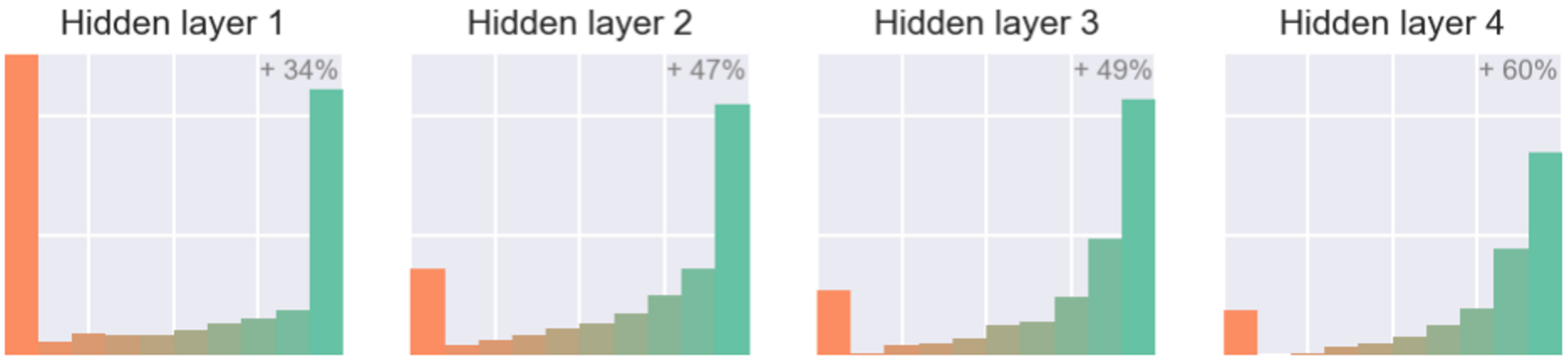
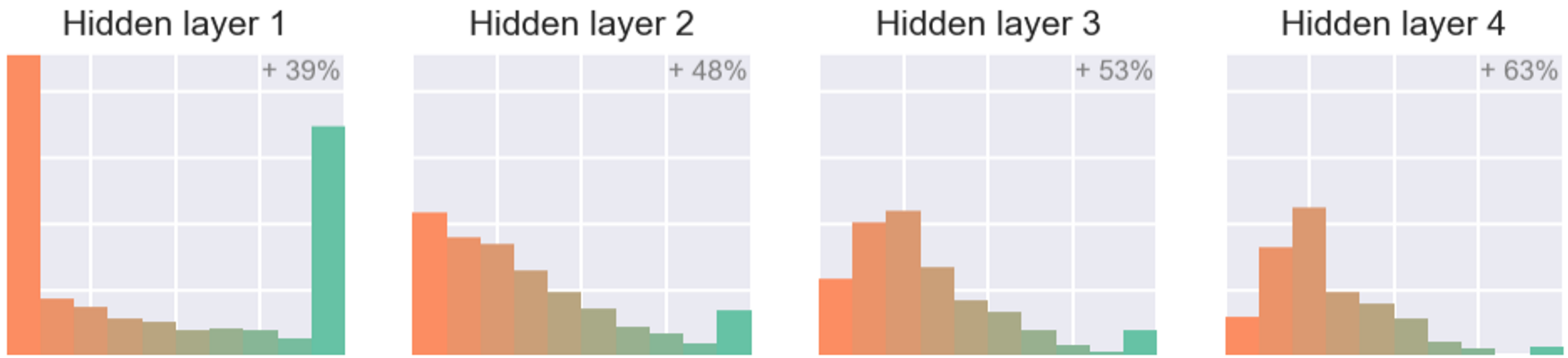


Cognitive flexibility: Untangling the connection between continual learning, neuronal selectivity and flexible cognitive control

aka

What makes neurons picky?

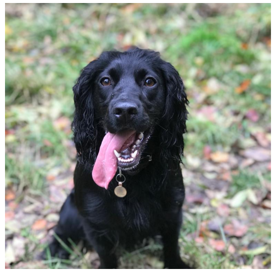
Tom George – Pehlevan Lab Meeting 05/13/2020



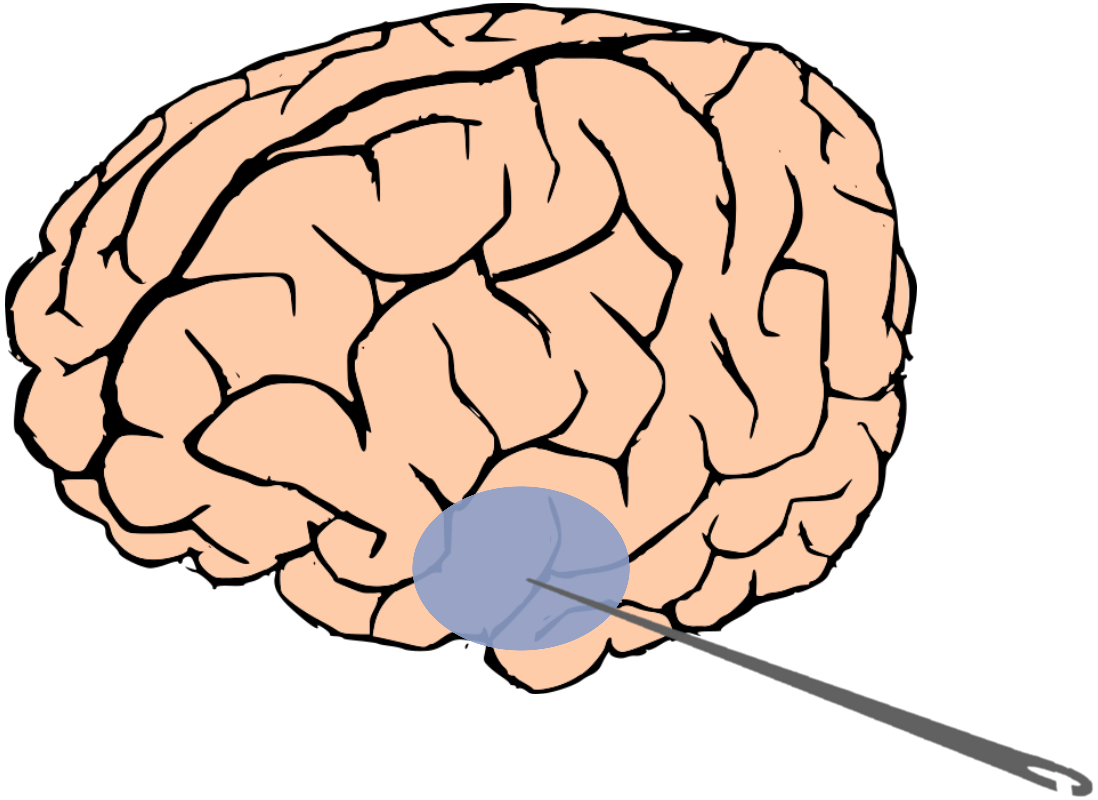
Some neurons are selective, whilst others are not

Task A: Look at a human face

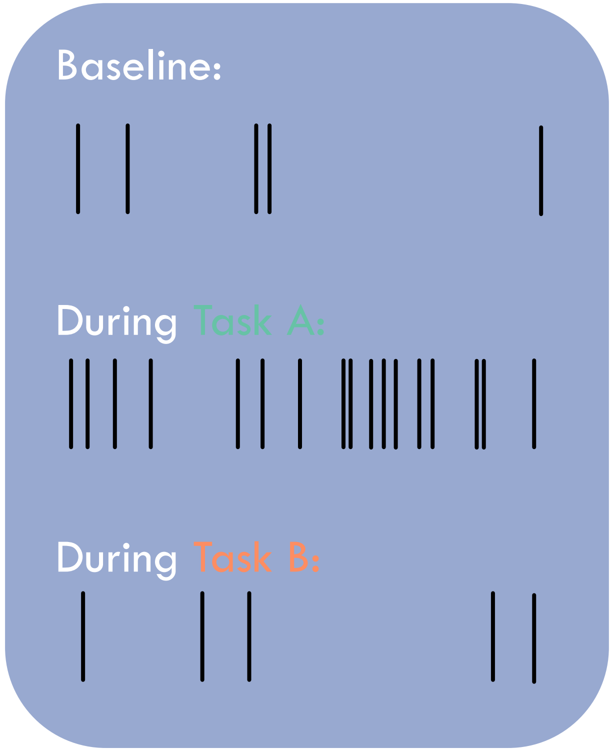
Task B: Look at a dog



VISUAL CORTEX



SELECTIVE to task A

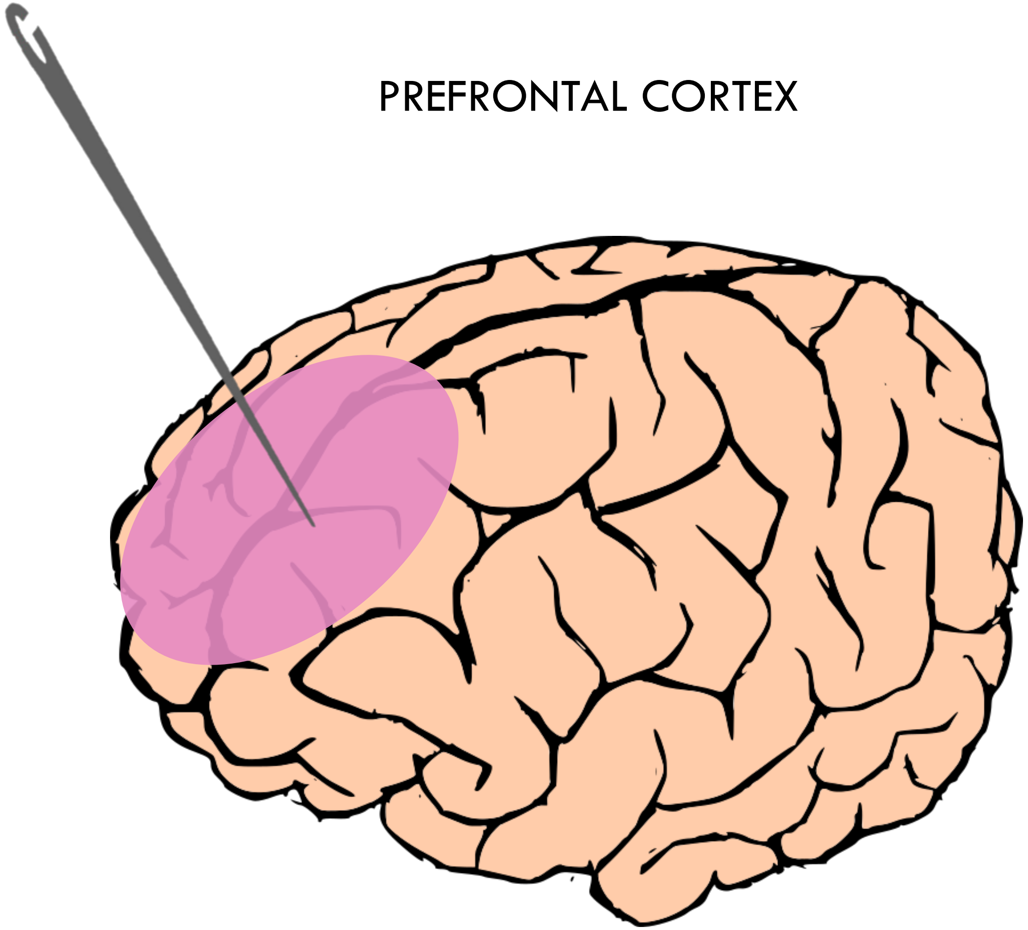


Some neurons are selective, **whilst others are not**

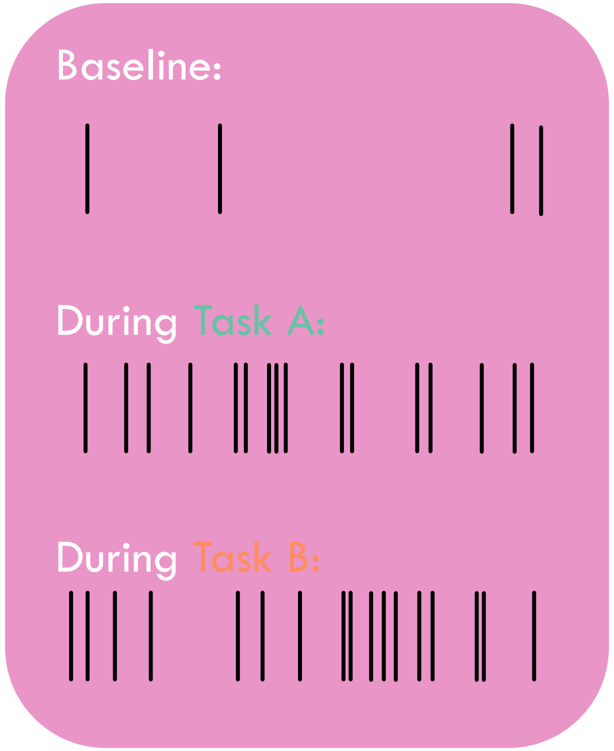
Task A:
Remembering a telephone number just read to you

Task B:
Deciding when to cross a busy road

PREFRONTAL CORTEX



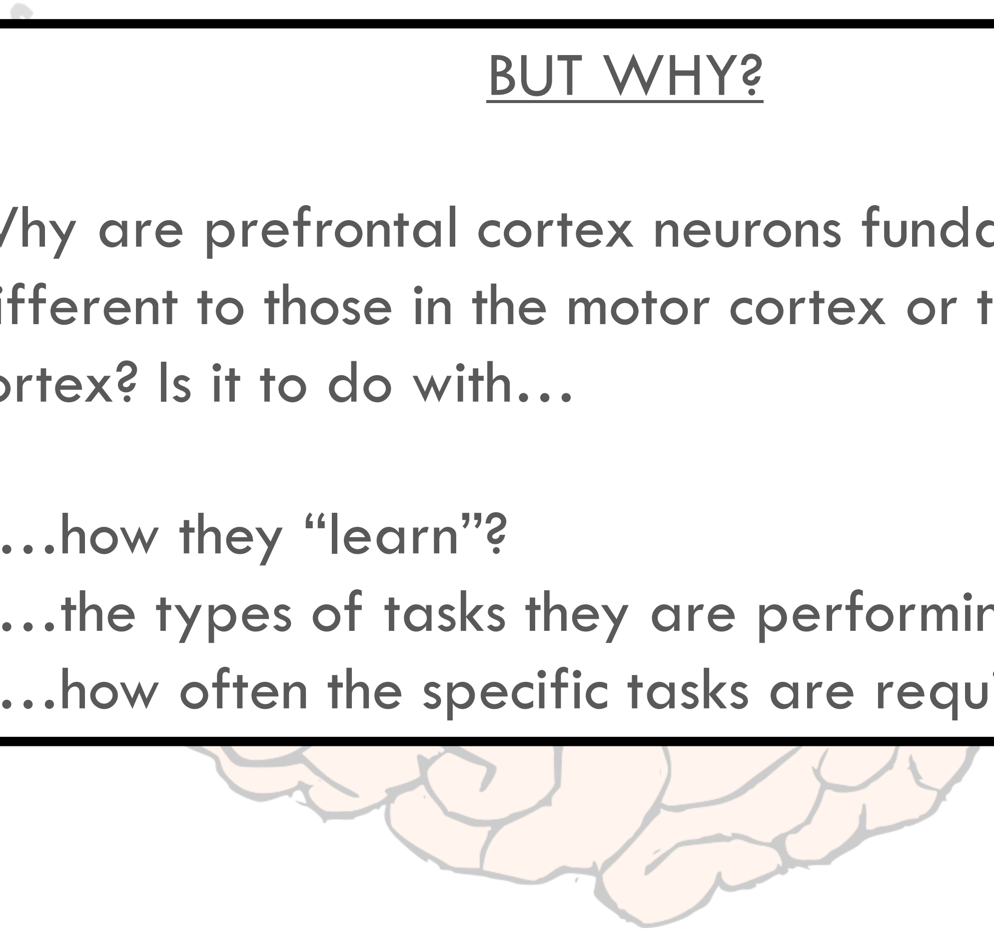
MIXED SELECTIVITY



Some neurons are selective, **whilst others are not**

Task A:
Remembering a telephone number just read to you

Task B:
Deciding when to cross a busy road

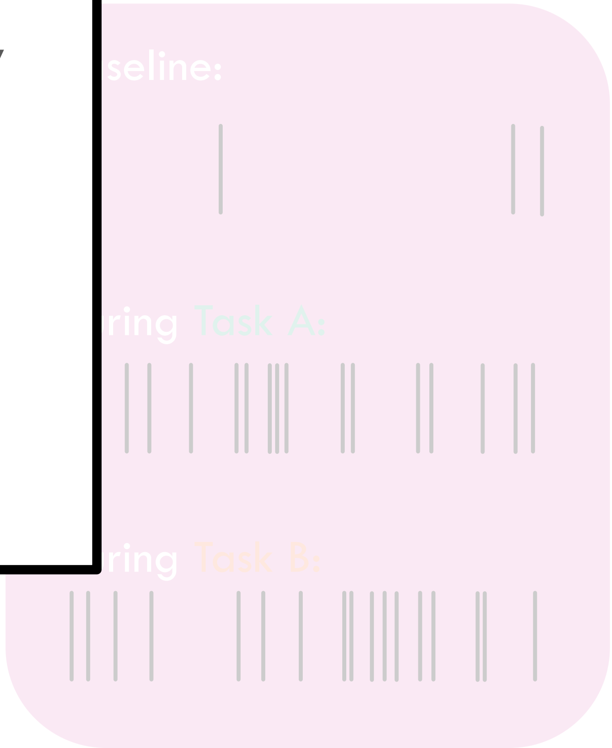


BUT WHY?

Why are prefrontal cortex neurons fundamentally different to those in the motor cortex or the visual cortex? Is it to do with…

- …how they "learn"?
- …the types of tasks they are performing?
- …how often the specific tasks are required?

MIXED SELECTIVITY



Roadmap

1. Ideas from the literature
2. A simple model trained on simple tasks
3. A more complex model trained on MNIST tasks
4. Conclusions

Roadmap

1. Ideas from the literature
2. A simple model trained on simple tasks
3. A more complex model trained on MNIST tasks
4. Conclusions

Mixed selectivity can be computationally advantageous

Rigotti et al. (2013) make a convincing argument for mixed selectivity in the PFC:

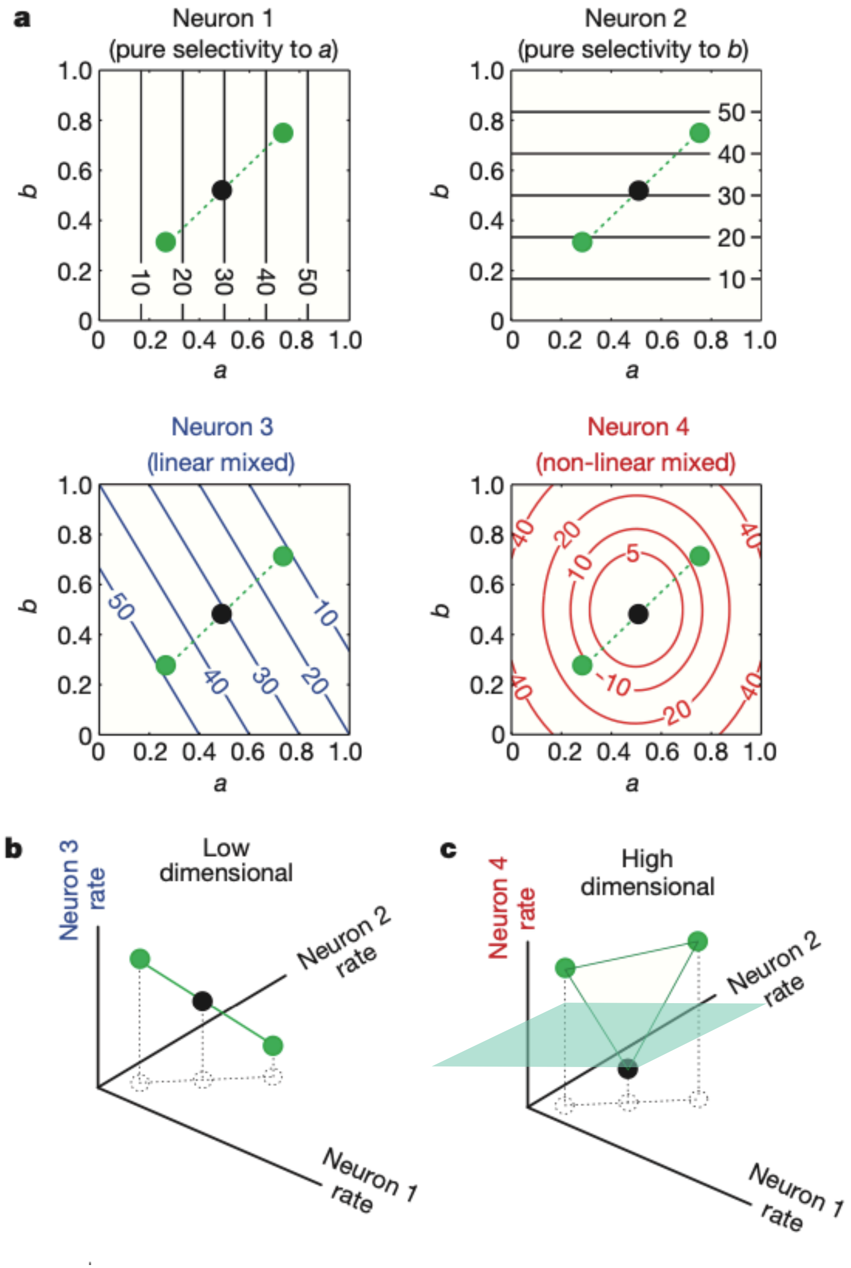
Mixed Selectivity → High dimensional representations → Large capacity for input-output functions → Good for the PFC

But **why** does the PFC need a large input-output function capacity?

Miller and Cohen (2001): "the PFC modifies responses to sensory data given changing contexts or goals".

The cognitive tasks it must perform span an infinite range:
Complex tasks can be composed **recursively** from simpler tasks.

Compare to vision: visual scenes (although rich and varied) are generally built from a basic set of polygons, colours and textures.



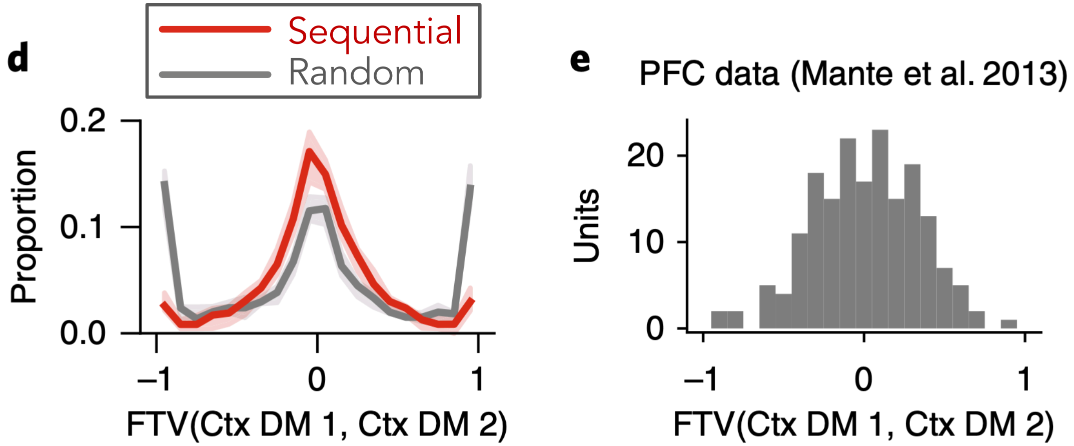
…or mixed selectivity could be a feature of *how* the PFC learns

Yang et al.(2019): trained a complex RNN model of the PFC on 20 'complex' cognitive tasks and found:

- **highly mixed selective** representations when the network was trained **sequentially (Task 1…then task 2…then task 3…)**. Matching what is found in the PFC, *Mante et al. (2013)*.
- **selective representations** when trained **randomly** on all tasks

Perhaps, then, mixed selectivity in the PFC is due to how we learn cognitive skills as children - in a blocked, sequential fashion.

Compare to vision: from the day we are born we are presented with many visual scenes in a random order and begin to 'learn' them, mostly unsupervised



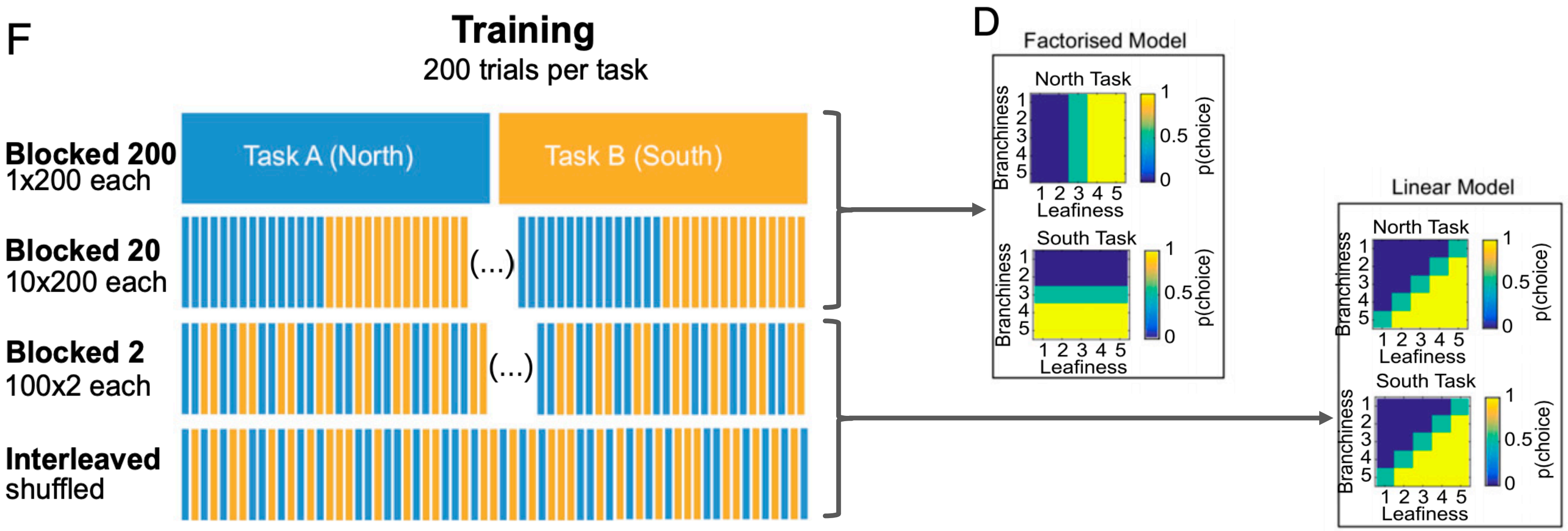
However, it's still not clear

Flesch & Summerfield et al. (2018) studied multiple task learning in humans and computers.

Analysis of human results suggest that blocked (aka sequential) training results in more factorized (potentially interpretable as selective) task representations.

This is opposite to the Yang et al. result.

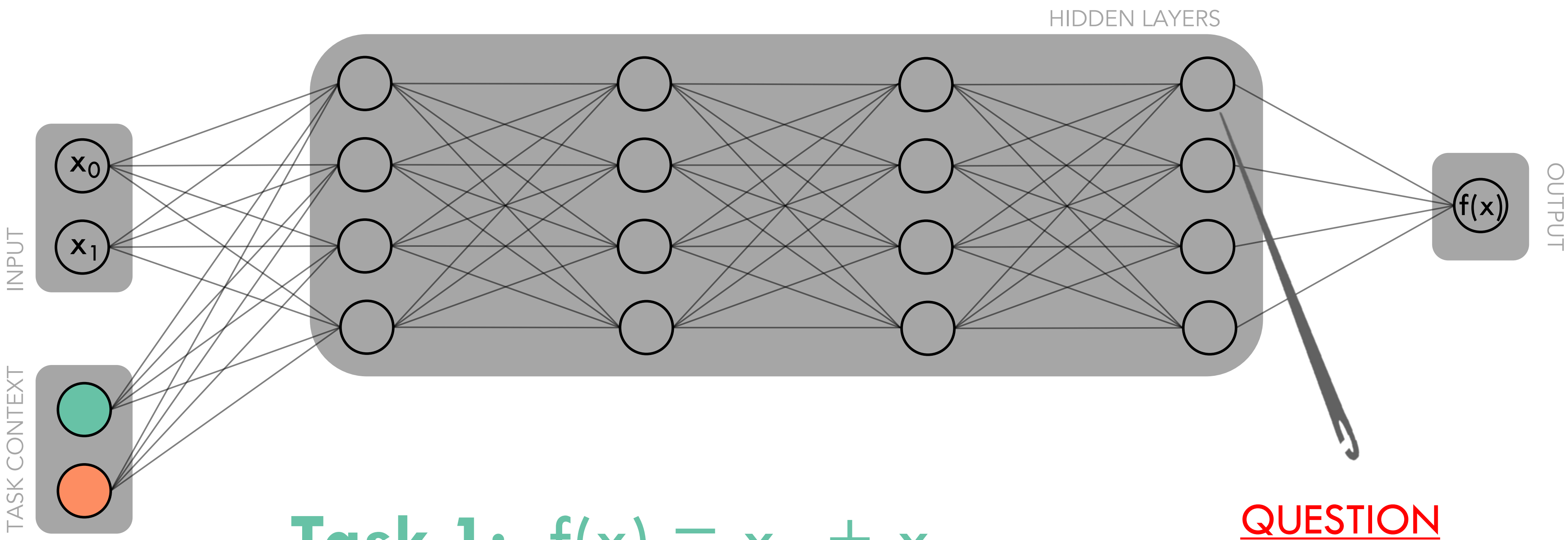
Either way: training style seems to be important.



Roadmap

1. Ideas from the literature
2. A simple model trained on simple tasks
3. A more complex model trained on MNIST tasks
4. Conclusions

Start simple: Train a basic deep network to learn two different tasks



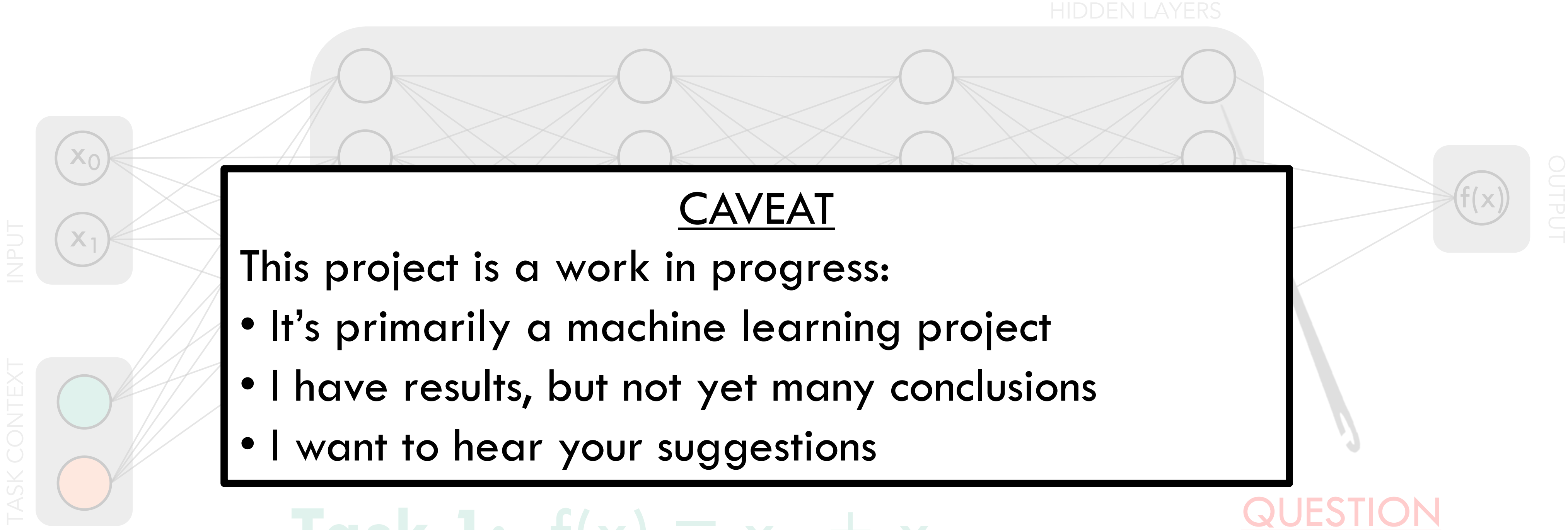
Task 1: $f(x) = x_0 + x_1$

Task 2: $f(x) = x_0 x_1$

QUESTION

- Will this neuron be equally important in both tasks? Or selective
- What happens when we vary architecture, tasks, training style etc…?

Start simple: Train a basic deep network to learn two different tasks



<u>CAVEAT</u>

This project is a work in progress:

- It's primarily a machine learning project
- I have results, but not yet many conclusions
- I want to hear your suggestions

Task 1: $f(x) = x_0 + x_1$

Task 2: $f(x) = x_0 x_1$

<u>QUESTION</u>

- Will this neuron be equally important in both tasks? Or selective
- What happens when we vary architecture, tasks, training style etc…?

Change in loss function tells us how ‘important’ a neuron is

Measure “importance” by how much the expected loss over some test set changes when neuron is lesioned:

$$\mathcal{I}_i(A) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_\mathcal{A}}\left[\left(\ell(z;\mathbf{h}) - \ell(z;\mathbf{h}|h_i=0)\right)^2\right]$$

Where **h** is a vector containing the state of the hidden neurons.
Taylor expanding gives:

$$\ell(z;\mathbf{h}|h_i=0) = \ell(z;\mathbf{h}) + (\mathbf{h}_{\setminus i} - \mathbf{h})^\mathsf{T}\frac{\partial \ell}{\partial \mathbf{h}} + \frac{1}{2}(\mathbf{h}_{\setminus i} - \mathbf{h})^\mathsf{T}\mathsf{H}(\mathbf{h}_{\setminus i} - \mathbf{h}) + \ldots$$

So, to first order,

$$\boxed{\mathcal{I}_i(A) \approx \mathop{\mathbb{E}}_{z\sim\mathcal{D}_\mathcal{A}}\left[\left(h_i \cdot \frac{\partial \ell}{\partial h_i}\right)^2\right]}$$

The importance of hidden neuron *i* for task *A*

Change in loss function tells us how 'important' a neuron is

Measure "importance" by how much the expected loss over some test set changes when neuron is lesioned:

$$\mathcal{I}_i(A) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_\mathcal{A}}\left[\left(\ell(z;\mathbf{h}) - \ell(z;\mathbf{h}|h_i = 0)\right)^2\right]$$

Where $\mathbf{h}$ is a vector

trivial to calculate in pytorch, tensorflow etc...

Taylor expanding gives:

$$\ell(z;\mathbf{h}|h_i = 0) = \ell(z;\mathbf{h}) + (\mathbf{h}_{\setminus i} - \mathbf{h})^\top \frac{\partial \ell}{\partial \mathbf{h}} + \frac{1}{2}(\mathbf{h}_{\setminus i} - \mathbf{h})^\top \mathrm{H}(\mathbf{h}_{\setminus i} - \mathbf{h}) + \ldots$$

So, to first order,

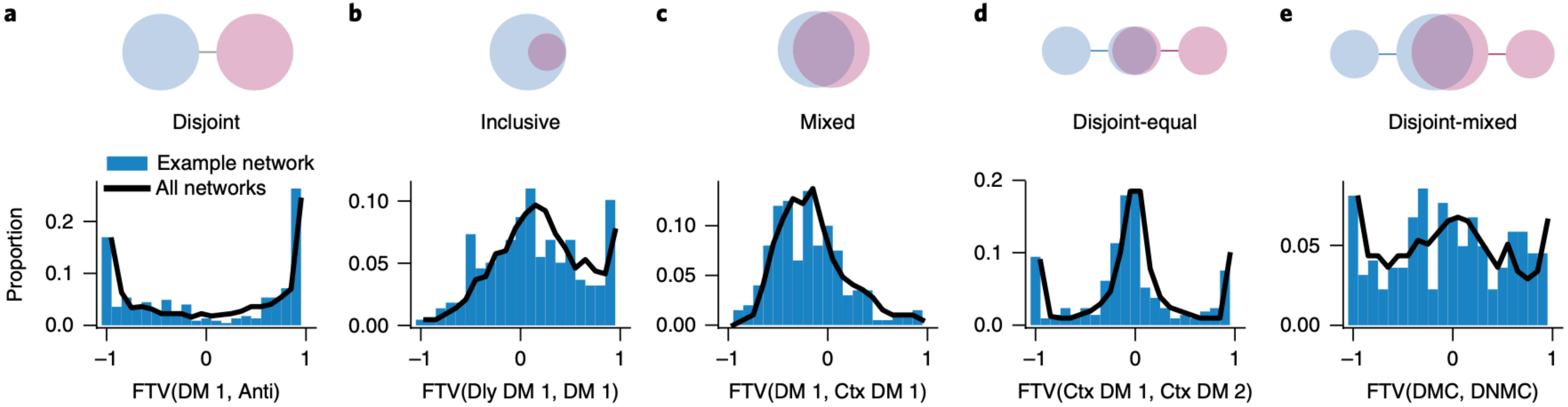
$$\boxed{\mathcal{I}_i(A) \approx \mathop{\mathbb{E}}_{z\sim\mathcal{D}_\mathcal{A}}\left[\left(h_i \cdot \frac{\partial \ell}{\partial h_i}\right)^2\right]}$$

The importance of hidden neuron *i* for task *A*

‘Relative Importance’ tells us about a neuron’s selectivity

$$\mathcal{RI}_i(A,B) = \frac{\mathcal{I}_i(A) - \mathcal{I}_i(B)}{\mathcal{I}_i(A) + \mathcal{I}_i(B)}$$

- ~1 means neuron is important for task A but not task B
- ~-1 means neuron is important for task B but not task A
- ~0 means neuron is equally important for both tasks



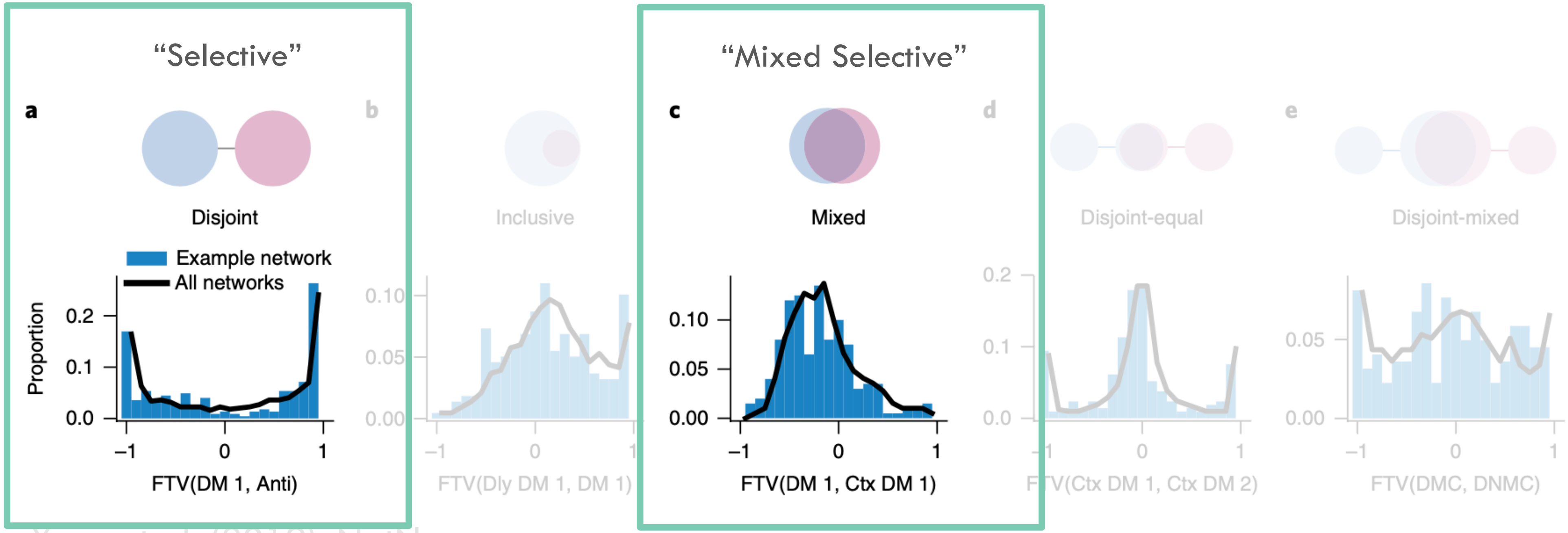
Plotting these as histograms over all hidden neurons gives very good indication about how the **network** represents the **tasks**

Yang et al. (2019), NatNeuro

‘Relative Importance’ tells us about a neuron’s selectivity

$$\mathcal{RI}_i(A, B) = \frac{\mathcal{I}_i(A) - \mathcal{I}_i(B)}{\mathcal{I}_i(A) + \mathcal{I}_i(B)}$$

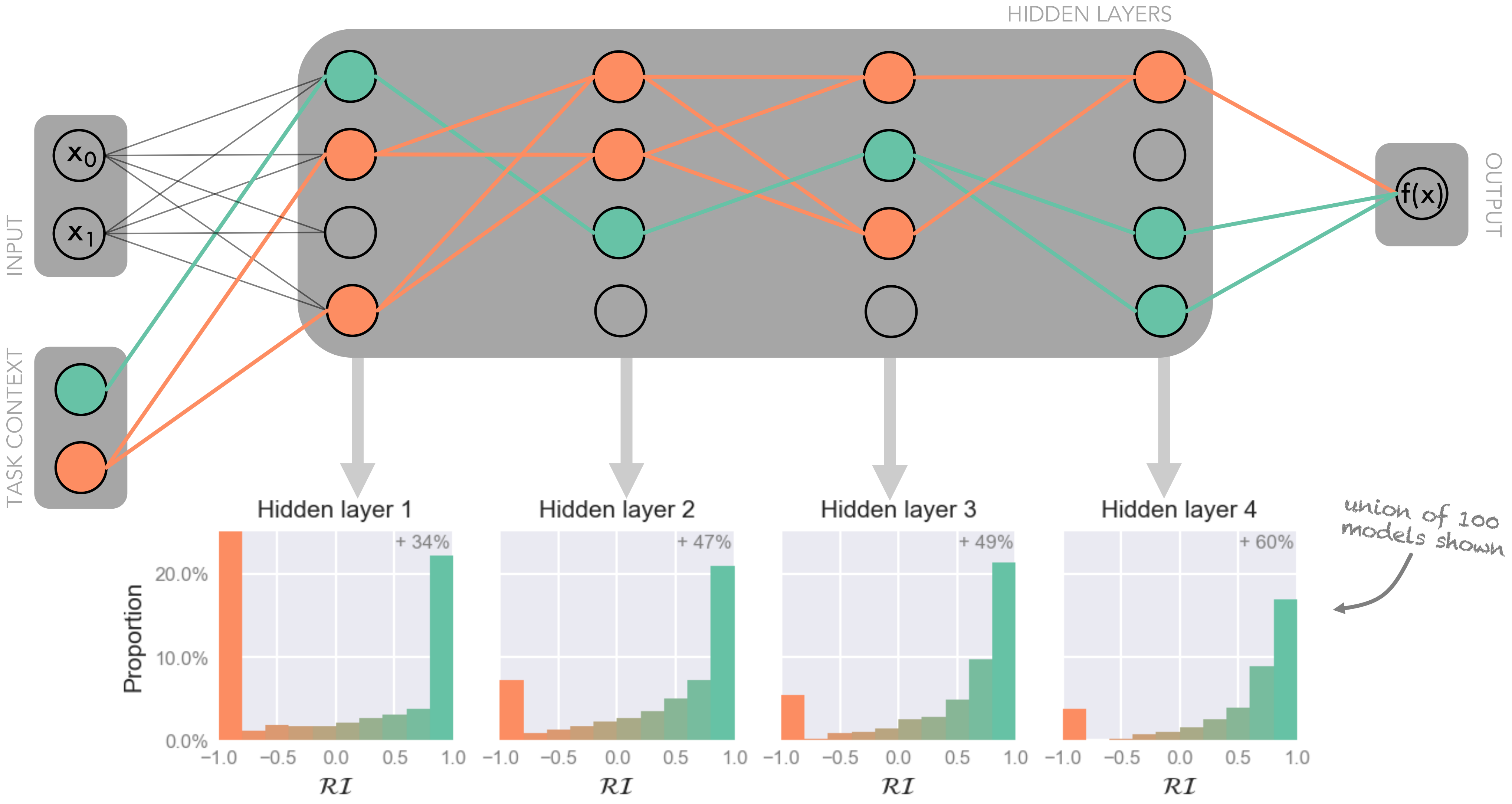
- ~1 means neuron is important for task A but not task B
- ~-1 means neuron is important for task B but not task A
- ~0 means neuron is equally important for both tasks



Plotting these as histograms over all hidden neurons gives very good indication about how the **network** represents the **tasks**

Yang et al. (2019), NatNeuro

1) Task context splits network into distinct subnetworks

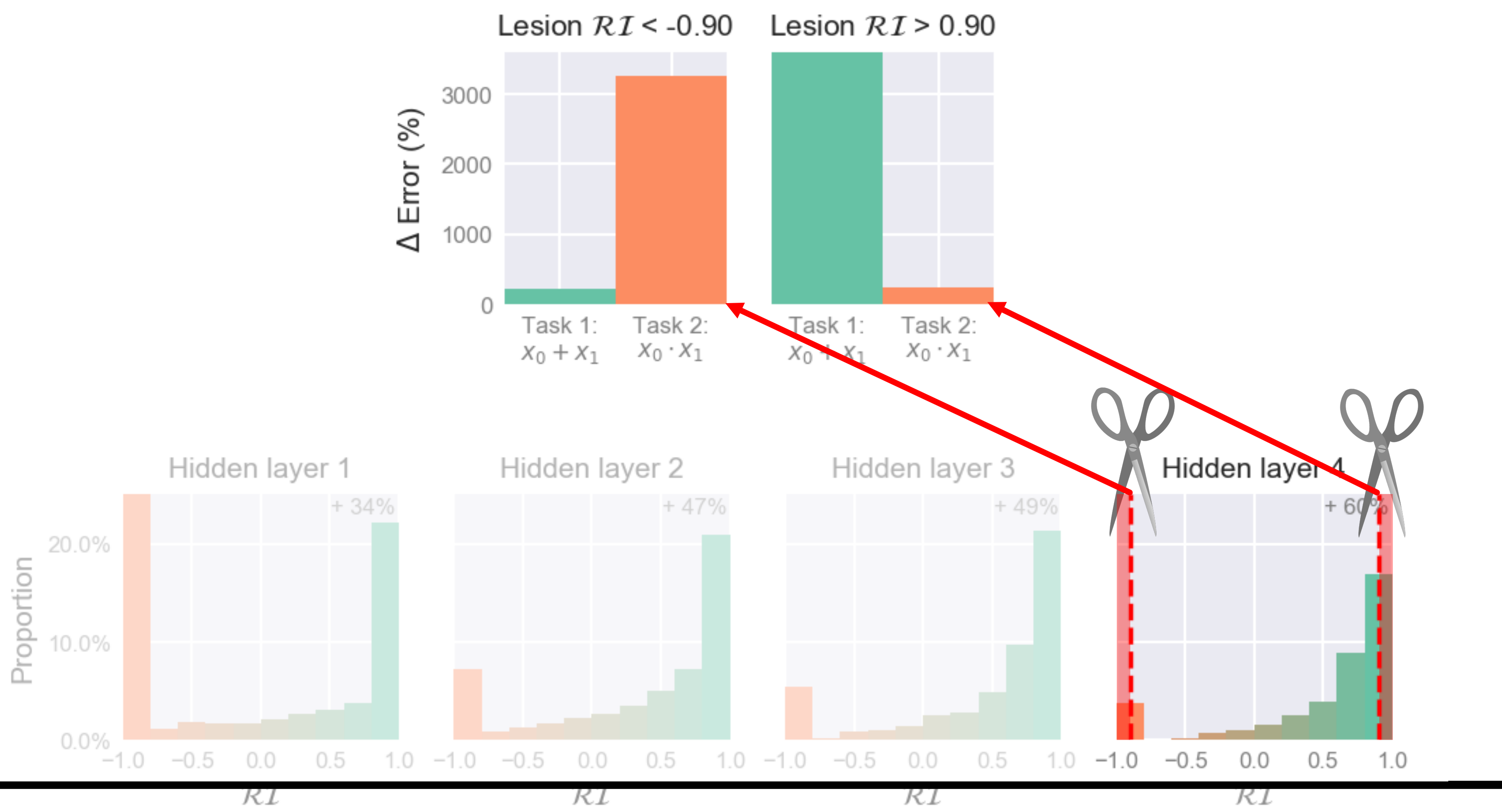


Take home:
One architecture … two networks.
Task context is the switch.

1) T

SANITY CHECK

'Relative importance' is a good indicator of neuronal selectivity, as we see by these lesion experiments



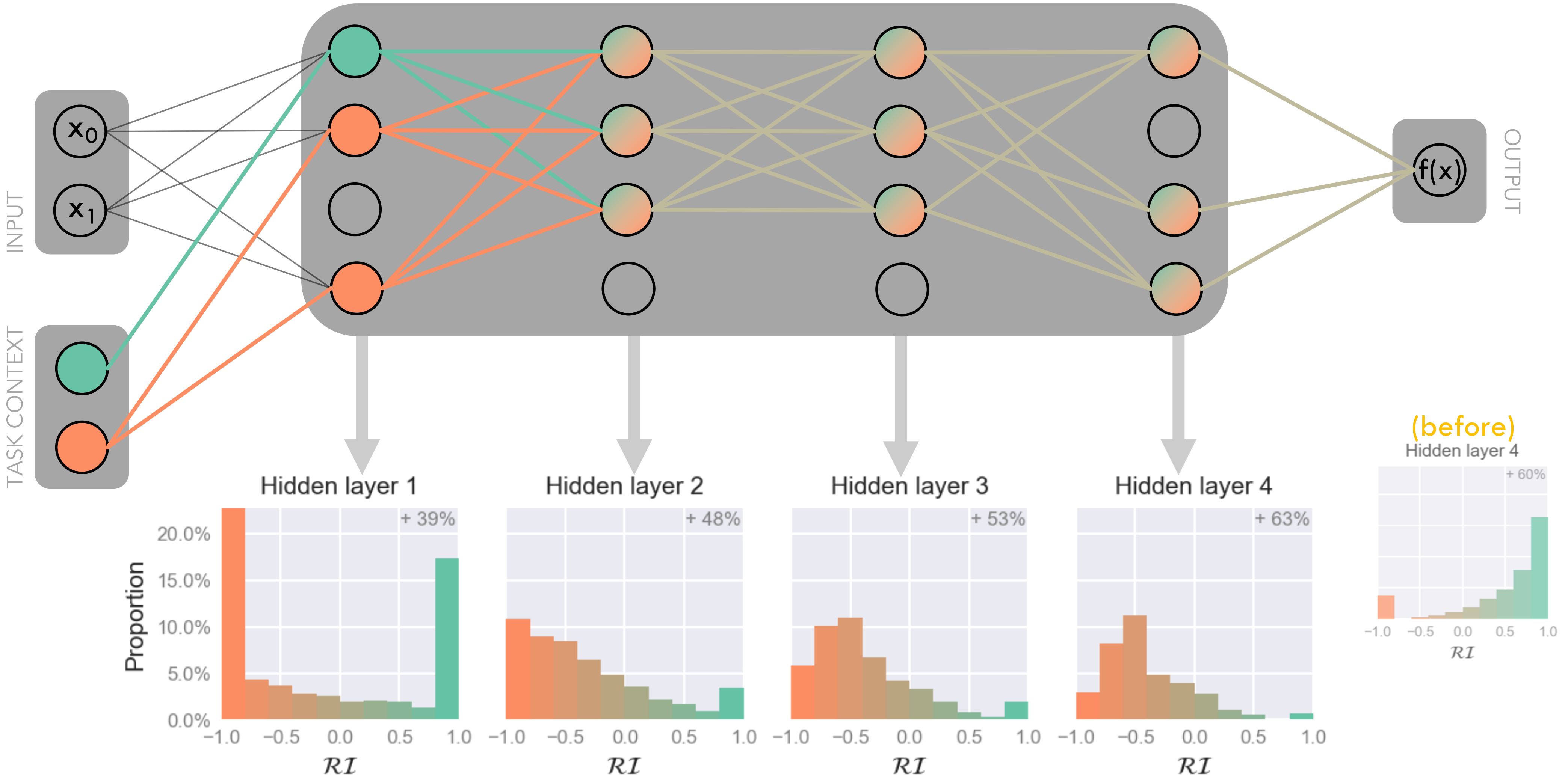
Take I
One
archit
…two
netwo

2) How 'similar' tasks are matter a lot

Task 1: $f(x) = x_0 + x_1$

~~**Task 2:** $f(x) = x_0 x_1$~~

Task 2: $f(x) = x_0 + 1.5x_1$

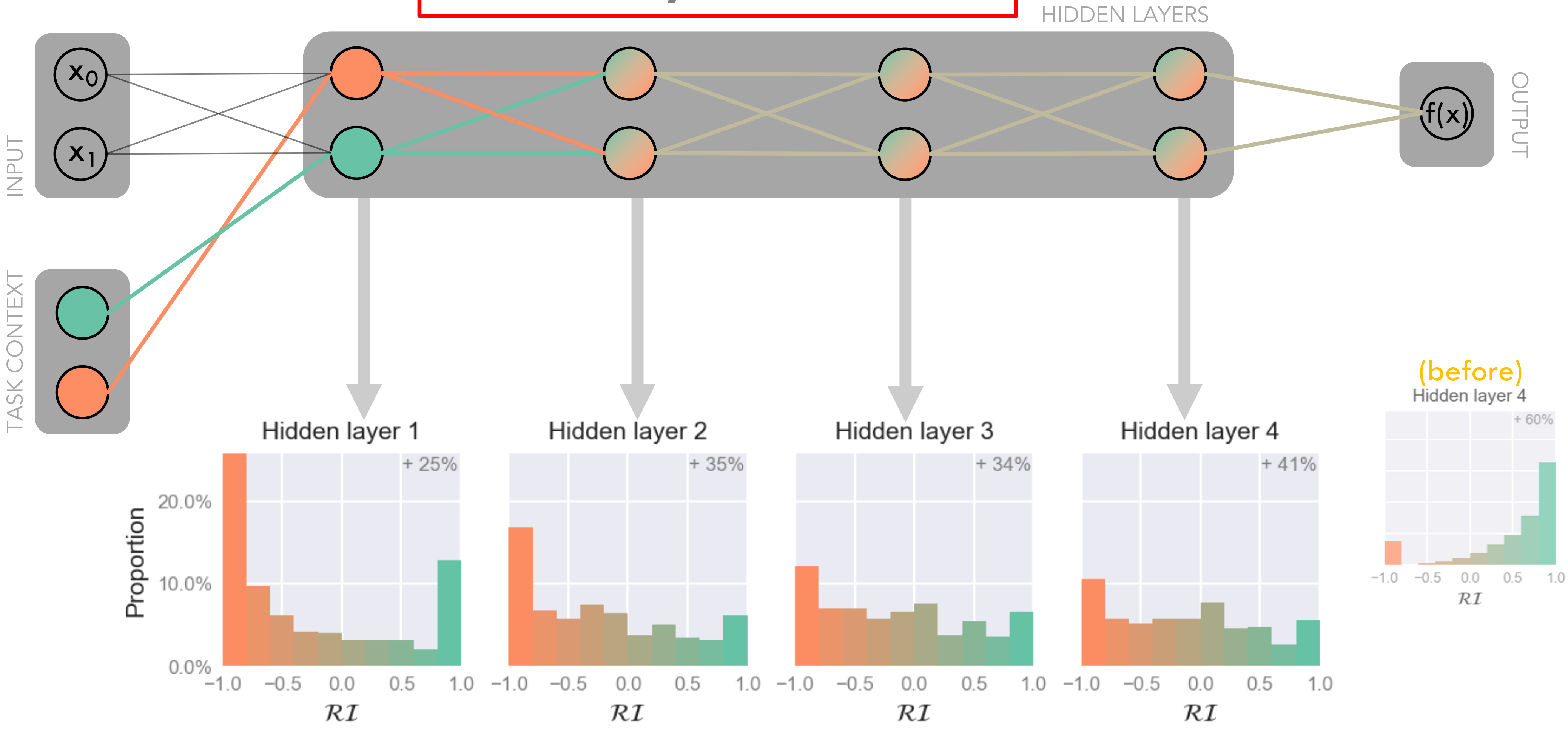


Take home:
Networks can recognize and exploit when tasks are similar

3) Constraining the network forces mixed selectivity

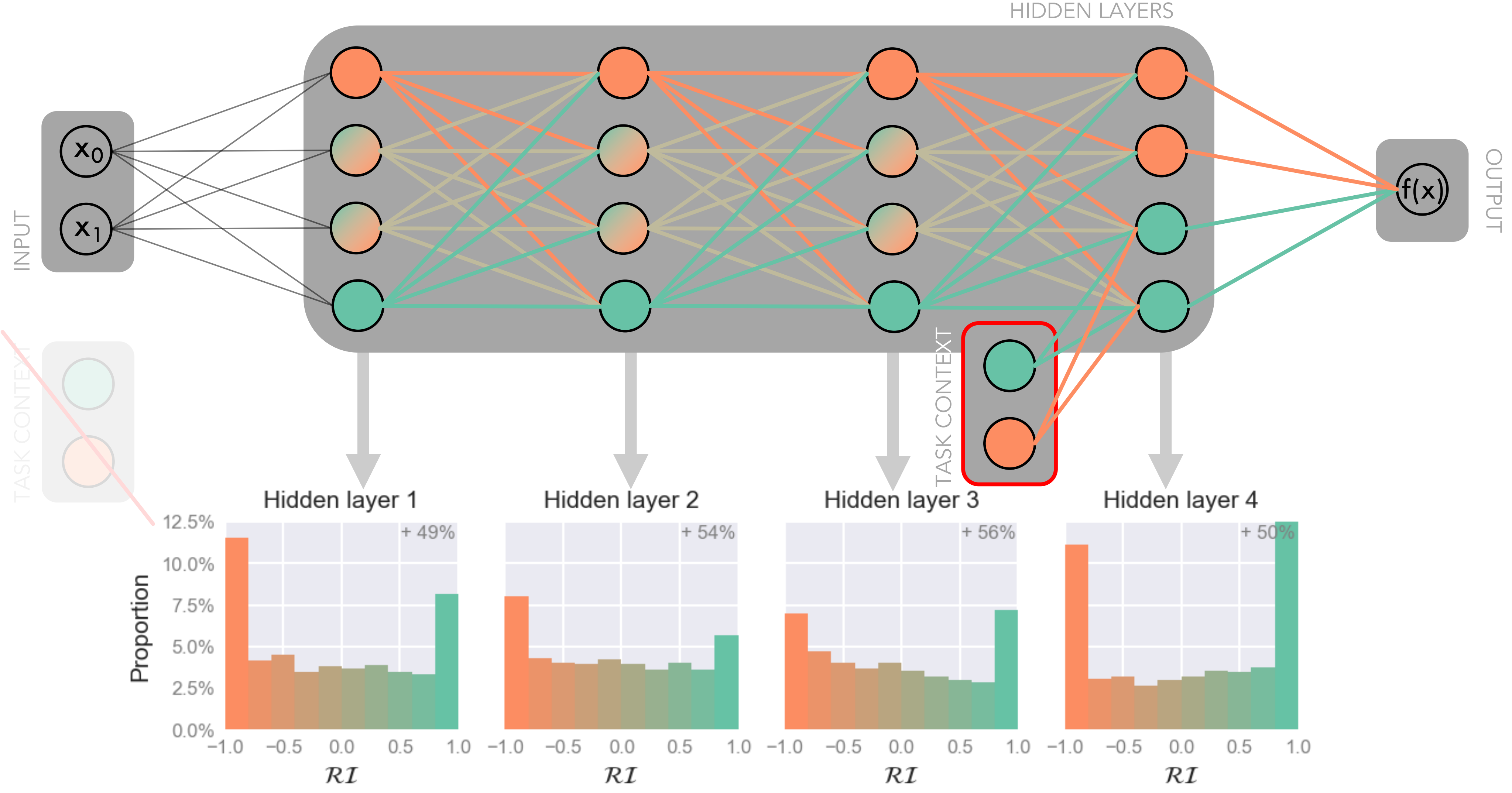
~~Hidden layer size = 100~~

Hidden layer size = 5



Take home:
Networks capacity matters

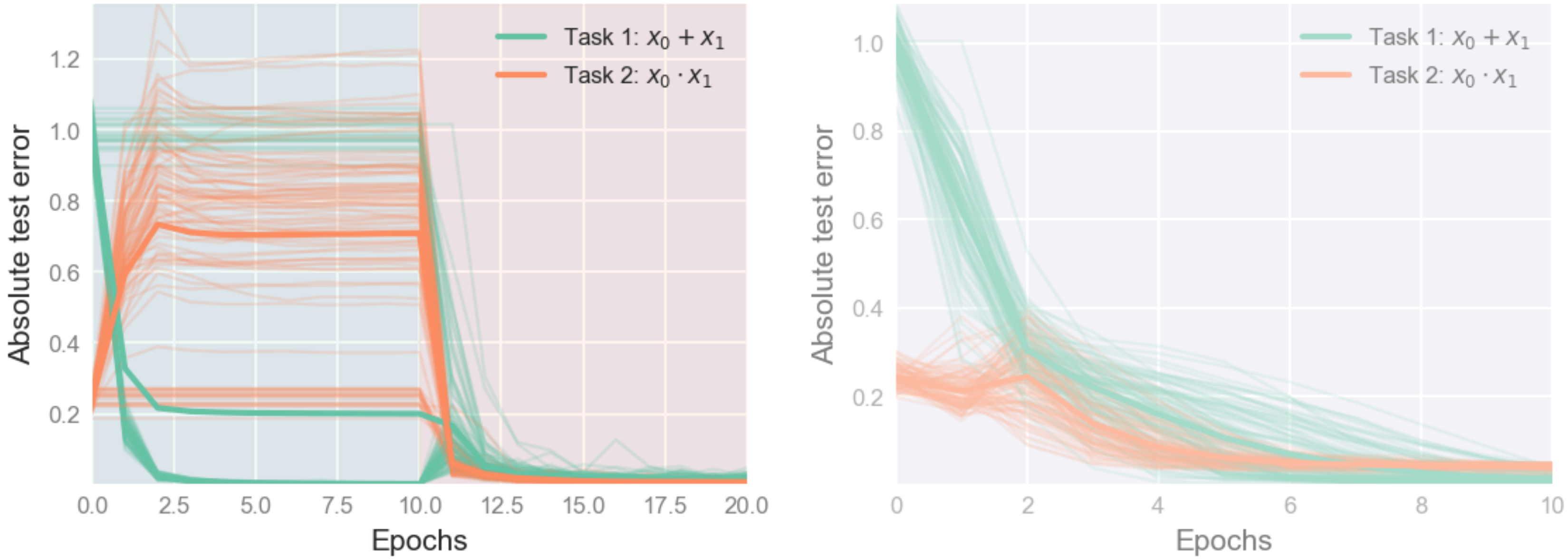
4) Which-task information can flow backwards

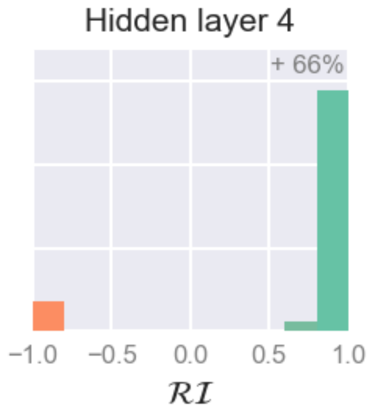


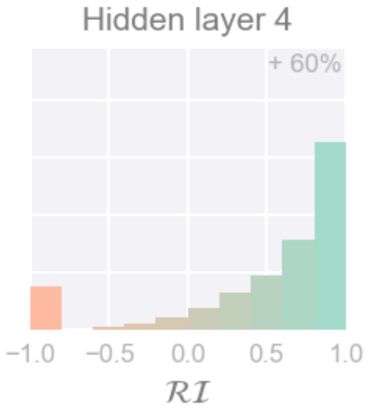
<u>Take home:</u>
Early layers are not task-independent feature extractors

5) 'Replay' style learning encourages selectivity

(before)







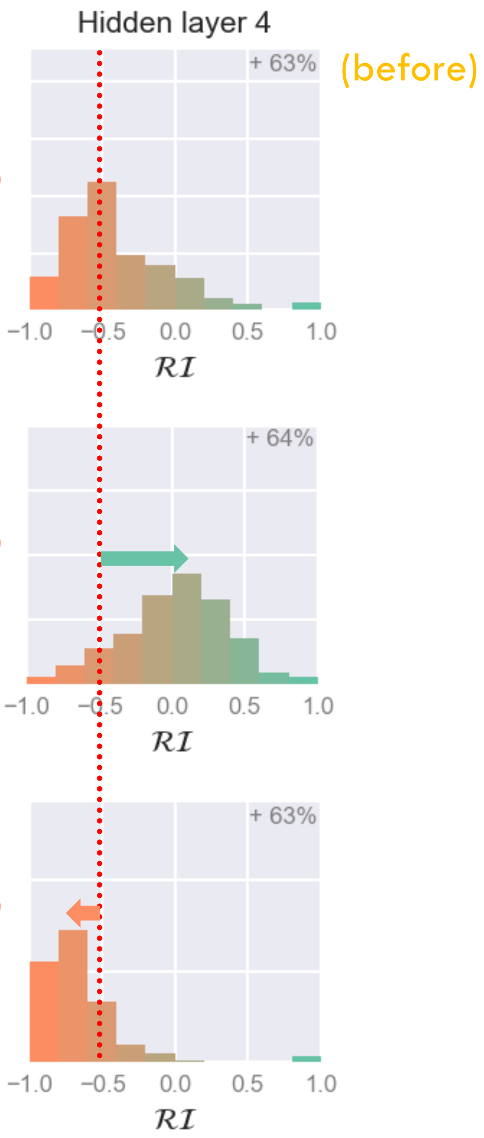
<u>Take home:</u>
Training order definitely matters

6) Biased learning encourages neurons to prioritize the more infrequent task

Task 1 50:50 Task 2

Task 1 20:80 Task 2

Task 1 80:20 Task 2



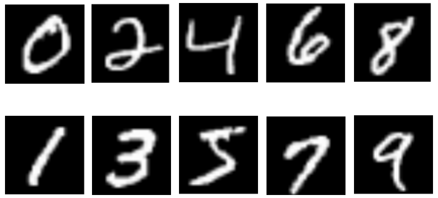
Take home:
We need fewer neurons to perform common tasks, not more

Roadmap

1. Ideas from the literature
2. A simple model trained on simple tasks
3. A more complex model trained on MNIST tasks
4. Conclusions

A more complex model: CNN + MNIST subsets

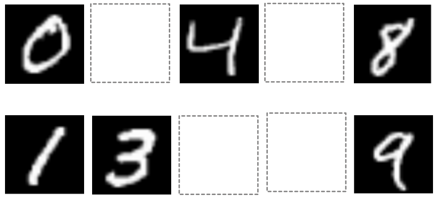
Task 1:
Odds vs
Evens

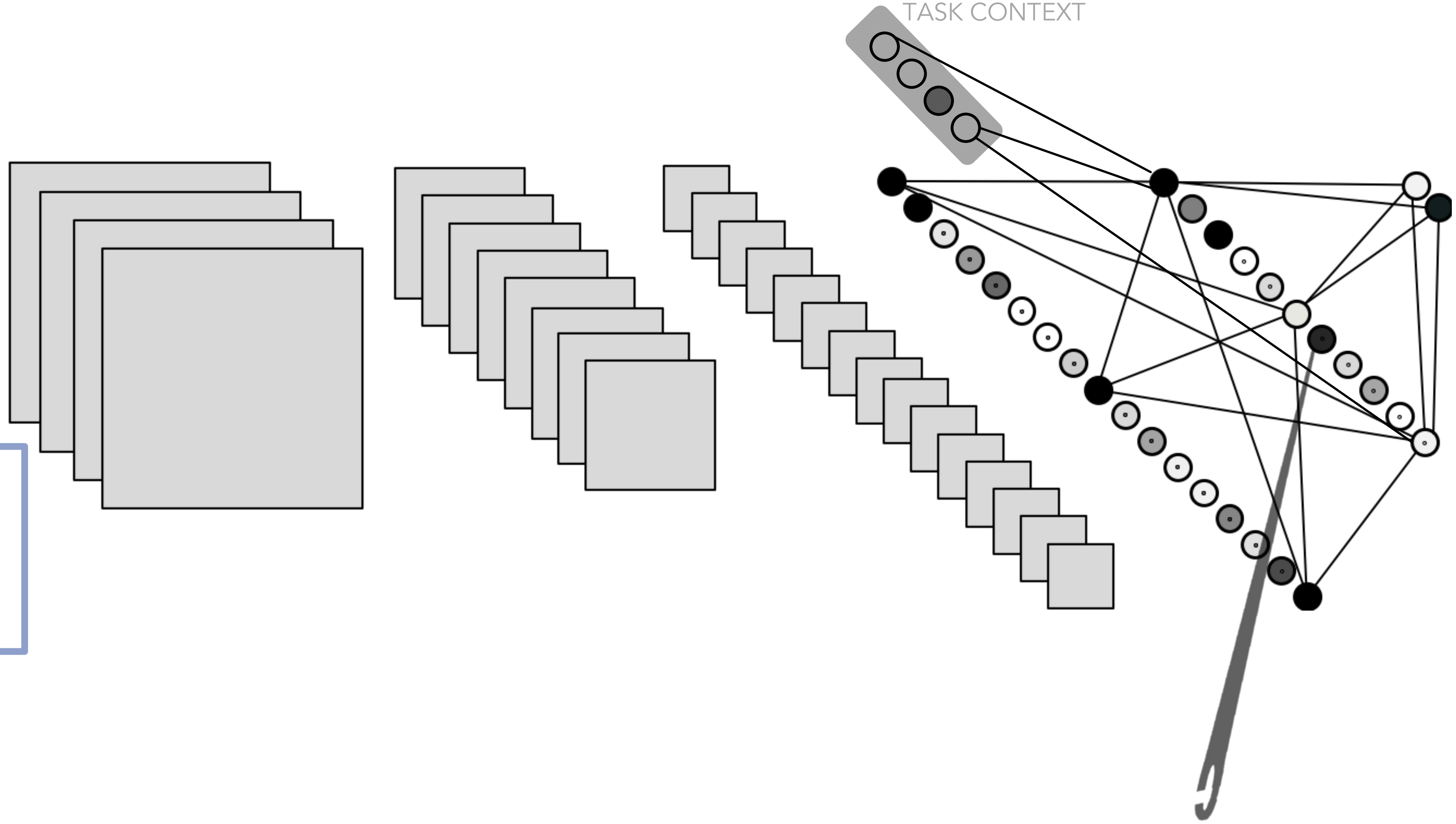


Task 2:
<5 vs
>=5

Task 3:
Prime vs
non-prime

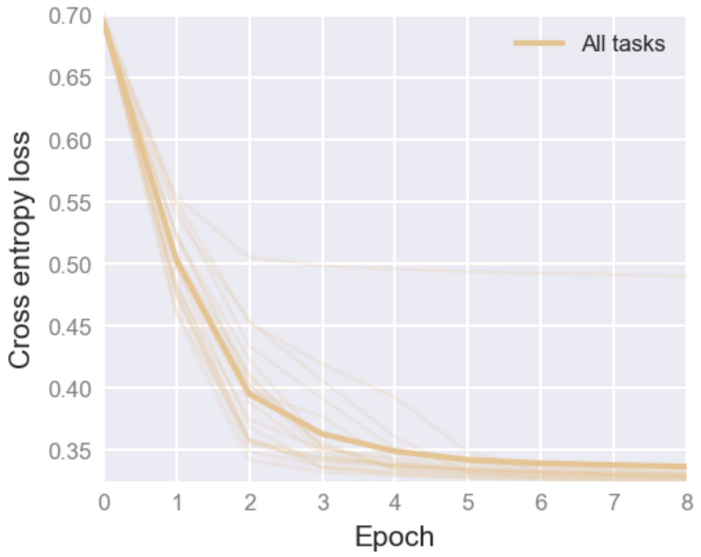
Task 4:
$\subset$ task 1

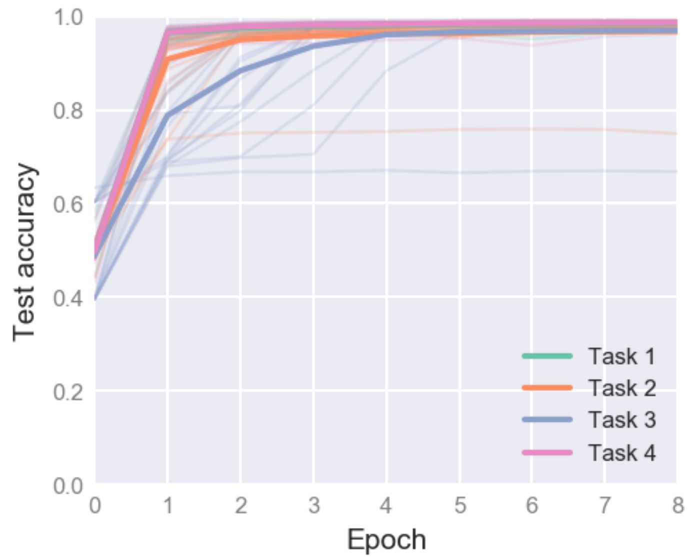


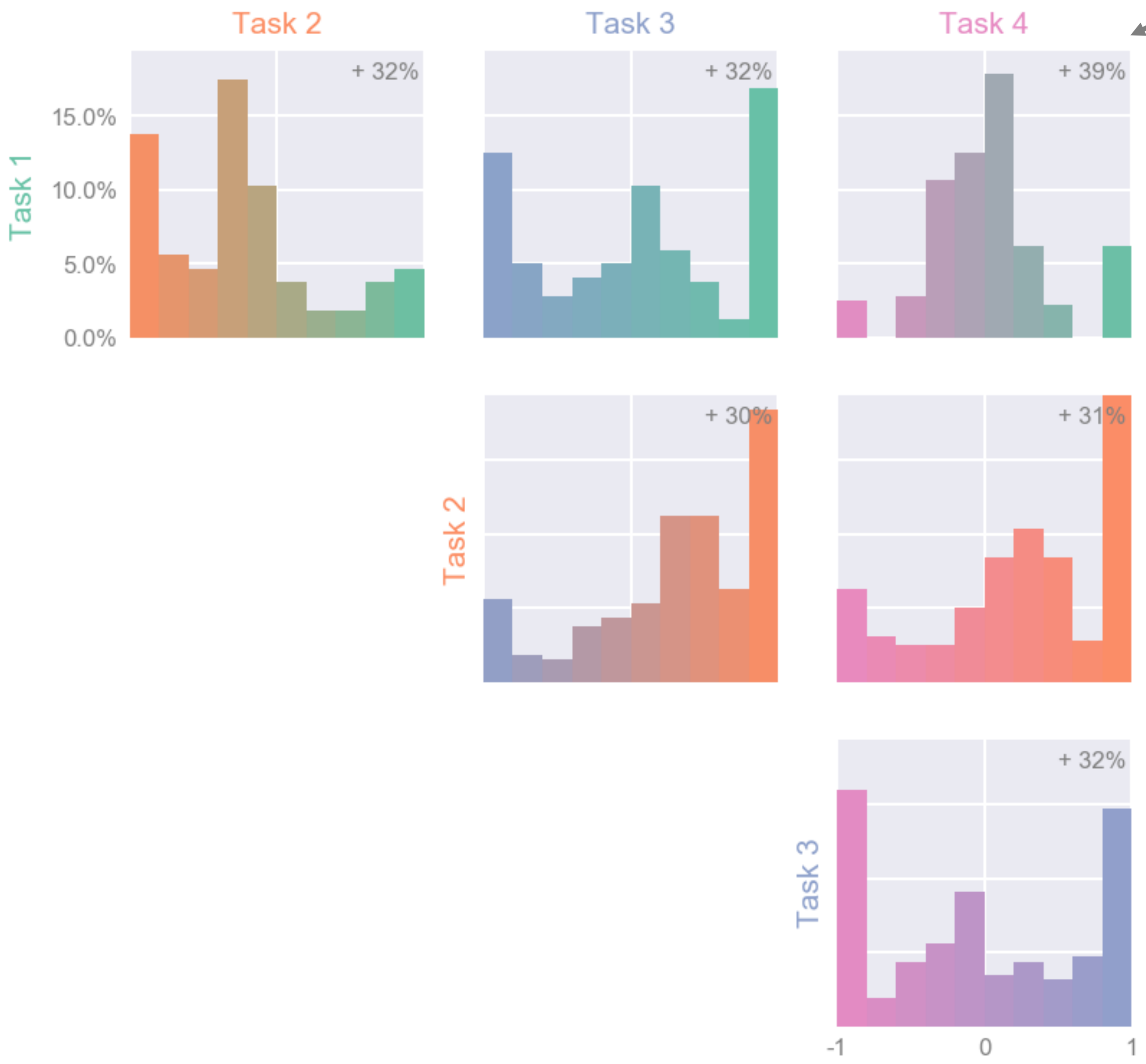


Randomly training on all tasks

union of 20 models shown



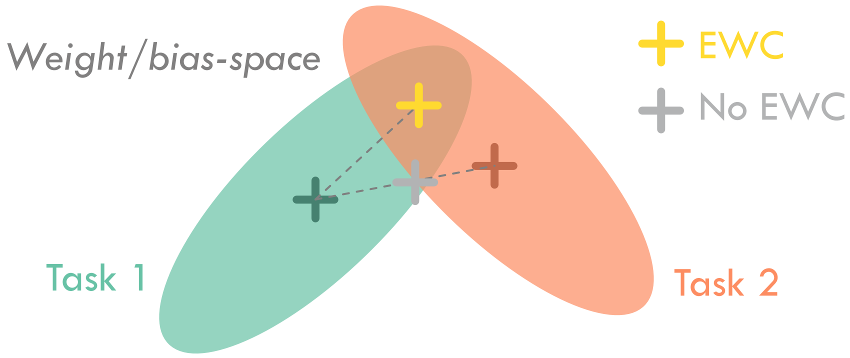




Randomly training on all tasks

Randomly training on all tasks

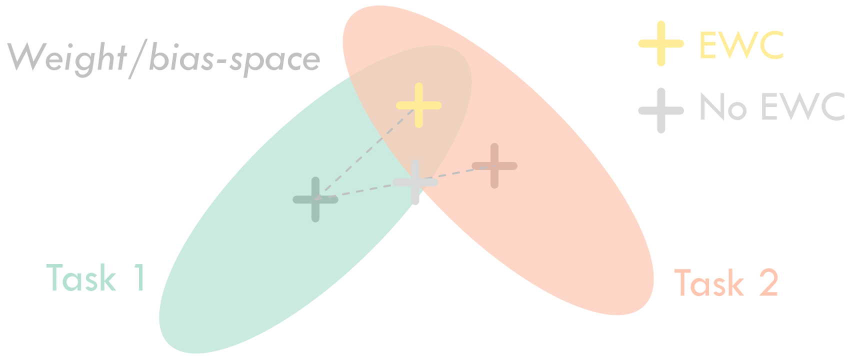
Continual learning via Elastic Weights Consolidation



$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{i,A})^2$$

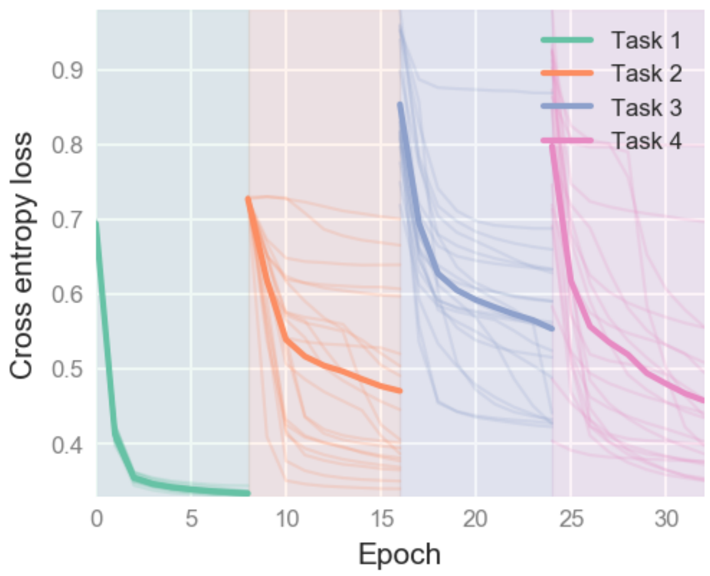
Quadratic weight penalty penalizes any changes in weights which are 'important' for previous tasks

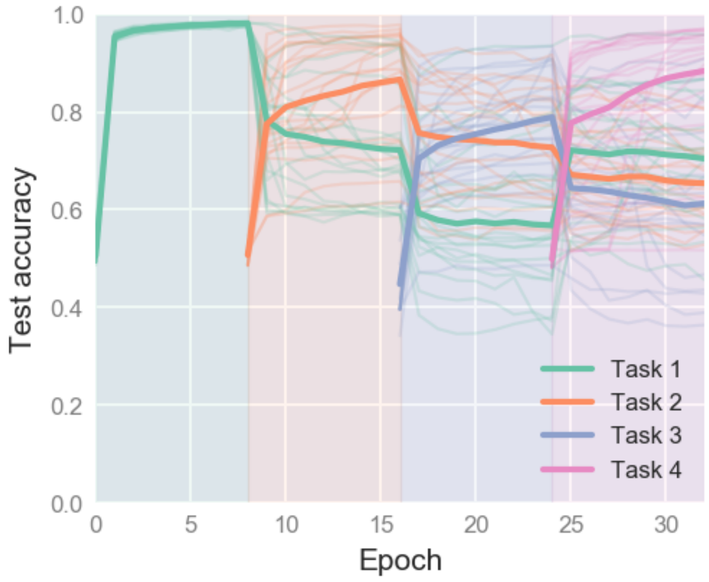
Continual learning via Elastic Weights Consolidation

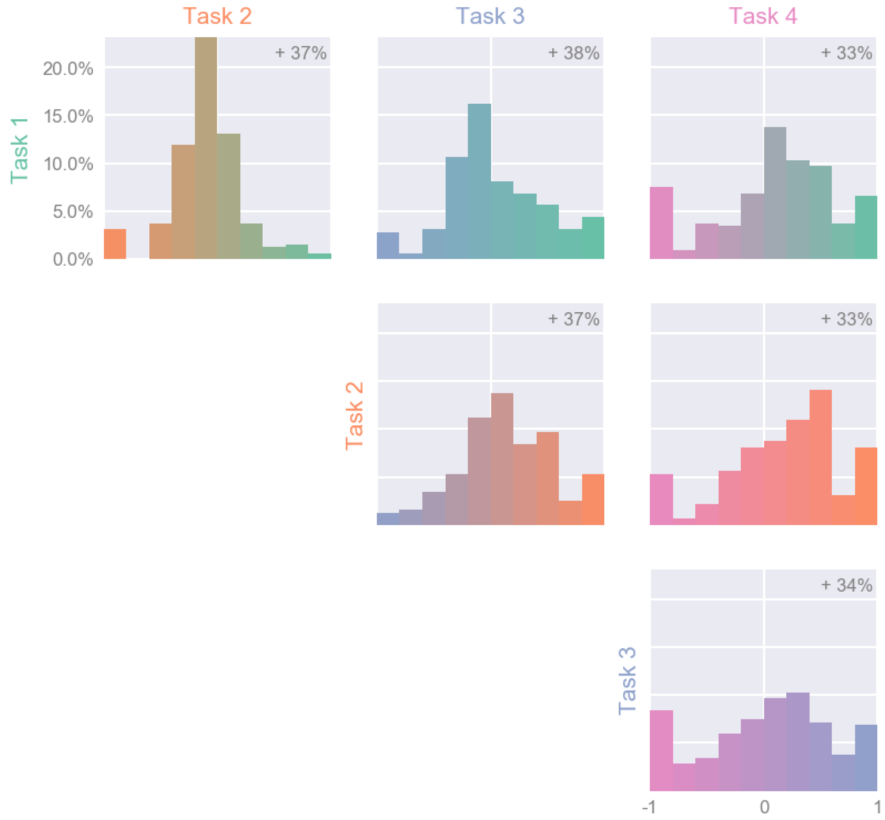


$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{i,A})^2$$

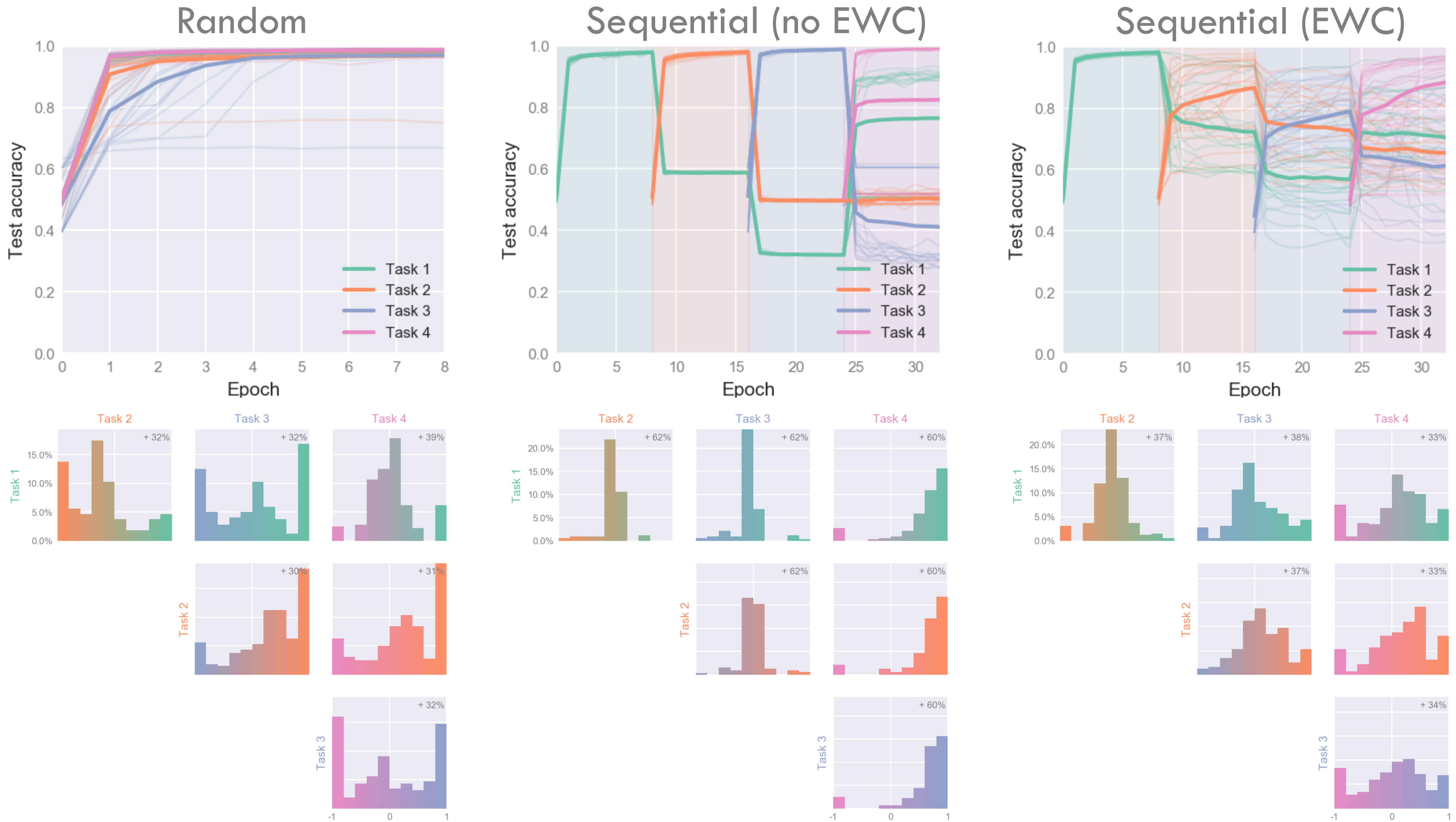
Quadratic weight penalty penalizes any changes in weights which are 'important' for previous tasks







Comparison of training styles



Comparison of training styles

Random

Sequential (no EWC)

Sequential (EWC)

When training is sequential neurons become mixed selective among all early tasks

Hypothesis: Selectivity, although **optimal**, is **unstable** and can't survive the overwriting process of sequential learning

Roadmap

1. Ideas from the literature
2. A simple model trained on simple tasks
3. A more complex model trained on MNIST tasks
4. Conclusions

Conclusions

1. Networks **recognize and exploit task similarities** by developing mixed-selective neurons.

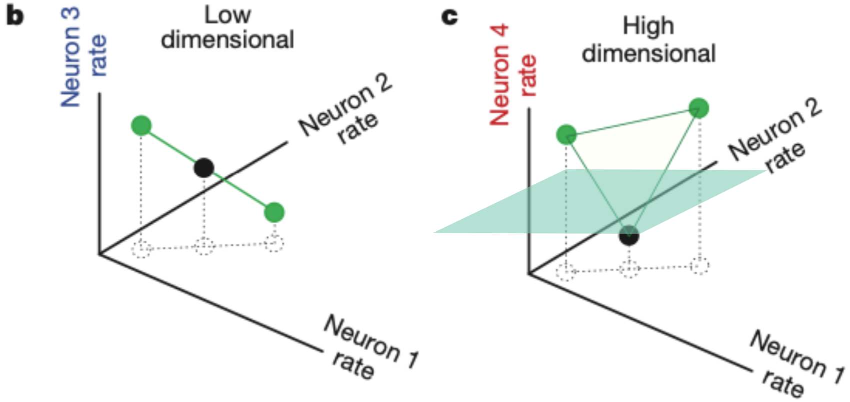
2. **Tasks, architecture and learning style** can all affect selectivity.

3. Neurons **specialize *in favour* of rare tasks.**

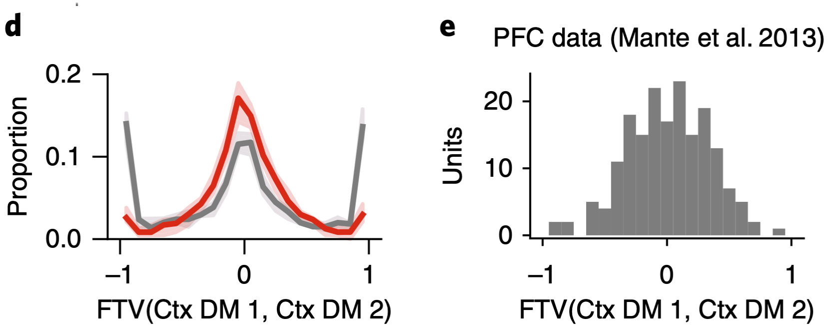
4. PFC neurons could be mixed because cognitive skills are learned in a more **blocked fashion** than, visual or motor skills [a highly debatable point in itself].
   - Neurons can't maintain selectivity to a task if they are later trained on many others.

5. Capacity constraints force neurons to be **mixed selective to "save space".**

So how does this fit in to the literature?

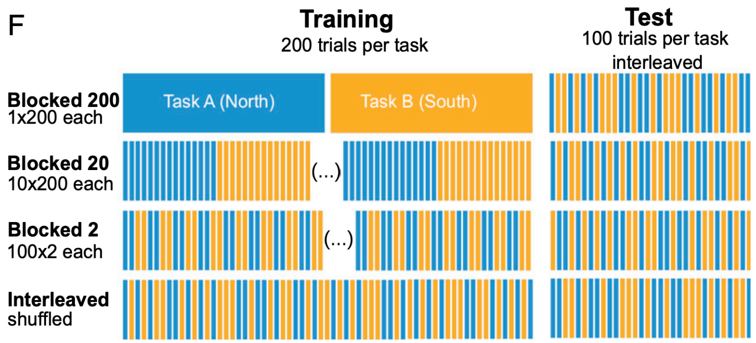


Rigotti et al. (2013): Our results support this study



Yang et al. (2019)

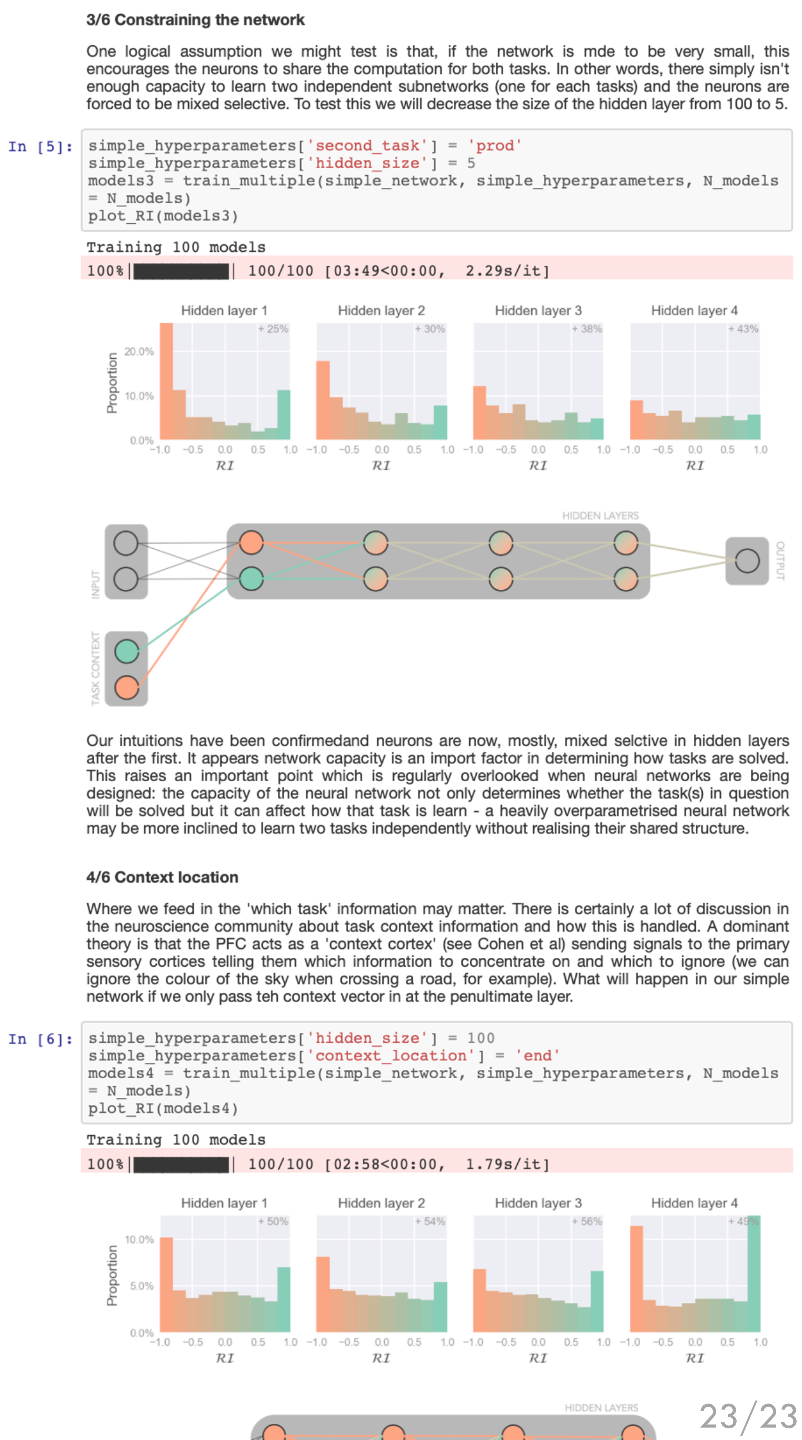
Our results tentatively support Yang.



Flesch et al. (2018)

Code

Code available on my Github page:
github.com/TomGeorge1234


3/6 Constraining the network

One logical assumption we might test is that, if the network is mde to be very small, this encourages the neurons to share the computation for both tasks. In other words, there simply isn't enough capacity to learn two independent subnetworks (one for each tasks) and the neurons are forced to be mixed selective. To test this we will decrease the size of the hidden layer from 100 to 5.

In [5]:
```
simple_hyperparameters['second_task'] = 'prod'
simple_hyperparameters['hidden_size'] = 5
models3 = train_multiple(simple_network, simple_hyperparameters, N_models
= N_models)
plot_RI(models3)
```

```
Training 100 models
100%|██████████| 100/100 [03:49<00:00,  2.29s/it]
```


Our intuitions have been confirmedand neurons are now, mostly, mixed selctive in hidden layers after the first. It appears network capacity is an import factor in determining how tasks are solved. This raises an important point which is regularly overlooked when neural networks are being designed: the capacity of the neural network not only determines whether the task(s) in question will be solved but it can affect how that task is learn - a heavily overparametrised neural network may be more inclined to learn two tasks independently without realising their shared structure.

4/6 Context location

Where we feed in the 'which task' information may matter. There is certainly a lot of discussion in the neuroscience community about task context information and how this is handled. A dominant theory is that the PFC acts as a 'context cortex' (see Cohen et al) sending signals to the primary sensory cortices telling them which information to concentrate on and which to ignore (we can ignore the colour of the sky when crossing a road, for example). What will happen in our simple network if we only pass teh context vector in at the penultimate layer.

In [6]:
```
simple_hyperparameters['hidden_size'] = 100
simple_hyperparameters['context_location'] = 'end'
models4 = train_multiple(simple_network, simple_hyperparameters, N_models
= N_models)
plot_RI(models4)
```

```
Training 100 models
100%|██████████| 100/100 [02:58<00:00,  1.79s/it]
```


Important References

- A. Ramirez-Cardenas and P. Viswanathan. The role of prefrontal mixed selectivity in cognitive control. Journal of Neuroscience, 36(35):9013–9015, August 2016
- Timo Flesch, Jan Balaguer, Ronald Dekker, Hamed Nili, and Christopher Summerfield. Comparing continual task learning in minds and machines. Proceedings of the National Academy of Sciences, 115(44):E10313–E10322, October 2018
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guil-laume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ra-malho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks, 2016.
- Earl K. Miller and Jonathan D. Cohen. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience, 24(1):167–202, March 2001.
- Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning, 2019.
- Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance ofmixed selectivity in complex cognitive tasks. Nature, 497(7451):585–590, May 2013.
- Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. Nature Neuroscience, 22(2):297–306, January 2019.