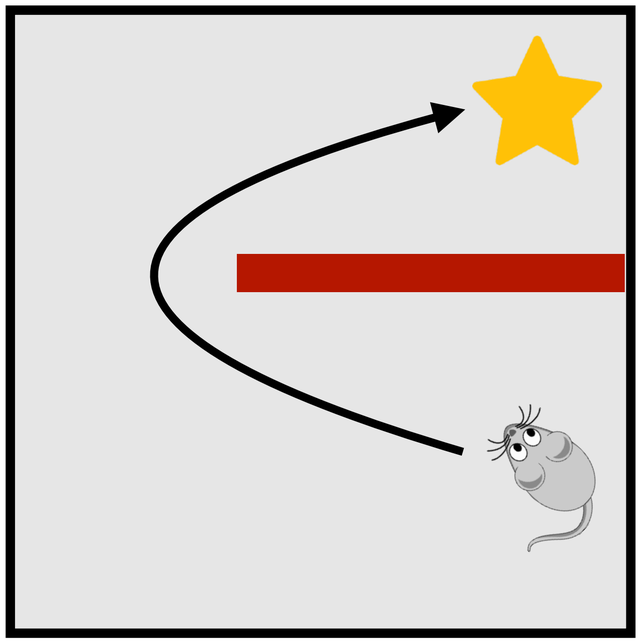
How to build cognitive maps

James Whittington

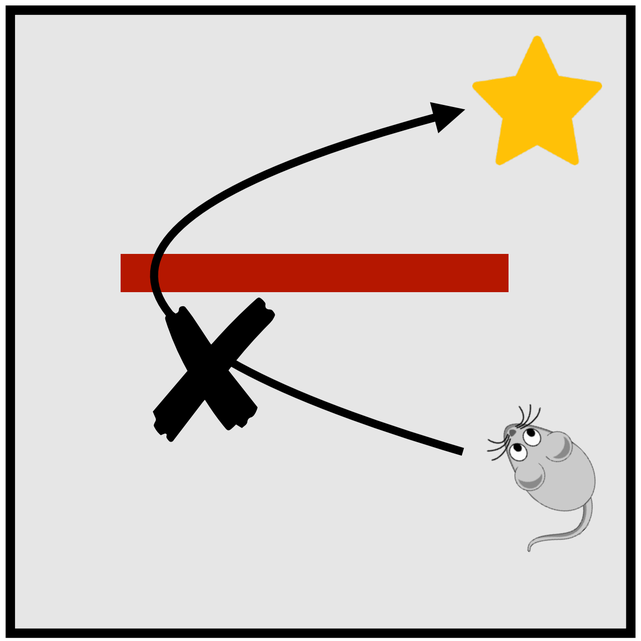
What is a cognitive map? Representating knowledge for flexible behaviour

What is a cognitive map? Representating knowledge for flexible behaviour



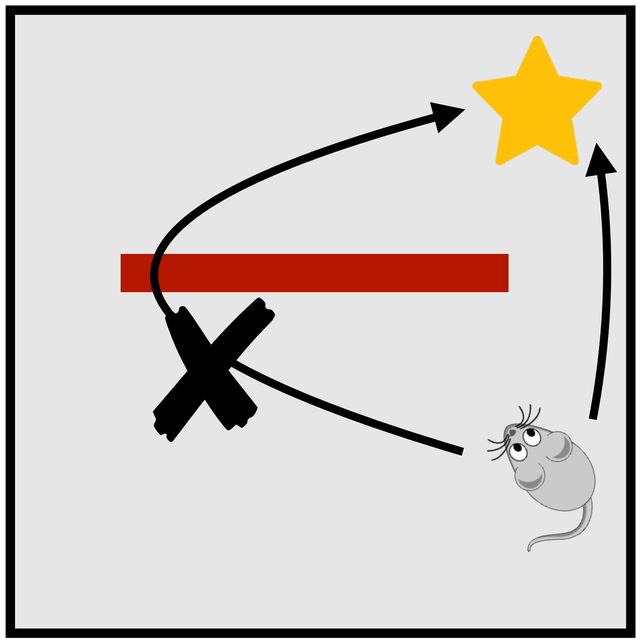
What is a cognitive map? Representating knowledge for flexible behaviour

Repeating old actions isn’t optimal



What is a cognitive map? Representating knowledge for flexible behaviour

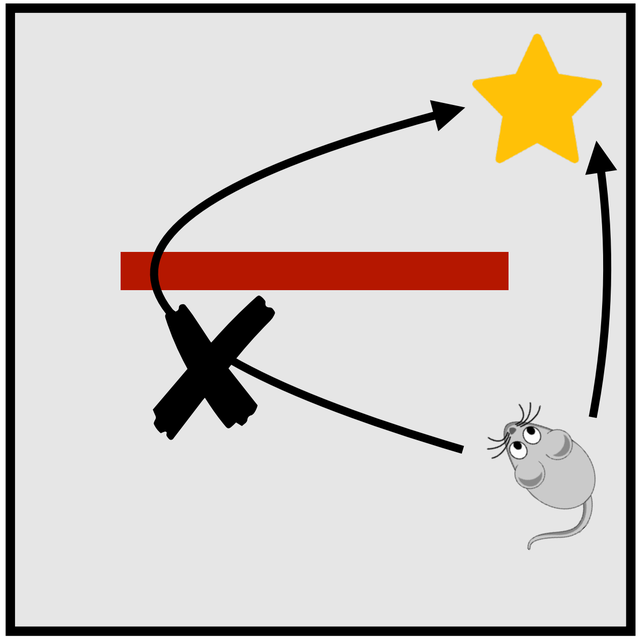
Repeating old actions isn’t optimal



Can take shortcuts with a world model

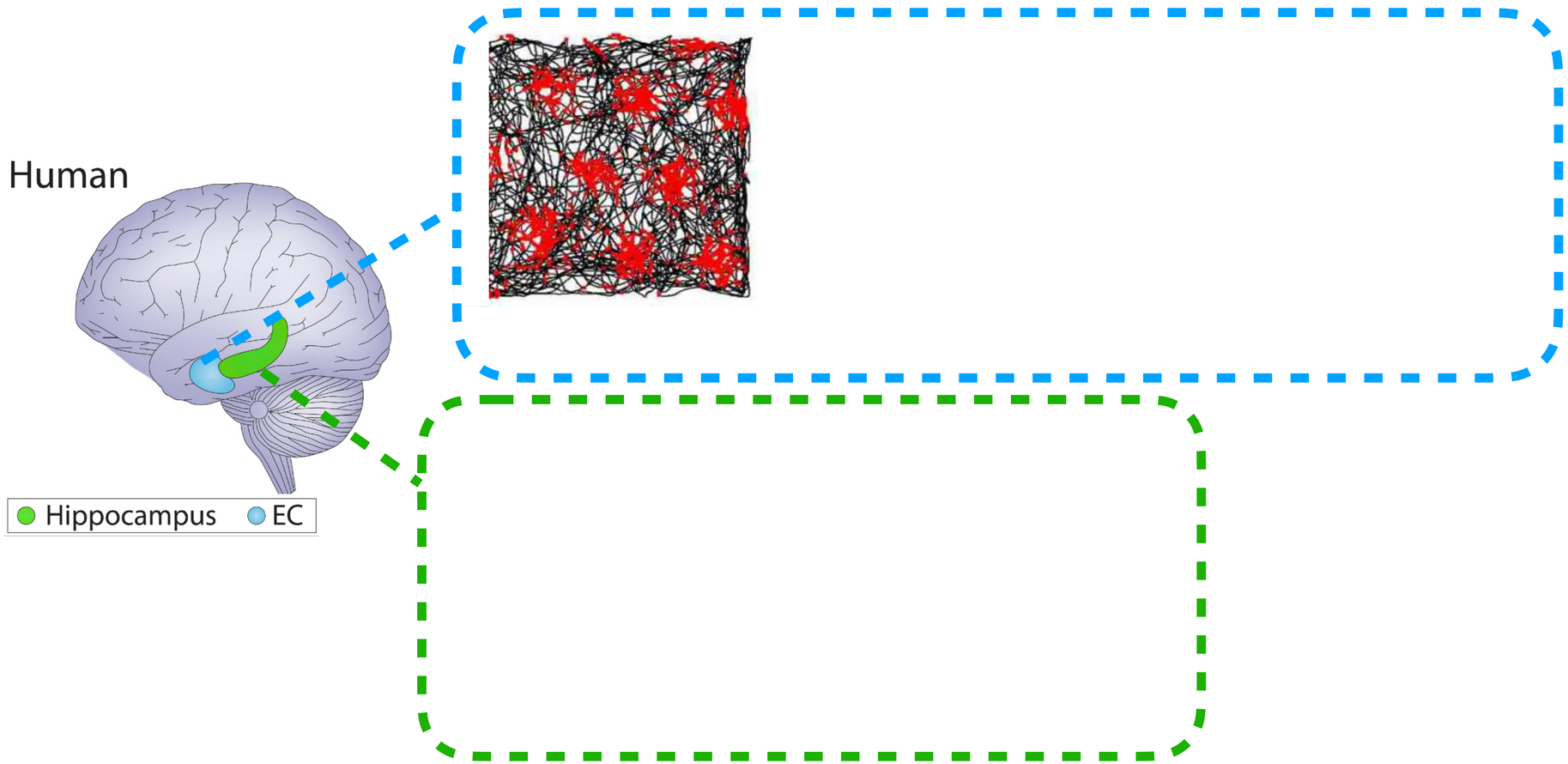
What is a cognitive map? Representating knowledge for flexible behaviour

Repeating old actions isn’t optimal

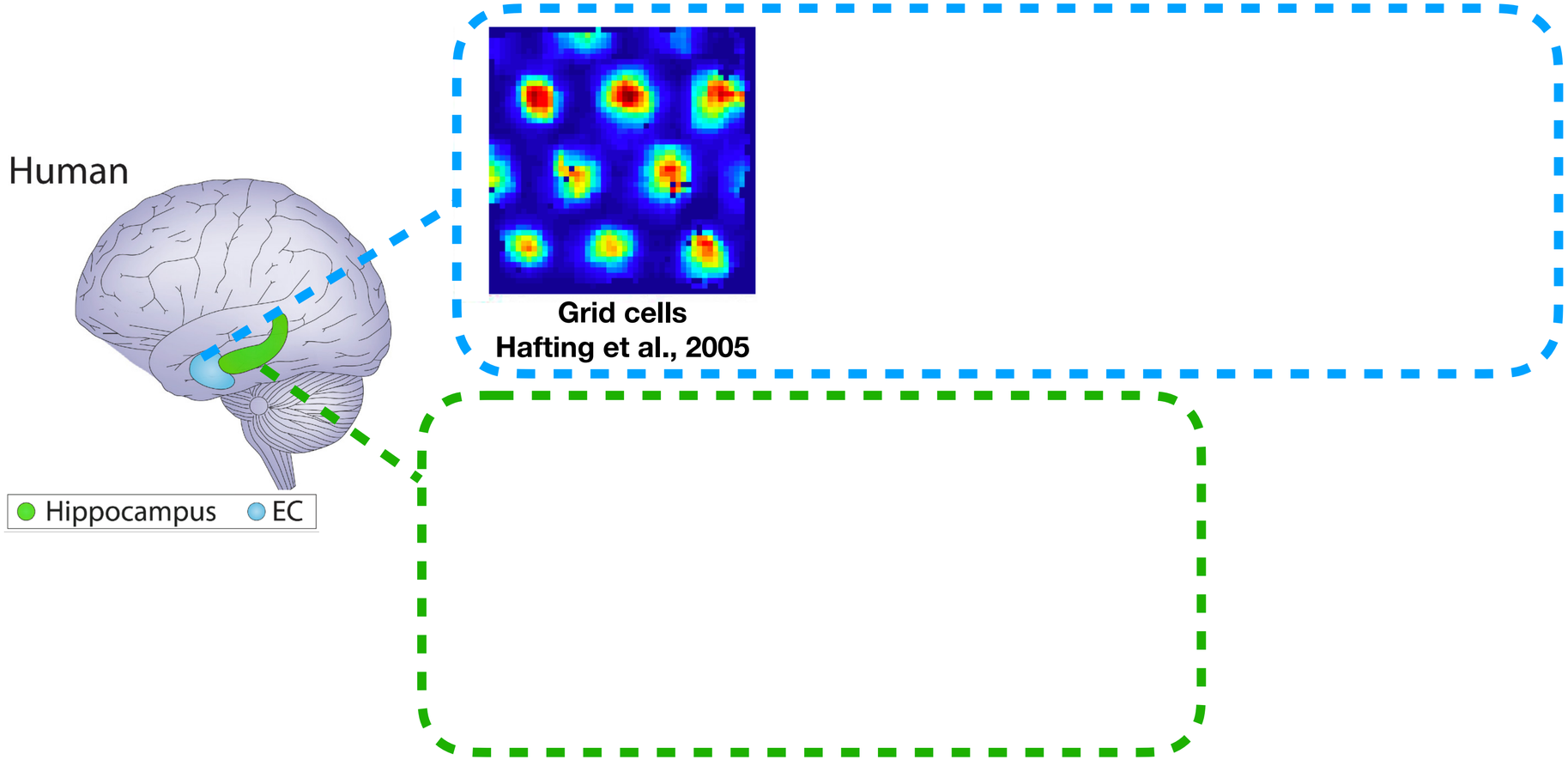


Can take shortcuts with a world model

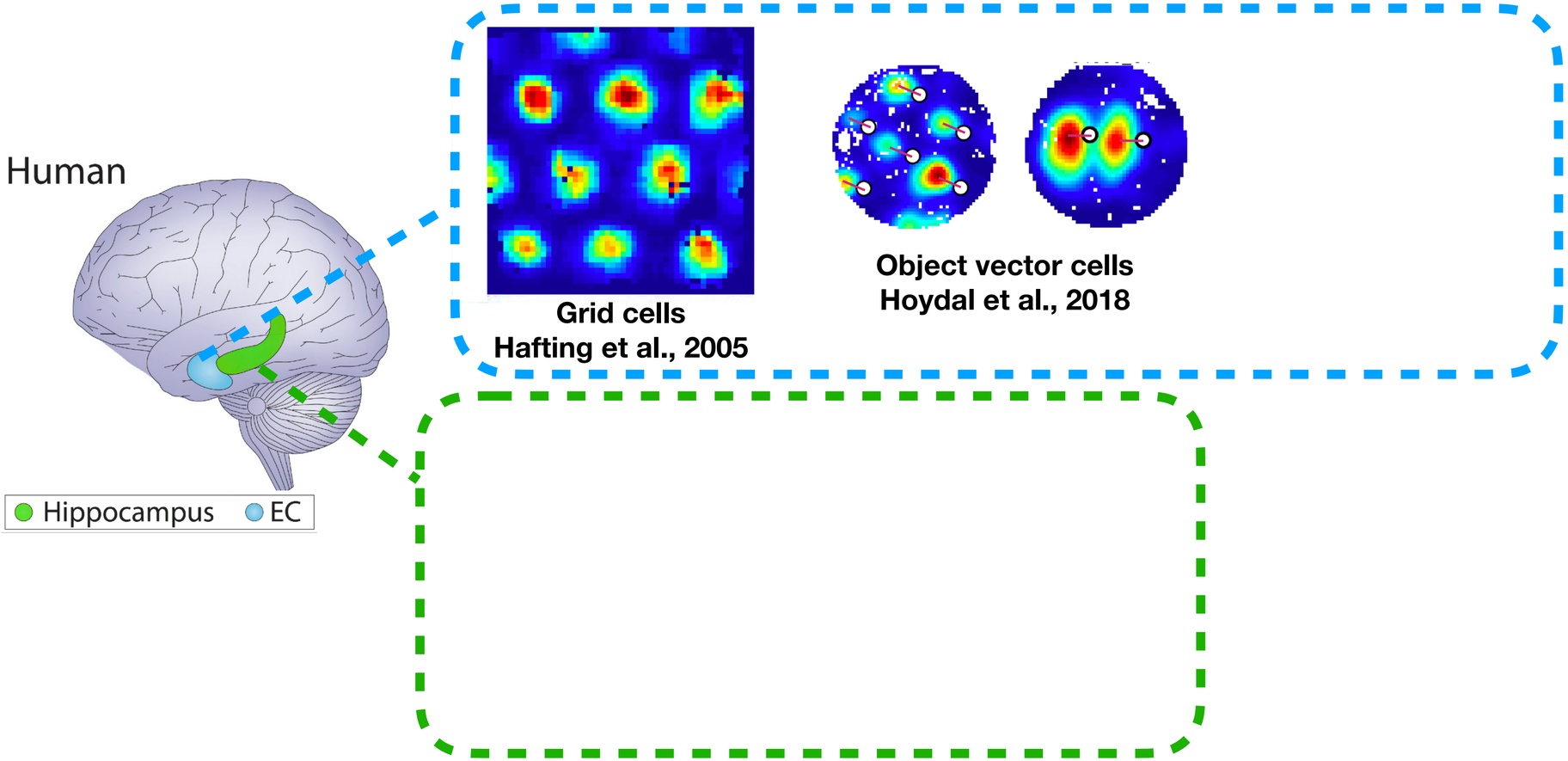
The hippocampal and entorhinal cognitive map



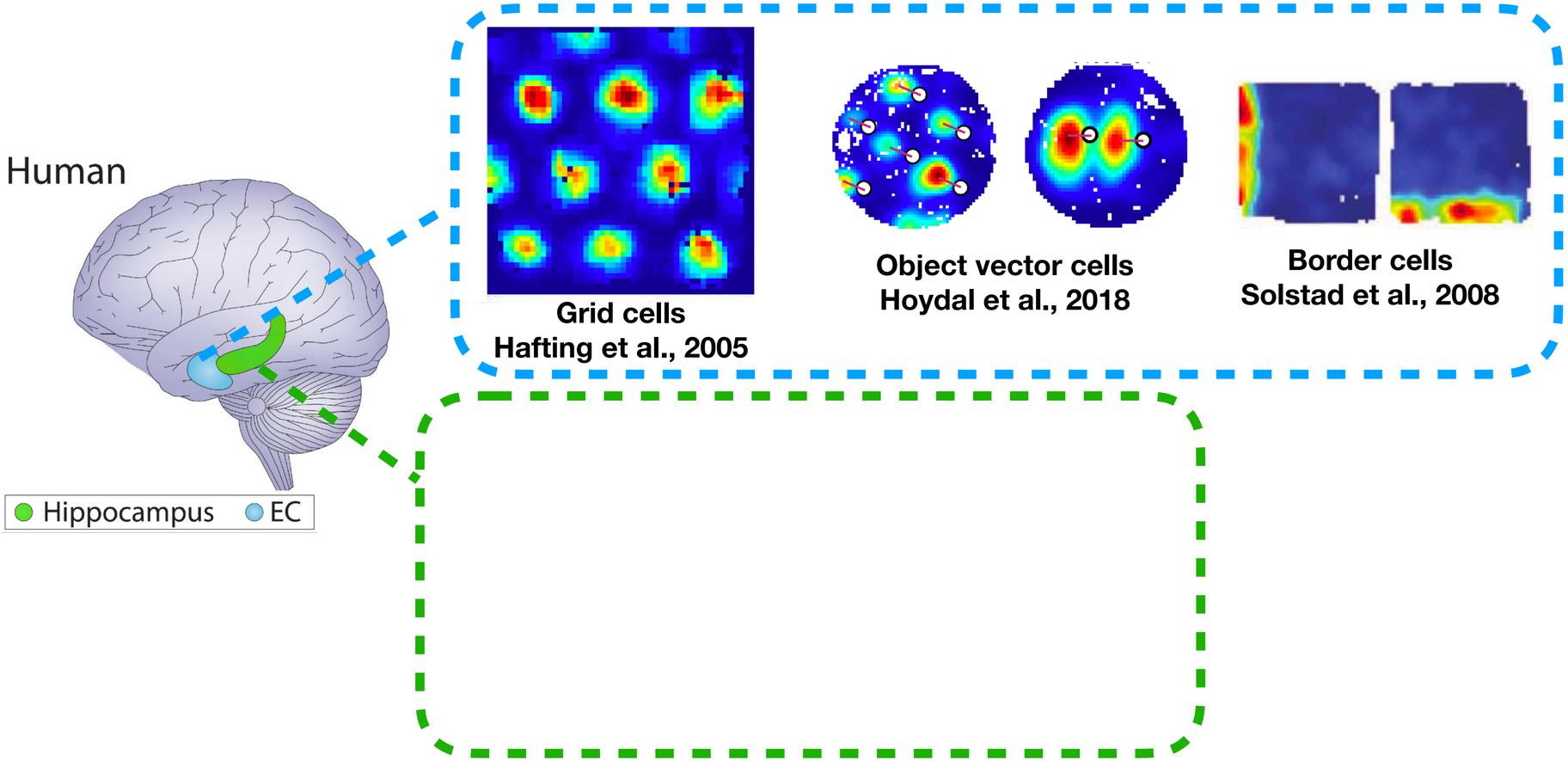
The hippocampal and entorhinal cognitive map



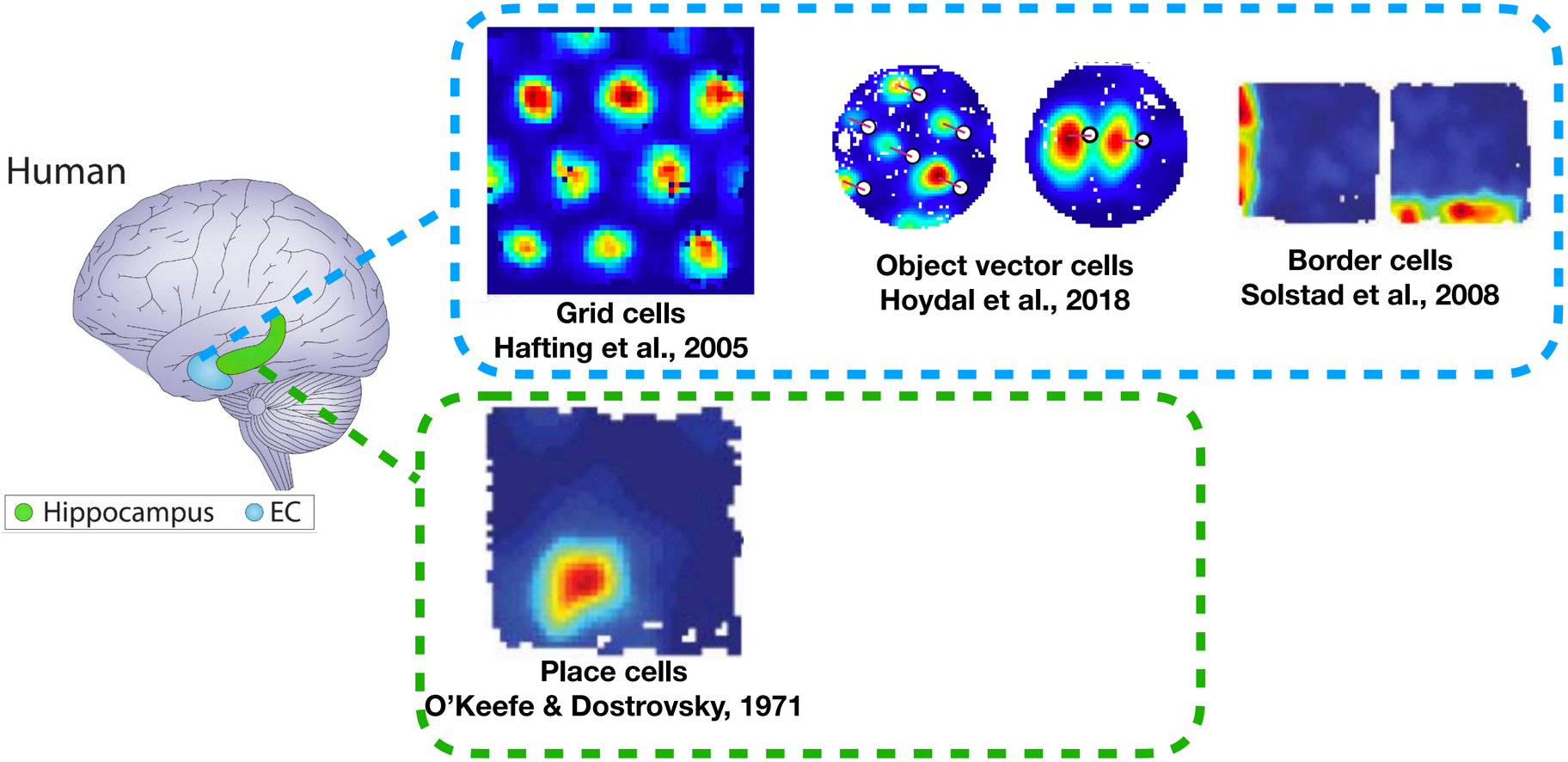
The hippocampal and entorhinal cognitive map



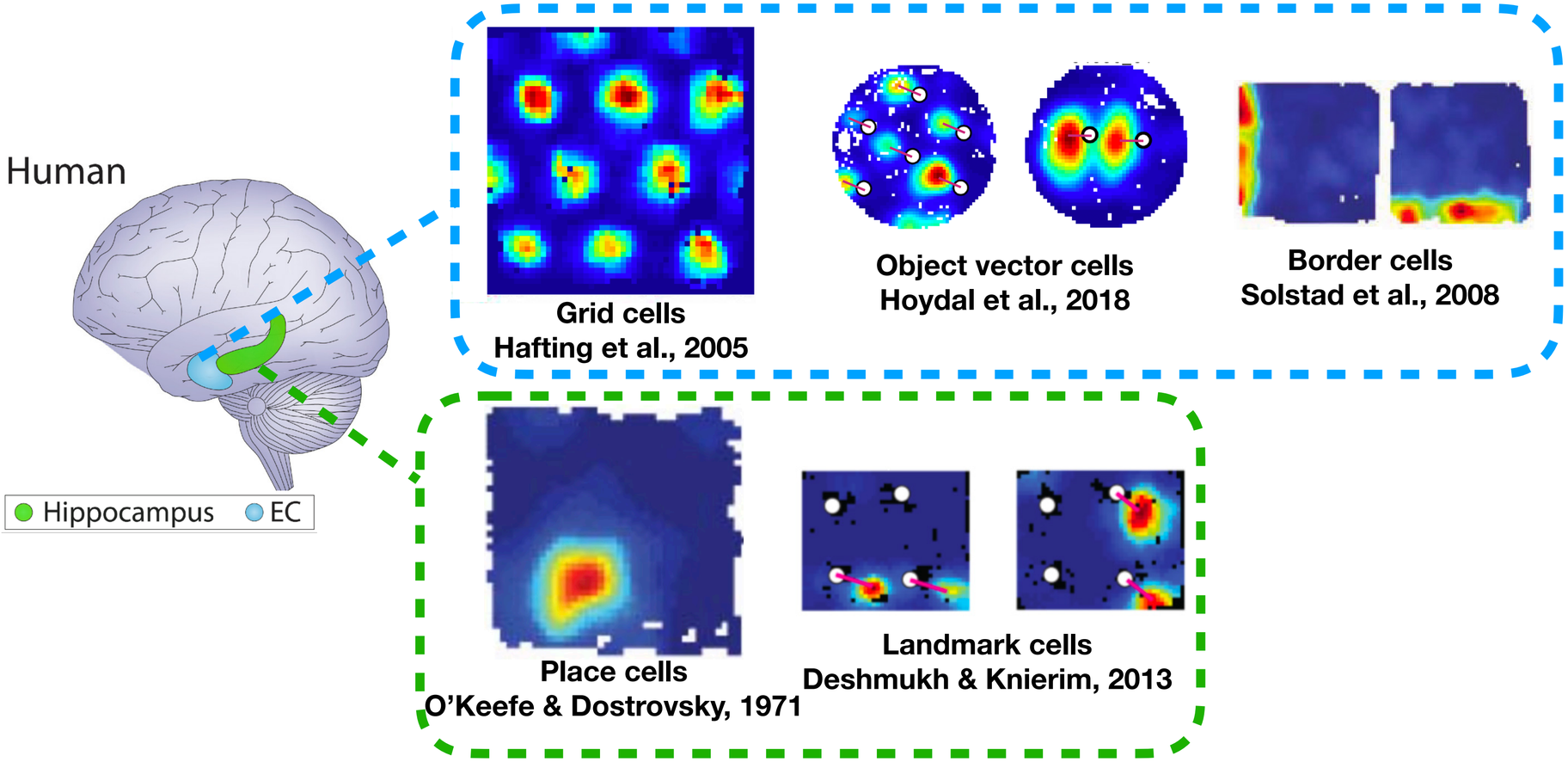
The hippocampal and entorhinal cognitive map



The hippocampal and entorhinal cognitive map

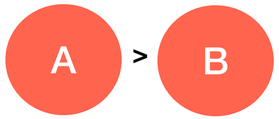


The hippocampal and entorhinal cognitive map



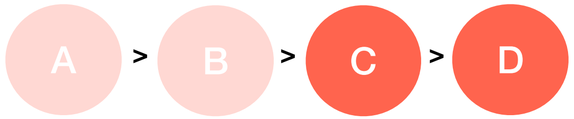
These brain areas are critical for non-spatial ‘shortcuts’

These brain areas are critical for non-spatial ‘shortcuts’



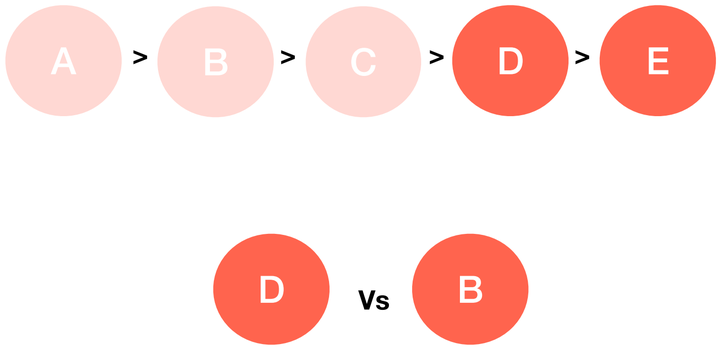
These brain areas are critical for non-spatial ‘shortcuts’

These brain areas are critical for non-spatial ‘shortcuts’

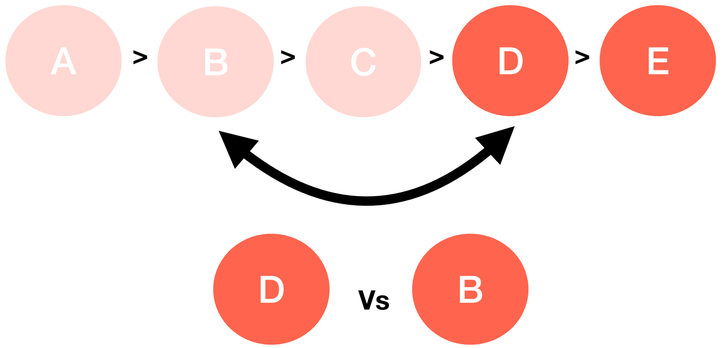


These brain areas are critical for non-spatial ‘shortcuts’

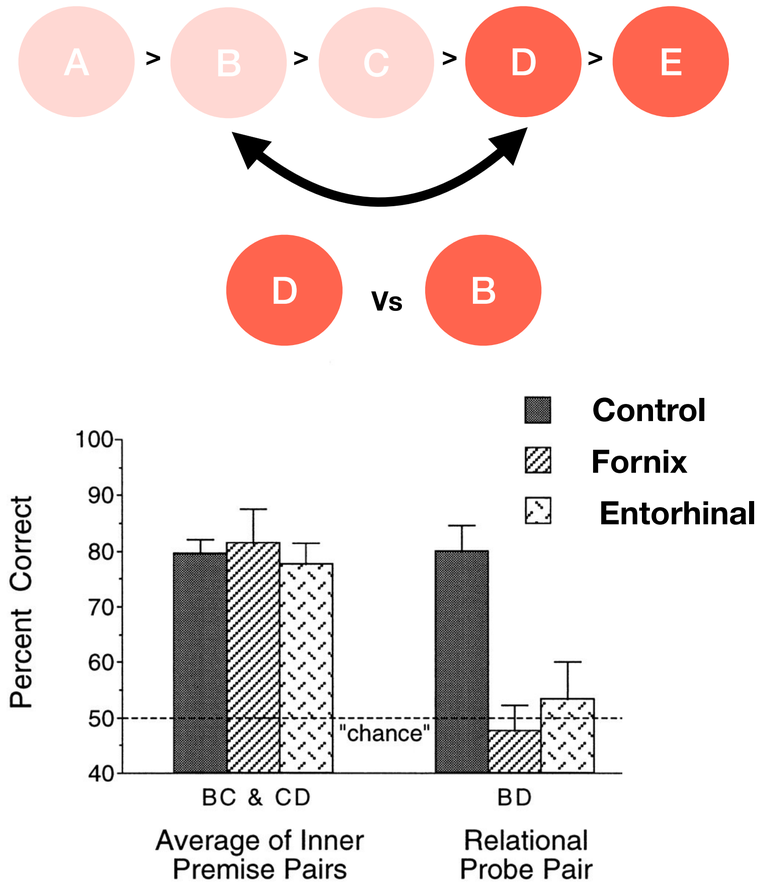
These brain areas are critical for non-spatial ‘shortcuts’



These brain areas are critical for non-spatial ‘shortcuts’

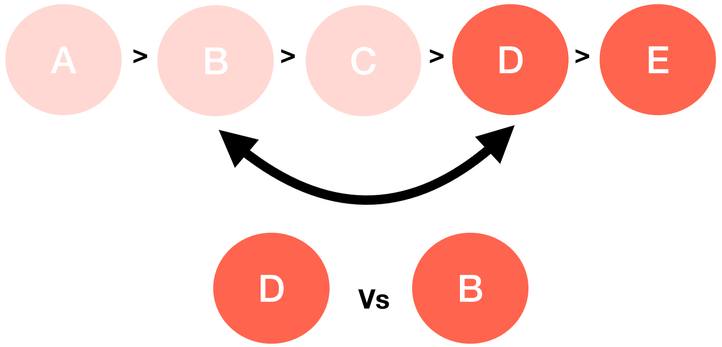


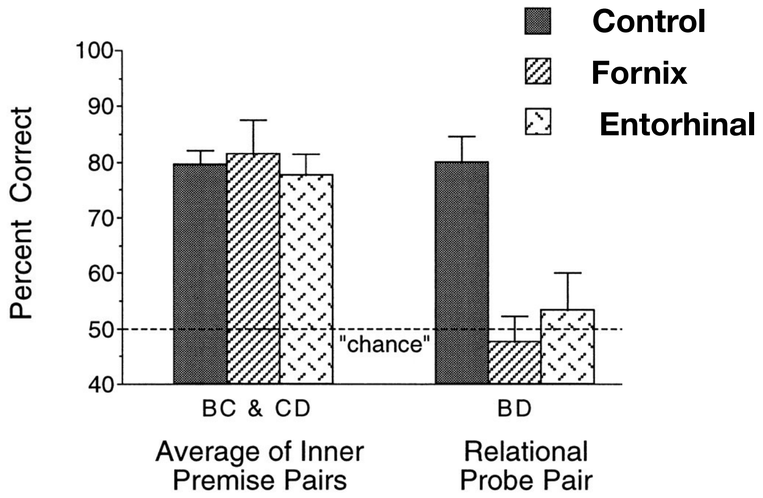
These brain areas are critical for non-spatial ‘shortcuts’



Dusek et al. PNAS 1997

These brain areas are critical for non-spatial ‘shortcuts’

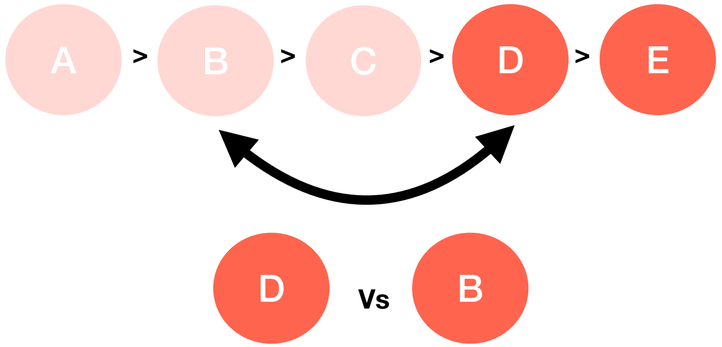


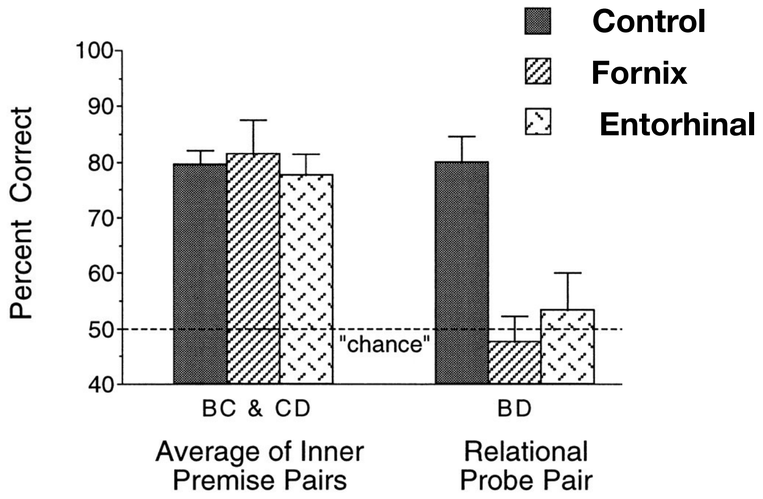


Dusek et al. PNAS 1997

“Spatial” cells seem to do similar things in non-spatial problems.

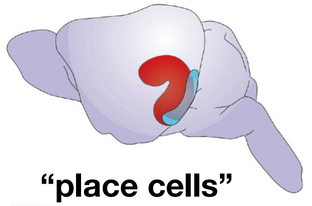
These brain areas are critical for non-spatial ‘shortcuts’



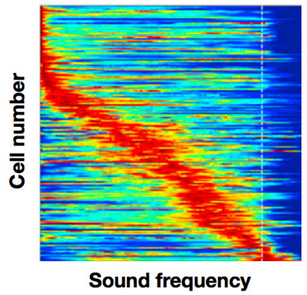


Dusek et al. PNAS 1997

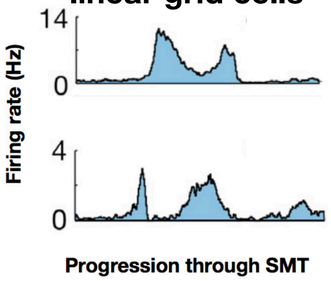
“Spatial” cells seem to do similar things in non-spatial problems.



“place cells”

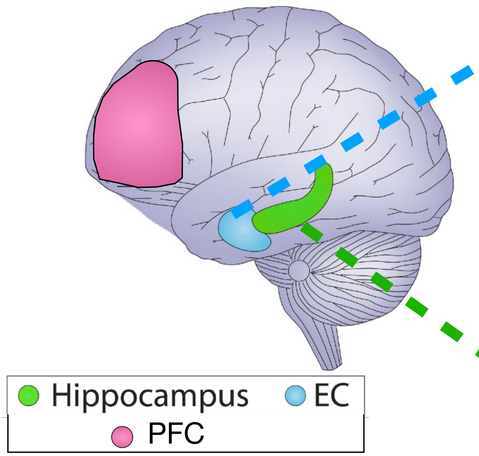


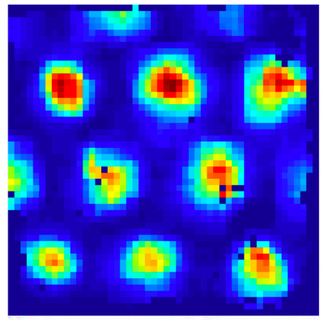
“linear grid cells”



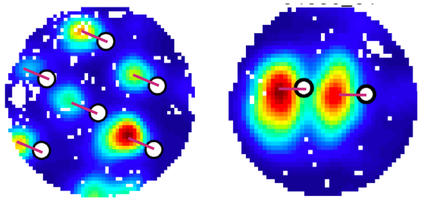
Aronov et al 2017

But prefrontal cortex solves this problem in a different way

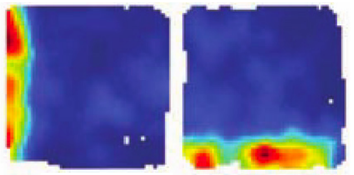




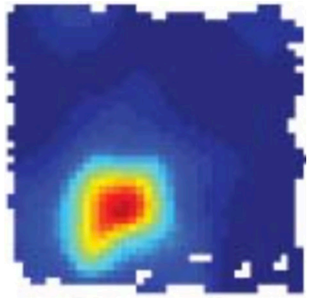
Grid cells
Hafting et al., 2005



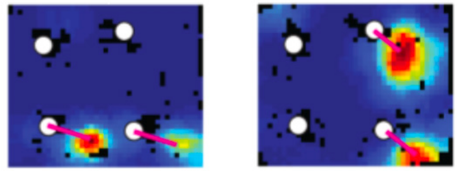
Object vector cells
Hoydal et al., 2018



Border cells
Solstad et al., 2008

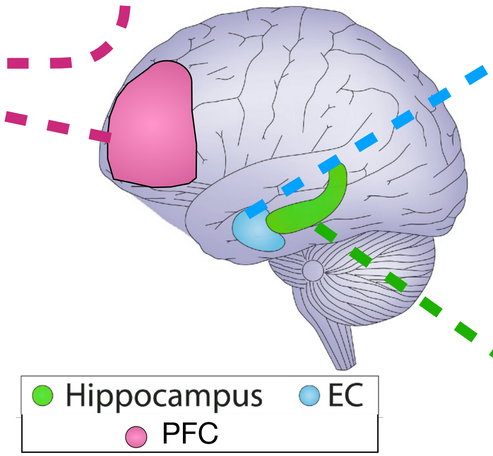


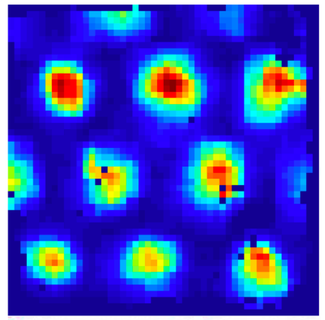
Place cells
O'Keefe & Dostrovsky, 1971



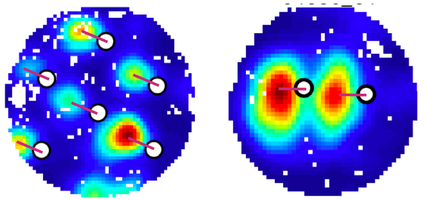
Landmark cells
Deshmukh & Knierim, 2013

But prefrontal cortex solves this problem in a different way

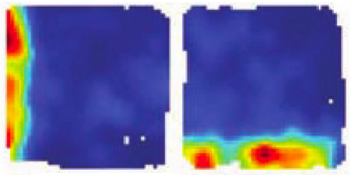




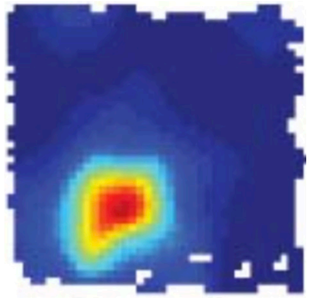
Grid cells
Hafting et al., 2005



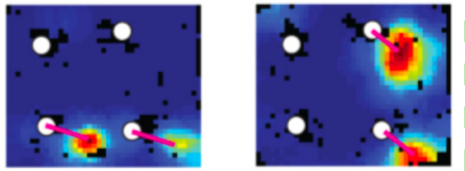
Object vector cells
Hoydal et al., 2018



Border cells
Solstad et al., 2008

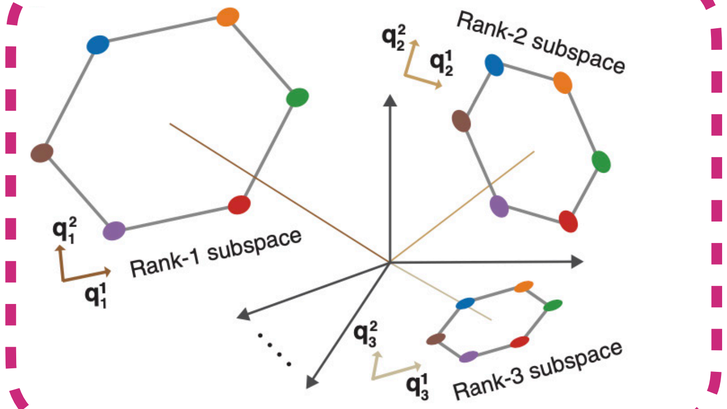


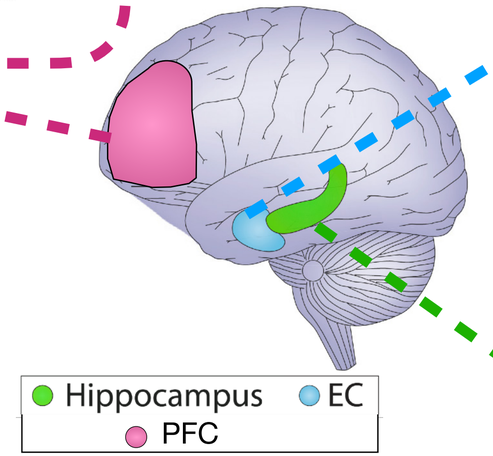
Place cells
O'Keefe & Dostrovsky, 1971

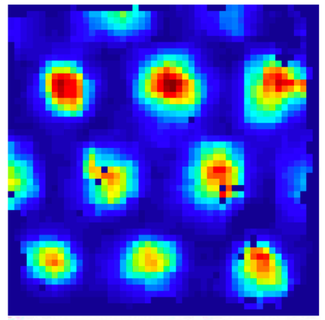


Landmark cells
Deshmukh & Knierim, 2013

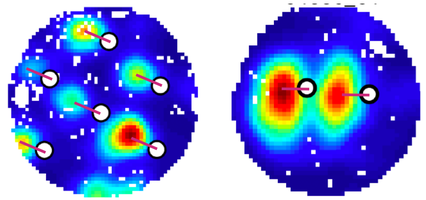
But prefrontal cortex solves this problem in a different way



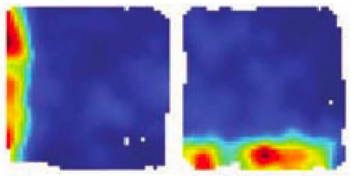




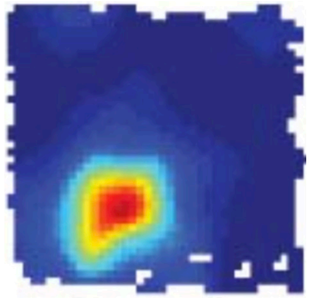
Grid cells
Hafting et al., 2005



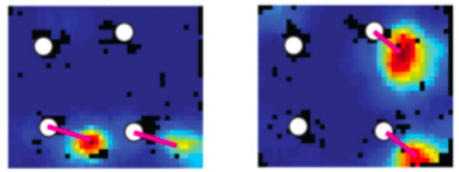
Object vector cells
Hoydal et al., 2018



Border cells
Solstad et al., 2008

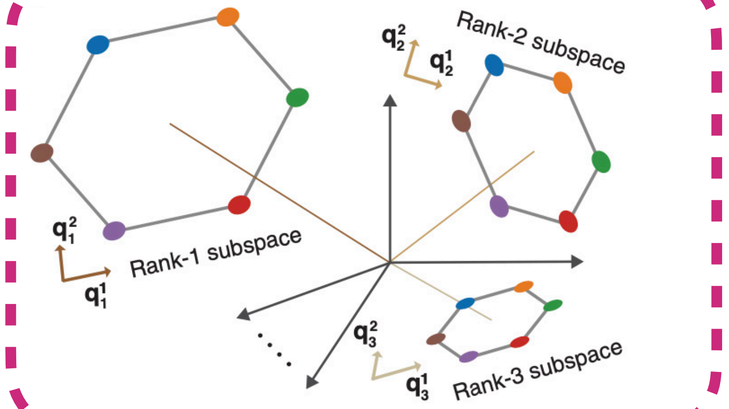


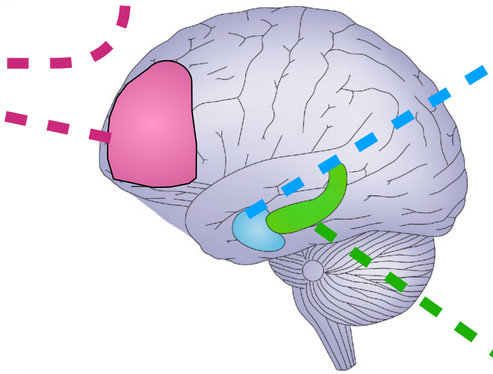
Place cells
O'Keefe & Dostrovsky, 1971



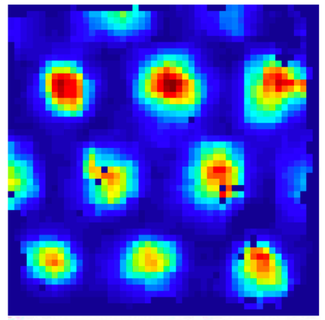
Landmark cells
Deshmukh & Knierim, 2013

But prefrontal cortex solves this problem in a different way

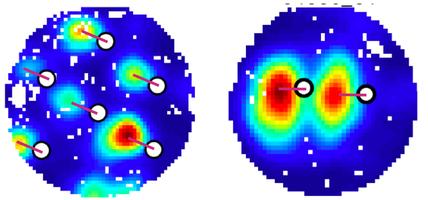




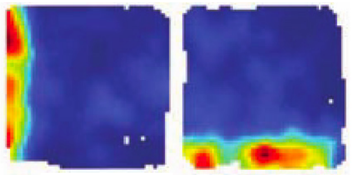
These cells tell you where you are or what you're seeing right now



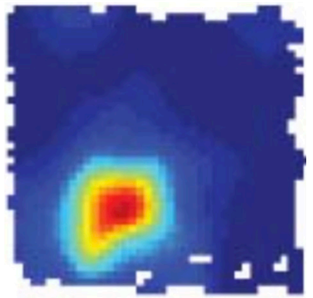
Grid cells
Hafting et al., 2005



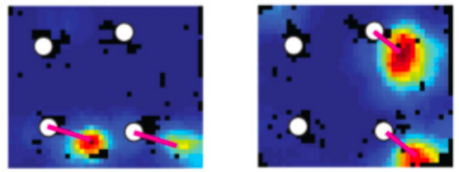
Object vector cells
Hoydal et al., 2018



Border cells
Solstad et al., 2008

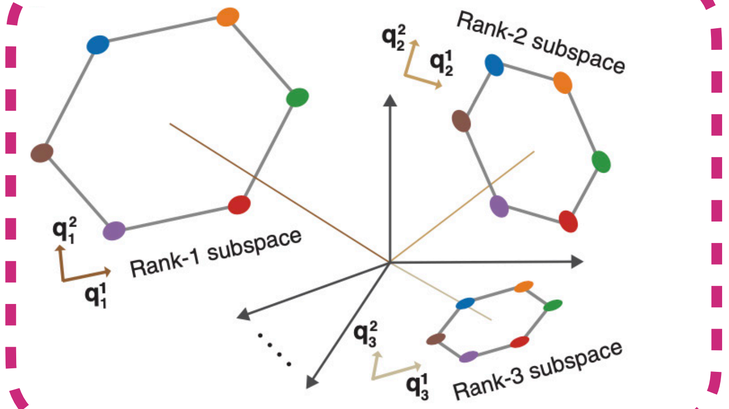


Place cells
O'Keefe & Dostrovsky, 1971

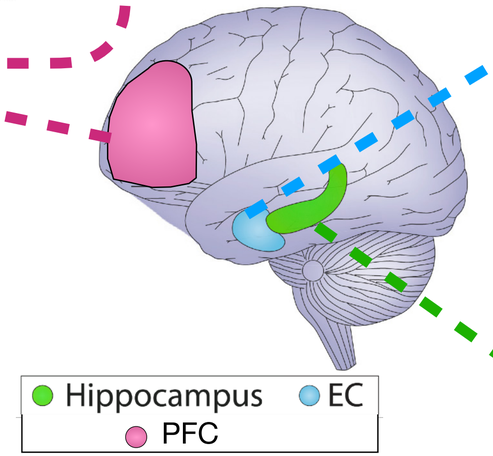


Landmark cells
Deshmukh & Knierim, 2013

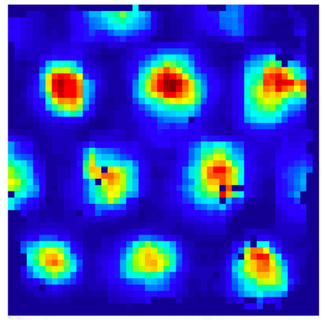
But prefrontal cortex solves this problem in a different way



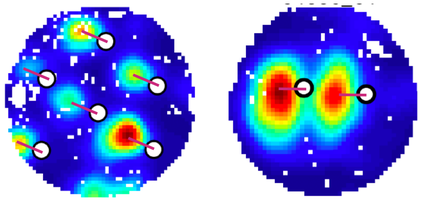
These cells include the past and future



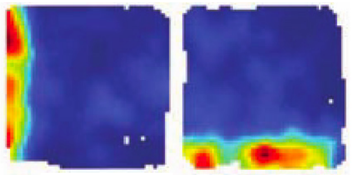
These cells tell you where you are or what you're seeing right now



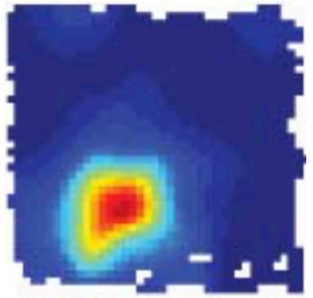
Grid cells
Hafting et al., 2005



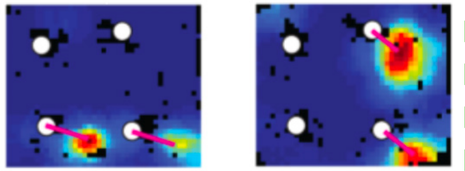
Object vector cells
Hoydal et al., 2018



Border cells
Solstad et al., 2008



Place cells
O'Keefe & Dostrovsky, 1971



Landmark cells
Deshmukh & Knierim, 2013

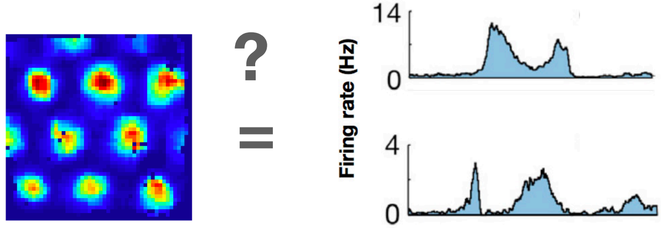
Puzzles of cognitive maps in the brain

Puzzles of cognitive maps in the brain

How does the same system do space and non-space?

Puzzles of cognitive maps in the brain

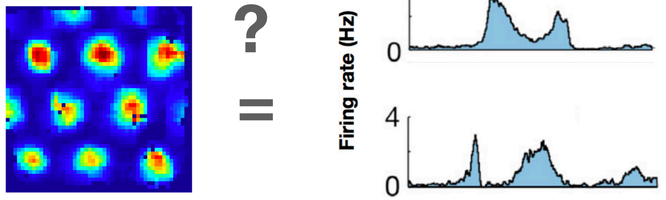
How does the same system do space and non-space?



How can brains learn these maps?

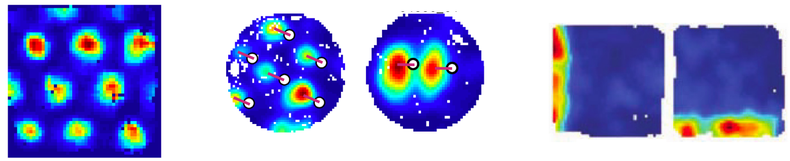
Puzzles of cognitive maps in the brain

How does the same system do space and non-space?



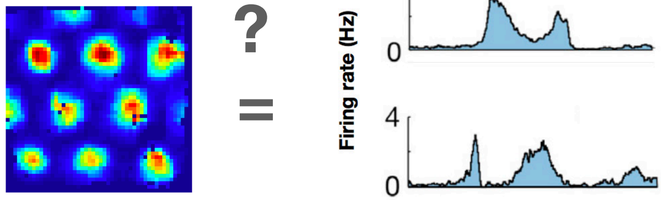
How can brains learn these maps?

Why do the neuron look the ways they do?



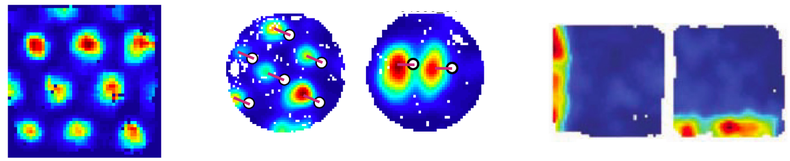
Puzzles of cognitive maps in the brain

How does the same system do space and non-space?

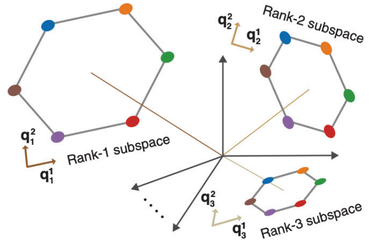


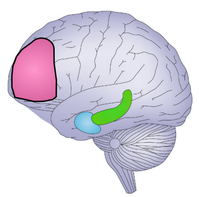
How can brains learn these maps?

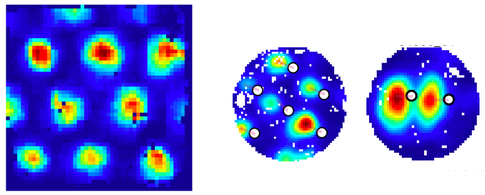
Why do the neuron look the ways they do?



How do different brain regions solve the same problem in different ways?





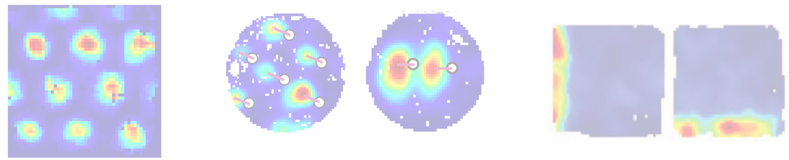


Puzzles of cognitive maps in the brain

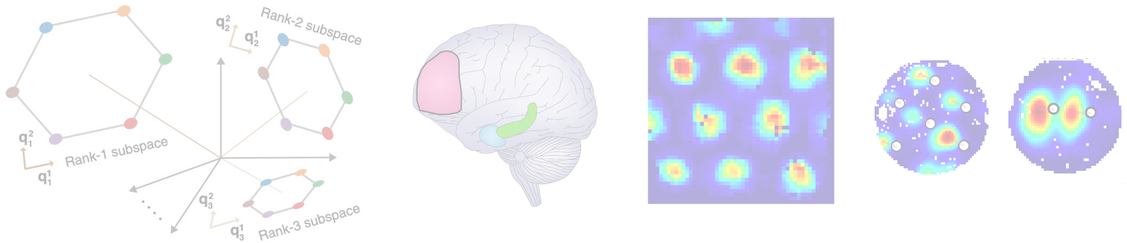
How does the same system do space and non-space?

How can brains learn these maps?

Why do the neuron look the ways they do?

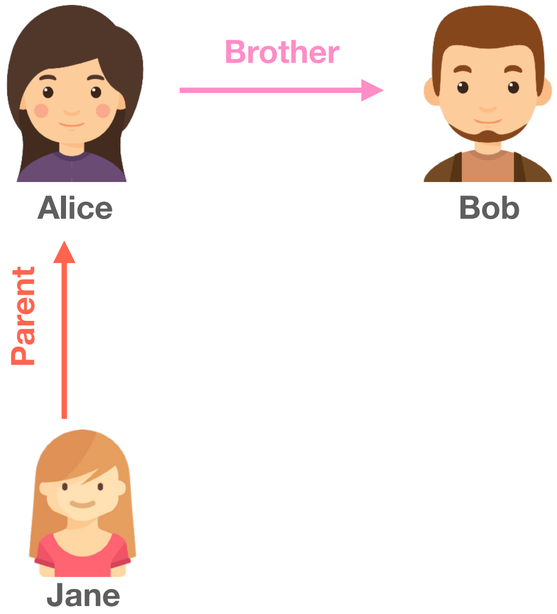


How do different brain regions solve the same problem in different ways?

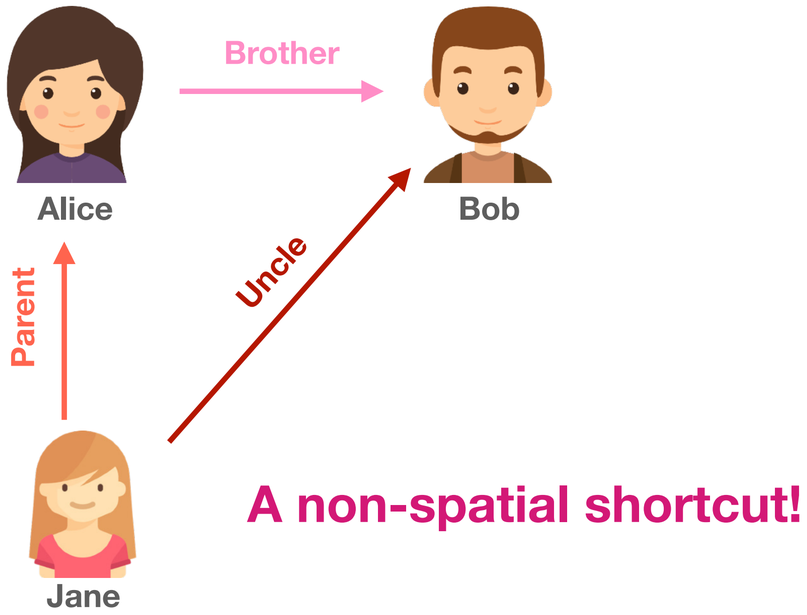


Space and non-space unified by abstracting relationships

Space and non-space unified by abstracting relationships

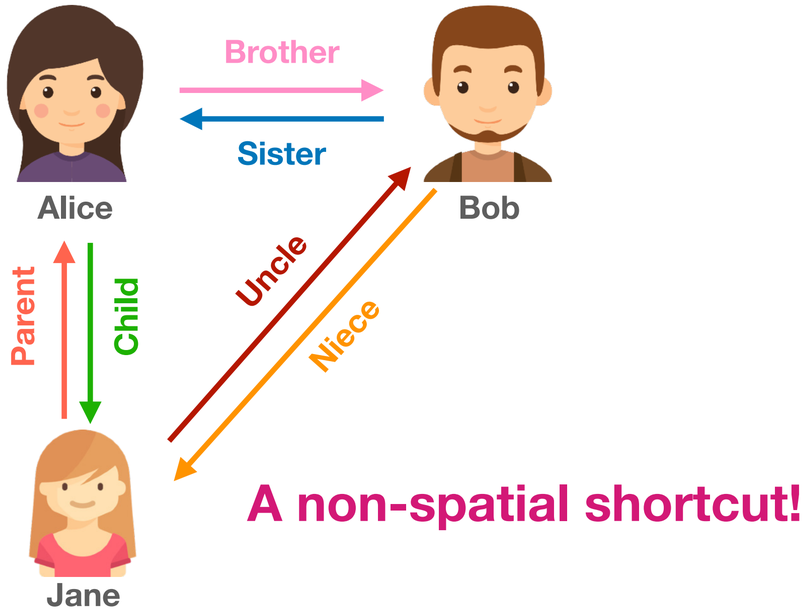


Space and non-space unified by abstracting relationships



A non-spatial shortcut!

Space and non-space unified by abstracting relationships

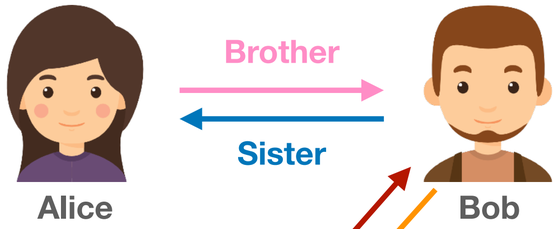


A non-spatial shortcut!

Parent + Brother = Uncle

Know which relationships add up to the same thing = path integration

Space and non-space unified by abstracting relationships



Brother

Sister

Parent

Child

Uncle

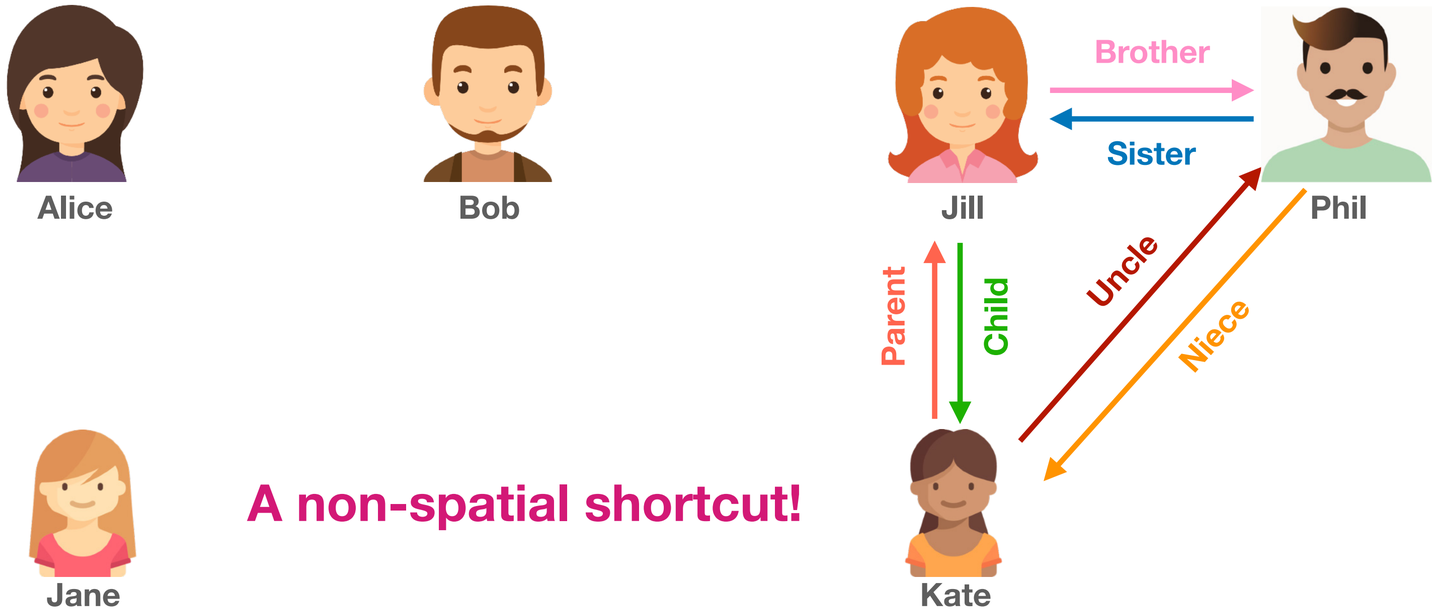
Niece

A non-spatial shortcut!

Parent + Brother = Uncle

Know which relationships add up to the same thing = path integration

Space and non-space unified by abstracting relationships



A non-spatial shortcut!

Parent + Brother = Uncle

Know which relationships add up to the same thing = path integration

Generalising abstract relationships to different scenarios

Space and non-space unified by abstracting relationships

Alice

Bob

Jill

Brother

Sister

Phil

Parent

Child

Uncle

Niece

Jane

A non-spatial shortcut!

Kate

London

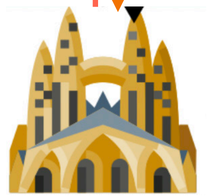
East

West

Berlin

Parent

South



Barcelona

East

West

Parent

South

Rome

Parent + Brother = Uncle

North + East + South + West = 0

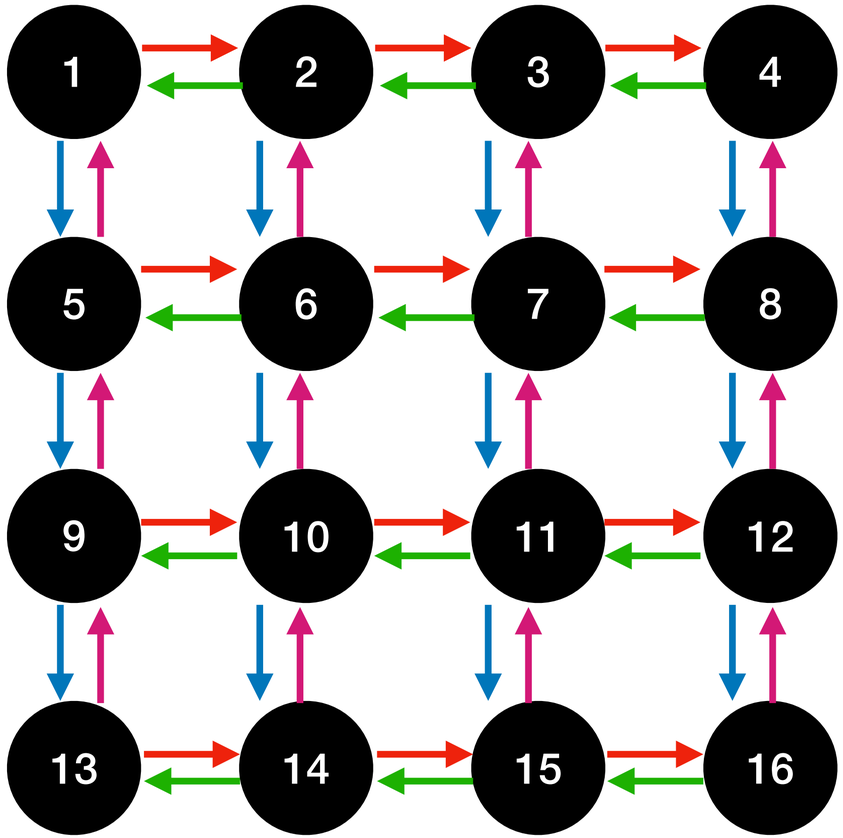
Know which relationships add up to the same thing = path integration

Generalising abstract relationships to different scenarios

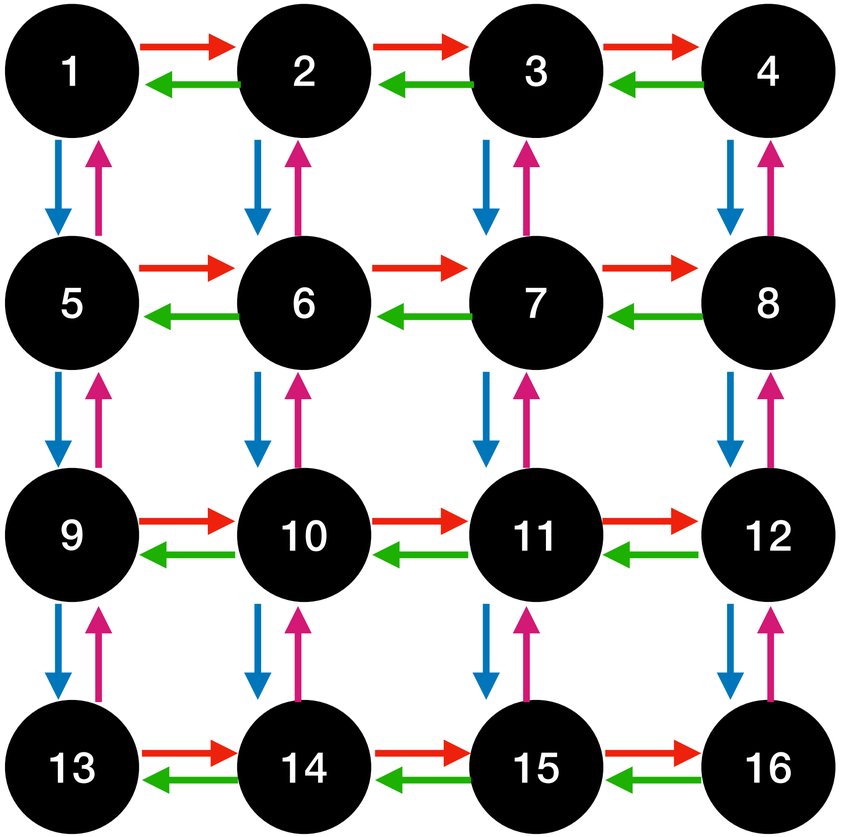
This is just like traditional path integration in space

Formalising relationships using graphs

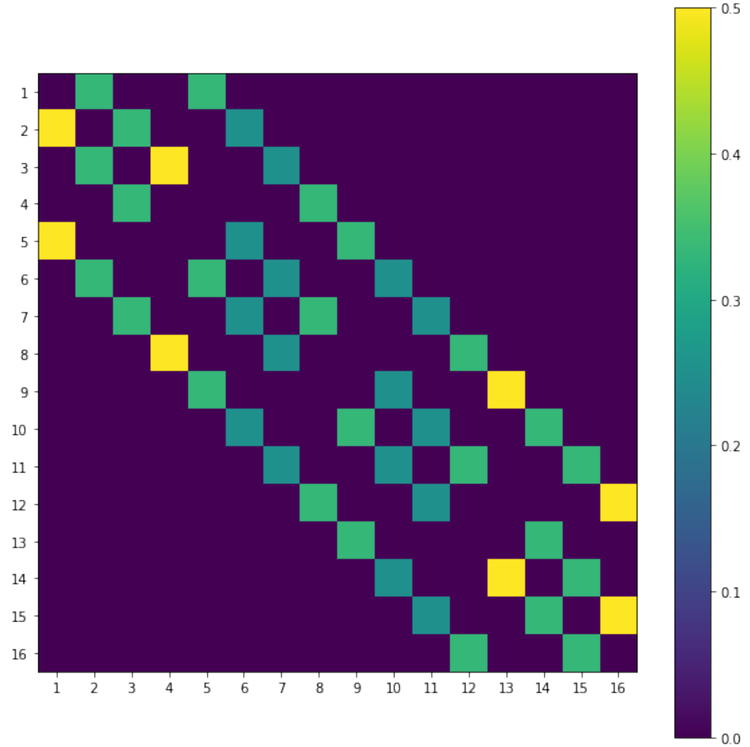
Formalising relationships using graphs



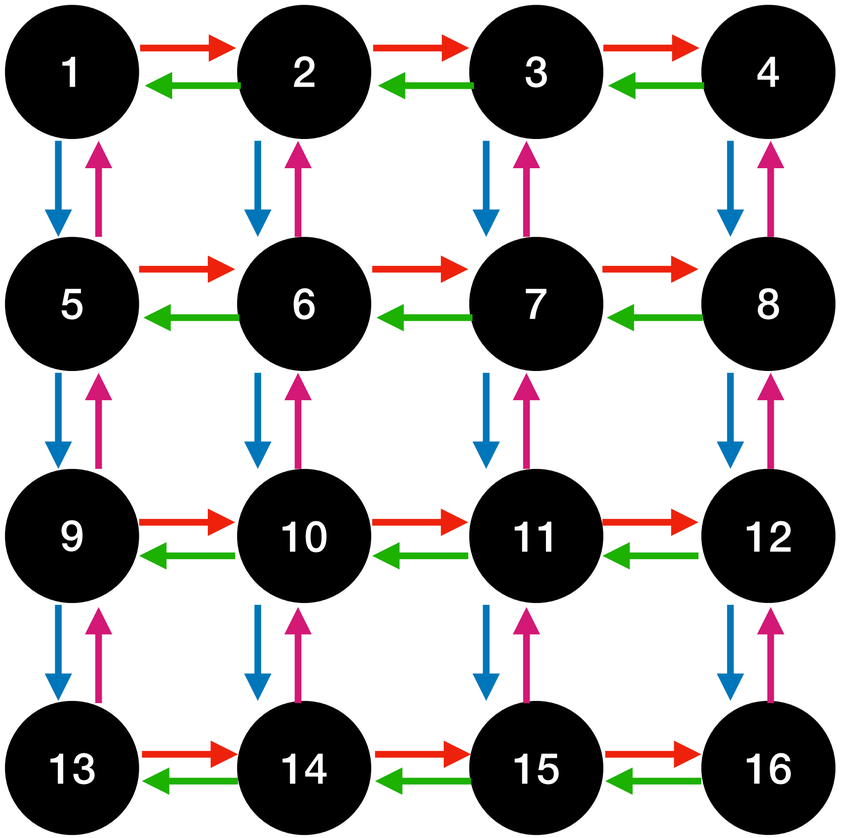
Formalising relationships using graphs



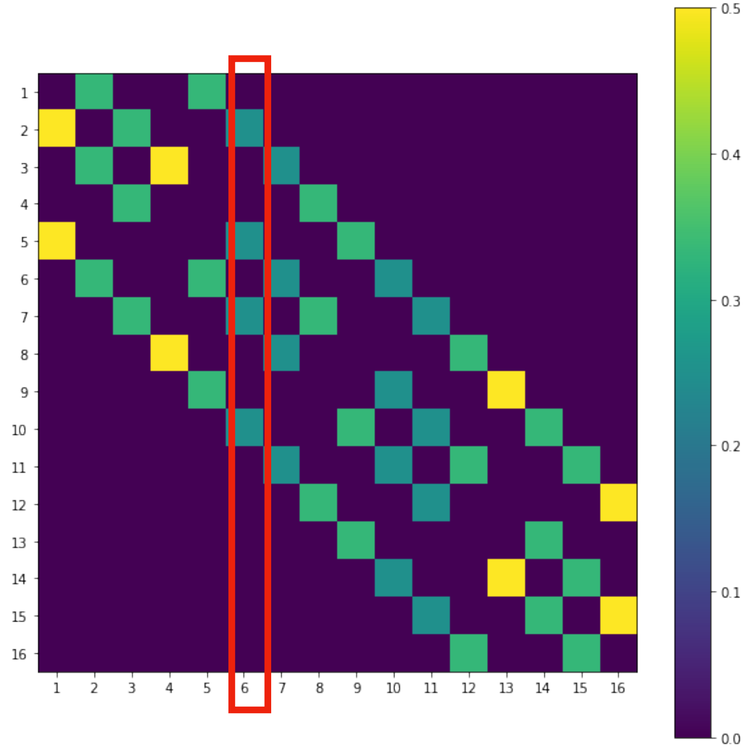
Transition matrix T_{ij}
= probability of going from state j to state I



Formalising relationships using graphs

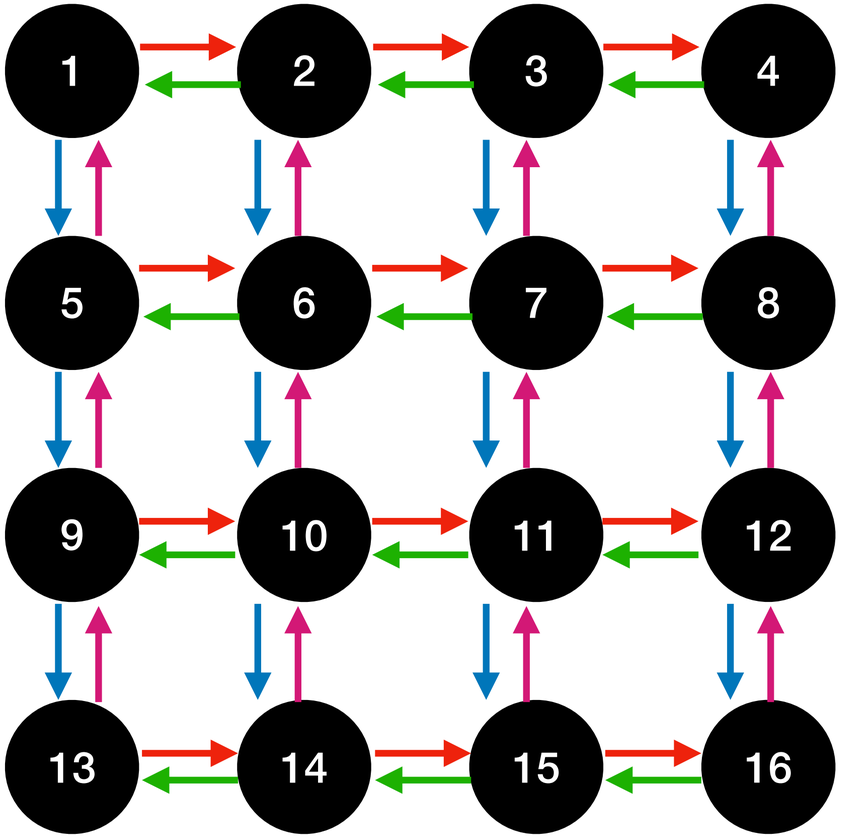


Transition matrix T_{ij}
= probability of going from state j to state I

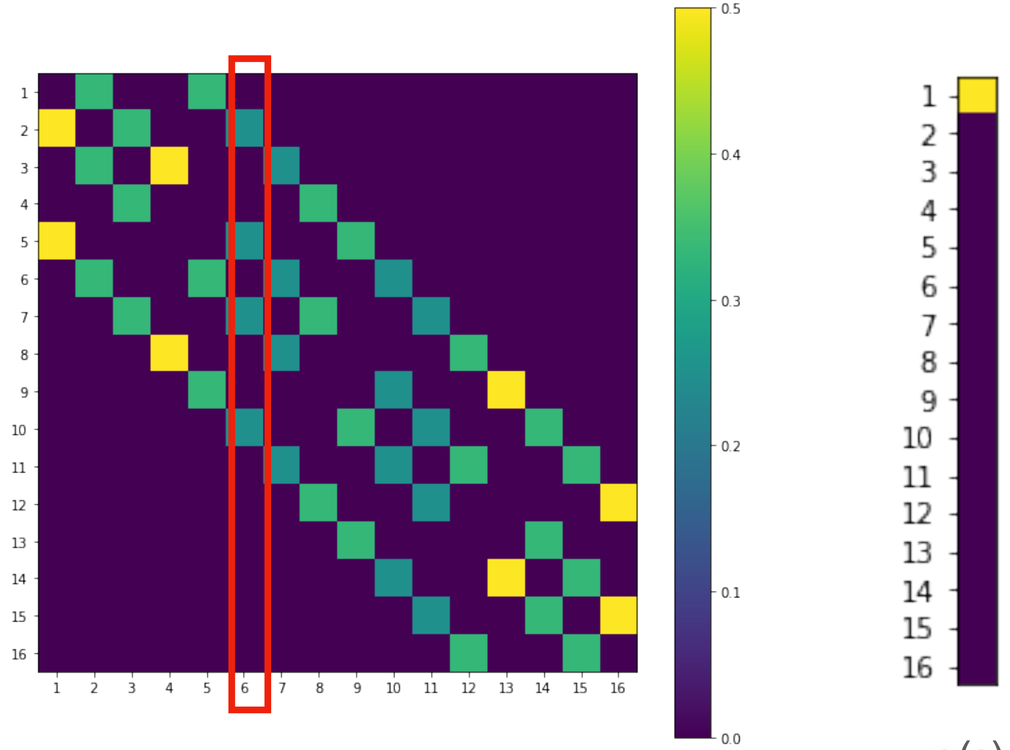


If I am in state 6, where will I be next?

Formalising relationships using graphs



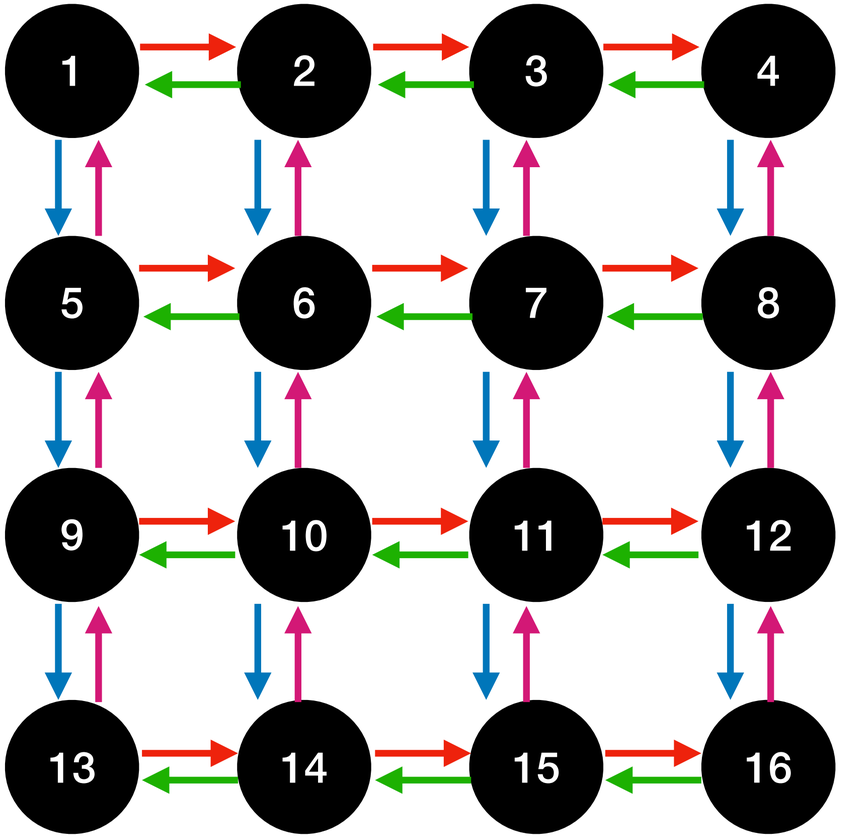
Transition matrix T_{ij}
= probability of going from state j to state I



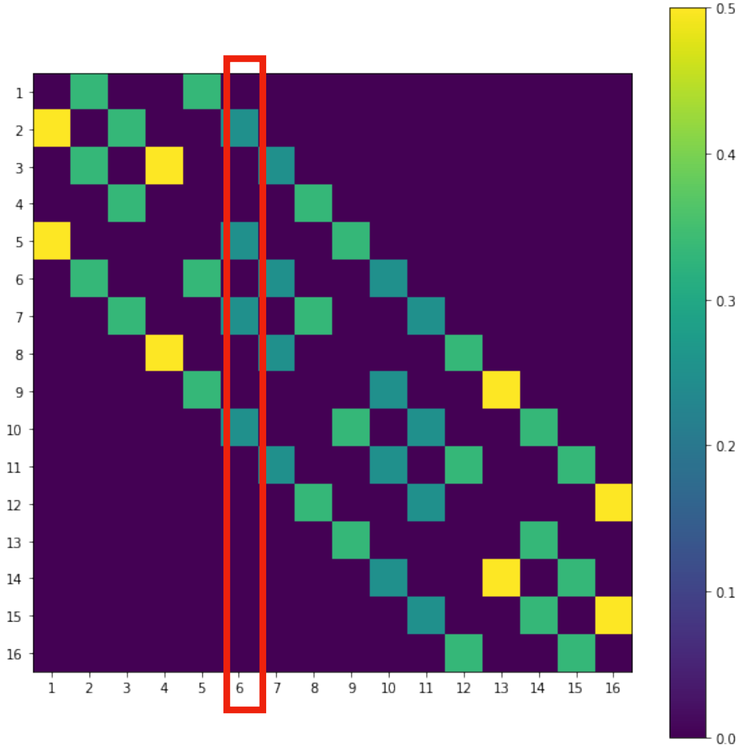
p(s)

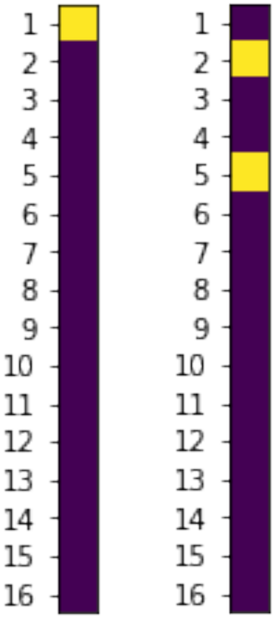
If I am in state 6, where will I be next?

Formalising relationships using graphs



Transition matrix T_{ij}
= probability of going from state j to state I



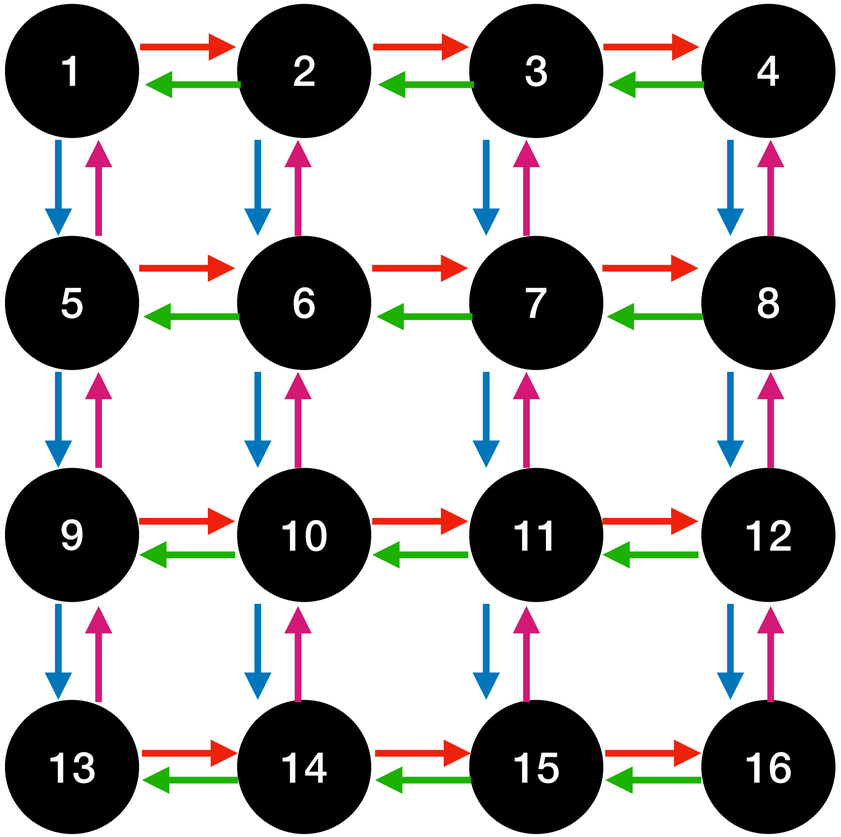


p(s)

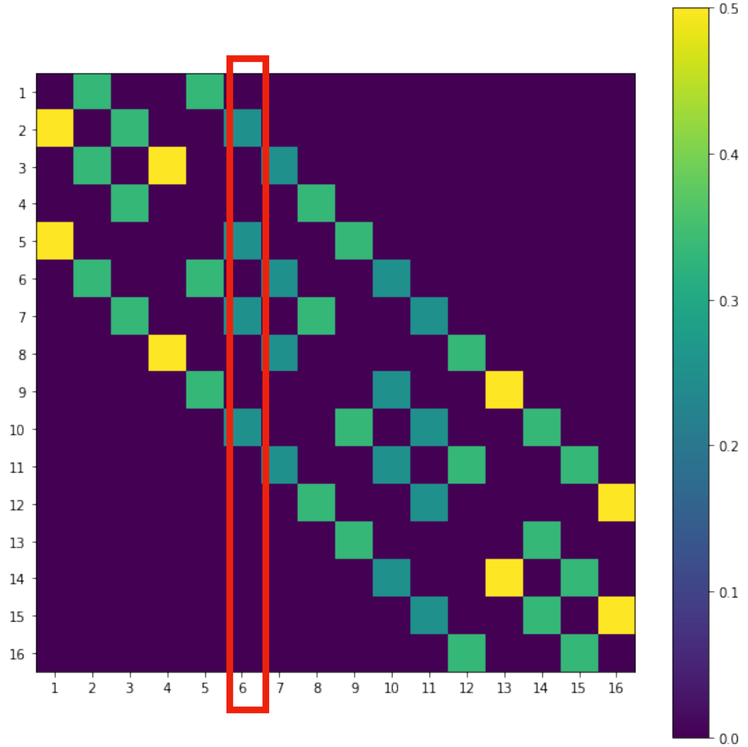
Tp(s)

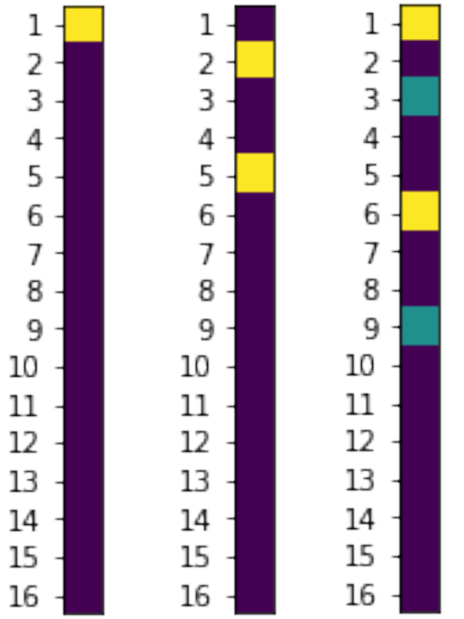
If I am in state 6, where will I be next?

Formalising relationships using graphs



Transition matrix T_{ij}
= probability of going from state j to state I



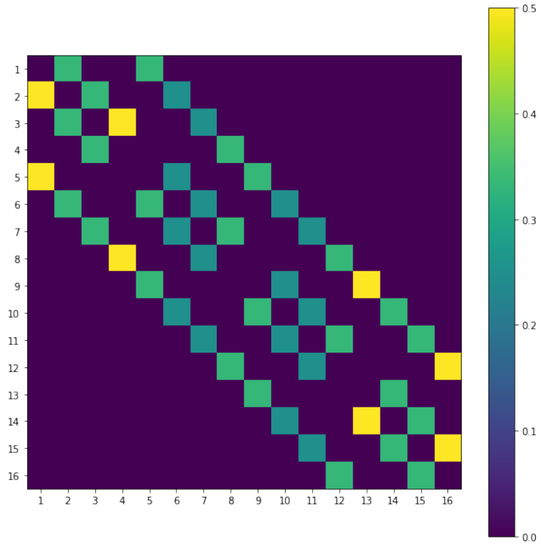


p(s) Tp(s) $T^2p(s)$

If I am in state 6, where will I be next?

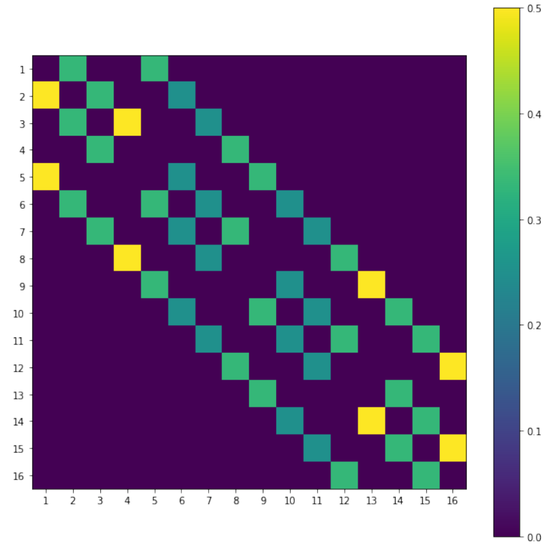
The successor representation

The transition matrix says where you'll likely be in 1 step's time



The successor representation

The transition matrix says where you'll likely be in 1 step's time

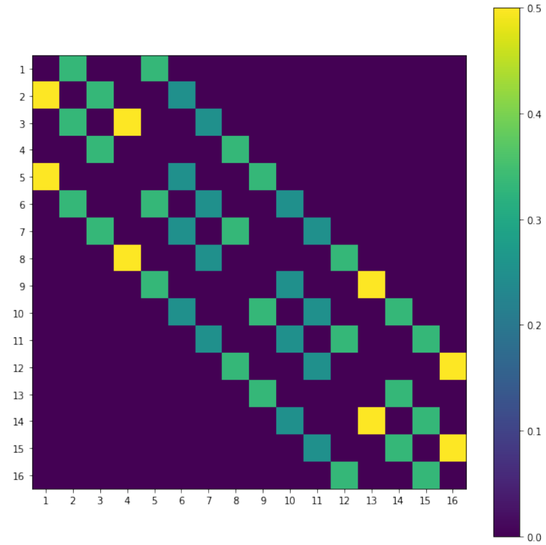


The successor representation says where you'll likely be over all future time steps

$$SR = I + T + T^2 + T^3 + \ldots$$

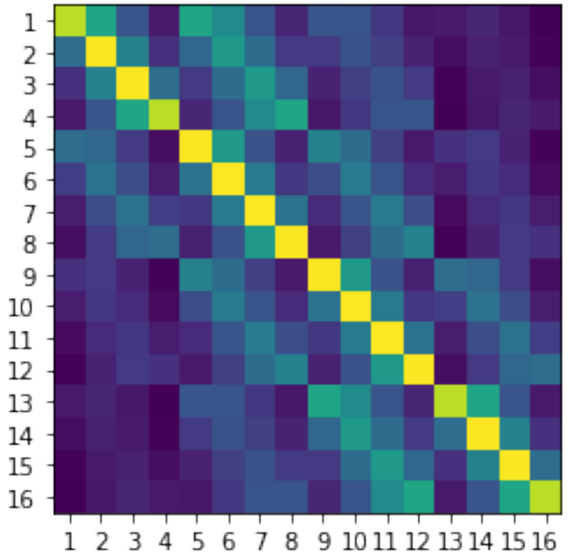
The successor representation

The transition matrix says where you'll likely be in 1 step's time



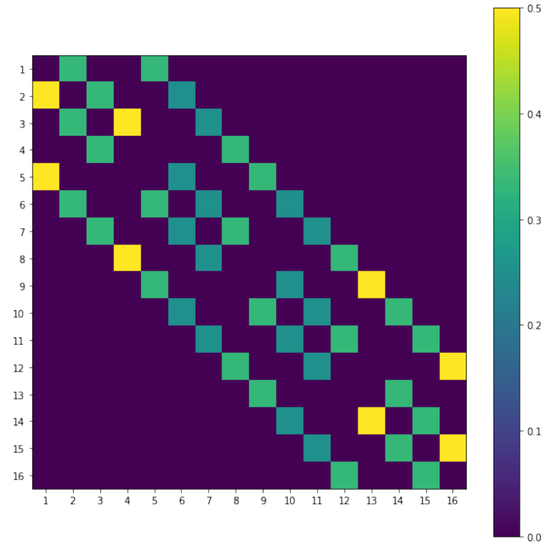
The successor representation says where you'll likely be over all future time steps

$$SR = I + T + T^2 + T^3 + ...$$



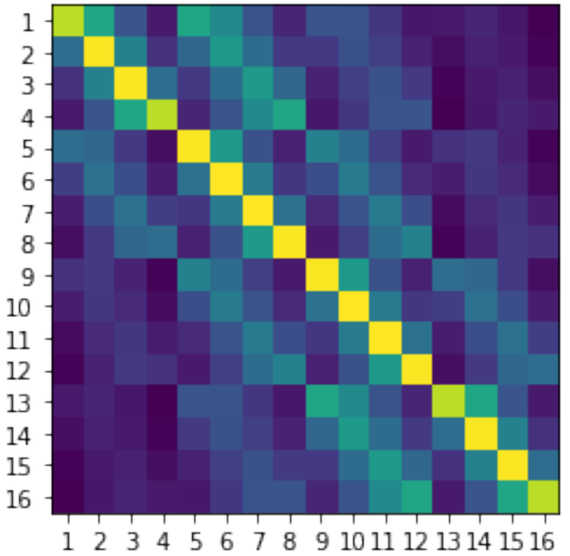
The successor representation

The transition matrix says where you'll likely be in 1 step's time



The successor representation says where you'll likely be over all future time steps

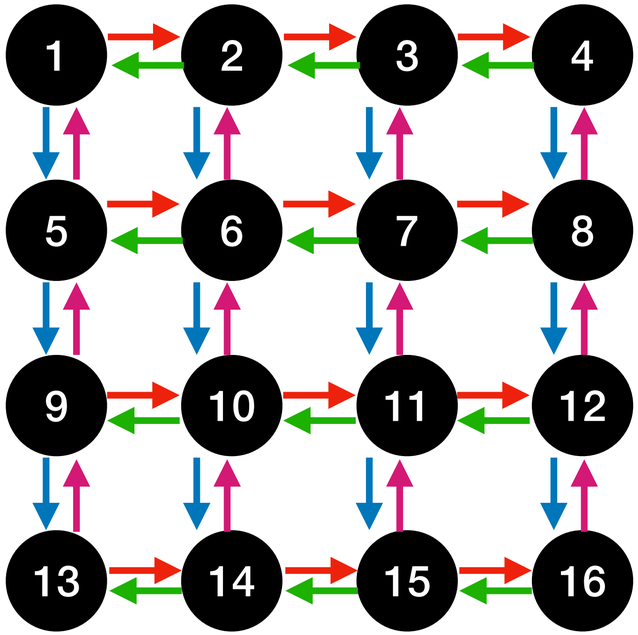
$$SR = I + T + T^2 + T^3 + ...$$



Often discounted so the sum converges

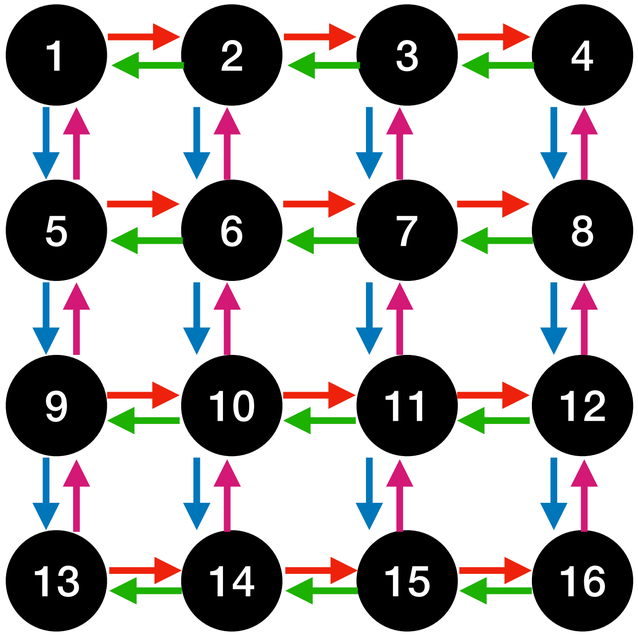
$$SR = I + \gamma T + \gamma^2 T^2 + \gamma^3 T^3 + ...$$

It looks a bit like place cells…

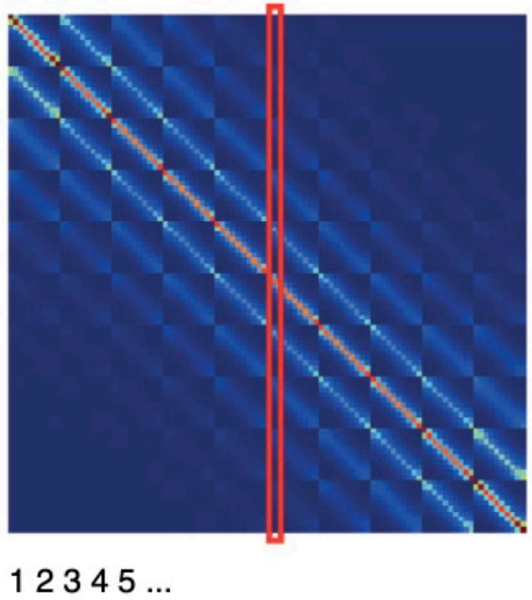


Stachenfeld, Botvinick, Gershman (2017)

It looks a bit like place cells…

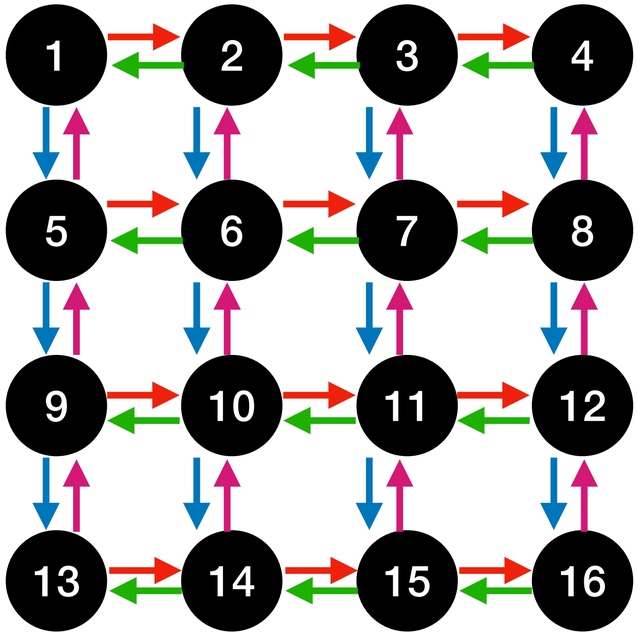


The successor representation

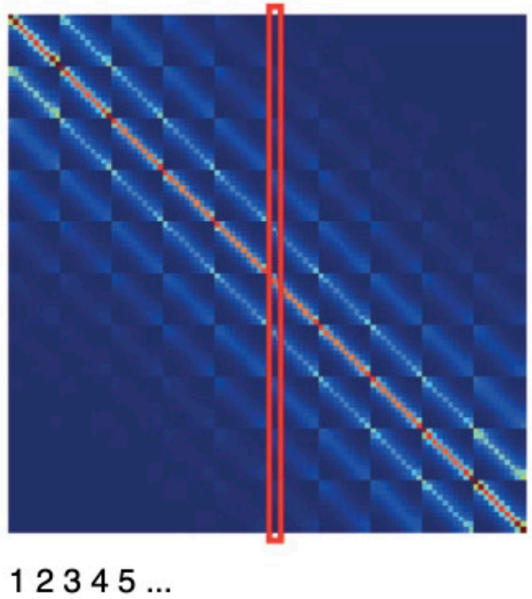


Stachenfeld, Botvinick, Gershman (2017)

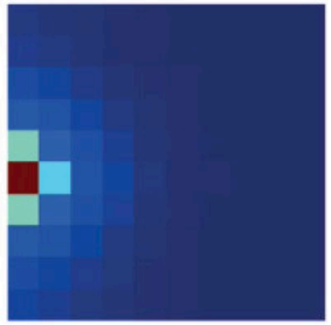
It looks a bit like place cells…



The successor representation

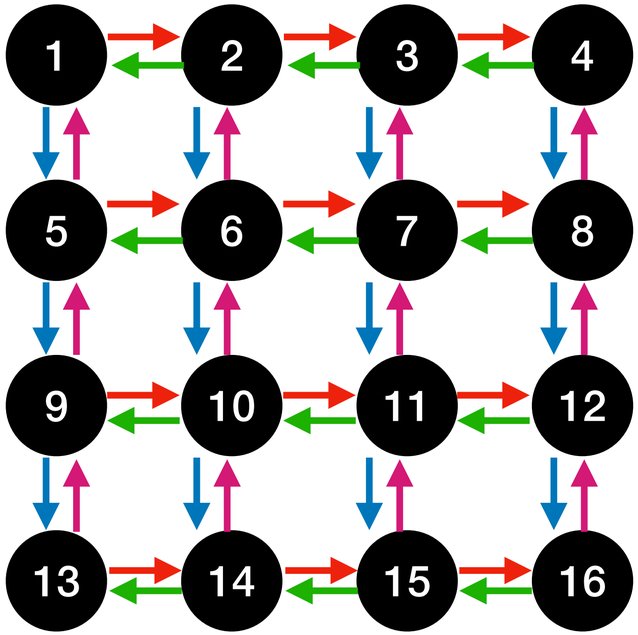


The column looks like a place cell

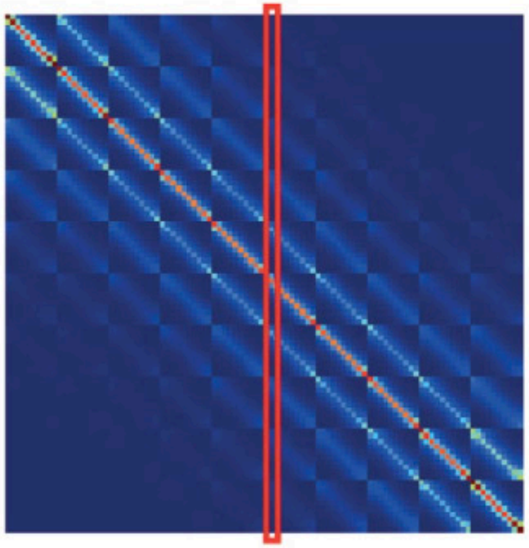


Stachenfeld, Botvinick, Gershman (2017)

It looks a bit like place cells…

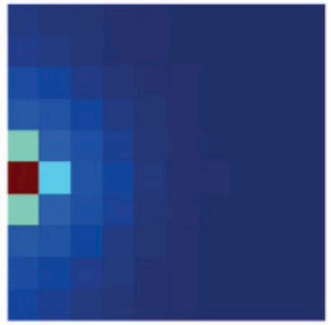


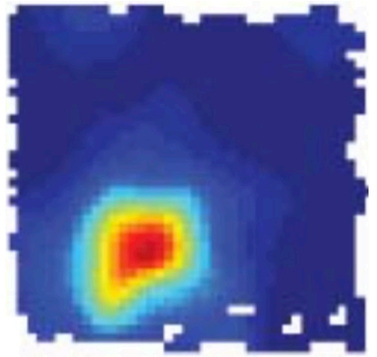
The successor representation



1 2 3 4 5 ...

The column looks like a place cell





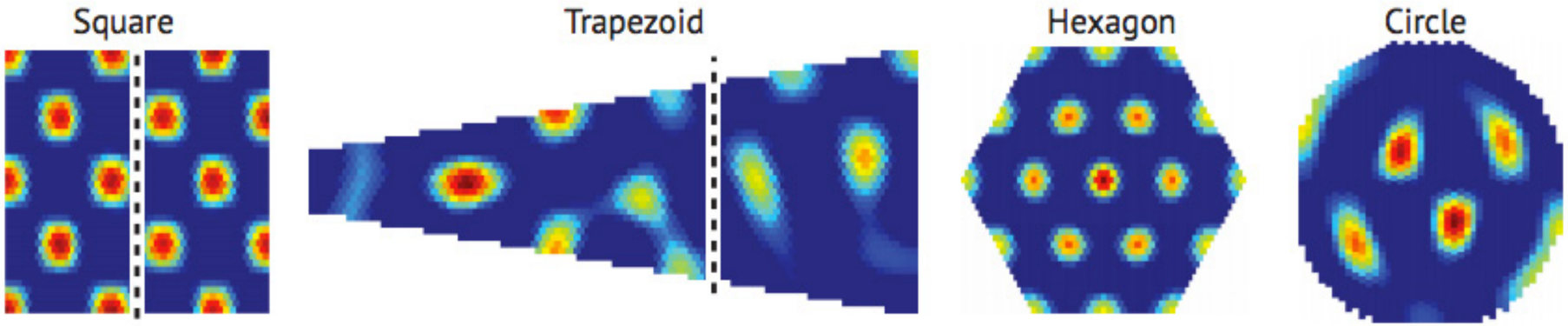
Place cells
O'Keefe & Dostrovsky, 1971

Stachenfeld, Botvinick, Gershman (2017)

Eigenvectors of the successor representation look grid cells

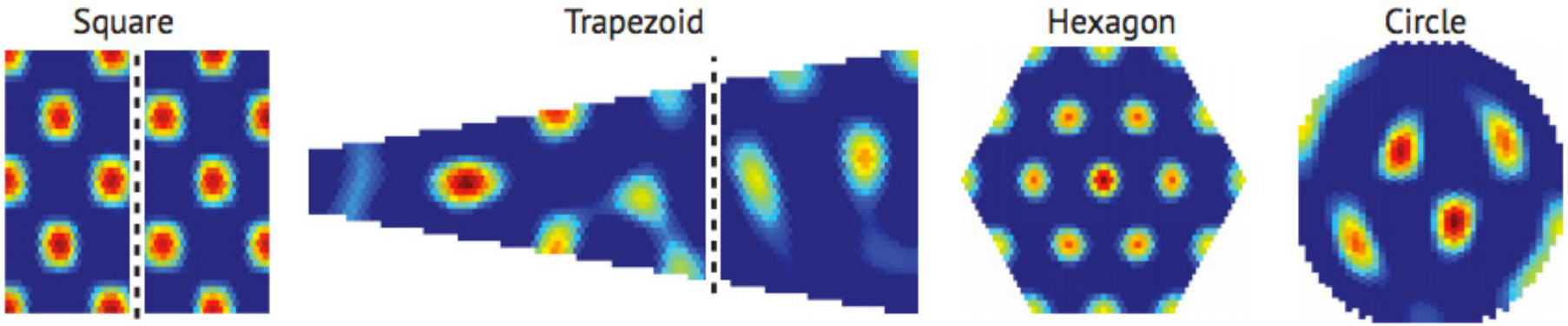
Stachenfeld, Botvinick, Gershman (2017)

Eigenvectors of the successor representation look grid cells

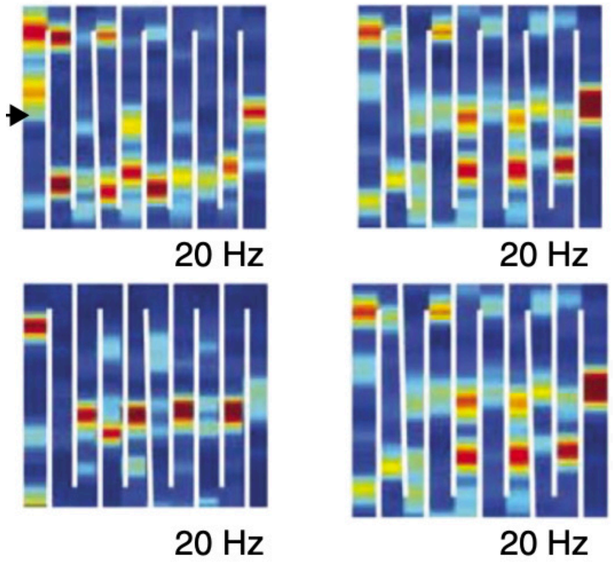


Stachenfeld, Botvinick, Gershman (2017)

Eigenvectors of the successor representation look grid cells

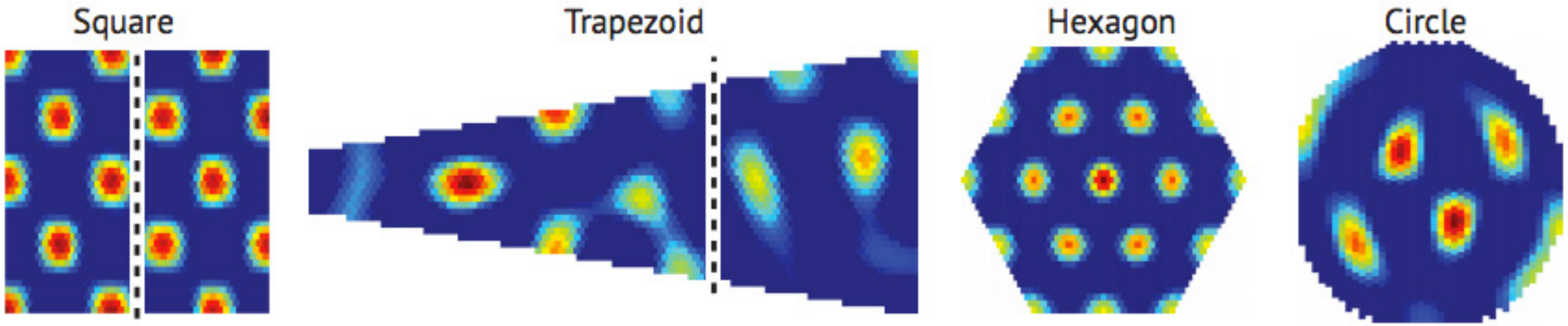


Grid cells in hairpin mazes ‘fragment’

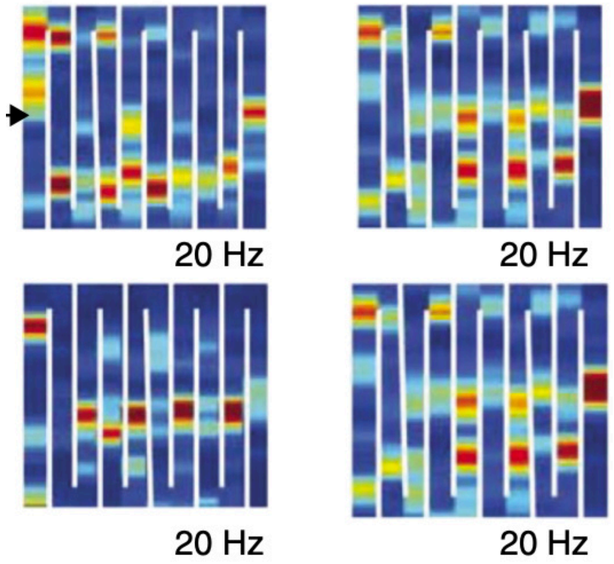


Stachenfeld, Botvinick, Gershman (2017)

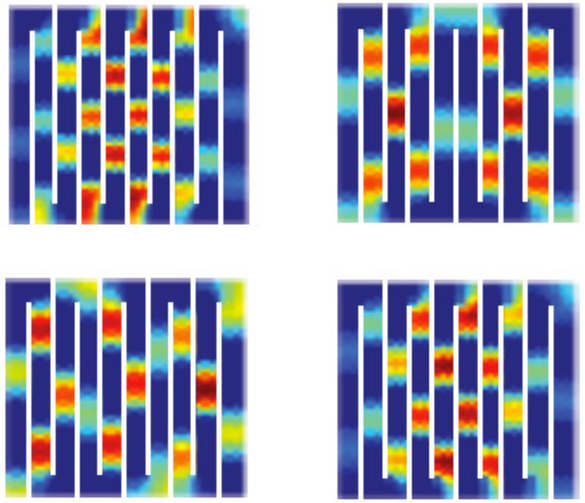
Eigenvectors of the successor representation look grid cells



Grid cells in hairpin mazes ‘fragment’



So too do the SR grid cells

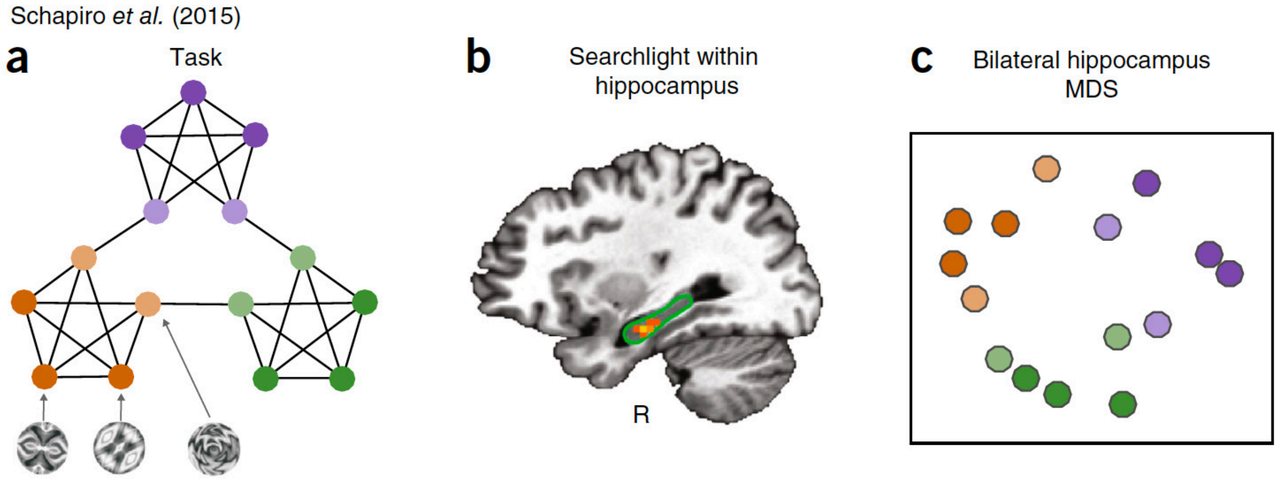


Stachenfeld, Botvinick, Gershman (2017)

Accounts for non-spatial brain representations too

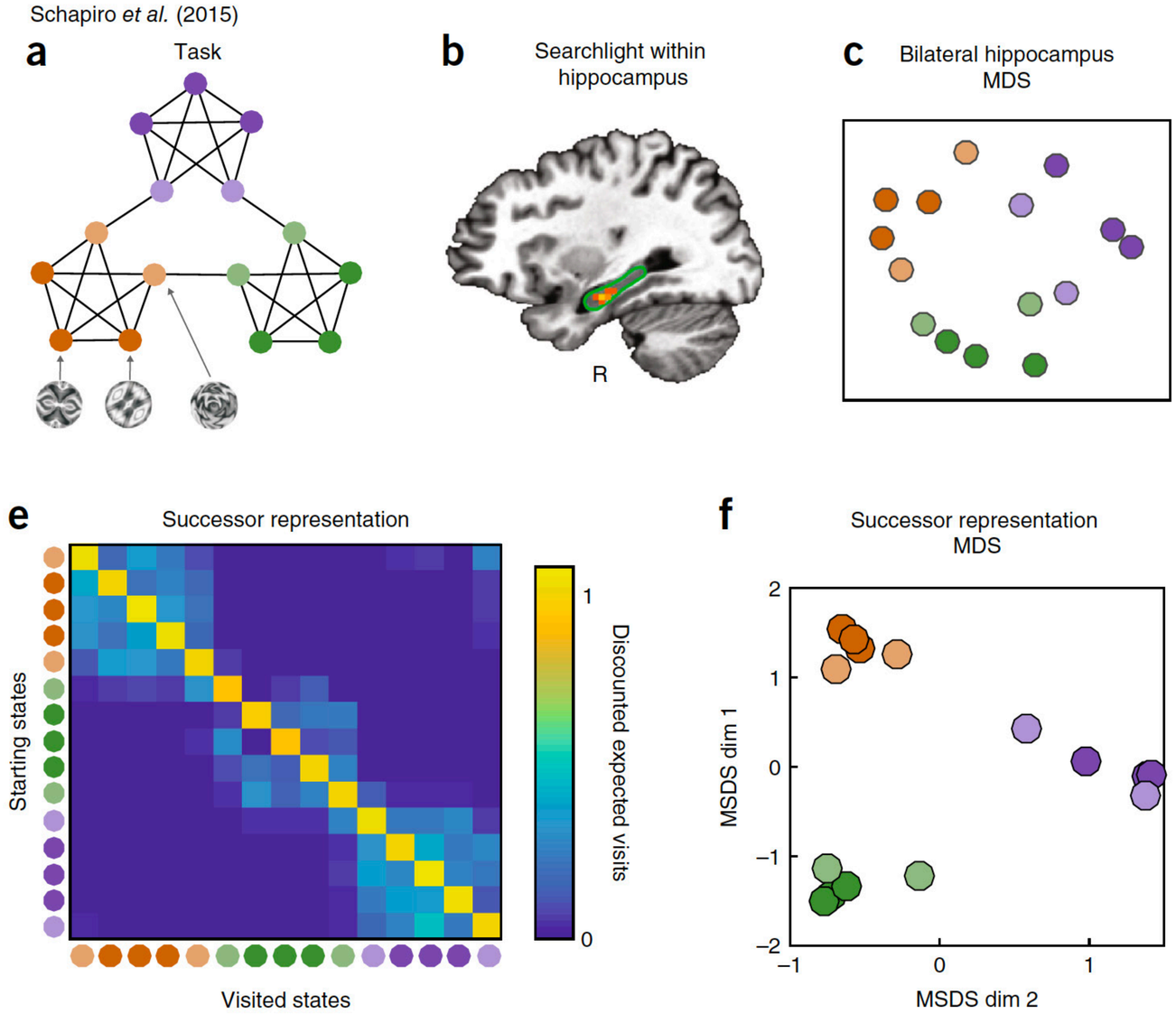
Stachenfeld, Botvinick, Gershman (2017)

Accounts for non-spatial brain representations too



Stachenfeld, Botvinick, Gershman (2017)

Accounts for non-spatial brain representations too



Stachenfeld, Botvinick, Gershman (2017)

We’ve understood a lot already

We’ve understood a lot already

… but the brain uses neurons

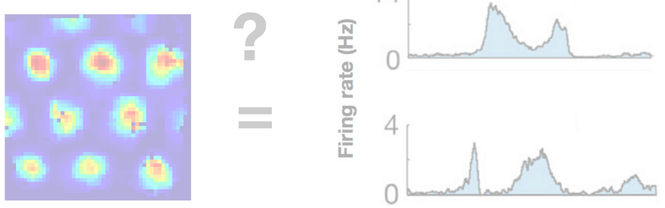
We've understood a lot already

… but the brain uses neurons

… and learns from sequential experience

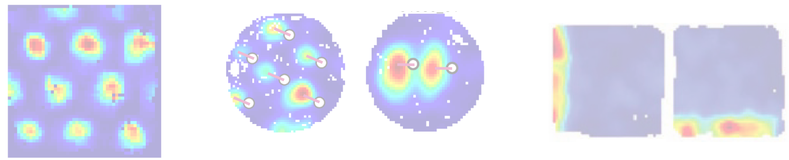
Puzzles of cognitive maps in the brain

How does the same system do space and non-space?

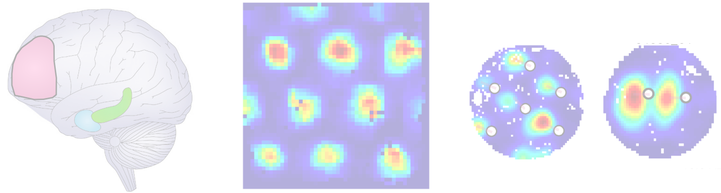


How can brains learn these maps?

Why do the neuron look the ways they do?

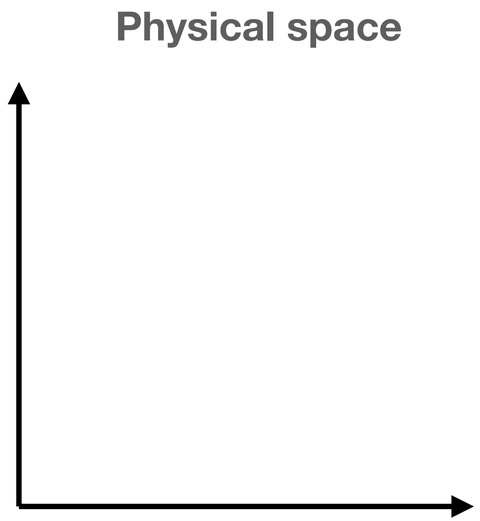


How do different brain regions solve the same problem in different ways?

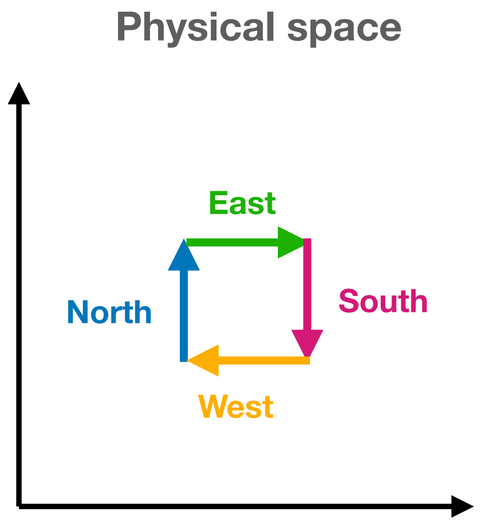


Recurrent neural networks learn from sequences

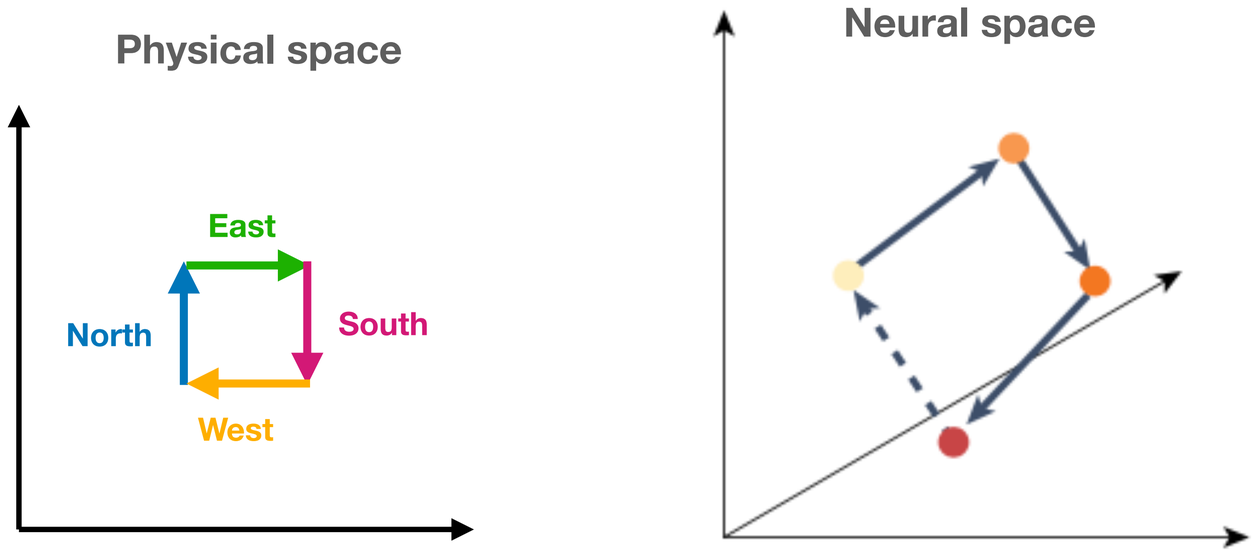
Recurrent neural networks learn from sequences



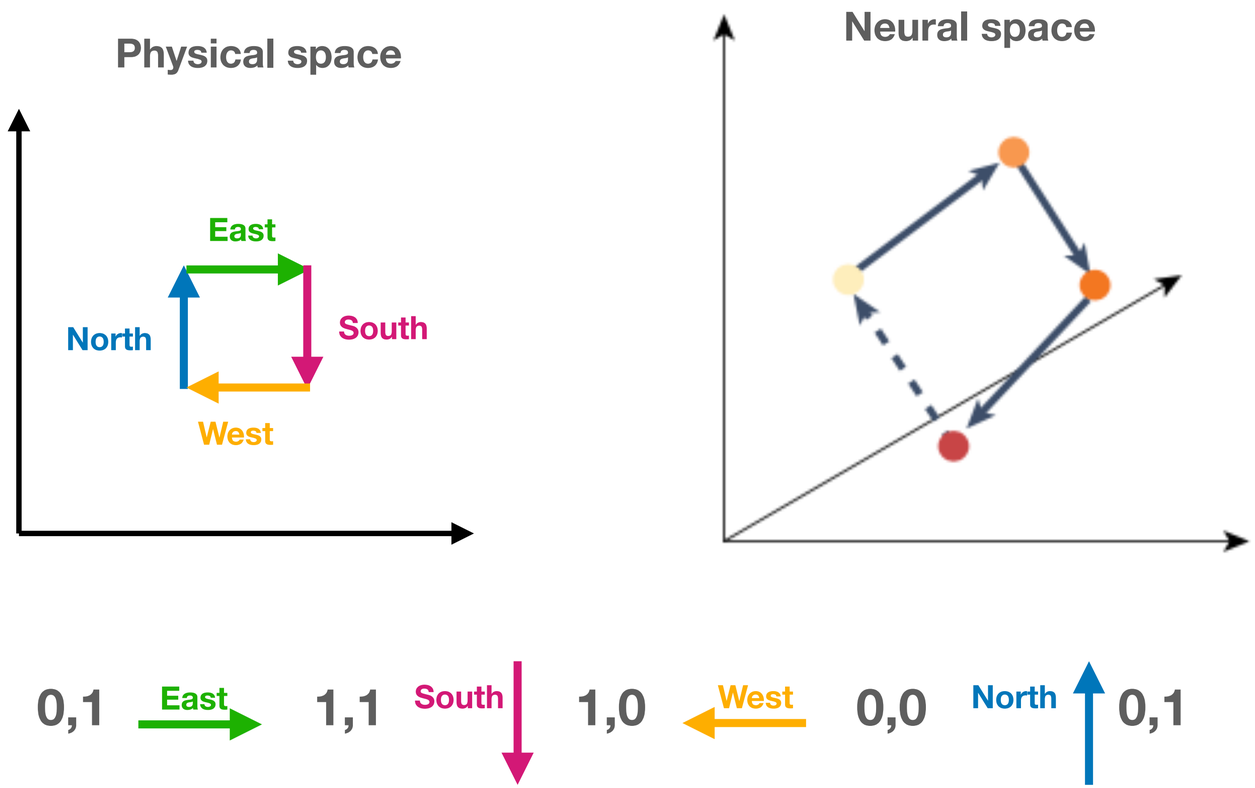
Recurrent neural networks learn from sequences



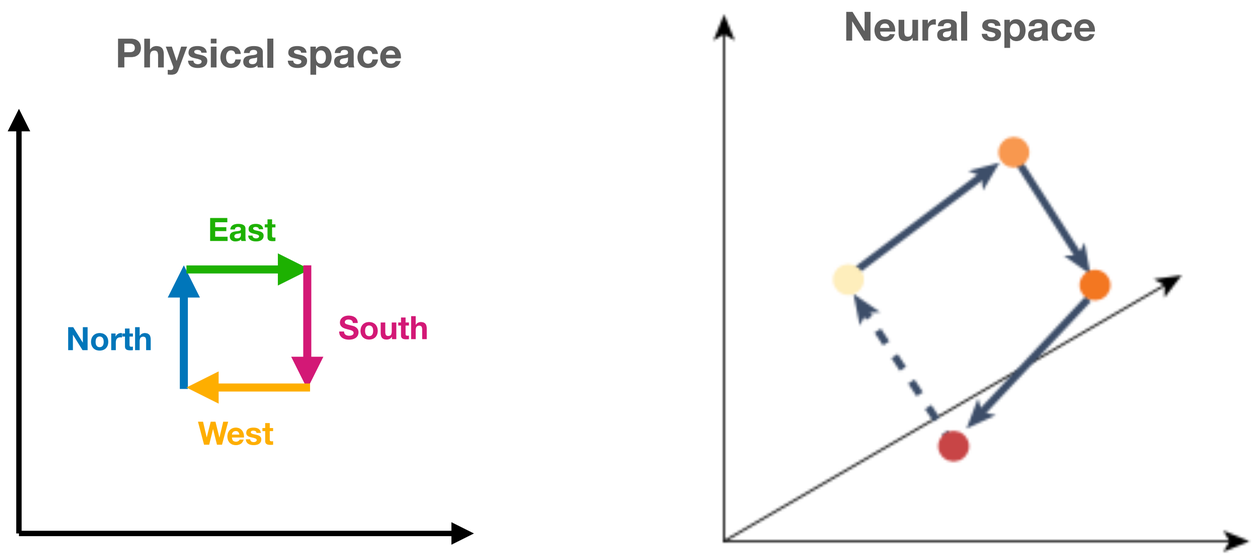
Recurrent neural networks learn from sequences

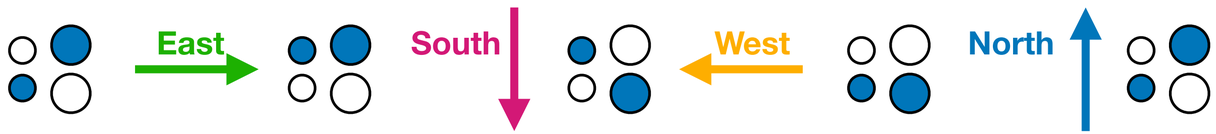


Recurrent neural networks learn from sequences

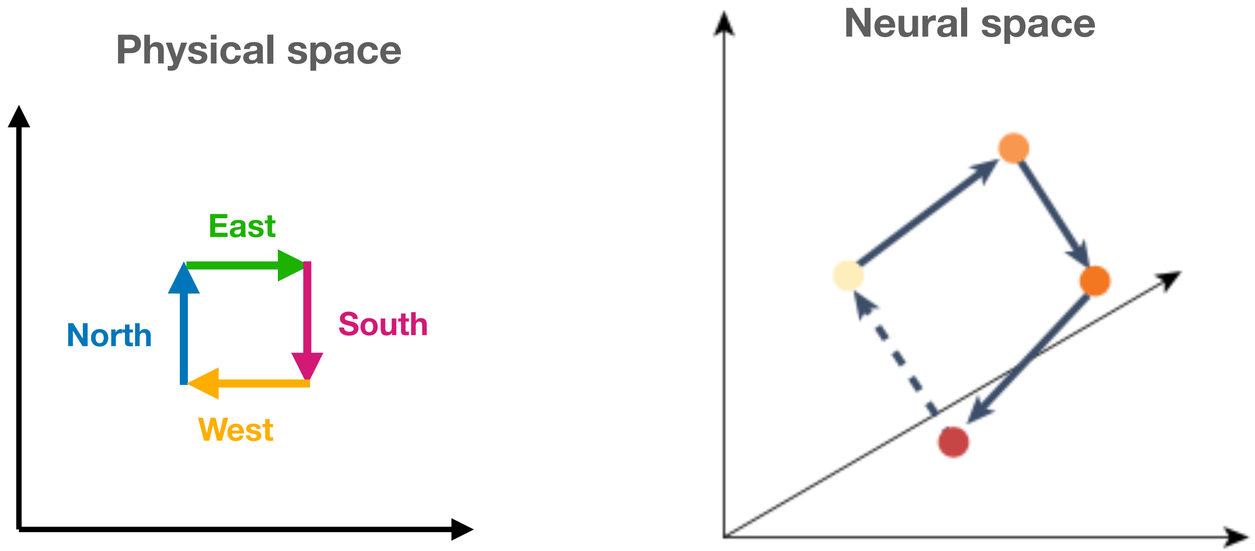


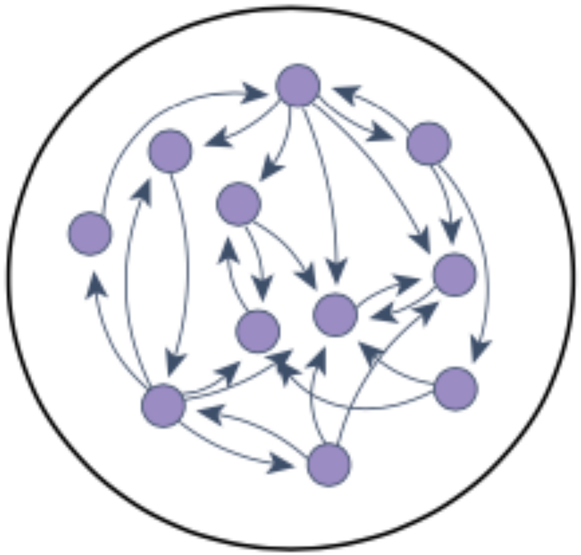
Recurrent neural networks learn from sequences

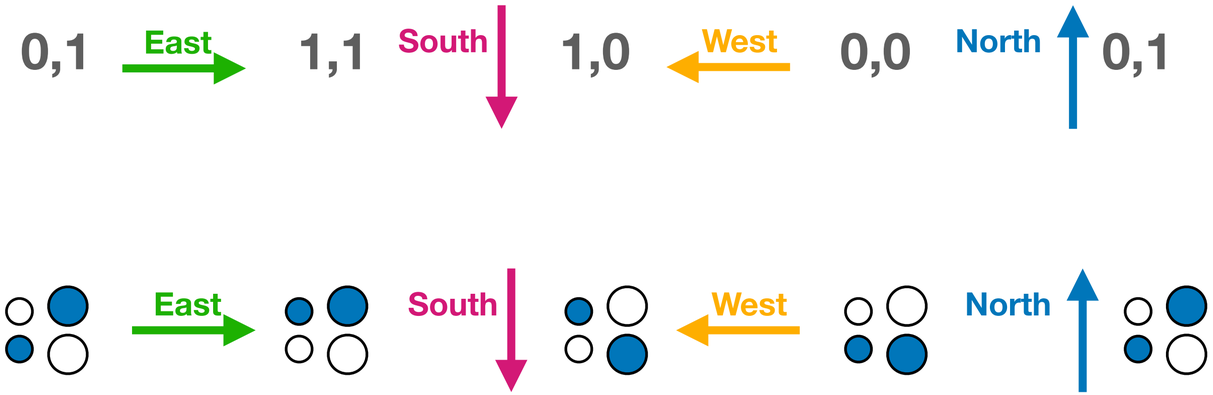




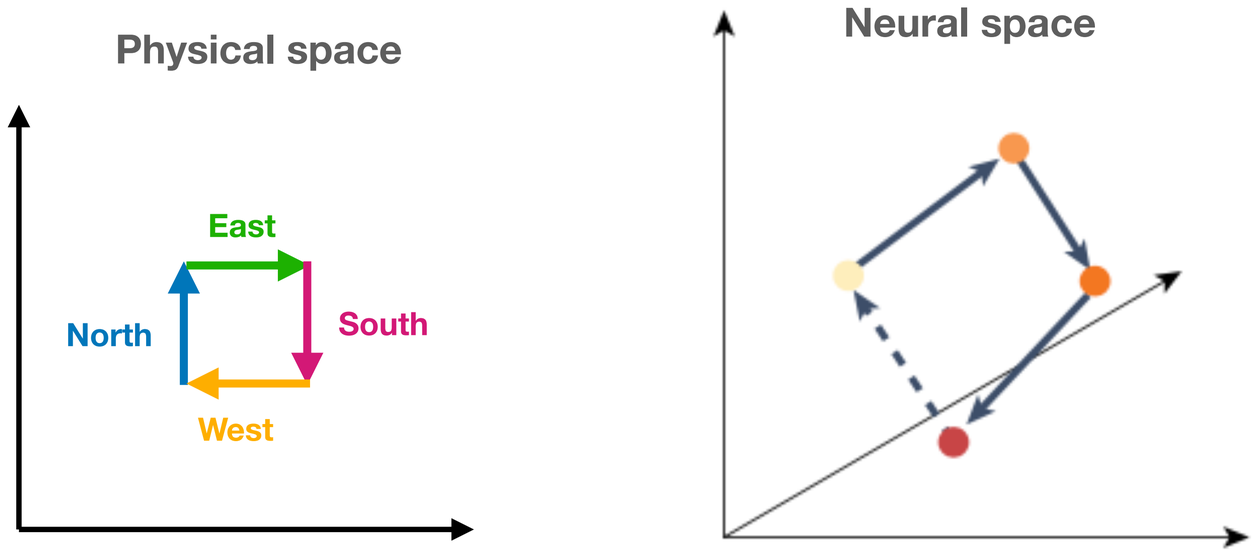
Recurrent neural networks learn from sequences

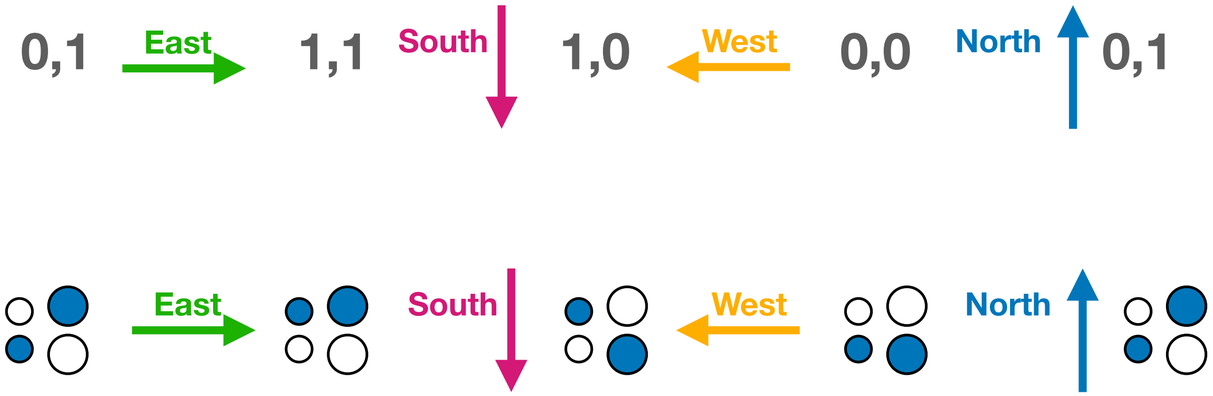


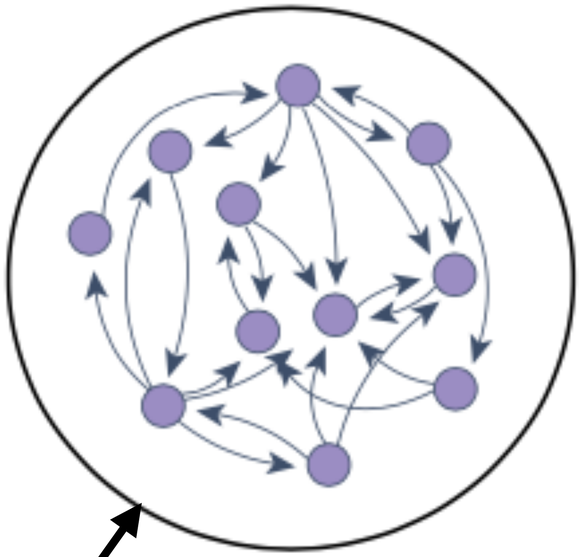




Recurrent neural networks learn from sequences

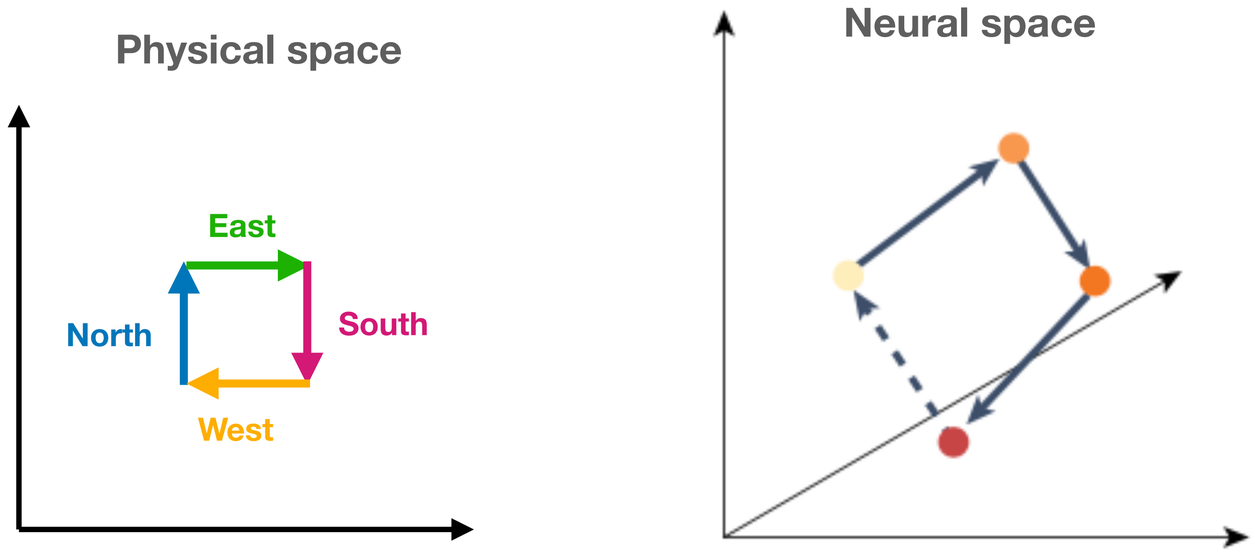


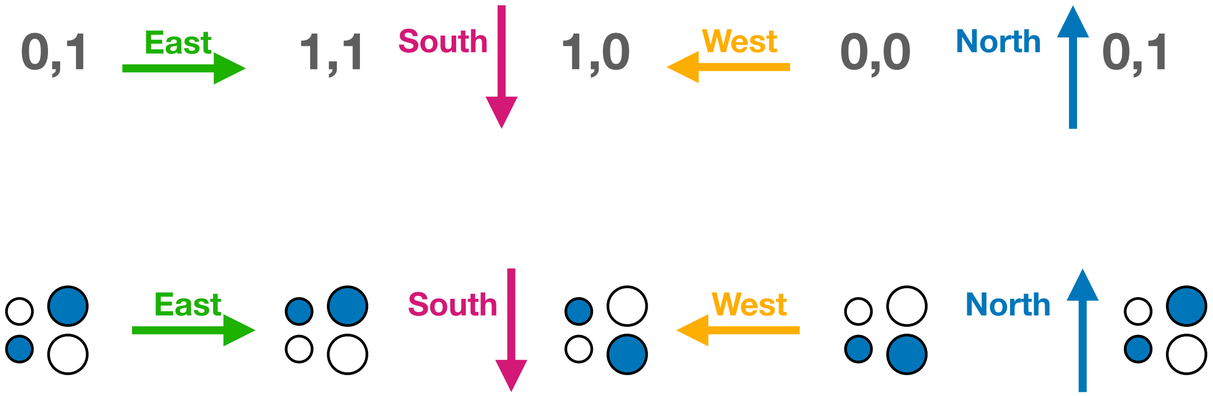


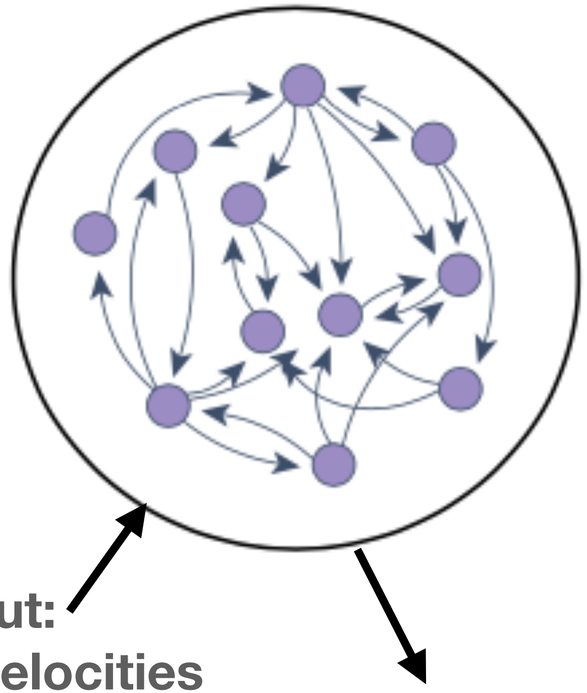


Input:
Actions/velocities

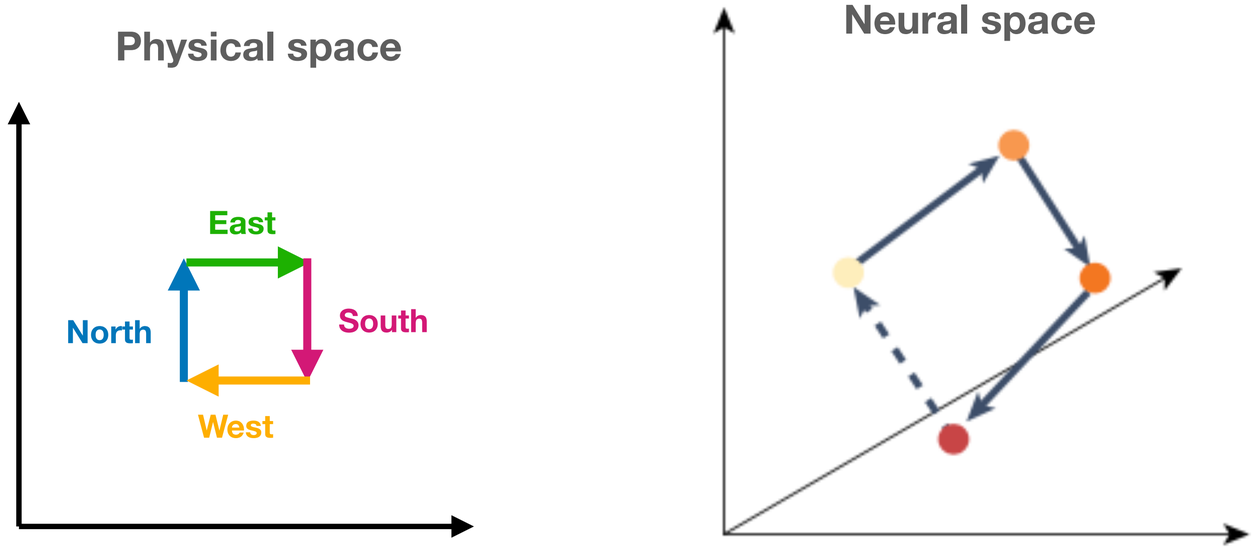
Recurrent neural networks learn from sequences



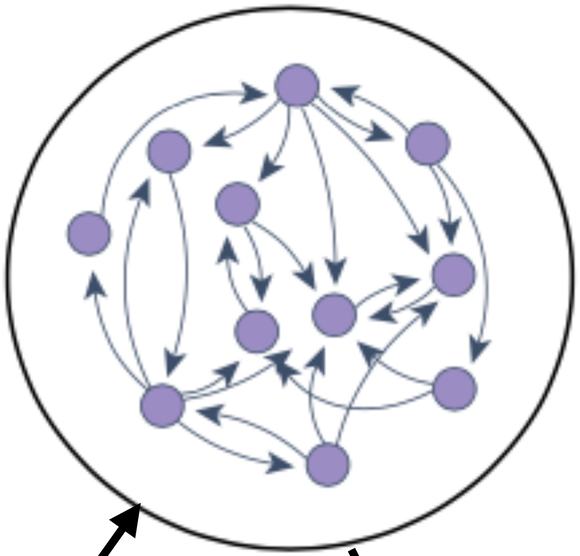




Recurrent neural networks learn from sequences

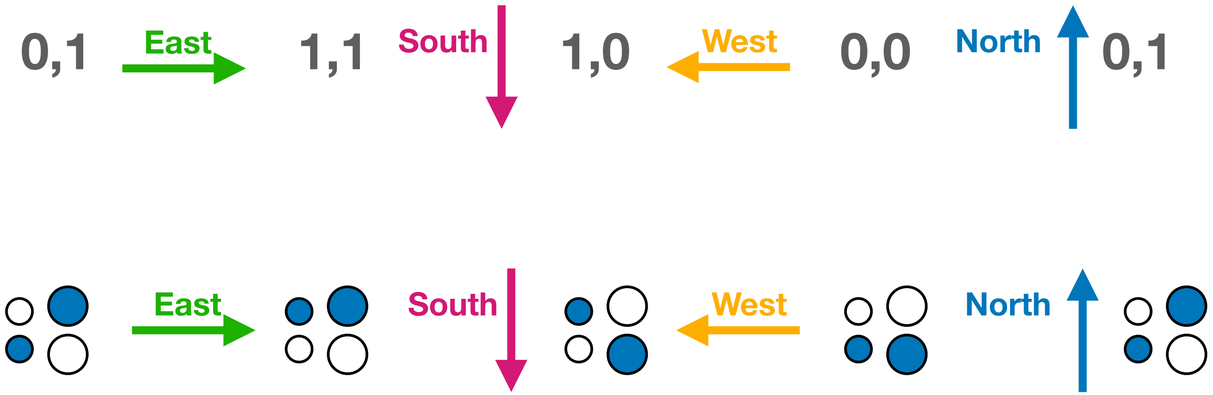


$$\frac{d\bar{s}}{dt} = -\bar{s} + f(W\bar{s} + B\bar{a})$$



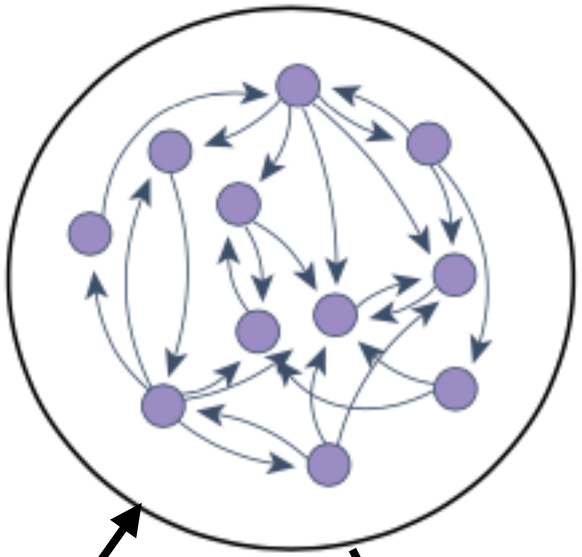
Input:
Actions/velocities

Output:
Where in space, e.g., 1,0



These networks learn grid cells

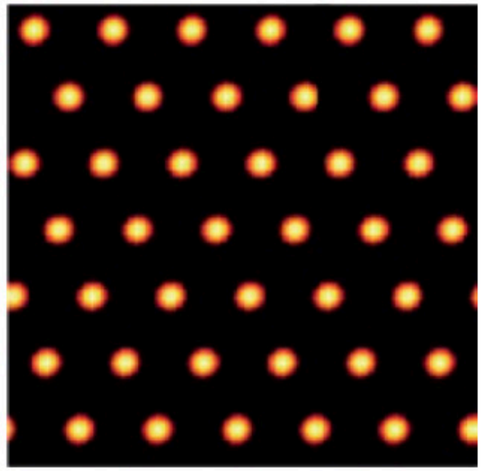
$$\frac{d\bar{s}}{dt} = -\bar{s} + f(W\bar{s} + B\bar{a})$$



Input:
Actions/velocities

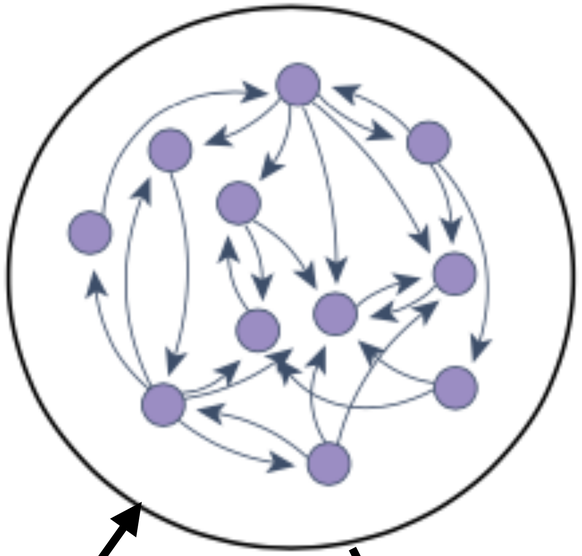
Output:
Where in space, e.g., 1,0

These networks learn grid cells



Burak & Feite (2009)

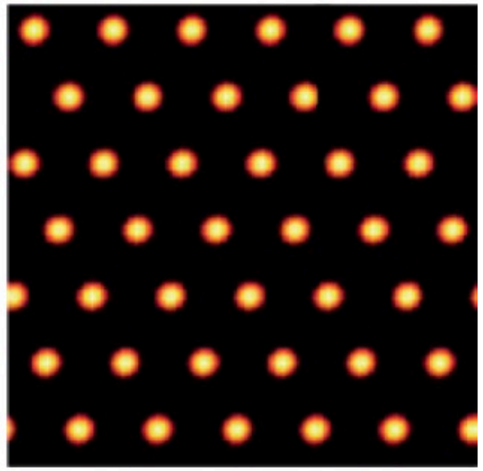
$$\frac{d\bar{s}}{dt} = -\bar{s} + f(W\bar{s} + B\bar{a})$$



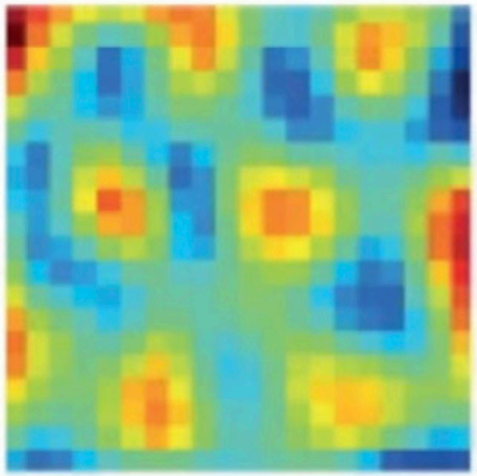
Input:
Actions/velocities

Output:
Where in space, e.g., 1,0

These networks learn grid cells

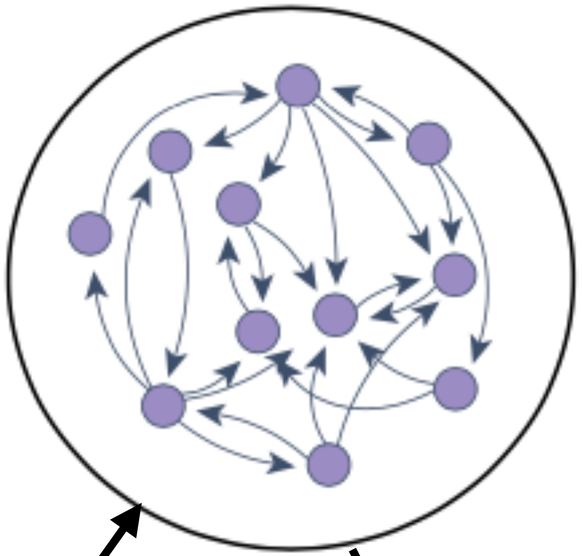


Burak & Feite (2009)



Banino et al., (2018)

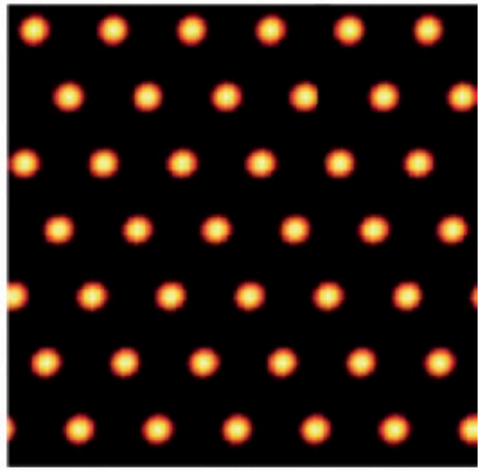
$$\frac{d\bar{s}}{dt} = -\bar{s} + f(W\bar{s} + B\bar{a})$$



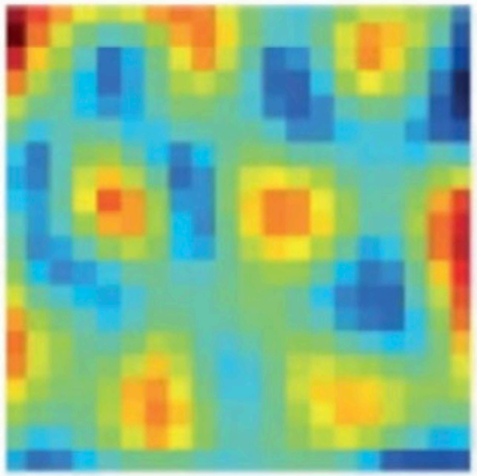
Input:
Actions/velocities

Output:
Where in space, e.g., 1,0

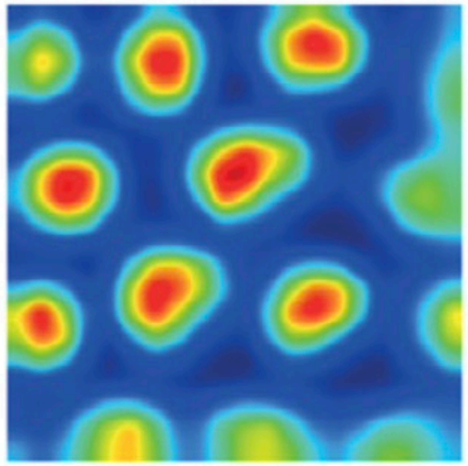
These networks learn grid cells



Burak & Feite (2009)

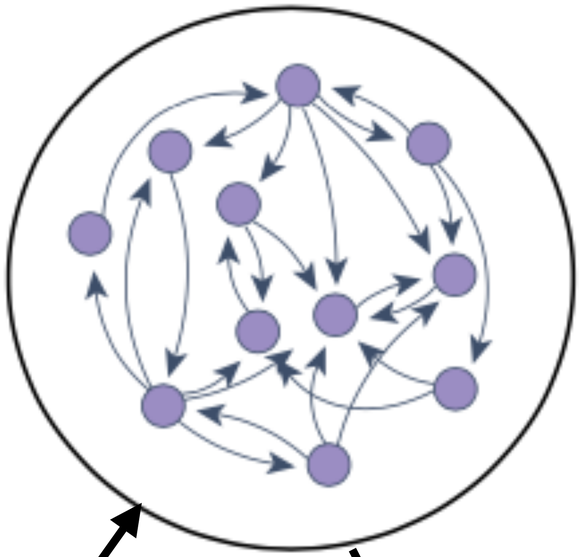


Banino et al., (2018)



Sorscher et al., (2023)

$$\frac{d\bar{s}}{dt} = -\bar{s} + f(W\bar{s} + B\bar{a})$$



Input:
Actions/velocities

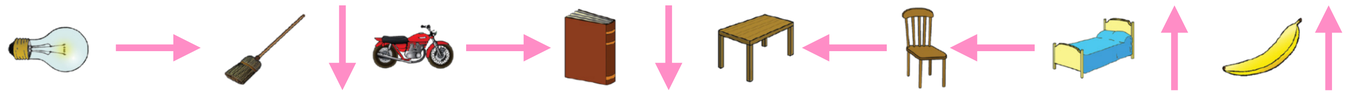
Output:
Where in space, e.g., 1,0

But these sequences are not what the brain sees

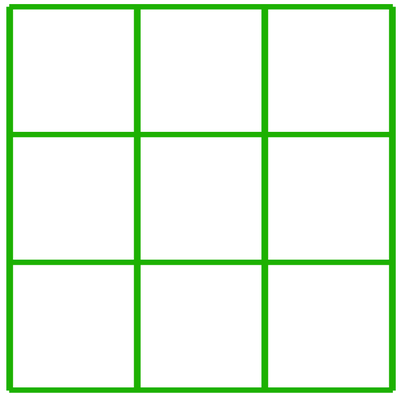
The brain sees sequences of external observations

The brain sees sequences of external observations

Given a stream of sensory stimuli and relations

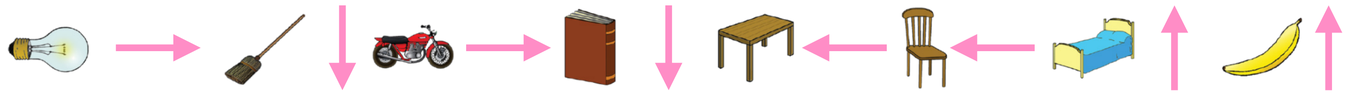


?

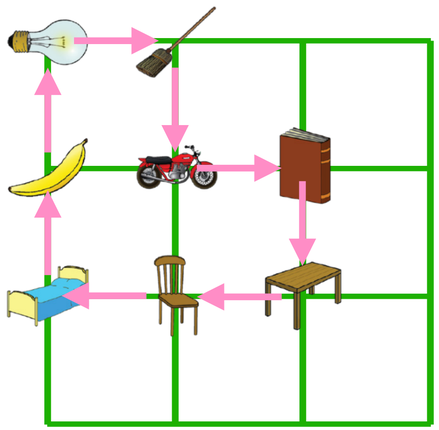


The brain sees sequences of external observations

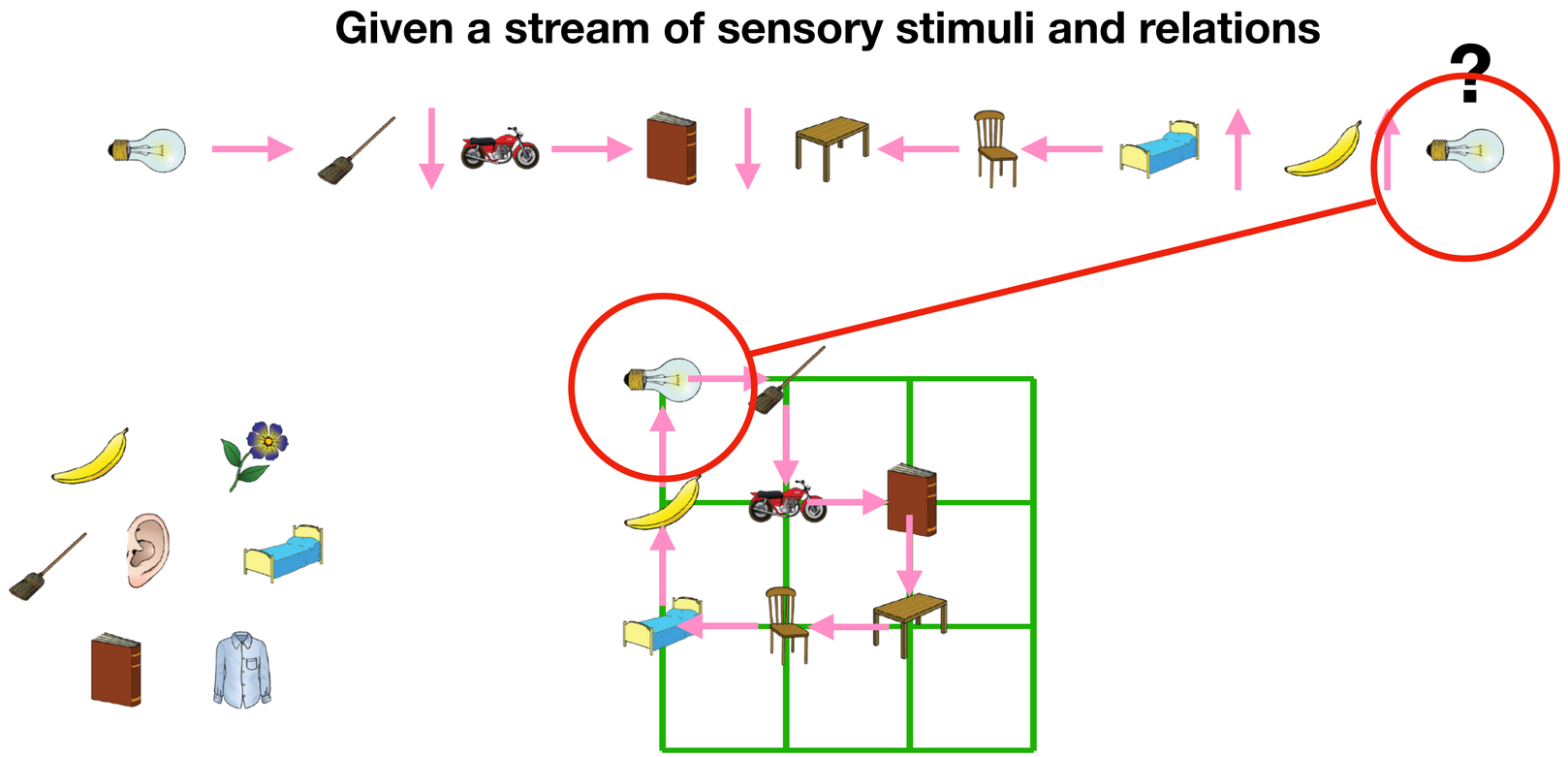
Given a stream of sensory stimuli and relations



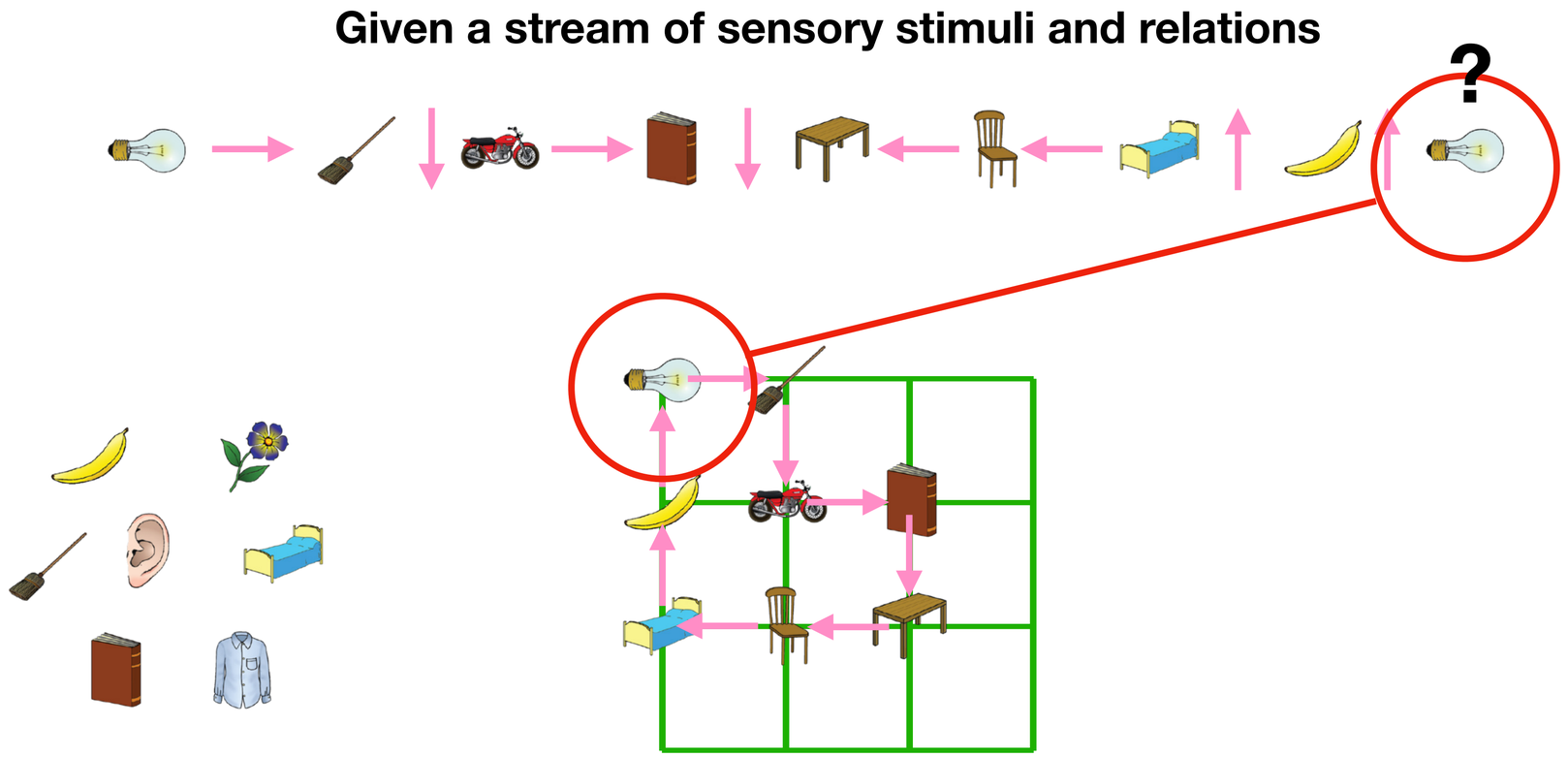
?



The brain sees sequences of external observations

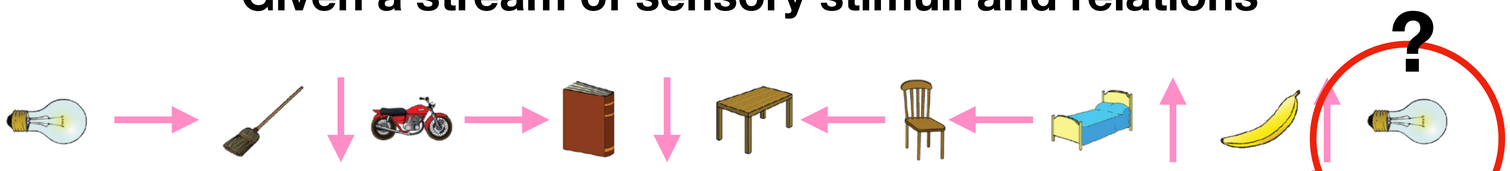


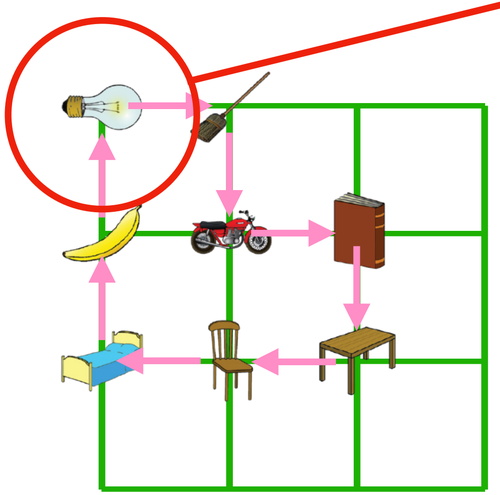
The brain sees sequences of external observations

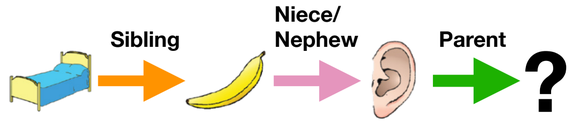


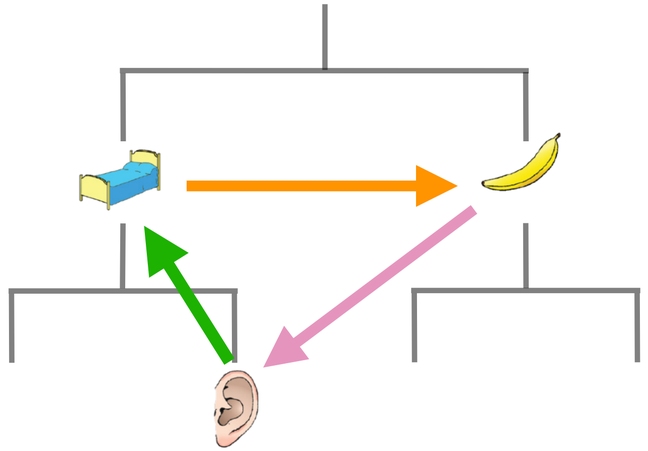
The brain sees sequences of external observations

Given a stream of sensory stimuli and relations



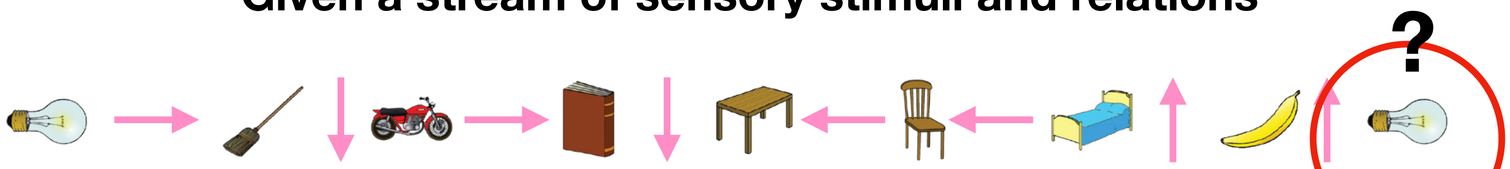


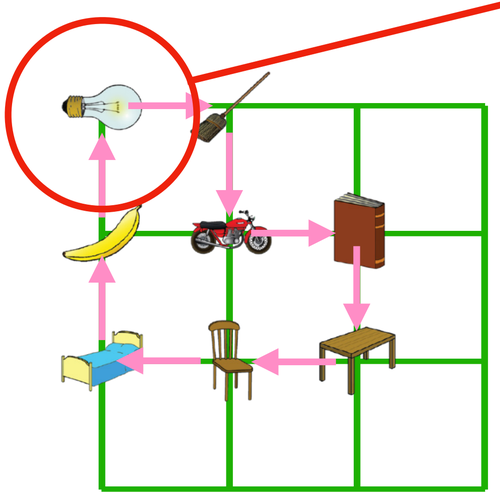




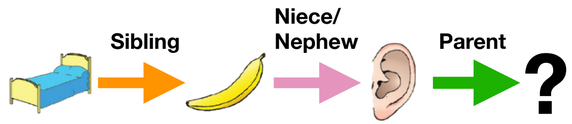
The brain sees sequences of external observations

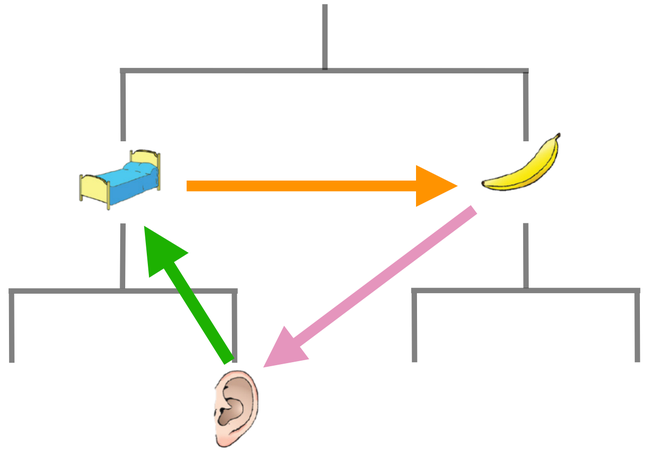
Given a stream of sensory stimuli and relations





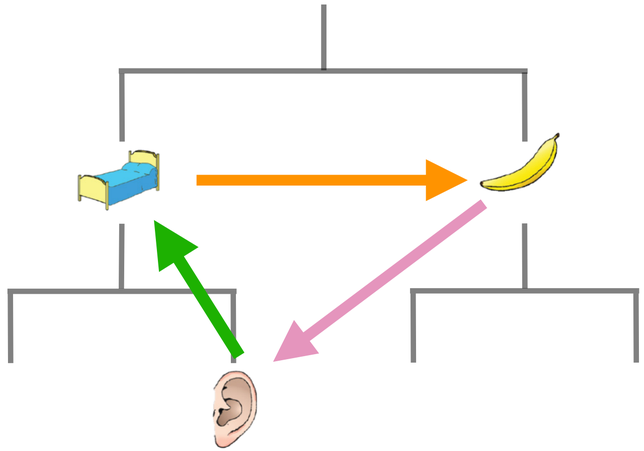
Knowing the map is only way to correctly predict these shortcut inferences zero shot



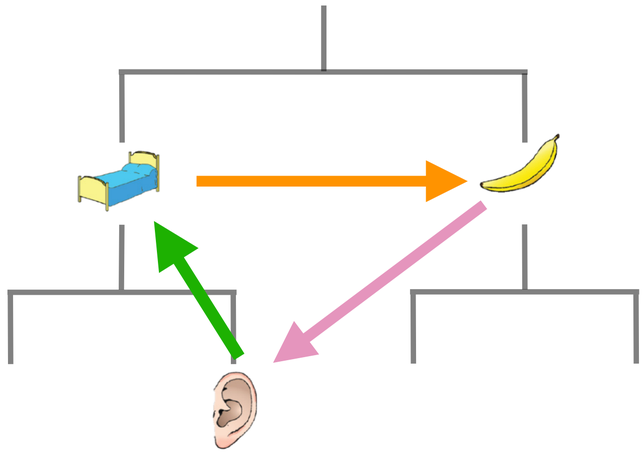


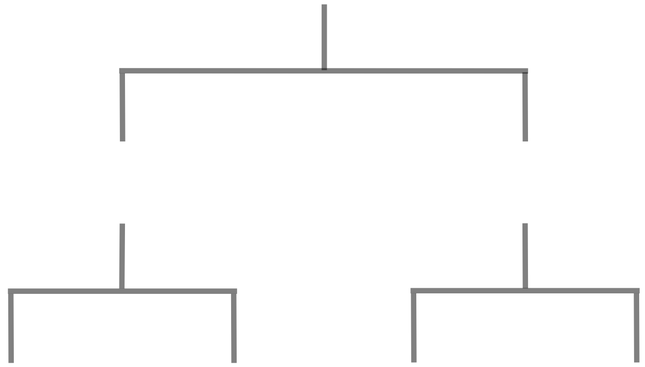
Representing abstract location and making conjunctive memories

Representing abstract location and making conjunctive memories

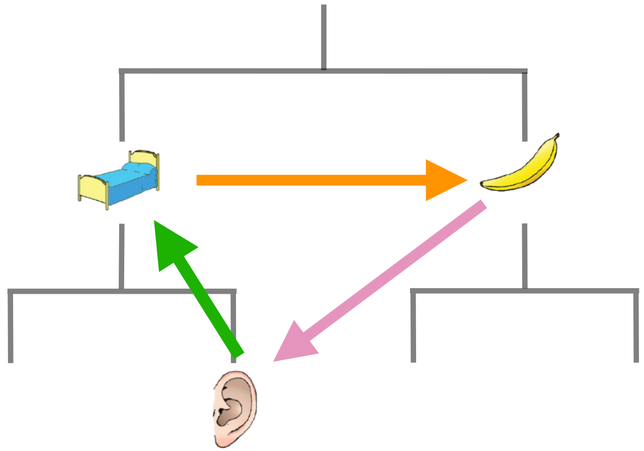


Representing abstract location and making conjunctive memories

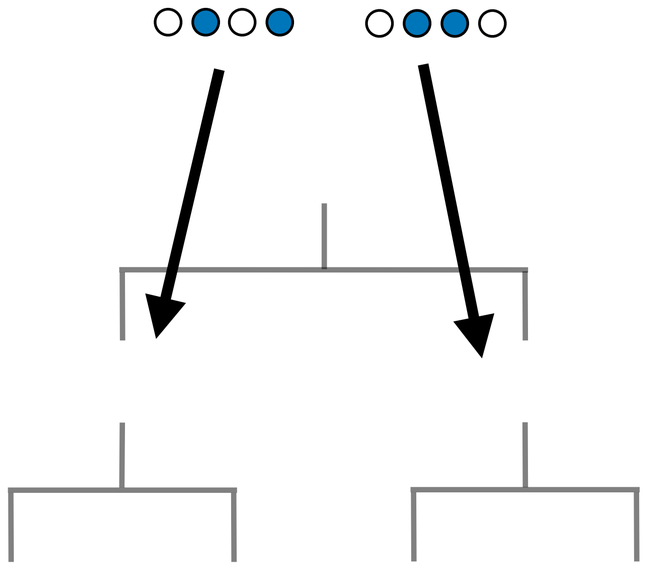




Representing abstract location and making conjunctive memories



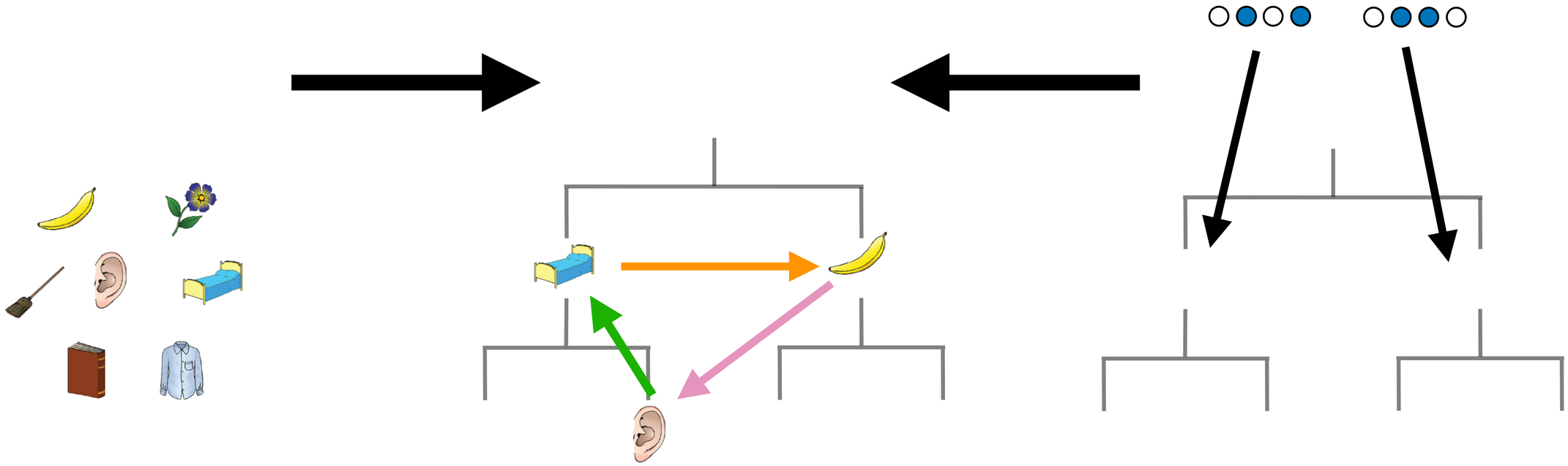
Represent **location** in
abstract structure



Representing abstract location and making conjunctive memories

Compose building blocks and form **memories** to know who is where in each family

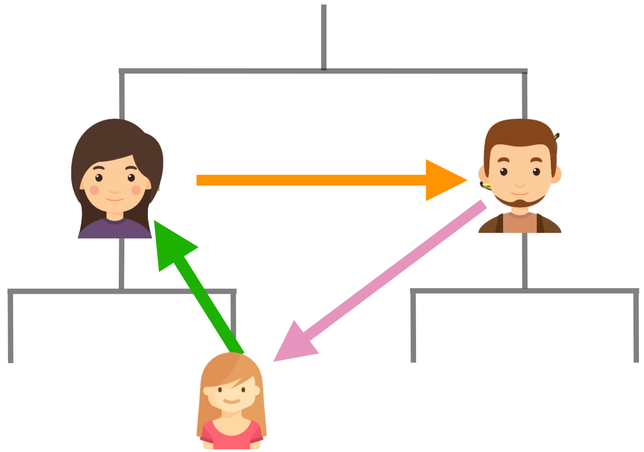
Represent **location** in abstract structure

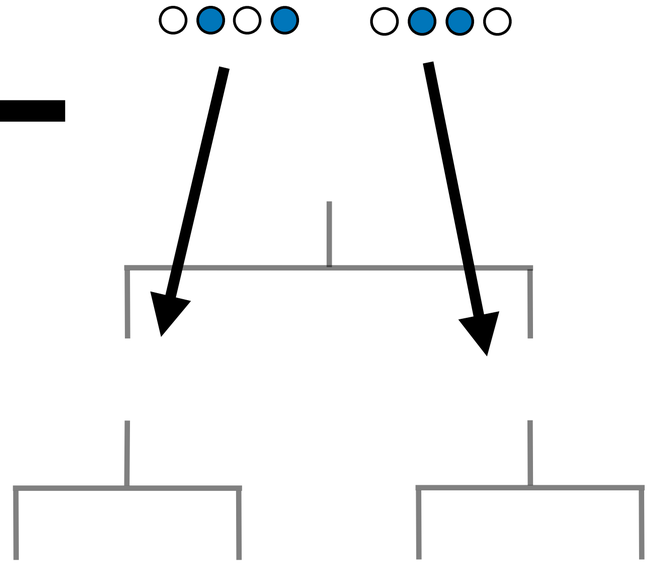


Representing abstract location and making conjunctive memories

Compose building blocks and form **memories** to know who is where in each family

Represent **location** in abstract structure

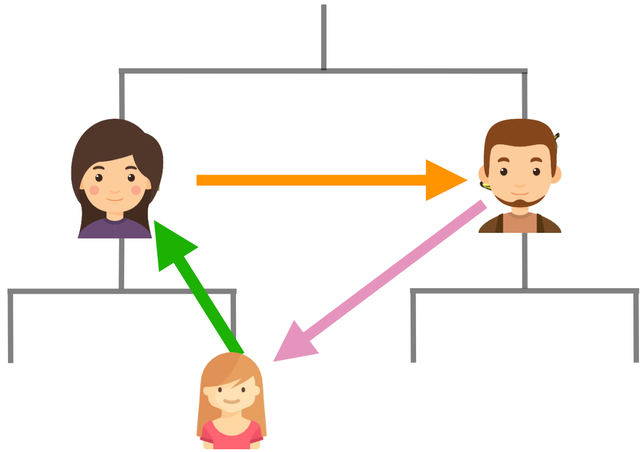


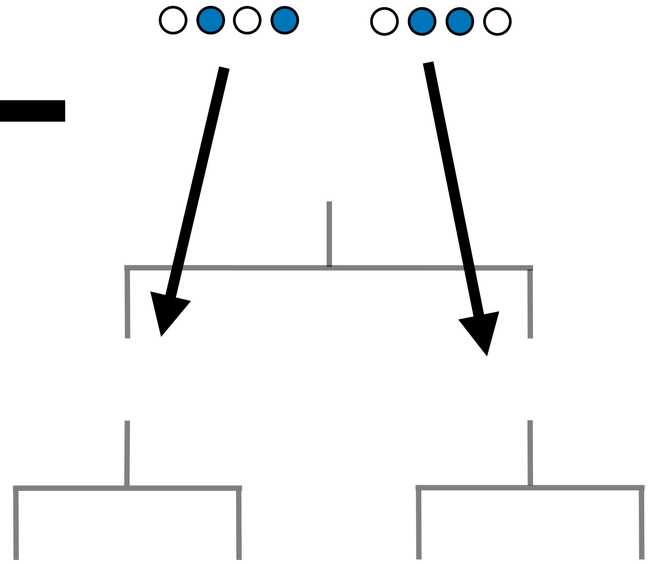


Representing abstract location and making conjunctive memories

Compose building blocks and form **memories** to know who is where in each family

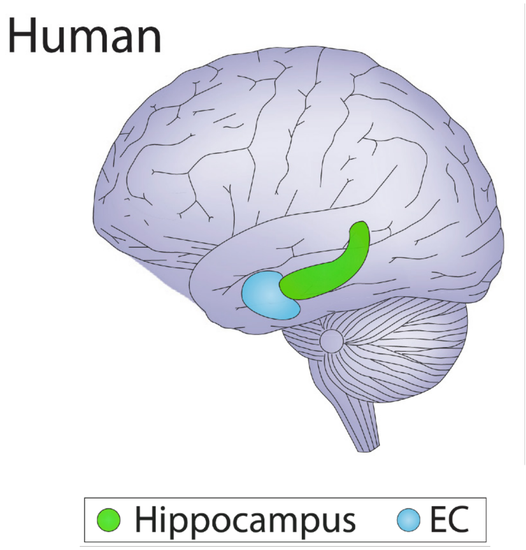
Represent **location** in abstract structure



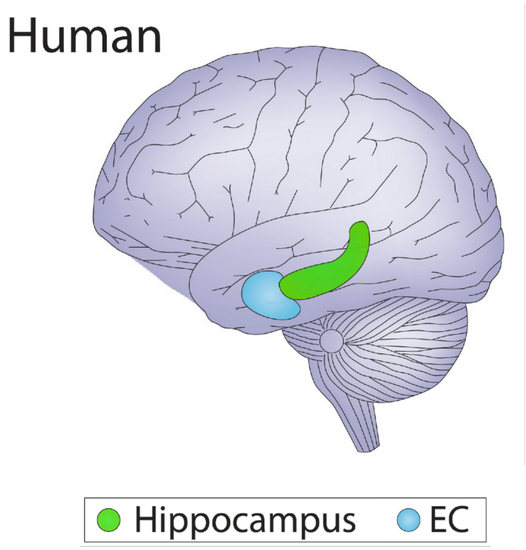


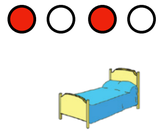
How do we do this and what on earth will the representations look like?

TEM learns abstract location representations by prediction errors

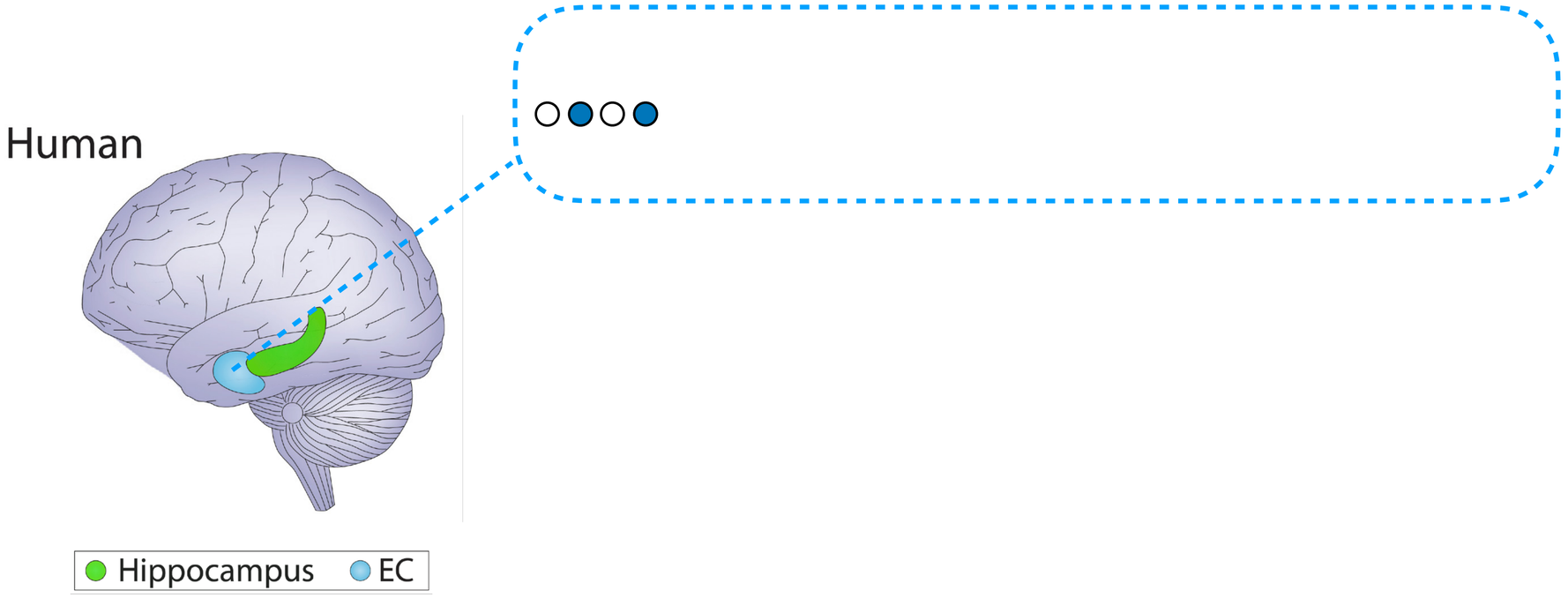


TEM learns abstract location representations by prediction errors

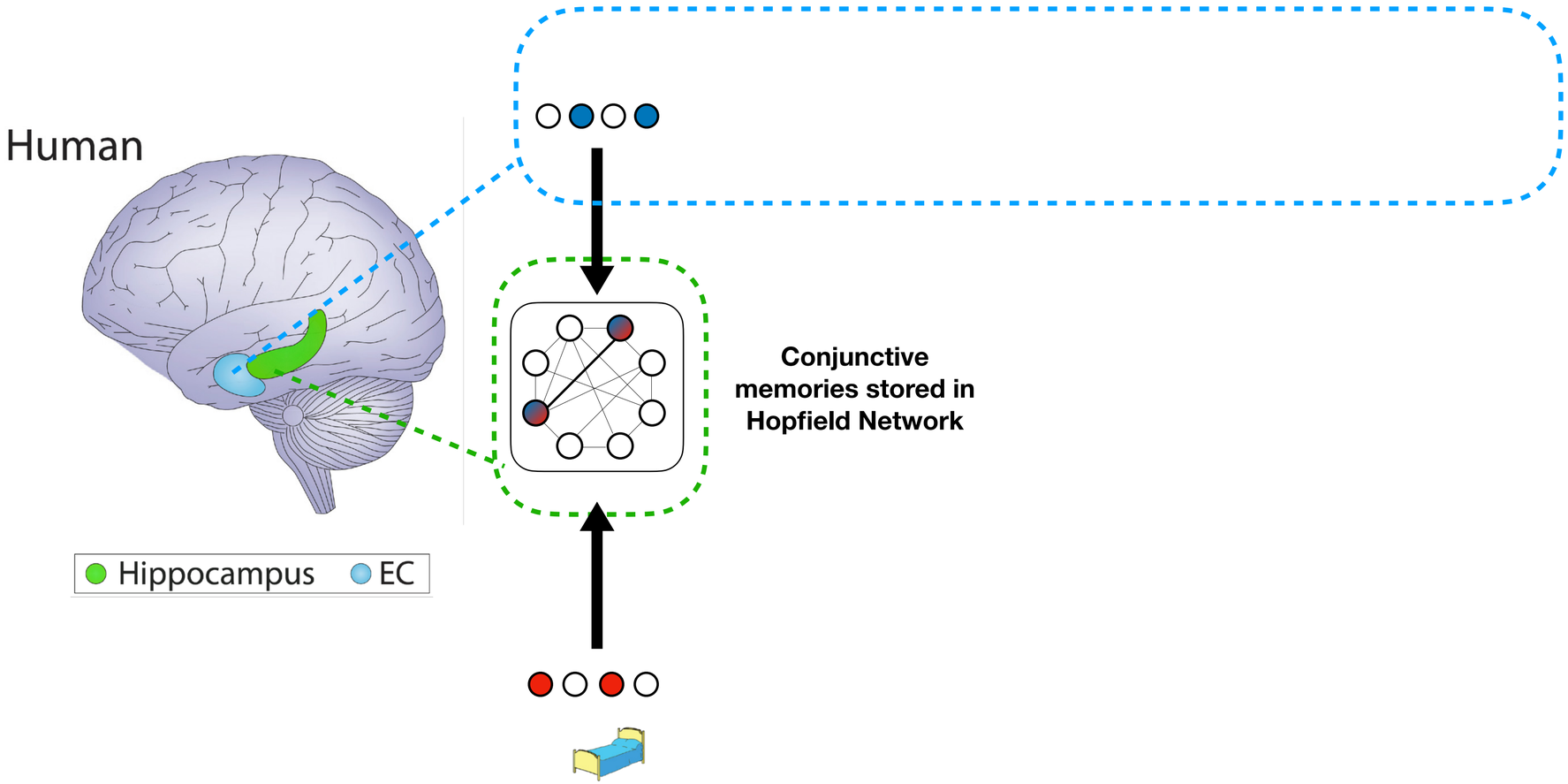




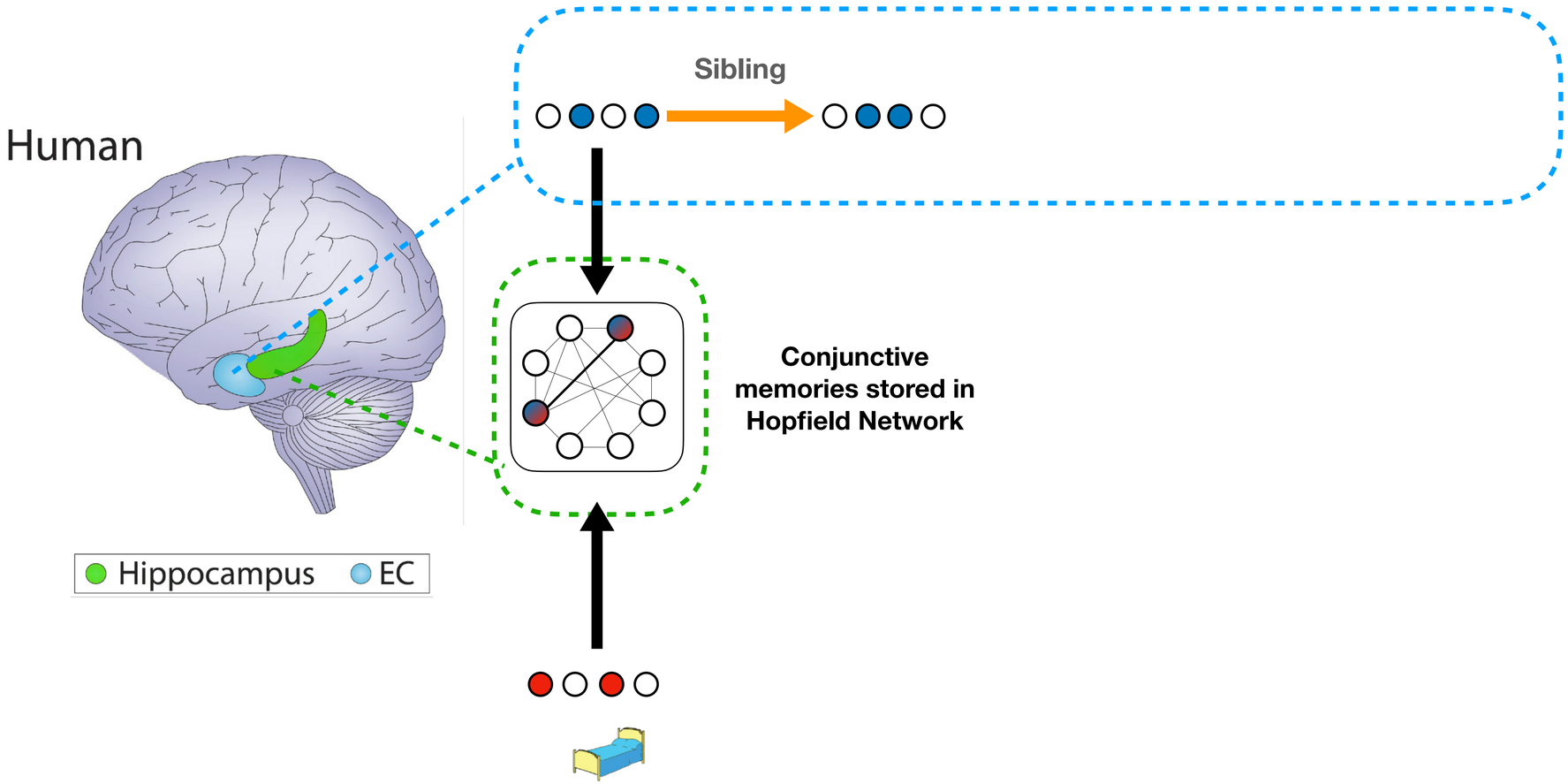
TEM learns abstract location representations by prediction errors



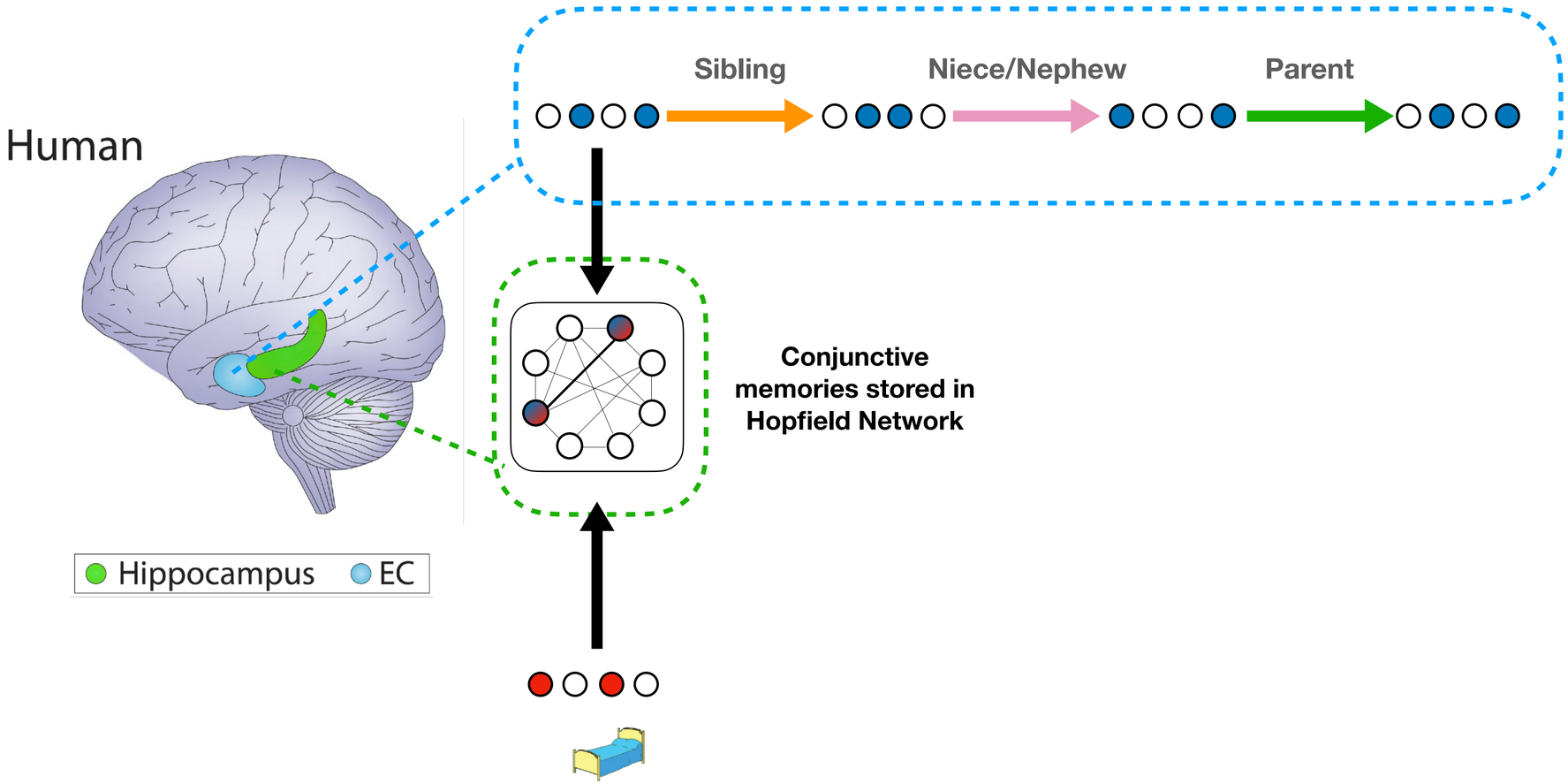
TEM learns abstract location representations by prediction errors



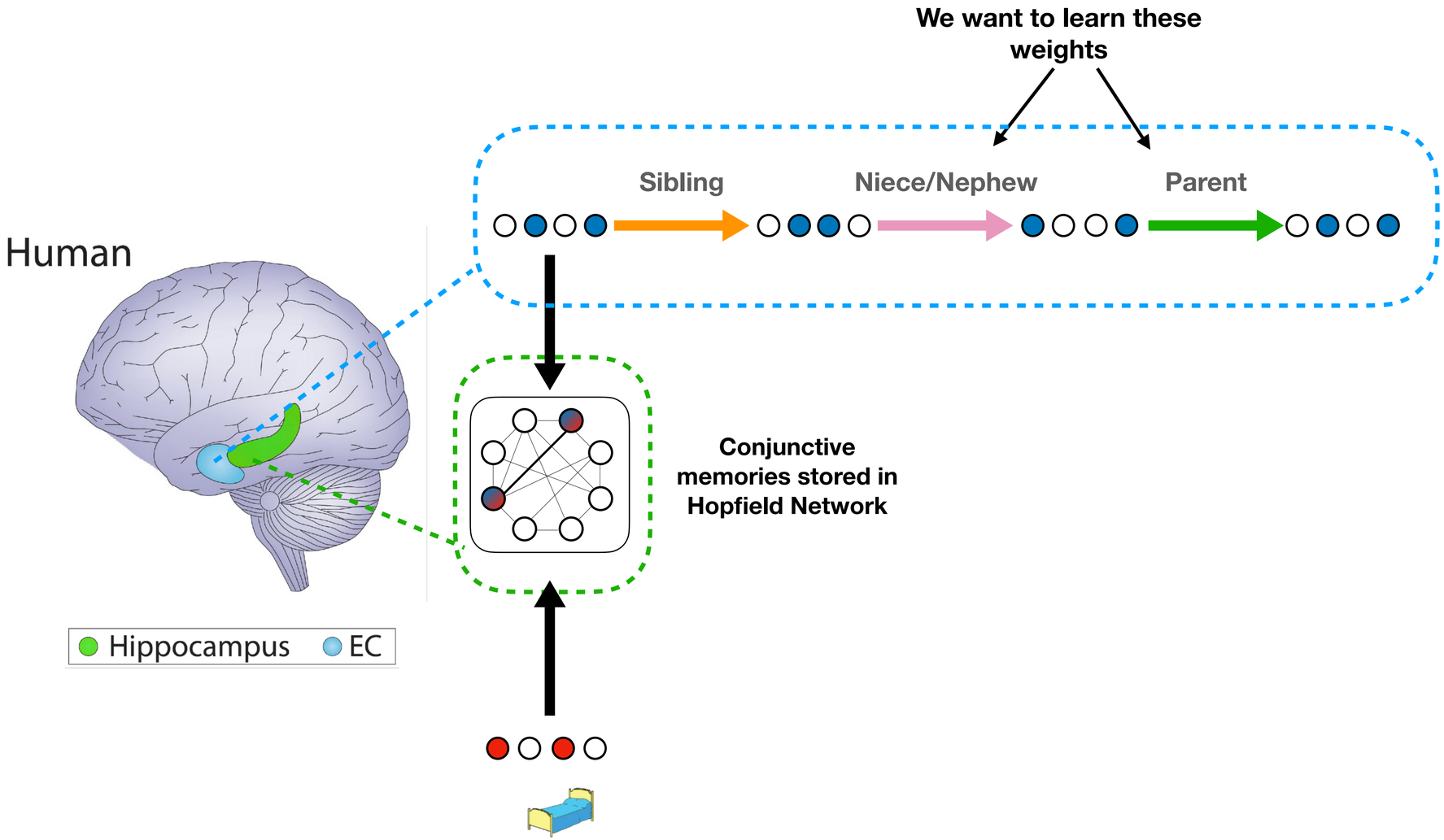
TEM learns abstract location representations by prediction errors



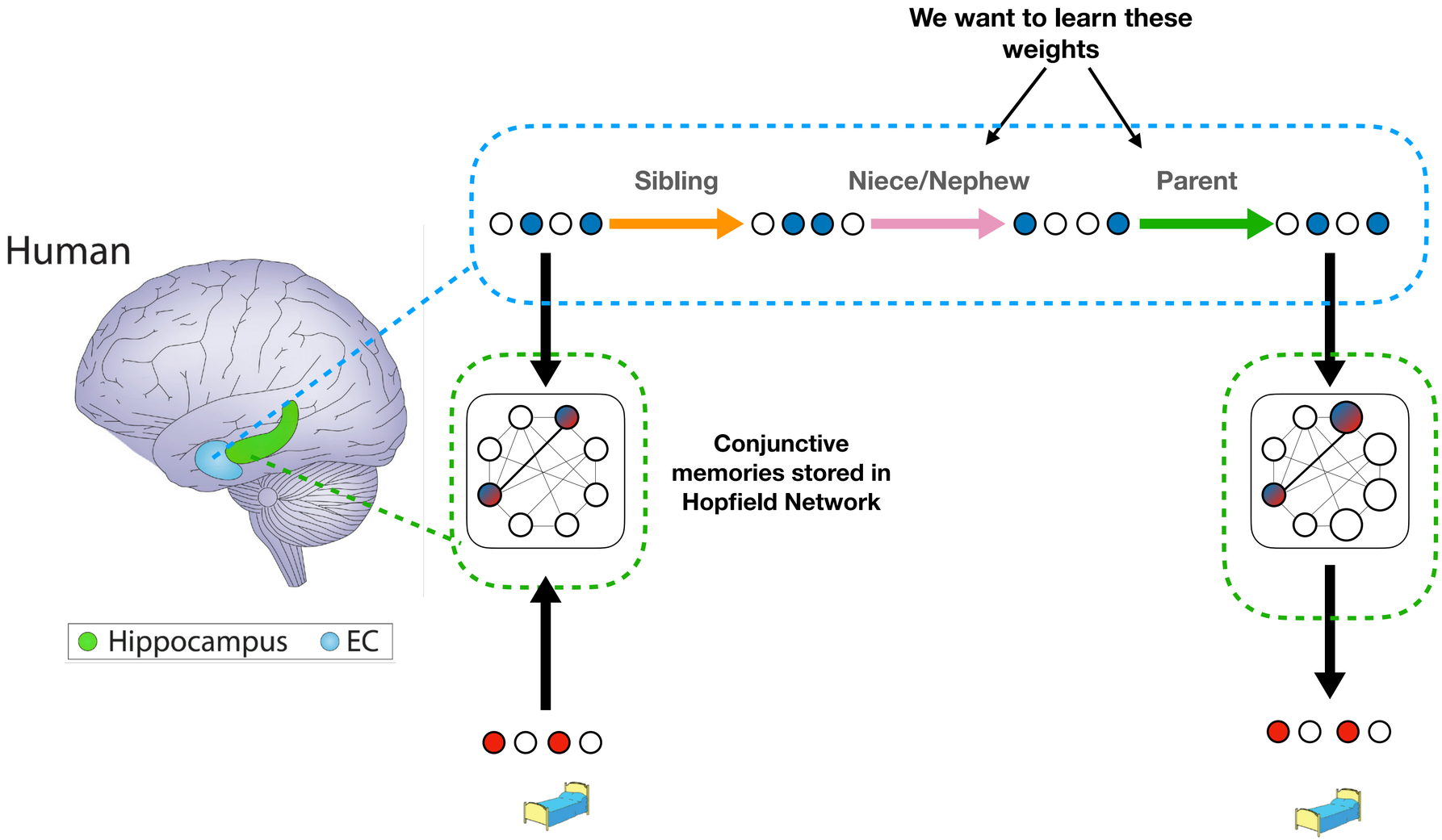
TEM learns abstract location representations by prediction errors



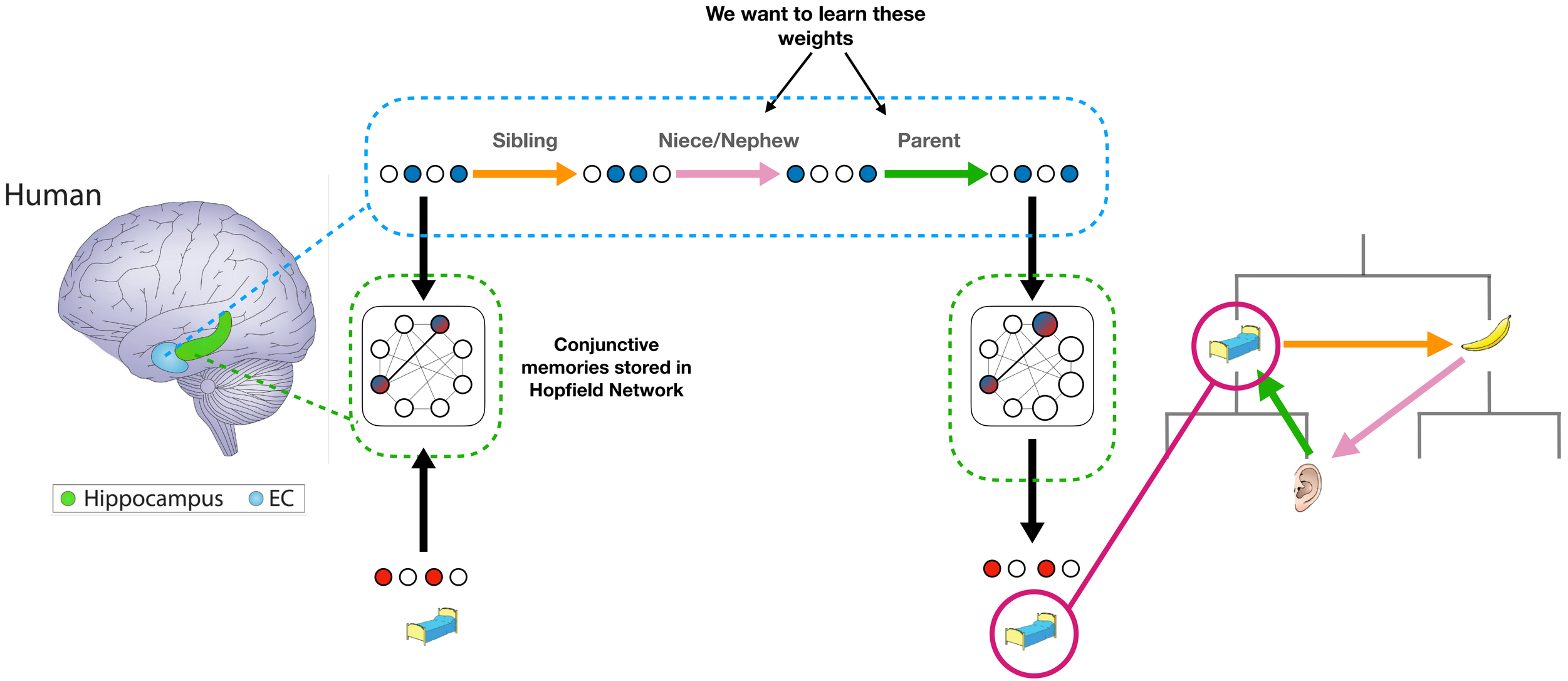
TEM learns abstract location representations by prediction errors



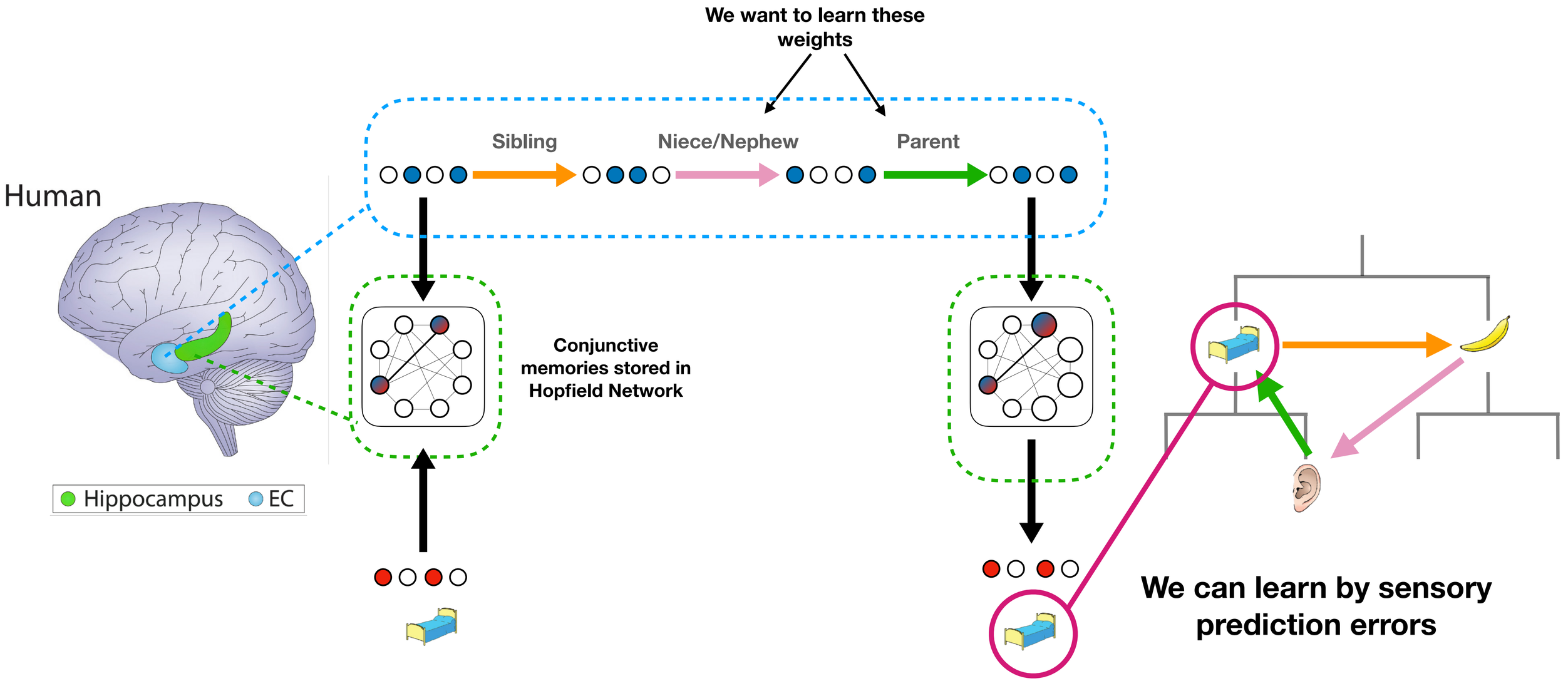
TEM learns abstract location representations by prediction errors



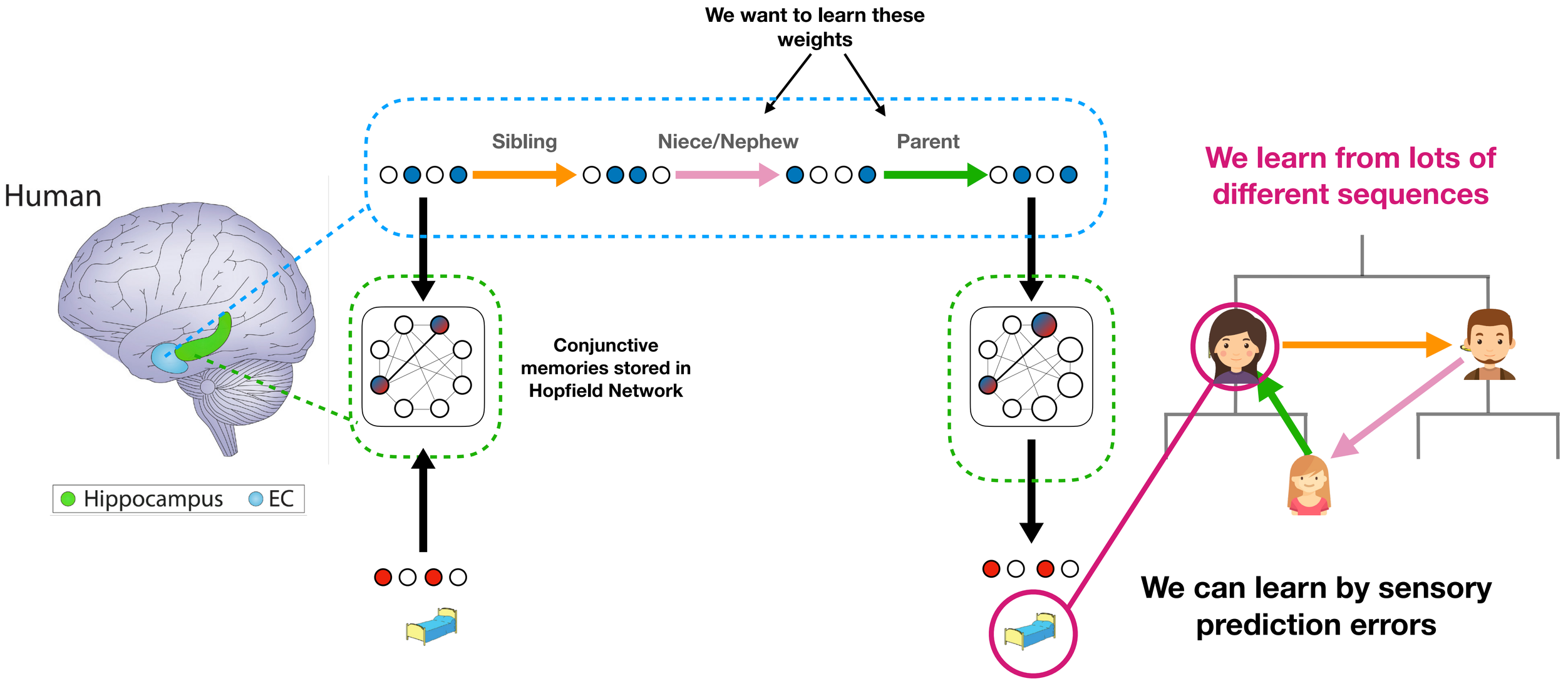
TEM learns abstract location representations by prediction errors



TEM learns abstract location representations by prediction errors



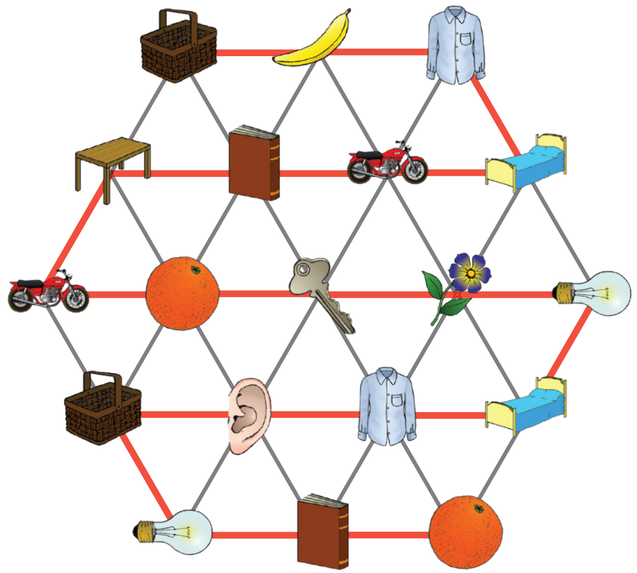
TEM learns abstract location representations by prediction errors



Does it work?

TEM infers relationships when it sees a new observation

Testing on a novel spatial environment

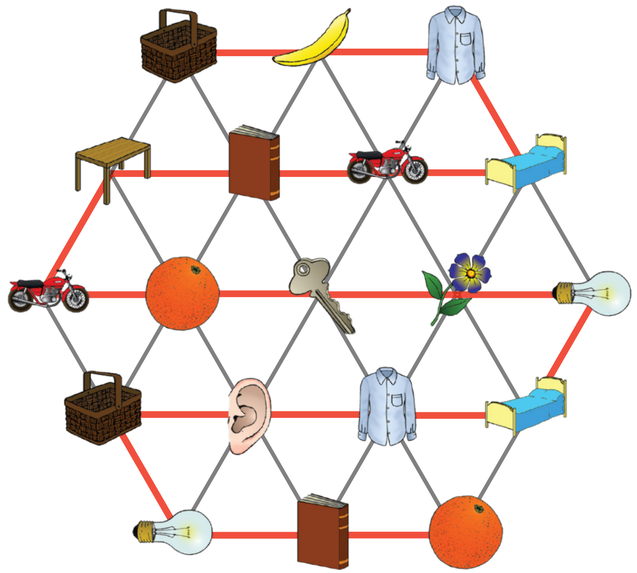


Performance in new worlds should depend on number of **nodes** it has seen, not number of **edges**

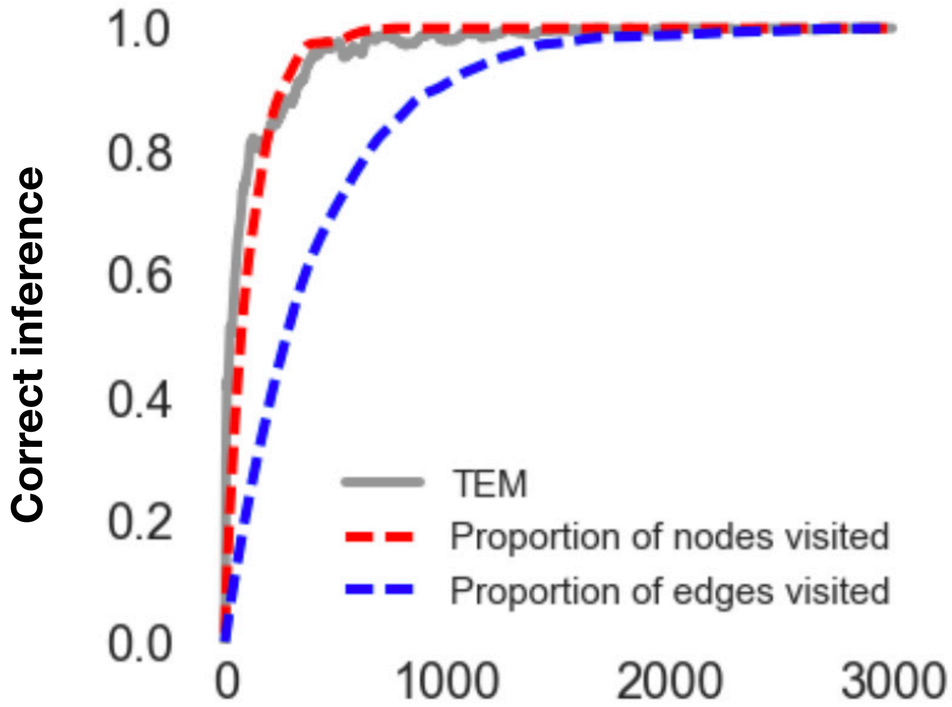
Whittington et al., 2020, *Cell*

TEM infers relationships when it sees a new observation

Testing on a novel spatial environment

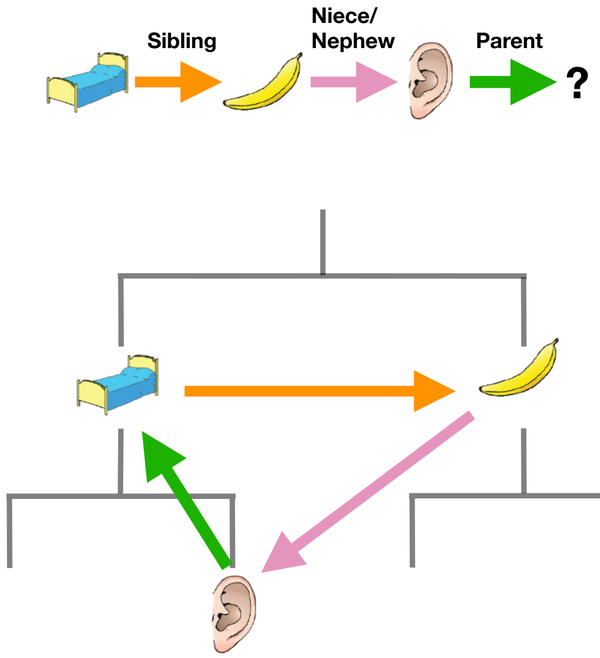


Performance in new worlds should depend on number of **nodes** it has seen, not number of **edges**



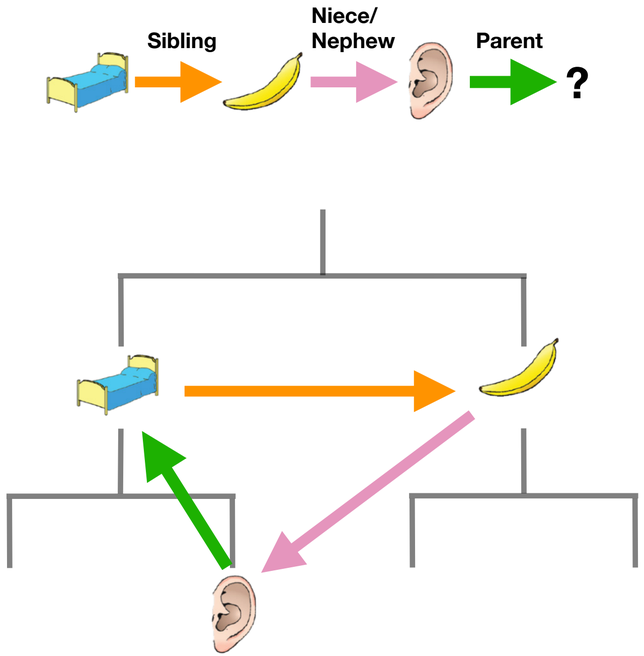
Whittington et al., 2020, *Cell*

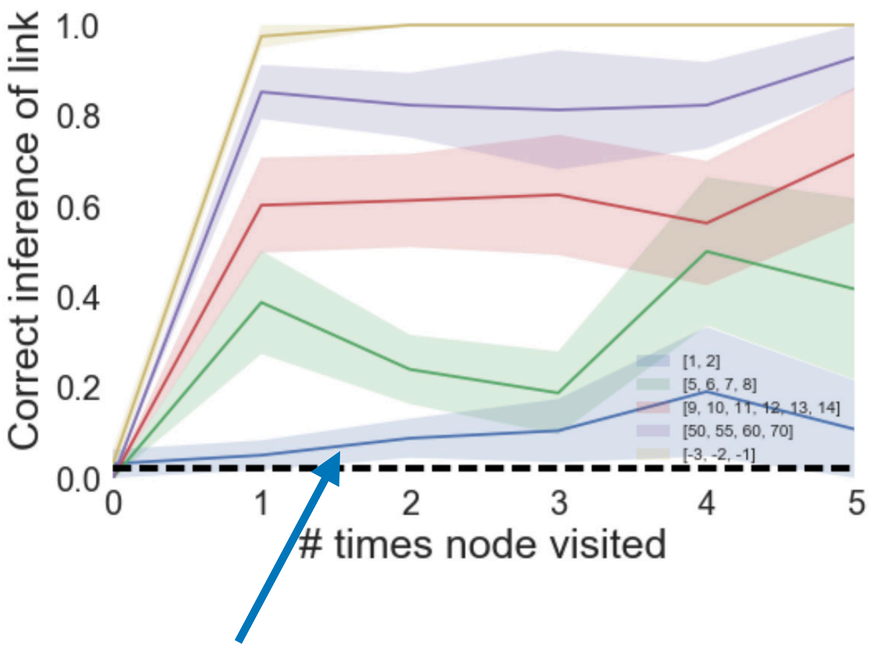
TEM makes inferences because it has learned in similar structures



Whittington et al., 2020, *Cell*

TEM makes inferences because it has learned in similar structures

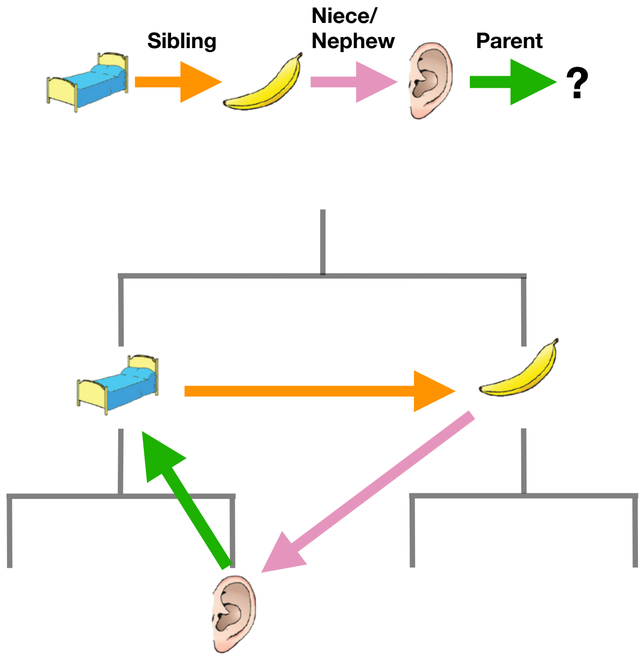


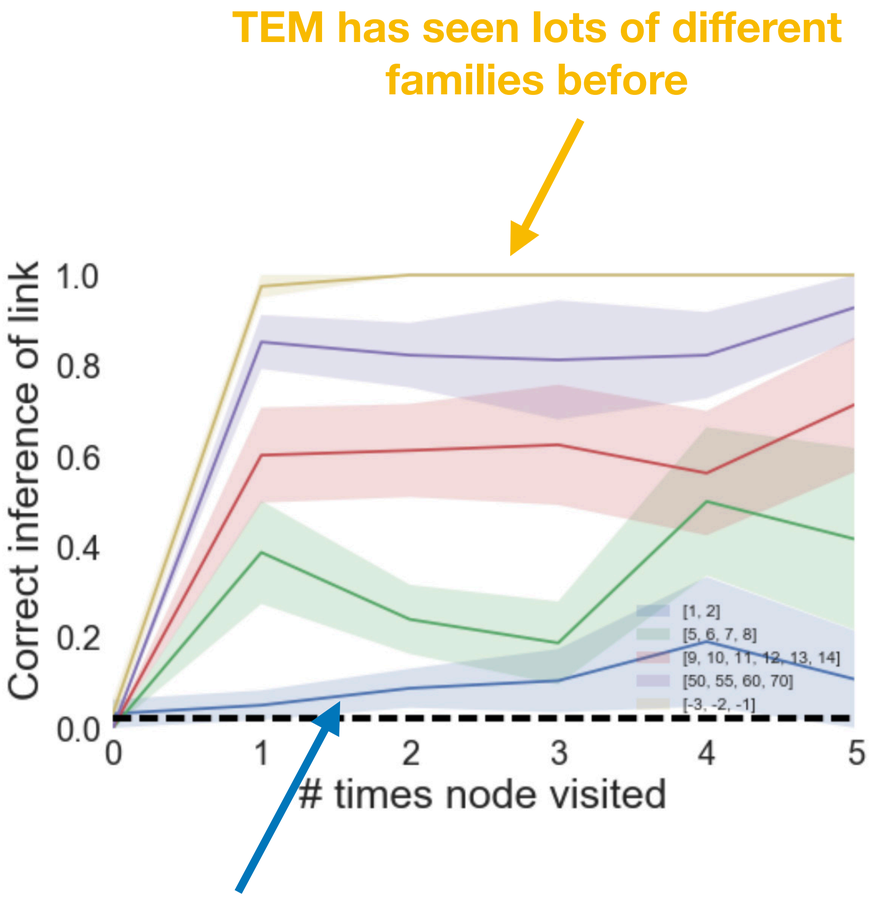


First time TEM ever saw a family

Whittington et al., 2020, *Cell*

TEM makes inferences because it has learned in similar structures





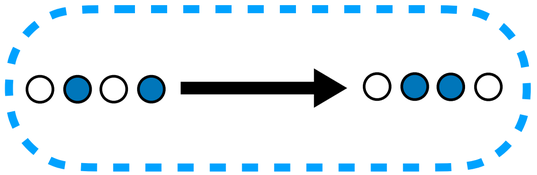
Whittington et al., 2020, *Cell*

Learned building block representations

Whittington et al., 2020, *Cell*

Learned building block representations

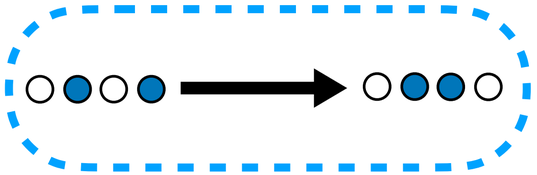
Path integrating RNN

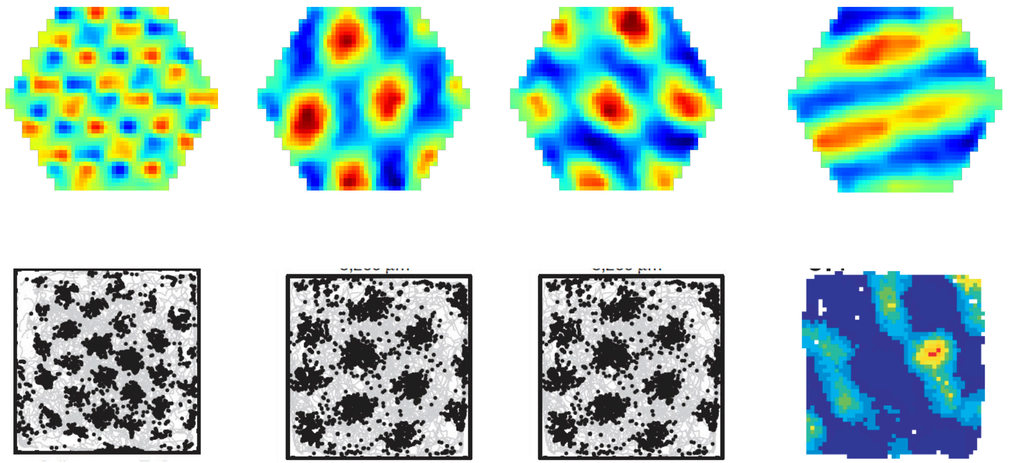


Whittington et al., 2020, *Cell*

Learned building block representations

Path integrating RNN





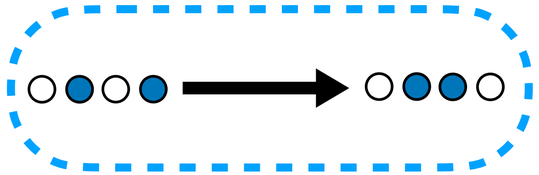
Grid cells
Hafting et al., 2005

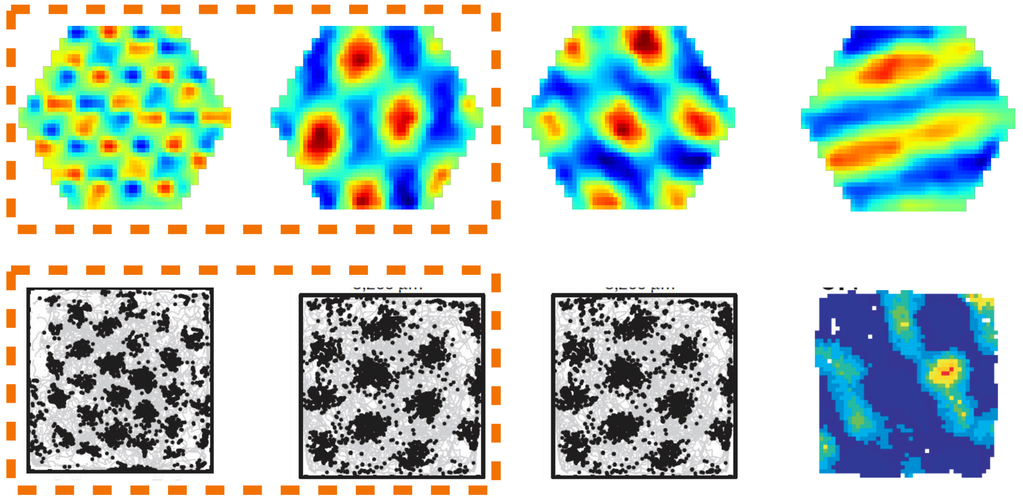
Band cells
Krupic et al., 2012

Whittington et al., 2020, *Cell*

Learned building block representations

Path integrating RNN





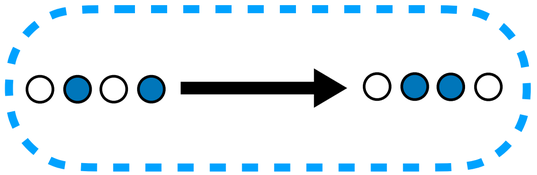
Grid cells
Hafting et al., 2005

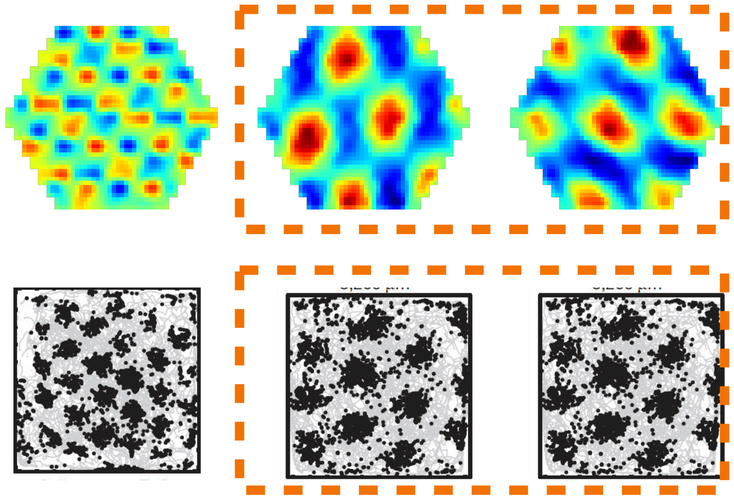
Band cells
Krupic et al., 2012

Whittington et al., 2020, *Cell*

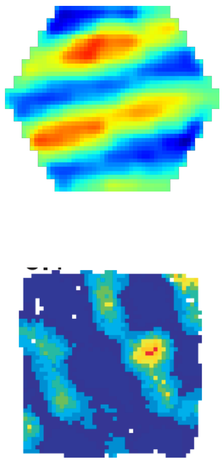
Learned building block representations

Path integrating RNN





Grid cells
Hafting et al., 2005

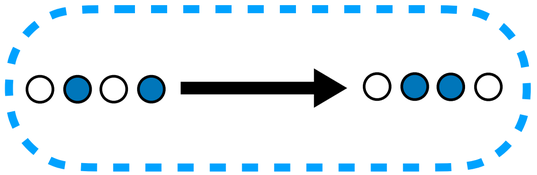


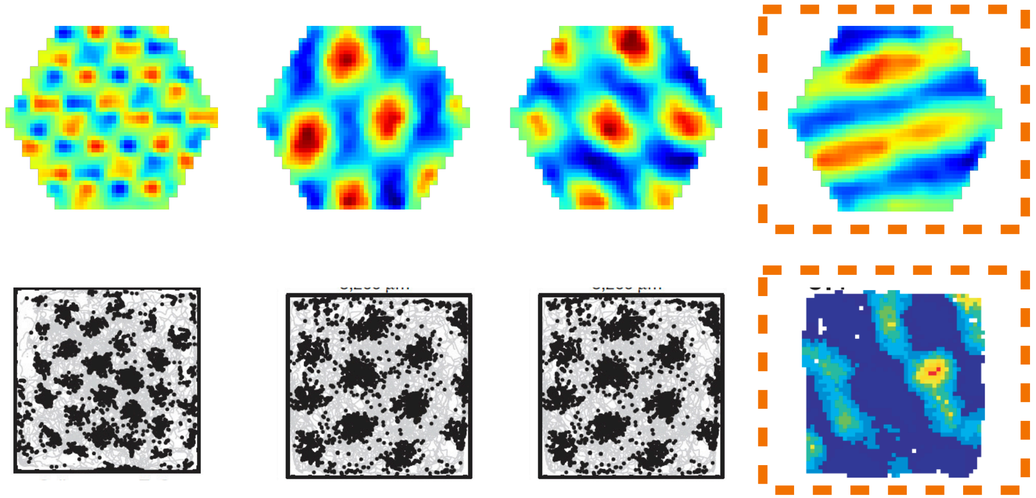
Band cells
Krupic et al., 2012

Whittington et al., 2020, *Cell*

Learned building block representations

Path integrating RNN





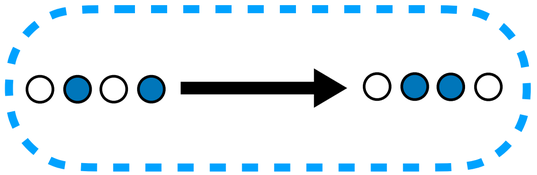
Grid cells
Hafting et al., 2005

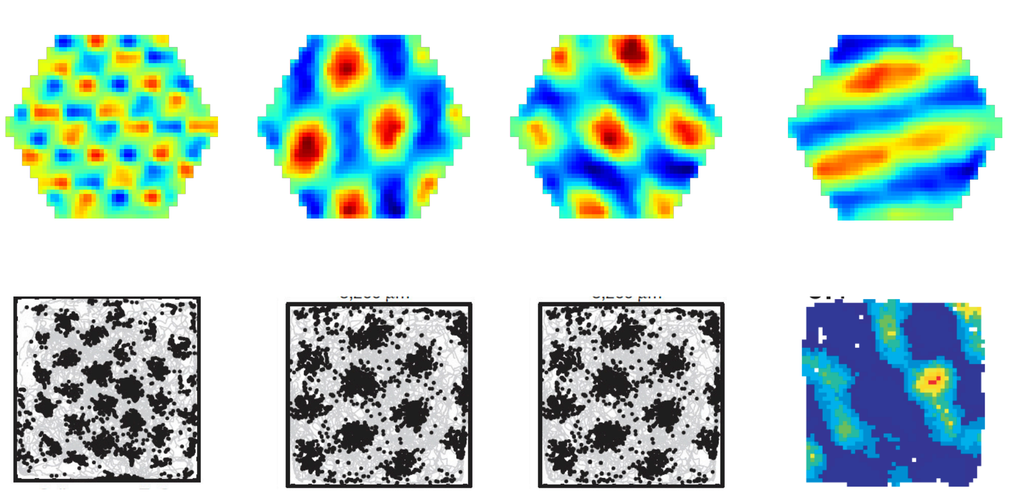
Band cells
Krupic et al., 2012

Whittington et al., 2020, *Cell*

Learned building block representations

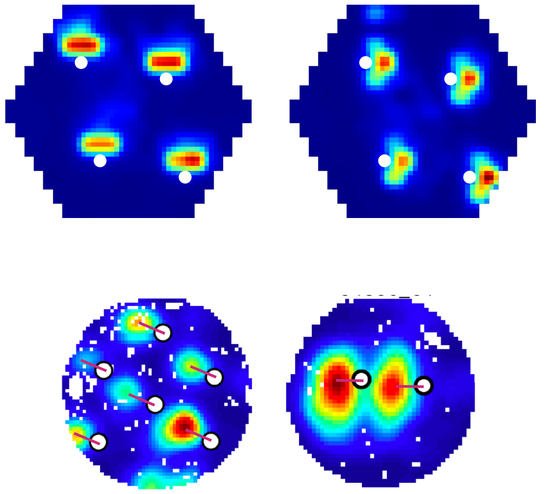
Path integrating RNN





Grid cells
Hafting et al., 2005

Band cells
Krupic et al., 2012

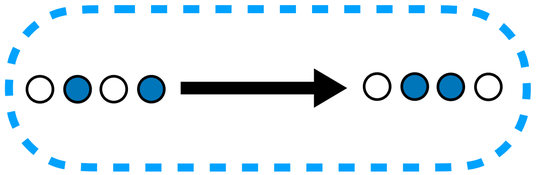


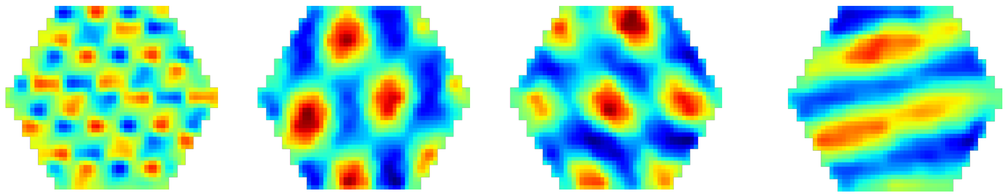
Object vector cells
Hoydal et al., 2018

Whittington et al., 2020, *Cell*

Learned building block representations

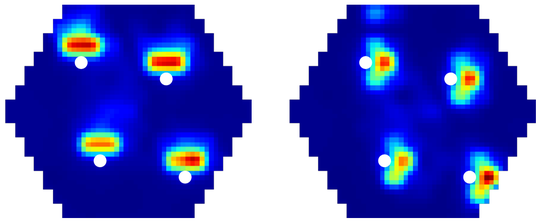
Path integrating RNN



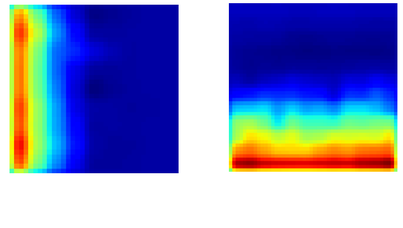


Grid cells
Hafting et al., 2005

Band cells
Krupic et al., 2012

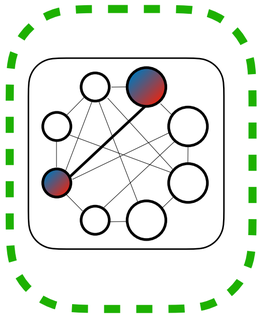


Object vector cells
Hoydal et al., 2018



Border cells
Solstad et al., 2008

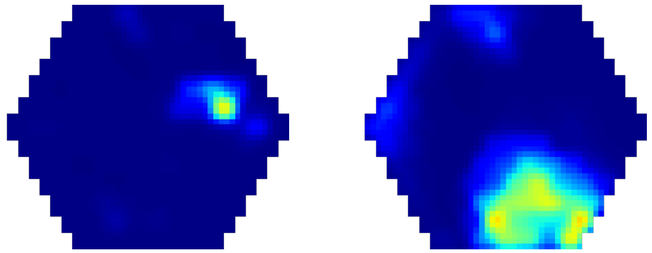
Whittington et al., 2020, *Cell*



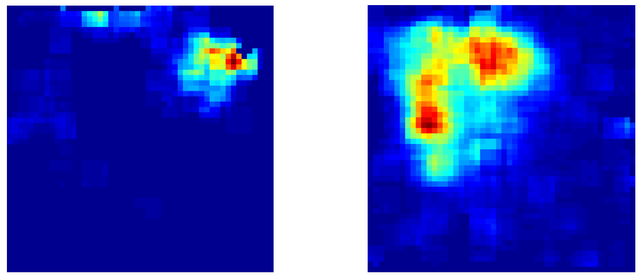
Hopfield
Network

Learned Hippocampal representations

Example TEM cells

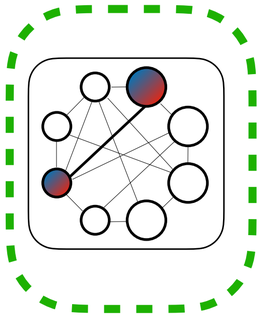


Example real cells



Place cells
O'Keefe & Dostrovsky, 1971

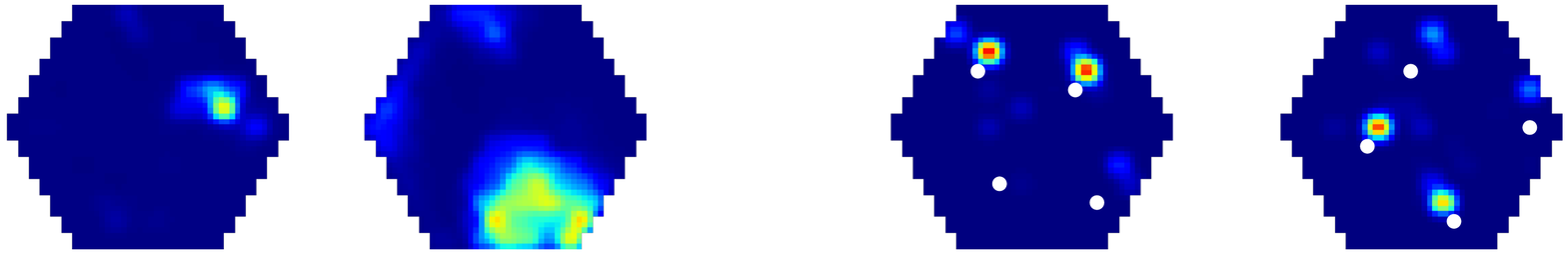
Whittington et al., 2020, *Cell*



Hopfield Network

Learned Hippocampal representations

Example TEM cells



Example real cells

Place cells
O'Keefe & Dostrovsky, 1971

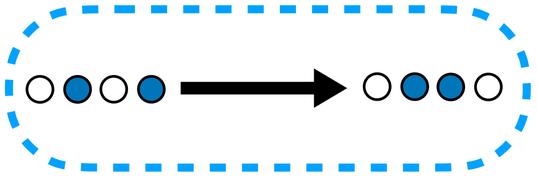
Landmark cells
Deshmukh & Knierim, 2013

Whittington et al., 2020, *Cell*

How do these representations generalise?

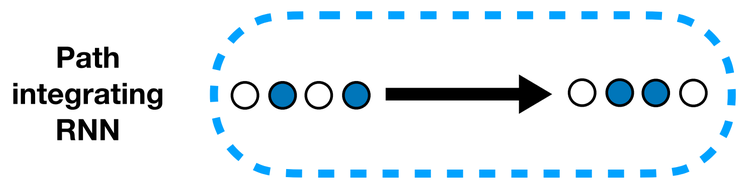
Building block remapping

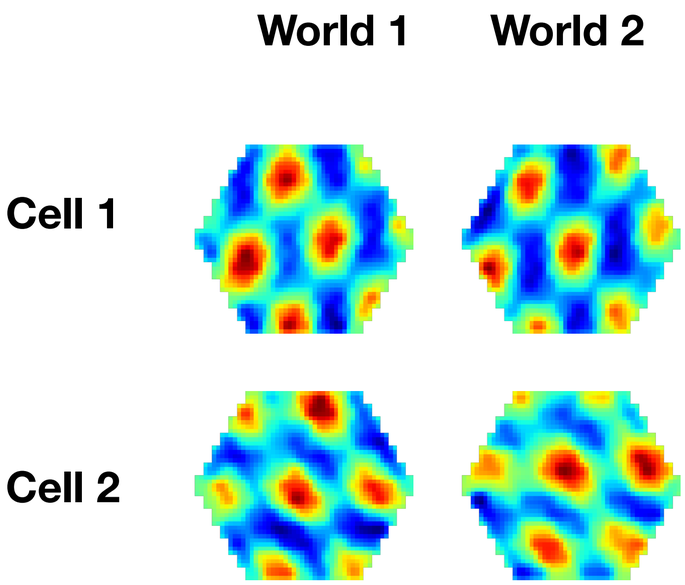
Path integrating RNN



Whittington et al., 2020, *Cell*

Building block remapping

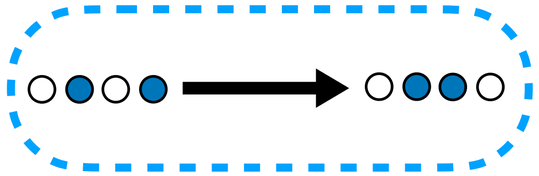


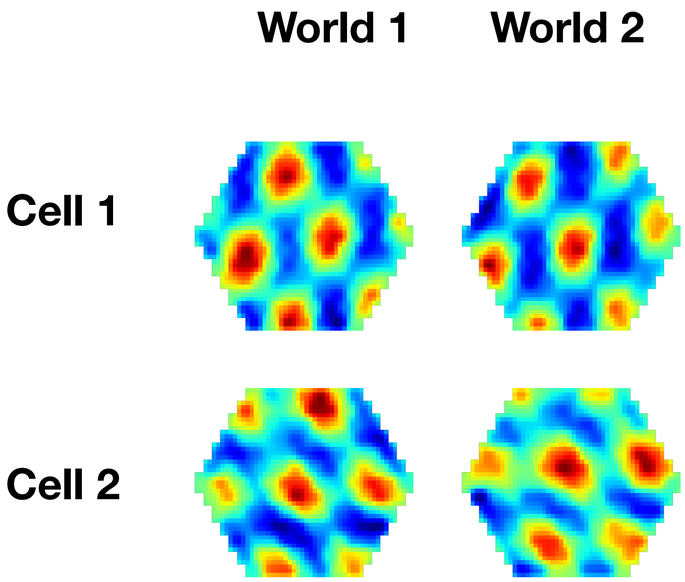


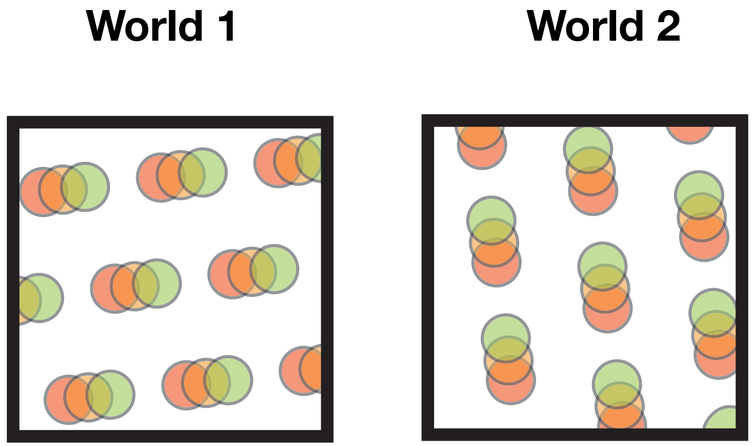
Whittington et al., 2020, *Cell*

Building block remapping

Path integrating RNN

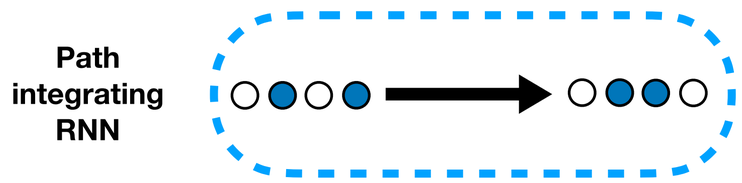


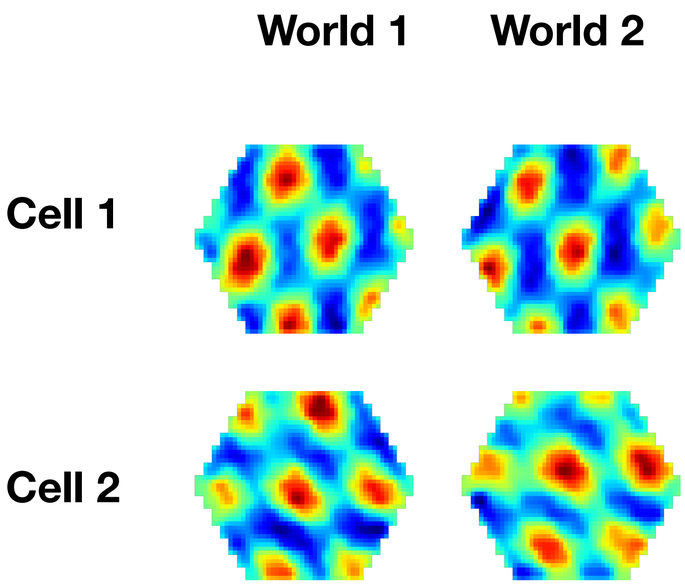


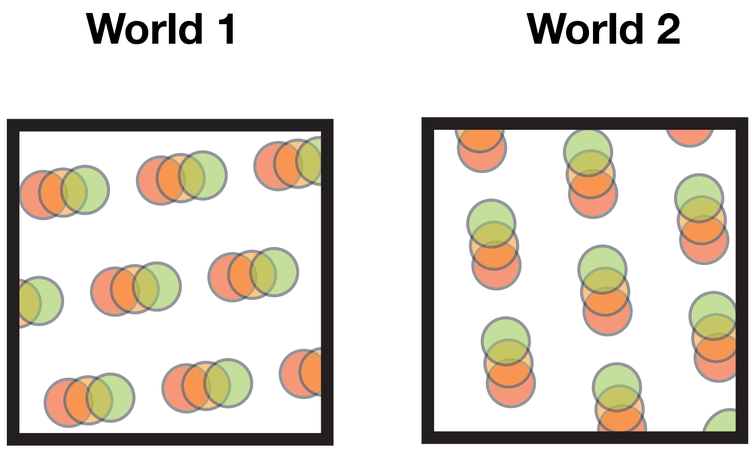


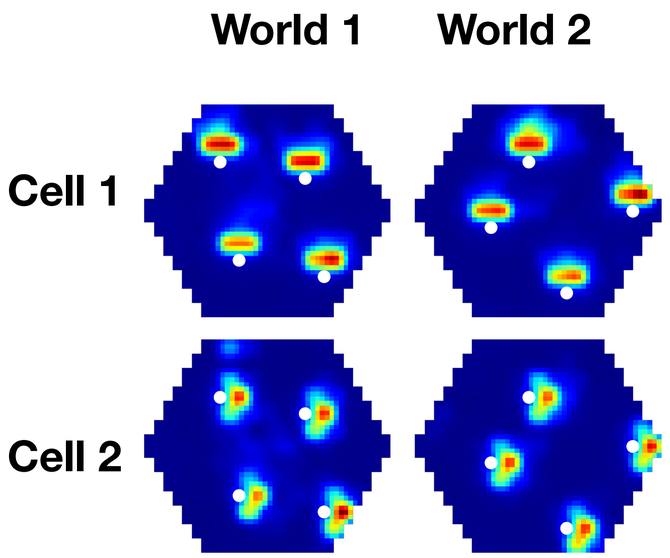
Whittington et al., 2020, *Cell*

Building block remapping

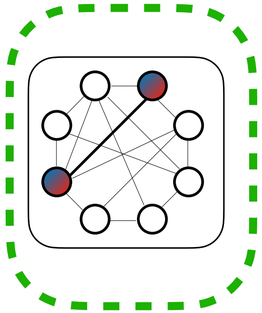








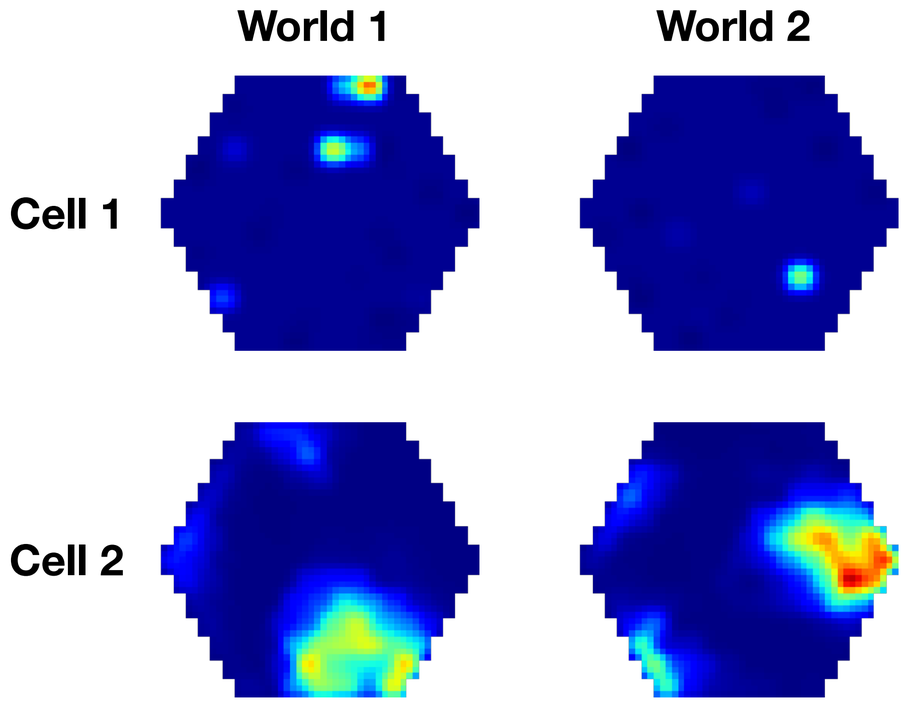
Whittington et al., 2020, *Cell*



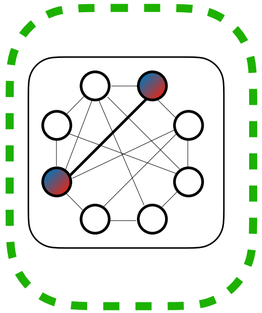
Hopfield
Network

Hippocampal remapping

Model cells



Whittington et al., 2020, *Cell*

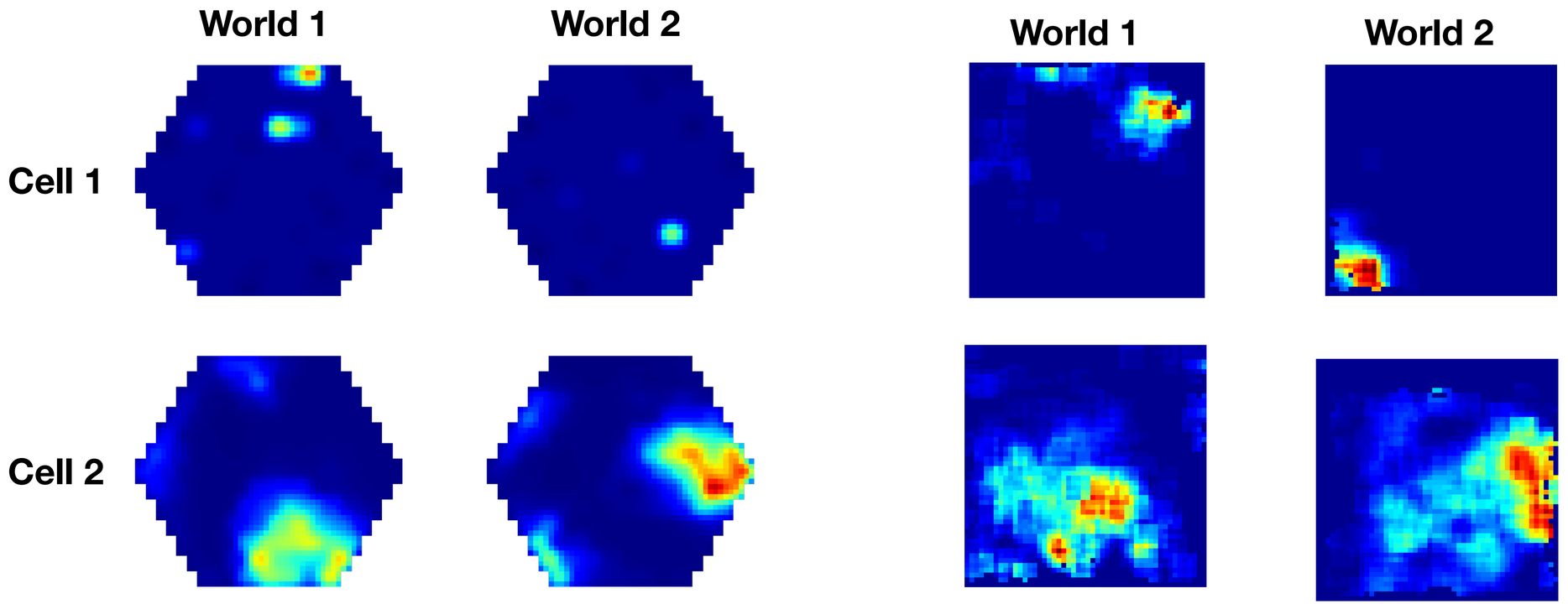


Hopfield
Network

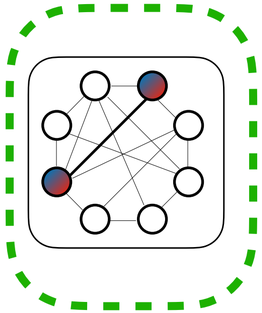
Hippocampal remapping

Model cells

Real cells



Whittington et al., 2020, *Cell*

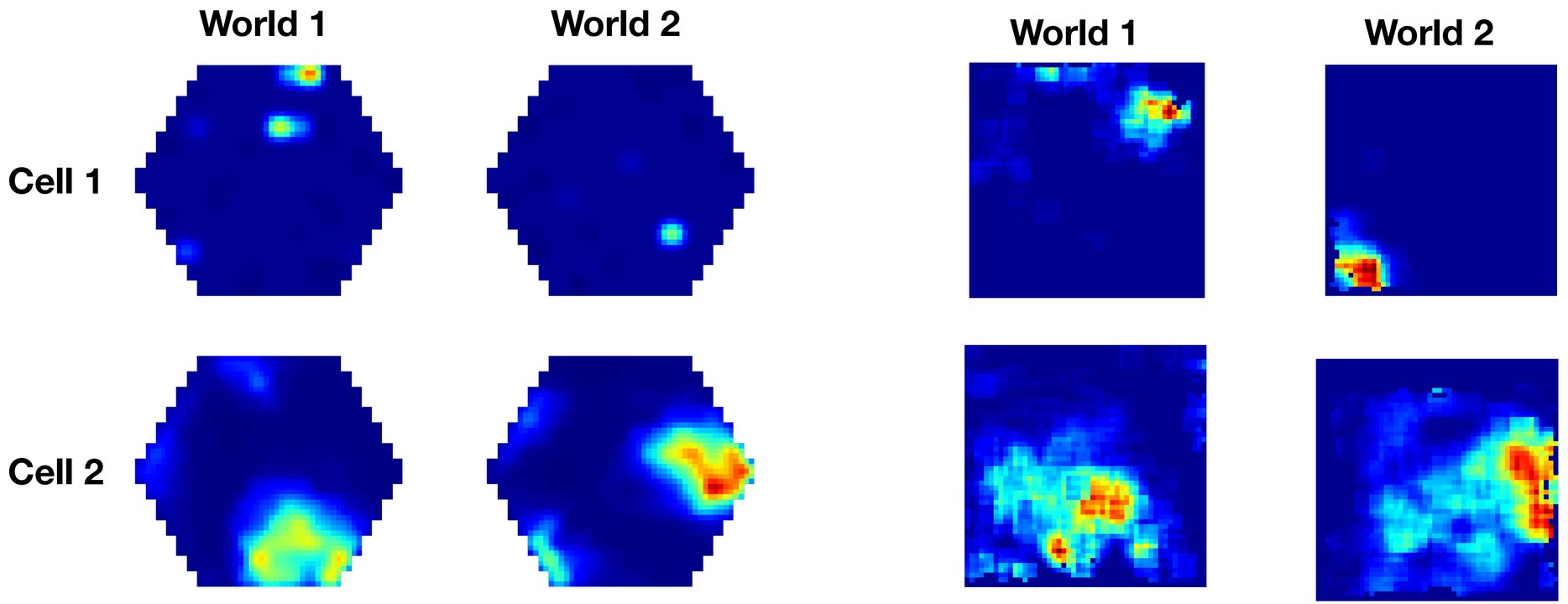


Hopfield
Network

Hippocampal remapping

Model cells

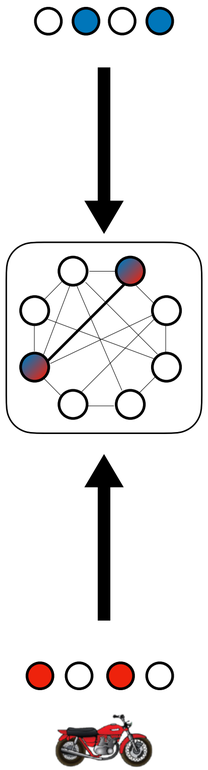
Real cells



Looks random but is it really?

Whittington et al., 2020, *Cell*

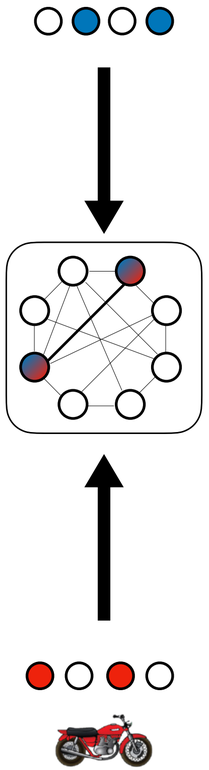
Remember that place cells are compositions of building blocks



Hippocampal cells are a conjunction of structural and sensory representations

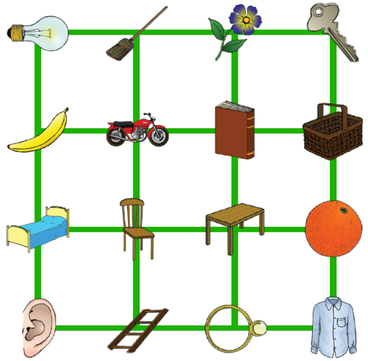
Whittington et al., 2020, *Cell*

Remember that place cells are compositions of building blocks

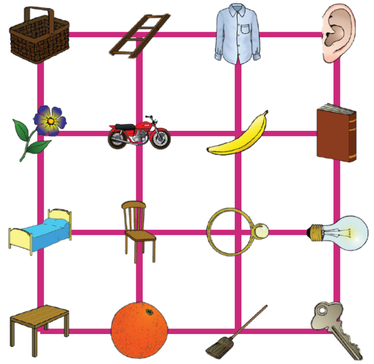


Hippocampal cells are a conjunction of structural and sensory representations

World 1



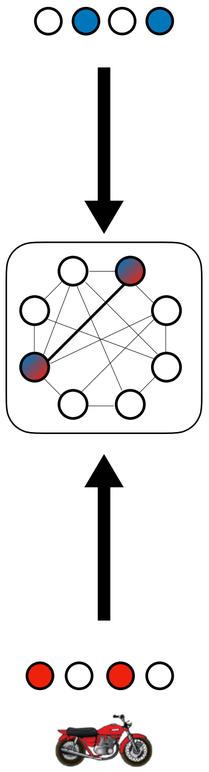
World 2



Building blocks are just in different combinations in different worlds

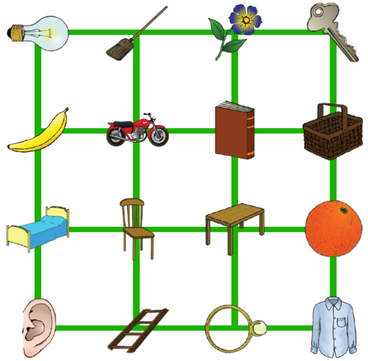
Whittington et al., 2020, *Cell*

Remember that place cells are compositions of building blocks

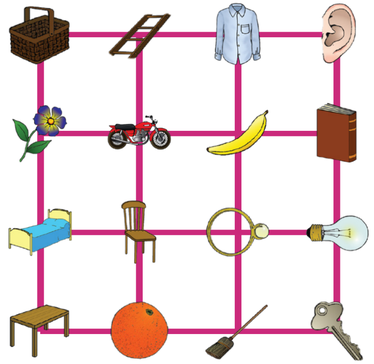


Hippocampal cells are a conjunction of structural and sensory representations

World 1



World 2



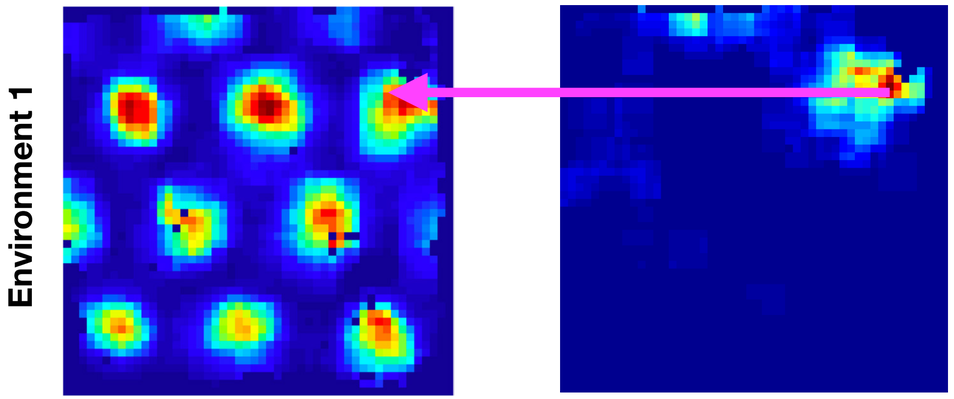
Building blocks are just in different combinations in different worlds

Hippocampal cells should remap consistent with their building block inputs

Whittington et al., 2020, *Cell*

Prediction: Grid cells control place cell remapping

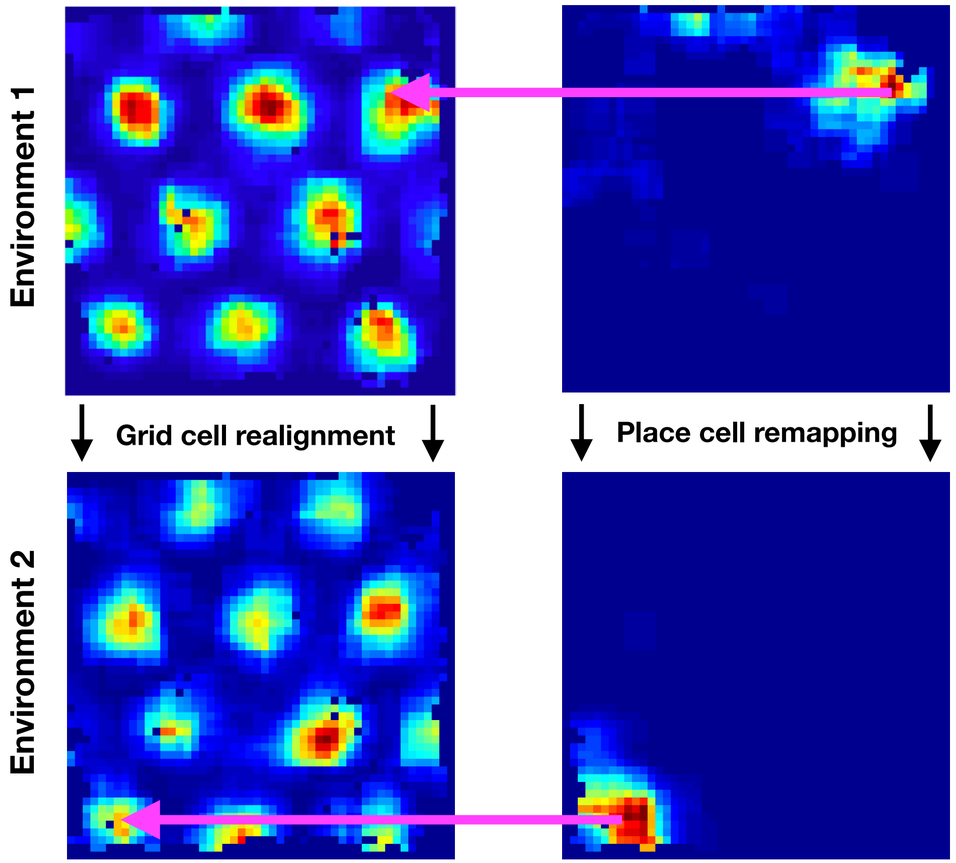
Place cells locations are constrained by grid cells



Whittington et al., 2020, *Cell*

Prediction: Grid cells control place cell remapping

Place cells locations are constrained by grid cells

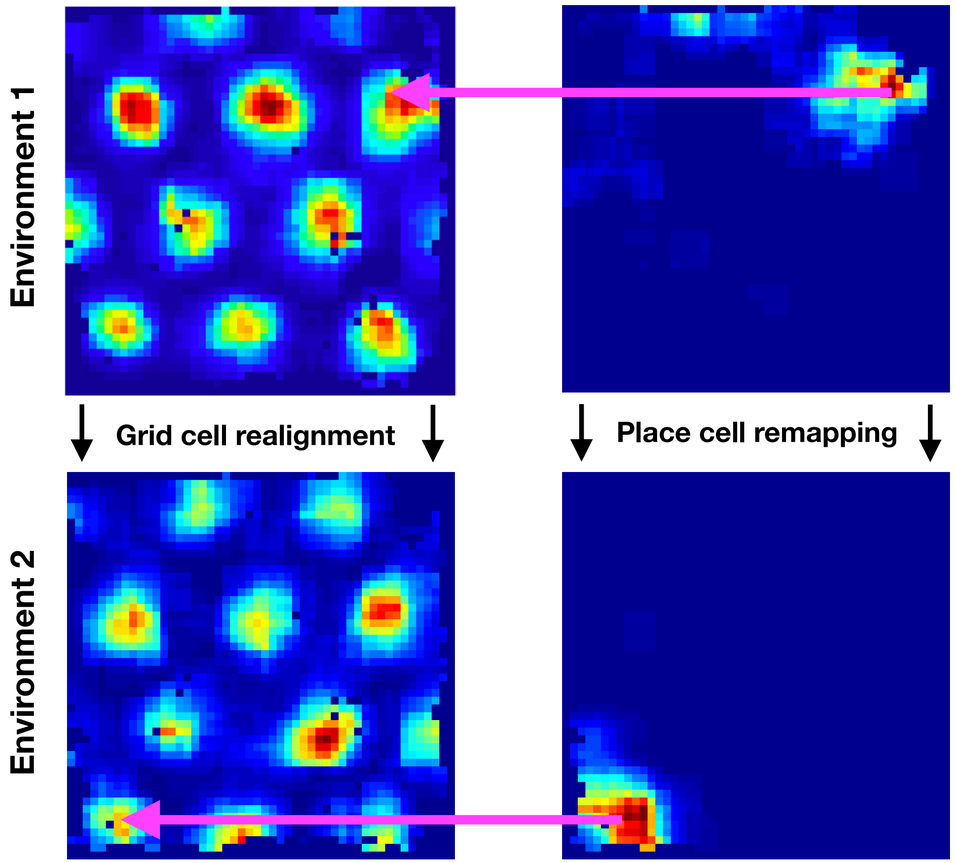


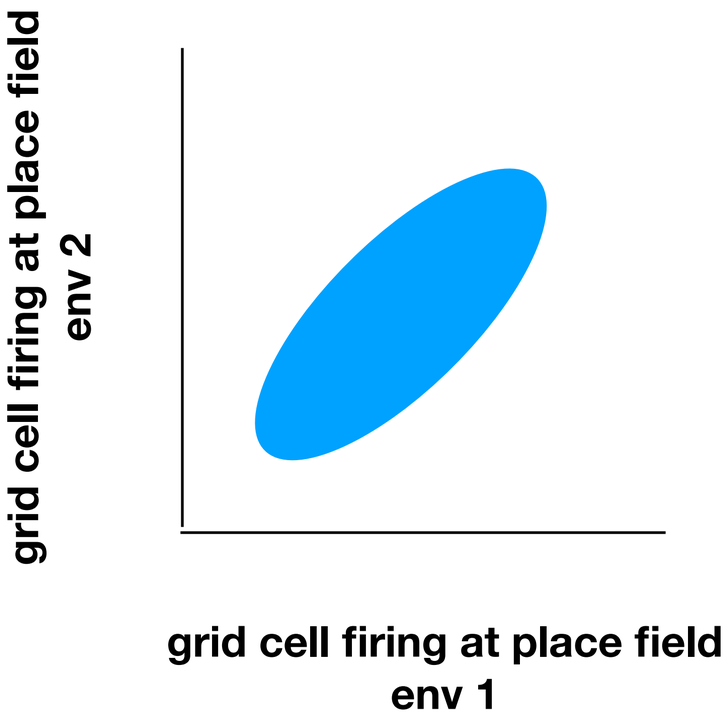
So they should remap to locations consistent with that grid cell

Whittington et al., 2020, *Cell*

Prediction: Grid cells control place cell remapping

Place cells locations are constrained by grid cells

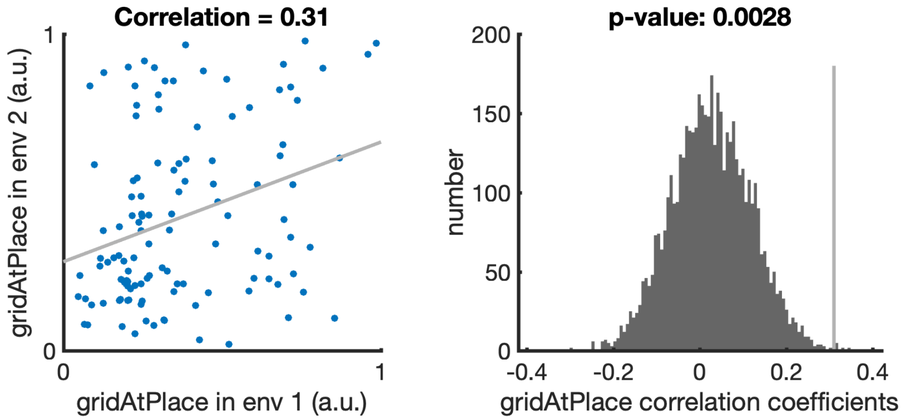




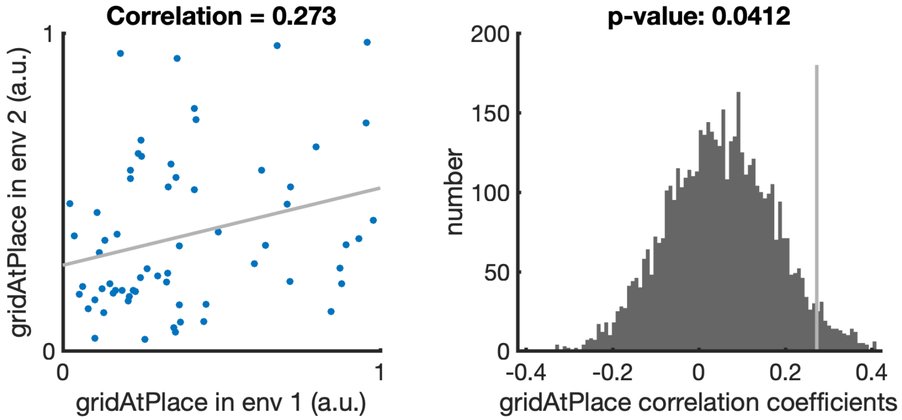
So they should remap to locations consistent with that grid cell

Whittington et al., 2020, *Cell*

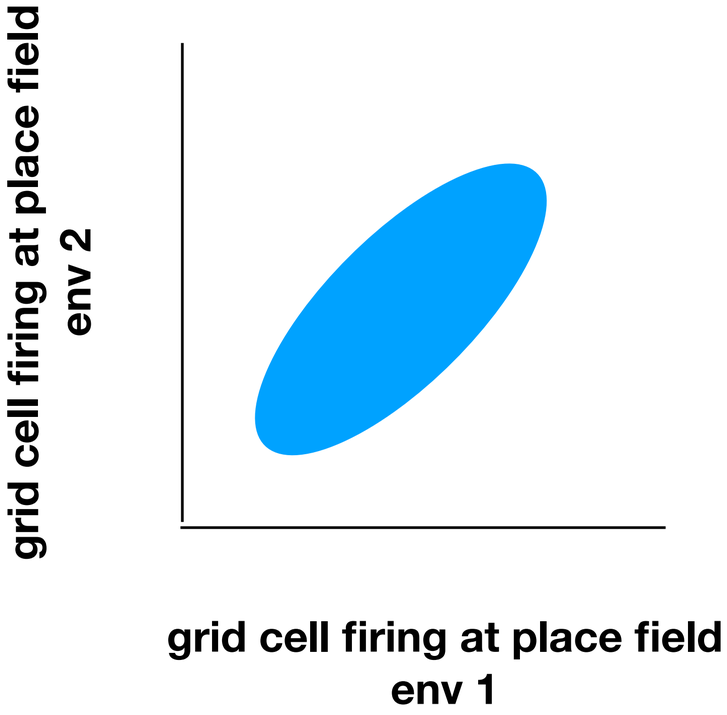
Prediction: Grid cells control place cell remapping



Data from Barry et al Current Biol. 2007

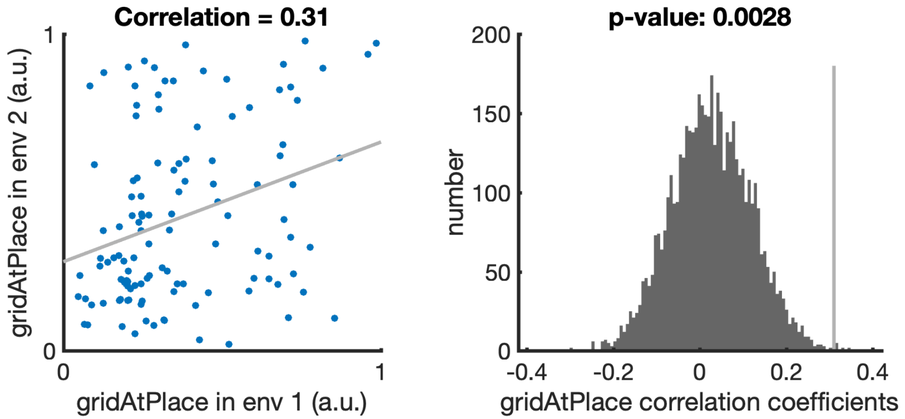


Data from Chen et al eLife 2018

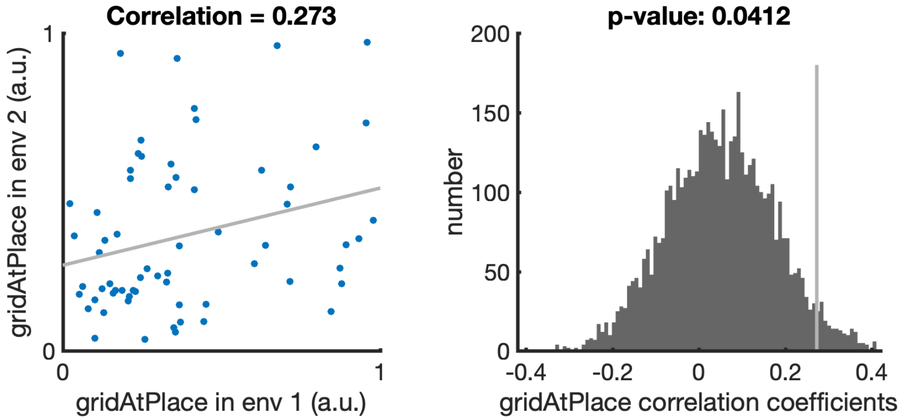


Whittington et al., 2020, *Cell*

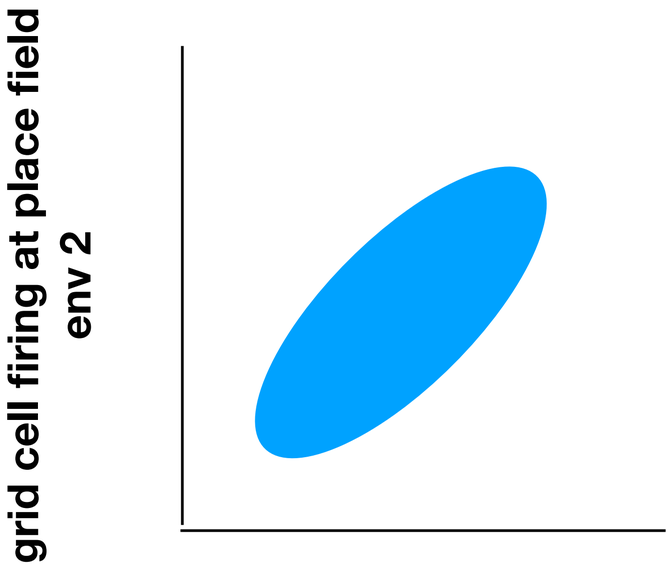
Prediction: Grid cells control place cell remapping



Data from Barry et al Current Biol. 2007



Data from Chen et al eLife 2018



Place cells don't remap randomly as previously thought!

Whittington et al., 2020, *Cell*

Representing more complex tasks

Lots of other hippocampal representations accounted for

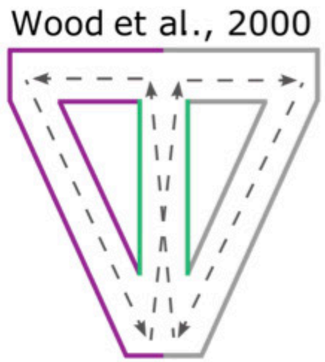
Whittington et al., 2022, *Nature Neuroscience*

Lots of other hippocampal representations accounted for

Wood et al., 2000

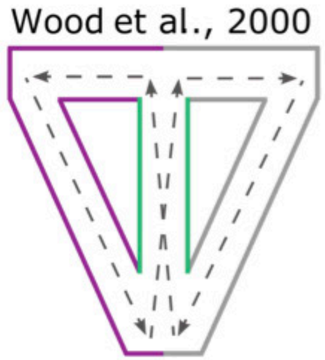
Whittington et al., 2022, *Nature Neuroscience*

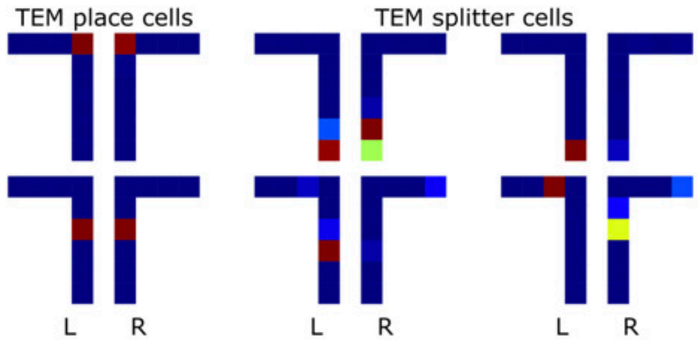
Lots of other hippocampal representations accounted for



Whittington et al., 2022, *Nature Neuroscience*

Lots of other hippocampal representations accounted for



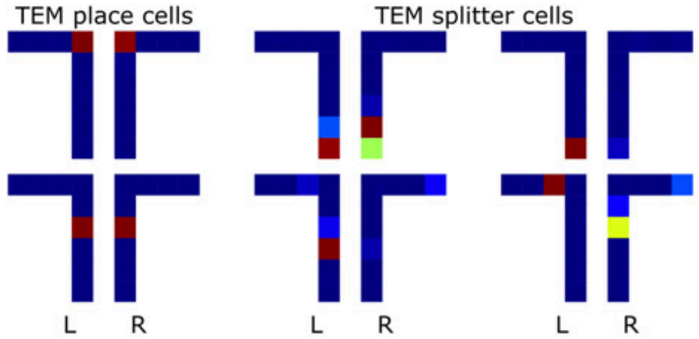


Whittington et al., 2022, *Nature Neuroscience*

Lots of other hippocampal representations accounted for

Wood et al., 2000

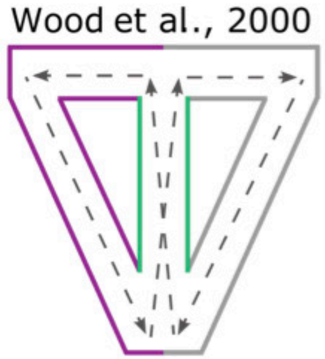
Simultaneous representation of spatial and task location

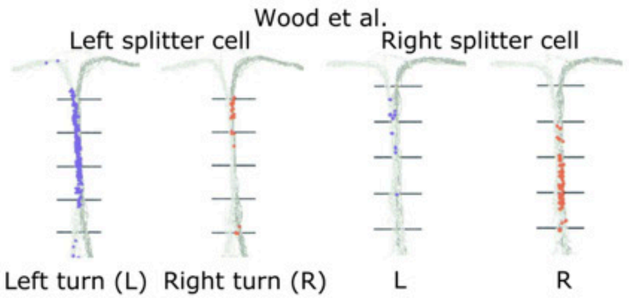


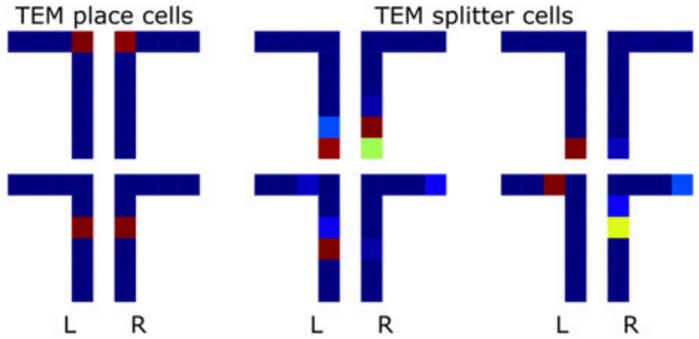
Whittington et al., 2022, *Nature Neuroscience*

Lots of other hippocampal representations accounted for

Simultaneous representation of spatial and task location



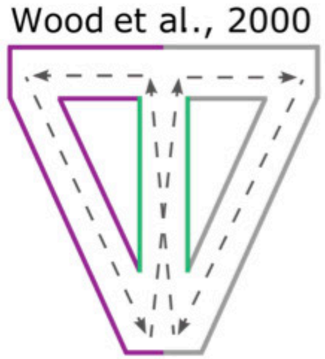


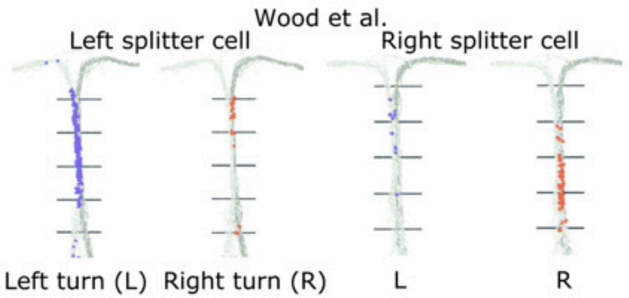


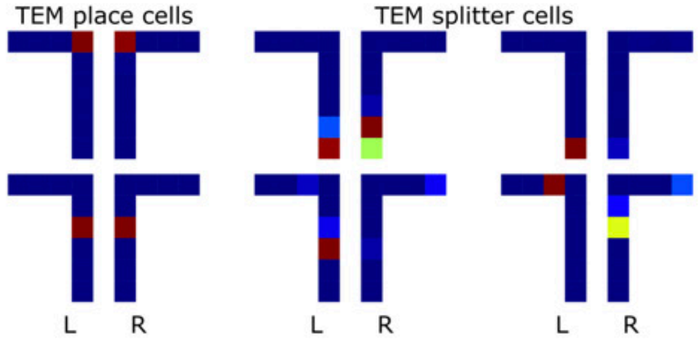
Whittington et al., 2022, *Nature Neuroscience*

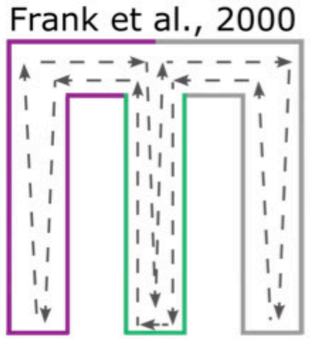
Lots of other hippocampal representations accounted for

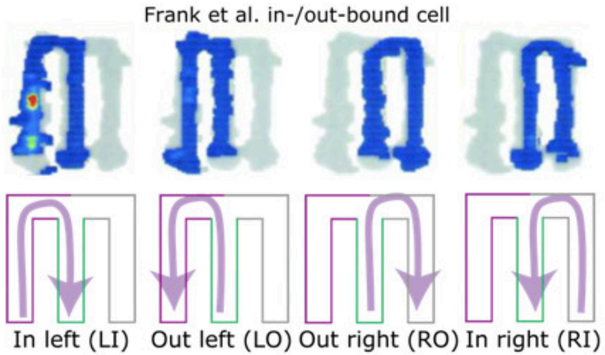
Simultaneous representation of spatial and task location

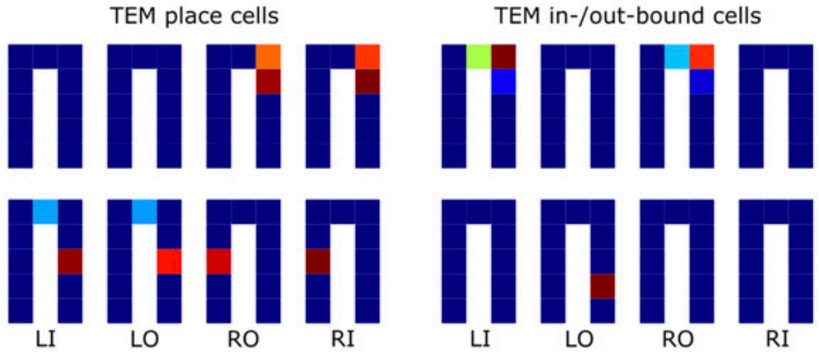










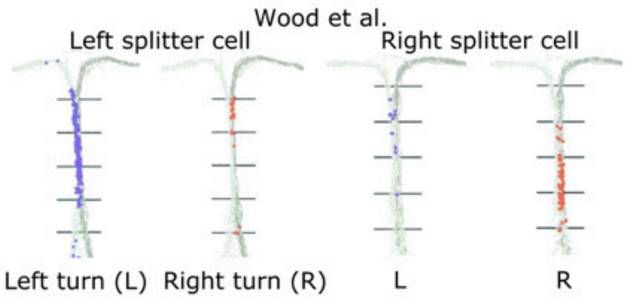


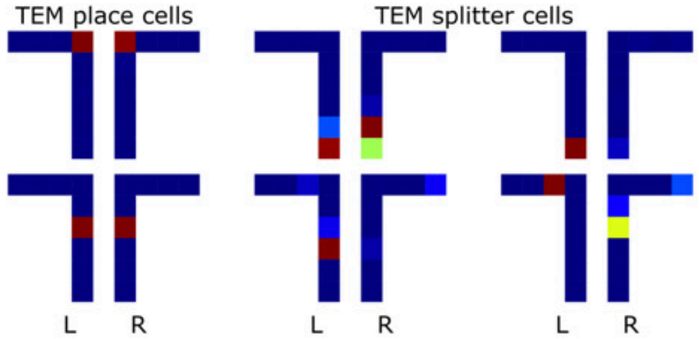
Whittington et al., 2022, *Nature Neuroscience*

Lots of other hippocampal representations accounted for

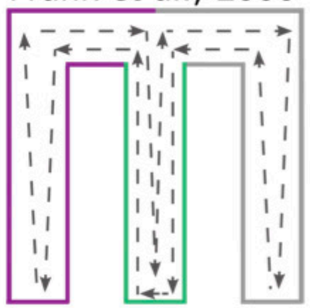
Simultaneous representation of spatial and task location

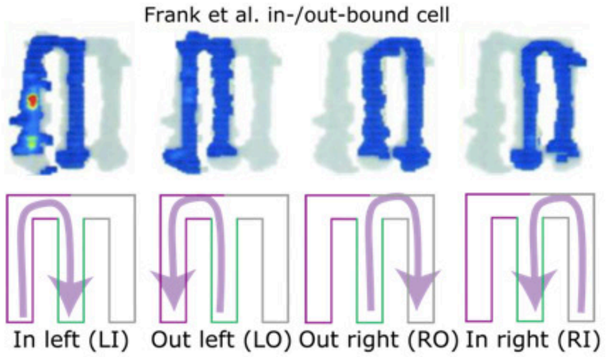
Wood et al., 2000

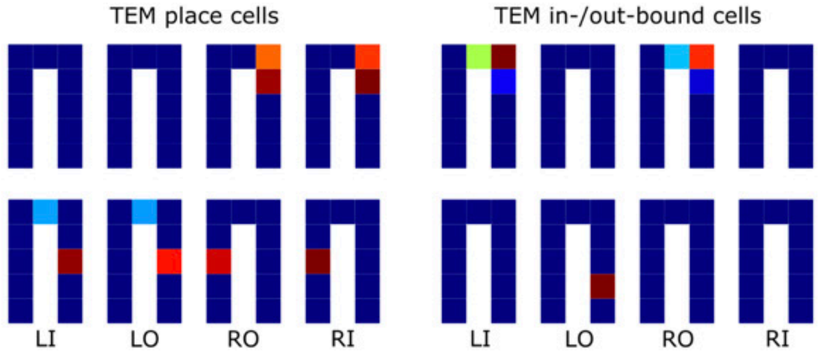




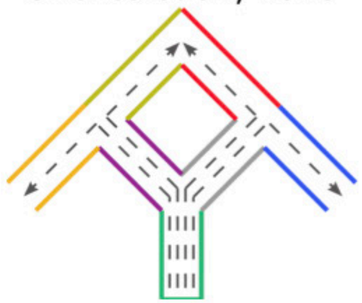
Frank et al., 2000

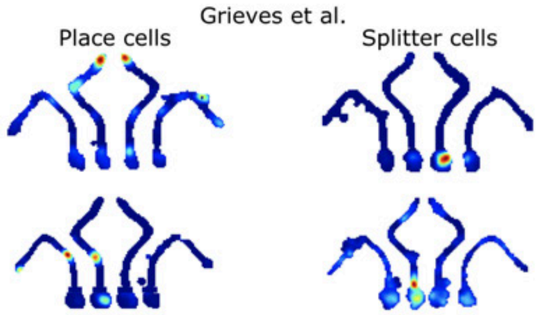


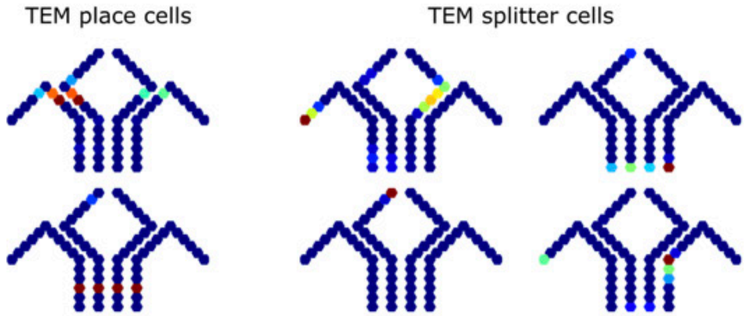




Grieves et al., 2016







Whittington et al., 2022, *Nature Neuroscience*

Lots of other hippocampal representations accounted for

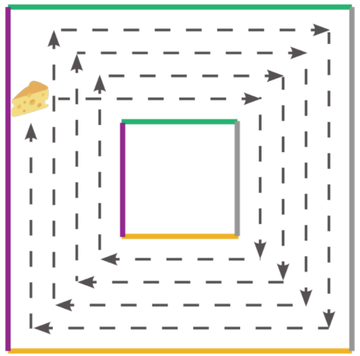
Simultaneous representation of spatial and task location

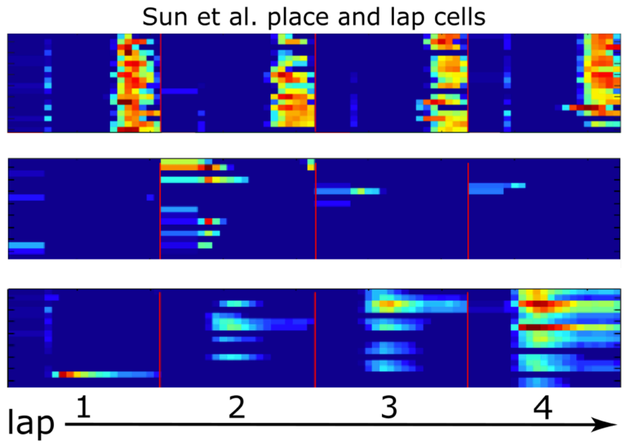
Whittington et al., 2022, *Nature Neuroscience*

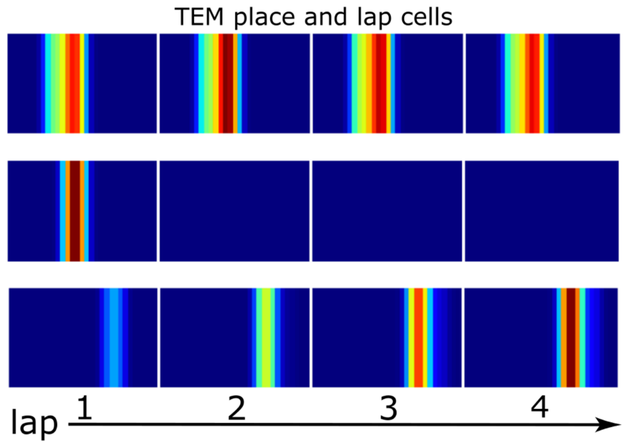
Lots of other hippocampal representations accounted for

Simultaneous representation of spatial and task location

Sun et al., 2020





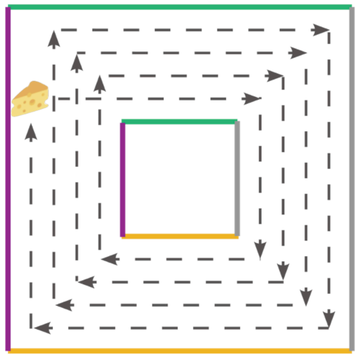


Whittington et al., 2022, *Nature Neuroscience*

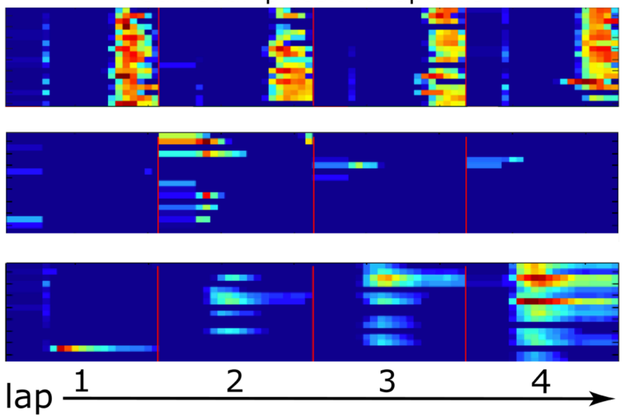
Lots of other hippocampal representations accounted for

Simultaneous representation of spatial and task location

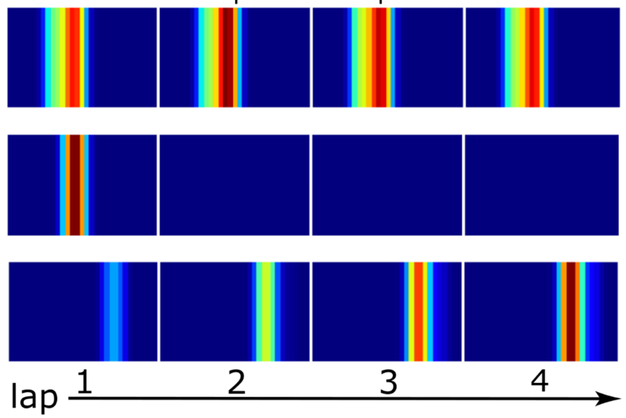
Sun et al., 2020



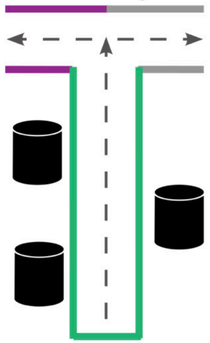
Sun et al. place and lap cells



TEM place and lap cells

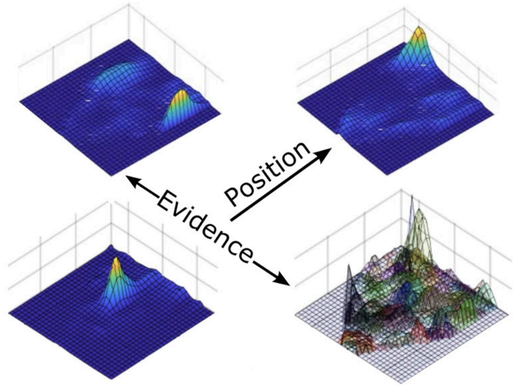


Nieh et al., 2021



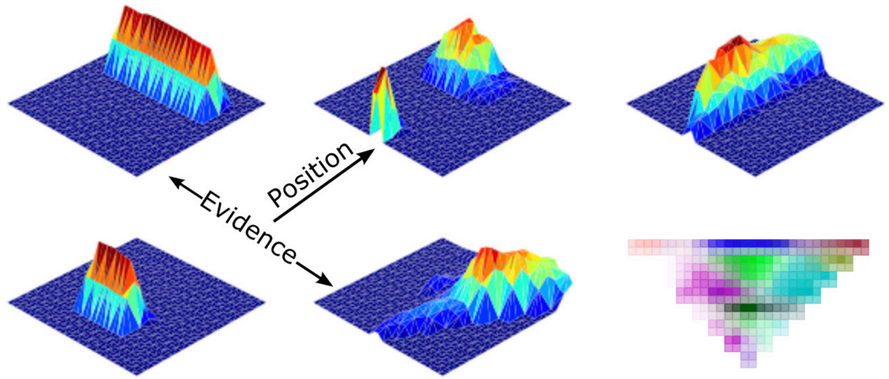
Rodent chooses side with more cues

Nieh et al. abstract position/evidence cells



TEM place cells

TEM abstract position/evidence cells

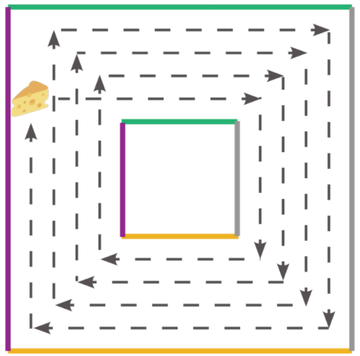


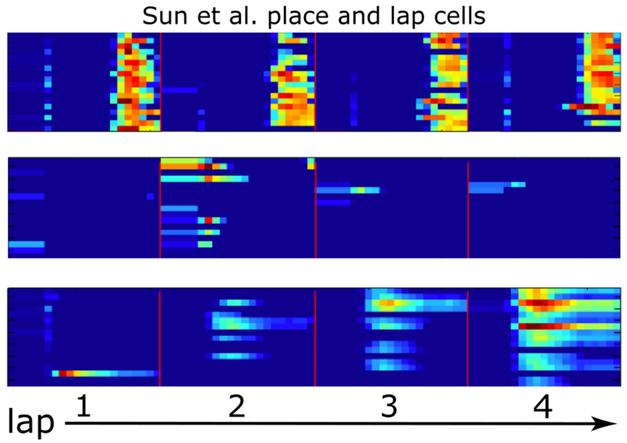
Whittington et al., 2022, *Nature Neuroscience*

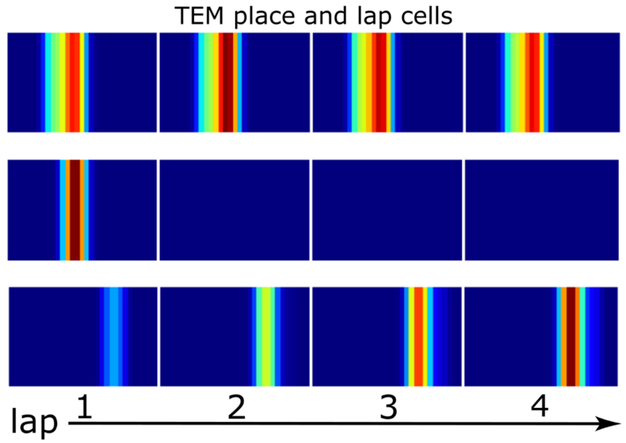
Lots of other hippocampal representations accounted for

Simultaneous representation of spatial and task location

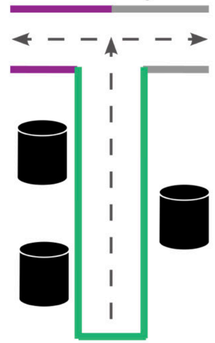
Sun et al., 2020



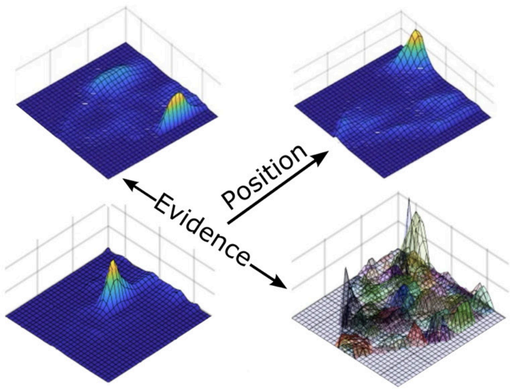


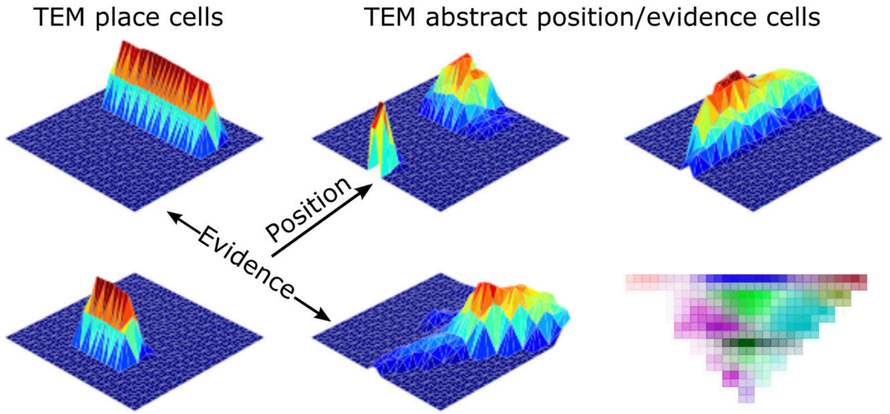


Nieh et al., 2021



Rodent chooses side with more cues

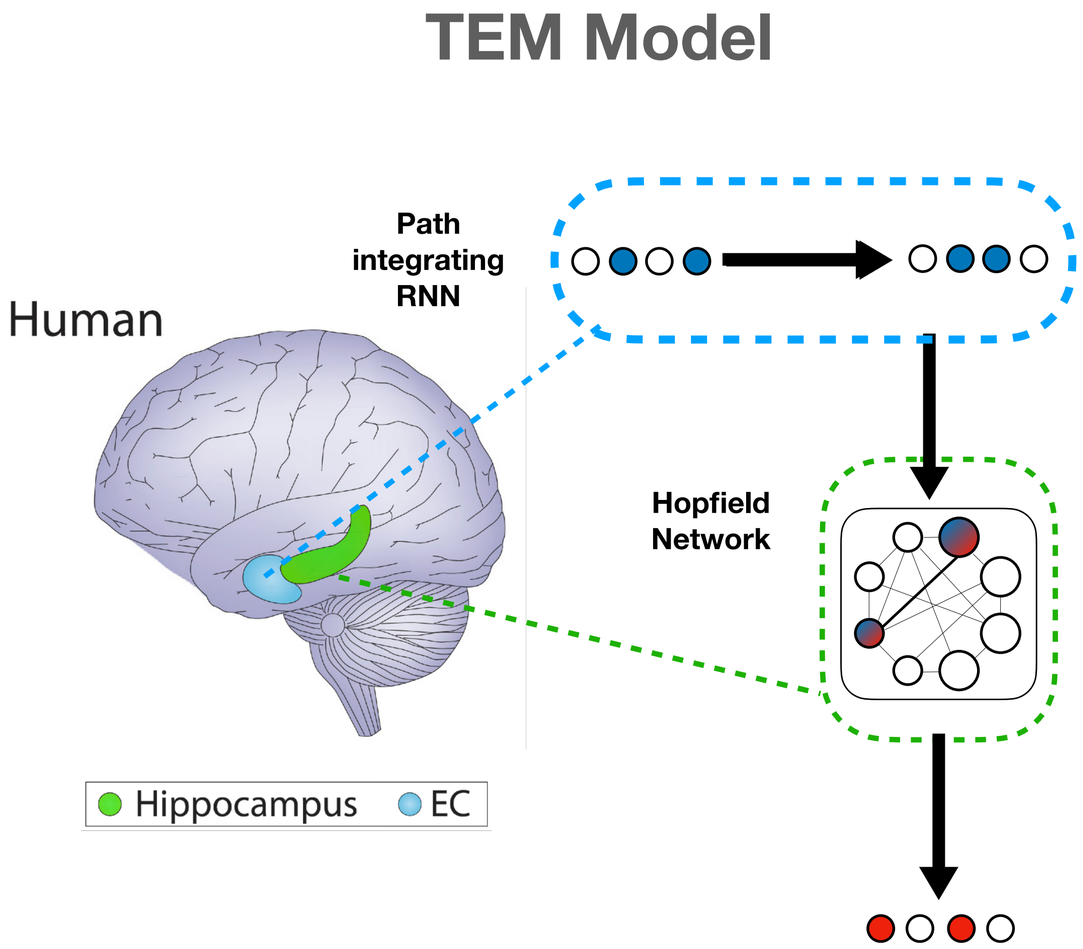




We've explained lots of cells, spatial and non-spatial, by the same principles!

Whittington et al., 2022, *Nature Neuroscience*

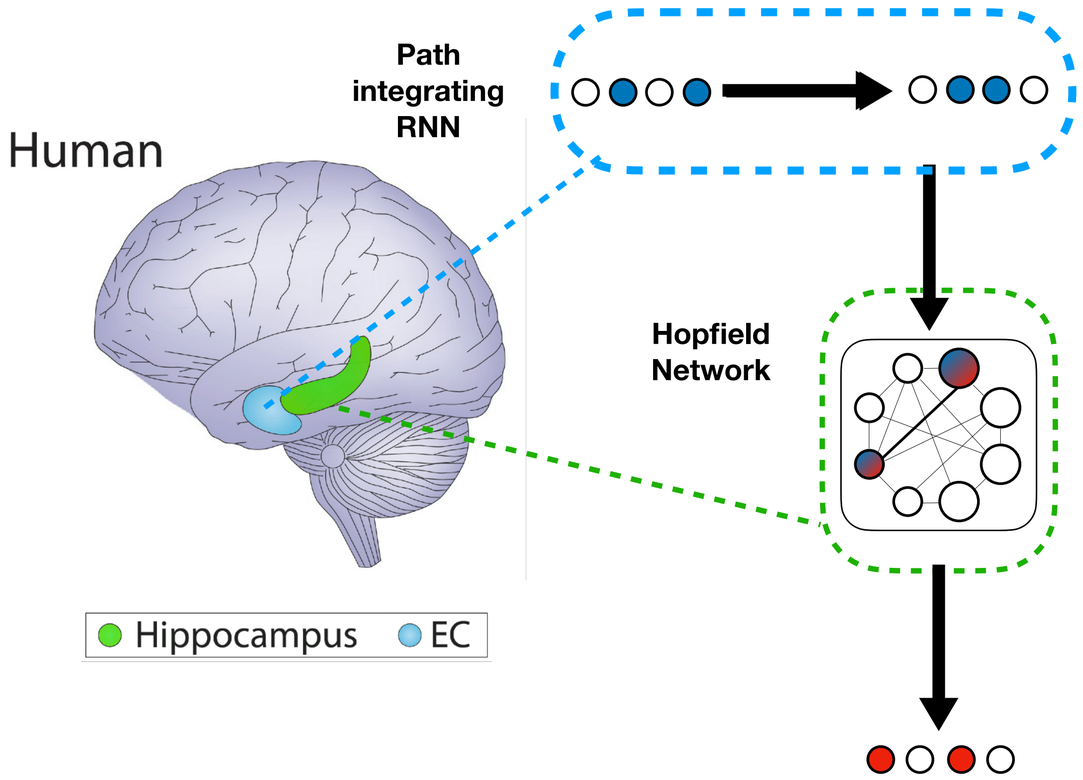
Side note: TEM is mathematically related to transformers



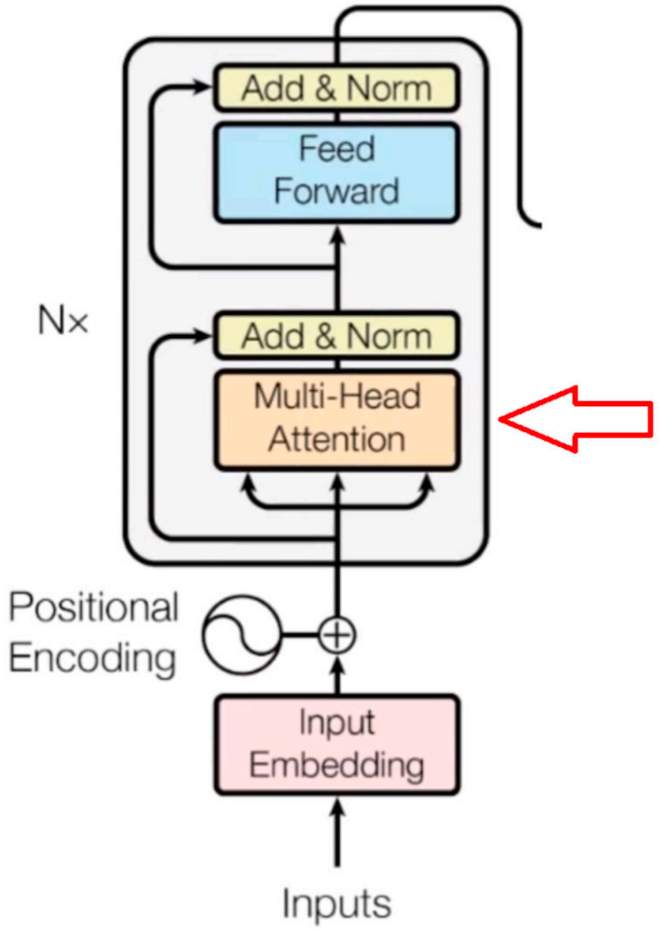
Whittington et al., 2022, *ICLR*

Side note: TEM is mathematically related to transformers

TEM Model



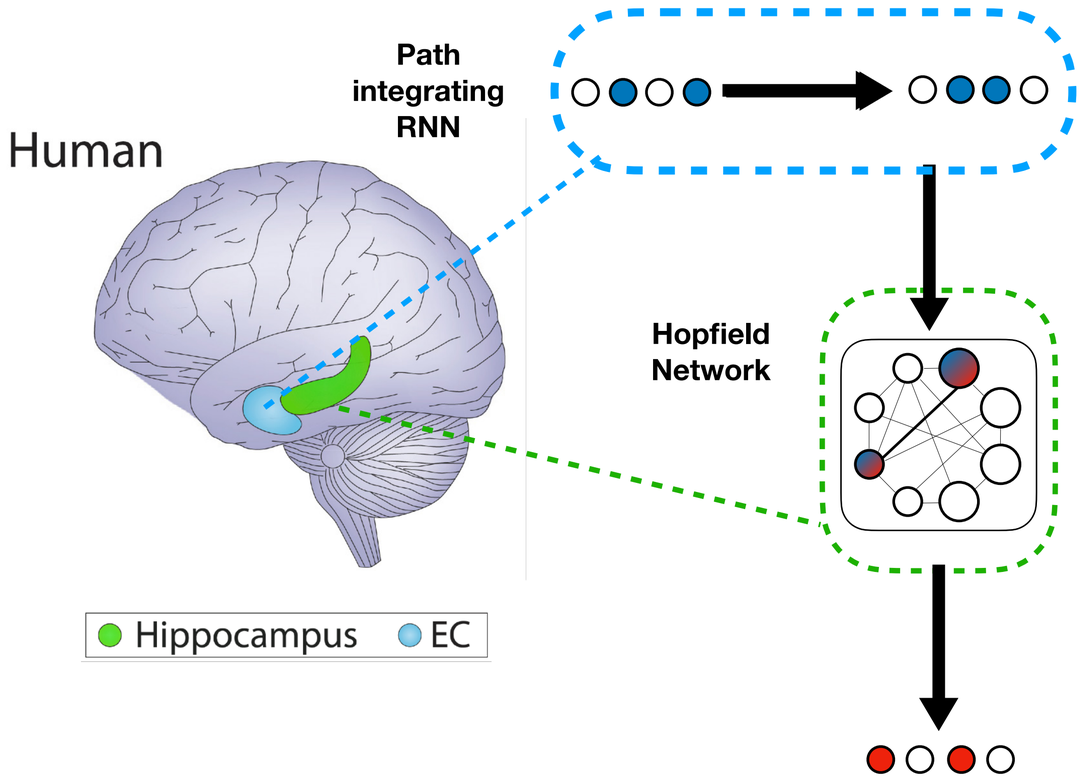
Transformer neural networks



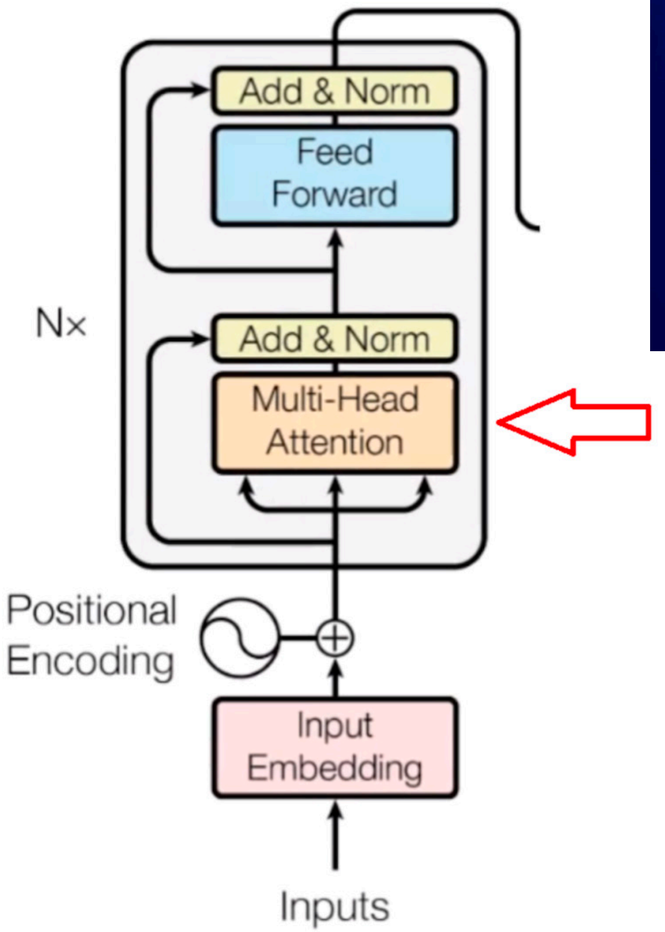
Whittington et al., 2022, *ICLR*

Side note: TEM is mathematically related to transformers

TEM Model



Transformer neural networks

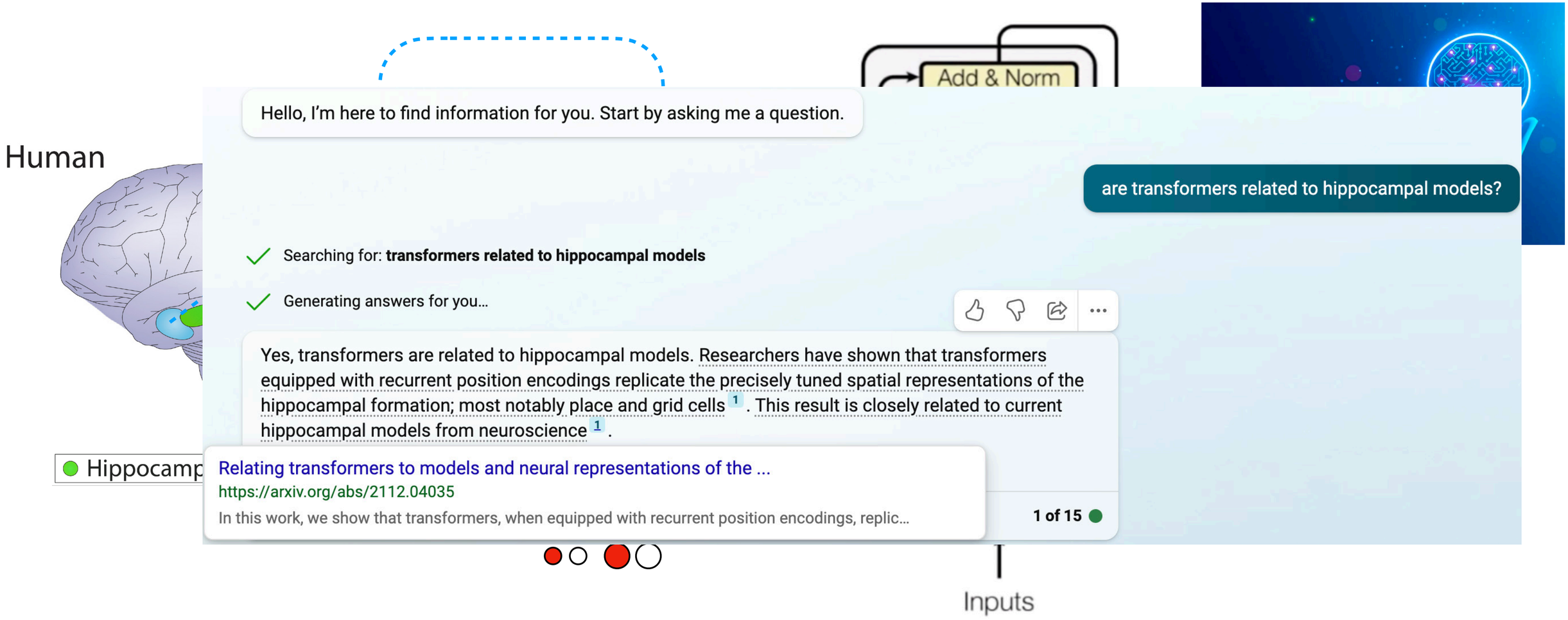


Whittington et al., 2022, *ICLR*

Side note: TEM is mathematically related to transformers

TEM Model

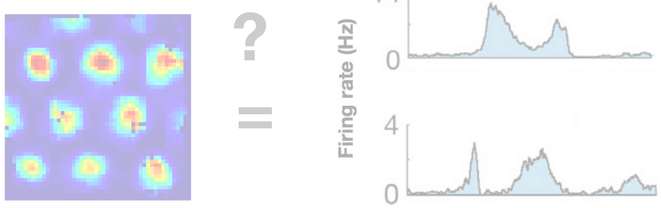
Transformer neural networks



Whittington et al., 2022, *ICLR*

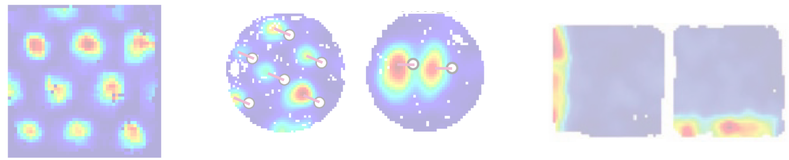
Puzzles of cognitive maps in the brain

How does the same system do space and non-space?

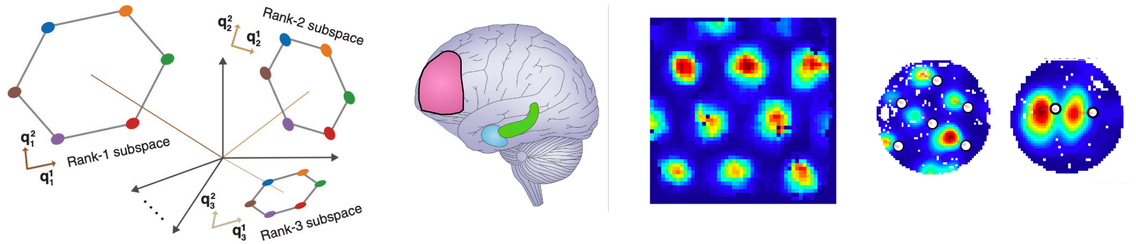


How can brains learn these maps?

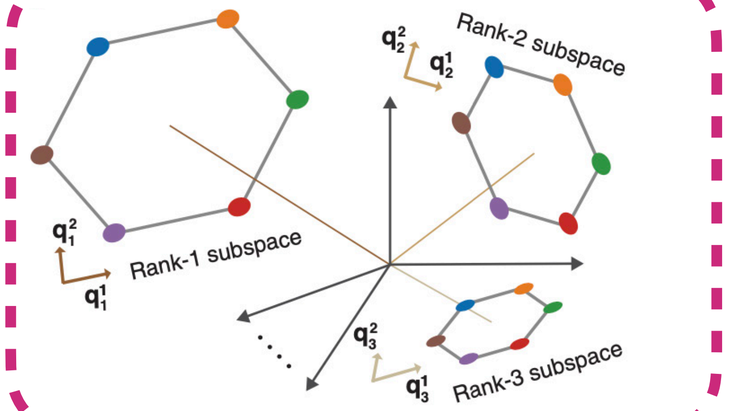
Why do the neuron look the ways they do?



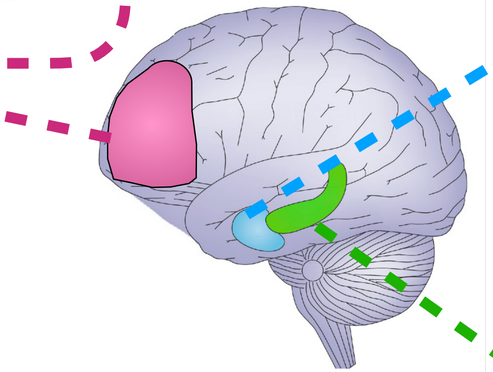
How do different brain regions solve the same problem in different ways?



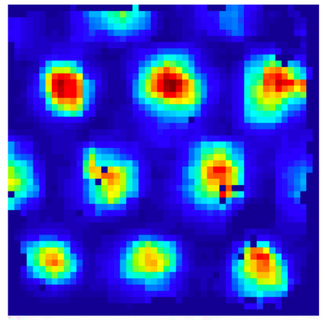
These models don’t capture anything about prefrontal cortex



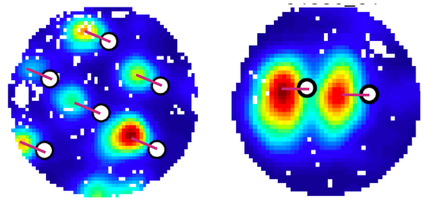
These cells include the past and future



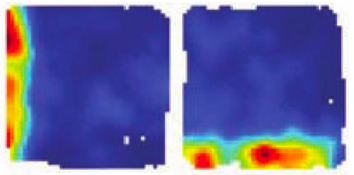
These cells tell you where you are or what you’re seeing right now



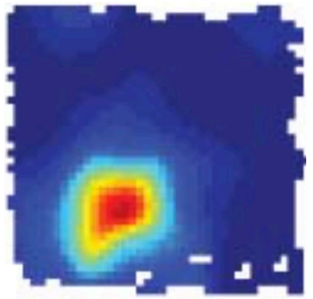
Grid cells
Hafting et al., 2005



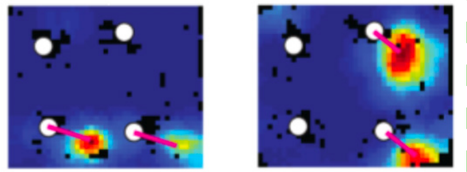
Object vector cells
Hoydal et al., 2018



Border cells
Solstad et al., 2008



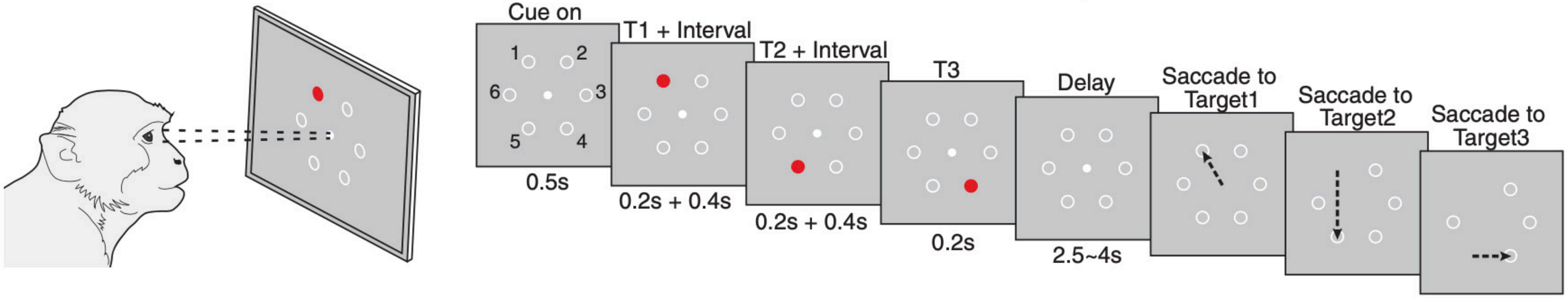
Place cells
O’Keefe & Dostrovsky, 1971



Landmark cells
Deshmukh & Knierim, 2013

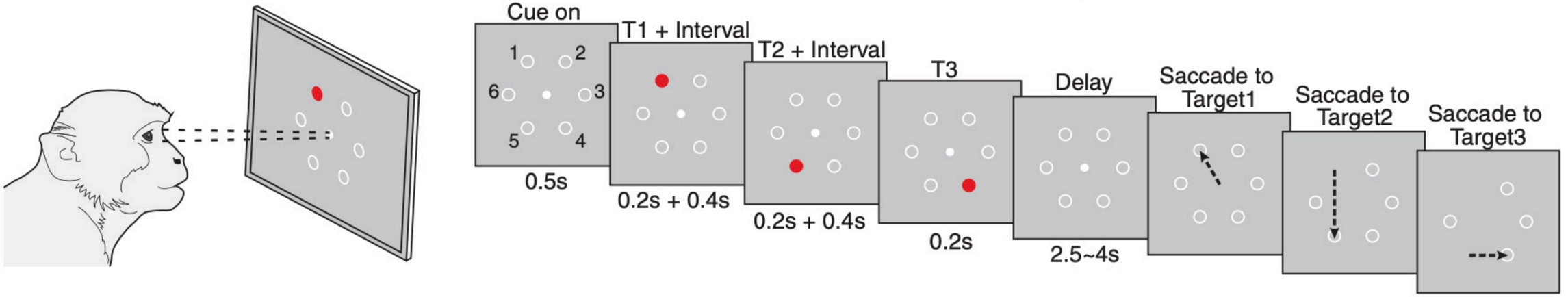
Prefrontal cells represent the whole history at any given time

Prefrontal cells represent the whole history at any given time

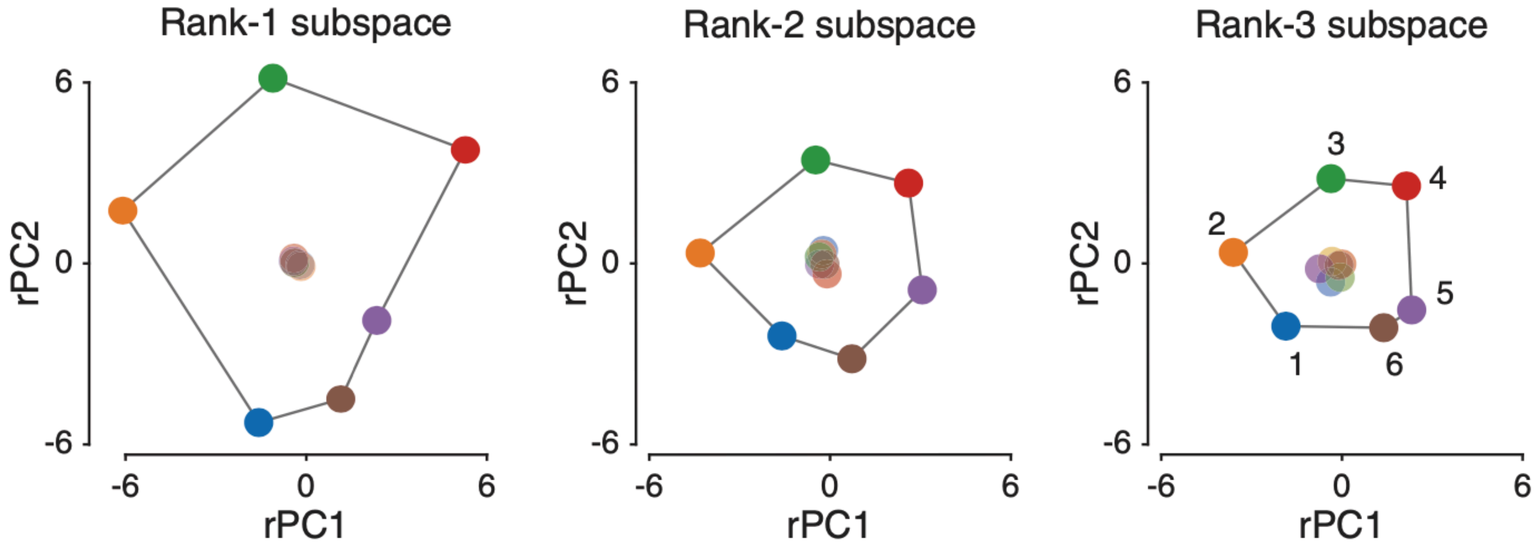


Xie et al., 2022

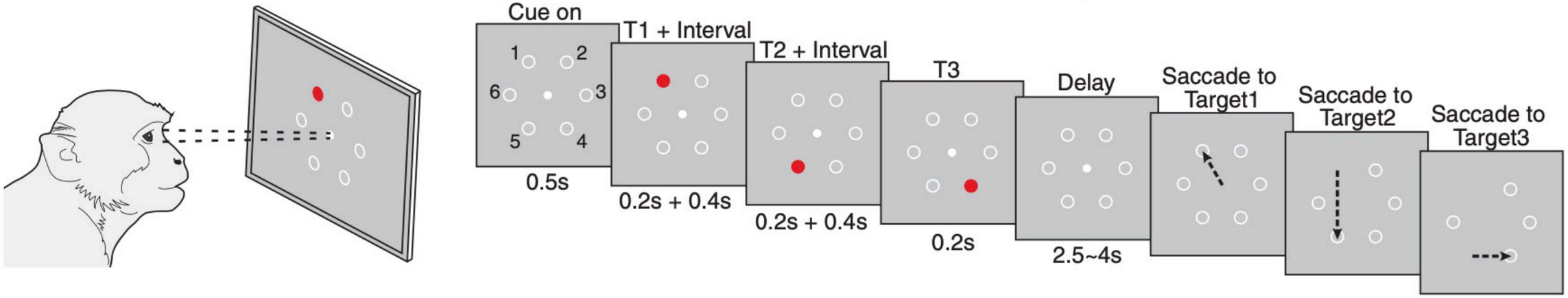
Prefrontal cells represent the whole history at any given time



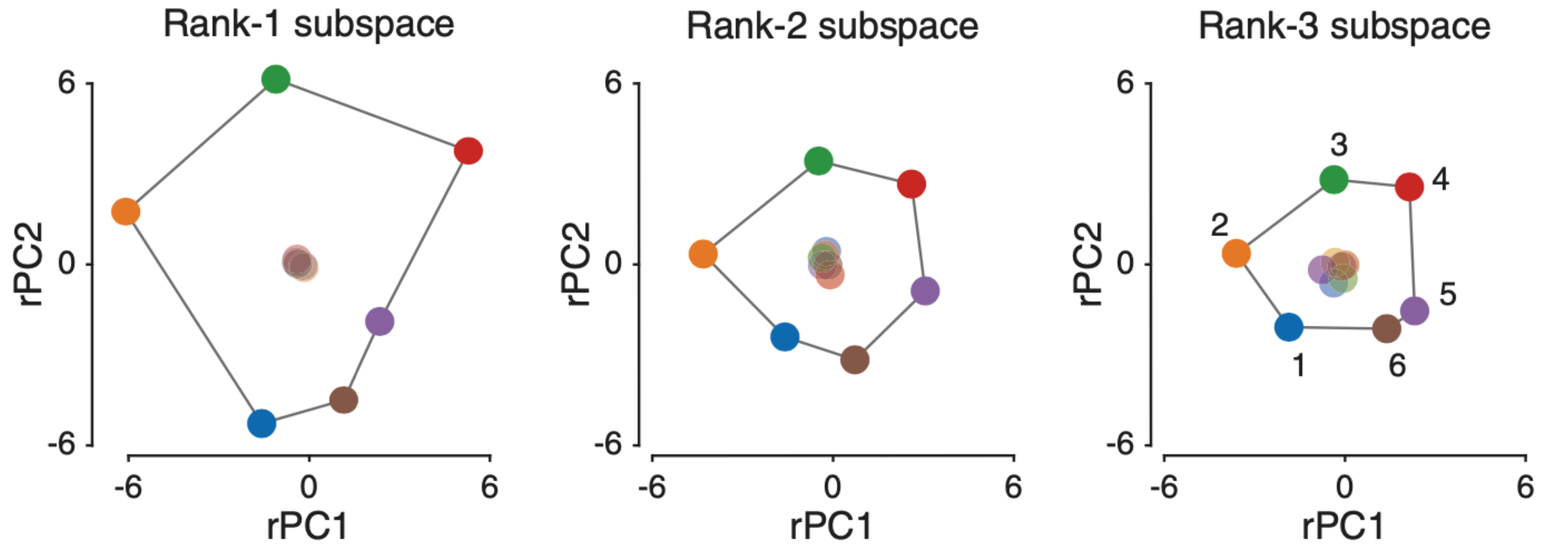
Xie et al., 2022



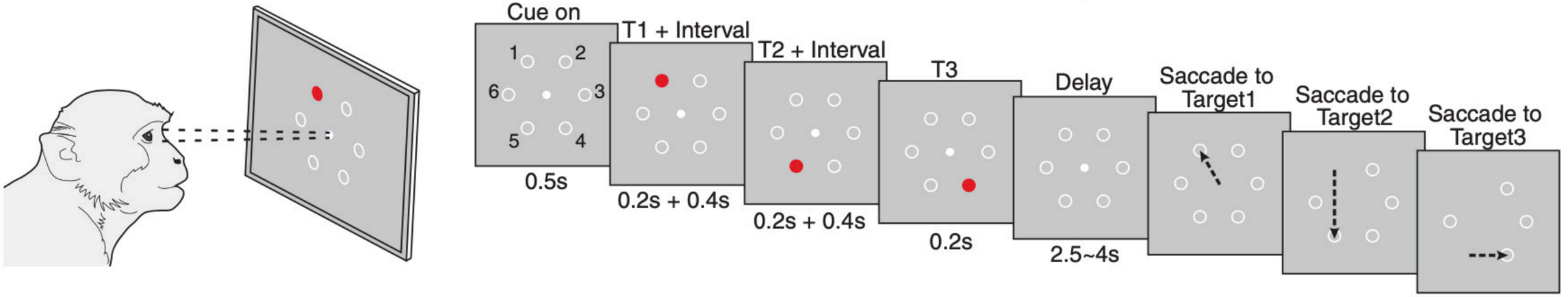
Prefrontal cells represent the whole history at any given time



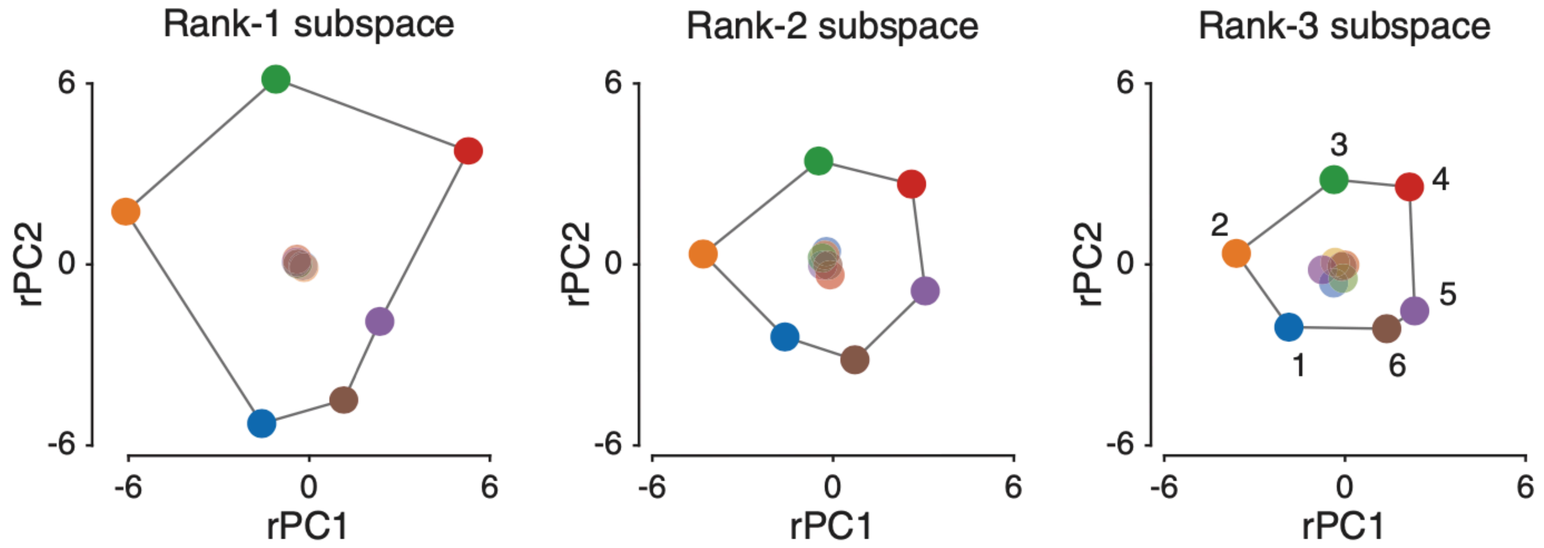
Xie et al., 2022



Prefrontal cells represent the whole history at any given time

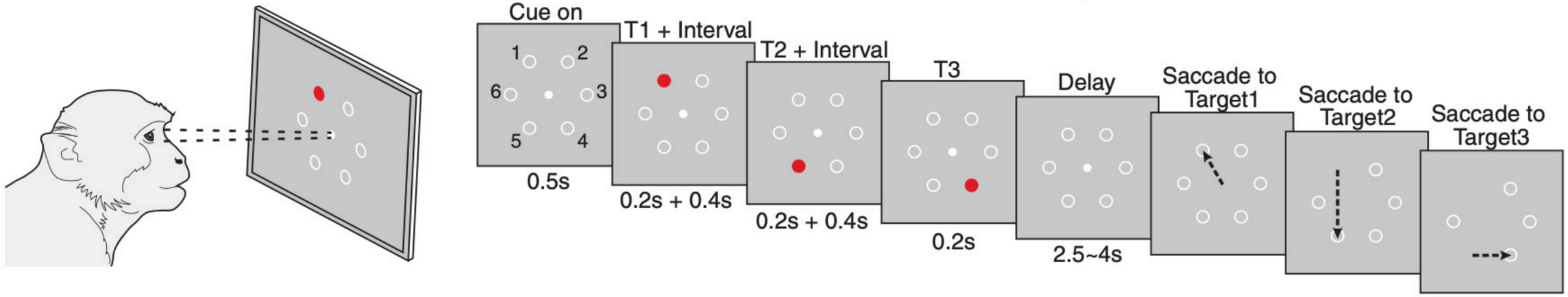


Xie et al., 2022

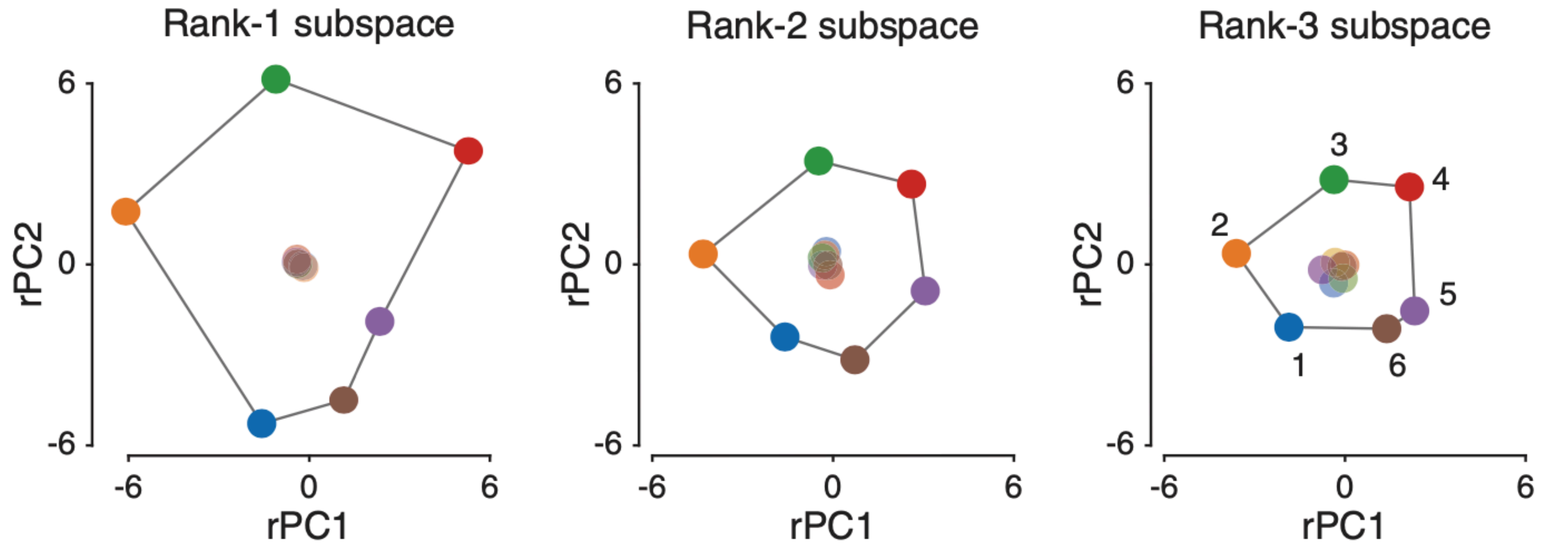


For sequence 1,3,4

Prefrontal cells represent the whole history at any given time

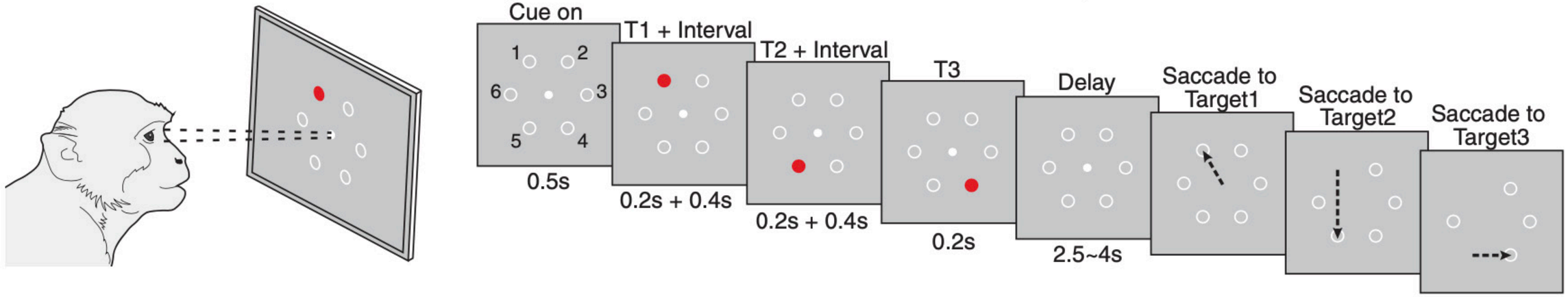


Xie et al., 2022

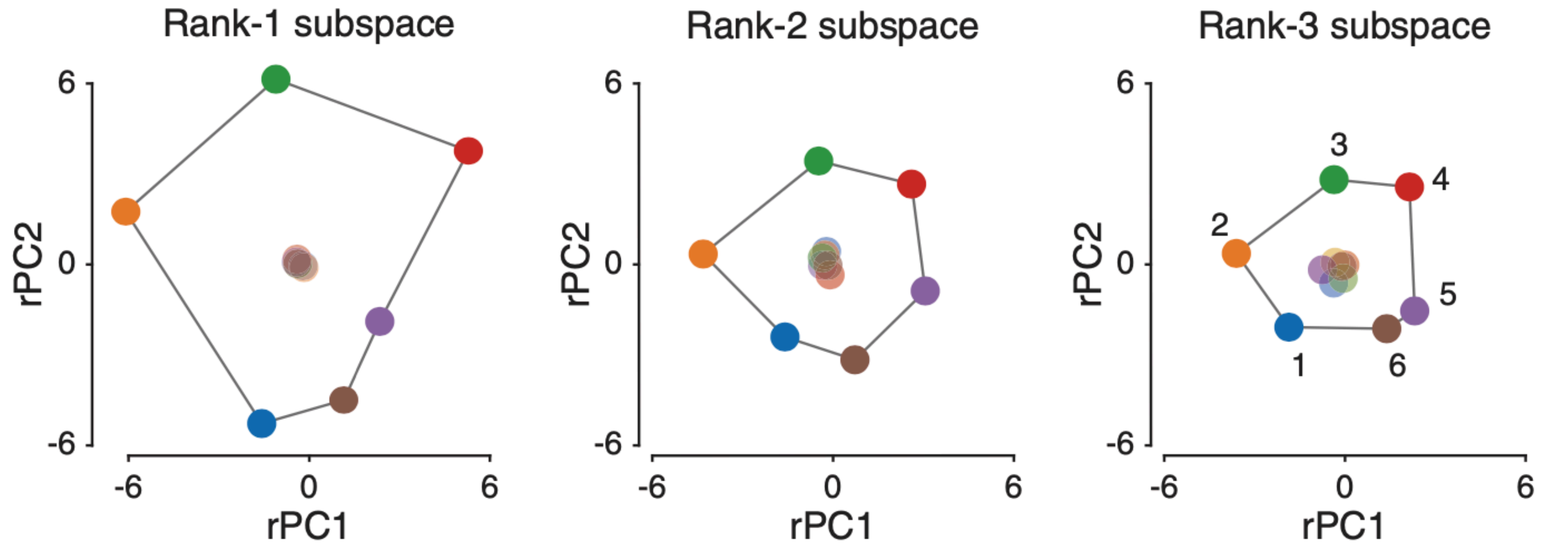


For sequence 5,2,6

Prefrontal cells represent the whole history at any given time



Xie et al., 2022



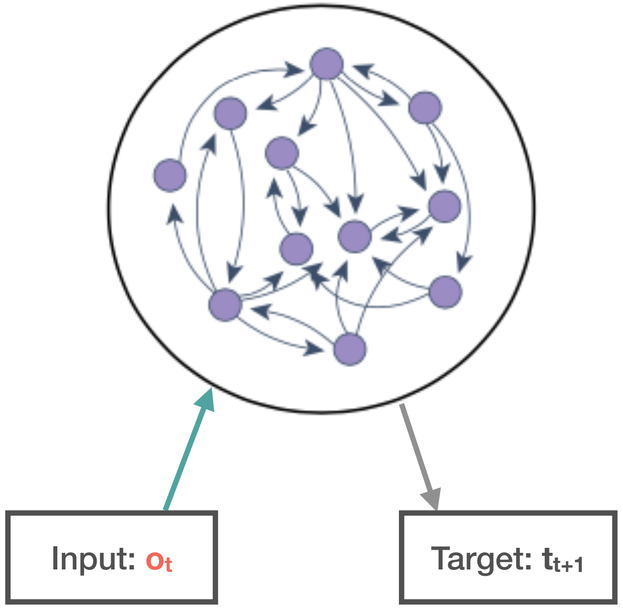
For sequence 5,2,6

These neurons keep a memory of the past!

Recurrent neural networks have internal memory

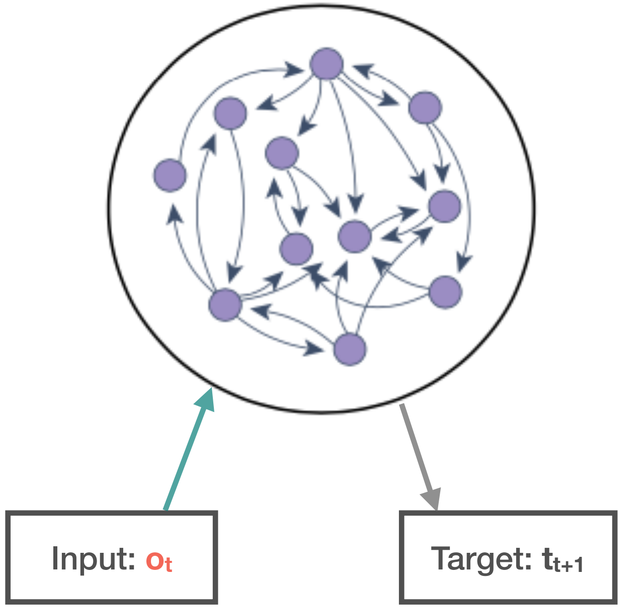
For sequence 5,2,6

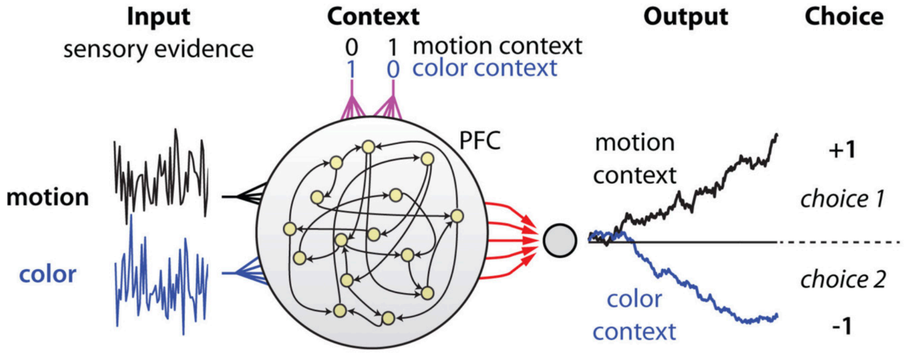
Recurrent neural networks have internal memory



For sequence 5,2,6

Recurrent neural networks have internal memory

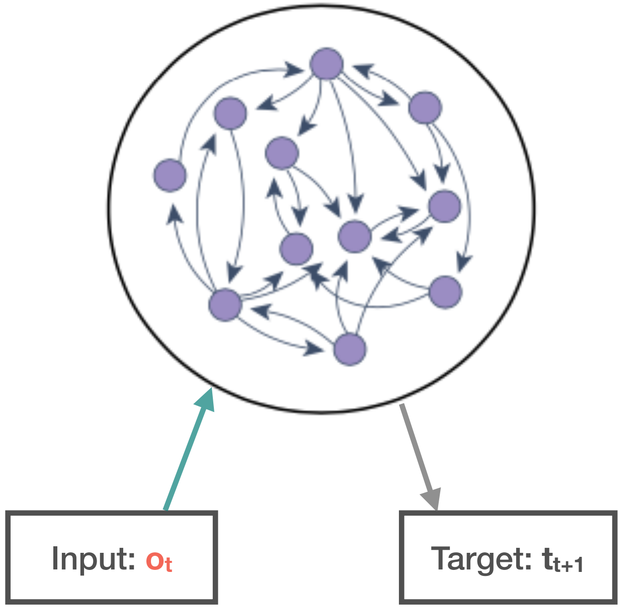


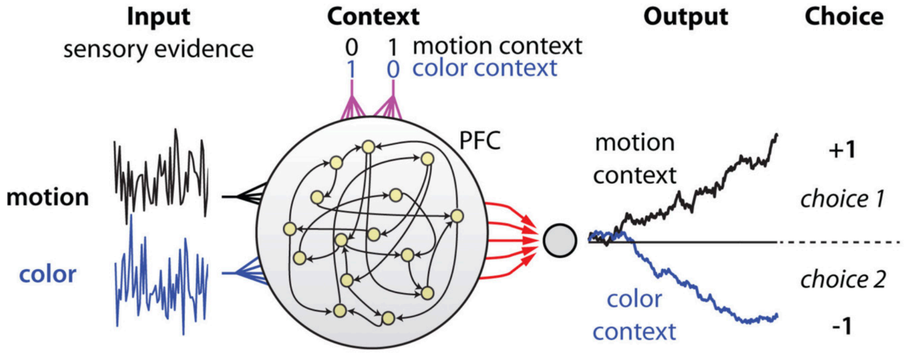


Mante et al., 2013

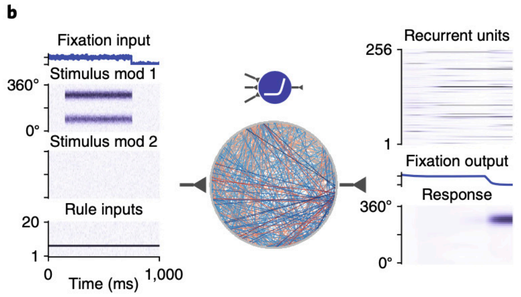
For sequence 5,2,6

Recurrent neural networks have internal memory





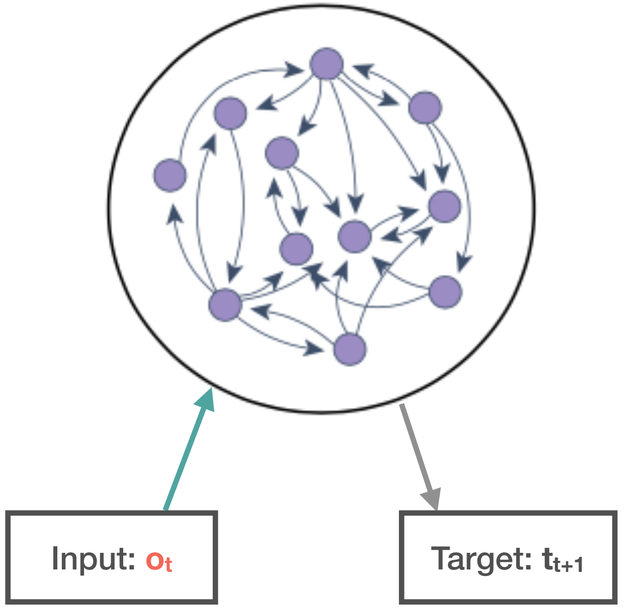
Mante et al., 2013

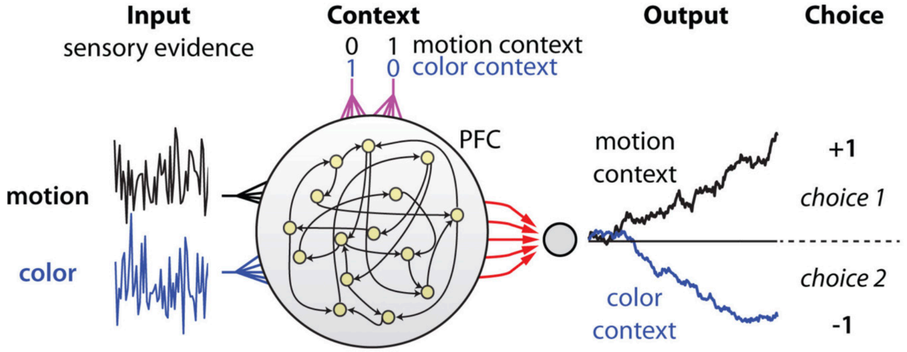


Yang et al., 2019

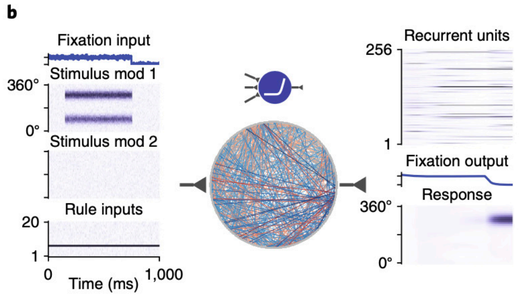
For sequence 5,2,6

Recurrent neural networks have internal memory

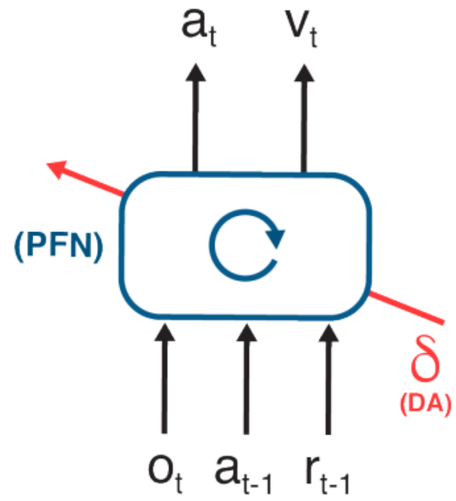




Mante et al., 2013



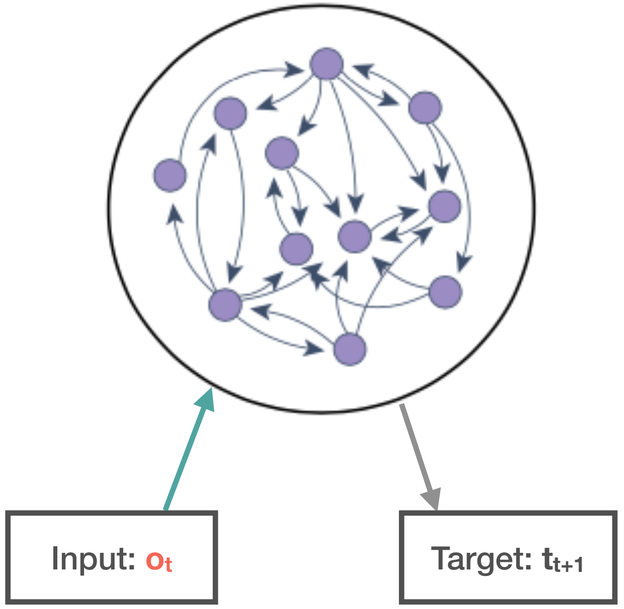
Yang et al., 2019

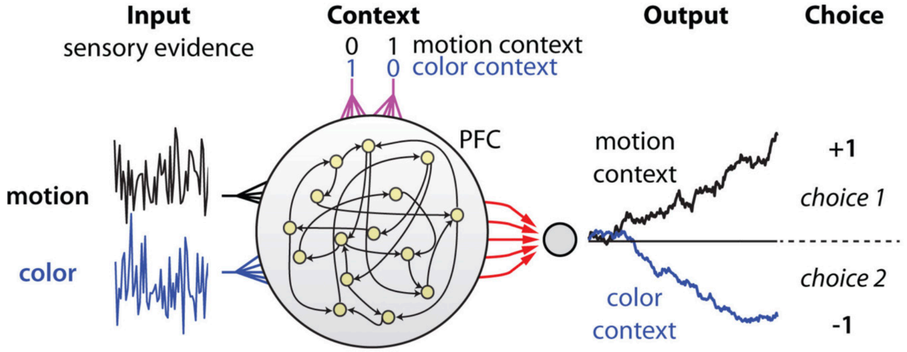


Wang et al., 2018

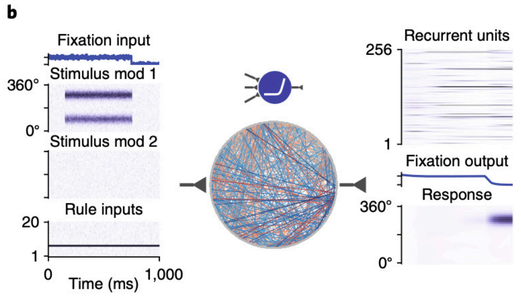
For sequence 5,2,6

Recurrent neural networks have internal memory

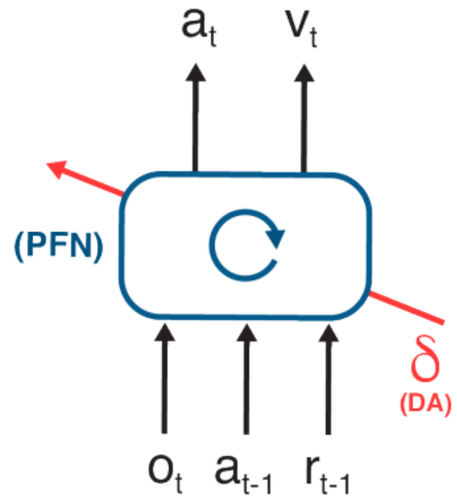




Mante et al., 2013



Yang et al., 2019

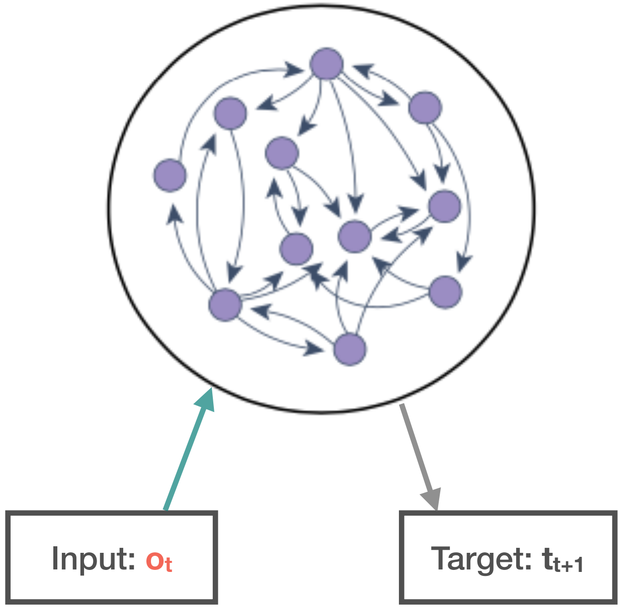


Wang et al., 2018

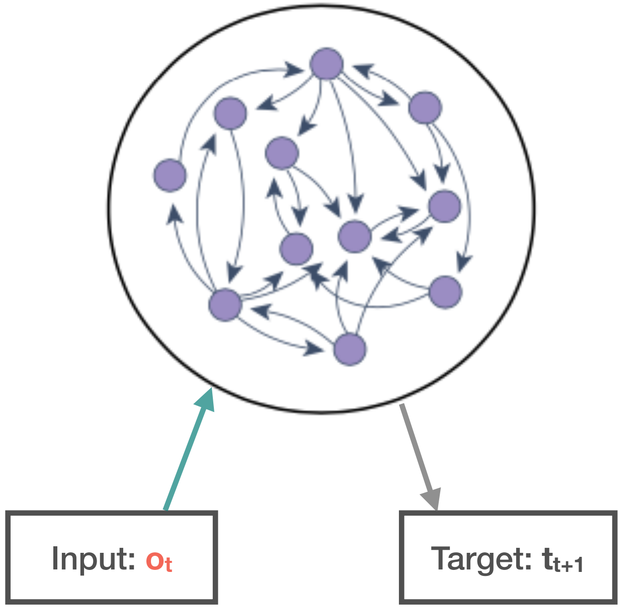
We want to understand the mechanism of these RNNs!

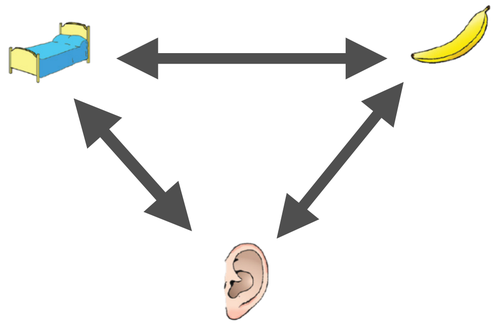
For sequence 5,2,6

Hypothesis: Sequence working memory using neural circuits of structured slots

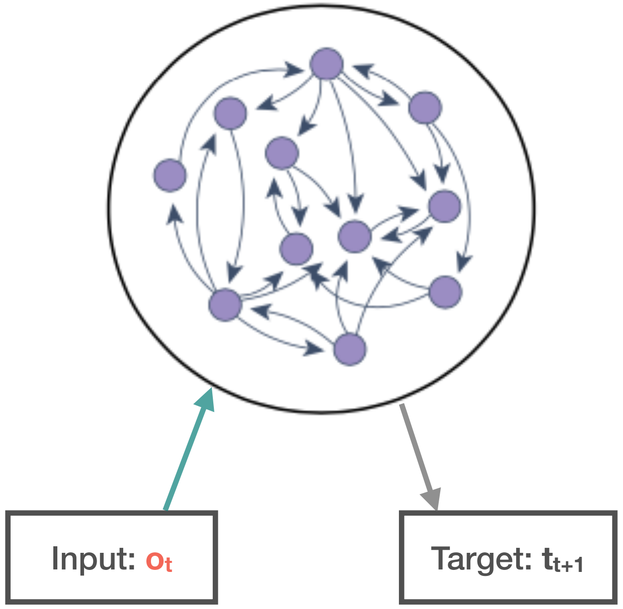


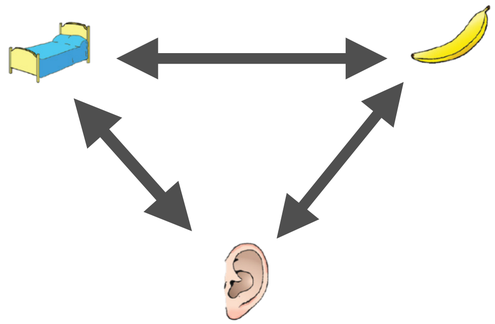
Hypothesis: Sequence working memory using neural circuits of structured slots

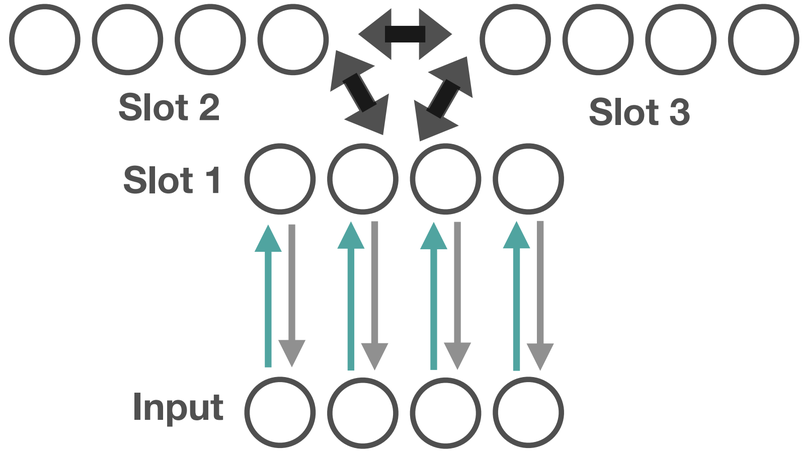




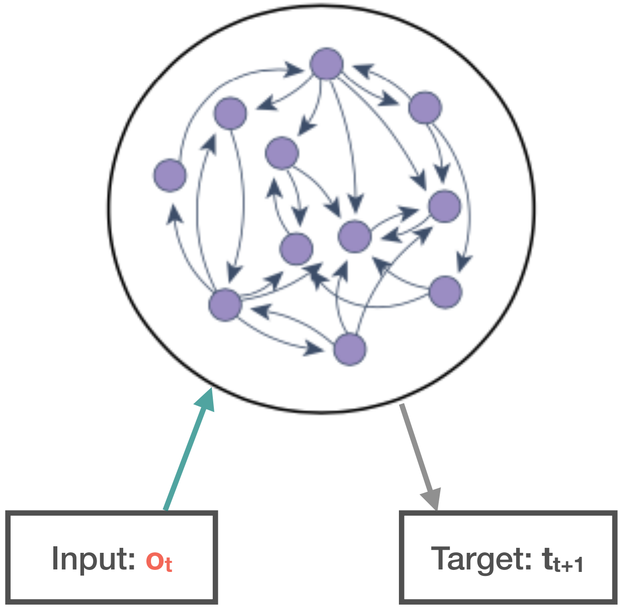
Hypothesis: Sequence working memory using neural circuits of structured slots

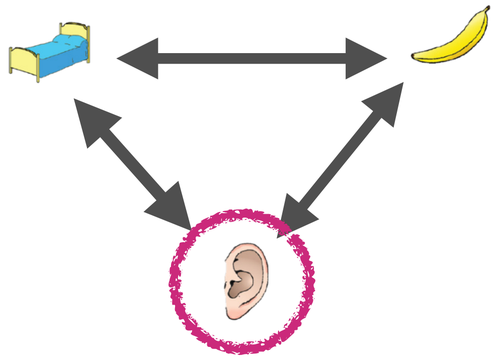


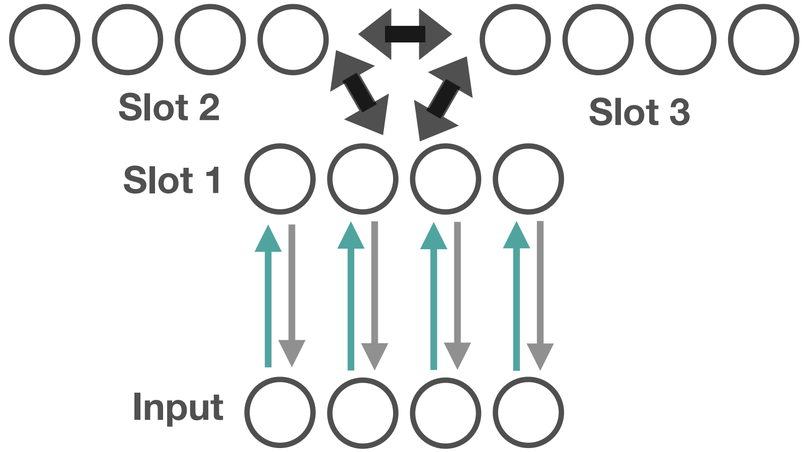




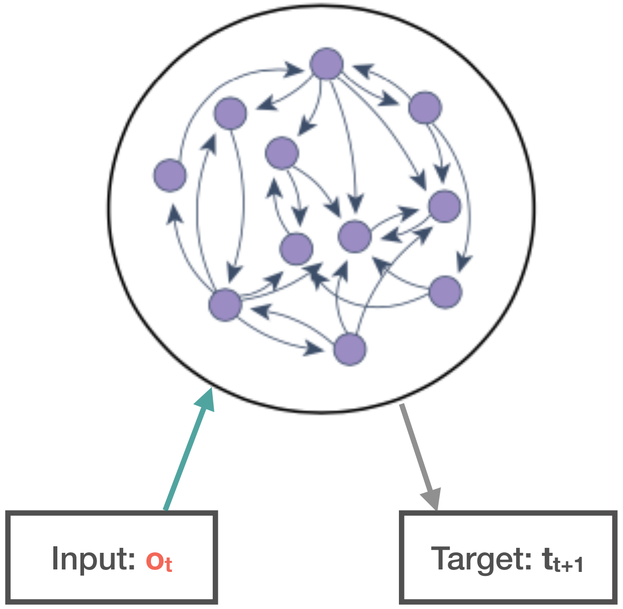
Hypothesis: Sequence working memory using neural circuits of structured slots

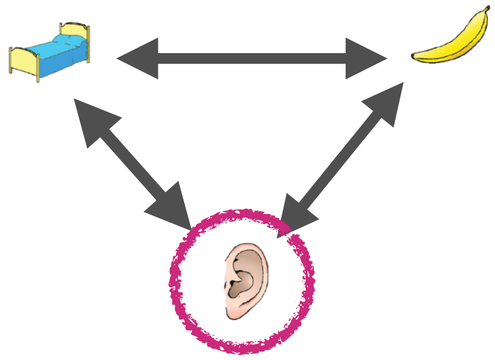


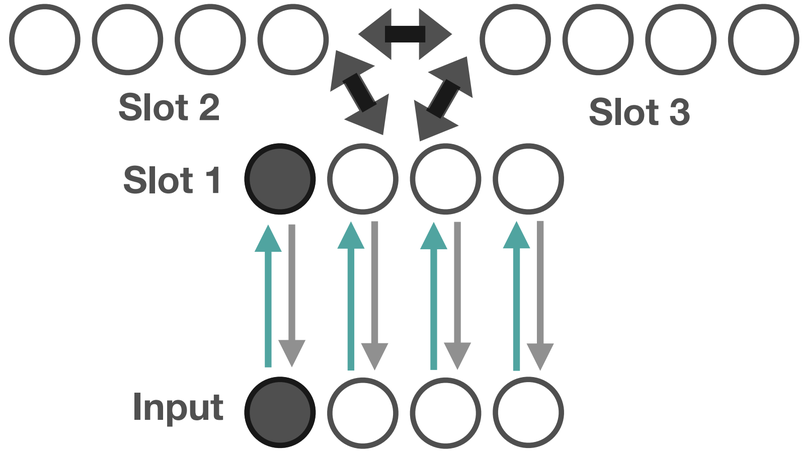




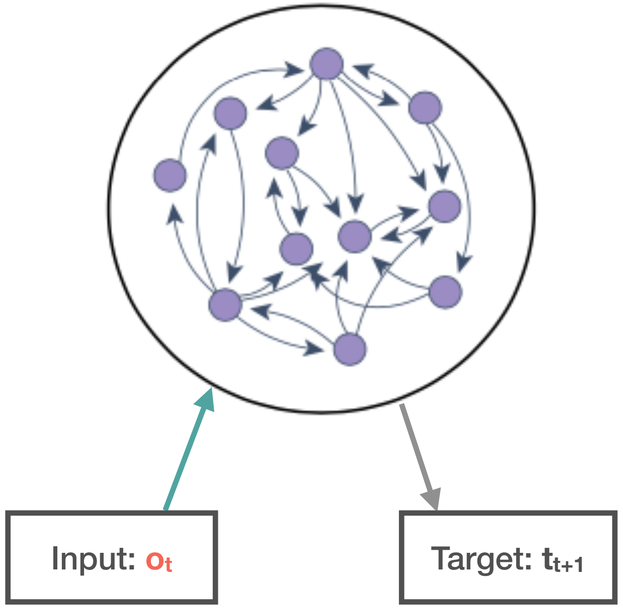
Hypothesis: Sequence working memory using neural circuits of structured slots

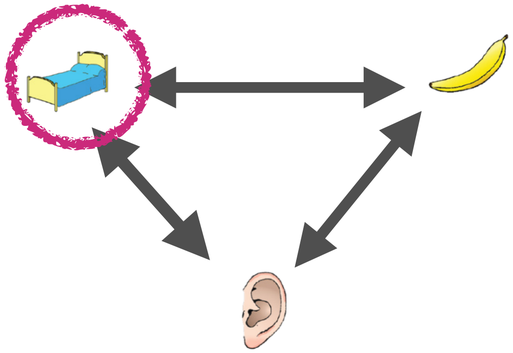


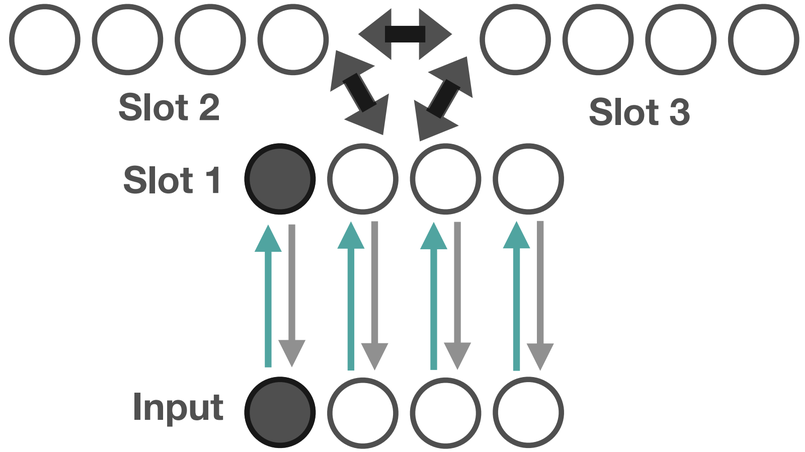




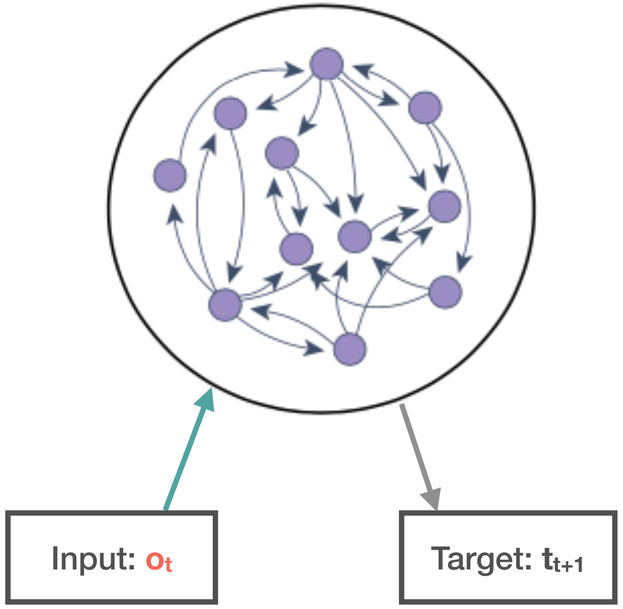
Hypothesis: Sequence working memory using neural circuits of structured slots

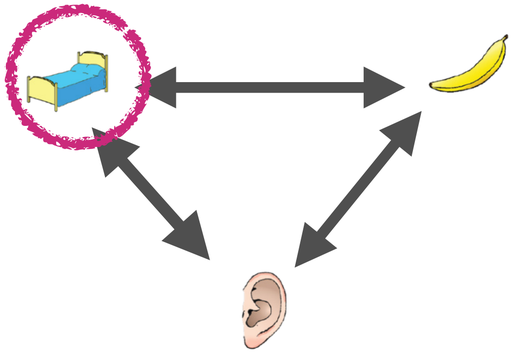


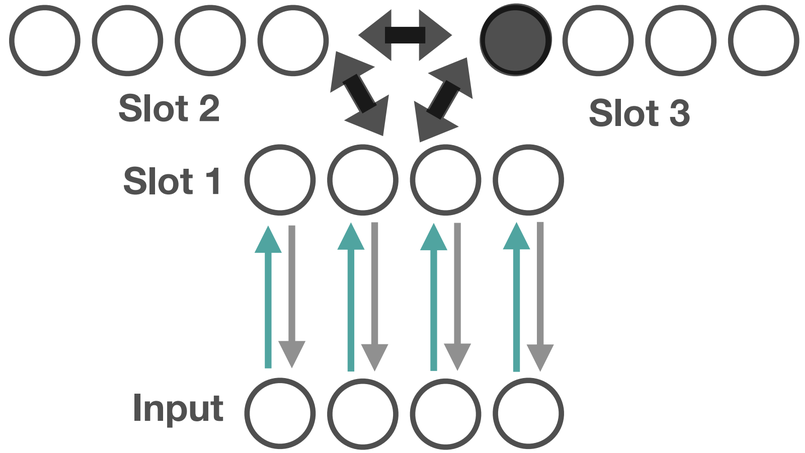




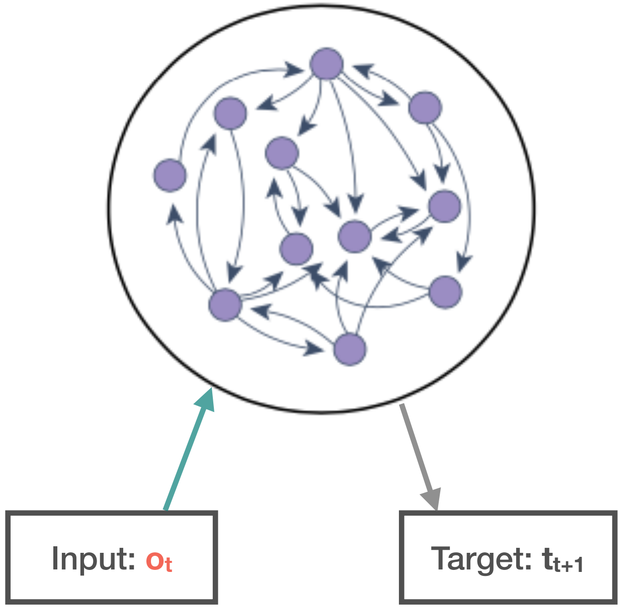
Hypothesis: Sequence working memory using neural circuits of structured slots

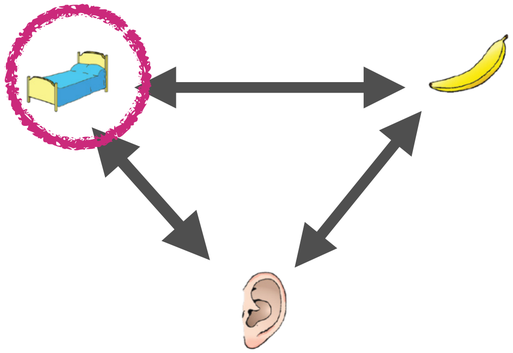


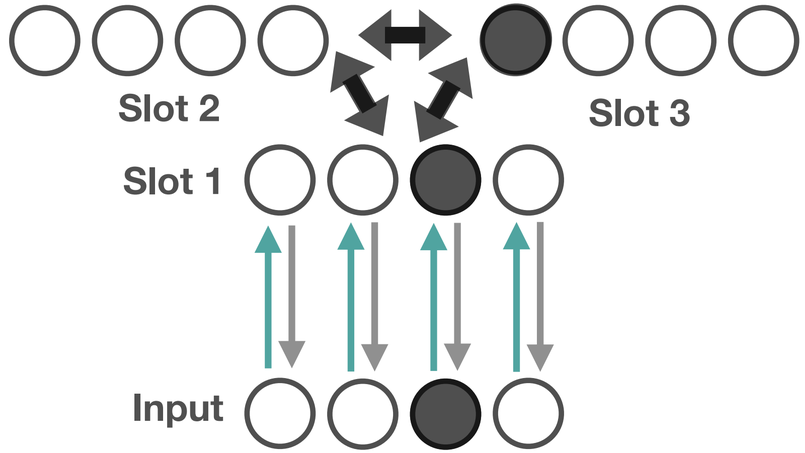




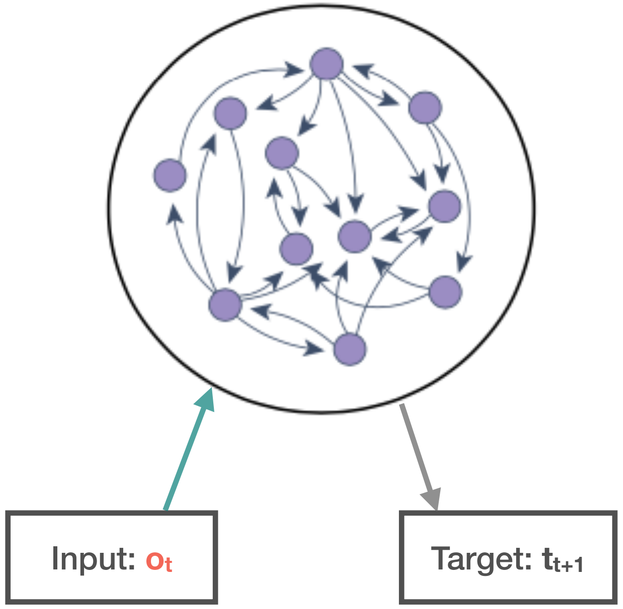
Hypothesis: Sequence working memory using neural circuits of structured slots

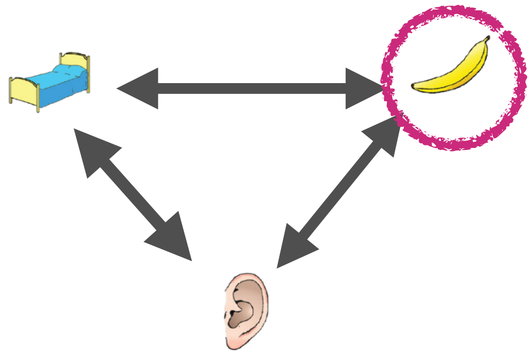


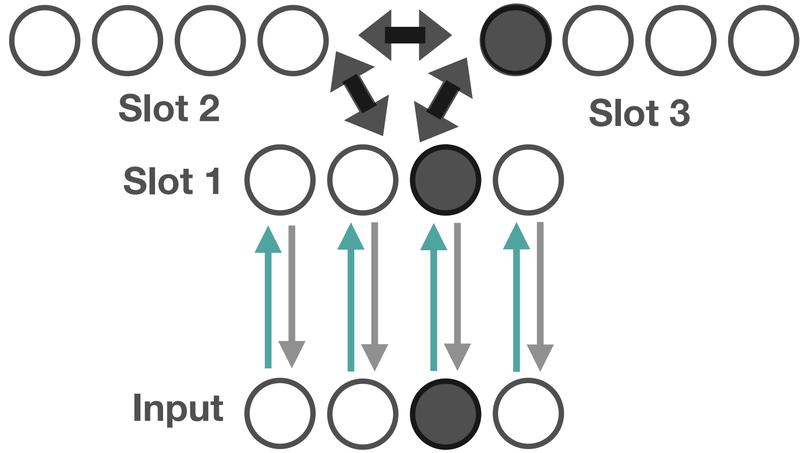




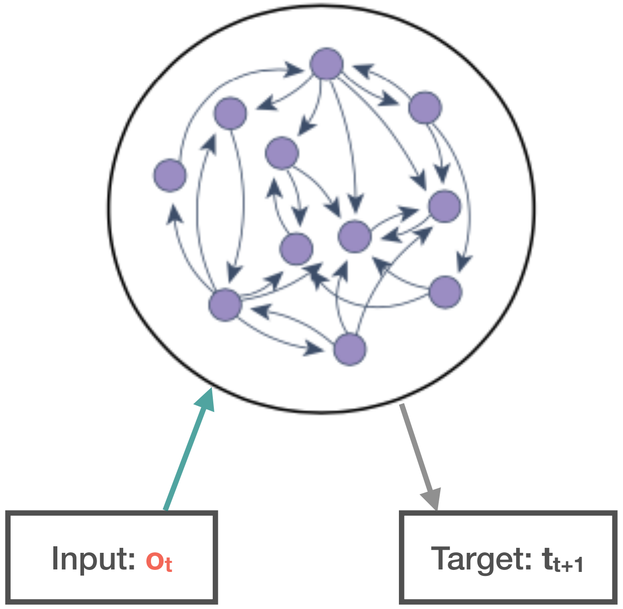
Hypothesis: Sequence working memory using neural circuits of structured slots

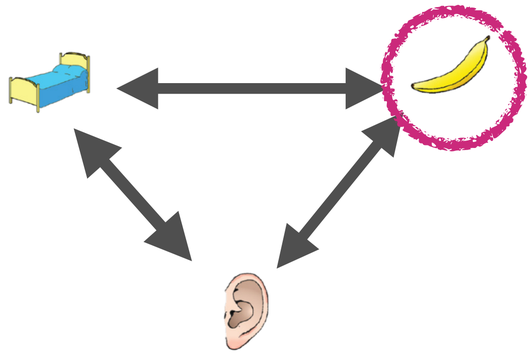


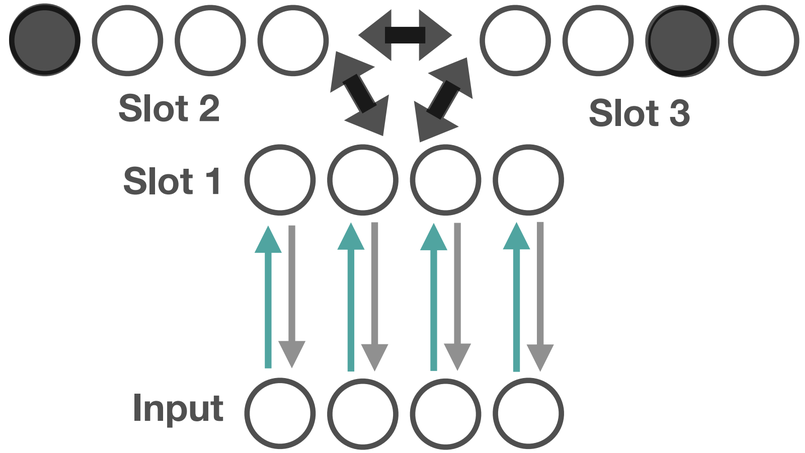




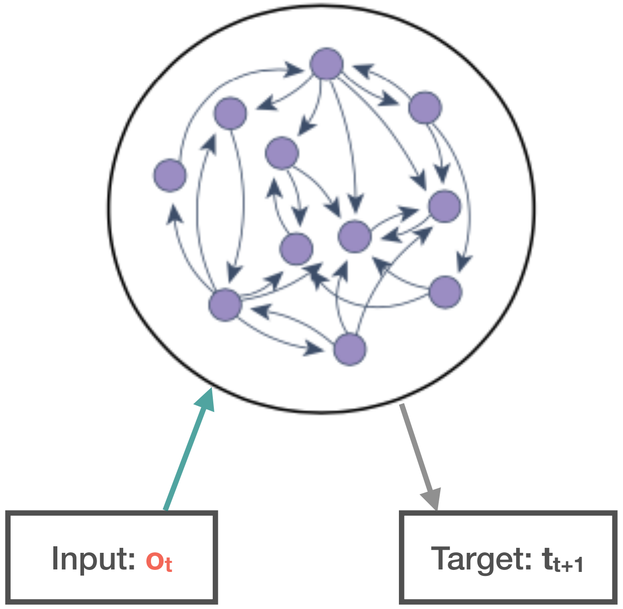
Hypothesis: Sequence working memory using neural circuits of structured slots

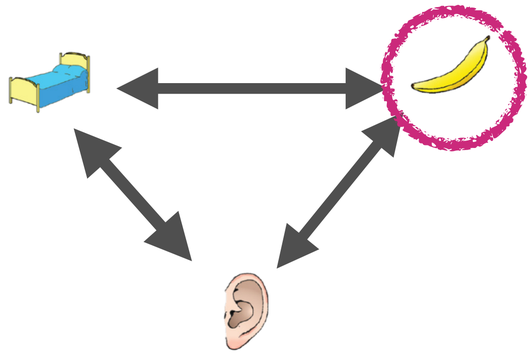


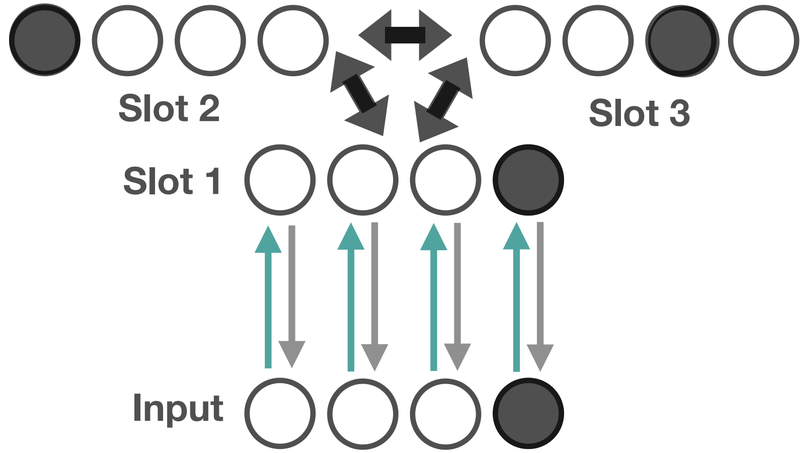




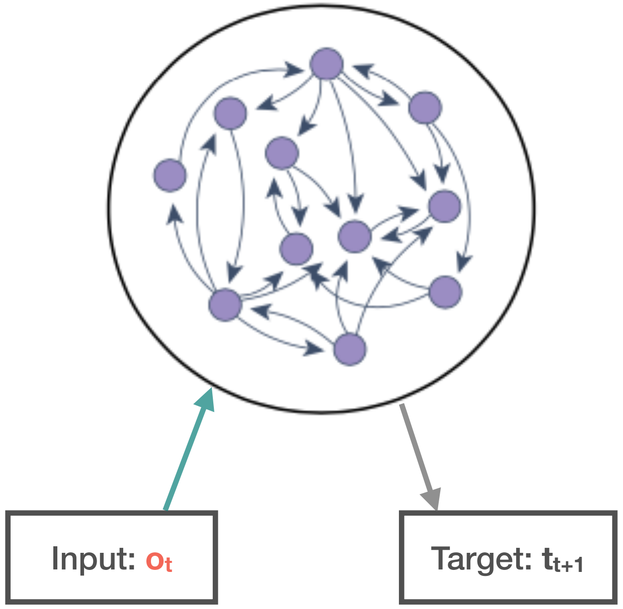
Hypothesis: Sequence working memory using neural circuits of structured slots

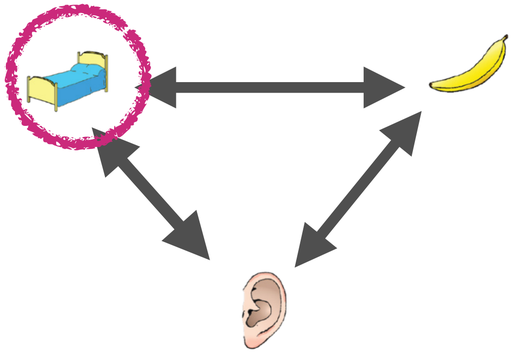


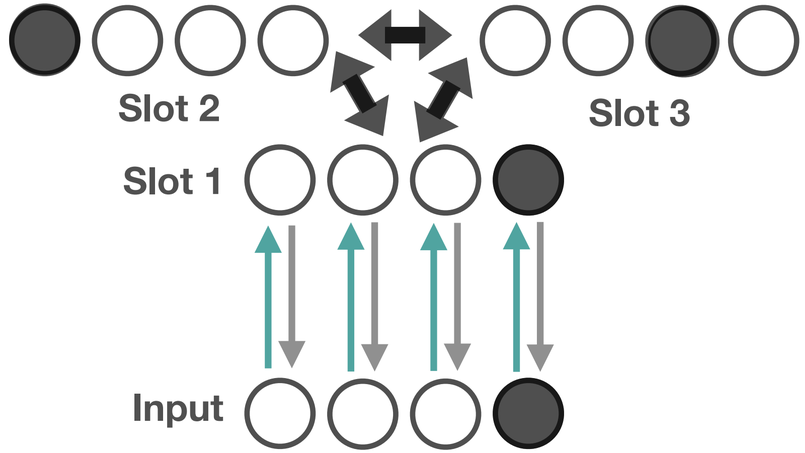




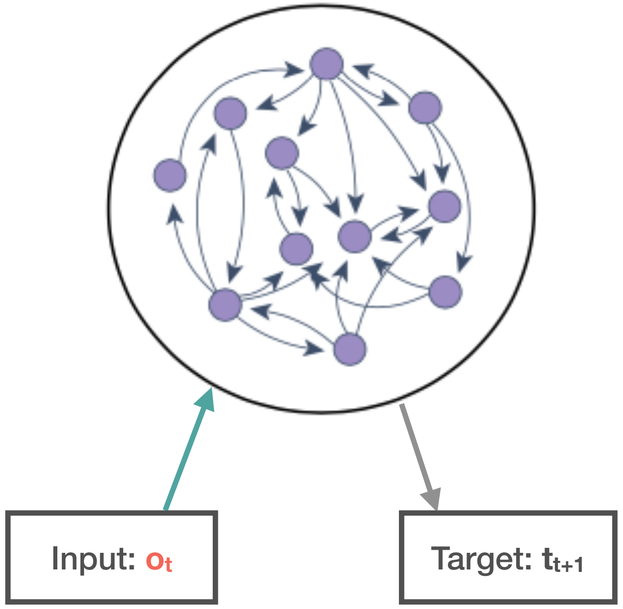
Hypothesis: Sequence working memory using neural circuits of structured slots

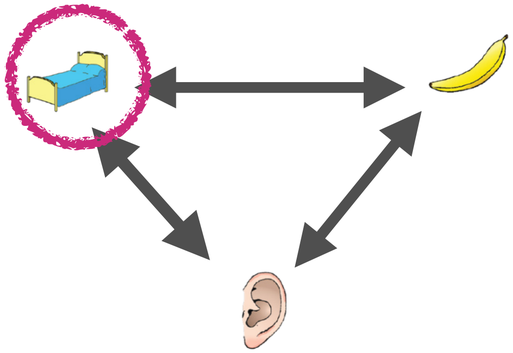


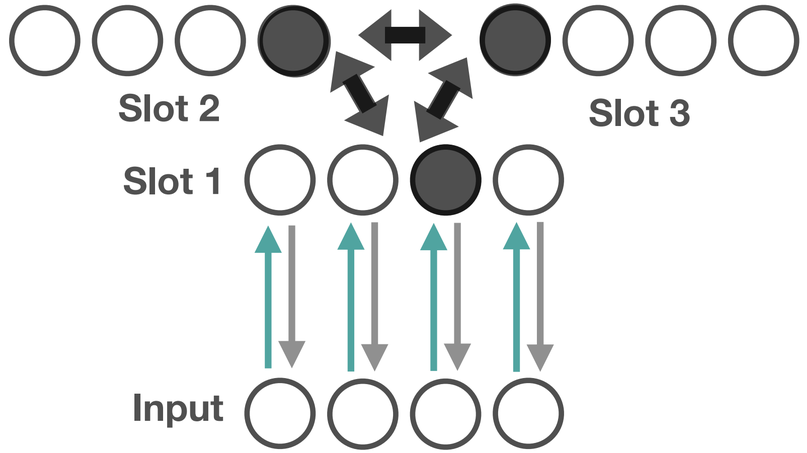




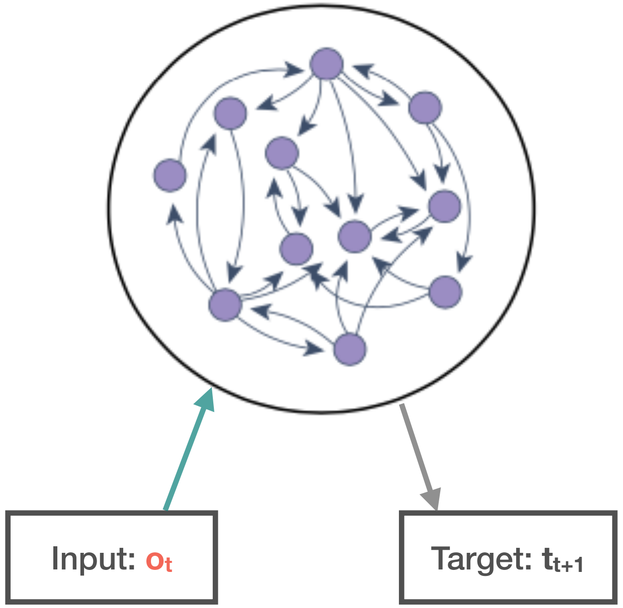
Hypothesis: Sequence working memory using neural circuits of structured slots

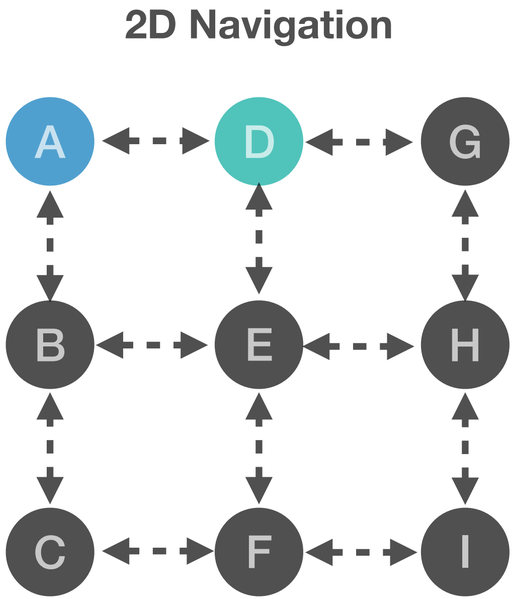


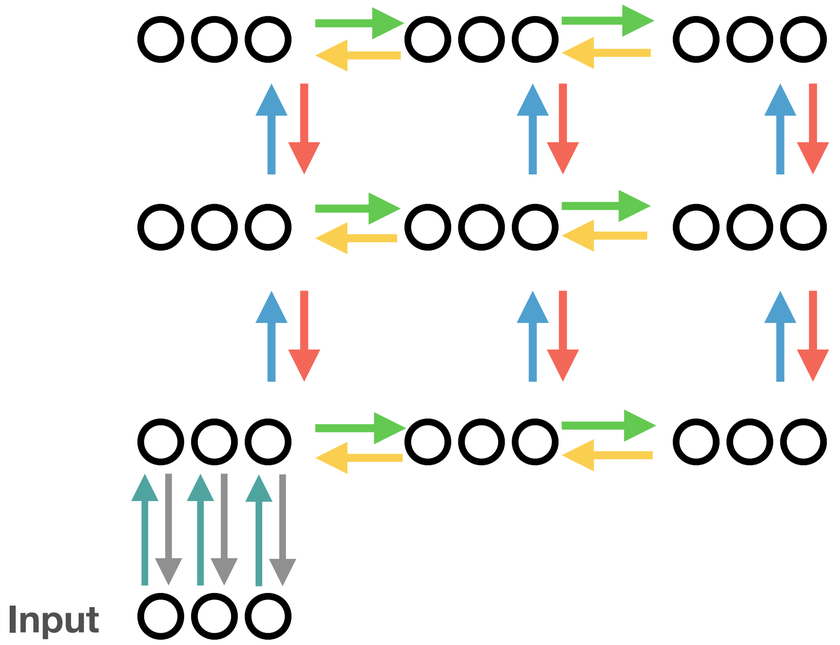




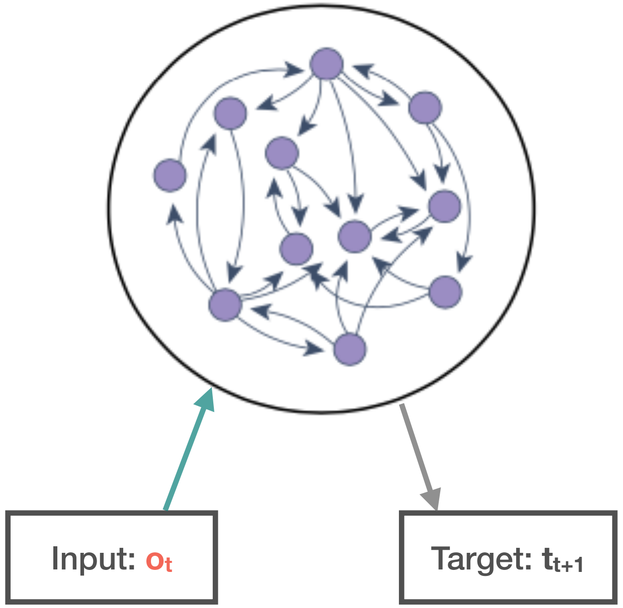
Different problems have different slot structure

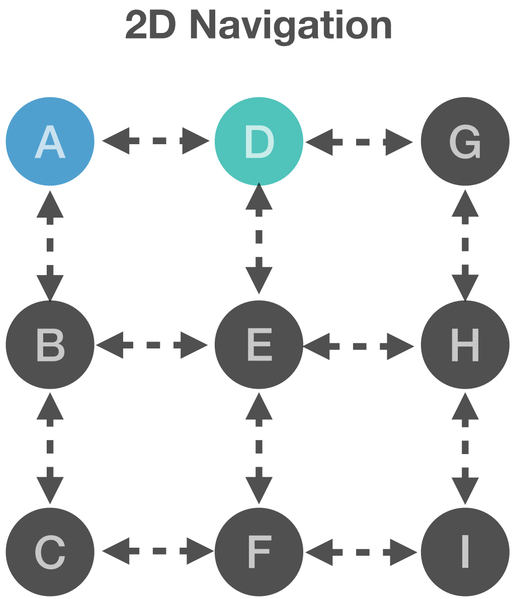


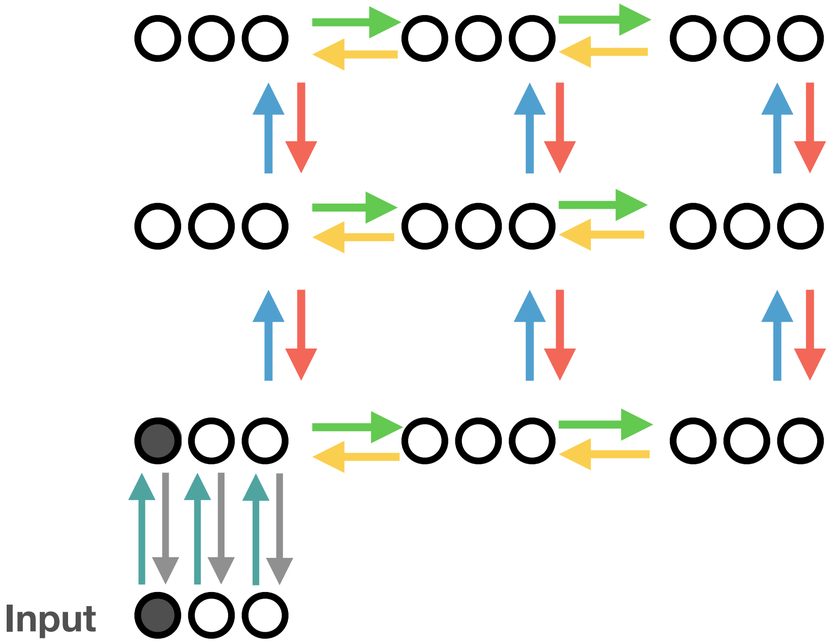




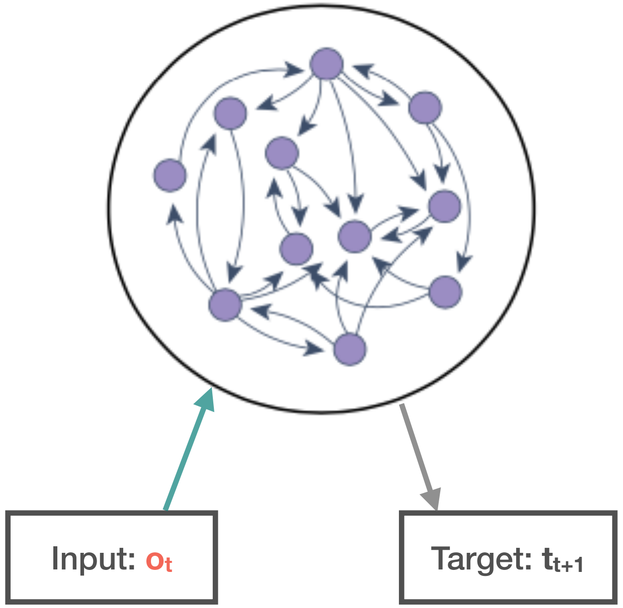
Different problems have different slot structure

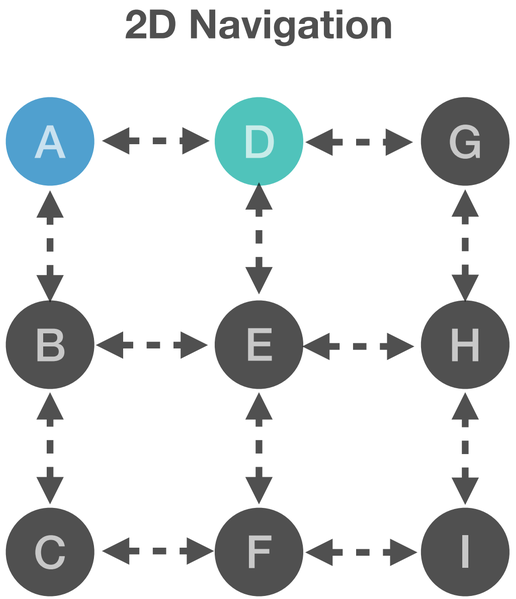


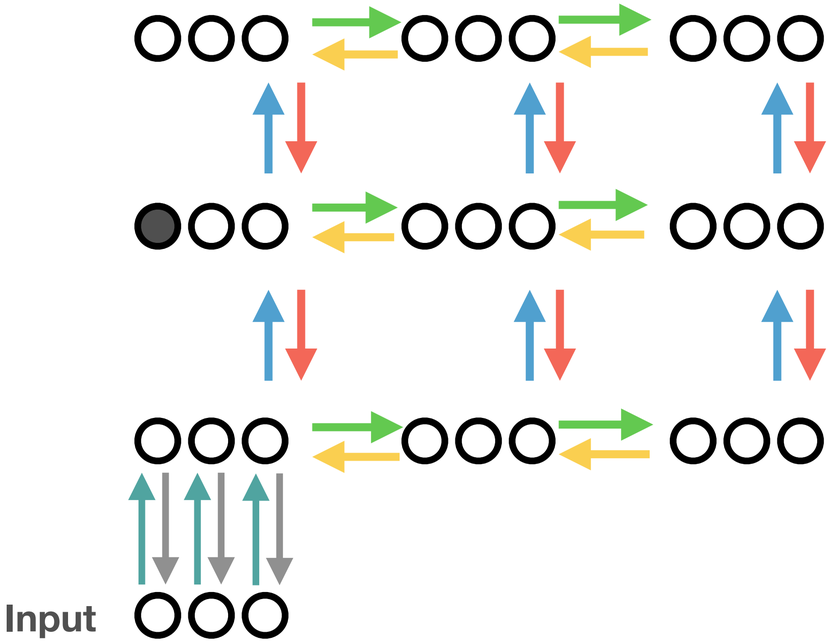




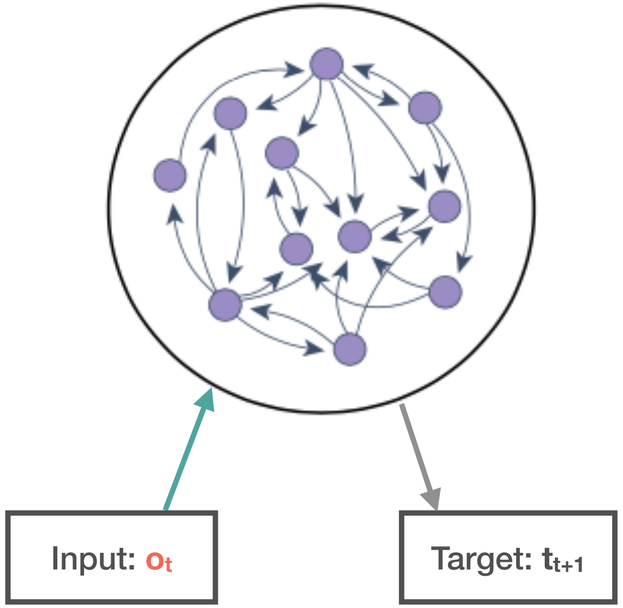
Different problems have different slot structure

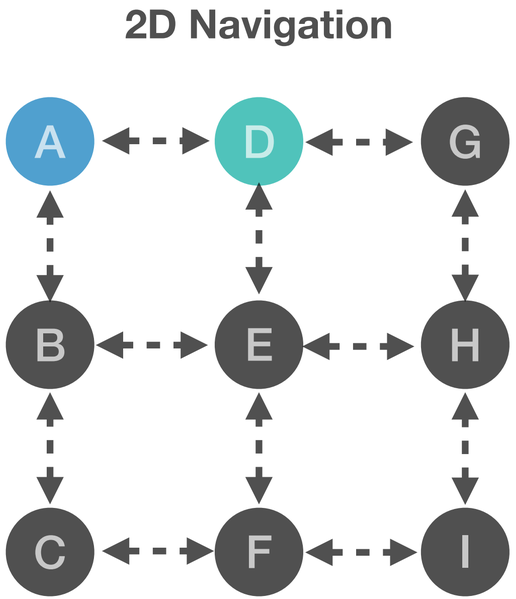


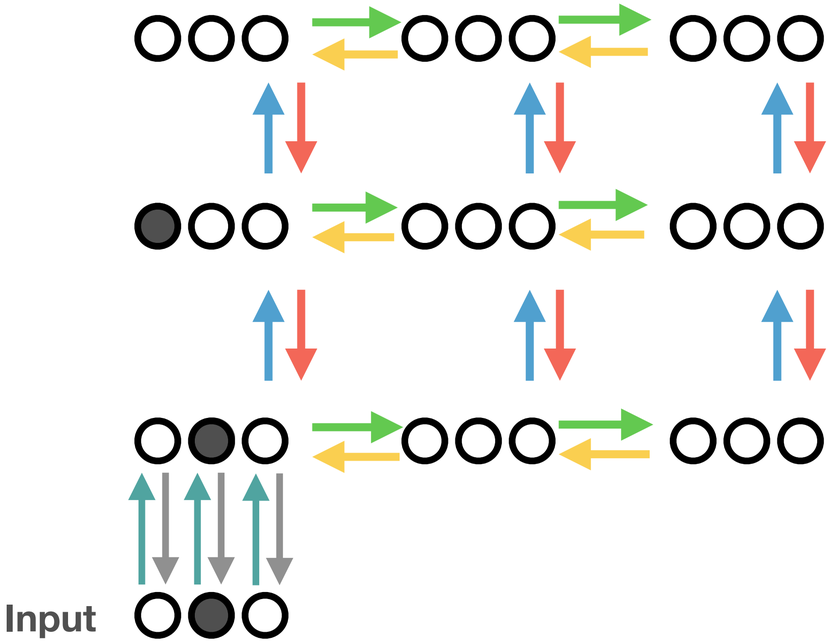




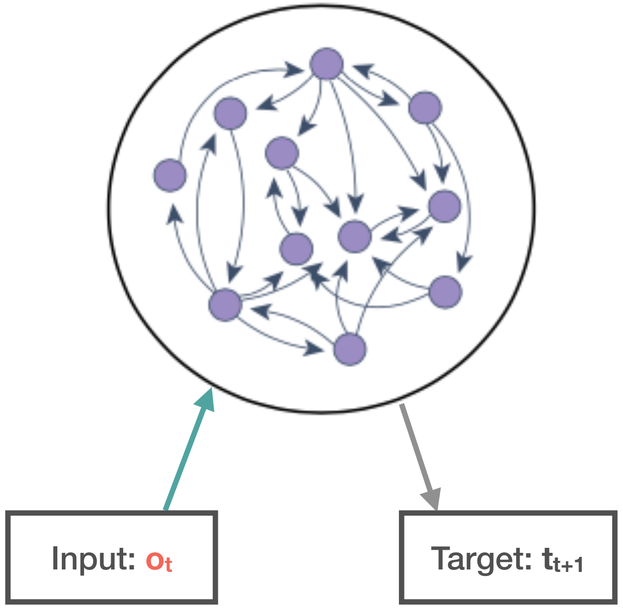
Different problems have different slot structure

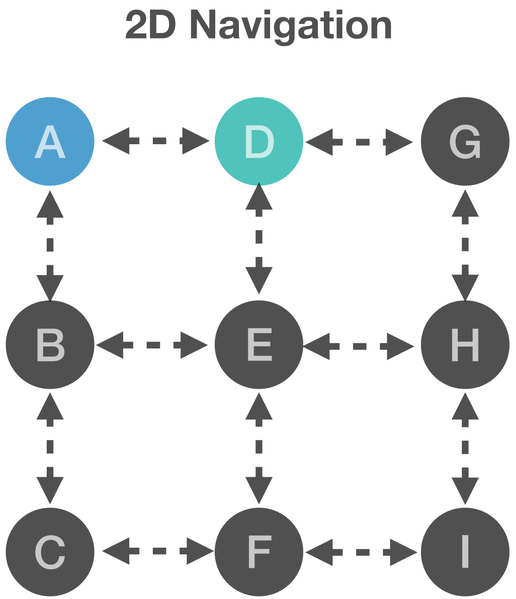


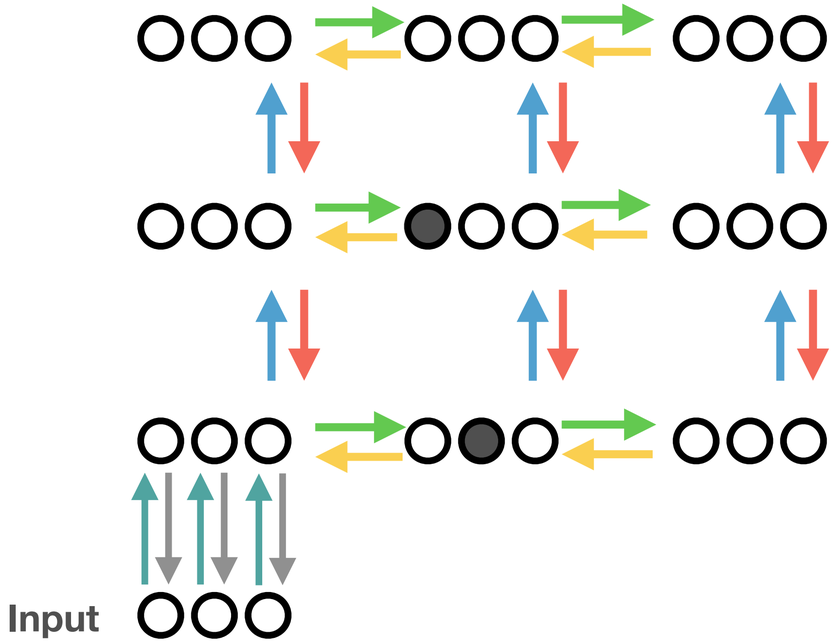




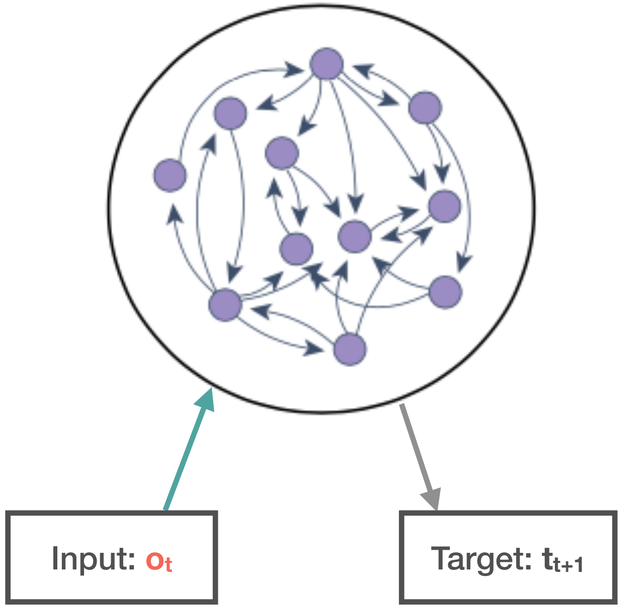
Different problems have different slot structure

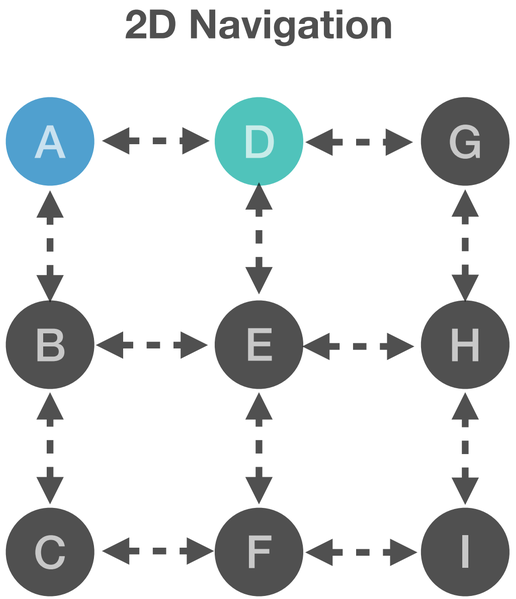


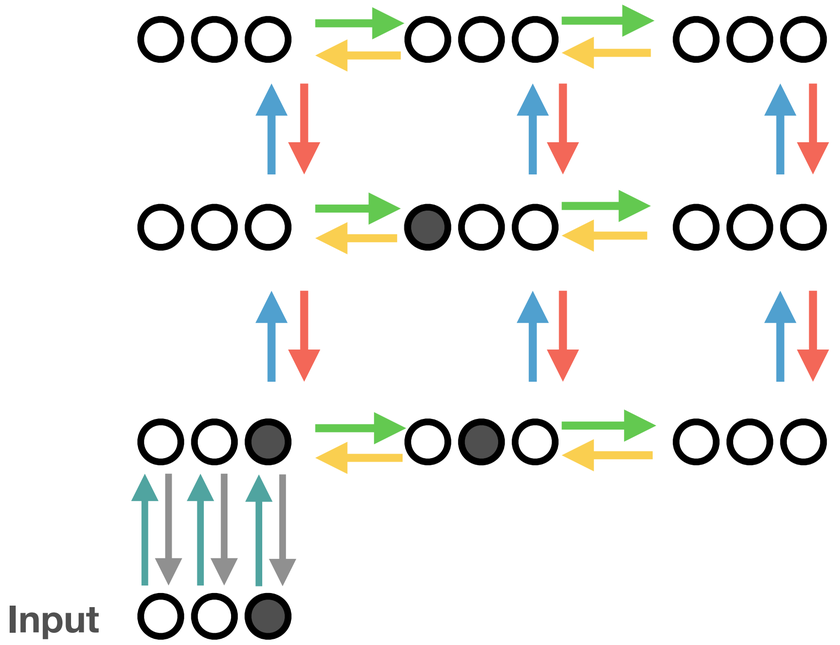




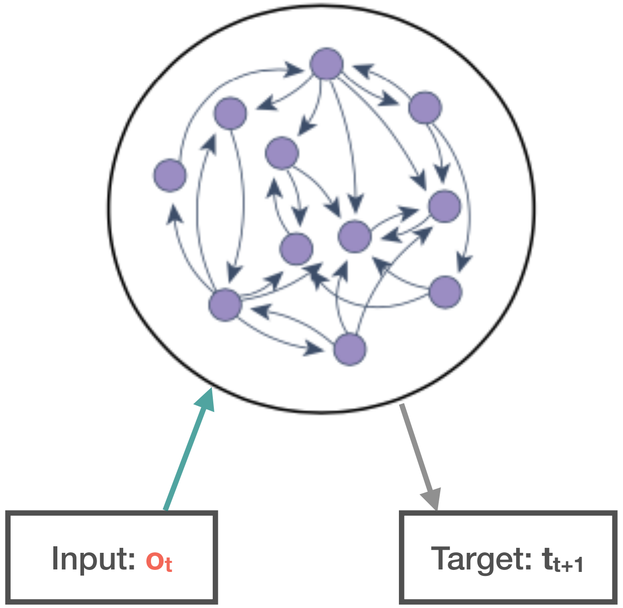
Different problems have different slot structure

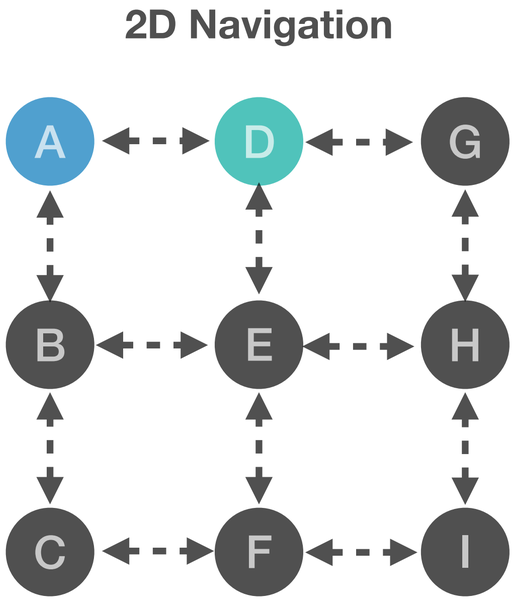


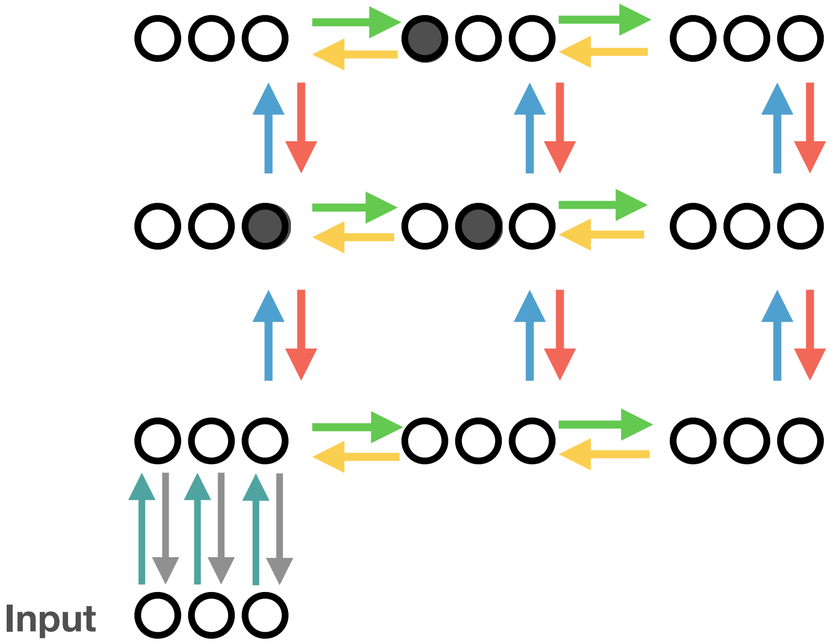




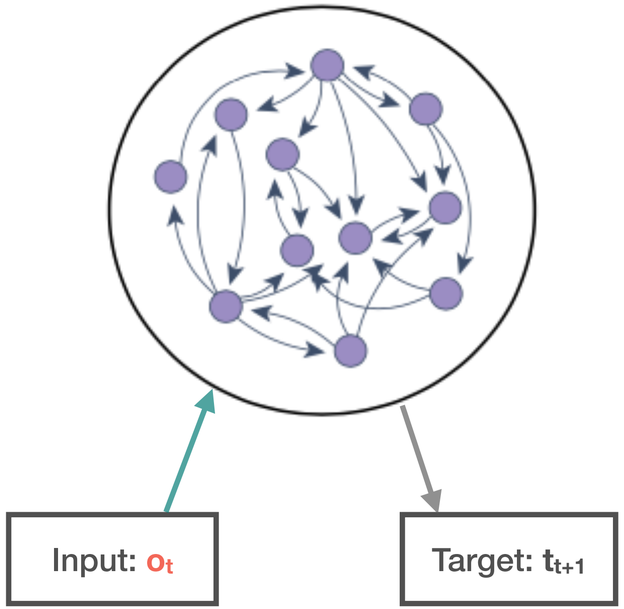
Different problems have different slot structure



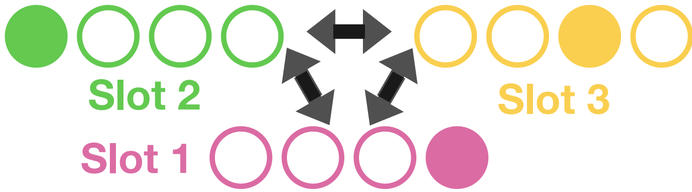




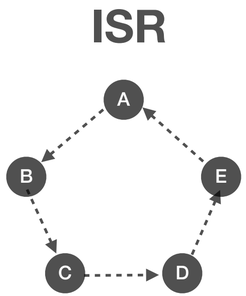
Do RNNs actually learn this mechanistic model?



?

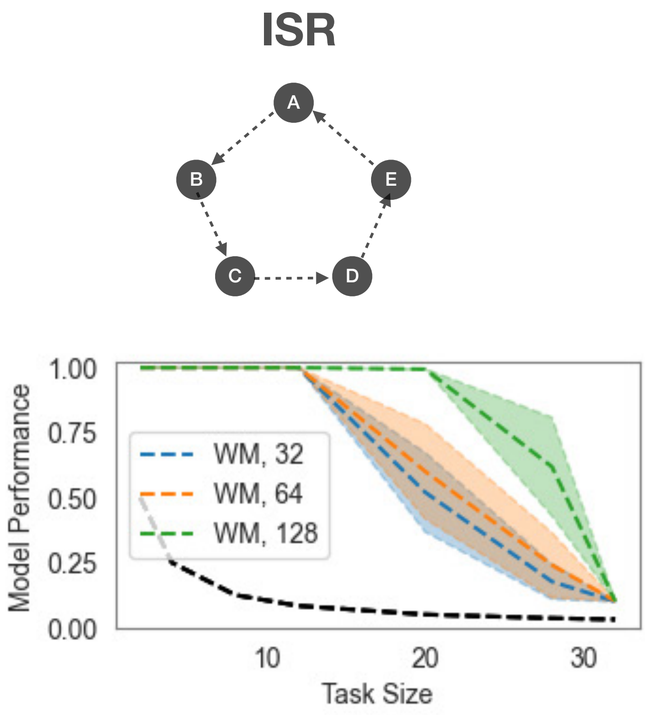


Frist, RNNs can actually solve the task



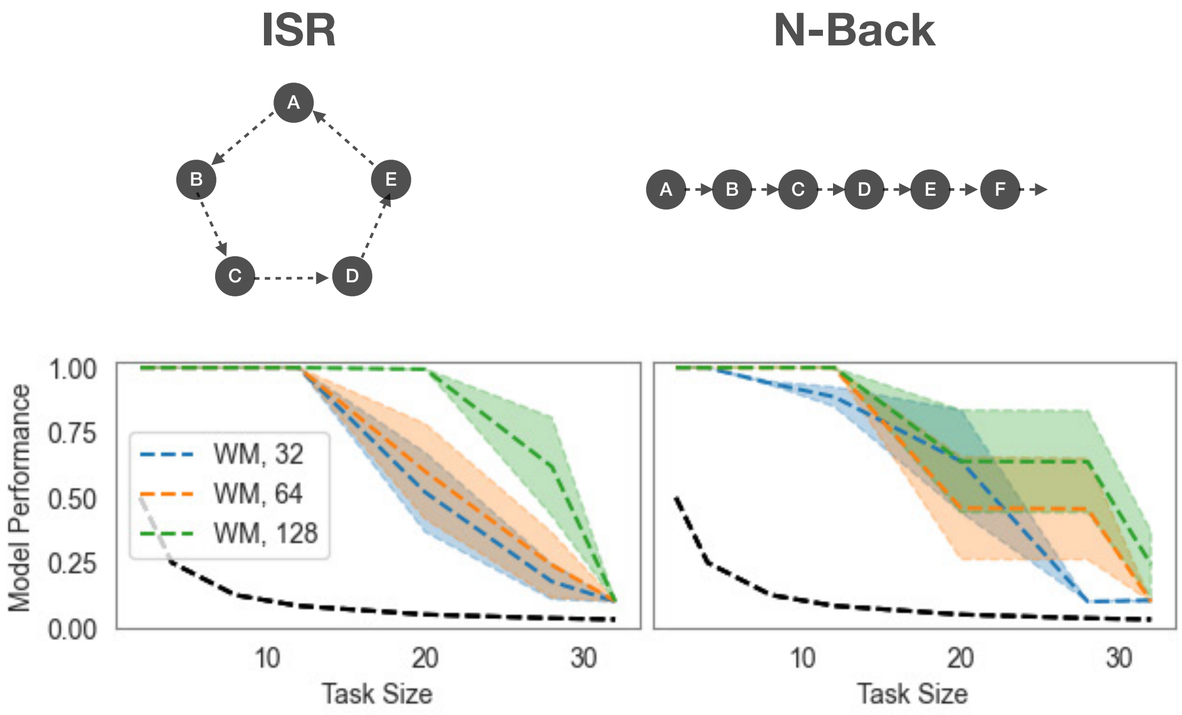
Whittington et al., 2023, *bioRxiv*

Frist, RNNs can actually solve the task



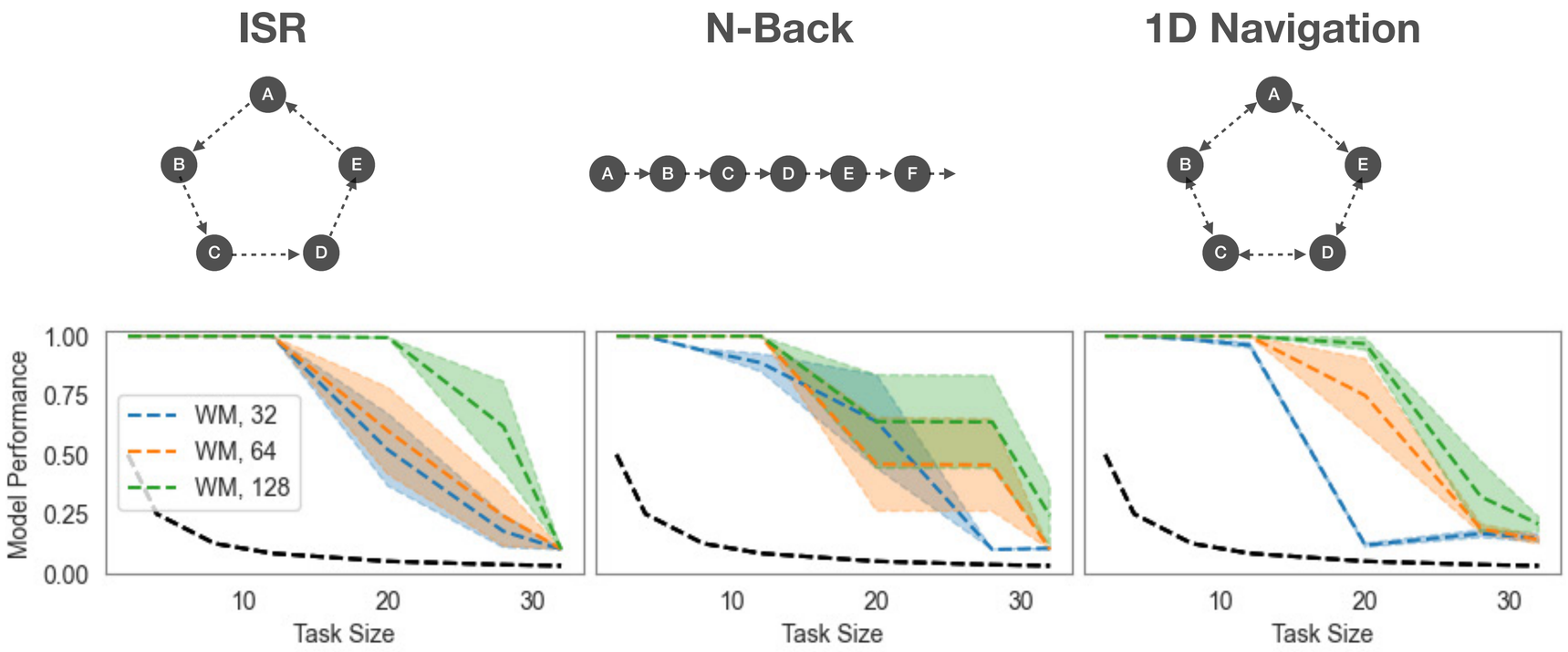
Whittington et al., 2023, *bioRxiv*

Frist, RNNs can actually solve the task



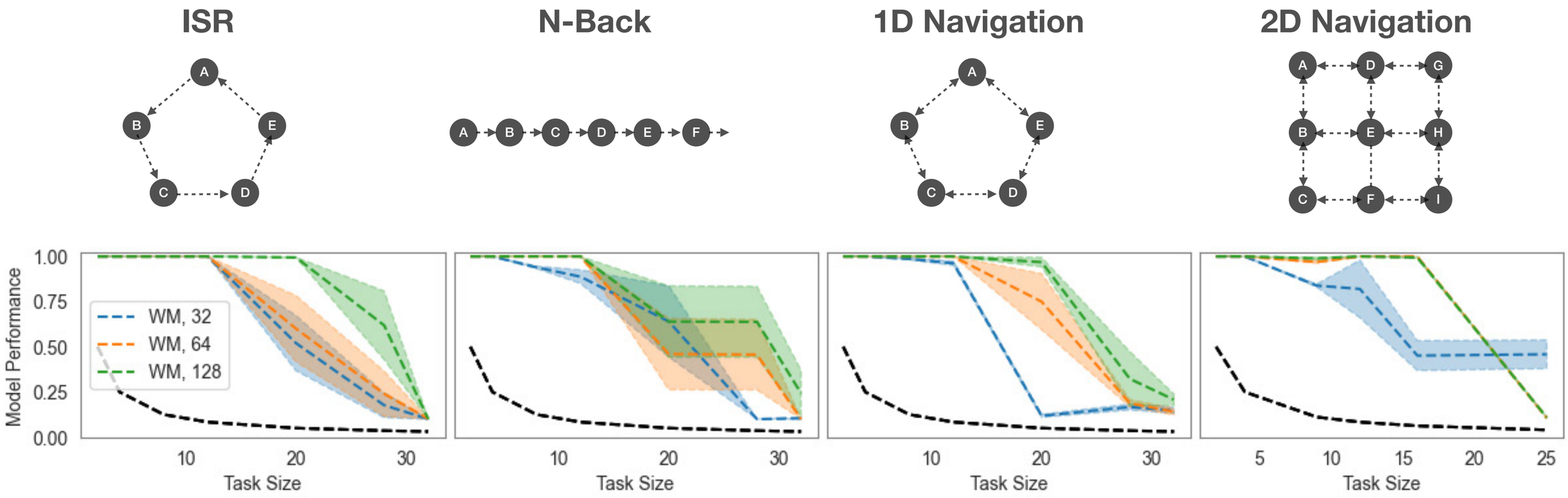
Whittington et al., 2023, *bioRxiv*

Frist, RNNs can actually solve the task



Whittington et al., 2023, *bioRxiv*

Frist, RNNs can actually solve the task



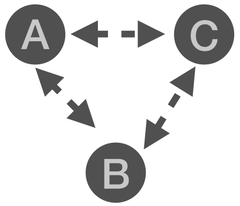
Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

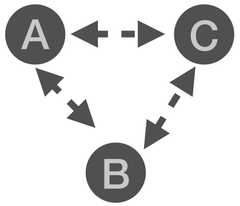
1D Navigation



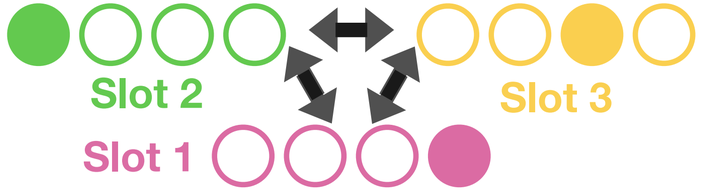
Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

1D Navigation



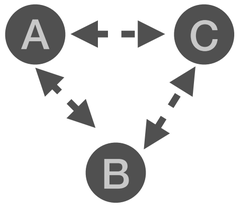
1D slot prediction



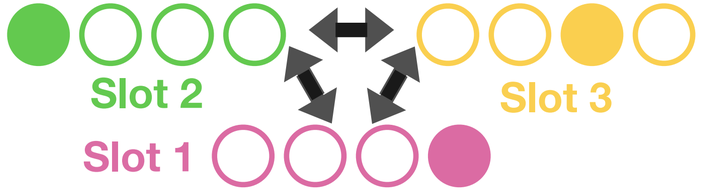
Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

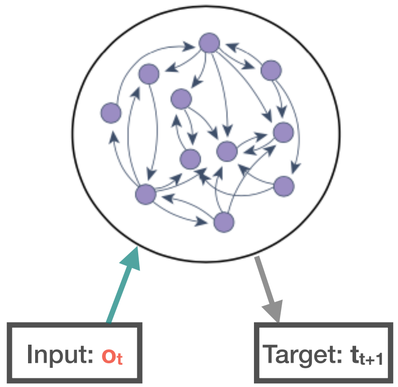
1D Navigation



1D slot prediction



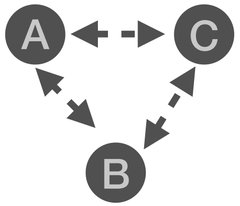
Should be able to decode the predicted contents of each slot at every tilmestep



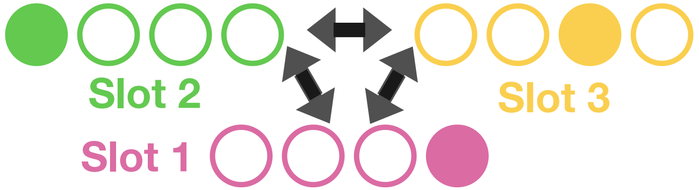
Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

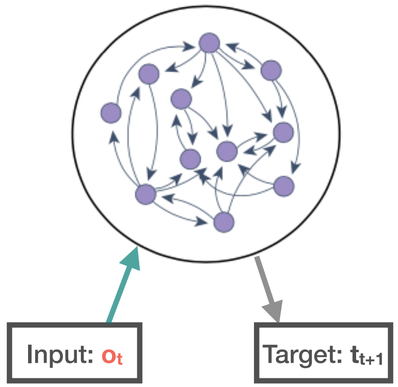
1D Navigation

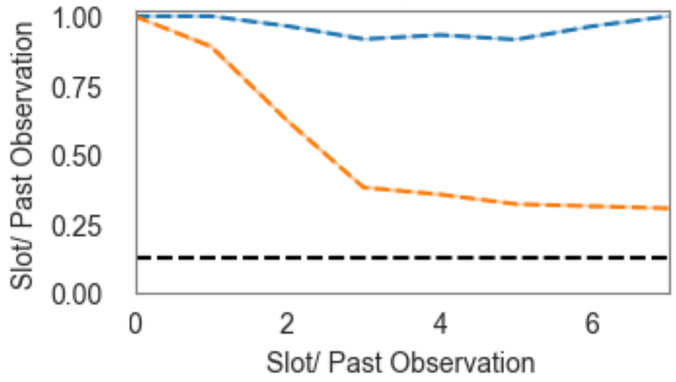


1D slot prediction



Should be able to decode the predicted contents of each slot at every tilmestep

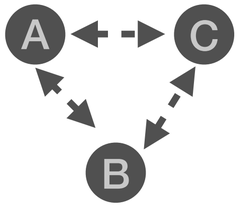




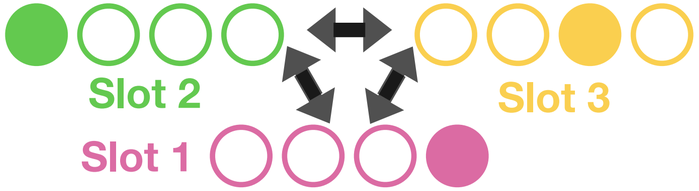
Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

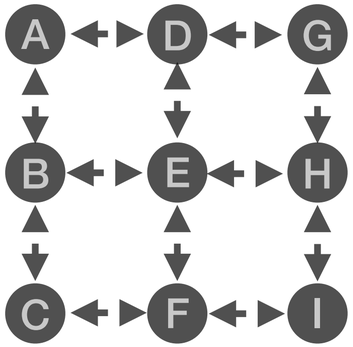
1D Navigation



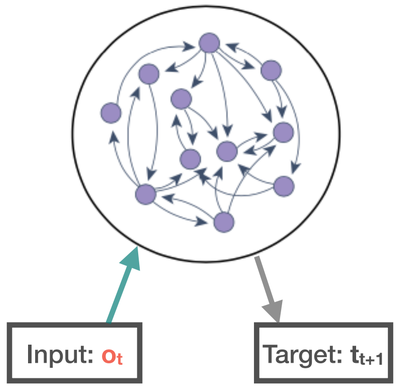
1D slot prediction

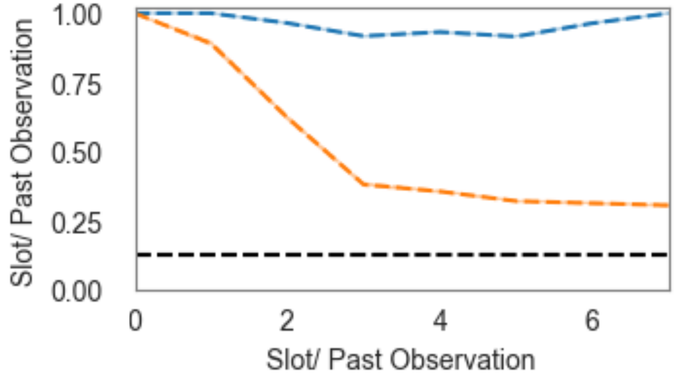


2D Navigation



Should be able to decode the predicted contents of each slot at every tilmestep

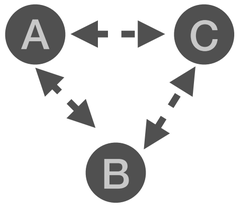




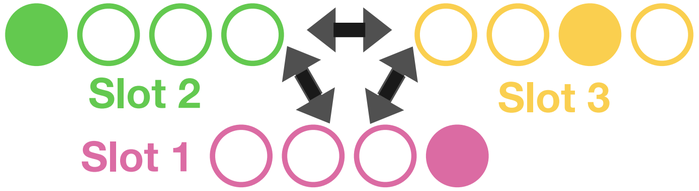
Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

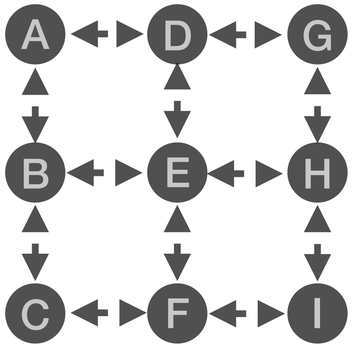
1D Navigation



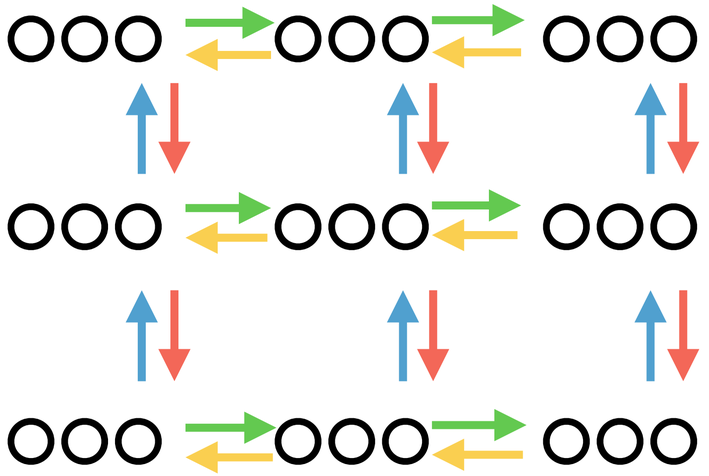
1D slot prediction



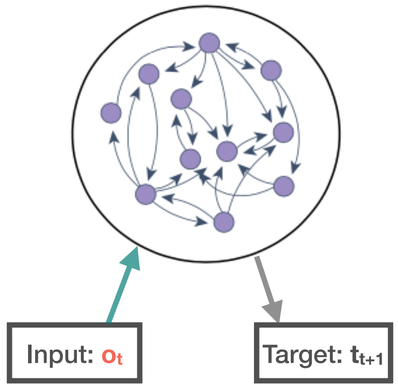
2D Navigation

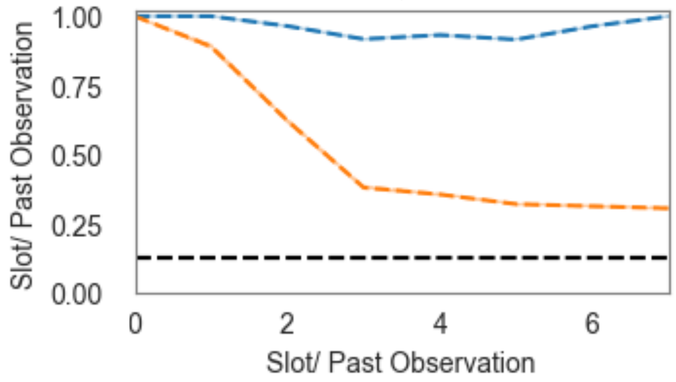


2D slot prediction



Should be able to decode the predicted contents of each slot at every tilmestep

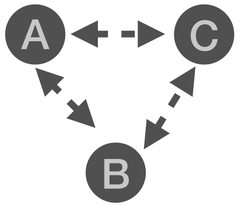




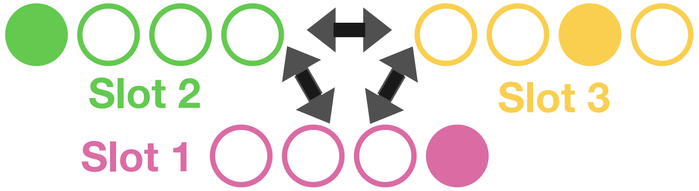
Whittington et al., 2023, *bioRxiv*

RNNs do use structured slots

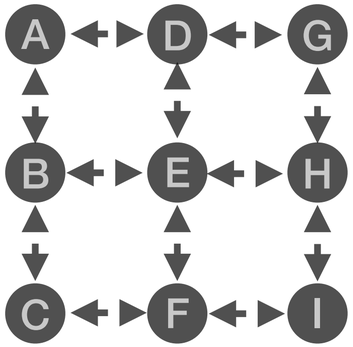
1D Navigation



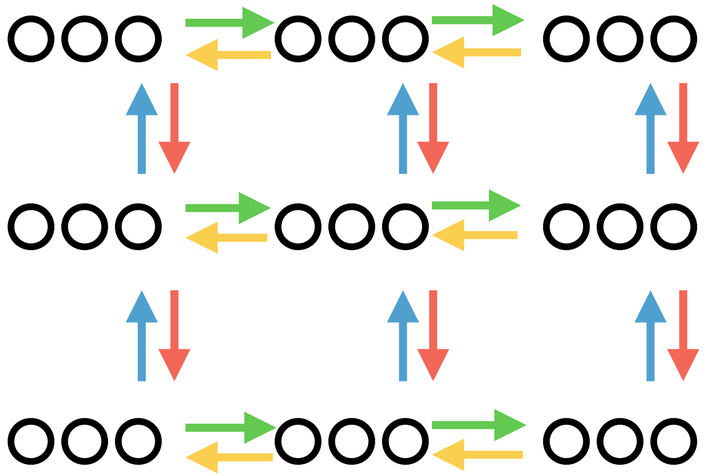
1D slot prediction



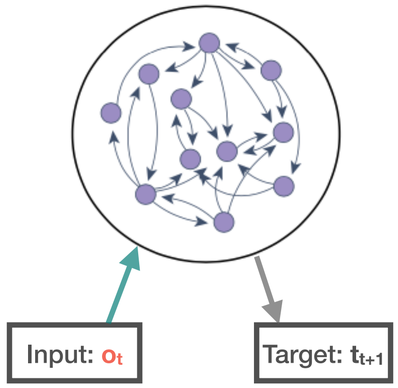
2D Navigation

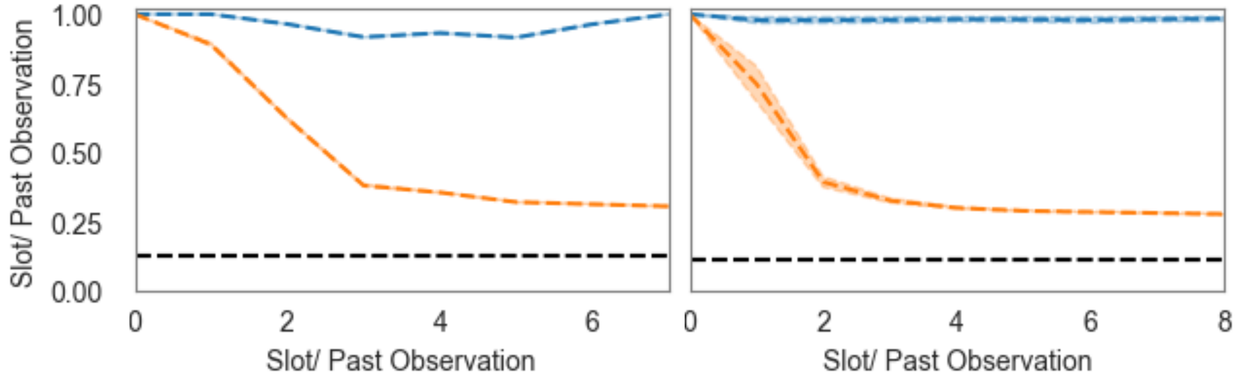


2D slot prediction



Should be able to decode the predicted contents of each slot at every tilmestep





Whittington et al., 2023, *bioRxiv*

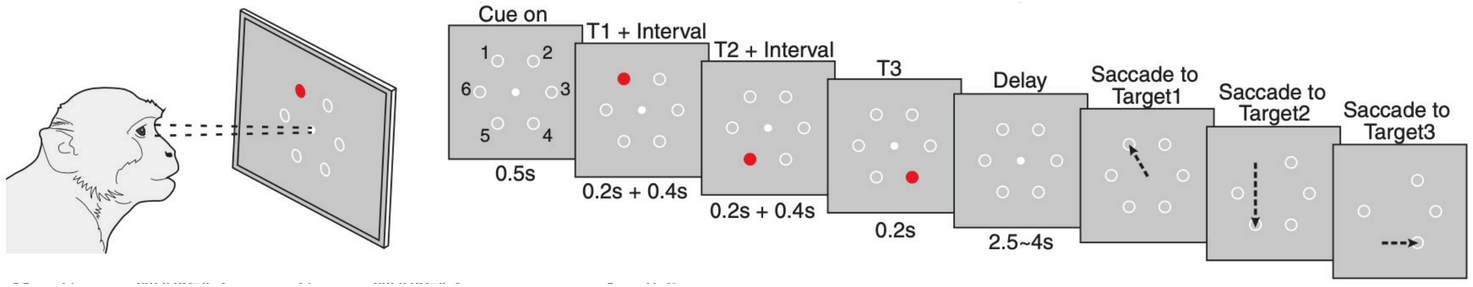
Does this match mechanistic interpretation match PFC data?

A slot based understanding unifies many representations in prefrontal cortex

Xie et al., 2022

Whittington et al., 2023, *bioRxiv*

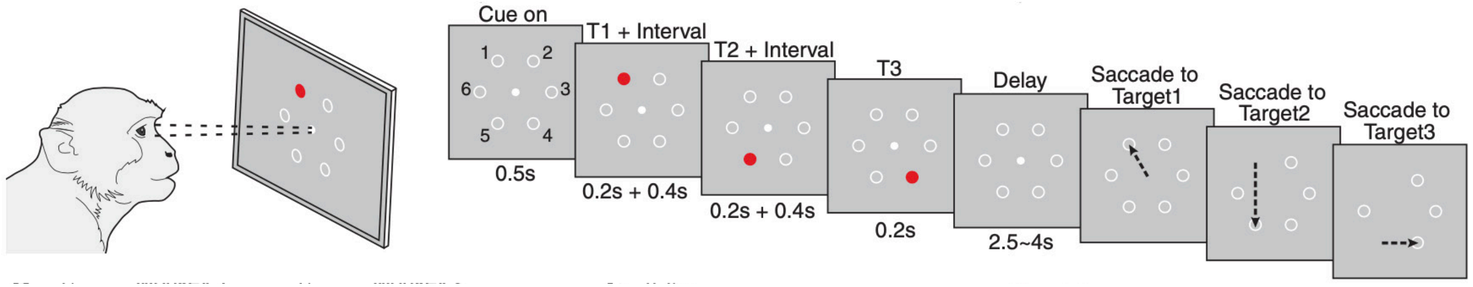
A slot based understanding unifies many representations in prefrontal cortex



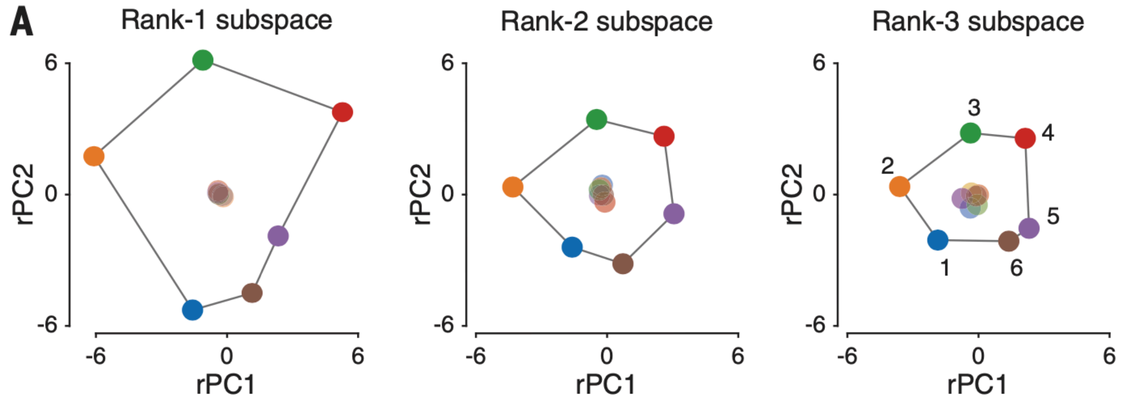
Xie et al., 2022

Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex

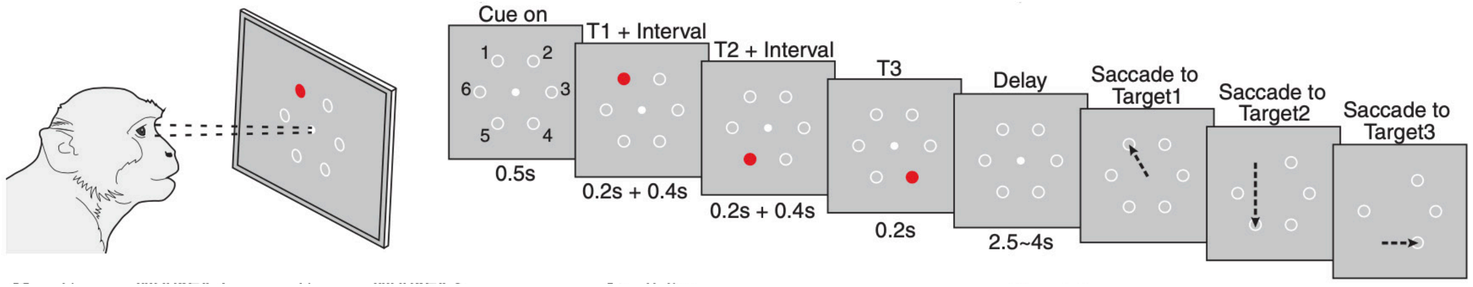


Xie et al., 2022

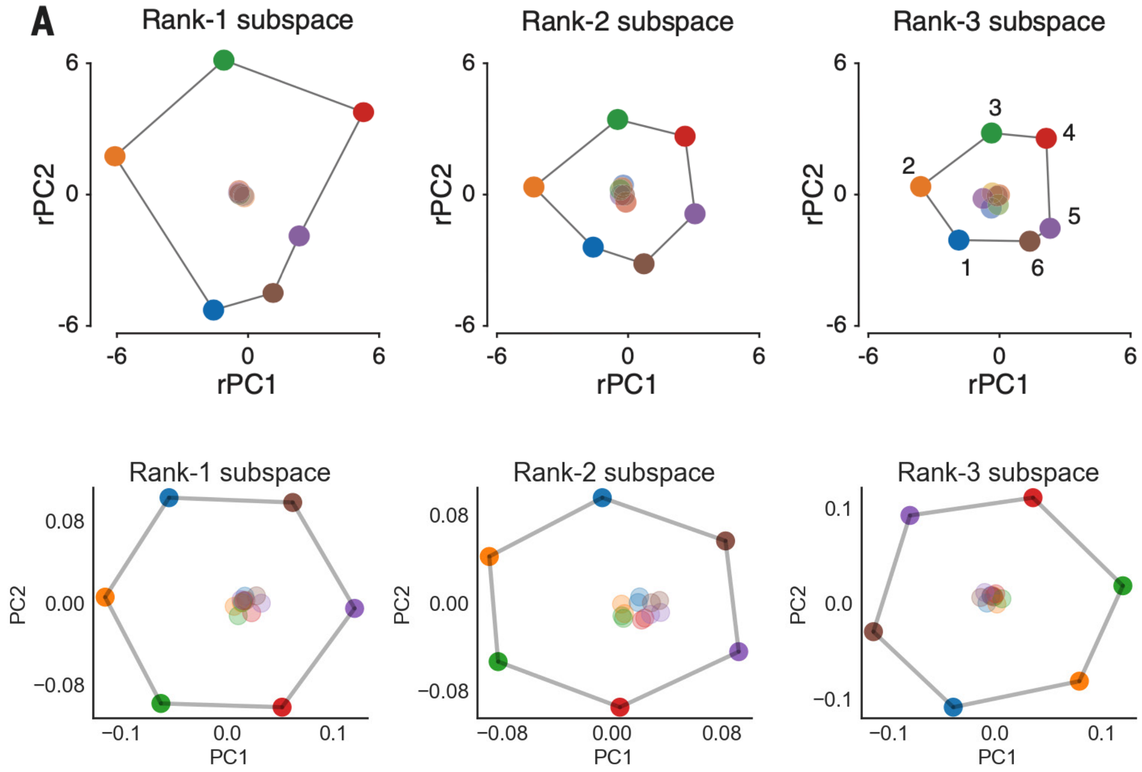


Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex

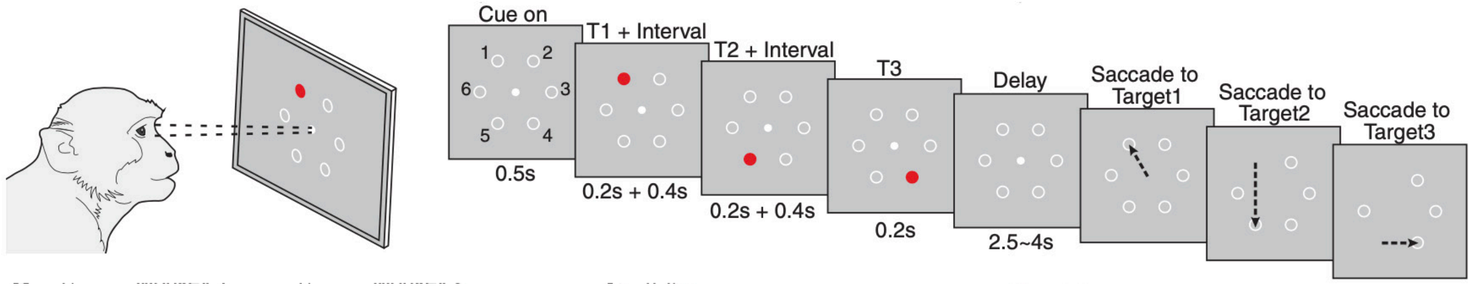


Xie et al., 2022

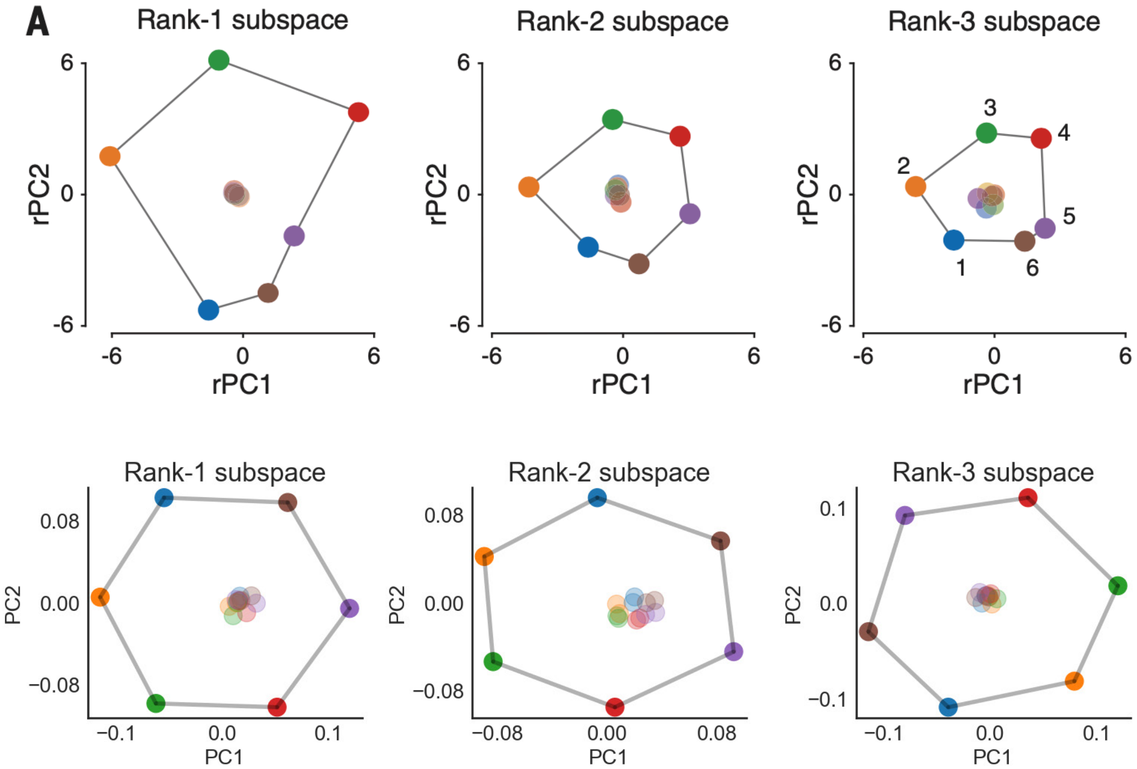


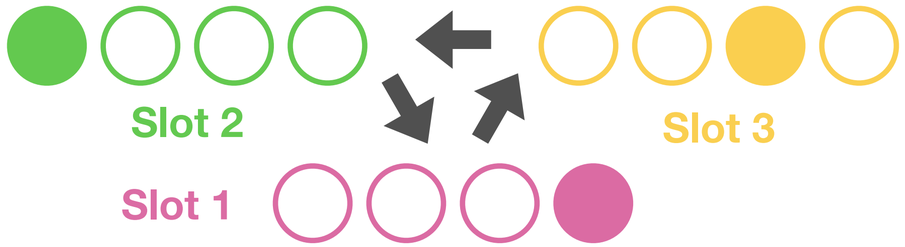
Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex



Xie et al., 2022





Whittington et al., 2023, *bioRxiv*

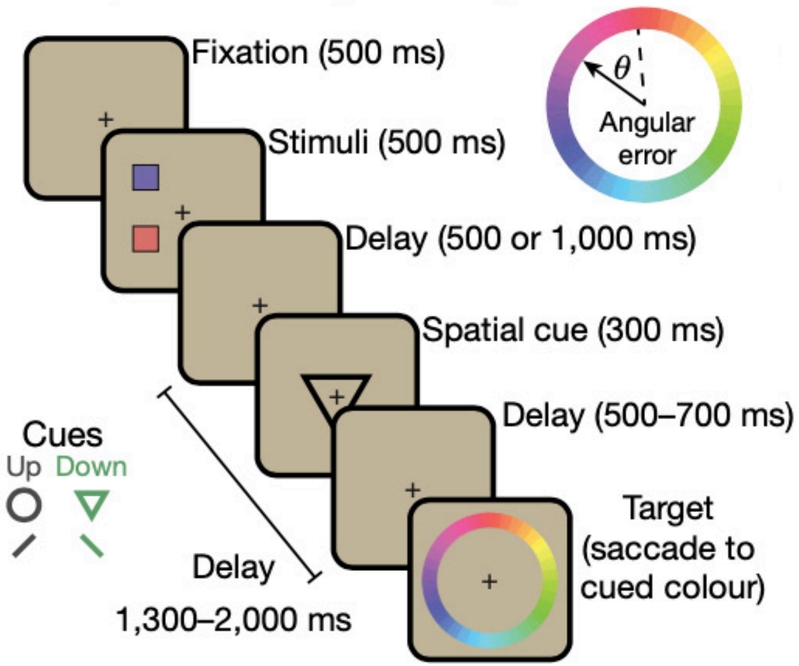
A slot based understanding unifies many representations in prefrontal cortex

Panichello & Buschman 2021

Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex

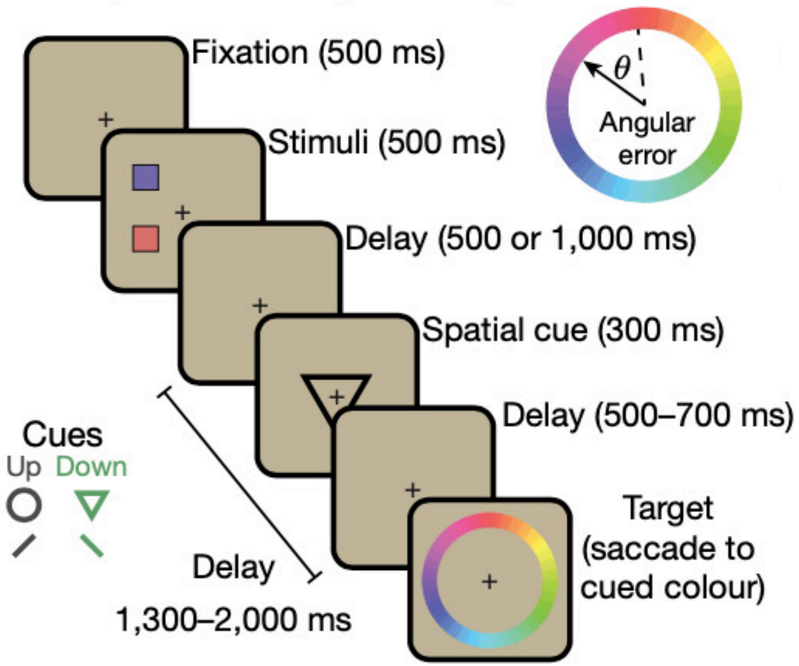
Panichello & Buschman 2021

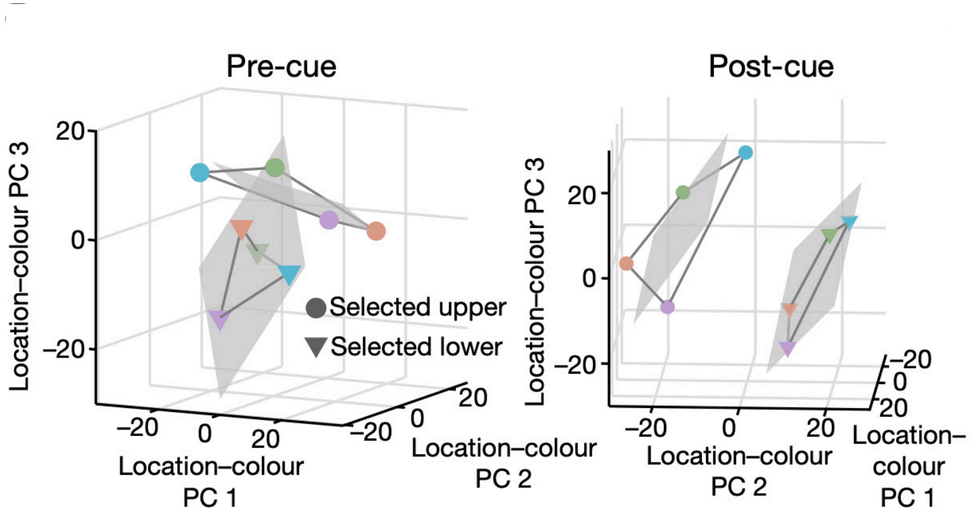


Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex

Panichello & Buschman 2021

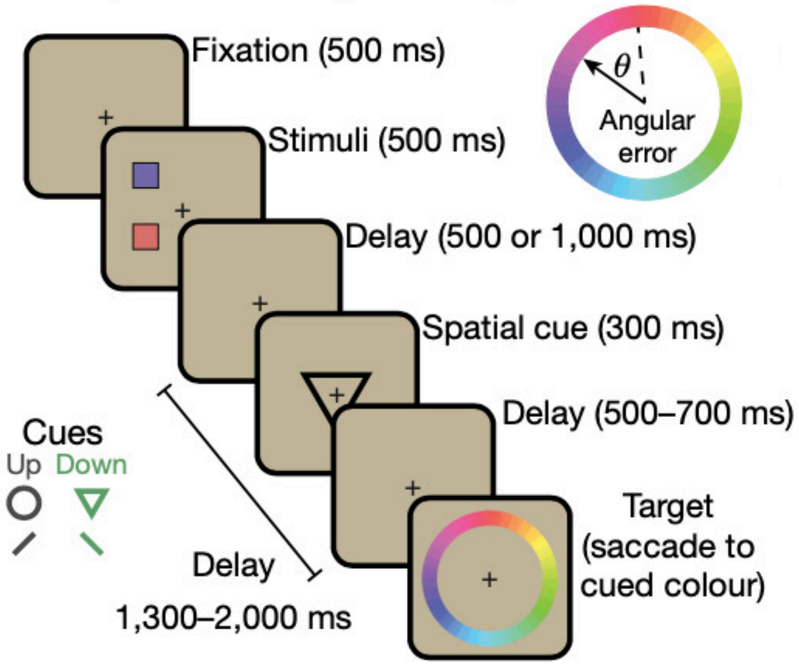


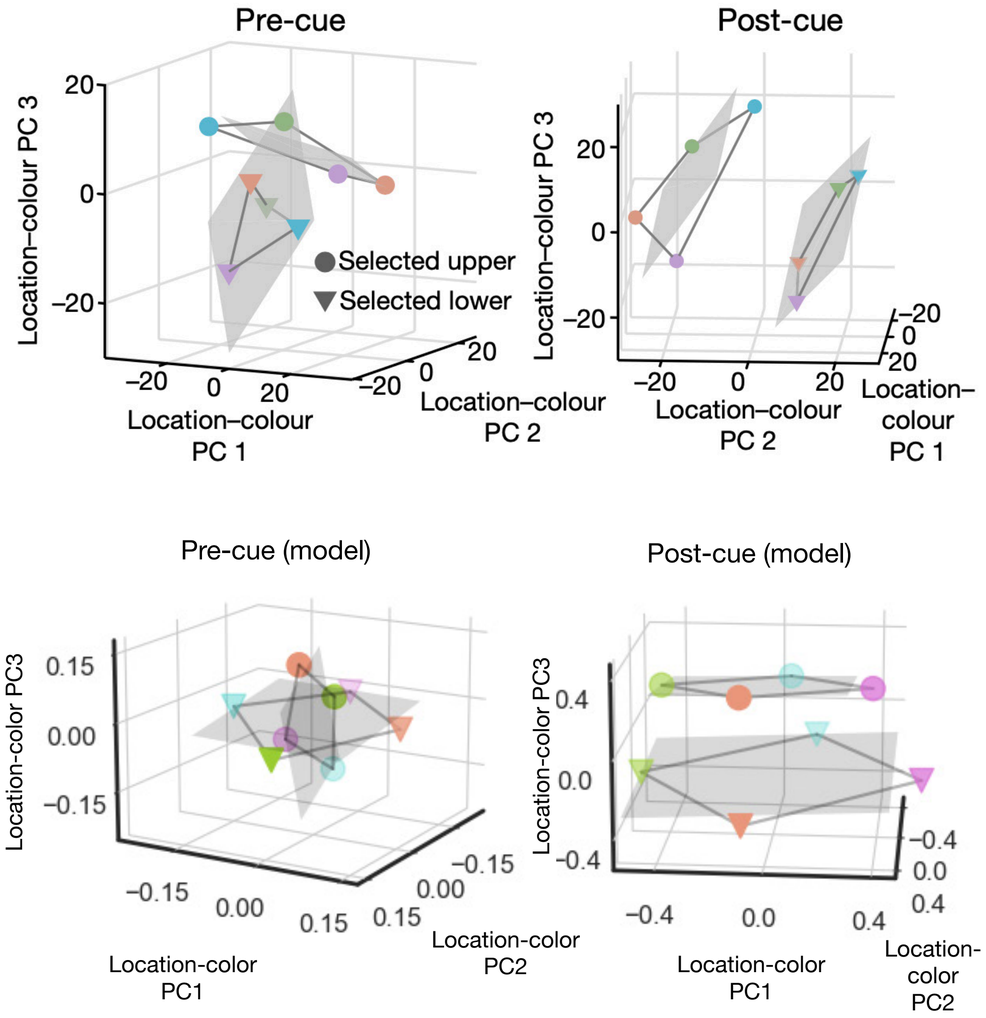


Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex

Panichello & Buschman 2021

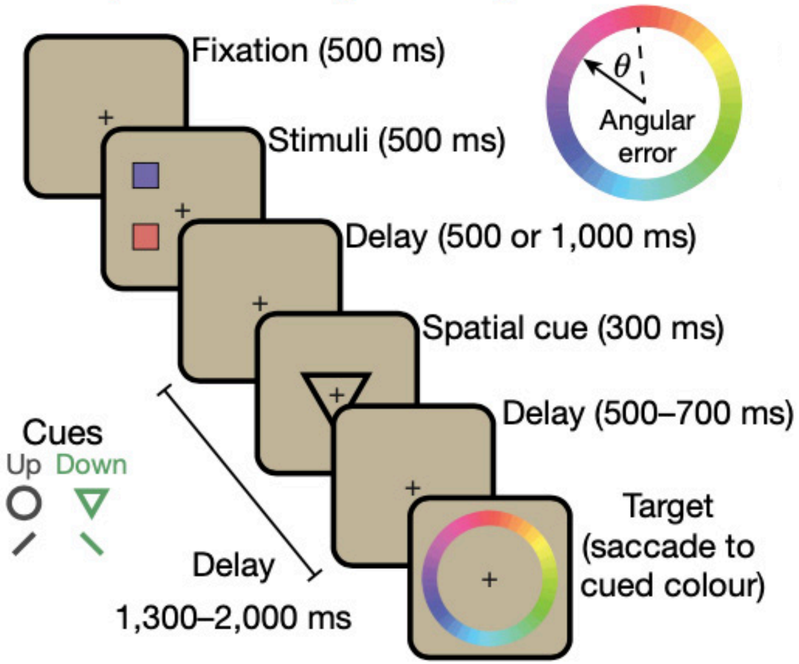


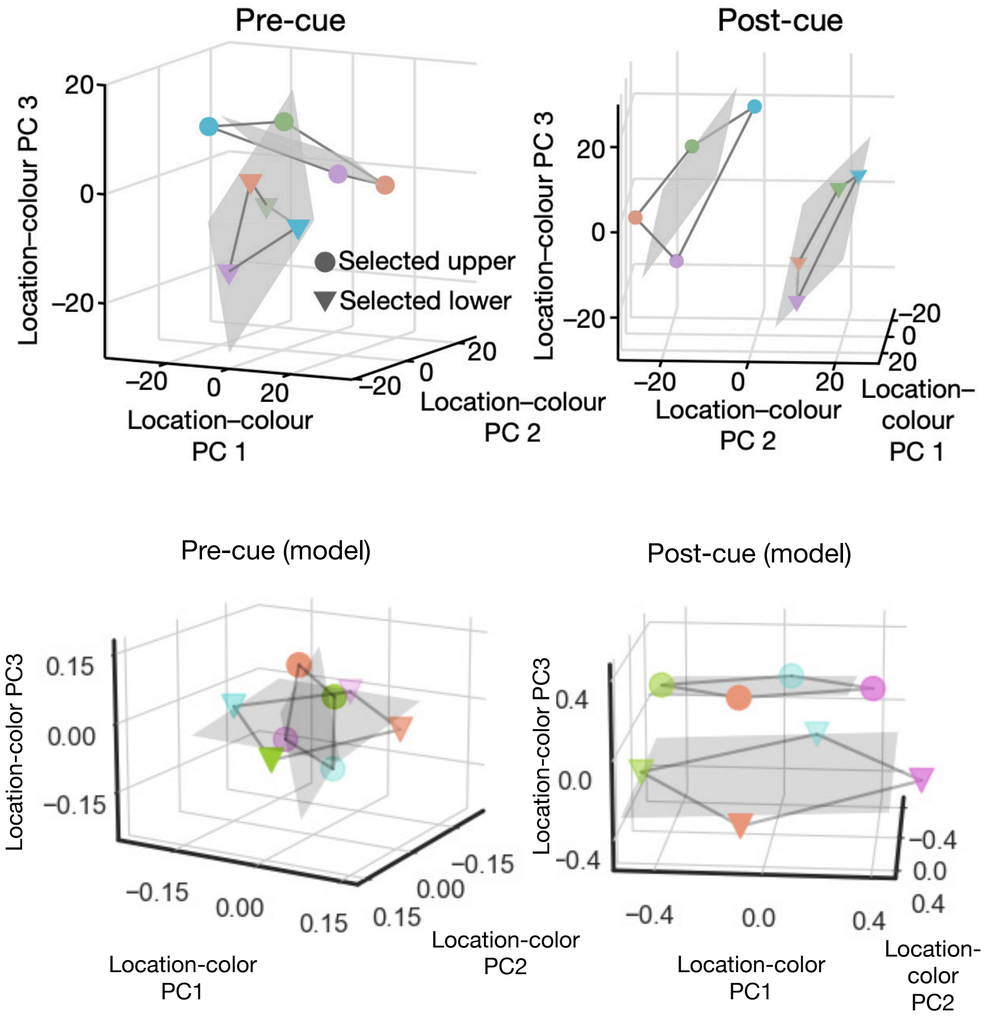


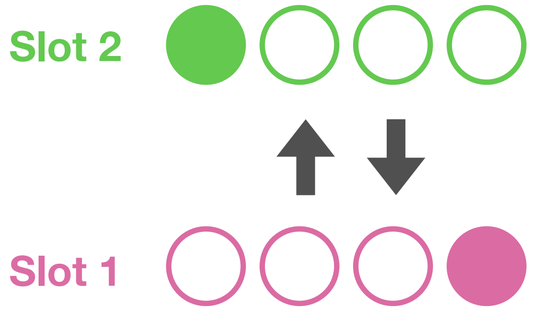
Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex

Panichello & Buschman 2021



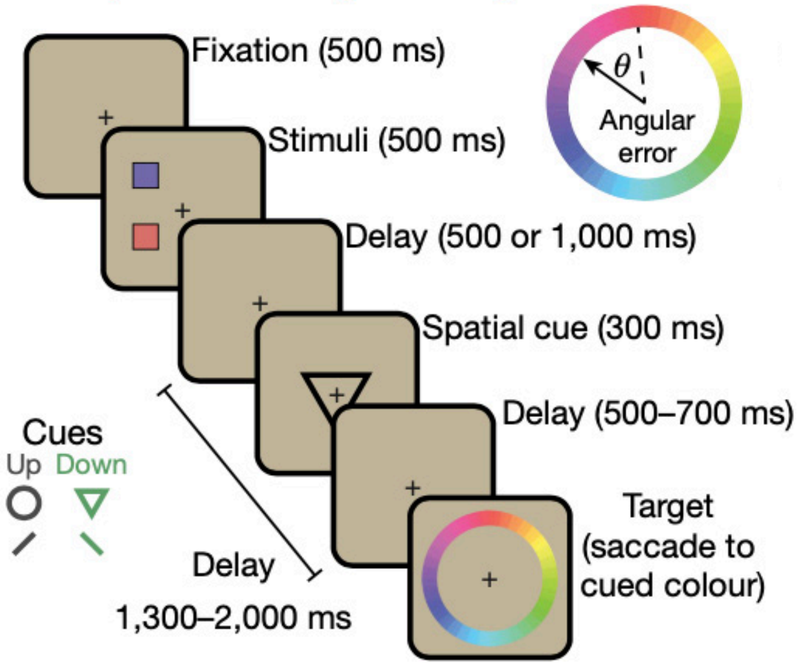


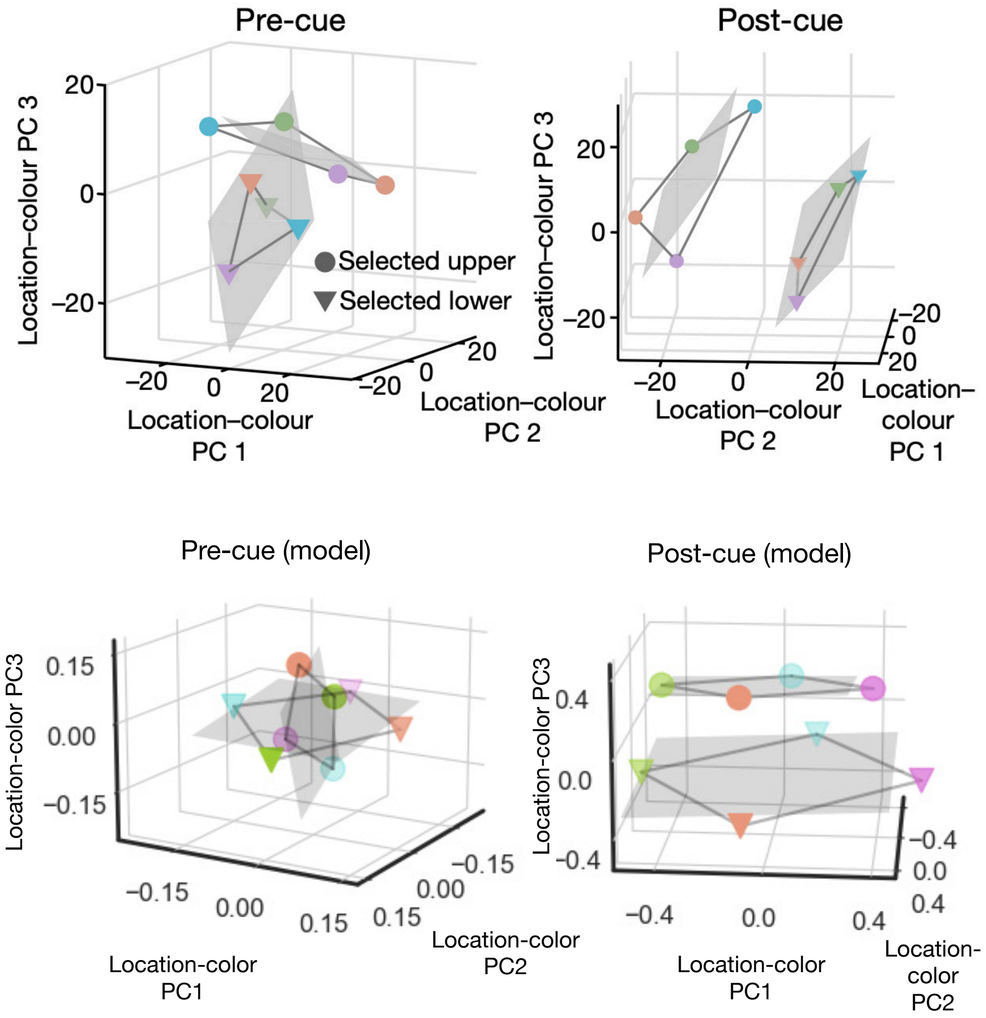


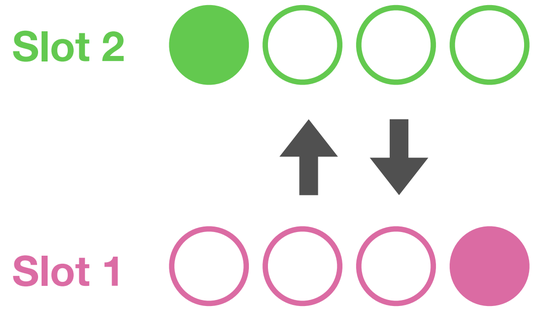
Whittington et al., 2023, *bioRxiv*

A slot based understanding unifies many representations in prefrontal cortex

Panichello & Buschman 2021







The cue behaves like an action signal!

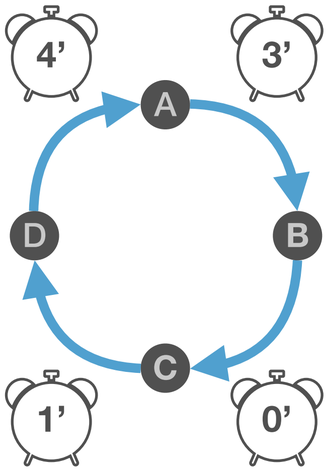
Whittington et al., 2023, *bioRxiv*

Makes novel predictions we tested in prefrontal data

Whittington et al., 2023, *bioRxiv*

El-Gaby et al., 2023, *bioRxiv*

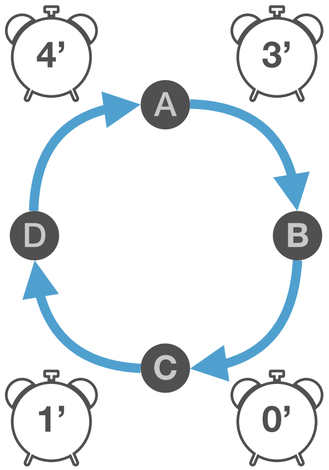
Makes novel predictions we tested in prefrontal data



Whittington et al., 2023, *bioRxiv*

El-Gaby et al., 2023, *bioRxiv*

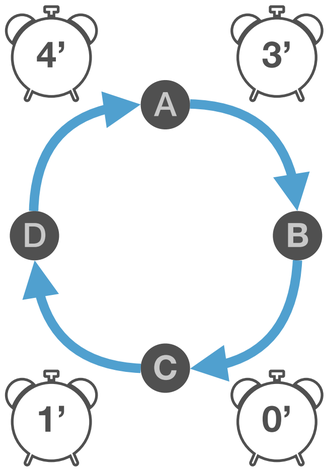
Makes novel predictions we tested in prefrontal data



Whittington et al., 2023, *bioRxiv*

El-Gaby et al., 2023, *bioRxiv*

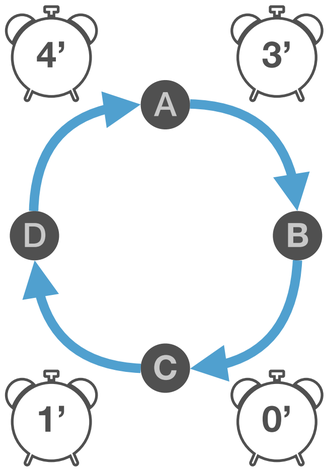
Makes novel predictions we tested in prefrontal data

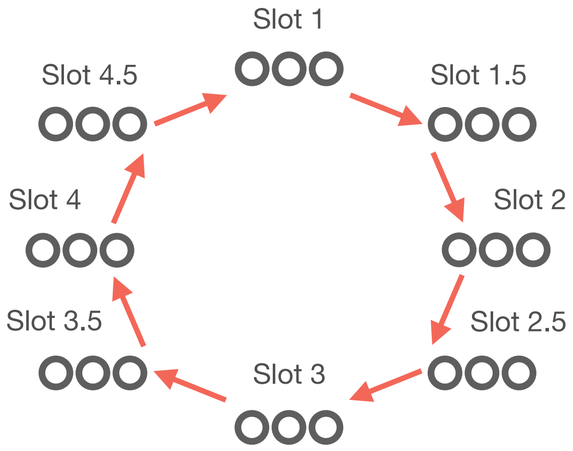


Whittington et al., 2023, *bioRxiv*

El-Gaby et al., 2023, *bioRxiv*

Makes novel predictions we tested in prefrontal data

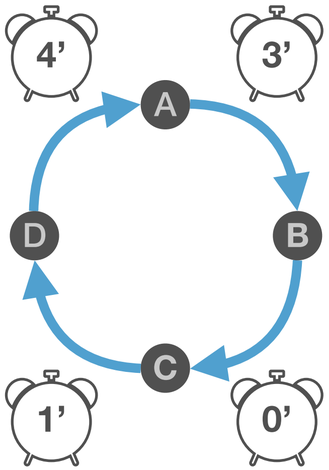




Whittington et al., 2023, *bioRxiv*

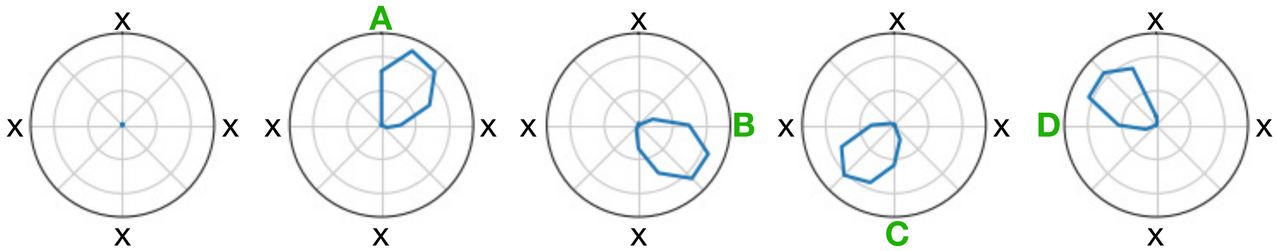
El-Gaby et al., 2023, *bioRxiv*

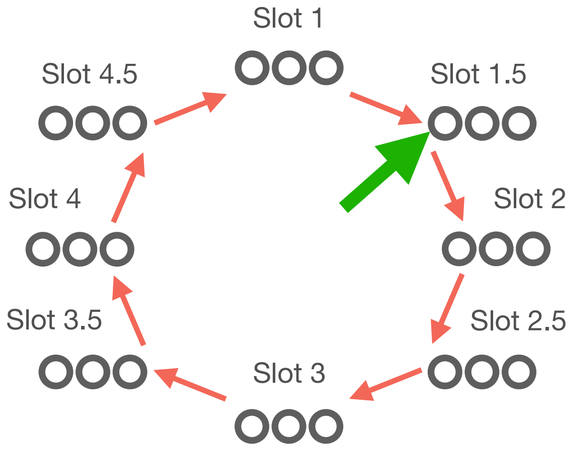
Makes novel predictions we tested in prefrontal data



Task progress slot cells

Cell #3

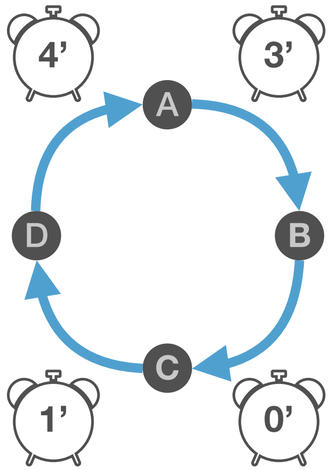




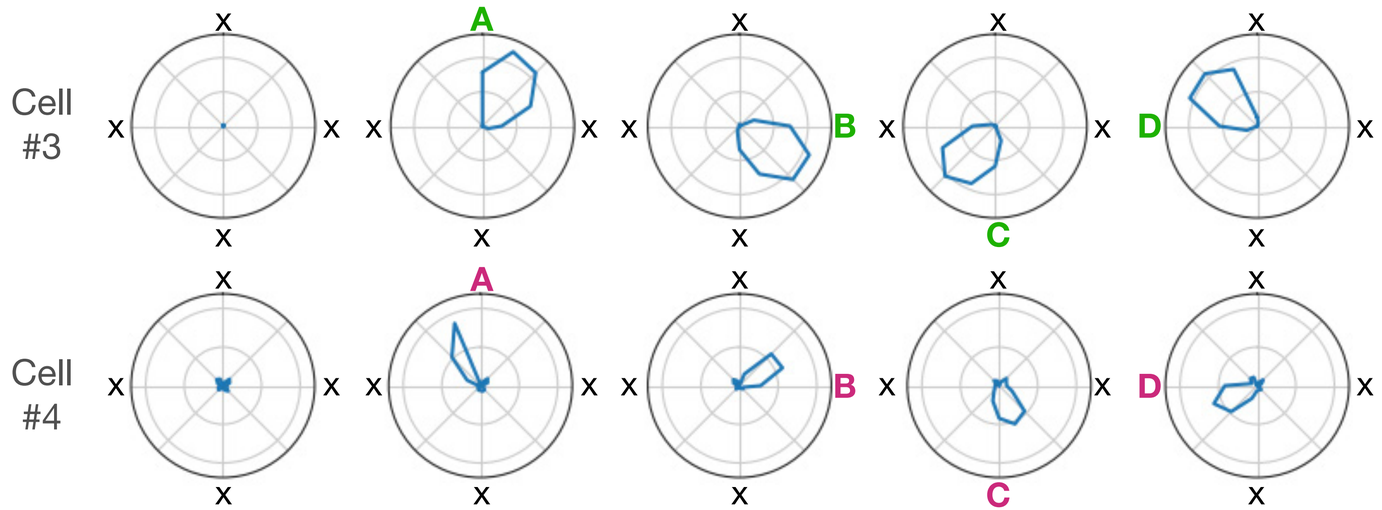
Whittington et al., 2023, *bioRxiv*

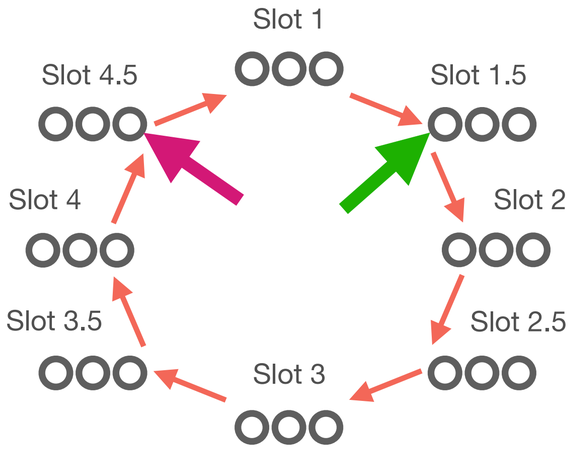
El-Gaby et al., 2023, *bioRxiv*

Makes novel predictions we tested in prefrontal data



Task progress slot cells

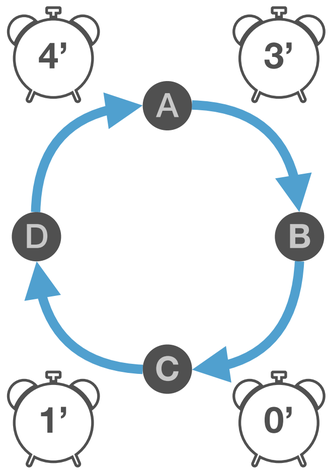




Whittington et al., 2023, *bioRxiv*

El-Gaby et al., 2023, *bioRxiv*

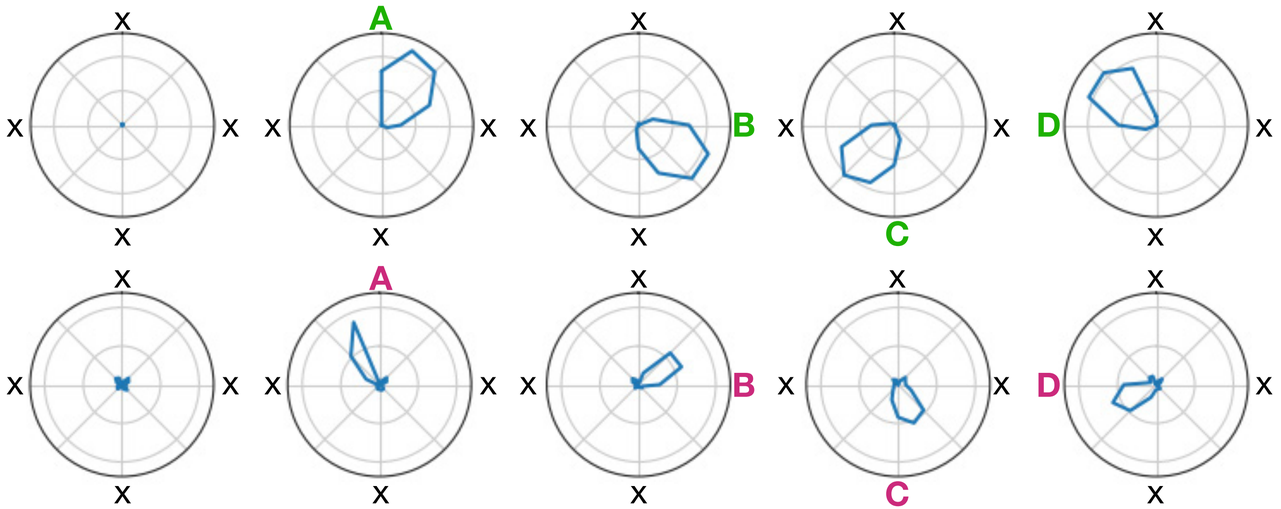
Makes novel predictions we tested in prefrontal data

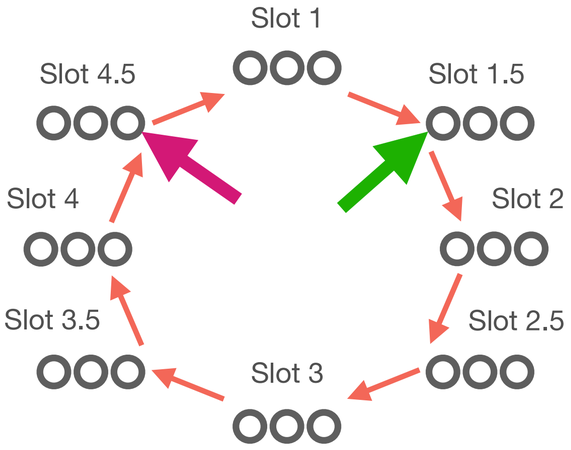


Task progress slot cells

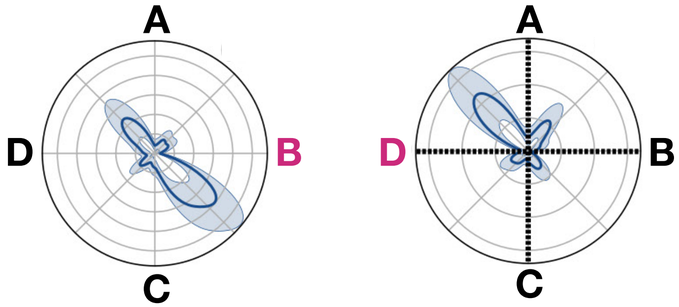
Cell #3

Cell #4





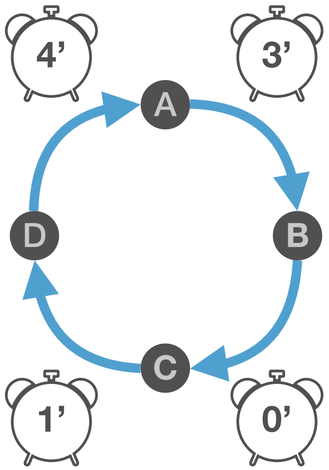
PFC real ‘purple 50% lag’ neuron



Whittington et al., 2023, *bioRxiv*

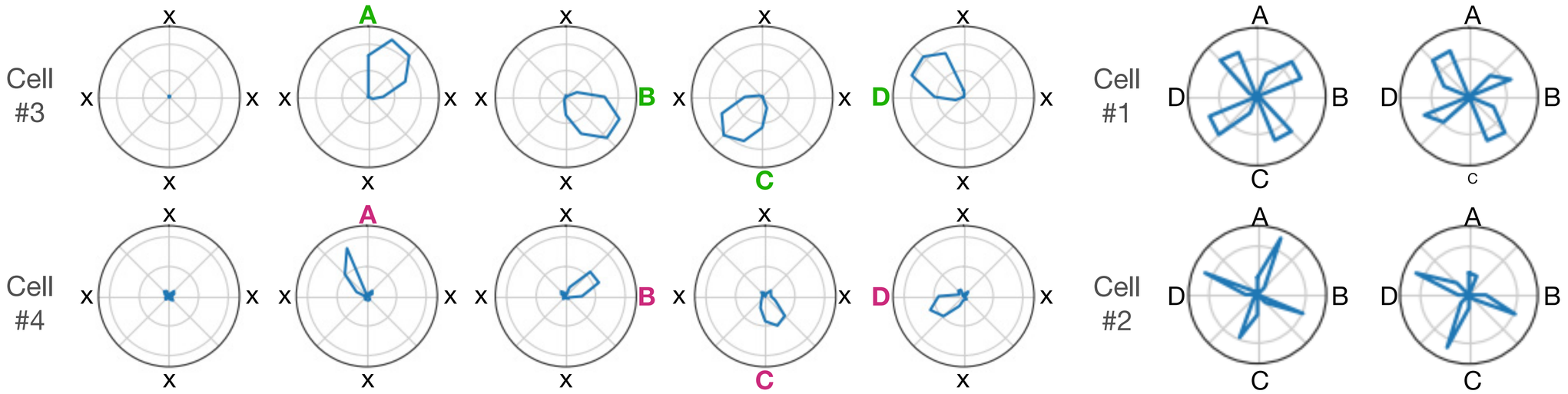
El-Gaby et al., 2023, *bioRxiv*

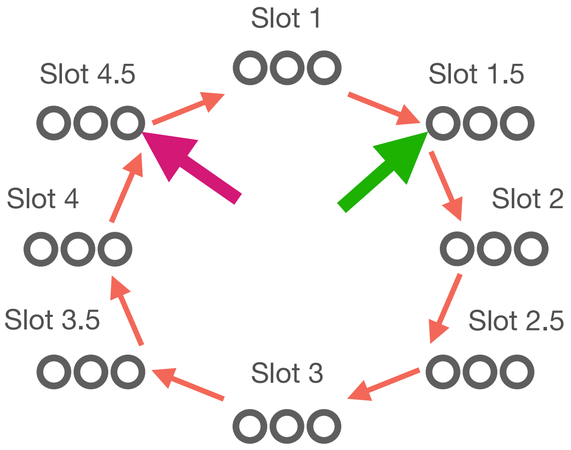
Makes novel predictions we tested in prefrontal data



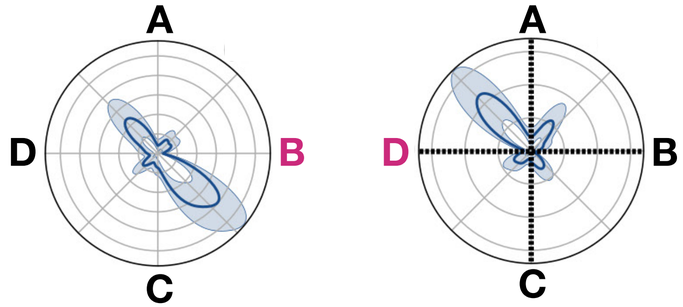
Task progress slot cells

Progress cells





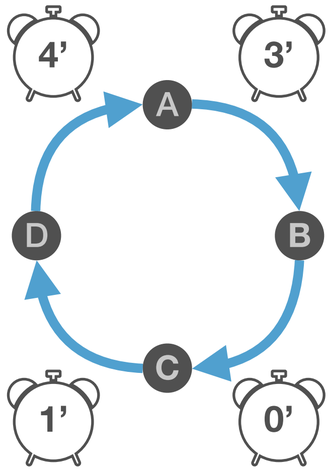
PFC real 'purple 50% lag' neuron



Whittington et al., 2023, *bioRxiv*

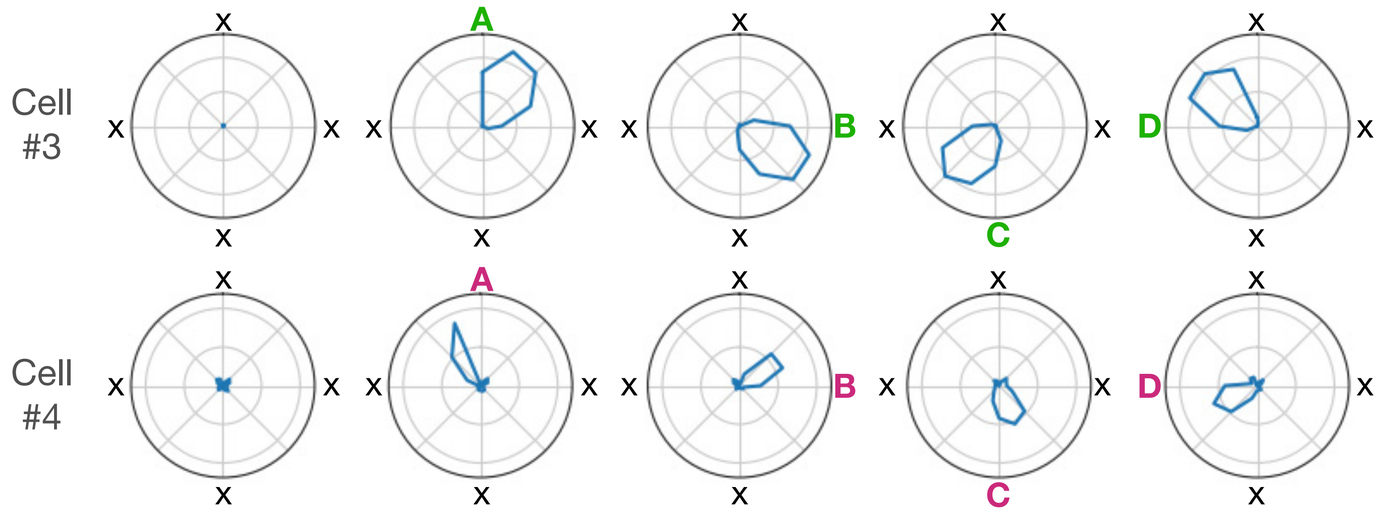
El-Gaby et al., 2023, *bioRxiv*

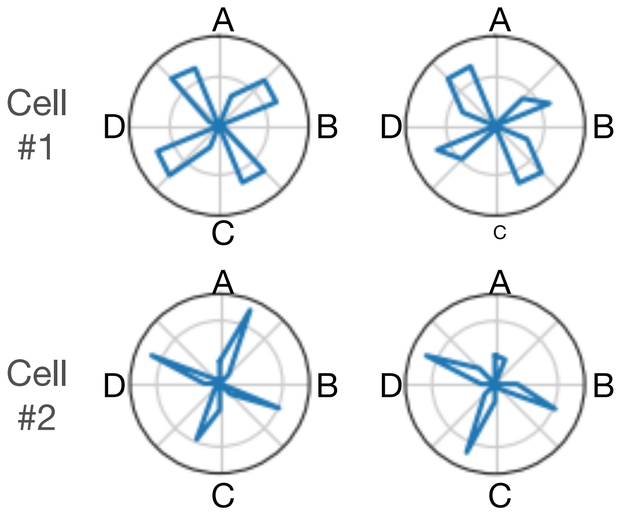
Makes novel predictions we tested in prefrontal data

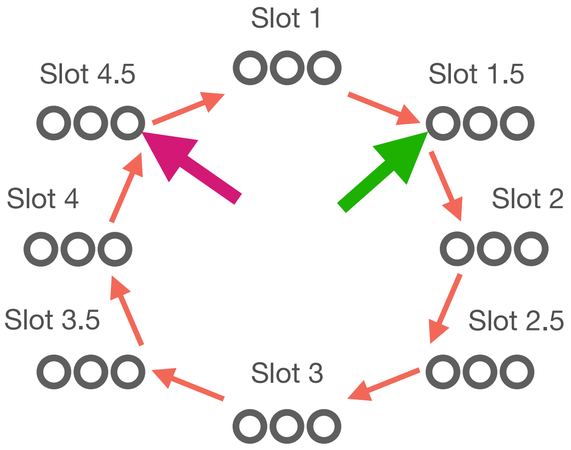


Task progress slot cells

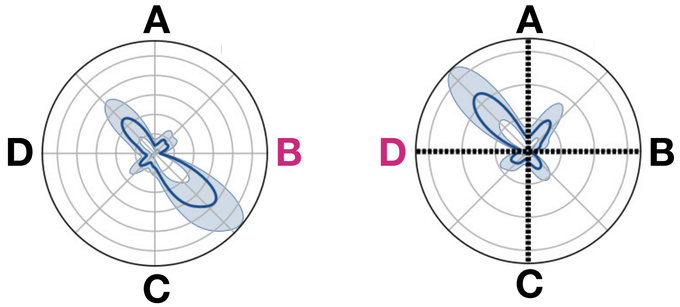
Progress cells

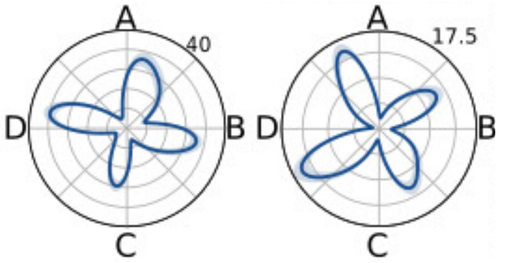






PFC real 'purple 50% lag' neuron

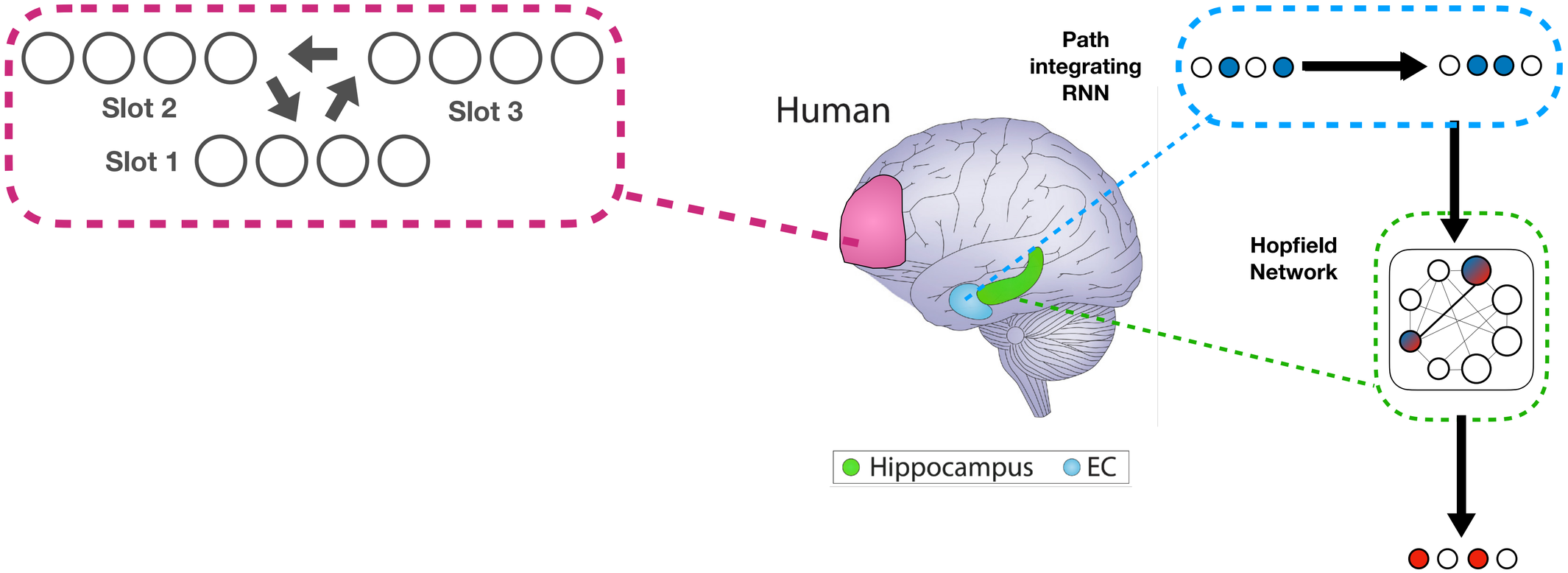




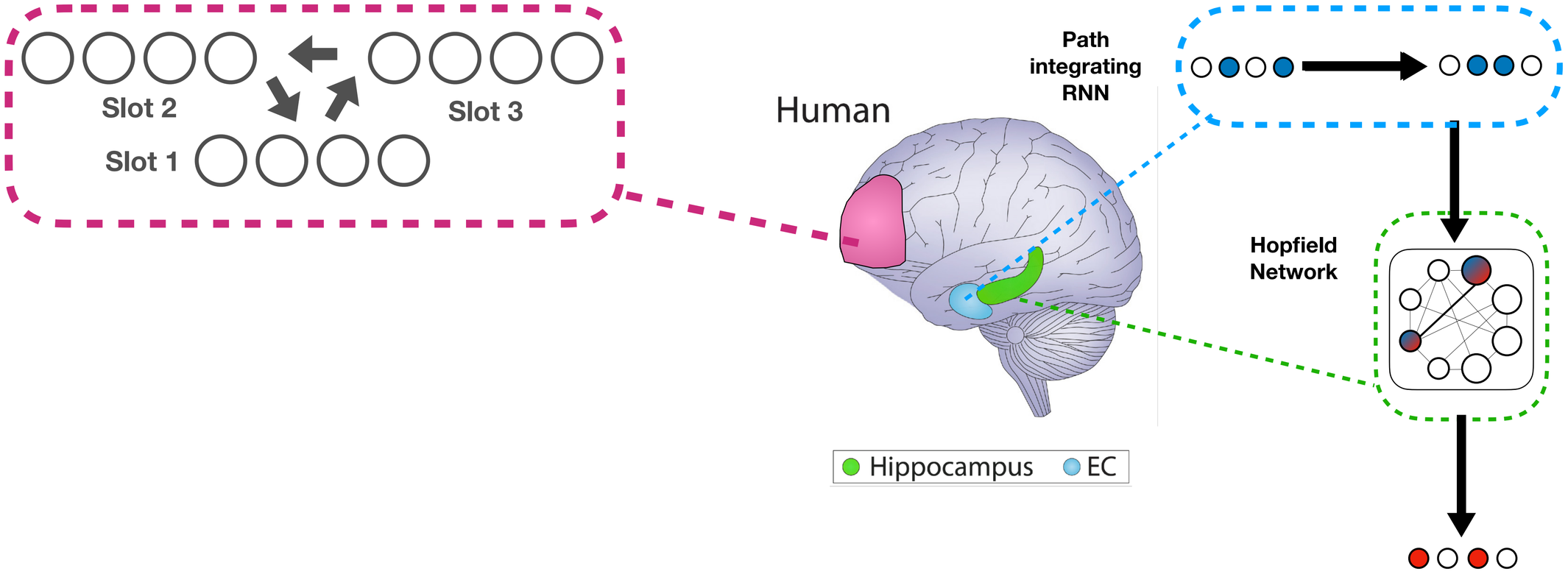
Whittington et al., 2023, *bioRxiv*

El-Gaby et al., 2023, *bioRxiv*

Lastly, the two algorithms are mathematically equivalent



Lastly, the two algorithms are mathematically equivalent



One stores memories in weights, and the other stores memories in neural activities

Puzzles of cognitive maps in the brain

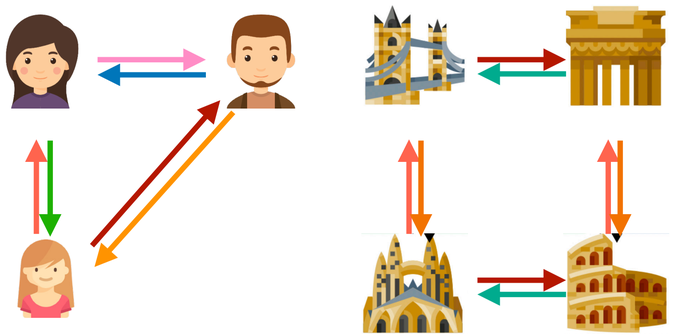
How does the same system do space and non-space?

How can brains learn these maps?

Why do the neuron look the ways they do?

How do different brain regions solve the same problem in different ways?

Puzzles of cognitive maps in the brain



How does the same system do space and non-space?

How can brains learn these maps?

Why do the neuron look the ways they do?

How do different brain regions solve the same problem in different ways?

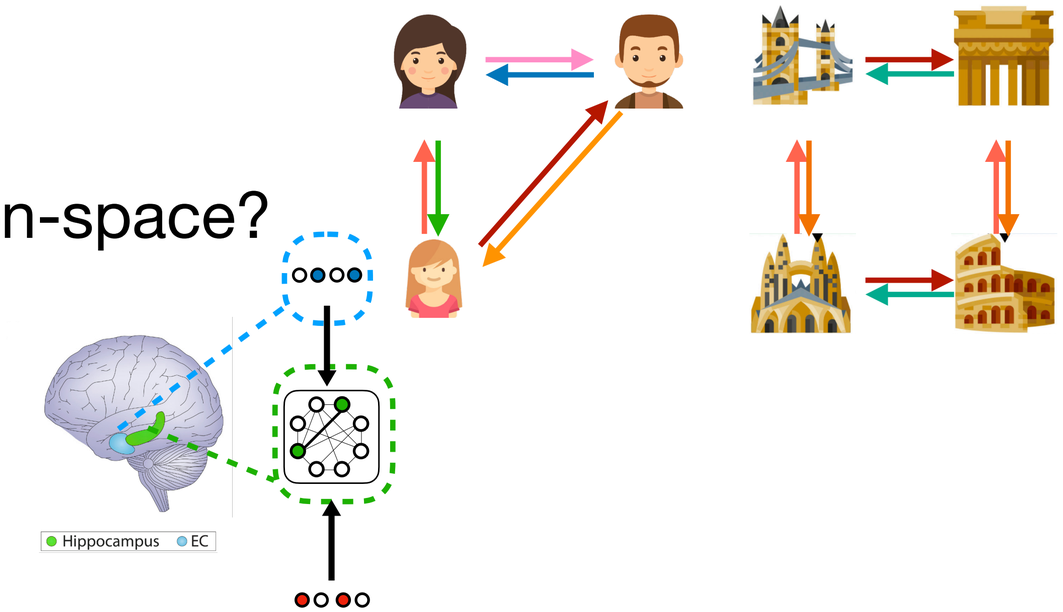
Puzzles of cognitive maps in the brain

How does the same system do space and non-space?

How can brains learn these maps?

Why do the neuron look the ways they do?

How do different brain regions solve the same problem in different ways?



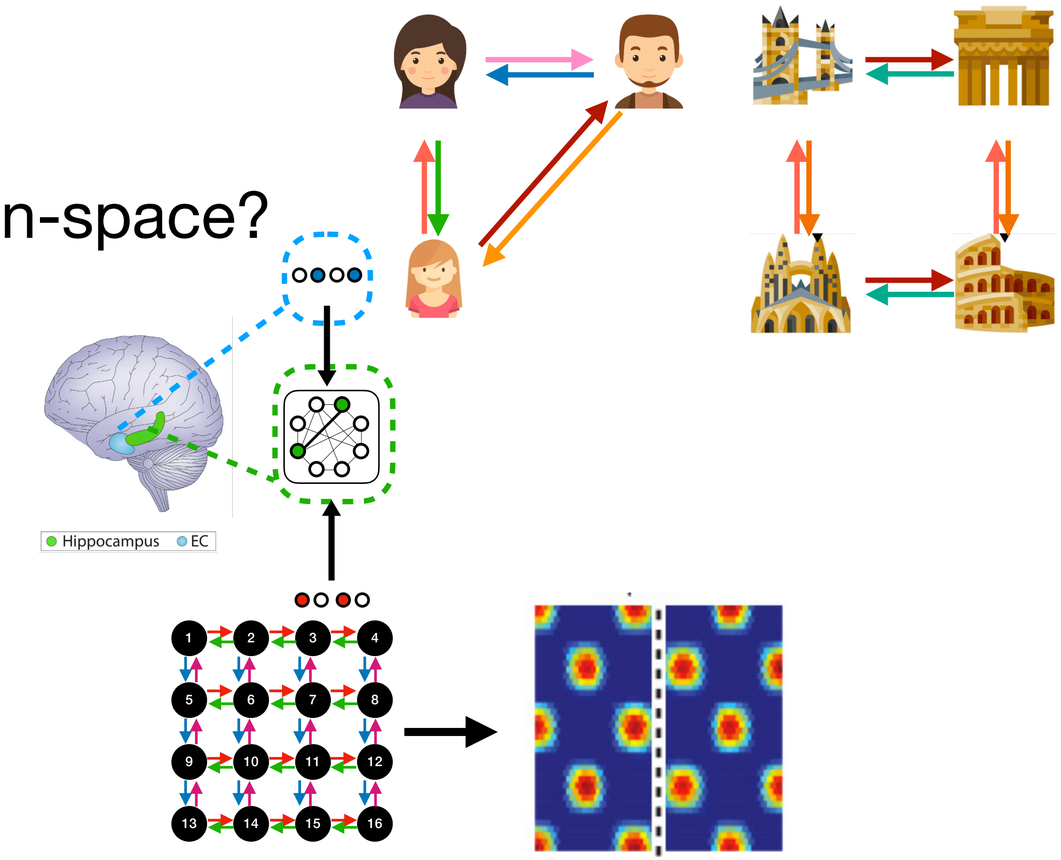
Puzzles of cognitive maps in the brain

How does the same system do space and non-space?

How can brains learn these maps?

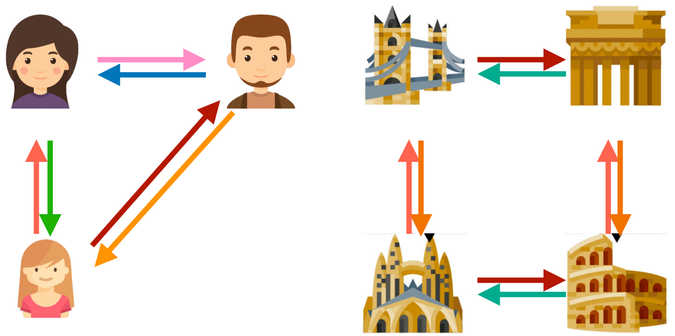
Why do the neuron look the ways they do?

How do different brain regions solve the same problem in different ways?

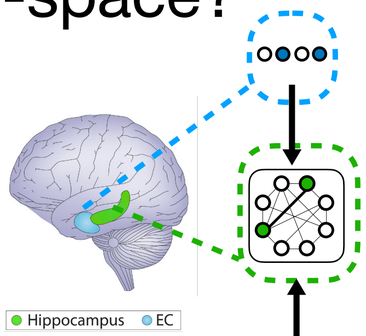


Puzzles of cognitive maps in the brain

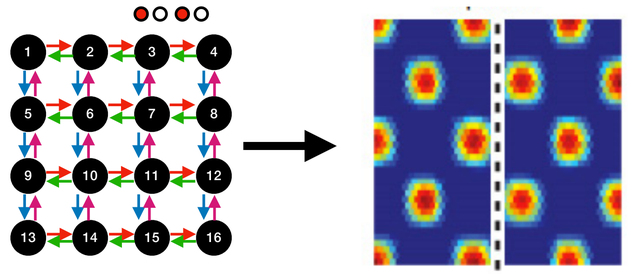
How does the same system do space and non-space?



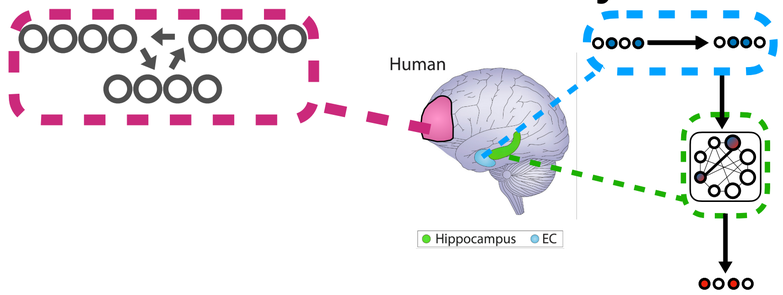
How can brains learn these maps?



Why do the neuron look the ways they do?

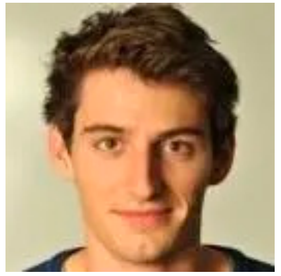


How do different brain regions solve the same problem in different ways?



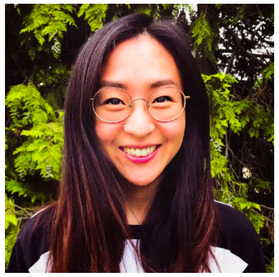
Thanks!

Jacob Bakermans



Tim Muller

Chongyu Qin



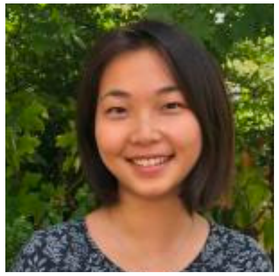
Will Dorrell

Shirley Mark

Guifen Chen

Jo Warren

Jiali Zhang



Mohamady El-Gaby

Tim Behrens

Surya Ganguli

wellcome

UNIVERSITY OF OXFORD

Stanford University

ZYPHRA