

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.



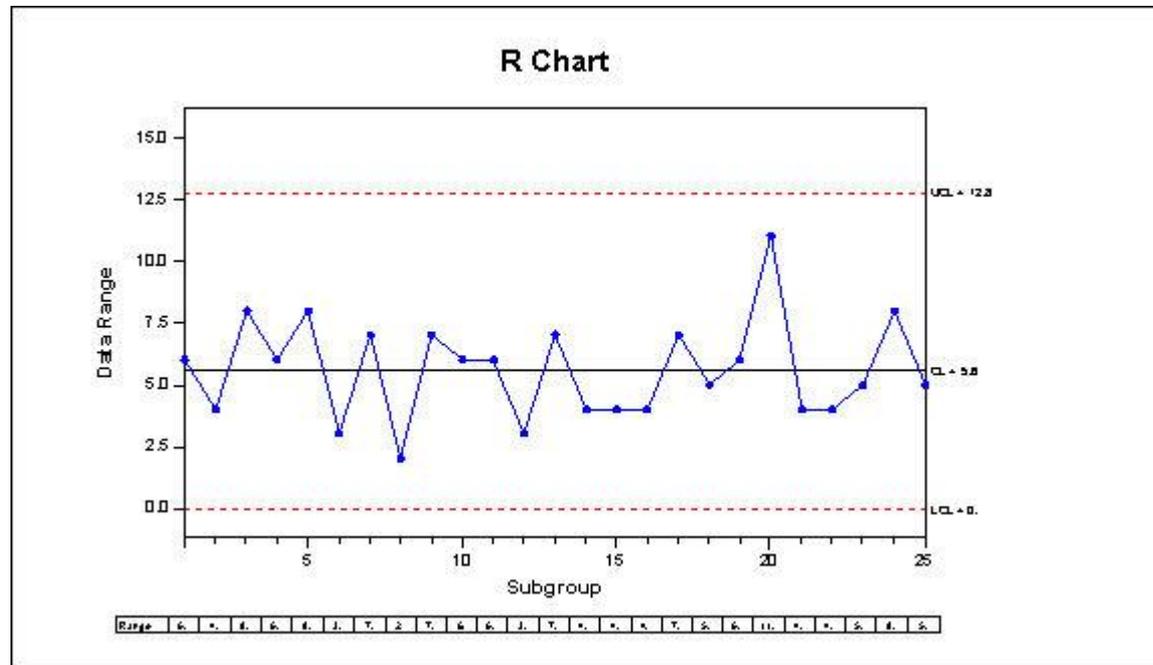
Introduction - Why R?

- ⊕ 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!**

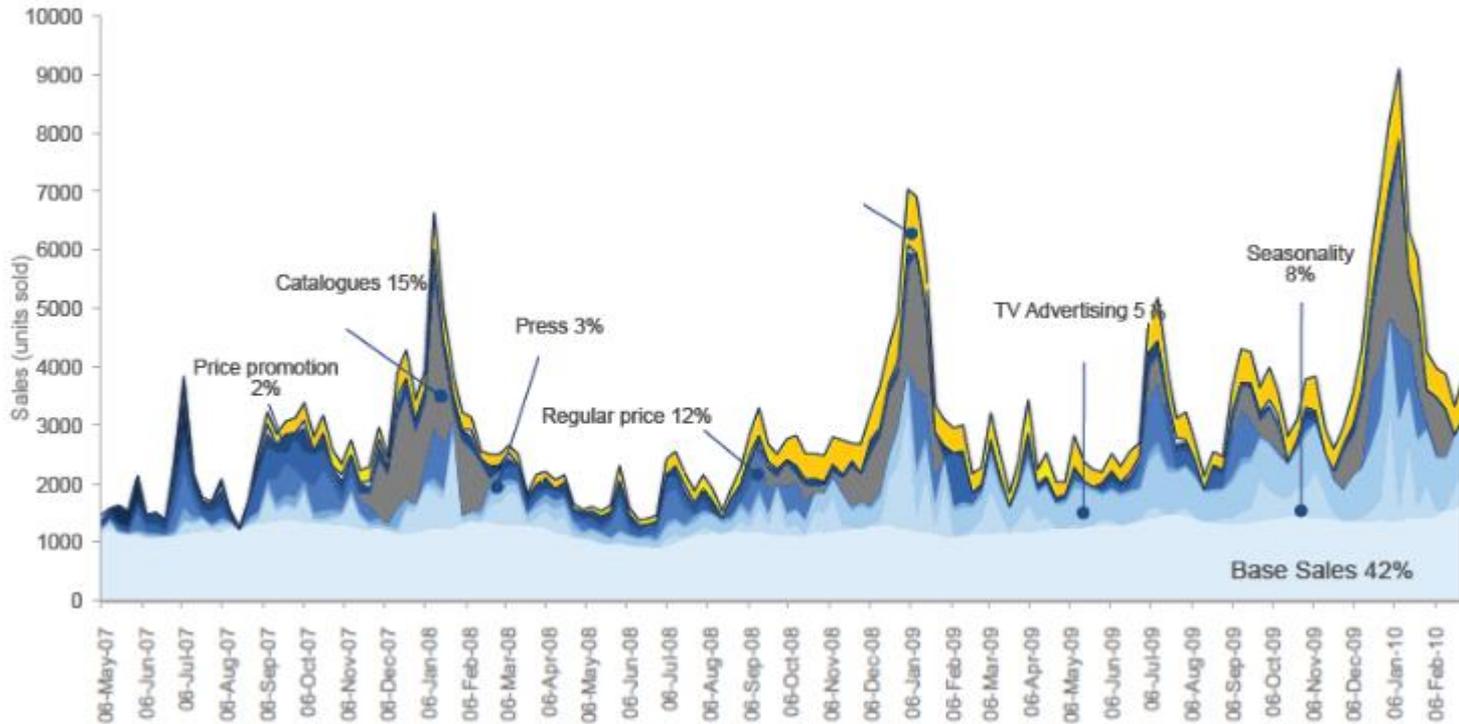
Introduction - Use of R

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



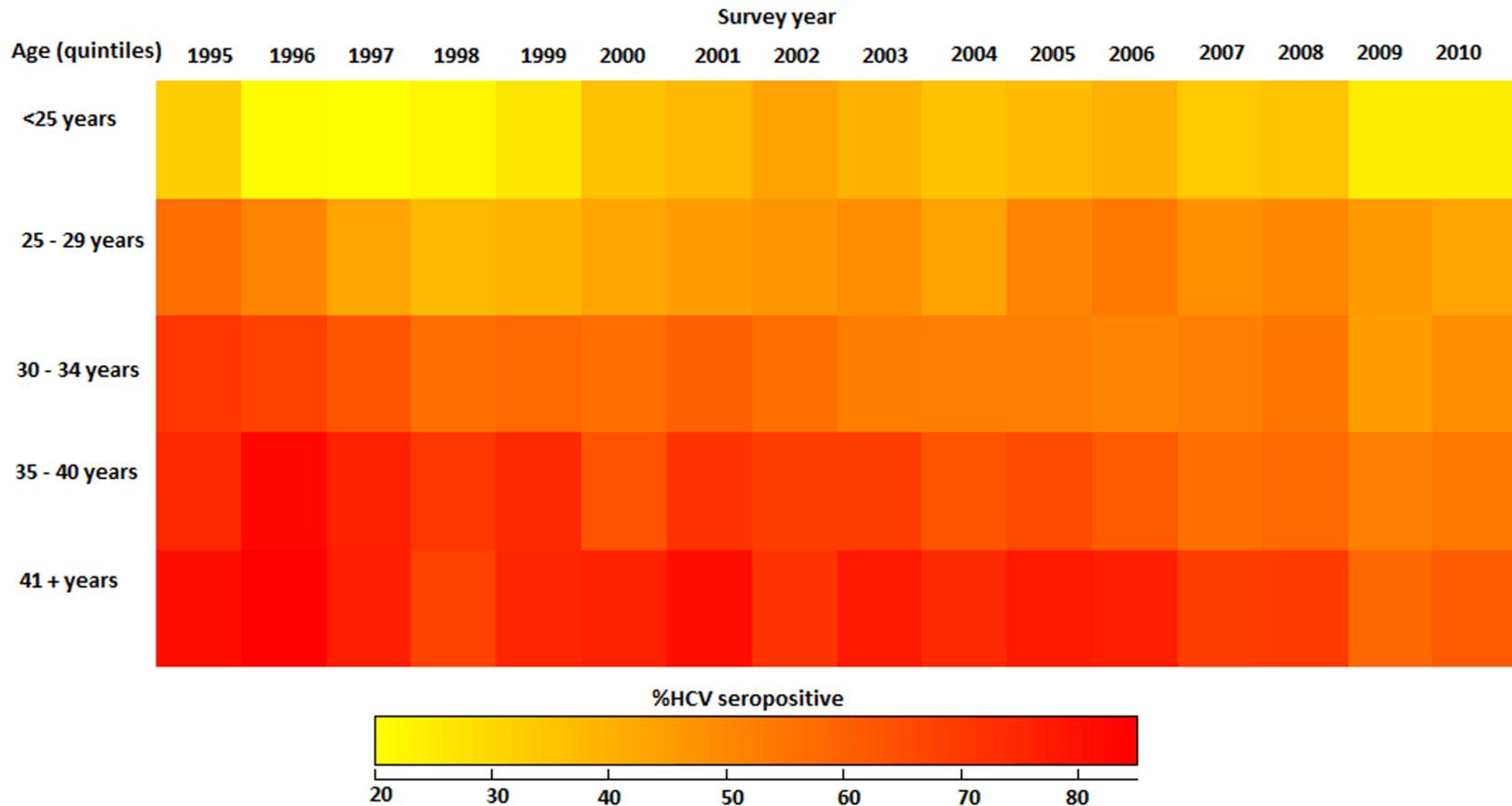
Introduction - Use of R

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



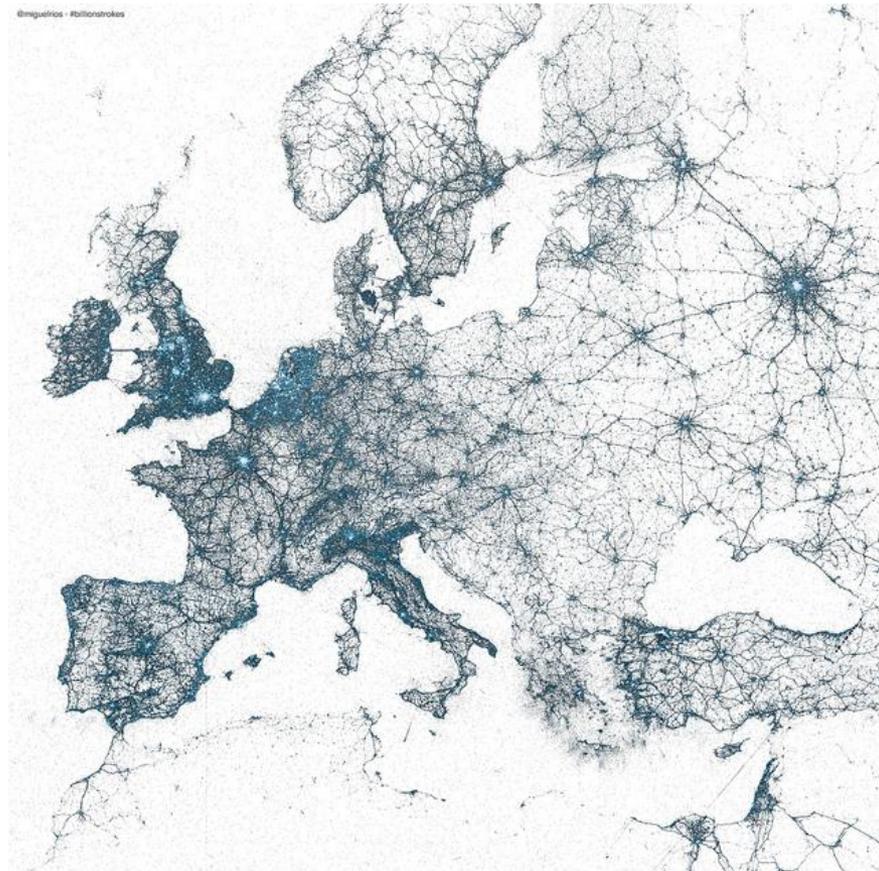
Introduction - Use of R

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



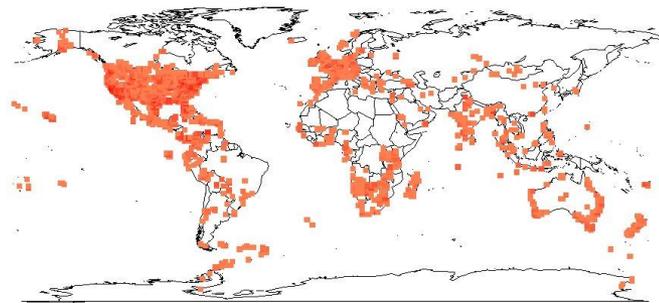
Introduction - Use of R

- Geo Packages: maps, google, GIS connection.



Introduction -R in Biodiversity

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



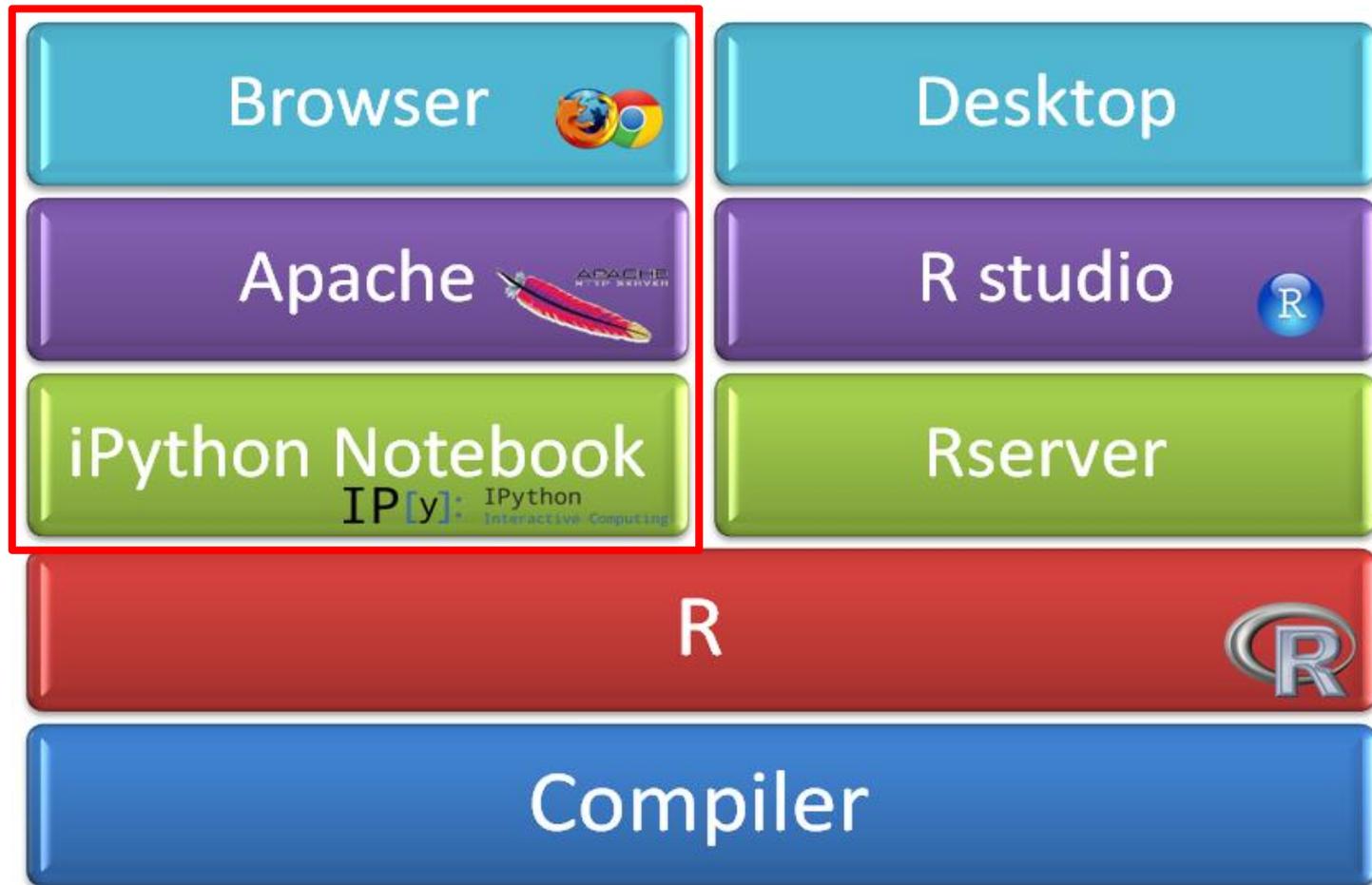
Lifewatch

- ❖ 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

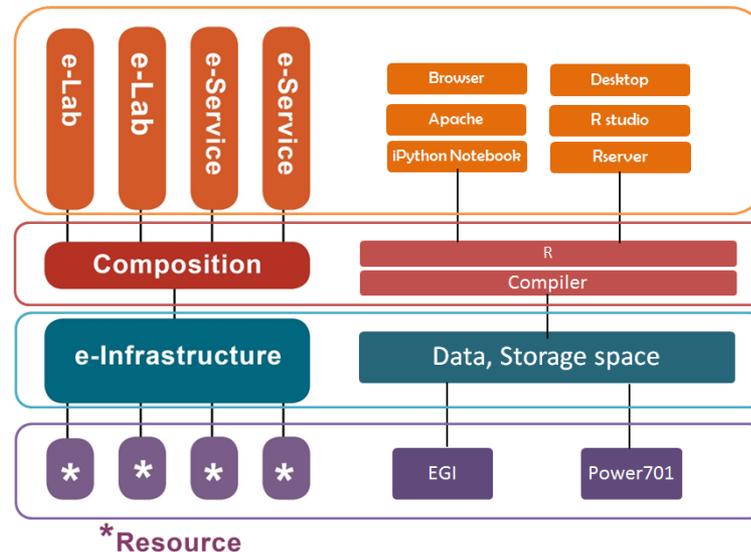
- ✿ Support the requirements of the community in Biodiversity and Ecosystems research.
- ✿ Establishing a direct collaboration between EGI.eu and the ESFRI LifeWatch to address specific needs.
- ✿ Four Mini-Projects:
 - ✦ Exploitation of the EGI infrastructure by the LifeWatch user community.
 - ✦ Tools 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



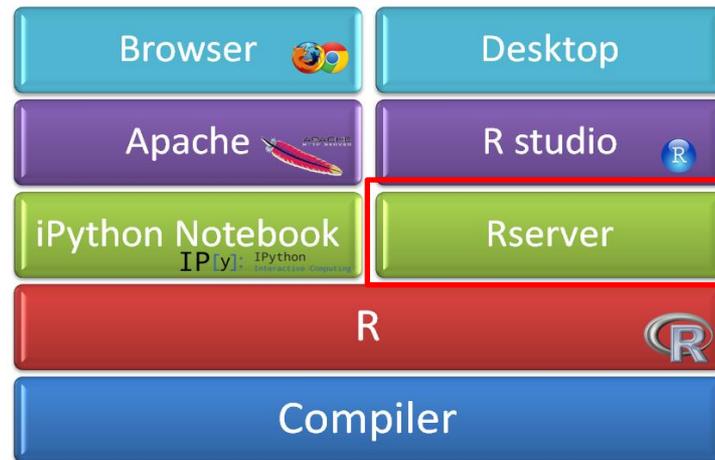
Architecture - Bottom Layer

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



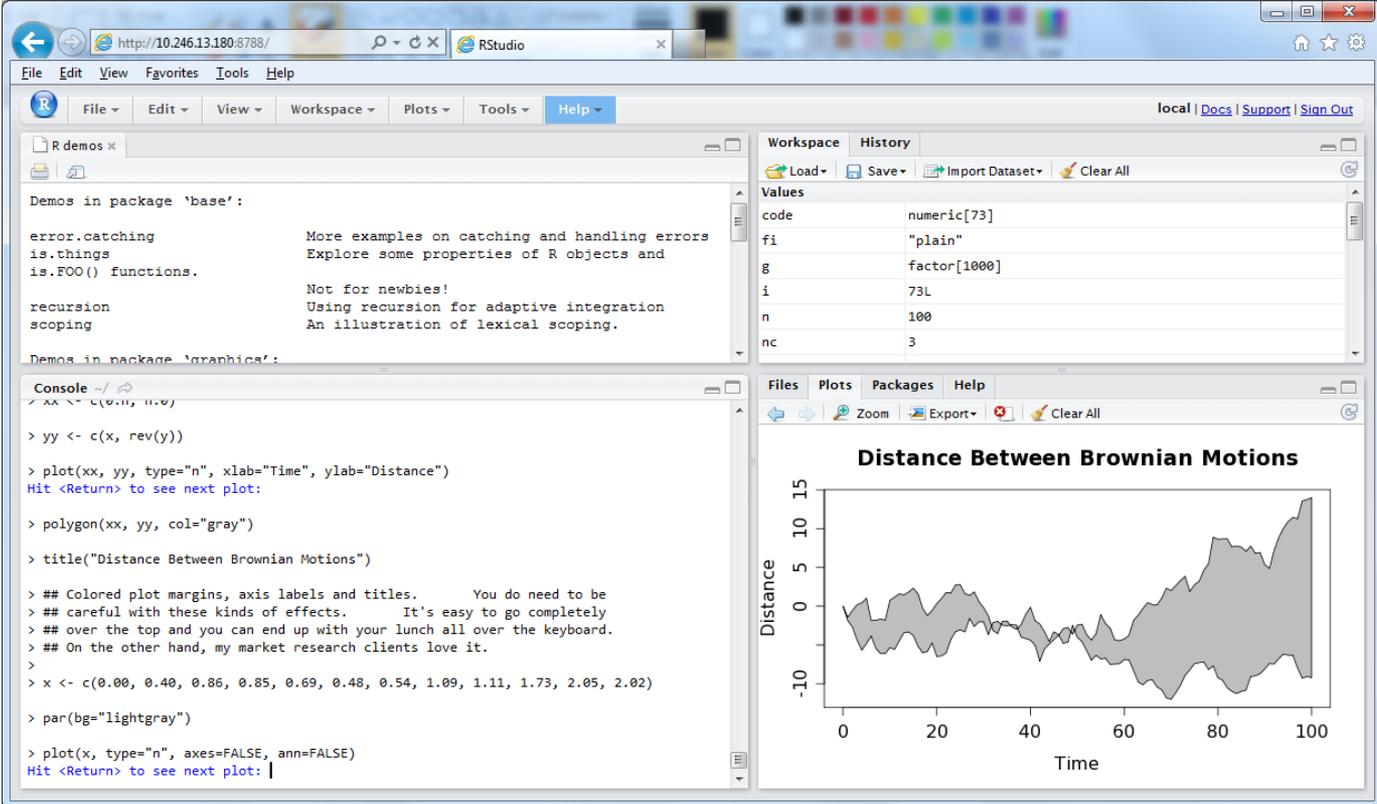
Architecture - Medium Layer

- ⊕ 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

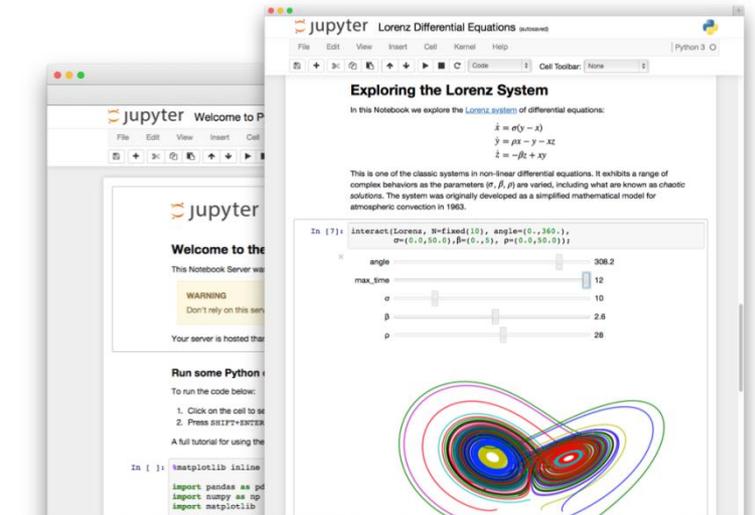


The screenshot displays the RStudio Server interface in a web browser. The browser address bar shows the URL `http://10.246.13.180:8788/`. The RStudio interface includes a menu bar (File, Edit, View, Favorites, Tools, Help), a toolbar, and several panels:

- Workspace:** Shows a table of values for variables: `code` (numeric[73]), `fi` ("plain"), `g` (factor[1000]), `i` (73L), `n` (100), and `nc` (3).
- Console:** Contains R code for generating a plot. The code includes comments about plot margins and axis labels, and defines variables `x` and `yy`.
- Plots:** Displays a plot titled "Distance Between Brownian Motions". The plot shows a gray shaded area representing the distance over time, with the x-axis labeled "Time" (0 to 100) and the y-axis labeled "Distance" (-10 to 15).

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.ipynthon
```

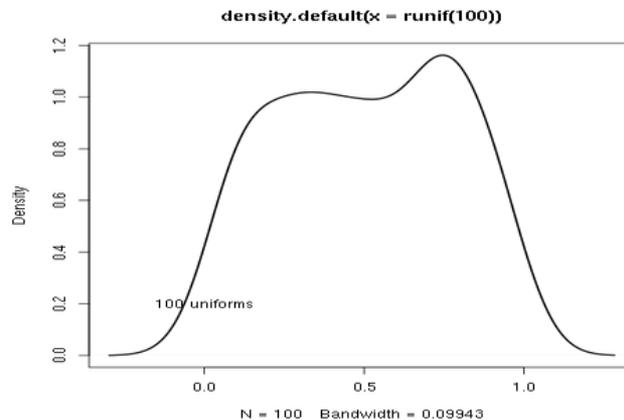
```
In [5]: %R getwd()
```

```
Out[5]: array(['/home/ipyb/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(0,0.01,1,0.01);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.

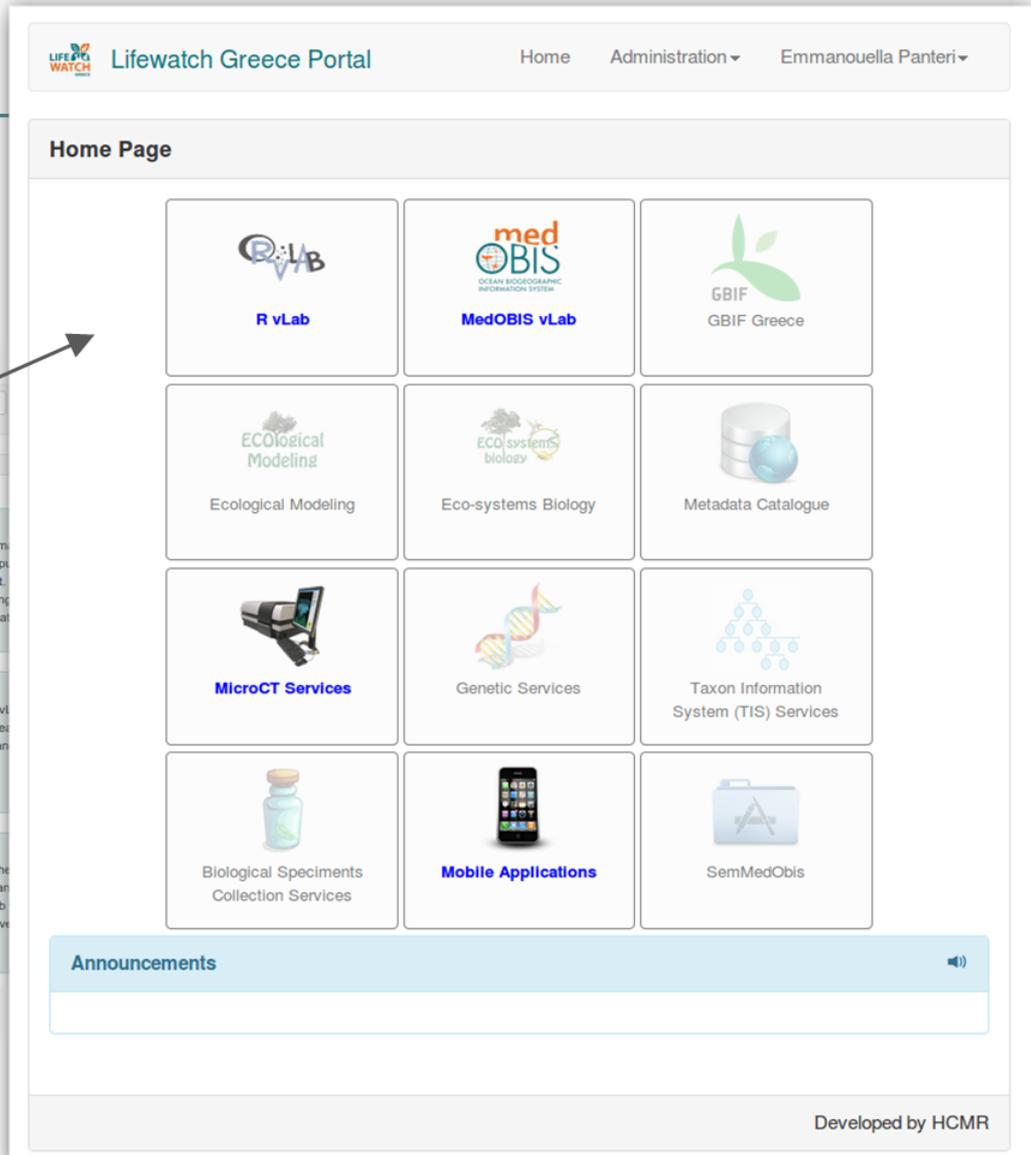
Main Objective:

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

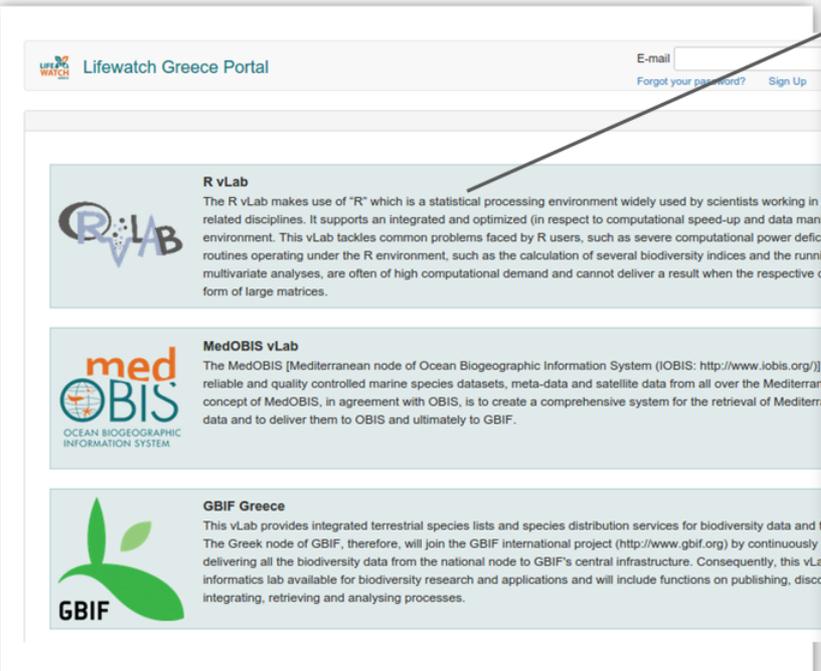
More specific issues addressed are:

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

portal.lifewatchgreece.eu



The screenshot shows the Lifewatch Greece Portal interface. At the top, there is a navigation bar with the logo, the text "Lifewatch Greece Portal", and links for "Home", "Administration", and "Emmanouella Panteri". Below this is a "Home Page" section featuring a 3x3 grid of service tiles. Each tile contains an icon, a title, and a brief description. The tiles are: R vLab, MedOBIS vLab, GBIF Greece, Ecological Modeling, Eco-systems Biology, Metadata Catalogue, MicroCT Services, Genetic Services, Taxon Information System (TIS) Services, Biological Specimens Collection Services, Mobile Applications, and SemMedObis. At the bottom of the grid is an "Announcements" section with a speaker icon. The footer of the page states "Developed by HCMR".



This is a detailed view of the R vLab service page. It features the R vLab logo and a paragraph of text: "The R vLab makes use of 'R' which is a statistical processing environment widely used by scientists working in 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 form of large matrices." Below this text are sections for "MedOBIS vLab" and "GBIF Greece", each with their respective logos and brief descriptions of their services.

R vLab
Help

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%

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

RvLab
Help

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%

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

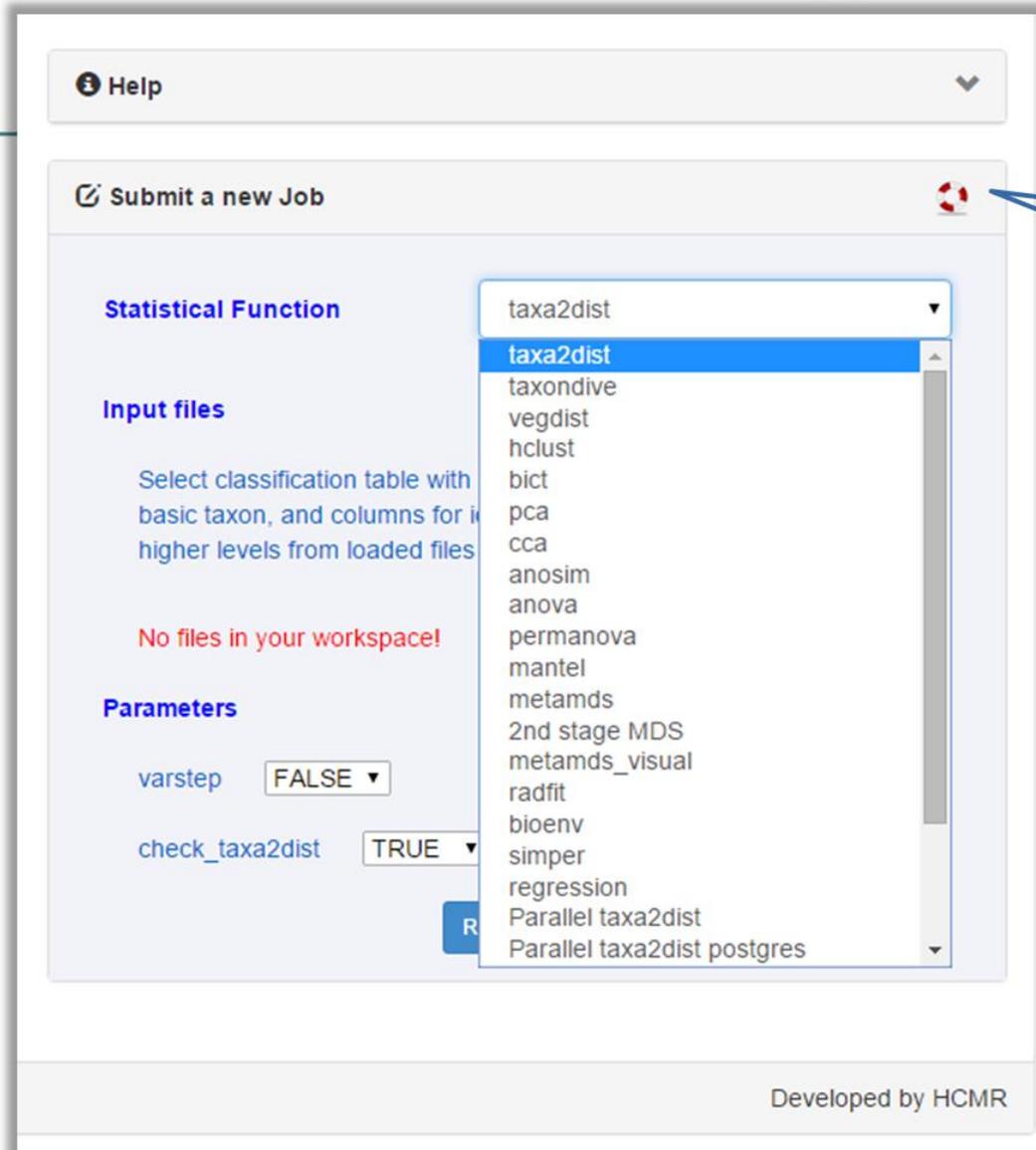
varstep FALSE

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



The screenshot shows the Rvlab web interface. At the top, there is a 'Help' button. Below it is a 'Submit a new Job' button with a red circular icon containing a white 'R' logo. The main area is divided into sections: 'Statistical Function', 'Input files', and 'Parameters'. The 'Statistical Function' section has a dropdown menu open, listing various functions. The 'Input files' section contains instructions and a message 'No files in your workspace!'. The 'Parameters' section has two dropdown menus: 'varstep' set to 'FALSE' and 'check_taxa2dist' set to 'TRUE'. At the bottom right, it says 'Developed by HCMR'.

Help

Submit a new Job

Statistical Function

taxa2dist

taxa2dist

taxondive

vegdist

hclust

bict

pca

cca

anosim

anova

permanova

mantel

metamds

2nd stage MDS

metamds_visual

radfit

bioenv

simper

regression

Parallel taxa2dist

Parallel taxa2dist postgres

Input files

Select classification table with basic taxon, and columns for higher levels from loaded files

No files in your workspace!

Parameters

varstep FALSE

check_taxa2dist TRUE

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 explore 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

