# Continuous Latent Variable Models for Neuroscience Data

Scott Linderman

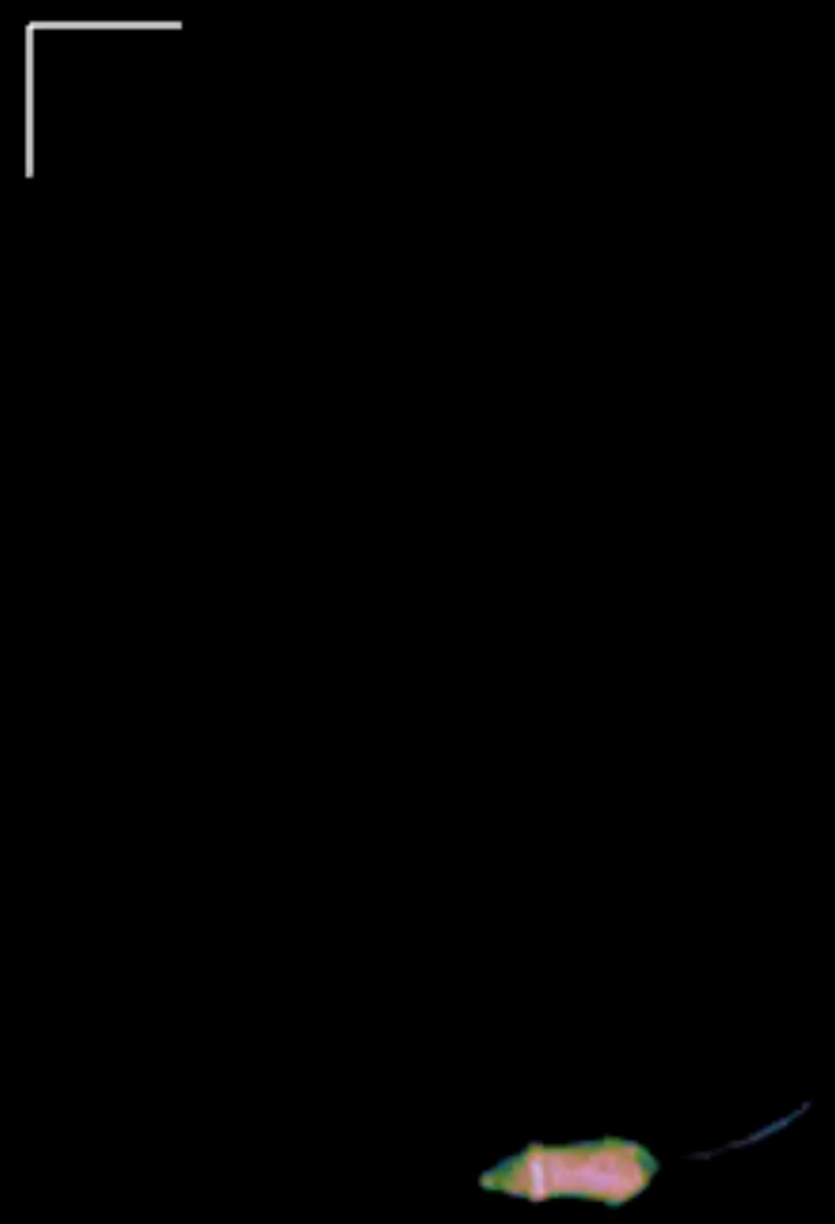
Stanford University

Department of Statistics

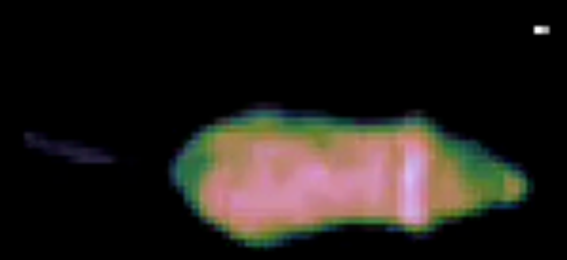
Wu Tsai Neurosciences Institute

# Example: Depth video of a freely moving mouse

Raw data



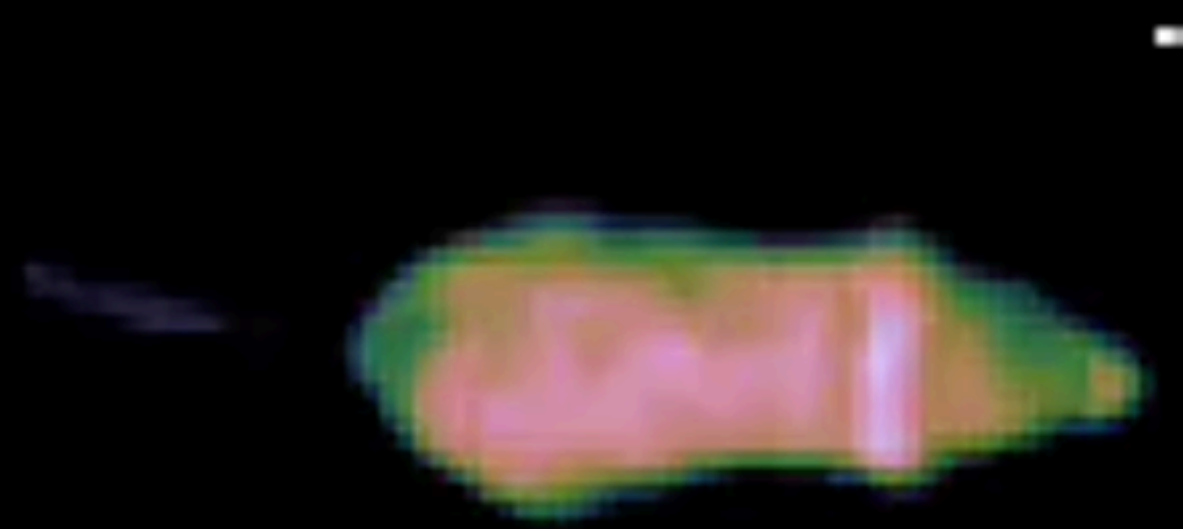
Cropped and Rotated



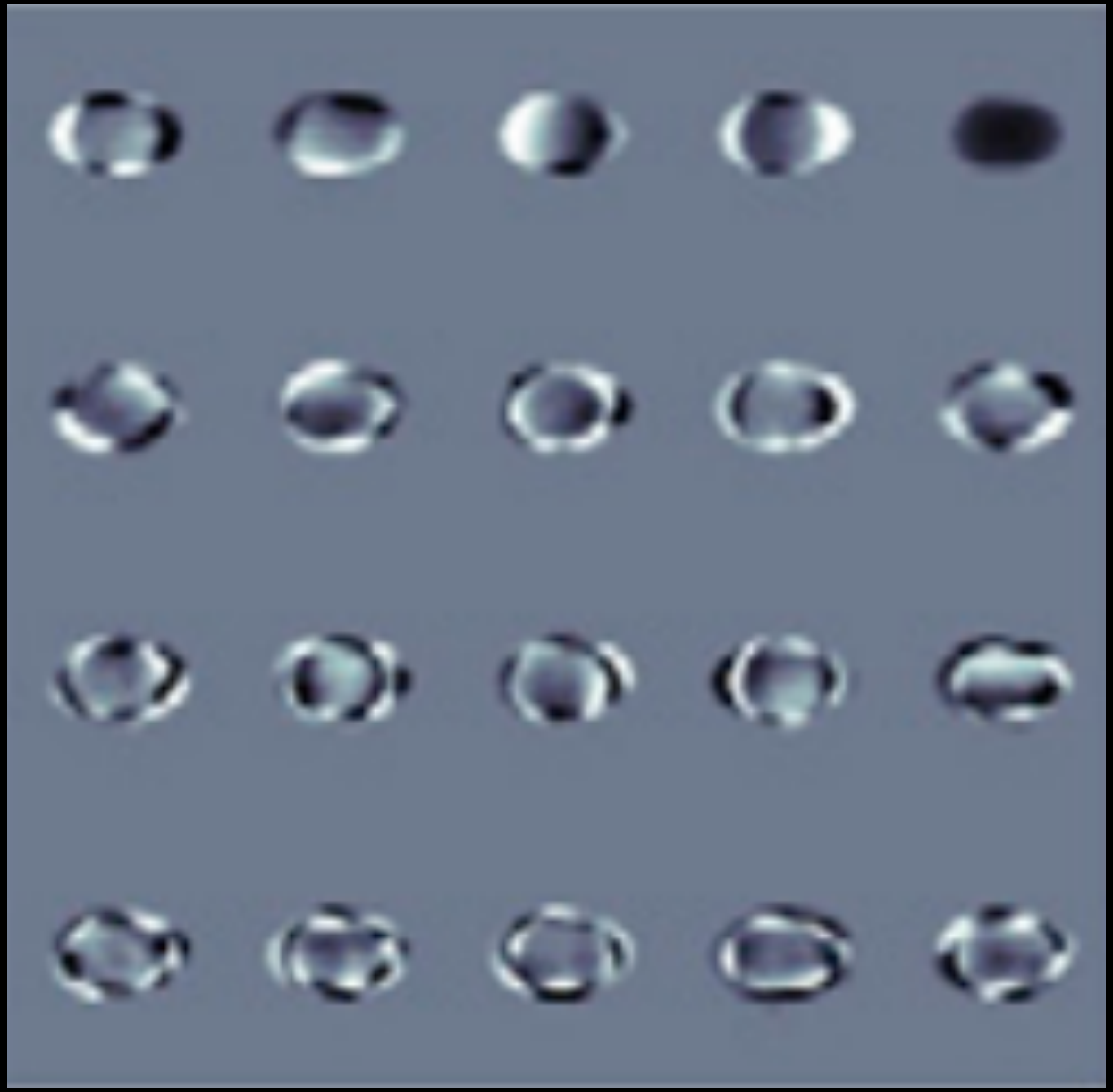
Wiltschko et al. (2015)
Markowitz et al. (2018)

# Principal Components Analysis (PCA) of video frames

Single frame

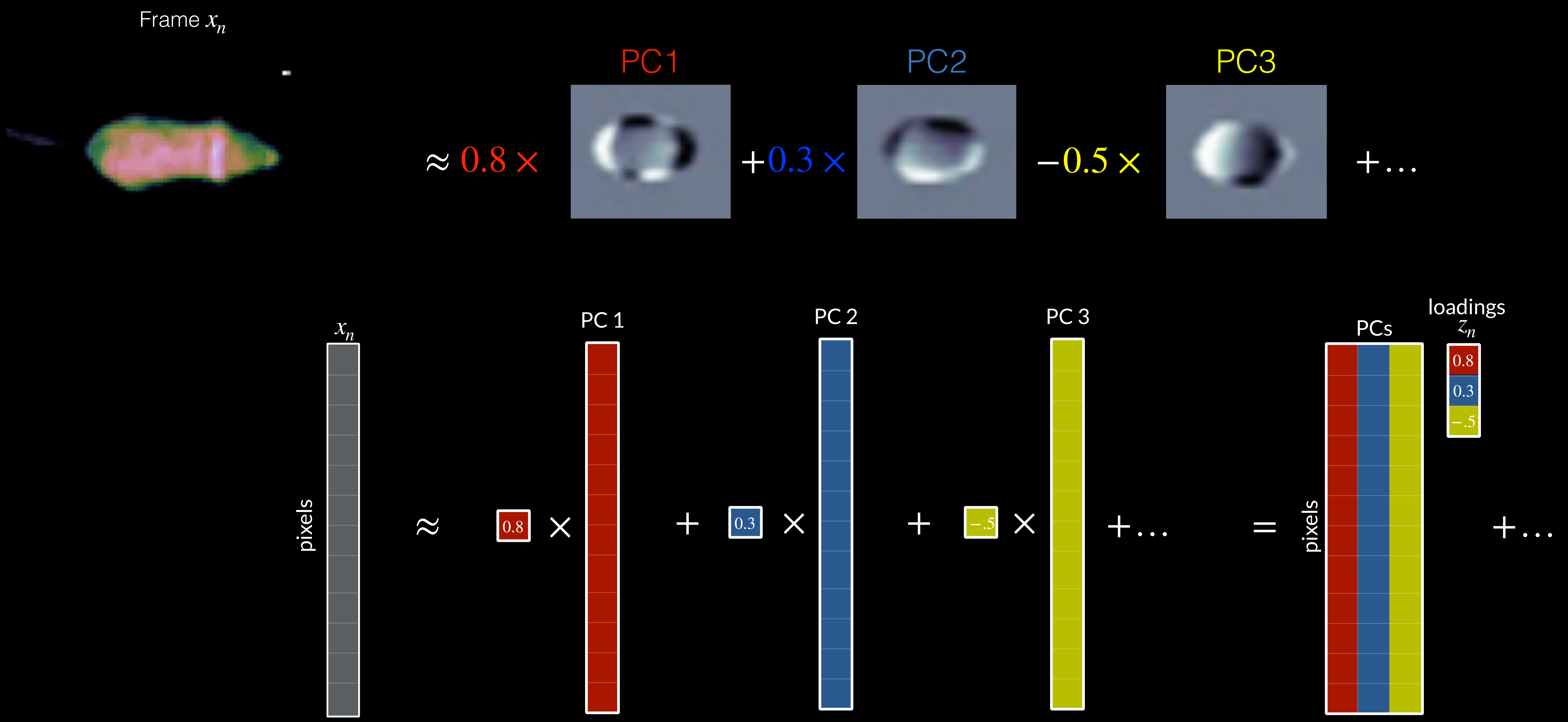


Principal Components



Wiltschko et al. (2015)
Markowitz et al. (2018)

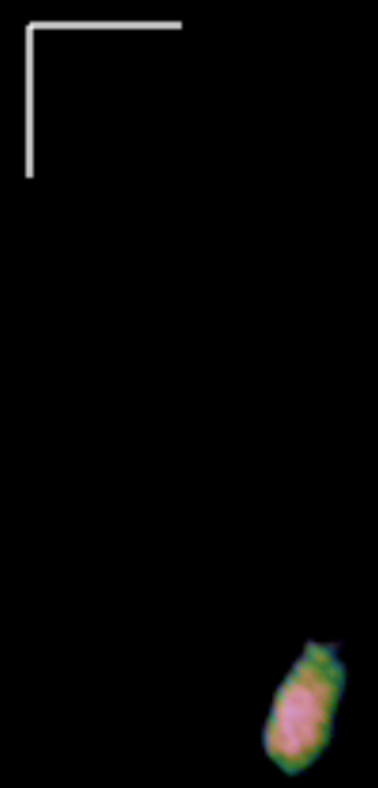
# Principal Components Analysis (PCA) of video frames



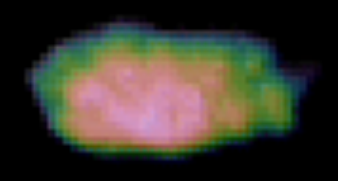

Wiltschko et al. (2015)
Markowitz et al. (2018)

# Depth video of a freely moving mouse

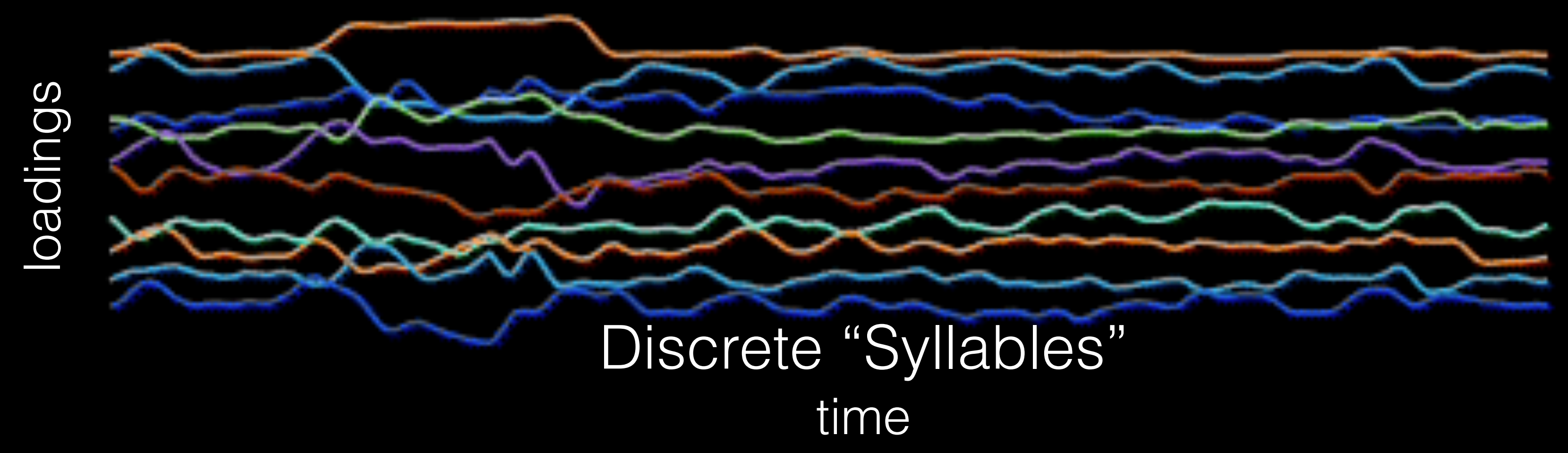
Raw data



Cropped and Rotated



Principal Components



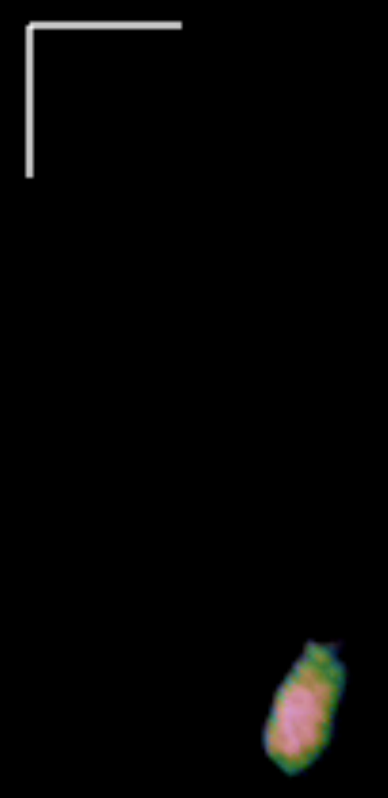

Discrete “Syllables”

time

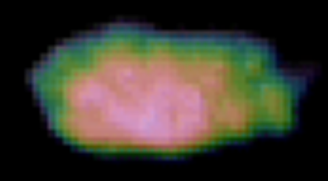
Wiltschko et al. (2015)
Markowitz et al. (2018)

# Depth video of a freely moving mouse

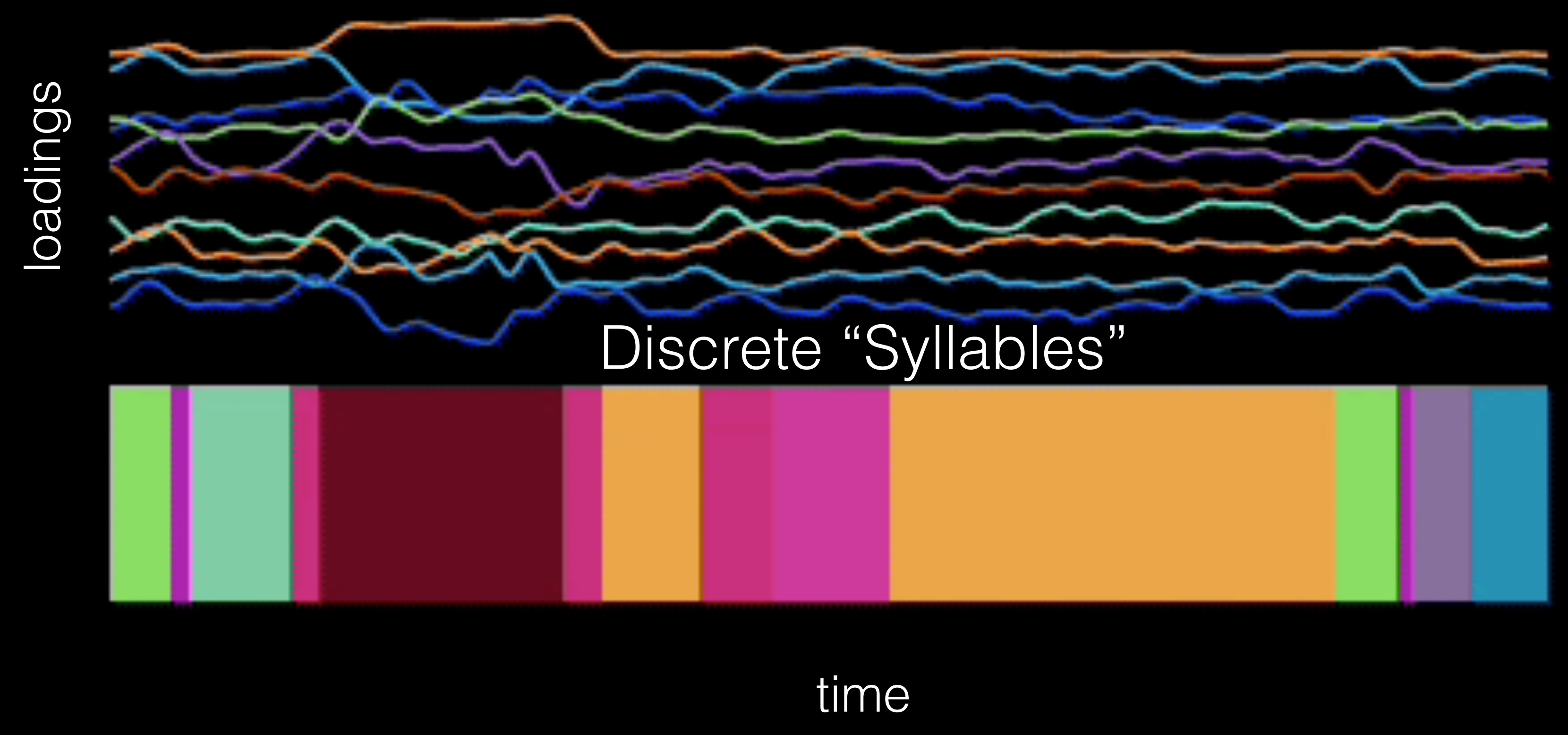
Raw data



Cropped and Rotated

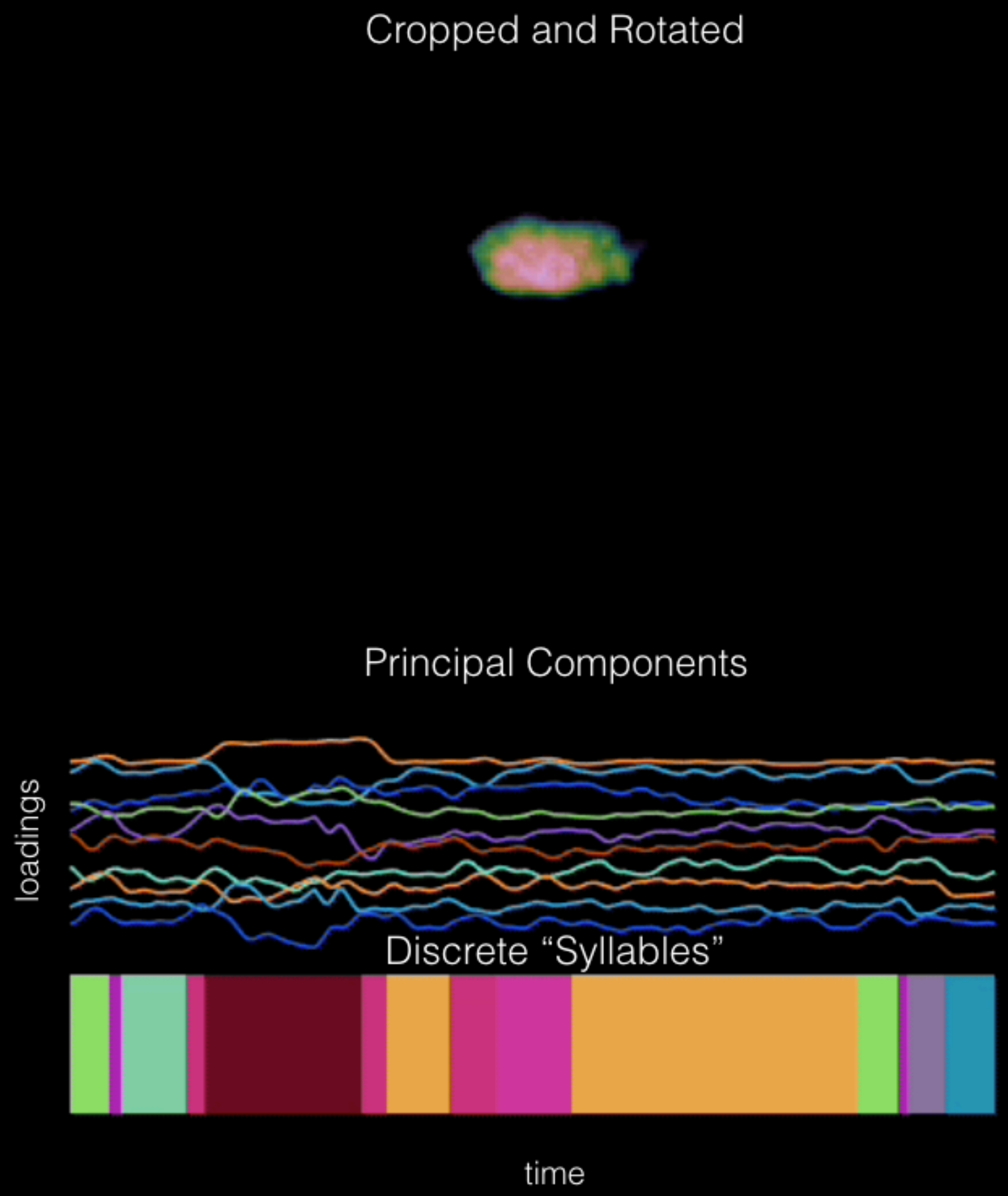


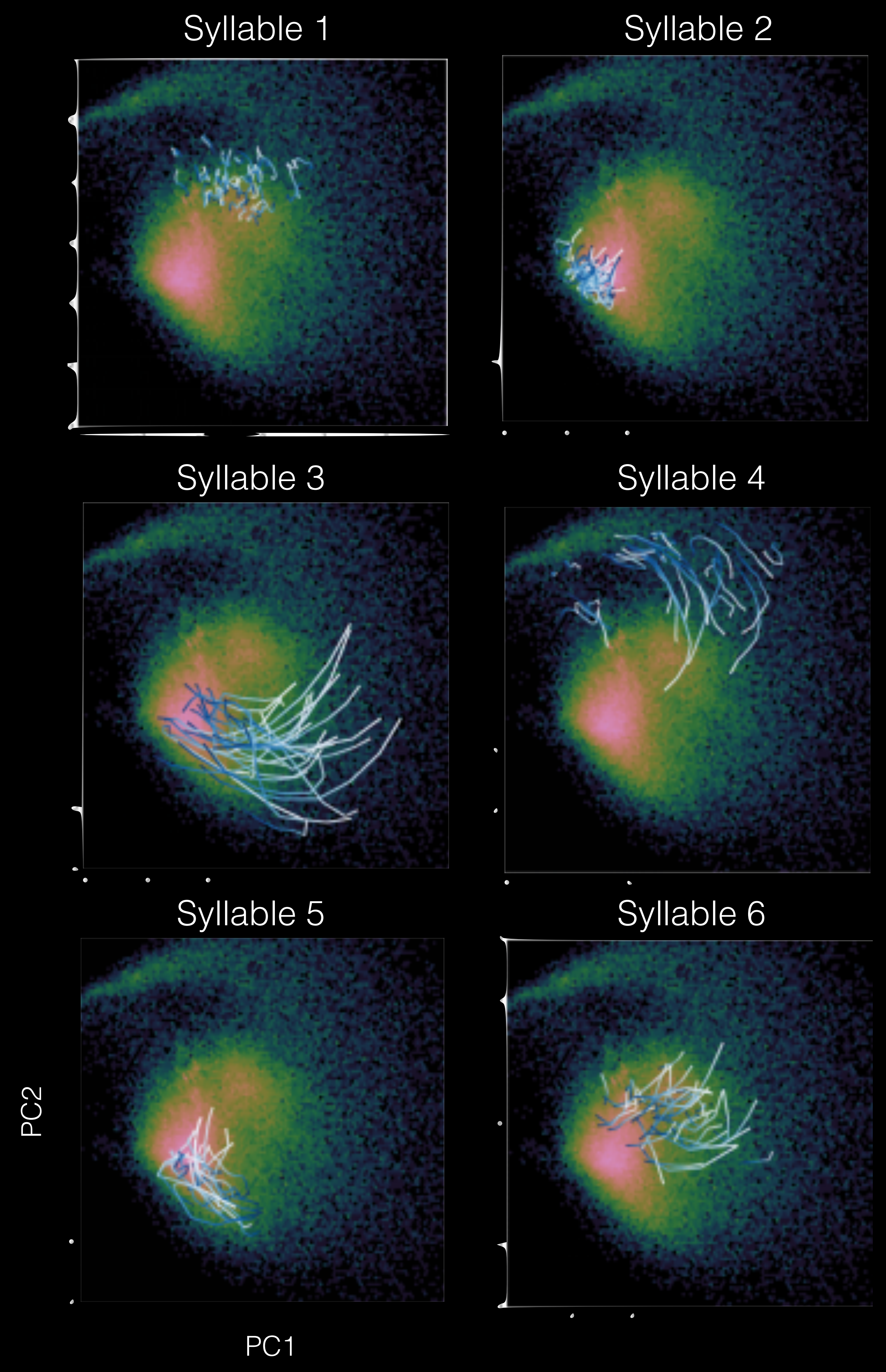
Principal Components




Wiltschko et al. (2015)
Markowitz et al. (2018)

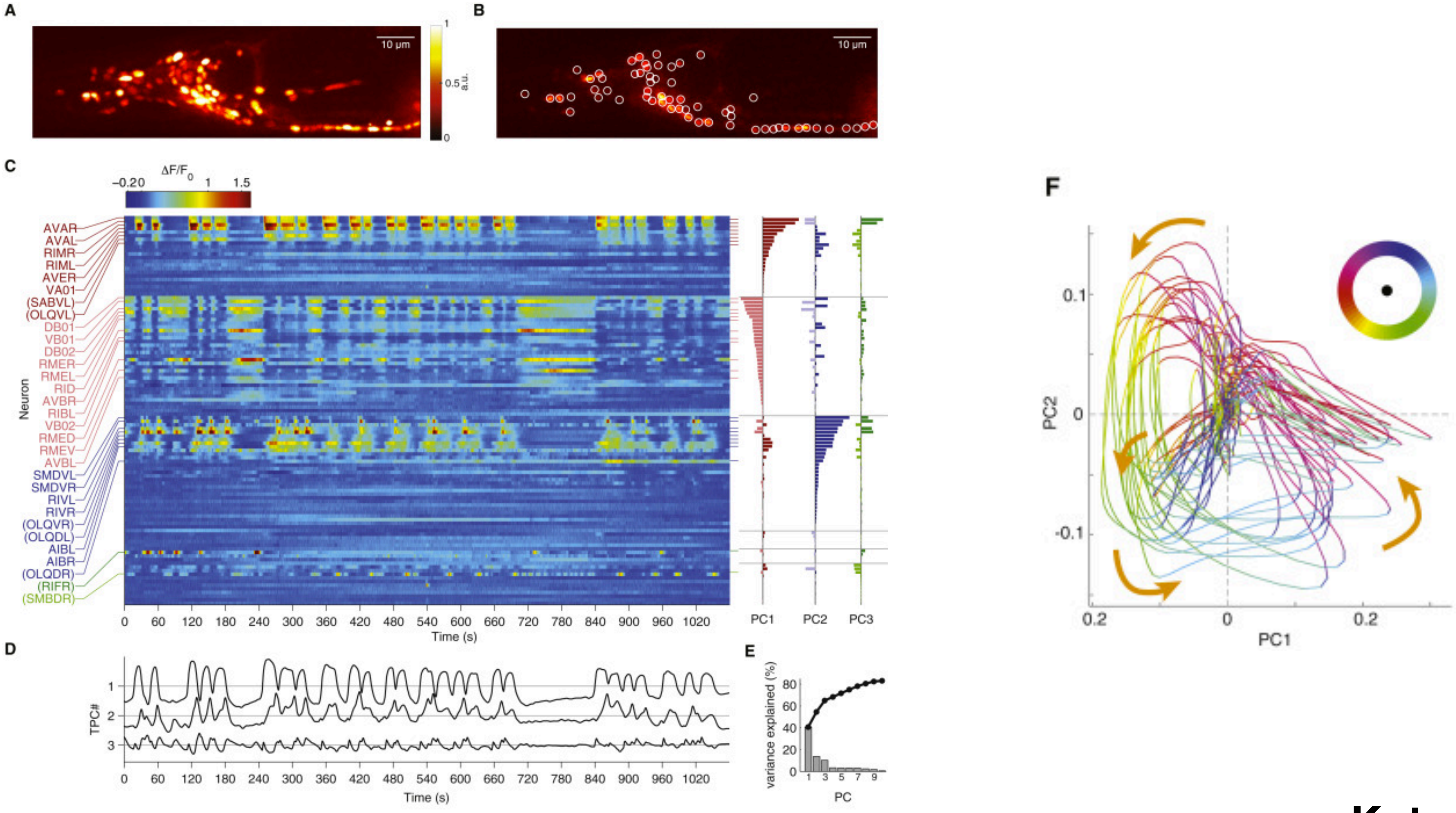
# Syllables as trajectories in PCA space



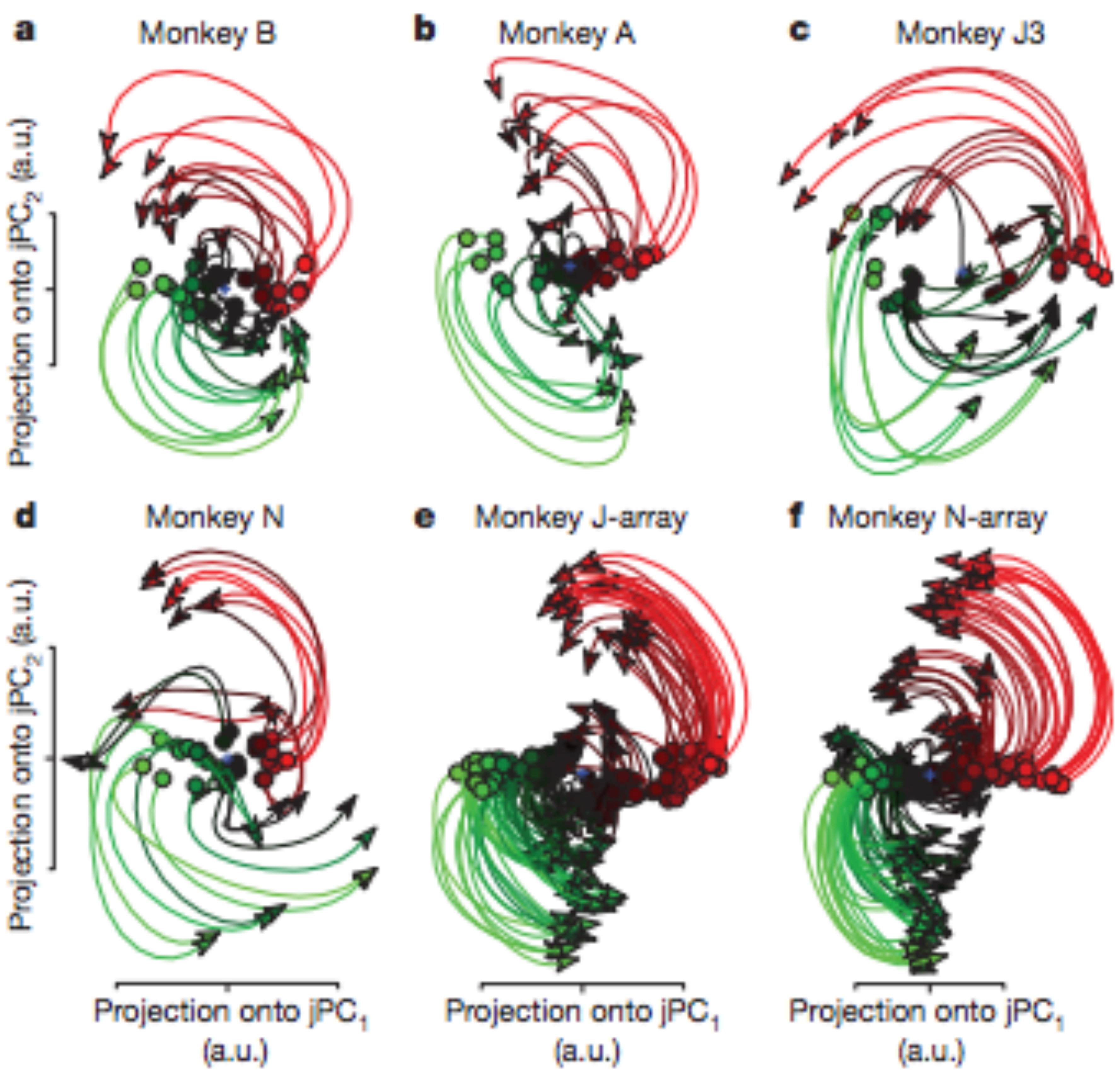

**Wiltschko et al. (2015)**
**Markowitz et al. (2018)**

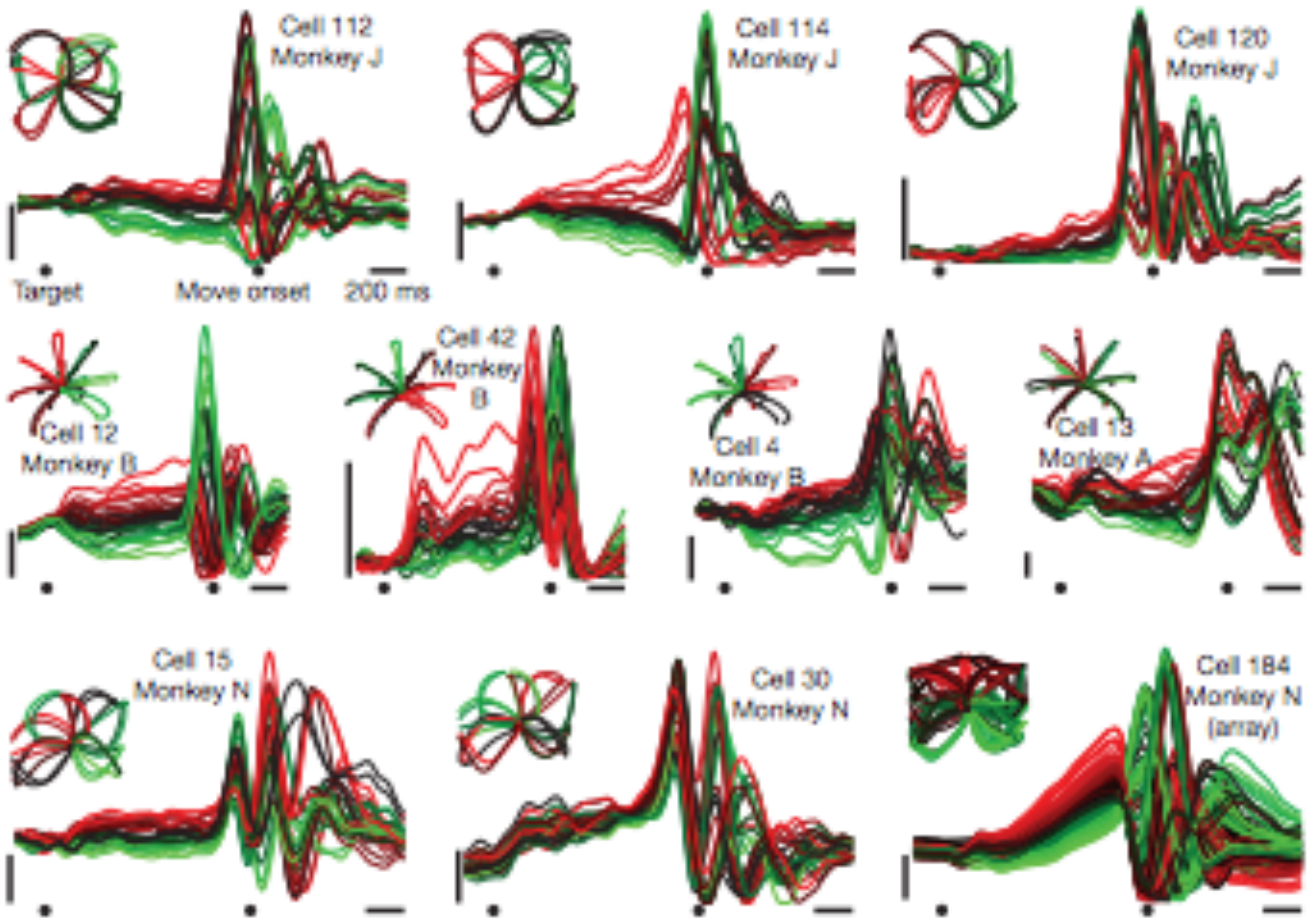
# Example: Calcium dF/F traces in C. elegans




**Kato et al (2015)**

# Example: Trial-averaged firing rates in motor cortex



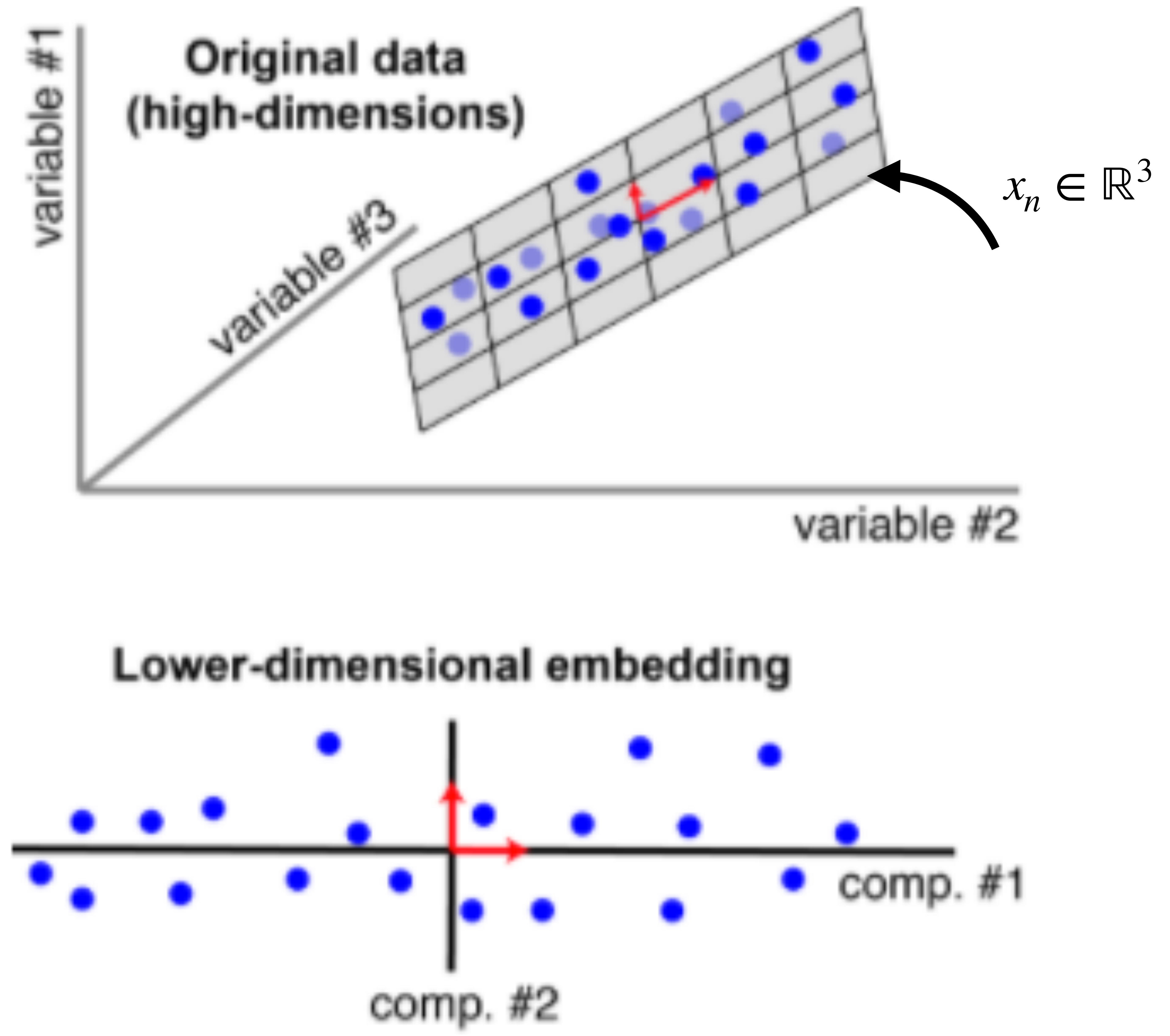




**Churchland et al (2012)**

# PCA: Finding dimensions of maximal variance

- PCA finds direction vectors (i.e. *components*) $c$ along which the variance of the data is largest.
- Projection of $x_n$ onto $c$ is given by $c^\top x_n$. Solve for

$$\max_c \ \mathrm{Var}[c^\top x] \text{ subject to } \|c\|_2 = 1.$$




Alex Williams

# Aside: Probability refresher

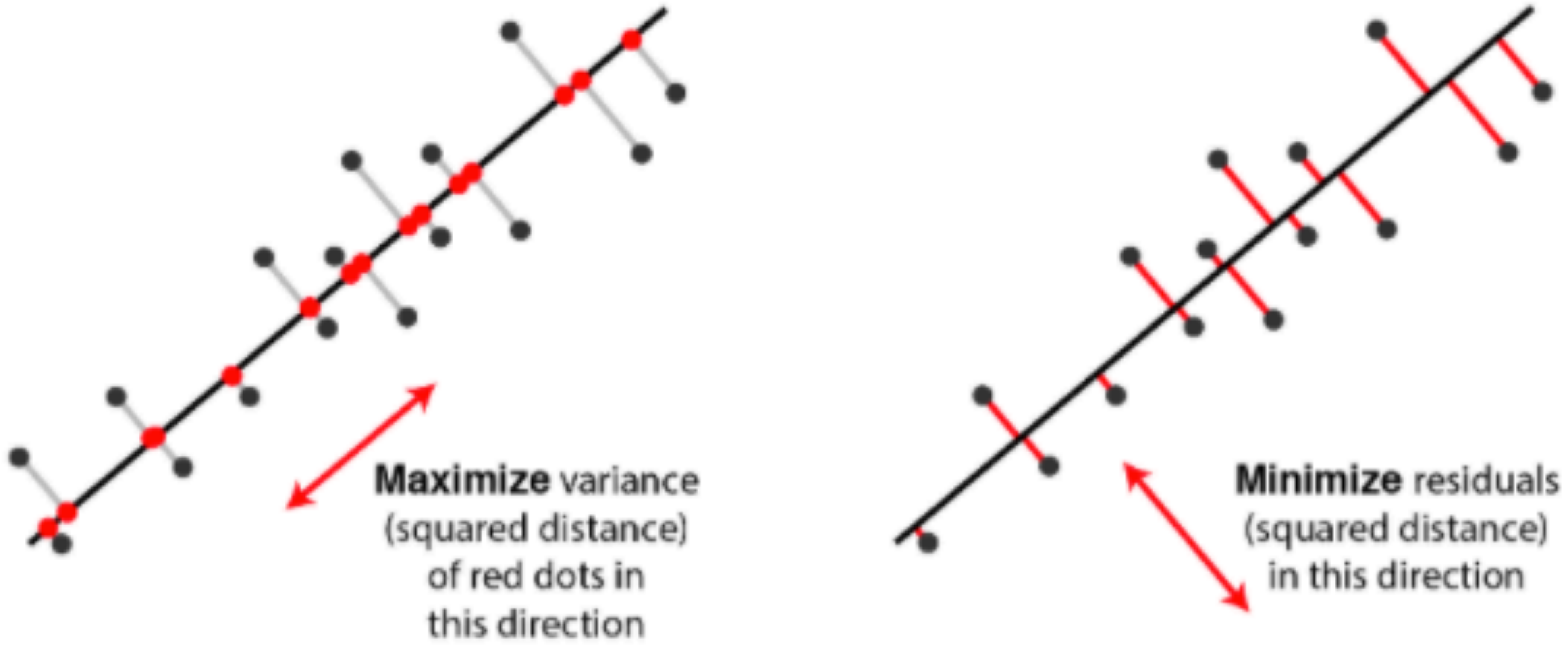
- **The big picture**
- **Random variables, probability densities**

# Aside: Probability refresher

- **Expectations**
- **Mean and variance**
- **Estimators**

# PCA: Minimizing reconstruction error

- Approximate $x_n \in \mathbb{R}^2$ with $\hat{x}_n = c_1 \cdot z_n$, where $c_1 \in \mathbb{R}^2$ is a unit vector (i.e. $\|c_1\|_2 = 1$) and $z_n \in \mathbb{R}$ is a scalar.
- For fixed $c_1$, find the value of $z_n$ that minimizes the error $\|x_n - \hat{x}_n\|_2$.



Two equivalent views of principal component analysis.

# PCA: Minimizing reconstruction error

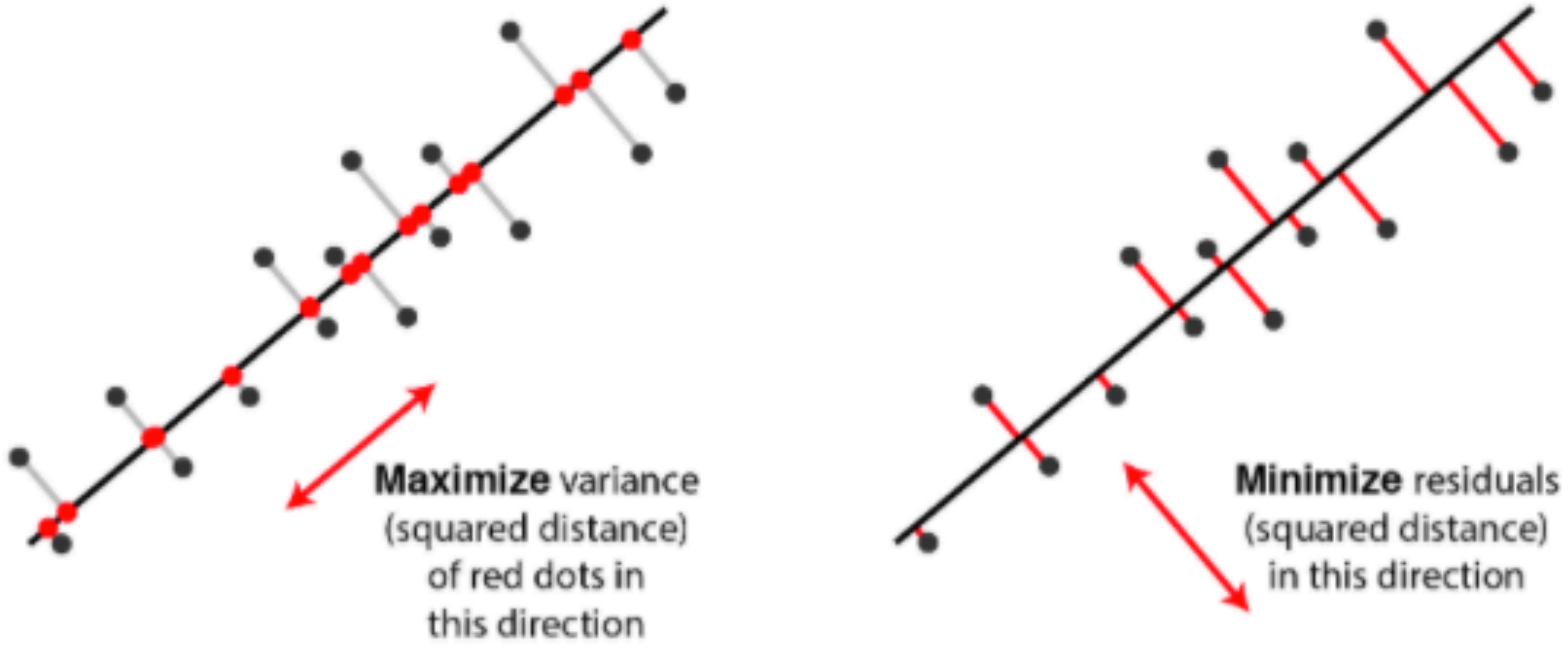
- Each datapoint can be represented as

$$x_n = \hat{x}_n + (x_n - \hat{x}_n)$$
$$= c_1(x_n^\top c_1) + c_2(x_n^\top c_2)$$

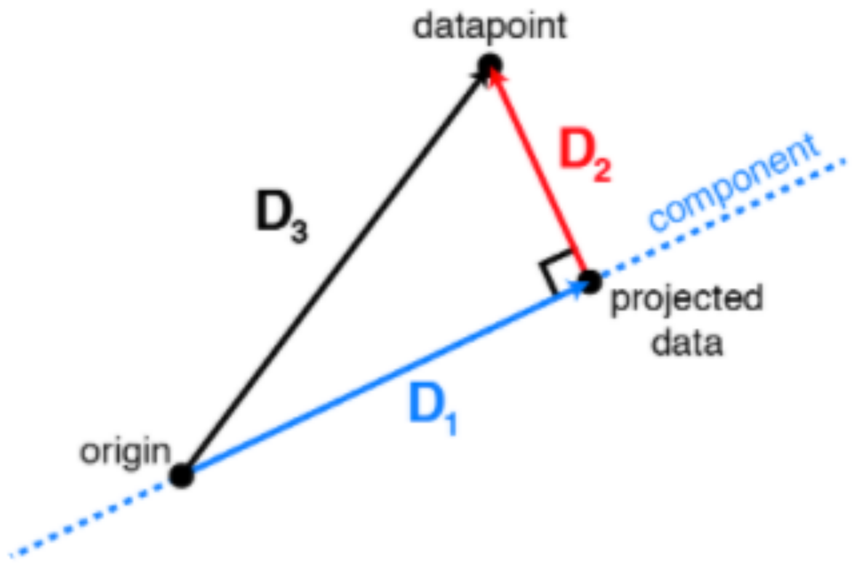
where $c_1$ and $c_2$ are orthogonal unit vectors. Note that we only need to solve for one of these two vectors.

- Minimize the sum of squared residuals wrt $c_2$

$$\min_{c_2} \frac{1}{N} \sum_{n=1}^{N} \|x_n - \hat{x}_n\|_2^2 \text{ subject to } \|c_2\|_2 = 1.$$




Two equivalent views of principal component analysis.




Alex Williams

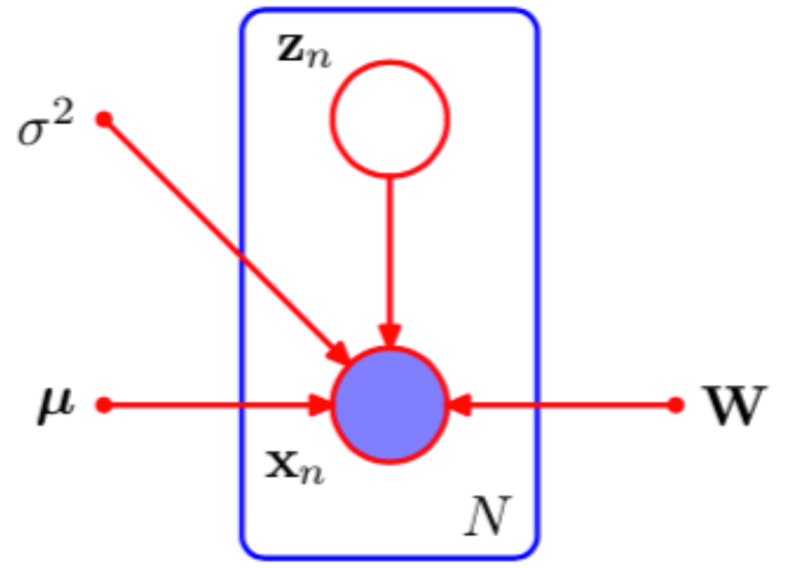
# Probabilistic PCA: A Generative model

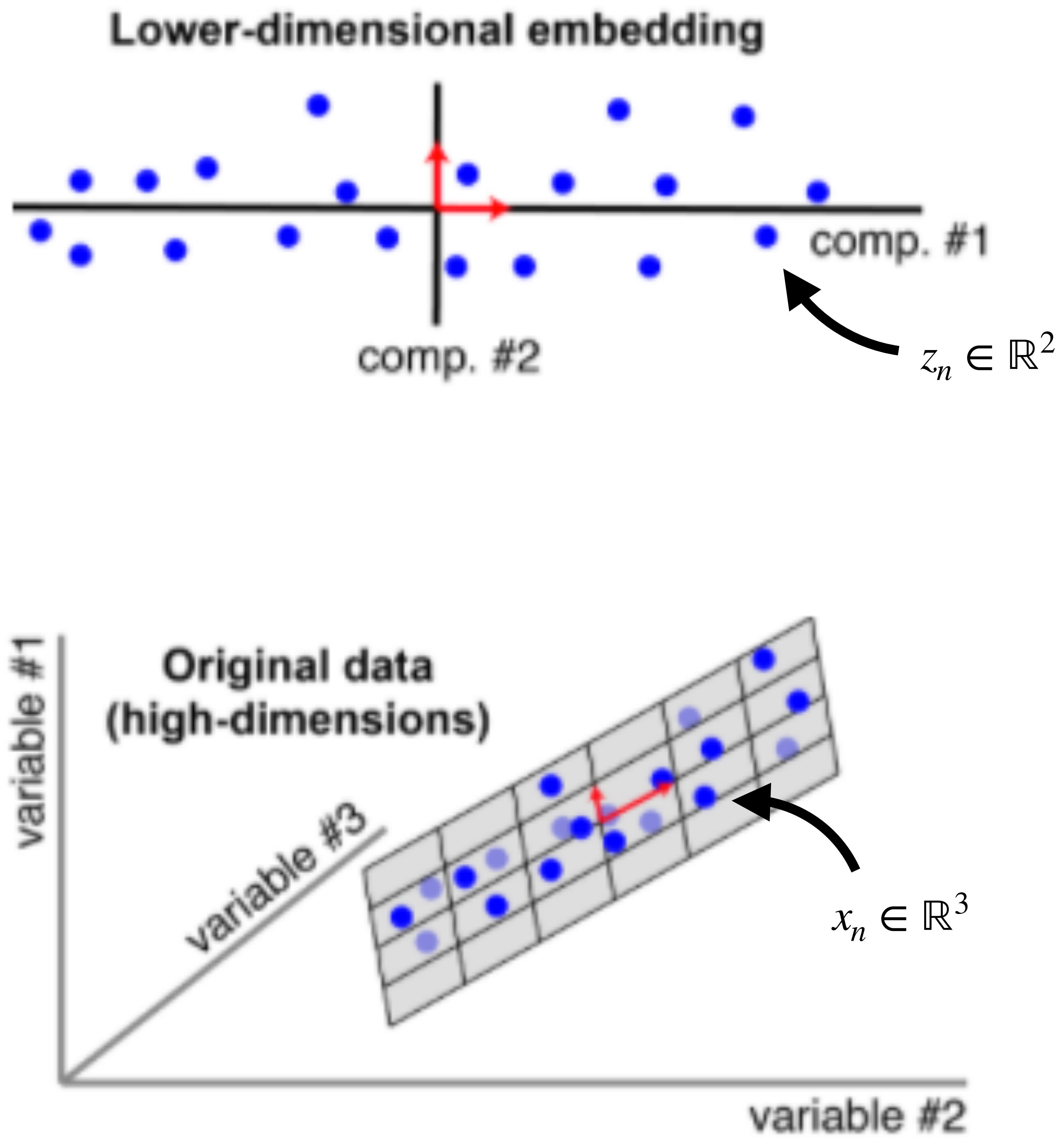
- **Step 1:** Sample low-dimensional loadings $z_n$ from Gaussian distribution.

$$z_n \sim \mathcal{N}(0, I)$$

- **Step 2:** Project into high dimensional space with matrix $W$, add an offset $\mu$, then add isotropic Gaussian noise to get $x_n$.

$$x_n = Wz_n + \mu + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$$



# Aside: Probability refresher II

- **Joint distribution**
- **Fundamental rules of probability**
- **Graphical models**
- **Bayes rule**

# Aside: The multivariate Gaussian distribution

- PDF
- Linear transformations
- Sums of Gaussians
- Marginals
- Conditionals

# Probabilistic PCA: A Generative model

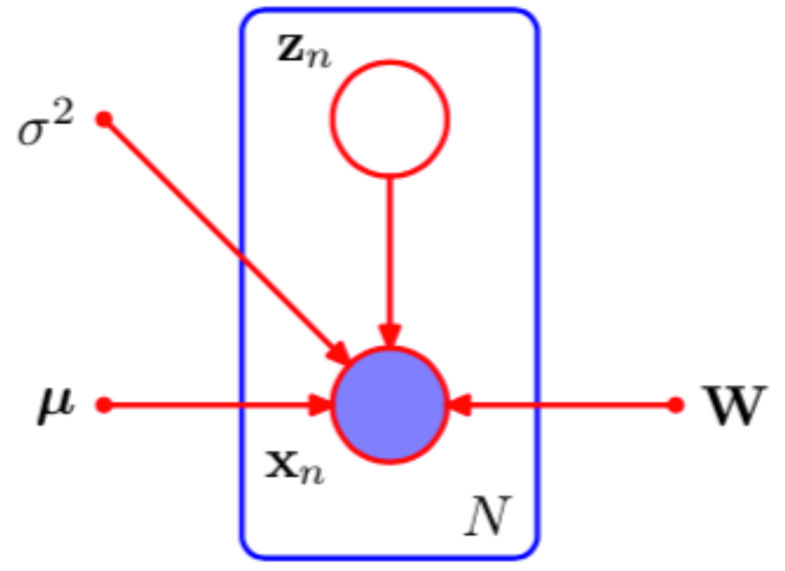
- **Step 1:** Sample low-dimensional loadings $z_n$ from Gaussian distribution.

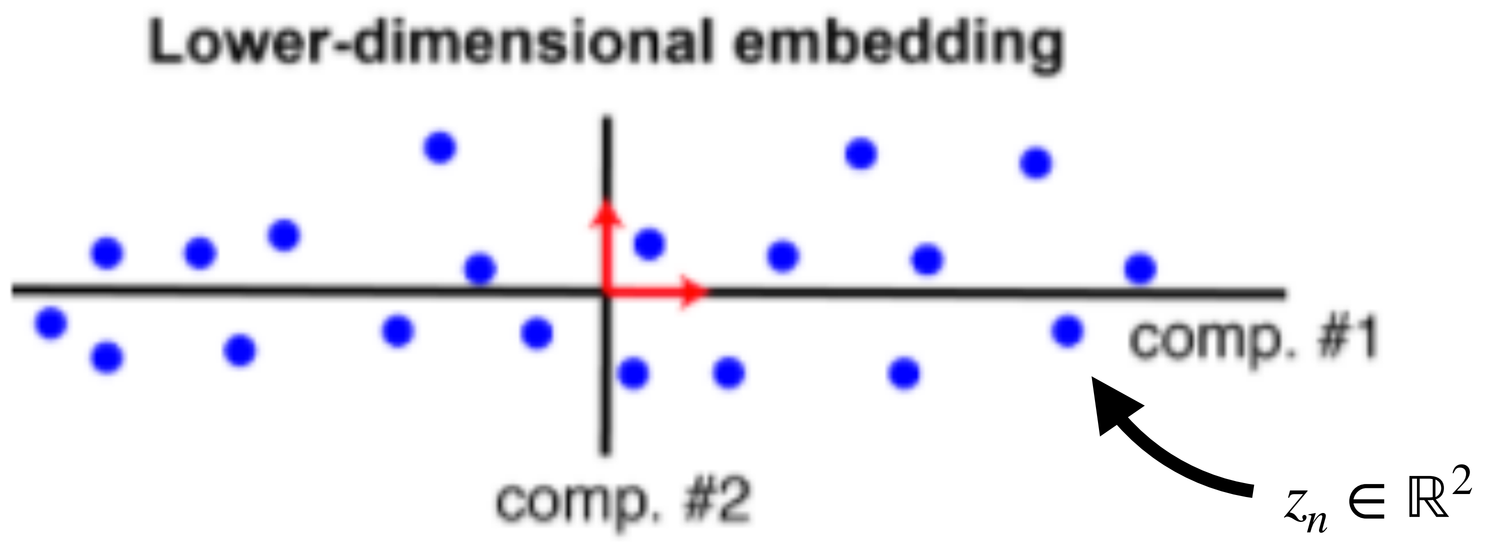
$$z_n \sim \mathcal{N}(0, I)$$

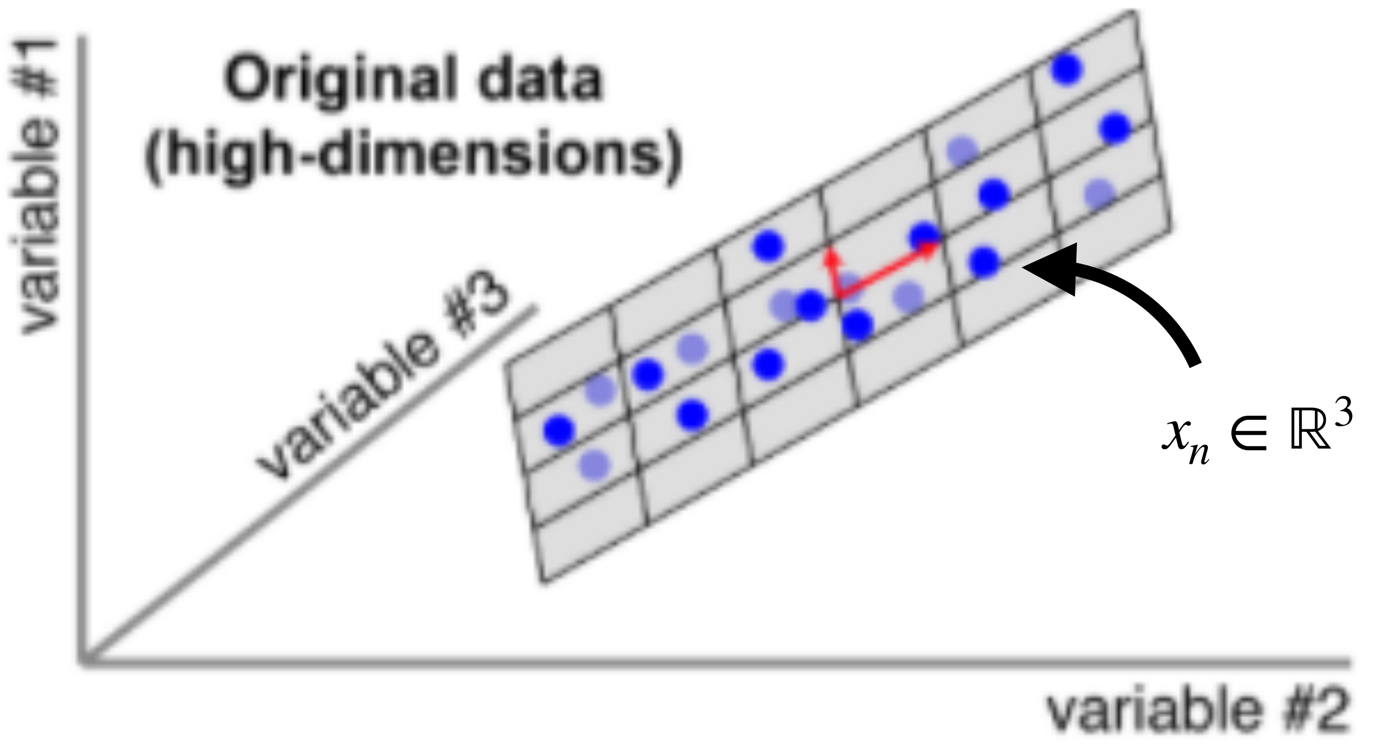
- **Step 2:** Project into high dimensional space with matrix $W$, add an offset $\mu$, then add isotropic Gaussian noise to get $x_n$.

$$x_n = Wz_n + \mu + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

- **What is the marginal distribution of $x_n$?**



# Fitting Probabilistic PCA with MLE

- **Maximum Likelihood Estimation (MLE) of $W$:** Tipping and Bishop (1999) showed that,

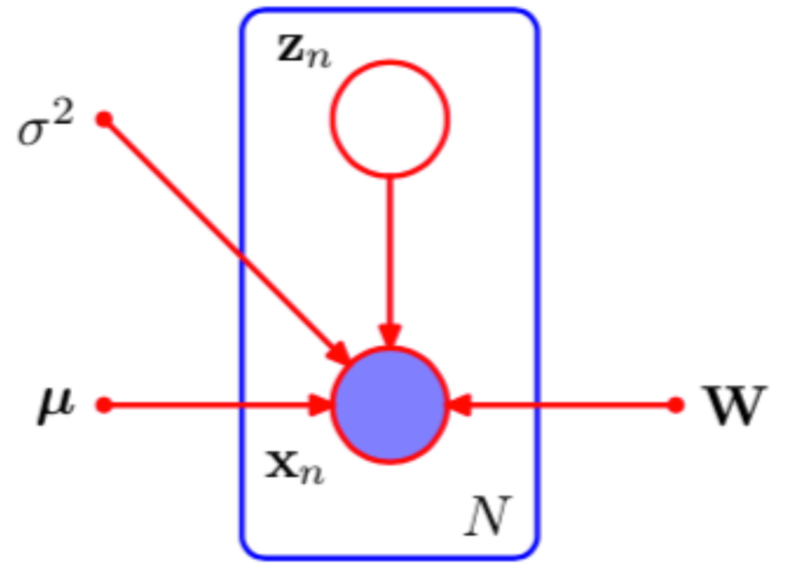
$$\hat{W} = C(\Lambda - \sigma^2 I)^{1/2} R.$$

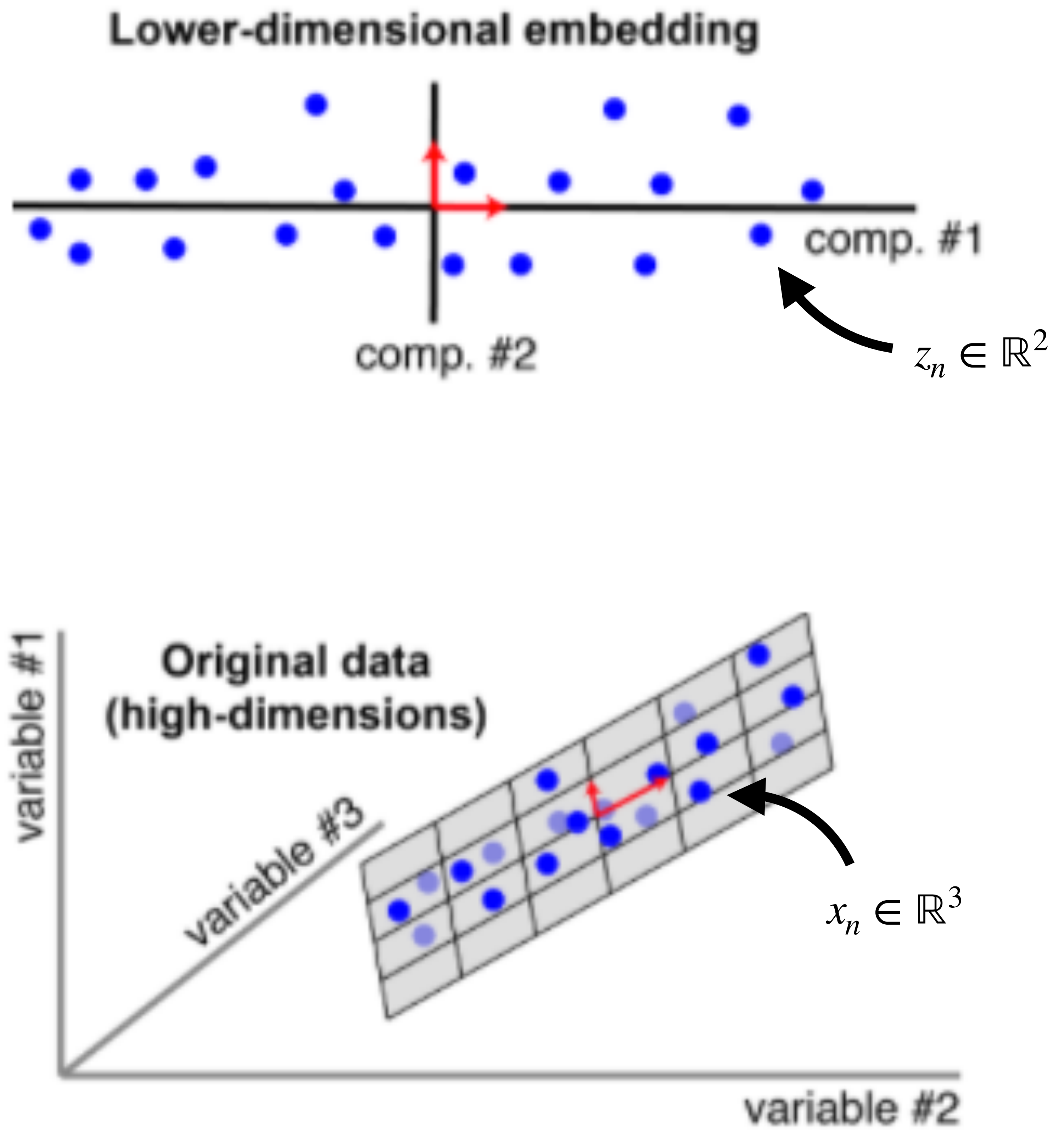
  where $C$ are eigenvectors of the covariance matrix (principal components), $\Lambda$ is a diagonal matrix with the corresponding eigenvalues, and $R$ is an arbitrary orthogonal matrix.

- As $\sigma^2 \to 0$ we recover the PCA solution, up to rotation.

- Alternatively, we can estimate $\sigma^2$ via MLE too:

$$\hat{\sigma}^2 = \frac{1}{D-K} \sum_{d=K+1}^{D} \lambda_d.$$

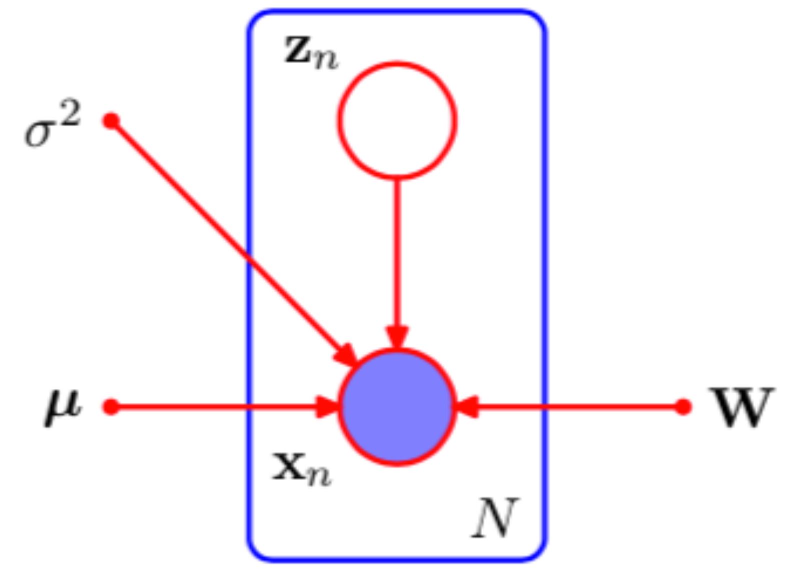


# Probabilistic PCA: Posterior on latent embedding

- What is the posterior distribution $p(z_n \mid x_n)$? By Bayes' rule:

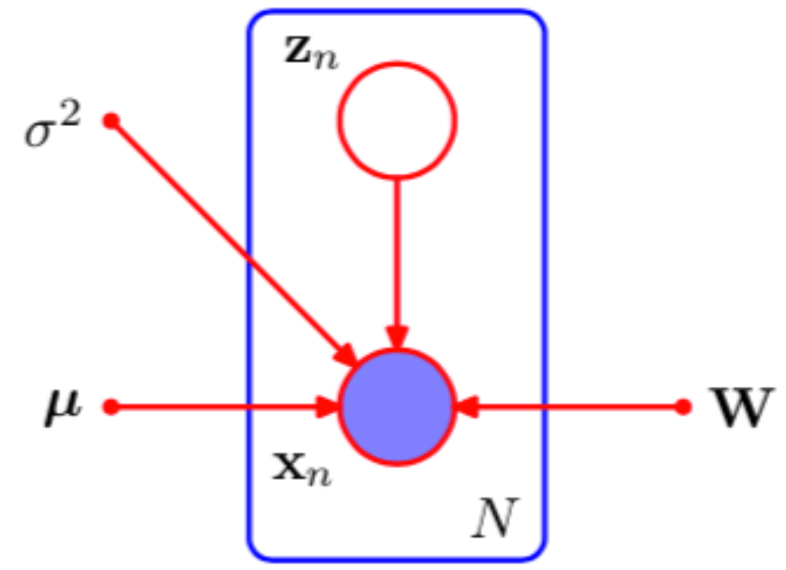
$$p(z_n \mid x_n) = \frac{p(z_n, x_n)}{p(x_n)} = \frac{p(z_n)\, p(x_n \mid z_n)}{p(x_n)} \propto p(z_n)\, p(x_n \mid z_n).$$



# Probabilistic PCA: Posterior on latent embedding

- What is the posterior distribution $p(z_n \mid x_n)$? By Bayes' rule:

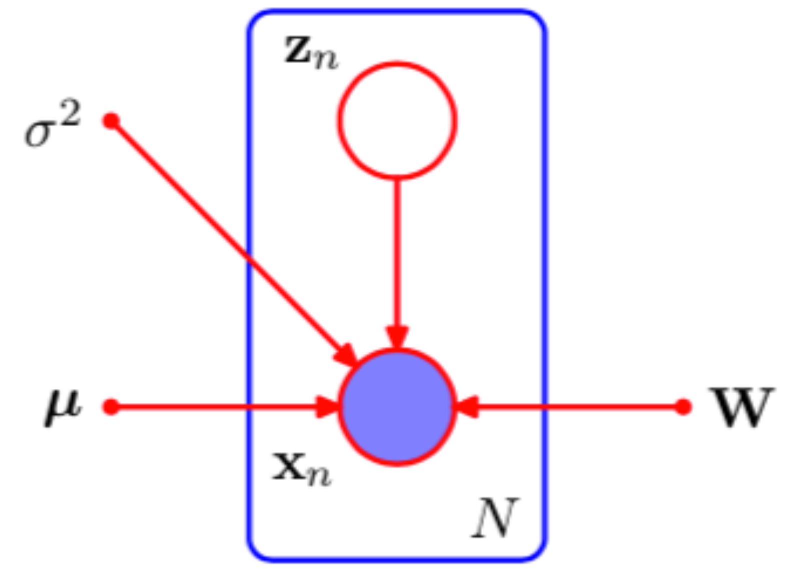
$$p(z_n \mid x_n) = \frac{p(z_n, x_n)}{p(x_n)} = \frac{p(z_n)\, p(x_n \mid z_n)}{p(x_n)} \propto p(z_n)\, p(x_n \mid z_n).$$



# Expectation-Maximization for Probabilistic PCA

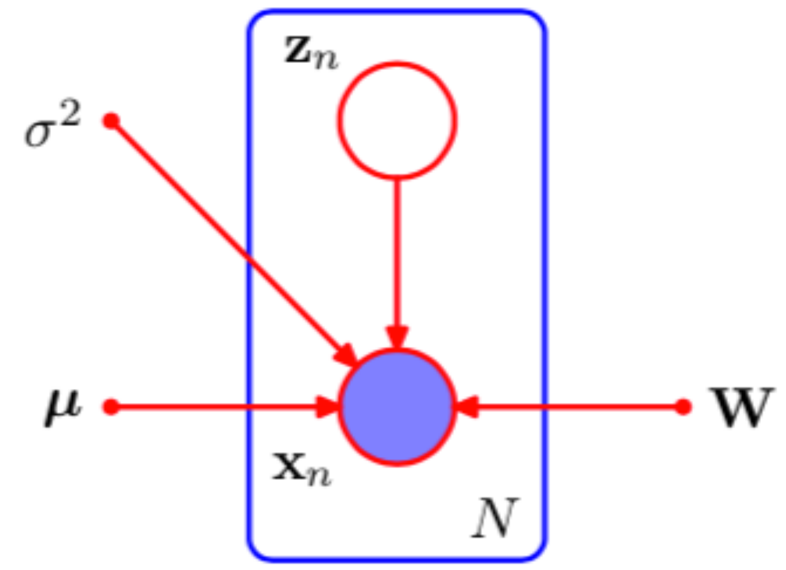
- Iterate between "updating $z_n$" and "updating the parameters $W, \mu, \sigma^2$."
- **E step**: Compute $p(z_n \mid x_n)$ using current parameters.
- **M step:** Find parameters that maximize expected log probability:

$$\mathscr{L}(W, \mu, \sigma^2) = \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n)}\left[\log p(z_n, x_n; W, \mu, \sigma^2)\right].$$



# Expectation-Maximization for Probabilistic PCA

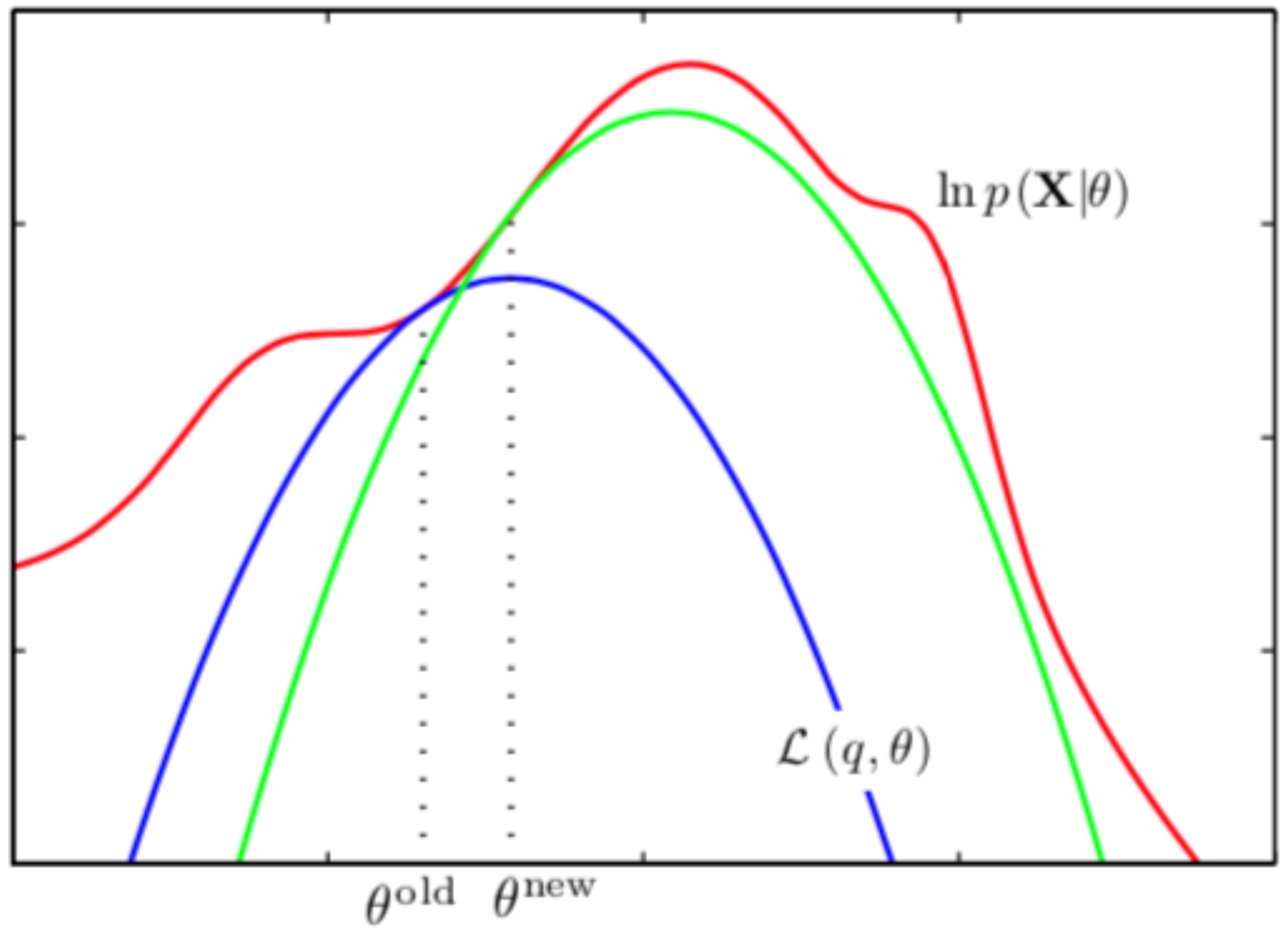
$$\mathscr{L}(W, \mu, \sigma^2) = \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n)} \left[\log p(z_n, x_n; W, \mu, \sigma^2)\right].$$



# Why does EM work?

- It maximizes a lower bound on the **marginal likelihood:**

$$\log p(x;\theta) = \sum_{n=1}^{N} \log p(x_n;\theta) = \sum_{n=1}^{N} \log \int p(x_n, z_n;\theta)\, \mathrm{d}z_n \ldots$$



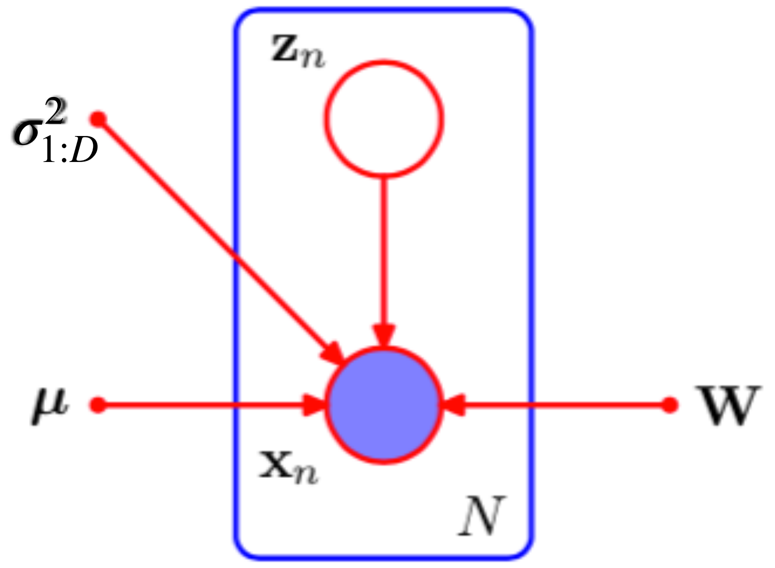
# Factor Analysis

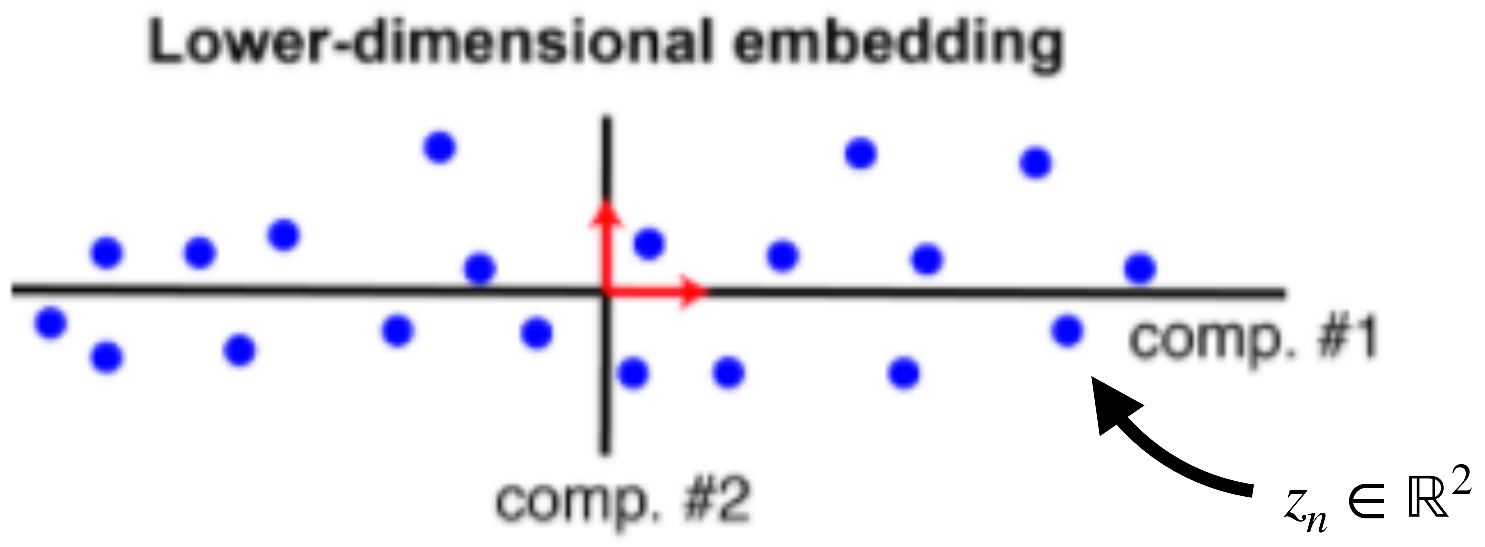
- **Same as Probabilistic PCA but with per-dimension noise!**

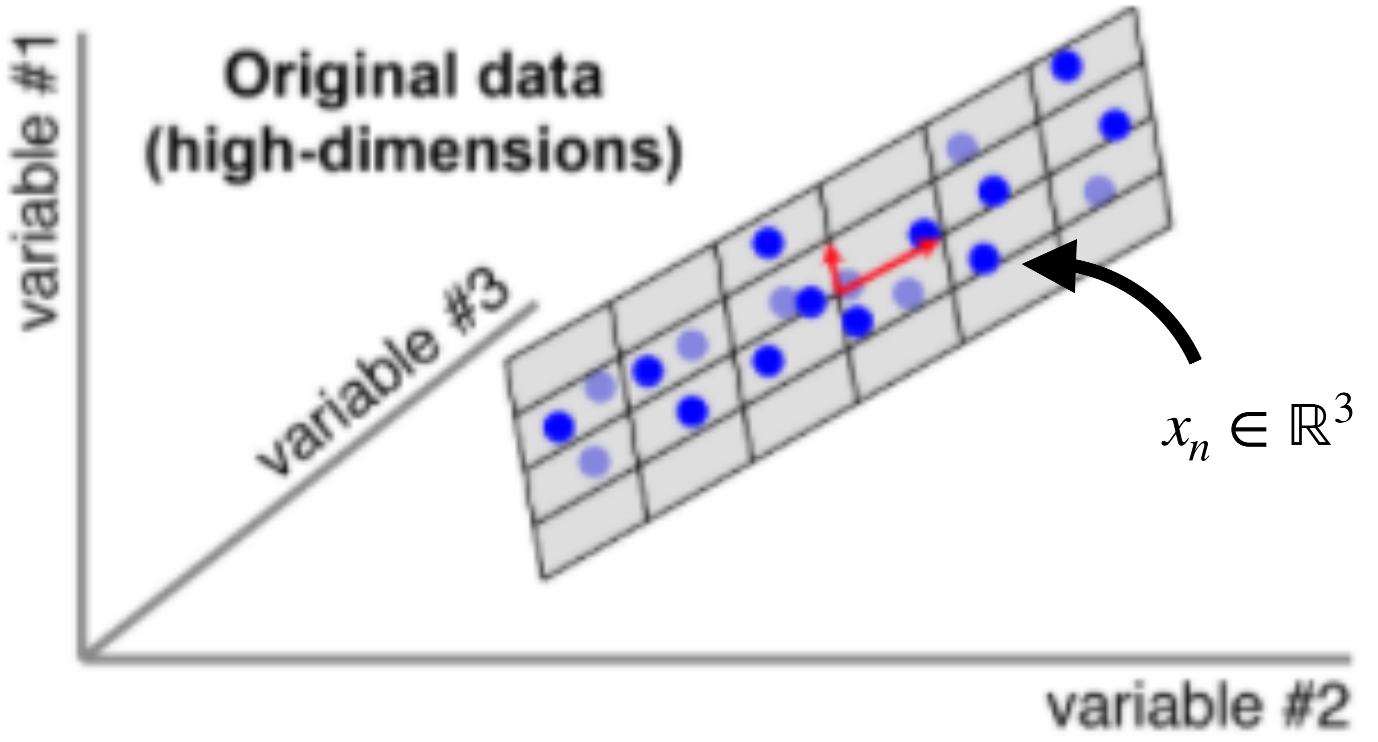
$$z_n \sim \mathcal{N}(0, I)$$
$$x_n = Wz_n + \mu + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2))$$

Equivalently

$$z_n \sim \mathcal{N}(0, I)$$
$$x_{nd} \sim \mathcal{N}(w_d^\top z_n + \mu_d, \sigma_d^2) \quad \text{for } d = 1, \ldots, D$$



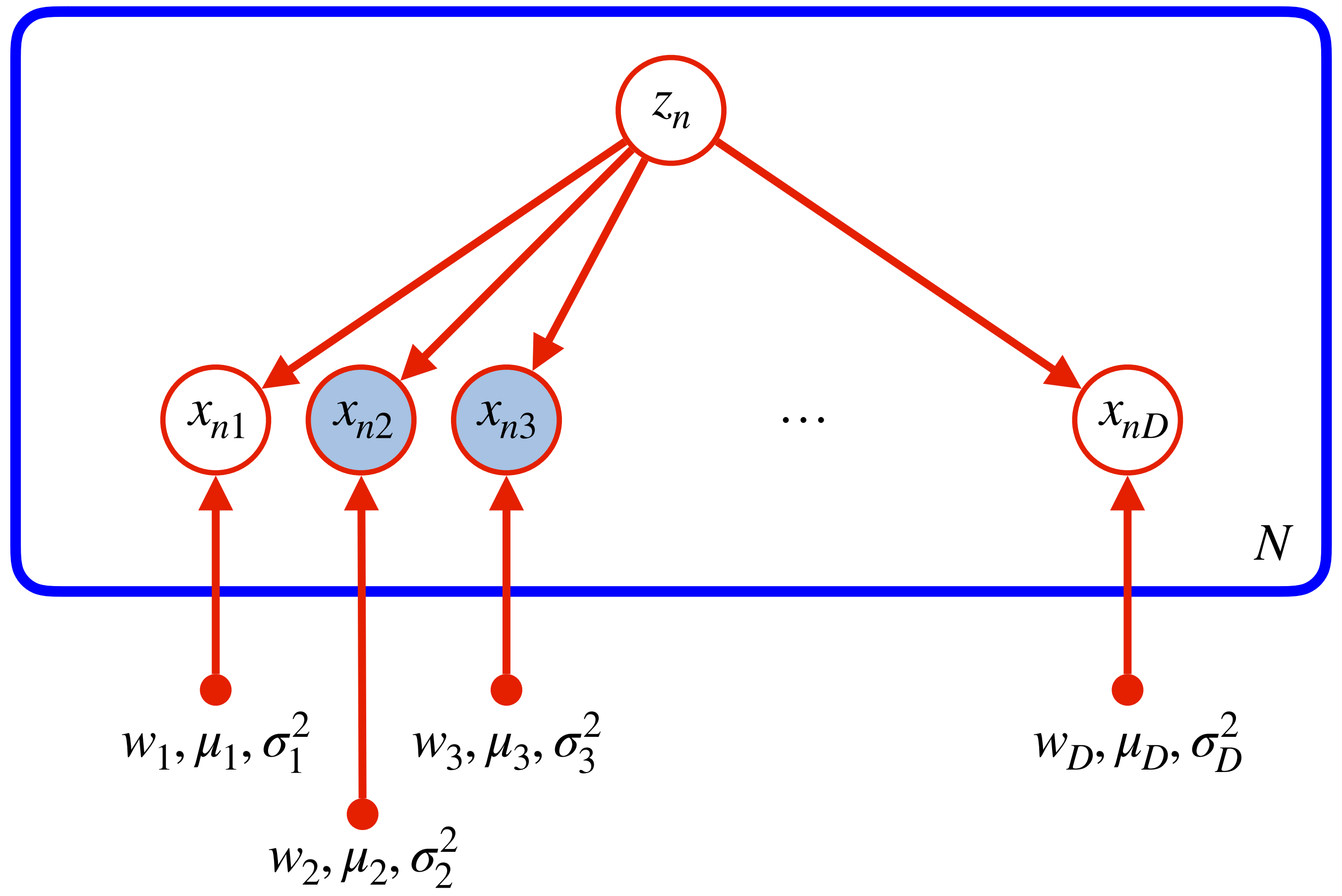



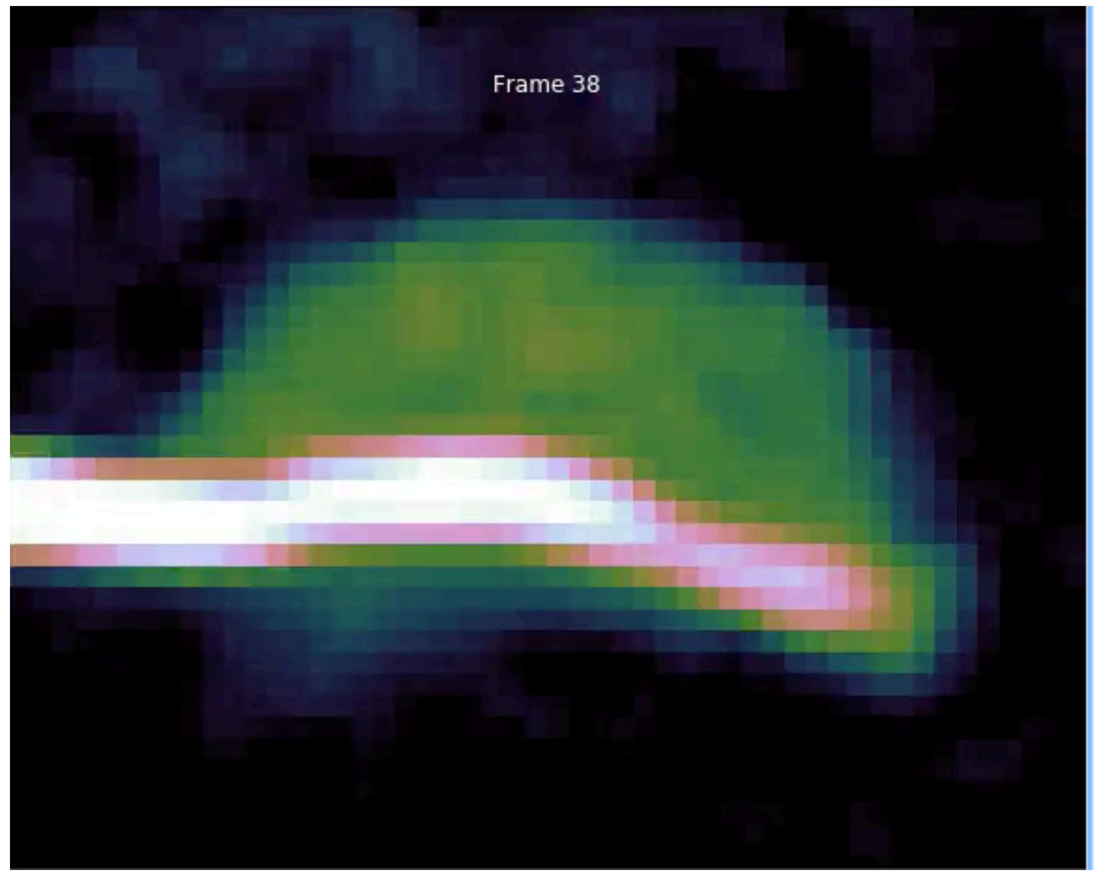



# Real World Example

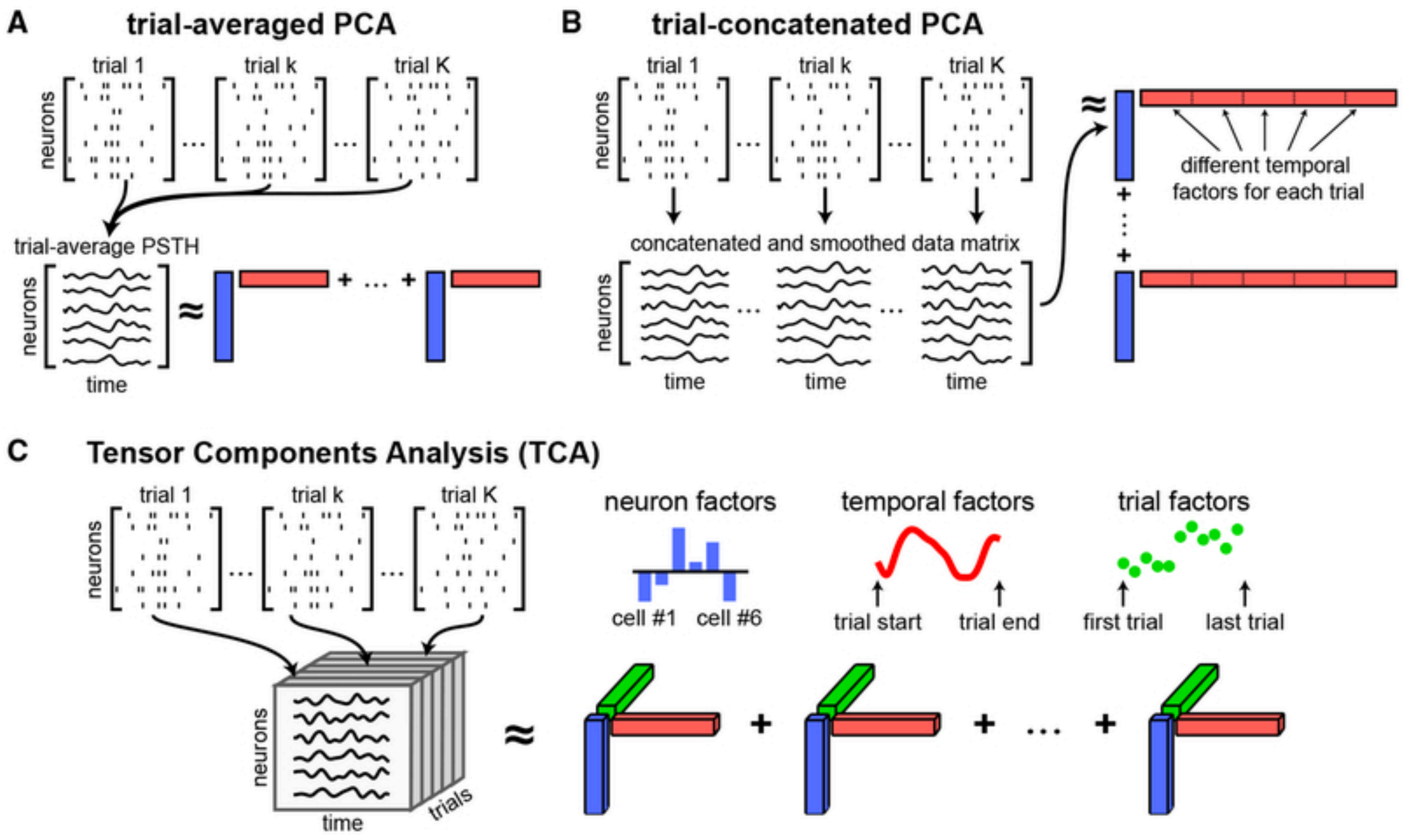
# Factor Analysis with Missing Data

- What is the posterior distribution $p(z_n \mid \{x_{nd} : x_{nd} \text{ not missing}\})$?



# Beyond Factor Analysis: Tensors



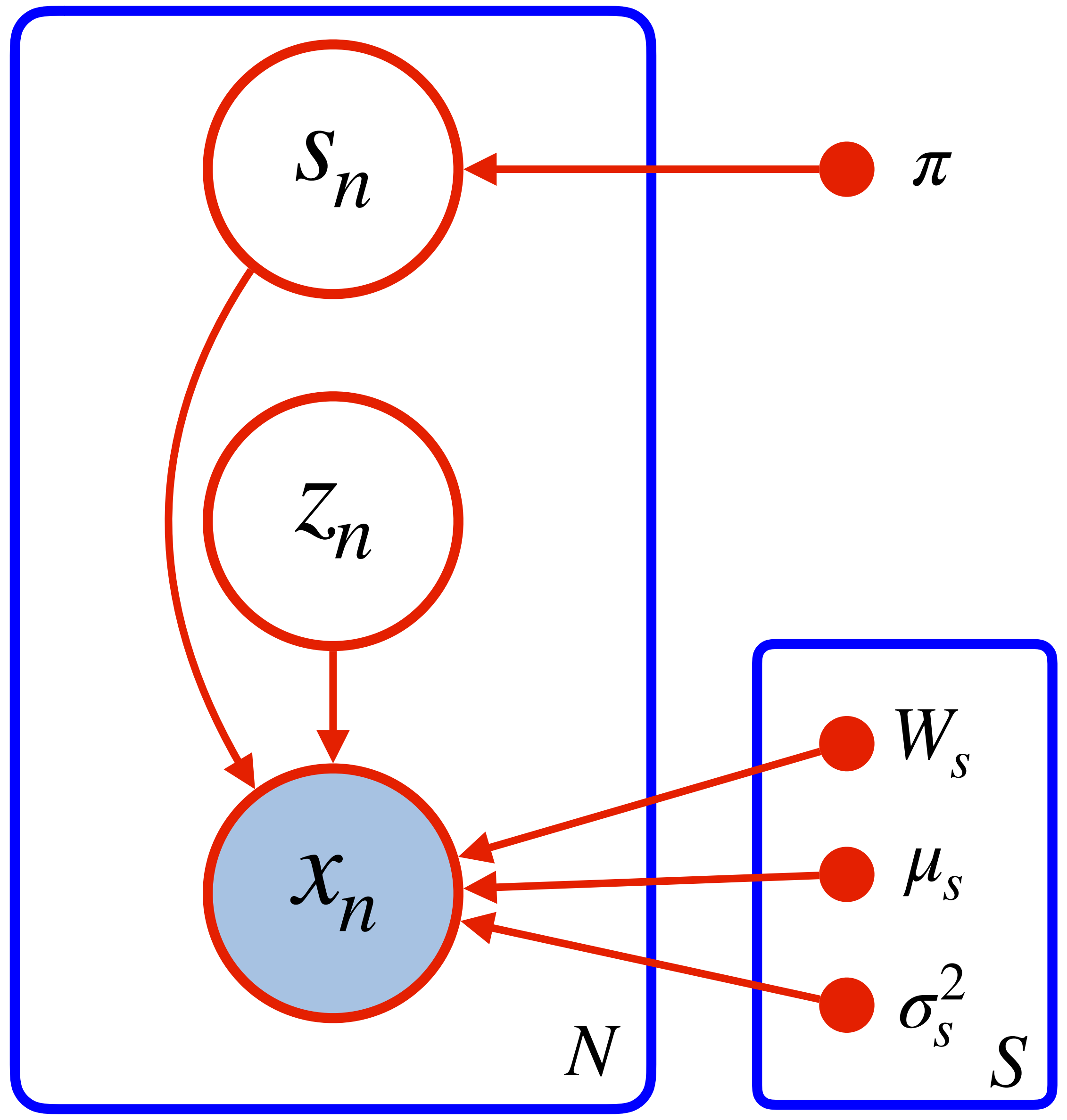

**Williams et al (2018)**

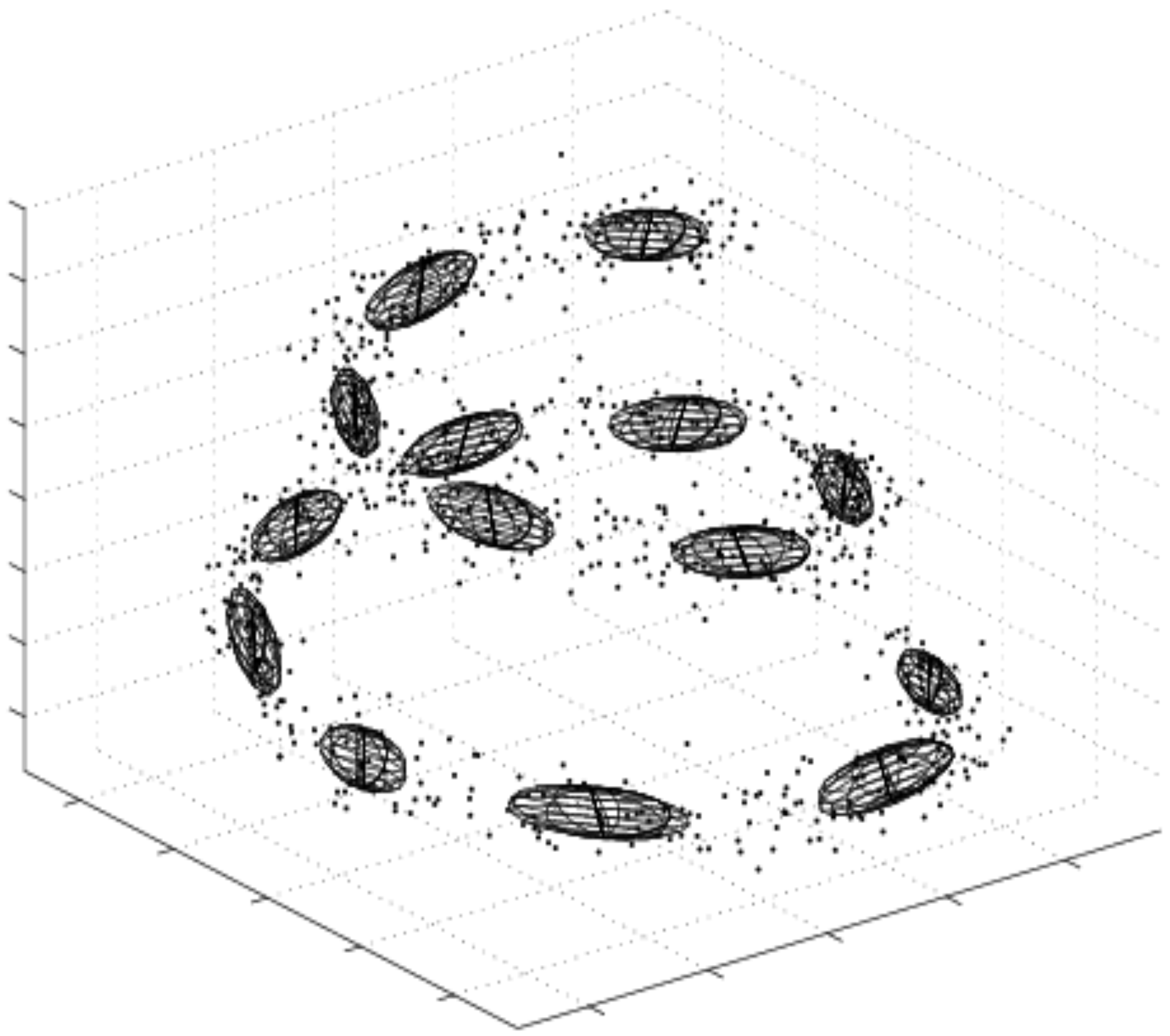
# Beyond Factor Analysis: Mixtures of Factor Models

Discrete
Latent States

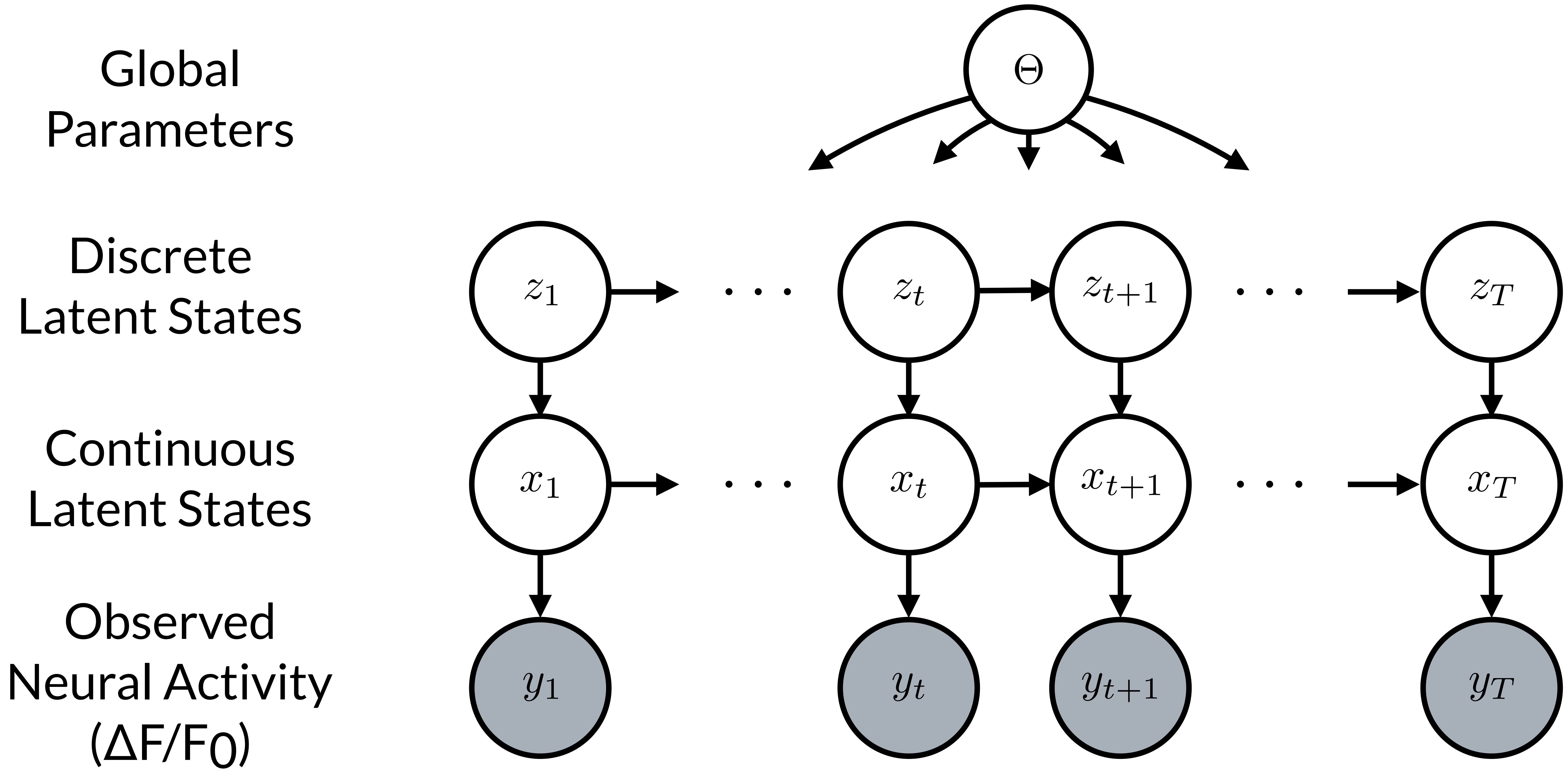
Continuous
Embedding

Observed
Data



# Beyond Factor Analysis: Time Series



# Beyond Factor Analysis: Nonlinear Latent Variable Models