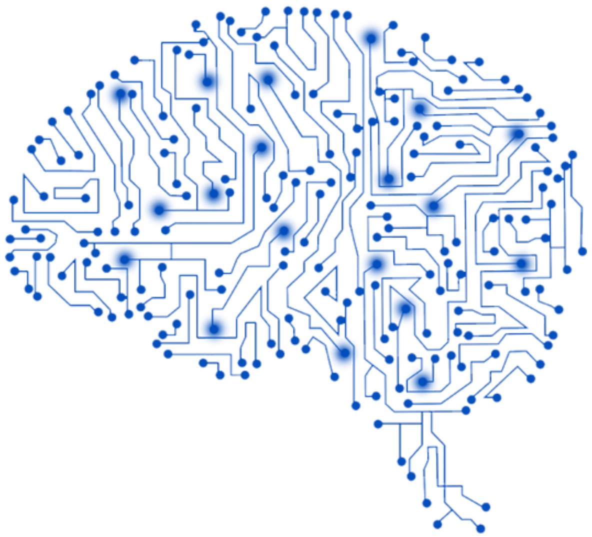
Understand The Brain Using Interpretable Machine Learning Models

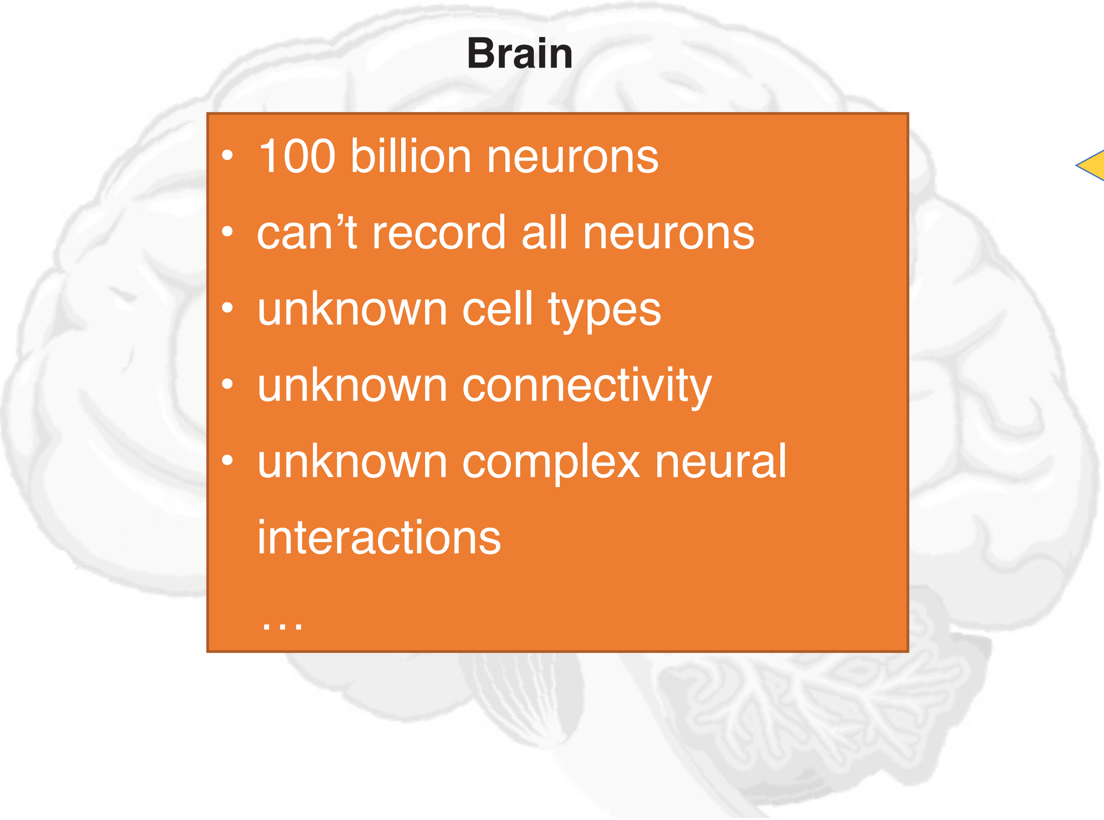


Anqi Wu

School of Computational Science and Engineering
College of Computing
Georgia Institute of Technology

Fundamental challenges in neuroscience

Brain



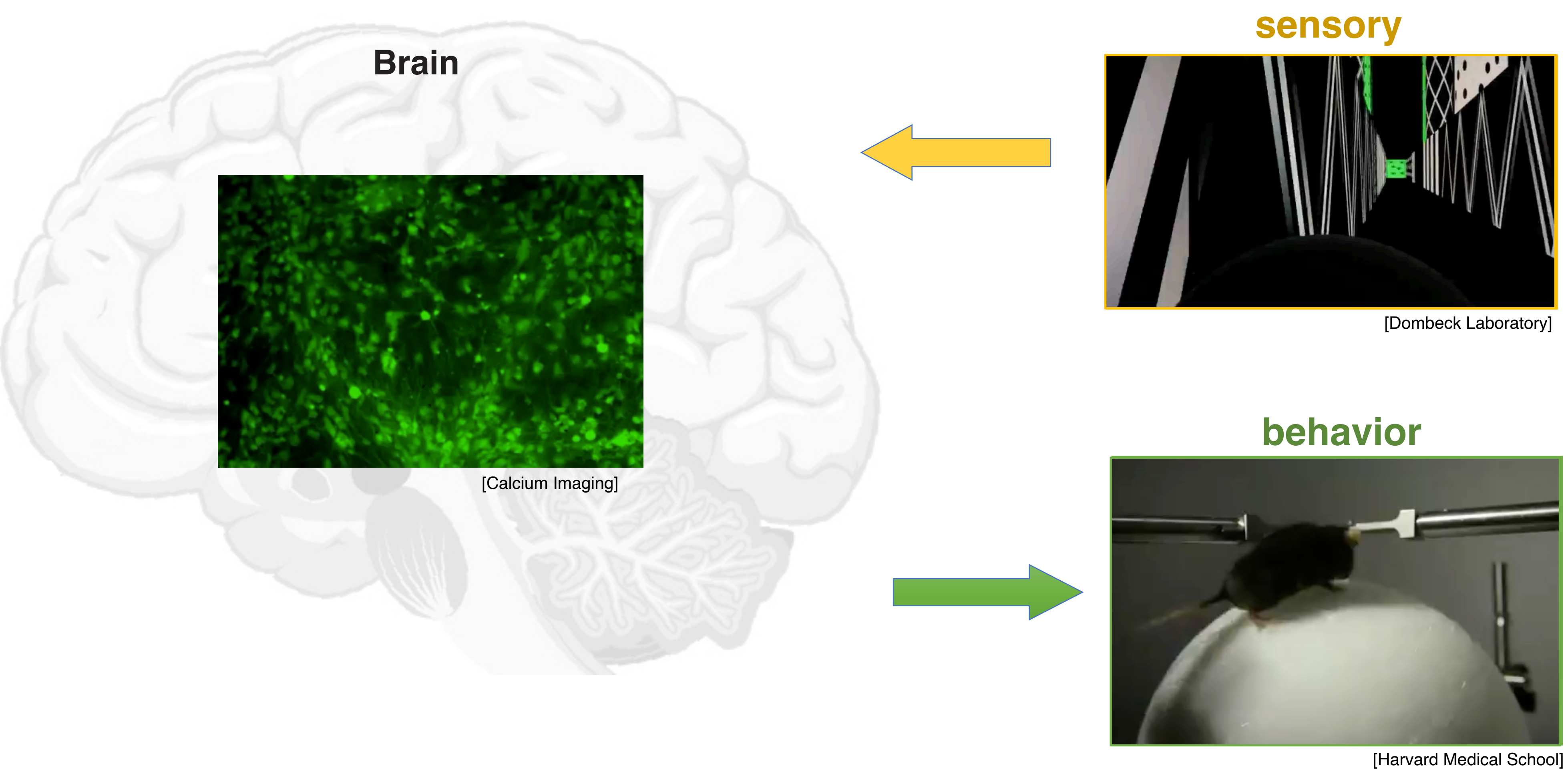
- 100 billion neurons
- can't record all neurons
- unknown cell types
- unknown connectivity
- unknown complex neural interactions

…

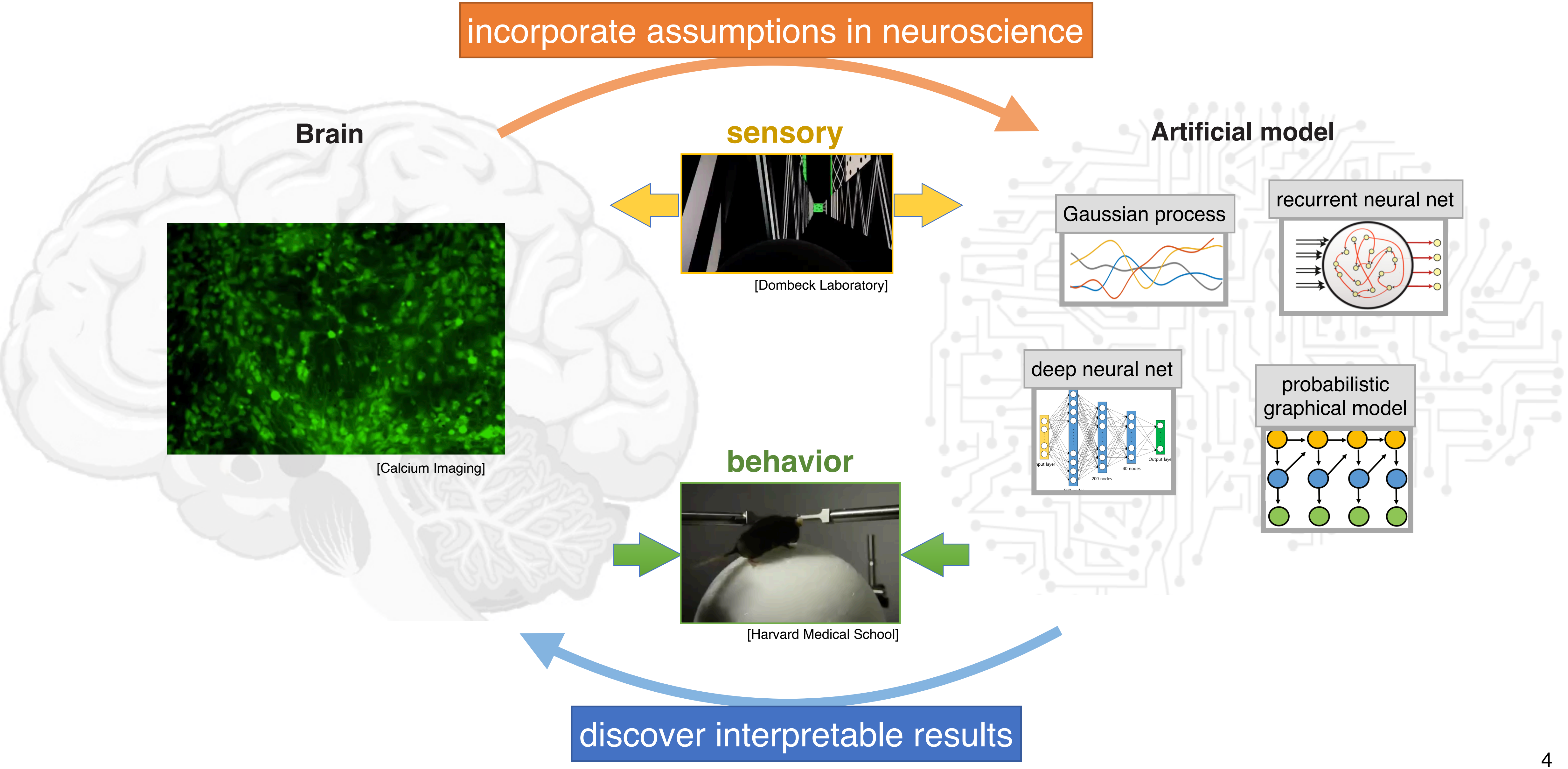
sensory

behavior

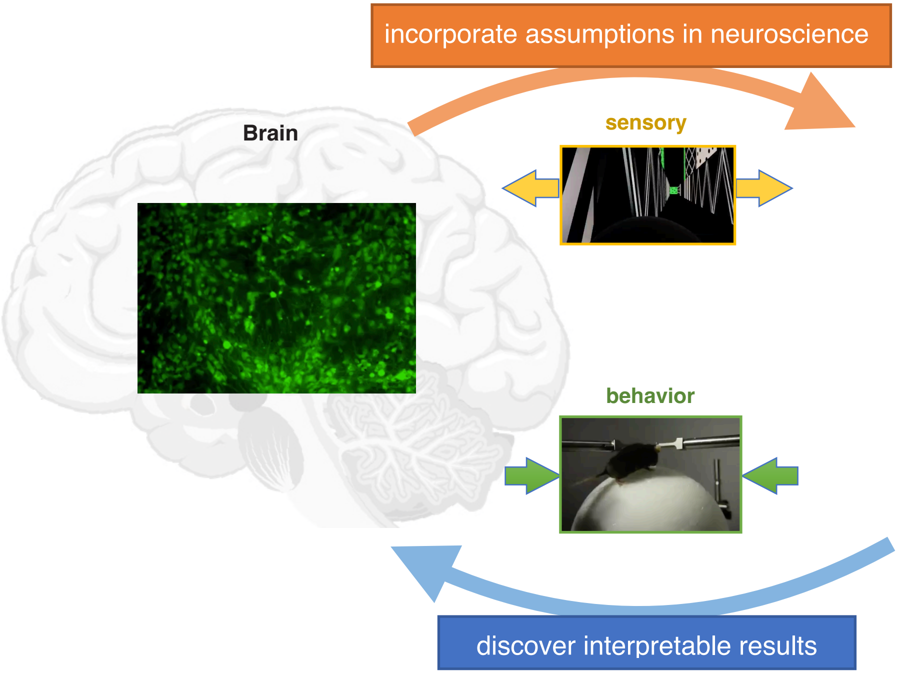
Fortunately, we have large-scale neuroscience data



Data-driven machine learning for neuroscience



Outline



Artificial model

- Latent structure discovery for neural recordings
- Structured priors for fMRI brain decoding
- Semi-supervised learning for animal behavior analysis and understanding

Outline

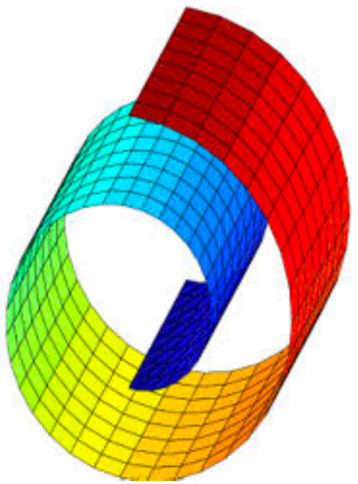
Latent structure discovery
for neural recordings

Structured priors for fMRI
brain decoding

Semi-supervised learning for
animal behavior analysis and
understanding

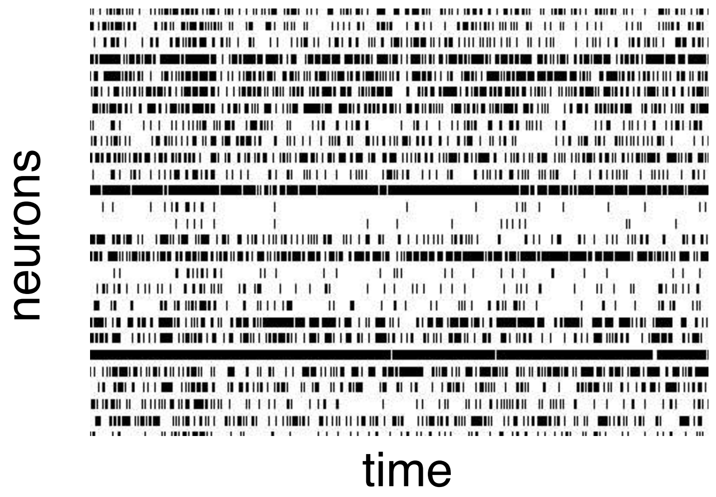
Problem: discover latent structure from neural spike trains

low-D structure



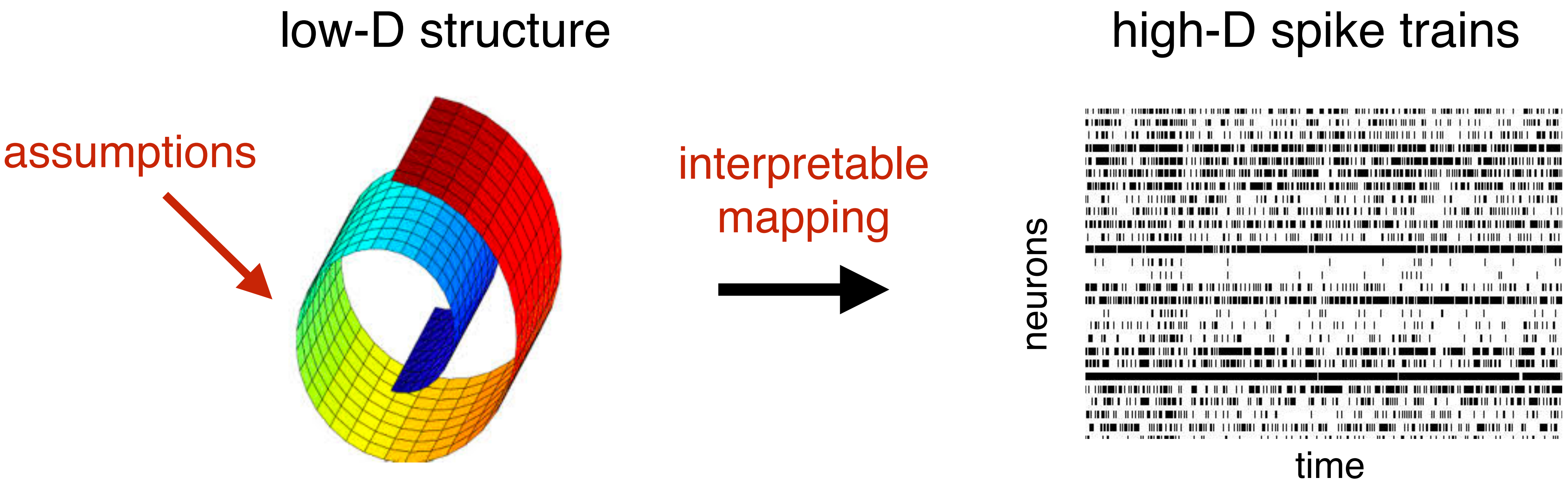
latent dimensionality (low-D) X time

high-D spike trains



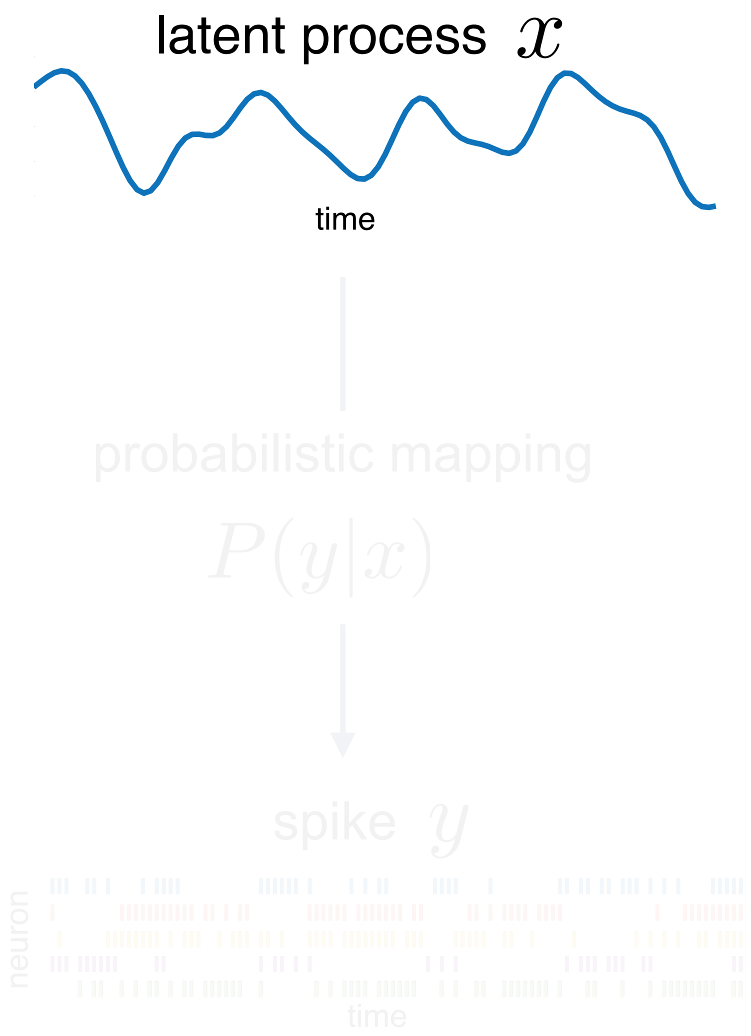
neurons (high-D) X time

Why build our own model?



- Generative model for population spike trains
- Incorporate assumptions inspired from neuroscience
- Explicit manifold assumption over latent structure
- Interpretable nonlinear mapping functions inspired from real neurons

Latent Manifold Tuning (LMT)



prior over latent

$$P(x)$$

Assumption I: latent process evolves smoothly

Gaussian process (GP) prior

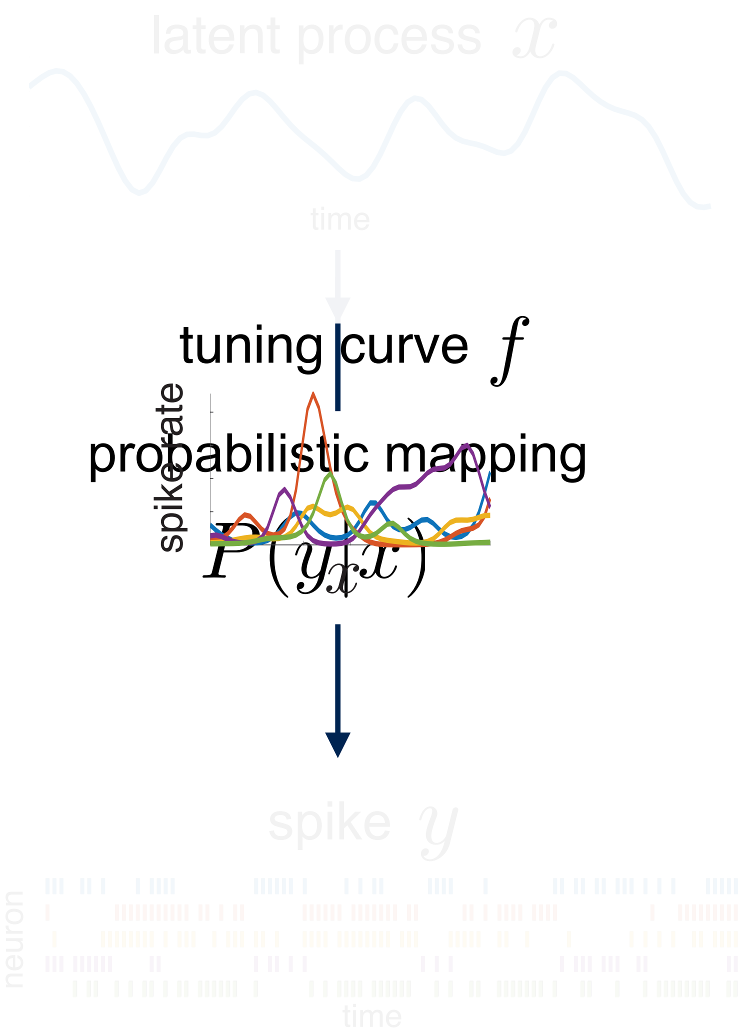
$$x(t) \sim GP(0, K_t)$$

governing smoothness of x(t)

ensuring temporal continuity

[Wu et al NeurIPS2017]

Latent Manifold Tuning (LMT): a generative model



prior over latent

$x(t) \sim GP(0, K_t)$ ensuring temporal continuity

tuning curve:

characterize relation between ~~input stimuli~~ latent variable and spike rate

Assumption II: outputs of tuning curves are non-negative

Assumption III: neurons fire smoothly over latent space

for neuron i

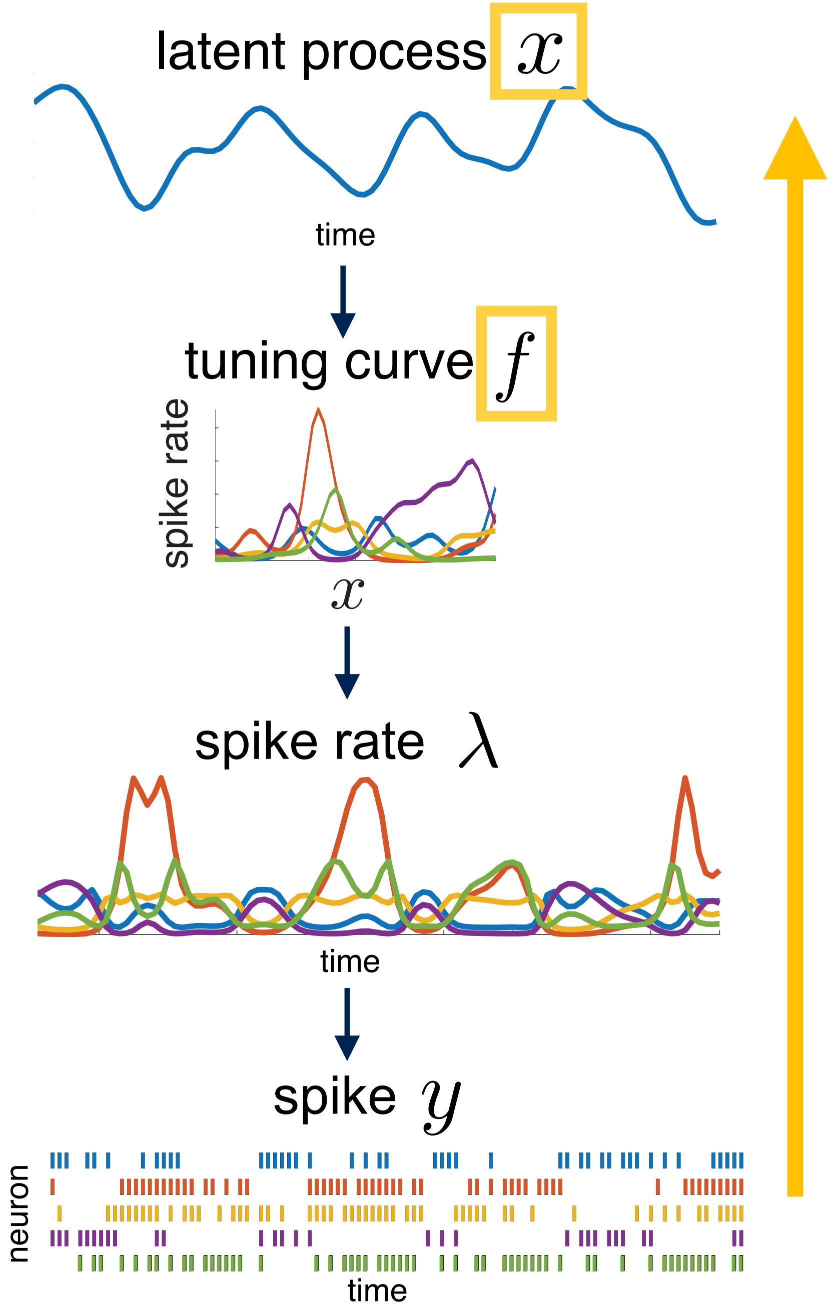
$$\log f_i(x) \sim GP(0, K_x)$$

governing smoothness of f(x)

non-negative smooth tuning curve

[Wu et al NeurIPS2017]

Latent Manifold Tuning (LMT): a generative model



prior over latent

$x(t) \sim GP(0, K_t)$ ensuring temporal continuity

probabilistic mapping

for neuron i

$\log f_i(x) \sim GP(0, K_x)$ non-negative smooth tuning curve

spike rate

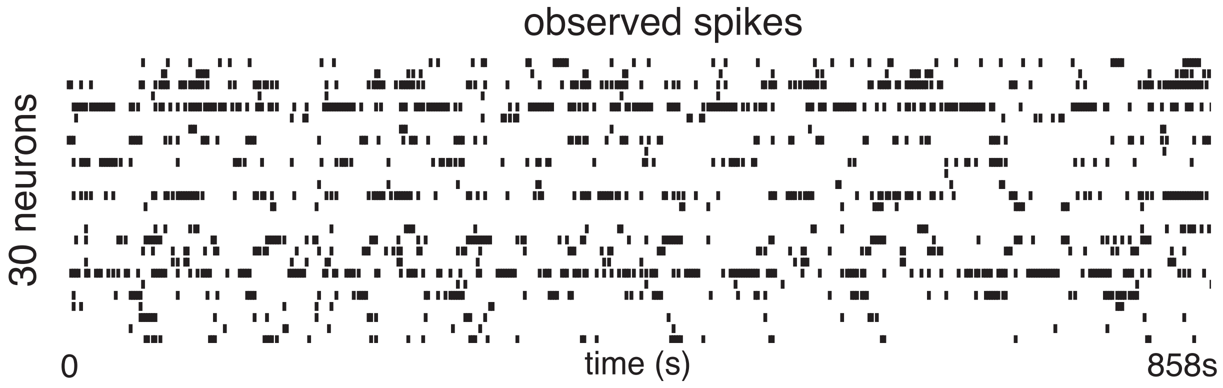
$\lambda_i(t) = f_i(x(t))$ non-negative continuous

Poisson spiking

$y_i(t) \sim \mathrm{Poiss}(\lambda_i)$ non-negative discrete count

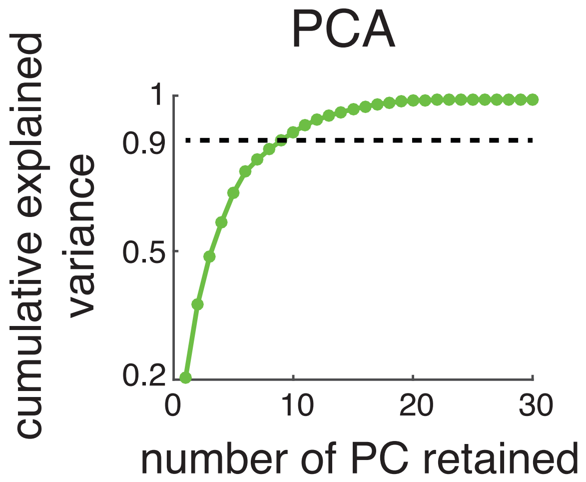
[Wu et al NeurIPS2017]

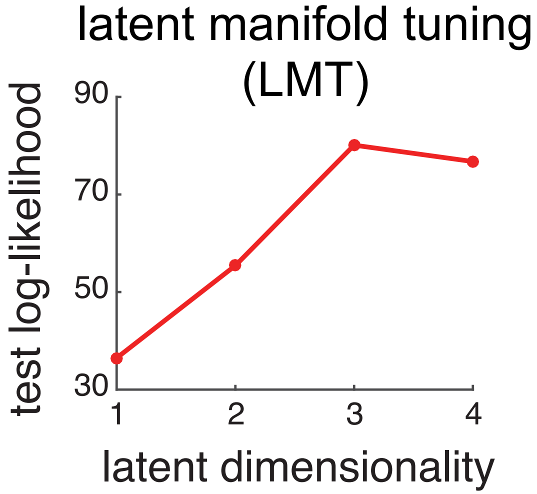
Application to rat hippocampus



[Karlsson et al 2008]

Place cells in hippocampus encode 2D spatial locations.

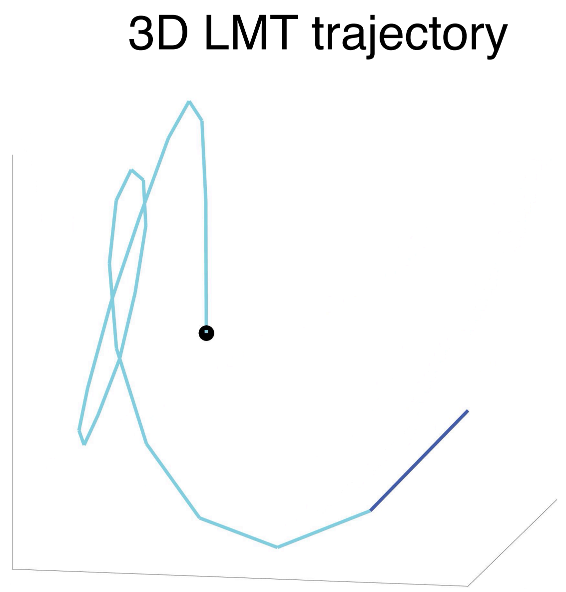




high-dimensional linear space

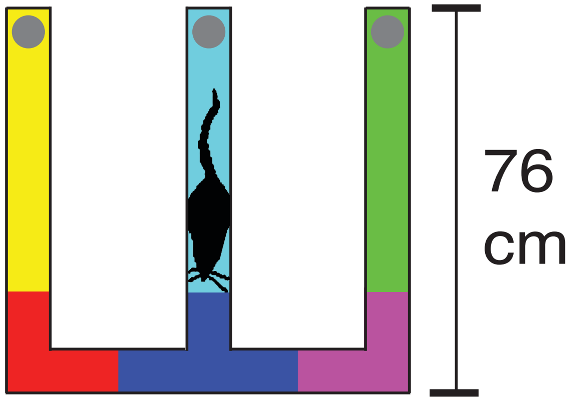
low-dimensional nonlinear manifold

Latent trajectories inferred from spike trains alone

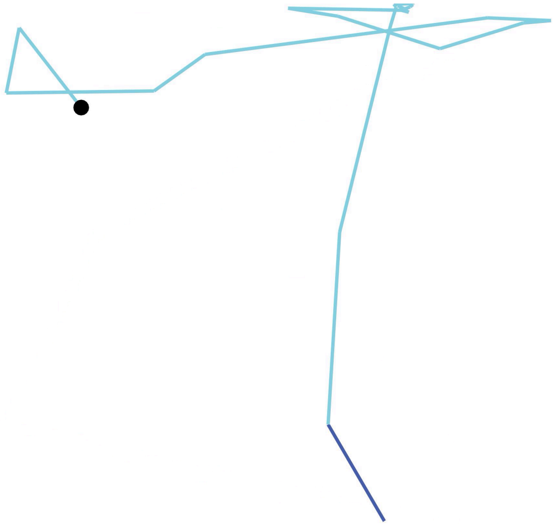


Latent trajectories inferred from spike trains alone

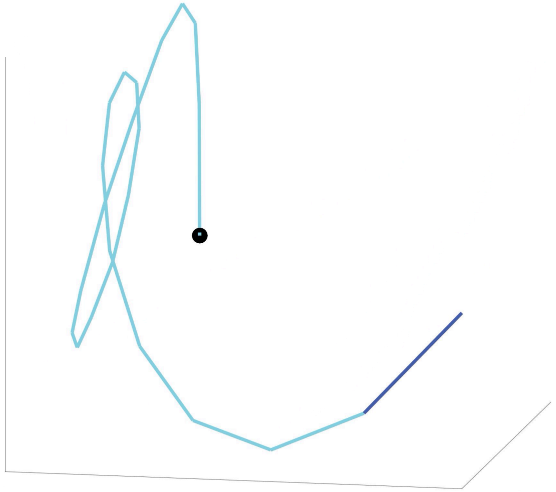
W-shaped maze



2D trajectory

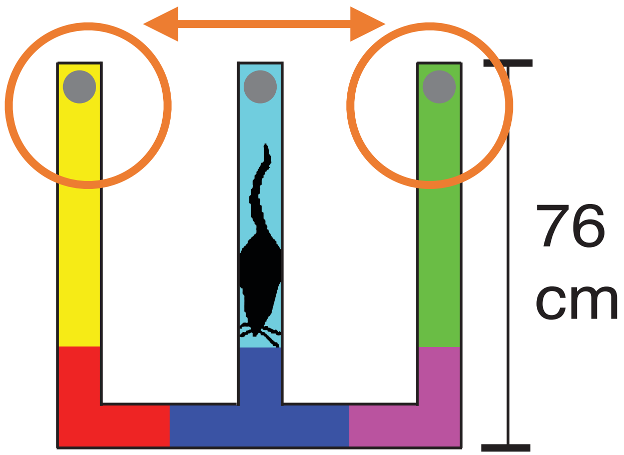


3D LMT trajectory

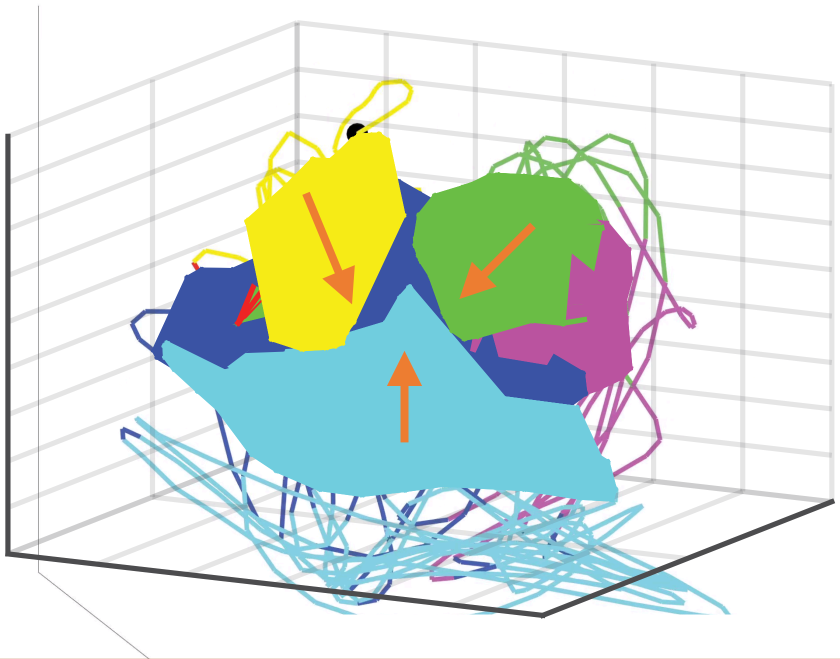


Highlight: interpretable latent manifold

W-shaped maze



3D LMT manifold

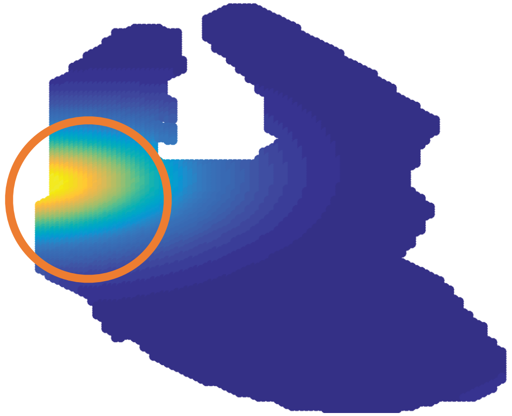


Question: what information is encoded in the 3rd dimension beyond 2 dimensions for 2D spatial locations?

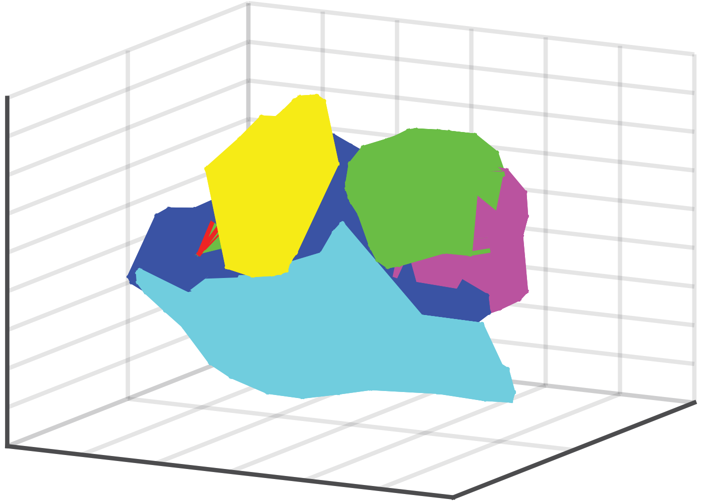
Place cells encode a strong correlation among all three arms in the 3rd dimension.

Tuning curve

estimated tuning curve
of a single place cell



3D LMT manifold

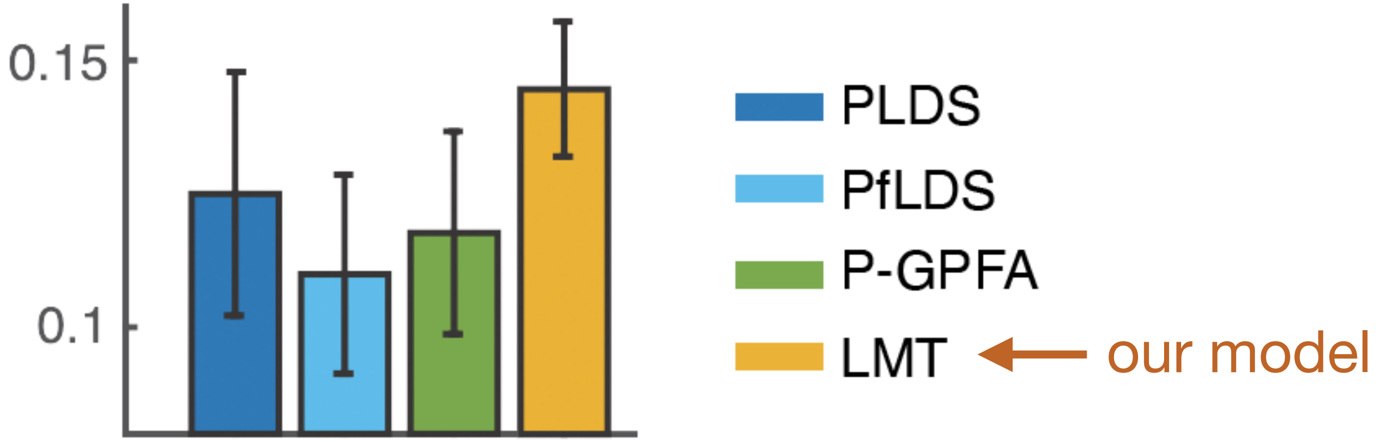


nonlinear mapping

spike trains

Latent manifold tuning model outperforms alternatives

predictive log-likelihood

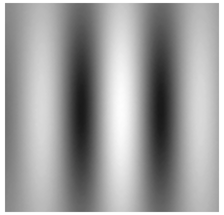


The higher, the better.

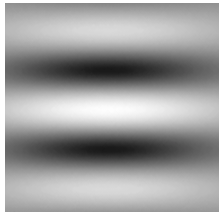
Rotational trajectories underlying cells in visual cortex

drifting sinusoidal stimuli

orientation 270°

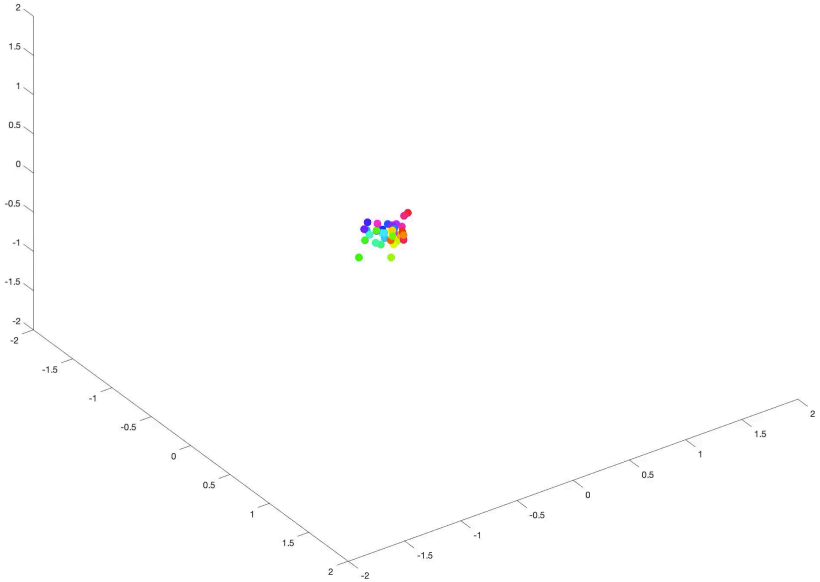


orientation 360°



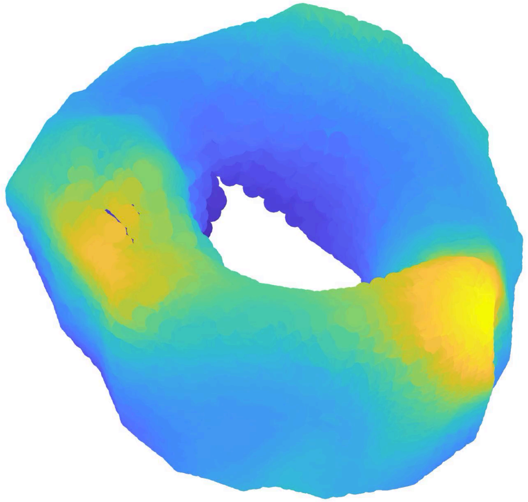
[Graf et al 2011]

latent trajectories



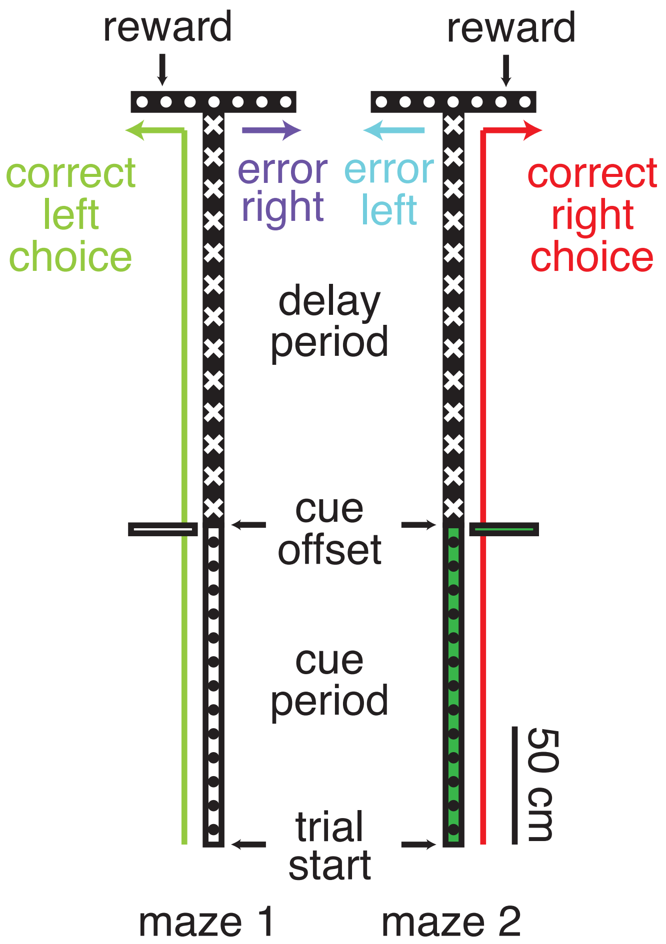
tuning curves

neuron 1

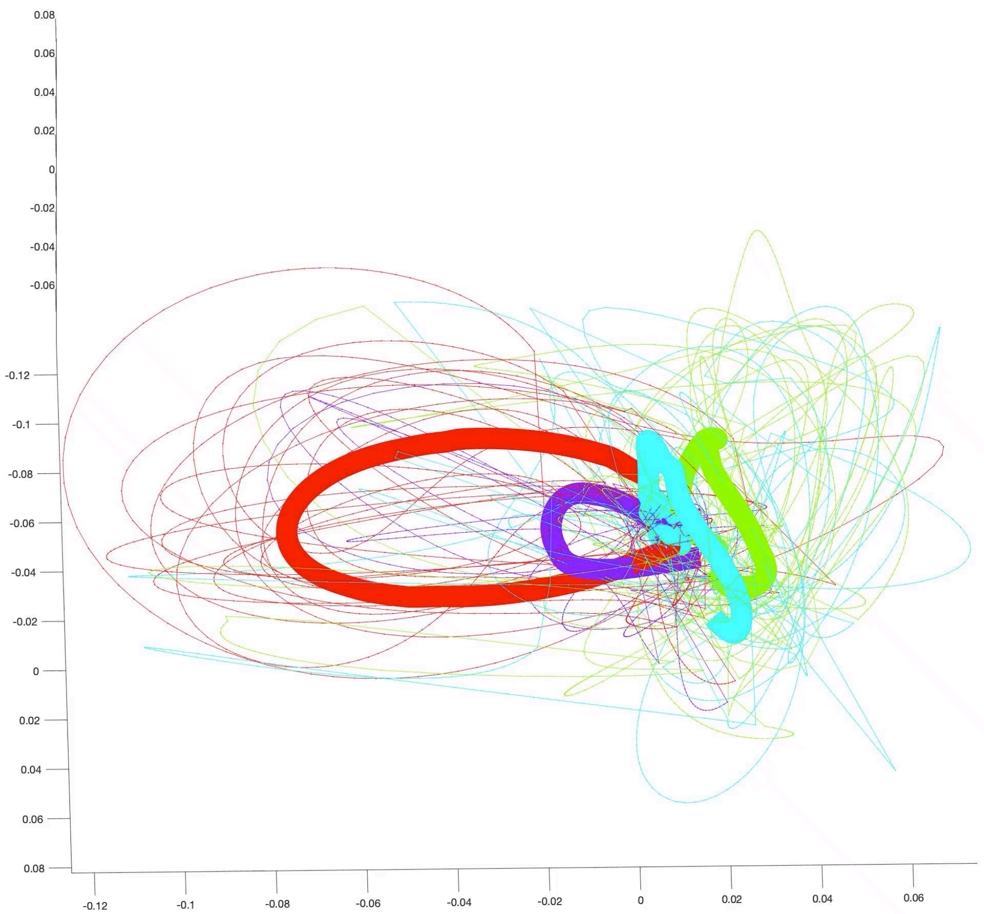


Choice-specific neuronal circuit trajectories in (posterior parietal cortex) PPC

perceptual decision-making



latent trajectories



[Harvey et al 2012]

tuning curves

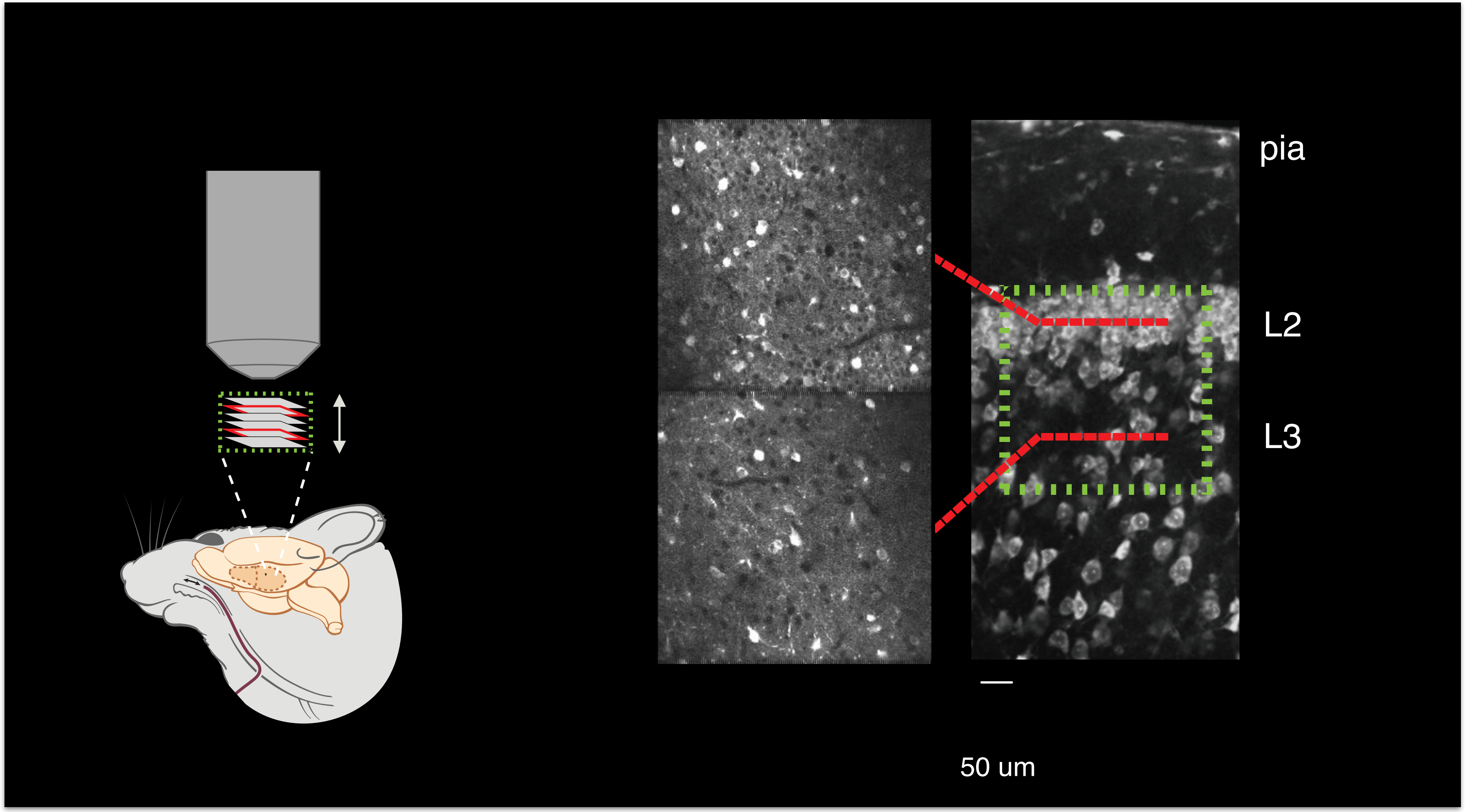
2D manifold

neuron 1

neuron 2

Olfactory topography in piriform cortex

Neural population imaging in piriform cortex



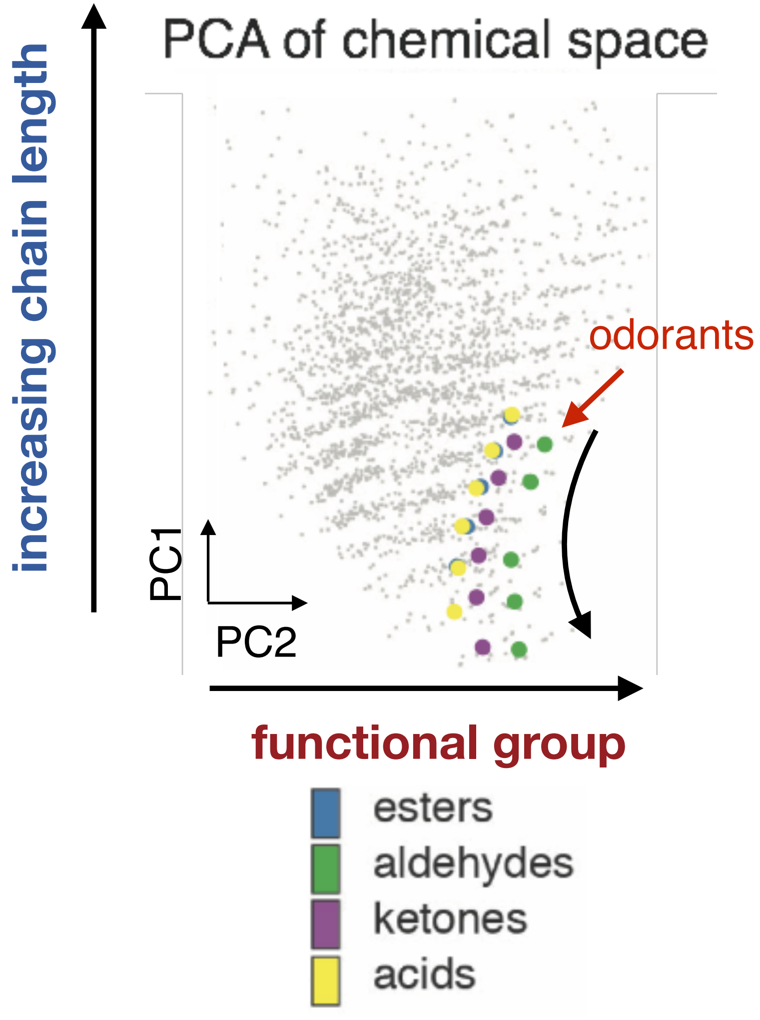
[Bob Datta lab at Harvard]

A variant of LMT:

- ~~spike trains~~
  Calcium imaging
- ~~latent trajectory~~
  latent topography

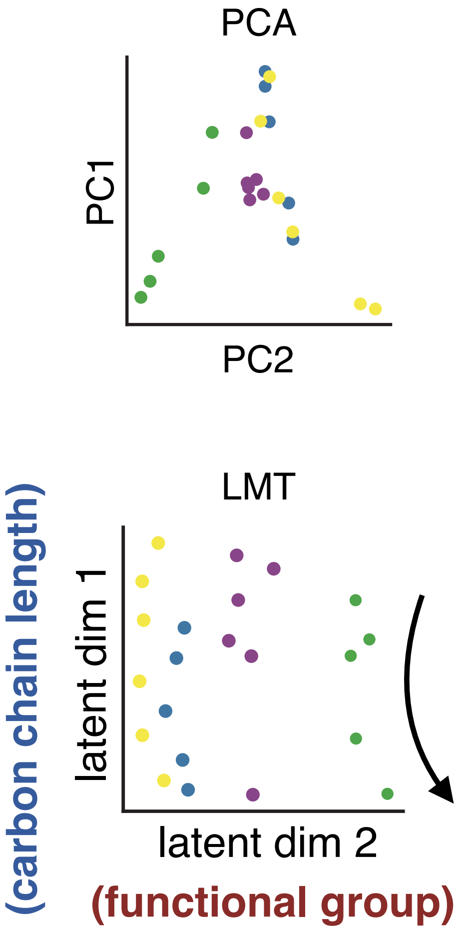
Olfactory topography in piriform cortex

odorant stimuli

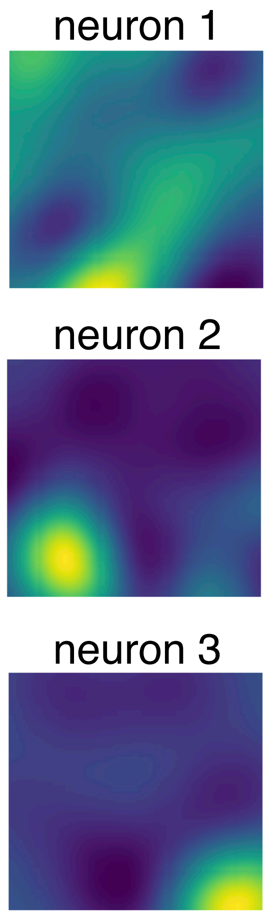


latent topography

(using neural population response only)



tuning curves



[Wu et al NeurIPS2018]

Contributions

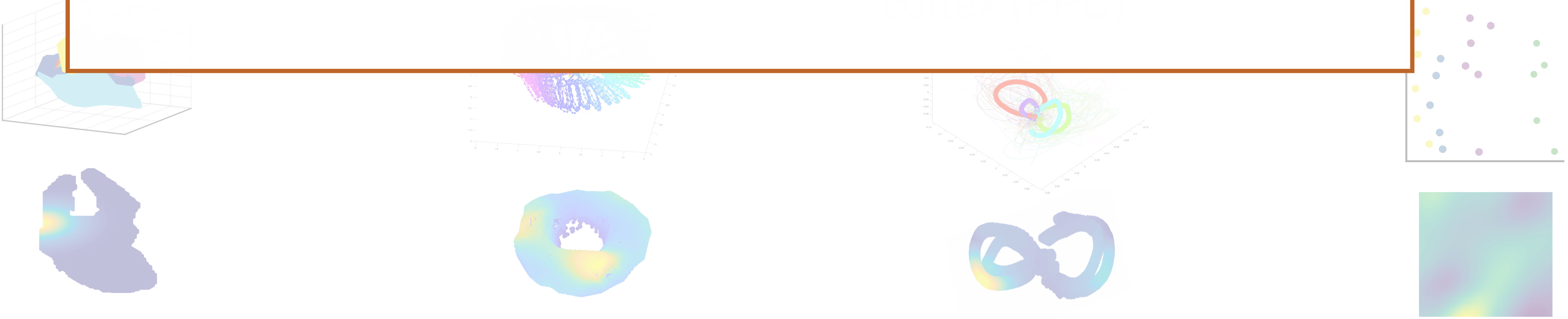
- Translate neuro-inspired assumptions into probabilistic modeling:

Highlight

Interpretable latent structure from high-dimensional neural activity alone

- Generativ
- Explicit ma
- Interpretab ent space

- Extract lo
- place cells i form cortex



Outline

Latent structure discovery for neural recordings

Structured priors for fMRI brain decoding

Semi-supervised learning for animal behavior analysis and understanding

Problem of interest in fMRI analysis

Regression (reverse inference / brain decoding)

response (e.g. reaction time, object label) — brain activity — decoding weights

$$Y = X \cdot \vec{w} + \textit{noise}$$

$$\text{n observations}\left\{\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix}\right. = \text{n observations}\ \overbrace{\begin{bmatrix} \text{———} \; x_1 \; \text{———} \\ \text{———} \; x_2 \; \text{———} \\ \vdots \\ \text{———} \; x_n \; \text{———} \end{bmatrix}}^{\text{p (voxels)}} \left.\begin{matrix} | \\ \vec{w} \\ | \end{matrix}\right\}\text{p dimensions}$$

Problem of interest in fMRI analysis

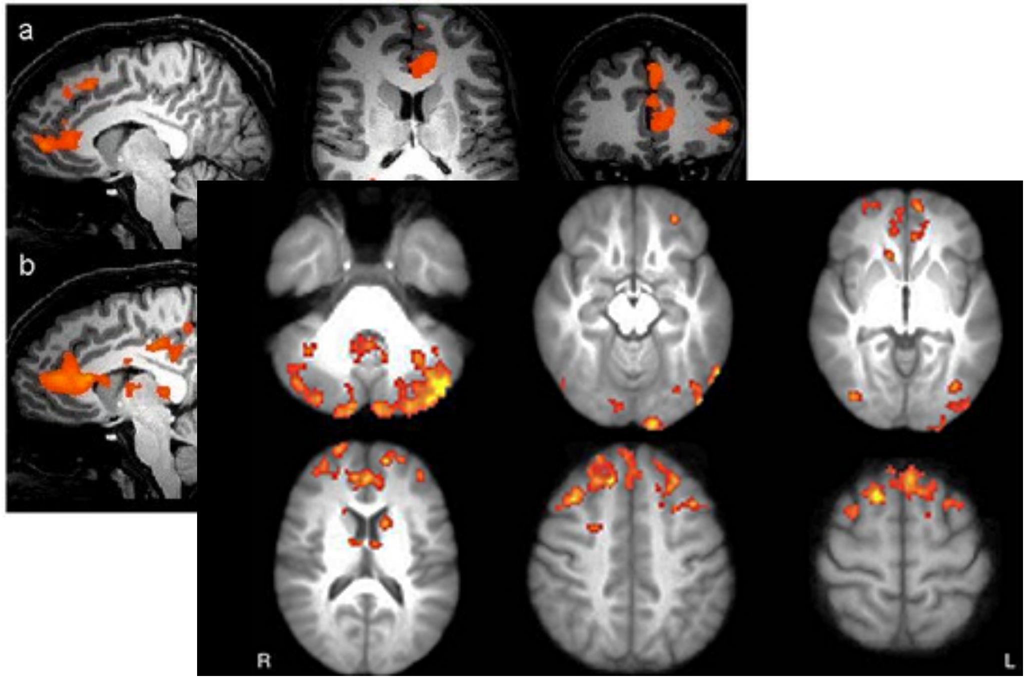
Regression (reverse inference / brain decoding)

response
(e.g. reaction time,
object label)

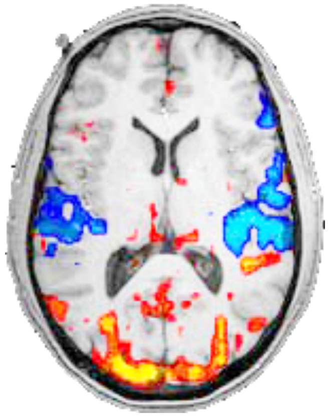
brain activity

decoding
weights

$$Y = X \cdot \vec{w} + \textit{noise}$$



visualize $\vec{w}$ as
the brain map



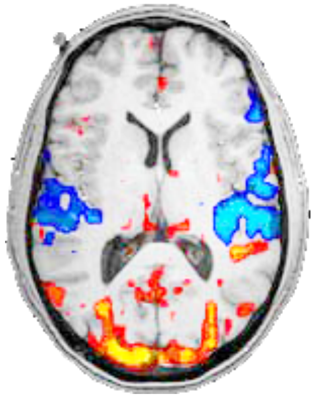
Our idea: a structured sparse and smooth prior

1. **structured sparsity**: non-zero decoding weights tend to be nearby each other

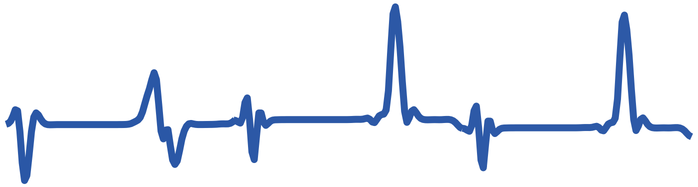
2. **smooth:** weights within a neighborhood are correlated with each other

Goal: construct a Bayesian prior that can generate $\vec{w}$ samples like this.

brain map of $\vec{w}$



vector of $\vec{w}$

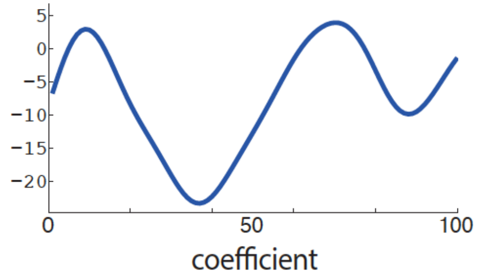


Dependent relevance determination

1. DRD: prior for structured sparsity

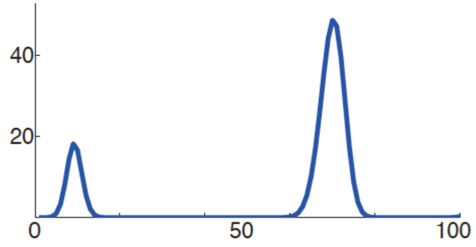
Gaussian process prior

$$\boxed{\mathbf{u}} \sim \mathrm{GP}(b\mathbf{1}, K)$$



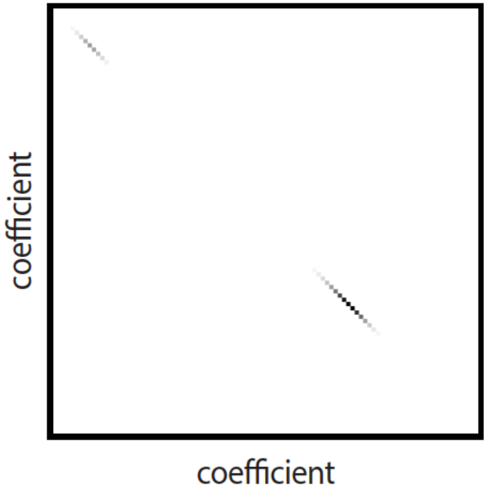
latent u from Gaussian Process

$$g = \exp(\mathbf{u})$$

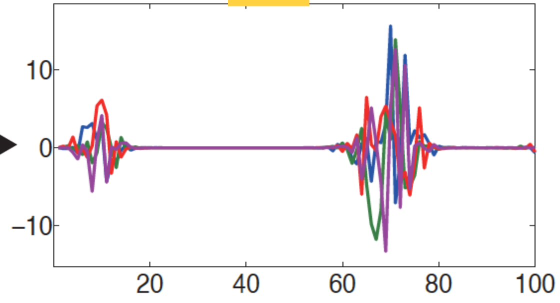


exponentiate

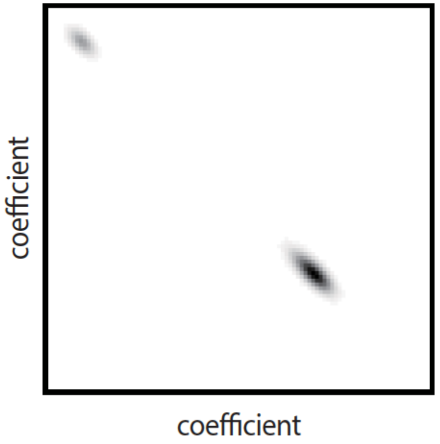
$$C \;\; = \mathrm{diag}(g)$$

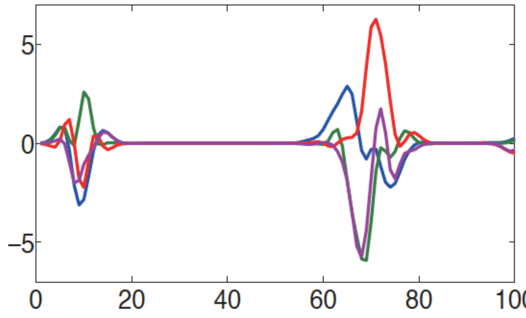


samples $\boxed{w} \sim \mathcal{N}(0, C\;\;)$



2. Smooth DRD: DRD with smoothness



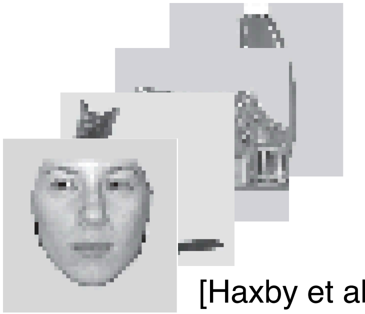


[Wu et al NeurIPS2014]
[Wu et al JMLR2019]
[Cai & **Wu** et al Neuropsychologia2020]

Application to fMRI data: classification (visual recognition task)

$\mathbf{Y} = \mathbf{X} \cdot \vec{w} + noise$

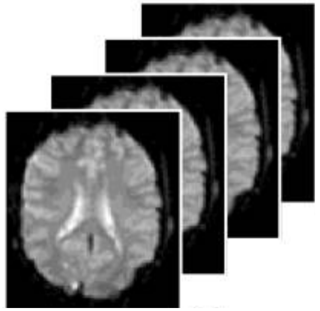
?



[Haxby et al 2001]

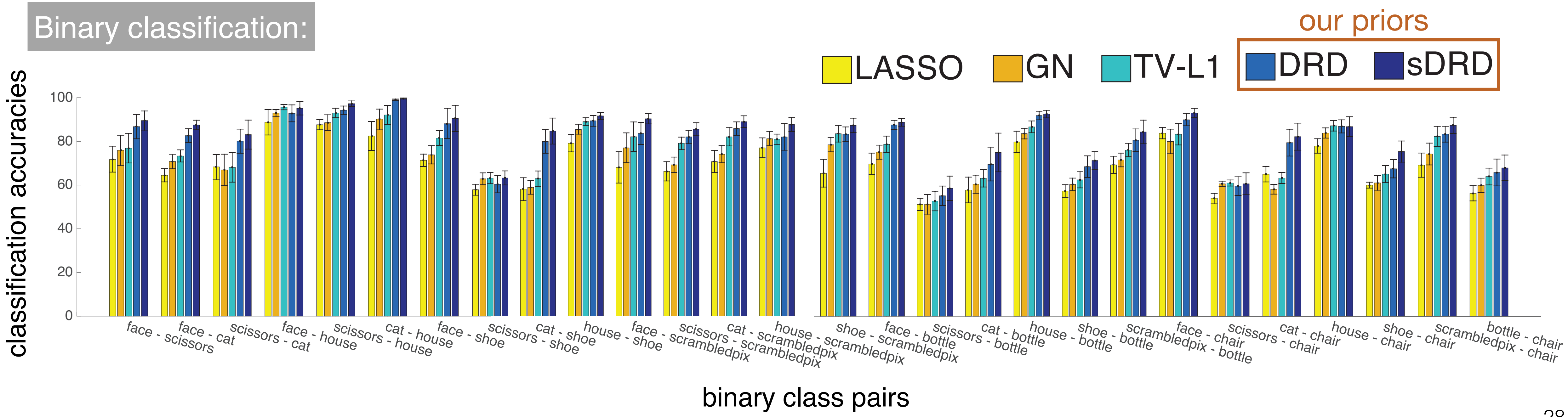
8 types of objects
(house, bottle, face, etc.)

decoding



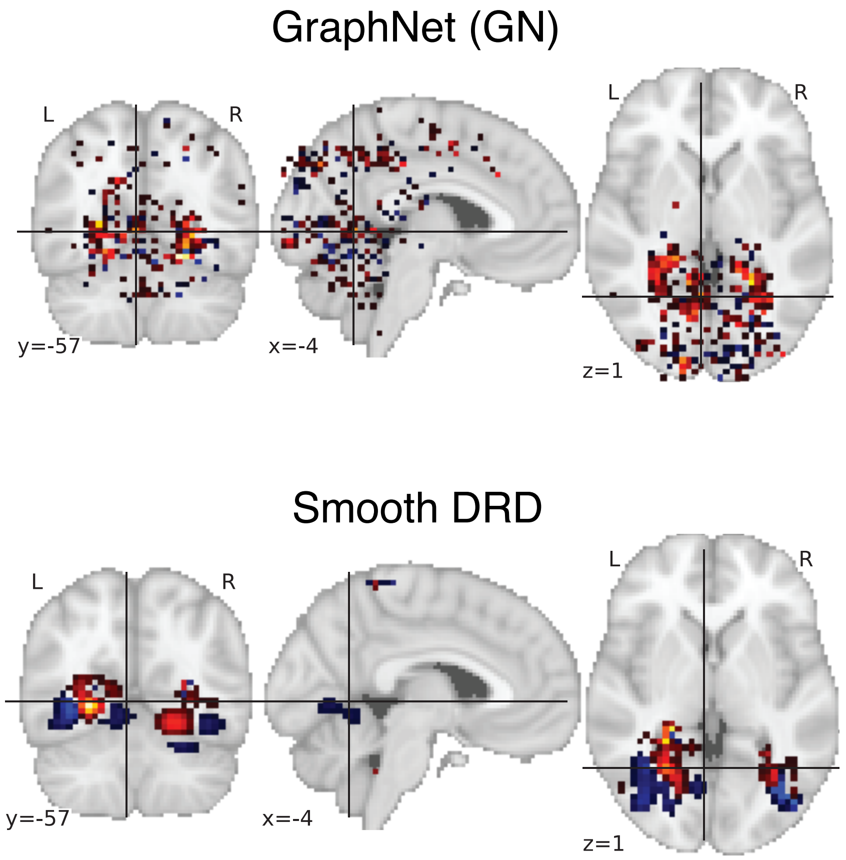
5k volumes
24k valid voxels in 40 × 64 × 64 3D volume

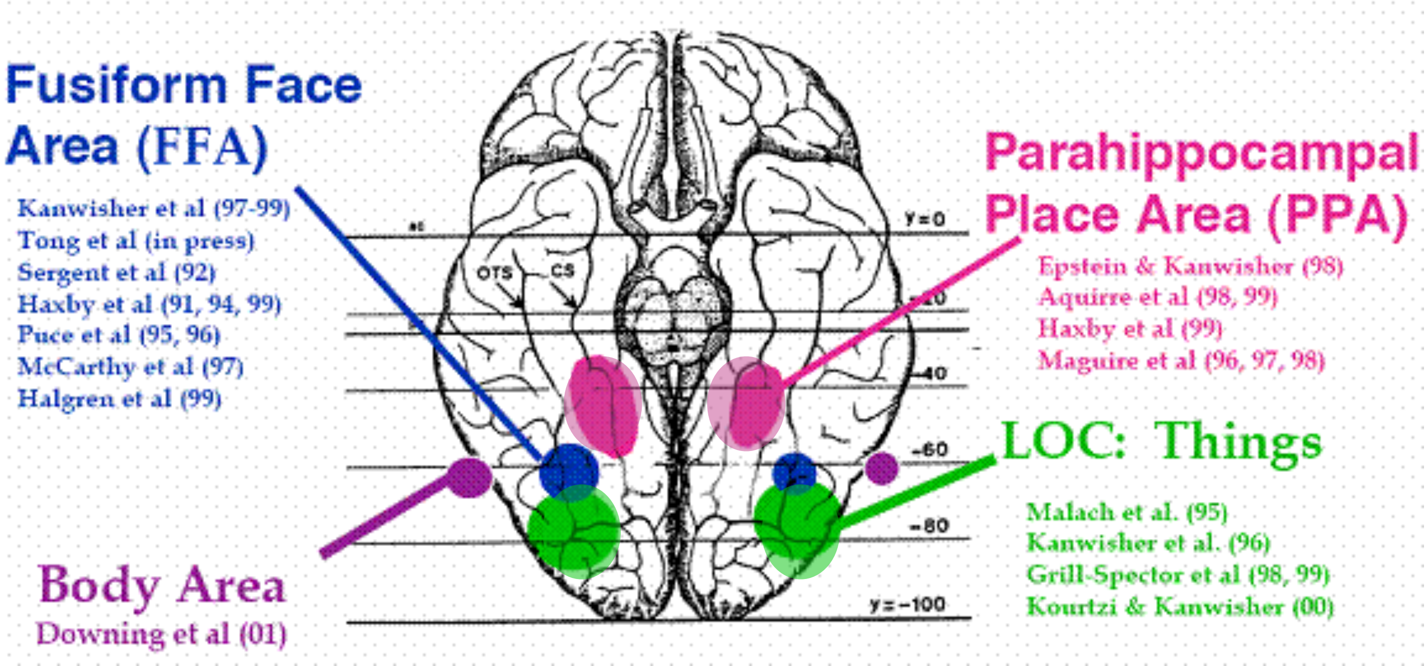
Binary classification:



Highlight: interpretable brain map of decoding weights

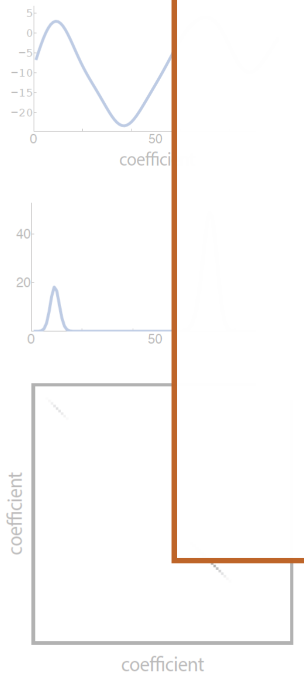
$\vec{w}$ for house vs bottle

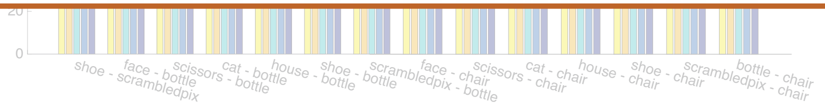




Contributions

- A hierarchical model for structured, sparse and smo… with Gaussian pr…
- Improving brain decoding tasks
- Identifying more interpretable

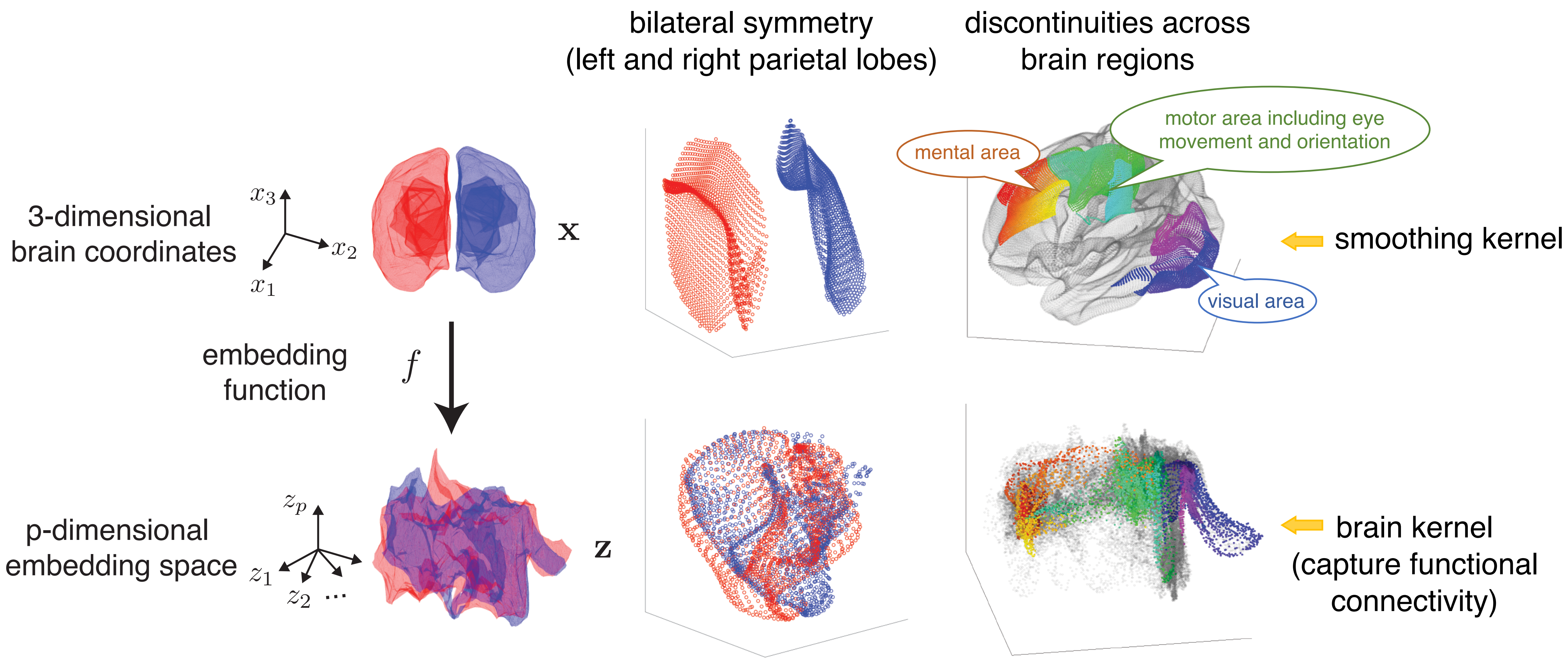




Highlight

Interpretable brain map of decoding weights for fMRI analysis

Brain kernel for non-stationary spatial correlation



[Wu et al NeuroImage2021]

Outline

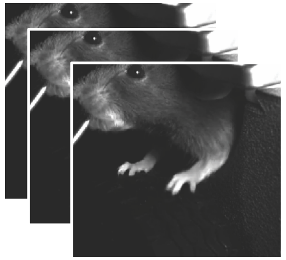
Latent structure discovery for neural recordings

Structured priors for fMRI brain decoding

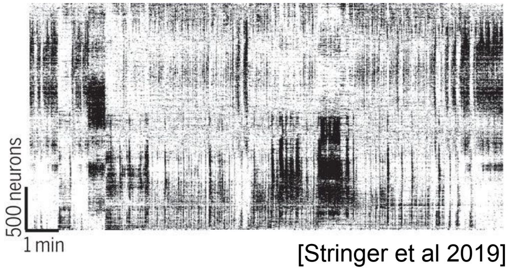
Semi-supervised learning for animal behavior analysis and understanding

Problem: what do we want to get out of behavioral videos?

behavioral video

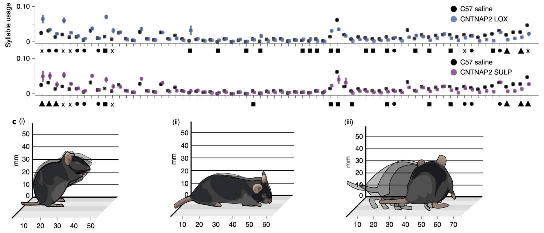


neural activity



[Stringer et al 2019]

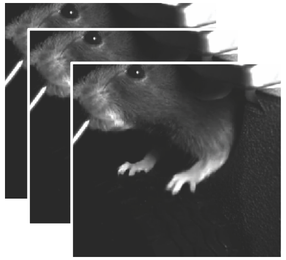
behavior analysis



[Wiltschko et al 2020]

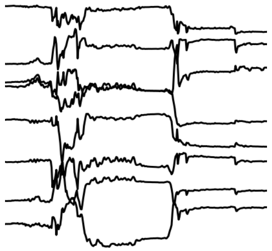
Problem: what do we want to get out of behavioral videos?

behavioral video



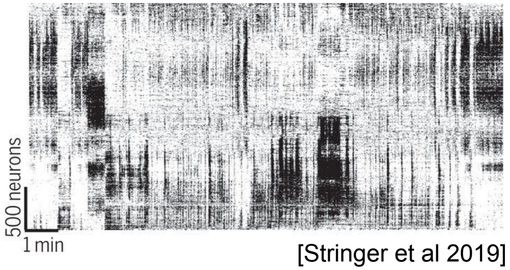
tracking
compression
segmentation

interpretable low-dimensional representation



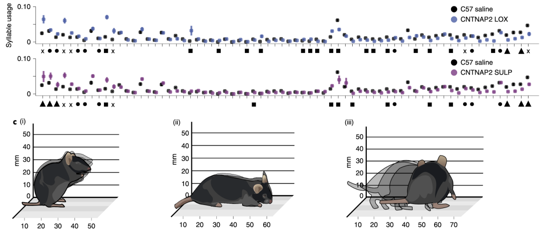
encoding
decoding

neural activity



[Stringer et al 2019]

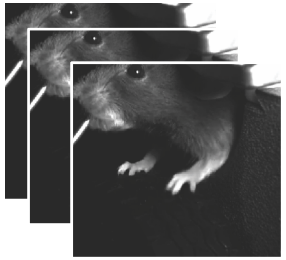
behavior analysis



[Wiltschko et al 2020]

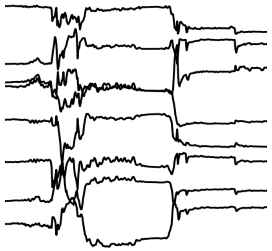
Problem: what do we want to get out of behavioral videos?

behavioral video



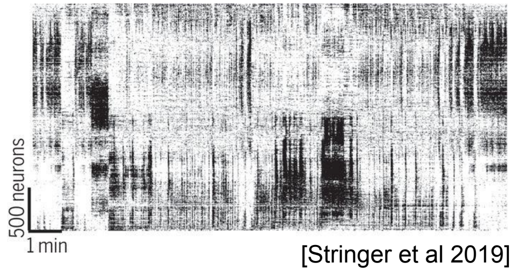
tracking
compression
segmentation

interpretable
low-dimensional representation



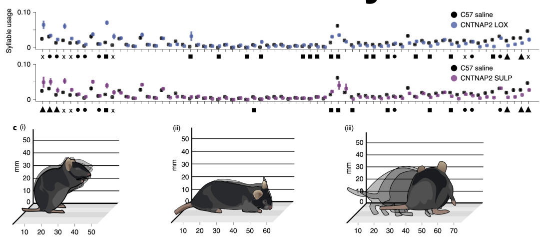
encoding
decoding

neural activity



[Stringer et al 2019]

behavior analysis



[Wiltschko et al 2020]

Scarce data issue: limited human-labeled data in behavioral analysis

Our solution: semi-supervised learning with both limited labeled data and vast amount of unlabeled data

labeled data

unlabeled data

semi-supervised learning

tracking
compression
segmentation

Tracking: animal pose estimation

fish [Norovich et al 2019]

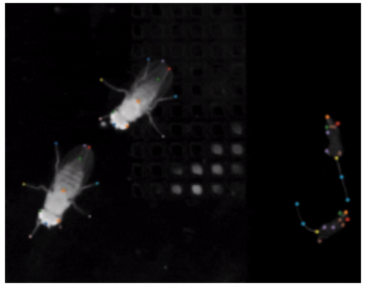
mouse [Meijer et al 2019]

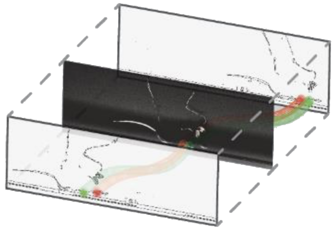
state-of-the-art animal pose estimation

DeepLabCut: a software package for animal pose estimation

DeepPoseKit

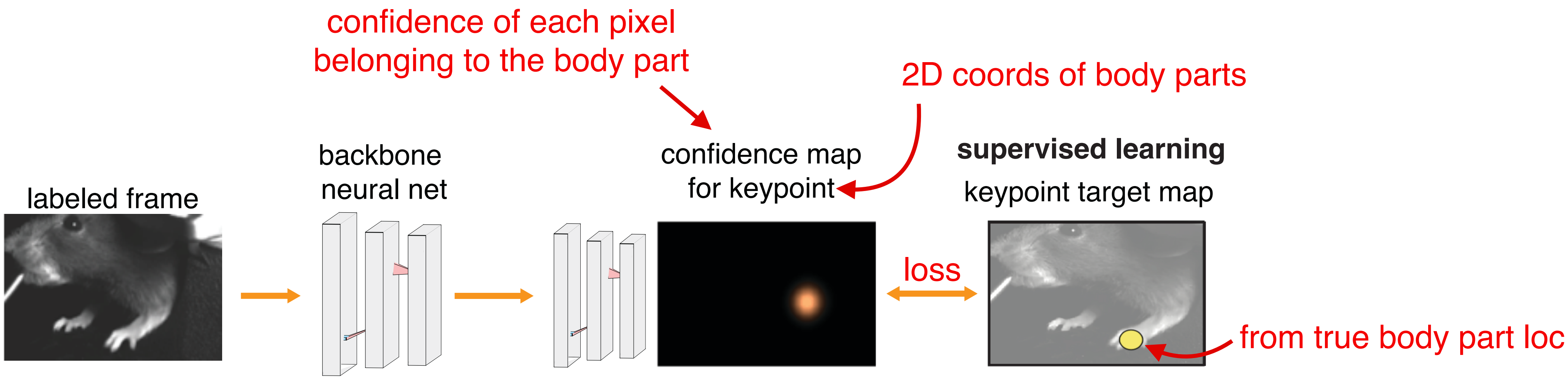
Social LEAP Estimates





OptiFlex: Video-Based animal pose estimation by optical flow.

Base model for animal pose estimation



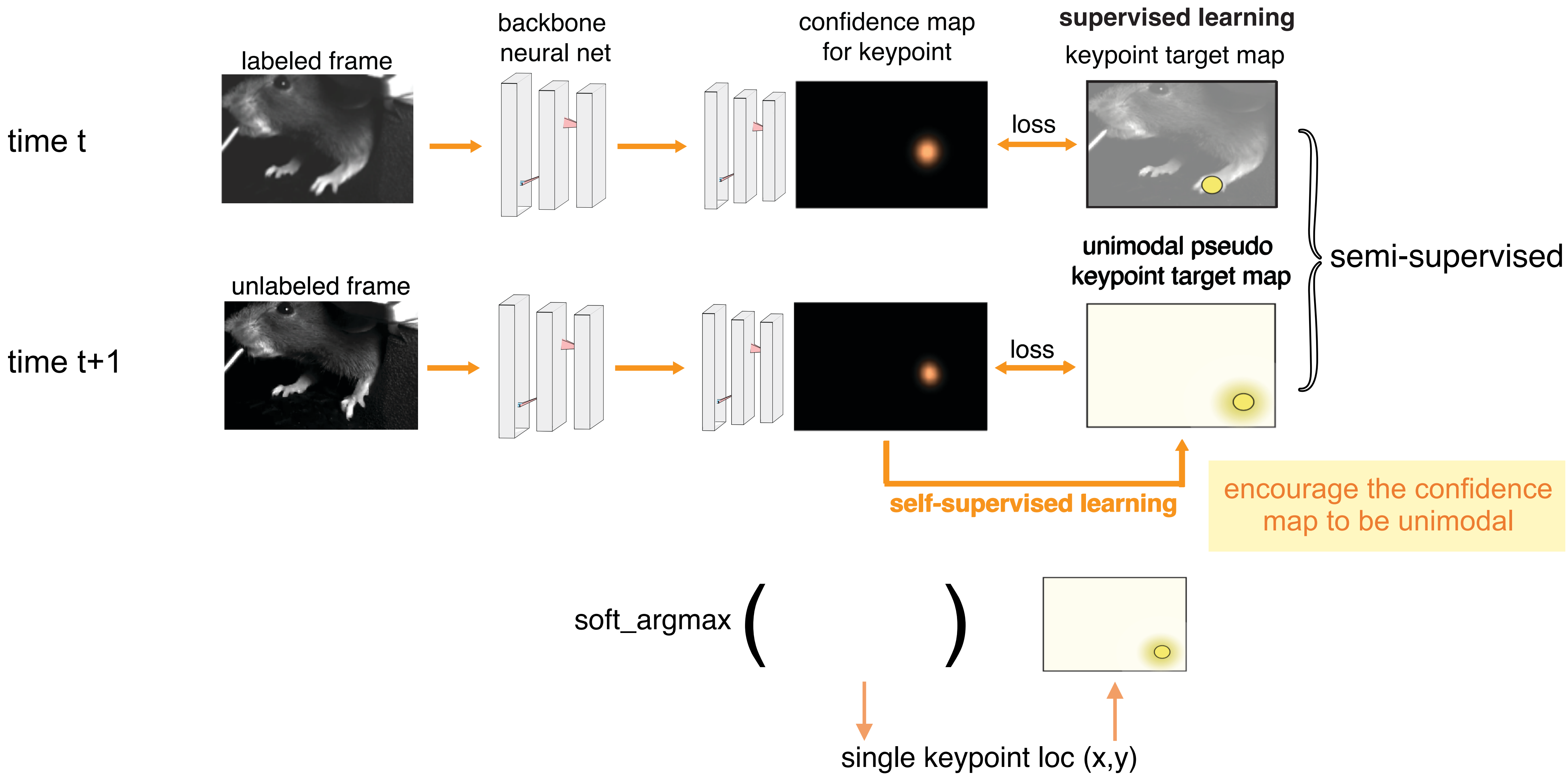
Limitations:

1. train only on labeled frames, ignoring information in unlabeled frames
2. occasional "glitches" (tracking is lost)

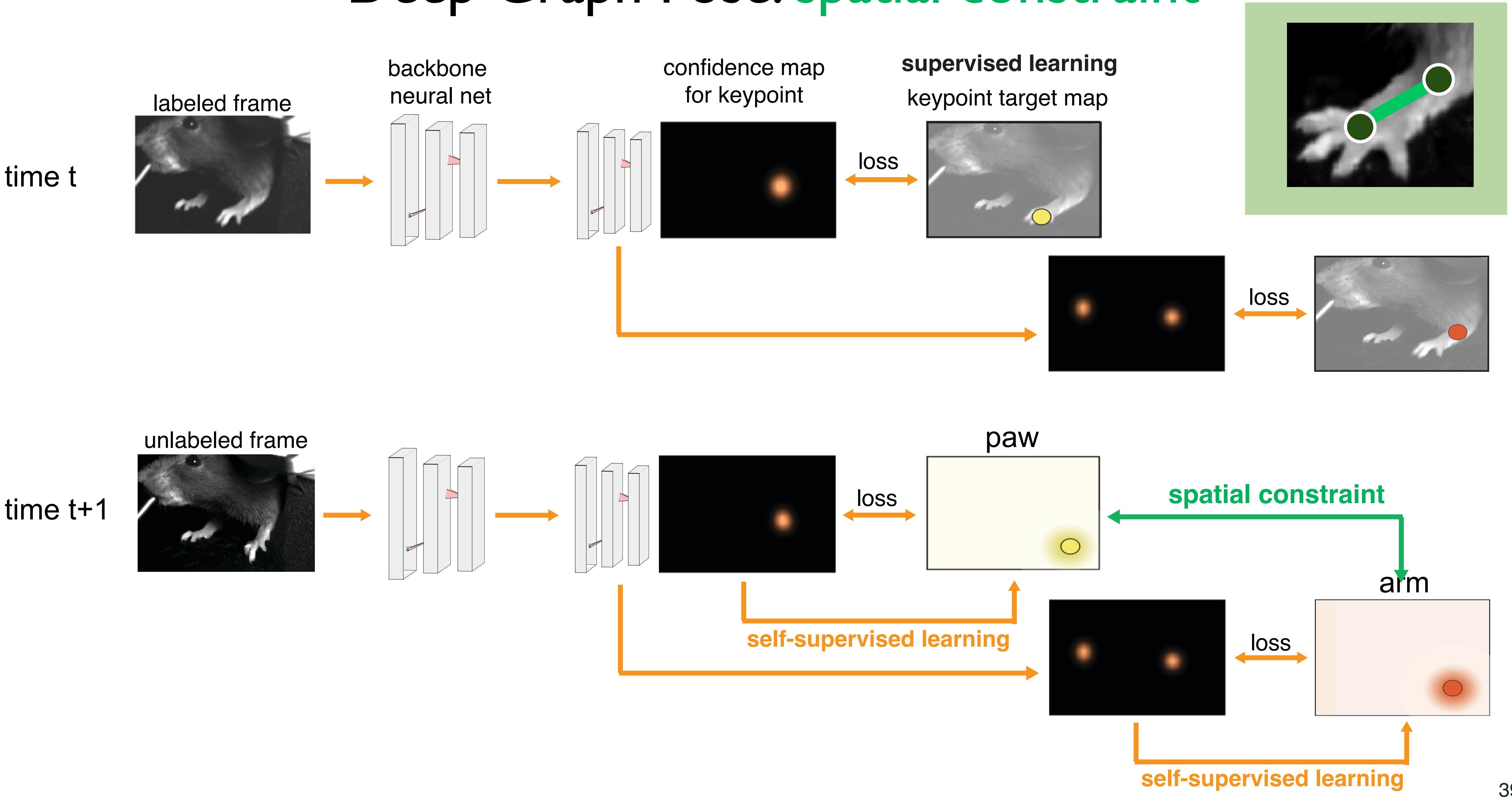
Our solutions:

1. unlabeled frames should help
2. temporal and spatial constraints

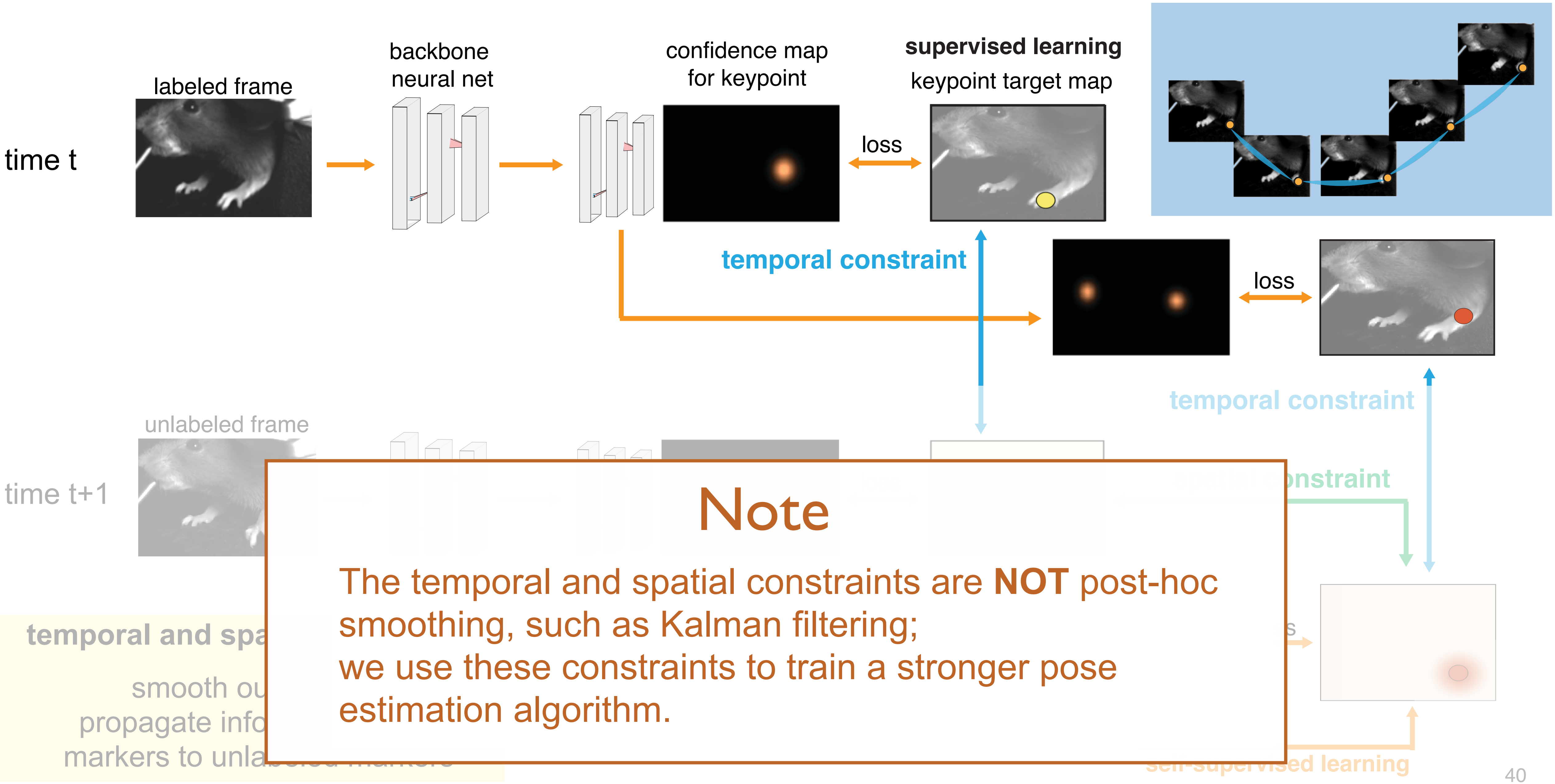
Deep Graph Pose: semi-supervised



Deep Graph Pose: spatial constraint



Deep Graph Pose: temporal constraint



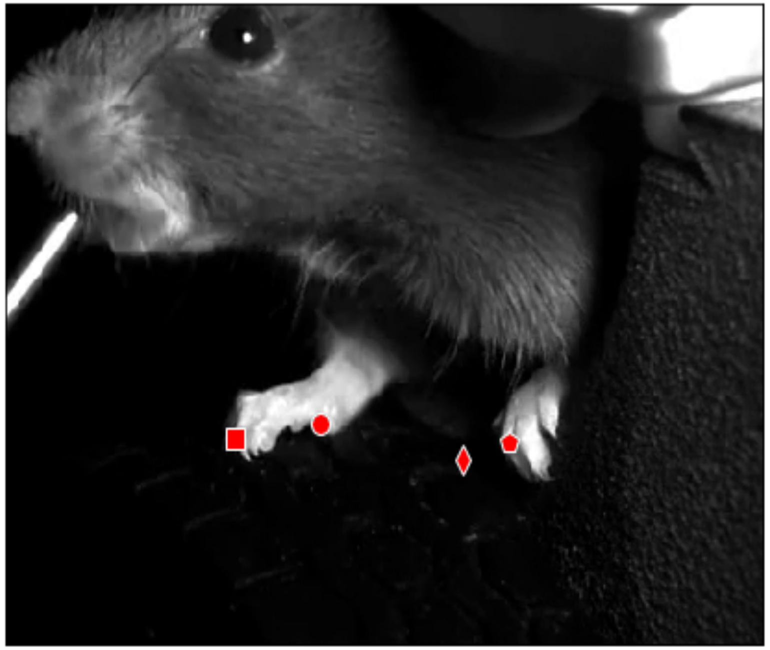
Note

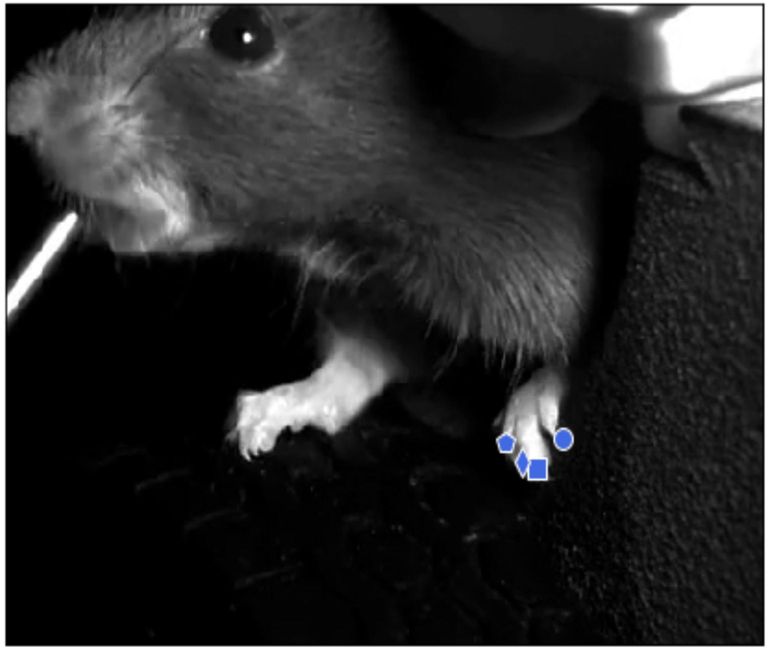
The temporal and spatial constraints are **NOT** post-hoc smoothing, such as Kalman filtering;
we use these constraints to train a stronger pose estimation algorithm.

DGP performance on a mouse dataset

DeepLabCut
(DLC)

DeepGraphPose
(DGP)





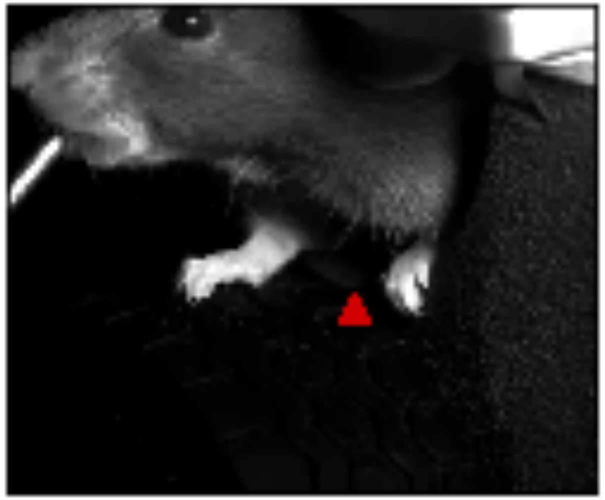
[Mathis et al 2018]

Dataset	Number of labeled frames / total number of frames
Meijer et al 2019	55/1000

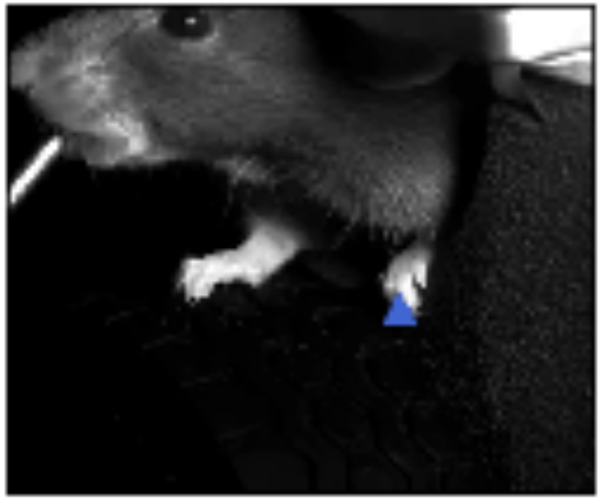
[**Wu**, Buchanan, Whiteway et al NeurIPS2020]

DGP encourages unimodal confidence maps

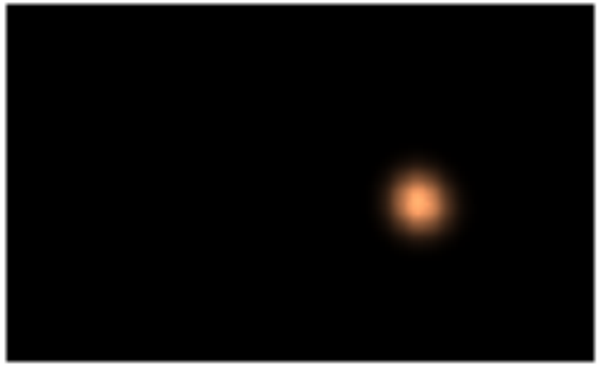
DeepLabCut
(DLC)

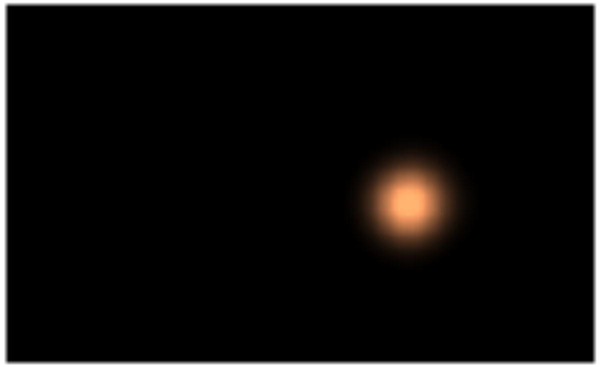


DeepGraphPose
(DGP)



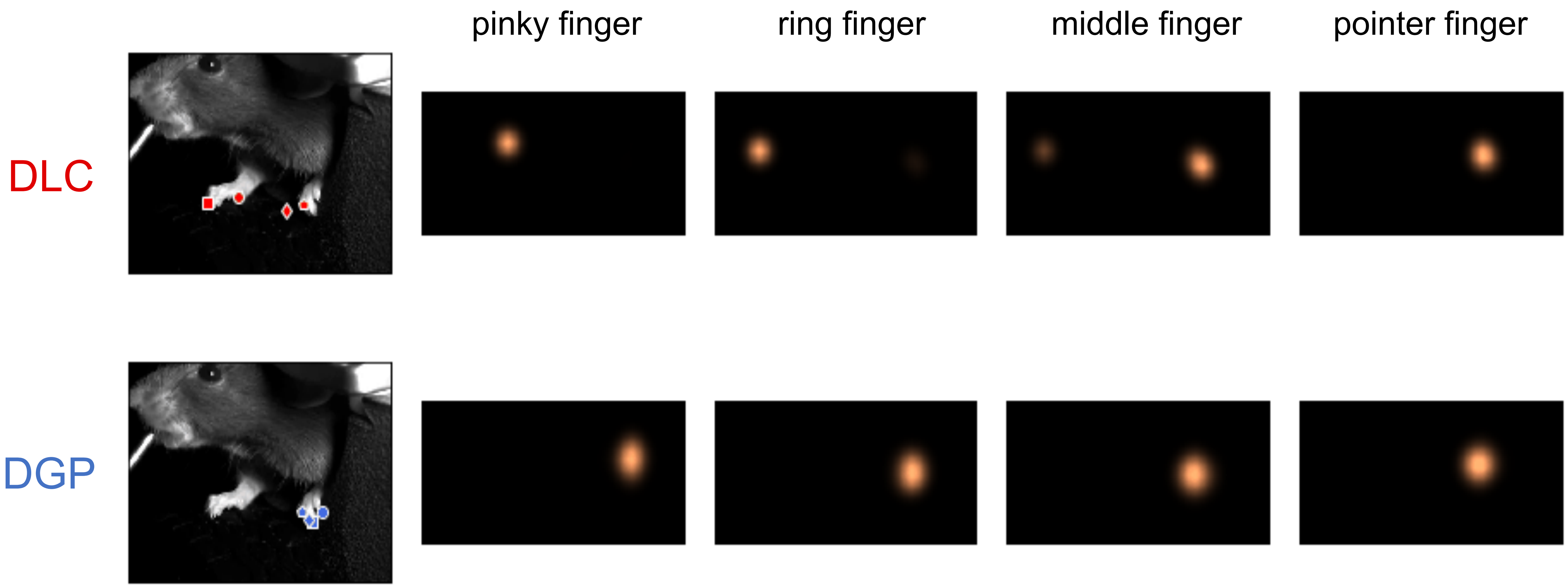
confidence map
for
middle finger





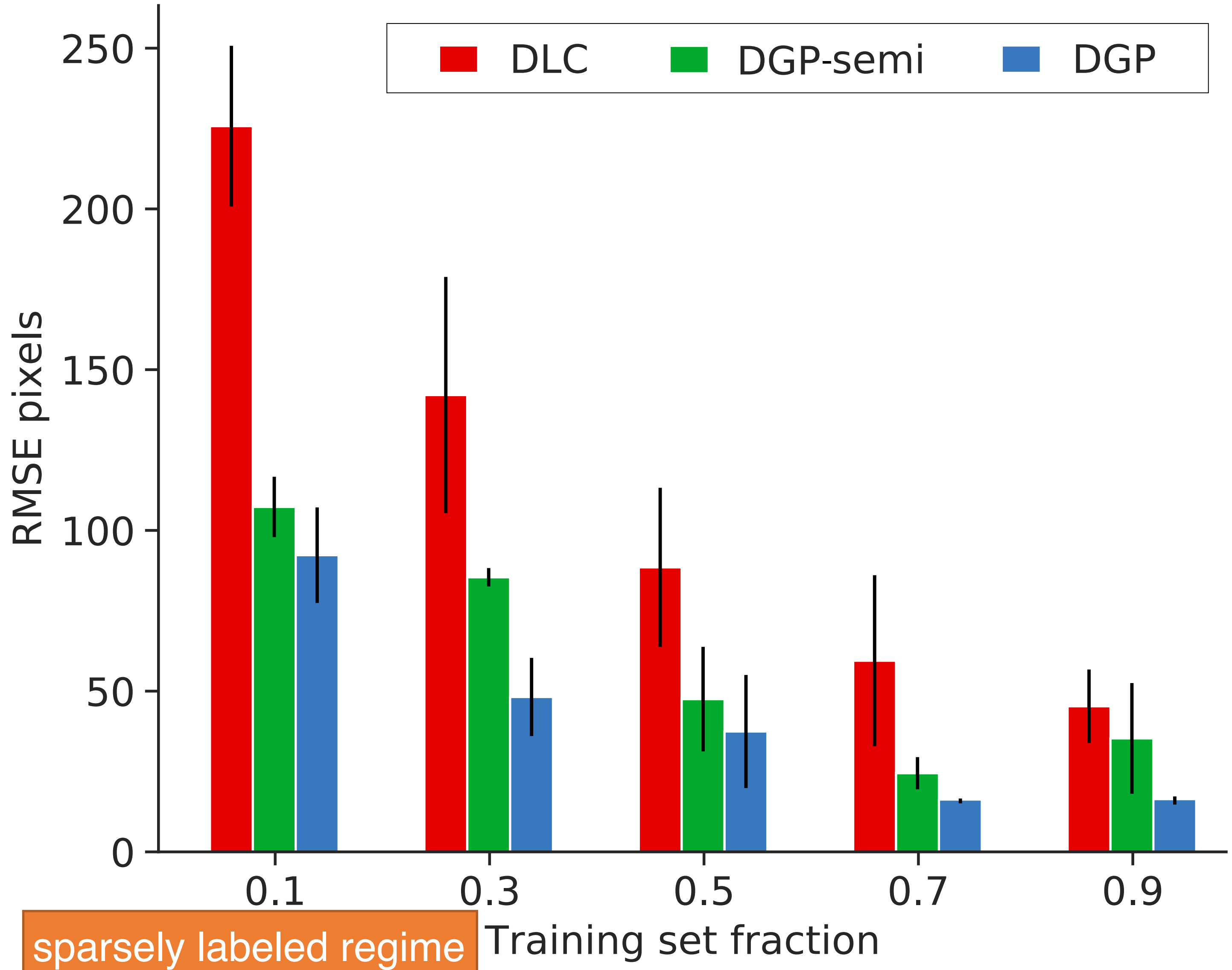
[**Wu**, Buchanan, Whiteway et al NeurIPS2020]

DGP encourages unimodal confidence maps



[**Wu**, Buchanan, Whiteway et al NeurIPS2020]

DGP outperforms DLC with fewer labeled frames



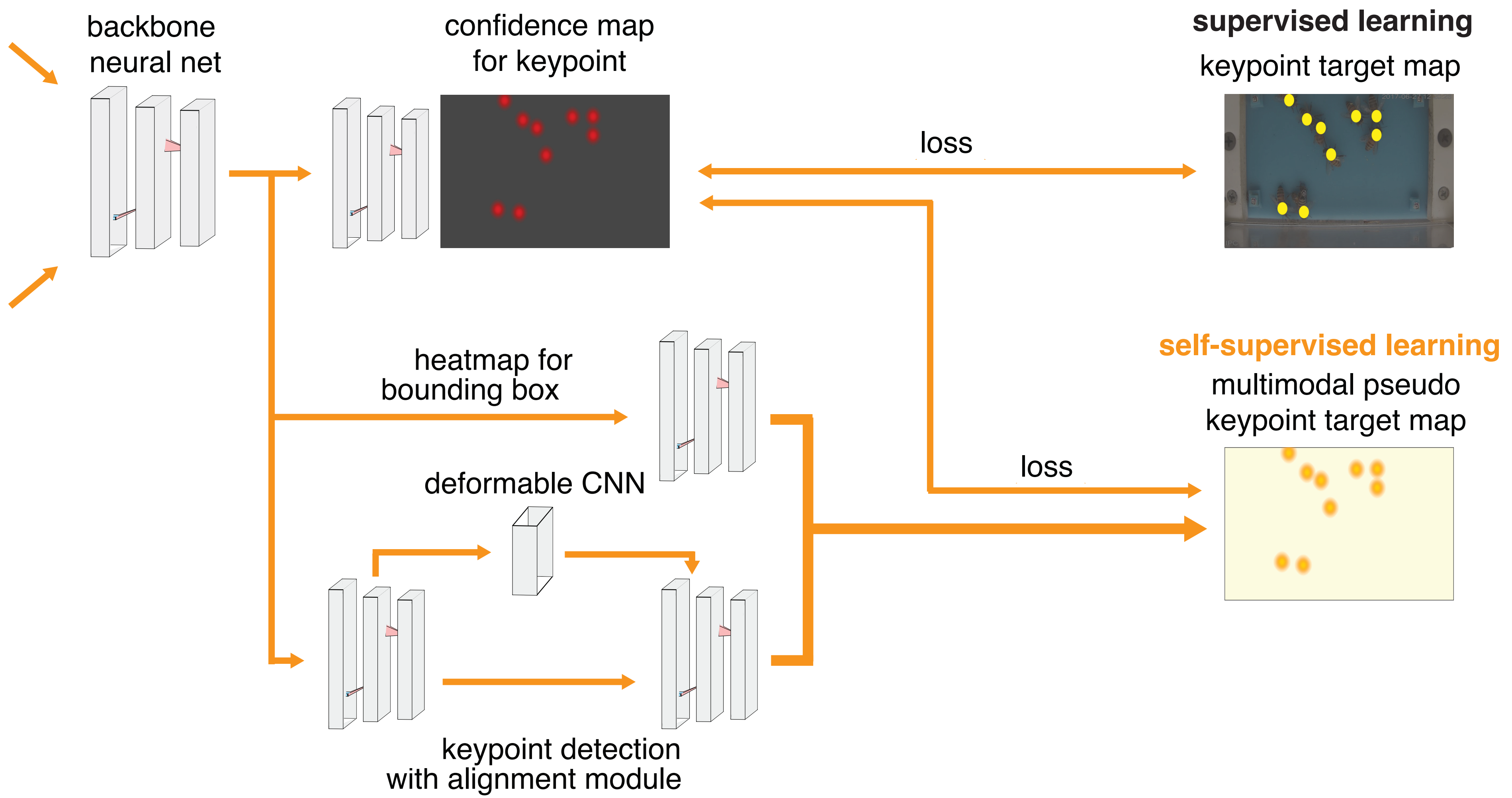
sparsely labeled regime

DGP-semi
- semi-supervised learning

DGP
- semi-supervised learning
- temporal constraints
- spatial constraints

[**Wu**, Buchanan, Whiteway et al NeurIPS2020]

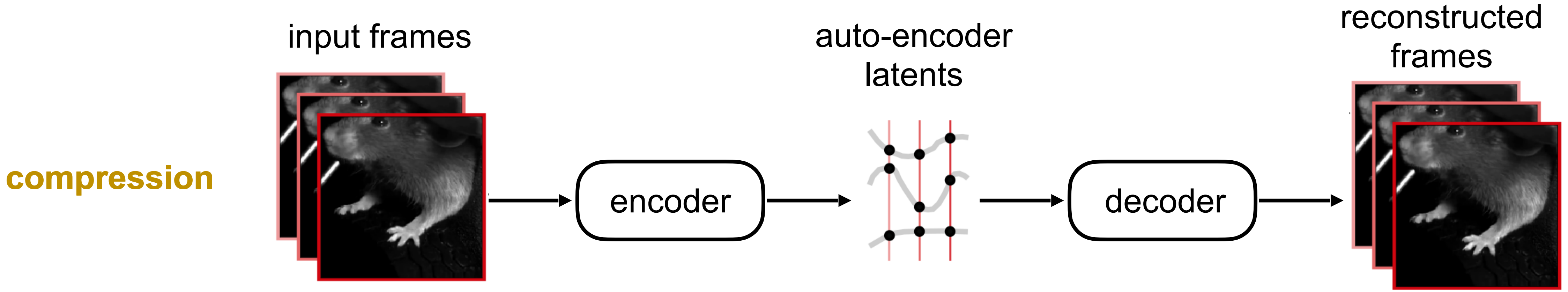
SemiMultiPose: Multi-animal Pose Estimation

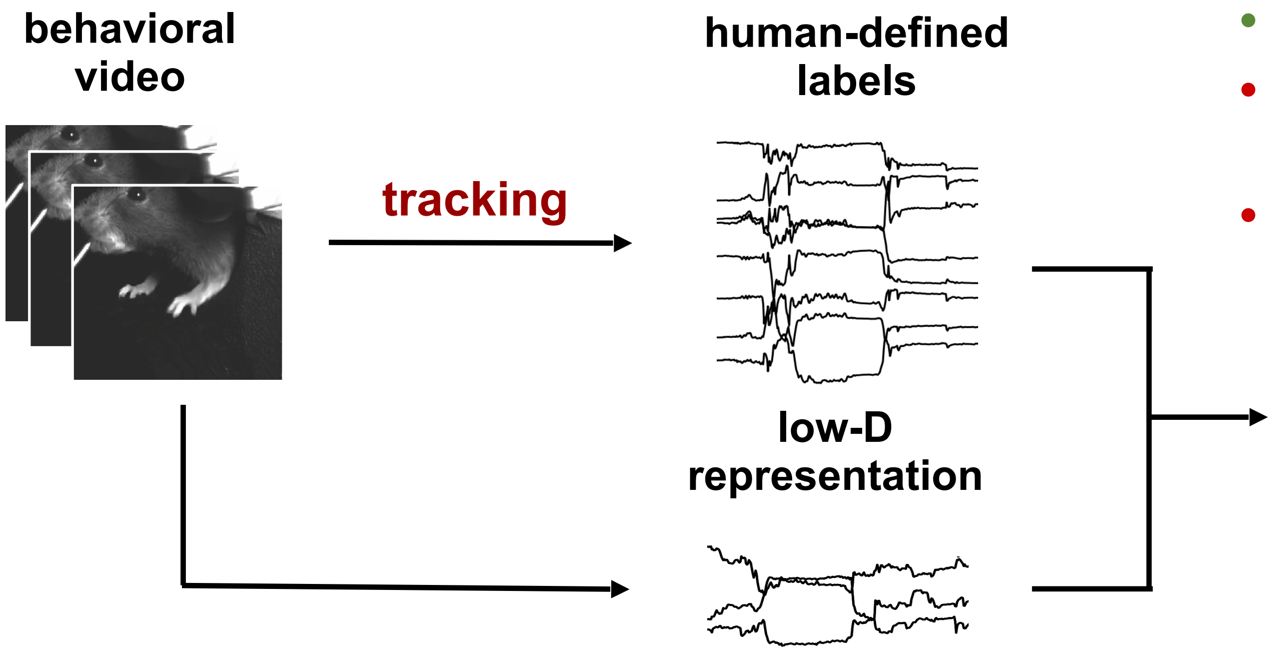


Ari Blau (Ph.D. @ Columbia)

[Blau & **Wu** et al under review]

Merging tracking and compression for improved behavioral representations



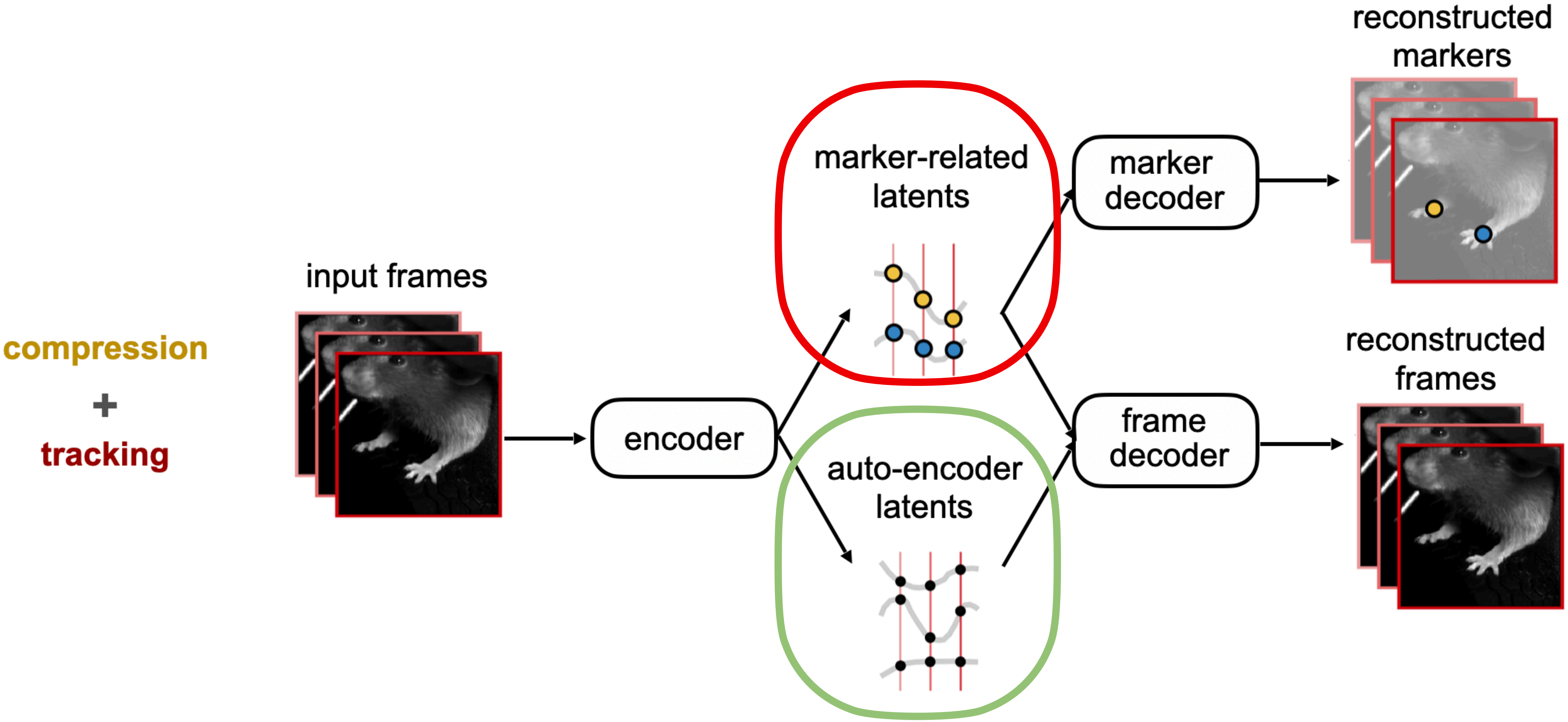


- **Pro**: no human labels required
- **Con**: low-d representation is difficult to interpret
- **Con**: does not allow experimenters to mark features of interest

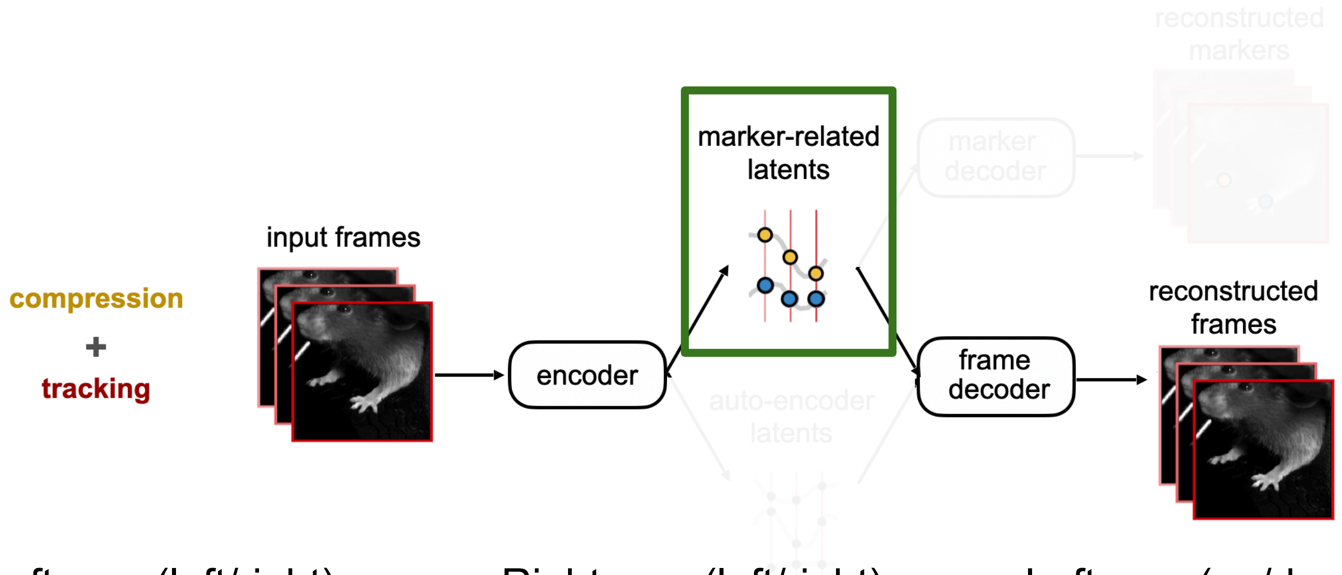
interpretable low-D representations

[**Wu**, Buchanan, Whiteway et al NeurIPS2020]
[Whiteway & **Wu** et al PLOSCompBio2021]

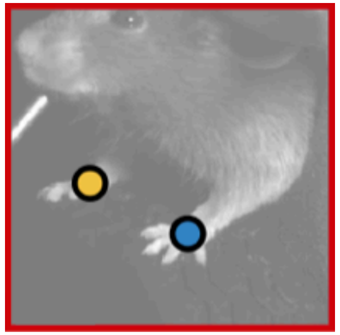
Semi-supervised compression: Partitioned Subspace VAE



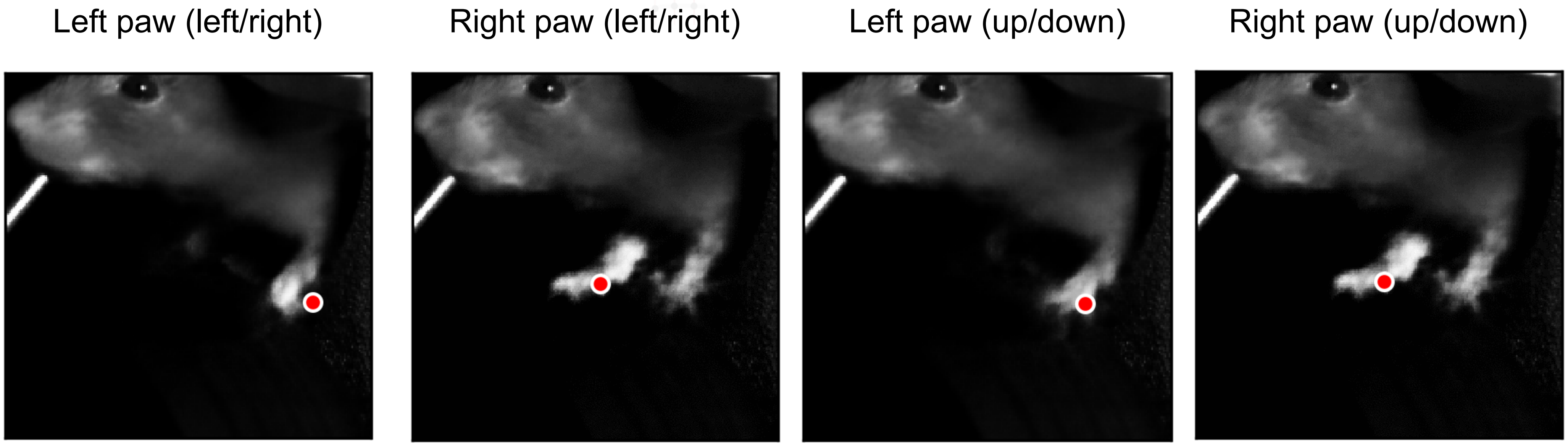
Highlight: disentangled marker-related latents



left and right paws



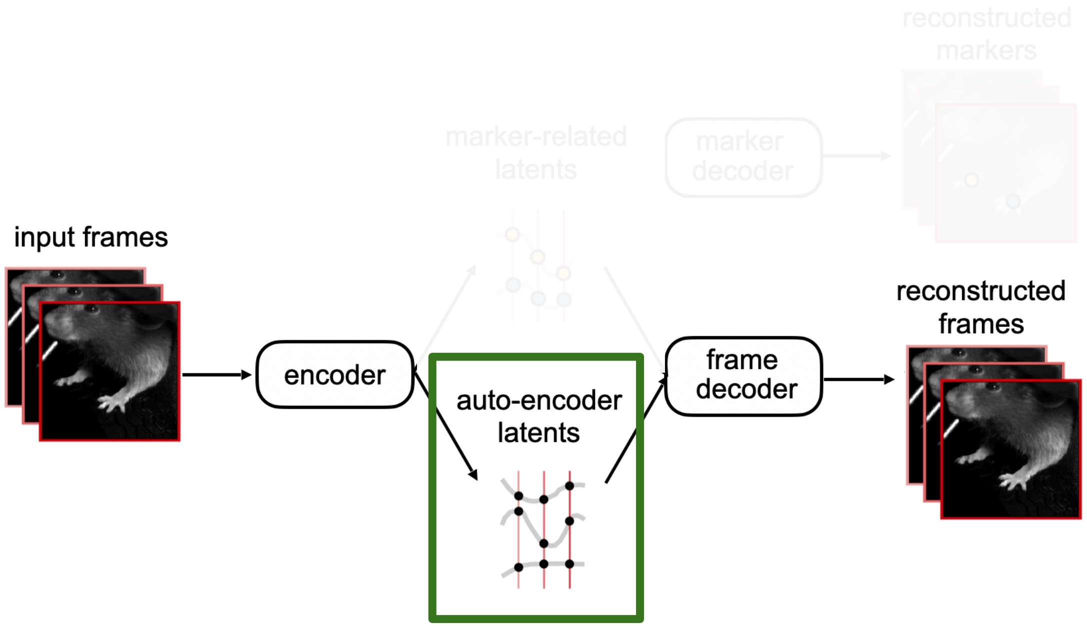
disentangled latents



[**Wu**, Buchanan, Whiteway et al NeurIPS2020]
[Whiteway & **Wu** et al PLOSCompBio2021]

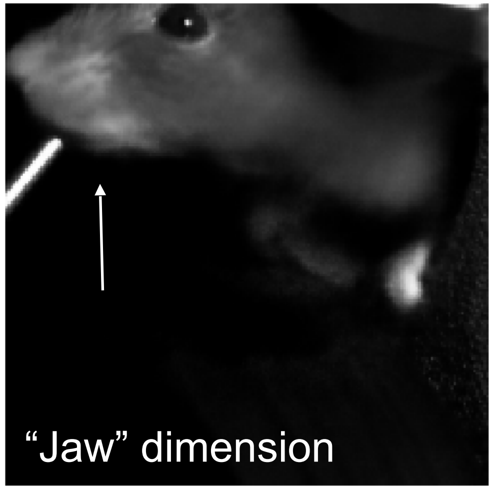
Highlight: disentangled auto-encoder latents

compression
+
tracking

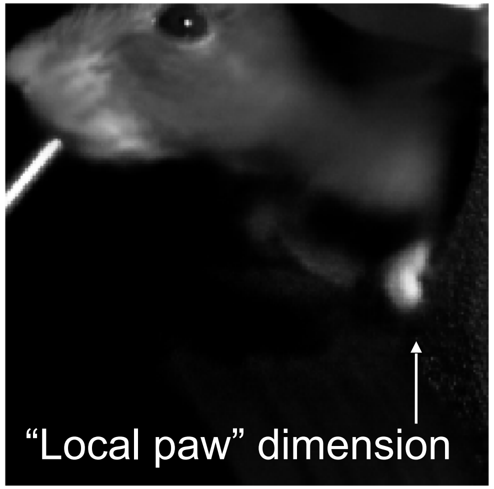


disentangled latents

Latent 1



Latent 2



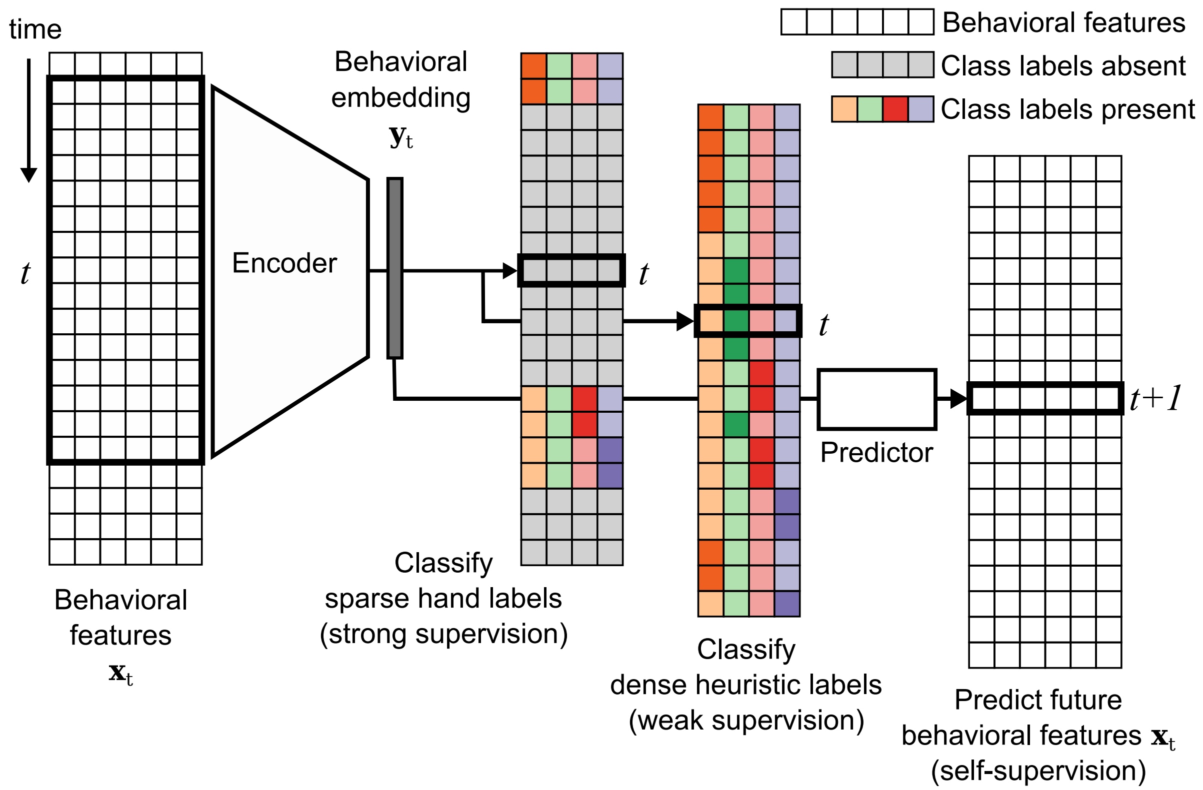
What can we do with these latents?

- Segment a sub-region of video (e.g. paw, or face)
- Decode individual dimensions of behavior
- More interpretable predictors for neural decoding

……

[**Wu**, Buchanan, Whiteway et al NeurIPS2020]
[Whiteway & **Wu** et al PLOSCompBio2021]

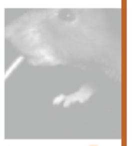
Semi-supervised behavioral segmentation

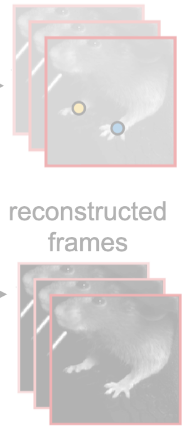


[Whiteway, Schaffer, **Wu** et al CV4Animals@CVPR2021]

Contributions

- New probabilistic graphical model for **tracking**: Deep Graph Pose (DGP)
- New **compression** method with DGP tracking





Highlight

Interpretable low-dimensional representation of behaviors

Outline

Latent structure discovery for neural recordings

Structured priors for fMRI brain decoding

Semi-supervised learning for animal behavior analysis and understanding

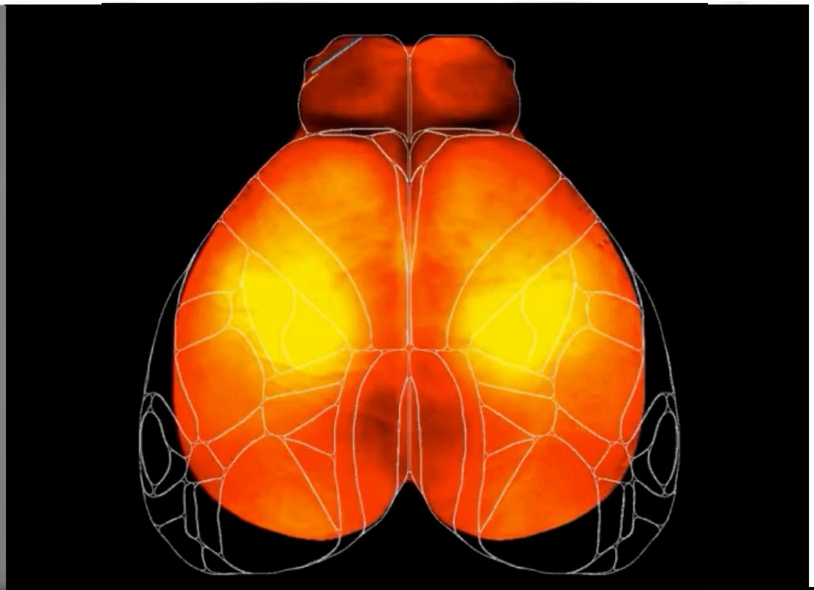
More complex neural and behavior data now and in the future

behavior

neural

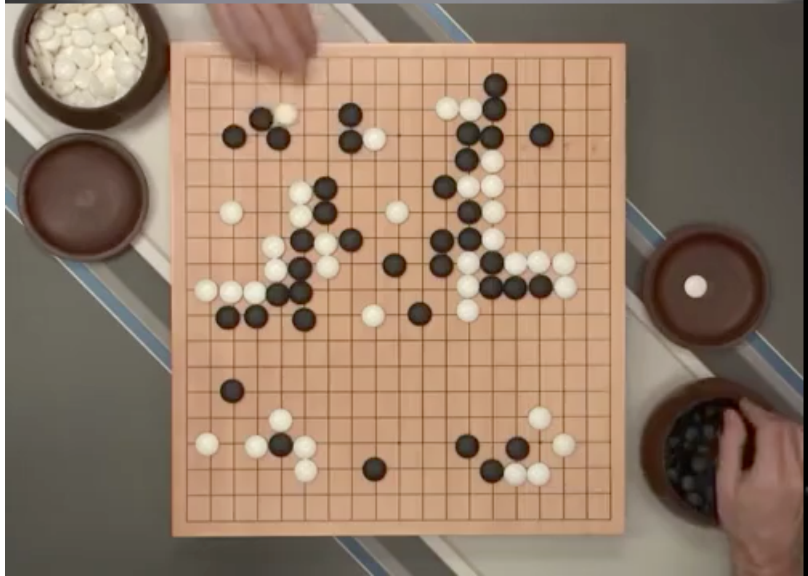
animal

[Choe el al 2017]



[Musall el al 2019]

human

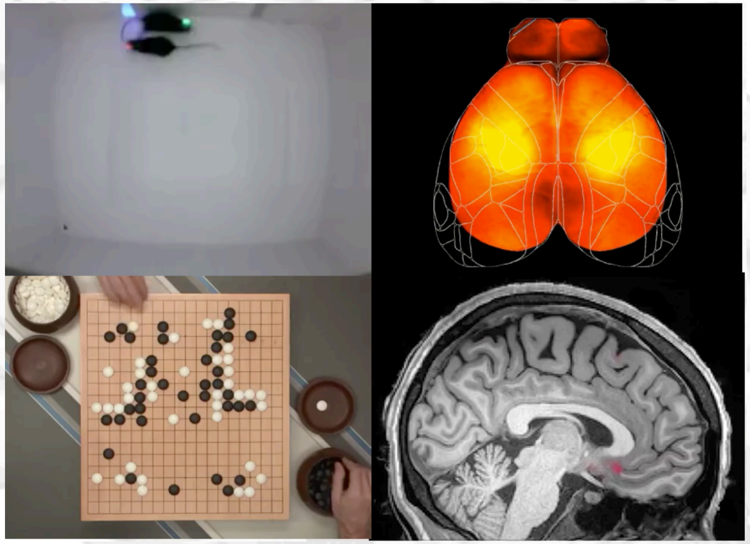


[Deepmind]

[Lewis et al 2019]

Neuroscience and AI

Brain



Artificial model

advanced quantitative models

neuro-inspired AI

deep learning

computer vision

computer graphics

imitation learning

probabilistic modeling

continual learning

active learning

reinforcement learning

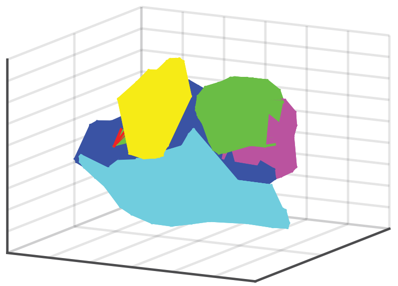
online learning

robotics

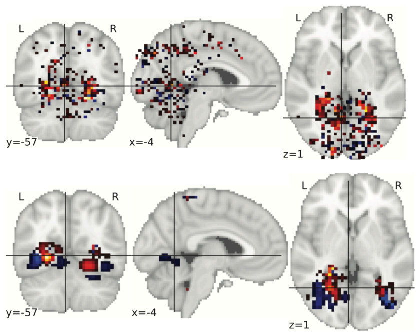
…

Plenty of interdisciplinary research opportunities between neuroscience and artificial intelligence!!

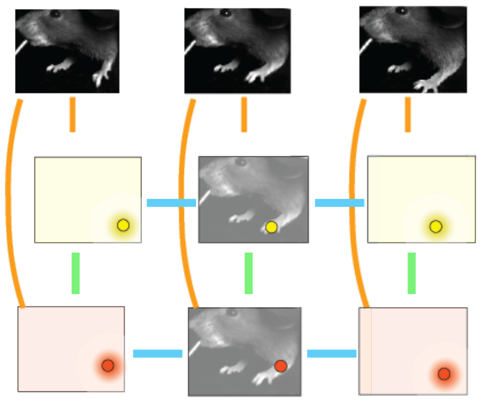
Acknowledgment



Nick Roy
Stephen Keeley
Stan Pashkovski
Bob Datta
Jonathan Pillow



Mijung Park
Oluwasanmi O. Koyejo
Samuel A. Nastase
Christopher A. Baldassano
Nicholas B. Turk-Browne
Kenneth A. Norman
Barbara E. Engelhardt
Jonathan W. Pillow



Estefany Kelly Buchanan
Matthew R. Whiteway
Michael Schartner
Guido Meijer
Jean-Paul Noel
Erica Rodriguez
Claire Everett
Amy Norovich
Evan Schaffer
Neeli Mishra
C. Daniel Salzman
Dora Angelaki
Andrés Bendesky
The International Brain Laboratory
John Patrick Cunningham
Liam Paninski