

R Computing services as SaaS in the Cloud

Fernando Aguilar

aguilarf@ifca.unican.es

Instituto de Física de Cantabria (IFCA)

Santander - Spain

Introduction

- ✿ R is a programming language and software environment for statistical computing and graphics, widely used among statisticians and data miners for developing statistical software and data analysis.
- ✿ R is a very useful language for those researchers that need to analyse data and are not IT experts.
- ✿ R community is pretty wide, so there are a number of plugins and modules to enrich the use.
- ✿ R is open.



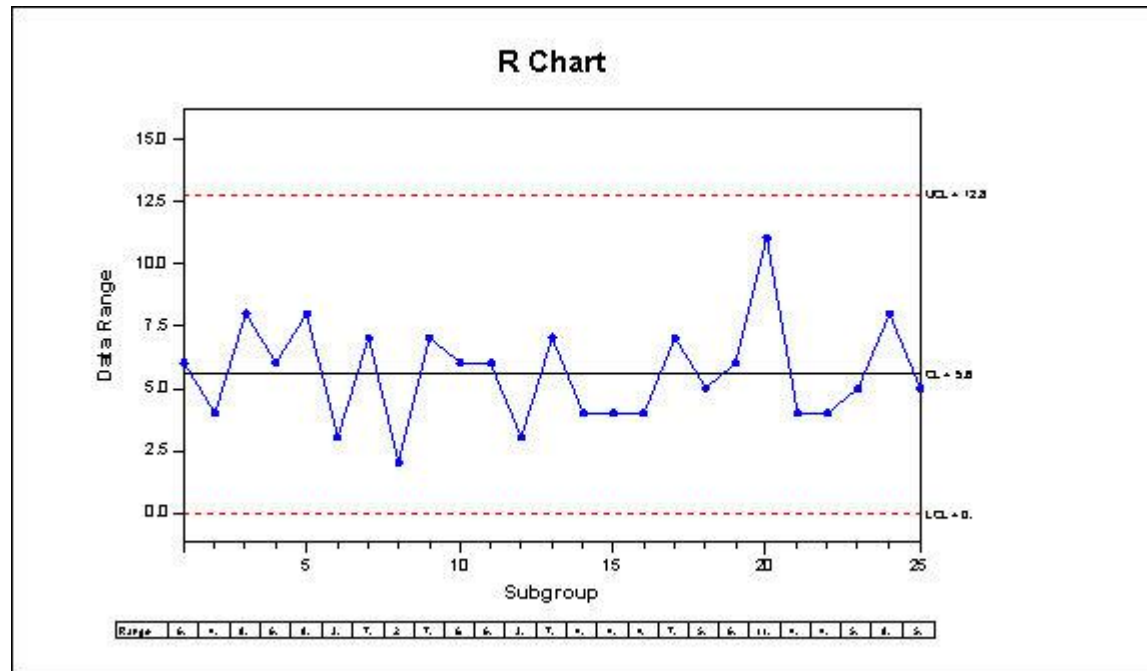
Introduction - Why R?

- ✿ Easy for beginners. Powerful for experts (integration with others, data sources, etc.)
- ✿ Thousands of packages.
- ✿ Explicit parallelism is straightforward in R.
- ✿ Growing community of users.

use **R**!

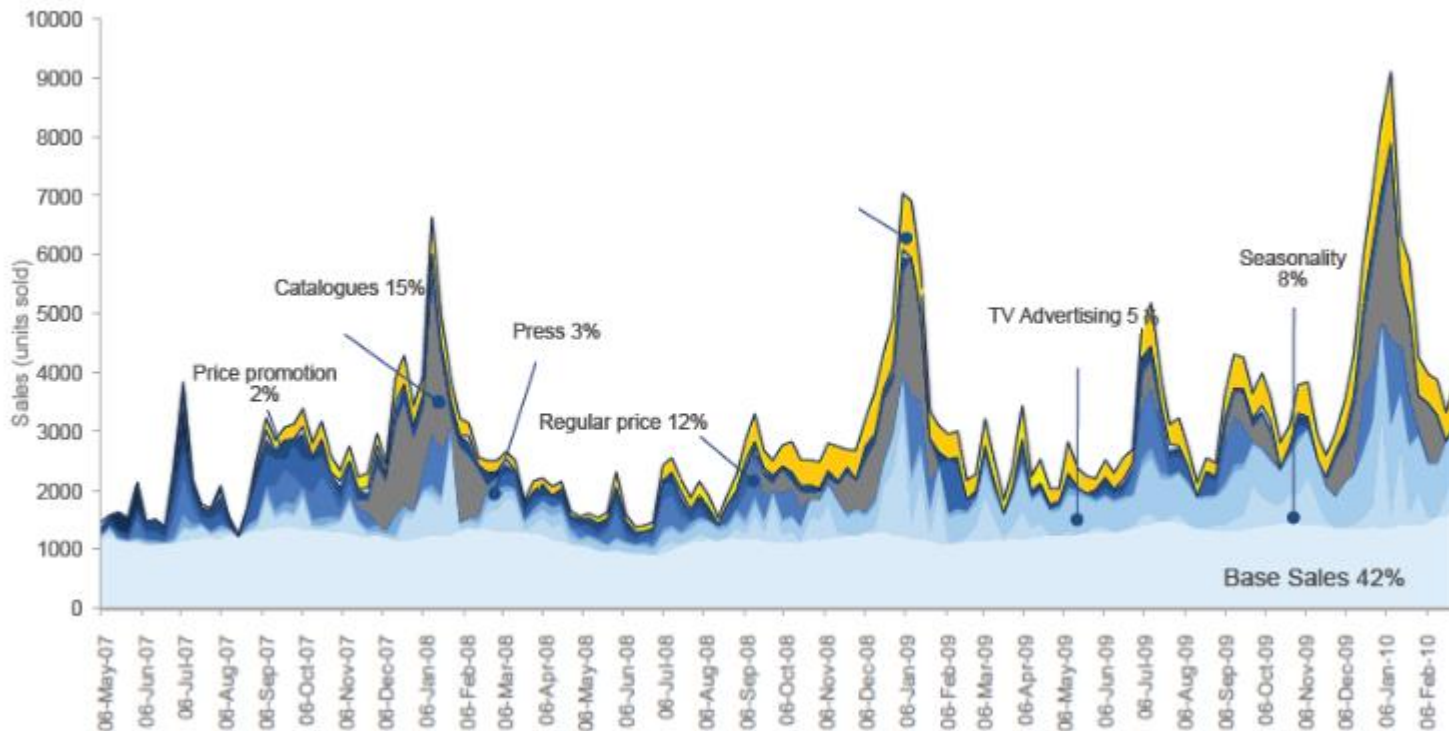
Introduction - Use of R

- Fields: data exploitation, data analysis, data mining, etc.
- From basic to complex:



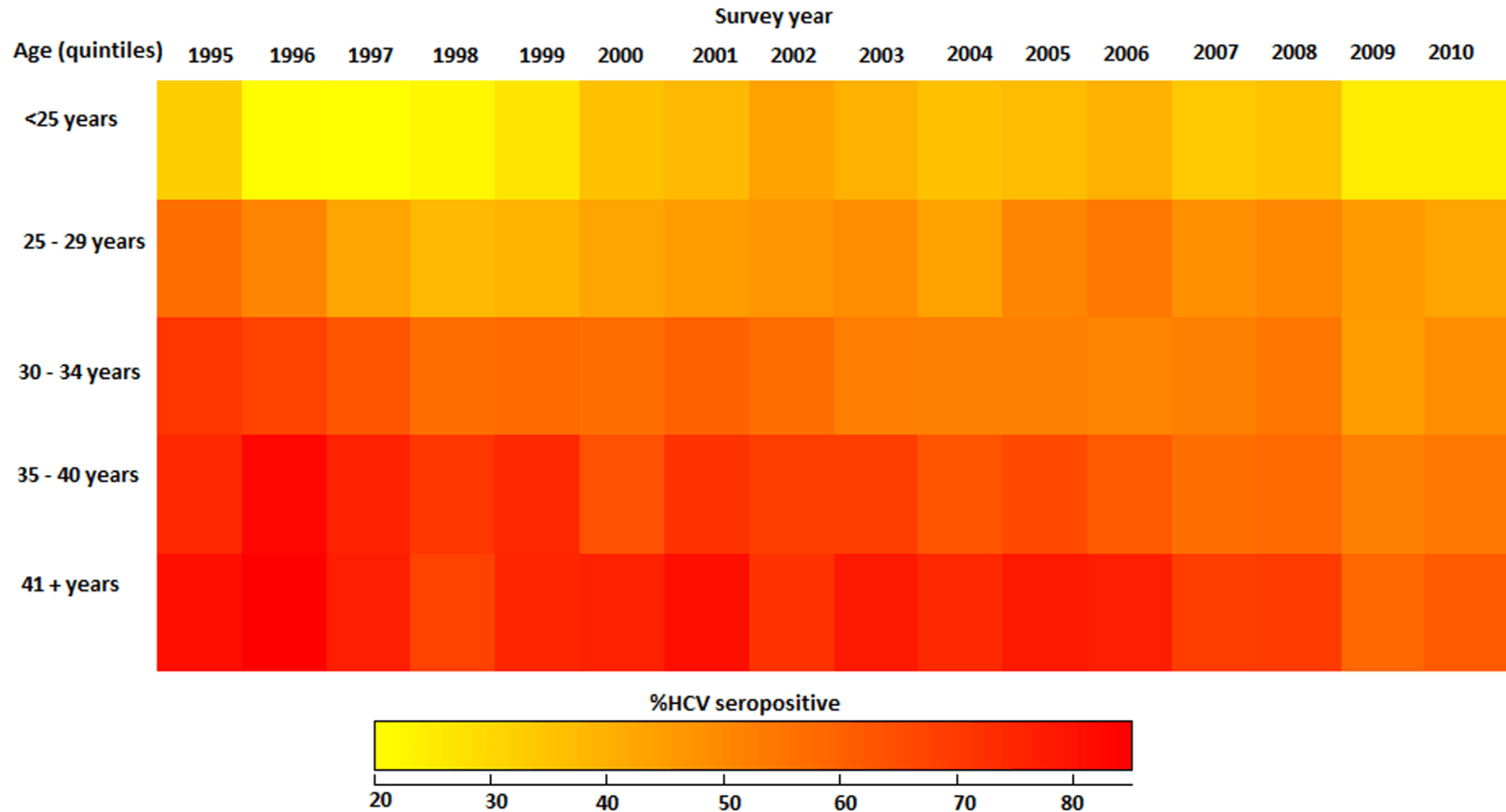
Introduction - Use of R

- ✪ Different sources, formats (DB, csv, etc.).



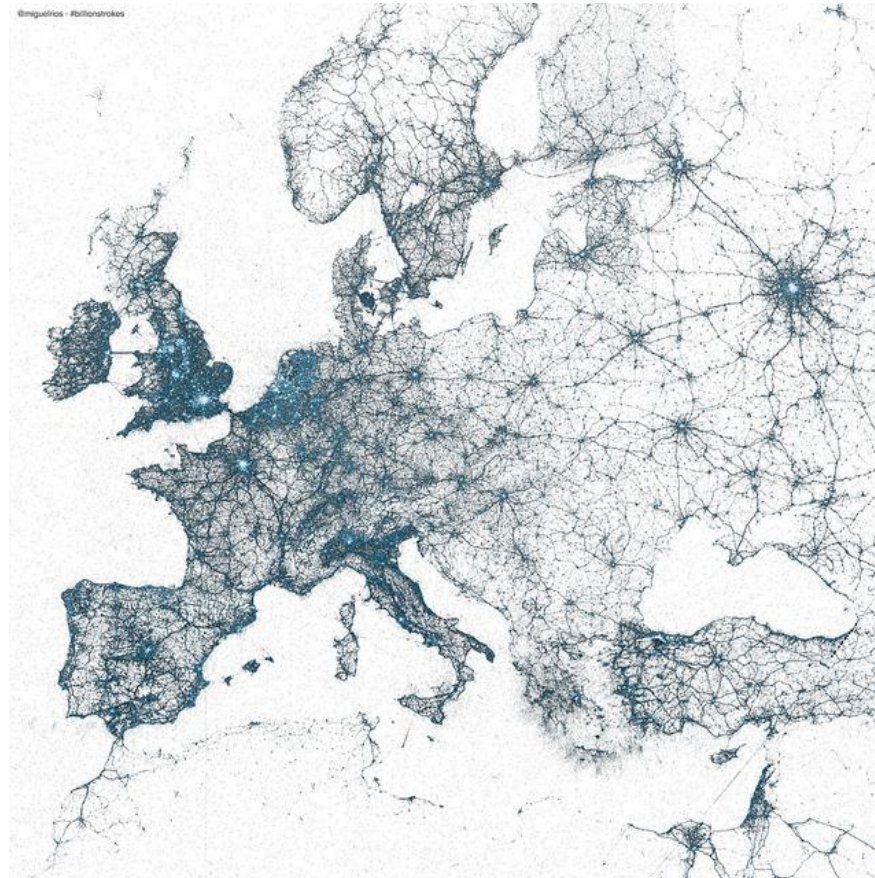
Introduction - Use of R

- More complex charts, more than one-two params.



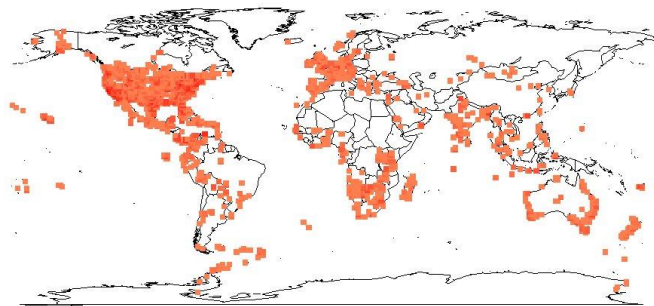
Introduction - Use of R

- 📍 Geo Packages: maps, google, GIS connection.



Introduction -R in Biodiversity

- ⊕ Species distribution (geo formats)
- ⊕ Data analysis
- ⊕ Data filtering
- ⊕ Satellite data
- ⊕ Image analysis



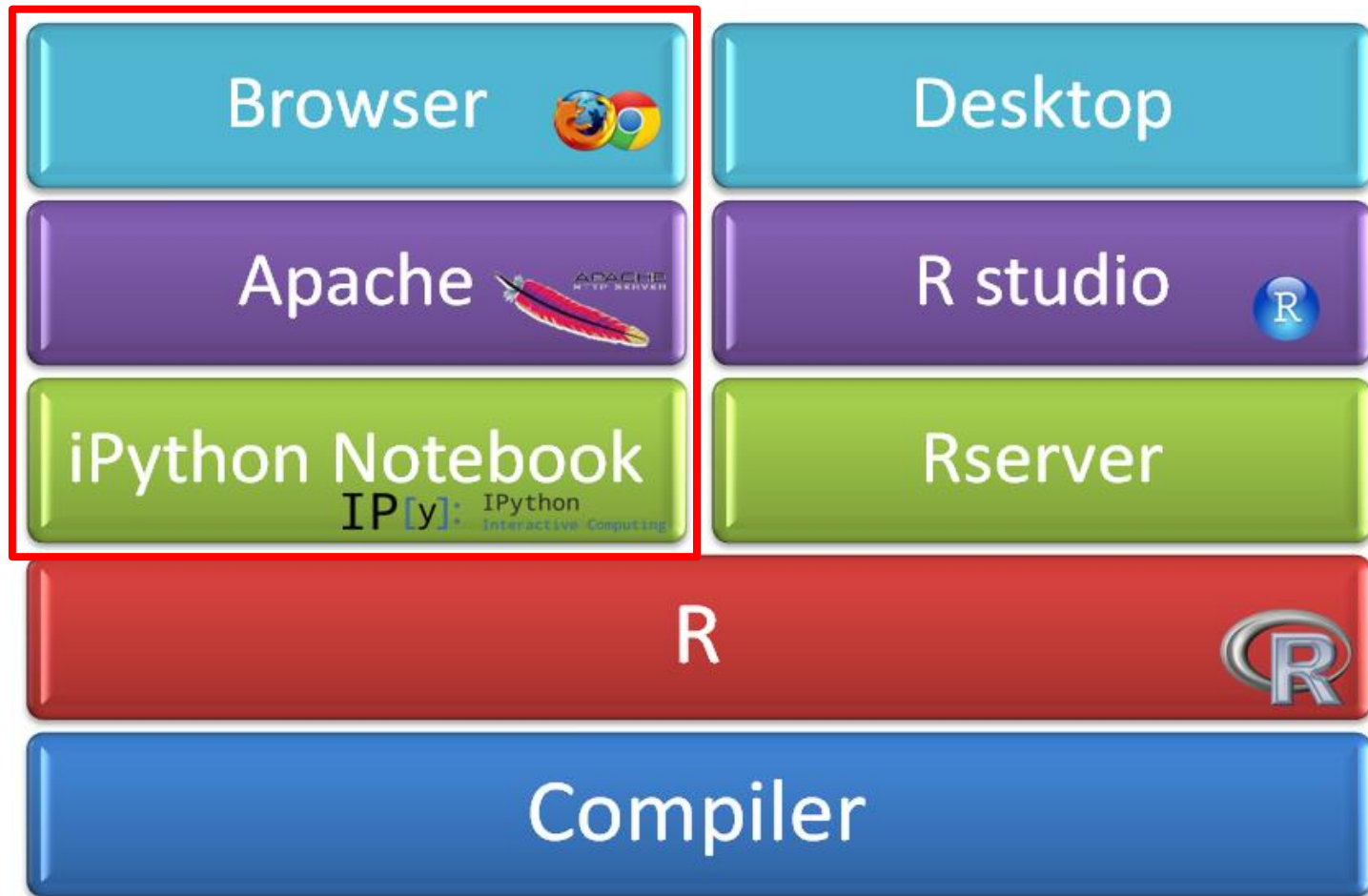
Lifewatch

- ❖ LifeWatch is the European e-Science infrastructure for biodiversity and ecosystem research. ESFRI
- ❖ Aims to provide advanced capabilities for research on the complex biodiversity system.
- ❖ e-Science infrastructures capitalize existing resources and data from physical infrastructures, distributed centers and single research groups.
- ❖ The capabilities offered by LifeWatch, as a e-Science infrastructure, allow users to tackle the big basic questions in biodiversity, as well to address the urgent societal challenges concerning biodiversity, ecosystems and other crosscutting issues.

EGL-Lifewatch Competence Centre

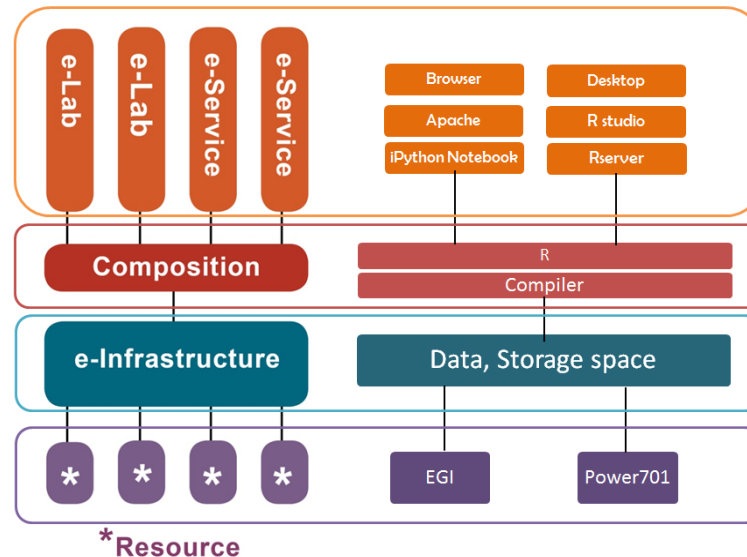
- ✿ Support the requirements of the community in Biodiversity and Ecosystems research.
- ✿ Establishing a direct collaboration between EGI.eu and the ESFRI LifeWatch to address specific needs.
- ✿ Four Mini-Projects:
 - ✦ Exploitation of the EGI infrastructure by the LifeWatch user community.
 - ✦ Tools required to support data management, data processing and modeling.
 - ✦ Integrate in EGI FedCloud framework, workflows, Vlabs.
 - ✦ Citizen Science in EGI e-infrastructure.
- ✿ Working Groups: R

Architecture



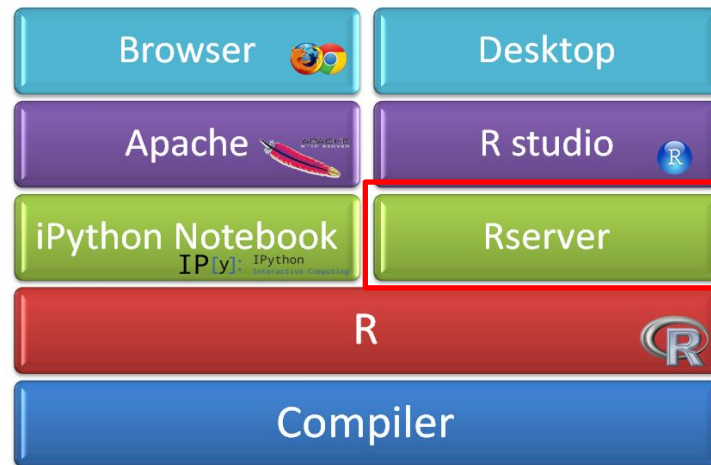
Architecture - Bottom Layer

- ✿ R instances, compiler, computing layer.
- ✿ Different choices:
 - ✦ HPC: Power701, improve performing.
 - ✦ Cloud: Load balance, different R version (package dependant), container.



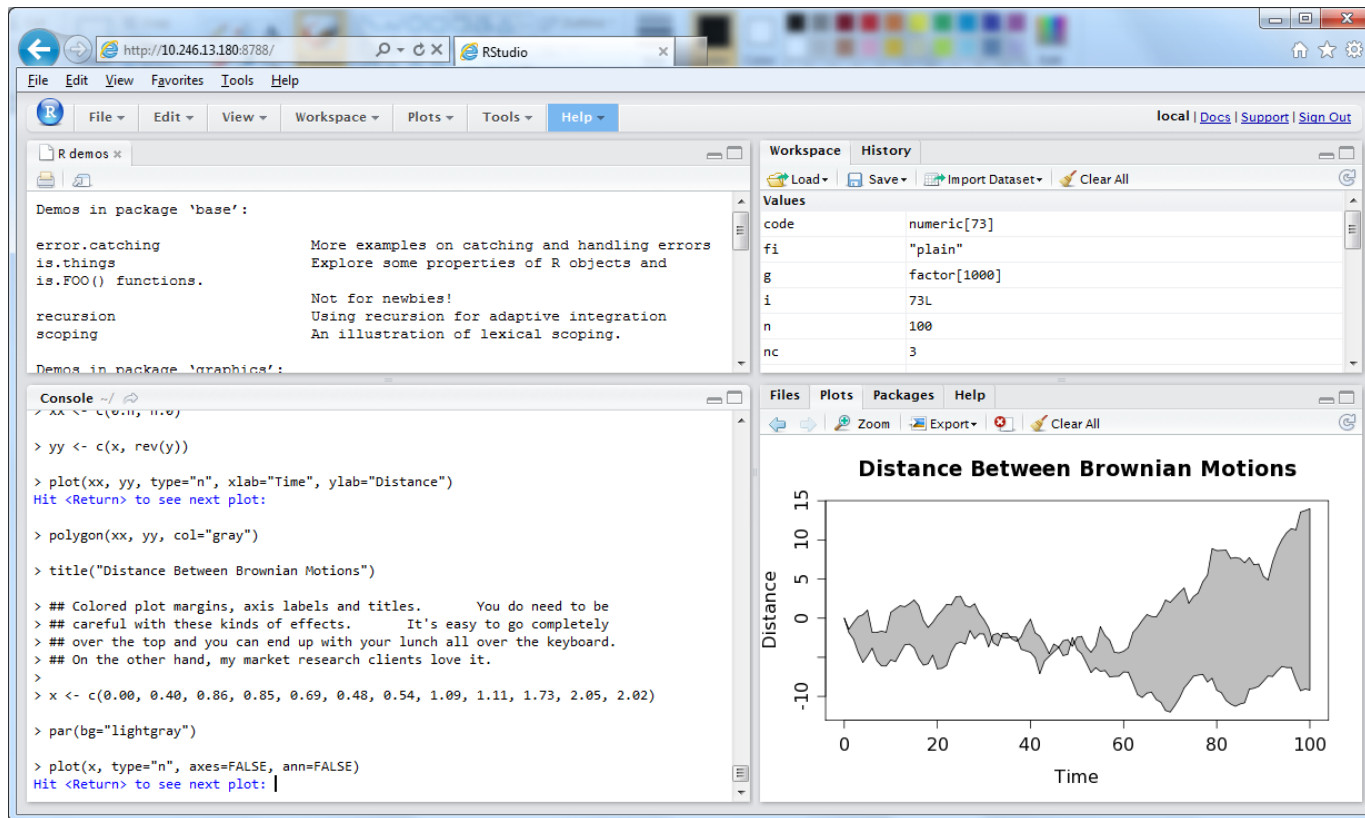
Architecture - Medium Layer

- ✿ R server - Interface between Computing and GUI.
- ✿ Not always needed.
- ✿ Client contact with server, that is an R package that connects both.
- ✿ Needs a desktop client.



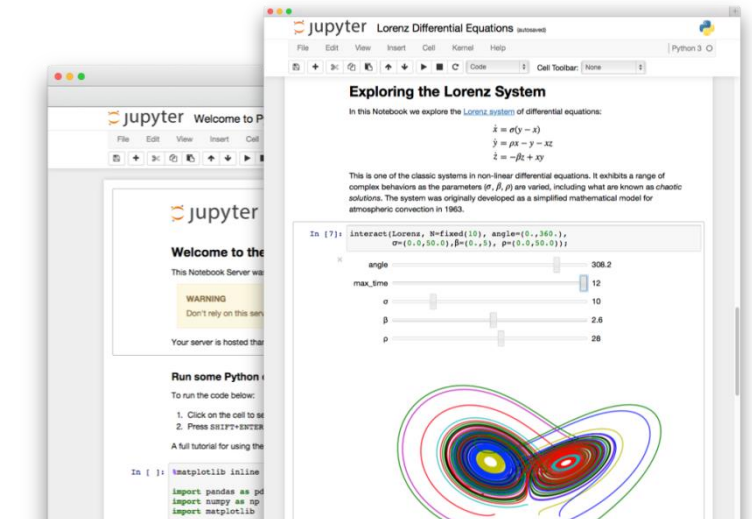
Architecture - SaaS

Rstudio Server



Architecture - SaaS

- ❁ The **Jupyter Notebook** is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.
- ❁ Over 40 languages.
- ❁ Sharable notebooks.
- ❁ Jupyter hub.



Architecture - SaaS

Jupyter Untitled0 Last Checkpoint: 05/26/2015 (autosaved)



File Edit View Insert Cell Kernel Help

Python 2

Code Cell Toolbar: None

```
In [1]: %load_ext rpy2.ipynon
```

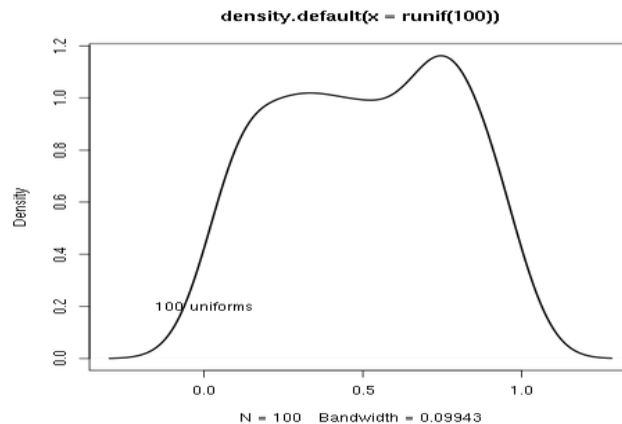
```
In [5]: %R getwd()
```

```
Out[5]: array(['/home/ipython/notebooks'],  
            dtype='|S21')
```

```
In [4]: %R library(parallel); detectCores(); system("grep MemFree /proc/meminfo", intern = TRUE)
```

```
Out[4]: array(['MemFree:      28163456 kB'],  
            dtype='|S27')
```

```
In [10]: # Let me take some pains on the 1st  
%R plot(density(runif(100)), lwd=2); text(x=0, y=0.2, "100 uniforms") # Showing you how to place text at will  
%R x=seq(0.01,1,0.01);par(col="blue") # default colour to blue.  
%R plot(x, sin(x), type="l"); lines(x, cos(x), type="l", col="red")
```



Architecture - SaaS - Exploitation

- ⊕ Workflows: R, python scripts, OpenShift.
- ⊕ User space.
- ⊕ Lifewatch Data Portal: dataset usage, experiment reanalysis, reproducibility.
- ⊕ Not only R, but other languages.

- Ecological/biological interest
- Offer statistical and visualization tools for LifeWatch Project, using R statistical language.

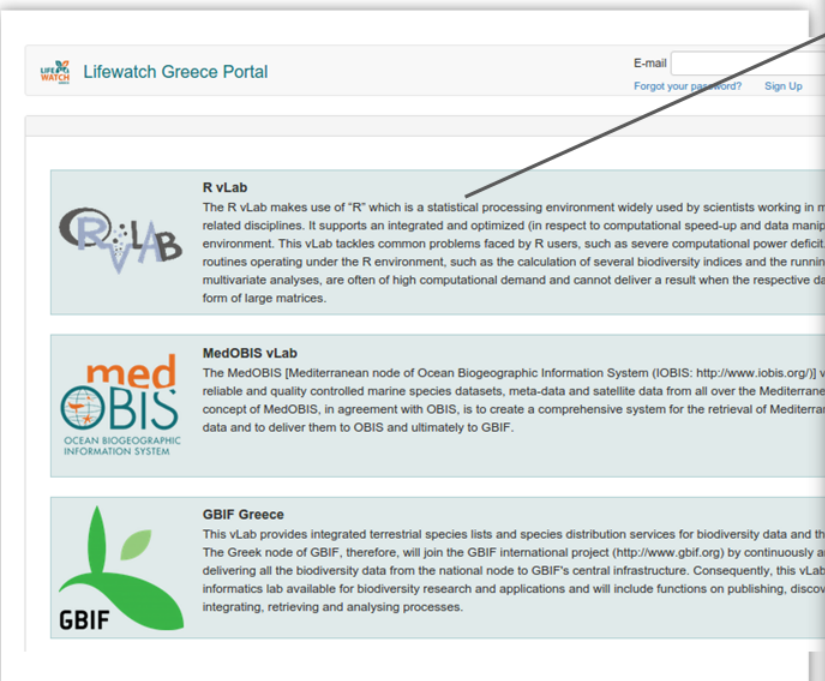
Main Objective:

- Optimize certain VEGAN package functions (Community Ecology Package), which supports, ordination methods, diversity analysis and other functions for community and vegetation ecologists.

More specific issues addressed are:

1. Big data manipulation (overcome memory barriers)
 2. Computational time speed-up (task segmentation multi-cores, cluster computing environment at HCMR – recently upgraded from Lifewatch)
- Develop an efficient and friendly user interface for analysis of ecological community data.

portal.lifewatchgreece.eu



Lifewatch Greece Portal

E-mail

[Forgot your password?](#) [Sign Up](#)

R vLab

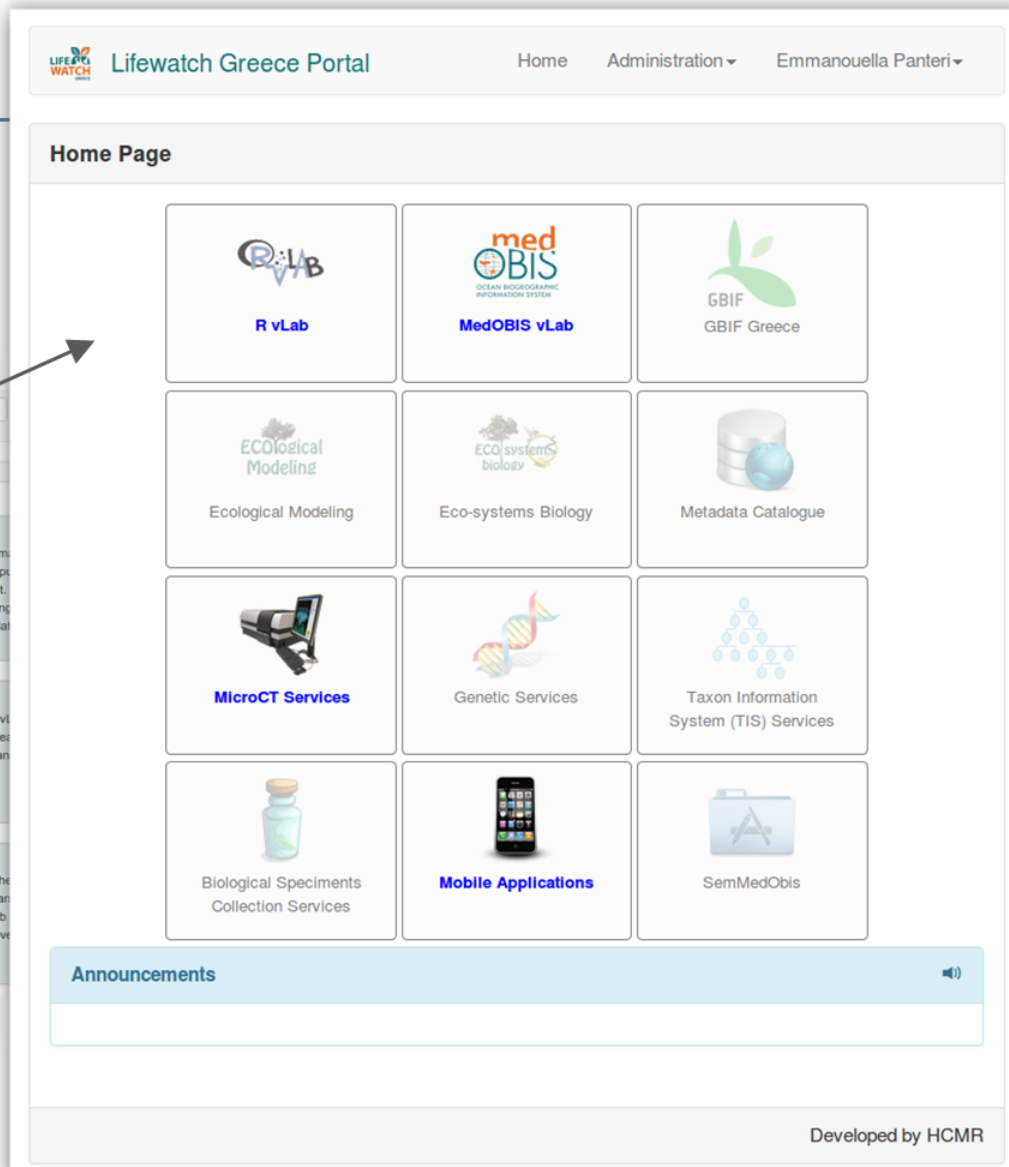
The R vLab makes use of "R" which is a statistical processing environment widely used by scientists working in many related disciplines. It supports an integrated and optimized (in respect to computational speed-up and data manipulation) environment. This vLab tackles common problems faced by R users, such as severe computational power deficit, routines operating under the R environment, such as the calculation of several biodiversity indices and the running of multivariate analyses, are often of high computational demand and cannot deliver a result when the respective data are in the form of large matrices.

medOBIS

The MedOBIS (Mediterranean node of Ocean Biogeographic Information System (OBIS: <http://www.obis.org/>)) vLab provides reliable and quality controlled marine species datasets, meta-data and satellite data from all over the Mediterranean. The concept of MedOBIS, in agreement with OBIS, is to create a comprehensive system for the retrieval of Mediterranean data and to deliver them to OBIS and ultimately to GBIF.













GBIF Greece


This vLab provides integrated terrestrial species lists and species distribution services for biodiversity data and the Greek node of GBIF, therefore, will join the GBIF international project (<http://www.gbif.org>) by continuously and delivering all the biodiversity data from the national node to GBIF's central infrastructure. Consequently, this vLab provides informatics lab available for biodiversity research and applications and will include functions on publishing, discovering, integrating, retrieving and analysing processes.




Lifewatch Greece Portal Home Administration ▾ Emmanouella Panteri ▾

Home Page

 R vLab	 MedOBIS vLab	 GBIF GBIF Greece
 Ecological Modeling	 Eco-systems Biology	 Metadata Catalogue
 MicroCT Services	 Genetic Services	 Taxon Information System (TIS) Services
 Biological Specimens Collection Services	 Mobile Applications	 SemMedObs

Announcements 

Developed by HCMR


R vLab

Workspace File Management

Available input files:

- softlagoonabundance.csv ✖
- softLagoonAbundance.csv ✖
- softlagoonaggregation.csv ✖
- softLagoonAggregation.csv ✖
- softlagoonenv.csv ✖
- softLagoonEnv.csv ✖
- softlagoonfactors.csv ✖
- softLagoonFactors.csv ✖

Upload new input files:

User's Storage Utilization: (396.00 KB)

0.0%

Recent Jobs:

Job ID	Function	Status	Submitted At	
Job312	taxa2dist	Completed	2015-08-22 21:04:25	✖
Job339	taxondive	Failed	2015-08-31 14:21:39	✖
Job340	vegdist	Completed	2015-08-31 14:21:55	✖
Job341	taxa2dist	Completed	2015-08-31 14:22:13	✖
Job342	anova	Failed	2015-08-31 14:22:39	✖
Job344	taxondive	Failed	2015-08-31 14:52:47	✖

Help

Submit a new Job

Statistical Function taxa2dist

Input files

Select classification table with a row for each species or other basic taxon, and columns for identifiers of its classification at higher levels from loaded files


- ☐ softlagoonabundance.csv
- ☐ softLagoonAbundance.csv
- ☐ softlagoonaggregation.csv
- ☐ softLagoonAggregation.csv
- ☐ softlagoonenv.csv
- ☐ softLagoonEnv.csv
- ☐ softlagoonfactors.csv
- ☐ softLagoonFactors.csv

Parameters

varstep FALSE

check TRUE

Developed by HCMR


R vLab

Workspace File Management

Available input files:

- softlagoonabundance.csv ✖
- softLagoonAbundance.csv ✖
- softlagoonaggregation.csv ✖
- softLagoonAggregation.csv ✖
- softlagoonenv.csv ✖
- softLagoonEnv.csv ✖
- softlagoonfactors.csv ✖
- softLagoonFactors.csv ✖

Upload new input files:

User's Storage Utilization: (396.00 KB)

0.0%

Help

Submit a new Job

Statistical Function taxa2dist

Input files

Select classification table with a row for each species or other basic taxon, and columns for identifiers of its classification at higher levels from loaded files

- ☐ softlagoonabundance.csv
- ☐ softLagoonAbundance.csv
- ☐ softlagoonaggregation.csv
- ☐ softLagoonAggregation.csv
- ☐ softlagoonenv.csv
- ☐ softLagoonEnv.csv
- ☐ softlagoonfactors.csv
- ☐ softLagoonFactors.csv

Parameters

varstep FALSE

check TRUE

Recent Jobs:

Job ID	Function	Status	Submitted At	
Job312	taxa2dist	Completed	2015-08-22 21:04:25	✖
Job339	taxondive	Failed	2015-08-31 14:21:39	✖
Job340	vegdist	Completed	2015-08-31 14:21:55	✖
Job341	taxa2dist	Completed	2015-08-31 14:22:13	✖
Job342	anova	Failed	2015-08-31 14:22:39	✖
Job344	taxondive	Failed	2015-08-31 14:52:47	✖

Developed by HCMR

Function documentation

Conclusions

- ✿ R is one of the best language for data analysis, managing, etc. For experts and non-experts.
- ✿ SaaS is the best approach for non IT researchers. Biodiversity.
- ✿ SaaS solutions that can explote FedCloud resources:
 - ✦ R oriented - RStudioServer
 - ✦ Jupyter - More open, more functionalities
 - ✦ LFW Greece VLab
- ✿ Lifewatch Open Science Framework integrates preservation of the whole data lifecycle with jupyter to provide user a complete environment for data managing.

Thanks for your attention



Fernando Aguilar
aguilarf@ifca.unican.es
Instituto de Física de Cantabria (IFCA)
Santander - Spain

