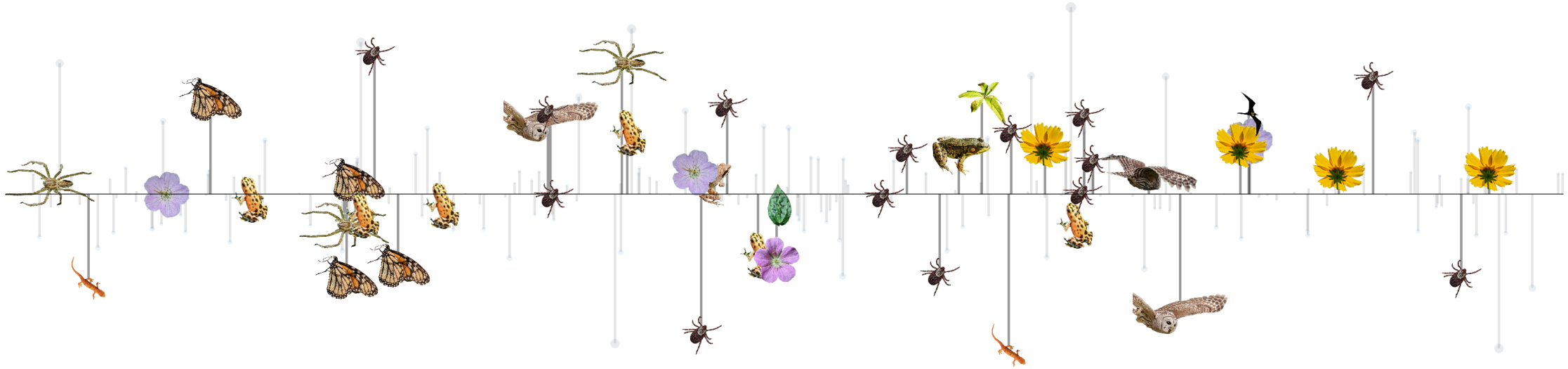
Intro to Quantitative Ecology
UMass Amherst – Michael France Nelson

Deck 3: Data Management



Announcements: Excel Analysis Toolpak

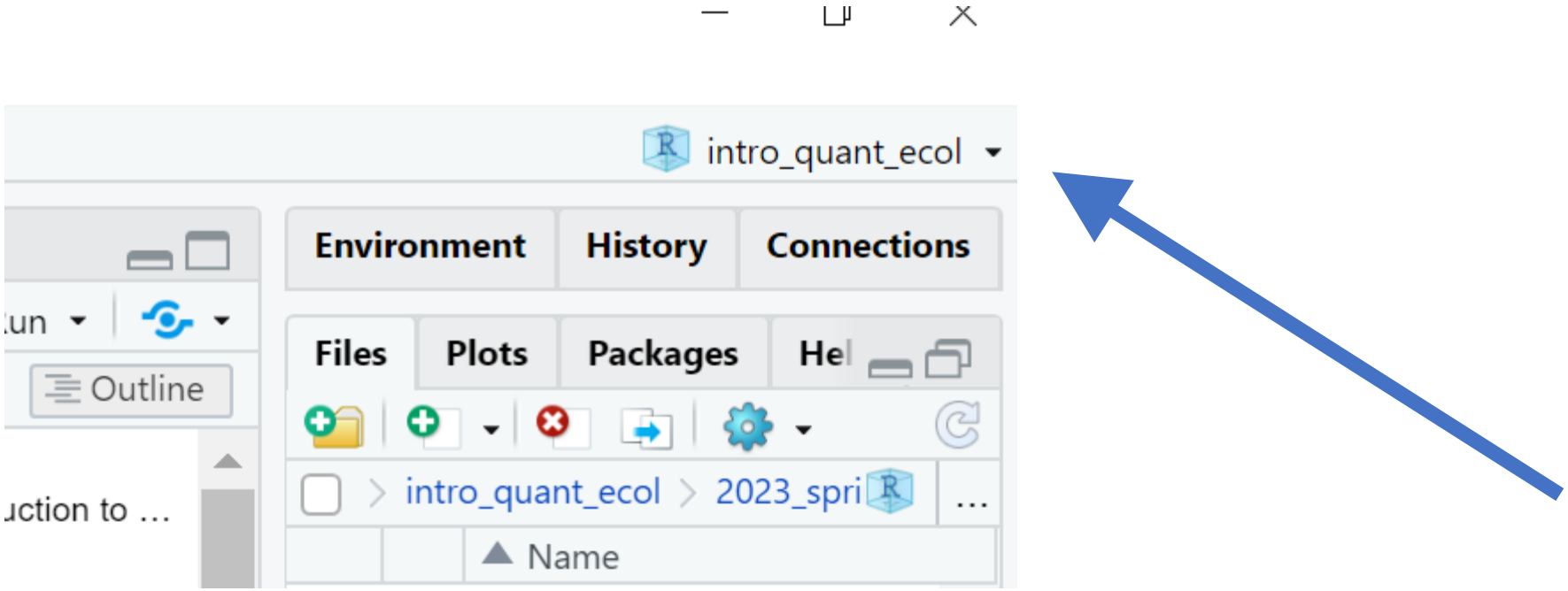
1. **Do NOT** install the Excel Analysis Toolpak.
   - The Analysis Toolpak is not available for the browser-based version of Excel.
   - As needed, I will provide updated information and instructions for passages in the book the refer to analyses in Excel.

ECHO360

- Auto-Capture is working!

RStudio and RProjects

- Make sure your RProject is loaded!



Announcements: Assignment Submission Conventions

- Submit 1 document per assignment (not individual screenshot files, for example)
- When you submit your assignments:
  - Answer the questions in order.
  - Clearly indicate which answer goes with which question

Slight Schedule Change

- Slight change in schedule (now updated on the website): We'll do the logical subsetting in-class exercise on Thursday. Today we'll have time for Desert Shrubs in-class.
  - Make sure you ask for explanations and demos in class!

Software Setup and Syllabus/Website Assignments

- You may revise/resubmit until you get full credit.
- Software setup question 3, hello world.
  - Original text of the question didn't emphasize the need for quotation marks. This has been updated on the assignment page.

Refresh Course Pages

- As the semester progresses, I (or Ana, or students) notice typos and ambiguous language on the various course pages.
- I try to fix these right away, but the changes won't appear in your browser unless you refresh the pages!
- TLDR: Refresh the course pages on github frequently!

Desert Shrubs Clarifications

- Sampling: every group will choose a sampling method.
- Sampling: every group will complete 9 sample counts.
- Your data sheet will have 312(!) lines.
  - You'll count a lot of quadrats.

Questions for me?

The Row Data Paradigm

All About Data

But… what's a paradigm?

Essential Meaning of *paradigm*

formal

1 : a model or pattern for something that may be copied

// Her recent book provides us with a new *paradigm* for modern biography.

2 : a theory or a group of ideas about how something should be done, made, or thought about

// the Freudian *paradigm* of psychoanalysis

// a new study that challenges the current evolutionary *paradigm*

Thanks Merriam-Webster!

The Row-Data Paradigm: Sampling Units

What's a sampling unit?

- The level of focus of our inquiry: an entity of interest
- The particular type of 'thing' that we want to know about.
  - How we define a sampling unit may change based on our particular question.

The Row-Data Paradigm: Sampling Units

Sampling units have **attributes**

- Examples:
  - Individual biological units: a salamander
  - Individual habitat: a lake, a patch, a pond, a mountaintop
  - Aggregate biological entity: a population of salamanders
- Sampling units are rows, their attributes in columns.

The Row Data Paradigm

Rows

- A row is a **sampling unit**
- Represents a single observation
- Examples
  - A salamander
  - A garlic mustard plant
  - A vernal pool
  - A forest patch

Columns

- A column is an **attribute,** i.e. a **variable**
- Represents a property of an observation
- Examples:
  - Salamander snout to vent length (SVL)
  - Salamander body mass
  - Number of flowers or seeds on a plant
  - Salamander population in a vernal pool
  - Water pH in a vernal pool

Data: some key concepts

Data

Data consist of observations that we hope to convert from information to understanding or knowledge.

Datasets consist of observations and attributes

The distinction of data and metadata isn't always clear, for example:

- The identity of the observer may be different for each observation, or it may apply to an entire dataset

Metadata

But… what is metadata?

- Applies to the whole dataset: metadata helps us understand the whole dataset
- Examples:
  - Sampling location
  - Units of measurement
  - Weather at the time of sampling
  - Unusual conditions
  - What is the spatial data projection?
  - What is the sampling unit?
- "Data without metadata is meaningless"

Recording Data

Important data recording concepts

What are some important questions your data sheets should answer?

- Who, what, where, when
- Variables (there may be multiple variables in each category)
- Notes: unusual conditions, etc.

Which of the above could be considered metadata?

It's better to record all the data and other observations that could be relevant later, even if you don't end up using all of the information for your analyses. It may be useful to future researchers

Variables: key concepts and terminology

Several pairs of terms are often used to describe different types of variables:

- Independent/dependent
- Predictor/response
- Explanatory/response

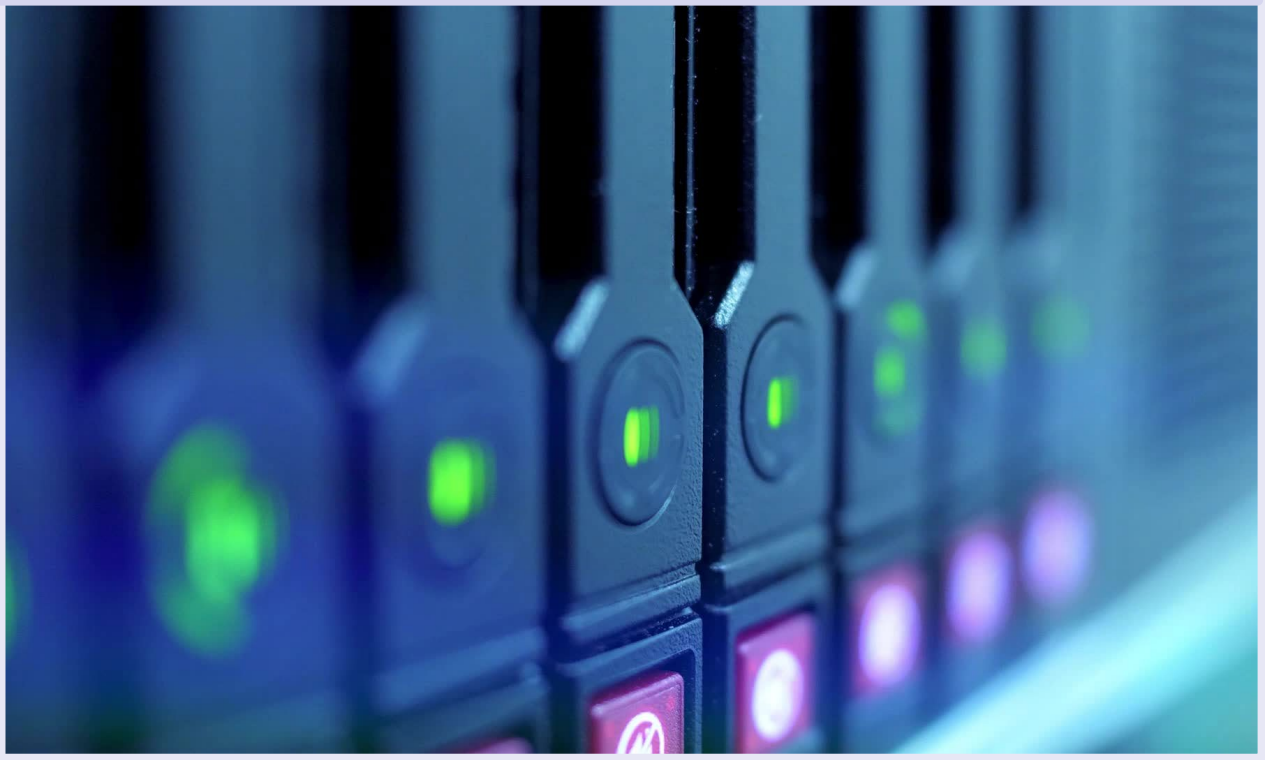
What do they mean?

What are some important data types?

What to record? Metadata

What is Metadata?

- Who recorded the data?
- What is the *sampling unit*?
- Where were the data recorded?
- When were the data recorded?
- How to decode numeric codes?

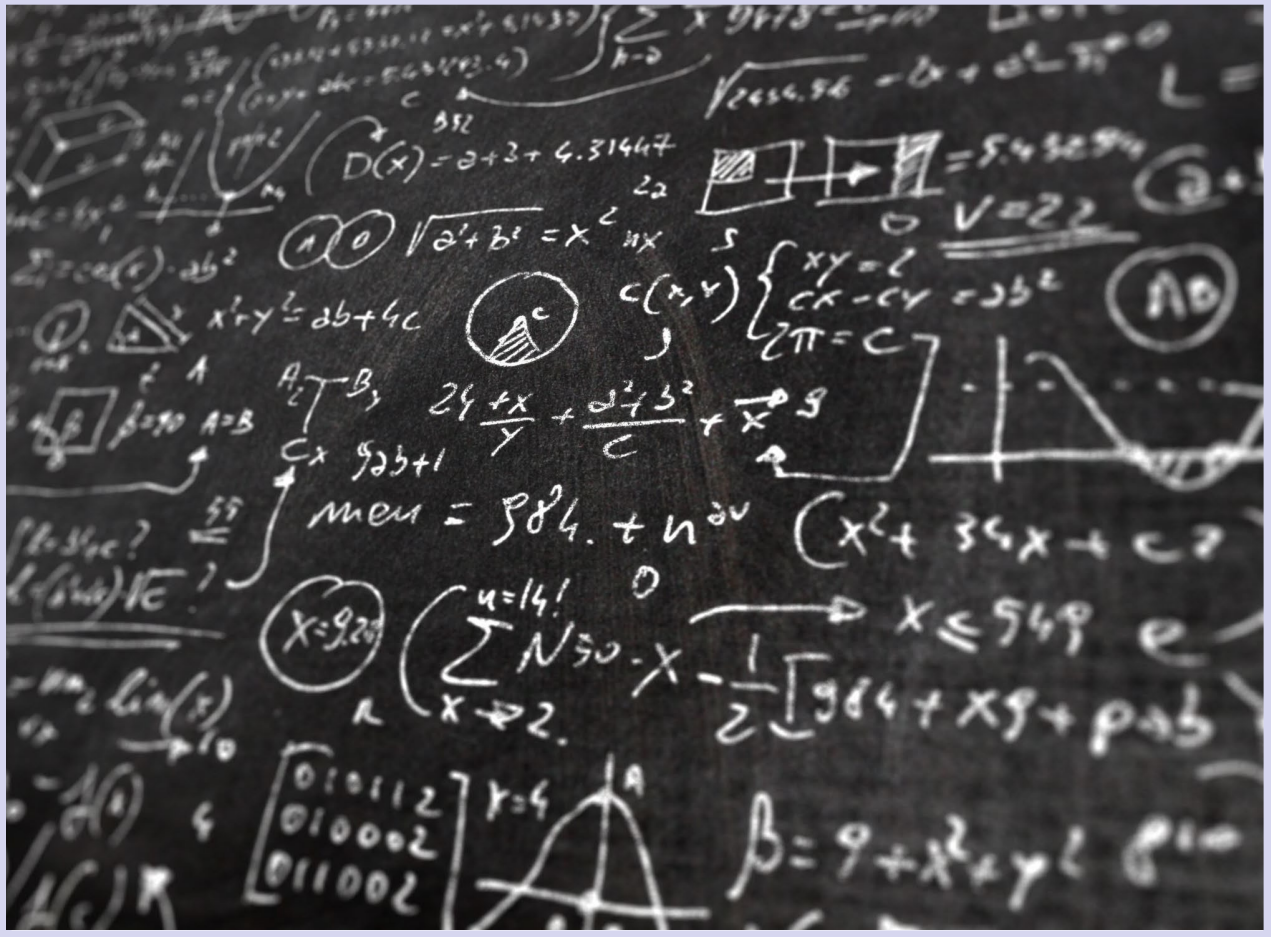


What to record?

The data: variables – response and predictors

Response Variable

Predictor Variables

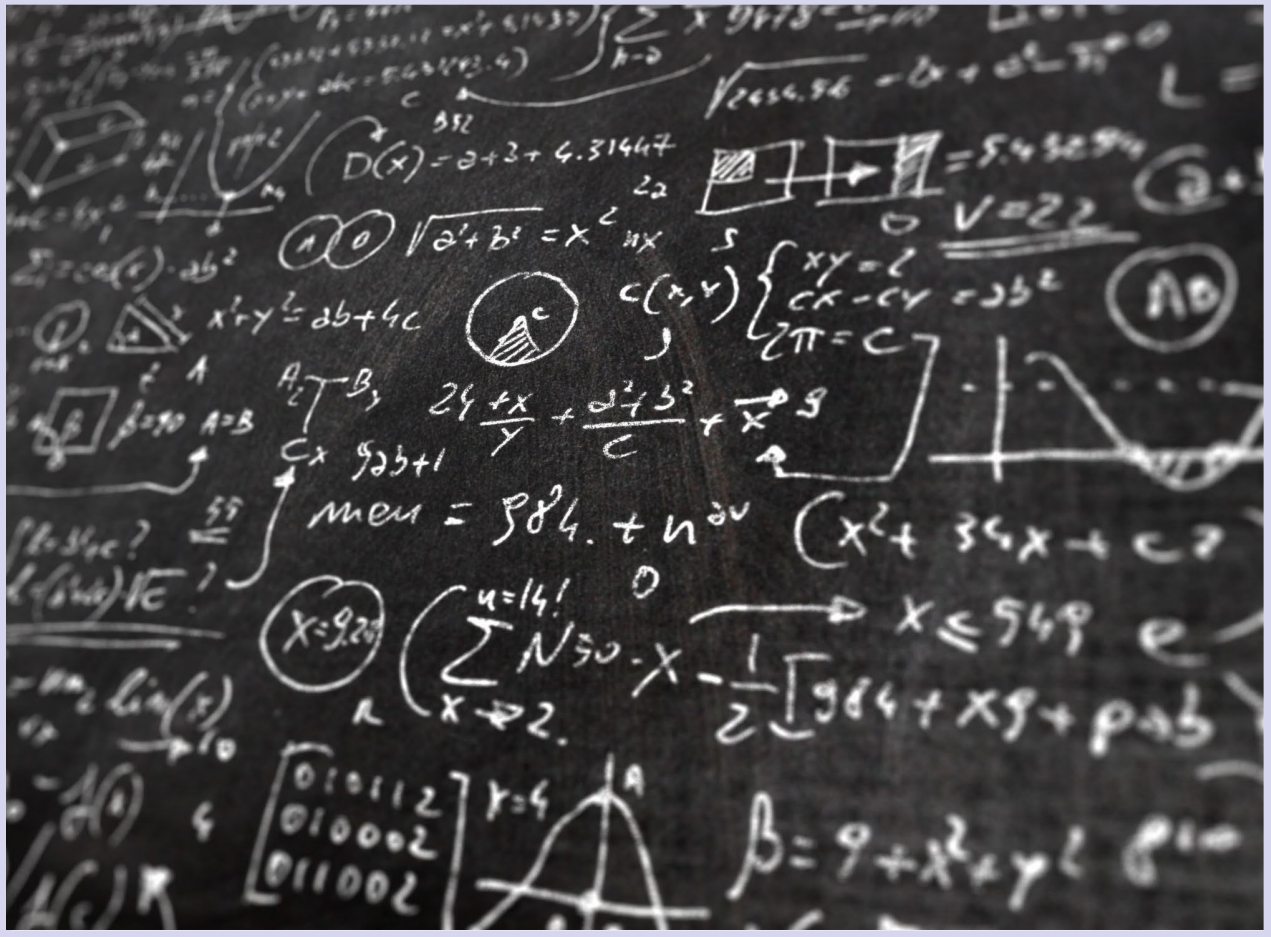


What to record?

The data: variables – response and predictors

Response Variable

- the variable of interest you are trying to estimate
- similar to/or sometimes called the “dependent variable”
- There is often only **one** response variable.

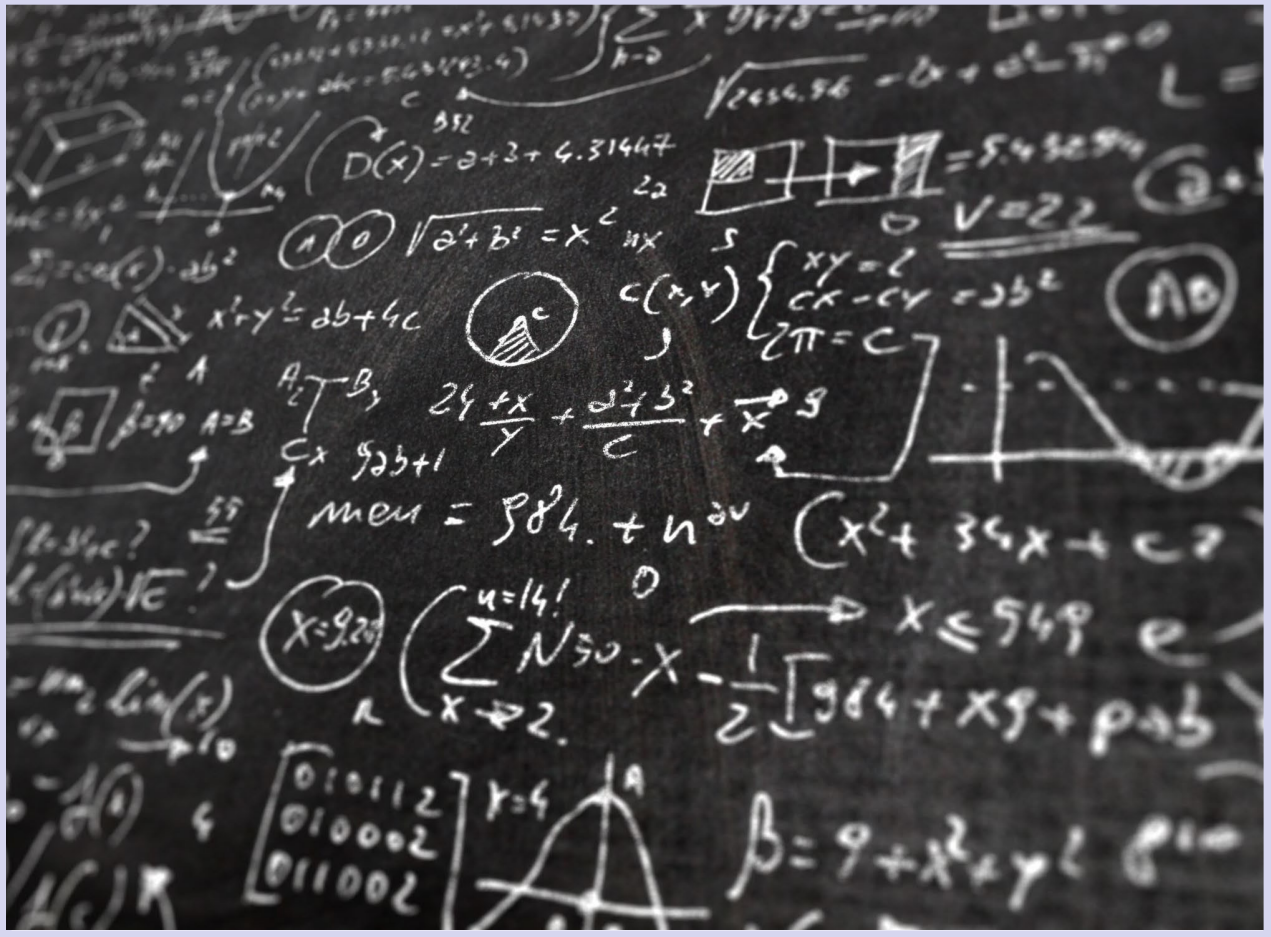


What to record?

The data: variables – response and predictors

Predictor Variables

- Variables that we think **might** influence the response variable
- similar to/or sometimes called the “independent variables” or “predictor variable”
- There can be **many** predictor variables.



In-Class Time For Desert Shrubs Assignment.

Questions for me before we start?

Variables: Predictors and Responses

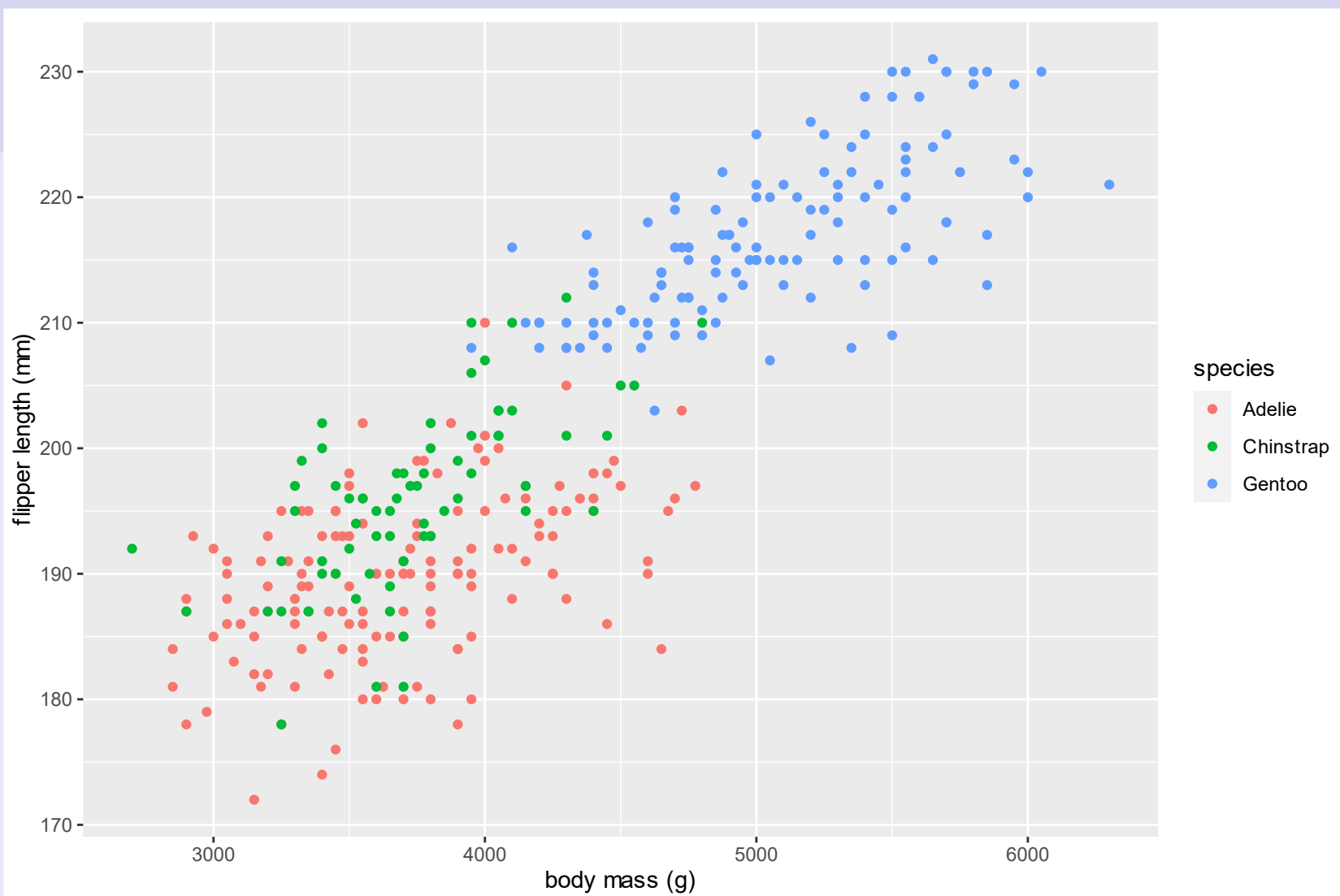
Predictors and Responses

Predictor Variables	Response Variables
• A.K.A. independent variables • Also called explanatory variables • Variables that we manipulate, in an experiment • Variables that we ant to use to understand changes in another quantity (the response variable) • Typically plotted on the x-axis • There may be several predictors	• A.K.A. dependent or response variables. • Variable that contains a quantity we want to understand. • We don't manipulate the values of the dependent variable; we just record them. • Typically, there is only one response variable. • The response variable is usually plotted on they-axis

Variables example: Can you identify the predictor and response in these graphs?

What is the?

- Response
- Predictor
- Grouping factor

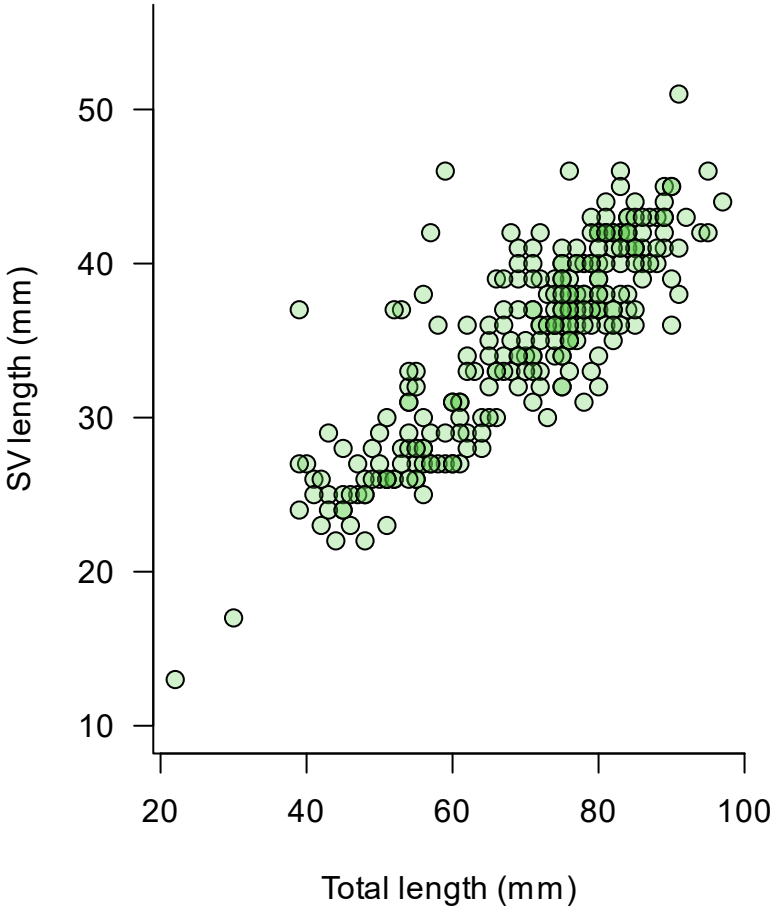


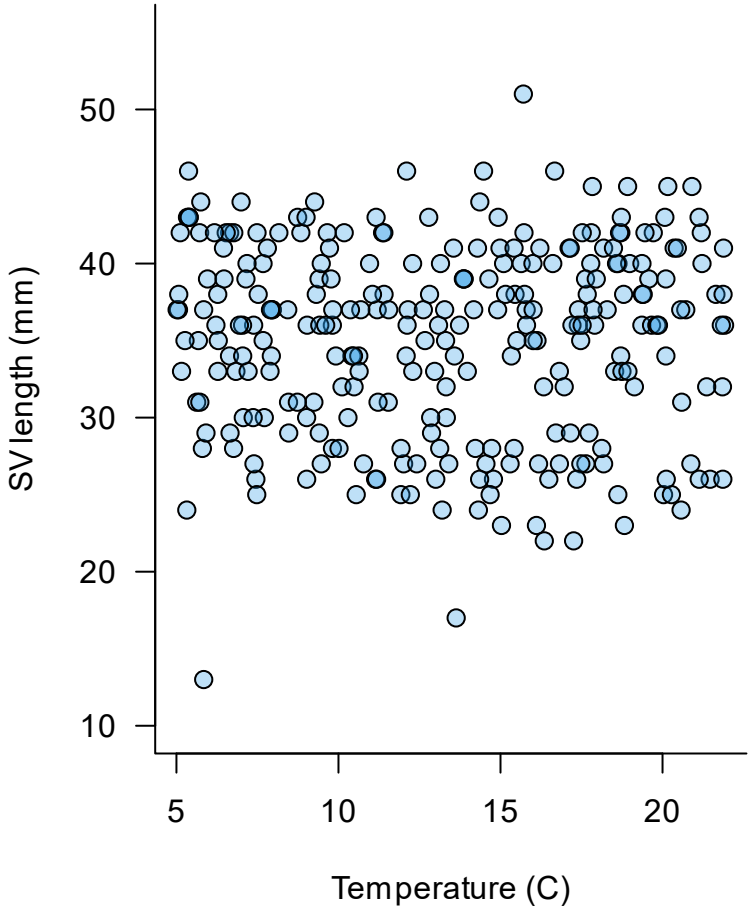
Variables example: Can you identify the predictor and response in these graphs?

What is the?

- Response
- Predictor

SV length = snout – vent length of salamanders





Some examples of variables

Can you identify the predictor and response variables in these *models*?

From high school math, you might remember the equation for a line: $y = mx + b$

We'll learn all about linear statistical models that look like this:

$$y_i = \alpha + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \epsilon$$

Model Thinking: Why did I call these equations models?

Subsetting by Name and Position

Subsetting a Data Frame

By Position (index)	By Name
• Recall the 2D data structure. • Uses square brackets. [,] • First entry is row • Second entry is column • Row or column may be blank to select all.	• All about column names! • Uses the dollar sign • Extracts an entire column. • Position of the column doesn't matter, as long as it has a name.

Subsetting Using Logical Tests

Subsetting Using Logical Tests

What are logical tests? In R there are several common tests:

- Equality
- Non-equality
- Strictly greater or less than
- Greater/lesser or equal to
- Does an item appear in a set?

Logical Test Syntax

Equality: 'double equals'

```
> # Are two numbers equal?
>
> 1 == 1
[1] TRUE
> 1 == 2
[1] FALSE
>
> # Are two strings equal?
>
> "abc" == "abc"
[1] TRUE
> "Abc" == "abc"
[1] FALSE
>
```

Equality for variables and vectors

- Equality tests are 'vectorized'

```
> # Create a numberic vector called a
> a = 1:4
> a1 = 1:4
> a2 = 1:5
>
> a == 1:4
[1] TRUE TRUE TRUE TRUE
> a == a1
[1] TRUE TRUE TRUE TRUE
> a == a2
[1]  TRUE  TRUE  TRUE  TRUE FALSE
Warning message:
In a == a2 :
  longer object length is not a multiple
 of shorter object length
```

Logical Test Syntax

Inequality: !=

It's like the inverse of ==

```
> # Are numbers unequal?
>
> 1 != 1
[1] FALSE
> 1 != 2
[1] TRUE
>
> # Are two strings unequal?
>
> "abc" != "abc"
[1] FALSE
> "Abc" != "abc"
[1] TRUE
```

Inquality for variables and vectors

- Equality tests are 'vectorized'

```
> # Create a numberic vector called a
> a  = 1:4
> a1 = 1:4
> a2 = 1:5
> a  != 1:4
[1] FALSE FALSE FALSE FALSE
> a  != a1
[1] FALSE FALSE FALSE FALSE
> a1 != a2
[1] FALSE FALSE FALSE FALSE  TRUE
Warning message:
In a1 != a2 :
  longer object length is not a multiple
 of shorter object length
```

The subset() Function

Subsetting is an art and a science

- Subsetting forces us to think about the structure of our data
- Subsetting forces us to thing about the content of our data.
- Logical operations are your best friend!

The penguin data:

```
> head(penguins)
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm
  <fct>   <fct>              <dbl>         <dbl>             <int>
1 Adelie  Torgersen           39.1          18.7               181
2 Adelie  Torgersen           39.5          17.4               186
3 Adelie  Torgersen           40.3          18                 195
4 Adelie  Torgersen           NA            NA                  NA
5 Adelie  Torgersen           36.7          19.3               193
6 Adelie  Torgersen           39.3          20.6               190
# ... with 3 more variables: body_mass_g <int>, sex <fct>,
#   year <int>
```

How Can I Subset the Penguin Data?

Strategies	Syntax
• Step back and think about what subset of the data you want. • Do you need all female penguins? • Do you need penguins above a certain body mass? • Do you need female penguins from a particular island	```# retrieve female penguins subset(penguins, sex == "female") # Retrieve heavy penguins subset(penguins, body_mass_g > 5000) # Retrieve Adelie penguins on Biscoe island subset(penguins, (island == "Biscoe") & (sex == "female"))```

Data Management

Data Cleaning

No dataset is perfect!

You should always inspect your data before using it, R has some helpful functions:

- head() will preview the first 6 rows of a 2D data structure (like a matrix or data frame)
- Summary() will give you a 5 number summary of each column. Will tell you the number of NA values.

Data Cleaning

Things to look out for:

- Missing data: R will code these as NA values
- Data recorded in the wrong column: these are harder to catch. If you have trouble running routine operations on your data in R, this is a good thing to look out for.
  - Check for values that seem unreasonable, for example a plant height of 100 meters.
  - Check for values recorded in the wrong format: text in a column that should be numeric

Data File Tips

Use intuitive filenames and directory structures.

- Create a ‘data’ directory in your project’s directory structure
  - Where have we heard this before?
- Avoid generic filenames like “New File(73).csv”
- Avoid spaces in filenames: use underscores instead.
  - “mander data.csv” should be “mander_data.csv”.
  - Many, but not all, modern software tools can recognize spaces in filenames, but they can cause insidious and hard to diagnose problems (I’m looking at you ArcGIS!).

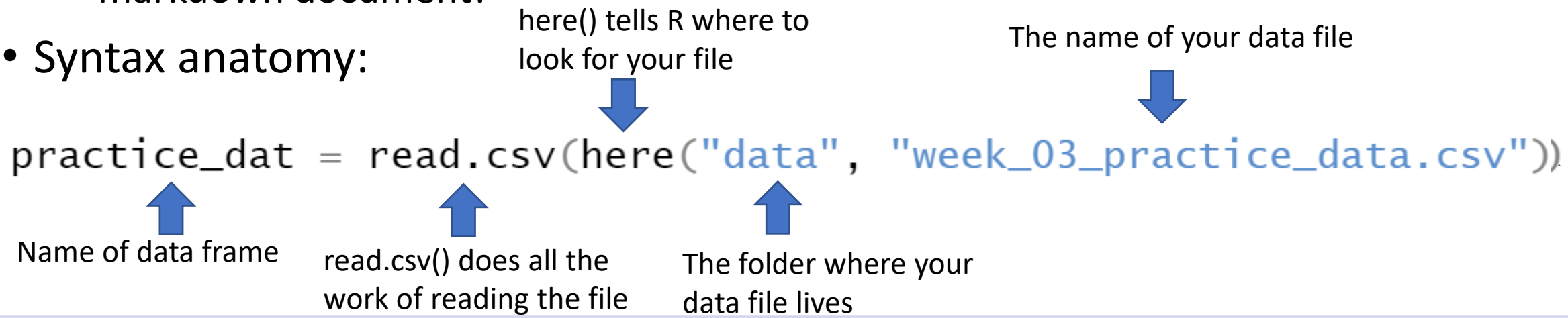
Data File Tips

- Filenames can be metadata:
  - mander_data_2003.csv, mander_data_2004.csv, etc.
- Avoid proprietary or binary formats, when possible. Prefer text-based formats like .csv or .txt.

Getting Data Into R

Review and Pointers

- Review the video and text walkthroughs if needed.
- Key pointers:
  - Always check that your RProject is open. This is the cause of about 50% of failures to read data!
  - Make sure your directory and filename are spelled correctly. It's easy to mistake an uppercase for a lowercase letter.
  - Do not use file.choose(). This method is not reproducible, and you can't use it in a markdown document!
- Syntax anatomy:



In-Class Data Import and Subsetting Exercise