R in Biomedicine

Juan R Gonzalez
Bioinformatics Research Group in Epidemiology (BRGE), ISGlobal
Department of Mathematics, UAB

URL: http://brge.isglobal.org
e-mail: juanr.gonzalez@isglobal.org

ISGlobal Barcelona Institute for Global Health

Institució CERCA Centres de Recerca de Catalunya

A partnership of:

"la Caixa" Foundation

CLÍNIC BARCELONA Hospital Universitari · Parc de Salut Barcelona MAR · Universitat de Barcelona · upf. Universitat Pompeu Fabra Barcelona · Generalitat de Catalunya · Gobierno de España · Ajuntament de Barcelona · Fundación Ramón Areces

Barcelona Institute for Global Health

Biomedical Research Park of Barcelona, PRBB
Center for Genomic Regulation, **CRG**
European Molecular Biology Laboratory Barcelona, **EMBL-Barcelona**
University Pompeu Fabra, **CEXS-UPF**
Hospital del Mar Medical Research Institute, **IMIM**
Institute of Evolutionary Biology, **IBE (UPF-CSIC)**
Barcelona Institute for Global Health**, ISGlobal**

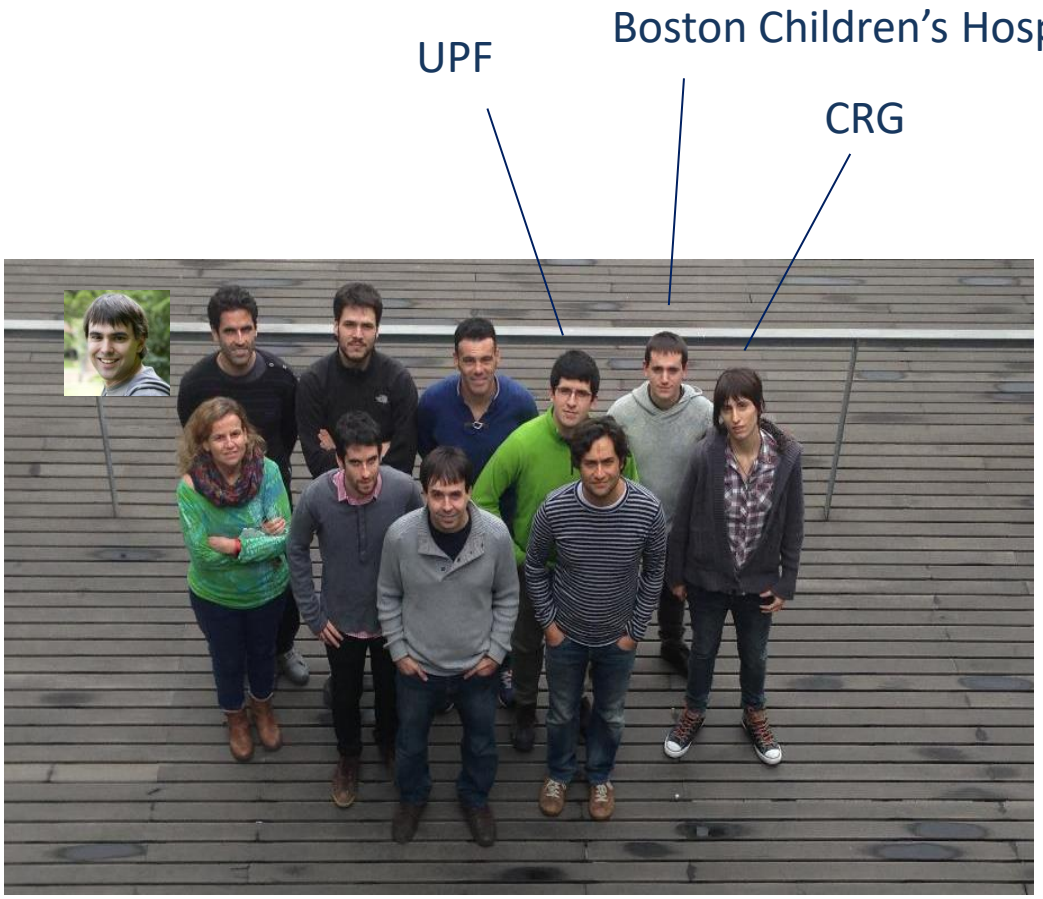
Bioinformatic Research Group in Epidemiology

ISGlobal
Alejandro Cáceres (Post-doc)
Carlos Ruiz (PhD Student)
Anna Montaner (PhD Student - qGenomics)
Dietmar Fernández (Technician – Bioinf)
Lara Nonell (PhD Student, IMIM)
Isaac Subirana (Post doc, IMIM)
Juan R González, PI

MIR
Julieta Polti

Alumni
10 MSc students (UAB, Uvic)
12 BSc students (Stats, Maths, …)

2 current BSc students
Nuria García (UPF – practicum)



Bioinformatic Research Group in Epidemiology

AIM: To study unexplored genetic variability in common complex diseases

- Genetic association studies (GWAS) [**SNPassoc**]
- Copy number variants (CNVs) [R-GADA, CNVassoc]
- Genetic mosaicims [MAD]
- Loss of chromosome Y [MADloy]
- Genetic inversions [inveRsion, invClust, scoreInvHap]
- Expsome data analysis [rexposome, omicRexposome]
- Post-omic data analysis [CTDquerier, psyGenet2R]
- Others (LOD, mGGA, …)

- **Databases**: dbGaP, EGA, EGCUT, UK Biobank, …

http://github.org/isglobal-brge

Search GitHub | Pull requests | Issues | Marketplace | Explore

Bioinformatic Research Group in Epidemiology (BRGE)

isglobal-brge

Bioinformatics, biostatistics, omic data integration, advanced statistical methods, R programming

Barcelona Institute for Global H...

Barcelona, Spain

https://brge.isglobal.org

Overview | Repositories 38 | Stars 0 | Followers 7 | Following 0

Pinned repositories — Customize your pinned repositories

CTDquerier

R package to make queries to Comparative Toxicogenomics Database and download the obtained results.

R ★ 1

MultiDataSet

Bioconductor package to encapsulate multiple datasets (including BioC omic data types)

R ★ 1

rexposome

Bioconductor package to characterize, analyze and integrate exposome with omic and disease data

R ★ 1

madloy_v1

Get summarized LRR to analyze LRR in association studies

R

MEAL

Methylation and Expression AnaLysis

R ★ 1

SNPassoc

Genetic association studies

R

333 contributions in the last year — Contribution settings

https://isglobal-brge.github.io/rexposome/

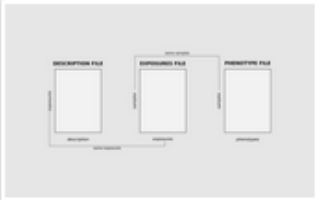
rexposome project

rexposome project is a set of R packages developed at *Bioinformatics Research Group in Epidemiology* (BRGE) from *Barcelona Global Health Institute* (ISGlobal). The project aims to provide a framework to incorporate **exposome** data in R/Bioconductor pipelines, with the goal of describing and analyzin exposome data.

Currently, the Bioconductor packages included in **rexposome project** are `rexposome` and `omicRexposome`. The first package, `rexposome`, offers an interface to load exposome data into Bioconductor-like objects and functions to describe and characterize the exposome. It also includes functions to perform **Exposome-Wide Association Studies**. The second package, `omicRexposome`, contains functions to perform exposome-omic association studies and multi-omic integration of exposome and omic data-sets. A third package, `postRexposome`, that facilitates post exposome data analyses aiming providing biological insights of positive findings, will be included soon in the project.

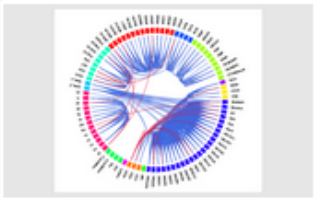
BRGE | ISGlobal | HELIX | R | Bioconductor | rexposome framework

Package Features



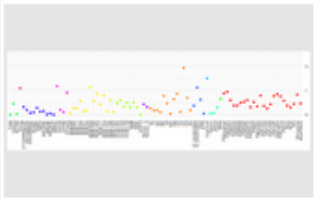
exposome data loading

`rexposome` implements `readExposome` and `loadExposome` functions. The first one is in charge to load into R exposome data stored in raw file and



exposome data exploration

Several functions are implemented in `rexposome` to explore and describe exposome data. The method `plotMissings` allow to analyse the



exposome data analysis

`rexposome` allows to perform EXposome-Wide Association Study (EXWAS) in both univariate and multivariate ways using standard

exposome-omic data analysis

`omicRexposome` implement the functions `associations` and `integration`. The fist allows to test the association between exposures and omic features (genes, CpGs...). The second allows to perform an integration of multiple layers, including exposome data.

Statistical Methods in Biomedicine

GLMs: Is smoking associated with lung cancer?

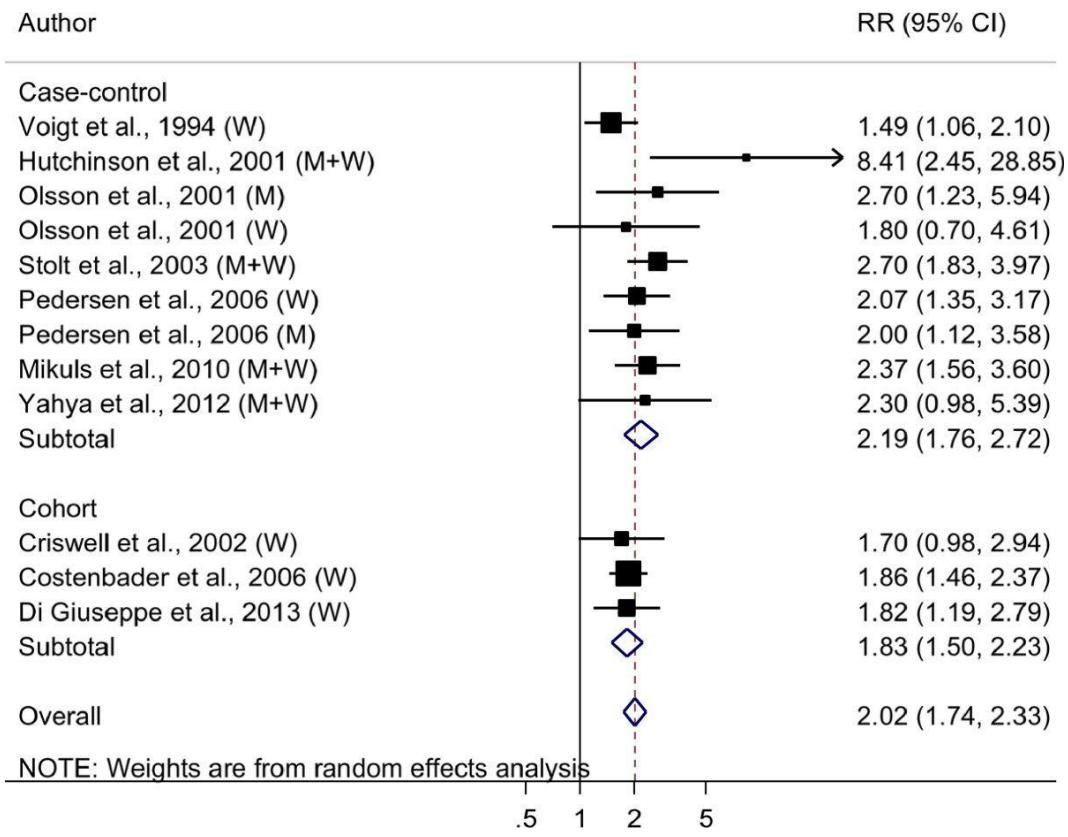
Table 4. Relative Risks of Smoking-Related Lung Cancer among Current and Former Smokers, According to the Level of Smoking.*

Smoking Level	African American	Native Hawaiian	Latino	Japanese American	White	Global P Value
≤10 Cigarettes/day						
Relative risk (95% CI)†	1.00	0.88 (0.60–1.29)	0.21 (0.14–0.31)	0.25 (0.18–0.36)	0.45 (0.34–0.60)	
P value		>0.5	<0.001	<0.001	<0.001	<0.001
Cases of lung cancer	215	34	52	50	54	
No. of participants	9886	2745	12,831	8378	7650	
11–20 Cigarettes/day						
Relative risk (95% CI)†	1.00	0.90 (0.74–1.12)	0.36 (0.29–0.44)	0.39 (0.32–0.47)	0.57 (0.49–0.68)	
P value		0.37	<0.001	<0.001	<0.001	<0.001
Cases of lung cancer	240	65	80	136	180	
No. of participants	6514	3062	4932	10,680	9877	
21–30 Cigarettes/day						
Relative risk (95% CI)†	1.00	0.93 (0.72–1.21)	0.61 (0.46–0.79)	0.61 (0.49–0.74)	0.73 (0.61–0.88)	
P value		>0.5	<0.001	<0.001	0.07	<0.001
Cases of lung cancer	65	24	27	102	157	
No. of participants	1671	1419	1406	4715	6062	
≥31 Cigarettes/day						
Relative risk (95% CI)†	1.00	0.95 (0.66–1.35)	0.79 (0.55–1.13)	0.75 (0.57–1.00)	0.82 (0.64–1.05)	0.31
P value		>0.5	0.38	0.31	>0.5	
Cases of lung cancer	45	35	26	64	124	
No. of participants	759	788	800	2305	3970	

* Global P values for racial and ethnic differences in risk for each smoking level were calculated with the use of the likelihood ratio test.
† Relative risks were adjusted for the duration of smoking, sex, and the time since quitting. Information about the model used to estimate relative risks is provided in the Supplementary Appendix. African Americans served as the reference group. CI denotes confidence interval.

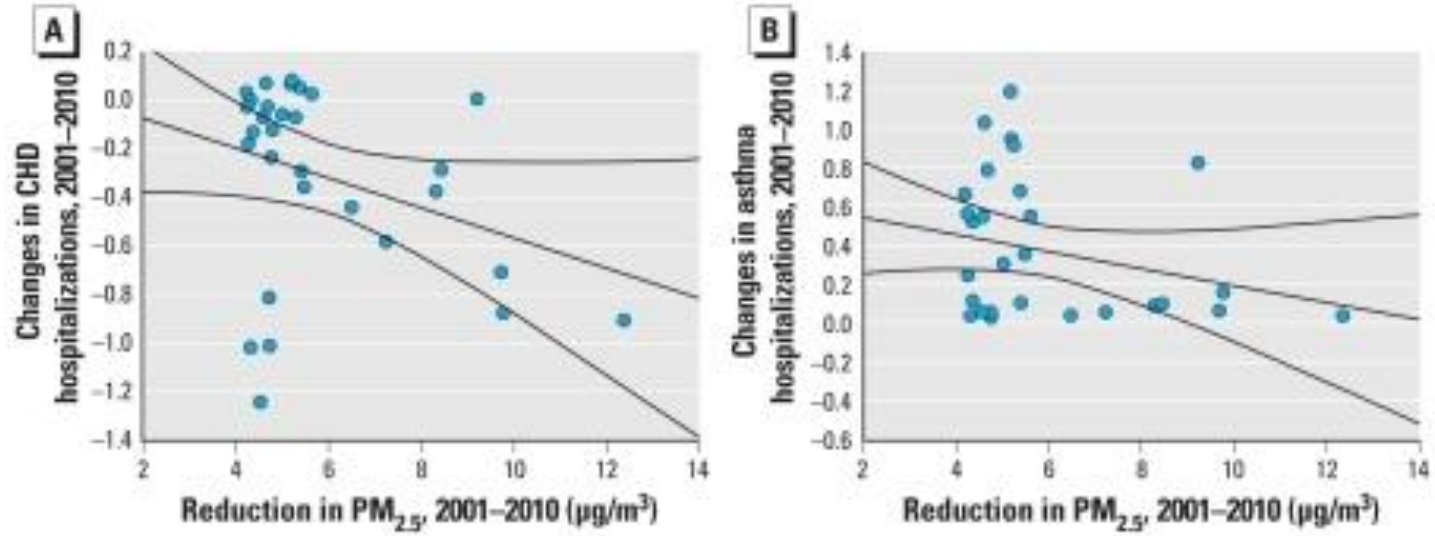
Statistical Methods in Biomedicine

Meta-analysis: Is smoking associated with lung cancer?

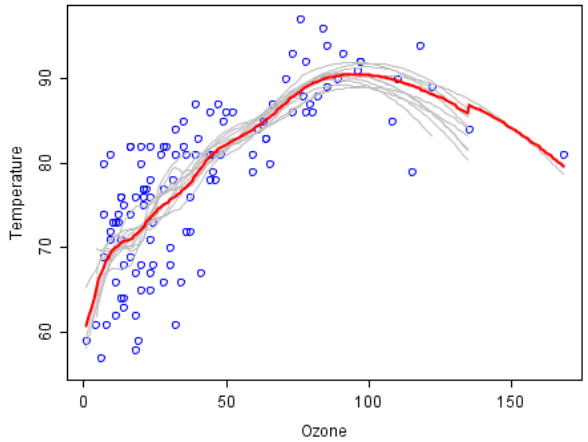


Statistical Methods in Biomedicine

GAMs: Is air pollution associated with asthma?

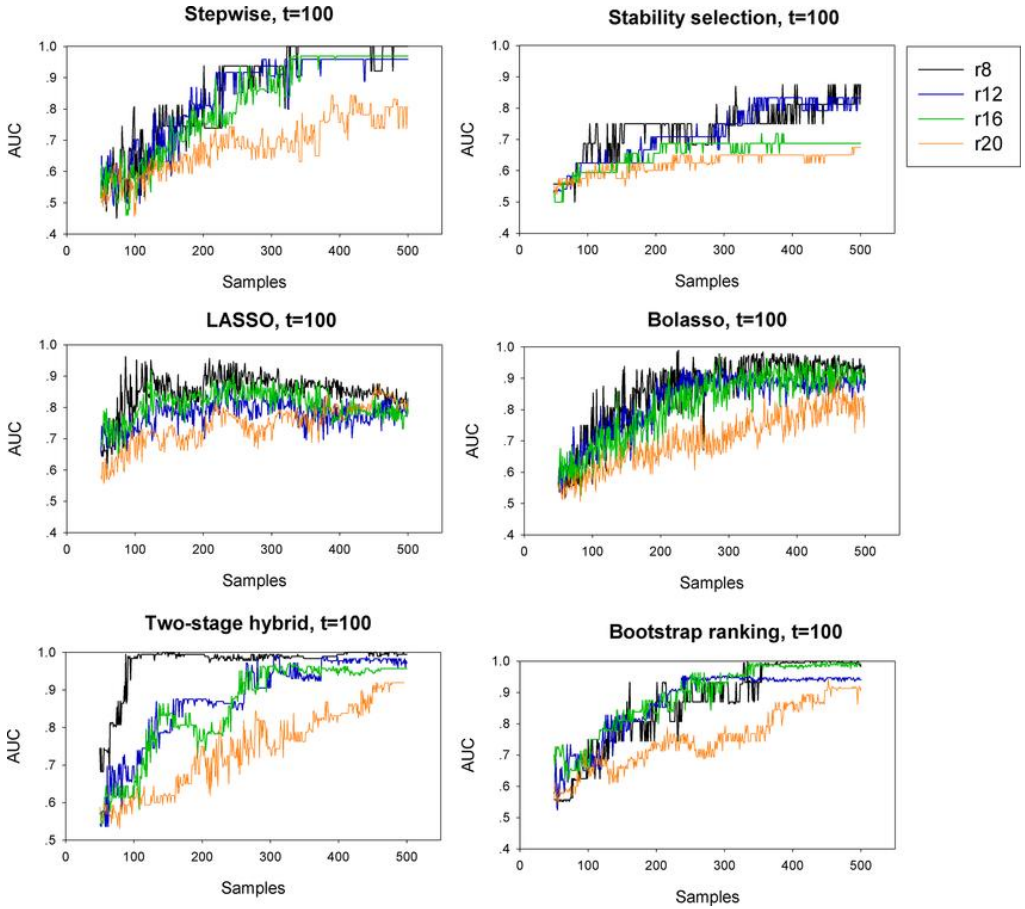


GAM: Generalized additive models



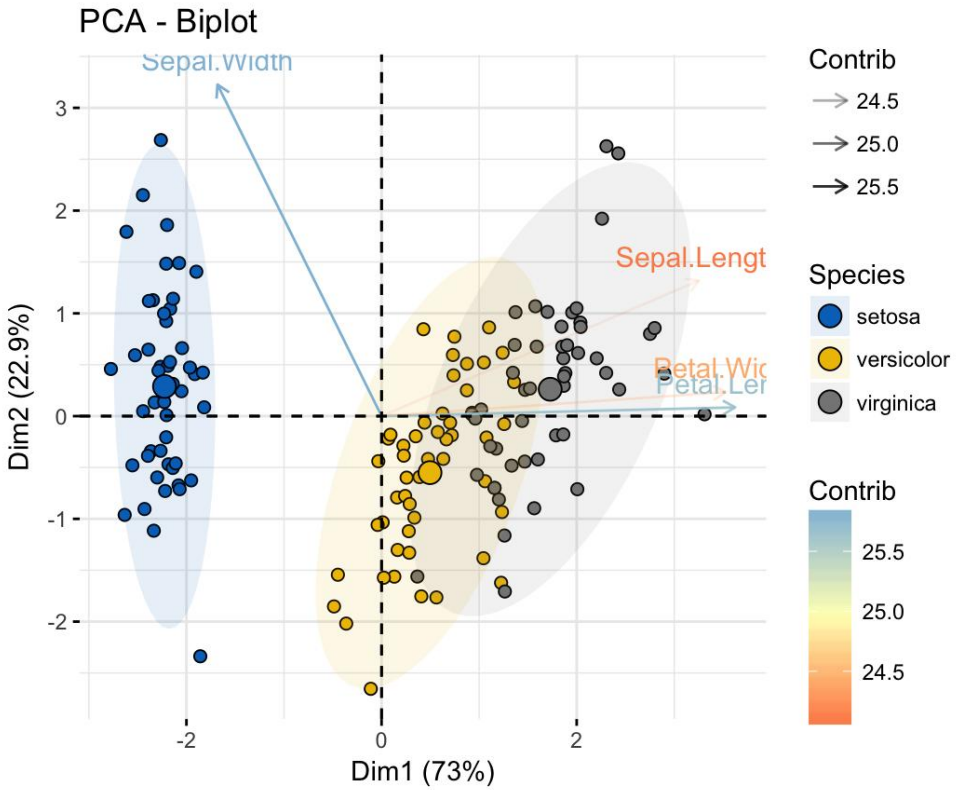
Statistical Methods in Biomedicine

Variable selection: Which factors are associated with cardiovascular disease?



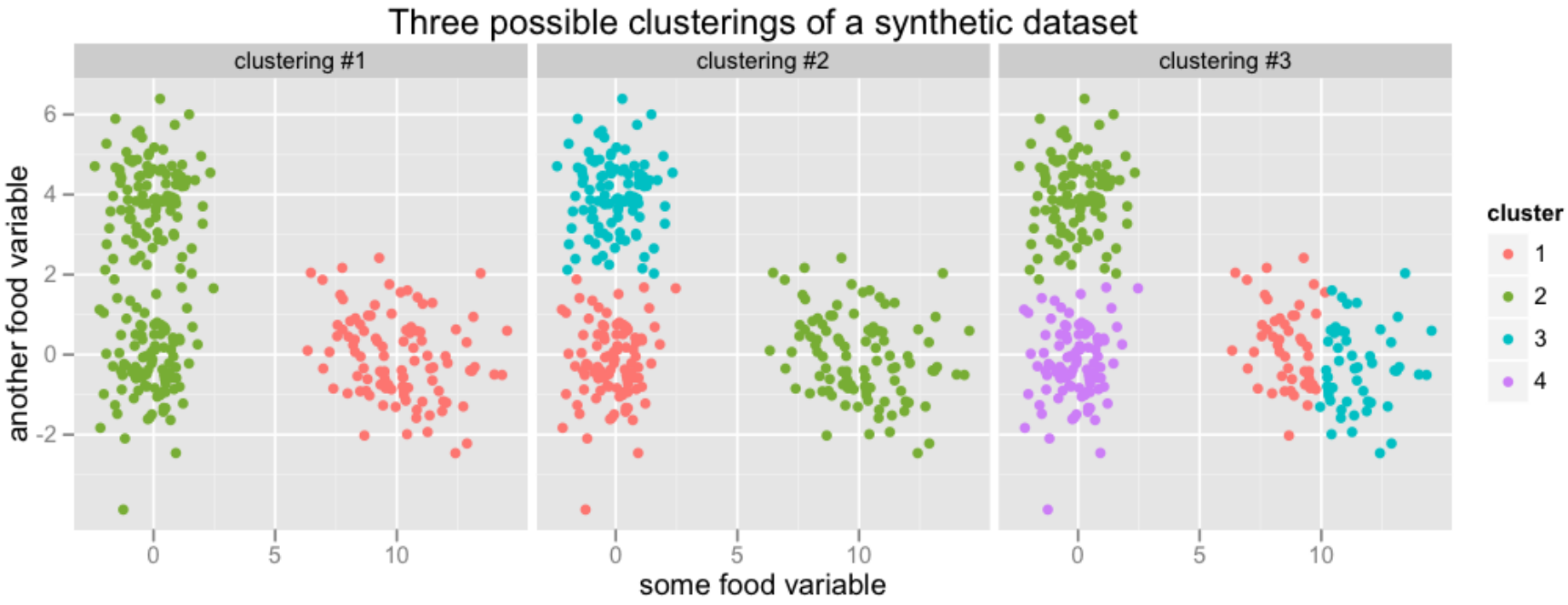
Statistical Methods in Biomedicine

Data Reduction (PCA, latent class analysis, …): Which factors explain the vast majority of observe variability?



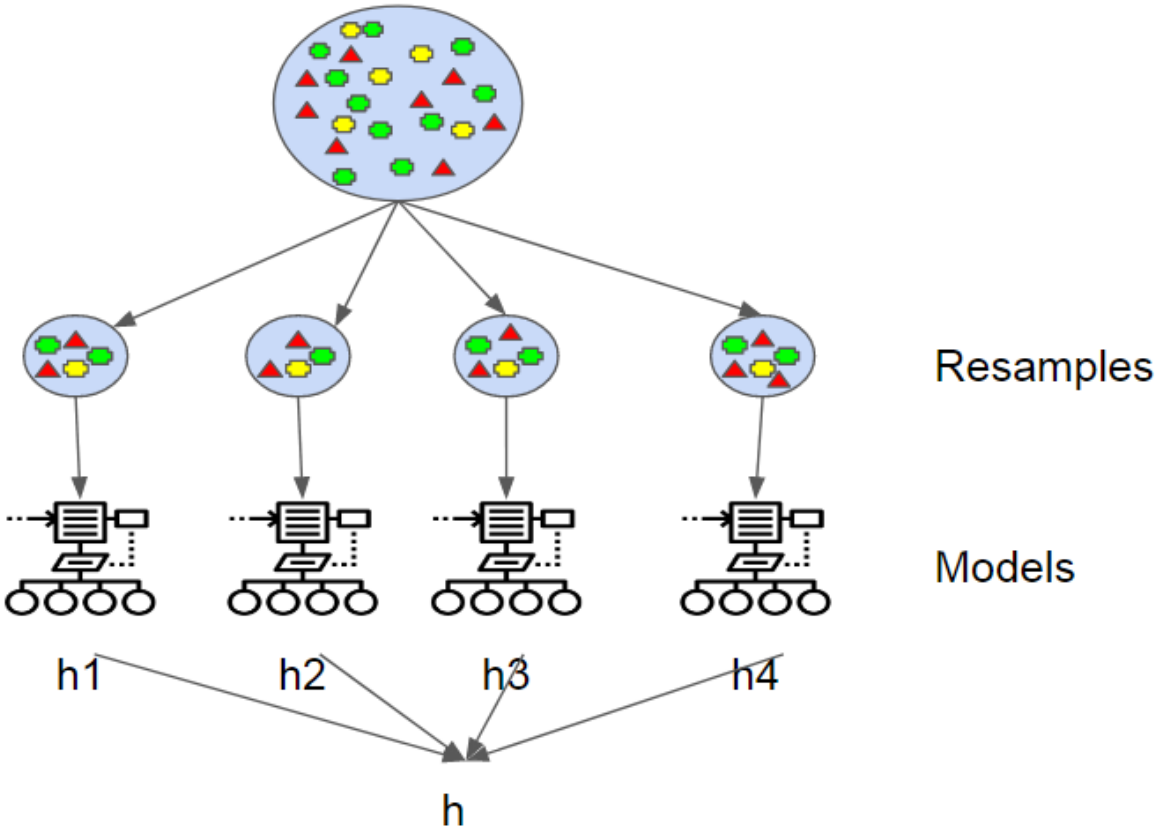
Statistical Methods in Biomedicine

Unsupervised Machine Learning (clustering, LDA, …):
How individuals cluster depending on their food consumption?

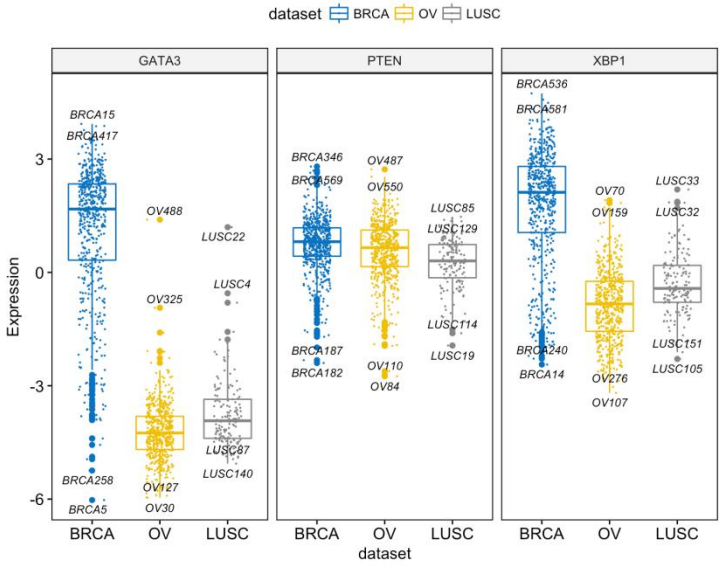


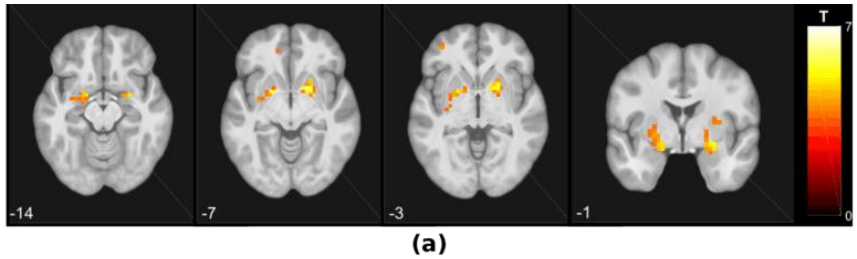
Statistical Methods in Biomedicine

Supervised Machine Learning (Boosting, Neural Networks, …): Model prediction

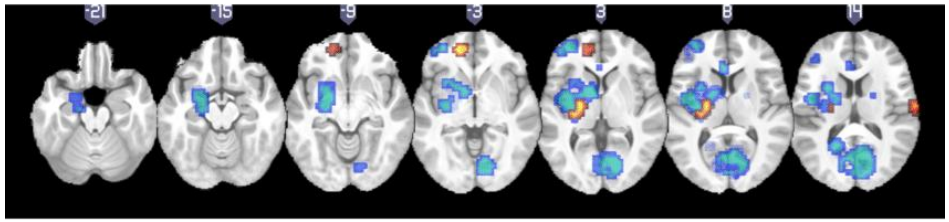


Data Visualization

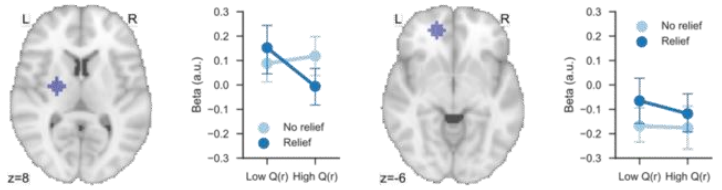




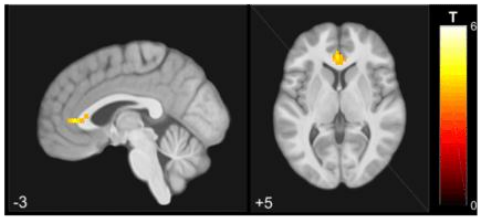
(a)



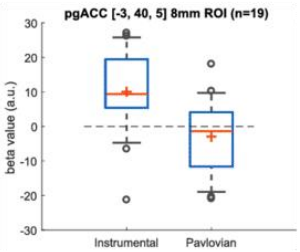
(b)



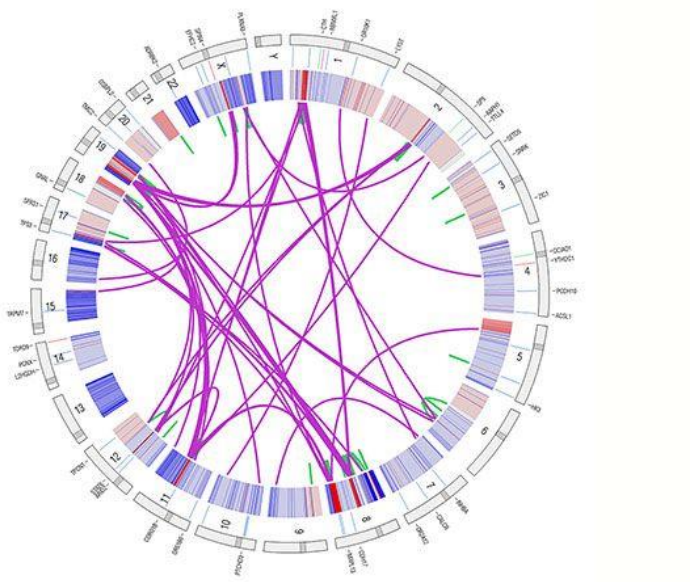
(c)



(d)



(e)



What's R

- Provide access to powerful **statistical and graphical** methods for the analysis of genomic data.
- Facilitate the **integration of biological metadata** (GenBank, GO, LocusLink, PubMed) in the analysis of experimental data.
- Allow the rapid development of **extensible**, **interoperable**, and **scalable** software.
- Promote high-quality documentation and **reproducible research**.
- Provide **training** in computational and statistical methods

Installing R: CRAN

CRAN: https://cran.rstudio.com/

CRAN
Mirrors
What's new?
Task Views
Search

The Comprehensive R Archive Network

Download and Install R

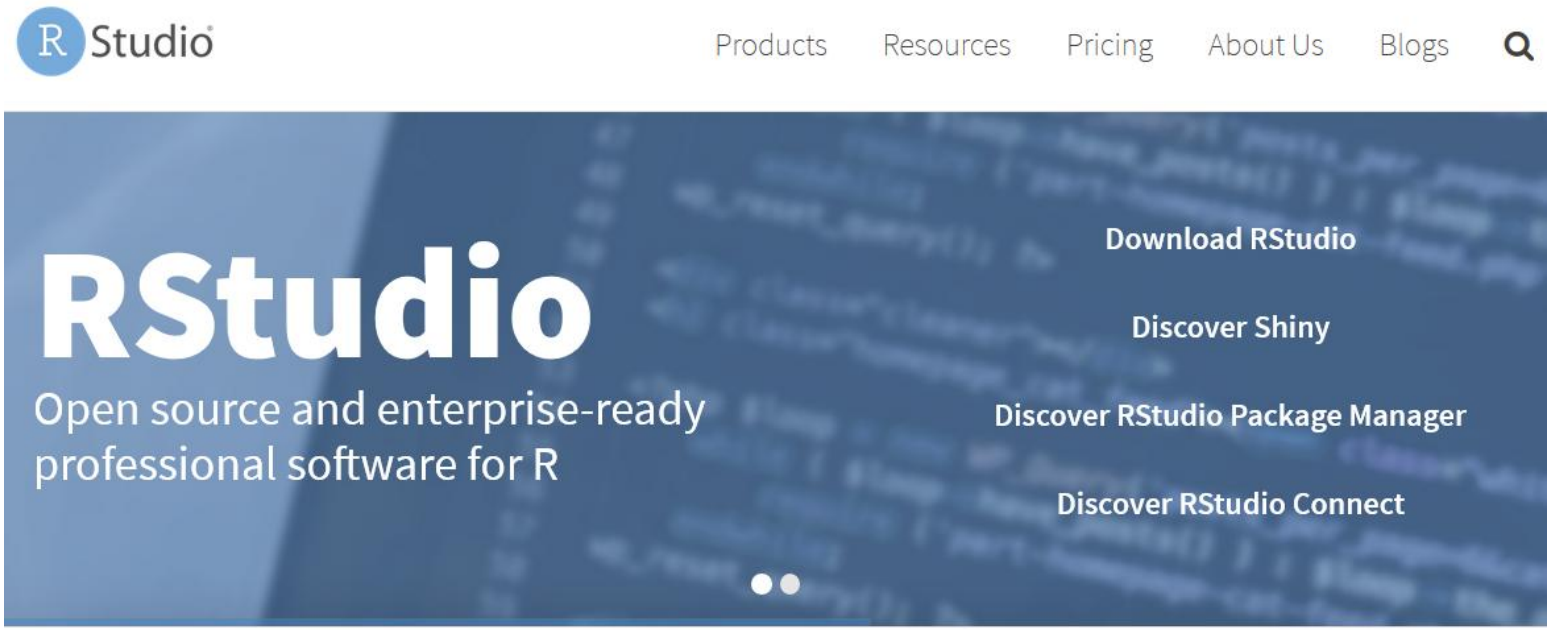
Precompiled binary distributions of the base system and contributed packages,
and Mac users most likely want one of these versions of R:

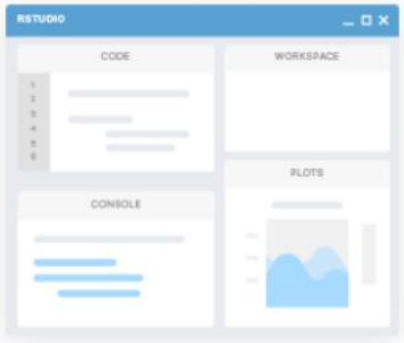
- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

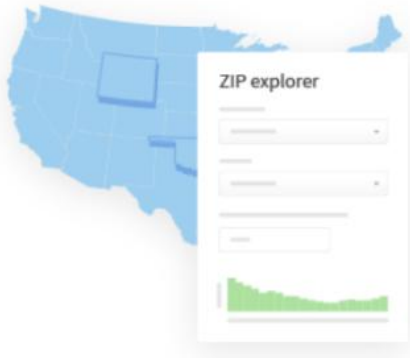
R is part of many Linux distributions, you should check with your Linux packa
management system in addition to the link above.

Rstudio

https://www.rstudio.com/







Packages

install.packages("nameOfThePackage")

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Contributed Packages

Available Packages

Currently, the CRAN package repository features 13664 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for infor
this repository. The manual R Installation and Administration (also containe
process in detail.

CRAN Task Views allow you to browse packages by topic and provide tools
special areas of interest. Currently, 39 views are available.

Package Check Results

All packages are tested regularly on machines running Debian GNU/Linux,

Tasks

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

CRAN Task Views

CRAN task views aim to provide some guidance which packages on CRAN are rele
topic. They give a brief overview of the included packages and can be automatically
The views are intended to have a sharp focus so that it is sufficiently clear which pa
excluded) - and they are *not* meant to endorse the "best" packages for a given task.

- To automatically install the views, the ctv package needs to be installed, e.g.,
  `install.packages("ctv")`
  and then the views can be installed via `install.views` or `update.views` (wh
  packages are not installed and up-to-date), e.g.,
  `ctv::install.views("Econometrics")`
  `ctv::update.views("Econometrics")`
- The task views are maintained by volunteers. You can help them by suggestir
  included in their task views. The contact e-mail addresses are listed on the in
- For general concerns regarding task views contact the ctv package maintaine

Topics

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
Databases	Databases with R
DifferentialEquations	Differential Equations
Distributions	Probability Distributions

Tasks

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

CRAN Task View: Machine Learning & Statistical Learning

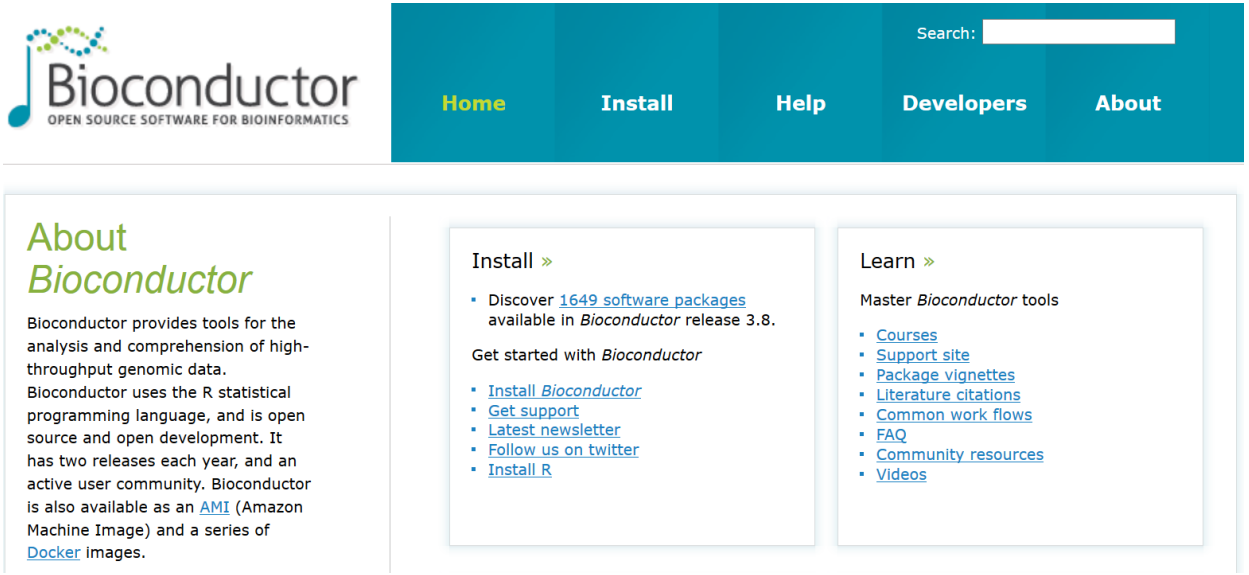
Maintainer: Torsten Hothorn
Contact: Torsten.Hothorn at R-project.org
Version: 2018-08-05
URL: https://CRAN.R-project.org/view=MachineLearning

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually referred to as machine learning. The packages can be roughly structured into the following topics:

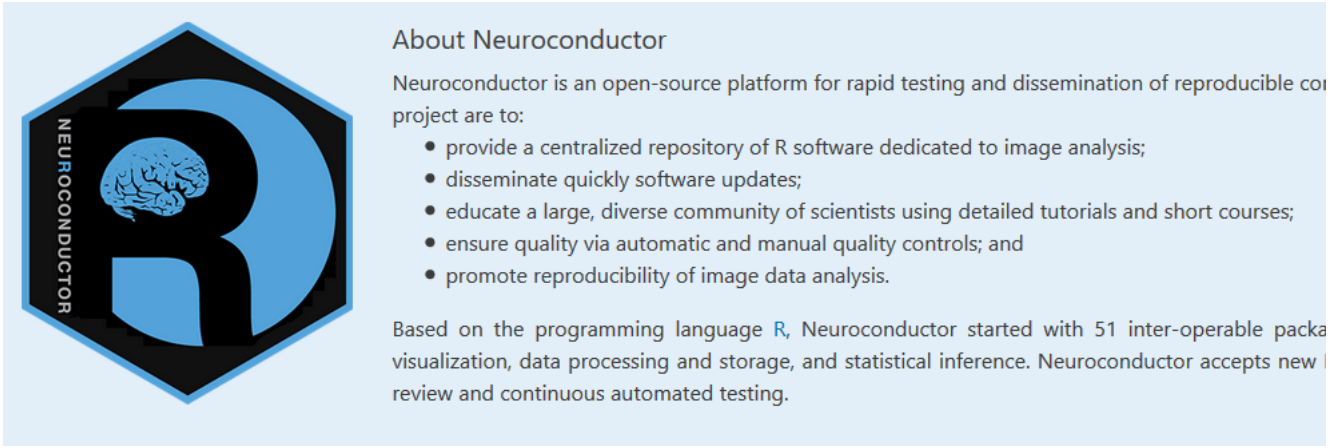
- *Neural Networks and Deep Learning* : Single-hidden-layer neural network are implemented in package nnet (shipped with base R). Package RSNNS offers an interface to the Stuttgart Neural Network Simulator (SNNS). rnn implements recurrent neural networks. Packages implementing deep learning flavours of neural networks include deepnet (feed-forward neural network, restricted Boltzmann machine, deep belief network, stacked autoencoders), RcppDL (denoising autoencoder, stacked denoising autoencoder, restricted Boltzmann machine, deep belief network) and h2o (feed-forward neural network, deep autoencoders). An interface to tensorflow is available in tensorflow.
- *Recursive Partitioning* : Tree-structured models for regression, classification and survival analysis, following the ideas in the CART book, are implemented in rpart (shipped with base R) and tree. Package rpart is recommended for computing CART-like trees. A rich toolbox of partitioning algorithms is available in Weka , package RWeka provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The

Parallel projects

Bioconductor: http://bioconductor.org



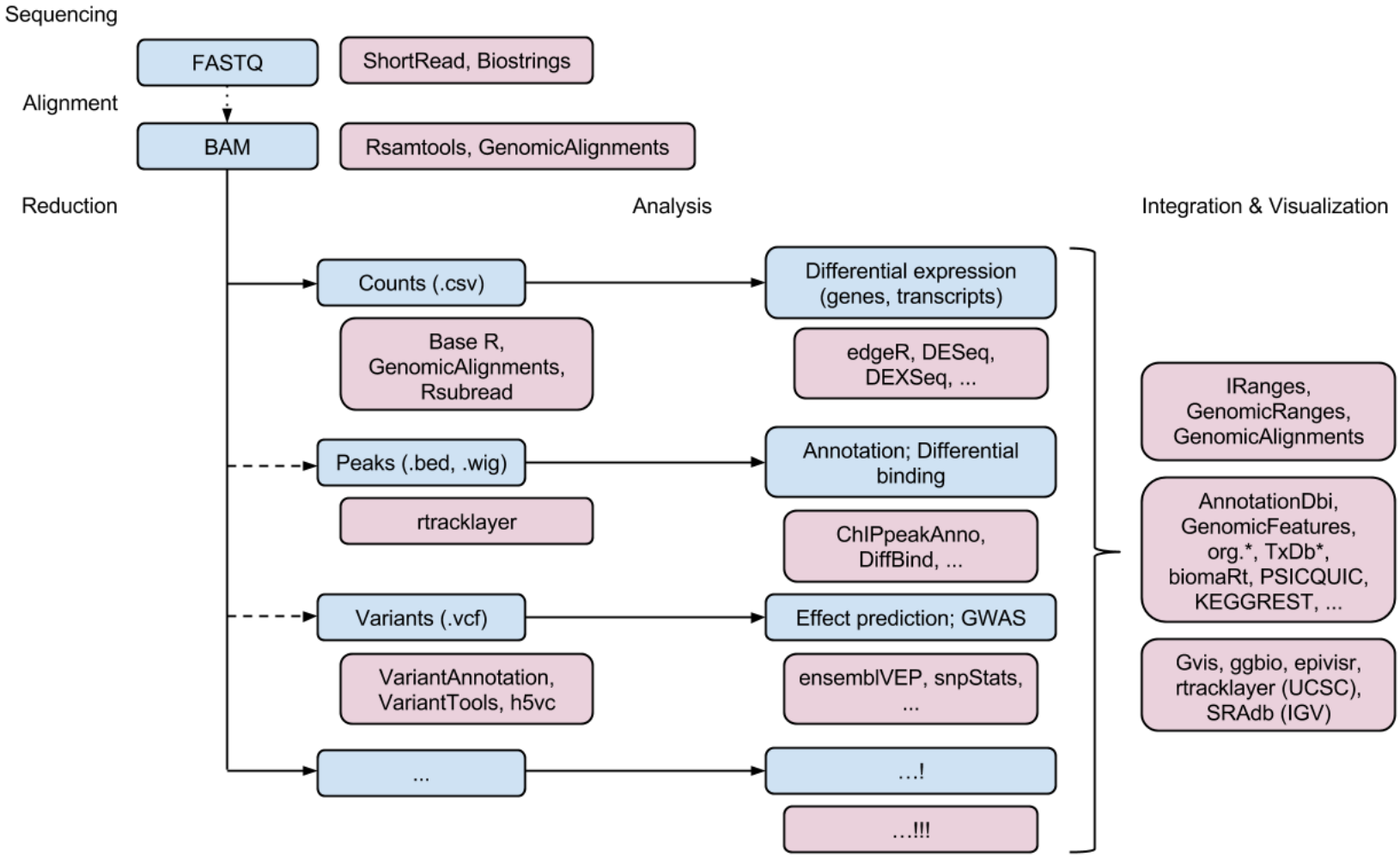
Neuroconductor: http://neuroconductor.org



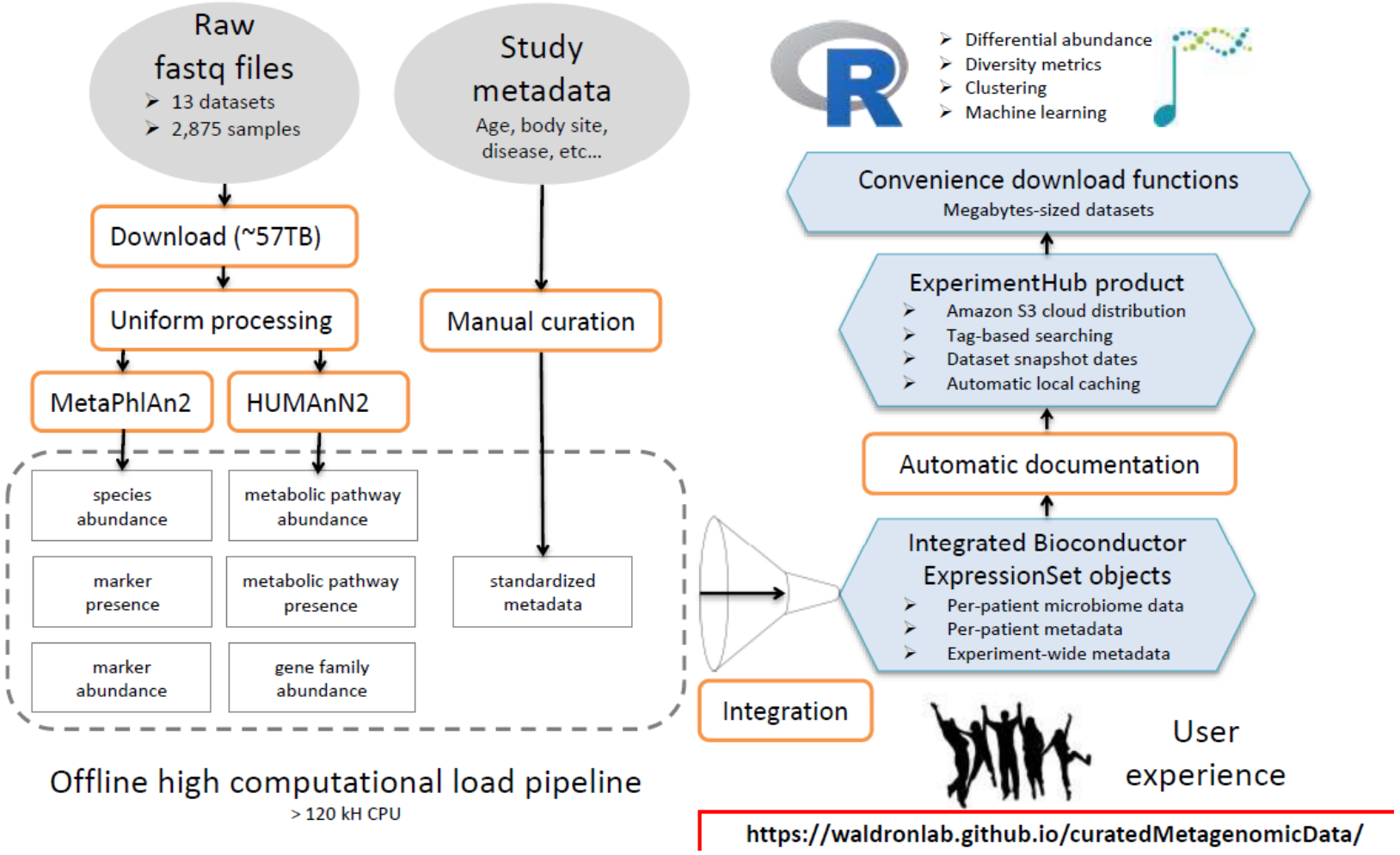
Bioconductor

- Provide access to powerful **statistical and graphical methods** for the analysis of **genomic** data.
- Facilitate the **integration of biological metadata** (GenBank, GO, LocusLink, PubMed) in the analysis of experimental data.
- Allow the rapid **development** of extensible, **interoperable**, and **scalable** software.
- Promote high-quality **documentation** and reproducible research.
- Provide **training** in computational and statistical methods

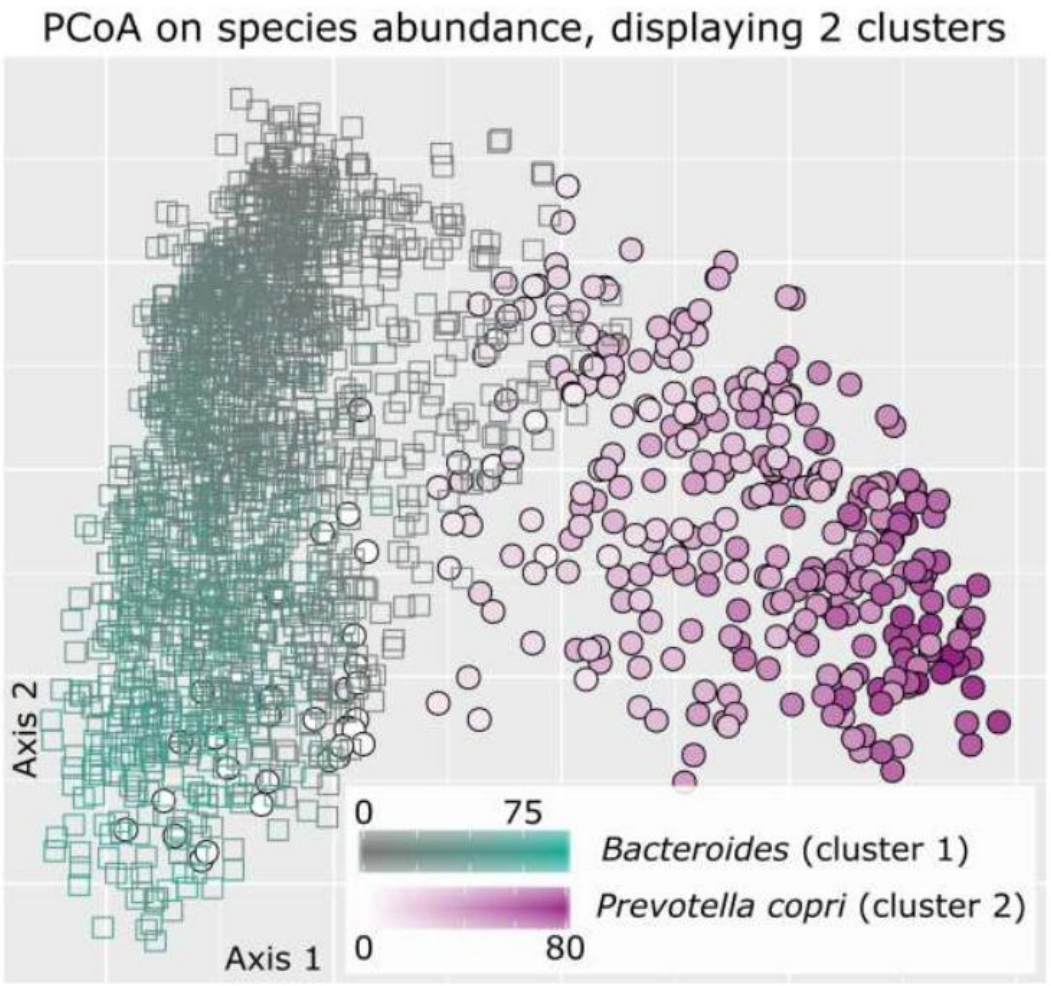
Bioconductor Ecosystem



Metagenomics



Metagenomics



Training

https://www.bioconductor.org/help/course-materials/

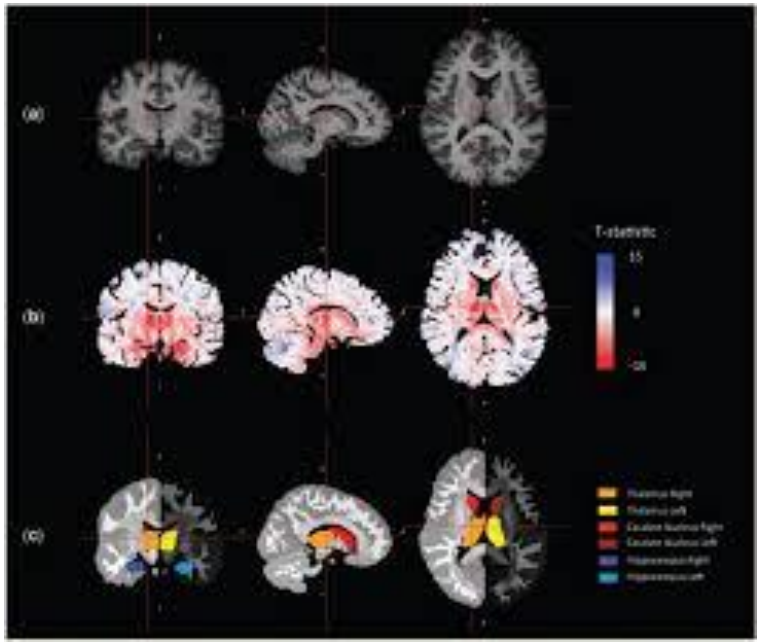
Keyword	Title	Course	Materials	Date	Bioc/R Version
Talk	Conducting Genomic Symphonies with *Bioconductor*, Michael Lawrence	BioInfoSummer17	pdf, zip	2017-12-4	3.6/3.4
Conference	Day 2 European Bioconductor Meeting 2017, Various authors	EuroBioc2017	YouTube	2017-12-5	3.6/3.4
Conference	Day 1 European Bioconductor Meeting 2017, Various authors	EuroBioc2017	YouTube	2017-12-5	3.6/3.4
Talk	File Management: BiocFileCache, AnnotationHub, ExperimentHub, Lori Shepherd	EuroBioc2017	Google Slides	2017-12-5	3.6/3.4
Talk	The *Bioconductor* Project: Current Status, Martin Morgan	EuroBioc2017	pdf, github	2017-12-5	3.6/3.4
Talk	The *Bioconductor* Project: Current Status, Martin Morgan	BiocAsia2017	pdf, github	2017-11-17	3.6/3.4
Devel	*Bioconductor* Masterclass: Package Development, Martin Morgan	BiocAsia2017	html, R, Rmd, github	2017-11-16	3.6/3.4
Intro	*Bioconductor* Masterclass: *Bioconductor* Essentials, Martin Morgan	BiocAsia2017	html, R, Rmd, github	2017-11-16	3.6/3.4
Intro / RNA-seq	*R* and *Bioconductor* for Genomic Analysis, Martin Morgan	OSU	Using R html, Rmd; Data Input and Manipulation html, Rmd; Statistics and Graphics html, Rmd; Introduction to	2017-09-11	3.6/3.4

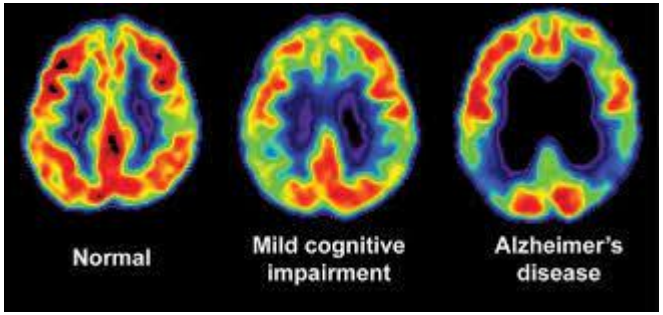
Neuroconductor

- Provide a centralized repository of R software dedicated to **image** analysis
- Disseminate quickly **software** updates
- **Educate** a large, diverse community of scientists using detailed tutorials and **short courses**
- Ensure quality via automatic and manual **quality controls**
- Promote **reproducibility** of image data analysis.

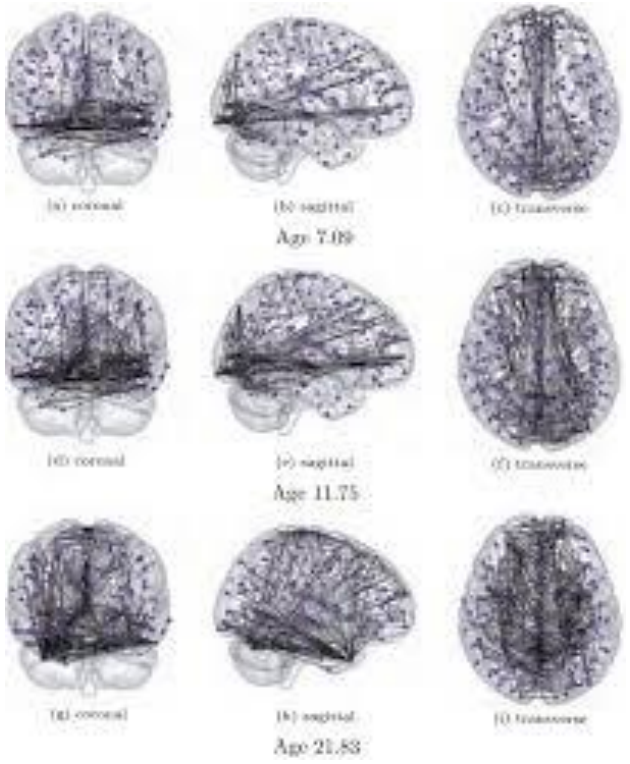
Neuroconductor

Functional imaging: volume activation





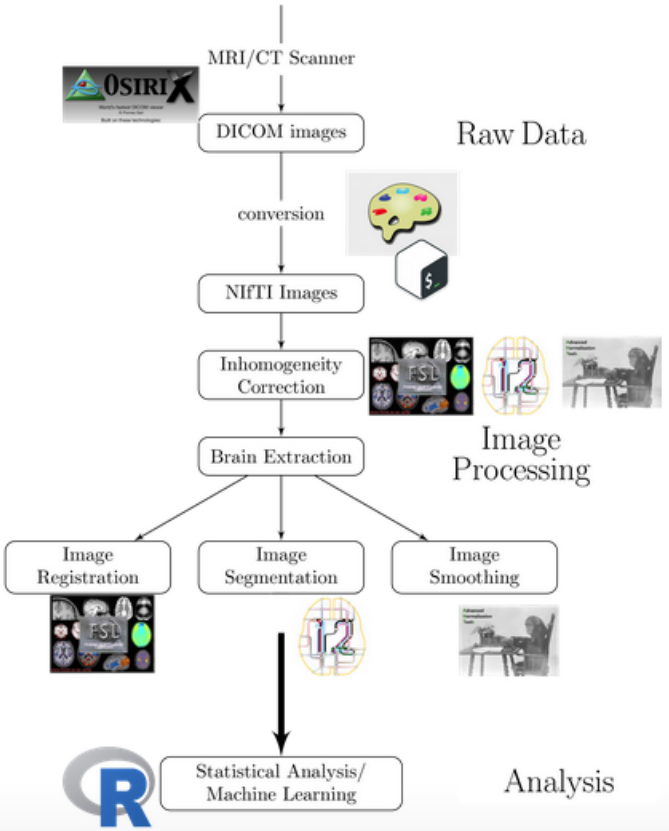
Neuron connections



Neuroconductor

Workflow for an Analysis

- bash
- FSL
- ANTs
- MRIcroGL
- OsiriX
- SPM 12



Neuroconductor

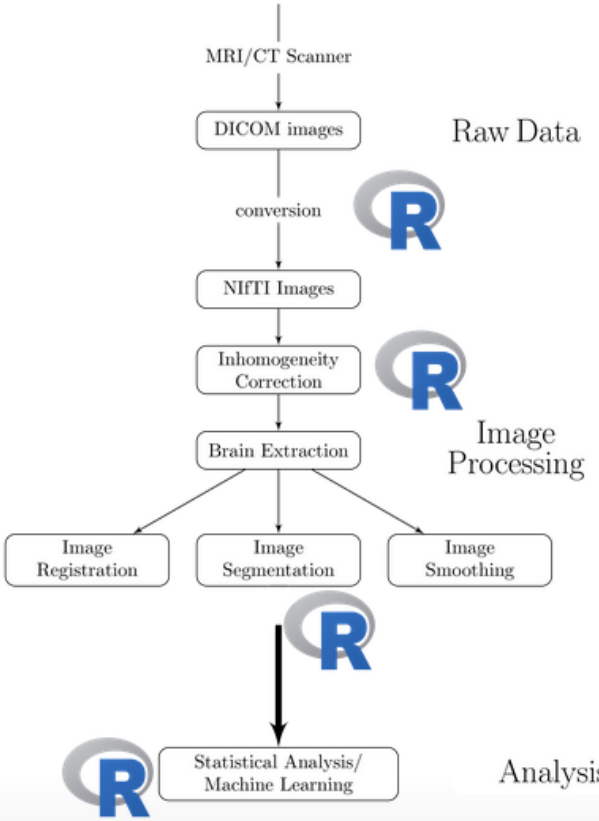
Goal:

Lower the bar to entry

- all R code
  - pipeline tool
  - "native" R code

Complete pipeline

- preprocessing and analysis



Training

Coursera Course:
Introduction to
Neurohacking In R
https://www.coursera.org/learn/neurohacking/

High performance computing

- for is not efficient: **apply** family
- C++, Fortran, … **RcppArmadillo**, **RcppEigen**, ….
- Parallel computing
- **Large memory** and **out-of-memory data:**

data.table, bigmemory, biglars, ff, ffbase

- Statistical methods for large datasets
- Integration with other languages
- Profiling tools: **microbenchmark**

Parallel computing

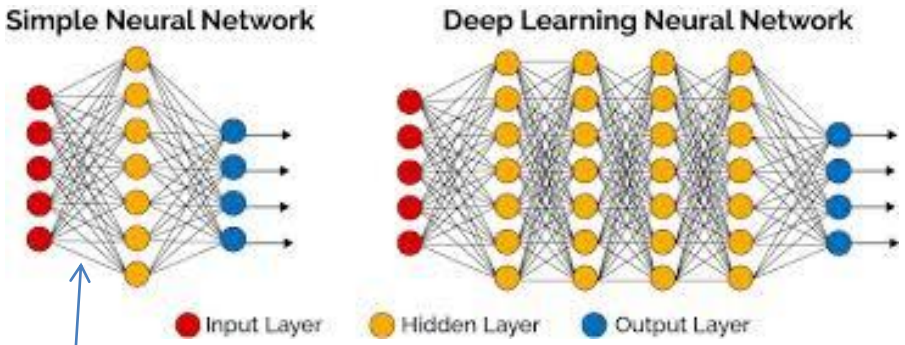
- **multicore, parallel**, **foreach** and **BiocParallel** functions distributes for loop to resident cores (high-level)
- **RcppParallel** parallize C++ code (low-level)
- **mclapply** applies any function to each element of a vector in parallel
- **h2o** facilitates machine learning (e.g. RFs, ANNs) in a parallel environment
- CRAN **HadoopStream** & hive serve **MapReduce** in Hadoop environment
- CRAN **cloudRmpi** serves MPI and
- Gputools, magma & OpenCl serve **GPU clusters**
- **RevoScaleR** integrates R with Microsoft SQL Server

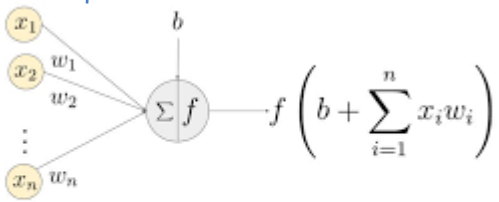
R package development

- Git **version control** system: store changes of files

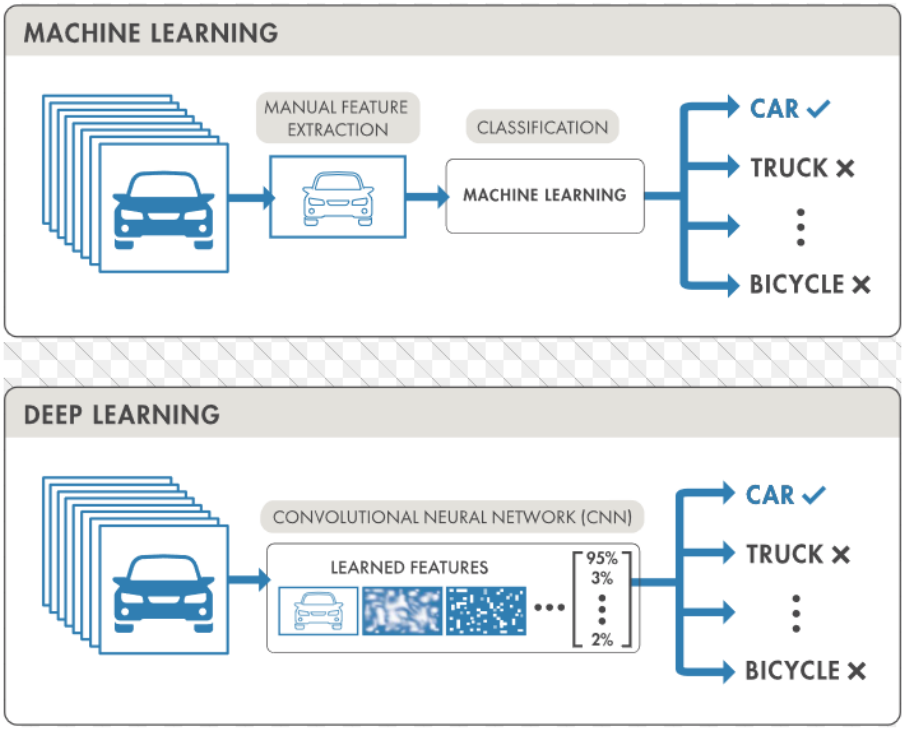
- GitHub is an **online** server of repositories
- **Distribute** packages and install them via `devtools::install_github`

R deep neural networks





An example of a neuron showing the input (x_1 - x_n), their corresponding weights (w_1 - w_n), a bias (b) and the activation function f applied to the weighted sum of the inputs.



R deep neural networks

nature genetics

Perspective | Published: 26 November 2018

A primer on deep learning in genomics

James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani & Amalio Telenti

Nature Genetics **51**, 12–18 (2019) | Download Citation

Abstract

Deep learning methods are a class of machine learning techniques capable of identifying highly complex patterns in large datasets. Here, we provide a perspective and primer on deep learning applications for genome analysis. We discuss successful applications in the fields of regulatory genomics, variant calling and pathogenicity scores. We include general guidance for how to effectively use deep learning methods as well as a practical guide to tools and resources. This primer

nature biotechnology

Perspective | Published: 06 September 2018

Deep learning in biomedicine

Michael Wainberg, Daniele Merico, Andrew Delong & Brendan J Frey

Nature Biotechnology **36**, 829–838 (2018) | Download Citation

Abstract

Deep learning is beginning to impact biological research and biomedical applications as a result of its ability to integrate vast datasets, learn arbitrarily complex relationships and incorporate existing knowledge.

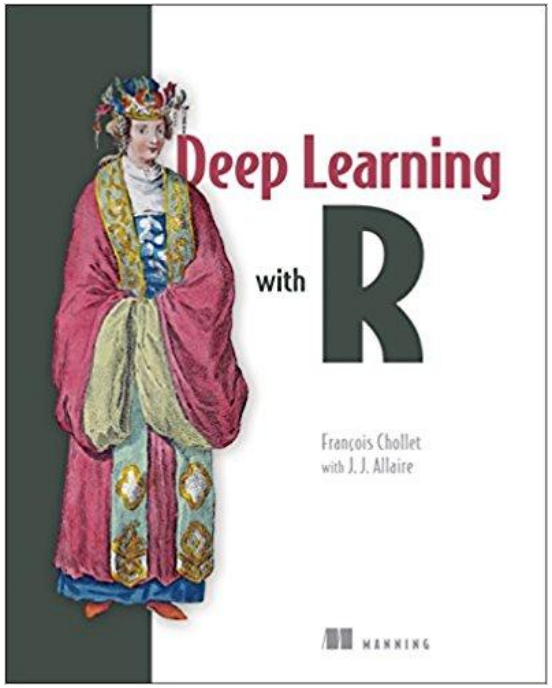
OPINION ARTICLE

Front. Neuroinform., 26 April 2018 | https://doi.org/10.3389/fninf.2018.00023

Deep Learning Methods to Process fMRI Data and Their Application in the Diagnosis of Cognitive Impairment: A Brief Overview and Our Opinion

Dong Wen[1,2], Zhenhao Wei[1,2], Yanhong Zhou[3], Guolin Li[4], Xu Zhang[1,2] and Wei Han[1,2*]

R deep neural networks

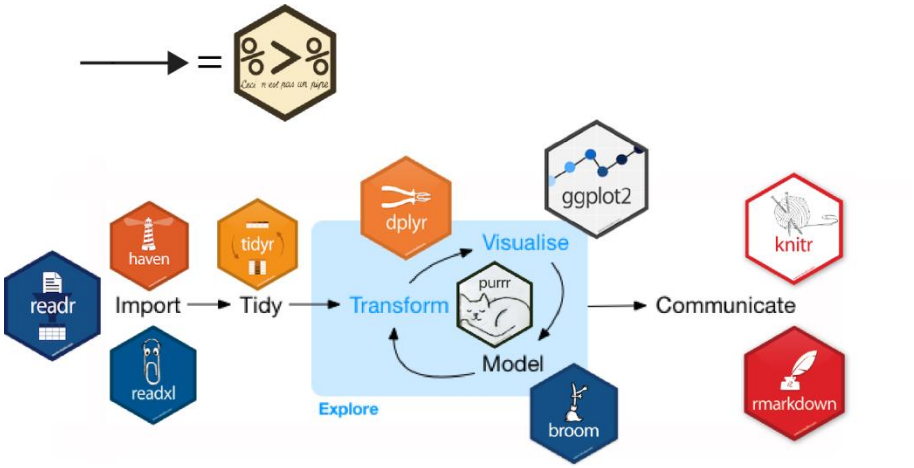


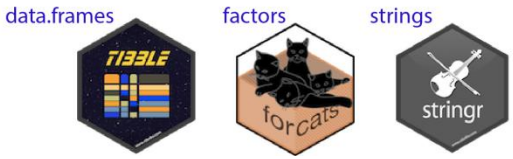
Introduction to basic concepts of machine learning
and deep learning:
Deep Learning with R in motion (https://bit.ly/2oPtXWv)

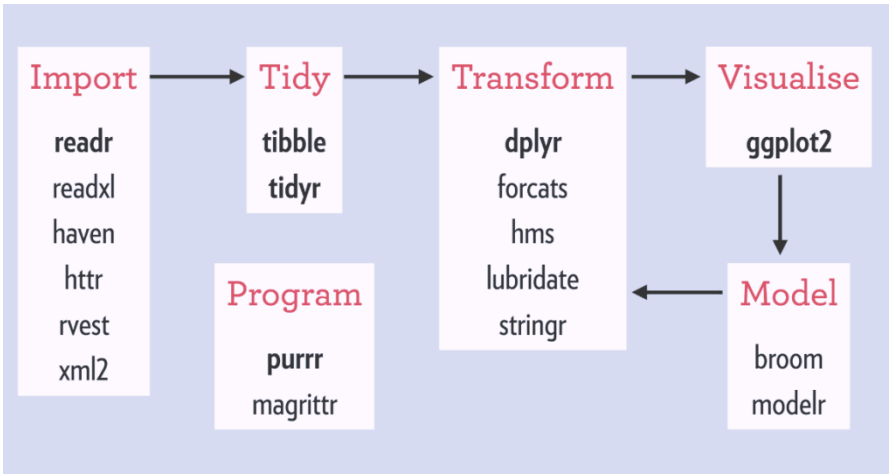
Excellent examples https://blogs.rstudio.com/tensorflow/

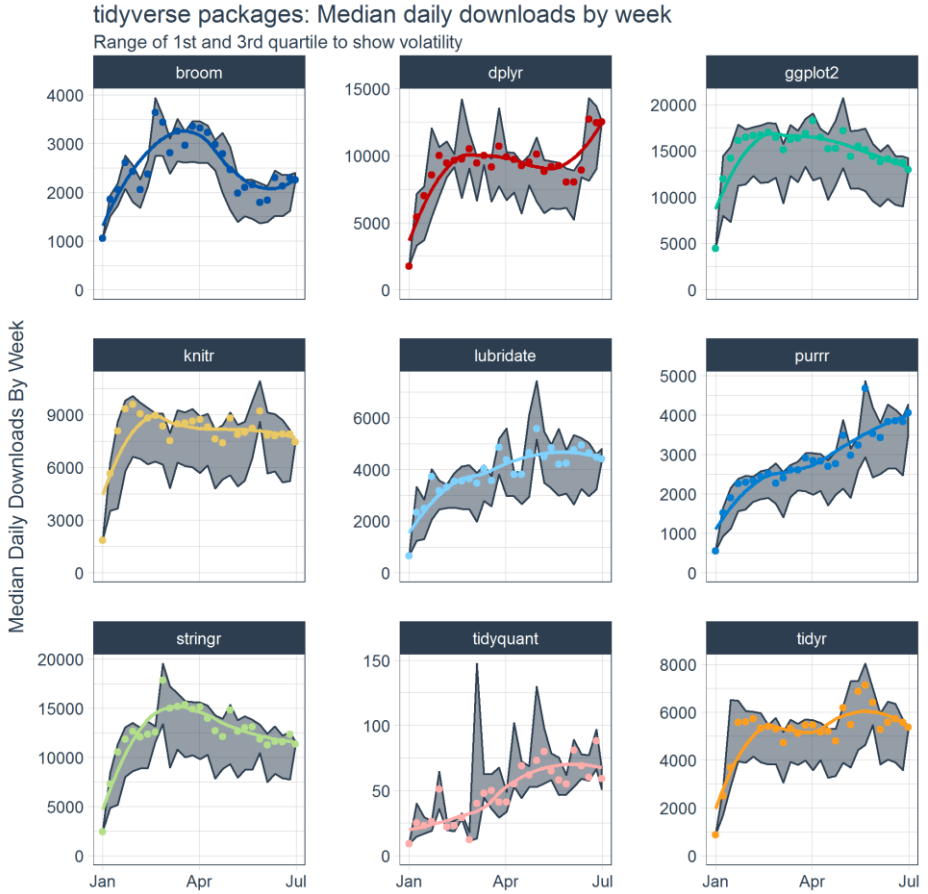
H20 project
https://www.h2o.ai/wp-content/uploads/2018/01/RBooklet.pdf

R tidyverse


readr
haven
tidyr
readxl
dplyr
ggplot2
knitr
rmarkdown
broom
purrr
Import
Tidy
Transform
Visualise
Model
Communicate
Explore


data.frames
factors
strings
tibble
forcats
stringr


Import
readr
readxl
haven
httr
rvest
xml2
Tidy
tibble
tidyr
Transform
dplyr
forcats
hms
lubridate
stringr
Visualise
ggplot2
Model
broom
modelr
Program
purrr
magrittr


tidyverse packages: Median daily downloads by week
Range of 1st and 3rd quartile to show volatility
Median Daily Downloads By Week
broom
dplyr
ggplot2
knitr
lubridate
purrr
stringr
tidyquant
tidyr
Jan
Apr
Jul

Course outline

- **January, 12:** R in biomedicine
- **January, 25 (3h):** R introduction (**data.table**, **mclapply, R markdown**)
- **February, 22 (2.5h):** R tidyverse
- **March, 1 (2.5h):** R tidyverse
- **March, 8 (2.5h):** Omic data: Bioconductor, GWAS
- **March, 15 (2.5h):** Omic data: GWAS (**lasso, H2O**)
- **March, 22 (4h):** Omic data: Multi-omic integration (**sPCA**, **GCCA**) + Visualization (Gviz)

Methodology: 30' talk + 30' exercises + Homework