The Geometry of Integration in Text Classification RNNs

Kyle Aitken*
Univ. of Washington

Vinay Ramasesh*
Blueshift, Alphabet

Ankush Garg
Google

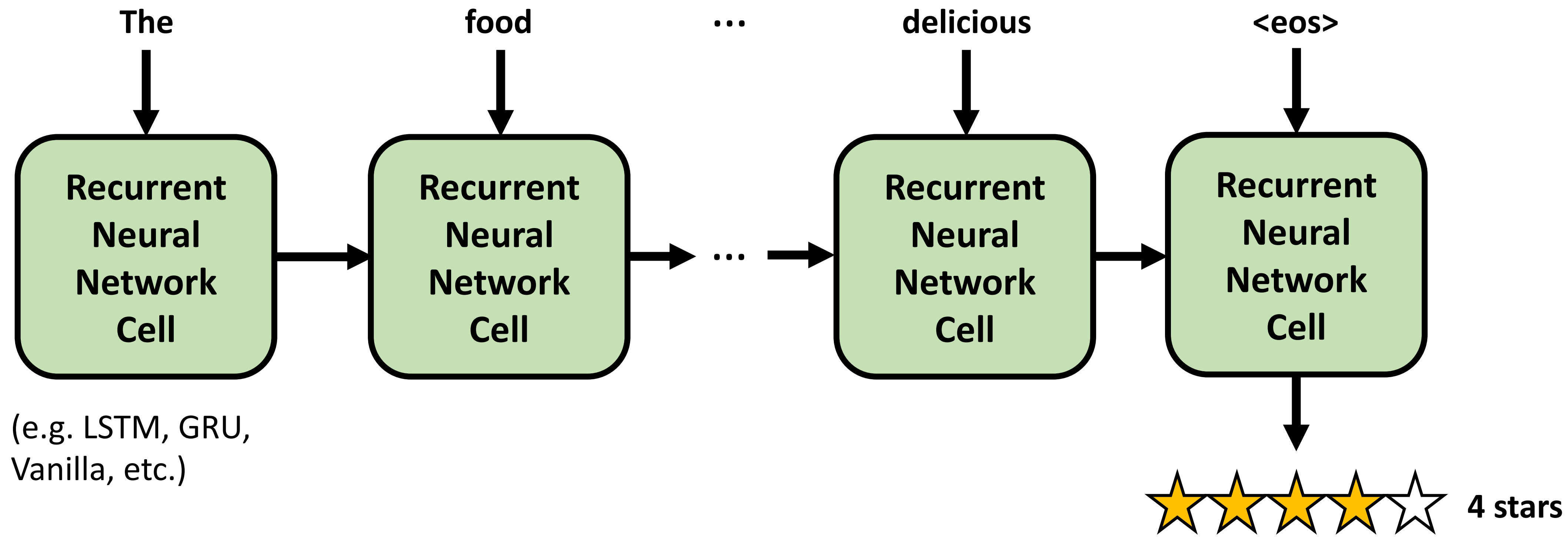
Yuan Cao
Google

David Sussillo
Google (now at Facebook)

Niru Maheswaranathan
Google

Motivation

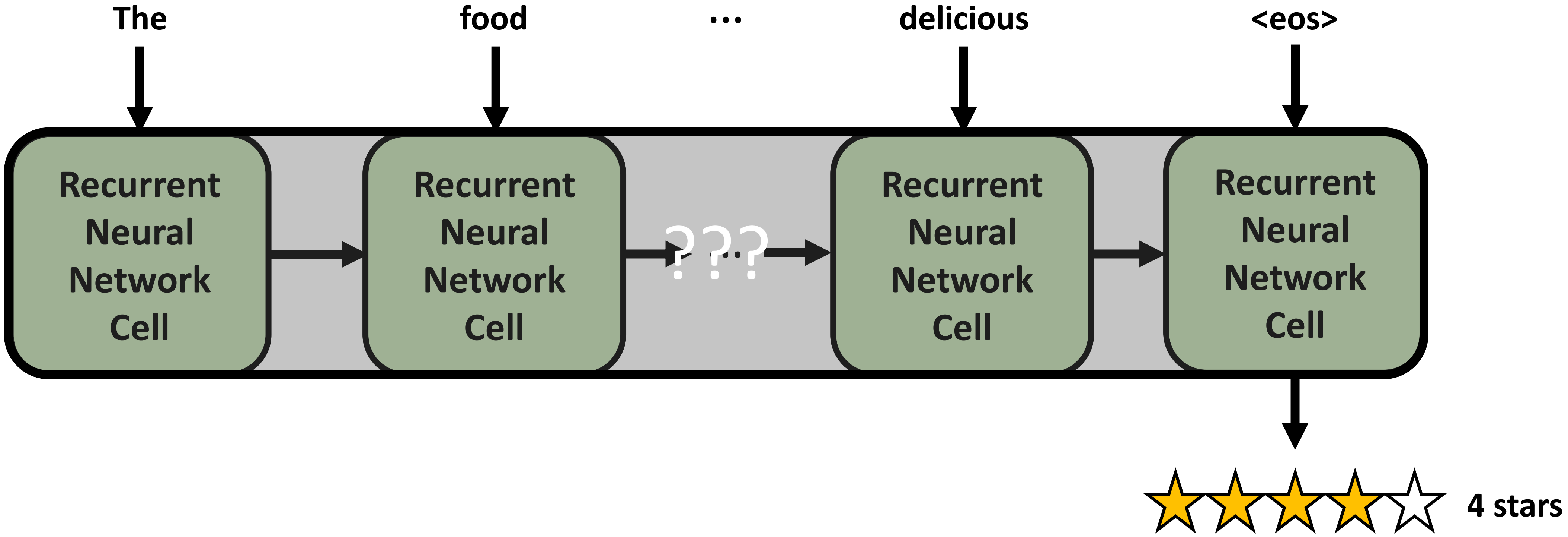
Sequential Input:
“The food was warm and delicious.”


The
food
...
delicious
<eos>
Recurrent Neural Network Cell
Recurrent Neural Network Cell
...
Recurrent Neural Network Cell
Recurrent Neural Network Cell
4 stars

(e.g. LSTM, GRU, Vanilla, etc.)

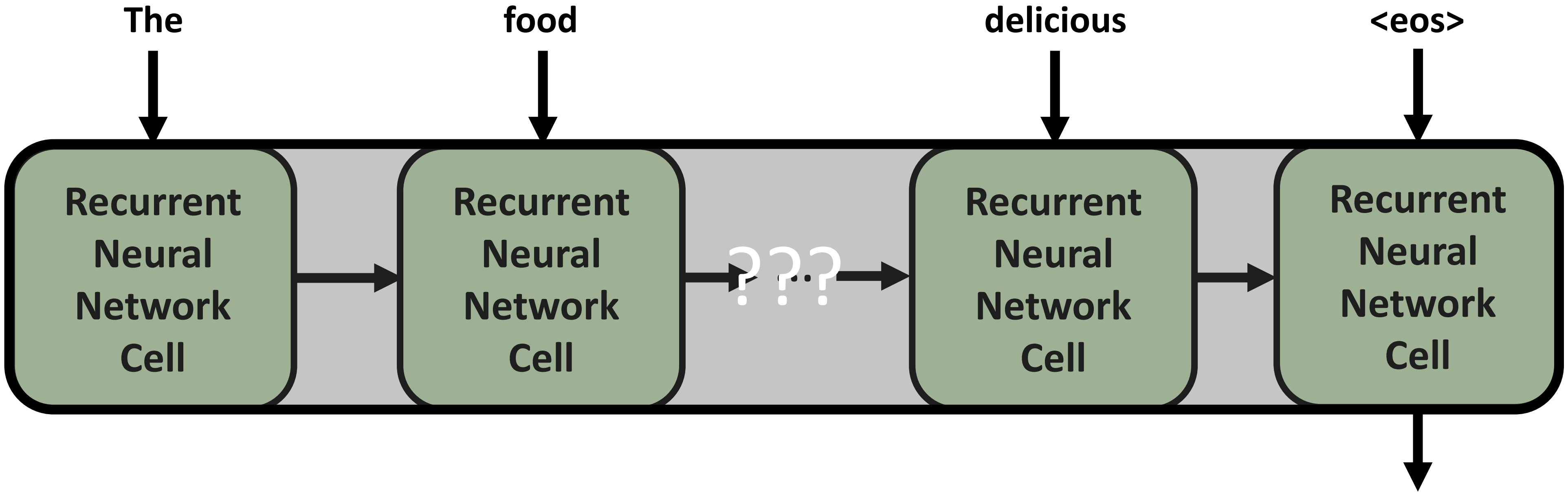
Motivation

Sequential Input:
“The food was warm and delicious.”


The
food
…
delicious
<eos>
Recurrent Neural Network Cell
Recurrent Neural Network Cell
???
…
Recurrent Neural Network Cell
Recurrent Neural Network Cell
★★★★☆
4 stars

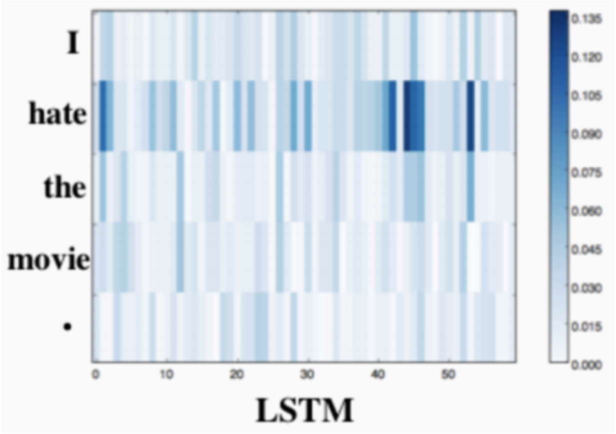
Motivation

Sequential Input:
“The food was warm and delicious.”

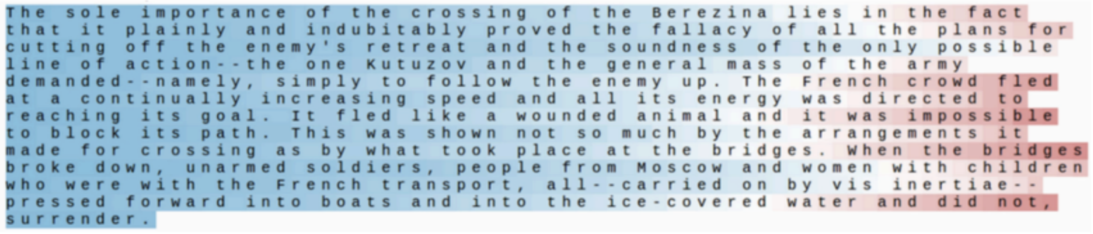

The
food
delicious
<eos>
Recurrent Neural Network Cell
Recurrent Neural Network Cell
???...
Recurrent Neural Network Cell
Recurrent Neural Network Cell

★★★★☆ **4 stars**

Saliency maps
(Li et al., 2015)

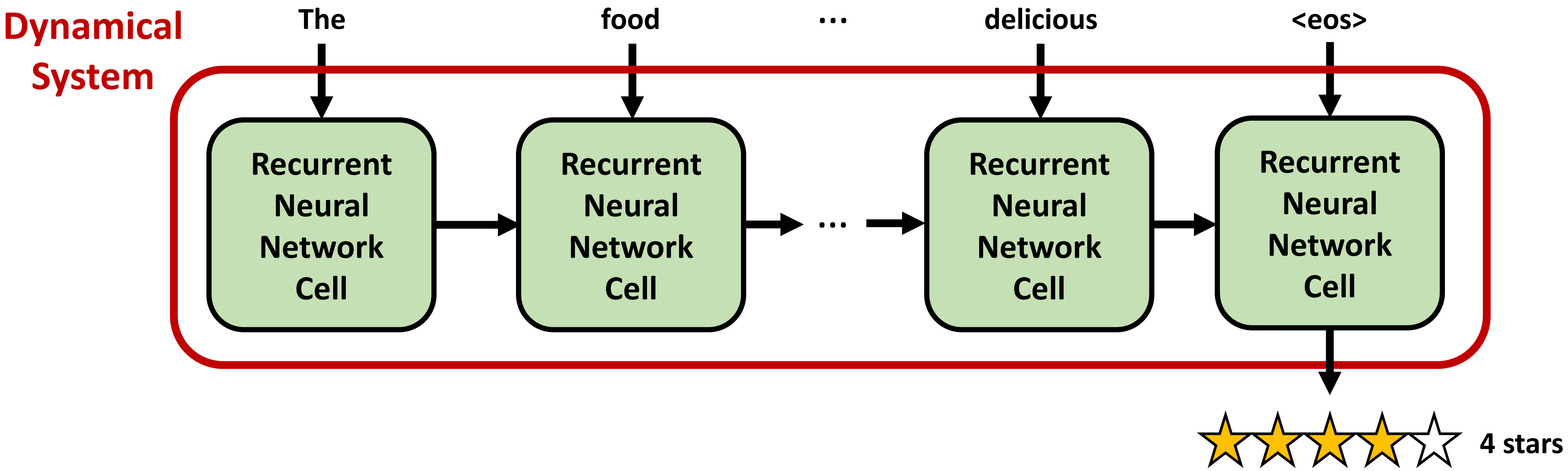


Inspect individual RNN units
(Karpathy*, Johnson* & Li 2015)


The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.

Motivation

Sequential Input:
“The food was warm and delicious.”



Dynamical System

The

food

…

delicious

<eos>

Recurrent Neural Network Cell

Recurrent Neural Network Cell

…

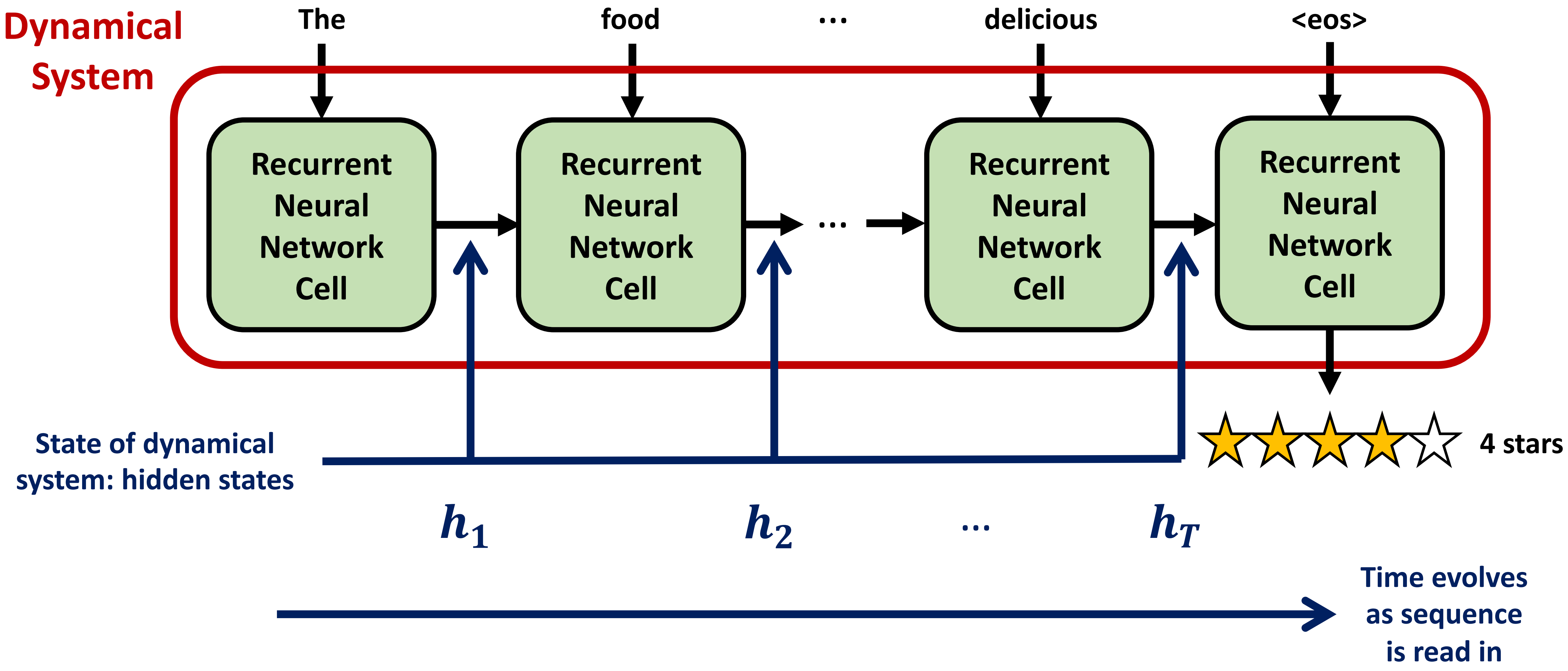
Recurrent Neural Network Cell

Recurrent Neural Network Cell

★★★★☆ 4 stars

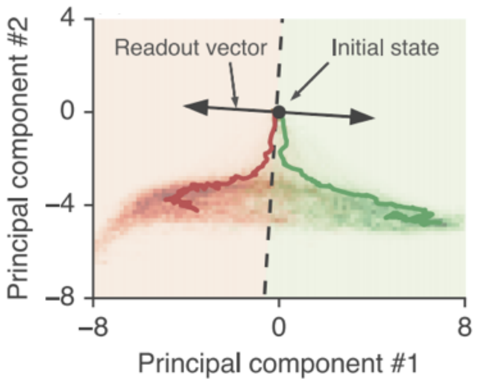
Motivation

Sequential Input:
“The food was warm and delicious.”


Dynamical System
The
food
…
delicious
<eos>
Recurrent Neural Network Cell
Recurrent Neural Network Cell
…
Recurrent Neural Network Cell
Recurrent Neural Network Cell
4 stars
State of dynamical system: hidden states
h_1
h_2
…
h_T
Time evolves as sequence is read in

Motivation (cont.)

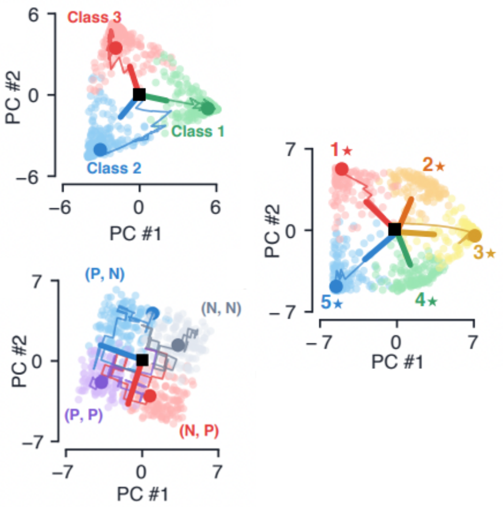
Broad Goal: Understand how recurrent neural networks perform various tasks by using tools from the study of dynamical systems (e.g., linearization, stability analysis).



Maheswaranathan, Williams et al., NeurIPS 2019
Maheswaranathan & Sussillo, ICML 2020

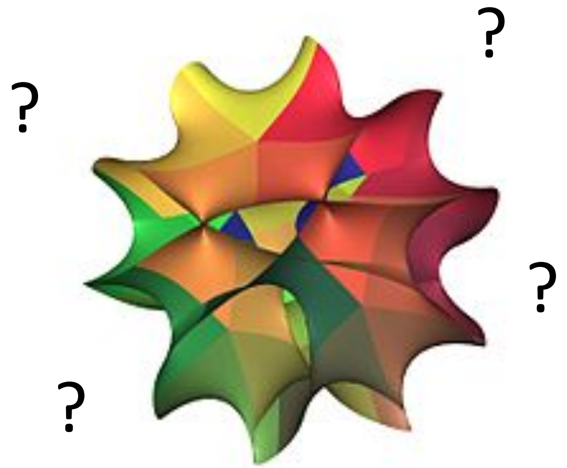
Prior work:

Binary sentiment classification



Our work:

Text classification tasks
Categorical, ordered, multi-labeled



Wikipedia

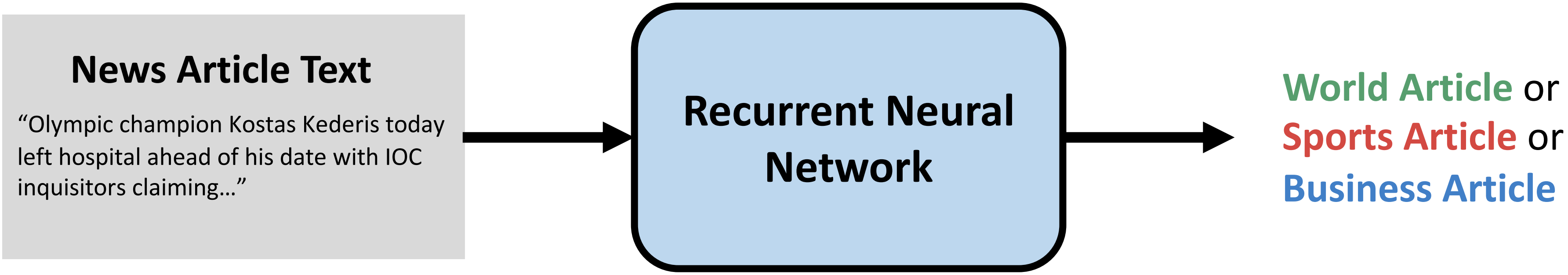
Future work:

Language modeling, translation, natural-language inference, etc.

Conclusion: Text-classification RNNs learn low-dimensional, interpretable dynamical systems.

3-Class Categorical Classification

Example: 3-Class AG News



News Article Text

"Olympic champion Kostas Kederis today left hospital ahead of his date with IOC inquisitors claiming…"

Recurrent Neural Network

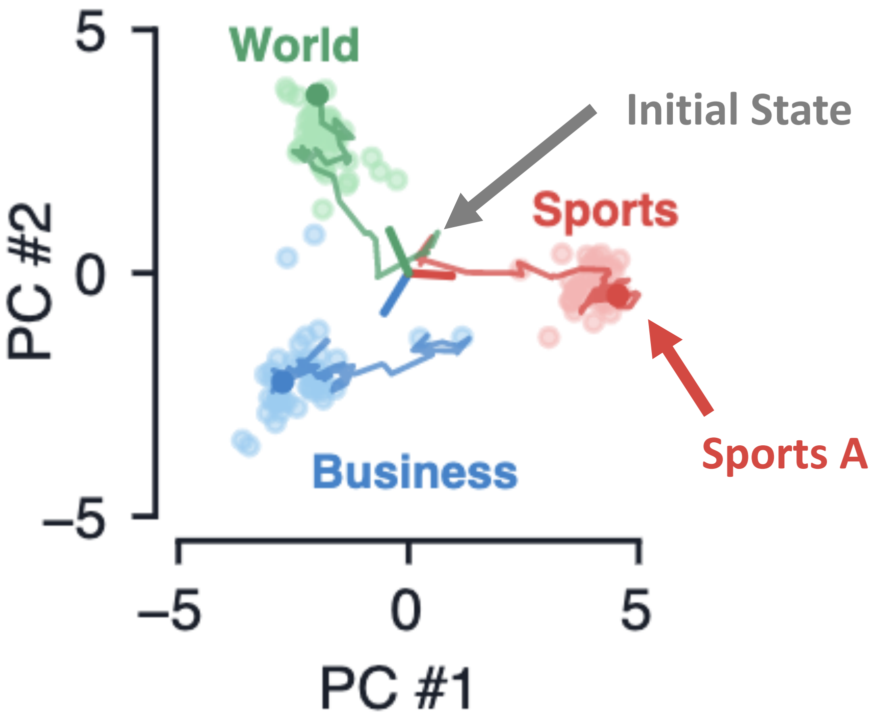
World Article or

Sports Article or

Business Article

3-Class Categorical Classification (cont.)

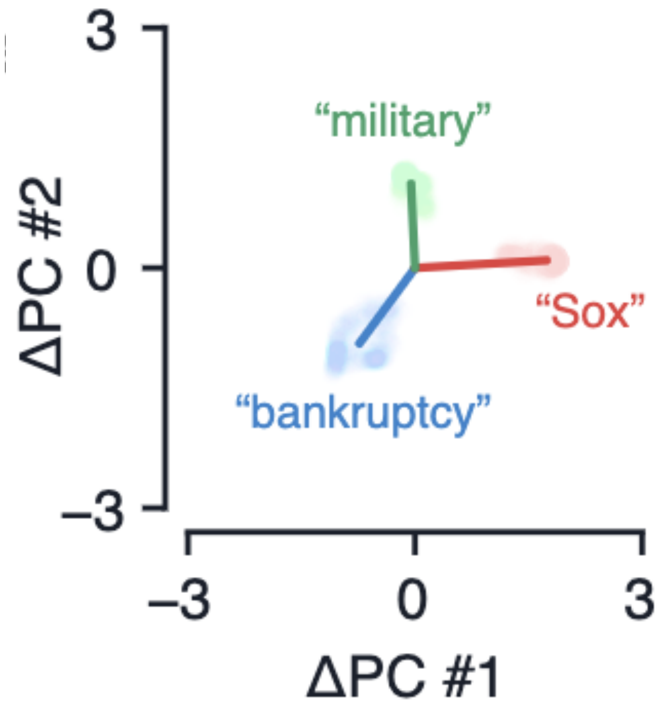
Final hidden states:
(with example trajectories and readouts)



Initial State

Sports Article trajectory

Word Deflections:



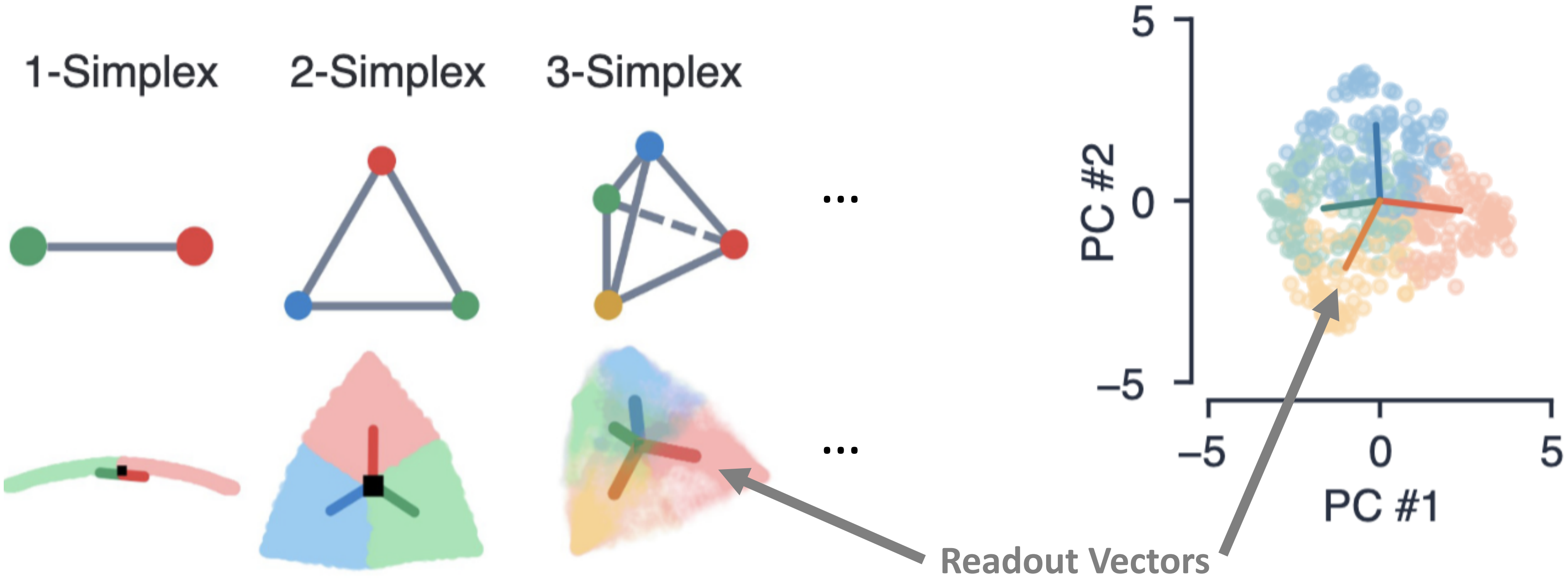
Mechanism: RNN uses 2 dimensions to accumulate and store relative scores for each class

N-Class Categorical Classification

Classes **2-Class** **3-Class** **4-Class** …

4-Class AG News

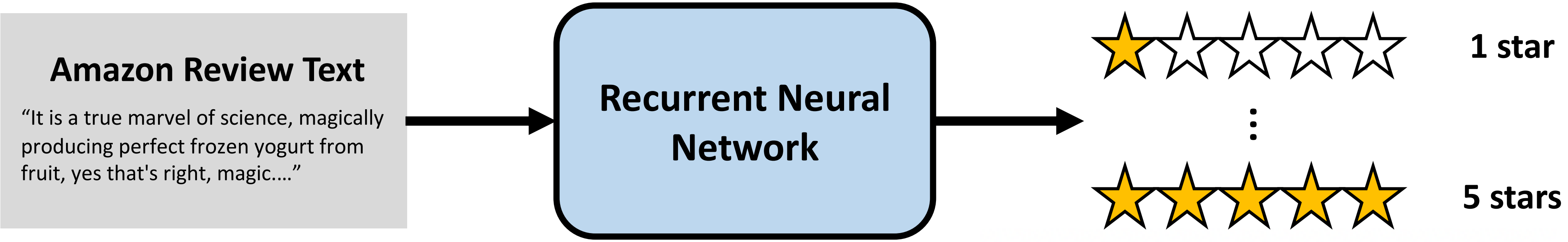
Hidden State/Fixed Point Geometry



Readout Vectors

Mechanism: RNN uses N-1 dimensions to accumulate and store relative scores for each class

5-Class Ordered Classification



Amazon Review Text

"It is a true marvel of science, magically producing perfect frozen yogurt from fruit, yes that's right, magic...."

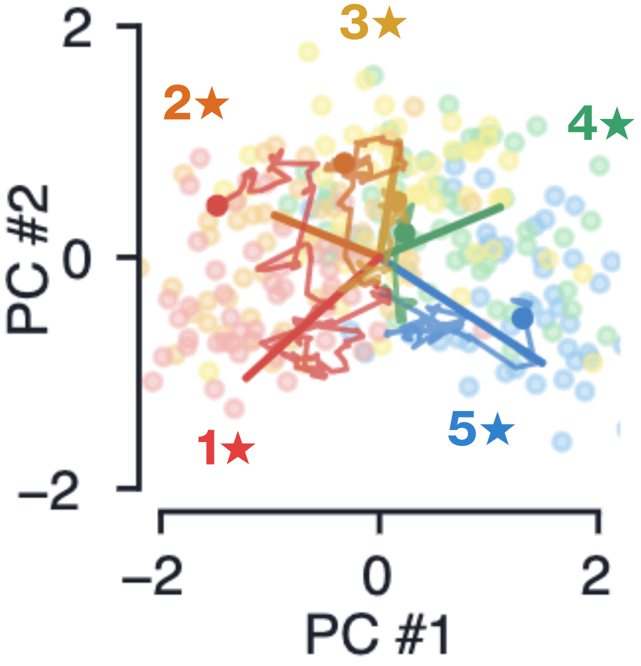
Recurrent Neural Network

1 star

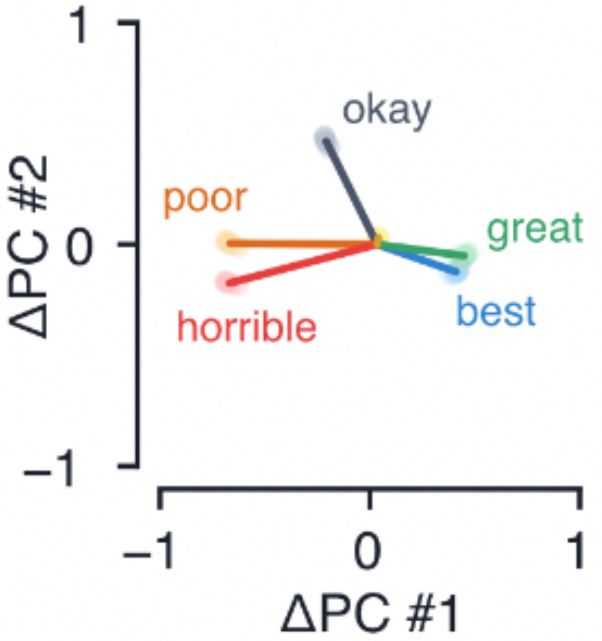
⋮

5 stars

Final hidden states:
(with example trajectories and readouts)



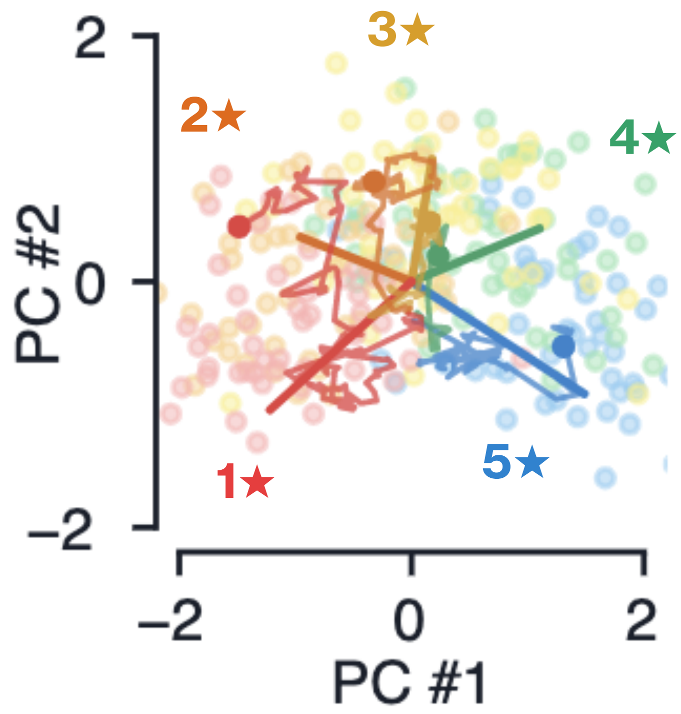
Word Deflections:



5-Class Ordered Classification (cont.)

	5-Class Categorical	**5-Class Ordered**
Geometry:	4-Simplex (4d)	Plane (2d)

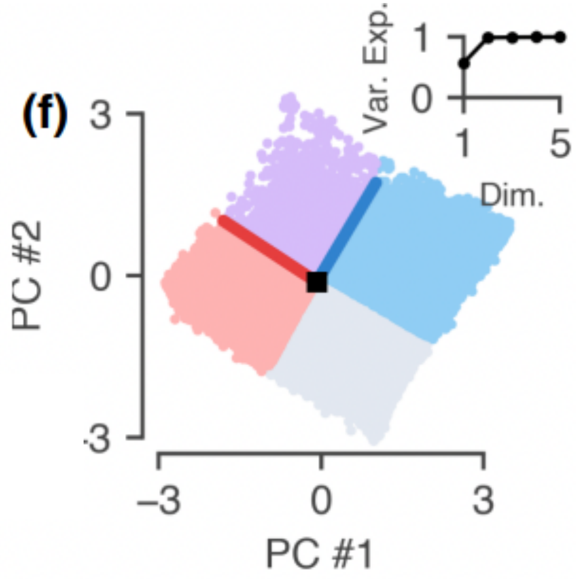
Take away: Correlations between classes can alter manifold dimensionality.



Also in the paper

- Multi-label classification (e.g. emotion tagging)
- Stability analysis underlying hidden state geometry
- Method of predicting dimensionality pre-training
- Simplified synthetic analogs for all three types of datasets
- Universality across RNN types, i.e. LSTMs, GRUs, UGRNNs

Multi-label Geometry



Conclusion: Text-classification RNNs, across architectures, learn low-dimensional, interpretable dynamical systems, with the geometry reflecting dataset statistics.

Ongoing/future work

- Extensions to RNNs for language models and translation models
- How does attention change the underlying dynamics? Can this be used to understand transformers? (Submitted to ICML 2021)

Come check out our booth at ICLR (poster session 5) or get in contact via: kaitken@uw.edu