Basic Statistics with R

Alex Sanchez, Miriam Mota, Ricardo Gonzalo and Santiago Perez-Hoyos

Statistics and Bioinformatics Unit.
Vall d'Hebron Institut de Recerca

Readme

- License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License http://creativecommons.org/licenses/by-nc-sa/4.0/
- You are free to:
  - **Share** : copy and redistribute the material
  - **Adapt** : rebuild and transform the material
- Under the following conditions:
  - **Attribution** : You must give appropriate credit, provide a link to the license, and indicate if changes were made.
  - **NonCommercial** : You may not use this work for commercial purposes.
  - **Share Alike** : If you remix, transform, or build upon this work, you must distribute your contributions under the same license to this one.

Section 1

Introduction

Outline

- Introduction
  - Who are we ("we"=teachers & students)
  - Why are we here (Why learn R?)
- How will we proceed: Methodology
- HW Data Science approach to using R
- References & Resources

Who are we (1): The Statistics and Bioinformatics Unit

www.ueb.vhir.org

Welcome to VHIR's Statistics and Bioinformatics Unit

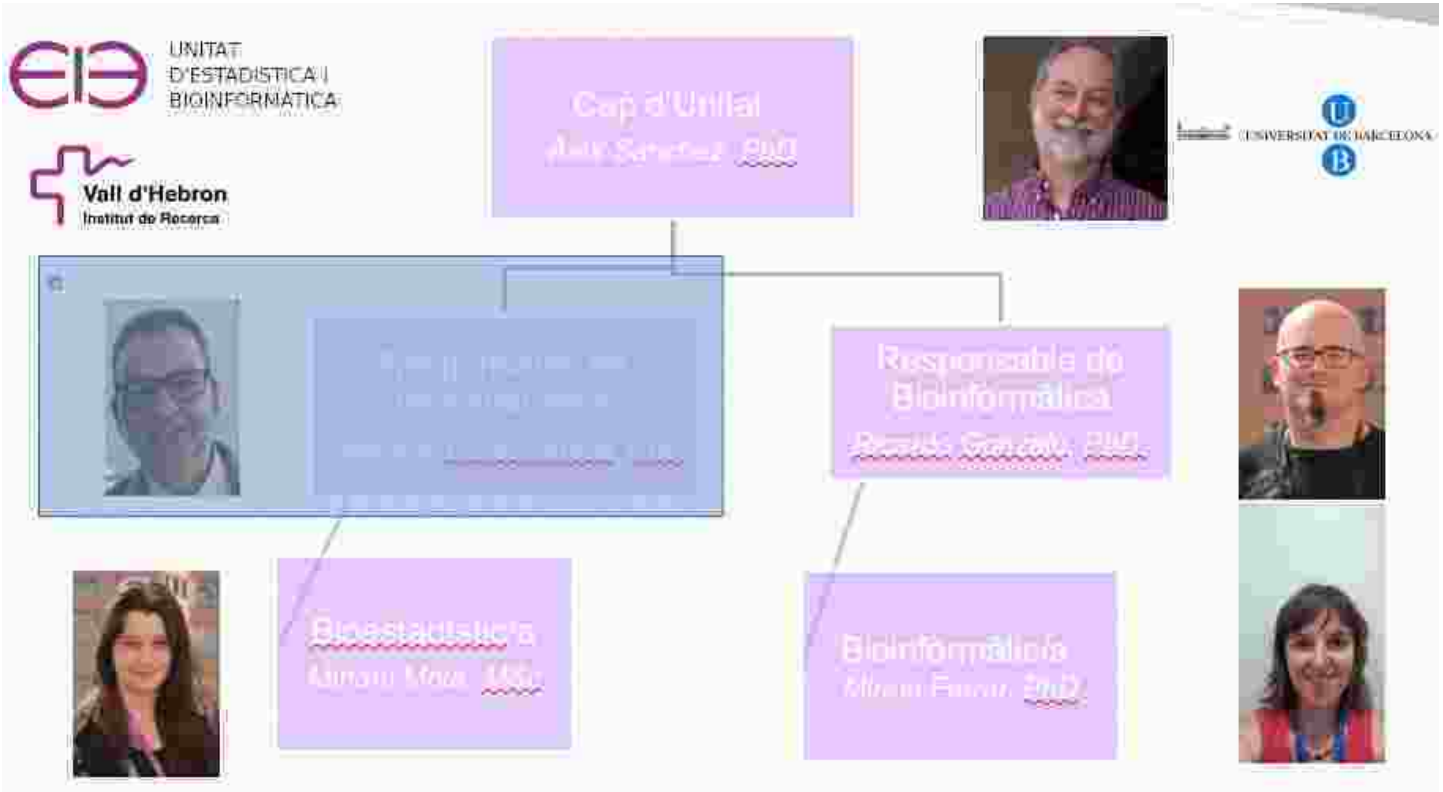
Who we are

The Statistics and Bioinformatics Unit (UEB-USMIB) is a service unit from the Scientific Support Area of the Vall d'Hebron Research Institute (VHIR - www.vhir.org)

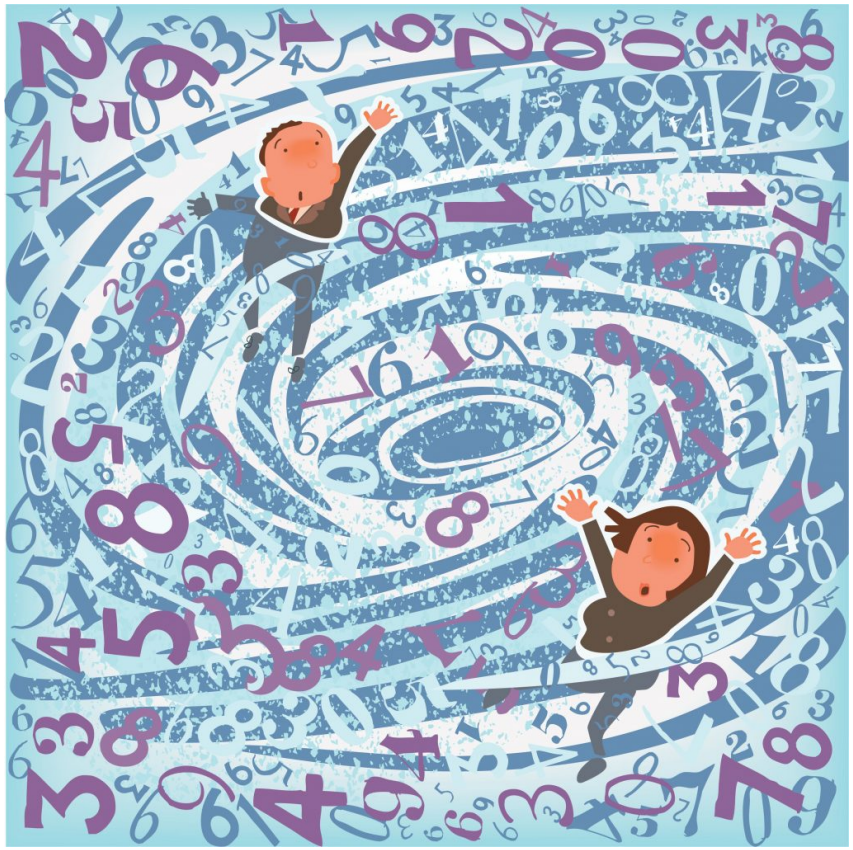
The UEB was created in 2008 within the Research Institute of the Hospital Vall d'Hebron in order to promote the use and development of modern statistical and bioinformatics resources on research performed in its environment.

Nowadays, the Statistics and Bioinformatics Unit includes the former Support Unit in Methodology for Biomedical Research (USMIB) and, as part of the Scientific and Technical Support Area of the Vall d'Hebron Research Institute, has the mission to provide expert advice, services and training for clinical and biomedical research.

Who are we (2): Teachers



Why this course (1)



"We are drowning in information but starving for knowledge"

Why this course (2)

- (Biomedical) research, as well as many other human activities (social networks, sports, COVID . . .) generate huge quantities of -often complex- data.
  - Although sometimes we will also have small datasets
- We believe that data leads to information that leads to knowledge, but we need to be able to extract one from the other.
- This can be attempted in many ways, artificial intelligence, machine learning, data science or something which is common to all of them: **plain statistics**!!!

What are our goals (1)

- The main goal of this course is to introduce a variety of statistical methods and tools, which is good enough to:
  - Help you analyze your own data when it makes sense
  - Suggests you when the analysis is complex enough to contact an expert statistician (such us those in the UEB)
  - Help you to distiguish one from the other
- A secondary, but not least important objective: Show how to do it using R

Why learn R

- Most people in most jobs have to *manage* information in their every day work.
- "Managing" may mean different things such as:
  - *retrieving*
  - *manipulating*
  - *visualizing*
  - *analyzing*
  - *reporting*
- R is a powerful tool that can be used to facilitate, improve or automate tasks such as those described above.

Why doing statistics with R

- R has become a "de facto" standard for statistical analysis
- Practically all existing statistical methods available
- Powerful graphics that can be used interactively to explore data
- Possibility of scripting analysis $\longrightarrow$ **Reproducibility**

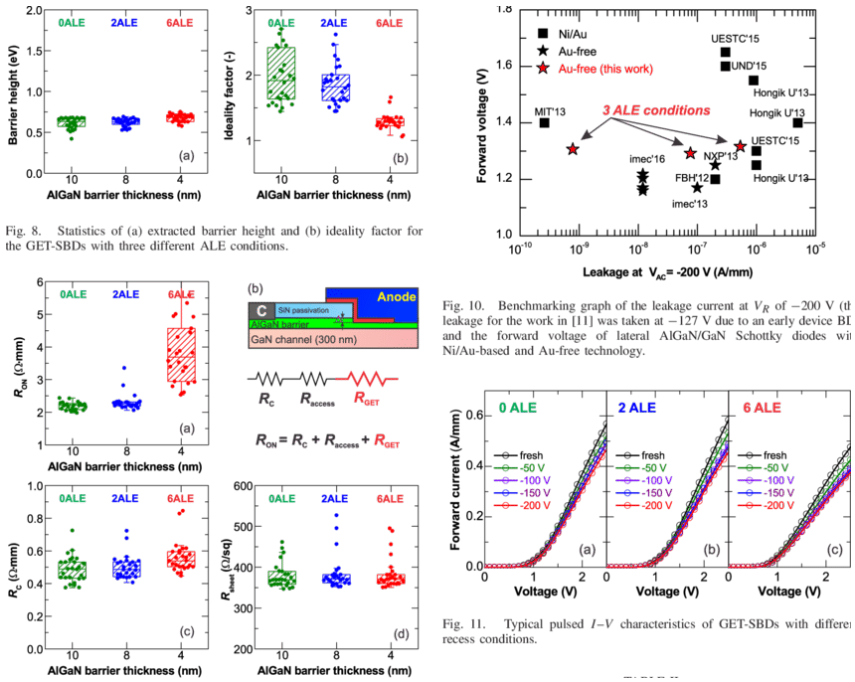
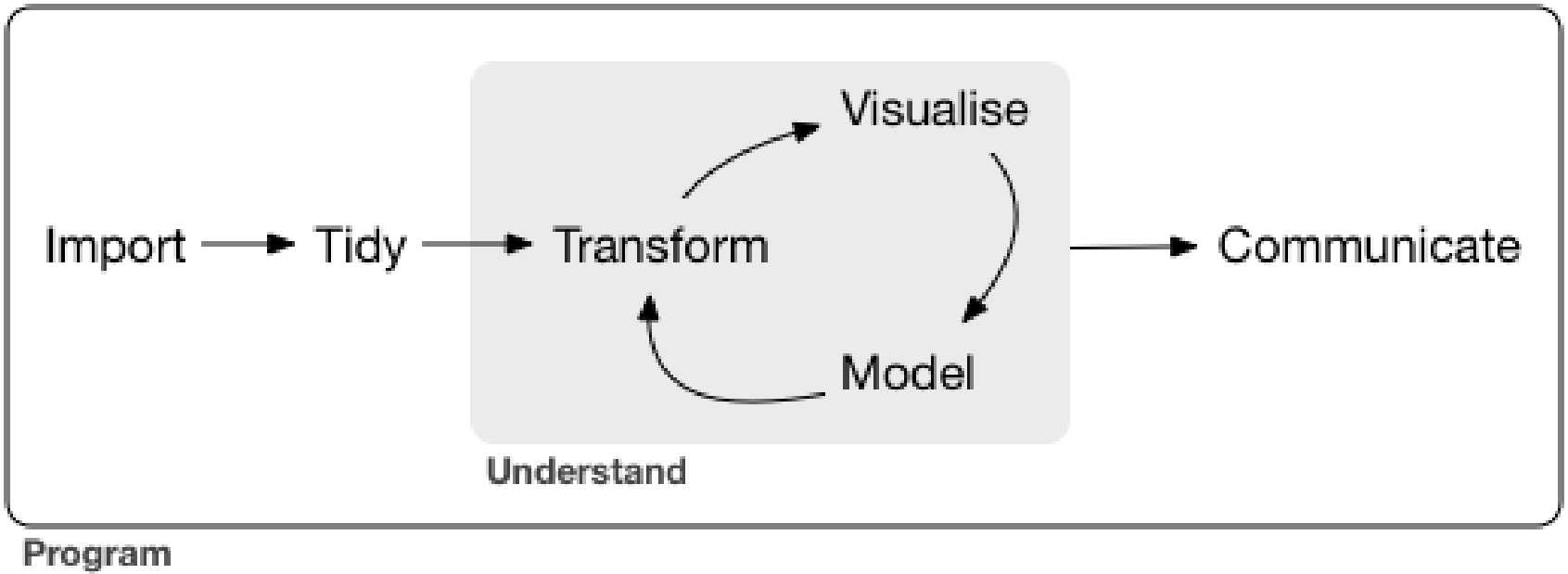


Figure 1: R for statistical Analysis and Graphics

Hadley Wickam's approach to learning and applying Data Sciendce



Your turn

- Provide examples of informations you may wish to manage
- Describe briefly
  - what this information is about
  - how it is stored
  - what you may wish to do with it
    - Transformations
    - Computations
    - Reports

How we will work

- Mastering R requires as many other disciplines
  1. Time
  2. Study, and
  3. Practice.
- Our lectures will have the following structure (all but the first)
  - 1st part: Discuss the work you have done during the week
  - 2nd part: We introduce a few new ideas
  - 3rd part: Practice exercises and start working on the case study suggested/your data.

Resources and references

- There is a huge variety of resources to learn R, books, tutorials, free online courses, etc.
- This course is based on the book Data Science for R.
- Other interesting books
  - Using R and RStudio for Data Management, Statistical Analysis, and Graphics, 2nd edition
- Online courses
  - Coursera's Data Science Specialization
- A list of R tutorials and courses