The R Community: An Insider’s Perspective

Achim Zeileis

https://eeecon.uibk.ac.at/~zeileis/
Overview

**R:**
- System for statistical computing.
- Open-source software under General Public License (GPL).
- [https://www.R-project.org/](https://www.R-project.org/)

**Insider:** Achim Zeileis.
- Statistician.
- Co-editor: Journal of Statistical Software.
- Ordinary member: R Foundation.
- Co-creator: useR! conference, R-Forge, ...
What is R?

**Based on:** ACM award-winning S language (core of commercial S-PLUS).

**Early 1990s:** Ross Ihaka and Robert Gentleman start reimplementation, eventually called **R**.

**Since 1997:**
- Base system developed by R Core Team.
- Highly extensible through packages.
- Openly shared through Comprehensive R Archive Network.

**Since 2000s:** Lingua franca in statistics. Around \(~100\) CRAN packages in 2000, more than 11,000 today (\(\sim 28\%\) nominal growth rate per year).

What is R?

Vantage points:
- Data analysis vs. programming.
- Statistics vs. data science.
- Community vs. app.
- Science vs. commerce.
What is R used for?

Classically: Statistics and graphics.
Linear regression, two-sample tests, scatter plots, bar charts, ...
What is R used for?

**Diversified methods:** Machine learning, social network analysis, econometrics, environmetrics, psychometrics, ...
**What is R used for?**

**Data structures:** Genomic data, spatial and space-time data, surveys, text corpora, connections to databases, ...
What is R used for?

**Specific applications:** Bioinformatics, business analytics, atmospheric sciences, finance, natural language processing, …
Why is R so successful?

• Open source.
• By statisticians for statisticians (in a very broad sense).
• Highly modular and extensible.
• Many subcommunities.
• Spillovers through joint journals, conferences, ...
• “Big Data Science.”
How does the R community work?

R Core/Foundation
Base system
CRAN
Mailing lists
How does the R community work?

**R Core/Foundation**
- Base system
- CRAN
- Mailing lists

**Scientific journals**
- *Journal of Statistical Software*

**Scientific conferences**

Other players
How does the R community work?

- **R Core/Foundation**
  - Base system
  - CRAN
  - Mailing lists

- **Scientific journals**
  - *Journal of Statistical Software*

- **(Scientific) conferences**

- **Other players**
How does the R community work?

R Core/Foundation
- Base system
- CRAN
- Mailing lists

 Scientiﬁc journals
- Journal of Statistical Software

The R Journal

Code collaboration

(Scientific) conferences

Bioconductor

How does the R community work?

- **R Core/Foundation**
  - Base system
  - CRAN
  - Mailing lists

- **Scientific journals**
  - Journal of Statistical Software
  - The R Journal

- **Code collaboration**
  - Bioconductor
  - R-Forge

- **(Scientific) conferences**

- **Use R!**
How does the R community work?

**R Core/Foundation**
- Base system
- CRAN
- Mailing lists

**Scientific journals**
- Journal of Statistical Software
- The R Journal

**Scientific (Scientific) conferences**

**Code collaboration**
- Bioconductor
- R-Forge
- GitHub
How does the R community work?

**R Core/Foundation**
- Base system
- CRAN
- Mailing lists

**Scientific conferences**

**Scientific journals**
- Journal of Statistical Software

**The R Journal**

**Code collaboration**
- Bioconductor
- R-Forge
- GitHub

**Communication**
- R-bloggers
How does the R community work?

**R Core/Foundation**
- Base system
- CRAN
- Mailing lists

**Scientific conferences**

**Scientific journals**
- *Journal of Statistical Software*
- The *R* Journal

**Code collaboration**
- Bioconductor
- R-Forge
- GitHub

**Communication**
- R-bloggers
- stackoverflow
- #rstats
How does the R community work?

R Core/Foundation
- Base system
- CRAN
- Mailing lists

Scientific journals
- Journal of Statistical Software
- The R Journal

Scientific conferences

Code collaboration
- Bioconductor
- R-Forge
- GitHub

Communication
- R-bloggers
- stackoverflow
- #rstats

Other players
- Microsoft
- RStudio
- DataCamp
- R OpenSci
How does the R community work?

**R Core/Foundation**
- Base system
- CRAN
- Mailing lists

**Scientific journals**
- *Journal of Statistical Software*
- The *R* Journal

**Code collaboration**
- Bioconductor
- R-Forge
- GitHub

**Communication**
- R-bloggers
- stackoverflow
- Twitter: #rstats

**Other players**
- Microsoft
- DataCamp
- RStudio
- R OpenSci
- R consortia

**(Scientific) conferences**
Why do you contribute to the R community?

In 1999: Undergraduate.
- “Why do you use R? We do have an S-PLUS license.”
- Open source!
Why do you contribute to the R community?

**In 1999:** Undergraduate.
- “*Why do you use R? We do have an S-PLUS license.*”
- Open source!

**In 2002:** PhD student.
- “*Why do you publish in online-only journals? That’s just like a technical report.*”
- Open access (free for everyone)!
Why do you contribute to the R community?

In 1999: Undergraduate.
- “Why do you use R? We do have an S-PLUS license.”
- Open source!

In 2002: PhD student.
- “Why do you publish in online-only journals? That’s just like a technical report.”
- Open access (free for everyone)!

Since 2004: Postdoc onwards.
- “Why do you volunteer to edit a free journal and organize conferences? You should make some money.”
- Open and reproducible science!
Why do others contribute to R?

Drivers: For participation in packages/conferences/mailing lists.

- Hybrid form of motivation:
  Moderated intrinsic motivation; well-internalized extrinsic motivation.

- Social characteristics of the work design:
  Feedback; social inclusion; building reputation.
Why do others contribute to R?

Drivers: For participation in packages/conferences/mailing lists.

- *Hybrid form of motivation:*
  Moderated intrinsic motivation; well-internalized extrinsic motivation.

- *Social characteristics of the work design:*
  Feedback; social inclusion; building reputation.

---

**R motivation survey**

What are interesting case studies?

**Weather forecasting**

What are interesting case studies?

### Weather forecasting


### Natural language processing

Precipitation forecasting in Tyrol

Input

Data from global forecast model (ECMWF):
GRIB/NCDF files.
Precipitation forecasting in Tyrol

Input

Data from global forecast model (ECMWF): GRIB/NCDF files.

Data wrangling

Spatiotemporal data: raster, ncdf4, rgdal, sp, zoo.

Database: RMySQL, RSQLite.
Precipitation forecasting in Tyrol

Input

Data from global forecast model (ECMWF):
GRIB/NCDF files.

Data wrangling

Spatiotemporal data: raster, ncdf4, rgdal, sp, zoo.
Database: RMySQL, RSQLite.

Statistical post-processing

Flexible probabilistic regression modeling:
mgcv, crch, bamlss.

Visualization

Weather maps: sp, leaflet.

Deployment

Web server with R interface: shiny, shinyjs.
Precipitation forecasting in Tyrol

**Input**

Data from global forecast model (ECMWF):
GRIB/NCDF files.

**Data wrangling**

Spatiotemporal data: raster, ncdf4, rgdal, sp, zoo.
Database: RMySQL, RSQLite.

**Statistical post-processing**

Flexible probabilistic regression modeling:
mgcv, crch, bamlss.

**Visualization**

Weather maps:
sp, leaflet.
Precipitation forecasting in Tyrol

Input

Data from global forecast model (ECMWF):
GRIB/NCDF files.

Data wrangling

Spatiotemporal data: raster, ncdf4, rgdal, sp, zoo.

Database: RMySQL, RSQLite.

Statistical post-processing

Flexible probabilistic regression modeling:
mgcv, crch, bamlss.

Visualization

Weather maps:
sp, leaflet.

Deployment

Web server with R interface:
shiny, shinyjs.
Precipitation forecasting in Tyrol

Input
Data from global forecast model (ECMWF): GRIB/NCDF files.

Data wrangling
Spatiotemporal data: raster, ncdf4, rgdal, sp, zoo.
Database: RMySQL, RSQlLite.

Statistical post-processing
Flexible probabilistic regression modeling:
mgcv, crch, bamlss.

Visualization
Weather maps:
sp, leaflet.

Deployment
Web server with R interface:
shiny, shinyjs.
Precipitation forecasting in Tyrol
Precipitation forecasting in Tyrol
Precipitation forecasting in Tyrol

- **Location:** 6,846
- **Scale:** 1,578
- **POP:** 100%
- **Expectation:** 10.03 mm/d
- **Ensemble mean:** 10.94 mm/d
- **Longitude:** 11.3928
- **Latitude:** 47.2672

**Forecast**
- **Tue 2017-07-11 to Wed 2017-07-12**
- Map by OpenStreetMap, CC BY-SA.
Text mining of Republican voter statements

Input

Republican faces (http://www.GOP.com/)
“I’m a Republican, because ...”
Text mining of Republican voter statements

Input

Republican faces (http://www.GOP.com/)
“I’m a Republican, because …”

Data wrangling

Web scraping: XML.
Corpus computation: tm.
Text mining of Republican voter statements

Input

Republican faces (http://www.GOP.com/)
“I’m a Republican, because …”

Data wrangling

Web scraping: XML.
Corpus computation: tm.

Statistical modeling

Scaling and clustering: smacof, hclust.
Social network analysis: igraph, ape.
Text mining of Republican voter statements

Input

Republican faces (http://www.GOP.com/)
“I’m a Republican, because …”

Data wrangling

Web scraping: XML.
Corpus computation: tm.

Statistical modeling

Scaling and clustering: smacof, hclust.
Social network analysis: igraph, ape.

Visualization

Dendrograms: ggplot2.
Graphs: igraph.
Text mining of Republican voter statements

Input

Republican faces (http://www.GOP.com/)
“I’m a Republican, because …”

Data wrangling

Web scraping: XML.
Corpus computation: tm.

Statistical modeling

Scaling and clustering: smacof, hclust.
Social network analysis: igraph, ape.

Visualization

Dendrograms: ggplot2.
Graphs: igraph.
Text mining of Republican voter statements
Text mining of Republican voter statements
Text mining of Republican voter statements
<table>
<thead>
<tr>
<th></th>
<th>Helpful</th>
<th>Harmful</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal</strong></td>
<td><strong>Strengths</strong></td>
<td><strong>Weaknesses</strong></td>
</tr>
<tr>
<td></td>
<td>Rich network of packages.</td>
<td>Scaling (e.g., CRAN, useR!).</td>
</tr>
<tr>
<td></td>
<td>Broad and active community.</td>
<td>Little centralized consolidation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and coordination.</td>
</tr>
<tr>
<td><strong>External</strong></td>
<td><strong>Opportunities</strong></td>
<td><strong>Threats</strong></td>
</tr>
<tr>
<td></td>
<td>More challenging data.</td>
<td>Fragmentation.</td>
</tr>
<tr>
<td></td>
<td>Need for data-driven methods.</td>
<td>Players with different agendas.</td>
</tr>
</tbody>
</table>
Where are we going from here?

**Quite certainly:** More growth and more diversity.

**Unclear:** Whether “one” R community will persist.

**Crucial:** Communication and exchange within and beyond the community.

**High potential:** Exciting and innovative collaborations across disciplines.