

# **Machine Learning Methods for Neural Data Analysis**

## **Demixing and Deconvolving Calcium Imaging Data**

Scott Linderman

*STATS 220/320 (NBIO220, CS339N). Winter 2023.*

# Announcements

- Lab 1 due tomorrow at midnight.
  - Please upload **just the pdf** on Gradescope.
  - Please add a short paragraph at the end about **author contributions**
  - Please see Ed discussion for minor errata on Problems 2e and 3c.
- Lab 2 will be released tomorrow evening so you can read through it before the lab on Friday.
- Team assignments will be announced tomorrow night as well.

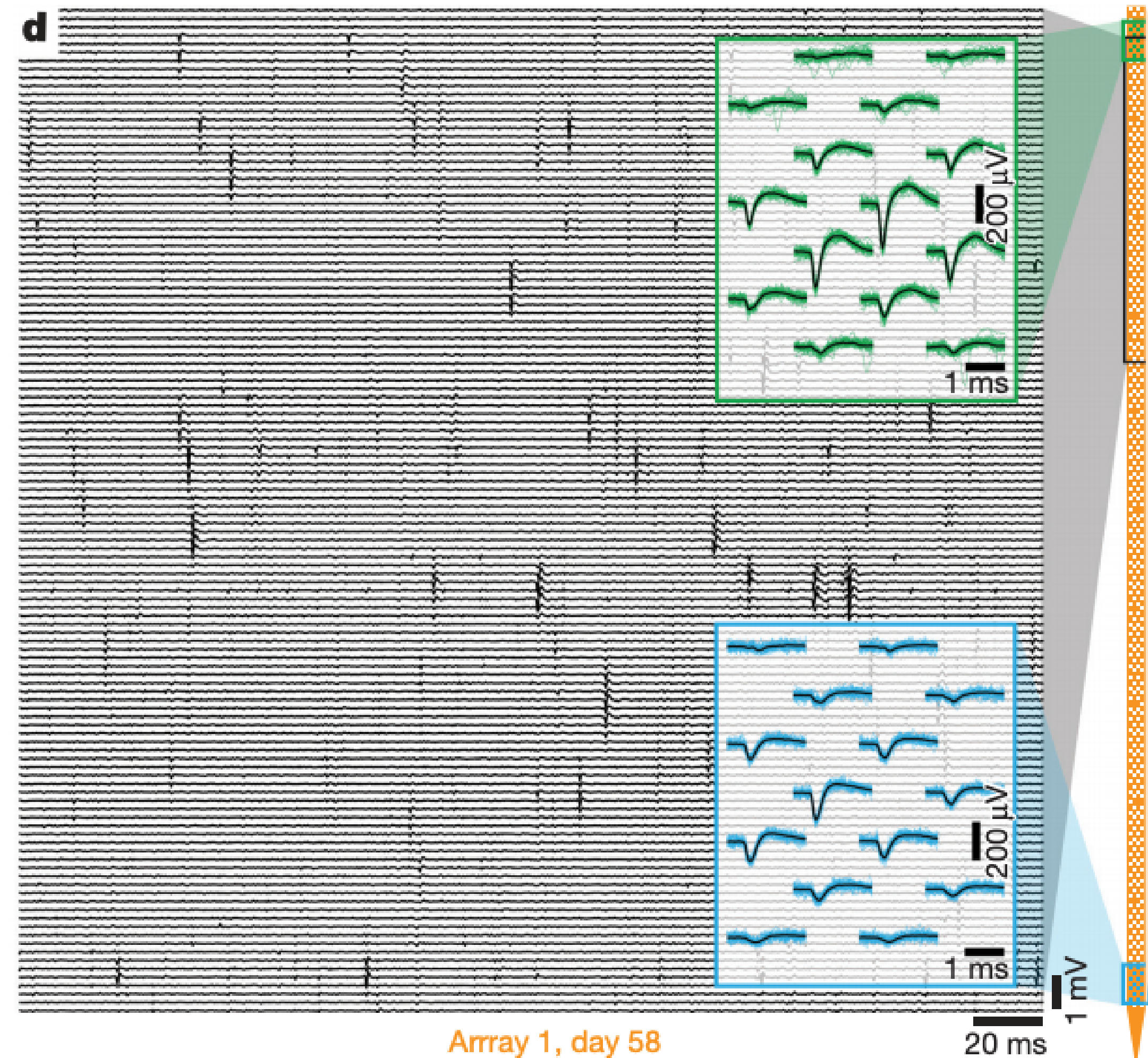
# Agenda

1. Optical physiology
2. Constrained Non-negative Matrix Factorization (CNMF)

# Recap

## Electrophysiology

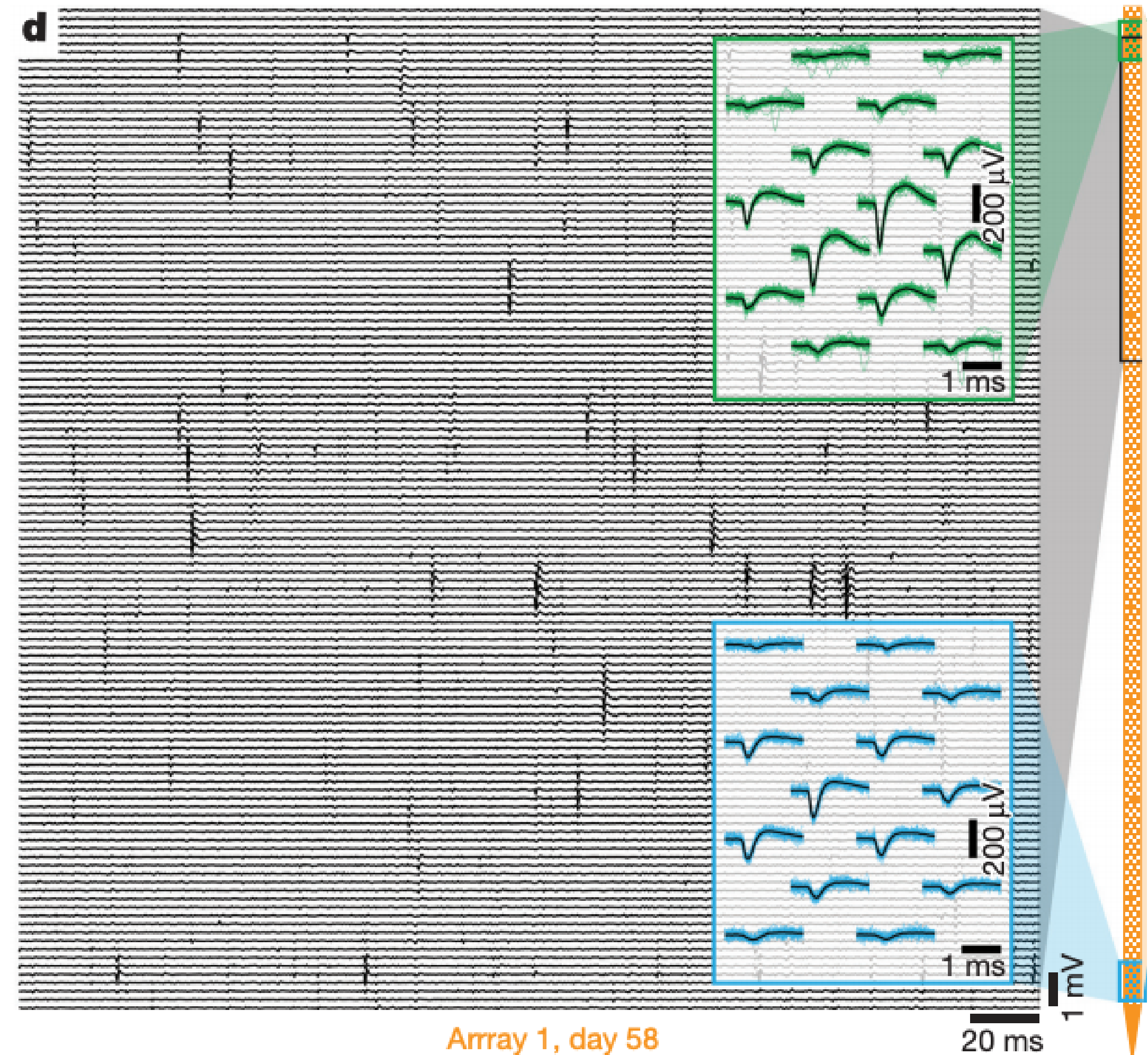
- So far, we've study electrophysiological ("ephys") recordings with high density probes.
- The raw data is a **multidimensional time series of voltage measurements**, one for each recording site on the probe.
- When neurons near the probe fire an **action potential**, it registers a **spike in the voltage** on nearby channels.
- Typical recordings detect spikes from **O(100) neurons**.



# Recap

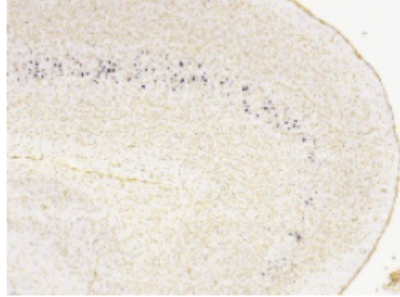
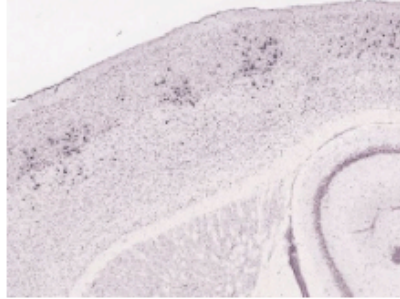
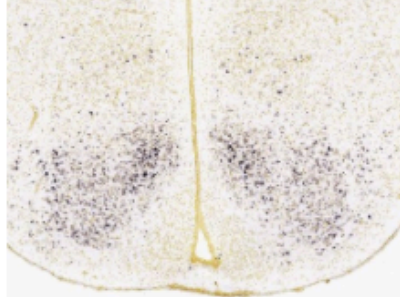
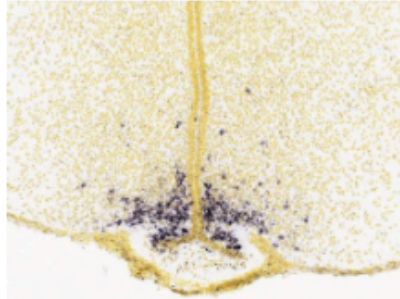
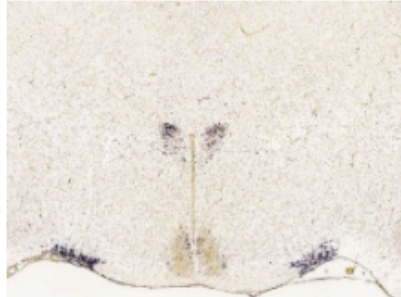
## Electrophysiology Limitations

- It's hard to detect neurons that fire rarely and produce low amplitude EAPs.
- More generally, you only pick up cells that happen to be close to the narrow probe.
- No cell-type specificity.
- In particular, ephys does not leverage our powerful genetic toolkits for certain model organisms.



# Genetic tools

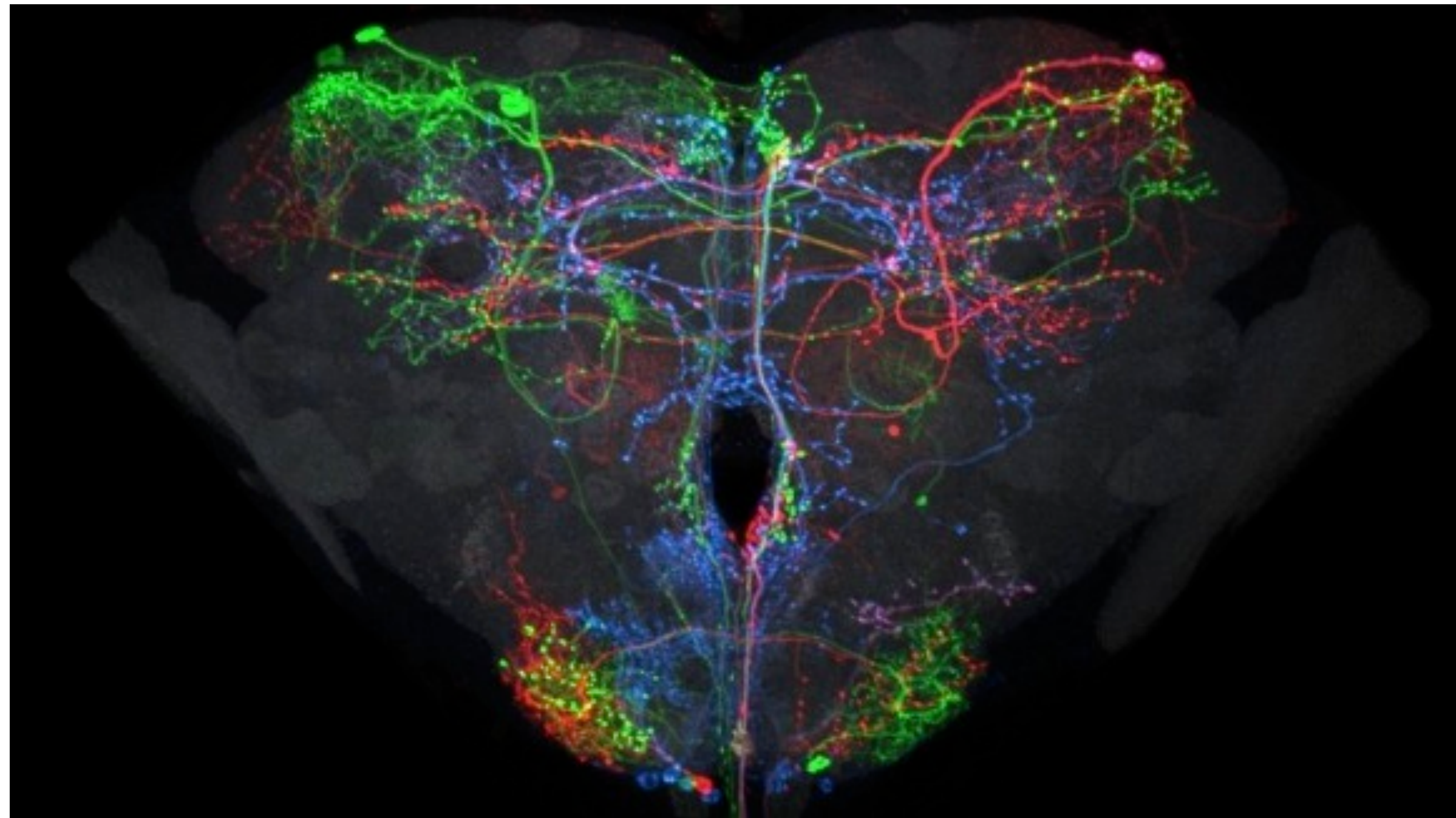
## Cre driver lines in mice

Drivers		
Drivers	Reporters	
<p>Data detailing transgene expression in Cre and other driver lines for adult and developing brain. Experiments include colorimetric in situ hybridization, fluorescent in situ hybridization and other histological methods.</p>		
Line Name	Example Image	Expression Pattern Summary
<a href="#">A930038C07Rik-Tg1-Cre</a> Allen Institute for Brain Science		<p>Widespread expression of reporter gene throughout the brain. Enriched in restricted populations within the olfactory areas, piriform cortex, hippocampus, and cerebellum. Adult Cre expression observed in restricted populations of striatum, layer 5 neocortex, hypothalamus, pons and medulla. This is different from the A930038C07Rik gene itself which is specifically expressed in layer 1.</p>
<a href="#">A930038C07Rik-Tg4-Cre</a> Allen Institute for Brain Science		<p>Scattered populations within cortical layers 4 and 5, septum, thalamus, and midbrain. In the cortex, unlike A930038C07Rik which is specifically expressed in layer 1, Cre-directed reporter expression is found enriched in a scattered population of cells in layer 5.</p>
<a href="#">Adcyap1-2A-Cre</a> Allen Institute for Brain Science		<p>Cre expression is enriched in restricted populations within the olfactory areas, hippocampus, striatum, thalamus, midbrain, pons, and medulla. Expression is scattered within the isocortex and hypothalamus. Reporter expression is widespread.</p>
<a href="#">Agrp-IRES-Cre</a> Bradford Lowell		<p>Enriched in the arcuate nucleus of the hypothalamus.</p>
<a href="#">Avp-IRES2-Cre</a> Allen Institute for Brain Science		<p>Expressed in restricted populations within the hypothalamus.</p>

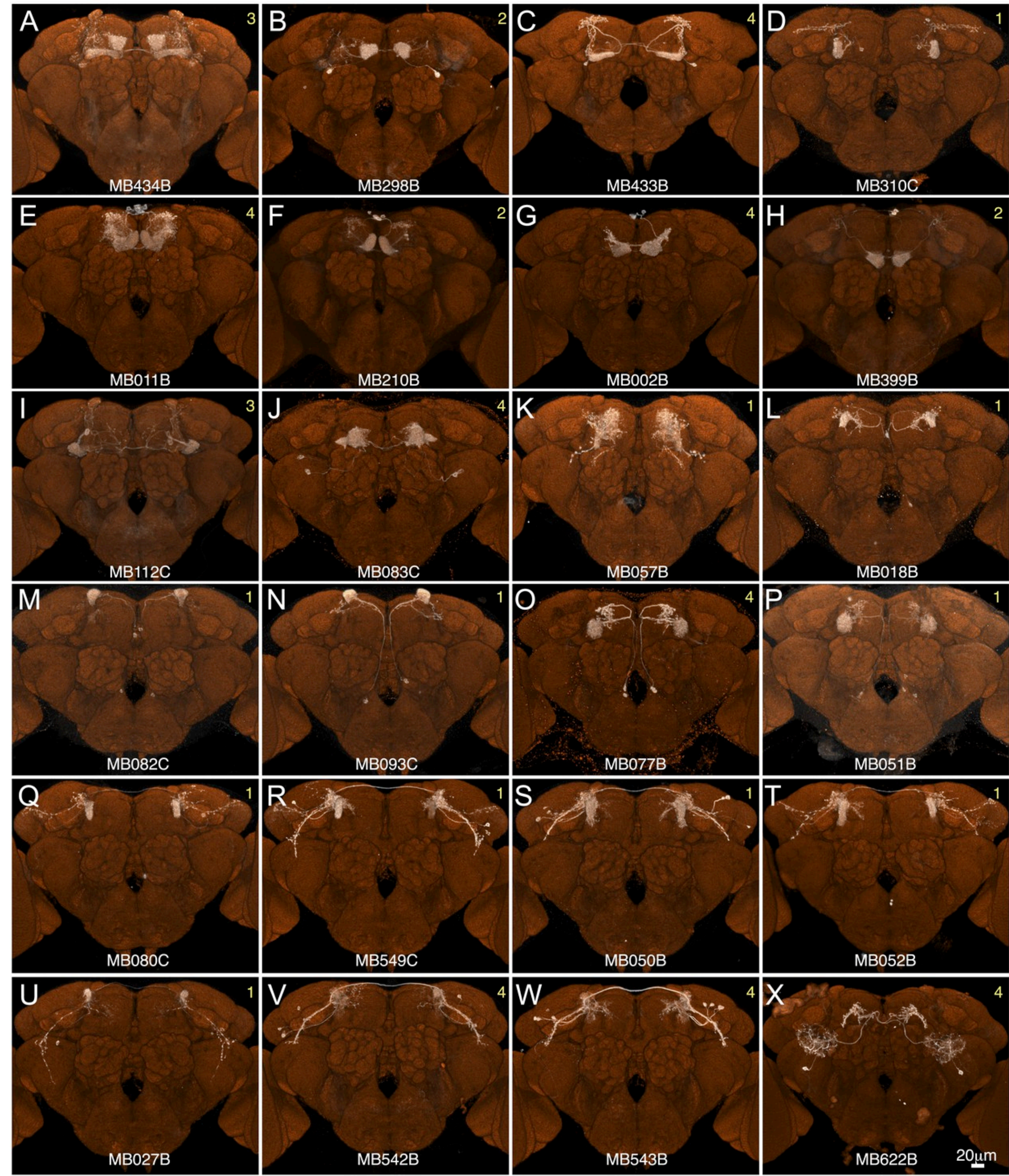
# Genetic tools

## GAL4 lines in flies

<https://www.janelia.org/node/45217>



Split-GAL4 lines for MBONs



# Genetically encoded indicators of neural activity

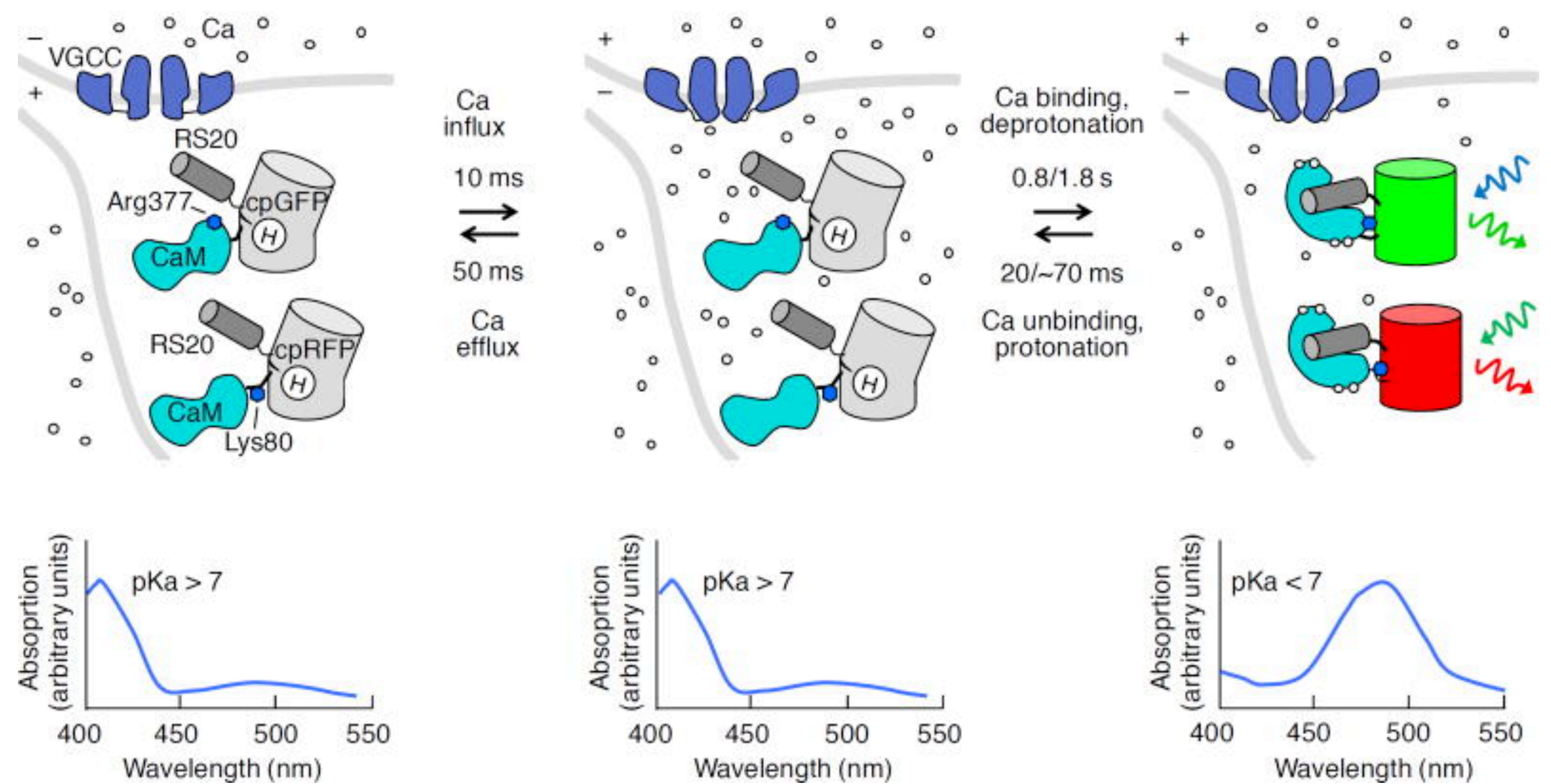
How can we make cells fluoresce only when they spike?

1. Look for a side effect of spiking.
2. Engineer a protein that fluoresces when that side effect is detected.
3. Modify the DNA of (subsets of) neurons to produce that protein.
4. Use a microscope to measure fluorescence in the genetically modified organism.



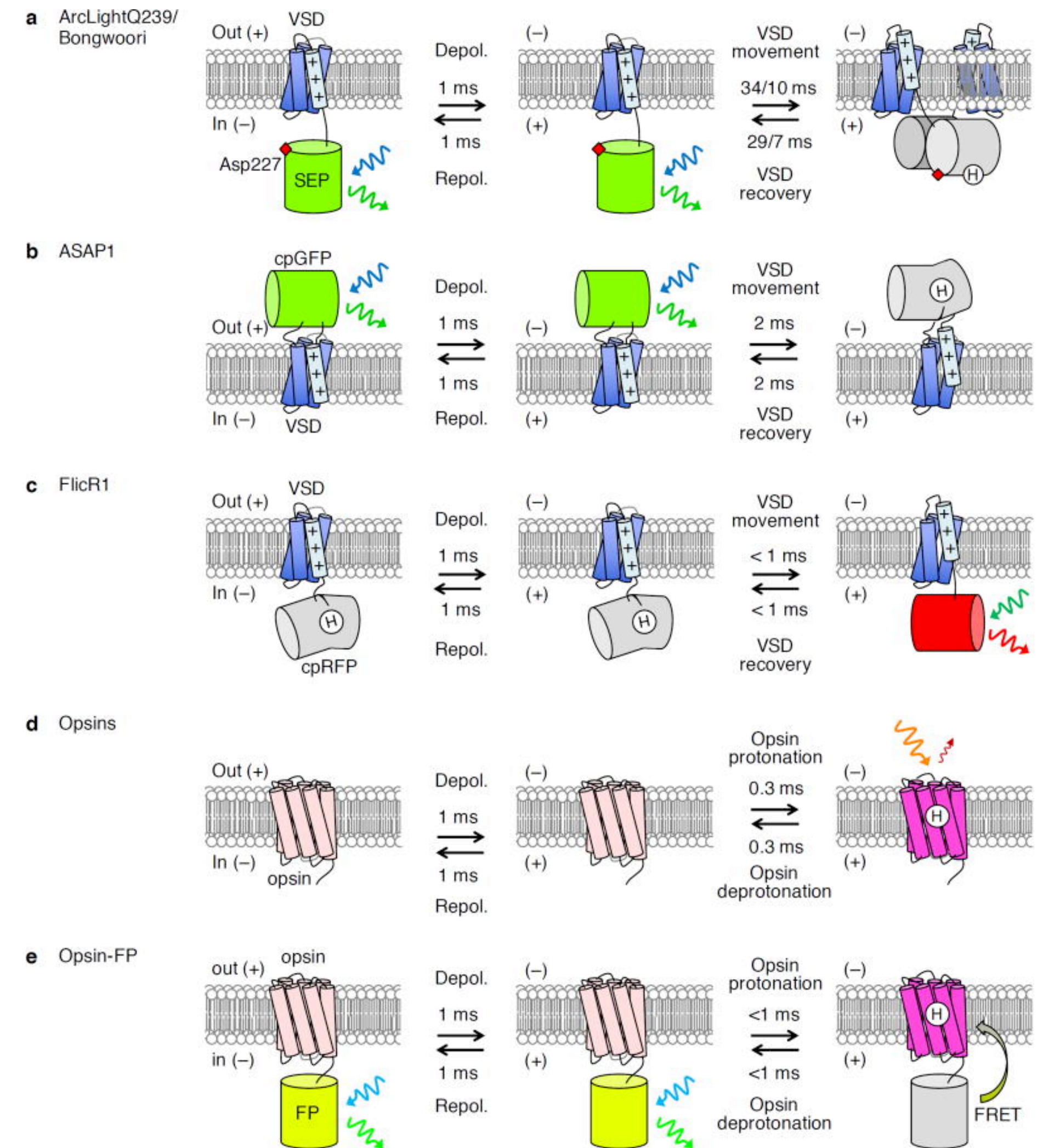
# Genetically encoded calcium indicators (GECIs)

- When neurons spike, voltage gated calcium channels (VGCCs) open and allow a rapid **influx of calcium ions** ( $\text{Ca}^{2+}$ ).
- Genetically encoded calcium indicators (GECIs) like **GCaMP** bind to these calcium ions and become fluorescent.
- The increased fluorescence decays as the calcium unbinds, producing a transient fluorescence indicative of neural spiking.
- Using driver lines, **GECIs can be targeted to specific cell types**.
- In some cases, **multiple GECIs** with different fluorescence wavelengths can be encoded simultaneously in **different subpopulations**.



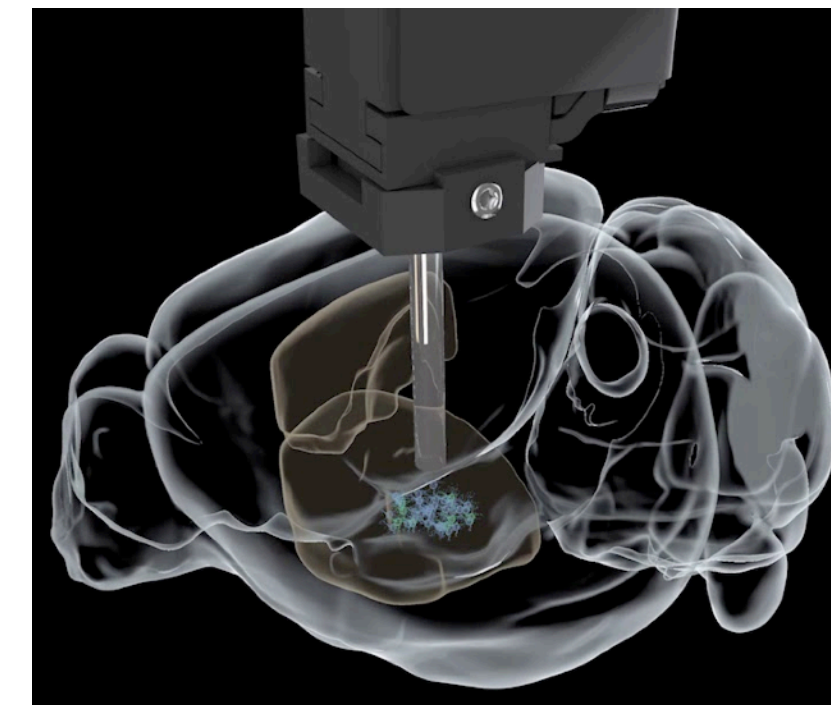
# Genetically encoded voltage indicators (GEVIs)

- Calcium is an indirect measure of spiking. Genetically encoded **voltage indicators** modulate fluorescence as a function of membrane potential.
- **Lots of designs:** fusing voltage sensing domains (e.g. from voltage-gated ion channels) to fluorescent proteins; harnessing natural opsins from microbes or algae.
- GEVIs are much more established, but great progress in GEVIs has been made in recent years.

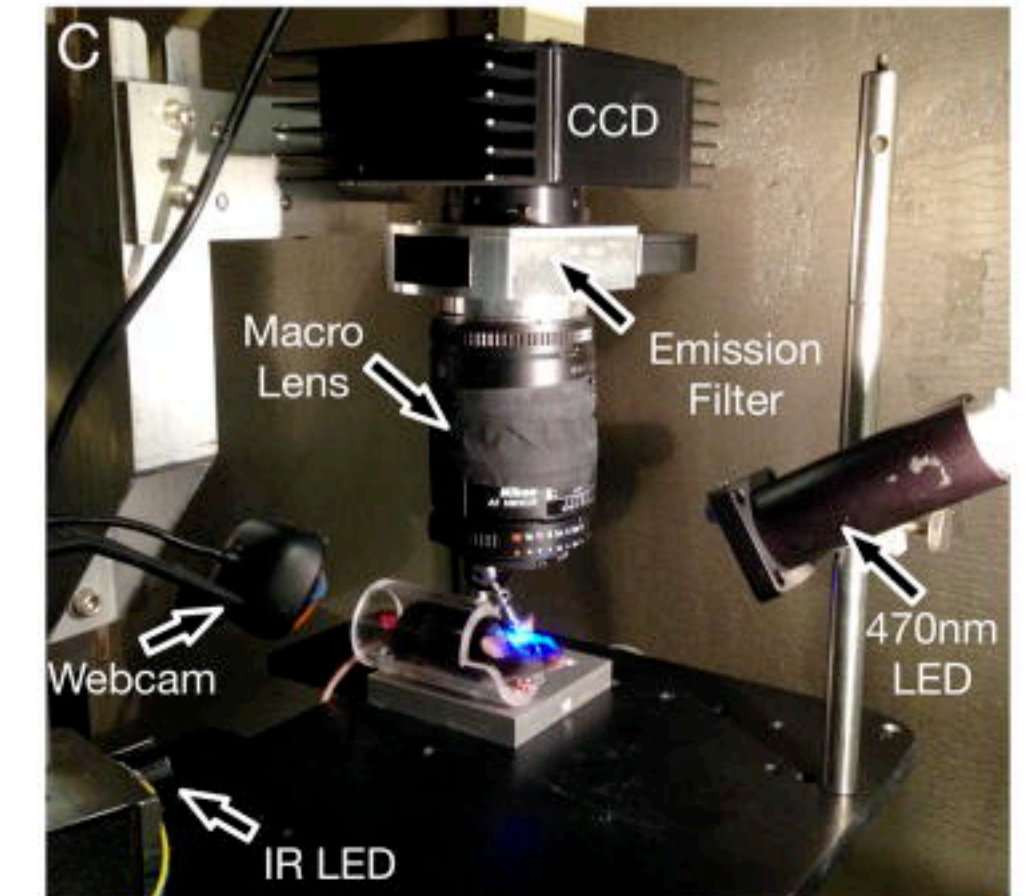


# Microscopy

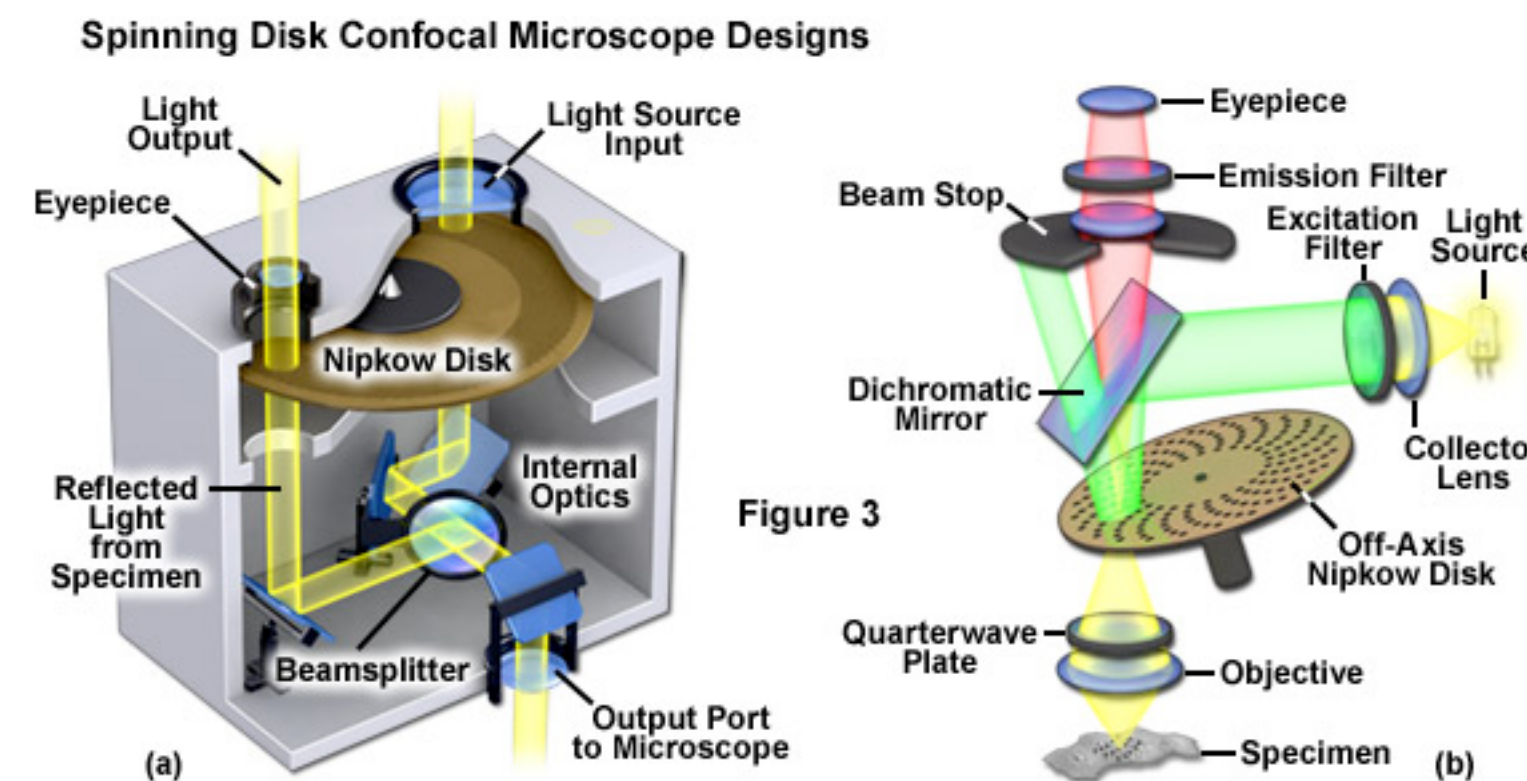
- Expressing the genetically encoded indicator is only half the battle.
- You still need to stimulate the cells with a light source and measure the resulting fluorescence.
- Again, there are lots of approaches: wide-field imaging, **2-photon microscopy**, laser scanning and spinning disk confocal microscopy, miniaturized GRIN lenses, fiber photometry.



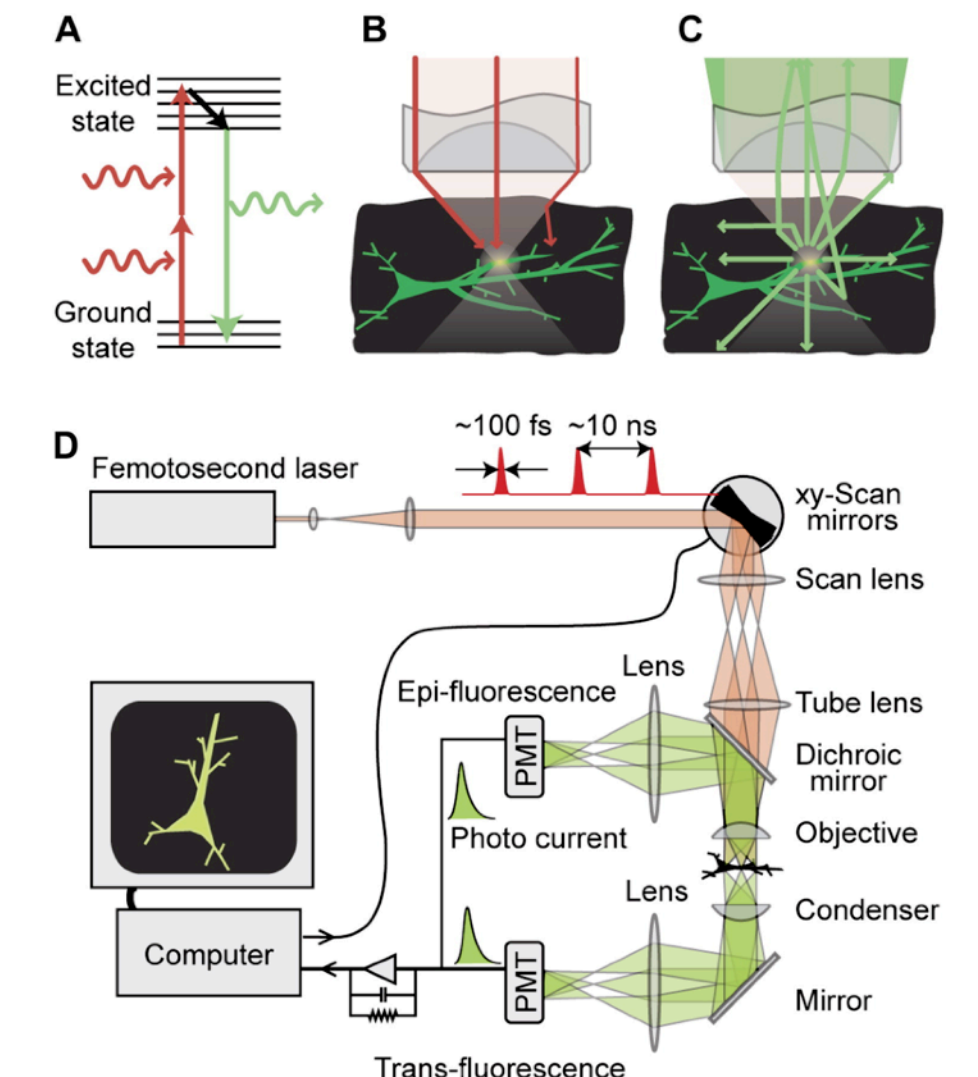
[www.inscopix.com](http://www.inscopix.com)



Silasi et al. J Neurosci Methods. 2016

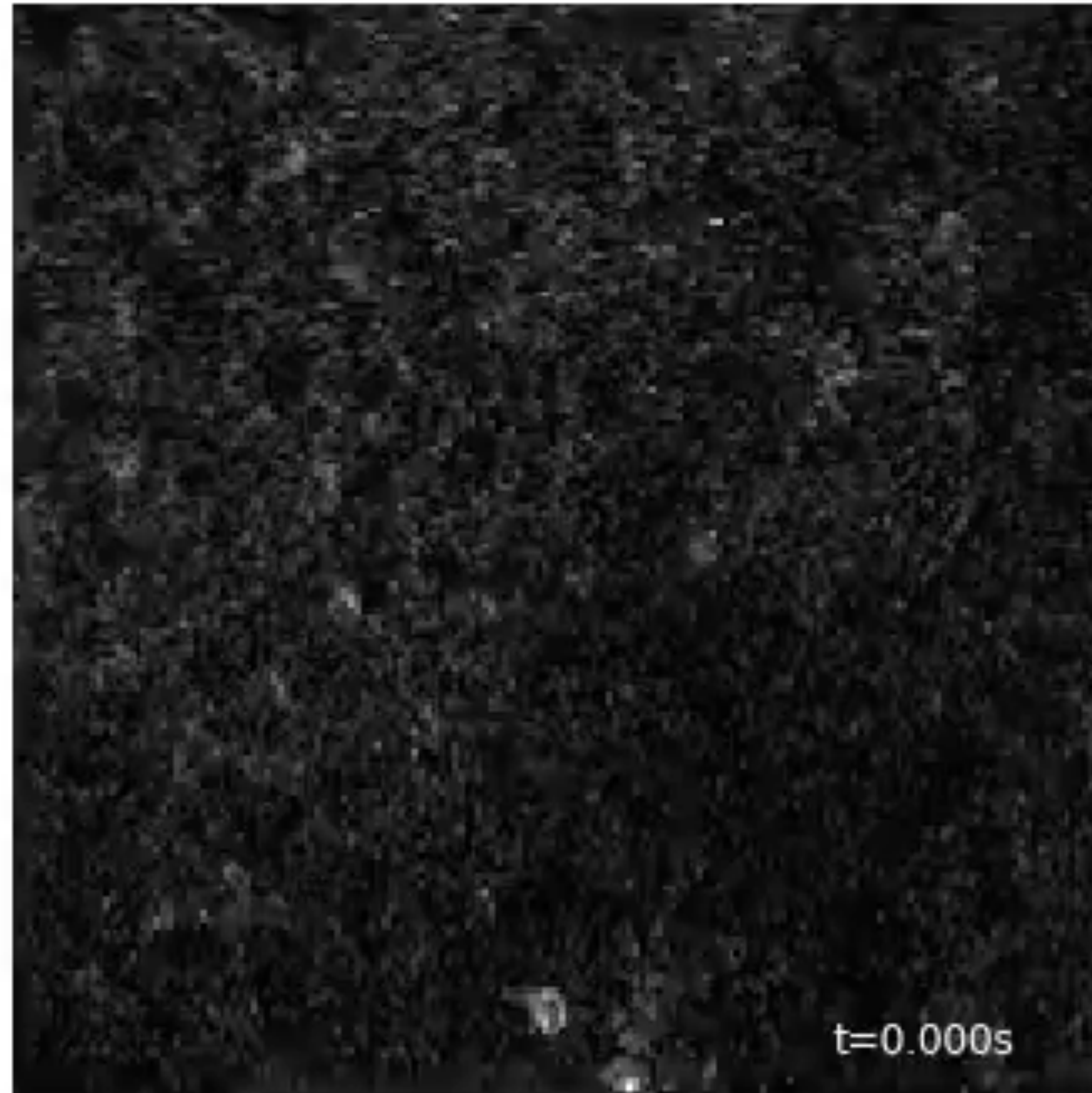


<http://zeiss-campus.magnet.fsu.edu/articles/spinningdisk/introduction.html>



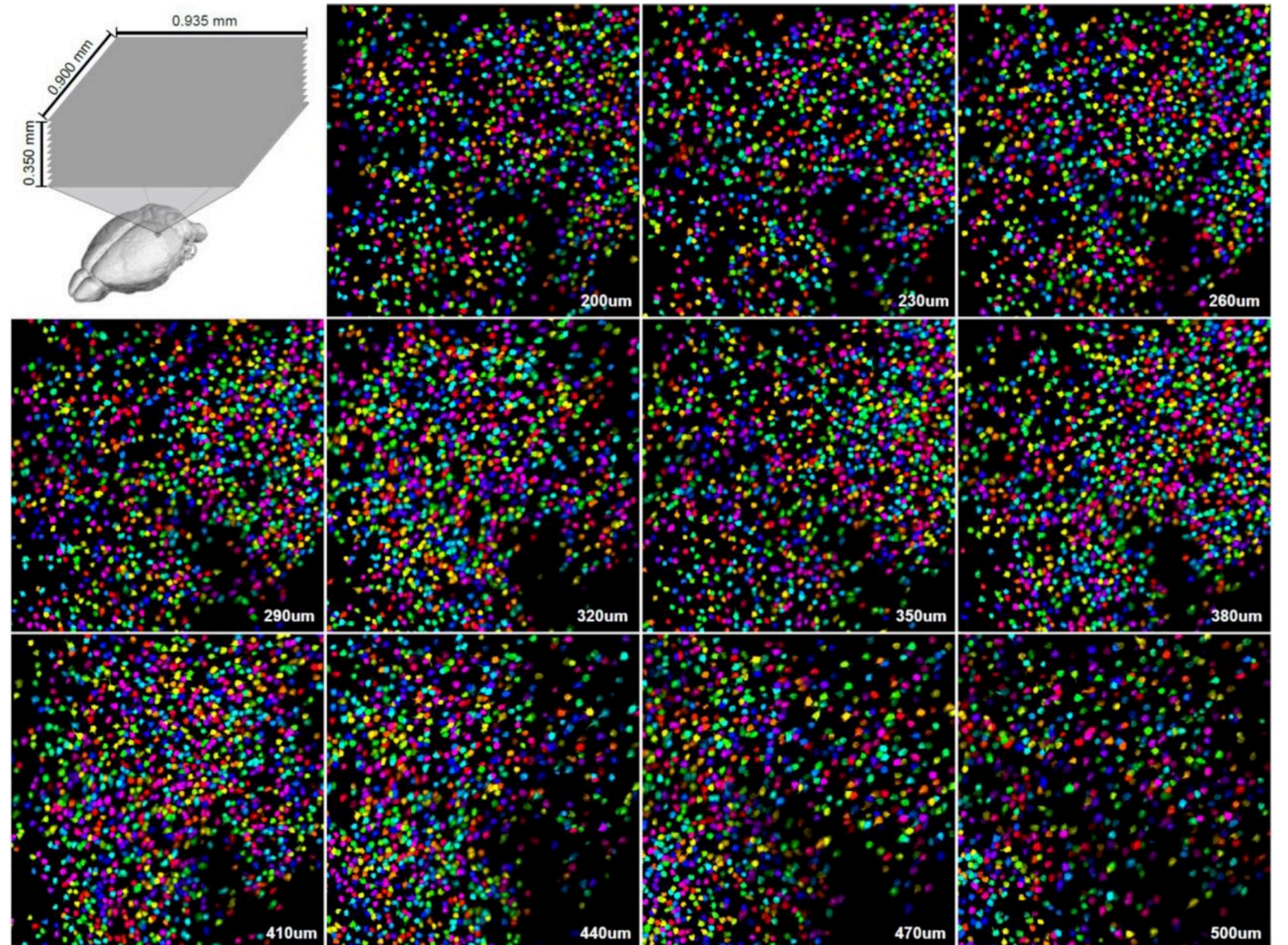
Svoboda and Yasuda. Neuron, 2006.

# 2 photon calcium imaging



# 2 photon calcium imaging

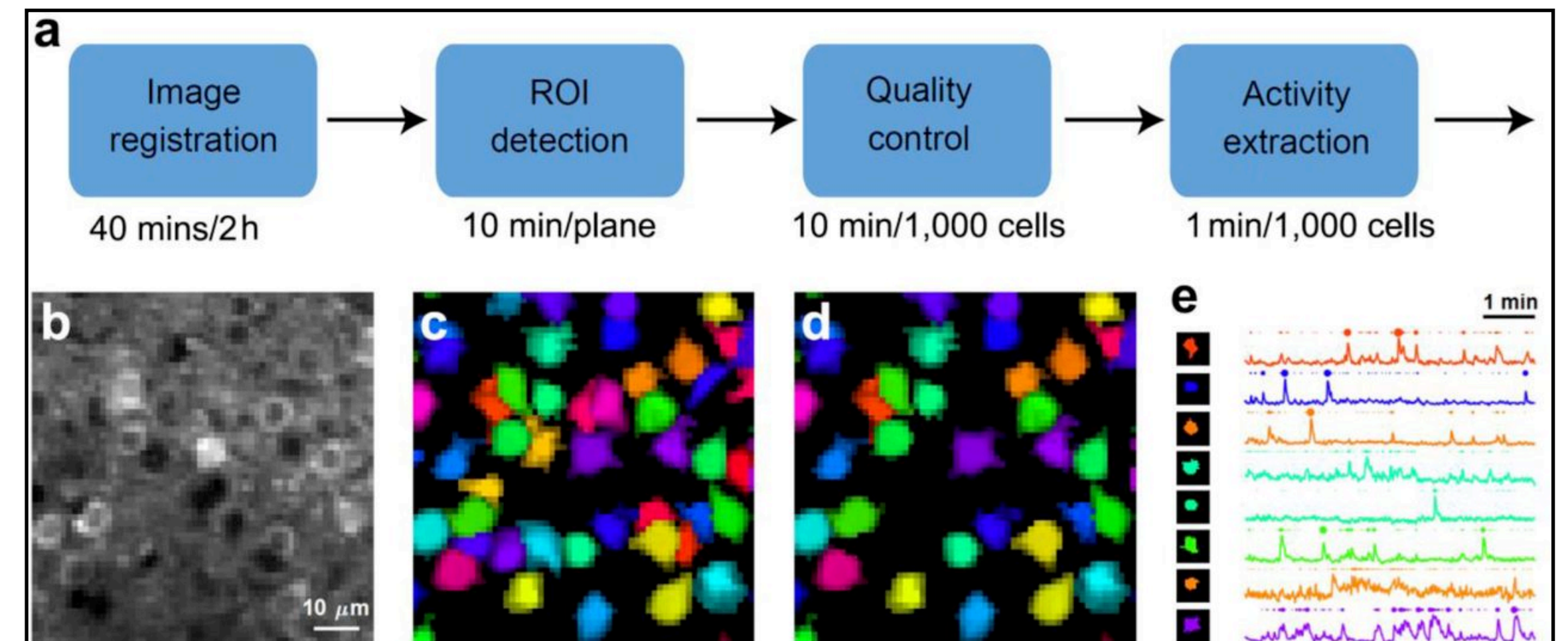
Over 10,000 cells



# Data analysis pipelines for 2P imaging

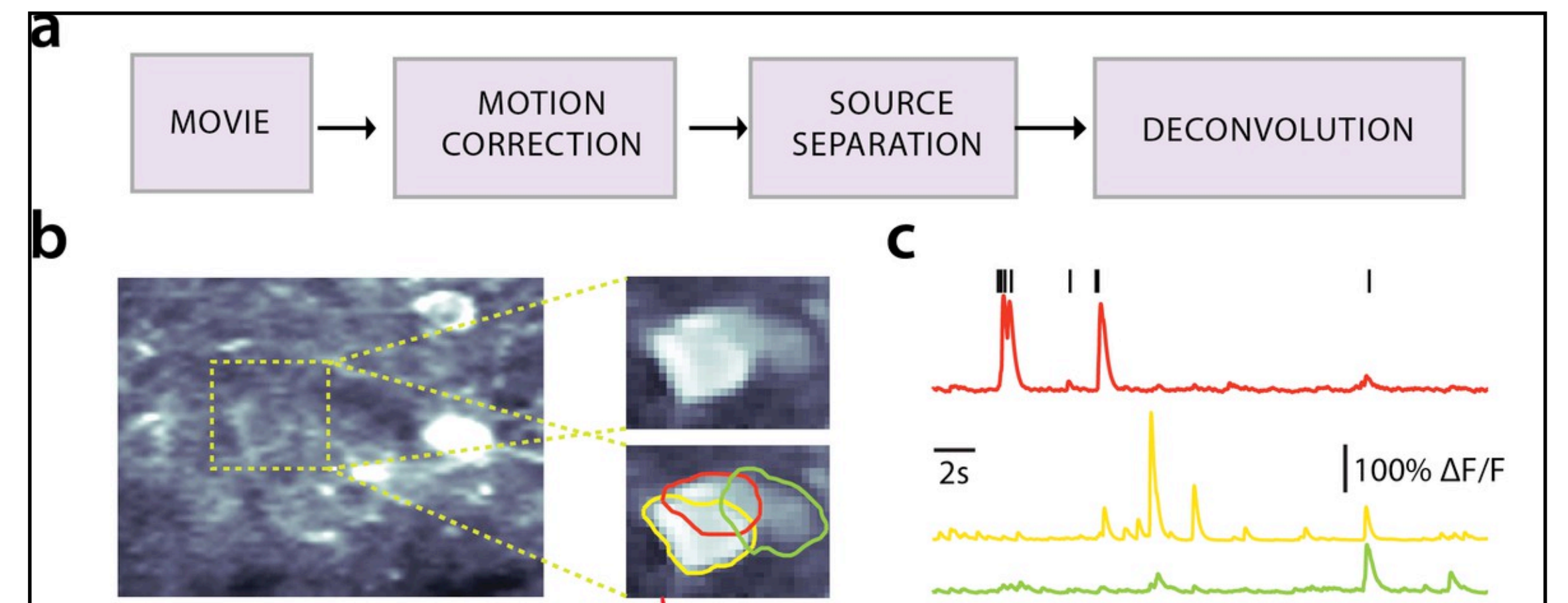
- Modern packages like Suite2P and CalmAn go through a few key steps to extract fluorescence traces.
- The key challenges are:
  - Correcting for motion artifacts.
  - Separating overlapping cells.
  - Accounting for background fluorescence.
  - Deconvolving spikes from fluorescence traces.

Suite2P



Pachitariu et al, bioRxiv 2017

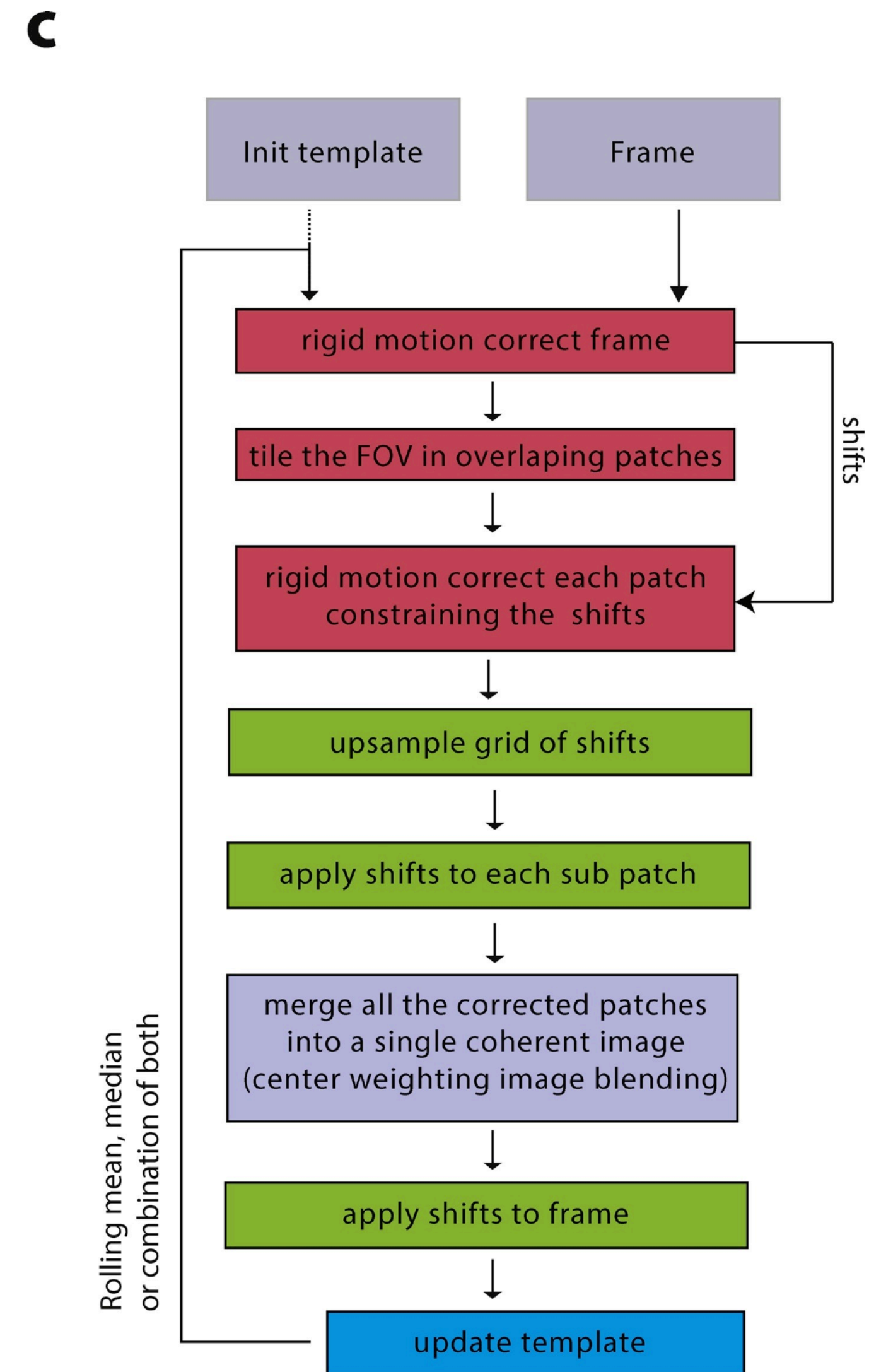
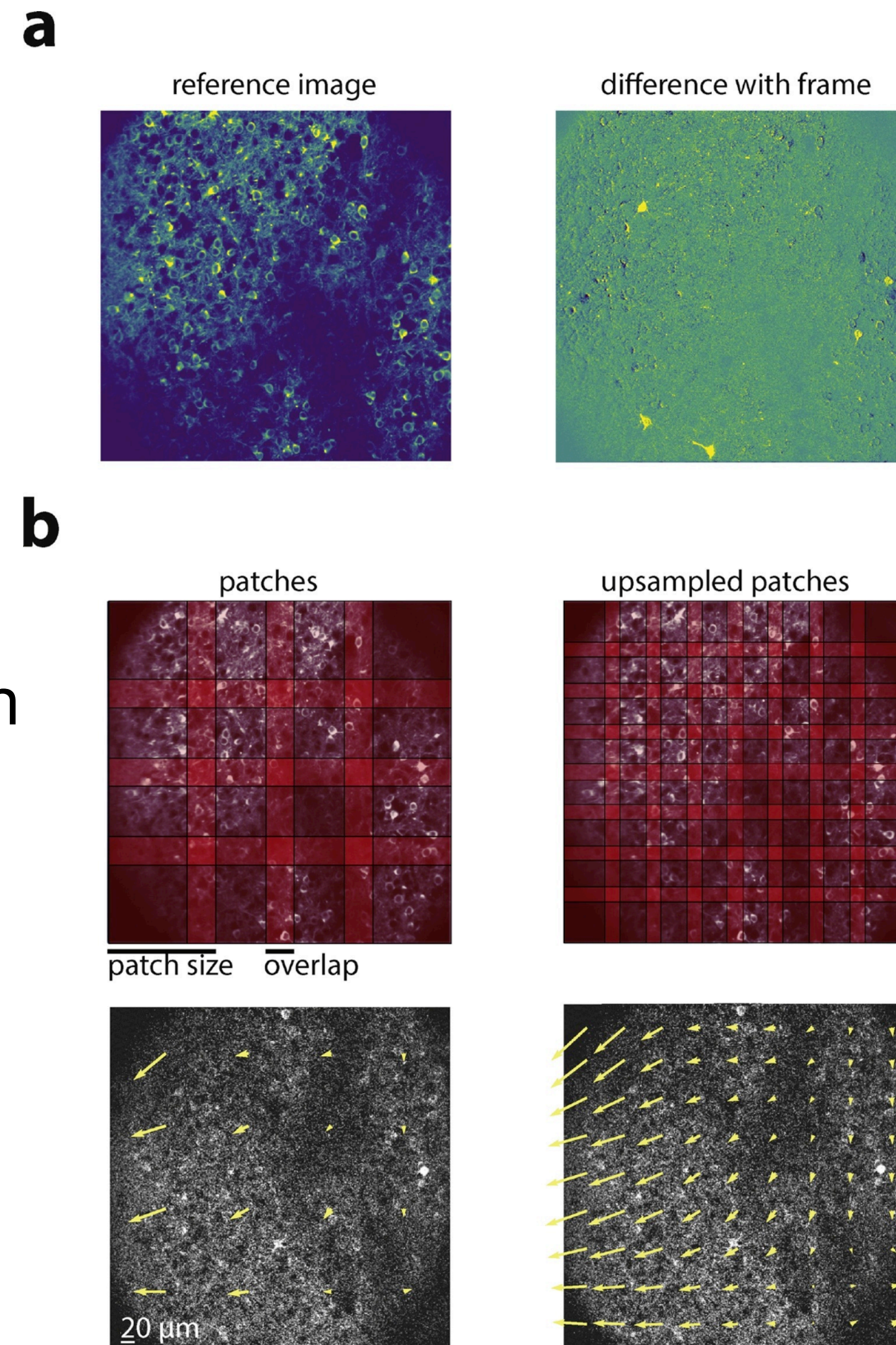
CalmAn



Giovanucci et al, eLife 2017

# Motion correction

- The brain is squishy and it moves in non-rigid ways in 3D during experiments.
- A variety of non-rigid motion correction algorithms have been proposed:
  - NoRMCorre (Pnevmatikakis and Giovannucci, 2017), used in CalmAn.
  - Phase correlation + kriging (Pachitariu et al, 2017) used in Suite2P.

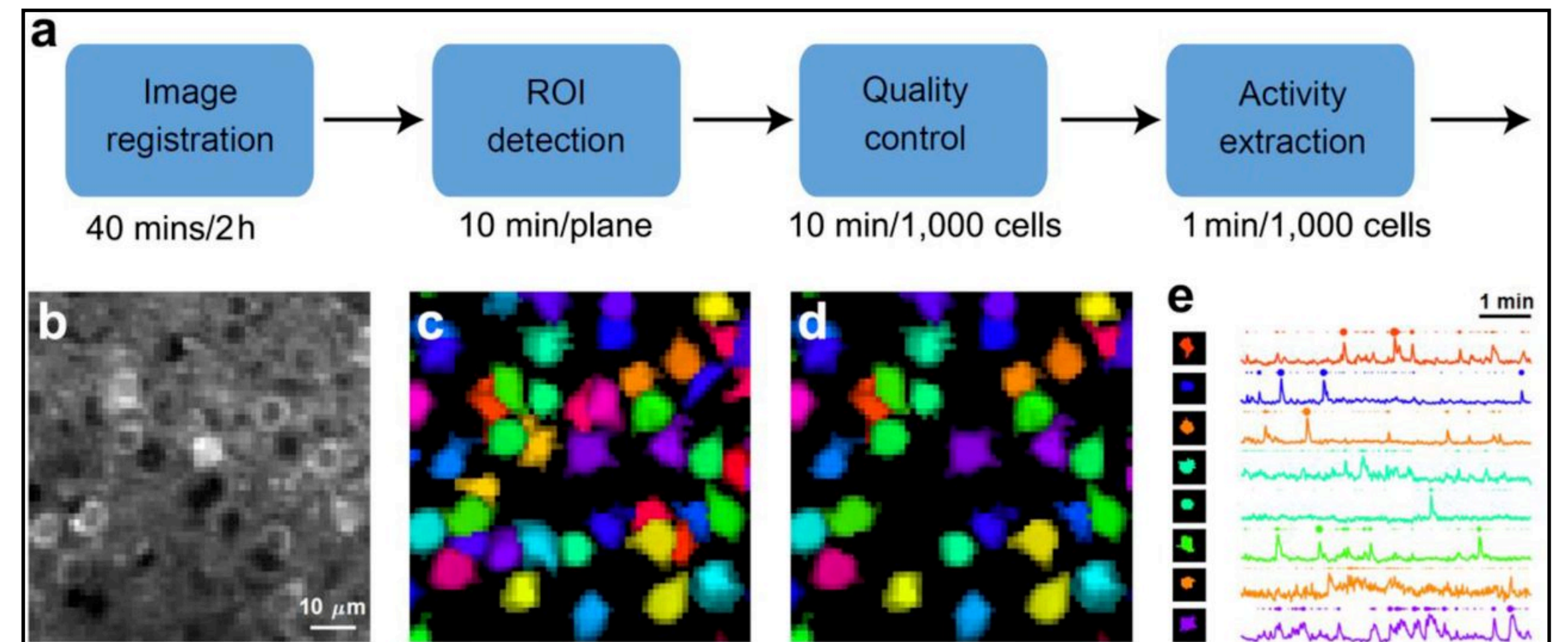


# Data analysis pipelines for 2P imaging

- Modern packages like Suite2P and CalmAn go through a few key steps to extract fluorescence traces.
- The key challenges are:
  - Correcting for motion artifacts.
  - Separating overlapping cells.
  - Accounting for background fluorescence.
  - Deconvolving spikes from fluorescence traces.

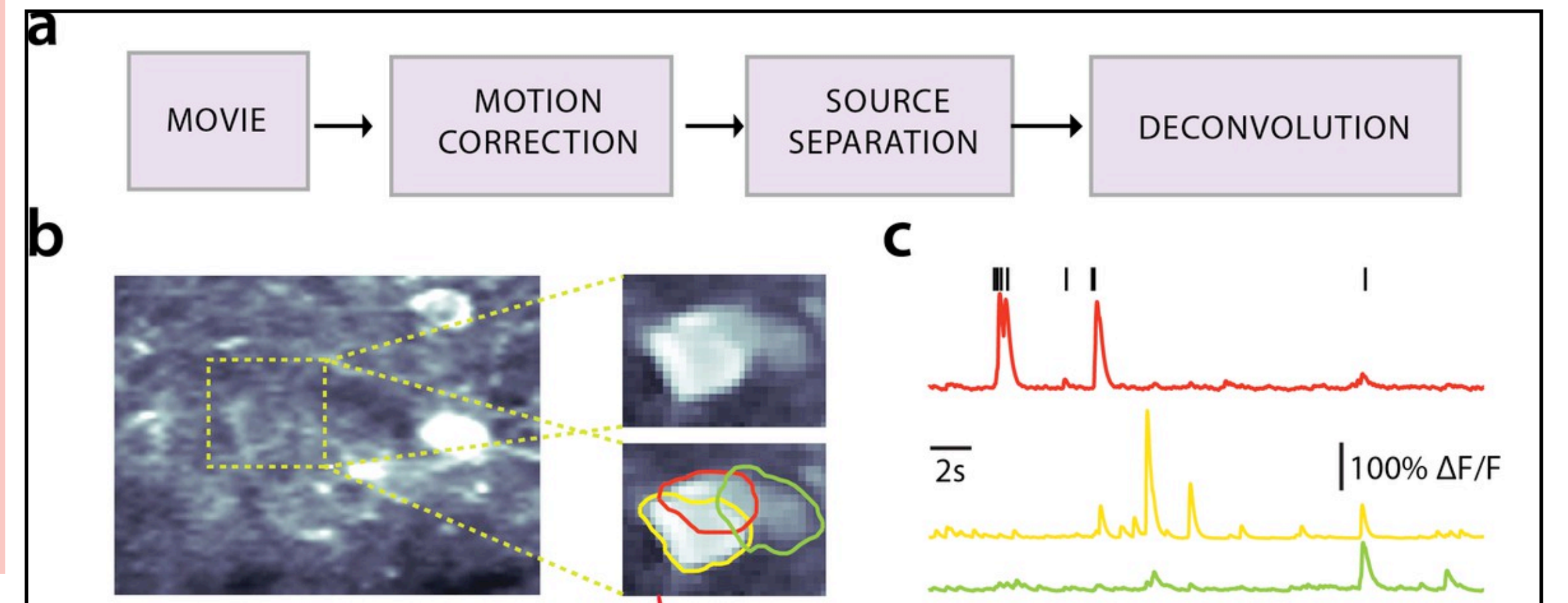
CNMF

Suite2P



Pachitariu et al, bioRxiv 2017

CalmAn



Giovanucci et al, eLife 2017

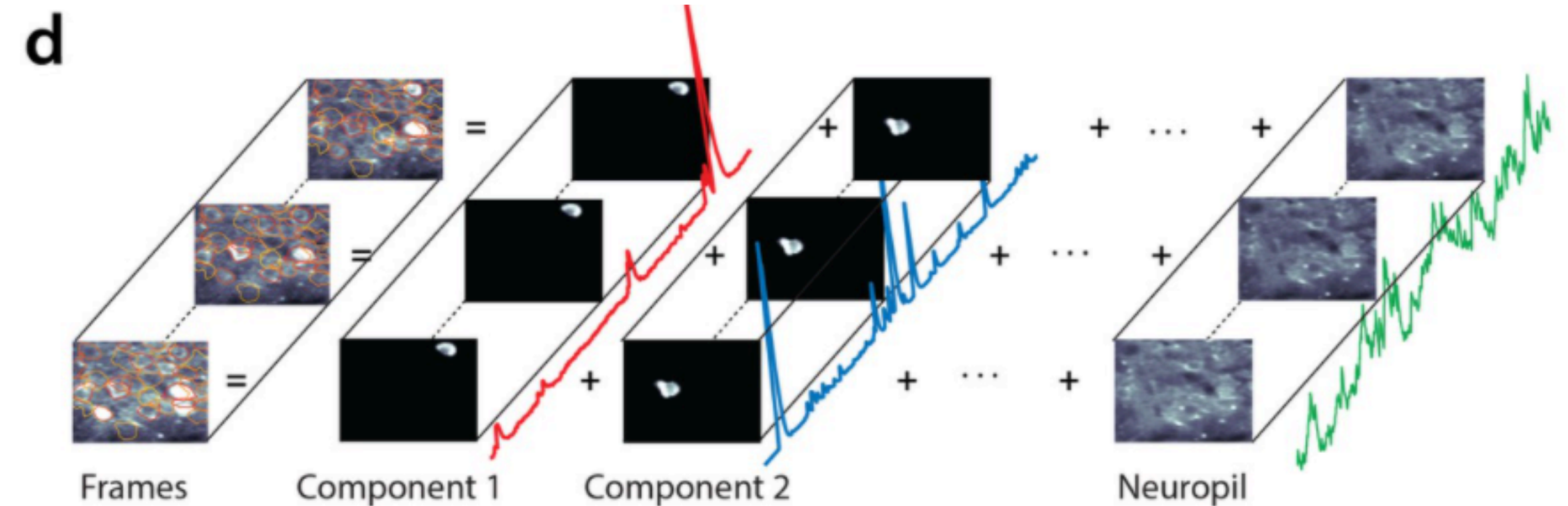


# Constrained Non-negative Matrix Factorization (CNMF)

*Pnevmatikakis et al, Neuron 2016.*

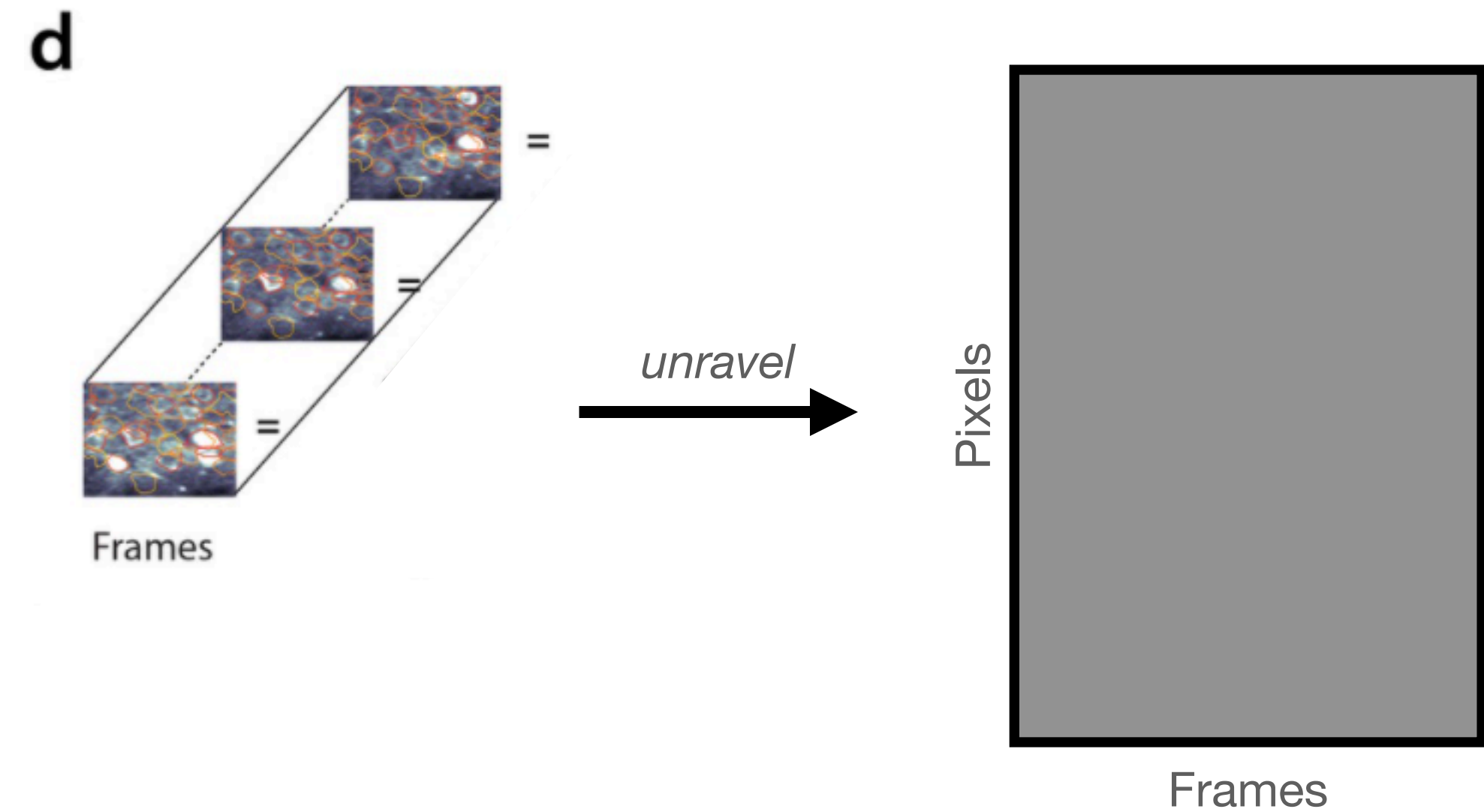
# CNMF

- Model the motion corrected movie as a superposition of fluorescence traces from multiple neurons, plus background.
- We can pose this as another convolutional matrix factorization problem.
- **Punchline:** *it's nearly the same as what we did for spike sorting!*



# Constants

- Let  $T$  denote the number of **frames** in the movie.
- $N$  denote the number of **pixels**.
- $D$  denote the **duration** (in frames) of a calcium spike.
- $K$  denote the (unknown) number of **neurons** that generated the spikes.



# Data and Latent Variables

- **Data:**

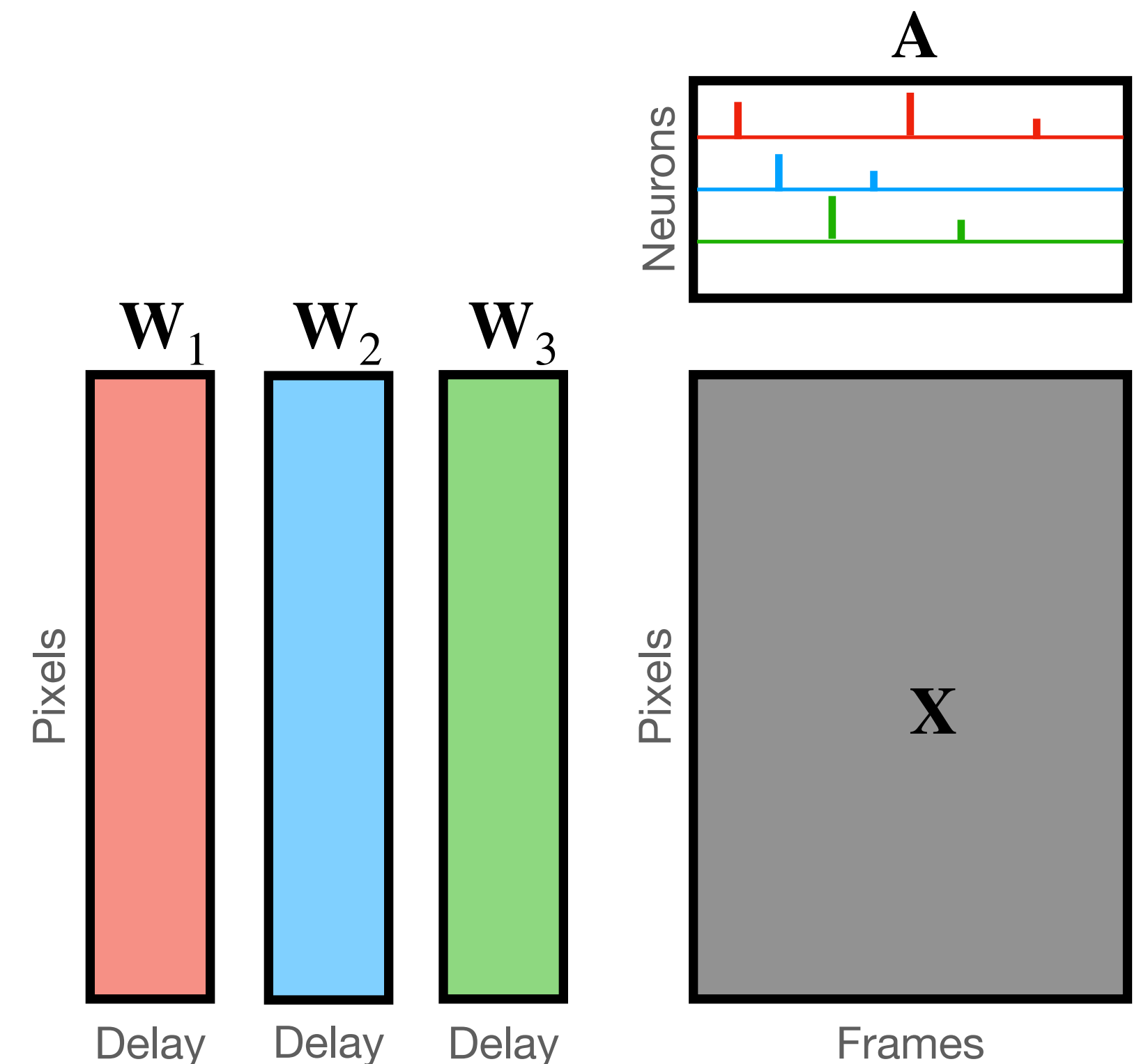
- Let  $\mathbf{X} \in \mathbb{R}^{N \times T}$  denote the motion corrected and unraveled video.

- **Latent Variables:**

- Let  $\mathbf{A} \in \mathbb{R}_+^{K \times T}$  denote the time series of spike amplitudes for each neuron.

- **Parameters:**

- Let  $\mathbf{W} \in \mathbb{R}^{K \times N \times D}$  denote the array of calcium responses for each neuron.

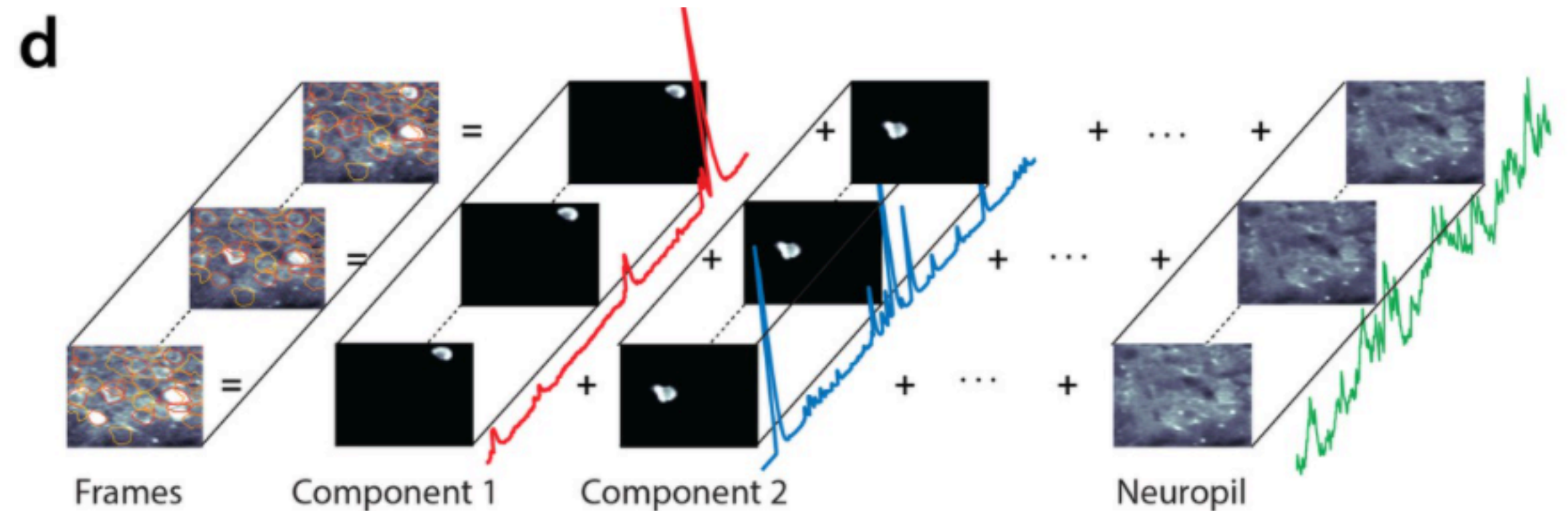


# Probabilistic Model

## Likelihood

Like last week, assume each spike induces a scaled calcium response in the video.

$$p(\mathbf{X} | \mathbf{A}, \mathbf{W}) = \prod_{t=1}^T \mathcal{N} \left( \mathbf{x}_t \mid \sum_{k=1}^K [\mathbf{a}_k \circledast \mathbf{W}_k]_t + \mathbf{u}_0 \mathbf{c}_{0,t}, \sigma^2 \mathbf{I} \right)$$



# Calcium response model

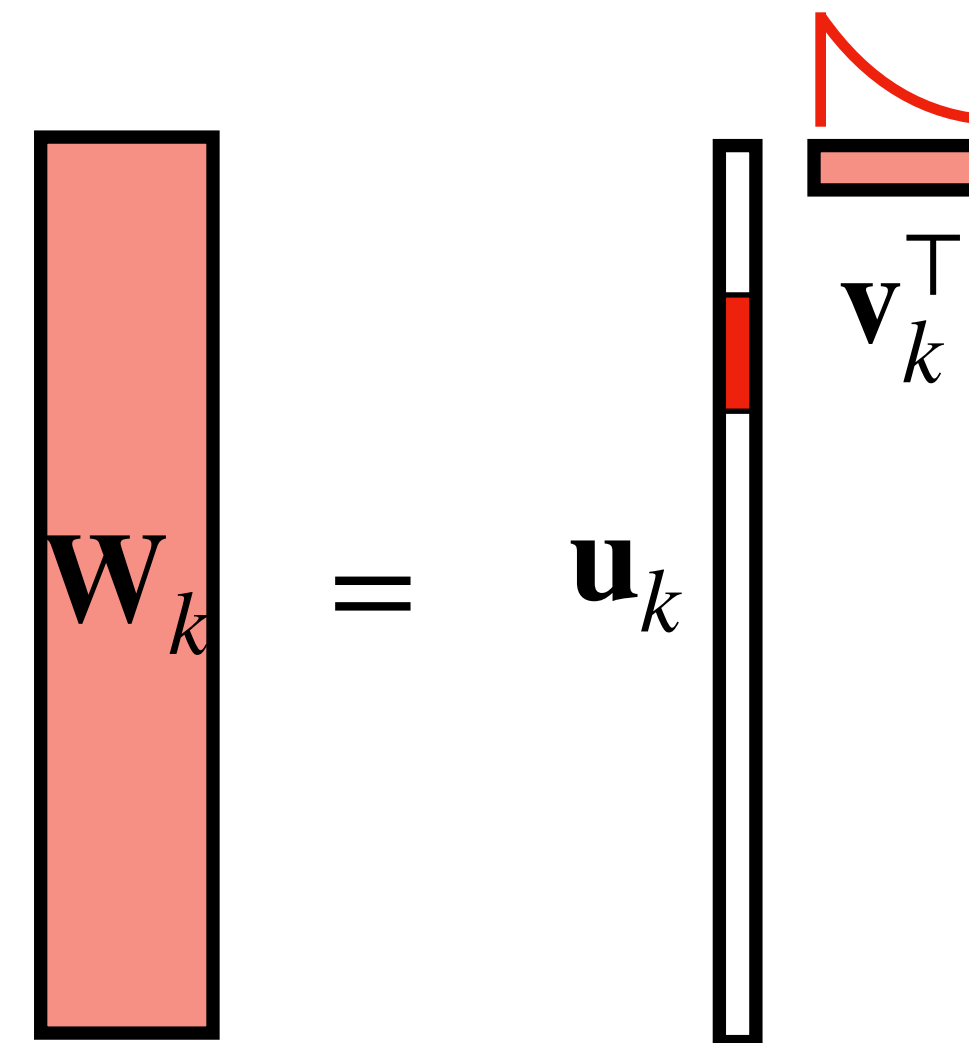
- Assume the calcium responses factor into spatial and temporal components.

$$\mathbf{W}_k = \mathbf{u}_k \mathbf{v}_k^\top$$

- Spatial factor  $\mathbf{u}_k$  specifies which pixels correspond to neuron  $k$ .
- Constrain the temporal components to be exponential decays.

$$v_{k,d} = e^{-d/\tau}$$

- Time constant of the decay is a function of the indicator;  $O(100\text{ms})$ .



# Calcium response model

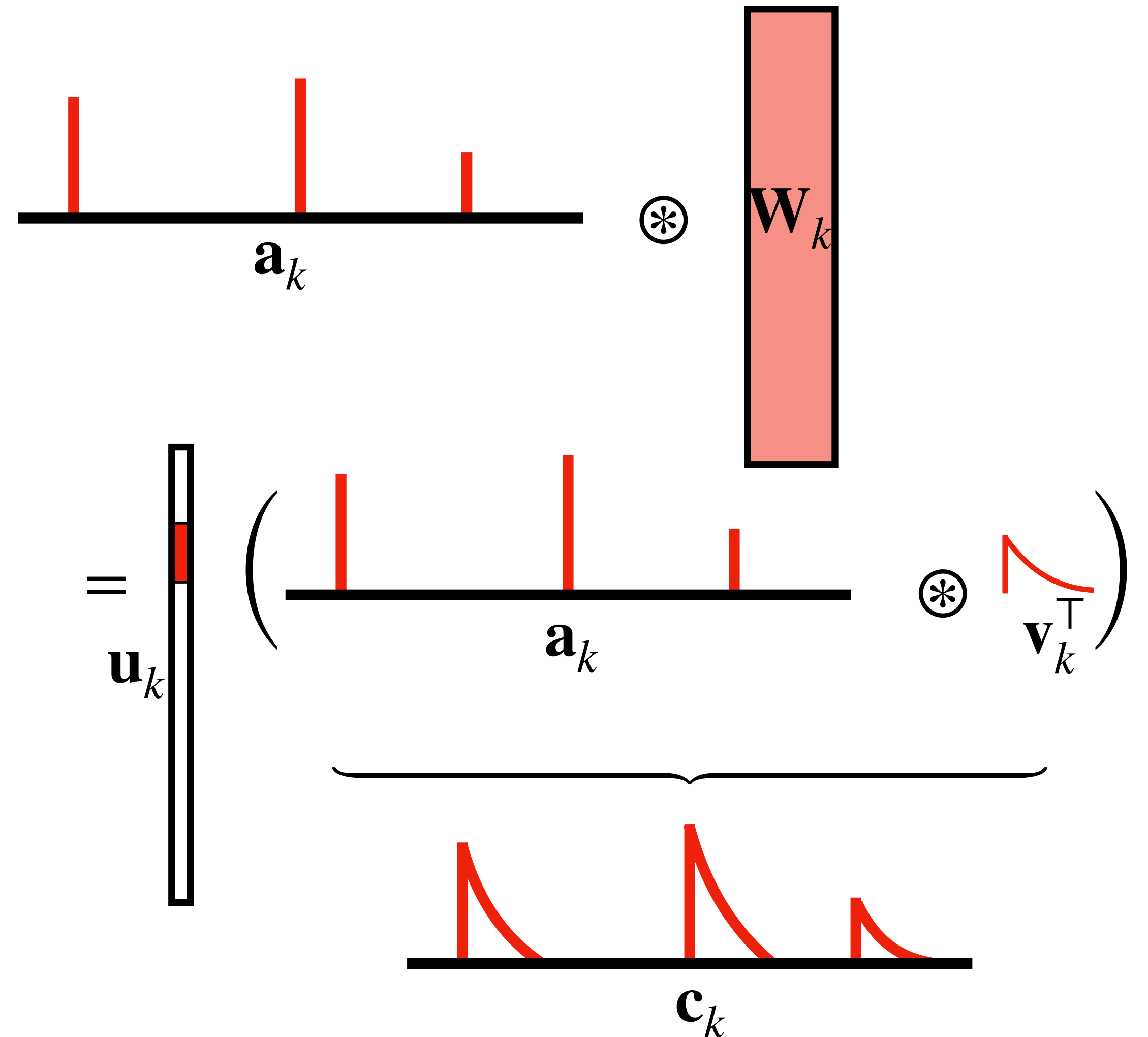
Then

$$[\mathbf{a}_k \circledast \mathbf{W}_k]_t = \mathbf{u}_k [\mathbf{a}_k \circledast \mathbf{v}_k]_t \triangleq \mathbf{u}_k \mathbf{c}_{k,t}$$

where

$$\mathbf{c}_{k,t} \triangleq [\mathbf{a}_k \circledast \mathbf{v}_k]_t = \sum_{d=0}^D a_{k,t-d} v_{k,d} = \sum_{d=0}^D a_{k,t-d} e^{-d/\tau}$$

is the calcium trace of neuron  $k$



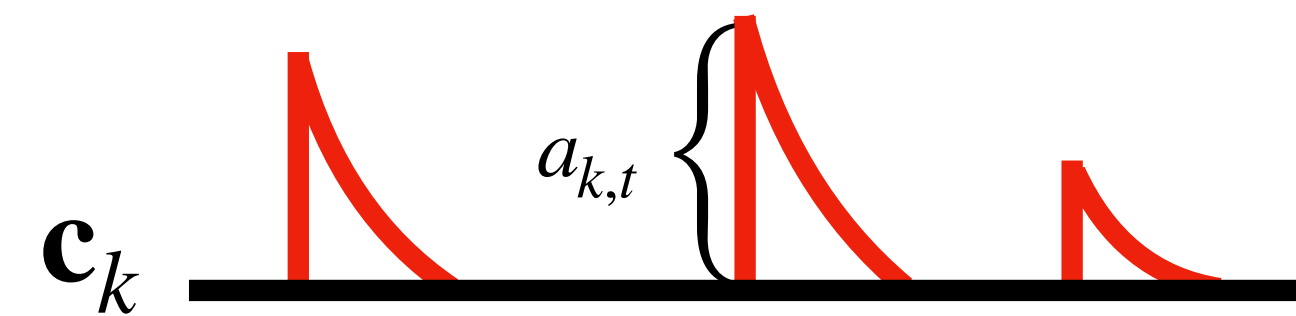
# Recursive formulation

The calcium response can be written recursively, thanks to the **exponential response**:

$$\begin{aligned}c_{k,t} &= \sum_{d=0}^D a_{k,t-d} e^{-d/\tau} \\ &= a_{k,t} + \sum_{d=1}^{t-1} a_{k,t-d} e^{-d/\tau} \\ &= a_{k,t} + \sum_{d=0}^{t-2} a_{k,t-(d+1)} e^{-(d+1)/\tau} \\ &= a_{k,t} + e^{-1/\tau} c_{k,t-1},\end{aligned}$$

Equivalently,  $a_{k,t} = c_{k,t} - e^{-1/\tau} c_{k,t-1}$ .

(Note, we took  $D \rightarrow \infty$  and zero-padded  $\mathbf{a}_k$  on the left.)

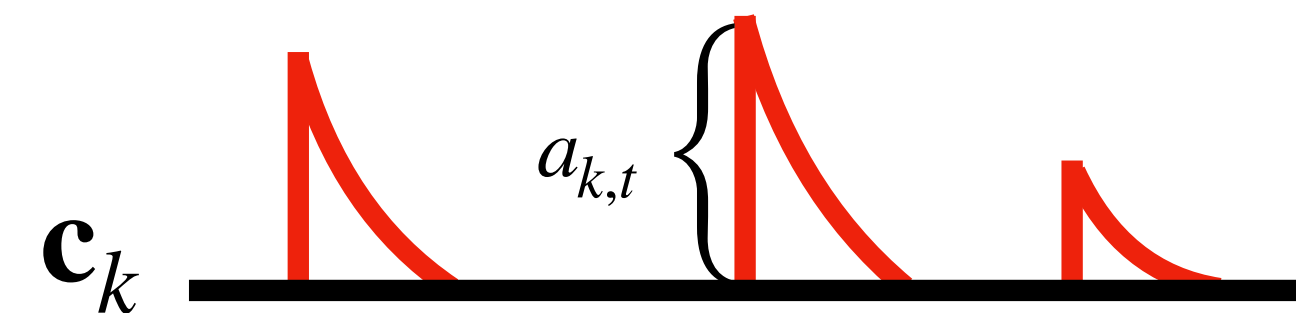




# Recursive formulation

In matrix form,

$$\mathbf{a}_k = \mathbf{G}\mathbf{c}_k \quad \mathbf{G} = \begin{bmatrix} 1 & & & & \\ e^{-1/\tau} & 1 & & & \\ & e^{-1/\tau} & 1 & & \\ & & & \ddots & \\ & & & & \ddots \end{bmatrix}.$$



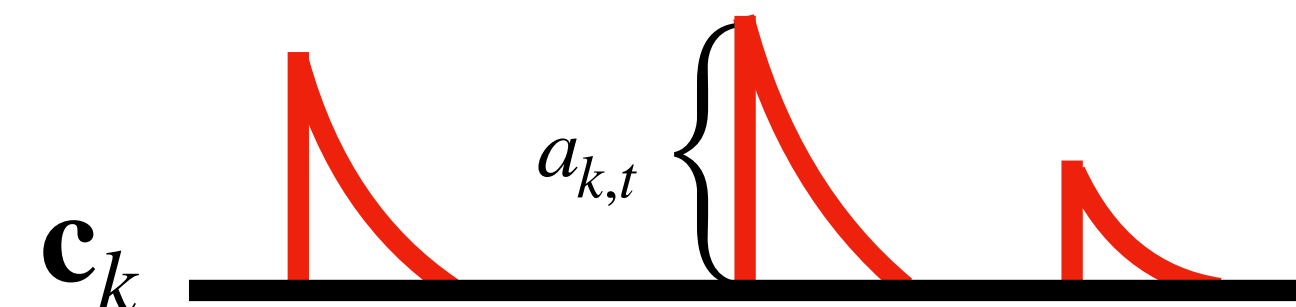
# Prior on calcium traces

## Via a prior on amplitudes

Suppose  $a_{k,t} \sim \text{Exp}(\lambda)$ , as in the spike sorting model.

We can derive the probability of  $\mathbf{c}_k$  using the change of measure formula,

$$\begin{aligned} p(\mathbf{c}_k) &= \left| \frac{d\mathbf{a}_k}{d\mathbf{c}_k} \right| \prod_{t=1}^T \text{Exp}(c_{k,t} - e^{-1/\tau} c_{k,t-1}; \lambda) \\ &= |\mathbf{G}| \prod_{t=1}^T \text{Exp}(c_{k,t} - e^{-1/\tau} c_{k,t-1}; \lambda) \end{aligned}$$



# Prior on calcium traces

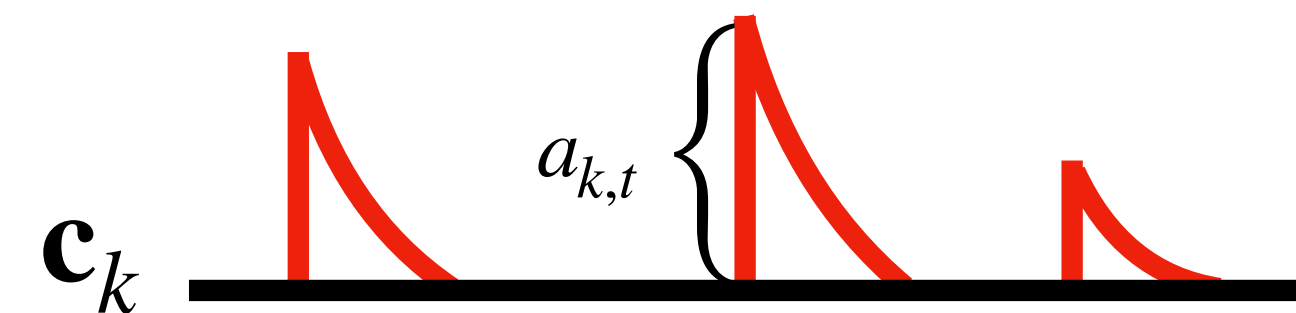
## Via a prior on amplitudes

Since  $\mathbf{G}$  is lower triangular, its determinant is the product of its diagonal; i.e. 1.

Thus,

$$p(\mathbf{c}_k) = \prod_{t=1}^T \text{Exp}(c_{k,t} - e^{-1/\tau} c_{k,t-1}; \lambda)$$

It's just the probabilities of the “jumps”  $a_{k,t}$ .



# Optimizing the calcium traces

Following the same steps as last week, we end up with the following objective for optimizing the calcium traces:

$$\mathcal{L}(\mathbf{c}_k) = -\frac{1}{2\sigma^2} \|\mathbf{c}_k - \boldsymbol{\mu}_k\|_2^2 + \lambda \sum_{t=1}^T (c_{k,t} - e^{-1/\tau} c_{k,t-1}),$$

where

$$\boldsymbol{\mu}_k = \mathbf{R}^\top \mathbf{u}_k$$

is the residual projected onto the spatial factor for this neuron.

# Optimizing the calcium traces

More compactly,

$$\begin{aligned}\mathcal{L}(\mathbf{c}_k) &= -\frac{1}{2\sigma^2} \|\mathbf{c}_k - \boldsymbol{\mu}_k\|_2^2 + \lambda \sum_{t=1}^T (c_{k,t} - e^{-1/\tau} c_{k,t-1}) \\ &= -\frac{1}{2\sigma^2} \|\mathbf{c}_k - \boldsymbol{\mu}_k\|_2^2 + \lambda \|\mathbf{G}\mathbf{c}_k\|_1.\end{aligned}$$

We still have that pesky hyperparameter  $\lambda$ ...

# Optimizing the calcium traces

## Dual formulation

- Maximizing  $\mathcal{L}(\mathbf{c}_k)$  is equivalent to solving the following convex optimization problem,

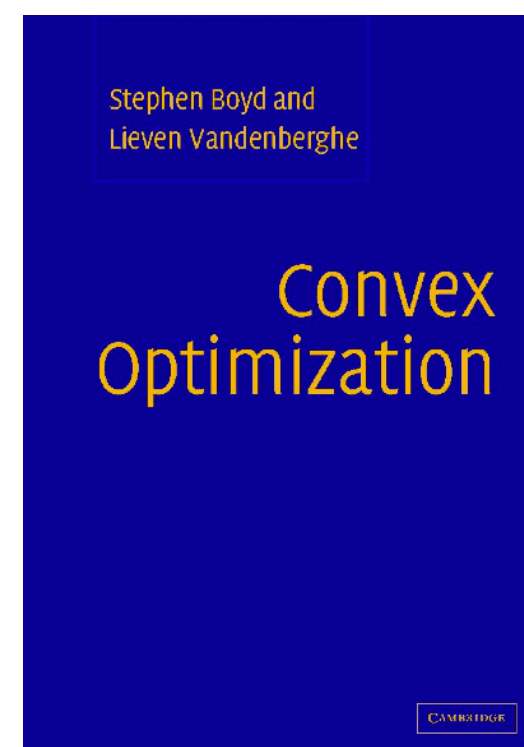
$$\hat{\mathbf{c}}_k = \arg \min_{\mathbf{c}_k} \|\mathbf{G}\mathbf{c}_k\|_1 \quad \text{subject to} \quad \|\mathbf{c}_k - \boldsymbol{\mu}_k\|_2 \leq \theta, \quad \mathbf{G}\mathbf{c}_K \geq 0,$$

for some threshold  $\theta$ .

- Under the model,  $c_{k,t} - \mu_{k,t} \sim \mathcal{N}(0, \sigma^2)$ , and  $z_{k,t} = \frac{c_{k,t} - \mu_{k,t}}{\sigma} \sim \mathcal{N}(0, 1)$ .
- $\|z\|_2$  is the norm of a vector of iid Gaussians. It follows a chi ( $\chi$ ) distribution.
- **Idea:** for large  $T$ , the chi distribution concentrates around  $\sqrt{T}$ . So set  $\theta = (1 + \epsilon)\sigma\sqrt{T}$ .
- How to get  $\sigma$ ? We can estimate the noise at each pixel by high-pass filtering the data, then standardize the data by dividing by the noise standard deviation so that in our model  $\sigma = 1$ .

# CVXPY

- CVXPY is a powerful library for convex optimization in Python, based on the CVX package from Grant and Boyd.
- It's ideally suited to solving these types of problems.
- If you want to learn more, take Prof. Boyd's course, EE364, and read his book!



## CVXPY

 Star 4,295

### Navigation

[Install](#)

[User Guide](#)

[Examples](#)

[API Documentation](#)

[FAQ](#)

[Citing CVXPY](#)

[Contributing](#)

[Related Projects](#)

[Changes to CVXPY](#)

[CVXPY Short Course](#)

[License](#)

### Quick search

### Version selector

## Welcome to CVXPY 1.3

Convex optimization, for everyone.

We are building a CVXPY community [on Discord](#). Join the conversation!

CVXPY is an open source Python-embedded modeling language for convex optimization problems. It lets you express your problem in a natural way that follows the math, rather than in the restrictive standard form required by solvers.

For example, the following code solves a least-squares problem with box constraints:

```
import cvxpy as cp
import numpy as np

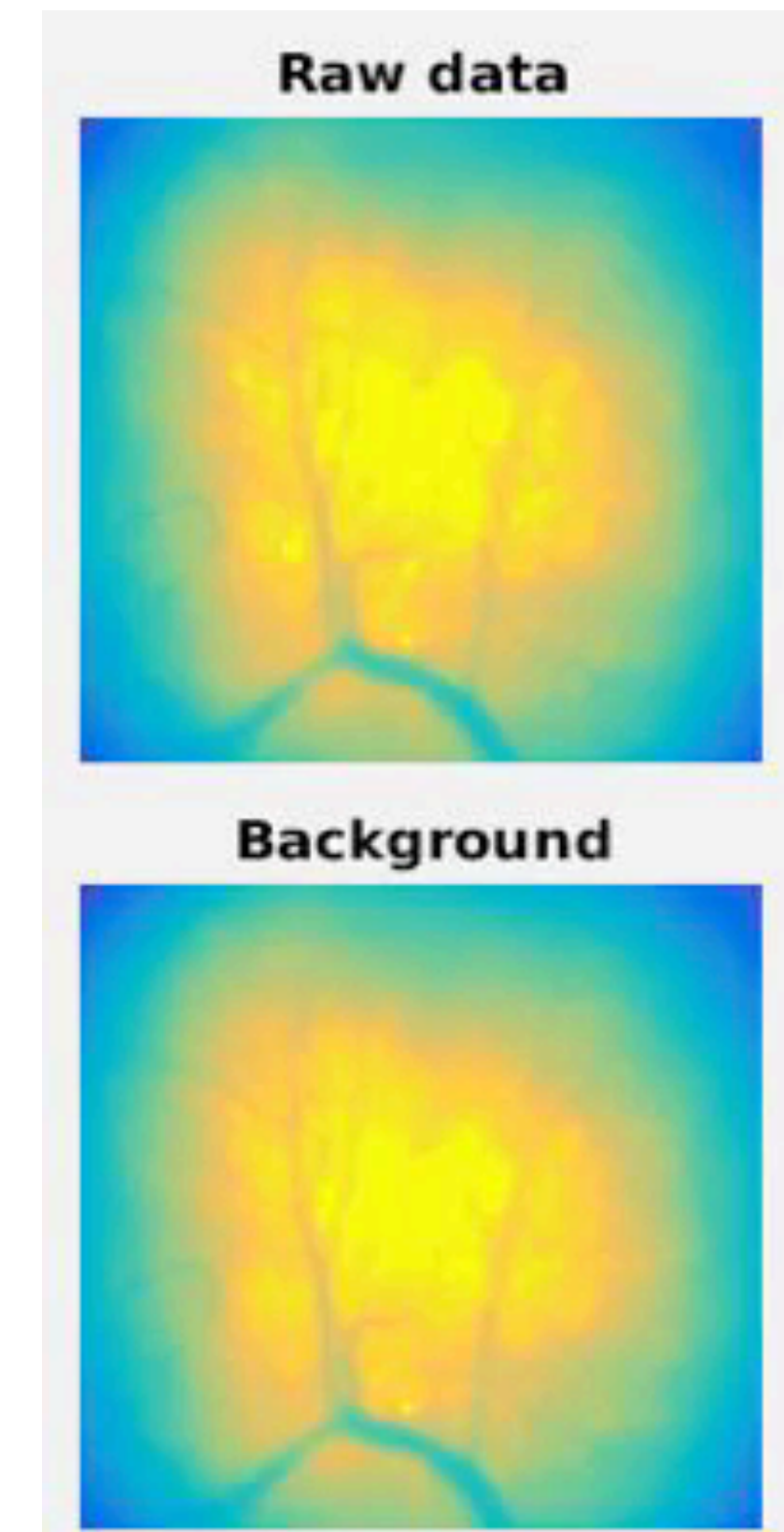
# Problem data.
m = 30
n = 20
np.random.seed(1)
A = np.random.randn(m, n)
b = np.random.randn(m)

# Construct the problem.
x = cp.Variable(n)
objective = cp.Minimize(cp.sum_squares(A @ x - b))
constraints = [0 <= x, x <= 1]
prob = cp.Problem(objective, constraints)

# The optimal objective value is returned by `prob.solve()`.
result = prob.solve()
# The optimal value for x is stored in `x.value`.
print(x.value)
# The optimal Lagrange multiplier for a constraint is stored in
# `constraint.dual_value`.
print(constraints[0].dual_value)
```

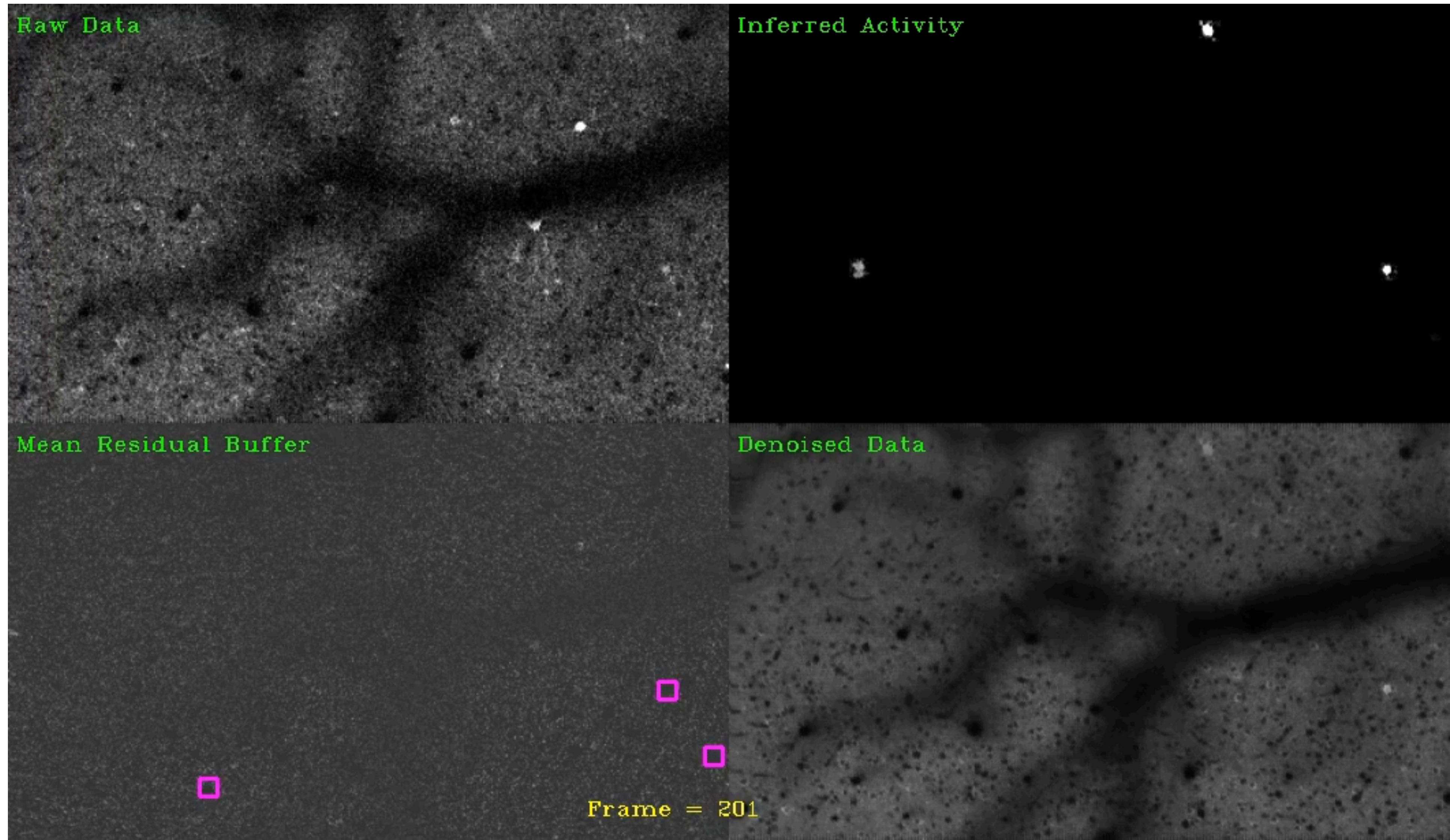
# Miscellanea

- We typically constrain the **spatial factors to be non-negative** too, unlike in spike sorting.
- We need to account for **background fluorescence** from out-of-focus cells.
- Typically, assume **rank-1** or **spatially smooth** background. See notes.
- As always, **preprocessing is important** for finding candidate neurons and characterizing noise. More on this in the lab.

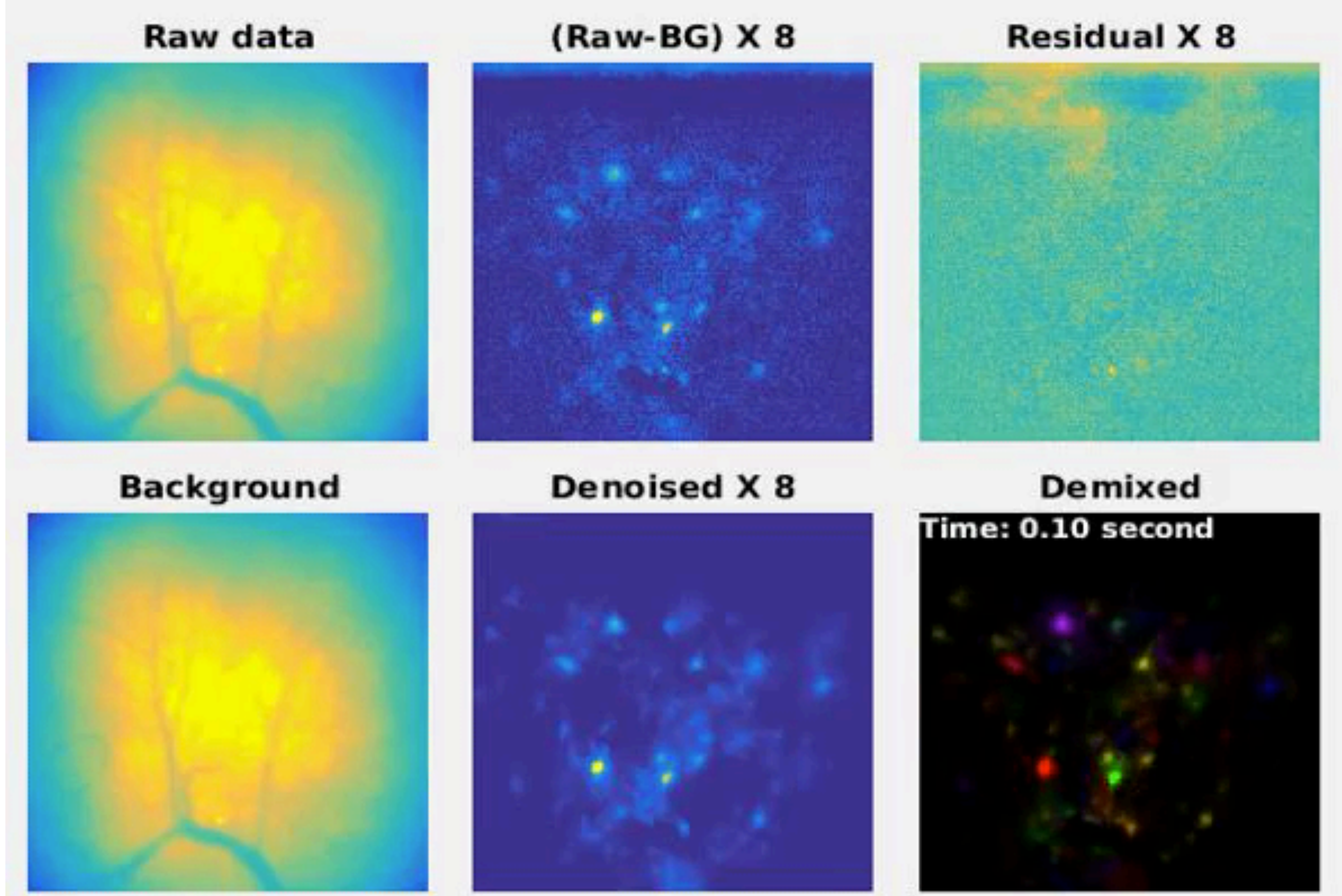


CNMF-E; Zhou et al, eLife 20:





OnACID; Giovanucci et al, NIPS 2017. Mesoscope data from A. Tolias lab



CNMF-E; Zhou et al, eLife 2018. Very different background model required for 1p data

# Conclusion

- **Optical physiology** offers a powerful and complementary toolkit for measuring neural activity in genetically defined cells.
- Methods for extracting calcium fluorescence traces are very similar to those for spike sorting. It's all **convolutional matrix factorization with constraints**.
- If we have an estimate of the noise, we can use it to **set hyper parameters (i.e. thresholds) automatically**.
- **Next time:** we'll dive deeper into the deconvolution problem of inferring spike times and amplitudes from calcium traces.

# Further reading

- Lin, Michael Z., and Mark J. Schnitzer. 2016. “Genetically Encoded Indicators of Neuronal Activity.” *Nature Neuroscience* 19 (9): 1142–53.
- Pnevmatikakis EA, Soudry D, Gao Y, et al. Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron*. 2016;89(2):285-299. doi:10.1016/j.neuron.2015.11.037
- Pachitariu, Marius, Carsen Stringer, Mario Dipoppa, Sylvia Schröder, L. Federico Rossi, Henry Dalgleish, Matteo Carandini, and Kenneth D. Harris. 2017. “Suite2p: Beyond 10,000 Neurons with Standard Two-Photon Microscopy.” Cold Spring Harbor Laboratory. <https://doi.org/10.1101/061507>.
- Zhou, Pengcheng, Shanna L. Resendez, Jose Rodriguez-Romaguera, Jessica C. Jimenez, Shay Q. Neufeld, Andrea Giovannucci, Johannes Friedrich, et al. 2018. “Efficient and Accurate Extraction of in Vivo Calcium Signals from Microendoscopic Video Data.” *eLife* 7 (February): e28728.
- Giovannucci, Andrea, Johannes Friedrich, Pat Gunn, Jérémie Kalfon, Brandon L. Brown, Sue Ann Koay, Jiannis Taxidis, et al. 2019. “CalmAn an Open Source Tool for Scalable Calcium Imaging Data Analysis.” *eLife* 8 (January). <https://doi.org/10.7554/eLife.38173>.