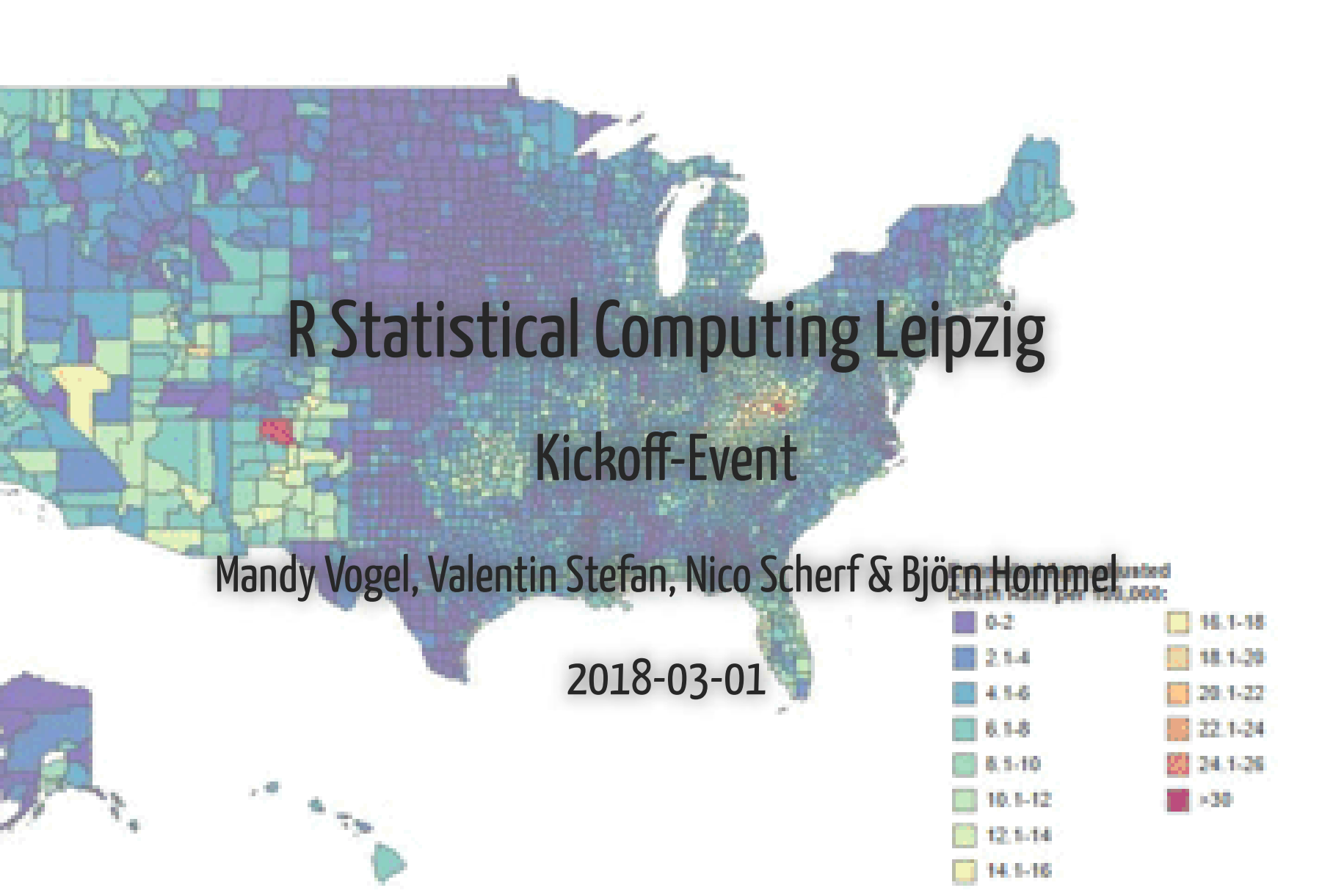


R Statistical Computing Leipzig

Kickoff-Event

Mandy Vogel, Valentin Stefan, Nico Scherf & Björn Hommel

2018-03-01

Agenda

1. Brief introduction to R
2. Real-world applications
3. Meetups in 2018
4. Open end

What topics are you interested in?

Please provide as many ideas as possible to help us make future meetups fun and engaging!

	x	freq
5	Data Visualization	23
6	Exploratory Data Analysis	19
18	Time Series Analysis	15
4	Data Screening & Cleaning	13
12	R and Databases	13
9	Machine Learning	12
20	Reporting with R Markdown/Sweave/Reproducible Research	12
1	Bayes: R and BUGS/JAGS/Stan	10
7	Image Processing	10
17	Statistical Inference (e.g. Natural Language Processing)	10

R-Leipzig Member Survey

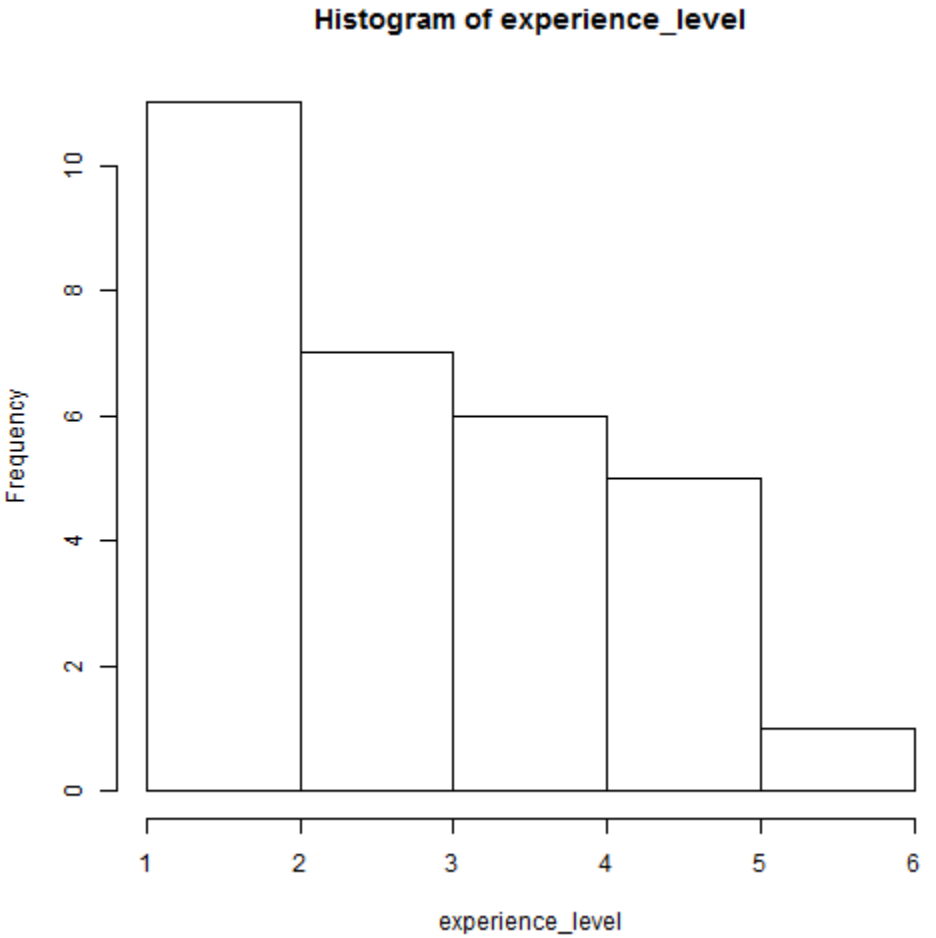
```
require(RCurl)
require(plyr)

dataCSV <- getURL('https://docs.google.com/spreadsheets/d/e/2PACX-1vSWD6_CABX
df <- read.csv(textConnection(dataCSV), stringsAsFactors = FALSE)

topics <- unlist(df$topics)
topics <- unlist(strsplit(topics, ','))
topics <- plyr::count(topics)
topics <- topics[order(topics$freq, decreasing = TRUE),]

knitr::kable(head(topics, 10), format = 'html')
```

How would you rate your experience level?



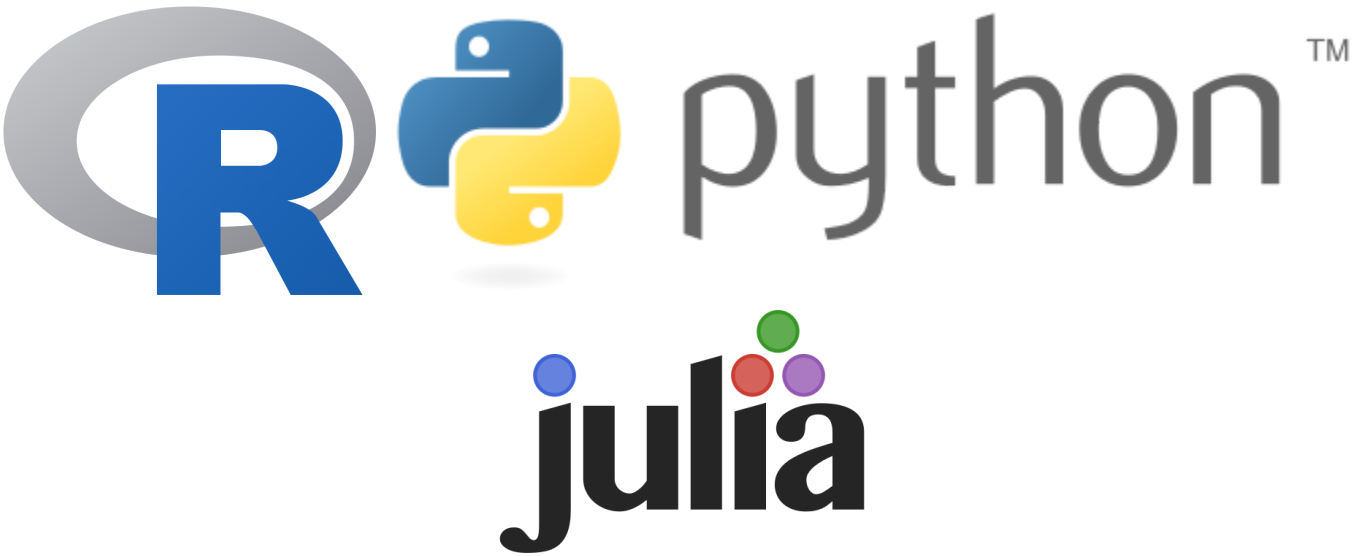
Do you have a background in one or more of the following sciences?

	x	freq
3	Data Sciences	10
7	Healthcare)	10
10	Logic)	9
15	Formal Sciences (e.g. Mathematics	9
5	Genetics)	8
4	Economics)	6
12	Sociology	6
17	Life Sciences (e.g. Biochemistry	6
19	Social Sciences (e.g. Psychology	6
1	Applied Sciences (e.g. Engineering	5

Introduction to R

Why use R?

A quick comparison to other languages



> "The best thing about `R` is that it was written by statisticians."

> "The best thing about `R` is that it was written by statisticians."

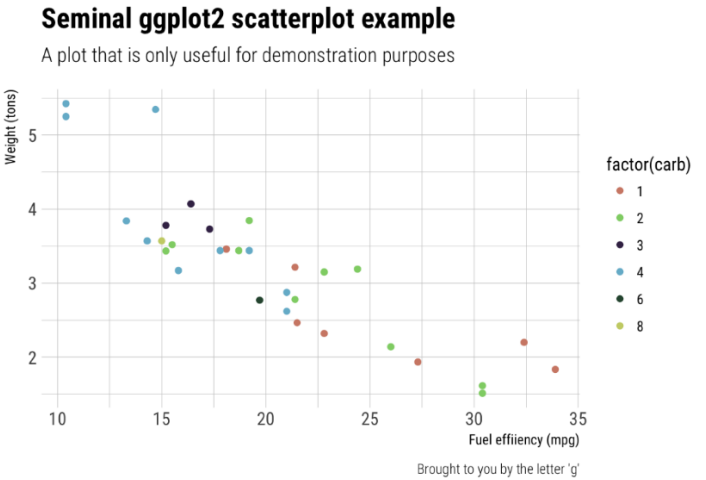
while

> "The worst thing about `R` is that it was written by statisticians."

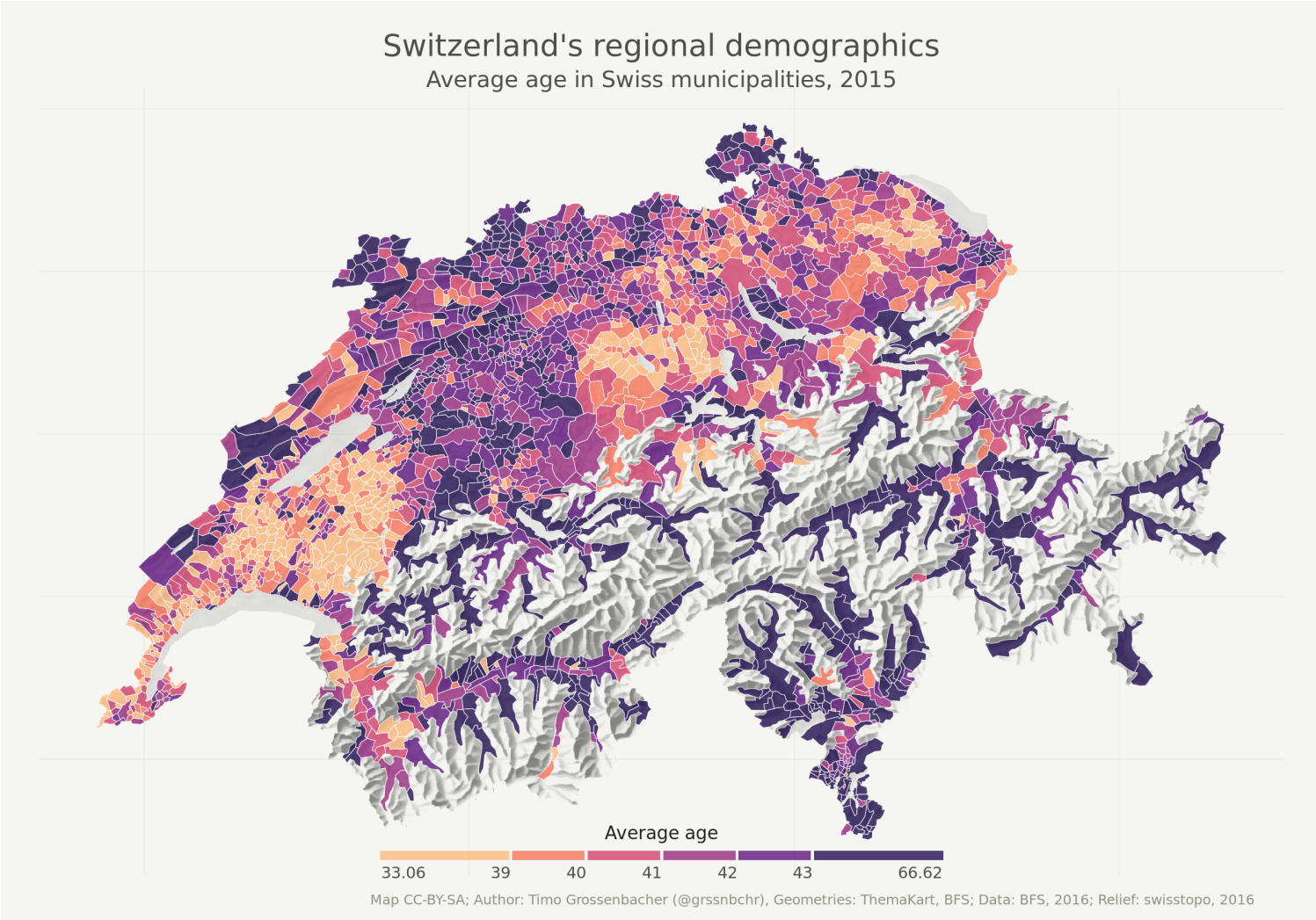
- better, user friendly data analysis
- easy to use complex formula and advanced models
- steep learning curve (not hard for experienced programmers)
- good for standalone analysis
- the closer you are to statistics and research the more you might prefer R

- productivity and code reusability
- more flexible to program new stuff
- easy to learn for beginners
- good for analysis integrated into larger frameworks
- the closer you are to engineering, the more you might prefer python

- **Visualisation**
- Lingua franca of statistics
- Traditional computational statistics (machine learning)
- **Data exploration**
- R Studio
- Shiny



https://hrbrmstr.github.io/hrbrthemes/



https://timogrossenbacher.ch/2016/12/beautiful-thematic-maps-with-ggplot2-only/

- slow (and sometimes ugly) code
- steep learning curve

- producing nice visualisations can be a pain
- immature functionality for data analysis

Open Source

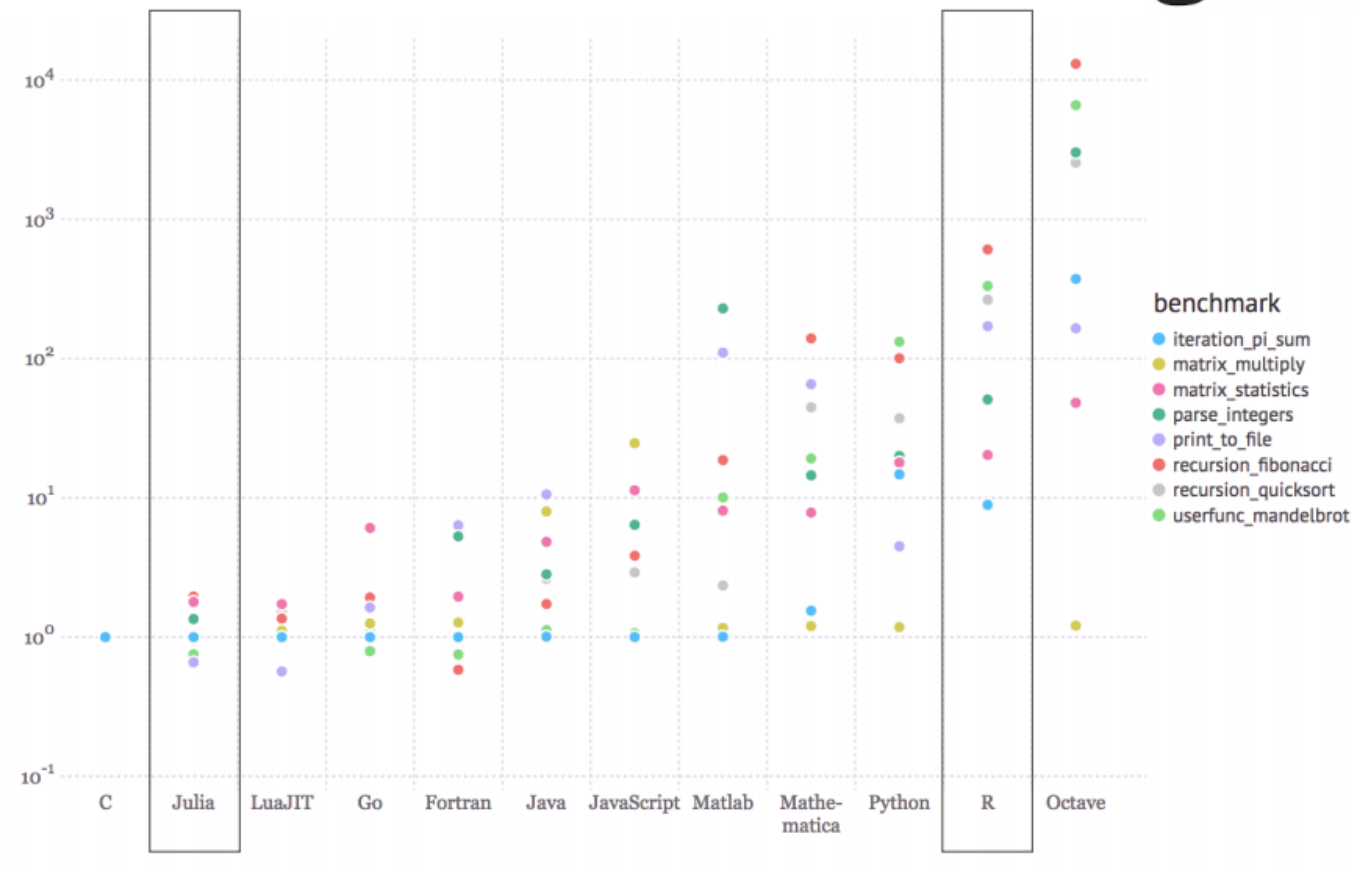
Advanced Tools

Online Communities

Jobs

- *interesting* language design
- good for statistics heavy projects
- large ecosystem of packages
- mature system
- native code rather slow

- clean language design
- number crunching
- developing ecosystem
- < v1.0
- native code fast

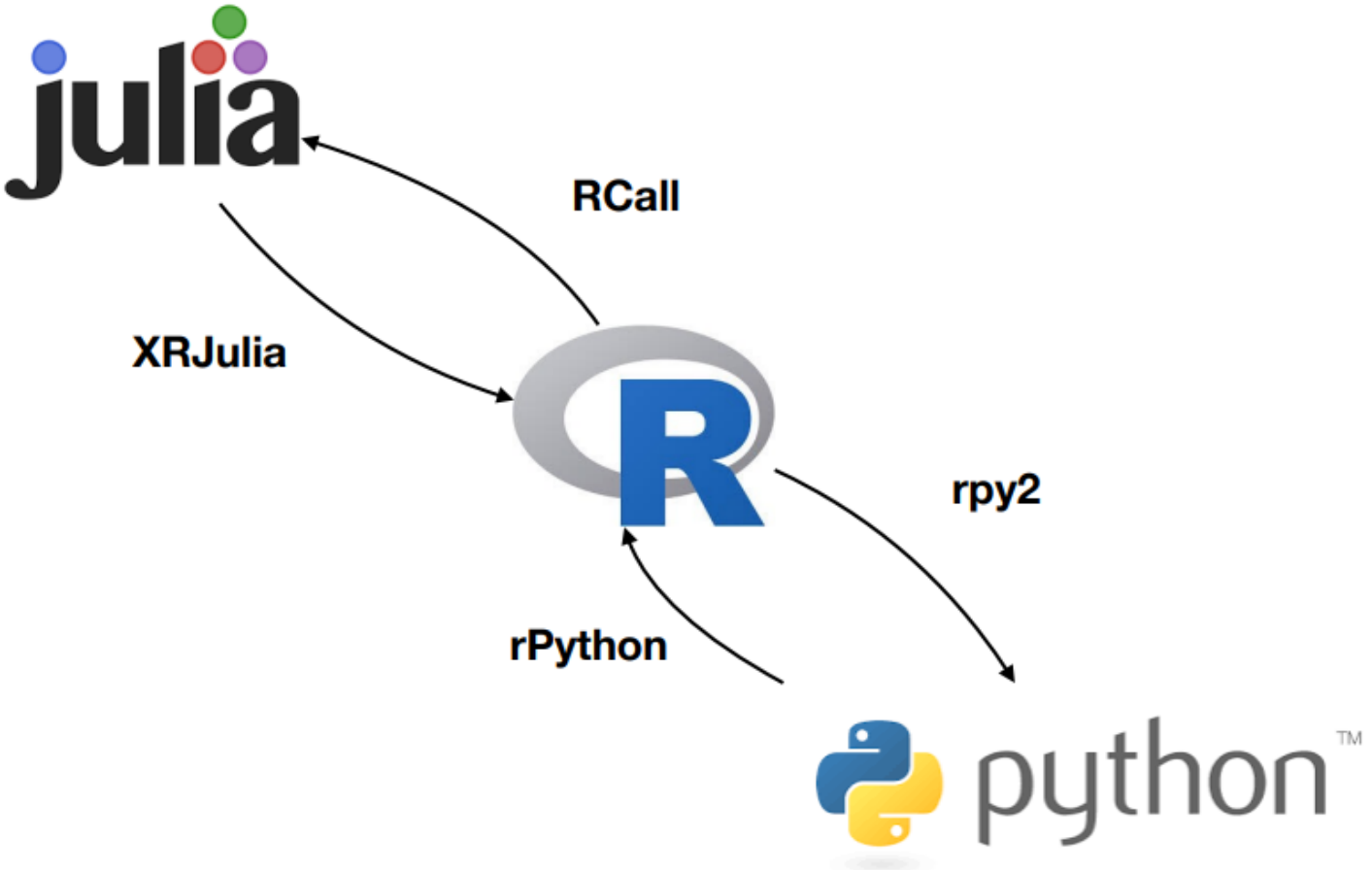

10^4
10^3
10^2
10^1
10^0
10^{-1}
C
Julia
LuaJIT
Go
Fortran
Java
JavaScript
Matlab
Mathe-
matica
Python
R
Octave
benchmark
iteration_pi_sum
matrix_multiply
matrix_statistics
parse_integers
print_to_file
recursion_fibonacci
recursion_quicksort
userfunc_mandelbrot

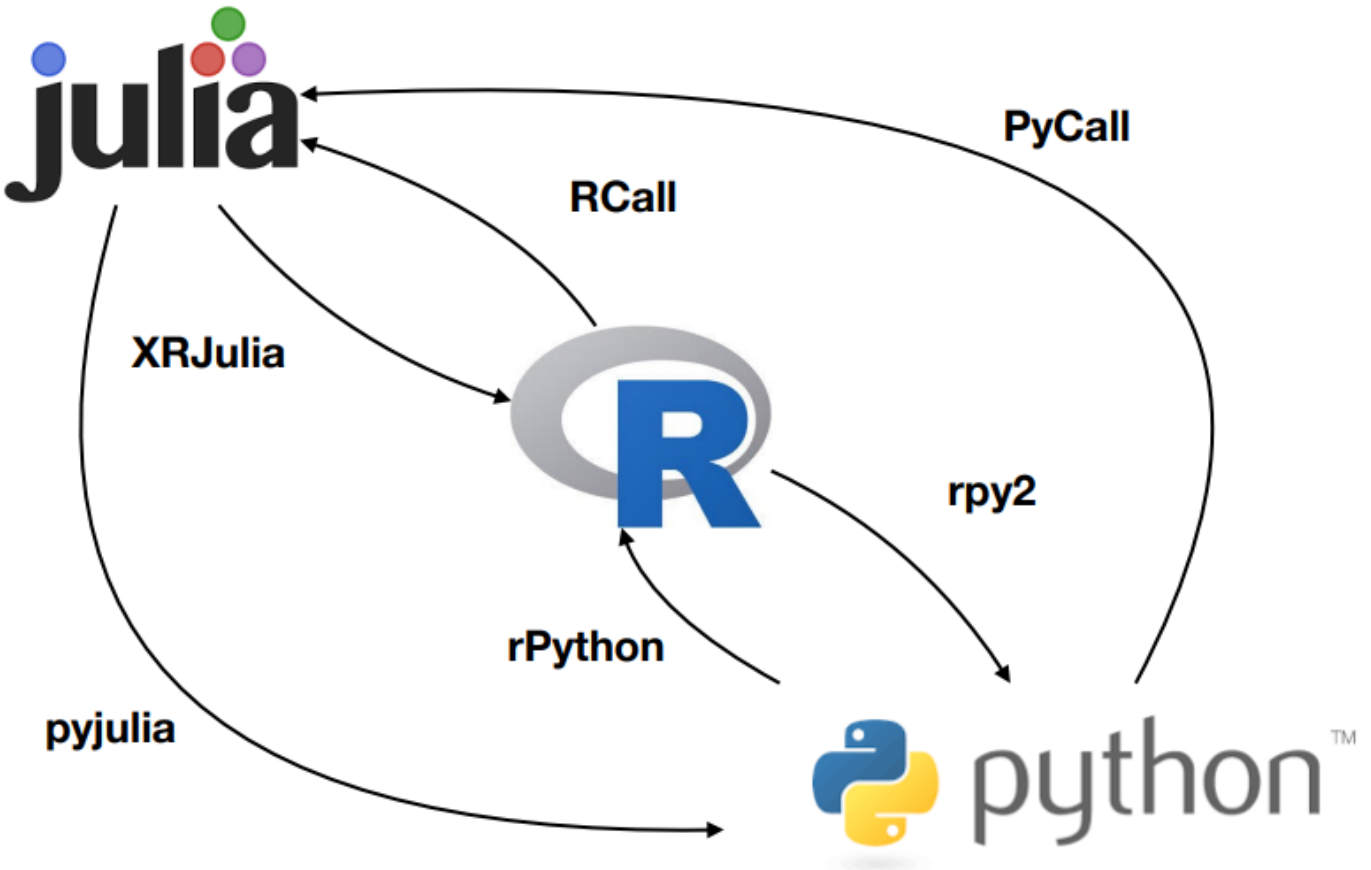
From my perspective

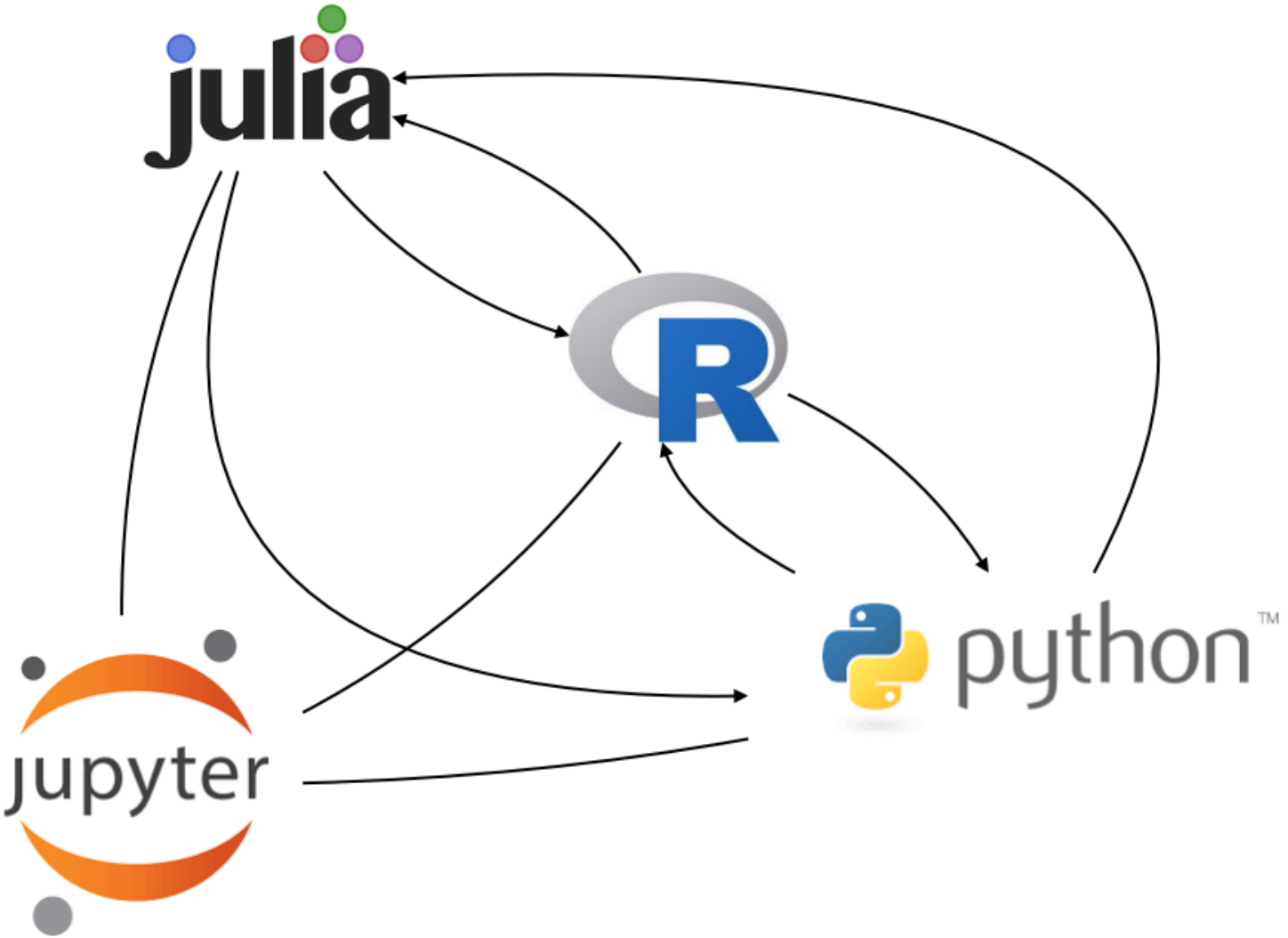
 for reliable statistical analysis (smaller scale)

 for larger data analysis/Deep Learning projects

 for new, experimental ideas (and fun)


julia
RCall
XRJulia
R
rpy2
rPython
python™

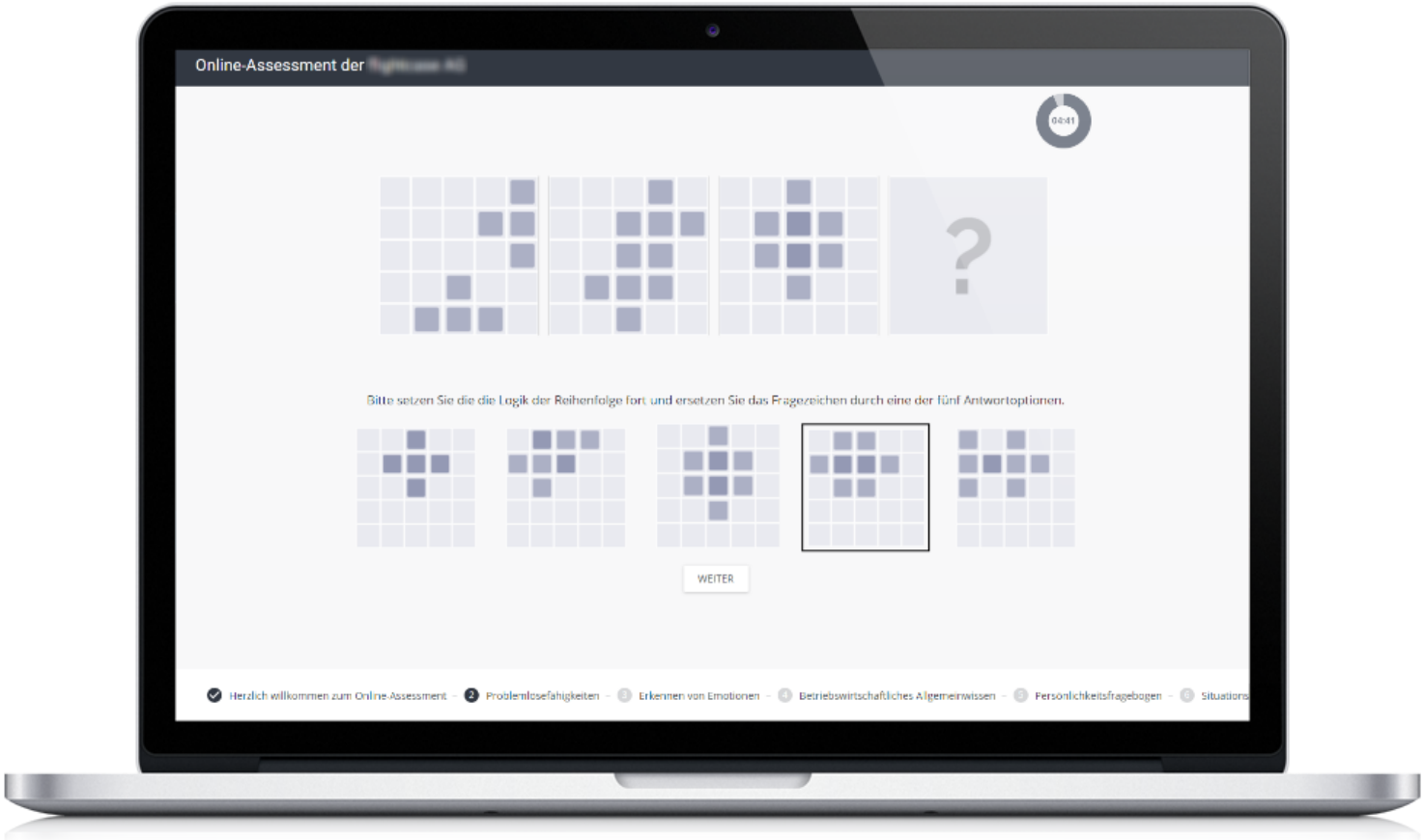

julia
PyCall
RCall
XRJulia
R
rpy2
rPython
pyjulia
python™


julia
R
python
jupyter

Real-world applications

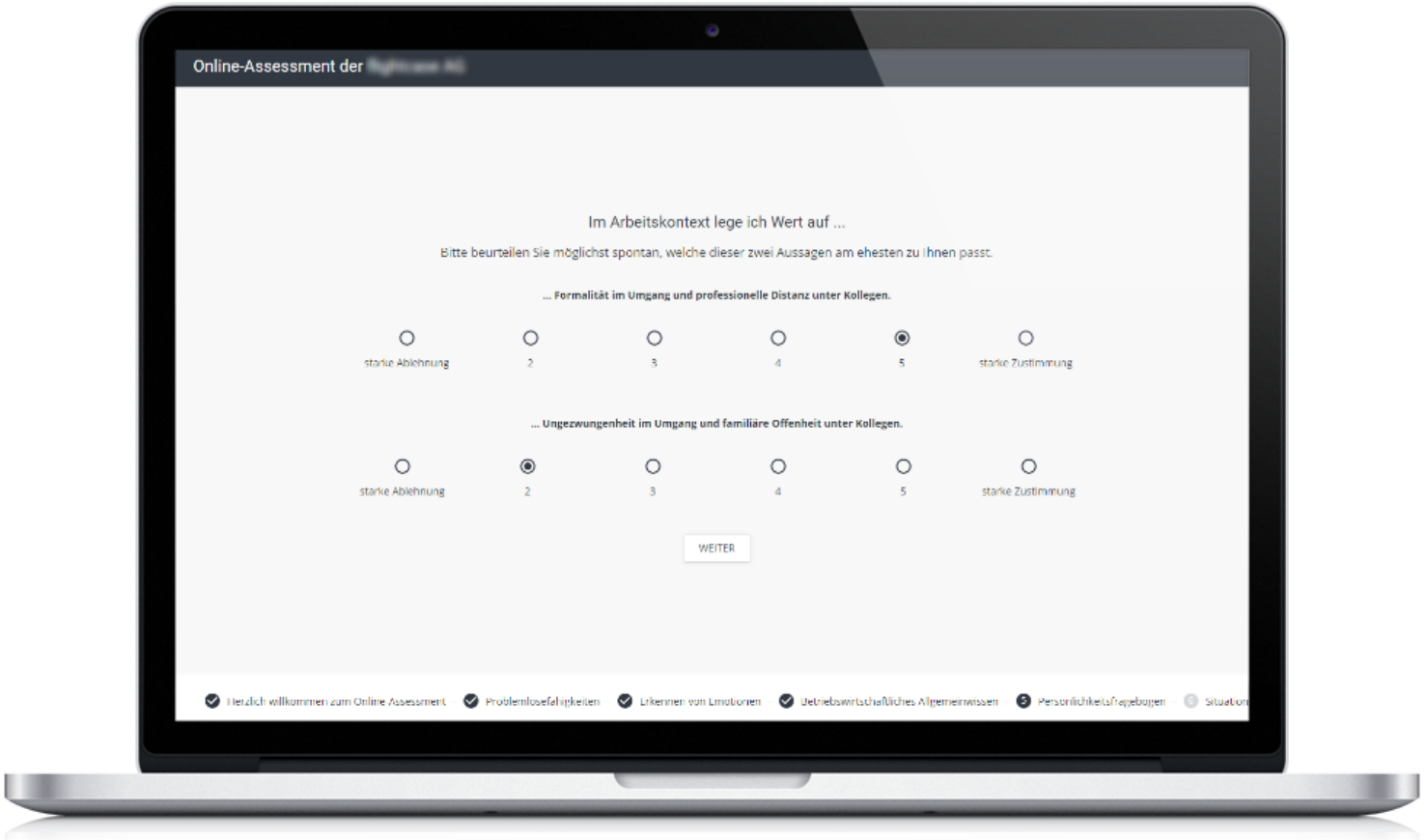
magnolia
psychometrics

Case 1: magnolia psychometrics predictive personality testing



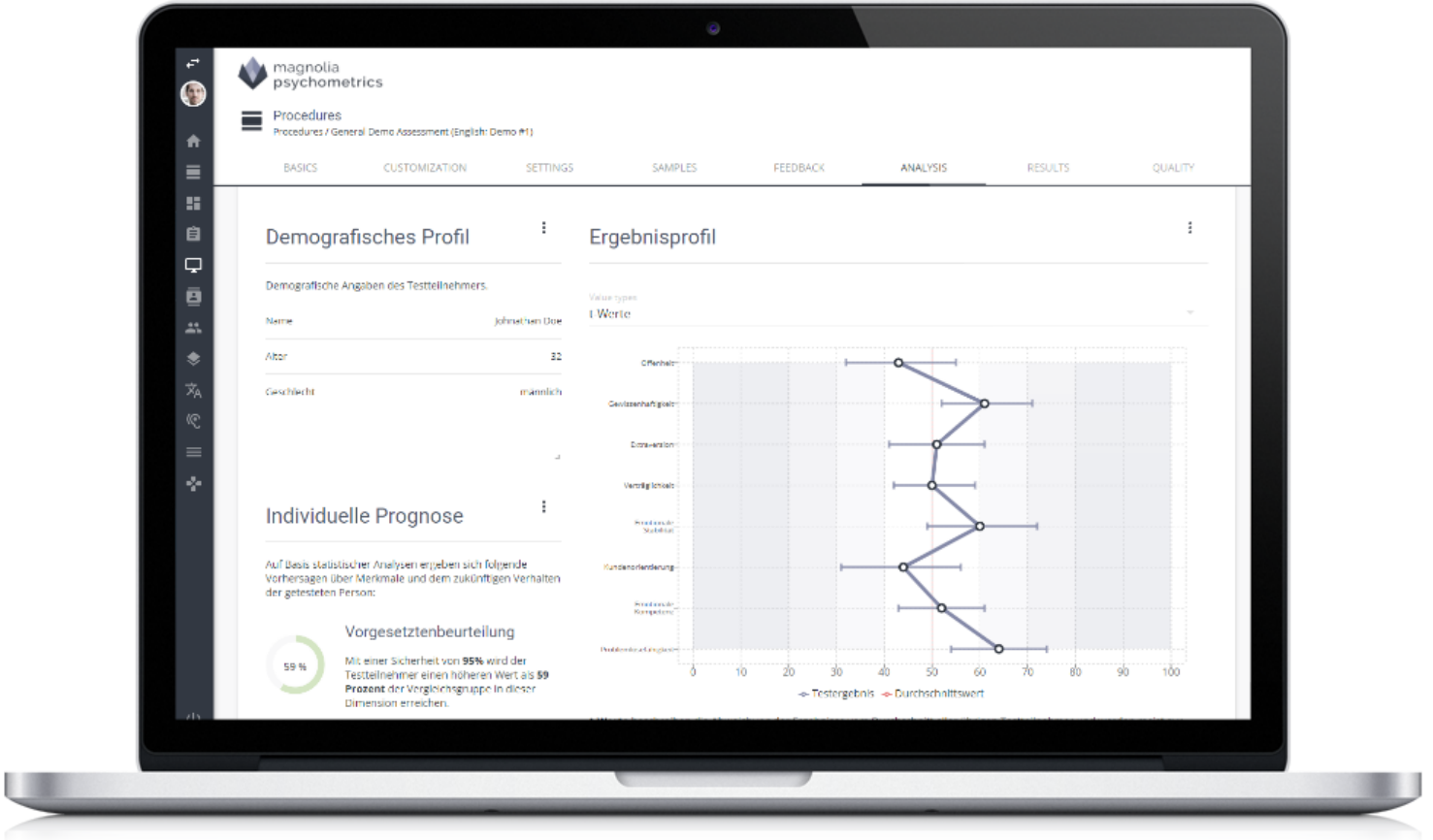
magnolia
psychometrics

Case 1: magnolia psychometrics predictive personality testing



magnolia
psychometrics

Case 1: magnolia psychometrics predictive personality testing



magnolia
psychometrics

Case 1: magnolia psychometrics predictive personality testing

Case 2: Using R to prevent food poisoning in Chicago

Foodborne illness in the U.S.

- **48 million** people get sick
- **128,000** get hospitalized
- **3,000** die

according to the center for disease control and prevention. The situation in the City of **Chicago** was particularly bad. Something had to be done...

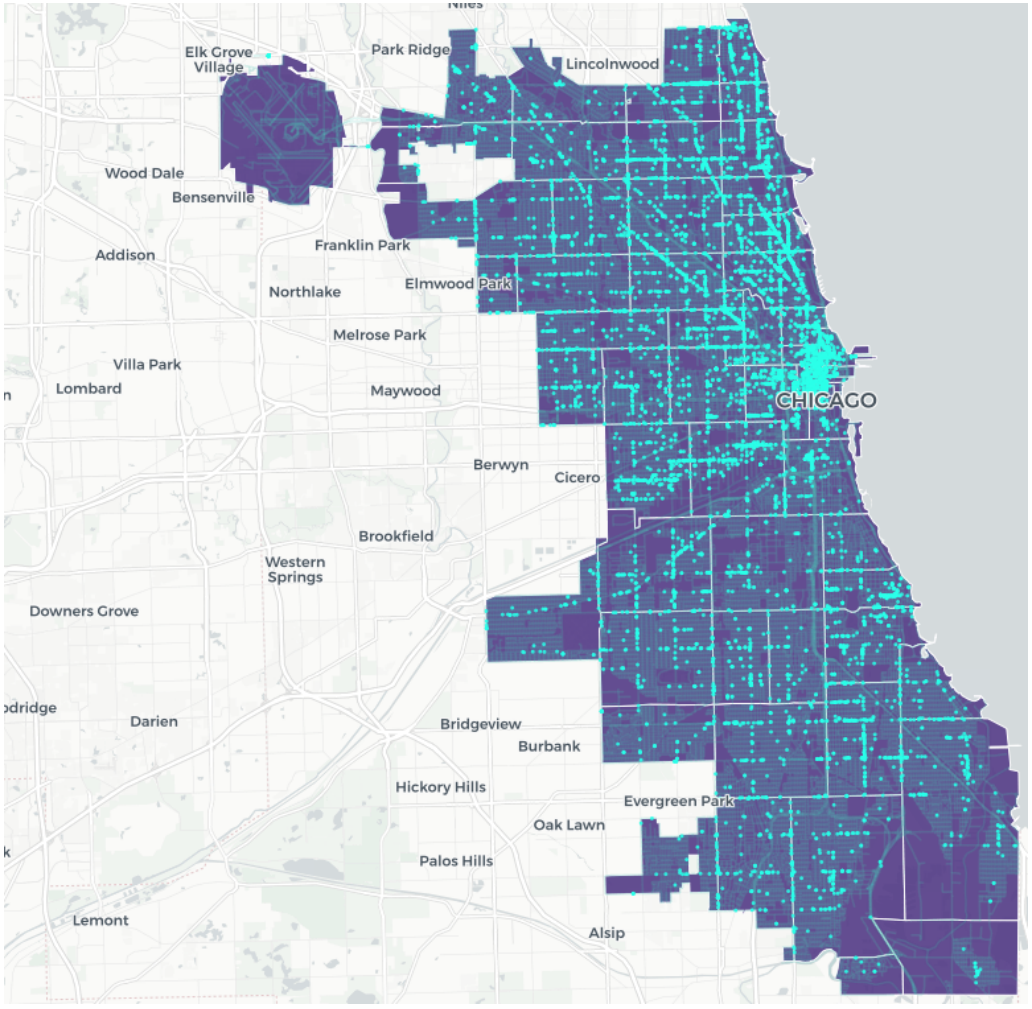
Case 2: Using R to prevent food poisoning in Chicago

To predict the probability of food establishmentsto have critical violations, data scientists looked at...

```
Business Licenses
Food Inspections
Crime
Garbage Cart Complaints
Sanitation Complaints
Weather
Sanitarian Information
```

Case 2: Using R to prevent food poisoning in Chicago

...and increased the rate of detecting violations by **25%**



Case 2: Using R to prevent food poisoning in Chicago

For data processing the code uses extensivelythe `data.table` package.

```
## Calcualte time since last inspection.
## If the time is NA, this means it's the first inspection; add an i
## variable to indicate that it's the first inspection.
dat_model[i = TRUE ,
          j = timeSinceLast := as.numeric(
              Inspection_Date - shift(Inspection_Date, -1, NA)) / 365
          by = License]
dat_model[ , firstRecord := 0]
dat_model[is.na(timeSinceLast), firstRecord := 1]
dat_model[is.na(timeSinceLast), timeSinceLast := 2]
dat_model[ , timeSinceLast := pmin(timeSinceLast, 2)]
```

All code available on GitHub

R-Meetups in 2018

What to expect from R-Meetups

Goals

- Networking
- Interdisciplinary Insights
- Practical Learning

What to expect from R-Meetups

Goals

- Networking
- Interdisciplinary Insights
- Practical Learning

Outline

Topic introduction	*see scheduled topics*	(~15 mins)
Code practice	*attempt to solve the problem in small groups*	(~60 mins)
Stand-ups	*tell us about your projects and/or probelms*	-
Open end	*code, network or drink beer*	-

Dataset

Speed Dating Experiment: *What attributes influence the selection of a romantic partner?*

Data from experimental speed dating events from 2002-2004, availible on kaggle (N ~ 8,000 observations)

Dataset

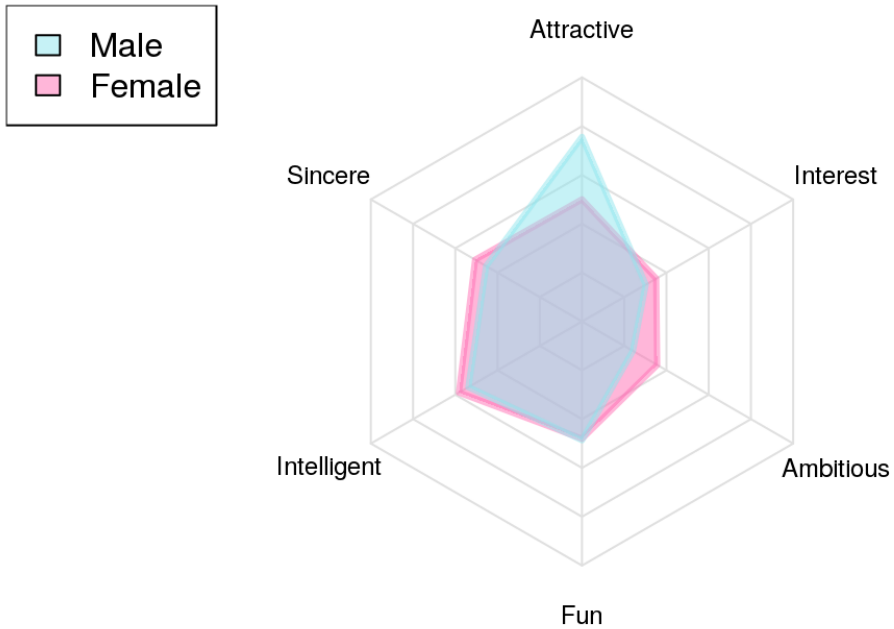
Speed Dating Experiment: *What attributes influence the selection of a romantic partner?*

Data from experimental speed dating events from 2002-2004, availible on kaggle (N ~ 8,000 observations)

Data Exploration Ideas

- What are the least desirable attributes in a male partner? Does this differ for female partners?
- Can people accurately predict their own perceived value in the dating market?
- ...

"The Ugly Truth of People Decisions in Speed Dating"



Insighty by James Hwang, Lucas Cadalzo

Save the date

topic	date	moderator
Data Visualization	16th April, 19:00	by Mandy Vogel
Exploratory Data Analysis	21th May, 19:00	by Björn Hommel
Machiene Learning	18th June, 19:00	by Nico Scherf
RMarkdown	16th July, 19:00	by Valentin Stefan
yet to be decided	-	by *you?*
yet to be decided	-	by *you?*
yet to be decided	-	by *you?*

- Volunteers?

Resources

- Updates about meetups on meetup.com
- r-leipzig's projects & presentations on GitHub
- r-leipzig on slack for r-related exchange
- Speed-Dating dataset on kaggle

Thanks!