Biodiversity Informatics

Introduction to our tools

Alban Sagouis, 17.04.2023

Questions

Who am I?

- Who am I?

Questions

Who are you?

- Who wants to work in private sector?
- Who wants to work in public academia?
- Who wants to work in ecology/biodiversity?

Questions

Who are you?

- Who has heard of R?

Questions

Who are you?

- Who knows what is functional programming?
- Who ever wrote a function in R or Python?

Questions

Who are you?

- What is your primary programming language?

Questions

Who are you?

- Who uses git?
- And GitHub/GitLab?

Questions

Who are you?

- Tell me more

Research process

- (Data acquisition)
- Checking and cleaning data
- Wrangling data
- Exploring data
- Statistical analyses
- Creating maps and plots
- (Writing articles)

Biodiversity data

What are we talking about?

- Origins
- Shapes of data
  - Categorical

Biodiversity data

What are we talking about?

- Origins
- Shapes of data
  - Categorical
  - Quantitative

Biodiversity data

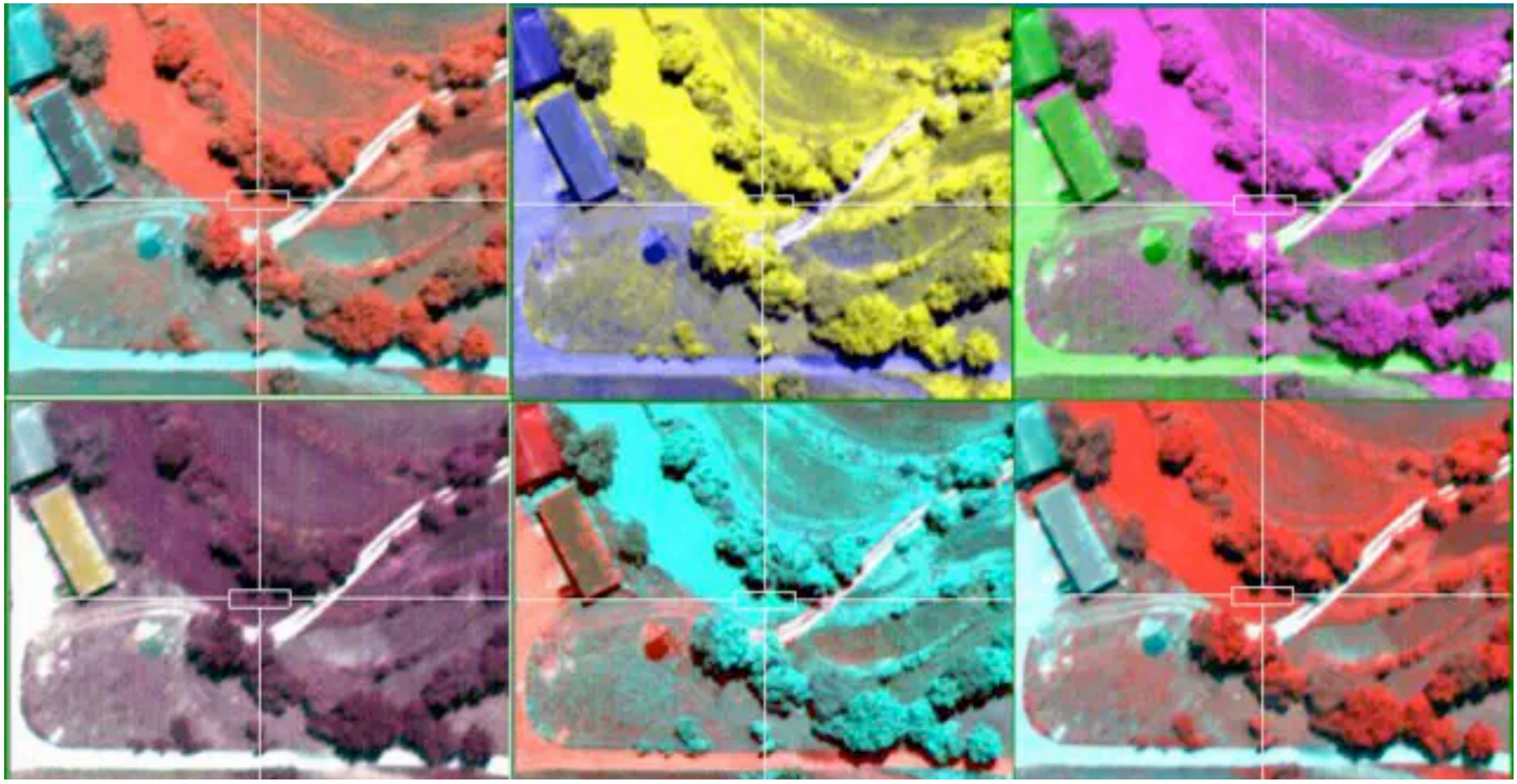
What are we talking about?

- Origins
- Shapes of data
  - Categorical
  - Quantitative
  - Graphical

Biodiversity data

What are we talking about?

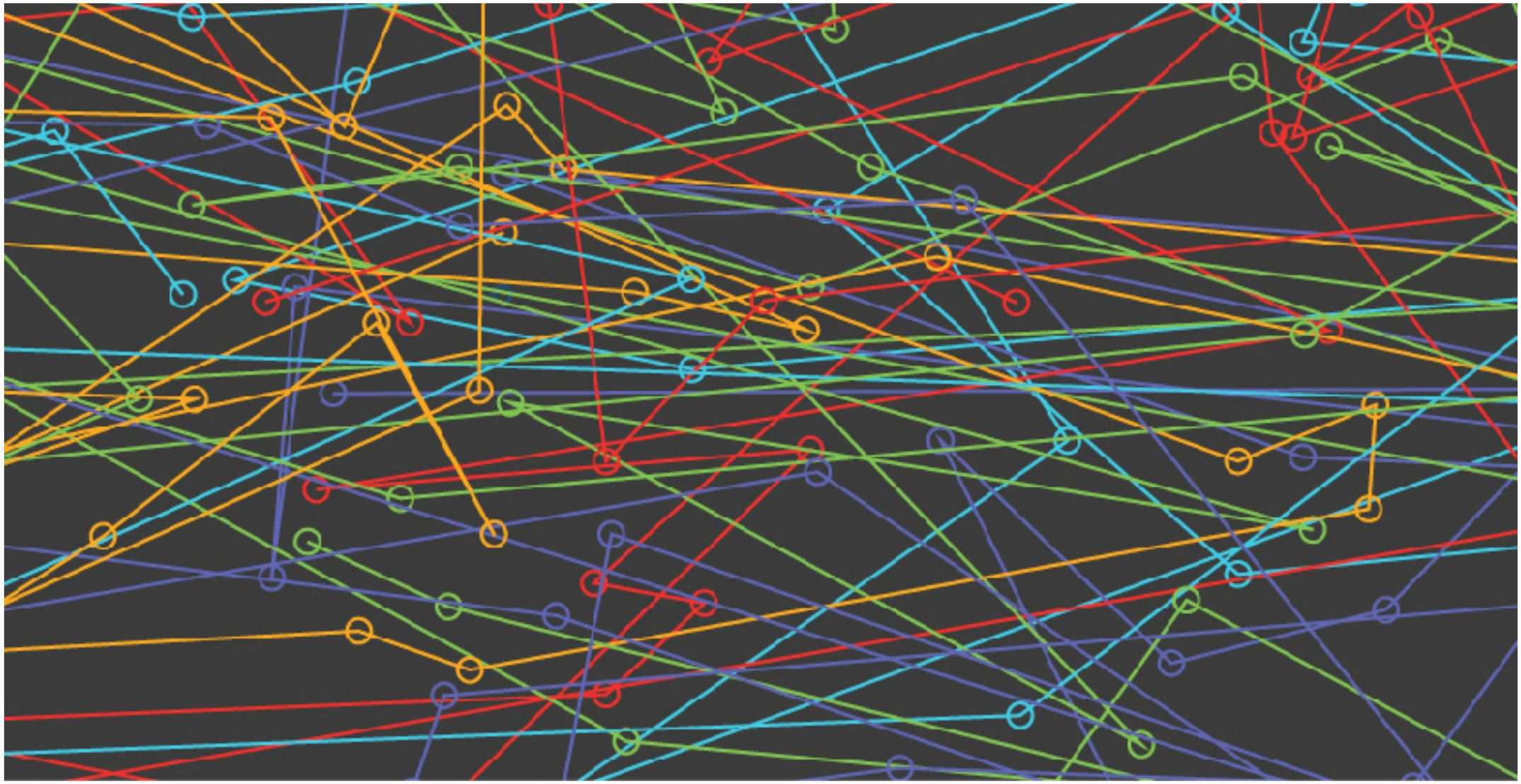
- Images / Rasters



Biodiversity data

What are we talking about?

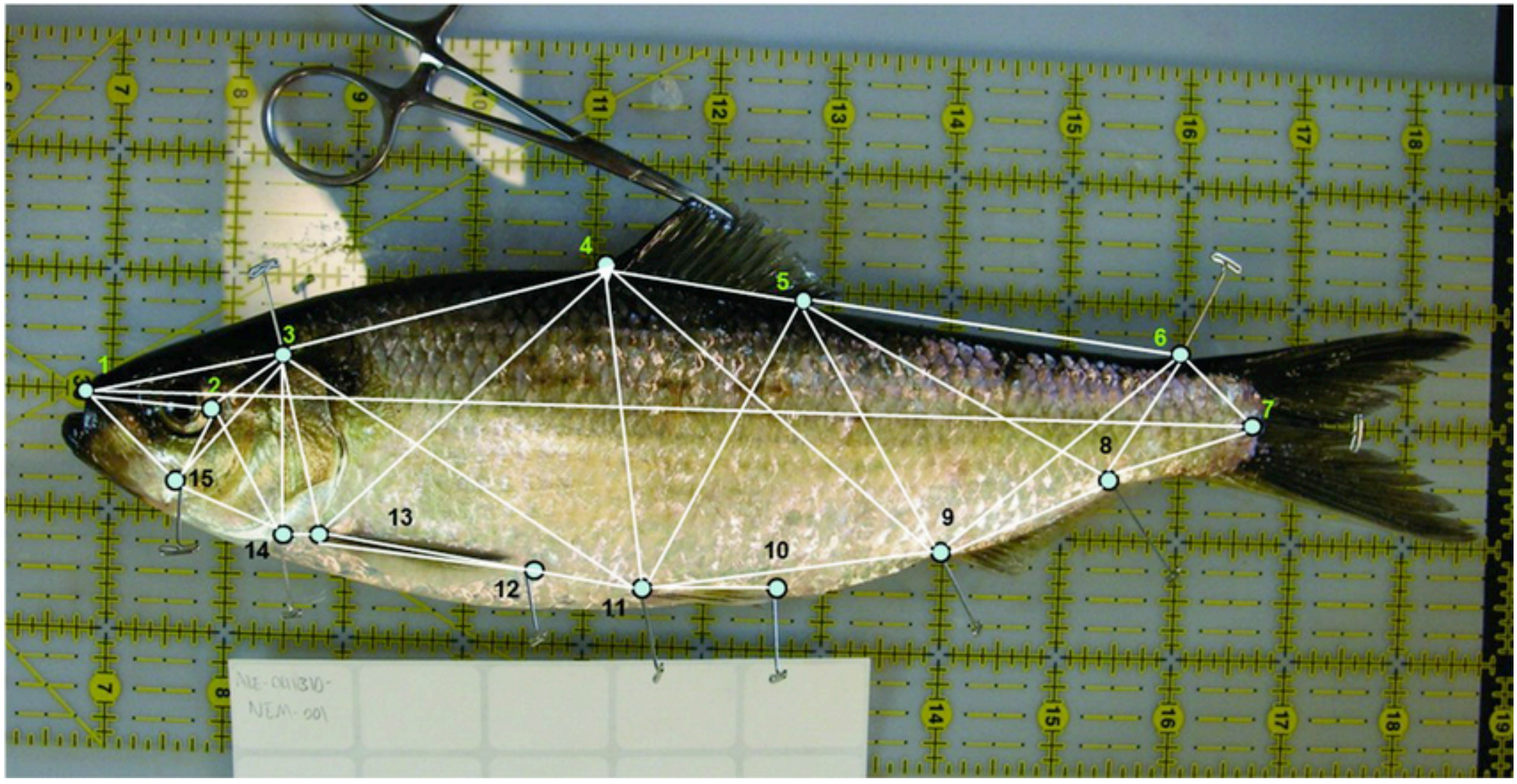
- Vectorised data



Biodiversity data

What are we talking about?

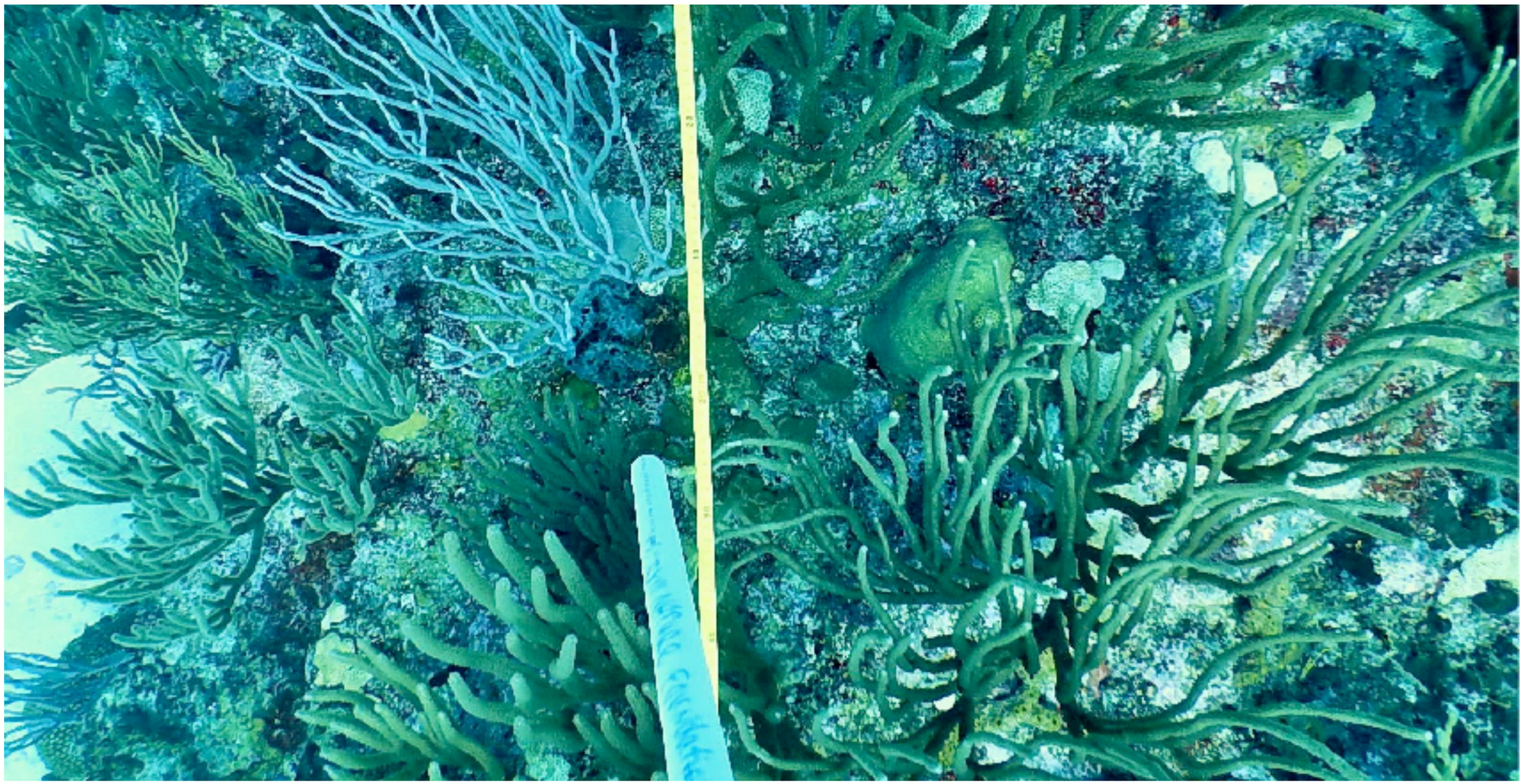
- Individual pictures



Biodiversity data

What are we talking about?

- Ecosystem picture



Biodiversity data

What are we talking about?

- Scales
  - Individual measures
  - Population
  - Community
  - Biodiversity metric
- Environment

Biodiversity data

What are we talking about?

- Meta-analyses
- Synthesis

Challenges

The everyday struggle

- Missing data

Challenges

The everyday struggle

- Missing data
- Unbalanced sampling

	Site A	Site B	Site C
2001	5	*NA*	8
2002	4	6	*NA*
2003	1	10	*NA*
2004	7	*NA*	9

Challenges

The everyday struggle

- Errors / problems
  - Missing metadata
    - Units
    - Coordinates: 43°32’10”, 43°54356, 43.54356, 7653454...NSWE
    - Taxonomical reference
- Encoding

```
Œμ–§,É,ÍA,Ë,±,Ë,±,Í‹C,Ã,¢,¼,ç¶,Ü,ê,Ä,¼,©,çA
’a¶“ú,Í•s–¾,±,±i,Ë,±,g,oj,ªo—ˆ,Ä,©,ç,Q,O”N,Æ,È,è,Ü,·B
```

An example of mojibake Japanese text misread as Latin-1 - Source

What we are working with

A summary

- Many different types of measures, units
- Many scales
- Many formats
- Many chances to get things wrong


R in ecology

R in ecology

- The R Project for Statistical Computing
  - R Core Team
    - First published in August 1993
  - R foundation
  - CRAN
    - CRAN originally had three mirrors and 12 contributed packages. As of December 2022, it has 103 mirrors and 18,976 contributed packages.
  - R-Hub

R in ecology

- Open source and free
- Interpreted (vs compiled) and easy
- Out of the box
  - Text manipulation
  - Mathematical operations
  - Basic statistical tests

R in ecology

- Very large collection of packages
  - Data treatment, GIS, image treatment, sound, video, coordinates...
  - Statistical analyses
- Possibility to make and share more

R in ecology

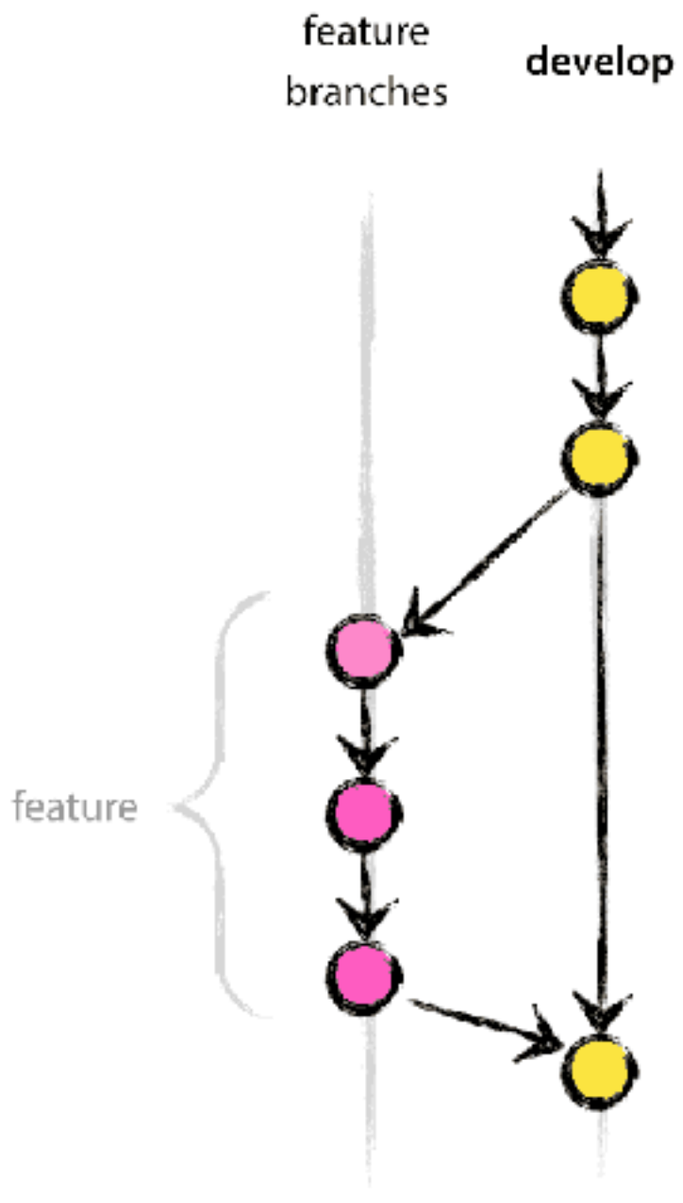
- Two ways to use it
  - Interactive session
  - Scripted way

Questions?

Git, GitHub and research in ecology

Git

- Development began in 2005
- Pushed by Linus Torvalds
- Open source and free
- Version Control System
- CLI
- Branching types
- GUI



GitHub

- Online repository
  - Allows people collaborating on code
  - Not meant for archiving
- Collaboration tools

Questions?

Scientific reliability

Introduction

- Scientific reliability
- Accuracy
- Reproducibility
- Transparency

Scientific reliability

- Continuously checking new decisions by
  - Testing code

Scientific reliability

Test your work

- Size of the data set
- Number of cases/rows increased or decreased
- Number of variable/columns changed

Scientific reliability

Test your work

- Size of the data set
- Number of cases/rows increased or decreased
- Number of variable/columns changed
- Quality of the data set
- Are all Realm values one of Freshwater, Terrestrial or Marine?
- Are all abundances of integer type?

Scientific reliability

Test your work

- Comparing models
- New data was added
- New parameters were added

Scientific reliability

Test your work

- Ensuring reproducibility
- For other users
  - On Windows, MacOS or Linux systems

Scientific reliability

Test your work

- Documentation
- Containerisation
- Archiving

Developing packages

- R packages should be checked on all operating systems (Linux, Mac OSX, Windows) when they contain:
  - Compiled code
  - Java dependencies
  - Dependencies on other languages
  - Packages with system calls
  - Anything depending on encoding
  - Anything with file system / path calls

Questions?

Research process

- What I do

Exercises

Git & GitHub

- Let's explore GitHub.com/chase-lab
- Get the data from the biodiversity_informatics_2023
- Download the repo
- Clone it locally
  - Using GitHub Desktop
  - Using git

Exercises

R basics

- Download and install R
- I recommend using RStudio as an IDE (Integrated Development Environment)

Exercises

R basics

- For the first contact with R: https://github.com/hbctraining/Intro-to-R-flipped/blob/master/lessons/01_introR-R-and-RStudio.md#interacting-with-r
- For an introduction to R basics: https://informatics.fas.harvard.edu/introduction-to-r-workshop.html#getting-help
- For an introduction to the tidyverse: https://jhudatascience.org/tidyversecourse/intro.html#the-tidyverse-ecosystem
- For an introduction to the package data.table: https://rdatatable.gitlab.io/data.table/articles/datatable-intro.html
- For an introduction to Git: https://phoenixnap.com/kb/how-to-use-git
- For an introduction to GitHub: https://github.com/skills/introduction-to-github