Introduction to Data Science

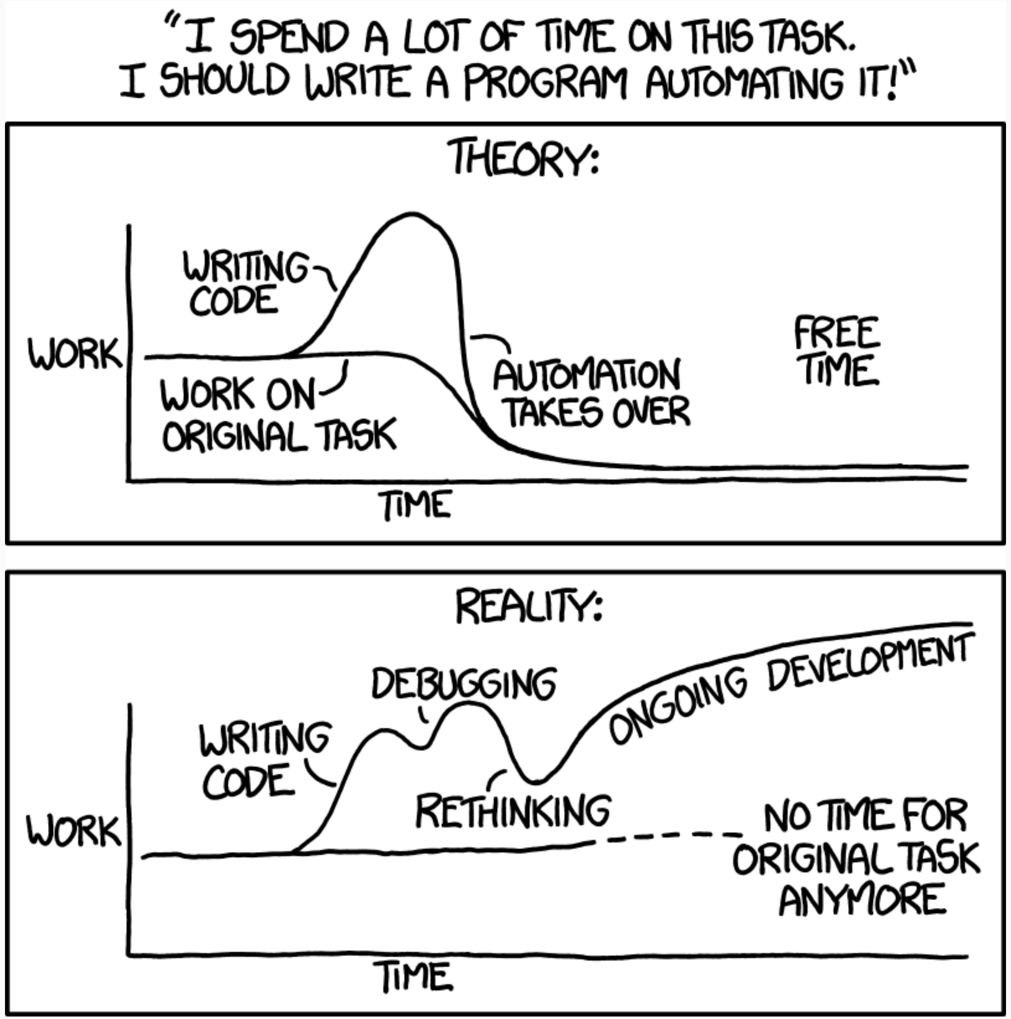
Session 14: Automation, scheduling, and packages

Simon Munzert
Hertie School | GRAD-C11/E1339

Table of contents

Automation and scripting

Automation



Credit Randall Munroe/xkcd 1319

Automation

Motivation

- We spend too much time on repetitive tasks.
- We're already automating using scripts that bundle multiple commands! Next step: The pipeline as a series of scripts and commands.
- Good pipelines are modular. But you don't want to trigger 10 scripts sequentially by hand.
- Some tasks are to be repeated on a regular basis (schedule).

When automation makes sense

- The input is variable but the process of turning input into output is highly standardized.
- You use a diverse set of software to produce the output.
- Others (humans, machines) are supposed to run the analyses.
- Time saved by automation >> Time needed to automate.

Different ways of doing it

We will consider automation

- using **R**

Thinking in pipelines

Key characteristics

- Pipelines make complex projects easier to handle because they break up a monolithic script into **discrete, manageable chunks**.
- If properly done, each stage of the pipeline defines its input and its outputs.
- Pipeline modules **do not modify their inputs** (*idempotence*). Rerunning one module produces the same results as the previous run.

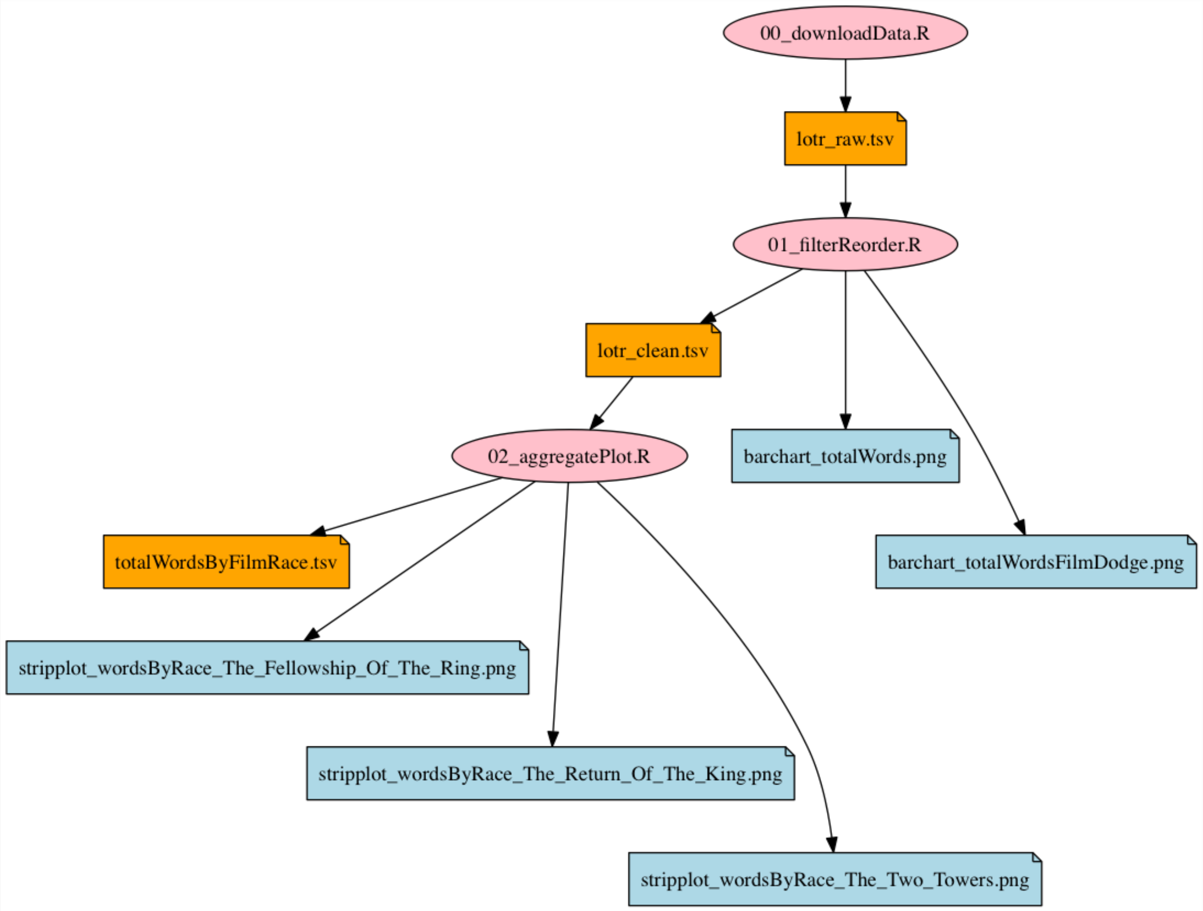
Key advantages

- When you modify one stage of the pipeline, you only have to rerun the downstream, dependent stages.
- Division of labor is straightforward.
- Modules tend to be a lot easier to debug.

A data science pipeline is a graph

Wait what

- Scripts and data files are vertices of the graph.
- Dependencies between stages are edges of the graph.
- Pipelines are not necessarily DAGS. Recursive routines are imaginable (but to be avoided?).
- Also, scripts are not necessarily hierarchical (e.g., multiple different modeling approaches of the same data in different scripts).
- An automation script gives *one* order in which you can successfully run the pipeline.



An example pipeline

In the following, we will work with this toy pipeline:

[1]Courtesy of Jenny Bryan.

An example pipeline

In the following, we will work with this toy pipeline:

- `00-packages.R` loads the packages necessary for analysis,

`00-packages.R`:

```
R> # install packages from CRAN
R> p_needed ← c("tidyverse" # tidyverse packages
+ )
R> packages ← rownames(installed.packages())
R> p_to_install ← p_needed[!(p_needed %in% packages)]
R> if (length(p_to_install) > 0) {
+   install.packages(p_to_install)
+ }
R> lapply(p_needed, require, character.only = TRUE)
```

An example pipeline

In the following, we will work with this toy pipeline:

- `00-packages.R` loads the packages necessary for analysis,
- `01-download-data.R` downloads a spreadsheet, which is stored as `lotr_raw.tsv`,

`01-download-data.R`:

```
R> ## download raw data
R> download.file(url = "http://bit.ly/lotr_raw-tsv",
+                     destfile = "lotr_raw.tsv")
```

An example pipeline

In the following, we will work with this toy pipeline:

- `00-packages.R` loads the packages necessary for analysis,
- `01-download-data.R` downloads a spreadsheet, which is stored as `lotr_raw.tsv`,
- `02-process-data.R` imports and processes the data and exports a clean spreadsheet as `lotr_clean.tsv`, and

`02-process-data.R`:

```
R> ## import raw data
R> lotr_dat ← read_tsv("lotr_raw.tsv")
R>
R> ## reorder Film factor levels based on story
R> old_levels ← levels(as.factor(lotr_dat$Film))
R> j_order ← sapply(c("Fellowship", "Towers", "Return"),
+                       function(x) grep(x, old_levels))
R> new_levels ← old_levels[j_order]
R>
R> ## process data set
R> lotr_dat ← lotr_dat %>%
+   # apply new factor levels to Film
+     mutate(Film = factor(as.character(Film), new_levels),
+     # revalue Race
+     Race = recode(Race, `Ainur` = "Wizard", `Men` = "Man")) %>%
+ ## <skipping some steps here to avoid slide overflow>
+
+ ## write data to file
+ write_tsv(lotr_dat, "lotr_clean.tsv")
```

An example pipeline

In the following, we will work with this toy pipeline:

- `00-packages.R` loads the packages necessary for analysis,
- `01-download-data.R` downloads a spreadsheet, which is stored as `lotr_raw.tsv`,
- `02-process-data.R` imports and processes the data and exports a clean spreadsheet as `lotr_clean.tsv`, and
- `03-plot.R` imports the clean dataset, produces a figure and exports it as `barchart-words-by-race.png`.

`03-plot.R`:

```
R> ## import clean data
R> lotr_dat <- read_tsv("lotr_clean.tsv") %>%
+ # reorder Race based on words spoken
+ mutate(Race = reorder(Race, Words, sum))
R>
R> ## make a plot
R> p <- ggplot(lotr_dat, aes(x = Race, weight = Words)) + geom_bar()
R> ggsave("barchart-words-by-race.png", p)
```

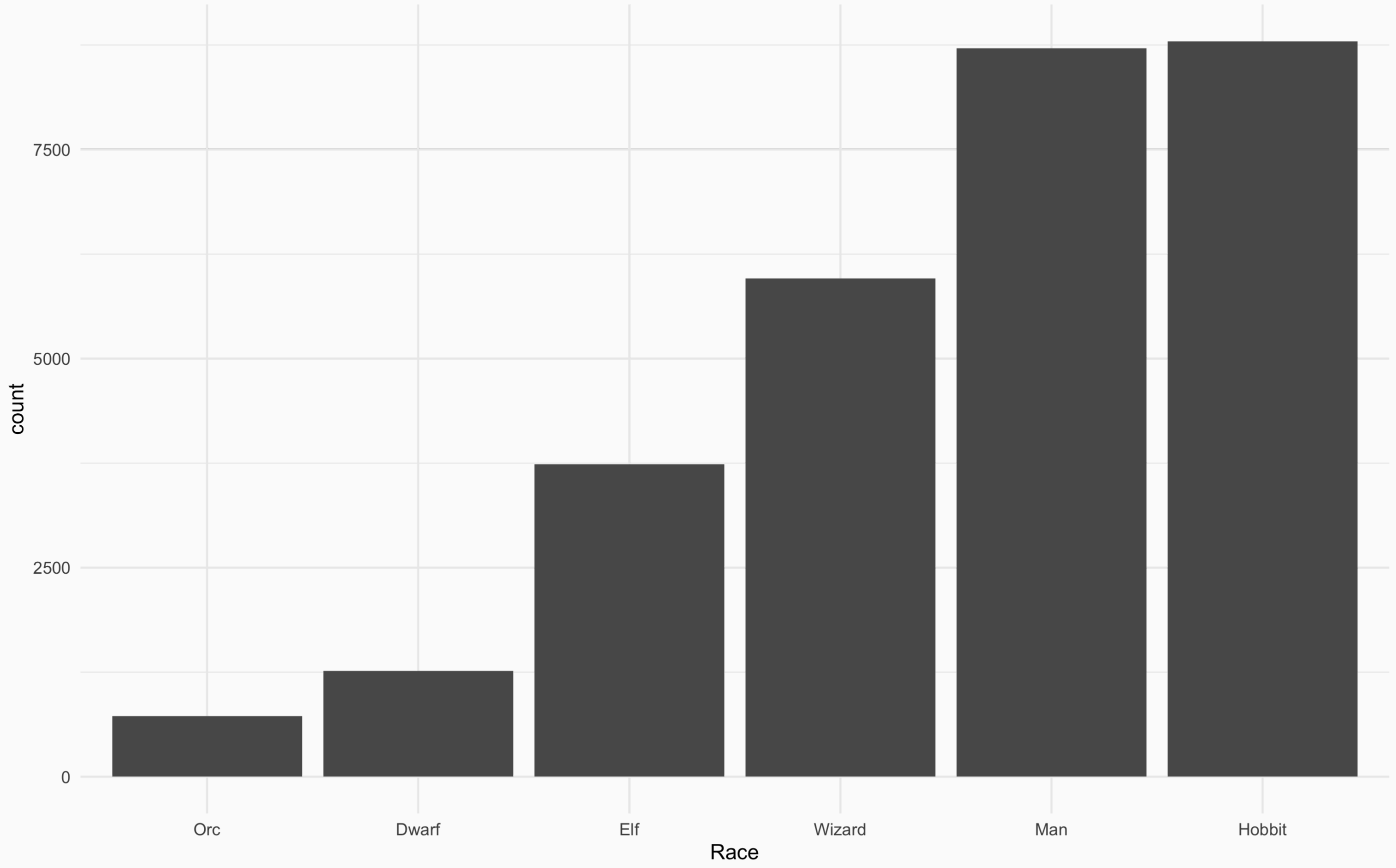
An example pipeline

```
R> slice_sample(lotr_dat, n = 10)
```

```
# A tibble: 10 × 5
   Film                        Chapter                      Character Race  Words
   <chr>                       <chr>                        <chr>     <chr> <dbl>
 1 The Return Of The King      64: The Mouth Of Sauron      Aragorn   Man      23
 2 The Fellowship Of The Ring  36: The Bridge Of Khazad-…   Frodo     Hobb…     4
 3 The Two Towers              36: Isengard Unleashed       Saruman   Wiza…    50
 4 The Fellowship Of The Ring  42: The Great River          Sam       Hobb…    37
 5 The Return Of The King      42: Breaking The Gate Of …   Gandalf   Wiza…    21
 6 The Two Towers              45: The Glittering Caves     Legolas   Elf      36
 7 The Two Towers              35: Helm's Deep              Rohan Wa… Man      22
 8 The Fellowship Of The Ring  33: Moria                    Aragorn   Man      31
 9 The Fellowship Of The Ring  43: Parth Galen              Aragorn   Man      79
10 The Return Of The King      24: Courage Is The Best D…   Gothmog   Orc       4
```

An example pipeline

```
R> p ← ggplot(lotr_dat, aes(x = Race, weight = Words)) +
+   geom_bar() + theme_minimal()
```



Automation using pipelines in R

Motivation and usage

- The `source()` function reads and parses R code from a file or connection.
- We can build a pipeline by sourcing scripts sequentially.
- This pipeline is usually stored in a "master" script.
- The removal of previous work is optional and maybe redundant. Often the data is overwritten by default.
- It is recommended that the individual scripts are (partial) standalones, i.e. that they import all data they need by default (loading the packages could be considered an exception).
- Note that as long as the environment is not reset, it remains intact across scripts, which is a potential source of error and confusion.

Example

The master script `master.R`:

```
R> ## clean out any previous work
R> outputs <- c("lotr_raw.tsv",
+               "lotr_clean.tsv",
+               list.files(pattern = "*.png$"))
R> file.remove(outputs)
R>
R> ## run scripts
R> source("00-packages.R")
R> source("01-download-data.R")
R> source("02-process-data.R")
R> source("03-plot.R")
```

Automation using the Shell and Rscript

Motivation and usage

- Alternatively to using an R master script, we can also run the pipeline from the command line.
- Note that here, the environments don't carry over across `Rscript` calls. The scripts definitely have to run in a standalone fashion (i.e., load packages, import all necessary data, etc.).
- The working directory should be set either in the script(s) or in the shell with `cd`.

Example

The master script `master.sh`:

```
#!/bin/sh
cd /Users/simonmunzert/github/examples/02-automation
set -eux
Rscript 01-download-data.R
Rscript 02-process-data.R
Rscript 03-plot.R
```

The `set` command allows to adjust some base shell parameters:

- `-e`: Stop at first error
- `-u`: Undefined variables are an error
- `-x`: Print each command as it is run

For more information on `set`, see here.

Automation using the Shell and Rscript

Motivation and usage

- Alternatively to using an R master script, we can also run the pipeline from the command line.
- Note that here, the environments don't carry over across `Rscript` calls. The scripts definitely have to run in a standalone fashion (i.e., load packages, import all necessary data, etc.).
- The working directory should be set either in the script(s) or in the shell with `cd`.
- One advantage of this approach is that it can be easily coupled with other command line tools, building a **polyglot pipeline**.

Example

The master script `master.sh`:

```
#!/bin/sh
cd /Users/simonmunzert/github/examples/02-automation
set -eux
curl -L http://bit.ly/lotr_raw-tsv > lotr_raw.tsv
Rscript 02-process-data.R
Rscript 03-plot.R
```

The `set` command allows to adjust some base shell parameters:

- `-e`: Stop at first error
- `-u`: Undefined variables are an error
- `-x`: Print each command as it is run

For more information on `set`, see here.

Automation using Make

Motivation and usage

- Make is an automation tool that allows us to specify and manage build processes.
- It is commonly run via the shell.
- At the heart of a make operation is the `makefile` (or `Makefile`, `GNUmakefile`), a script which serves as a recipe for the building process.
- A `makefile` is written following a particular syntax and in a declarative fashion.
- Conceptually, the recipe describes which files are built how and using what input.

Advantages of Make

- It looks at which files you have and automatically figures out how to create the files that you have. For complex pipelines this "automation of the automation process" can be very helpful.
- While shell scripts give *one* order in which you can successfully run the pipeline, Make will figure out the parts of the pipeline (and their order) that are needed to build a desired target.

GNU Make

Automation using Make (cont.)

Basic syntax

Each batch of lines indicates

- a file to be created (the target),
- the files it depends on (the prerequisites), and
- set of commands needed to construct the target from the dependent files.

Dependencies propagate.

- To create any of the `png` figures, we need `lotr_clean.tsv`.
- If this file changes, the `png`s change as well when they're built.

Example `makefile`

```
all: lotr_clean.tsv barchart-words-by-race.png words-histogram.png

lotr_raw.tsv:
    curl -L http://bit.ly/lotr_raw-tsv > lotr_raw.tsv

lotr_clean.tsv: lotr_raw.tsv 02-process-data.R
    Rscript 02-process-data.R

barchart-words-by-race.png: lotr_clean.tsv 03-plot.R
    Rscript 03-plot.R

words-histogram.png: lotr_clean.tsv
    Rscript -e 'library(ggplot2);
    qplot(Words, data = read.delim("$<"), geom = "histogram");
    ggsave("$@")'
    rm Rplots.pdf

clean:
    rm -f lotr_raw.tsv lotr_clean.tsv *.png
```

Automation using Make (cont.)

Getting Make to run

- Using the command line, go into the directory for your project.
- Create the `Makefile` file.[1]
- The most basic Make commands are `make all` and `make clean` which builds (or deletes) all output as specified in the script.

Example `makefile`

```
all: lotr_clean.tsv barchart-words-by-race.png words-histogram.png

lotr_raw.tsv:
    curl -L http://bit.ly/lotr_raw-tsv > lotr_raw.tsv

lotr_clean.tsv: lotr_raw.tsv 02-process-data.R
    Rscript 02-process-data.R

barchart-words-by-race.png: lotr_clean.tsv 03-plot.R
    Rscript 03-plot.R

words-histogram.png: lotr_clean.tsv
    Rscript -e 'library(ggplot2);
    qplot(Words, data = read.delim("$<"), geom = "histogram");
    ggsave("$@")'
    rm Rplots.pdf

clean:
    rm -f lotr_raw.tsv lotr_clean.tsv *.png
```

[1] While the basic syntax is simple (see right), the devil's in the detail. Check out resources listed on the next slide if you want to learn more.

Automation using Make - FAQ

Does it work on Windows?

To install an run `make` on Windows, check out these instructions.

Where can I learn more?

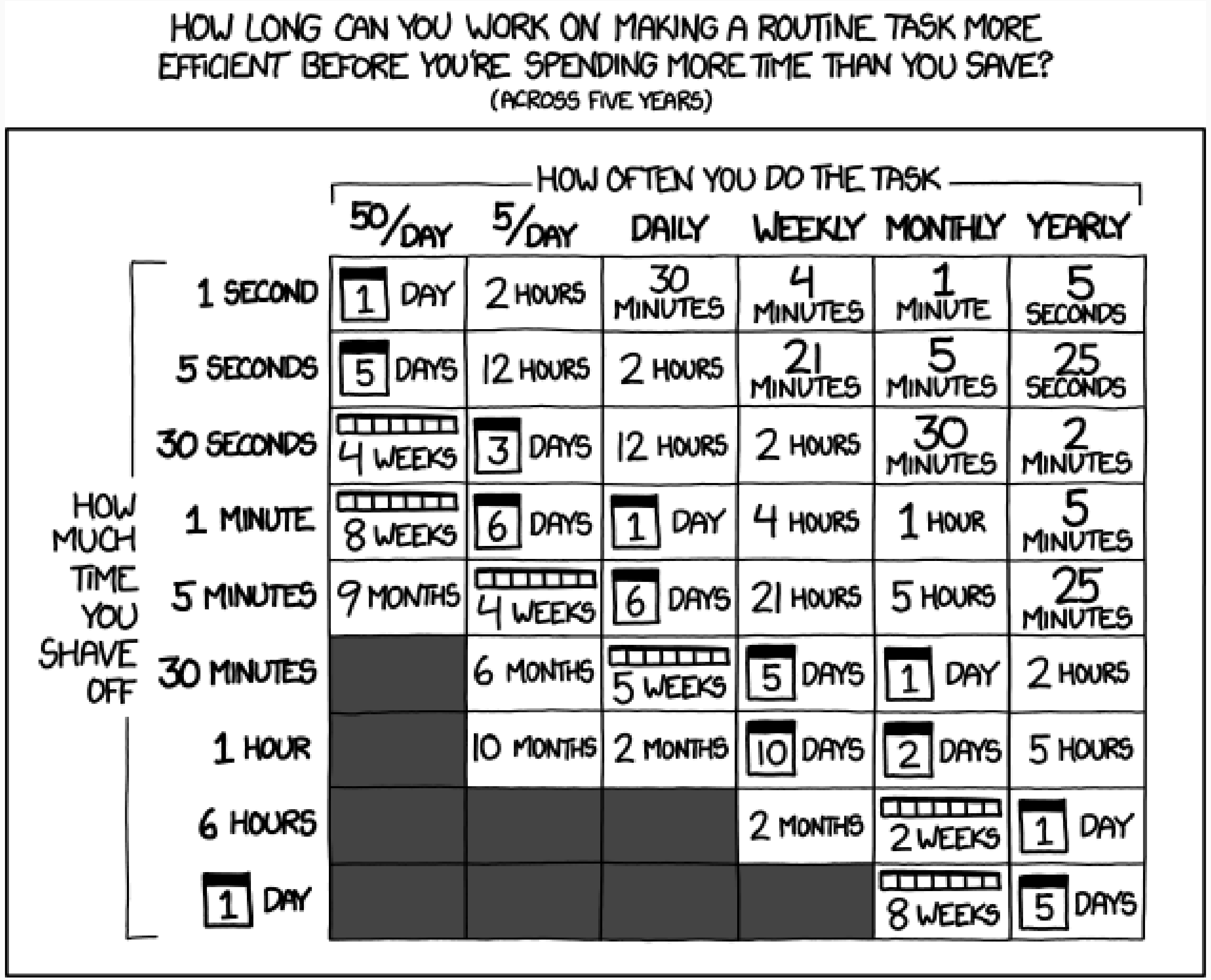
If you consider working with Make, check out the official manual, this helpful tutorial, Karl Broman's excellent minimal make introduction, or this Stat545 piece.

This is dusty technology. Are there alternatives?

In the context of data science with R, the `targets` package is an interesting option. It provides R functionality to define a Make-stype pipeline. Check out the overview and manual.

Scheduling

Scheduling



Credit Randall Munroe/xkcd 1205

Scheduling scripts and processes

Motivation

- So far, we have automated data science pipelines.
- But the execution of these pipelines still needs to be triggered.
- In some cases, it is desirable to also **automate the initialization** of R scripts (or any processes for that matter) **on a regular basis**, e.g. weekly, daily, on logon, etc.
- This makes particular sense when you have moving parts in your pipeline (most likely: data).

Common scenarios for scheduling

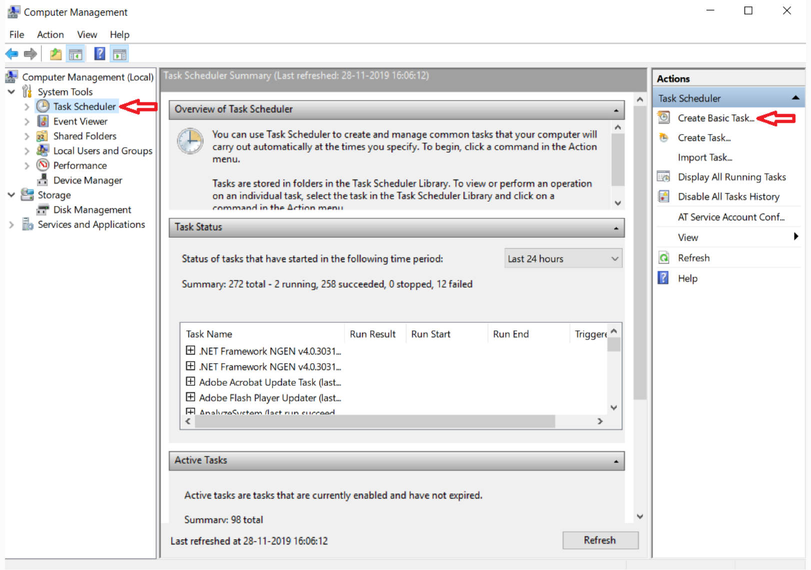
1. You fetch data from the web on a regular basis (e.g., via scraping scripts or APIs).
2. You generate daily/weekly/monthly reports/tweets based on changing data.
3. You build an alert control system informing you about anomalies in a database.

`Credit` Simone Giertz

Scheduling scripts and processes on Windows

Scheduling options

- Processes on Windows can be scheduled with the Windows Task Scheduler.
- Manage them via a GUI (⟶ Control Panel) or the command line using `schtasks.exe`.
- The R package `taskscheduleR` provides a programmable R interface to the WTS.



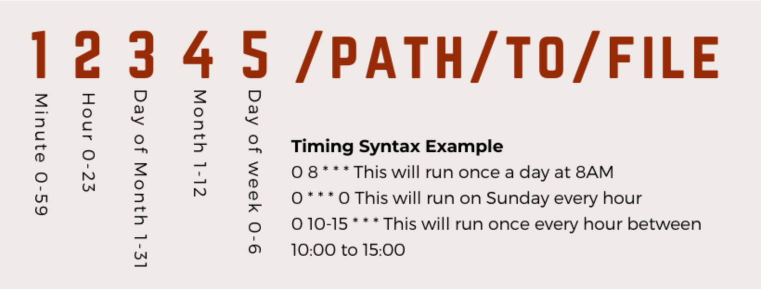
`taskscheduleR` example

```
R> library(taskscheduleR)
```

Scheduling scripts and processes on a Mac

Scheduling options

- On macOS you can schedule background jobs using `cron` and `launchd`.
- `launchd`[1] was created by Apple as a replacement for the popular Linux utility `cron` (deprecated but still usable).
- The R package `cronR` provides a programmable R interface.
- `cron` syntax for more complex scheduling:



`cronR` example

```
R> library(cronR)
R> myscript <- "examples/scrape-wiki.R"
R> # Create bash code for crontab to execute R script
R> cmd <- cron_rscript(myscript)
R>
```

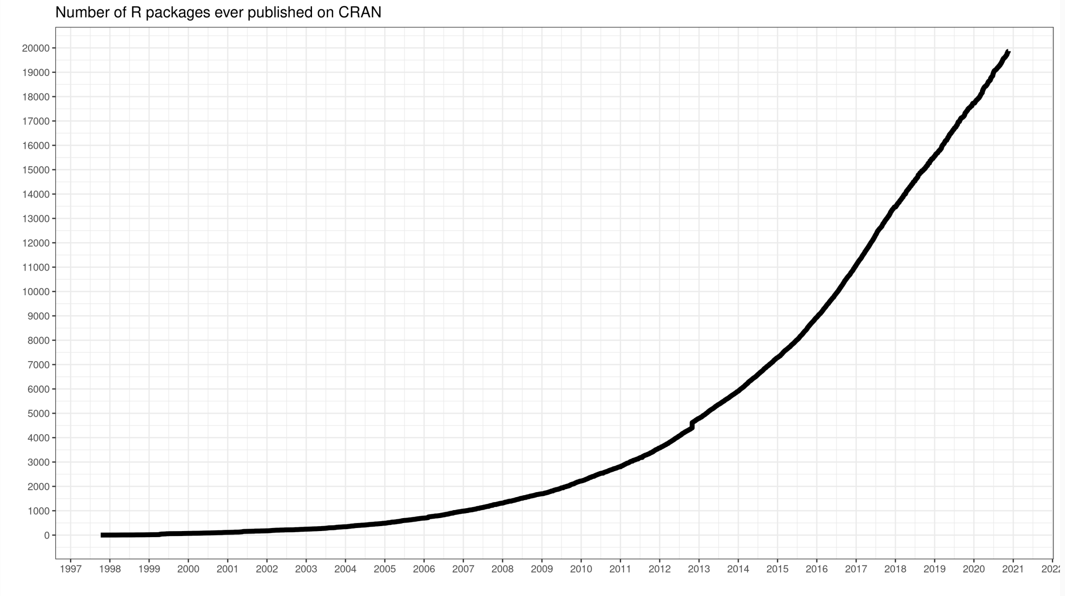
[1] For more resources on scheduling with `launchd`, check out this and this and this.

R packages

Writing an R package

The state of the R package ecosystem

- As of November 2021, the CRAN package repository features more than 18,000 packages.
- Many, many more are available on GitHub and other code sharing platforms.
- R has a vivid community that continuous to create and build extensions and maintain the existing environment. Many of them have much more training and time to invest in software development.
- So, why should we (and with that I mean YOU) write yet another R package?



`Credit` daroczig

Why create another R package?

1. **Thinking in functions.** R is a functional programming language, and packages bundle functions. Thinking of projects as packages is consistent with a functional mindset.

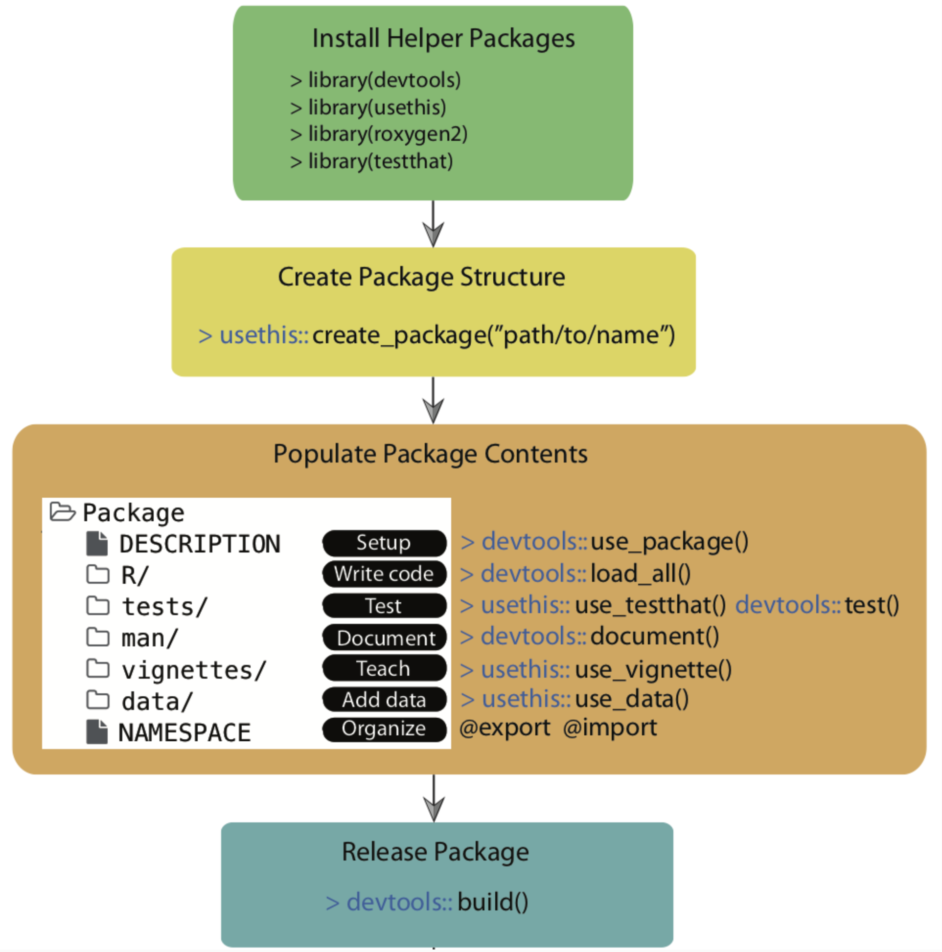
2. **Automation and transportability.** By turning tasks into functions you save repetitive typing, keep frequently-used code together, and let code travel across projects.

3. **Collaboration and transparency.** Packages are ideal to make functionality available to others, but also to let others contribute. As a side effect, it nudges you to document your functions properly and gives you the opportunity to let others review and improve your code easily.

4. **Visibility and productization.** Publishing code in packages is potentially giving you project a big boost in visibility. Also, it is more likely to be perceived as a product than an insular project.

Creating a package from start to finish

1. Choose a package name
2. Set up your package with RStudio (and GitHub)
3. Fill your package with life
    - Add functions
    - Write help files
    - Write a `DESCRIPTION`
    - Add internal data
4. Check your package
    - Write tests
    - Check on various operating systems
    - Check for good coding practice
5. Submit to CRAN (or GitHub early in the process)
6. Promotion
    - Write a vignette
    - Build a package website



`Credit` Simo Goshev, Steve Worthington

Tools to get you started

devtools

- The workhorse of package development in R
- Provides functions that simplify common tasks, such as package setup, simulating installs, compiling from source

usethis

- Provides workflow utilities for project development (loaded by `devtools`)
- Many `use_*()` functions to help create package tests, data, description, etc.

testthat

- Provides functions that make it easy to describe what you expect a function to do, including catching errors, warnings, and messages.

roxygen2

- Provides functions to streamline/automate the documentation of your packages and functions

An example walkthrough

In the following we will briefly study the process of creating a package.

The example is taken from Methods Bites, the Blog of the MZES Social Science Data Lab, and developed by Cosima Meyer and Dennis Hammerschmidt.

The idea is to create a package `overviewR` that helps you to get an overview – hence, the name – of your data with particular emphasis on the extent that your distinct units of observation are covered for the entire time frame of your data set.

The package is real and lives on both CRAN and GitHub. Check out the vignette.

Step 1: Idea and name

Idea

- I'll leave you alone with that one.
- … but you might want to check out the over 18k existing ones that live on CRAN.

Name

- Package names can only be letters and numbers and must start with a letter.
- The package `available` helps you — both with getting inspiration for a name and with checking whether your name is available.

Example

```
R> library(available)
R> # Check for potential names
R> available::suggest("Easily extract information about sample")
```

```
  easilyr
```

```
R> # Check whether it's available
R> available::available("overviewR", browse = FALSE)
```

Step 2: Set up your package

Option 1: via RStudio and GitHub

- Use RStudio's Project Wizard and click on `File` > `New Project ...` > `New Directory` > `R Package`.
- Check the box `Create a git` to set up a local git.

Option 2: `usethis`

- Use `usethis::create_package()`, which will set up a template package directory in the specified folder.
- You have to take care of version control yourself (recommendation: initiate project on GitHub first).

Example

```
R> create_package("overviewR", open = FALSE)
```

```
✓ Creating 'overviewR/'
✓ Setting active project to '/Users/simonmunzert/github/intro-to-data-science-21/lectures/10-debugging-automatio
✓ Creating 'R/'
✓ Writing 'DESCRIPTION'
Package: overviewR
Title: What the Package Does (One Line, Title Case)
Version: 0.0.0.9000
```

Step 2: Set up your package (cont.)

Basic components

1. The `DESCRIPTION` file
    - stores metadata about the package
    - lists dependencies if any
    - is pre-generated by `roxygen2`

Example

```
Package: overviewR
Title: What the Package Does (One Line, Title Case)
Version: 0.0.0.9000
Authors@R:
    person(given = "First",
           family = "Last",
           role = c("aut", "cre"),
           email = "first.last@example.com",
           comment = c(ORCID = "YOUR-ORCID-ID"))
Description: What the package does (one paragraph).
License: `use_mit_license()`, `use_gpl3_license()` or friends to pick a
    license
```

Step 2: Set up your package (cont.)

Basic components

1. The `DESCRIPTION` file
    - stores metadata about the package
    - lists dependencies if any
    - is pre-generated by `roxygen2`
    - it will later look like this

Example

```
Type: Package
Package: overviewR
Title: Easily Extracting Information About Your Data
Version: 0.0.2
Authors@R: c(
    person("Cosima", "Meyer", email = "XX@XX.com", role = c("cre","aut")),
    person("Dennis", "Hammerschmidt", email = "XX@XX.com", role = "aut"))
Description: Makes it easy to display descriptive information on
    a data set.  Getting an easy overview of a data set by displaying and
    visualizing sample information in different tables (e.g., time and
    scope conditions).  The package also provides publishable TeX code to
```

Step 2: Set up your package (cont.)

Basic components

1. The `DESCRIPTION` file
    - stores metadata about the package
    - lists dependencies if any
    - is pre-generated by `roxygen2`
    - it will later look like this
    - and displayed online like this

Example

overviewR: Easily Extracting Information About Your Data

Makes it easy to display descriptive information on a data set. Getting an easy overview of a data set by displaying and visualizing sample information in different tables (e.g., time and scope conditions). The package also provides publishable 'LaTeX' code to present the sample information.

Version:	0.0.7
Depends:	R (≥ 3.5.0)
Imports:	dplyr (≥ 1.0.0), ggplot2 (≥ 3.3.2), tibble (≥ 3.0.1)
Suggests:	covr, devtools, knitr, pkgdown, rmarkdown, spelling, testthat
Published:	2020-11-23
Author:	Cosima Meyer [cre, aut], Dennis Hammerschmidt [aut]
Maintainer:	Cosima Meyer <cosima.meyer at gmail.com>

Step 2: Set up your package (cont.)

Basic components

1. The `DESCRIPTION` file
    - stores metadata about the package
    - lists dependencies if any
    - is pre-generated by `roxygen2`
    - it will later look like this
    - and displayed online like this
2. The `NAMESPACE` file
    - will later contain information on exported and imported functions.
    - helps you manage (and avoid) function clashes
    - will be populated automatically using `devtools::document()`

Example

```
# Generated by roxygen2: do not edit by hand

export(overview_crossplot)
export(overview_crosstab)
export(overview_heat)
```

Step 2: Set up your package (cont.)

Basic components

1. The `DESCRIPTION` file
   - stores metadata about the package
   - lists dependencies if any
   - is pre-generated by `roxygen2`
   - it will later look like this
   - and displayed online like this
2. The `NAMESPACE` file
   - will later contain information on exported and imported functions.
   - helps you manage (and avoid) function clashes
   - will be populated automatically using `devtools::document()`
3. The **R** folder
   - this is where all the functions you will create go

Step 3: Fill your package with life

Adding functions

The folder **R** contains all your functions and each function is saved in a new R file where the function name and the file name are the same.

In the preamble of this file, we can add information on the function. This information will be used to render the help files.

Example

```
#' @title overview_tab
#'
#' @description Provides an overview table for the time and scope conditions of
#'     a data set
#'
#' @param dat A data set object
#' @param id Scope (e.g., country codes or individual IDs)
#' @param time Time (e.g., time periods are given by years, months, ... )
#'
#' @return A data frame object that contains a summary of a sample that
#'     can later be converted to a TeX output using \code{overview_print}
#' @examples
```

Step 3: Fill your package with life (cont.)

Adding functions

The folder **R** contains all your functions and each function is saved in a new R file where the function name and the file name are the same.

In the preamble of this file, we can add information on the function. This information will be used to render the help files.

When you execute `devtools::document()`, R automatically generates the respective help file in man as well as the new `NAMESPACE` file.

Example

overview_tab {overviewR} R Documentation

overview_tab

Description

Provides an overview table for the time and scope conditions of a data set

Usage

```
overview_tab(dat, id, time)
```

Step 6: Install your package!

Installing a local package

We are now ready to load a developmental version of the package. This works with `devtools::install()`, which will also try to install dependencies of the package from CRAN, if they're not already installed.

You need to run this from the parent working directory that contains the package folder.

We're now ready to call functions from the package.

Example

```
R> install("overviewR")
```

```
✓  checking for file '/Users/simonmunzert/github/intro-to-data-science-21/lectures/10-debugging-automation/examp
─  preparing 'overviewR':
✓  checking DESCRIPTION meta-information ...
─  checking for LF line-endings in source and make files and shell scripts
─  checking for empty or unneeded directories
   Omitted 'LazyData' from DESCRIPTION
─  building 'overviewR_0.0.0.9000.tar.gz'

Running /Library/Frameworks/R.framework/Resources/bin/R CMD INSTALL \
```

Steps 3-6

We skipped a couple of important (and some optional) steps now, including:

- Build and check a package, clean up → `devtools::check()`
- Iterative loading and testing → `devtools::load_all()`
- Adding unit tests → `usethis::use_testthat()`
- Import functions from other packages (CRAN package dependency) → `usethis::use_package()`
- Git version control and collaboration → `usethis::use_github()`
- Add a proper public description → `usethis::use_readme_rmd()`
- Build PDF manual → `devtools::build_manual()`
- Add vignettes → `usethis::use_vignette()`
- Add a licence → `usethis::use_gpl_license()`, `usethis::use_mit_license()`, ...
- Convert into a single bundled file (binary or zipped) → `devtools::build()`
- Submit to CRAN → `devtools::release()`
- Build website for your package → `pkgdown::build_site()`

Be sure to check out the motivating example and more resources (next slide).

Writing R packages - FAQ

Is learning this worth the time?

Yes.

Where can I learn more?

Glad that you're asking! There's tons of materials out there. Apart from the used tutorial and the R packages book, have a look at the devtools cheatsheet and another overview over at RStudio. Knowing how to turn a package into a website within minutes is fascinating, too.

When do we need a package, and when is a GitHub repo simply enough?

Do you think of your work as a project or a product? If it's the latter, maybe a package is right for you. (But... a research paper is also a product, right? 🤯)