Introduction to open source R

Edmondo Di Giuseppe

Institute of Biometeorology, CNR-National Research Council

MedCOF Training Workshop

Course, 15-18 November 2016

Outline

At the Mall

Product

R is a free software environment for data analysis, graphics and statistical computing

The **R** name is based on the (first) names of the first two developers **R**obert Gentleman and **R**oss Ihaka, playing on the name of the **S** language, from which R is derived.

At 2016:
- 2 millions of users
- 9488 available packages
- 2000 package developers
- 1200 open projects on R-forge

Product

R is a free software environment for data analysis, graphics and statistical computing

The **R** name is based on the (first) names of the first two developers **R**obert Gentleman and **R**oss Ihaka, playing on the name of the **S** language, from which R is derived.

At 2016:
- 2 millions of users
- 9488 available packages
- 2000 package developers
- 1200 open projects on R-forge

A bit of history

- late 70s John Chambers and others at Bell Laboratories develop a statistical programming language named **S**
- late 80s Richard A. Becker, John M. Chambers and Allan R. Wilks publish *The New S Language* in which the system is rewritten in C
- early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland
- 1995 R is available under **General Public License** (GPL)
- 1997 **R core** group of programming developers is founded
- 2002 **R foundation for Statistical computing** is established in Vienna

A bit of history

late 70s John Chambers and others at Bell Laboratories develop a statistical programming language named **S**

late 80s Richard A. Becker, John M. Chambers and Allan R. Wilks publish *The New S Language* in which the system is rewritten in C

early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland

1995 R is available under **General Public License** (GPL)

1997 **R core** group of programming developers is founded

2002 **R foundation for Statistical computing** is established in Vienna

A bit of history

- late 70s John Chambers and others at Bell Laboratories develop a statistical programming language named **S**
- late 80s Richard A. Becker, John M. Chambers and Allan R. Wilks publish *The New S Language* in which the system is rewritten in C
- early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland
- 1995 R is available under **General Public License** (GPL)
- 1997 **R core** group of programming developers is founded
- 2002 **R foundation for Statistical computing** is established in Vienna

A bit of history

- late 70s John Chambers and others at Bell Laboratories develop a statistical programming language named **S**
- late 80s Richard A. Becker, John M. Chambers and Allan R. Wilks publish *The New S Language* in which the system is rewritten in C
- early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland
- 1995 R is available under **General Public License** (GPL)
- 1997 **R core** group of programming developers is founded
- 2002 **R foundation for Statistical computing** is established in Vienna

A bit of history

- late 70s John Chambers and others at Bell Laboratories develop a statistical programming language named **S**
- late 80s Richard A. Becker, John M. Chambers and Allan R. Wilks publish *The New S Language* in which the system is rewritten in C
- early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland
- 1995 R is available under **General Public License** (GPL)
- 1997 **R core** group of programming developers is founded
- 2002 **R foundation for Statistical computing** is established in Vienna

A bit of history

- late 70s John Chambers and others at Bell Laboratories develop a statistical programming language named **S**
- late 80s Richard A. Becker, John M. Chambers and Allan R. Wilks publish *The New S Language* in which the system is rewritten in C
- early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland
- 1995 R is available under **General Public License** (GPL)
- 1997 **R core** group of programming developers is founded
- 2002 **R foundation for Statistical computing** is established in Vienna

Buyers

Trainees Programming Skills

1 Which Operating System (OS) do you prevalently use?

Response	Average	Total
Windows	67%	22
Linux	24%	8
Unix	9%	3
Total	100%	33/23

Buyers

Trainees Programming Skills

2 What is your expertise in programming (1 none; 2 little; 3 fair; 4 much; 5 expert)?

	1	2	3	4	5	Average rank ⇓
Experience in programming						2.7

Responses	1	2	3	4	5	Total
Experience in programming	3 (13%)	**8** (35%)	6 (26%)	5 (22%)	1 (4%)	**23**

Buyers

Trainees Programming Skills

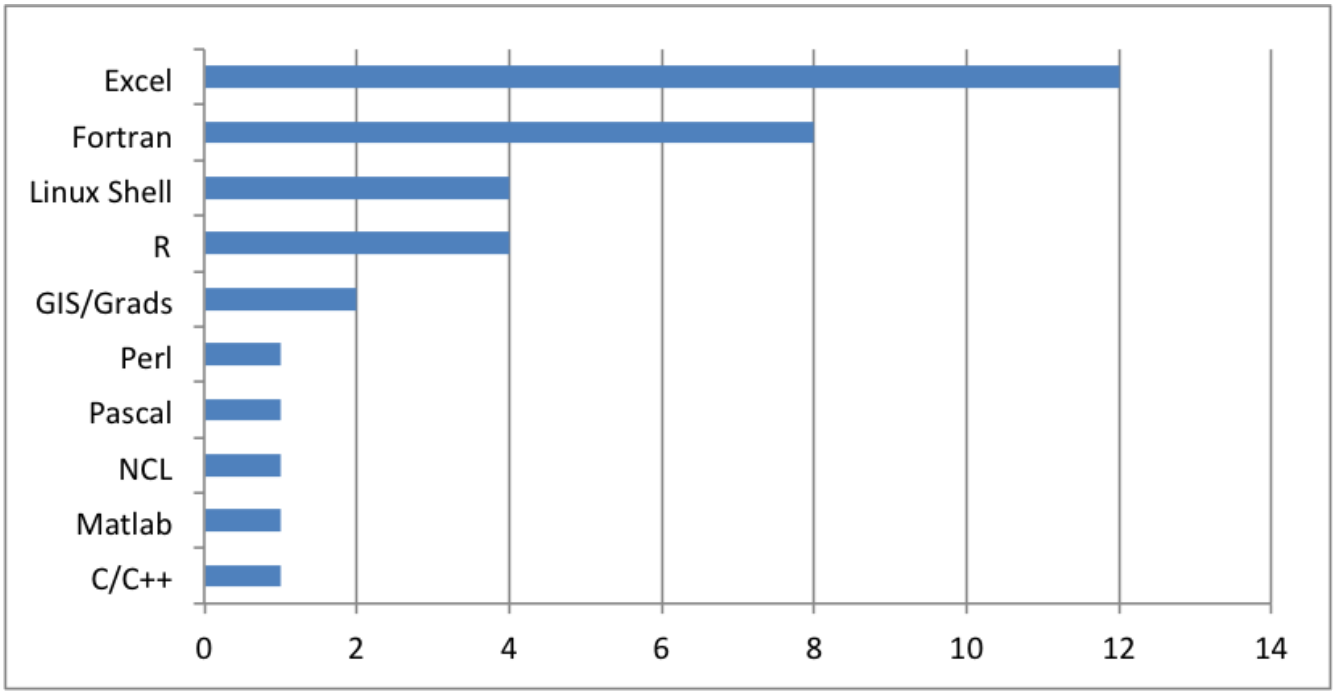
3 Can you manipulate and analyzing data with tools other than spreadsheet?

Response	Average	Total
Yes	52%	12
No	48%	11
Total	100%	23/23

Buyers

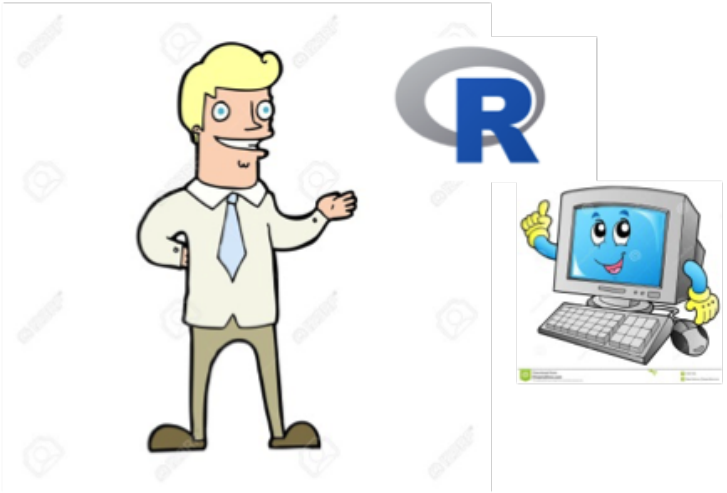
Trainees Programming Skills

4 Which software or language are you familiar with?



Merchant

He is a Statistician



Do not trust in statisticians

Outline

That is the question

What is really R:

- a tool for data mining?
- a programming language for data mining?

"Self-service analytics": R vs Competitors

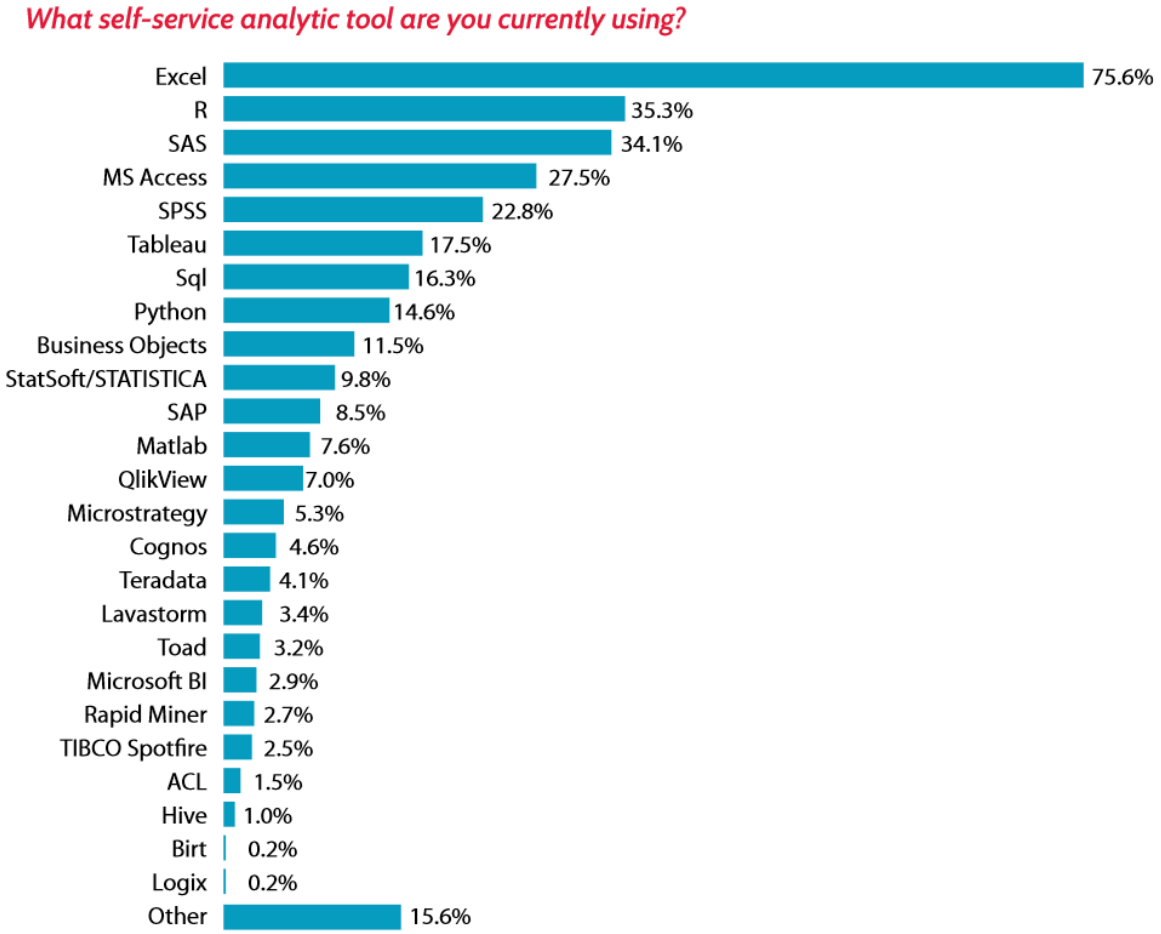


Figure: Lavastorm, Inc. survey of analytic communities. The results were published in **March, 2013**

R is an interpreted language

It means that **CODE** entered into the **R CONSOLE** (or run as R script in batch mode) is executed by the **interpreter**, a program within the R system.

Object oriented programming

Data are stored in objects.

- vector
- matrix
- array
- factors
- dataframe
- list
- ...

Every object has a **type** (double, complex, etc.) and is a member of a **class** (numeric, character, etc.).

R code is composed of

a series of expressions

- arithmetic expressions
- assignment statements
- conditional statements
- loops
- ...

functions

- pre-loaded
- via packages to be installed and loaded
- own

“Programming languages”: Competitors vs R

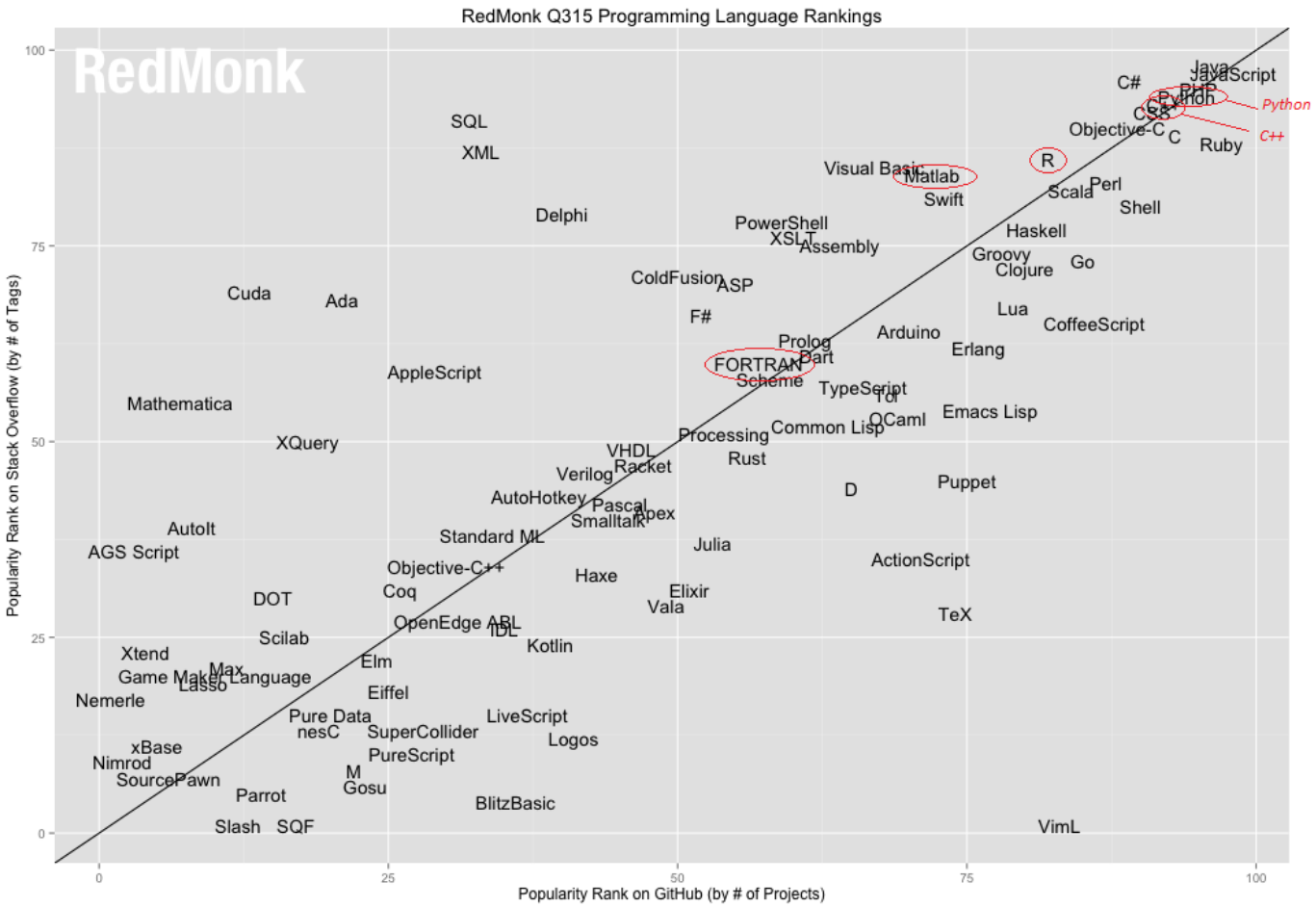


Figure: RedMonk survey: Programming language popularity as measured by **number of projects on GitHub** and the amount of **discussion on StackOverflow**.

"Tools" or "Programming languages" for data mining: R vs Competitors

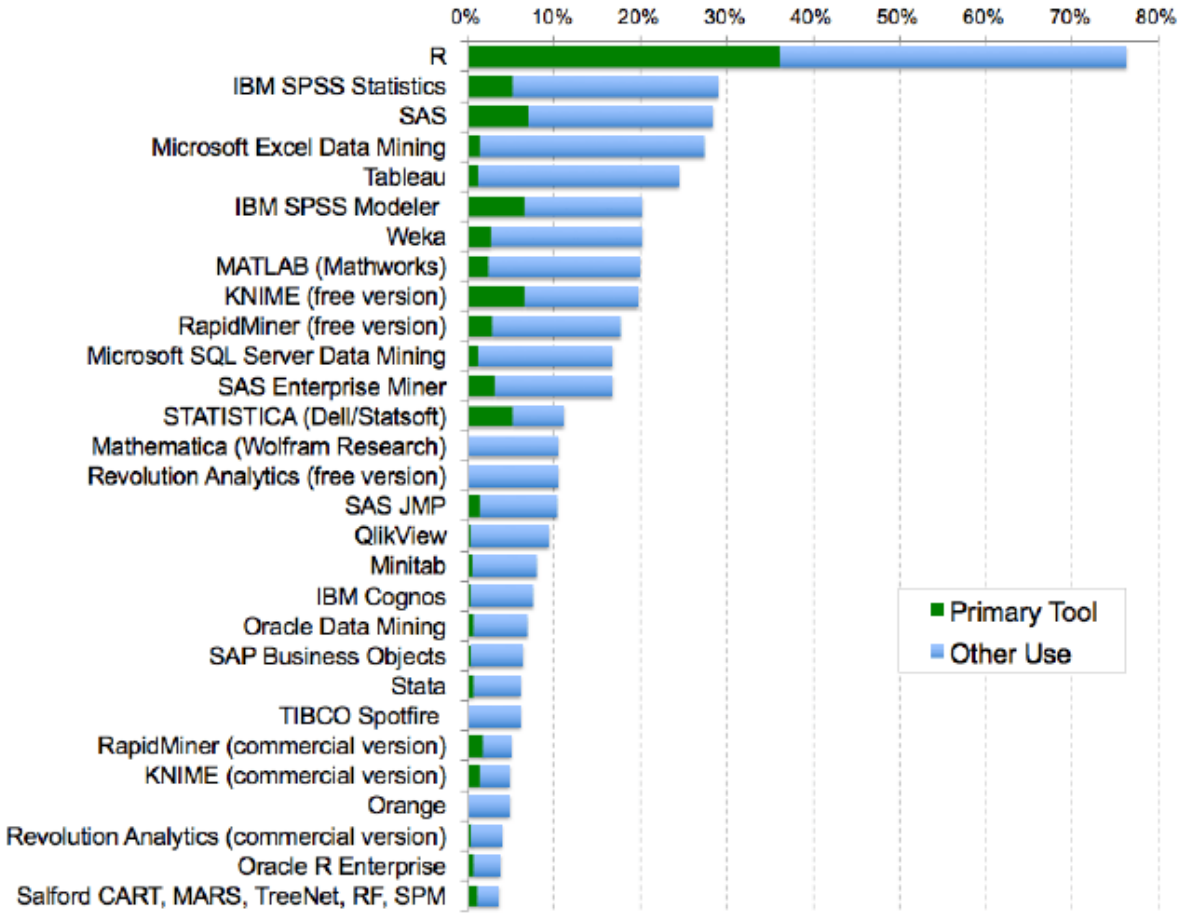


Figure: Analytics tools used by **1**, **220** respondents to the 2015 Rexer Analytics Survey. In this view, each respondent was free to check multiple tools.

“Data science software”: R vs Competitors

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	% users in 2015 % users in 2014 % users in 2013
R (1293), 3.6% alone	46.9% 38.5% 37.4%
RapidMiner (870), 13.7% alone	31.5% 44.2% 39.2%
SQL (853), 0% alone	30.9% 25.3% na
Python (837), 0% alone	30.3% 19.5% 13.3%
Excel (631), 0% alone	22.9% 25.8% 28.0%
KNIME (553), 6.7% alone	20% 15.0% 5.9%
Hadoop (507), 0% alone	18.4% 12.7% 9.3%
Tableau (341), 0% alone	12.4% 9.1% 6.3%
SAS base (313), 0.6% alone	11.3% 10.9% 10.7%
Spark (311), 0% alone	11.3% 2.6% na
Weka (310), 0% alone	11.2% 17.0% 14.3%
SAS Enterprise Miner (302), 3.6% alone	10.9% 7.2% 5.9%

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	% users in 2015 % users in 2014 % users in 2013
Microsoft SQL Server (268), 0% alone	9.7% 10.5% 7.0%
MATLAB (243), 0% alone	8.8% 8.4% 9.9%
scikit-learn (229), 0% alone	8.3% na na
Unix shell/awk/gawk (221), 0% alone	8.0% 5.8% na
IBM SPSS Statistics (213), 0% alone	7.7% 7.7% 8.7%
IBM SPSS Modeler (197), 7.1% alone	7.1% 5.7% 6.1%
Alteryx (155), 39.4% alone	5.6% 3.1% 0.3%
Pig (150), 0% alone	5.4% 3.5% na
Other programming languages (140), 0% alone	5.1% 3.0% na
Other free analytics/data mining tools (138), 0% alone	5.0% 5.1% 3.4%
Other Hadoop/HDFS-based tools (125), 0% alone	4.5% 3.9% na

Figure: KDnuggets 2015 poll

"Papers using data analytics tools": R vs Competitors

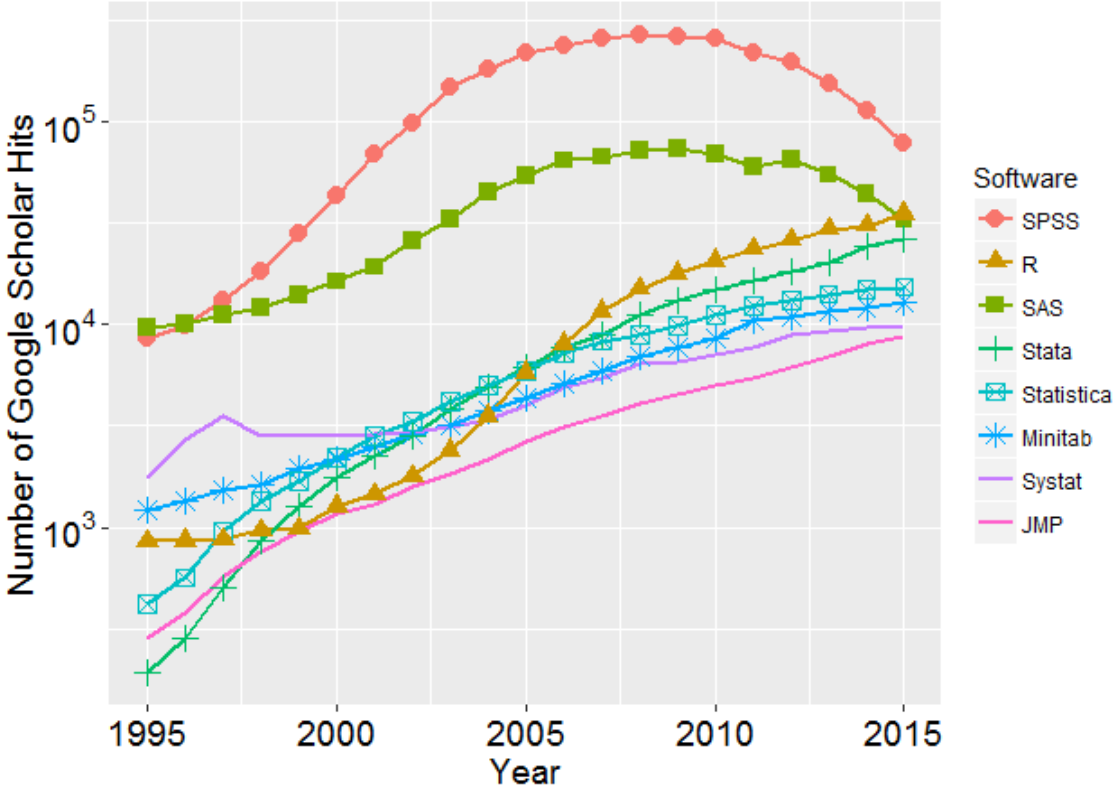


Figure: A logarithmic view of the number of scholarly articles found in each year by Google Scholar. This combines the previous two figures into one by compressing the y-axis with a base 10 logarithm.

Outline

Statistical modeling

Most statistical theory focuses on `data modeling`, `prediction` and `inference`.

Most known statistical models are:

- Time series analysis
- Linear regression
- Non parametric regression
- Cluster Analysis
- Principal Component Analysis
- Bayesian modeling
- ...

A simple example: Multivariate Linear Regression

Let $\mathbb{Y}$ be the **DJF cumulate precipitation** of a specific grid area (cell), and (X_1, X_2, X_3) the indexes North Atlantic Oscillation (NAO), Atlantic multidecadal oscillation (AMO) and El Niño-Southern Oscillation (ENSO), respectively.

We envision a regression model where $\mathbb{Y}$ **depends on large scale drivers** and we want to compute the estimate of model coefficients in order to use them for forecasting DJF precipitation:

$$Y_{prec}^{DJF} = \beta_0 + \beta_1 X_1^{NAO} + \beta_2 X_2^{AMO} + \beta_3 X_3^{ENSO} + \epsilon$$

We should decide whether there is any SIGNIFICANT RELATIONSHIP between the dependent variable y and any of the independent variables $X_k \quad (k = 1, 2, ..., p)$.

A simple example: Multivariate Linear Regression

Let $\mathbb{Y}$ be the **DJF cumulate precipitation** of a specific grid area (cell), and the indexes North Atlantic Oscillation (NAO), Atlantic multidecadal oscillation (AMO) and El Niño-Southern Oscillation (ENSO).

We envision a regression model where $\mathbb{Y}$ **depends on large scale drivers** and we want to compute the estimate of model coefficients in order to use them for forecasting DJF precipitation:

$$Y_{prec}^{DJF} = \beta_0 + \beta_1 X_1^{NAO} + \beta_2 X_2^{AMO} + \beta_3 X_3^{ENSO} + \epsilon$$

We should decide whether there is any SIGNIFICANT RELATIONSHIP between the dependent variable y and any of the independent variables $X_k \quad (k = 1, 2, ..., p)$.

A simple example: Multivariate Linear Regression

Excuse me, did you checked if the assumption that the error term ϵ in the MLR model is independent of X_k $(k = 1, 2, ..., p)$, and is normally distributed, such as

$$\epsilon \sim N(0, c)$$

with zero mean and constant variance holds?

Maybe you already know, in R you find a simple function for quantile-quantile plot, which is

```
qqnorm()
```

and also a package that contains several normality test

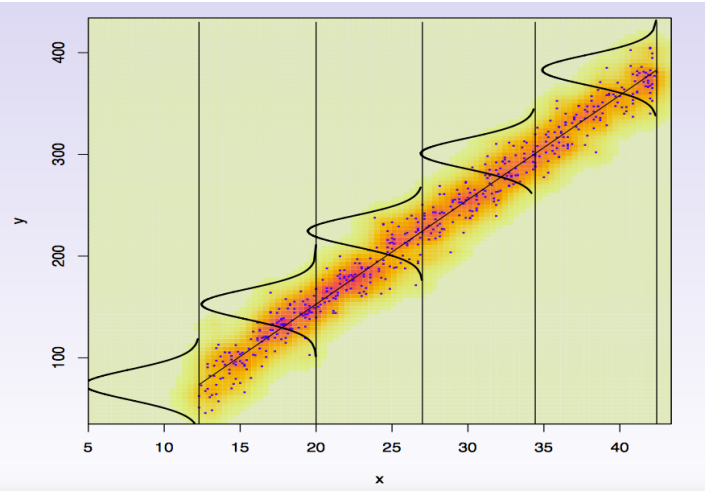
```
library(nortest)
cvm.test()     # Cramer-von Mises normality test
ad.test()        # Anderson-Darling test for normality
lillie.test()      # Lilliefors (Kolmogorov-Smirnov) test for normality
...
```

A more sophisticated example: GAMLSS

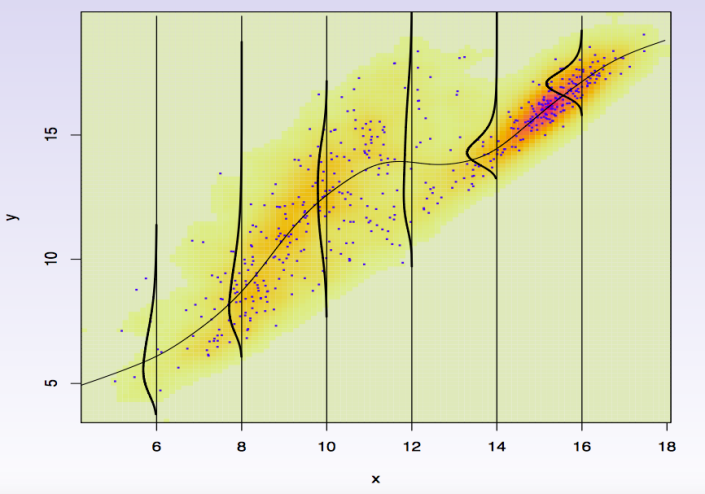
Generalized Additive Models for Location, Scale and Shape

`http://www.gamlss.org`

GAMLSS is a modern distribution-based approach to (semi-parametric) regression models, where all the parameters of the assumed distribution for the response can be modeled as additive functions of the explanatory variables



(a)



(b)

Figure: a) Linear regression modeling; b) Gamlss regression modeling

Outline

Data import

Supported data files

- .xls
- .csv
- .txt
- NetCDF
- .....

Connection to databases

- SAS
- Microsoft Access
- MySQL
- STATA
- .....

Data manipulation: `dplyr` package

Data manipulation is the process of **arranging data** in order to make data analysis process, such as visualizing and modeling.

```
filter(airquality, Temp > 70)

  Ozone Solar.R Wind Temp Month Day
1    36     118  8.0   72     5   2
2    12     149 12.6   74     5   3
3     7      NA  6.9   74     5  11
4    11     320 16.6   73     5  22
5    45     252 14.9   81     5  29
6   115     223  5.7   79     5  30
...
```

(a)

```
mutate(airquality, TempInC = (Temp - 32) * 5 / 9)

  Ozone Solar.R Wind Temp Month Day  TempInC
1    41     190  7.4   67     5   1 19.44444
2    36     118  8.0   72     5   2 22.22222
3    12     149 12.6   74     5   3 23.33333
4    18     313 11.5   62     5   4 16.66667
5    NA      NA 14.3   56     5   5 13.33333
...
```

(b)

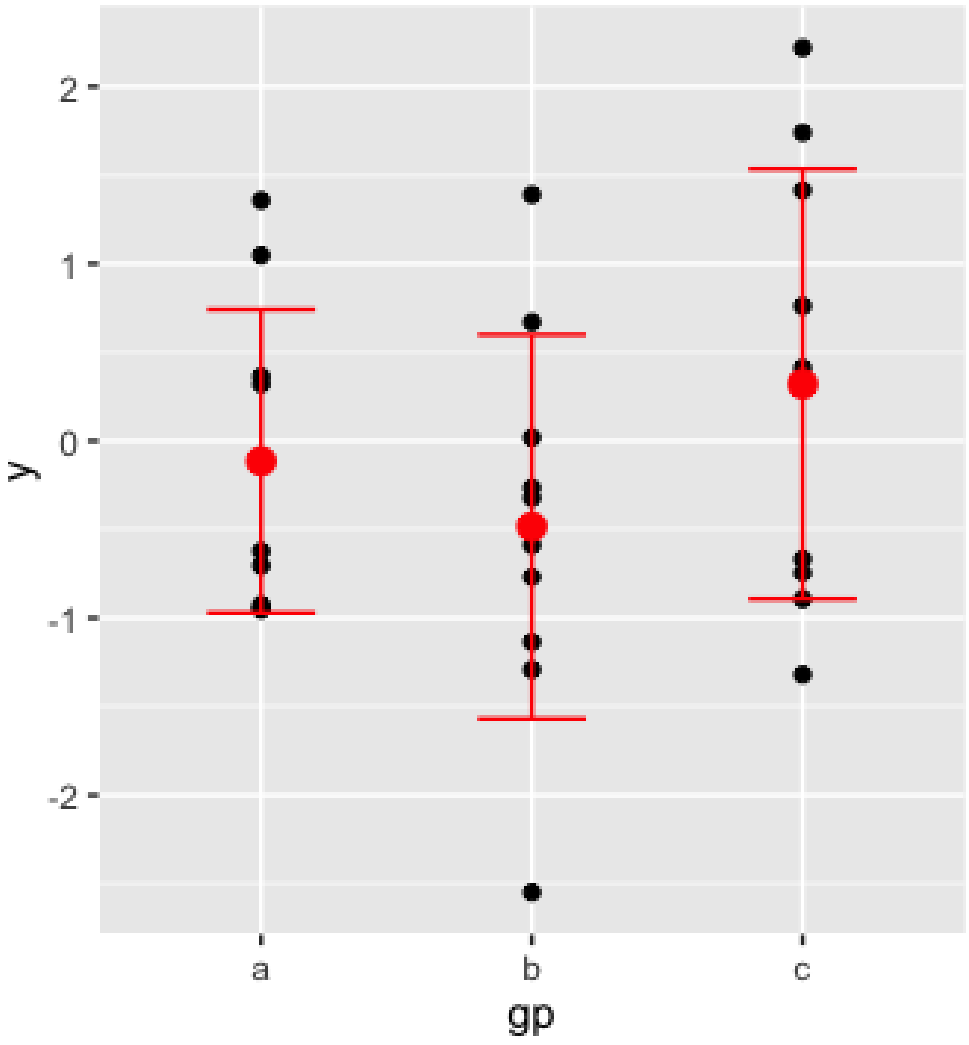
```
summarise(group_by(airquality, Month), mean(Temp, na.rm = TRUE))

  Month mean(Temp)
1     5   65.54839
2     6   79.10000
3     7   83.90323
4     8   83.96774
5     9   76.90000
```

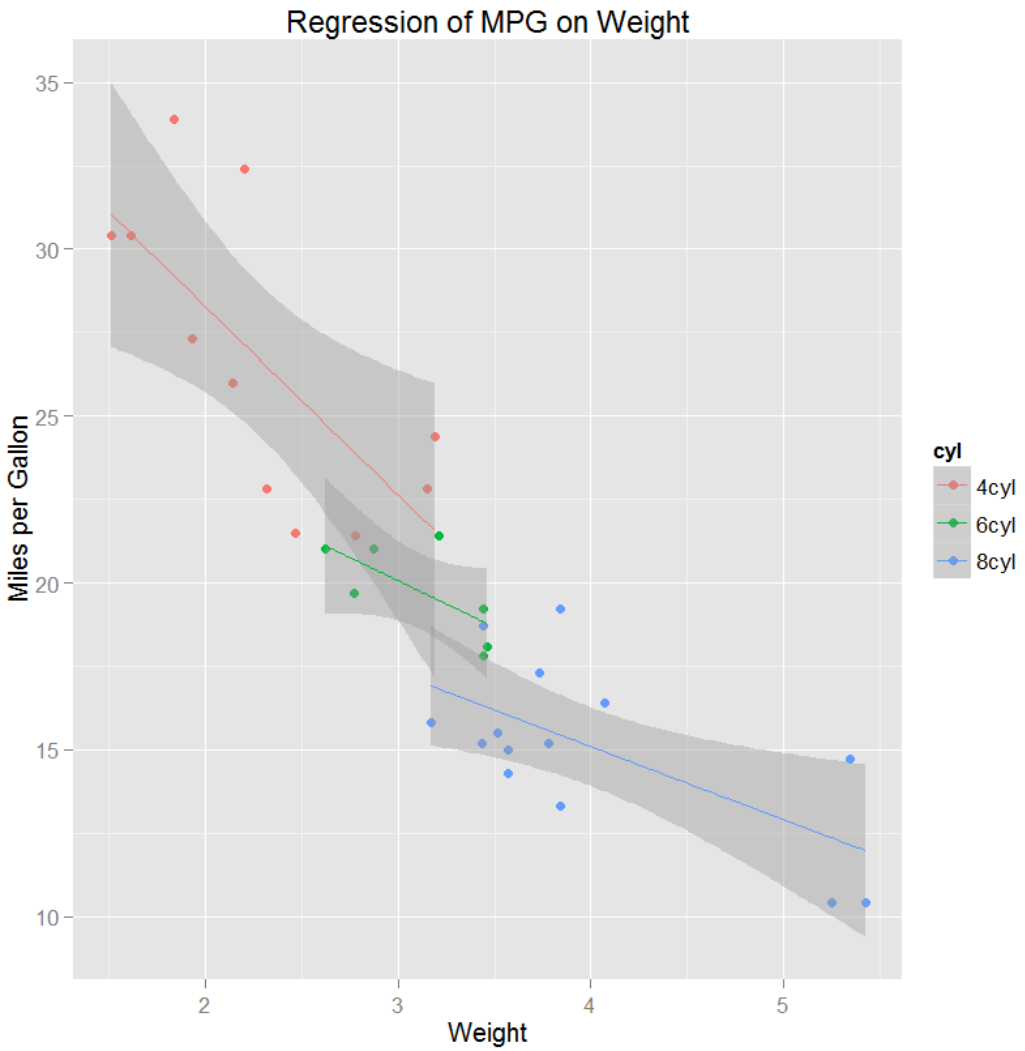
(c)

Outline

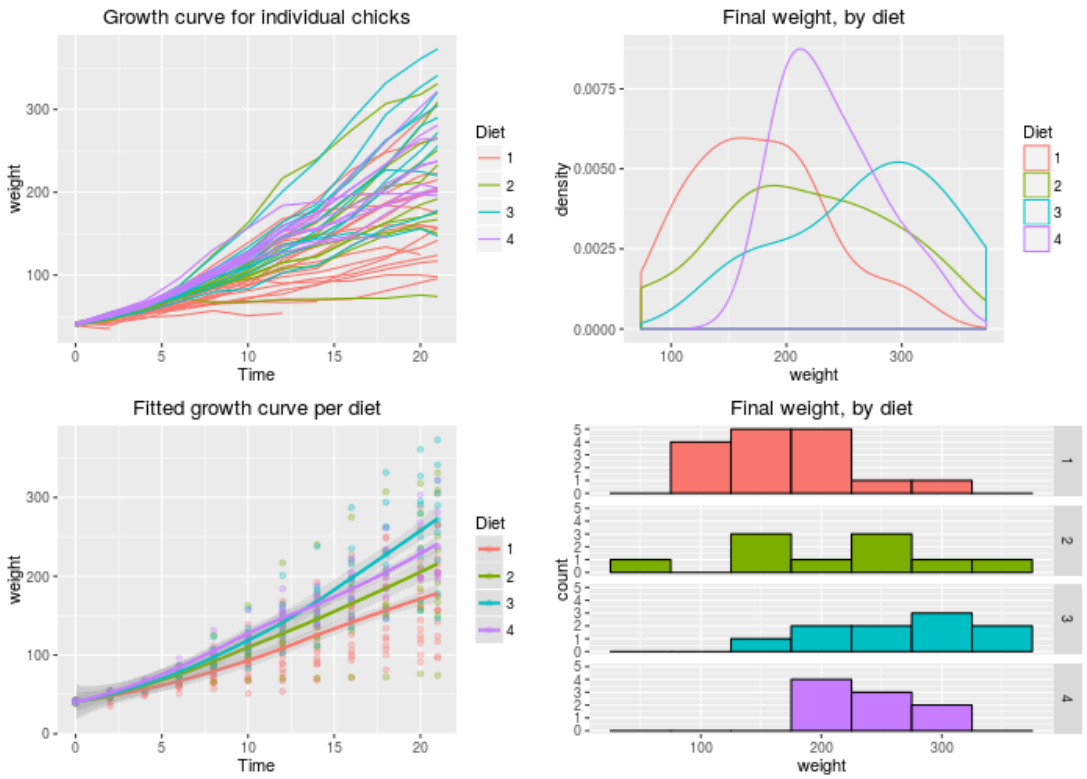
Visualizing data: `ggplot2` package



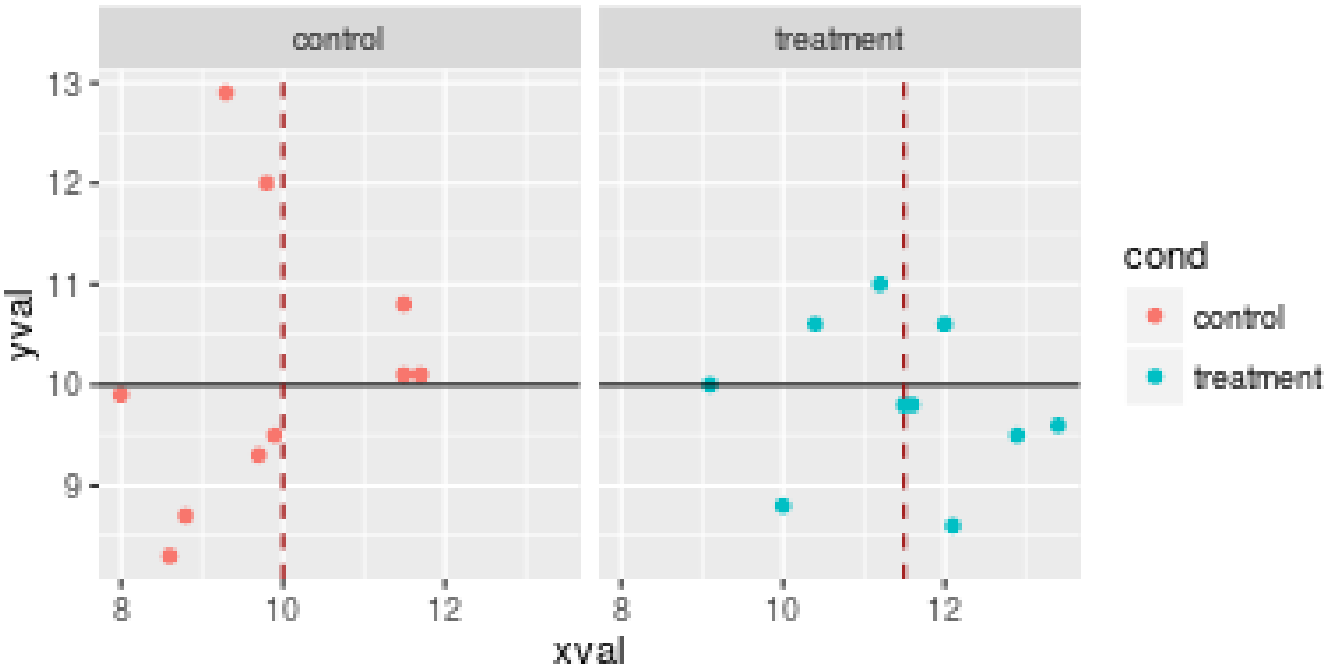
Visualizing data: `ggplot2` package



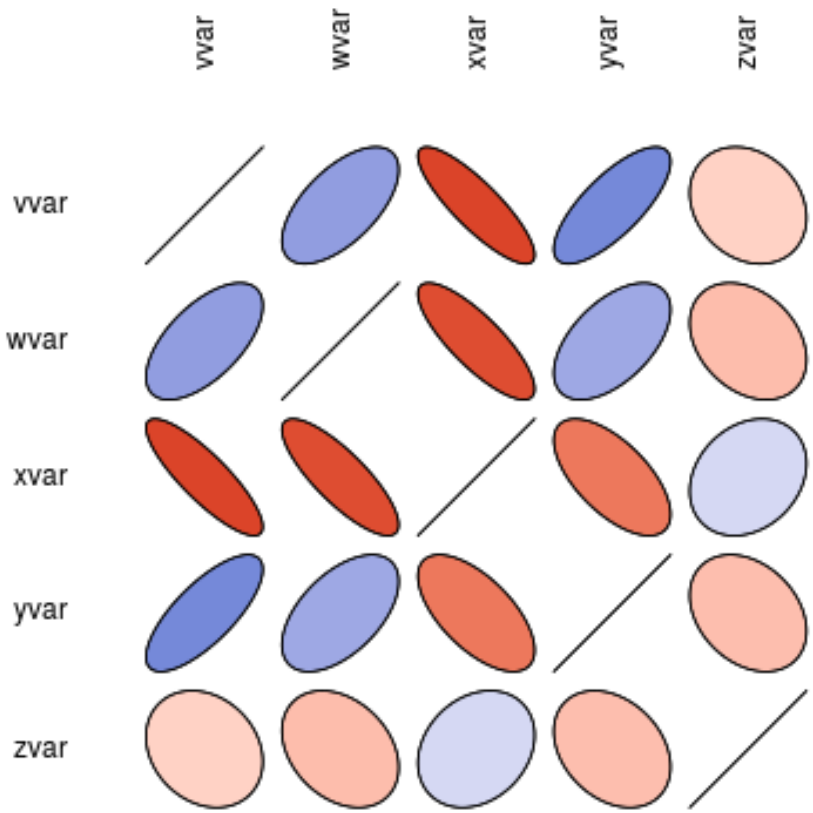
Visualizing data: `ggplot2` package



Visualizing data: `ggplot2` package



Visualizing data: `ggplot2` package



Outline

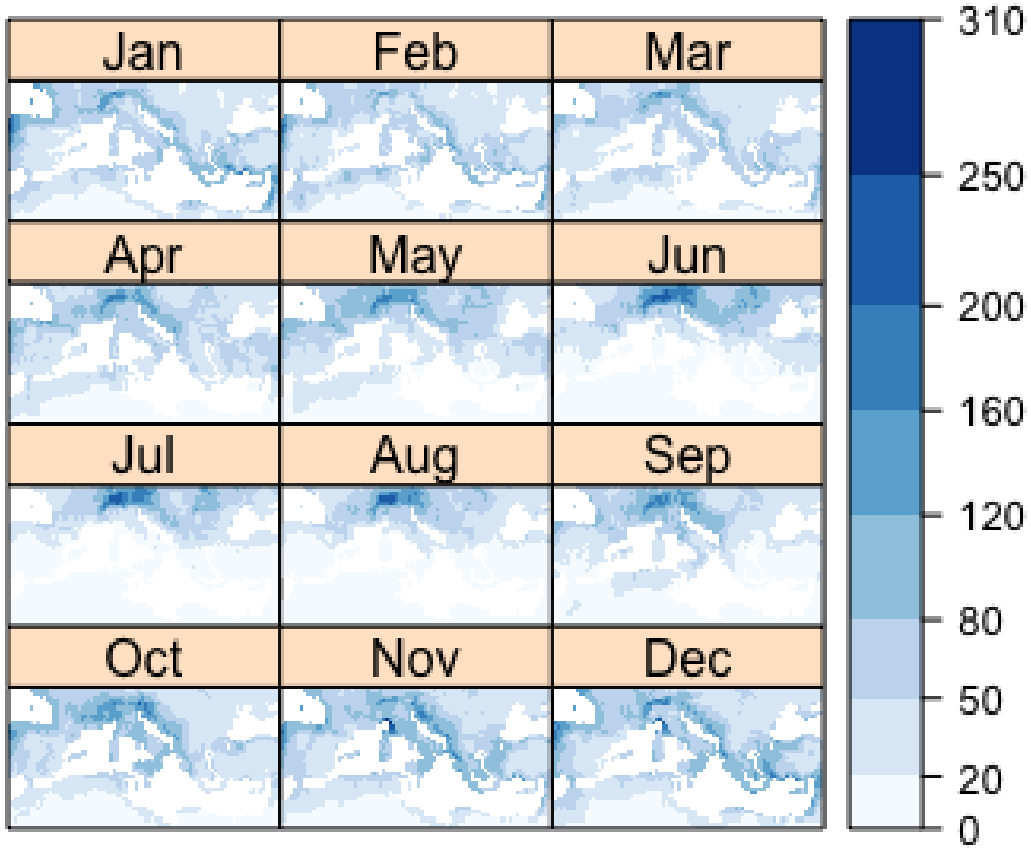
The raster package 1/3

The introduction of the raster package to R has been *a **revolution** for geo processing and analysis using R*.
Robert J. Hijmans is the original developer of the package. Among other things the

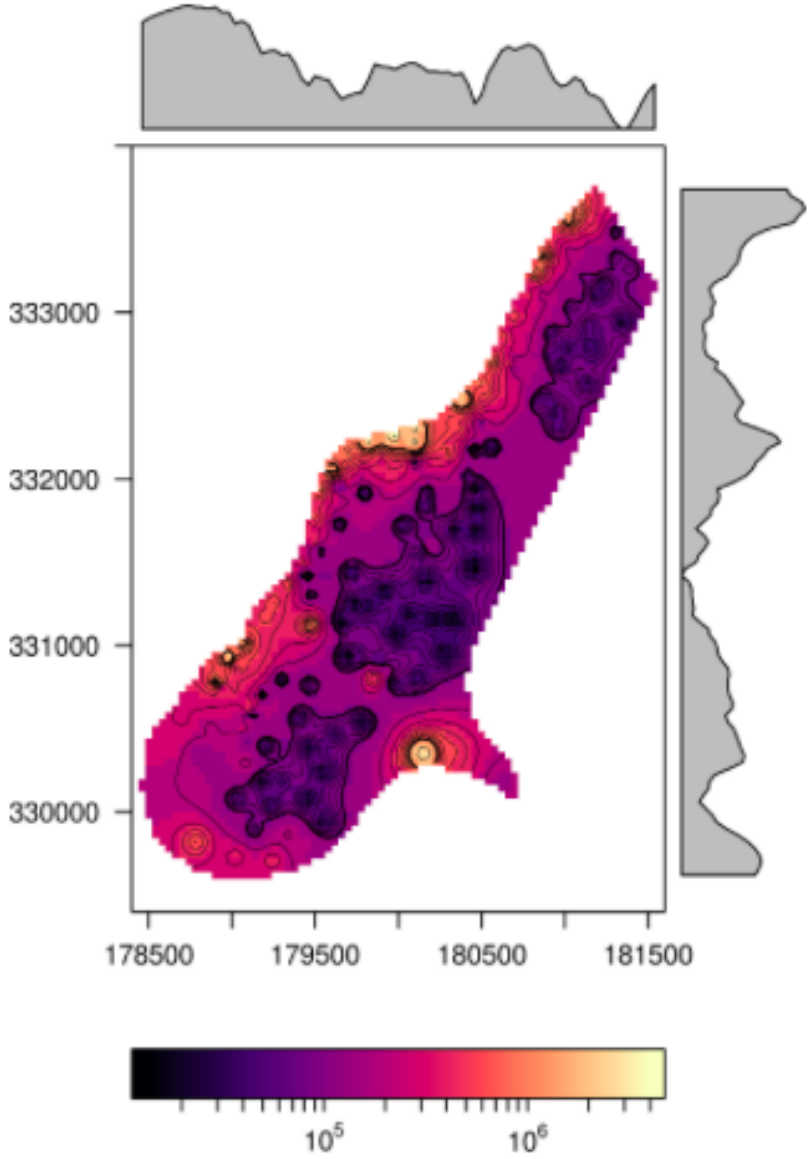
raster package allows to:

- Read and write raster data of **most commonly used** format (thanks to extensive use of rgdal)
- Perform most **raster operations** (creation of raster objects, performing spatial/geometric operations (re- projections, resampling, etc), filtering and raster calculations)
- Work on **large raster datasets** thanks to its built-in block processing functionalities
- **Visualize** and **interact** with the data
- etc...

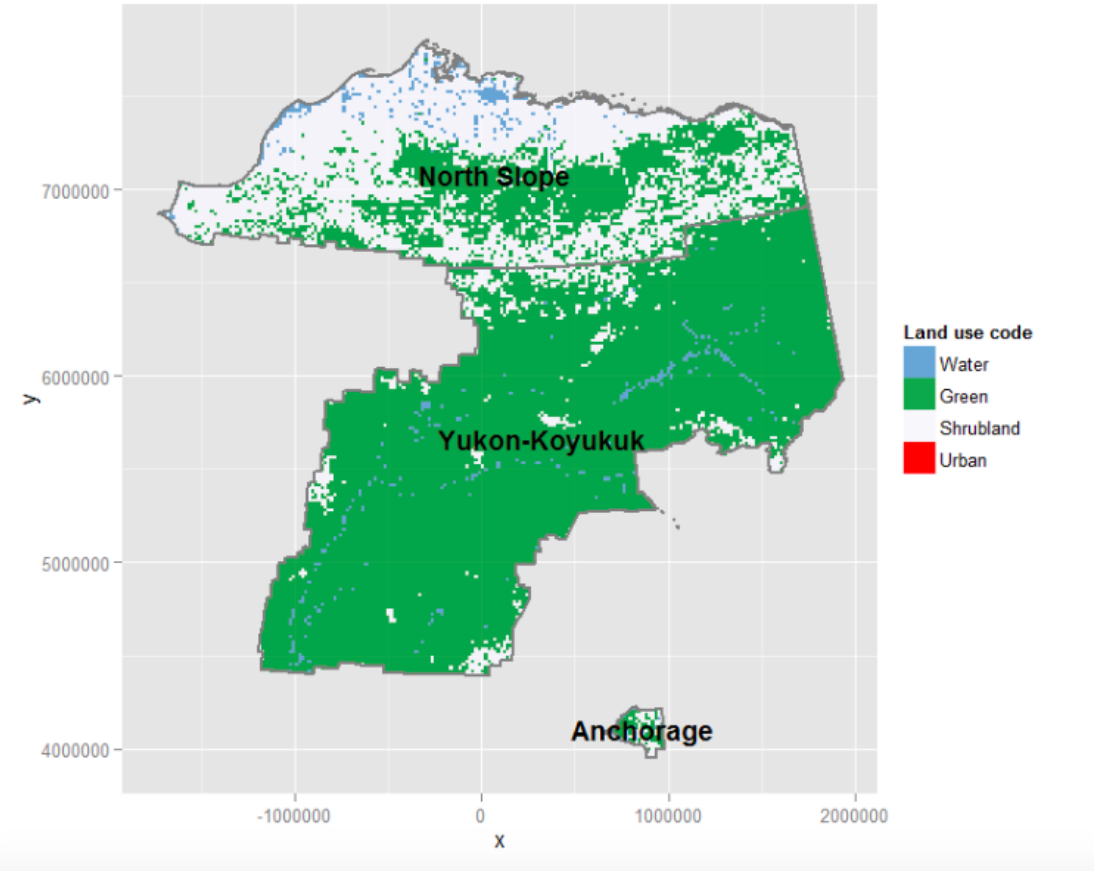
The raster package 2/3



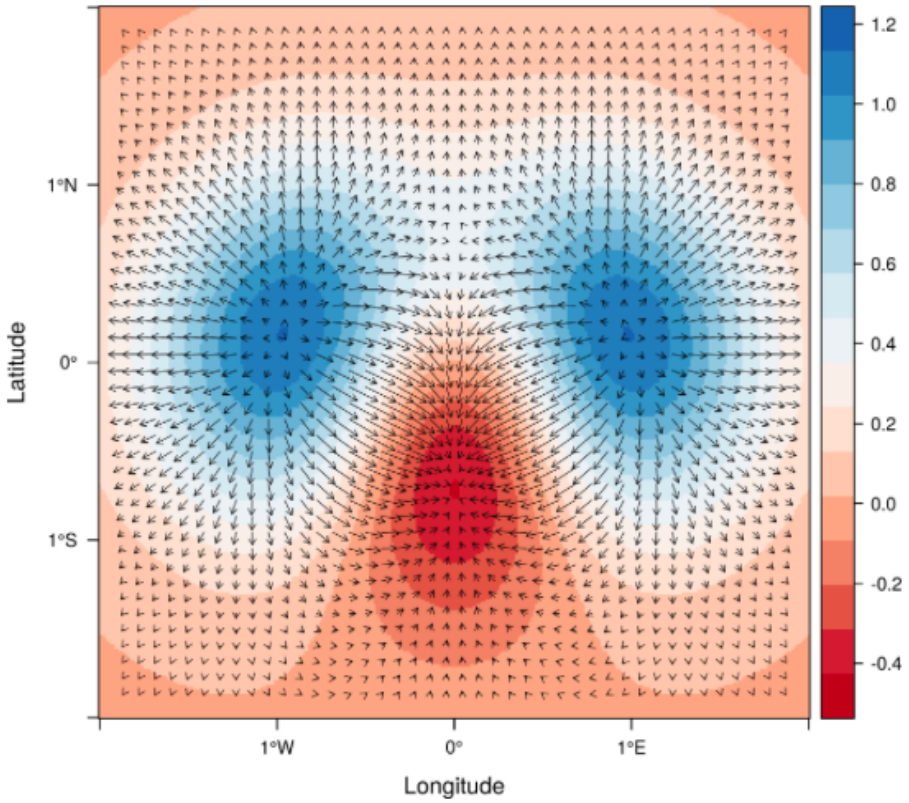
The raster package 2/3



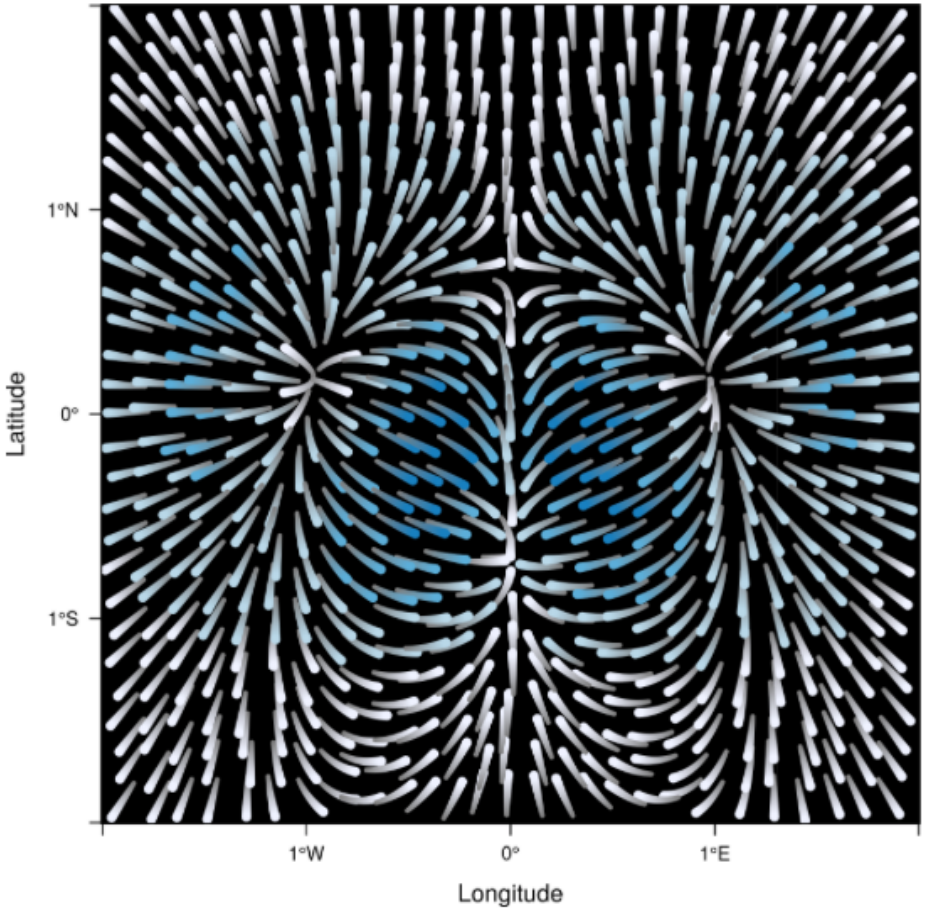
The raster package 2/3



The raster package 2/3



The raster package 2/3

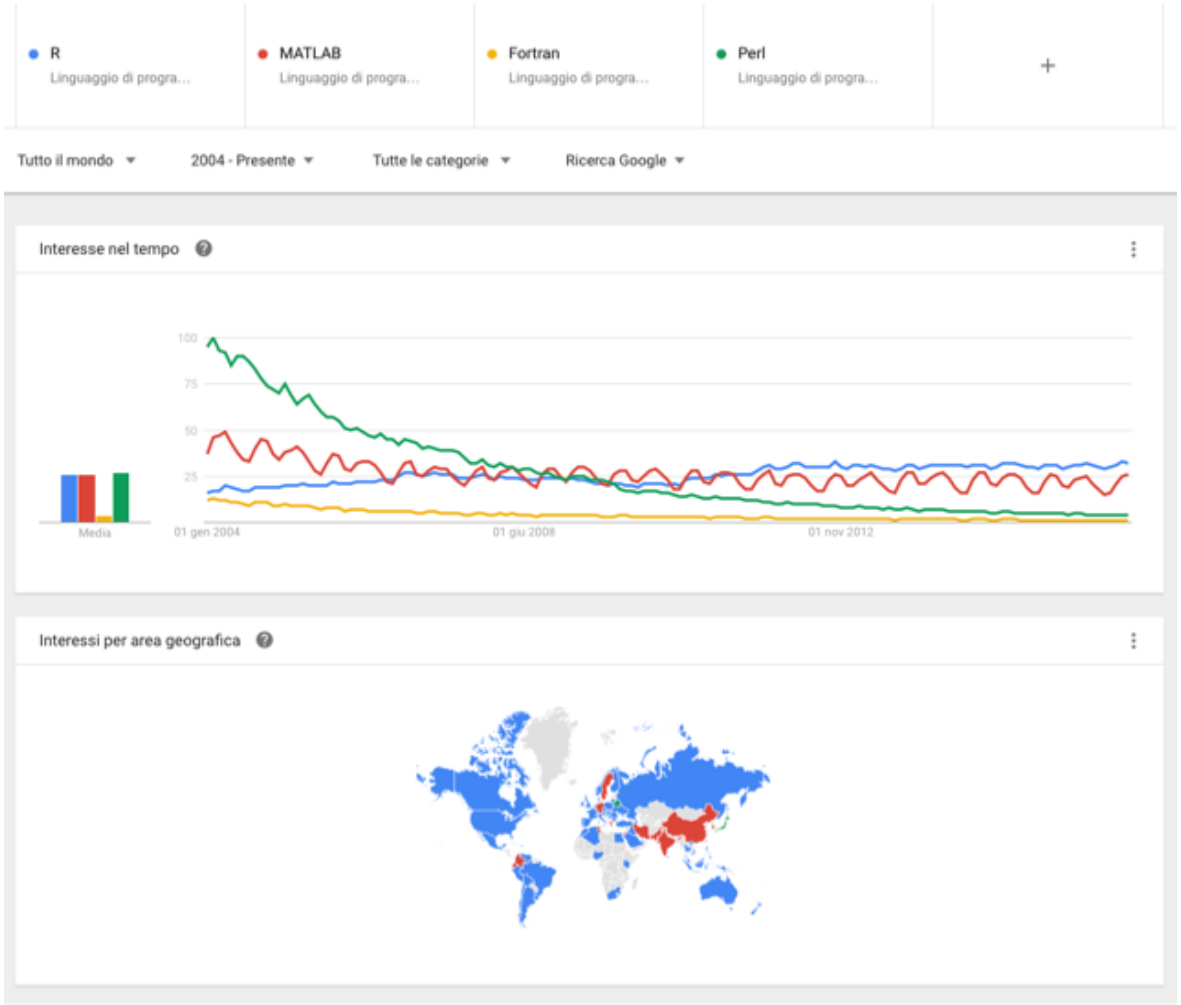


The raster package 3/3

CRAN Task View: Analysis of Spatial Data
(`https://cran.r-project.org/web/views/Spatial.html`)

`raster + rasterVis + sp + ggplot2 + ...`

R is *Google trendy*



To summarize

- R is not as powerful and generic as a programming language (it is a kind of DIALECT)
- R is native for statistical analysis and modeling
- R has a widespread support, however it is NON-COMMERCIAL

Time of decision