Technische Universität München

Business Analytics

Tutorial 7: Data Preparation and Modelling Workflow

Decision Sciences & Systems (DSS)

Department of Informatics

Technical University of Munich

Analytics Cup – Preview

Dates

- **Start**: January 7 **Submission Deadline**: February 1
- 18 - 22 January: tutor and TA office hours for Analytics Cup support (with individual team)
- By February 8: Notification of Winning Teams (prepare to present your solutions in the lectures)
  Check-In with teams with problematic solutions
- February 11, 2020: Analytics Cup End presentation in the final lecture (probably live, tbc)
  (well performing teams are expected to present, if asked)

Regulations *(Details will follow)*

- Groups of 1-4 students, no cooperation outside your group.
- Well defined analytical task and data set, you will submit solutions via an online platform
- There will be specific rules what you are and are not allowed to do. (Details on January 7th)

Grading *(Details will follow)*

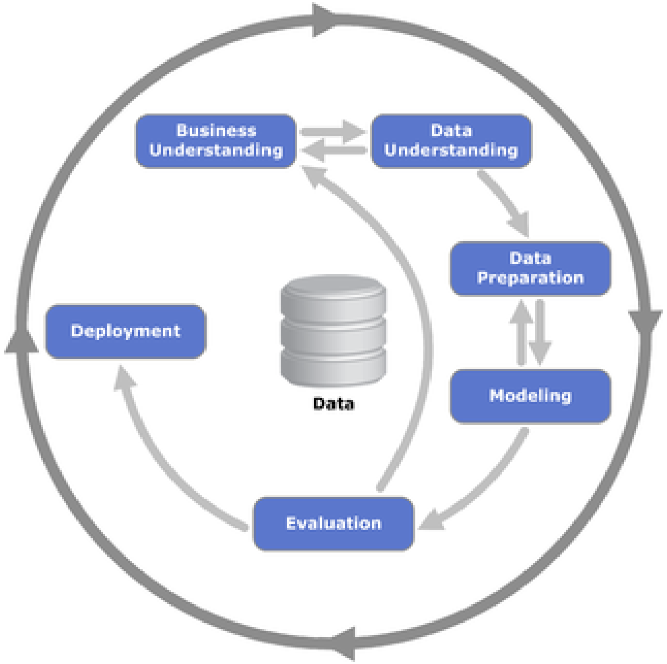
- Completing the project can *only* improve your grade, it will never be a disadvantage
- If AC grade ("bonus") is better than your Exam Grade, your final grade will be 67% exam, 33% AC
- Bonus only valid for an exam in 2021 ***passed on your first try*** (can be **either** Final **or** Deferred exam)!
- Bonus only counts toward exams *this* semester. AC bonuses achieved in earlier years do not count.
  (Some Covid-19 related exceptions for bonus from 2019. Talk to us in advance, if you were affected.)

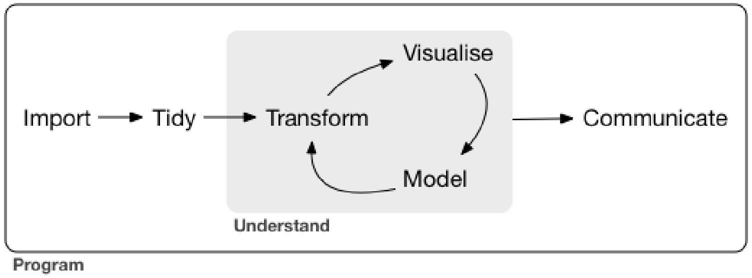
Detailed rules will be announced on January 7th.

Topics this Week

- Goal: Provide tools for complete Data Analytics workflow and prepare you for the Analytics Cup
- **Data Cleaning and Preparation**
  - Recap of Week 1 concepts
  - Tidy Data and Pivoting
  - Relational Data and Joins
- **Meta-Machine Learning in R with tidymodels**
  - Building easily reproducible and modifyable analytics pipelines using tidymodels packages.

Data Analytics Process





Crisp-DM Process

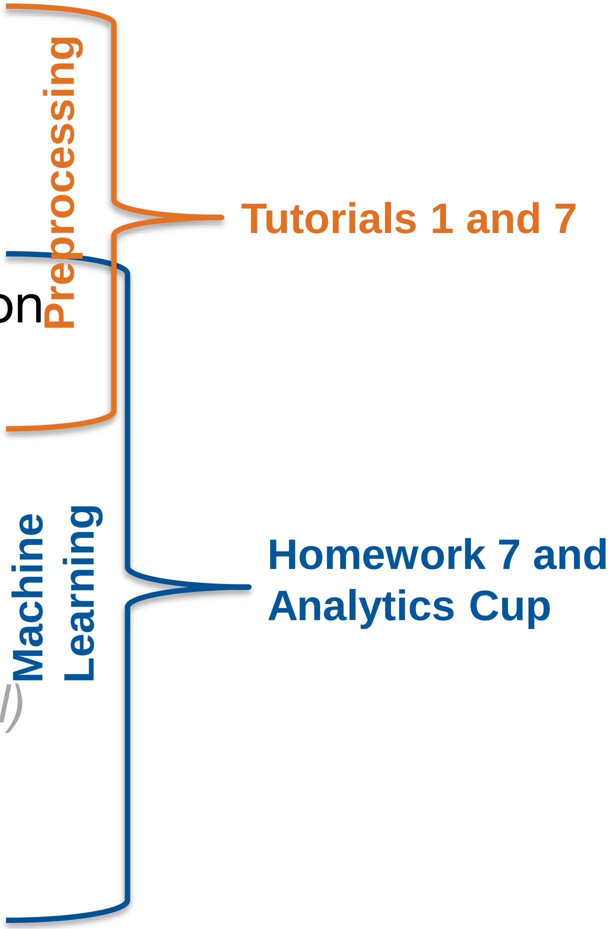
https://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png

Data Science Workflow

https://r4ds.had.co.nz/introduction.html

Data Analytics / Machine Learning Programming Workflow

1. Data Loading (week 1)
2. Data Exploration (week 1)
3. Data Cleaning, Preparation and Imputation
4. Feature-Selection and -Engineering
5. Modeling
   - Task, Algorithm, Resampling strategy
6. Training and Evaluating the Model
7. Tuning and Refining *(not shown in tutorial)*
8. Predict on unseen data, write output



Preprocessing

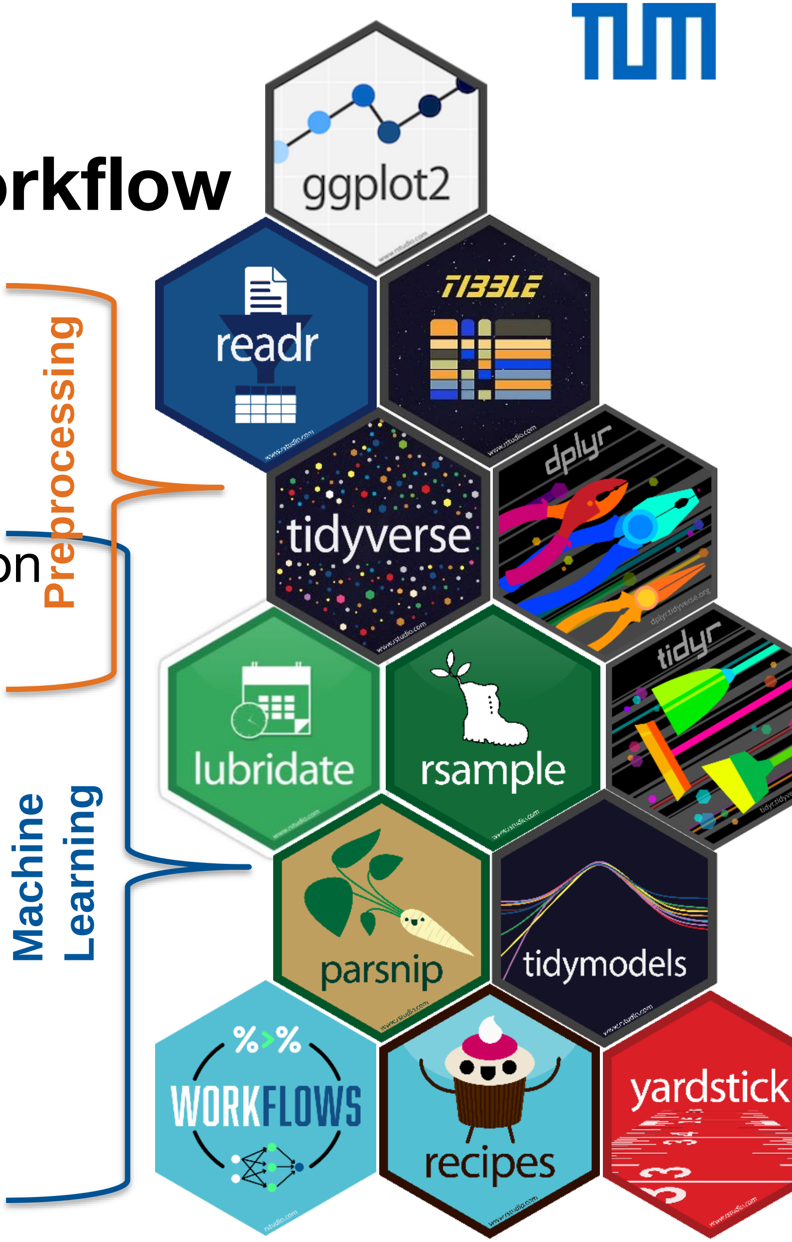
Tutorials 1 and 7

Machine Learning

Homework 7 and Analytics Cup

Tidyverse and Tidymodels packages in the Modelling workflow

1. Data Loading
2. Data Exploration *(not part of tutorial)*
3. Data Cleaning, Preparation and Imputation
4. Feature-Selection and -Engineering
5. Modeling
   - Task, Algorithm, Resampling strategy
6. Training and Evaluating the Model
7. Tuning and Refining *(not part of tutorial)*
8. Predict on unseen data, write output



Preprocessing

Machine Learning

Further reading:
Wickham and Goremund: R for Data Science https://r4ds.had.co.nz/
Tidymodels Documentation: https://tidymodels.org

Data Loading (compare in Week 1 Tutorials)

tibble is a wrapper around R’s data.frame and provides:

- Better printing
- Better debugging
  (warnings for type safety etc)
- Interfaces to other backends with familiar API
  (not relevant for us)
  (Databases, Spark, data.table)

*Every tibble **is** a data.frame. Everything you’ve learned to do with data.frames also works with tibbles.*

More at: https://tibble.tidyverse.org/

readr provides alternative implementations of io operations, e.g. read_csv() to replace read.csv()
Readr functions

- Are **faster** than base R’s counterparts
- Have sensible defaults (no row names, StringsAsFactors=FALSE)
- Better type detection (e.g. date parsing)
- Create tibbles instead of data.frames

More at: https://readr.tidyverse.org/

Data Cleaning + Wrangling

Dplyr and **tidyr** provide tools for easy, consistent and efficient transformation of tabular data
(such as tibbles and data.frames)

Lubridate provides convenience functions for easier working with dates, e.g.

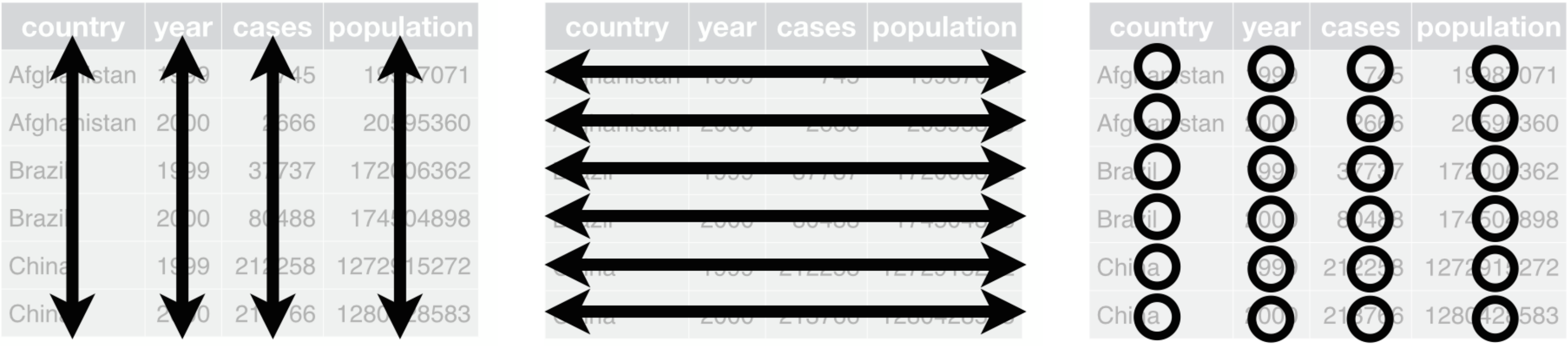
- Parsing dates and times from Strings
- Calculating time intervals and differences
- Time Zone Conversion
- …

Further Reading: R4DS, Section “Wrangle”
Cheat Sheet:
https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf

Cheat Sheet:
https://github.com/rstudio/cheatsheets/raw/master/lubridate.pdf

Tidy Data

Recommended Reading: R4DS, Chapter 12



variables observations values

https://r4ds.had.co.nz/tidy-data.html

Tabular data is *tidy* if:

- 1. each variable/feature is in a single column
- 2. each observation/instance is in a single row
- 3. each value is in a single cell

- Tidy data is required for most analysis and modeling tasks
- Tidying up a dataset is usually the first step of data cleaning.
- It can sometimes be ambiguous what constitutes a feature based on context: e.g. “Address” vs “Street | Number | ZIP | City | Country”

Untidy Data

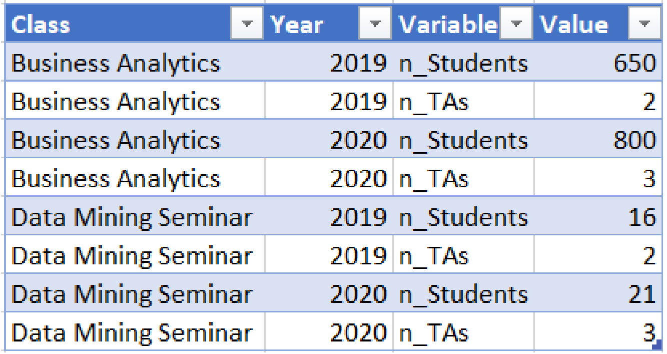
Reasons why most data in practice isn't tidy:

- Bad design
- Problems in the data collection process
- Table is optimized for something other than analysis (e.g. data entry, storage, fast processing, compliance with required formats, …)
- Pragmatic violations sometimes desirable
- …

"Long Format"

Advantages:

- Well suited for high throughput Big Data processing tasks, nontabular storage models
- Often useful in Vizualisation as intermediate result when creating a single plot comparing variables

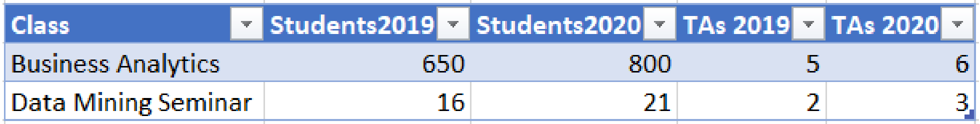


Class	Year	Variable	Value
Business Analytics	2019	n_Students	650
Business Analytics	2019	n_TAs	2
Business Analytics	2020	n_Students	800
Business Analytics	2020	n_TAs	3
Data Mining Seminar	2019	n_Students	16
Data Mining Seminar	2019	n_TAs	2
Data Mining Seminar	2020	n_Students	21
Data Mining Seminar	2020	n_TAs	3

"Wide Format"

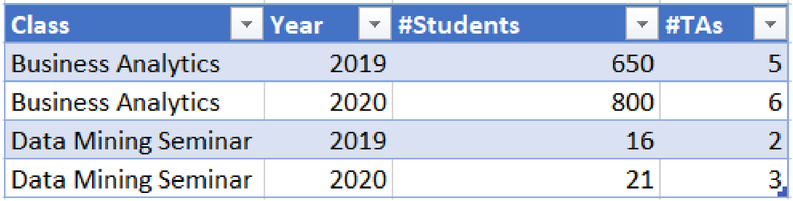
Advantages:

- Requires less storage space / smaller file sizes
- For small datasets, often more human-readible
- Easier manual data entry



Class	Students2019	Students2020	TAs 2019	TAs 2020
Business Analytics	650	800	5	6
Data Mining Seminar	16	21	2	3

"Tidy Format"



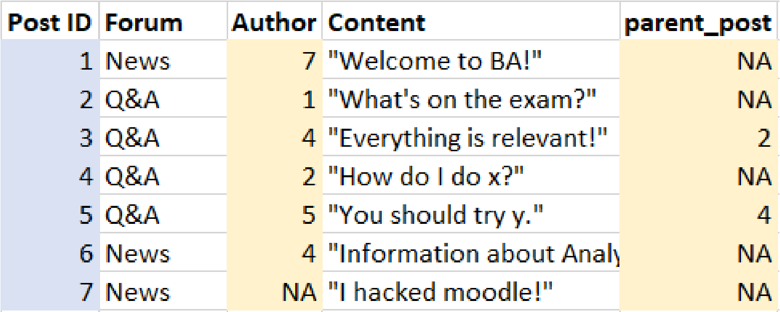
Class	Year	#Students	#TAs
Business Analytics	2019	650	5
Business Analytics	2020	800	6
Data Mining Seminar	2019	16	2
Data Mining Seminar	2020	21	3

Converting between these formats is called **pivoting**. In R, you can use the \`pivot_longer\` and \`pivot_wider\` Functions from the \`tidyr\` package.

Relational Data

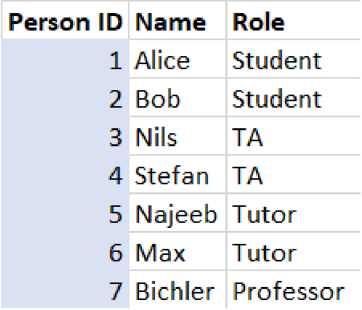
Recommended Reading: R4DS, Chapter 13

moodle_posts.csv



Post ID	Forum	Author	Content	parent_post
1	News	7	"Welcome to BA!"	NA
2	Q&A	1	"What's on the exam?"	NA
3	Q&A	4	"Everything is relevant!"	2
4	Q&A	2	"How do I do x?"	NA
5	Q&A	5	"You should try y."	4
6	News	4	"Information about Analy	NA
7	News	NA	"I hacked moodle!"	NA

participants.csv



Person ID	Name	Role
1	Alice	Student
2	Bob	Student
3	Nils	TA
4	Stefan	TA
5	Najeeb	Tutor
6	Max	Tutor
7	Bichler	Professor

Often, data is spread over multiple tables. **Join** operations let you combine them.

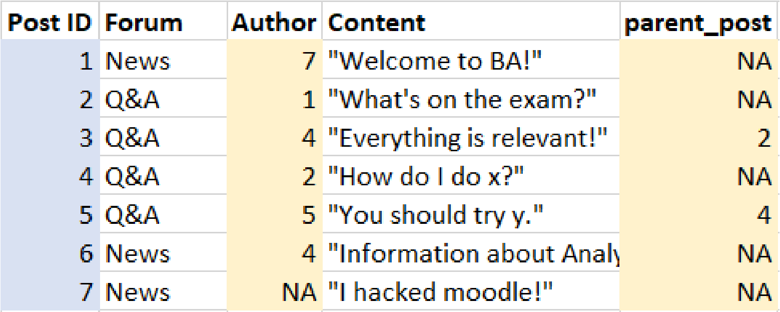
Relational data has columns that are **primary keys** (uniquely identify observation in same table) or **foreign keys** (refer to an observation in another table) that can be used to combine tables.

Not all data is explicitly relational. One can also join on non-key attributes.

Relational Data

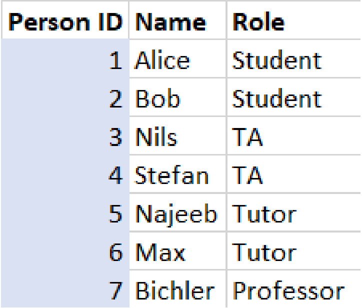
Recommended Reading: R4DS, Chapter 13

moodle_posts.csv



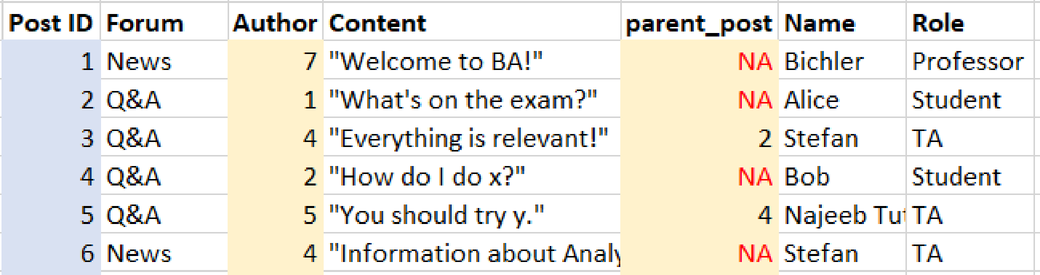
Post ID	Forum	Author	Content	parent_post
1	News	7	"Welcome to BA!"	NA
2	Q&A	1	"What's on the exam?"	NA
3	Q&A	4	"Everything is relevant!"	2
4	Q&A	2	"How do I do x?"	NA
5	Q&A	5	"You should try y."	4
6	News	4	"Information about Analy	NA
7	News	NA	"I hacked moodle!"	NA

participants.csv



Person ID	Name	Role
1	Alice	Student
2	Bob	Student
3	Nils	TA
4	Stefan	TA
5	Najeeb	Tutor
6	Max	Tutor
7	Bichler	Professor

inner_join(moodle_posts, participants, by=c(“Author” = “Person ID”))

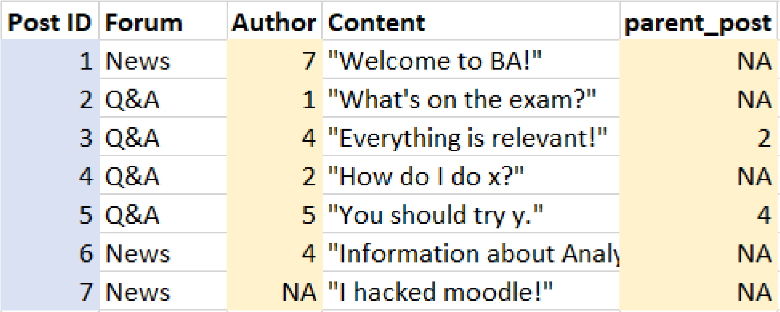


Post ID	Forum	Author	Content	parent_post	Name	Role
1	News	7	"Welcome to BA!"	NA	Bichler	Professor
2	Q&A	1	"What's on the exam?"	NA	Alice	Student
3	Q&A	4	"Everything is relevant!"	2	Stefan	TA
4	Q&A	2	"How do I do x?"	NA	Bob	Student
5	Q&A	5	"You should try y."	4	Najeeb Tu	TA
6	News	4	"Information about Analy	NA	Stefan	TA

Relational Data

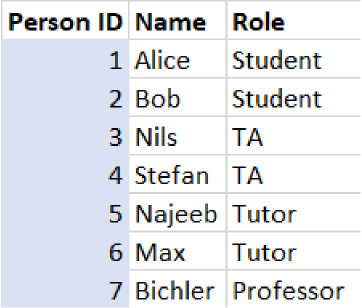
Recommended Reading: R4DS, Chapter 13

moodle_posts.csv



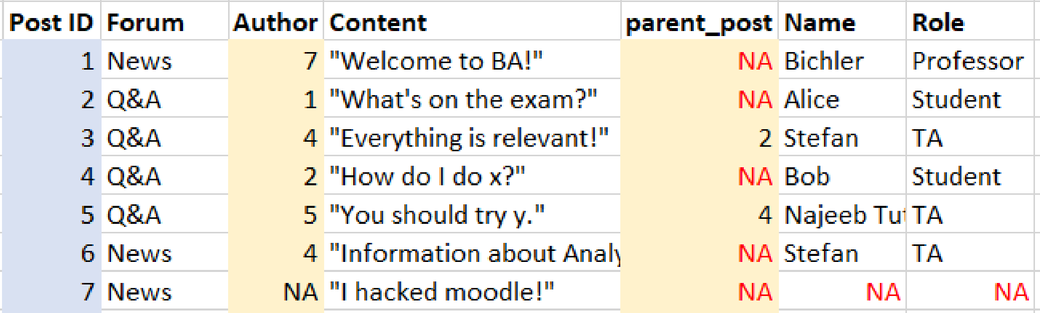
Post ID	Forum	Author	Content	parent_post
1	News	7	"Welcome to BA!"	NA
2	Q&A	1	"What's on the exam?"	NA
3	Q&A	4	"Everything is relevant!"	2
4	Q&A	2	"How do I do x?"	NA
5	Q&A	5	"You should try y."	4
6	News	4	"Information about Analy	NA
7	News	NA	"I hacked moodle!"	NA

participants.csv



Person ID	Name	Role
1	Alice	Student
2	Bob	Student
3	Nils	TA
4	Stefan	TA
5	Najeeb	Tutor
6	Max	Tutor
7	Bichler	Professor

left_join(moodle_posts, participants, by=c("Author" = "Person ID"))

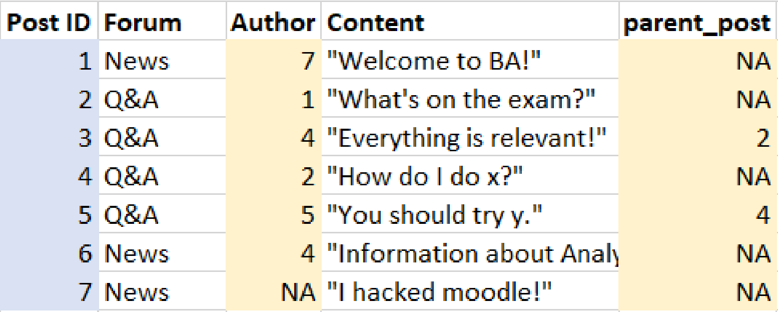


Post ID	Forum	Author	Content	parent_post	Name	Role
1	News	7	"Welcome to BA!"	NA	Bichler	Professor
2	Q&A	1	"What's on the exam?"	NA	Alice	Student
3	Q&A	4	"Everything is relevant!"	2	Stefan	TA
4	Q&A	2	"How do I do x?"	NA	Bob	Student
5	Q&A	5	"You should try y."	4	Najeeb Tu	TA
6	News	4	"Information about Analy	NA	Stefan	TA
7	News	NA	"I hacked moodle!"	NA	NA	NA

Relational Data

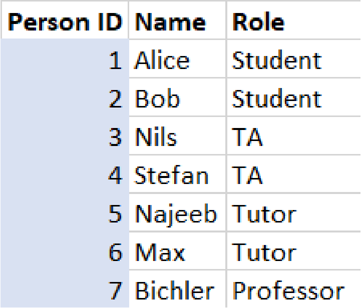
Recommended Reading: R4DS, Chapter 13

moodle_posts.csv



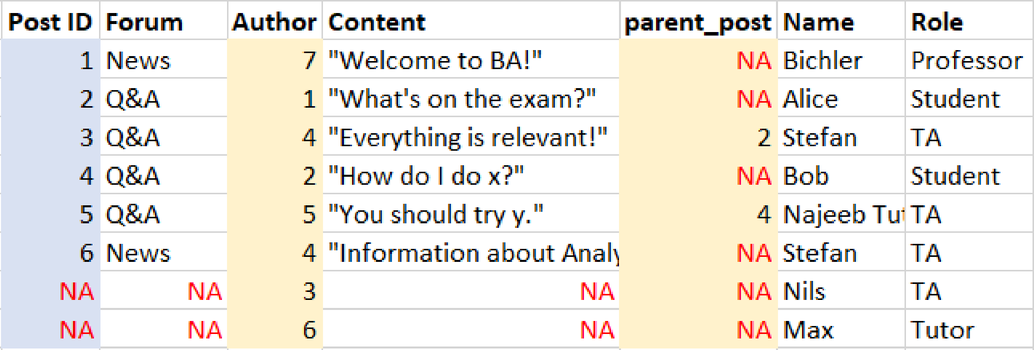
Post ID	Forum	Author	Content	parent_post
1	News	7	"Welcome to BA!"	NA
2	Q&A	1	"What's on the exam?"	NA
3	Q&A	4	"Everything is relevant!"	2
4	Q&A	2	"How do I do x?"	NA
5	Q&A	5	"You should try y."	4
6	News	4	"Information about Analy	NA
7	News	NA	"I hacked moodle!"	NA

participants.csv



Person ID	Name	Role
1	Alice	Student
2	Bob	Student
3	Nils	TA
4	Stefan	TA
5	Najeeb	Tutor
6	Max	Tutor
7	Bichler	Professor

right_join(moodle_posts, participants, by=c(“Author” = “Person ID”))

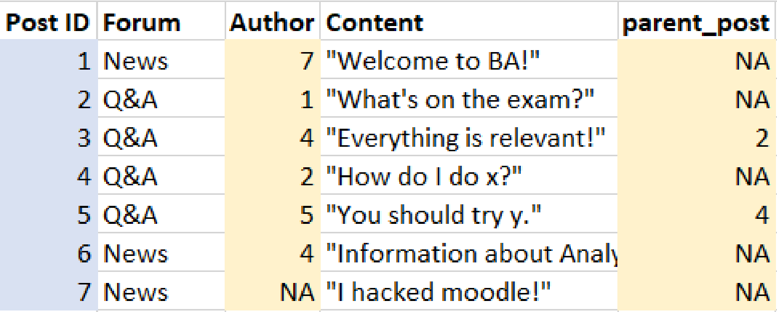


Post ID	Forum	Author	Content	parent_post	Name	Role
1	News	7	"Welcome to BA!"	NA	Bichler	Professor
2	Q&A	1	"What's on the exam?"	NA	Alice	Student
3	Q&A	4	"Everything is relevant!"	2	Stefan	TA
4	Q&A	2	"How do I do x?"	NA	Bob	Student
5	Q&A	5	"You should try y."	4	Najeeb Tu	TA
6	News	4	"Information about Analy	NA	Stefan	TA
NA	NA	3	NA	NA	Nils	TA
NA	NA	6	NA	NA	Max	Tutor

Relational Data

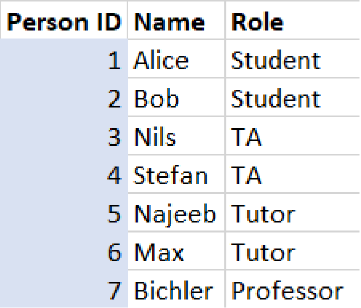
Recommended Reading: R4DS, Chapter 13

moodle_posts.csv



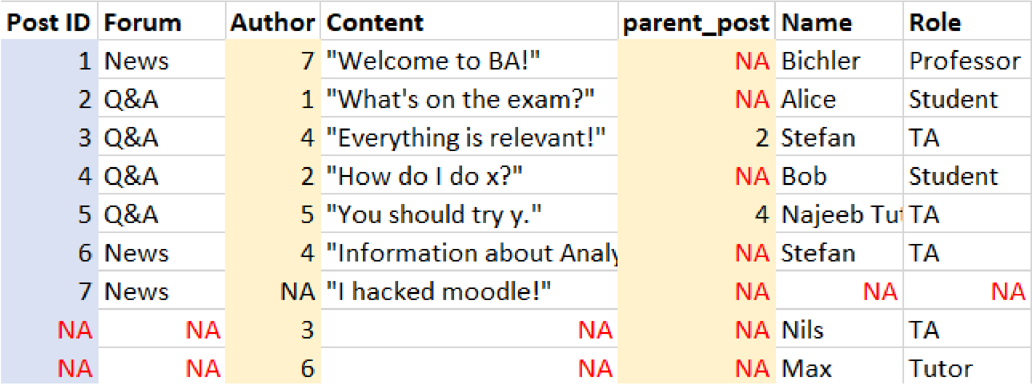
Post ID	Forum	Author	Content	parent_post
1	News	7	"Welcome to BA!"	NA
2	Q&A	1	"What's on the exam?"	NA
3	Q&A	4	"Everything is relevant!"	2
4	Q&A	2	"How do I do x?"	NA
5	Q&A	5	"You should try y."	4
6	News	4	"Information about Analy	NA
7	News	NA	"I hacked moodle!"	NA

participants.csv



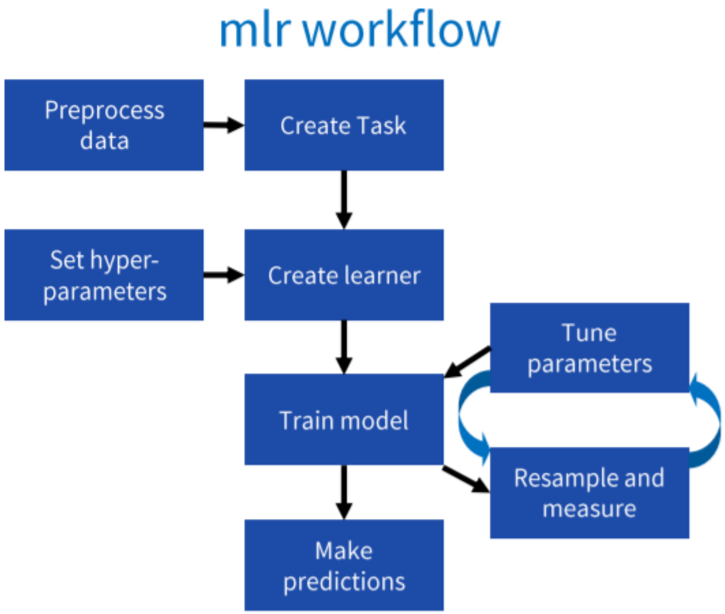
Person ID	Name	Role
1	Alice	Student
2	Bob	Student
3	Nils	TA
4	Stefan	TA
5	Najeeb	Tutor
6	Max	Tutor
7	Bichler	Professor

full_join(moodle_posts, participants, by=c(“Author” = “Person ID”))



Post ID	Forum	Author	Content	parent_post	Name	Role
1	News	7	"Welcome to BA!"	NA	Bichler	Professor
2	Q&A	1	"What's on the exam?"	NA	Alice	Student
3	Q&A	4	"Everything is relevant!"	2	Stefan	TA
4	Q&A	2	"How do I do x?"	NA	Bob	Student
5	Q&A	5	"You should try y."	4	Najeeb Tu	TA
6	News	4	"Information about Analy	NA	Stefan	TA
7	News	NA	"I hacked moodle!"	NA	NA	NA
NA	NA	3	NA	NA	Nils	TA
NA	NA	6	NA	NA	Max	Tutor

Meta Machine Learning



Source: https://mlr.mlr-org.com/

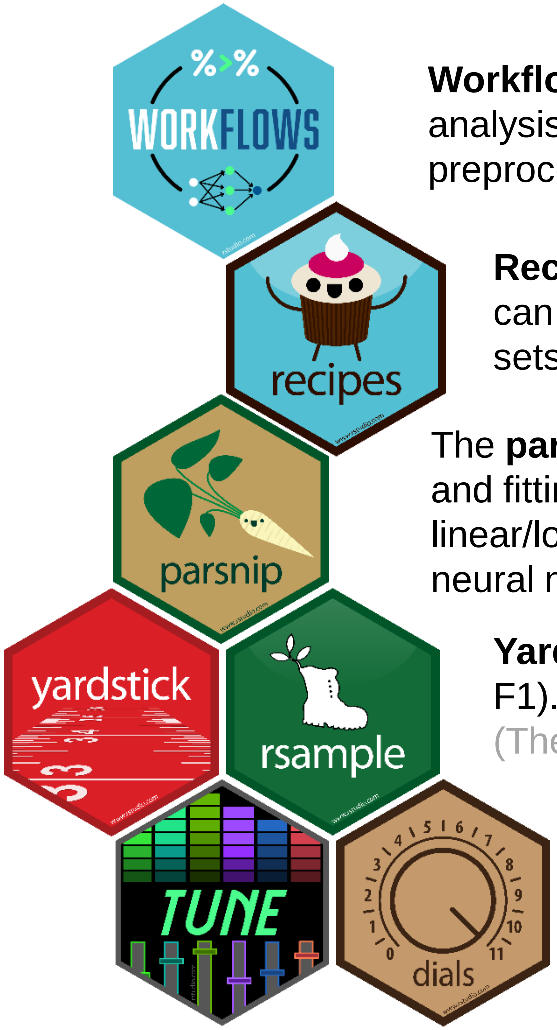
- **Problem:** implementation of specific functionality (models / algorithms, resampling, hyperparameter tuning) is spread across 100s of packages, each with their own specific interface

- **Meta machine learning** frameworks provide a unified user view. Common features:
  - Wrappers around third party backend packages, providing a unified interface and making it easy to switch out individual parts
  - Ability to create reproducible pipelines that can be consistently applied to different data without duplicate code

- Meta ML frameworks in R:
  - taught here: **tidymodels**
  - Alternatives: caret, mlr, mlr3, h2o, …

Homework: tidymodels case study

Further Reading: Documentation at https://www.tidymodels.org/



Workflows are rich objects that persist throughout the data analysis and keep track of components (tasks, data, preprocessing steps, model specifications, trained models).

Recipes define reproducible data preprocessing steps that can be applied to multiple data sources (e.g. train / test sets).

The **parsnip** package provides a unified model specification and fitting interface for ~36 backend packages, e.g. linear/logistig regression, decision trees, random forests, neural networsk, gradient boosting …

Yardstick provides easy to use methods for model evaluation (e.g. roc, F1). **Rsample** provides resampling methods (e.g. Cross-Validation). (These methods will be covered in week 8.)

Tune and **dials** provide methods to optimize your models settings / hyperparameters for the best possible performance. (We will not cover these in the homework, but you may want to use them in the Analytics Cup.)