



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/>

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: **R**oss Ihaka and **R**obert 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 (~ 28% nominal growth rate per year).

Since 2010s: Popular programming language (#5, IEEE Spectrum 2016), especially for data science (KDnuggets 2015–2017, Top 2: Python & R).

What is R?

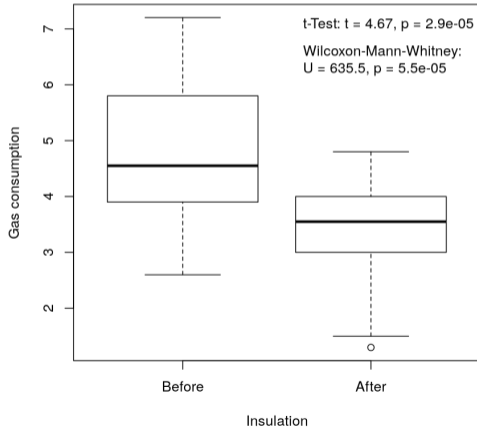
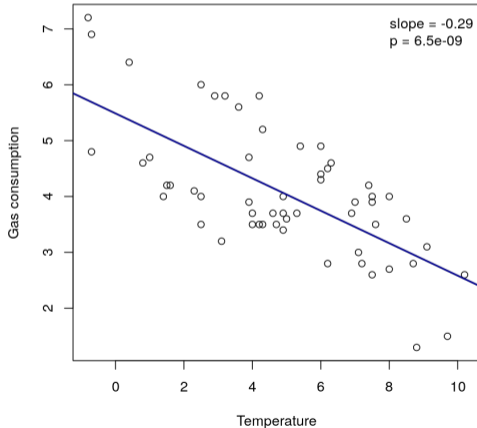
Vantage points:

- Data analysis vs. programming.
- Statistics vs. data science.
- Community vs. app.
- Science vs. commerce.

What is R used for?

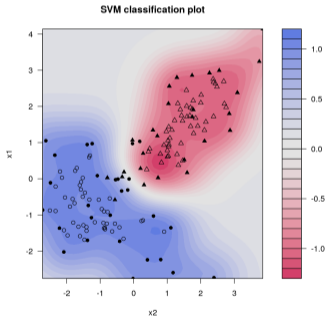
Classically: Statistics and graphics.

Linear regression, two-sample tests, scatter plots, bar charts, ...

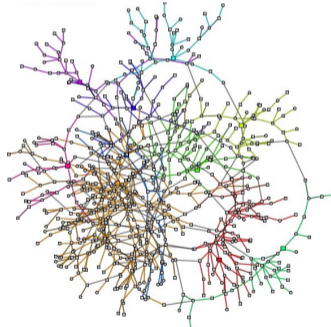


What is R used for?

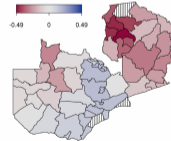
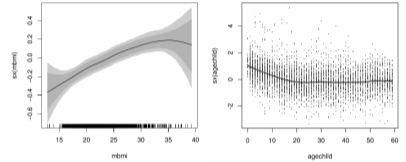
Diversified methods: Machine learning, social network analysis, econometrics, environmetrics, psychometrics, ...



doi:10.18637/jss.v015.i09



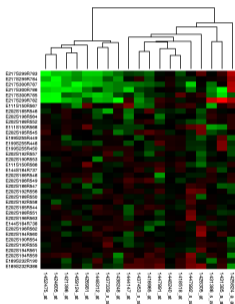
doi:10.18637/jss.v024.i07



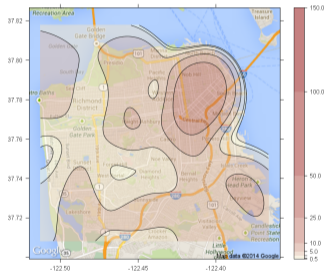
doi:10.18637/jss.v063.i121

What is R used for?

Data structures: Genomic data, spatial and space-time data, surveys, text corpora, connections to databases, ...

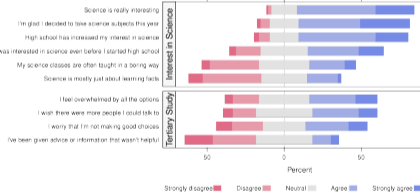


https://en.wikipedia.org/wiki/Heat_map



doi:10.18637/jss.v063.i04

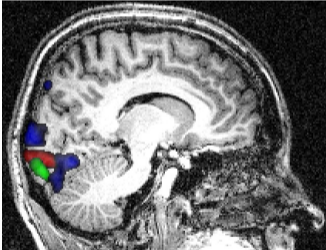
New Zealand Students Still Taking Science in Year 13



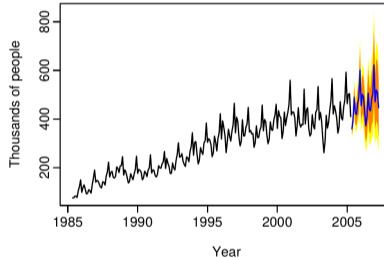
doi:10.18637/jss.v057.i05

What is R used for?

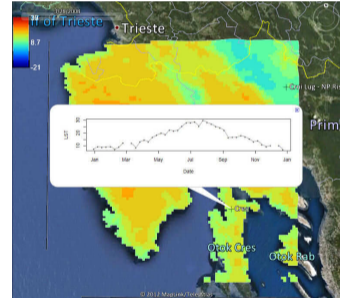
Specific applications: Bioinformatics, business analytics, atmospheric sciences, finance, natural language processing, ...



doi:10.18637/jss.v044.i09



doi:10.18637/jss.v027.i03



doi:10.18637/jss.v063.i05

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



How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

Scientific journals



*Journal of
Statistical Software*

The  Journal

(Scientific) conferences



useR!

How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

Scientific journals



*Journal of
Statistical Software*

The  Journal

(Scientific) conferences



Code collaboration



How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

Scientific journals



*Journal of
Statistical Software*

The  Journal

(Scientific) conferences



useR!

Code collaboration



How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

Scientific journals



*Journal of
Statistical Software*

The  Journal

(Scientific) conferences



useR!

Code collaboration



How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

Scientific journals



*Journal of
Statistical Software*

The  Journal

(Scientific) conferences



useR!

Code collaboration



Communication



How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

(Scientific) conferences



Scientific journals



*Journal of
Statistical Software*

The  Journal

Code collaboration



Communication



How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

Scientific journals



*Journal of
Statistical Software*

The  Journal

(Scientific) conferences



Code collaboration



Communication



Other players



DataCamp



How does the R community work?



R Core/Foundation

Base system

CRAN

Mailing lists

Scientific journals



*Journal of
Statistical Software*

The  Journal

(Scientific) conferences



Code collaboration



Communication



Other players



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

Mair P, Hofmann E, Gruber K, Hatzinger R, Zeileis A, Hornik K (2015).

“Motivation, Values, and Work Design as Drivers of Participation in the R Open Source Project for Statistical Computing.” *PNAS – Proceedings of the National Academy of Sciences of the United States of America*, **112**(48), 14788–14792.
doi:10.1073/pnas.1506047112

What are interesting case studies?

Weather forecasting

Stauffer R, Umlauf N, Messner JW, Mayr GJ, Zeileis A (2017).

“Ensemble Post-Processing of Daily Precipitation Sums over Complex Terrain Using Censored High-Resolution Standardized Anomalies.”

Monthly Weather Review, **45**(3), 955–969. doi:10.1175/MWR-D-16-0260.1

What are interesting case studies?

Weather forecasting

Stauffer R, Umlauf N, Messner JW, Mayr GJ, Zeileis A (2017).

“Ensemble Post-Processing of Daily Precipitation Sums over Complex Terrain Using Censored High-Resolution Standardized Anomalies.”

Monthly Weather Review, **45**(3), 955–969. doi:10.1175/MWR-D-16-0260.1

Natural language processing

Mair P, Rusch T, Hornik K (2014).

“The Grand Old Party – A Party of Values?” *SpringerPlus*, **3**(697), 1–10.

doi:10.1186/2193-1801-3-697

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.

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, RSQLite.

Statistical post-
processing



Flexible probabilistic regression modeling:
mgcv, **crch**, **bamlss**.

Visualization



Weather maps:
sp, leaflet.

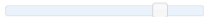
Deployment



Web server with R interface:
shiny, shinyjs.

Precipitation forecasting in Tyrol

Overlay Opacity



Type selection

- RAW ENS
- MOS forecast

Product selection

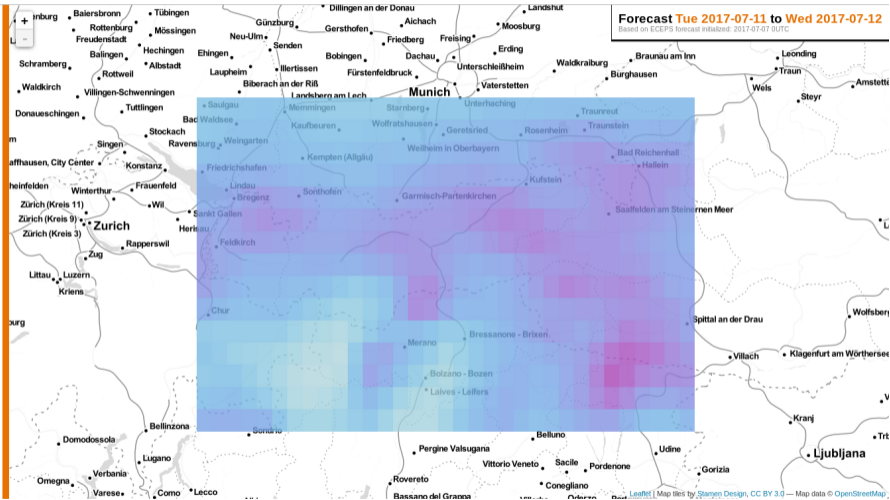
- **Expectation**
- Probability($r > 0\text{mm}/24\text{h}$)
- Probability($r > 1\text{mm}/24\text{h}$)
- Probability($r > 5\text{mm}/24\text{h}$)
- Probability($r > 10\text{mm}/24\text{h}$)

Forecast horizon

- Day 1 (+6 to +30)
- Day 2 (+30 to +54)
- Day 3 (+54 to +78)
- Day 4 (+78 to +102)
- **Day 5 (+102 to +126)**
- Day 6 (+126 to +150)

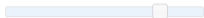
Points of Interest

- Innsbruck
- Hafelekar
- Moosweg, Rum
- St. Anton a.A.
- Lienz



Precipitation forecasting in Tyrol

Overlay Opacity



Type selection

- RAW ENS
- MOS forecast

Product selection

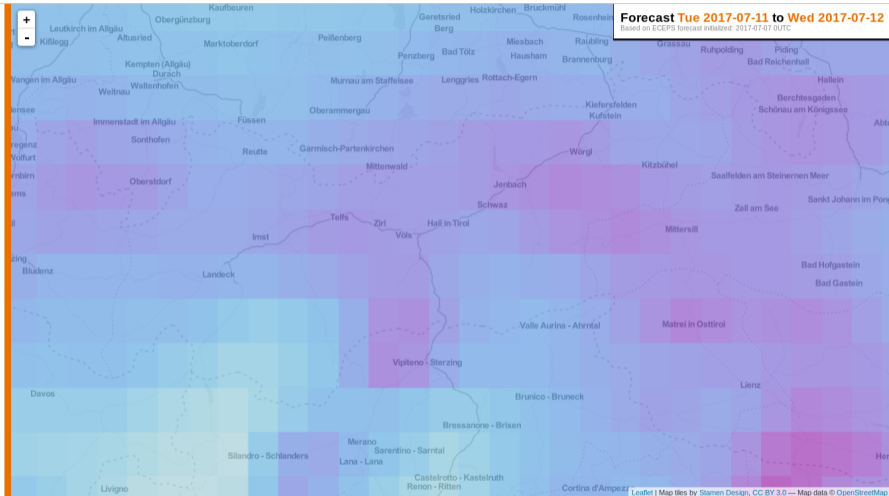
- Expectation
- Probability($r > 0\text{mm}/24\text{h}$)
- Probability($r > 1\text{mm}/24\text{h}$)
- Probability($r > 5\text{mm}/24\text{h}$)
- Probability($r > 10\text{mm}/24\text{h}$)

Forecast horizon

- Day 1 (+6 to +30)
- Day 2 (+30 to +54)
- Day 3 (+54 to +78)
- Day 4 (+78 to +102)
- Day 5 (+102 to +126)
- Day 6 (+126 to +150)

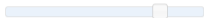
Points of Interest

- Innsbruck
- Hafelekar
- Moosweg, Rum
- St. Anton a.A.
- Lienz



Precipitation forecasting in Tyrol

Overlay Opacity



Type selection

- RAW ENS
- MOS forecast

Product selection

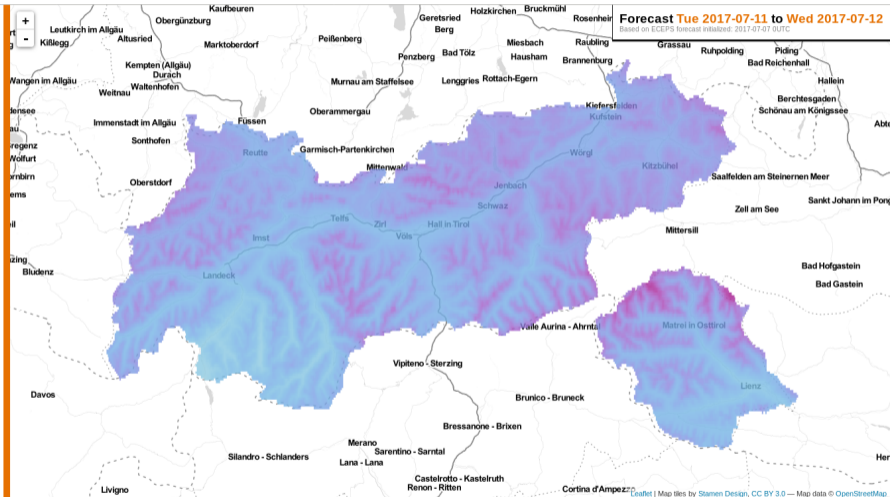
- **Expectation**
- Probability($r > 0\text{mm}/24\text{h}$)
- Probability($r > 1\text{mm}/24\text{h}$)
- Probability($r > 5\text{mm}/24\text{h}$)
- Probability($r > 10\text{mm}/24\text{h}$)

Forecast horizon

- Day 1 (+6 to +30)
- Day 2 (+30 to +54)
- Day 3 (+54 to +78)
- Day 4 (+78 to +102)
- **Day 5 (+102 to +126)**
- Day 6 (+126 to +150)

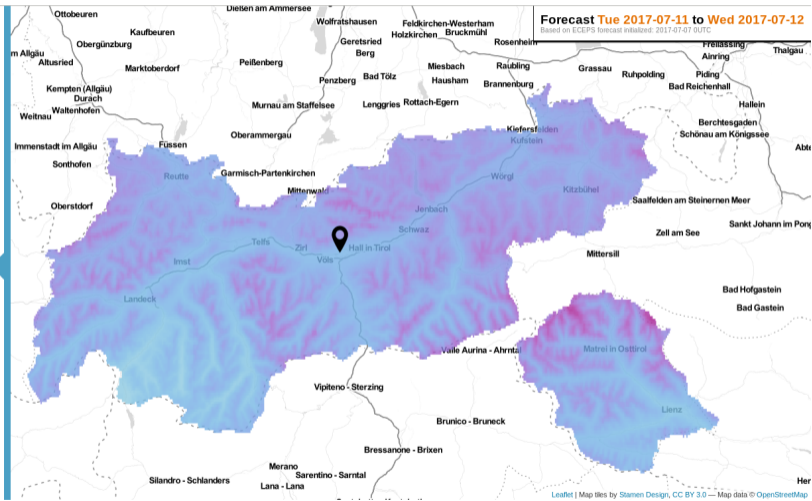
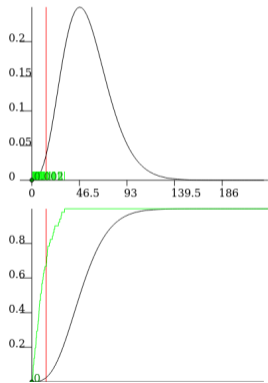
Points of Interest

- Innsbruck
- Hafelekar
- Moosweg, Rum
- St. Anton a.A.
- Lienz



Precipitation forecasting in Tyrol

Location: 6.846
Scale: 1.578
POP: 100 %
Expectation: 14.03 mm/d
Ensemble mean: 10.94 mm/d
Longitude: 11.3928
Latitude: 47.2672



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.

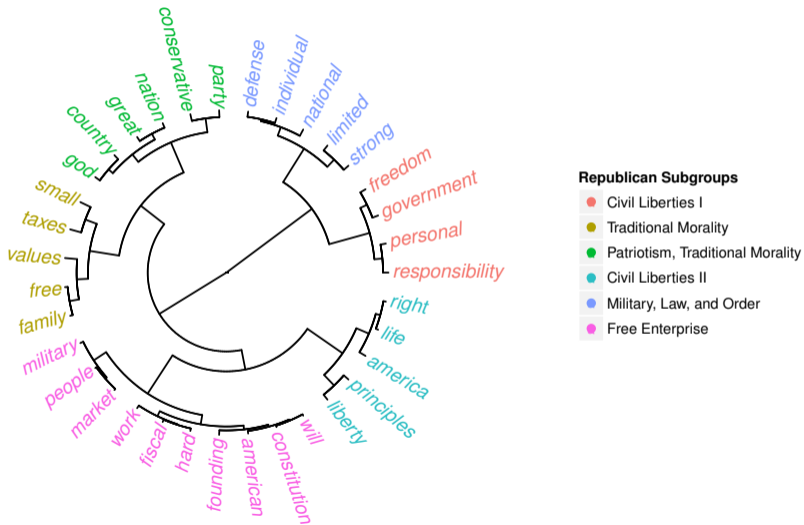
Visualization



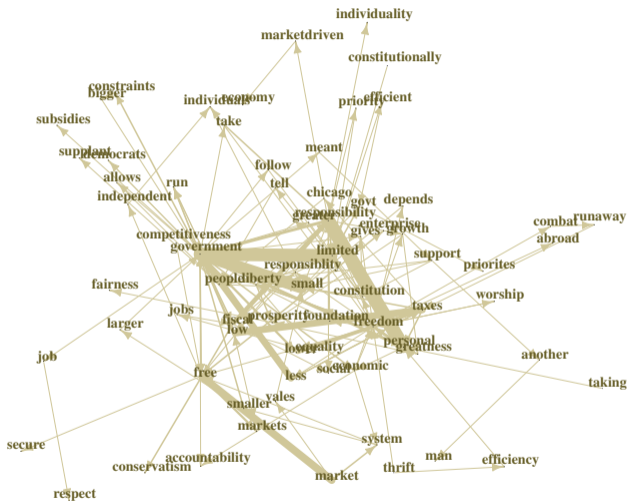
Dendrograms: ggplot2.

Graphs: igraph.

Text mining of Republican voter statements



Text mining of Republican voter statements



Where are we going from here?

Helpful

Harmful

Internal

Strengths

Rich network of packages.
Broad and active community.

Weaknesses

Scaling (e.g., CRAN, useR!).
Little centralized consolidation
and coordination.

External

Opportunities

More challenging data.
Need for data-driven methods.

Threats

Fragmentation.
Players with different agendas.

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.