Tanjona Ramiadantsoa (with materials from Cara Brook)

Data and models

E2M2 2020
Valbio, Ranomafana
January 7th, 2020

Adjunct professor,
University of Fianarantsoa
Department of Life Science
Department of Mathematics

Honorary Fellow,
University of Wisconsin-Madison
Department of Integrative Biology

Forewords

- I tend to talk fast, please interrupt and remind me all the time

Forewords

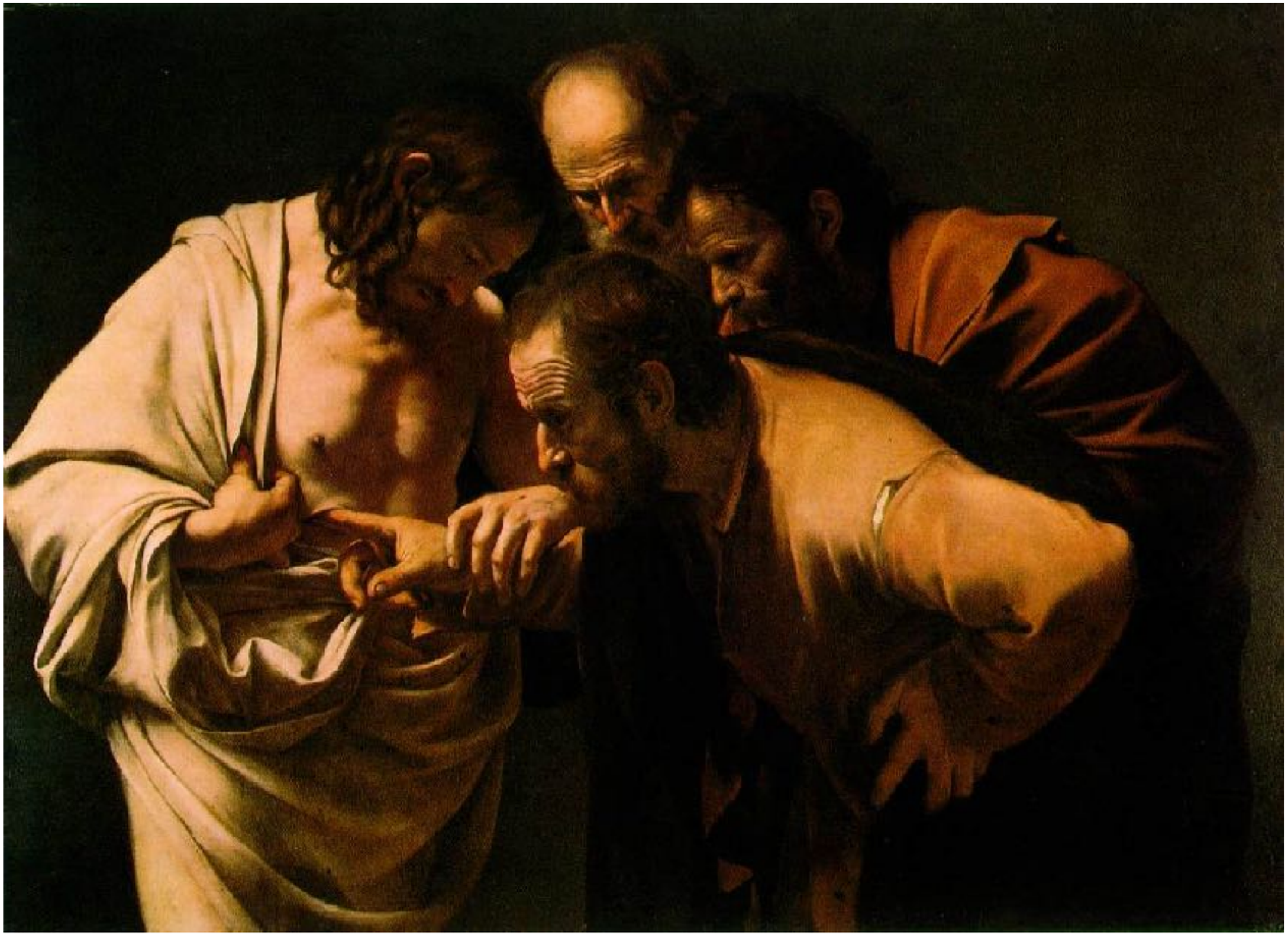
- I tend to talk fast, please interrupt and remind me all the time
- This workshop is for you!!!

Forewords

- I tend to talk fast, please interrupt and remind me all the time
- This workshop is for you!!!
- I use a lot of Disney and underwear references (no particular reason)

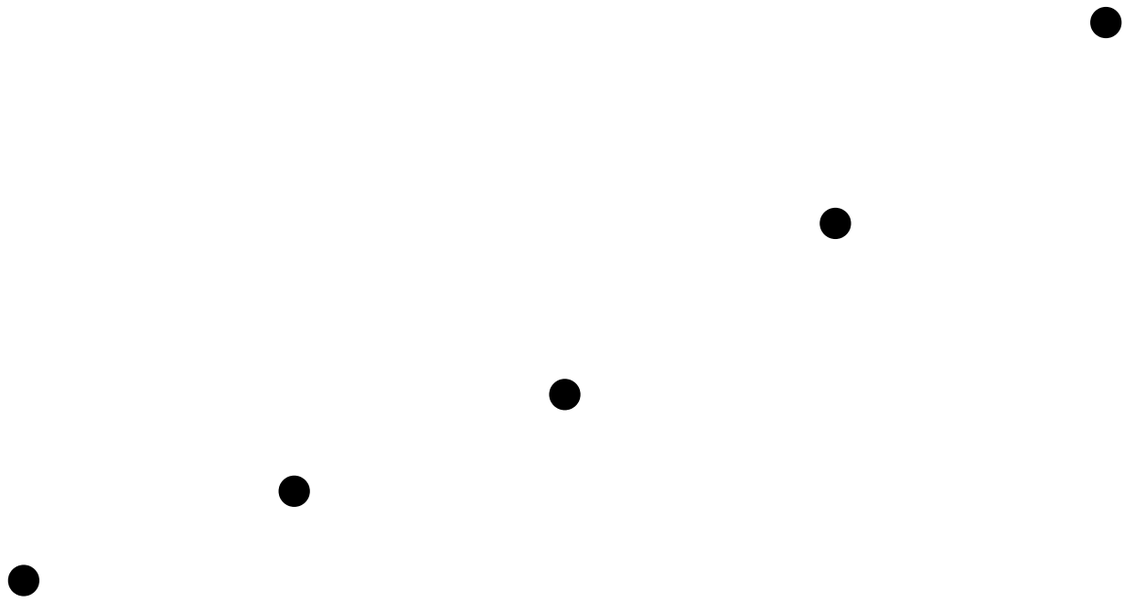
What is a scientist?

(St.) Thomas

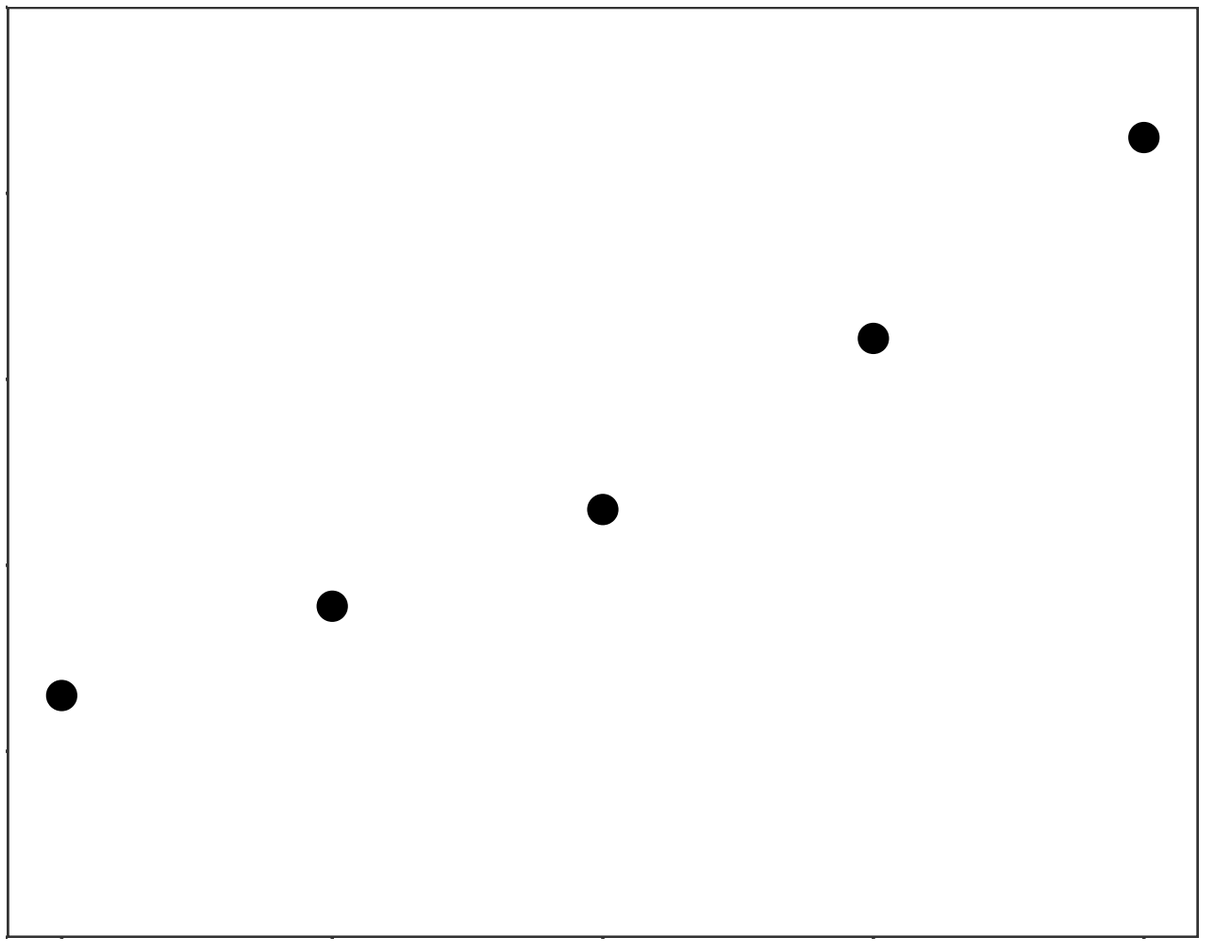


The ability to correctly collect and interpret data makes a difference between the scientifically literate and illiterate

Is this data?

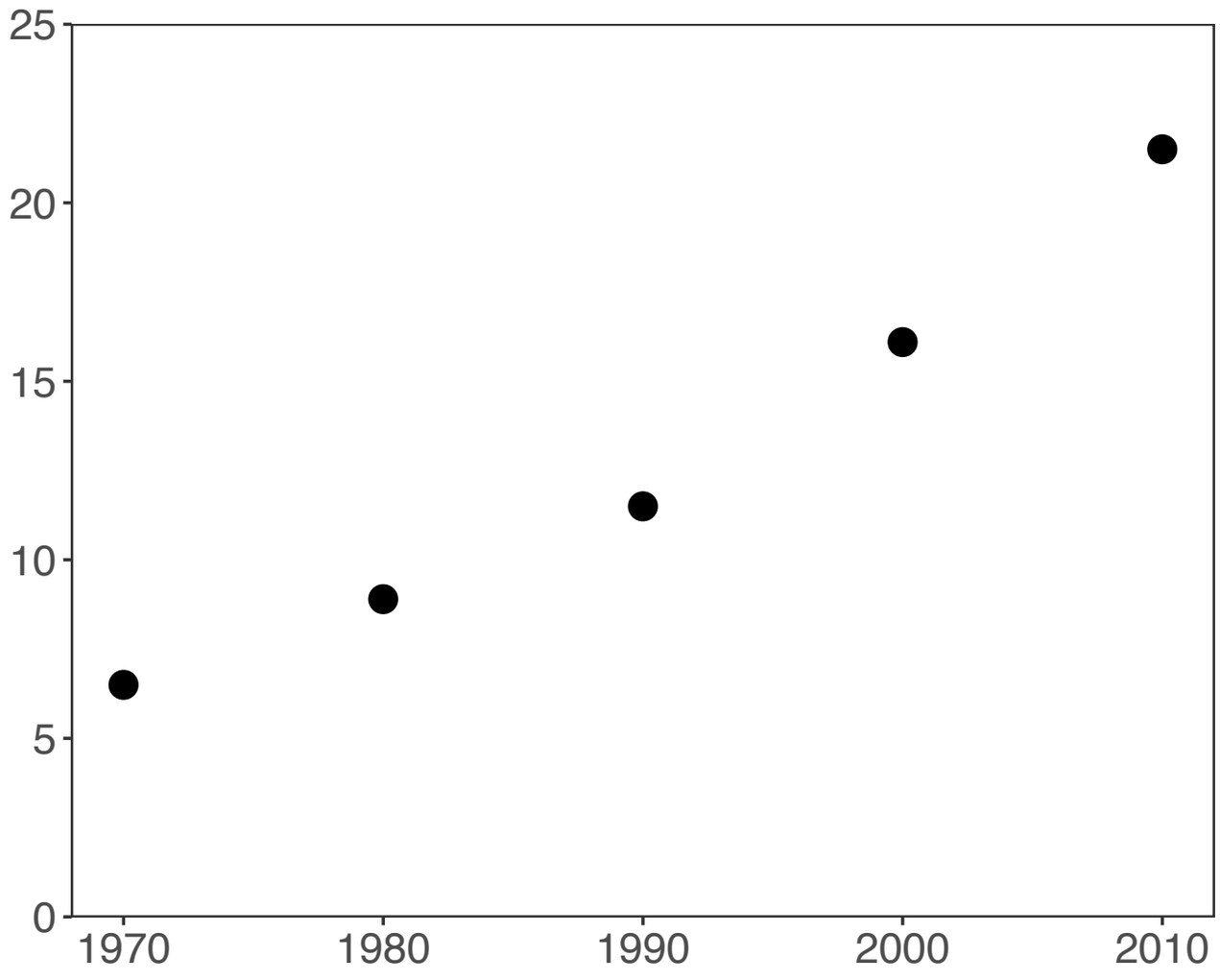


Is this data?

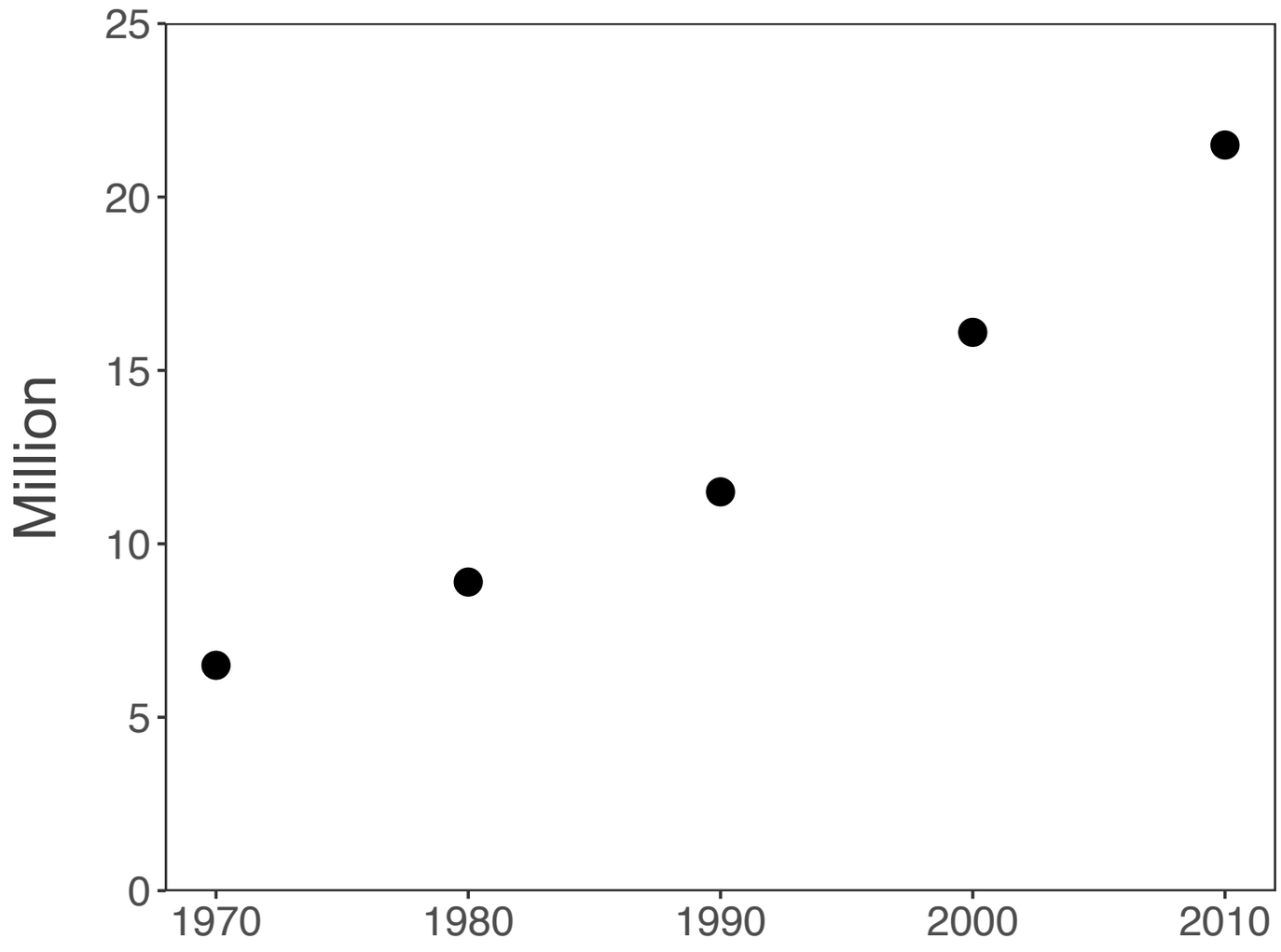


What do we need to make this a data?

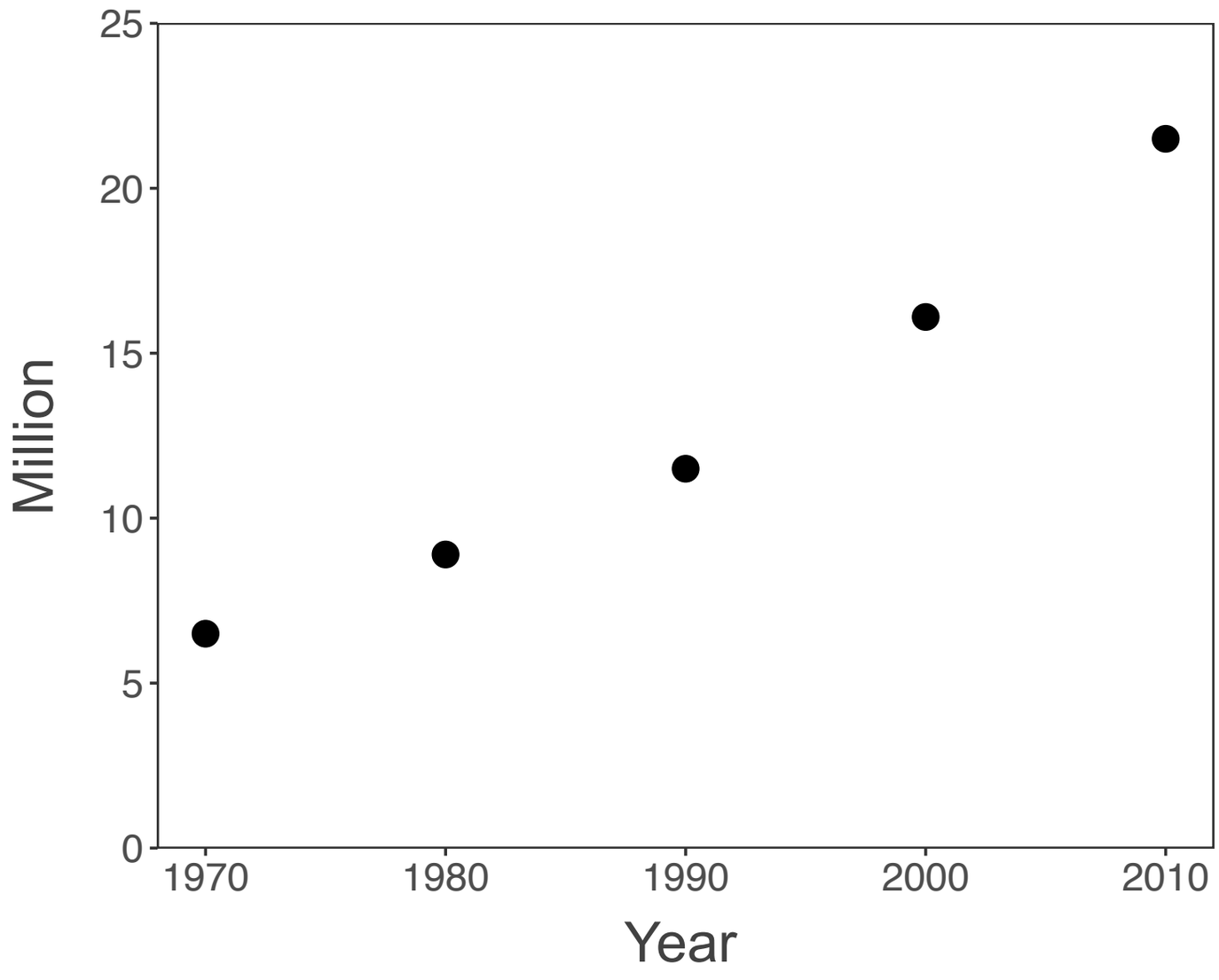
Is this data?



Is this data?

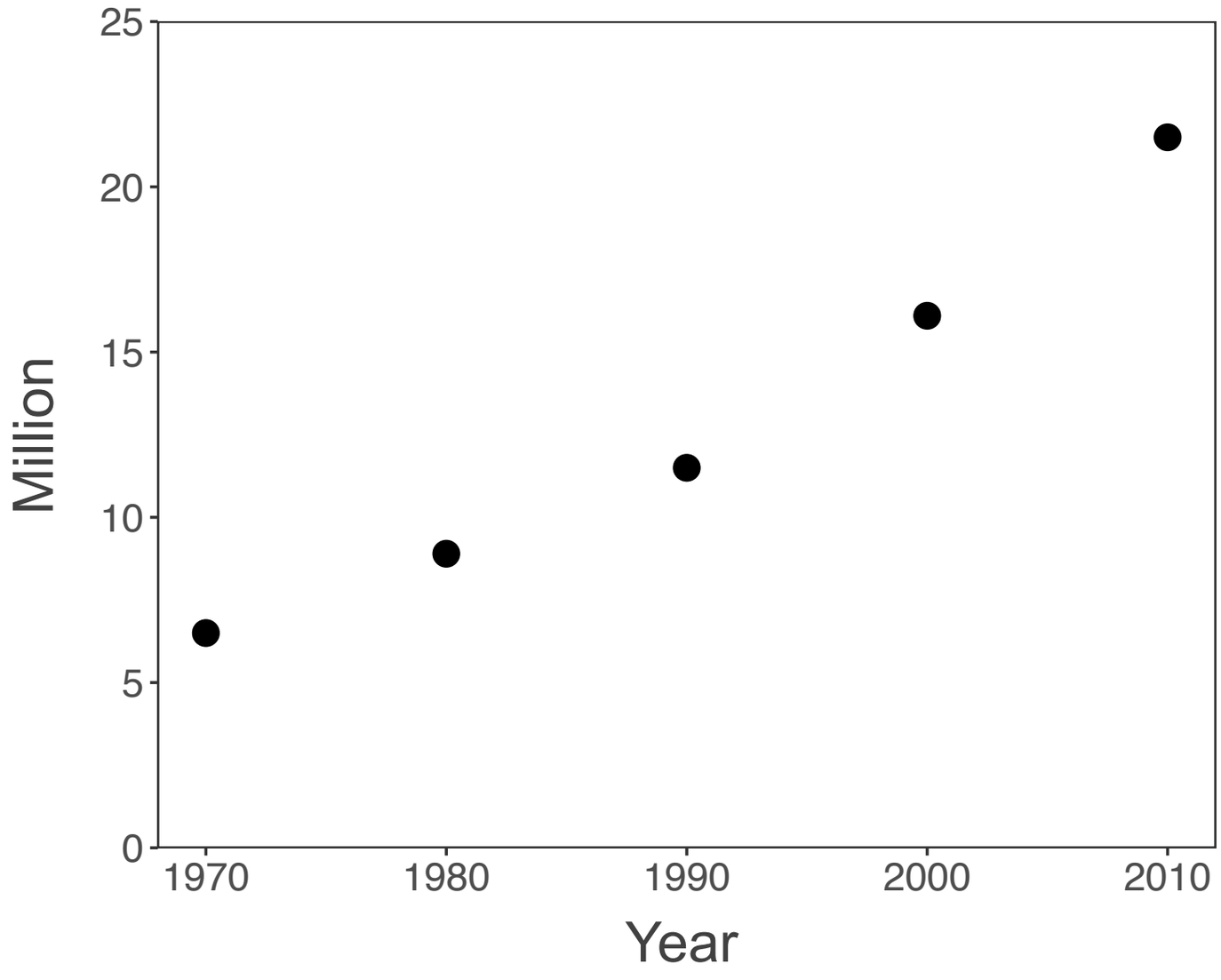


Is this data?

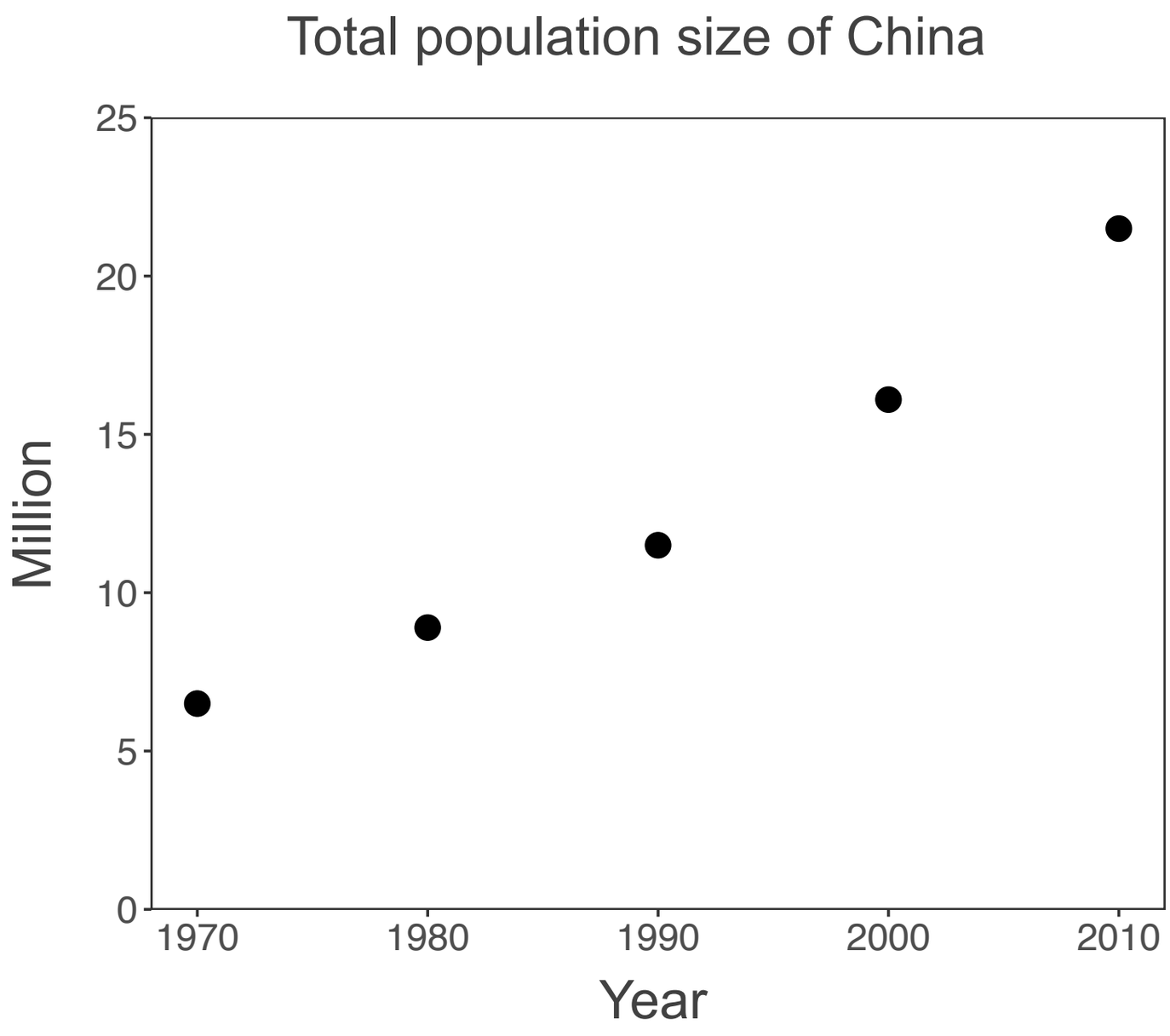


Is this data?

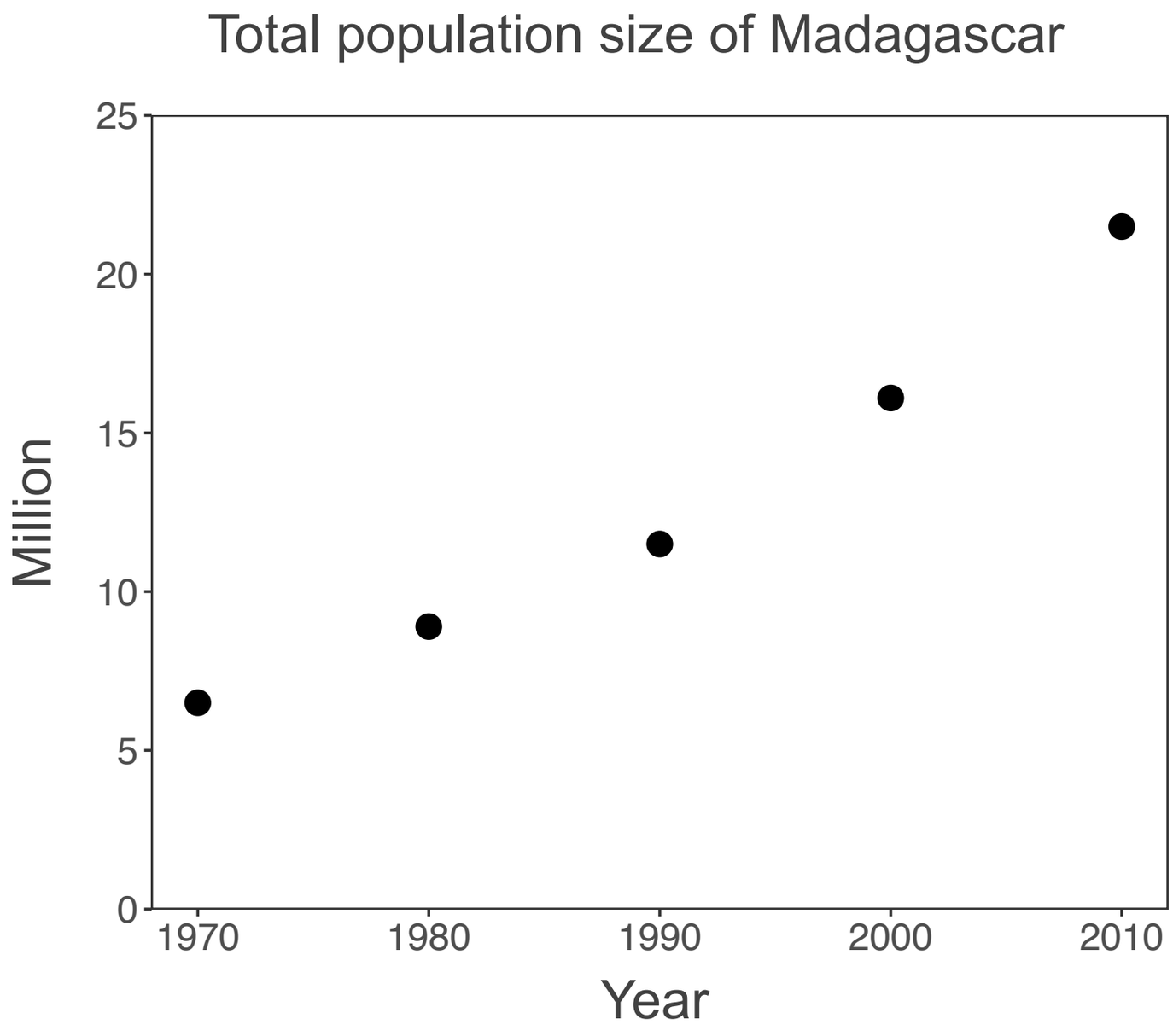
Number of views for Justin Bieber's "Baby" video on Youtube



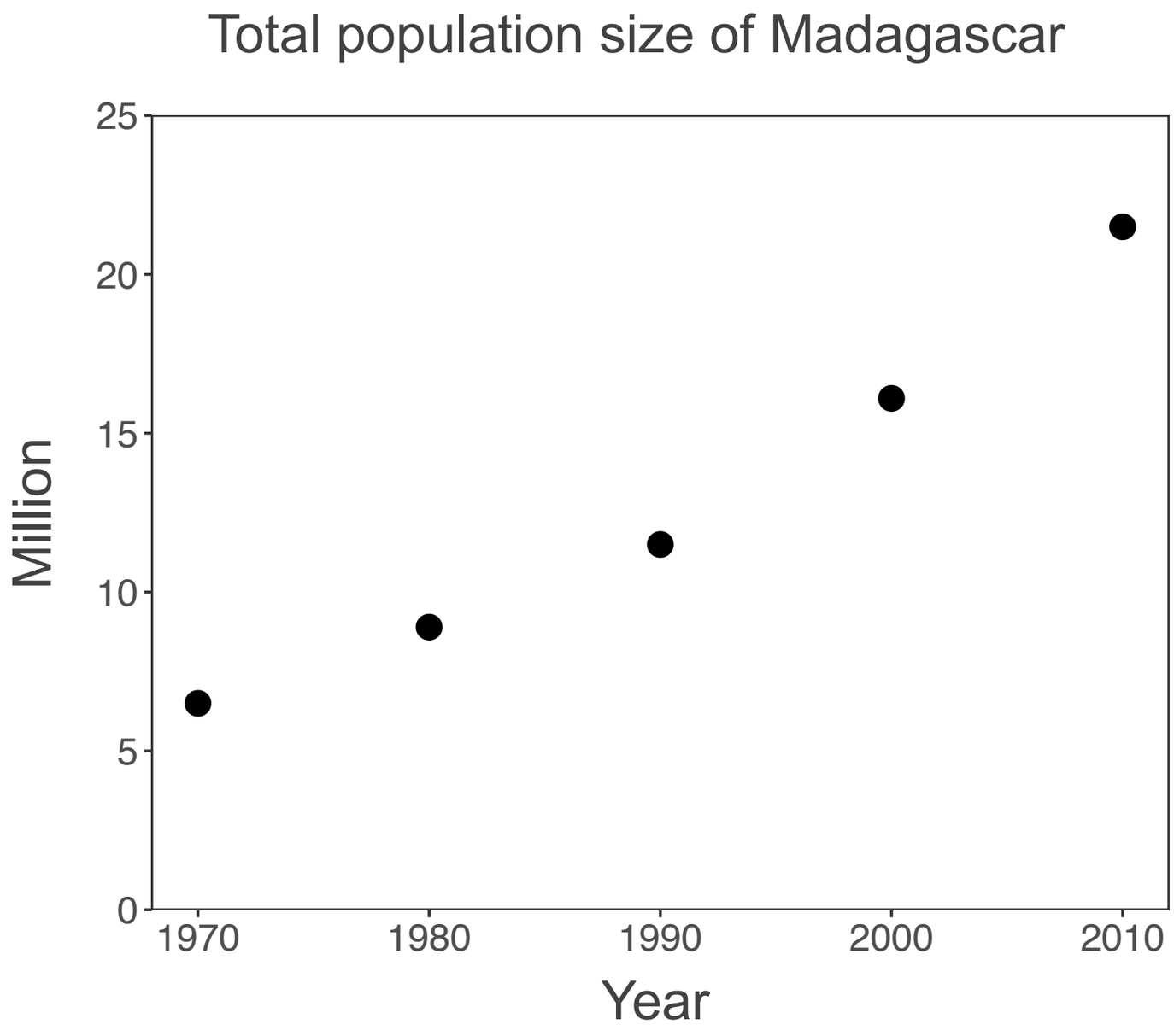
Is this data?


Total population size of China
Million
25
20
15
10
5
0
1970
1980
1990
2000
2010
Year

Is this data?


Total population size of Madagascar
Million
25
20
15
10
5
0
1970
1980
1990
2000
2010
Year

Is this data?



Source: World Bank (accessed 2017)

Notes

- Asking what is a data can become extremely philosophical
- A data should have at least one explanatory (X) and one response variable (Y) and be precise
- More practically, data are evidences to **support claims**

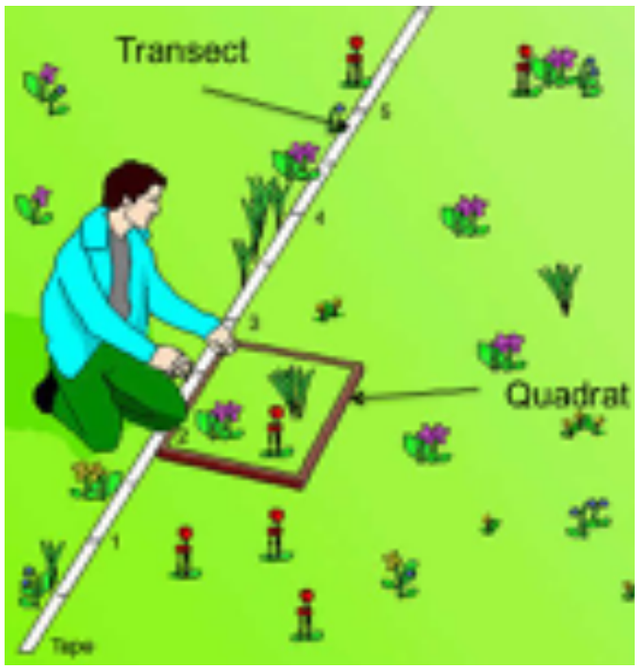
Data: X and Y values

- **Numerical**
  - A variable is numerical when you can transform it with mathematical operation
  - Examples: integer, real number, multi-dimensional number

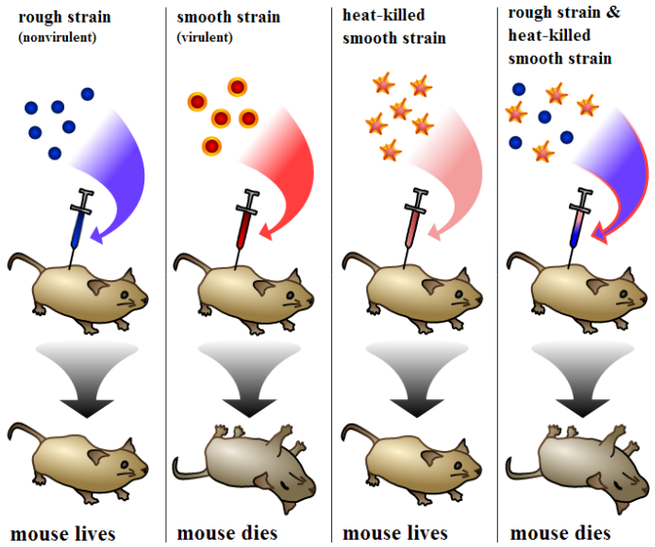
- **Categorical ('factors' in R language)**
  - A variable is categorical when it is not numerical but a categorical can be numerical?
  - Examples: colors, (blood) types, species name

Source of data

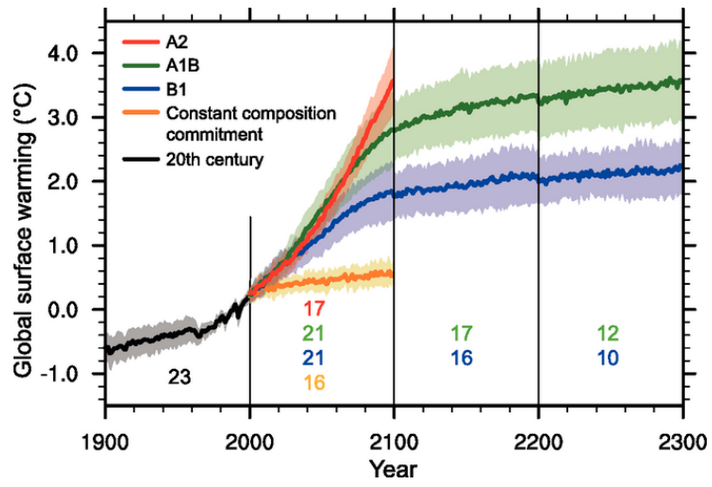
Observational



Experimental



Simulated



Empirical data

What is X and what is Y?

Things to consider

- Data acquisition
  - Impossible
  - Theoretically possible but practically unfeasible
- Data quality and quantity
  - Trade-off with time, money, human effort …
- Reproducibility
  - Spatial and temporal
- Measurement errors
- Open access?

Thoughts on data?

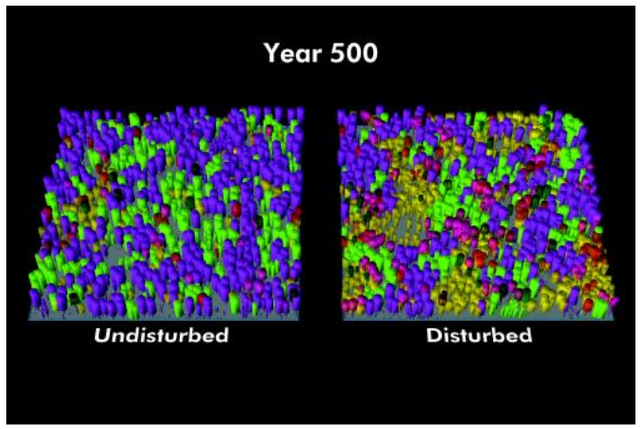
Models

Is this a model?

Human

Car

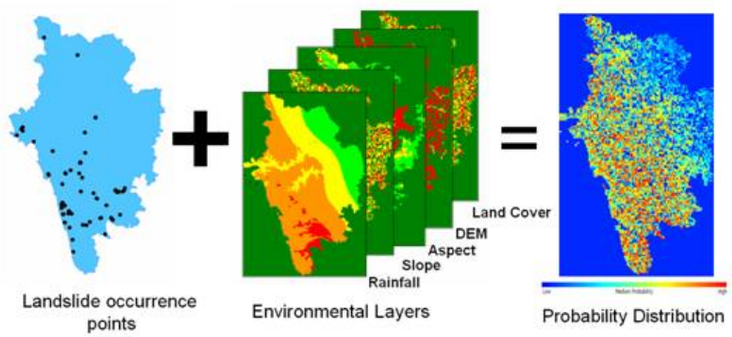
Forest simulator



Ecology & Evolution

Zoonotic pathogen

Niche mapping



Models

A model := a simplified representation of a phenomenon

During E2M2, you will learn steps for a modeling project

Steps to a modeling project

- Research question
- Formulate a hypothesis
- Test the hypothesis with data
- (Models are scientific procedures to generate data and / or analyze data)

Two broad classes of models

Statistical

Correlative

Mechanistic

Causative

Statistical vs. mechanistic model

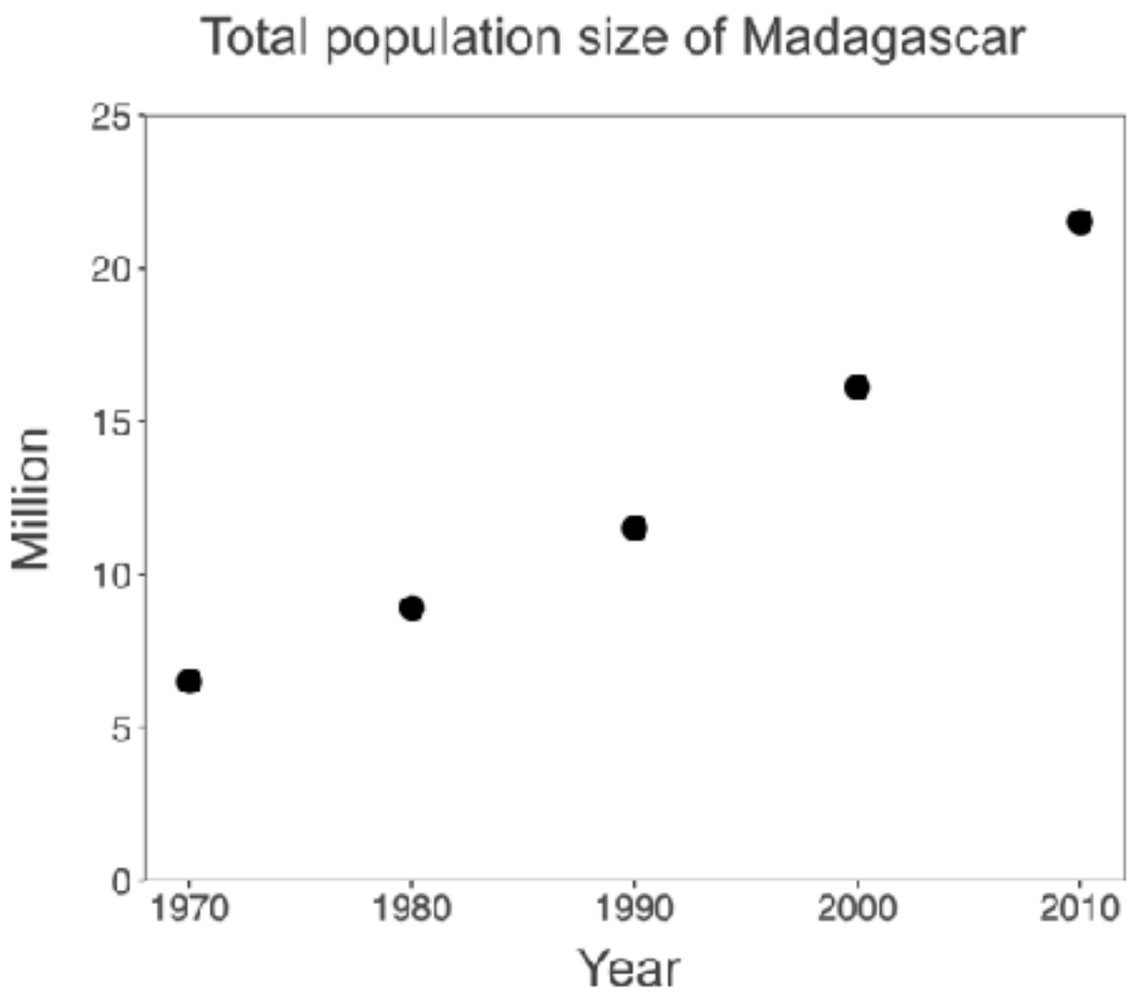
- Statistical:
  - You ask: What?
- Mechanistic
  - You ask: How?

Statistical vs. mechanistic model

- Statistical:
  - You ask: What?
  - Require data collection
- Mechanistic
  - You ask: How?
  - Make your own data

Statistical vs. mechanistic model

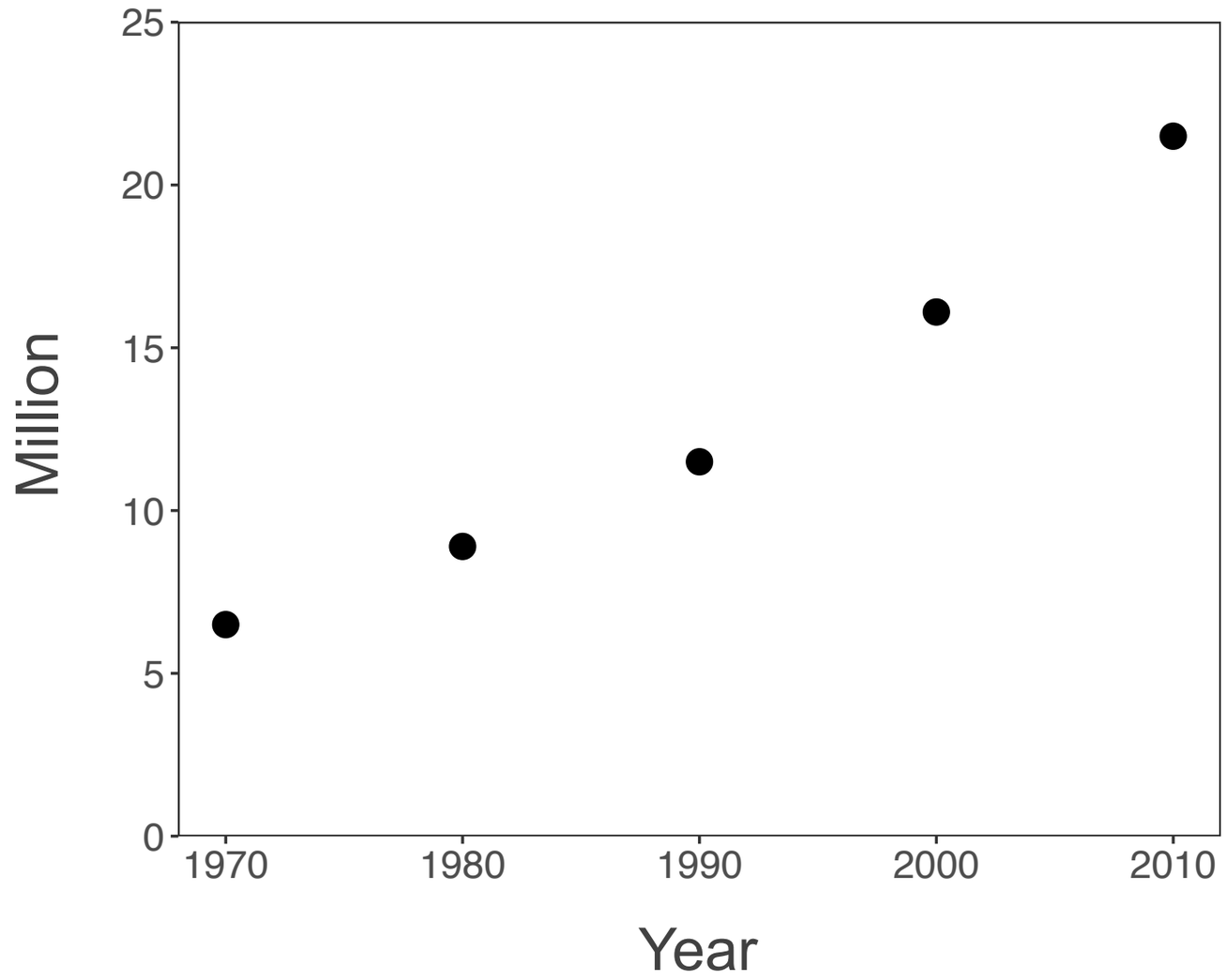
- Statistical:
  - You ask: What?
  - Require data collection
  - Fit a function to the data (estimate parameter values)
- Mechanistic
  - You ask: How?
  - Make your own data
  - Process generating a phenomenon (explore the effect of changing parameter values)



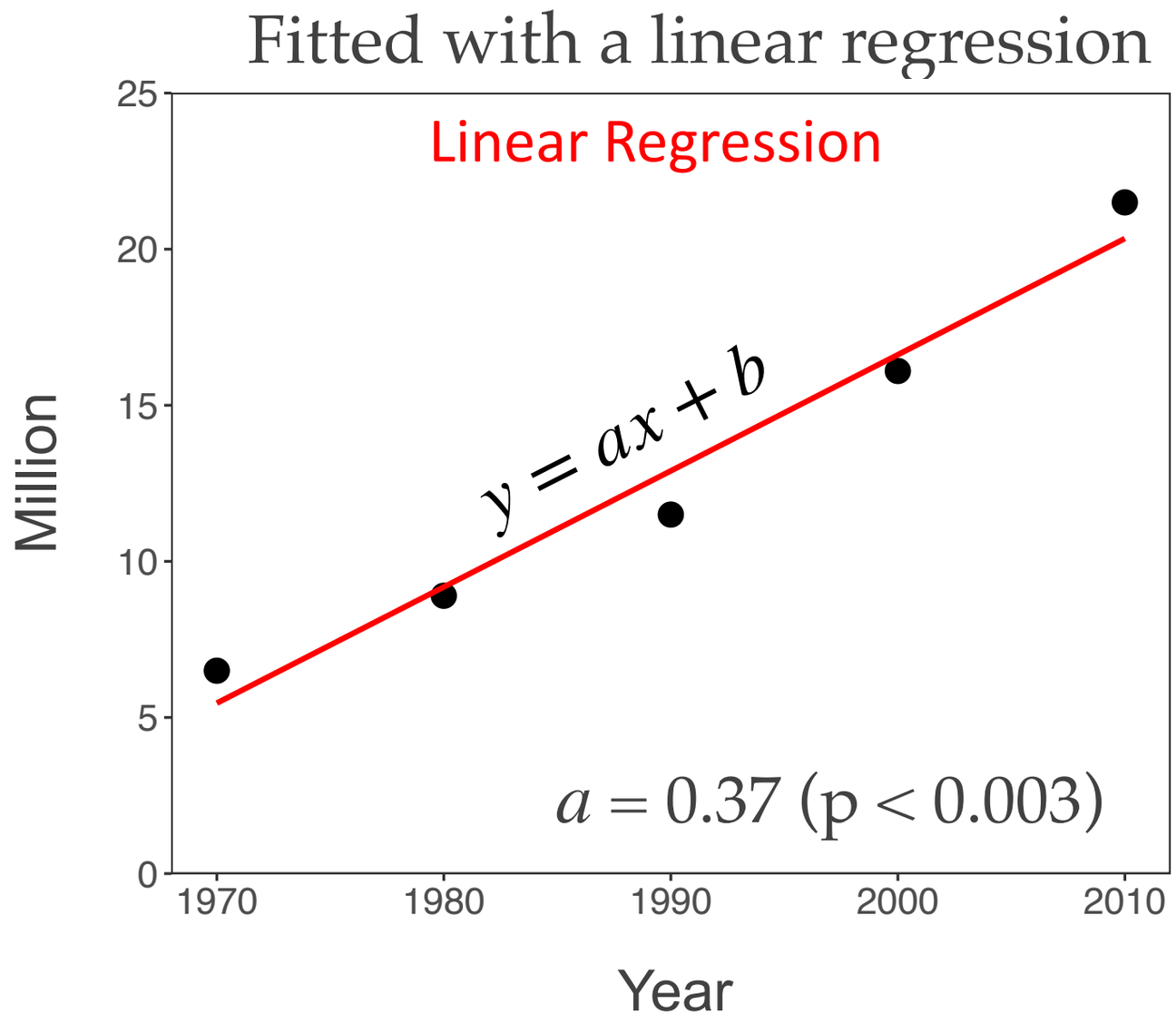
During E2M2, you will look at the figure in two ways: empirical or simulated data

Statistical model

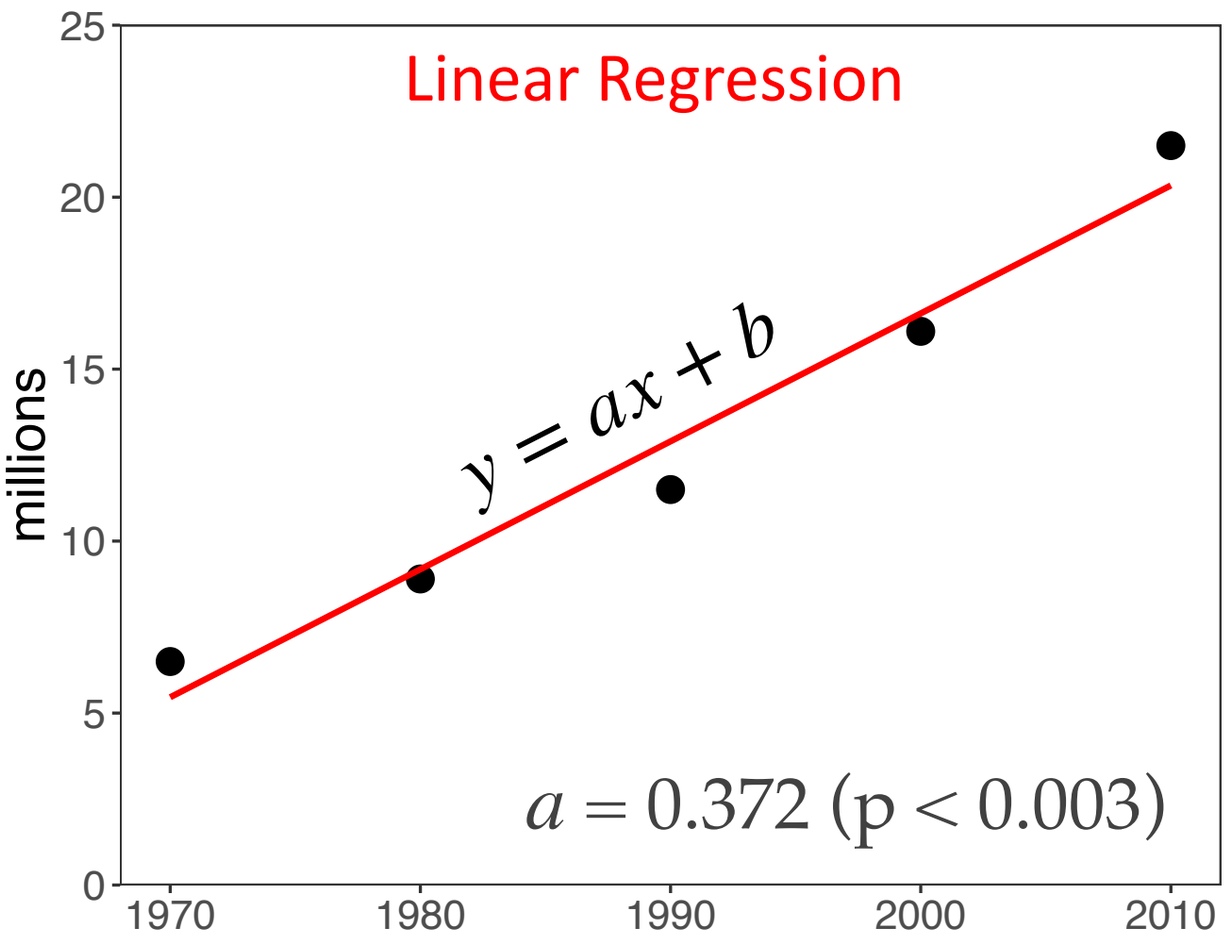
This is an empirical data

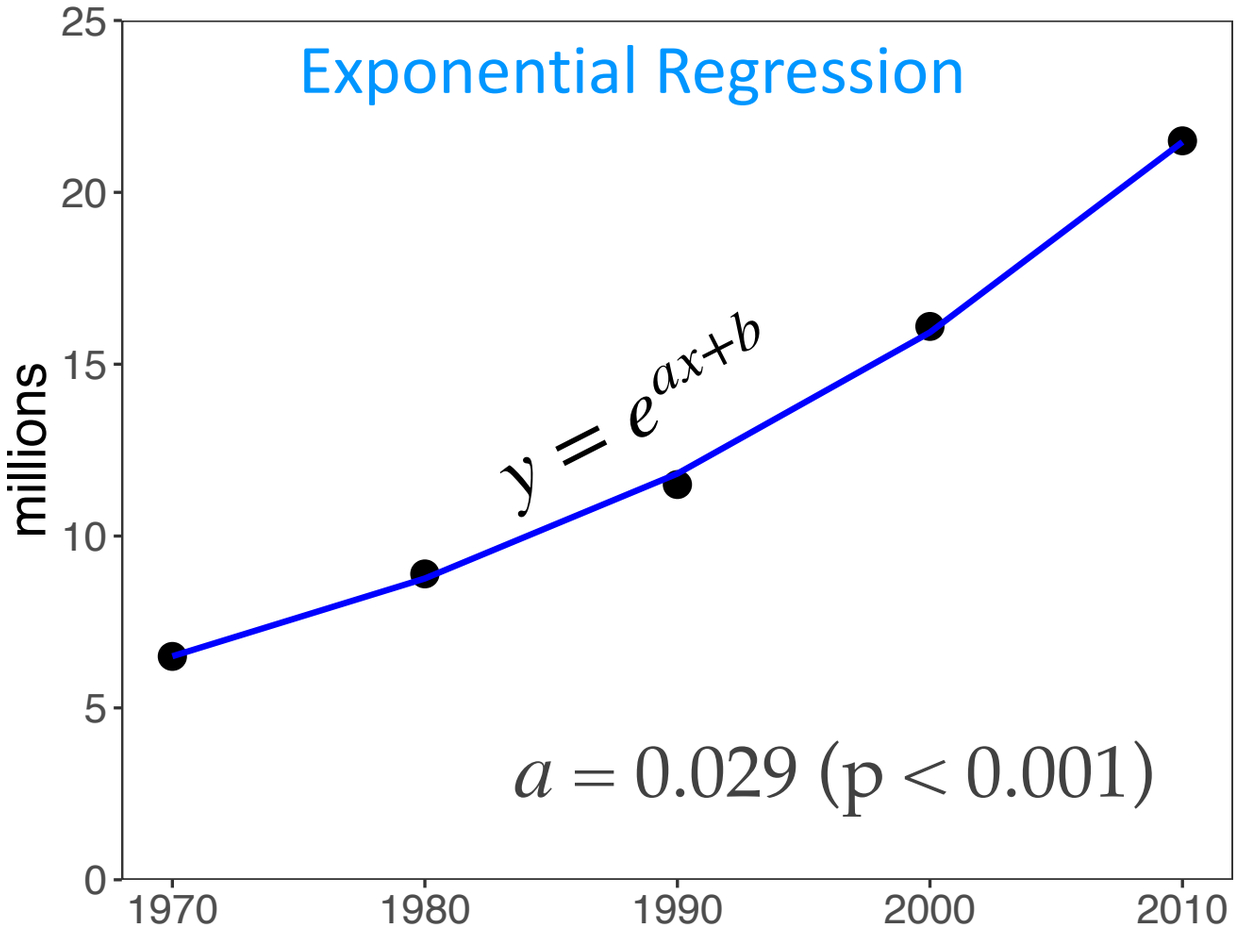


Using a statistical model



Two statistical models





That was wrong!!!

What is the research question? What is the hypothesis?

A model is a tool not an endpoint

Remember the steps

- Research question
- Formulate a hypothesis
- Test the hypothesis with the data
- (Models are tools to test your hypothesis)

During this week, you will **spend a lot of time** writing research questions and formulating hypotheses

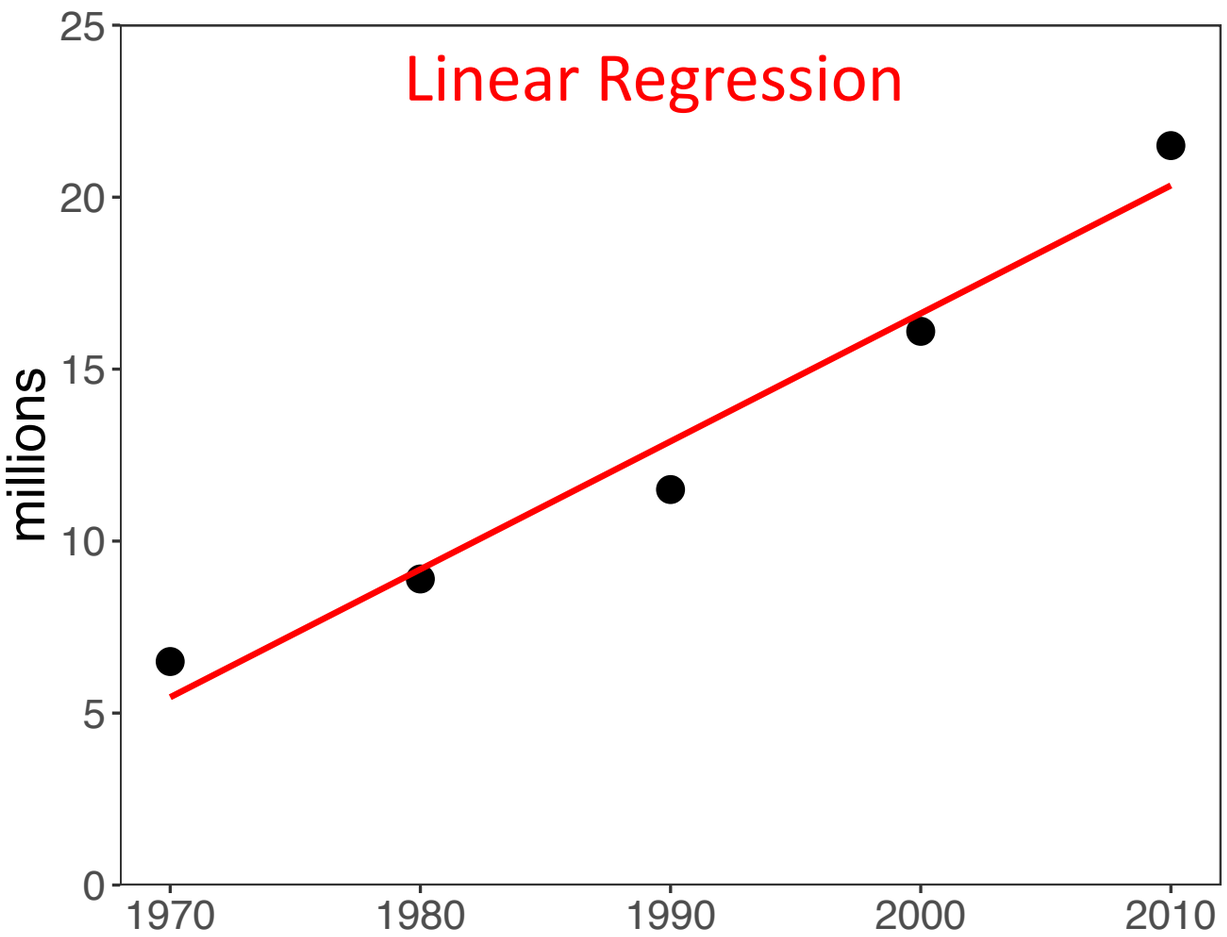
Tuesday, Jan 7: "Understanding Your System"

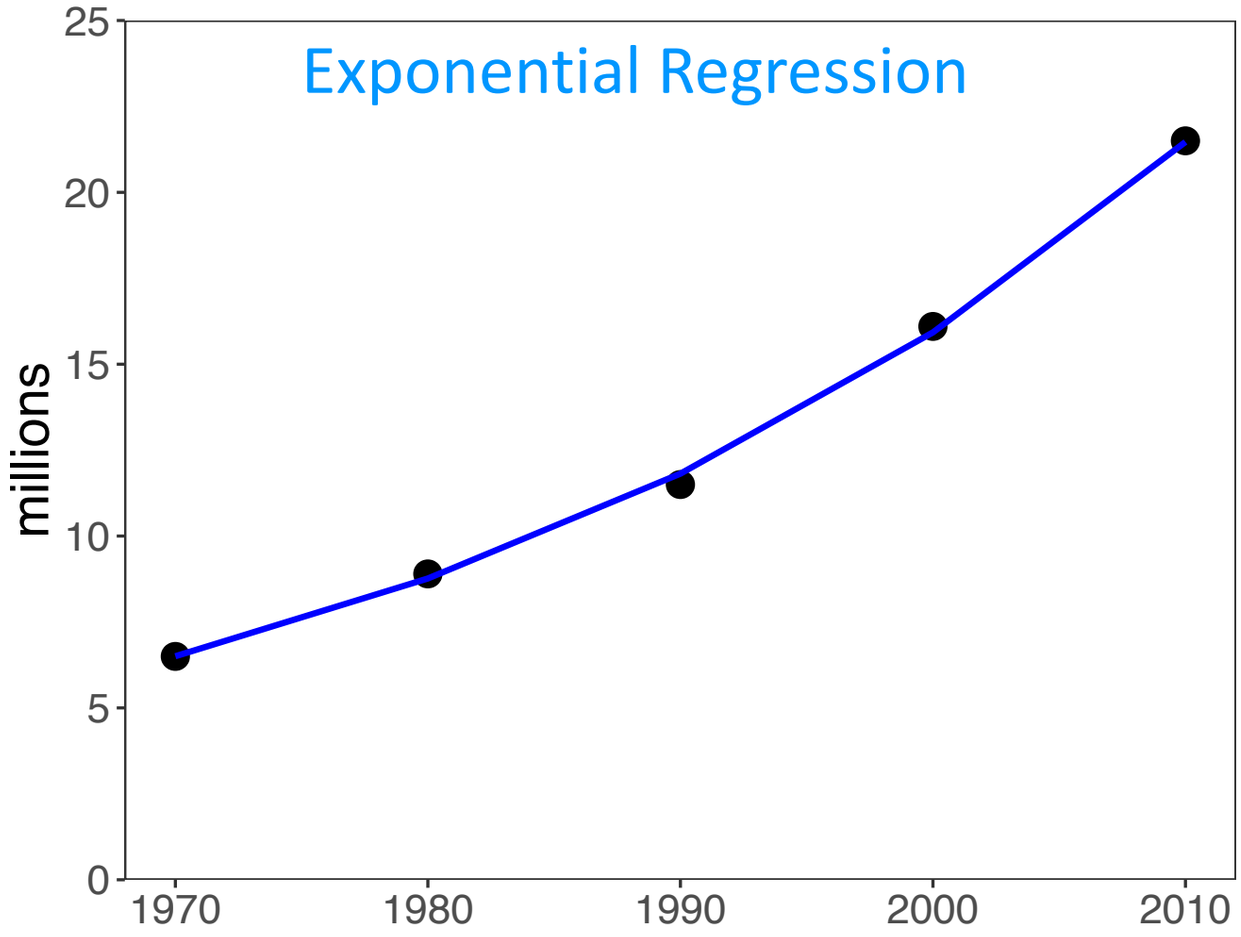
- 6:30-8:00am: Breakfast
- 8:00-8:30am: Road Map and Daily Agenda. Brief Introductions **(Cara)**
- 8:30-9:00: Lecture: Ecology Meets Epidemiology **(Cara)**
- 9:00 - 10:00am: Lecture: What are We Doing Here? Data and Models **(Tanjona)**
- 10:00-10:30am: Coffee Break
- 10:30-12:00pm: Reviewing Key Concepts in Mathematics **(Tanjona)**
- 12:00-1:00pm: Lunch
- 1:00-2:00pm: Introductions: 2-min student/mentor/instructor introductions **(Sarah)**
- 2:30-4:00pm: Formulating research questions **(Cara)**
    - For HW, write your own research question
- 4:00-4:30pm: Coffee Break
- 4:30-6:30pm: Lecture/Tutorial: Study Design and Data Collection **(Sarah)**
- 6:30-7:30pm: Dinner

tics **(Andres, Fara)**

- 12:00-1:00pm: Lunch
- 1:00-2:30pm: Exercise + Discussion: Dynamical Fever **(Christian, Fara, Mentors)**
- 2:30-3:00pm: Coffee Break
- 3:00-4:30pm: Lecture/Tutorial: Introduction to Compartmental Models and Differential Equations **(Cara)**
- 4:30-5:00pm: Small Group Sessions: Refining research questions for modeling **(all instructors lead small groups)**
    - All students should have a workable statistical and mechanistic question by the end of the session
    - For HW, assign: Creating a model world to address a research question
- 5:00-6:30pm: Mid-session feedback **(Program Evaluator)**
- 6:30-7:30pm: Dinner

Transition to mechanistic models

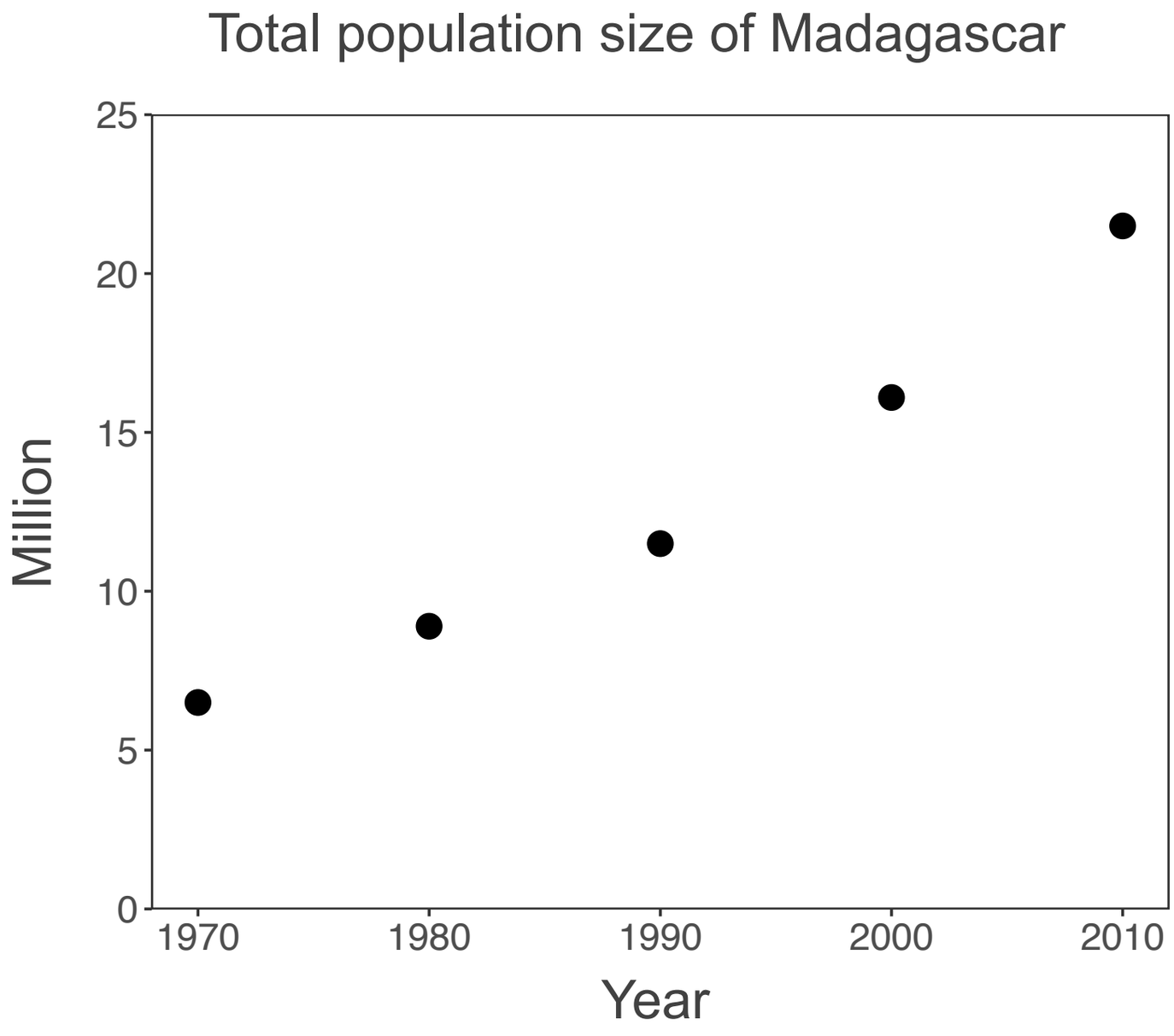




Tells nothing about why population is increasing

Mechanistic model

This is a simulated data



Source: World Bank (accessed 2017)

Mechanistic

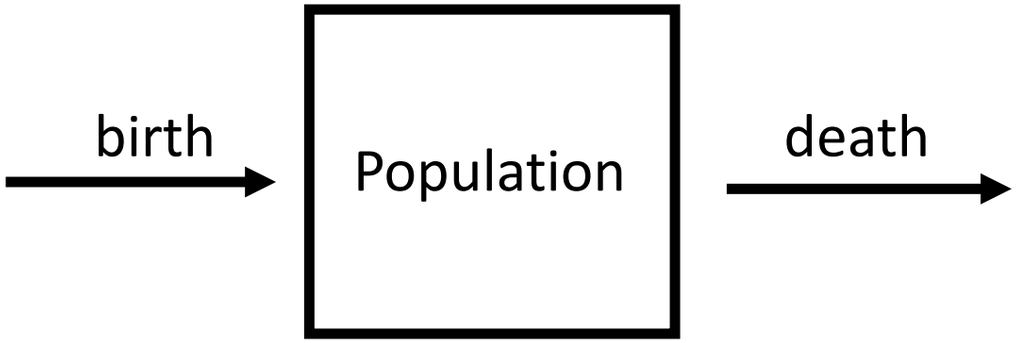
- Mechanistic model allows you to **make your own data**, generate data through time
- You think about the type of processes that can generate the phenomenon you are interested in
- Translate the processes into mathematical equations (your equation is your model)
- Solve the equation!

Mechanistic model

- Research question: How does a population change if birth and death rates are constant?
- Formulate a hypothesis: Population increases exponentially if birth > death, and goes extinct if birth < death
- Build the model that translate your hypothesis into equation
- Assess if the solution of the equation agrees with the observation in the research question

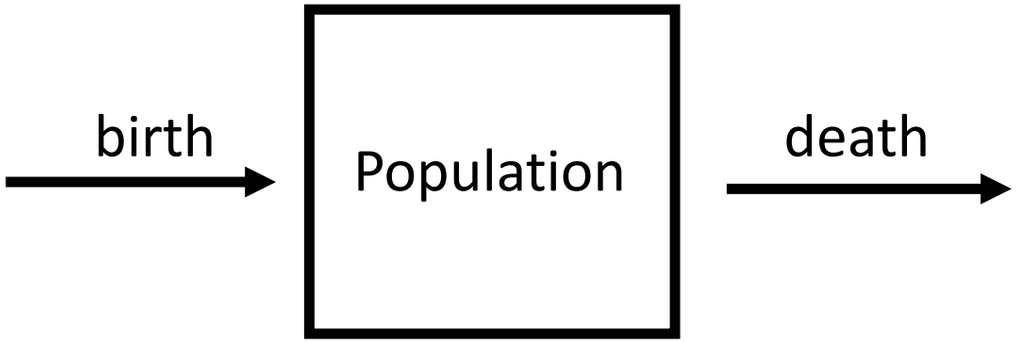
Mechanistic model

Compartmental model



Mechanistic model

Compartmental model

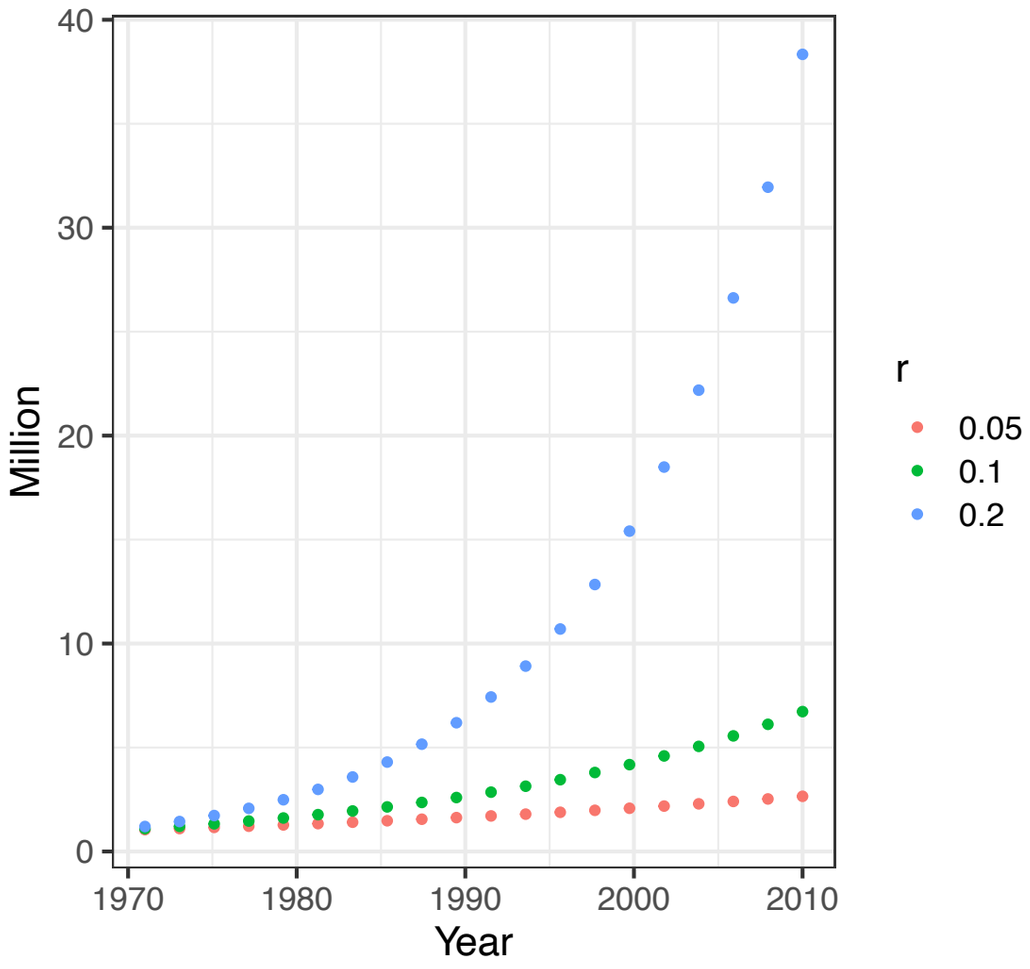


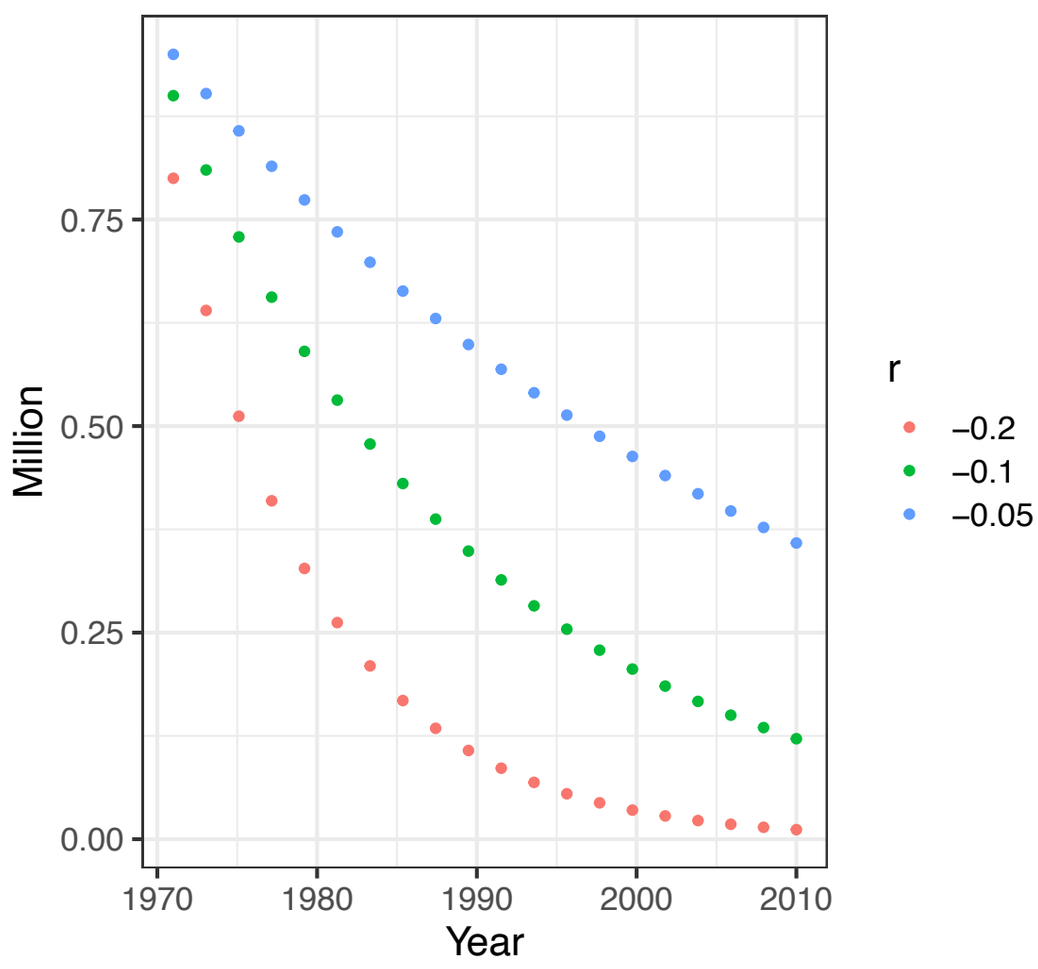
In terms of equation :

$$p_{t+1} = p_t + bp_t - dp_t = (1 + r)p_t$$

Comparing with the hypothesis

- Population increases exponentially if birth > death
- Population goes extinct if birth < death





Mechanistic model

- Research question: How does a population change if birth and death rates are constant?
- Formulate a hypothesis: Population increases exponentially if birth > death, and goes extinct if birth < death
- Build the model that translate your hypothesis into equation: $p_{t+1} = (1 + b - d)p_t = (1 + r)p_t$
- Assess if the solution of the equation agrees with the observation in the research question: it agrees

Two broad classes of models

Statistical

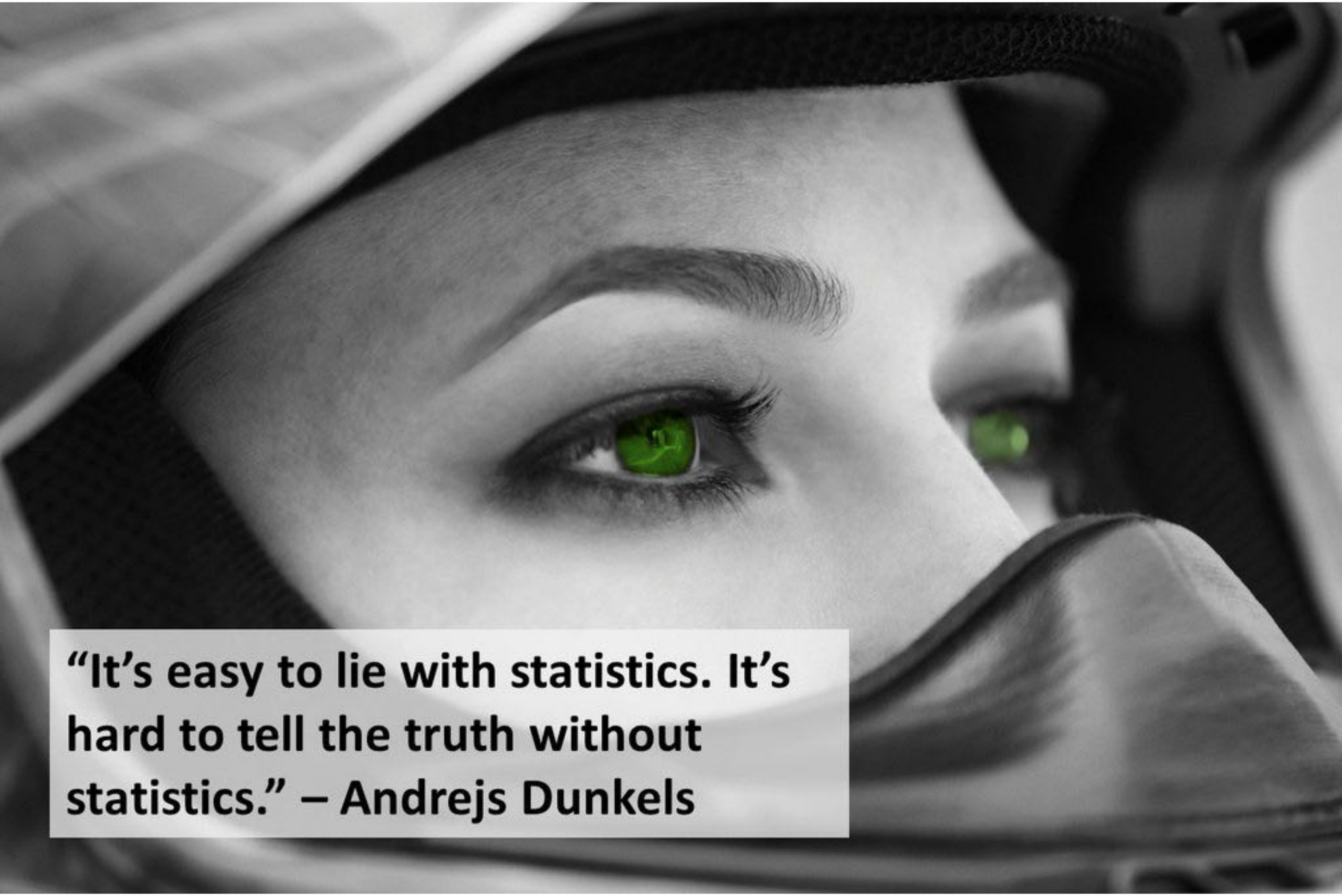
Correlative

Mechanistic

Causative

Statistical model: beware!!!

- Statistical models are based on specific assumptions
  - Make sure your data does not violate these assumptions
  - What happened if you violate those assumptions?
- There are so many statistical models
  - There is not necessarily a single best approach
- Advances in computational powers, statistical packages, and data size often attracts people to use **complex statistical** models which **are not necessarily better**


“It’s easy to lie with statistics. It’s hard to tell the truth without statistics.” – Andrejs Dunkels

Mechanistic model: beware !!!

- Parameters used in the mechanistic models sometimes are not measurable
- Simulations can be computationally intensive
- Advances in computational power often inspire the development of **more complex models** which are **not necessarily better**

Coralie Fritsch

Principle

- When building a model, you include elements that you feel is most important (parsimony) to explain a phenomenon
- The details depend on how similar do you want your model to reproduce the real-world data

R is just a tool

R is a good (not the best) program
for mechanistic and statistical modeling

Take homes

- Research question and hypothesis are more important that models
- Any data needs context: the X and Y should be clear
- Models are rigorous tools to assess how the data support the claim
  - There are figuratively an infinite number of models
  - Statistical model works with question starting with **what**
  - Mechanistic model **generates data** and works with question starting with **how**

E2M2

- Ask questions
- Perform classical statistical model
- Create simple mechanistic model
- Use R as a tool