Admin stuff

- **Time:** 5 weeks for Segment 1, no classes March 15 (!)
- **TAs:** Reka Borbely + Katya Maksimova
- **Recitations:** usual weekly slot to discuss HW, slots for additional material, no slot this week
- **Lecture notes:** PDF to appear online on Moodle
- **Style:** Mostly slides
  (i) worked out examples on a single dataset
  (ii) "excursions" for selected papers (**BLUE BOXES** in the notes)
- **Homeworks:** small assignments + mini-project (**GRAY BOXES** in the notes), TAs will provide more instruction on the deadlines / how to upload the assignments

Topic of Segment 1:

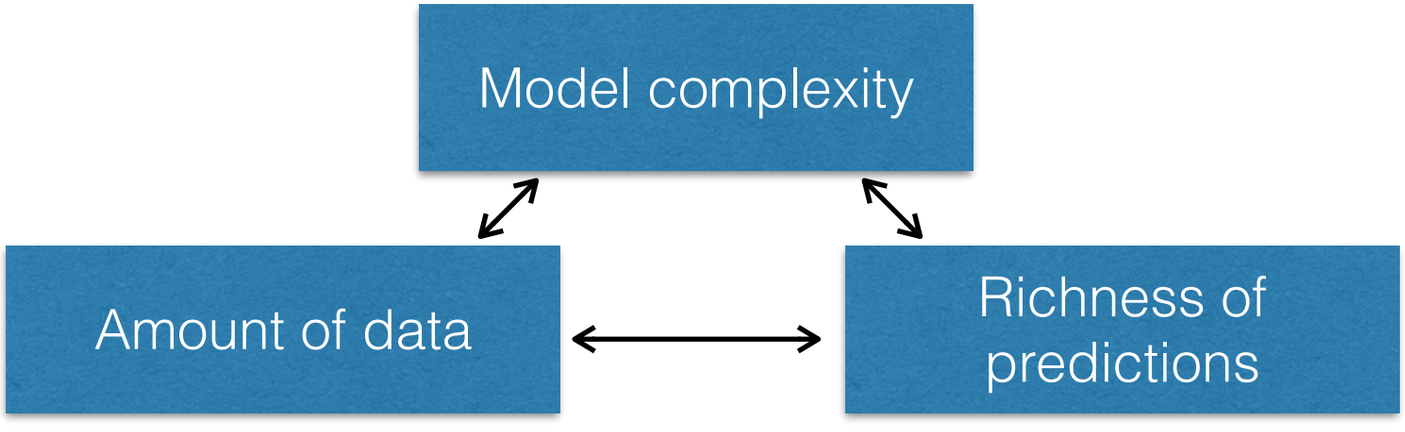
Understanding and visualizing data

Goal: <u>Hands-on / applied</u> understanding of the basic tools used to characterize "a new dataset".
Usefulness of understanding "low order statistical structure" in the data.
This is **neither** a rigorous statistics / math course **nor** "intro to modeling" for life scientists.

- One-point statistics (histograms, PDFs, moments, …)
- Statistical dependencies, correlations, PCA
- Bootstrap / subsampling error bar estimation
- Stationarity, temporal correlations, basic temporal filtering
- Discovering higher-order statistics in the data: *K-means, t-SNE, MDS, ICA*

Data analysis / modeling

- **Predictive power**
  Model complexity depends on the data, interpretability not essential; performance on withheld / future data generated by the same process
- **Hypothesis testing**
  Hypothesis-driven research is only one possible route towards scientific understanding; excellent understanding of statistical assumptions behind testing
- **Principled understanding**
  There exist “true” models describing measurement-device independent quantities; parts of the model (e.g., parameters) should be interpretable even at the cost of (somewhat) worse in-sample generalization; generalization of model predictions to out-of-sample data (intervention / perturbation, counterfactual etc predictions)



Why are you building a model at all?

This is not obvious and depends on the assumptions of different fields.

The recommended approach and its tradeoffs crucially depend on the answer.

What is considered a “model”?

Statistics / ML

Null models / hyp. tests

Linear models $y_i = \mathbf{A}\mathbf{x}_i$

Universal approx. $y_i = f(\mathbf{x}_i; \theta)$

hypothesis explicit, data not limiting
data “is given”, less discussion about
whether in “relevant” form

hard-to-interpret, large # params,
but powerful

Biology

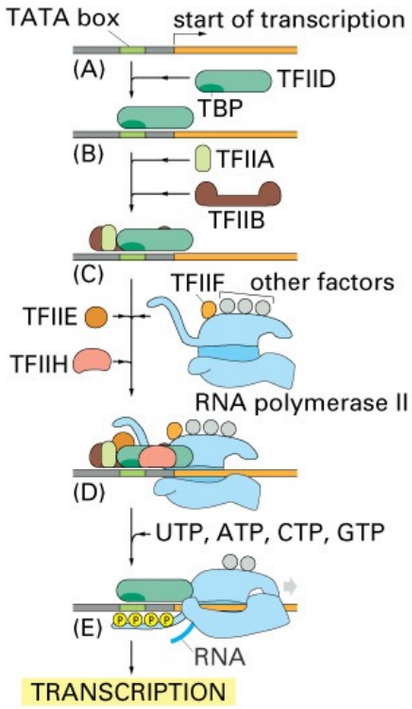


Figure 8-10 Essential Cell Biology, 2/e. (© 2004 Garland Science)

c.f. engineering block diagrams
typically no quantitative predictions,
but can make strong qualitative ones
(= for perturbation experiments)
some models can be turned into
e.g., graphical or ODE-type models
(but insufficient data to fit those)
exploring the vast space of dependencies

Physics

$$F_g = G\frac{m_1 m_2}{r^2}$$

simple equation,
but identified relevant quantities essential
internal consistency, nearly perfect
generalization performance (“law of nature”)

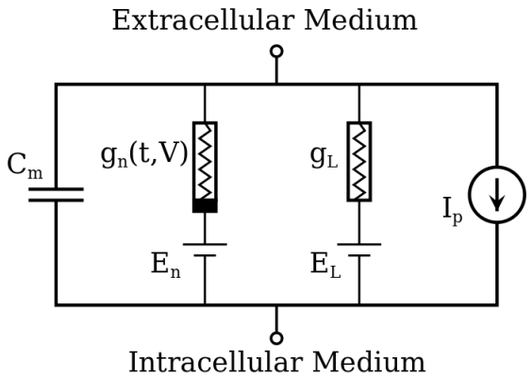
Appropriate “scale” for the model

- **Mechanistically detailed models**
  The observation results are explained in terms of “elementary processes” given by some underlying theory or knowledge; usually “lots” of parameters.

- **Phenomenological models**
  “Elementary processes” are coarse-grained away into an effective model with low number of parameters

  - colliding hard spheres in a box
  - metal atoms in a crystal lattice
  - all-atom protein dynamics
  - Hodgkin-Huxley model

  - ideal gas
  - Ohm’s law, ising magnet, …
  - stochastic conformational transitions
  - leaky-integrate-and-fire model



$$I = C_m \frac{\mathrm{d}V_m}{\mathrm{d}t} + \bar{g}_{\mathrm{K}} n^4 (V_m - V_K) + \bar{g}_{\mathrm{Na}} m^3 h (V_m - V_{Na}) + \bar{g}_l (V_m - V_l),$$

$$\frac{dn}{dt} = \alpha_n(V_m)(1-n) - \beta_n(V_m) n$$

$$\frac{dm}{dt} = \alpha_m(V_m)(1-m) - \beta_m(V_m) m$$

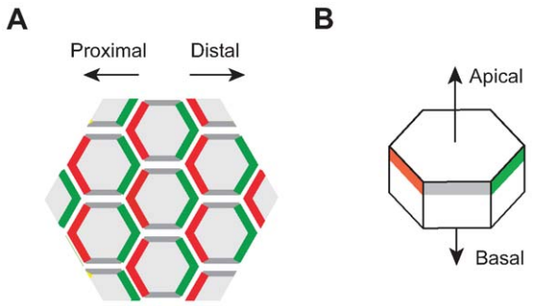
$$\frac{dh}{dt} = \alpha_h(V_m)(1-h) - \beta_h(V_m) h$$

$$I(t) - \frac{V_{\mathrm{m}}(t)}{R_{\mathrm{m}}} = C_{\mathrm{m}} \frac{dV_{\mathrm{m}}(t)}{dt}$$

+spike-and-reset,
whenever Vm > Threshold

Often, phenomenological models can be more precise / predictive for a **selected question.** Sometimes, their structure can be strongly constrained by microscopic knowledge.

Appropriate “scale” for the model



Shraiman & Burak *PLOS Comp Biol*: 2 equations, ~5 parameters

Axelrod, Tomlin et al, *Science*: ~40 parameters

$$\mathrm{Dsh} + \mathrm{Fz} \underset{A_1 B \lambda_1}{\overset{R_1}{\rightleftarrows}} \mathrm{DshFz}$$

$$\mathrm{Fz}^\dagger + \mathrm{Vang} \underset{\lambda_2}{\overset{R_2}{\rightleftarrows}} \mathrm{FzVang}$$

$$\mathrm{Vang} + \mathrm{Pk} \underset{\lambda_3}{\overset{R_3}{\rightleftarrows}} \mathrm{VangPk}$$

$$\mathrm{DshFz}^\dagger + \mathrm{Vang} \underset{\lambda_4}{\overset{R_4}{\rightleftarrows}} \mathrm{DshFzVang}$$

$$\mathrm{Dsh}^\dagger + \mathrm{FzVang} \underset{A_1^\dagger B^\dagger \lambda_5}{\overset{R_5}{\rightleftarrows}} \mathrm{DshFzVang}$$

$$\mathrm{Fz}^\dagger + \mathrm{VangPk} \underset{\lambda_6}{\overset{R_6}{\rightleftarrows}} \mathrm{FzVangPk}$$

$$\mathrm{FzVang} + \mathrm{Pk} \underset{\lambda_7}{\overset{R_7}{\rightleftarrows}} \mathrm{FzVangPk}$$

$$\mathrm{Dsh}^\dagger + \mathrm{FzVangPk} \underset{A_1^\dagger B^\dagger \lambda_8}{\overset{R_8}{\rightleftarrows}} \mathrm{DshFzVangPk}$$

$$\mathrm{DshFz}^\dagger + \mathrm{VangPk} \underset{\lambda_9}{\overset{R_9}{\rightleftarrows}} \mathrm{DshFzVangPk}$$

$$\mathrm{DshFzVang} + \mathrm{Pk} \underset{\lambda_{10}}{\overset{R_{10}}{\rightleftarrows}} \mathrm{DshFzVangPk}$$

$$\frac{\partial[\mathrm{Dsh}]}{\partial t} = -P_1 - P_5^\dagger - P_8^\dagger + \mu_{\mathrm{Dsh}} \nabla^2 [\mathrm{Dsh}]$$

$$\frac{\partial[\mathrm{Pk}]}{\partial t} = -P_3 - P_7 - P_{10} + \mu_{\mathrm{Pk}} \nabla^2 [\mathrm{Pk}]$$

$$\frac{\partial[\mathrm{Fz}]}{\partial t} = -P_1 - P_2^\dagger - P_6^\dagger + \mu_{\mathrm{Fz}} \nabla^2 [\mathrm{Fz}]_D$$

$$\frac{\partial[\mathrm{Vang}]}{\partial t} = -P_2 - P_3 - P_4 + \mu_{\mathrm{Vang}} \nabla^2 [\mathrm{Vang}]$$

$$\frac{\partial[\mathrm{DshFz}]}{\partial t} = P_1 - P_4^\dagger - P_9^\dagger + \mu_{\mathrm{DshFz}} \nabla^2 [\mathrm{DshFz}]_D$$

$$\frac{\partial[\mathrm{VangPk}]}{\partial t} = P_3 - P_6 - P_9 + \mu_{\mathrm{VangPk}} \nabla^2 [\mathrm{VangPk}]$$

$$\frac{\partial[\mathrm{FzVang}]}{\partial t} = P_2 - P_5 - P_7 + \mu_{\mathrm{FzVang}} \nabla_s^2 [\mathrm{FzVang}]_D$$

$$\frac{\partial[\mathrm{DshFzVang}]}{\partial t} = P_4 + P_5 - P_{10} + \mu_{\mathrm{DshFzVang}} \nabla_s^2 [\mathrm{DshFzVang}]_D$$

$$\frac{\partial[\mathrm{FzVangPk}]}{\partial t} = P_6 + P_7 - P_8 + \mu_{\mathrm{FzVangPk}} \nabla_s^2 [\mathrm{FzVangPk}]_D$$

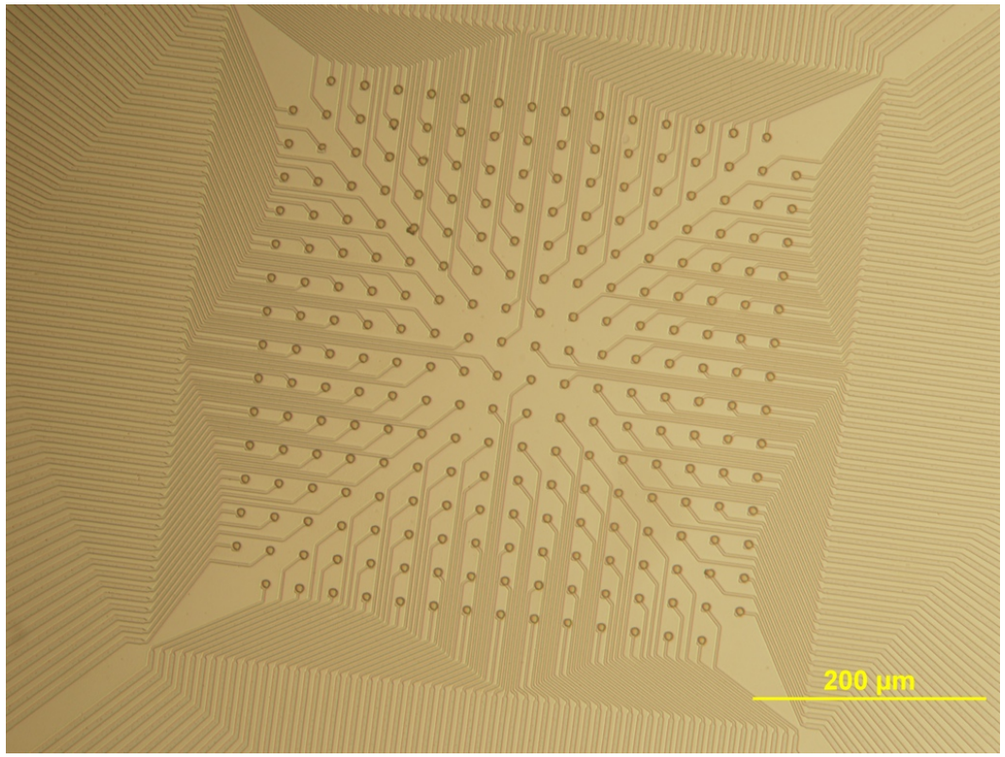
$$\frac{\partial[\mathrm{DshFzVangPk}]}{\partial t} = P_8 + P_9 + P_{10} + \mu_{\mathrm{DshFzVangPk}} \nabla_s^2 [\mathrm{DshFzVangP}$$

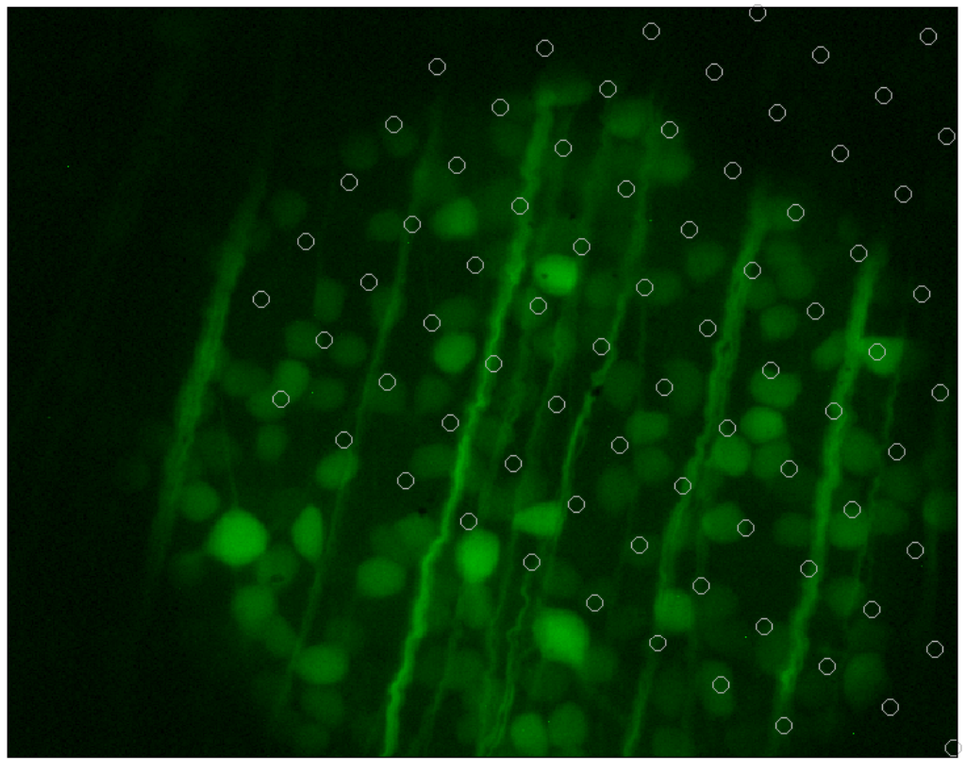
Diverging schools of thought about what is a “proper modeling scale” in life science, less so in physical sciences (RG, TD) or in ML (“nothing special” about physical reality, so long as predictions are good the model is good)

Issues with the phenomenological model: how to relate to perturbation experiments, dependence on microscopic processes lost

Issues with mechanistic model: non-identifiability, what are the relevant processes, overfitting

Studying complete neural populations



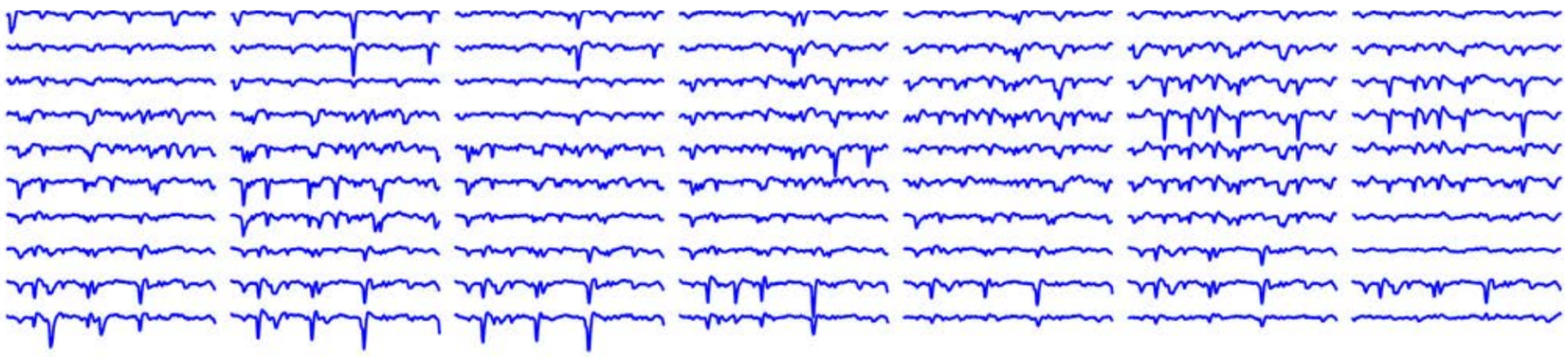


O Marre et al, J Neurosci (2012)

- 252 electrode array, dense spacing, salamander (also rat, guinea pig) retina
- able to record ~200-300 RGCs in a dense patch simultaneously if you can **“spike sort” the signals**
- >90 % coverage
- This is (almost) a complete population encoding the stimuli in a small visual angle

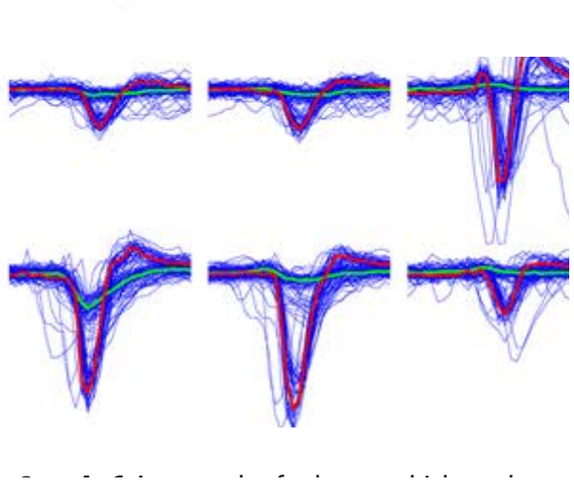
Spike sorting

Example: 70 electrodes, 40ms segment

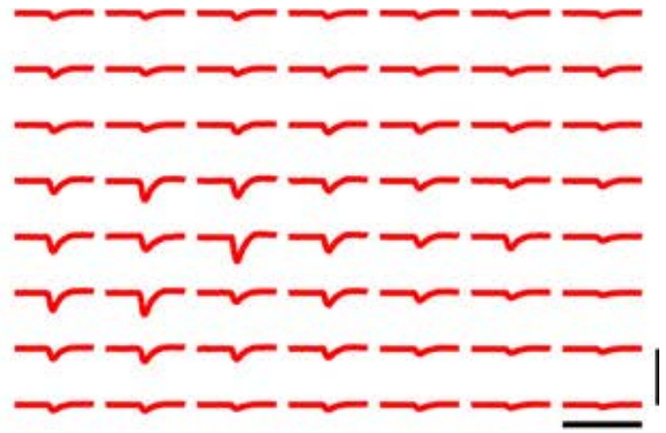


Decompose 256-electrode signal ***s*** into spike templates ***w*** (for neuron *j*) happening at times t_i, corrupted by non-gaussian noise ***e***

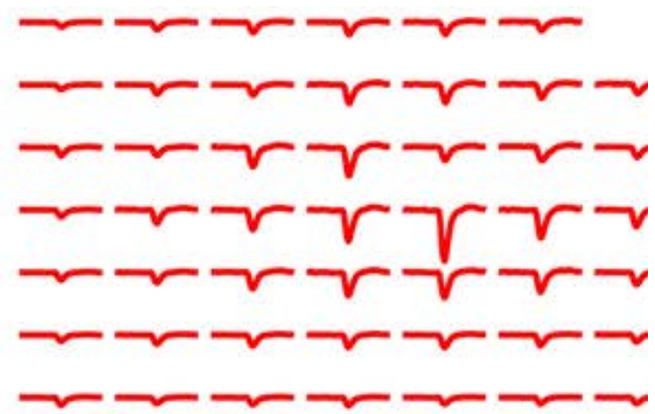
$$\mathbf{s}(t) = \sum_{ij} a_{ij}\mathbf{w}_j(t - t_i) + \mathbf{e}(t)$$



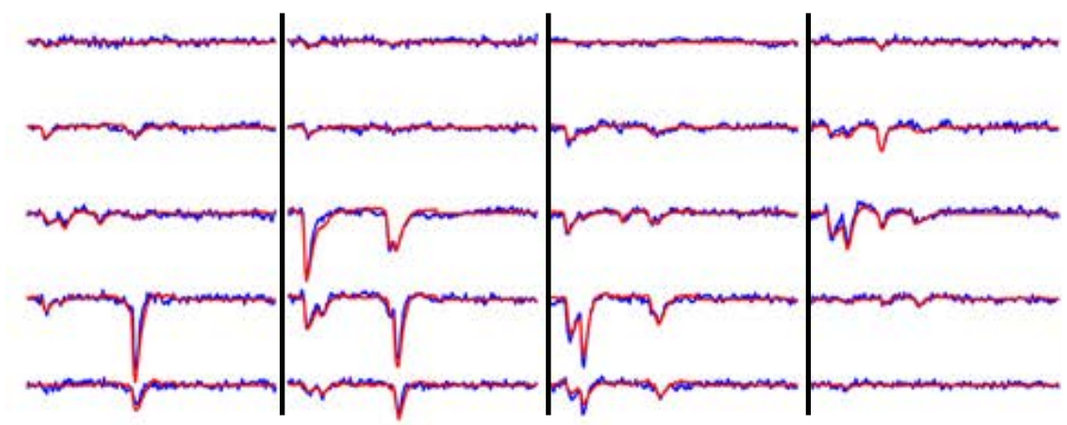
Spike template 1



Spike template 2



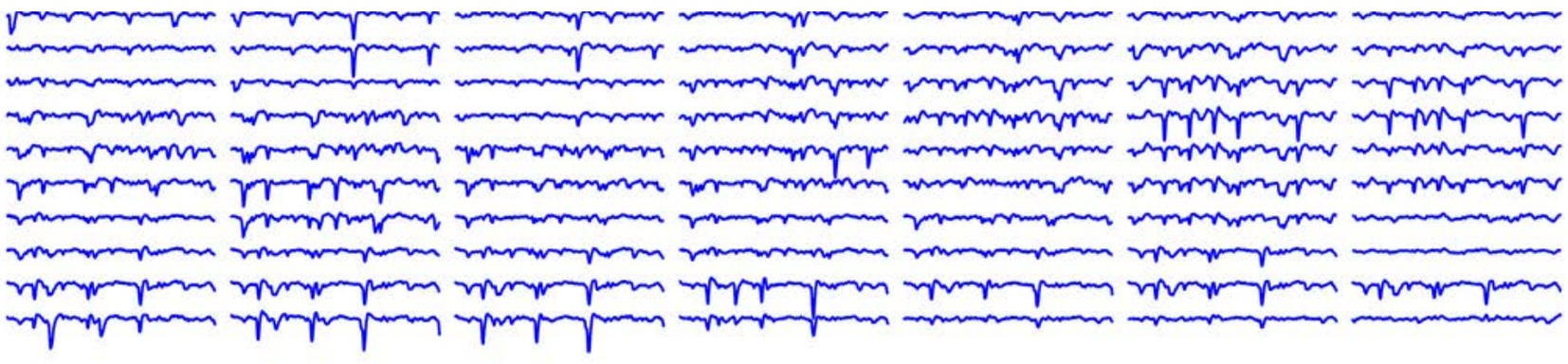
4 examples of signal reconstruction



O Marre et al, J Neurosci (2012)

Spike sorting

Example: 70 electrodes, 40ms segment



Decompose 256-electrode signal ***s*** into spike templates ***w*** (for neuron *j*) happening at times t_i, corrupted by non-gaussian noise ***e***

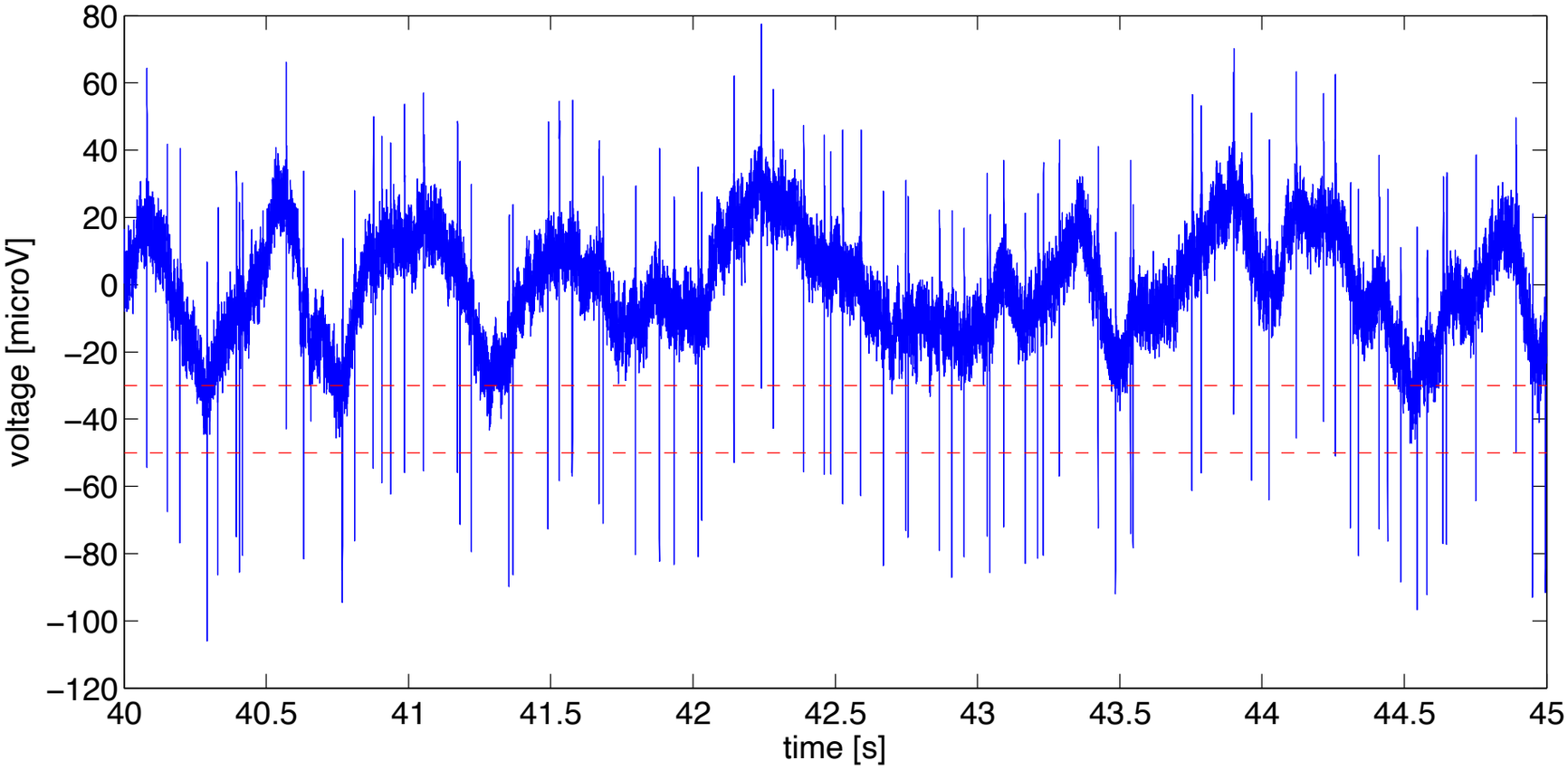
$$\mathbf{s}(t) = \sum_{ij} a_{ij}\mathbf{w}_j(t - t_i) + \mathbf{e}(t)$$

- Joint clustering (identifying templates ***w***) and fitting (identifying t_i, a_{ij})
- Spike signals can overlap in time and space
- Templates (cluster prototypes) vary spike-by-spike
- Non-gaussian, correlated noise
- Physical constraints (spatial locality, additivity of voltage…) need to be taken into account explicitly
- Essentially limited data (by biological constraints): spike-sort each experiment separately
- No extensive “gold truth” or training data, only implicit quality checks on the putative result

Not an out-of-the-box problem in ML, statistics, or signal processing.
Problem-specific solution required (requires lots of details + time) -
and **the best existing solutions are not the cleanest / general ones!**

Looking at low order statistics of the data

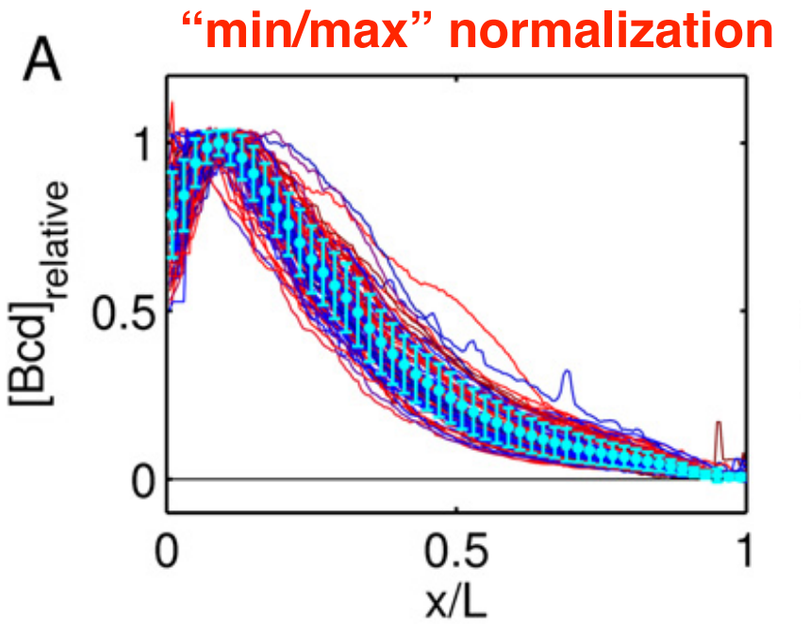
Raw data in detail



- 20 kHz sampling, 2M data points (**trace1**) ~ 100 s worth of data
- fast “noise” fluctuations, slow “baseline” fluctuation, “spikes” (zoom in) [make all these notions more precise later]
- **sanity checks:** beginning / end of data, gaps etc; this already requires some understanding of the data
- do quantities have physical units?
- do absolute zero and y-units have any particular meaning or can they always be undone by **“data centering”** and **“normalization”** (or hidden into [a.u.])?

Excursion: About offsets and arbitrary units

- Try to measure the concentration profile *c(x)* in many embryos, but hard to do directly
- How to undo the measurement-by-measurement "arbitrary" offset/background and scale? [Note a wide range of similar problems: people assigning grades, most measurements of quantitative biology, digital signals etc]



Houchmandzadeh et al, *Nature* (2002)

- **Option 1:** data normalization / alignment (to mean? to zero?)
- **Option 2:** calibration (measure a known zero and a reference signal, or over the whole dynamic range if unsure about systematics) — hard experimentally, but physics standard
- **Option 3:** design analyses that don't depend on such normalization — difficult inference

Basic considerations - what the data actually is about and how it is represented - can matter a lot!
Take the time to understand how exactly the data was generated and represented.

Descriptive statistics

Data $\mathcal{D} = \{\mathbf{x}^t\}, \;\; t = 1, \ldots, N$

Empirical distribution $P_{\text{emp}}(\mathbf{x}) = \frac{1}{N} \sum_{t=1}^{N} \delta\left(\mathbf{x} - \mathbf{x}^t\right)$ (complete summary of data if samples IID)

"True" distribution that generated the data $P(\mathbf{x})$

(underlying notion of a repeatable experiment)

(we often assume a parametric model for this, e.g., Gaussian with *mean* and *variance)*

(Sample) statistic

Some function *f* of the **data** that is informative of the properties of the generating distribution; **estimator** — approximation for a parameter of the distribution

e.g. "mean" estimator $\bar{x} = \frac{1}{N} \sum_{t=1}^{N} x^t$

$$m = \int dx \; xP(x)$$

if I repeatedly sample **IID** data from underlying *P* which has a **true expectation** *m,* then as *N* gets large,

$$\langle \bar{x} - m \rangle = 0 \qquad \text{Std}(\bar{x}) \sim N^{-1/2}$$

unbiased estimator

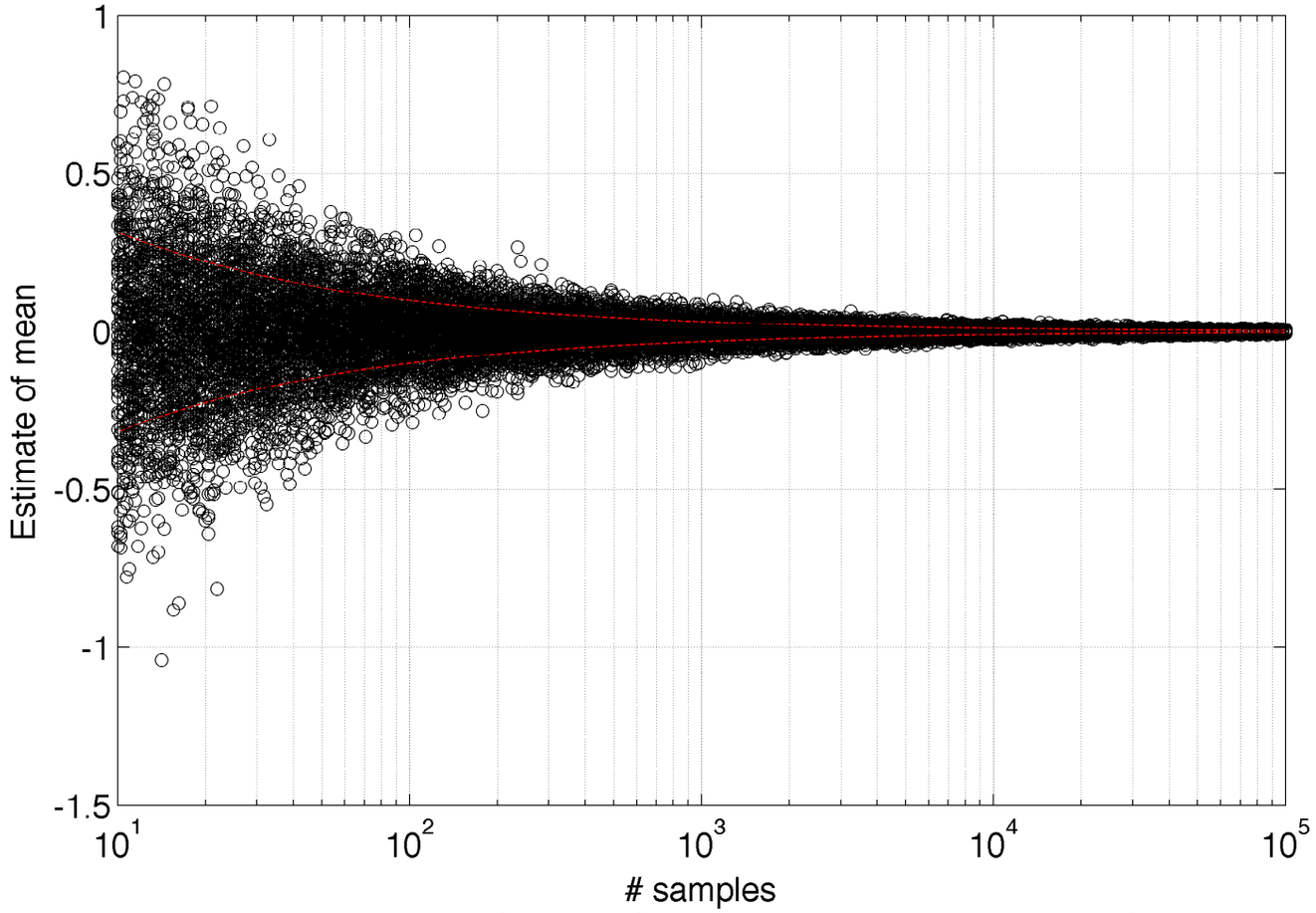
efficient estimator

Descriptive statistics

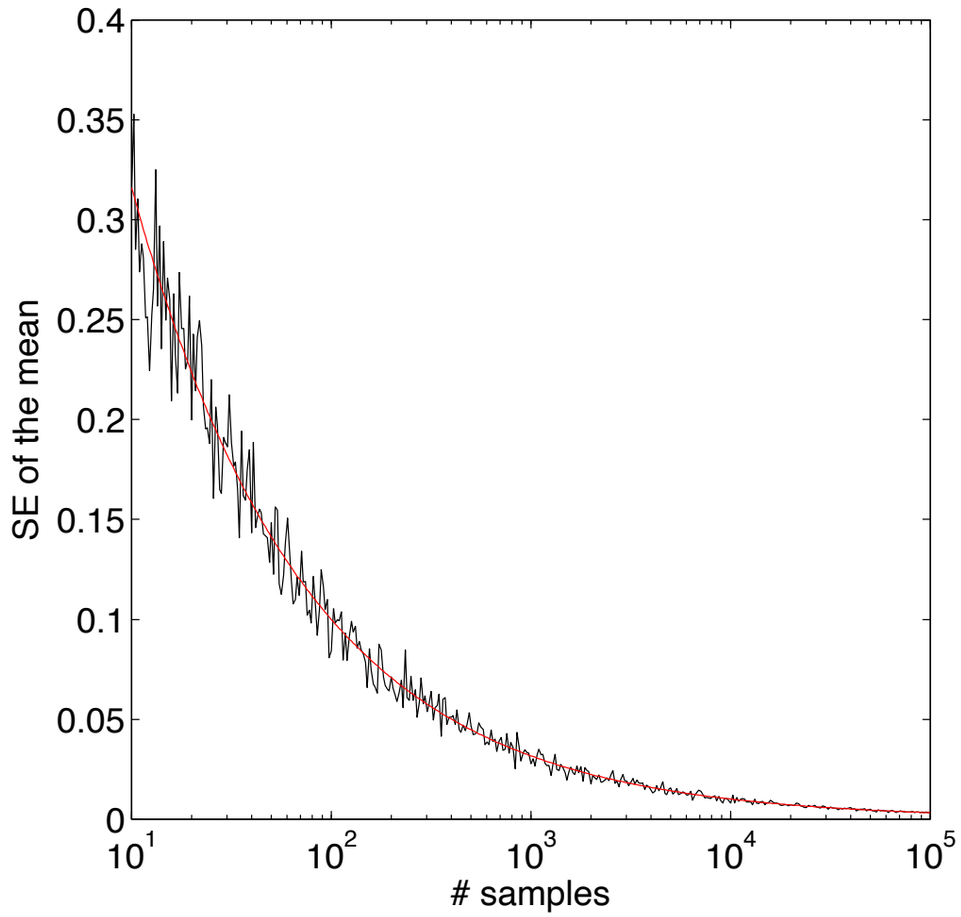
In short, **sample statistics** are functions of the data that have a particular value and an error bar!
(mean and SE, variance, moments, histogram and other counting estimates, entropy, etc.)

A boring but illustrative example: $P(x) \sim N(0,1)$

estimator for the mean approaches 0, as it should



its error drops as *sqrt(N)*, as it should



Recall: variance (or std) is a sample statistic (and a parameter of the Gaussian distribution), while SE is a property of a particular estimator

For Gaussian distributions (and when CLT applies), SE of the mean and variance are related.

Histograms, distributions, PDFs

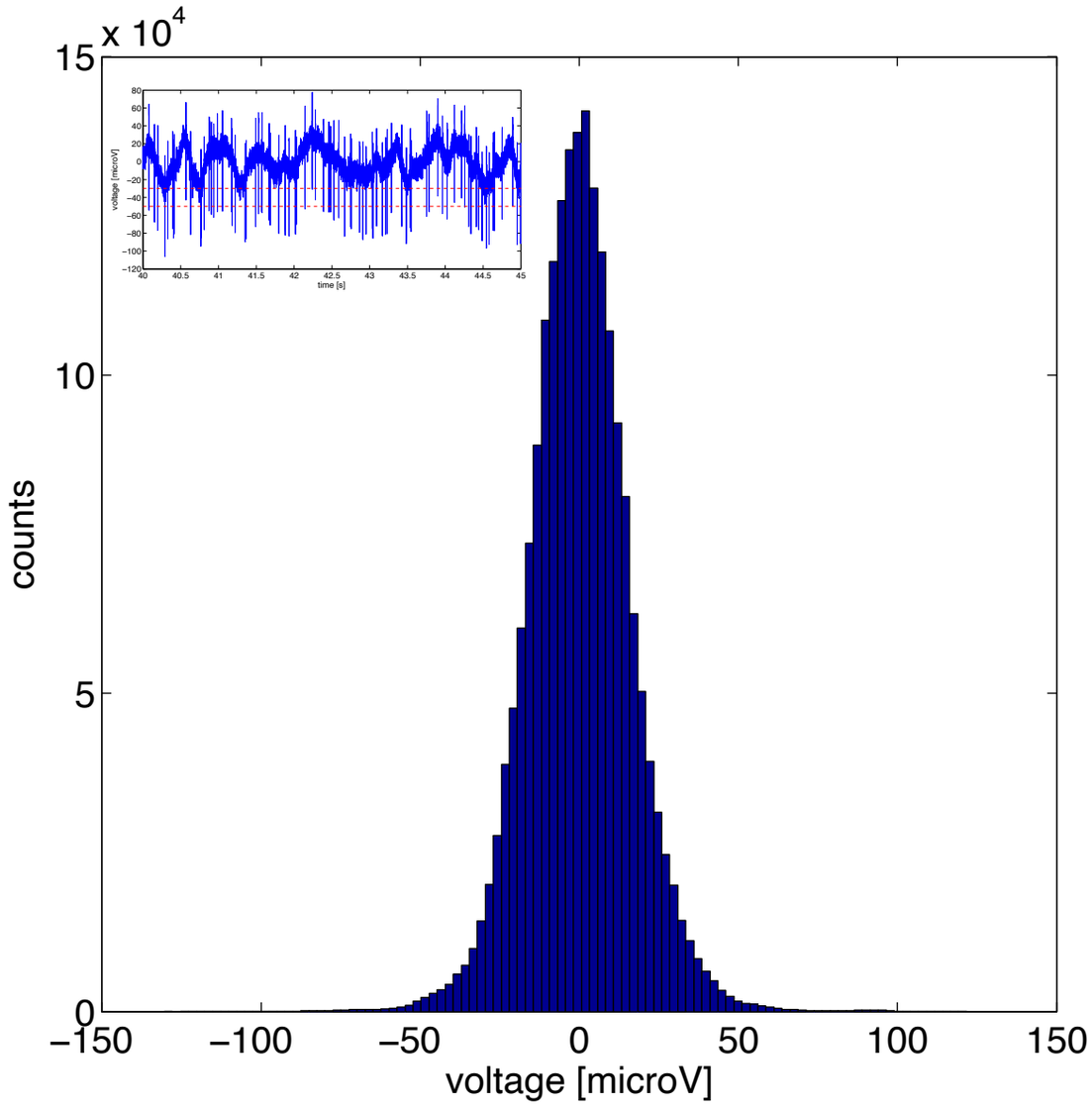
Make sure you know the difference!

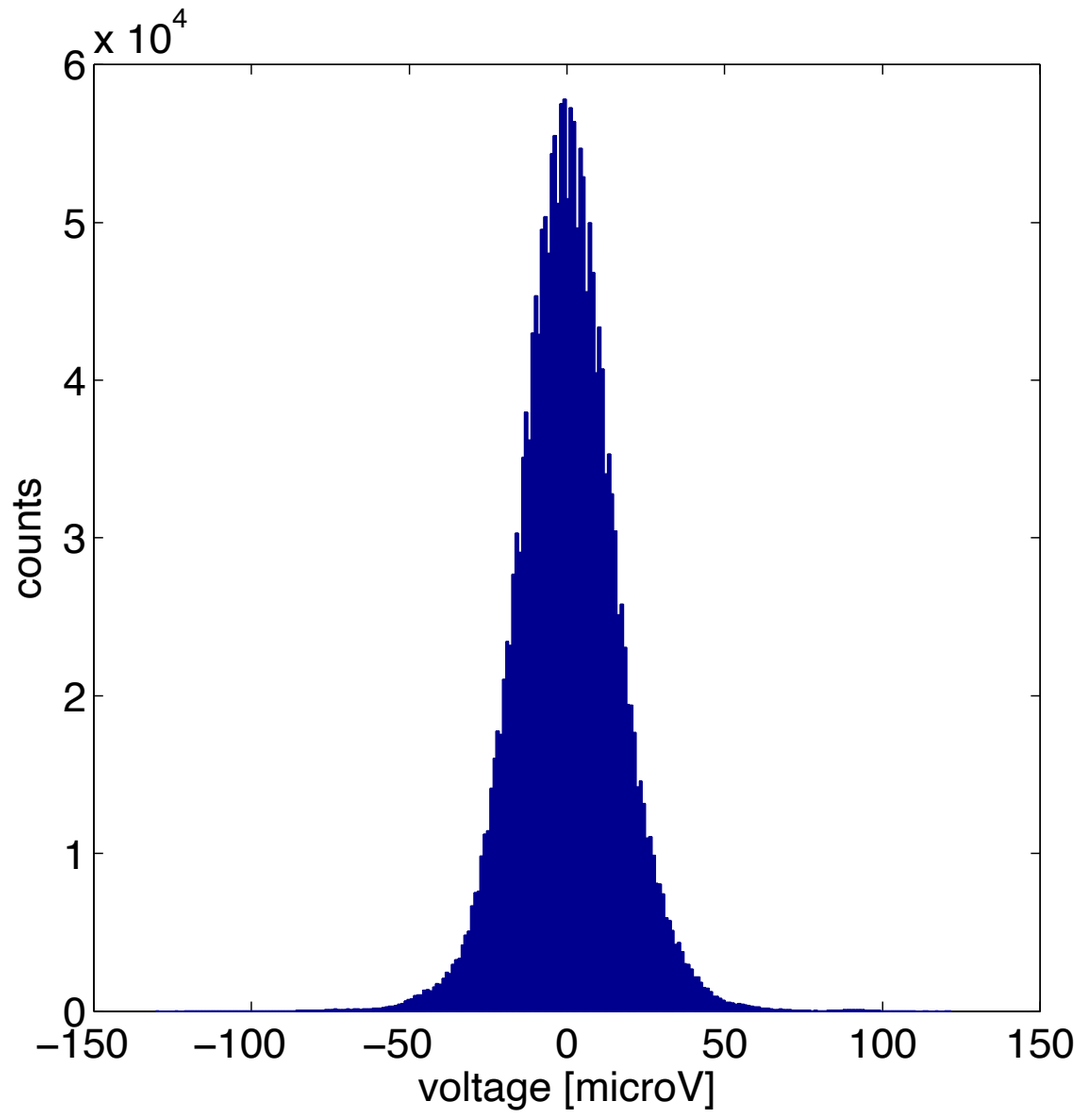
Statistics are all functions of the data and thus of the empirical distribution, P_{emp}.

Thus, before computing *any* statistic (and making a table of numbers), maybe one should visualise this distribution first and then *decide* what statistic is interesting / appropriate.

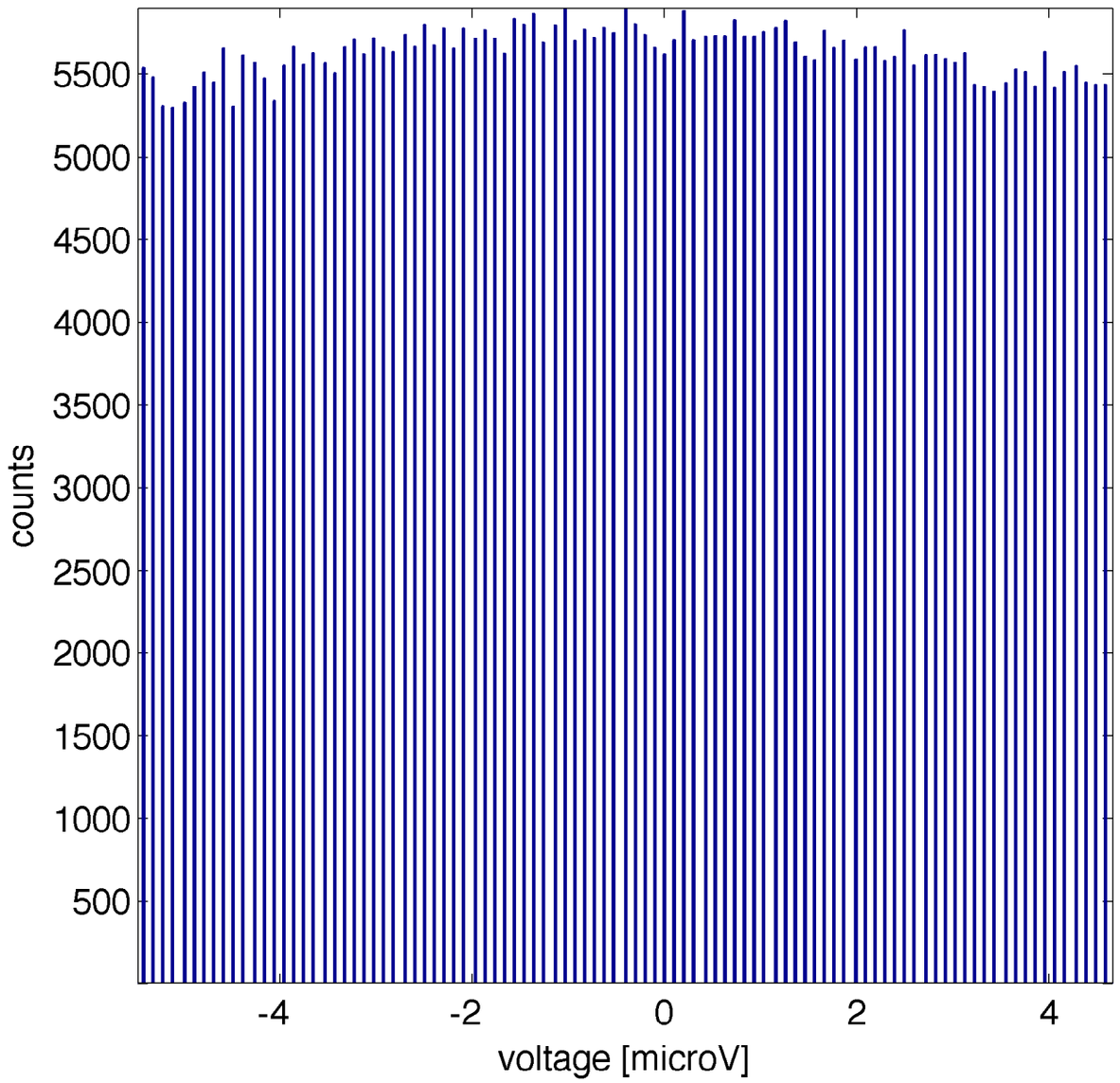
How does one construct and plot empirical distributions?

Histograms



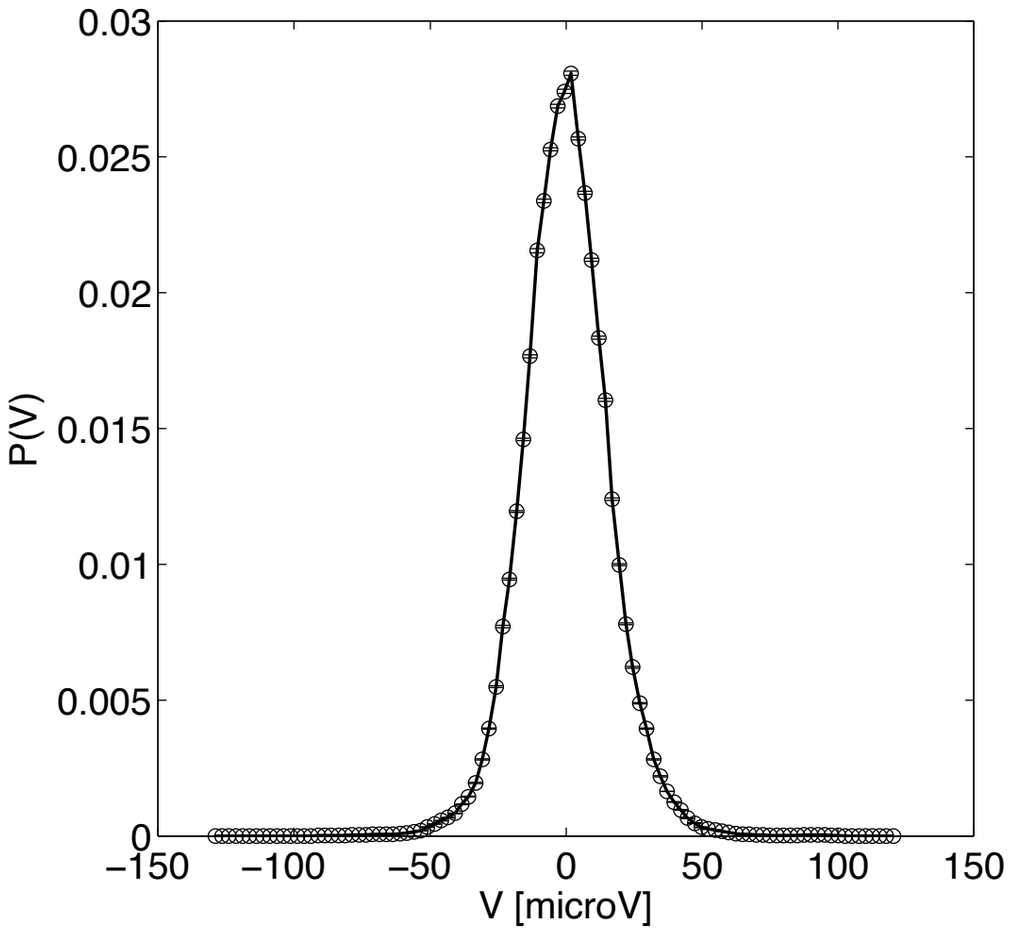


- histograms: depend on x-binning, ~OK if variable is discrete or uniform binning, otherwise?
- **(+)** get a sense of the statistical power
  (-) cannot compare y-values quickly
- in the right histogram (1000 bins), why are there steps at the peak? Maybe we should zoom in..



- So the data really is discrete…?! Not something you'd guess from looking at the time trace.
- numel(unique(data))=2022
- Truly continuous (math only?), ~continuous (experimental sampling), and essentially discrete (e.g., DNA sequences, categorical labels) data and their histograms

Plot empirical PDFs instead of histograms



- **Normalized PDF:** comparable y-values independent of binning!
  (BTW: What are the PDF units?)

- bin centers vs bin edges: use common sense!

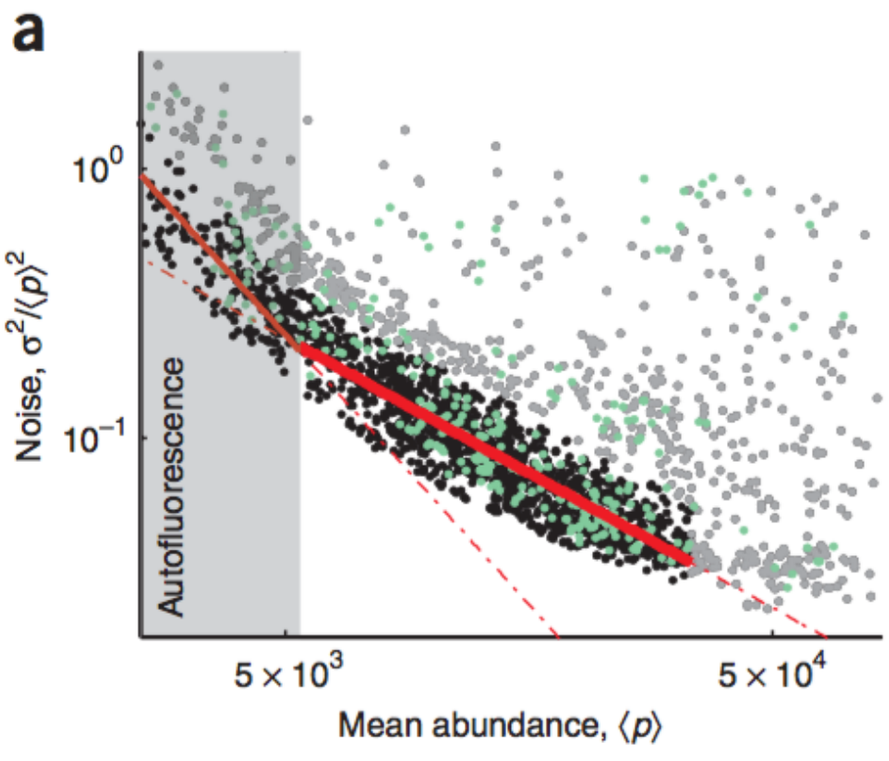
- Plotting distribution on log scale

Excursion: How you visualize the data is very important

nature genetics

Noise in protein expression scales with natural protein abundance

Arren Bar-Even[1], Johan Paulsson[2,3], Narendra Maheshri[4], Miri Carmi[1], Erin O'Shea[4], Yitzhak Pilpel[1] & Naama Barkai[1,5]



Main paper

"Gray points were excluded from the fitting process (Methods)"

Methods

"To obtain the best-fitted curve and define its prefactor, we combined all data points and used an iterative procedure that discards outliers (i.e., points whose distance to the best-fitted line is > 0.5). This process was done using a fixed slope…"

"The precise position of the line separating $1/\langle p\rangle^2$ versus $1/\langle p\rangle$ regions was (arbitrarily) chosen so that the two fitted curves coincide. The right border of the $1/\langle p\rangle$ was chosen by eyeballing."

Comparing empirical PDF to common distributions

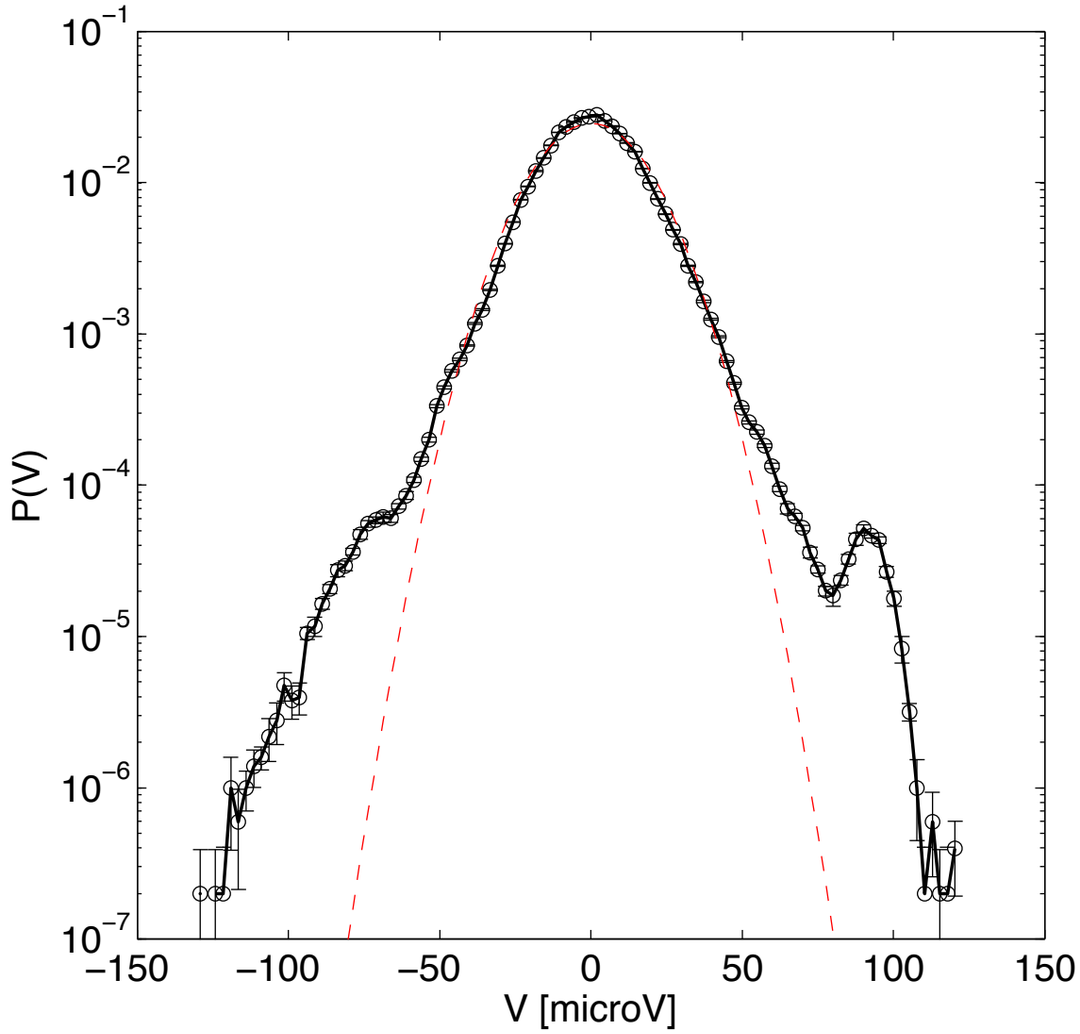
$$\mathcal{N}(x;\bar{x},\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \qquad \sigma^2 = \frac{1}{N}\sum_{t=1}^{N}(x^t - \bar{x})^2$$

variance

$$\mathcal{N}(\mathbf{x};\bar{\mathbf{x}},\mathbf{C}) = (2\pi)^{-D/2}|C|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})\mathbf{C}^{-1}(\mathbf{x}-\bar{\mathbf{x}})}$$

$$M_k = \frac{1}{N}\sum_{t=1}^{N}\frac{(x^t - \bar{x})^k}{\sigma^k}$$

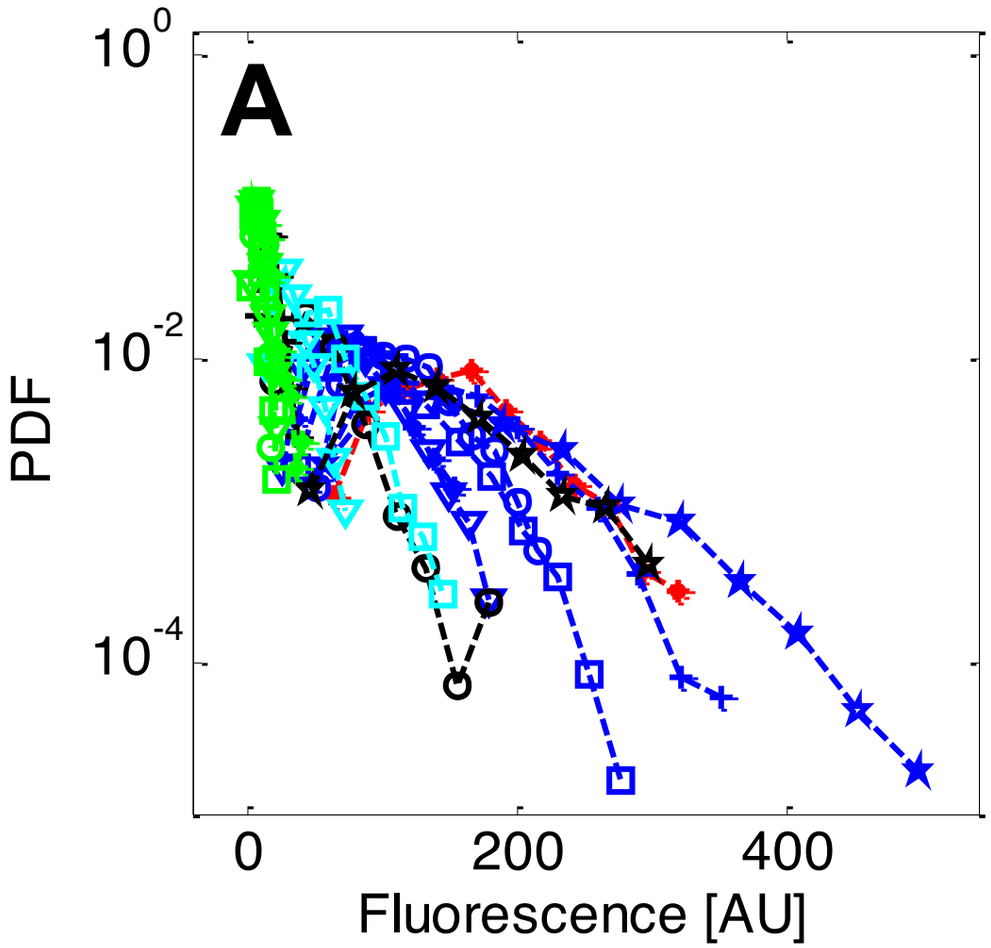
- Symmetric, single-peaked (strong central tendency)
- Gaussian distributions are characterized by first and second moments (mean, variance); higher-order moments can be expressed in terms of low order moments (odd order = 0)
- Central limit theorem
- Maxent distribution (most random distributions consistent with 1st + 2nd order moments)
- Integrals are analytically tractable; marginals are Gaussian
- 1-sigma ~ 68%, 2-sigma ~ 95% weight



Heavier tails compared with Gaussian distribution. Useful baseline reference: once-per-experiment.

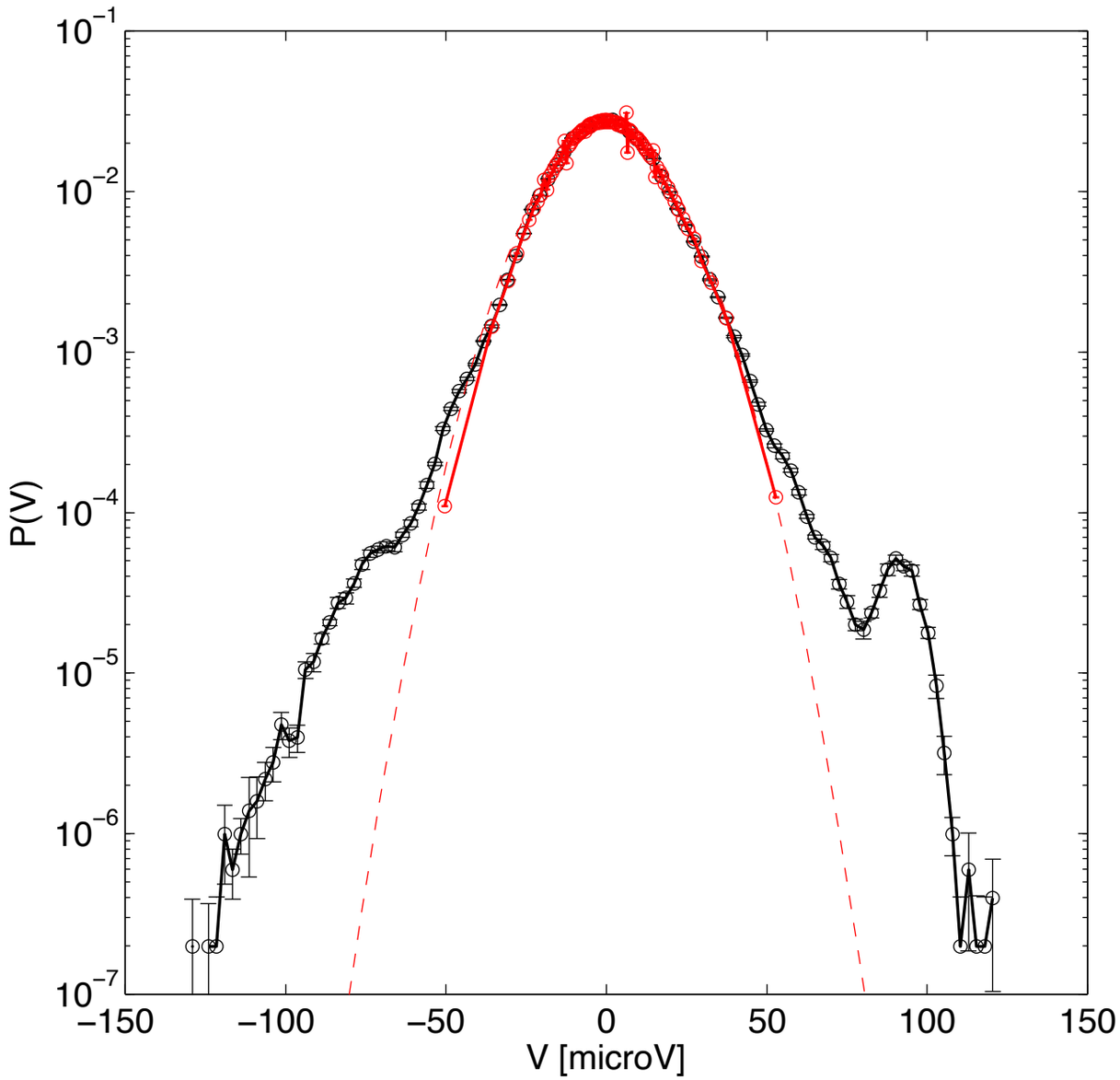
Excursion: Data collapse / non-Gaussianity

- What if we have multiple traces, and thus look at multiple histograms? Example data by Naama Brenner et al (arxiv.org: 1503.01046; *Eur Phys J E 38,* 2015).
- Scale each distribution by the first two moments (mean, variance) - “z-scoring data”…
- Gaussian: standard centered parabola on the semilogy plot

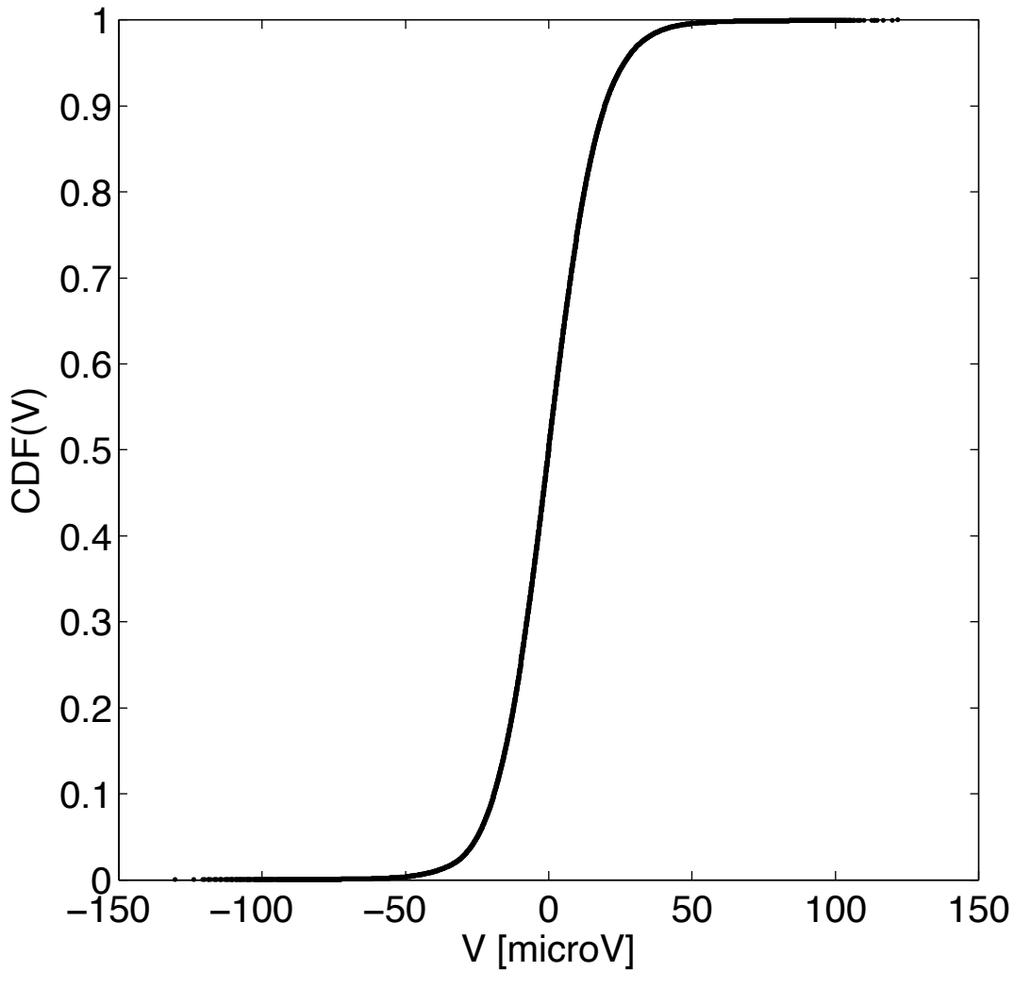


- Useful if:
  (i) experimental variability in mean/variance (cf. scale + offset case in gradients & grading)
  (ii) data is generated by some universal underlying process that is “scaled” or modulated by extrinsic factors (e.g., natural image luminance distributions)
  (iii) data is generated by a hierarchy of stochastic processes

- Another way of binning: equi-populated bins
- now normalized PDF is crucial!
- **equalizes the errors along the variable range**
- (+) for some applications this is desired, e.g., 2D/3D histograms, certain statistical tests
- (-) hard to see what's going on in the tails (sampling + how do you choose the bin centers)



- Can we get rid of binning? What about distributions that contain point masses?
- What does this do: **plot(sort(data,'ascend'),(1:numel(data))./numel(data),'k.');**
- Consider CDFs (cumulative distribution functions)… (CDF is free of units!)



- look at 1-CDF(V) on the log plot to examine the tails
- Easy to read off **quantile statistics (median, 75% quantile etc)**