



A journey through the Tidyverse

Thomas Hütter • Explorer of data



A journey through the Tidyverse

Thomas Hütter

Data Saturday Guatemala - virtual 2021



A journey through the TidyveRse

Thomas Hütter, Diplom-Betriebswirt

- Application developer, consultant, accidental DBA, author
- Worked at consultancies, ISVs, end user companies
- SQL Server > 6.5, former „Navision“ > 3.0, R > 3.1.2
- Speaker at SQL events around Europe



 @DerFredo <https://twitter.com/DerFredo>

 de.linkedin.com/in/derfredo

 www.xing.com/profile/Thomas_Huetter



JOIN !
THE DATABASE CONFERENCE



Pure expertise at the
SQL Server Konferenz 2018
FEB 26th - 28th, 2018 | DARMSTADT

Sign up now at sqlkonferenz.de



Agenda

- Prerequisites: R base system, IDE, Tidyverse packages
- The Tidyverse concept: why and what?
- The Tidyverse components: packages and demos
- Wrap-up, ressources & credits, Q&A

Pre-Prerequisites

- Programming language for statistical computing and visualization, widely used by statisticians, data miners, analysts, data scientists
- Created by **R**oss Ihaka and **R**obert Gentleman, Uni Auckland, in 1993 as an open source implementation of the (1970s) S language
- GNU project, maintained by the R Foundation for Statistical Computing, compiled builds für Mac OS, Linux, Windows, supported by R Consortium
- Extensible through user-created packages, > 17000 available on CRAN
- Commercial support, e.g. since 2007 by Revolution Analytics, acquired by Microsoft in 2015, now provide Microsoft R Open, R Server
- IDEs: R.App, RStudio, MS R Tools for VisualStudio (< Version 2019)
- Support for R now in SQL Server (R services), Power BI, Azure ML



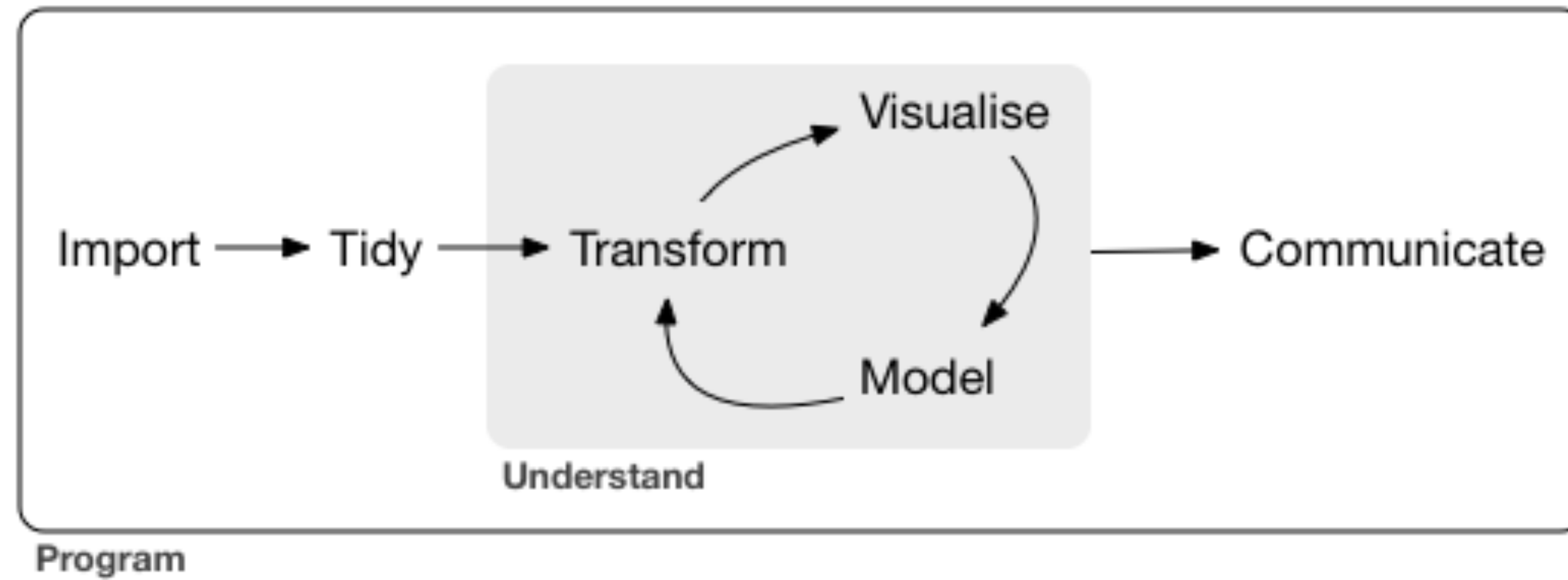
Prerequisites



- You already have an idea what R can be used for
- Install the R base system, available from <https://cran.r-project.org/index.html>
- Get the IDE of your choice, in my case RStudio: <https://www.rstudio.com/products/rstudio/download/>
- Of course we'll need the Tidyverse `install.packages("tidyverse")`
then `library(tidyverse)` will load the core packages
- Let the fun begin!

The Tidyverse concept: why and what?

What a typical data analysis/data science project may look like



The components of the Tidyverse cover these tasks and can help you to accomplish them in a concise manner.

The Tidyverse concept: why and what?

- „The goal of these principles is to provide a uniform interface so that Tidyverse packages work together naturally“. *)
- Tidy data is data stored in a consistent, reusable structure, preferably in rectangular datasets, where ideally:
1 row = 1 observation and 1 column = 1 variable.
- No need for conversions in the middle of analysis.
- You can concentrate on your data!

*) © 2017 Wickham: „The tidy tools manifesto“

The Tidyverse components

- Import: readr, readxl, DBI, haven, httr, XML2, rvest, jsonlite
- Tidy: tibble, tidyr
- Transform: dplyr, stringr, forcats, lubridate, hms, blob
- Visualize: ggplot2
- Model: tidymodels package collection (replacing modelr)
- Communicate: *R Markdown*, ggplot2, *Shiny*
- Program: purrr, magrittr, glue

Packages are: Core, additional, *non-Tidyverse*

Import

- readr: mainly imports flat files like csv and others
- readxl: import Excel files into R (xls and xlsx)
- DBI: database interface, encapsulates low-level driver work
- haven: import/export files from SPSS, Stata, SAS systems
- httr: handles http requests as GET() and POST()
- XML2: parse XML files
- rvest: scrape („harvest“) web pages; wraps httr and XML2
- jsonlite: JSON anyone? Parse, generate, stream, ...

Transform

- dplyr: „A grammar of data manipulation“, provides functions according to the verbs of basic data manipulation: select, filter, arrange, mutate, summarize ...
- stringr: simple, consistent wrapper for string operations
- forcats: tools for working with factors (reordering levels etc.)
- lubridate: functions to work with date-times and time-spans
- hms: a „pretty“ time-of-day class
- blob: for storing binary („blob“) data

Tidy

- tibble: „Tibbles are a modern take on data frames“
 - never change input types (strings <-> factors)
 - never adjust variable names (allow crazy names)
 - no row.names()
 - prettier print output
- tidyr: easily tidy data mainly with these functions:
 - gather() collapses multiple columns into key-value pairs
converts wide -> long
 - spread() does the inverse of gather()
converts long -> wide
 - watch out for pivot_longer() and pivot_wider functions !

Visualize

- ggplot2: create elegant data visualizations using the „grammar of graphics“
 - initialize a plot stating the data frame to be used
 - define the aesthetic mappings per plot or per layer
 - add layers of geometric representation of the data
 - optionally add other options: scales, themes, facets

Model

The tidymodels collection of packages (former modelr):

- rsample: efficient data splitting and resampling
- parsnip: common API to modeling/analysis functions
- recipes: tidy interface to data pre-processing tools
- workflows: bundle pre-processing, modeling, post-processing
- tune: helps optimizing hyperparameters
- yardstick: performance metrics for models
- broom: converts statistical info into user-friendly formats
- dials: creates and manages tuning parameters

Communicate

- *R Markdown*: package and tool to render markdown files to (X)HTML, pdf or other output formats
- ggplot2: see „Visualize“ section
- *Shiny*: a framework for easily building interactive web applications in R with minimal effort

Program

- `magrittr`: the forward pipe operator `%>%` for R, chaining of commands by forwarding the result of one function/expression into the next function call
- `purrr`: tools for functional programming, e.g.
 - using `map_*()` functions instead of loops or `apply()`
 - error handling: `safely()`, `possibly()`, `quietly()`
- `glue`: provides alternatives to `paste()` for easier combination of data and strings

Tidyverse wrap-up

- „Tidy datasets are all alike, but every messy dataset is messy in it's own way“ (Hadley Wickham)
- To avoid this, engage the tidy data philosophy and tools
- So preferably convert 'messy' to tidy data, where tidy means:
 - one variable per column
 - one observation per row
 - each type of observational unit is a tibble
- Easier passing of data between the tools / packages
- Make the tools work together in a natural way

Resources & credits

- The Tidyverse web site <http://www.tidyverse.org/>
- R for Data Science, Hadley Wickham & Garrett Grolemund, O'Reilly, ISBN 978-1491910399, also at <http://r4ds.had.co.nz>
- The tidy tools manifesto
<https://mran.microsoft.com/web/packages/tidyverse/vignettes/manifesto.html>
- More on the Shiny framework <http://shiny.rstudio.com/>
and on R markdown <http://rmarkdown.rstudio.com/>
- World economic outlook database: International monetary fund
<http://www.imf.org/external/pubs/ft/weo/2019/02/weodata/download.aspx>
- F1 data from: <http://www.formel1.de/rennergebnisse/wm-stand/2016/>

A journey through the Tidyverse

Thank you for your interest & keep in touch:

 @DerFredo <https://twitter.com/DerFredo>

 de.linkedin.com/in/derfredo

 www.xing.com/profile/Thomas_Huetter



This file and all demo scripts can be found at:

<http://j.mp/DerFredoGuatemala2021>