HOST
Hochschule Stralsund

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

TARGOBANK
@TARGOBANK

Folgen

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

03:21 - 13. März 2019

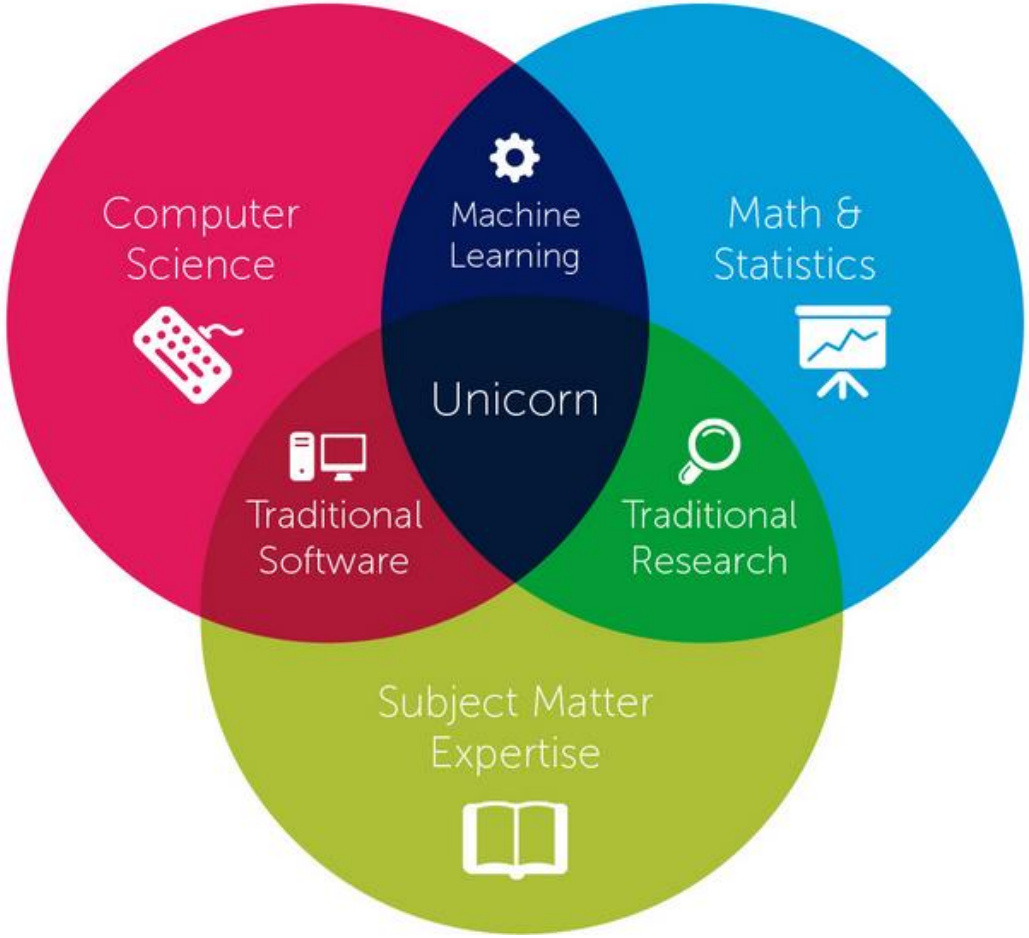
4 Retweets **30** „Gefällt mir"-Angaben

2 4 30

apoBank @apoBank · 13. März
Antwort an @TARGOBANK
Wir freuen uns auch über Eure Einsen 😉

11

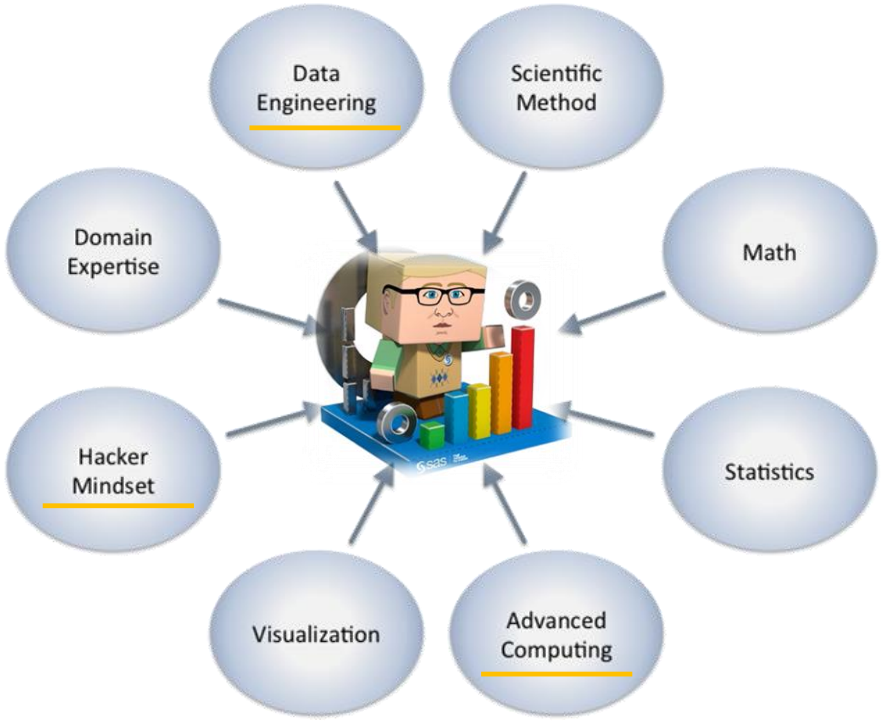
Data Science



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 neuster 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

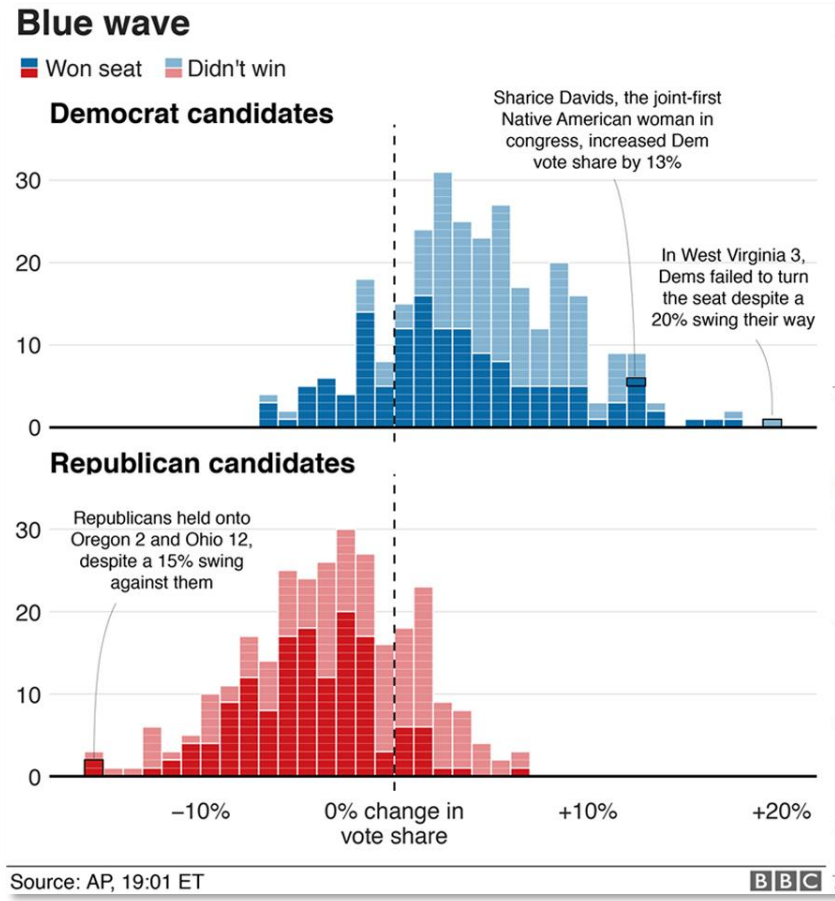
- Verständnis der Geschäftszusammenhänge
- Erfahrung mit explorativer Datenanalyse
- Statistisches Grundverständnis
- Visualisierungs- und Kommunikations-techniken

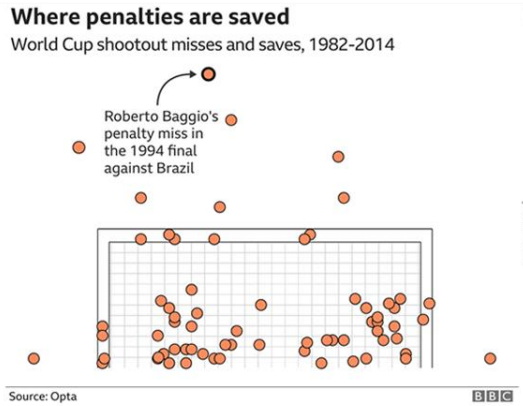
Data Science Team bestehend aus 15 internen Experten und einem externen Coach

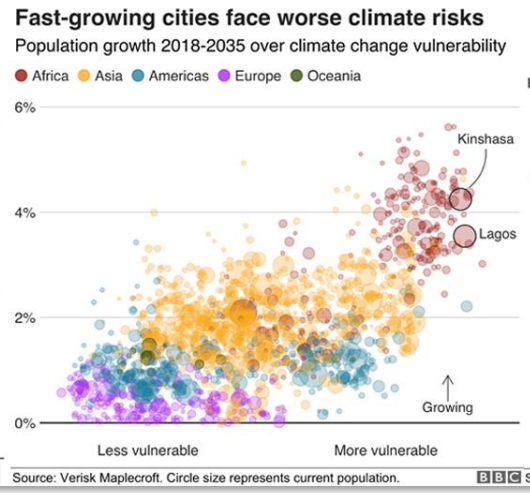
Fraport

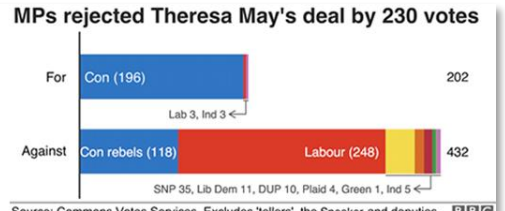
If statistics programs/languages were cars...

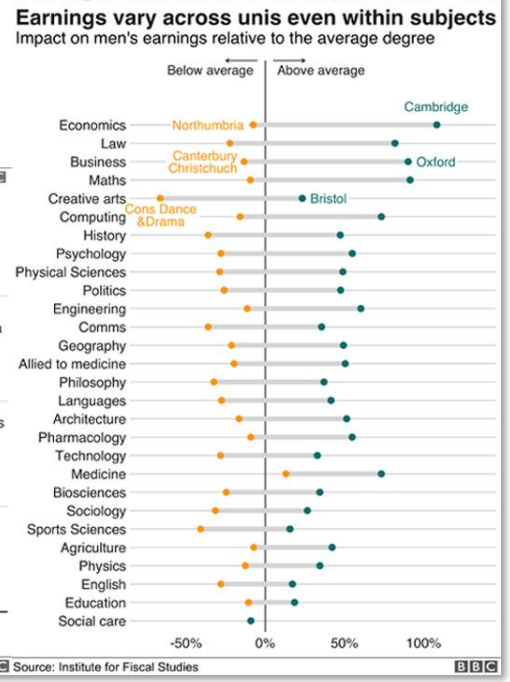
BBC NEWS







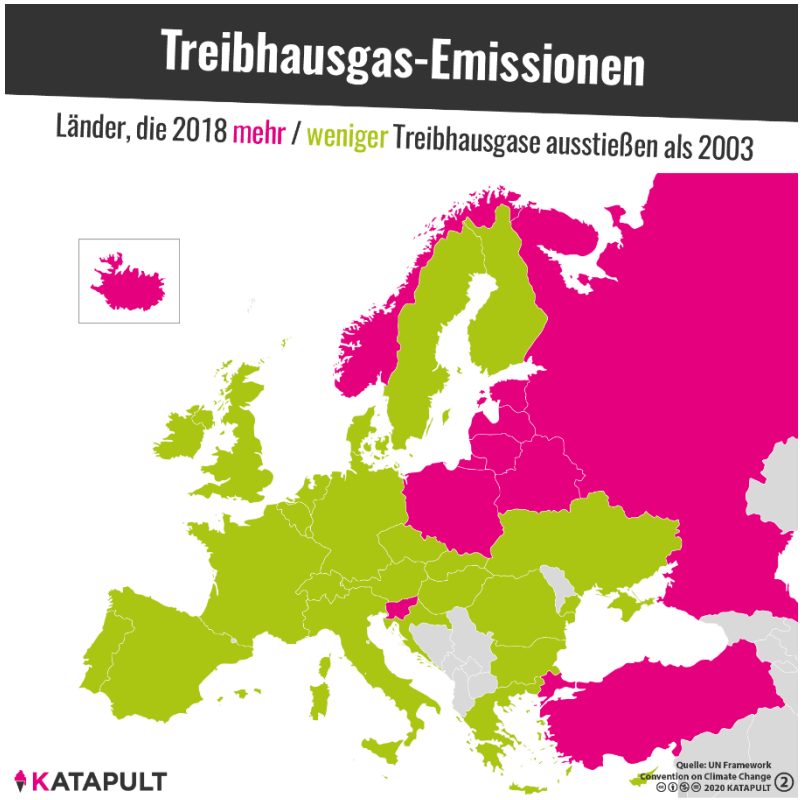




https://bbc.github.io/rcookbook/

KATAPULT

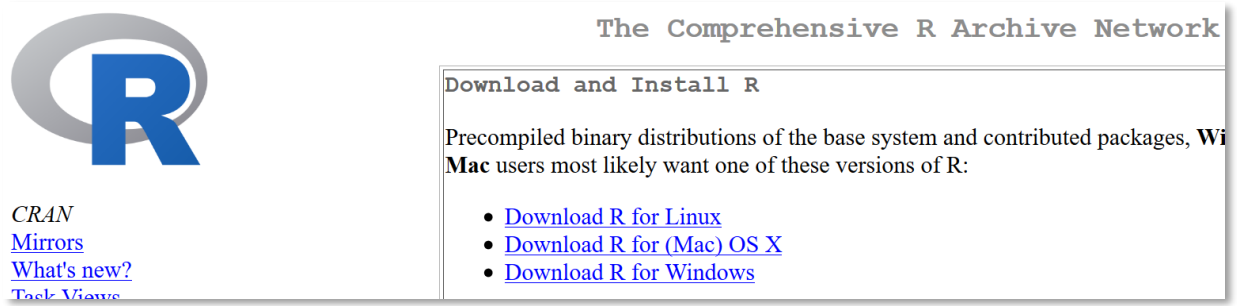
Magazin für ~~Eis,~~ Kartografik und Sozialwissenschaft



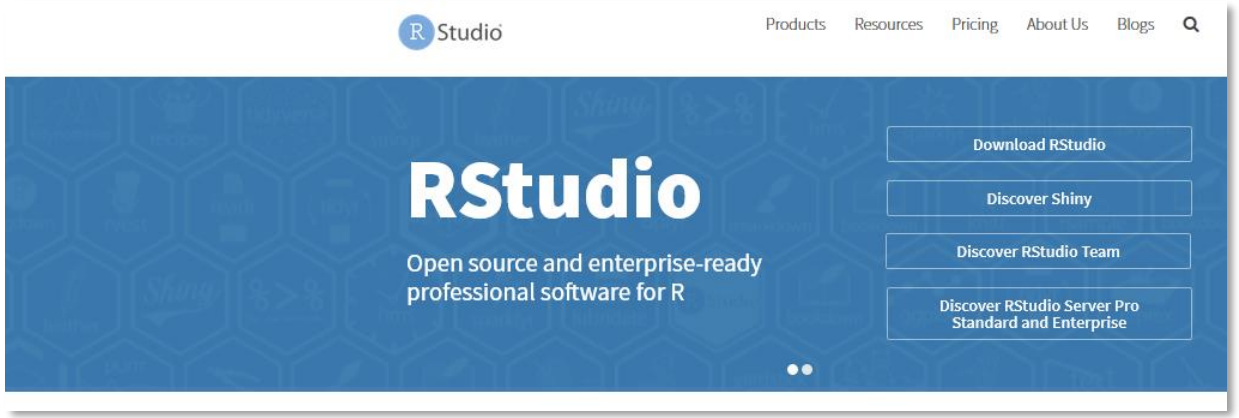
https://katapult-magazin.de

Installation

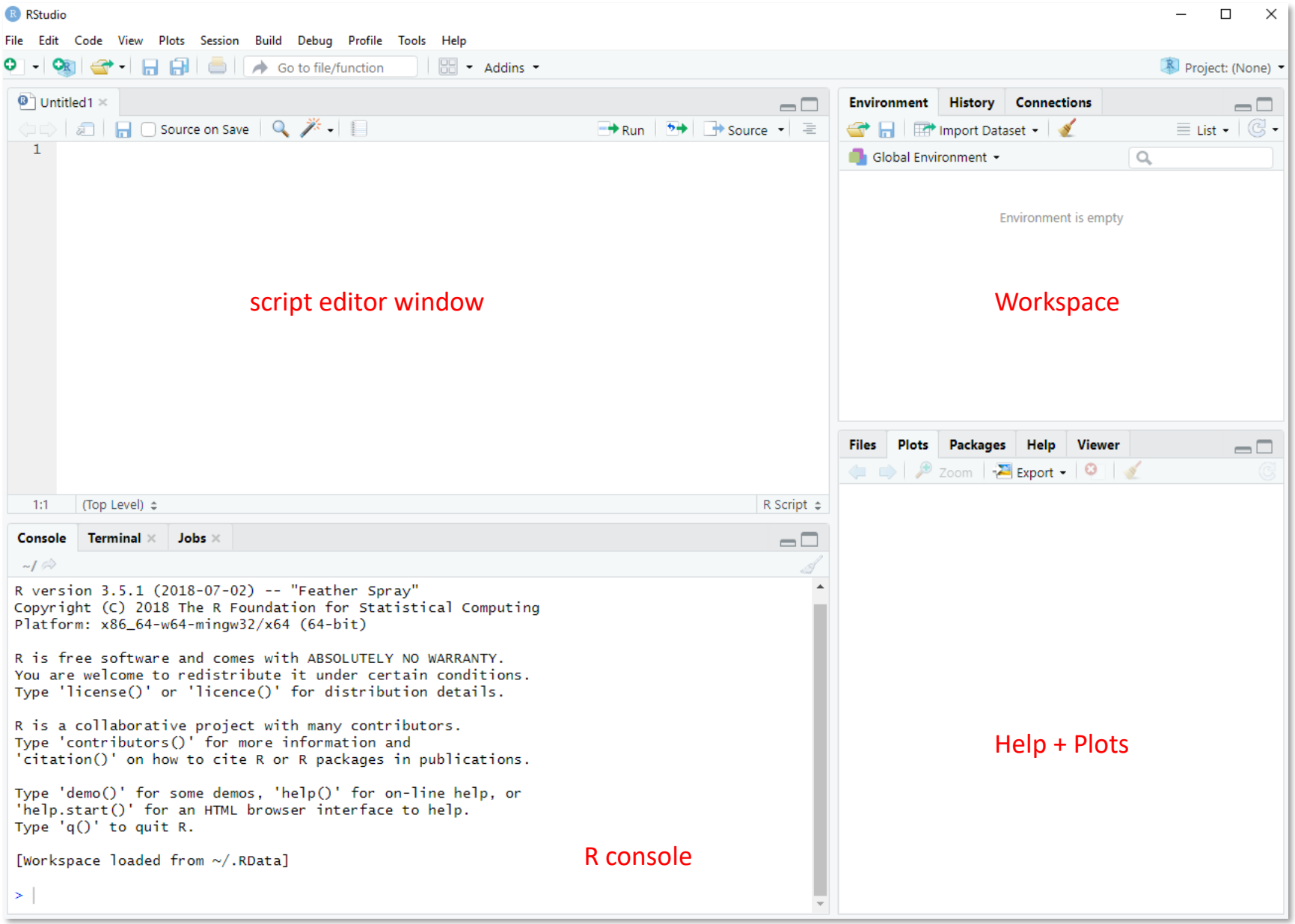
https://www.r-project.org/



https://www.rstudio.com/



R Studio



<CTRL> + <Enter>
to execute code

Git and Github…

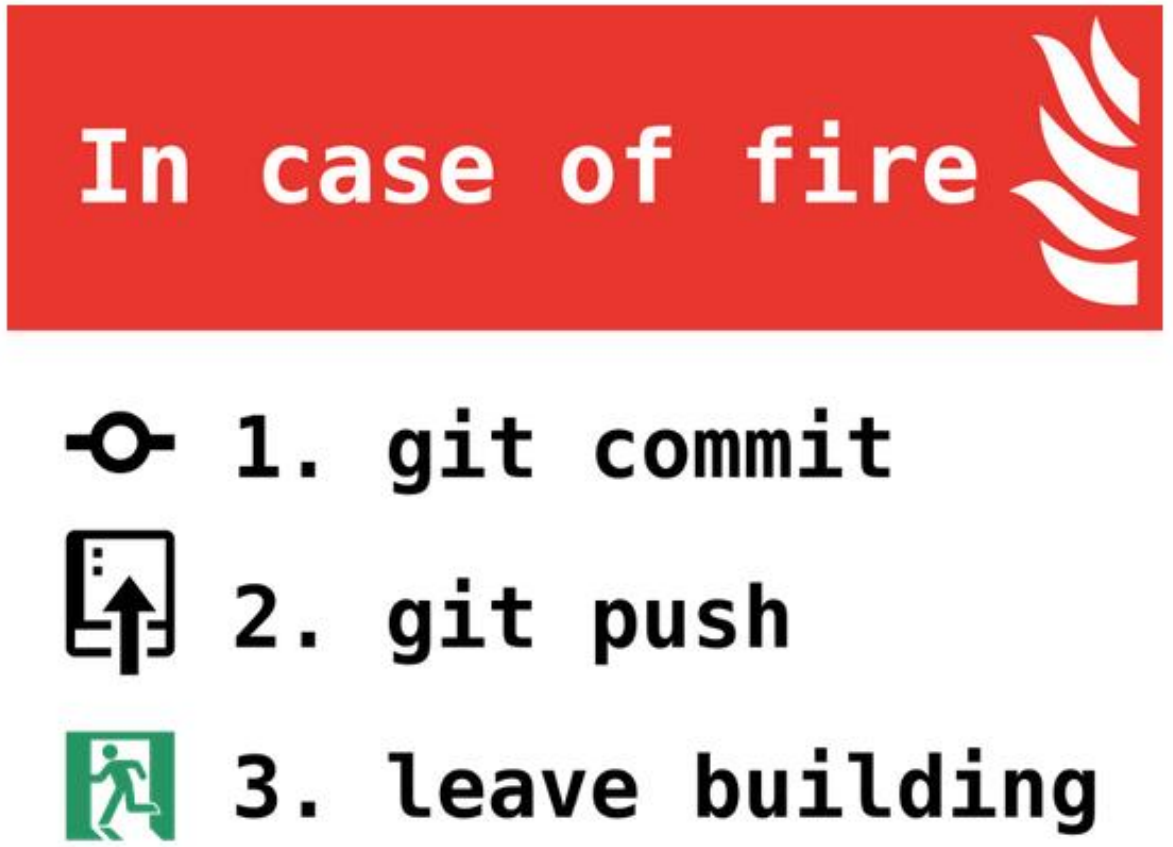
Happy Git and GitHub for the useR

Jenny Bryan, the STAT 545 TAs, Jim Hester

Let's Git started

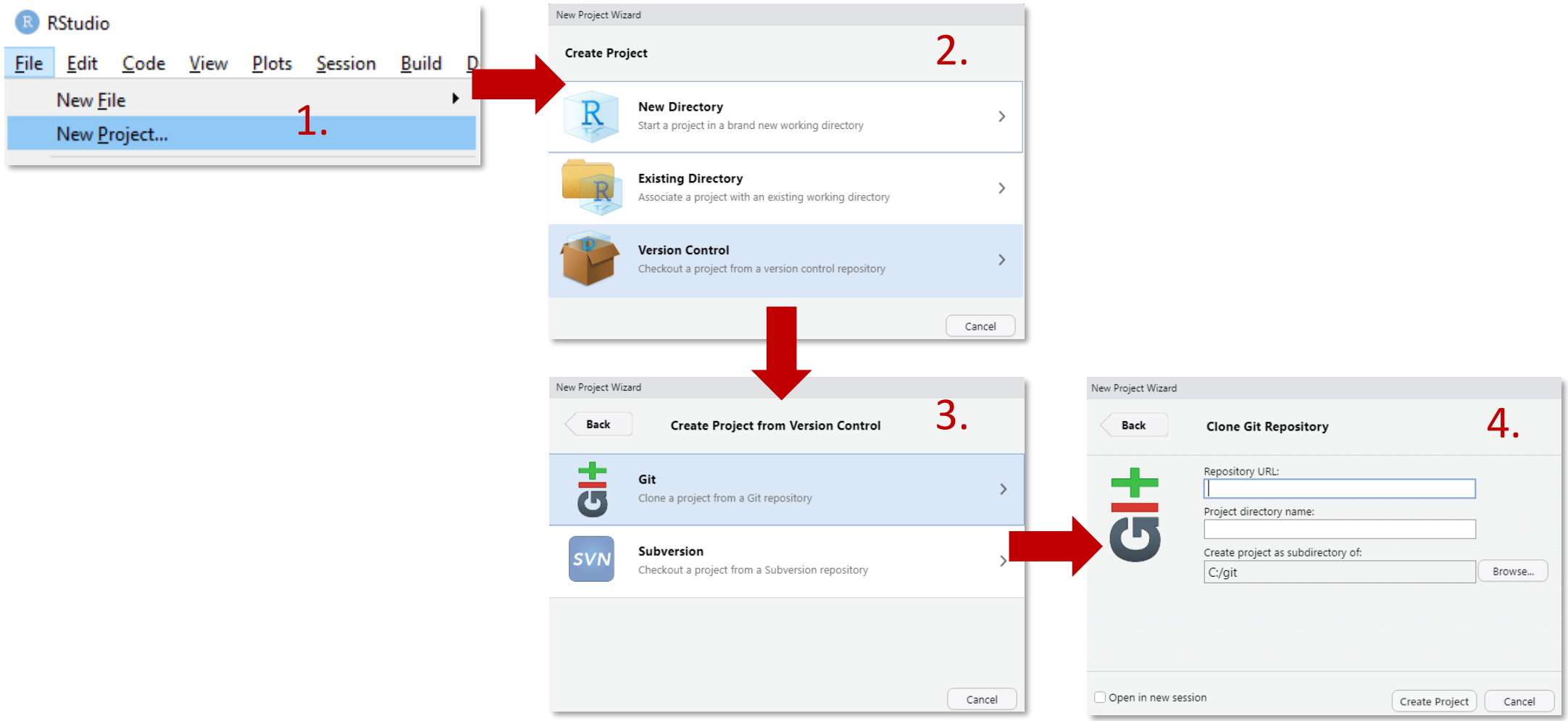
https://happygitwithr.com/

Pull, Push and Commit…



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

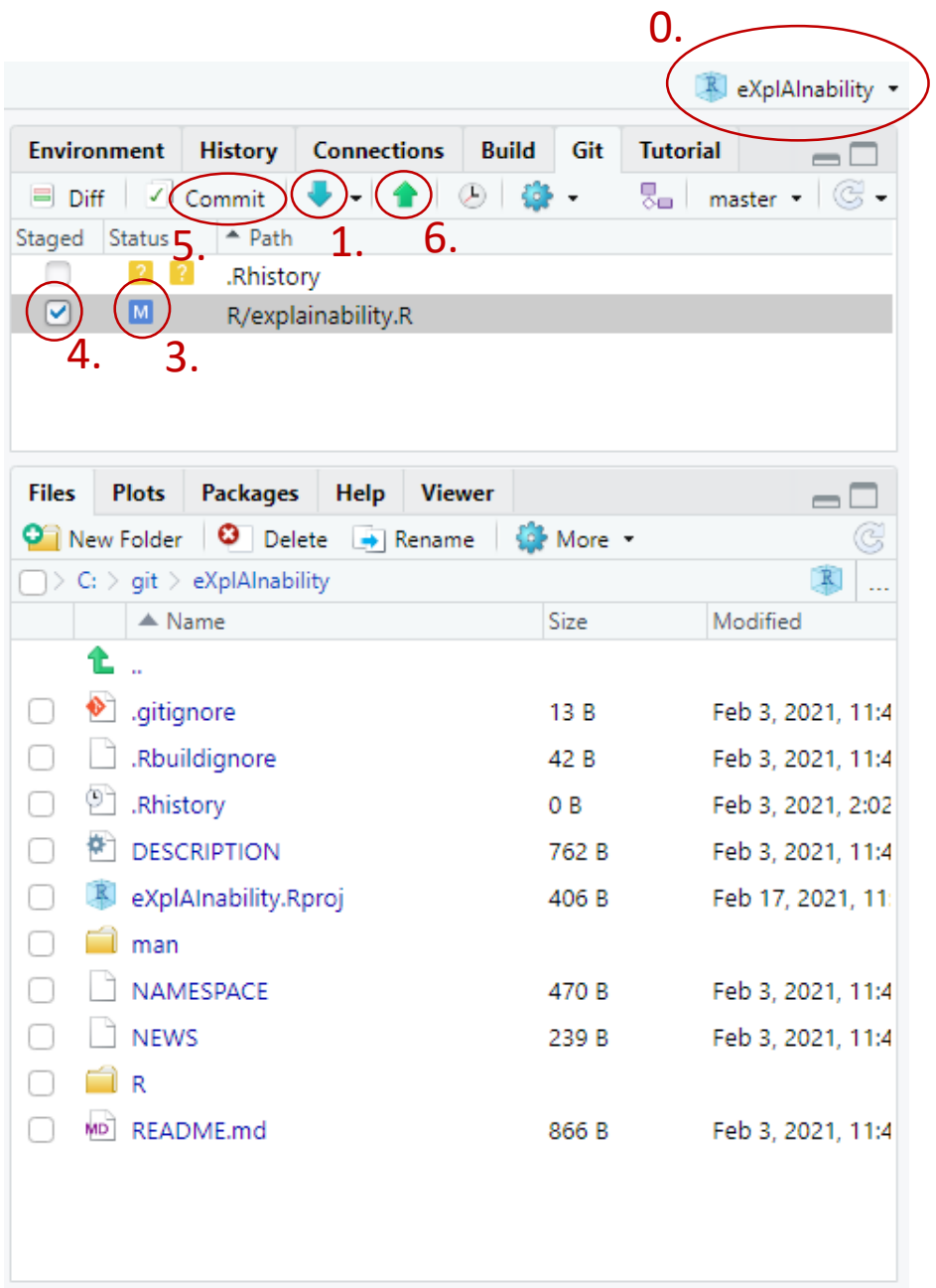
Creating a local copy of a Project from Github



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 commited
5. Commit (+ add comment if requested)
6. Push to Github

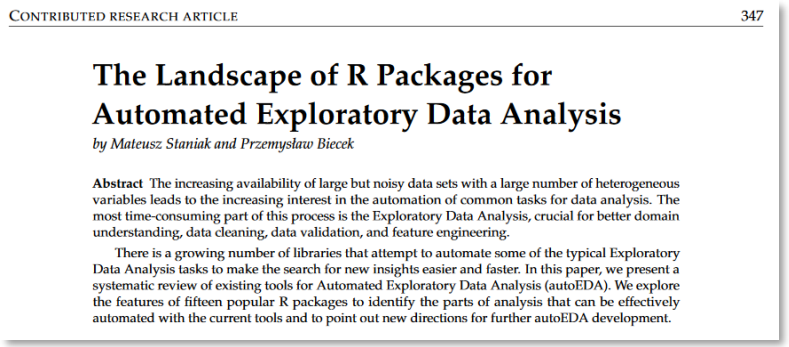


Prüfungsleistung

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. Upload Findings

3. **Predictive Modelling Project**
   I. Assignment of Teams
   II. Kaggle competition
   III. Deployment of your model
   IV. Presentation of your results


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