Current Trends in Artificial Intelligence

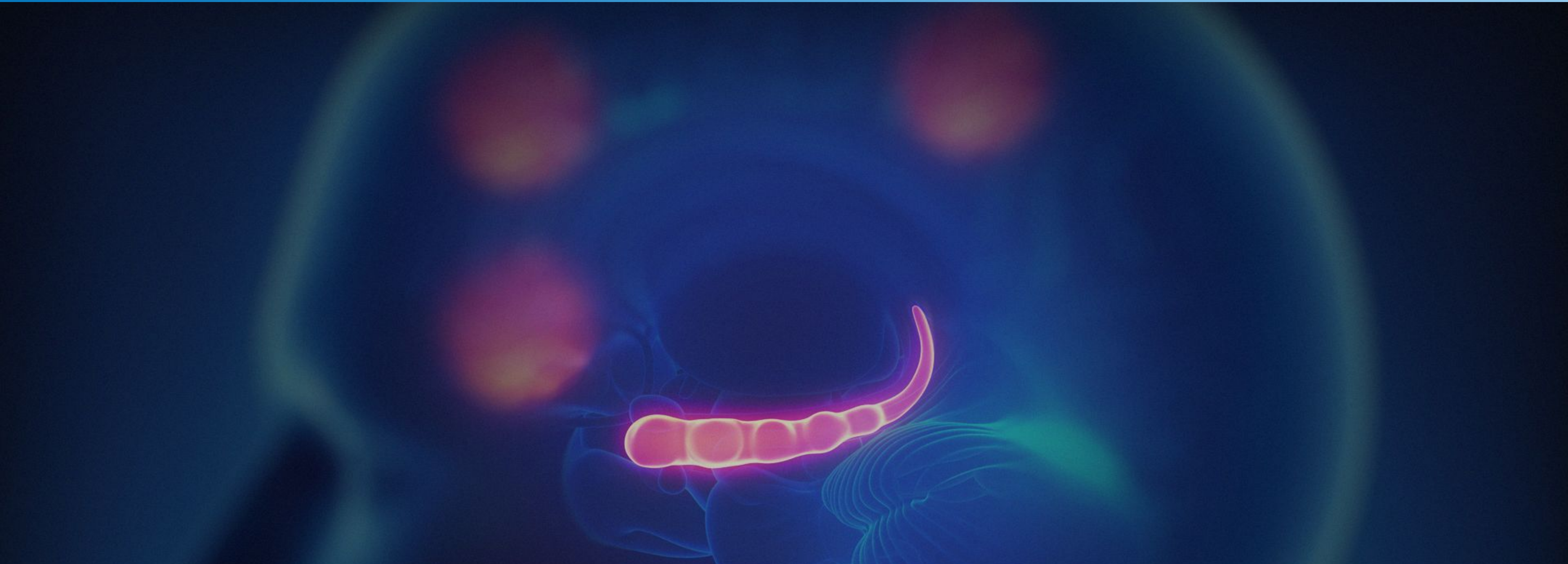
A few lessons from Complementary Learning Systems

DeepMind

Ulrich Paquet @ 2nd DTU Compute Workshop on Current Trends in Artificial Intelligence

How can we be creatively inspired by neuroscience? How can it help us think out of the box, to make progress in artificial intelligence? **

** Asked by a layman, myself :)



If neocortex is a neural net that learns online, from a stream of experience, then how is it possible to learn about specific things that only ever happen once?

Creative inspiration

Complementary Learning Systems Theory

Intelligent agents must possess (at least) **two learning systems**

In mammals they are the **neocortex** and **hippocampus**

- Neocortex gradually acquires structured knowledge representation
- Hippocampus quickly learns specifics of individual experiences

Idea: Learning specific arbitrary knowledge happens through hippocampus replaying it to neocortex over and over again during sleep, interleaved with other experiences

Complementary Learning Systems Theory

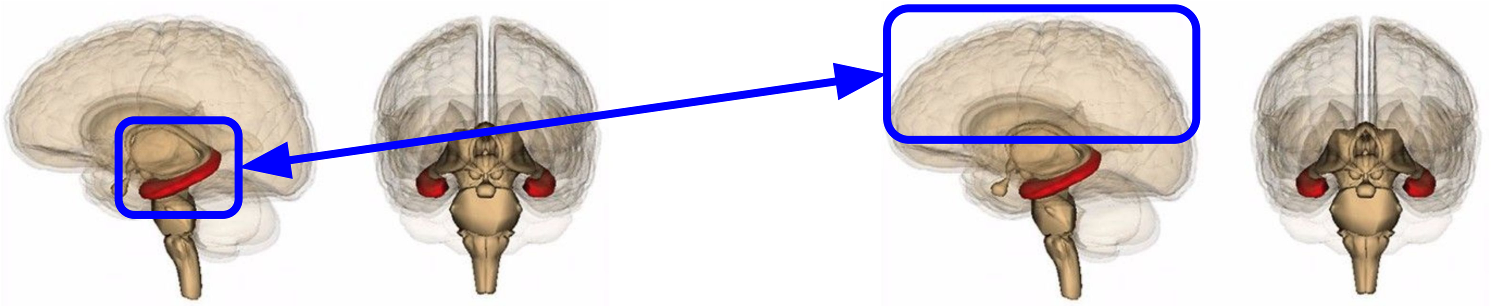
Wherefrom, and whereto?

The goal is to tie together a wide range of empirical data from different parts of neuroscience in a common connectionist theoretical framework.

- The original citation is McClelland, McNaughton & O'Reilly (1995)
- This talk draws from Kumaran, Hassabis, and McClelland (2016)
  - Update the theory in light of new empirical data
  - Argue that CLS provides **useful principles for machine learning**

Complementary Learning Systems Theory

Hippocampus	Neocortex



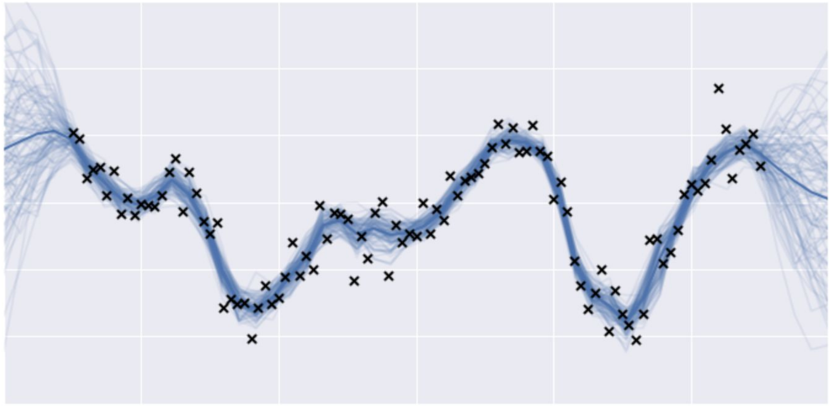
Hippocampus quickly learns specifics of individual experiences

Neocortex gradually acquires structured knowledge representation

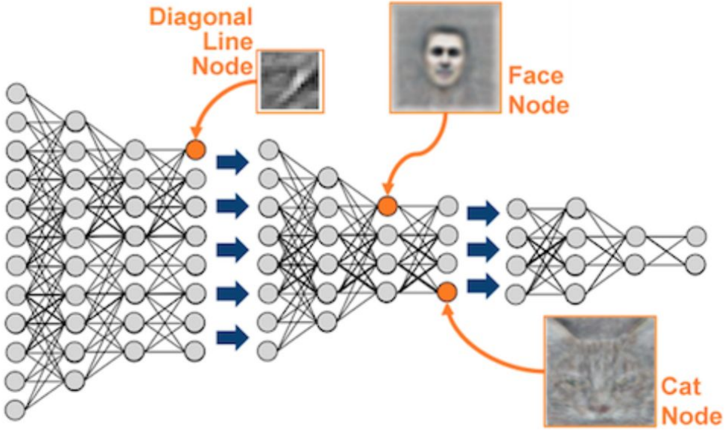
Source: Wikimedia/Life Sciences Database

Parallels with machine learning

Non-parametric methods	**Parametric methods**



Instance-based learning, where each experience has its **own coordinates**, capacity can be **increased** as required and parameters can grow with data.



Optimal parametric characterization of the **statistics of the environment** by learning gradually through **repeated**, interleaved exposure to large numbers of training examples.

Deep Networks

Composition of multiple processing layers

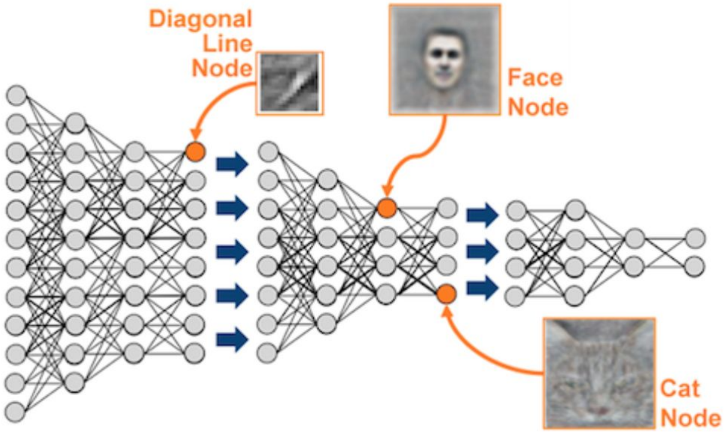
State of the art in image and speech recognition

Learns successively more abstract representations from sensory data

Oriented edges → Edge combinations → Object parts

Generalization

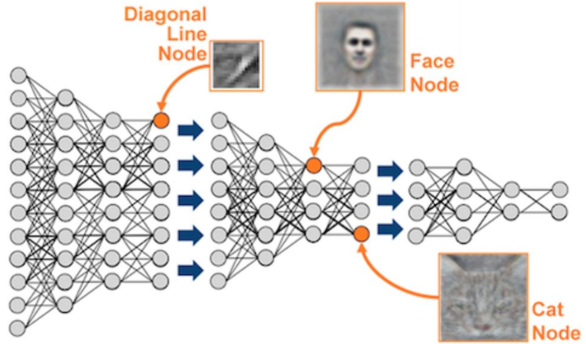
Parametric methods

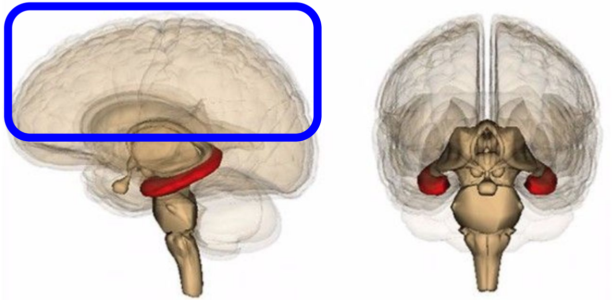


Optimal parametric characterization of the **statistics of the environment** by learning gradually through **repeated**, interleaved exposure to large numbers of training examples.

Structured knowledge representation in neocortex

Learning such a system has **limitations**





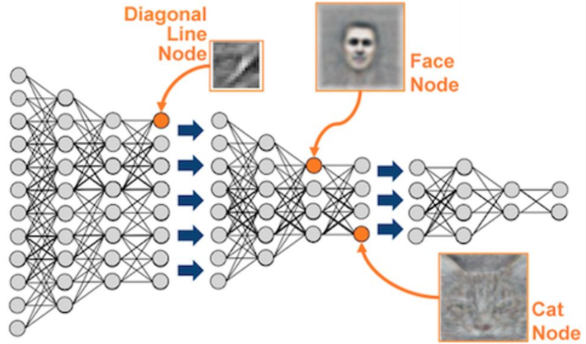
Structured knowledge representation in neocortex

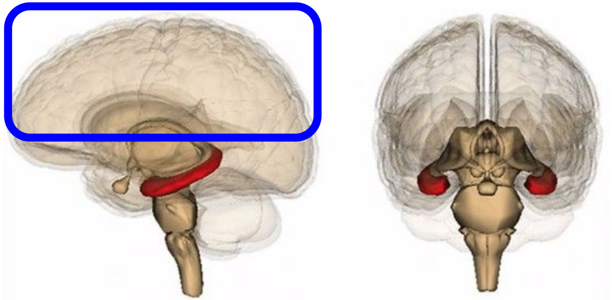
Learning such a system has **limitations**

- **A single experience counts!** Say a life-threatening situation
- Rapid adjustment of connection weights to accommodate new information can severely **disrupt** the representation of existing knowledge

“Catastrophic interference”

“Stability-plasticity dilemma”

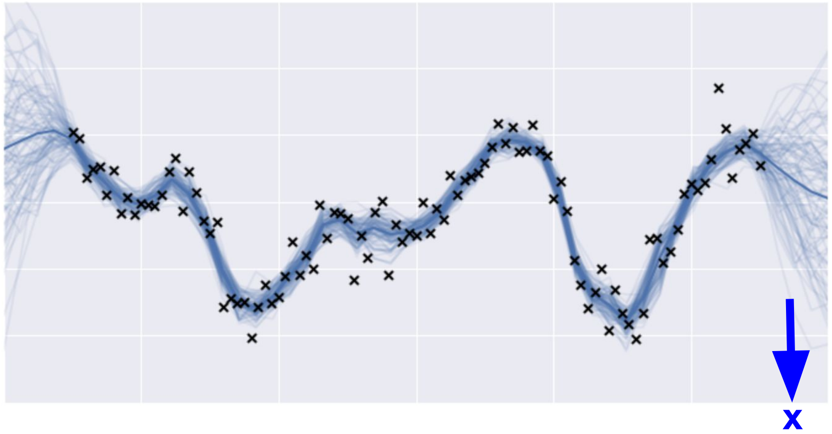




Gaussian Processes

A single experience counts

Non-parametric methods



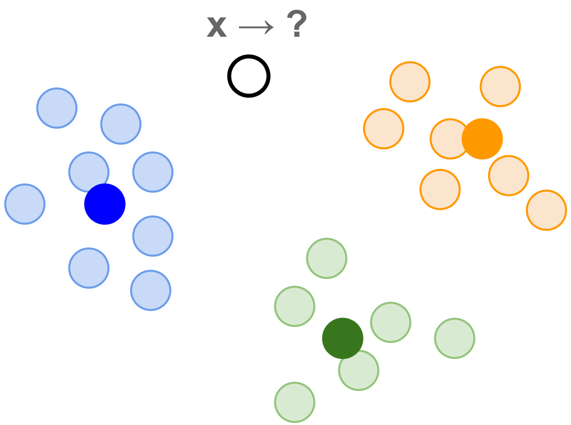
Instance-based learning, where each experience has its **own coordinates**, capacity can be **increased** as required and parameters can grow with data.

- **One-shot** accumulation of experience and learning as a new input **x** becomes a template to be compared against
- Predictions by comparing an input with known templates
- No successively more abstract representations

Dirichlet Process Mixture Models

A single experience counts

Non-parametric methods

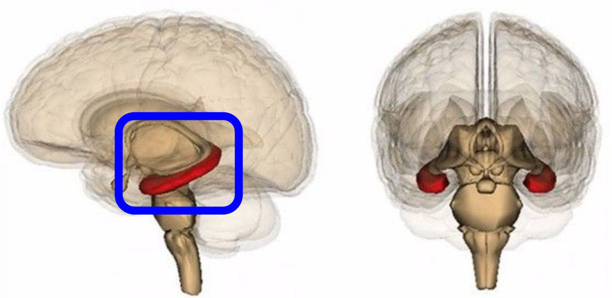


Instance-based learning, where each experience has its **own coordinates**, capacity can be **increased** as required and parameters can grow with data.

- Model data density through templates (cluster centers)
- Can **dynamically** “on the fly” **allocate more capacity** for a surprising new input input **x**
- Uses existing capacity for familiar inputs

Complementary Learning Systems Theory

Hippocampus



Hippocampus quickly learns specifics of individual experiences

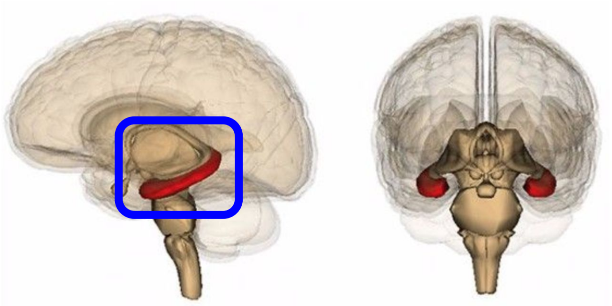
Instance based representation in the hippocampal system

- **Rapid** and relatively individuated storage of information about individual items or experiences
- CLS proposes that the HC and MTL structures support the **initial** storage of item-specific information.
- Role in **recognition memory** for specific items

Source: Wikimedia/Life Sciences Database

Complementary Learning Systems Theory

Hippocampus



Hippocampus quickly learns specifics of individual experiences

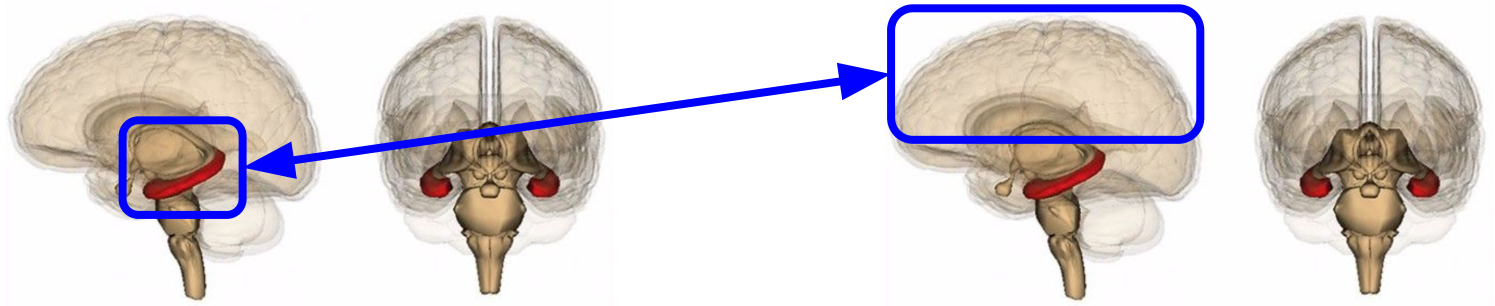
Involved in the formation of **episodic memory** as well as **spatial memory** used in navigation

- Navigation - **linkage of spatial locations**
- Episodic memory - **linkage of events**
- Both may depend critically on temporal sequence encoding (a code that captures time ordering)

Source: Wikimedia/Life Sciences Database

Complementary Learning Systems Theory

Hippocampus	Neocortex



Systems-level consolidation

- Idea: **Gradual** cortical learning driven by **replay** of new information
- Interleaved with other activity to minimize disruption of existing knowledge during the integration of new information

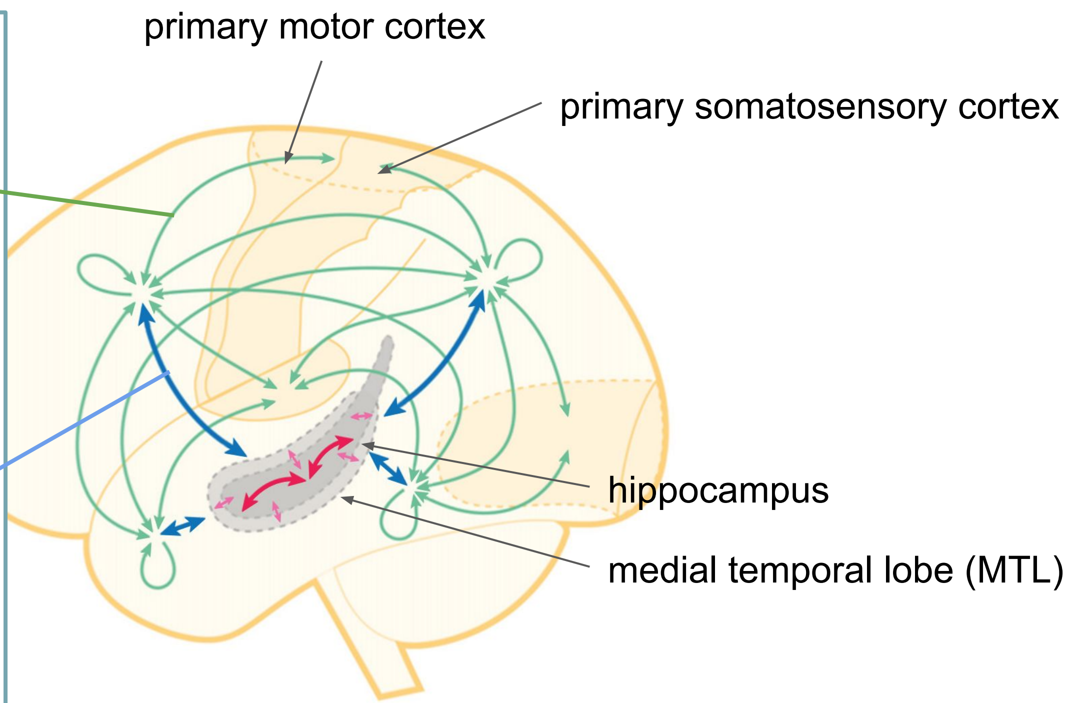
Source: Wikimedia/Life Sciences Database

CLSs and their interactions

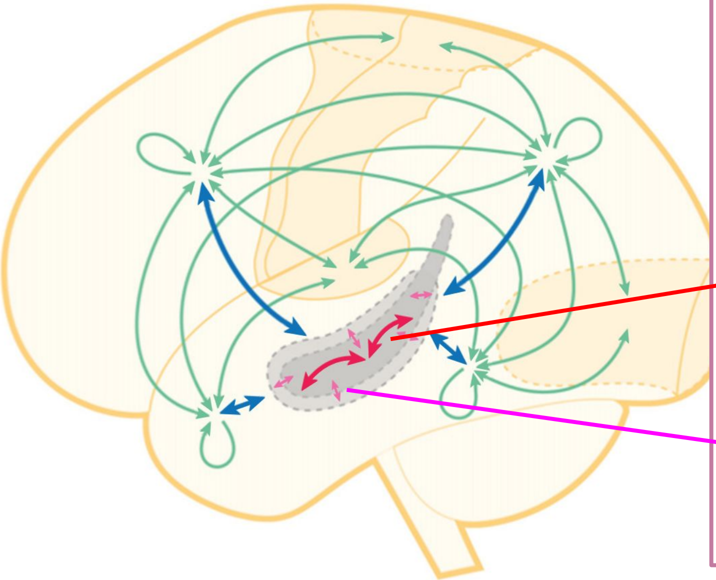
bidirectional connections within and between integrative neocortical association areas, for gradual acquisition of **structured knowledge** through interleaved learning

bidirectional connections between neocortical areas and the MTL for **storage, retrieval and replay**

= part of the structure-specific neocortical learning system in CLS theory



CLSs and their interactions



rapid synaptic plasticity crucial for the rapid **binding** of the elements of an event into an integrated hippocampal representation

initial learning of arbitrary new information

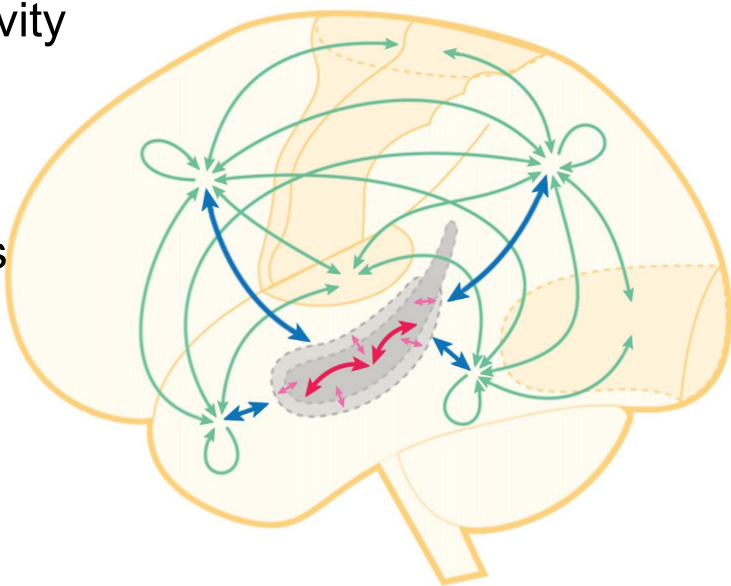
connections within hippocampus

connections between hippocampus and surrounding MTL cortices

CLSs and their interactions

systems-level consolidation
=

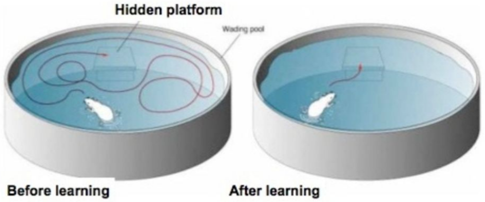
1. hippocampal activity during **replay**
2. → neocortical association areas
3. → learning within intra-neocortical connections



systems-level consolidation is **complete** when memory retrieval can occur without the hippocampus

memory retrieval = **reactivation** of the relevant set of neocortical representations

(later in talk)



Evidence supporting CLS theory #1

Hippocampal replay

- Replay of recent experiences occurs during **offline** periods

sleep + rest

- The hippocampus and neocortex **interact** during replay in **systems-level consolidation**

(Optogenetic blockage of CA3 output in transgenic mice after learning in the contextual fear paradigm specifically reduces sharp-wave ripple (SWR) complexes in CA1 and impairs consolidation…)

Place cells and replay in rodent hippocampus

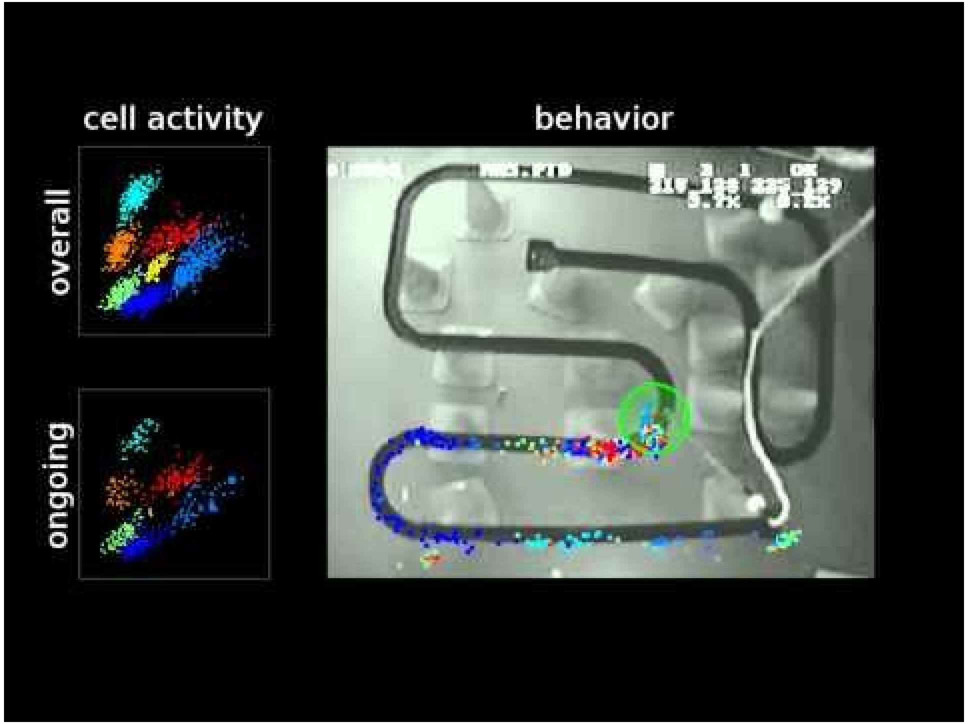
https://www.youtube.com/watch?v=4LnTWixQbbs

Empirical evidence of replay

- Sleep → hippocampal neurons exhibit **large irregular activity** (LIA) patterns that are distinct from the activity patterns observed during active states
- LIA = synchronous discharges (though to be) initiated in CA3 produce **sharp-wave ripples** (SWRs) which are propagated to neocortex
- SWR → **reactivation** of recent experiences, expressed as the **sequential firing of place cells**
- Replay events are **time-compressed by a factor of 20**
- Single event **replayed many times** during a single sleep period

SWR = spontaneous neural activity occurring within the hippocampus during periods of rest and slow wave sleep, evident as negative potentials (i.e. sharp waves). Transient high-frequency (~150Hz) oscillations (i.e. ripples) occur within these sharp waves, which can reflect the reactivation of activity patterns that occurred during actual experience, sped up by an order of magnitude.

Hippocampal place cells recorded in the Wilson lab at MIT



https://www.youtube.com/watch?v=lfNVv0A8QvI

Circumventing the statistics of the environment

Hippocampus **“marks”** **salient but statistically infrequent experiences**

Why? So that...

- such events are **not swamped** by the wealth of **typical experiences**
- such events are preferentially stabilized and replayed into the neocortex, allowing knowledge structures to incorporate this new information

Idea of **adaptive reweighting**

Why **adapt**? Maladaptive consequences like post-traumatic stress disorder. A unique aversive experience may be transformed into a persistent and dominant representation through a runaway process of repeated reactivation

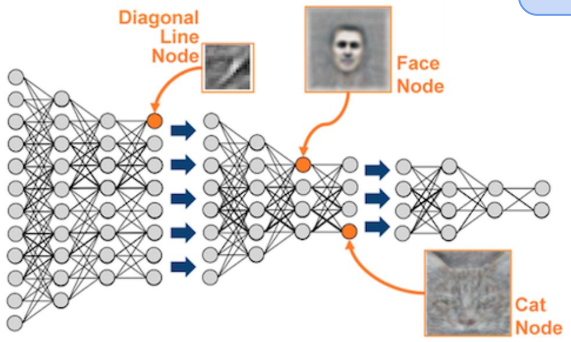
Role of replay of hippocampal memories

Big picture:

- Replay allows **goal-dependent weighting of experience statistics**
- The neocortex does **not** have to be a slave to the **statistics of its environment**

reweighting: surprising / novel / high in reward value (positive or negative) / high in information content (reducing uncertainty about best action in a given state)

recap of today's "big tool"...



Optimal parametric characterization of the **statistics of the environment** by learning gradually through **repeated**, interleaved exposure to large numbers of training examples.

Evidence supporting CLS theory #2

The role of the hippocampus in memory

- Bilateral **damage** to the hippocampus profoundly affects **memory for new information**.
- Language + reading + general knowledge + acquired cognitive skills = **intact**
- New types of learning are hippocampus dependent

Role of hippocampus in memory

Hippocampal lesion

- Lost ability to form new memories
- Remote memories spared
- Perceptual and motor skills spared
- Systems consolidation

https://www.youtube.com/watch?v=c62C_yTUyVg

Evidence supporting CLS theory #3

Hippocampus supports core <u>computations</u> and <u>representations</u> of a fast-learning episodic memory system

- Episodic memory = the collection of past personal experiences that occurred at a **particular time and place**.
- Remembering a 6th birthday party: EM allows an individual to figuratively travel back in time to remember the event that took place at that particular time and place

Evidence supporting CLS theory #3

Hippocampus supports core <u>computations</u> and <u>representations</u> of a fast-learning episodic memory system

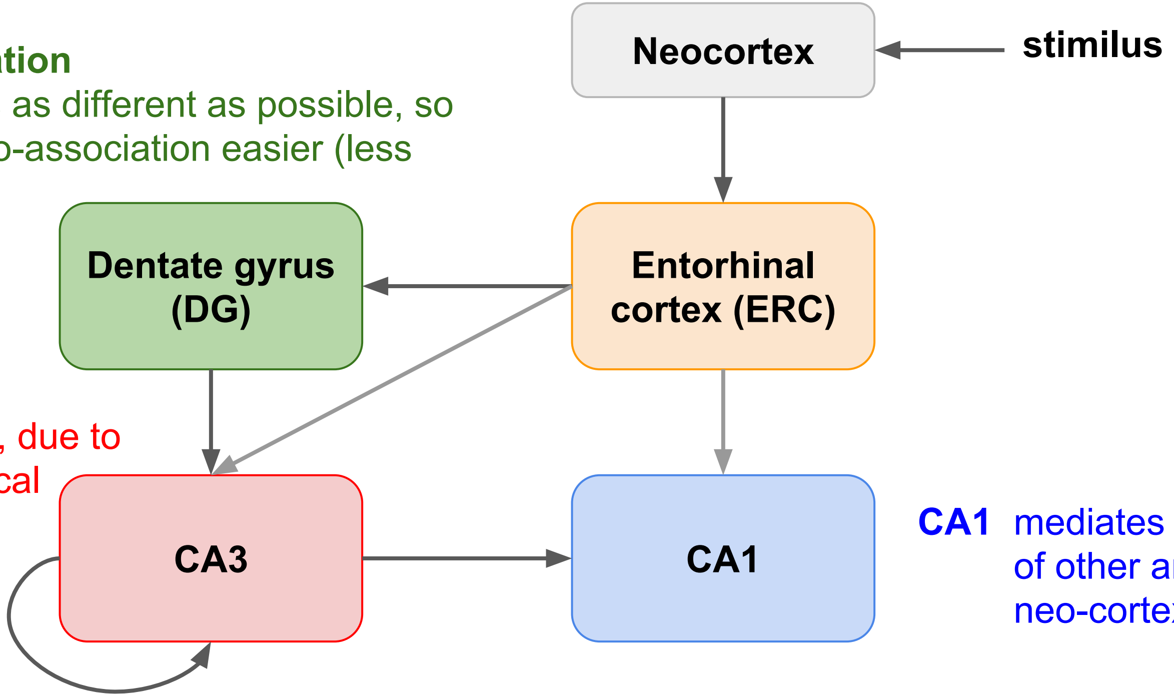
Episodic memory depends on hippocampus

It is helped by a capacity to **bind together** (auto-associate) **diverse inputs** from different brain areas that represent **parts** of an event

- Episodic memory = the collection of past personal experiences that occurred at a **particular time and place**.
- Remembering a 6th birthday party: EM allows an individual to figuratively travel back in time to remember the event that took place at that particular time and place

Pattern separation and completion

DG **pattern separation**
makes patterns as different as possible, so as to make auto-association easier (less confusing)



CA3 **auto-associator**, due to its dense reciprocal connections

→ **pattern completion**

CA1 mediates the activation of other areas of the neo-cortex

Pattern separation and completion

- Idea: parts of event -- **spatial** (place) and **non-spatial** (what happened) -- are **processed in parallel** before converging in the hippocampus in DG/CA3 subregions

- **Pattern separation** + **pattern completion** = central to hippocampus for storing details of experiences

Pattern separation and completion

Pattern separation

(Idea): Dentate gyrus subregion in HC performs pattern separation, orthogonalizes incoming inputs before…

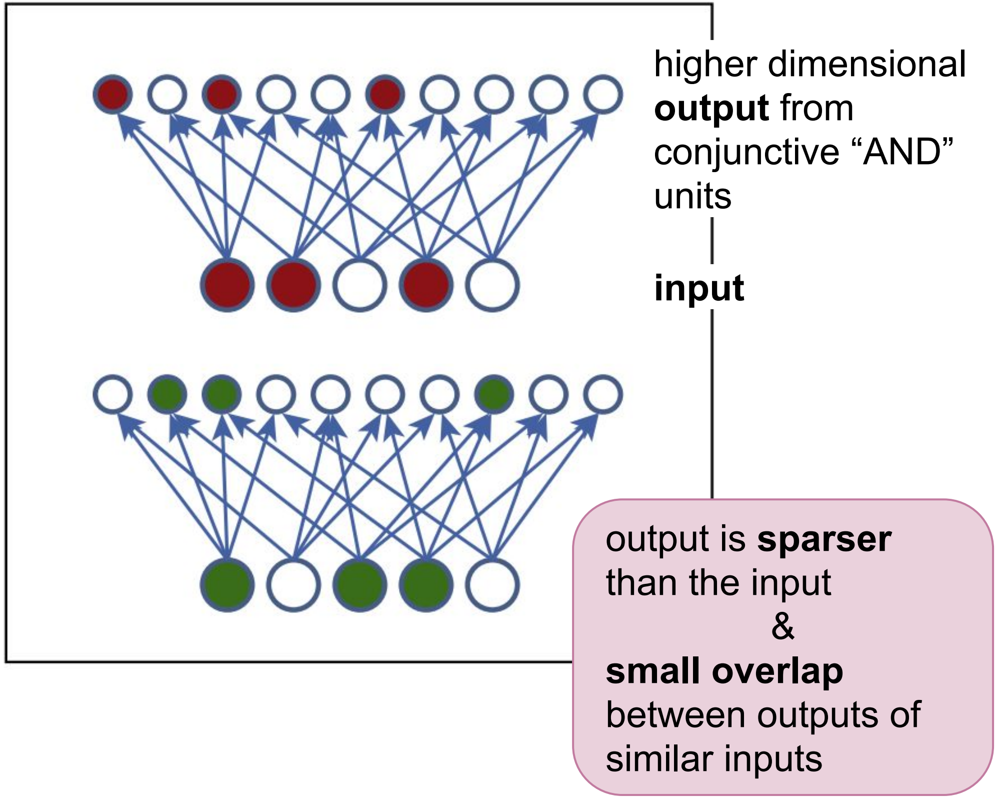
Pattern completion

(Idea): ...auto-associative storage in the CA3 region

Output of an **entire stored pattern** (e.g. corresponding to an entire episodic memory) from a **partial input**. Functions as an attractor network.

Pattern separation in dentate gyrus (?)

- Pattern separation → similar input patterns result in more distinct output patterns
- The result is a **conjunctive code**
- Thought to be implemented in **DG**

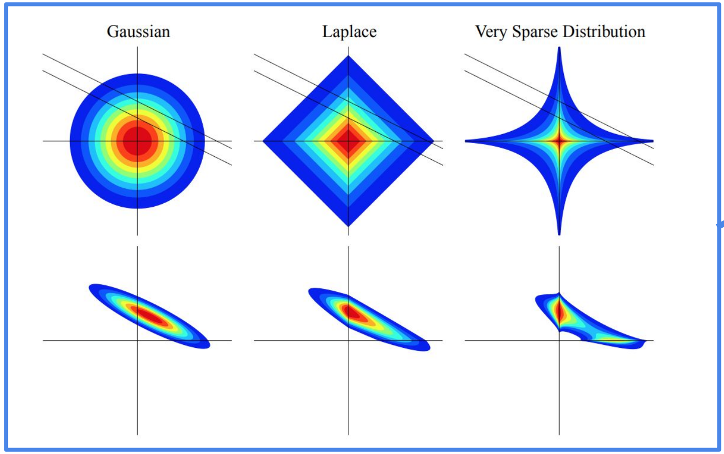


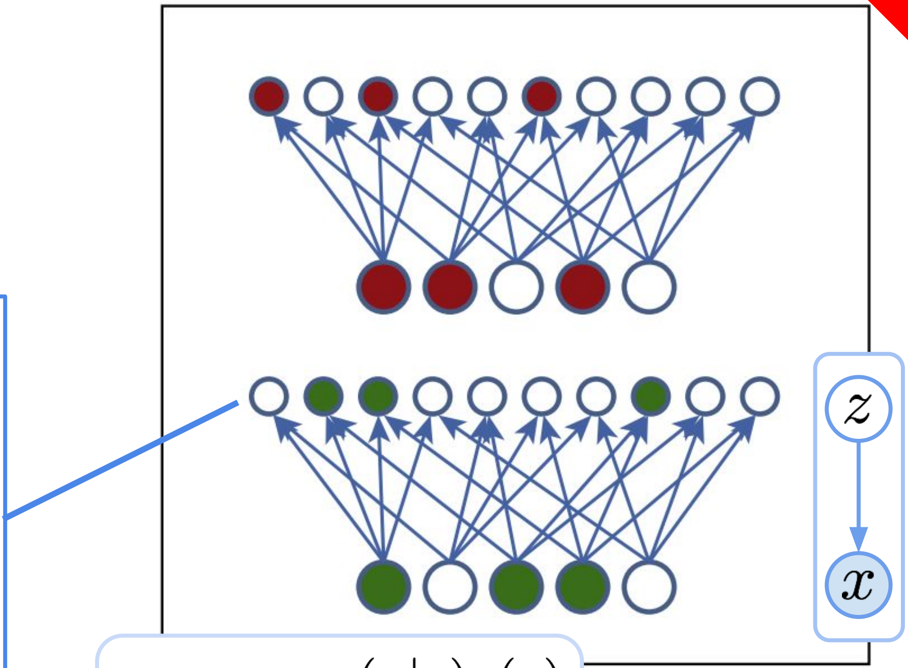
Sparsity priors, compressed sensing

In the machine learning world

- Sparse, overcomplete representation
- “Sparse linear models”
- **Inference and learning is difficult**

Sparsity prior on a much larger latent variable



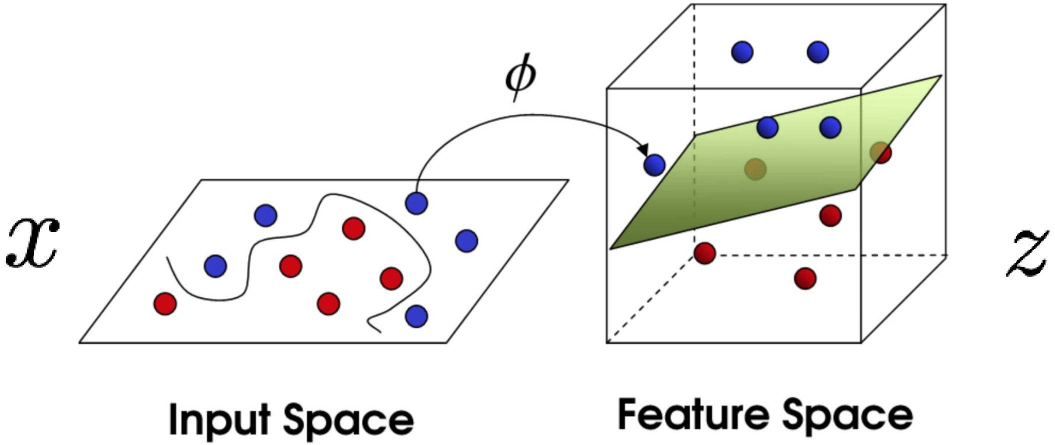


$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

RKHS

In the machine learning world

- Support vector machines
- **Kernel trick** and the reproducing kernel Hilbert space (RKHS)
- **Implicitly** maps inputs to a higher (**infinite**!) dimensional space
- No sparse representation, though...



Pattern completion

In the machine learning world

Associative memories

Can it be a sparse code, like DG → CA3?

42.3: Definition of the continuous Hopfield network

(a)

`moscow------russia`
`lima----------peru`
`london-----england`
`tokyo--------japan`
`edinburgh-scotland`
`ottawa------canada`
`oslo--------norway`
`stockholm---sweden`
`paris-------france`

A list of desired memories

z

(b) `moscow---::::::::: ⟹ moscow------russia`
`::::::::::--canada ⟹ ottawa------canada`

pattern completion, given a partial pattern

(c) `otowa-------canada ⟹ ottawa------canada`
`egindurrh-sxotland ⟹ edinburgh-scotland`

z

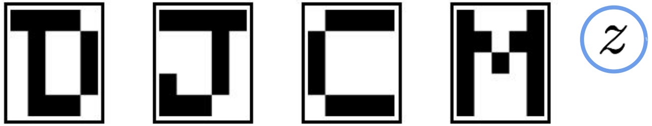
error correction

From David MacKay's *Information Theory, Inference, and Learning Algorithms*

Pattern completion

In the machine learning world

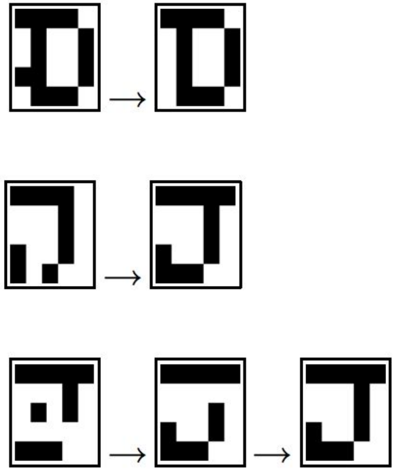
Associative memories



A list of desired memories

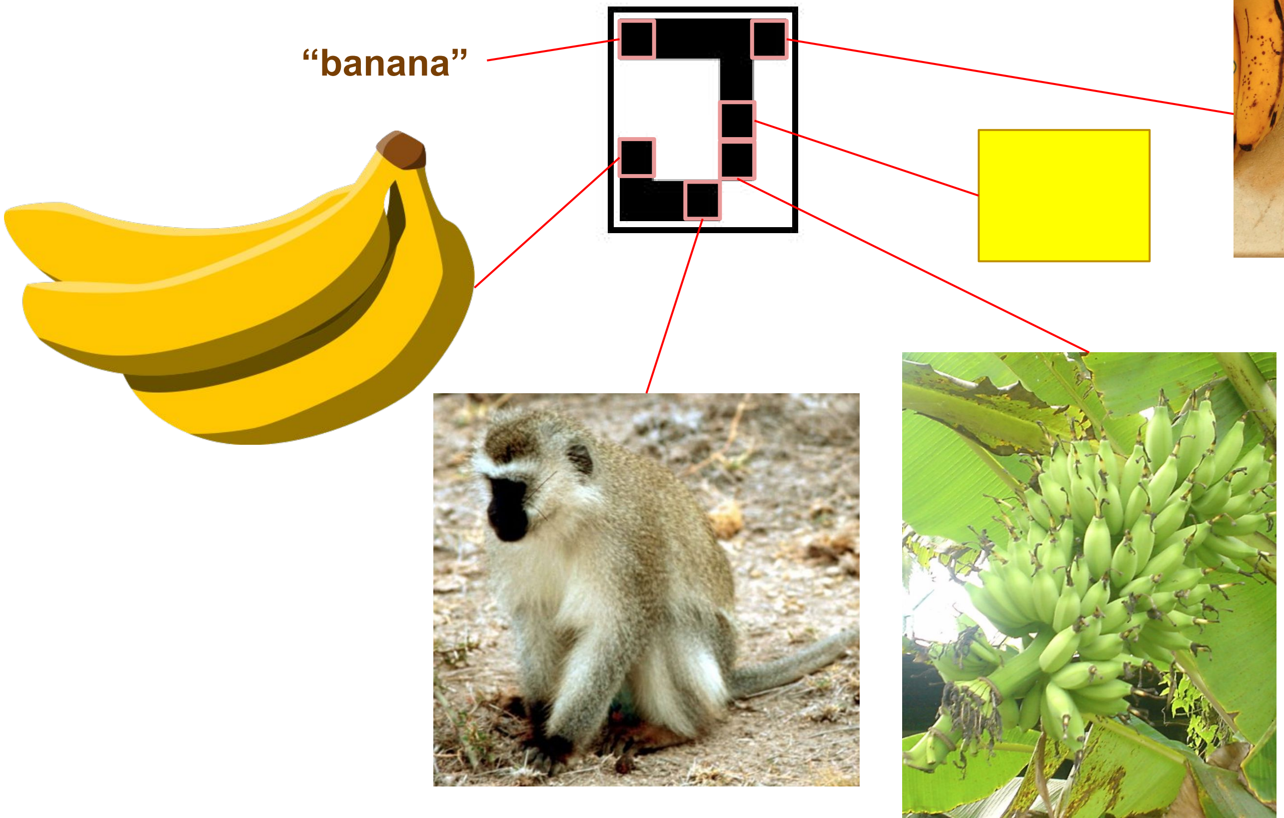
Can it be a sparse code, like DG → CA3?

initial states restored to a desired memory via an attractor network



From David MacKay's *Information Theory, Inference, and Learning Algorithms*

Auto-associative at a concept-level? Hierarchical?



Memories in topological fashion like graphs?

Images: Wikimedia

New patterns

Similar input pattern via entorhinal cortex **to previous pattern** (memory retrieval)

- CA3 outputs a pattern closer to the one it previously used for this ERC pattern

Low overlap to previously stored patterns (memory formation)

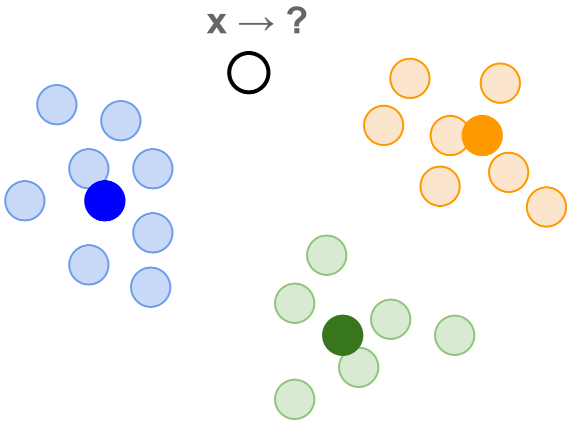
- **DG creates a new, statistically independent cell population** / neurogenesis
- Pattern separation!
    - Non-parametric Bayes

(Amount of overlap required for pattern completion may differ across the hippocampus)

New patterns

In the machine learning world

Non-parametric methods



Instance-based learning, where each experience has its **own coordinates**, capacity can be **increased** as required and parameters can grow with data.

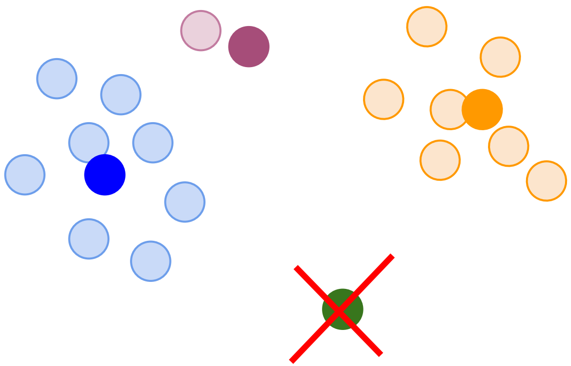
Dirichlet process mixture models

- Models data density through templates (cluster centers)
- Can **dynamically** “on the fly” **allocate more capacity** for a surprising new input input **x**
- Uses existing capacity for familiar inputs

New patterns

In the machine learning world

Non-parametric methods



Instance-based learning, where each experience has its **own coordinates**, capacity can be **increased** as required and parameters can grow with data.

Dirichlet process mixture models

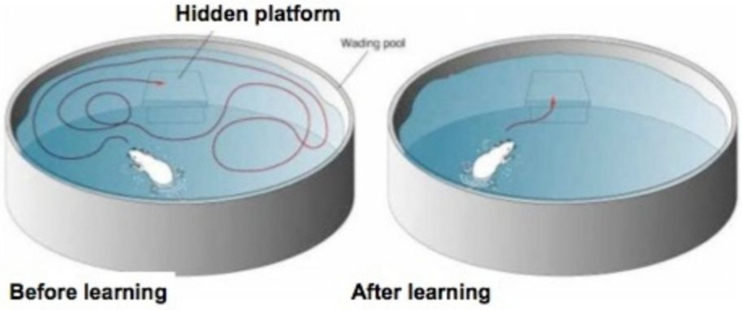
- Forgetting: **Capacity can be removed** if no input x is associated with it (say over a time window)

Evidence supporting CLS theory #4

The hippocampus and neocortex support quantitatively different forms of representation

Rat behaviour in Morris water maze

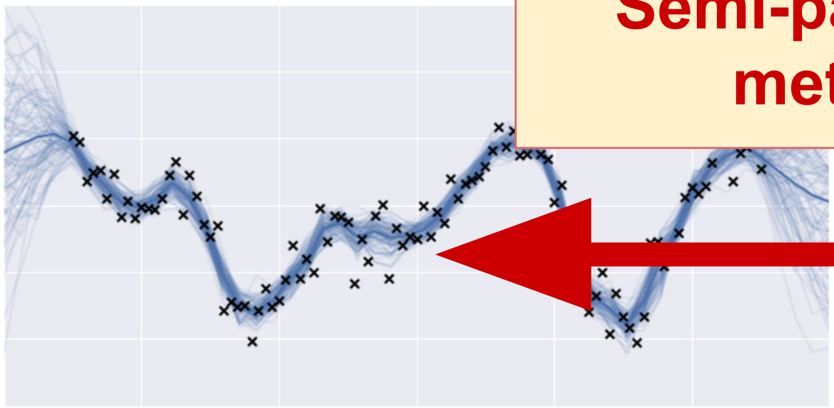
- Early on appeared to reflect **individual episodic traces** (i.e. an instance-based non-parametric representation)
- Was later (28 days after learning) consistent with the use of a **parametric representation putatively housed in the neocortex**

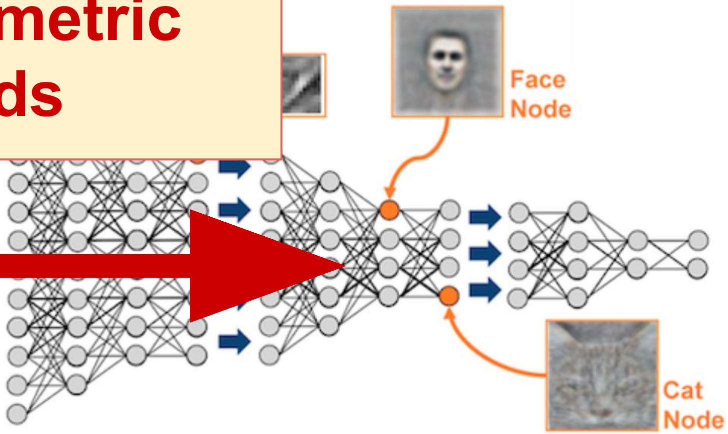


Trends, trends, trends

Non-parametric methods | **Parametric methods**

Semi-parametric methods





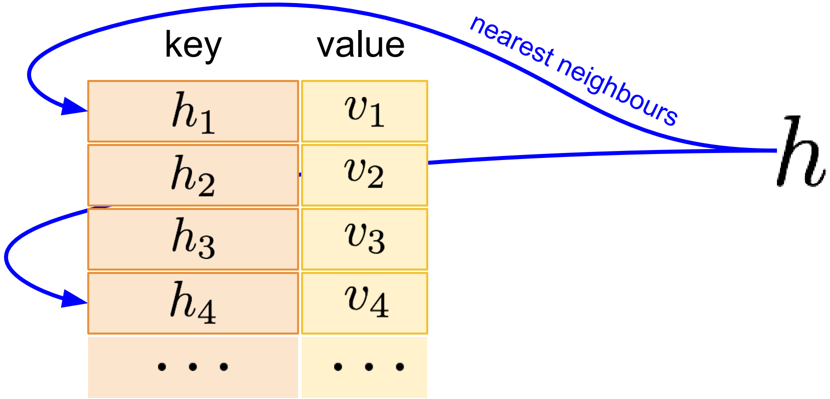
Instance-based learning, where each experience has its **own coordinates**, capacity can be **increased** as required and parameters can grow with data.

Optimal parametric characterization of the **statistics of the environment** by learning gradually through **repeated**, interleaved exposure to large numbers of training examples.

Neural Episodic Control

K-nearest neighbours on keys

Parametric methods

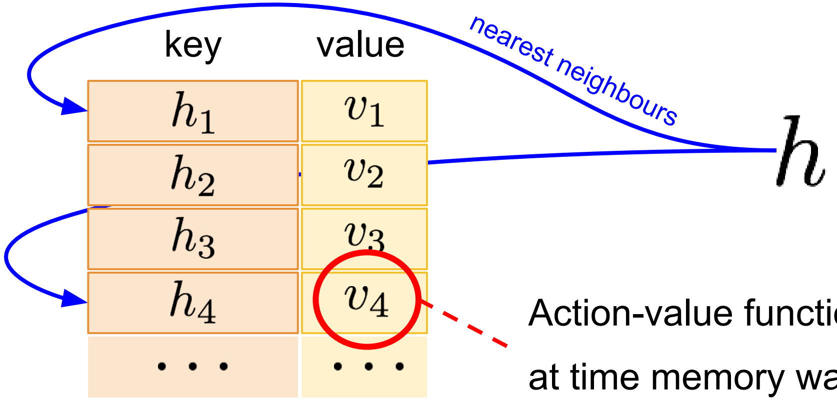


Memory module
for each action a

Neural Episodic Control

K-nearest neighbours on keys

Parametric methods



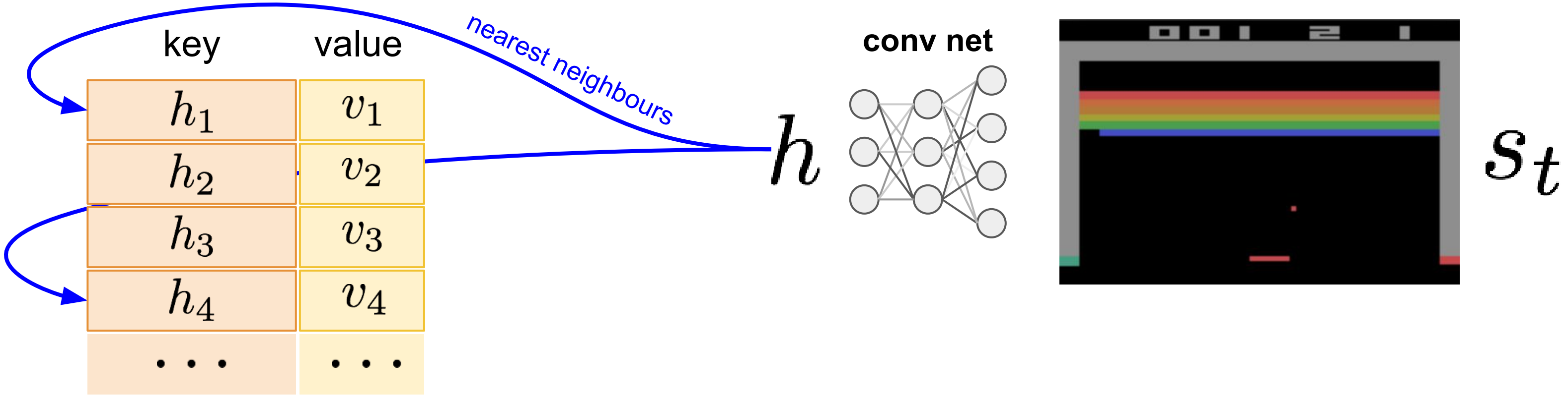
Memory module
for each action a

Action-value function estimate $Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_t \gamma^t r_t \,\middle|\, s, a\right]$
at time memory was written

Neural Episodic Control

K-nearest neighbours on keys

Parametric methods

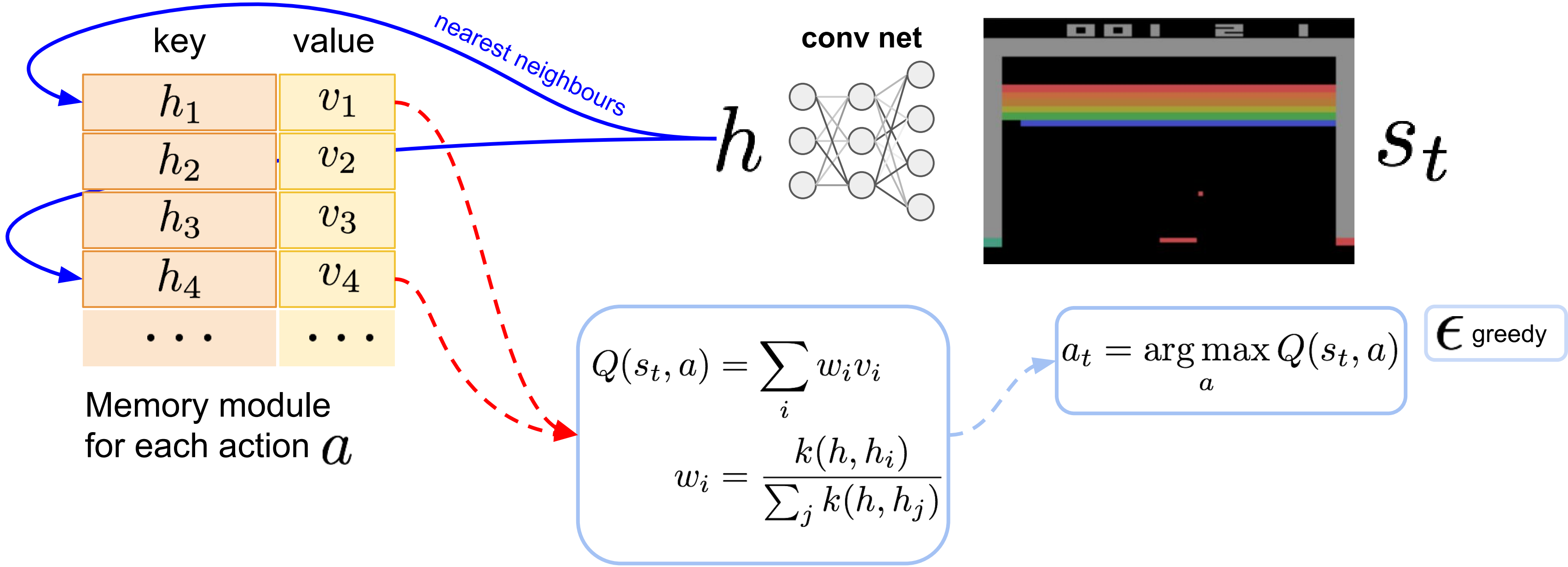


Memory module
for each action a

Neural Episodic Control

K-nearest neighbours on keys

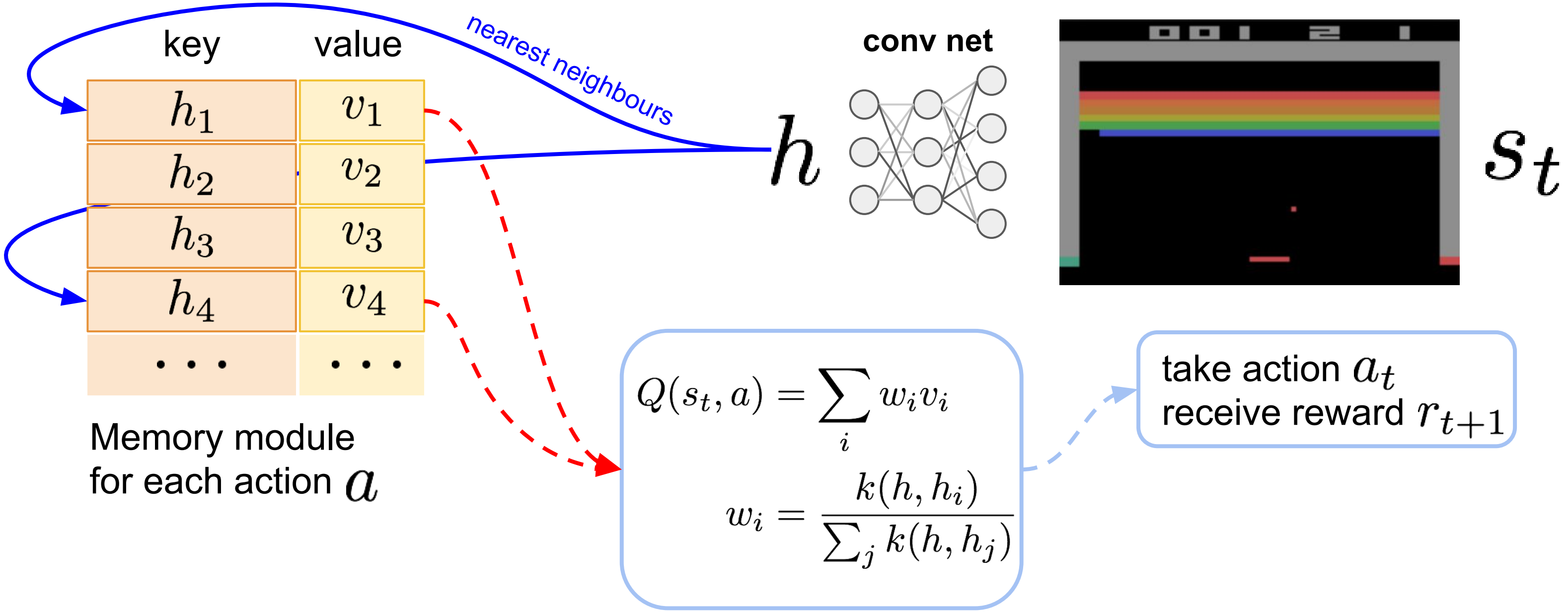
Parametric methods



Neural Episodic Control

K-nearest neighbours on keys

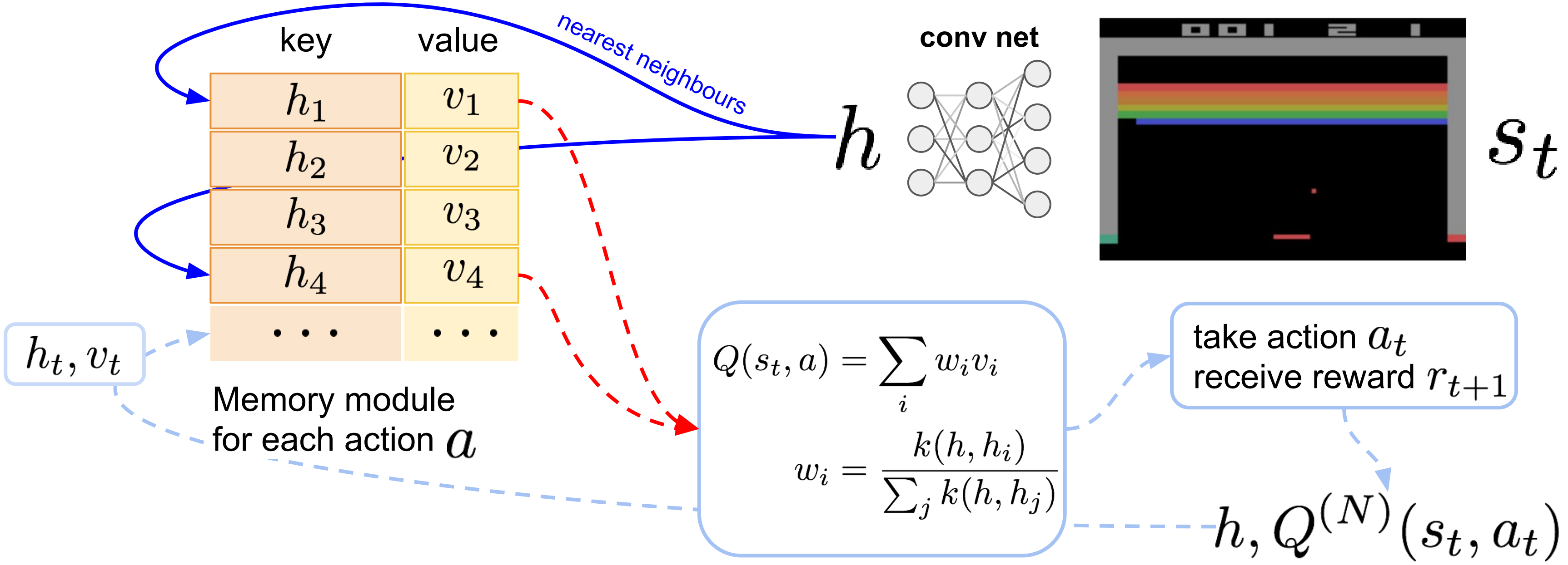
Parametric methods



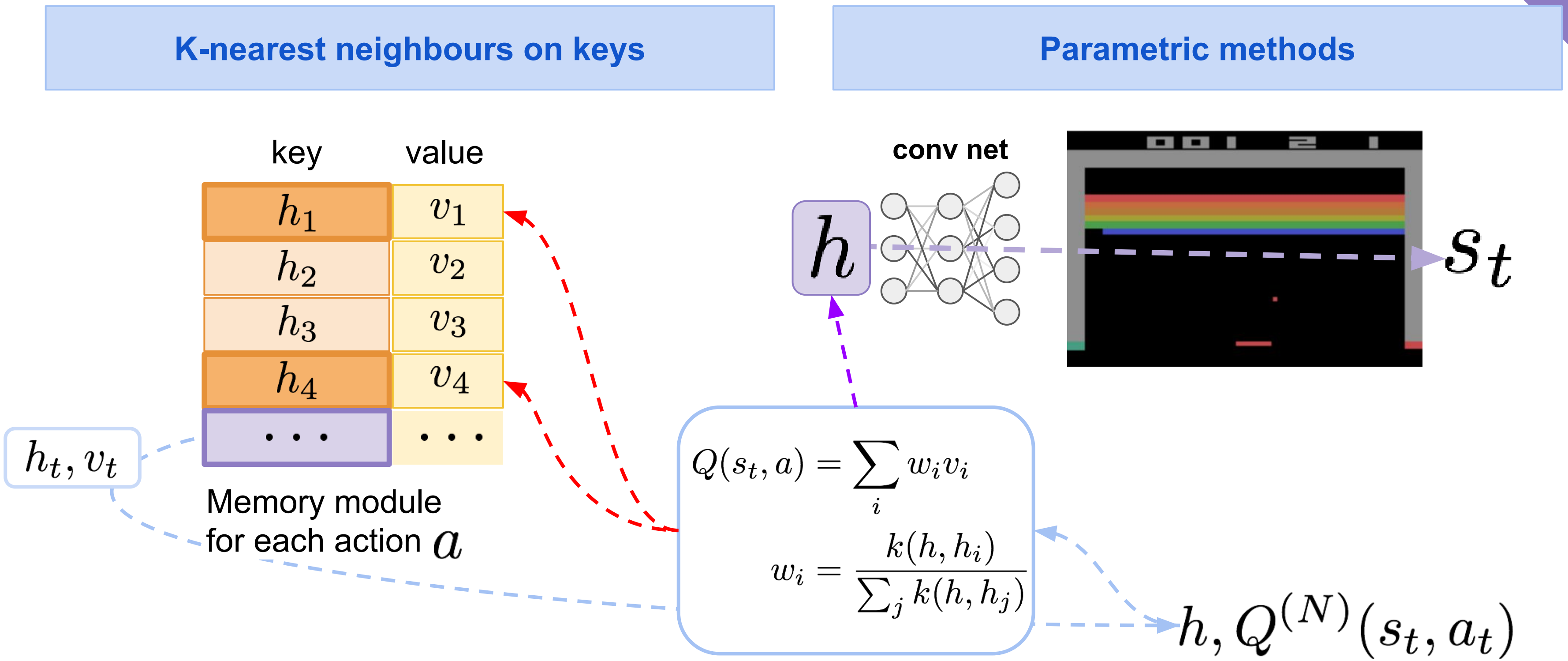
Neural Episodic Control

K-nearest neighbours on keys

Parametric methods



Neural Episodic Control

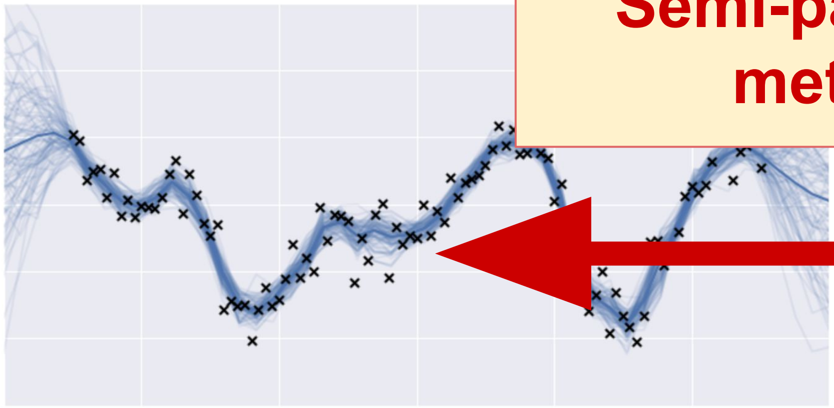


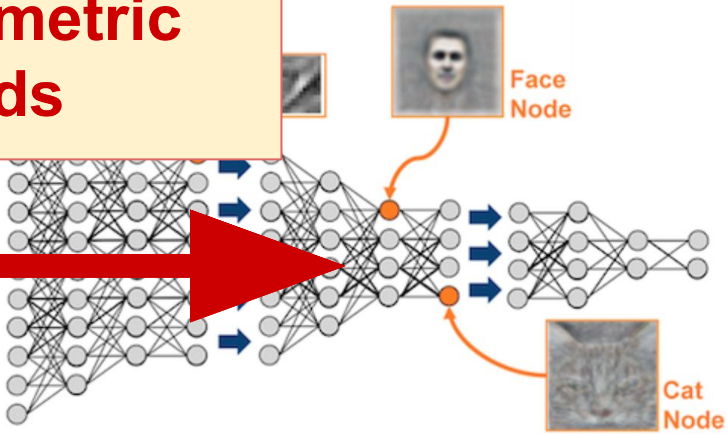
Conclusion

Non-parametric methods

Parametric methods

Semi-parametric methods



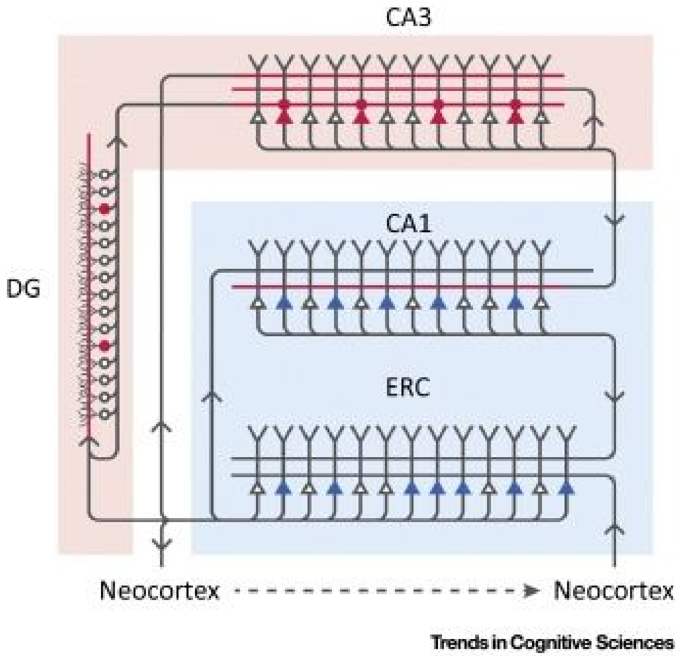


be creative :)

DeepMind

www.deepmind.com

Hippocampal Subregions, Connectivity, and Representation



Entorhinal cortex (ERC): grid cells

Dentate gyrus(DG): pattern separation, very sparse, adult neurogenesis

CA1: place cells

CA3: pattern completion, highly recurrent

Neural Episodic Control

Action-value function $Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_t \gamma^t r_t \,\middle|\, s, a \right]$

N-step Q-value estimate $Q^{(N)}(s_t, a) = \sum_{j=0}^{N-1} \gamma^j r_{t+j} + \gamma^N \max_{a'} Q(s_{t+N}, a')$

Adds N **on-policy rewards** and bootstraps the sum of discounted rewards for the rest of the trajectory, **off-policy**.

Train on random minibatch from replay memory