Models

Review

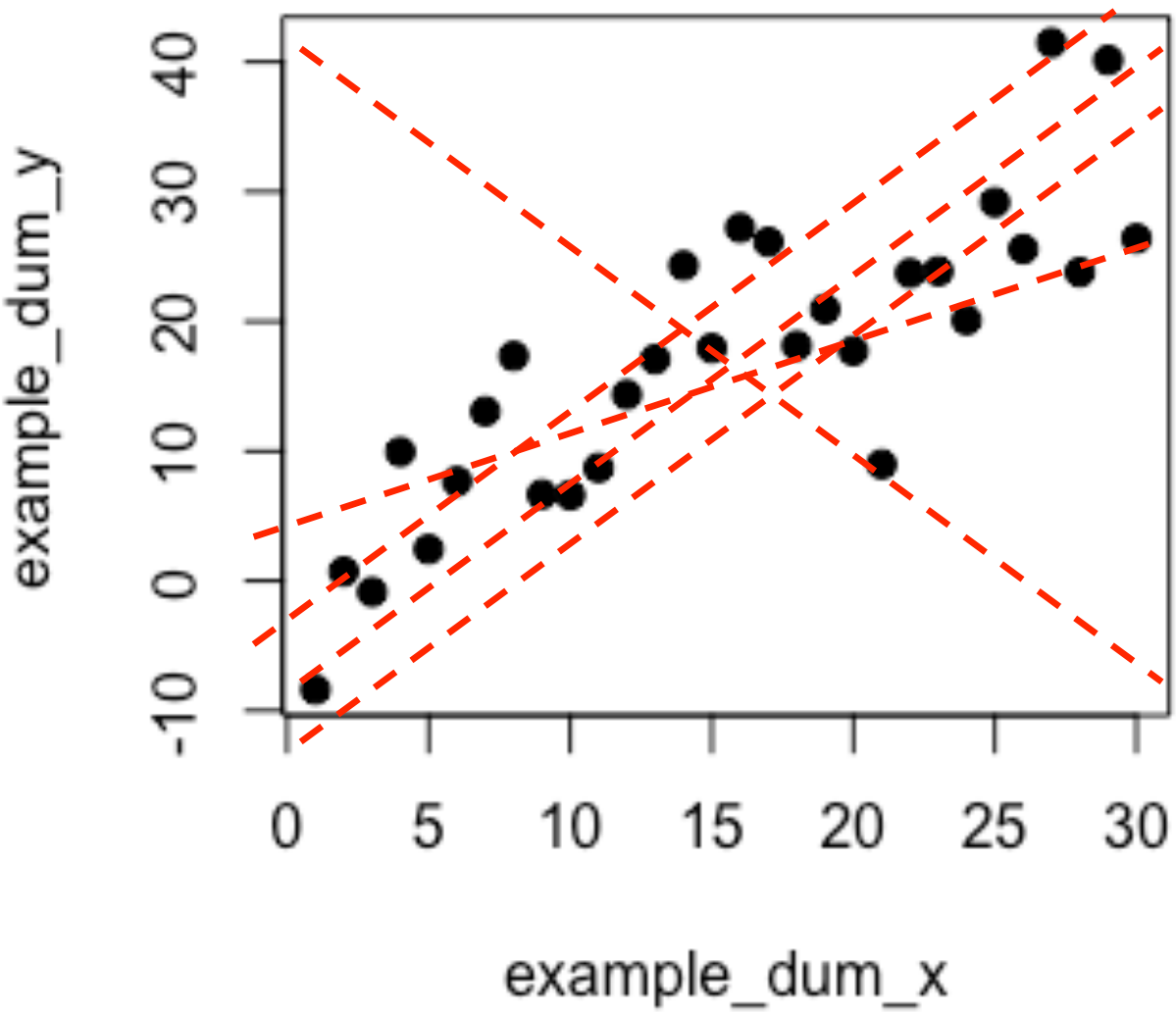
- Chi-squared
- Student’s T
- ANOVA
- Regression
- General linear models

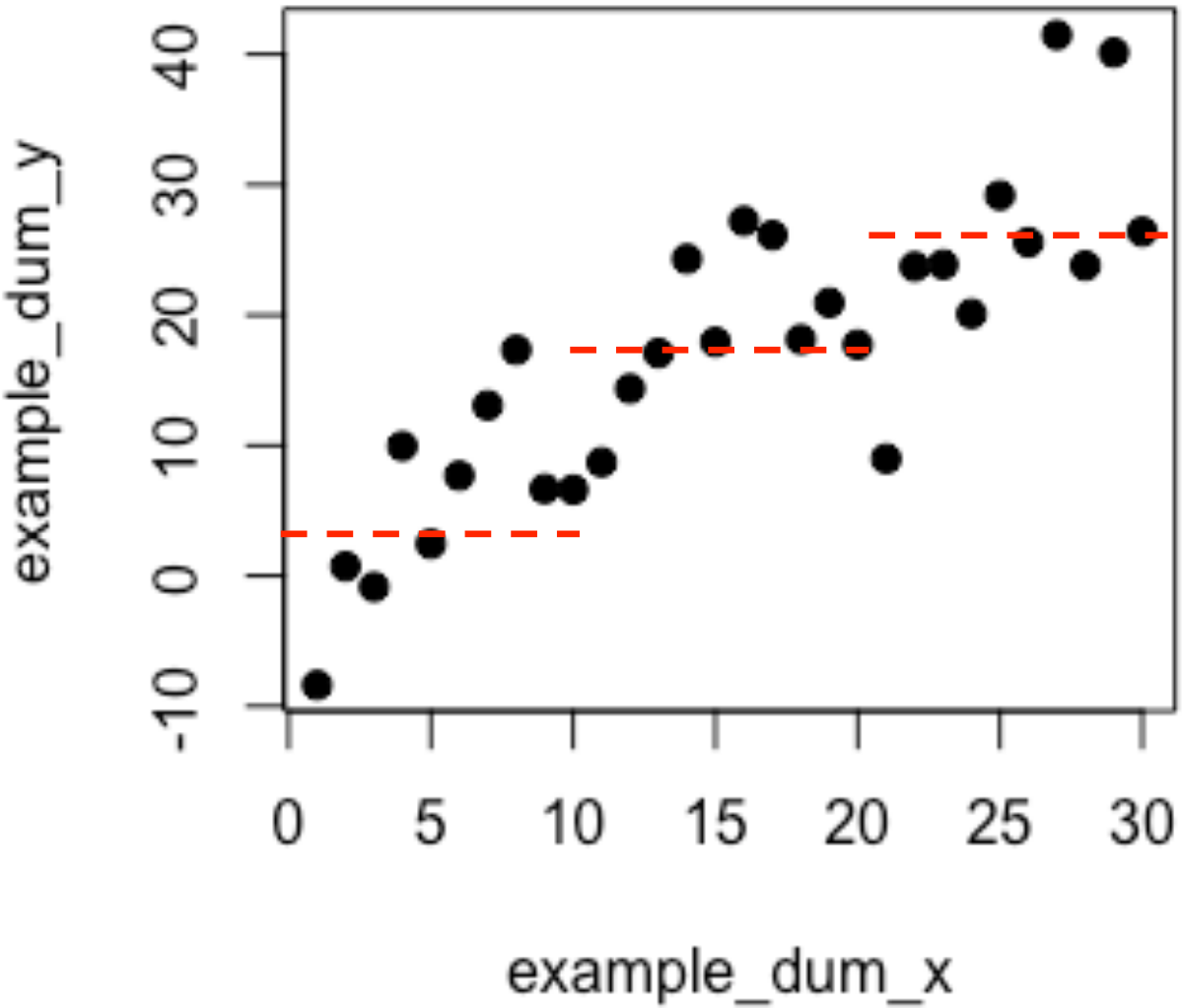
Review

- Chi-squared → evenly spread
- Student's T → one or two means
- ANOVA → multiple means
- Regression → straight line ($y \sim a + bx$)
- General linear models → combinations

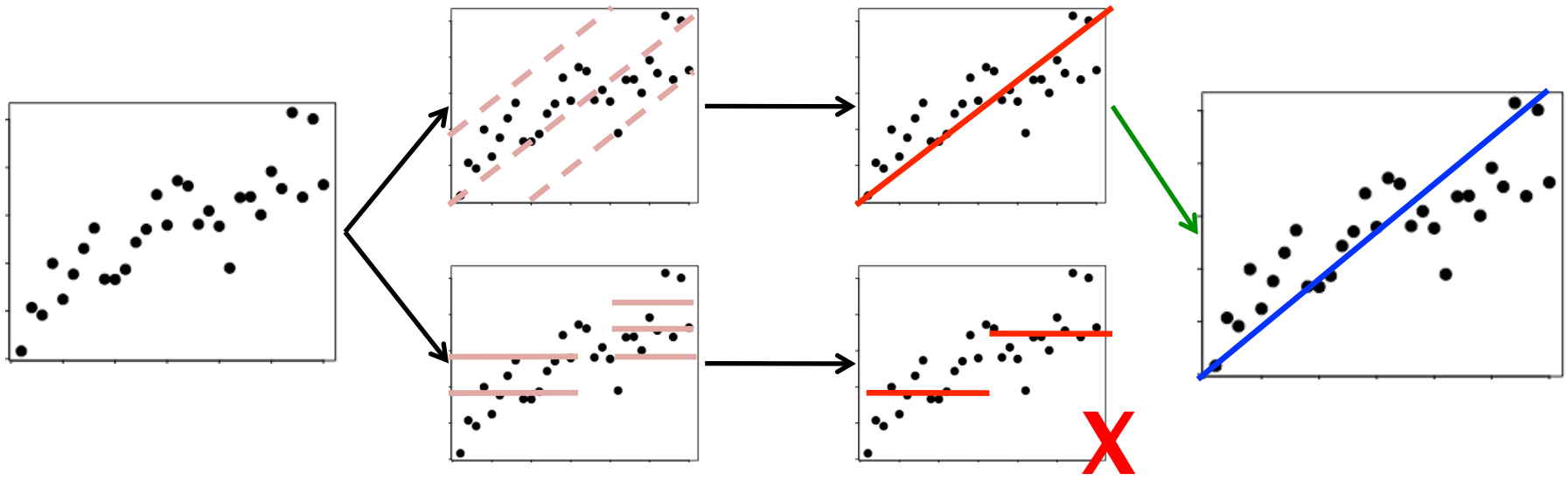
Review

- These are all ***models***
- They reflect our guesses about underlying phenomena, e.g.:
  - Chi-sq (smoking): *Cancer rates equal*
  - T-test (Mile End heights): *rich vs. poor lifestyles*
  - ANOVA (fertiliser): *Fertiliser A supplies Ca^{+}, Fert. B doesn't*
  - Regression (height~rain): *more watering → more growth*
- Which to use?


example_dum_y
-10
0
10
20
30
40
0
5
10
15
20
25
30
example_dum_x

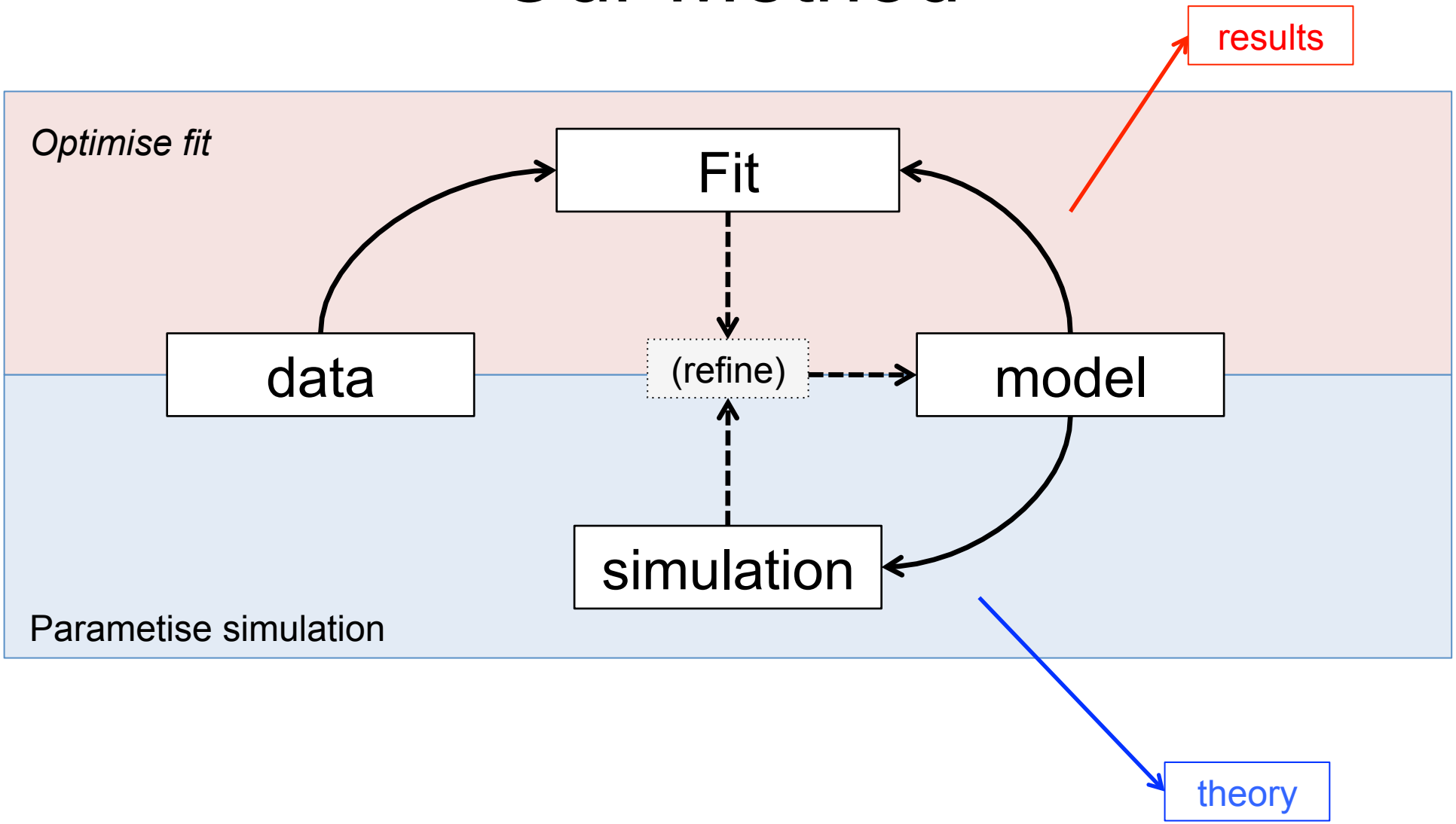

example_dum_y
-10
0
10
20
30
40
0
5
10
15
20
25
30
example_dum_x

Selection and fitting

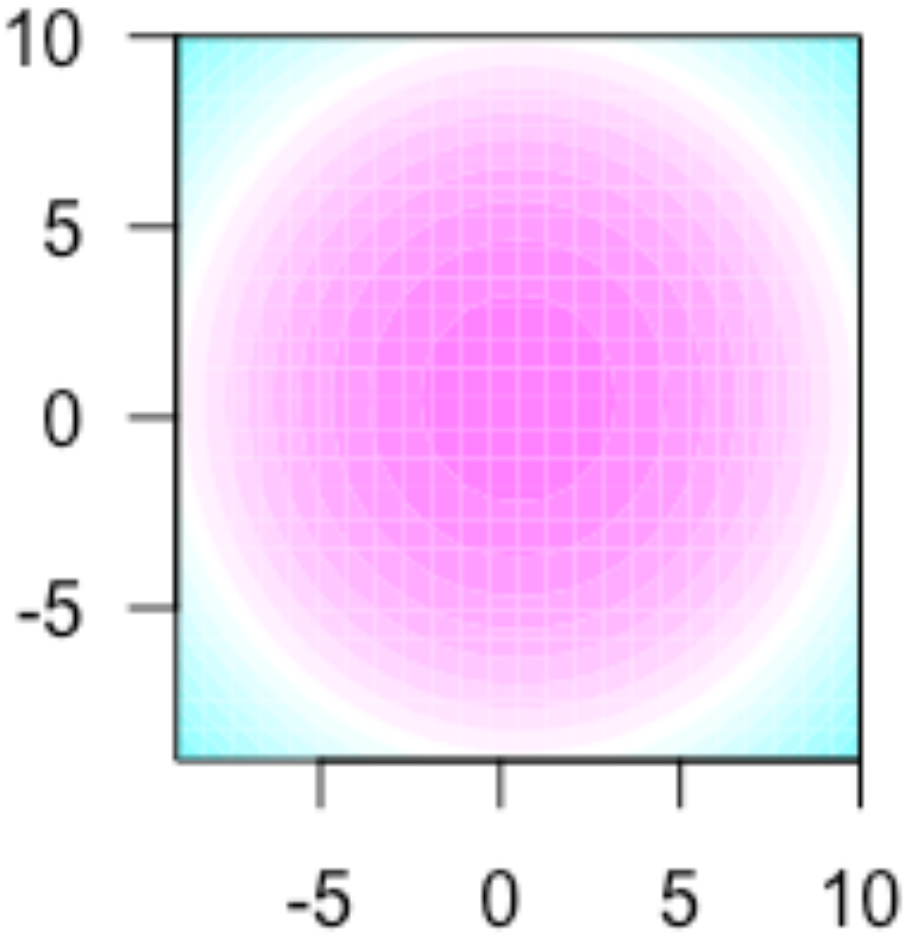


Choose → fit → compare → result

Our Method

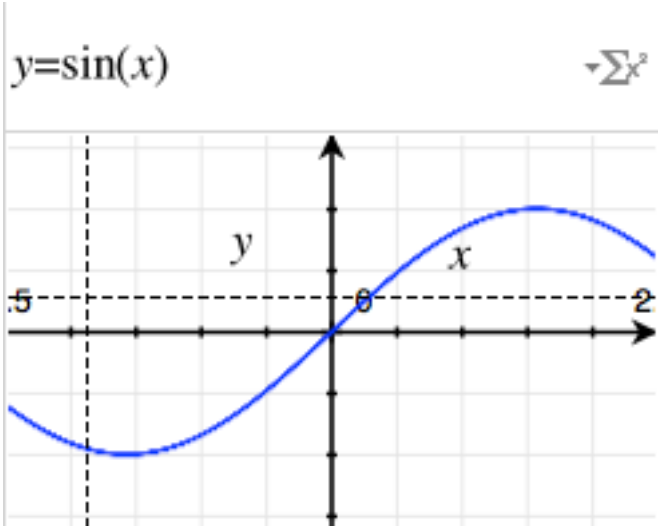

results
Optimise fit
Fit
data
(refine)
model
simulation
Parametise simulation
theory

Other models are available

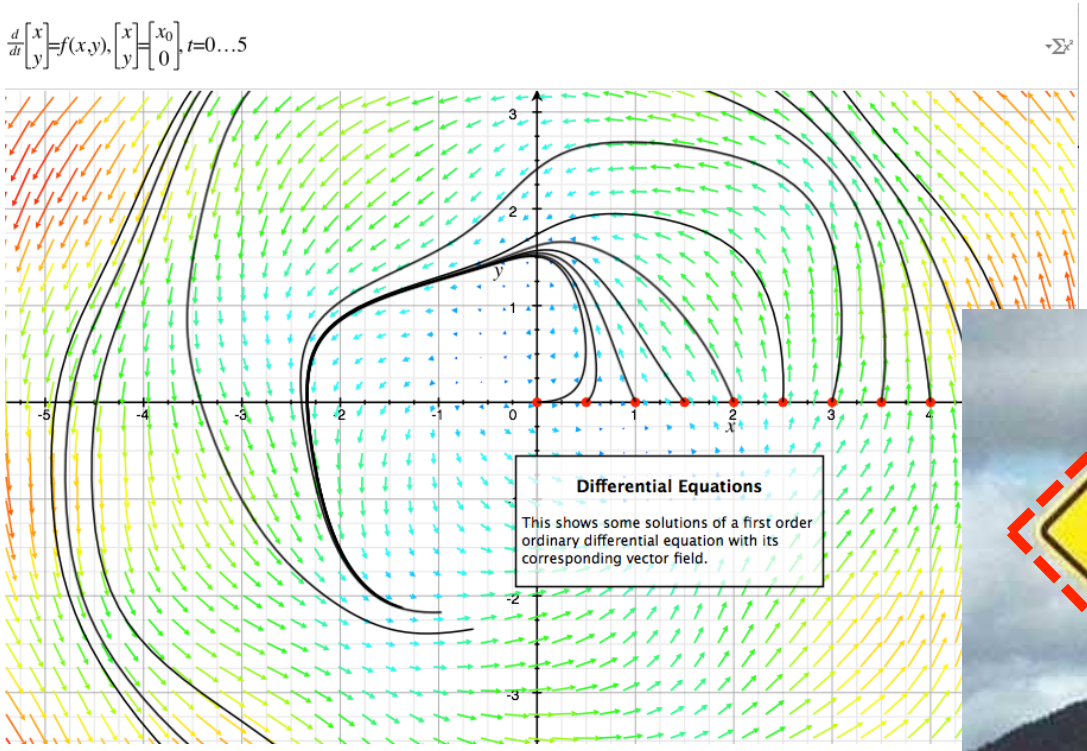


Left: $x^2 + y^2 \leq 1$

Below: $y = \sin(x)$



Other models are available



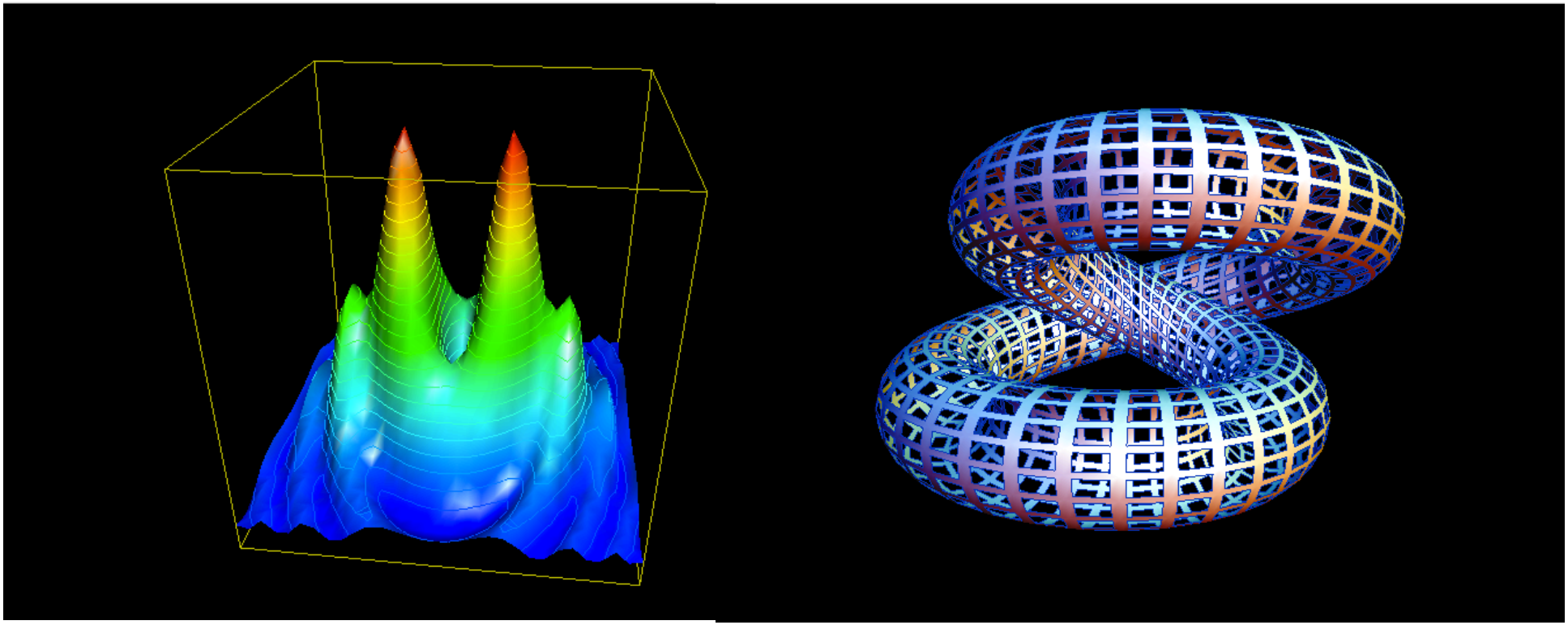
Above: particle motion

Right: regular shapes

Other models are available

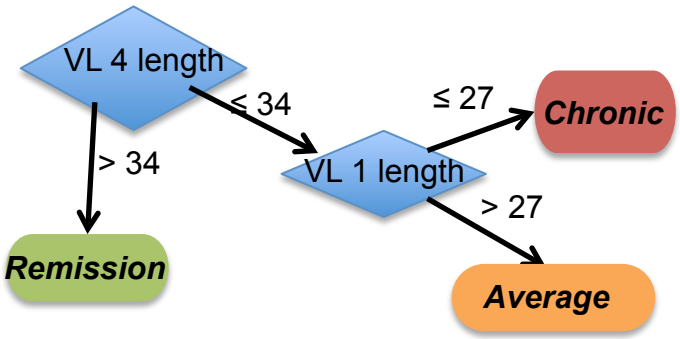
$$=\frac{\sin(x^2+3y^2)}{0.1+r^2}+(x^2+5y^2)\cdot\frac{\exp(1-r^2)}{2},\ r=\sqrt{x^2+y^2}$$

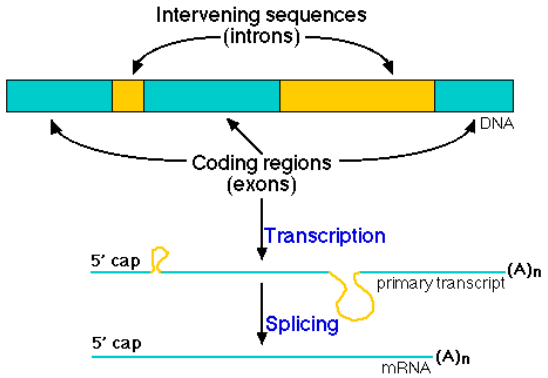
$$\begin{bmatrix} r_0 \\ \theta \\ z \end{bmatrix}=\begin{bmatrix} 3+\sin t+\cos u \\ 2t \\ \sin u+2\cos t \end{bmatrix},\ t=0\ldots 2\pi,\ u=0\ldots 2\pi$$

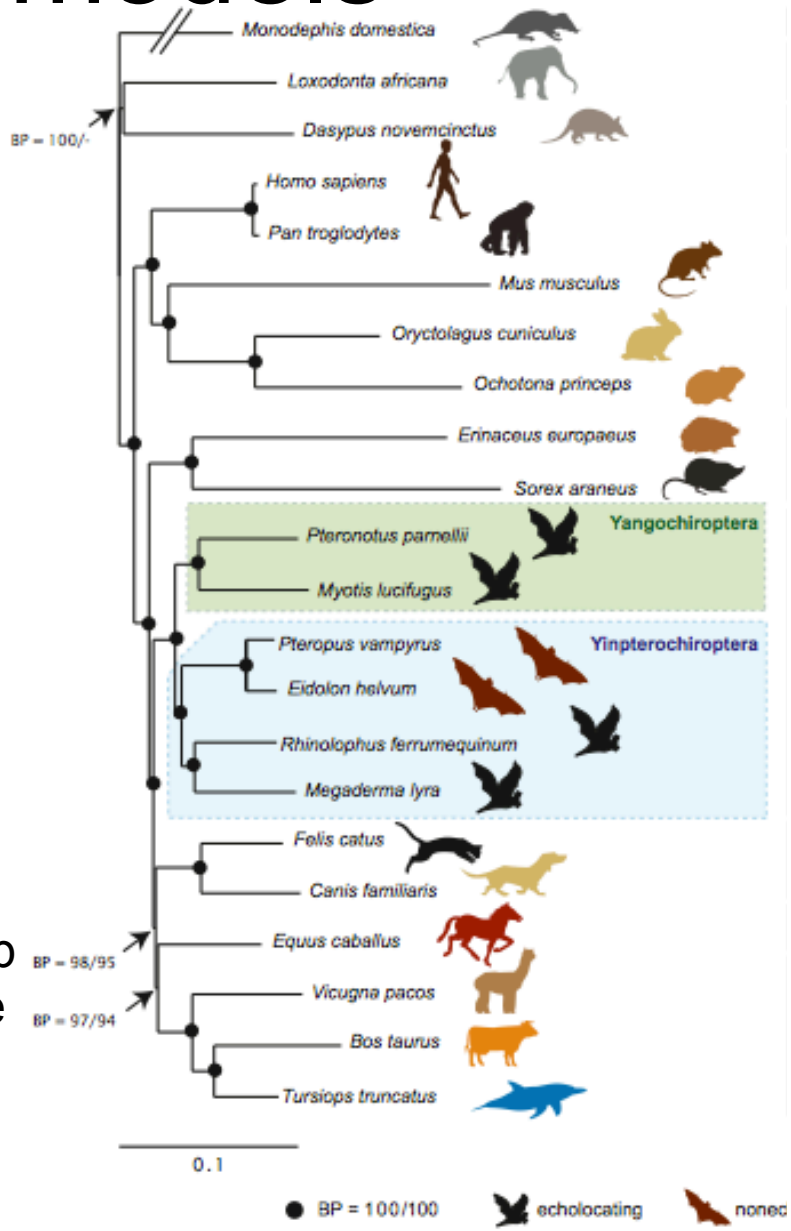


Patterns in 2, 3, 4 or *n*-dimensional space

Biological models







Biological models (*clockwise from top*): Decision model: HIV virus hypervariable loop vs. health; phylogenetic tree: DNA sequence vs. evolutionary history; annotation: DNA sequence vs. transcription role

What is a model?

- A model is any statement that explains structure in the data
- Usually related to our guess about underlying phenomena
- Might describe:
  - Numerical observations, space, time
  - Genome structure, evolutionary relationships
  - Human language, anything else
- Usually expressed mathematically

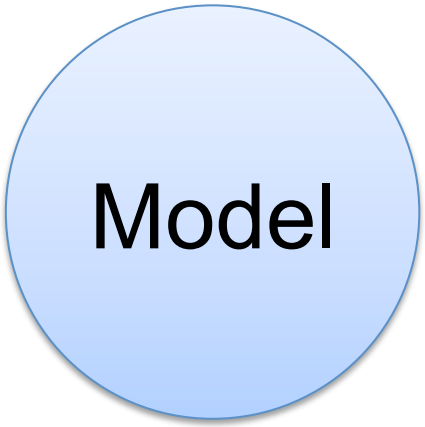
Errors and residuals

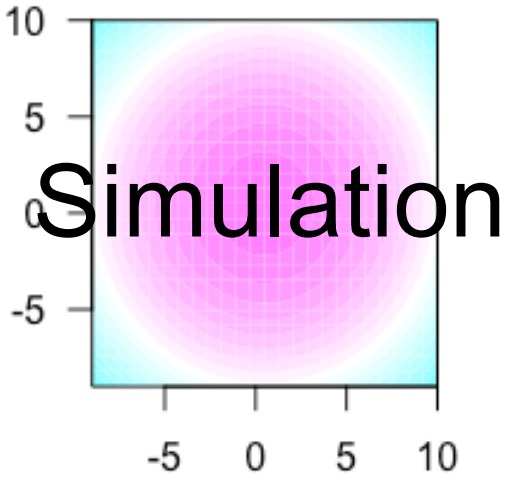
- The real world is noisy
- This is why spotting models is *hard*
- And why **fitting parameters** is *also hard*

Probability *gives us a language to express our belief that a particular sequence of events* (the **data**) *have occurred due to some underlying process* (the **model**), *accounting for random disturbances* (the **error)**

Simulation

- Simulation, modelling, error and inference are intimately linked
- “Does Model A describe x?”
- “Does Model A-simulated data resemble x?”





Parameters

- ‘A’ model really only makes general statements about the nature of the phenomenon
- To make specific, **predictive** statements, we need to **parametise the model**
- E.g. collect and interpret height data:
  - →“Height is normally distributed”
  - →$Height_{MILE\ END} = 1.78m \pm 15.2cm$
  - →95% of people chosen randomly between 1.63m and 1.93m
  - →“Don’t stock XXS or XXL jeans”

Null hypotheses

- The '***null hypothesis***' is usually how we infer and test a model
- It is not special!
- Usually a minimal 'business as usual'-type description of the data
- But for complex data types (genomes; molecular interactions; evolution) even simple null models will still be *very* complicated

Model choice

- Sadly, relationships are rarely clear-cut
- Which model to use?
  - *Distribution: uniform? normal? exponential?*
  - *How many means? 1? 2? k?*
  - *Spatial shape: circle? polygon? irregular?*
  - *Substitution model: equal-rates? codon-bias?*
- Model selection doesn't just influence results; it **is** the results

Model comparison

- *We've met the basic principles*
  - **Efficiency** – we like models that are simple to understand
  - **Tractability** – we like models that are simple to fit/parametise
  - **Significance** – we like models that explain variation
- **Model comparison**: statistically trading competing models off

Model comparison

- Number of d.f. / parameters
- Total variation / variance
- Mean variation
- Test statistics (Chi-sq, T, F)
- Information criteria, odds-ratios (AIC)

(…. and: *Bayesian vs. frequentist traditions….)*

Algorithms

- Some models are straightforward to fit/ parametise, because mathematicians have derived solutions, e.g.:

$$\hat{\alpha} = \bar{y} - \ddot{\beta}\,\bar{x},$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\operatorname{Cov}(x, y)}{\operatorname{Var}(x)}$$

- Usually life is harder
- Especially in many dimensions

Algorithms

We use algorithms for two purposes, then:

1. **Model parametisation:** if we have a model to fit to complex data, we need a scheme to determine model coefficients
2. **Model selection:** even if parametising a given model is trivial, there is an omnishambles of *possible models* to choose from

Search strategies

- Numerical answer (or approximation) easily derived
- Brute-force
- Stepwise / incremental search with gradient descent
- Random search (walk)
- Markov-chain Monte Carlo
- Combinations

Probability and confidence

- *p* – Probability of seeing data, *if* H_0 correct:

$$p = \Pr(D|H_0)$$

- Not perfect but we're stuck with it
- For simple models *p* is straightforward to calculate; e.g. for coin-toss ($P_{heads}=0.5$) a sequence of k heads from n tosses $\{H,T,H,H\}$ k determined from the binomial distribution:
-

$$\Pr(k,n \mid P_{heads} = 0.5) = \binom{n}{k}$$

Probability and confidence

- For other models (normally-distributed, and related) solutions to exact or approximate probabilities are available (T, F etc) allowing us to construct *p*-values and "accept or reject H_0"
- For more complex models (complicated distributions, high dimensions) calculating probability exactly becomes harder or impossible
- *Reminder: Variances, confidence intervals and p all depend on our ability to quantify belief*

Estimating confidence

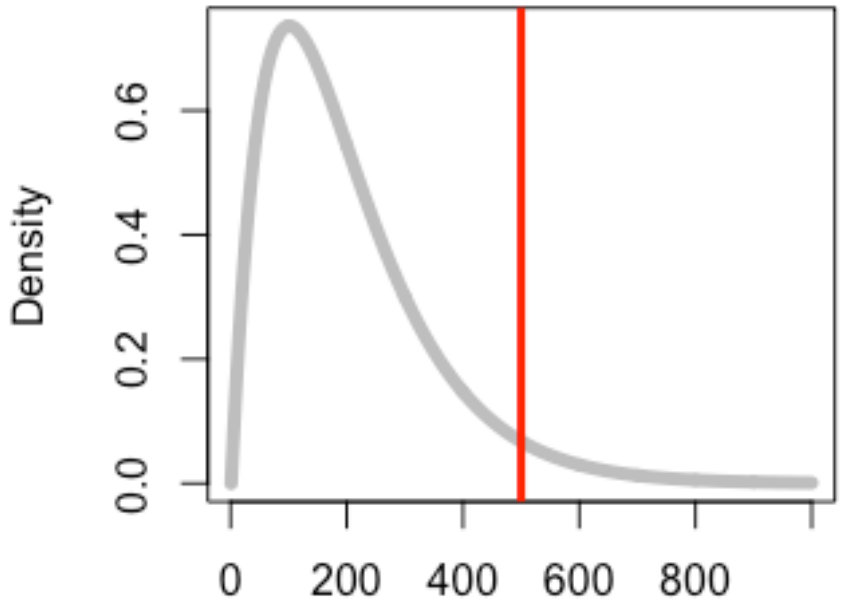
Where we can't calculate directly we may have to try our best:

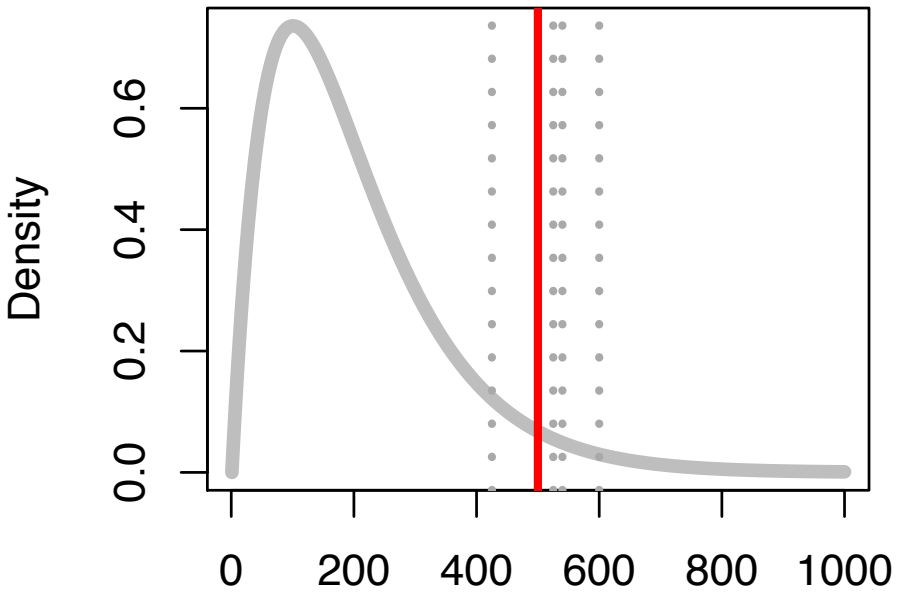
- *How sensitive is this estimate to the data?* (bootstrap estimates)
- *How sensitive is this estimate to the model?* (vary parameters, empirically)
- *How closely does this dataset resemble the null?* (simulations)

Estimating confidence

Where we can't calculate directly we may have to try our best:

- *How sensitive is this estimate to the data?* (bootstrap estimates)
- *How sensitive is this estimate to the model?* (vary parameters, empirically)
- *How closely does this dataset resemble the null?* (simulations)

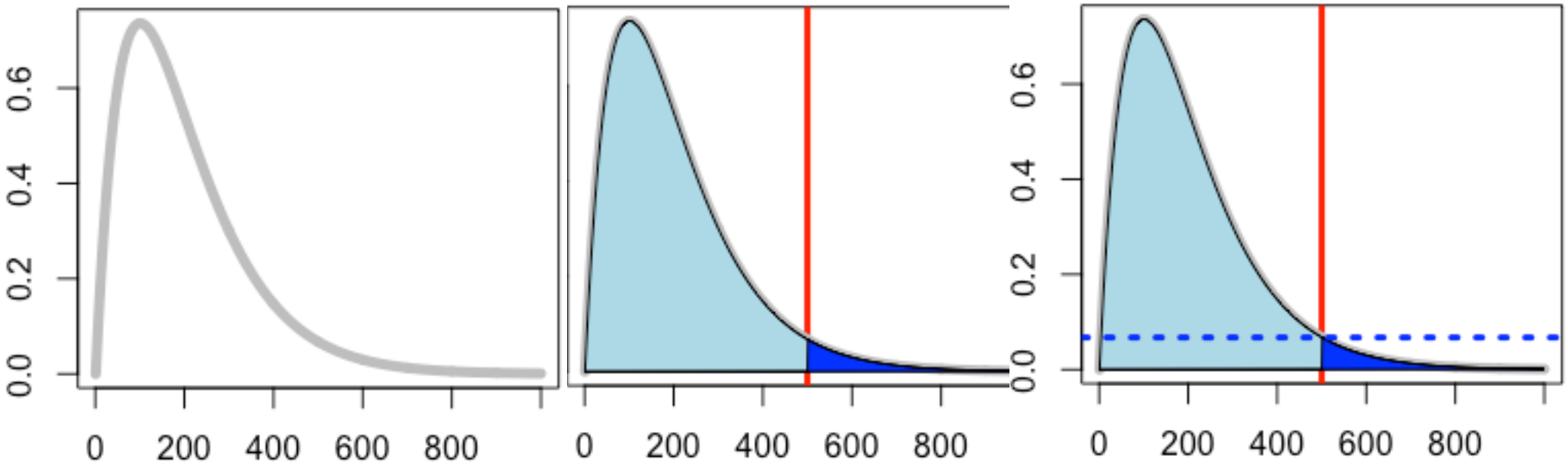




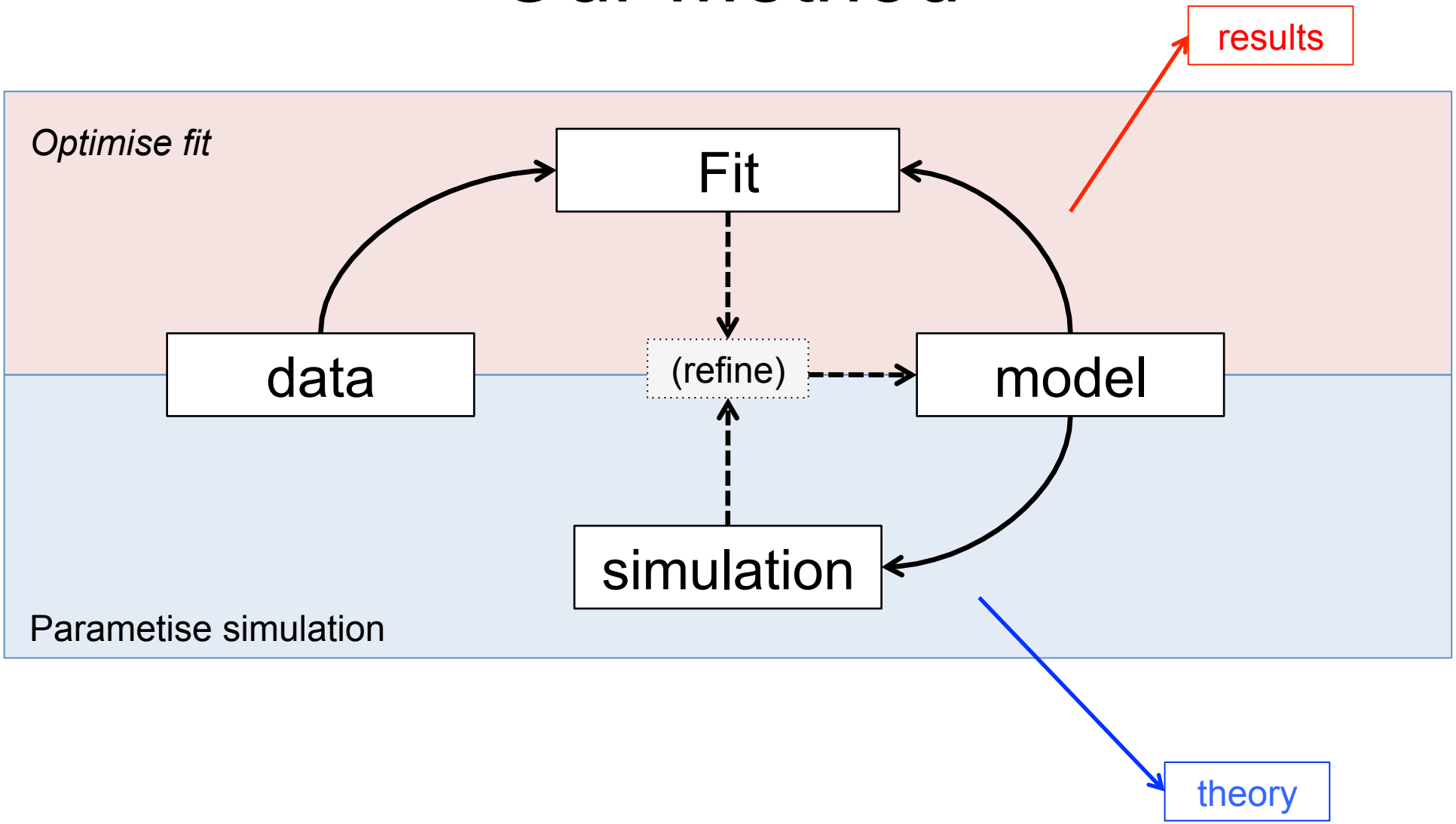
Estimating confidence

Where we can't calculate directly we may have to try our best:

- *How sensitive is this estimate to the data?* (bootstrap estimates)
- *How sensitive is this estimate to the model?* (vary parameters, empirically)
- *How closely does this dataset resemble the null?* (simulations)

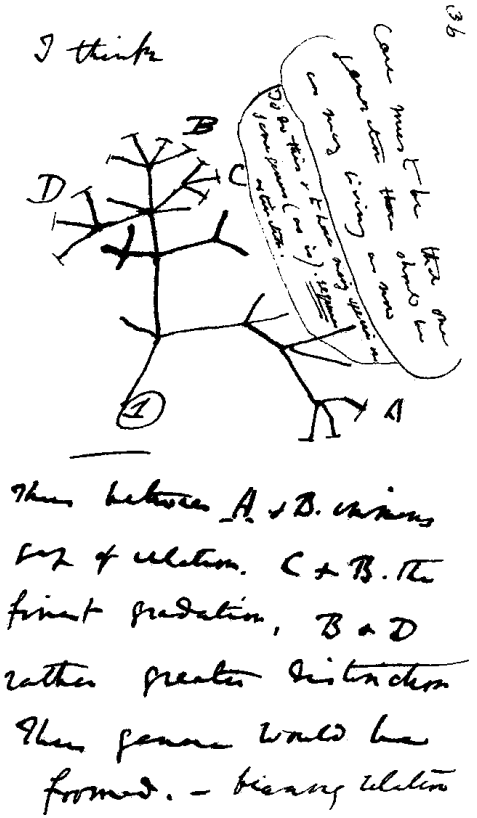


Our Method


results
Optimise fit
Fit
data
(refine)
model
simulation
Parametise simulation
theory

Example: Phylogenetics

- An evolutionary tree is a hypothesis
- e.g., a *model,* of DNA sequence evolution
- Many trees are *possible*
- For neutral sites, approximate as a stochastic process on a *fixed* tree, then compare to successive *proposed* trees (Felsenstein*, 1981)*



Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach

Joseph Felsenstein

Department of Genetics, University of Washington, Seattle, Washington 98195, USA

Summary. The application of maximum likelihood techniques to the estimation of evolutionary trees from nucleic acid sequence data is discussed. A computa-

produced by parsimony methods (Edwards 1963; Edwards and Cavalli-Sforza 1964; Camin and Sokal 1965). These methods implicitly assume that change is improb-

Example: Phylogenetics

- Consider two DNA sequences (A,B) with n orthologous sites, over time t. Assume mutations follow iid. Total will be a product:

$$\Pr(B \mid A,t) = \prod_{k=1}^{n} \Pr(B_k \mid A_k,t)$$

- At each site k, probability that a base i mutates j to is given by:

$$P_{i\to j}(t) = e^{-ut}\delta_{i\to j} + (1-e^{-ut})\pi_j$$

where: u denotes substitution rate ('speed'); delta the change function; pi the equilibrium frequency

After Felsenstein (1981)

Example: Phylogenetics

Felsenstein (1981)

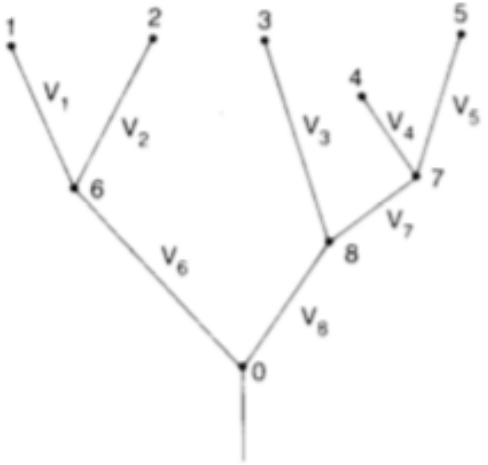


Fig. 1. The tree used in the discussion of computing the likelihood. The v's are the lengths of the segments

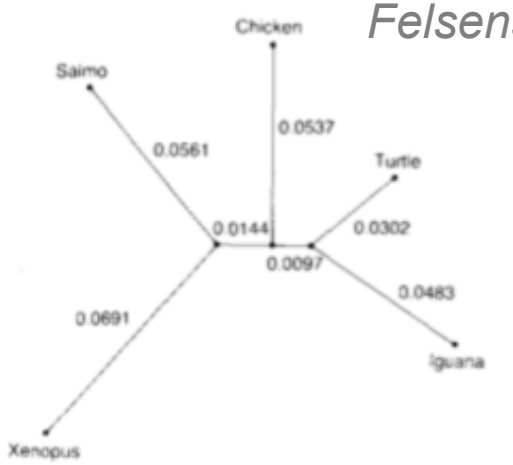


Fig. 4. The maximum likelihood estimate of the phylogeny for 5S RNA sequences from five vertebrate species

- Simply divide the tree up into segments
- Sequences for unobserved (ancestral) taxa are unknown – so we compute for all possible states
- Sum of possible states: site likelihood
- Product of site likelihoods: tree likelihood (or sum, in lnL)
- Simple algorithm:
  1. For starting tree, parametise branches to optimise (maximise) likelihood
  2. Propose new trees and accept if they improve on the current one

Summary

- Statistics is the proposal and fitting of models to explain data
- Some models are simple, others complex. Some, we might invent ourselves
- Models must be parametised, or fitted, to be useful
- Null hypotheses are just models, too
- With infinitely many possible models to choose from, they have to be compared to give the ‘best’ for our data
- Goodness-of-fit and likelihood cannot always be computed exactly – we may need to approximate, measure as ratios, or guess