

Universality and Individuality in Recurrent Neural Networks

Niru Maheswaranathan, Alex Williams, Matthew D Golub, Surya Ganguli, David Sussillo

Google Brain & Stanford University

NeurIPS 2019

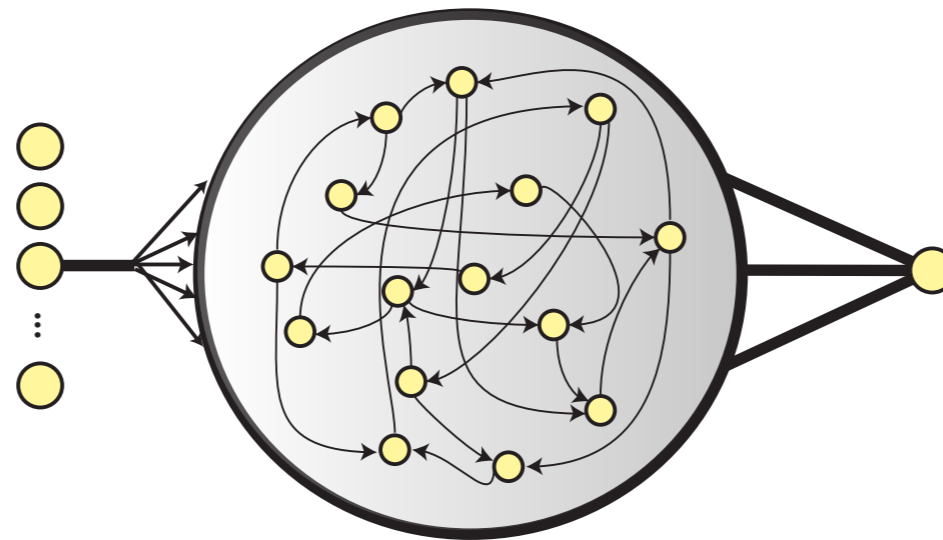
Poster #179

arxiv:1907.08549

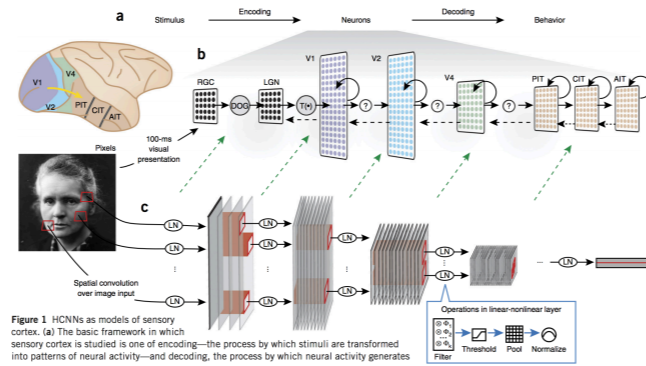
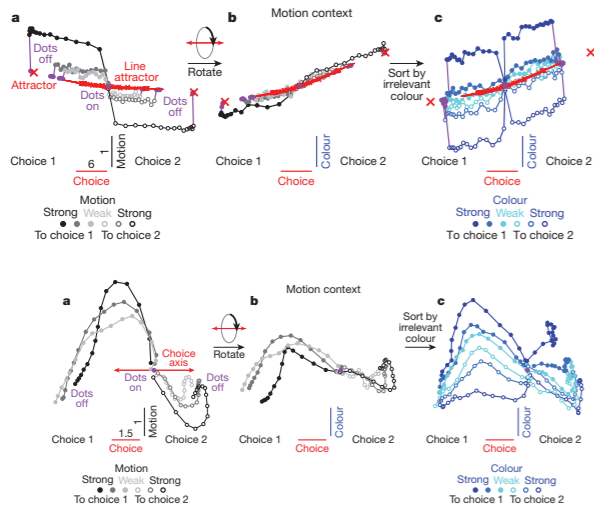
Artificial neural networks in neuroscience

Advantages:

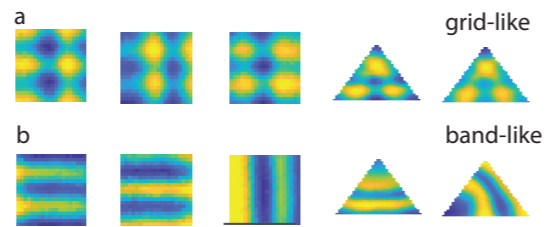
- Can train ANNs to accomplish tasks analogous to those studied in animals.
- Can inspect/probe/dissect artificial networks very easily.
- Can easily initiate a huge number of *in silico* studies



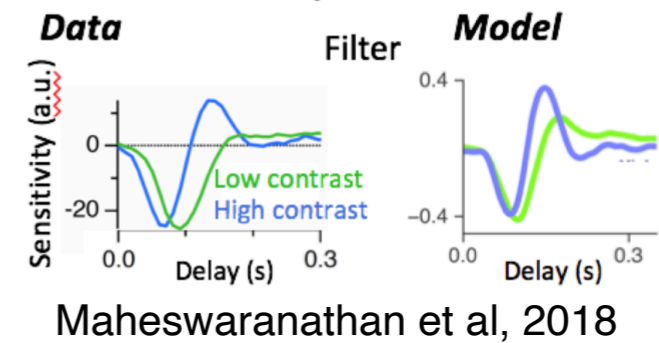
Artificial & biological neural networks



Yamins & DiCarlo, 2014

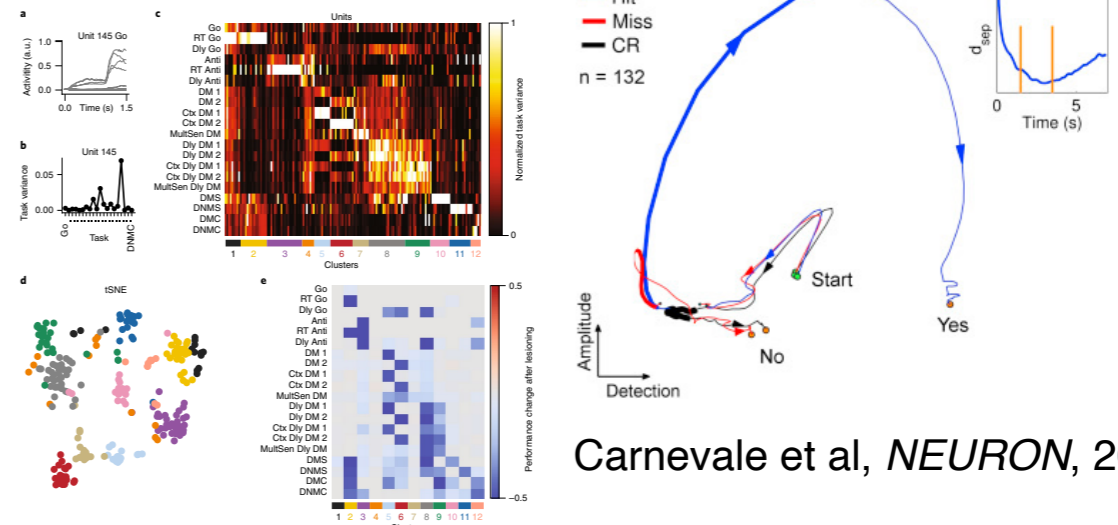


Cueva & Wei, ICLR, 2018



Maheswaranathan et al, 2018

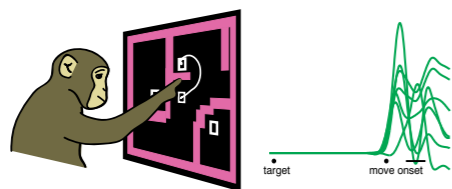
Mante & Sussillo et al. *Nature* 2013



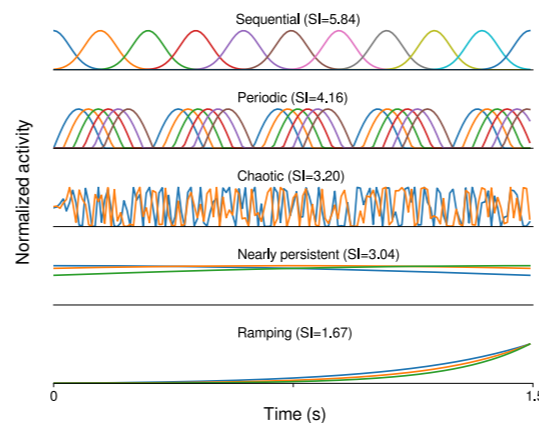
Carnevale et al, *NEURON*, 2015

Yang, et al., *Nature Neuroscience* 2019

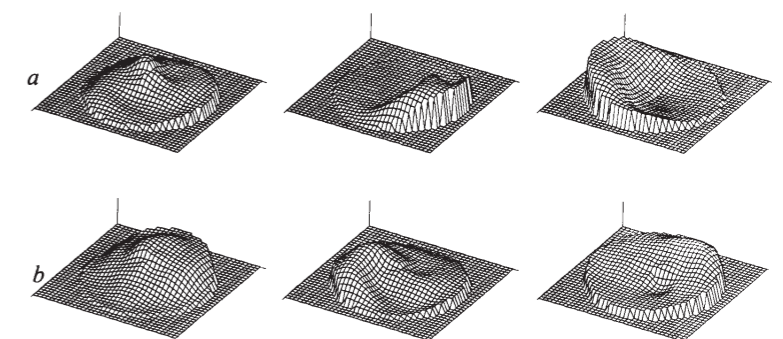
Rajan, Harvey, Tank, *Neuron*, 2016



Sussillo et al., *Nature Neuroscience*, 2015



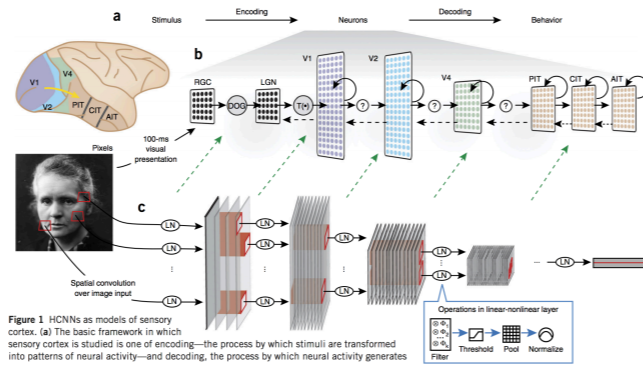
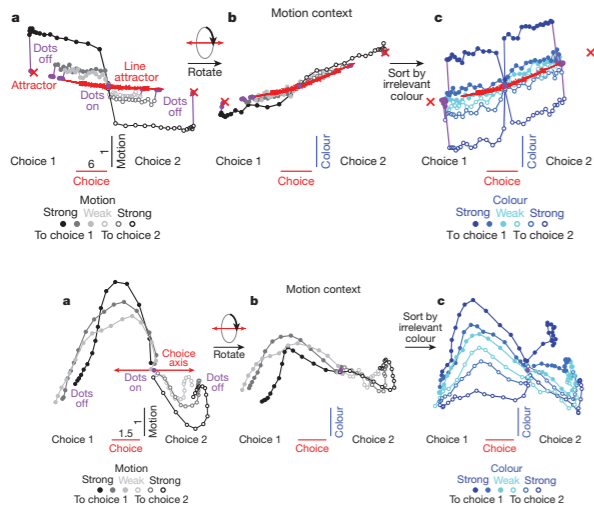
Orhan & Ma, *Nature Neuroscience*, 2019



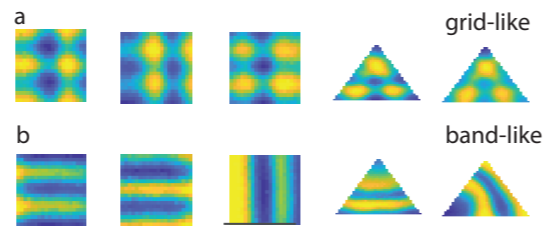
Zipser & Andersen, *Science*, 1988

Artificial & biological neural networks

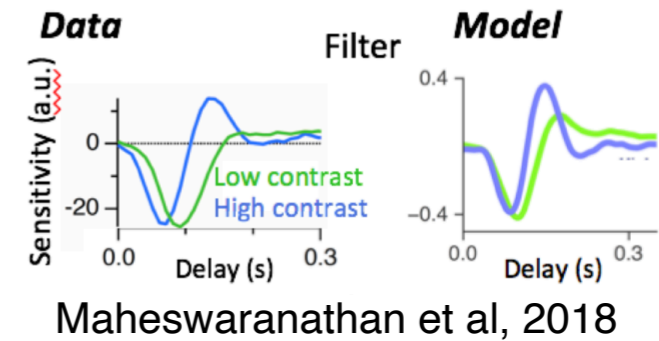
Networks have surprisingly similar representations...



Yamins & DiCarlo, 2014

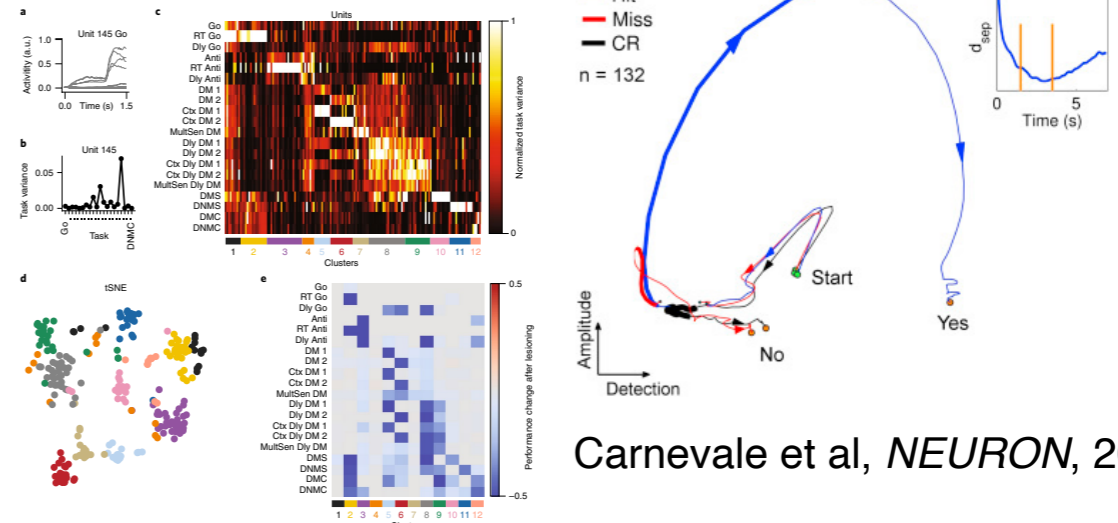


Cueva & Wei, ICLR, 2018



Maheswaranathan et al, 2018

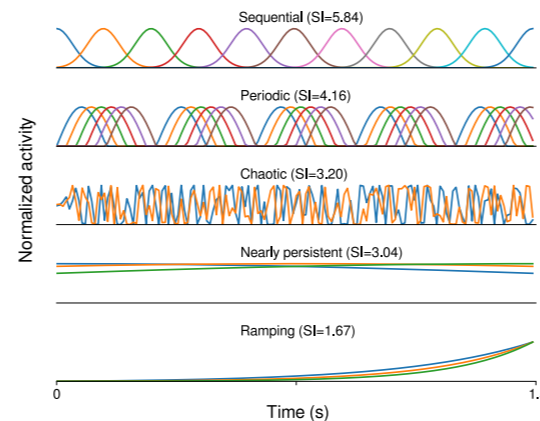
Mante & Sussillo et al. *Nature* 2013



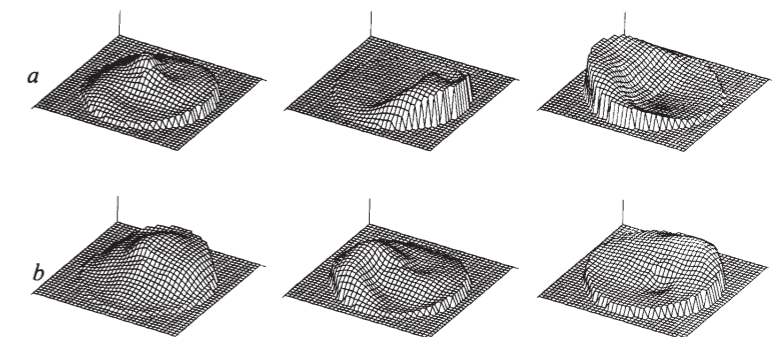
Carnevale et al, *NEURON*, 2015

Yang, et al., *Nature Neuroscience* 2019

Rajan, Harvey, Tank, *Neuron*, 2016

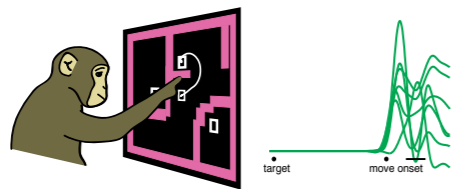


Orhan & Ma, *Nature Neuroscience*, 2019



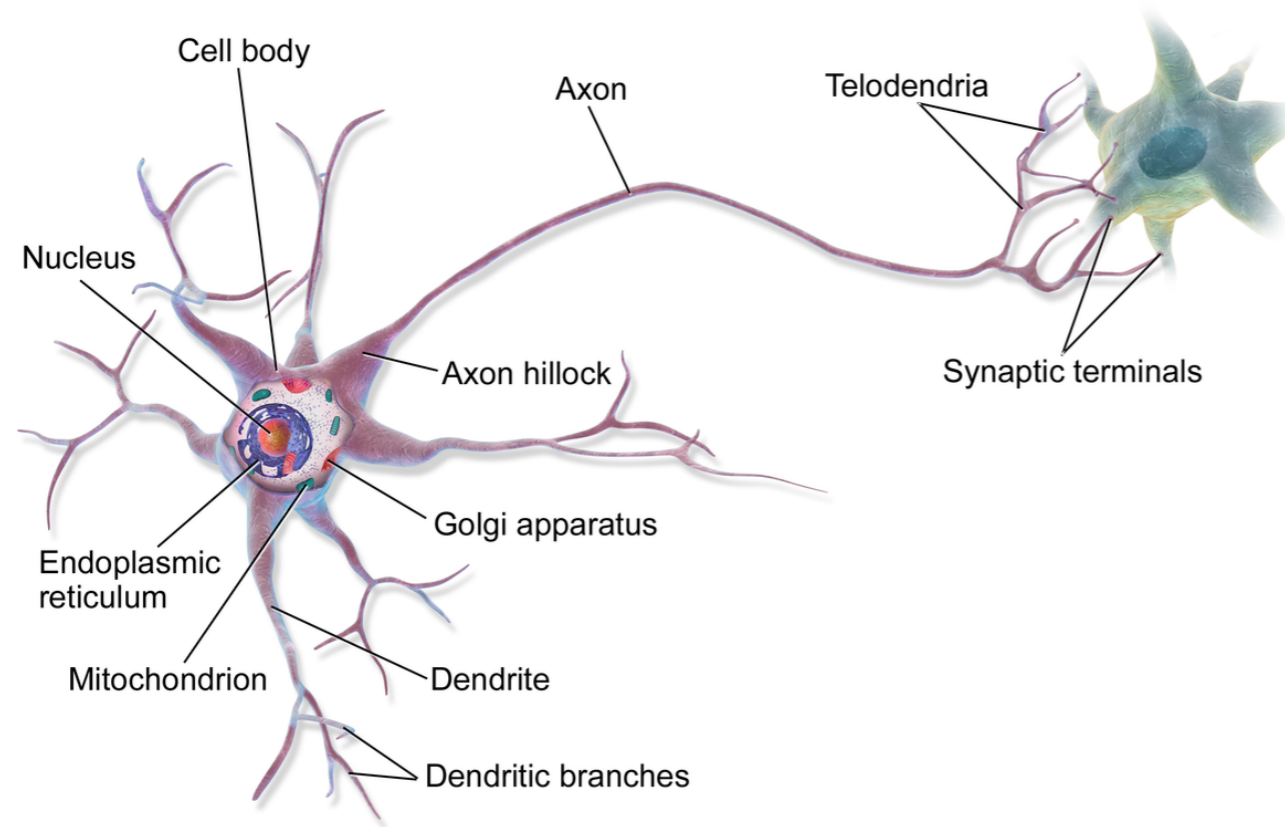
Zipser & Andersen, *Science*, 1988

Sussillo et al., *Nature Neuroscience*, 2015



Artificial & biological neural networks

...but are composed of drastically different elements!



Biological neuron

$$y_i = \tanh\left(\sum_i W_{ij} x_j\right)$$

Artificial neuron

Central question

When trained to perform the same **task**, why should we expect artificial and biological networks to be **similar**, given the drastic **differences in underlying mechanism**?

This work: an empirical approach

Network mechanisms

RNN architectures (e.g. LSTMs, GRUs, ...)

Nonlinearities (e.g. ReLU, tanh)

...

Similarity measures

Canonical correlation analysis (CCA)

Centered kernel alignment (CKA)

Tasks

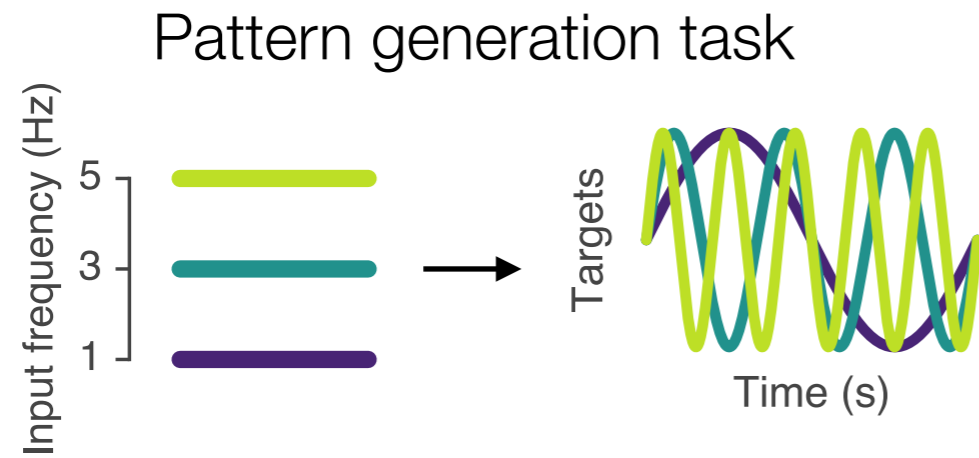
Decision making

Pattern generation

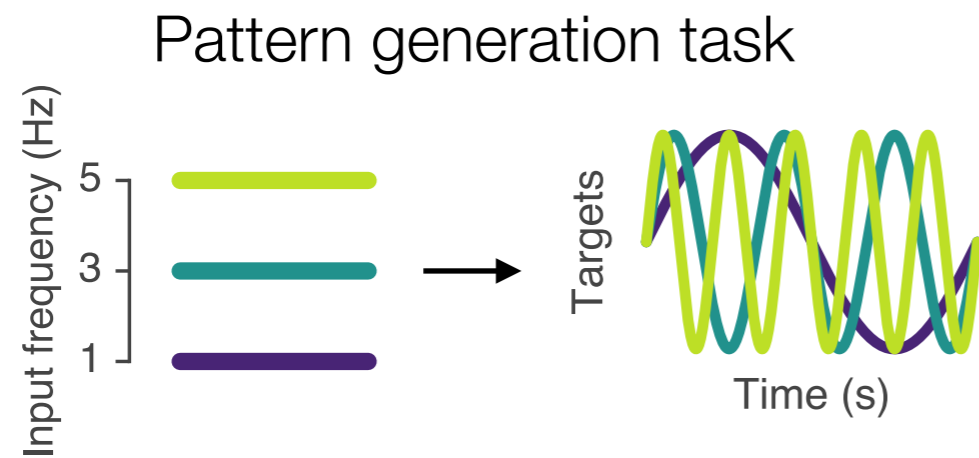
Working memory

Evidence of both *universality* and *individuality*

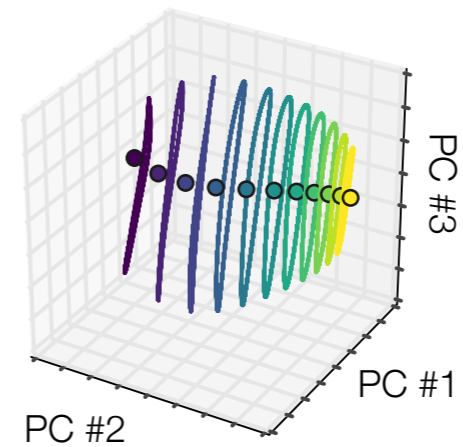
Evidence of both *universality* and *individuality*



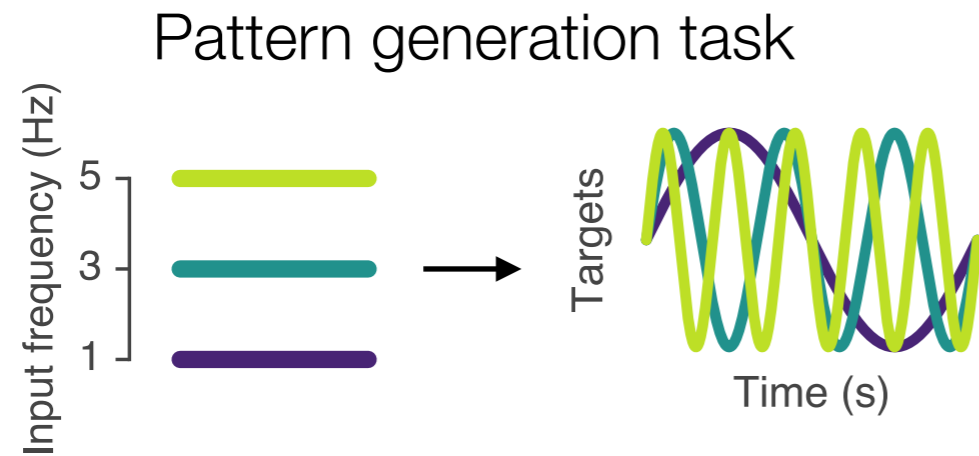
Evidence of both *universality* and *individuality*



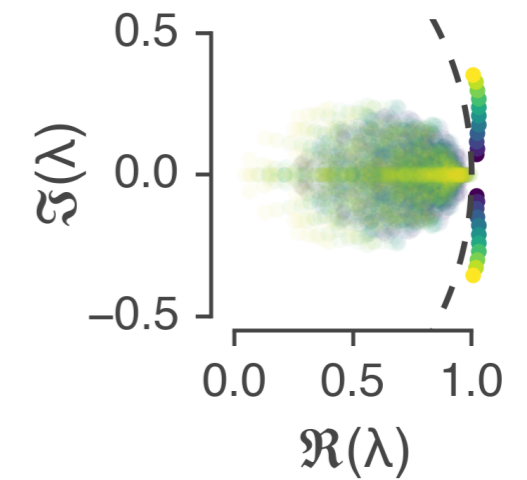
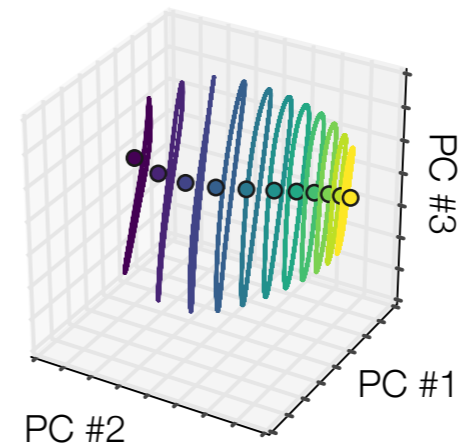
Analyzing trained networks



Evidence of both *universality* and *individuality*

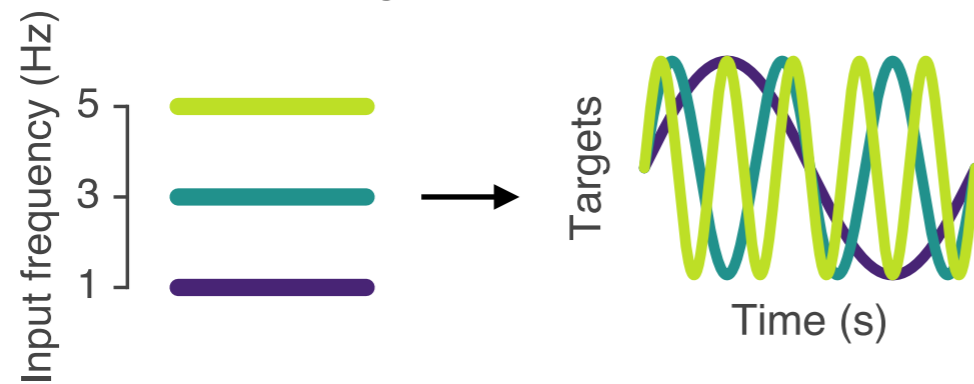


Analyzing trained networks

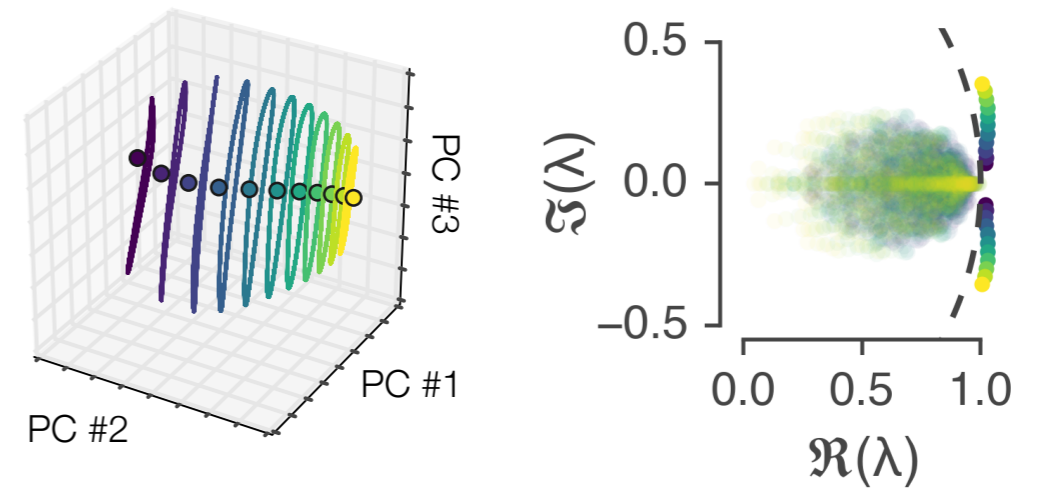


Evidence of both *universality* and *individuality*

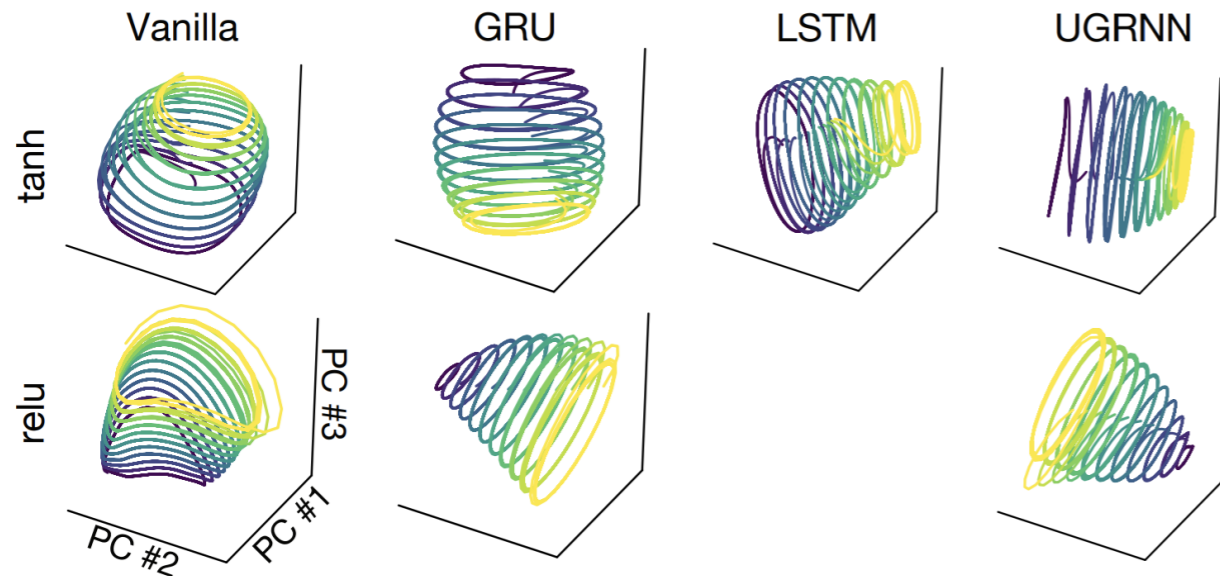
Pattern generation task



Analyzing trained networks

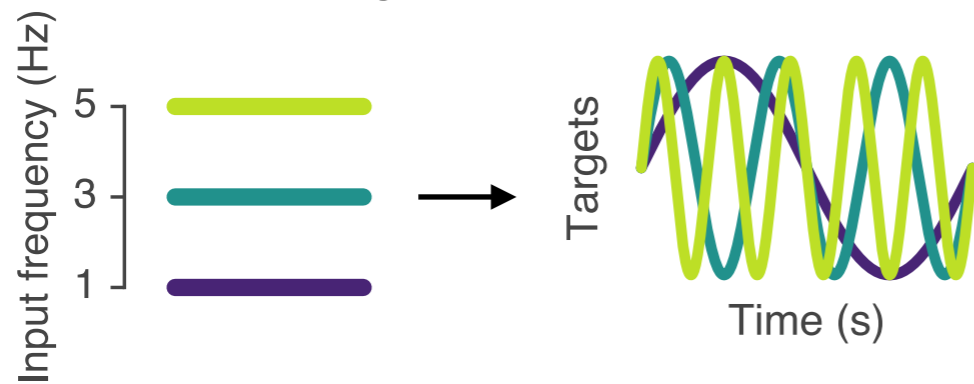


Network representations show individuality

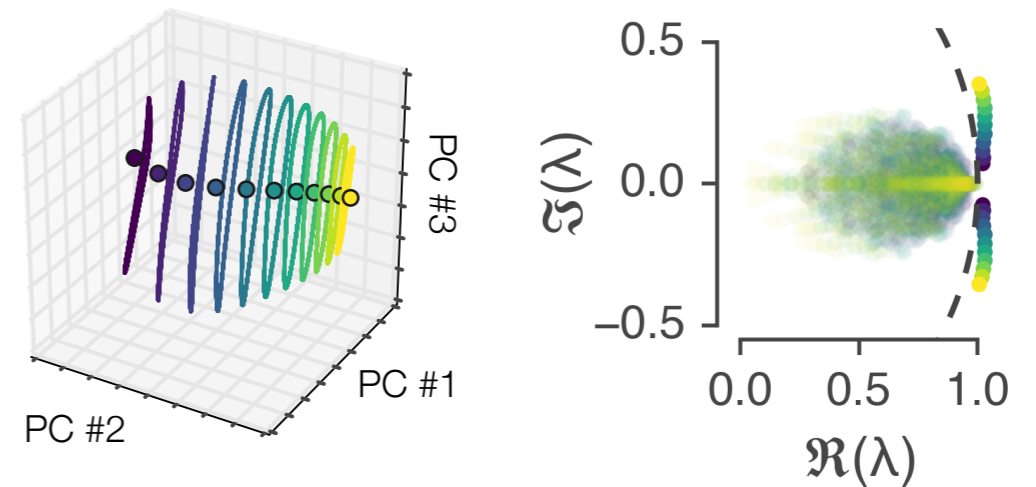


Evidence of both *universality* and *individuality*

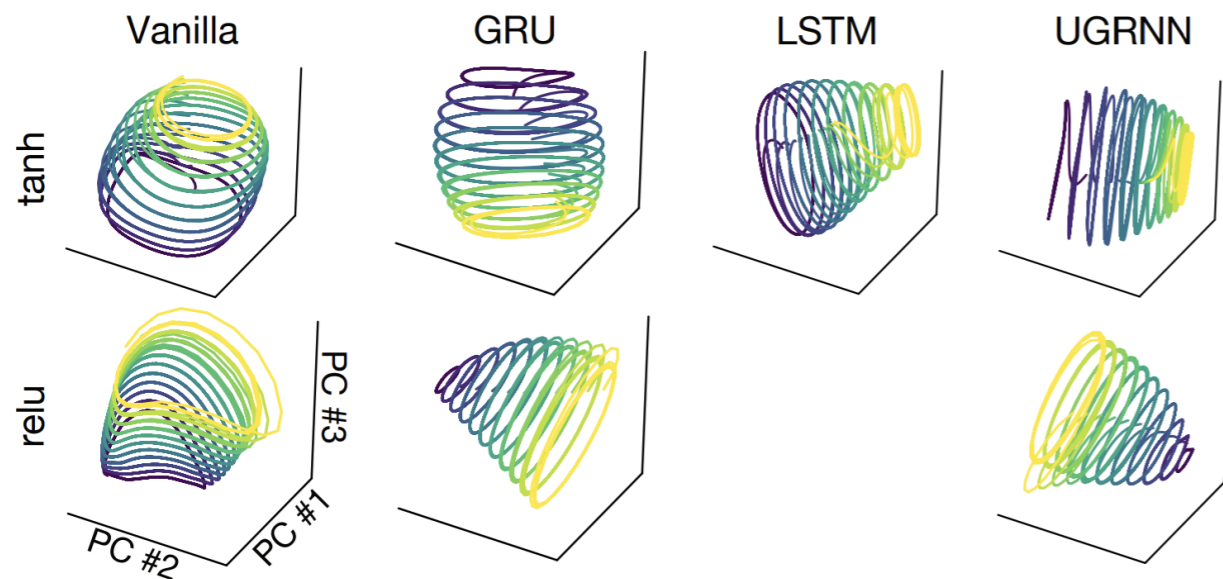
Pattern generation task



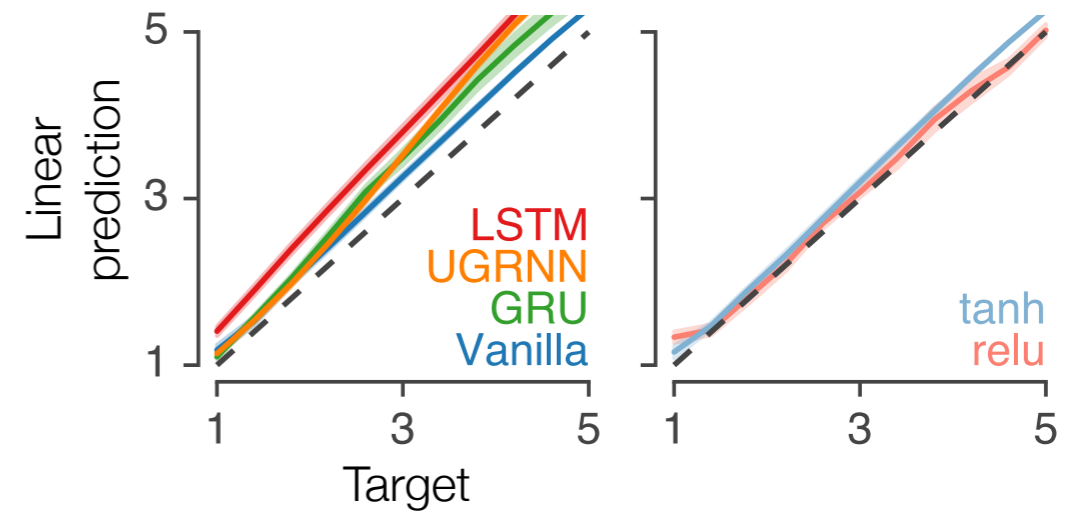
Analyzing trained networks



Network representations show individuality



but aspects of the computation are universal



Learn more at Poster #179

Extracting universal algorithmic principles from large populations of recurrent networks

1907.08549
arXiv

Niru Maheswaranathan*, Alex H. Williams*, Matthew D. Golub*, Surya Ganguli*, David Sussillo* // Google Brain & Stanford University

Motivation

Comparing brains and artificial neural networks (ANNs) [1-4]:

- Why should we trust comparisons between biological brains and artificial neural networks?
- Do networks trained on the same task find universal or unique solutions?
- Are these comparisons sensitive to modeling choices, such as neural network architectures?

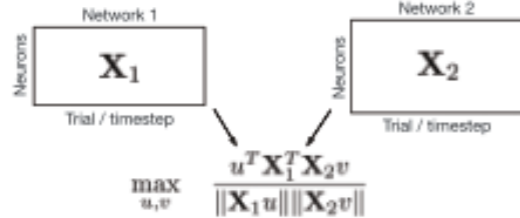


Methods

Trained ~1000s of networks with different architectures on a several tasks.

Vanilla RNN (Colins et al 2017) Update Gate RNN (Cho et al 2014) Gated Recurrent Unit (GRU) (Hochreiter & Schmidhuber 1997) Long-short term memory (LSTM)

1. Comparing representations: canonical correlation analysis (CCA) [5,6]



2. Comparing algorithms: fixed point analysis [7,8]

RNN defines a nonlinear map: $h^{t+1} = F(h^t, x^t)$

Computing (approximate) fixed points: $\arg\min_h \|F(h, 0) - h\|_2^2$

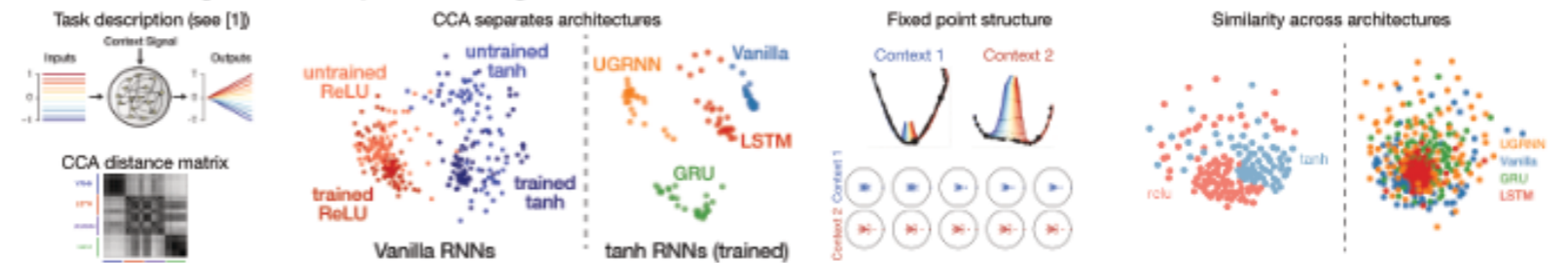
Linearizing around fixed points: $h^{t+1} - h^* \approx J(h^t - h^*)$

References

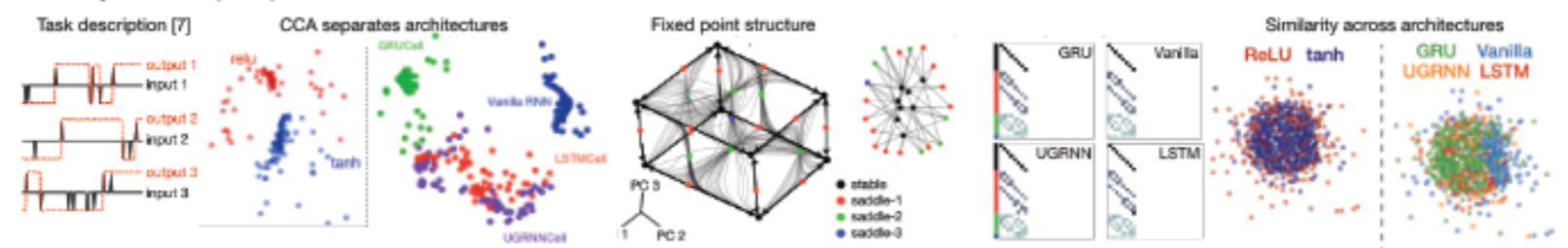
- [1] Mante, Sussillo et al. (2013) Nature. 503: 78-84.
- [2] Yamins et al. (2014). Proc Natl Acad Sci USA. 111(23): 8619.
- [3] Kriegeskorte (2015). Annual Rev of Vis Sci. 1:417.
- [4] Barak (2017). Curr Opin Neurobiol. 46:1.
- [5] Raghu et. al. (2017). Neural Information Processing Systems. arxiv.org/abs/1705.05806
- [6] Wang et. al. (2018). arxiv.org/abs/1812.02598
- [7] Sussillo & Barak. (2013). Neural computation 25.3: 626-649.
- [8] Golub & Sussillo. JOSS. (2018): 1003.

Results

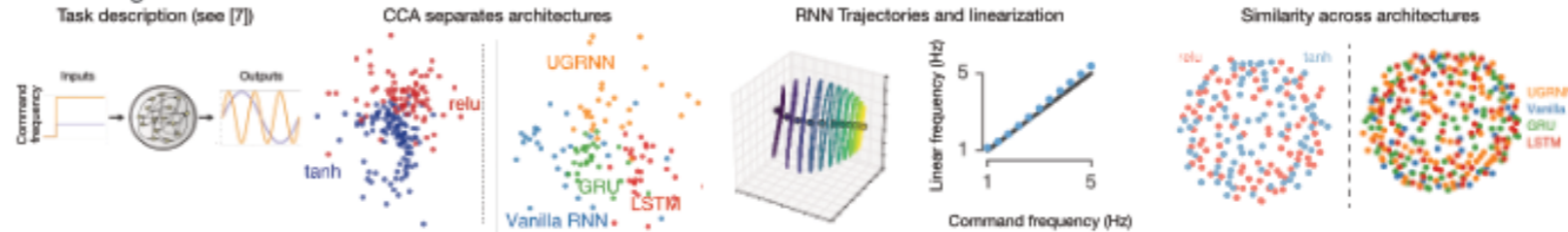
Decision making: context-dependent integration



Memory: N-bit flip flop



Pattern generation: sine waves



Conclusions

- Depending on how you quantify similarity, networks trained on the same task can look identical or different!
- Geometry of artificial neural representations are architecture dependent, but algorithms are more universal

Contact

nirum@google.com
@niru_m