How recurrent networks implement contextual processing in sentiment analysis

Niru Maheswaranathan and David Sussillo
Google Research
ICML 2020

@niru_m

Sentiment classification using RNNs

Sentiment classification using RNNs

"That restaurant is amazing! I love it!" → **positive**

"I cannot stand that place. Terrible food." → **negative**

Sentiment classification using RNNs

"That restaurant is amazing! I love it!" → **positive**

"I cannot stand that place. Terrible food." → **negative**

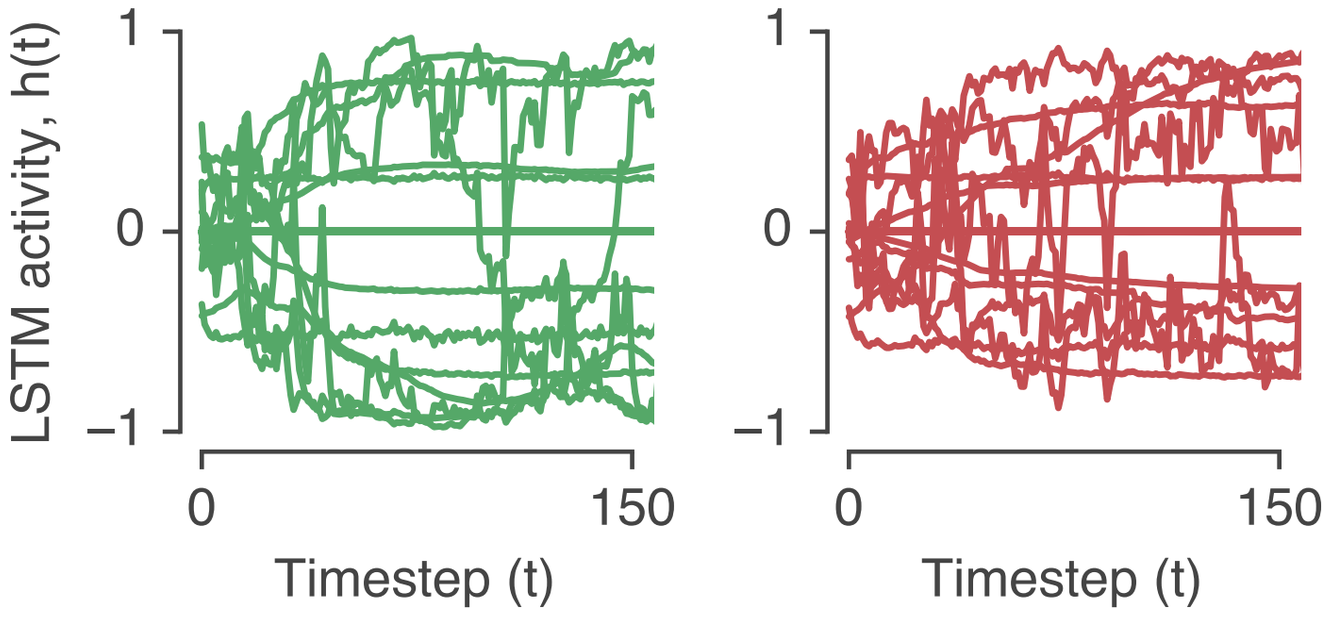
RNNs solve the task, but it's hard to understand ***how*** they do it

Sentiment classification using RNNs

"That restaurant is amazing! I love it!" → **positive**

"I cannot stand that place. Terrible food." → **negative**

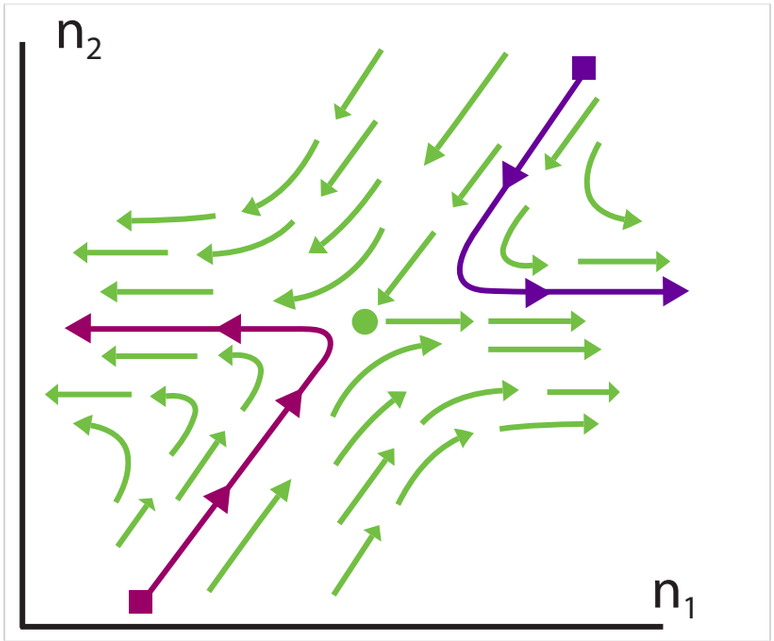
RNNs solve the task, but it's hard to understand ***how*** they do it



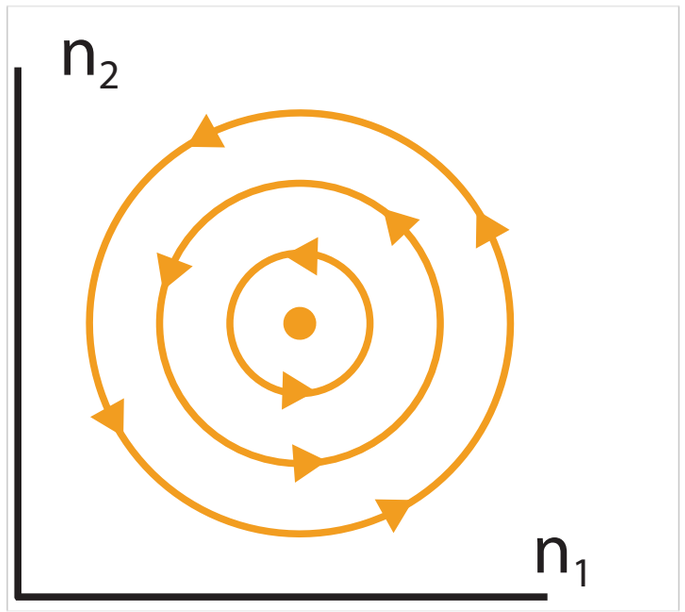
Understanding RNN dynamics through linearization

Understanding RNN dynamics through linearization

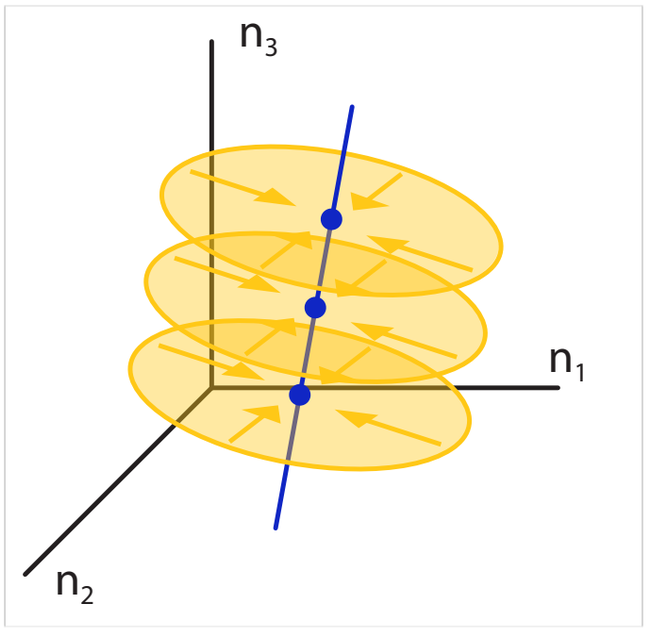
Saddle Point



Oscillations

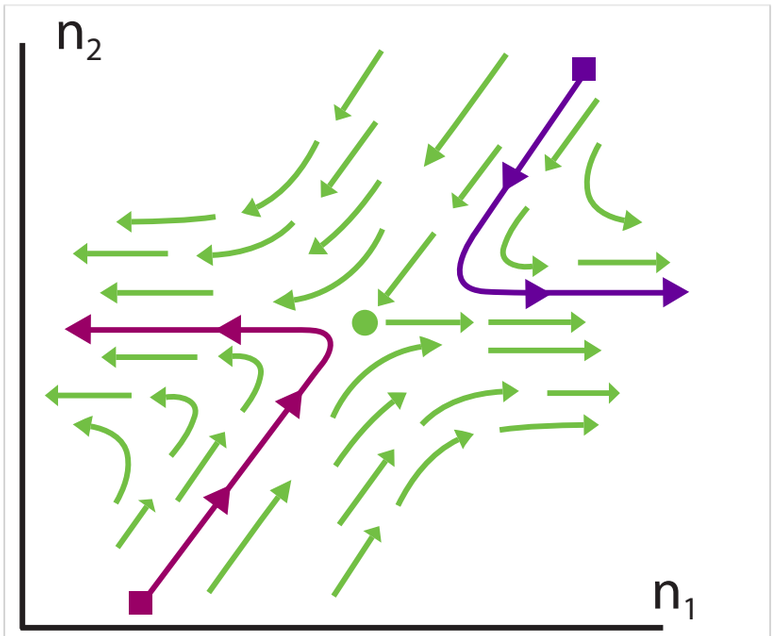


Line Attractor

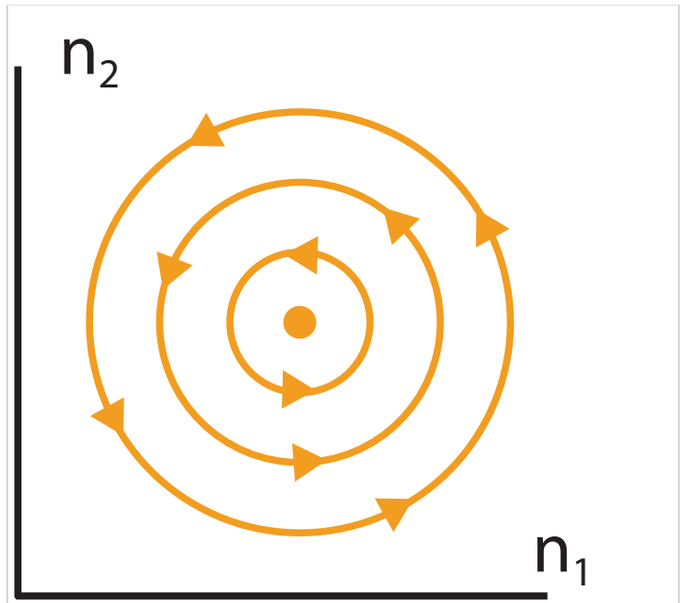


Understanding RNN dynamics through linearization

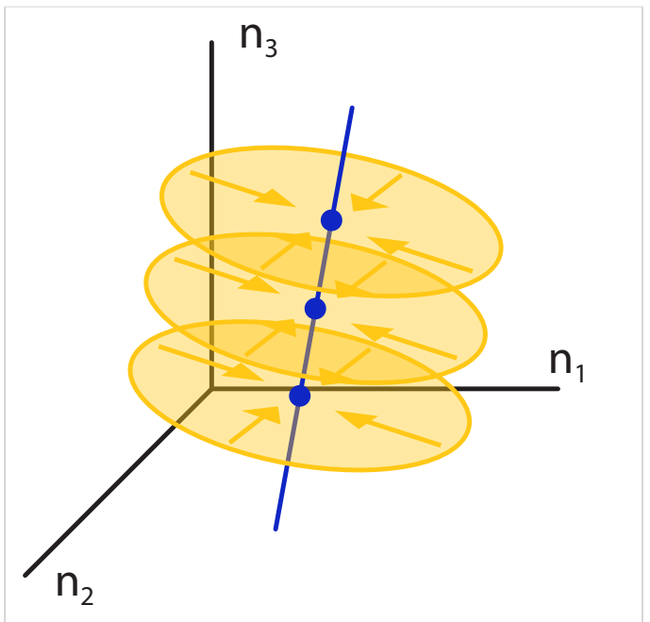
Saddle Point



Oscillations

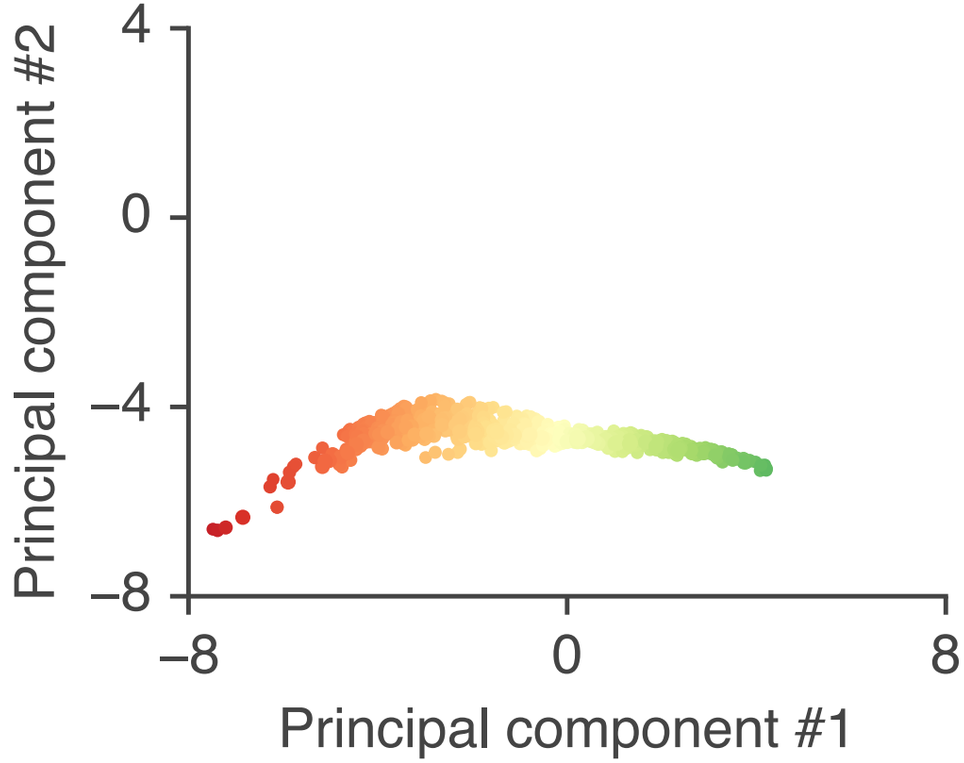


Line Attractor



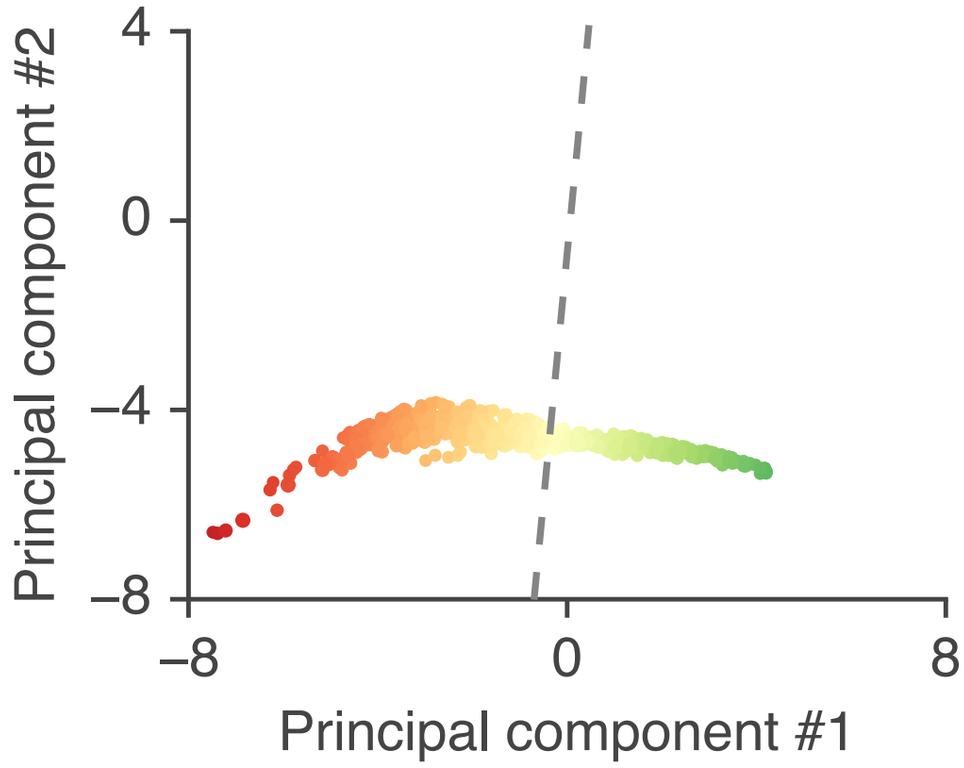
Line attractor dynamics in trained RNNs

Line attractor dynamics in trained RNNs



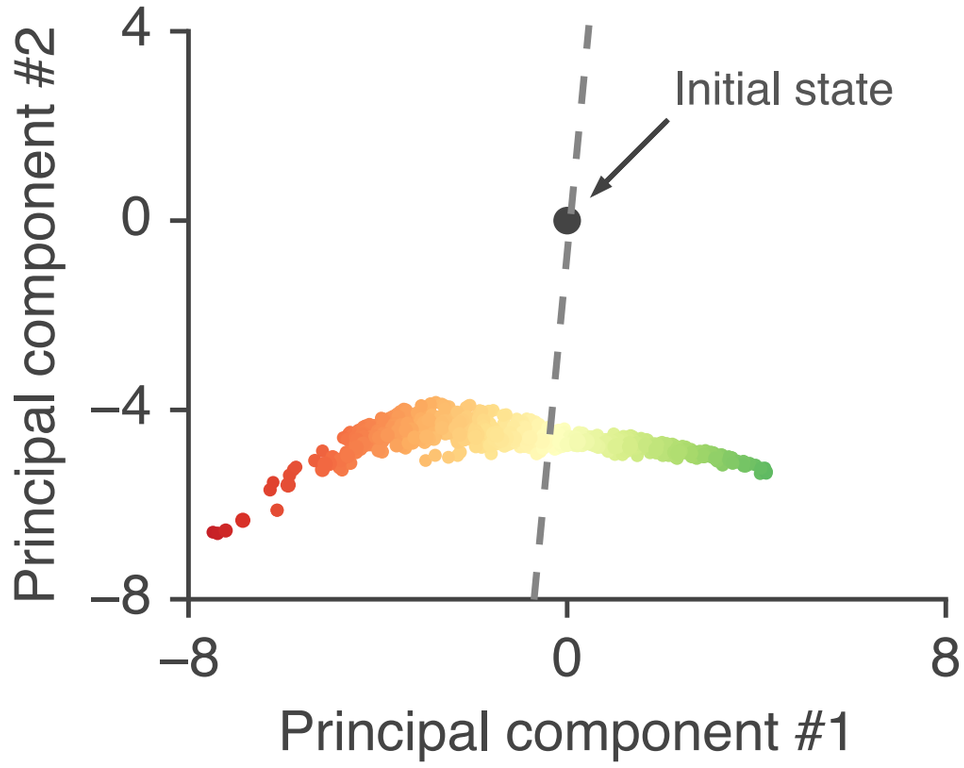
Maheswaranathan*, Williams* et al, NeurIPS 2019

Line attractor dynamics in trained RNNs



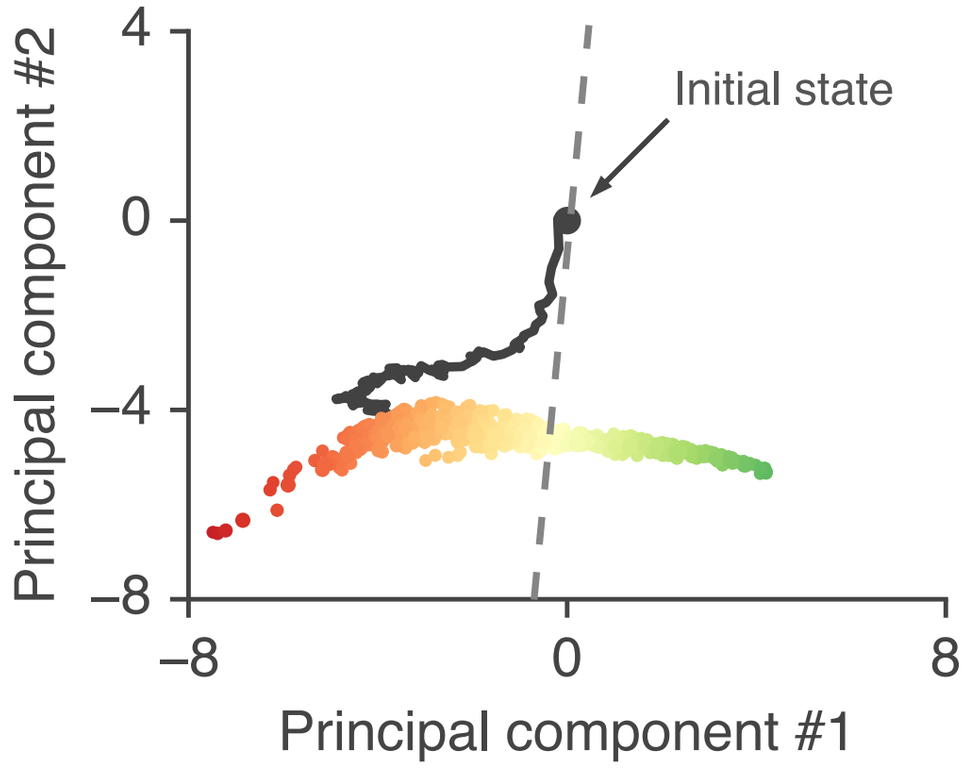
Maheswaranathan*, Williams* et al, NeurIPS 2019

Line attractor dynamics in trained RNNs



Maheswaranathan*, Williams* et al, NeurIPS 2019

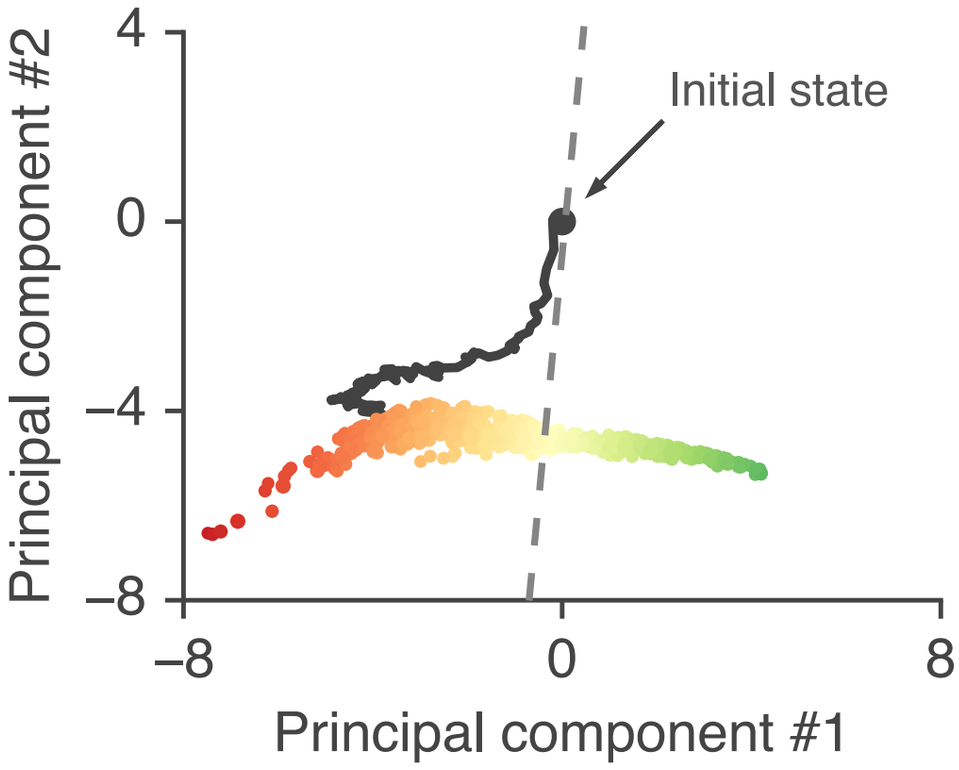
Line attractor dynamics in trained RNNs



Maheswaranathan*, Williams* et al, NeurIPS 2019

Line attractor dynamics in trained RNNs

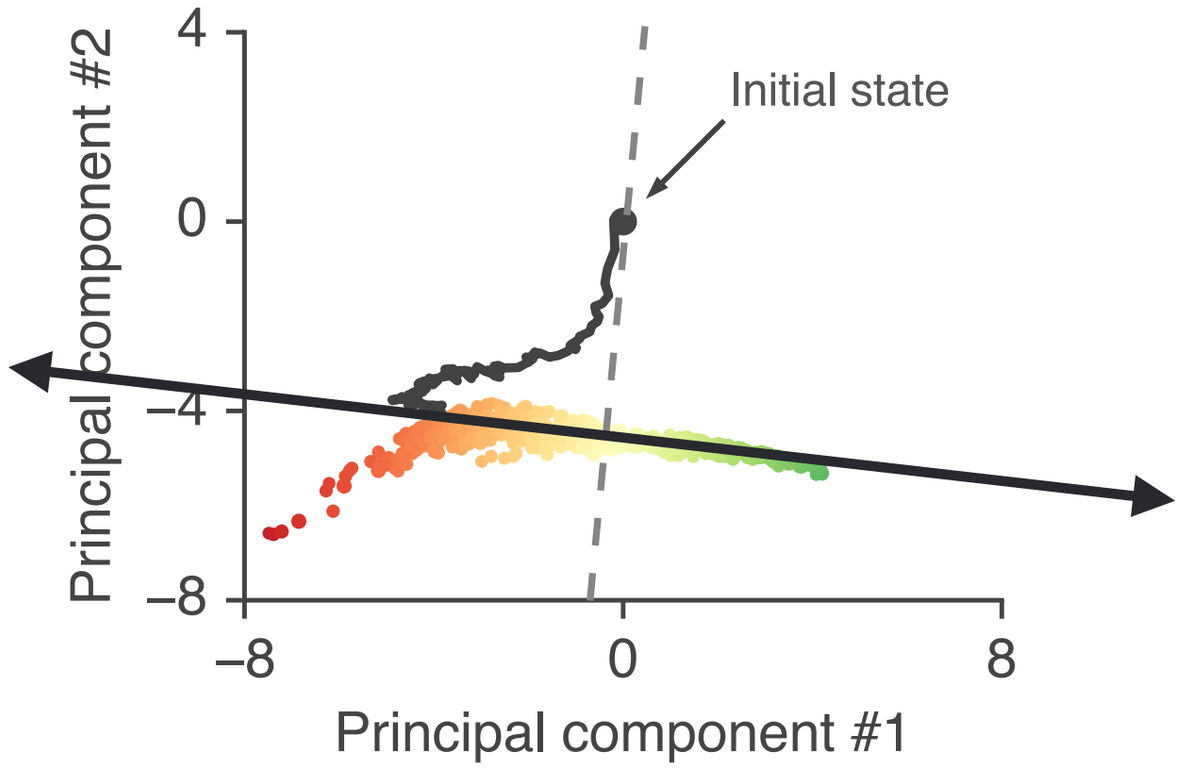
Approximate line attractor dynamics explain the most of the RNN's performance



Maheswaranathan*, Williams* et al, NeurIPS 2019

Line attractor dynamics in trained RNNs

Approximate line attractor dynamics explain the most of the RNN's performance



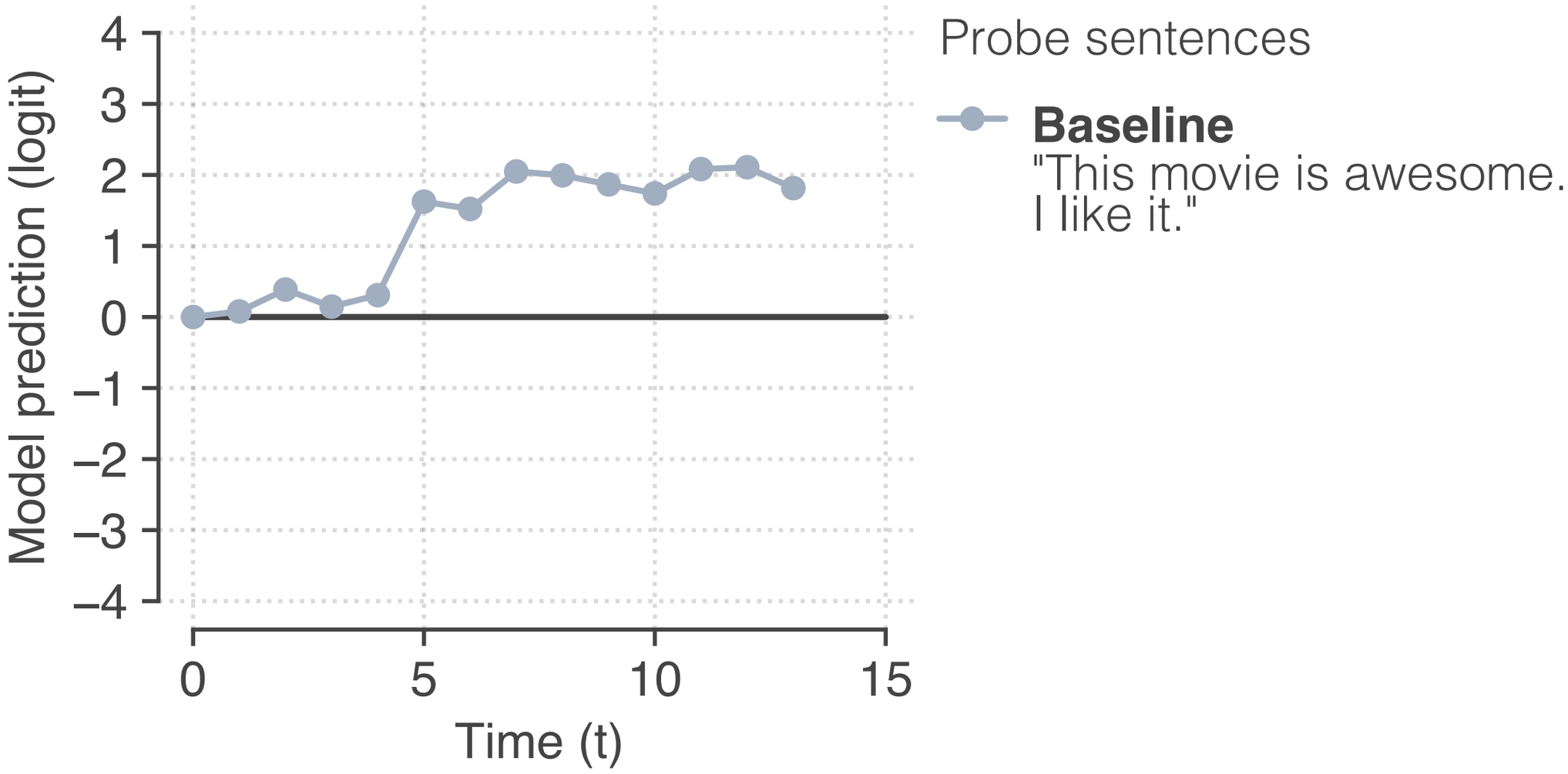
Maheswaranathan*, Williams* et al, NeurIPS 2019

Line attractor dynamics in trained RNNs

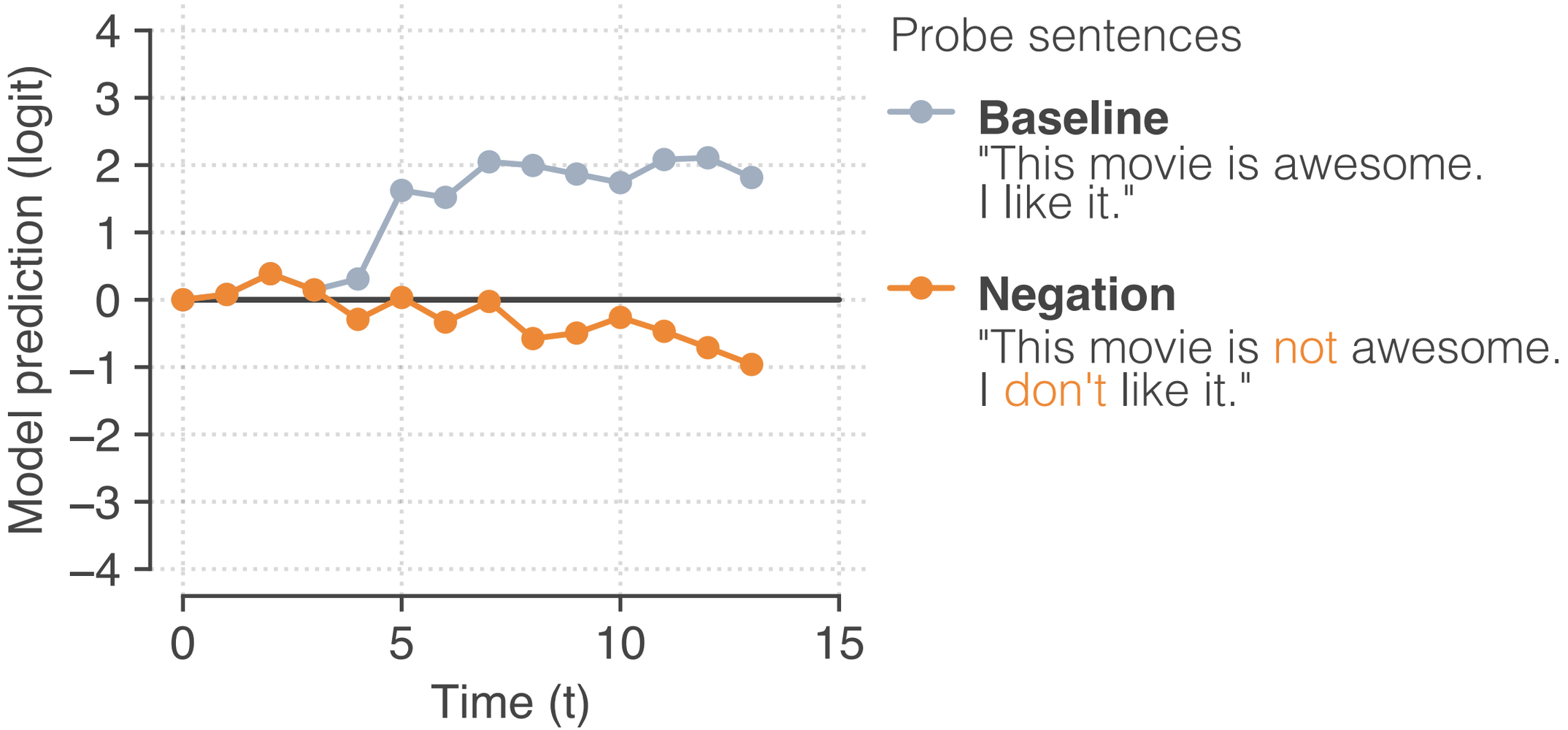
Approximate line attractor dynamics explain the most of the RNN's performance

A remaining puzzle…

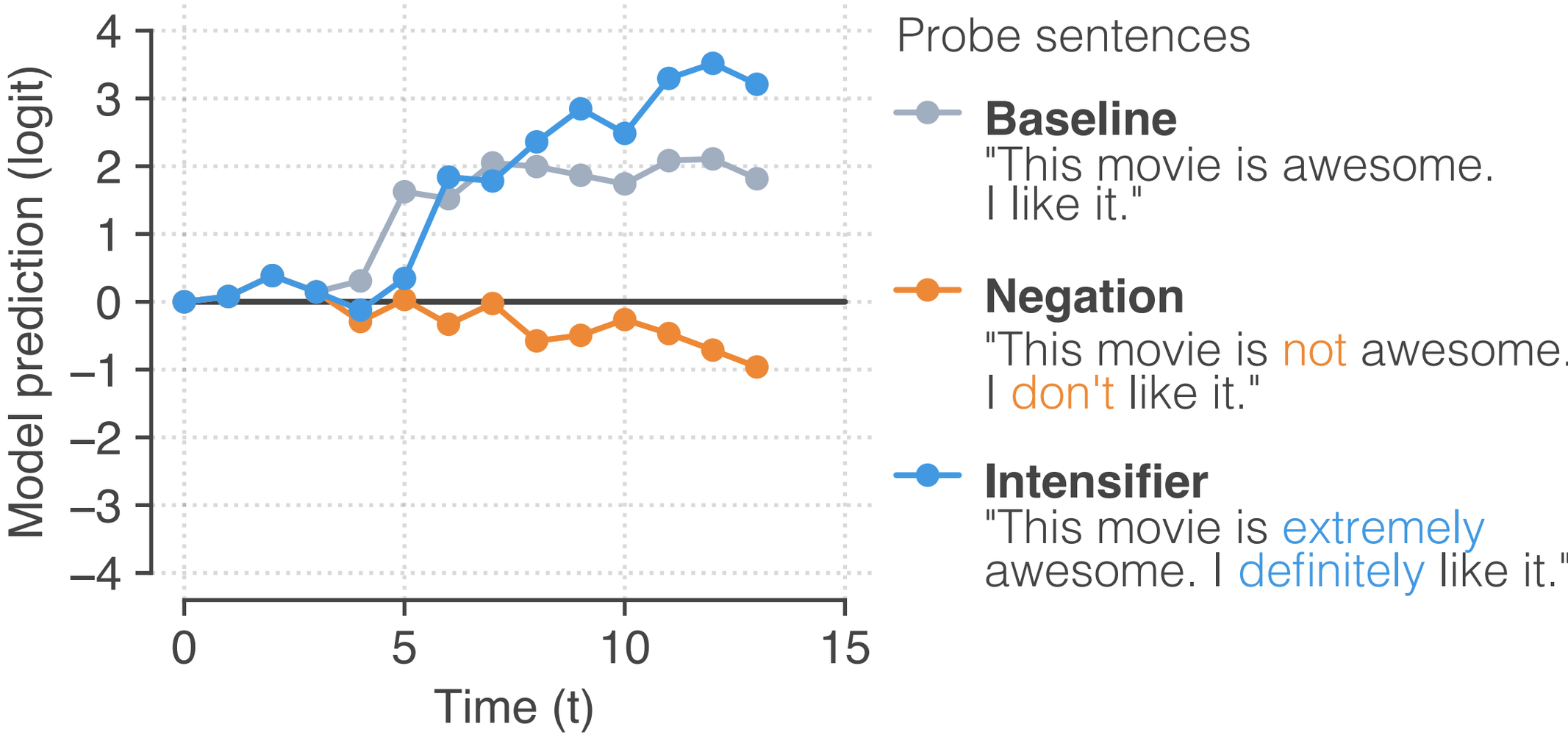
A remaining puzzle…



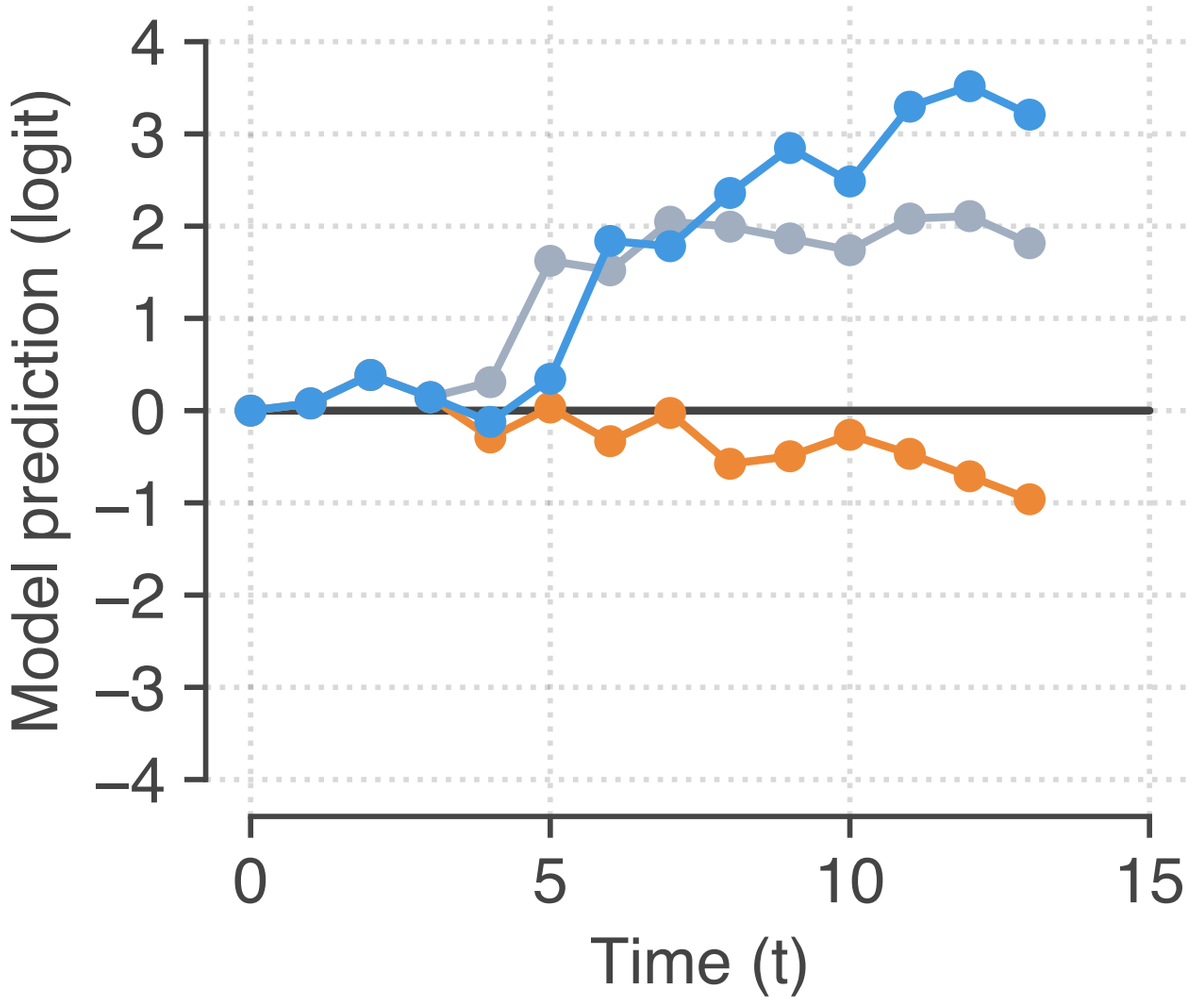
A remaining puzzle…



A remaining puzzle…



Contextual processing in RNNs



Probe sentences

Baseline
"This movie is awesome. I like it."

Negation
"This movie is not awesome. I don't like it."

Intensifier
"This movie is extremely awesome. I definitely like it."

Contextual processing in RNNs

Contributions of our work

Contextual processing in RNNs

Contributions of our work

- Data-driven method to identify contextual inputs

Contextual processing in RNNs

Contributions of our work

- Data-driven method to identify contextual inputs
- Analysis of the strength and timing of modifier effects

Contextual processing in RNNs

Contributions of our work

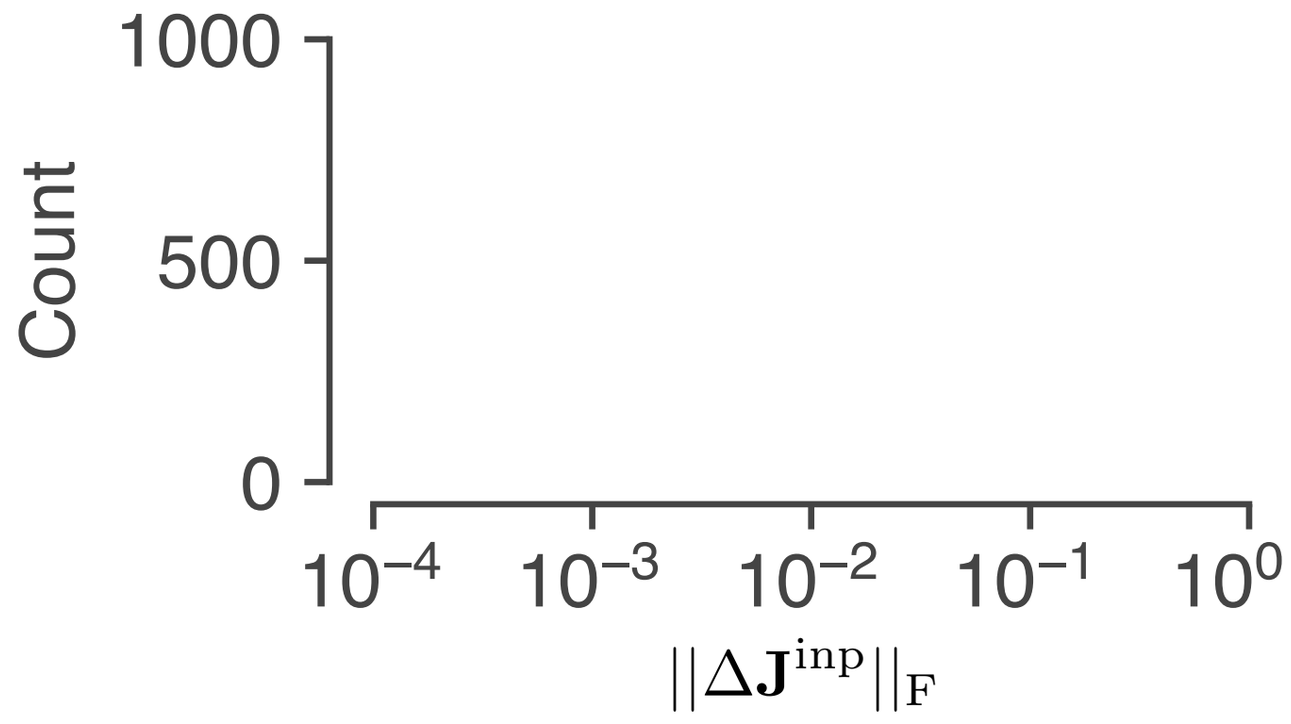
- Data-driven method to identify contextual inputs
- Analysis of the strength and timing of modifier effects
- Experiments that demonstrate the identified mechanisms are necessary and sufficient for RNN performance

Identifying contextual processing

Use the **change in input sensitivity** as a measure of contextual processing

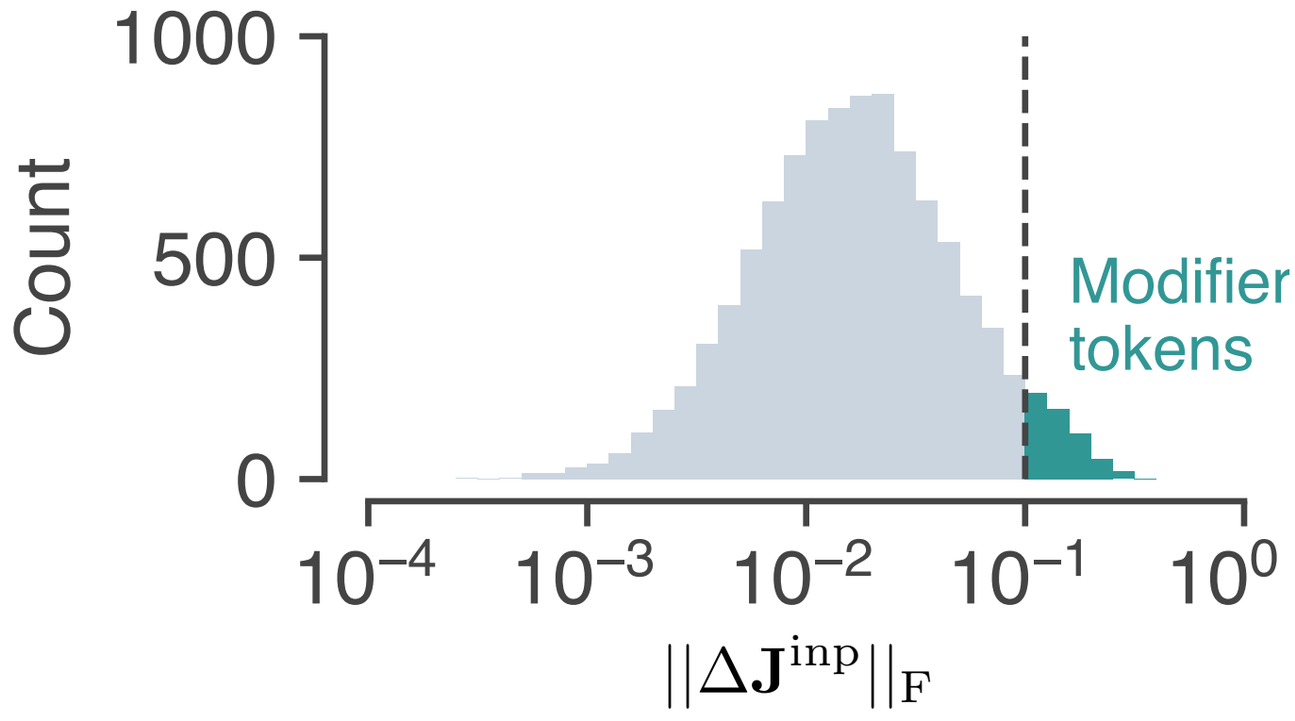
Identifying contextual processing

Use the **change in input sensitivity** as a measure of contextual processing

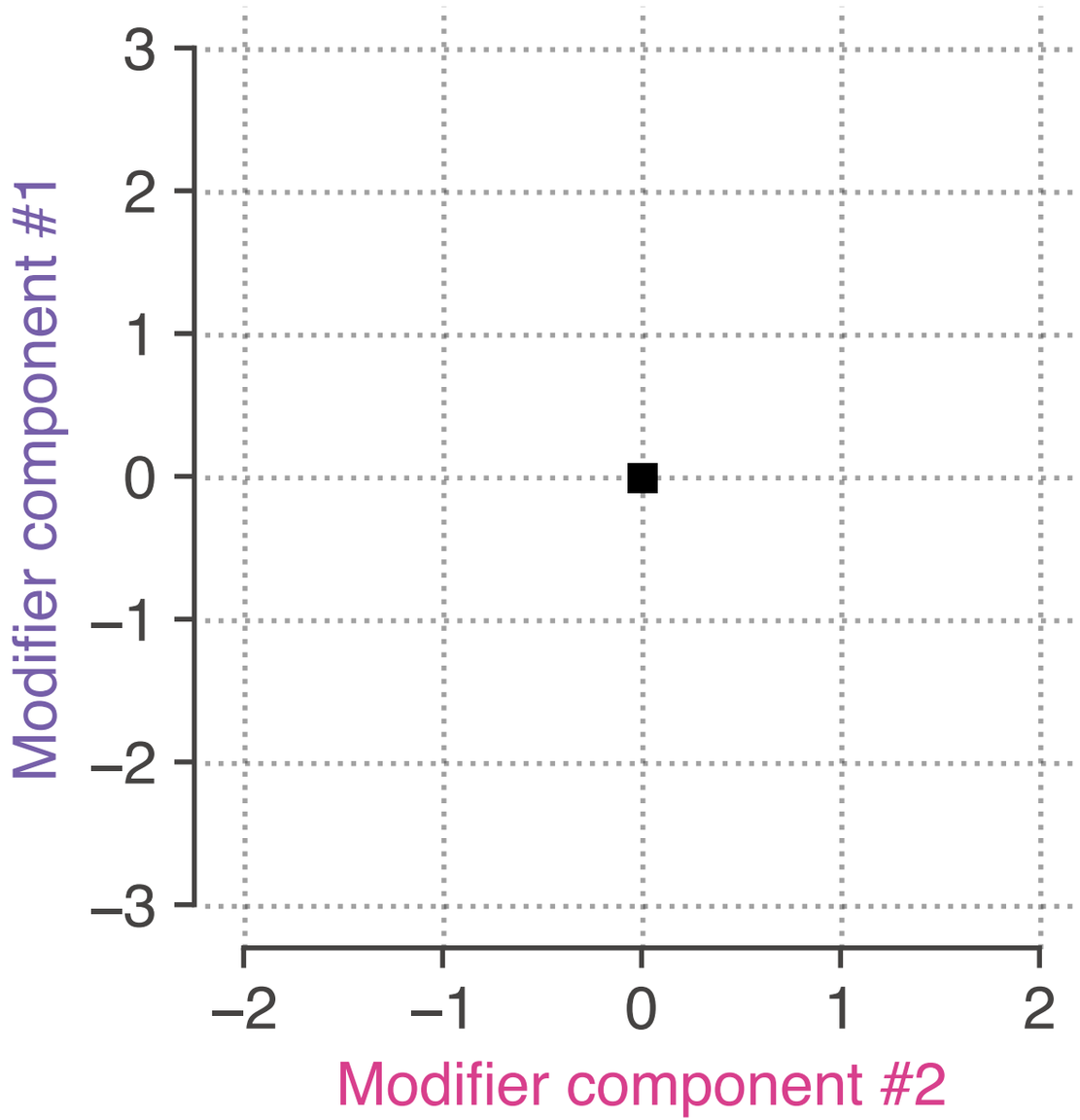


Identifying contextual processing

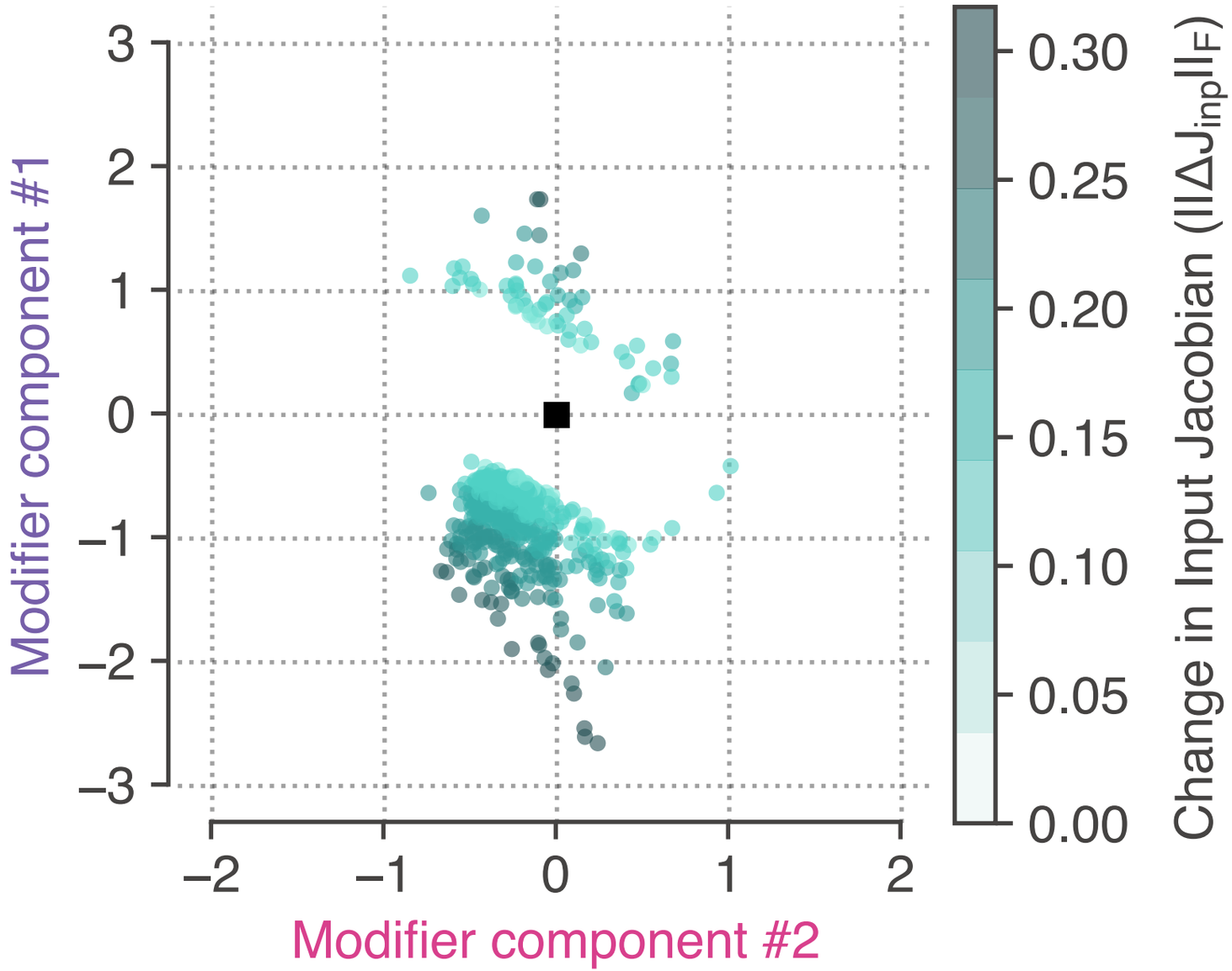
Allows us to identify **modifier inputs**



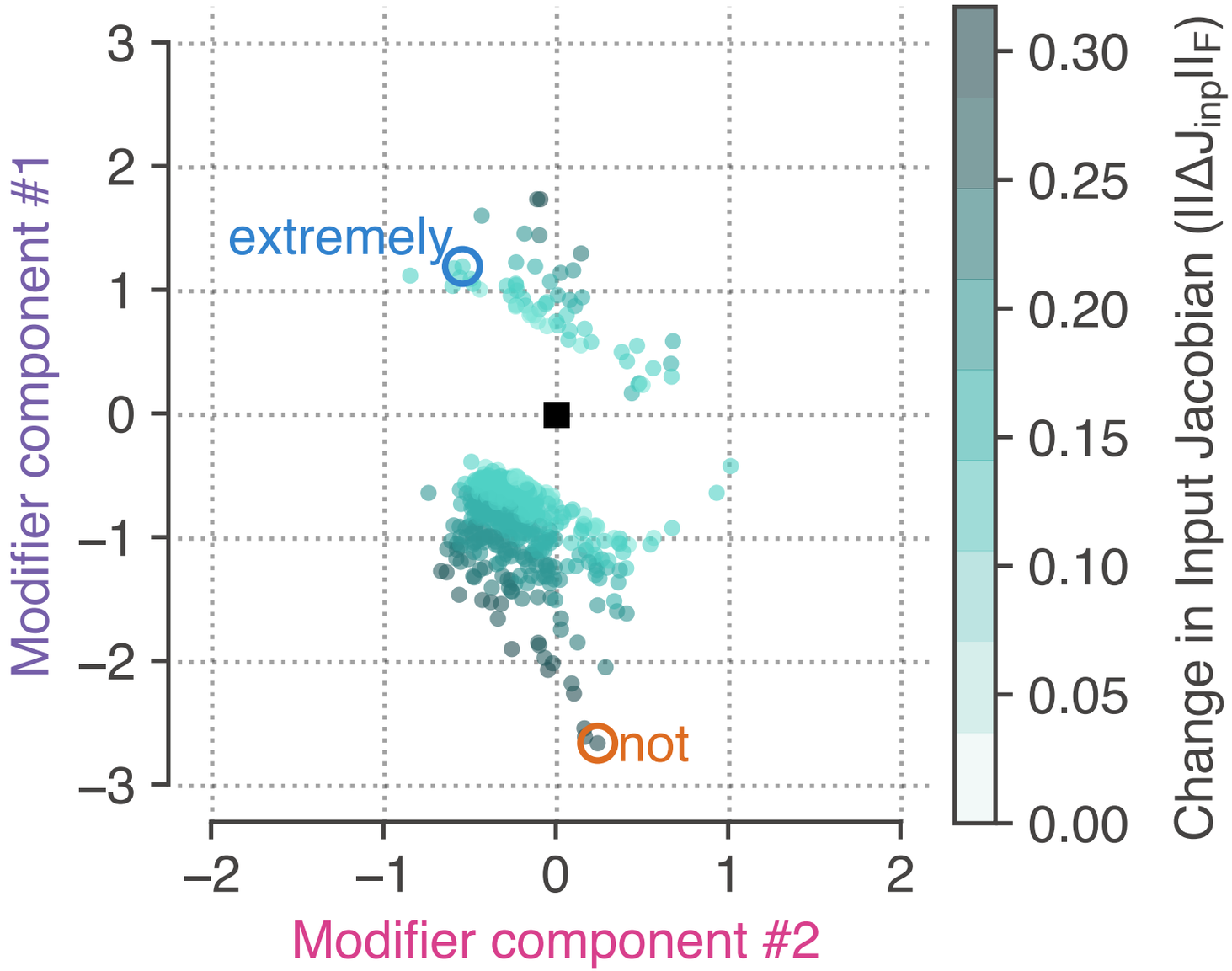
Modifier subspace



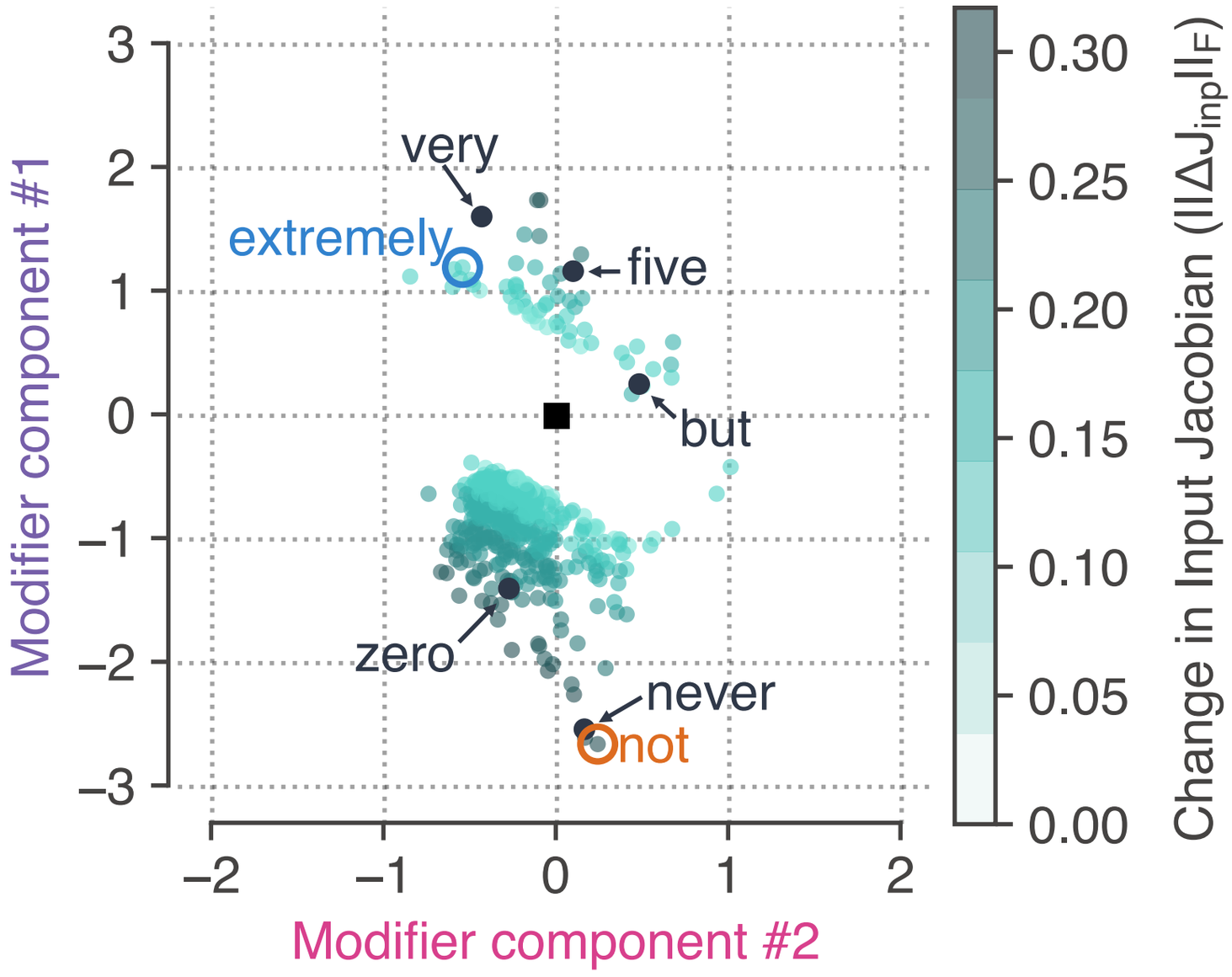
Modifier subspace



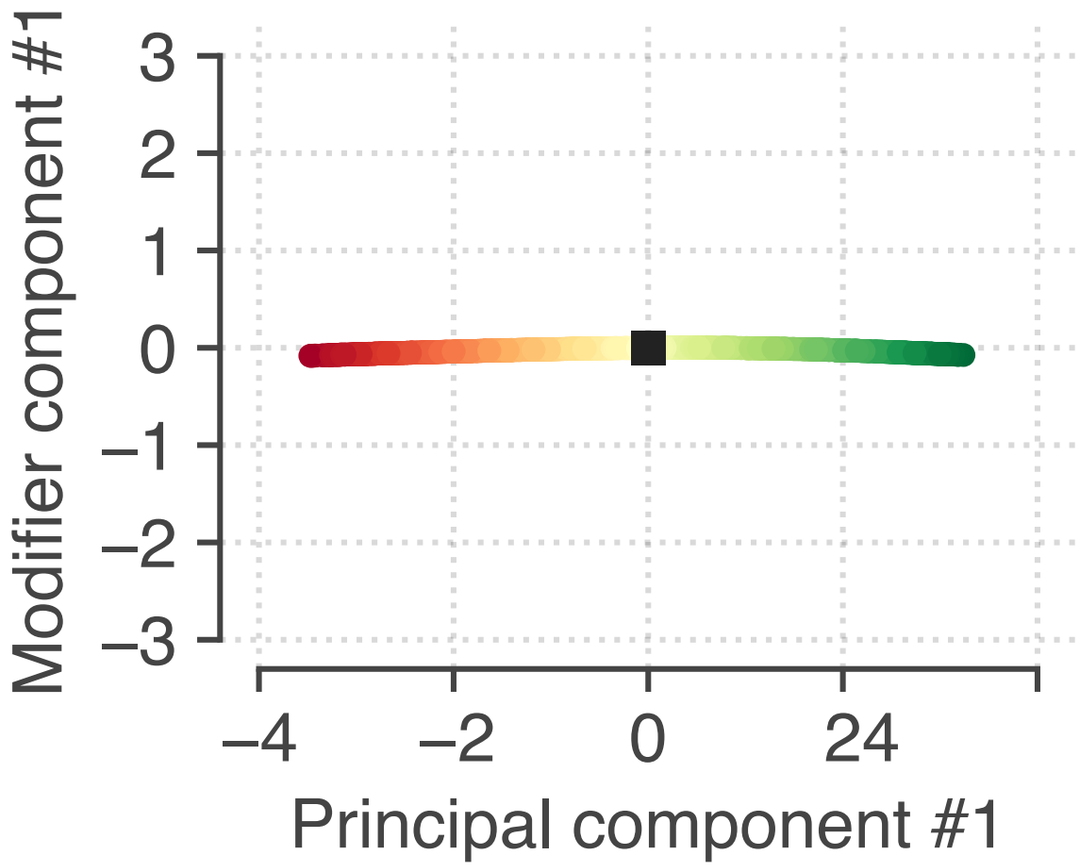
Modifier subspace



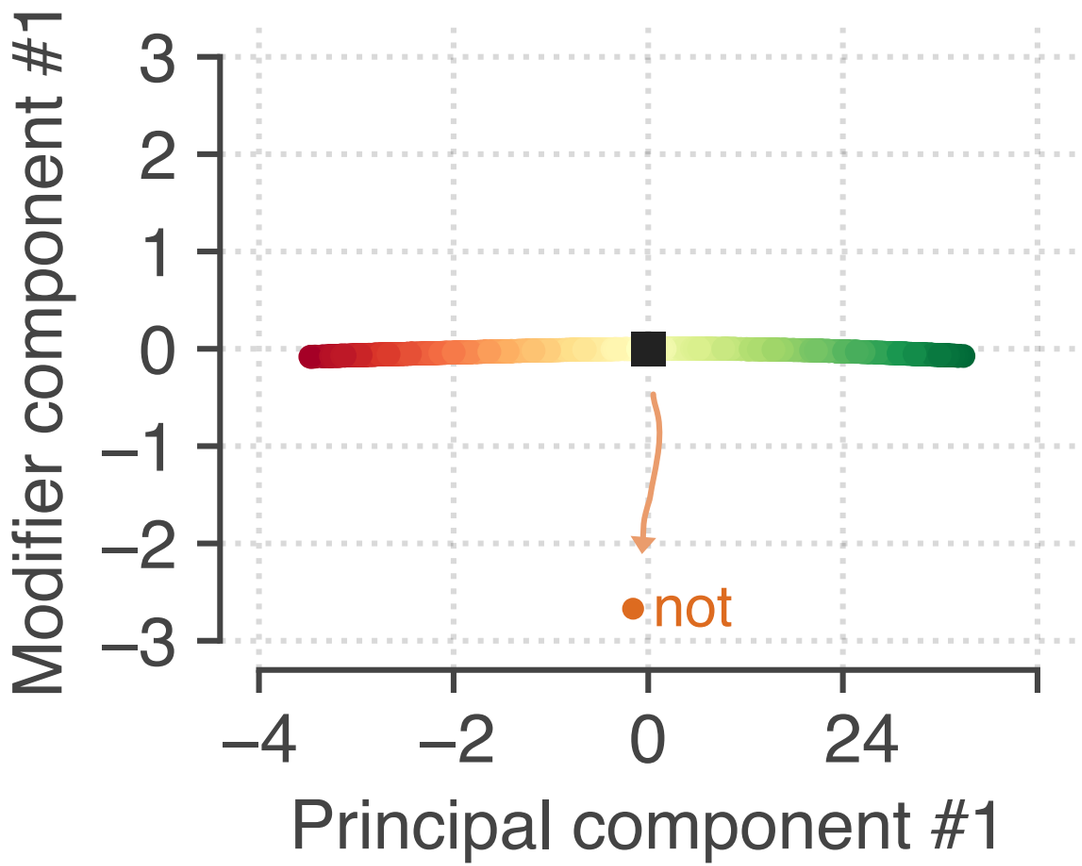
Modifier subspace



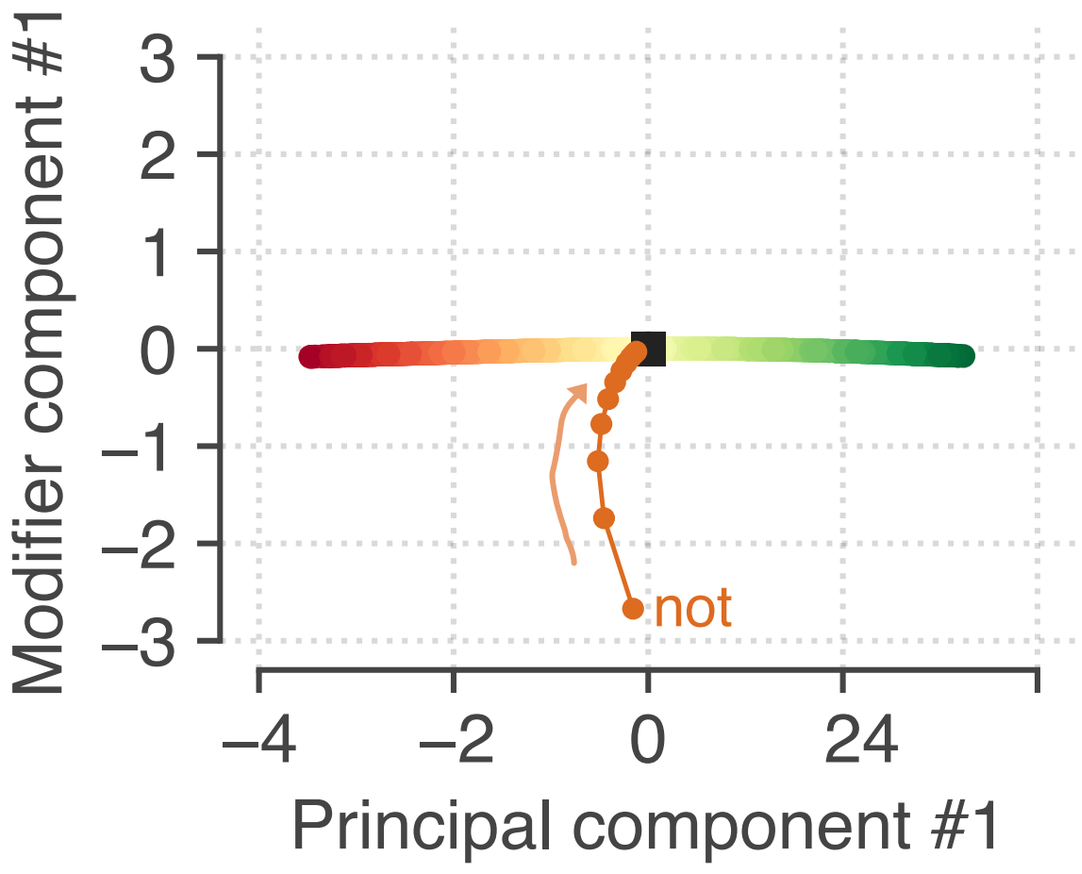
Modifier dynamics



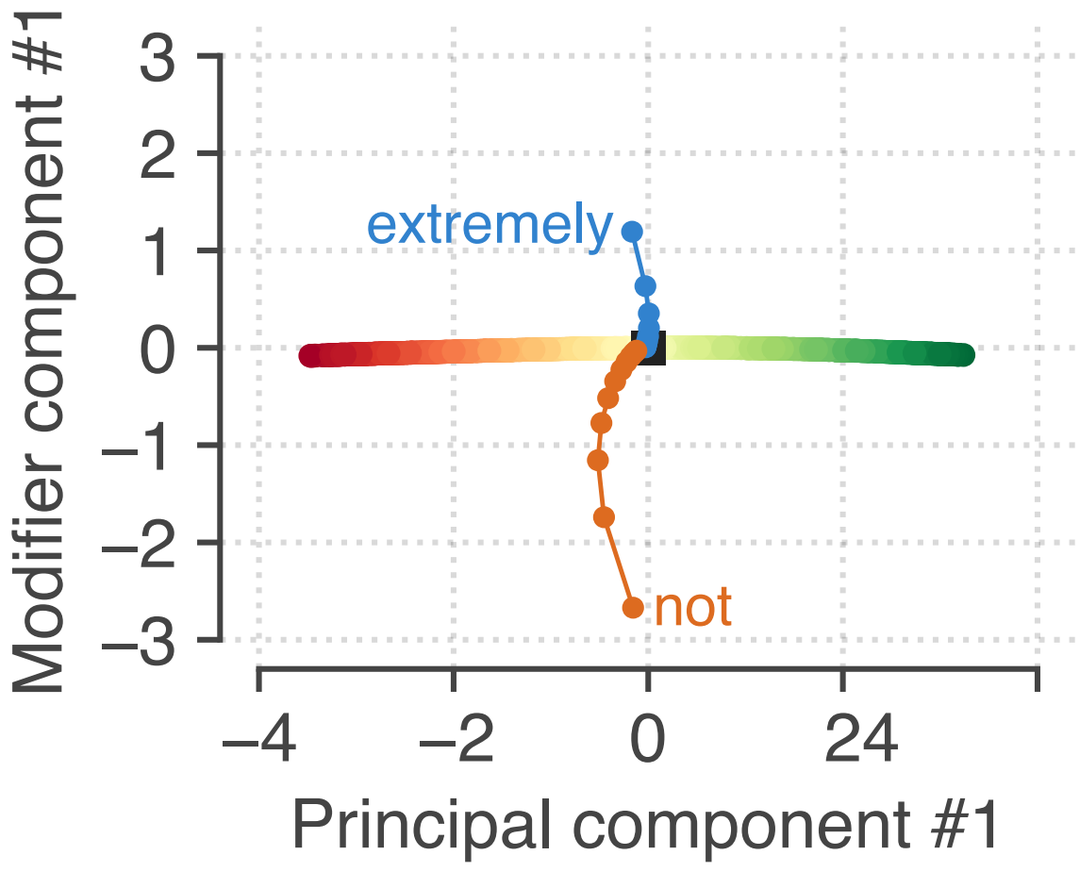
Modifier dynamics



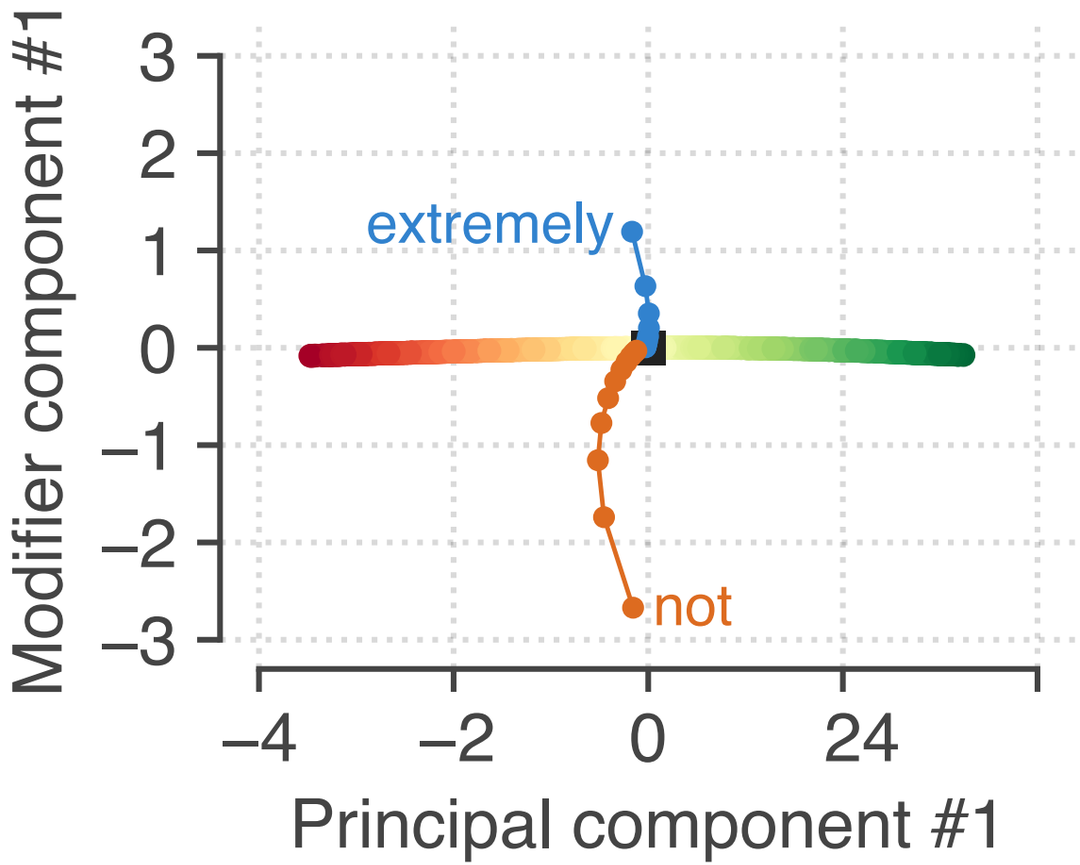
Modifier dynamics

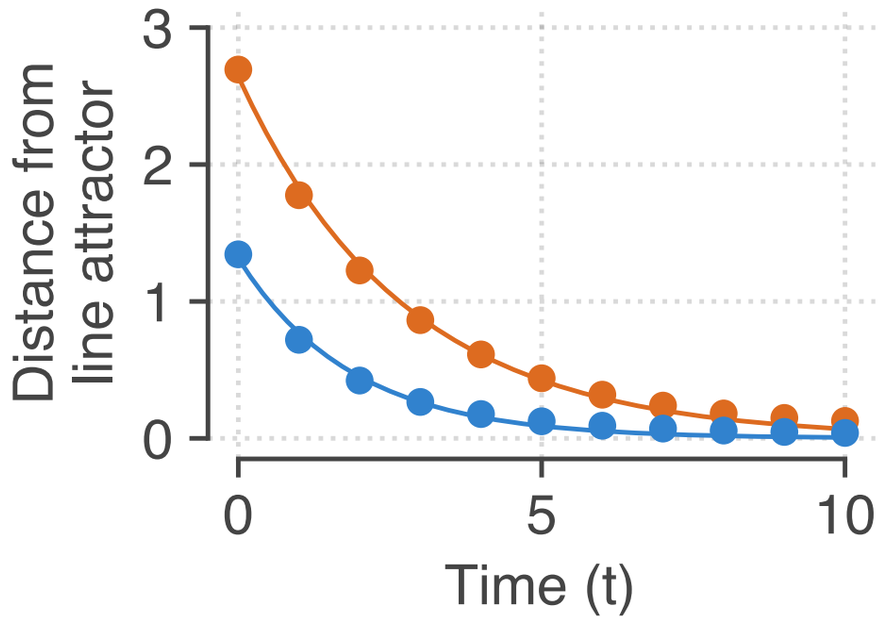


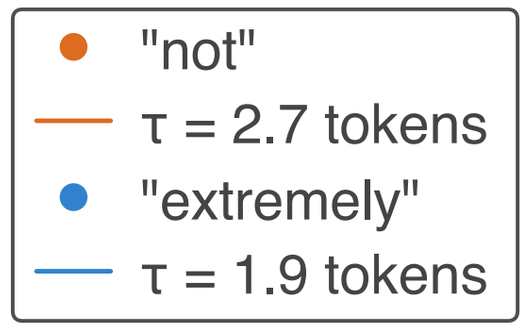
Modifier dynamics



Modifier dynamics

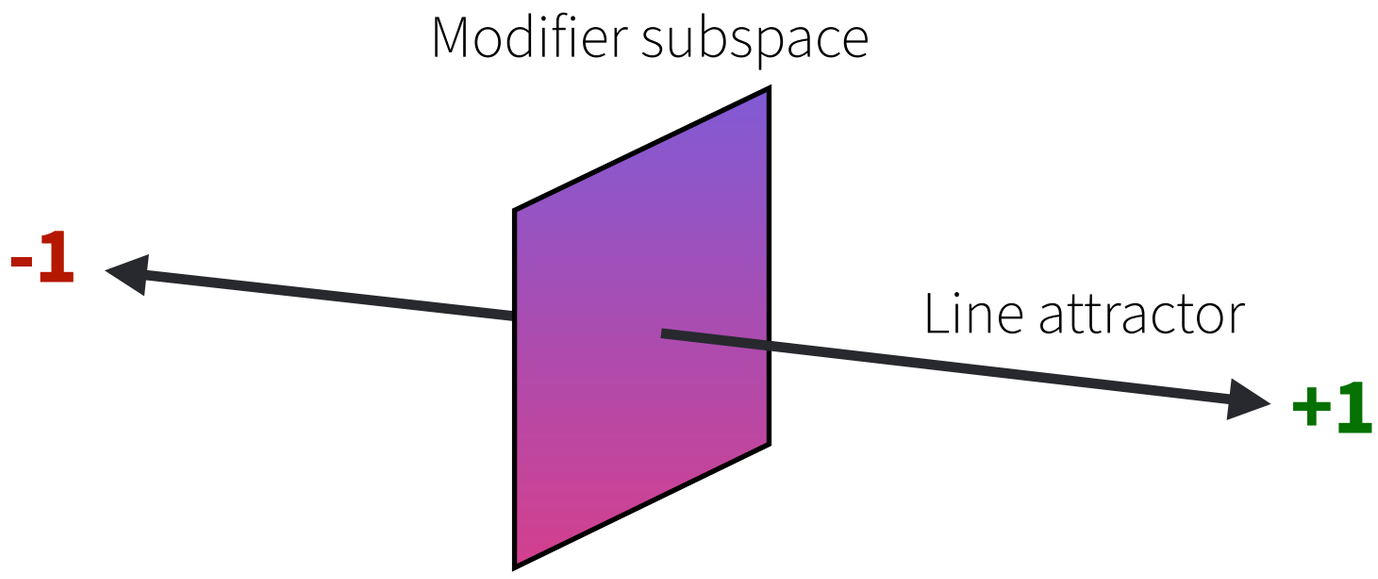




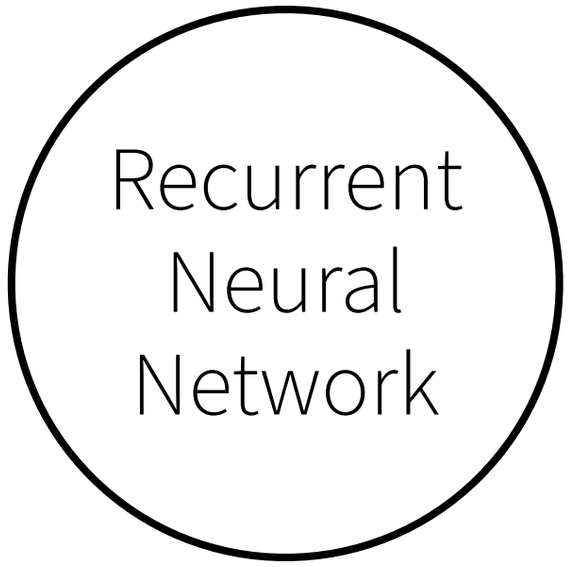


Synthesizing our new understanding

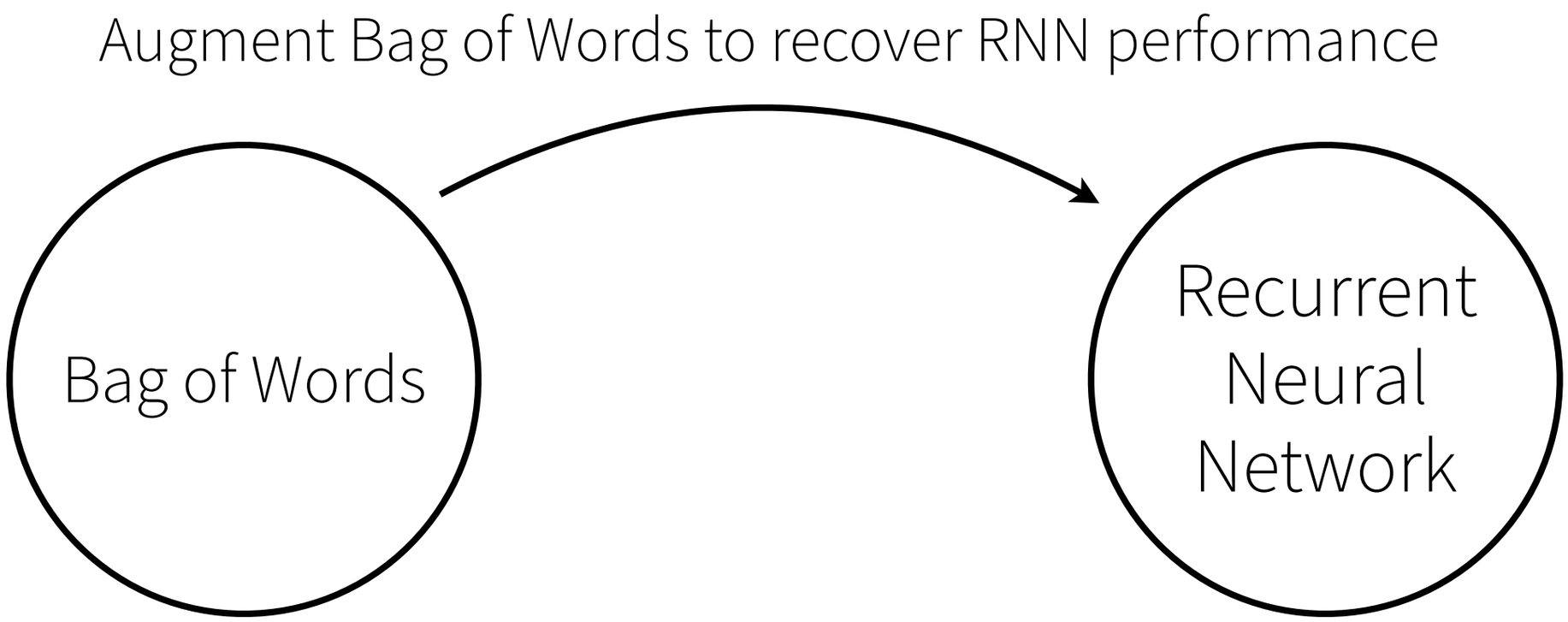
Synthesizing our new understanding



Synthesizing our new understanding



Synthesizing our new understanding



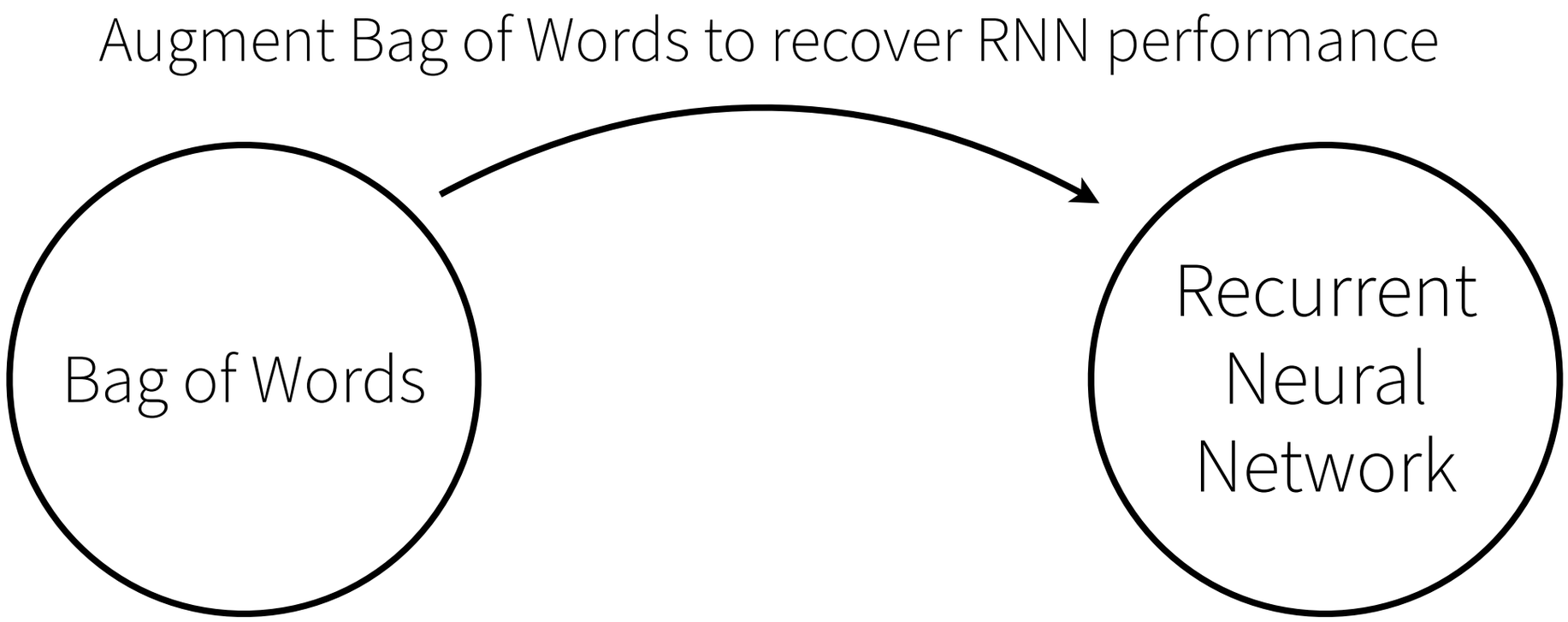
Augmented bag-of-words model recovers RNN performance

Model	Accuracy
Bag of Words (Baseline)	93.6%
RNN (GRU)	95.8%

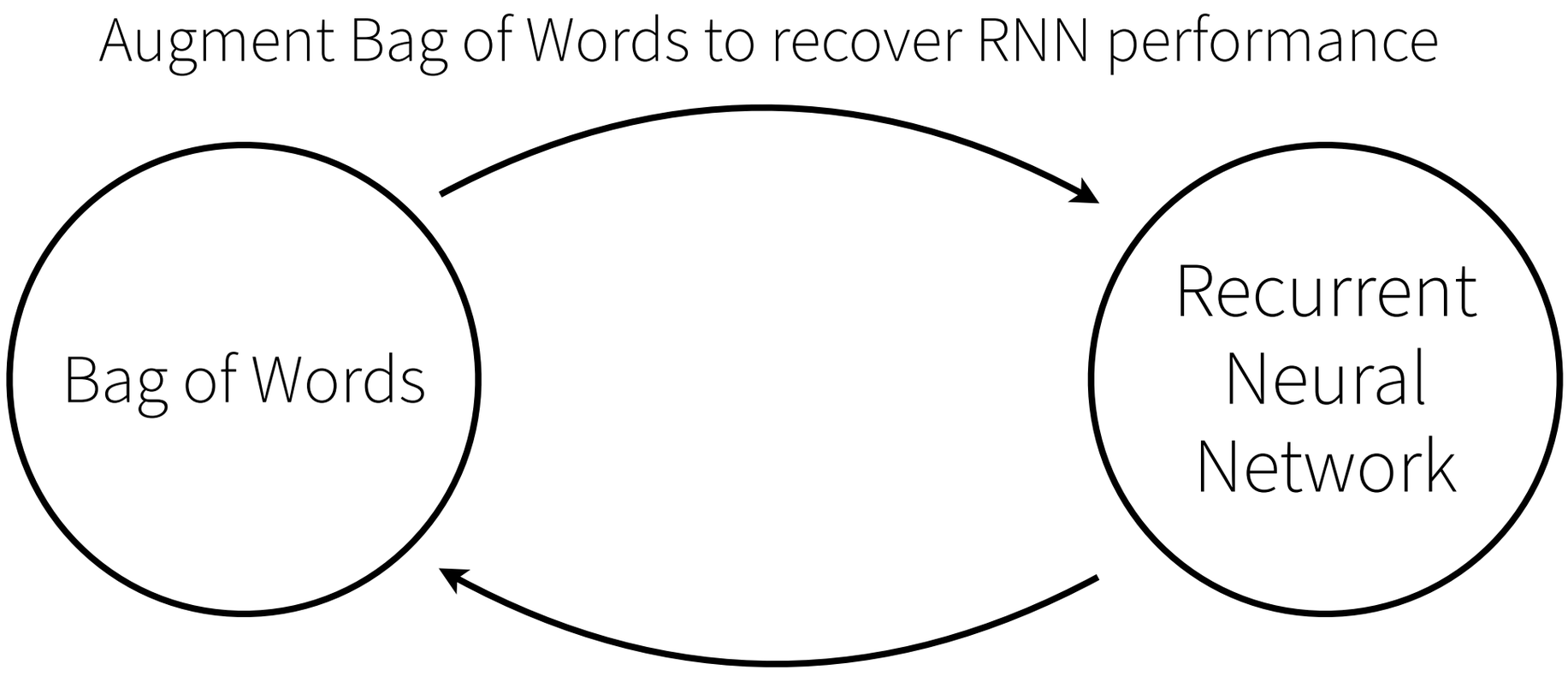
Augmented bag-of-words model recovers RNN performance

Model	Accuracy
Bag of Words (Baseline)	93.6%
Augmented Bag-of-Words (includes modifier effects)	95.5%
RNN (GRU)	95.8%

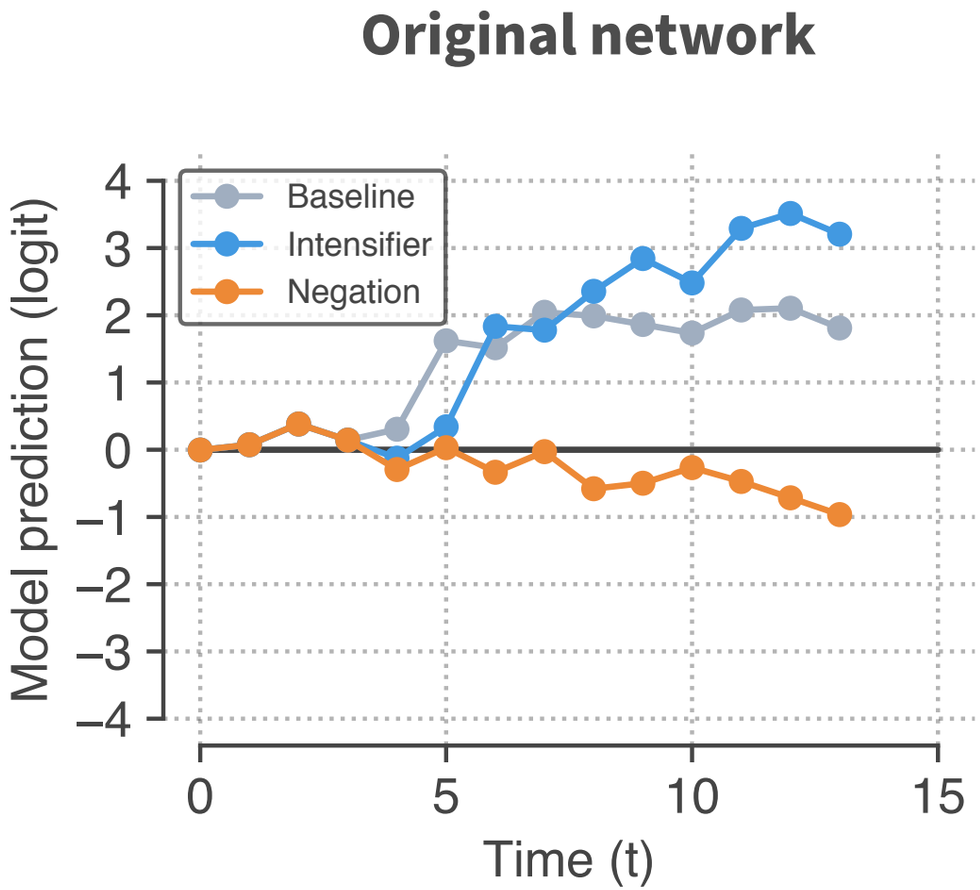
Synthesizing our new understanding



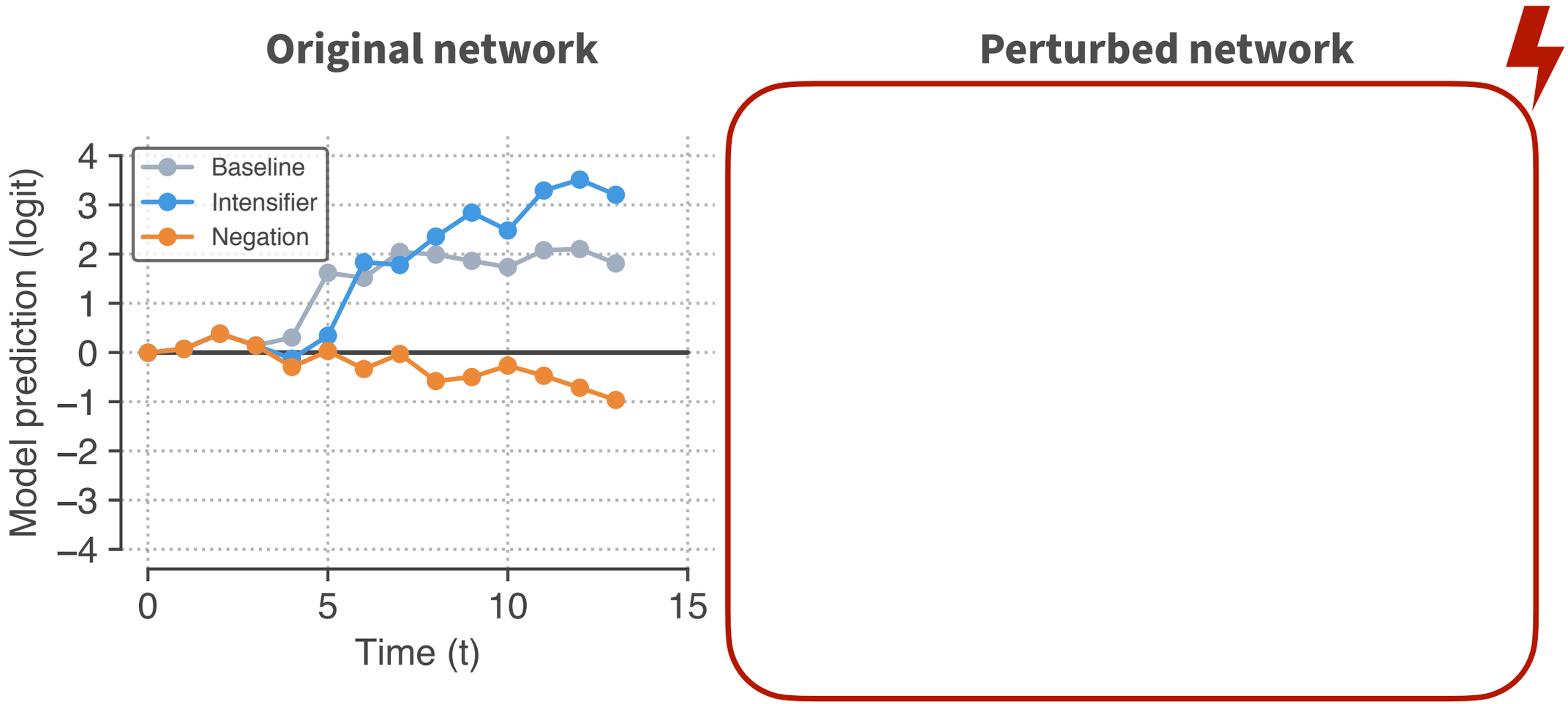
Synthesizing our new understanding



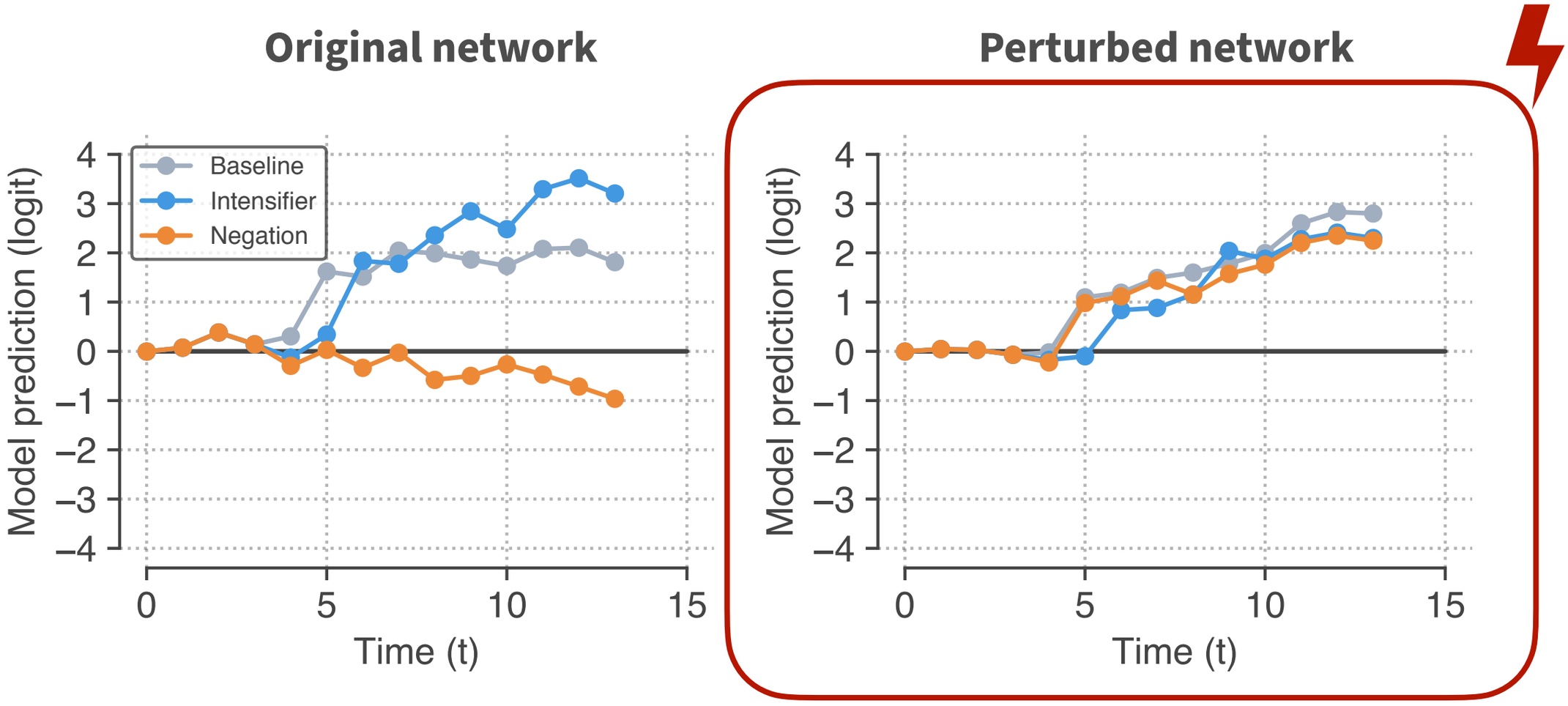
Perturbation experiment removes modifier effects



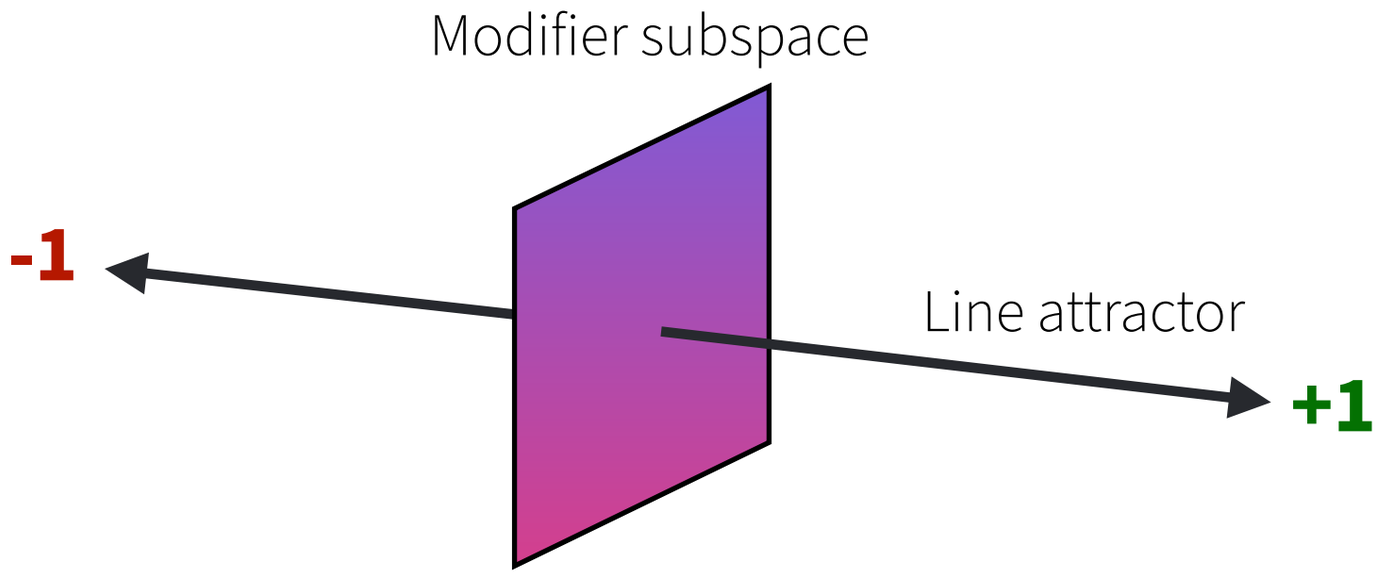
Perturbation experiment removes modifier effects



Perturbation experiment removes modifier effects



Thank you!



Paper

arxiv.org/abs/2004.08013

Niru Maheswaranathan

@niru_m

nirum@google.com