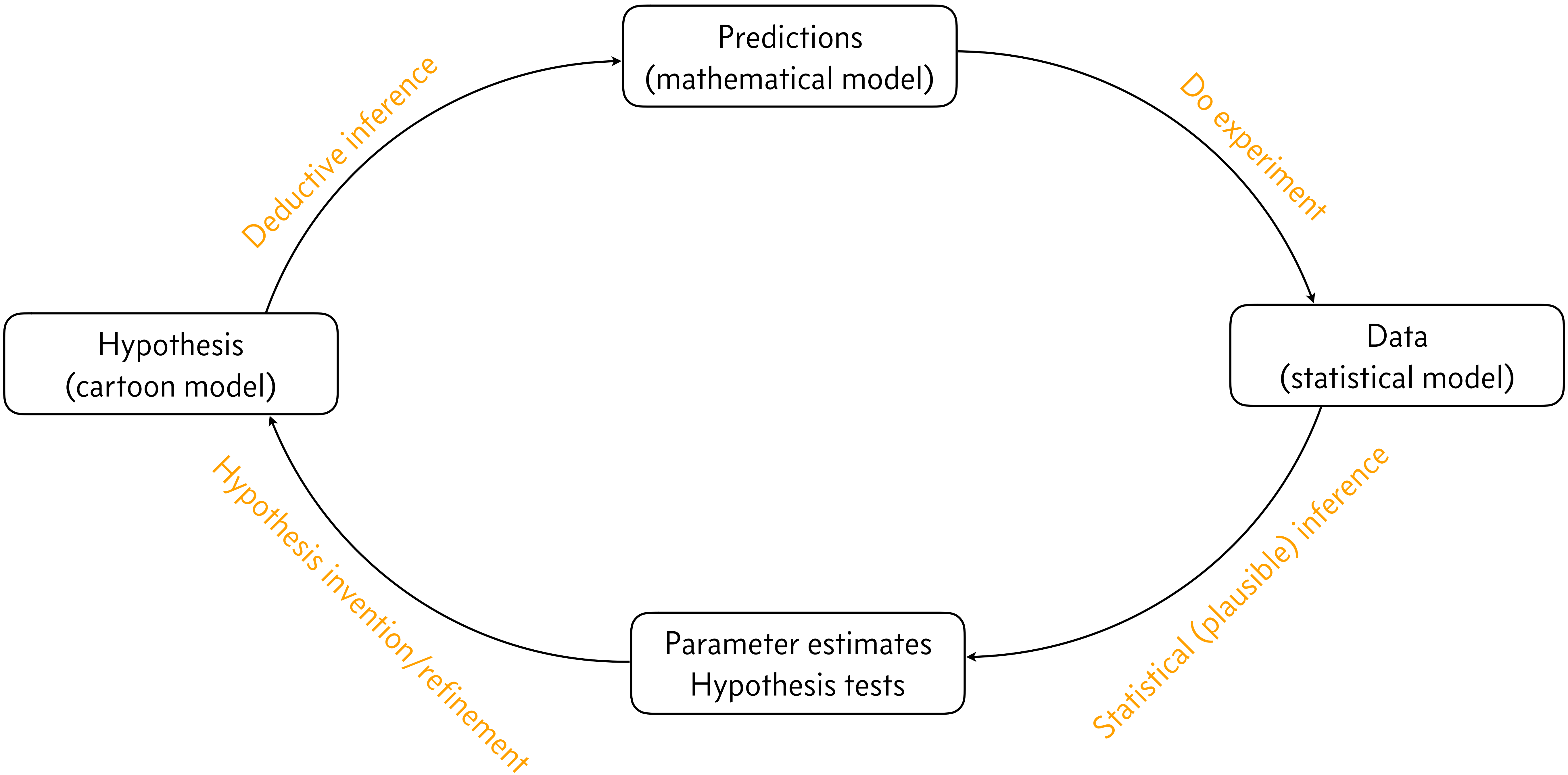
BE/Bi 103

Data Analysis in the Biological Sciences

Fall term, 2017

The scientific method



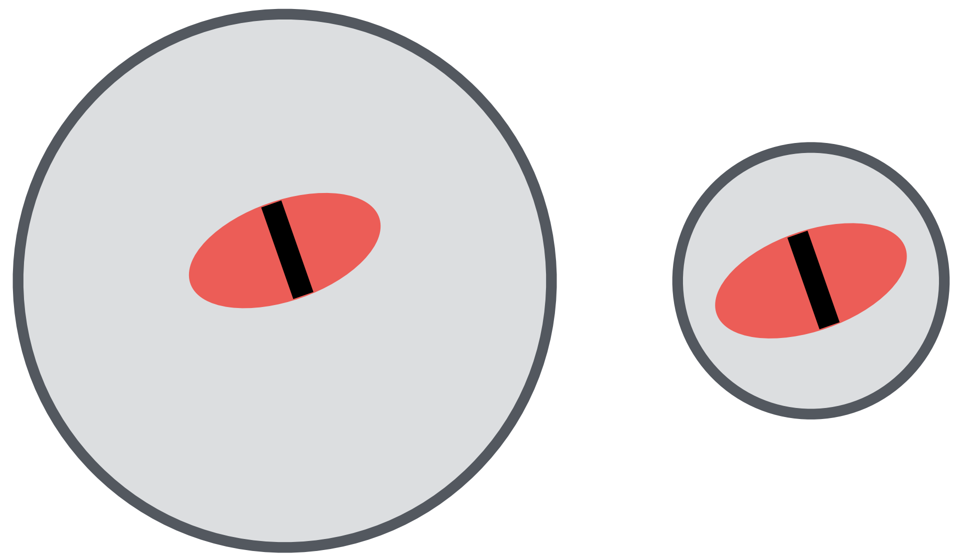
Statistical inference requires a probability theory

Bayes's theorem for parameter estimation:

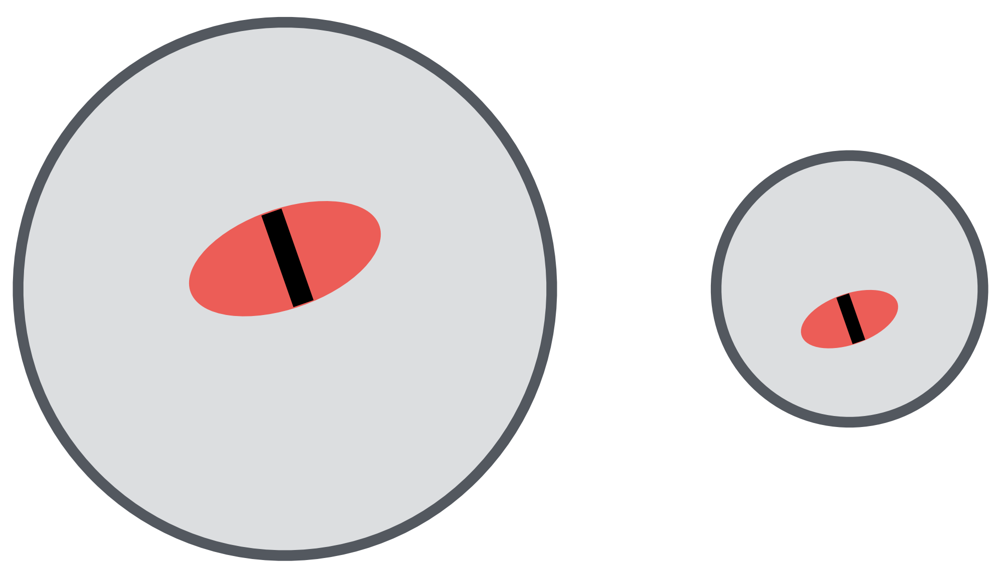
$$\text{posterior} = g(\theta|D, M) = \frac{f(D|\theta, M)\; g(\theta|M)}{f(D|M)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Cartoon models shape our thinking

Model a:



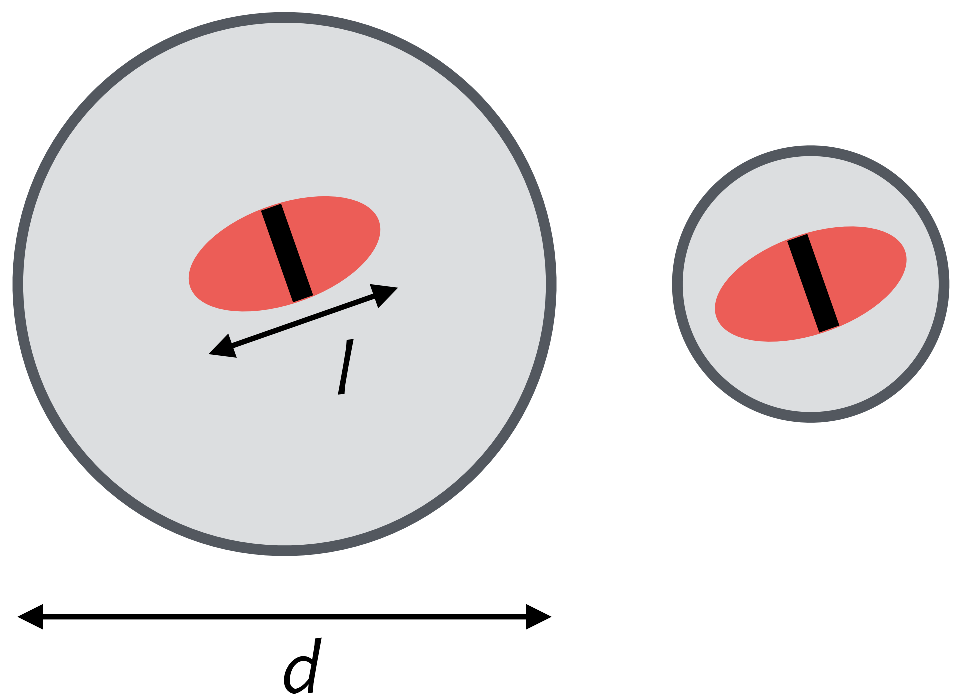
Model b:



Gunawardena, *MBoC*, **25**, 3441, 2014

Mathematical models identify parameters

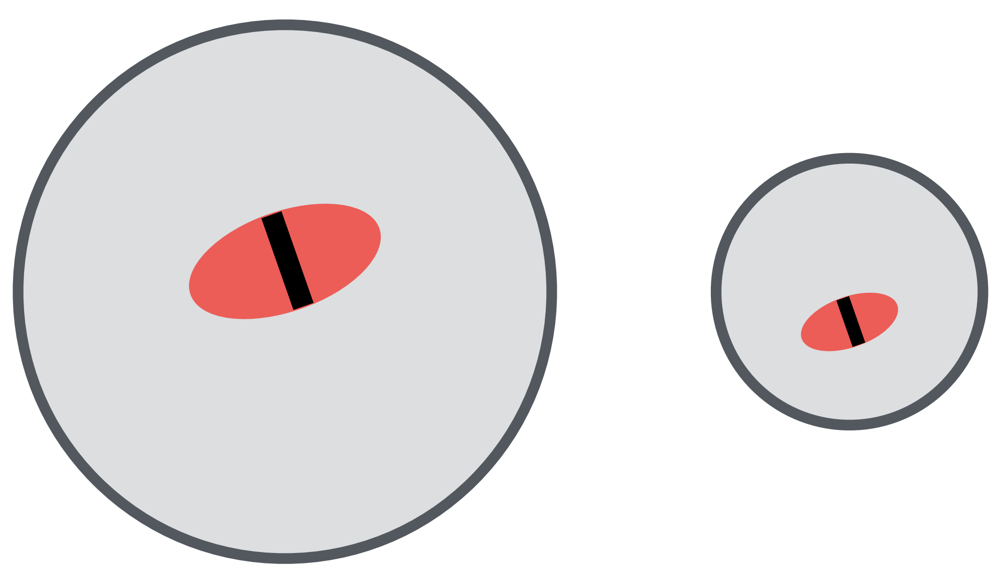
Model a:



$l \neq l(d)$

$l = l_s$

Model b:

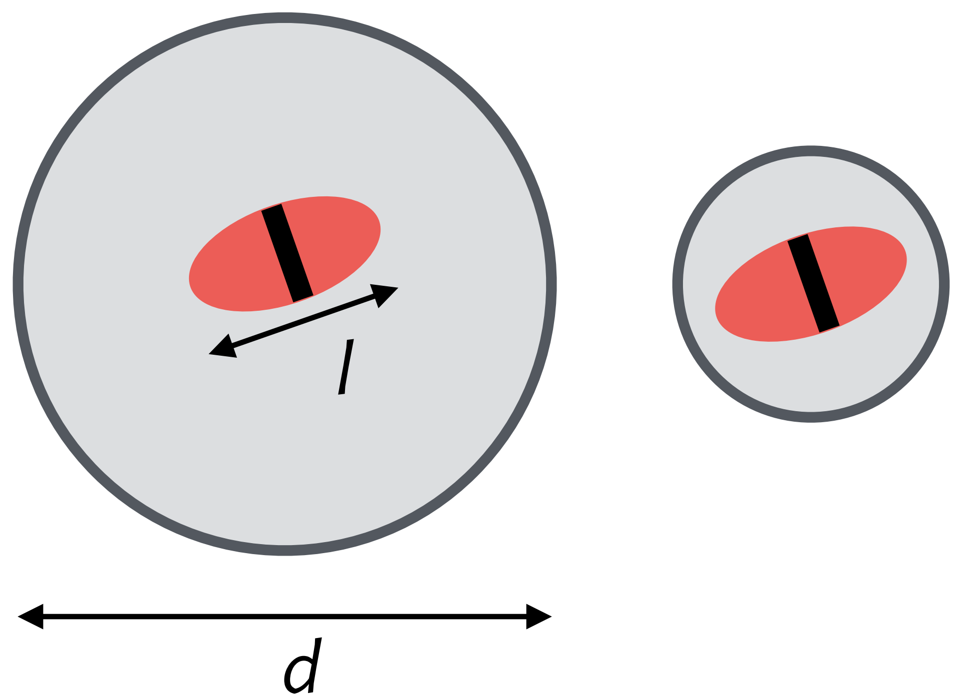


$$l(d;\gamma,\phi) = \frac{\gamma d}{(1+(d/\phi)^3)^{1/3}}$$

Good, et al., *Science*, **342**, 856, 2013

Statistical models are *generative*

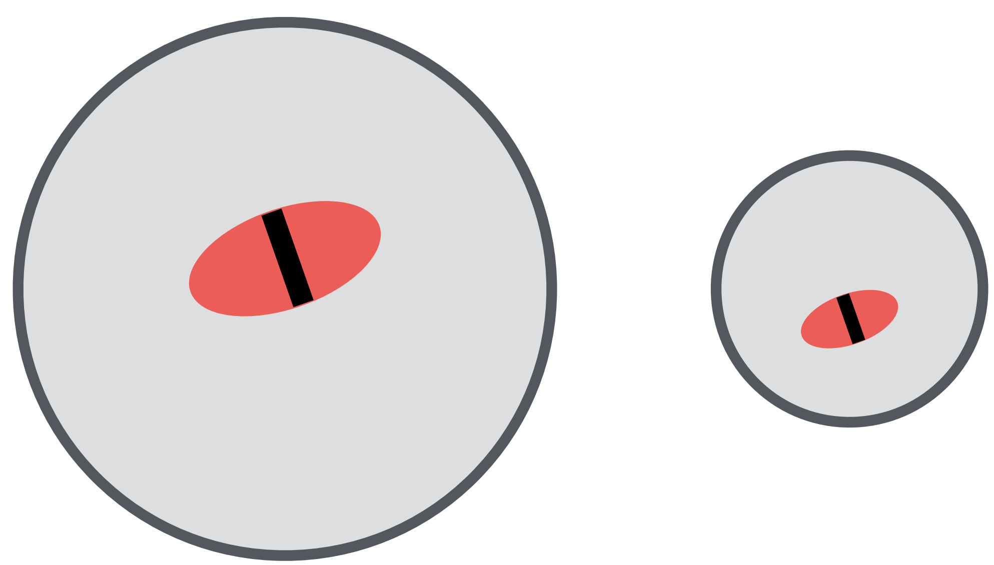
Model a:



$l_i \mid l_s, \sigma \sim \mathrm{Norm}(l_s, \sigma)\ \forall\ i$

$$l(d;\gamma,\phi) = \frac{\gamma d}{(1+(d/\phi)^3)^{1/3}}$$

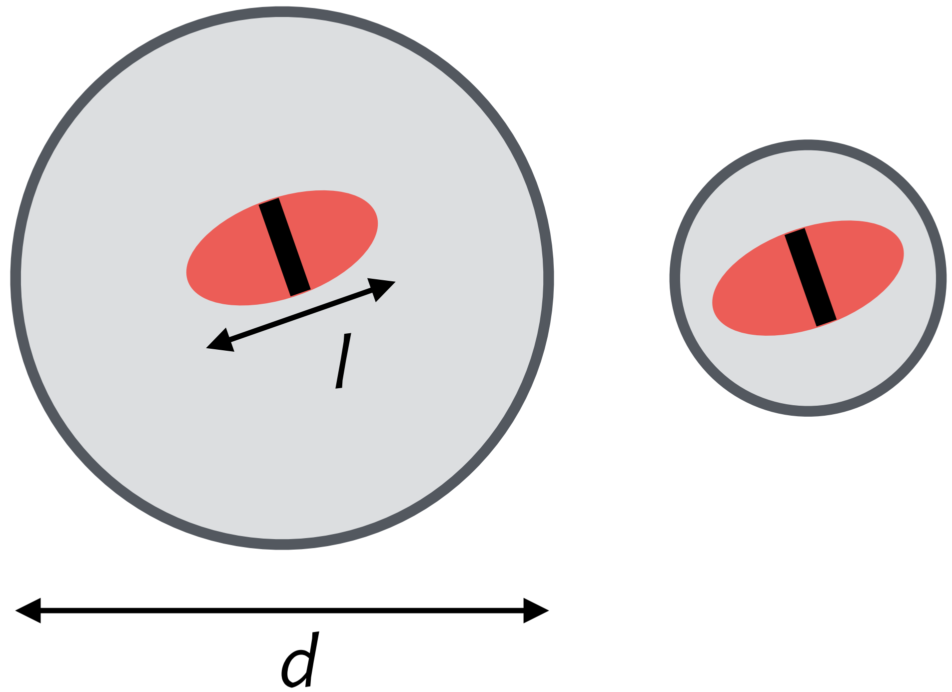
Model b:



$l_i, d_i \mid \gamma, \phi, \sigma \sim \mathrm{Norm}(\mathrm{l}(\mathrm{d};\gamma,\phi), \sigma)\ \forall\ i$

Statistical models need a prior

Model a:

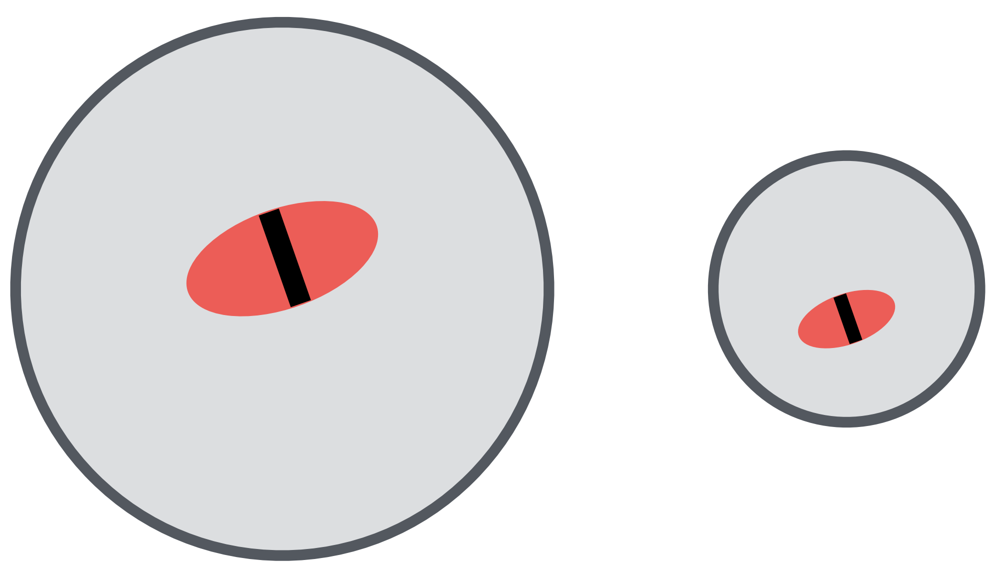


$l_s \sim \text{Uniform}(0, 1\text{ mm})$

$\sigma \sim \text{Jeffreys}$

$l_i \,|\, l_s, \sigma \sim \text{Norm}(l_s, \sigma)\ \forall\, i$

Model b:



$\phi \sim \text{Uniform}(0, 1\text{ mm})$

$\gamma \sim \text{Uniform}(0, 1)$

$\sigma \sim \text{Jeffreys}$

$l_i, d_i \,|\, \gamma, \phi, \sigma \sim \text{Norm}(\text{l}(\text{d};\gamma,\phi), \sigma)\ \forall\, i$

Allen Downey @AllenDowney · Nov 17

If I tell you my likelihoods are based on a truckload of subjective modeling decisions, nobody panics. But when I say that my prior is based on one little assumption, everyone loses their minds!

> **Data Science Fact** @DataSciFact
>
> "By the time we've reached thinking about priors, we are already two or three levels of ad hociness down the hole. What's a little more?" — Matt Briggs

1 12 26

Given the statistical model and the data,
the posterior is completely determined.

All of the “work” of inference is computing it!

We can sometimes express the posterior analytically

Repeated measurements

$$f(\mathbf{x} \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

$$g(\mu) = \begin{cases} (\mu_{\text{max}} - \mu_{\text{min}})^{-1} & \mu_{\text{min}} < \mu < \mu_{\text{max}}, \\ 0 & \text{otherwise,} \end{cases}$$

$$g(\sigma \mid I) = \begin{cases} (\ln(\sigma_{\text{max}}/\sigma_{\text{min}})\,\sigma)^{-1} & \sigma_{\text{min}} < \sigma < \sigma_{\text{max}} \\ 0 & \text{otherwise.} \end{cases}$$

We can sometimes express the posterior analytically

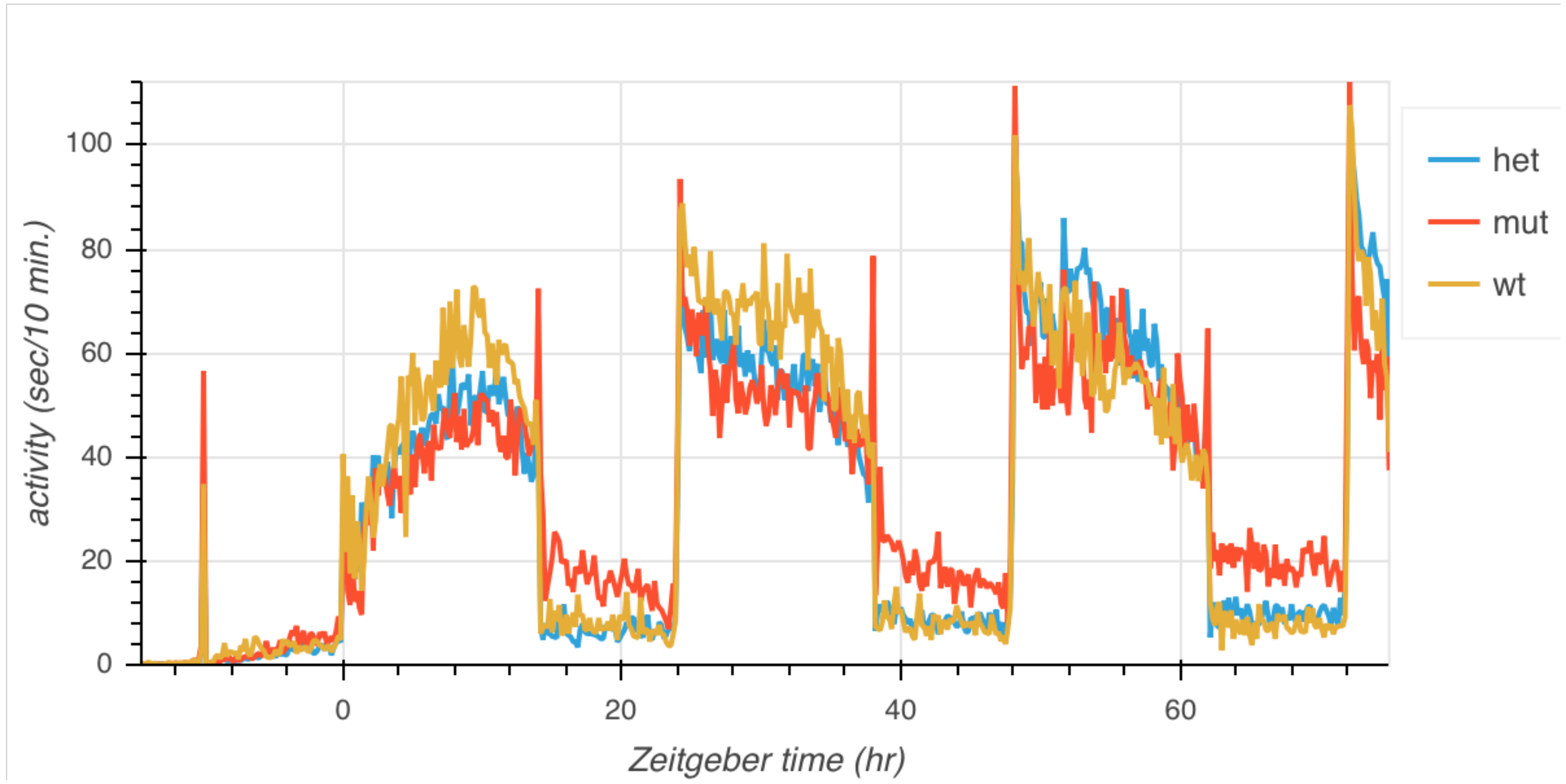
Repeated measurements

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

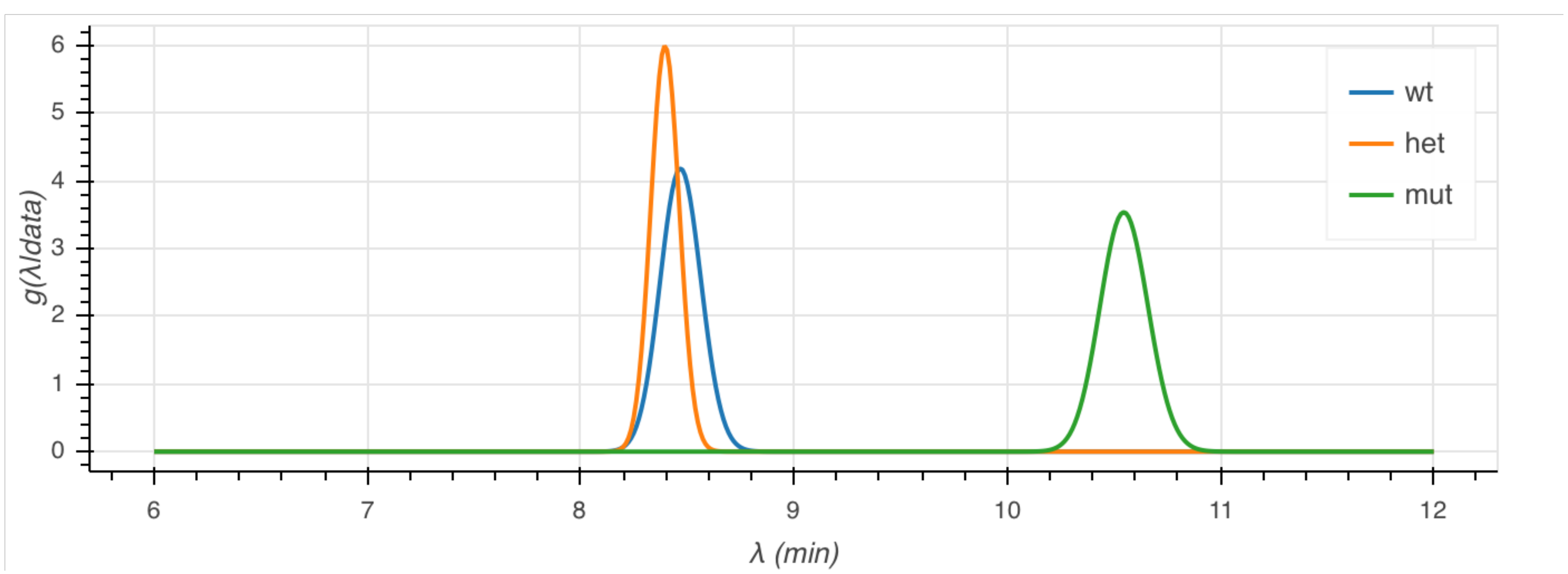
$$r^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$g(\mu \mid \mathbf{x}) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\,\frac{1}{r}\,\left(1 + \frac{(\mu - \bar{x})^2}{r^2}\right)^{-\frac{n}{2}}$$

$$g(\sigma \mid \mathbf{x}) = \frac{\left(nr^2\right)^{(n-1)/2}}{2^{(n-3)/2}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\,\exp\left[-\frac{nr^2}{2\sigma^2}\right]$$



activity (sec/10 min.)
Zeitgeber time (hr)
het
mut
wt



g(λ|data)
λ (min)
wt
het
mut

The posterior may sometimes be approximated as Gaussian

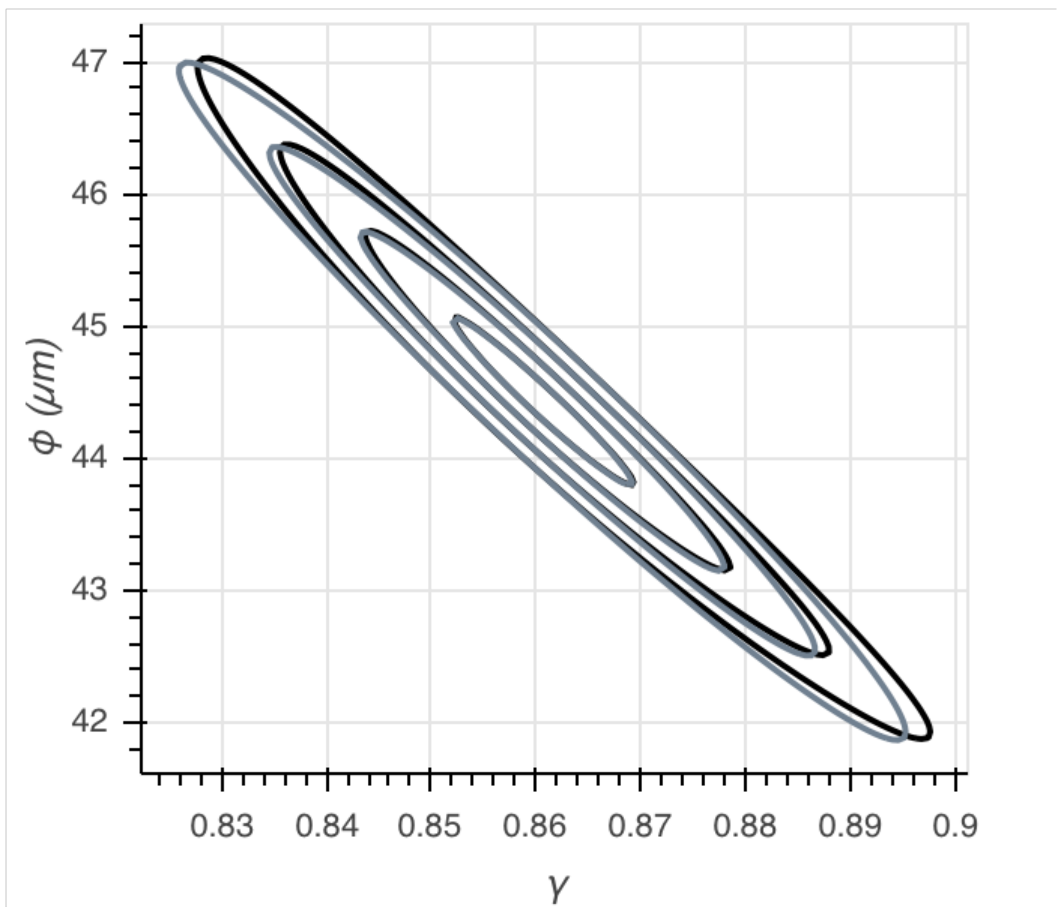
1. Find the most probable parameters θ^* (the MAP).

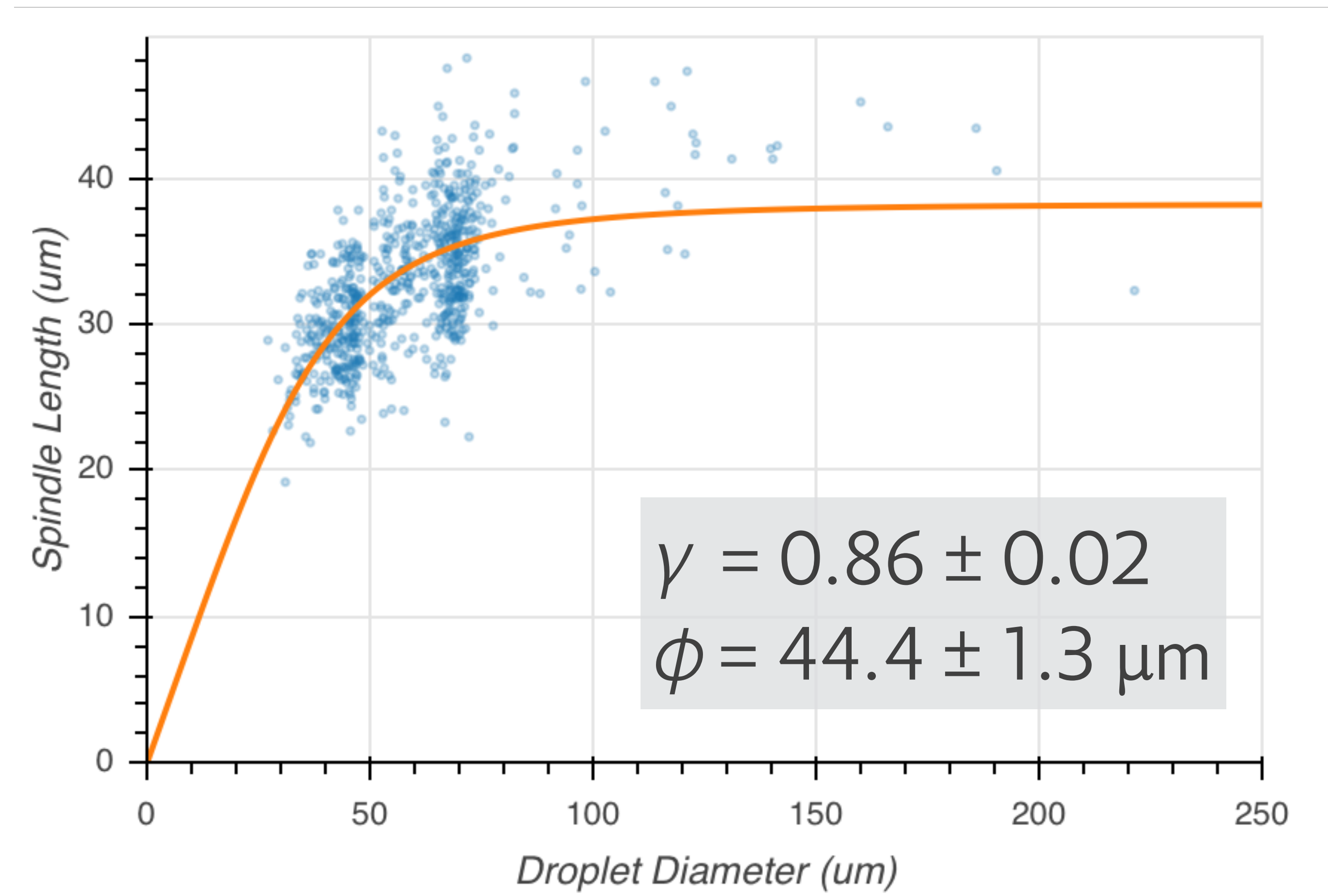
2. Approximate the posterior $g(\theta^*|D)$ as Gaussian by doing a Taylor expansion of $\ln g(\theta^*|D)$ about θ^*.

3. The covariance matrix is the negative inverse of the Hessian of $\ln g(\theta^*|D)$.

Obvious assumption: posterior is approximately Gaussian.

The posterior may sometimes be approximated as Gaussian



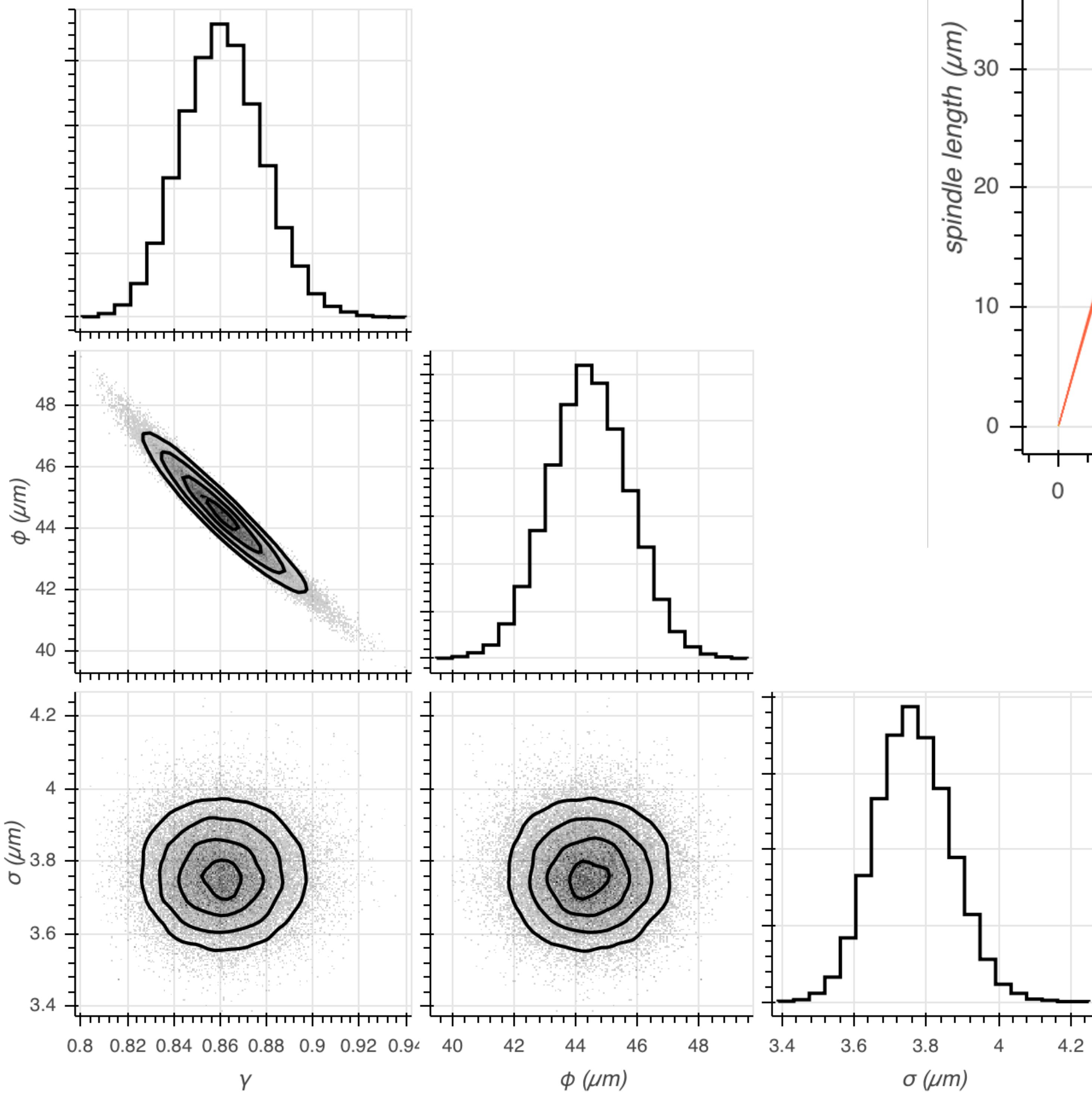


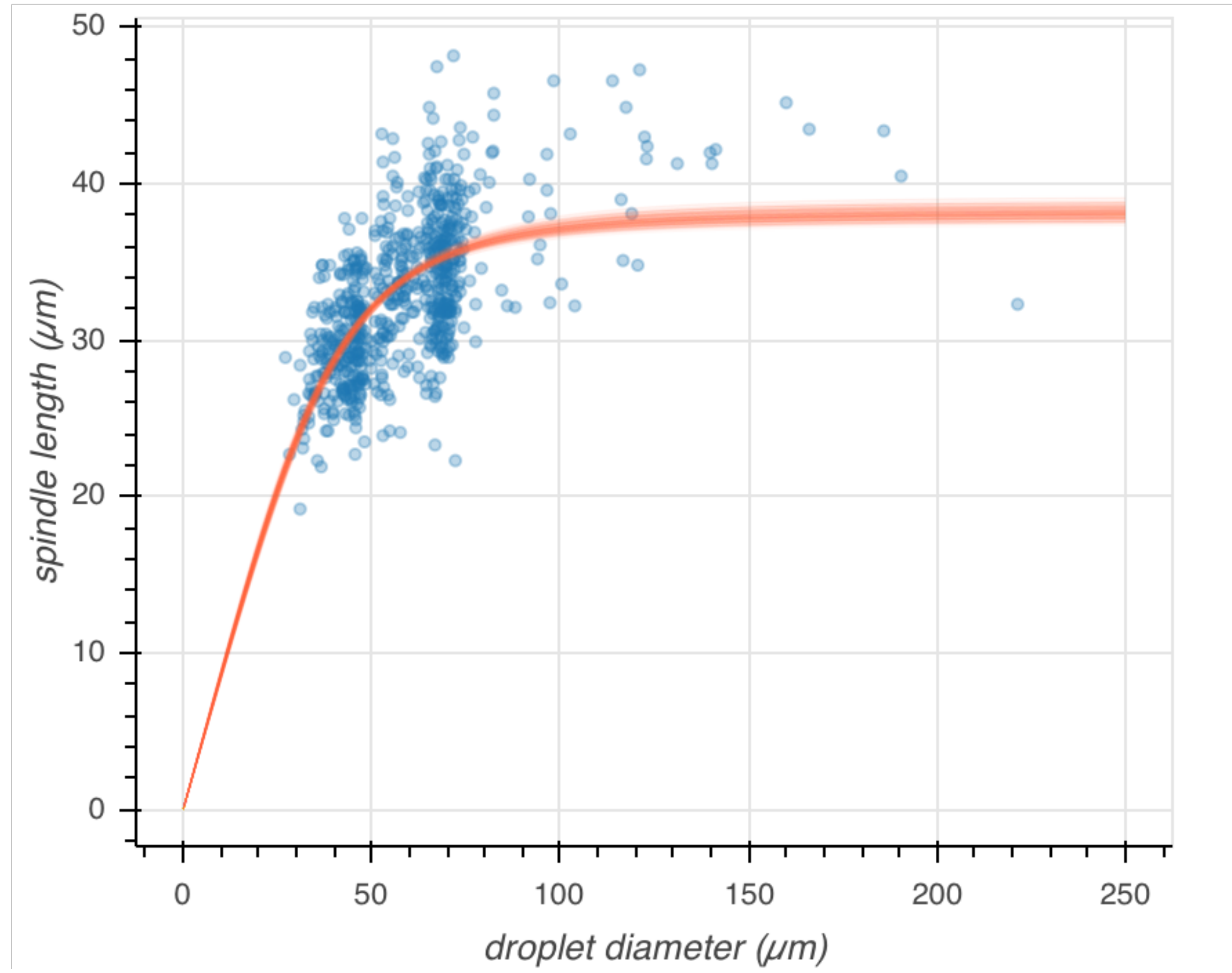
The posterior may be *sampled* using MCMC

1. Define the (log) posterior distribution

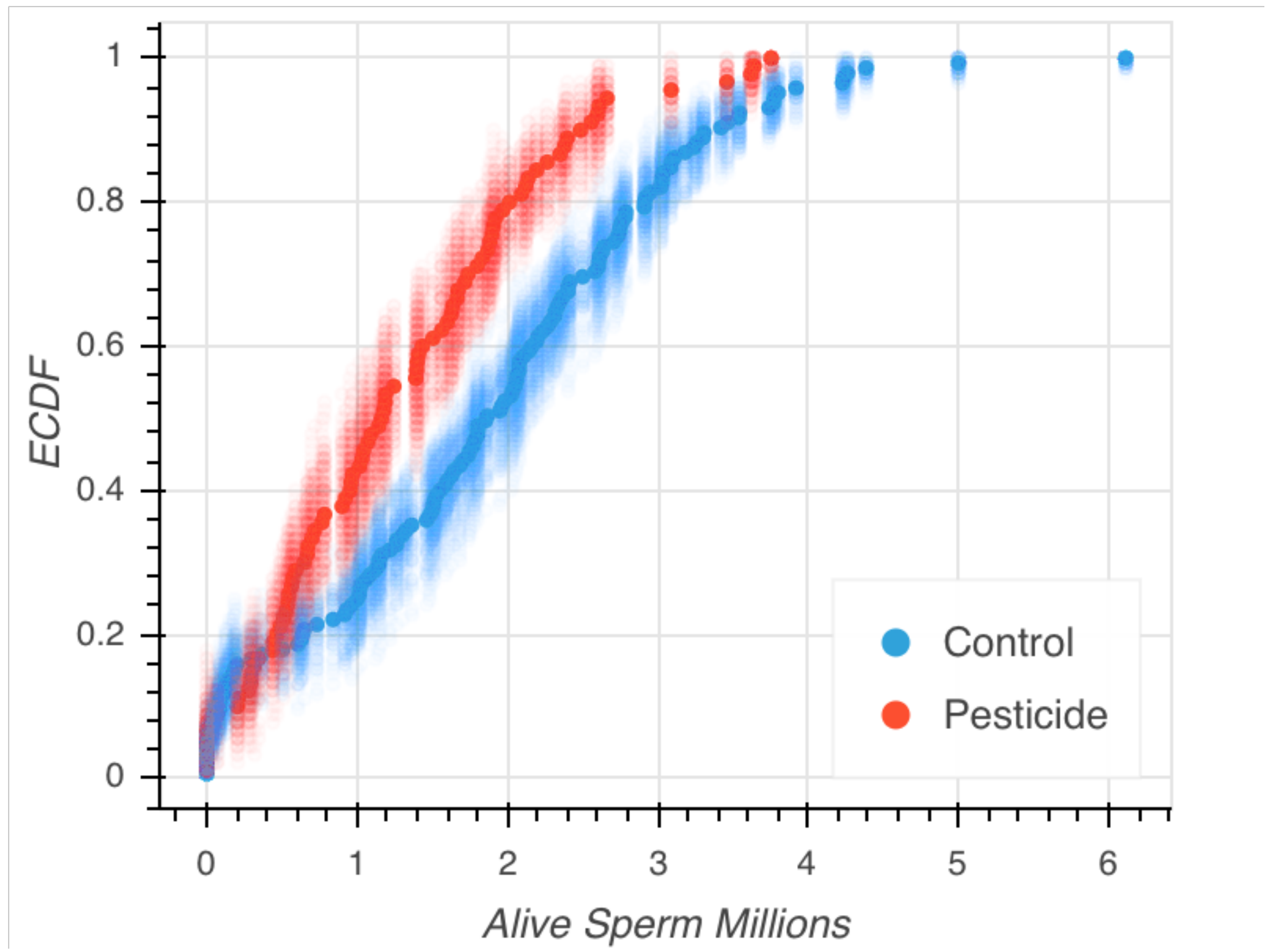
2. Efficiently sample the posterior with an ergodic, positively recurrent Markov chain

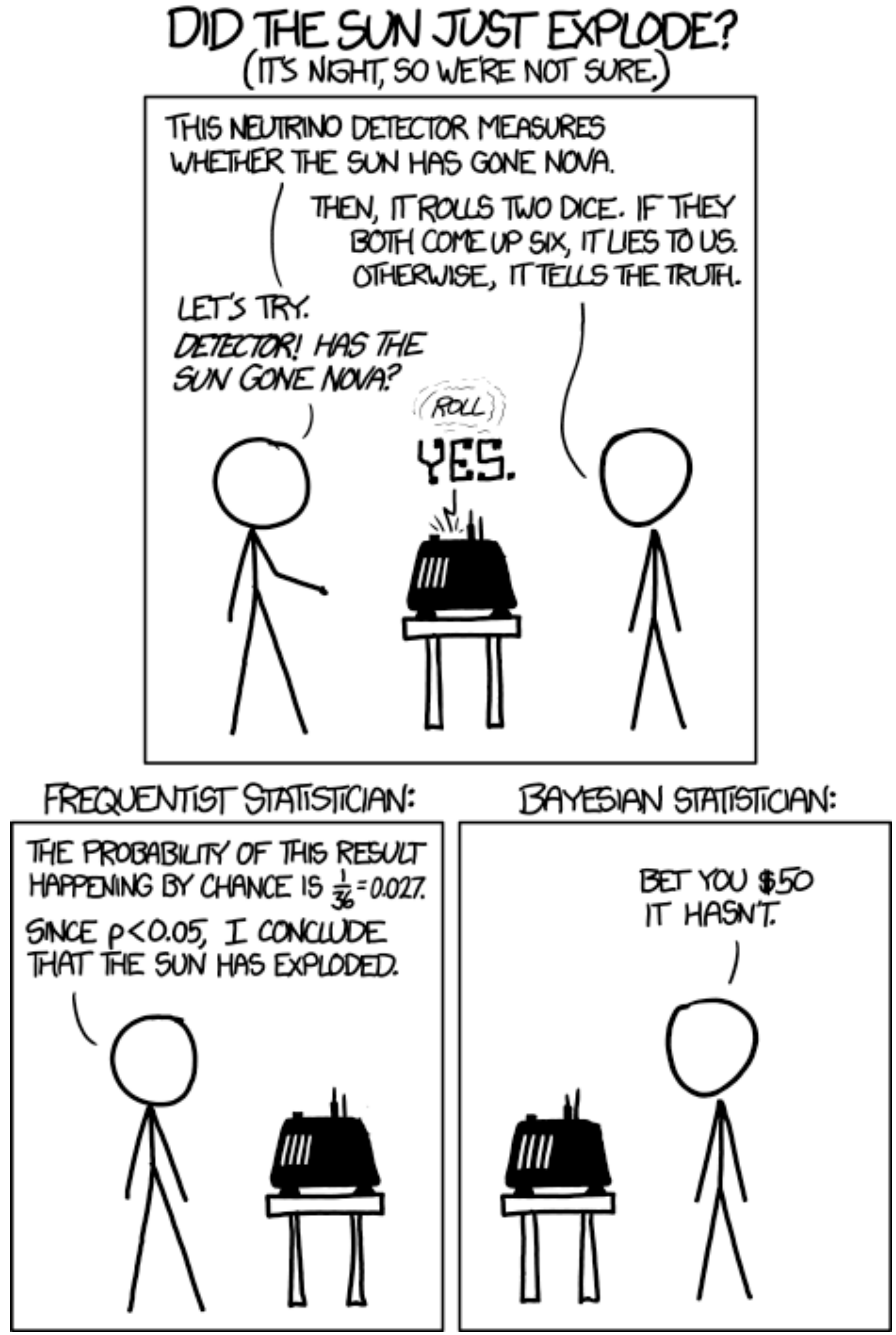
3. Obtain marginalized posterior by considering specific parameters.


48
46
44
42
40
φ (μm)
4.2
4
3.8
3.6
3.4
σ (μm)
0.8
0.82
0.84
0.86
0.88
0.9
0.92
γ
40
42
44
46
48
φ (μm)
3.4
3.6
3.8
4
4.2
σ (μm)

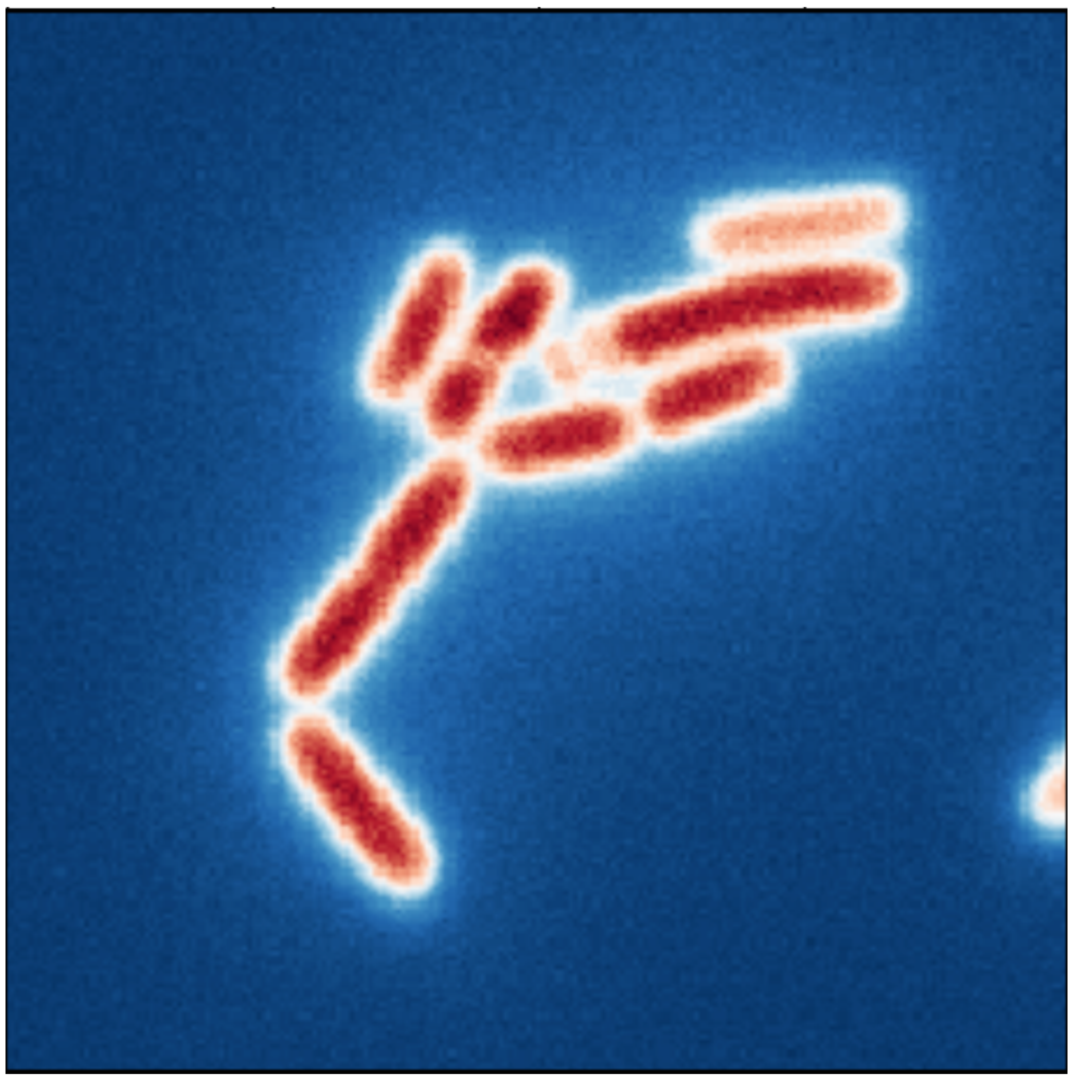

50
40
30
20
10
0
spindle length (μm)
0
50
100
150
200
250
droplet diameter (μm)

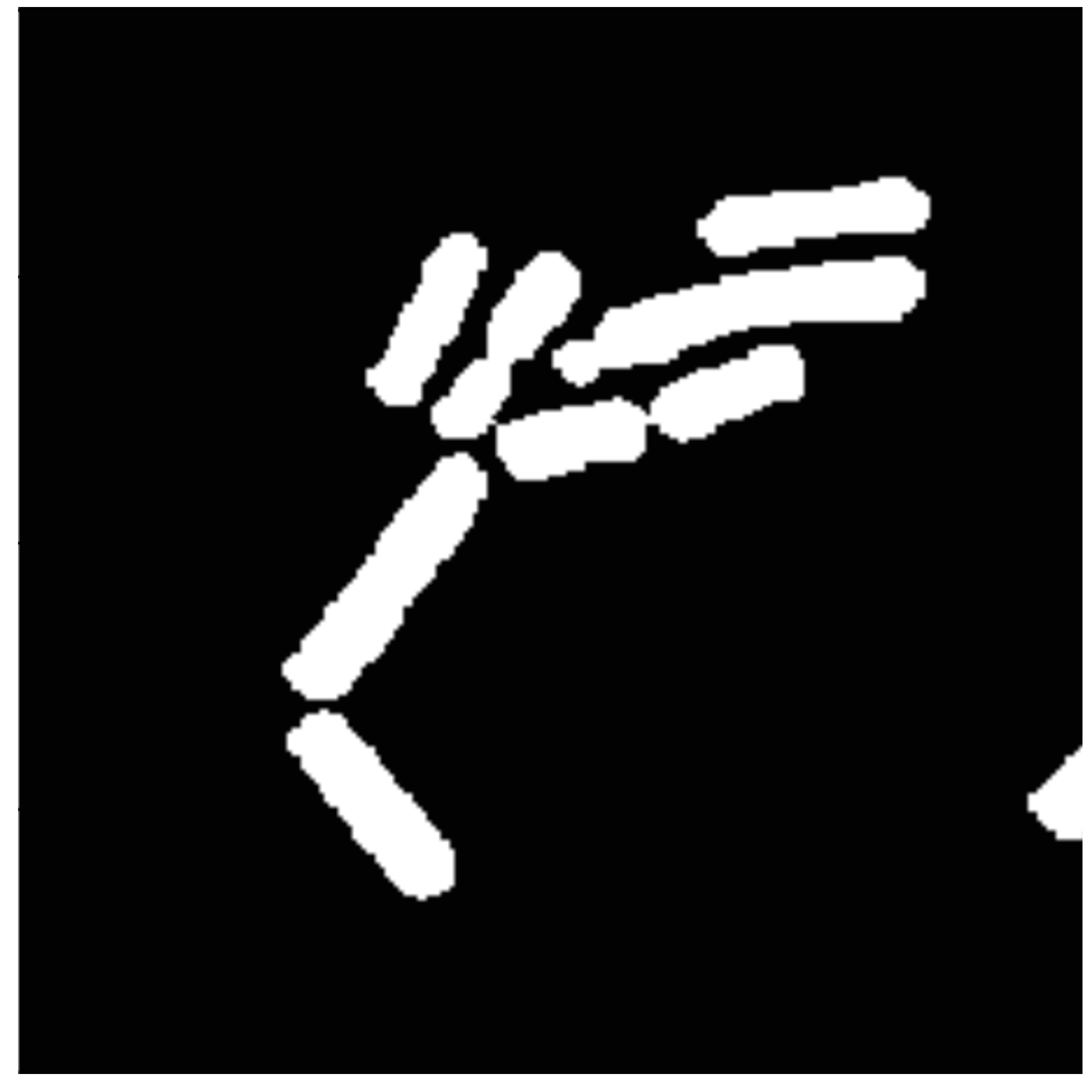
Frequentist approaches can be useful and easily implemented

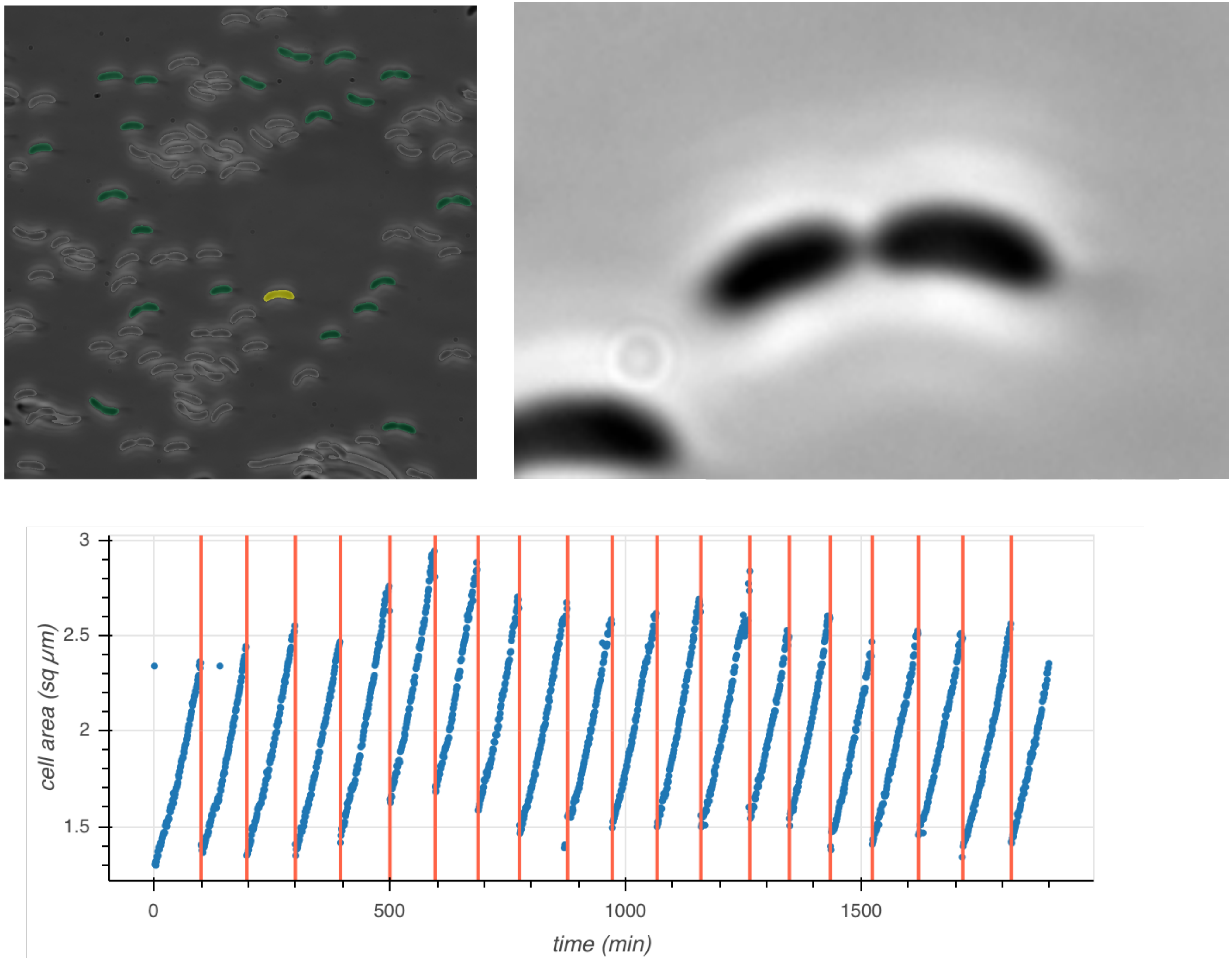



DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

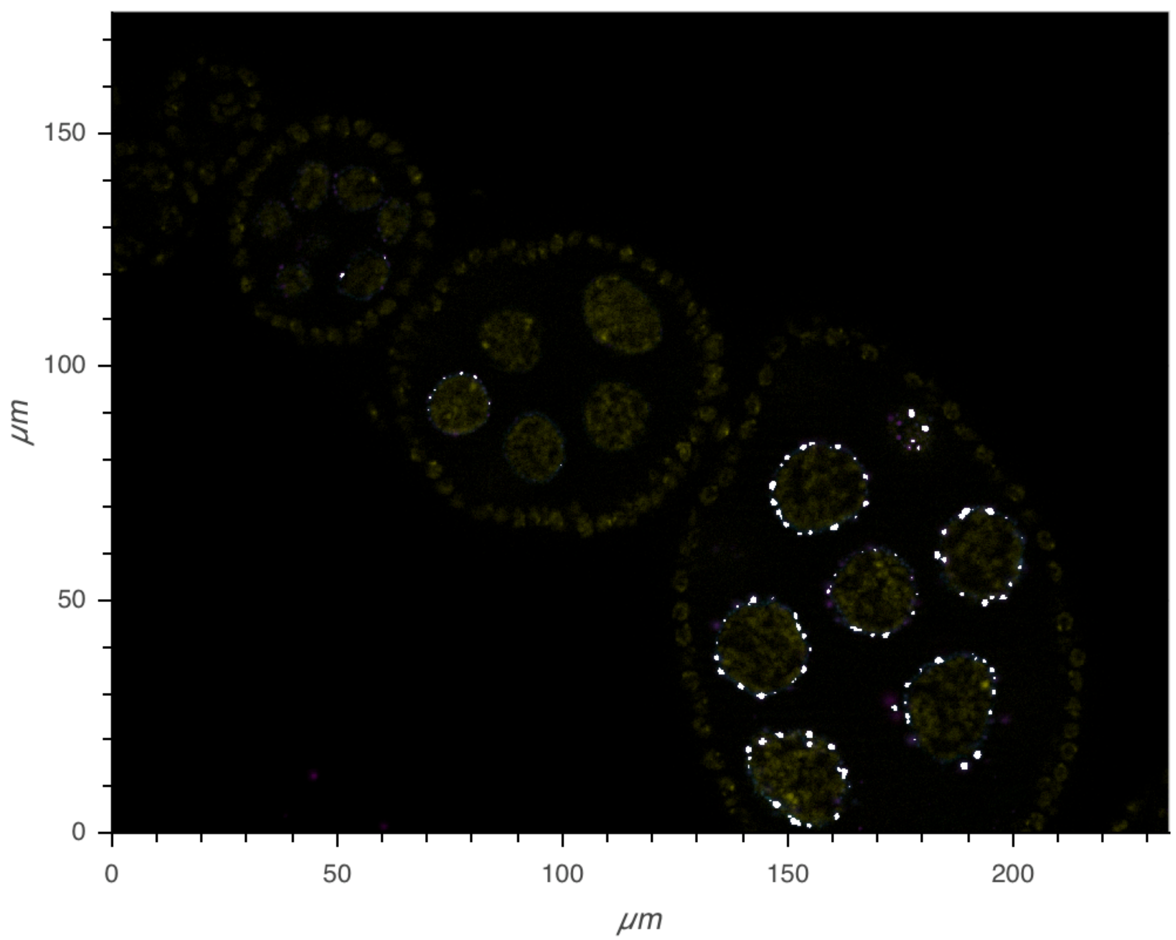
Your computer can see!

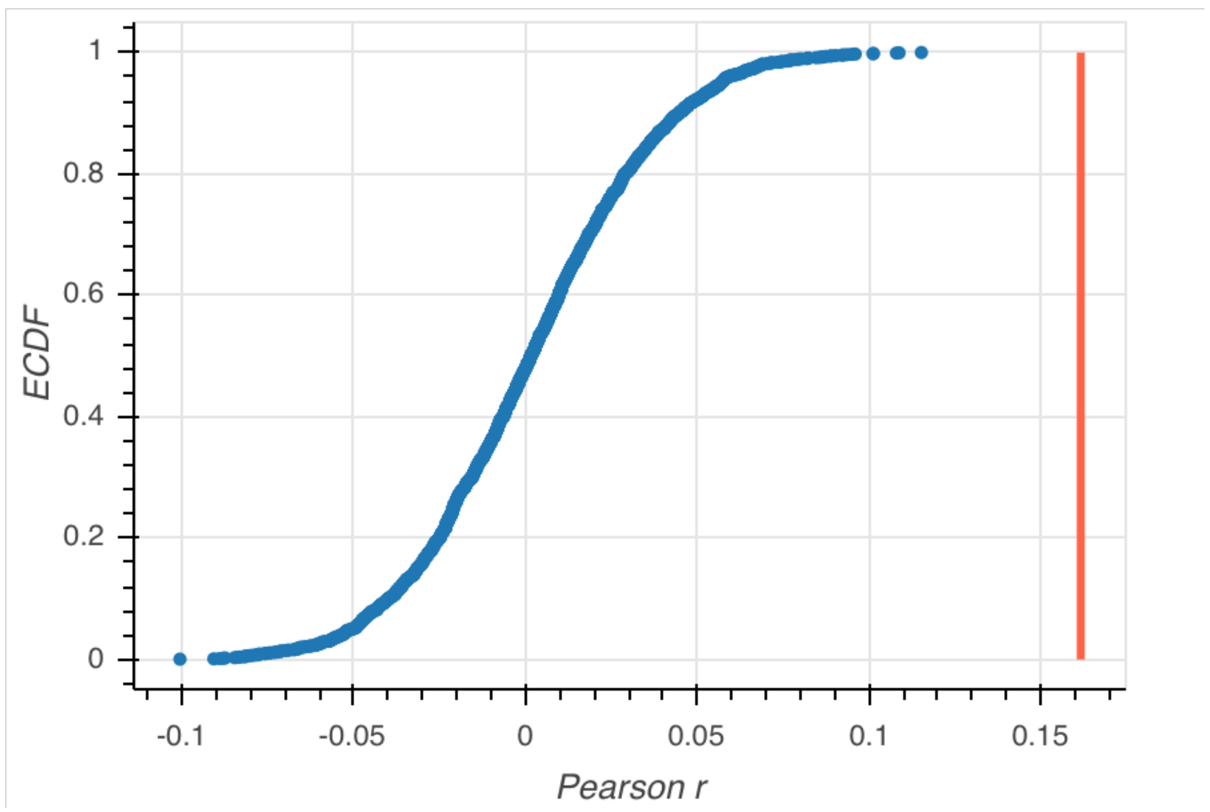




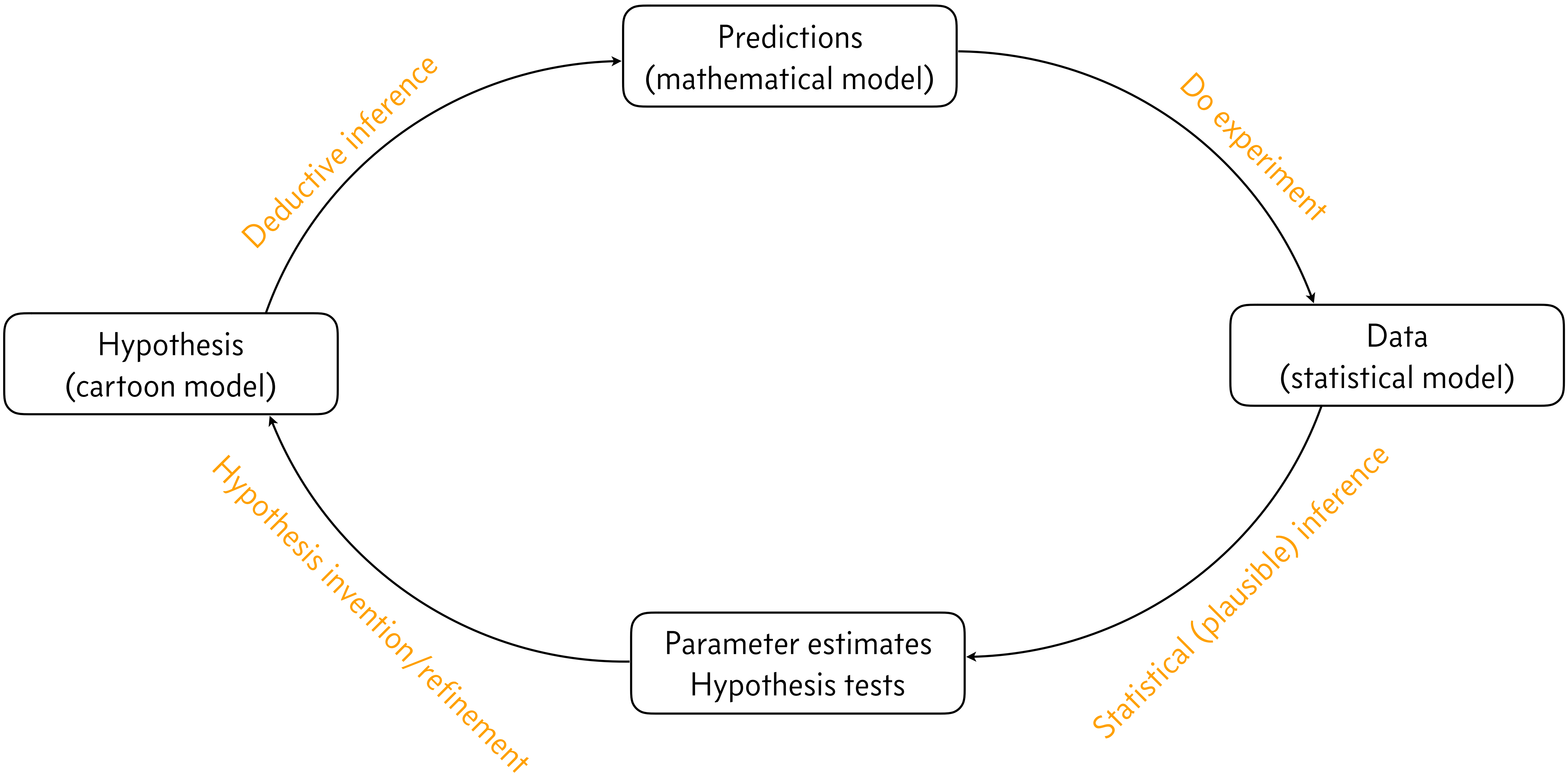

3
2.5
2
1.5
cell area (sq μm)
0
500
1000
1500
time (min)

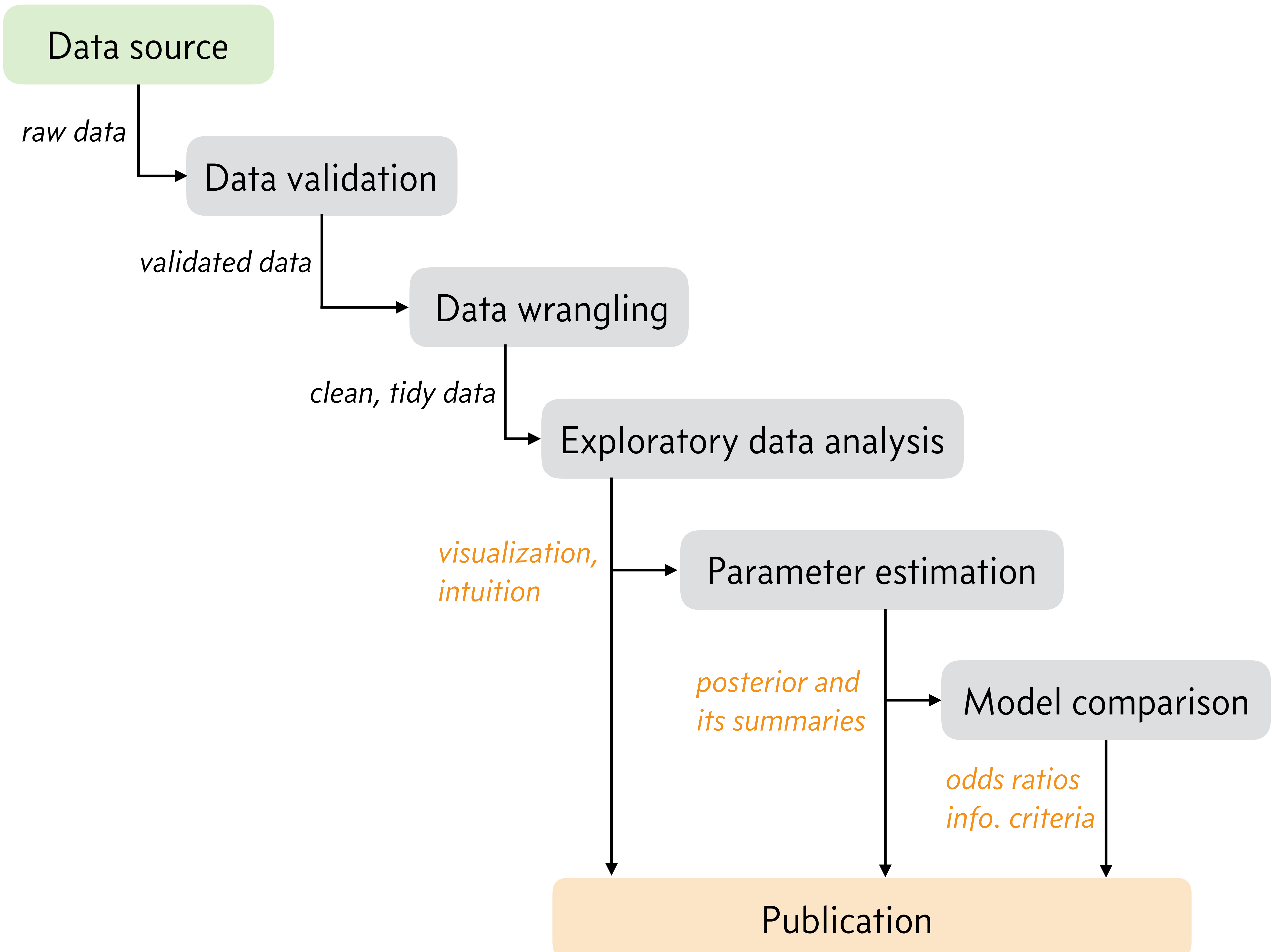
Colocalization can and should be quantified





The scientific method




Data source
raw data
Data validation
validated data
Data wrangling
clean, tidy data
Exploratory data analysis
visualization,
intuition
Parameter estimation
posterior and
its summaries
Model comparison
odds ratios
info. criteria
Publication

Reproducible research requirements

Protocols are complete, organized, and accessible.

Note instruments, firmware versions, all operating parameters

Data sets are complete, organized, and accessible.

Use standardized tools, include intermediate results, store sensibly

All processing is automated with open code.

Use open source tools, use version control, make your code public

Thank you

Thank you to the data sources

Caltech

- Avni Gandhi, Audrey Chen, Grigorios Oikonomou, and David Prober
- Greg Reeves, Nathanie Trisnadi, and Angela Stathopoulos
- Ravi Nath, Claire Bedbrook, Mike Abrams, and Lea Goentoro
- Jin Park and Michael Elowitz
- Zak Singer and Michael Elowitz
- Alex Webster and Alexei Aravin
- Dawna Bagherian, Kyu Lee, and Markus Meister
- Griffin Chure, Manuel Razo, and Rob Phillips
- Meaghan Sullivan, Kevin Yu, Jimmy Hamilton, and the students of Bi 1x
- Han Wang and Paul Sternberg
- Emily Blythe and Ray Deshaies
- Lior Pachter and contributors to SNPedia

Extramural

- Melissa Gardner (U Minnesota), Marija Zanic (Vanderbilt), and Joe Howard (Yale)
- Matt Good and Dan Fletcher (UC-Berkeley)
- Nate Goehring (Crick) and Stephan Grill (BIOTEC-Dresden)
- Charlie Wright, Srividya Iyer-Biswas, and Norbert Scherer (U Chicago)
- Peter and Rosemary Grant (Princeton)
- Thomas Kelinteich and Stanislav Gorb (Kiel)
- Lars Straub and Geoffrey Williams (U Bern)
- Alan Perelson (Santa Fe Institute)
- Yanping Chen and the UCR Time Series Classification Archive

Thank you

303 contributors to Jupyter notebook

268 contributors to Bokeh

29 contributors to HoloViews

237 contributors to scikit-image

973 contributors to scikit-learn

938 contributors to Pandas

582 contributors to Numpy

141 contributors to PyMC3

325 contributors to Theano

41 contributors to emcee/ptemcee

Contributors to the rest of the SciPy stack

python SOFTWARE FOUNDATION

Thank you

Junedh Amrute

Heidi Klumpe

James McGehee

Porfirio Quintero-Cadena

Christina Su

Thank you

All of you!

Go forth and…

Use what you have learned to do reproducible quantitative research.

Evangelize workflows for reproducible science.