Machine Learning Methods for Neural Data Analysis

Demixing and Deconvolving Calcium Imaging Data

Scott Linderman

STATS 220/320 (NBIO220, CS339N).

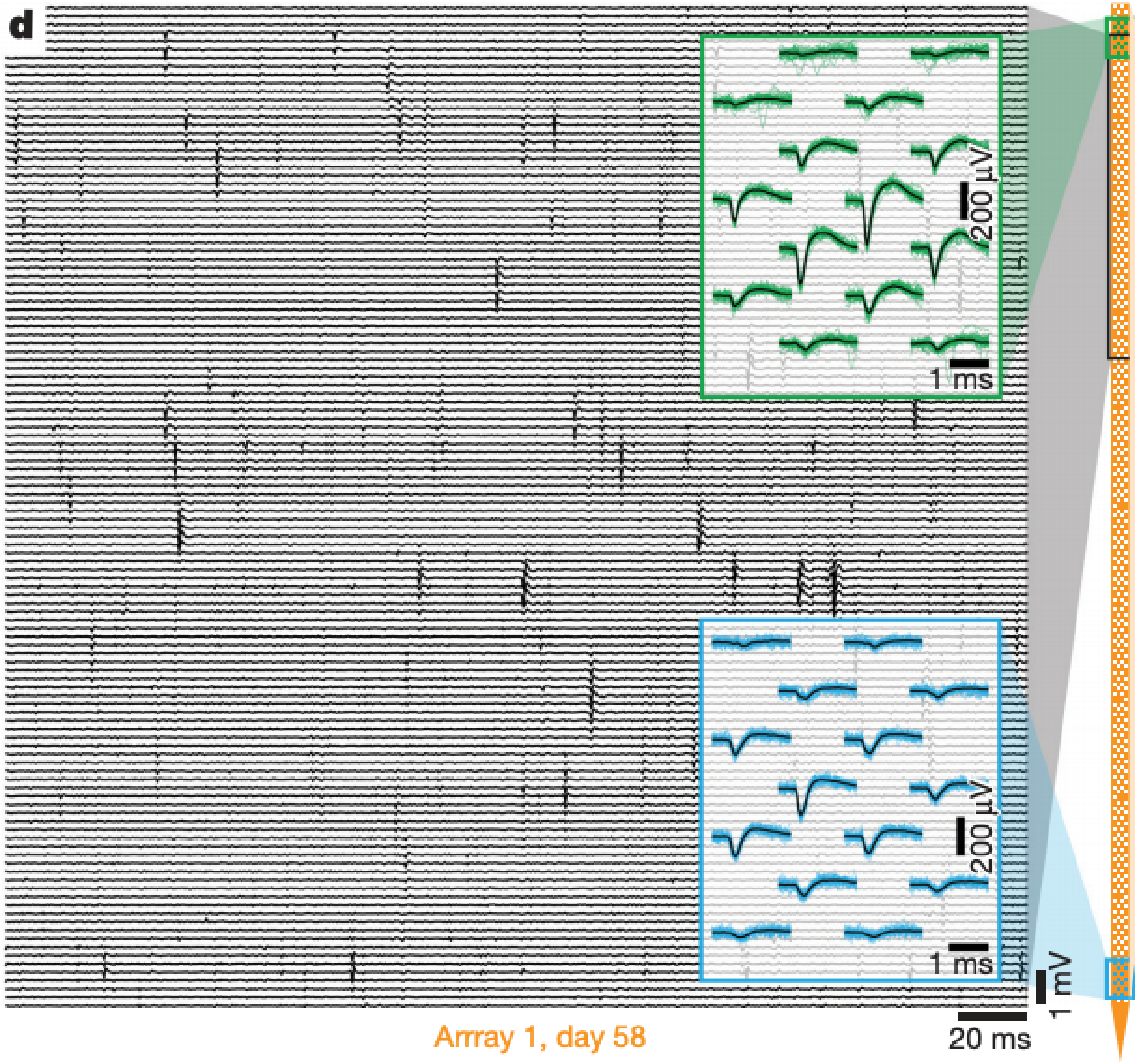
Agenda

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

Recap

Electrophysiology

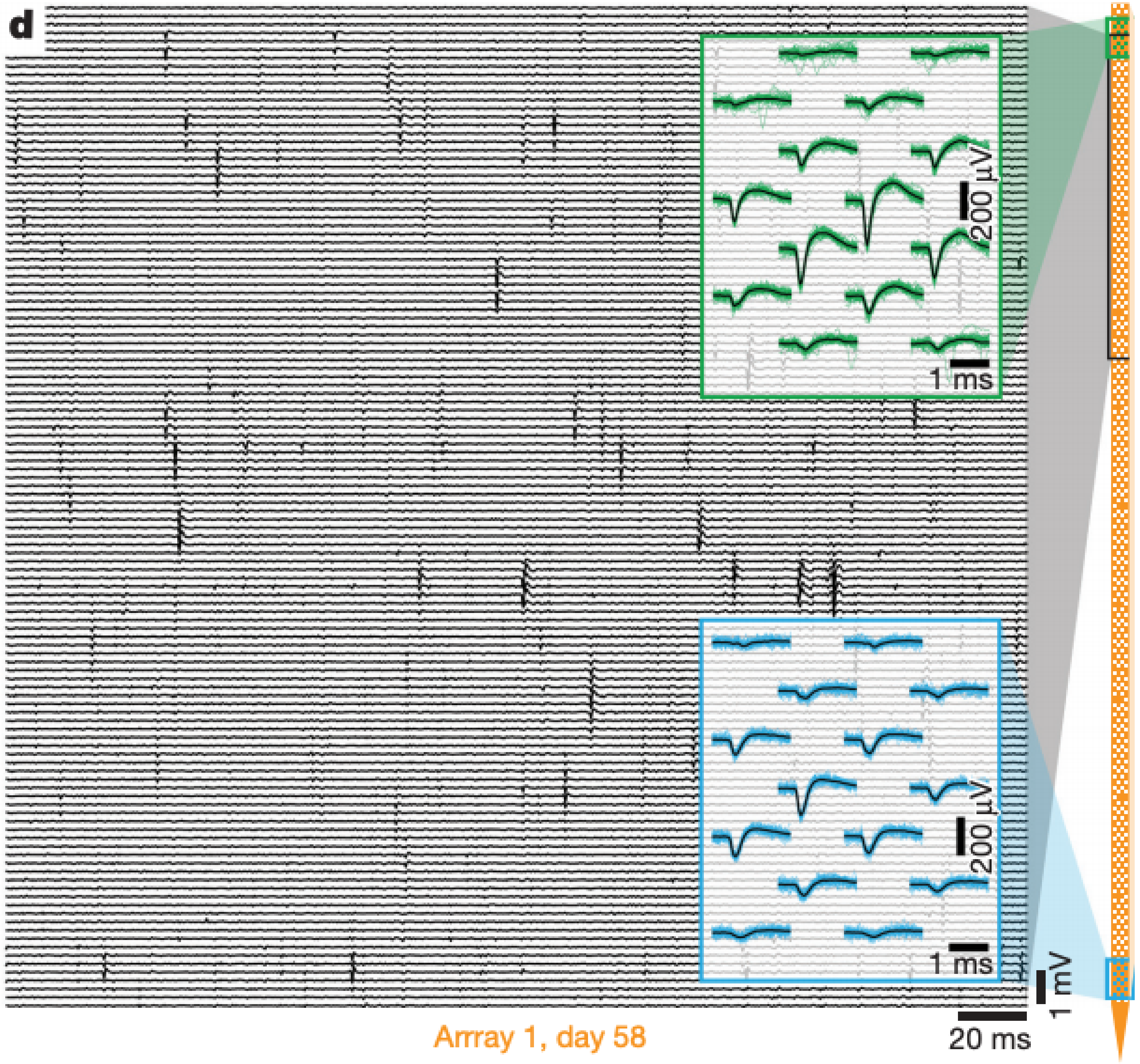
- So far, we've study electrophysiological ("ephys") recordings with tetrodes and high density probes.
- The raw data is a **multidimensional time series** of **voltage measurements**, one for each recording site on the probe.
- When neurons near the probe fire an **action potential**, it registers a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



Recap

Electrophysiology Limitations

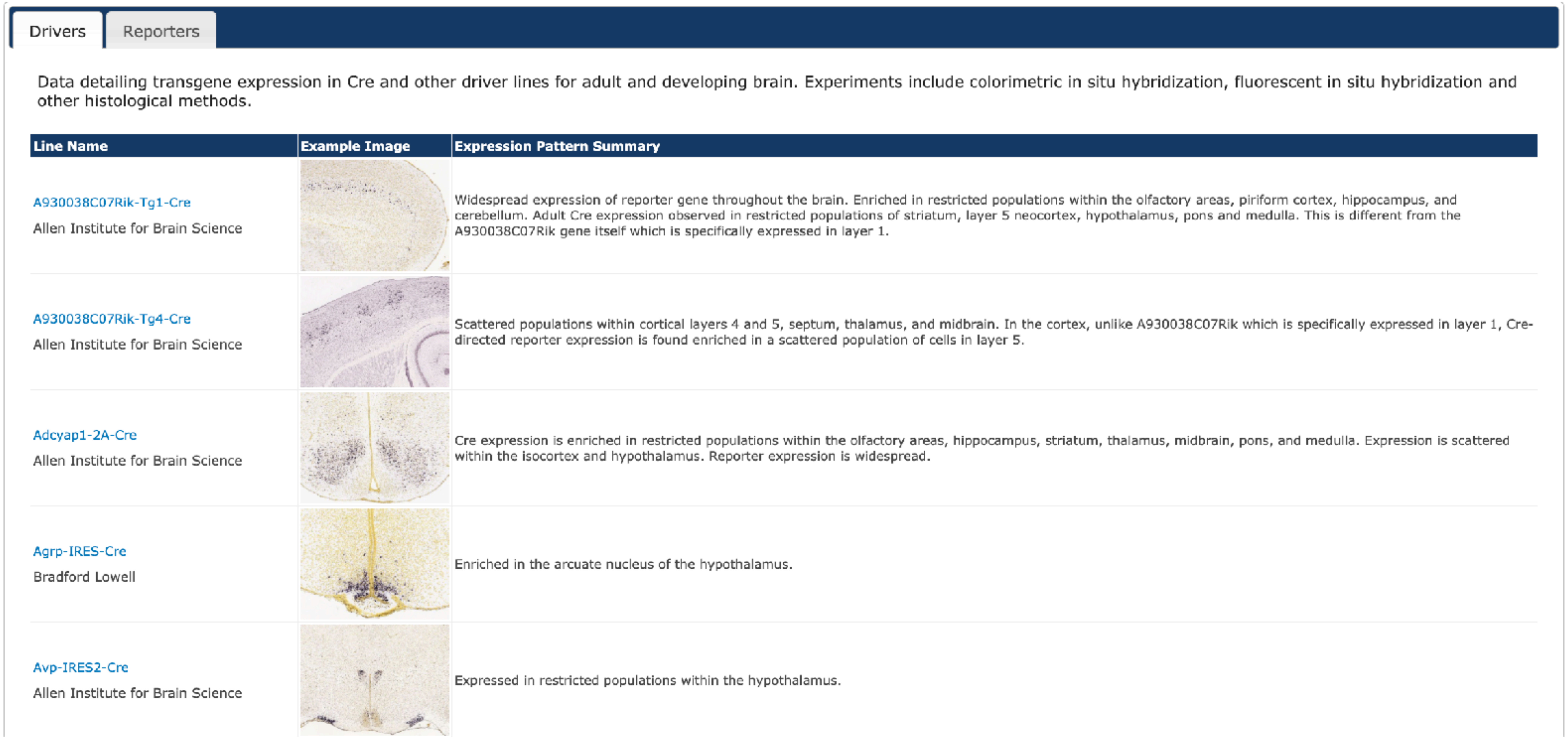
- It's hard to detect neurons that fire rarely and produce low amplitude EAPs.
- More generally, you only detect cells that happen to be close to the narrow probe.
- No cell-type specificity.
- In particular, ephys does not leverage our powerful genetic toolkits for certain model organisms.



Arrray 1, day 58

Genetic tools

Cre driver lines in mice



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.

Line Name	Example Image	Expression Pattern Summary
A930038C07Rik-Tg1-Cre Allen Institute for Brain Science		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.
A930038C07Rik-Tg4-Cre Allen Institute for Brain Science		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.
Adcyap1-2A-Cre Allen Institute for Brain Science		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.
Agrp-IRES-Cre Bradford Lowell		Enriched in the arcuate nucleus of the hypothalamus.
Avp-IRES2-Cre Allen Institute for Brain Science		Expressed in restricted populations within the hypothalamus.

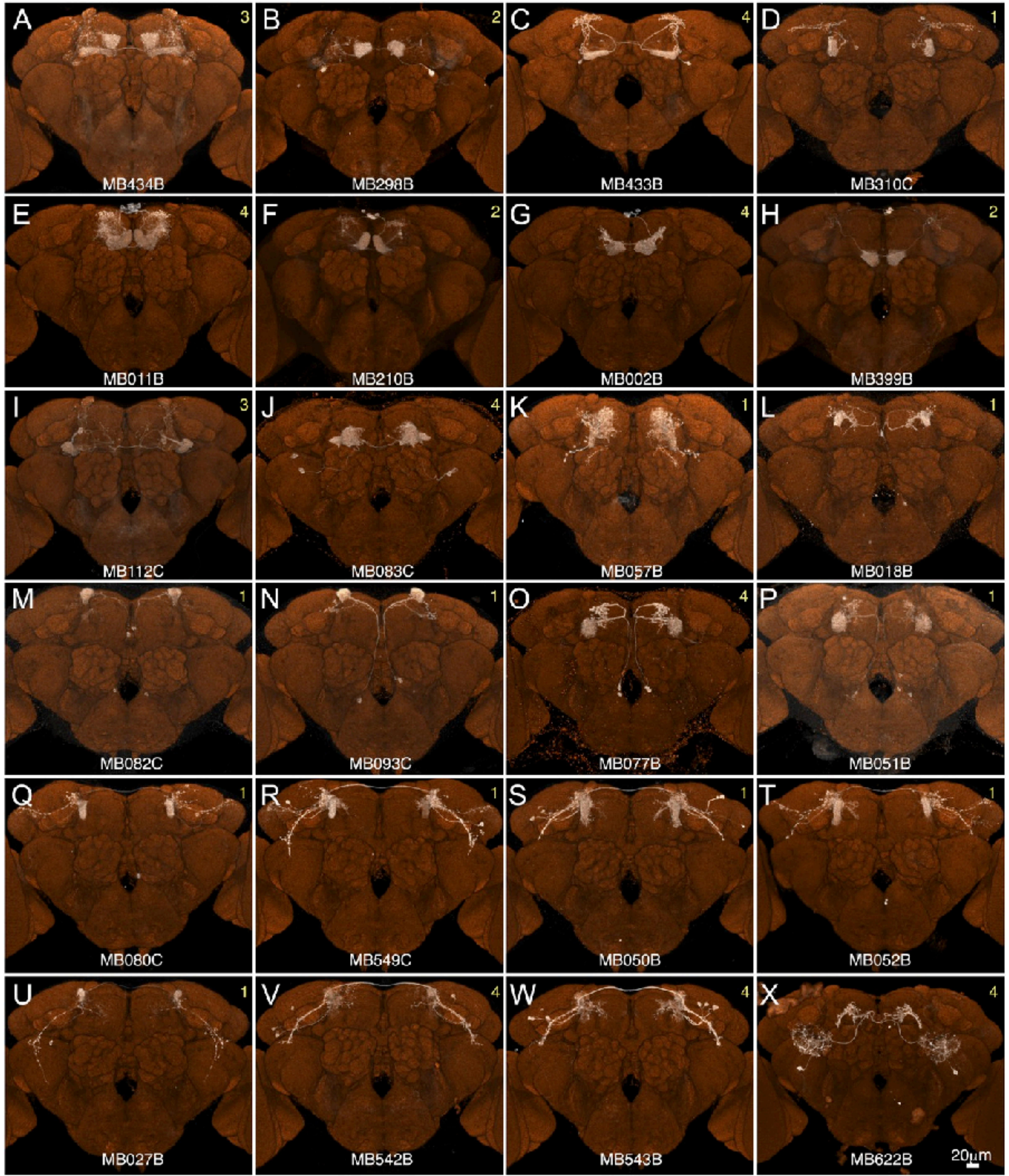
https://connectivity.brain-map.org/transgenic

Genetic tools

GAL4 lines in flies

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

Split-GAL4 lines for MBONs



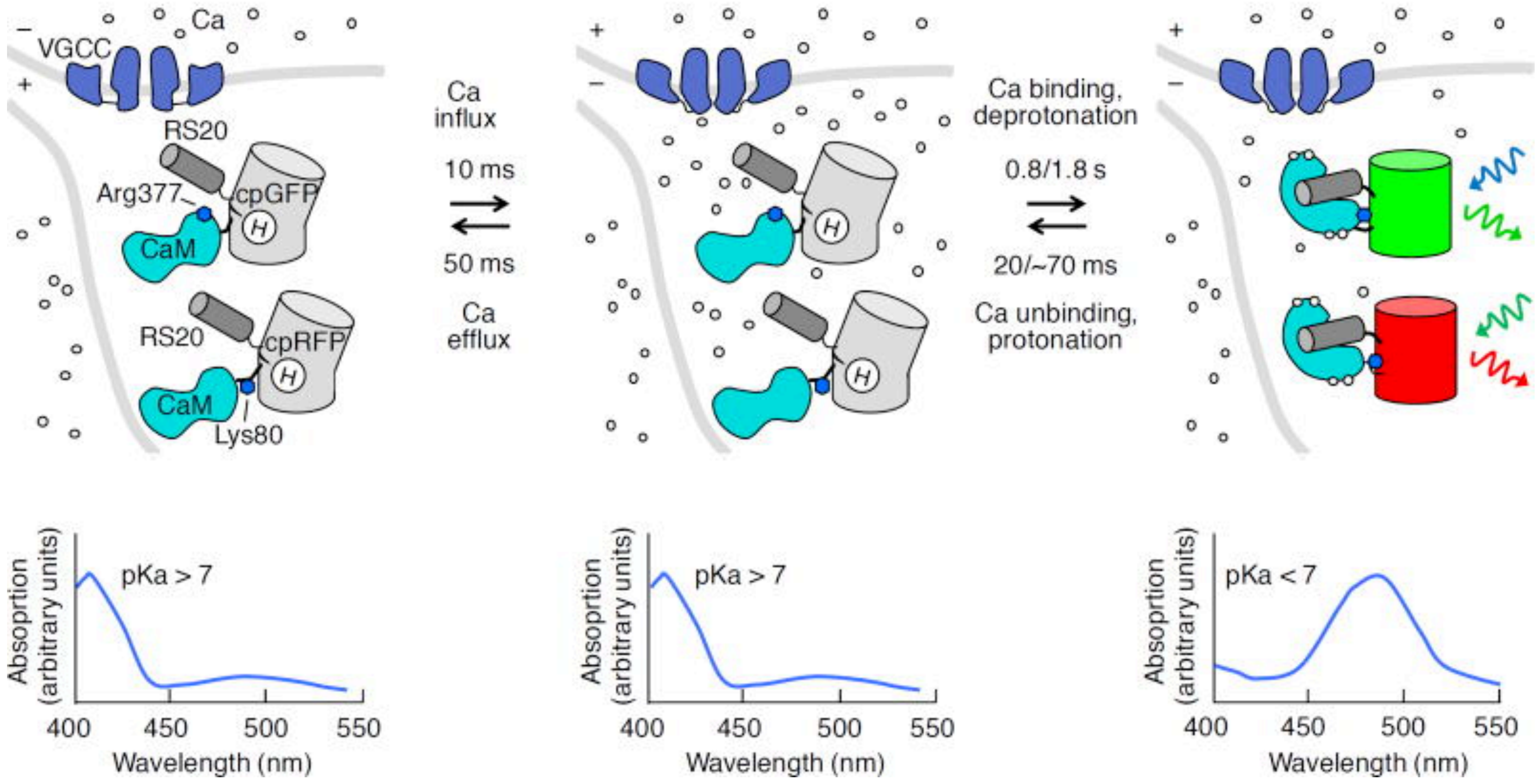
Genetically encoded indicators of neural activity

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)

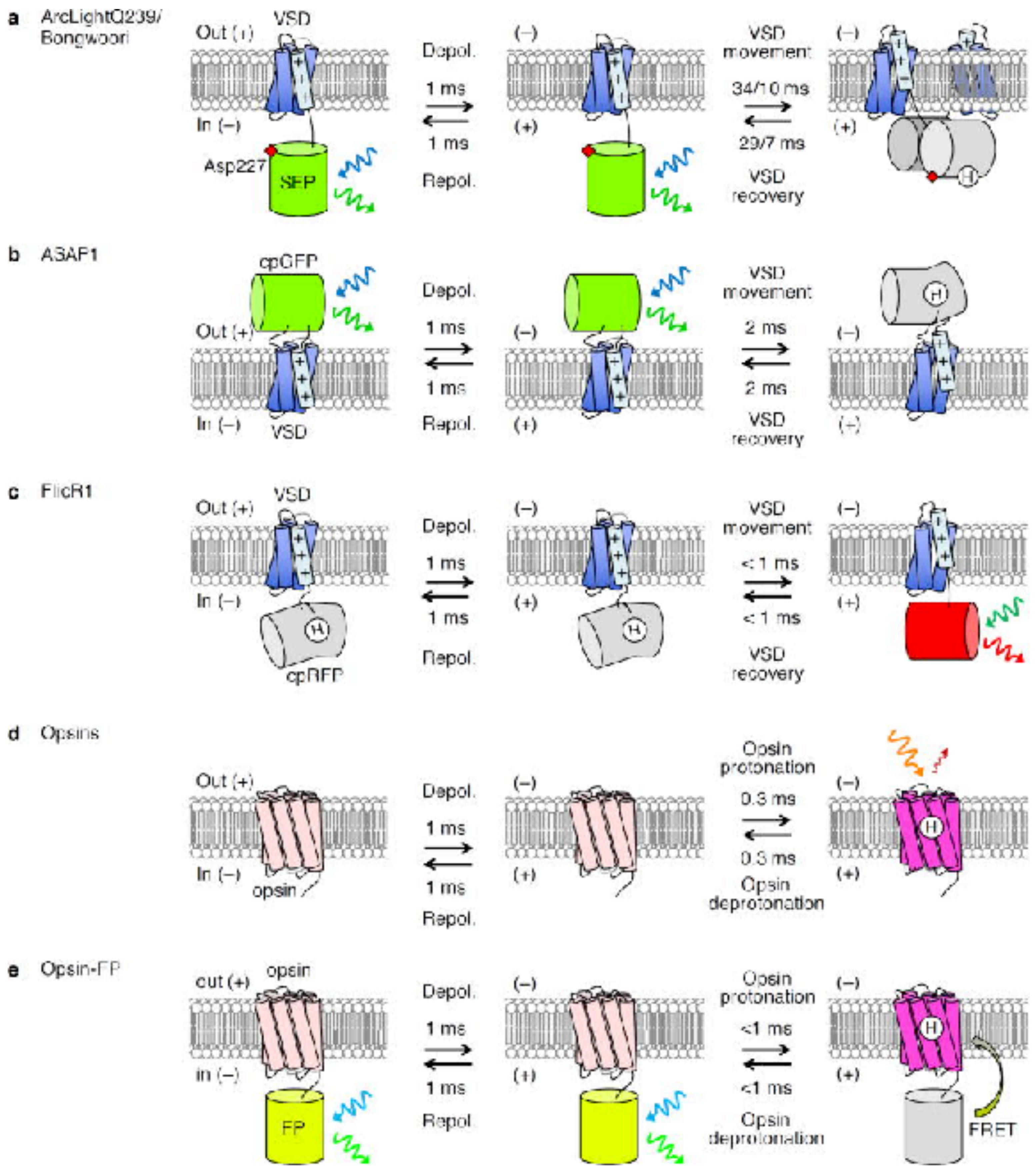
- When neurons spike, voltage gated calcium channels (VGCCs) open and allow a rapid **influx of calcium ions** (Ca^{2+}).
- Genetically encoded calcium indicators (GECIs) like **GCaMP** bind to these calcium ions and become fluorescent.
- The increased fluorescence decays as the calcium unbinds, producing 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.**



Lin, Michael Z., and Mark J. Schnitzer. 2016. “Genetically Encoded Indicators of Neuronal Activity.” *Nature Neuroscience* 19 (9): 1142–53.

Genetically encoded voltage indicators (GEVIs)

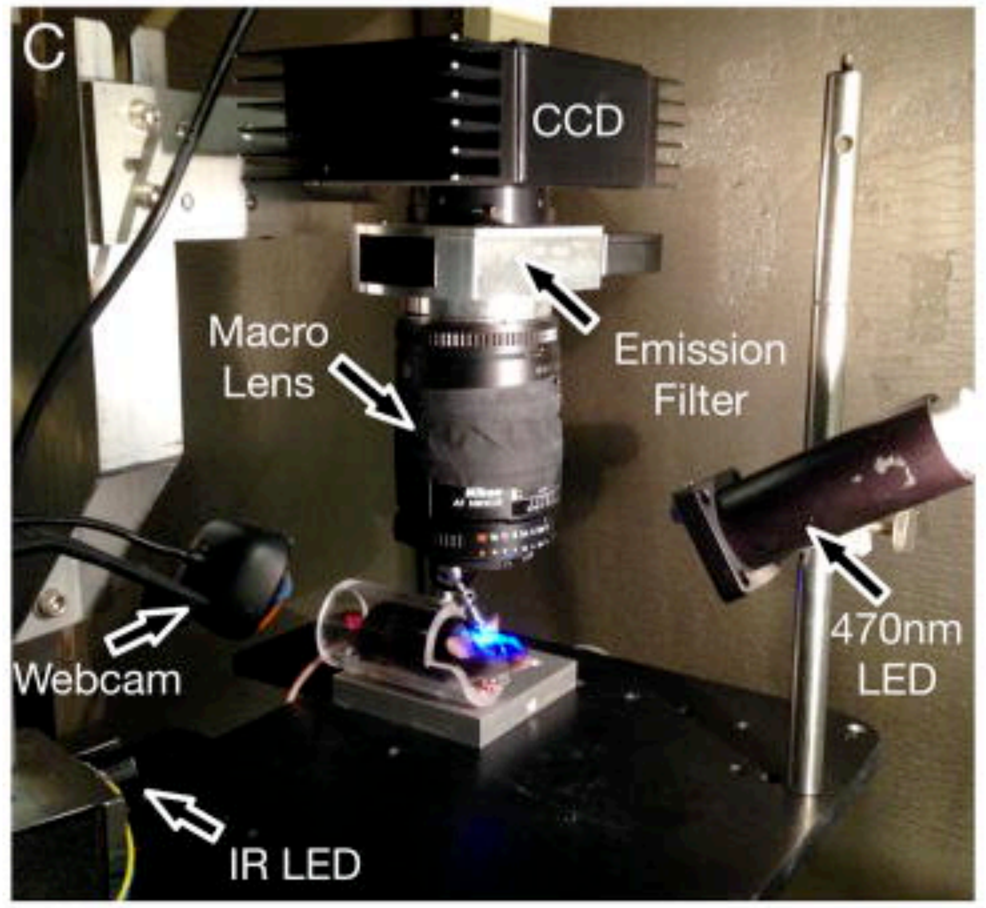
- Calcium is an indirect measure of spiking. Genetically encoded **voltage indicators** modulate fluorescence as a function of membrane potential.
- **Lots of designs**: fusing voltage sensing domains (e.g. from voltage-gated ion channels) to fluorescent proteins; harnessing natural opsins from microbes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



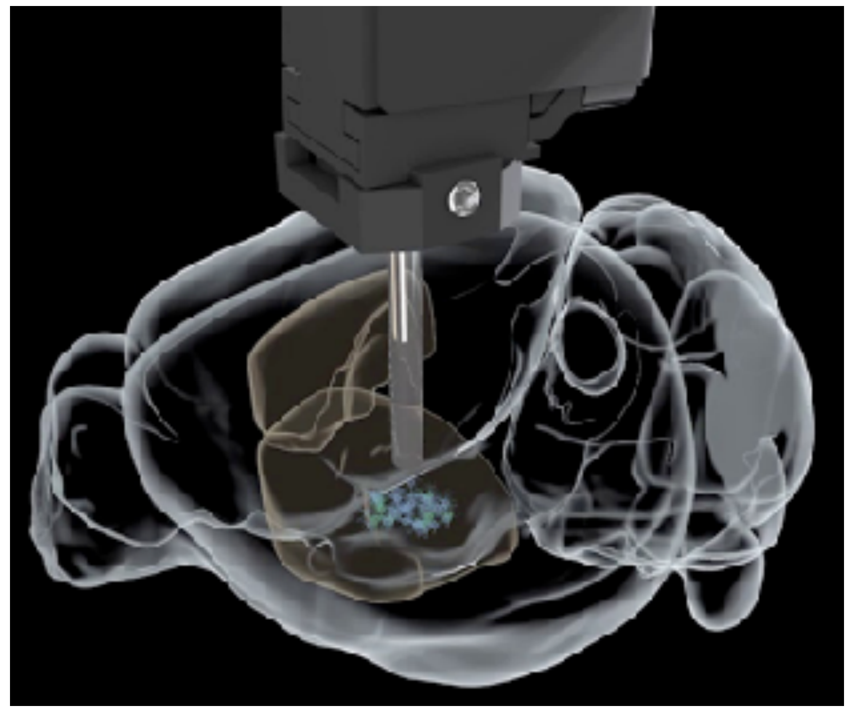
Lin, Michael Z., and Mark J. Schnitzer. 2016. “Genetically Encoded Indicators of Neuronal Activity.” *Nature Neuroscience* 19 (9): 1142–53.

Microscopy

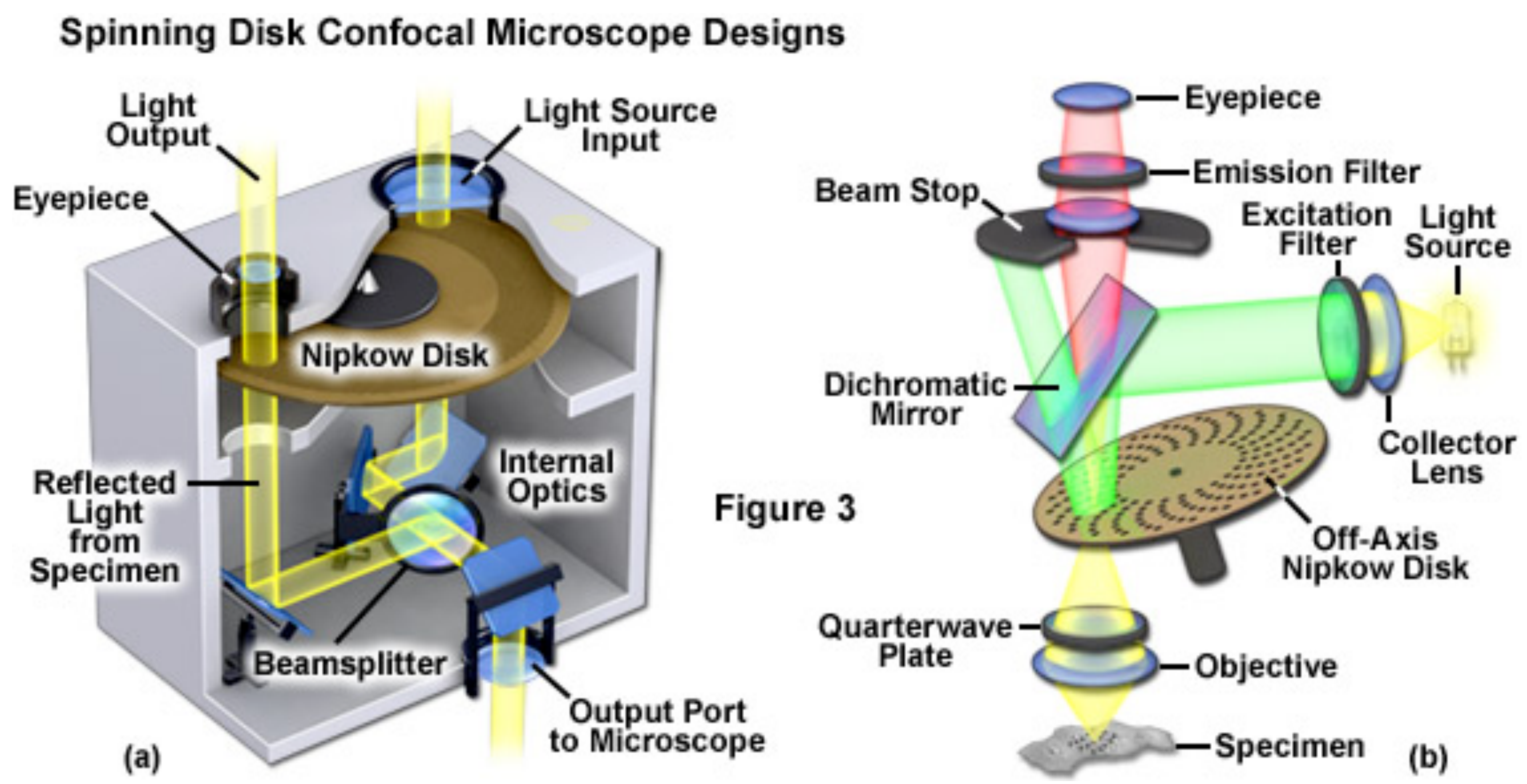
- Expressing the genetically encoded indicator is only half the battle.
- You still need to stimulate the cells with a light source and measure the resulting fluorescence.
- Again, there are lots of approaches: wide-field imaging, **2-photon microscopy**, laser scanning and spinning disk confocal microscopy,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



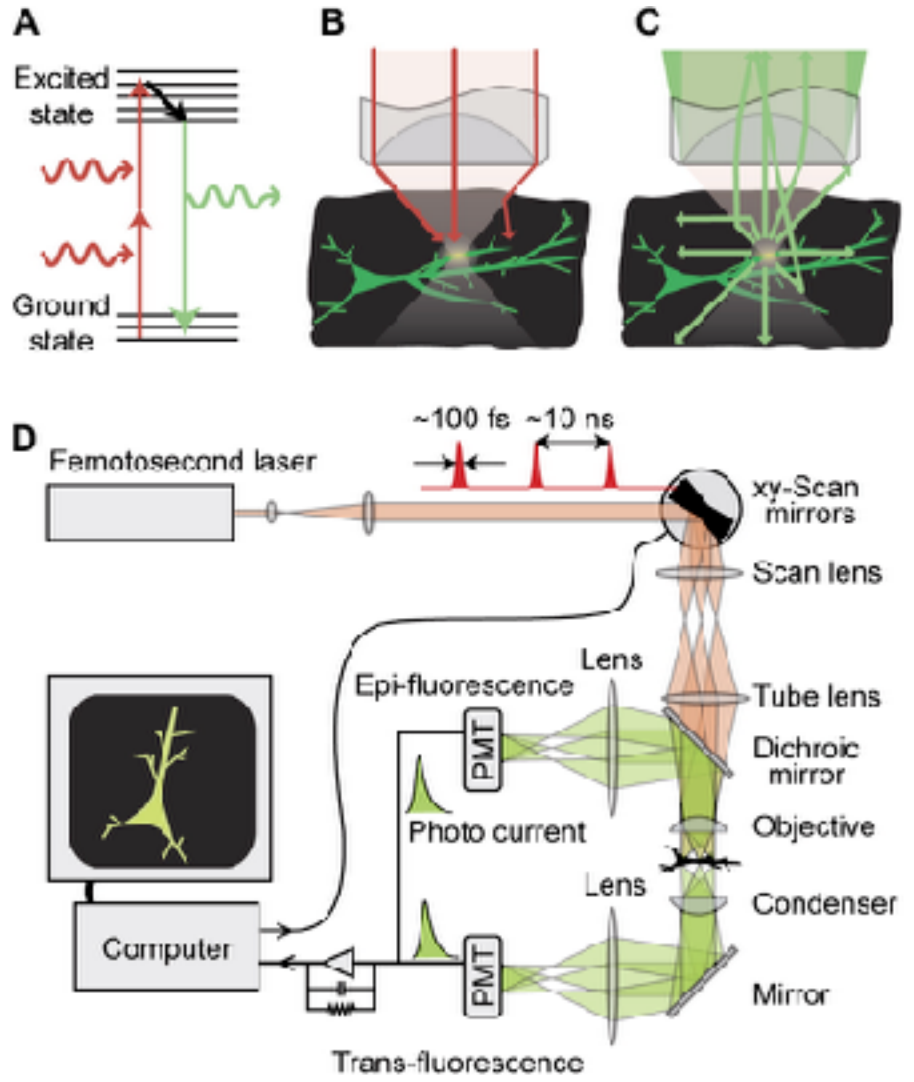
Silasi et al. J Neurosci Methods. 2016



www.inscopix.com

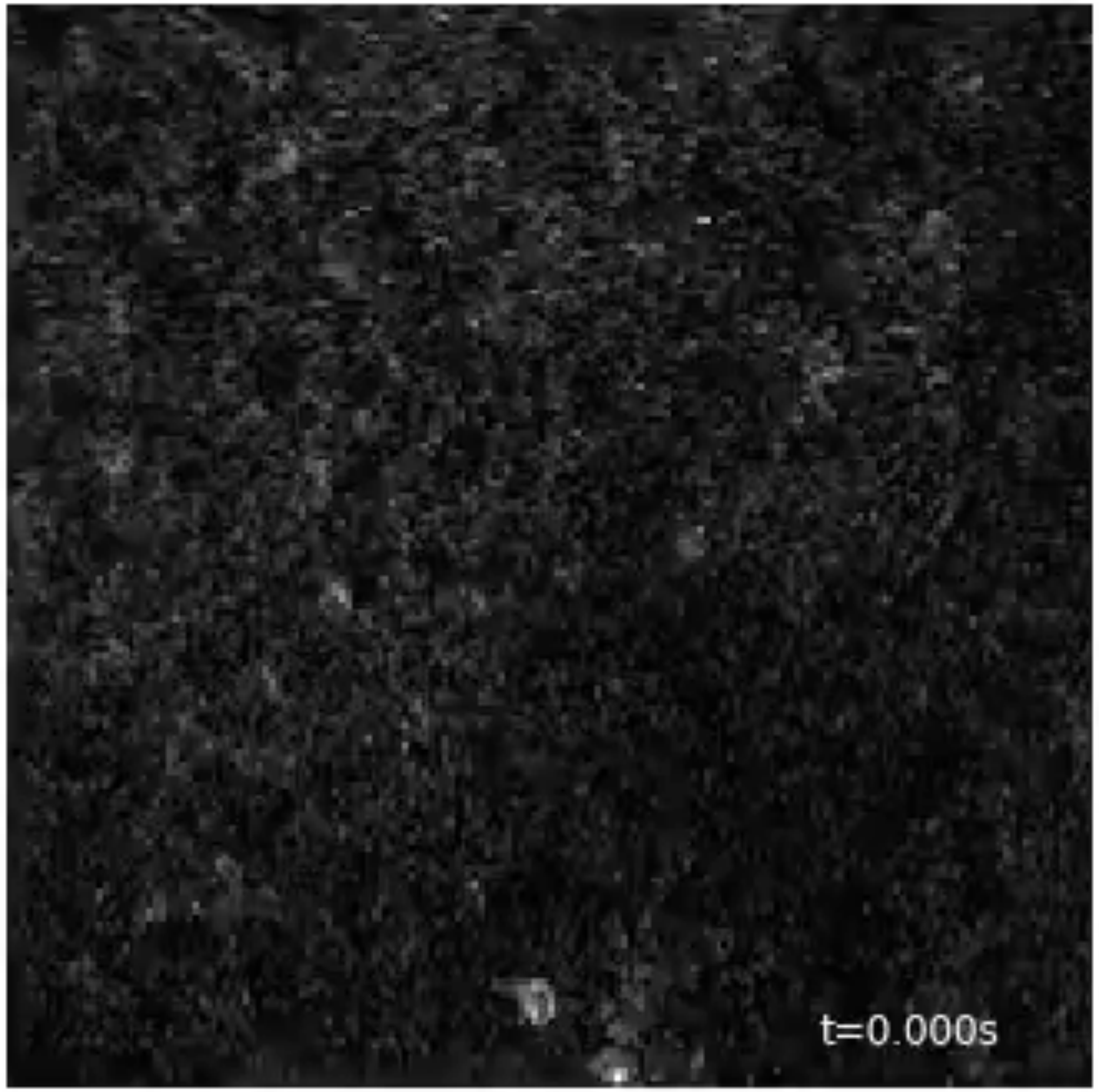


http://zeiss-campus.magnet.fsu.edu/articles/spinningdisk/introduction.html



Svoboda and Yasuda. Neuron, 2006.

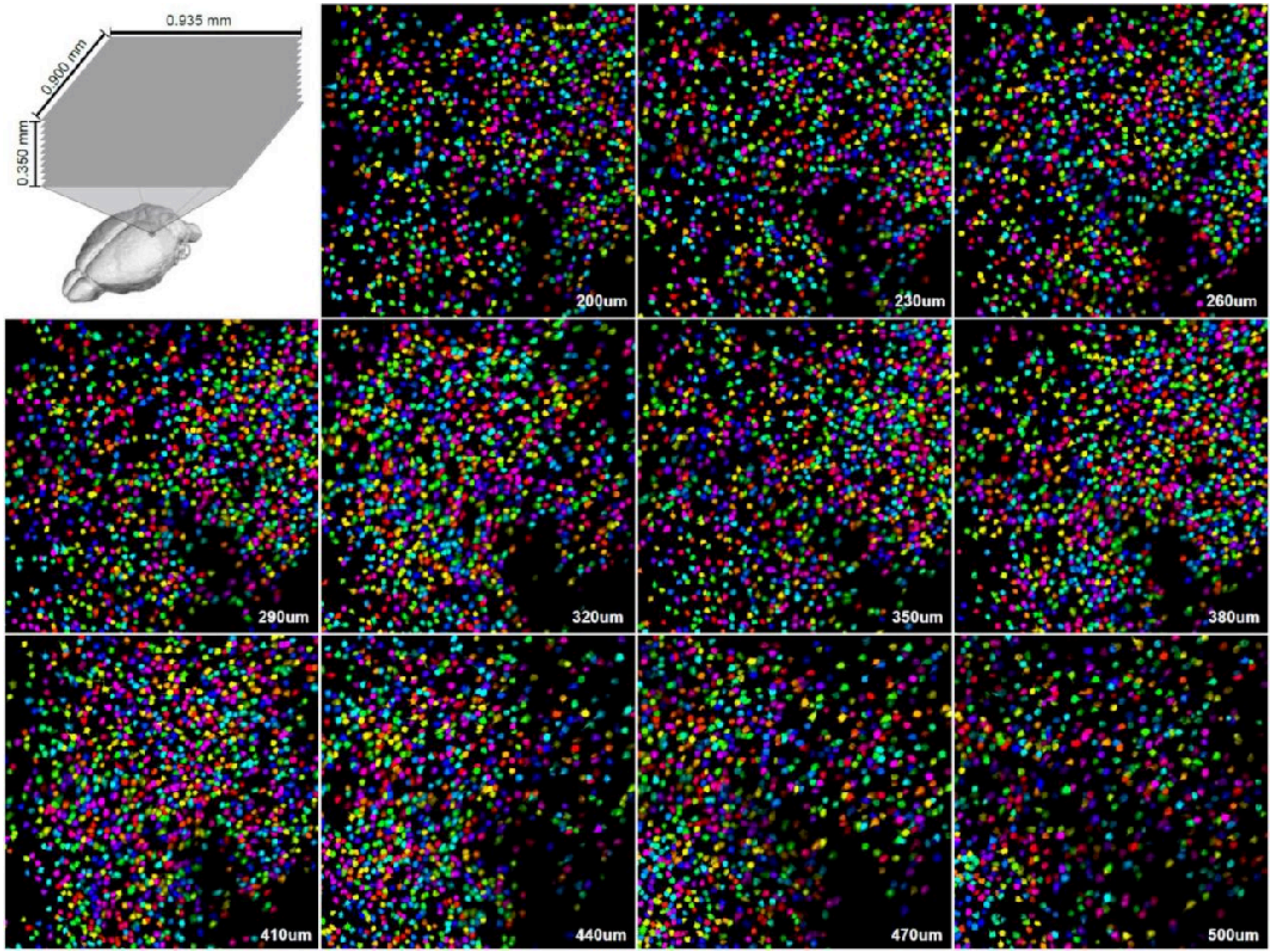
2 photon calcium imaging



Sue Ann Koay and David Tank

2 photon calcium imaging

Over 10,000 cells

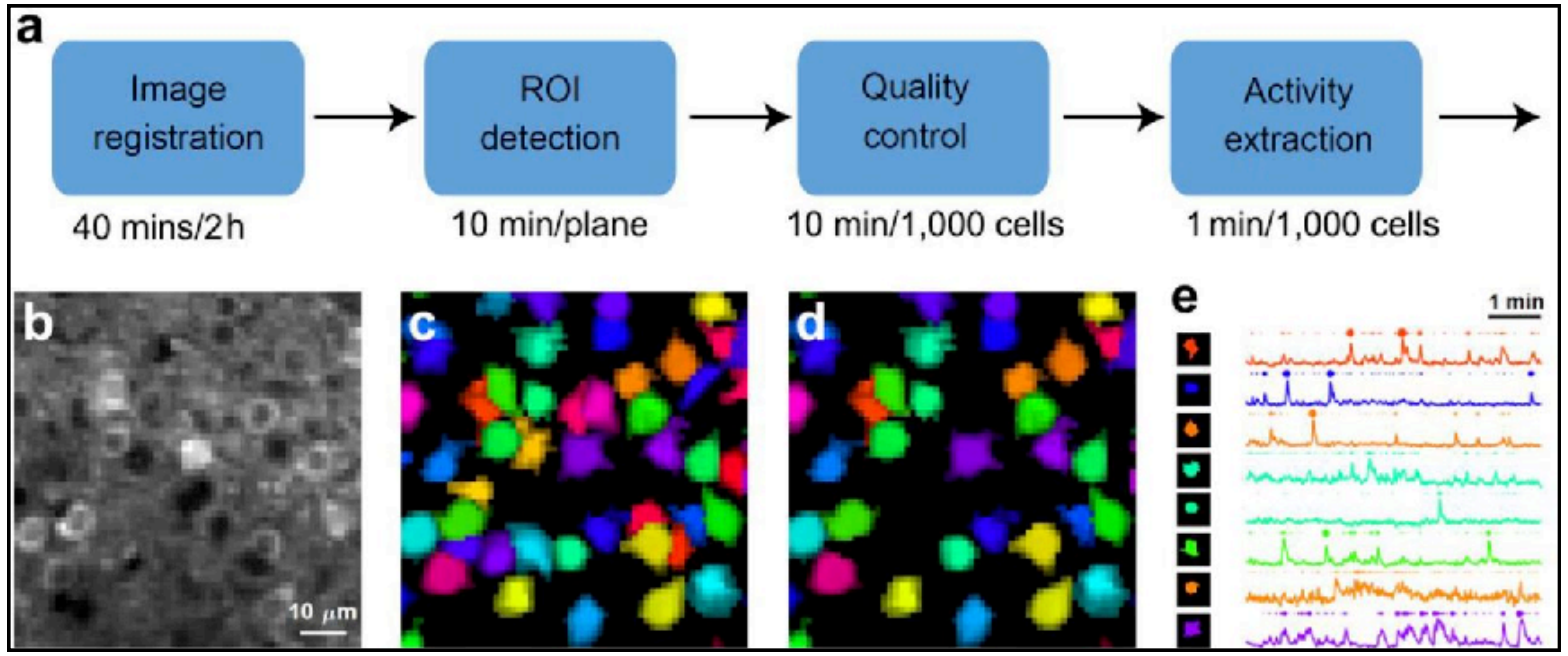


Pachitariu et al, bioRxiv 2017
Stringer et al, Nature 2019

Data analysis pipelines for 2P imaging

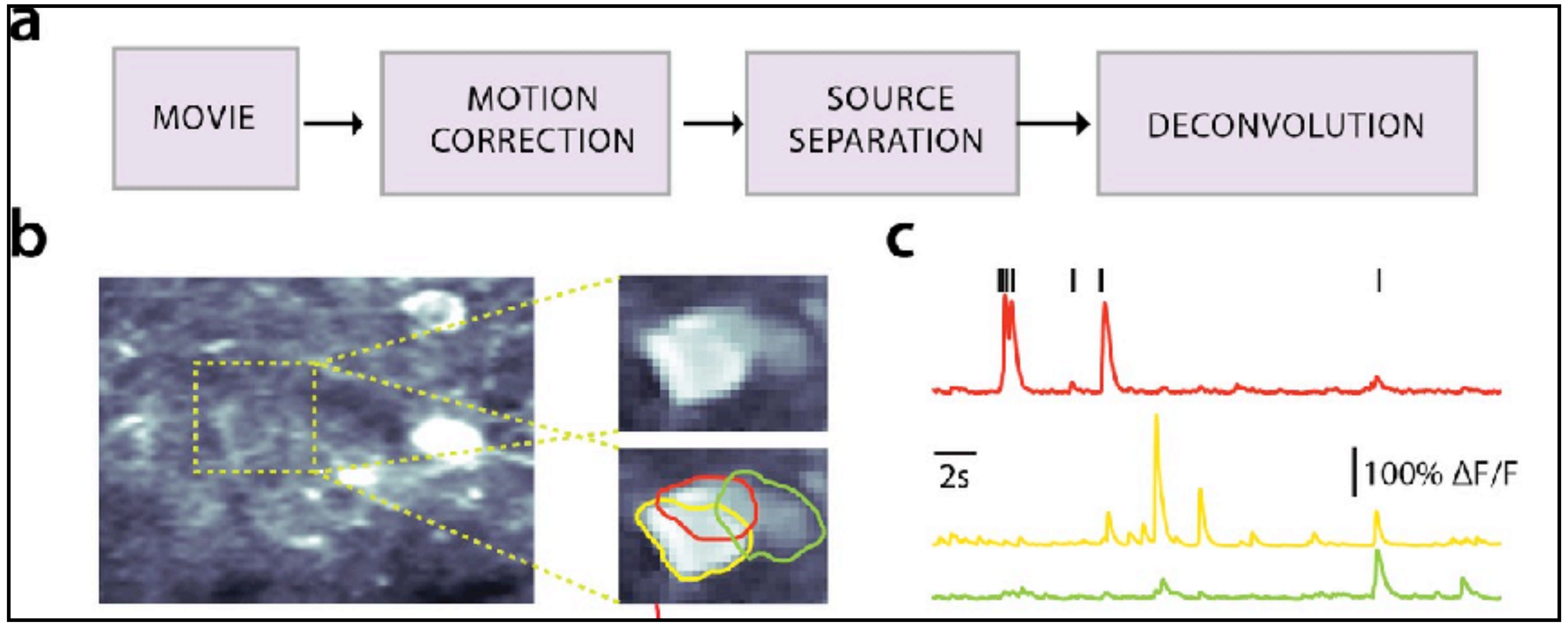
- Modern packages like Suite2P and CalmAn go through a few key steps to extract fluorescence traces.
- The key challenges are:
  - Correcting for motion artifacts.
  - Separating overlapping cells.
  - Accounting for background fluorescence.
  - Deconvolving spikes from fluorescence traces.

Suite2P



Pachitariu et al, bioRxiv 2017

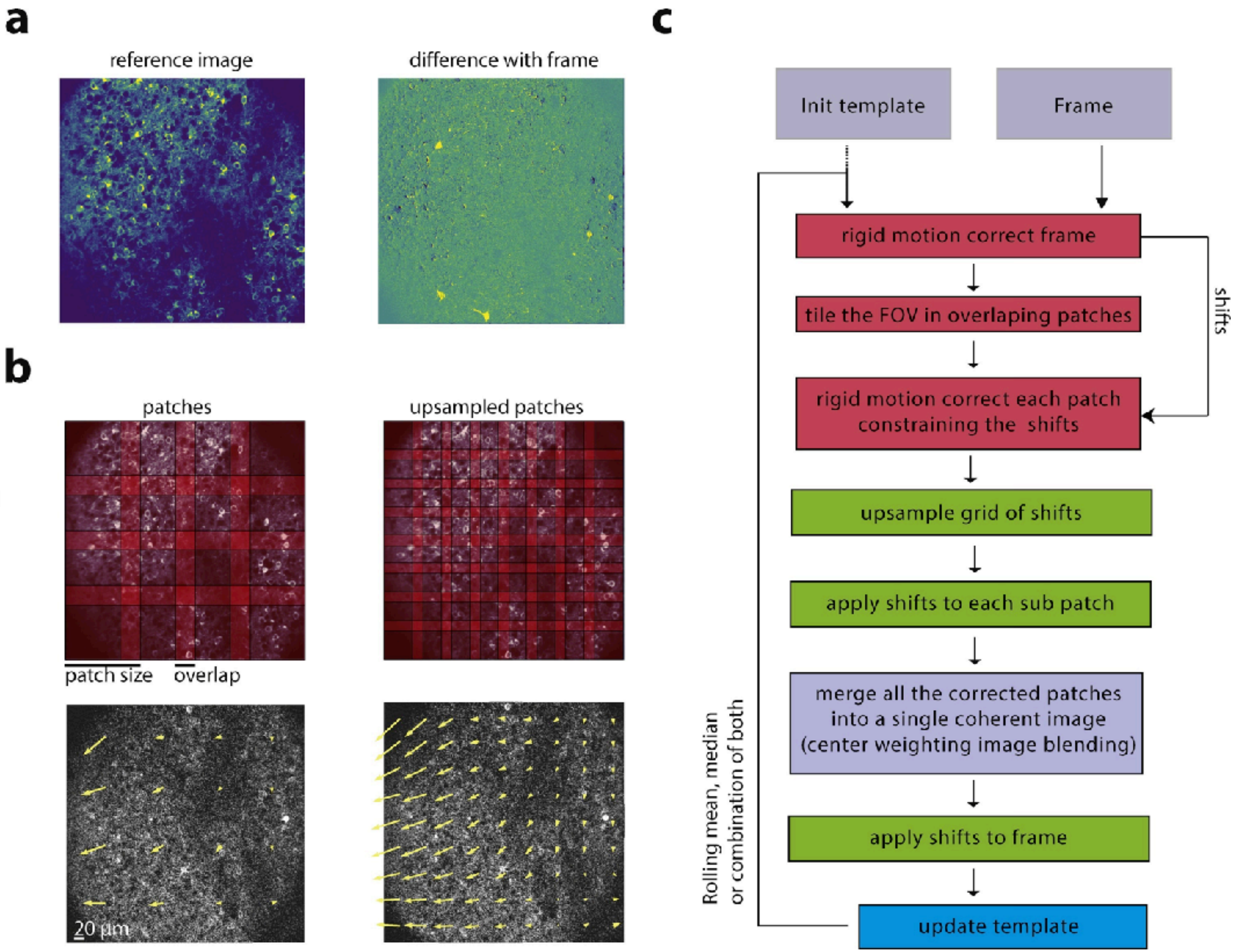
CalmAn



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 CaImAn.
  - Phase correlation + kriging (Pachitariu et al, 2017) used in Suite2P.



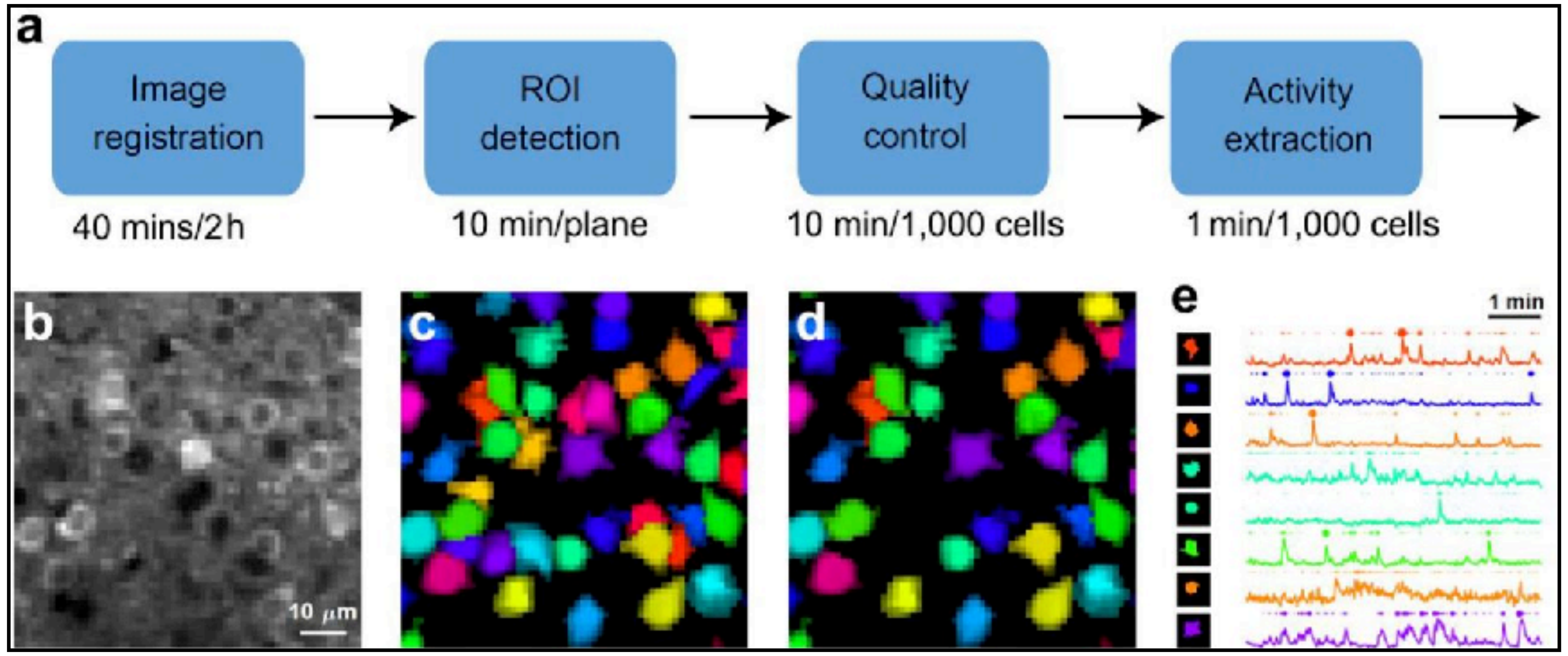
Pnevmatikakis and Giovannucci, 2017.

Data analysis pipelines for 2P imaging

- Modern packages like Suite2P and CaImAn 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.

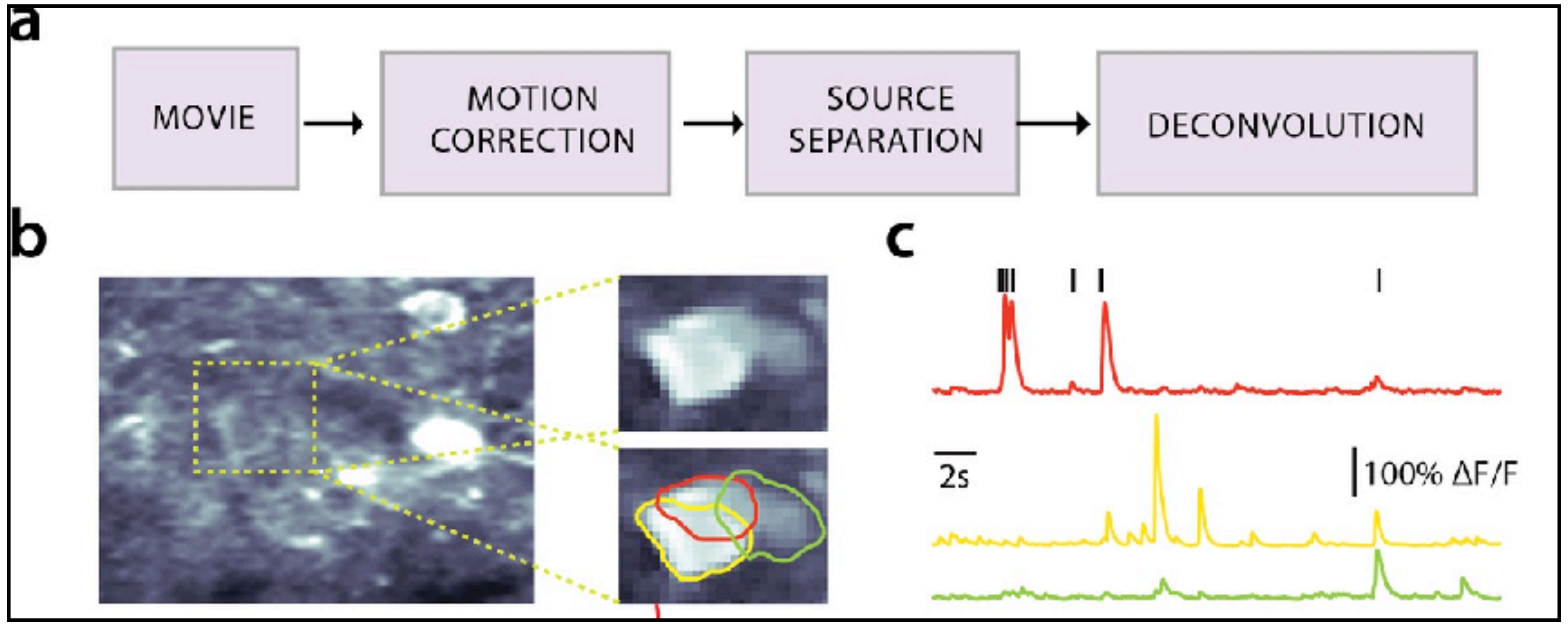
CNMF

Suite2P



Pachitariu et al, bioRxiv 2017

CaImAn



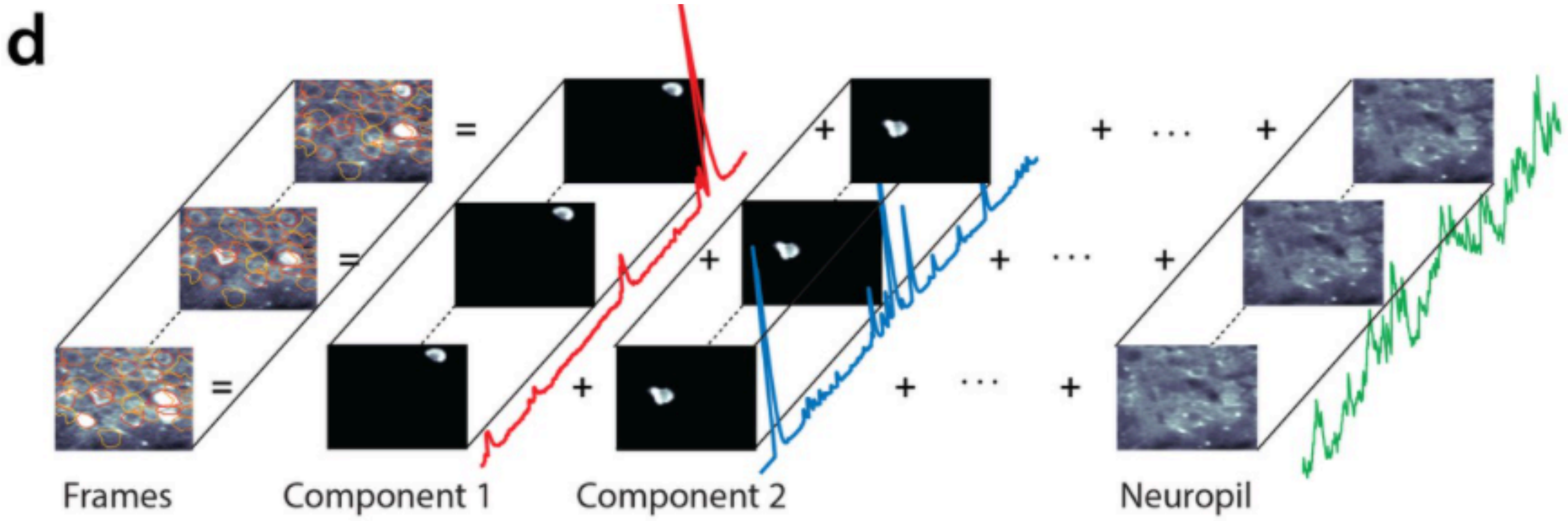
Giovanucci et al, eLife 2017

Constrained Non-negative Matrix Factorization (CNMF)

Pnevmatikakis et al, Neuron 2016.

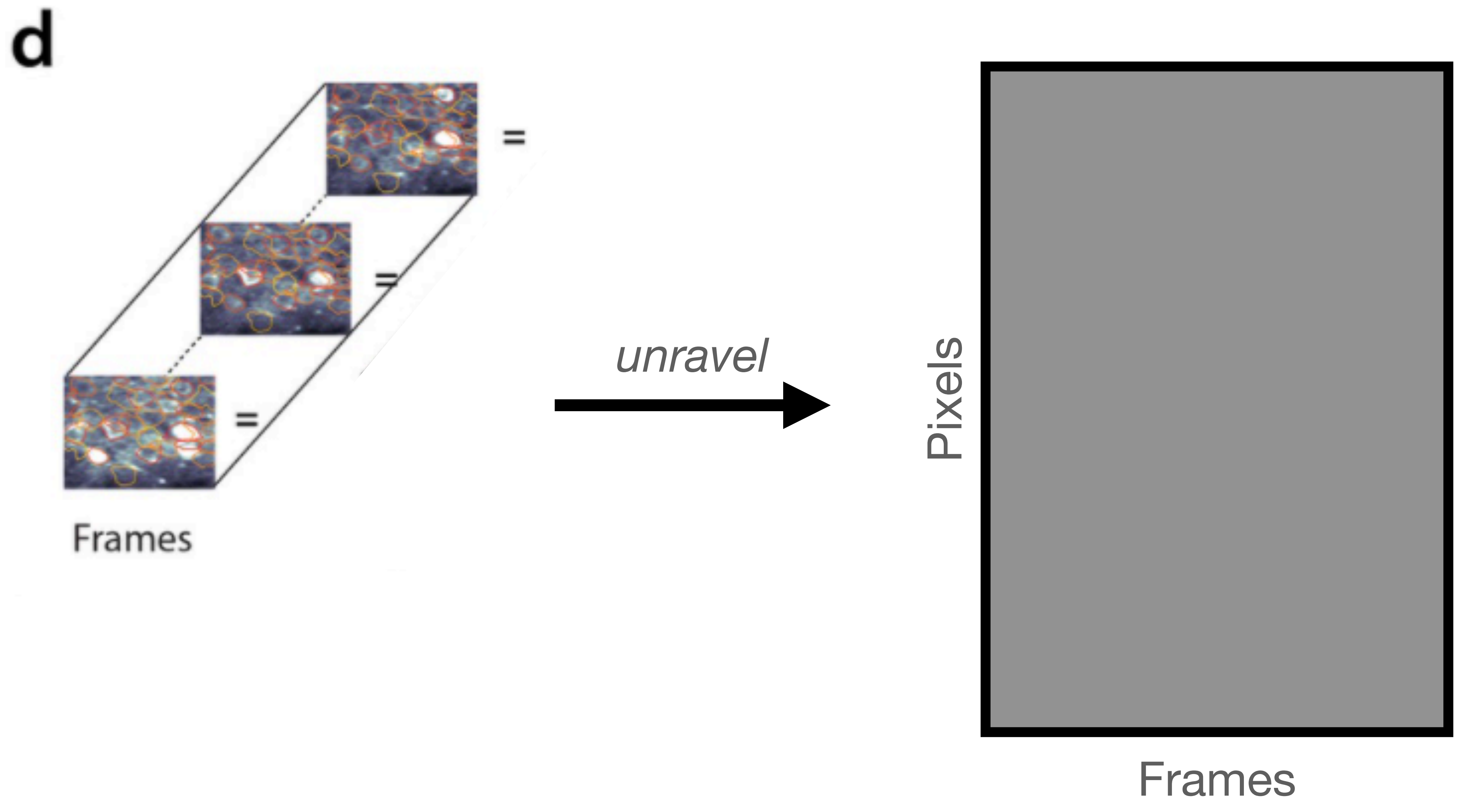
CNMF

- 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!*



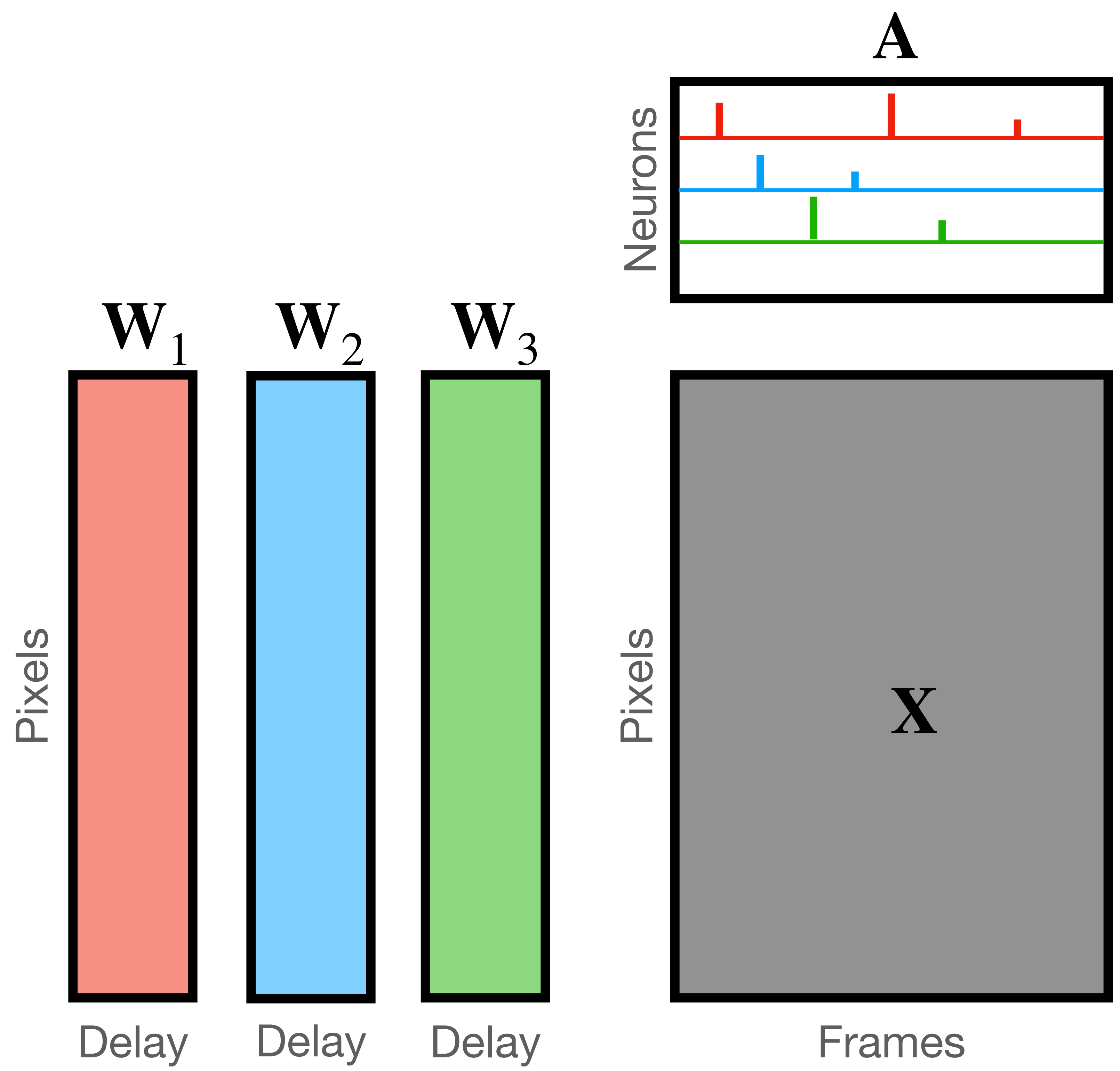
Constants

- 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 Model Parameters

- **Data:**
  - Let $\mathbf{X} \in \mathbb{R}^{N \times T}$ denote the motion corrected and unraveled video.
- **Parameters:**
  - Let $\mathbf{A} \in \mathbb{R}_+^{K \times T}$ denote the time series of spike amplitudes for each neuron.
  - Let $\mathbf{W} \in \mathbb{R}^{K \times N \times D}$ denote the array of calcium responses for each neuron.

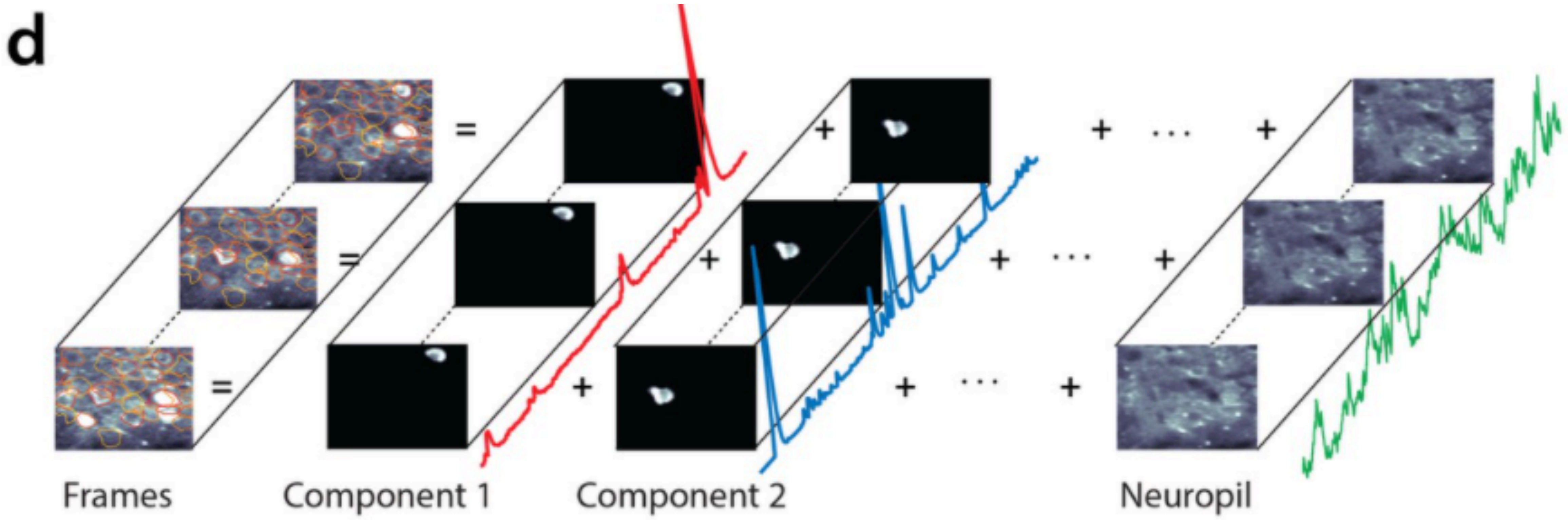


Probabilistic Model

Likelihood

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

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



Calcium response model

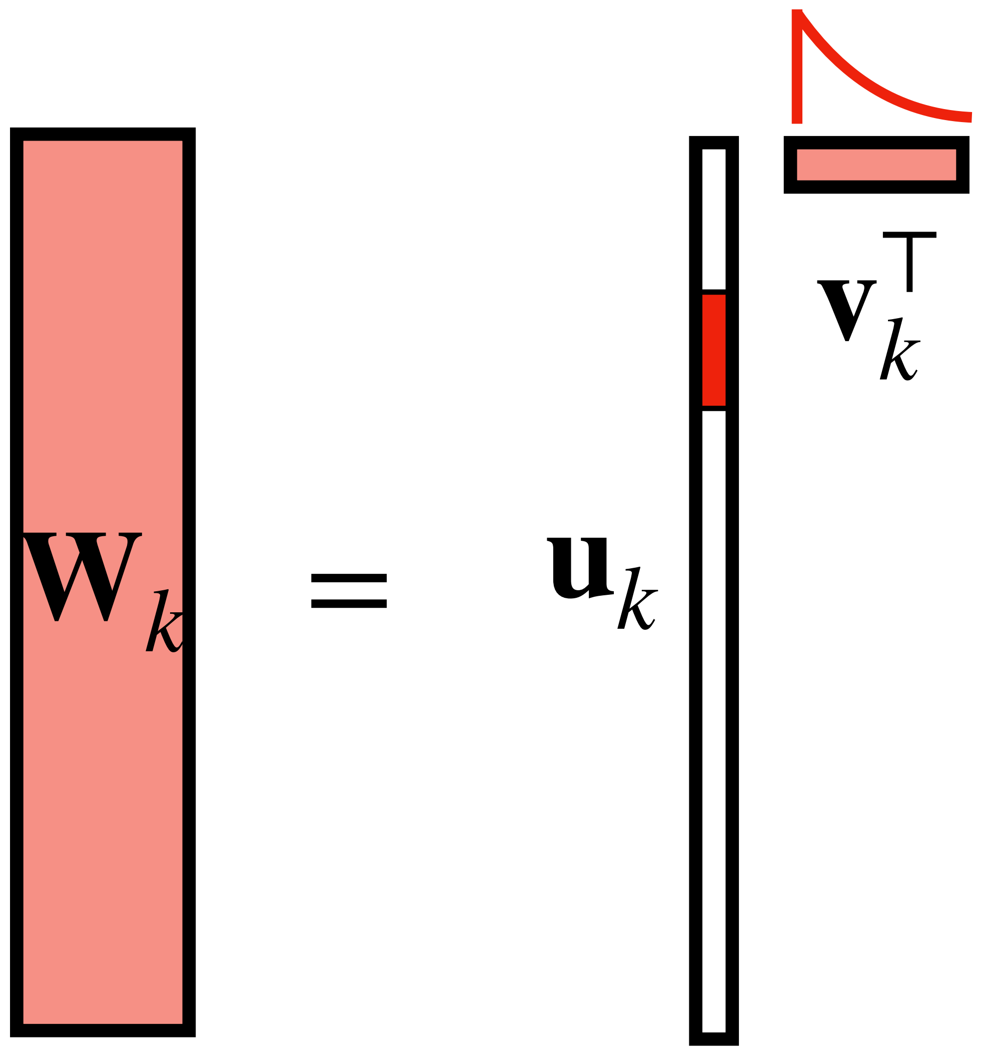
- Assume the calcium responses factor into spatial and temporal components.

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

- 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(100ms).



Calcium response model

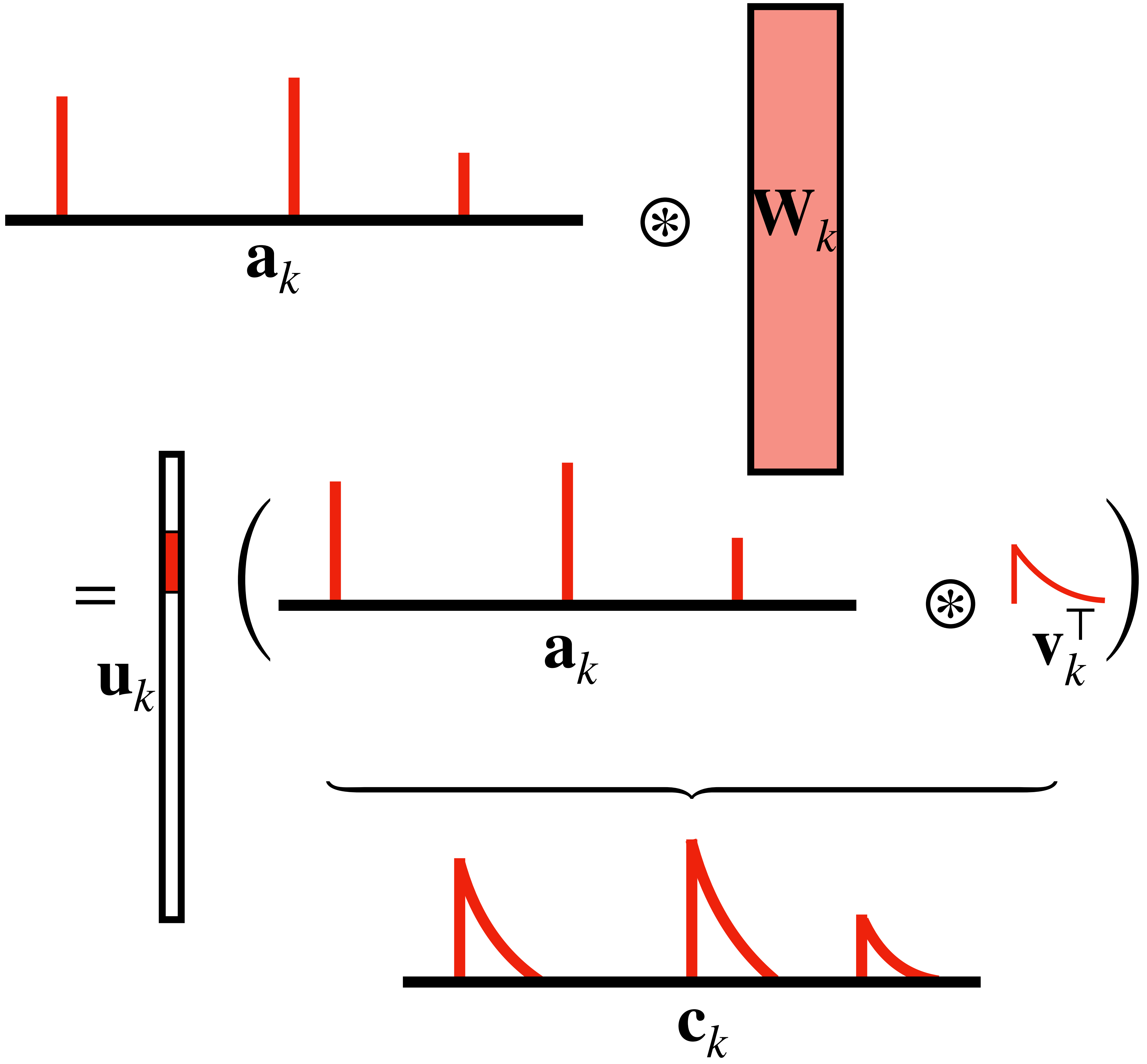
Then

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

where

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

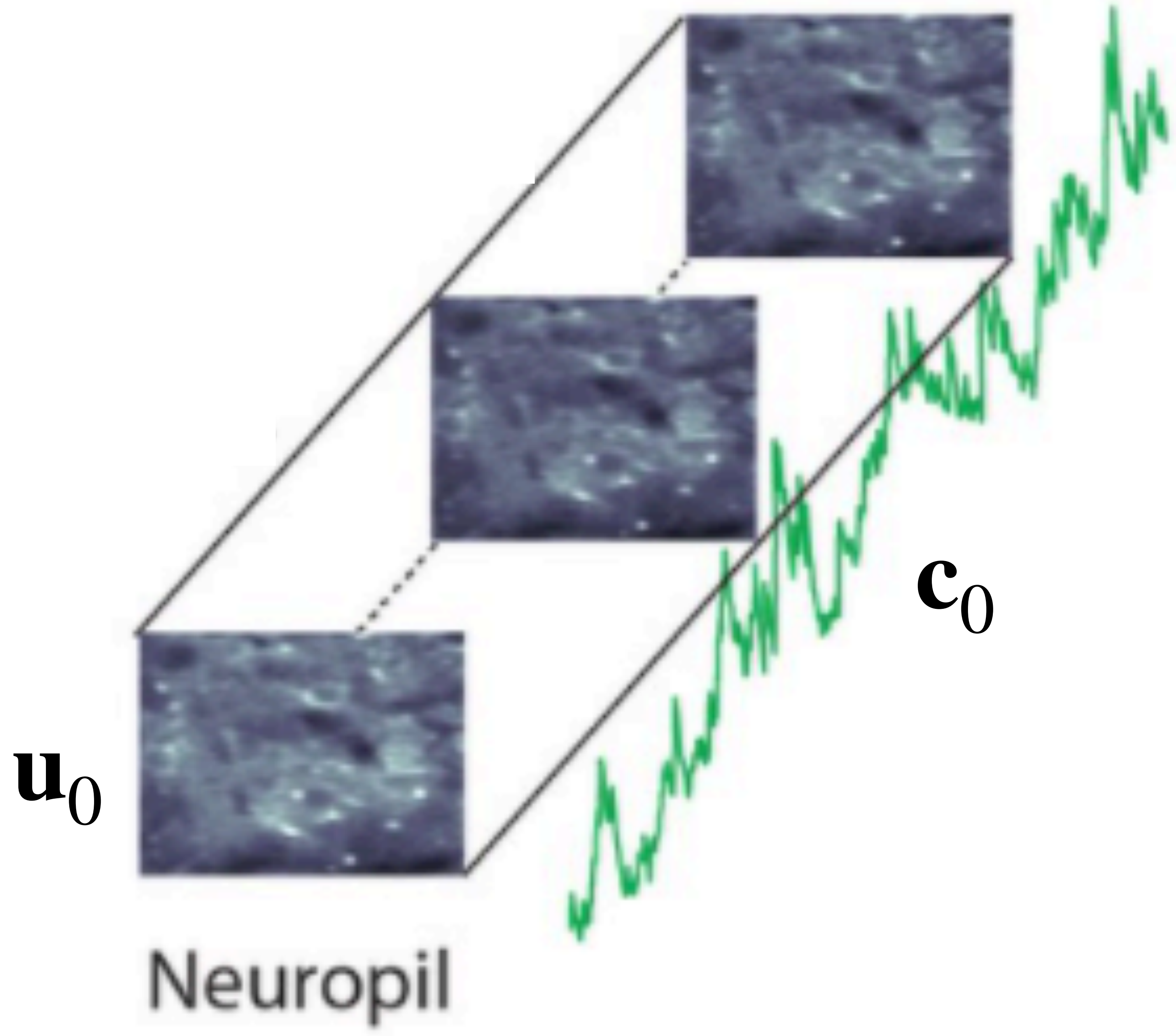
is the calcium trace of neuron k



Background model

Unlike last time, here we will explicitly model **background activity.**

Assume the background neuropil has its own spatial footprint $\mathbf{u}_0$ and time-varying intensity $\mathbf{c}_0 \in \mathbb{R}^T$.

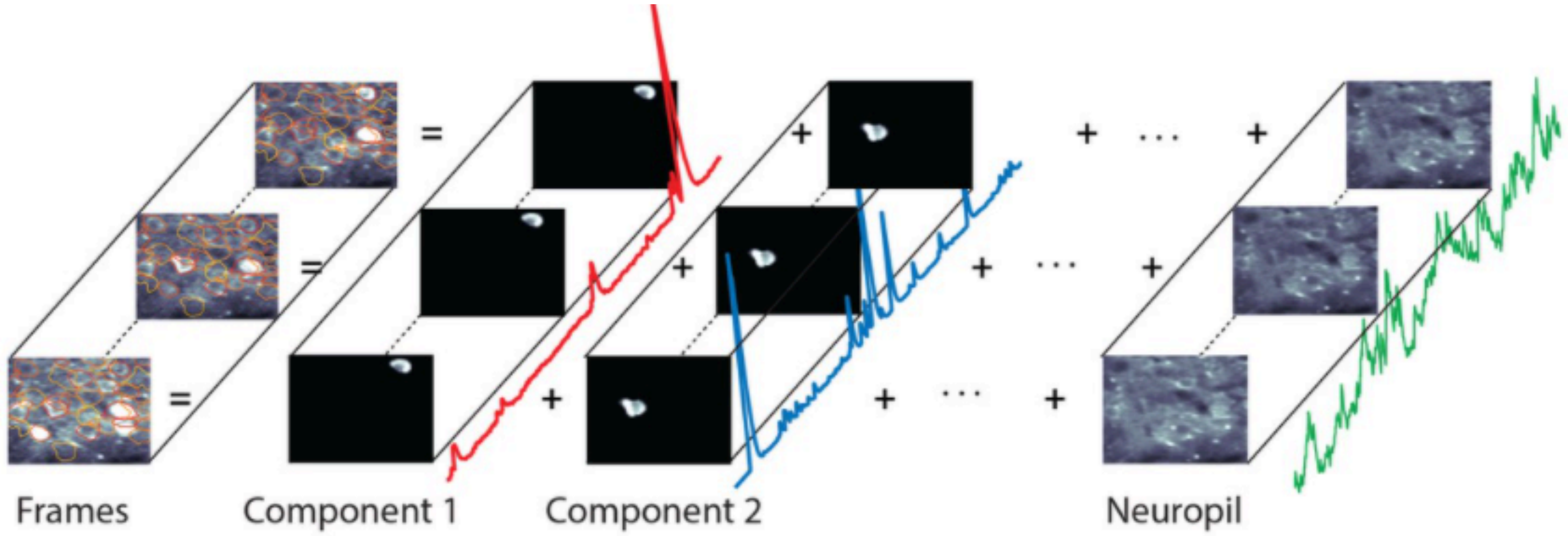


Probabilistic Model

Likelihood in terms of footprints and calcium traces

Putting it all together, the likelihood in terms of footprints and calcium traces is,

$$
\begin{aligned}
p(\mathbf{X} \mid \mathbf{U}, \mathbf{C}) &= \prod_{n=1}^{N} \prod_{t=1}^{T} \mathcal{N}\left(x_{n,t} \,\middle|\, \sum_{k=1}^{K} u_{k,n} c_{k,t} + u_{0,n} c_{0,t}, \sigma^2\right) \\
&= -\frac{1}{2\sigma^2} \|\mathbf{X} - \mathbf{U}\mathbf{C}^\top\|_\mathrm{F}^2 + \text{const}
\end{aligned}
$$



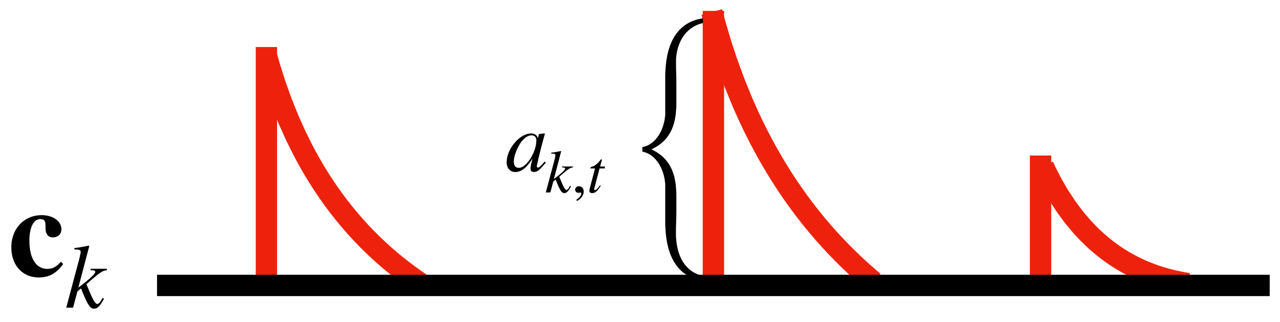
Recursive formulation

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

$$c_{k,t} = \sum_{d=0}^{D} a_{k,t-d} e^{-d/\tau}$$

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

(Technically, we assumed $D \gg \tau$.)

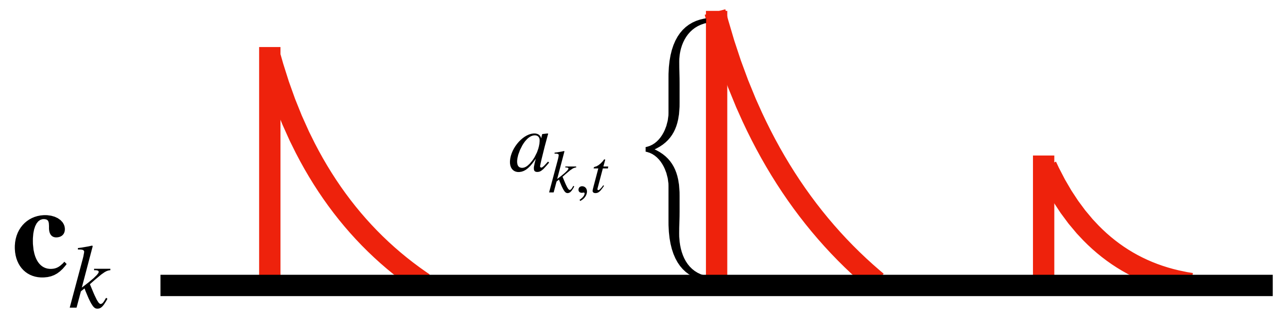


Equivalently, $a_{k,t} = c_{k,t} - e^{-1/\tau} c_{k,t-1}$.

Recursive formulation

In matrix form,

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



Prior on calcium traces

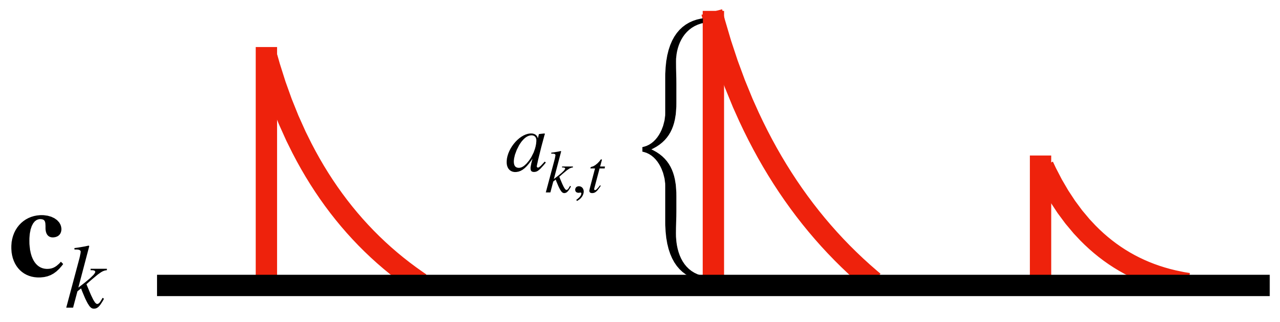
Via a prior on amplitudes

Note that $\mathbf{c}_k$ and $\mathbf{a}_k$ are in 1:1 correspondence. A prior on amplitudes is a prior on calcium traces.

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

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

(Technically, this relies on change of measure formula and the fact that $|\mathbf{G}| = 1$.)



Optimizing the calcium traces

Consider the log likelihood as a function of $\mathbf{c}_k$, and define the residual $\mathbf{R} = \mathbf{X} - \sum_{j \neq k} \mathbf{u}_j \mathbf{c}_j^\top - \mathbf{u}_0 \mathbf{c}_0^\top$ with columns $\mathbf{r}_t$,

$$\begin{aligned}
\log p(\mathbf{X} \mid \mathrm{U}, \mathbf{C}) &= \sum_{t=1}^{T} \log \mathcal{N}(\mathbf{r}_t \mid \mathbf{u}_k c_{k,t}, \sigma^2 \mathbf{I}) \\
&= -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (\mathbf{r}_t - \mathbf{u}_k c_{k,t})^\top (\mathbf{r}_t - \mathbf{u}_k c_{k,t}) \\
&= -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (\mathbf{r}_t^\top \mathbf{r}_t - \mathbf{r}_t^\top \mathbf{u}_k c_{k,t} + c_{k,t}^2 \mathbf{u}_k^\top \mathbf{u}_k) \\
&= -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (-\mathbf{r}_t^\top \mathbf{u}_k c_{k,t} + c_{k,t}^2) + c'
\end{aligned}$$

where we used the fact that $\mathbf{u}_k^\top \mathbf{u}_k = 1$.

Optimizing the calcium traces

Completing the square and simplifying, the log likelihood (up to an additive constant) is,

$$\log p(\mathbf{X} \mid \mathbf{U}, \mathbf{C}) = -\frac{1}{2\sigma^2} \|\mathbf{c}_k - \boldsymbol{\mu}_k\|_2^2$$

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

Finally, add the prior log probability to obtain the following objective for optimizing the calcium trace of neuron k, holding everything else fixed:

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

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,

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

For $\mathbf{c}_k \geq 0$.

This is a **convex optimization problem!**

Optimizing the calcium traces

Dual formulation

Maximizing $\mathscr{L}(\mathbf{c}_k)$ is equivalent to solving the following dual problem,

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

for some threshold θ.

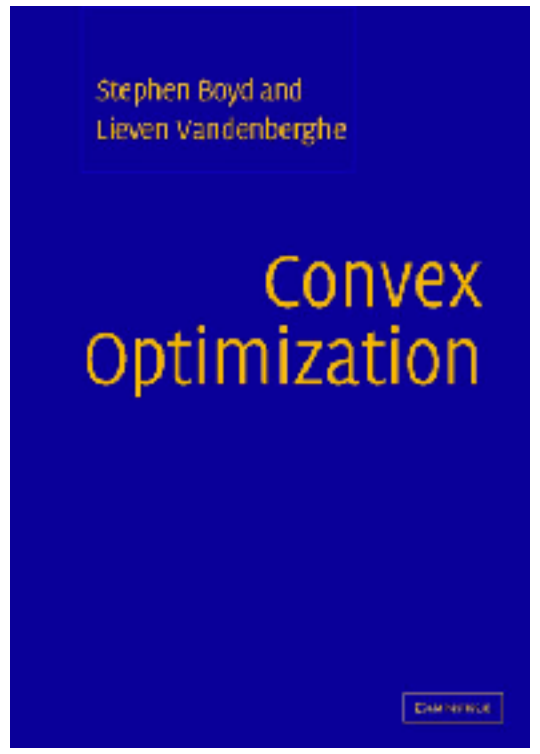
Optimizing the calcium traces

Setting the regularization hyperparameter

- In the primal form, we have a hyperparameter λ; in the dual we have a threshold θ. How should we set these?
- Under the model, $c_{k,t} - \mu_{k,t} \sim \mathcal{N}(0, \sigma^2)$, and $z_{k,t} = \dfrac{c_{k,t} - \mu_{k,t}}{\sigma} \sim \mathcal{N}(0, 1)$.
- $\|\mathbf{z}_k\|_2 = \sigma^{-1}\|\mathbf{c}_k - \boldsymbol{\mu}_k\|_2$ is the norm of a vector of iid Gaussians. It follows a chi (χ) distribution.
- **Idea:** for large T, the chi distribution concentrates around $\sqrt{T}$. So set $\theta = (1 + \epsilon)\sigma\sqrt{T}$.
- How to get σ? 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!



Welcome to CVXPY 1.3

Convex optimization, for everyone.

We are building a CVXPY community on Discord. Join the conversation!

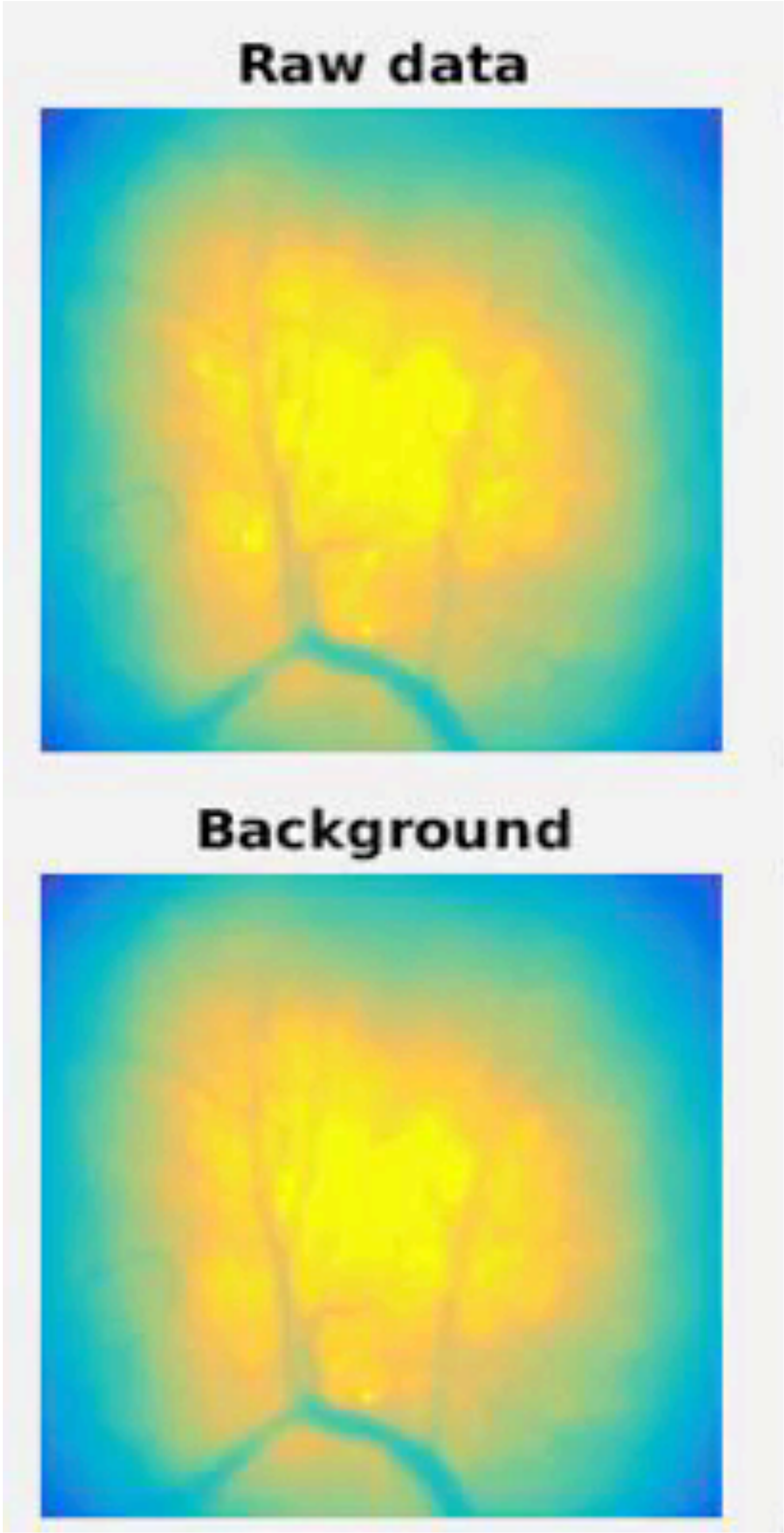
CVXPY is an open source Python-embedded modeling language for convex optimization problems. It lets you express your problem in a natural way that follows the math, rather than in the restrictive standard form required by solvers.

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

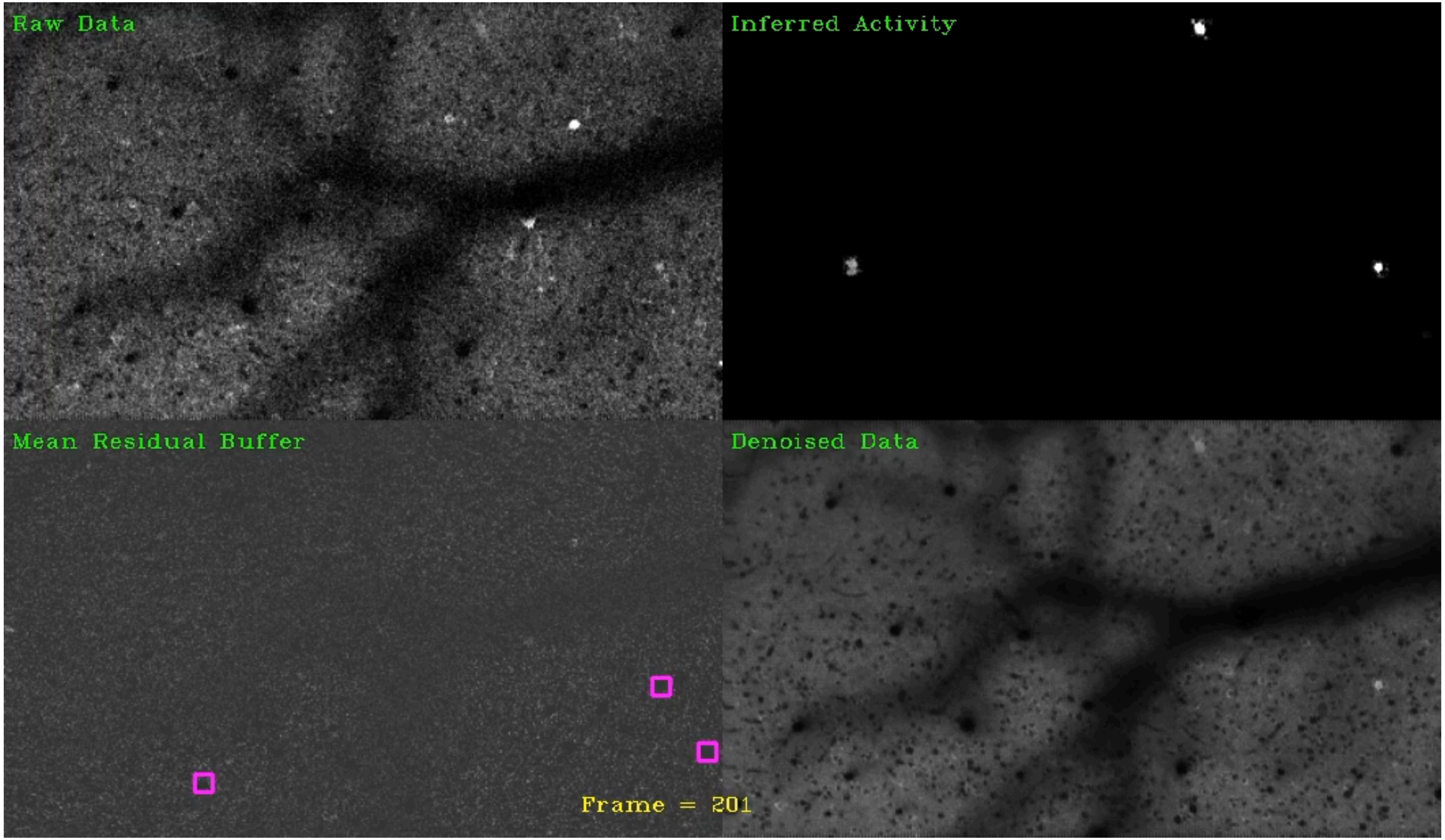
```python
import cvxpy as cp
import numpy as np

# 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

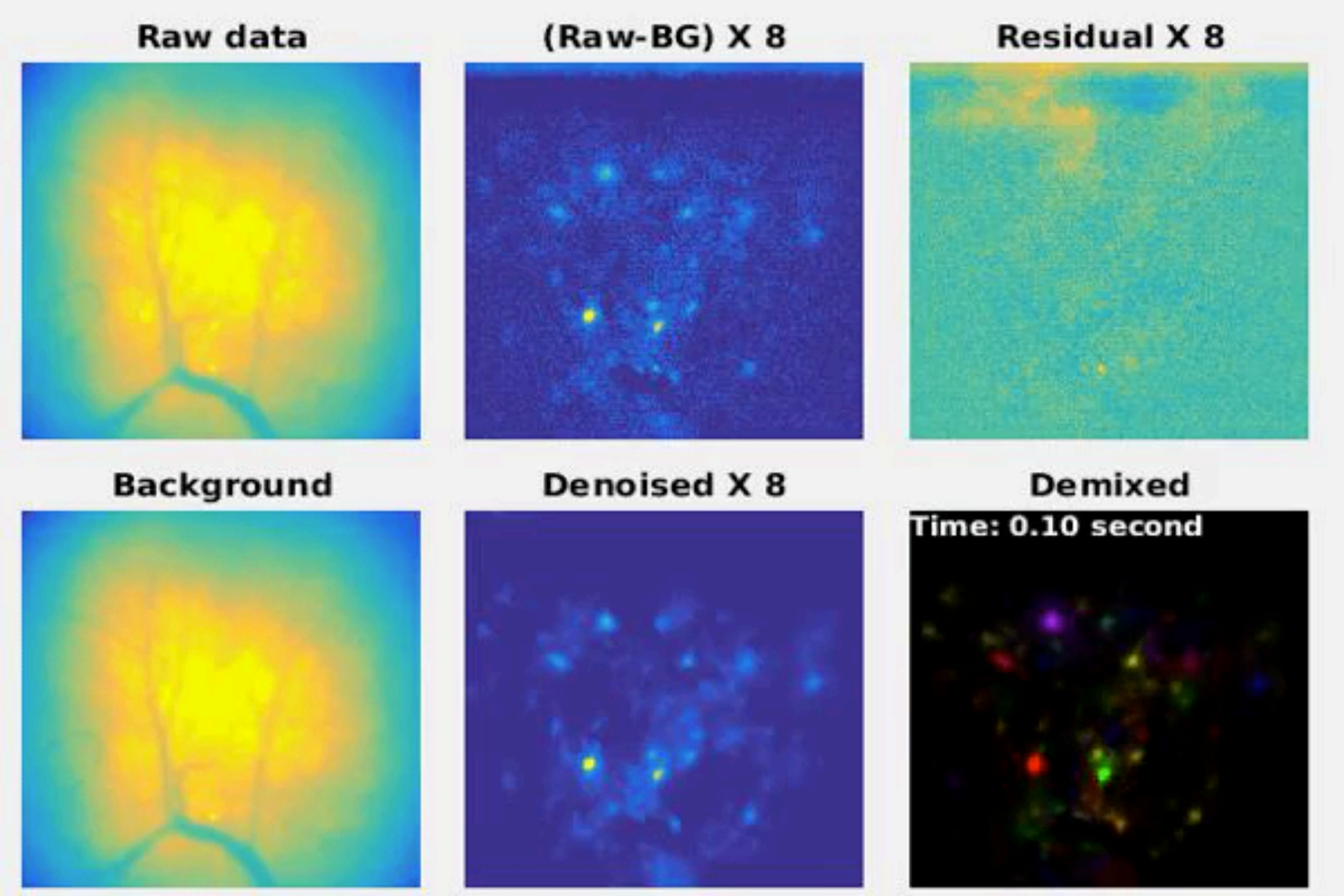
- We typically constrain the **spatial factors to be non-negative** too, unlike in spike sorting.
- We need to account for **background fluorescence** from out-of-focus cells.
- Typically, assume **rank-1** or **spatially smooth** background. See notes.
- As always, **preprocessing is important** for 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. “CaImAn an Open Source Tool for Scalable Calcium Imaging Data Analysis.” eLife 8 (January). https://doi.org/10.7554/eLife.38173.