Nonlinear Dynamics in Recurrent Neural Networks

Final Presentation | May 4, 2018

Joel Dapello
Elbert Gong
Kevin Stephen

Trajectory

Background
- What is a recurrent neural network?
- Why does it matter?

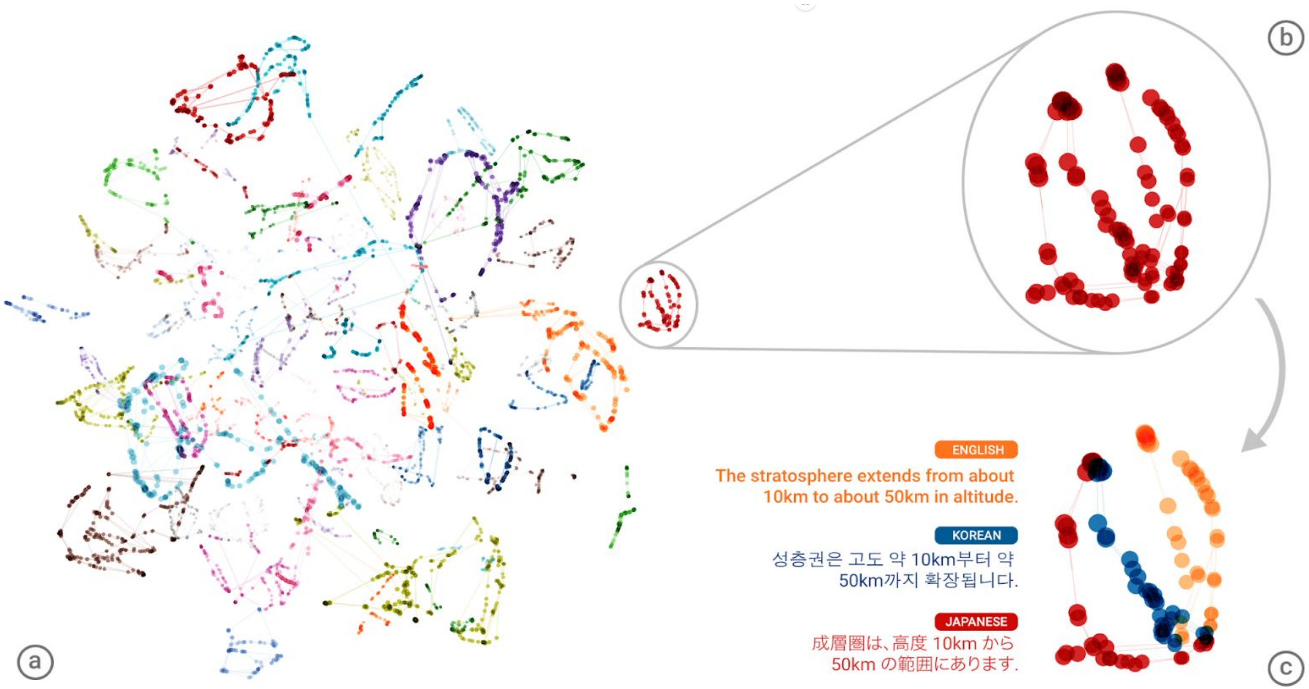
Motivation
- The RNN-Nonlinear Dynamics Connection
- How we get there: dimensionality reduction, optimization

Problem
- Three-Bit Flip Flop Problem
- Our replication of the system

Methods & Results
- Principal Component Analysis, NLPCA
- SINDy
- LSTMVis

Conclusions
- Language applications
- Findings, extensions, next steps

Background
Why are RNNs useful?

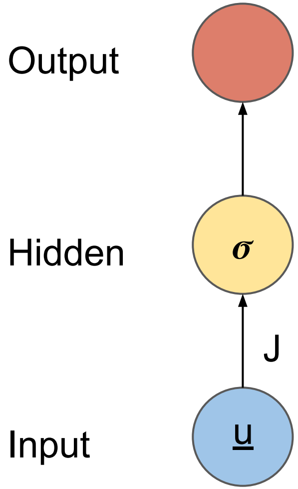


Shown here:
Google’s Multilingual Neural Machine Translation System.

Source: https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html

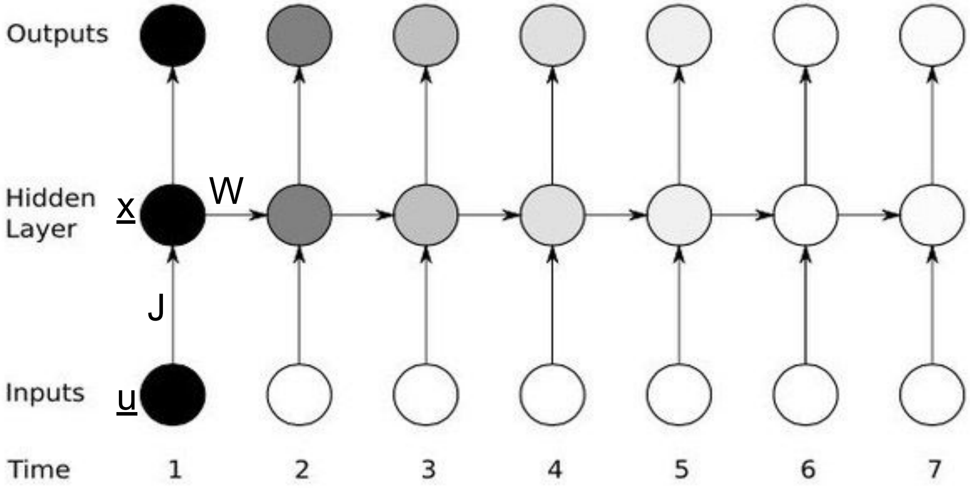
Recurrent Neural Network Structure

What is a recurrent neural network?



Basic Neural Network

$$\underline{x} = \tanh(J\underline{u})$$



Recurrent Neural Network

$$\underline{x}_{t+1} = \tanh(W\underline{x}_t + J\underline{u}_t)$$

Interpreting RNNs

Opening the black box…

by following hidden states

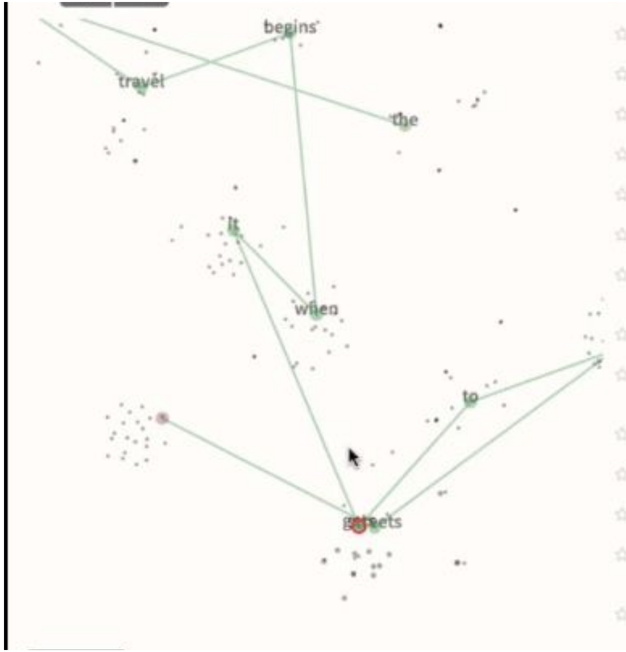
Command Prompt - python

```
 -2.1885
  0.8090
  7.7974
 -9.9696
 -2.4375
  4.6992
 -3.6019
 -6.3875
  8.3638
 -2.6013
  2.5896
  2.7510
  6.8370
  7.6563
 -0.6561
[torch.FloatTensor of size 100]
), ('output_layer.h2o.weight',

Columns 0 to 9
1.00000e-02 *
 -7.6379  5.6771 -7.8376  4.5422  7.3511 -2.8587  9.2204 -3.2472  3.7721 -6.2286
 -3.5290 -8.5306  6.7713 -8.3834  9.7036 -5.4028 -2.2963  1.5380  4.4052 -7.2753
 -8.7684  1.5421  6.5147 -1.8595 -9.0344 -6.5313 -1.9975 -0.5717 -7.9942  3.2839
  3.2077 -9.3507 -1.3442  0.0419 -9.3864 -5.1471 -8.5895  4.7379 -5.8487  6.5270
  6.2922 -7.1076 -1.6781 -3.7995 -5.5788 -0.9157 -8.2896 -2.8091 -9.5601 -9.2940
 -5.7974  5.5876  0.3699  6.0822  7.9787 -0.3705  7.5929  0.6846  9.7740  2.5377
 -1.9645 -9.0144 -2.2687 -8.2195 -4.0192  5.8962  0.8880  9.9428 -5.1578  4.9010
  7.9676 -6.2428  8.8580 -5.8865 -5.8712  7.7525 -8.5098 -0.3356 -0.4576 -8.7177

Columns 10 to 19
1.00000e-02 *
  0.3935 -6.9568 -4.3695 -0.0669 -4.4937  3.7733  2.9923 -9.2453 -9.9346  4.5047
 -4.8762  2.0678 -2.4409 -9.4717 -6.5144  5.4125  2.4286  9.9384 -0.7616 -2.0118
  8.3702 -0.1297  7.3290 -0.9227  6.5492 -7.4703 -5.1127 -7.2113 -3.2810  3.1991
  8.6627  9.9840 -3.7008  4.4609 -9.4855 -3.0064 -2.9788 -7.5221 -6.4145  3.0557
  6.2947 -9.2242  2.7828 -5.0139  8.7721 -2.0650 -7.4642  2.9307  9.9637  7.6365
  2.9322 -0.9936 -8.6592  4.6761 -0.7700 -3.1420  1.1847 -6.6662  7.6145 -3.6399
 -9.2400  8.0039  0.6627  1.4703 -2.8231 -8.3022 -6.7386  5.0310  0.9414  2.5686
  2.2691  9.9118 -2.5577  9.0104  9.5495  4.3467 -4.7656 -4.6596  5.0771  8.1915

Columns 20 to 29
1.00000e-02 *
  7.6629 -9.0370 -3.7781  1.4072  4.6685  3.0809 -5.8118 -5.6445  2.9414 -1.1333
 -2.1690  7.3585  1.9869 -2.5283  2.9617  8.1835 -8.6816  3.5277 -0.0416 -1.5931
```



RNNs and Nonlinear Dynamics: The Connection

The RNN IS the dynamical system!

Map equation:

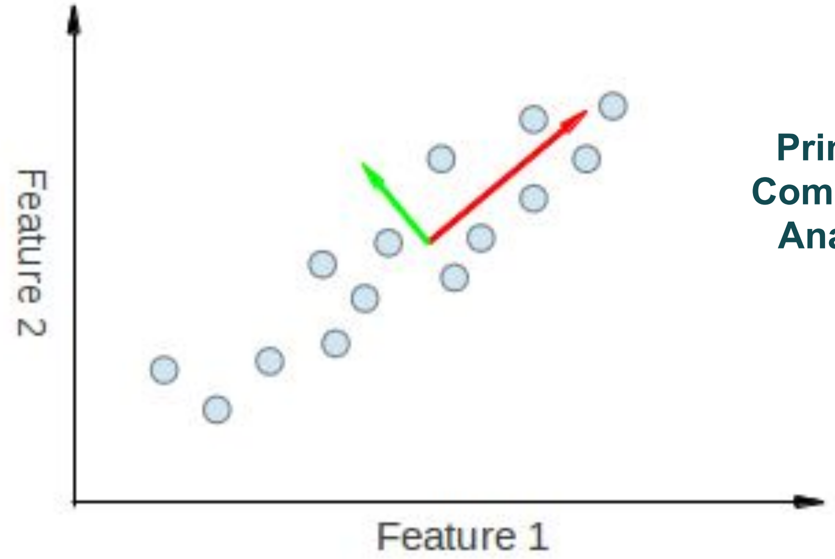
$$\underline{x}_{t+1} = \tanh(W\underline{x}_t + J\underline{u}_t)$$

Nonlinear Dynamical System:

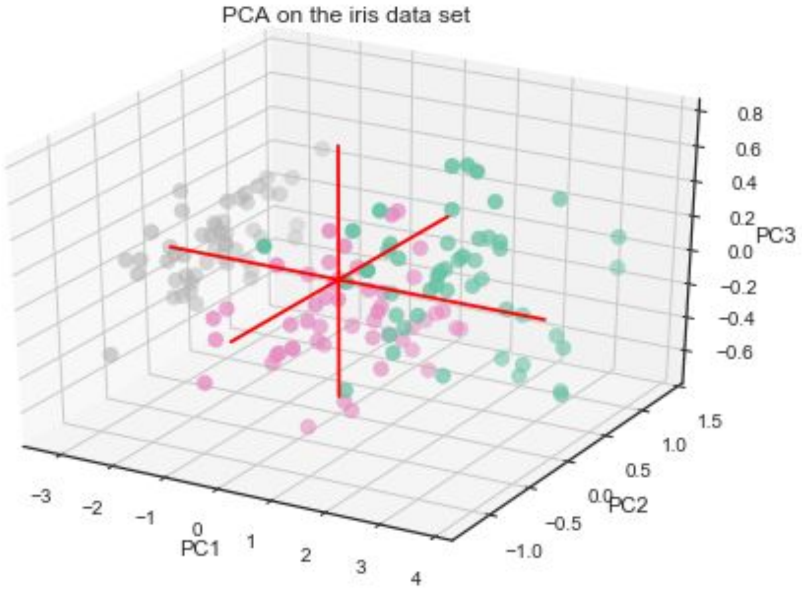
$$\underline{\dot{x}} = -\underline{x} + \tanh(W\underline{x} + J\underline{u})$$

How to get there: Dimensionality Reduction

How do we analyze 1000+ dimensional data with methods from our course?



Principal Component Analysis



Q Optimization for Finding FPs

How do we find FPs when we can't solve for them analytically?

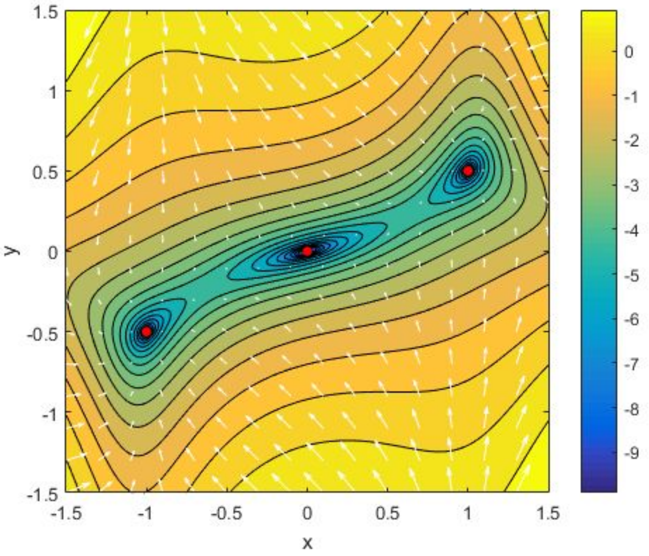
$$\dot{\underline{x}} = -\underline{x} + \tanh(W\underline{x} + J\underline{u})$$

$$q(\underline{x}) = \frac{1}{2}|\dot{\underline{x}}|^2$$

Find fixed points by minimizing q

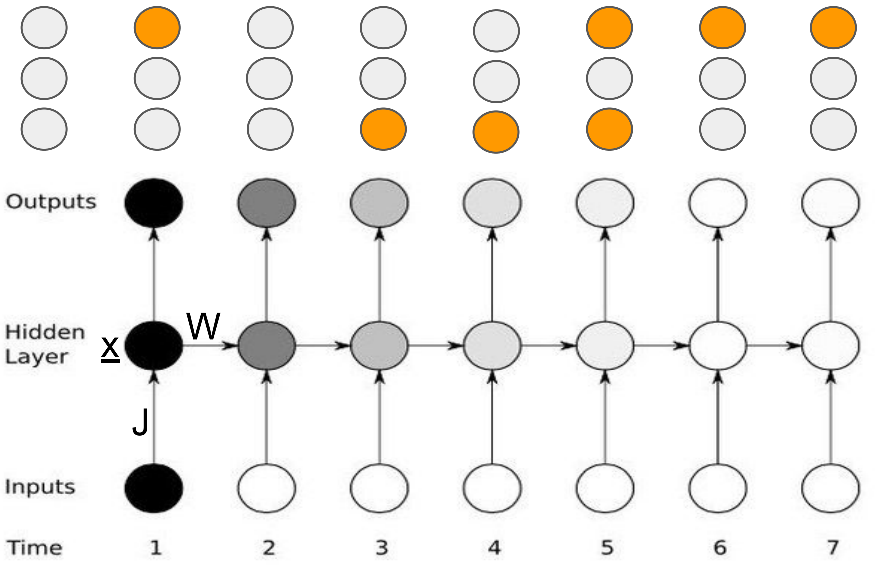
Example:

$$\dot{x} = (1 - x^2)y, \dot{y} = x/2 - y$$



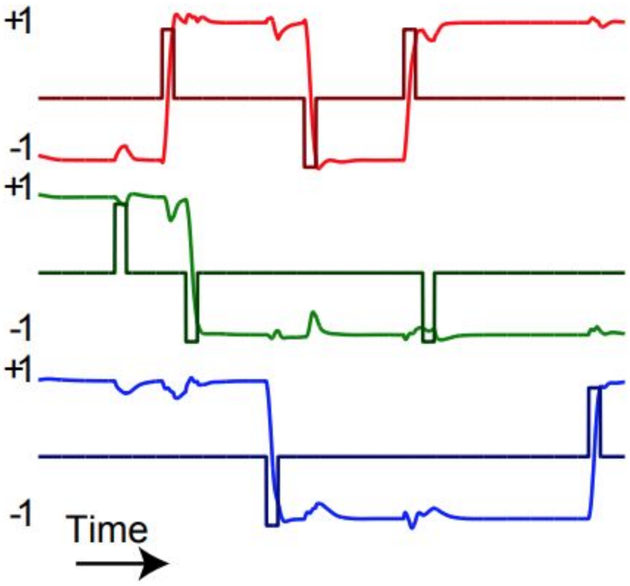
Three-Bit Flip-Flop Problem

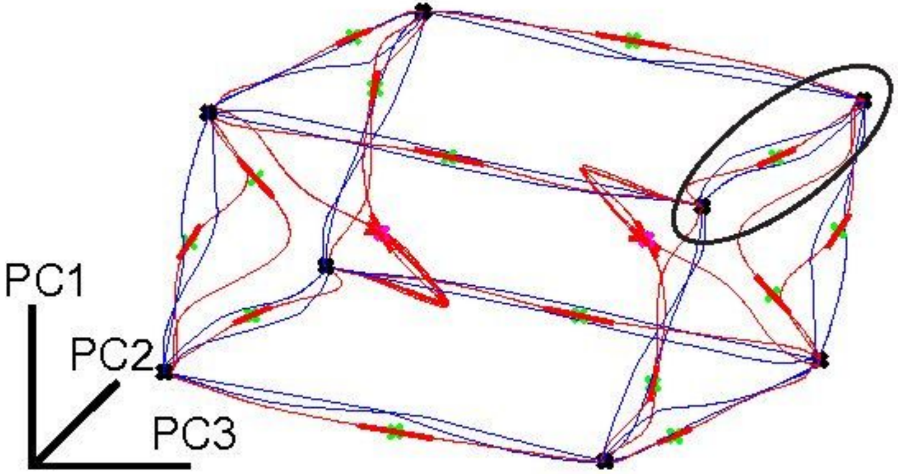
What is the three-bit flip-flop problem?



u	1	2	3	4	5	6	7
	1	-1	0	0	1	0	0
	0	0	0	-1	0	0	0
	0	0	1	0	0	-1	0

3-Bit Flip-Flop Analysis - Sussillo & Barak

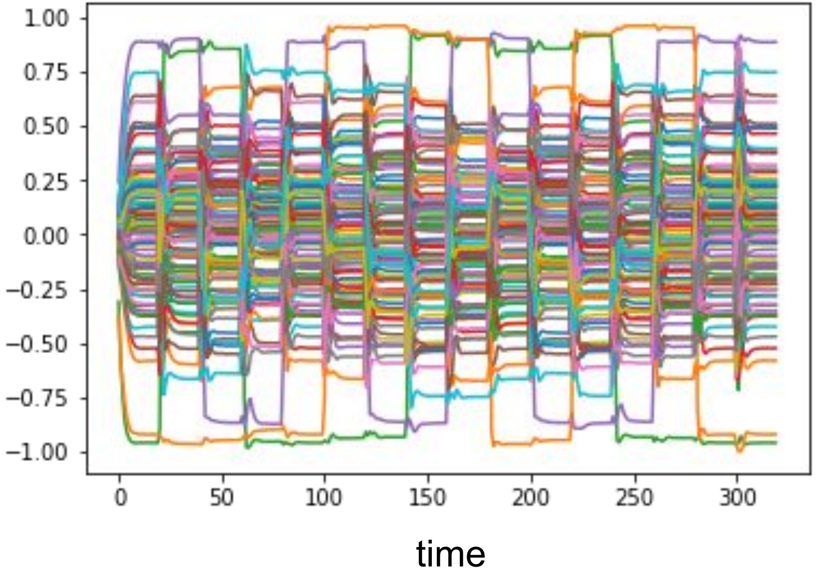




Sussillo, David and Barak, Omri. Opening the Black Box: low-dimensional dynamics in high-dimensional systems. Neural Comput. 2013 Mar; 25(3):626-49. doi: 10.1162/NECO_a_00409. Epub 2012 Dec 28.

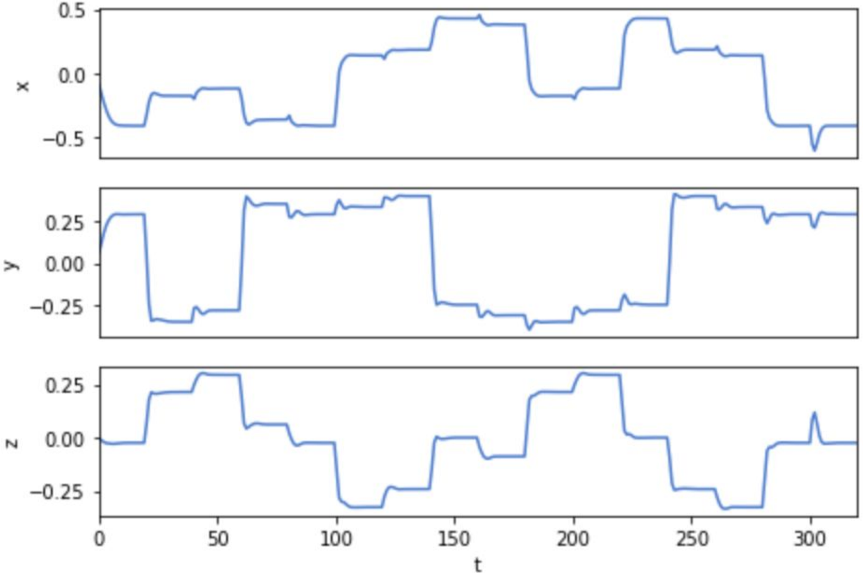
Our Replication

Trajectories of x, the hidden states

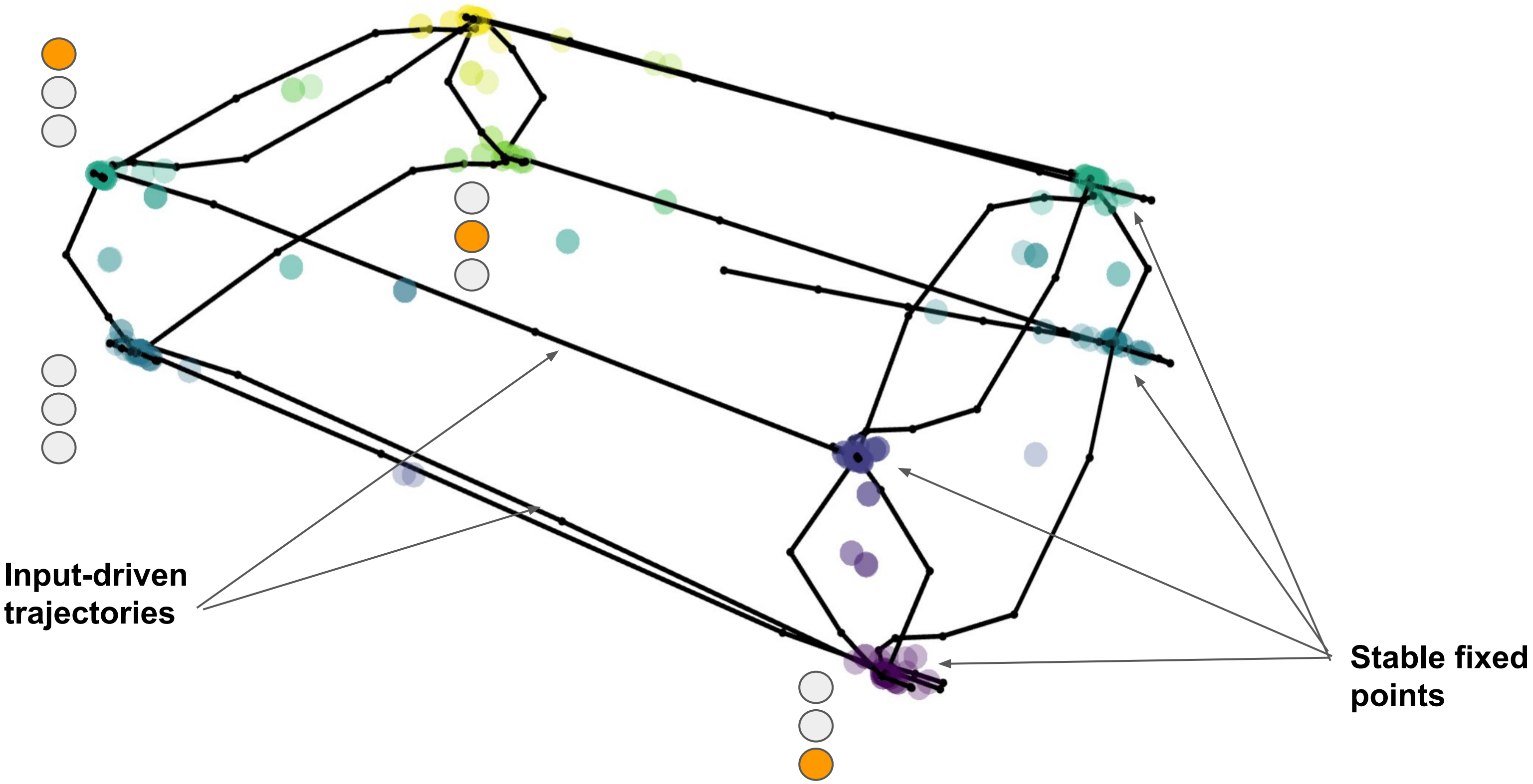


PCA

Trajectories of x projected onto first three principal components



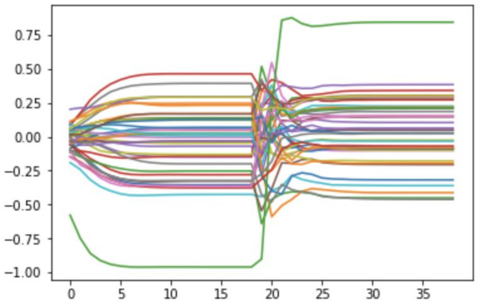
Our Replication

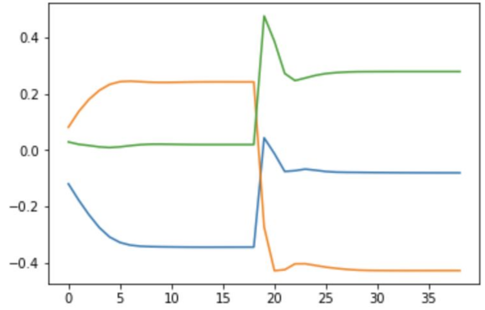


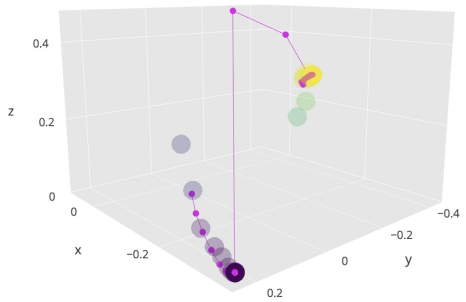
Our Replication

The FPs act like memory reservoirs.

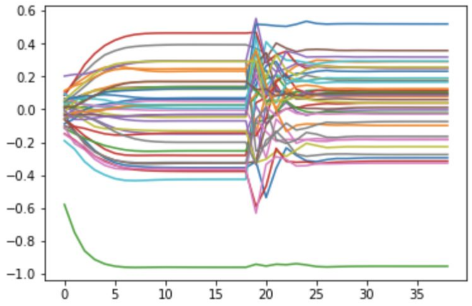
At time 20: Turn on traffic light #0

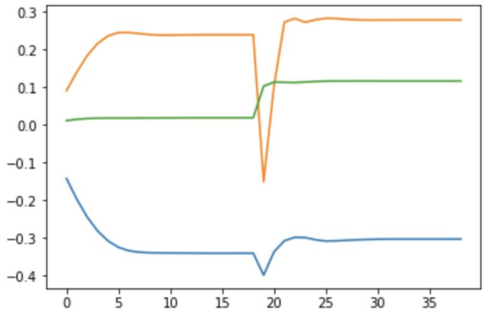


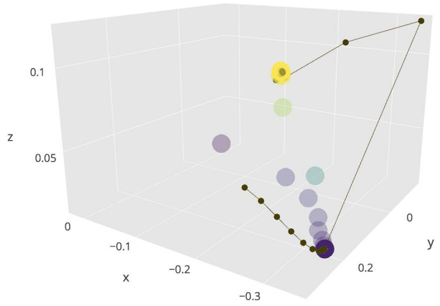




At time 20: Turn on traffic light #1

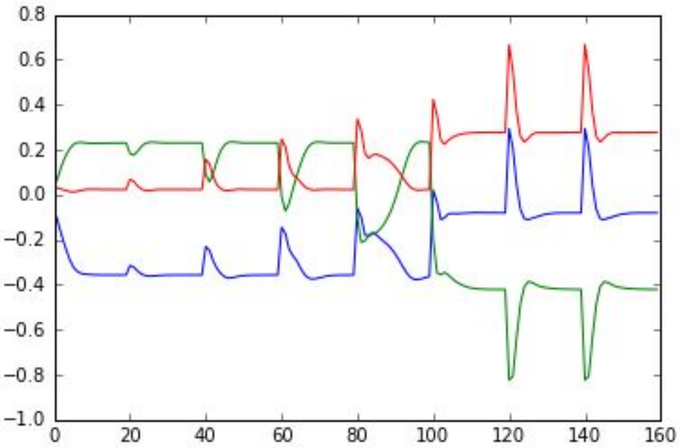


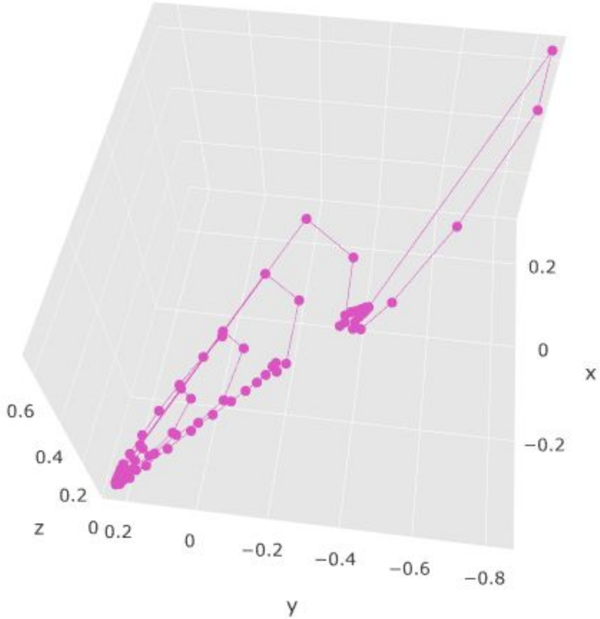




Our Replication

The FPs act like memory reservoirs.

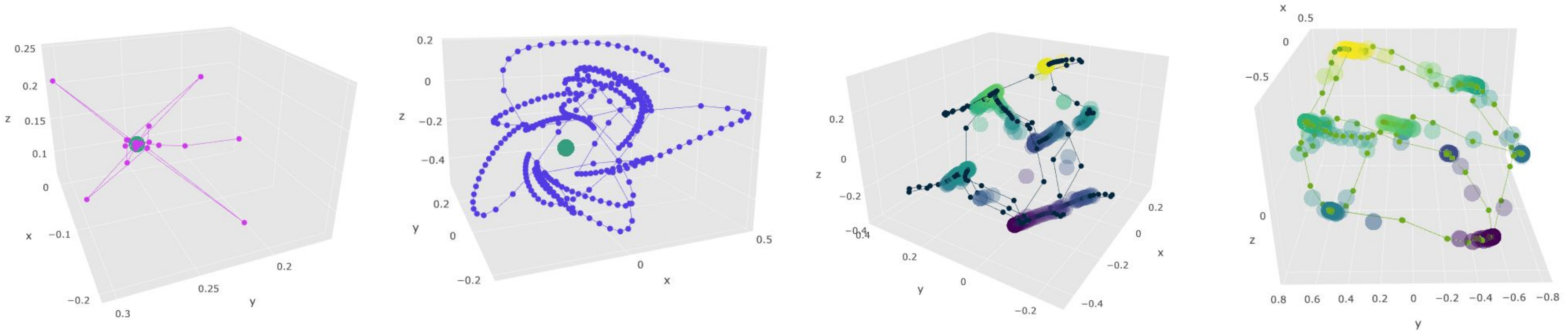




Homoclinic and heteroclinic orbits, driven by inputs.

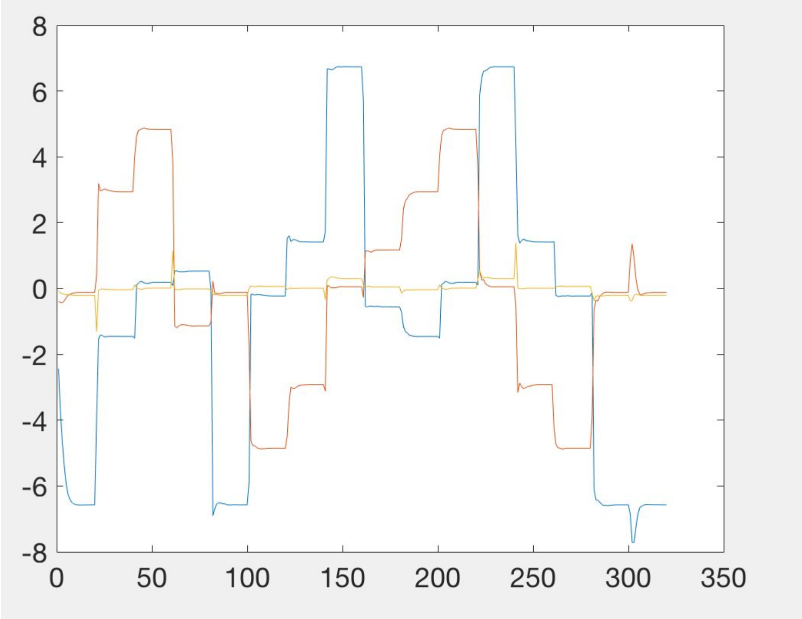
Evolution of System through Training

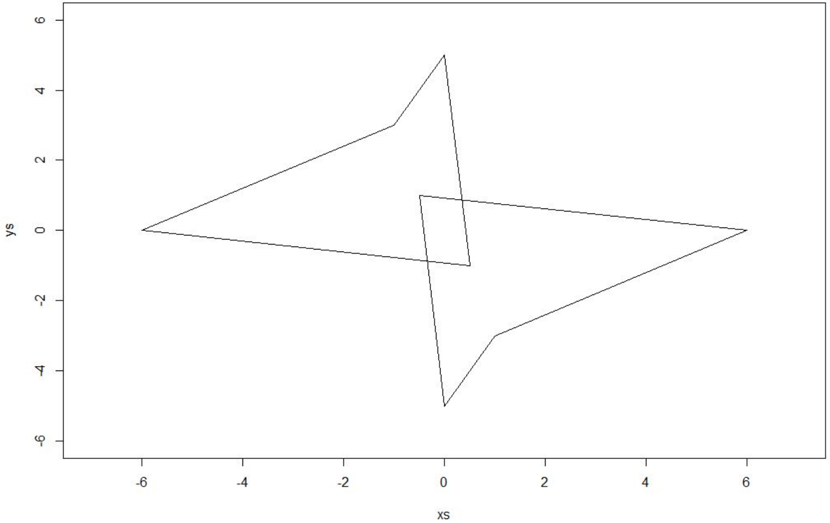
Bifurcations: gradual emergence of structure from scratch.



Nonlinear PCA

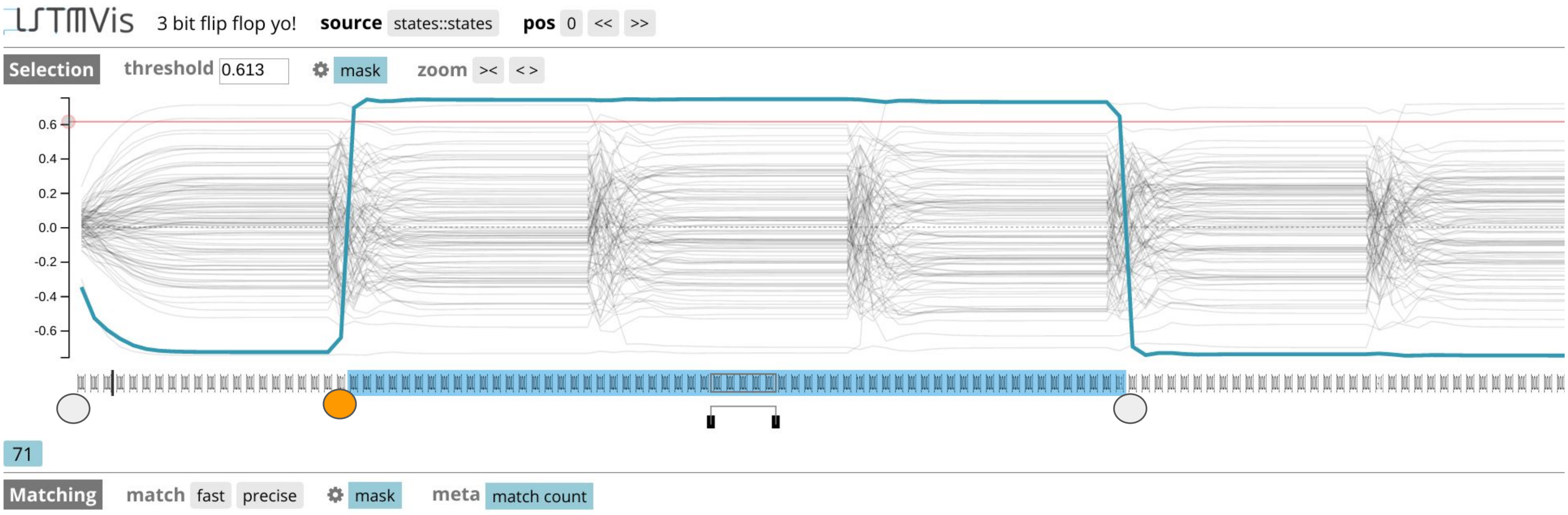
Nonlinear curves can store most of the data's signal more efficiently than PCA





LSTM Vis

Drilling down to the individual neuron level, sans PCA



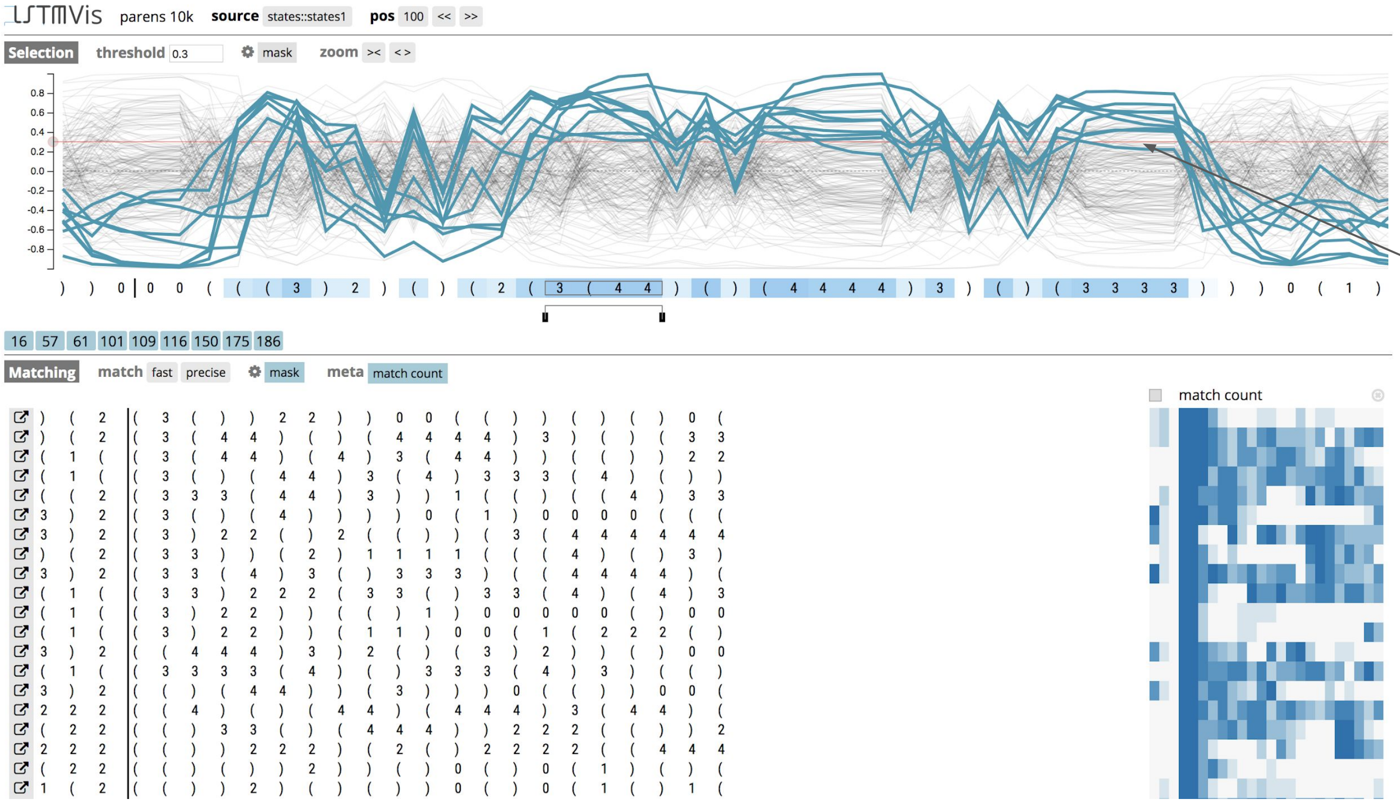
lstm.seas.harvard.edu

Simple Language-Like System: Parentheses Dataset

The Parentheses Language

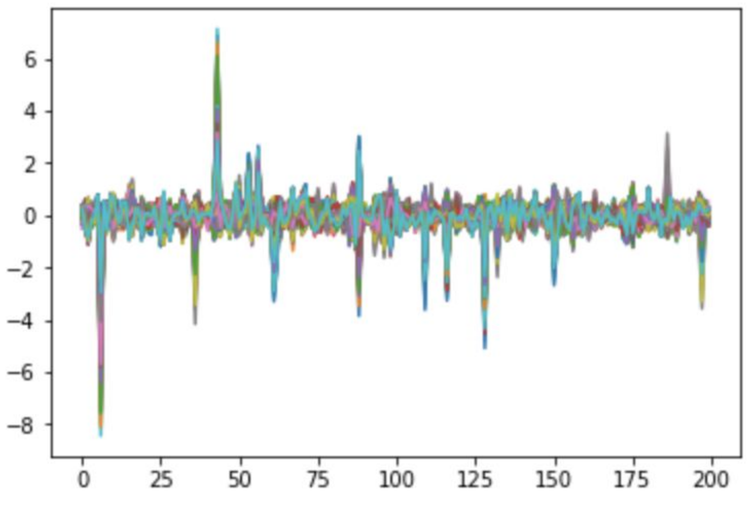
( )(((3) 2)) ((2 ()))

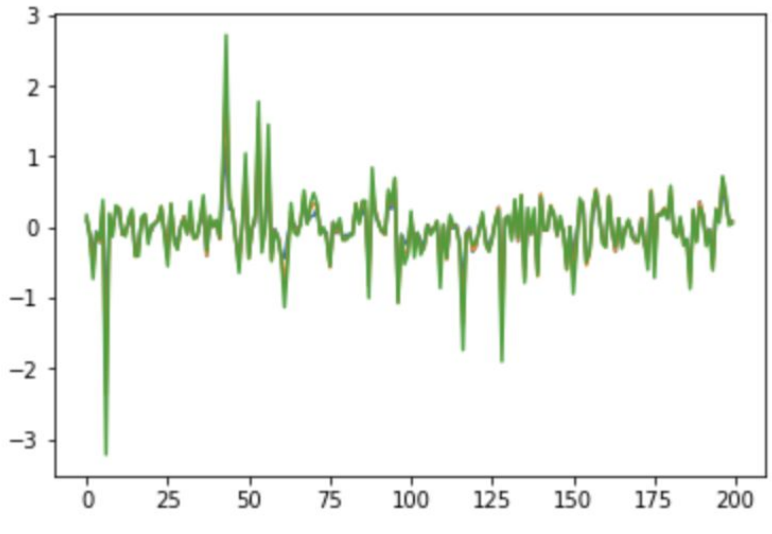
Step 1: LSTM Vis



Like the three-bit system, groups of neurons move together and memorize changes in the data.

Step 2: Hidden State Dimensionality Reduction

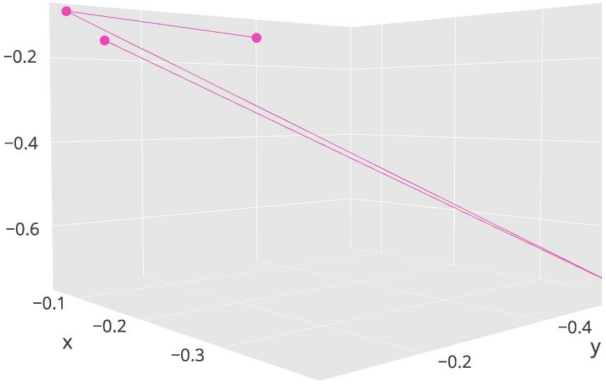




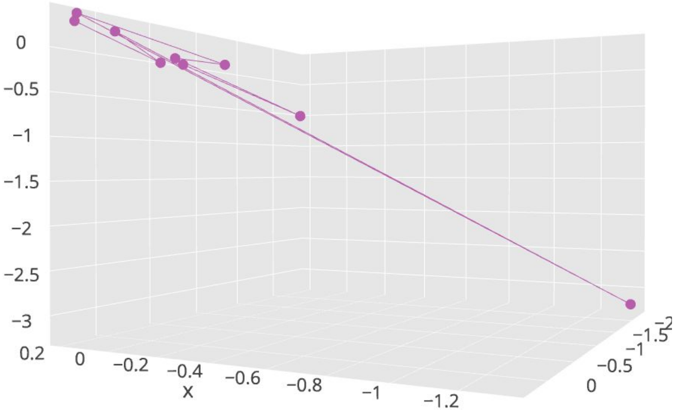
Takeaways:

1. This data can be stored in just one dimension!
2. The explanation of the majority of the variance in the data using PCA suggests that a dynamical approach will yield accurate results.

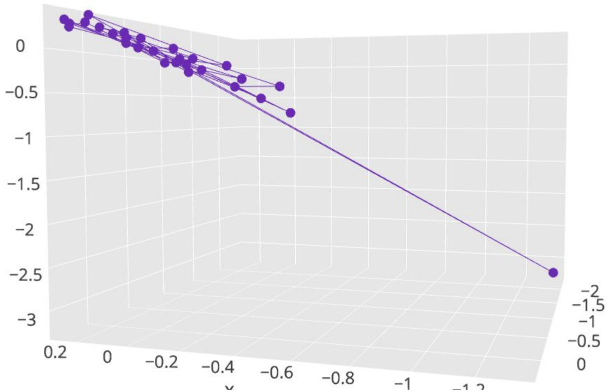
Step 3: Trajectory Analysis



t = 5

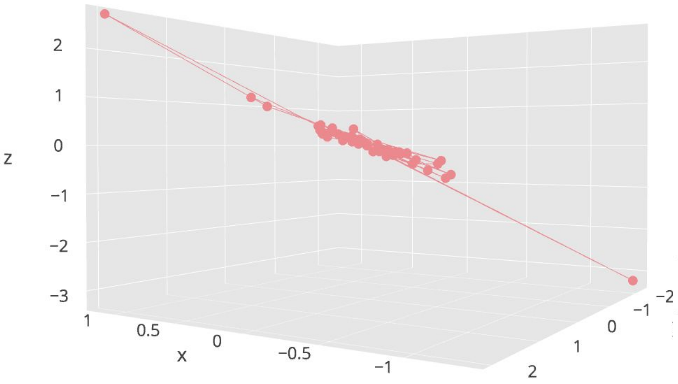


t = 10

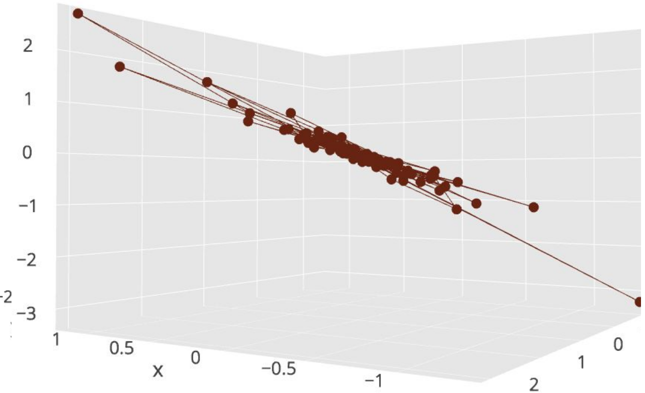


t = 20

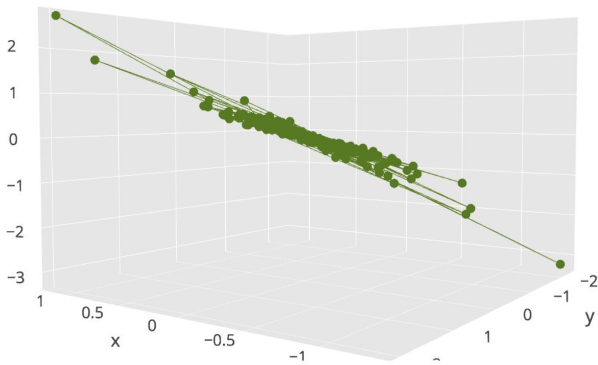
Step 3: Trajectory Analysis



t = 50



t = 100



t = 200

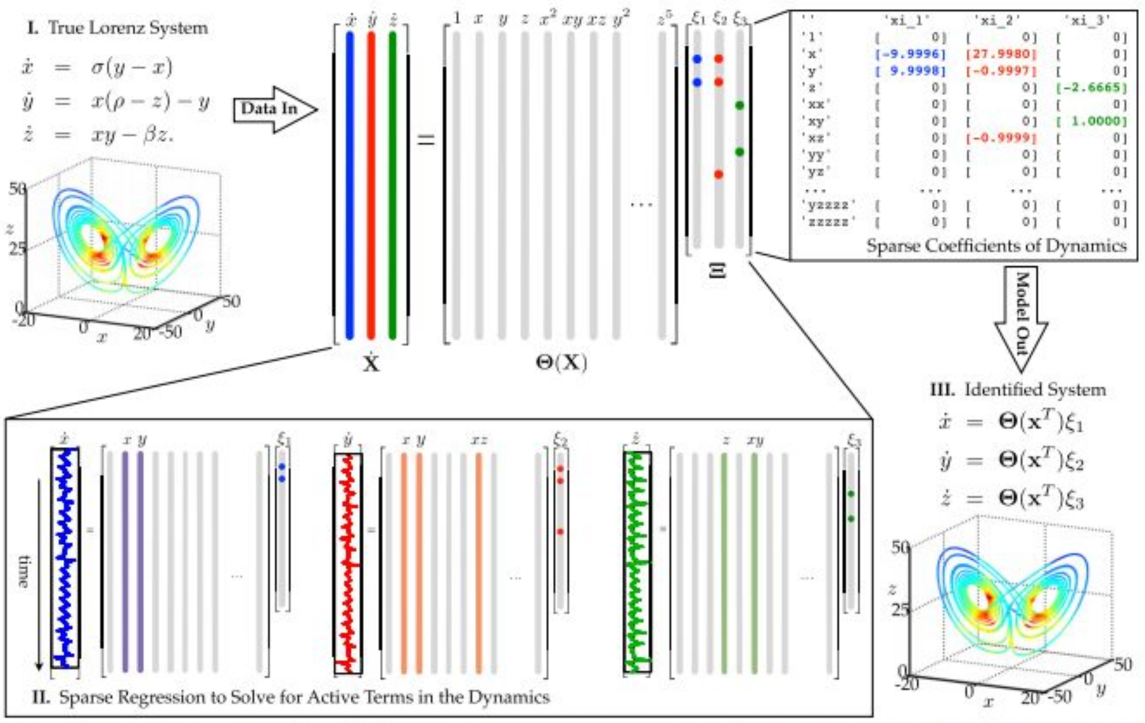
Takeaways:

1. Trajectories of the hidden states seem to converge to a stable attractor at the origin.
2. It's more difficult to see a clean orbit pattern as we did for the three-bit flip flop.
3. With some work, we have extended our approach to a very basic language set.

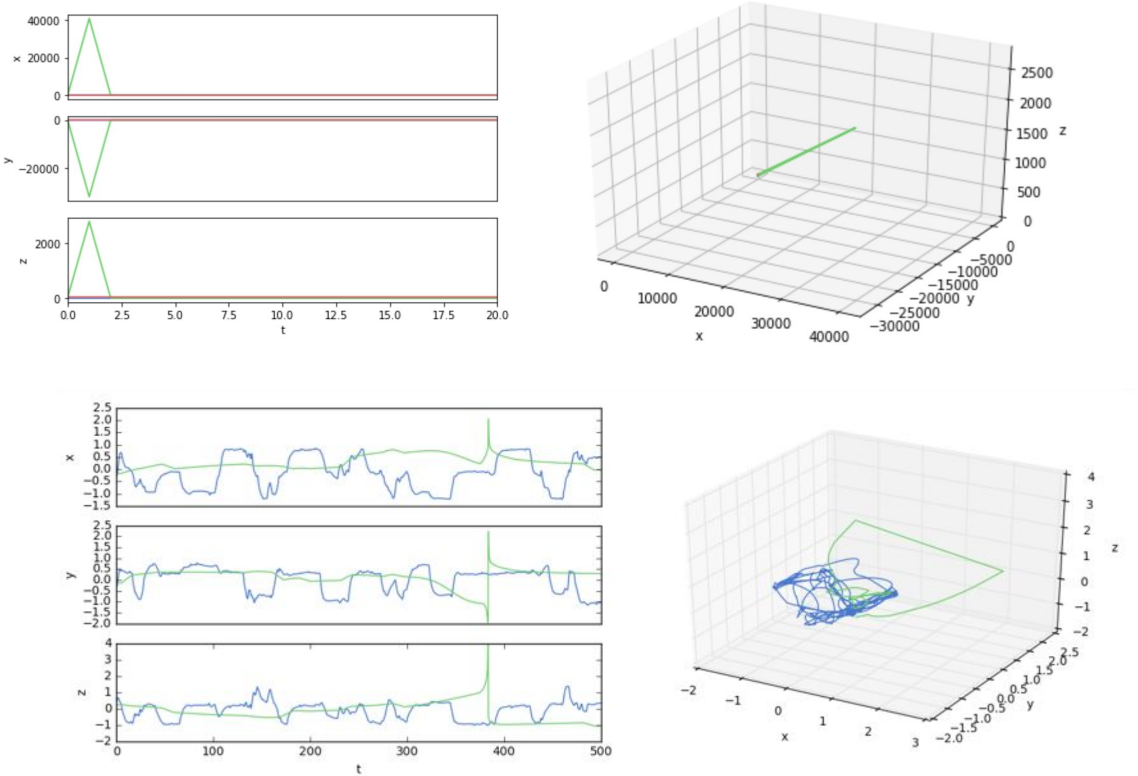
Future Explorations

- Beyond fixed points: attracting submanifolds
- Diagnosing chaos in RNNs
- Applying techniques to a full language dataset (more data, bigger nets)
    - Limit cycles = repetitive sentences?
- Other dimensionality reduction techniques (LFADS)
- Learning a LDS over word vectors

Sparse Identification of NLDS



Our Attempts So Far



Thank You!

Barak, Omri et al. From Fixed Points to Chaos. In Prog Neurobiol. 2013 Apr;103:214-22. doi: 10.1016/j.pneurobio.2013.02.002. Epub 2013 Feb 21.

Brunton, Steven L., Proctor, Joshua L., Kutz, J. Nathan. *Sparse identification of nonlinear dynamics.* Proceedings of the National Academy of Sciences Apr 2016, 113 (15) 3932-3937; DOI:10.1073/pnas.1517384113 (video)

Elman, Jeffrey. "Language as a Dynamical System." in Robert F. Port & T. van Gelder (Eds.) *Mind as Motion: Explorations in the Dynamics of Cognition.* Cambridge, MA: MIT Press, 1995. Pp. 195-223.

Gao, Peiran, Trautmann, Eric, Yu, Byron, Santhanam, Gopal, Ryu, Stephen, Shenoy, Krishna Shenoy, and Ganguli, Surya. *A Theory of Multineuronal Dimensionality, Dynamics and Measurement*. *A Theory of Multineuronal Dimensionality, Dynamics and Measurement*. Stanford University, 2017.

Jang, Eric. "Recurrent Neural Networks; A Dynamical Systems Perspective."

Laurent, Thomas, and James Von Brecht. *A Recurrent Neural Network without Chaos.* Department of Mathematics, Loyola Marymount University.

Mastrogiuseppe, Francesca, and Srdjan Ostojic. *Linking Connectivity, Dynamics and Computations in Recurrent Neural Networks*. *Linking Connectivity, Dynamics and Computations in Recurrent Neural Networks*, arxiv.org/pdf/1711.09672.pdf.

Olah, Chris. "Understanding LSTM Networks." *Github.io*, 27 Aug. 2015, colah.github.io.

Rush, Alexander; Strobelt, Hendrik; Gehrmann, Sebastian; Huber, Bernd; Pfister, Hanspeter. LSTMvis. Lstm.seas.harvard.edu.

Sussillo, David and Barak, Omri. Opening the Black Box: low-dimensional dynamics in high-dimensional systems. Neural Comput. 2013 Mar; 25(3):626-49. doi: 10.1162/NECO_a_00409. Epub 2012 Dec 28.

Sussillo, David; Jozefowicz, Rafal; Abbott, L. F.; Pandarinath Chethan. *LFADS - Latent Factor Analysis via Dynamical Systems*. arXiv:1608.06315. (video)