UNIVERSITY OF CALGARY

BioSci UofC

R Wizardry

Instructor: **Oscar Montoya**
(oemontoy@ucalgary.ca)

TA: **Ryan Tate**
(ryan.tate@ucalgary.ca)

Introduction to programming in R and R-Studio

Disclaim:

This is NOT a statistics course!

Learning to code

- Coding can be fun
  - Treat it as a game
  - Many ways to solve the same problem
  - R can be run line-by-line; play around!

- Coding is an investment
  - May be slow the first time but easy to repeat and modify
  - Add unique set of skills to your C.V.
  - Fastest growing job market

R: programming language and software

- R language is a dialect of S / S+
- Widely used for advanced statistic, plotting and data analysis
- Hundreds of books and (many are open source!) with examples and available data sets
- Several free courses (edx, coursera, Code School, Code Academy, DataCamp, among others)

- Other help resources for the R novice: video tutorials, The R Project (CRAN), Stack Overflow, R-Bloggers, and many more

- Another powerful reason to use R:

It's opensource

- Functions outside base R are organized as “packages” in a “library”, where each package is focused on handling a relatively narrow set of functions

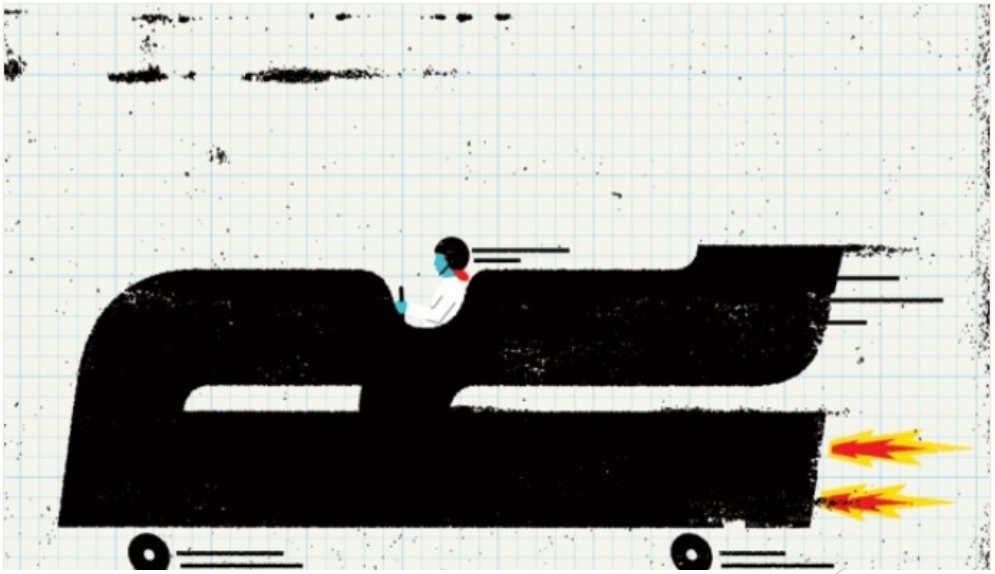
NATURE | TOOLBOX

Programming tools: Adventures with R

A guide to the popular, free statistics and visualization software that gives scientists control of their own data analysis.

Sylvia Tippmann

29 December 2014 | Clarified: 13 February 2015



R or Spreadsheet software? Actually, both!

When to use Excel-like software:

- When you have quick and dirty number crunching to do:
  - Small handful of descriptive stats
  - You need to look something up
  - Run a quick sort/filter
  - Quickly visualize some data (scatter plot, bar plots)

- Inputing long datasets is easier in Excel than in R.

R or Spreadsheet software? Actually, both!

Use R when you need:

- To explore data
- Serious statistical capabilities and reproducibility in the long run
- Have big datasets that require serious manipulation (R is more powerful and faster)
- State-of-the-art graphics

“If you are using R and you think you're in hell, this is a map for you” (Patrick Burns, The R Inferno) Codes can be recycled (“the best bioinformatician is the laziest one”)

Recommended articles:

http://www.r-bloggers.com/why-you-should-learn-r-first-for-data-science/

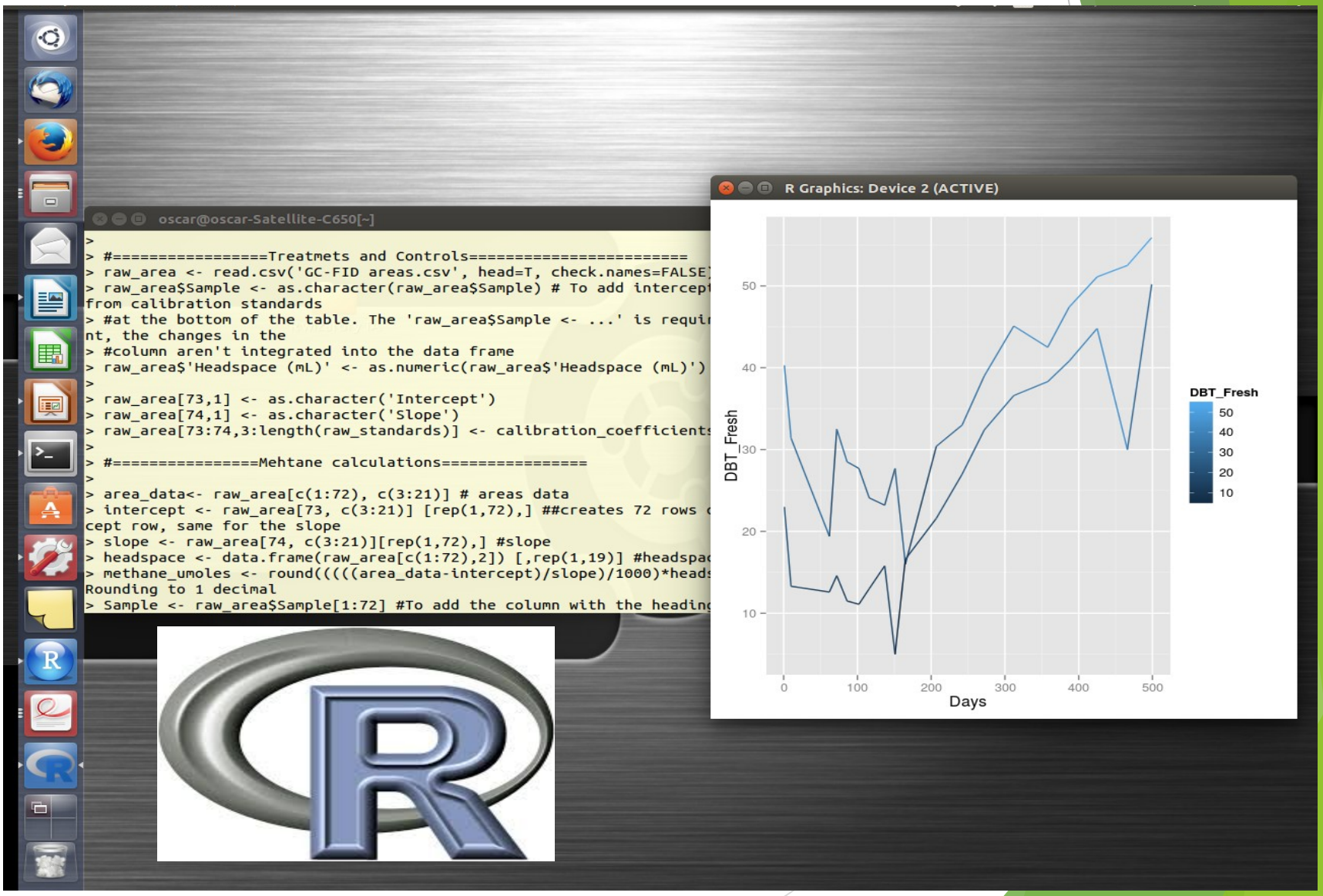
http://www.nature.com/news/programming-tools-adventures-with-r-1.16609

The R User Interface

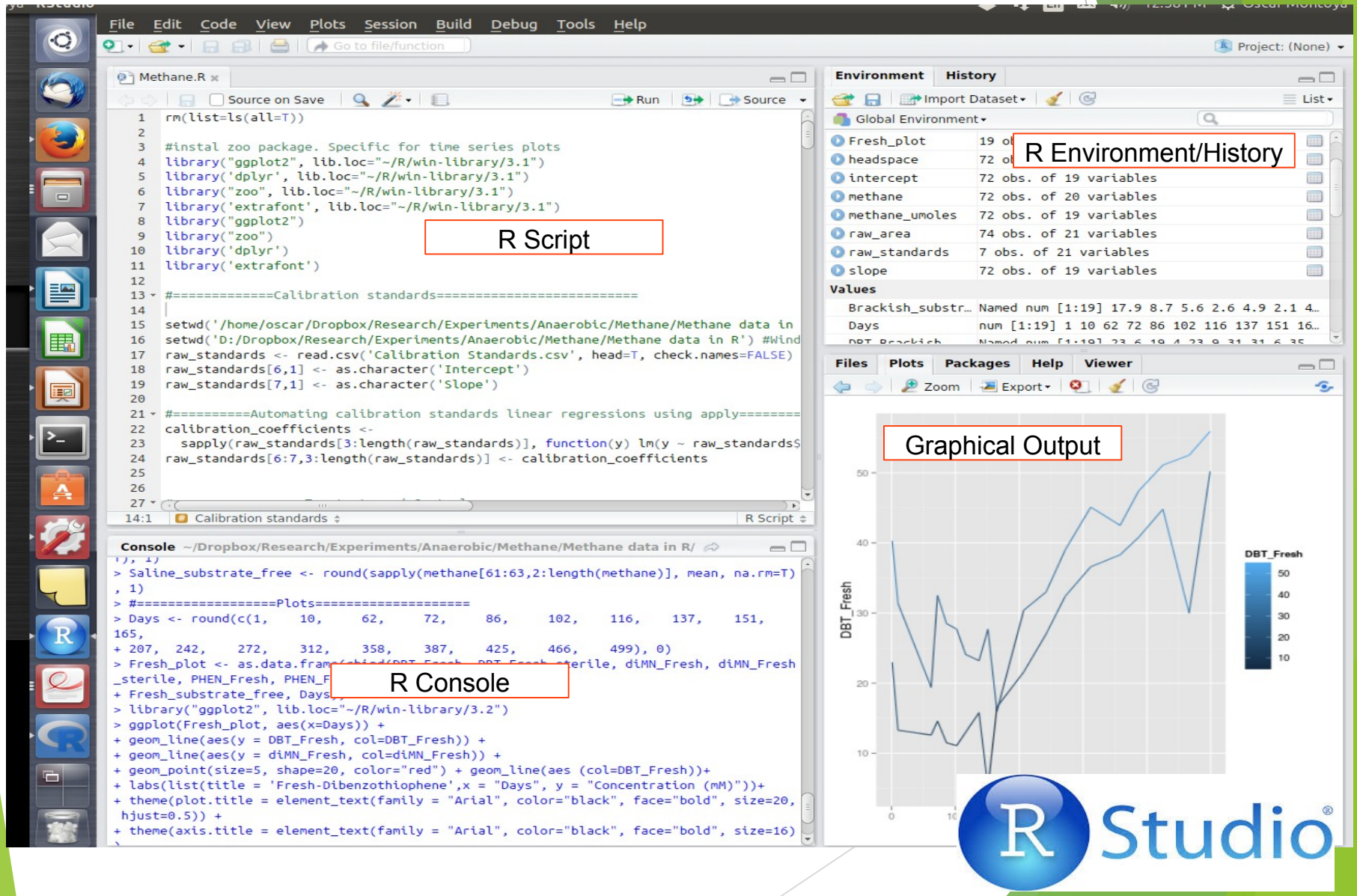
“Before you can ask your computer to save some numbers, you’ll need to know how to talk to it. That’s where R and RStudio come in. RStudio gives you a way to talk to your computer. R gives you a language to speak in.”

Although R can be ran from a terminal, integrated development environments like R-Studio, offer a Graphic User Interface (GUI) that simplifies the visualization of the coding.

Using R from terminal



R-Studio (has a GUI)



Type of data in biology

Numeric

- $0 \rightarrow +\infty$, integer (i.e., count data)
- $0 \rightarrow +\infty$, continuous
- $-\infty \rightarrow +\infty$, continuous
- $-\infty \rightarrow +\infty$, integer

Character

- Study site is: “Alberta” v. “British Columbia”
- Genetic data stored as basepairs: “GATTACA” vs. “CTGCCAC”

Factors

Special way to store character data with a ‘numeric’ rank: “Alberta” v. “British Columbia”

Computer interprets “Alberta” as 1, and “British Columbia” as 2

Types of objects in R

Vector: a sequence of at least one stored value(s):

List: form of a vector in which elements need not be of the same type; elements can be vectors, matrices, arrays, or lists themselves

Matrix: 2-d form of a vector that can be indexed by rows and columns; data must be of the same type

Array: N-dimensional form of a vector that can be indexed by rows, columns, depth, etc.

Data Frame: like a matrix, but rows/columns can vary in type; character and numeric data allowed

Data types

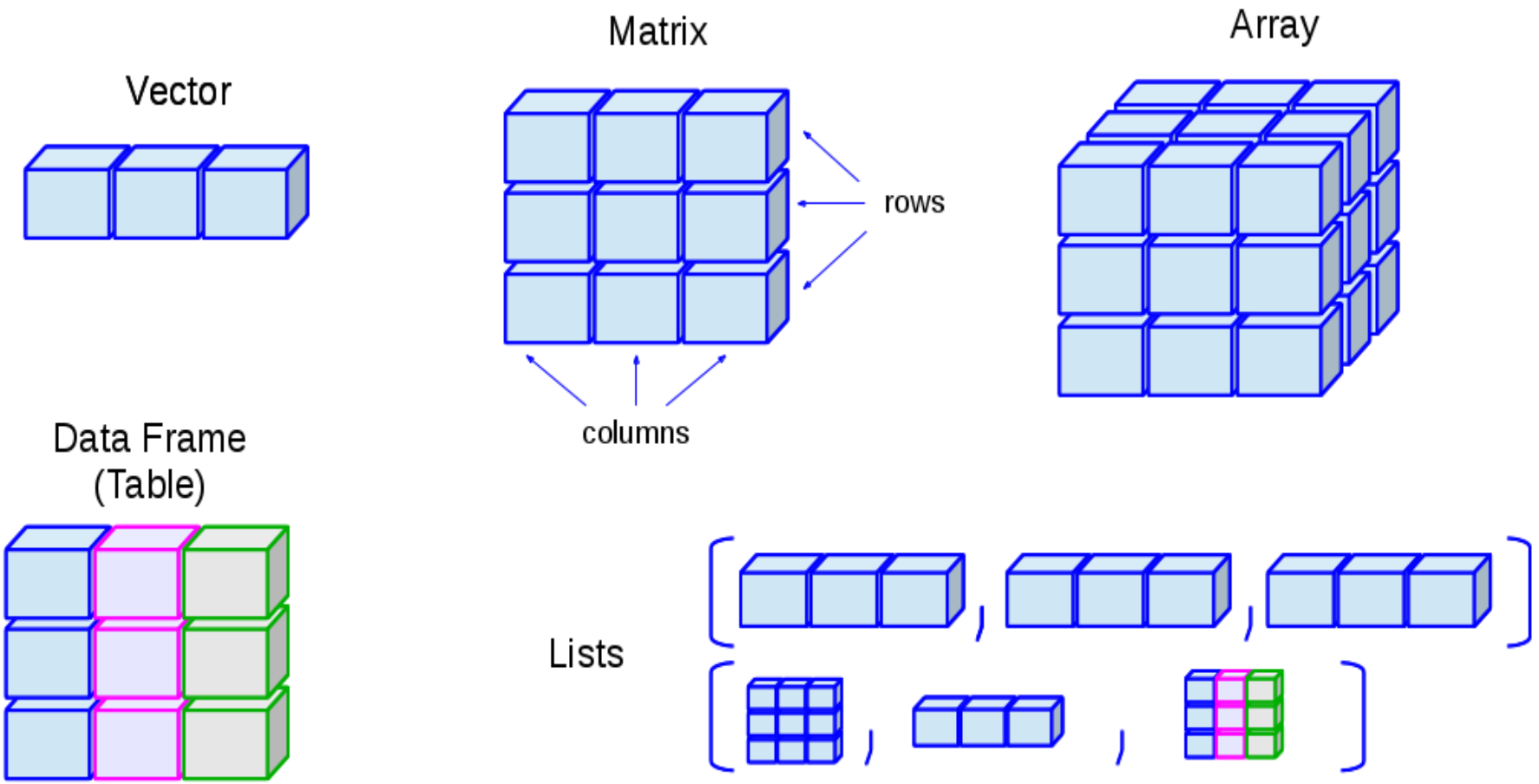
All data is either numeric or character

Hence:

- Methods you learn to deal with numeric data in R will be the same
  - Regardless if physiology, biochemical, genetic, or ecological data

- Methods you learn to deal with character data will be the same across all character data

- Some programming methods vary in efficiency
  - Analysis of 100,000+ bp sequences needs computational efficiency
- Programming languages are similar
  - C++, ADMB, R, S, SAS, etc.



http://venus.ifca.unican.es/Rintro/dataStruct.html

Types of objects