

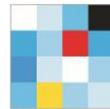
# Modul 3: Umsetzung - Implementierung der Stichprobenziehung im Kontext des Gesamtablaufs

Angewandte Datenanalyse für die öffentliche Verwaltung in Bayern (ADA Bayern)

[www.ada-oeffentliche-verwaltung.de](http://www.ada-oeffentliche-verwaltung.de)

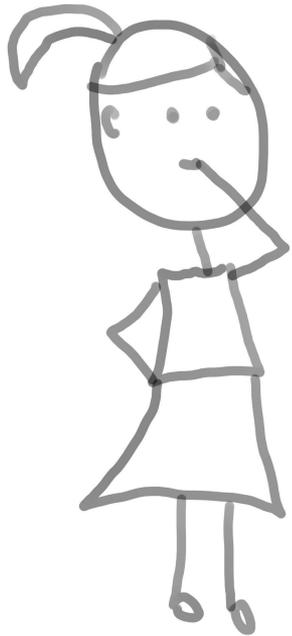


**BERD**  
@NFDI



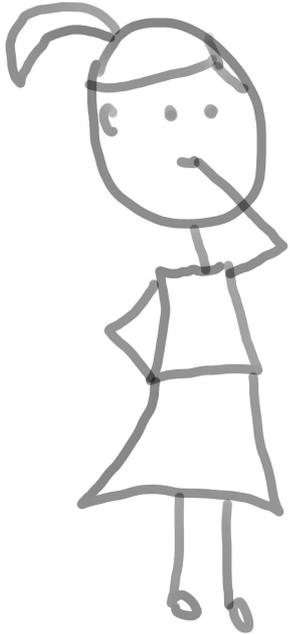
Bayerisches Staatsministerium  
für Digitales





*Worauf freuen Sie sich heute besonders?*

## Modul 3

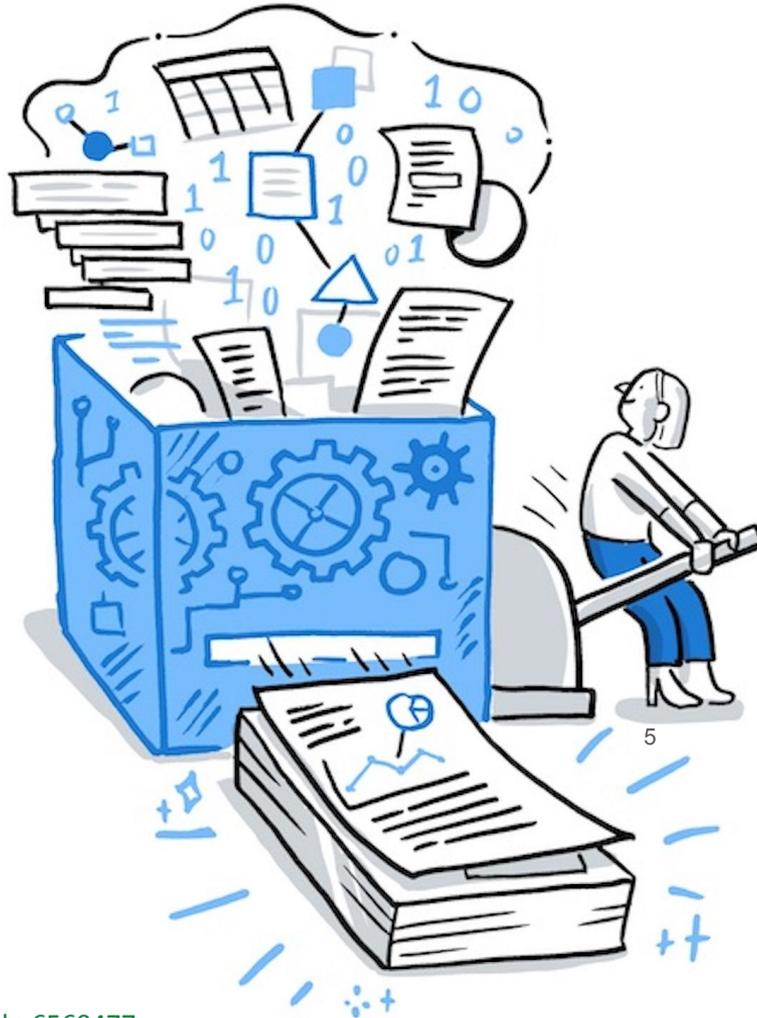


Am Ende dieses Moduls können Sie...

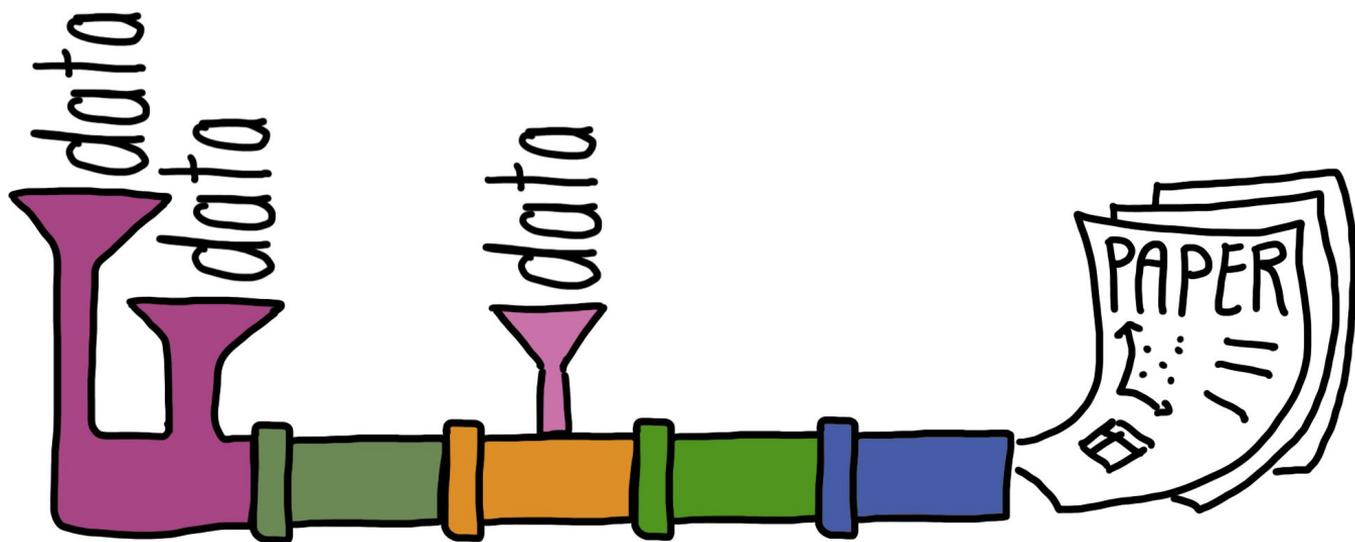
- ... die gewählte Strategie für Datensätze aus verschiedenen Jahren umsetzen.
- ... Lücken erkennen, die im Projekt zu schließen sind, um die Analysen in das Tagesgeschäft aufzunehmen.
- ... die Datenanalysen dokumentieren.

Einführung	10:00 - 10:15
Wie man Analysen wiederholbar macht	10:15 - 10:45
Pause	10:45 - 11:00
Umsetzung	11:00 - 11:30
Report aus den Gruppen	11:30 - 12:00
Mittagspause	12:00 - 13:00
Stand-up	13:00 - 13:15
Umsetzung	13:15 - 14:00
Pause	14:00 - 14:20
Umsetzung	14:20 - 15:20
Wrap-up und Ausblick	15:20 - 15:30

# Wie man Analysen wiederholbar macht



Scriberia 



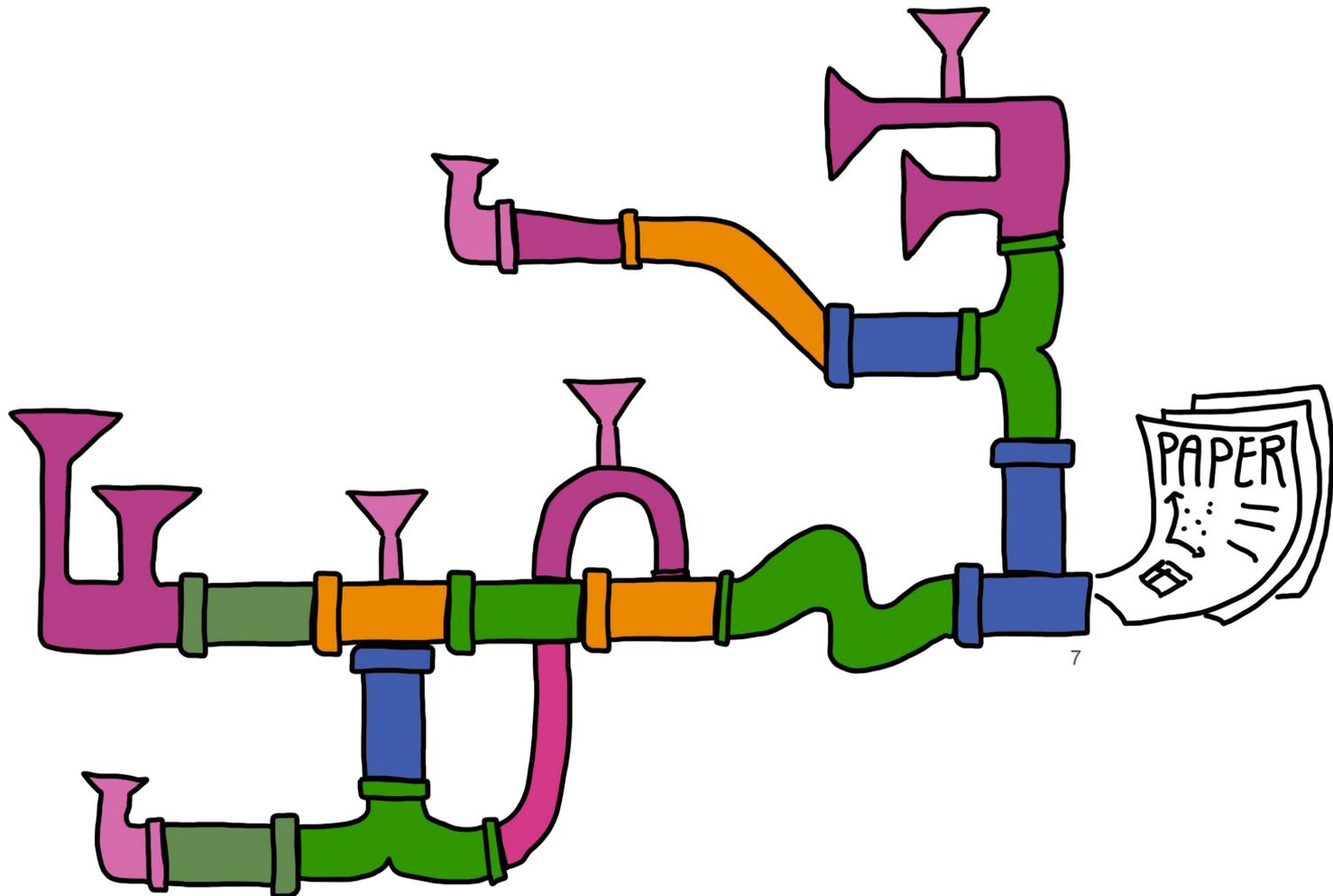
data  
cleaning

overview

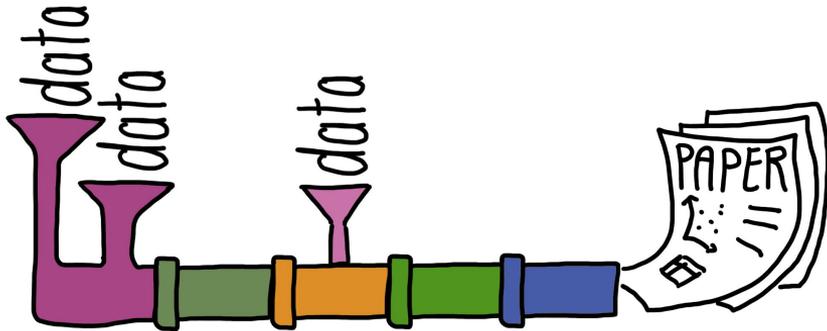
figures

modelling

text



# Reproduzierbare Datenanalysen



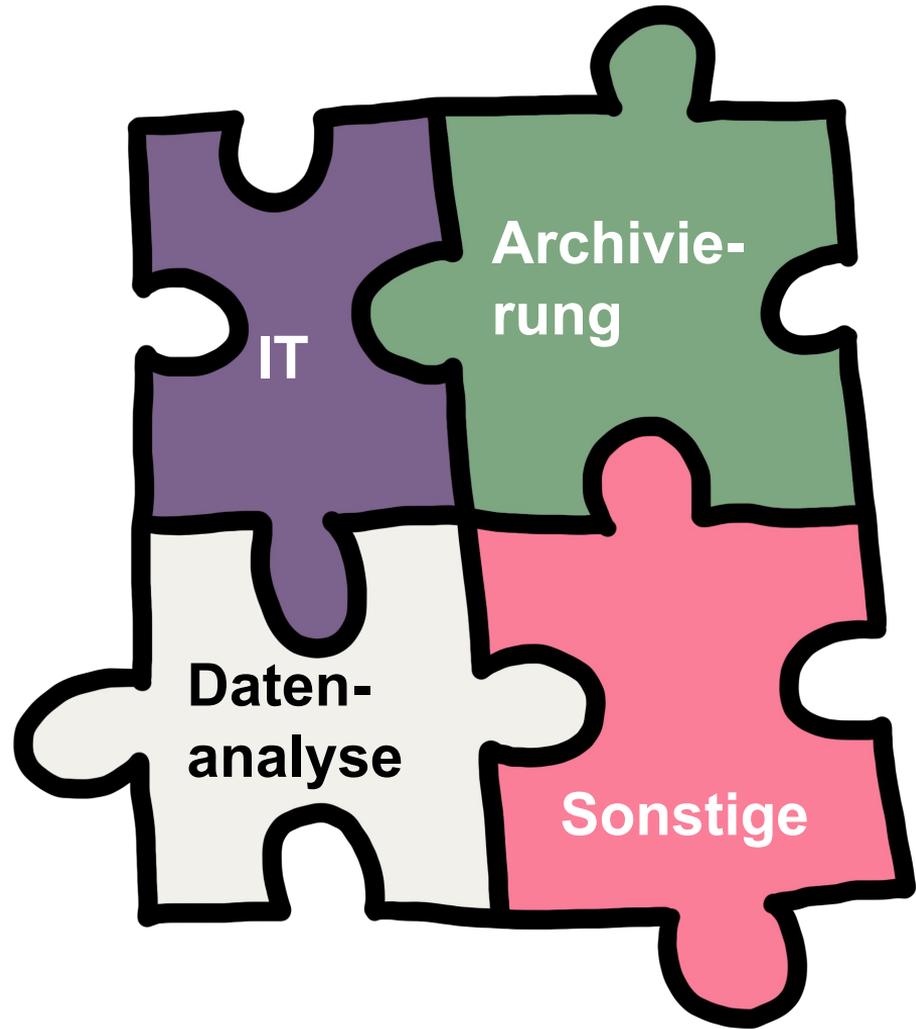
1. Code nutzen
2. Schritte dokumentieren  
(empfohlen: automatisieren)
3. Gute Organisation

# Warum entwickeln wir ein R Paket?

*“Seriously, it doesn’t have to be about sharing your code (although that is an added benefit!). It is about saving yourself time.”*

<https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/>

Data Science ist  
ein Team Sport!



# Die Entwicklung eines R Pakets hilft uns bei der ...

1. Organisation
2. Dokumentation
3. Teilbarkeit
4. Erweiterbarkeit

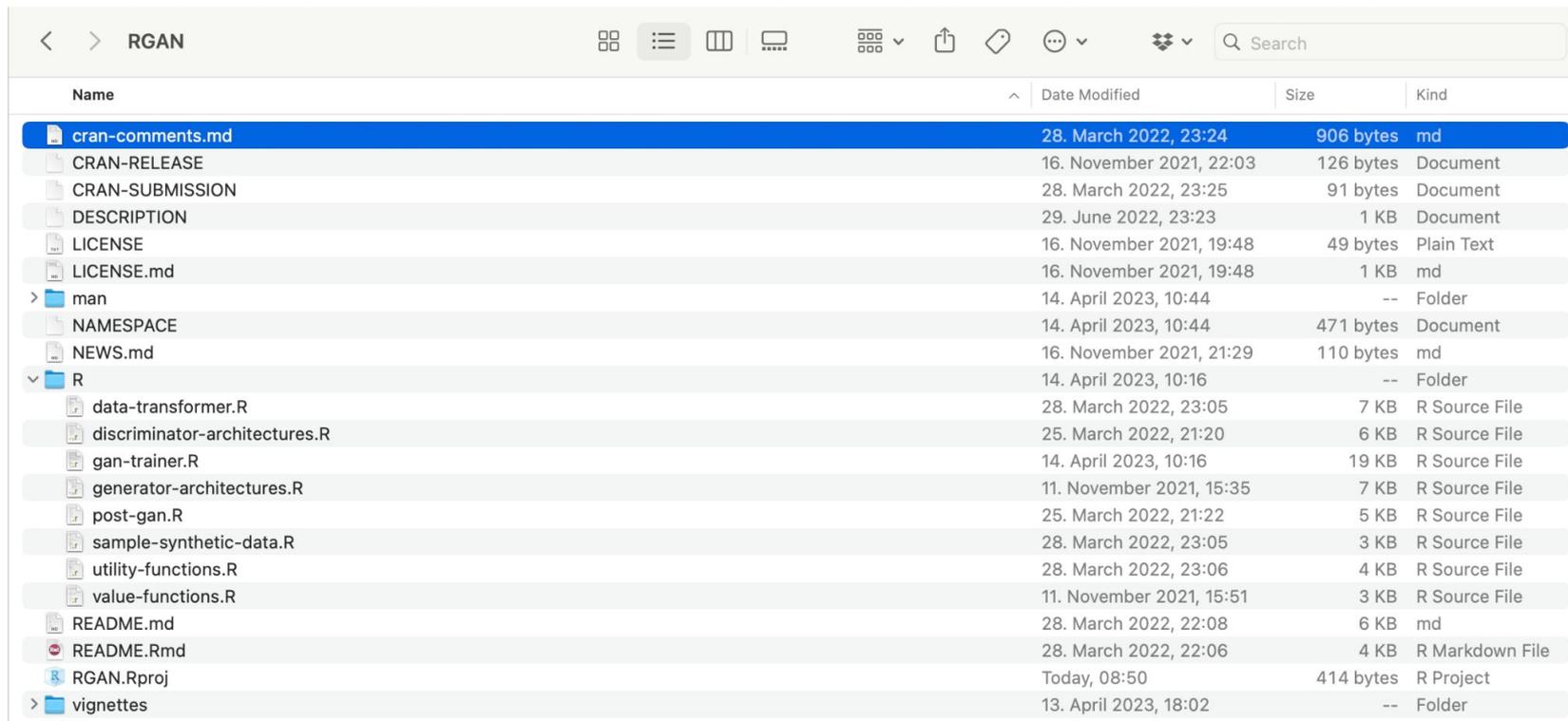
# Die Entwicklung eines R Pakets hilft uns bei der ...

## Organisation

1. R Pakete setzen eine standardisierte Struktur voraus.
2. Für uns nützliche Funktionen werden an einem Ort gebündelt.
3. Wir minimieren copy & paste code.

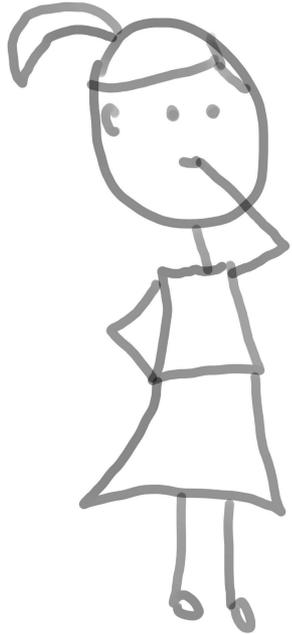
# Die Entwicklung eines R Pakets hilft uns bei der ...

## Organisation



Name	Date Modified	Size	Kind
cran-comments.md	28. March 2022, 23:24	906 bytes	md
CRAN-RELEASE	16. November 2021, 22:03	126 bytes	Document
CRAN-SUBMISSION	28. March 2022, 23:25	91 bytes	Document
DESCRIPTION	29. June 2022, 23:23	1 KB	Document
LICENSE	16. November 2021, 19:48	49 bytes	Plain Text
LICENSE.md	16. November 2021, 19:48	1 KB	md
man	14. April 2023, 10:44	--	Folder
NAMESPACE	14. April 2023, 10:44	471 bytes	Document
NEWS.md	16. November 2021, 21:29	110 bytes	md
R	14. April 2023, 10:16	--	Folder
data-transformer.R	28. March 2022, 23:05	7 KB	R Source File
discriminator-architectures.R	25. March 2022, 21:20	6 KB	R Source File
gan-trainer.R	14. April 2023, 10:16	19 KB	R Source File
generator-architectures.R	11. November 2021, 15:35	7 KB	R Source File
post-gan.R	25. March 2022, 21:22	5 KB	R Source File
sample-synthetic-data.R	28. March 2022, 23:05	3 KB	R Source File
utility-functions.R	28. March 2022, 23:06	4 KB	R Source File
value-functions.R	11. November 2021, 15:51	3 KB	R Source File
README.md	28. March 2022, 22:08	6 KB	md
README.Rmd	28. March 2022, 22:06	4 KB	R Markdown File
RGAN.Rproj	Today, 08:50	414 bytes	R Project
vignettes	13. April 2023, 18:02	--	Folder

Gemeinsam organisieren wir unser R Paket am besten.



*Welche Anforderungen haben Sie an den  
Funktionsumfang des R Pakets?*

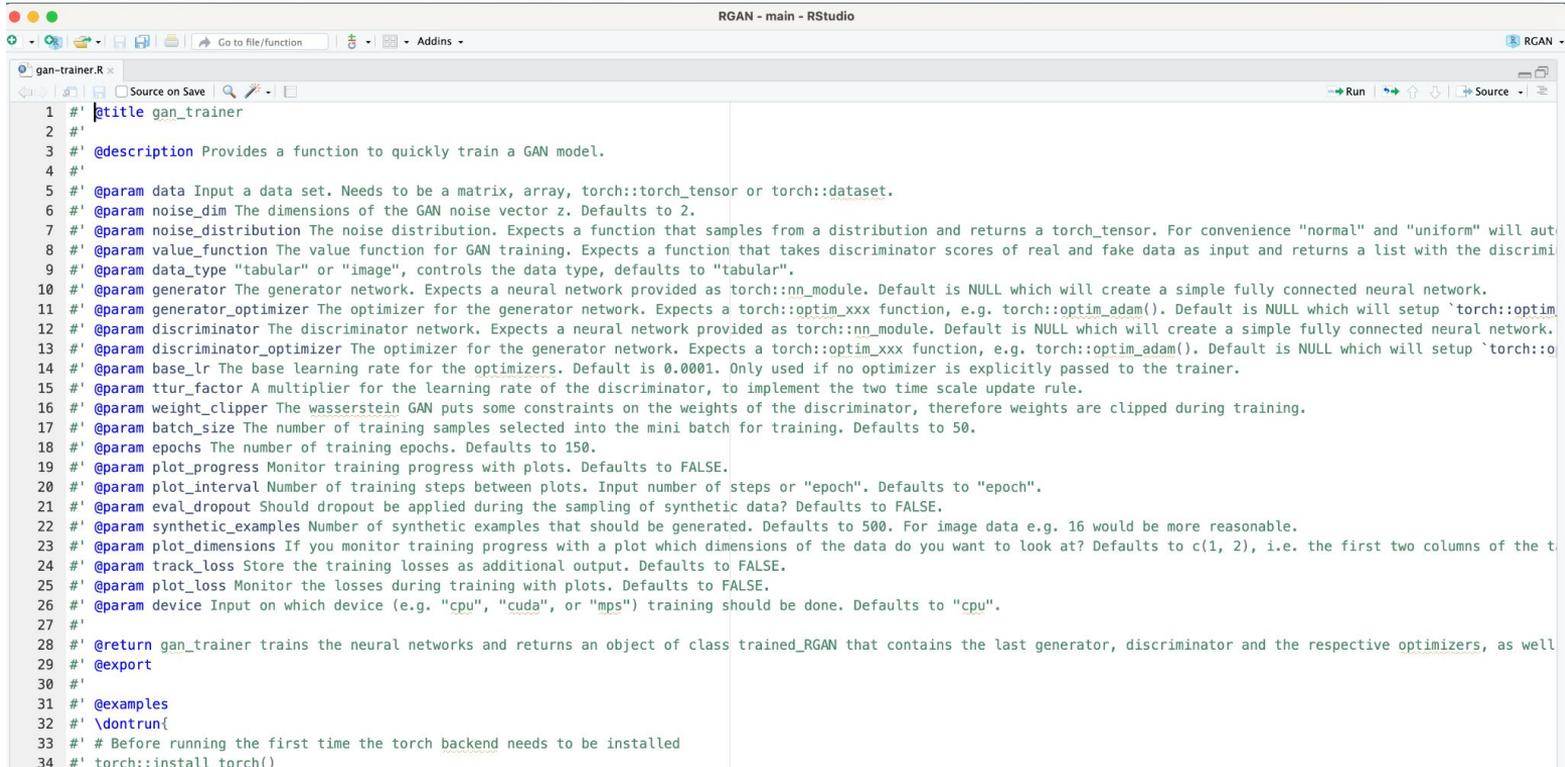
# Die Entwicklung eines R Pakets hilft uns bei der ...

## Dokumentation

1. R Pakete haben eine standardisierte Form der Dokumentation.
2. Wir entwickeln die Dokumentation passgenau im Team.
3. Code Beispiele (Vignetten) geben konkrete Anwendungsbeispiele.

# Die Entwicklung eines R Pakets hilft uns bei der ...

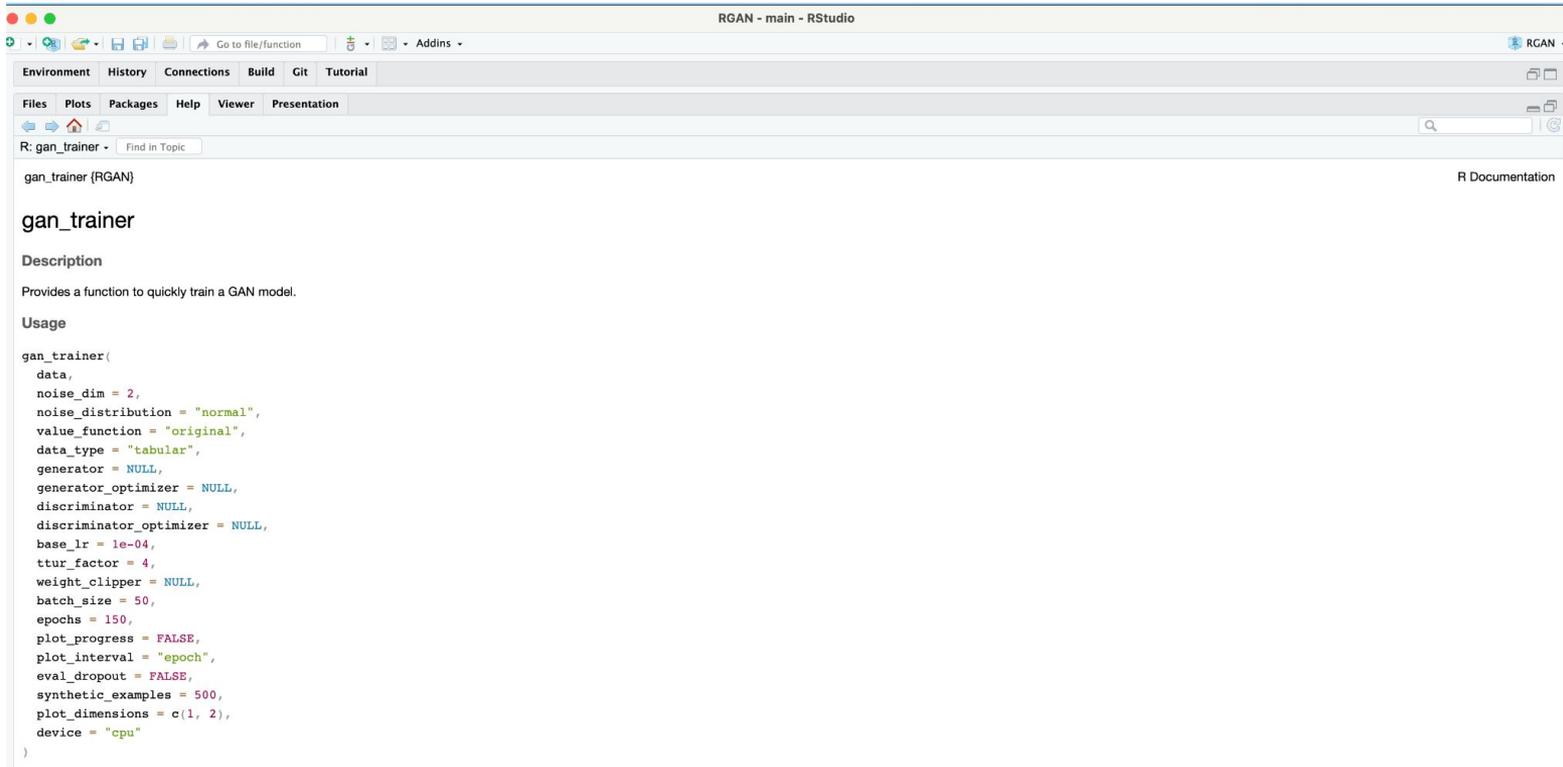
## Dokumentation



```
1 #' @title gan_trainer
2 #'
3 #' @description Provides a function to quickly train a GAN model.
4 #'
5 #' @param data Input a data set. Needs to be a matrix, array, torch::torch_tensor or torch::dataset.
6 #' @param noise_dim The dimensions of the GAN noise vector z. Defaults to 2.
7 #' @param noise_distribution The noise distribution. Expects a function that samples from a distribution and returns a torch_tensor. For convenience "normal" and "uniform" will aut
8 #' @param value_function The value function for GAN training. Expects a function that takes discriminator scores of real and fake data as input and returns a list with the discrimi
9 #' @param data_type "tabular" or "image", controls the data type, defaults to "tabular".
10 #' @param generator The generator network. Expects a neural network provided as torch::nn_module. Default is NULL which will create a simple fully connected neural network.
11 #' @param generator_optimizer The optimizer for the generator network. Expects a torch::optim_xxx function, e.g. torch::optim_adam(). Default is NULL which will setup `torch::optim
12 #' @param discriminator The discriminator network. Expects a neural network provided as torch::nn_module. Default is NULL which will create a simple fully connected neural network.
13 #' @param discriminator_optimizer The optimizer for the generator network. Expects a torch::optim_xxx function, e.g. torch::optim_adam(). Default is NULL which will setup `torch::o
14 #' @param base_lr The base learning rate for the optimizers. Default is 0.0001. Only used if no optimizer is explicitly passed to the trainer.
15 #' @param ttur_factor A multiplier for the learning rate of the discriminator, to implement the two time scale update rule.
16 #' @param weight_clipper The wasserstein GAN puts some constraints on the weights of the discriminator, therefore weights are clipped during training.
17 #' @param batch_size The number of training samples selected into the mini batch for training. Defaults to 50.
18 #' @param epochs The number of training epochs. Defaults to 150.
19 #' @param plot_progress Monitor training progress with plots. Defaults to FALSE.
20 #' @param plot_interval Number of training steps between plots. Input number of steps or "epoch". Defaults to "epoch".
21 #' @param eval_dropout Should dropout be applied during the sampling of synthetic data? Defaults to FALSE.
22 #' @param synthetic_examples Number of synthetic examples that should be generated. Defaults to 500. For image data e.g. 16 would be more reasonable.
23 #' @param plot_dimensions If you monitor training progress with a plot which dimensions of the data do you want to look at? Defaults to c(1, 2), i.e. the first two columns of the t
24 #' @param track_loss Store the training losses as additional output. Defaults to FALSE.
25 #' @param plot_loss Monitor the losses during training with plots. Defaults to FALSE.
26 #' @param device Input on which device (e.g. "cpu", "cuda", or "mps") training should be done. Defaults to "cpu".
27 #'
28 #' @return gan_trainer trains the neural networks and returns an object of class trained_RGAN that contains the last generator, discriminator and the respective optimizers, as well
29 #' @export
30 #'
31 #' @examples
32 #' \dontrun{
33 #' # Before running the first time the torch backend needs to be installed
34 #' torch::install_torch()
```

# Die Entwicklung eines R Pakets hilft uns bei der ...

## Dokumentation



The screenshot shows the RStudio interface with the documentation for the `gan_trainer` function. The window title is "RGAN - main - RStudio". The top menu bar includes "Environment", "History", "Connections", "Build", "Git", and "Tutorial". Below that is a toolbar with icons for "Files", "Plots", "Packages", "Help", "Viewer", and "Presentation". The main editor area displays the documentation for `gan_trainer (RGAN)`, which includes a description, usage, and a list of arguments.

gan\_trainer (RGAN) R Documentation

### gan\_trainer

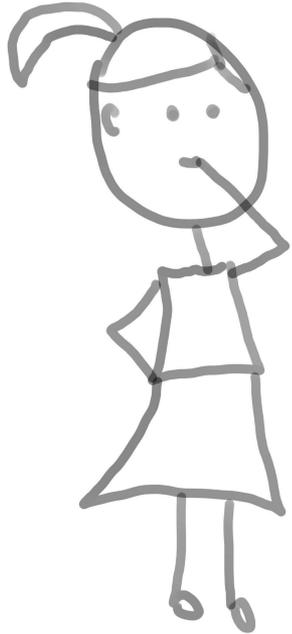
**Description**

Provides a function to quickly train a GAN model.

**Usage**

```
gan_trainer(  
  data,  
  noise_dim = 2,  
  noise_distribution = "normal",  
  value_function = "original",  
  data_type = "tabular",  
  generator = NULL,  
  generator_optimizer = NULL,  
  discriminator = NULL,  
  discriminator_optimizer = NULL,  
  base_lr = 1e-04,  
  ttur_factor = 4,  
  weight_clipper = NULL,  
  batch_size = 50,  
  epochs = 150,  
  plot_progress = FALSE,  
  plot_interval = "epoch",  
  eval_dropout = FALSE,  
  synthetic_examples = 500,  
  plot_dimensions = c(1, 2),  
  device = "cpu"  
)
```

Gemeinsam dokumentieren wir unser R Paket am besten.



*Welche Anforderungen haben Sie an eine gute Dokumentation?*

# Die Entwicklung eines R Pakets hilft uns bei der ...

## Teilbarkeit

1. R Pakete sind sehr einfach mit anderen R Nutzer:innen teilbar. Zum Beispiel über CRAN oder github.
2. Der Quellcode von R Paketen ist offen einsehbar.

# Wie man Analysen in R wiederholbar macht



CRAN  
[Mirrors](#)  
[What's new?](#)  
[Search](#)  
[CRAN Team](#)

About R  
[R Homepage](#)  
[The R Journal](#)

Software  
[R Sources](#)  
[R Binaries](#)  
[Packages](#)  
[Task Views](#)  
[Other](#)

Documentation  
[Manuals](#)  
[FAQs](#)  
[Contributed](#)

Donations  
[Donate](#)

## The Comprehensive R Archive Network

### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux \(Debian, Fedora/Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

### Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2023-10-31, Eye Holes) [R-4.3.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

### Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

### Supporting CRAN

- CRAN operations, most importantly hosting, checking, distributing, and archiving of R add-on packages for various platforms, crucially rely on technical, emotional, and financial support by the R community.  
Please consider making [financial contributions](#) to the R Foundation for Statistical Computing.

### What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](#) for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](#) nearest to you to minimize network load.

### Submitting to CRAN

To "submit" a package to CRAN, check that your submission meets the [CRAN Repository Policy](#) and then use the [web form](#).

# Wie man Analysen in R wiederholbar macht

## METACRAN: Search and browse all CRAN/R packages

 20,444  
active packages

 10,594  
package maintainers

 339  
updates last week

 37,122,314  
downloads last week

### Most downloaded

#### ragg

Graphic Devices Based on AGG  
1.2.7, published 2 months ago, by [Thomas Lin Pedersen](#)

#### textshaping

Bindings to the 'HarfBuzz' and 'Fribidi'  
Libraries for Text Shaping  
0.3.7, published 4 months ago, by [Thomas Lin Pedersen](#)

#### ggplot2

Create Elegant Data Visualisations Using the  
Grammar of Graphics  
3.4.4, published 4 months ago, by [Thomas Lin Pedersen](#)

#### rlang

Functions for Base Types and Core R and  
'Tidyverse' Features  
1.1.3, published a month ago, by [Lionel Henry](#)

#### dplyr

A Grammar of Data Manipulation  
1.1.4, published 3 months ago, by [Hadley Wickham](#)

#### cli

Helpers for Developing Command Line  
Interfaces  
3.6.2, published 2 months ago, by [Gábor Csárdi](#)

#### vctr

Vector Helpers  
0.6.5, published 3 months ago, by [Davis Vaughan](#)

#### lifecycle

Manage the Life Cycle of your Package  
Functions  
1.0.4, published 3 months ago, by [Lionel Henry](#)

#### devtools

Tools to Make Developing R Packages Easier  
2.4.5, published a year ago, by [Jennifer Bryan](#)

### Trending this week

# Die Entwicklung eines R Pakets hilft uns bei der ...

## **Erweiterbarkeit**

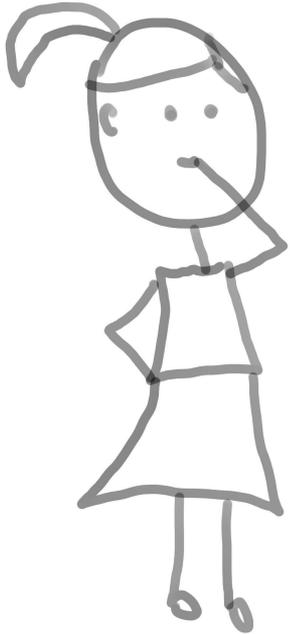
1. R Pakete können einfach um neue Funktionen erweitert werden.
2. Bei offener Entwicklung, zum Beispiel auf github, können andere Entwickler:innen zum Code beitragen.

Mit R können wir auch interaktive Apps entwickeln

<https://koala.stat.uni-muenchen.de/>

Pause

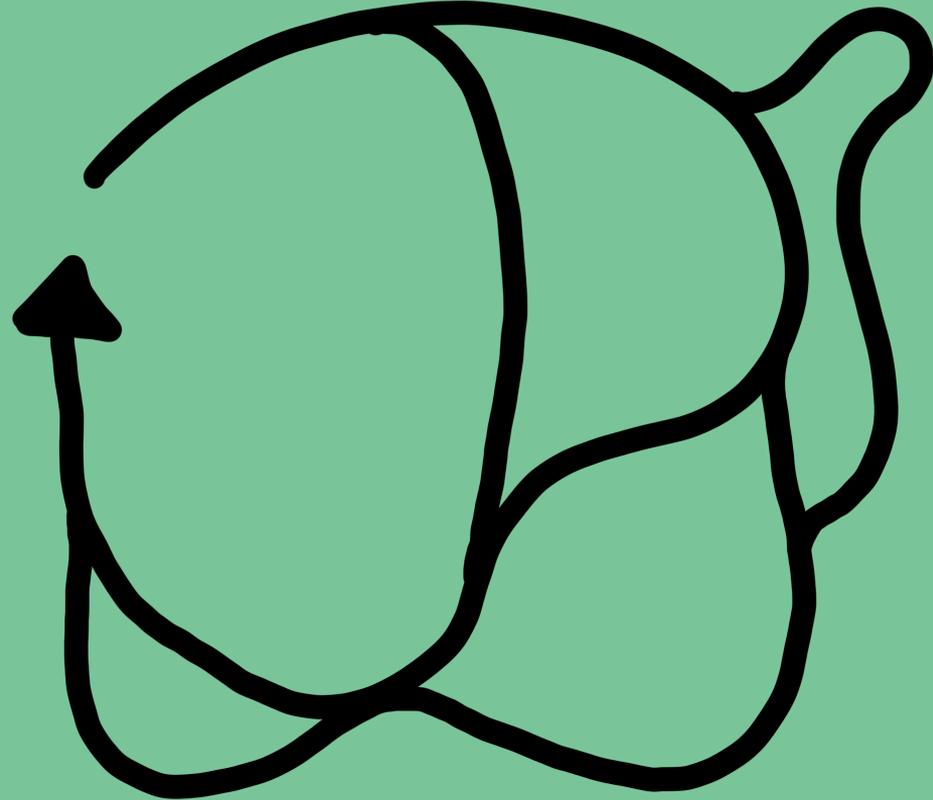
# Umsetzung



*Gibt es Fragen / Diskussionsbedarf, bevor wir wieder in die Umsetzung einsteigen?*

# Umsetzung

Prototyp Notebook entwickeln

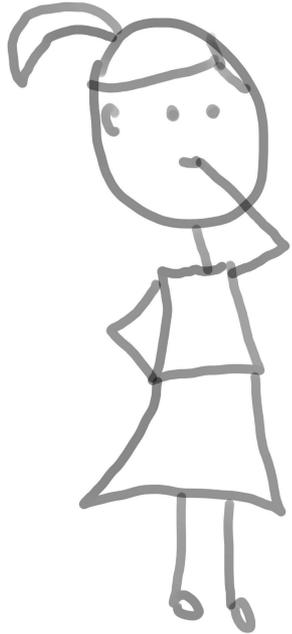


# Reports aus den Gruppen

*Zusatzfrage: Was wünschen Sie  
sich für den Nachmittag?*

Mittagspause

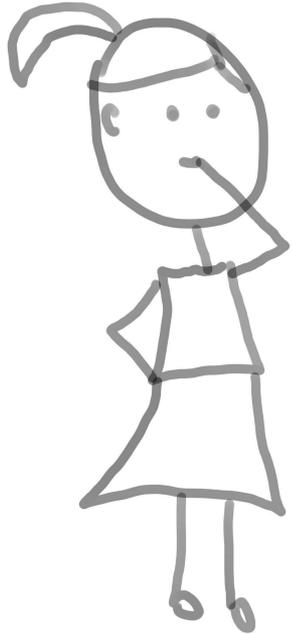
# Umsetzung



*Was möchten Sie bis morgen erreichen?  
(Ziel)*

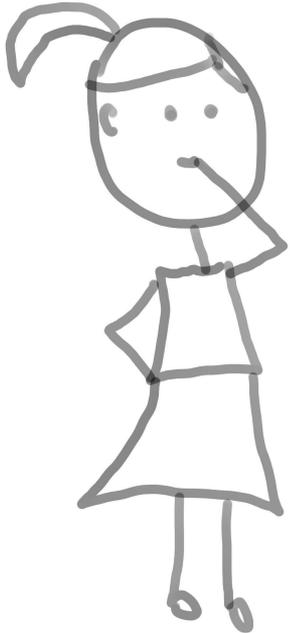
*Was könnten Hürden sein, die dieses Ziel  
schwerer erreichbar machen?*

# Umsetzung



*Welche Sorgen / Bedenken sind  
aufgekommen?*

# Ausblick: Modul 4



Am Ende dieses Moduls können Sie...

- ... einen nachhaltigen Prozess etablieren, wie die Daten übertragen, die Analysen durchgeführt und Archivierungsentscheidungen dokumentiert werden.
- ... ihr Wissen an andere weitergeben.
- ... einschätzen, wie die benötigte Infrastruktur erhalten werden kann.