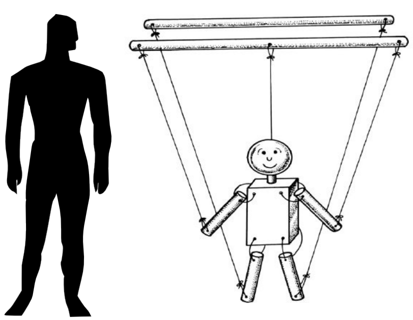
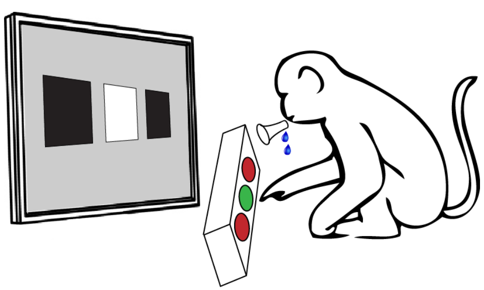
Learning to classify
From behavior to neural dynamics

Yarden Cohen
Advisors: Elad Schneidman. Rony Paz

Neurobiology department,
Weizmann Institute of Science, Israel



Behavior Modeling



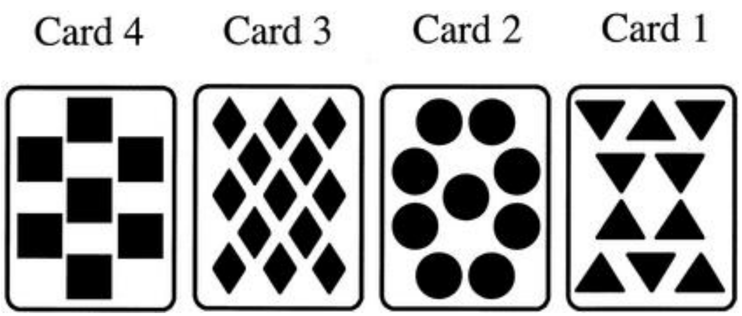
Electrophysiology

May 2015

Learning to classify

Will it rain today?

Experimental and modeling approaches to rule based learning



80% 

20% 

- Neurological disorders' effect on learning 'weather prediction'
- After training neurons reflect correct probabilities
- Complexity correlates with mean success on different rules
- Prior that people have on the task

Gluck et al. Learning and Memory, 2002
Yang&Shadlen, *Nature*, 2007
Feldman, *Nature* 2000
Goodman et al. *Cognitive Science*, 2008
Griffiths&Tenenbaum Behavioral and brain sciences 2001

How do individuals learn conceptually different (deterministic) rules?

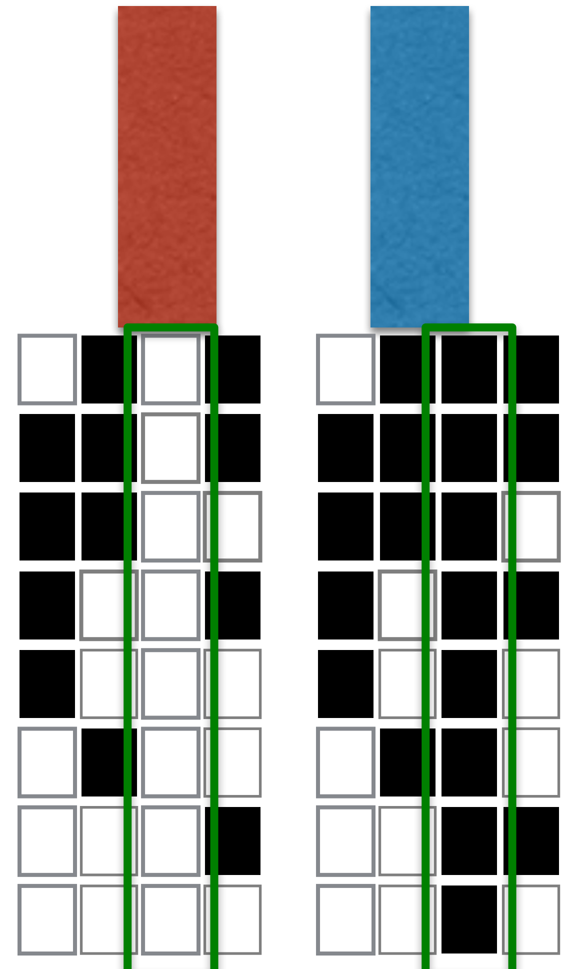
<u>A single framework that describes</u>:

- Learning dynamics
- Individual subjects
- Conceptually different rules

Deterministic binary classification task

pattern $\vec{x} = (x_1, x_2, x_3, x_4)$

label y

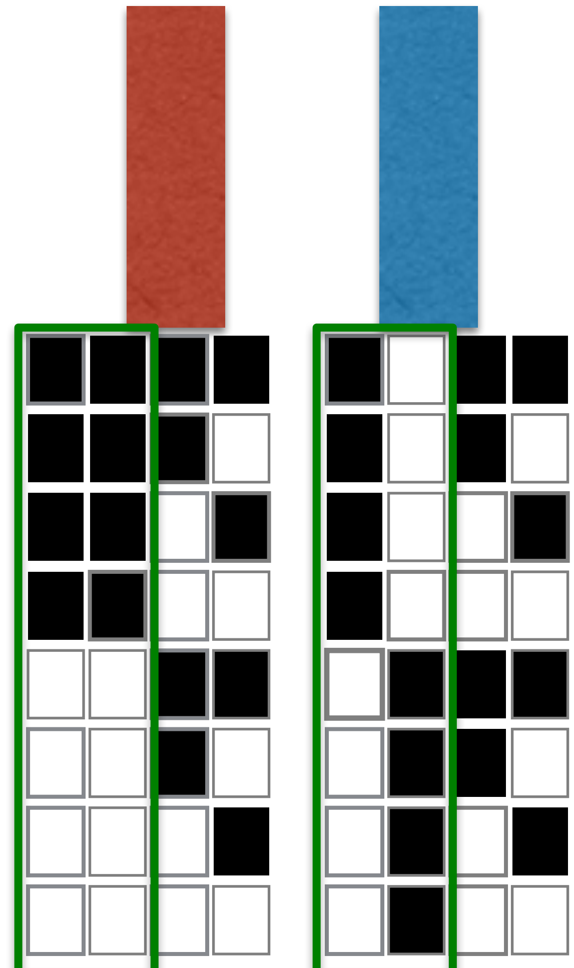


(Cohen & Schneidman, *PNAS*,2013)

Deterministic binary classification task

pattern $\vec{x} = (x_1, x_2, x_3, x_4)$

label y



(Cohen & Schneidman, *PNAS*,2013)

<u>For n-squares</u>

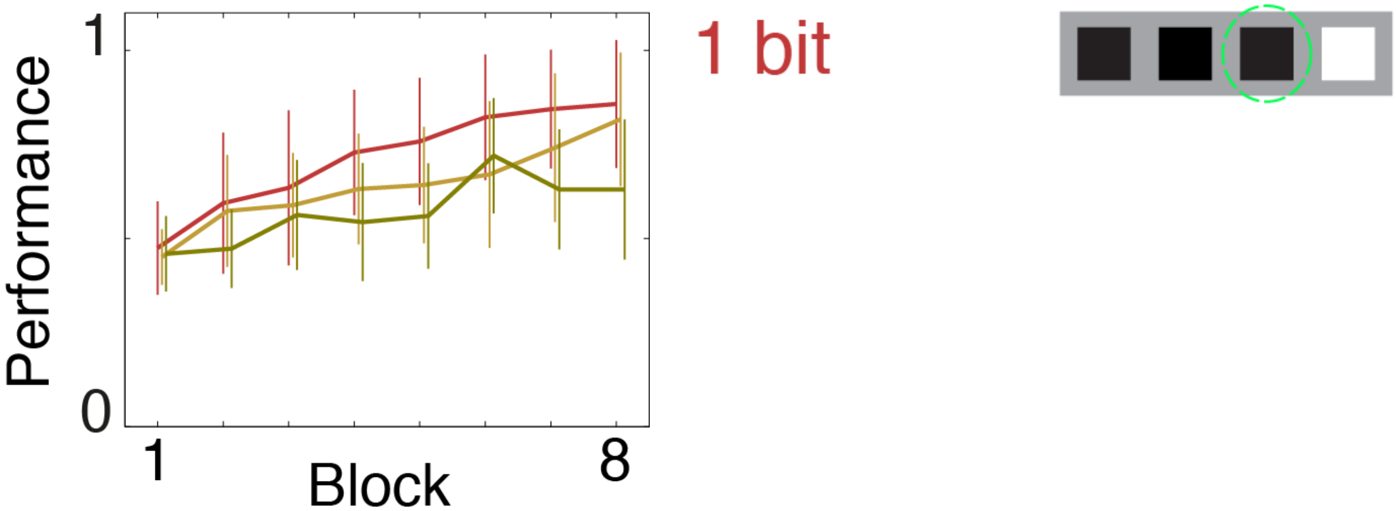
2^n patterns

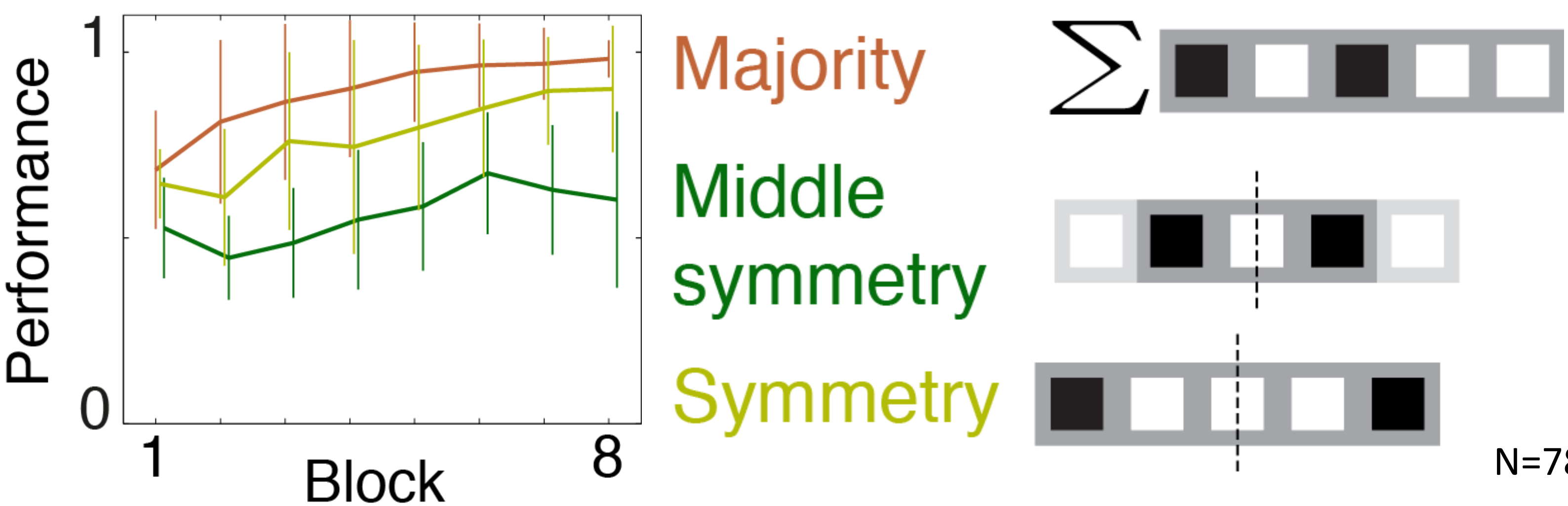
2^{2^n} potential (deterministic) rules

N=4 → >65,000 rules

N=5 → >9,000,000,000 rules

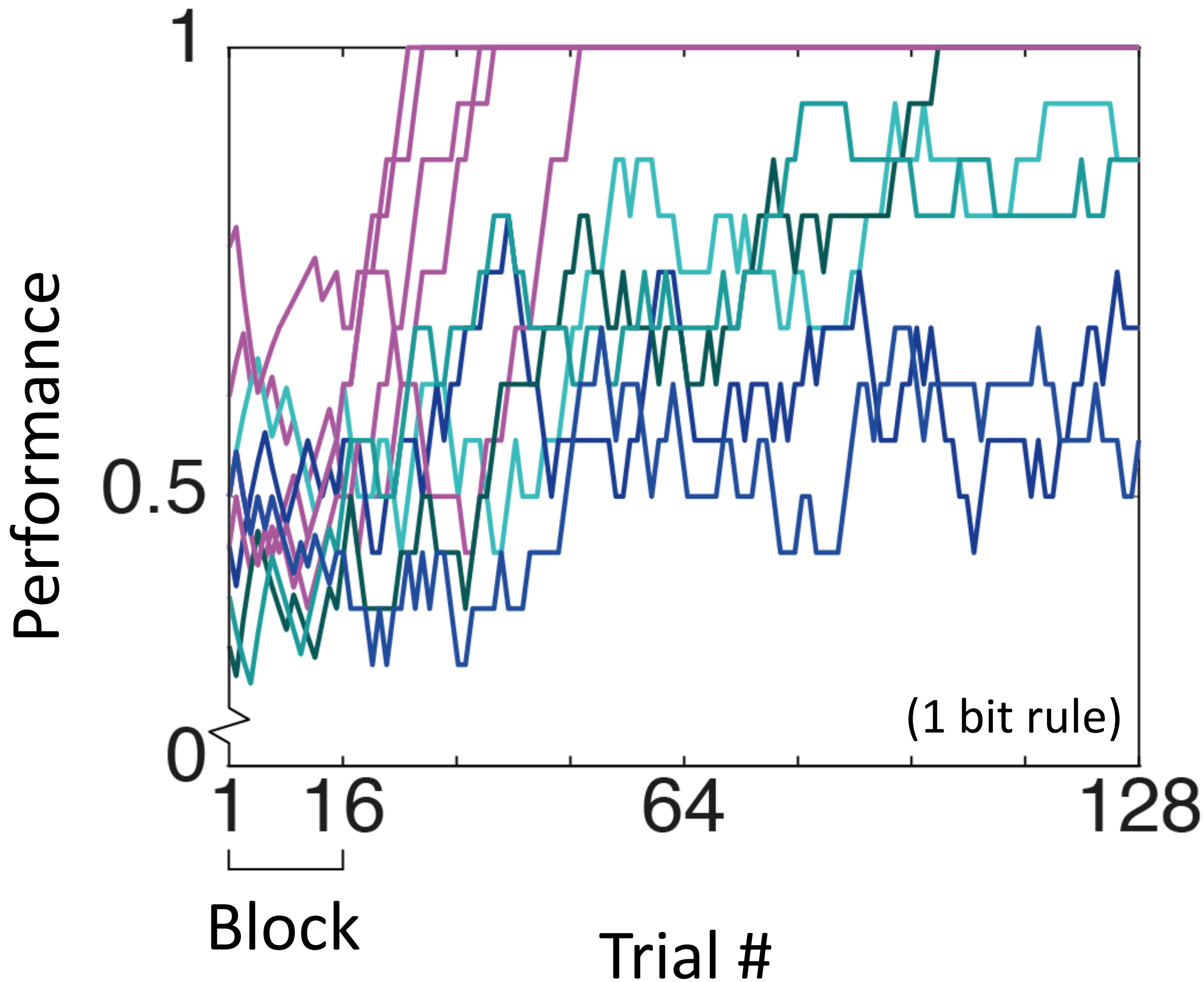
Average reflects rule complexity but poorly accounts for individual behavior



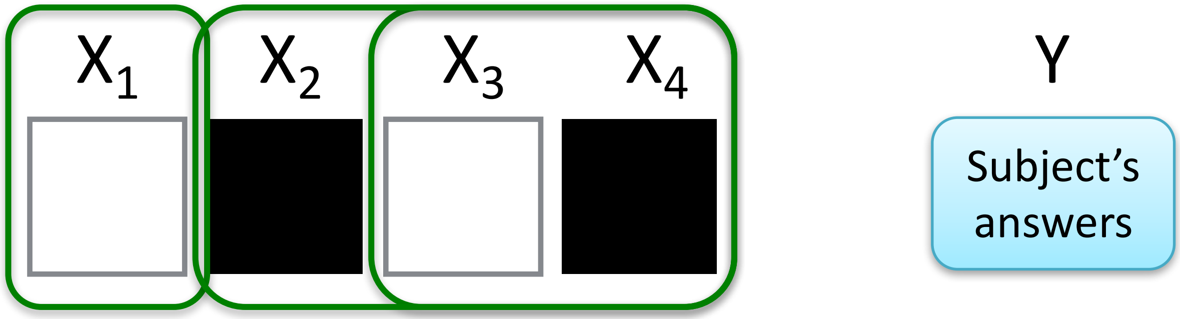


N=78 subjects, each learned 4 rules

Learning curves are very diverse



Directly measuring strategies rarely succeeds



Pattern features that span all rules

Black=-1
White=1

1 bit: $f(X_1,X_2,X_3,X_4)=X_1$

2 bit: $f(X_1,X_2,X_3,X_4)=X_3X_4$

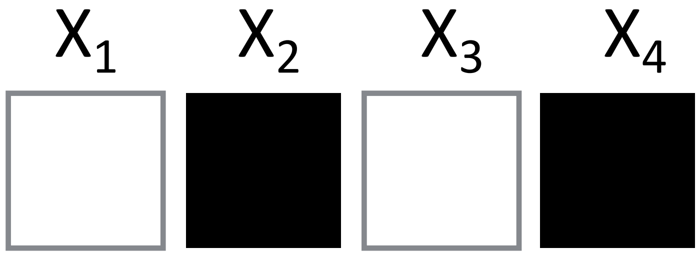
3 bit: $f(X_1,X_2,X_3,X_4)=X_2X_3X_4$

4 bit: $f(X_1,X_2,X_3,X_4)=X_1X_2X_3X_4$

Mutual information measures feature-answer relation

Directly measuring strategies rarely succeeds

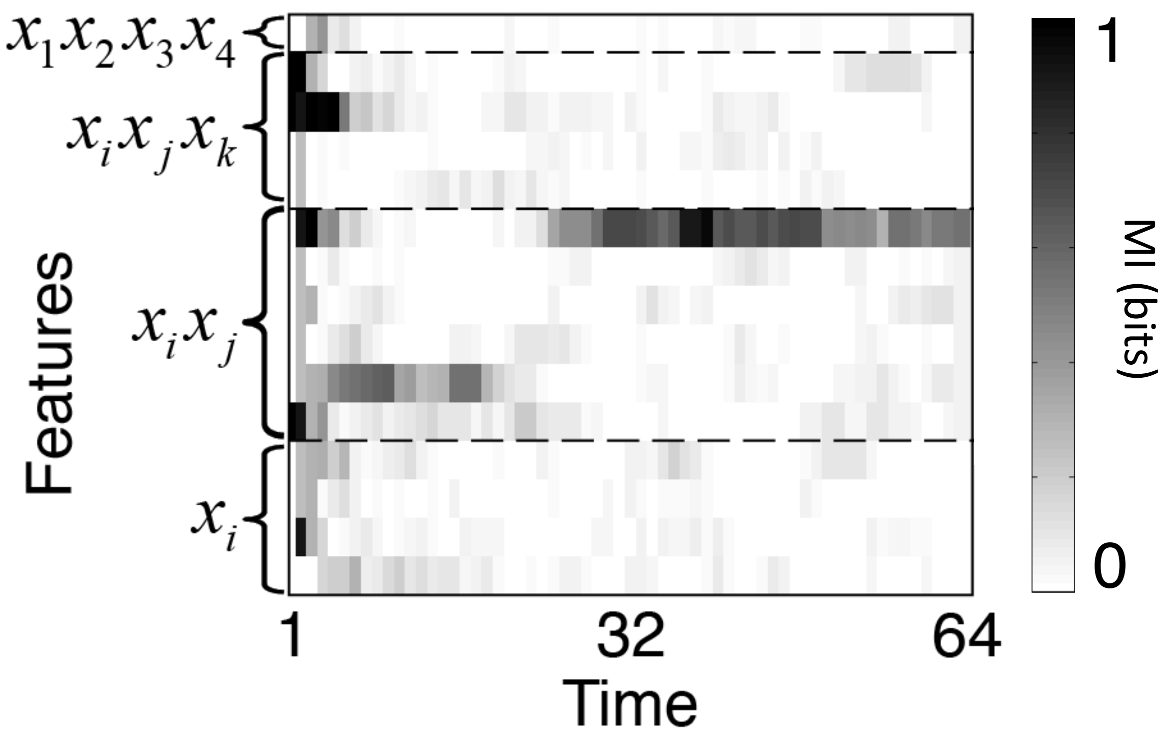
Black=-1
White=1



Y

Subject's answers

Compute $I(f(X);Y)$



Internal category models introduce features weighting

$$p(\vec{x}|y) = \qquad f_\mu(\vec{x})\}$$

features

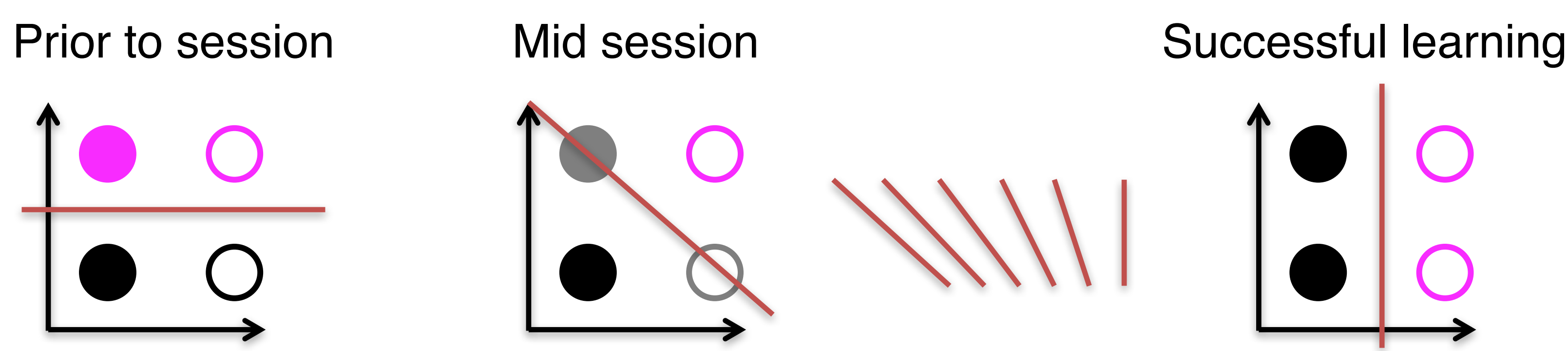
(Cohen & Schneidman, *PNAS*,2013)

Learning is a change in the feature weights

$$p(\vec{x}|y) = \frac{1}{Z}\exp\{\beta \sum_{\mu} \alpha_{\mu}(t) f_{\mu}(\vec{x})\}$$

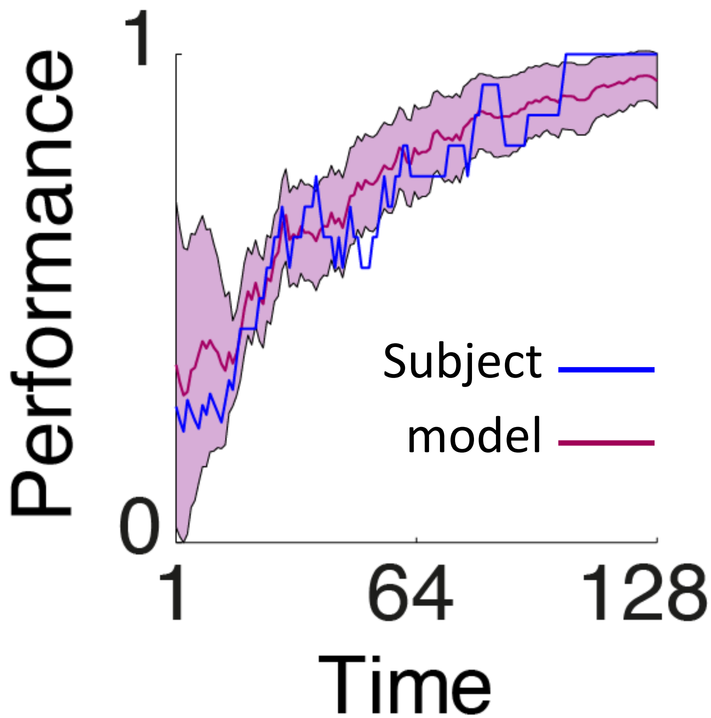
Learning rule

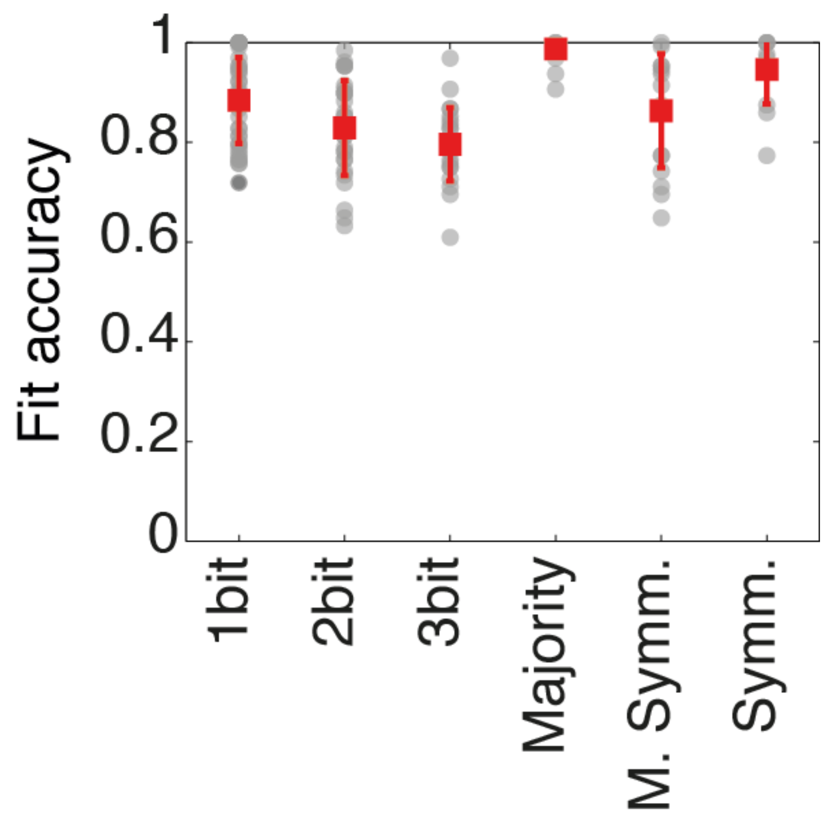
$$\Delta\alpha_{\mu} = \eta \cdot \frac{\partial p(y \mid \vec{x})}{\partial \alpha_{\mu}}$$



Models fit behavior well

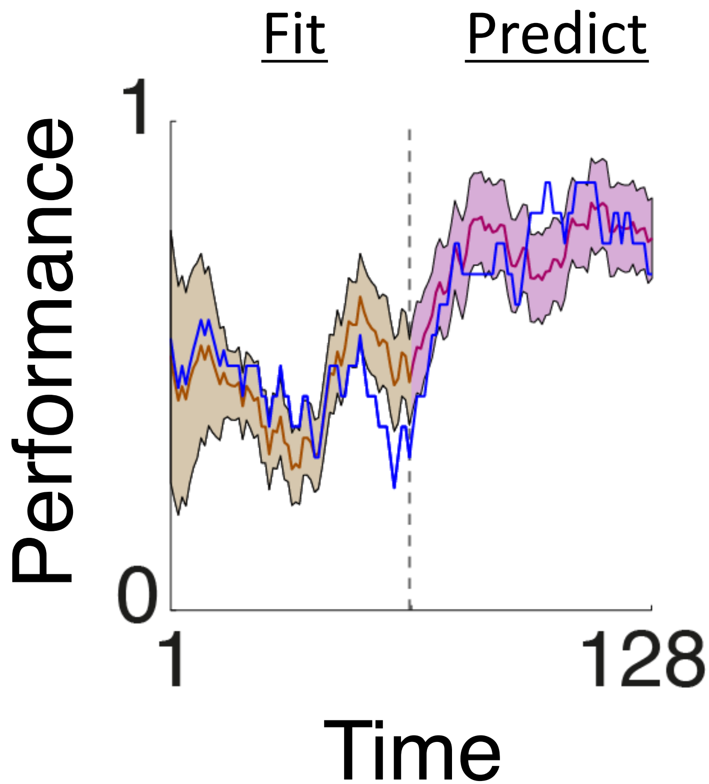
$$p(\vec{x}|y) = \frac{1}{Z}\exp\{\beta \sum_{\mu} \alpha_{\mu}(t) f_{\mu}(\vec{x})\}$$

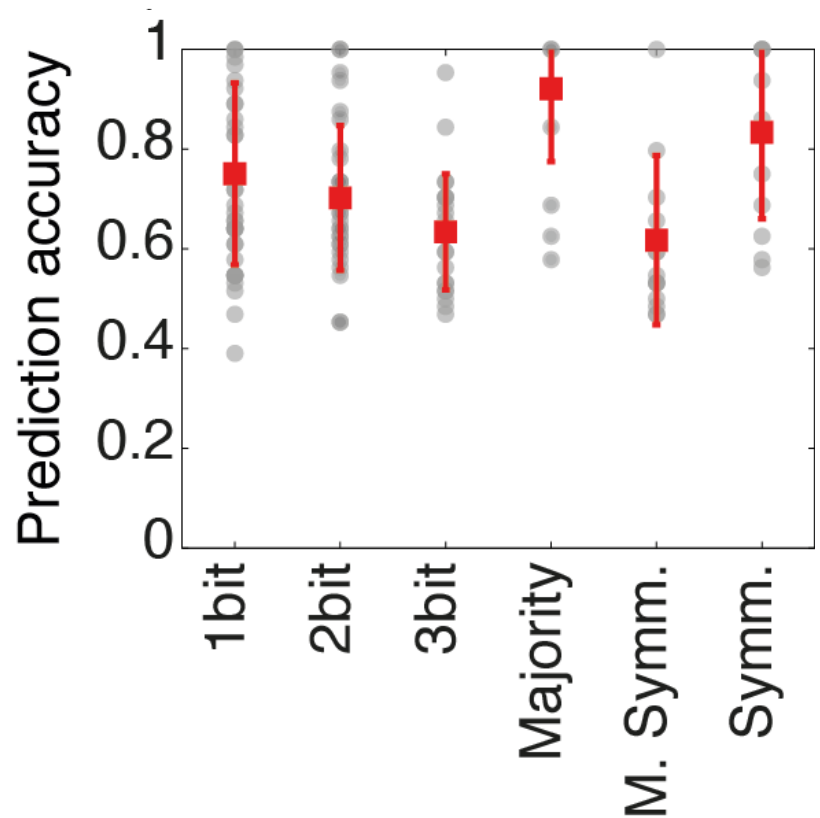




(Cohen & Schneidman, *PNAS*,2013)

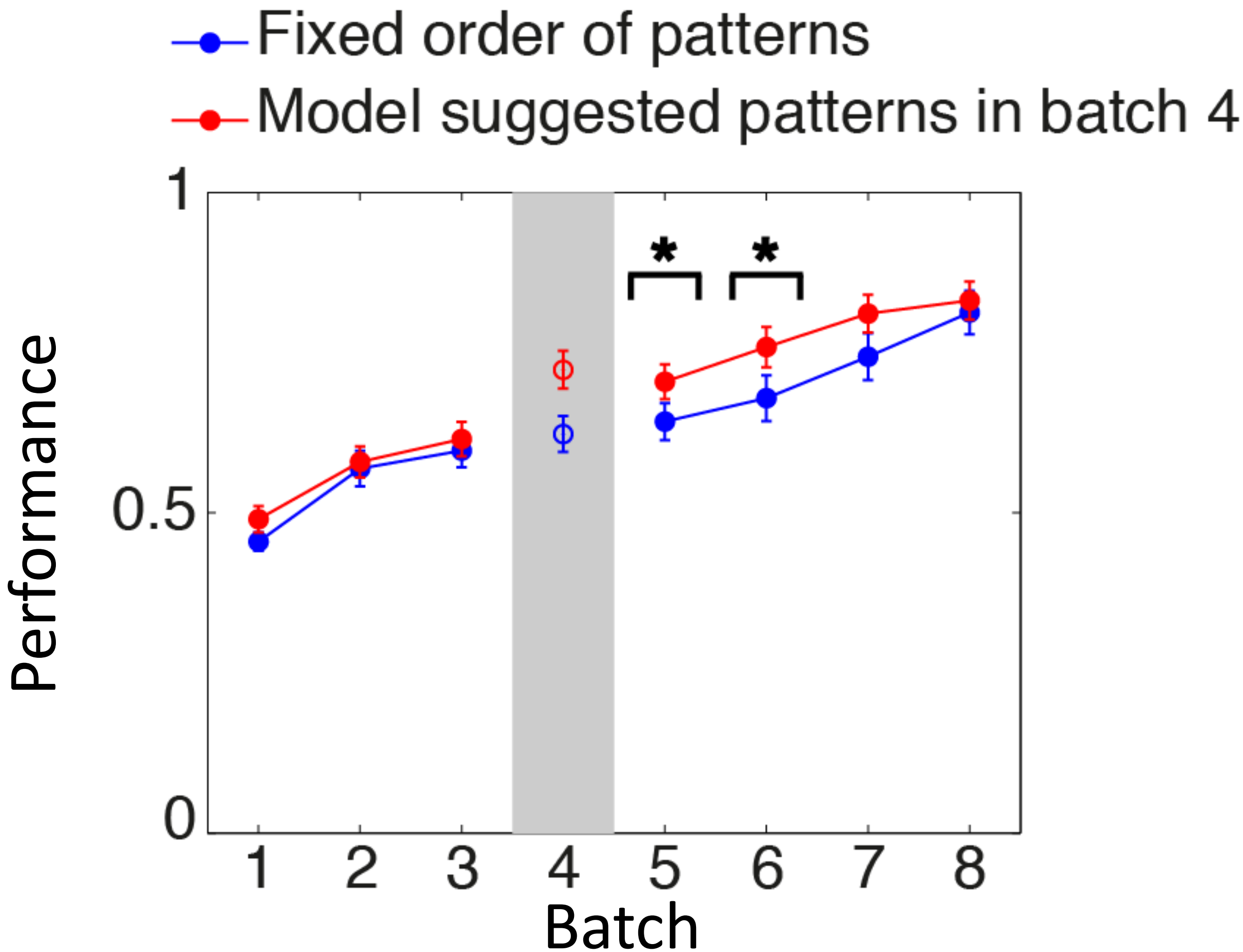
Models predict future answers





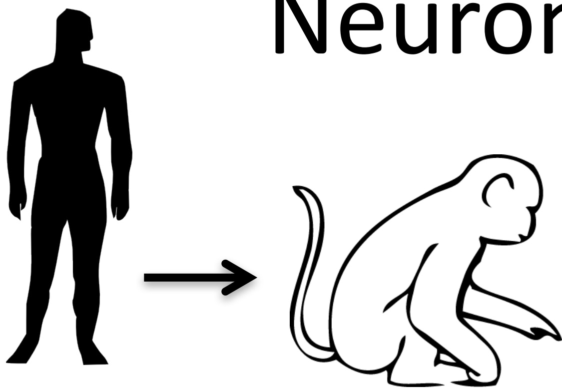
(Cohen & Schneidman, *PNAS*,2013)

Models can be used to improve learning



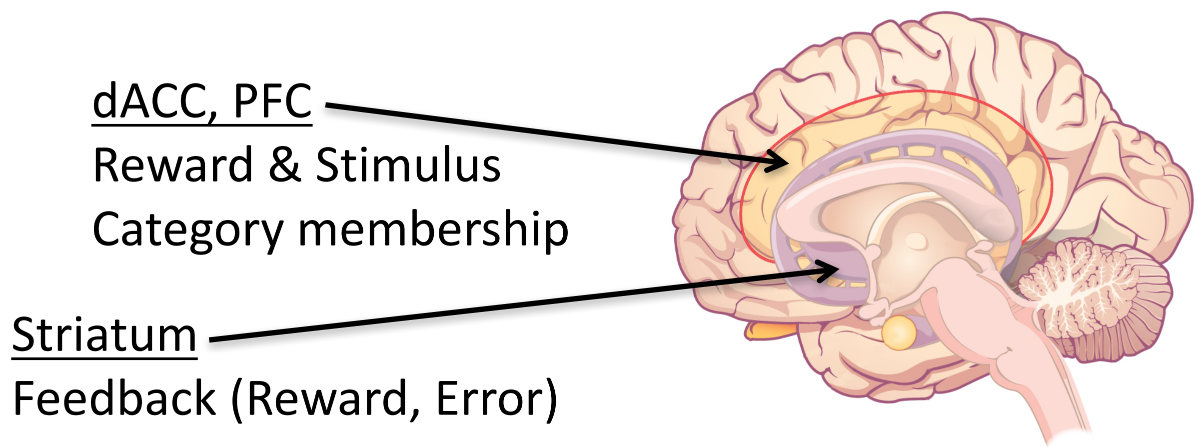
(Cohen & Schneidman, *PNAS*,2013)

Neuronal correlates of learning components

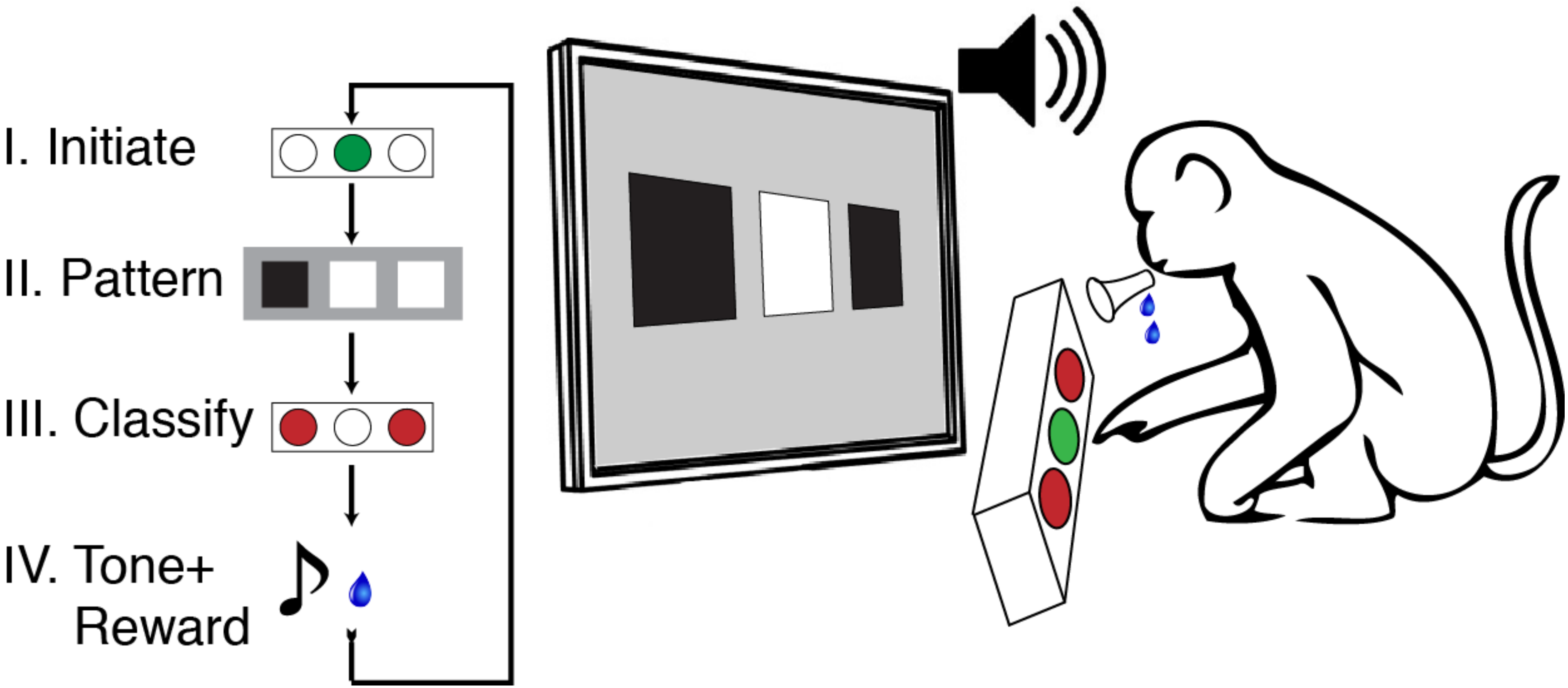


To study learning related dynamics:

- Record in acquisition of new complex rules
- Use conceptually different rules

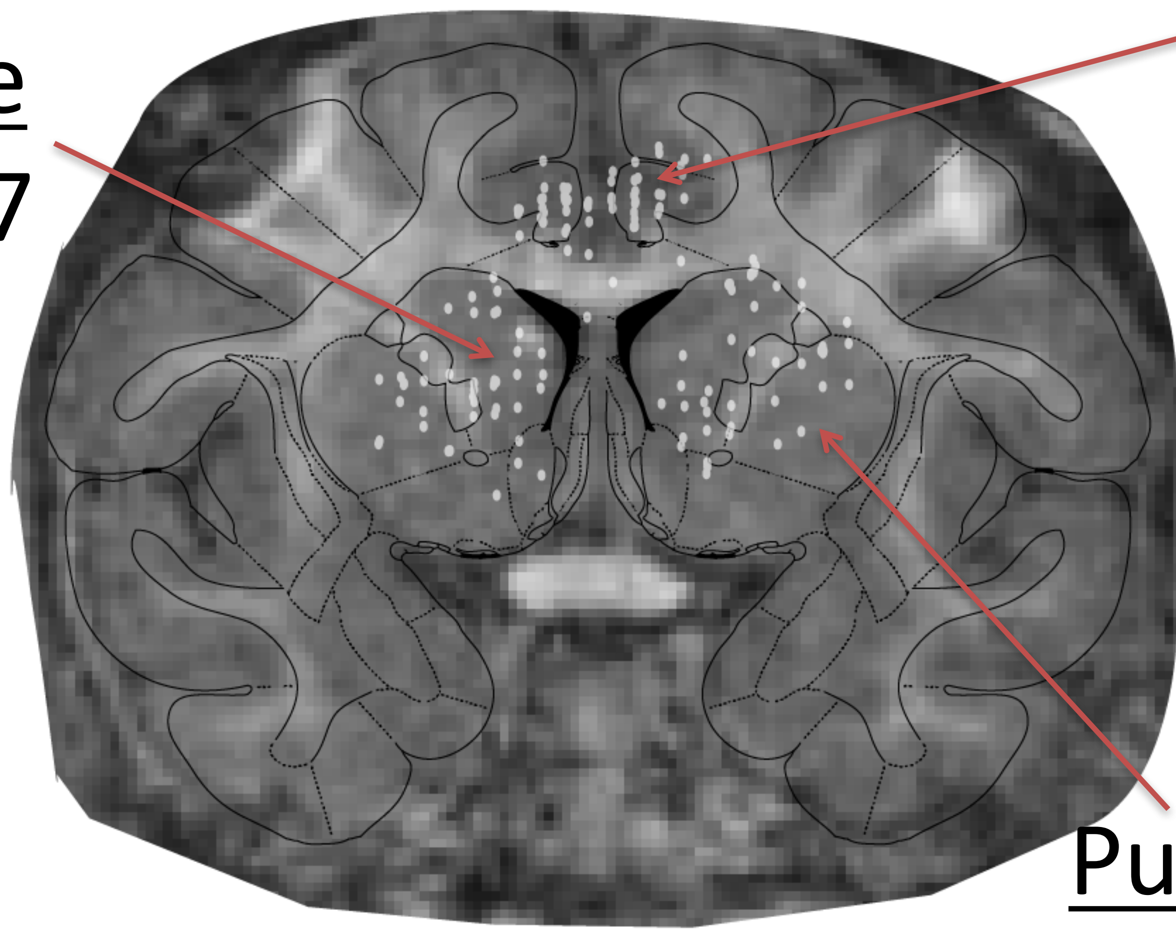


Monkeys learned to classify binary patterns



We recorded from dACC, Caudate and Putamen

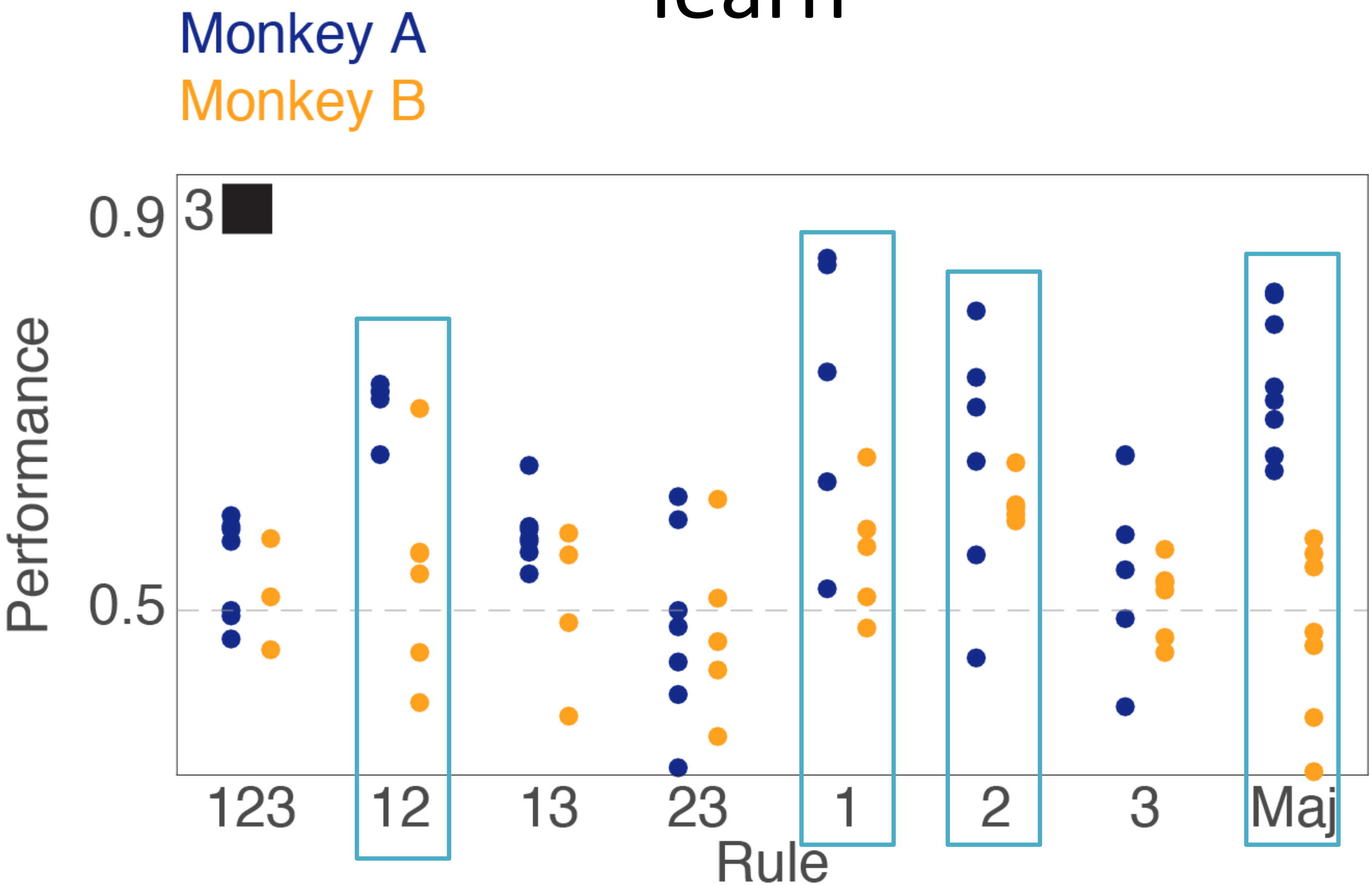
Caudate
N=98,97



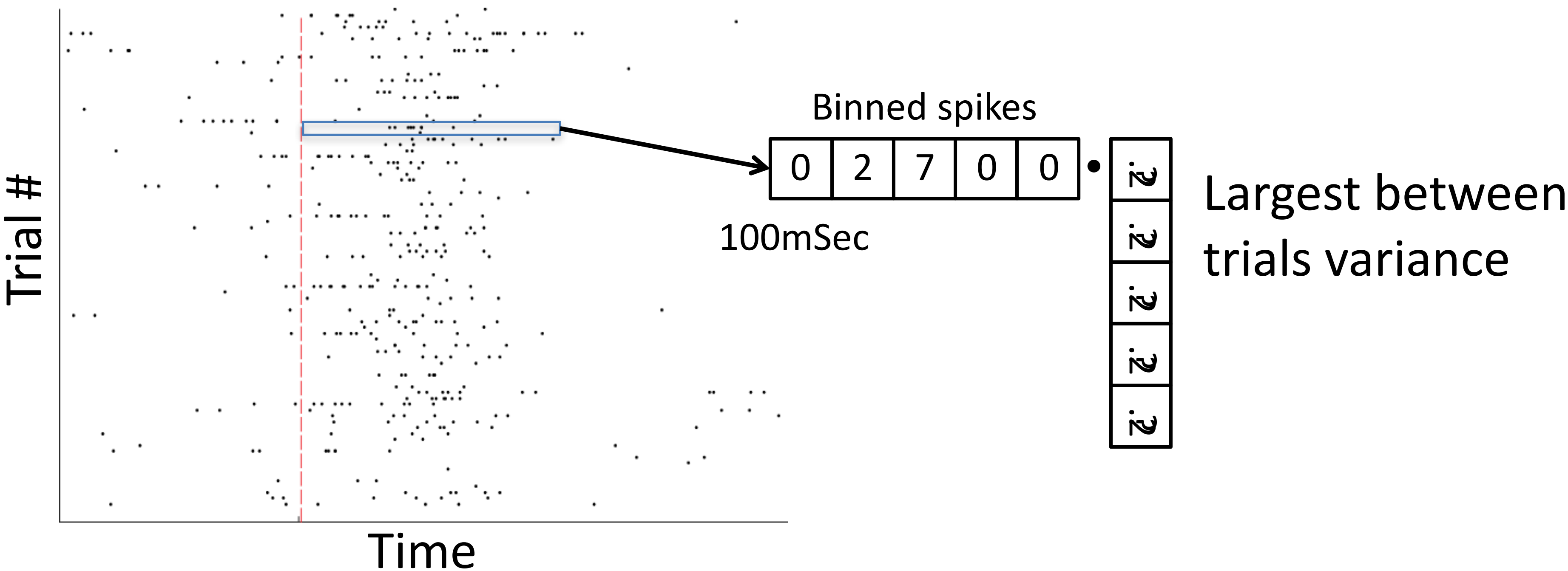
dACC
N=309, 440

Putamen
N=93,103

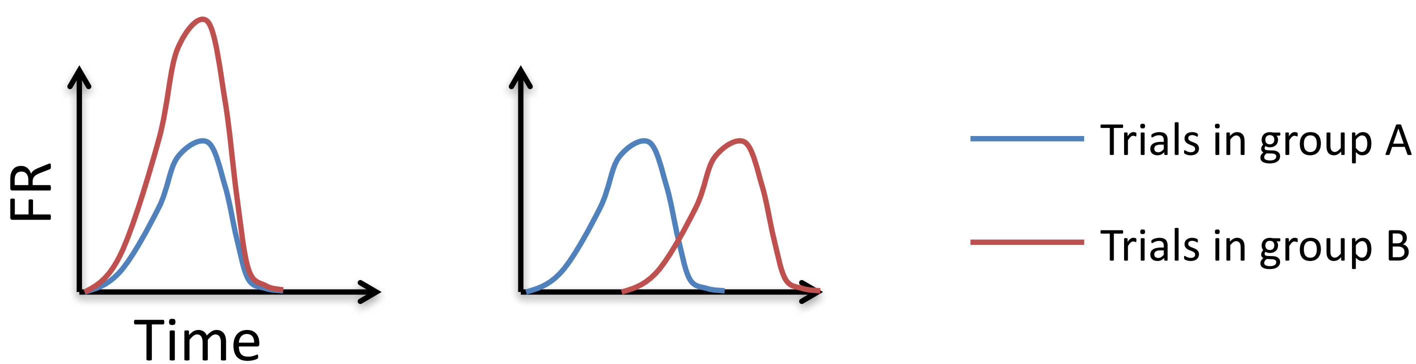
Monkeys were different but both could learn



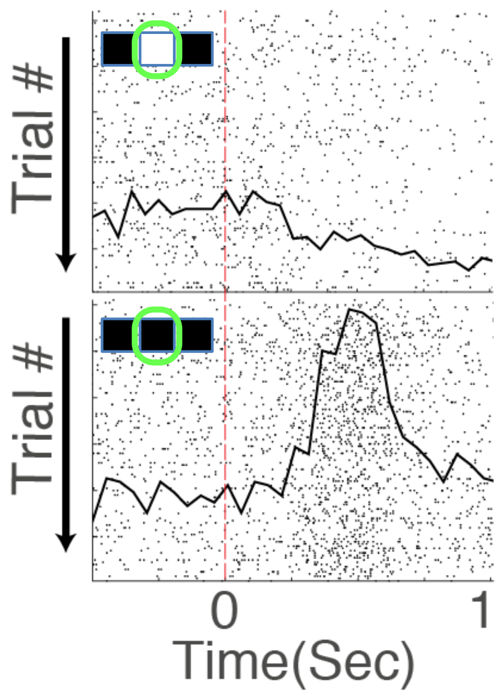
Spike train analysis for identifying feature selective neurons

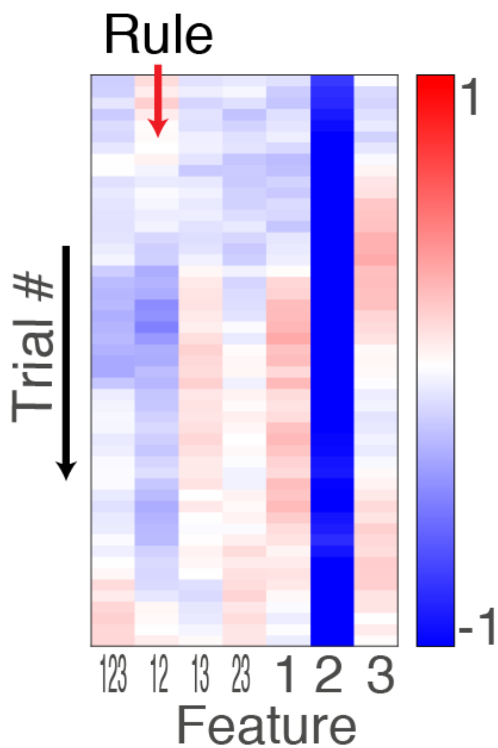


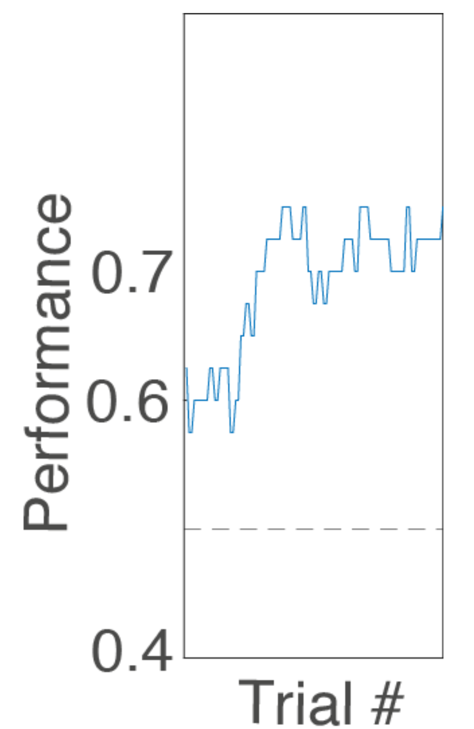
Feature sensitivity leads to variance in spiking

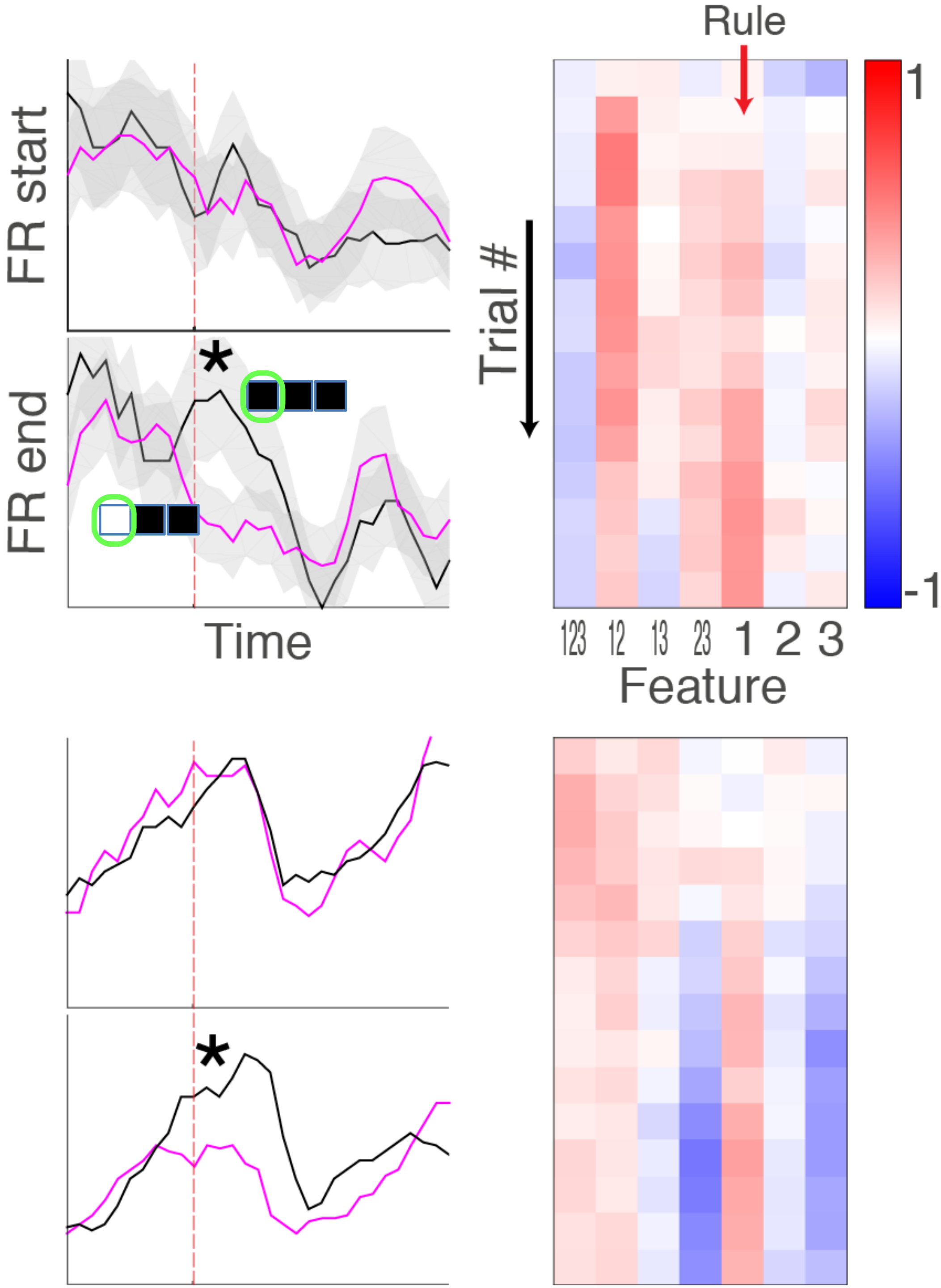


Neurons with stable feature correlations

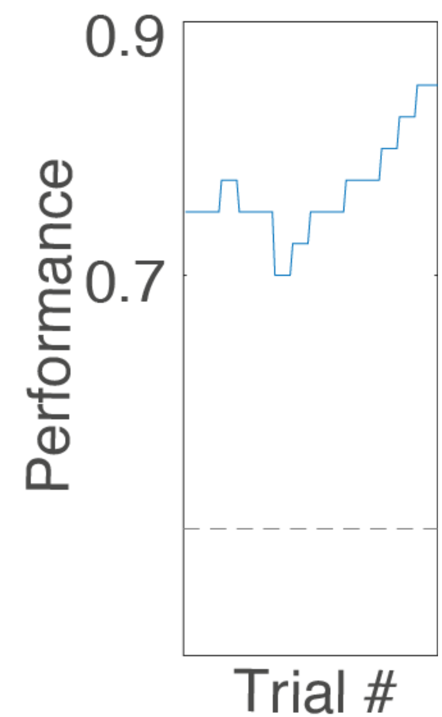




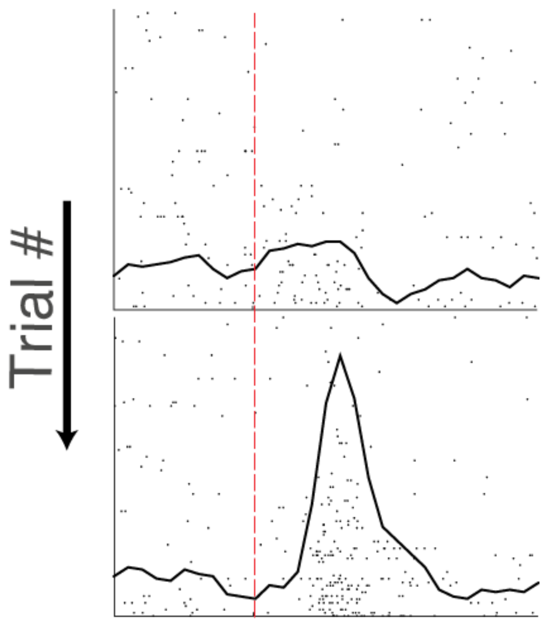


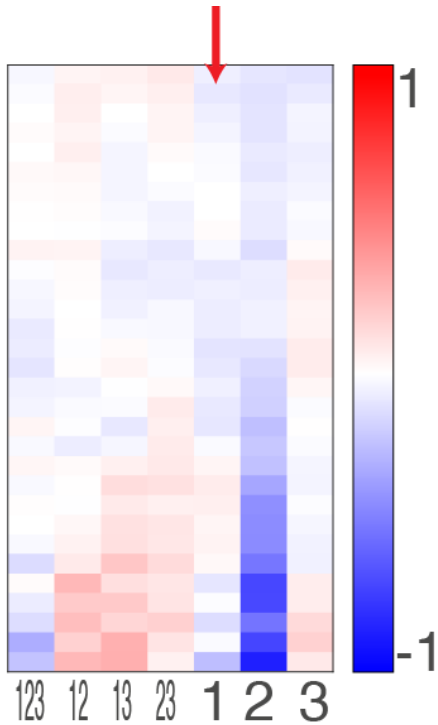


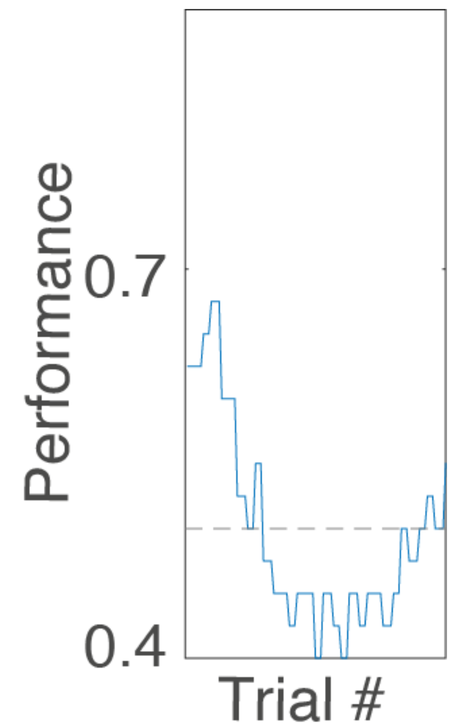
Moving feature correlations



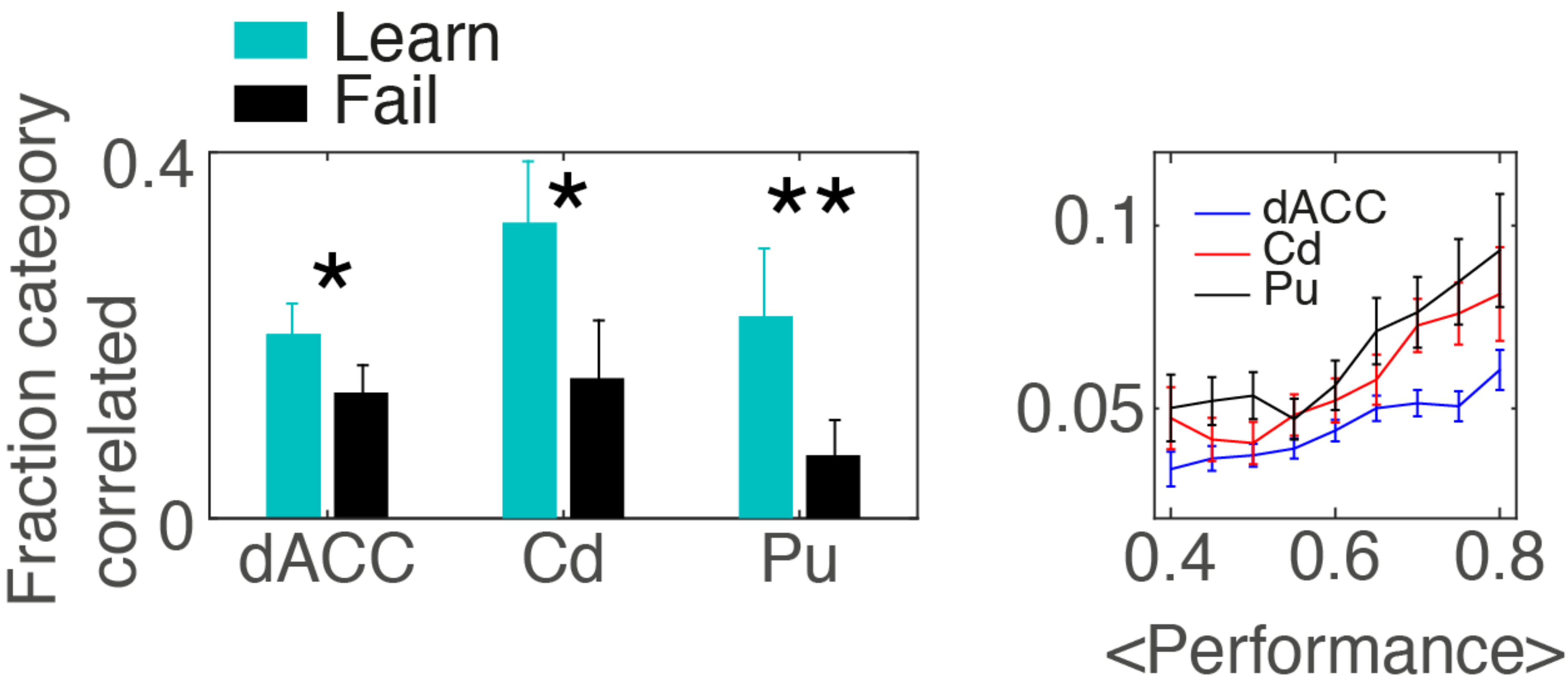
Moving feature correlations in failed sessions



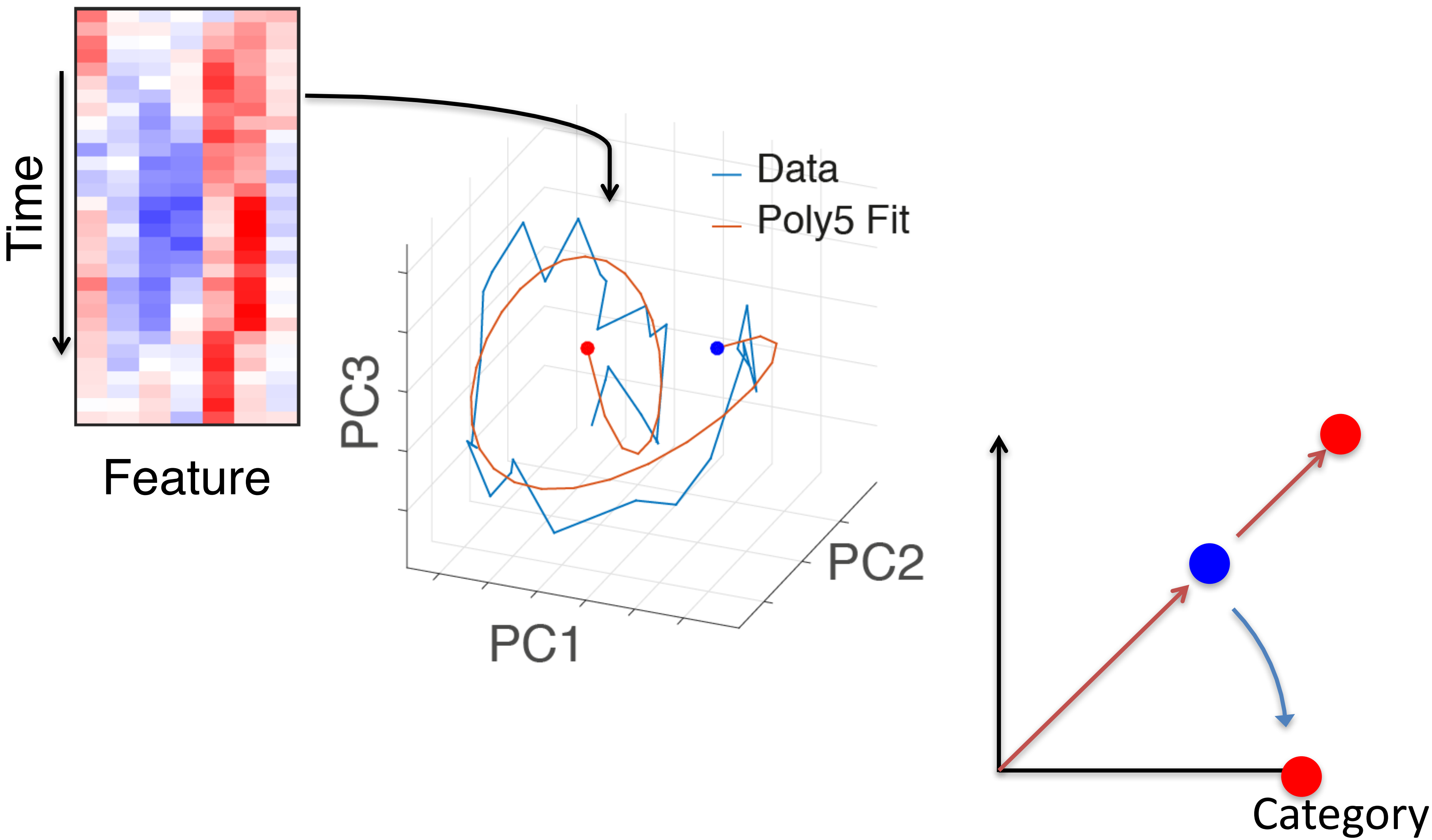




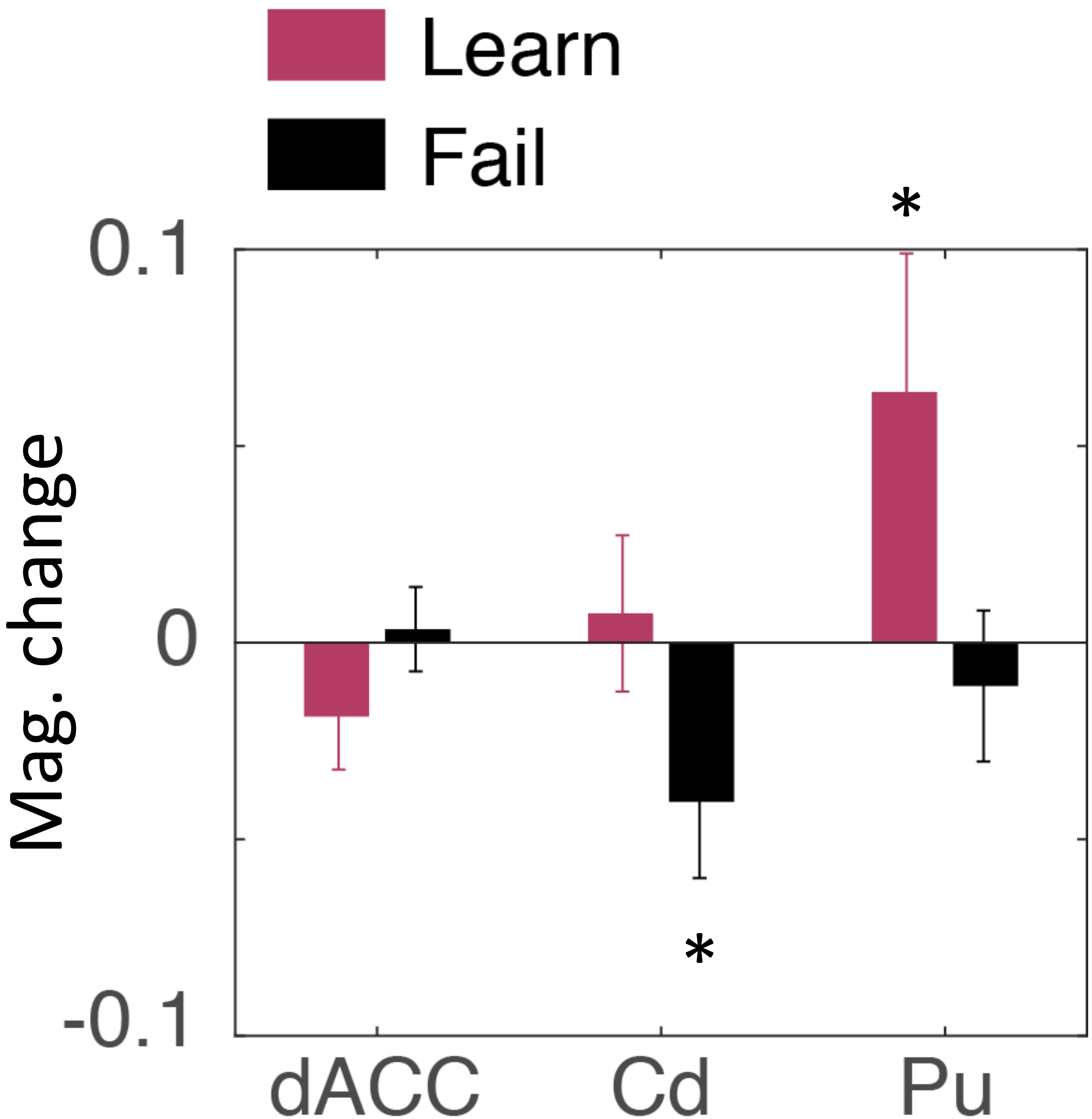
More category correlated neurons in learned rule and high performance

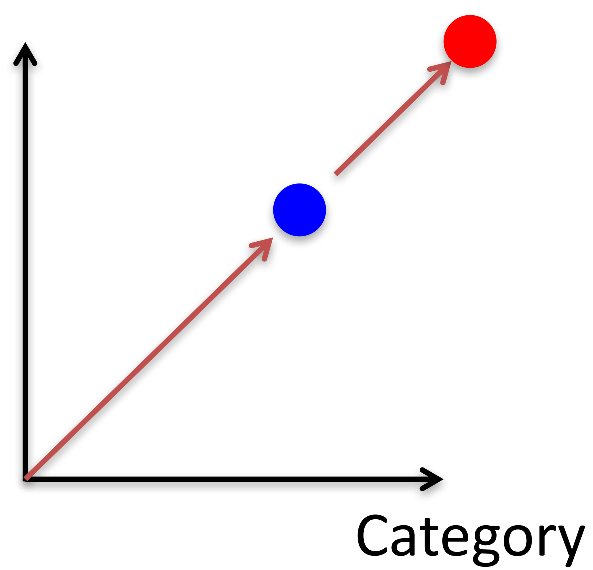


Analysis of high dimension trajectory

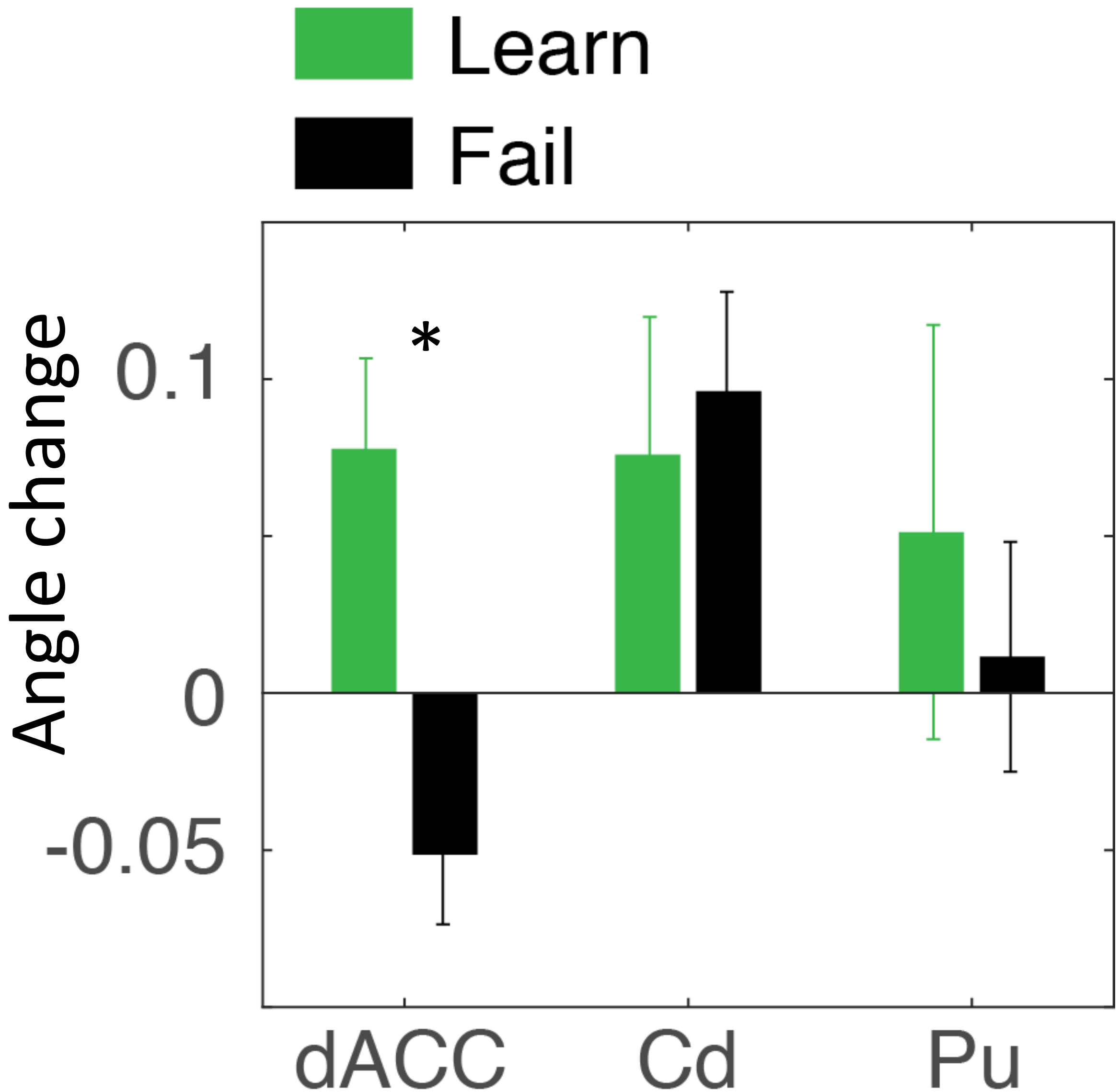


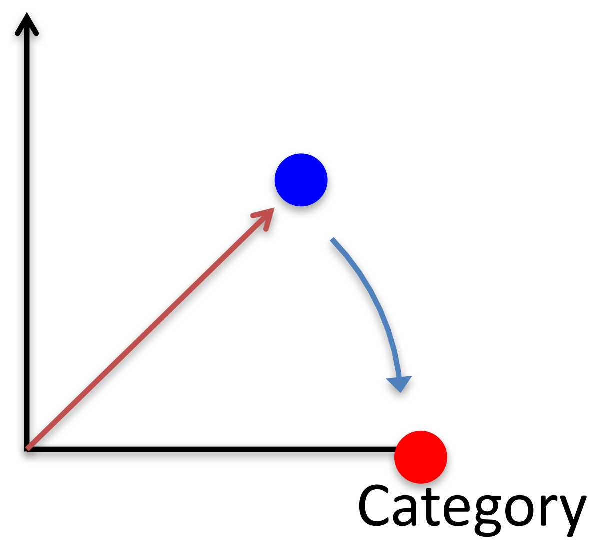
Magnitudes change in the Striatum



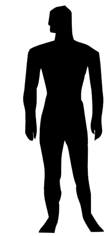


Directions change in dACC,Cd

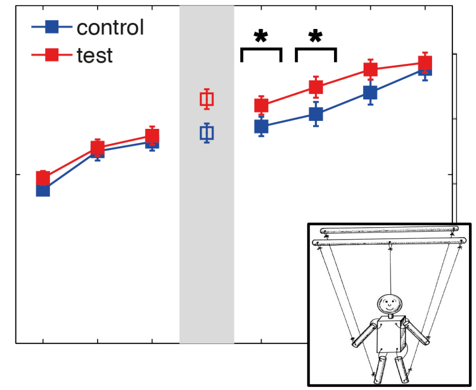




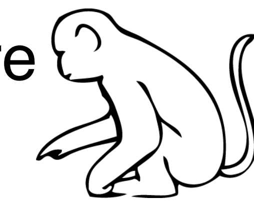
Conclusions



Feature based models predict individual behavior and enable personalized teaching

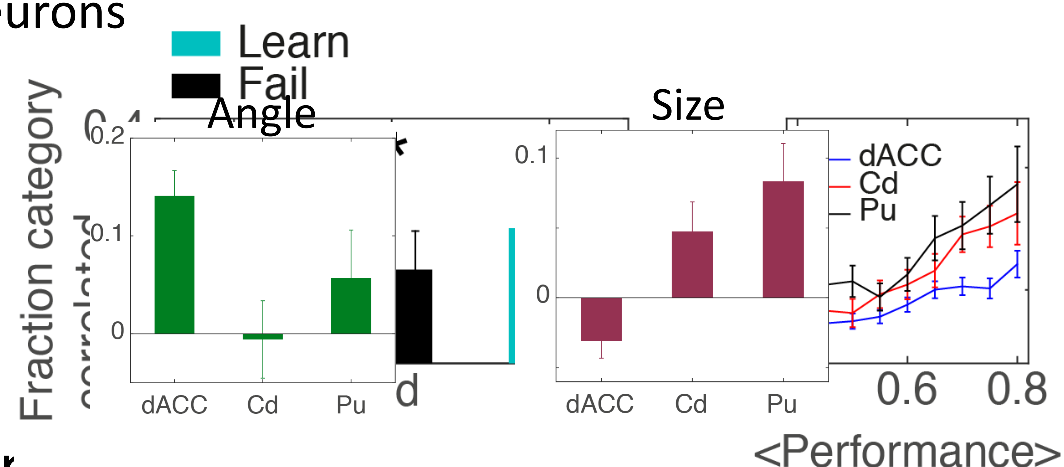


- Describe the broad range of behavior
- Separate the prior from simple learning dynamics
  - Predict behavior
  - Use models to choose personalized teaching sequence



Learning manifests in high dimensional dynamics of feature correlations that leads to increase in category correlation

- Fraction of category correlated neurons
  - Increases for learned rules
  - Increases with performance
- Vectors of feature correlation
  - Increase size in Putamen
  - Rotate in dACC



<u>Next:</u>

- Trajectory of single neuron
- How do neurons move together

Acknowledgments

Rony Paz
Yossi Shochat
Aryeh Taub
Eilat Kahana
Nir Samuel
Shahak Yariv
Yoav Kfir
Tamar Stolero
Netanel Ghatan
Tal Harmelech
Noga Cohen
Eyal Weinreb
Rita Perets
Vered Bezalel
Liran Szlak

Elad Schneidman
Rachel Ludmer
Oren Forkosh
Yair Shemesh
Roy Harpaz
Ori Maoz
Ehud Karpas
Tal Tamir
Amir Bar
Lior Baltiansky
Linor Balilti Torgman

Thank you!