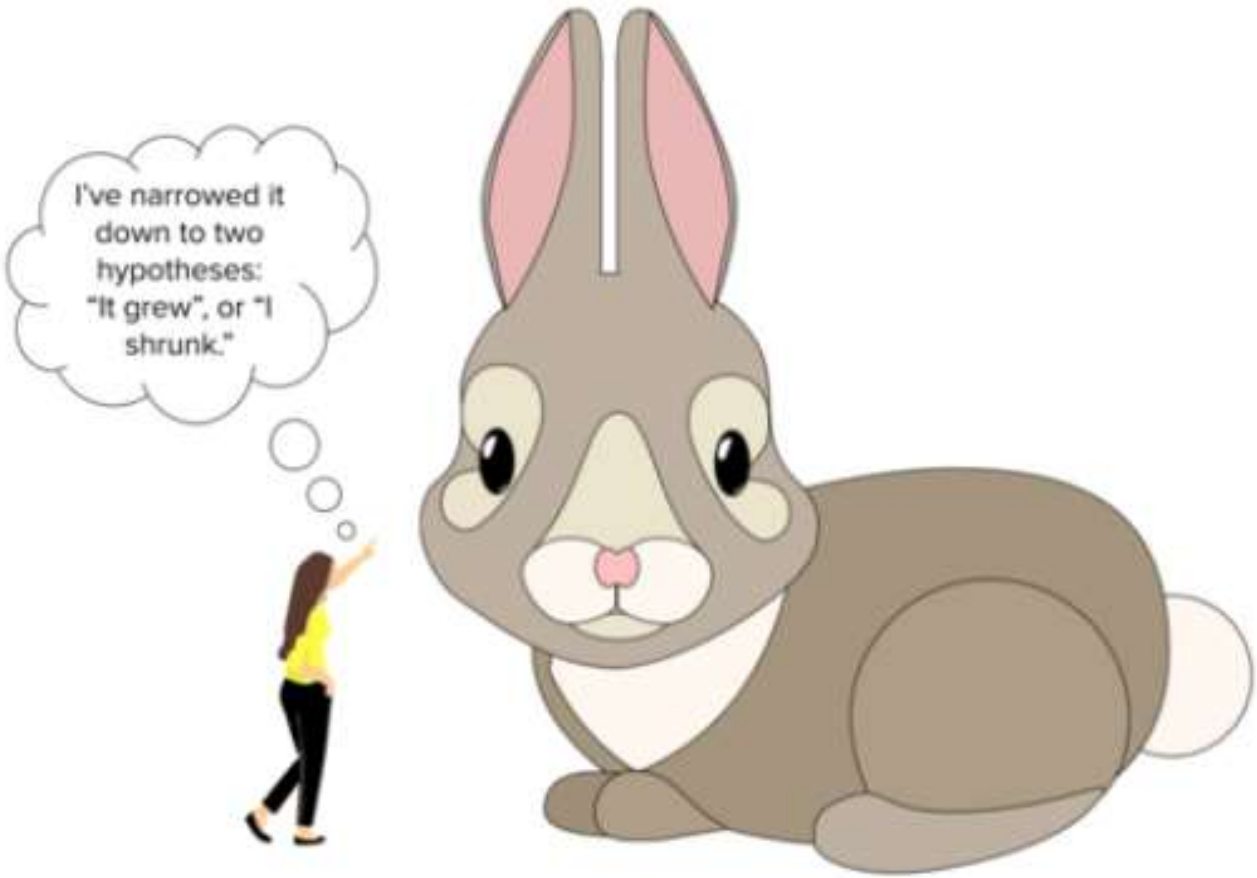
Introduction to statistics using R

Seminar series - session 3

Session 3 – Learning objectives

- What is a hypothesis?
- What is a prediction?
- Type I & type II statistical errors
- What is statistical modelling?
- Sampling methods
- Tips on building data bases

Hypothesis testing



Hypothesis testing

We are now making **inferences** about the **population** using information from our sample

Statistics

Descriptive

Organizing, summarizing, and presenting data

Inferential

Drawing conclusions about a population, based on data from a sample

What is a hypothesis

- A statement about how something works or why you observe something
- Can be tested with an **experiment**

H: birds take dust baths to maintain their plummage

What is a prediction?

- States the outcome of the experiment if the hypothesis is correct
- Tested with **models**

P1: Birds who dust bath more often have less oil excess on feathers

P2: Birds who dust bath have less ectoparasites

A hypothesis must be **testable** and **falsifiable**

Testability

"The gods walk among humans in an unobservable plan"

It may or may not be true, but it cannot be tested… So, not quite a hypothesis

Falsifiability

The Russell's Teapot: a small teapot is orbiting the sun between Earth and Mars.

Technically **testable**: we could send a spaceship that may come across it.

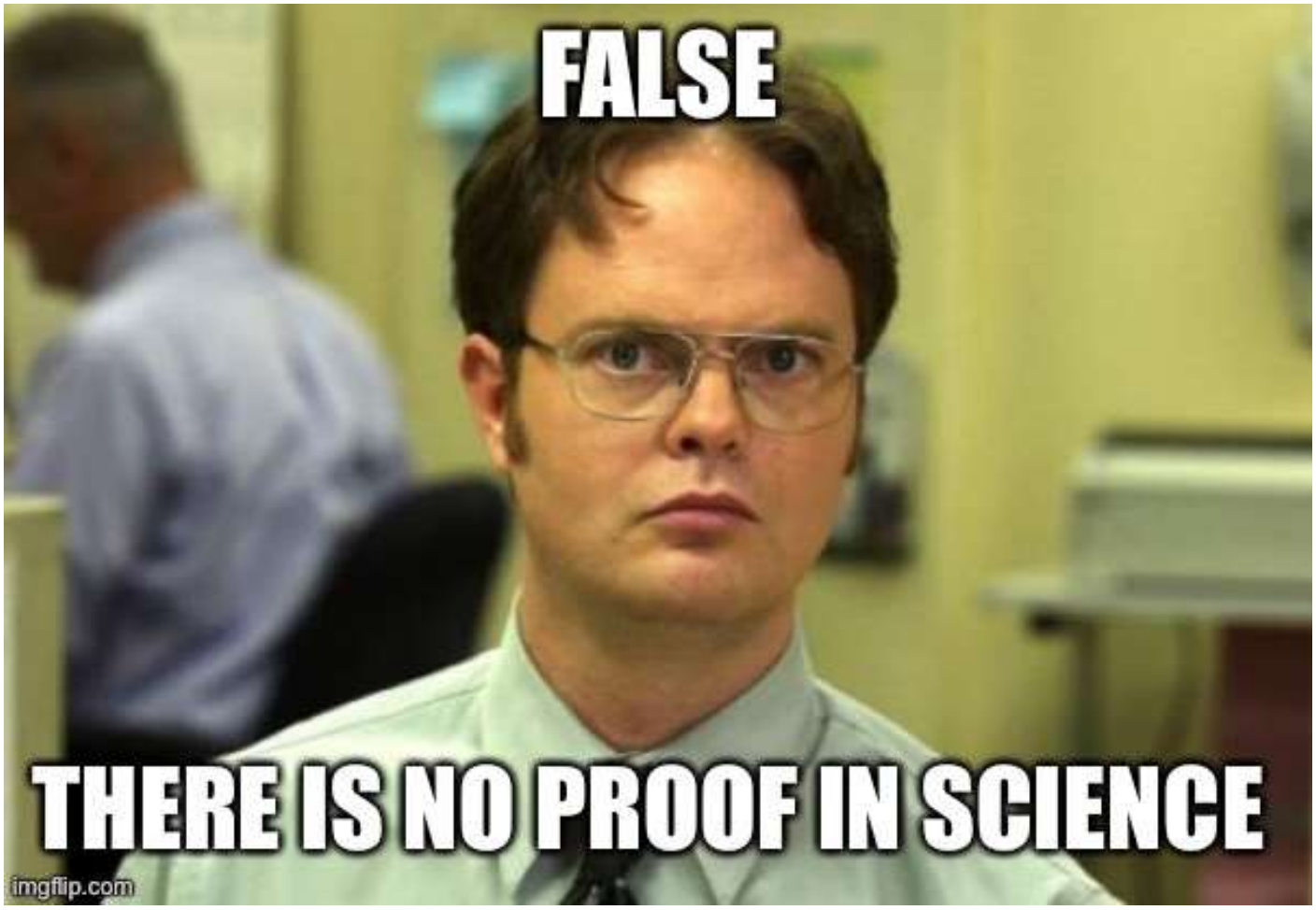
Falsifiability

But not **falsifiable**: the spaceship may never encounter it.
Yet, that doesn't mean that it's not there

Semantics!

With our study, we have proved that…

Semantics!



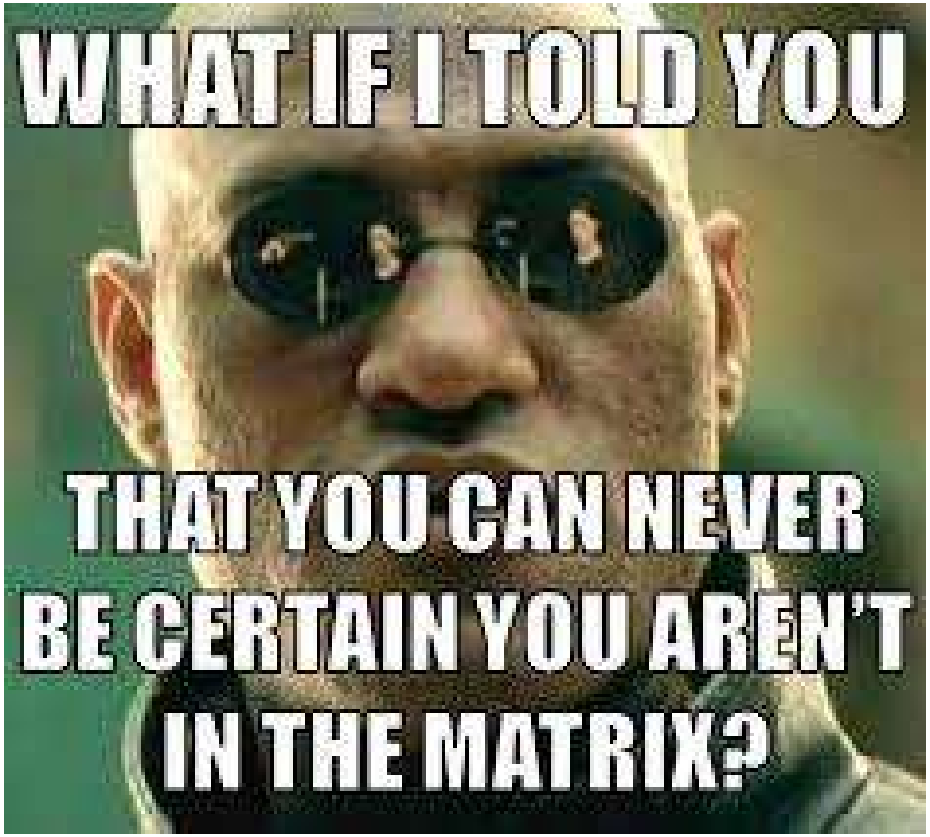
Semantics!

Why you should never use the word “proven” in science

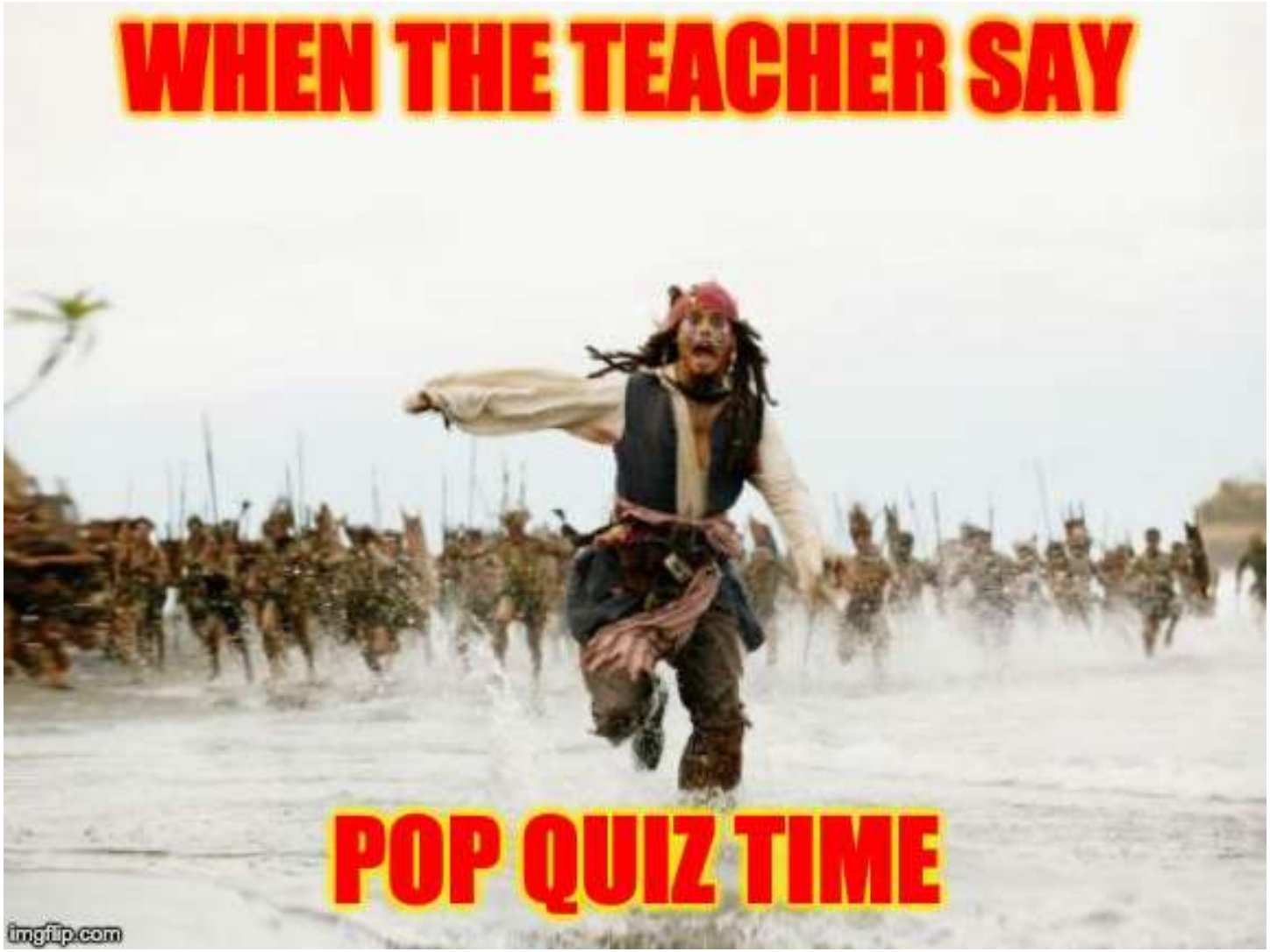
Semantics!

Proven is absolute

Semantics!


WHAT IF I TOLD YOU
THAT YOU CAN NEVER
BE CERTAIN YOU AREN'T
IN THE MATRIX?

Now let’s see if you get it!



What is a Statistical hypothesis?

The **statistical hypothesis** is the **scientific prediction**

What is a Statistical hypothesis?

Statistical hypothesis (or prediction) testing

H_0: A=B Null hypothesis

Observed difference or effect due to sampling/experimental error or chance

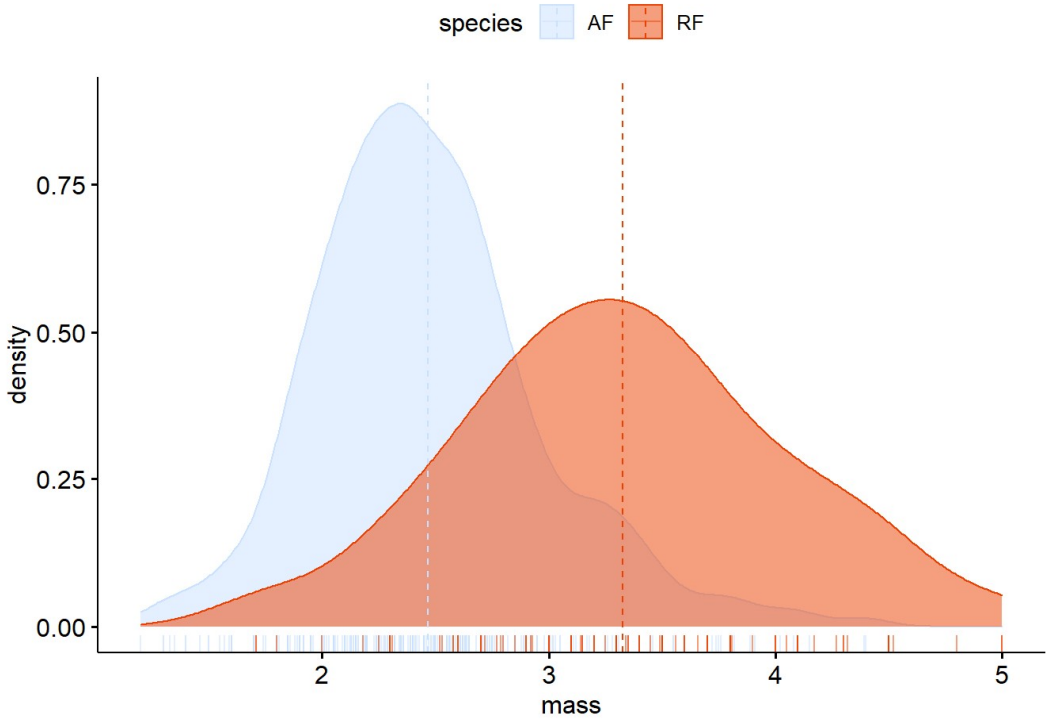
H_A: A≠B Alternative hypothesis

Chance alone cannot explain effect or difference observed

An example

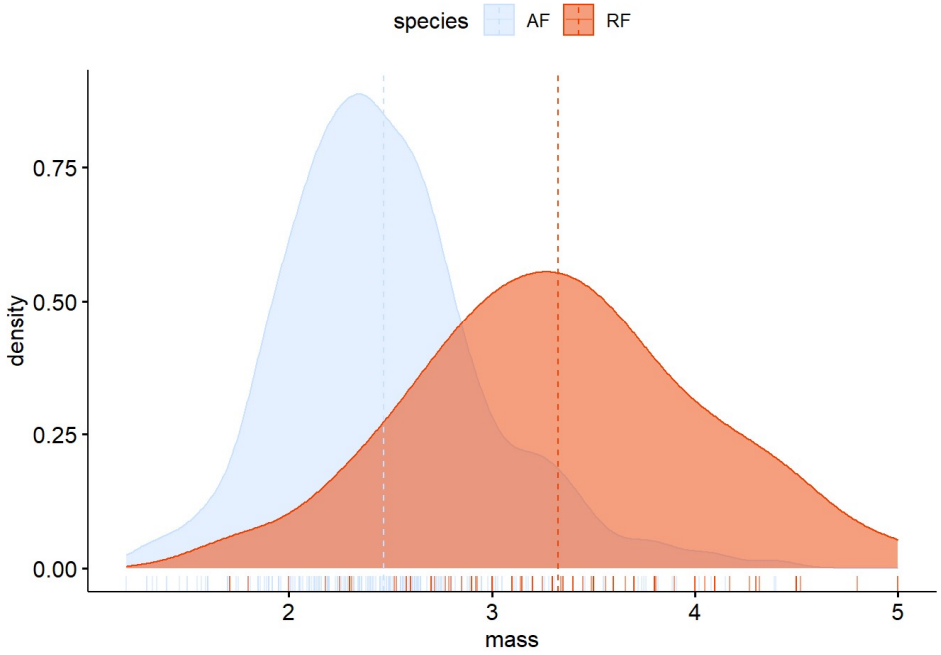
2 fox species: body mass comparison

RF: 3.3Kg [1.5 – 5]
AF: 2.5Kg [1.5 – 4.5]

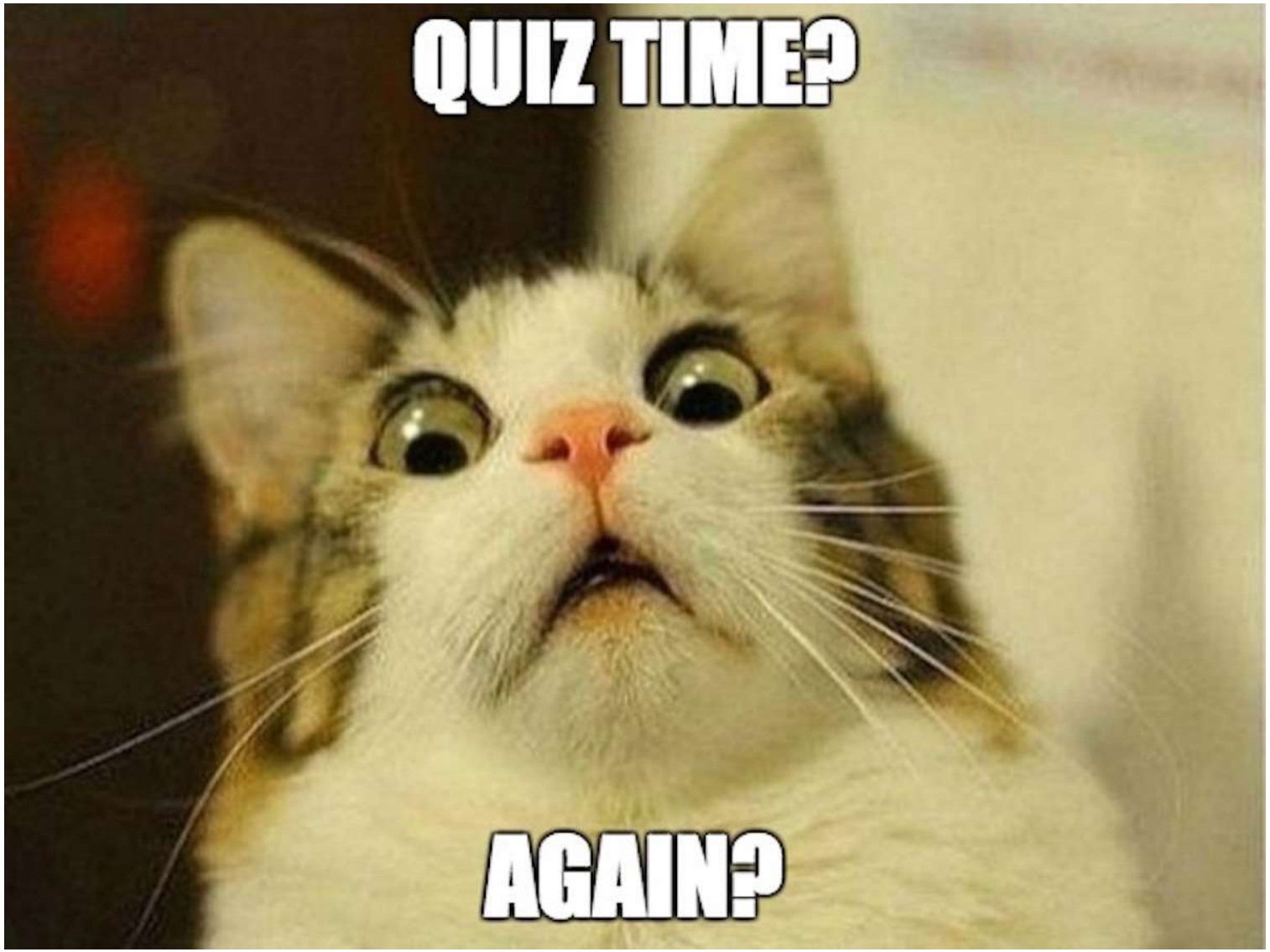


An example

- H_0: There is no difference in body mass between Arctic and red foxes
- H_A: body mass of Arctic and red foxes differs



Let’s see if you get it!



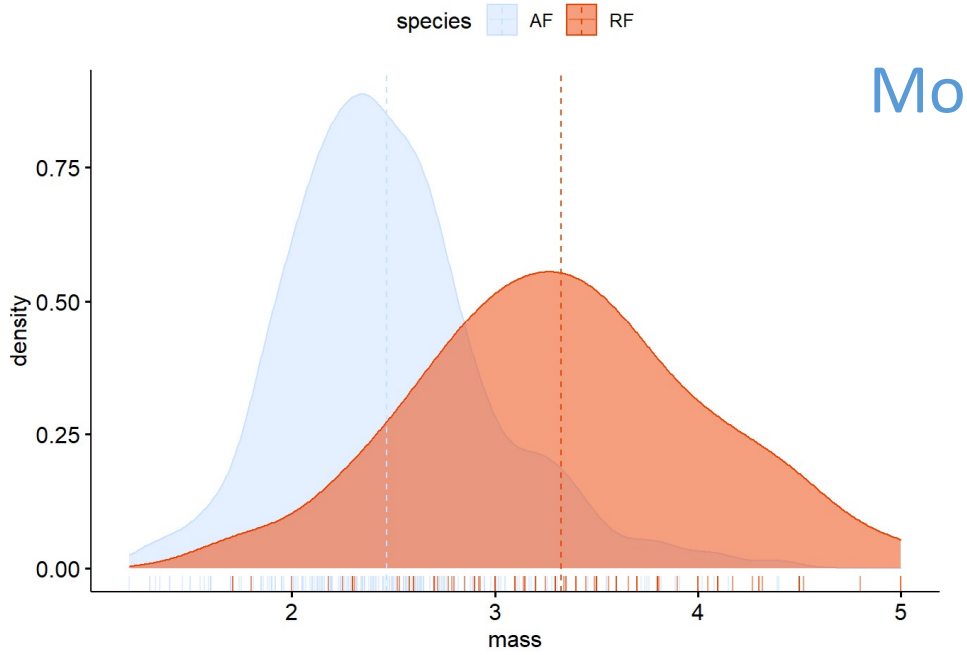
How to accept or reject H_0?

Many statistical tools, including but not limited to the famous P-value

- P-value
- confidence intervals of coefficients
- effect size
- Bayes factor
- AIC
- …

How to accept or reject H_0?

Let's use the p-value as an example:



Model: BM ~ species

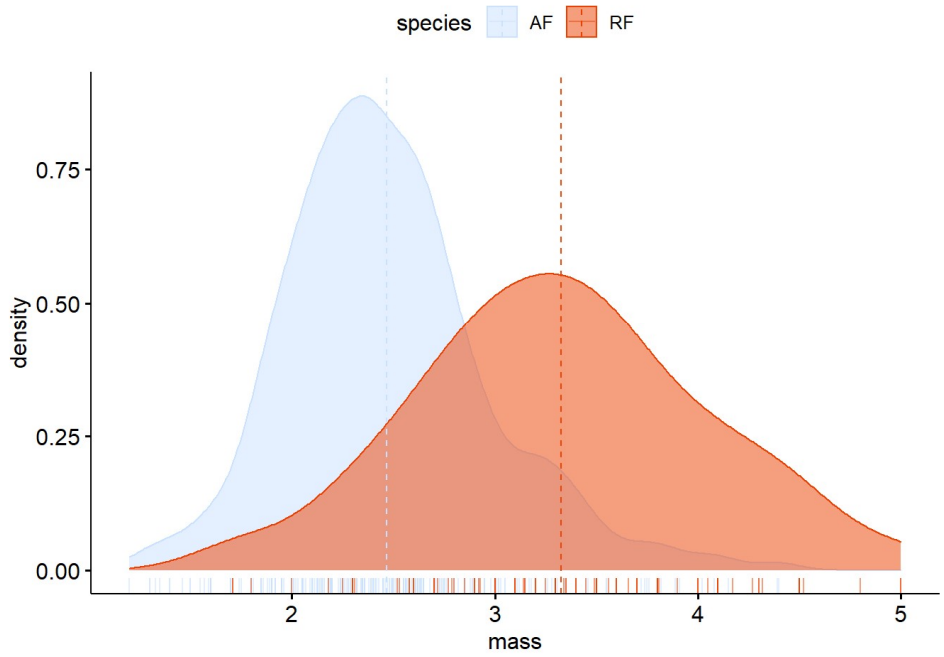
P = 0.003

α = 0.05

p-value is the probability given our model that the difference between mean body mass of red and Arctic foxes would be equal to or more extreme than the observed value

How to accept or reject H_0?

Let's use the p-value as an example:



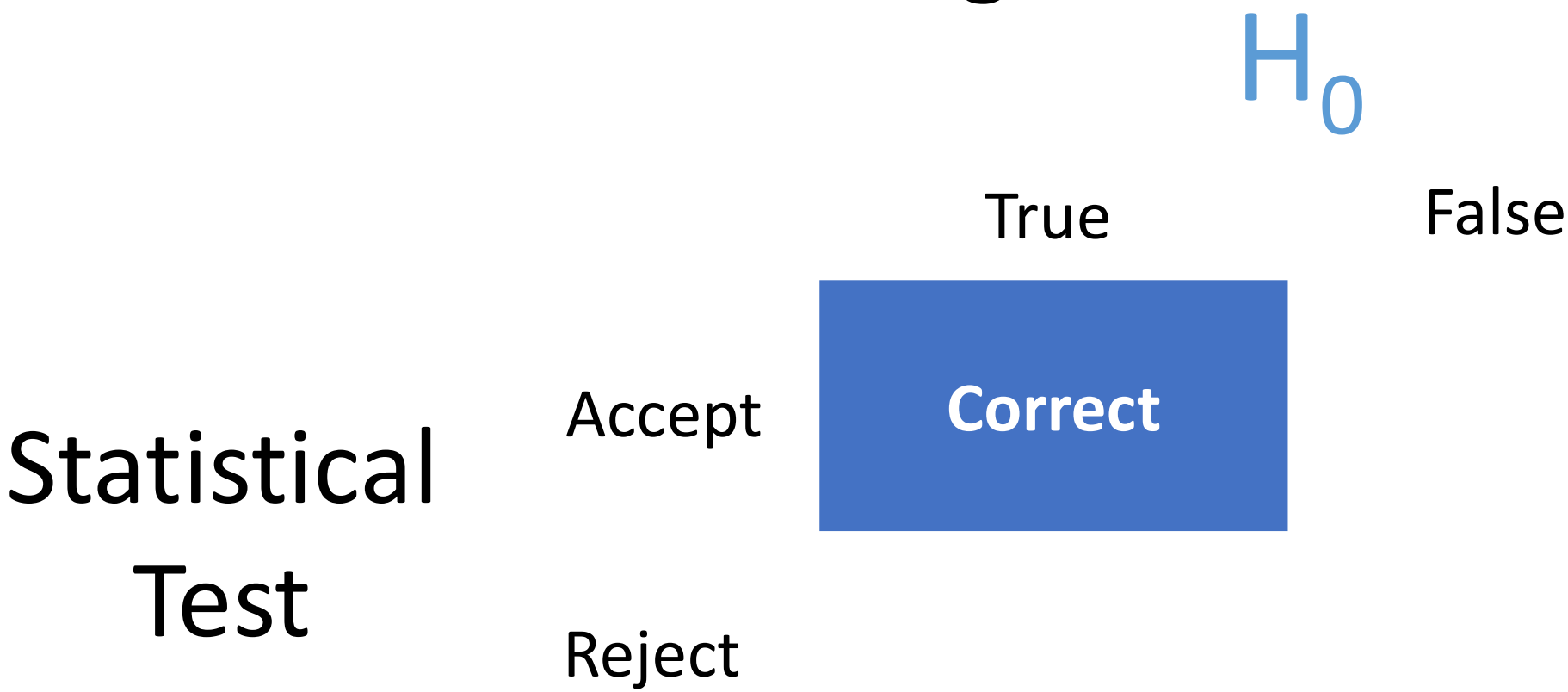
P = 0.003

Roughly, it's a measure of how likely is your data under H_0

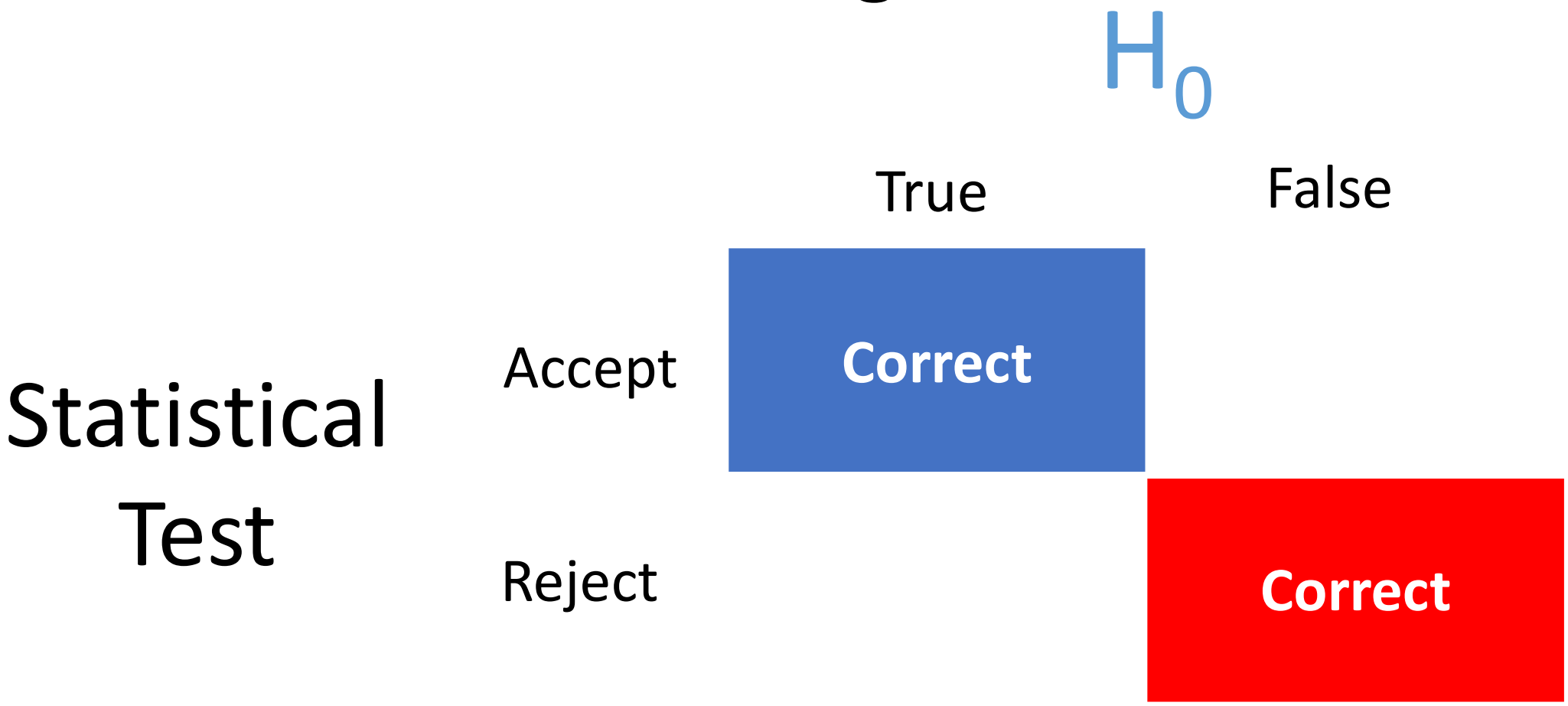
What if we're wrong?

Statistical Test	H_0 True	H_0 False
Accept		
Reject		

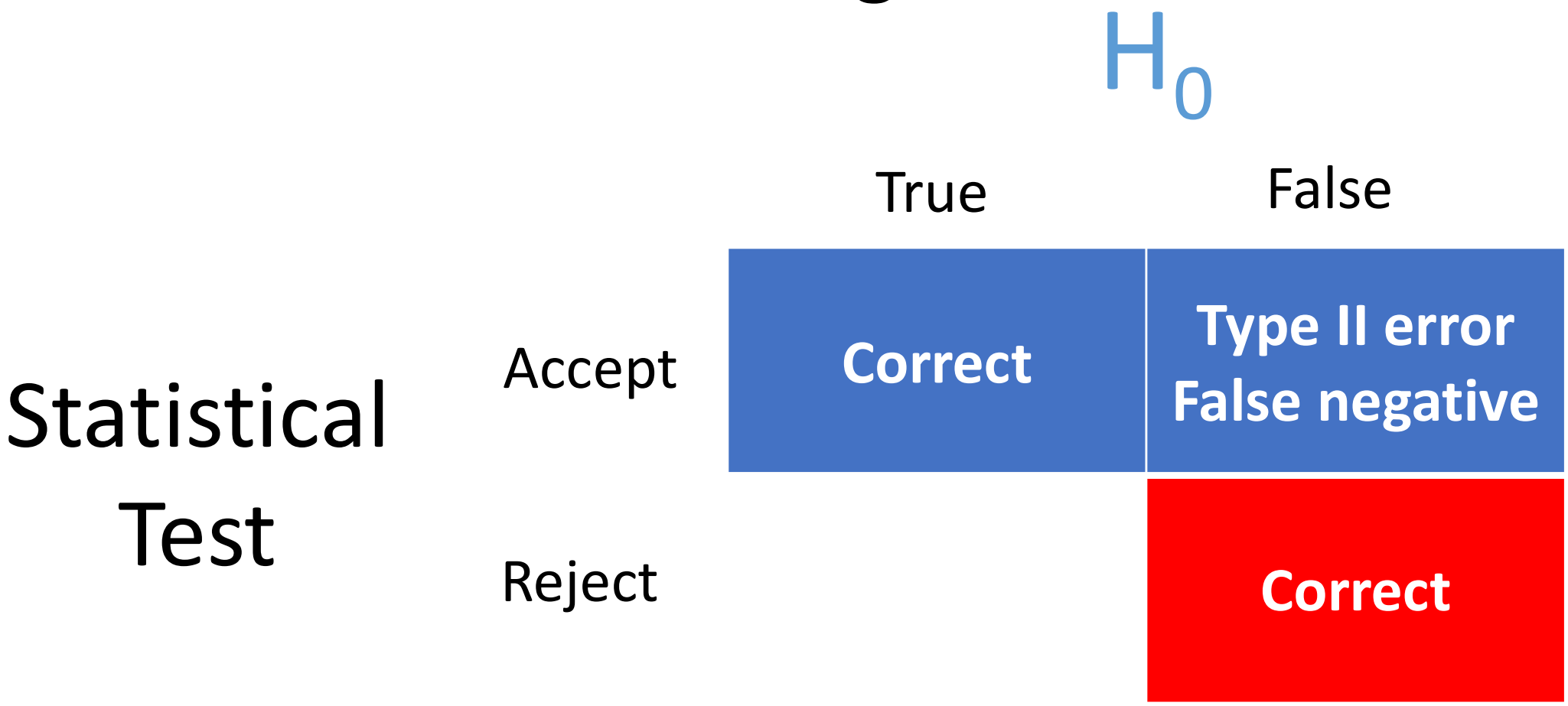
What if we're wrong?



What if we're wrong?



What if we're wrong?



Type II is the probability of not finding a **pattern that's there.**

What if we're wrong?

Statistical Test	H_0 True	H_0 False
Accept	**Correct**	**Type II error False negative**
Reject	**Type I error False positive**	**Correct**

Type II is the probability of not finding a **pattern that's there.**

Type I is the probability of finding a **pattern** that's **not there.**

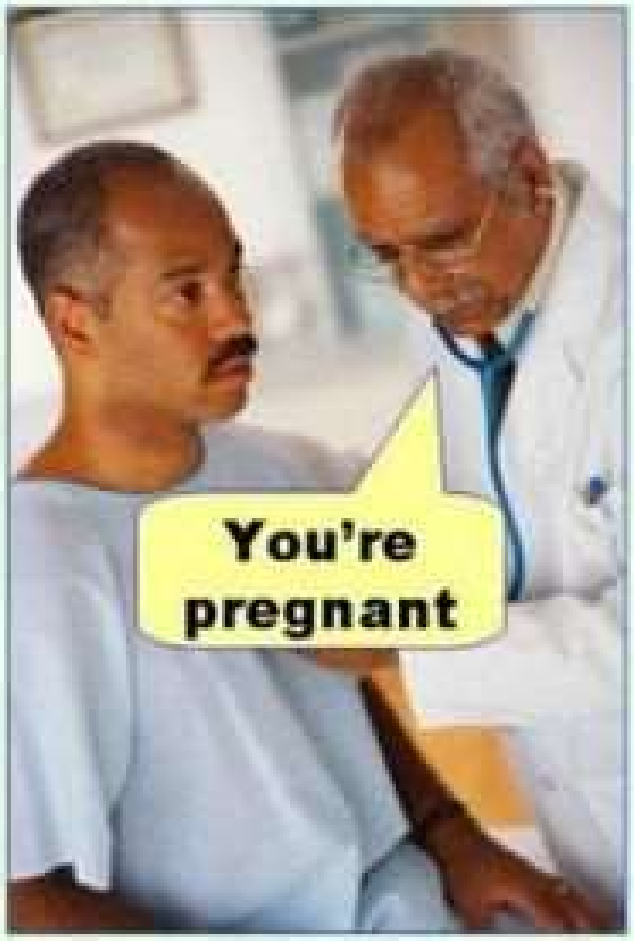
What if we're wrong?

Statistical Test	H_0 True	H_0 False
Accept	**Correct**	**Type II error False negative**
Reject	**Type I error False positive**	**Correct**

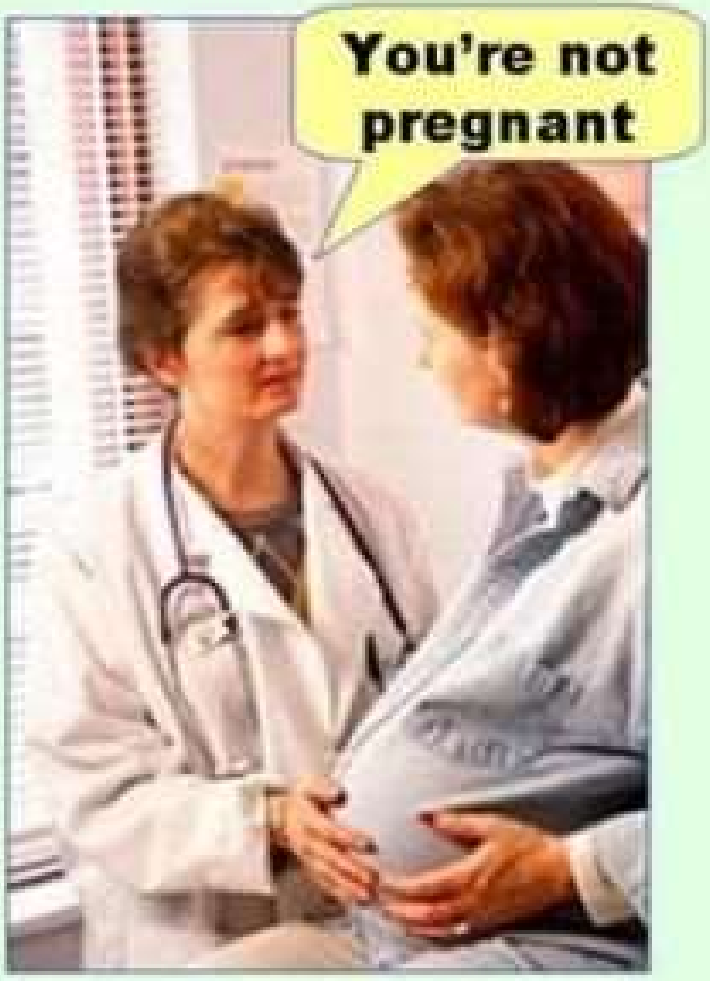
Ideally, type I (or α) and type II (or β) errors will both be very small, and similar

Statistical error:

Type I error
(false positive)

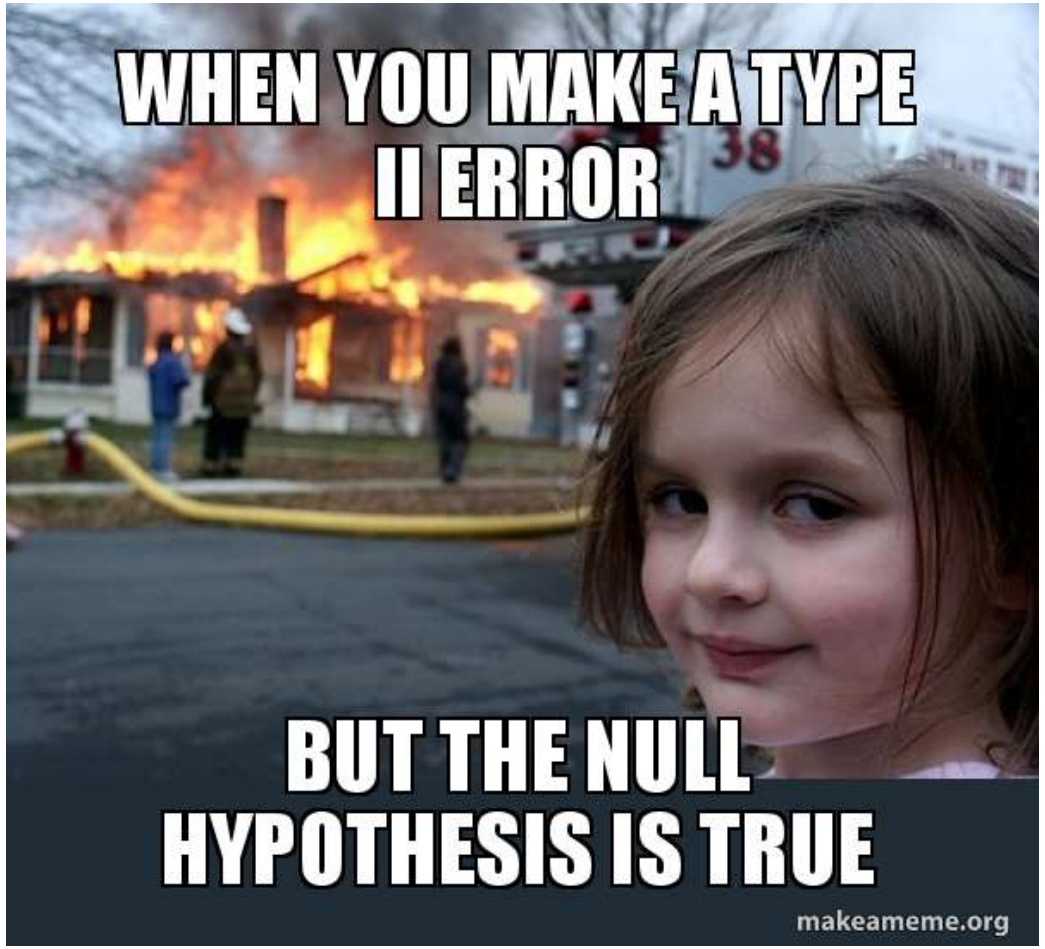


Type II error
(false negative)



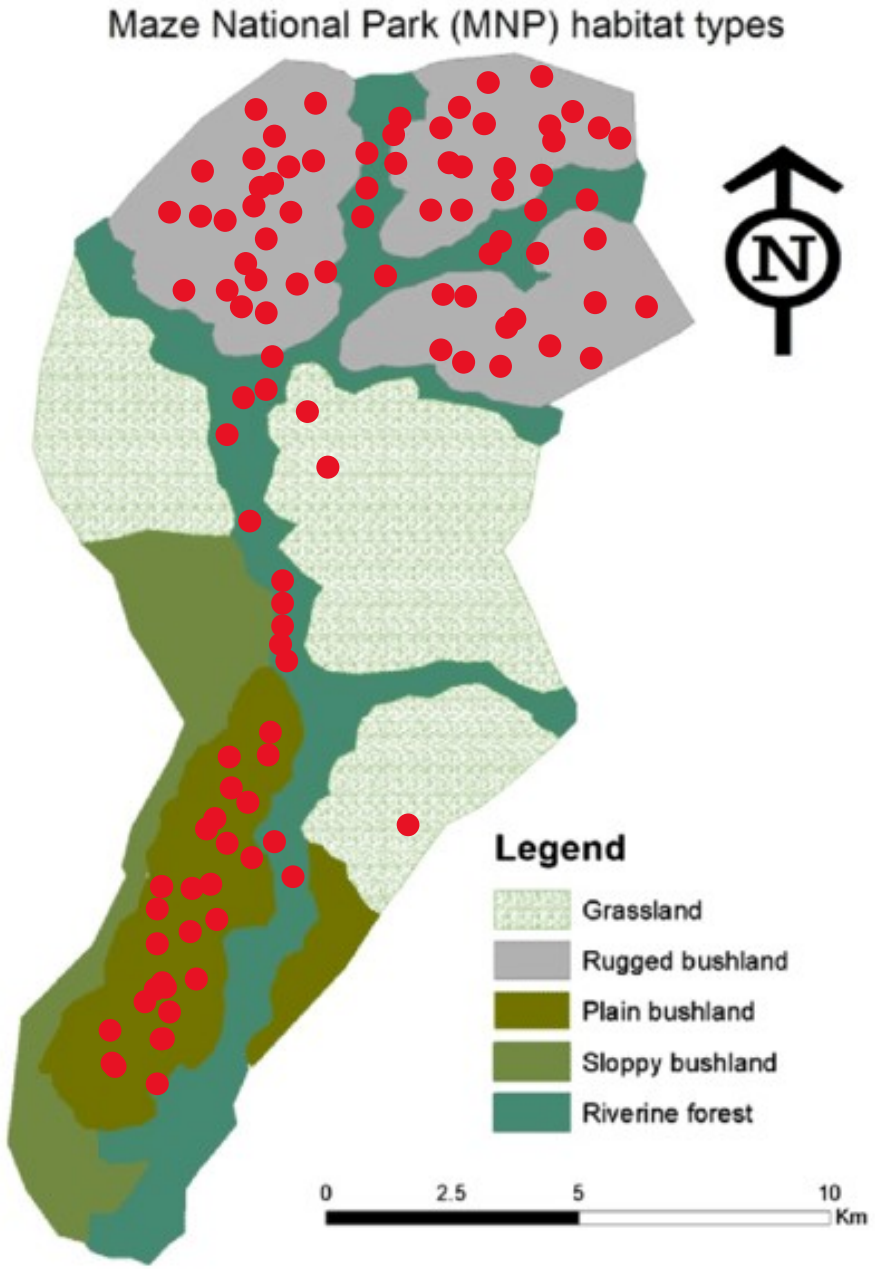
Which one is worse?

Depends on your objectives!



Which one is worse?

You create a protected area for species A, so you record A's habitat use to include enough important habitat

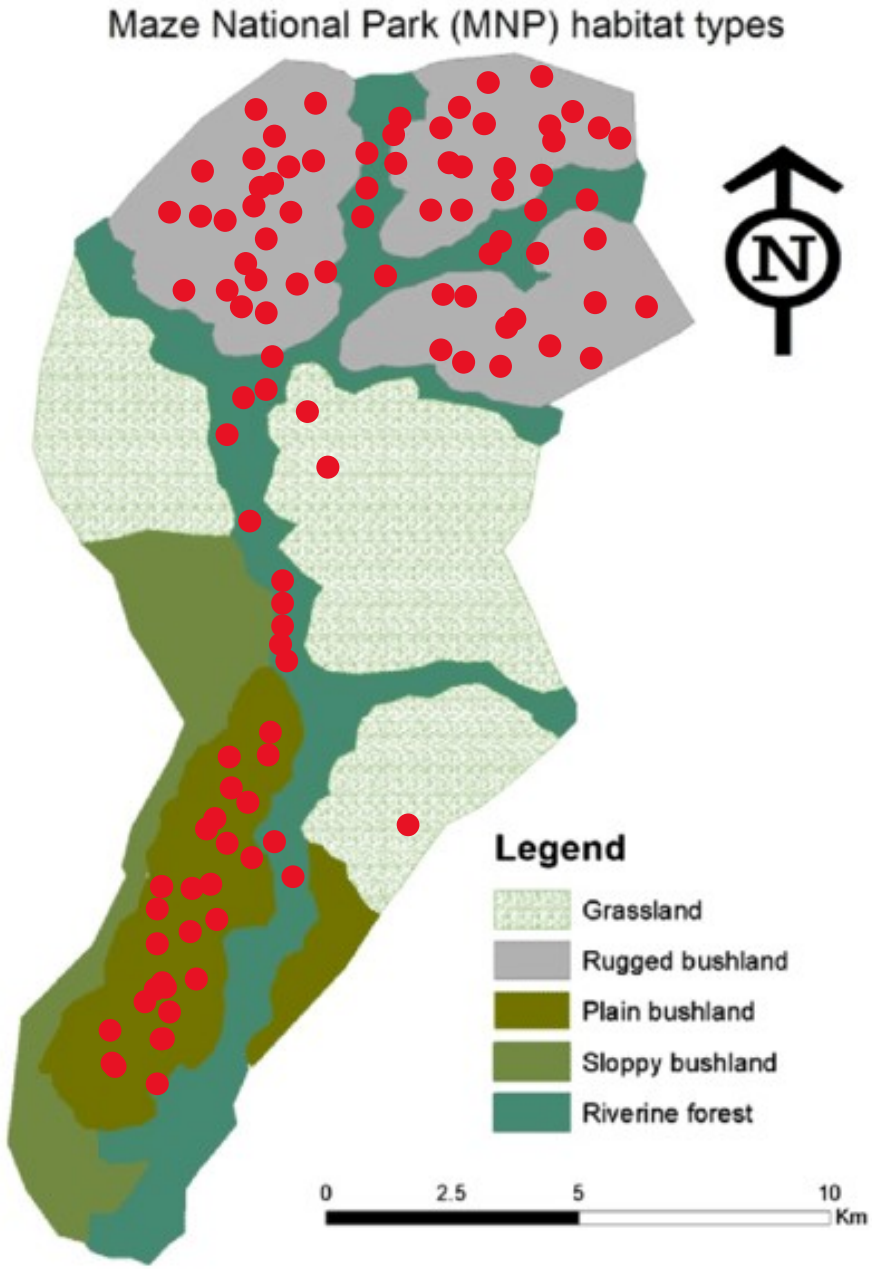


Modified from Tamrat et al. (2020)

Which one is worse?

A spends little time in Grassland, yet it's a key habitat for A reproduction.

A has many occurrences in Riverine forest, yet A doesn't depend on this habitat to persist



Modified from Tamrat et al. (2020)

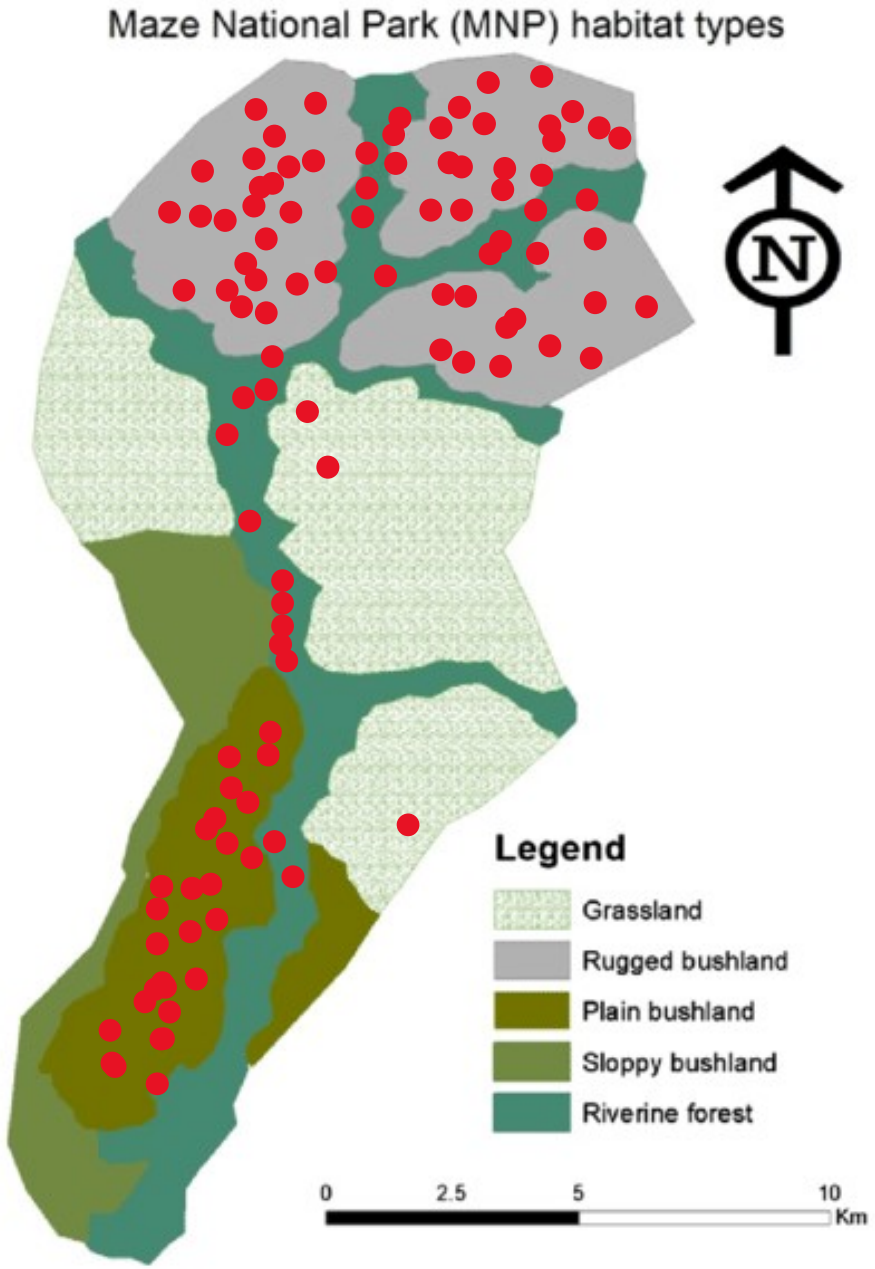
H_0: "Habitat" has no effect on A's persistence

Type I: You include Riverine forest, which has no effect on A's persistence

Reject H_0 when it's true

Type II: You don't include grassland, which is key to A's persistence

Accept H_0 when it's false



Modified from Tamrat et al. (2020)

Which one is worse?

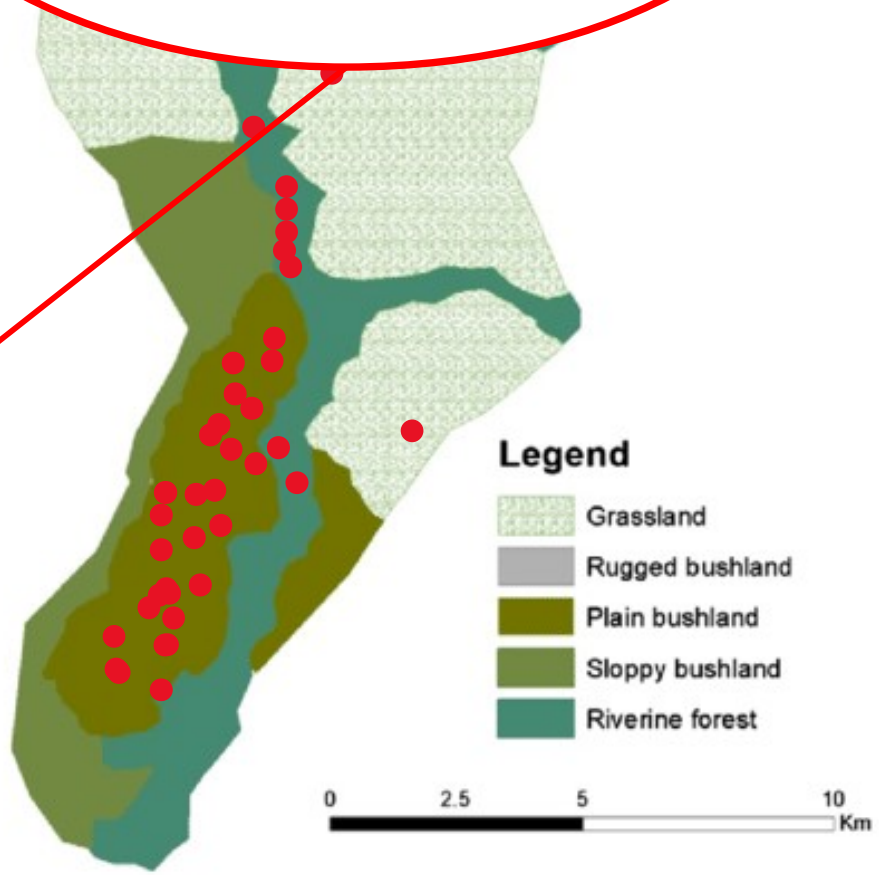
That's the worse, here

Type I: You include Riverine forest, which has no effect on A's persistence

Reject H_0

Type II: You don't include grassland, which is key to A's persistence

Accept H_0 when it's false



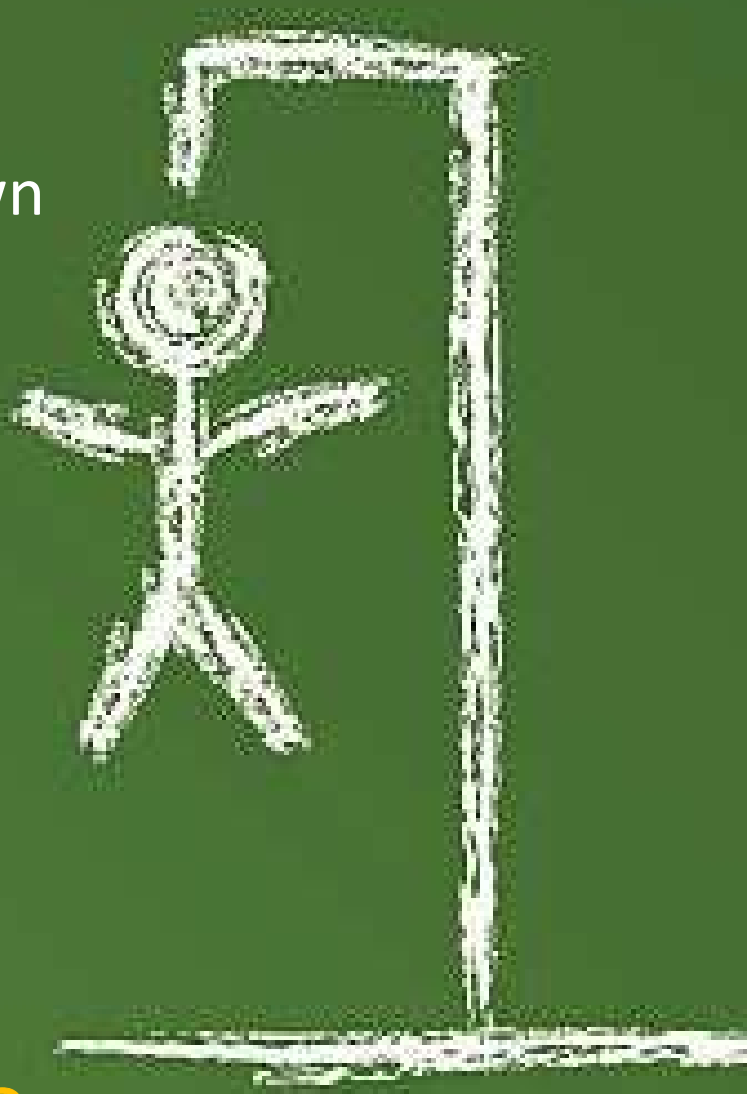
Modified from Tamrat et al. (2020)

William Wallace is put on trial for high treason

H_0: WW is not guilty

Type I: WW is condemned to be hanged, drawn and quartered, although he was not guilty

Type II: WW is freed although he's guilty



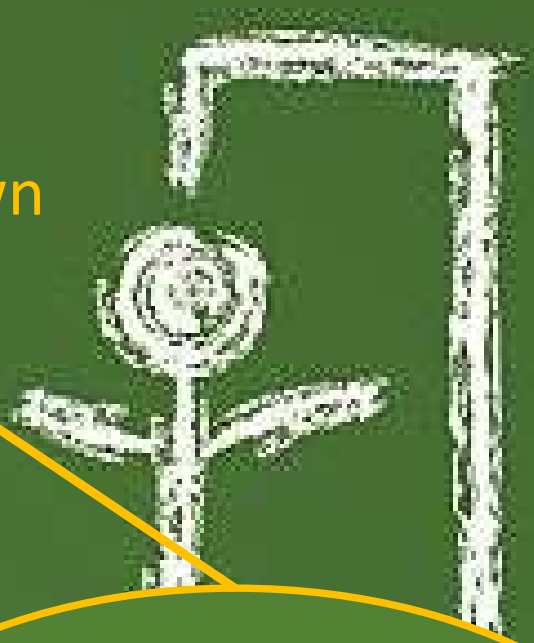
Which is worse?

William Wallace is put on trial for high treason

H_0: WW is not guilty

Type I: WW is condemned to be hanged, drawn and quartered, although he was not guilty

Type II: WW is freed although he's guilty



This time, type I is worse!

What is statistical modelling?

- mathematical relationship between random and non-random variables
- Applying statistical analysis to a dataset
- Ex.: Simple Linear regressions

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

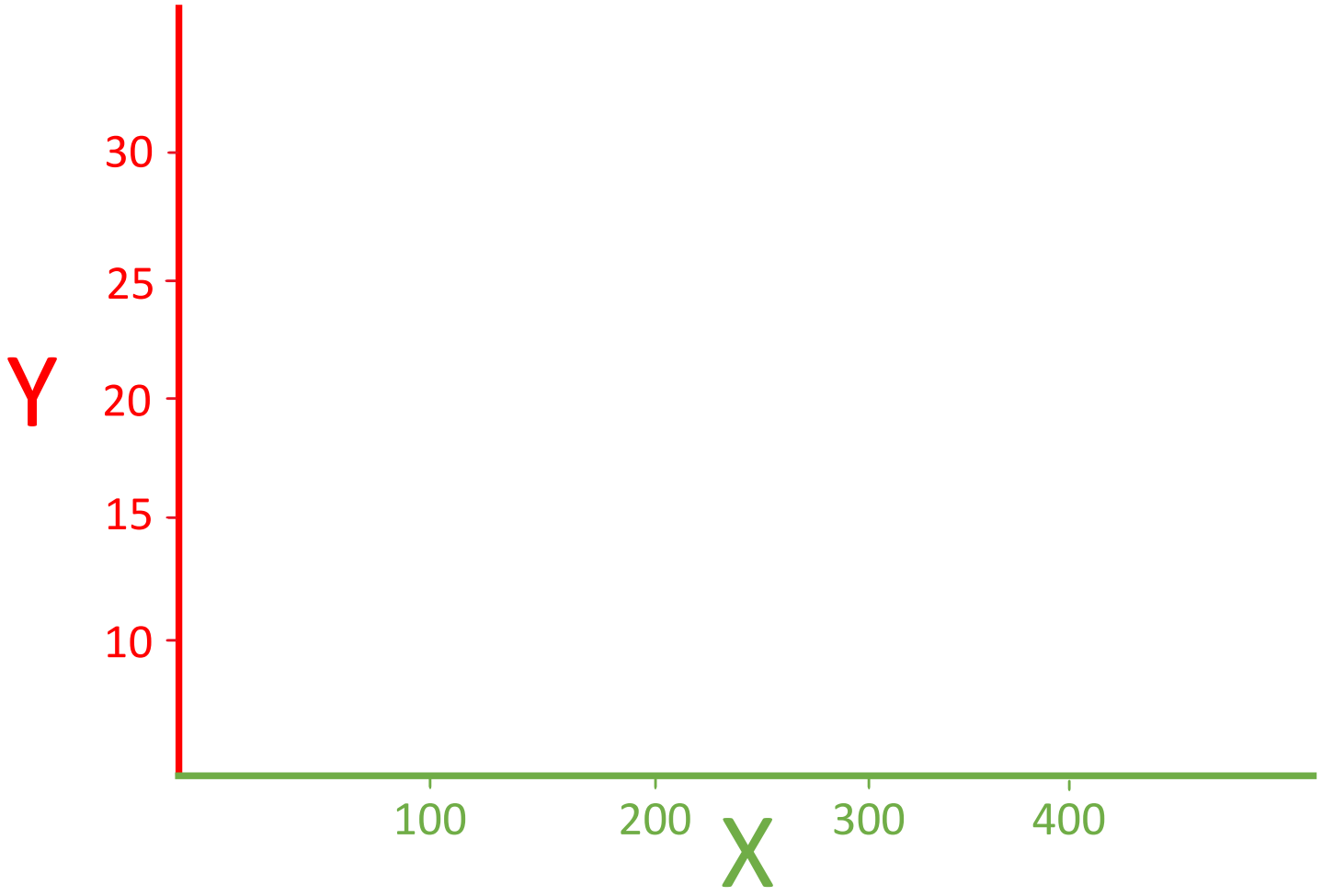
Y: response variable

β_0: intercept

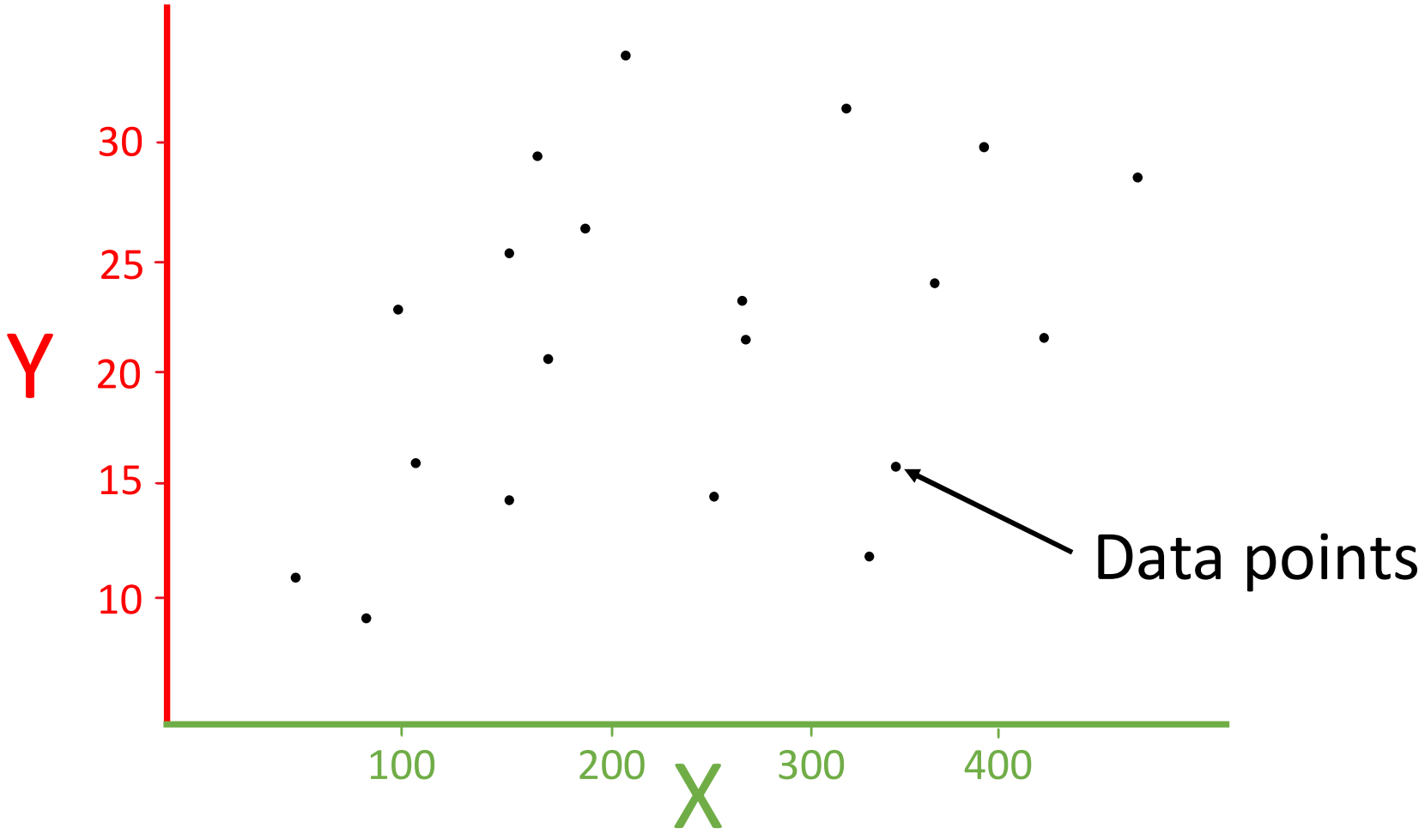
β_1: regression coefficient – how much Y changes for each unit increase in the explanatory variable X

ε: error term – margin of error in the model

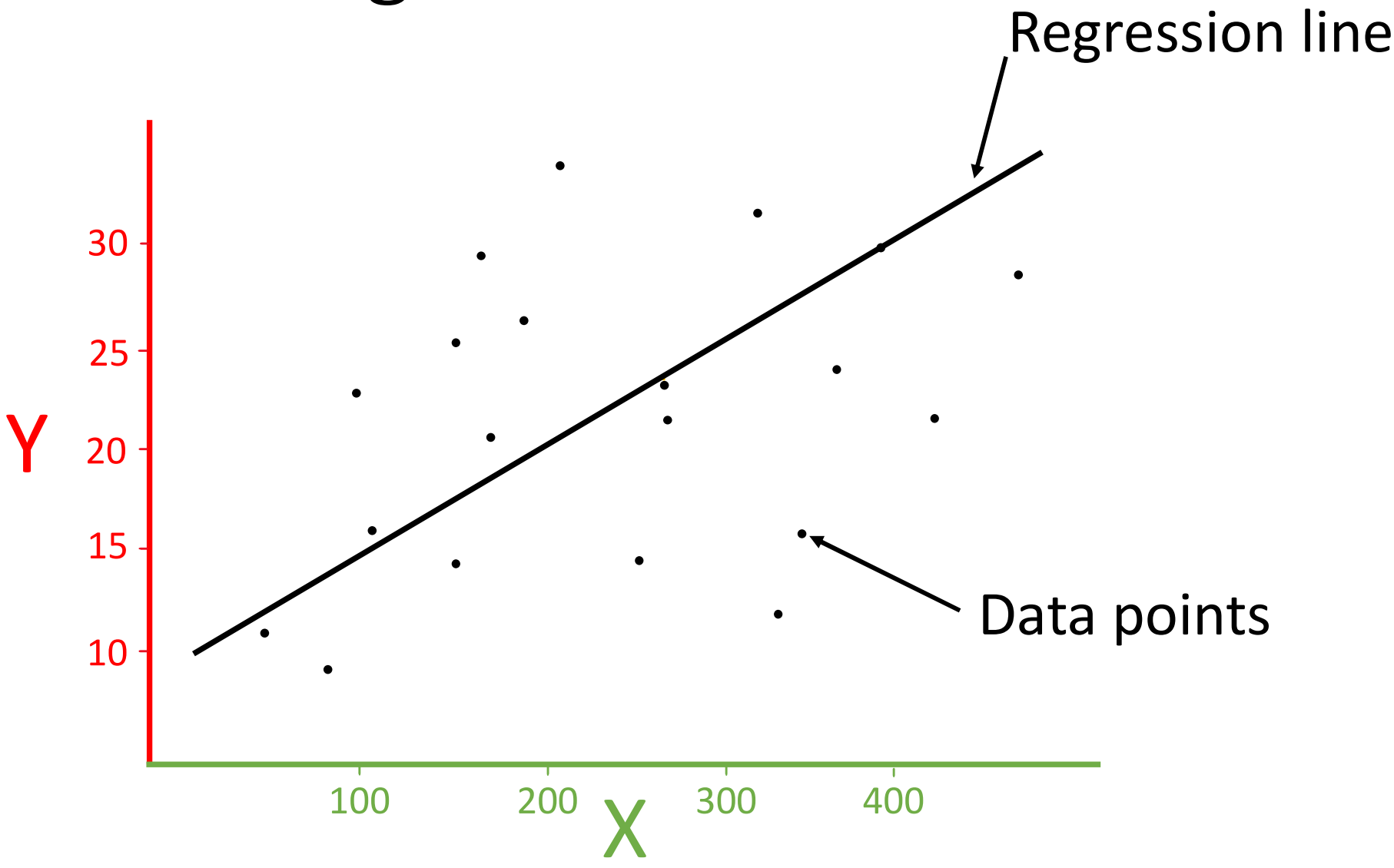
Linear regression



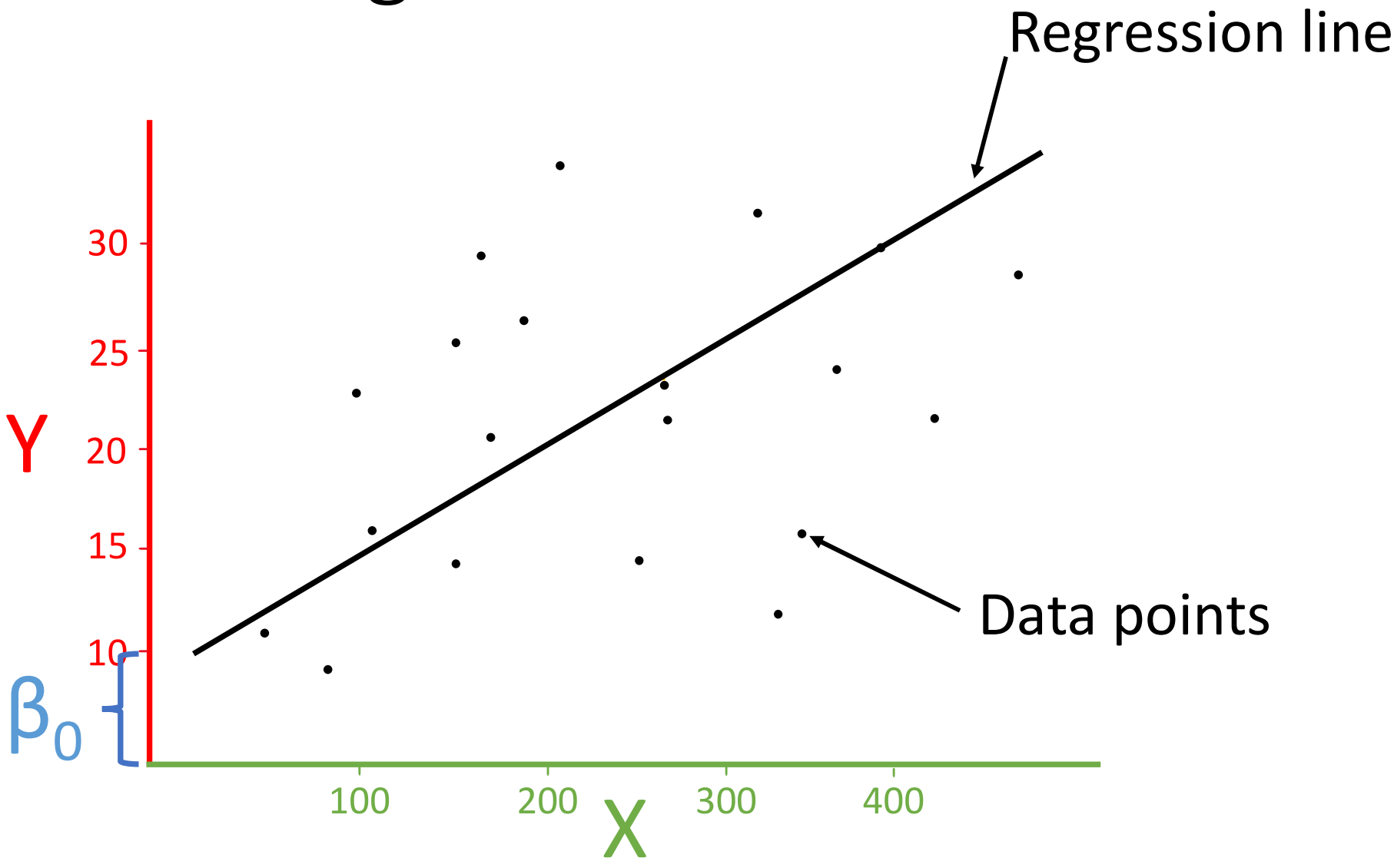
Linear regression



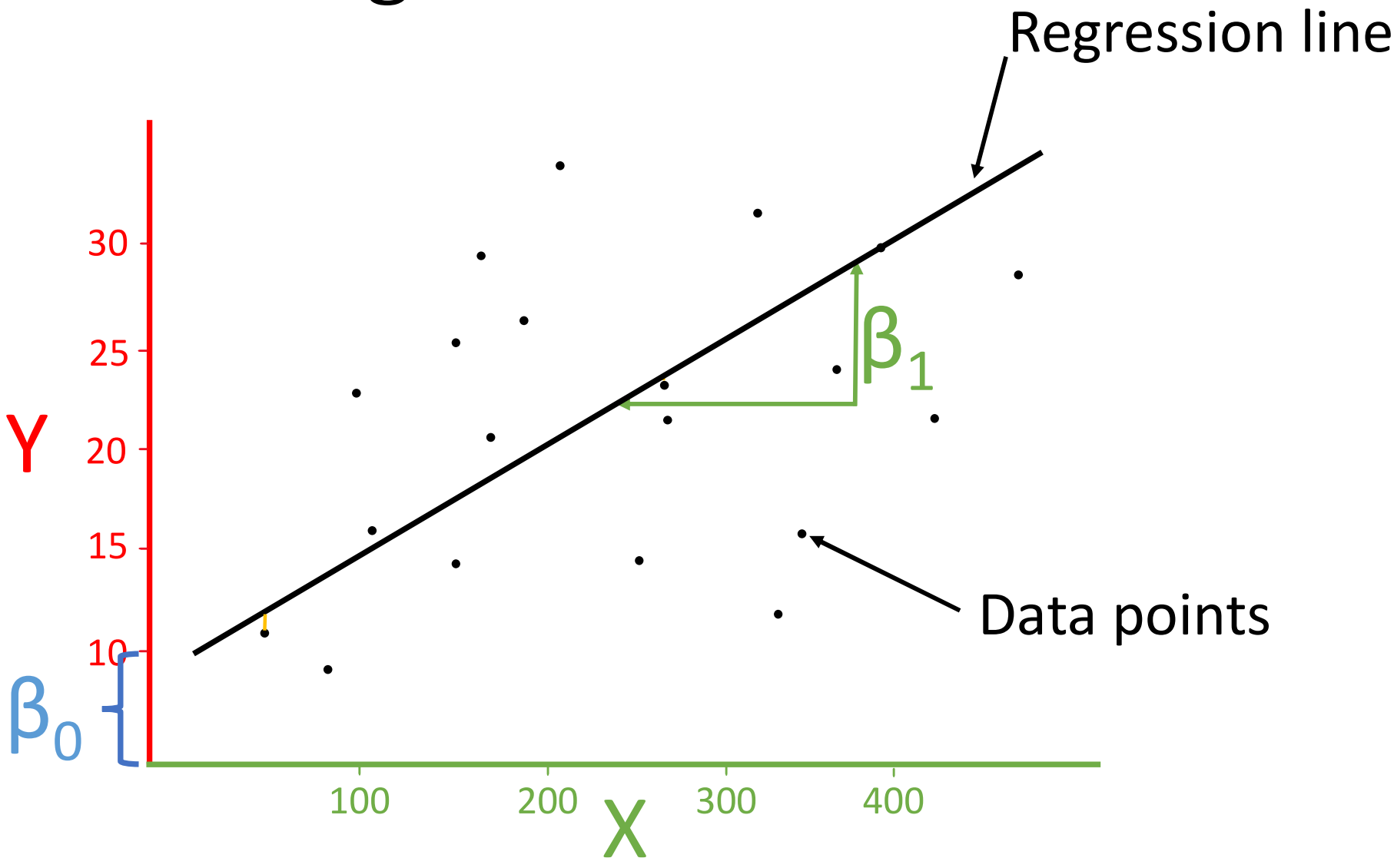
Linear regression



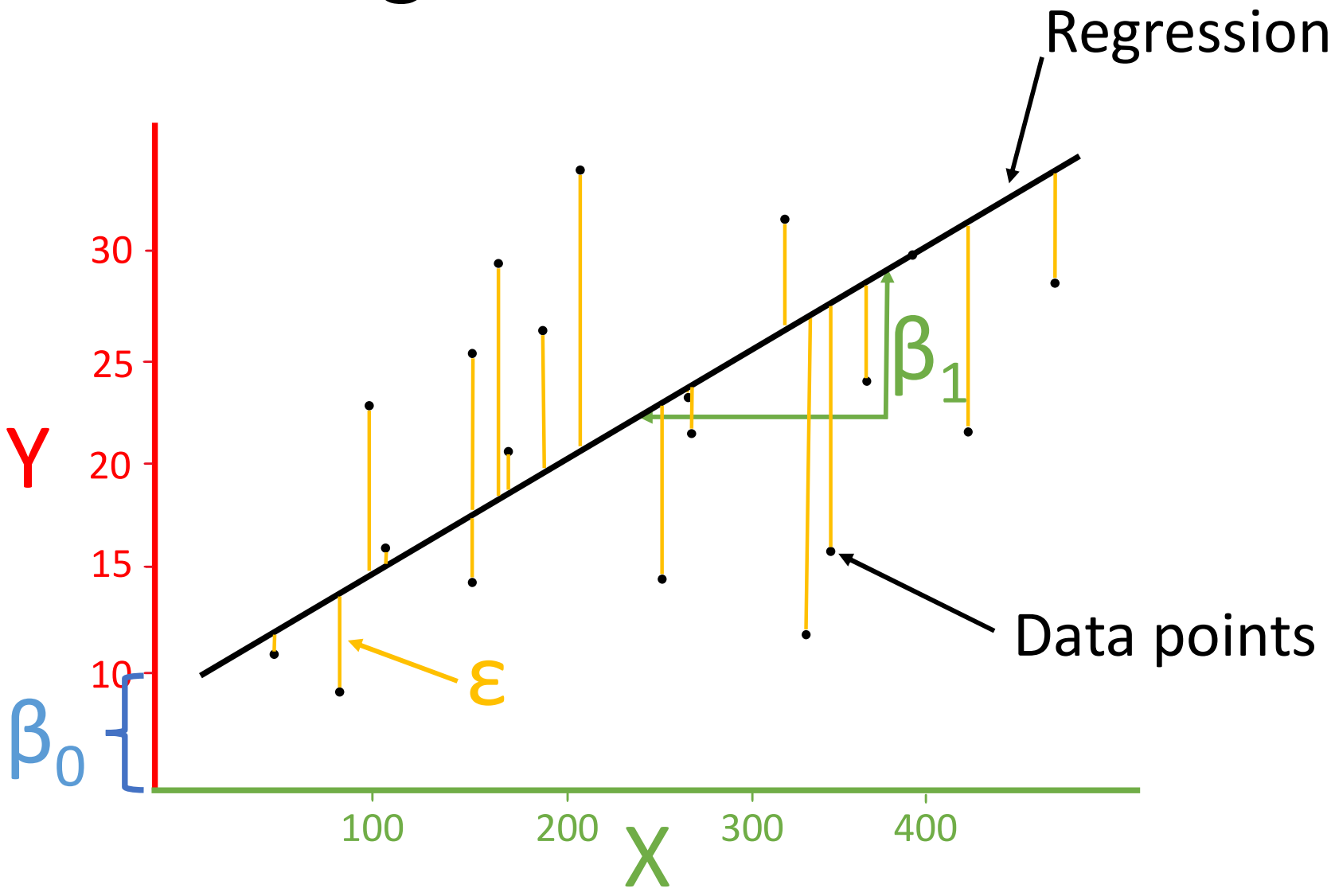
Linear regression



Linear regression

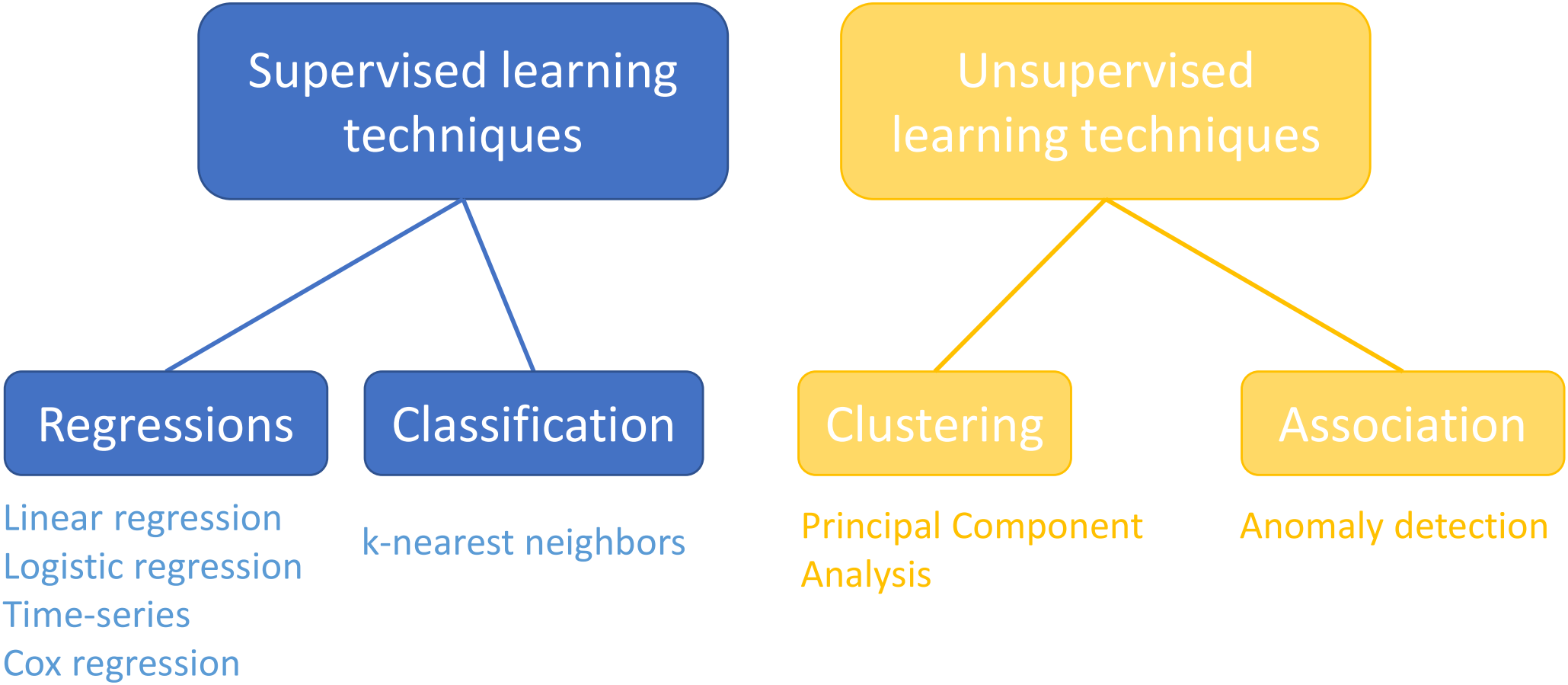


Linear regression

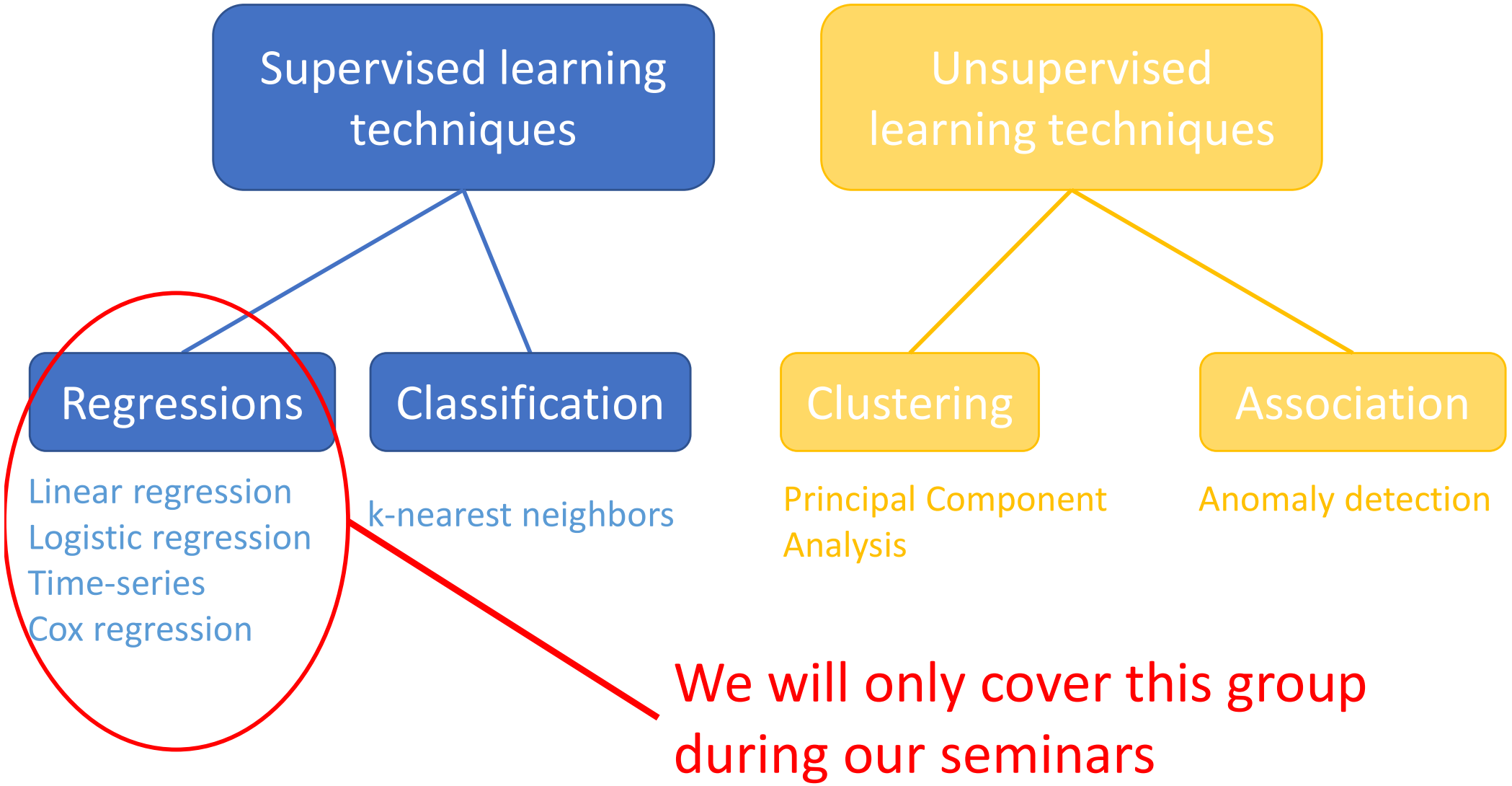


$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Other types of statistical models



Other types of statistical models



Sampling

Collect necessary data to test your predictions

Data must be:

- Replicated
- Independent

Most statistical test assume that occurrence of an observation does not affect the probability of another.

- Random

Easier said than done. Ensures estimates are not biased.

Replicates

Repetition of an experiment or observation in the same or similar conditions

- Populations without variability virtually don't exist
- You need to quantify that variability to draw sensible conclusions
- You need enough replicates to quantify biological random variability
- Statistical tests rely on replication
- **Ensures reliability of the estimates.**

Ex.: Replicated in space

Bird abundance in Forest?

Transect B

Transect A

Ex.: Replicated in time and space

Transect A

Transect B

Bird abundance in Forest?

Total sample size

Total sample size refers to the entire scope of your study

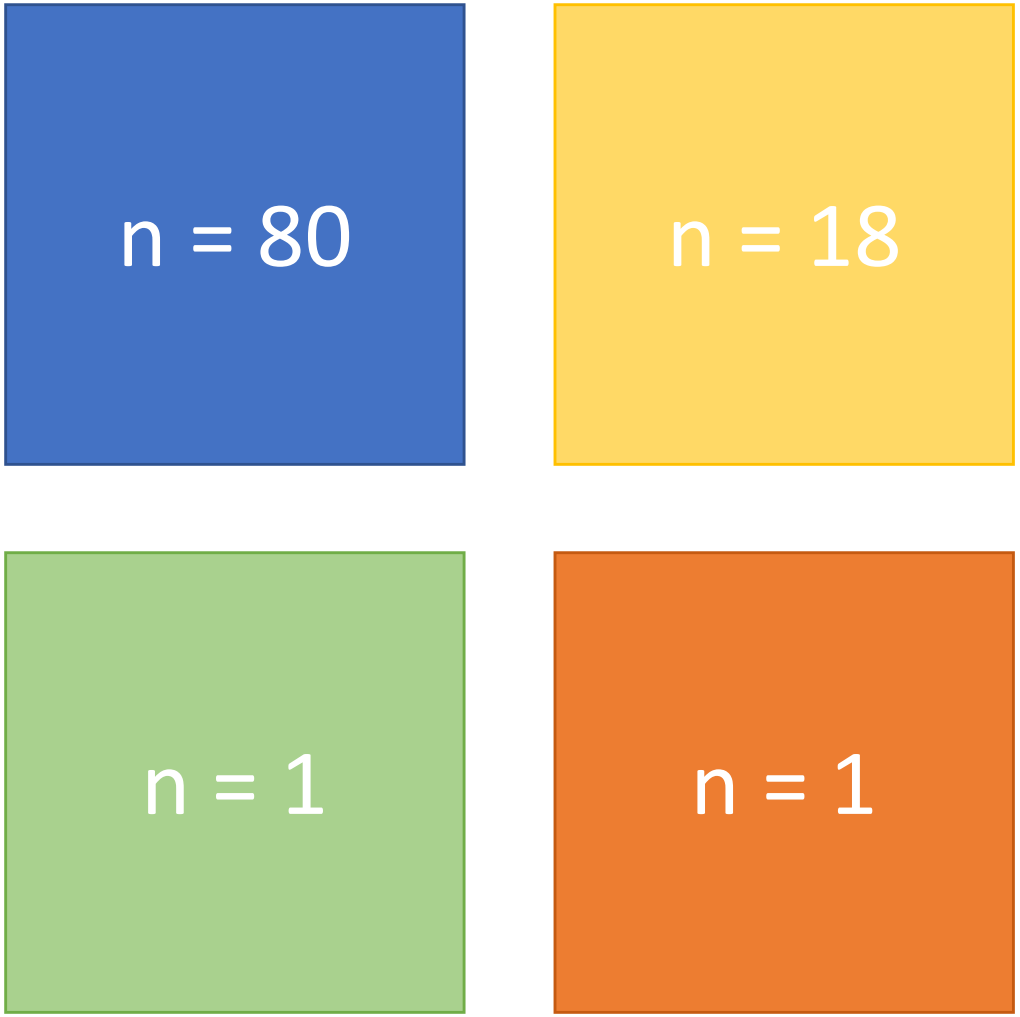
- Importance of balanced designed!
  - Effect of factor with low n may be hidden by factor with high n
  - Power of the test may be lower
- **Also ensures reliability of the estimates**

High total sample size does not ensure the design was good

4 treatments, N = 100

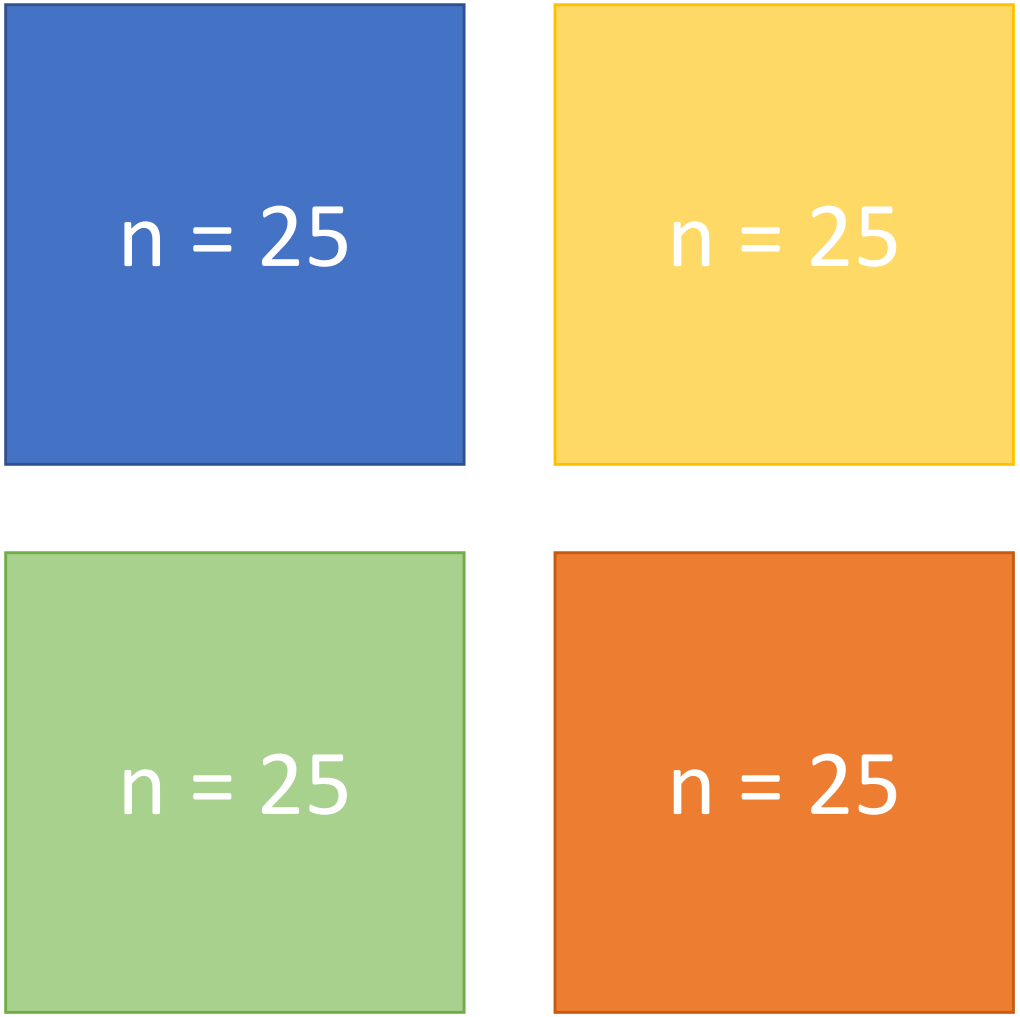
Is it a good sample size?

High total sample size does not ensure the design was good

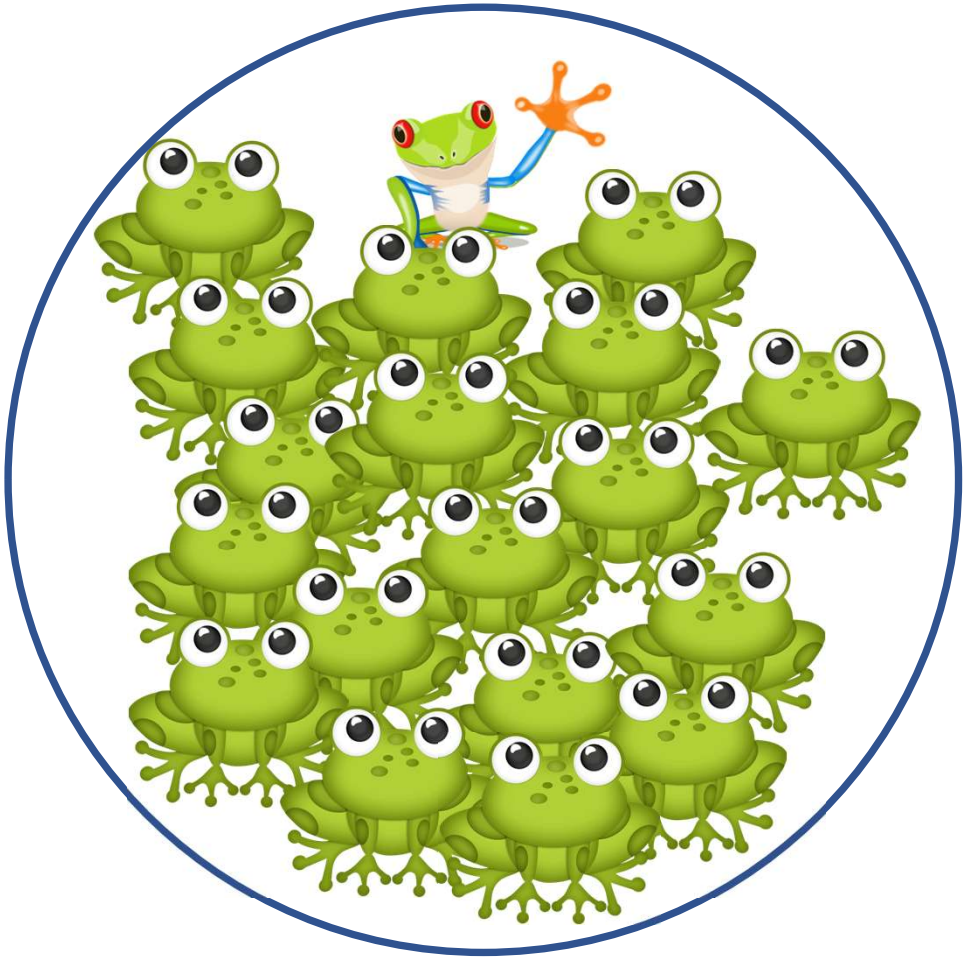


The design is highly imbalanced

High total sample size does not ensure the design was good

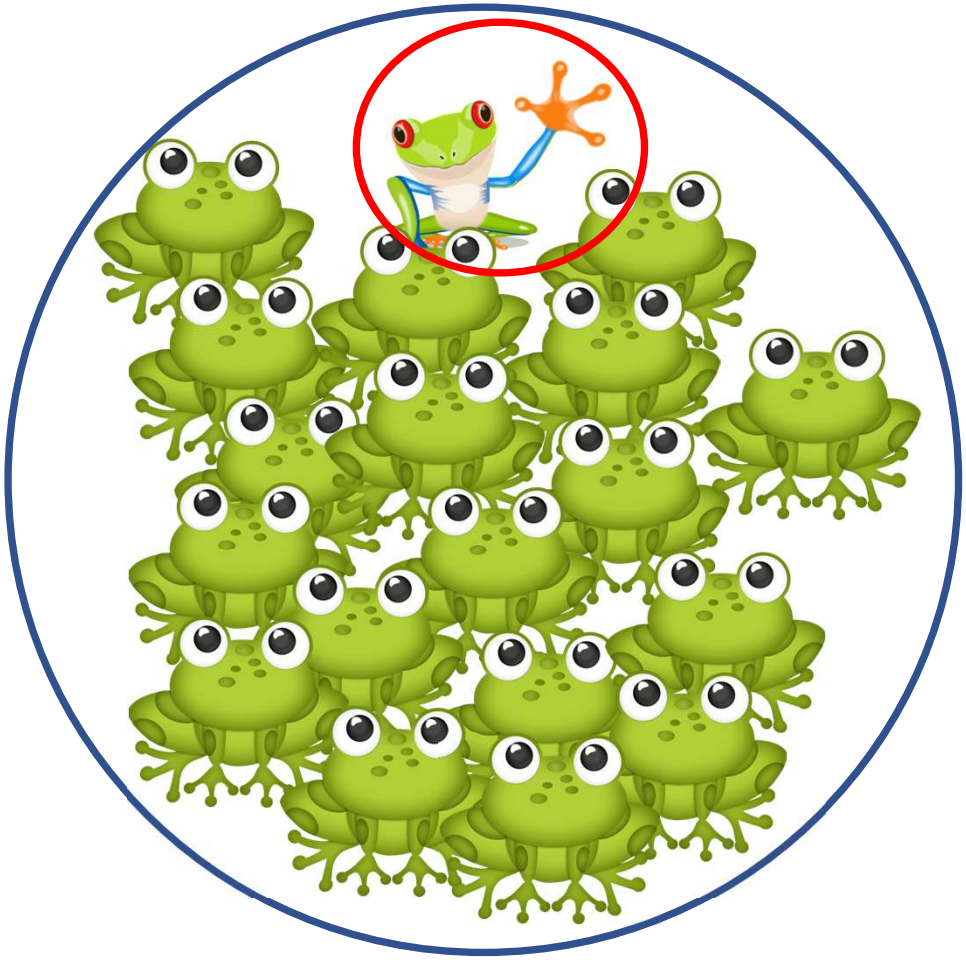


Now the design is balanced: Quality if often more important than quantity



Population

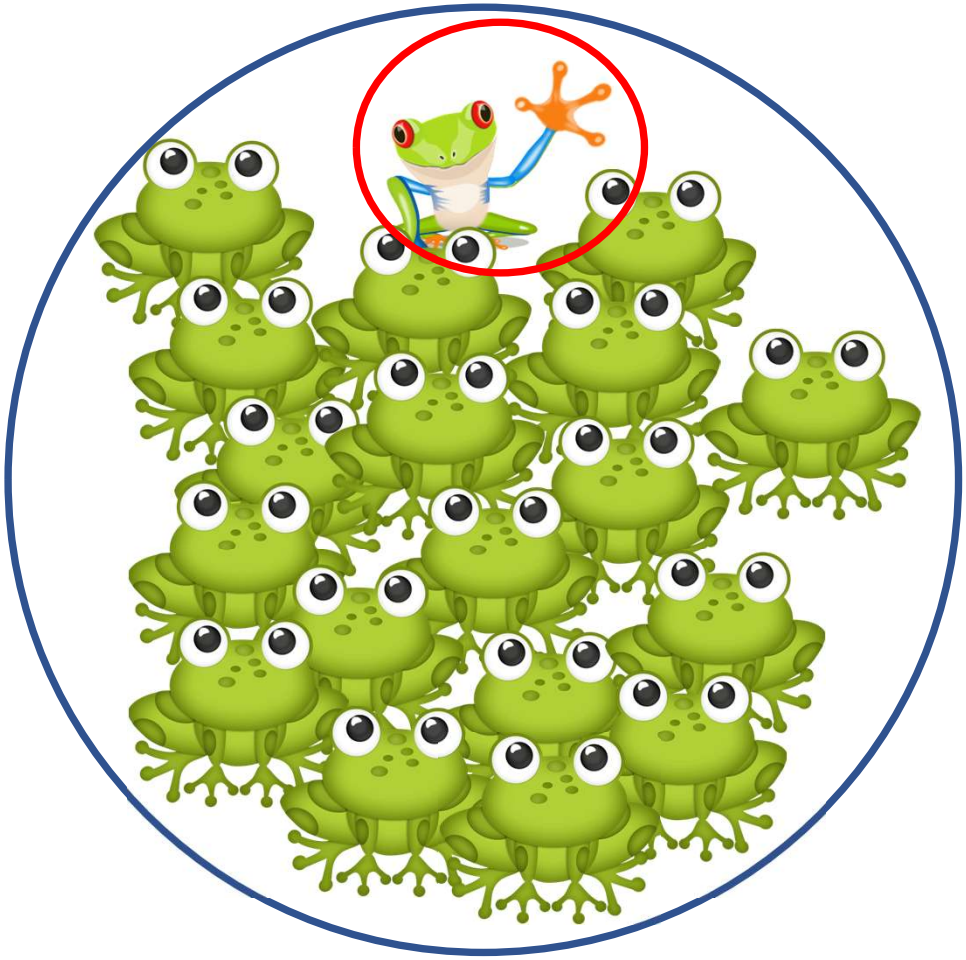
Reliability



Population

Outlier

Reliability



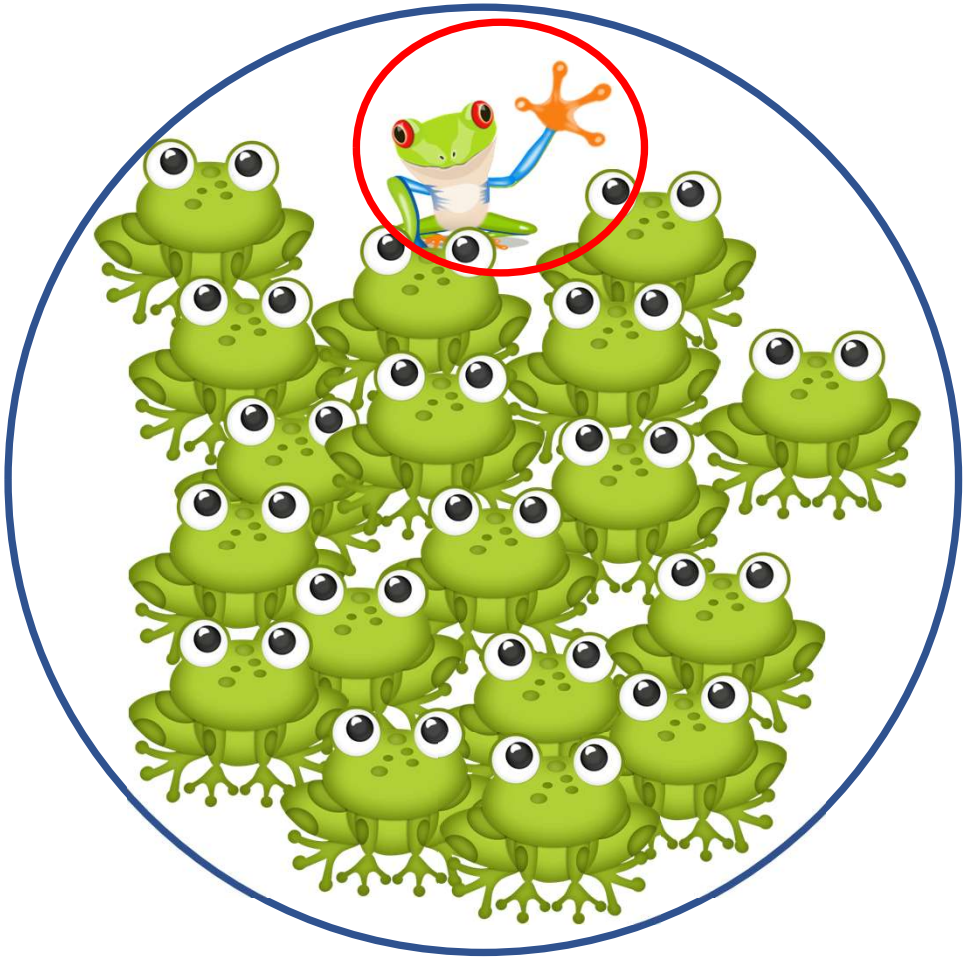
Population

Outlier

N = 1

N = 2

Reliability

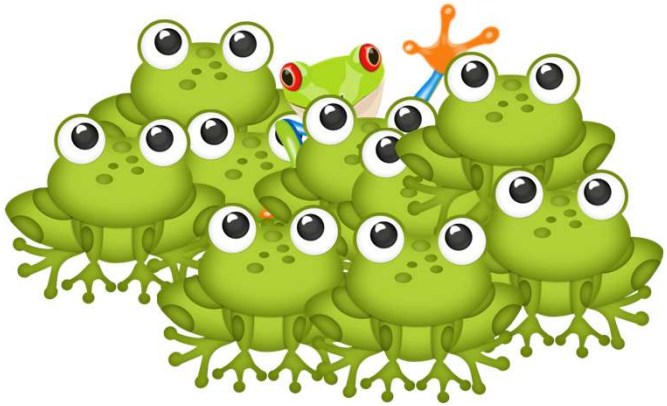


Population

Outlier

N = 1

N = 2



N = 10

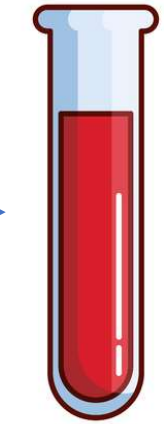
Reliability

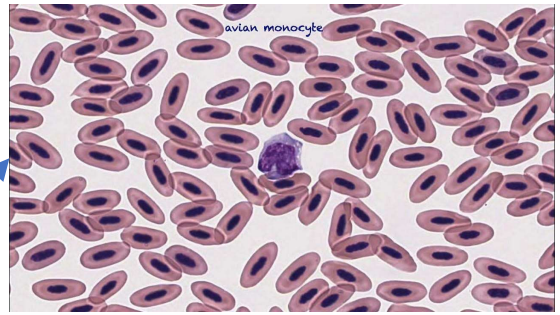
Replication

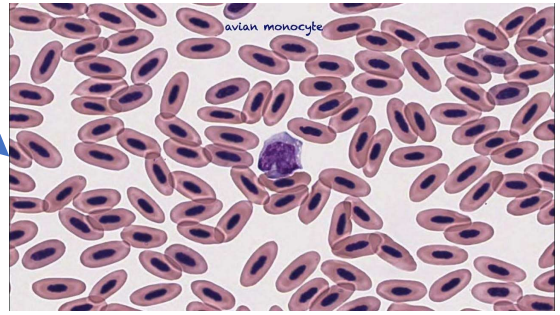
Repeated, Replicates, Reproducible, Replicable

Ex.: How many monocytes per mL of blood in sandgrouse?

Bird A



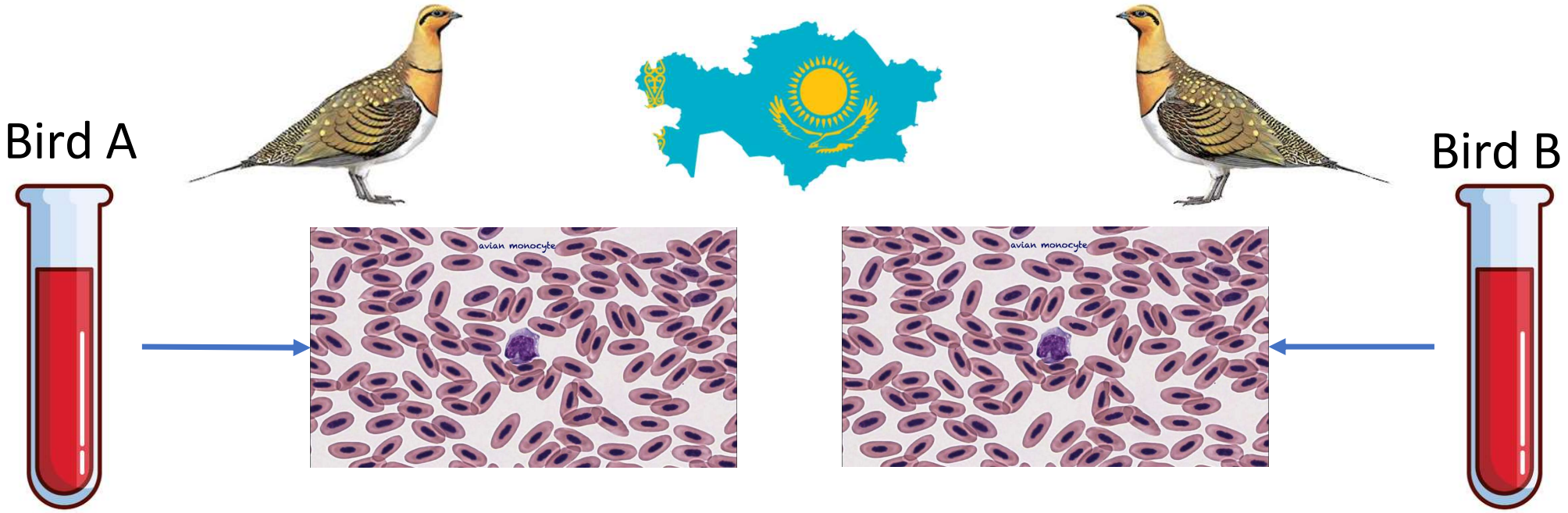




Repeated measurements

Repeated, Replicates, Reproducible, Replicable

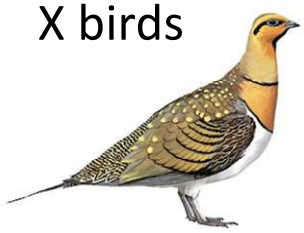
Ex.: How many monocytes per mL of blood in sandgrouse?

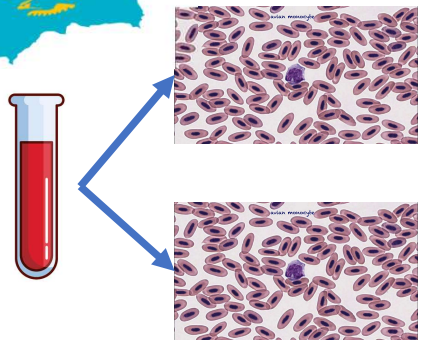


Replicates

Repeated, Replicates, Reproducible, Replicable

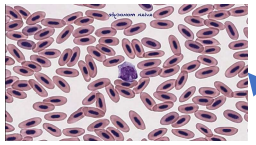
X birds

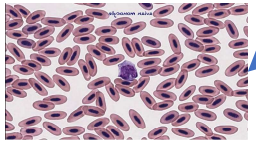




Publish your protocol

Researcher B **replicated** your **study**





X birds

Can compare to your results. Improve reliability, ID new factors, refine knowledge

Repeated, Replicates, Reproducible, Replicable

Ex.: How many monocytes per mL of blood in sandgrouse?

Publish your protocol

	A	B	C	D
1	0.30	0.91	0.73	0.51
2	0.18	0.26	0.28	0.97
3	0.95	0.45	0.13	0.78
4	0.79	0.34	0.68	0.70
5	0.12	0.71	0.75	0.54
6	0.83	0.95	0.72	0.31
7	0.53	0.34	0.48	0.07
8	0.29	0.87	0.39	0.92
9	0.31	0.85	1.00	0.18
10	0.29	0.84	0.97	0.13
11	0.28	0.95	0.35	0.41
12	0.66	0.11	0.93	0.46
13	0.03	0.80	0.63	0.81
14	0.92	0.28	0.78	0.16
15				

Make your data available

Now your study is **reproducible**

Ensure Replicability and reproducibility

1. Record everything you do
2. Be transparent and detailed
3. Publish your raw data
4. (Publish your code)

Repeated, Replicates, Reproducible, Replicable

- Repeated measurement: same experimental unit
- Replicates: different experimental units, same experiment
- Replicability: same methods, new data
- Reproducibility: same methods, same data, different researcher

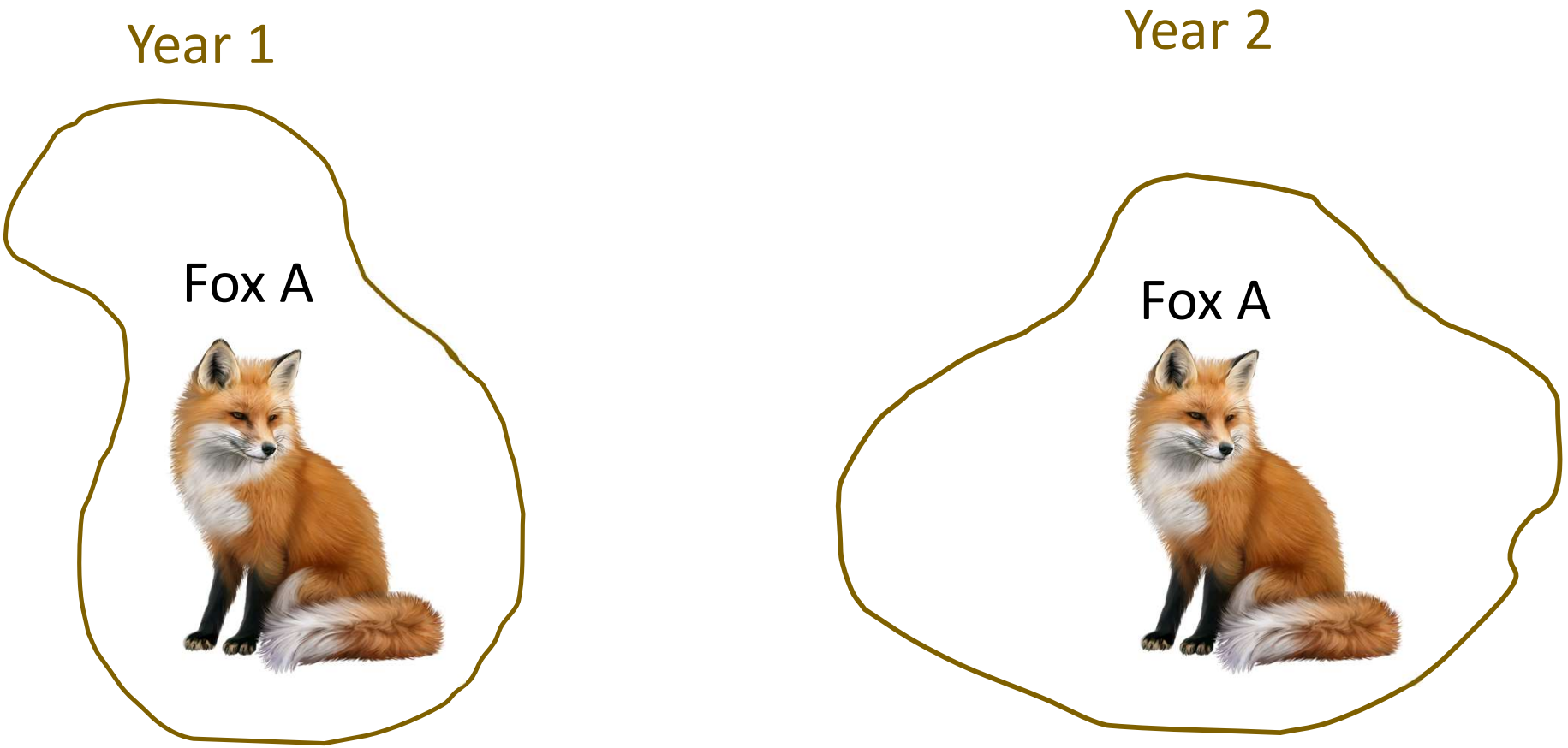
Independence of replicates

- Independence **eliminates bias** in data
- **Validity of your results:** if you violate assumption of your tests, your results may be biased
- **Consistency** of your results across samples
- **Simplify analyses**

- Dependence may be inherent to your study design: you **MUST** account for it in analyses

Pseudoreplication

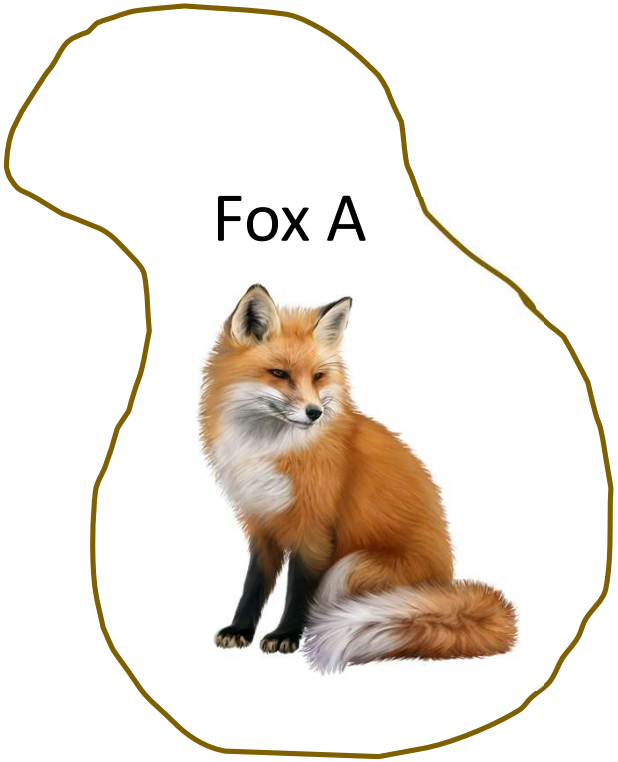
- Violate independence assumption

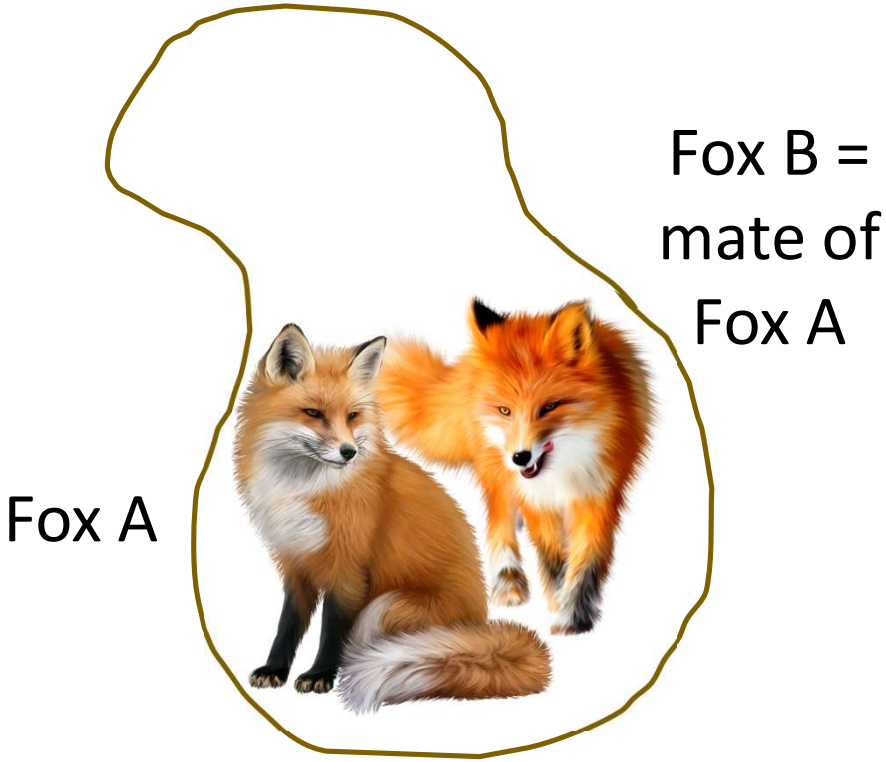


Pseudoreplication

- Violate independence assumption

Year 1

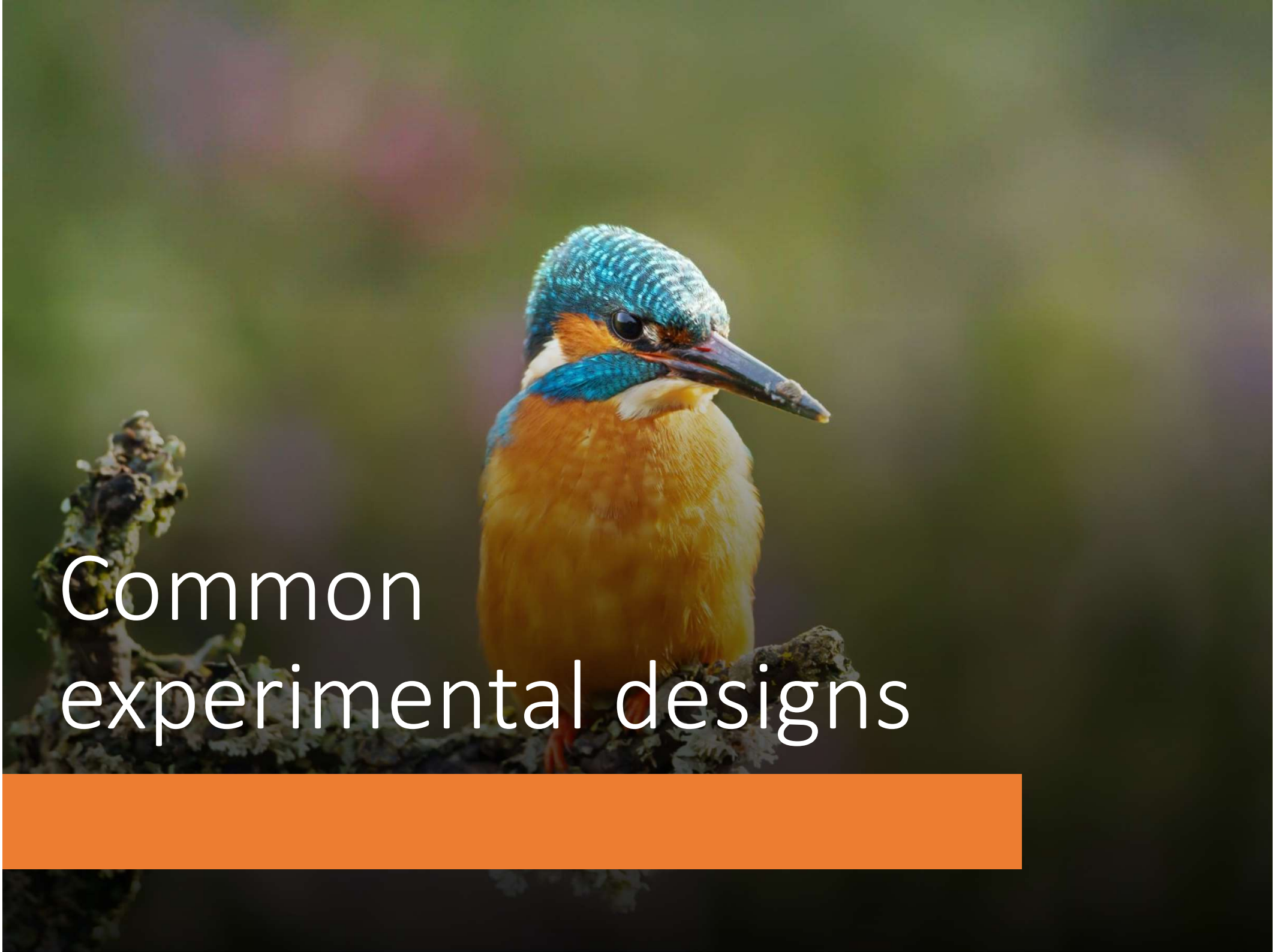




Use of a control group

- Need to assess how response variable behaves without treatment
- remove effect of all factors that are not the factor of interest

- allows to confirm that study results are due to manipulation of independent variables



Common experimental designs

Sampling

Probability vs Non-probability

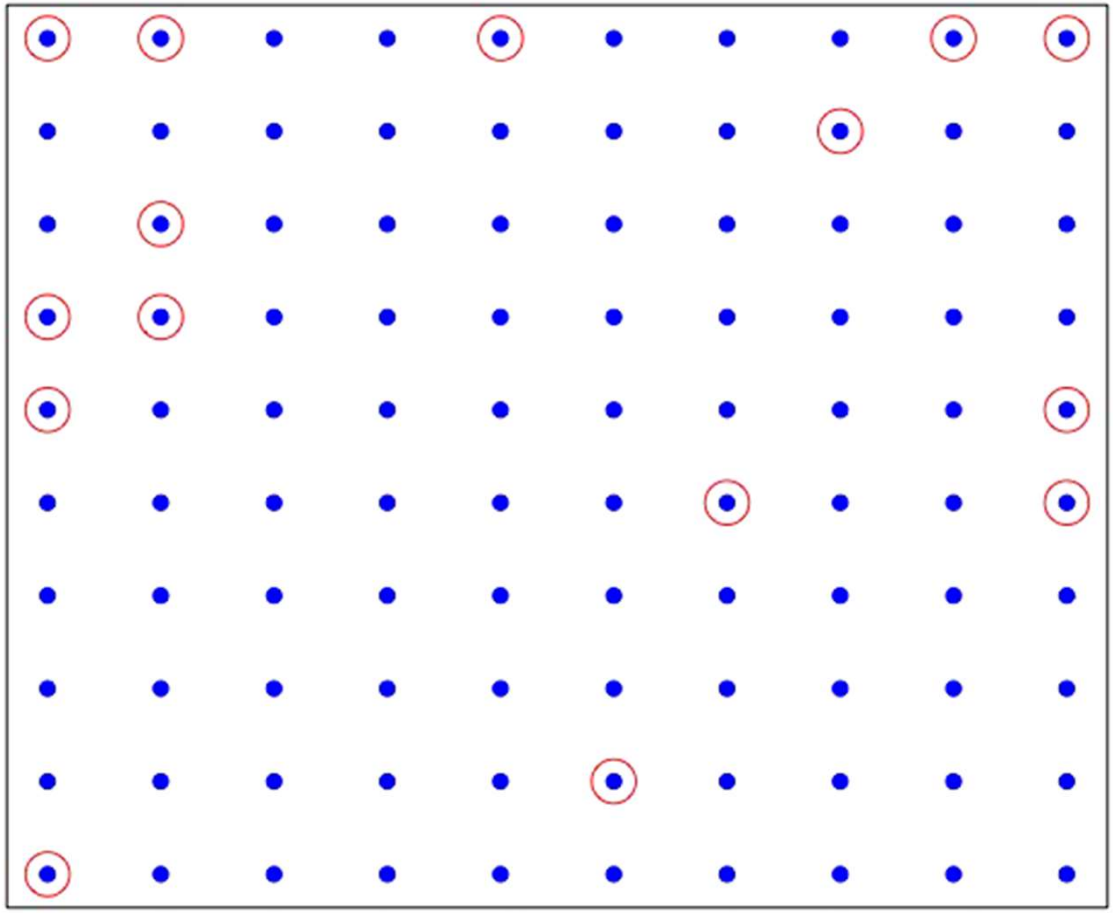
Probability	Non-probability
Random	Non-random
Each individual has equal chance of being sampled	Each individual **does not have** equal chance of being sampled
ex.: simple random sampling	ex.: convenience sampling

Sampling bias!

Common types of probability sampling

- Simple random
- Systematic
- Stratified
- clustered

Simple random sampling



Simple random sampling

Pros ✅

- Robust to bias
- Most robust sampling in inferential methods

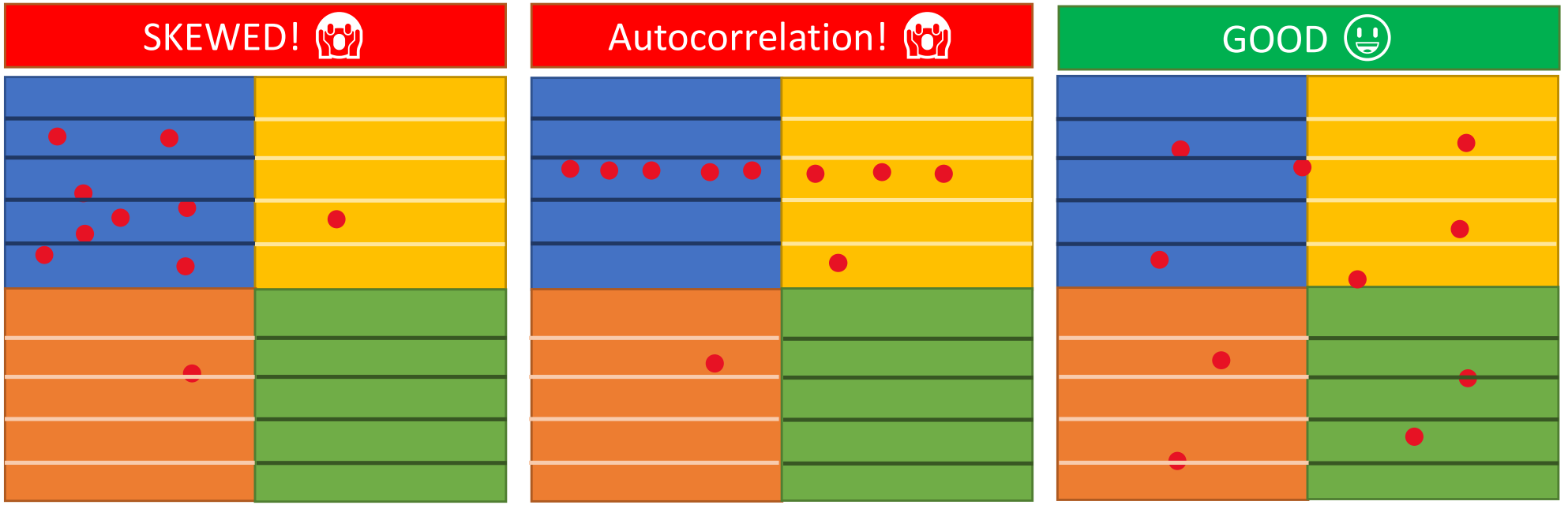
Cons ❌

- Costly
- Bias can still happen by chance (random cluster selection)

Simple random sampling

- But keep in mind!
- You can get bias by chance only

- Ex.: Generate 10 random sampling locations.



Random or haphazard?

- Random ≠ arbitrary
- Haphazard = arbitrary
- Haphazard: non-probability technique
  - Arbitrary
  - No rigorous method of sampling
  - No specific reason for including or excluding items
  - Subject to bias (even though no conscious bias)
  - unpredictable errors and risk of invalid results

Systematic sampling

- Units selected at regular interval with a random starting point
- Highly popular: time and cost efficient

- Make sure there is no underlying pattern in the way the population is ordered:
  - Temporal periodicity (e.g. day/night)
  - Spatial repetitive patterns (e.g. crops)

Systematic sampling

Pros ✅

- Simplicity
- Time/cost efficient
- Population evenly sampled
- Risk of clustered random selection avoided

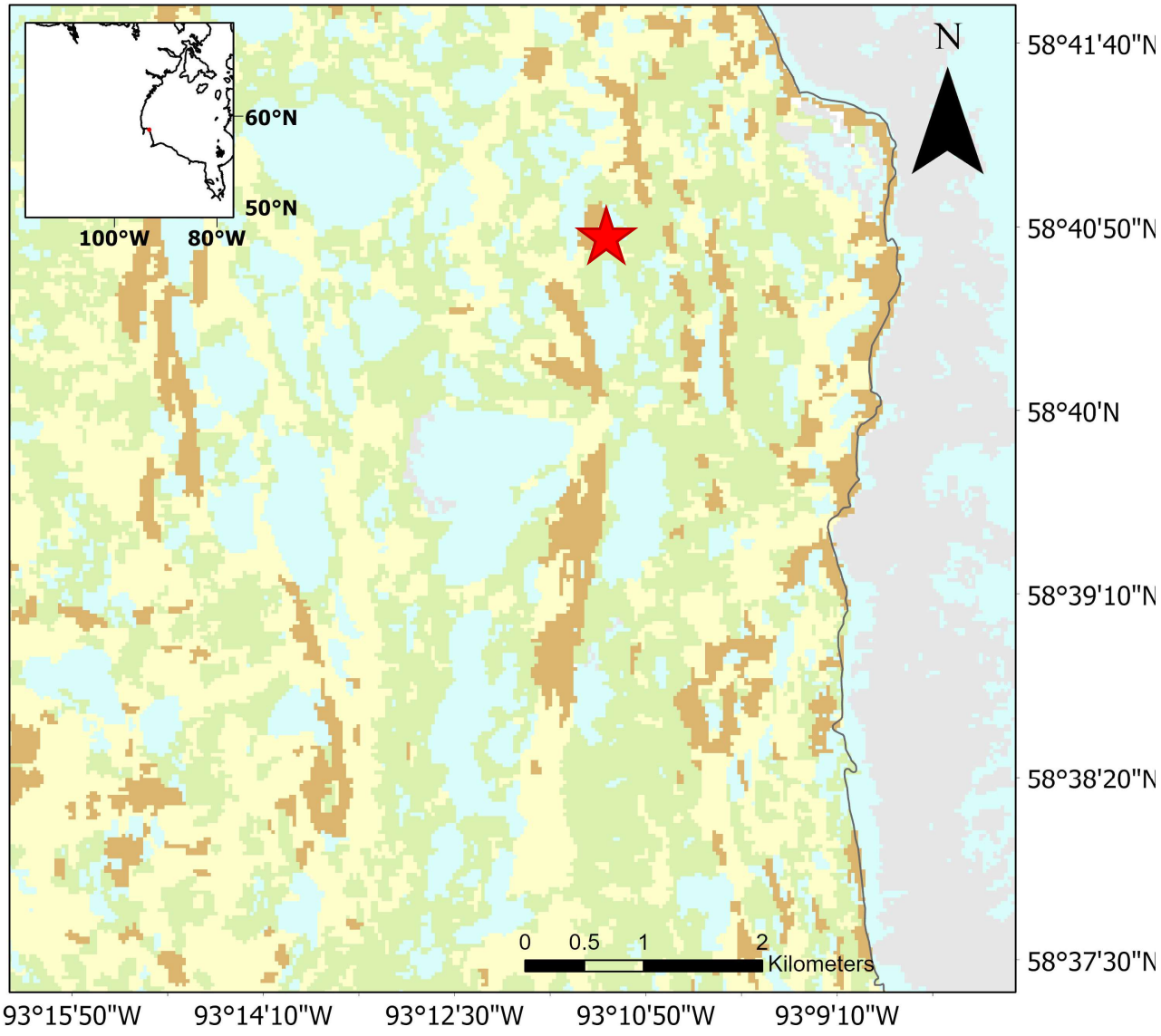
Cons ❌

- Population must show randomness along measured variable

Systematic sampling: real-life example

- **Objective:** You want to determine Goose abundance on the tundra
- **Method:** Line transect distance sampling
  - 15 2-Km Transects, East-West, inland/coastal
- **Implementation:**
  1. Generate a random starting point in your study area (e.g. using ArcGIS)

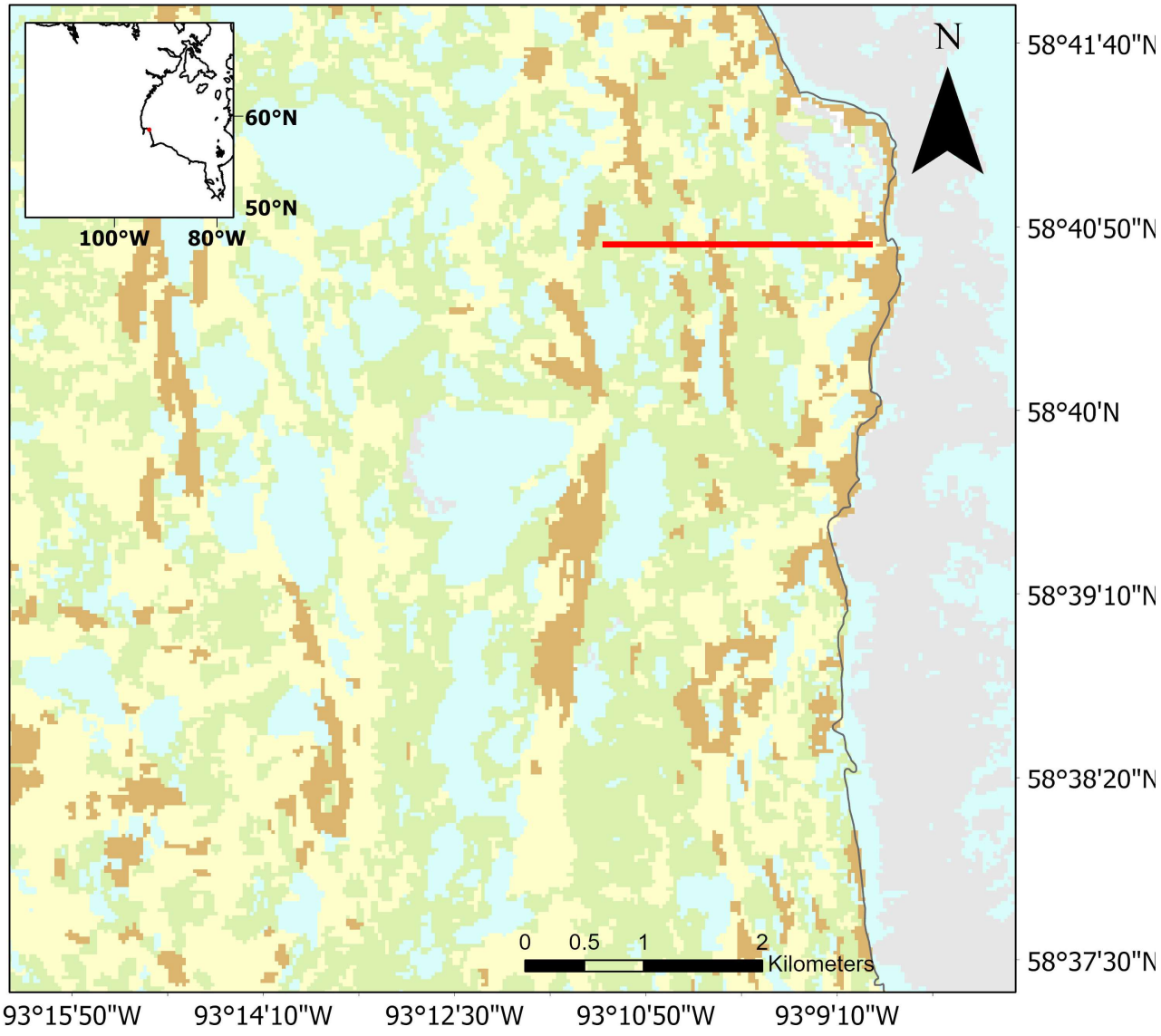
Systematic sampling: real-life example



Systematic sampling: real-life example

- **Objective:** You want to determine Goose abundance in the tundra
- **Method:** Line transect distance sampling
  - 15 2-Km Transects, East-West, inland/coastal
- **Implementation:**
  1. Generate a random starting point in your study area (e.g. using ArcGIS)
  2. Flip a coin to decide if your starting transect will be inland or coastal

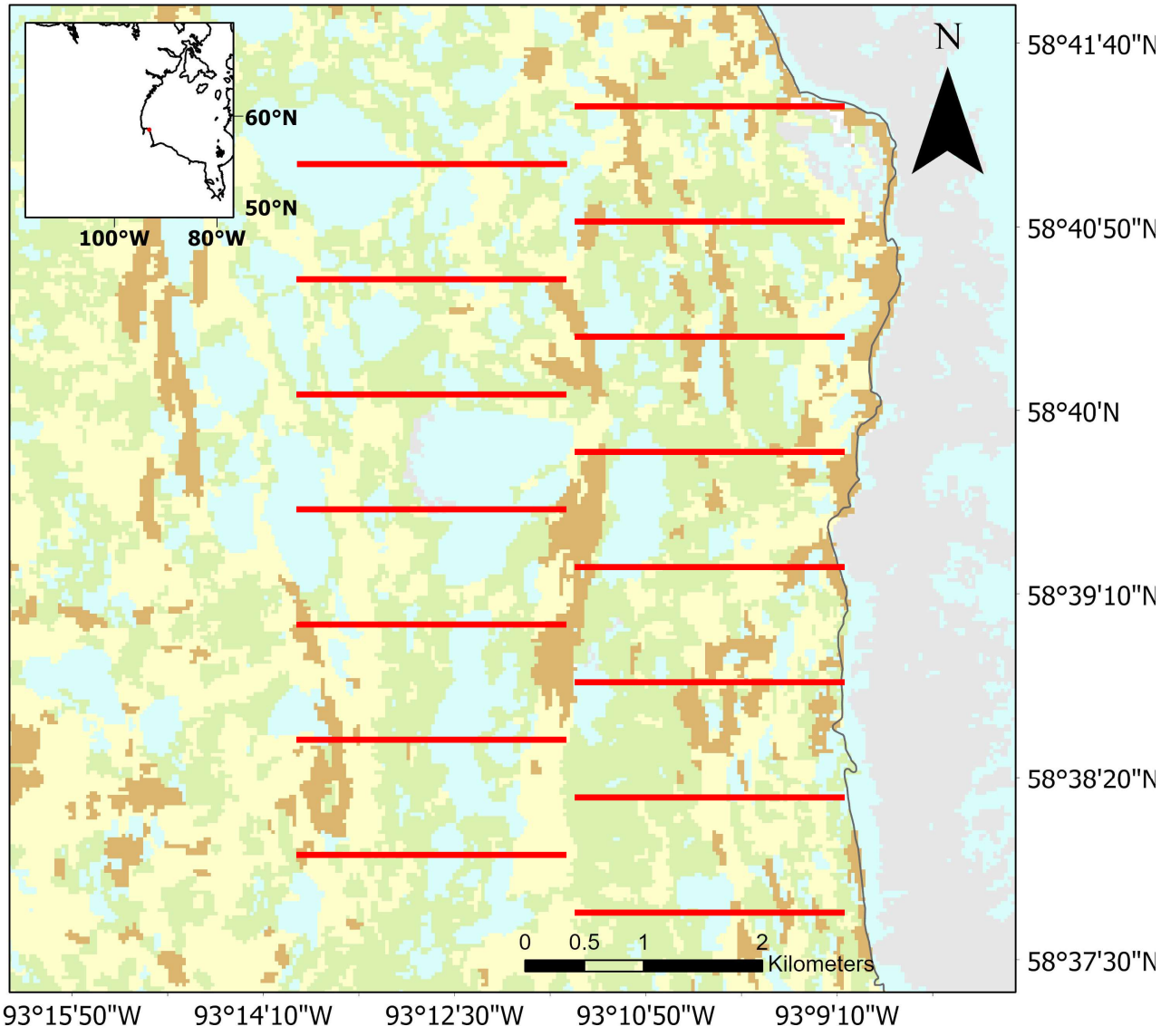
Systematic sampling: real-life example

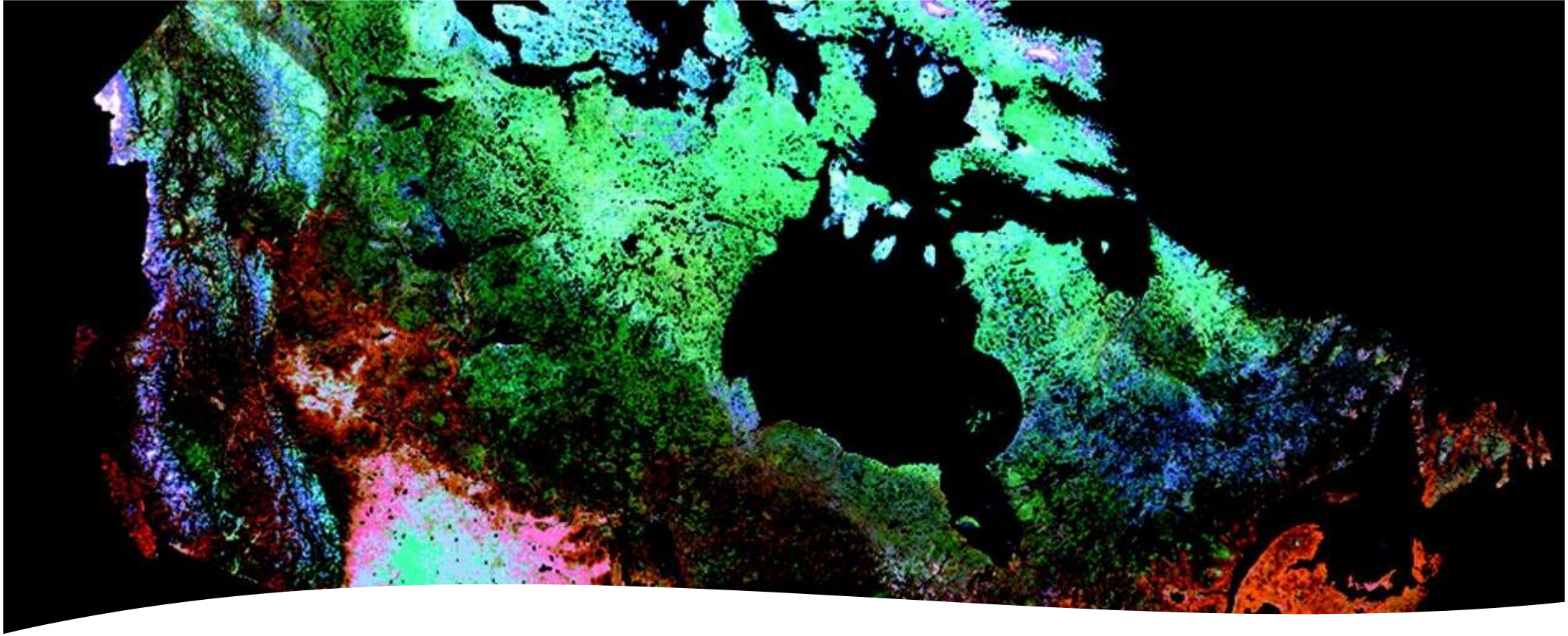


Systematic sampling: real-life example

- **Objective:** You want to determine Goose abundance in the tundra
- **Method:** Line transect distance sampling
  - 15 2-Km Transects, East-West, inland/coastal
- **Implementation:**
  1. Generate a random starting point in your study area (e.g. using ArcGIS)
  2. Flip a coin to decide if your starting transect will be inland or coastal
  3. Place your other transects every 1 Km alternatively inland or coastal

Systematic sampling: real-life example





Stratified sampling

- Define non-overlapping homogeneous strata
- n individuals randomly selected from each strata
- Strata can be used to control for sources of variation

Stratified sampling

Pros ✅

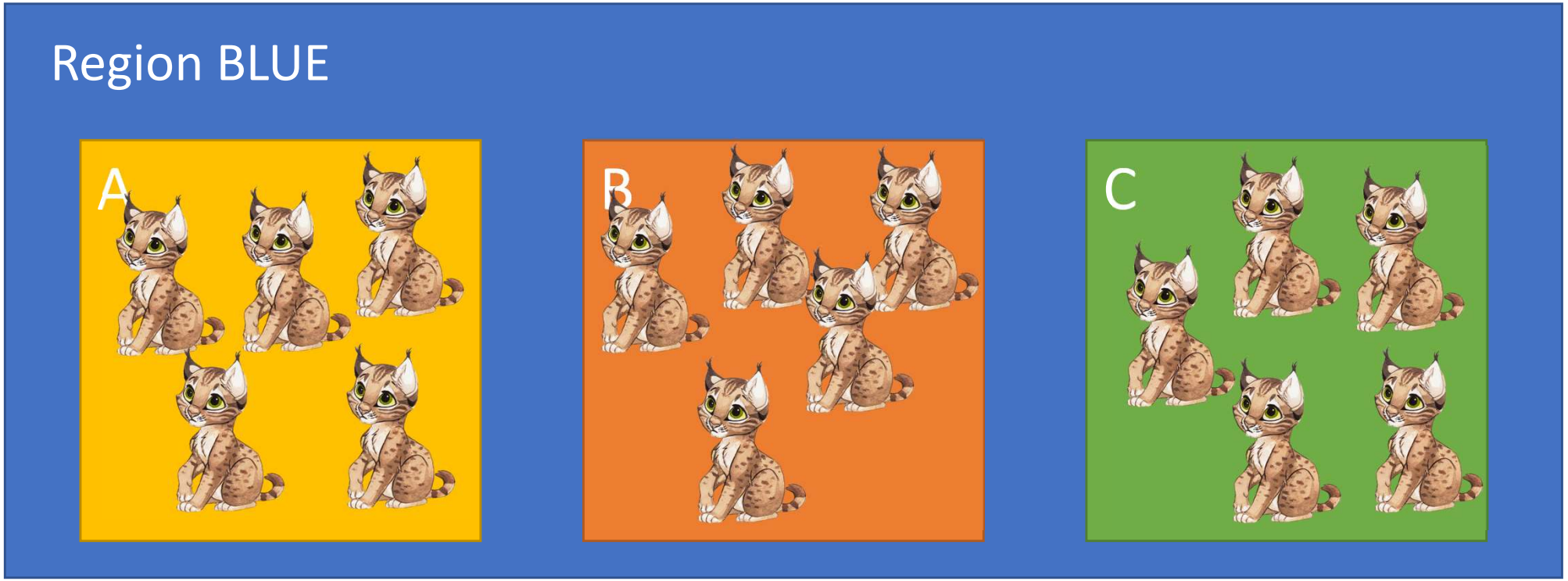
- Control for sources of variation
- Representative of the whole population allows for stronger inferences
- Includes minority groups

Cons ❌

- Stratum definition

Stratified sampling: example

- **Question:** What is lynx body mass in region BLUE
- **Method:** sample 15 lynx from BLUE
  - BLUE has 3 habitats (A,B,C) with different prey and climate conditions
  - Habitat may be a source of variation
  - Habitat = stratum, N= 5 per stratum



Clustered sampling

- Population naturally divided into homogeneous groups (clusters)
- Groups selected at random
  - One-stage cluster sampling: measure all individuals within clusters
  - Two-stage cluster sampling: randomly select n individuals per cluster
- Groups are sampling units

Clustered sampling

Pros ✅

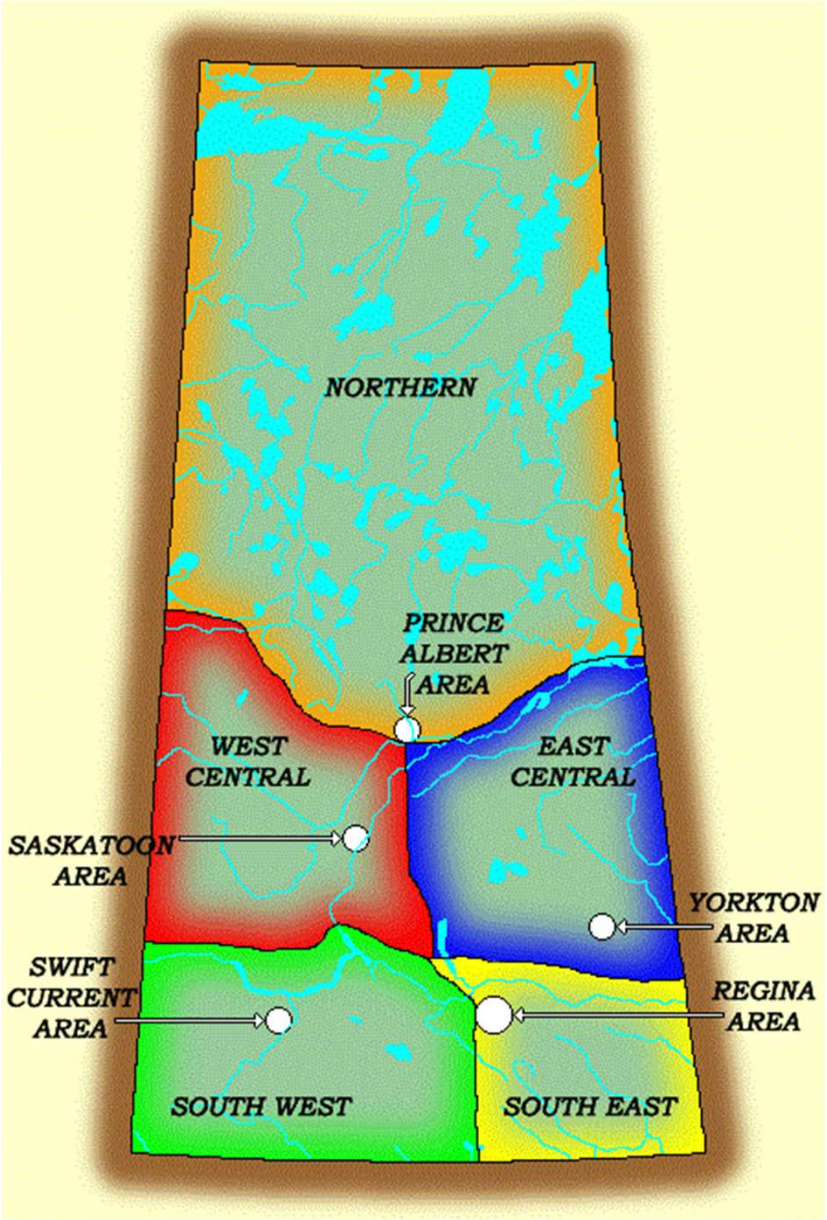
- Time/cost efficient (e.g., large geographic areas)
- Allows for inferences, if clustering not biased

Cons ❌

- Less robust than other methods for inferences
- Larger sampling error
- Cluster definition

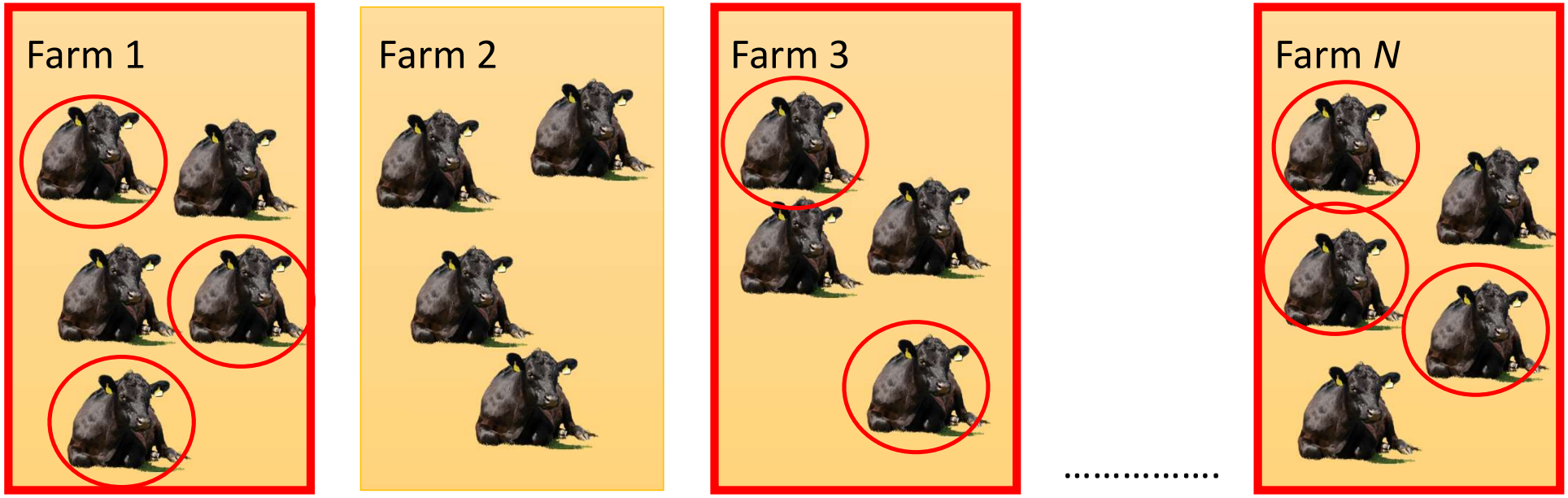
Clustered sampling: example

- **Question:** What is the prevalence of ectoparasite in cows in Saskatchewan?
- **Issues:**
  - large geographic area
  - Farms are spread out
  - Thousands of cows per farm



Clustered sampling: example

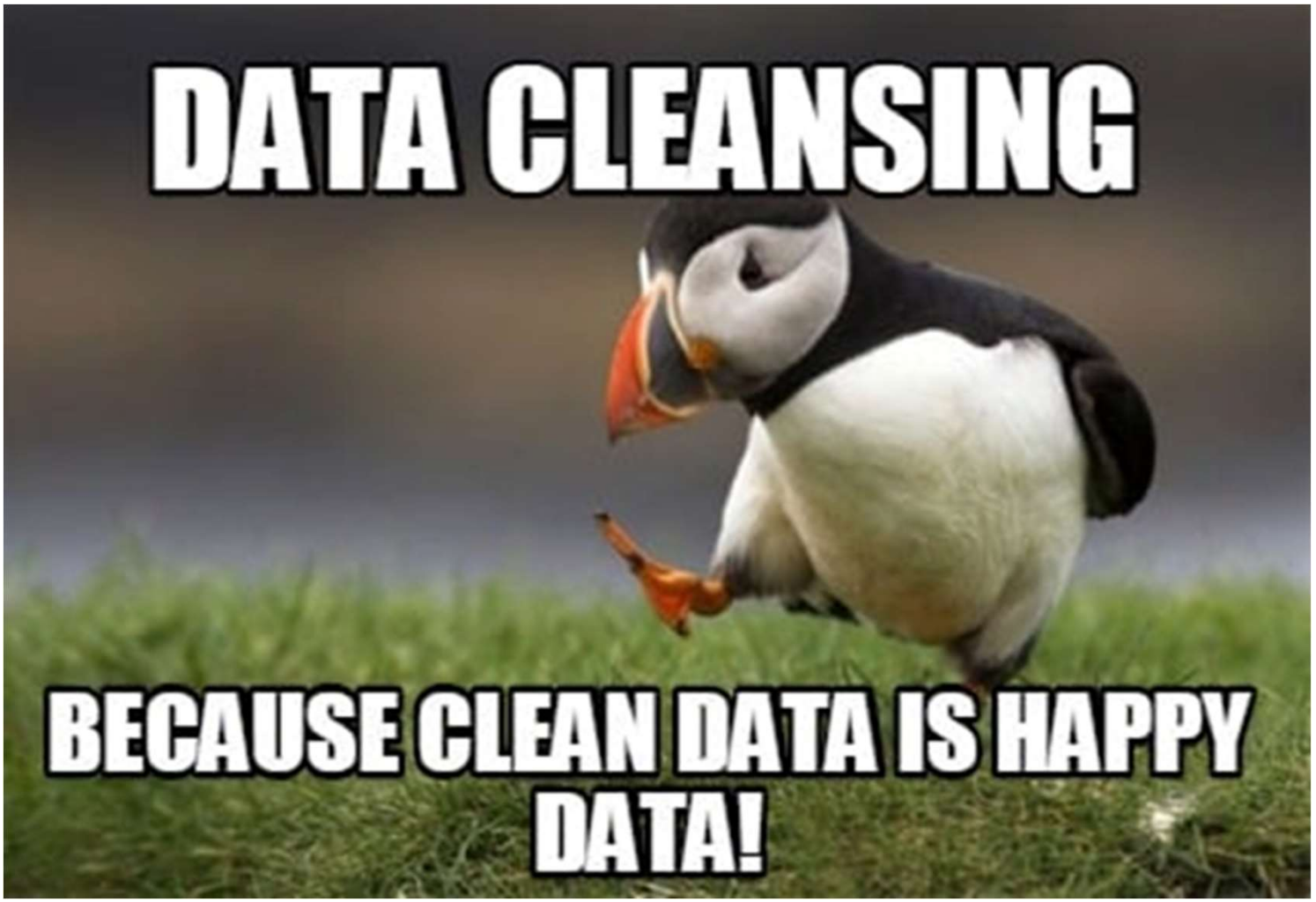
- **Question:** What is the prevalence of ectoparasite in cows in Saskatchewan?
- **Methods:**
  - Select farms at random
  - Two-staged cluster sampling: randomly select 50 cows per farm



While sampling, keep in mind:

- Your question!
- Level of **precision** you need
- Sample size (adjust as a function of expected noise and variability in the target population)
- Time of sampling (ex.: biological cycles, climate patterns…)
- Location of sampling

A few tips on building a data base



1. Have the question in mind

- If you build a database to answer specific questions:
  - Choose and collect your variable wisely (choice of the variable, precision of measurement…)
  - think about the structure of your model

- If you build a long-term project database (no specific question):
  - Make it in a format where it’s easy to extract information

1. Have the question in mind

- What does the data represent?
- How is it acquired?
- At what rate is it acquired?
- How much data is expected?
- how is it going to be used?

snake_case

Pros: Concise when it consists of a few words.
Cons: Redundant as hell when it gets longer.

```
push_something_to_first_queue, pop_what, get_whatever…
```

PascalCase

Pros: Seems neat.

```
GetItem, SetItem, Convert, ...
```

Cons: Barely used. (why?)

camelCase

Pros: Widely used in the programmer community.
Cons: Looks ugly when a few methods are n-worded.

```
push, reserve, beginBuilding, ...
```

skewer-case

Pros: Easy to type.

```
easier-than-capitals, easier-than-underscore, …
```

Cons: Any sane language freaks out when you try it.

SCREAMING_SNAKE_CASE

Pros: Can demonstrate your anger with text.
Cons: Makes your eyes deaf.

```
LOOK_AT_THIS, LOOK_AT_THAT, LOOK_HERE_YOU_MORON, …
```

nocase

Pros: Looks professional.
Cons: Misleading af.

```
supersexyhippothalamus, bool penisbig, …
```

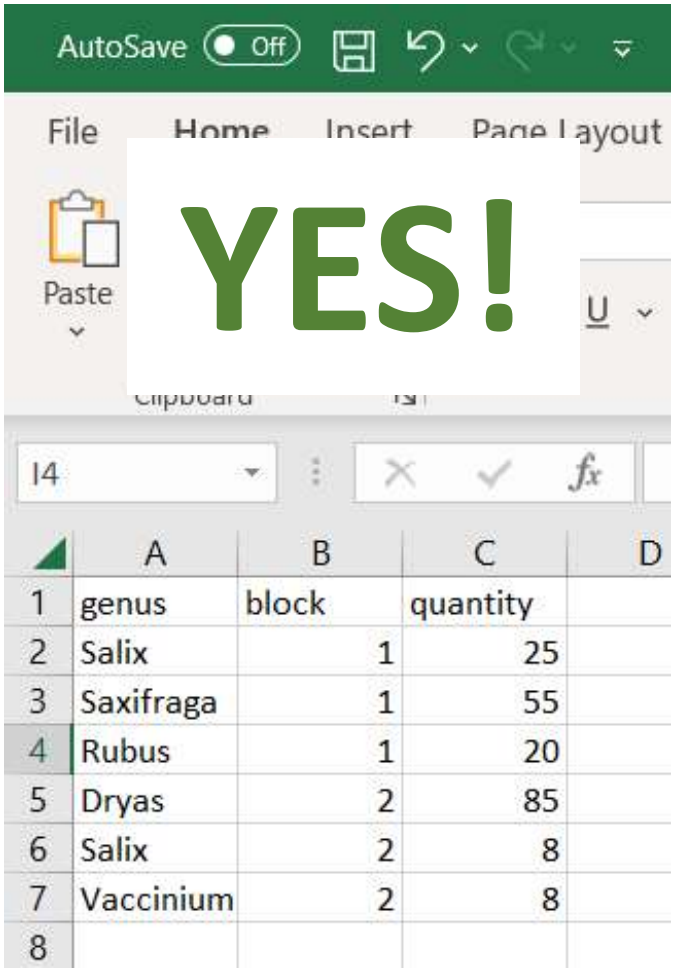
fUcKtHeCaSe

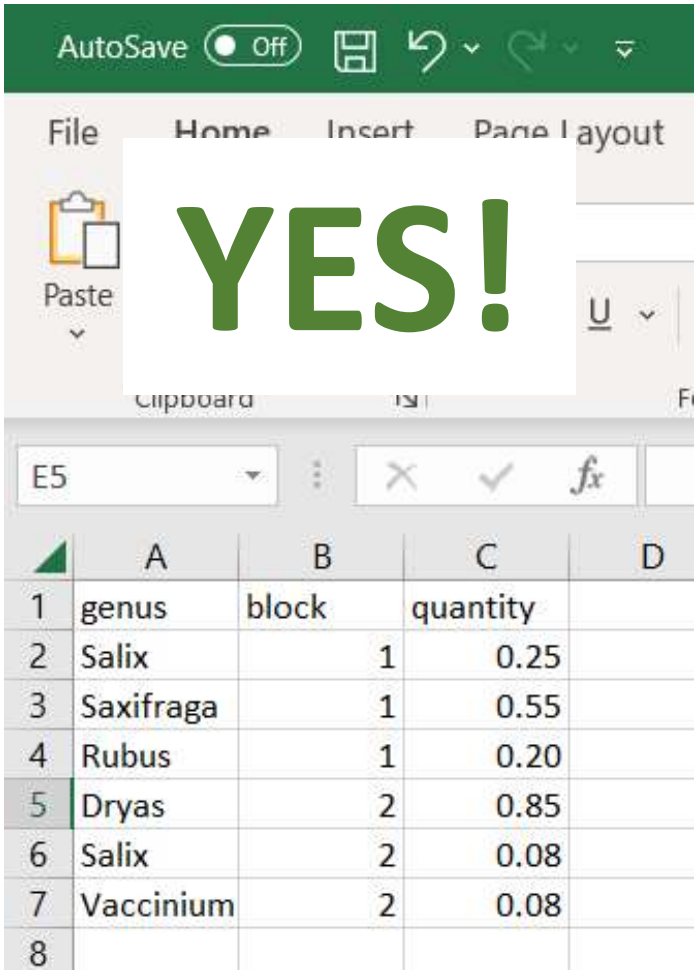
Pros: Can live outside of the law.
Cons: Can be out of a job.

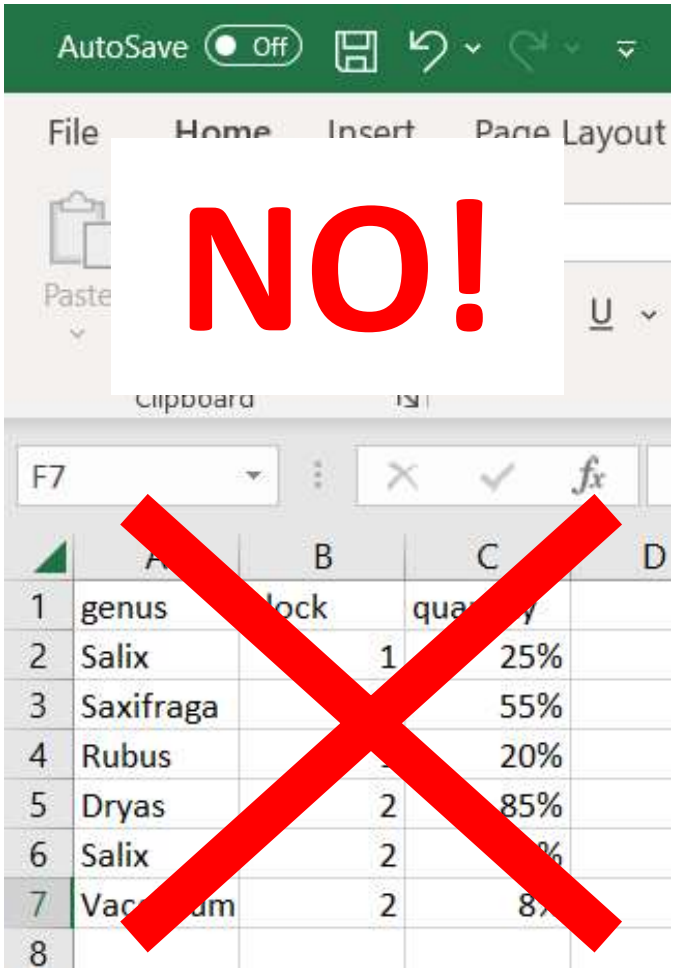
2. Naming variable

- Don’t use special characters (accents, spaces, parentheses, &, ?, ^, /, !, ñ…)
- No unit in column title
- If names are multi-worded use “.” or “_” or capitalize each word: my.df, my_df, MyDf
- Keep it short and informative (the best you can)
- Large databases: use a unique ID for each row (avoid redundant rows)

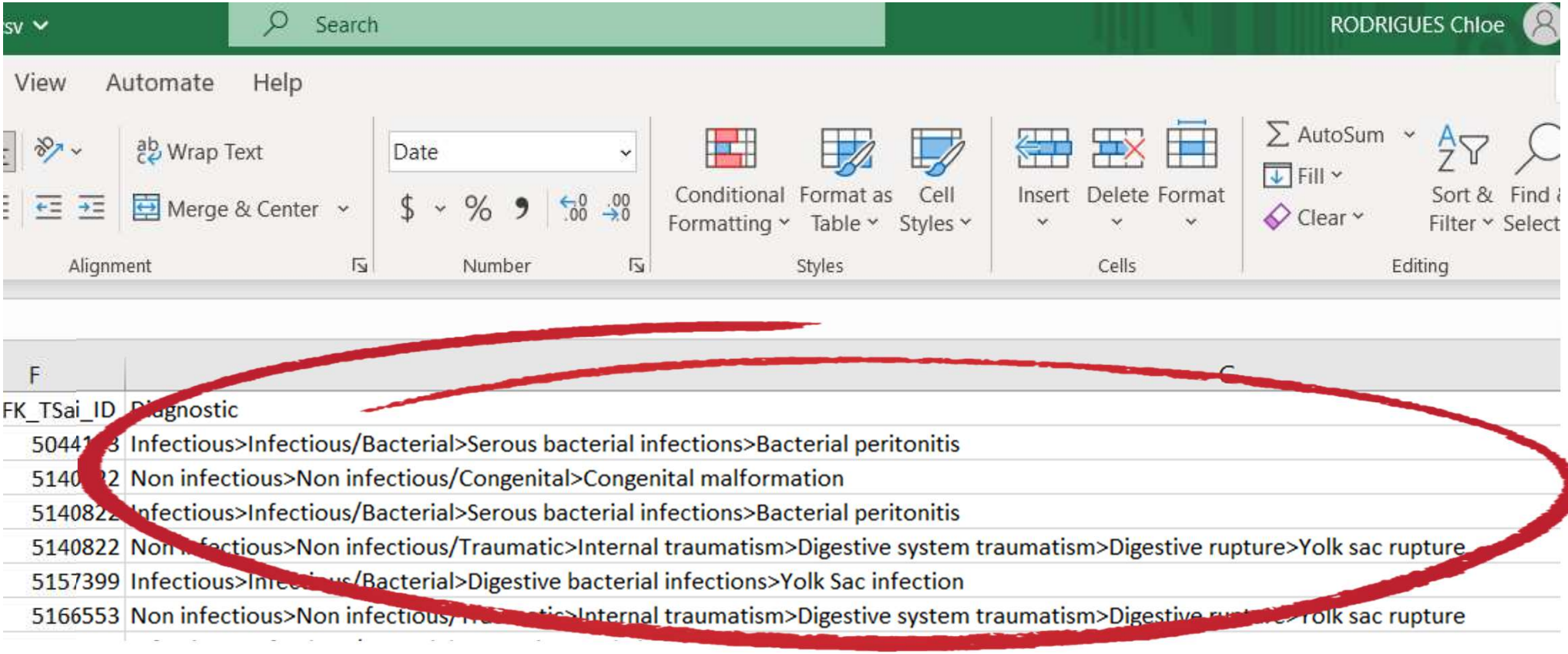
3. No unit after values for numerical variables!







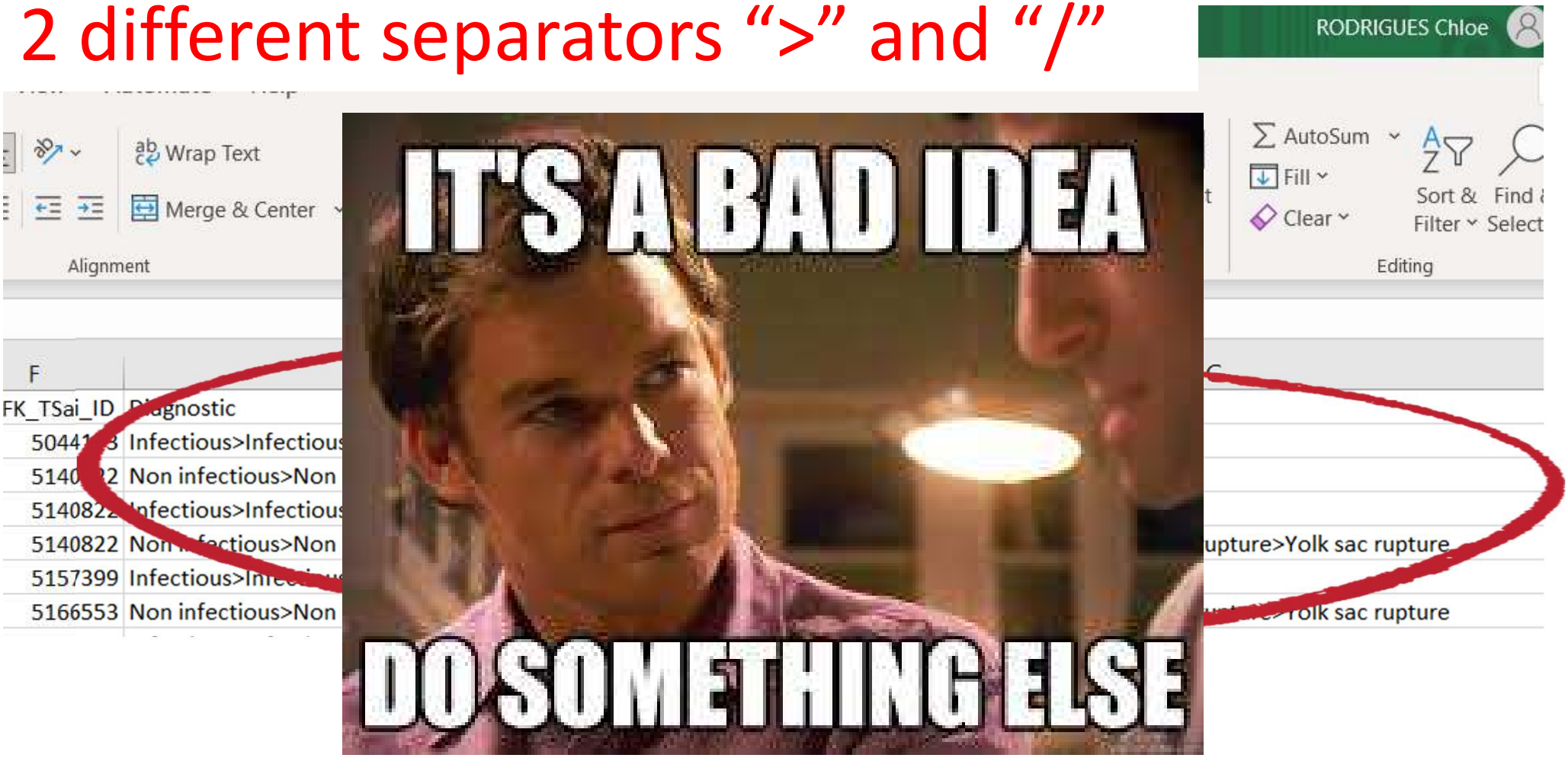
4. Never aggregate variables in the same column



4. Never aggregate variables in the same column

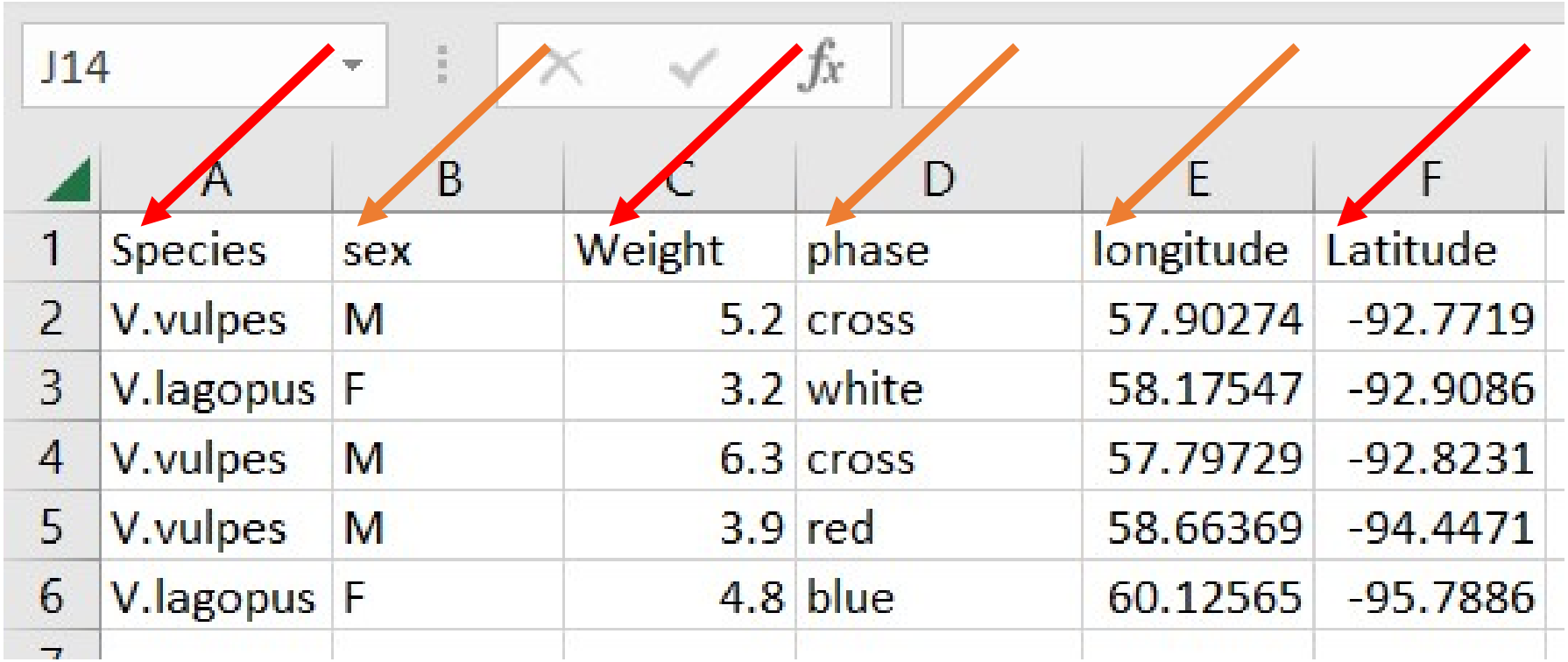
7 variables under the same column

2 different separators “>” and “/”



5. Be consistent when formatting

- When naming variables (columns)



J14

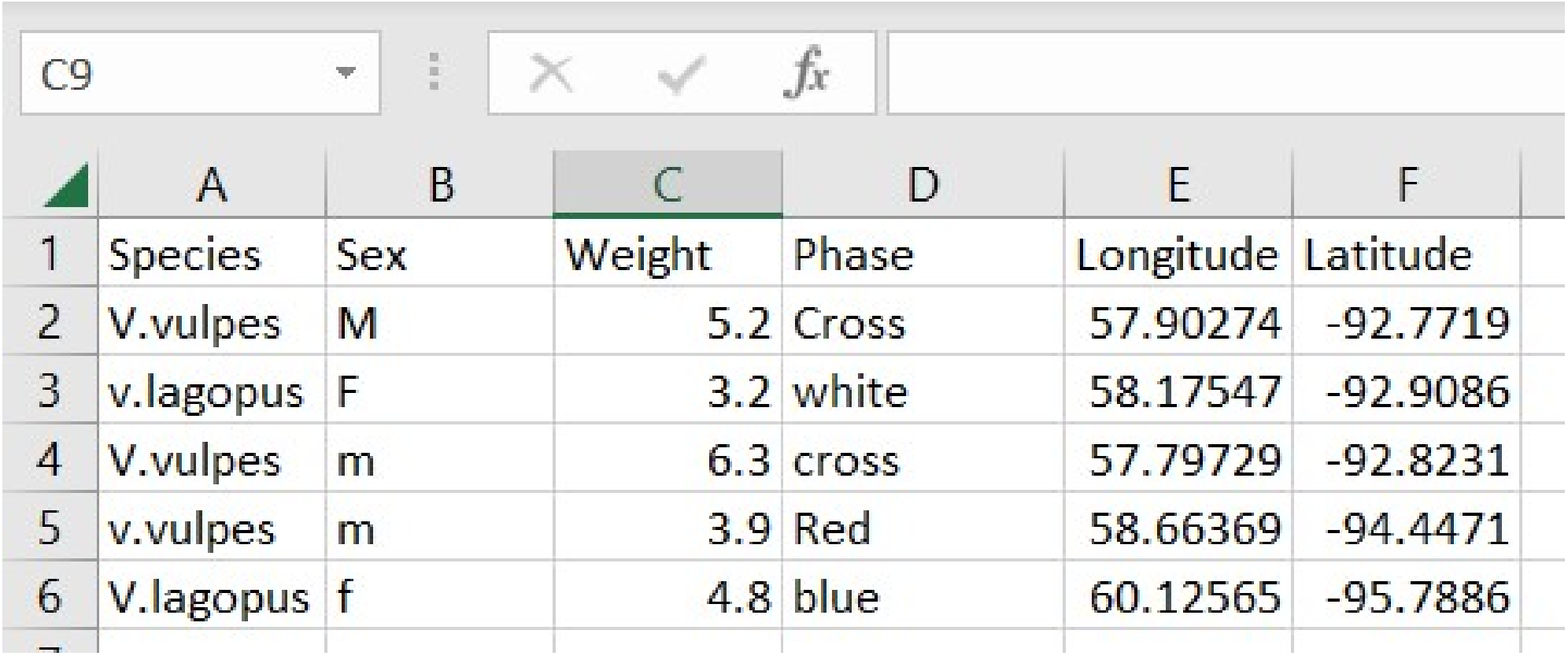
	A	B	C	D	E	F
1	Species	sex	Weight	phase	longitude	Latitude
2	V.vulpes	M	5.2	cross	57.90274	-92.7719
3	V.lagopus	F	3.2	white	58.17547	-92.9086
4	V.vulpes	M	6.3	cross	57.79729	-92.8231
5	V.vulpes	M	3.9	red	58.66369	-94.4471
6	V.lagopus	F	4.8	blue	60.12565	-95.7886

5. Be consistent when formatting

- When naming variables (columns)

5. Be consistent when formatting

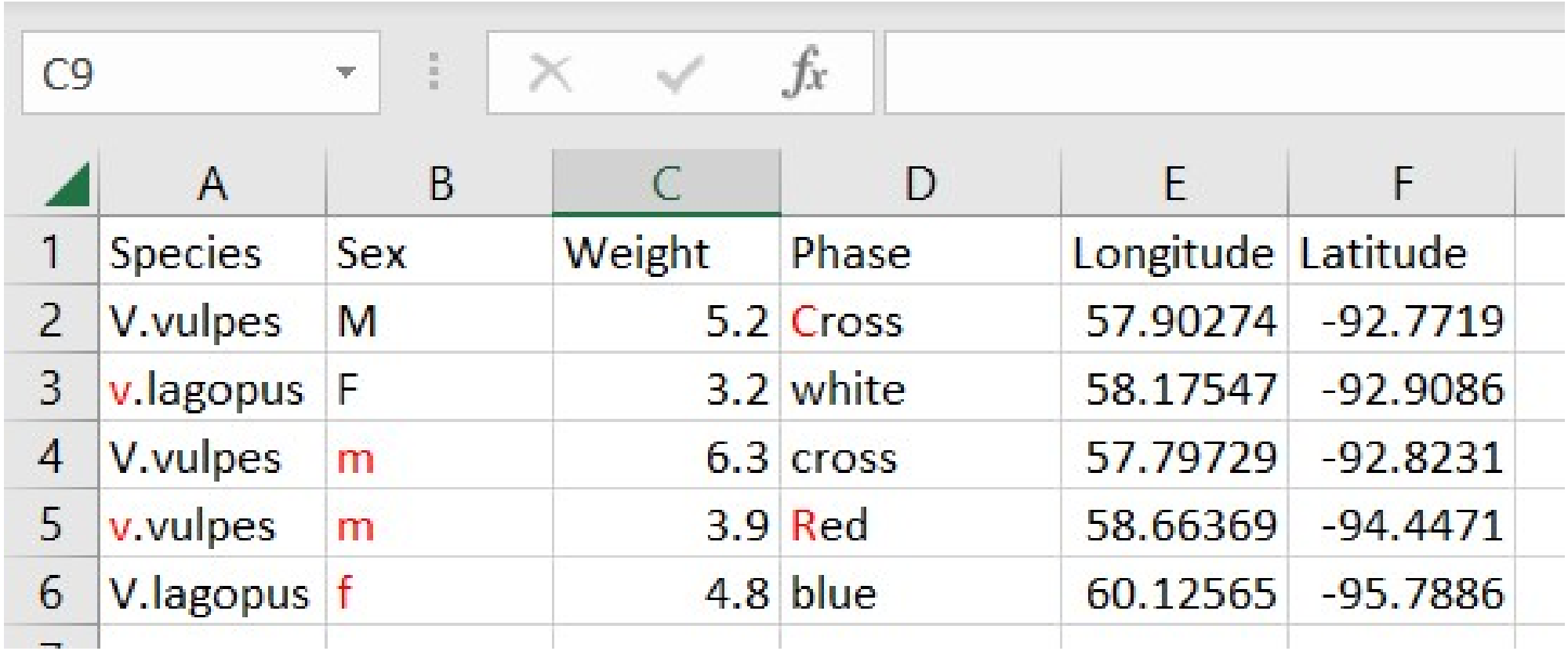
- When naming variables (columns)
- When naming levels of a factor



	A	B	C	D	E	F
1	Species	Sex	Weight	Phase	Longitude	Latitude
2	V.vulpes	M	5.2	Cross	57.90274	-92.7719
3	v.lagopus	F	3.2	white	58.17547	-92.9086
4	V.vulpes	m	6.3	cross	57.79729	-92.8231
5	v.vulpes	m	3.9	Red	58.66369	-94.4471
6	V.lagopus	f	4.8	blue	60.12565	-95.7886

5. Be consistent when formatting

- When naming variables (columns)
- When naming levels of a factor



	A	B	C	D	E	F
1	Species	Sex	Weight	Phase	Longitude	Latitude
2	V.vulpes	M	5.2	Cross	57.90274	-92.7719
3	v.lagopus	F	3.2	white	58.17547	-92.9086
4	V.vulpes	m	6.3	cross	57.79729	-92.8231
5	v.vulpes	m	3.9	Red	58.66369	-94.4471
6	V.lagopus	f	4.8	blue	60.12565	-95.7886

5. Be consistent when formatting

- When naming variables (columns)
- When naming levels of a factor

R is case sensitive!!!

						F
1	Species	Sex	Weight	Phase	Longitude	Latitude
2	V.vulpes	M	5.2	Cross	57.90274	-92.7719
3	v.lagopus	F	3.2	white	58.17547	-92.9086
4	V.vulpes	m	6.3	cross	57.79729	-92.8231
5	v.vulpes	m	3.9	Red	58.66369	-94.4471
6	V.lagopus	f	4.8	blue	60.12565	-95.7886

5. Be consistent when formatting

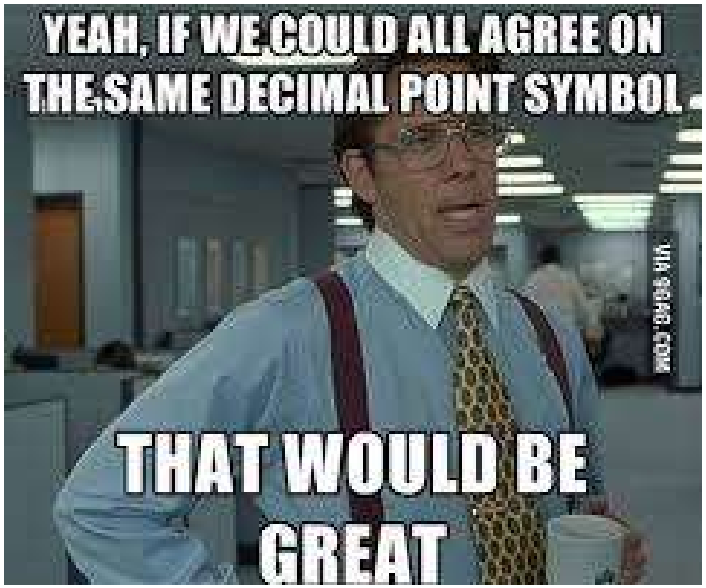
- When naming variables
- When naming levels of a

R is case se

1	Species	Sex	Weight
2	V.vulpes	M	5
3	v.lagopus	F	3
4	V.vulpes	m	6
5	v.vulpes	m	3
6	V.lagopus	f	4

6. Csv vs csv2 format

- Regional differences:
  - “.” vs “,” as a decimal separator (3.5 vs 3,5)
  - “,” vs “;” as a column delimiter
  - read.csv (English) vs read.csv2 (French or Spanish)



7. Last but not least: provide metadata

What are the data?

How were the data collected?

Where were the data collected?

When were the data collected?

(**Who** collected the data?)

Indicate **units** in the metadata

Really **anything relevant** for anyone to understand the data

7. Example: provide metadata

The variables associated with each HormoneBase entry are listed below:

Variable name	Variable definition	Data type/unit
Vert_Group	Taxonomic group (amphibian, bird, fish, mammal, reptile)	String
Genus	Genus	String
Species	Species	String
Common_name	Common name	String
Population_1	Name of first location at which samples were collected (city/region, state/province, country)	String
Population_2	If applicable, name of second location at which samples were collected	String