000

- *"The task for the years ahead is to produce a study of mental processes, grounded firmly in empirical neural science, yet still fully concerned with problems of how internal representations and states of mind are generated"*

  Kandel, Schwartz, Jessel, Siegelbaum, Hudspeth, *Principles of Neural Science* 2013

Where is the representation of the world?

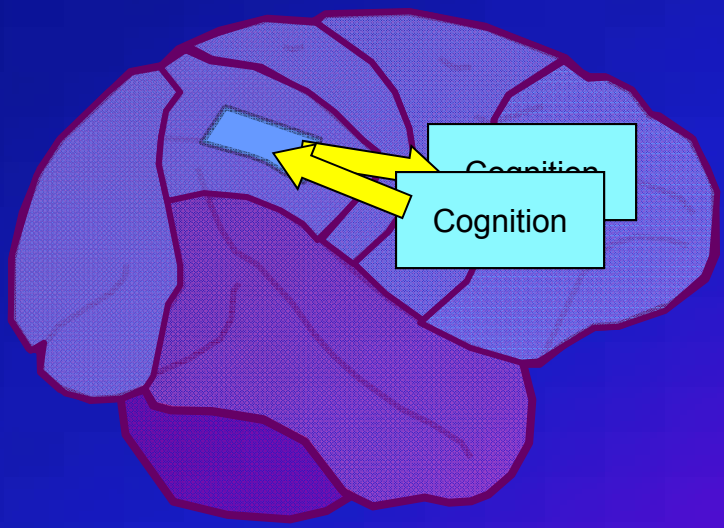
- The visual system
  - Two visual processing streams:
    - ventral “what”
    - dorsal “where”
  - Separate regions analyze color, motion, form, etc.
  - Separate regions for near and far space
- Binding problem
  - How to create the unified representation of the world that is needed as input for cognition?



Where are the perception, cognition, and action systems?

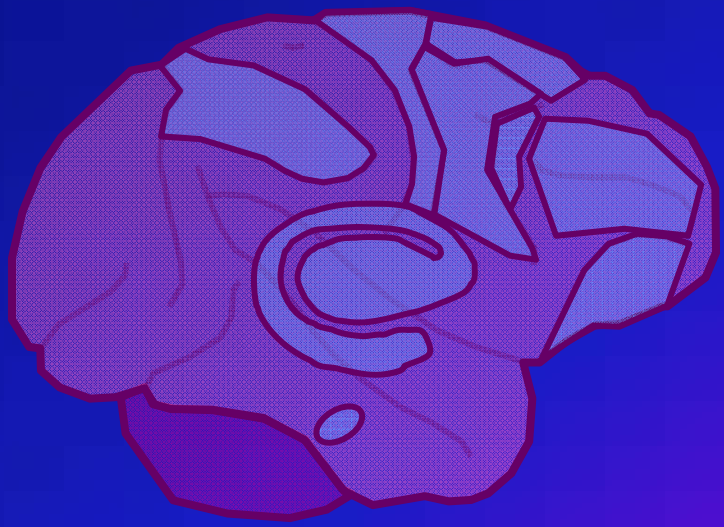
- Sensory and motor regions
- "Association" regions
  - Appear to first encode sensory, then motor representations
  - Even true for single cells

  Example: Lateral intraparietal area
  - Represents ***attended*** stimuli (before cognition, input)
  - Represents ***intended*** actions (after cognition, output)
  - How could it be both?
  - In what box does it belong?
- Similar questions for other parietal and frontal regions



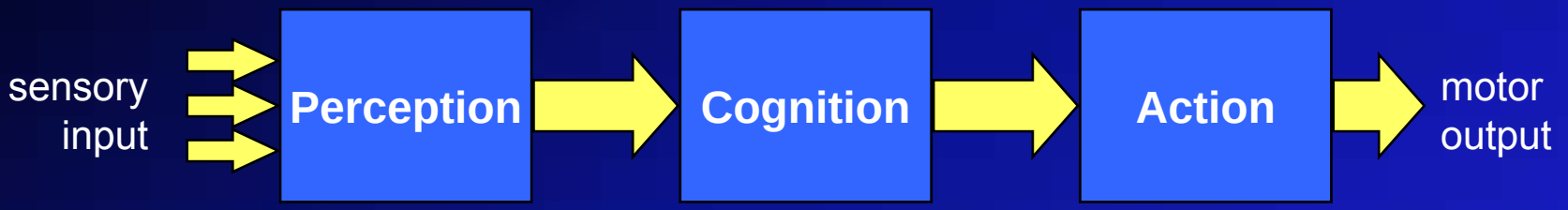
Where is a decision made?

- Neural correlates of decision variables
  - in prefrontal and orbitofrontal cortex
  - also in parietal cortex
  - premotor cortex
  - supplemental motor area
  - frontal eye fields
  - basal ganglia
  - even primary motor cortex
  - and the superior colliculus



- Timing of decisions
  - All these regions reflect decision almost simultaneously (~150ms)
    - Ledberg et al. (2007) *Cerebral Cortex*

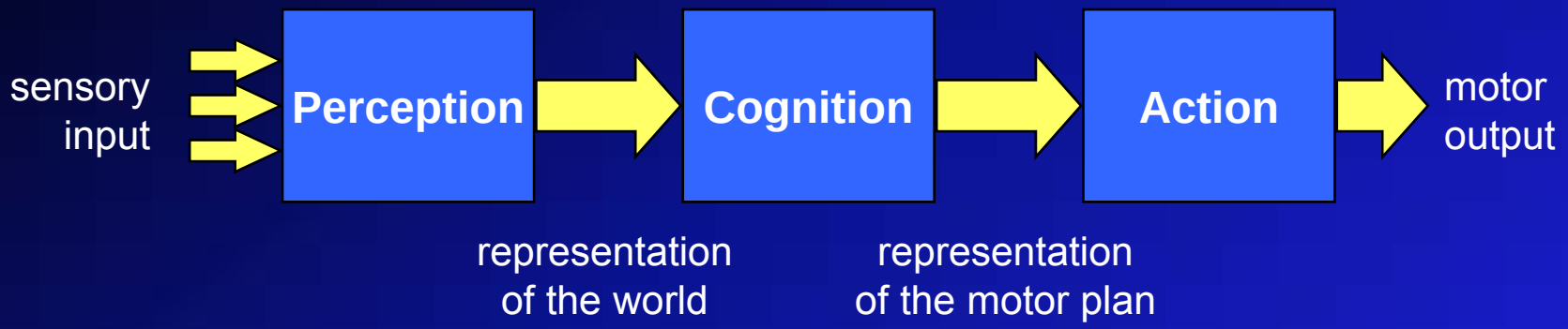
Psychological architecture for behavior



- Some observations:
  1. It inherits its structure from dualism – a view that everyone rejects
  2. Designed to explain abstract problem-solving – not all behavior
  3. Its concepts were developed under the explicit assumption that the substrate doesn't matter

- Perhaps it should not be surprising that this model has difficulty explaining neural data…

What is behavior?



- University courses
- Textbooks
- Journals
- Conferences
- Academic departments
- Grant review committees
- Scientists
- Questions we ask
- Theories we propose

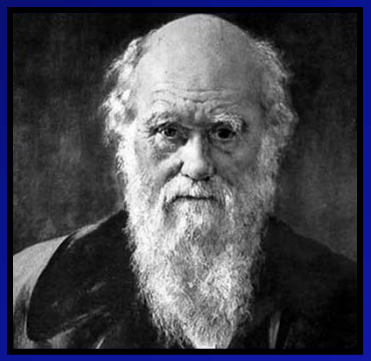
Q: From where does this view originate?

Q: What questions *should* we ask?

Evolution

"Nothing in biology makes sense except in the light of evolution"

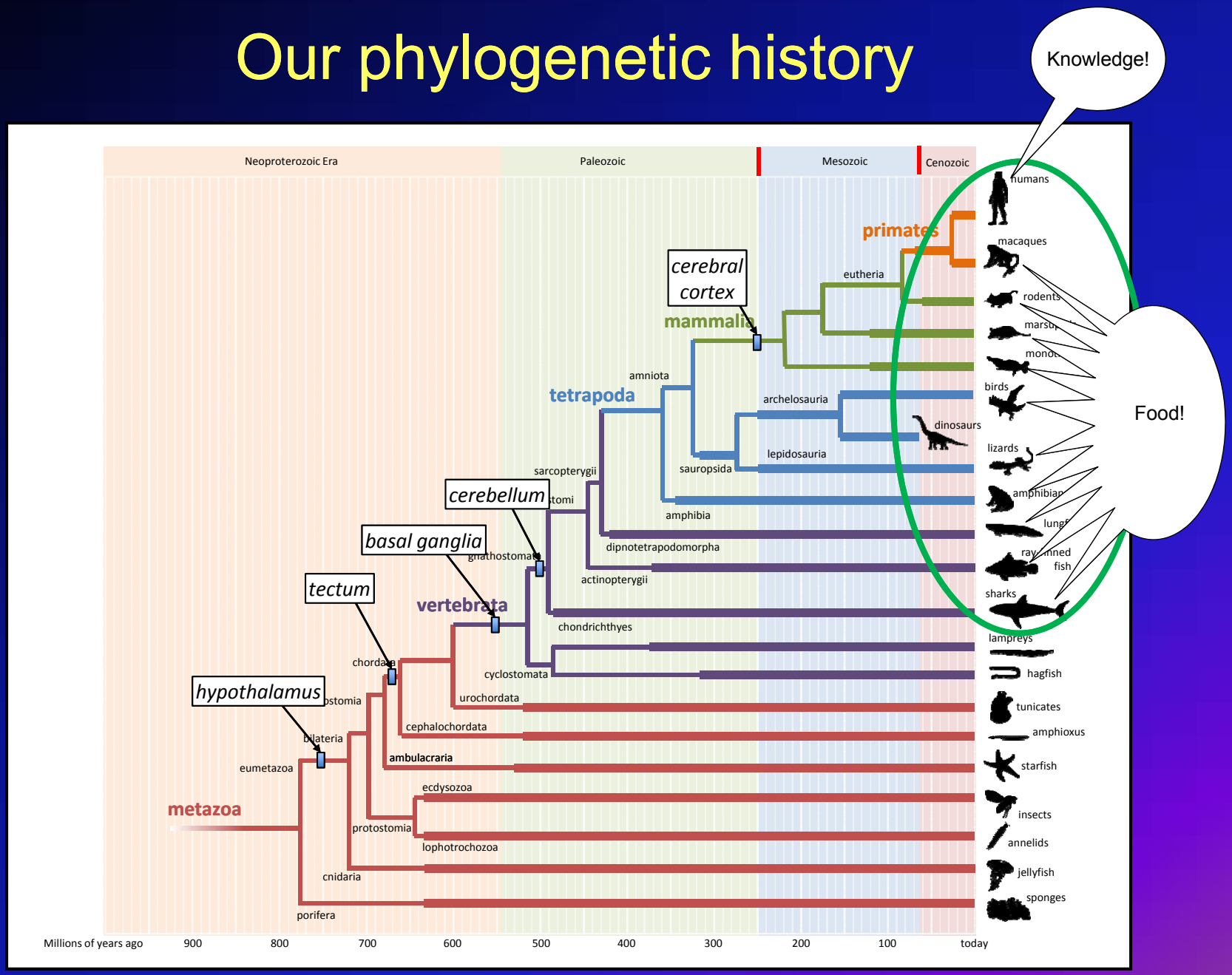
Theodosius Dobzhansky, 1973



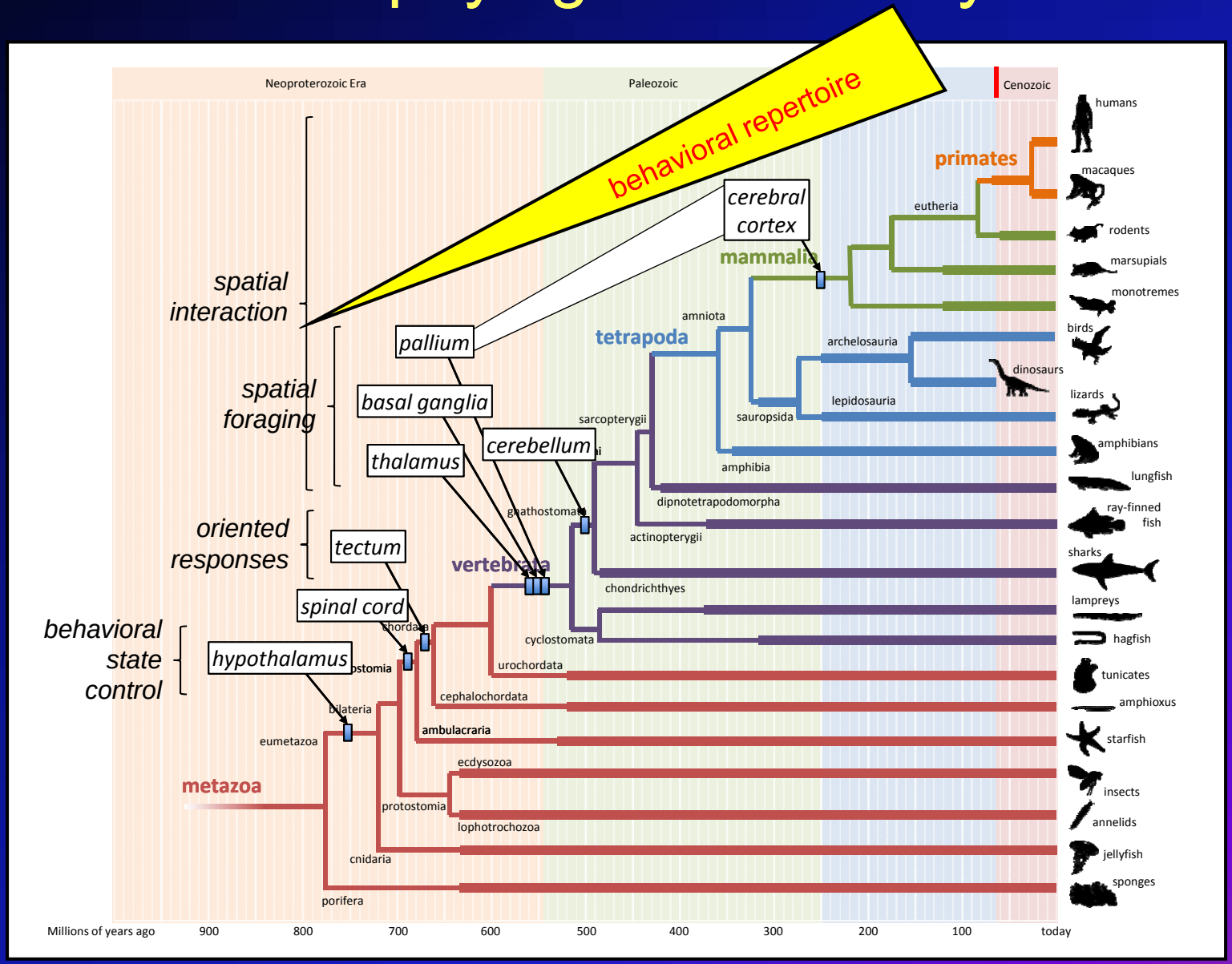
Darwin

- Two key concepts:
  - Natural selection
    - What is the selective advantage of a mechanism?
  - Descent with modification
    - What are its phylogenetic origins?

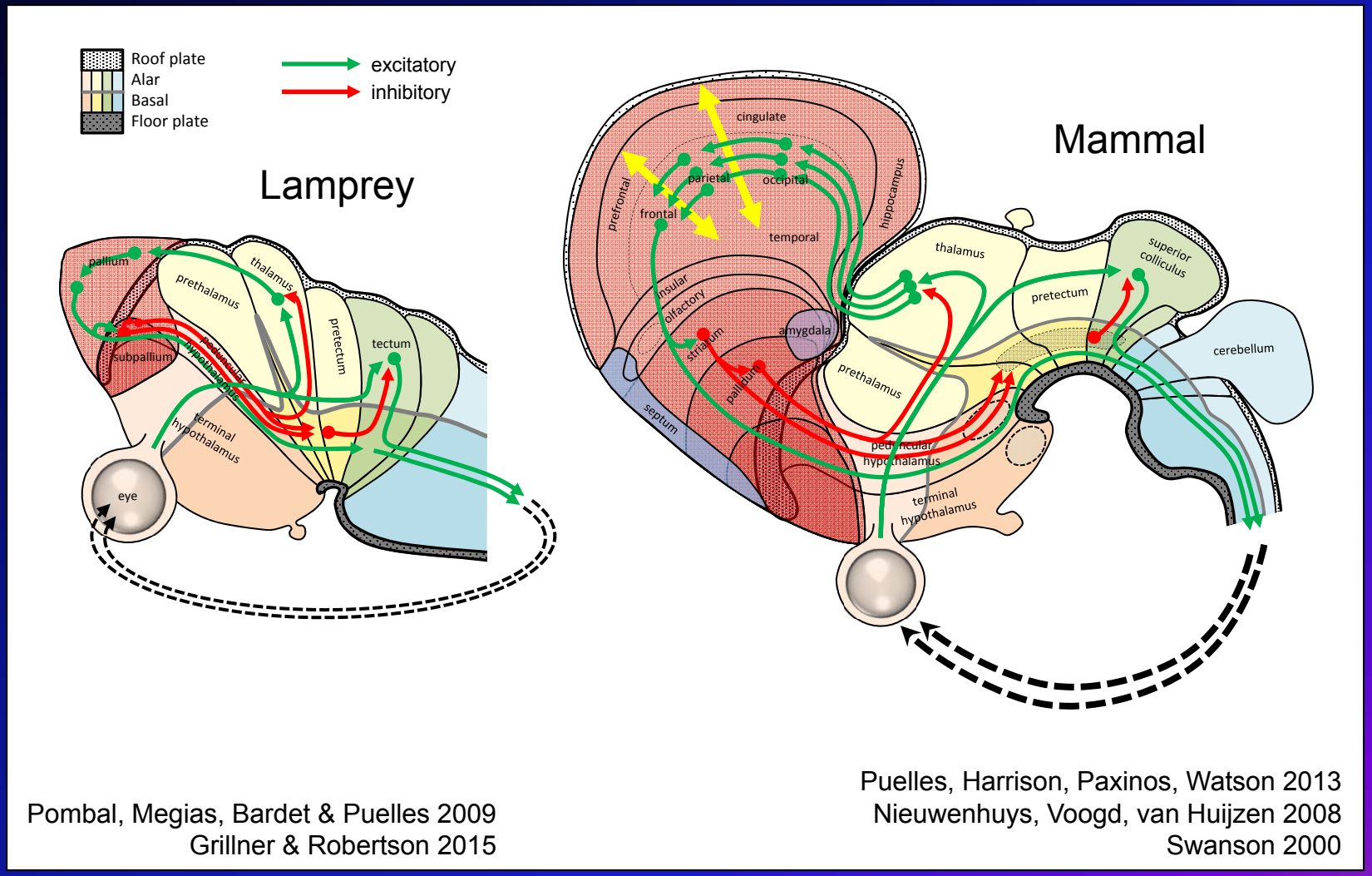
Our phylogenetic history



Our phylogenetic history



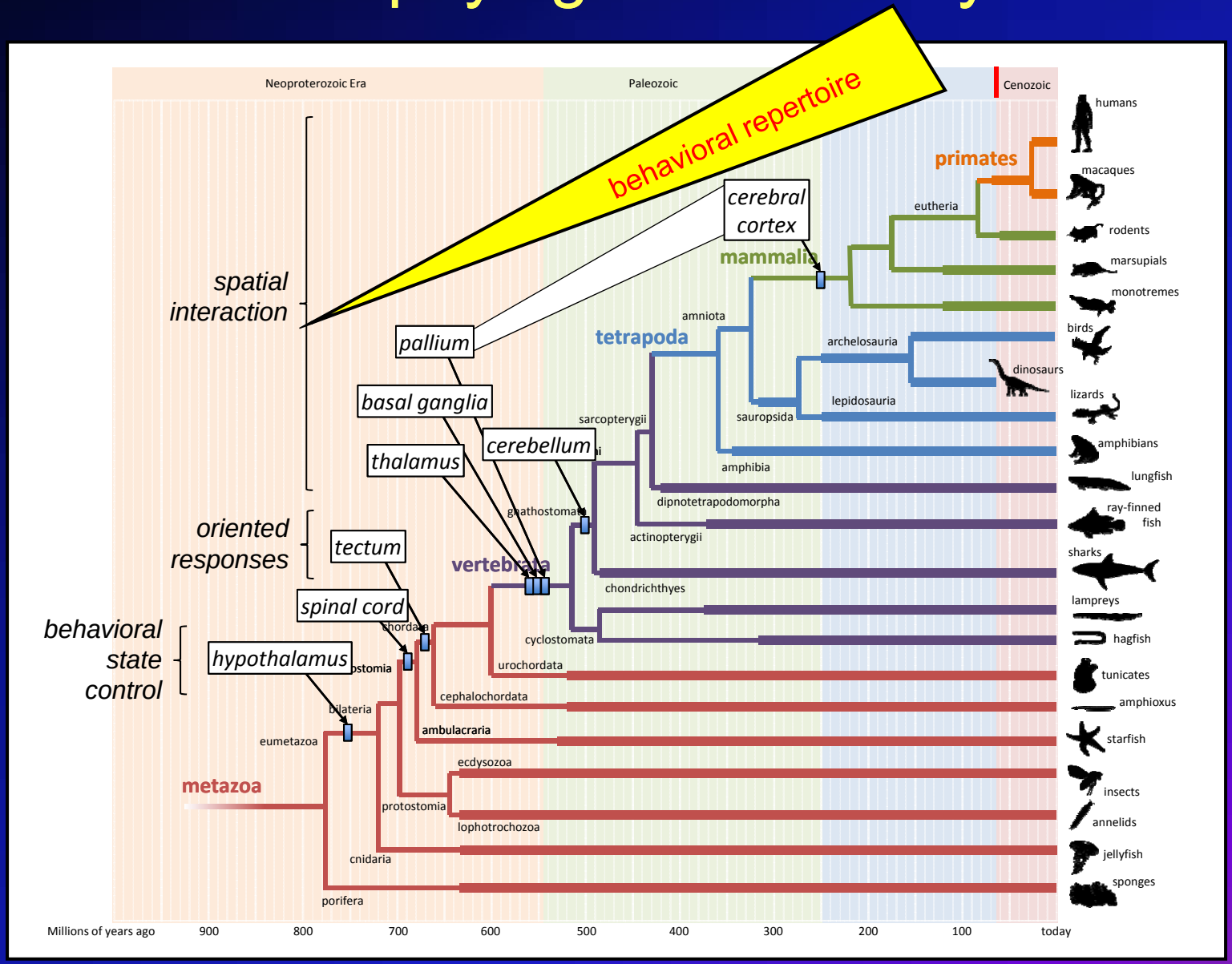
Basic vertebrate plan



Pombal, Megias, Bardet & Puelles 2009
Grillner & Robertson 2015

Puelles, Harrison, Paxinos, Watson 2013
Nieuwenhuys, Voogd, van Huijzen 2008
Swanson 2000

Our phylogenetic history



What are living systems?

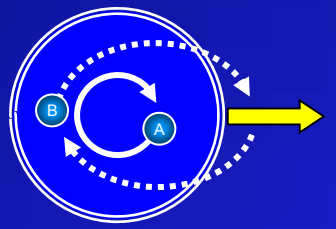
- Input-output systems?

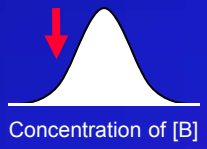
Control systems

- All living things are *control systems*
  - Ex: Biochemistry
    - Suppose there is some substance A necessary for survival
    - Suppose there's a catalyst for creating A whose action is regulated inversely by the concentration of A
    - Feedback control system
    - Exploits consistencies in the laws of chemistry
    - Control loop within the organism: "**Physiology**"

Control systems

- Control systems can extend beyond the skin
  - Ex: Kinesis
    - Suppose substance B cannot be produced within the body, must be absorbed from the world
    - If the local concentration of substance B is below desired levels, move randomly
    - Exploits statistics of nutrient distributions (assumes that there is more elsewhere)



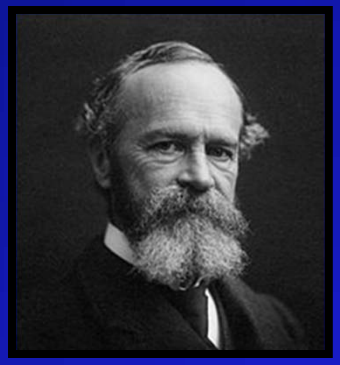


    - Control loop that extends outside the skin: "**Behavior**"
  - Reliable motor-sensory contingencies exist
    - Statistics of food distributions (move → find food)
    - Laws of optics and mechanics (contract muscle → arm moves)
    - Laws of external physics (push on obstacle → it yields)

Different ways of looking at behavior

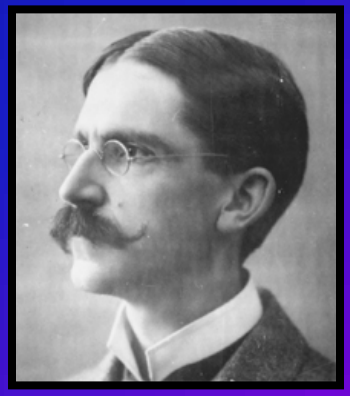
1. Given a perception, produce the best action
   - *"The whole neural organism … is, physiologically considered, but a machine for converting stimuli into reactions"*

     William James, 1890



2. Of the possible actions, produce that which results in the best perception
   - *"What we have is a circuit… the motor response determines the stimulus, just as truly as sensory stimulus determines movement"*

     John Dewey, 1896



Meaning is not a problem

- Interactive behavior always has meaning
  - Some states are good, some are not
  - Some stimuli indicate desirable states, some don't
  - Some actions produce good results, some don't

- Questions:
  - ~~How is meaning attached to symbols?~~
  - How do you control interactive behavior?

Different questions

- Ethology: Studies of animal behavior in the wild
  - What are the species-specific behavioral niches?
  - What are the “key stimuli” that animals use as cues for motivating different actions?
  - How does the brain implement “closed-loop” sensorimotor control?

Von Uexküll

Tinbergen

Lorenz & Von Holst

Different kinds of representations

- "Descriptive" representations
  - Capture ***knowledge*** about the world and the organism
  - Explicit
  - Objective, accurate to external reality, uncontaminated by internal states
  - Examples:
    - Reconstructed visual image
    - 3-D map of the world
    - Desired path of the hand in space
    - Expected value of a choice

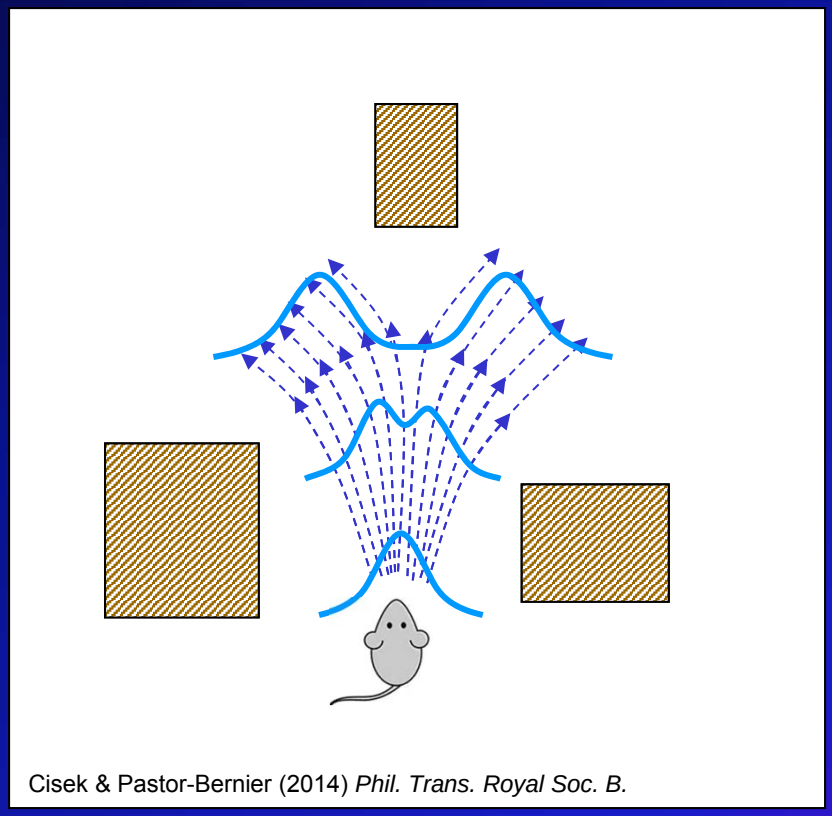
David Marr

- "Pragmatic" representations
  - Used to ***guide interaction*** between the world and the organism
  - Implicit
  - Subjective, mix external reality and internal state, often correlate with many variables at once
  - Examples:
    - Control signals guiding movement
    - Salience map
    - Subject-dependent opportunities for action ("***affordances***")

J.J. Gibson

Different kinds of behavior

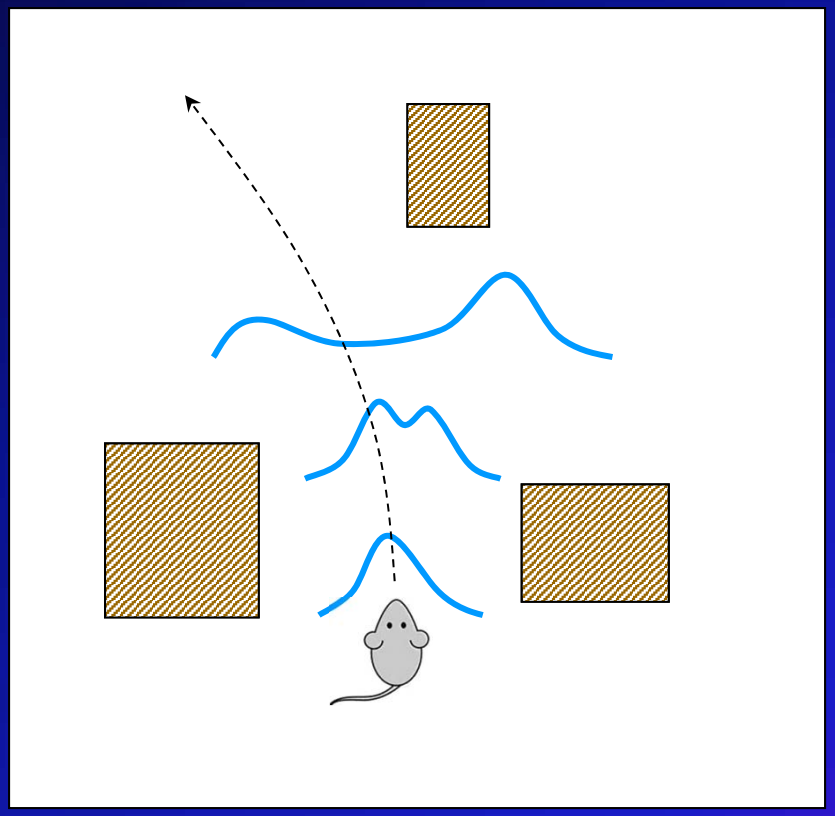
“affordances”



Cisek & Pastor-Bernier (2014) *Phil. Trans. Royal Soc. B.*

“desirability density function”

The choices themselves emerge from geometry



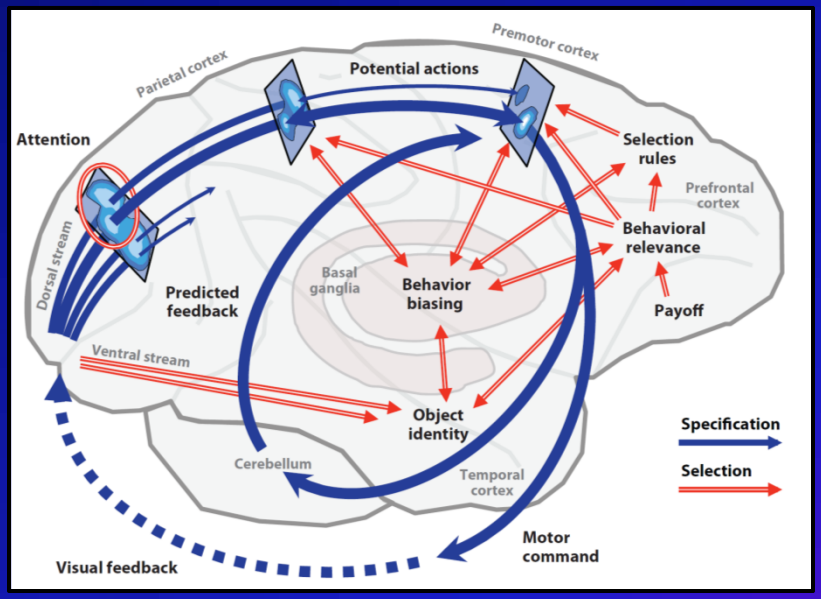
Must continuously specify and re-specify potential actions, evaluate costs/benefits, make decisions…

The choices themselves emerge from geometry

Everything is constantly changing

“Affordance competition hypothesis”

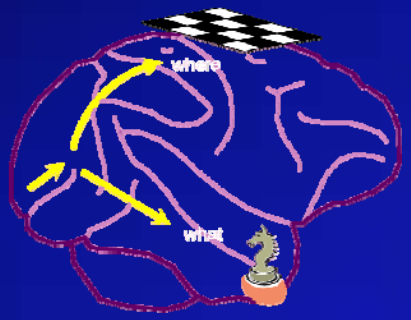
- **Action specification**
  - Dorsal visual stream (for visually guided action)
  - Multiple *competing* potential actions
- **Action selection**
  - Biasing that competition
  - Attentional selection
  - Decision-making
- **Execution**
  - Feedback control through the sensorimotor system
  - Forward prediction through cerebellum
- **All of these can occur in parallel during overt activity**



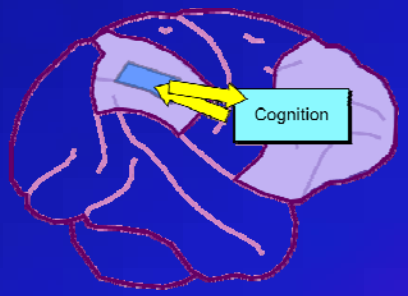
Cisek (2007) *Phil.Trans.Royal Soc. B.*
Cisek & Kalaska (2010) *Annual Review of Neurosci.*

Explaining neural data

- Two visuo-motor systems

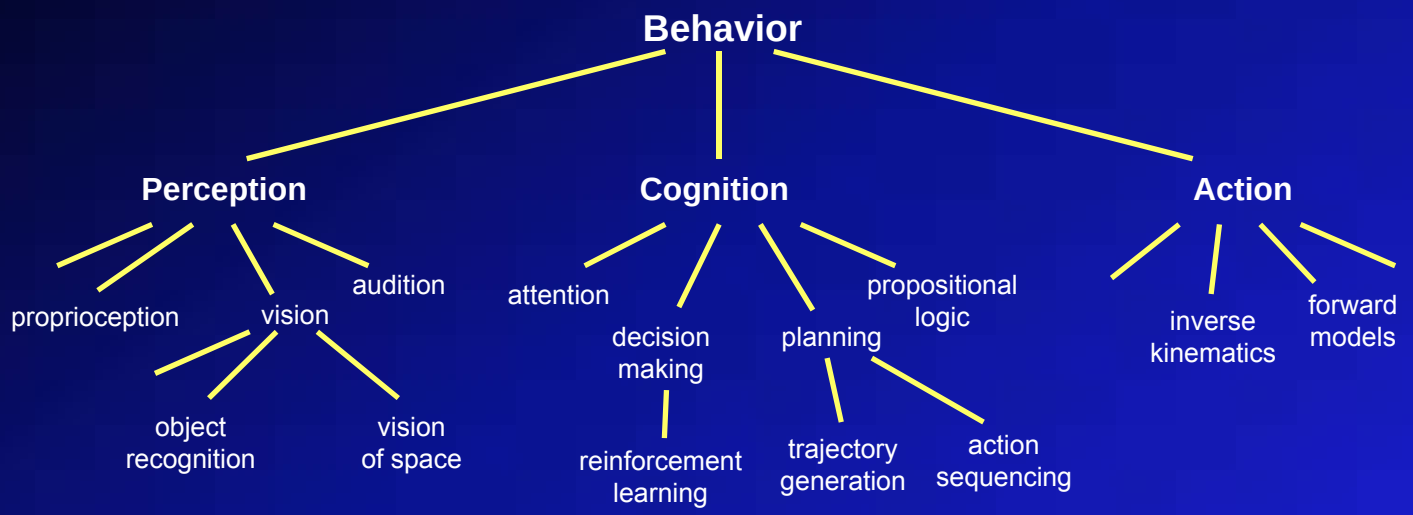


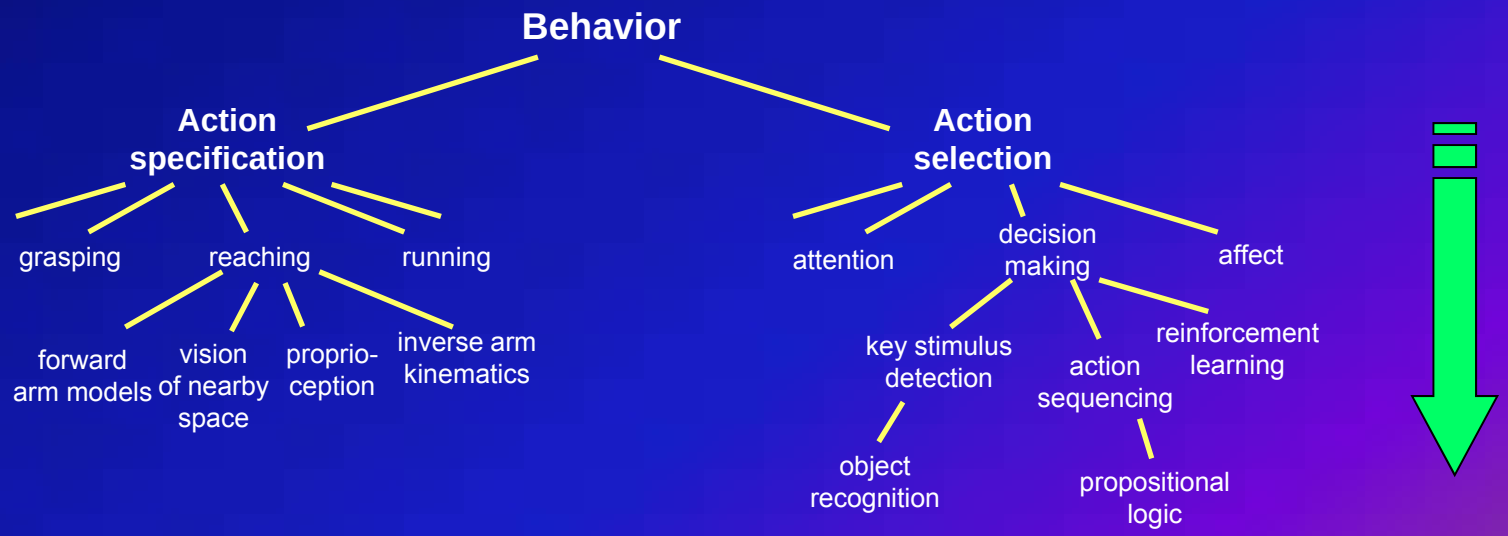
- Mixtures of sensory/motor/cognitive variables



- Decisions emerge in parallel, across a distributed network

A different functional decomposition





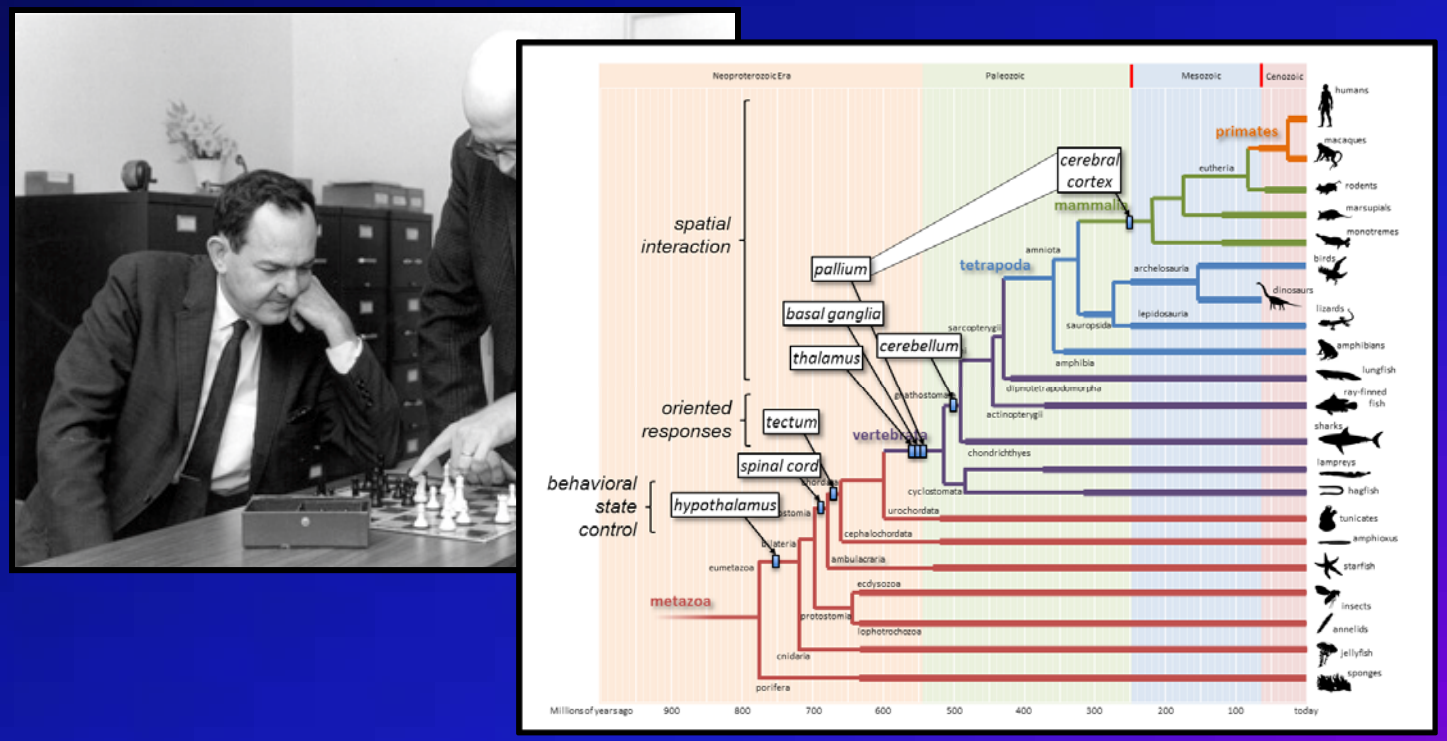
Meaning is not a problem

- Interactive behavior always has meaning
  - Some states are good, some are not
  - Some stimuli indicate desirable states, some don't
  - Some actions produce good results, some don't

- Questions:
  - ~~How is meaning attached to symbols?~~
  - How do you control interactive behavior?
  - How does cognition emerge from interactive behavior?

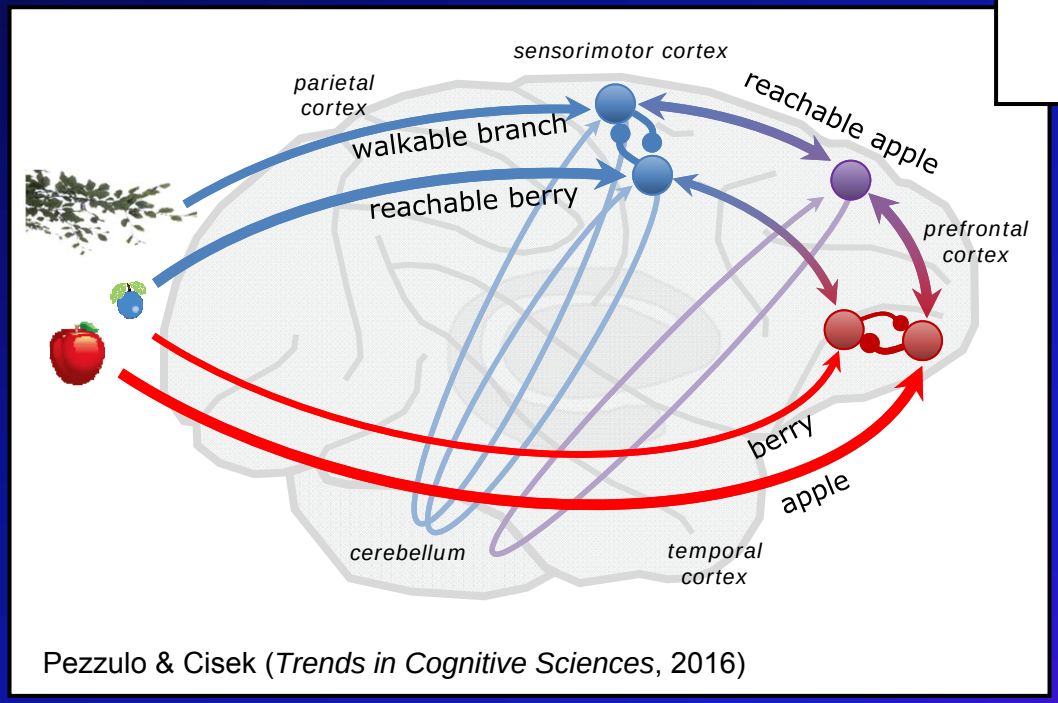
Evolution of cognition

- Humans *can* solve highly abstract tasks
  - A completely novel architecture? Serial cognitive model?
  - An elaboration of the ancestral parallel model?



Hierarchical affordance competition

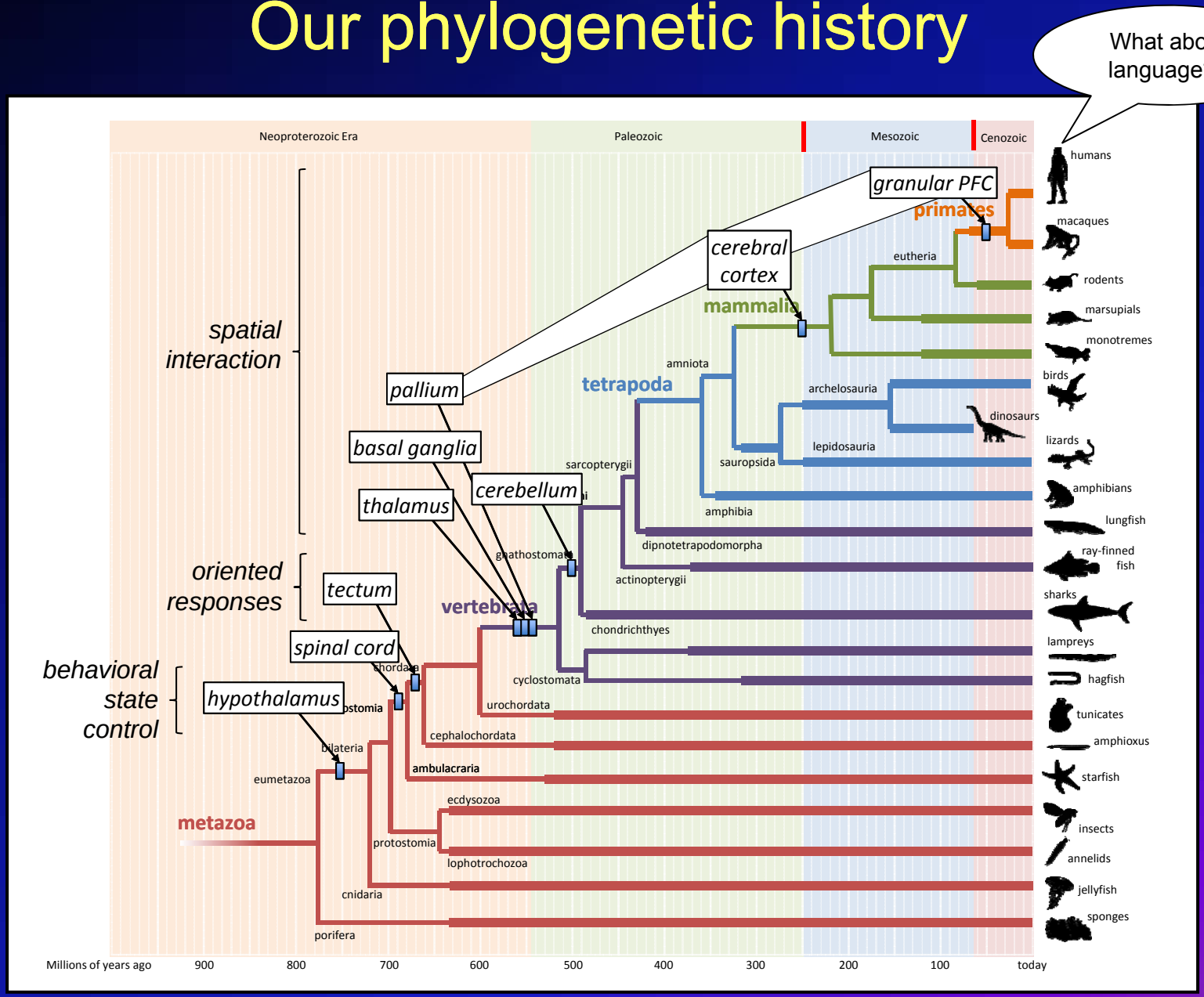
- Multi-level selection
  - High-level: goals
  - Low level: available affordances
  - Mid-level: predicted affordances



Pezzulo & Cisek (*Trends in Cognitive Sciences*, 2016)

Prediction of action consequences makes possible a linkage between levels

Our phylogenetic history

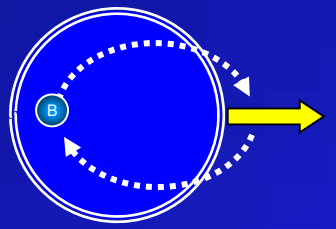


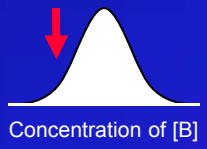
Control systems

- All living things are *control systems*
  - Ex: Biochemistry
    - Suppose there is some substance A necessary for survival
    - Suppose there's a catalyst for creating A whose action is regulated inversely by the concentration of A
    - Feedback control system
    - Exploits consistencies in the laws of chemistry
    - Control loop within the organism: "**Physiology**"

Control systems

- Control systems can extend beyond the skin
  - Ex: Kinesis
    - Suppose substance B cannot be produced within the body, must be absorbed from the world
    - If the local concentration of substance B is below desired levels, move randomly
    - Exploits statistics of nutrient distributions (assumes that there is more elsewhere)

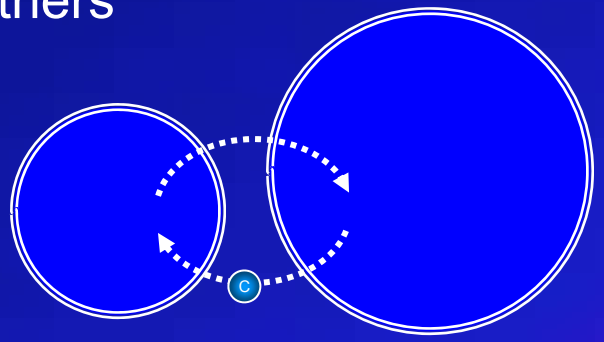




    - Control loop that extends outside the skin: "**Behavior**"
  - Reliable motor-sensory contingencies exist
    - Statistics of food distributions (move → find food)
    - Laws of optics and mechanics (contract muscle → arm moves)
    - Laws of external physics (push on obstacle → it yields)

Control systems

- Control systems can even extend through others
  - Suppose you're a helpless baby, and cannot obtain substance C yourself
  - Cry, mom will come to help
  - Laws of interaction exist
    - Baby cries → mother offers food
    - You show teeth → I back off
    - One gazelle runs → all run
  - Control loop that extends through others: "**Communication**"



- Communication is about persuasion (ex: this talk)

  Cisek "Beyond the computer metaphor" (1999) *Journal of Consciousness Studies*

- The question is not how meaning is attached to symbols, but how symbols are detached from meaningful interaction

  Pezzulo & Castelfranchi "The symbol detachment problem" (2009) *Cognitive Processing*

Back to AI

- How is any of this relevant to building artificial systems?
  - To build strong AI, must we know all about evolution??
- Insofar as neuroscience can inform AI research, we need a good model of the brain / behavior / thinking
  - Classical serial model is increasingly challenged by neural data
  - An alternative model is that the brain is a control system
    - Claude Bernard (1813-1878), Walter Cannon (1871-1945)
    - Cybernetics: Norbert Wiener, W. Ross Ashby
    - Perceptual control theory: William T. Powers
    - Ecological psychology: James J. Gibson
    - Active inference: Karl Friston
  - Promising fit with neural data
  - Meaning is not a problem

Back to AI

- The problem of semantics
  - Has *it* been solved?
  - It has been avoided
- The road to strong AI?
  - Systems that mediate interactive behavior
  - Reinforcement learning (e.g. Deep Q-networks)
    - Reward indicates desired states
    - Favors the emergence of representations that are useful
      - The features and categories are *functionally* meaningful
      - Predict future states
    - “Model free” versus “Model based”
    - Extending the control by discovering new interactions

Thank you

"The great end of life is not knowledge but action" T. H. Huxley (1825-1895)

"Your head is there to move you around" R.E.M. (1980-2011)

Current lab members
- Marie-Claude Labonté (technician)
- David Thura (postdoc)
- Matthew Carland (PhD student)
- Ayuno Nakahashi (PhD student)
- Julien Michalski (MSc student)
- Timothy Meehan (postdoc)

Alumni
- Jean-Philippe Thivierge (Ottawa)
- Thomas Michelet (U. Bordeaux)
- Valeriya Gritsenko (UWV)
- Ignasi Cos (Barcelona)
- Alexandre Pastor-Bernier (Cambridge)

Funding

Groupe de recherche sur le système nerveux central GRSNC

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

Fonds de la recherche en santé Québec

CIHR IRSC

NSERC CRSNG

THE EJLB FOUNDATION

Visiting & summer students
- Genevieve Aude Puskas
- Elisabeth Rounis (London)
- Stephany El-Murr
- Nicolas Belanger
- Julie Beauregard-Racine
- Charles-William Fradet
- Farid Medleg
- Elsa Tremblay
- Encarni Marcos (Barcelona)
- Jessica Trung
- Jean-François Cabana
- Albert Feghaly
- Gerard Derosiere (Brussels)
- Guido Guberman
- Philippe Castonguay

Collaborators
- John Kalaska (U. Montréal)
- Andrea Green (U. Montréal)
- Karim Jerbi (U. Montréal)
- Dang Nguyen (U. Montréal)
- Alain Dagher (McGill)
- Tom Shultz (McGill)
- Jean-Philippe Thivierge (Ottawa)
- Aaron Batista (U. Pittsburgh)
- Julie Duque (Brussels)
- Giovanni Pezzulo (Rome)
- Sven Bestmann (UCL)