R course
–
Getting started

12. – 16.09.2016

Max Planck Institute for Marine Microbiology

Christiane Hassenrück

Course overview

Monday:

- General background
- Getting familiar with R

Tuesday:

- Data handling
- Plotting

Wednesday:

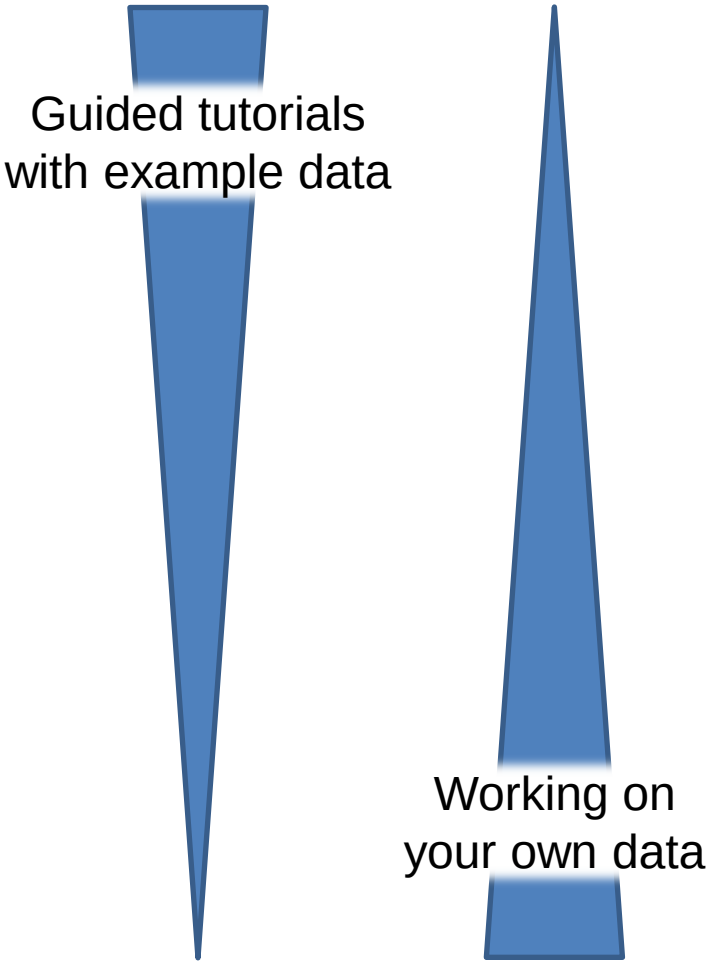
- Experimental design
- Statistics

Thursday

- Writing R functions

Friday

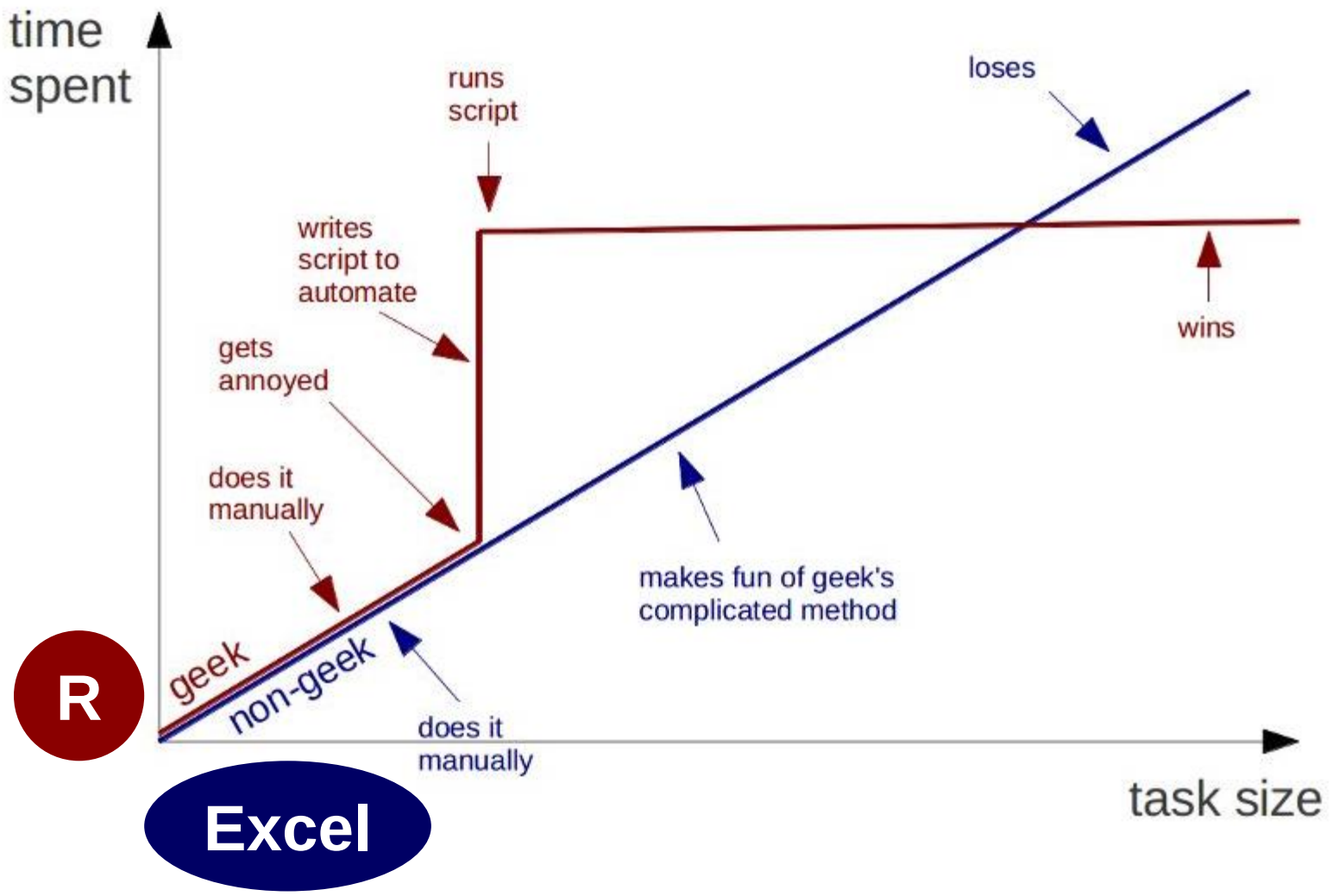
- TBD



"R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS." (https://www.r-project.org/)

Pro's	**Con's**
Versatile	Command-line
Platform-independent	Needs some getting used to…
Data exploration and visualization	Won't always tell you what to do…
Hypothesis testing	
Advanced graphics	
Large data sets	
Reproducible	
Open source	
Good documentation and online support	
… many more	

We are going to do something about that!



http://www.howtogeek.com/102420/geeks-versus-non-geeks-when-doing-repetitive-tasks-funny-chart/

R resources

R console:

Download R and R packages: **CRAN Mirrors**
(https://cran.r-project.org)

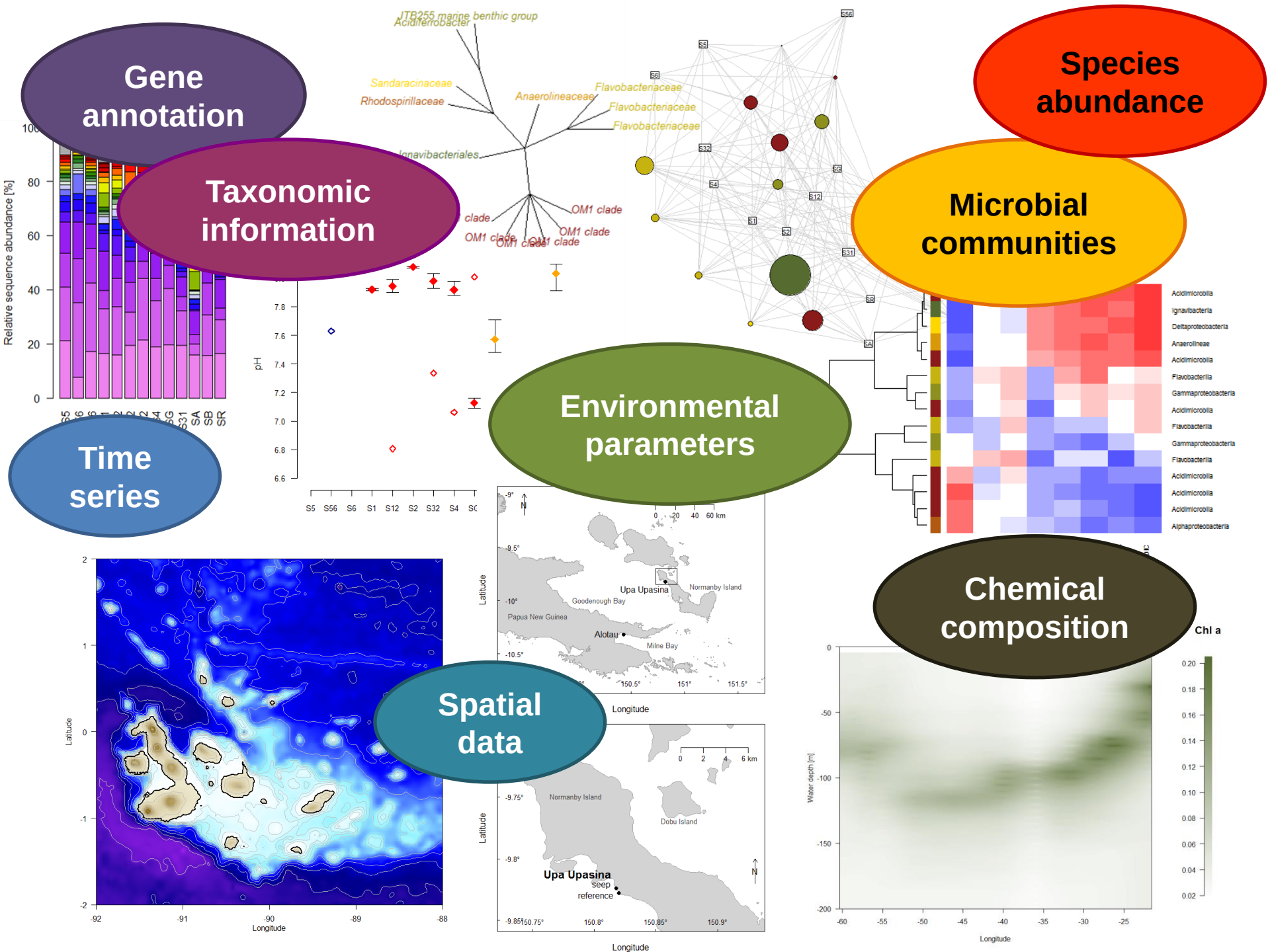
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
(https://www.bioconductor.org)

GitHub
(https://github.com)

More user-friendly implementation:

RStudio®

Tinn-R
Editor - GUI for R Language and Environment


Gene annotation
Taxonomic information
Species abundance
Microbial communities
Environmental parameters
Time series
Spatial data
Chemical composition
Relative sequence abundance [%]
pH
Acidiferrobacter
ITB255 marine benthic group
Sandaracinaceae
Rhodospirillaceae
Anaerolineaceae
Flavobacteriaceae
Ignavibacteriales
OM1 clade
Acidimicrobiia
Ignavibacteria
Deltaproteobacteria
Anaerolineae
Flavobacteriia
Gammaproteobacteria
Alphaproteobacteria
Upa Upasina
Normanby Island
Goodenough Bay
Papua New Guinea
Alotau
Milne Bay
Dobu Island
seep
reference
Latitude
Longitude
Water depth [m]
Chl a

Input formats

- Most common input format for tabular data:
  - .txt
  - .csv
  - .tsv

- Include variable names in first row (header)
- Values usually tab, space, or comma separated
- Avoid special characters and spaces in data values, variable names, and file names

	Bad	**Good**
Variable name	mean temperature mean-temperature mean_temperature	temperature.mean
Data value	day 1	day1 1 (variable name: day)

Input formats

- **Bad:**

reef	site	seep.influence	pH		
Illi	S1	medium	7.92	7.93	7.91
Illi	S12	medium	7.94	7.9	7.99
reef	site	seep.influence	SiO4		
Illi	S1	medium	4.47	4.245	4.956
Illi	S12	medium	2.08	2.15	1.836
reef	site	seep.influence	PO4		
Illi	S1	medium	0.11	0.107	0.107
Illi	S12	medium	0.09	0.083	0.093

- **Good:**

reef	site	seep.influence	pH	SiO4	PO4
Illi	S1	medium	7.92	4.471	0.109
Illi	S1	medium	7.93	4.245	0.107
Illi	S1	medium	7.91	4.956	0.107
Illi	S12	medium	7.94	2.076	0.09
Illi	S12	medium	7.9	2.15	0.083
Illi	S12	medium	7.99	1.836	0.093

Google's R Style Guide

information on file formatting and generating clean R code:
https://google.github.io/styleguide/Rguide.xml

Errors in R

Syntax errors

- When R doesn't understand you, because the command doesn't make sense…
- R returns an error message
- E.g.: Trying to calculate the mean of categorical data

Semantic errors

- When R doesn't do what you want, although the command makes sense…
- R will not return an error message, because the command is valid
- More dangerous errors
- E.g.: Calculating percentages over columns, and not rows

Google is your new best friend ☺

Example data sets

https://github.com/chassenr/

- Data set 1: CO_2 vents
  - Carbonate chemistry and nutrient concentrations from the bottom water
  - Microbial community composition in the sediment (sequence counts and taxonomic information)
  - Sediment microprofiles
  - Coastlines of Papua New Guinea
- Data set 2: Bathymetry of the Galapagos Islands
- Data set 3: Hydrochemistry across the Tropical North Atlantic
- Your data!