Data and models

Tanjona Ramiadantsoa

E^2M^2, ValBio Ranomafana

14 Jan 2018

With materials from:
- Steve Bellan, University of Austin, Texas
- Cara Brook, University of California, Berkley

Goals for this lecture

- To encourage you to participate, interact, and ask questions
- To acknowledge there are many types of models
- To differentiate between statistical model and mechanistic (mathematical) model

Data and models in the syllabus

- Sunday:
  - **1:00-2:00pm** Lecture: *Exploring & visualizing data in R* **(Christian + Cara)**
  - **4:30-5:30pm:** Tutorial: *Basic statistical modeling in R* **(Andres)**
- Monday:
  - **8:30am-9:30am:** Lecture: *Introduction to mixed modeling* **(Andres)**
  - **1:00-2:30pm:** Lecture w/Tutorial: *Introduction to occupancy modeling* (**Fidy**)
- Tuesday
  - **8:30-10:00am**: Lecture: *Introduction to Compartmental Models and Differential Equations* **(Jess)**
  - **10:30am-12:00pm:** Tutorial: *Building Mechanistic Models in R* **(Jess)**
- Wednesday:
  - **8:30-9:30am:** Lecture w/Tutorial: *Model Fitting in Practice – the Basic Concept* **(Cara)**
- Thursday:
  - **10:00-11:30am:** Lecture w/Tutorial: *Introduction to Network Modeling* (**Fidy**)
  - **1:00pm-2:00pm**: Lecture: *Introduction to Spatial Modeling* (**Amy**)

….

Outline

- Data or not data
- Statistical model
- Mechanistic (mathematical) model
- Combining mechanistic and statistical model
- R

Data is the backbone of science

- Data serve as evidences to support a claim
- Models are used to explain the data, and “predict”

Data or not data?

- 19
- 19: total number of fingers and toes
- 19: total number of fingers and toes of Brian

- 5, 14, 21
- 5, 14, 21: the number of children of Cara, Jess, and Fidy, respectively.
- Cara, Jess, and Fidy are the name of three tenrecs at the Duke Lemur Center

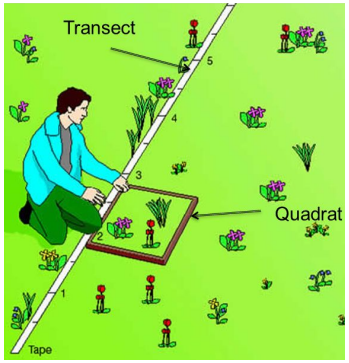
Data: general structure

- Data contain **a relationship between at least two variables**: x and y
  - x: explanatory, control, driver, independent variable(s)
  - y: response, dependent variable(s)
- x and y should be clearly defined (with respect to the question)
  - E.g.: 19: total number of fingers and toes

Data: sources of x and y

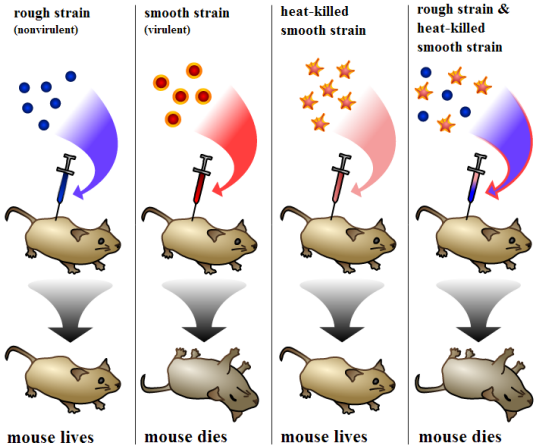
Observational

- Just measure x and y



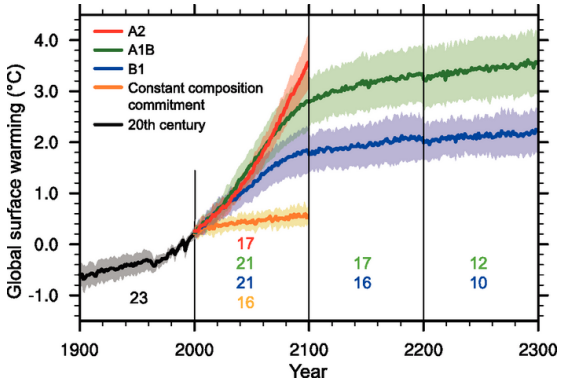
Experimental

- Interfere with x or the relationship between x and y



Simulated

- Create a relationship between x and y



Empirical data

Data: types

Numerical

- A variable is numerical when you can transform it with mathematical operation
- Examples?
- Integer, real number, multi-dimensional number

Categorical

- A variable is categorical when it is not numerical but a categorical can be numerical?
- Examples?
- Colors, (blood) types, species name

Can we always represent data in a table?

Things to consider

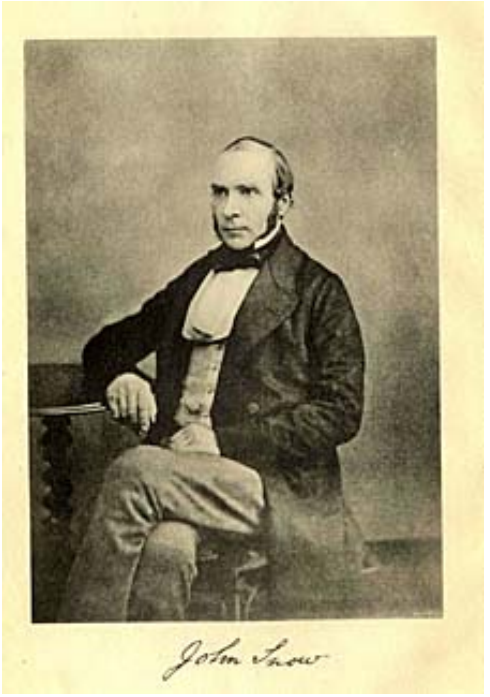
- Data acquisition
  - Impossible, example?
  - Theoretically possible but practically unfeasible, examples?
- Data quality and quantity
  - In practice there is always a trade-off
    - Example: monetary cost, human effort -> power analysis, sampling design etc.
- Reproducibility
- Measurement errors
  - Examples?

…

Visualization

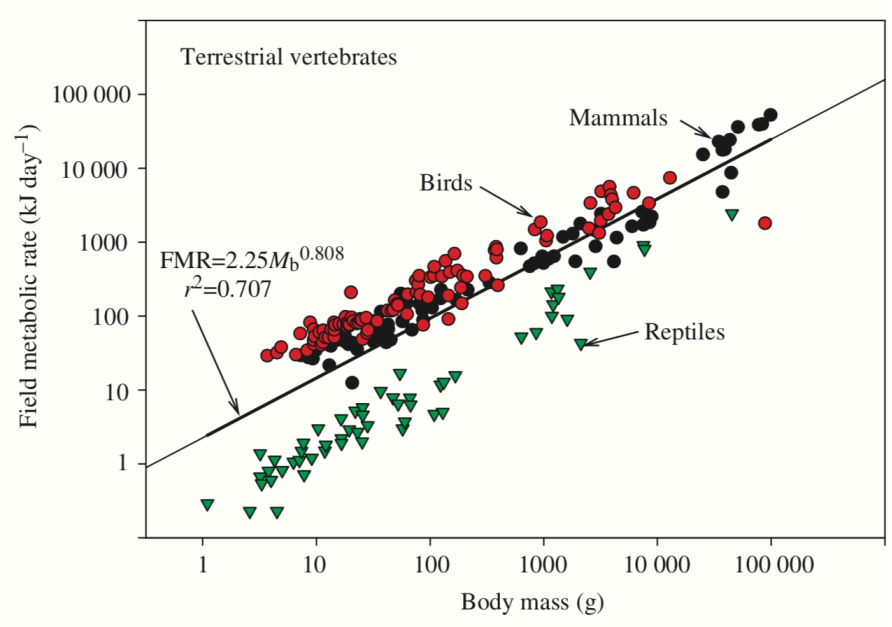
- Good visualization not only allows to clearly show the results but can reveal the answer

- Cholera outbreak in London 1847-1854

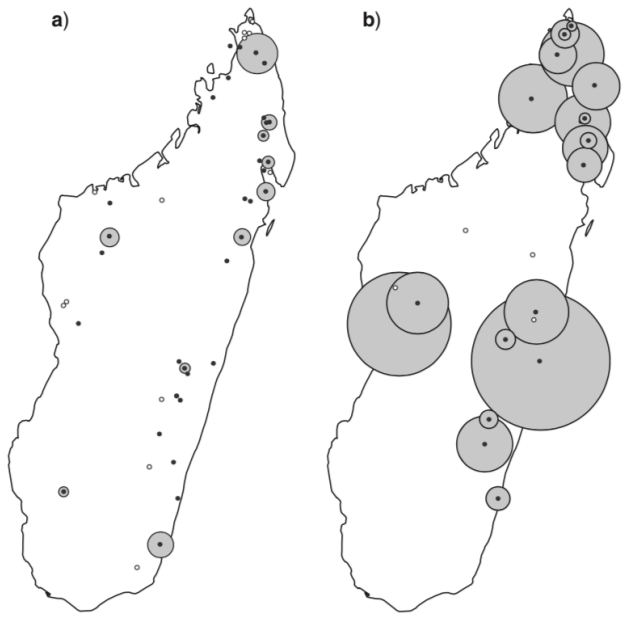


Visualization: examples

Scatter plot



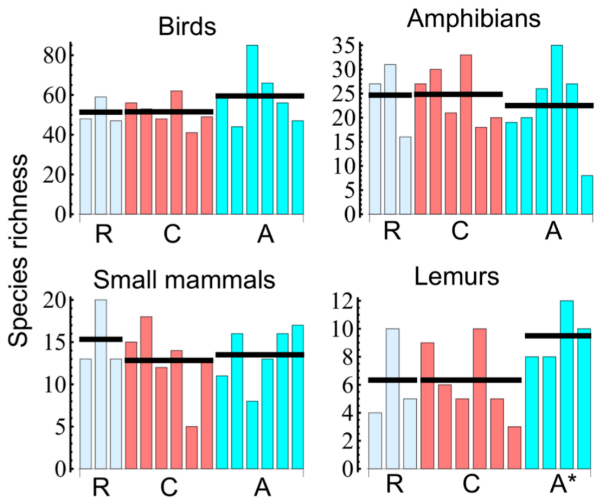
Map



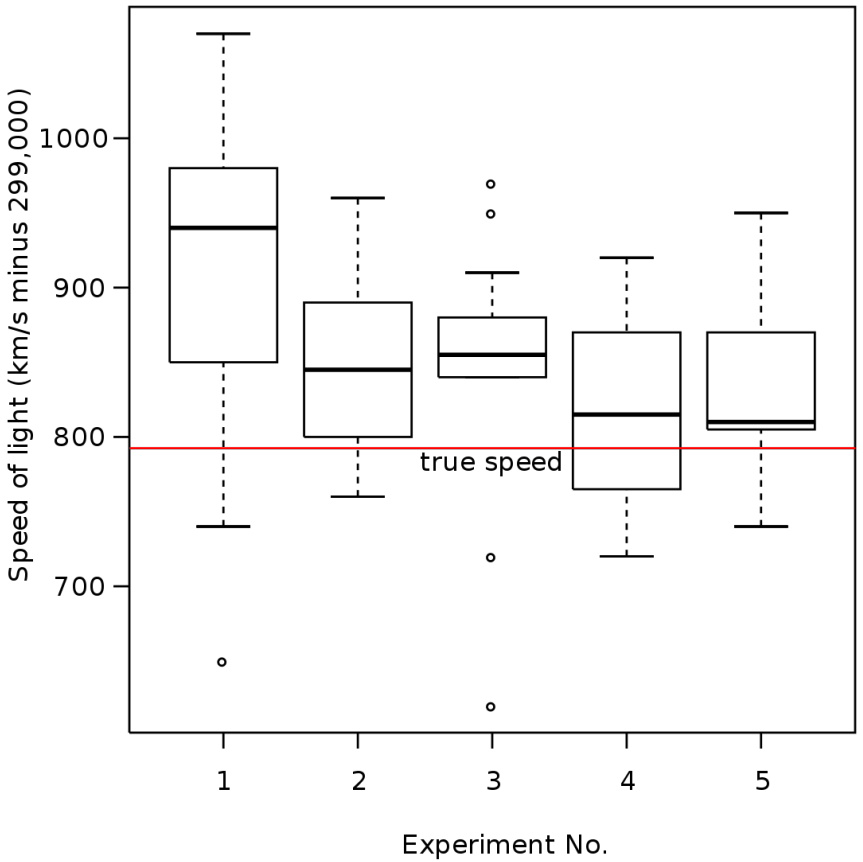
Biomass of dung beetles

Visualization: continued

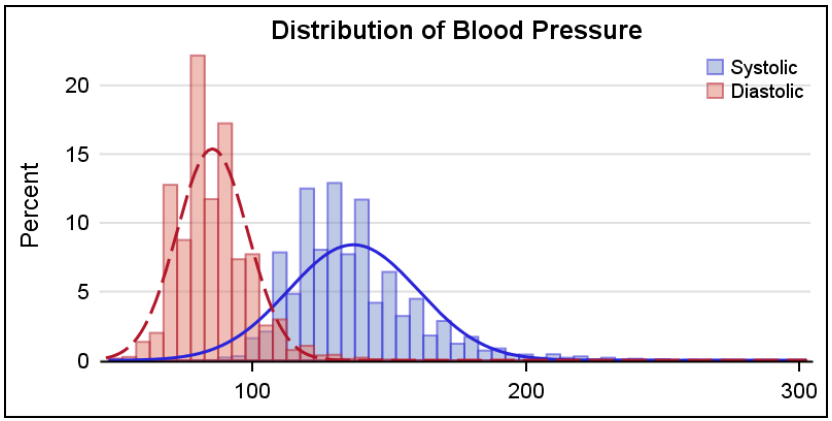
Bar chart



Box plot



Histogram



Visualization with R

- **Sun, Jan 14: "*Getting Started With Data*"**
- **6:30-8:00am:** Breakfast
- **8:00-8:30am:** Road Map and Daily Agenda **(Cara)**
- **8:30-9:30am:** Lecture: *Models and Data* **(Tanjona)**
- **9:30-10:30am:** Software installation and catch-up.
    - Mentors + instructors make sure all students have the proper materials installed and work through 4 tutorials with them
- **10:30am – 11:00am**: Break
- **11:00am-12:00pm:** 1-min student introductions and research presentations **(Cara)**
- **12:00-1:00pm:** Lunch
- **1:00-2:00pm** Lecture: *Exploring & visualizing data in R* **(Christian + Cara)**
- **2:00-3:00pm:** Tutorial: *Exploring & visualizing data in R* **(Christian + Cara)**
- **3:00-3:30pm:** Break
- **3:30pm-4:30pm:** Lecture: *Linear regression and simple statistics* **(Andres)**
- **4:30-5:30pm:** Tutorial: *Basic statistical modeling in R* **(Andres)**
- **5:30-6:30pm:** Free time
- **6:30-7:30pm:** Dinner

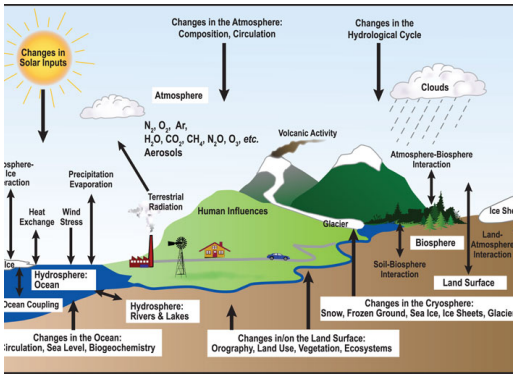
Models & modeling & modelers

A model is a simplified or/and an idealized version of reality

Human

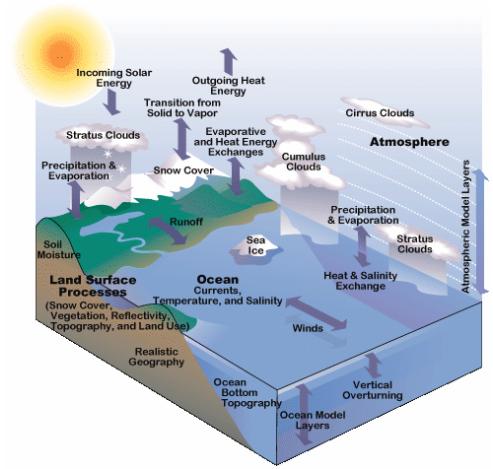
Car

Ecosystem

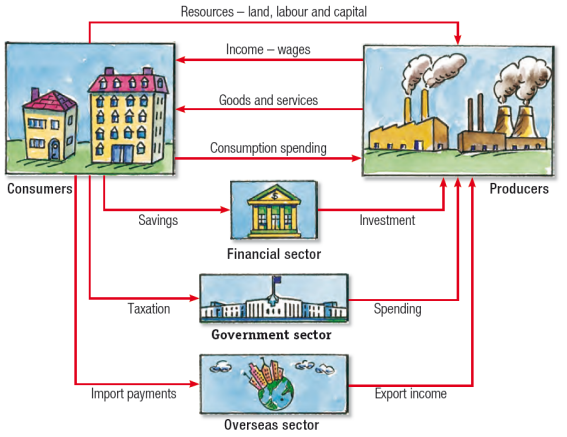


Ecology & Evolution

Climate



Economy



During this workshop

- Learn to **use and build models** to scientifically understand/infer the relationship between explanatory variable(s) (x) and response variable(s) (y) based on ecological or epidemiological questions
- Define x and y concisely
  - **Monday: 11:00am-12:00pm:** Writing Exercise: Formulating research questions (HW) **(Cara)**
- Distinguish between statistical and mechanistic model

Statistical modeling

Statistical model: correlation

- **Need data!!!**
- Models
  - t-test, Chi-square, ANOVA…
  - “Ordination families (PCA, CA, NMDS…)”
  - Regression families (LM, GLM, GLMM, GAM, NDLM (Marius)…)
  - Species distribution model and families (MaxEnt…)
  - …
- Some other classifications
  - Parametric vs. non-parametric
  - Frequentist vs. Bayesian
  - …

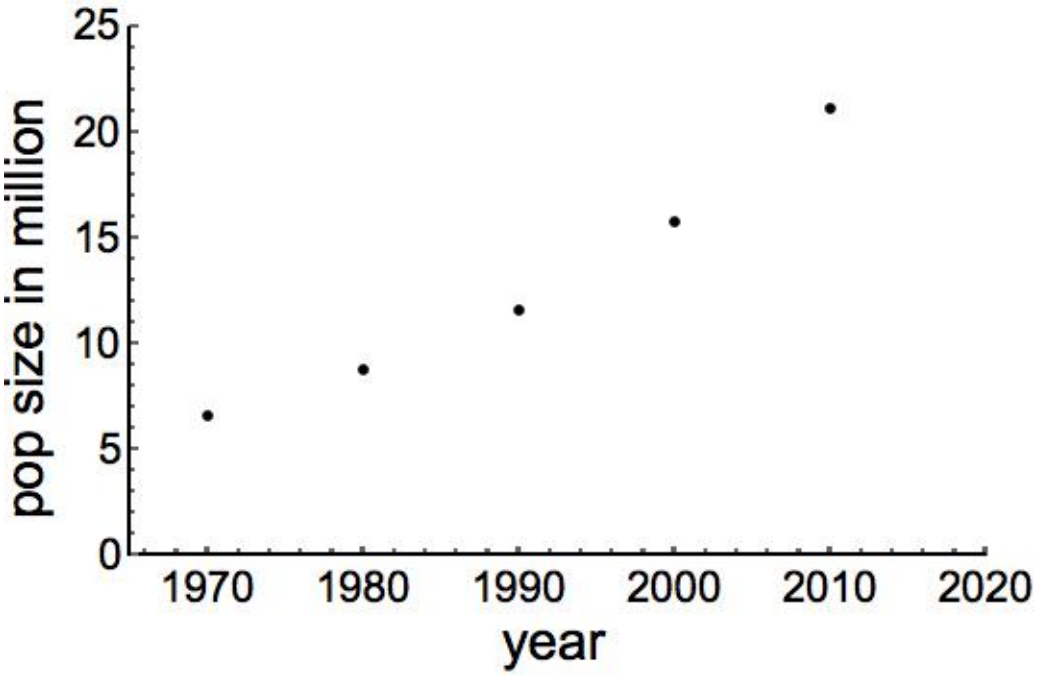
Philosophy

- To rigorously assess the **type** and **strength** of the relationship between x and y
  - Find a significant relationship and p-value mania
- Steps:
  - formulate a research question
  - make a null hypothesis (H0) and alternative hypothesis (H1)
  - obtain data
  - assess H0 with an appropriate statistical analysis (p-hacking!!!)

Concrete example

Questions: How does Malagasy population change through time?

Dynamics of Malagasy population



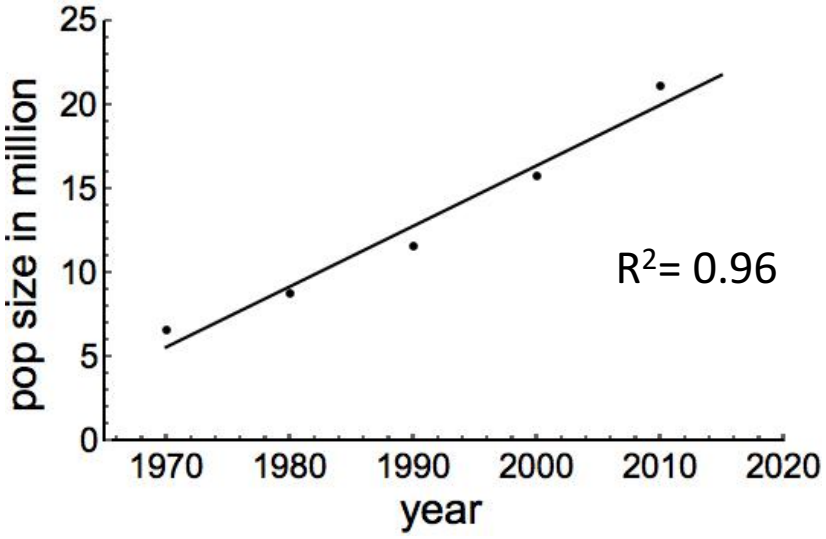
Source: world bank

Statistical model

Linear regression

$$pop = a * year + b$$

$$a = 0.024;\ p < 0.005$$



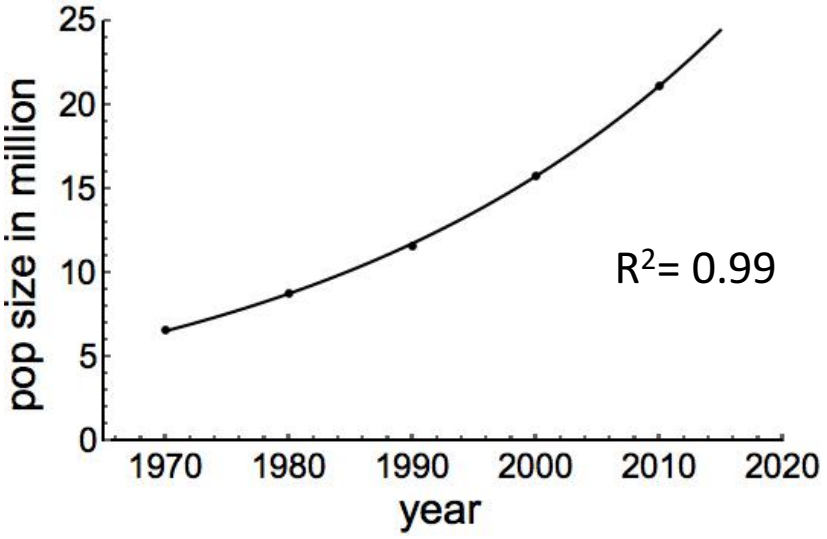
(in R: $pop \sim year$)

Conclusions?

Nonlinear regression

$$pop = e^{a * year} * b$$

$$a = 0.029;\ p < 10^{-6}$$



(in R:?)

Conclusions?

A final point on statistical modeling

- Pros and cons: to be determined
- Statistical models/test are based on specific assumptions (e.g., data should be normally distributed), assessing a model means you need to make sure **the assumptions are not violated**.
- There are so many statistical models…

Mechanistic modeling

Mechanistic model: causation

- **Create models to generate data!!!**
- Model types
  - Equation: Hardy-Weinberg equilibrium
  - Difference equations: Ricker, Beverton-Holt, **Logistic**..
  - Differential equations: Lotka-Voltera, Logistic…
  - Integro-differential equations
  - Individual-Based Model (IBM)
- Classifications
  - Deterministic vs. stochastic
  - Non-spatial vs. spatial
  - Population vs. community

Philosophy

- Think about the mechanisms (processes) that link x and y
- The simpler, the better

Concrete example

Questions: When does a population go extinct?

- Verbal assumptions

$$pop\ next\ year = pop\ this\ year + birth\ - death$$

- Compartment representation

- Mathematical translation

$$Y_{t+1} = Y_t + b * Y_t\ - d * Y_t = (1 + b\ - d) * Y_t$$

$$Y_t = Y_0 * (1 + b\ - d\ )^t$$

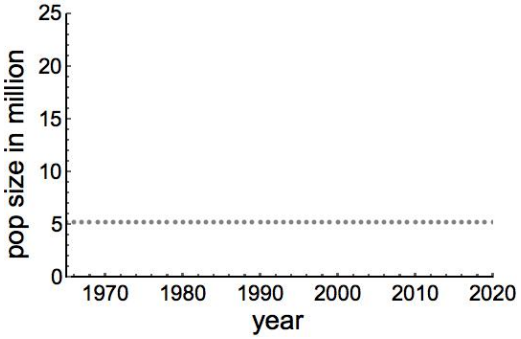
Concrete example

- Mathematical analysis and data generation

$$Y_t = Y_0 * (1 + b - d)^t$$

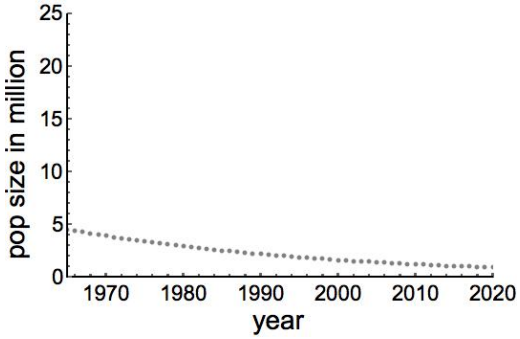
$b - d$: annual growth rate (r)

$b - d = 0$



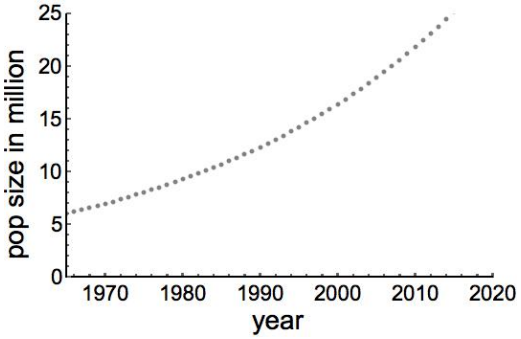
Negative

$b - d = -0.02 < 0$



Positive

$b - d = 0.02 > 0$

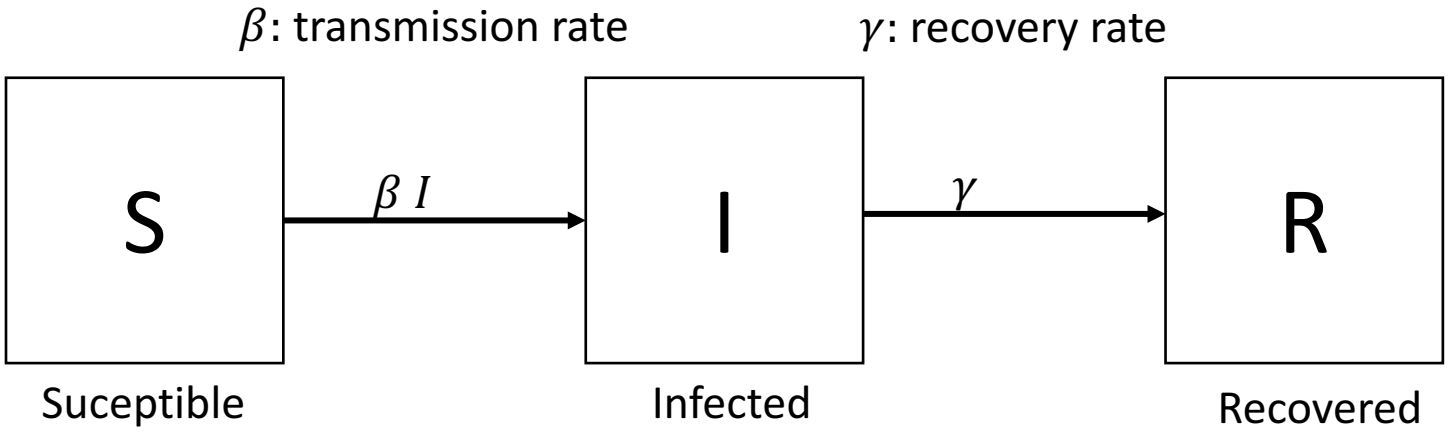


Notice **no empirical data** is involved

Is the causation a bit clearer?

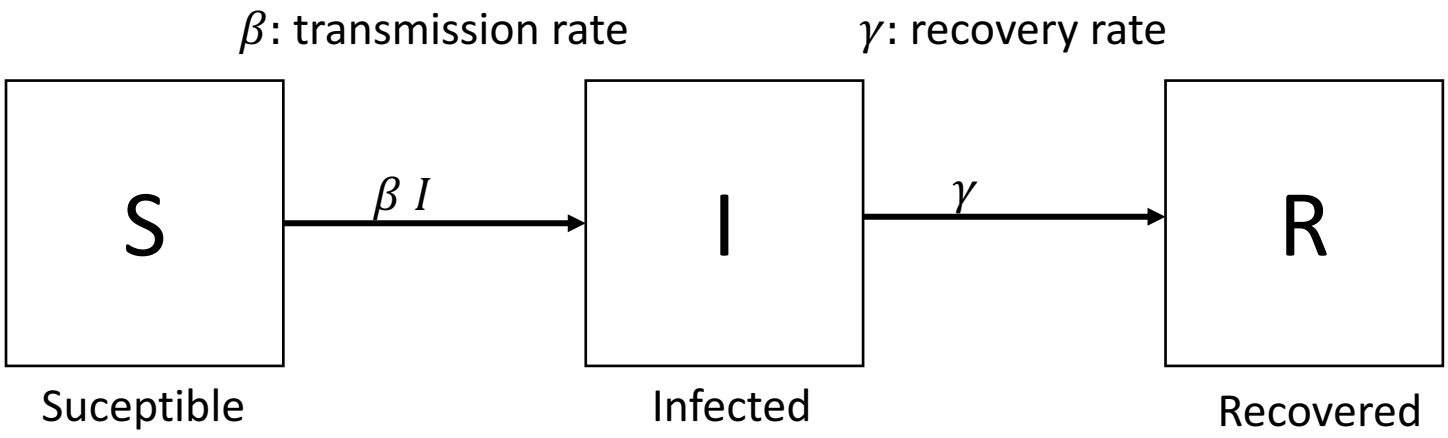
A simple epidemiological model

SIR model: S + I + R = N



(Kermack and McKendrick 1927)

SIR model: S + I + R = N

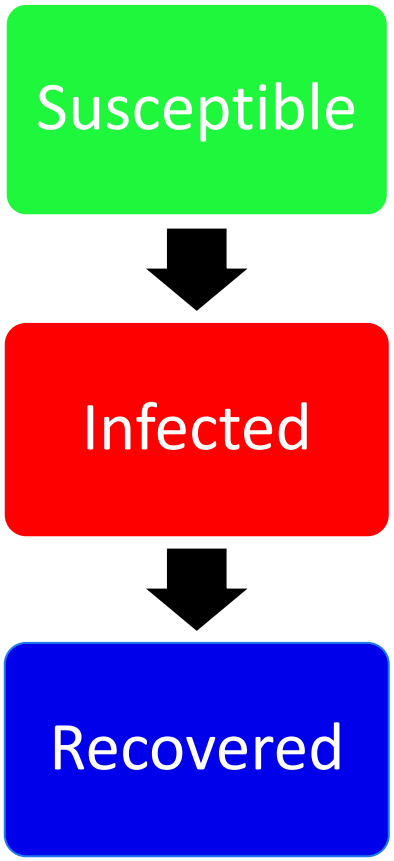


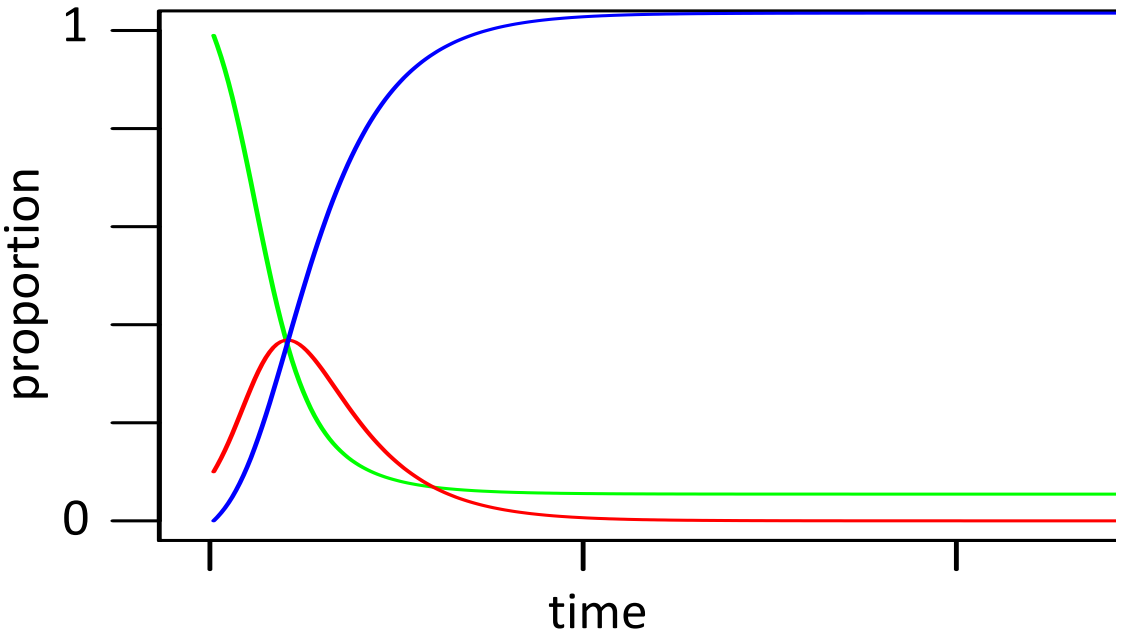
What is a rate?

Rate should always be measured with respect to a time unit!

(Kermack and McKendrick 1927)

Generating data with SIR model





Discuss the result?

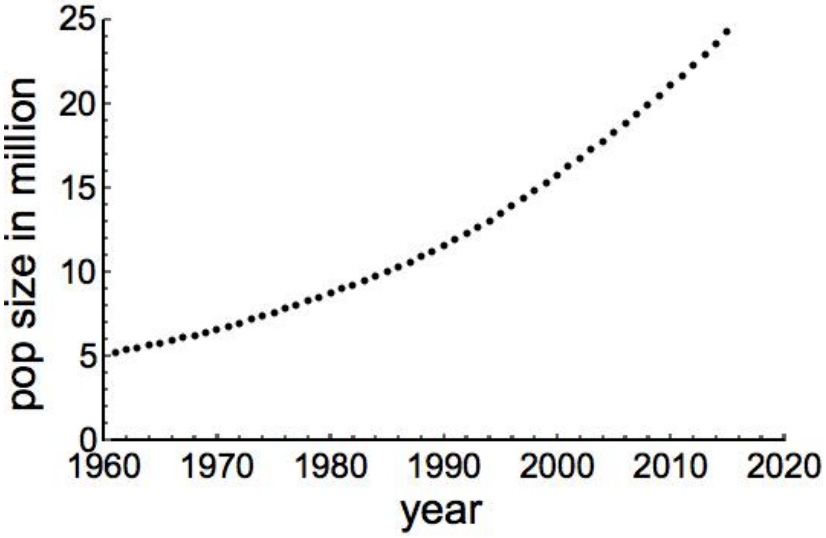
Pros and cons

- Parameters used in the mechanistic models sometimes are not measurable
- Simulations can be computationally intensive
- The increase in computation availability and power foster the use and the increase in the complexity of mechanistic model
- …

Combine statistical model and mechanistic model

Ideal when mechanistic and statistical models meet

Empirical data from world bank



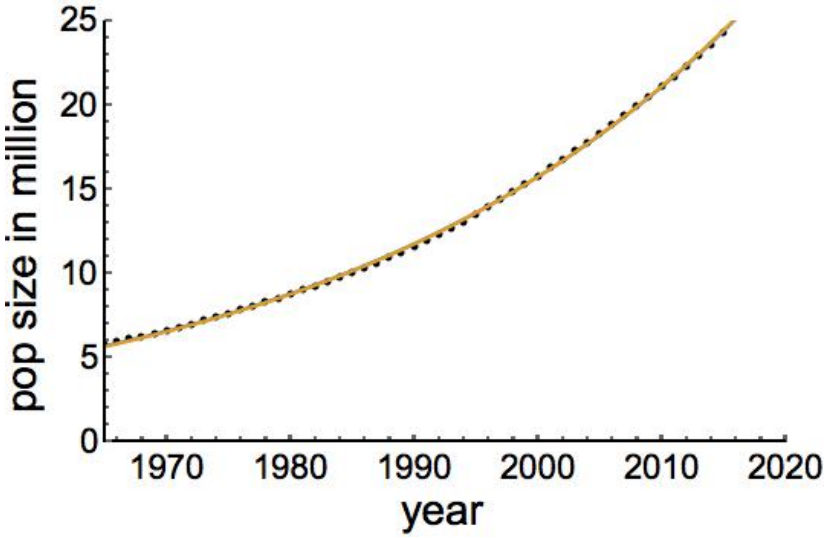
Mechanistic model of population growth

$$Y_t = Y_0 * (1 + b - d)^t$$
$$pop = pop_0 * (1 + r)^t$$

Fit the data using the mechanistically (mathematically) derived relationship

Ideal when math and stat models meet

Fit the data using the mechanistically (mathematically) derived relationship

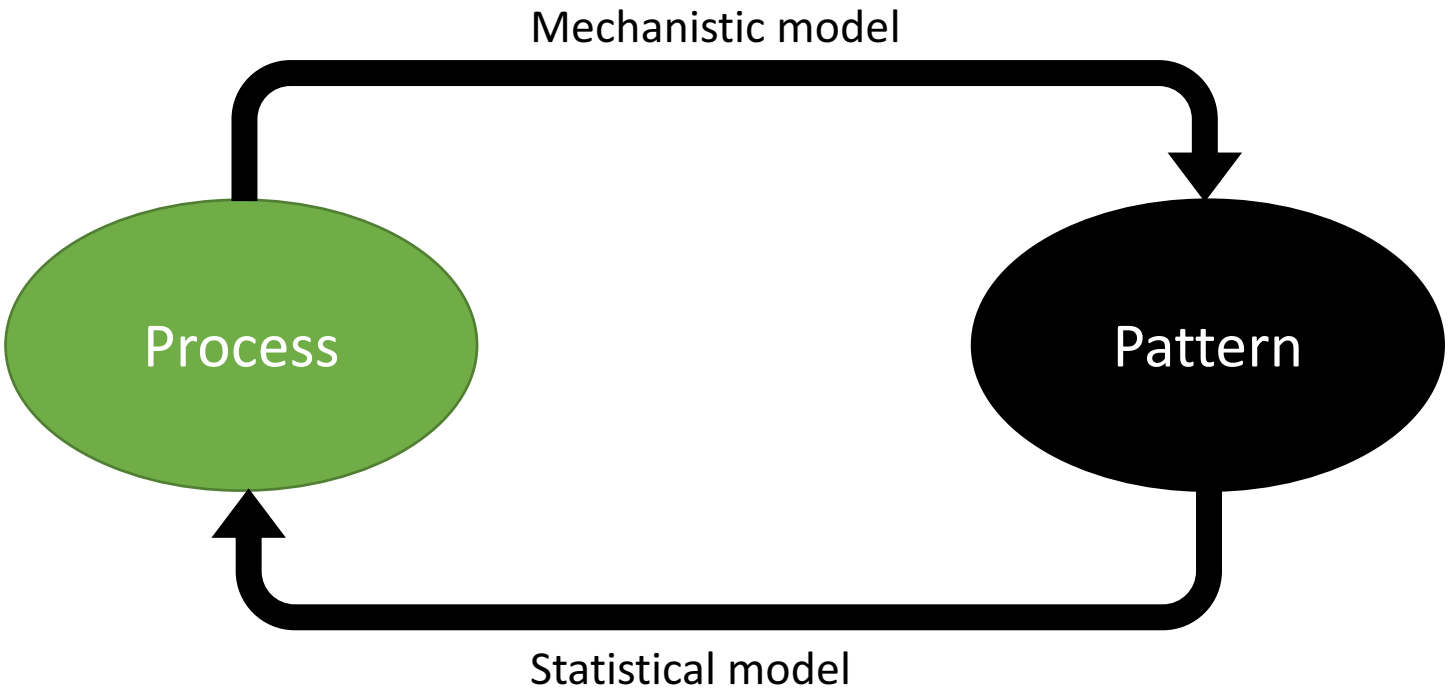


$$Y_t = Y_0 * (1 + b - d)^t$$
$$pop = pop_0 * (1 + r)^t$$

$$r = 2.9\%, (p < 10^{-172}), R^2 = 0.99$$

The annual growth rate for Malagasy population is 2.8 %

Conclusion



Tools

- Computer power keeps increasing
- language/software
  - Fortran, C, C++
  - Julia, Java, Python
  - Matlab, Maple, Mathematica,
  - SAS, SPSS, Stata
- Specific programs
  - Vortex, RAMAS, NetLogo for IBM
  - NicheMapper for physiology, iLand for forest dynamics
  - MaxEnt for species distribution modeling
  - Zonation for reserve selection etc…
- The compromise: R---very powerful for
  - Visualization
  - Data formatting and sorting
  - Statistical analyses
  - Simulation (mechanistic model)

Thank you for your attention!
Questions?