Partial observation can induce mechanistic mismatches in data-constrained models of neural dynamics

WILLIAM QIAN

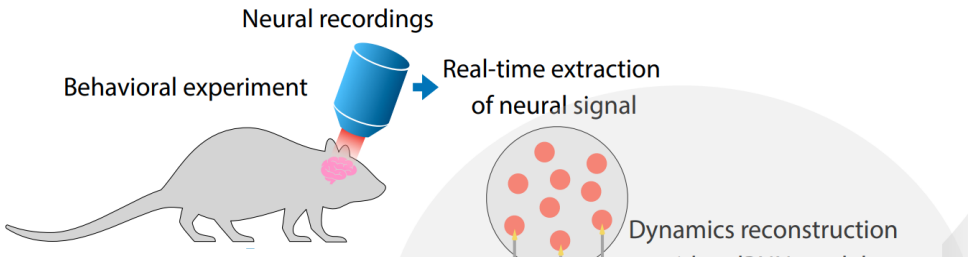
JACOB ZAVATONE-VETH

BENJAMIN S. RUBEN

CENGIZ PEHLEVAN

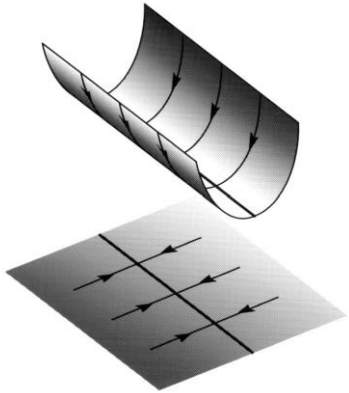
NeurIPS, 2024

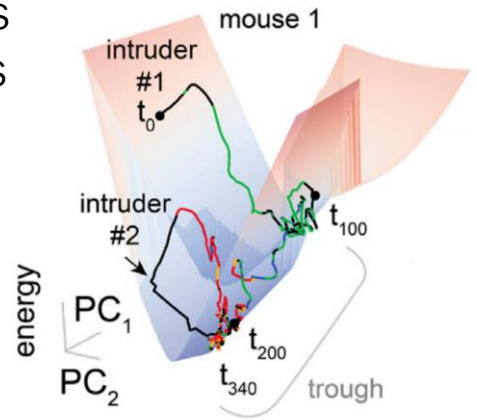
Uncovering mechanisms of neural circuits via data-driven modeling



Dynamical Systems Analysis

"mechanism"





Barak and Sussillo (2013)
Mante, Sussillo, Shenoy, Newsome (2013)
Rajan, Harvey, Tank (2016)
Perich et al. (2021)
Nair et al. (2023)
Vinograd et al. (2024)

Other figures:
1. Fatih Dinc, Adam Shai, Mark Schnitzer, Hidenori Tanaka. CORNN: Convex optimization of recurrent neural networks for rapid inference of neural dynamics
2. H.S. Seung, How the brain keeps the eyes still

Mechanistic Identifiability

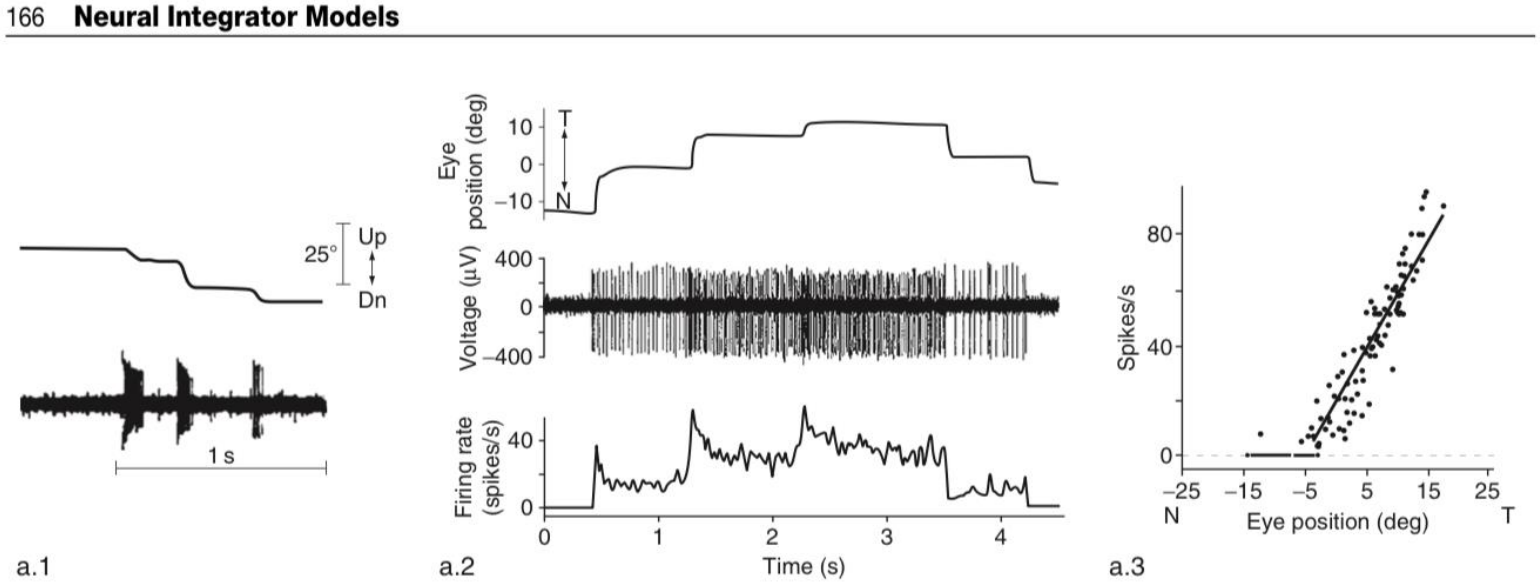
When can we trust the mechanistic insights gained from this procedure?

Is it possible that we fit the data perfectly but still get the mechanism wrong?

Possible issues

- Partially observed
- Observation noise
- Low-pass filtered activity (calcium imaging)
- Architecture mismatch
- …

Example: temporal integration of scalar inputs

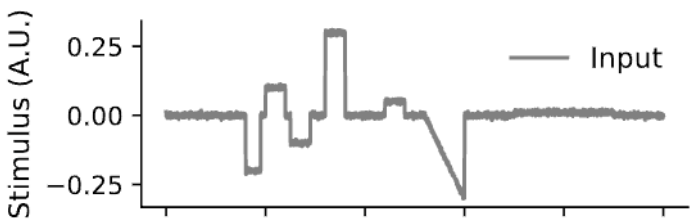


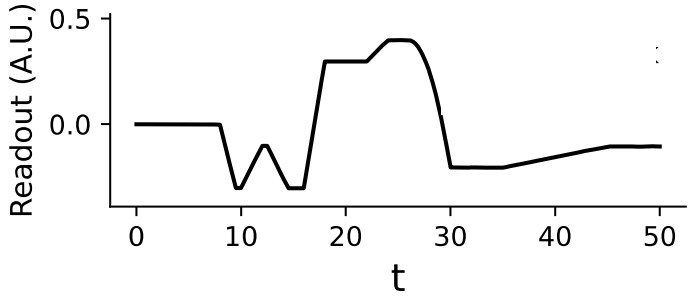
Kaneko CR et al. (1981)

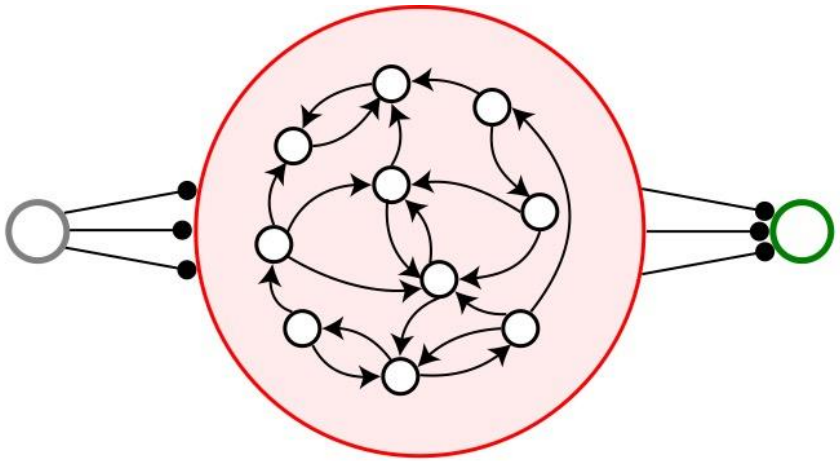
Aksay E et al. (2000)

The oculomotor system

Example: a neural integrator

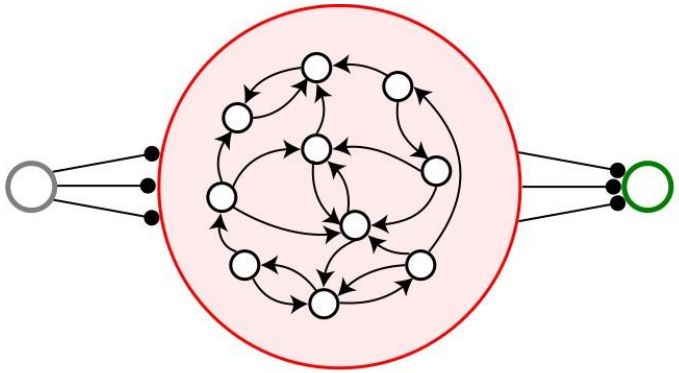


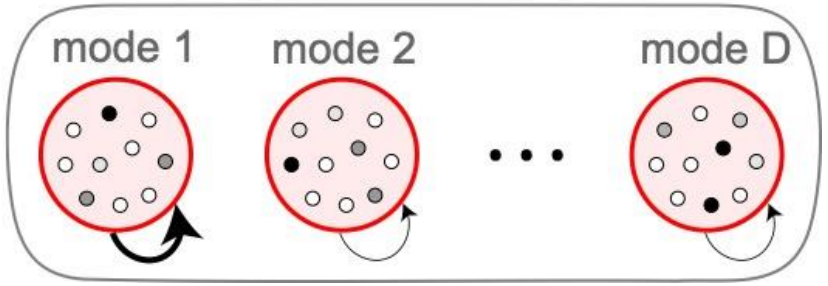




Many ways to construct such a circuit
Cannon et al., 1983; Seung,1996; Koulakov, 2002;
White et al, 2004; Goldman, 2009;

Mechanism 1: line attractor



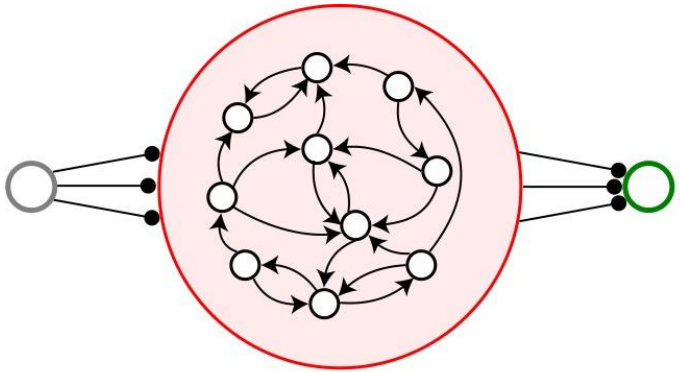


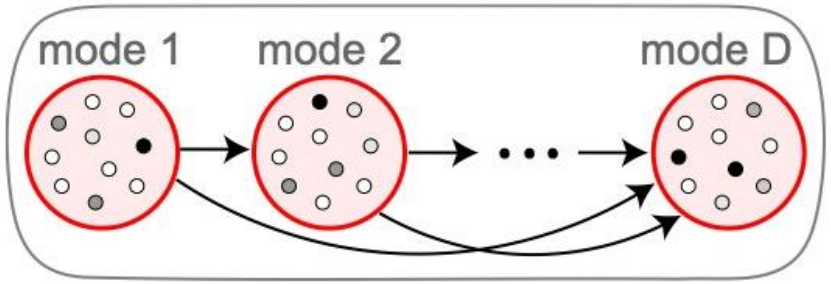
$$W = O\Lambda O^T$$

Top eigenvalue close to one.

Seung,1996

Mechanism 2: feedforward chain





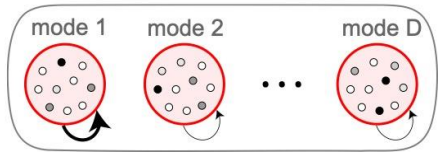
$$W = OUO^T$$

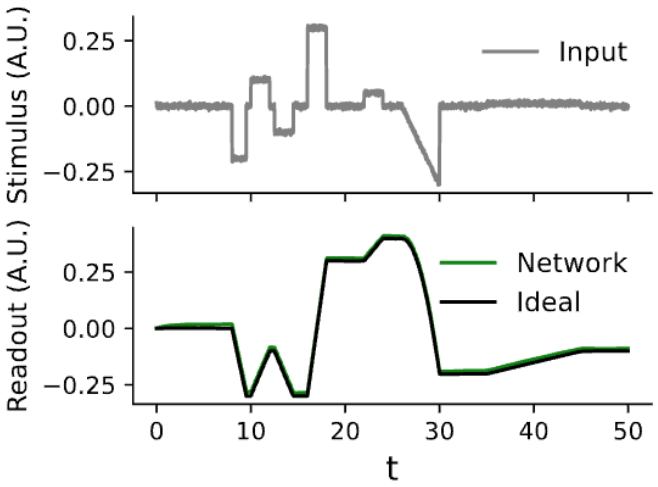
U (strictly) upper triangular
(all eigenvalues equal to zero)

Abeles, 1982; White et al, 2004;
Goldman, 2009;

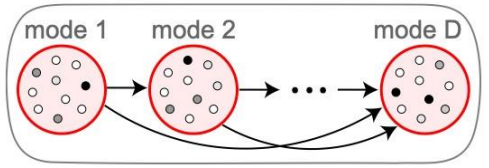
We can construct an integrator using either of these mechanisms

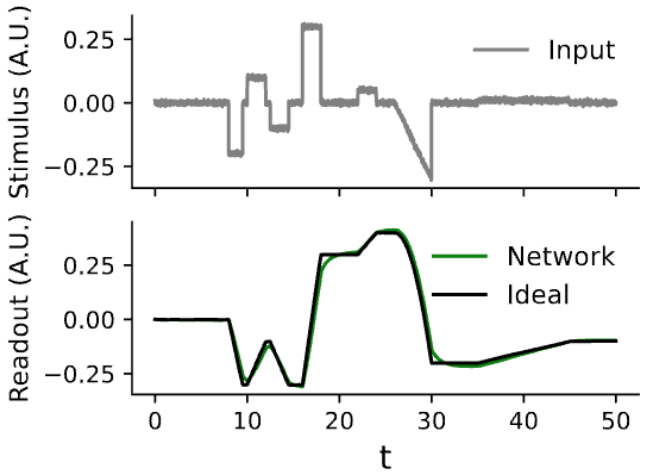
Mechanism 1: line attractor





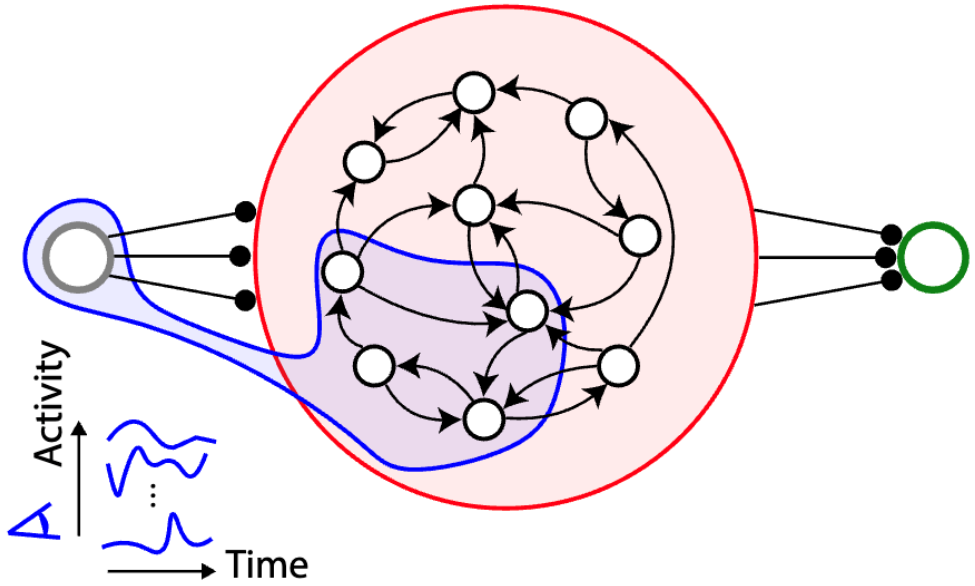
Mechanism 2: feedforward chain



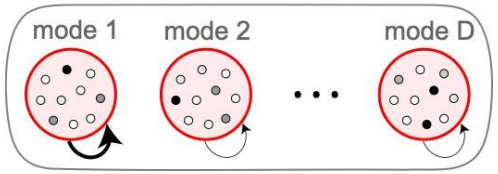


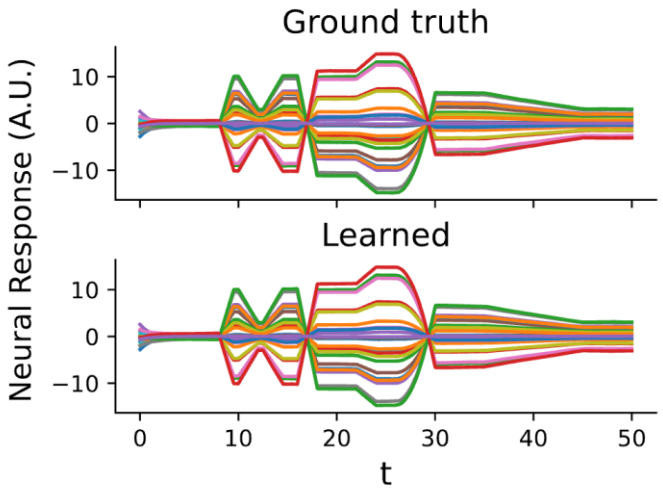
Question: Can we infer the mechanism that the network uses to integrate from “synthetic” neural recordings using the data-driven method?

Record a subset of neurons (5%) and fit a LDS model

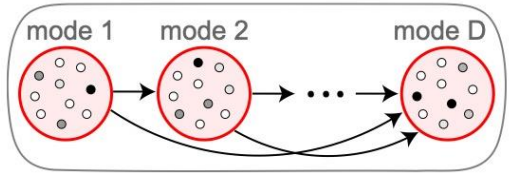


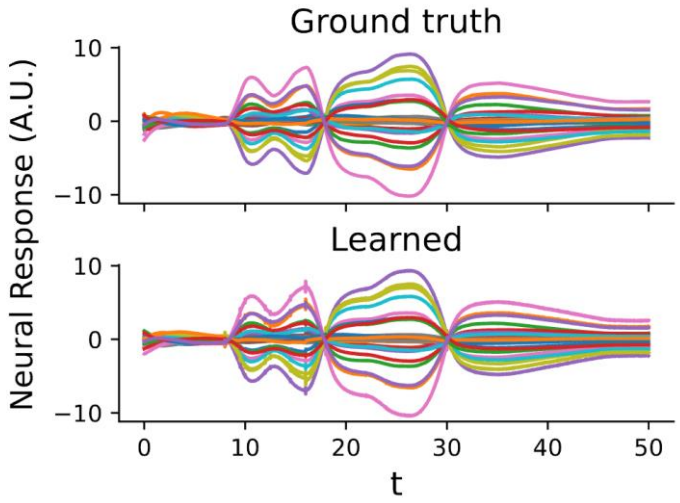
Mechanism 1: line attractor



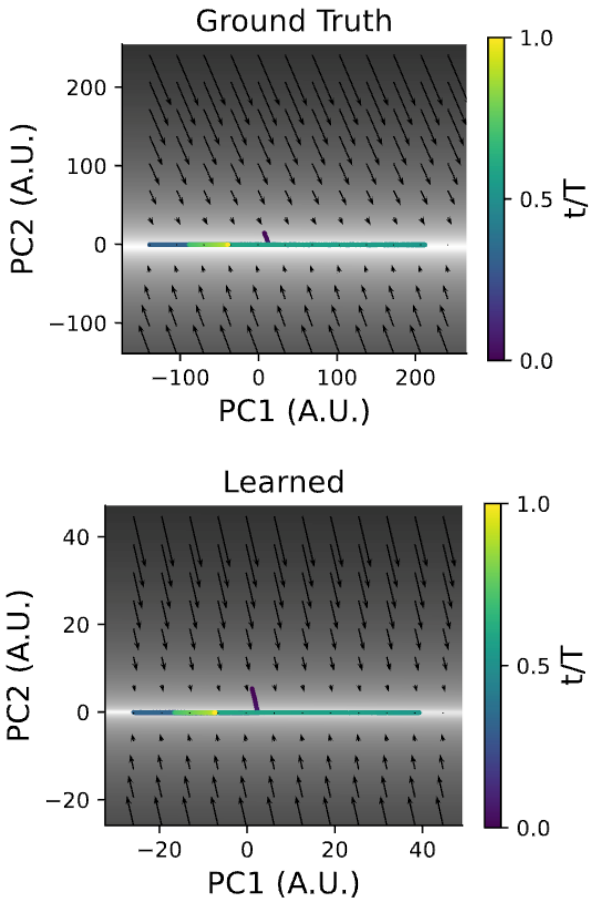


Mechanism 2: feedforward chain

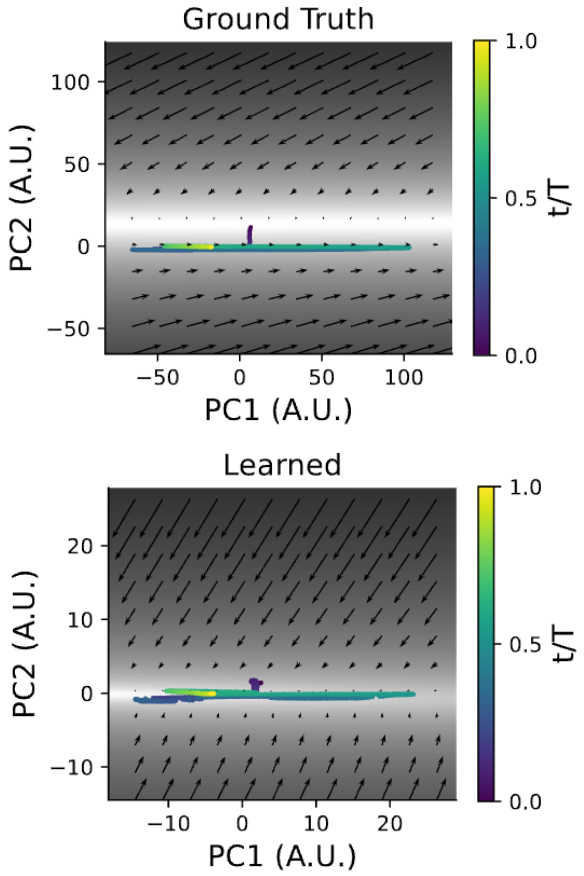




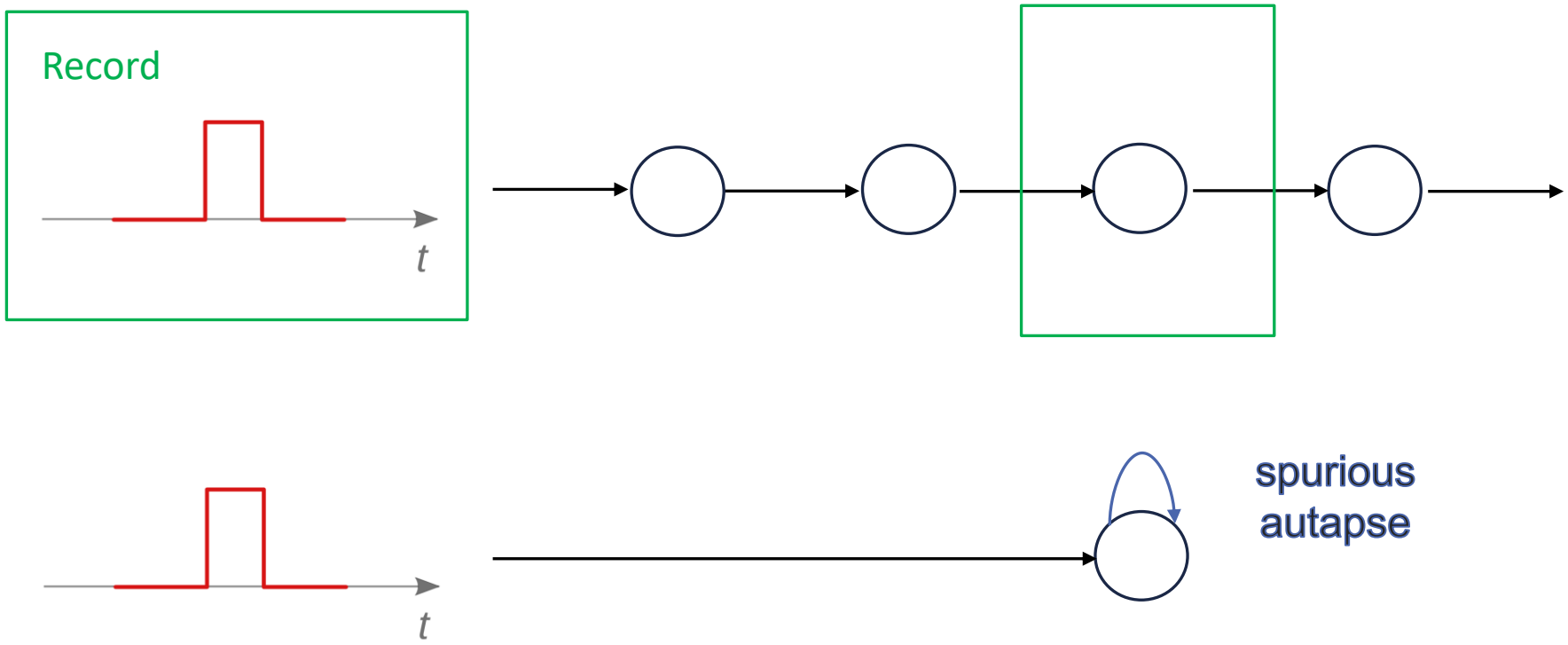
Mechanism 1: line attractor



Mechanism 2: feedforward chain



Why does this happen?

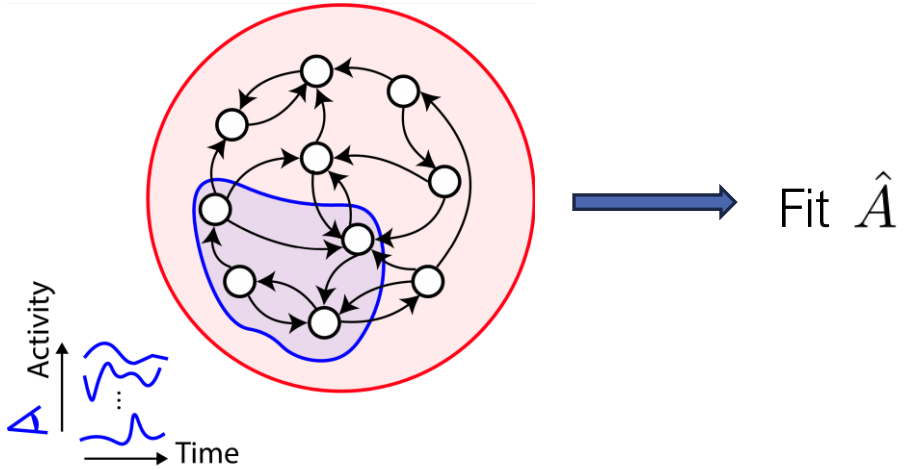


Analytically tractable setting:

Consider a "teacher" $\tau\dot{\mathbf{z}} = -\mathbf{z} + B\mathbf{z} + \boldsymbol{\xi}(t)$

a "student" $\tau\dot{\mathbf{x}} = -\mathbf{x} + A\mathbf{x} + \boldsymbol{\eta}(t)$

and partial observation: $\mathbf{x}^{\text{obs}}(t) = P\mathbf{z}(t)$



Summary of findings:

- (Informal) If *B* (teacher connectivity) is a normal matrix, recovered eigenspectra are qualitatively similar to ground truth
- (Informal) When B is non-normal, student may spuriously estimate large timescales of dynamics

Conclusions

- Latent Dynamical Systems (LDS) models incorrectly identify feedforward integrators performing a stimulus-integration task as line attractors
- Students imitating non-normal teacher dynamics can learn attractors not supported by the teacher, such as spurious line attractors, fixed points, and limit cycles
- See manuscript for the details!

Jacob

Ben

Cengiz Pehlevan