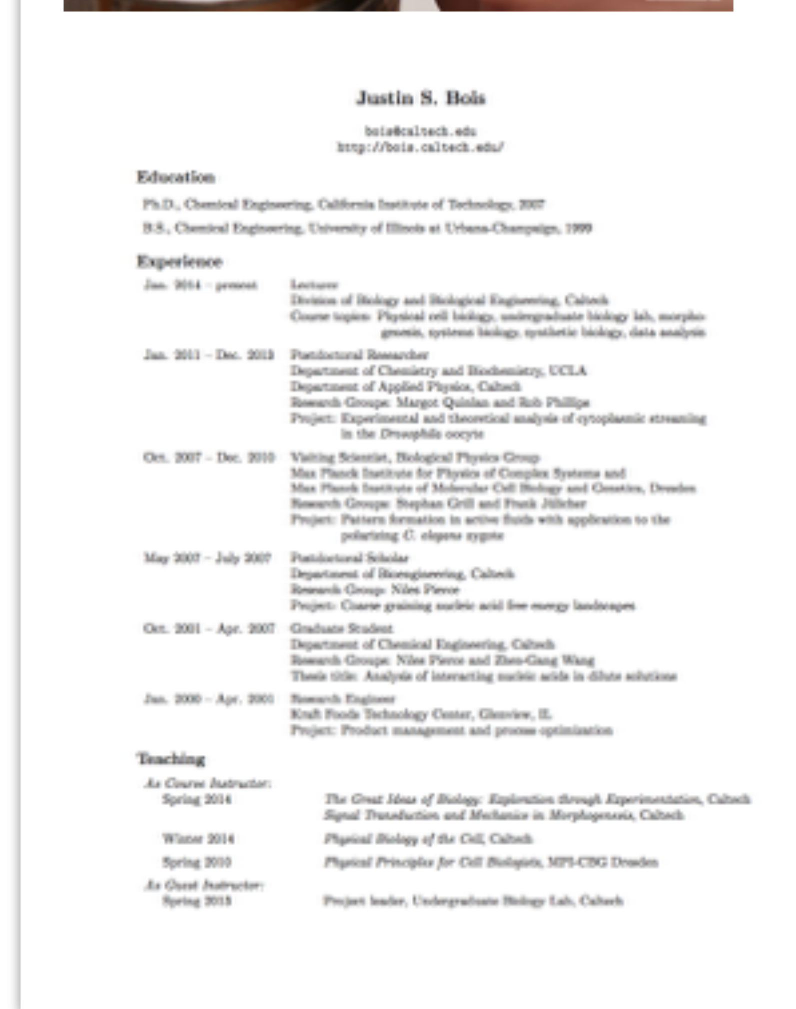


Justin S. Bois

bois@caltech.edu
http://bois.caltech.edu/

Education

Ph.D., Chemical Engineering, California Institute of Technology, 2007

B.S., Chemical Engineering, University of Illinois at Urbana-Champaign, 1999

Experience

Jan. 2014 – present	Lecturer Division of Biology and Biological Engineering, Caltech Course topics: Physical cell biology, undergraduate biology lab, morphogenesis, systems biology, synthetic biology, data analysis
Jan. 2011 – Dec. 2013	Postdoctoral Researcher Department of Chemistry and Biochemistry, UCLA Department of Applied Physics, Caltech Research Groups: Margot Quinlan and Rob Phillips Project: Experimental and theoretical analysis of cytoplasmic streaming in the *Drosophila* oocyte
Oct. 2007 – Dec. 2010	Visiting Scientist, Biological Physics Group Max Planck Institute for Physics of Complex Systems and Max Planck Institute of Molecular Cell Biology and Genetics, Dresden Research Groups: Stephan Grill and Frank Jülicher Project: Pattern formation in active fluids with application to the polarizing *C. elegans* zygote
May 2007 – July 2007	Postdoctoral Scholar Department of Bioengineering, Caltech Research Group: Niles Pierce Project: Coarse graining nucleic acid free energy landscapes
Oct. 2001 – Apr. 2007	Graduate Student Department of Chemical Engineering, Caltech Research Groups: Niles Pierce and Zhen-Gang Wang Thesis title: Analysis of interacting nucleic acids in dilute solutions
Jan. 2000 – Apr. 2001	Research Engineer Kraft Foods Technology Center, Glenview, IL Project: Product management and process optimization

Teaching

As Course Instructor:

Spring 2014	*The Great Ideas of Biology: Exploration through Experimentation*, Caltech *Signal Transduction and Mechanics in Morphogenesis*, Caltech
Winter 2014	*Physical Biology of the Cell*, Caltech
Spring 2010	*Physical Principles for Cell Biologists*, MPI-CBG Dresden

As Guest Instructor:

Spring 2013	Project leader, Undergraduate Biology Lab, Caltech

OpenCV

Useful MCMC packages: OpenBUGS, RJAGS, RStan

Useful plotting packages: ggplot2, shiny

Useful data management packages: dplyr2, tidyr

RStudio®

ANACONDA®

Useful MCMC packages: OpenBUGS, RJAGS, RStan

Useful plotting packages: ggplot2, shiny

Useful data management packages: dplyr2, tidyr

Jython

OpenCV

ilastik

the interactive learning and segmentation toolkit

CellProfiler

cell image analysis software

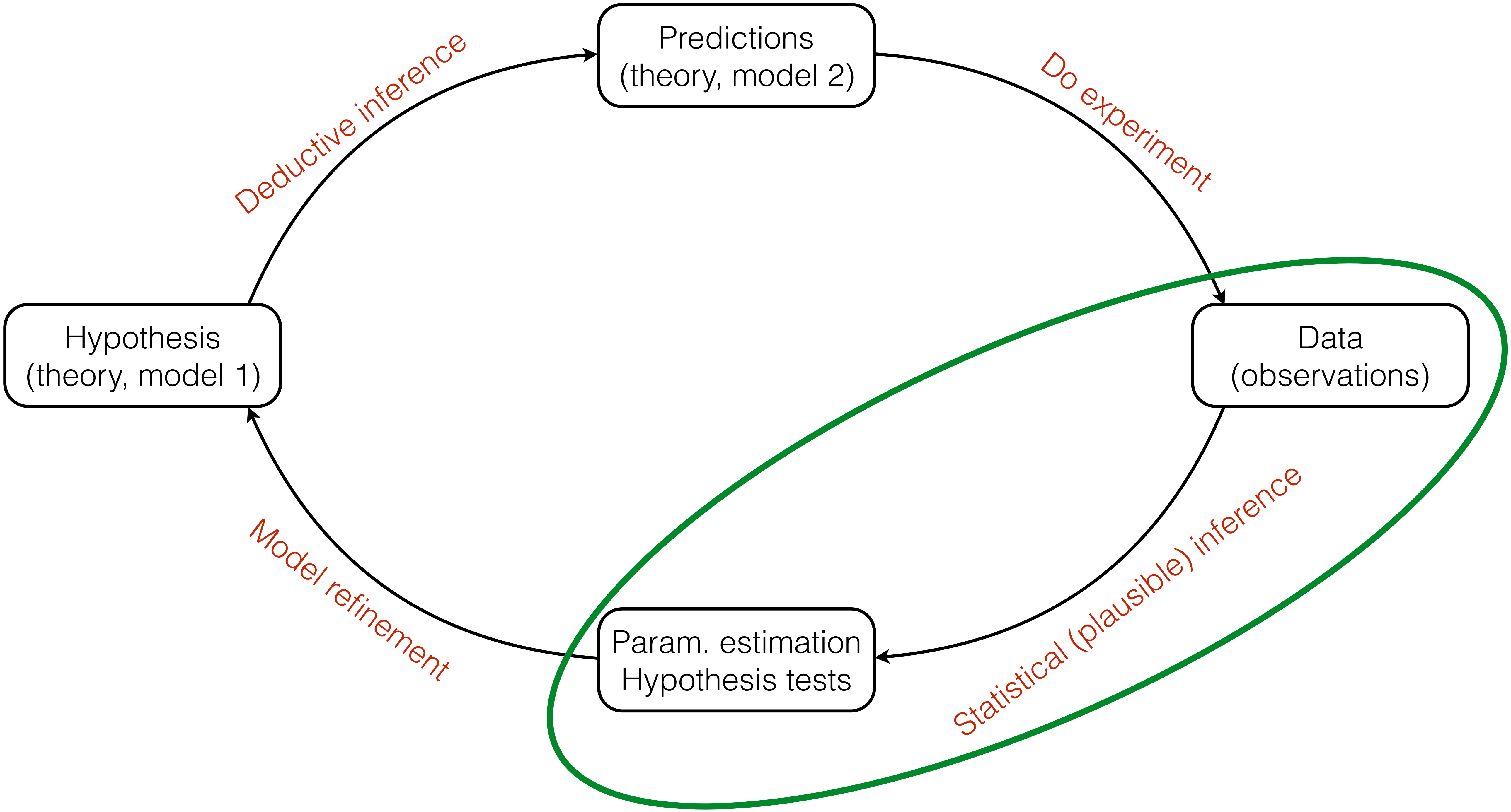
icy

BE/Bi 103

Data Analysis in the Biological Sciences

Fall term, 2015

The scientific method



Statistical inference requires a probability theory

M_i: model i

$\mathbf{a}_i$: the set of parameters associated with model i

D: the measured data

I: all other knowledge

Bayes's theorem for parameter estimation:

$$\text{posterior} = P(\mathbf{a}_i|D, M_i, I) = \frac{P(D|\mathbf{a}_i, M_i, I)P(\mathbf{a}_i|M_i, I)}{P(D|M_i, I)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Normalization of posterior (marginalization):

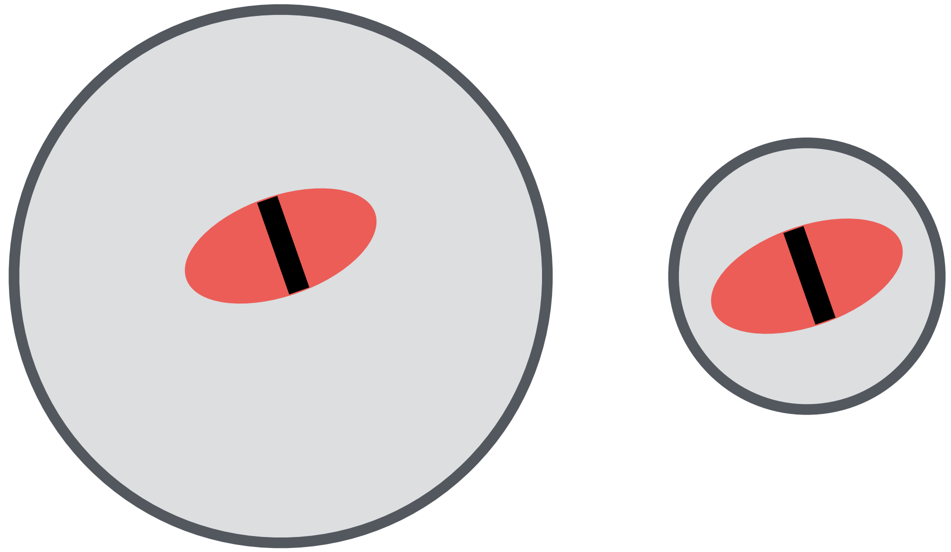
$$P(D|M_i, I) = \int \mathrm{d}\mathbf{a}\, P(D|\mathbf{a}_i, M_i, I)P(\mathbf{a}_i|M_i, I)$$

Bayes's theorem for model selection:

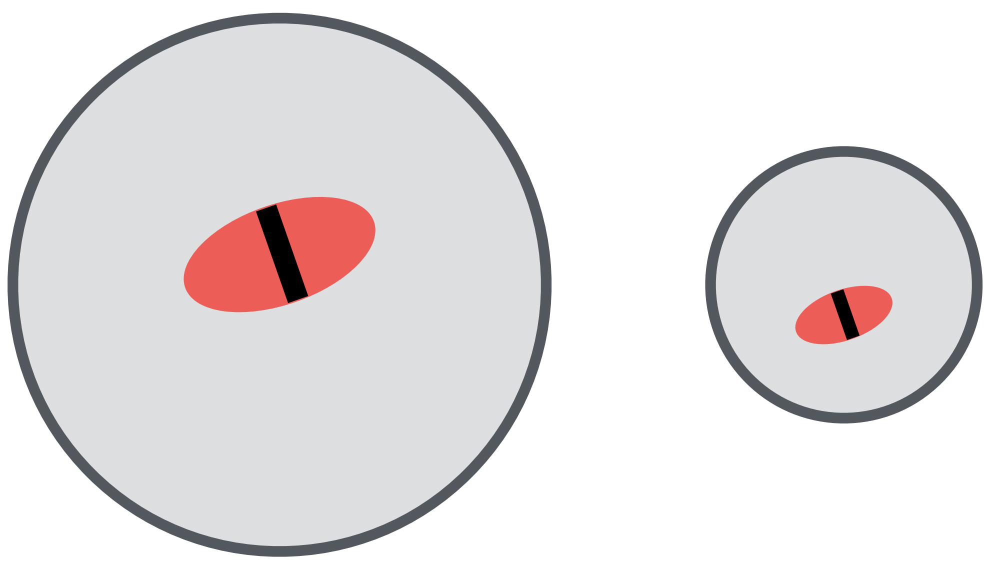
$$P(M_i|D, I) = \frac{P(D|M_i, I)\, P(M_i|I)}{P(D|I)}$$

Model type 1: Cartoons (informal)

Model a:



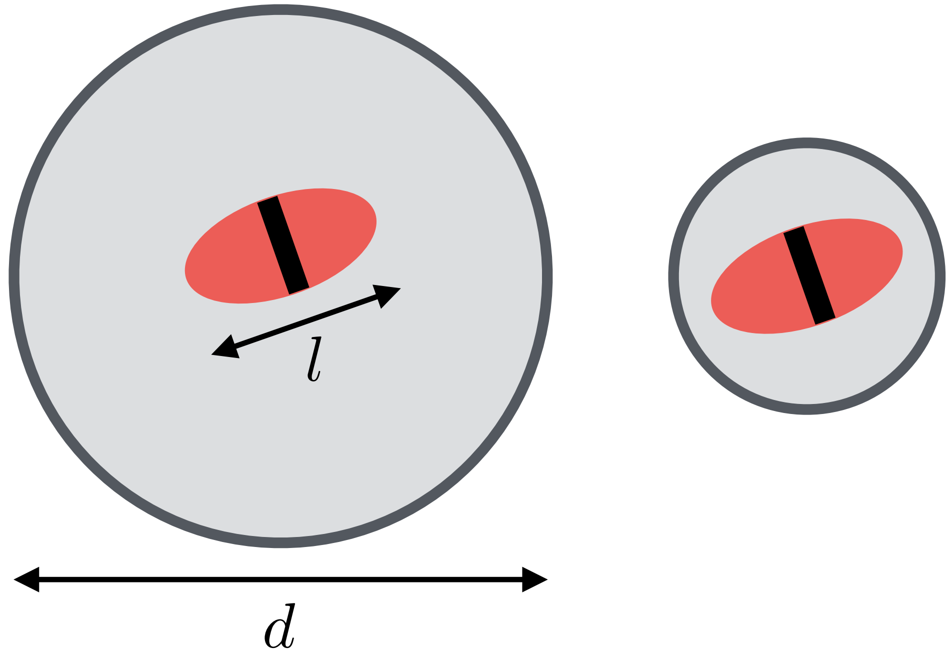
Model b:



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

Model type 2: Mathematized cartoons (formal)

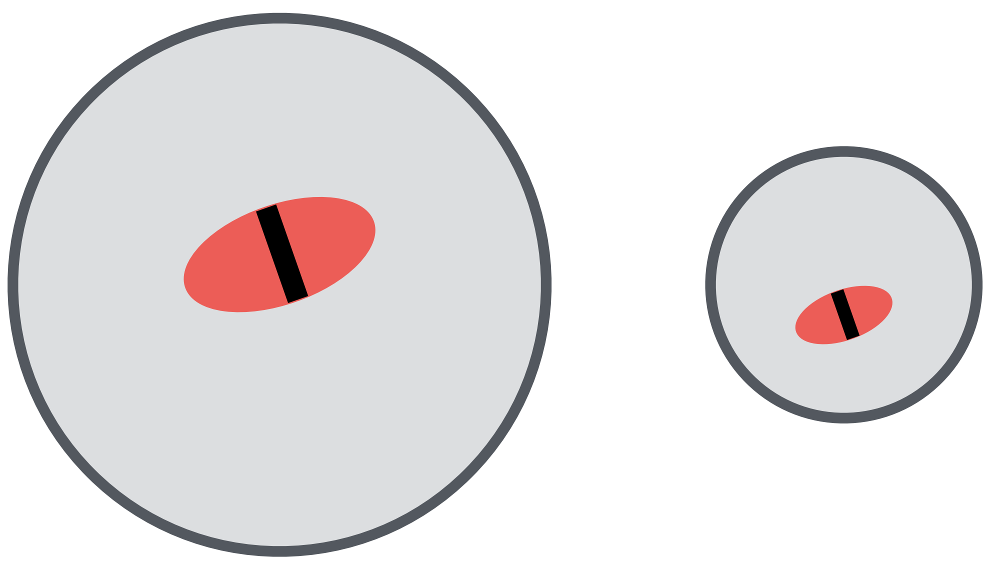
Model a:



$$l \neq l(d)$$

$$l = \theta$$

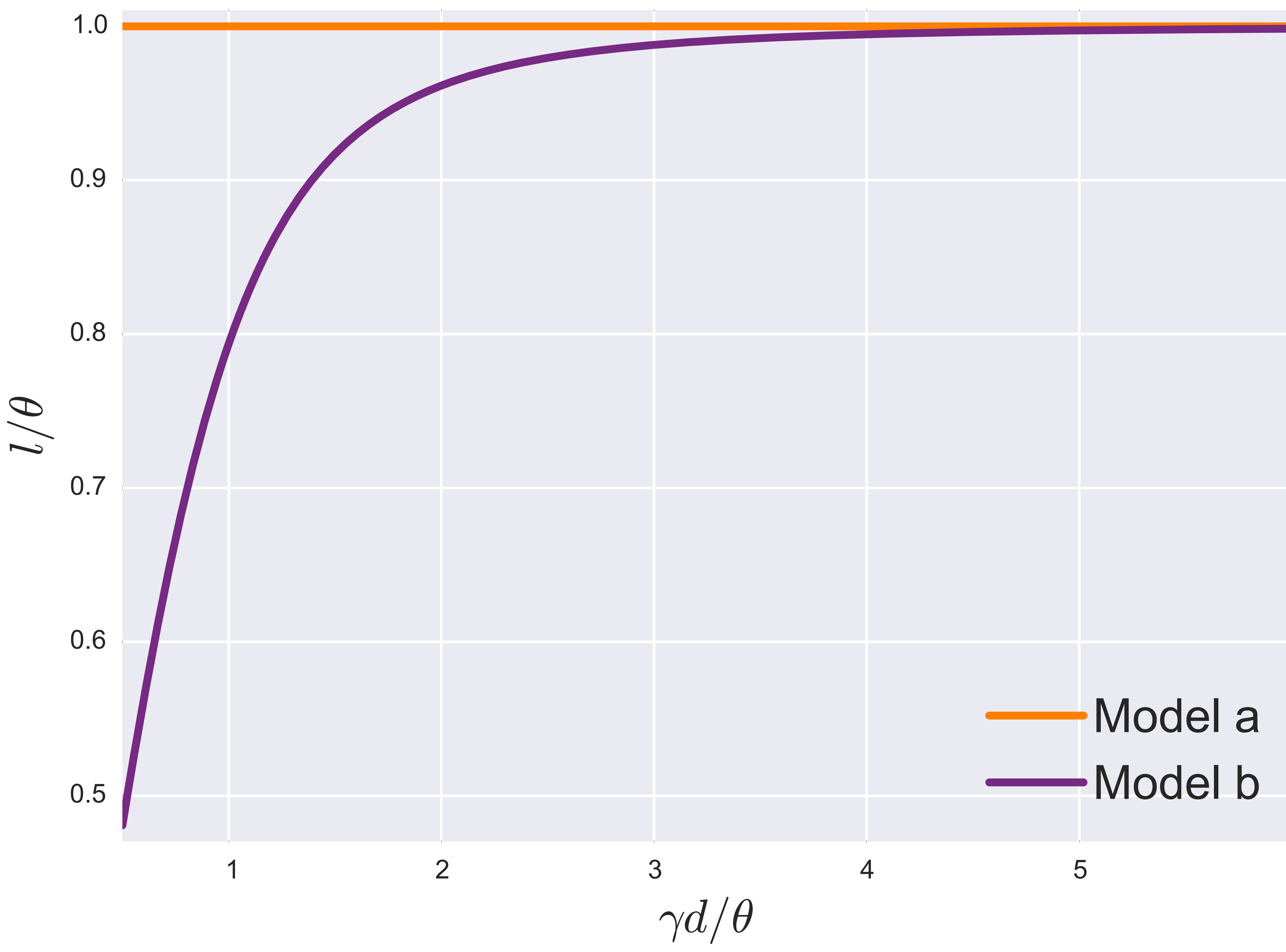
Model b:



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

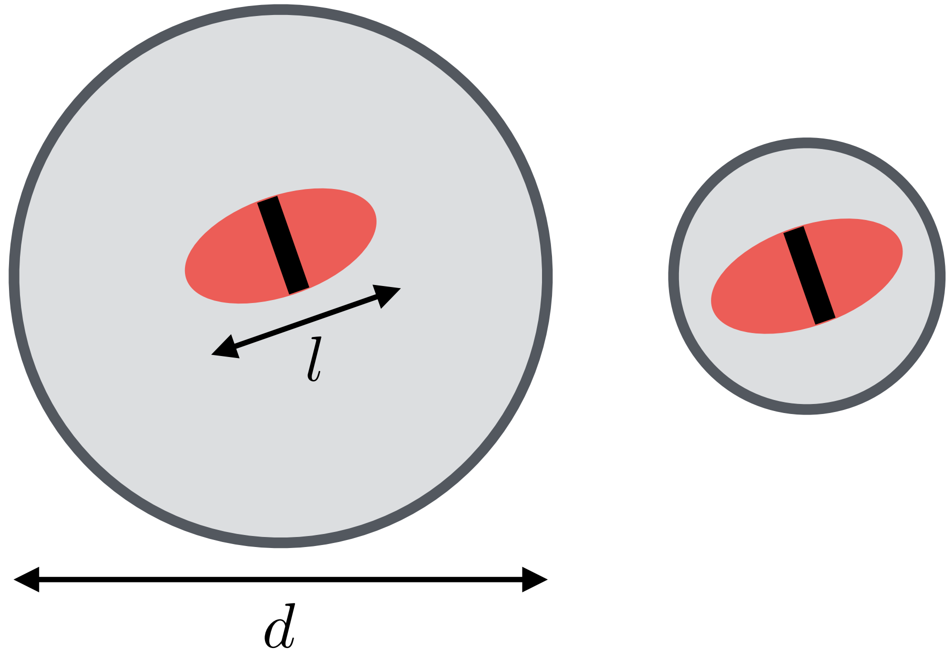
Good, et al., *Science*, **342**, 856, 2013

Model type 2: Mathematized cartoons (formal)



Model type 3: Model type 2 + data description

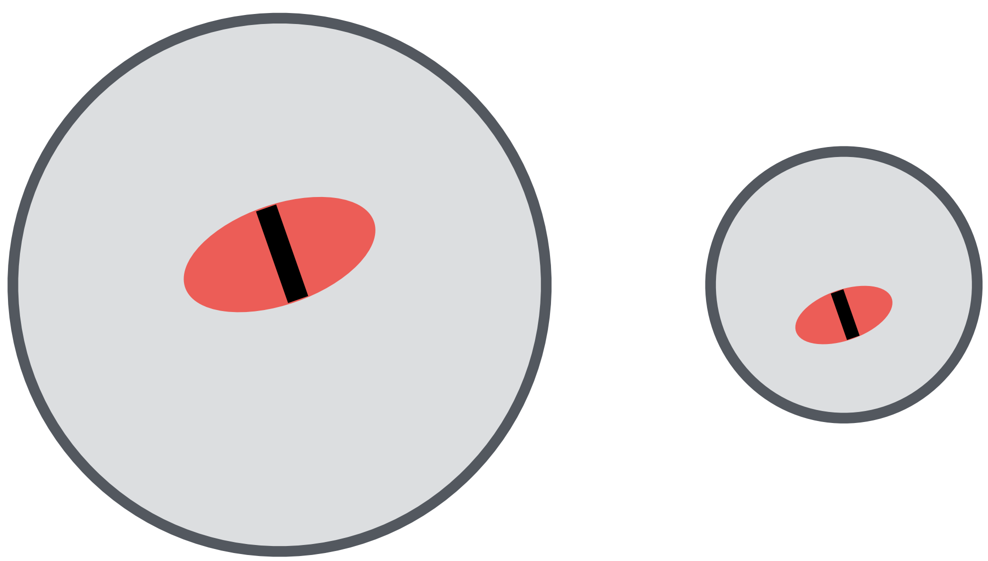
Model a:



$$l_i = \theta + e_i$$

e_i Gaussian distributed

Model b:



$$l_i = \frac{\gamma d_i}{(1 + (\gamma d_i/\theta)^3)^{\frac{1}{3}}} + e_i$$

e_i Gaussian distributed

Models (definition 3!)

M_i and I encode the functional form of the likelihood $P(D|\mathsf{a}_i, M_i, I)$ and prior $P(\mathsf{a}_i|M_i, I)$.

Prior $P(\mathsf{a}_i|M_i, I)$: Often chosen to be **uninformative**, e.g., uniform or Jeffreys.

Likelihood $P(D|\mathsf{a}_i, M_i, I)$: Depends on model, often independent Gaussians.

Given the model and all our previous knowledge, *the posterior is completely determined*. All of the "work" of inference is computing it!

Computing the posterior: analytical results

Multiple measurements of parameter μ with unknown variance σ^2.

$$P(\{x_i\}|\mu,\sigma,I) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-(x_i-\mu)^2/2\sigma^2}$$

$$P(\mu,\sigma|I) \propto \sigma^{-1}$$

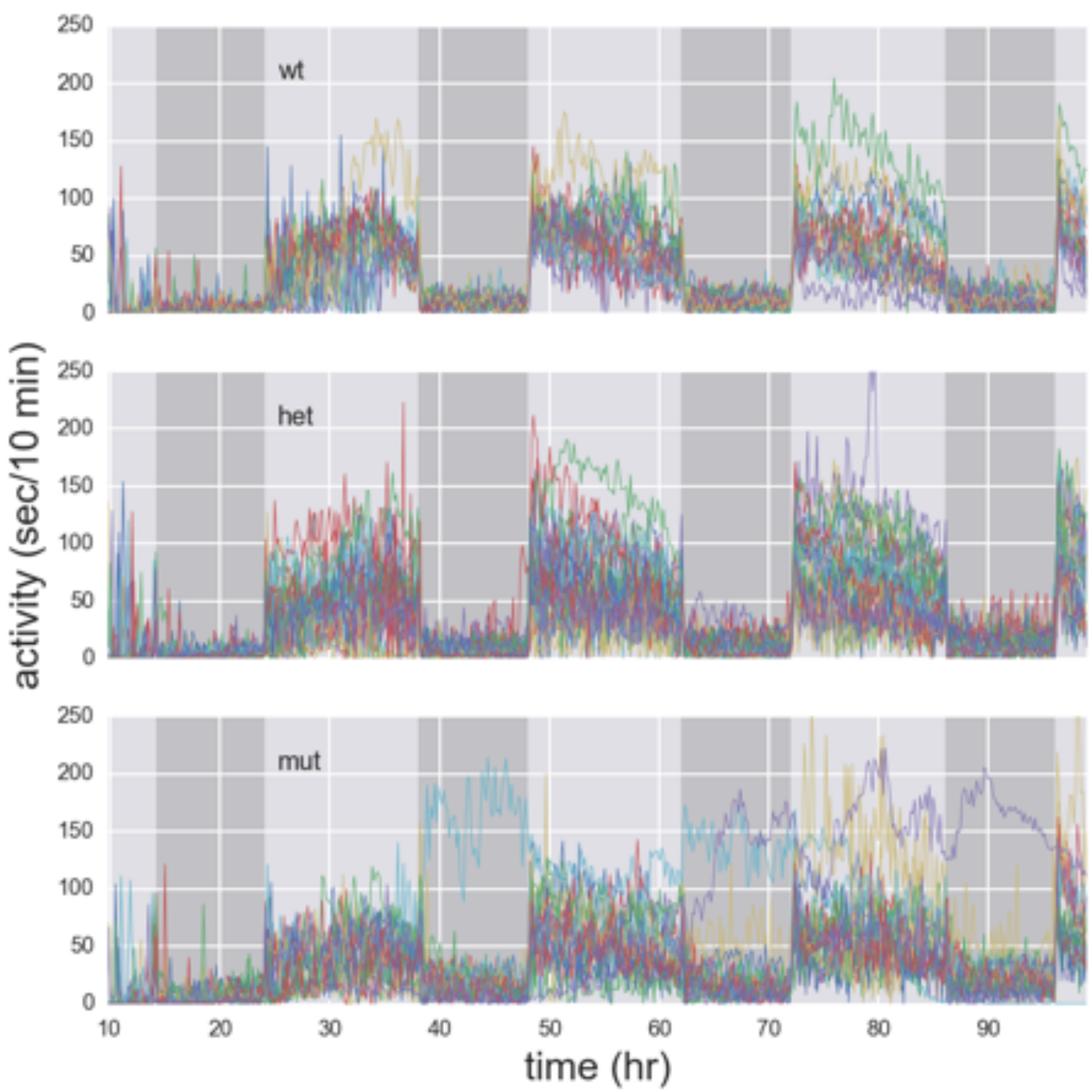
most probable $\mu = \bar{x} \equiv \frac{1}{n}\sum_i x_i$

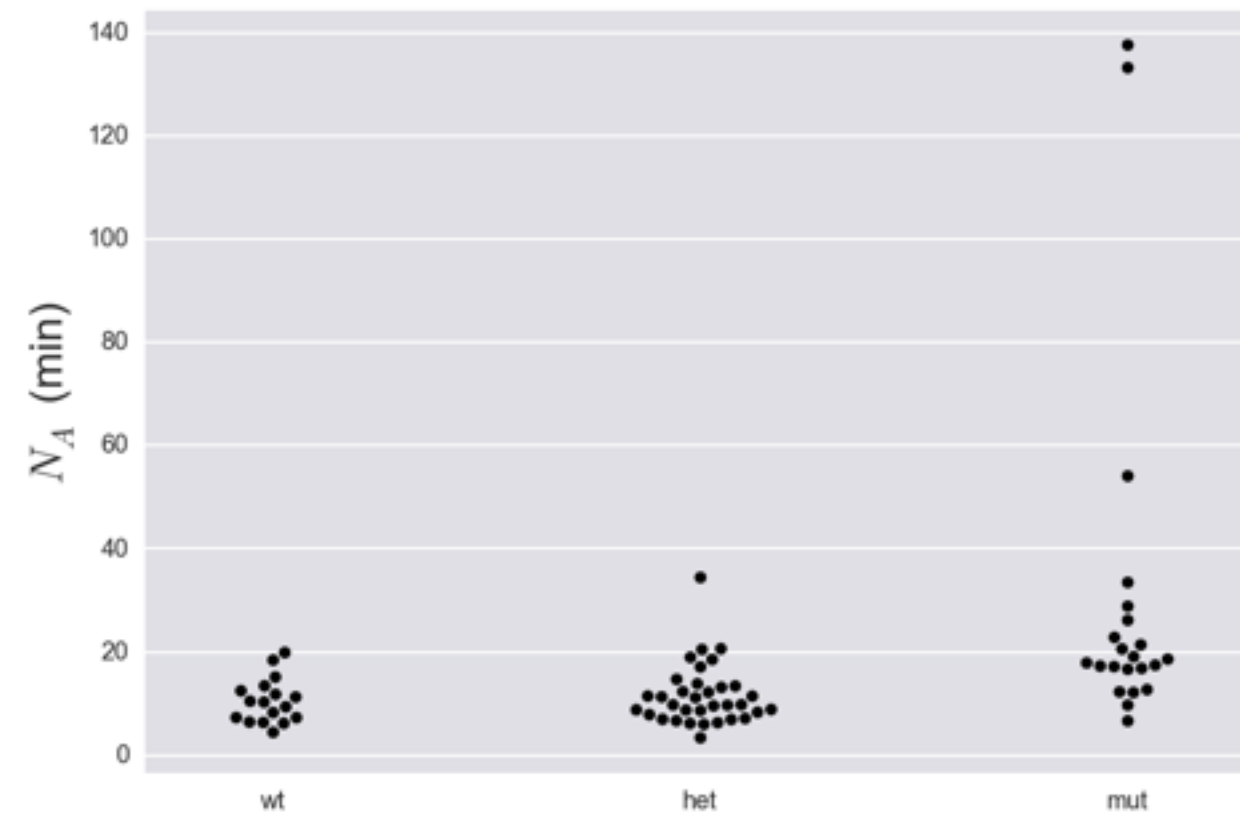
most probable $\sigma^2 = r^2 \equiv \frac{1}{n}\sum_i (x_i - \bar{x})^2$

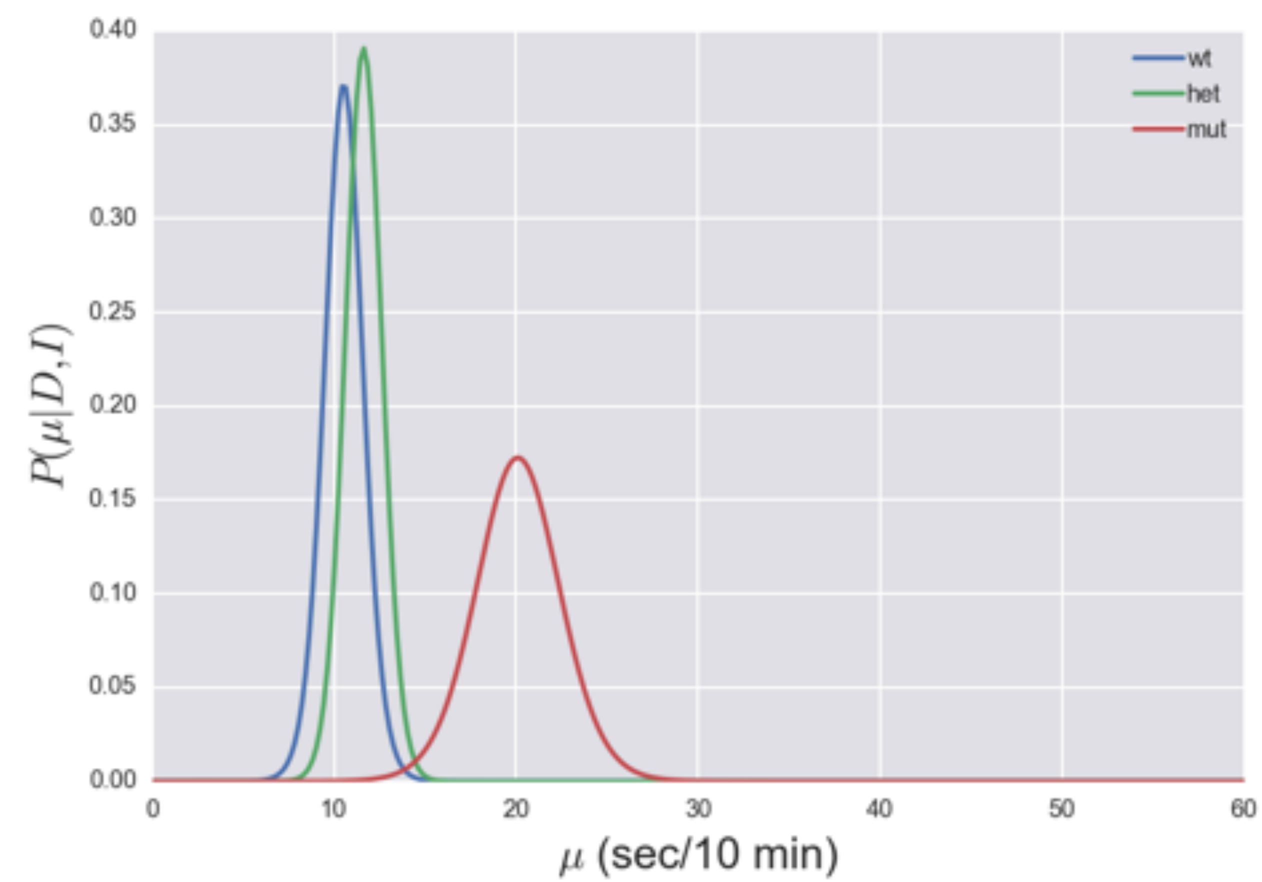
$$P(\mu|\{x_i\},I) \approx \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)}\frac{1}{r}\left(1+\frac{(\bar{x}-\mu)^2}{r^2}\right)^{-\frac{n}{2}} \quad \text{(Student-t)}$$

$$\mu \approx \bar{x} \pm r/\sqrt{n}$$

Computing the posterior: analytical results





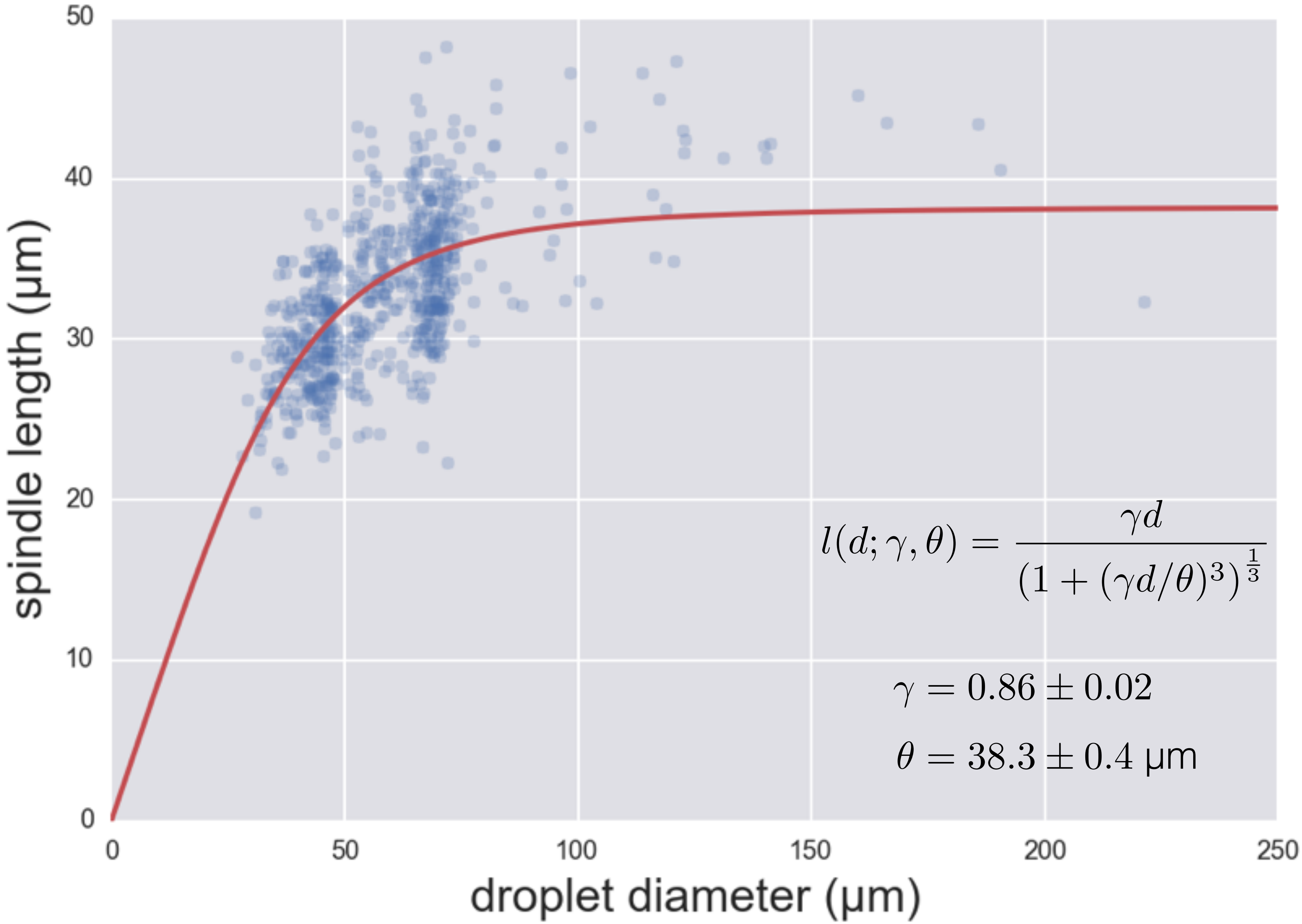


Computing the posterior: approximate summary

1. Find most probable parameters $\mathbf{a}^*$.
2. Approximate $P(\mathbf{a}|D, I)$ as Gaussian by doing a Taylor expansion of $\ln P(\mathbf{a}|D, I)$ about $\mathbf{a}^*$.
3. The covariance matrix is given by the negative inverse of the Hessian of $\ln P(\mathbf{a}|D, I)$.

Obvious assumption: posterior is approximately Gaussian.

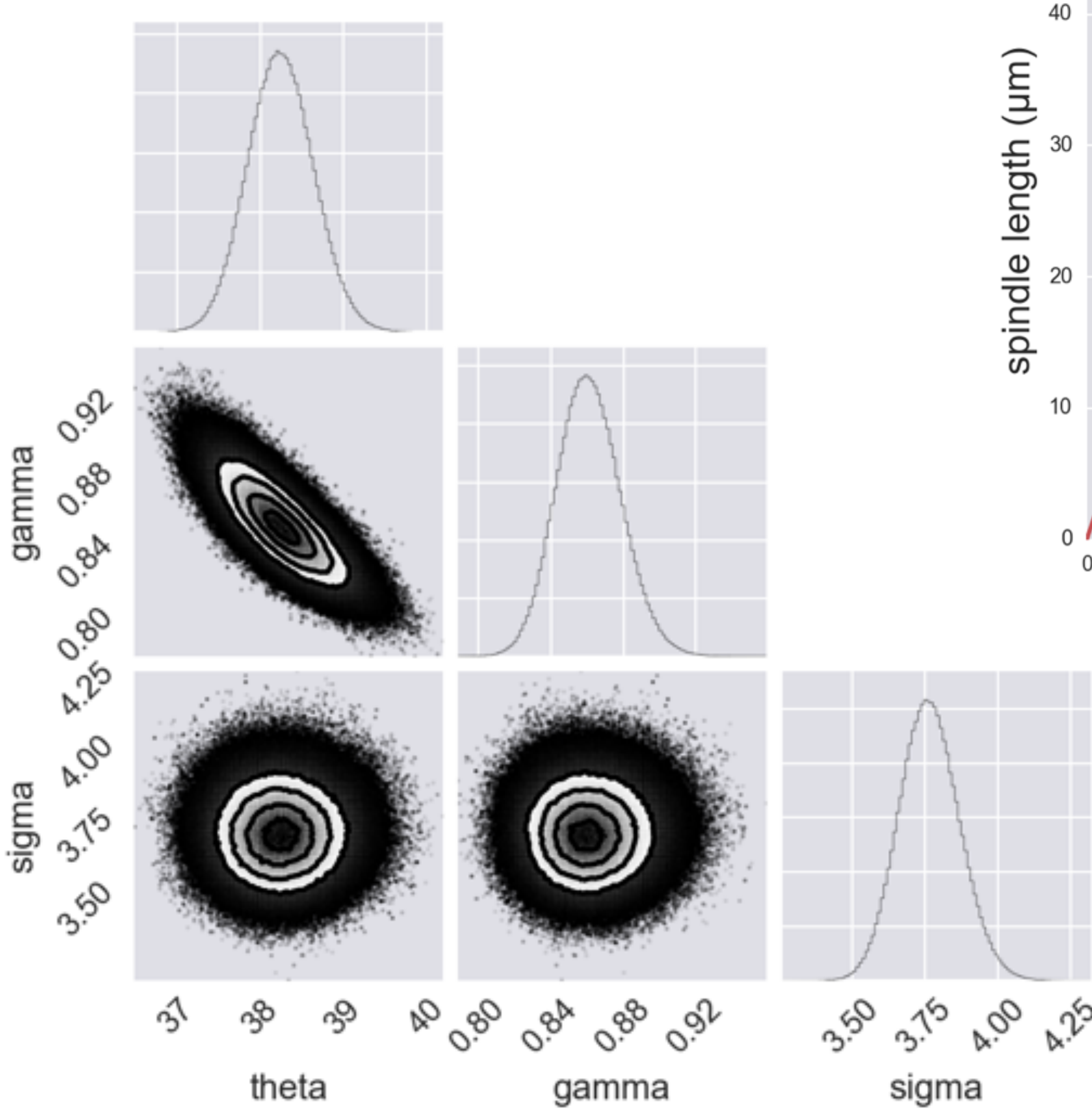
Computing the posterior: approximate summary

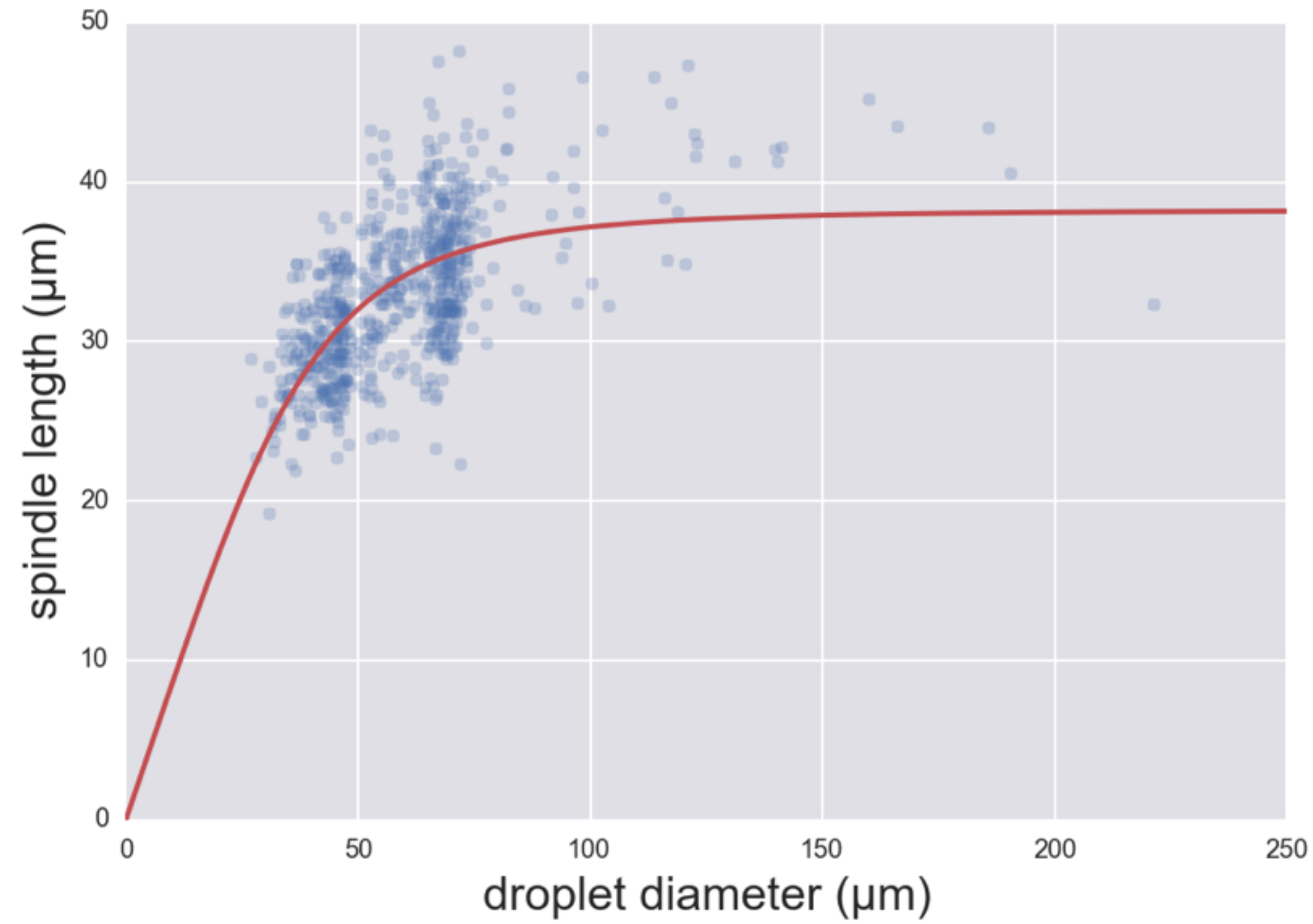


Computing the posterior: MCMC

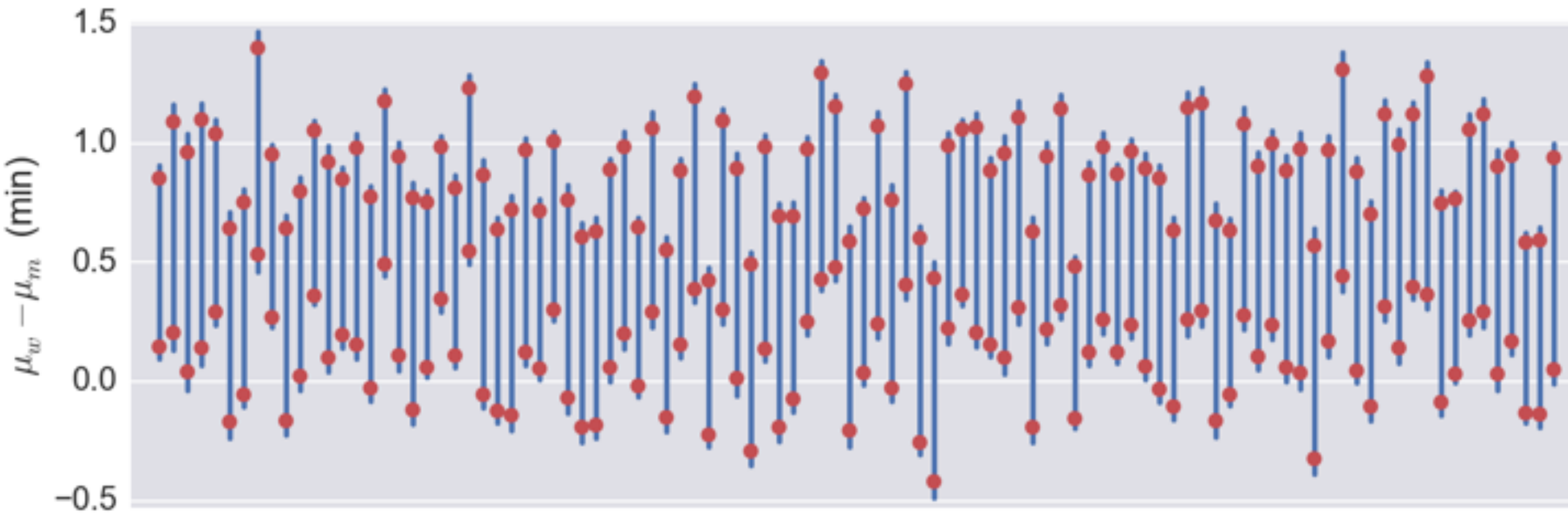
1. Define the (log) posterior distribution.
2. Efficiently sample the posterior with an ergodic, positively recurrent Markov chain.
3. Posterior is trivially marginalized by considering specific parameters.
4. Bin samples to get histograms describing posterior.

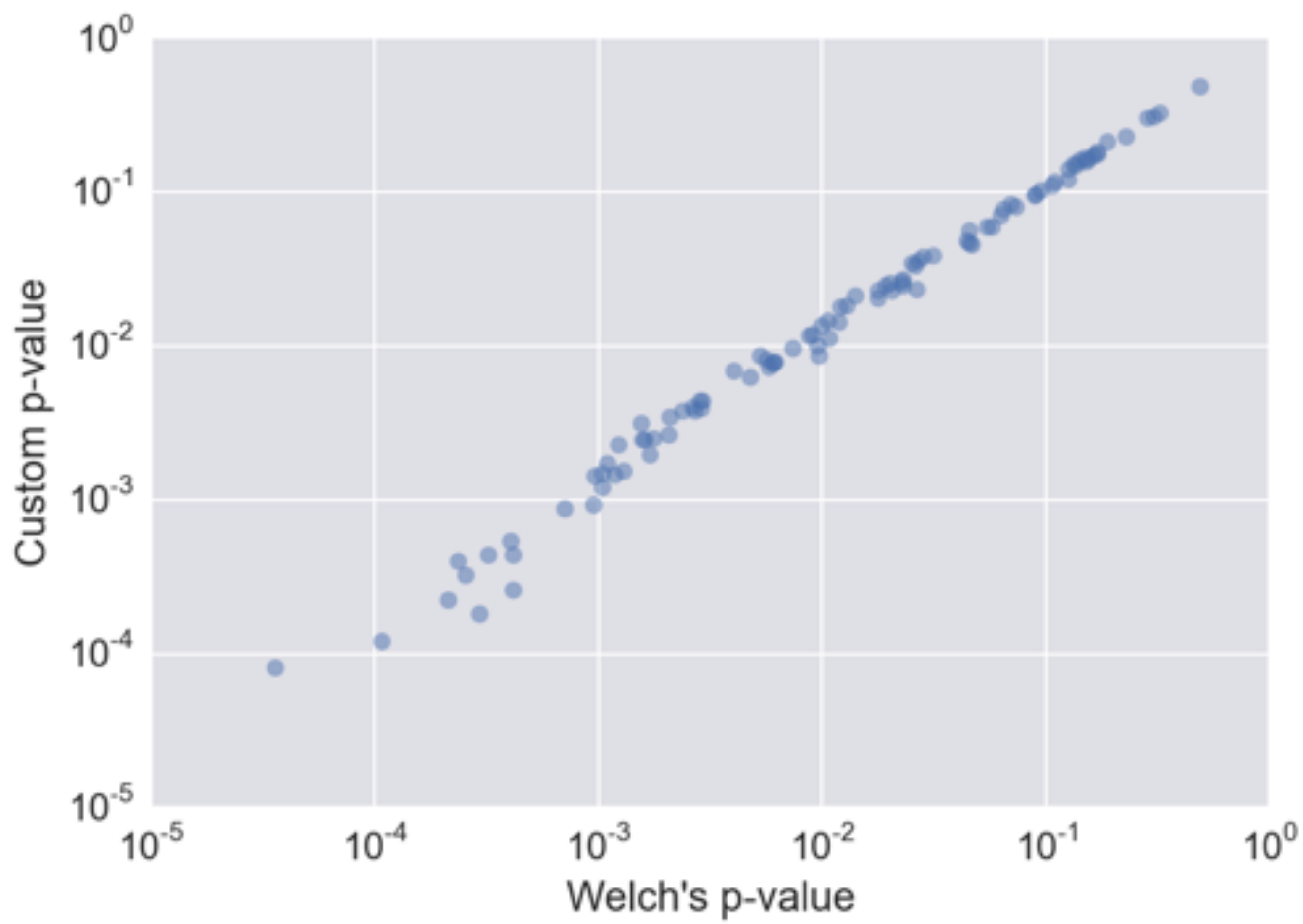
Computing the posterior: MCMC





Foray into frequentism





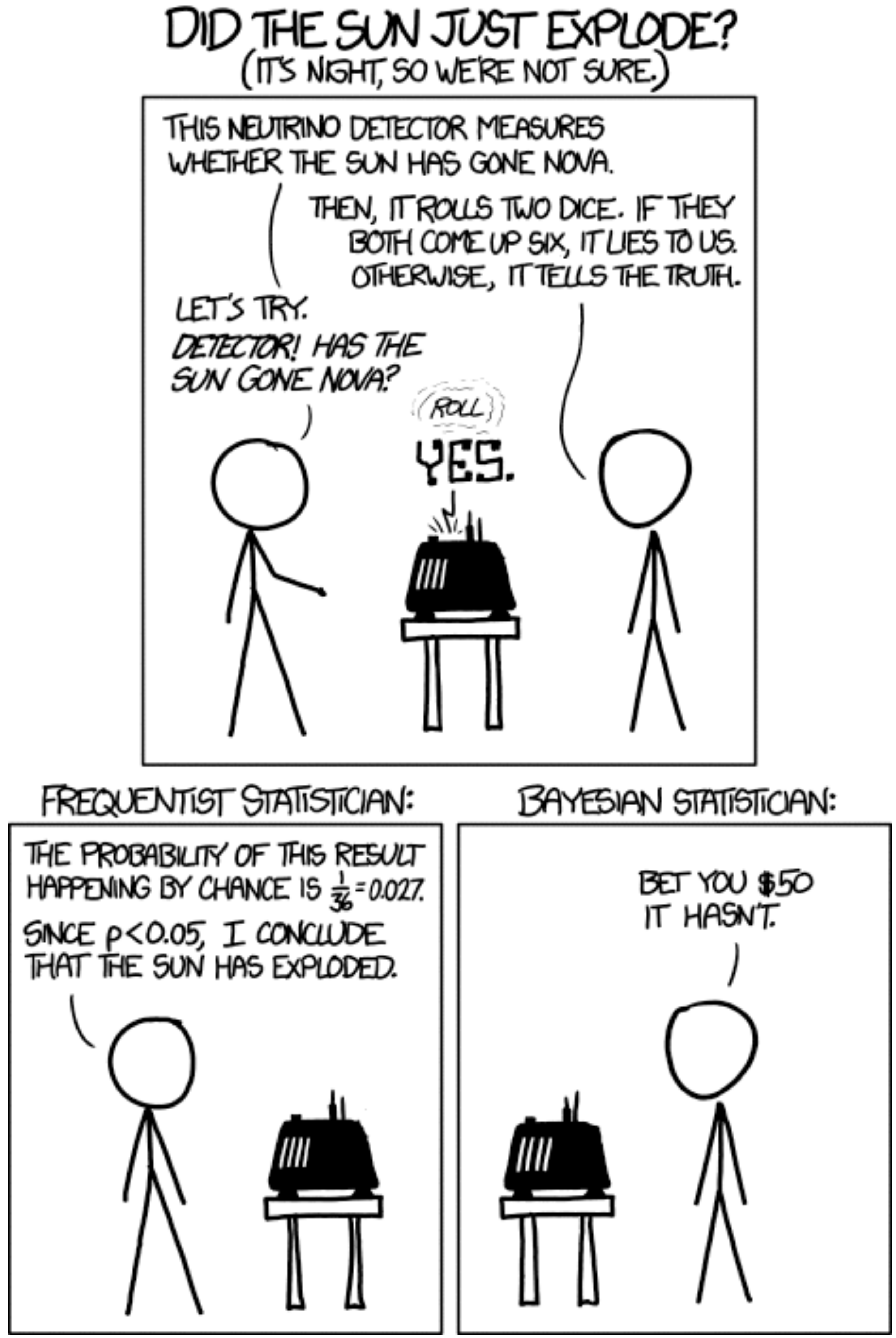
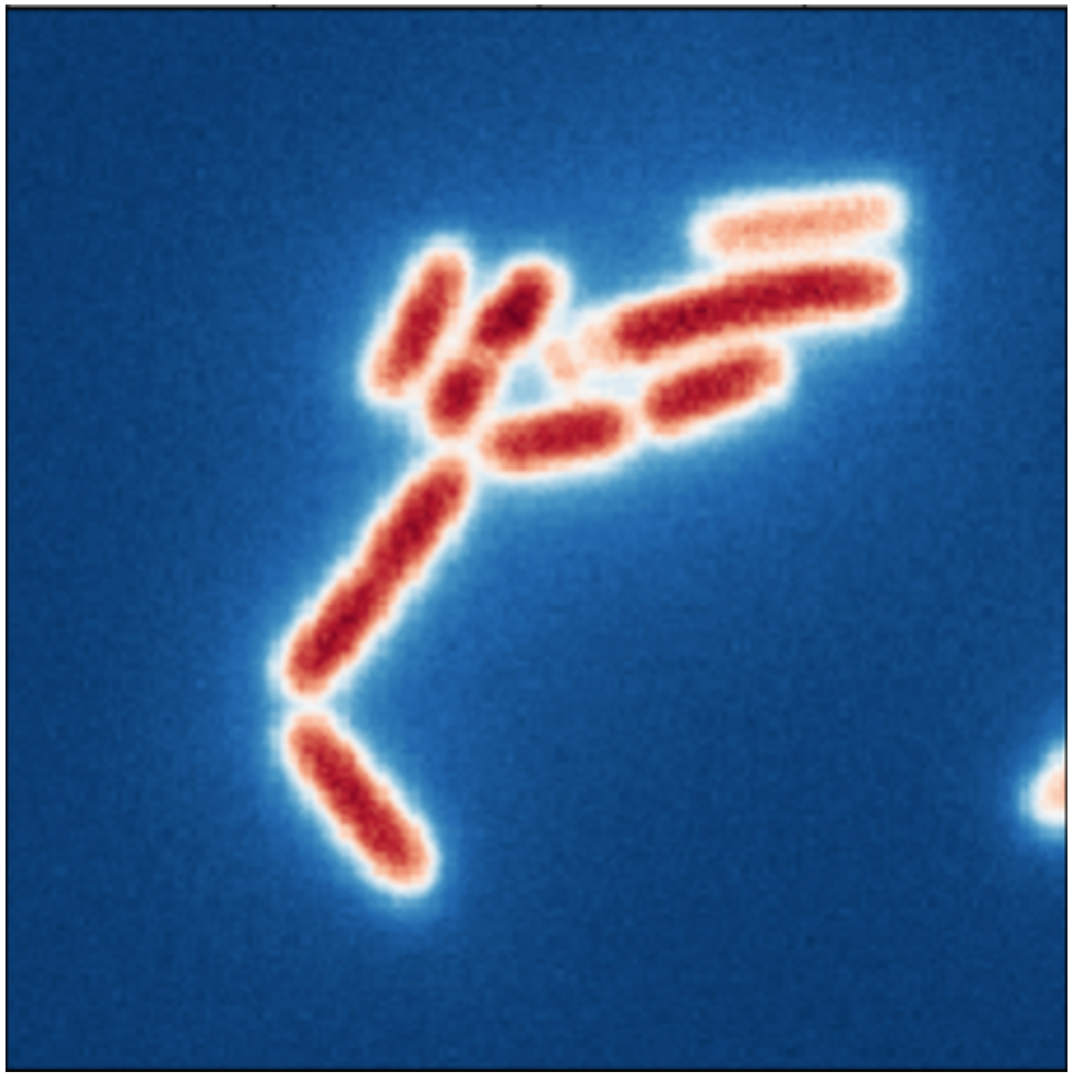

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE 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.

Image segmentation



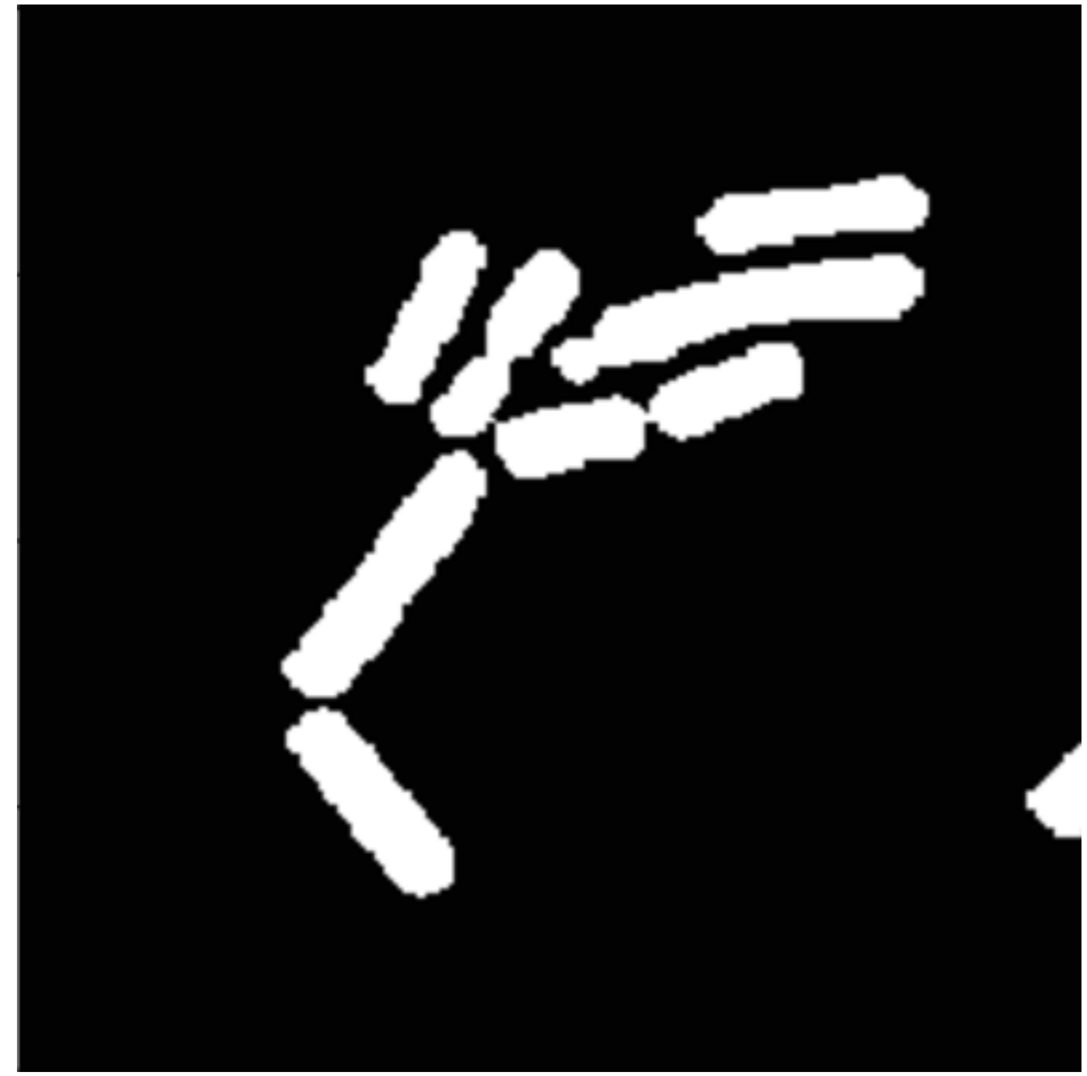
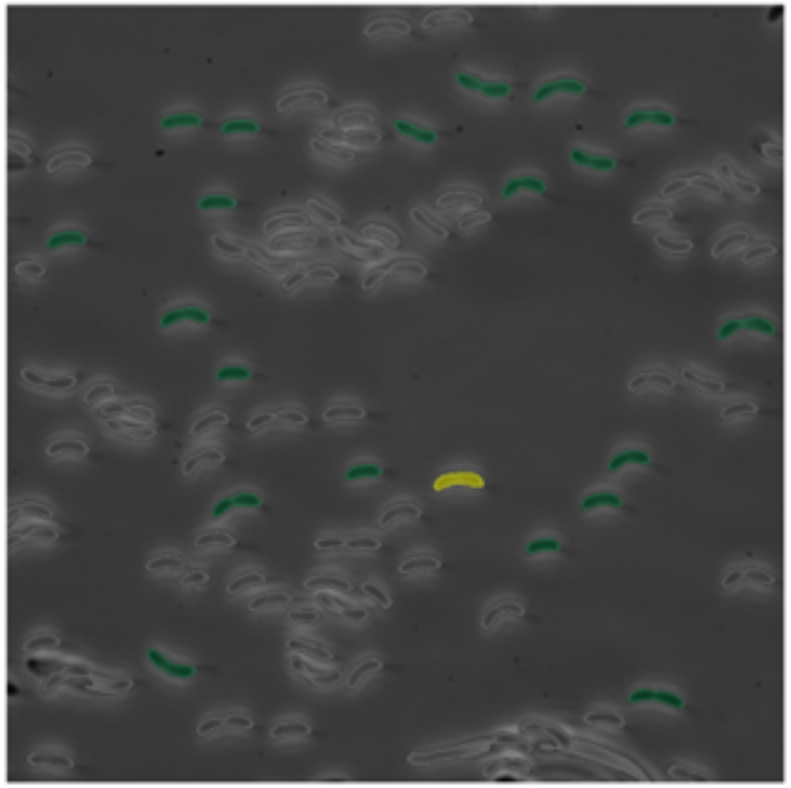
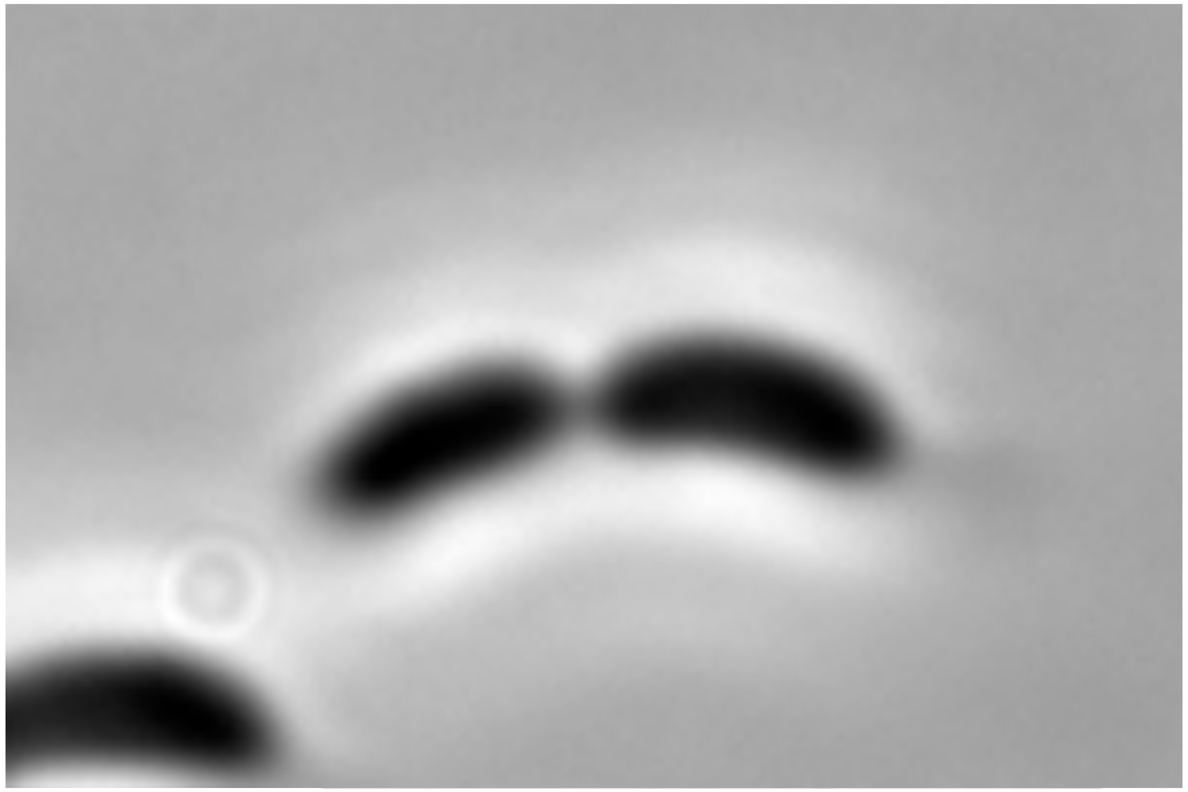
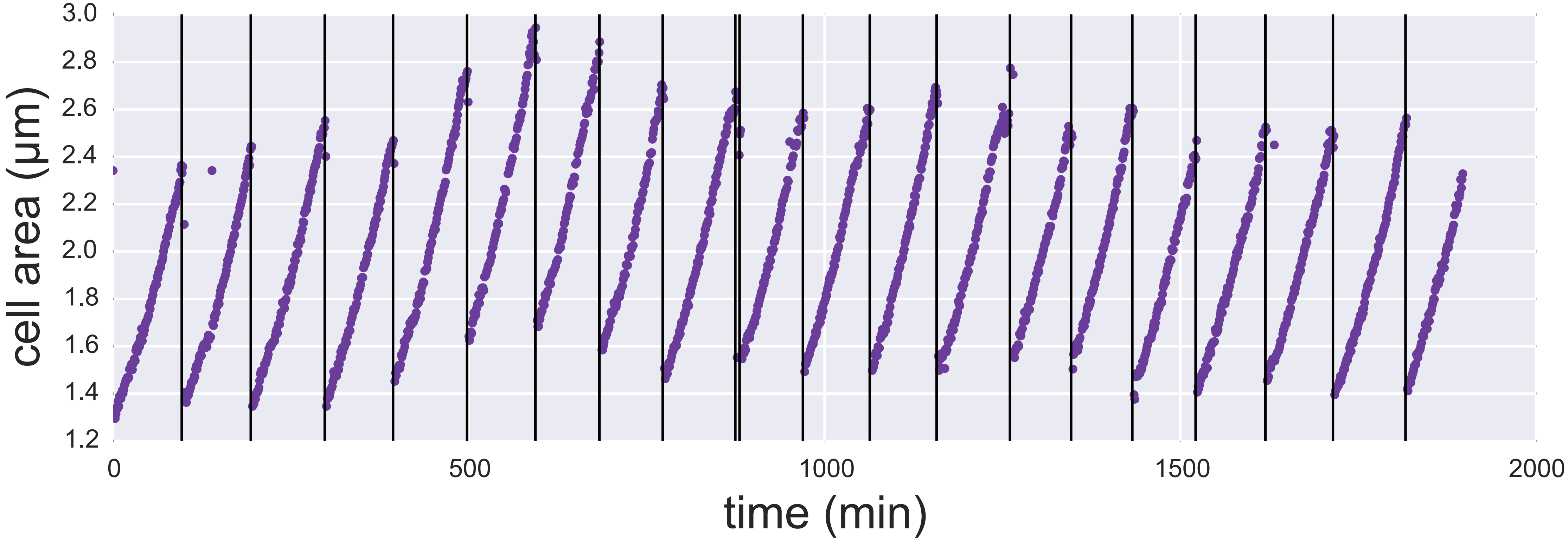


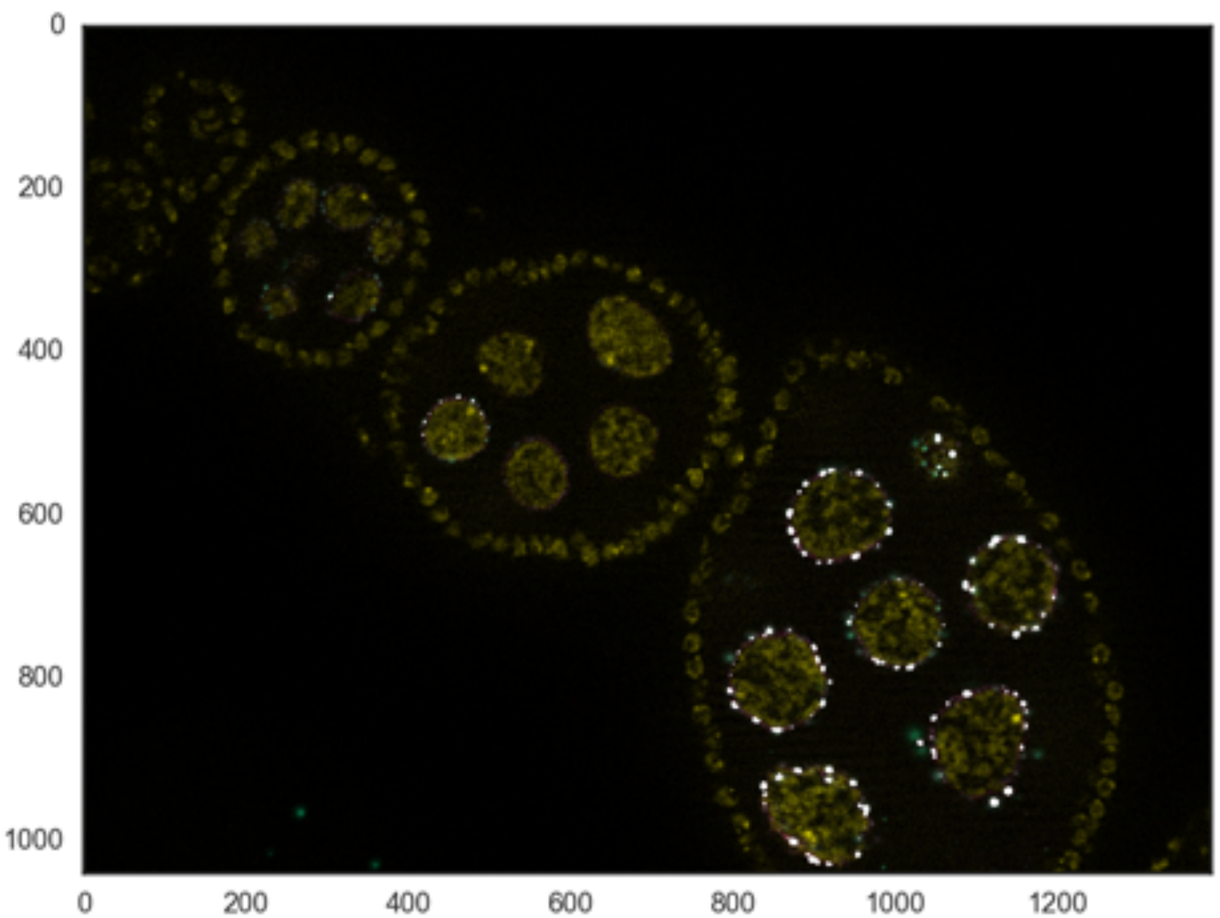
Image segmentation

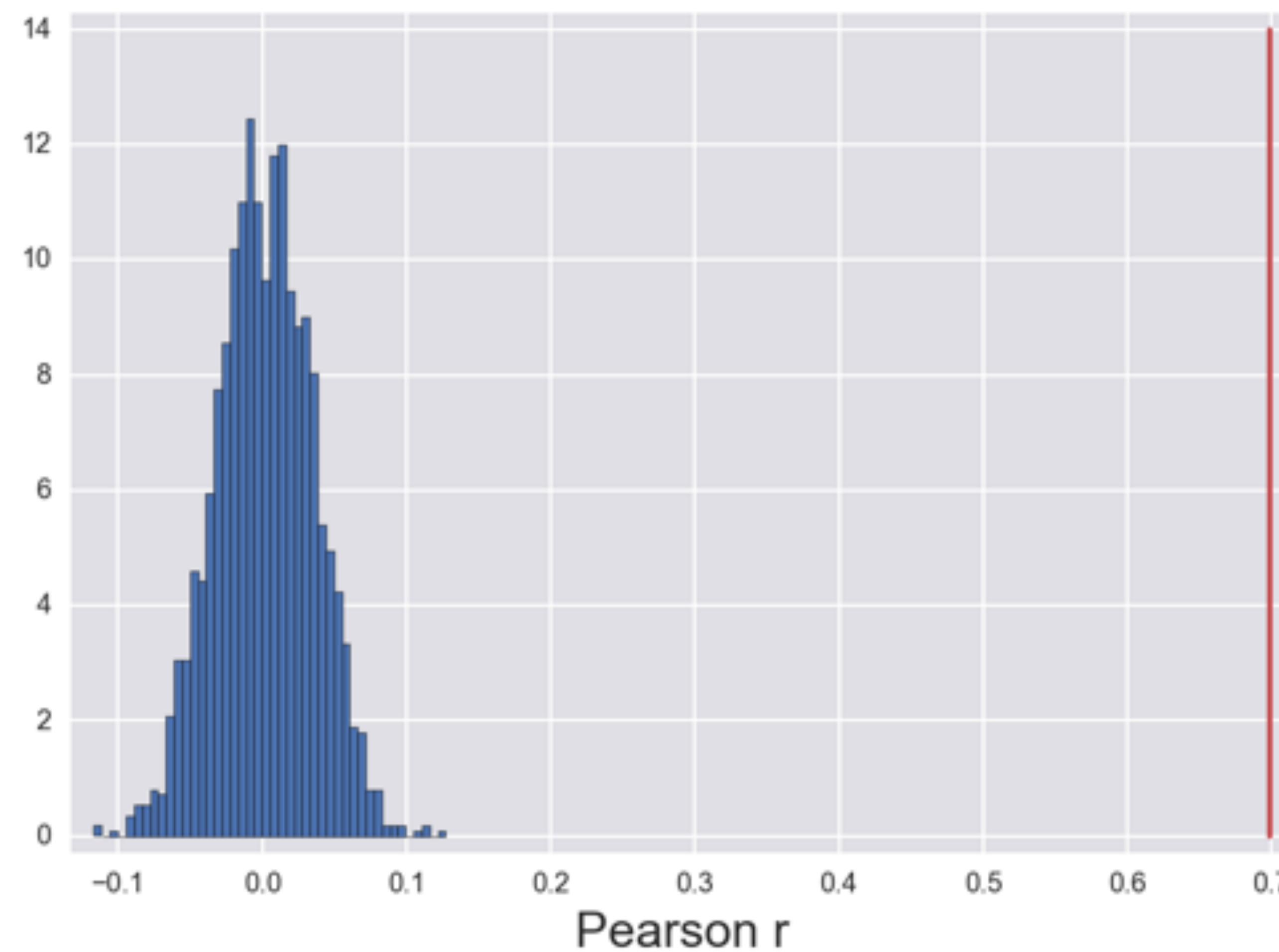




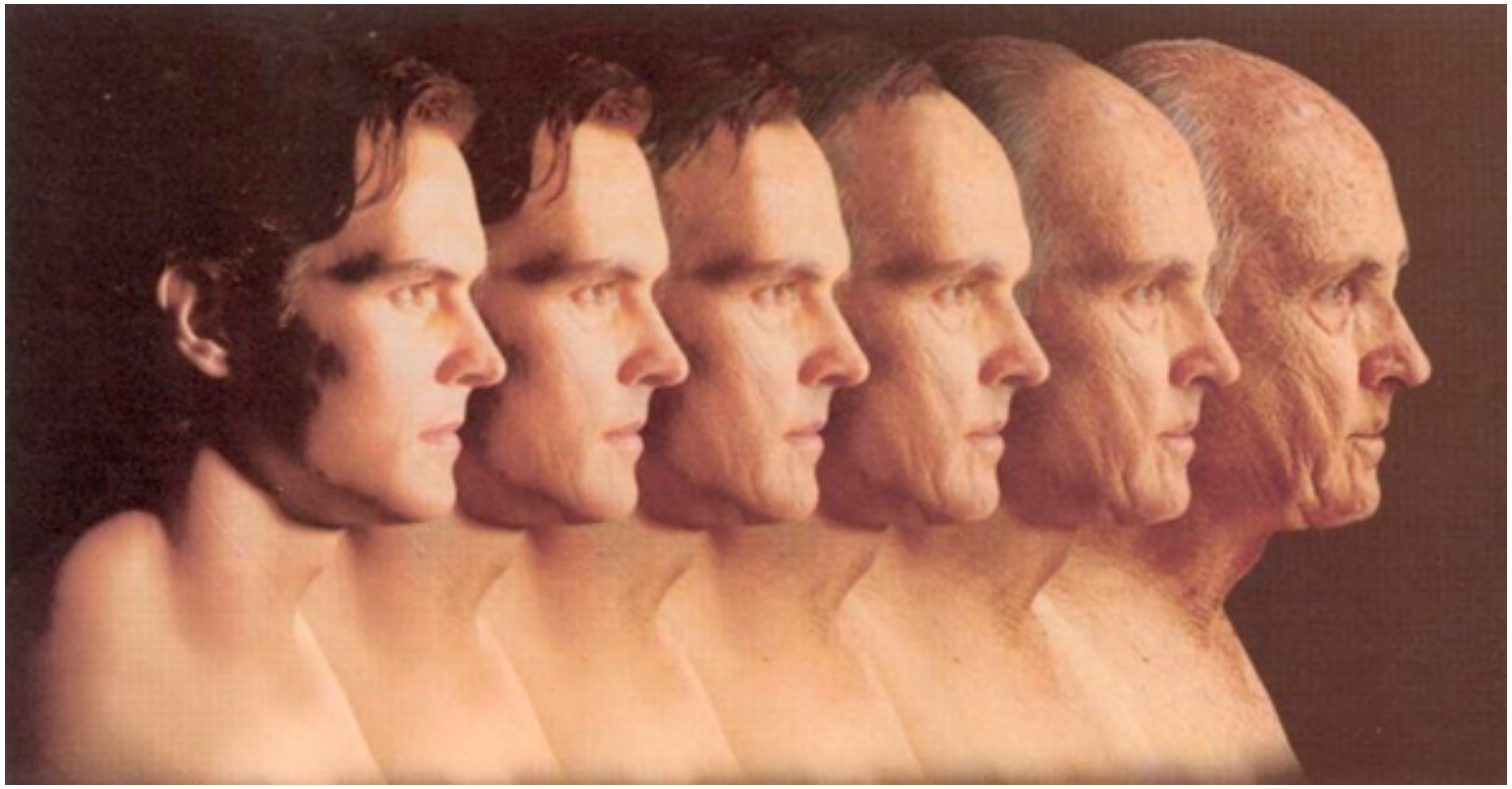


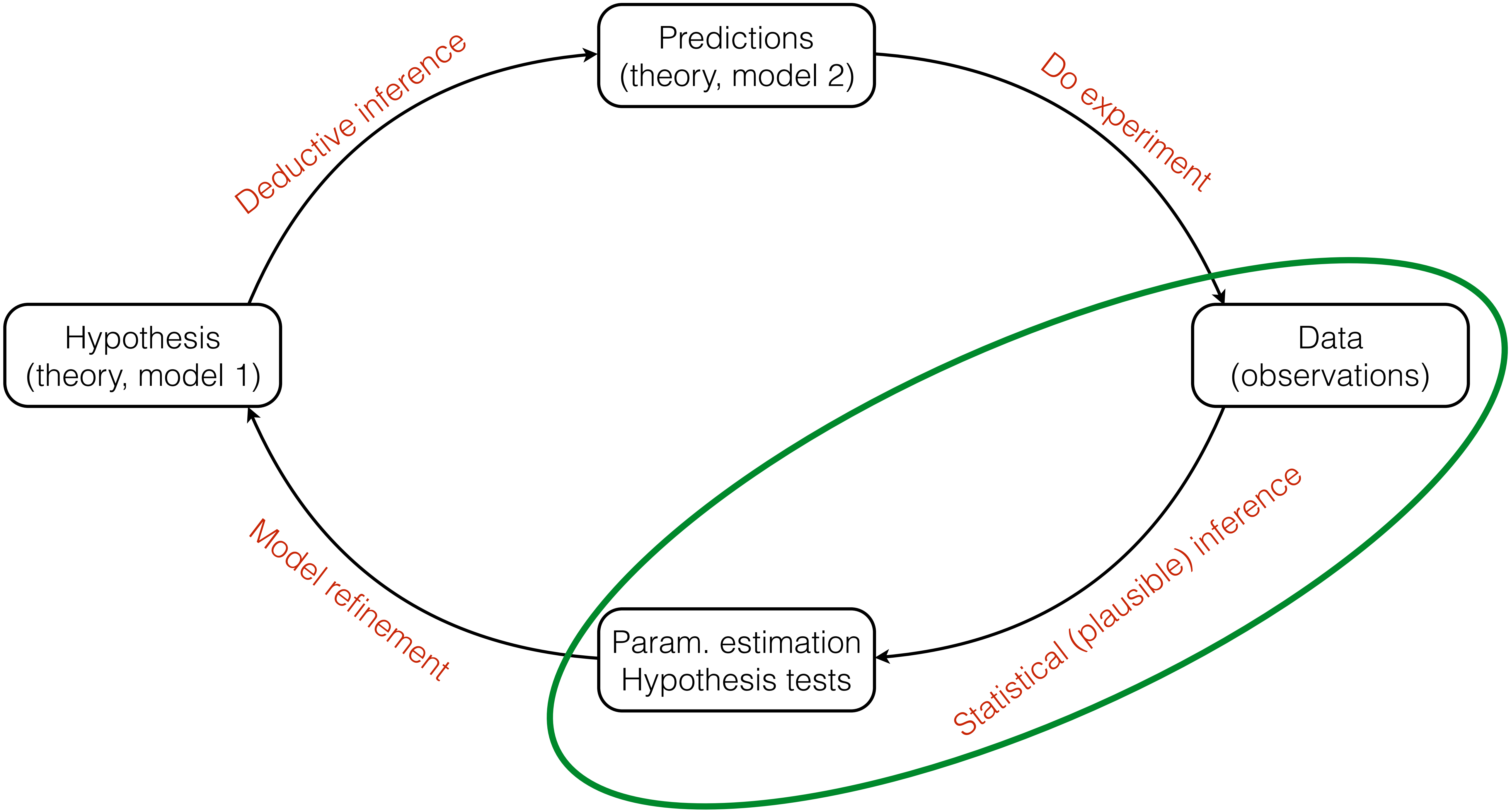
Colocalization






1 Litre
Dairy Farmers
No Fat
Skim
Skim milk that simply lasts longer




Predictions
(theory, model 2)
Deductive inference
Do experiment
Hypothesis
(theory, model 1)
Data
(observations)
Model refinement
Statistical (plausible) inference
Param. estimation
Hypothesis tests