



Marine Ecological Modelling Global Climate Change

Marine Data Science

Jorge Assis, PhD // jmassis@ualg.pt // jorgemfa.medium.com
2020, Centre of Marine Sciences, University of Algarve



Marine data science

"an exciting new discipline that **turns raw data** into **understanding, insight, and knowledge**".

(Grolemund & Wickham 2016).

Biology of species

Statistical modelling

Handling (big) data

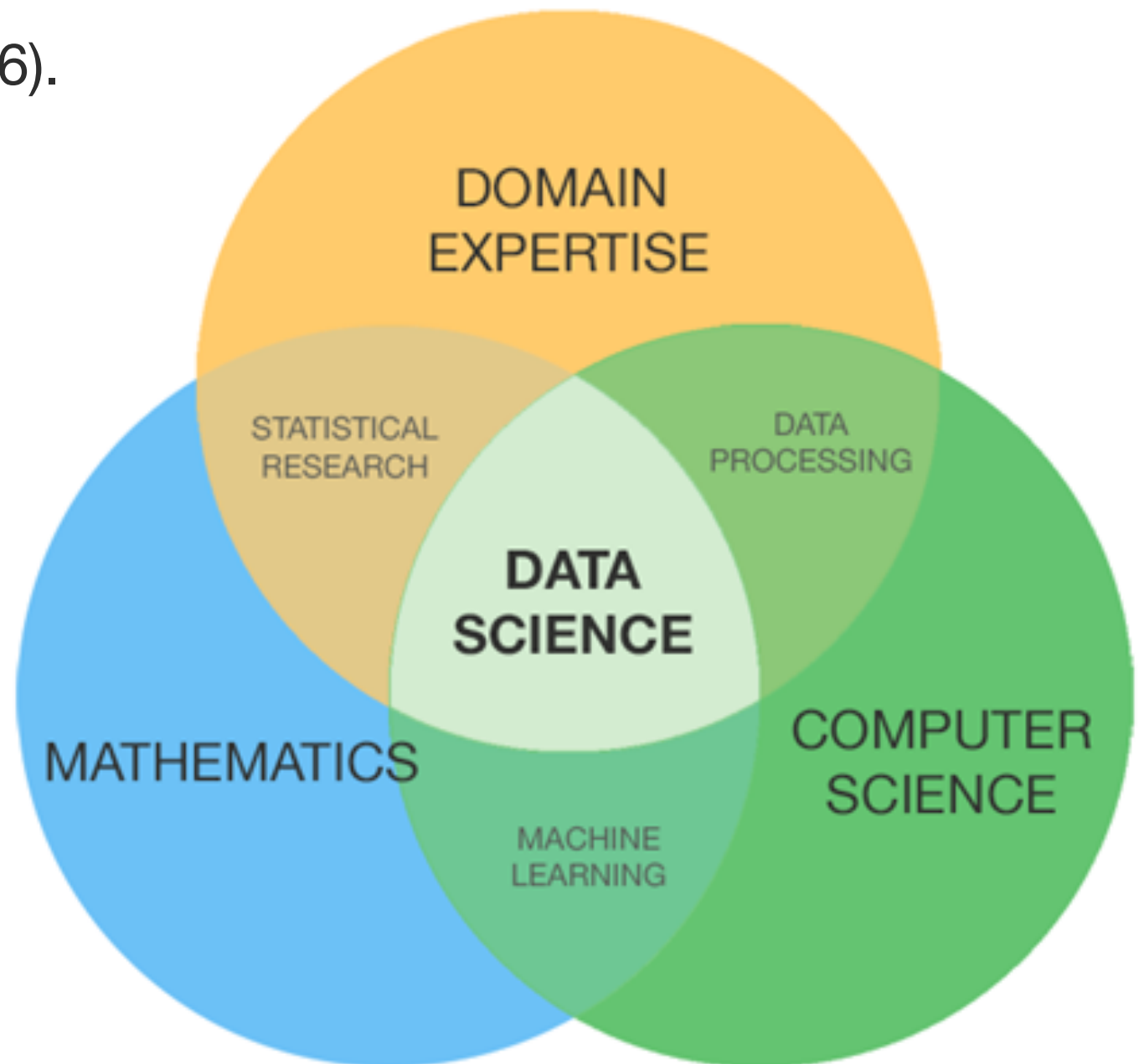
Data processing

Data visualization

Descriptive analyses

Machine learning

Simulation scenarios





Open science tools

“allows **transparency at all stages of the research process**, coupled with **free and open access to data, code, and papers**”.

(Hampton et al. 2014)

Coding language [**R language**]

Coding environment, editor, visualization and support [**R Studio**]

Organization, collaboration and version control [git; **GitHub**]

this talk: <https://github.com/jorgeassis/>

written in RStudio's RMarkdown

versioned with Git

shared with GitHub



Reproducibility and optimization

“According to interviews and expert estimates, **researchers spend up to 50 percent** of their time mired in the mundane labor of **organizing and preparing data**”. NYTimes (2014)

Transforming, rescaling, gap-filling, formatting, renaming, etc.

Underpins the scientific process.

Before

Manually (without coding);

Large Excel processes;

Internal documents and emails.

Now (reproducible)

Full coded process;

RMarkdown.



Collaboration and communication

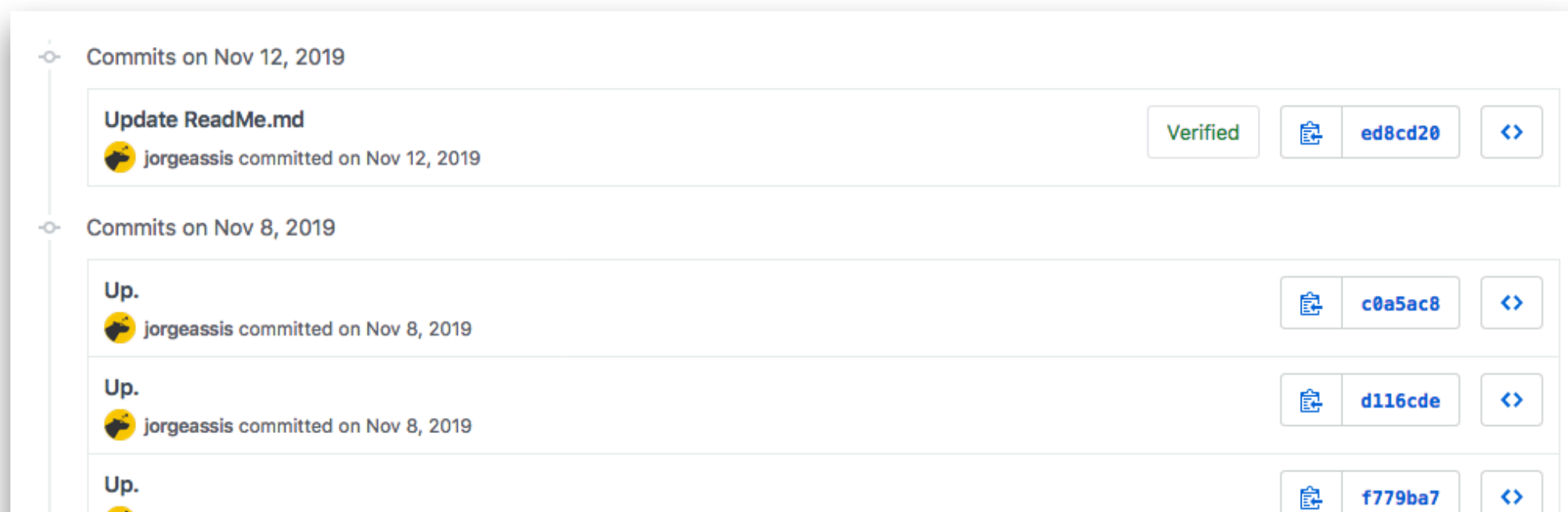
“For scientists, [Git] works like a notebook for scientific computing... it keeps a lasting record of events”. Nature 2016

Before

Filenames suffixed with dates, initials (e.g., final_JL-2016-08-05.csv);
Email chains (often forwarded).

Now (reproducible)

Version control with git;
Short messages accompany committed changes.





Sharing data

Before

Published manuscripts;

Data on personal FTP server (online?) and in supplementary information.

Now (reproducible)

Published manuscripts

Data open on GitHub, Figshare, etc.



jorgeassis / marineforestsDB

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Actions Projects 0 Wiki Security Insights Settings

A fine-tuned global distribution dataset of marine forests

Edit

r dataset marine records Manage topics

233 commits 1 branch 0 packages 0 releases 1 contributor

Branch: master New pull request Create new file Upload files Find file Clone or download

| | | | |
|-----------------------------|------------------|---------------|---------------------------------------|
| jorgeassis Update ReadMe.md | | | Latest commit ed8cd20 on Nov 12, 2019 |
| Data | Up. | 3 months ago | |
| Functions | Up. | 3 months ago | |
| .DS_Store | Up. | 11 months ago | |
| .gitignore | Up. | 3 months ago | |
| Git.Rproj | Up. | 3 months ago | |
| ReadMe.md | Update ReadMe.md | 3 months ago | |
| sourceMe.R | Up. | 3 months ago | |

ReadMe.md

A fine-tuned global distribution dataset of marine forests

J. Assis, E. Fragkopoulou, D. Frade, J. Neiva, A. Oliveira, D. Abecasis, S. Faugeron, E.A. Serrão

Abstract

Species distribution records are a prerequisite to follow climate-induced range shifts across space and time. However, synthesizing information from various sources such as peer-reviewed literature, herbaria, digital repositories and citizen science initiatives is not only costly and time consuming, but also challenging, as data may contain thematic and taxonomic errors and generally lack standardized formats. We address this gap for important marine ecosystem



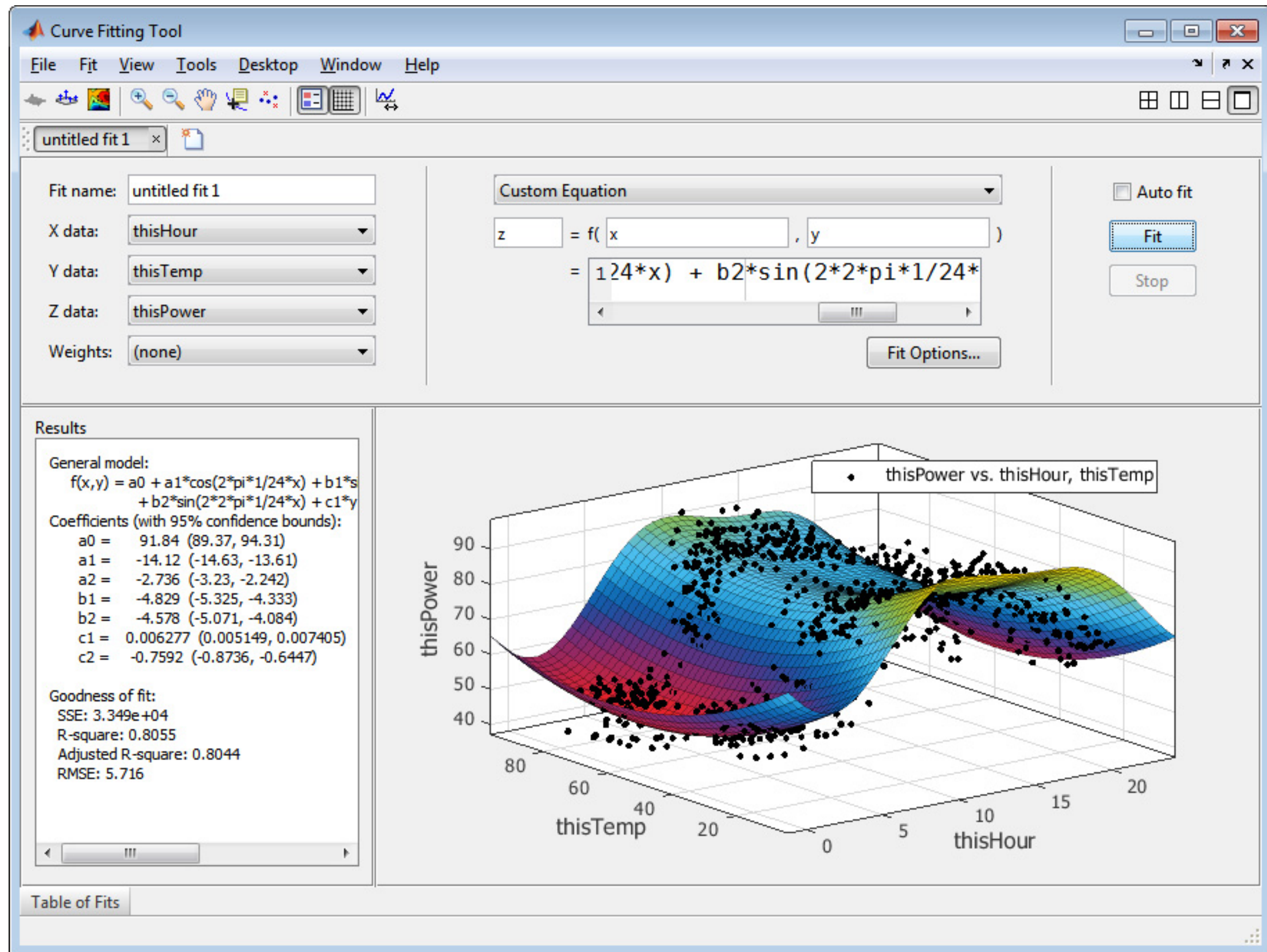
My programming origin

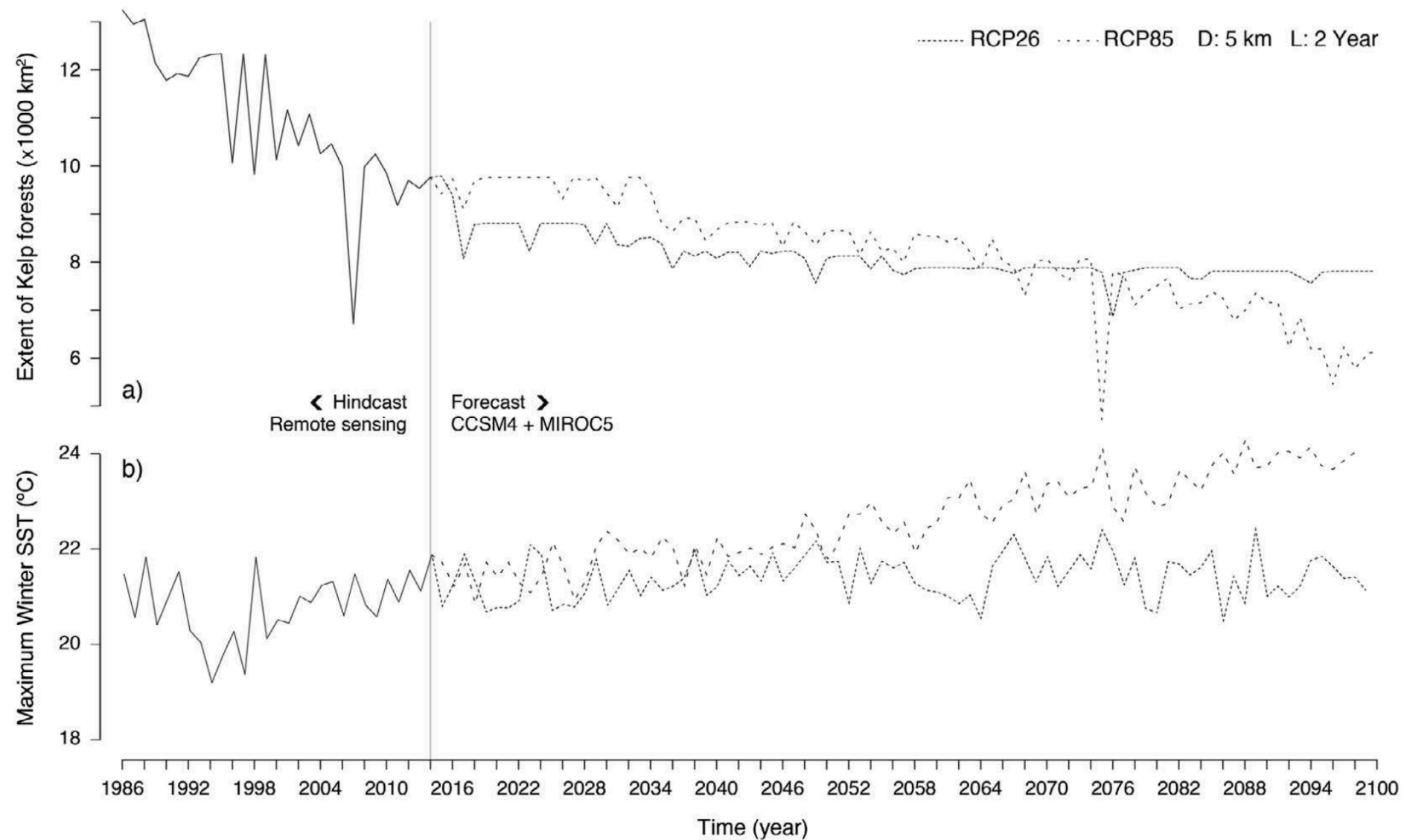
Being curious about how things were made.

Changing stuff and see what happens.

Trial and error, error, error, error, ...



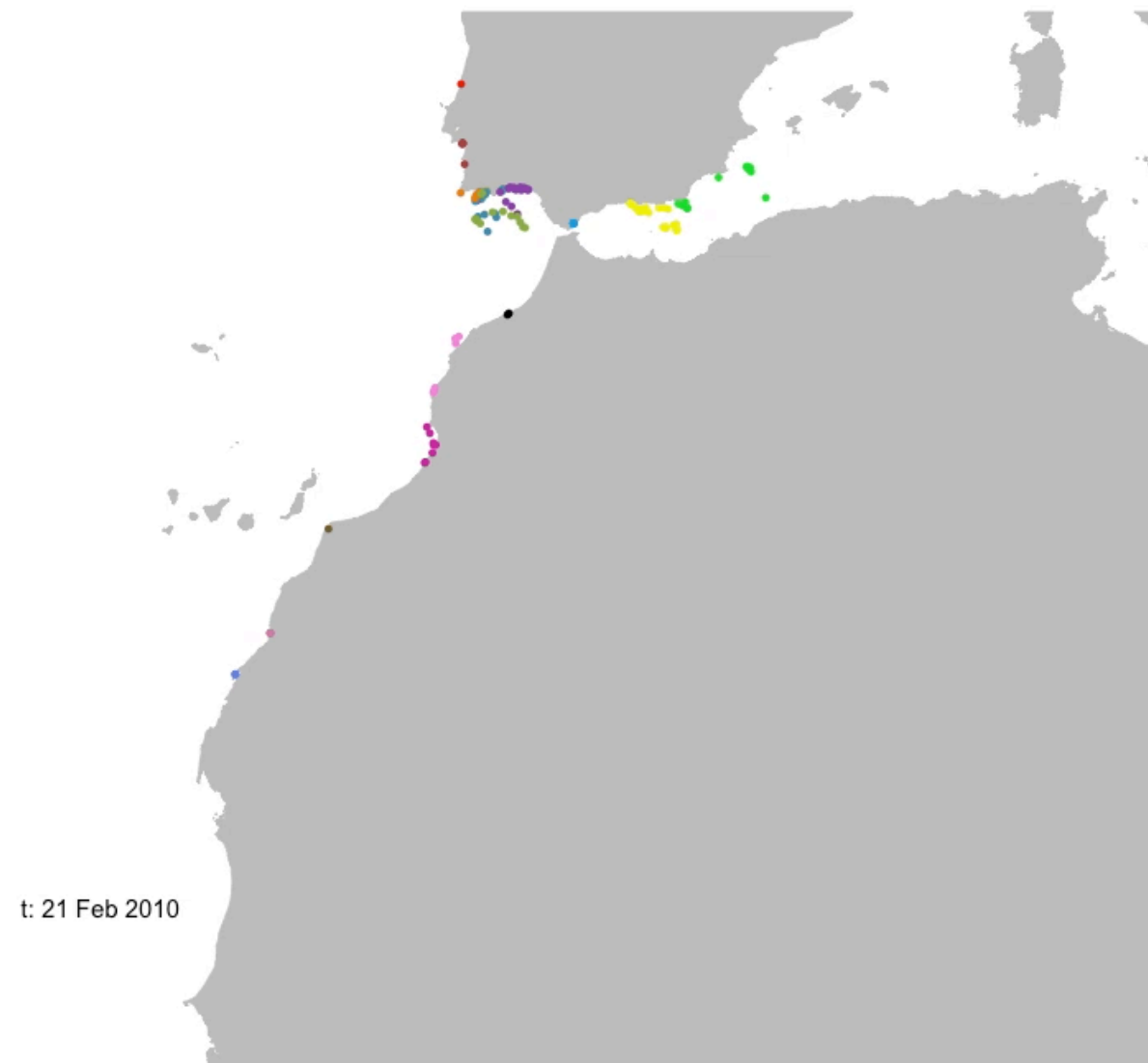




Modelling the potential distribution of species.

[different climate scenarios producing shifts in distributions]

> 2013



**Generating virtual oceans to address marine connectivity.
[ocean currents mediating population connectivity]**



Weak biodiversity connectivity in the European network of no-take marine protected areas

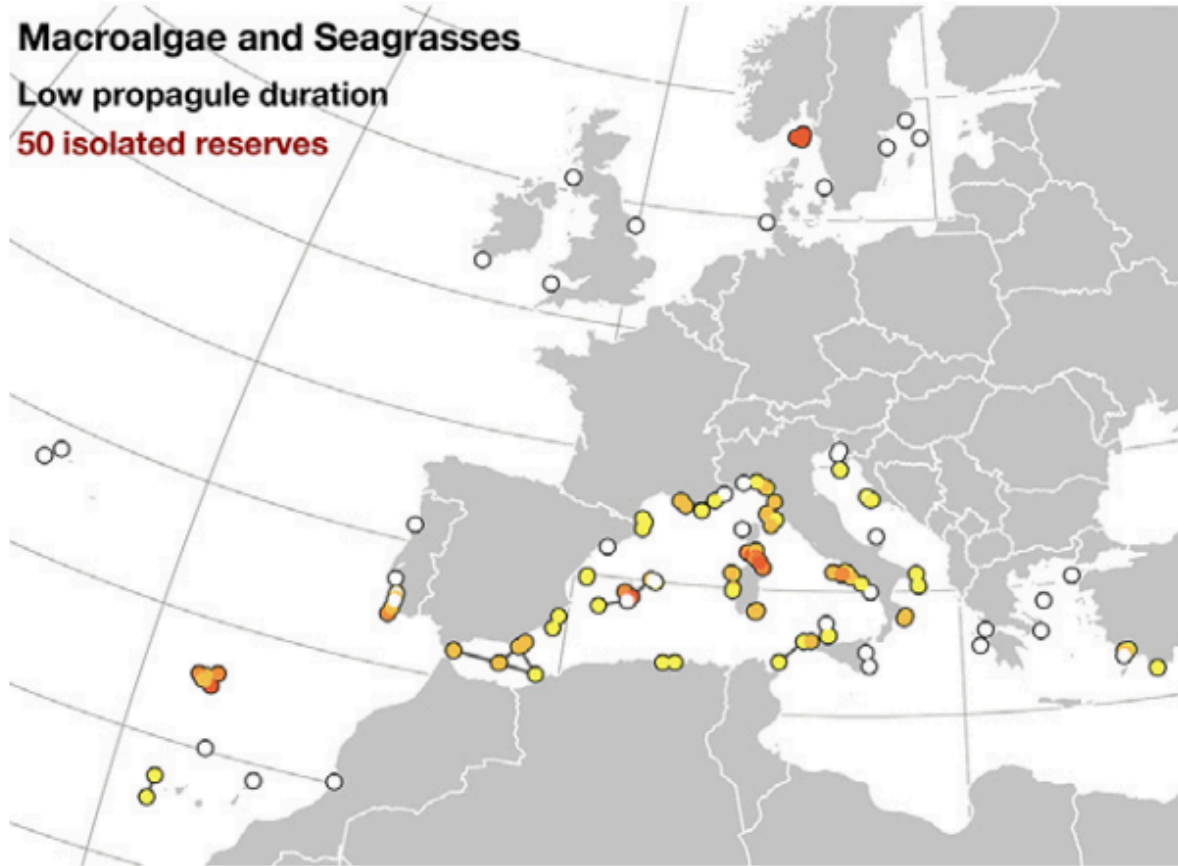
[Hydrosphere + Biosphere + Anthroposphere]

Connectivity degree: - +

Macroalgae and Seagrasses

Low propagule duration

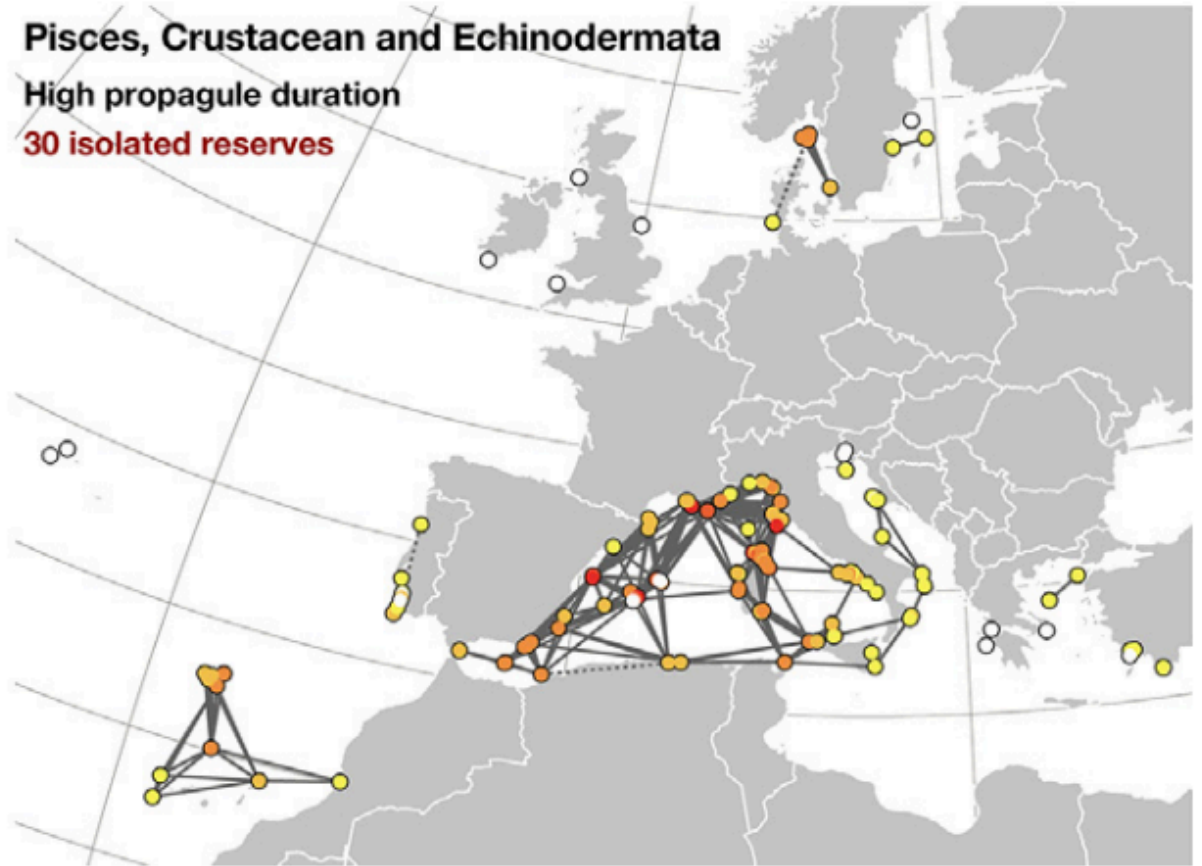
50 isolated reserves



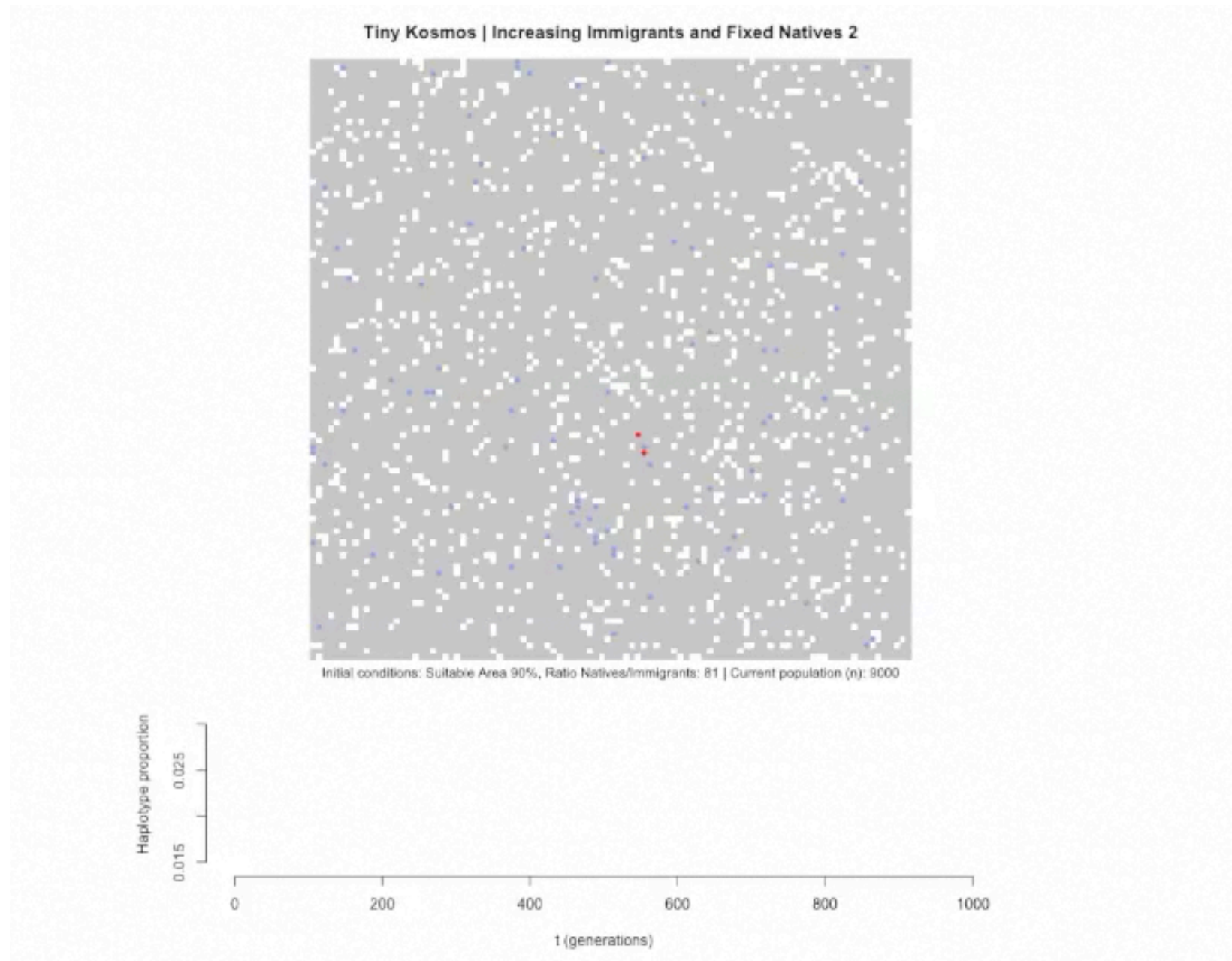
Pisces, Crustacean and Echinodermata

High propagule duration

30 isolated reserves



Generating virtual oceans to address marine connectivity.
[ocean currents mediating population connectivity]



Generating tiny cosmos for immigrant populations.
[test assumptions like the winner takes it all]



Errors are friendly

Computer errors are just like someone saying 'I didn't understand what you mean'. Google it (copy-and-paste!) or use Stack Overflow.

“Plotting a map using ggplot”



You can apply the same code logic you used to generate the single `inset_map`, before passing the results to `grid.arrange()`:

1



```
plot.list <- list(p1, p2, p3, p4)

plot.list %>%

  # add inset map to each plot in the list
  lapply(function(p) ggdraw() +
    draw_plot(p, 0, 0, 1, 1) +
    draw_plot(p.shp, 0.5, 0.52, 0.5, 0.4)) %>%

  # convert each plot in the list to grob
  lapply(ggplotGrob) %>%

  # arrange in grid, as before
  grid.arrange(grobs = ., ncol = 2)
```



Important research questions

How past climate changes mediated genetic diversity levels?

How future climate will structure the distribution of marine biodiversity?

What is the potential effect of wave disturbance in the global distribution of seagrasses?

lead to,

Important technical questions

How can I work with data too big for Excel?

How can I subset big data files (e.g., time periods or other attributes)?

How can I visualize my data?

How can I model the effect of ecological drivers on my response variable?



Innovative research questions have no pre-made packages or software environments

Recommendations

Get to your own scientific questions sooner;

Learn to code [in R with RStudio];

Code in every new project;

Use version control [git with GitHub].

Learn to program in an intentional way

~~in a panic~~ feeling empowered

~~for a single purpose~~ thinking ahead

~~in isolation~~ with a community



Why learn it with R and RStudio

R is free! (“Free as in free speech”);

Optimized for research (self-documenting, repeatable);

- Easier the next time
- Numerous Excel horror stories of scientific studies (TED Talk)

Scalable from small to large problems;

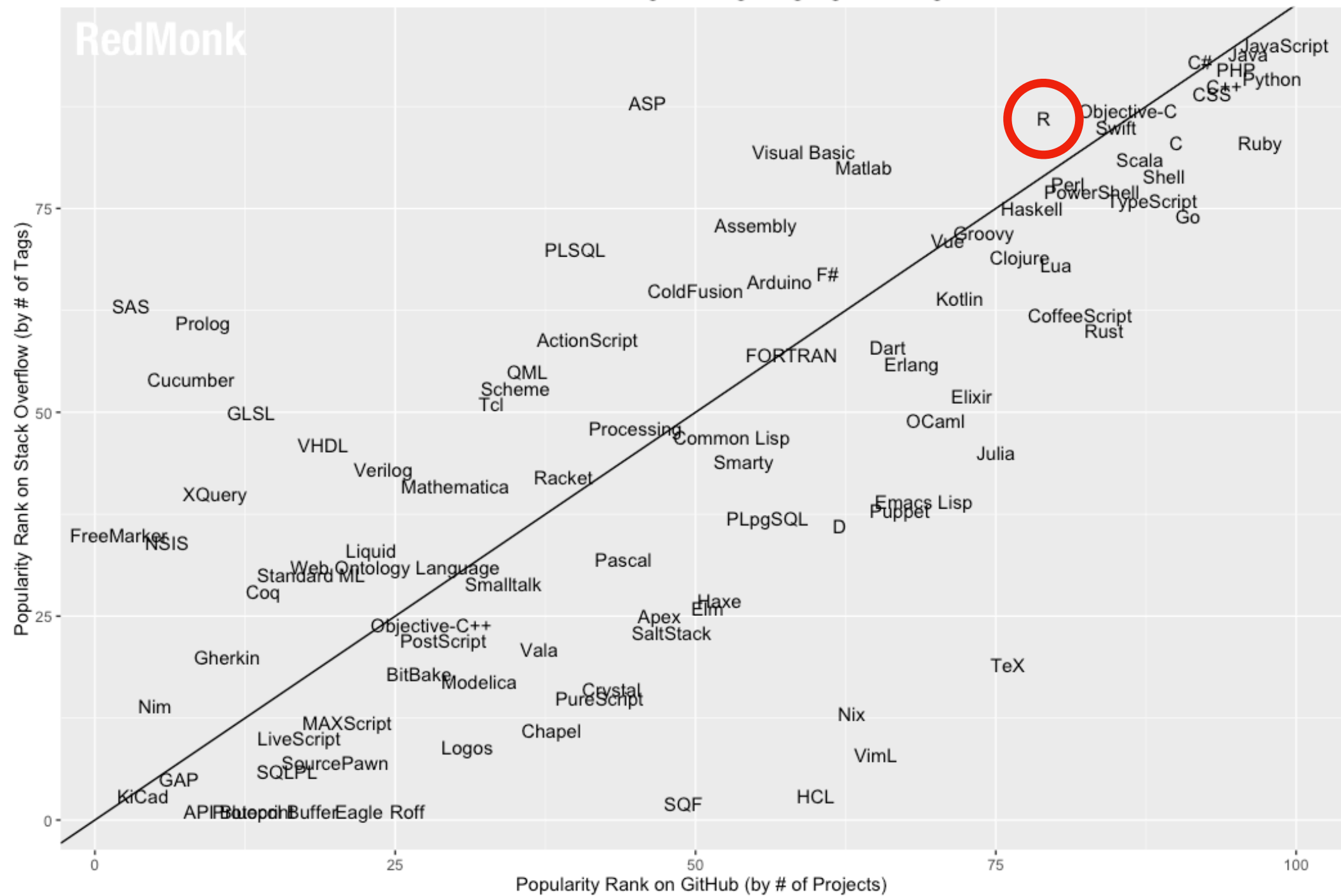
Many learning resources and communities;

- Stack Overflow
- Multiple R books (Free online)

R is ‘becoming’ one of the new norms (paradigm shift).



RedMonk Q318 Programming Language Rankings





Advantages of writing R scripts

Reproducibility and transparency :: Not only the results, but all steps of the analysis are made available.

Flexibility :: Some analyses need only a few code tweaks, from pre-existing scripts (e.g., Github).

Exchange :: In theory other R users can understand your script.

