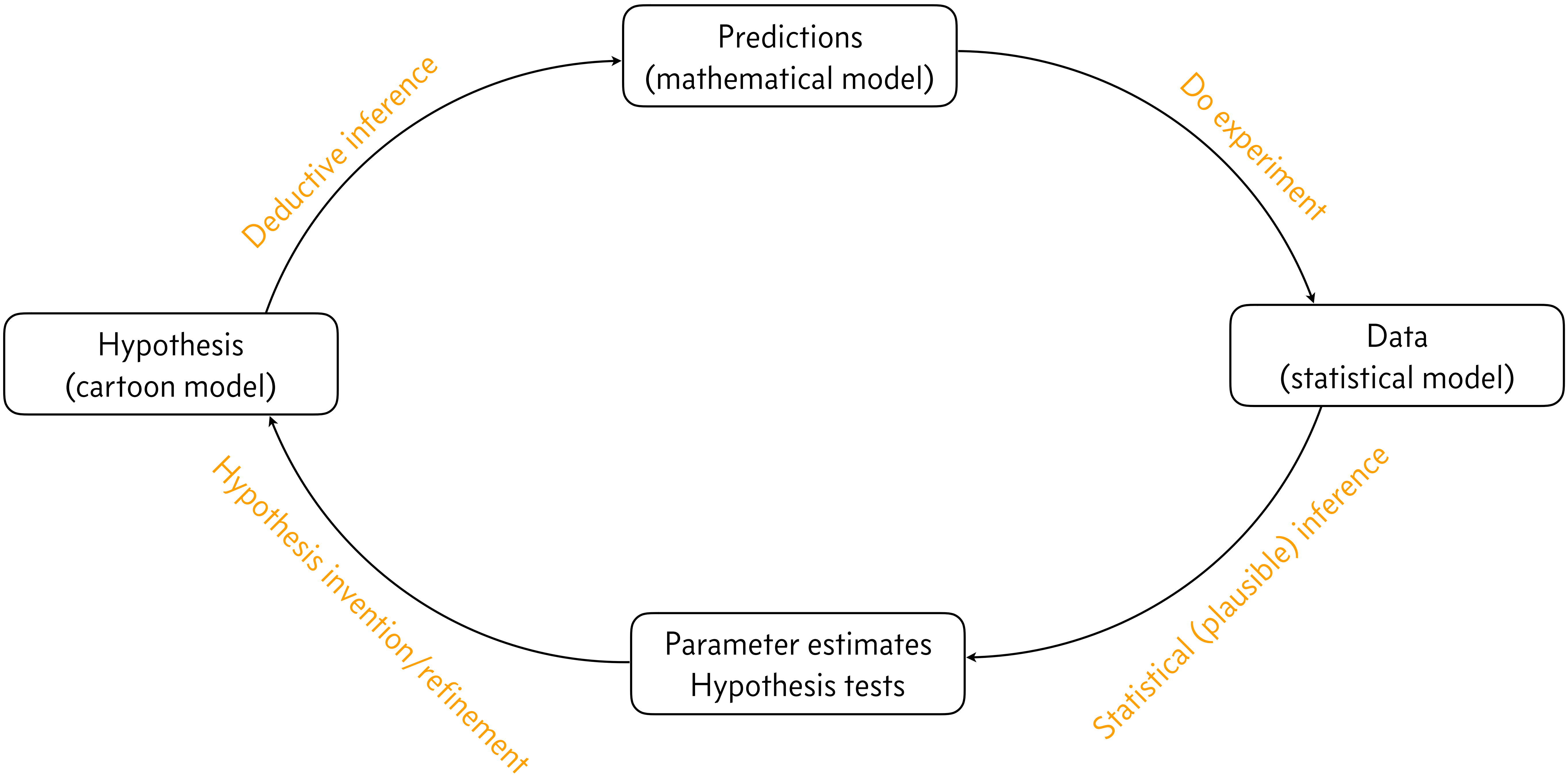
BE/Bi 103

Data Analysis in the Biological Sciences

Fall term, 2018

The scientific method



"Exploratory data analysis can never be the whole story,
but nothing else can serve as a foundation stone—as the first step."

John Tukey

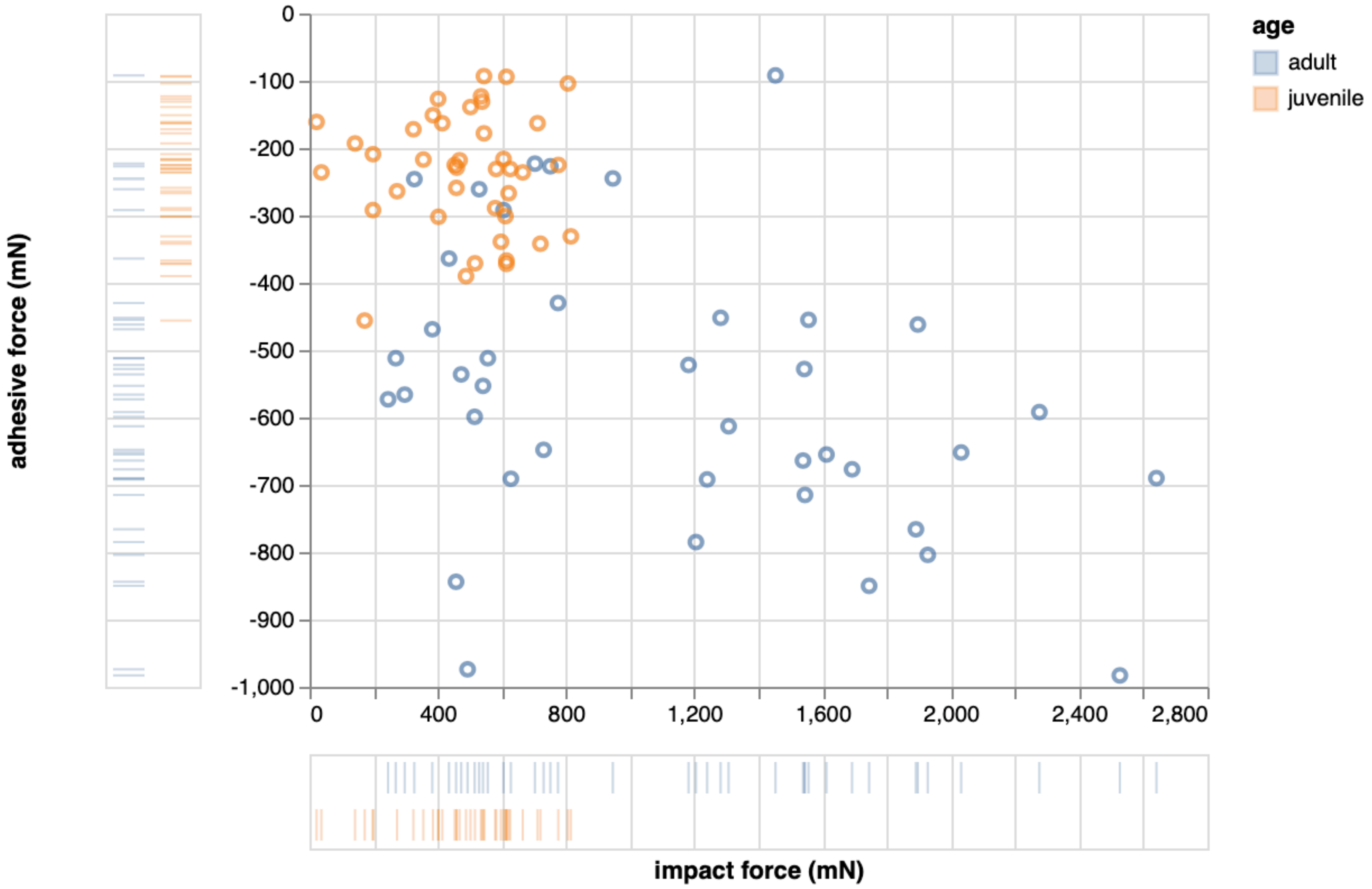
Tidy data enables rapid, logical data access

	location	activity	time	zeit	zeit_ind	day	genotype	light
0	1	0.6	2013-03-15 18:31:09	-14.480833	-869	4	het	True
1	1	1.9	2013-03-15 18:32:09	-14.464167	-868	4	het	True
2	1	1.9	2013-03-15 18:33:09	-14.447500	-867	4	het	True
3	1	13.4	2013-03-15 18:34:09	-14.430833	-866	4	het	True
4	1	15.4	2013-03-15 18:35:09	-14.414167	-865	4	het	True

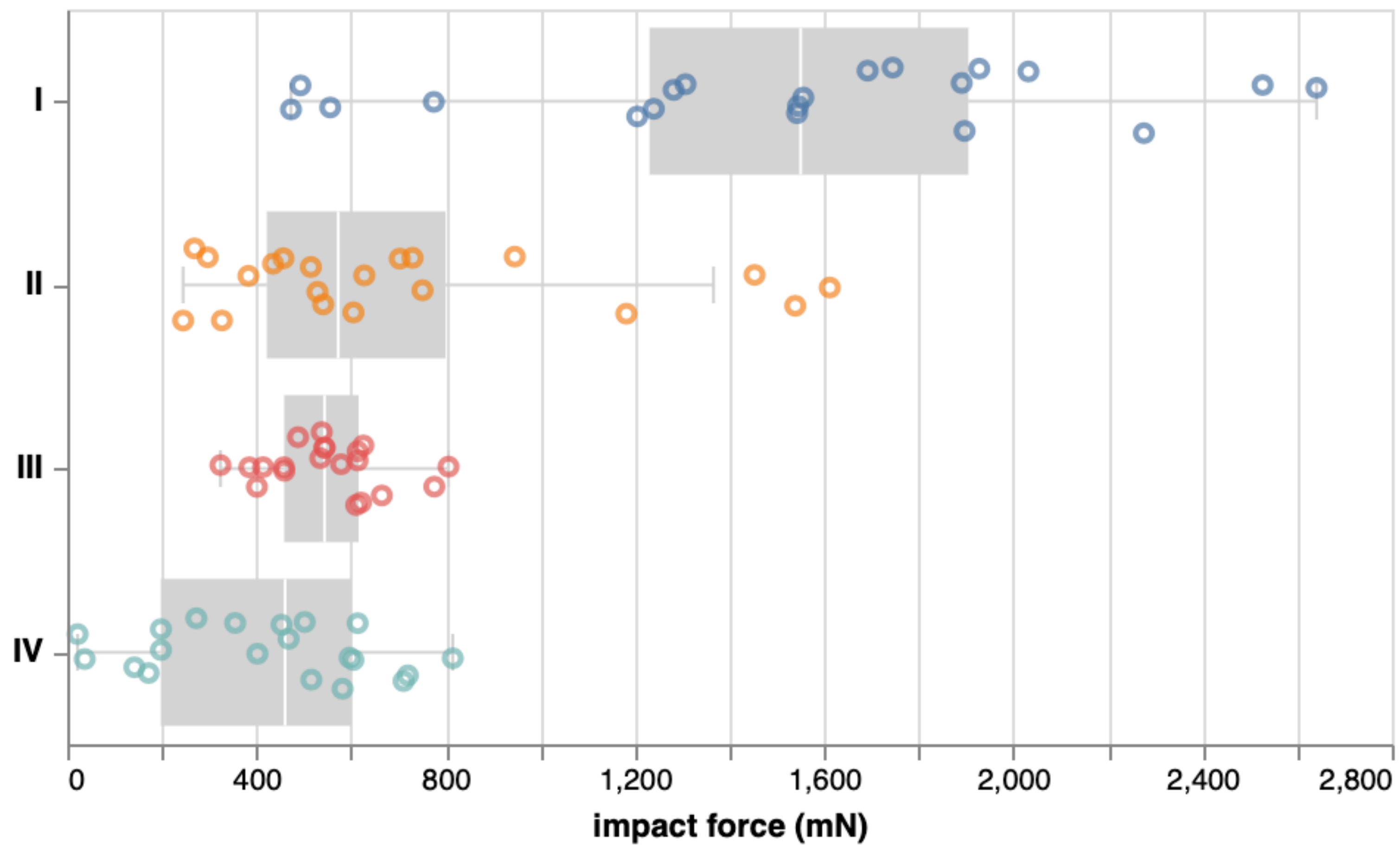
Split, apply, combine

Your new wash, rinse, repeat?

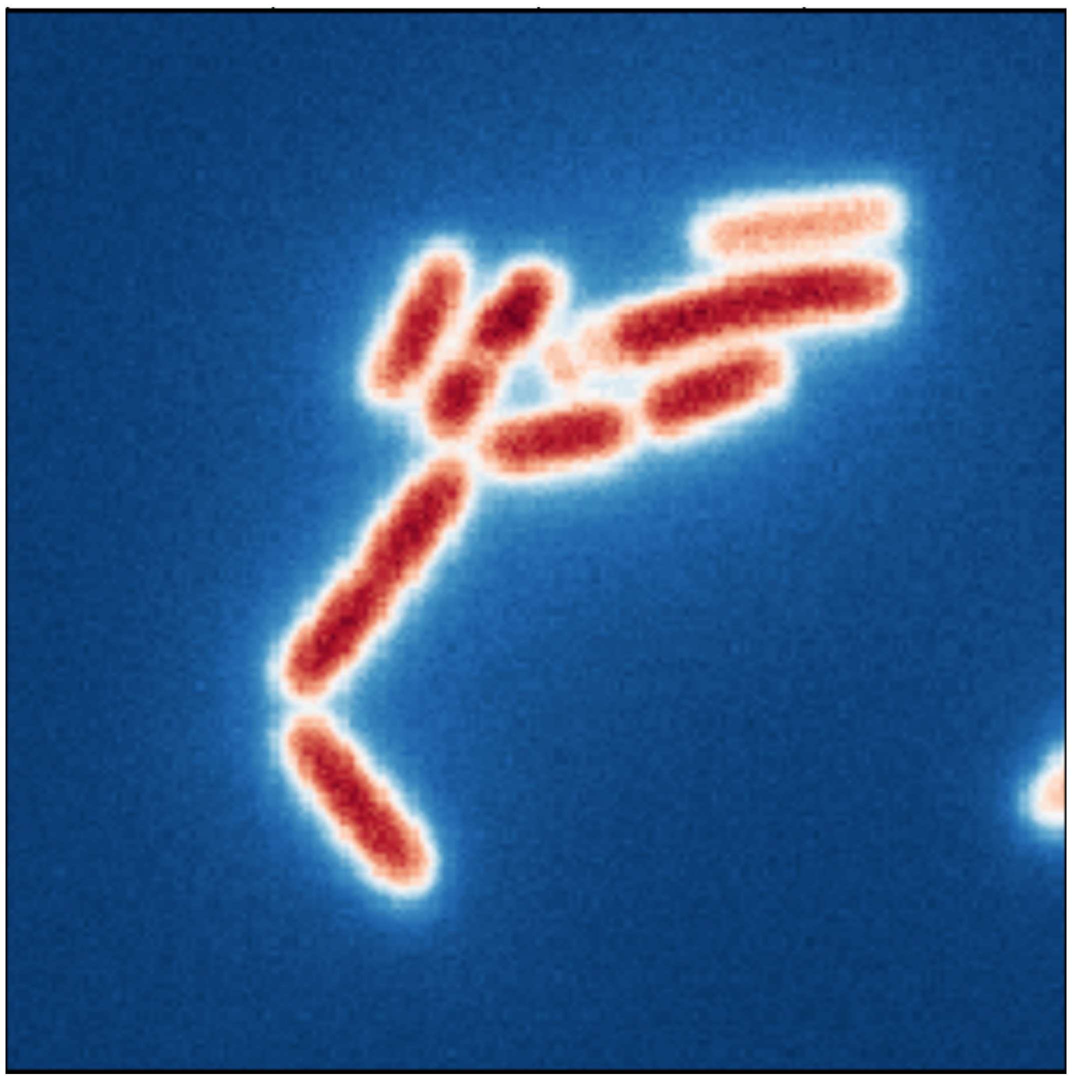
High-level plotting libraries enable rapid building of informative graphics

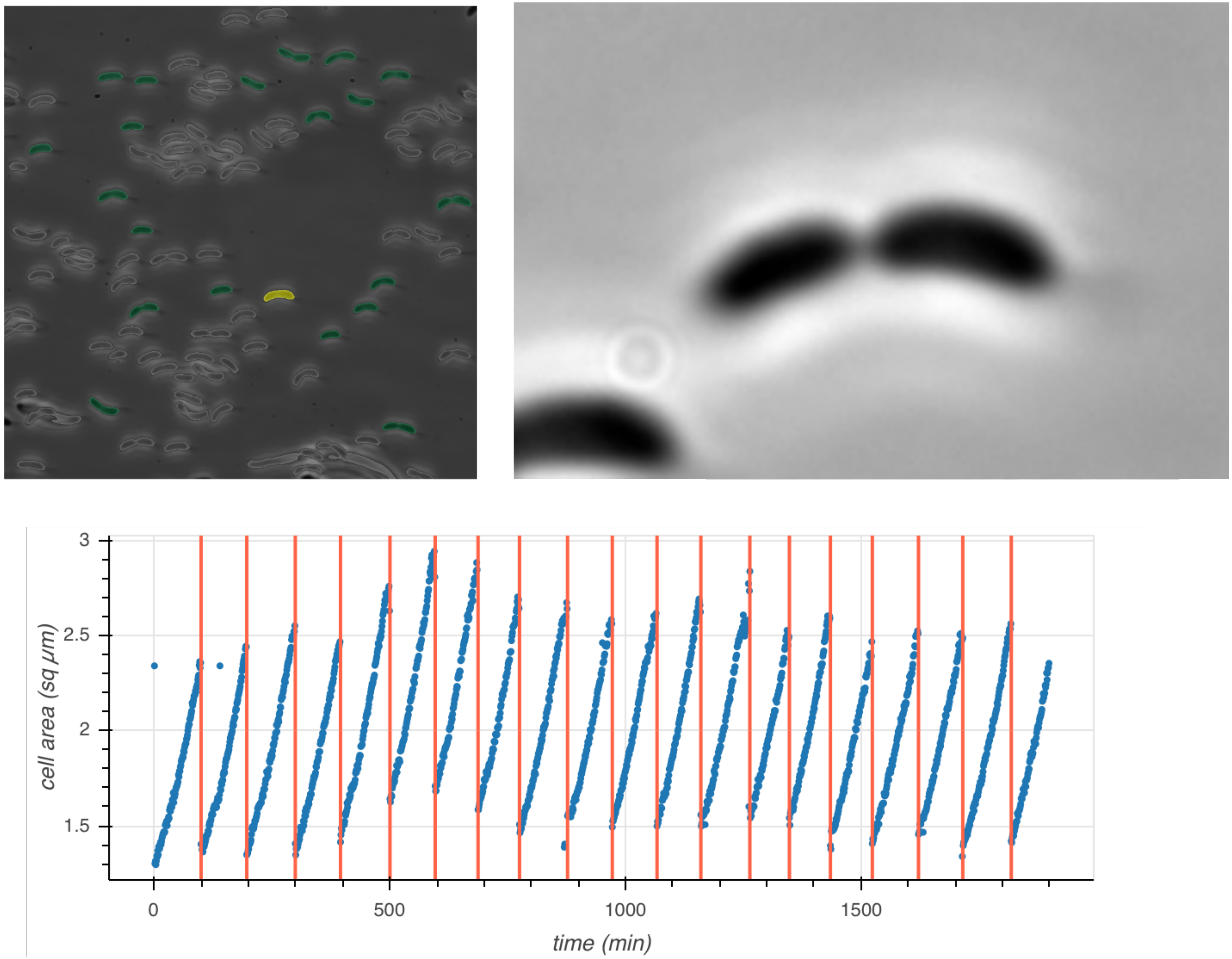


High-level plotting libraries enable rapid building of informative graphics



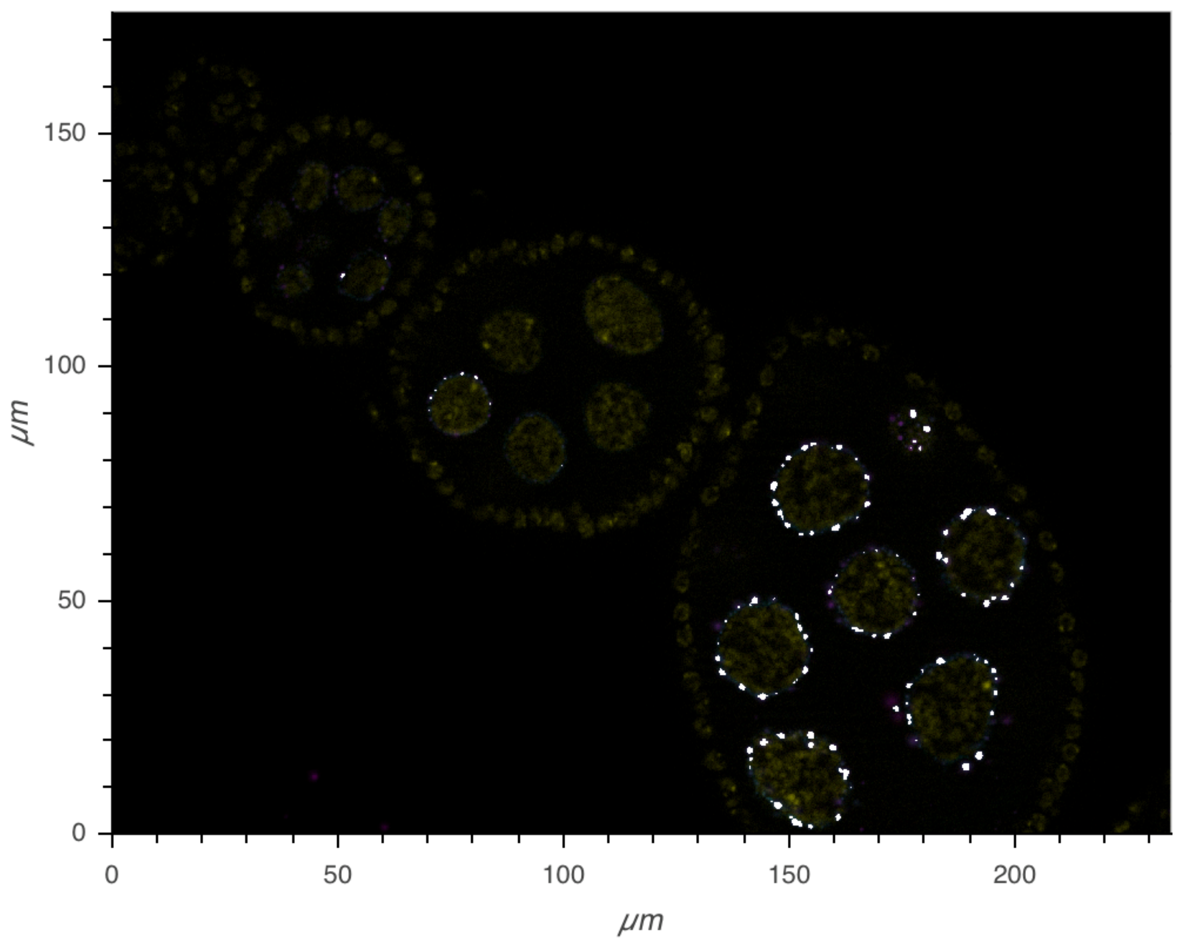
Your computer can see!

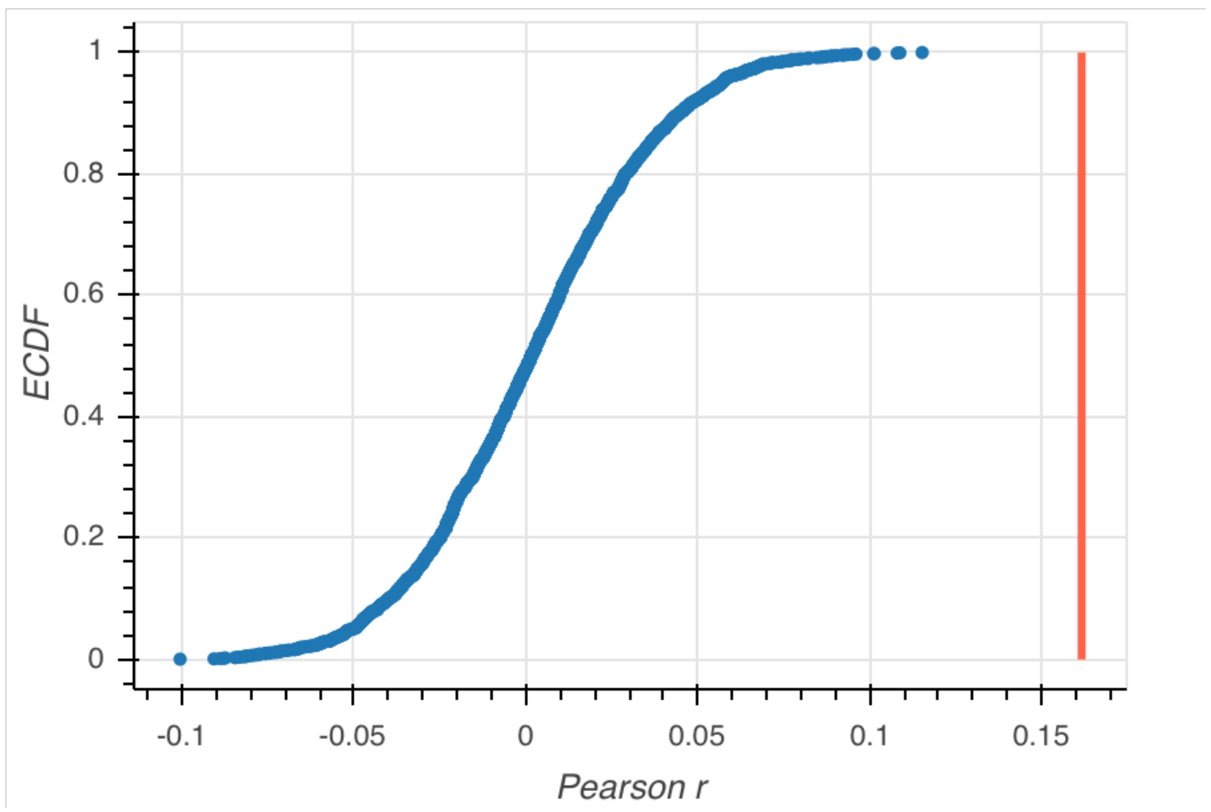




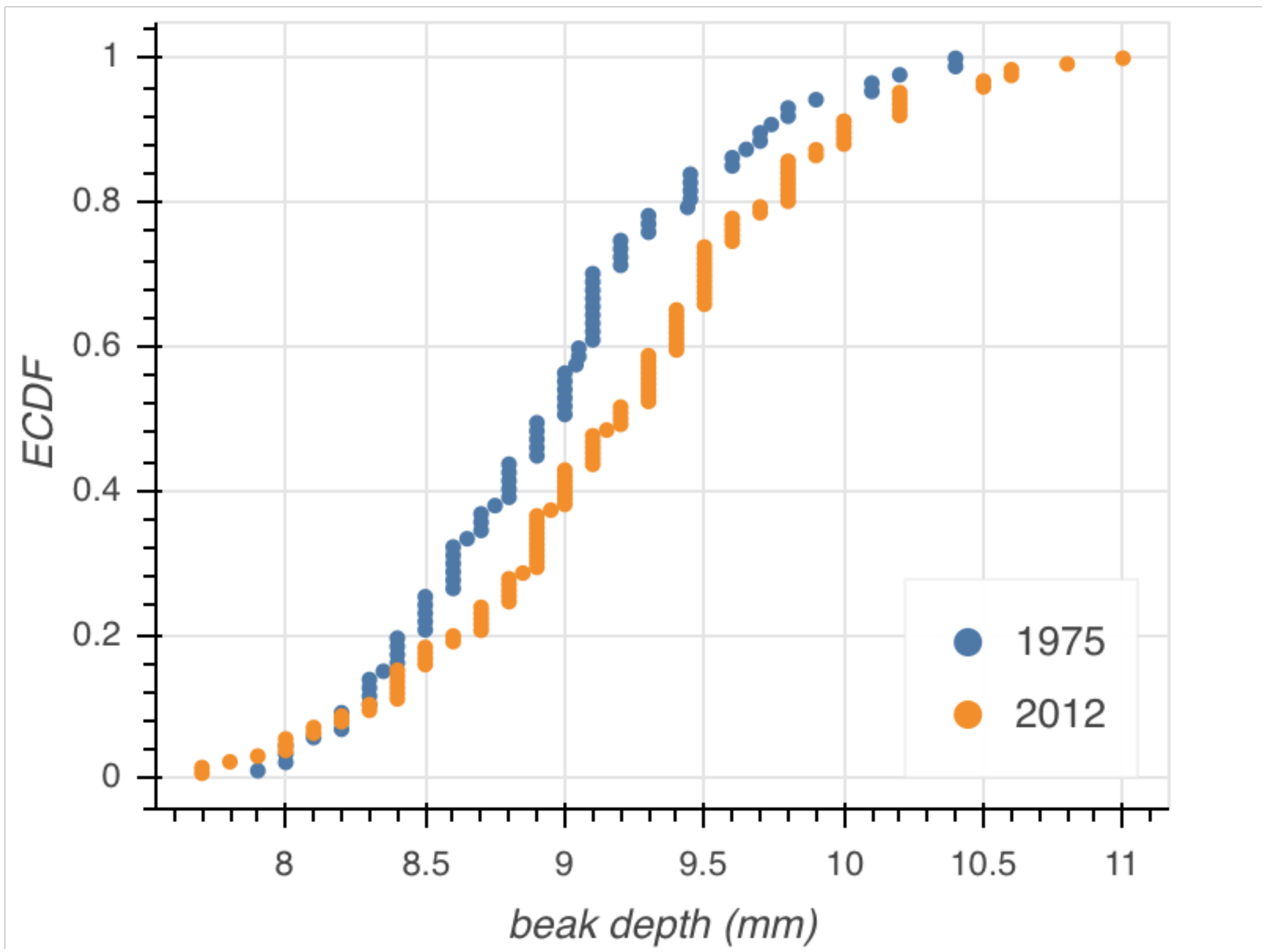
cell area (sq µm)
time (min)
0
500
1000
1500
1.5
2
2.5
3

Colocalization can and should be quantified





ECDFs allow visualization of the entire *distribution* of the data



Statistical inference requires a probability theory

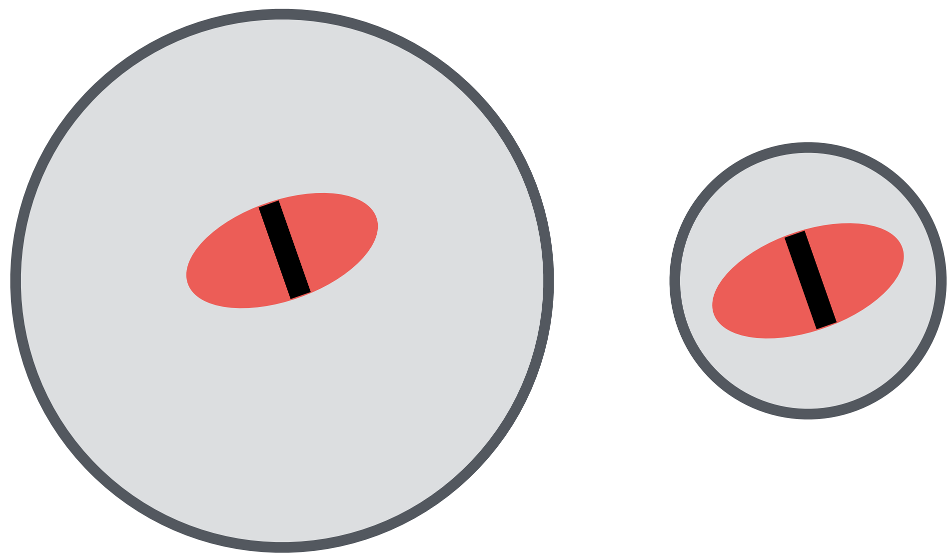
Generative joint distribution $\pi(y, \theta) = f(y \mid \theta)\ \ g(\theta)$

Bayes's theorem for parameter estimation

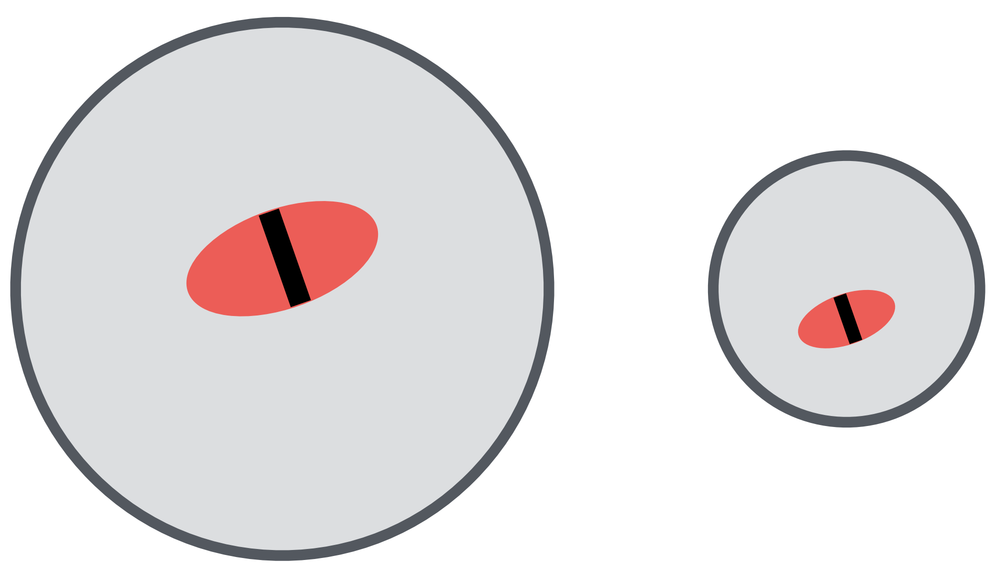
$$\text{posterior} = g(\theta \mid y) = \frac{f(y \mid \theta)\ \ g(\theta)}{f(y)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Cartoon models shape our thinking

Model a:



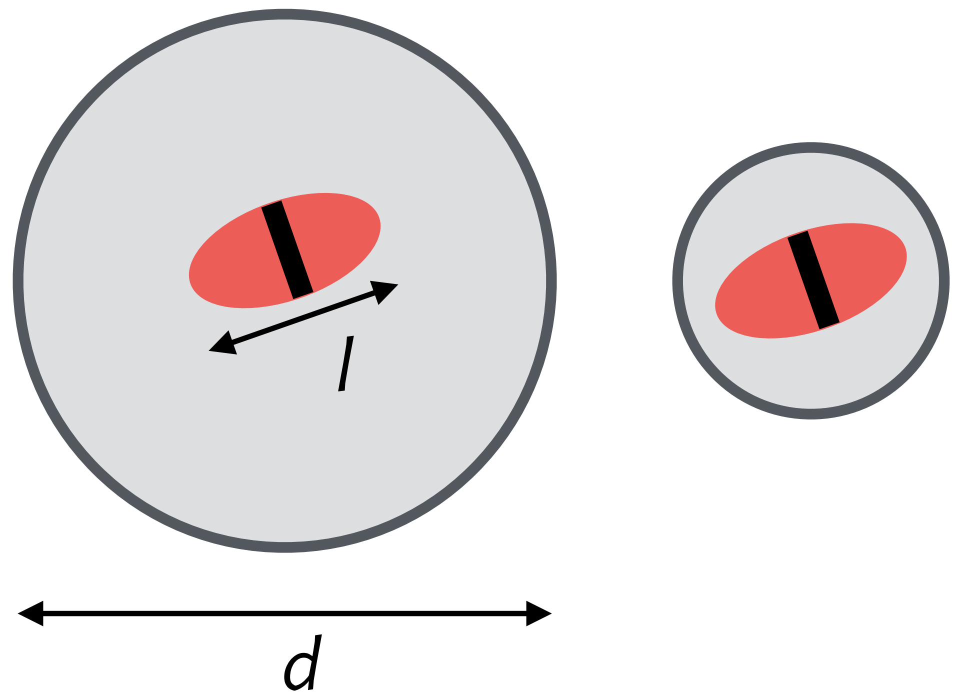
Model b:



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

Mathematical models identify parameters

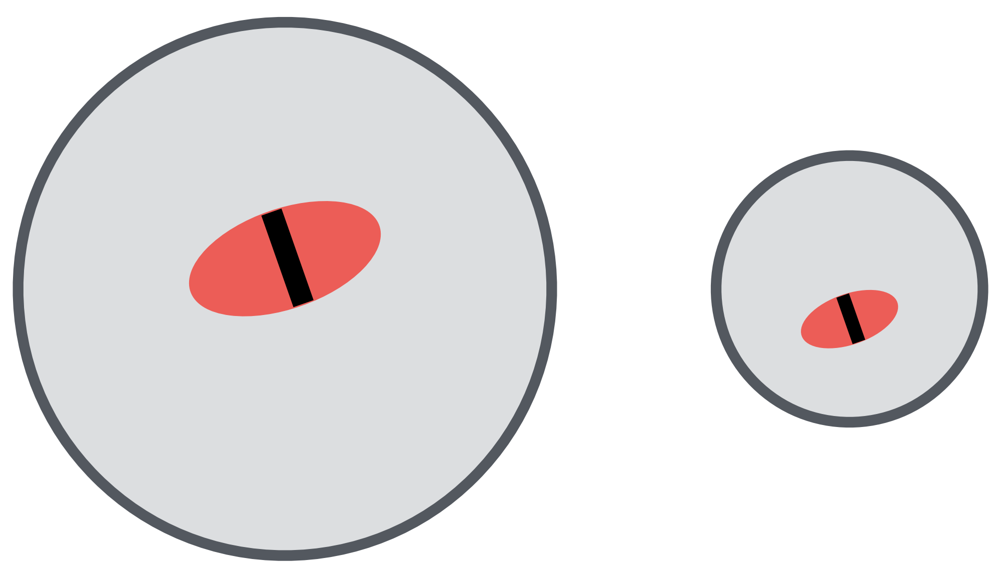
Model a:



$$l \neq l(d)$$

$$l = l_s$$

Model b:

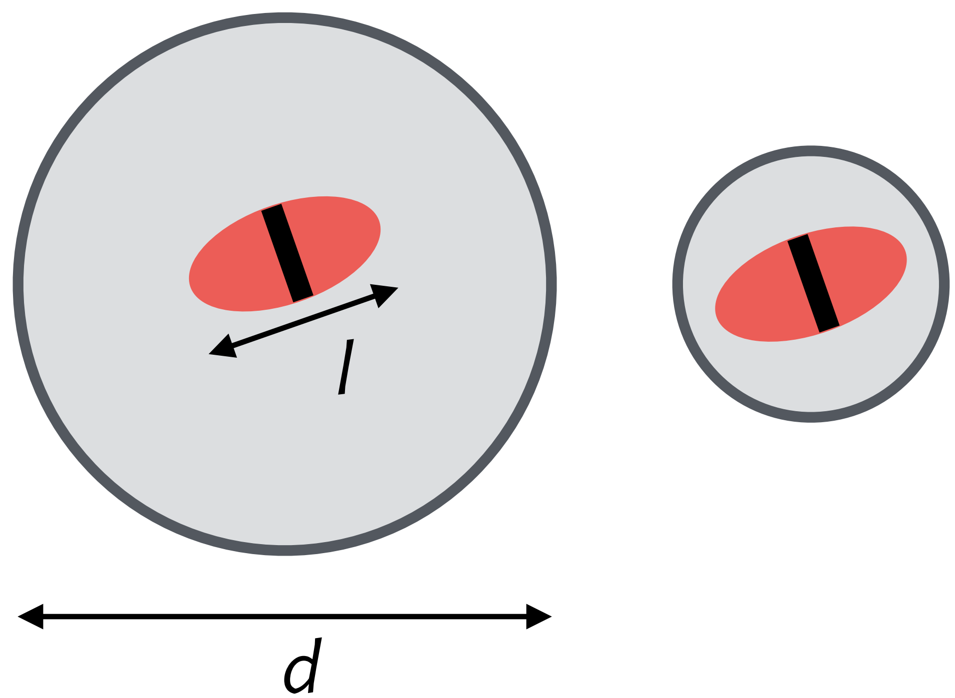


$$l(d;\gamma,\phi) = \frac{\gamma d}{(1+(d/\phi)^3)^{1/3}}$$

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

Statistical models are *generative*

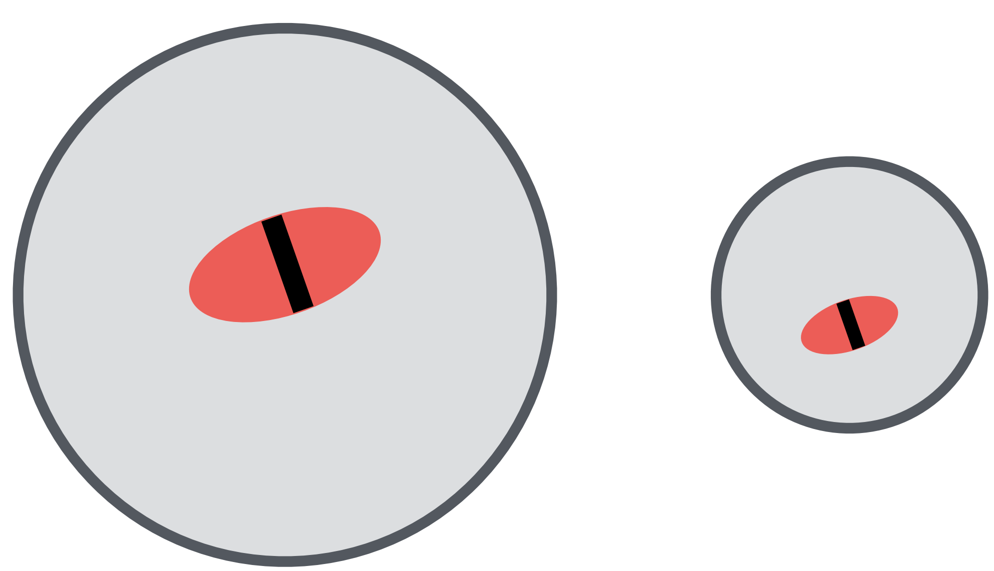
Model a:



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

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

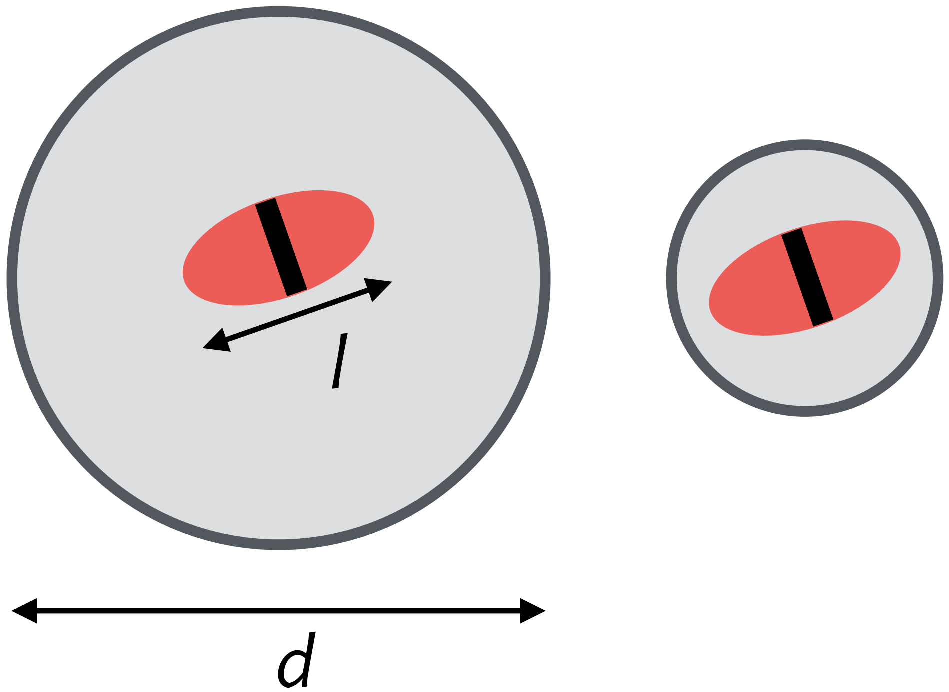
Model b:



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

Statistical models need priors

Model a:



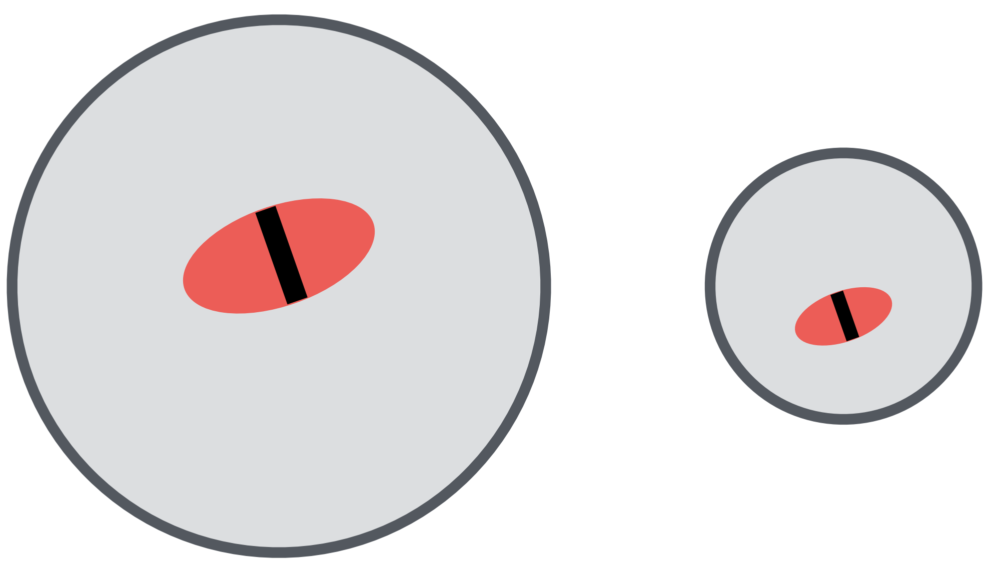
$\phi \sim \mathrm{LogNorm}(3, 0.75)$

$\sigma_0 \sim \mathrm{Gamma}(2, 10)$

$\sigma = \sigma_0 \, \phi$

$l_i \sim \mathrm{Norm}(\phi, \sigma) \; \forall \, i$

Model b:



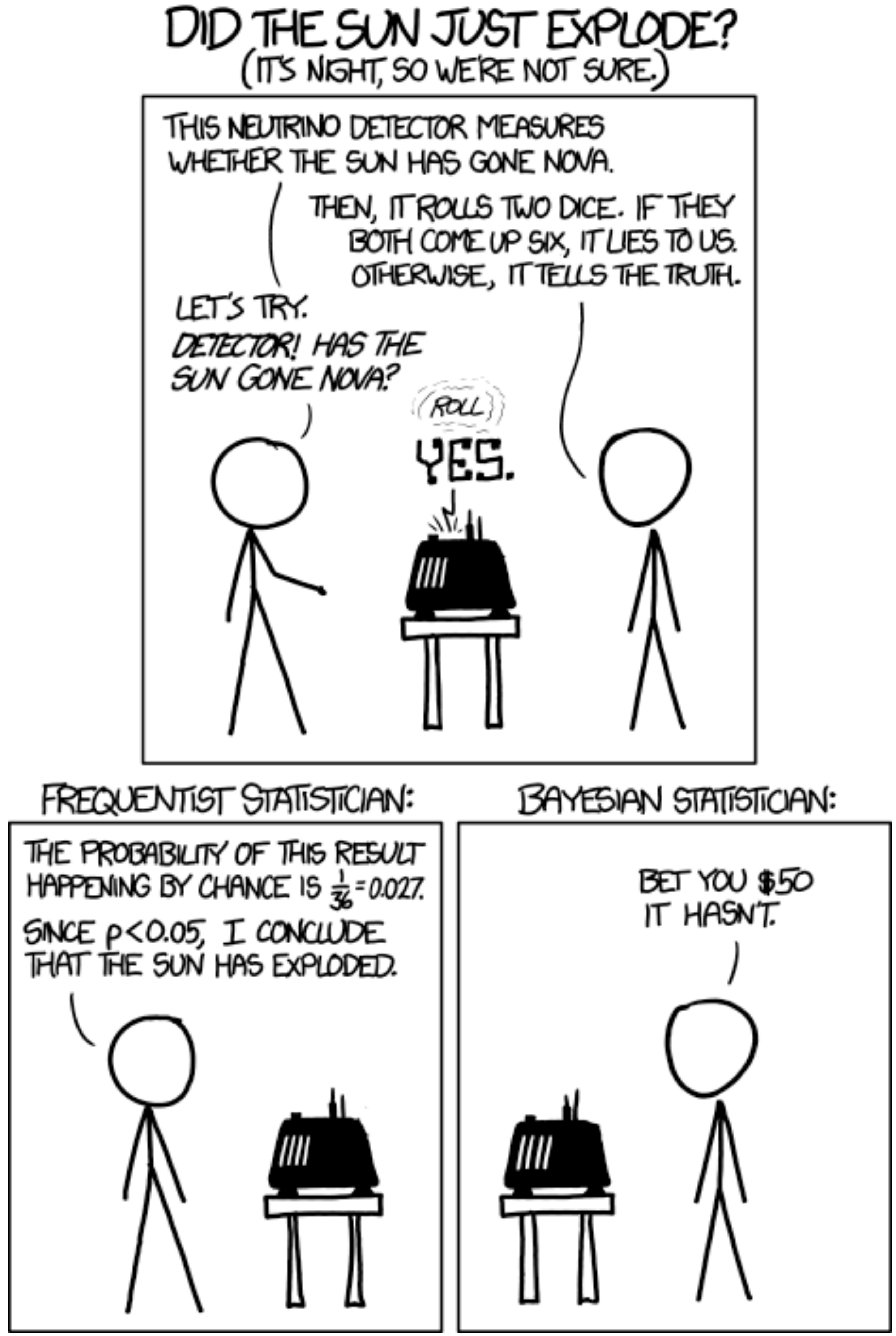
$\phi \sim \mathrm{LogNorm}(3, 0.75)$

$\gamma \sim \mathrm{Beta}(2,2)$

$\sigma_0 \sim \mathrm{Gamma}(2, 10)$

$\sigma = \sigma_0 \, \phi$

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


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 1/36 = 0.027.
SINCE p < 0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

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

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

The posterior may sometimes be approximated as Gaussian

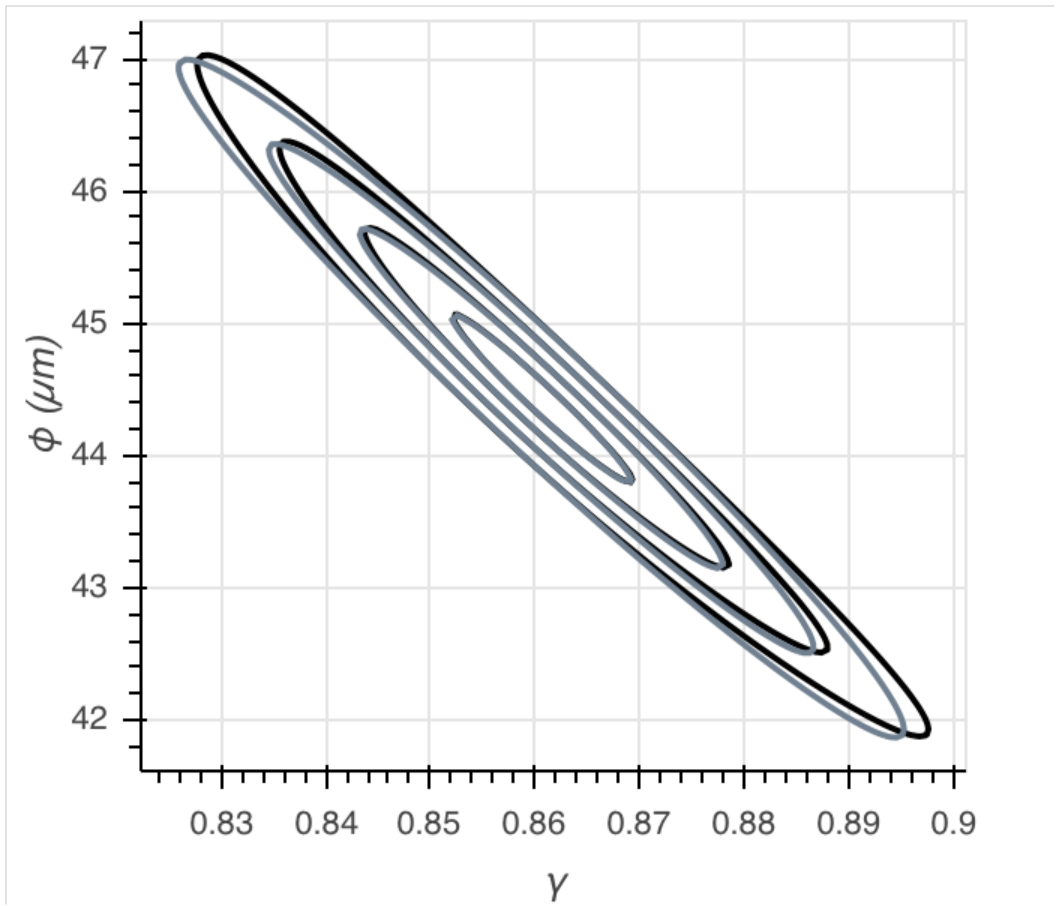
1. Find the most probable parameters θ^* (the MAP).

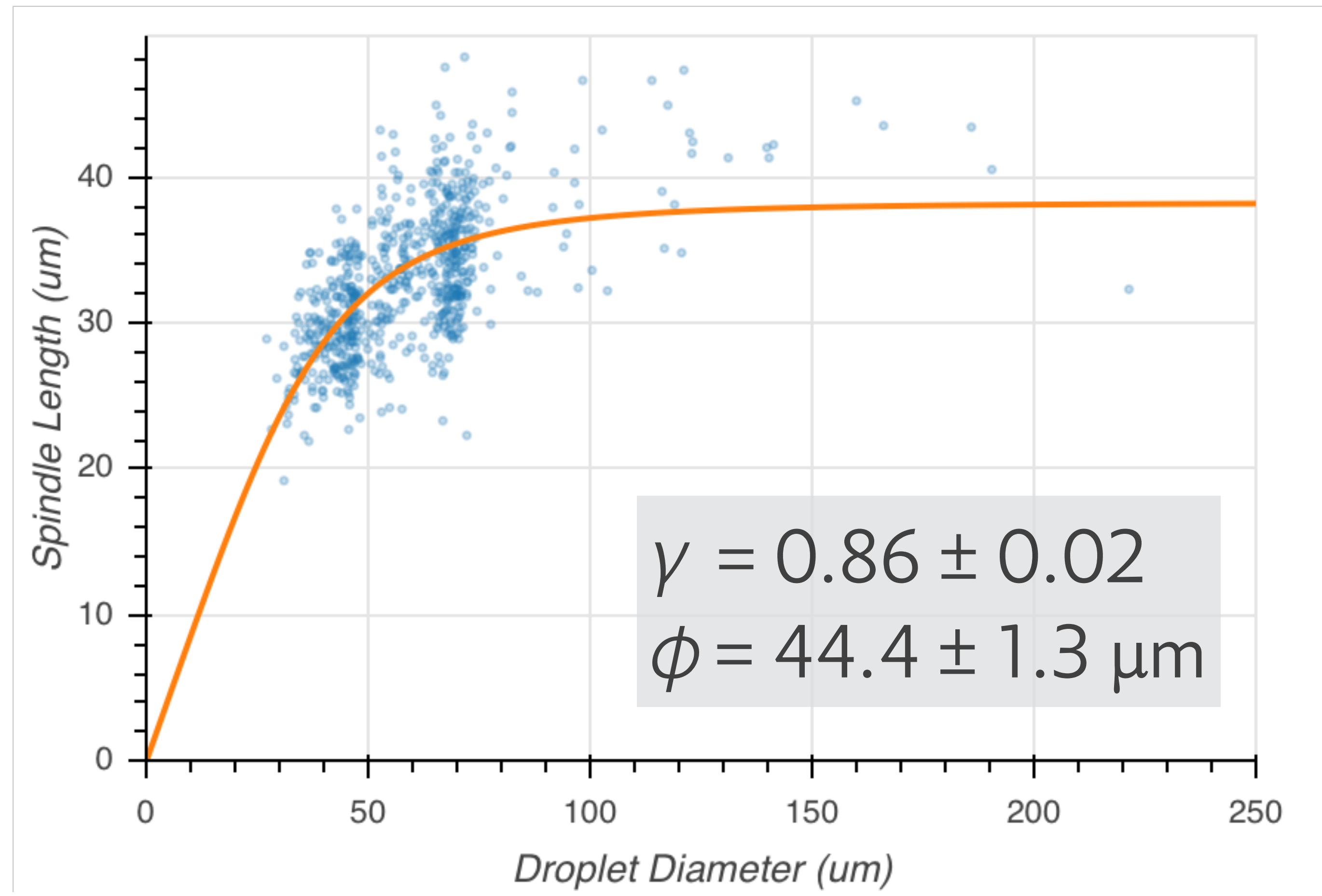
2. Approximate the posterior $g(\theta^*|y)$ as Gaussian by doing a Taylor expansion of $\ln g(\theta^*|y)$ about θ^*.

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

Obvious assumption: posterior is approximately Gaussian.

The posterior may sometimes be approximated as Gaussian





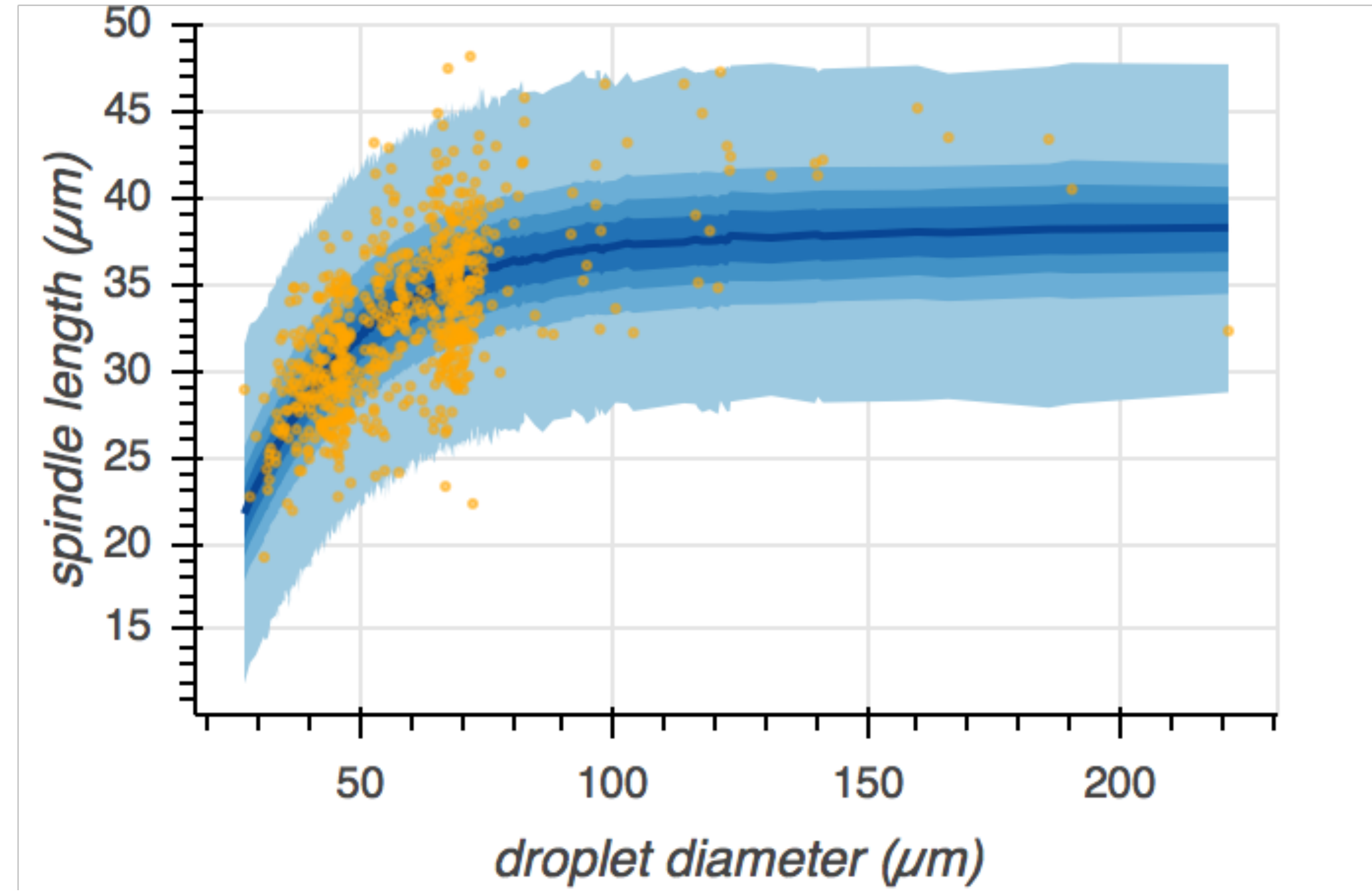
The posterior may be *sampled* using MCMC

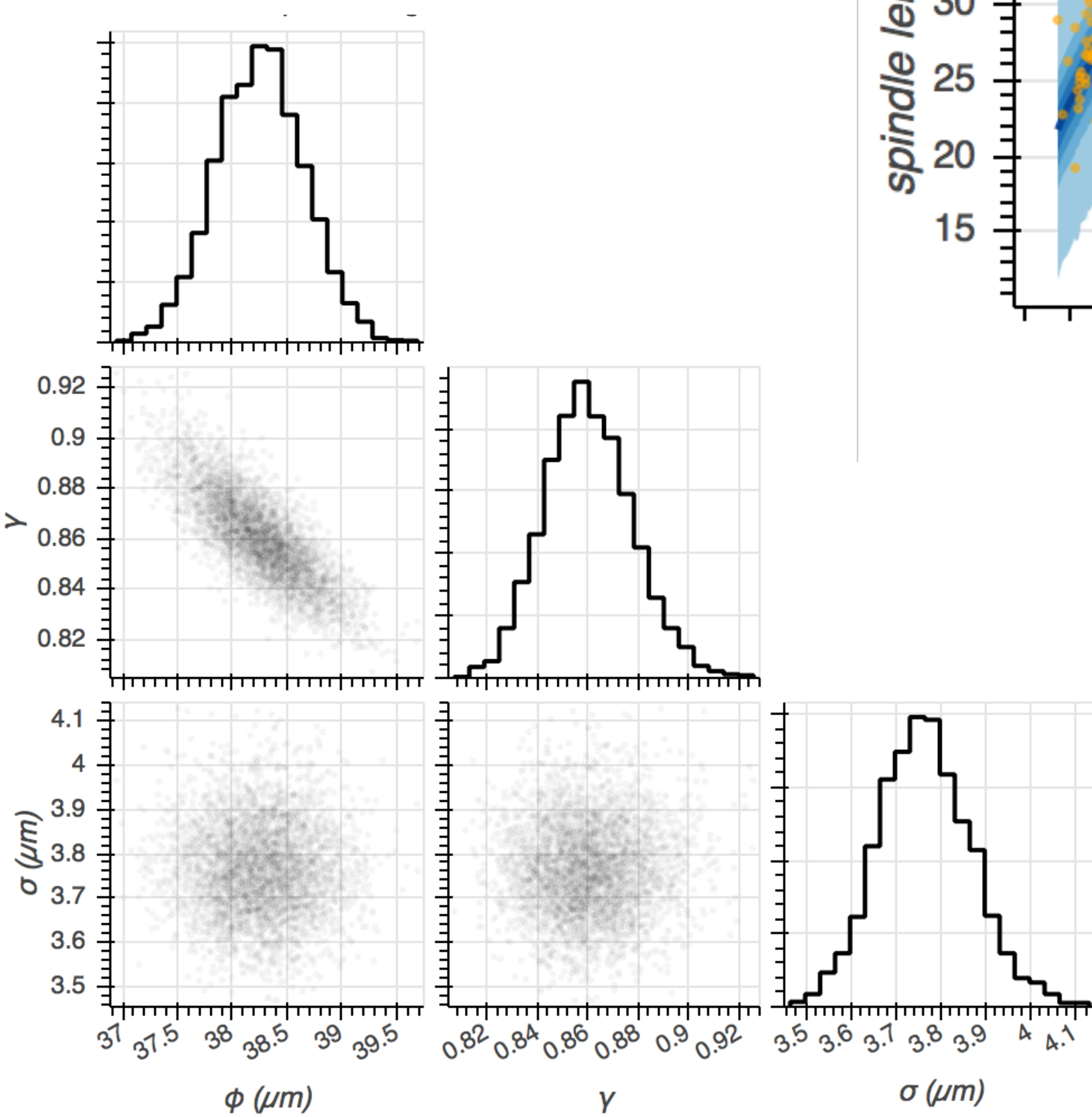
1. Define the (log) posterior distribution

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

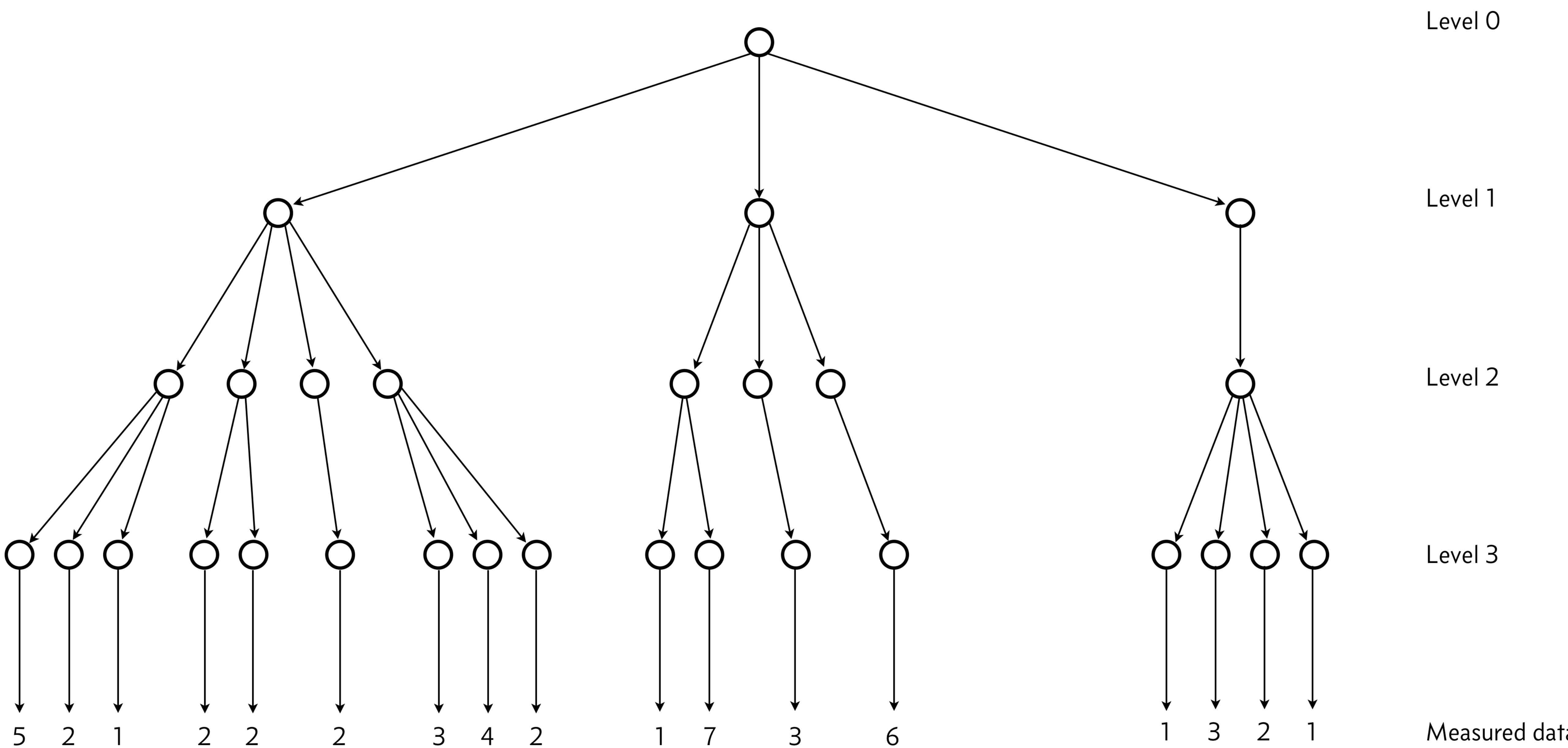
3. Obtain marginalized posterior by considering specific parameters.

®

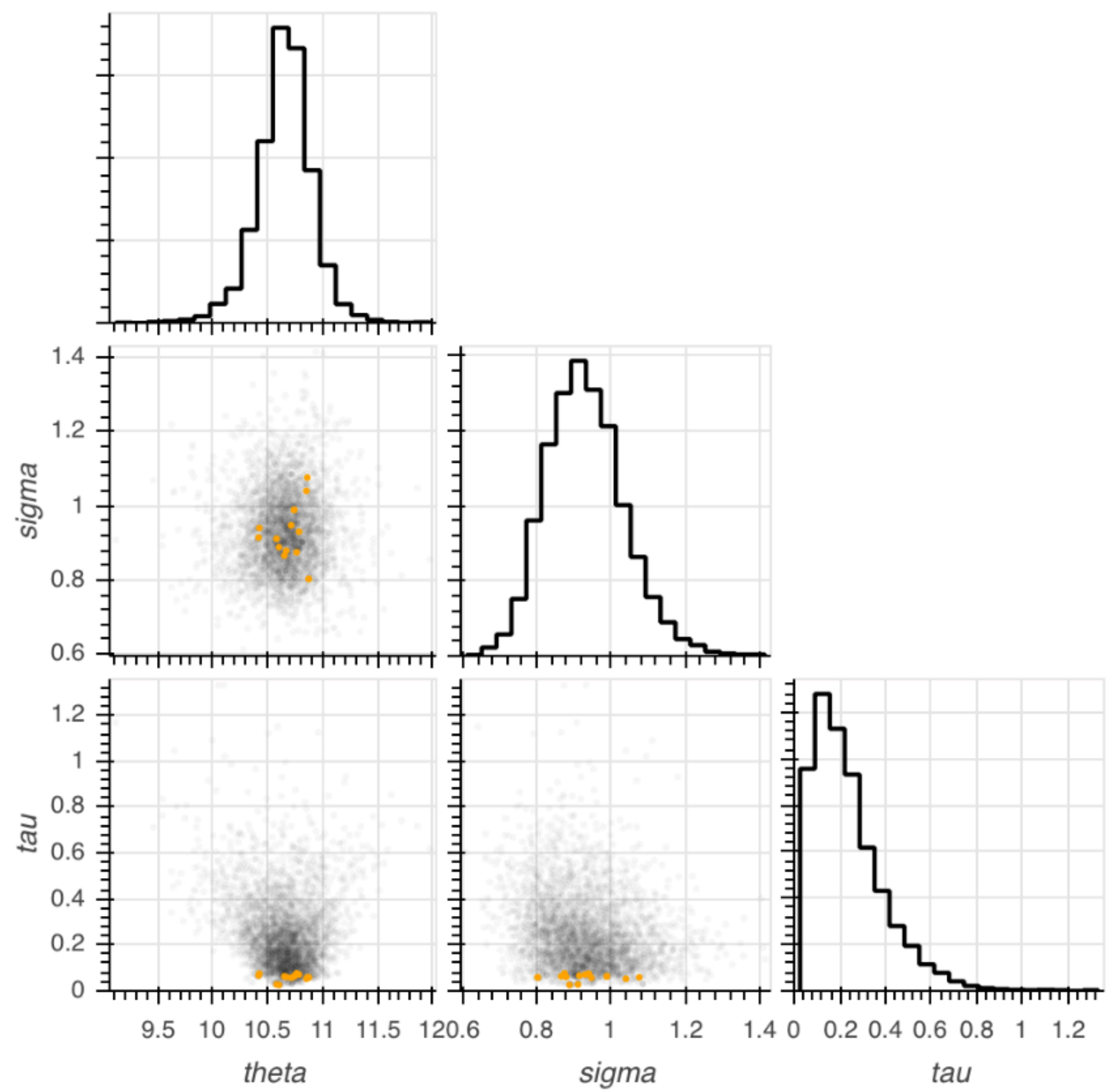

spindle length (µm)
droplet diameter (µm)
50
45
40
35
30
25
20
15
50
100
150
200


γ
σ (µm)
ϕ (µm)
γ
σ (µm)
0.92
0.9
0.88
0.86
0.84
0.82
4.1
4
3.9
3.8
3.7
3.6
3.5
37
37.5
38
38.5
39
39.5
0.82
0.84
0.86
0.88
0.9
0.92
3.5
3.6
3.7
3.8
3.9
4
4.1

Hierarchical models are an important class of generative models



Sampling of hierarchical models underscores the need for diagnostics



Principled model building

Use domain knowledge to build a simple generative model

Perform prior predictive checks

Perform simulation-based calibration

Check diagnostics, shrinkage, z-score, and self-consistency

Perform MCMC sampling of the posterior

Check diagnostics and make plots

Perform posterior predictive checks

And model comparison, if need be

Add complexity if necessary

Simple model should be a limit or special case

The generative model can *predict* new data

Generative joint distribution $\pi(\tilde{y}, \theta \mid y) = f(\tilde{y} \mid \theta)\; g(\theta \mid y)$

The generative model can *predict* new data

Generative joint distribution $\pi(\tilde{y}, \theta \mid y) = f(\tilde{y} \mid \theta)\; g(\theta \mid y)$

Posterior predictive distribution $\pi(\tilde{y} \mid y) = \int \mathrm{d}\theta\; f(\tilde{y} \mid \theta)\; g(\theta \mid y)$

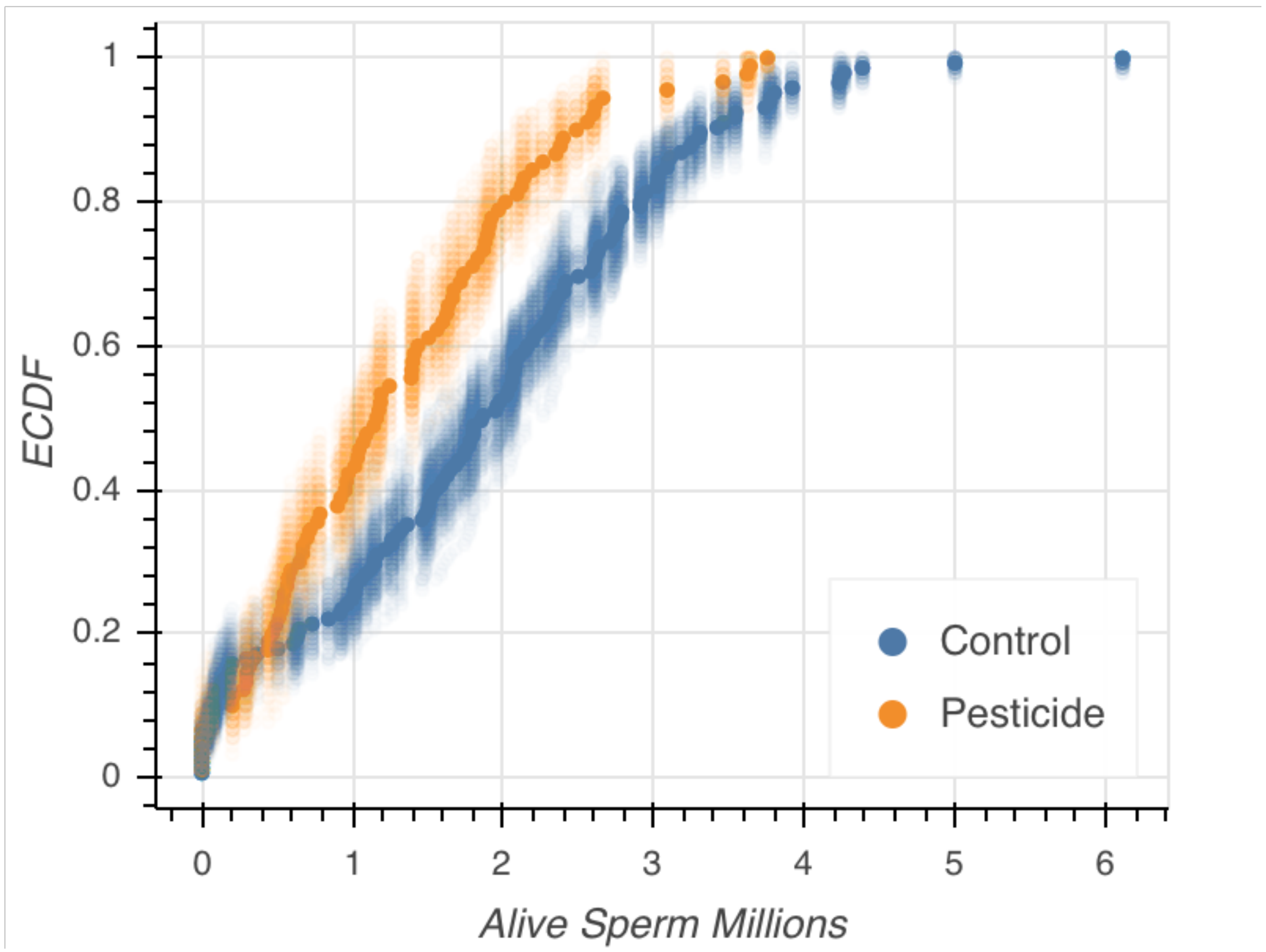
If we do not have a model, we can use the plug-in principle

Generative joint distribution $\pi(\tilde{y}, \theta \mid y) = f(\tilde{y} \mid \theta)\; g(\theta \mid y)$

Posterior predictive distribution $\pi(\tilde{y} \mid y) = \int \mathrm{d}\theta\; f(\tilde{y} \mid \theta)\; g(\theta \mid y)$

Plug-in predictive distribution $\pi(\tilde{y} \mid y) \approx \pi_{\mathrm{empirical}}(\tilde{y} \mid y)$

Bootstrap approaches can be useful and easily implemented



Many intriguing and useful topics await…

In previous editions of BE/Bi 103	Particle tracking Image correlation Particle image velocimetry (PIV) Watershed algorithms
2020 course from David Van Valen	Deep learning methods in image processing

Many intriguing and useful topics await…

In previous editions of BE/Bi 103

- Bayes factors
- Affine invariant MCMC
- Parallel tempering MCMC
- Approximate Bayesian computation
- Variational Bayesian inference

Many intriguing and useful topics await…

In future editions of BE/Bi 103?

Hidden Markov models

Gaussian processes

Integrated nested Laplace approximation

Expectation maximization

Nonparametric Bayes

Sparse regression

Missing data

Many intriguing and useful topics await…

In future editions of BE/Bi 103?

Handling big data (Dask, DataShader)

Building dashboards (Bokeh, HoloViews)

Many intriguing and useful topics await…

In previous editions of BE/Bi 103

- K-means clustering
- Support vector machines
- t-distributed stochastic neighbor embedding
- Kernel density estimation
- LASSO and ridge regression

CS/CNA/EE 156 a

CMS/CS/CNS/EE/IDS 155

- A world of machine learning

Many intriguing and useful topics await…

Bi/BE/CS 183

More in the future

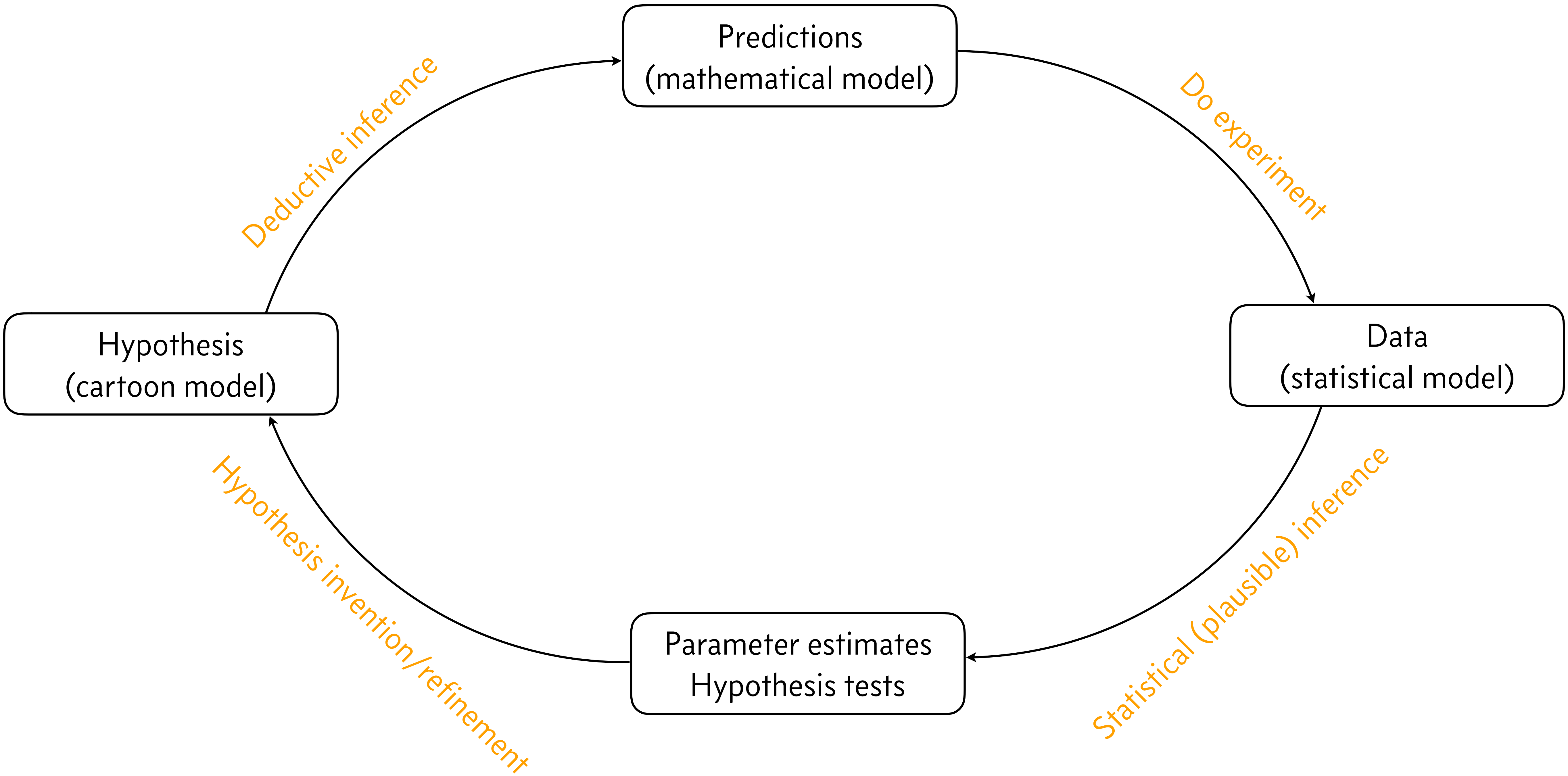
A world of bioinformatics

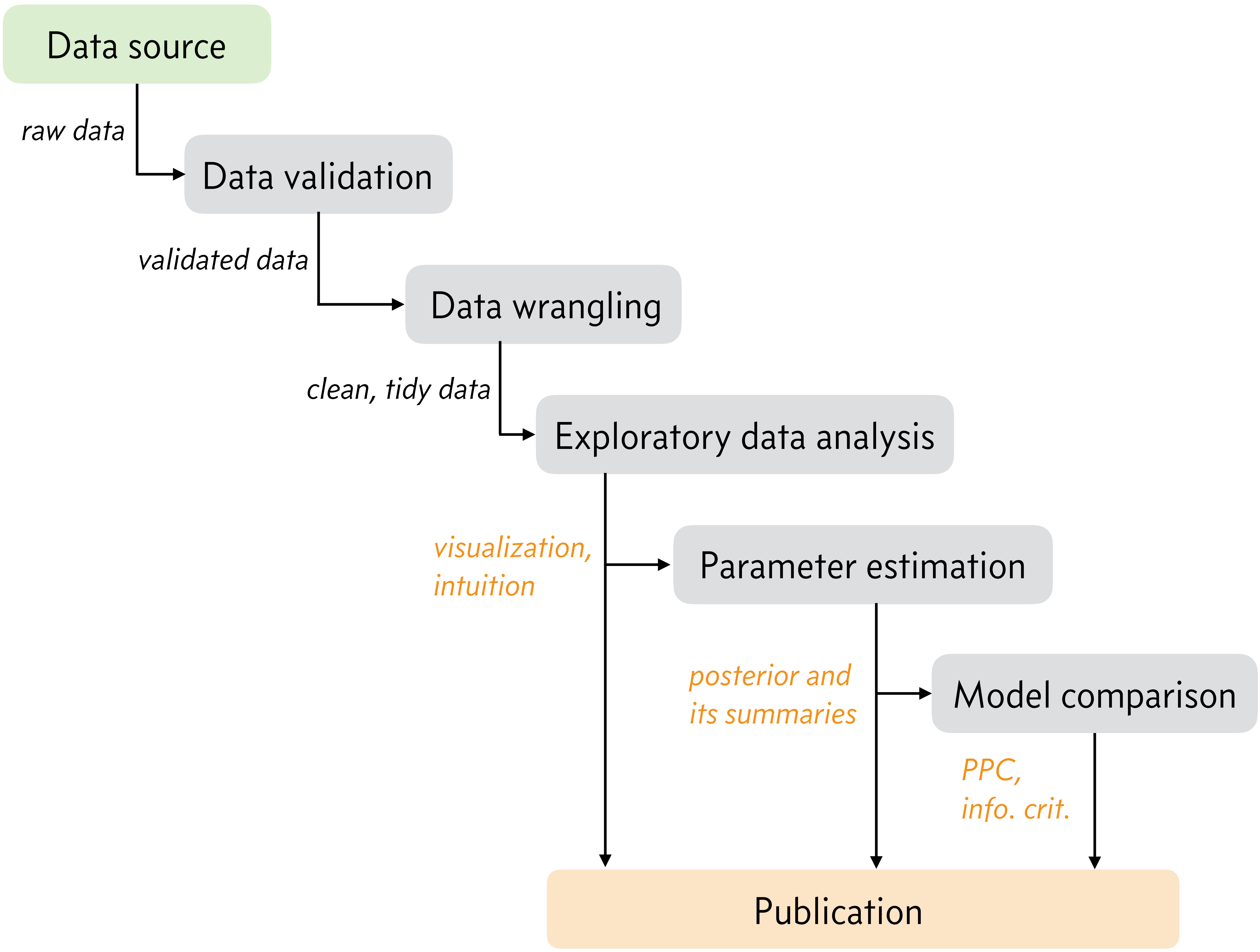
ACM/EE/IDS 116

ACM/CS/IDS 157

A world of frequentist statistics

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
PPC,
info. crit.
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

ANACONDA

aws

GitHub

Thank you to the data sources

Caltech

- Avni Gandhi, Audrey Chen, Grigorios Oikonomou, and David Prober
- 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
- Meaghan Sullivan, Kevin Yu, Jimmy Hamilton, and the students of Bi 1x
- Han Wang and Paul Sternberg
- Emily Blythe and Ray Deshaies
- Julian Wagner and Joe Parker
- Zach Shao and Julie Kornfield

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 (MPI-CBG)
- 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)
- Simon Harvey and Helen Orbidans, Christchurch

Thank you

203 contributors to JupyterLab

332 contributors to Bokeh

62 contributors to Altair

285 contributors to scikit-image

206 contributors to conda

1349 contributors to Pandas

709 contributors to Numpy

67 contributors to Stan

30 contributors to PyStan

16 contributors to ArviZ

Contributors to the rest of the open source software we've used

Thank you to Michael Betancourt

Thank you

John Ciemniecki

Sophie Miller

Christina Su

Julian Wagner

Thank you

All of you!

Go forth and…

Use what you have learned to do reproducible quantitative research.

Evangelize workflows for reproducible science.