

Data Science

Introduction

Gero Szepannek



Data

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)



Artwork: Tamar Cohen, Andrew J Buboltz, 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Liebe @apoBank, bei uns sitzen nur Einsen,
keine Nullen...! 😊 Viele Grüße aus #Duisburg



03:21 - 13. März 2019

4 Retweets 30 „Gefällt mir“-Angaben



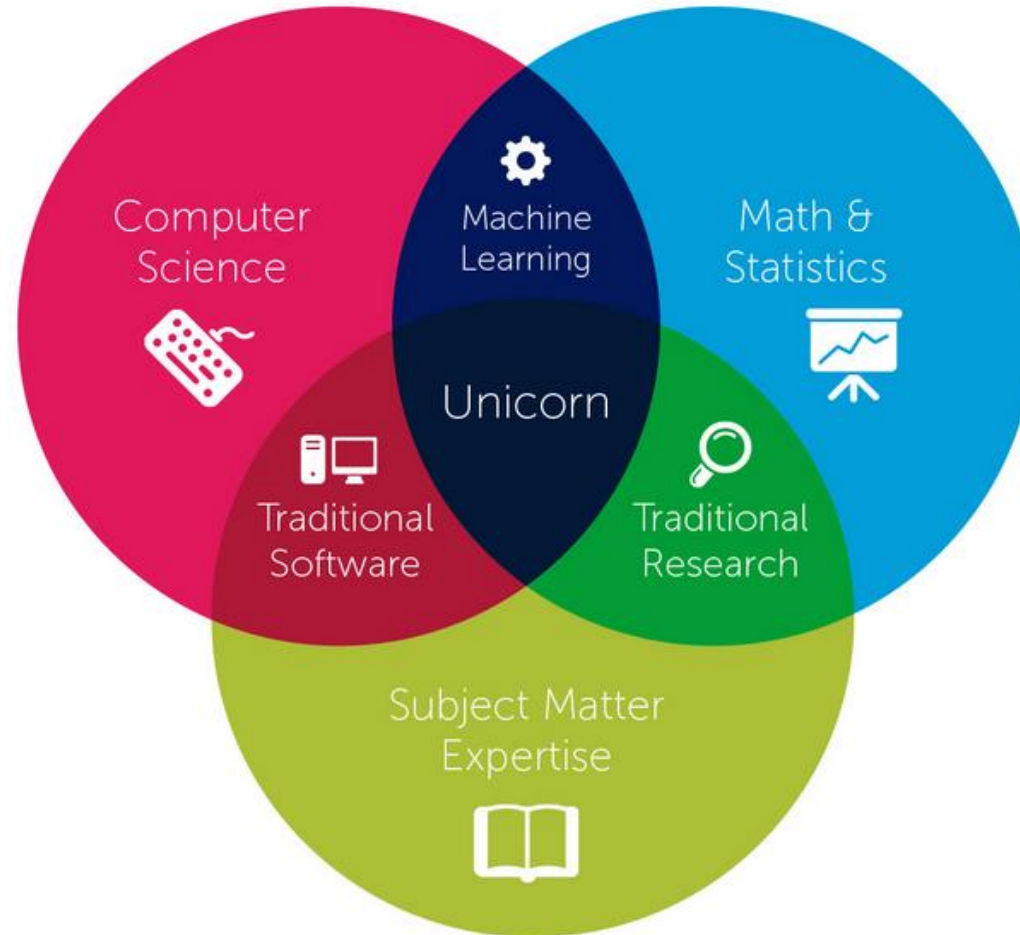
apoBank @apoBank · 13. März

Antwort an @TARGOBANK

Wir freuen uns auch über Eure Einsen 😊



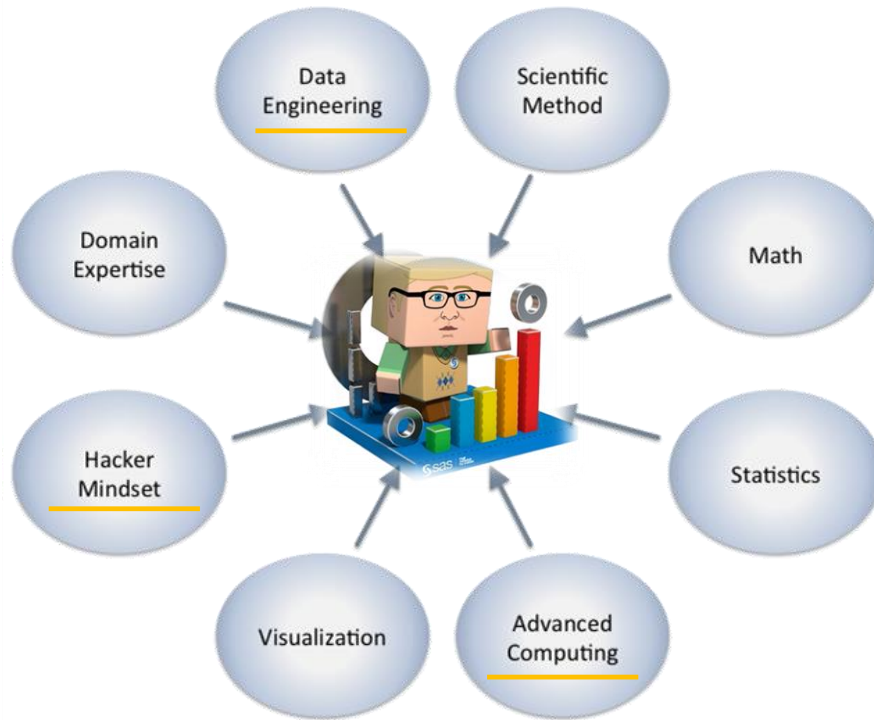
Data Science



<https://indico.io/blog/data-science-buzzwords-demystified/>

Kompetenzen im Data Science Team müssen (weiter)entwickelt werden!

Der „Data Scientist“ als Schlüsselrolle



Wenn es um die Anwendung neuester Analysemethoden geht, braucht es neue Kompetenzrollen.

Weitere wichtige Rollen



Prozessexperte

- Hohes Fachwissen
- Definition der Problemstellung
- Interpretation der Ergebnisse



Datenexperte

- Hohes Datenverständnis
- Verknüpfung von Fachwissen und Daten
- Datenvorbereitung und -aufbereitung



Business Analyst

- Verständnis der Geschäftszusammenhänge
- Erfahrung mit explorativer Datenanalyse
- Statistisches Grundverständnis
- Visualisierungs- und Kommunikationstechniken

Data Science Team bestehend aus 15 internen Experten und einem externen Coach



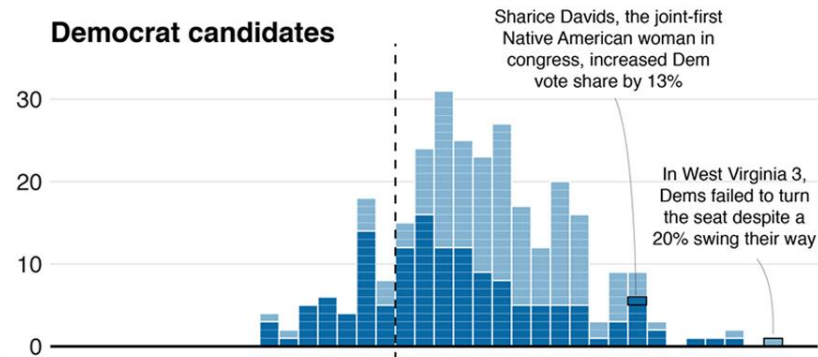
If statistics programs/languages were cars...



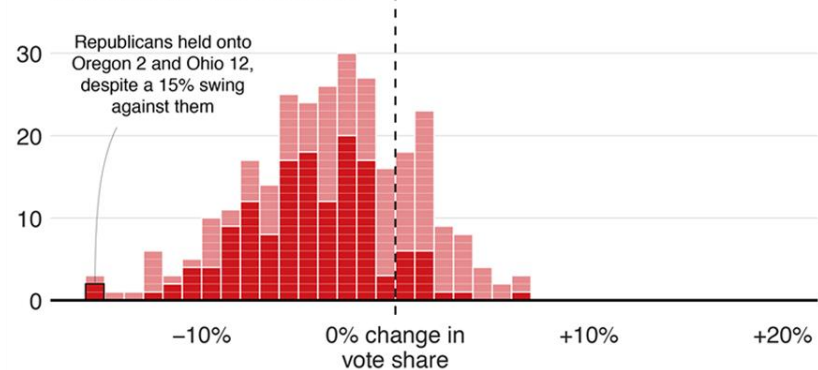
Blue wave

■ Won seat ■ Didn't win

Democrat candidates



Republican candidates

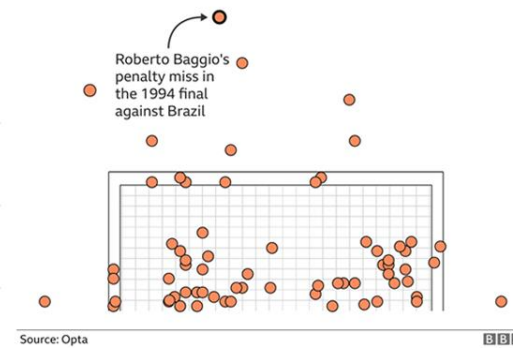


Source: AP, 19:01 ET

BBC

Where penalties are saved

World Cup shootout misses and saves, 1982-2014

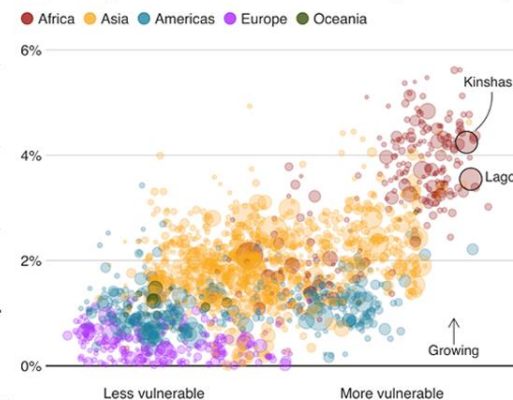


Source: Opta

BBC

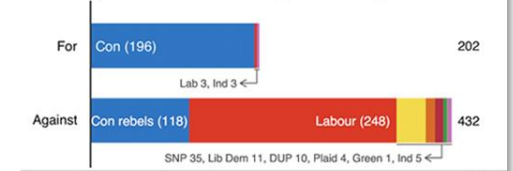
Fast-growing cities face worse climate risks

Population growth 2018-2035 over climate change vulnerability



Source: Verisk Maplecroft. Circle size represents current population.

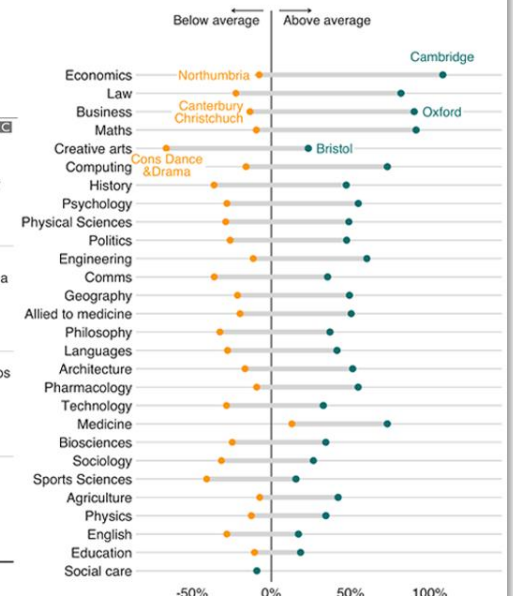
MPs rejected Theresa May's deal by 230 votes



Source: Commons Votes Services. Excludes 'tellers', the Speaker and deputies

Earnings vary across units even within subjects

Impact on men's earnings relative to the average degree



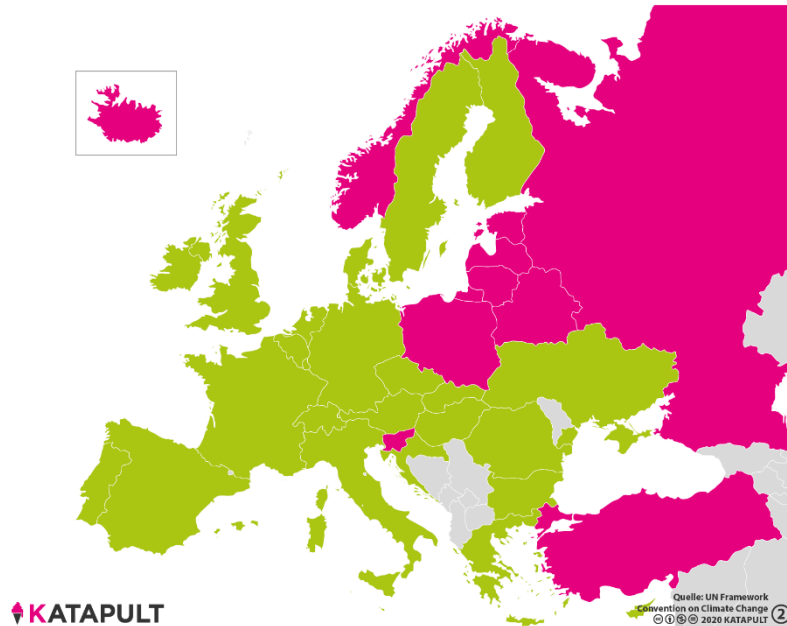
Source: Institute for Fiscal Studies

BBC

<https://bbc.github.io/rcookbook/>

Treibhausgas-Emissionen

Länder, die 2018 **mehr** / **weniger** Treibhausgase ausstießen als 2003



Säuger mit dem **dicksten Fell**

Wer es hat



FISCHOTTER

50.000

Haare pro cm²

Wer es braucht



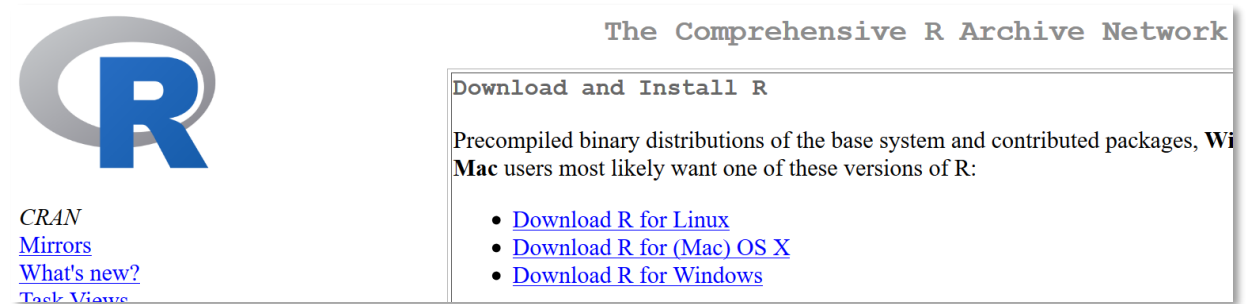
CHRISTIAN DROSTEN

120

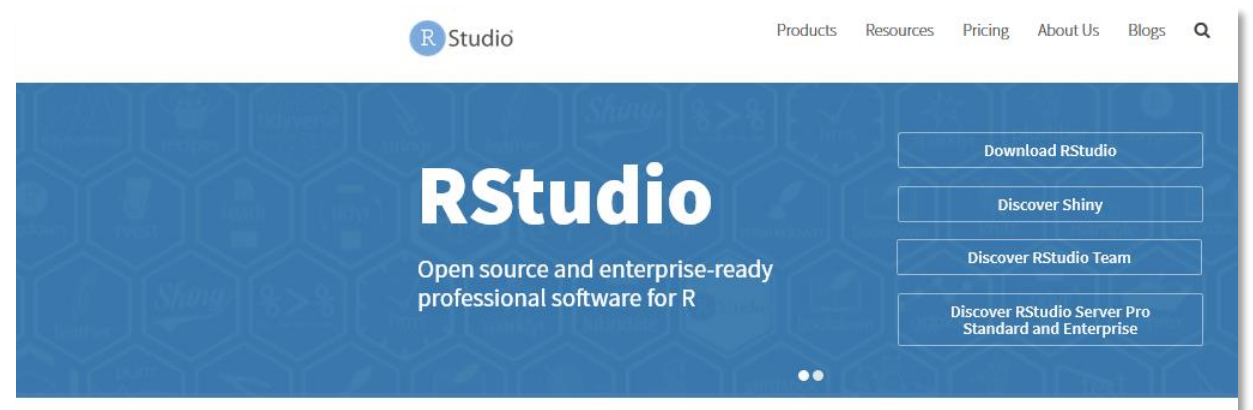
Haare pro cm²

Installation

<https://www.r-project.org/>

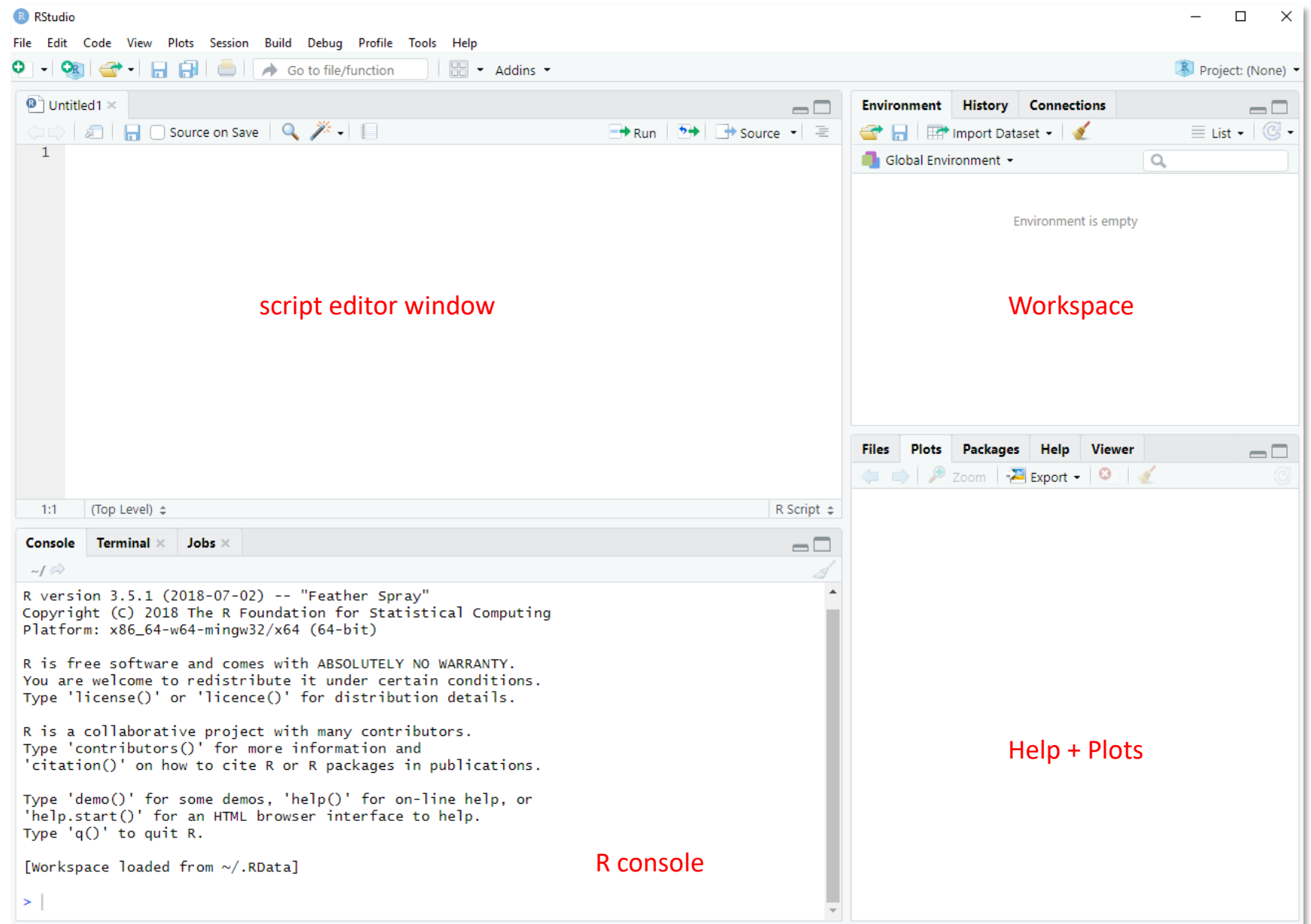


<https://www.rstudio.com/>



R Studio

<CTRL> + <Enter>
to execute code



Git and Github...



git



Happy Git and GitHub for the useR

Jenny Bryan, the STAT 545 TAs, Jim Hester

Let's Git started



<https://happygitwithr.com/>

Pull, Push and Commit...

In case of fire



1. git commit



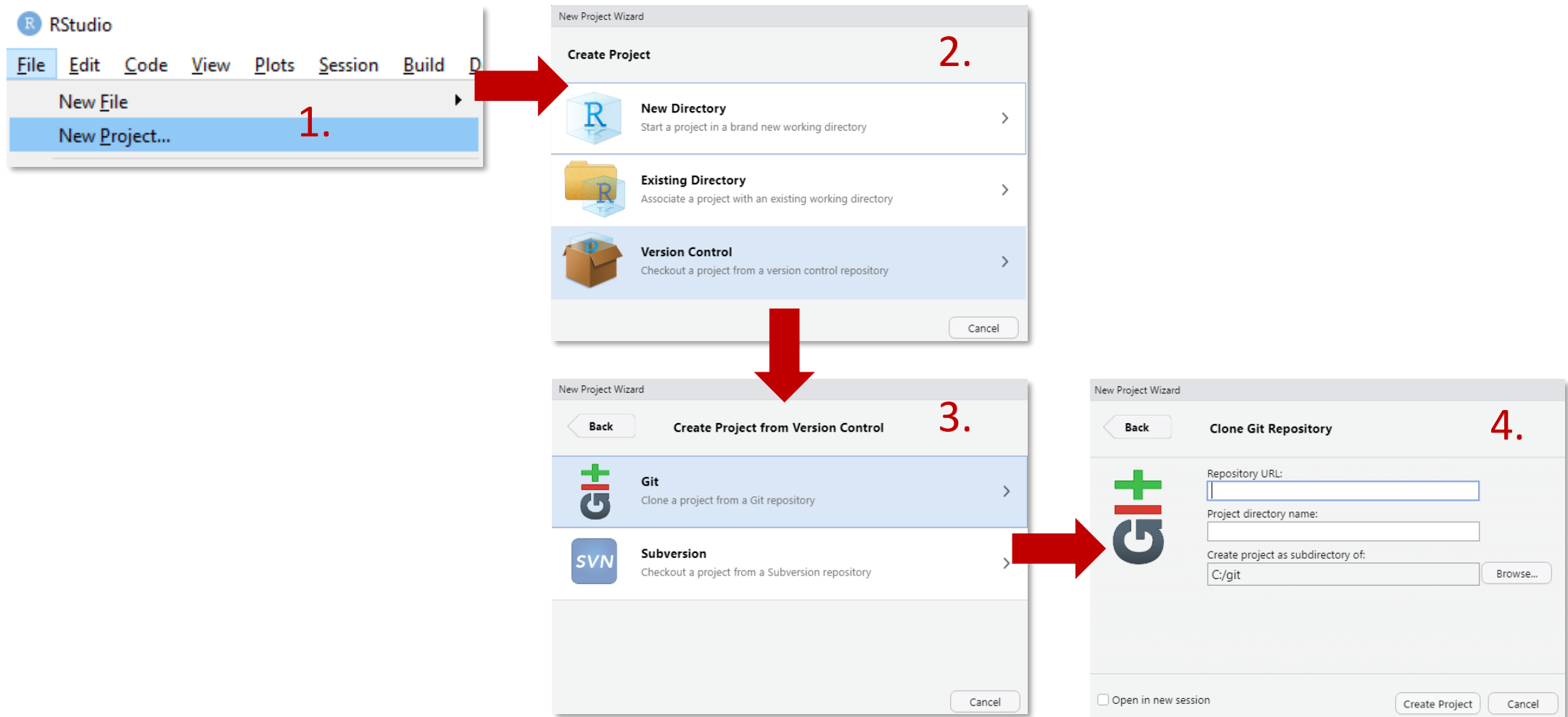
2. git push



3. leave building

<https://medium.com/mindorks/what-is-git-commit-push-pull-log-aliases-fetch-config-clone-56bc52a3601c>

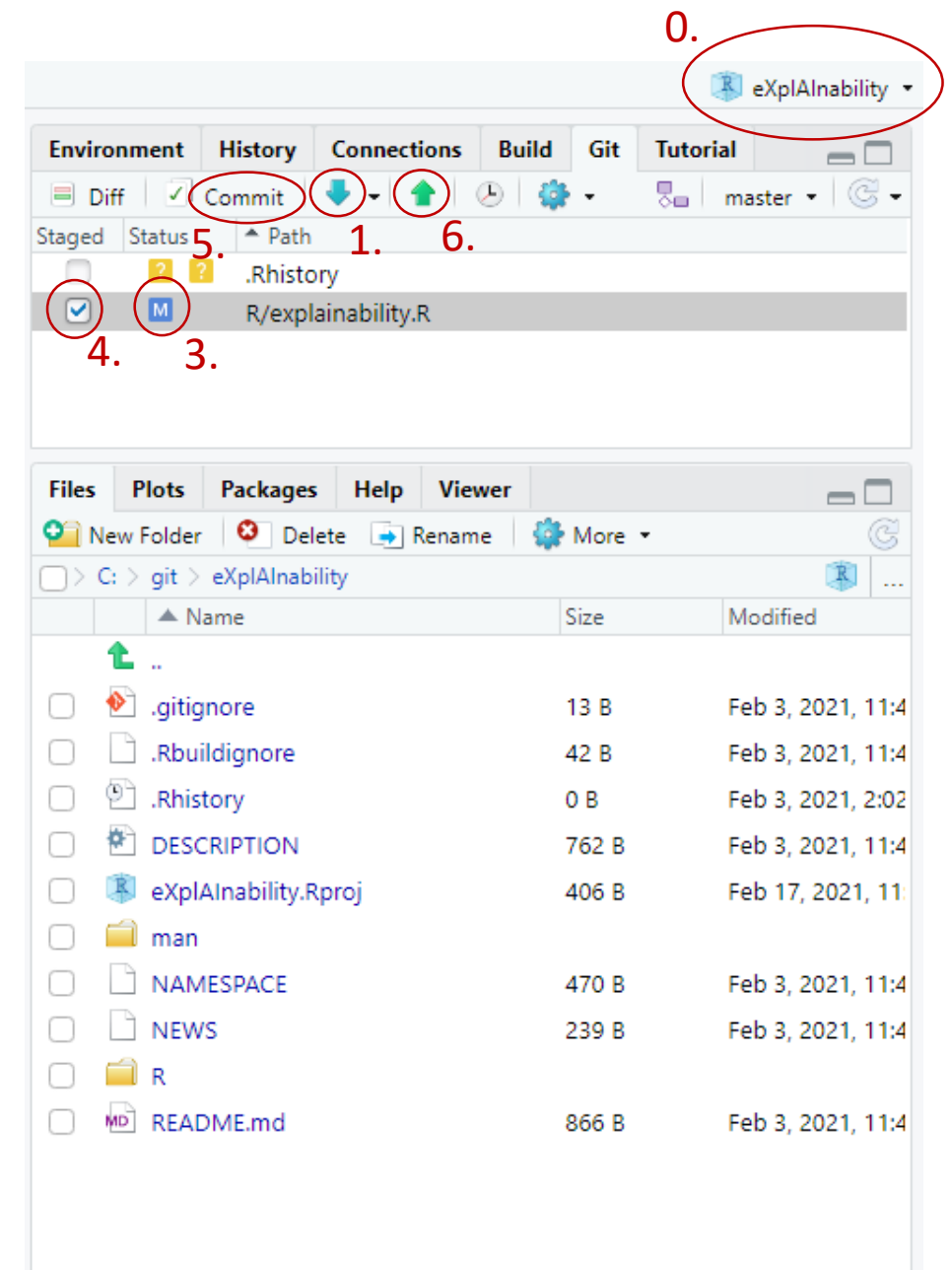
Creating a local copy of a Project from Github



Workflow for using Git

...from out of RStudio

0. Open project
1. Pull latest version from github
2. ...work...
3. ...modified files appear
4. Mark modified files that should be committed
5. Commit (+ add comment if requested)
6. Push to Github



Prüfungsleistung

1. Exploratory Data Analysis Presentation

- I. Assignment of teams ←
- II. Select a package for auto EDA from article
- III. Present package to the course
- IV. Code demo

2. Data Wrangling Task

- I. Assignment of teams
- II. Conduct EDA
- III. Mail executable *.Rmd with your findings

3. Predictive Modelling Project

- I. Assignment of Teams
- II. Kaggle competition
- III. Deployment of your model
- IV. Presentation of your results

kaggle

CONTRIBUTED RESEARCH ARTICLE

347

The Landscape of R Packages for Automated Exploratory Data Analysis

by Mateusz Staniak and Przemysław Biecek

Abstract The increasing availability of large but noisy data sets with a large number of heterogeneous variables leads to the increasing interest in the automation of common tasks for data analysis. The most time-consuming part of this process is the Exploratory Data Analysis, crucial for better domain understanding, data cleaning, data validation, and feature engineering.

There is a growing number of libraries that attempt to automate some of the typical Exploratory Data Analysis tasks to make the search for new insights easier and faster. In this paper, we present a systematic review of existing tools for Automated Exploratory Data Analysis (autoEDA). We explore the features of fifteen popular R packages to identify the parts of analysis that can be effectively automated with the current tools and to point out new directions for further autoEDA development.

<https://journal.r-project.org/archive/2019/RJ-2019-033/index.html>

Aufgaben (Task 1):

1. Datum: 6.5.2021
2. Machen Sie sich mit dem Paket vertraut
3. Funktionalitäten vorstellen
4. Besonders Nützliche Funktion in R demonstrieren
5. Pull request: *.R File mit R Code + Folien
6. Vortragsdauer: 10 Minuten / Team

Assignment of Teams...

Team	# 1	# 2	Paket
	1 Susann Kindler	Paul Fetscher	arsenal
	2 Luca Jetter	Jonas Gagelmann	summarytools
	3 Lena Fredrich	Felix Jaworski	dataMaid
	4 Johannes Krecklow	Lindemann, Timo	DataExplorer
	5 Sebastian Bergemann	Alexander Voltz	visdat
	6 Alexander Storbeck	Marcus Pradel	funModeling