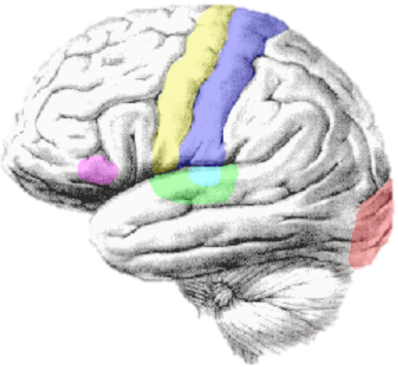
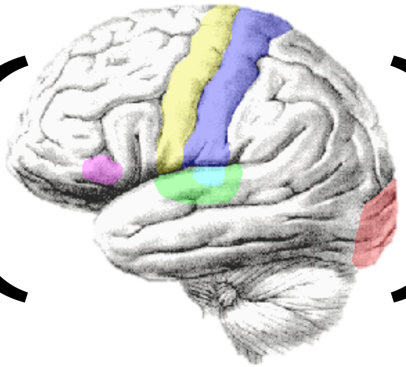
TECHNION Israel Institute of Technology
The Ruth and Bruce Rappaport Faculty of Medicine
Science and Engineering of Neural Systems
The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering
NETWORK BIOLOGY RESEARCH LABORATORIES

Learning from learning systems

Omri Barak

October, 2024

Models of the brain

$\frac{d}{dt}$  $= f\Big($  $\Big)$

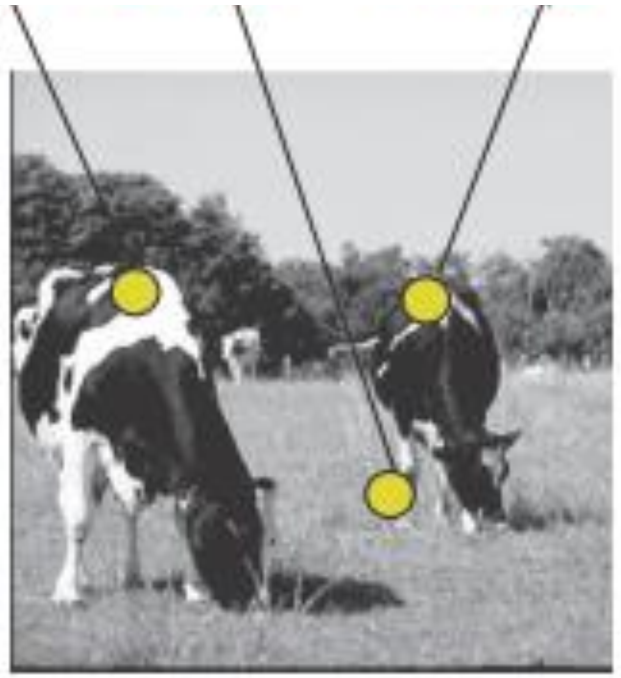
Models should be simple

- Simple models often explain a component of the system.
- Physical or functional
- The hope is that when it is embedded back into the system, functionality is retained.

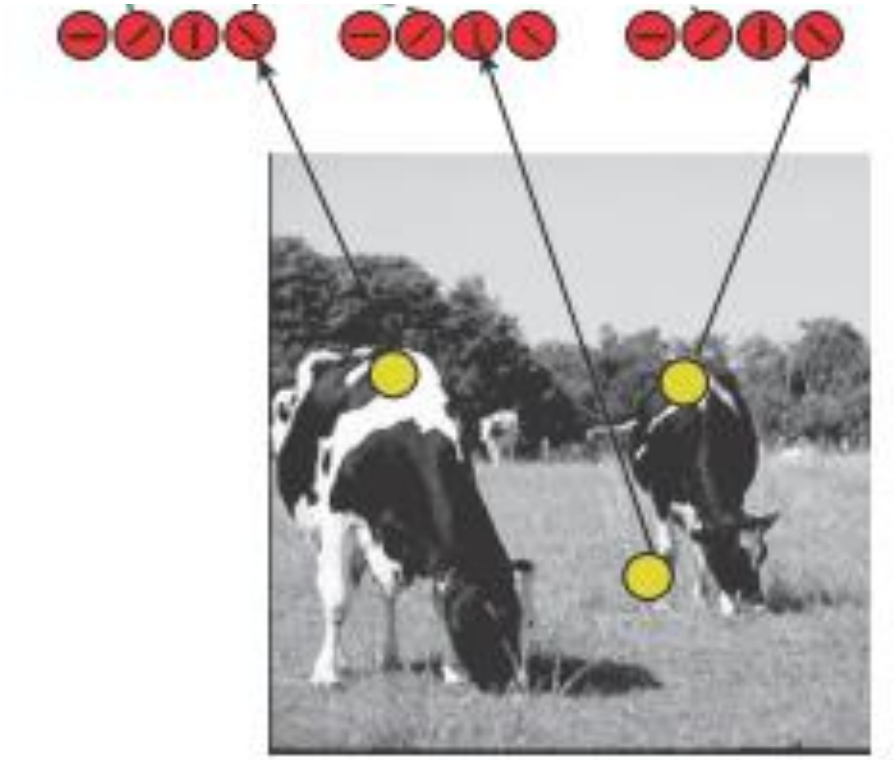
INTERIOR OF VAUCANSON'S AUTOMATIC DUCK.
A, clockwork; *B*, pump; *C*, mill for grinding grain; *F*, intestinal tube; *J*, bill; *H*, head; *M*, feet.

Descartes, 1662

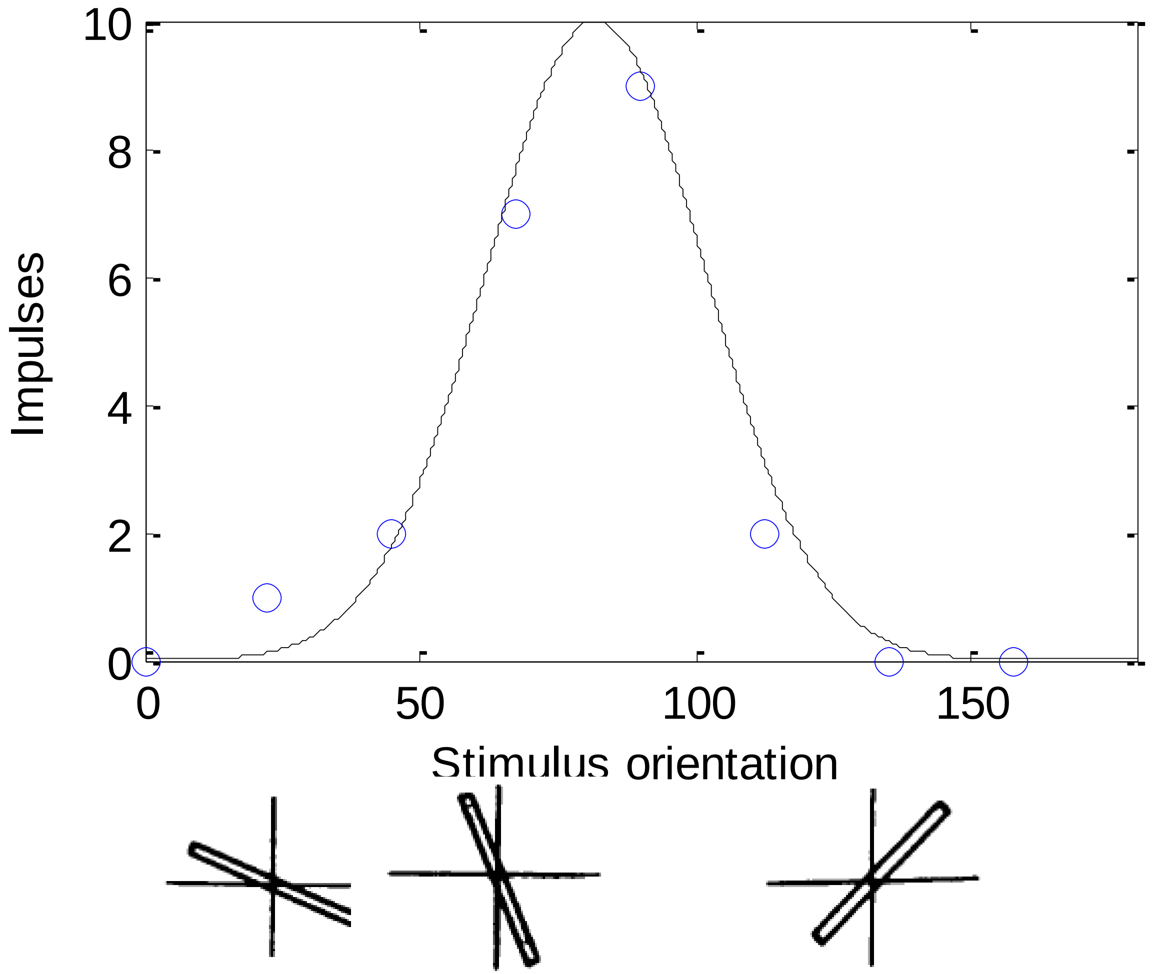
Let’s recognize a cow!



A cow is made of lines

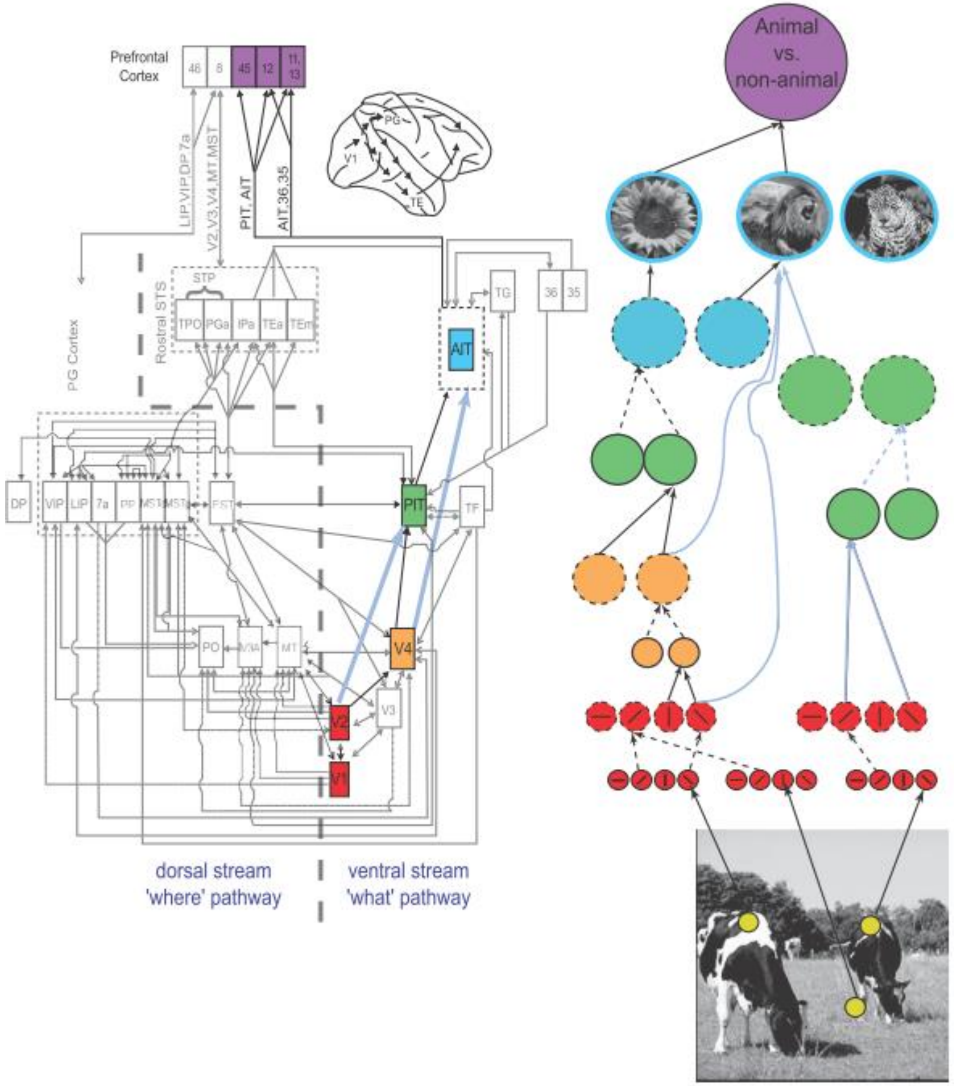


V1 cells like lines!



Primary visual cortex
Hubel and Wiesel, 1959

Oriented lines build images

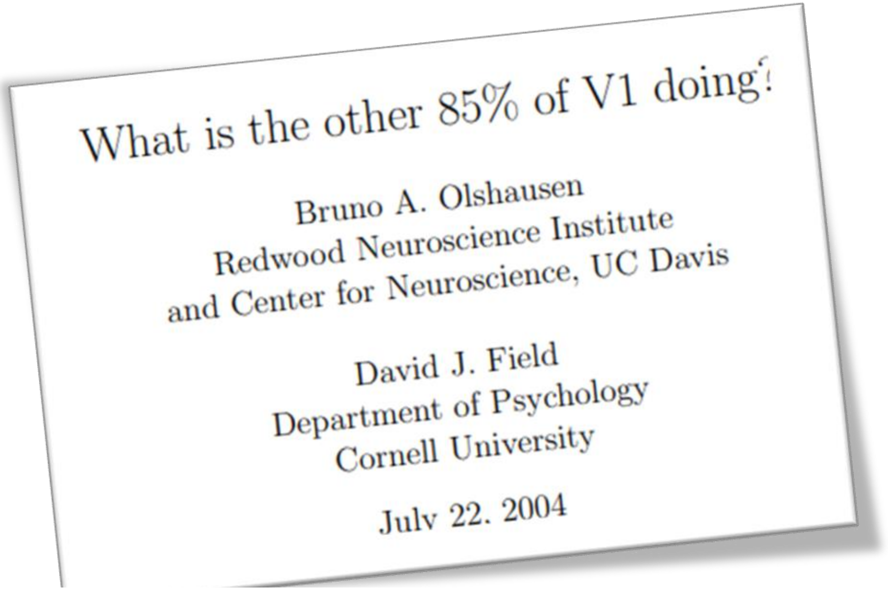


Serre, Oliva, Poggio 2007

Is functionality retained?

- Simple models often explain a component of the system.
- Physical or functional
- **The hope is that when it is embedded back into the system, functionality is retained.**

No

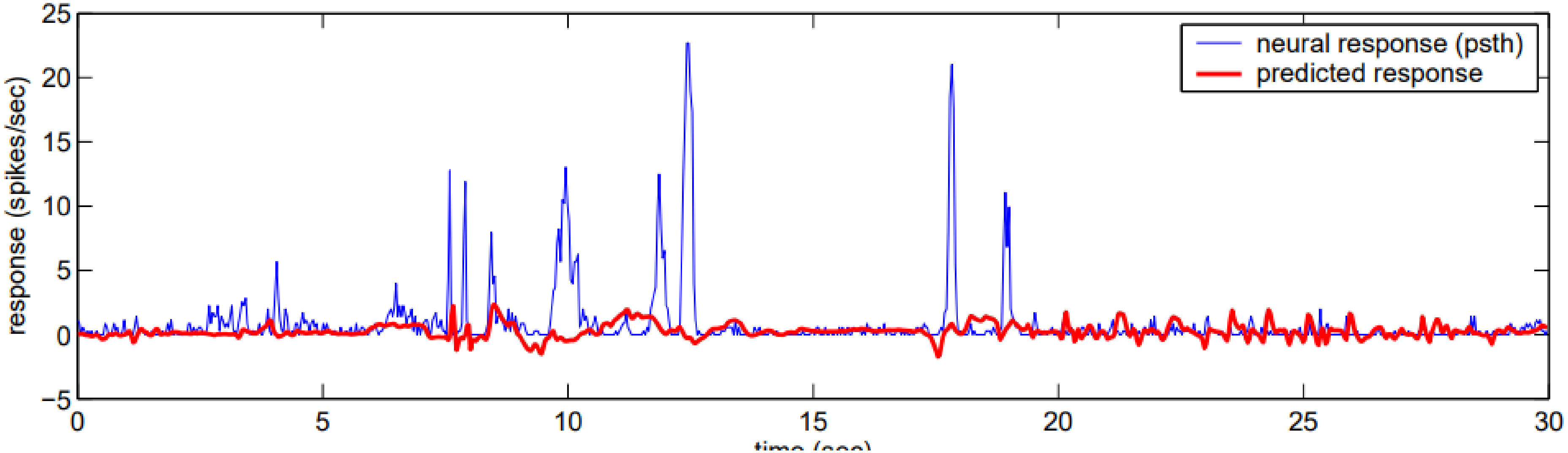


What is the other 85% of V1 doing?

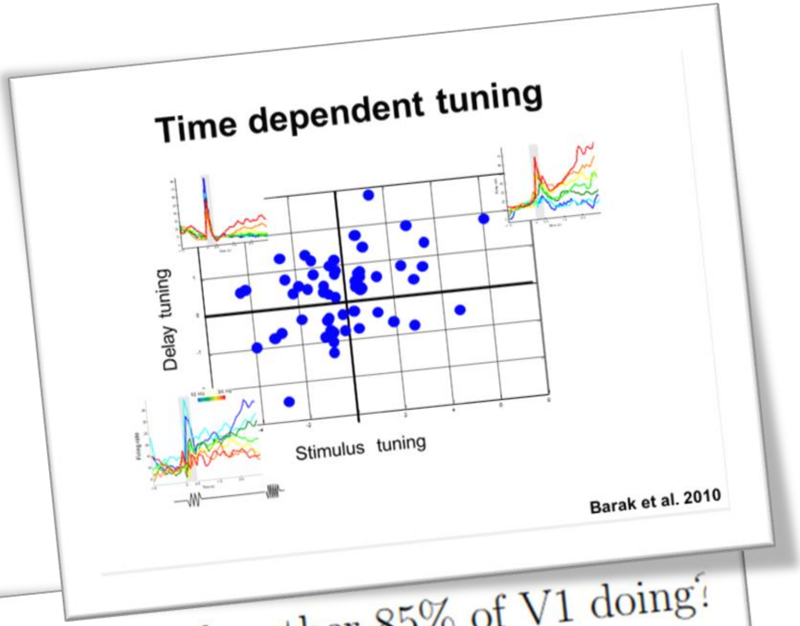
Bruno A. Olshausen
Redwood Neuroscience Institute
and Center for Neuroscience, UC Davis

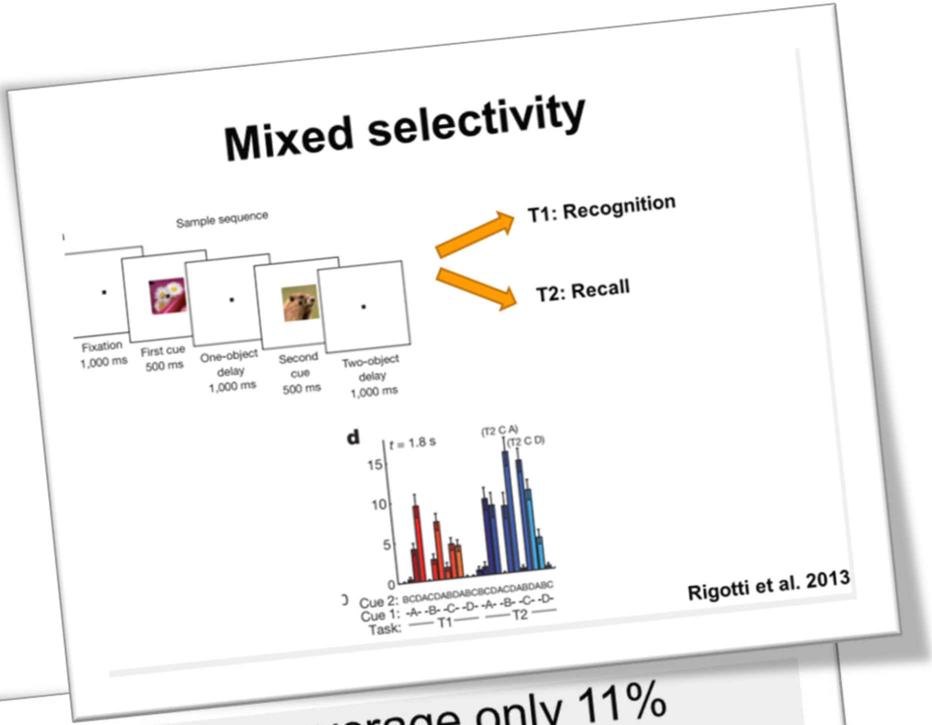
David J. Field
Department of Psychology
Cornell University

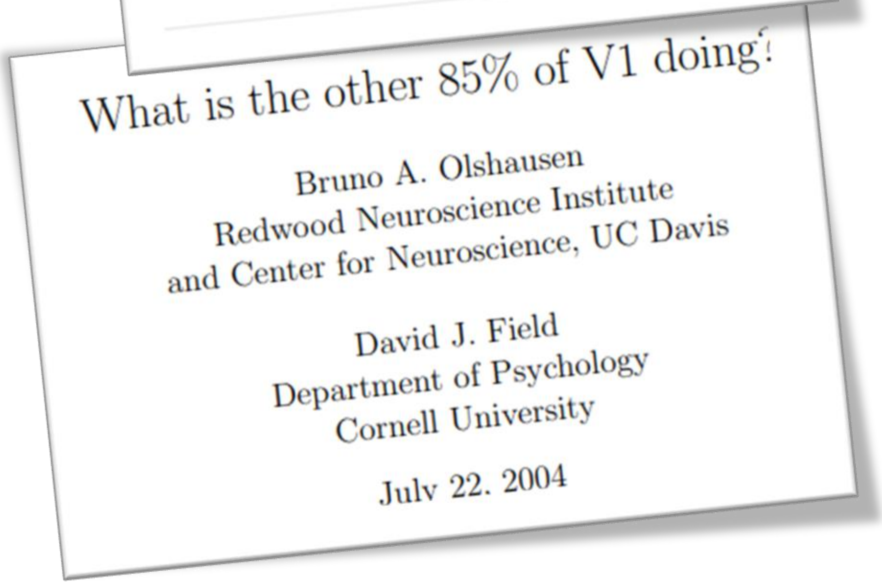
July 22, 2004



No



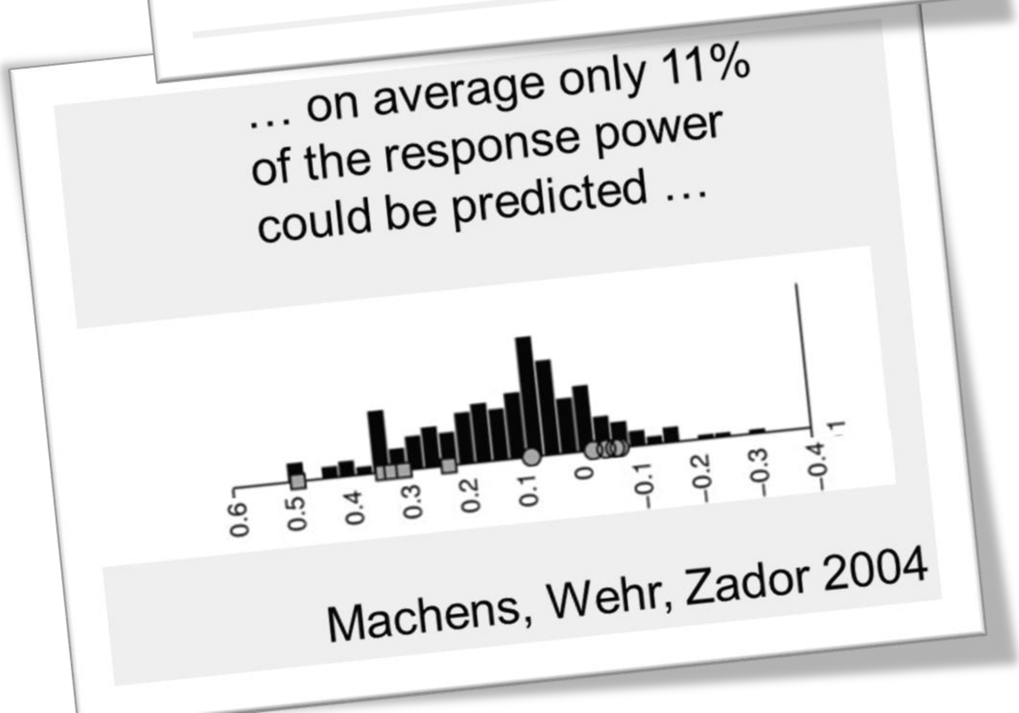



What is the other 85% of V1 doing?

Bruno A. Olshausen
Redwood Neuroscience Institute
and Center for Neuroscience, UC Davis

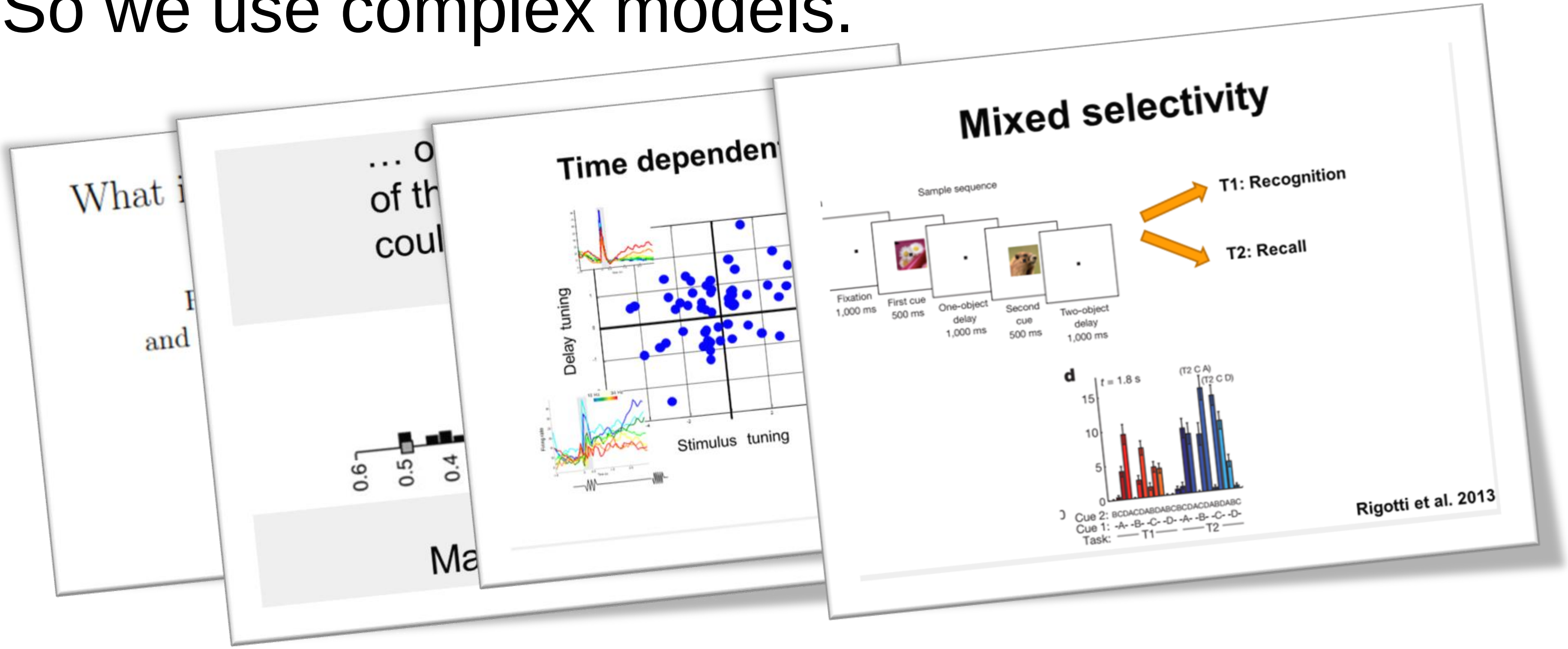
David J. Field
Department of Psychology
Cornell University

July 22, 2004



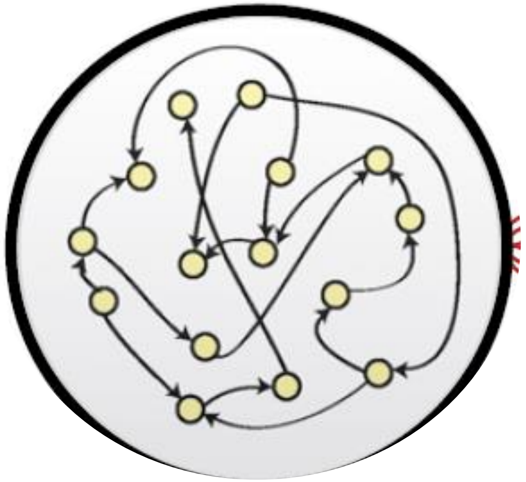
Complex data – complex models

- We would like simple models.
- The brain is not playing nice.
- So we use complex models.



Machine learning as a model

Model



Millions of parameters
Trained to match
input-output pairs

Experiment

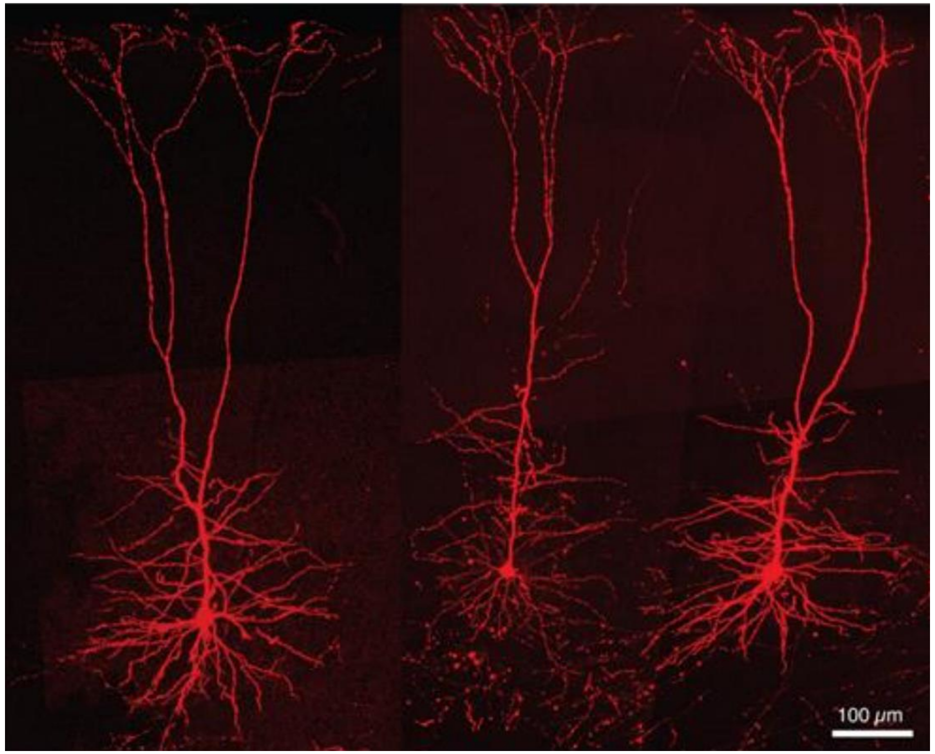
Requirement for a model system

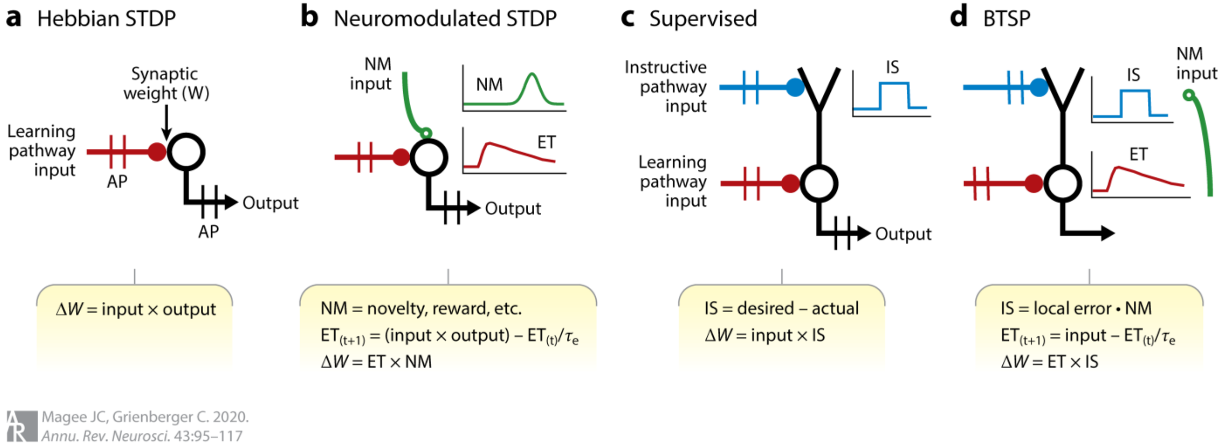
- Common features with target system
- Easier to manipulate / understand

Common features

$$\frac{dx_i}{dt} = -x_i + \sum_{j=1}^{N} W_{ij}\phi(x_j) + V_i u(t)$$

$$\frac{dW}{dt} = -\nabla_W L(W)$$



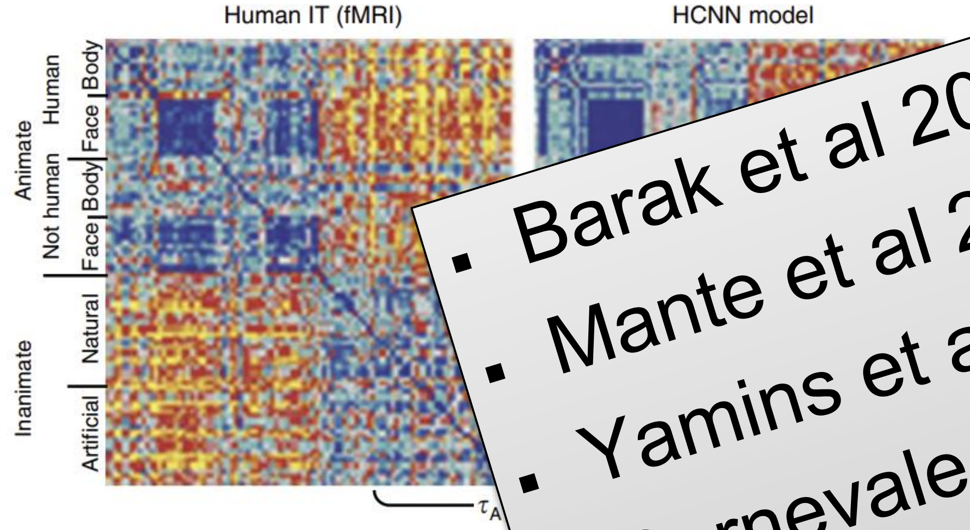


Otor et al. Science 2022

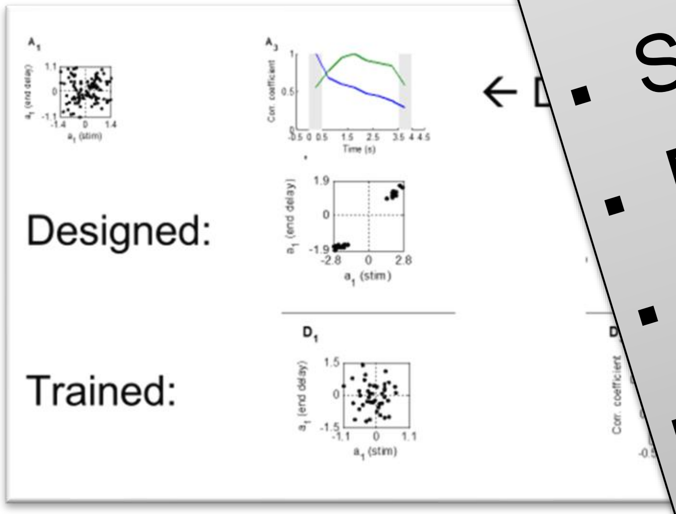
Magee et al. Ann rev neuro 2020

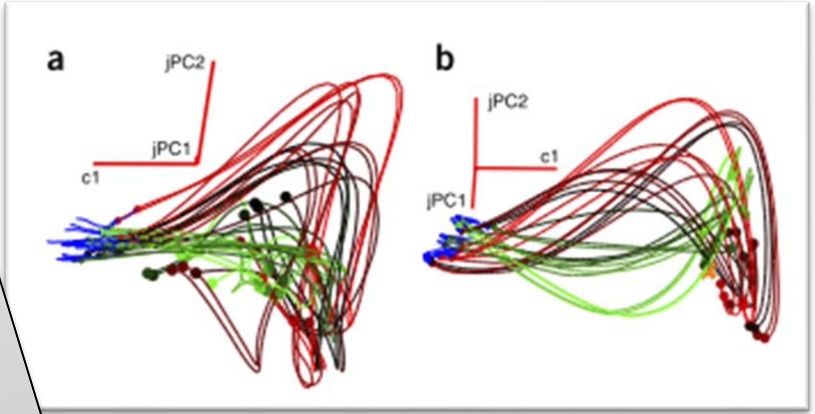
Common features

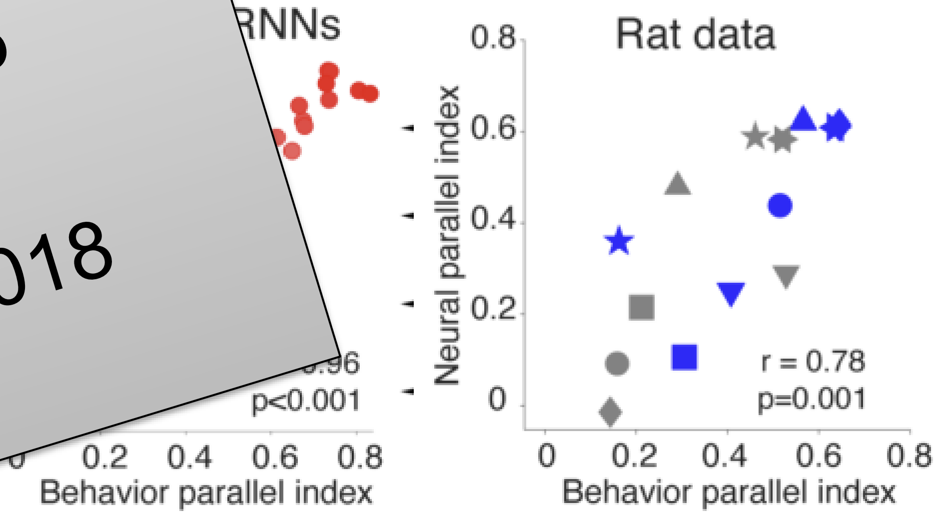
- Barak et al 2013
- Mante et al 2013
- Yamins et al 2014
- Carnevale et al 2015
- Song et al 2016
- McIntosh et al. 2016
- Russo et al 2018
- Remington et al 2018
- And more…



Yamins, DiCarlo 201







al 2013

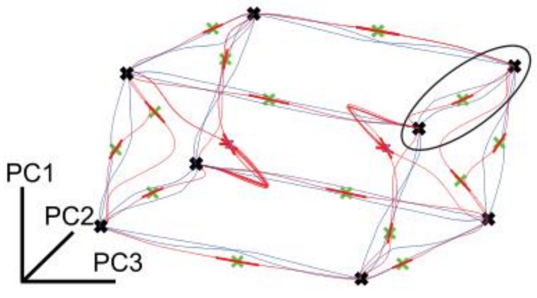
Pagan et al 2022

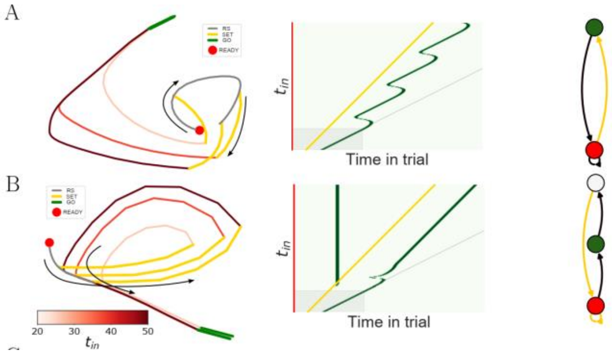
Requirement for a model system

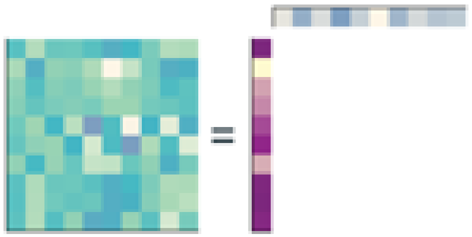
- Common features with target system
- Easier to manipulate / understand

Easier to manipulate/understand

- Fixed points
  - Sussillo & Barak 2013
  - Katz & Reggia 2017
- Graphs from trajectories
  - Turner et al 2021
  - Brennan et al 2023
- Low rank connectivity
  - Mastroguiseppe & Ostojic 2018
  - Schuessler et al 2020







Requirement for a model system

- Common features with target system
- Easier to manipulate / understand

Why does it work?

- Maybe the model is not as far from reality as we think
- Perhaps this similarity is because both are instances of complex systems that adapt to their environment.
- Learning from learning systems

Why does it work?

- Maybe the model is not as far from reality as we think
- Perhaps this similarity is because both are instances of complex systems that adapt to their environment.
- Learning from learning systems

Does it work?

- How to compare model and data?
- How to analyze models?
- What do we learn from this?
- Biological plausibility – should we care? How?

Three “short” stories

- **Representational drift, learning and implicit regularization**
  - Ratzon et al. eLife 2024
- **Explain vs. predict**
  - Dabholkar et al. Arxiv 2024
- **Most variance is not task-related**
  - Schuessler et al. eLife 2024

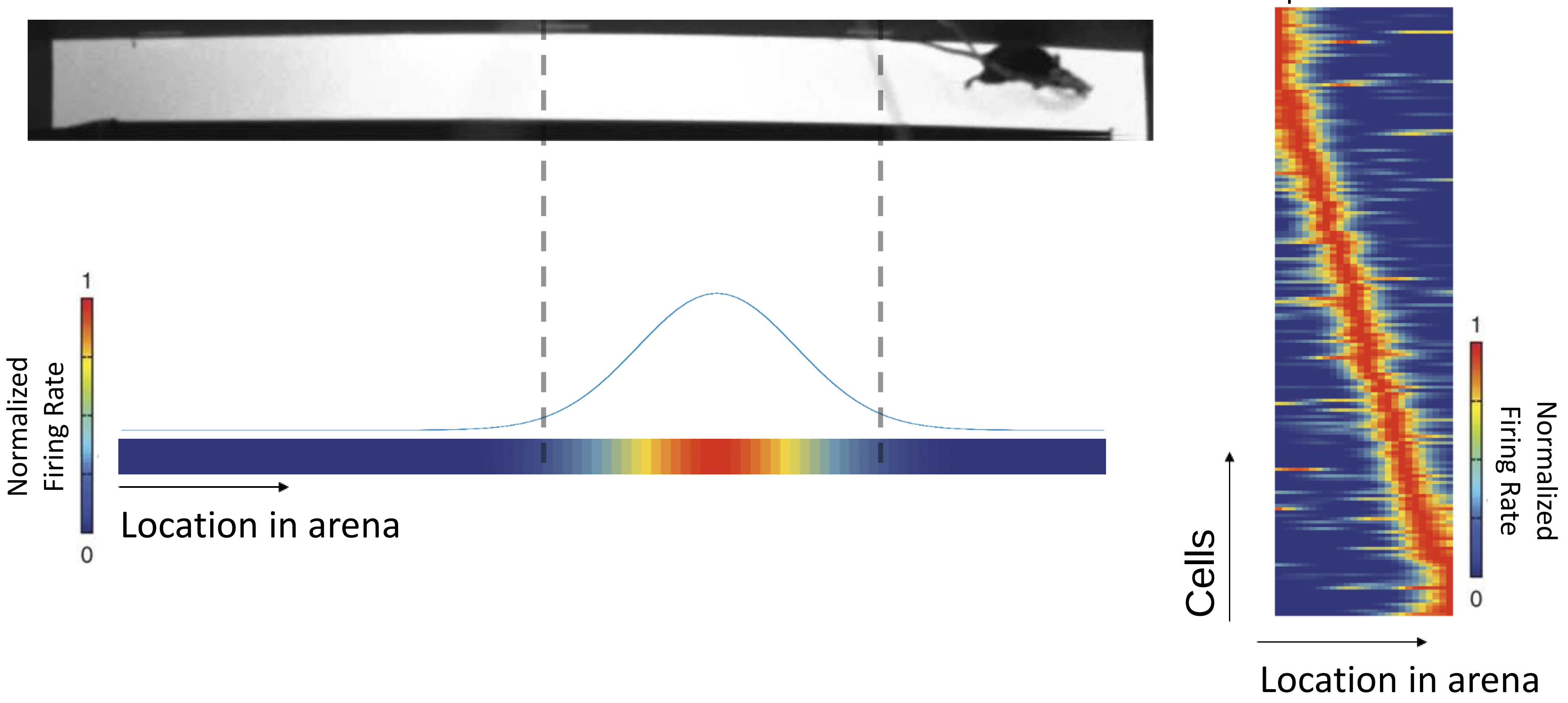
Drift vs learning

Aviv Ratzon

Dori Derdikman

Place Cells Are Spatially Tuned Neurons

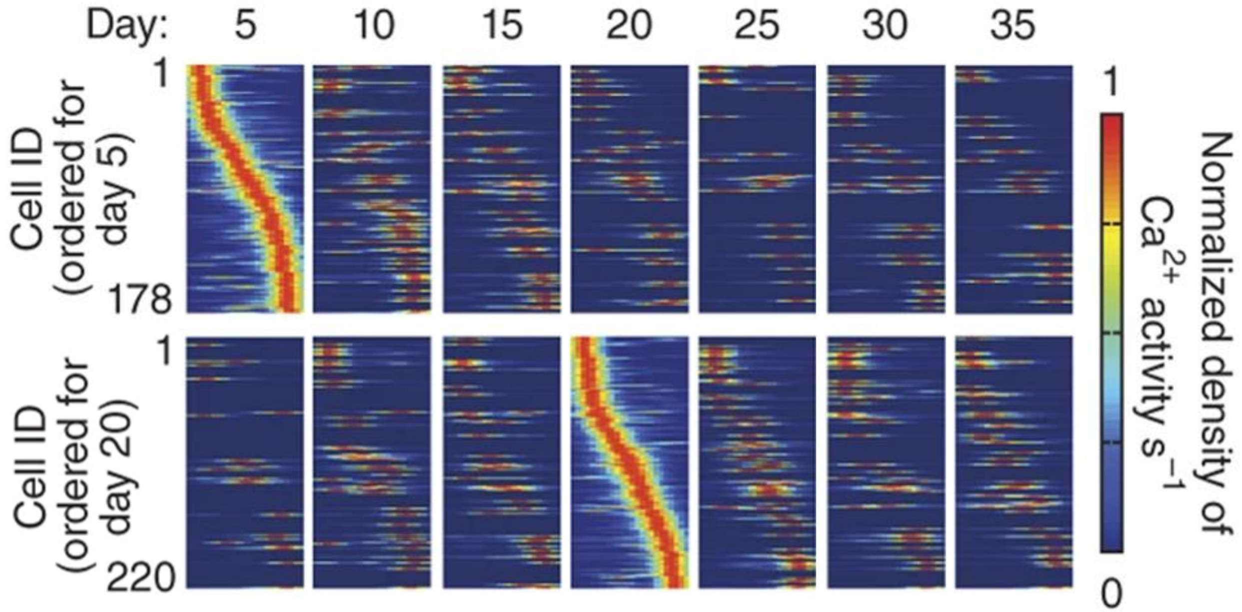
O'keefe & Dostrovsky (1971)



Rate maps change over time

Ziv et al. (2013)

Hippocampus CA1



- **Visual Cortex -** *Deitch et al. (2020)*
- **Olfactory Cortex -** *Schoonover et al. (2021)*
- **Parietal Cortex -** *Driscoll et al. (2017)*

Learning vs. drift

Learning	Drift
New behavior	Fixed behavior
Transient	Persistent
Directed	Undirected
Useful	Detrimental?

Hypothesis:

Manifestations of the same process

Framework

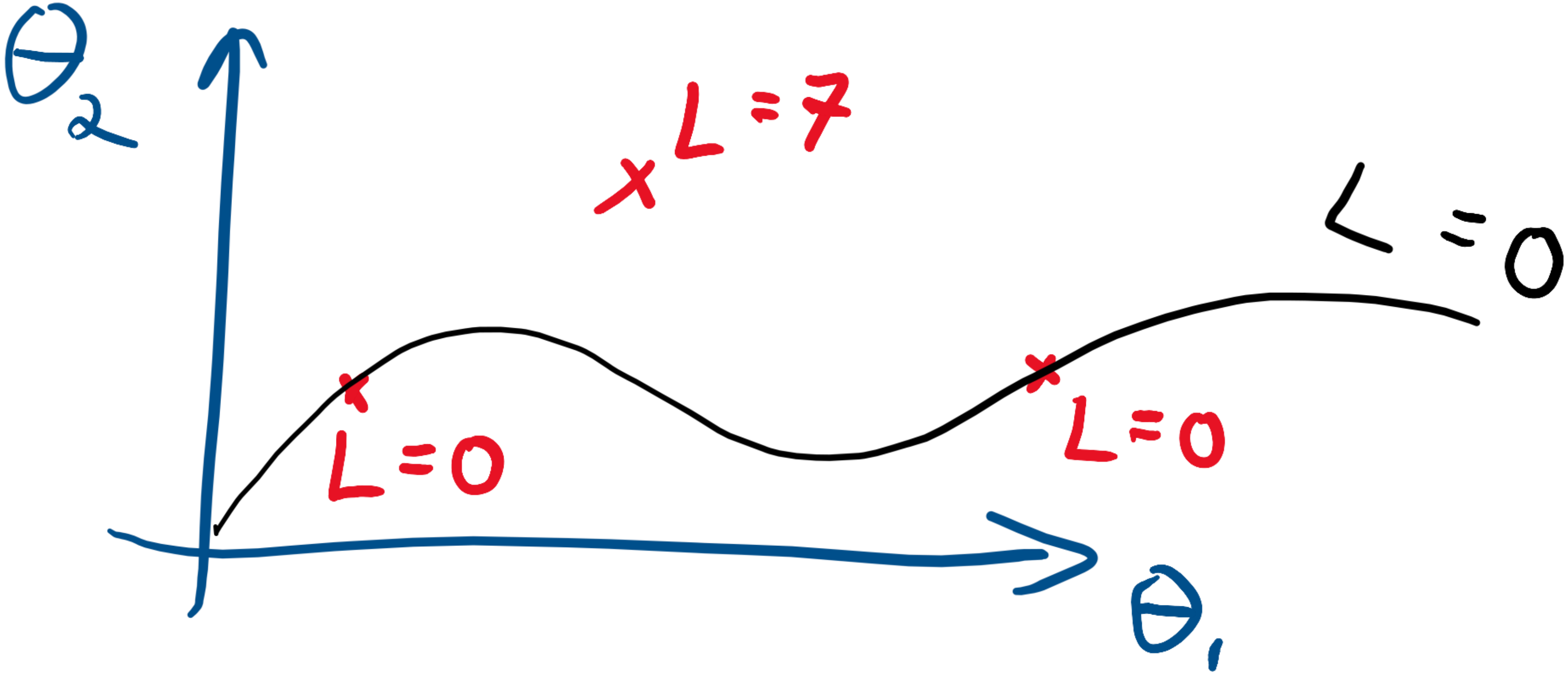
- Parameters: synapses, excitability, … $\theta_1, \theta_2, \ldots$
- Loss: measure of performance. L

- Loss is a function of parameters. $L(\theta)$
- Learning reduces loss by changing parameters.

Degeneracy

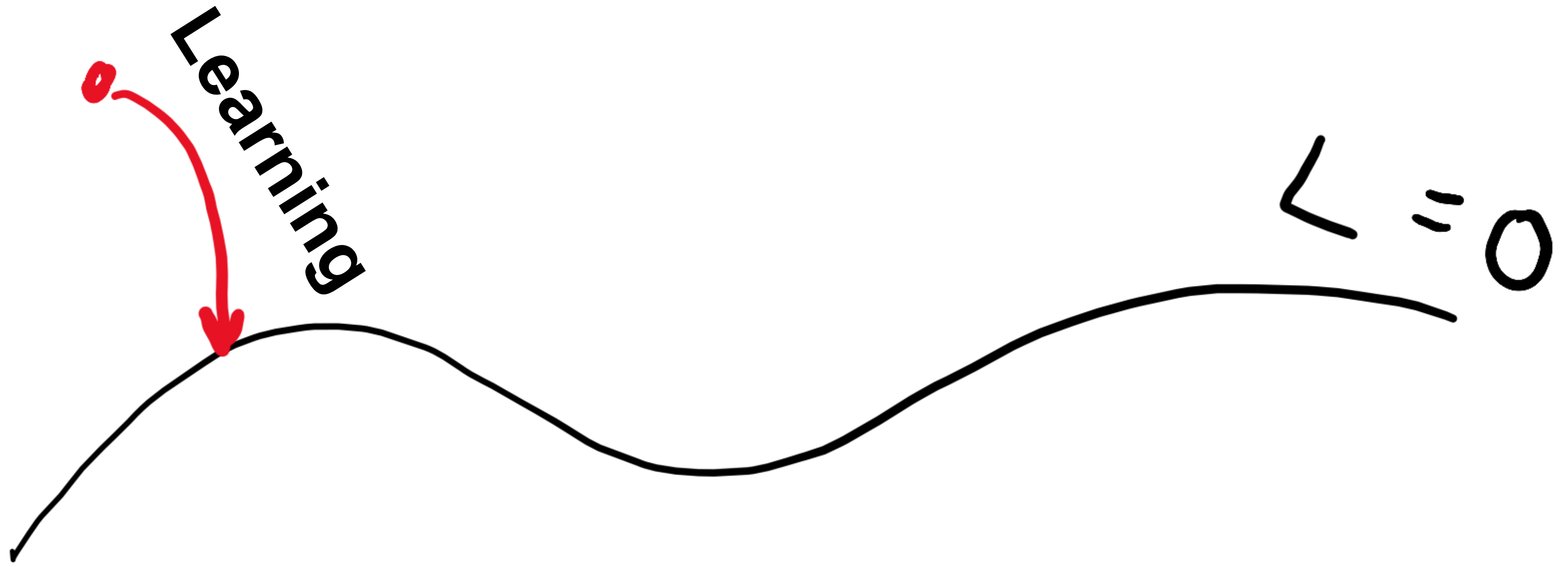
- Many parameter configurations
- Corresponding to neural representations
- Could lead to same behavior

Space of parameters

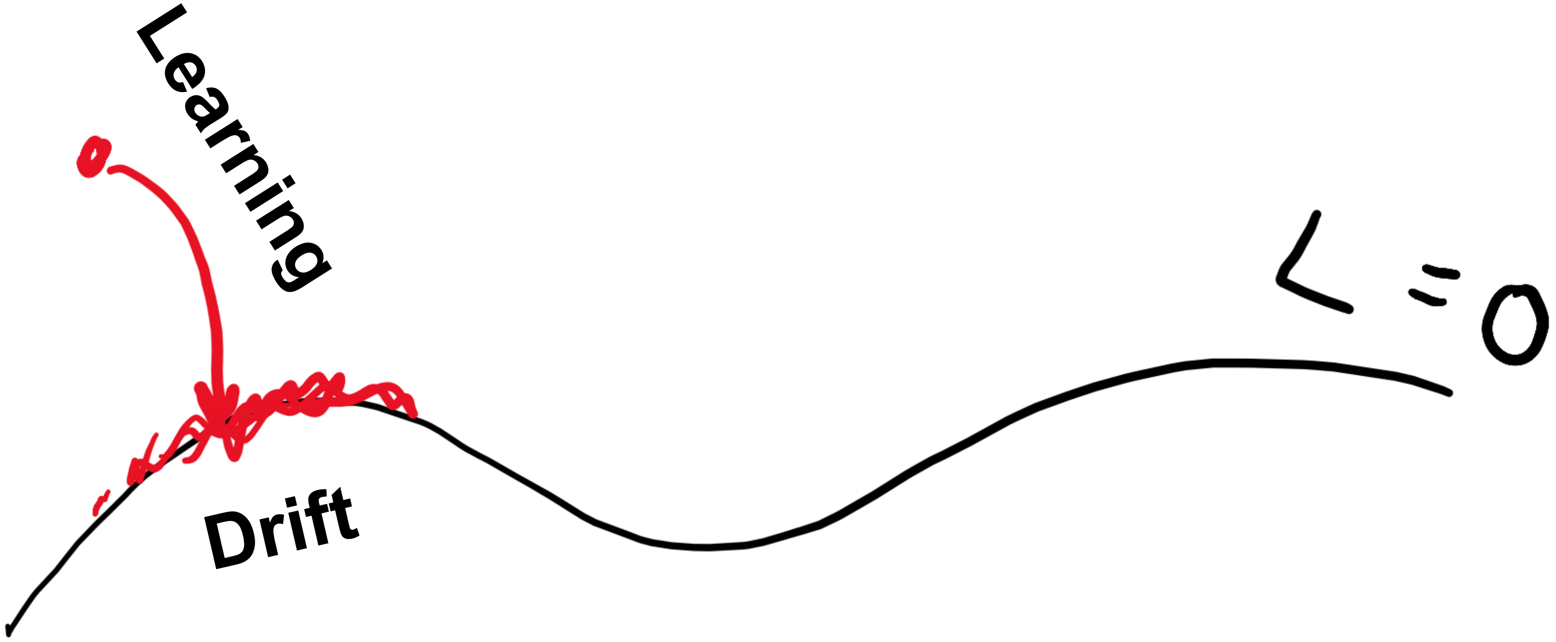


Space of parameters

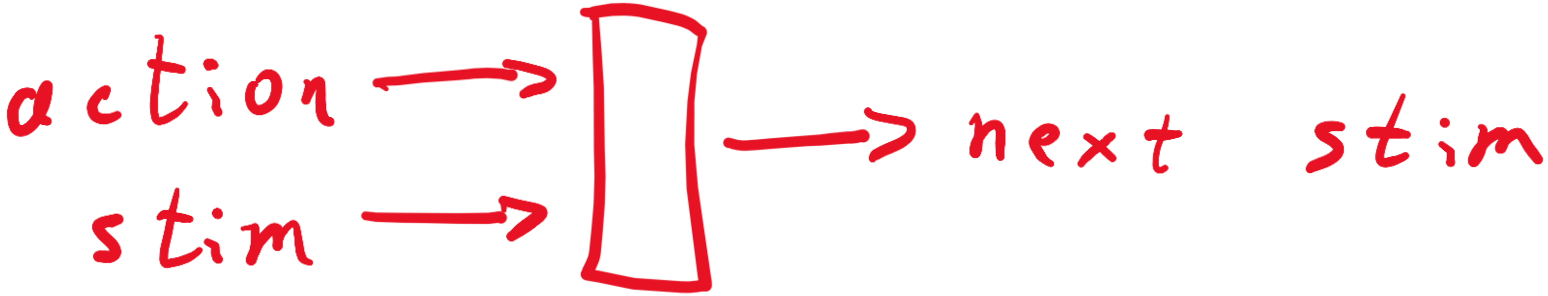
Learning reduces loss

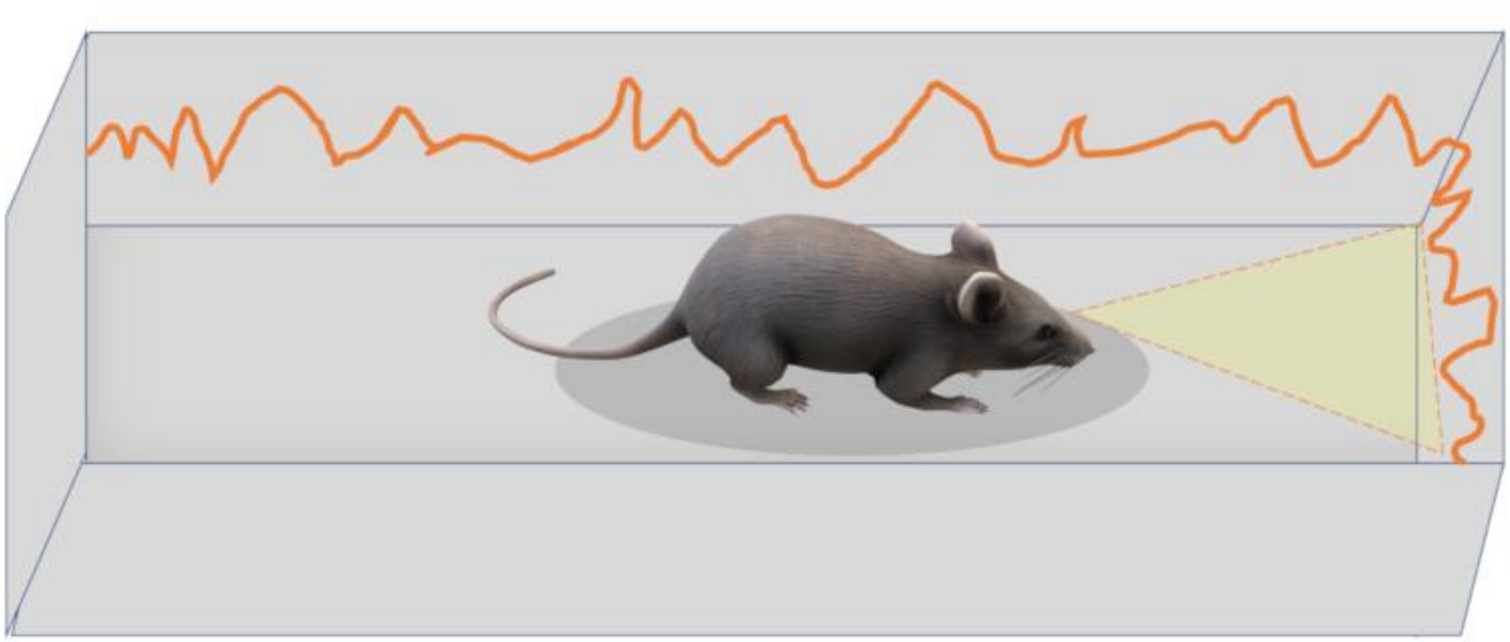


What happens if we never stop learning?

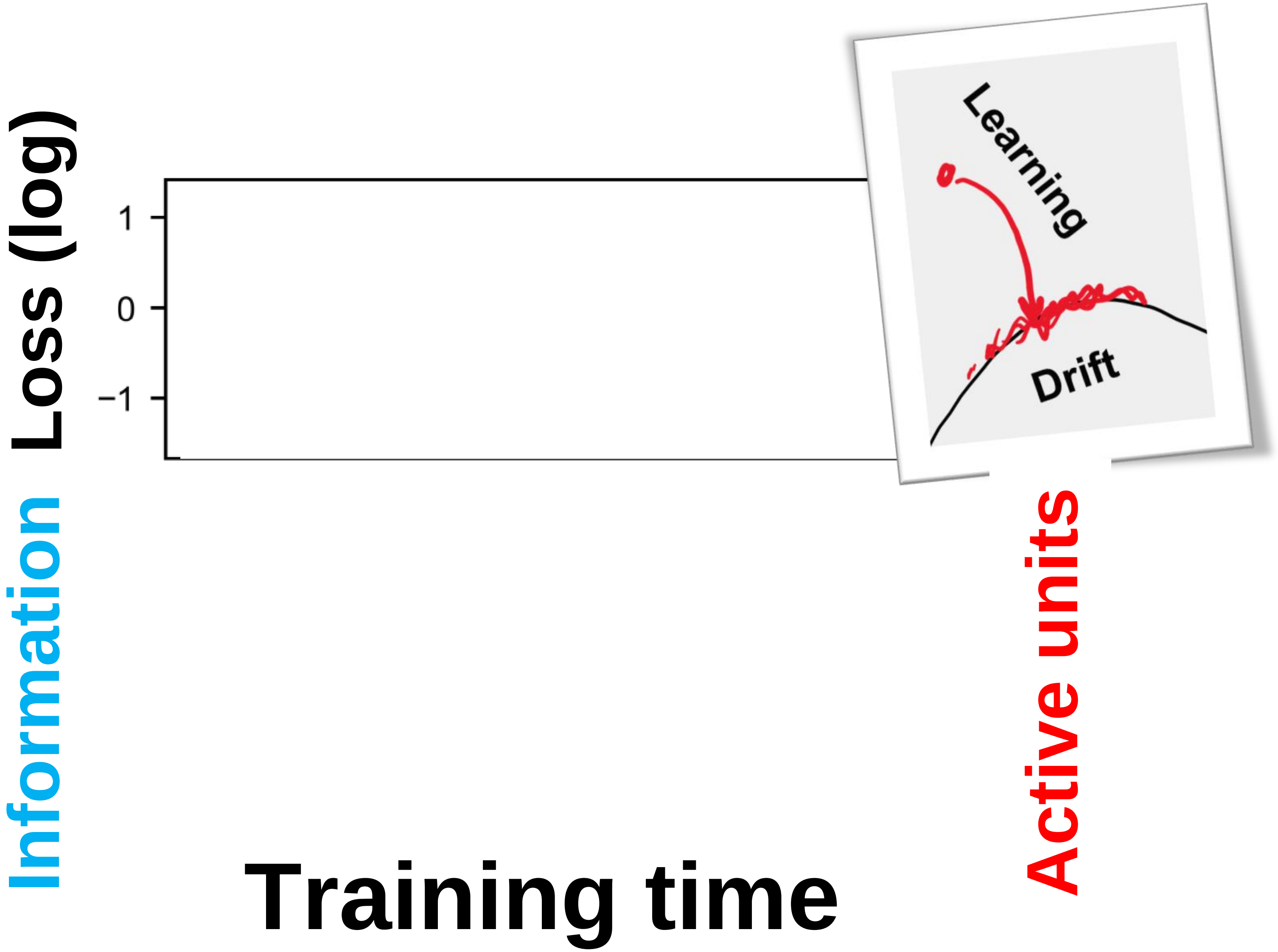


Simple model

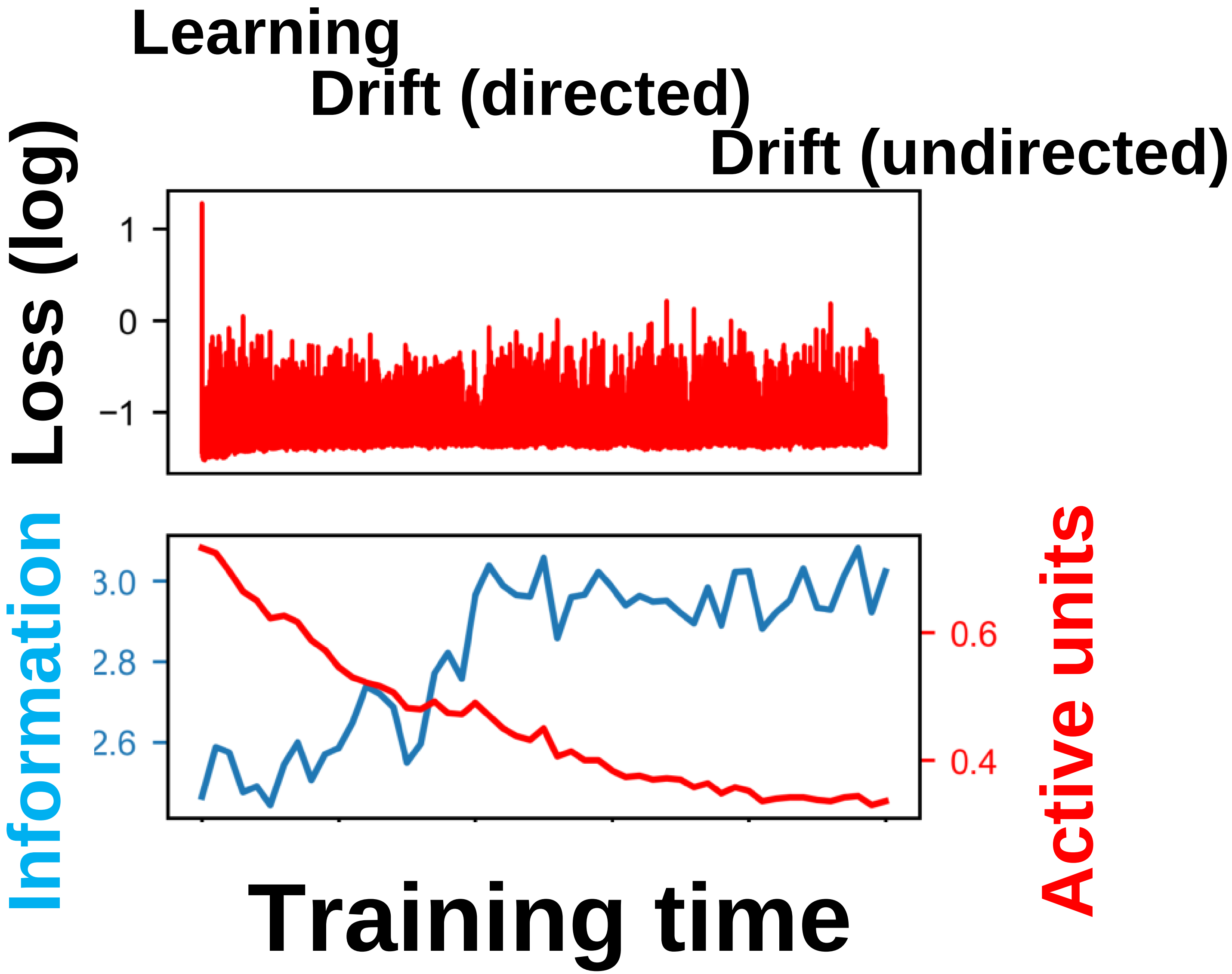




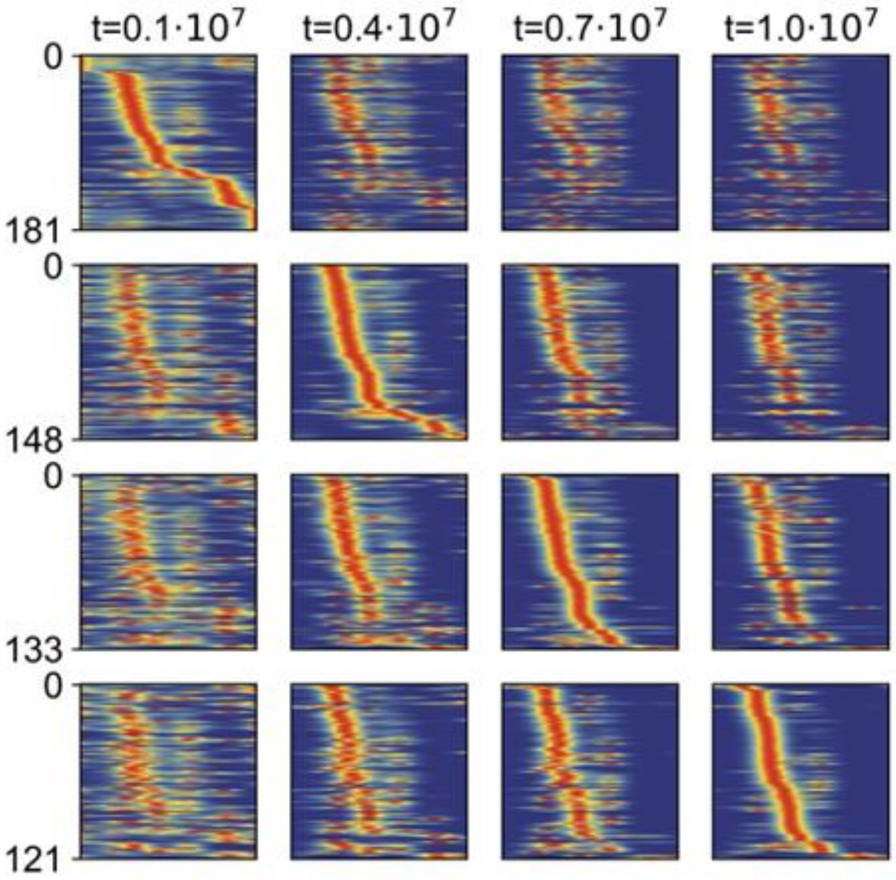
Never stop learning

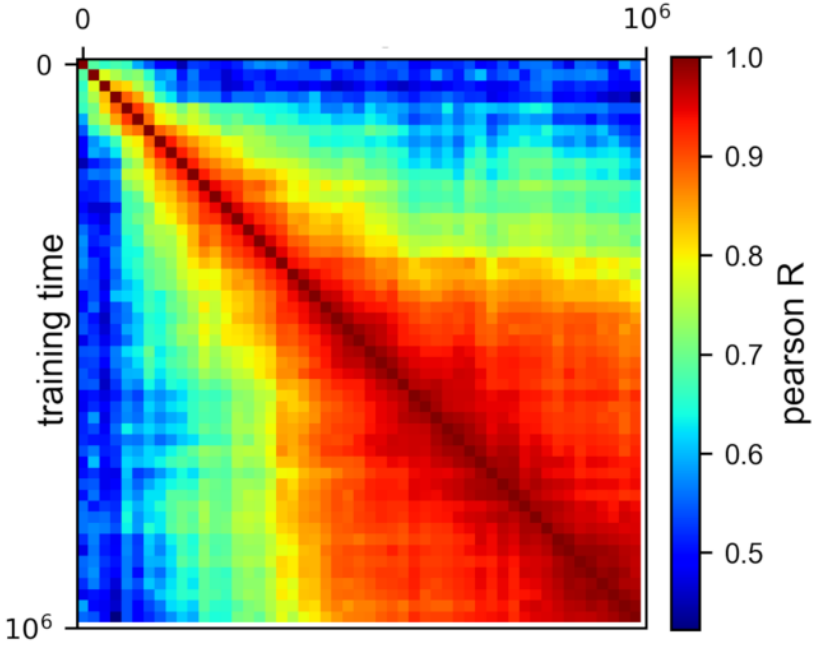


Three phases



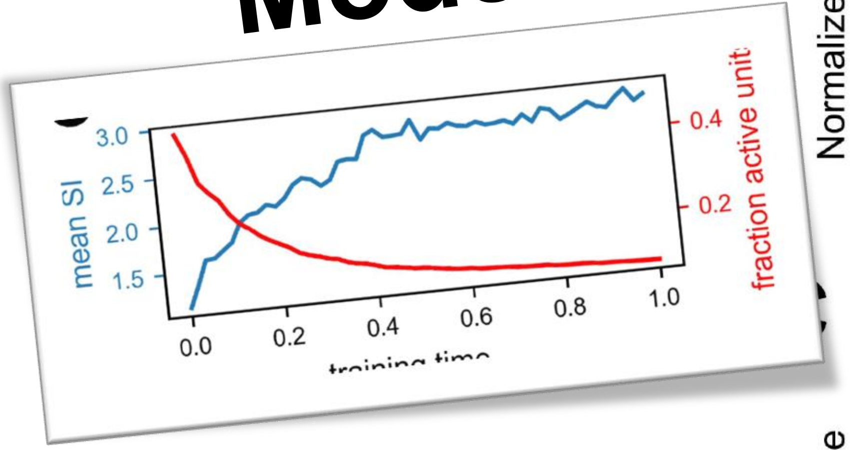
Representational drift



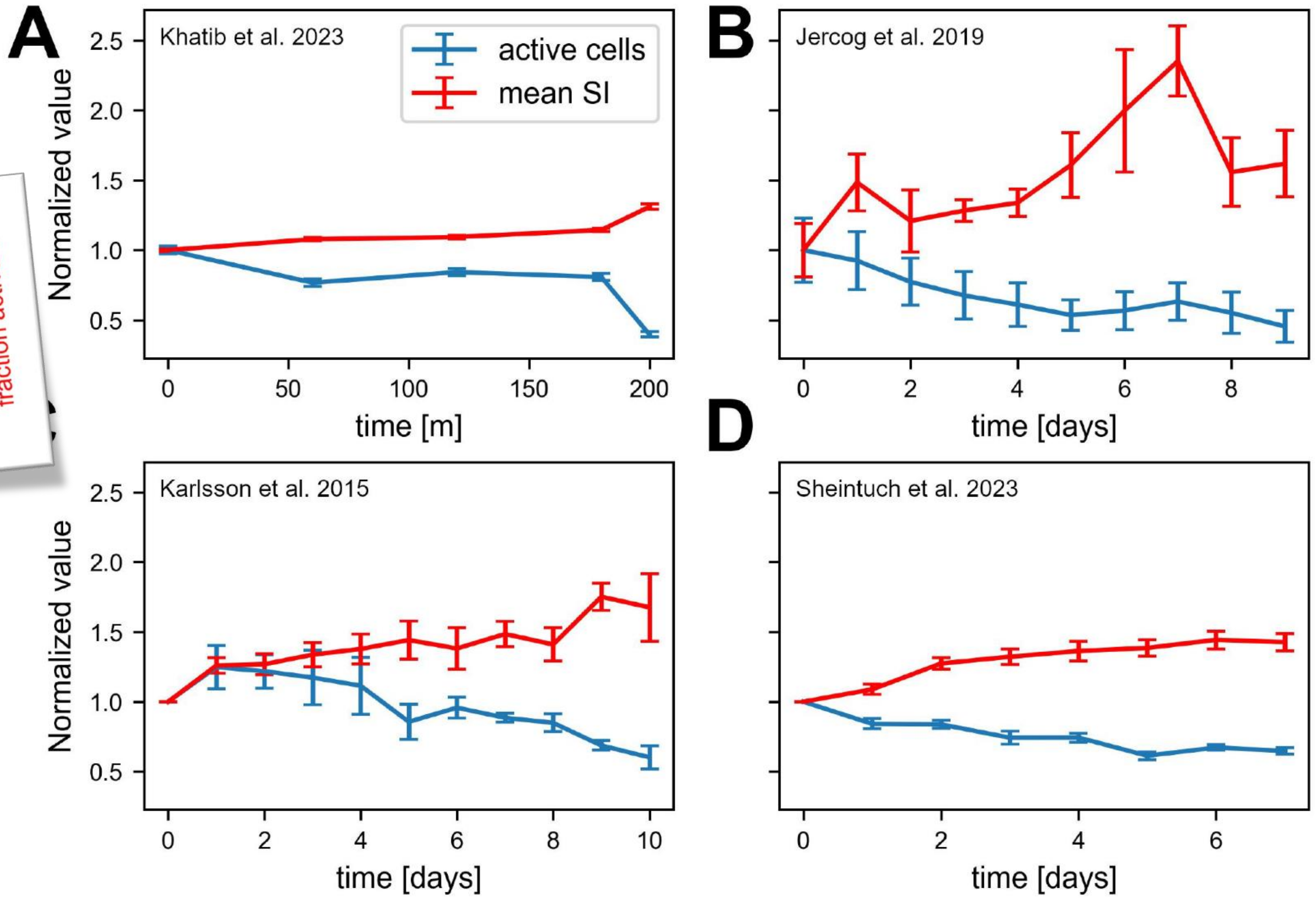


Data

Model

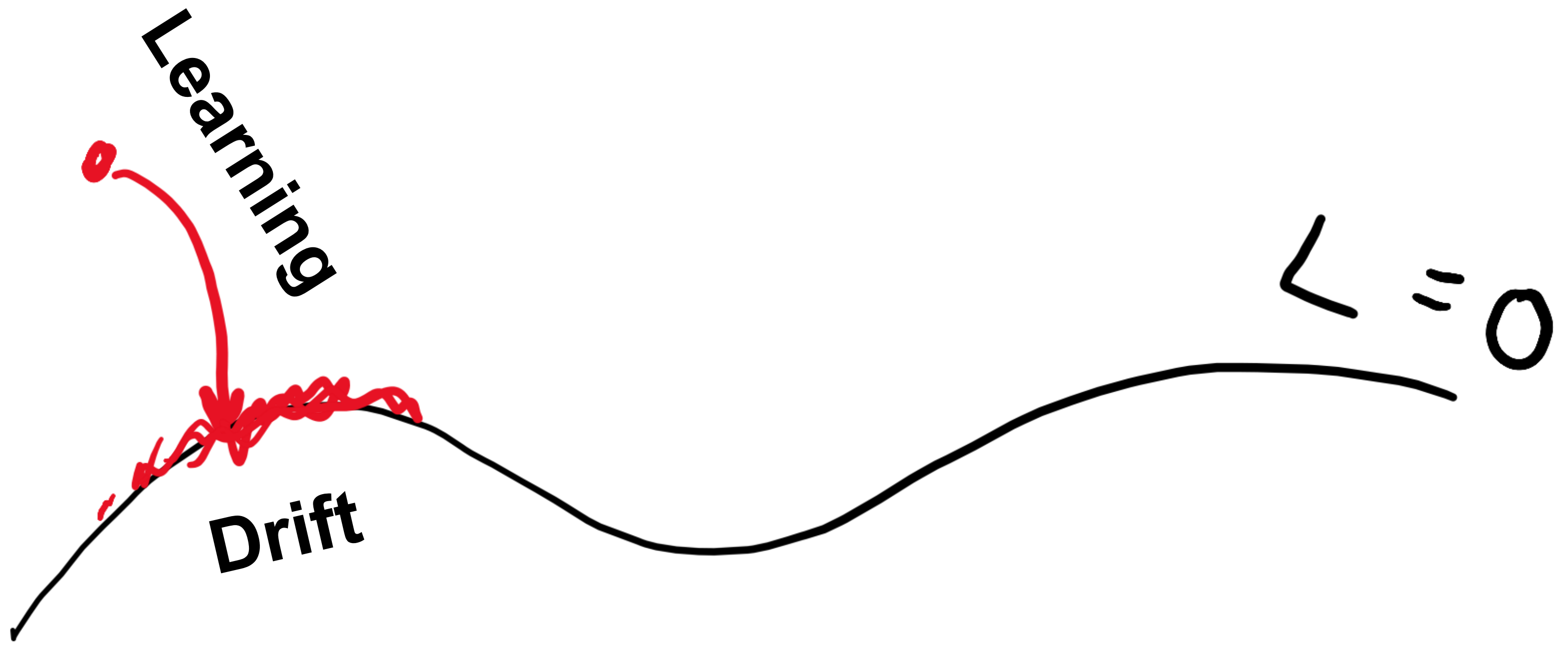


mean SI
3.0
2.5
2.0
1.5
0.4
0.2
fraction active unit
0.0
0.2
0.4
0.6
0.8
1.0

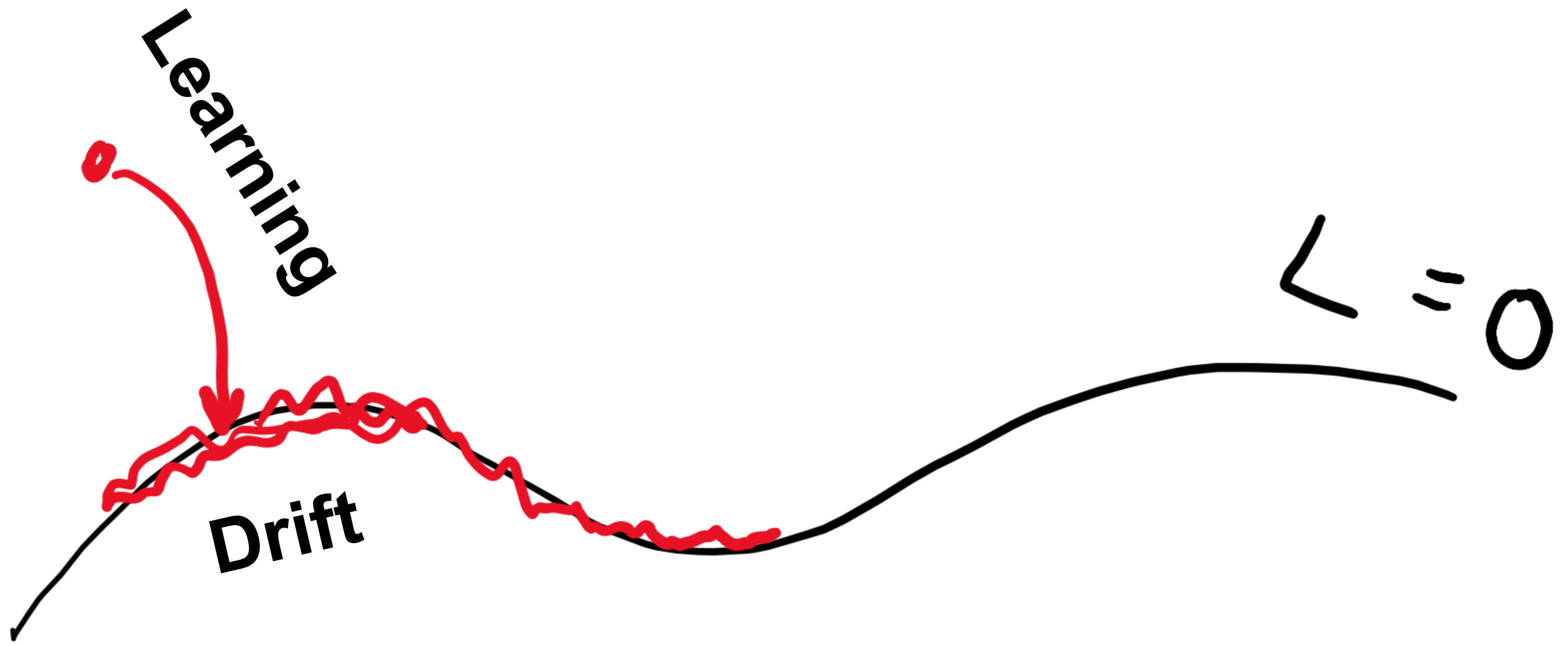


A
Khatib et al. 2023
active cells
mean SI
Normalized value
2.5
2.0
1.5
1.0
0.5
0
50
100
150
200
time [m]
B
Jercog et al. 2019
0
2
4
6
8
time [days]
C
Karlsson et al. 2015
Normalized value
2.5
2.0
1.5
1.0
0.5
0
2
4
6
8
10
time [days]
D
Sheintuch et al. 2023
0
2
4
6
time [days]

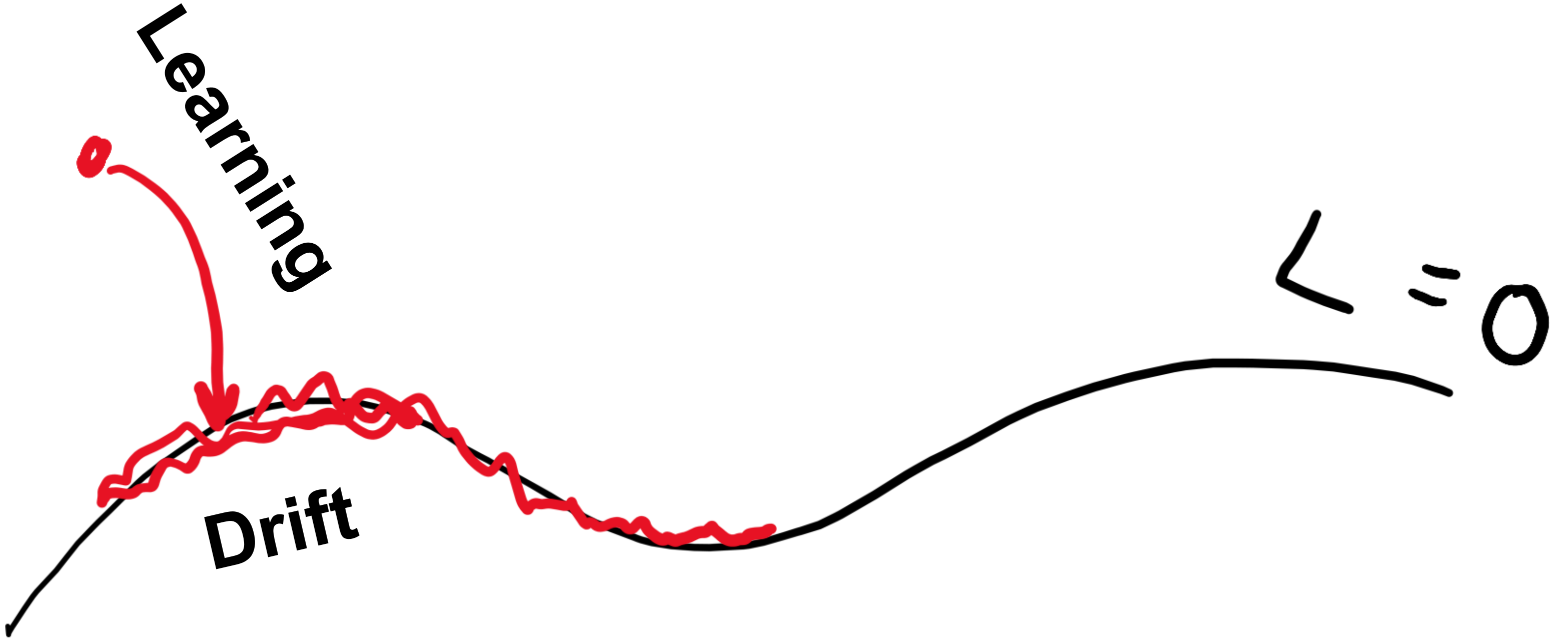
Directed drift



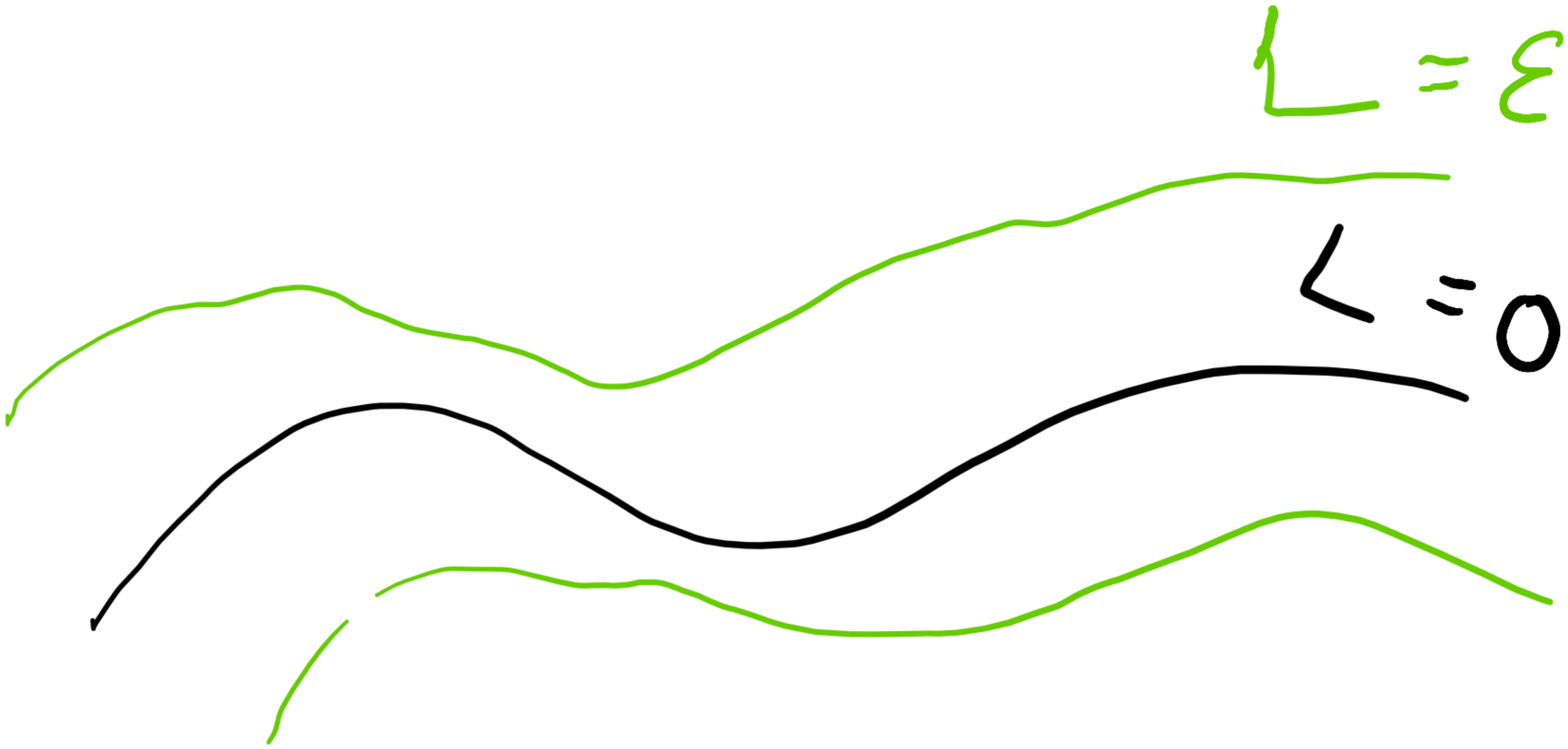
Directed drift



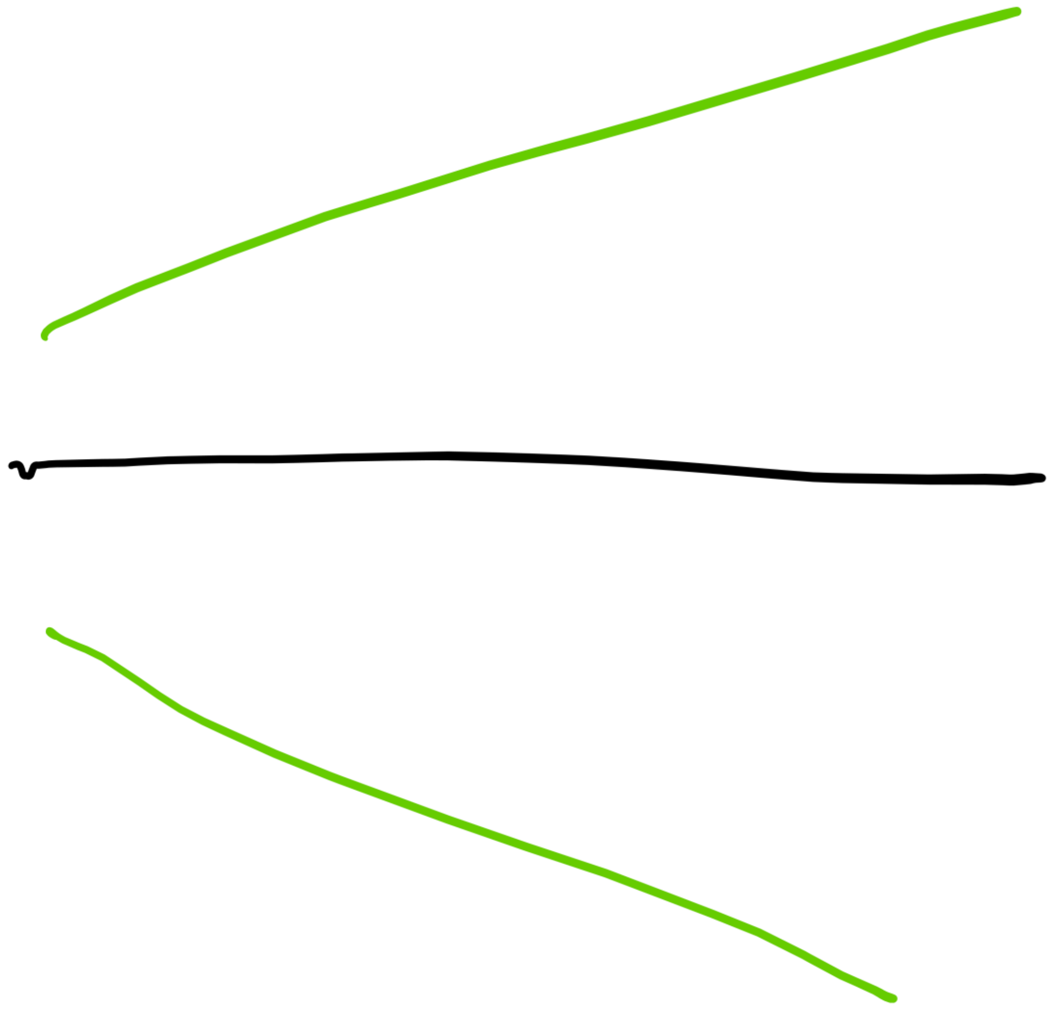
Why?



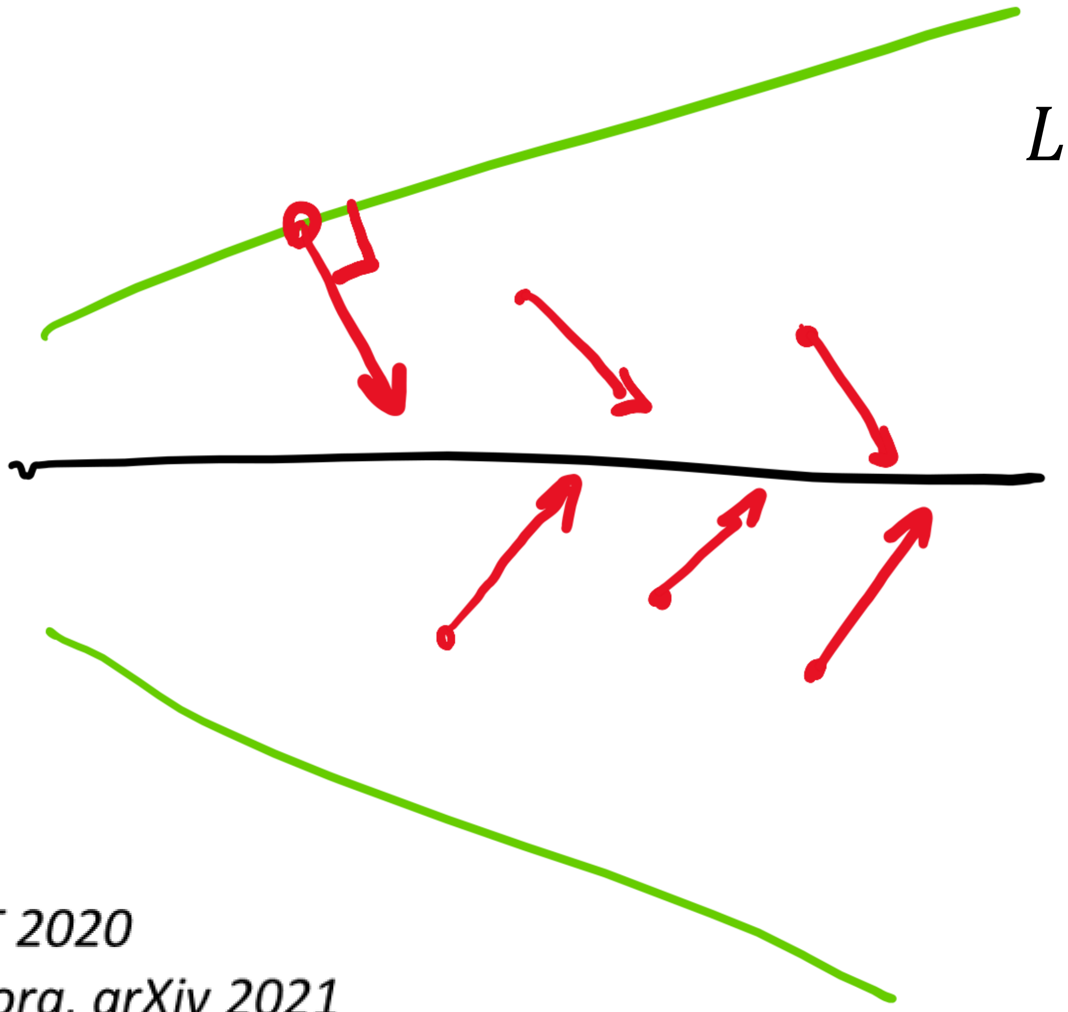
Contour lines



Zoom in



Gradient w/ noise



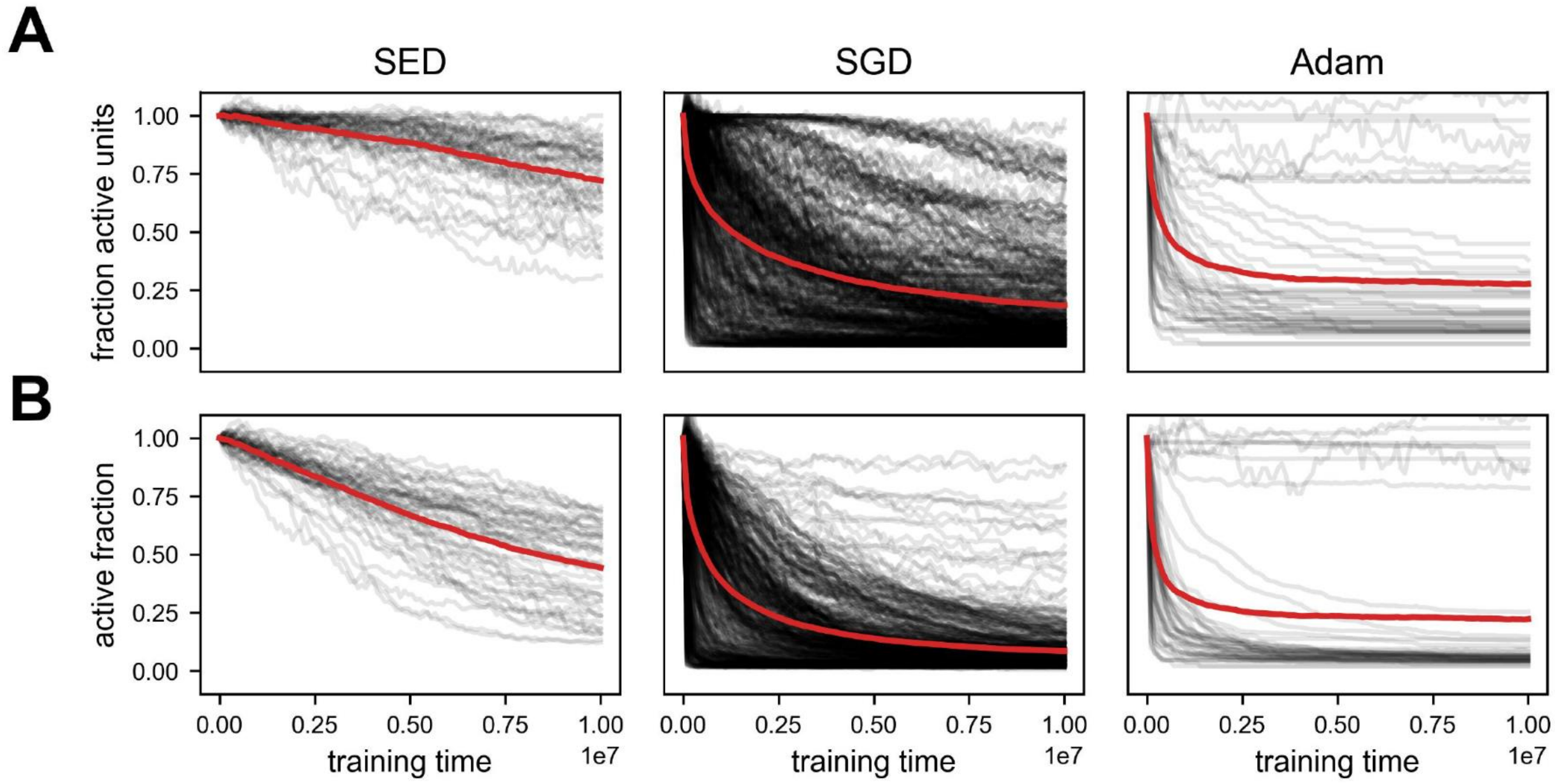
$$L(\theta) = 0 + 0 + L''(\theta)$$

$$L'(\theta) = \left(L''(\theta)\right)'$$

$$L(0) = 0$$

Blanc et al., COLT 2020
Li, Wang, and Arora, arXiv 2021

Infer learning rules



Does it work?

- **How to compare model and data?**
  - Qualitative phenomena
- **How to analyze models?**
  - Implicit regularization. Two phases
- **What do we learn from this?**
  - Link drift and learning
  - Infer learning rules
- **Biological plausibility – should we care? How?**

Quantitative comparisons

- Objective: Understand underlying dynamics
- Assume data is a projection from “true” dynamics
- Fit model to data.
- Hope that model dynamics similar to true dynamics
- Problem: no ground truth
- Solution: prediction.

Kabir Dabholkar

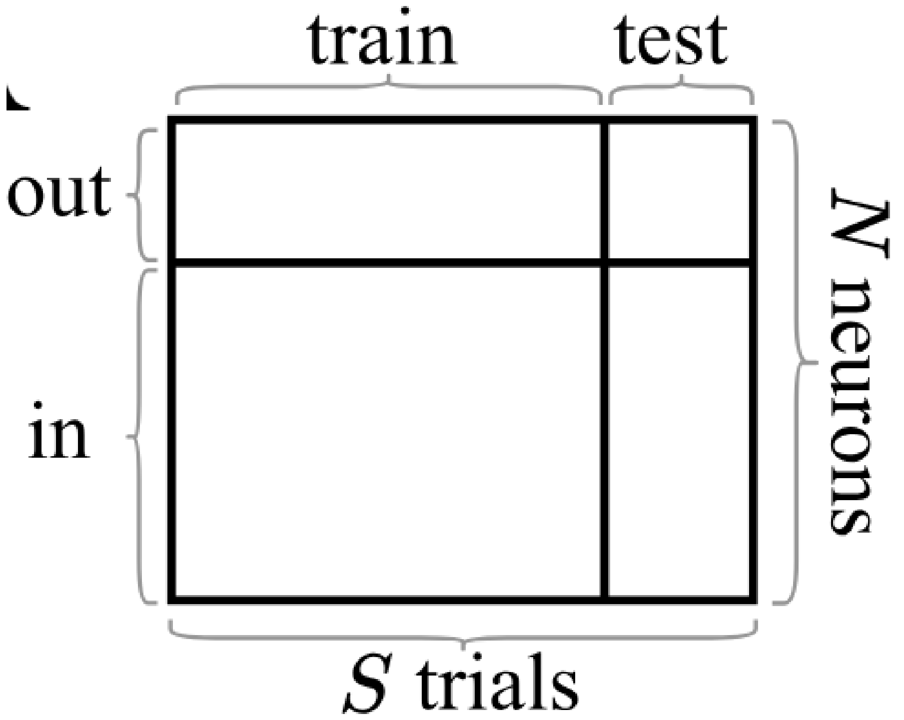
Framework

- Assume data is a projection from “true” dynamics.
- Train model to have a projection which is the data.

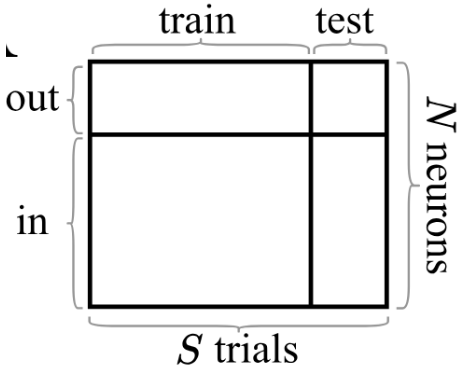
- Neural latents benchmark (Pei et al 2021)
- LFADS (Sussillo et al 2016)
- Neural transformers (Le et al 2022)
- Low rank (Valente et al 2022)
- RNN (Koppe et al 2019)

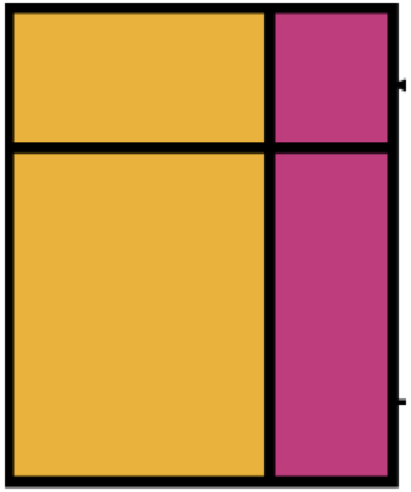
Framework

Data x

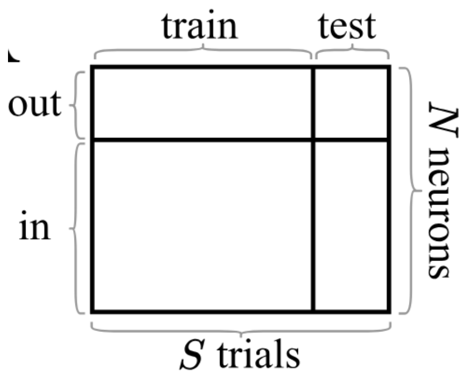


Framework

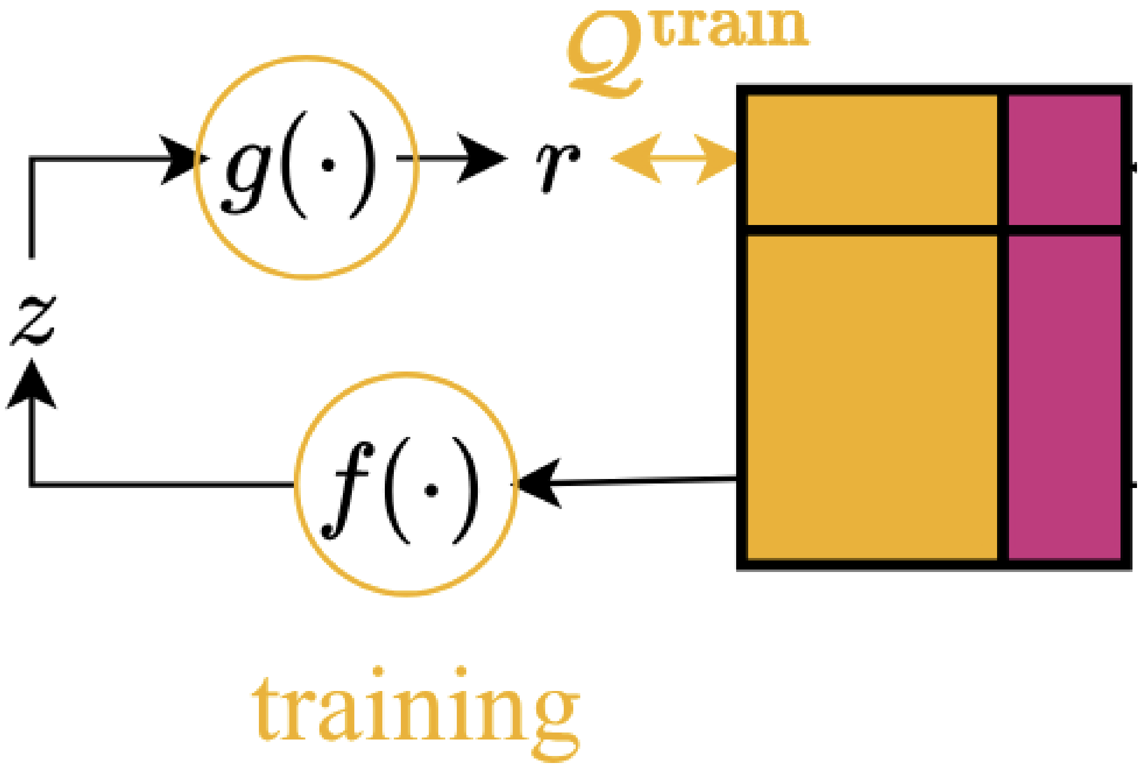




Framework

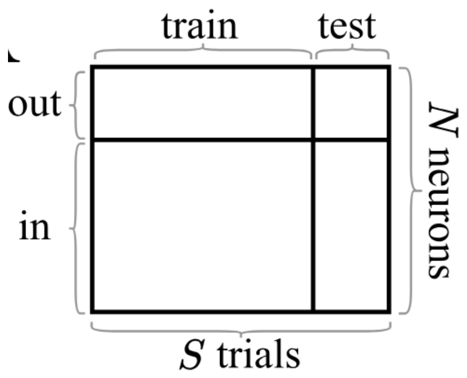


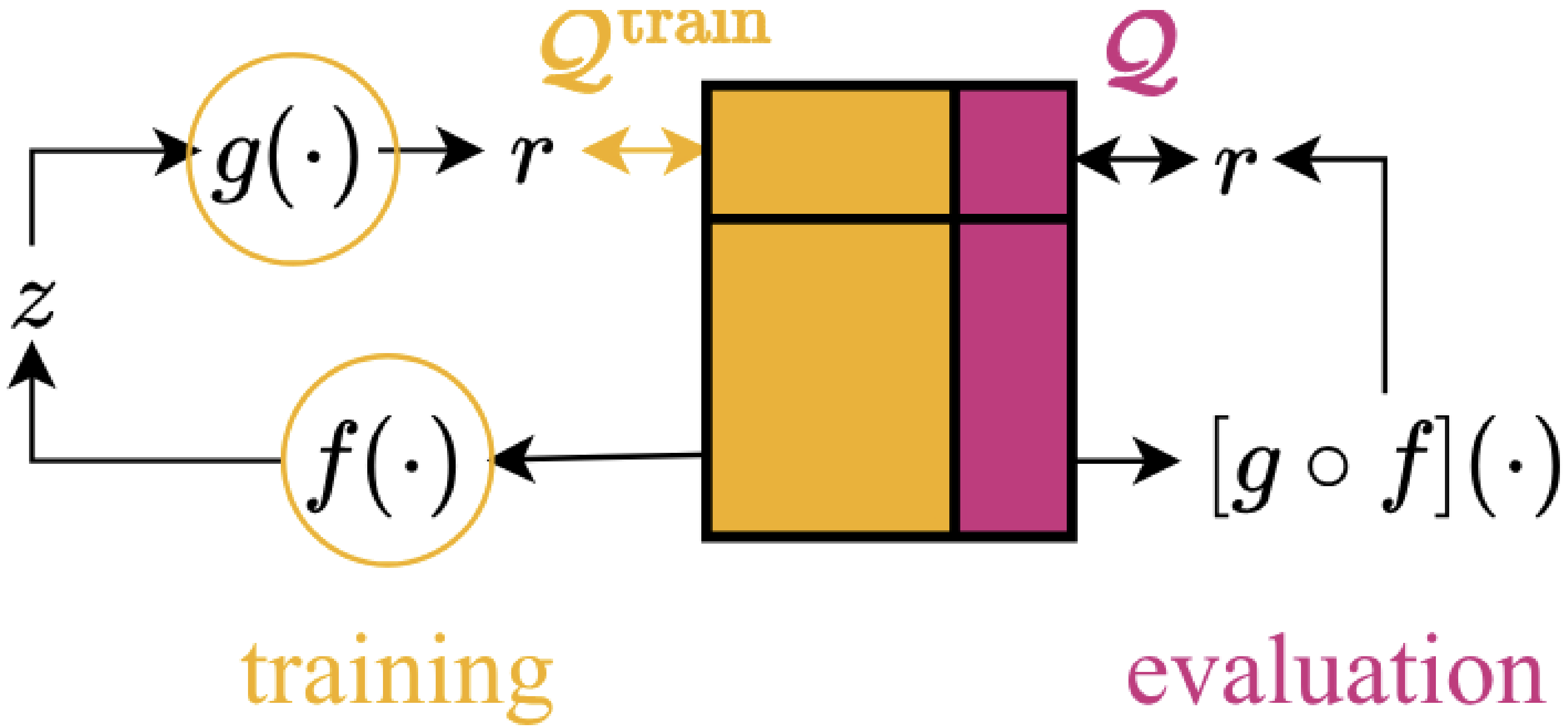
train
test
out
in
N neurons
S trials



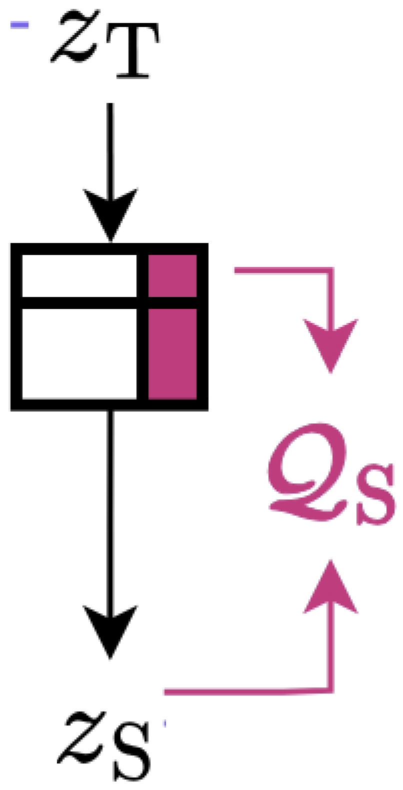
Q^{train}
$g(\cdot)$
r
z
$f(\cdot)$
training

Framework

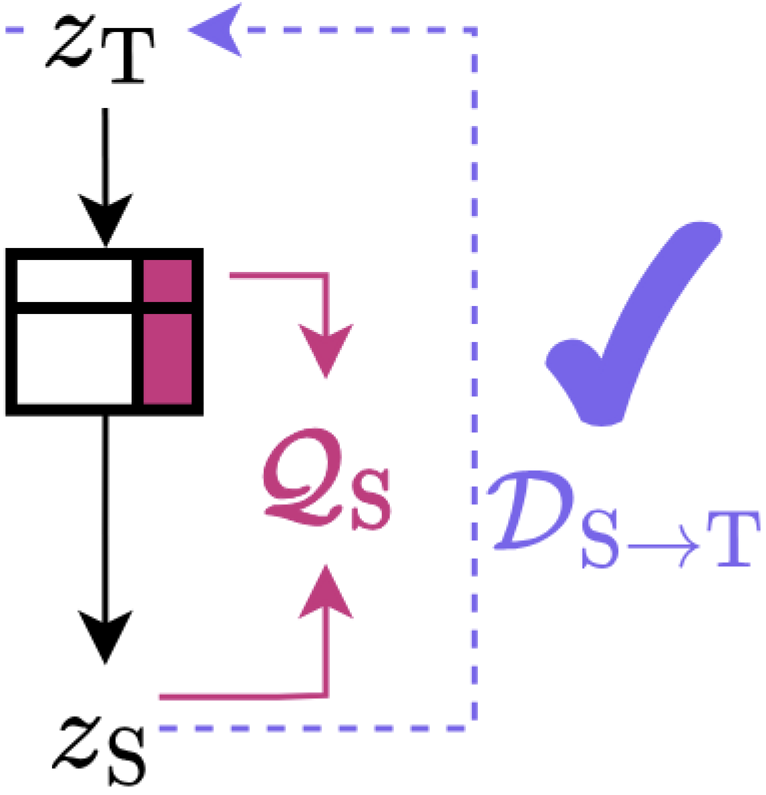

train
test
out
in
N neurons
S trials


Q^{train}
Q
$g(\cdot)$
r
r
z
$f(\cdot)$
$[g \circ f](\cdot)$
training
evaluation

The problem with prediction



The problem with prediction



Works!

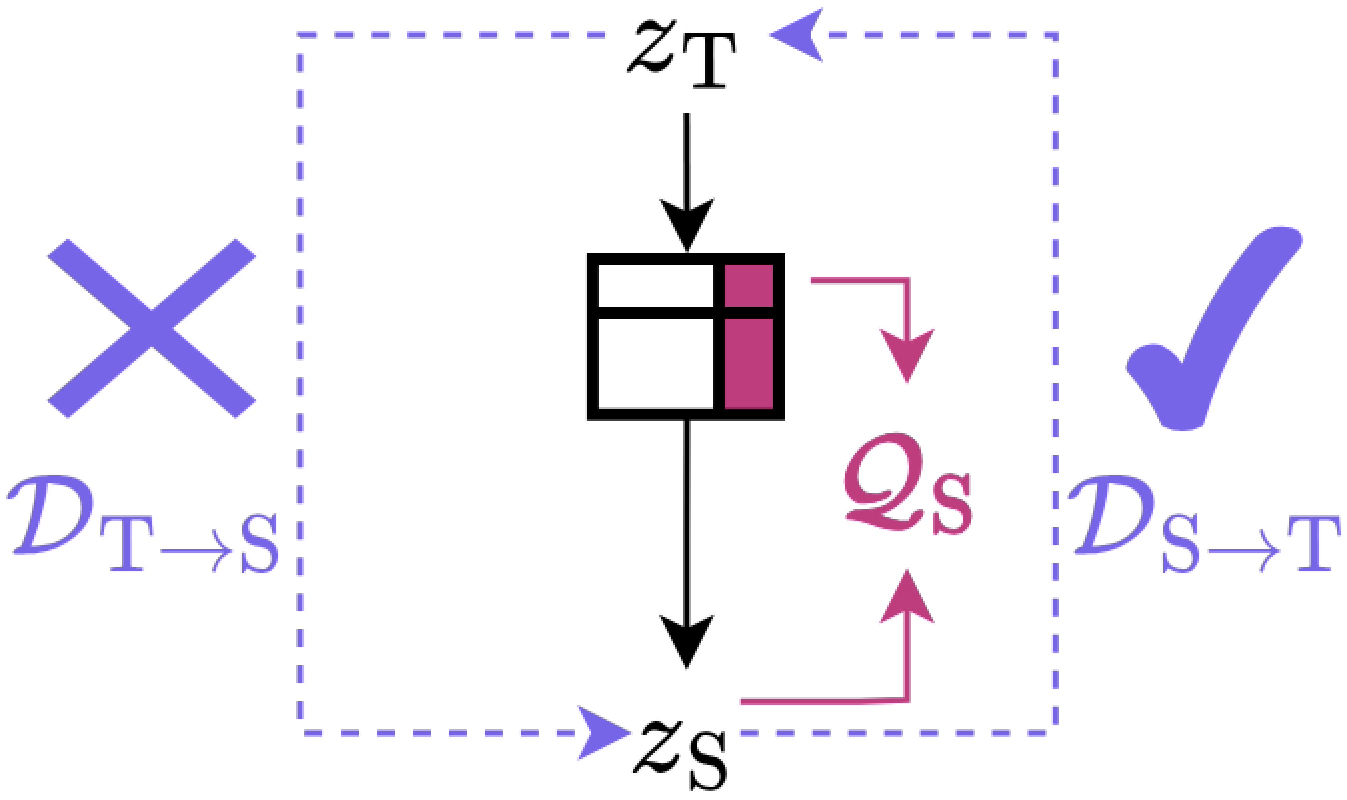
Definition
of good model*

*observable

The problem with prediction

No guarantee

Could be
overly complex



Works!

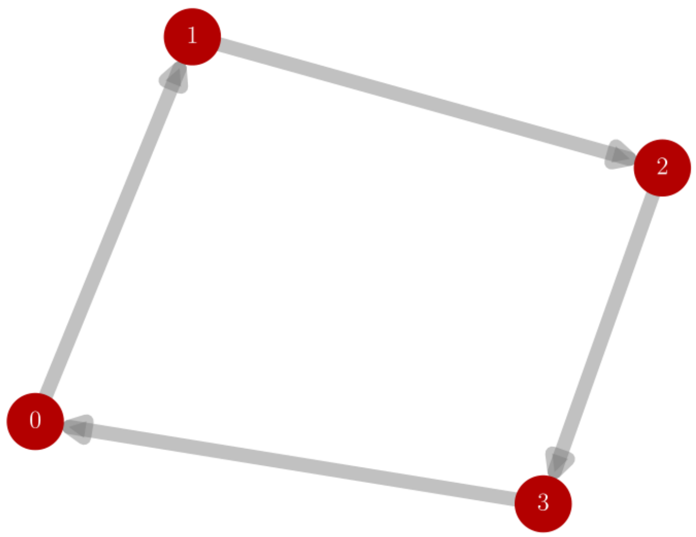
Definition
of good model*

*observable

Simple demonstration

- Student – teacher setting
- Hidden Markov Model (HMM)

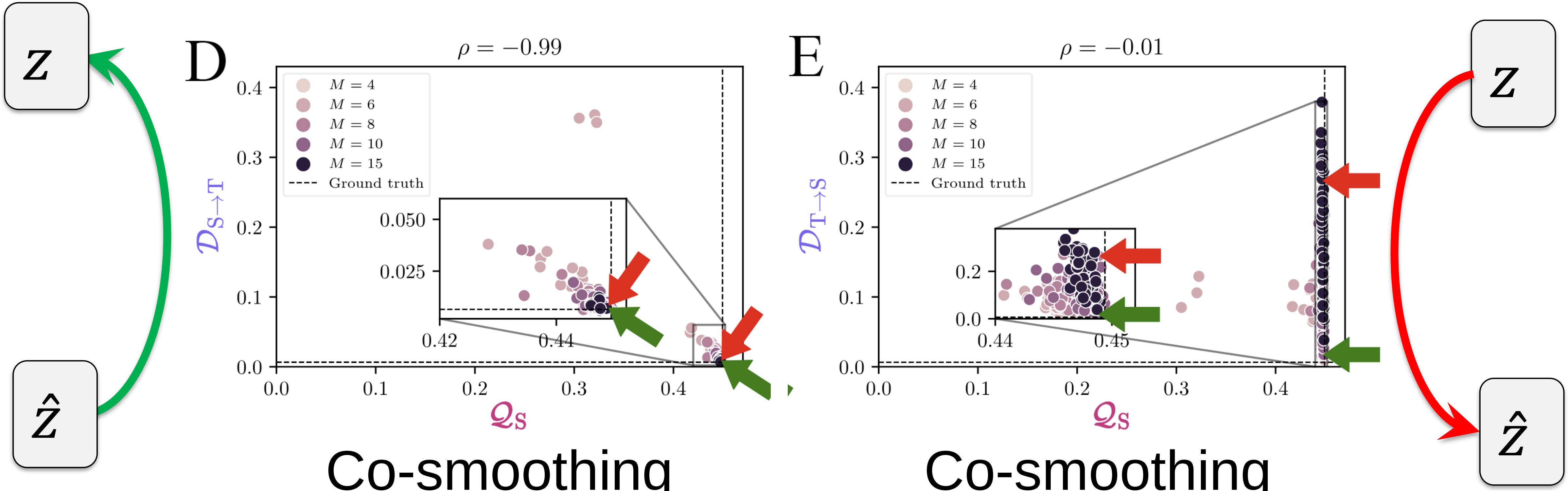
Teacher



Students:

- 400 HMMs
- 4 – 15 states
- Gradient-based optimization

Asymmetric decoding



Co-smoothing

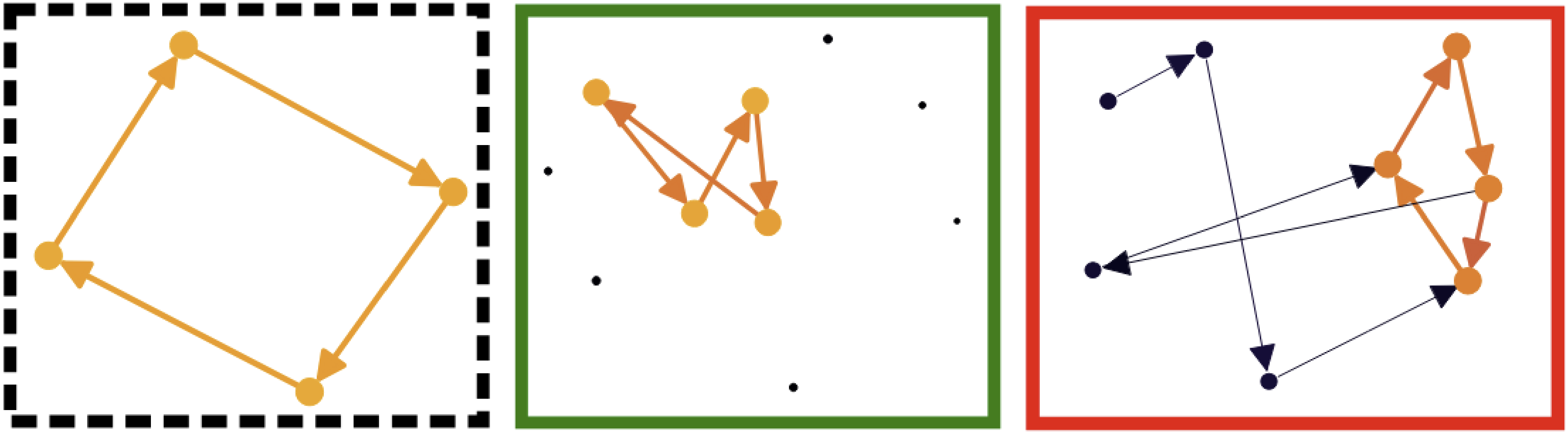
Co-smoothing

Why the discrepancy?

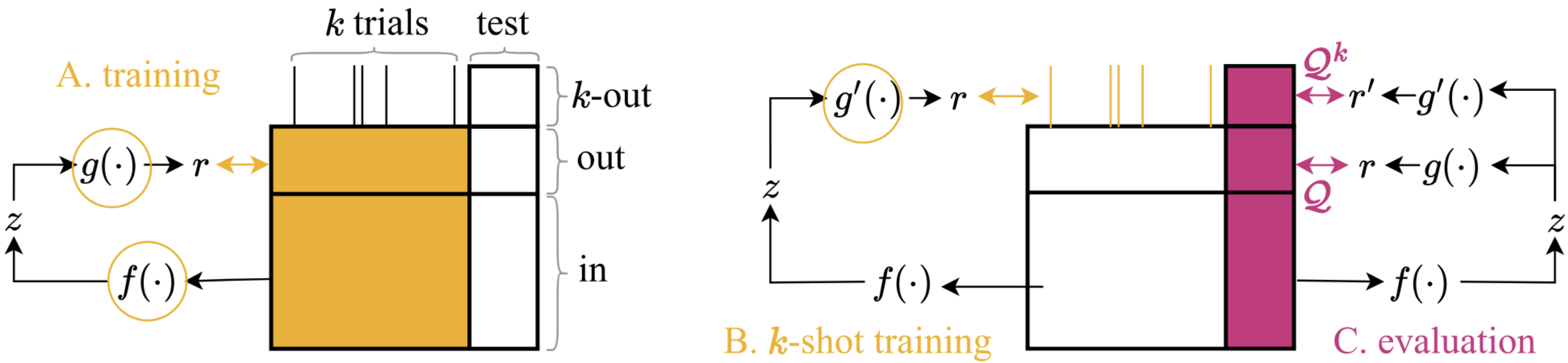
Teacher

Good student

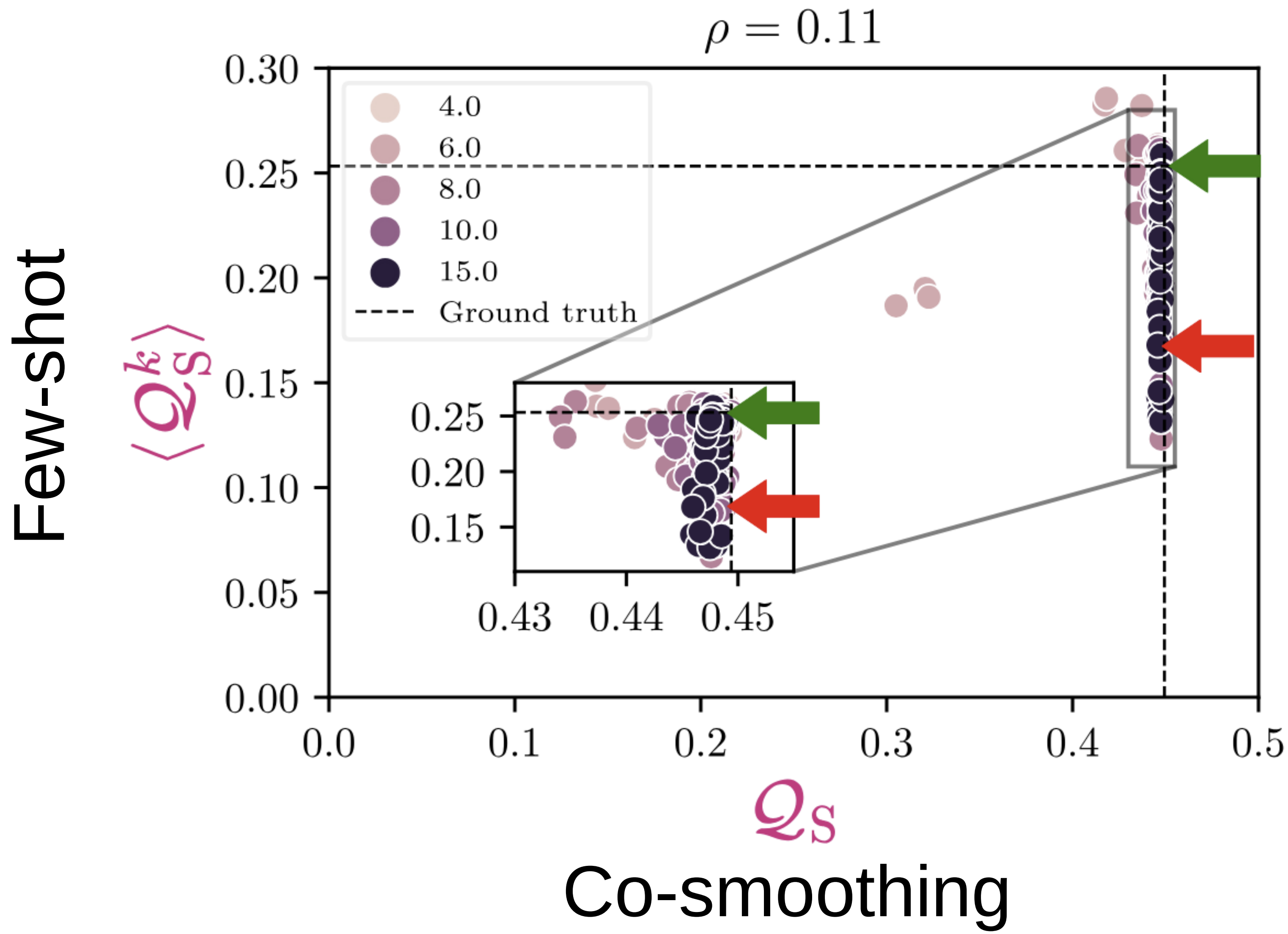
Bad student



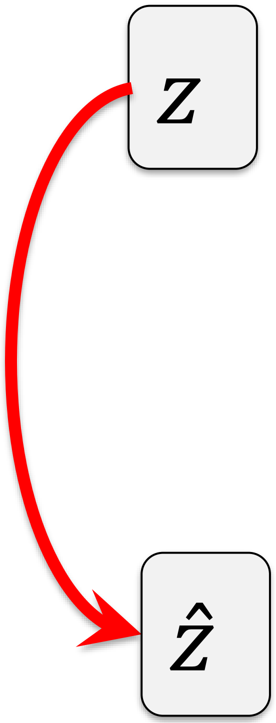
A remedy: few-shot prediction



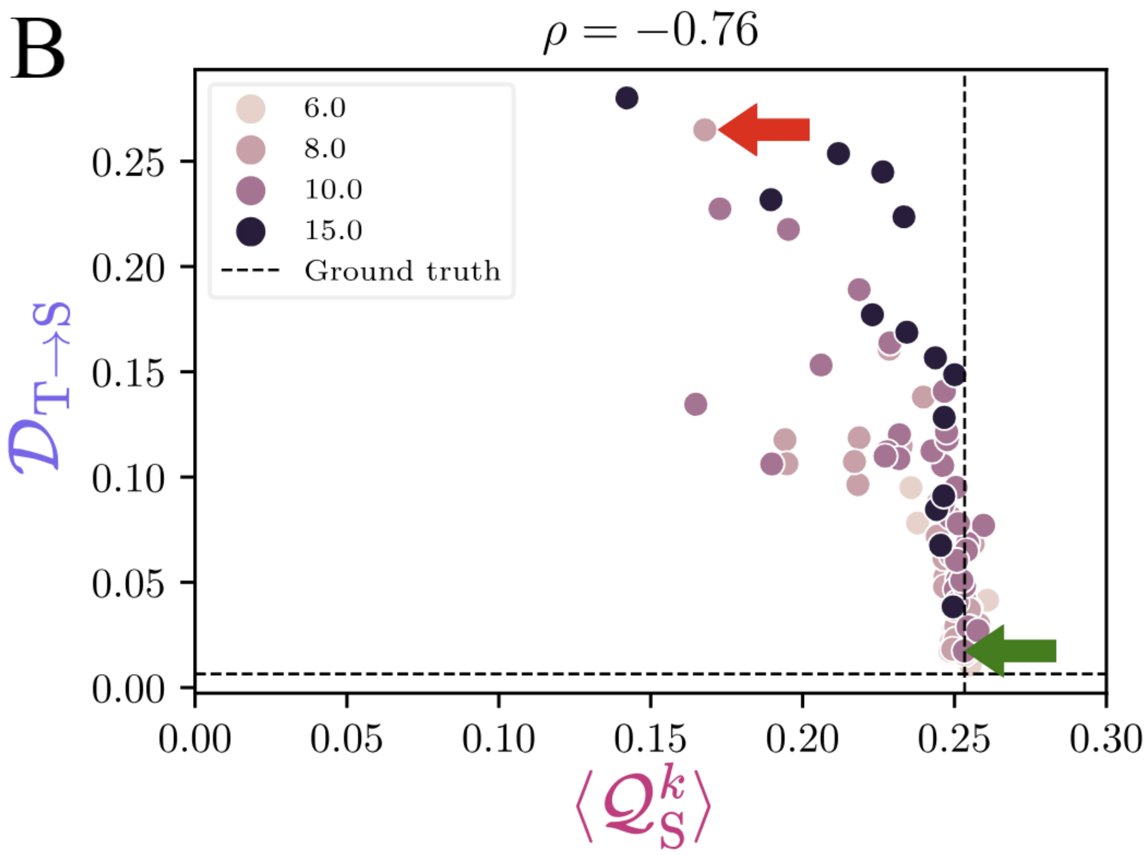
Measures something new



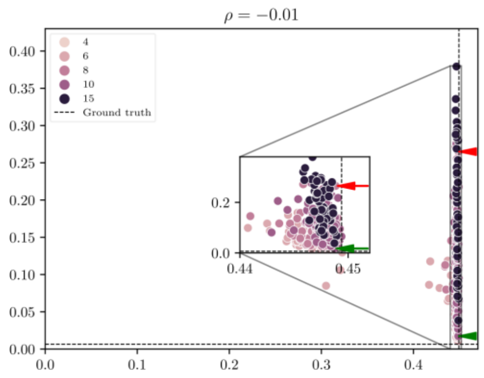
Correlates with ground truth



Decoding error



Few-shot



Co-smoothing

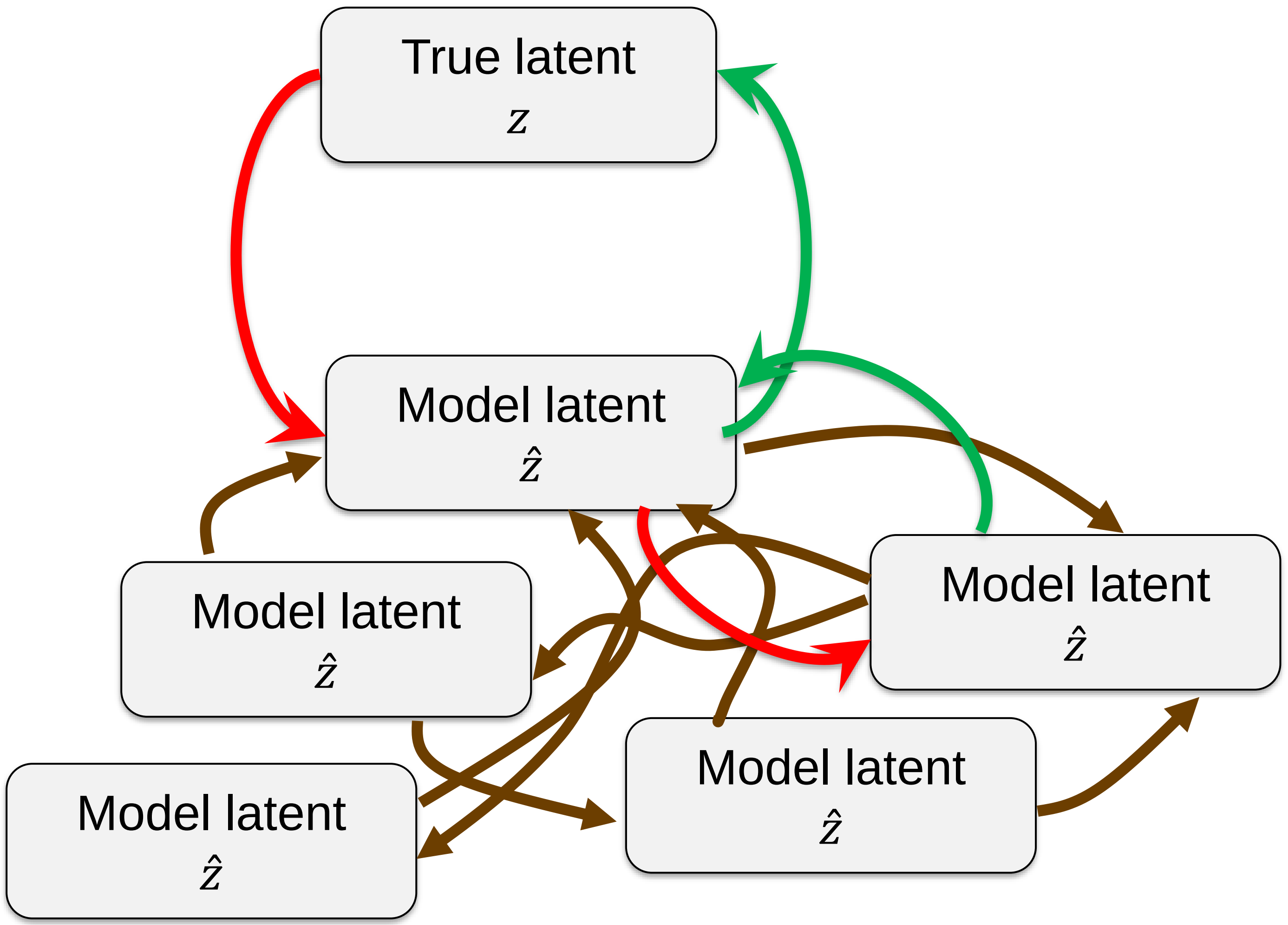
Why does it work?

- Data efficiency
- Information from observations is “spread” on fewer states.
- Can prove this for HMMs

Beyond HMMs

- State of the art models:
  - LFADS
  - Transformers (STNDT)
- MC_maze
  - Primary motor cortex, dorsal premotor cortex
  - Monkeys reaching in “maze” setting
  - Churchland et al 2010

No ground truth



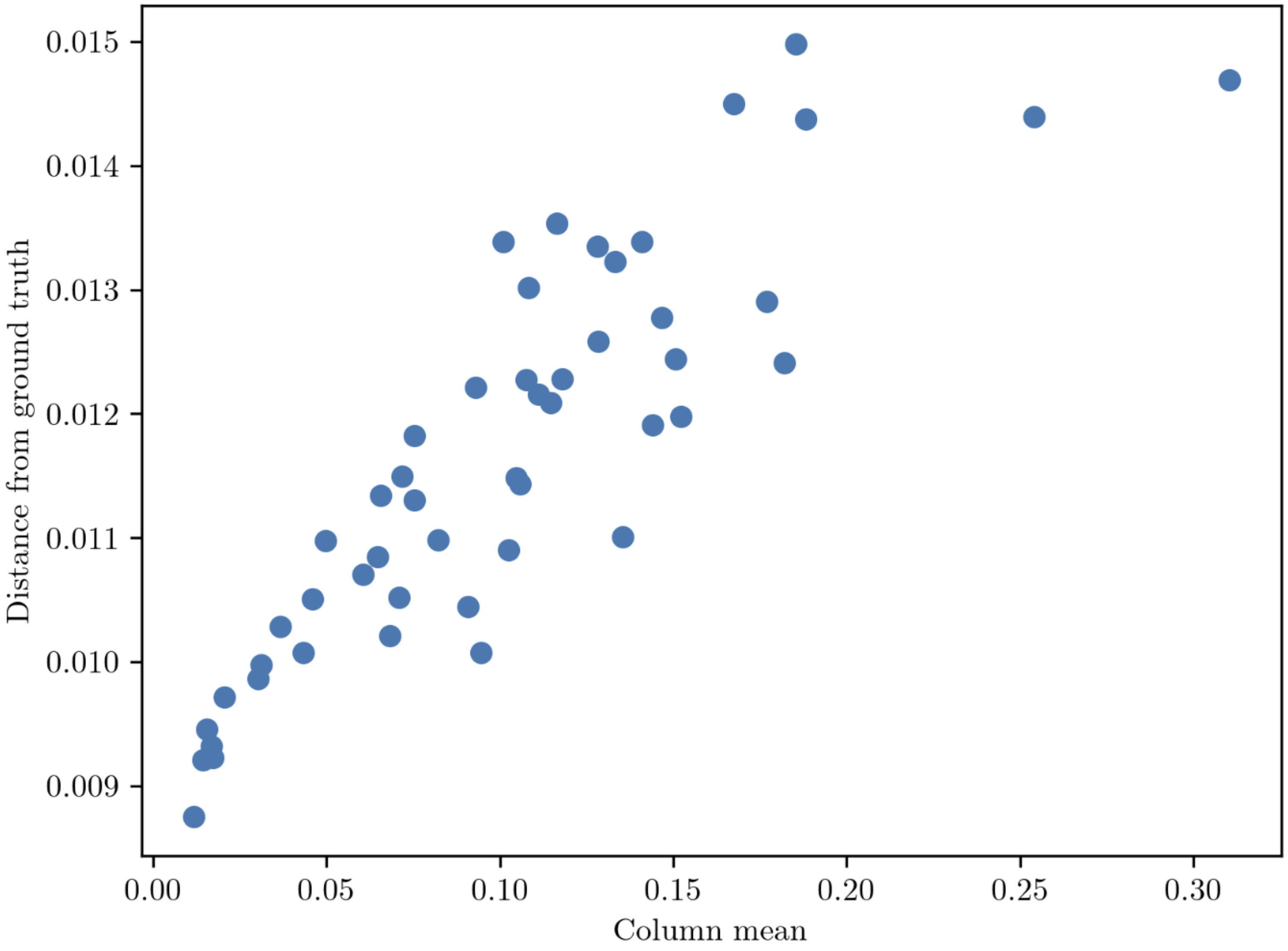
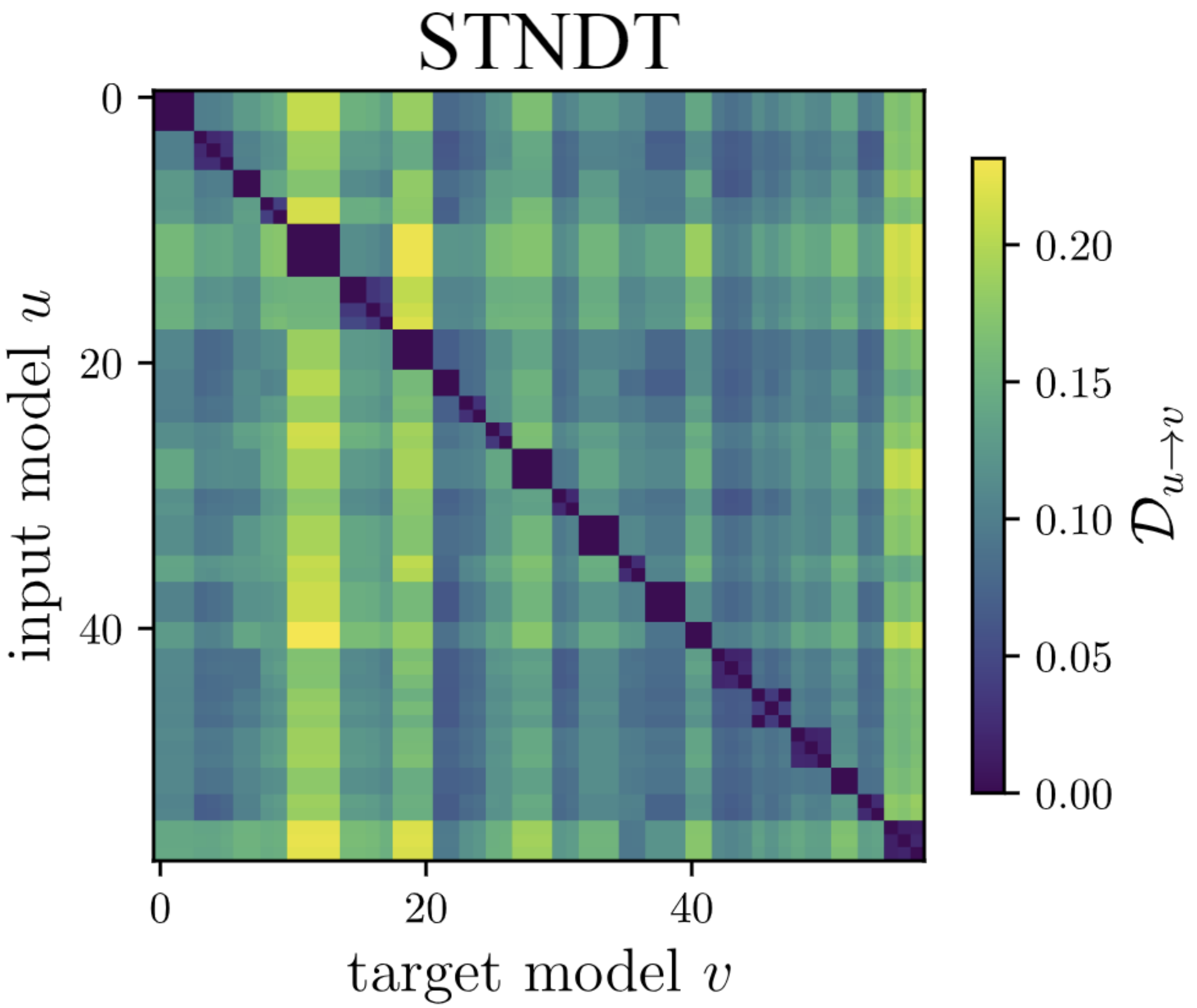


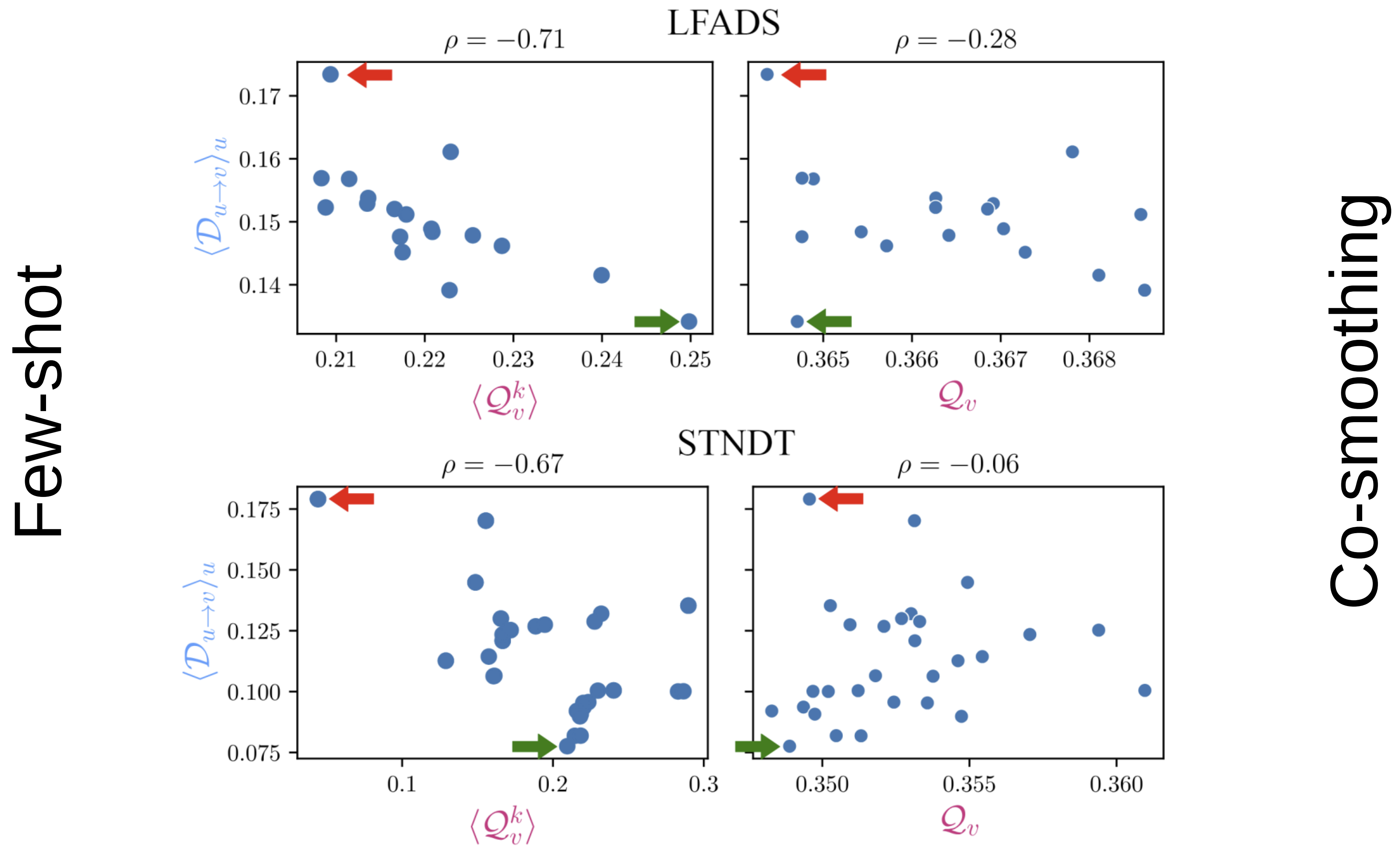
Figure 9: For HMM students with high co-smoothing $\mathcal{Q}_\mathrm{S} > \mathcal{Q}_\mathrm{T} - 10^{-3}$ (and therefore low $\mathcal{D}_{\mathrm{S}\to\mathrm{T}}$ 1D), the cross-decoding metric $\langle \mathcal{D}_{u\to v}\rangle_{u\in\text{students}}$ is correlated to ground truth distance $\mathcal{D}_{\mathrm{T}\to\mathrm{S}}$.

Cross decoding



$$< \mathcal{D}_{u \to v} >_u$$

Few-shot works



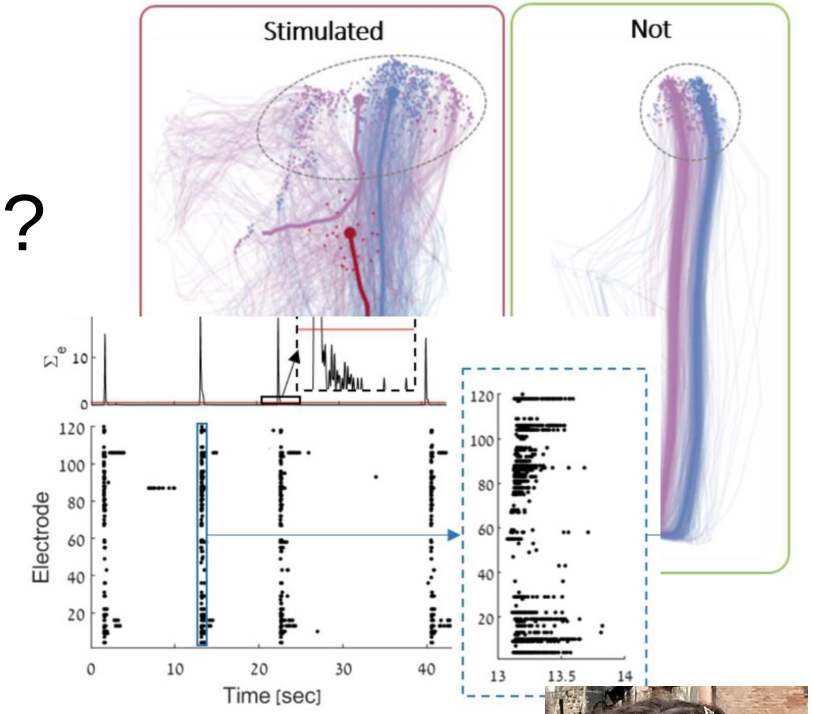
Cross-decoding distance

Does it work?

- **How to compare model and data?**
  - Quantitative fit
- **How to analyze models?**
  - HMMs as a tractable tool
- **What do we learn from this?**
  - Improve latents without looking at them.
  - Understand the relation predict-explain.
- **Biological plausibility – should we care? How?**

Biological plausibility

- In the eye of the beholder
- SGD is not realistic, but GD with noise?
- At what level should we compare?
- Fitting models to data
  - Model mismatch (more/less realistic)
- There’s still work to be done…



Beer and Barak 2024

Chen Beer

Conclusions

- Complex neural signatures suggest complex models
- Surprisingly, a good model for adapting complex systems (e.g. the brain) could be…
  - Adapting complex systems!
- Understanding these models requires theory.

Benefits of this approach

- Cross-field fertilization
  - Neuro: representational drift, task-irrelevant variance.
  - ML: Implicit regularization, rich/lazy regimes.
  - Evolution: Survival of the flattest
  - Genetics: scale free topology (Rivkind et al 2020)
  - Cancer: (Shomar et al 2022)
- Hypothesis generation
- And also many dangers and pitfalls

Lab members:

Chen Beer
Elia Turner
Kabir Dabholkar
Aviv Ratzon

Matthijs Pals
Nathan Cloos

Alumni

Oded Barzelay
Alexander Rivkind
Xu Tie
Lee Susman
Friedrich Schuessler
Aseel Shomar

International Human Frontier Science Program Organization

הקרן הלאומית למדע

Collaborators:

Naama Brenner
Srdjan Ostojic
Dori Derdikman
Jackie Schiller
Shimon Marom
Jakob Macke