Marine Ecological Modelling
Global Climate Change

Marine Data Science

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

Data science

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

(Grolemund & Wickham 2016).

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)

Data science

Biology of species

Statistics

Programming

Working with data

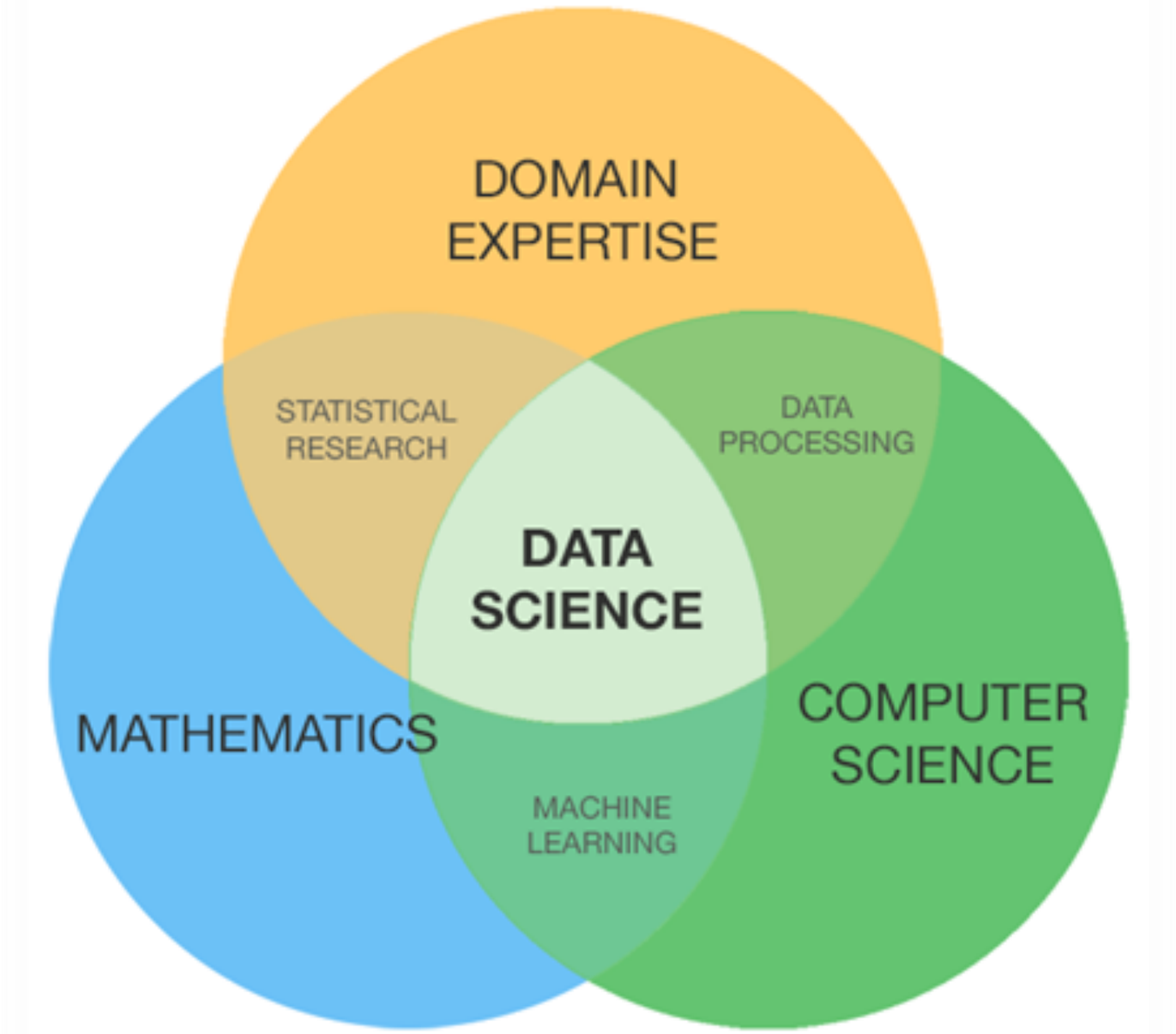
Descriptive statistics

Data visualization

Statistical modeling

Handling big data

Machine learning



Open science tools

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

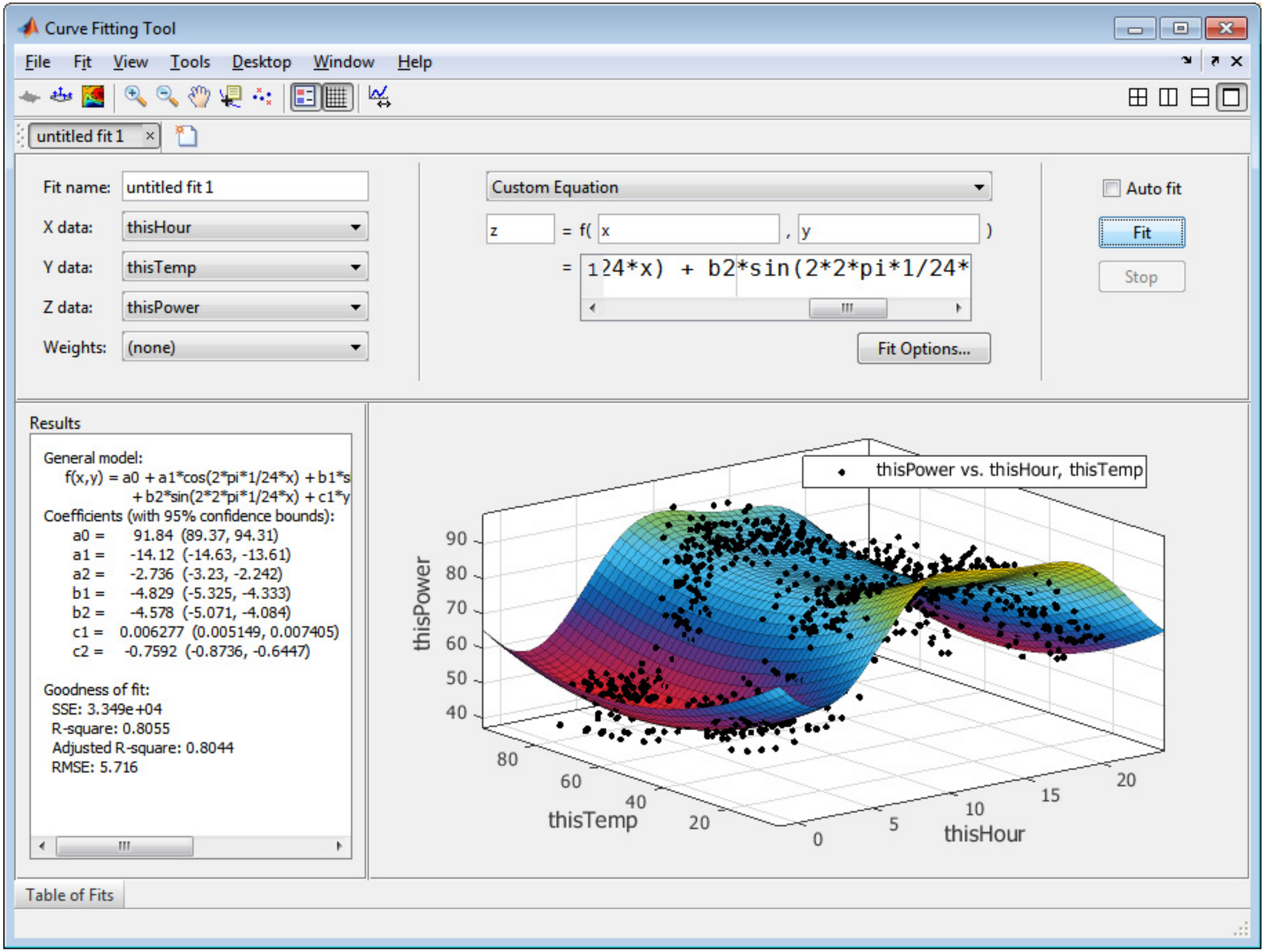
My programming origin story

Being curious and ask how things were developed.

Changing stuff and see what happens.

Trial and error, error, error, error, …

1983


Curve Fitting Tool
File Fit View Tools Desktop Window Help
untitled fit 1
Fit name: untitled fit 1
X data: thisHour
Y data: thisTemp
Z data: thisPower
Weights: (none)
Custom Equation
z = f(x , y)
= 1?24*x) + b2*sin(2*2*pi*1/24*
Fit Options...
Auto fit
Fit
Stop
Results
General model:
f(x,y) = a0 + a1*cos(2*pi*1/24*x) + b1*s
+ b2*sin(2*2*pi*1/24*x) + c1*y
Coefficients (with 95% confidence bounds):
a0 = 91.84 (89.37, 94.31)
a1 = -14.12 (-14.63, -13.61)
a2 = -2.736 (-3.23, -2.242)
b1 = -4.829 (-5.325, -4.333)
b2 = -4.578 (-5.071, -4.084)
c1 = 0.006277 (0.005149, 0.007405)
c2 = -0.7592 (-0.8736, -0.6447)
Goodness of fit:
SSE: 3.349e+04
R-square: 0.8055
Adjusted R-square: 0.8044
RMSE: 5.716
thisPower vs. thisHour, thisTemp
thisPower
90
80
70
60
50
40
thisTemp
80
60
40
20
thisHour
0
5
10
15
20
Table of Fits

2011

Errors are friendly

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

“Plotting a map using ggplot”

1

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

```
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?

Important technical questions

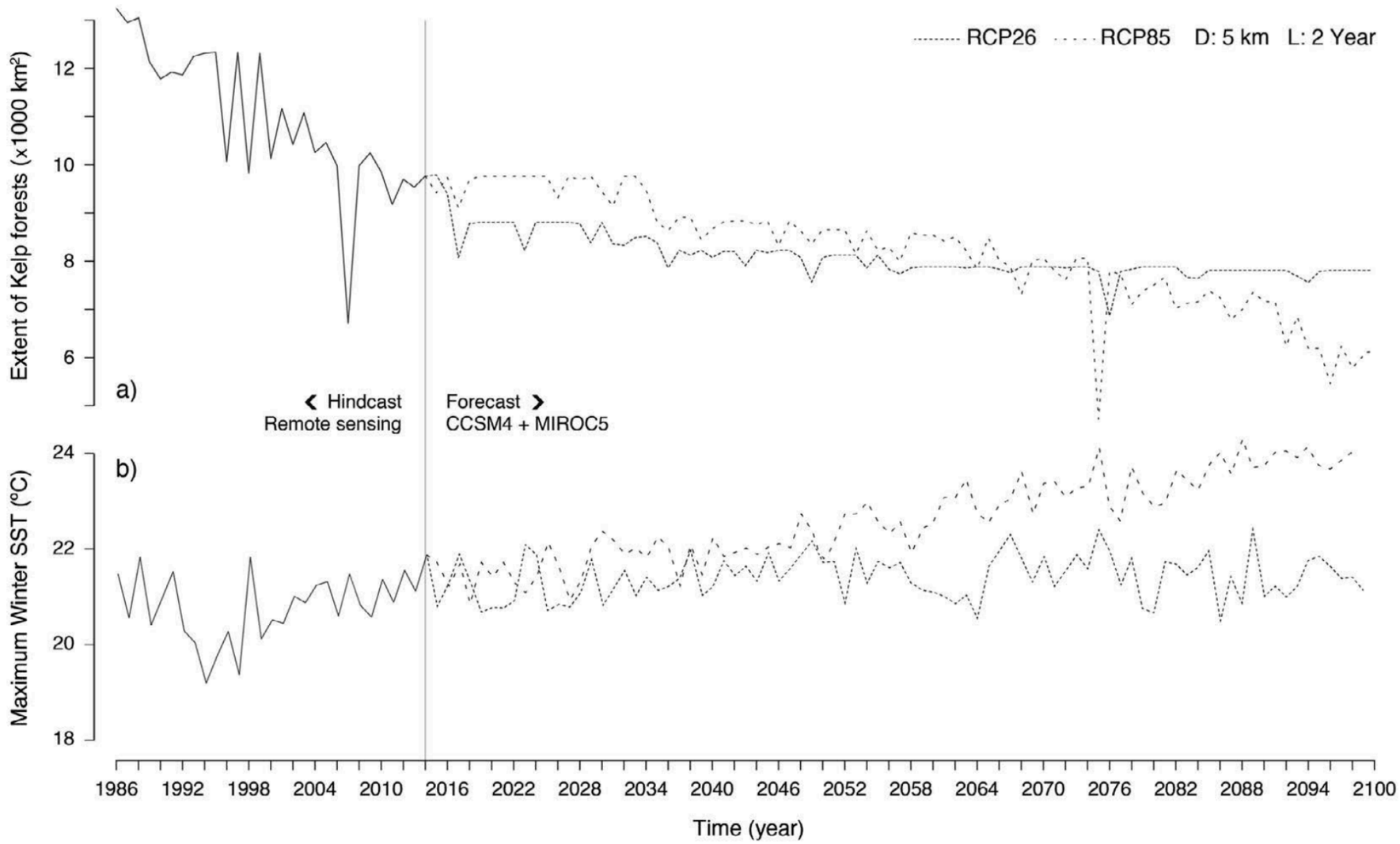
How do I work with data too big for Excel?

How do I subset years or other attributes?

How do I visualize temporal data?

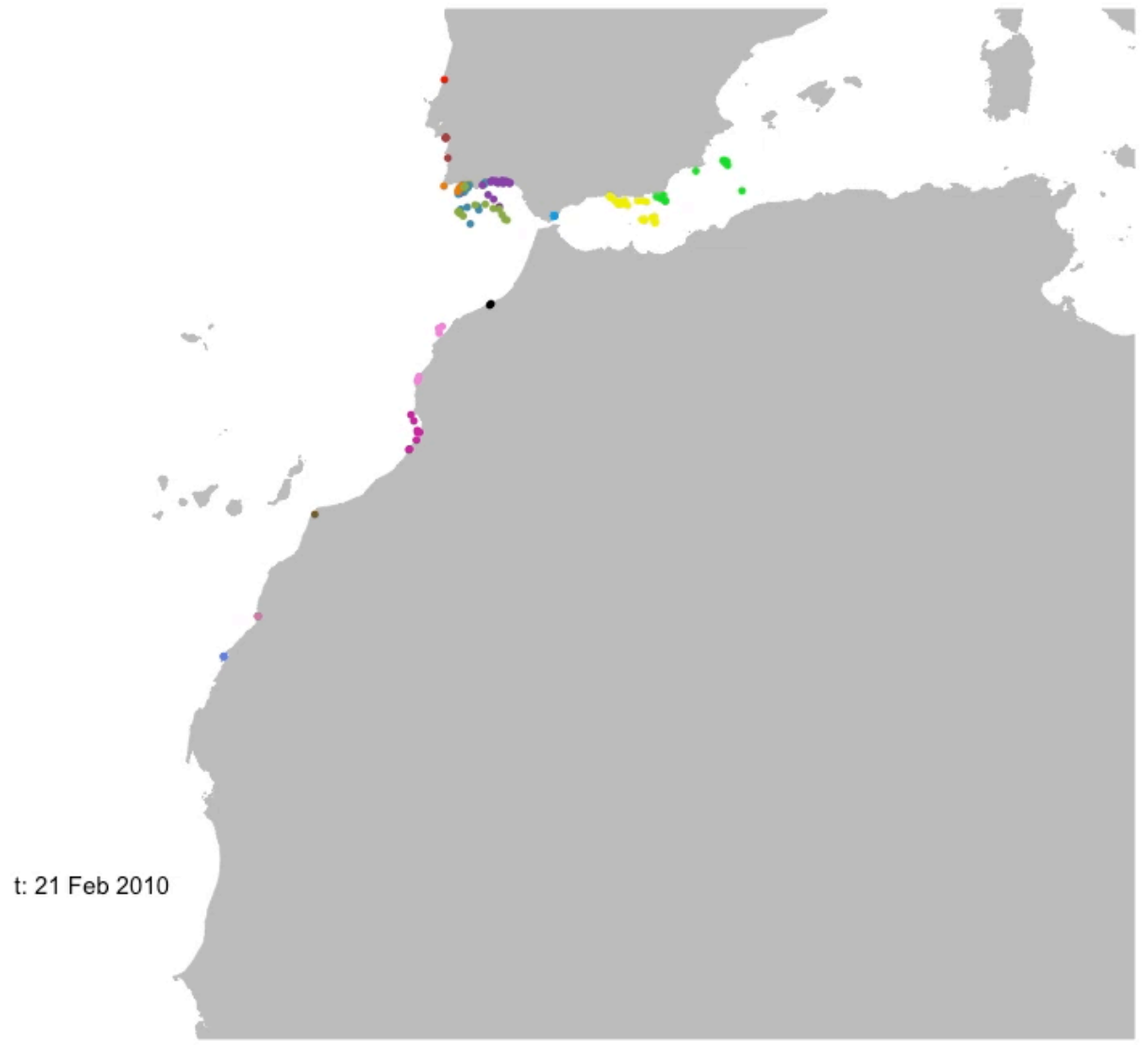
How on earth will I accomplish the task of my project?

Innovative research questions have no pre-made package



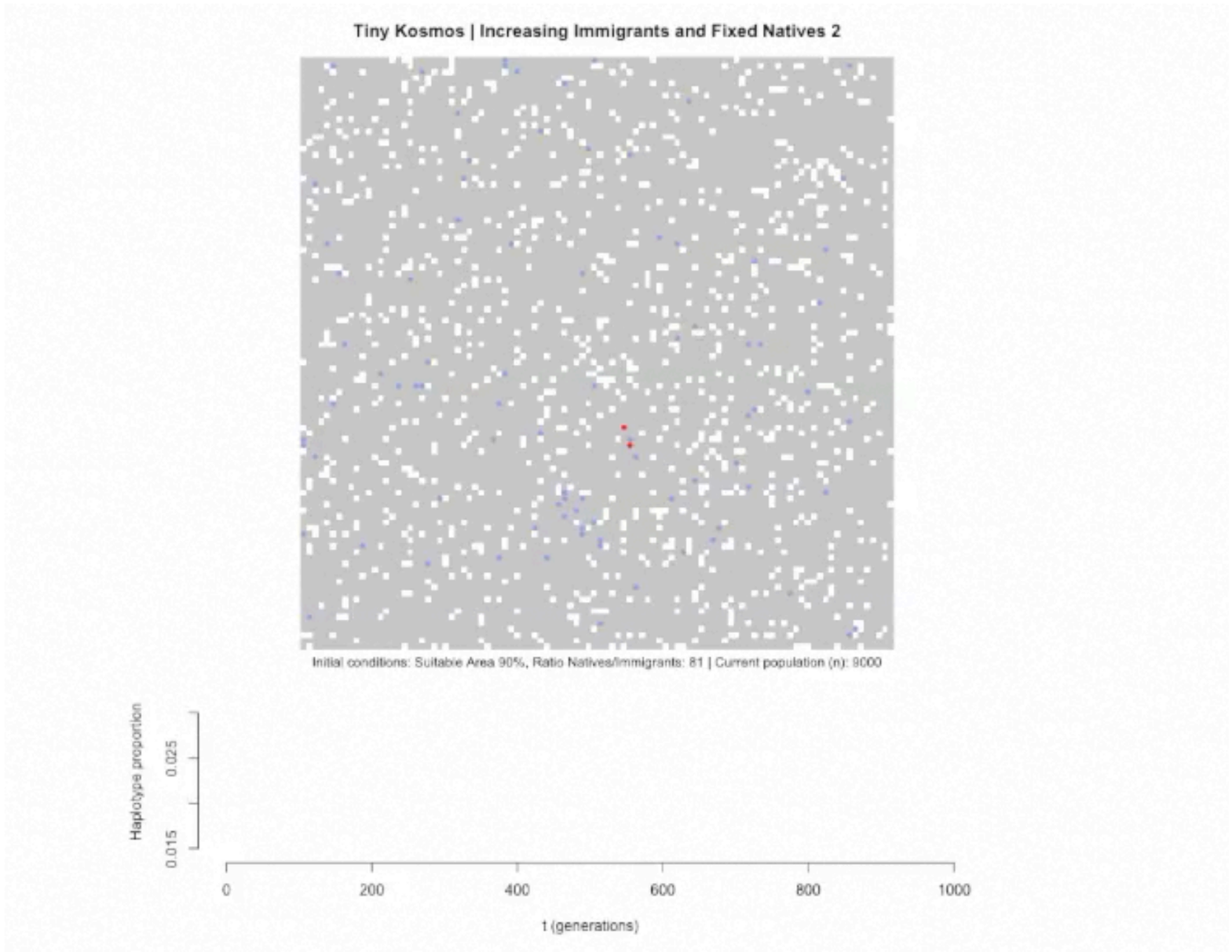
Modelling the coastal extent of kelp forests.

[different climate scenarios producing shifts in distributions]



Generating a virtual ocean with drifting propagules.

[ocean currents mediating population connectivity]



Playing with immigrants and reproduction in a tiny cosmos.

[the winner takes it all]

Reproducibility and optimization

"**Scientists**, according to interviews and expert estimates, **spend from 50 percent to 80 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

Full coded process;
R with documentation;
RMarkdown.

Collaboration and communication

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

Before

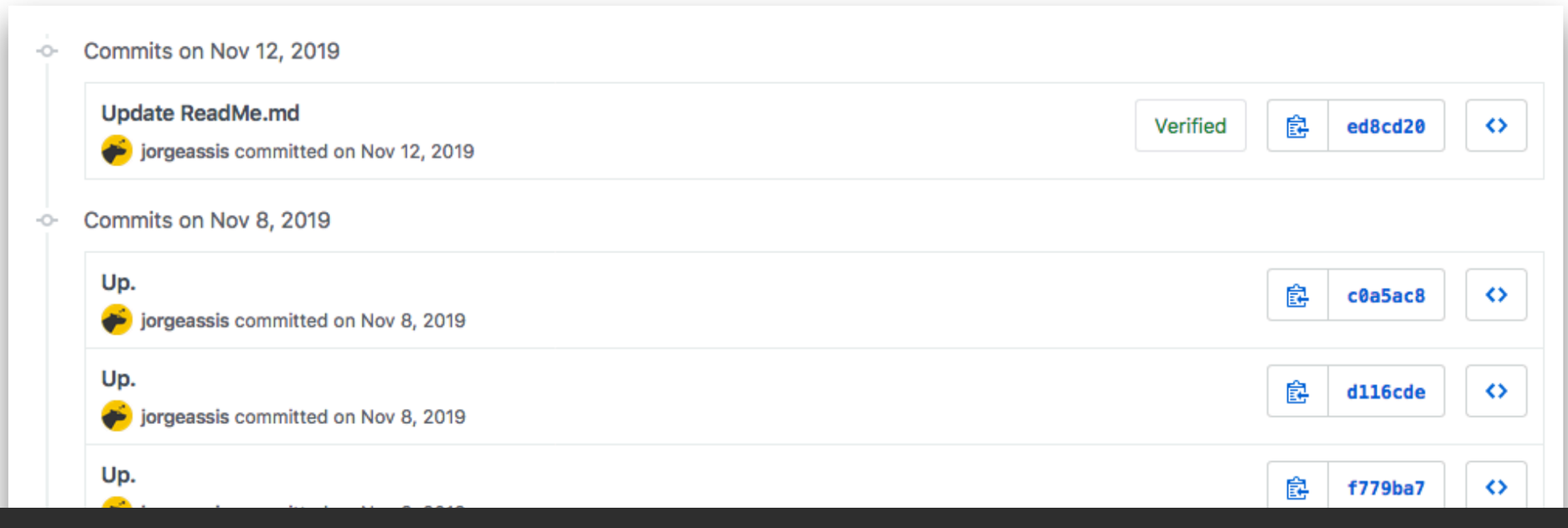
Filenames suffixed with dates, initials (e.g., final_JL-2016-08-05.csv);

Email chains (often forwarded).

Now

Version control with git;

Short messages accompany committed changes.



Sharing data

Before

Published manuscripts;

Data on FTP server and supplementary information.

Now

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-

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 R with RStudio

R is free! (“Free as in free speech, not free beer”);

Optimized for research (self-documenting, repeatable);

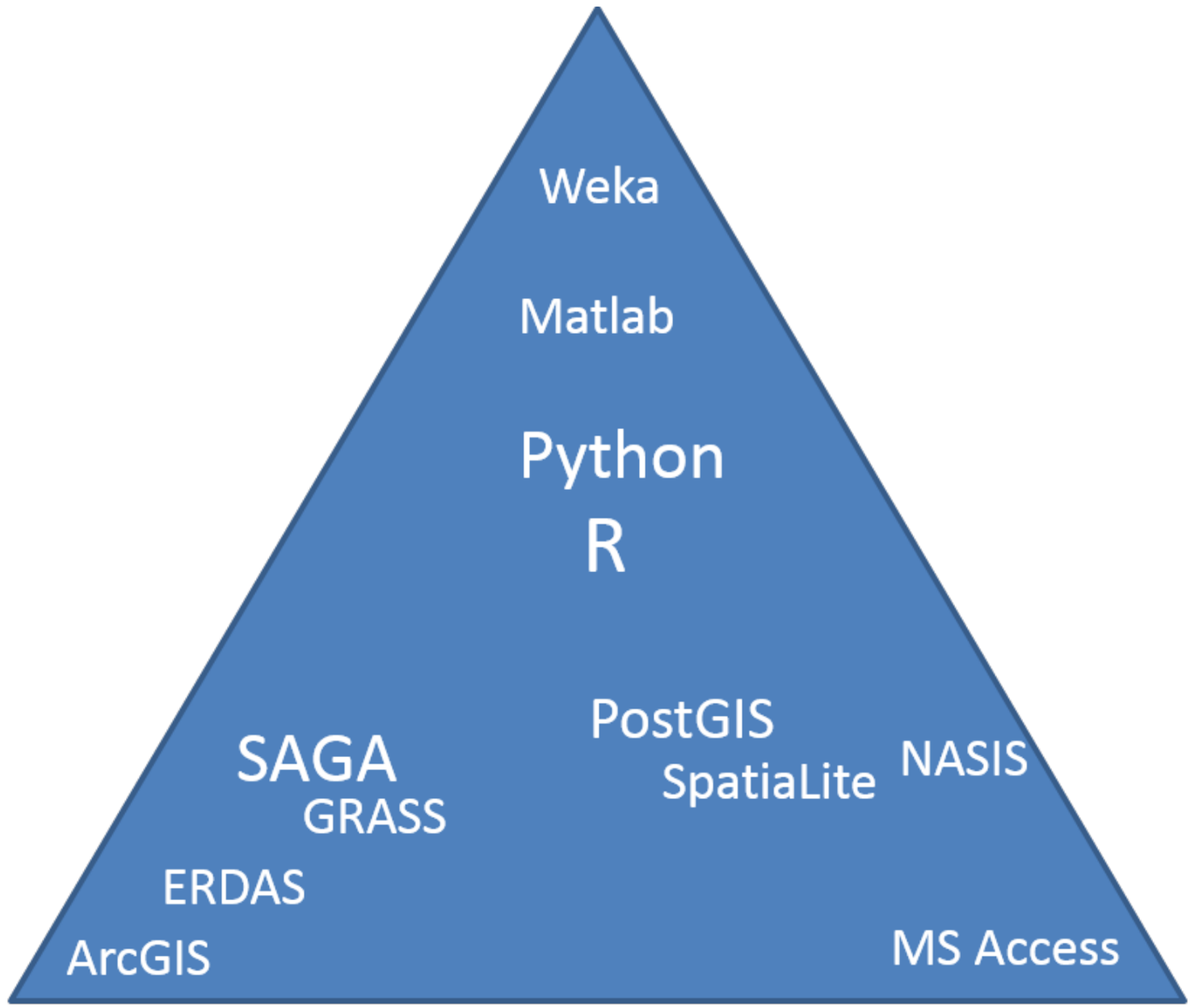
- Easier the next time
- Numerous Excel horror stories of scientific studies (TED Talk)

Scalable from small or large problems;

Many learning resources and communities;

- Stack Overflow
- R books (Free Online)

R is ‘becoming’ the new norm (paradigm shift?).


Statistics
Weka
Matlab
Python
R
PostGIS
SpatiaLite
SAGA
GRASS
ERDAS
ArcGIS
NASIS
MS Access
Spatial analyses
Database management