Sainsbury Wellcome Centre

UCL

Reservoir computing and its application to unsupervised temporal structure learning

aka. random nets process structured data

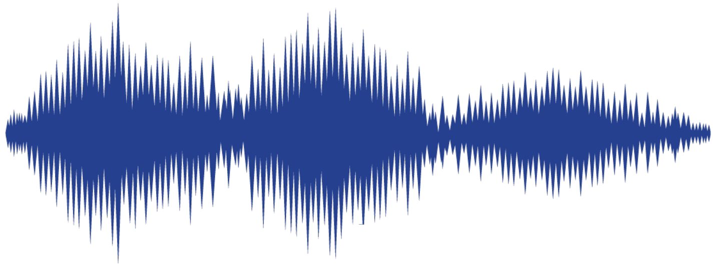
Tom George

Athena Akrami & Claudia Clopath

Sainsbury Wellcome Centre

Imperial College London

Reservoir computing and its application to unsupervised temporal structure learning



- Temporal Structure
  - It's all around us
  - We're great at learning it e.g. *Dehaene et al. (2015)*
- Unsupervised
  - Do we learn structure when it isn't task relevant?
  - Akrami lab experimental results suggest maybe. See also e.g. *Saffran et al. (1996)*
- Reservoir networks
  - Compared to RNNs, cheaper to train and fewer a priori constraints, e.g. *Jaeger et al. (2001)*
  - Architectural parallels to cortex e.g. *Szary et al. (2011)*

5 key taxonomies of temporal structure

"how does the brain encode temporal sequences of items, such that this knowledge can be used to ***retrieve a sequence from memory****, recognize it,* ***anticipate on forthcoming items****, and* ***generalize*** *this knowledge to* ***novel sequences*** *with a similar structure?" – Lashley (1951)*

1. Transition and timing knowledge

♪♪♪♪♪♪♪♪♪♪♪♪ ?

2. Chunking

gopilagikobatokibutokibugikobagopila

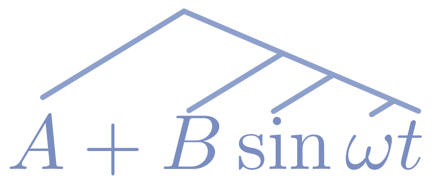
3. Ordinal knowledge

1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3

4. Algebraic patterns

mimitu	totobu	gagari	pesipe	pipigo
AAB	AAB	AAB	ABA	AAB

5. Nested tree structures generate by symbolic rules



Dehaene et al. (2015)

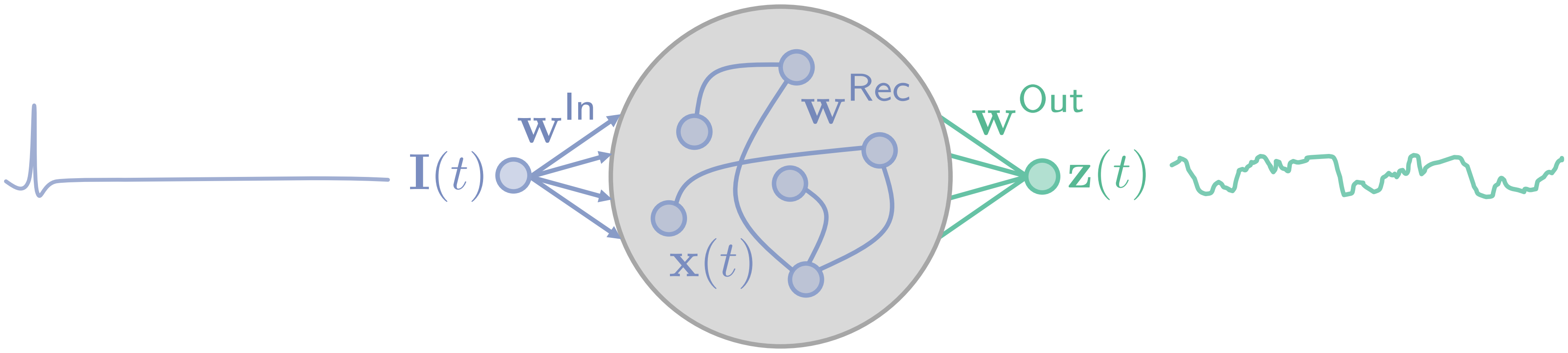
Roadmap

Roadmap

1. A reservoir network model for temporal structure learning
2. The role of chaos
3. Experimental results and modelling predictions
4. Conclusions

Reservoir networks are just random RNNs

We train the output weights only



- Random fixed recurrent weights → dynamics

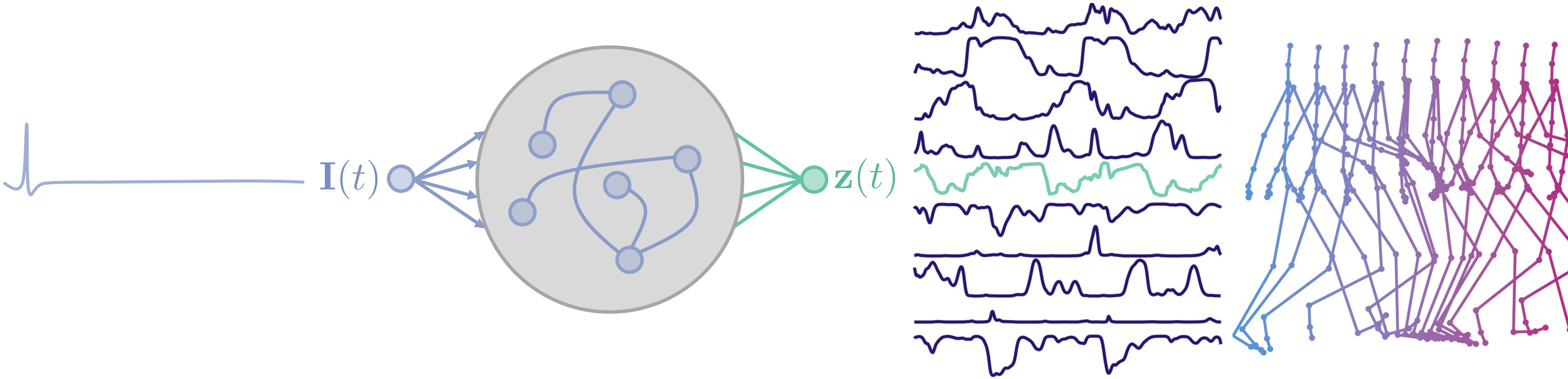
$$\tau\dot{\mathbf{x}} = -\mathbf{x} + \mathbf{W}^{\text{Rec}} \cdot \phi(\mathbf{x}) + \mathbf{W}^{\text{In}} \cdot \mathbf{I} + \cdots$$ e.g. noise + feedback

$$\mathbf{W}^{\text{Rec}}_{ij} \sim \mathcal{N}(0, \frac{g}{\sqrt{N}})$$

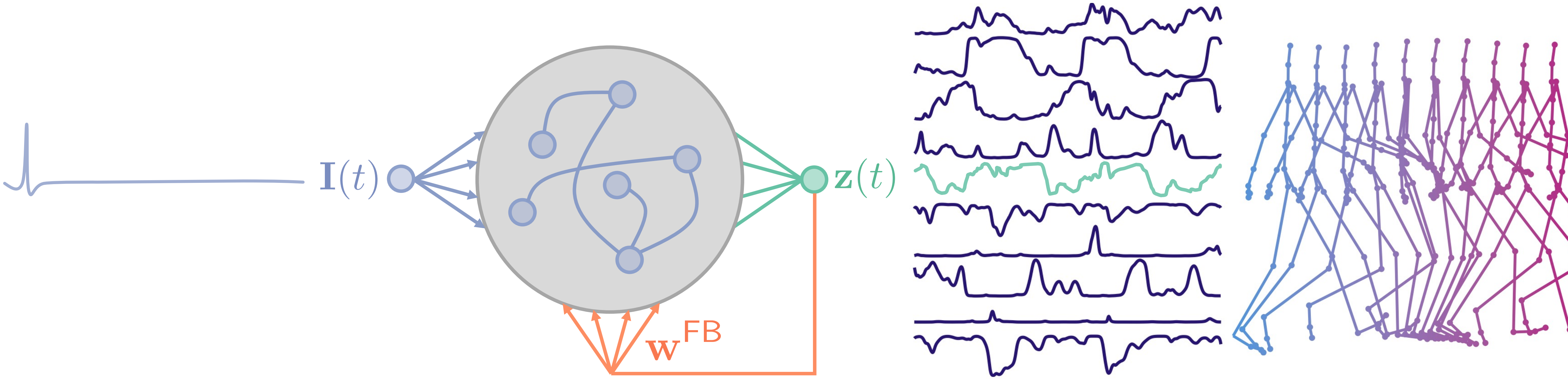
- Trainable linear weights → readout

$$\mathbf{z} = \mathbf{W}^{\text{Out}} \cdot \phi(\mathbf{x})$$

Reservoir networks are just random RNNs which can do non-random things

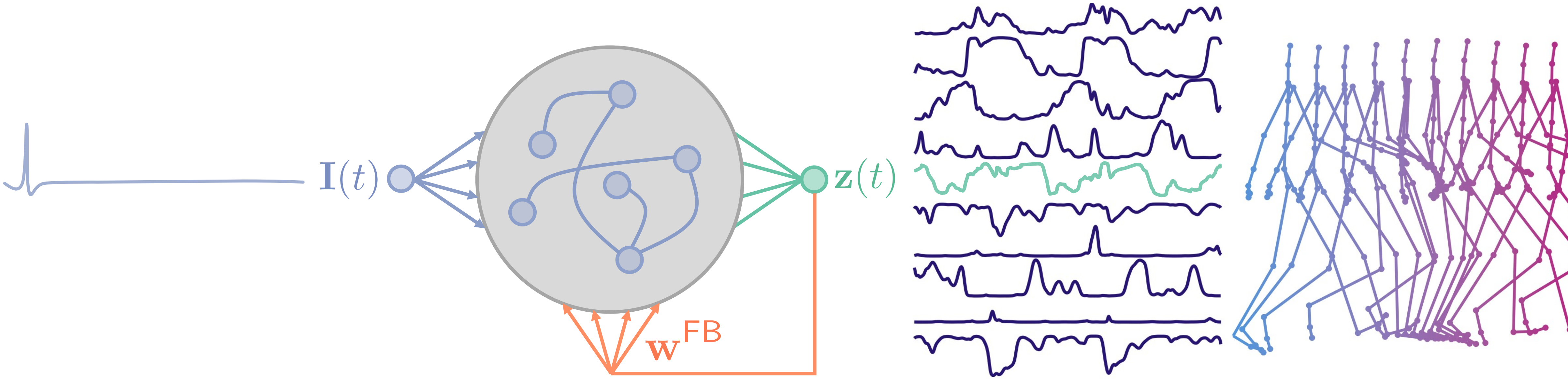


Reservoir networks are just random RNNs which can do non-random things



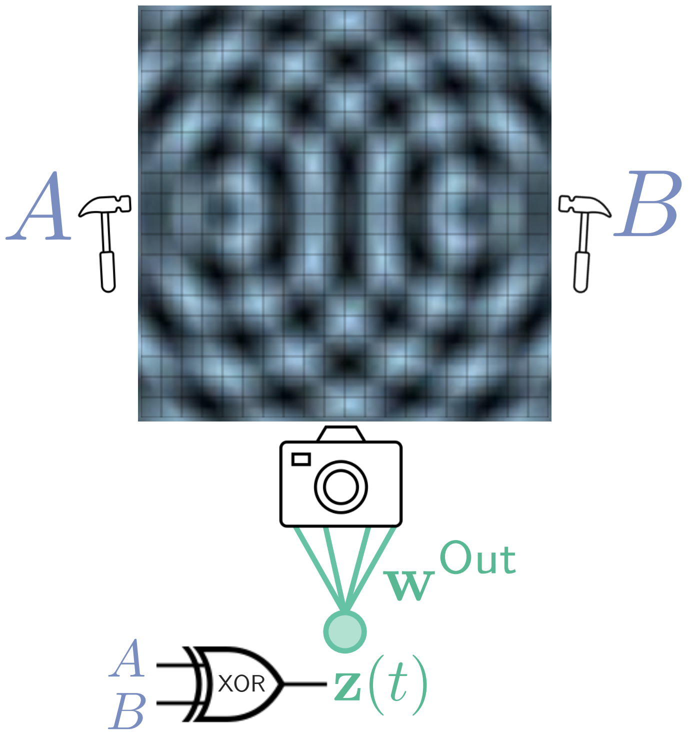
- **Pattern generation:** FORCE allowed feedback error during training of $\mathbf{w}^{\text{Out}}$ via RLS for pattern generation in, e.g., motor cortex. (Had a huge impact on the field.) *Sussillo and Abbott (2009).*

Reservoir networks are just random RNNs which can do non-random things



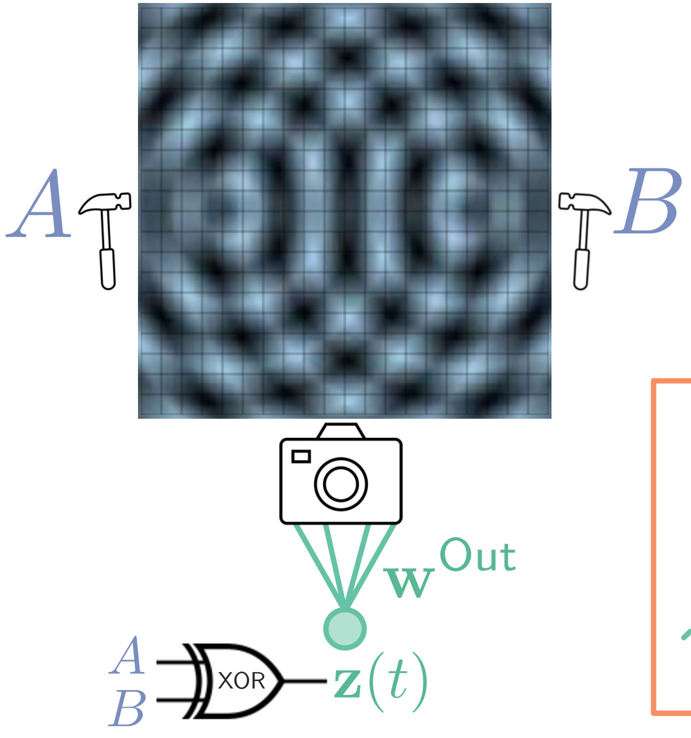
- Pattern generation: FORCE learning, *Sussillo and Abbott (2009)*.
- Robust timing: reservoir nets as the brain's 'stopwatch', *Laje and Buonomano (2013)*.
- Representations: History dependent mixed-selective representations in PFC, *Enel et al. (2016)*.
- Chunking/event segmentation: *Asabuki and Fukai (2018)*.

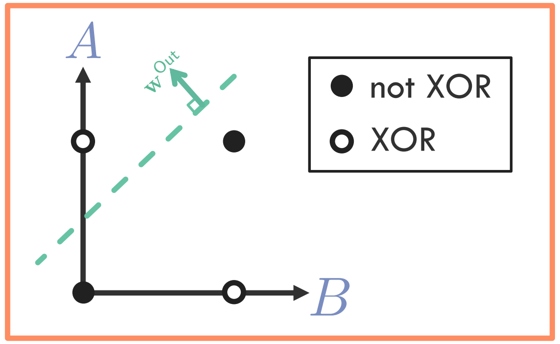
Reservoir computing with a bucket of water?



Fernando and Sojakka (2003)

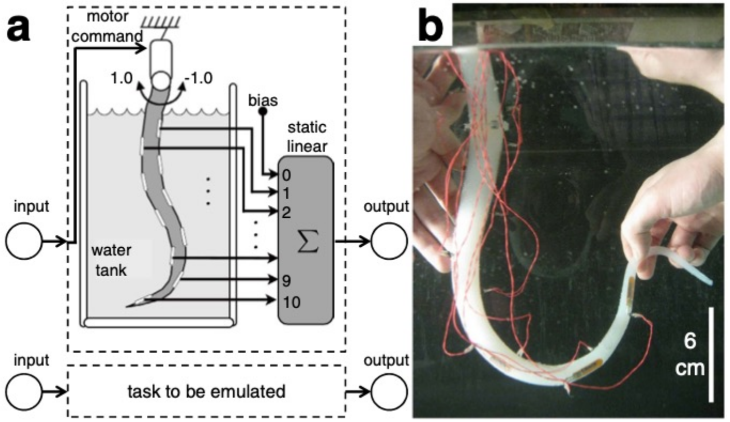
Reservoir computing with a bucket of water?



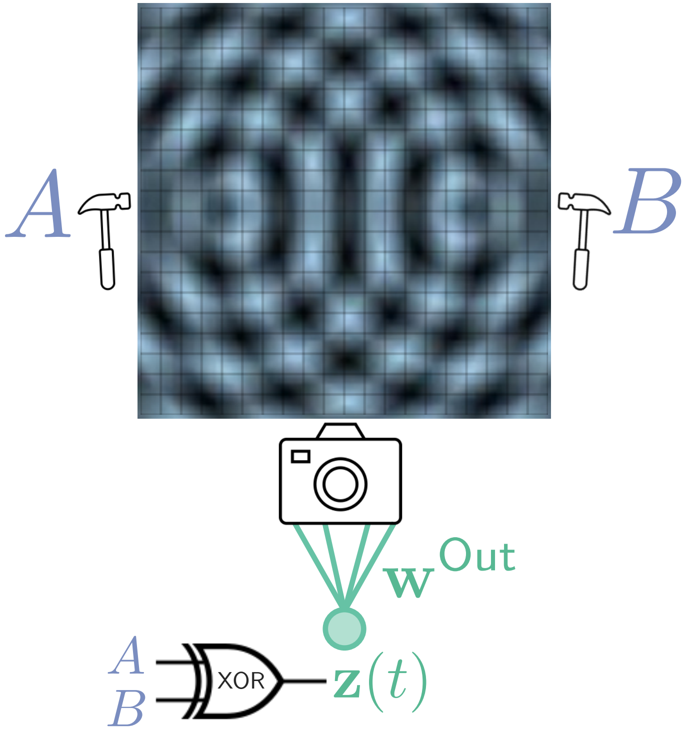


Fernando and Sojakka (2003)

Reservoir computing with a bucket of water?...an Octopus arm?!?!



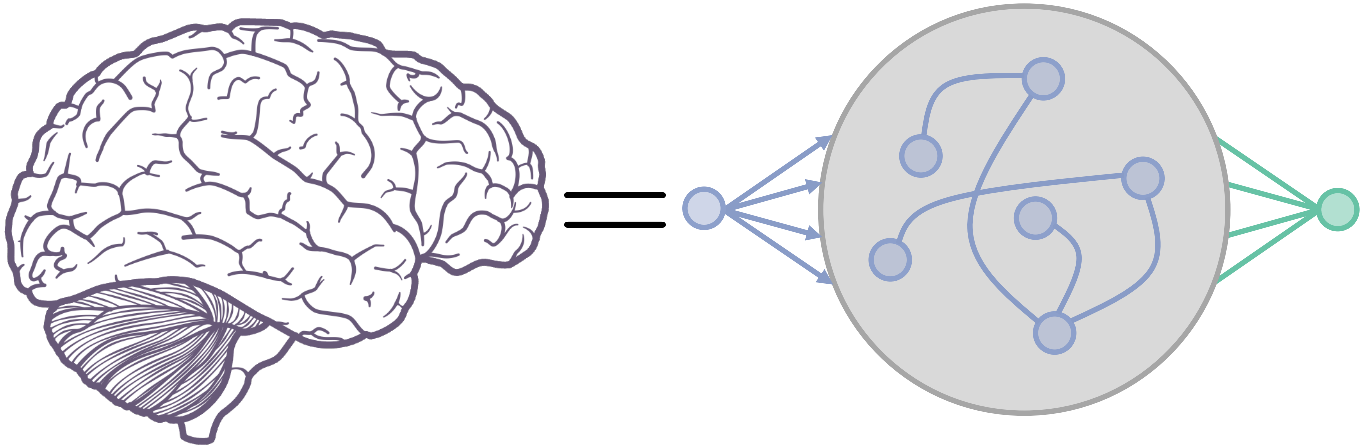
Nakajima et al., (2015)



Fernando and Sojakka (2003)

1. Nonlinearity
2. Dynamic
3. Many degree's of freedom

To first order, cortex is a sparse randomly connected RNN satisfying these requirements



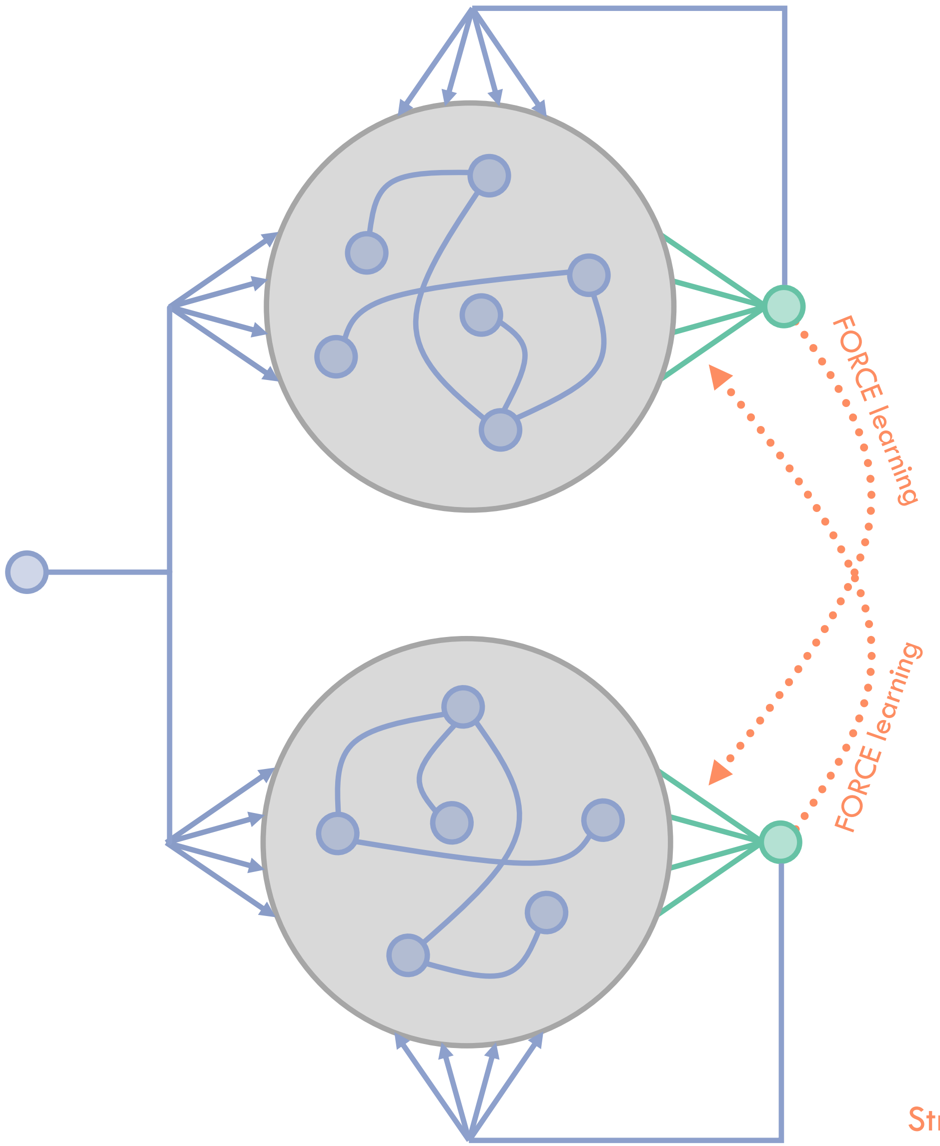
$+\mathcal{O}(\delta^2)$

1. Nonlinearity
2. Dynamic
3. Many degree's of freedom

Training rule: Two networks, each tries to predict the other

Weights are updated by FORCE

Temporally random inputs → Temporally unpredictable dynamics → Mutual prediction *impossible*



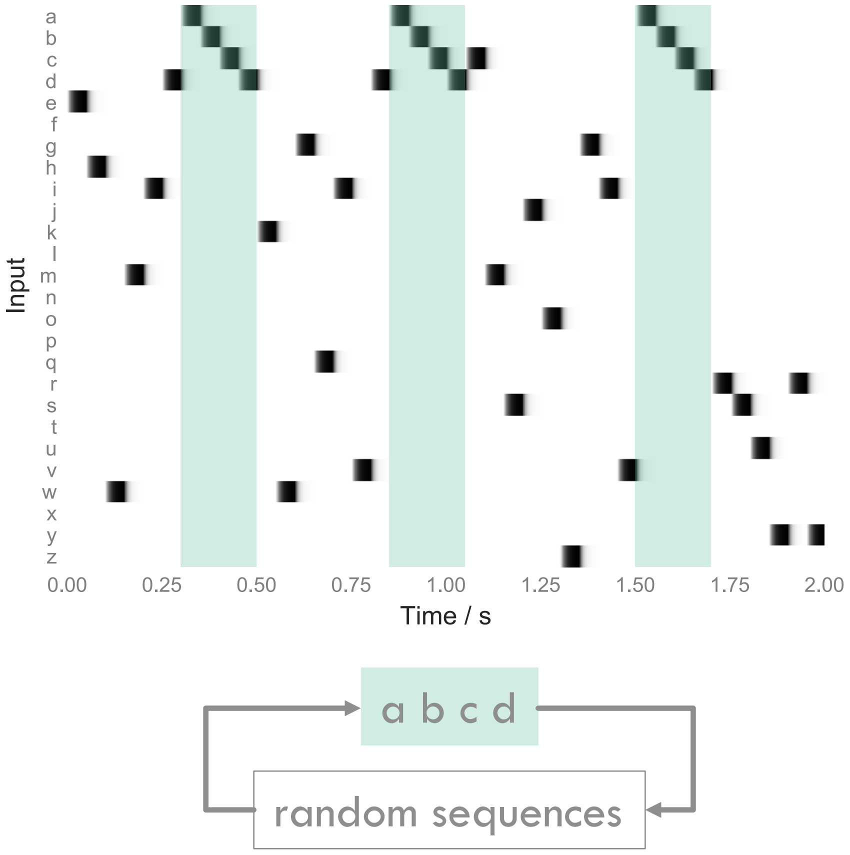
Temporally structured inputs → Temporally predictable dynamics → Mutual prediction *possible*

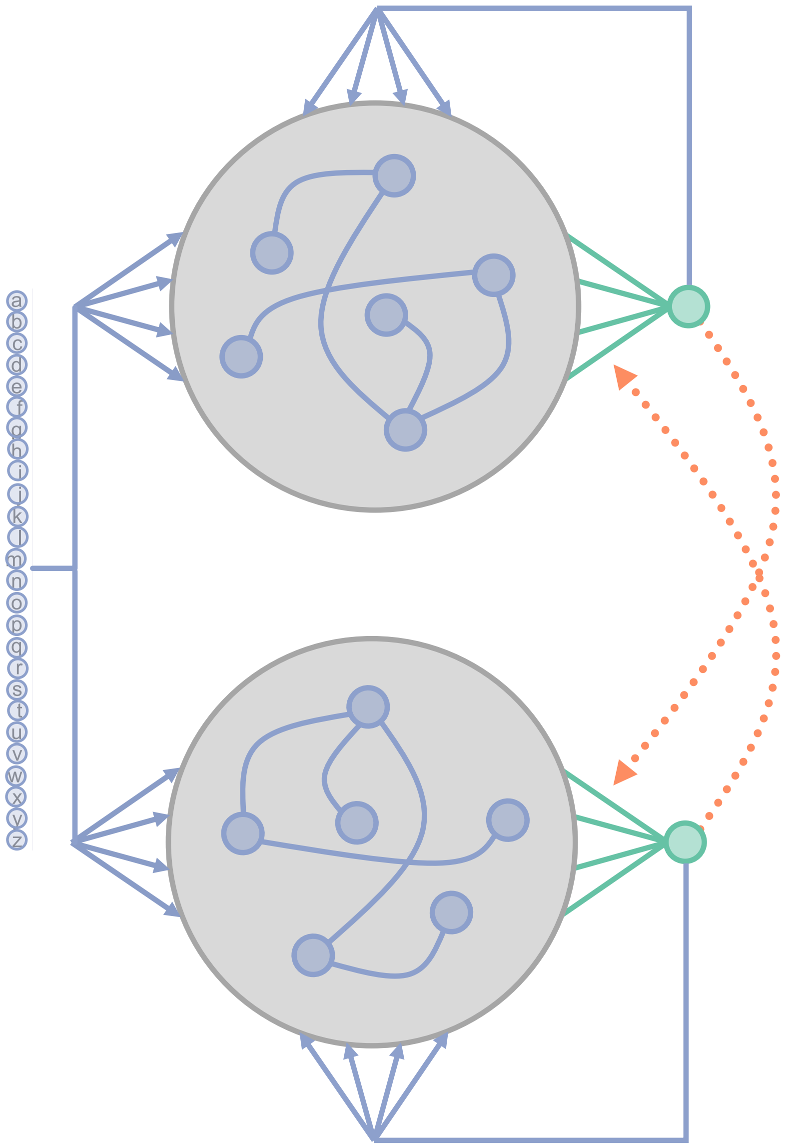
Strictly, the target functions are: $f_1(t) = \left[\tanh \hat{z}_2(t)\right]_+$ $f_2(t) = \left[\tanh \hat{z}_1(t)\right]_+$

Asabuki and Fukai, (2018)

Training rule: Two networks, each tries to predict the other

TRAINING

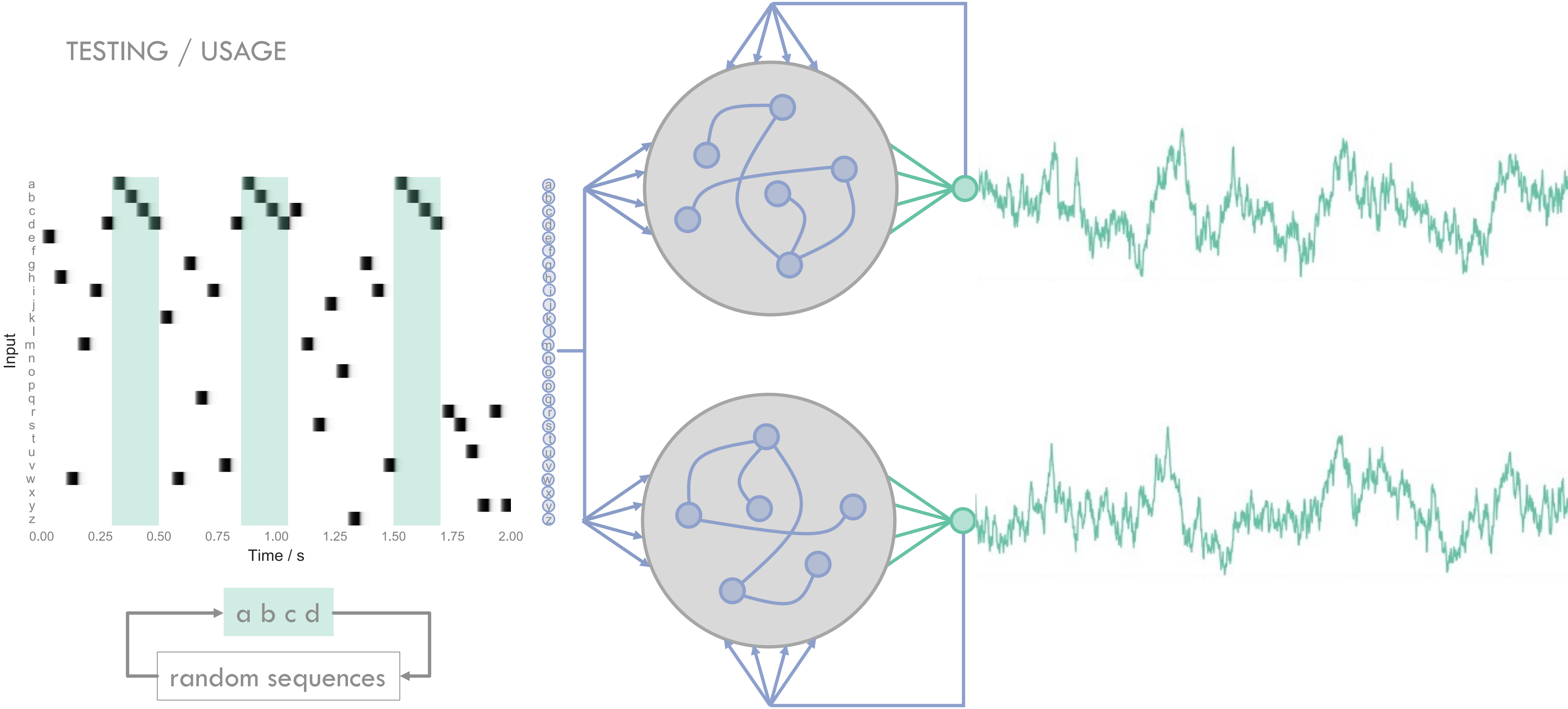




Asabuki and Fukai, (2018)

Once trained the reservoirs can act independently

TESTING / USAGE

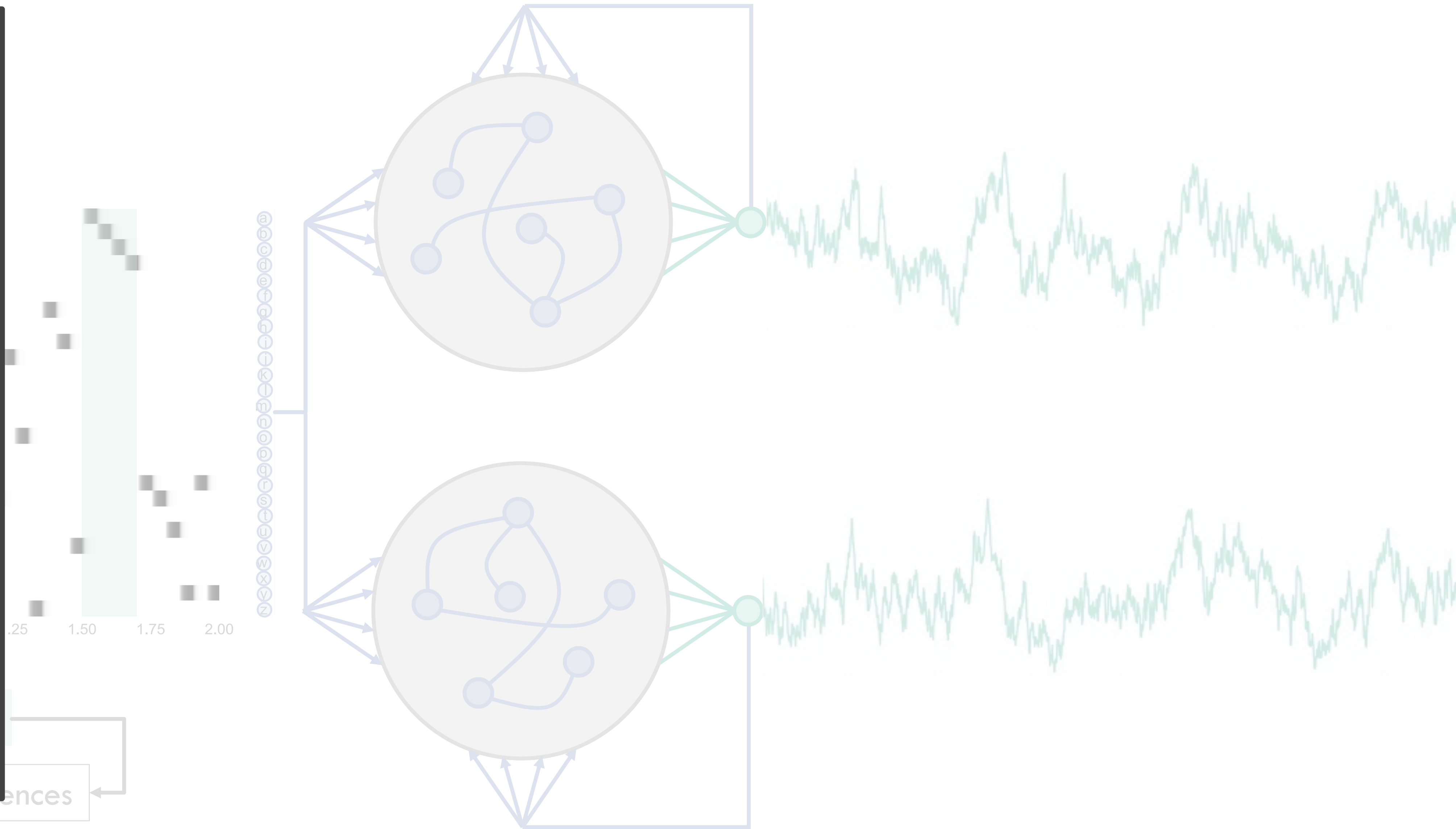


Asabuki and Fukai, (2018)

Once trained the reservoirs can act independently

Intuition for the training rule

- It's impossible to learn a random trajectory (can do no better than predict the mean → z = 0)
- It may* be possible to learn the stereotyped trajectory caused by a recurring sequence or 'chunk'



*It's not obvious why it would "want" to learn (notice $\mathbf{w}_1^{Out} = \mathbf{w}_2^{Out} = \mathbf{0}$ is a valid solution). I have some ideas we could discuss at the end.

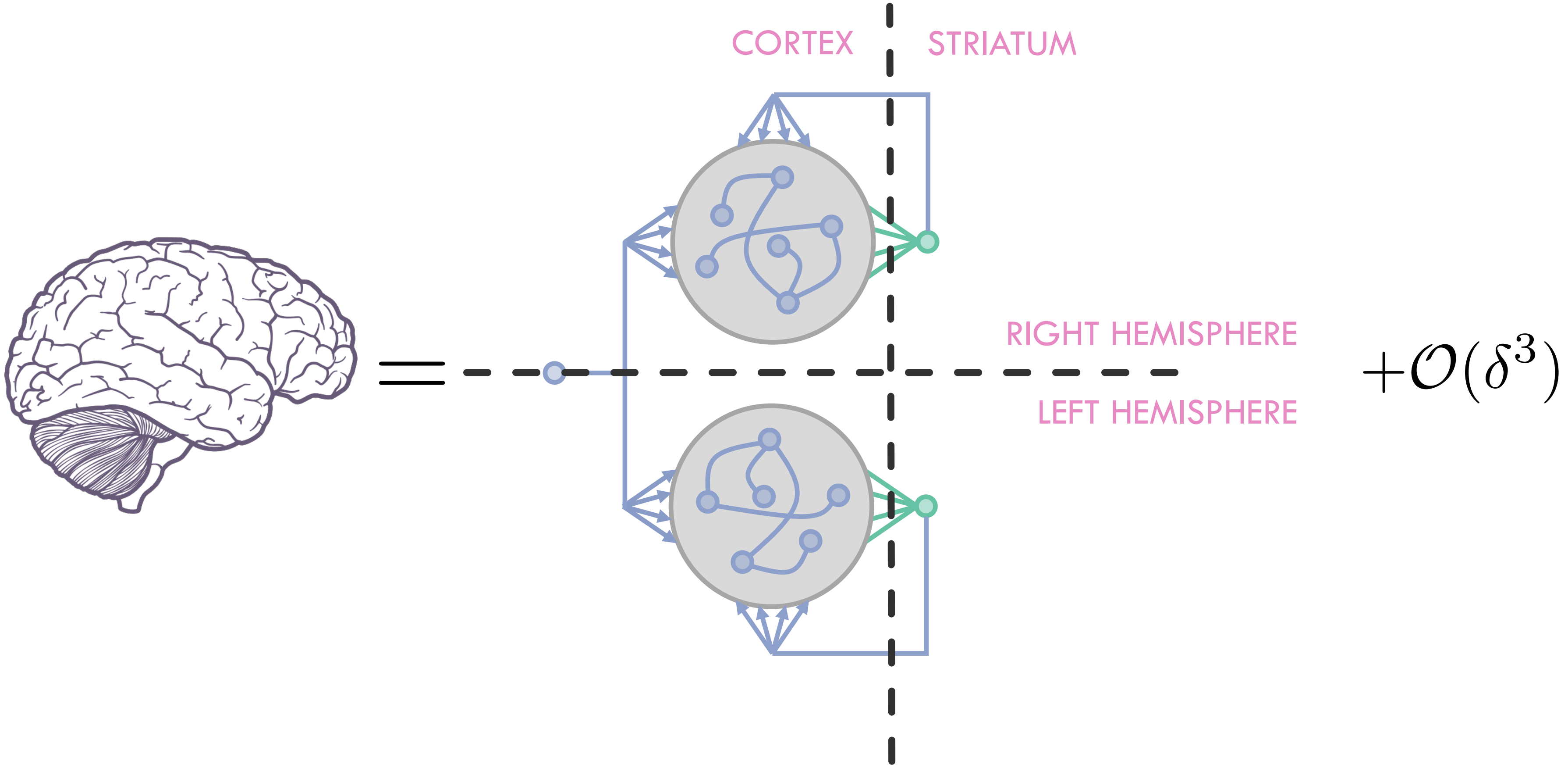
Once trained the reservoirs can act independently

Intuition for the training rule

- It's impossible to learn a random trajectory (can do no better than predict the mean → z = 0)
- It may* be possible to learn the stereotyped trajectory caused by a recurring sequence or 'chunk'

Inspiration for the training rule

If you squint, there's a similarity to cortico-basal ganglia loops



*It's not obvious why it would "want" to learn (notice $\mathbf{w}_1^{\text{Out}} = \mathbf{w}_2^{\text{Out}} = \mathbf{0}$ is a valid solution). I have some ideas we could discuss at the end.

Sainsbury Wellcome Centre

5 key taxonomies of temporal structure

*"how does the brain encode temporal sequences of items, such that this knowledge can be used to **retrieve a sequence from memory**, recognize it, **anticipate on forthcoming items**, and **generalize** this knowledge to **novel sequences** with a similar structure?" – Lashley (1951)*

1. Transition and timing knowledge — ♪♪♪♪♪♪♪♪♪♪♪♪ ?
2. Chunking — gopilagikobatokibutokibugikobagopila
3. Ordinal knowledge — 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
4. Algebraic patterns — mimitu (AAB) totobu (AAB) gagari (AAB) pesipe (ABA) pipigo (AAB)
5. Nested tree structures generate by symbolic rules — $A + B\sin\omega t$

Dehaene et al. (2015)

5 key taxonomies of temporal structure

"how does the brain encode temporal sequences of items, such that this knowledge can be used to ***retrieve a sequence from memory****, recognize it,* ***anticipate on forthcoming items****, and* ***generalize*** *this knowledge to* ***novel sequences*** *with a similar structure?" – Lashley (1951)*

✓ 1. Transition and timing knowledge — ♪♪♪♪♪♪♪♪♪♪♪♪ ?

✓ 2. Chunking — gopilagikobatokibutokibugikobagopila

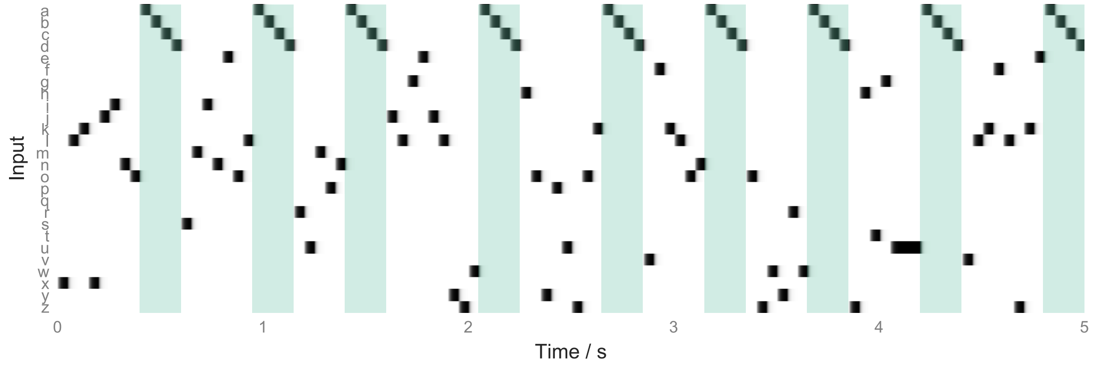
? 3. Ordinal knowledge — 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3

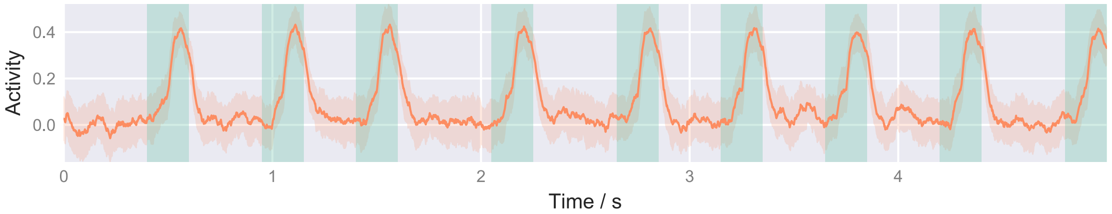
✗ 4. Algebraic patterns — mimitu (AAB) totobu (AAB) gagari (AAB) pesipe (ABA) pipigo (AAB)

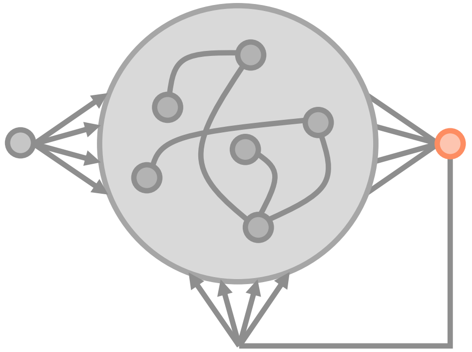
✗ 5. Nested tree structures generate by symbolic rules — $A + B \sin \omega t$

Dehaene et al. (2015)

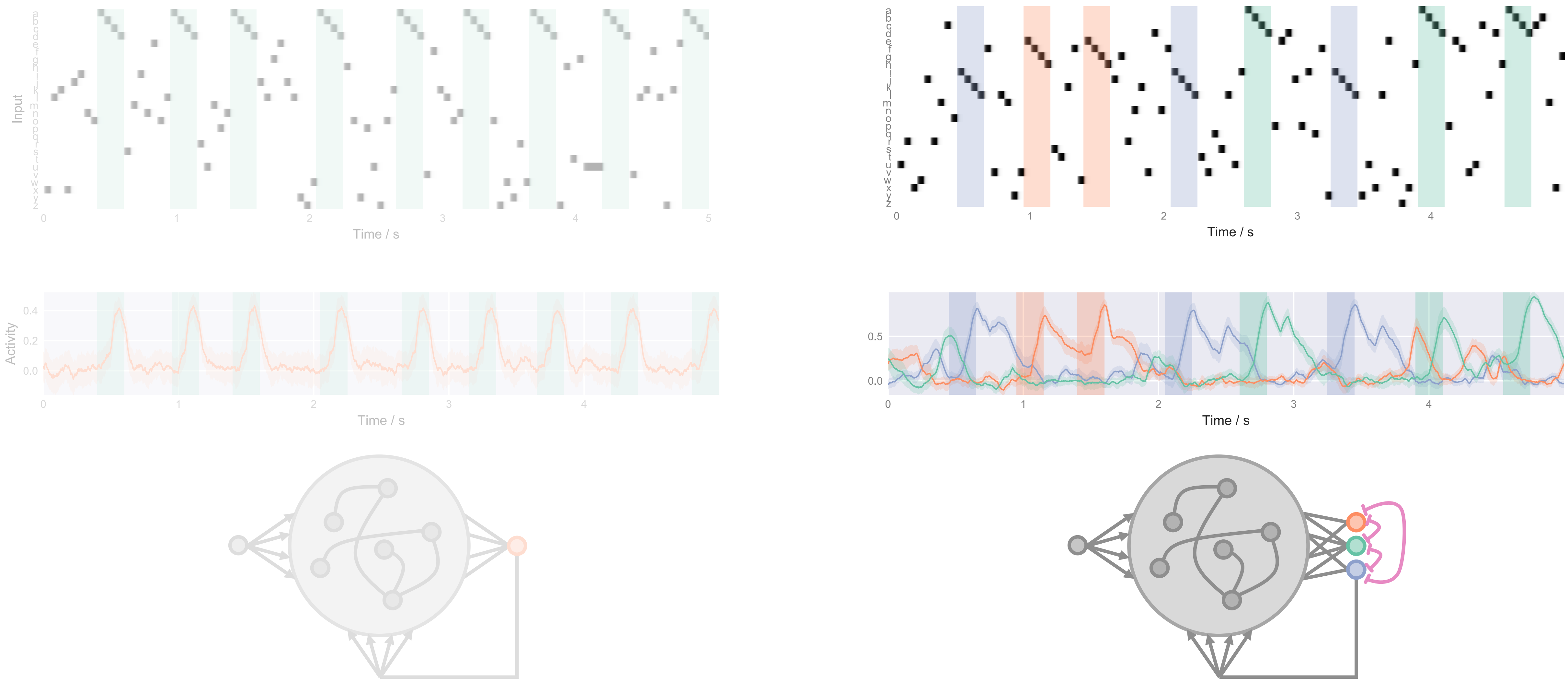
2. Chunking, aka 'event segmentation'



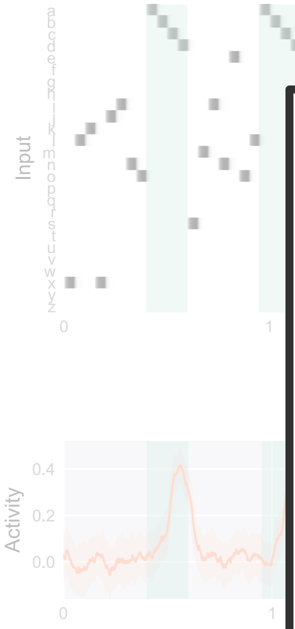




2. Chunking, aka ‘event segmentation’



2. Chunking, aka 'event segmentation'



Explanations of chunking:

1. Transition probability

   Saffran et al. (1996)

2. Temporal community structure

   Schapiro et al. (2013)

2. Chunking, aka 'event segmentation'

Explanations of chunking:

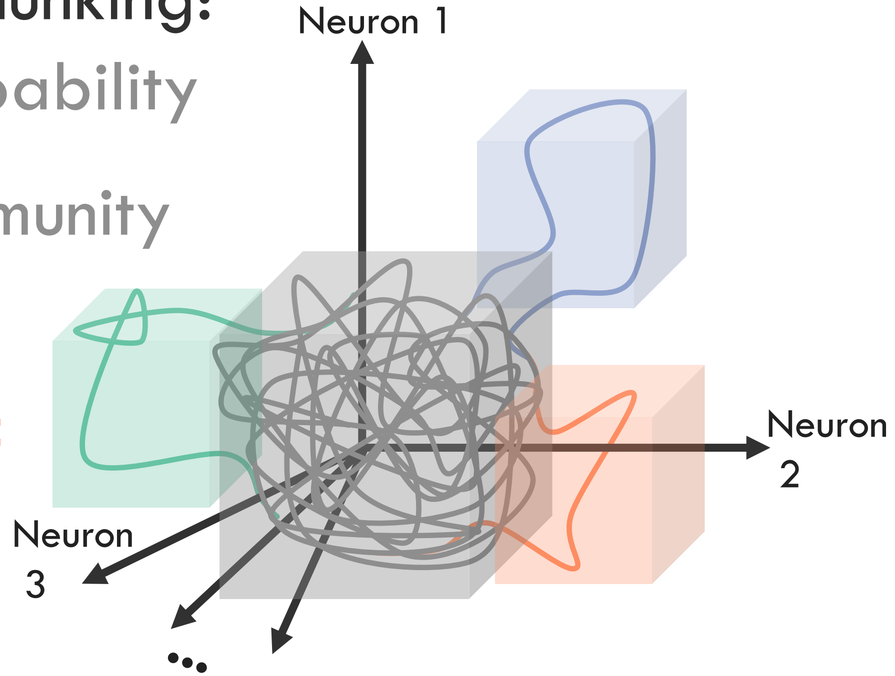
1. Transition probability

   Saffran et al. (1996)

2. Temporal community structure

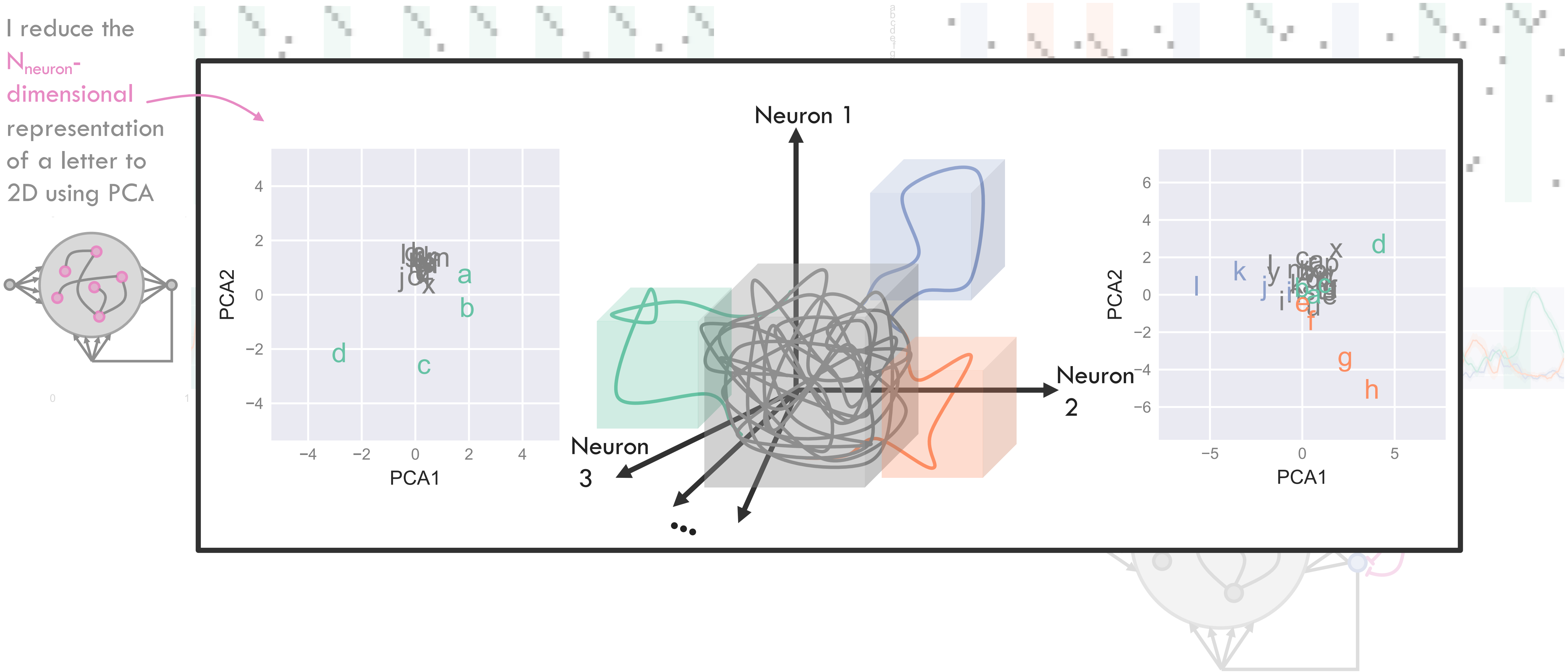
   Schapiro et al. (2013)

2a. Chunk-specific predictable trajectory



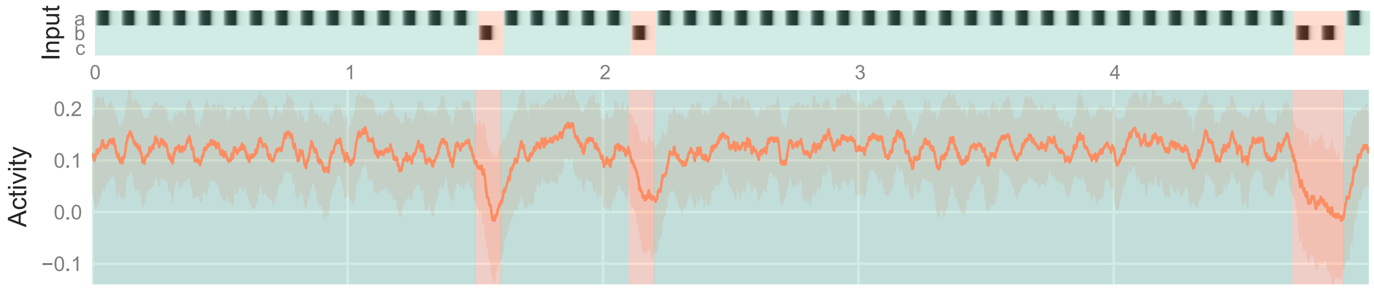
2. Chunking, aka ‘event segmentation’

I reduce the N_{neuron}-dimensional representation of a letter to 2D using PCA



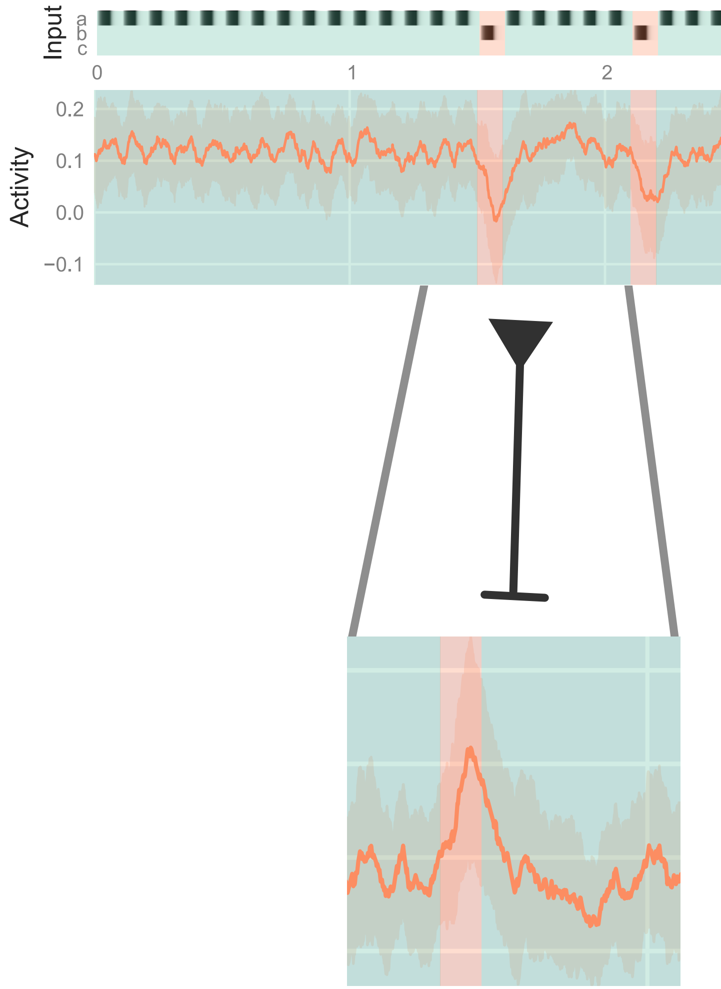
1. Transition and timing knowledge

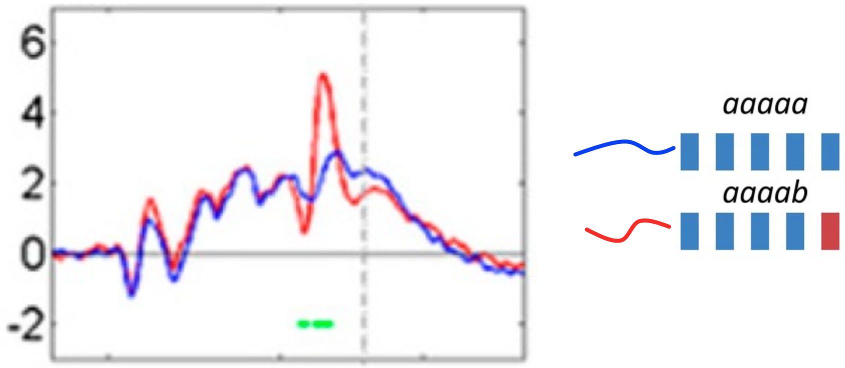
AAAAAAX



1. Transition and timing knowledge

AAAAAAX

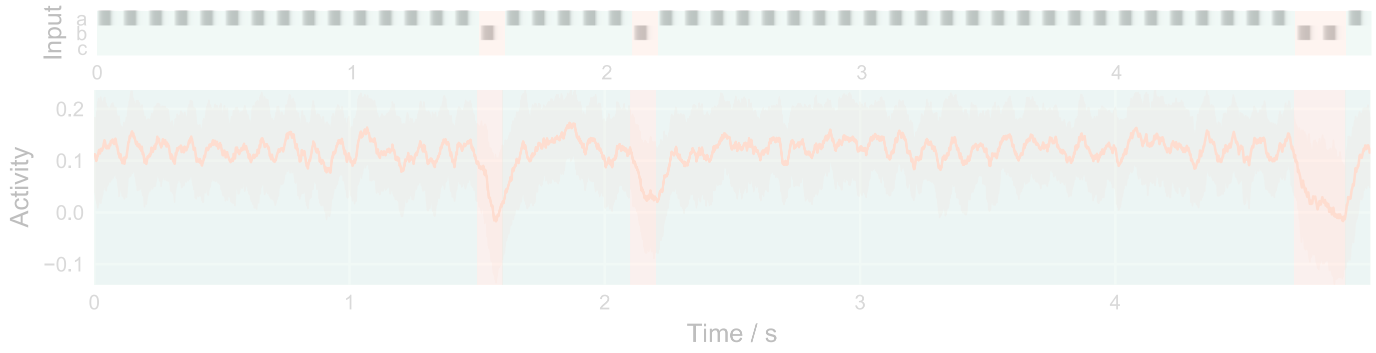




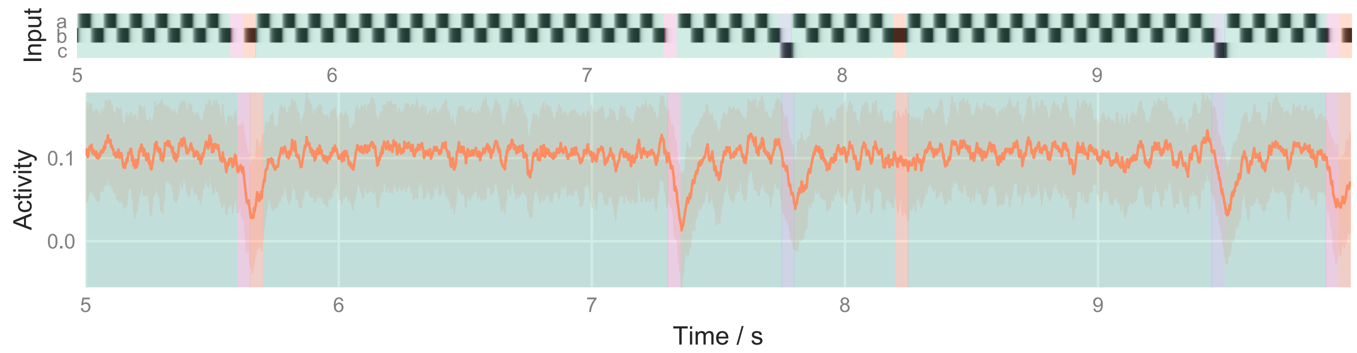
Strauss et al. (2015)

1. Transition and timing knowledge

AAAAAAX



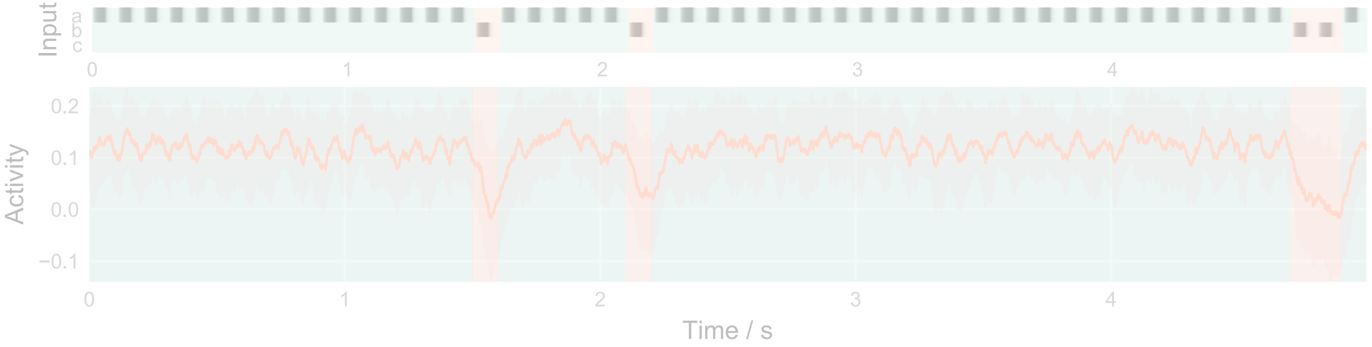
ABABABX



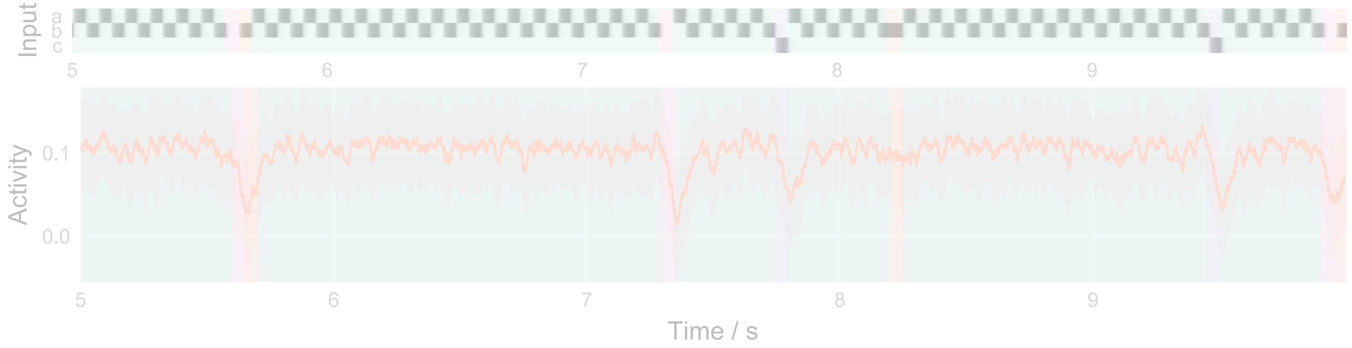
MMR response to unseen (but not unexpected) stimuli

1. Transition and timing knowledge

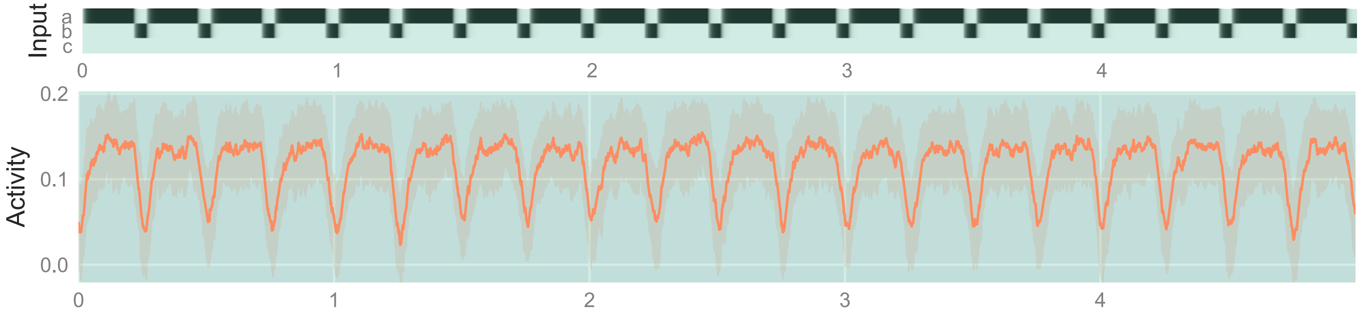
AAAAAAX



ABABABX



AAAABAAAAB



Continued presence of MMR to B replicates finding in *Strauss et al.* (2015).

1. Transition and timing knowledge

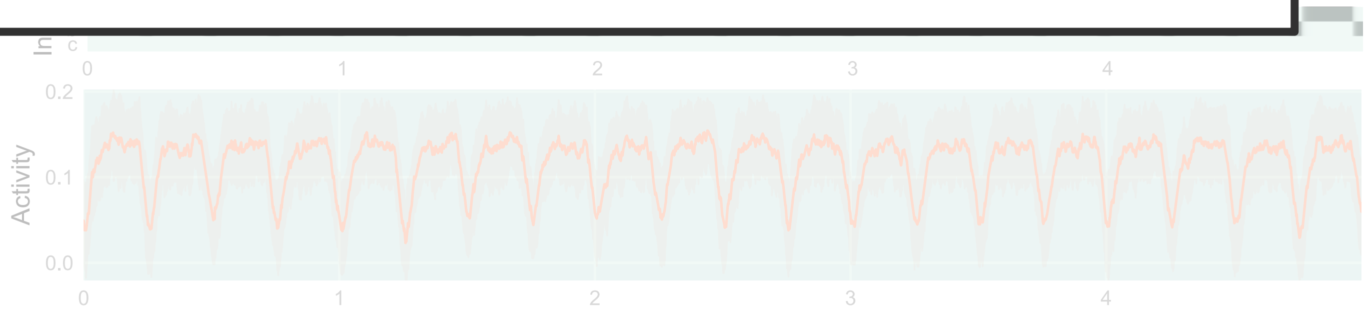
AAAAAAX

Explanations of MMR:

1. Stimulus-specific adaptation
   May et al. (2010)
2. Predictive coding
   Friston (2005)

ABABABX

AAAABAAAAB



Continued presence of MMR to B replicates finding in *Strauss et al. (2015)*.

1. Transition and timing knowledge

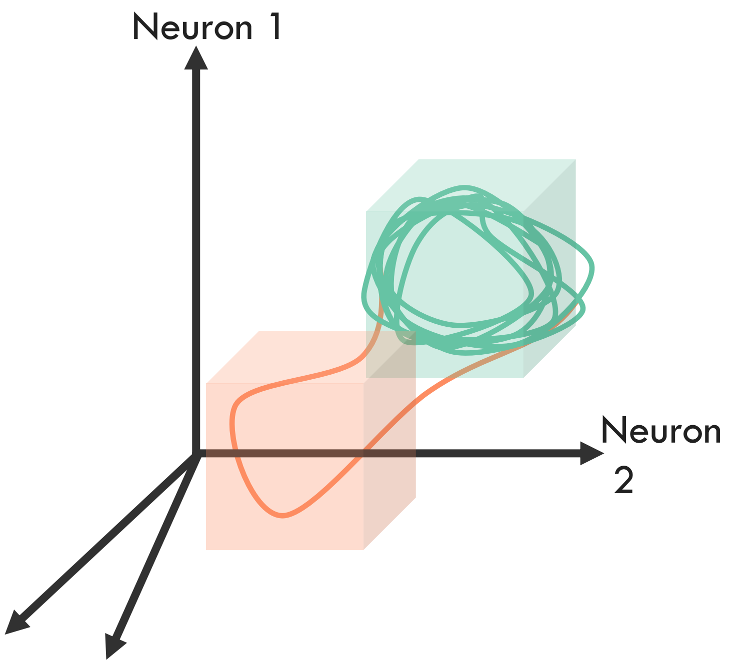
AAAAAAX

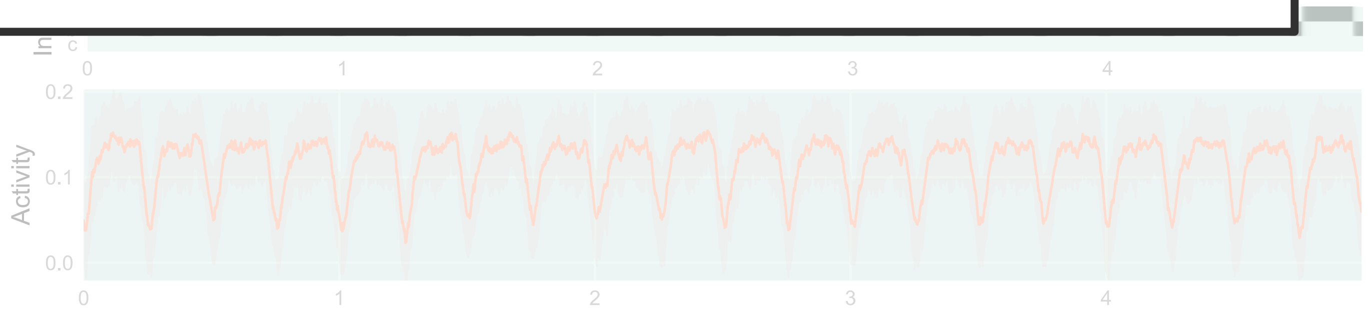
ABABABX

AAAABAAAAB

Explanations of MMR:

1. Stimulus-specific adaptation
   May et al. (2010)
2. Predictive coding
   Friston (2005)
3. (or 2a) Disruption of otherwise stabilised recurrent dynamics





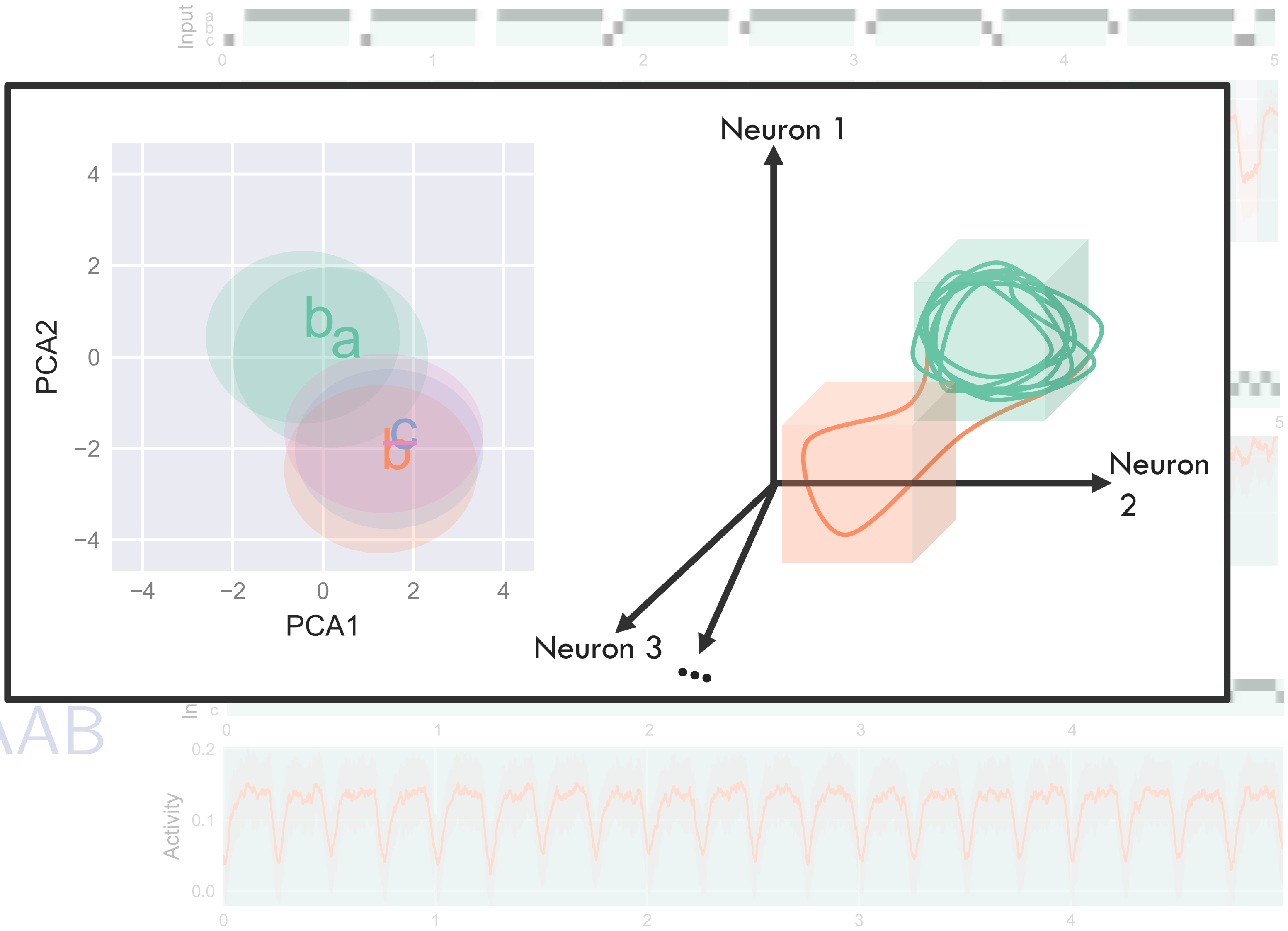
Continued presence of MMR to B replicates finding in *Strauss et al. (2015)*.

1. Transition and timing knowledge

AAAAAAX

ABABABX

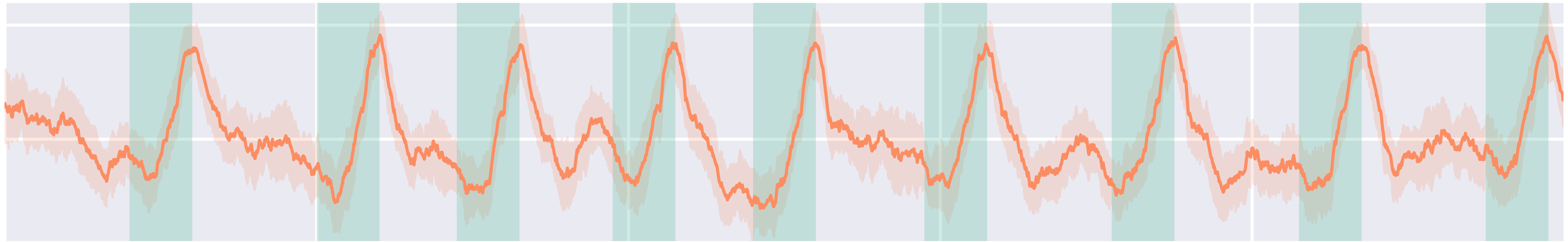
AAAABAAAAB



Continued presence of MMR to B replicates finding in *Strauss et al.* (2015).

3. Ordinal position

Ramping evidence in favour of chunk gives info on **last**, but not **first**, ordinal position



4. Algebraic patterns

mimitu totobu gagari pesipe pipigo

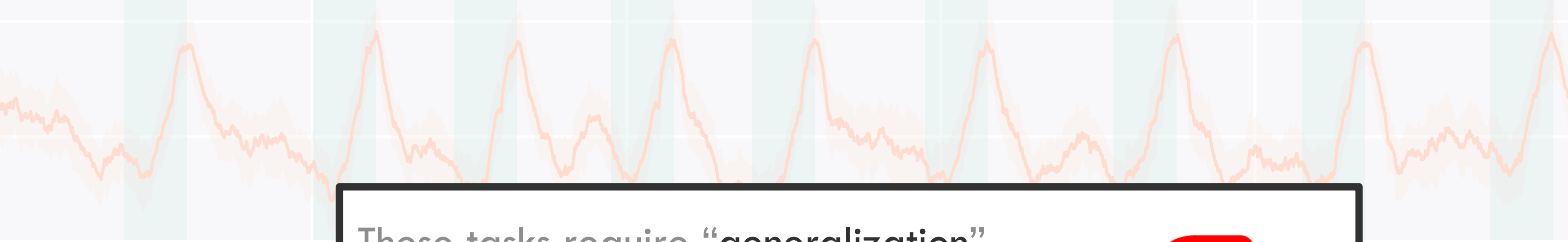
AAB AAB AAB ABA AAB

5. Nested tree structure

$$A + B \sin \omega t$$

3. Ordinal position

Ramping evidence in favour of chunk gives info on **last**, but not **first**, ordinal position



These tasks require "**generalization**"

This model isn't **expressive** enough to learn the **latent structure** required.

4. Algebraic pattern

mimitu totobu g

AAB AAB AAB ABA AAB

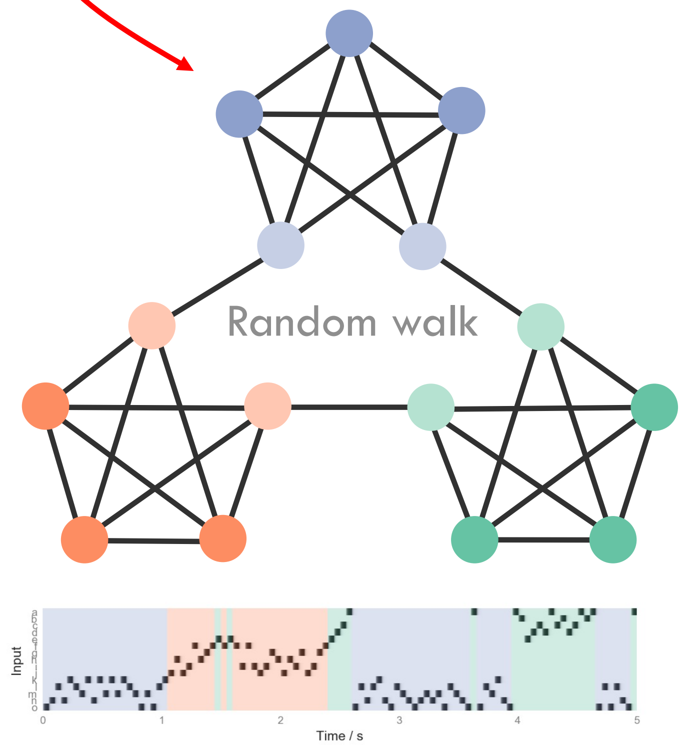
5. Nested tree structure

$A + B \sin \omega t$

Representations reflect temporal community structure… like in the brain

A naïve method for chunking: If your ability to predict what's coming next suddenly falls, it' probably because you're at the end of a chunk

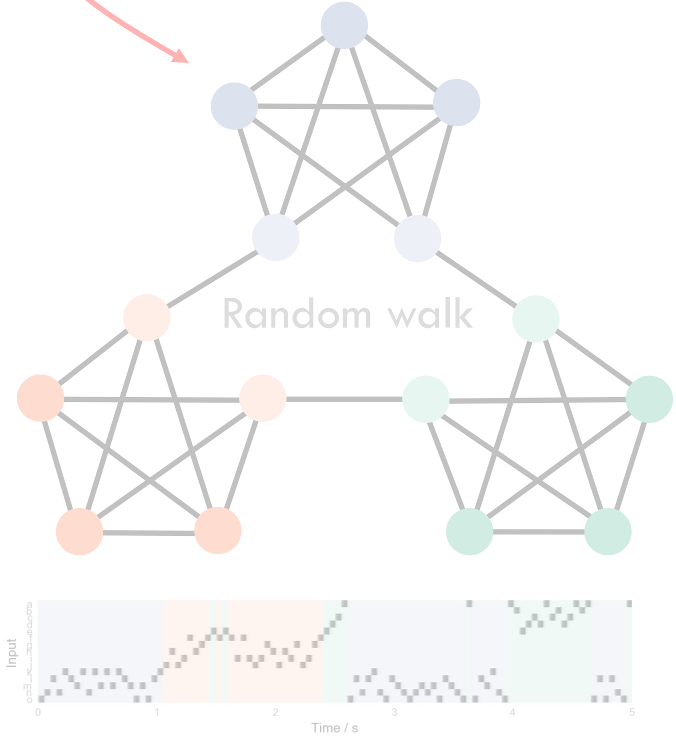
i.e. it fails here



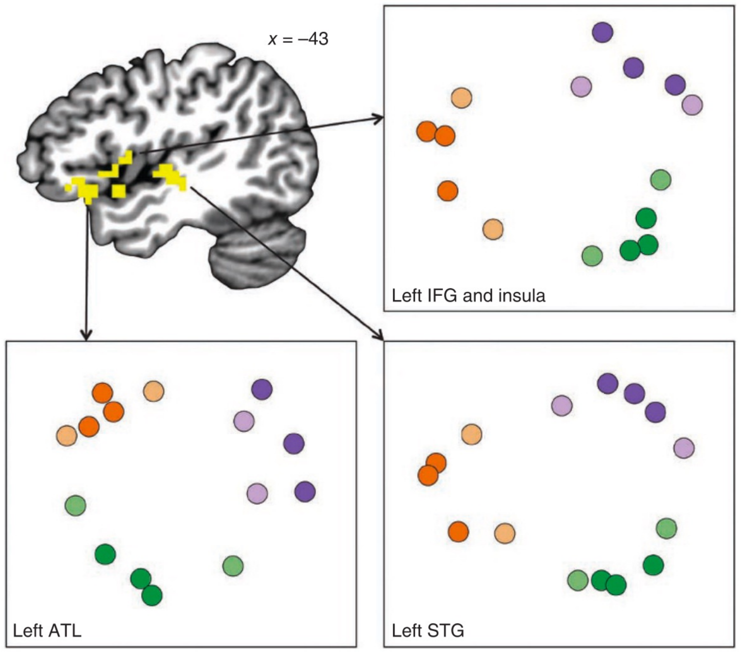
Representations reflect temporal community structure… like in the brain

A naïve method for chunking: If your ability to predict what's coming next suddenly falls, it' probably because you're at the end of a chunk

i.e. it fails here



THE BRAIN



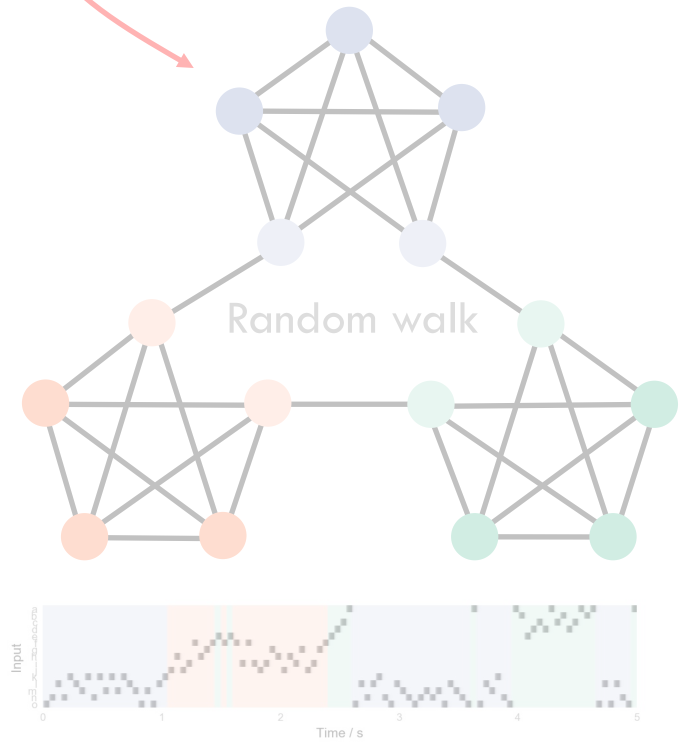
Schapiro et al. (2013)

An improved method for chunking: If two events repeatedly occur together in time, learn representations whose similarity respects this.

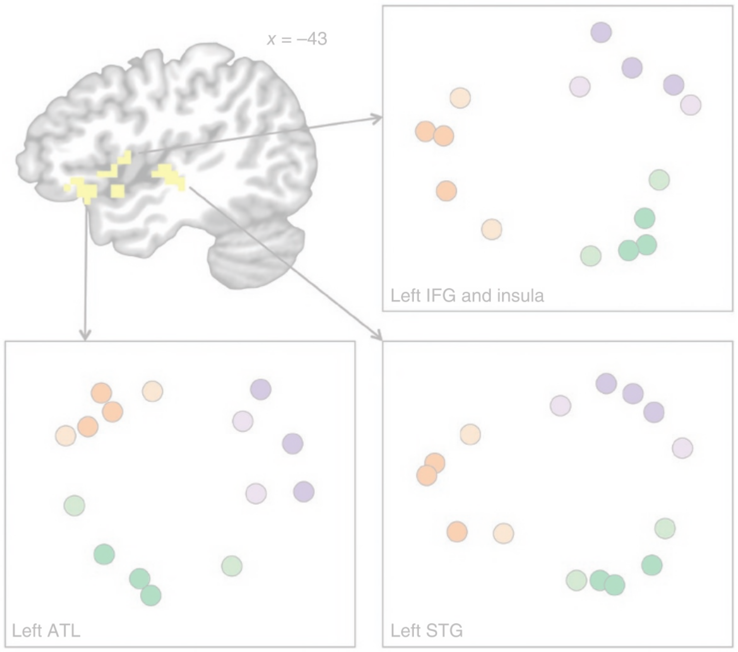
Representations reflect temporal community structure… like in the brain

A naïve method for chunking: If your ability to predict what's coming next suddenly falls, it' probably because you're at the end of a chunk

i.e. it fails here



THE BRAIN



Schapiro et al. (2013)

RESERVOIR NETWORK

✅ • Can it chunk the random walk?

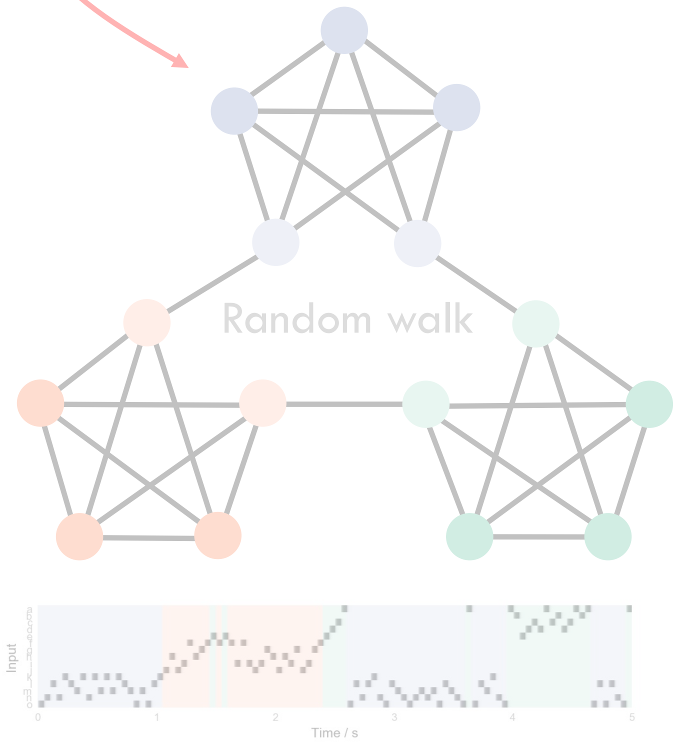
✅ • Will the representation respect temporal community structure…i.e. look like the brain?

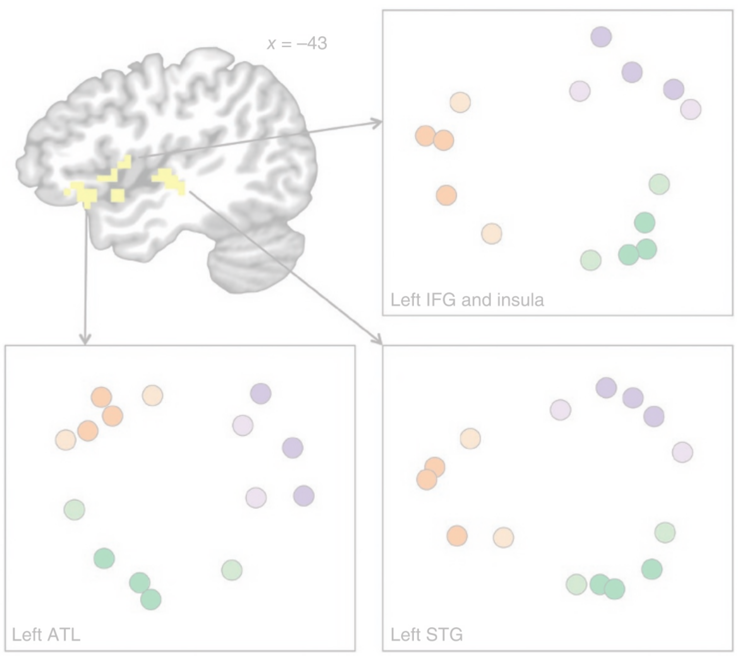
An improved method for chunking: If two events repeatedly occur together in time, learn representations whose similarity respects this.

Representations reflect temporal community structure… like in the brain

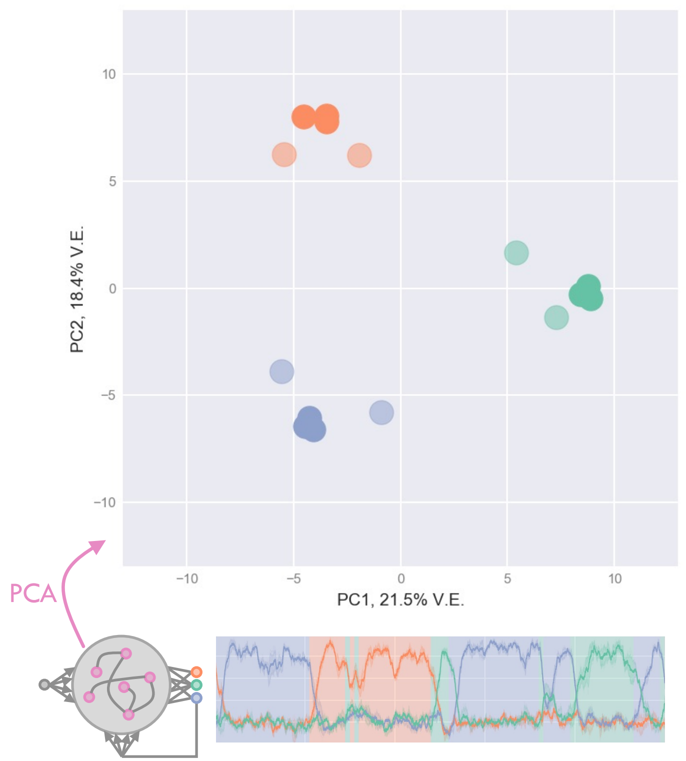
A naïve method for chunking: If your ability to predict what's coming next suddenly falls, it' probably because you're at the end of a chunk

i.e. it fails here



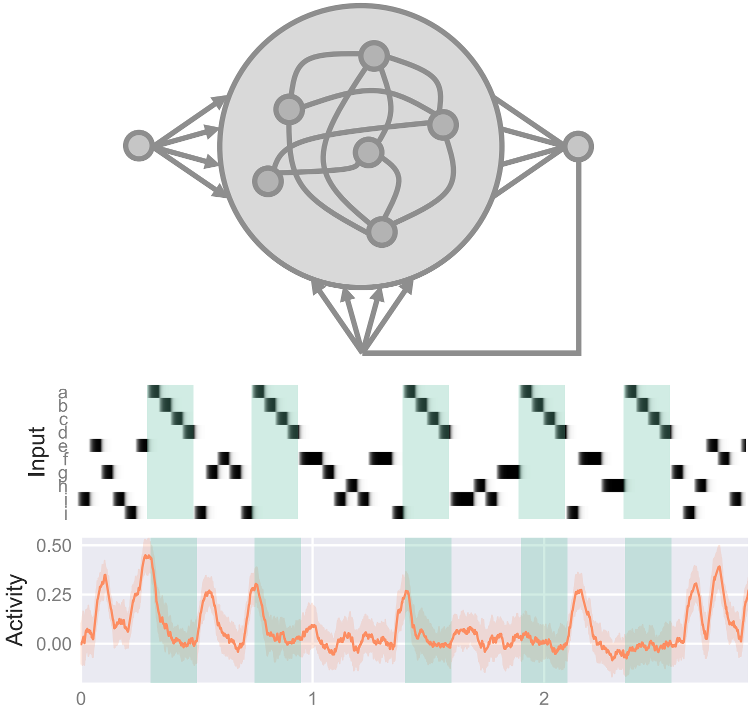


Schapiro et al. (2013)

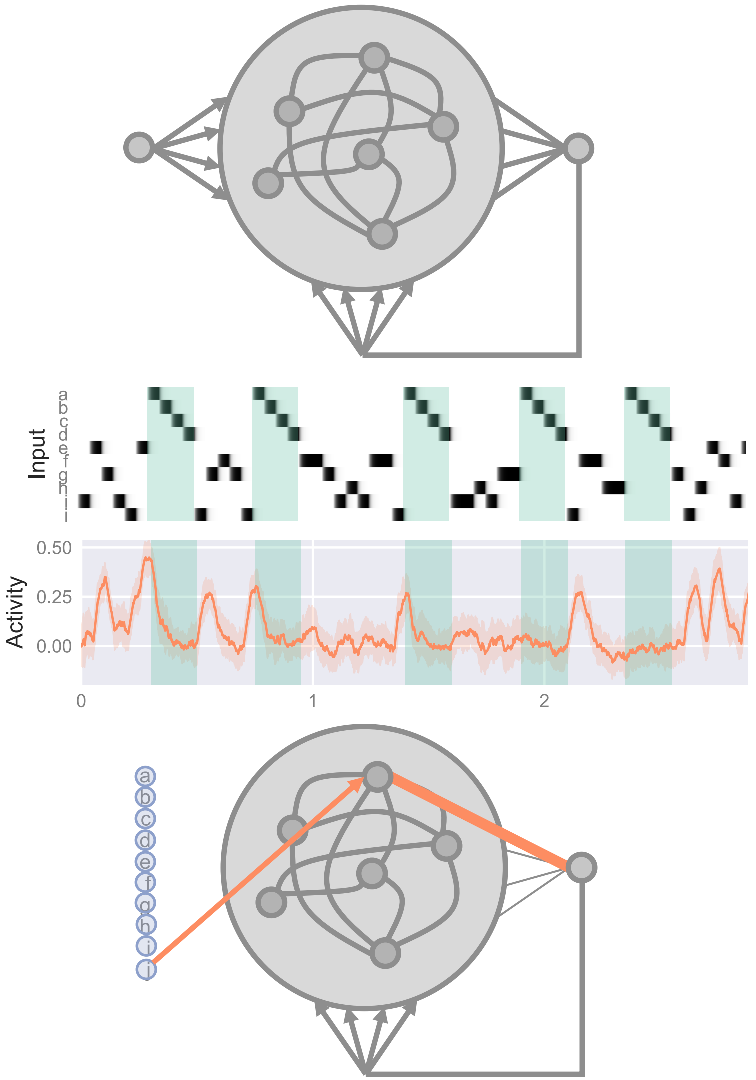


An improved method for chunking: If two events repeatedly occur together in time, learn representations whose similarity respects this.

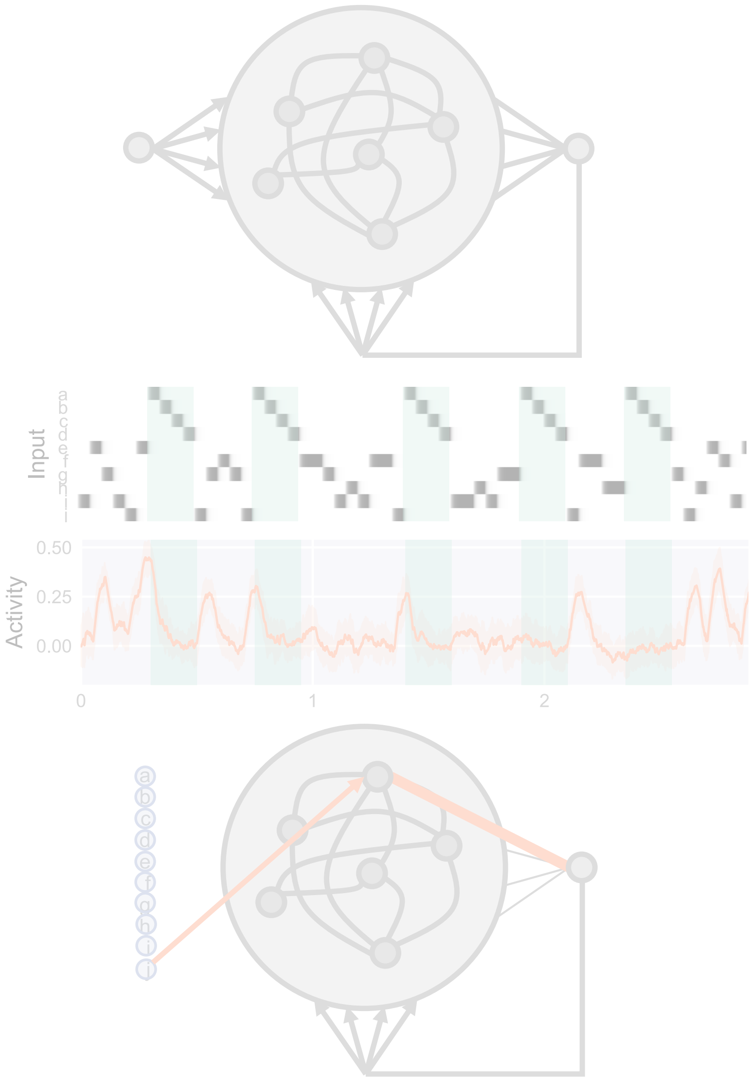
Chunking is improved when the network is forced to engage dynamics

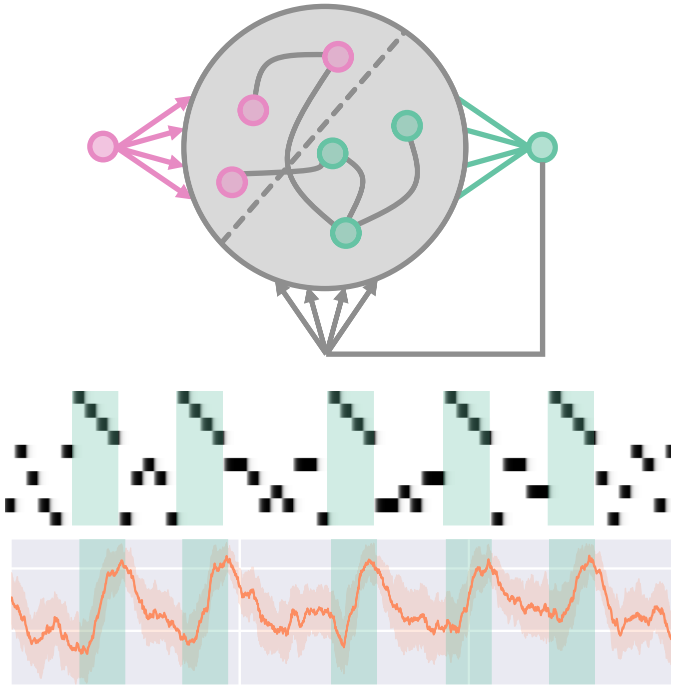


Chunking is improved when the network is forced to engage dynamics



Chunking is improved when the network is forced to engage dynamics

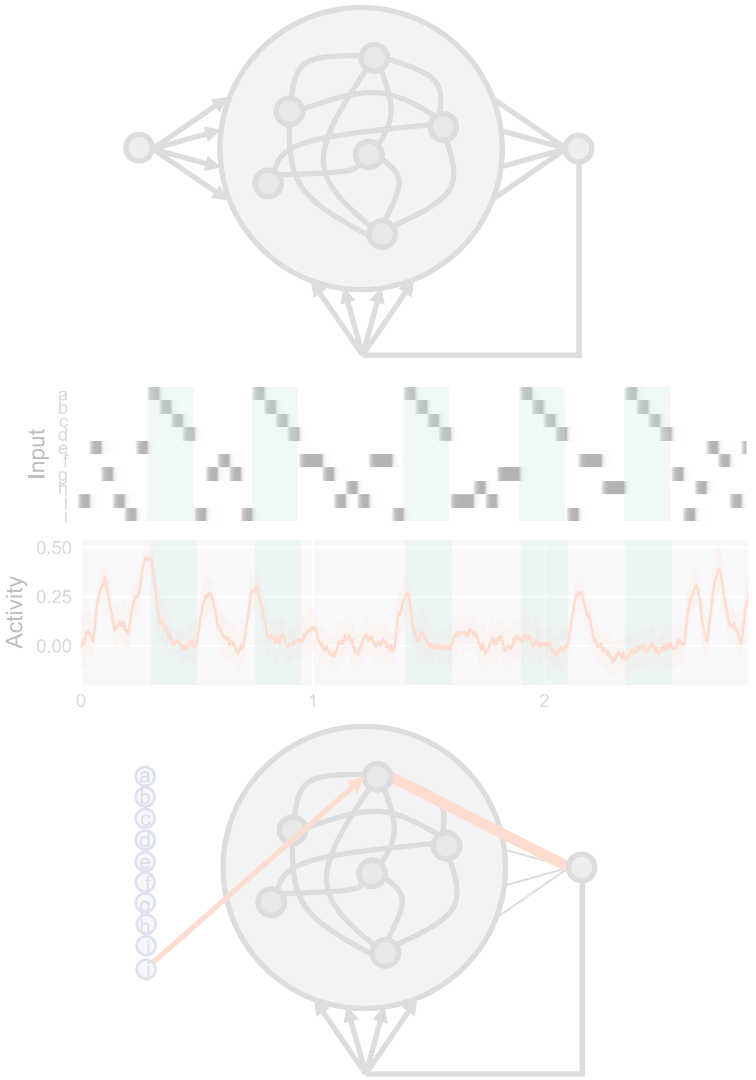


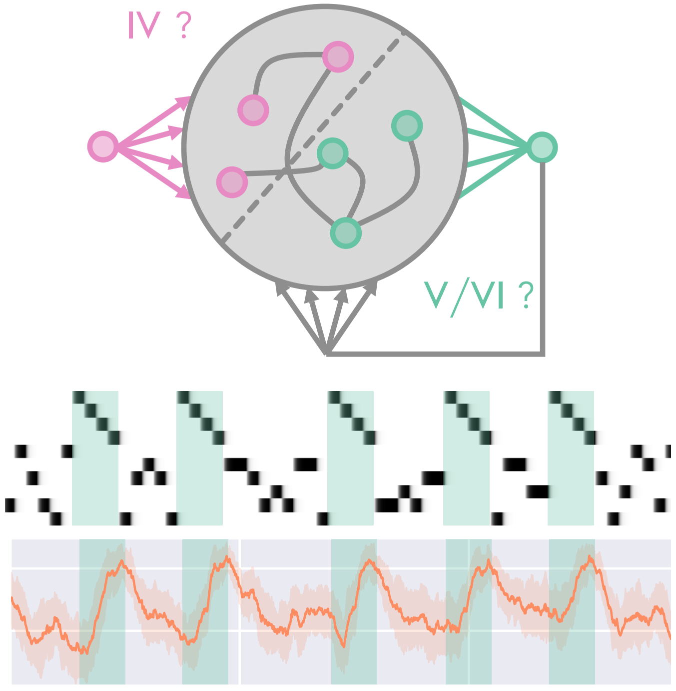


Encourage dynamics by:

- Increasing sparsity
- Splitting inputs and outputs

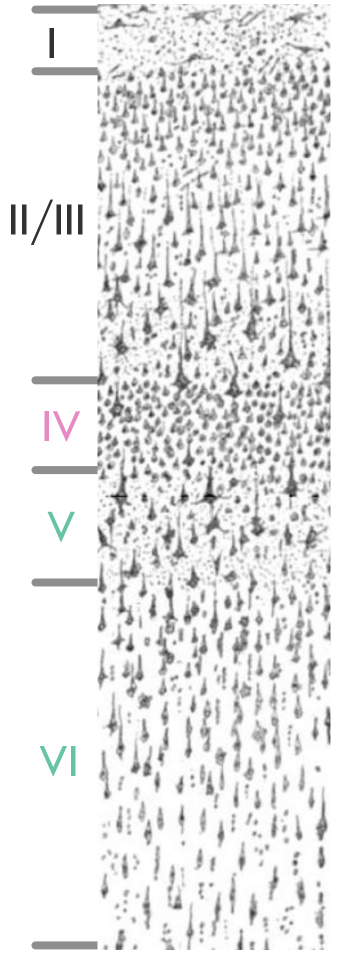
Chunking is improved when the network is forced to engage dynamics





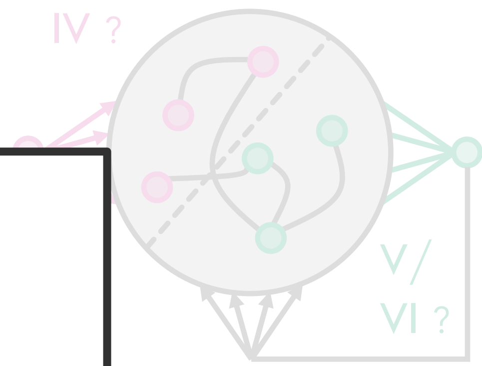
Encourage **dynamics** by:

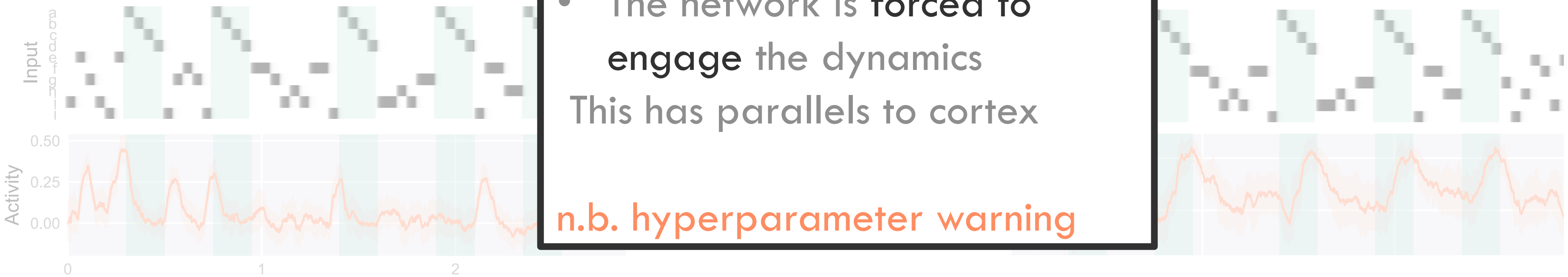
- Increasing **sparsity**
- **Splitting** inputs and outputs



Ramon y Cajal (1911)

Chunking is improved when the network is forced to engage dynamics





Summary:

Chunking is improved when

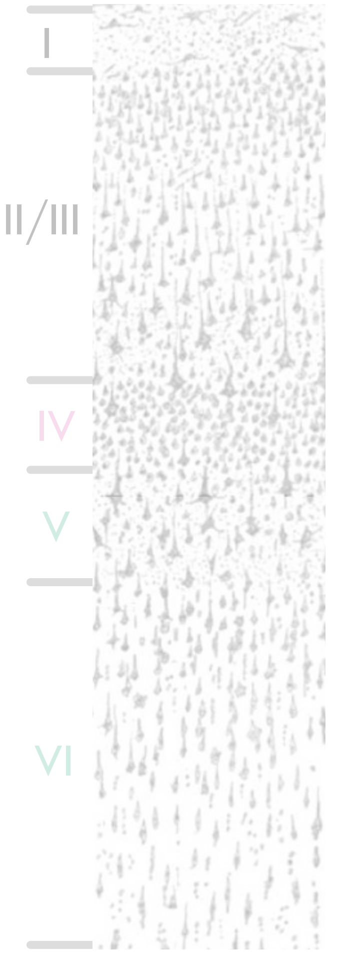
- There is **richer dynamics**
- The network is **forced to engage** the dynamics

This has parallels to cortex

n.b. hyperparameter warning

Encourage **dynamics** by:

- Increasing **sparsity**
- **Splitting** inputs and outputs

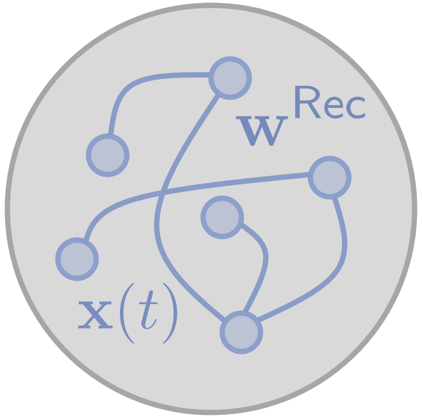


Ramon y Cajal (1911)

Roadmap

1. A reservoir network model for temporal structure learning
2. The role of chaos
3. Experimental results and modelling predictions
4. Conclusions

The role of chaos



$$\mathbf{W}_{ij}^{\text{Rec}} \sim \mathcal{N}(0, \frac{g}{\sqrt{N}})$$

g determines dynamics in a self-driven network

- g < 1 → only transient dynamics
- g > 1 → rich, possibly chaotic, dynamics

Sompolinksy, 1988

We choose g = 1.5

The role of chaos

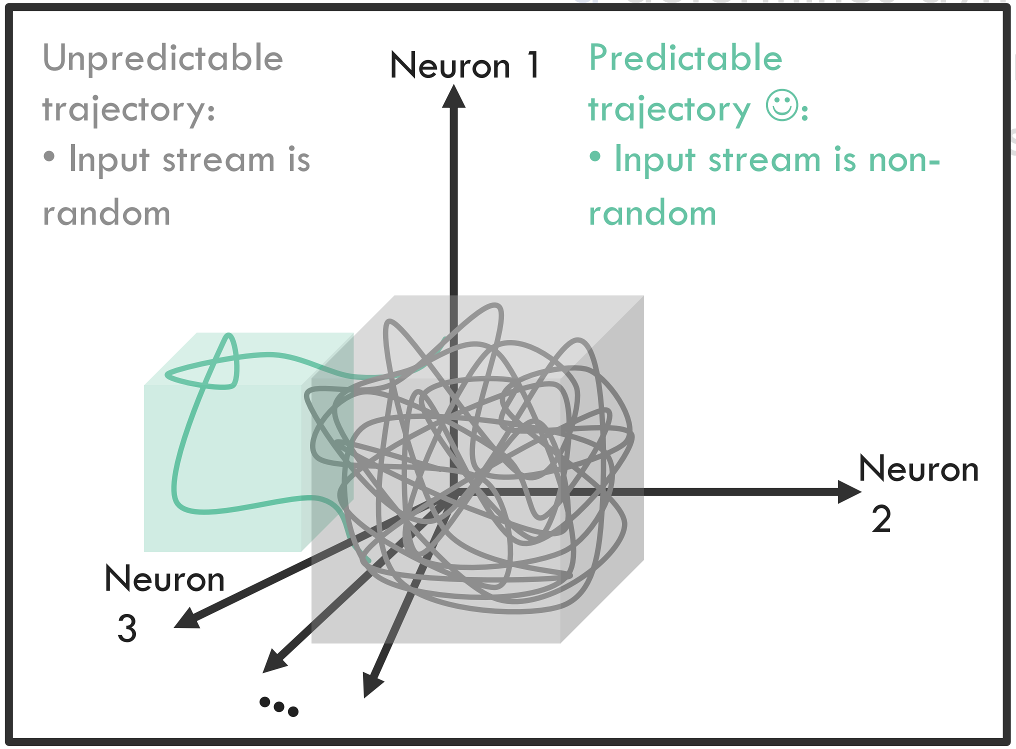
$\mathbf{W}_{ij}^{\text{Rec}}$

g determines dynamics in a self-driven network

...nsient dynamics

...ssibly chaotic, dynamics

Sompolinksy, 1988



The role of chaos

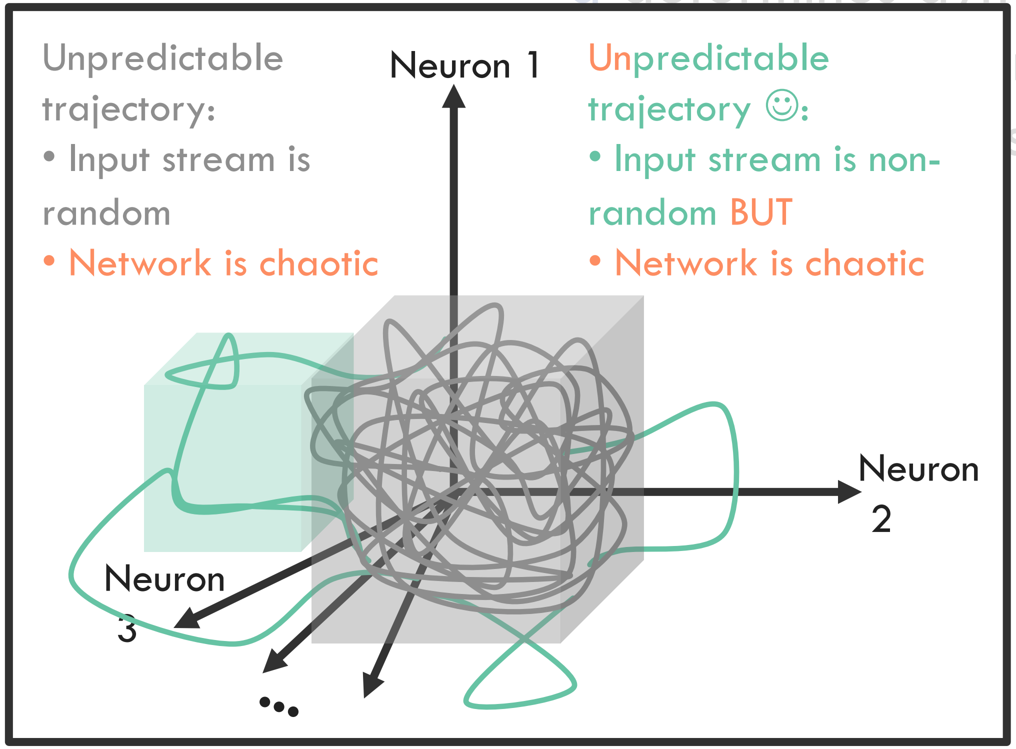
W_{ij}^{Rec}

g determines dynamics in a self-driven network

nsient dynamics

ssibly chaotic, dynamics

Sompolinksy, 1988



Sainsbury Wellcome Centre

UCL

The role of chaos

$$\mathbf{W}^{\text{Rec}}_{ij} \sim \mathcal{N}(0, \frac{g}{\sqrt{N}})$$

g determines dynamics in a self-driven network

- g < 1 → only transient dynamics
- g > 1 → rich, possibly chaotic, dynamics

Sompolinksy, 1988

We choose g = 1.5

Strong **inputs**, **noise** and **feedback** can all **suppress chaos.**

e.g. *Rajan et al. (2010)*, or Francesca's work. Intuition is that more external inputs and less recurrent 'self-talk' leads to more stable dynamics

The role of chaos

$$\mathbf{W}_{ij}^{\text{Rec}} \sim \mathcal{N}(0, \frac{g}{\sqrt{N}})$$

g determines dynamics in a self-driven network

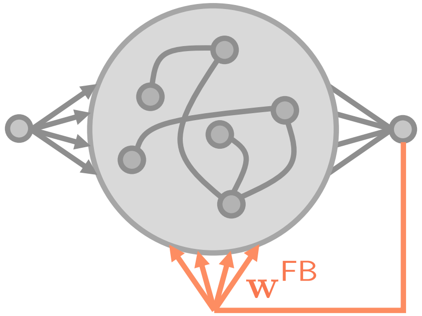
- g < 1 → only transient dynamics
- g > 1 → rich, possibly chaotic, dynamics

Sompolinksy, 1988

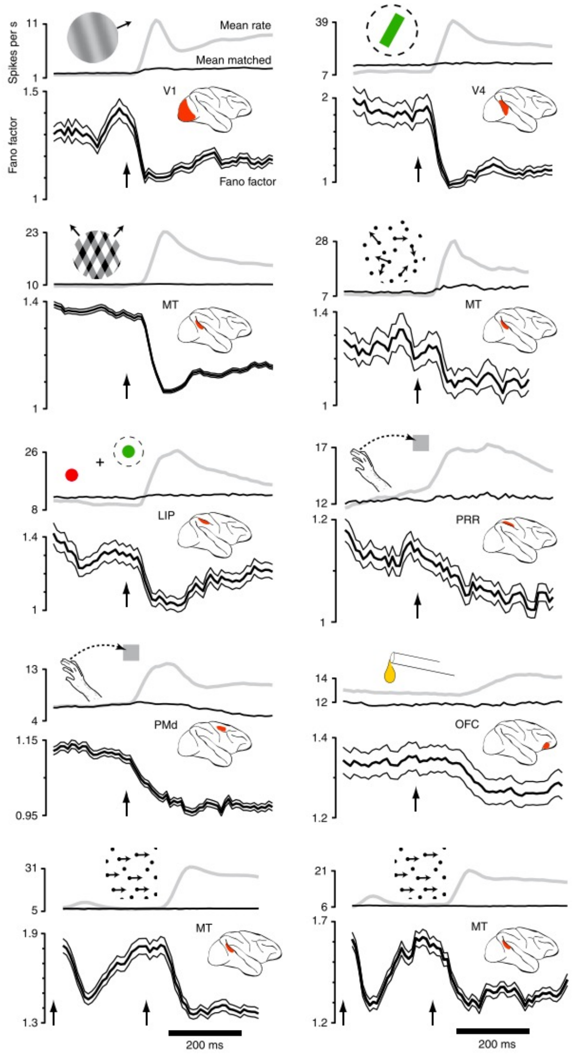
We choose g = 1.5

Strong inputs, noise and feedback can all suppress chaos.

e.g. Rajan et al. (2010), or Francesca's work. Intuition is that more external inputs and less recurrent 'self-talk' leads to more stable dynamics

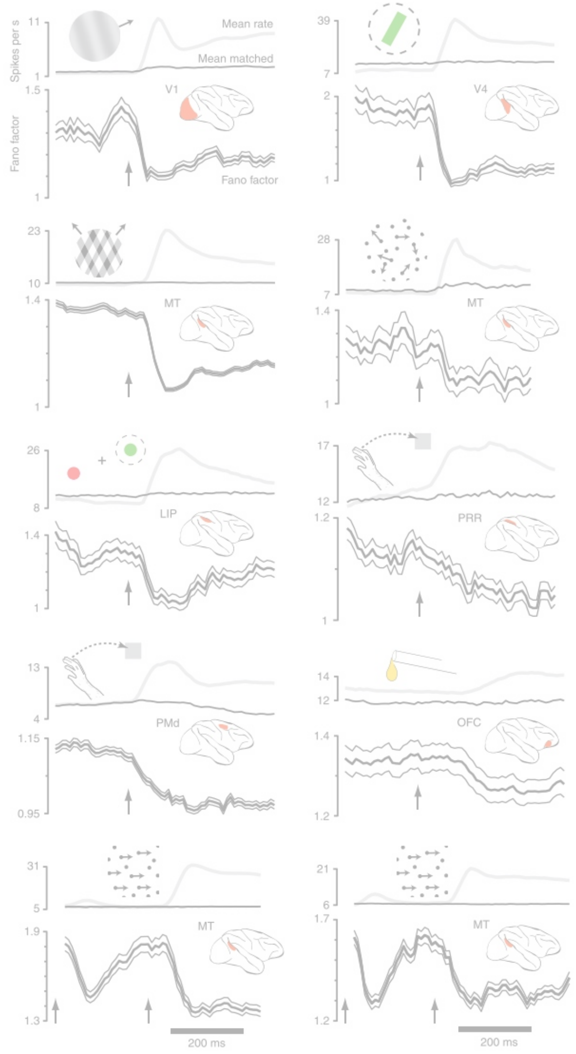


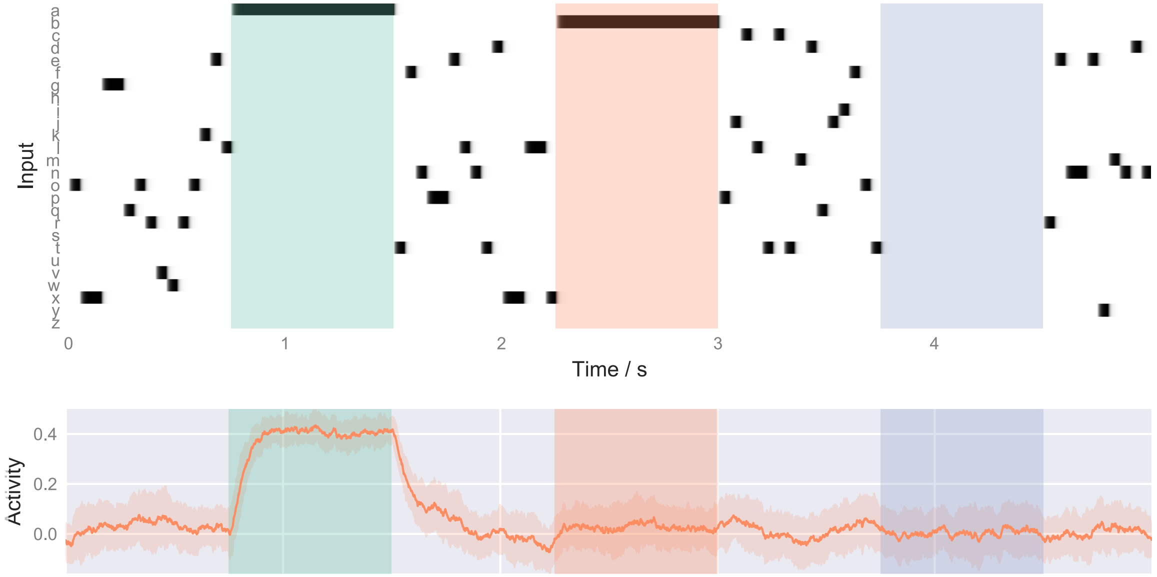
Stimulus onset quenches neural variability



Churchland…Sahani et al. (2010)

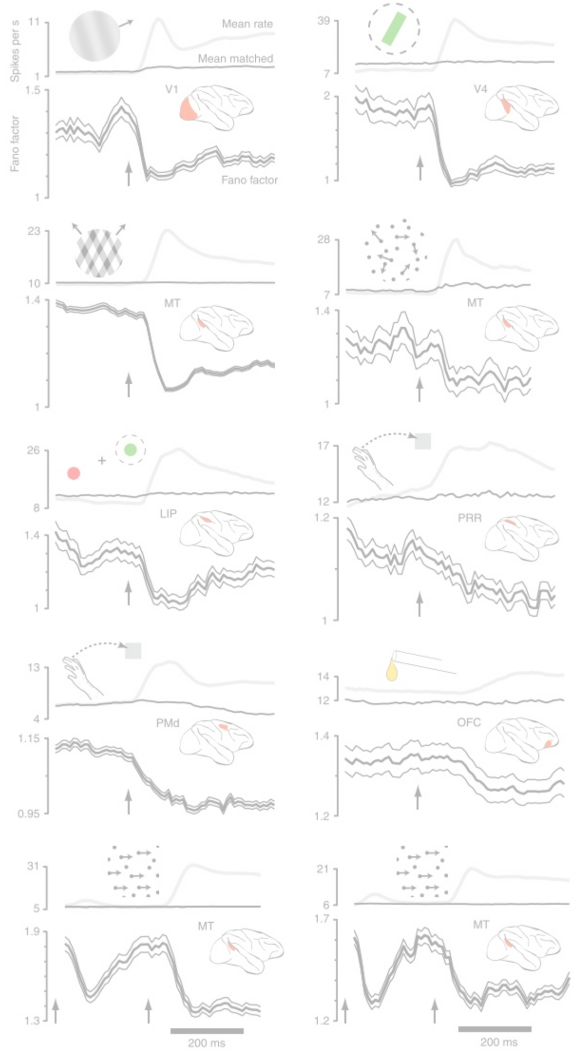
Stimulus onset quenches neural variability…but not indiscriminately



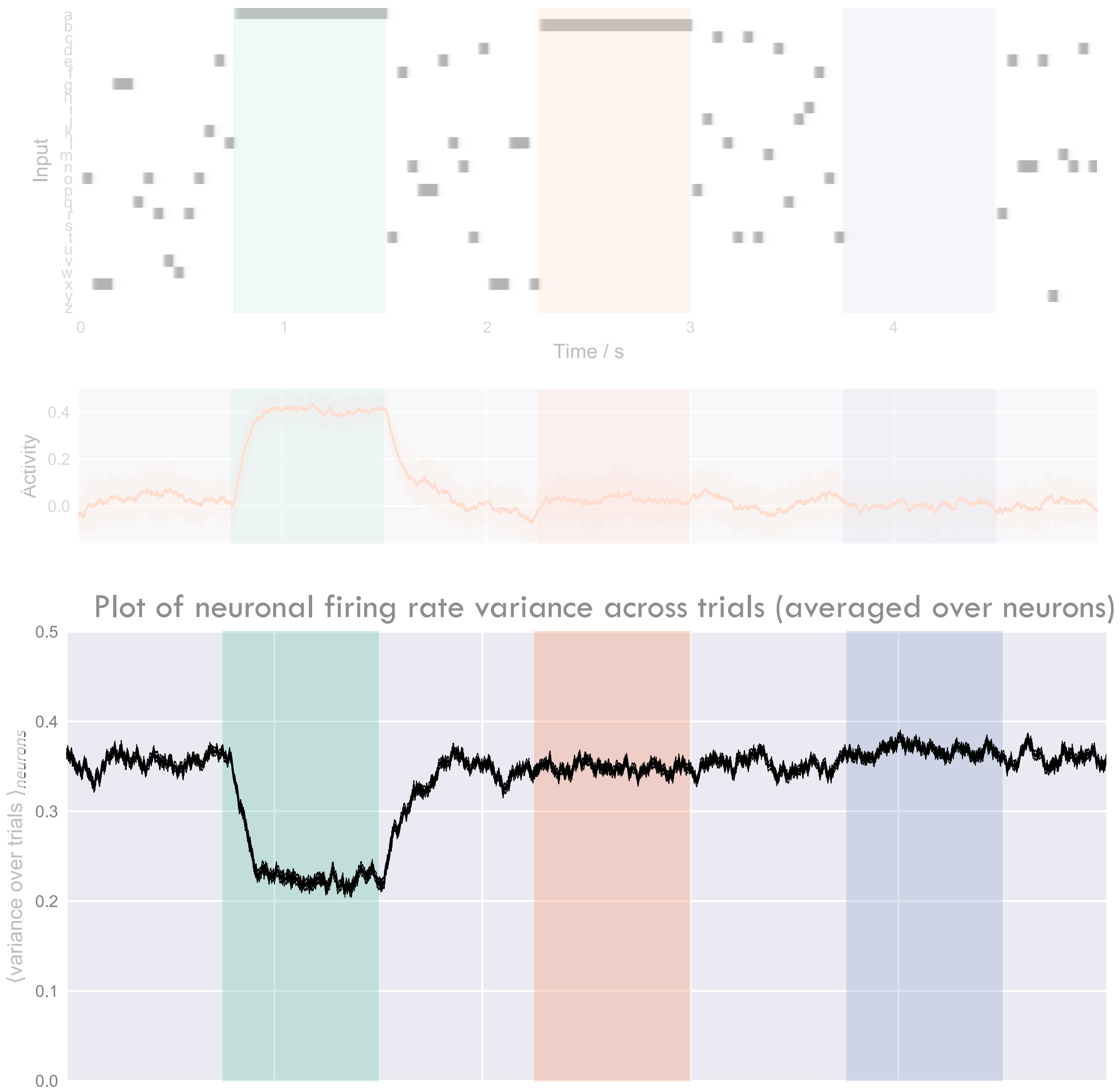


Churchland…Sahani et al. (2010)

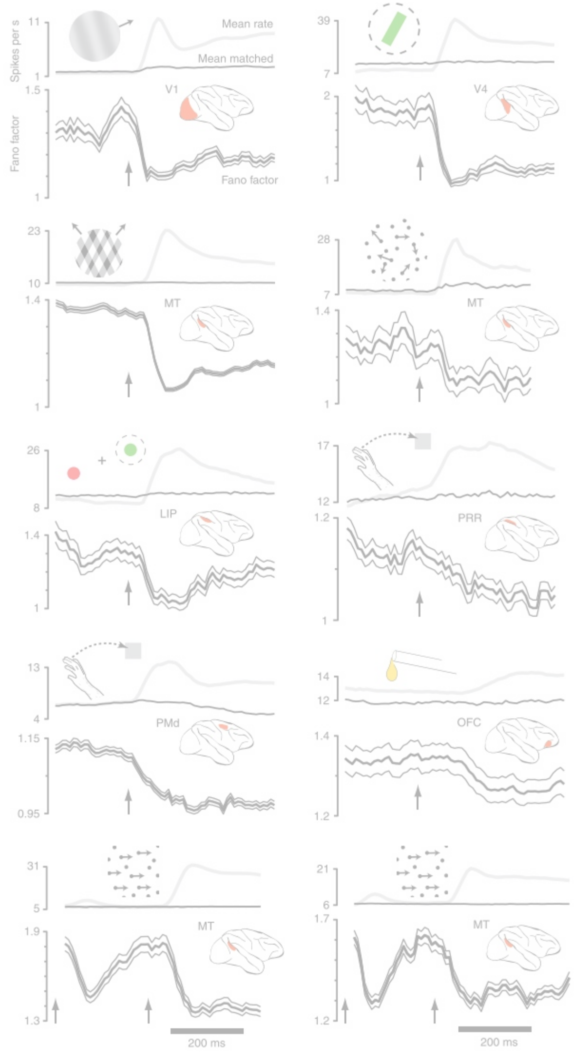
Stimulus onset quenches neural variability…but not indiscriminately



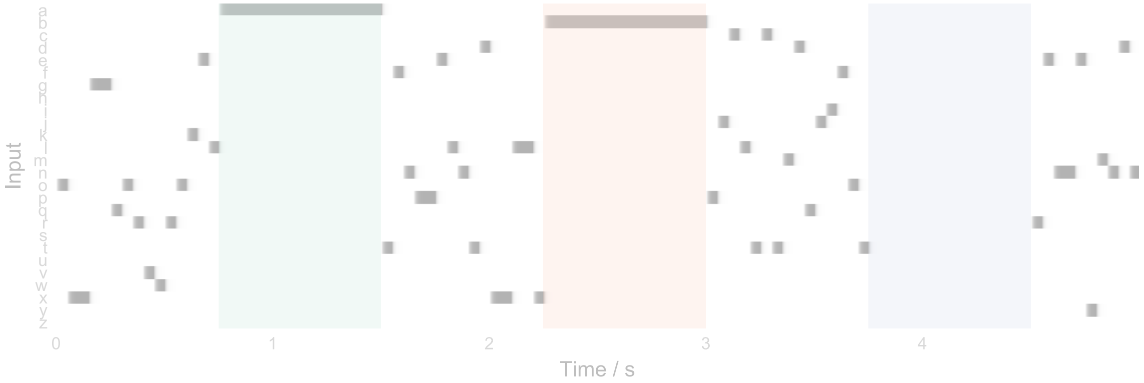
Churchland…Sahani et al. (2010)

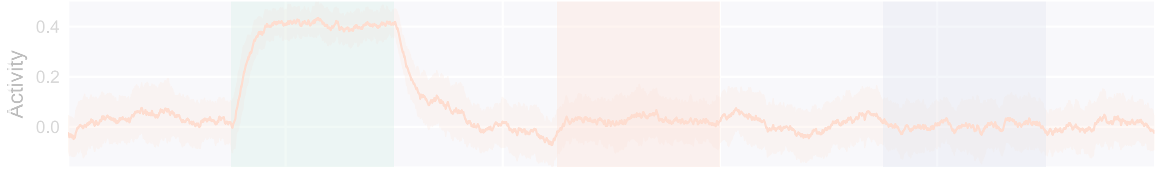


Stimulus onset quenches neural variability…but not indiscriminately

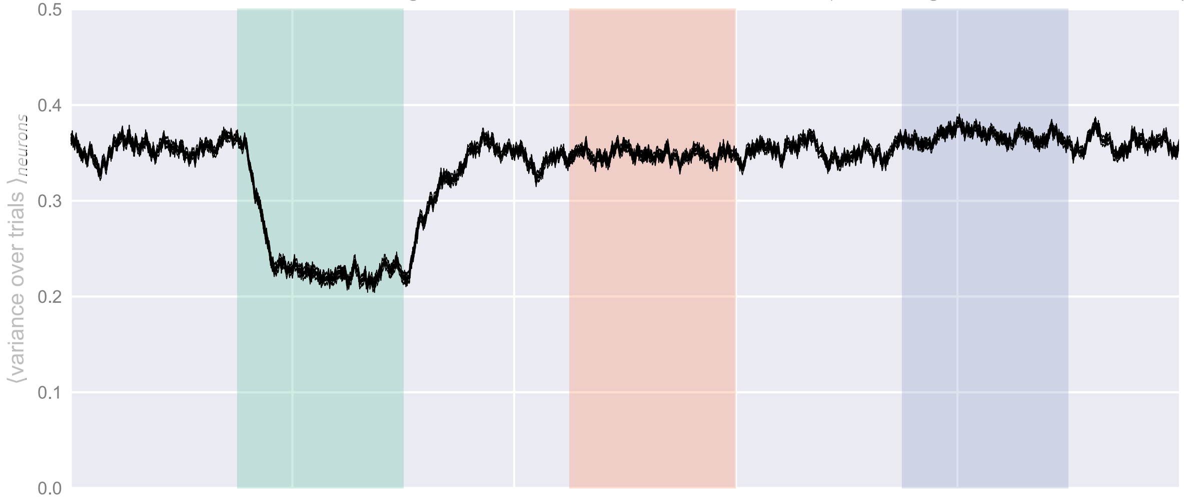


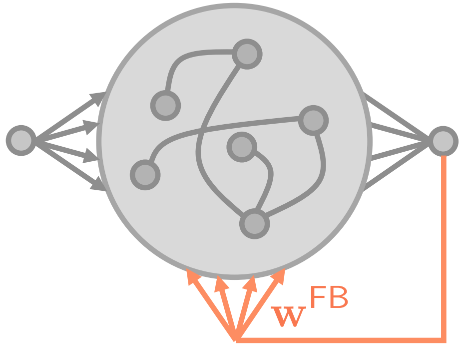
Churchland…Sahani et al. (2010)



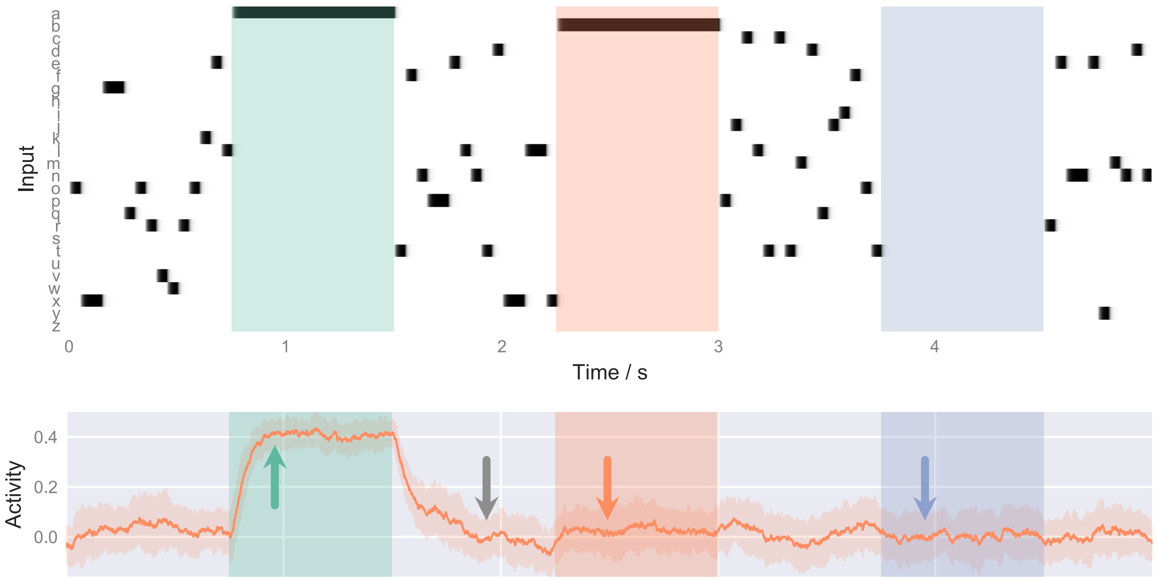


Plot of neuronal firing rate variance across trials (averaged over neurons)



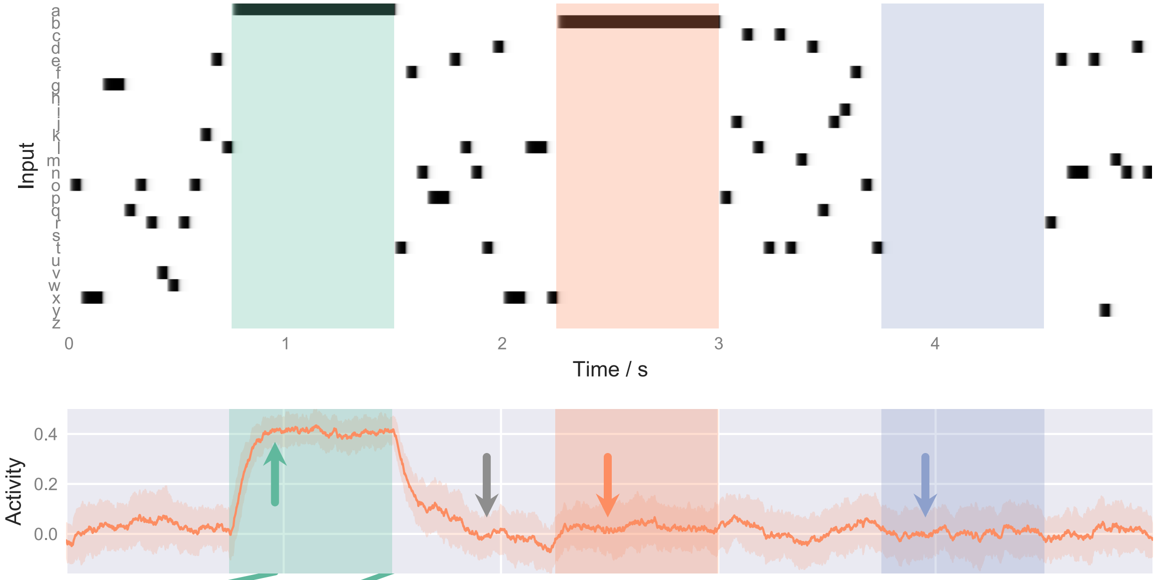


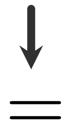
‘Chunking’ suppresses chaos in the internal dynamics



1. A copy network is made
2. A small perturbation applied to all neurons
3. Both networks left to evolve and magnitude of perturbation is tracked
4. This processed is repeated a few times

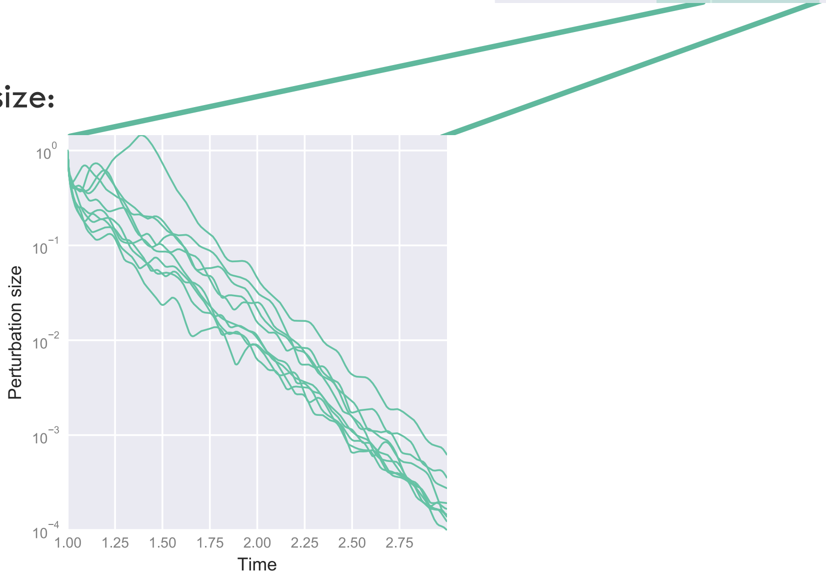
‘Chunking’ suppresses chaos in the internal dynamics



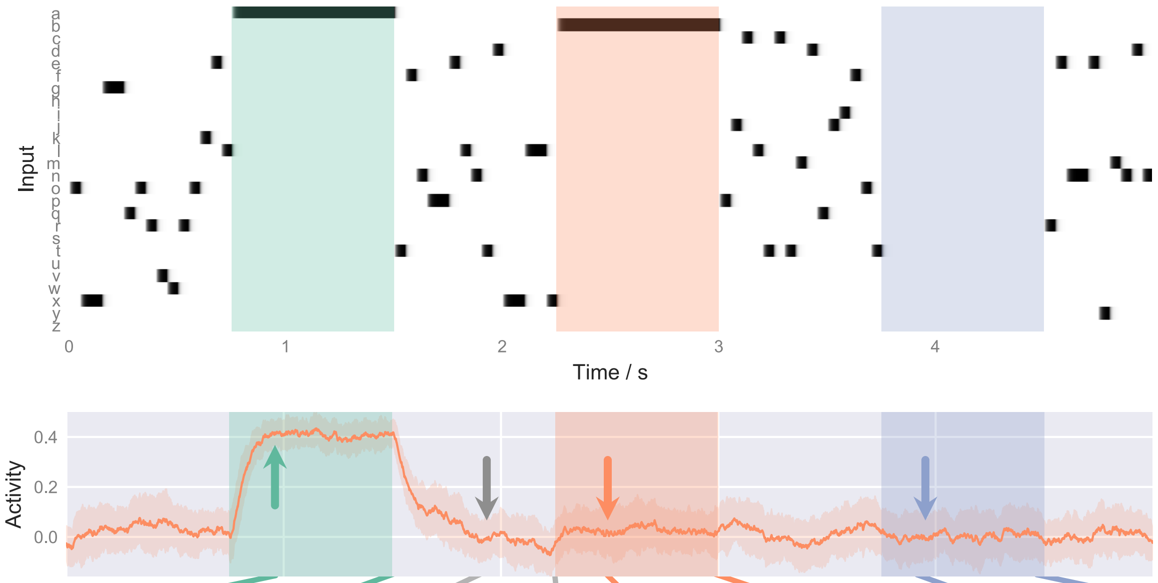


1. A copy network is made
2. A small perturbation applied to all neurons
3. Both networks left to evolve and magnitude of perturbation is tracked
4. This processed is repeated a few times

Evolution of perturbation size:

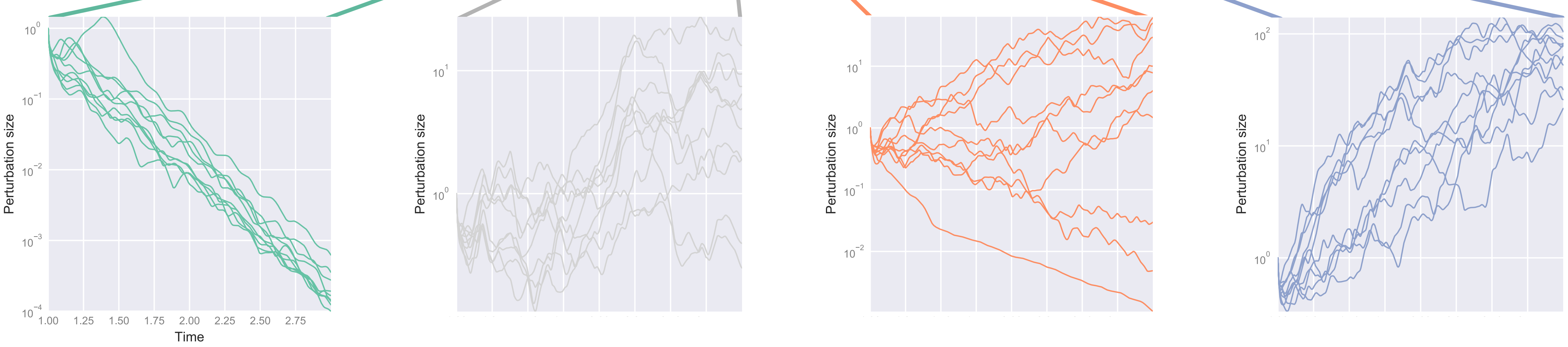


‘Chunking’ suppresses chaos in the internal dynamics

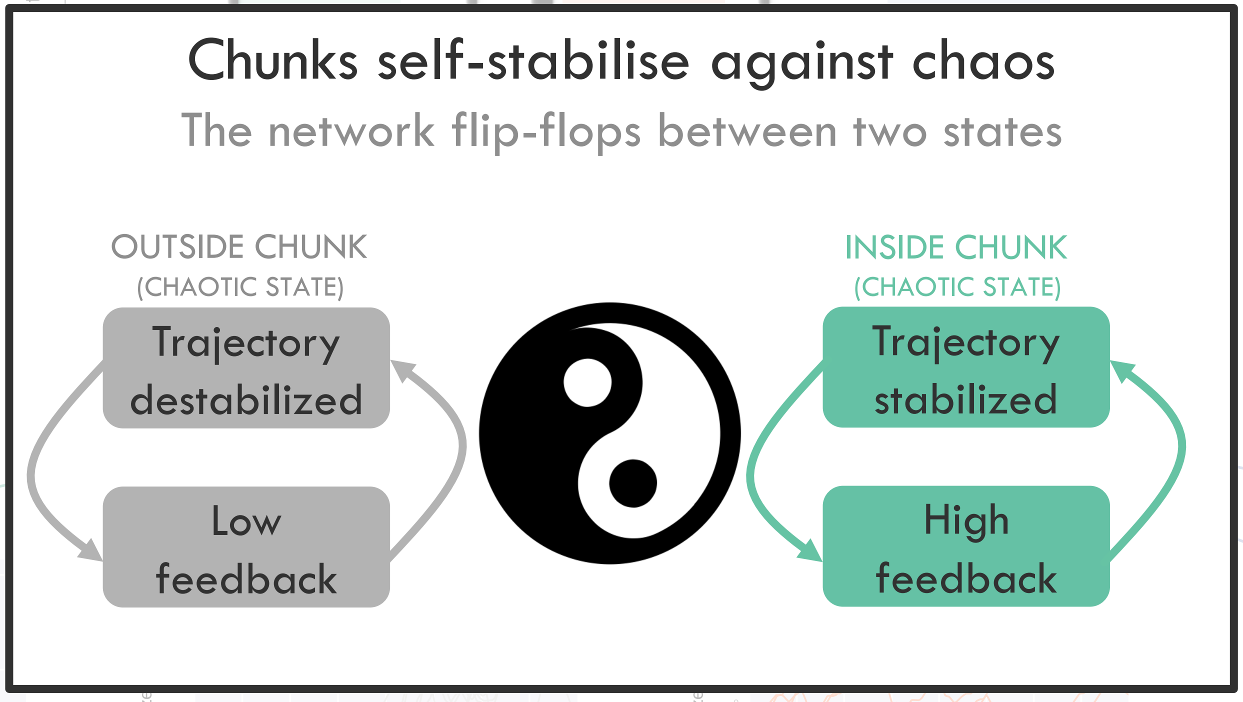


1. A copy network is made
2. A small perturbation applied to all neurons
3. Both networks left to evolve and magnitude of perturbation is tracked
4. This processed is repeated a few times

Evolution of perturbation size:

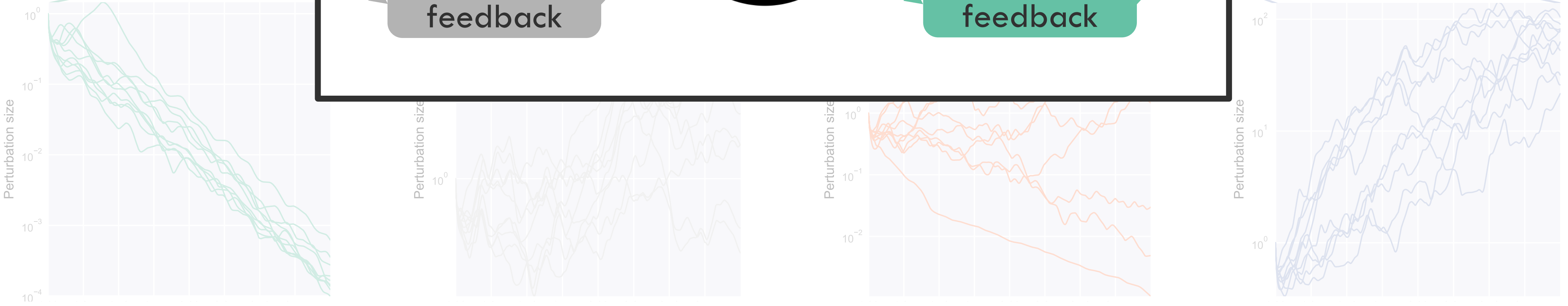


‘Chunking’ suppresses chaos in the internal dynamics



1. A copy network is made
2. A small perturbation applied to all neurons
3. Both networks left to evolve and magnitude of perturbation is tracked
4. This processed is repeated a few times

Evolution of perturbation size:



For an interesting (different) example of dual stable-chaotic networks see *Laje and Buonomano (2013)*

Roadmap

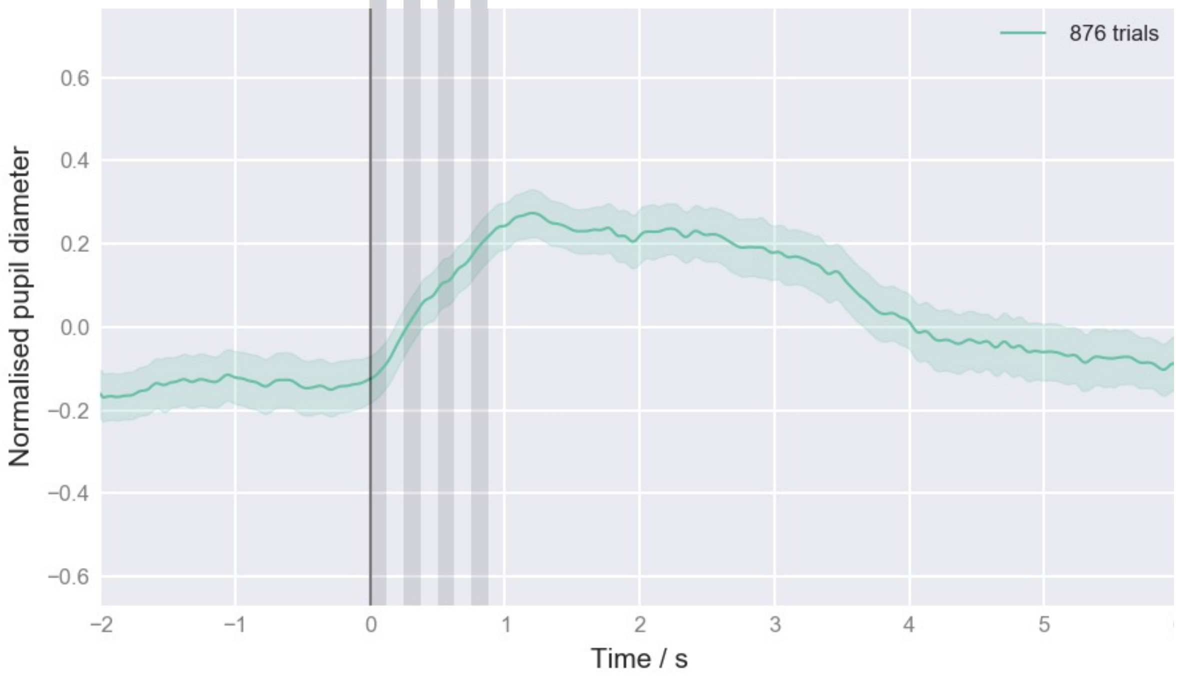
1. A reservoir network model for temporal structure learning
2. The role of chaos
3. Experimental results and modelling predictions
4. Conclusions

We trained people on a distractor task whilst playing them (secretly structured) tone sequences in the background

Dammy et al., *Nature*, 202X

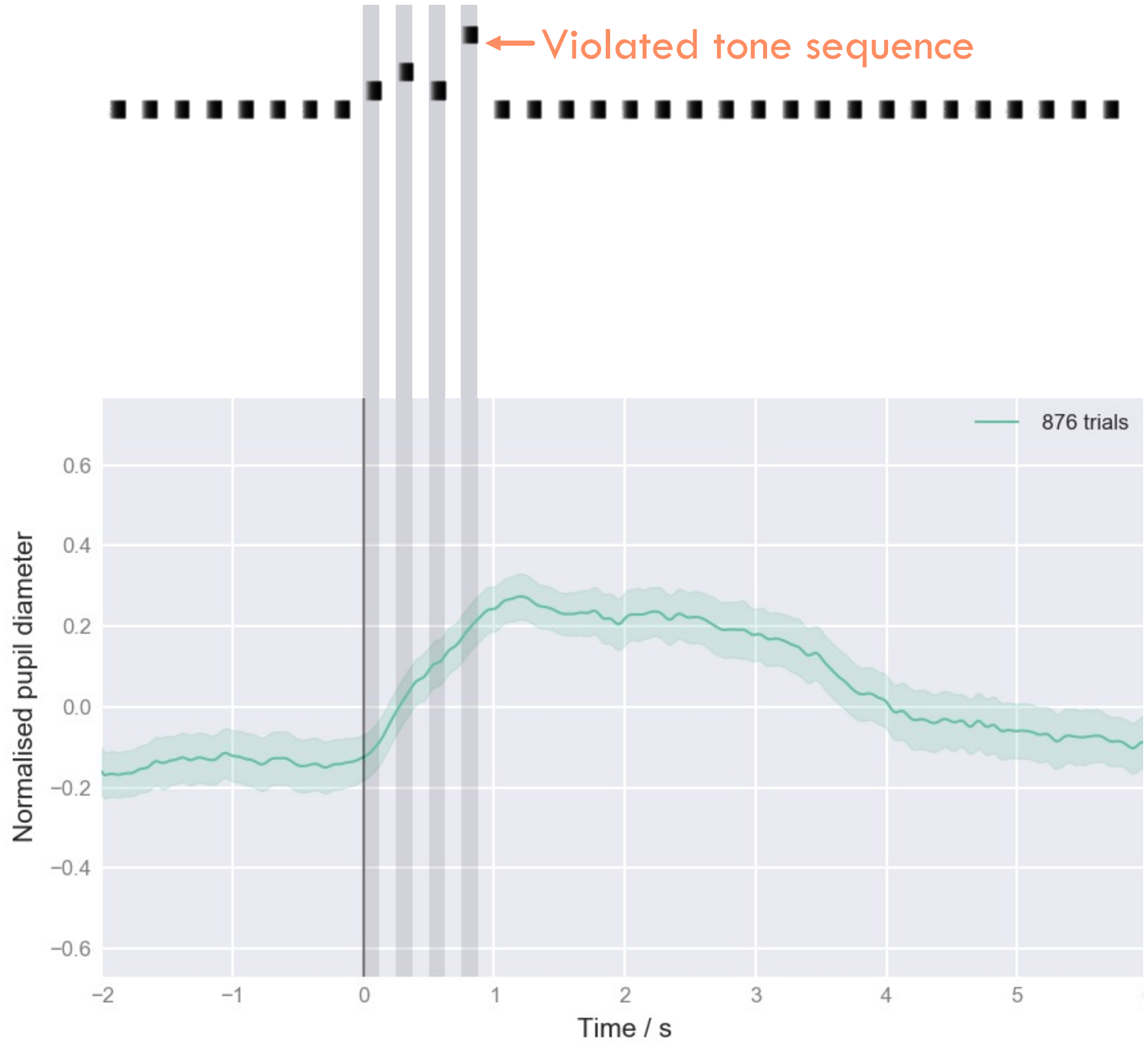
We trained people on a distractor task whilst playing them (secretly structured) tone sequences in the background

Pupil diameter shows chunking-like behaviour



Dammy et al., Nature, 202X

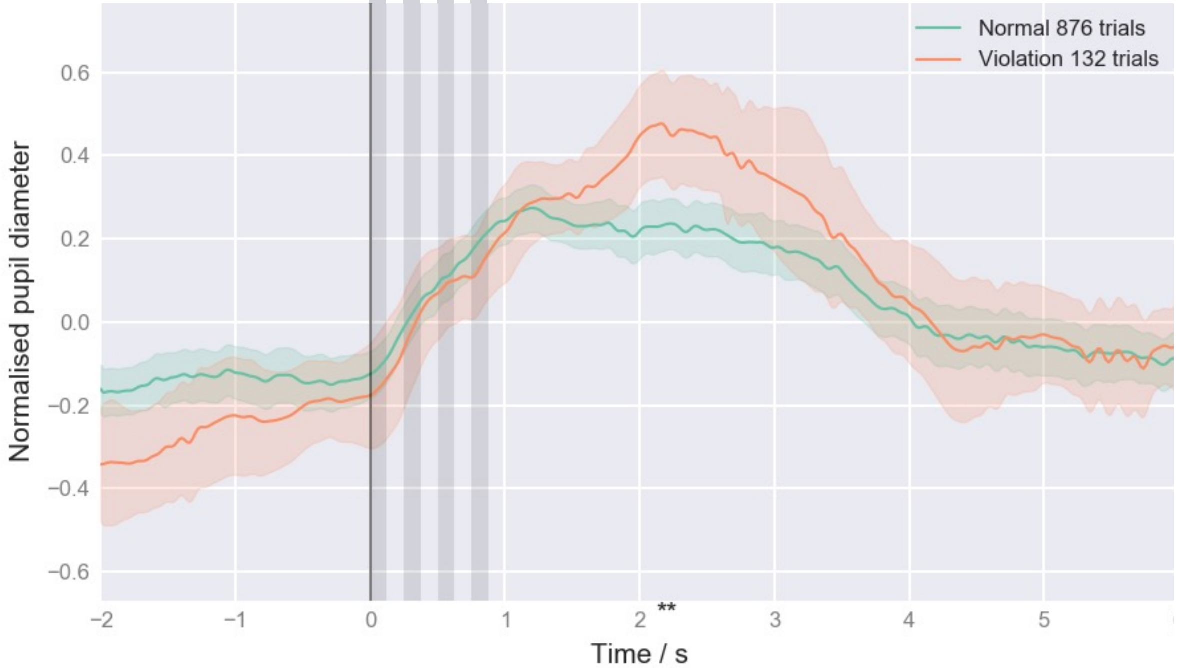
Occasionally we violated the sequence…



Dammy et al., Nature, 202X

Occasionally we violated the sequence…

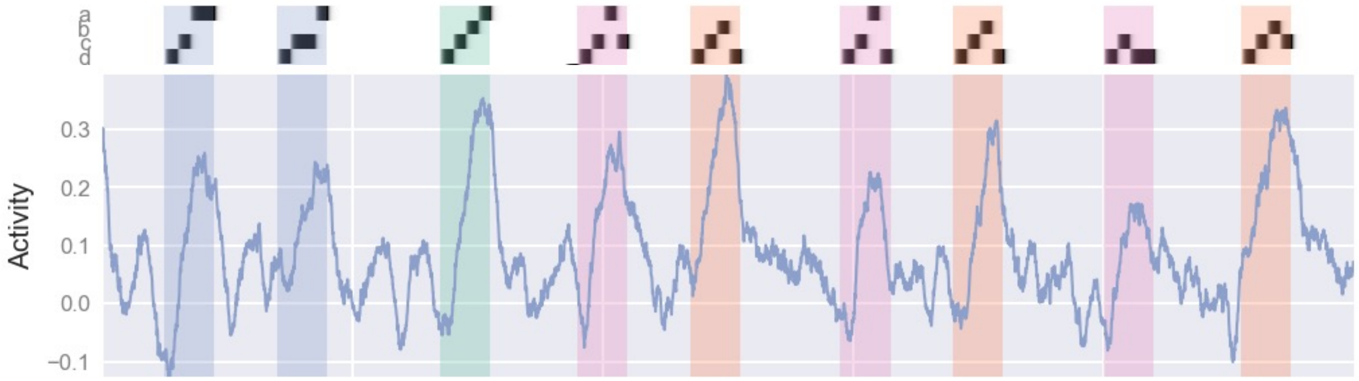
…revealing they “learned” the structure (even though they weren’t instructed to)



Dammy et al., Nature, 202X

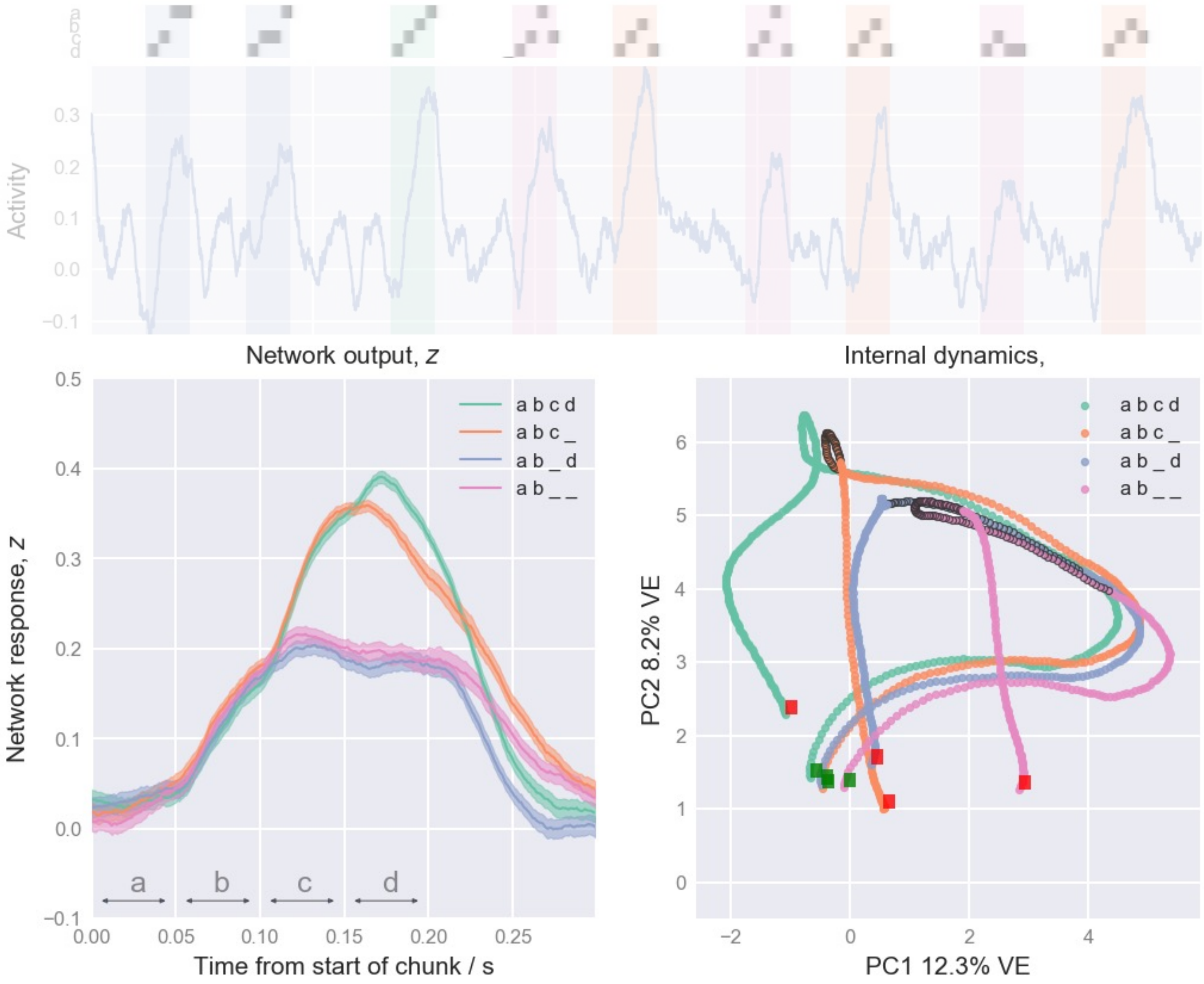
We can simulate a similar experiment on the reservoir model

Here we assume the network output is a proxy for pupil diameter



We can simulate a similar experiment on the reservoir model

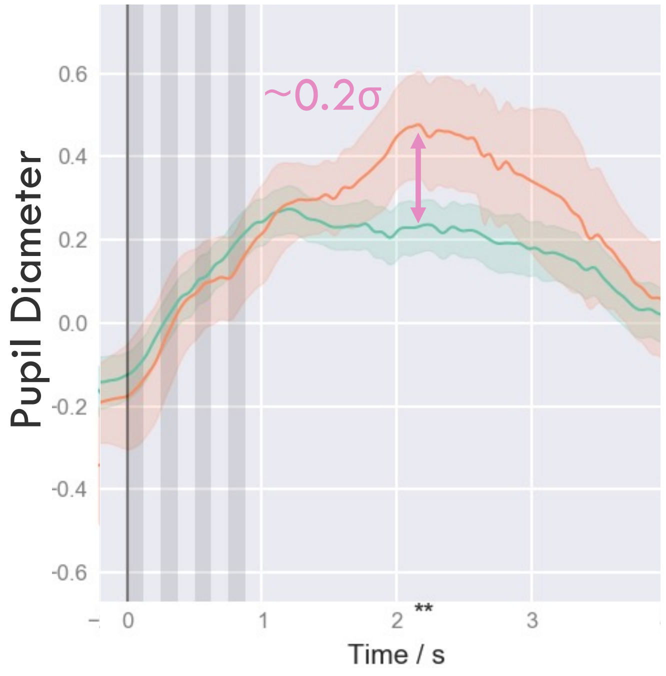
Here we assume the network output is a proxy for pupil diameter



Hallmarks of recurrent processing imparted on pupil data

From quite (left) to very (right) dubious

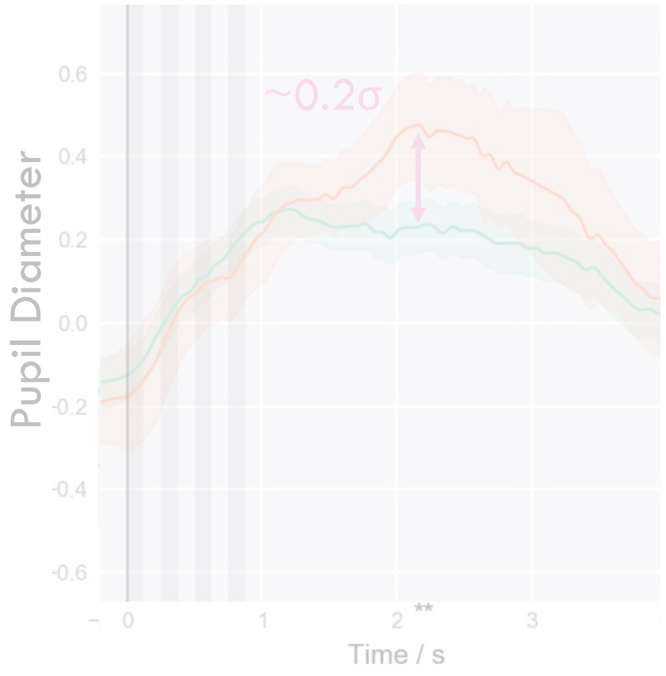
- The effect is tiny.
- Explained by the chunk self-stabilizing effect?



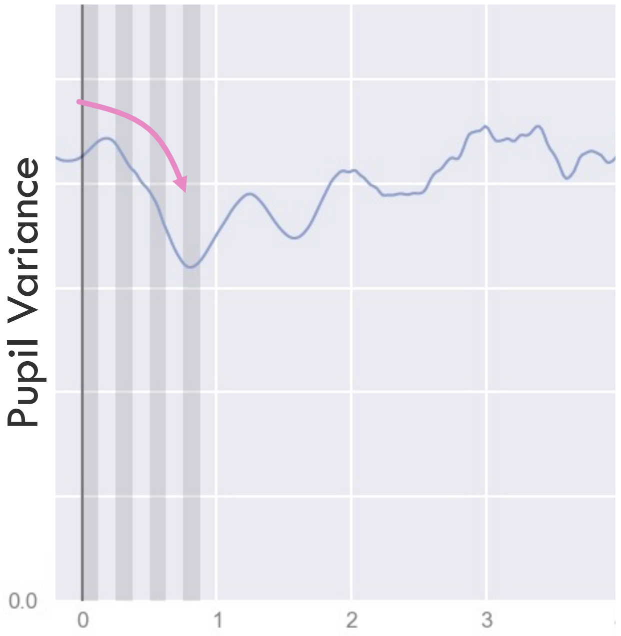
Hallmarks of recurrent processing imparted on pupil data

From quite (left) to very (right) dubious

- The effect is tiny.
- Explained by the chunk self-stabilizing effect?



Pupil variance decreases sharply after stimulus onset



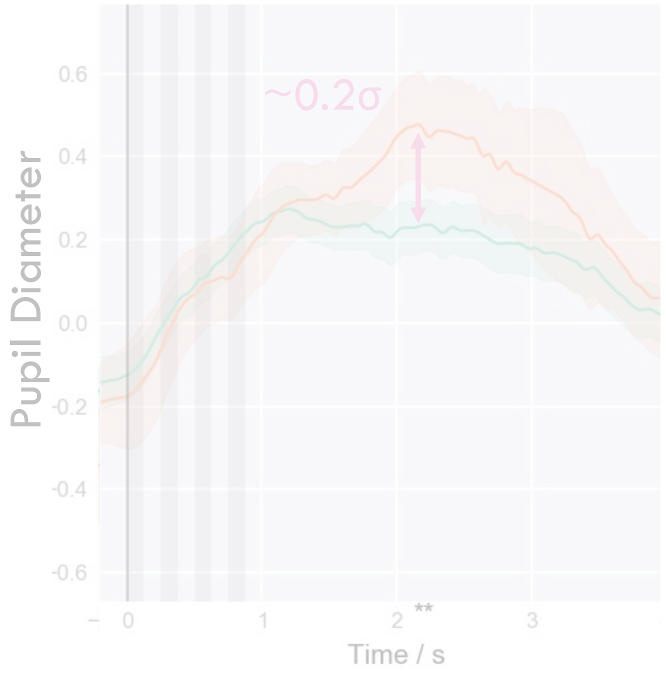
Sainsbury Wellcome Centre

UCL

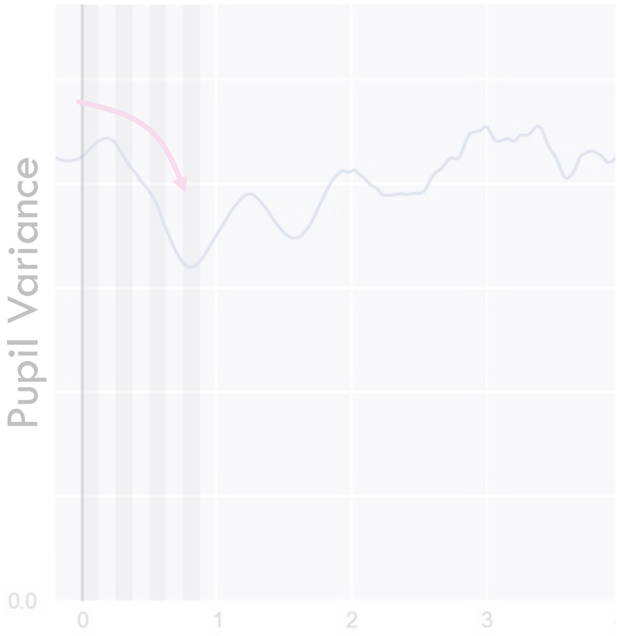
Hallmarks of recurrent processing imparted on pupil data

From quite (left) to very (right) dubious

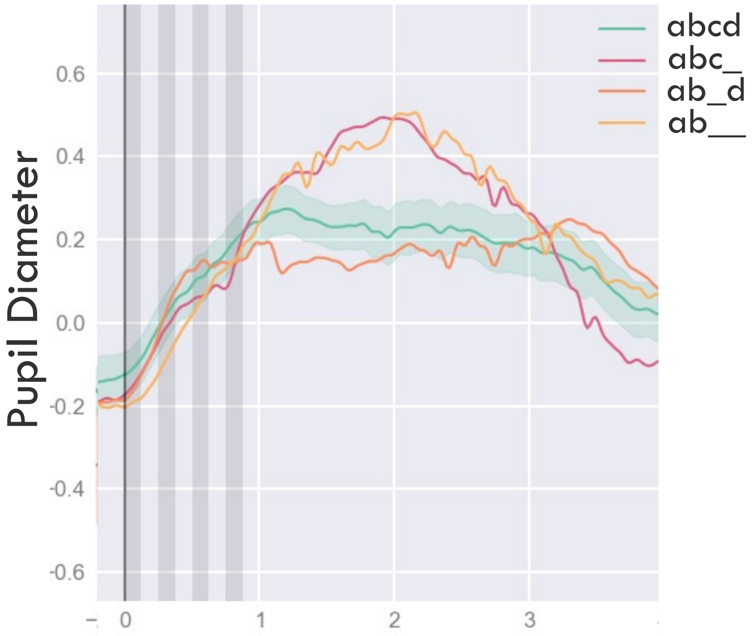
- The effect is tiny.
- Explained by the chunk self-stabilizing effect?



Pupil variance decreases sharply after stimulus onset



'Late' perturbations have longer effect as self-stabilizing effect is turned off

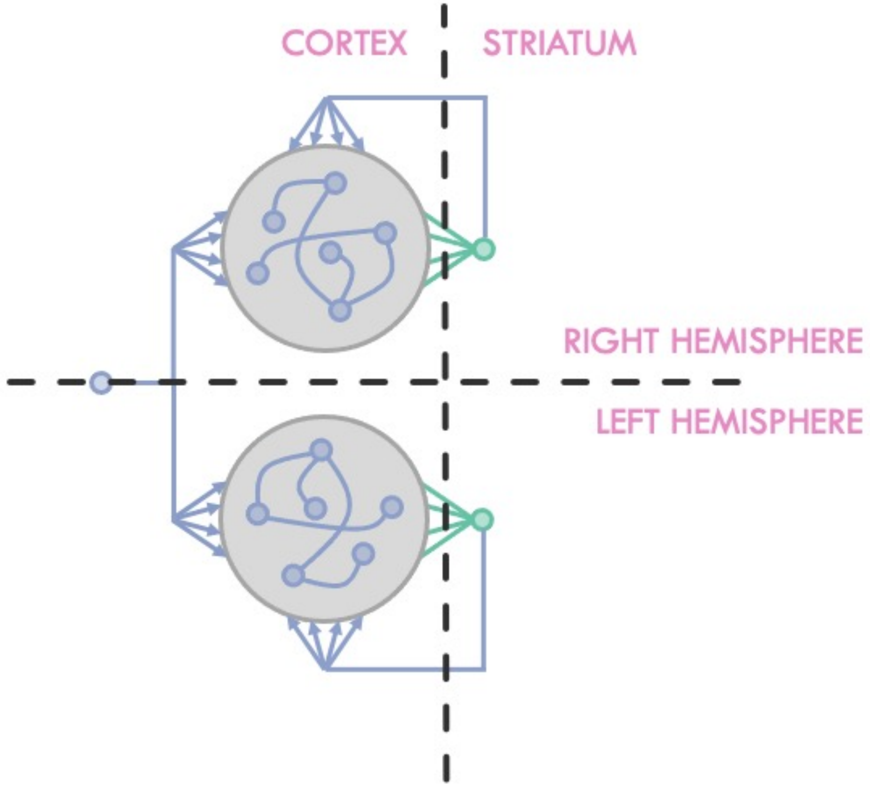


Roadmap

1. A reservoir network model for temporal structure learning
2. The role of chaos
3. Experimental results and modelling predictions
4. Conclusions

Conclusions

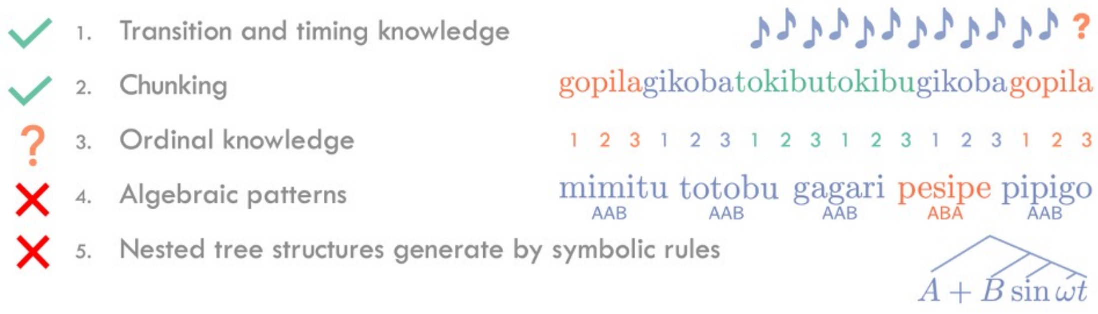
- Reservoir nets have architectural parallels to the brain (particularly cortex)



Conclusions

- Reservoir nets have architectural parallels to the brain (particularly cortex)

- Can explain basic temporal structure processing requiring short memory (~100x neuronal timescale)

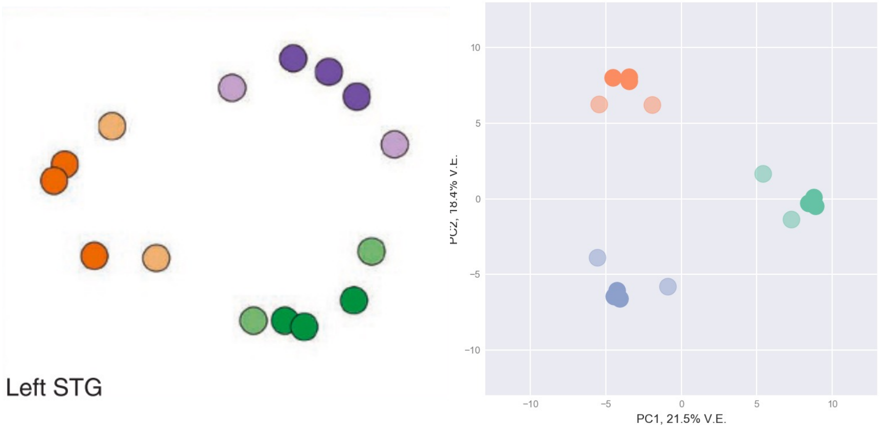


Conclusions

- Reservoir nets have architectural parallels to the brain (particularly cortex)

- Can explain basic temporal structure processing requiring short memory (~100x neuronal timescale)

- Representational similarity to cortex

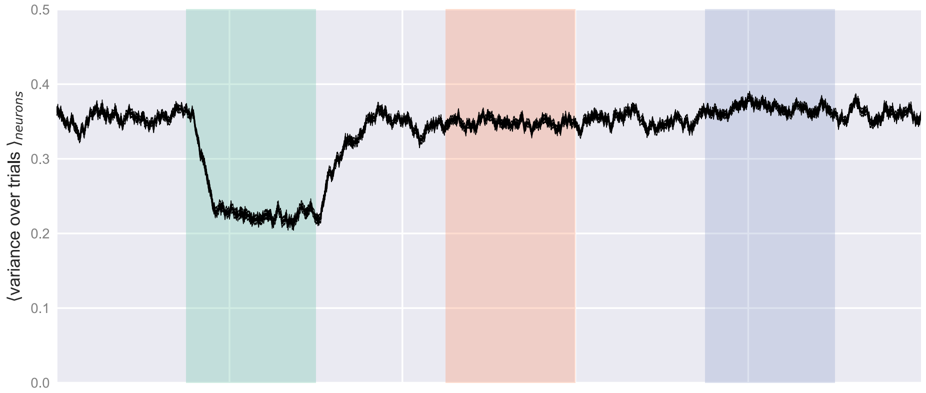


Sainsbury Wellcome Centre

UCL

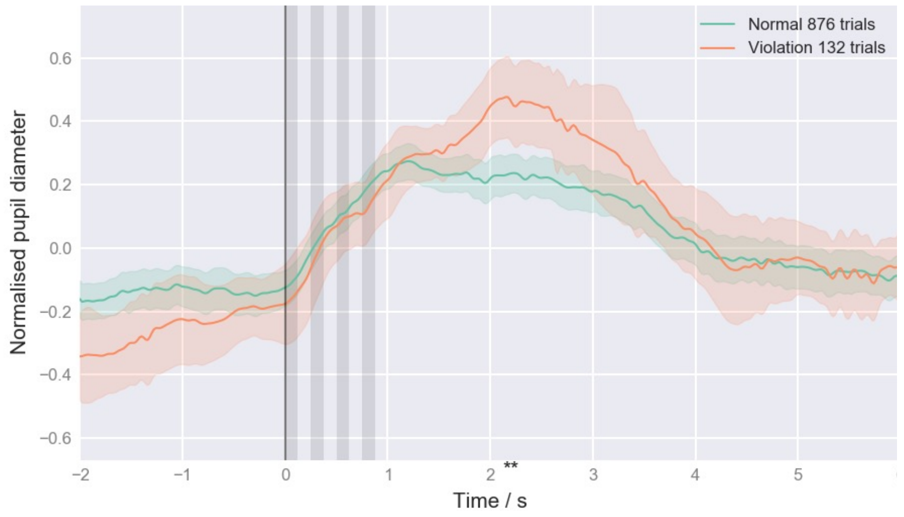
Conclusions

- Reservoir nets have architectural parallels to the brain (particularly cortex)
- Can explain basic temporal structure processing requiring short memory (~100x neuronal timescale)
- Representational similarity to cortex
- The network embraces chaos, dynamically suppressing it with feedback when needed.



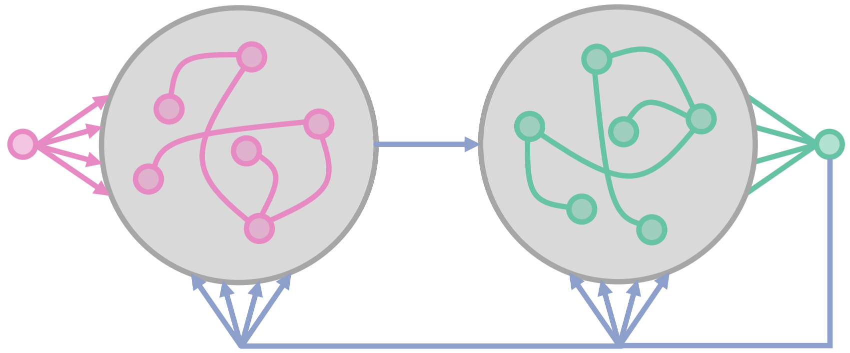
Conclusions

- Reservoir nets have architectural parallels to the brain (particularly cortex)
- Can explain basic temporal structure processing requiring short memory (~100x neuronal timescale)
- Representational similarity to cortex
- The network embraces chaos, dynamically suppressing it with feedback when needed.
- Hallmarks of recurrent processing are compatible with experimental data



Future Directions

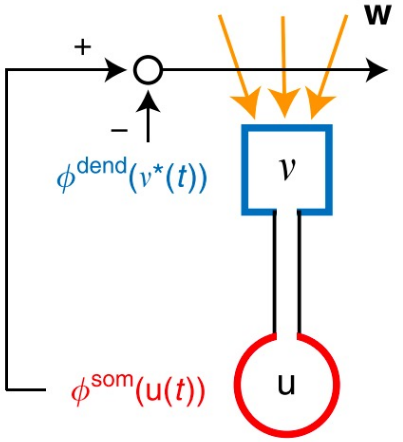
- Other architectures
  - Spiking?



Future Directions

- Other architectures
  - Spiking?

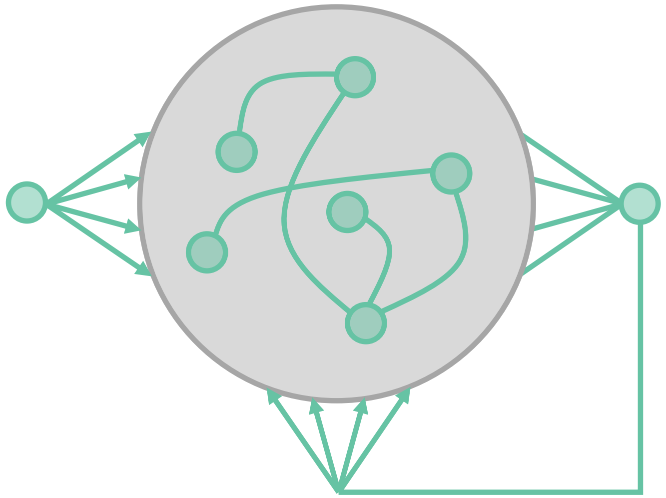
- Other learning rules
  - Minimize information loss e.g. *Asabuki et al. (2019)*



$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \int dt\, D_{KL}[\phi^{som}(u(t)) \,\|\, \phi^{dend}(v^*(t))]$$

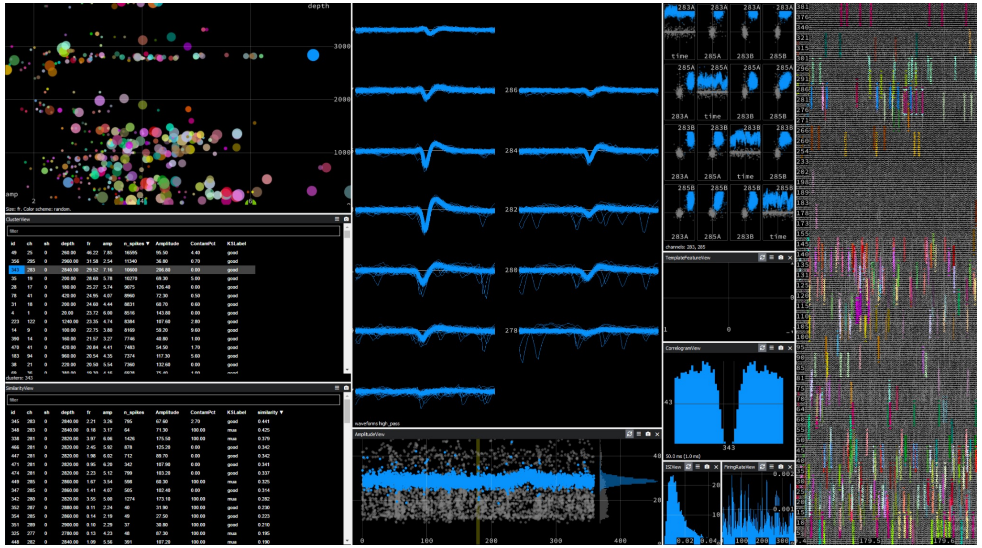
Future Directions

- Other architectures
  - Spiking?
- Other learning rules
  - Minimize information loss *e.g. Asabuki et al. (2019)*
- BPTT



Future Directions

- Other architectures
  - Spiking?
- Other learning rules
  - Minimize information loss *e.g. Asabuki et al. (2019)*
- BPTT
- Compare to neuronal data (@Dammy?)



Beautiful spike data di Elena

Thanks to:

LIM Lab — Learning Inference Memory

- Athena Akrami
- Dammy Onih
- Peter Vincent



- Claudia Clopath

OIST

- Tomoki Fukai
- Toshitake Asabuki


https://github.com/TomGeorge1234/ReservoirComputing
https://github.com/TomGeorge1234/PupillometryPipeline