# Statistical models for neural encoding

## Part 1: discrete-time models

Liam Paninski

Gatsby Computational Neuroscience Unit
University College London
http://www.gatsby.ucl.ac.uk/∼liam
*liam@gatsby.ucl.ac.uk*

November 2, 2004

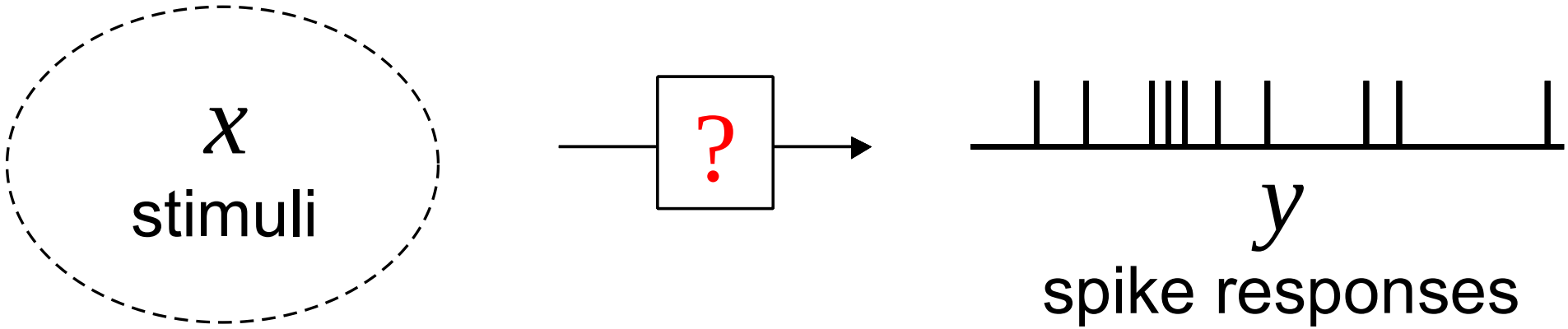
# Goals

Introduce some statistical methods for "reading the brain"

Give an intuitive feel for when these techniques will work (and why)

Discuss novel insights about neural coding obtained via statistical modeling efforts

# The neural code




Input-output relationship between

- External observables $x$ (sensory stimuli, motor responses...)
- Neural variables $y$ (spike trains, population activity...)

Probabilistic formulation: $p(y|x)$

# Basic goal

...reading the brain.

Fundamental question: how to estimate $p(y|x)$ from experimental data?

General problem is too hard — not enough data, too many possible inputs $x$

e.g., responses to natural scenes in IT cortex

# Avoiding the curse of insufficient data

Many approaches to make problem tractable:

**1**: Estimate some functional $f(p)$ instead

e.g., NIPS03 workshop on estimation of information-theoretic quantities

**2**: Select stimuli more efficiently

e.g., (Foldiak, 2001; Machens, 2002; Paninski, 2003b)

**3: Fit a model with small number of parameters**

# Neural encoding models

Good news: many methods available. Well-understood, easy to use.

Bad news: none are perfect.

Variety of approaches: different tools for different situations

# Encoding models

Main theme: want model to be flexible but not overly so

Flexibility vs. "fittability"

# Encoding models

We want $p(spike|\vec{x})$

0-th idea: in 1-dimension or finite $x$: just take $\hat{p}(spike|\vec{x}) =$ fraction of $\vec{x}$ which led to spike

Very flexible ("nonparametric"), but quickly overwhelmed for large-D $\vec{x}$

# Additive models

First idea: model $p(spike|\vec{x})$ as linear in $\vec{x}$:

$$p(spike|\vec{x}) = \vec{k} \cdot \vec{x} + b.$$

Fit coefficients by linear regression:

$$\vec{k} = (E(\vec{x}^t \vec{x}))^{-1} E(\vec{x} \cdot spike)$$

— Easy, but too simple. Neurons often code nonlinearly; besides, leads to negative firing rate predictions.

# Aside: regression

$$\vec{k} = (E(\vec{x}^t \vec{x}))^{-1} E(\vec{x} \cdot spike)$$

Where does this come from?

Optimal least-squares problem: choose $\vec{k}$ to minimize

$$Err = E[(\vec{k} \cdot \vec{x} - y)^2].$$

Take derivative w.r.t. $\vec{k}$, set to zero (exercise).

Is this automatically a minimum? If so, is it automatically the only minimum?

# Additive models

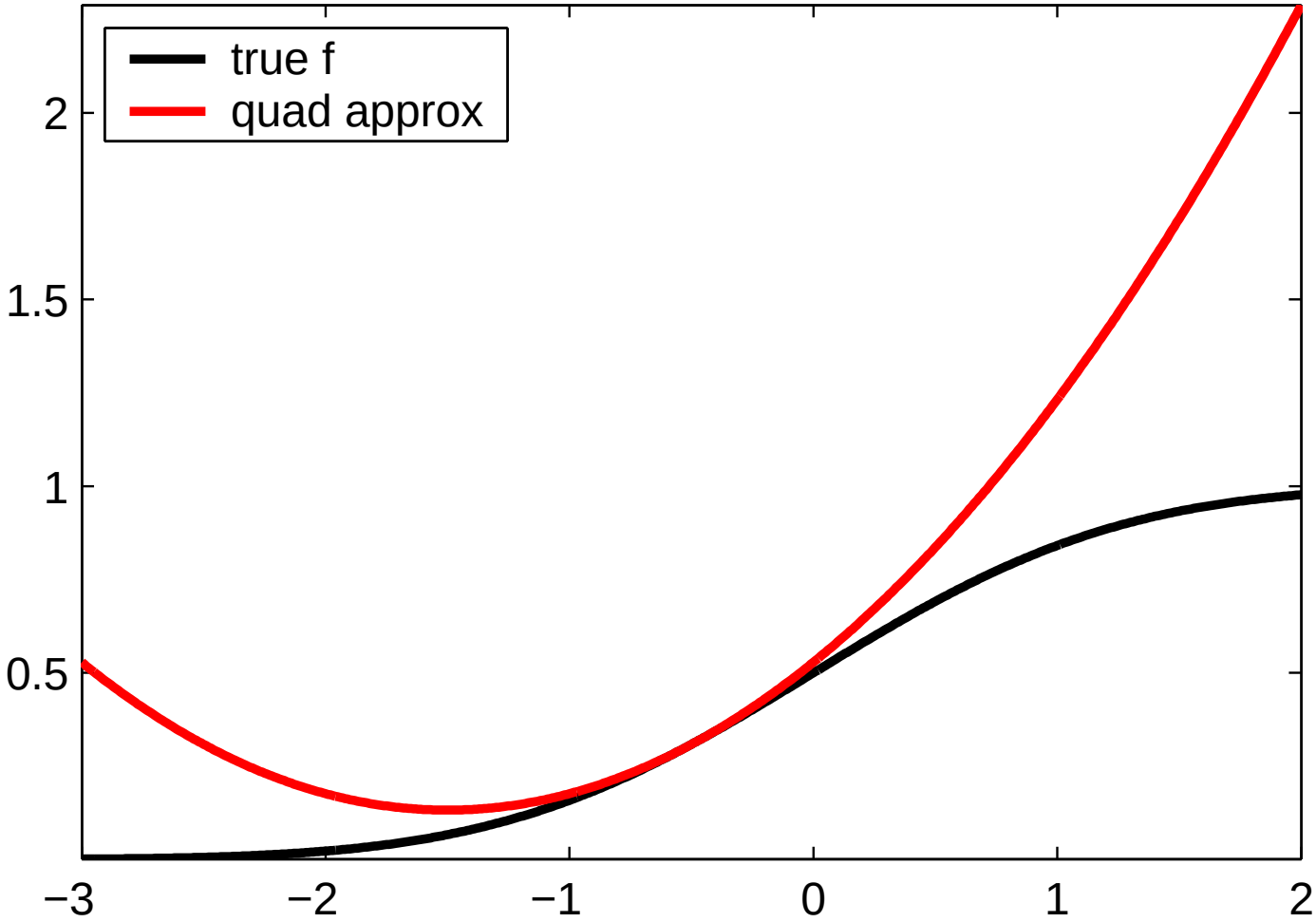
Next: polynomial (Volterra/Wiener) expansion.

$$p(spike|\vec{x}) = k_0 + \vec{k}_1 \cdot \vec{x} + \vec{x}^t K_2 \vec{x} + K_3 \vec{x}^3 + \ldots$$

- can be fit with regression methods...

# Volterra models

... but still not great. Quadratic approximation often poor; going beyond 2nd order requires much data



# Additive models

Natural direction: more general expansions.

$$p(spike|\vec{x}) = \sum_i k_i \mathcal{F}_i(\vec{x})$$

e.g., $\mathcal{F}_i$ = sigmoids, Gaussian bumps

Can still be fit by regression methods.

— not a bad idea, but not very commonly used. Requires:

1) good functions $\mathcal{F}_i$ to try

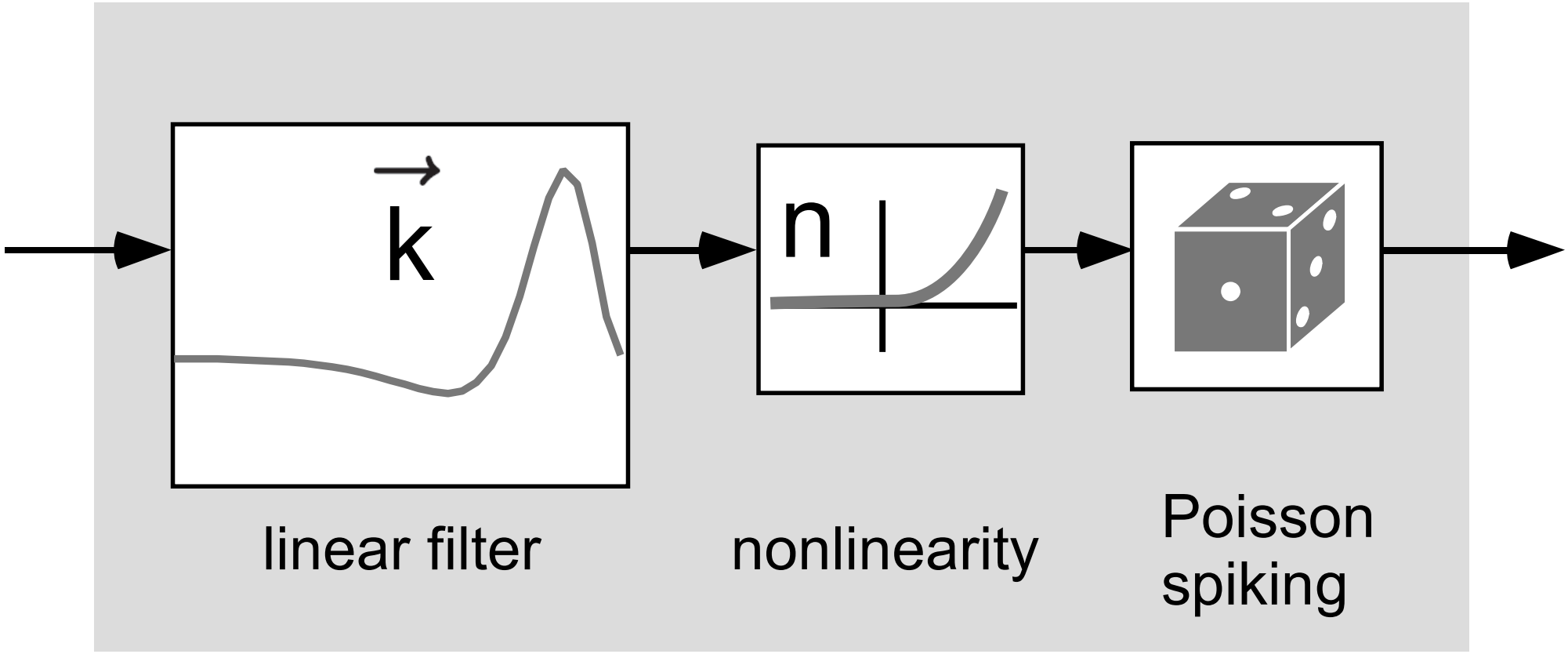
2) good way to control complexity (avoid overfitting)

# Different approach: Cascade models

Idea: dimensionality reduction

Try to pick out a few important linear filters $\{\vec{k}_i\}_{i \leq m}$ and then fit nonlinear models in this lower-D space

# LNP model



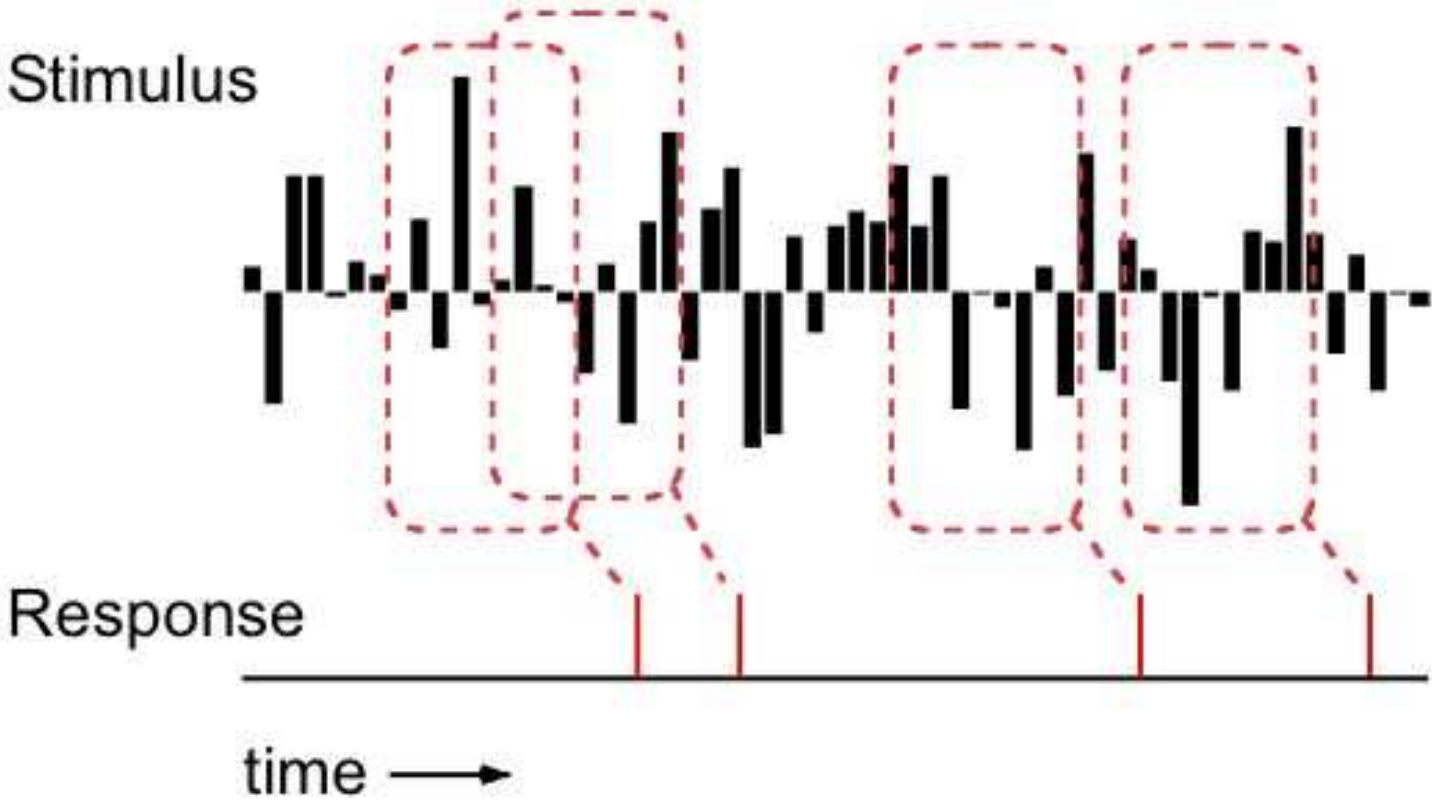

.

# LNP model

$\vec{k}$ could represent (quasilinear) presynaptic, dendritic filtering

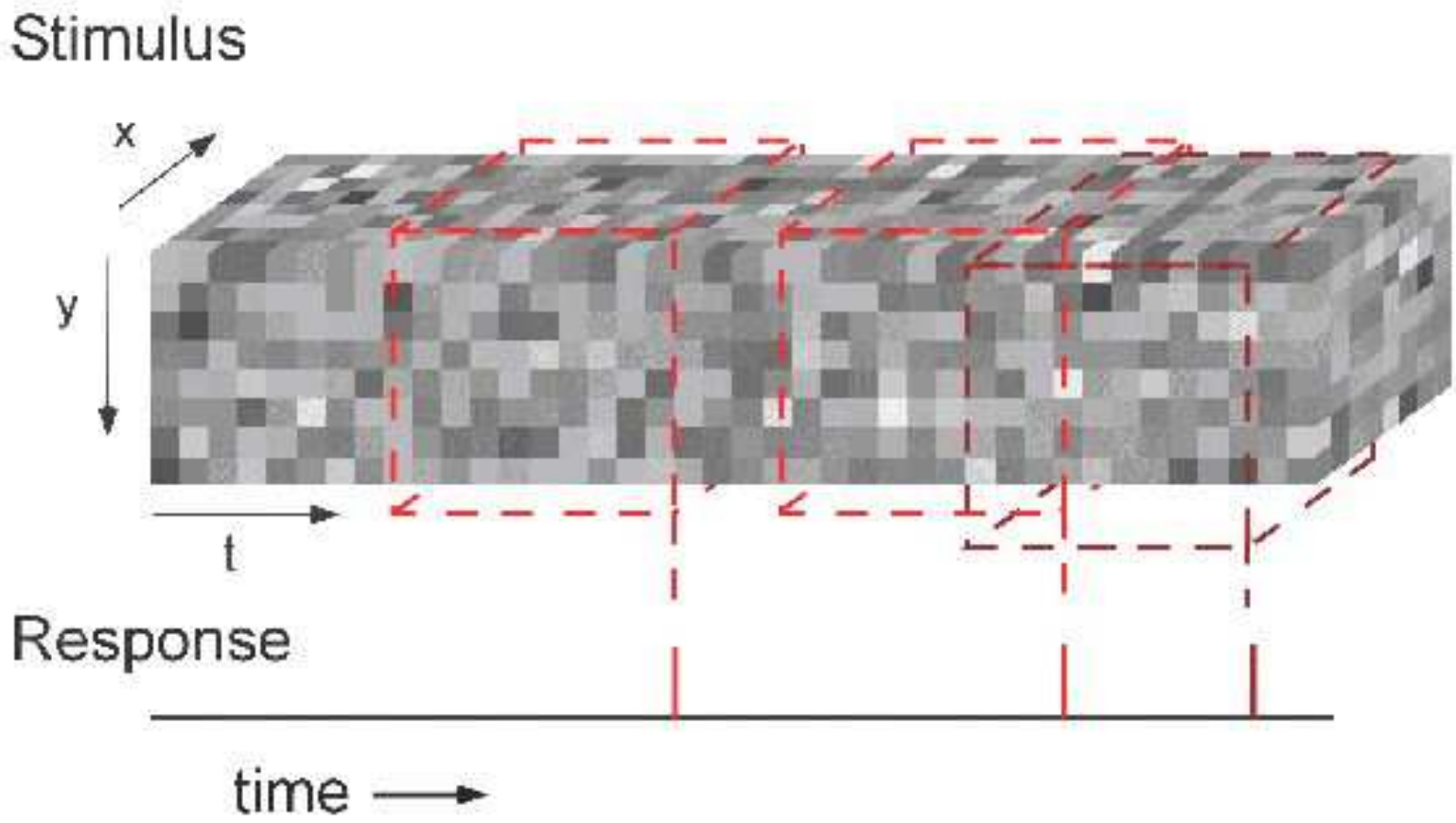
How to fit $\vec{k}$ and nonlinearity?

# Spike-triggered ensemble



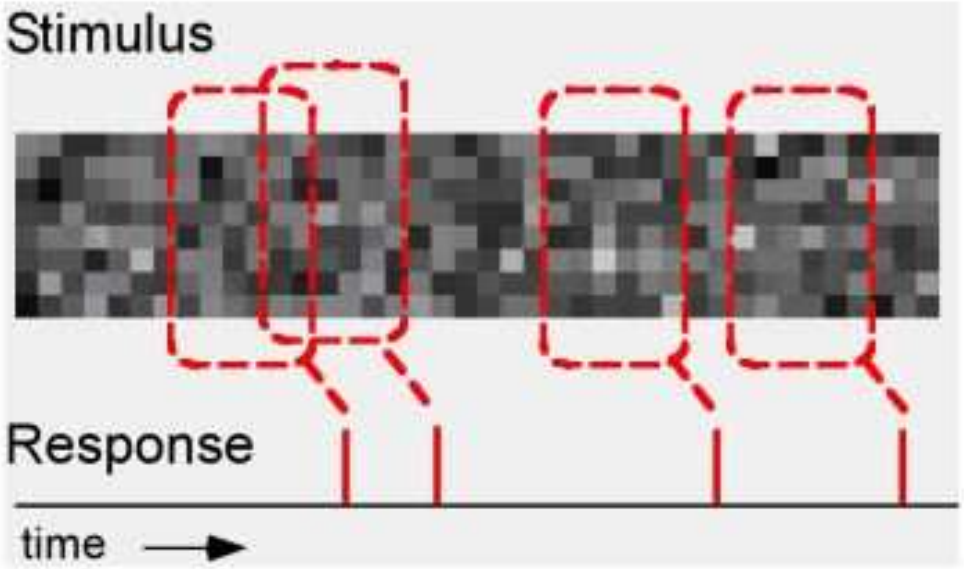

- 9-sample stimulus block

# Spike-triggered ensemble (3D stimulus)



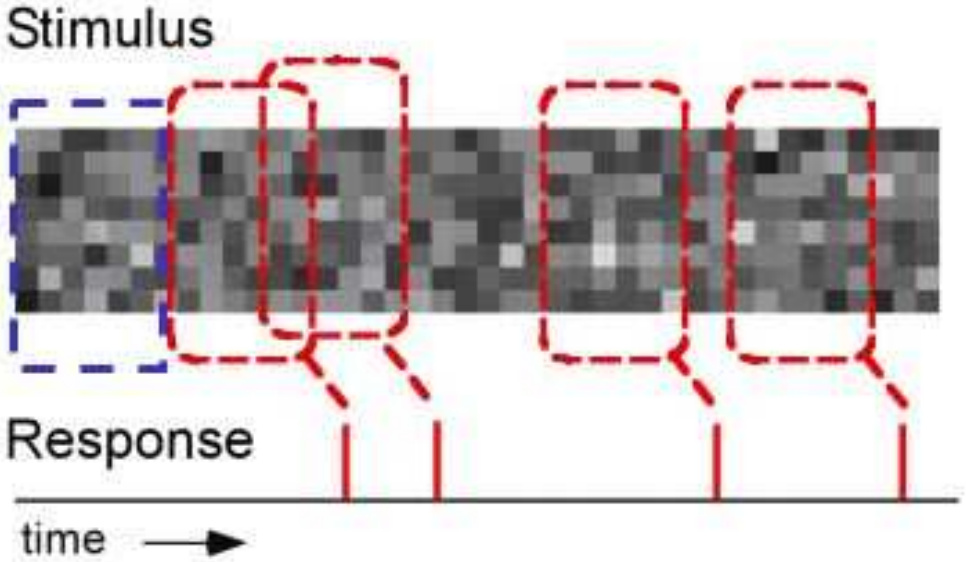

- 8 x 8 x 10 stimulus block

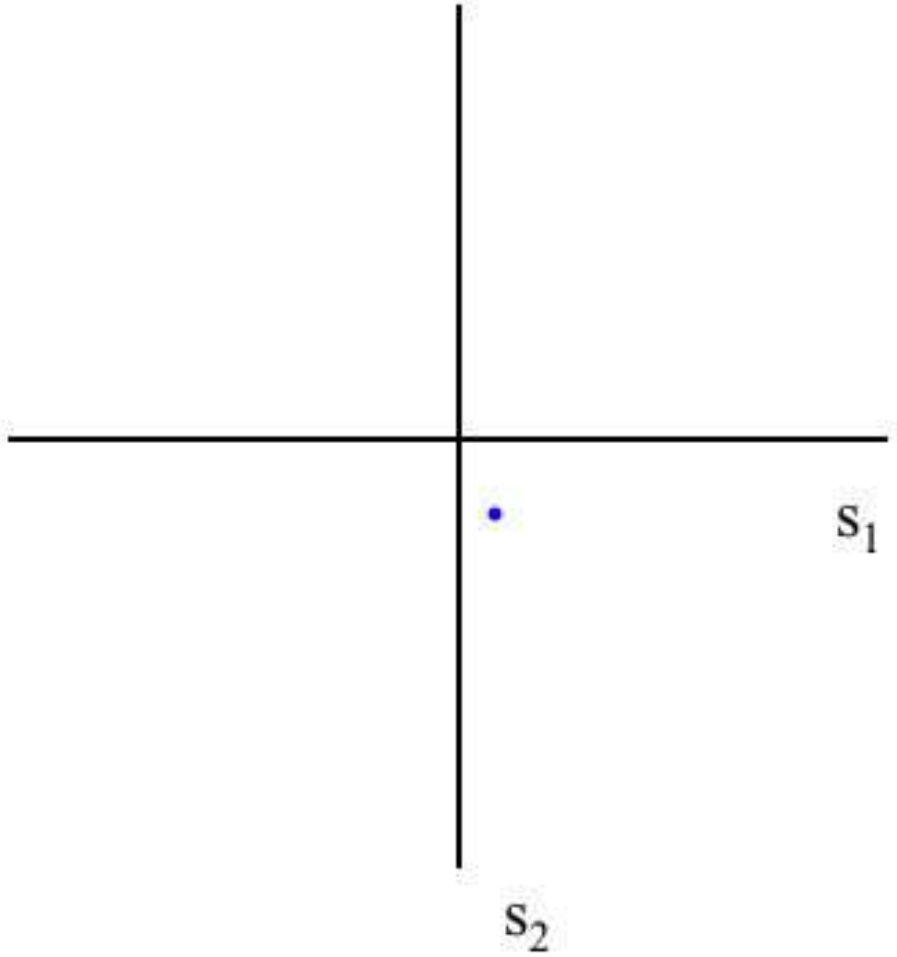
# 2D stimulus (flickering bars)




- 8 x 6 stimulus block
  = 48-dimensional vector

# Geometric picture



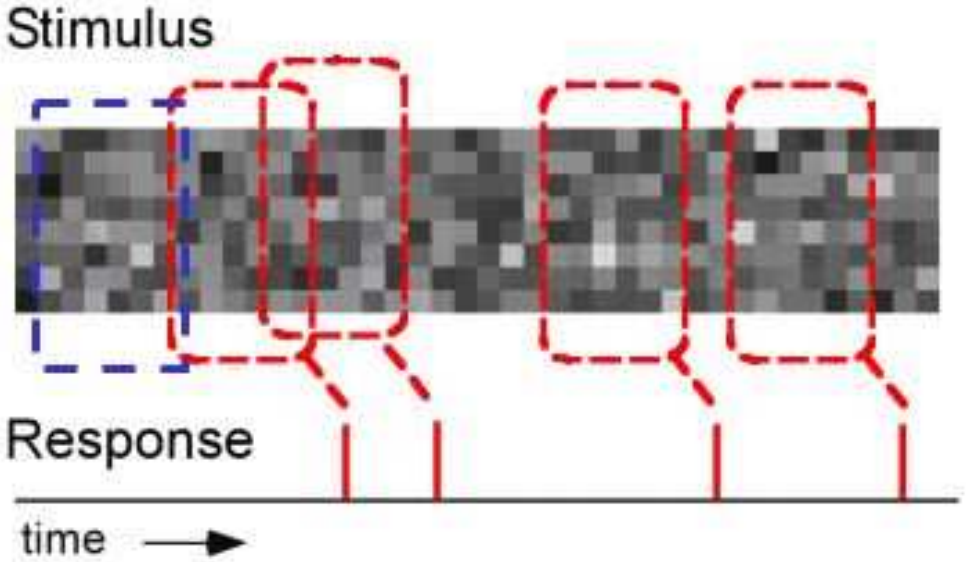



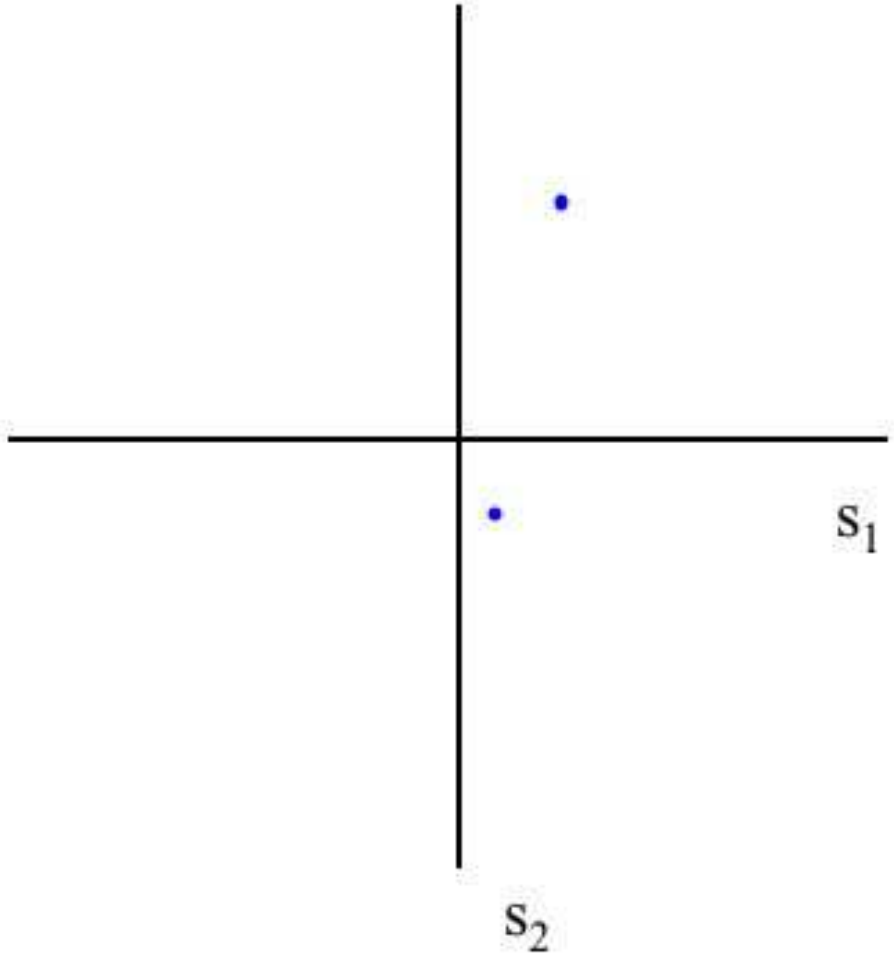

- 8 x 6 stimulus block
  = 48-dimensional vector

raw stimuli

spiking stimuli

# Geometric picture



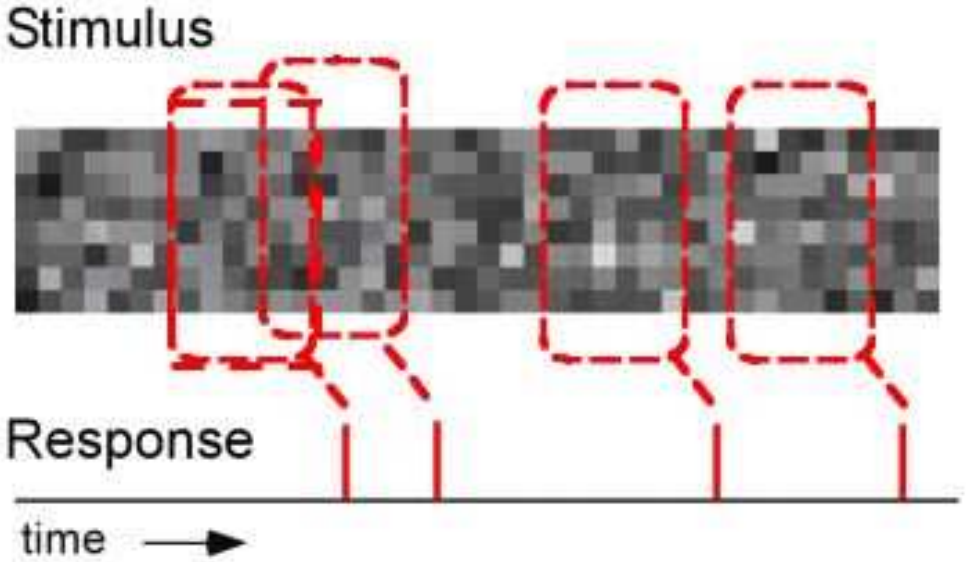



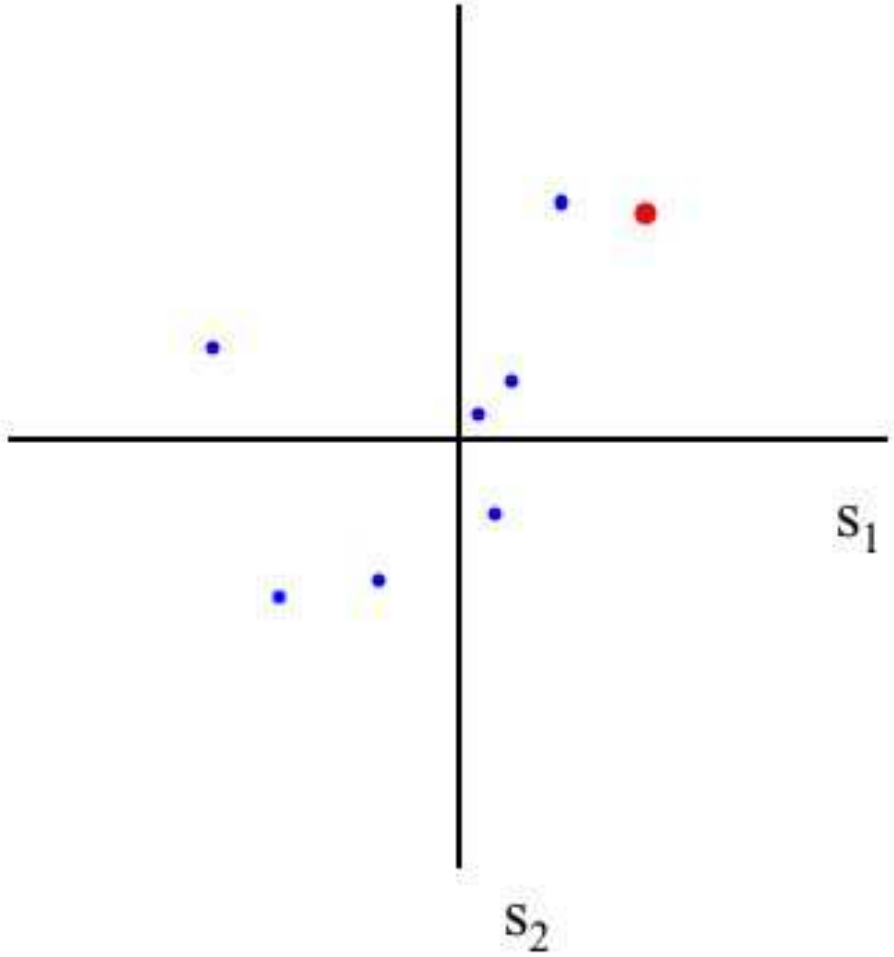

- 8 x 6 stimulus block
  = 48-dimensional vector

raw stimuli

spiking stimuli

# Geometric picture



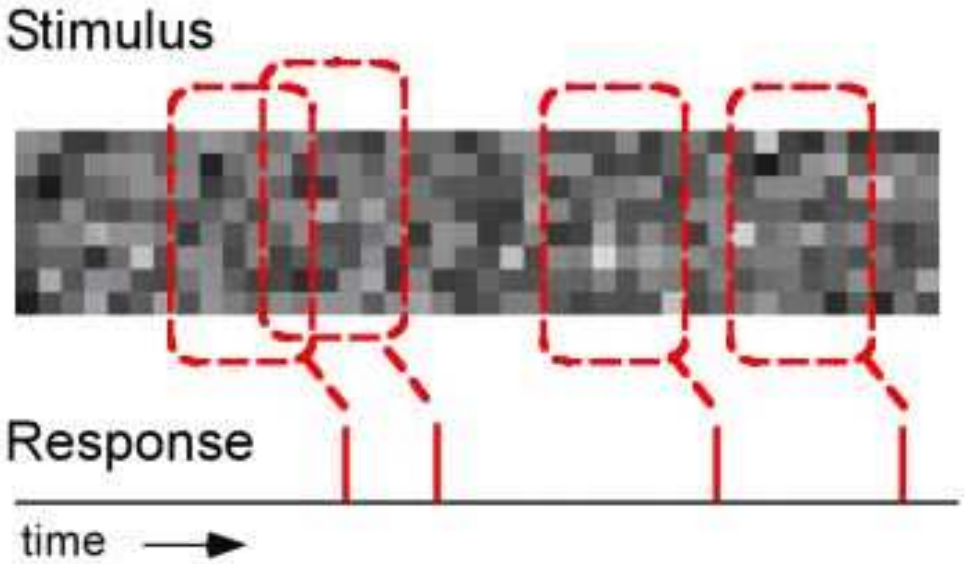



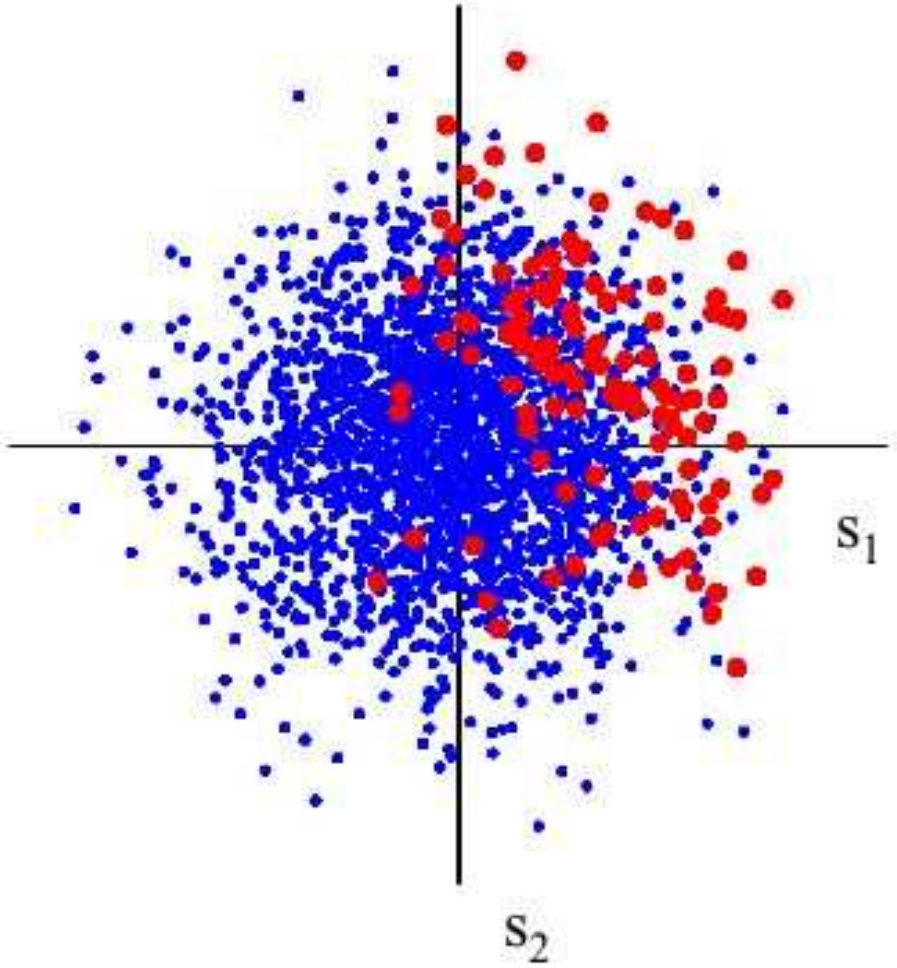

- 8 x 6 stimulus block
  = 48-dimensional vector

raw stimuli

spiking stimuli

# Geometric picture

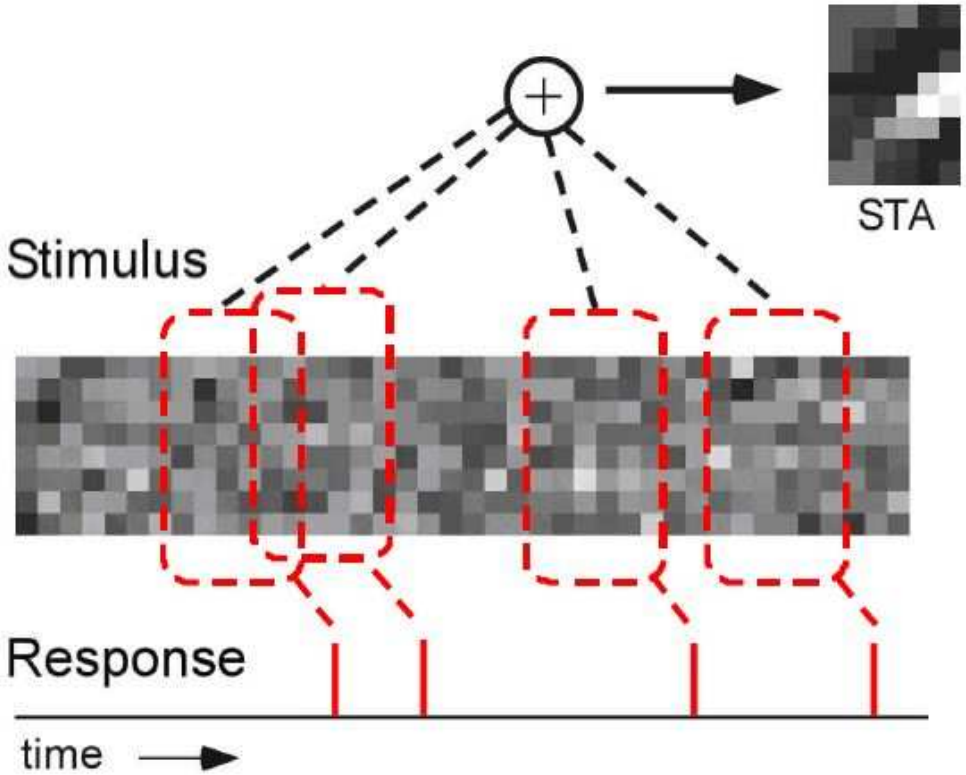


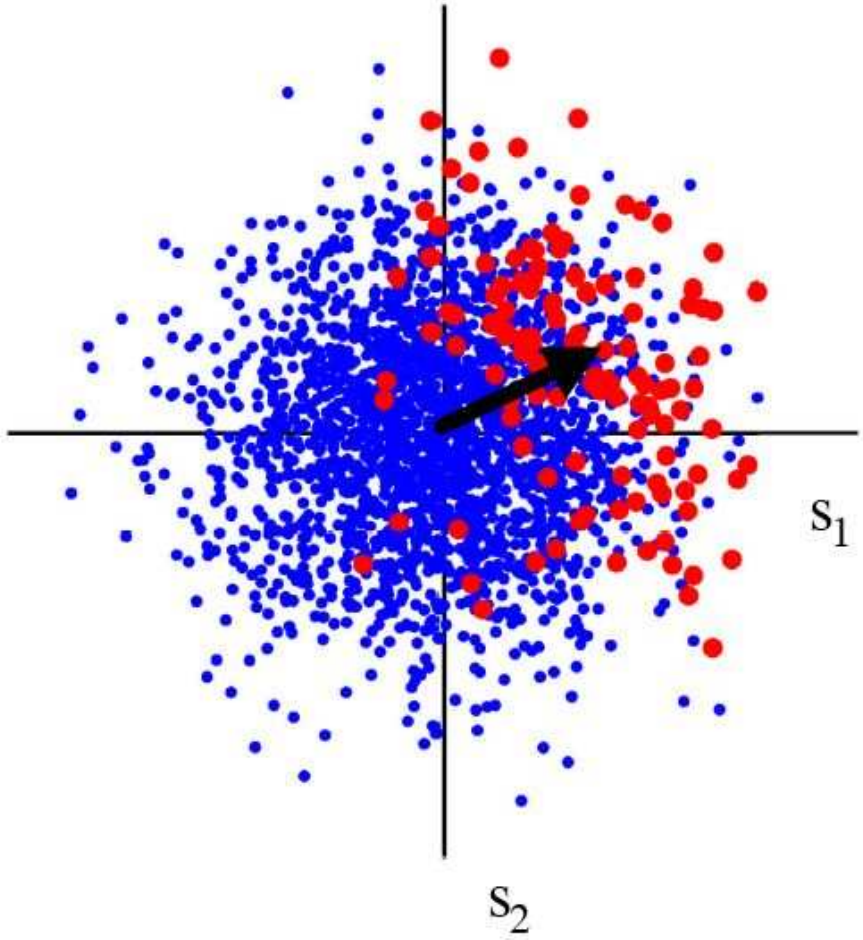




raw stimuli

spiking stimuli

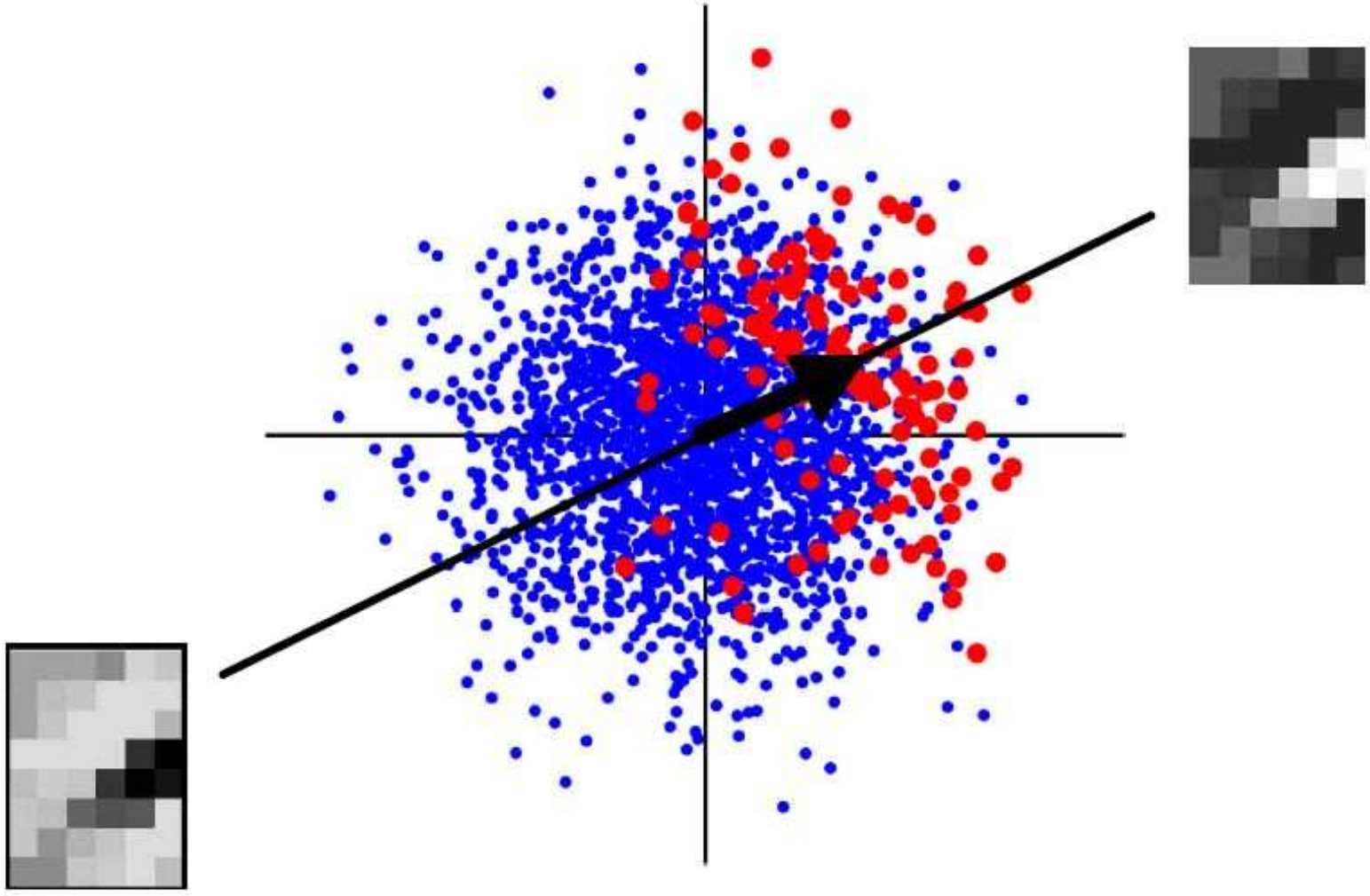
# Computing the STA




— just cross-correlate spikes and $\vec{x}$.

## STA defines a "direction" in stimulus space



STA is *unbiased* estimate of $\vec{k}$ if $p(stim)$ is radially symmetric: (Chichilnisky, 2001).

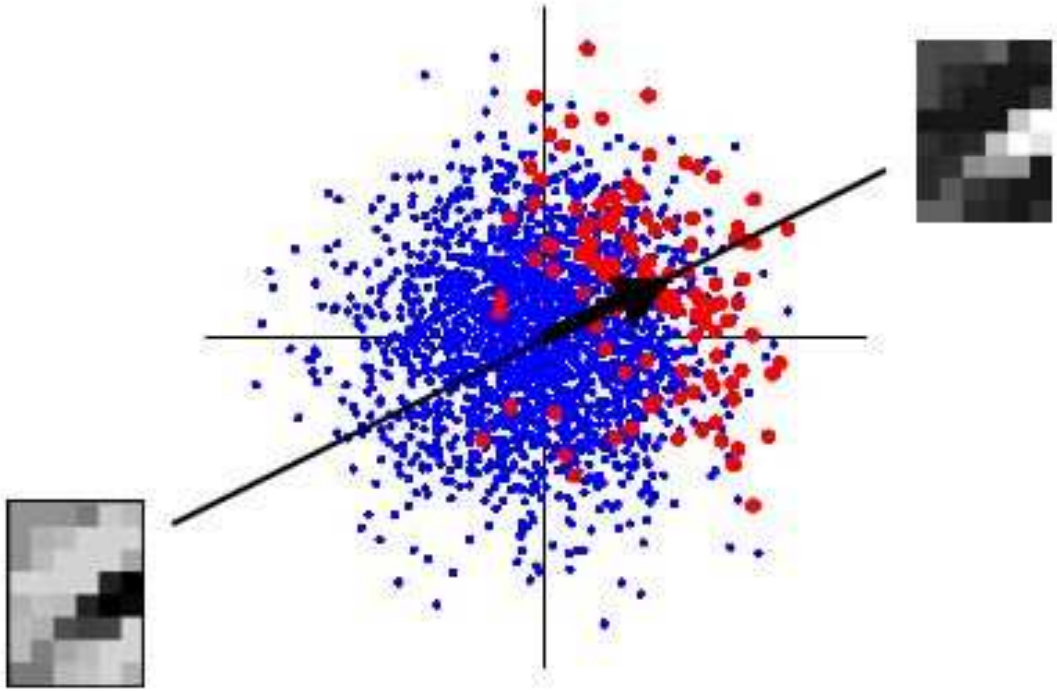
# Aside: bias and variance

In general, two kinds of estimation error to consider:

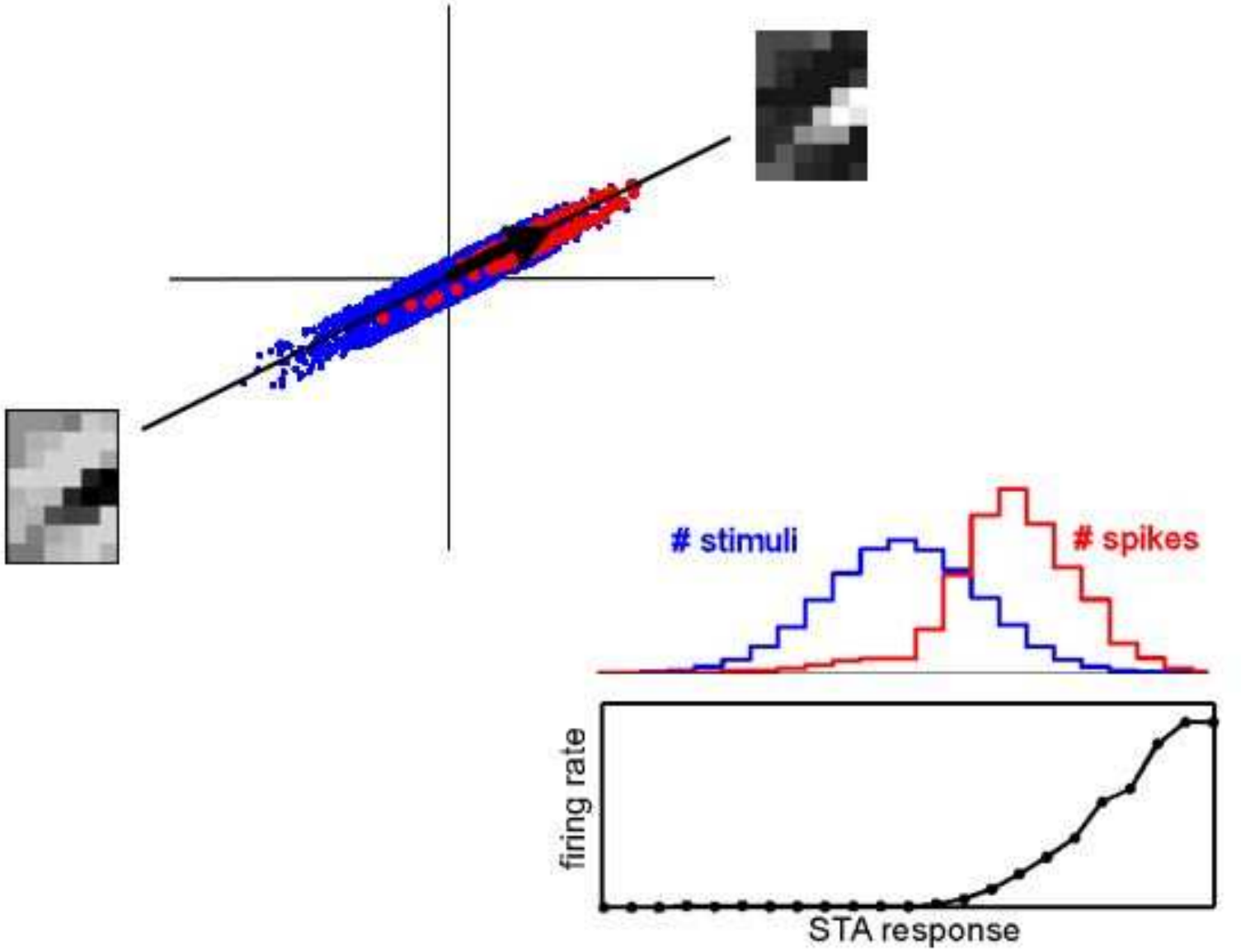
- **Bias**: average error $E(\hat{k}) - \vec{k}$; expected difference between estimate $\hat{k}$ and true $\vec{k}$
- **Variance**: spread around mean $E(\hat{k})$

— bias of STA is zero. But still need sufficient samples to make variance small (reduce noise in estimate).

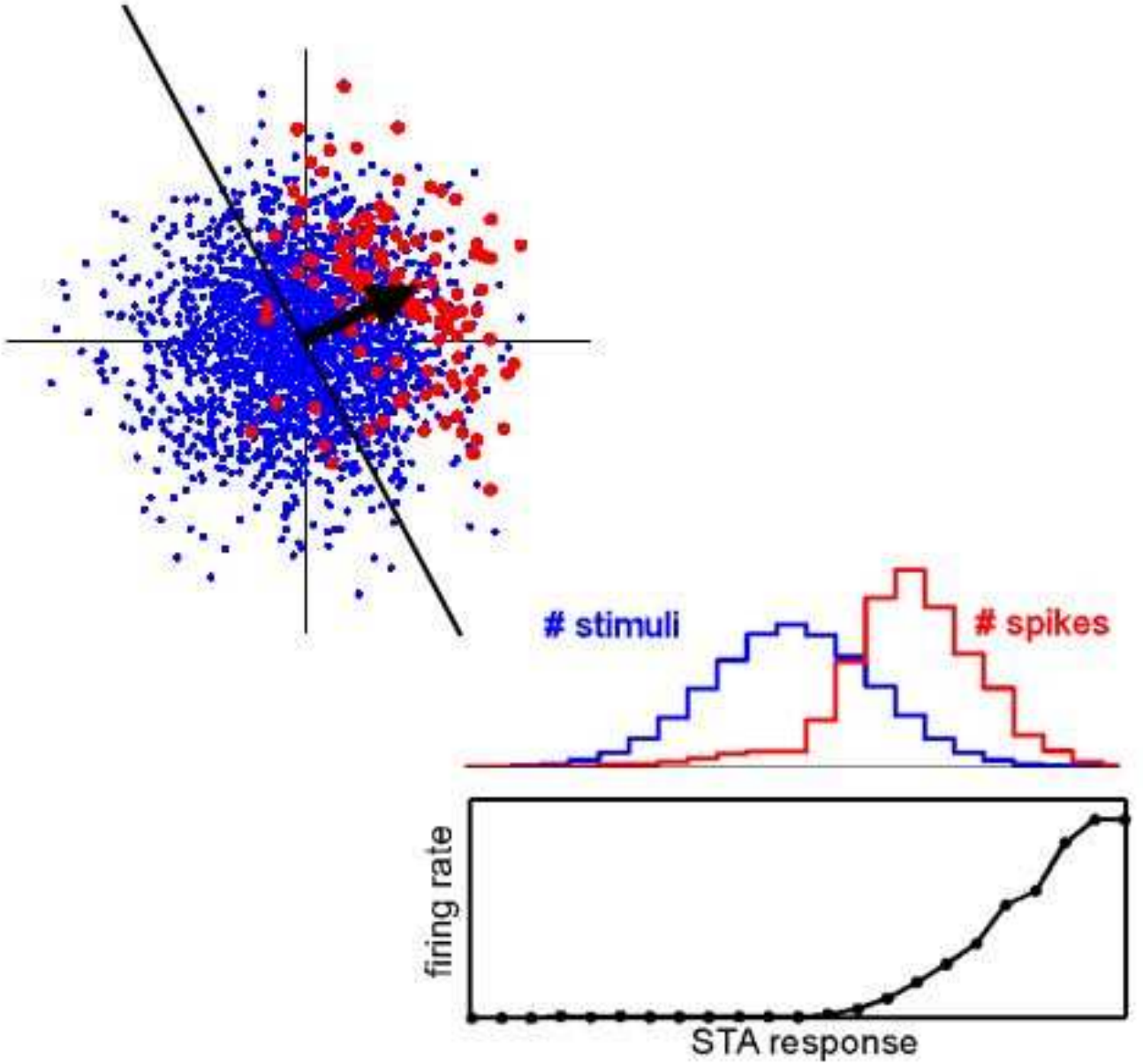
# Projecting onto the STA



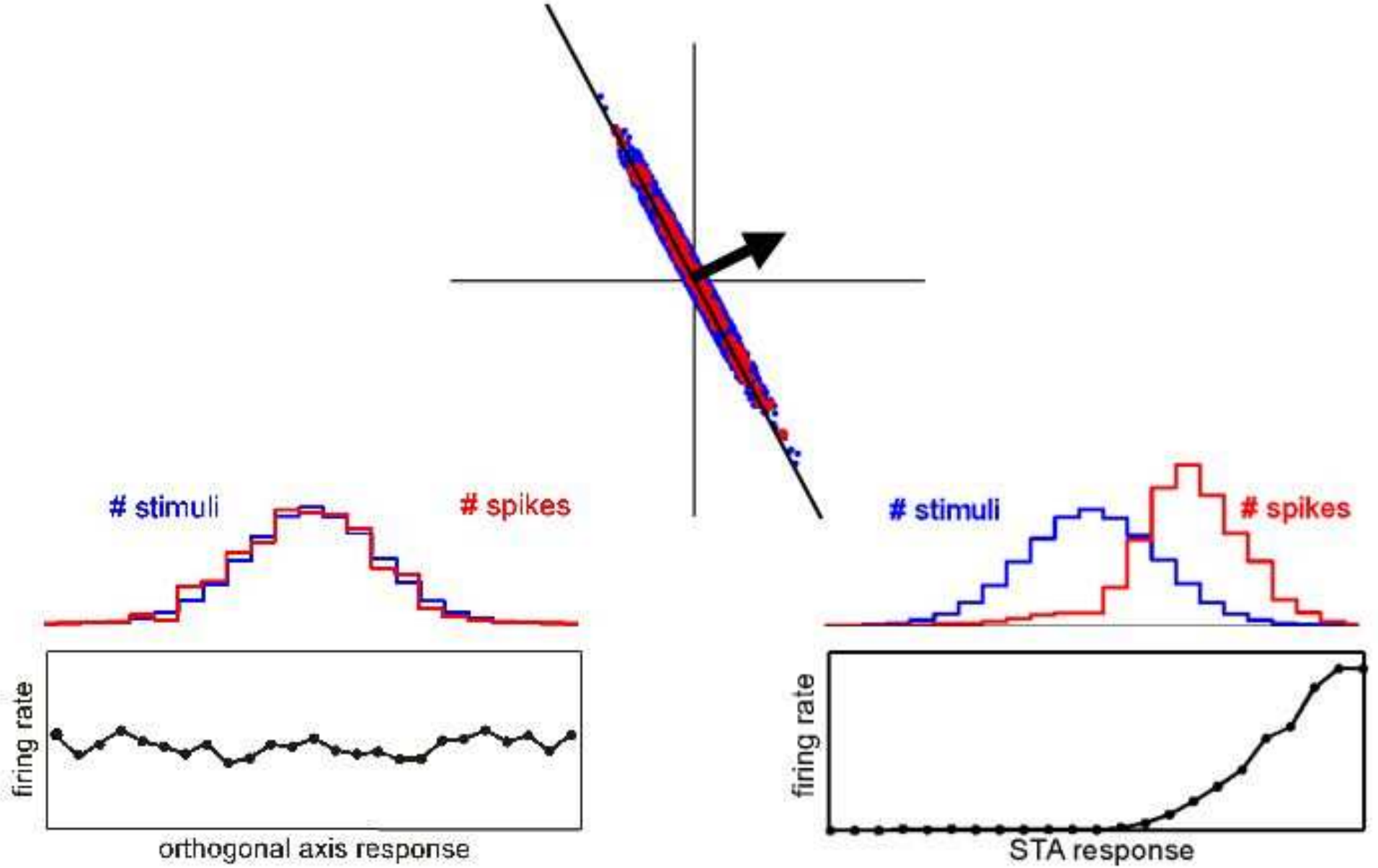
# Projecting onto the STA



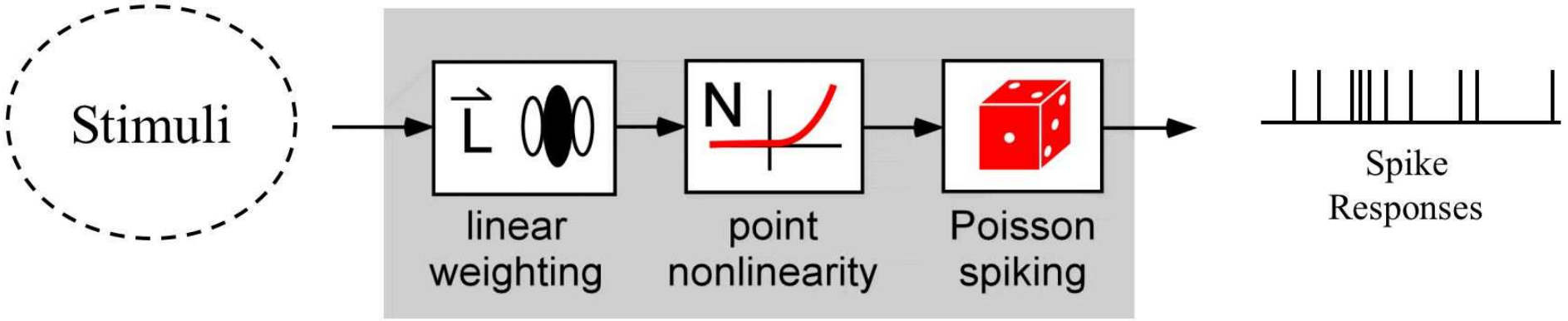
# Projecting onto an axis orthogonal to the STA

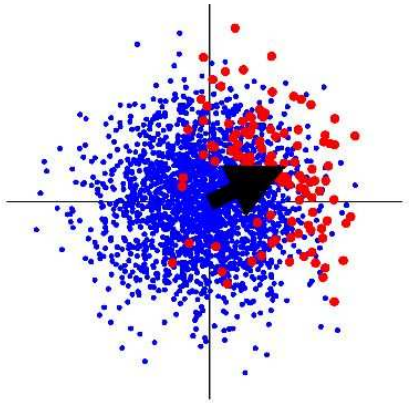


# Projecting onto an axis orthogonal to the STA

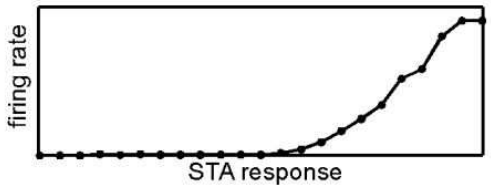


# LNP cascade model



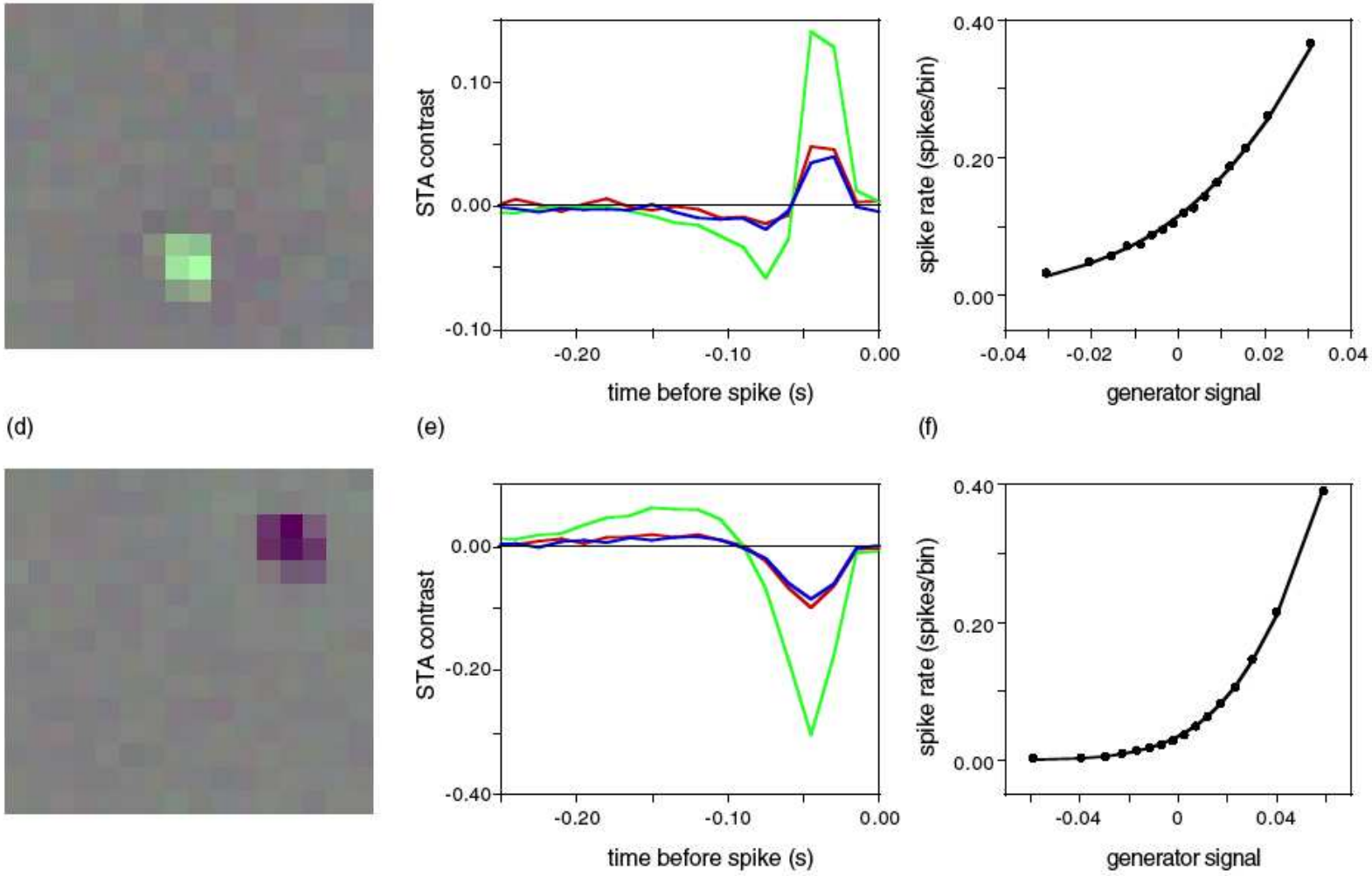



- STA provides unbiased estimate of $\vec{L}$ in radially symmetric case



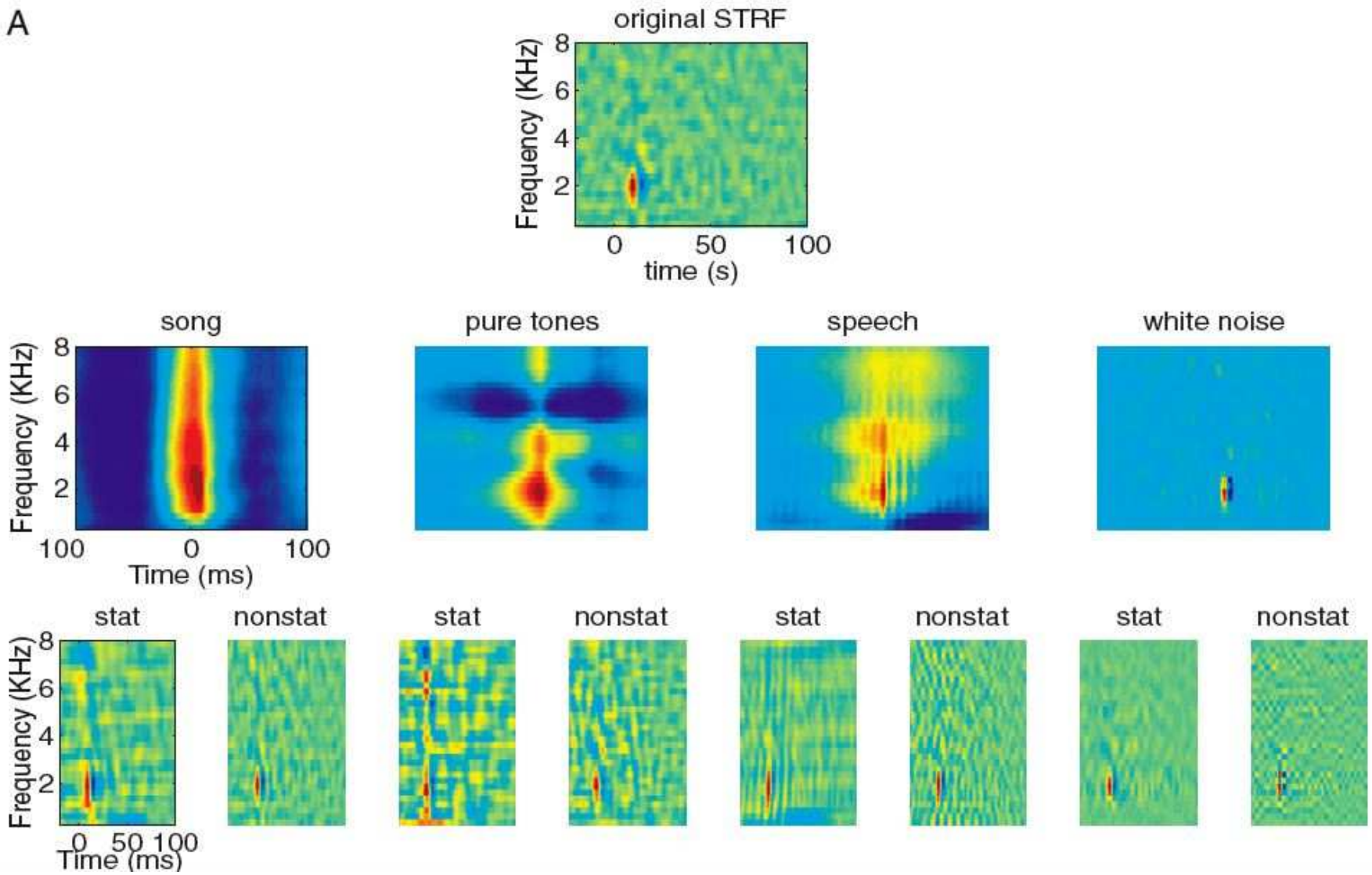

- projection onto STA provides estimate of N

# Retinal example




(Chichilnisky, 2001)

# Need to account for prior covariance!




(Theunissen et al., 2001)

# Accounting for prior covariance

Instead of taking raw STA, take

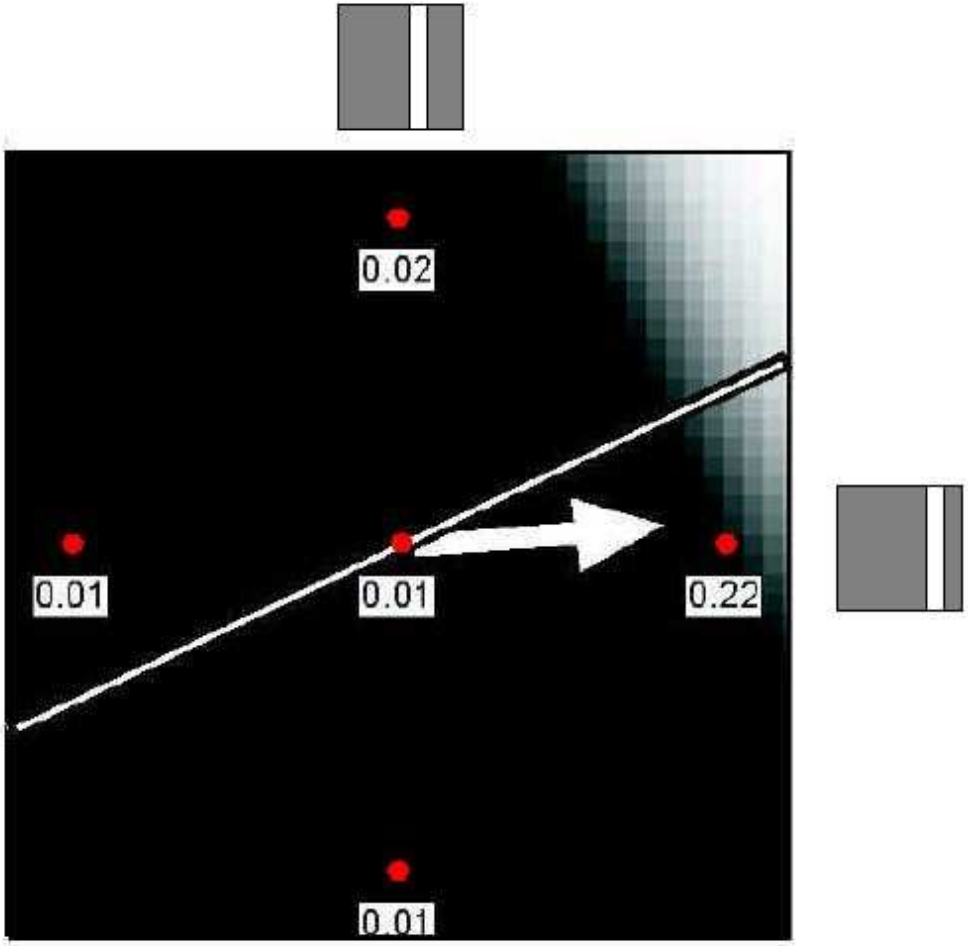
$$\vec{k} = C^{-1}\vec{k}_{STA}$$

$C$ = prior covariance matrix, $E(\vec{x}^t\vec{x})$.

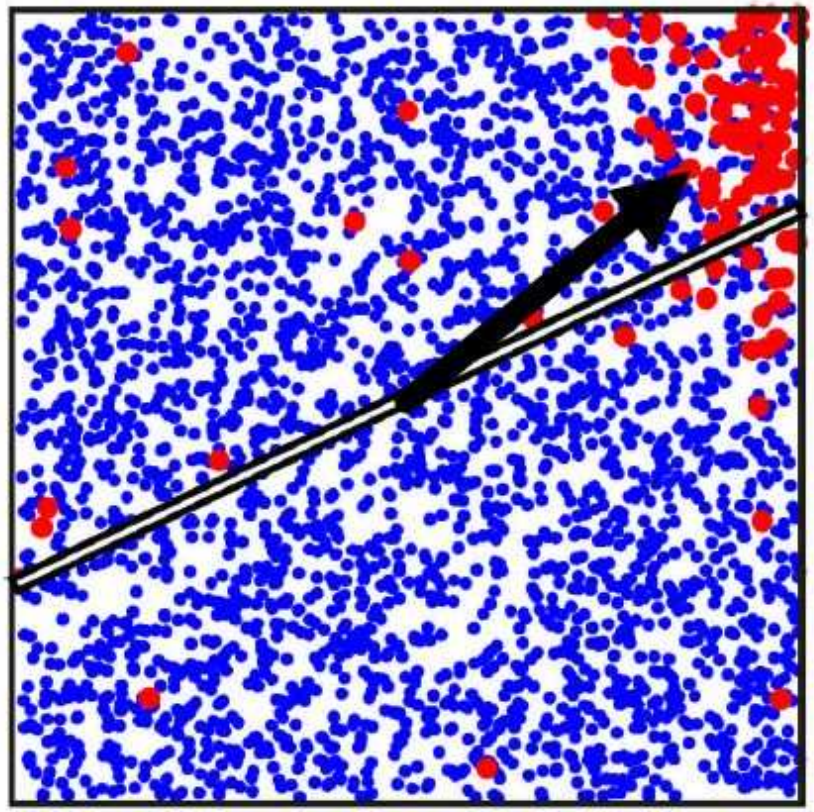
— Same as regression solution we used for additive model

— Unbiased for *elliptically* symmetric $p(\vec{x})$, not just radially symmetric (exercise: prove this.)

# Asymmetric examples: STA failures

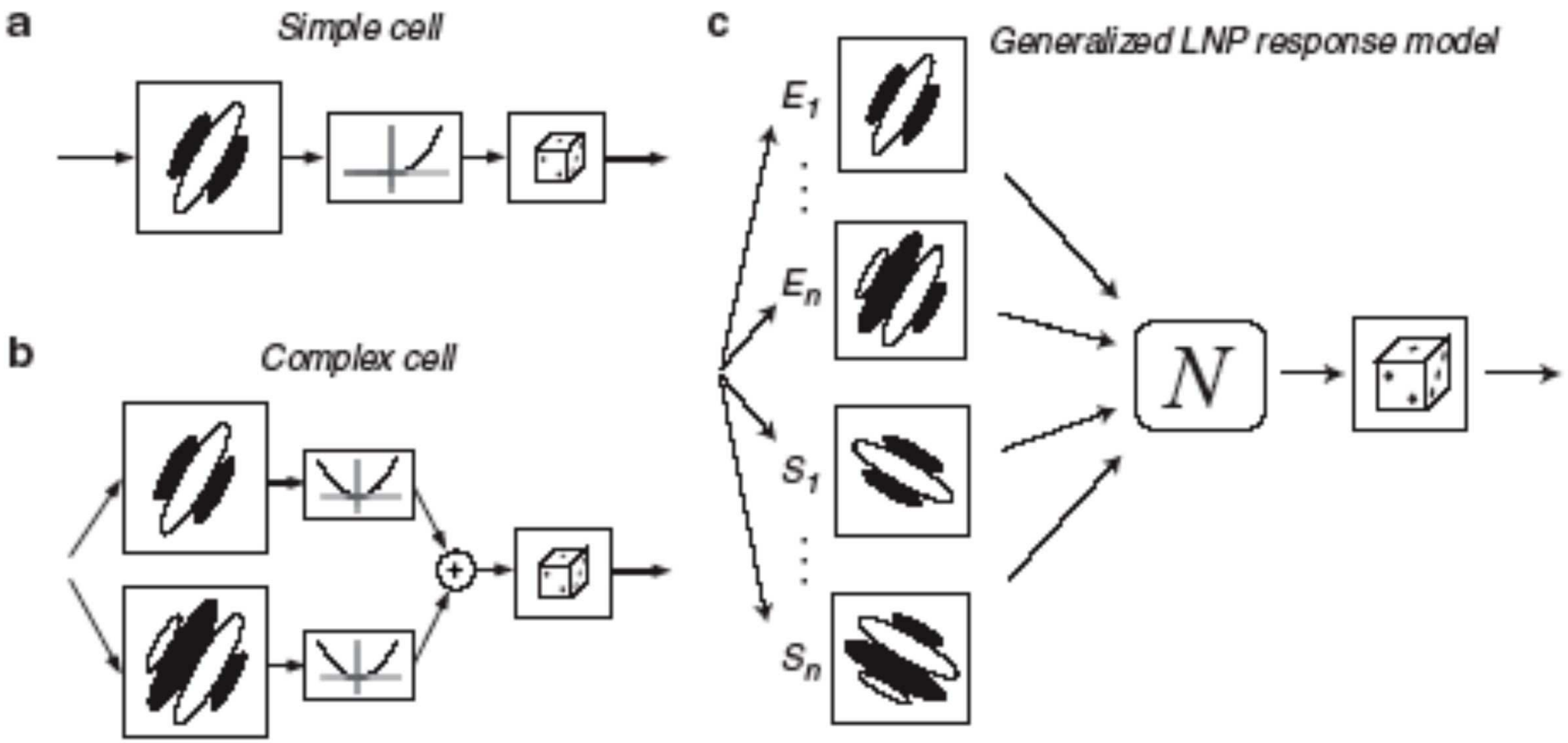



"sparse" noise

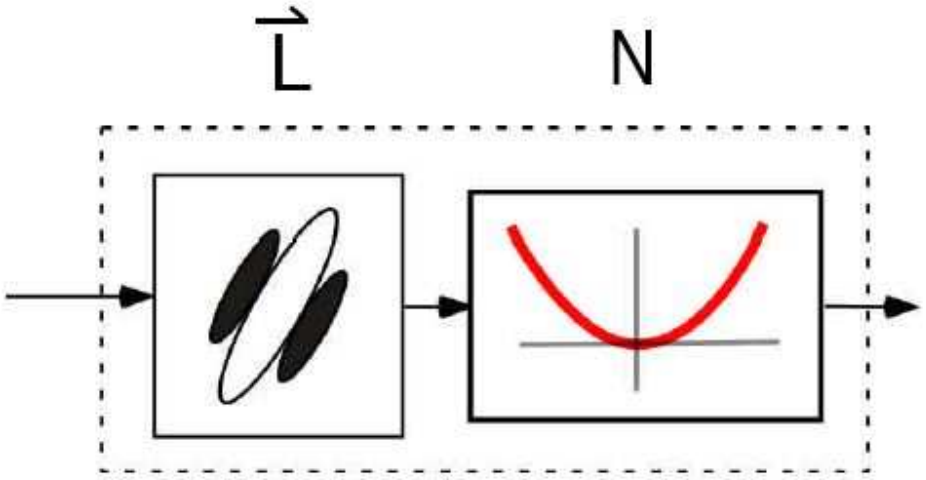


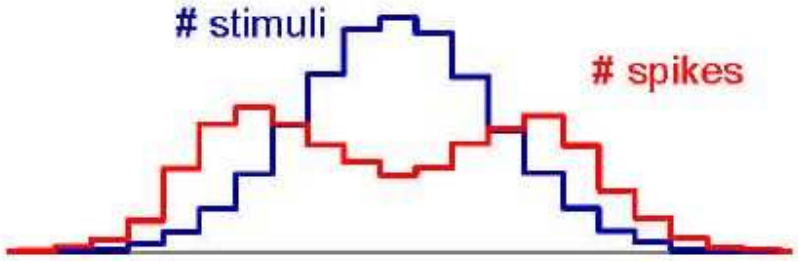
uniform noise

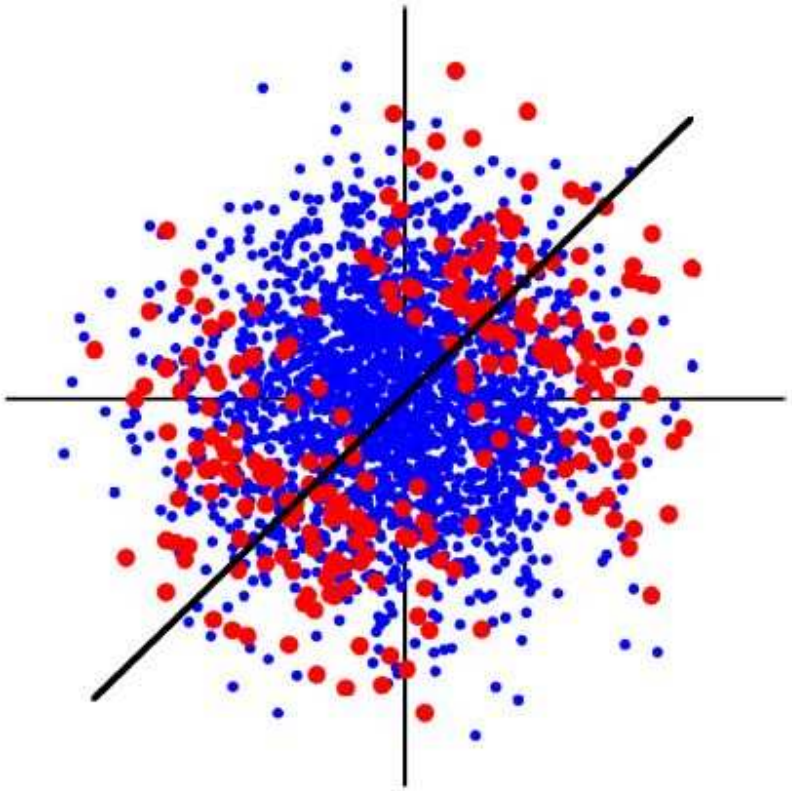
# What if symmetric nonlinearity? Or $> 1$ filter?



# Spike-triggered covariance



# Maximizing the change in variance

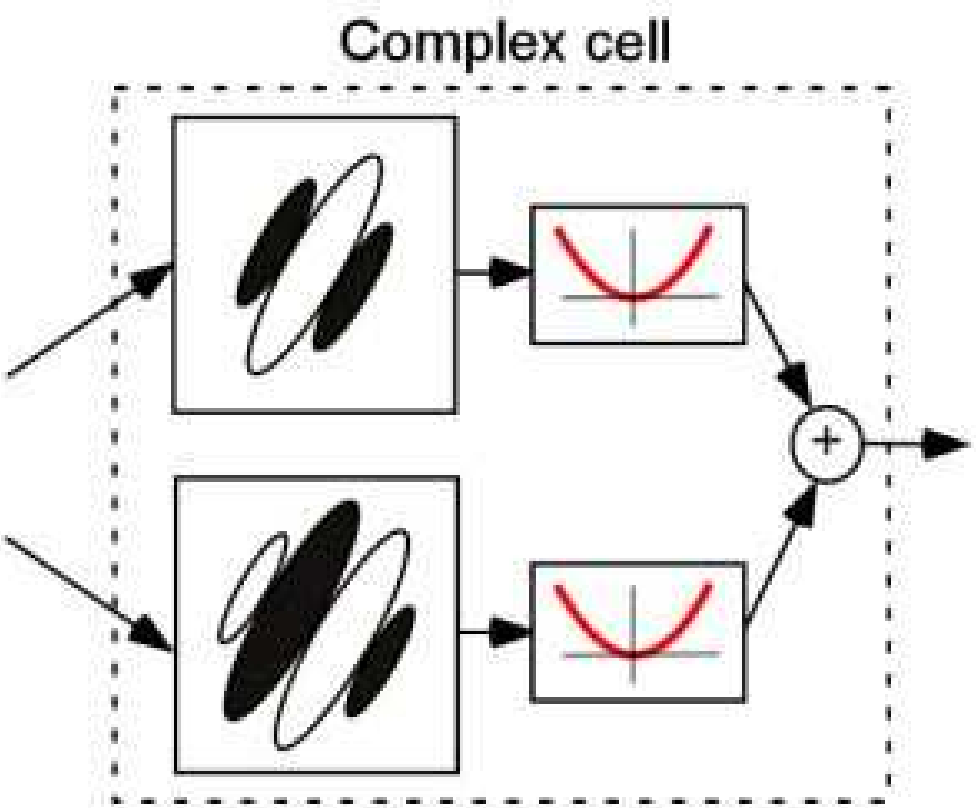
We want to pick the direction where change in variance is largest:

maximize $\vec{v}^t C_\Delta \vec{v}$ subject to $||\vec{v}||_2 = 1$;

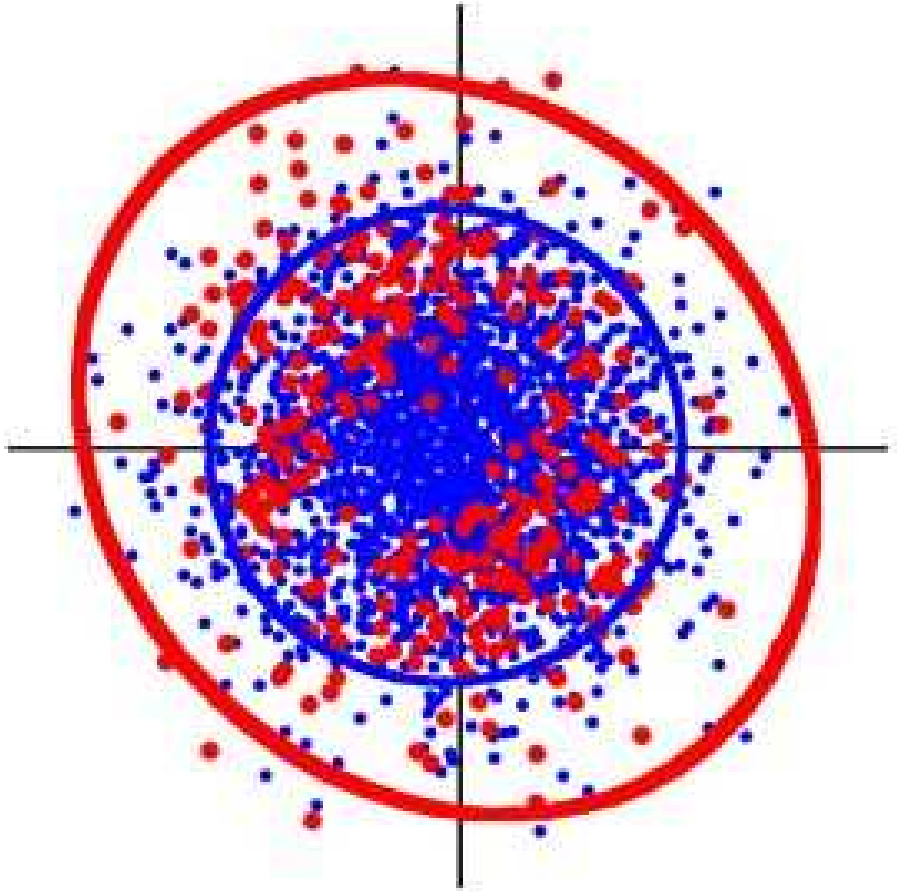
$C_\Delta = C_{prior} - C_{spike}$

$\Longrightarrow$ find eigenvectors of $C_\Delta$ with very large / small eigenvalues

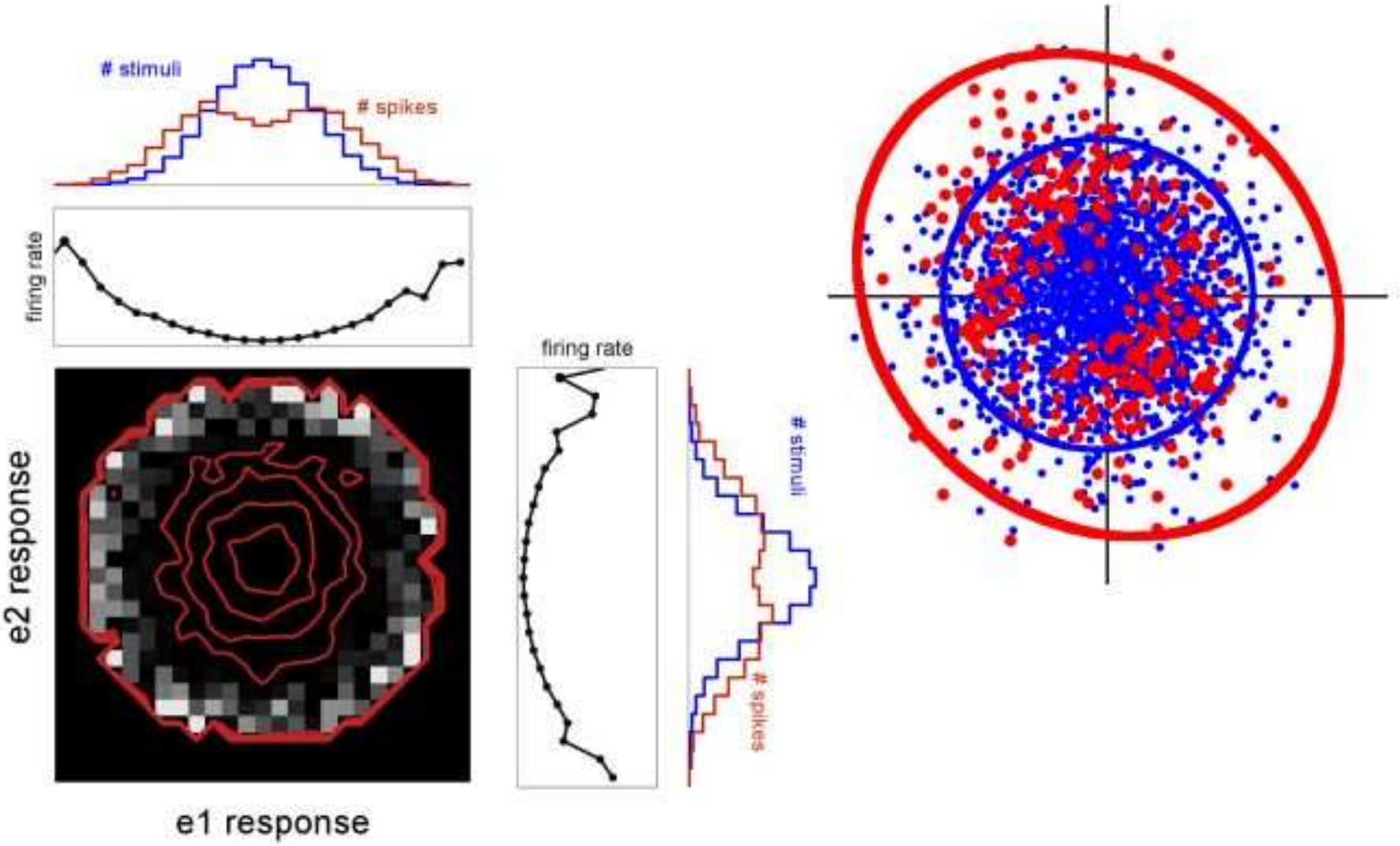
# Multiple linear filters



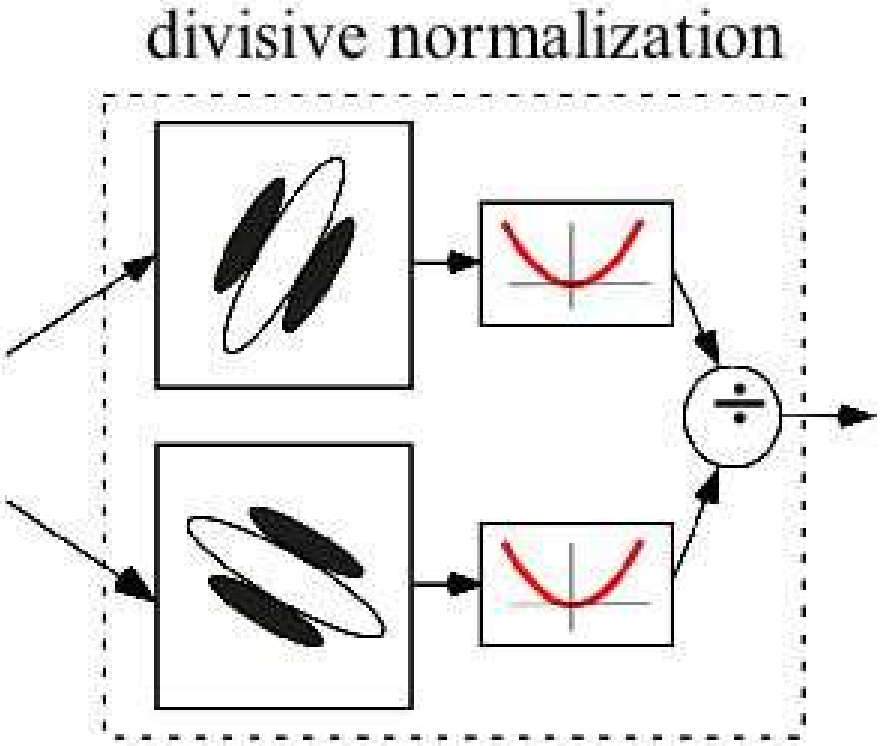

Adelson & Bergen 1985

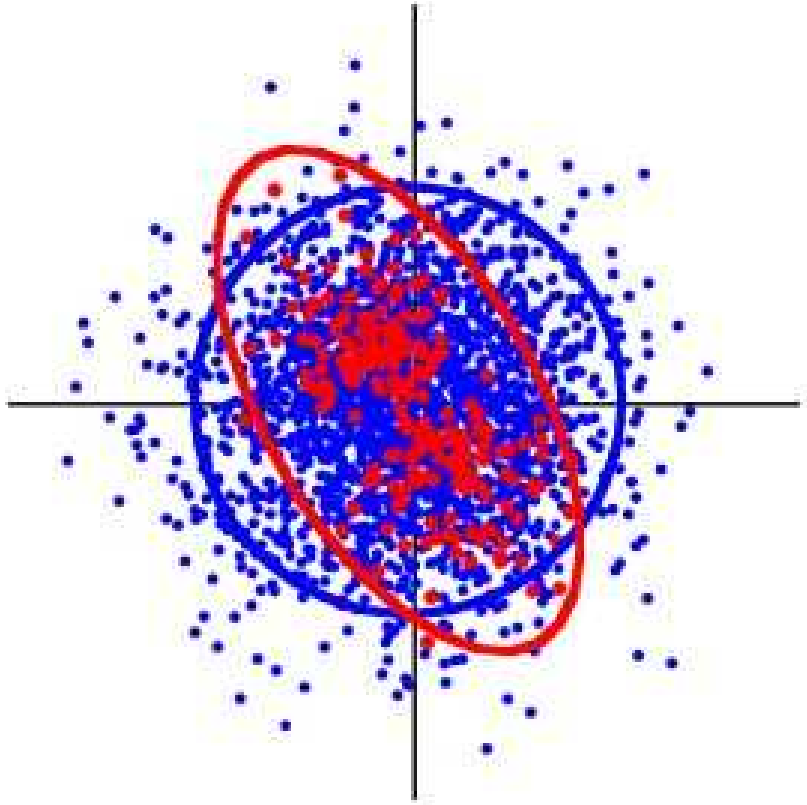


# Constructing the 2D nonlinearity

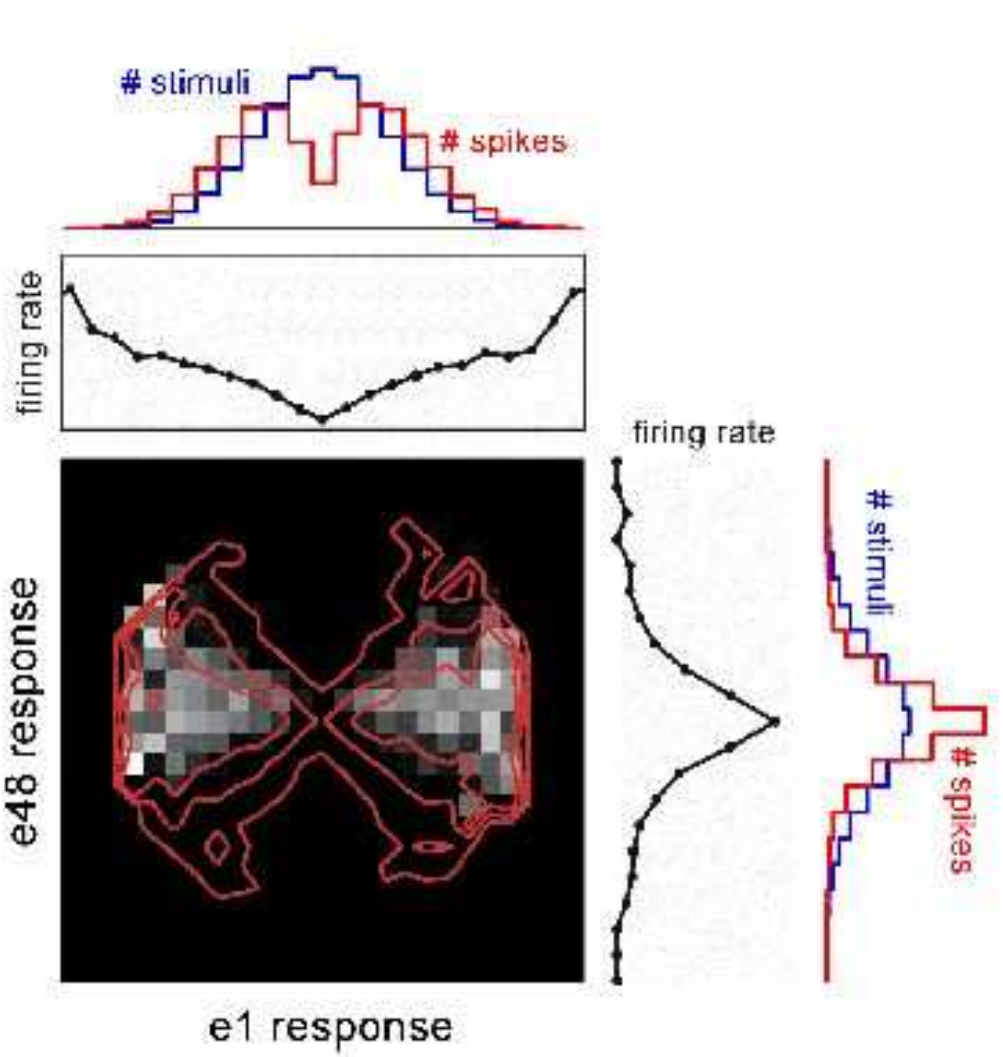


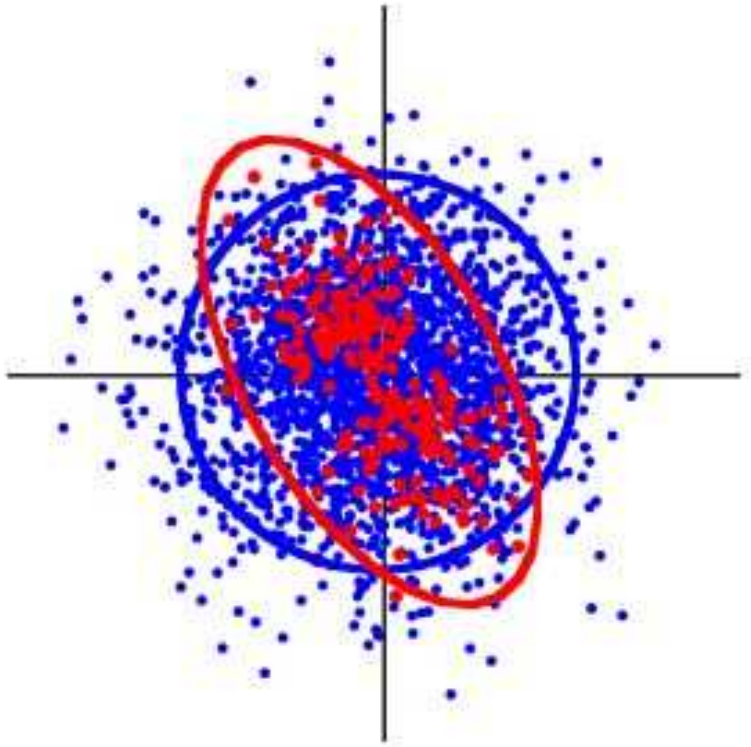
# Suppressive interactions



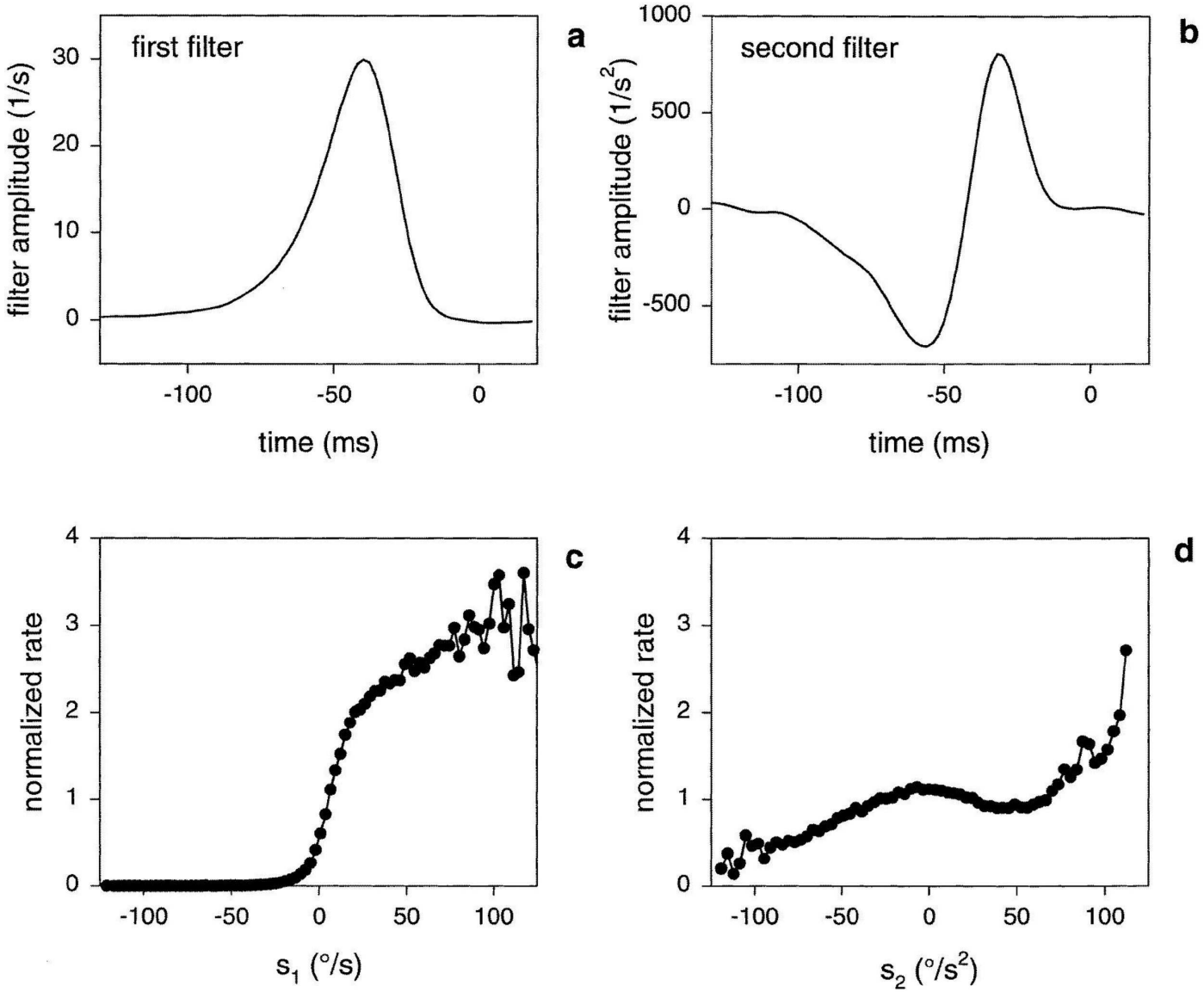



# STC: suppressive axes



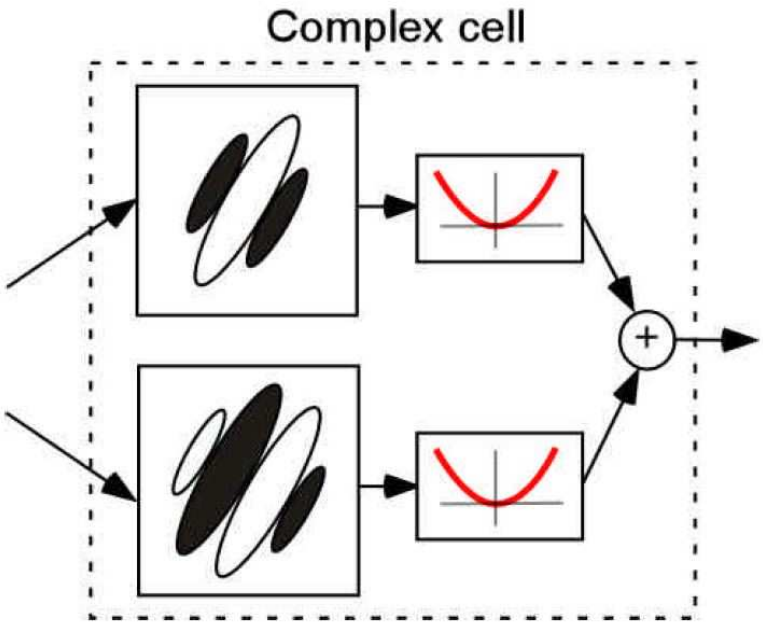



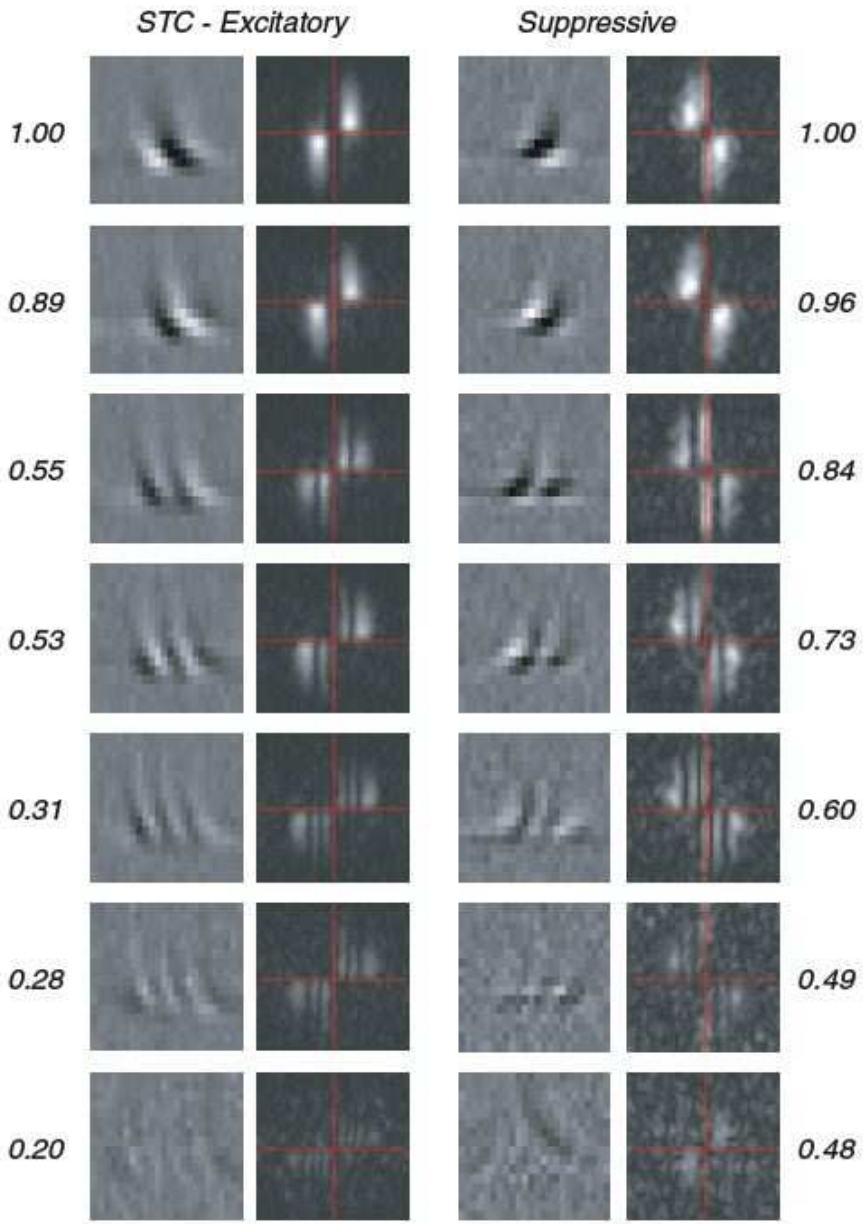
# Example from fly visual system




(Brenner et al., 2001)

# V1 cells aren't so simple




(Rust et al., 2003)

# What if neither mean nor covariance change?

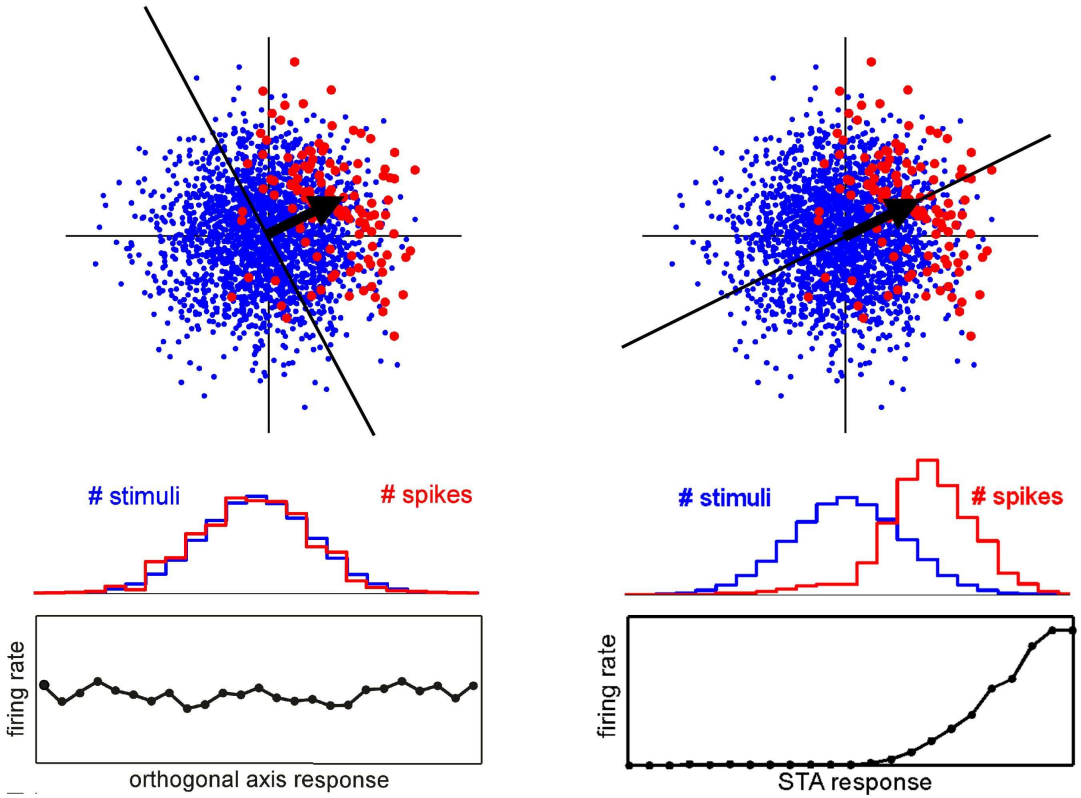
Fit L and N at same time.

Leads to unbiased estimators, no matter what $p(stim)$ is — symmetric priors unnecessary...

...but takes longer to compute.

(Weisberg and Welsh, 1994)

# Choose most modulatory direction




(Paninski, 2003a; Sharpee et al., 2004)

Proof that this gives correct $\vec{k}$ requires info theory - will come back to this in a couple lectures

# Other applications: response-triggered averaging

Psychophysics ("classification images")

fMRI

Optical imaging

...

# Interneuronal interactions

Above: firing rate given stimulus.

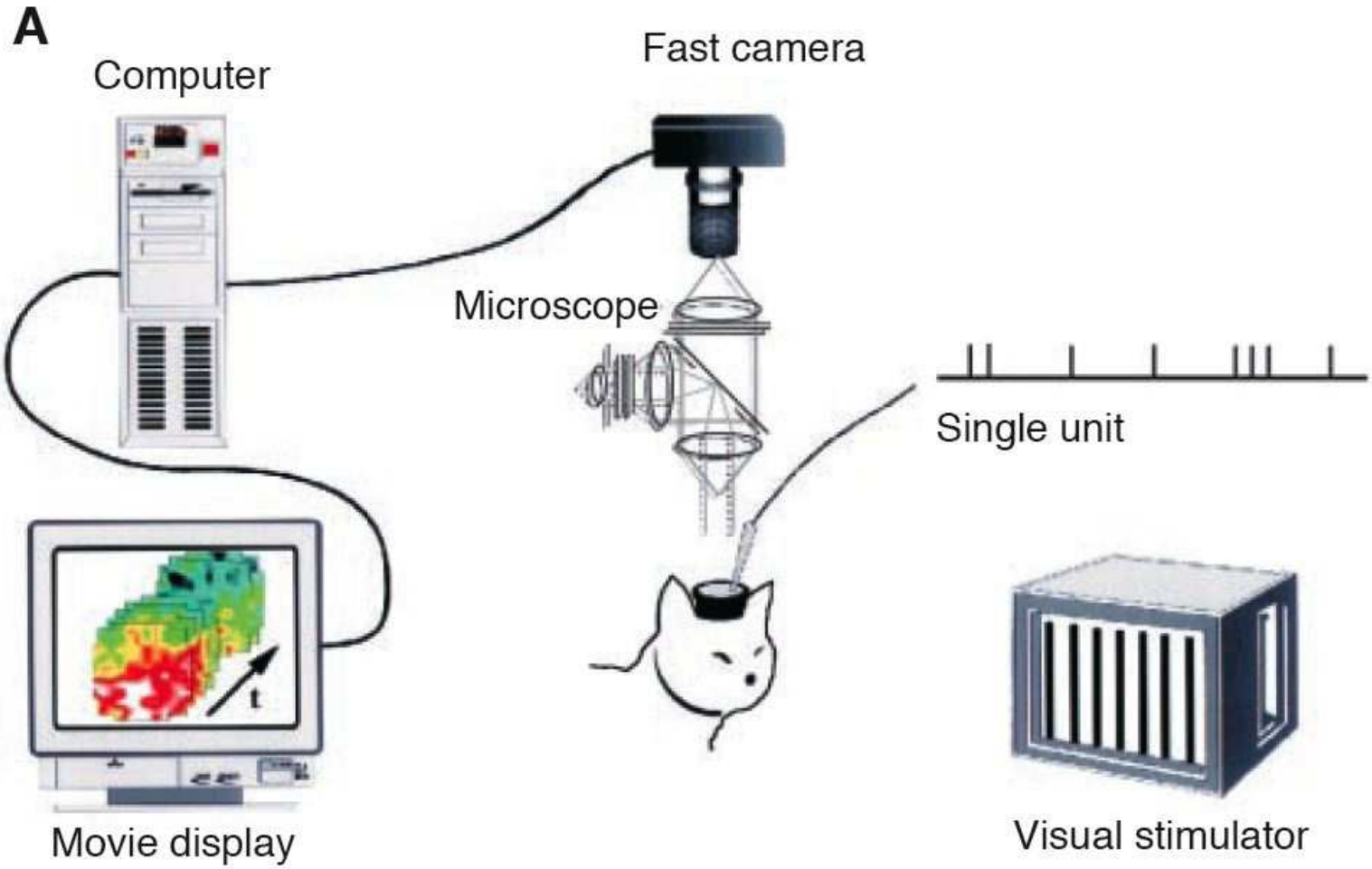
What if we want rate given activity of neighboring cells in network?

In a sense, just redefining what we mean by "stimulus"

$\Longrightarrow$ same mathematical techniques apply!

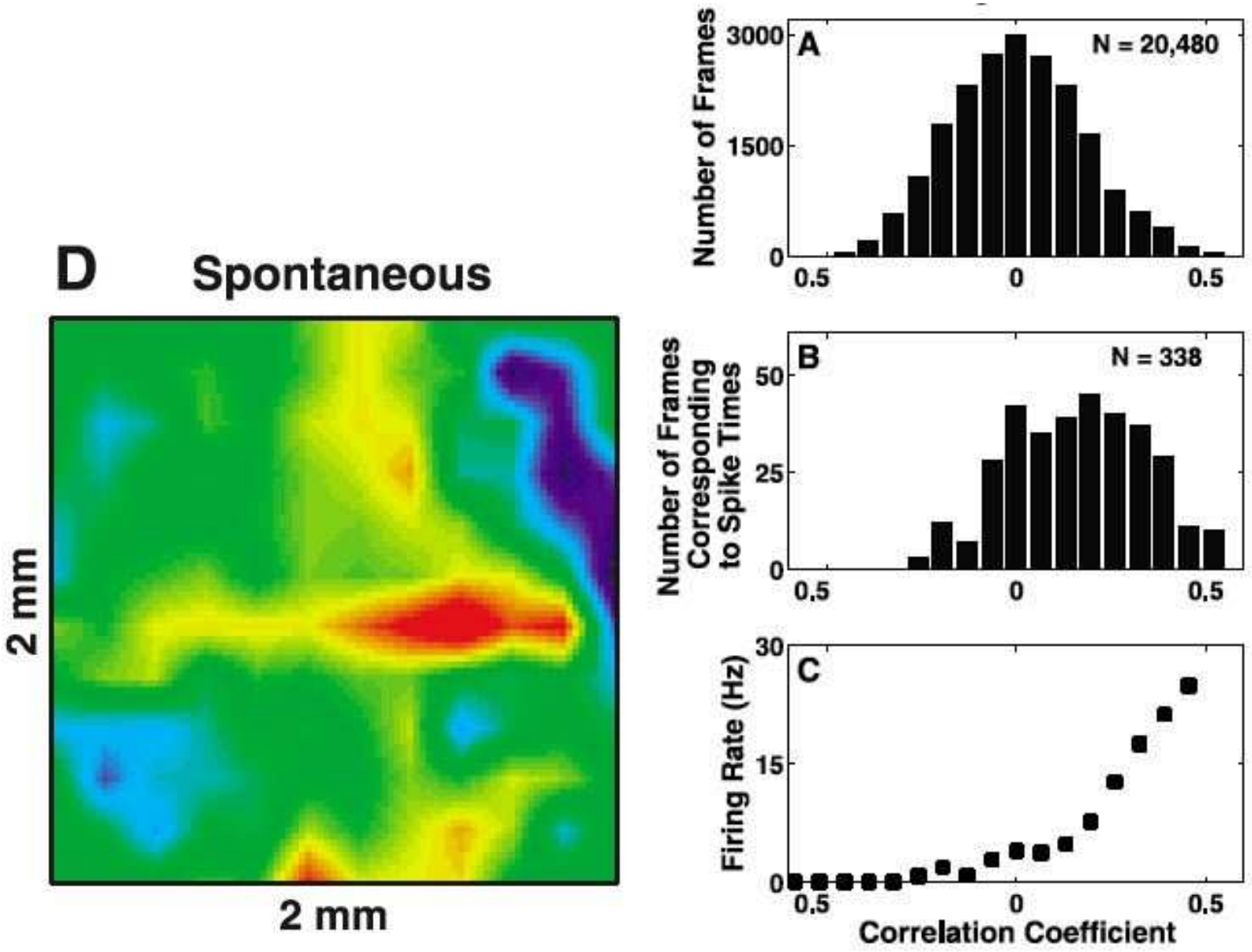
Can incorporate single- or multi-unit activity, or LFP, etc.

# Predictions from neighbor activity in V1



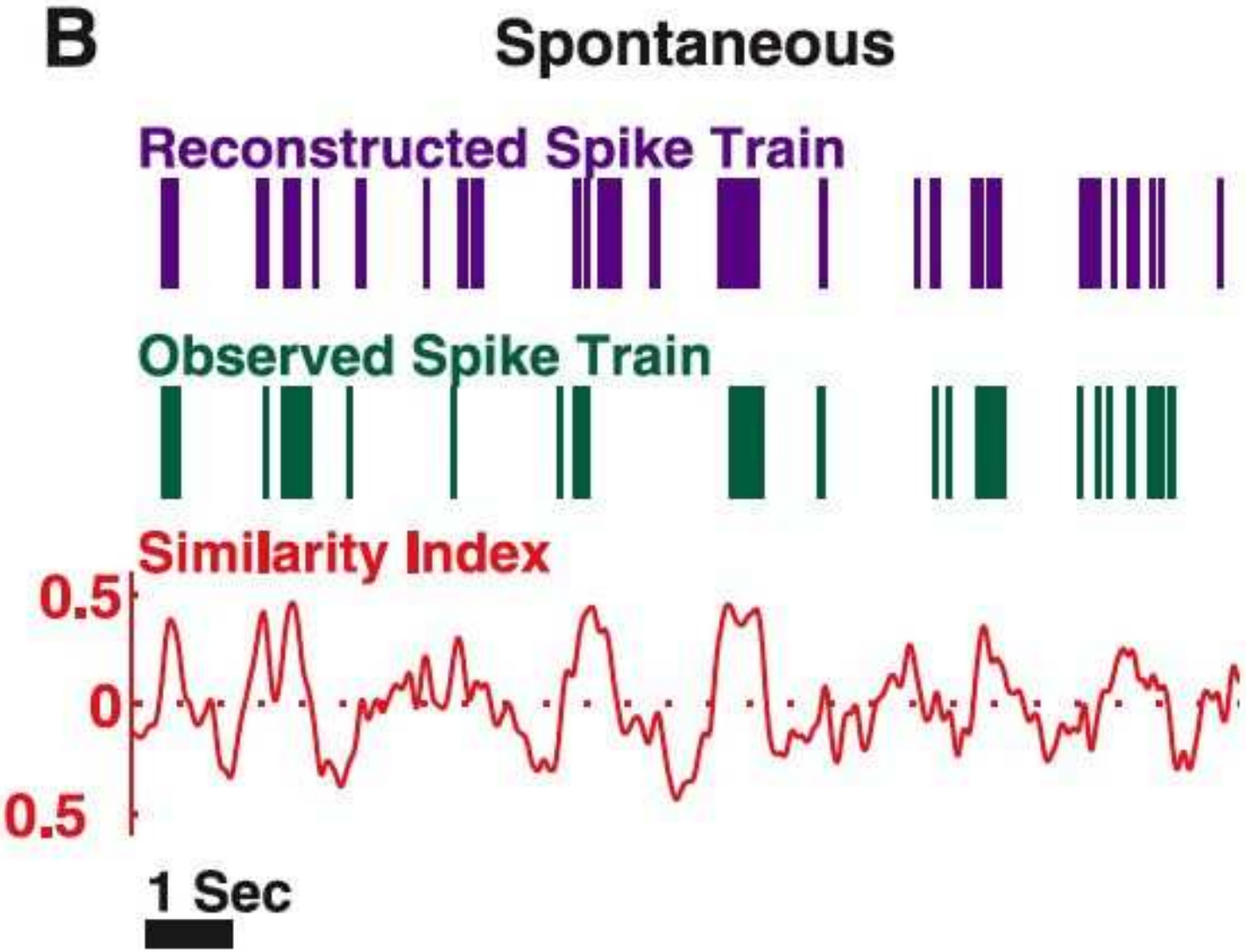

(Tsodyks et al., 1999)

# Predictions from neighbor activity in V1



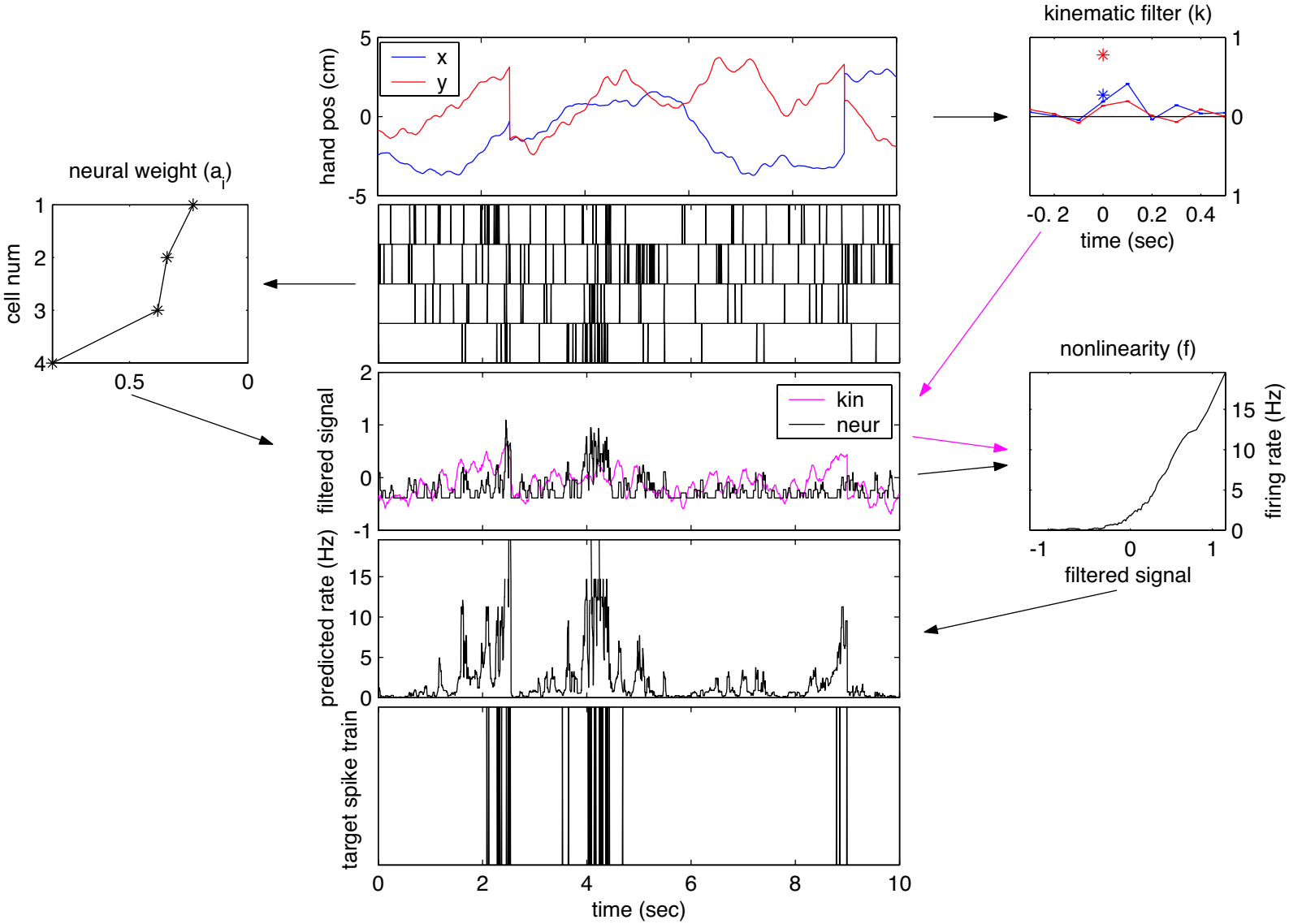

(Tsodyks et al., 1999)

# Predictions from neighbor activity in V1



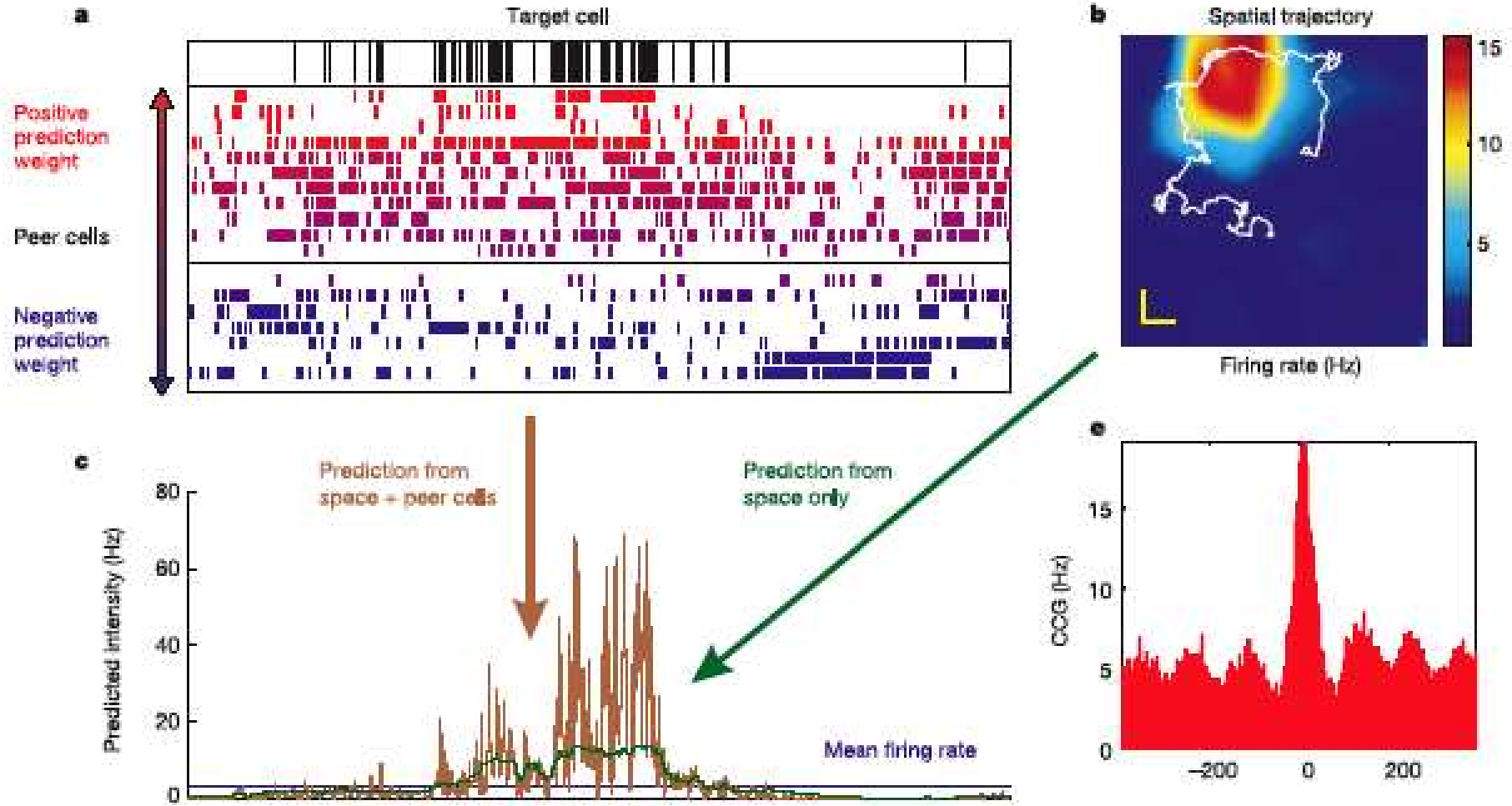

(Tsodyks et al., 1999)

# Combining kinematic, neighbor activity in MI




(Paninski et al., 2004)

# Combining place field and neighbor activity in hippocampus




(Harris et al., 2003)

# Regularization

Fitting stimulus, neighbor effects $\implies$ lots of parameters.

Lots of parameters + not enough data $\implies$ overfitting.

How to avoid this?

# Regularization, linear version

Instead of minimizing

$$Err = E[(\vec{k} \cdot \vec{x} - y)^2],$$

minimize

$$Err + \lambda_1 ||\vec{k}||_2 + \lambda_2 ||D\vec{k}||_2,$$

$||\vec{k}||_2 = \vec{k} \cdot \vec{k}$. Solution:

$$\vec{k}_{reg} = (C + \lambda_1 I + \lambda_2 \Delta)^{-1} \vec{k}_{STA}.$$

$\lambda_1, \lambda_2 \to 0 \implies$ no regularization

$\lambda_1 \to \infty \implies \vec{k}_{reg} \to \vec{0}$: shrinking

$\lambda_2 \to \infty \implies \vec{k}_{reg} \to \vec{1}$: smoothing

# Theoretical justification

Idea: introduce some bias to reduce variance

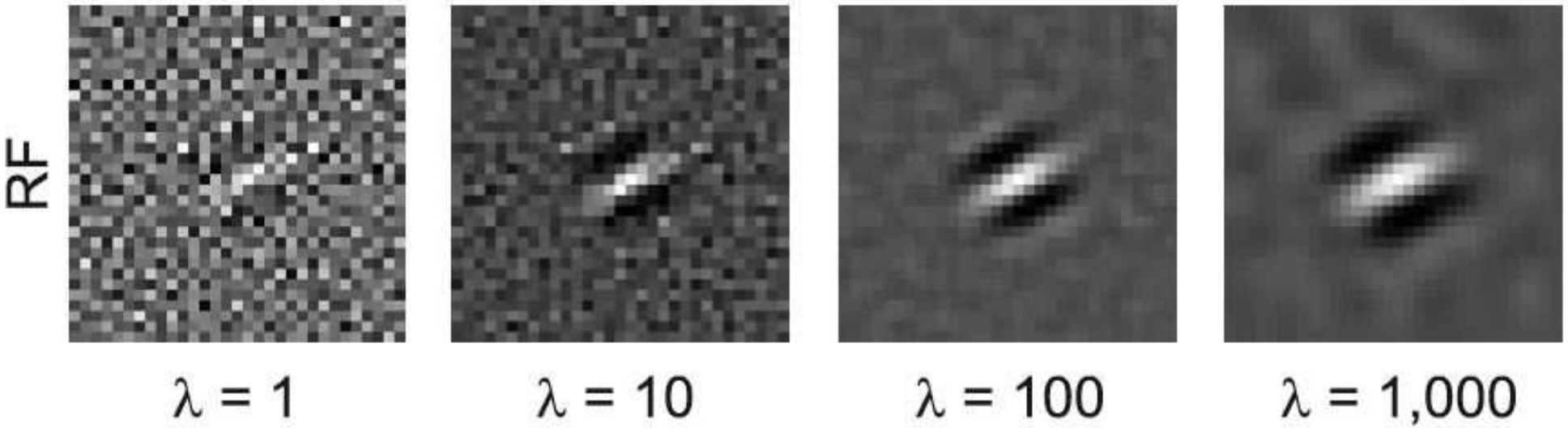
If $\lambda_1, \lambda_2$ chosen correctly, $\vec{k}_{reg}$ is a better estimate for $\vec{k}$ than $\vec{k}_{STA}$, for all underlying $\vec{k}$ (James and Stein, 1960)

# Bayesian interpretation

$\vec{k}_{reg}$ = posterior mean given data $\{\vec{x}, y\}$, under Gaussian errors, Gaussian prior on $\vec{k}$ (exercise)

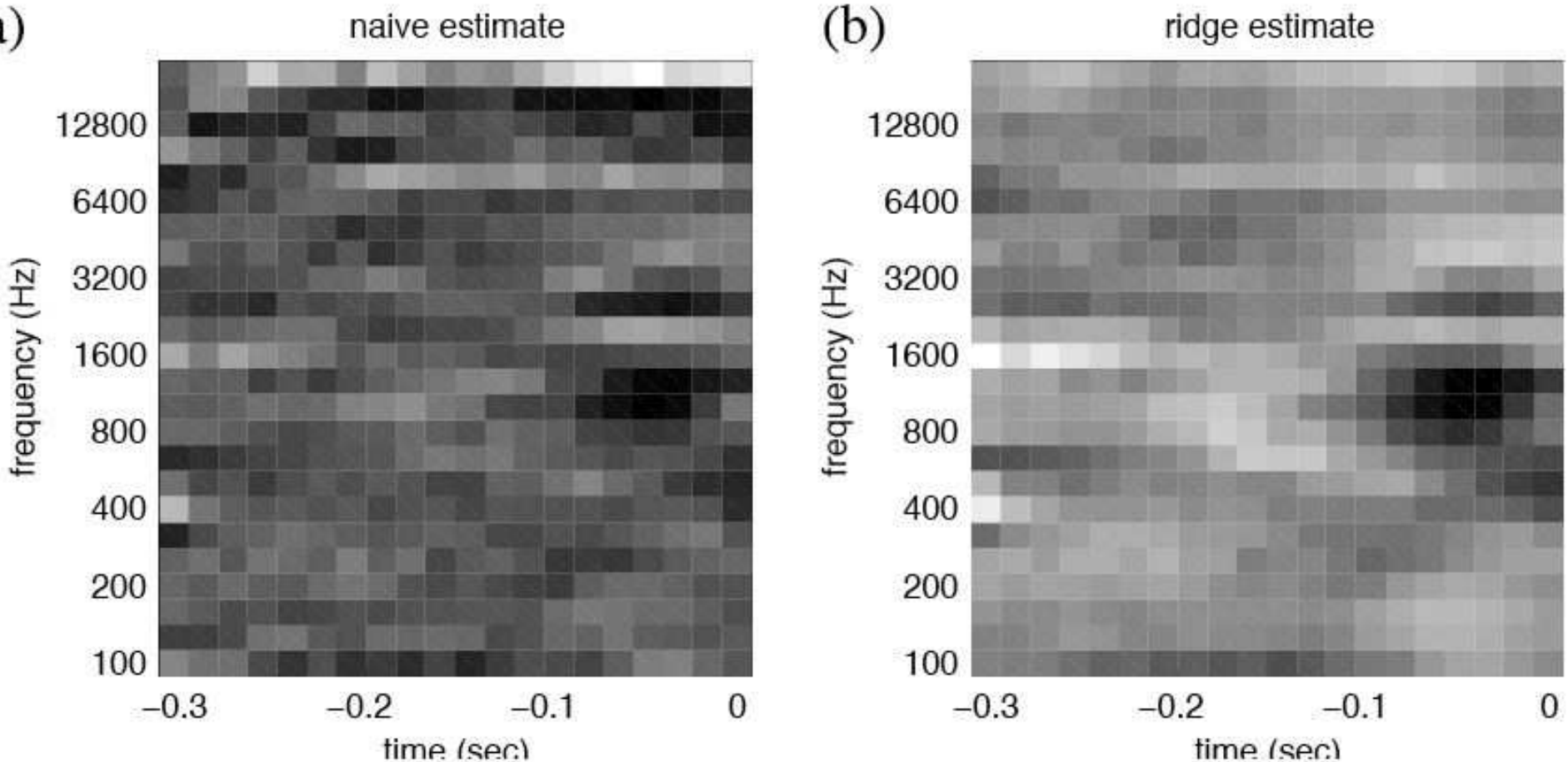
$\lambda_1, \lambda_2$ large $\implies$ we expect $\vec{k}$ to be smooth and small

# Regularization can improve estimates of V1 receptive fields



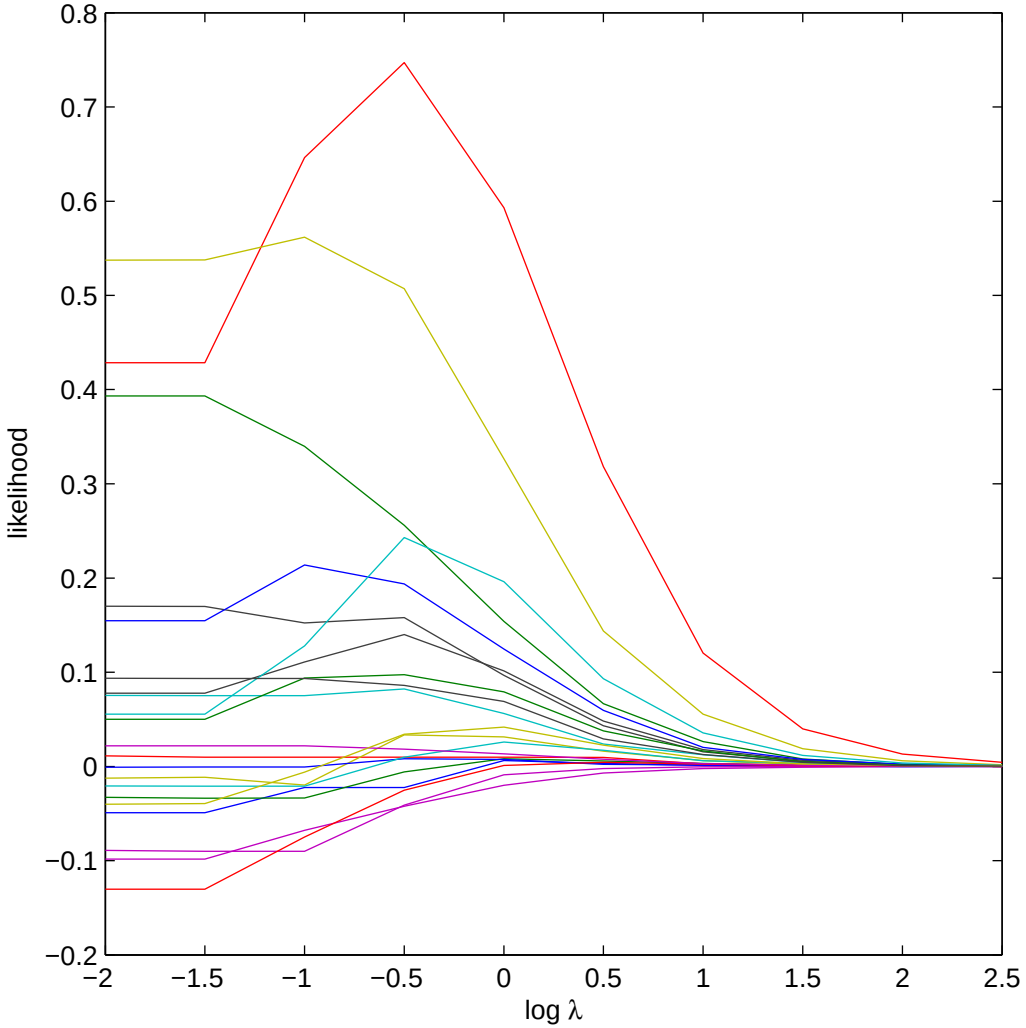

(Smyth et al., 2003)

# Regularization can improve estimates of A1 receptive fields




(Machens et al., 2003)

# Regularization can improve predictions given MI neighbor activity




(Paninski et al., 2003; Paninski, 2004)

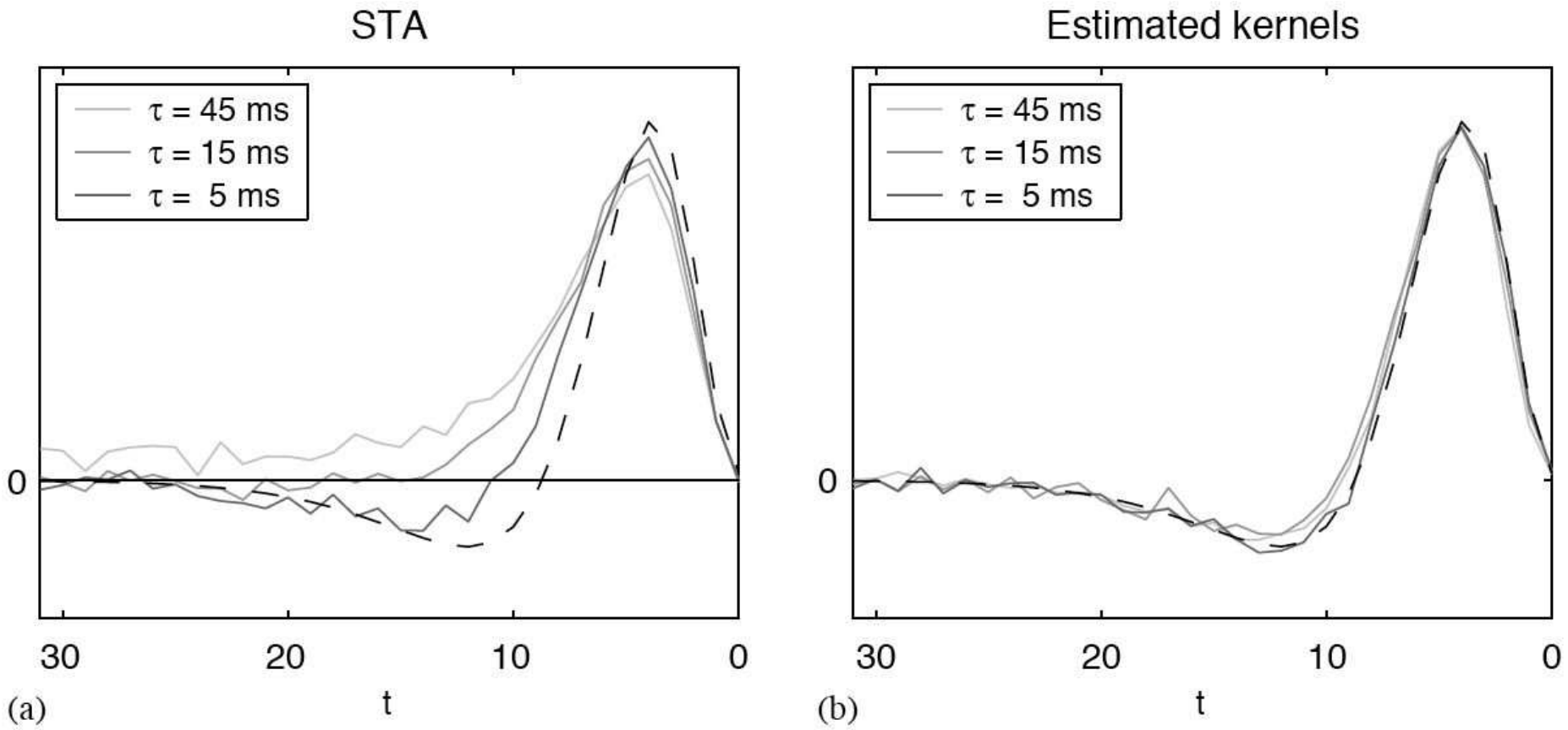
# One last problem...

We assumed above that spikes are *conditionally independent* given $\vec{x}$.

What if spikes are history dependent?

(they are... refractory period, adaptation, burstiness, etc.)

# STA is biased if spikes history-dependent




(Pillow and Simoncelli, 2003)

# Summary so far...

Introduced cascade idea, geometric intuition

Discussed STA, STC fitting procedures

Extended cascade idea to include interneuronal effects

Illustrated importance of "complexity control," reducing overfitting

Next: incorporating history dependence; continuous-time models

# References


Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2001). Adaptive rescaling optimizes information transmission. *Neuron*, 26:695–702.

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12:199–213.

Foldiak, P. (2001). Stimulus optimisation in primary visual cortex. *Neurocomputing*, 38–40:1217–1222.

Harris, K., Csicsvari, J., Hirase, H., Dragoi, G., and Buzsaki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature*, 424:552–556.

James, W. and Stein, C. (1960). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379.

Machens, C. (2002). Adaptive sampling by information maximization. *Physical Review Letters*, 88:228104–228107.

Machens, C., Wehr, M., and Zador, A. (2003). Spectro-temporal receptive fields of subthreshold responses in auditory cortex. *NIPS*.

Paninski, L. (2003a). Convergence properties of some spike-triggered analysis techniques. *Network: Computation in Neural Systems*, 14:437–464.

Paninski, L. (2003b). Design of experiments via information theory. *Advances in Neural Information Processing Systems*, 16.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15:243–262.

Paninski, L., Fellows, M., Shoham, S., Hatsopoulos, N., and Donoghue, J. (2004). Superlinear population encoding of dynamic hand trajectory in primary motor cortex. *Journal of Neuroscience*, 24:8551–8561.

Paninski, L., Lau, B., and Reyes, A. (2003). Noise-driven adaptation: in vitro and mathematical analysis. *Neurocomputing*, 52:877–883.

Pillow, J. and Simoncelli, E. (2003). Biases in white noise analysis due to non-Poisson spike generation. *Neurocomputing*, 52:109–115.

Rust, N., Schwartz, O., Movshon, A., and Simoncelli, E. (2003). Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey v1 directionally selective neurons. *Presented at the Annual Computational Neuroscience meeting, Alicante, Spain.*

Sharpee, T., Rust, N., and Bialek, W. (2004). Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Computation*, 16:223–250.

Smyth, D., Willmore, B., Baker, G., Thompson, I., and Tolhurst, D. (2003). The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *Journal of Neuroscience*, 23:4746–4759.

Theunissen, F., David, S., Singh, N., Hsu, A., Vinje, W., and Gallant, J. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, 12:289–316.

Tsodyks, M., Kenet, T., Grinvald, A., and Arieli, A. (1999). Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286:1943–1946.

Weisberg, S. and Welsh, A. (1994). Adapting for the missing link. *Annals of Statistics*, 22:1674–1700.