Data Handling: Import, Cleaning and Visualisation

Lecture 10:

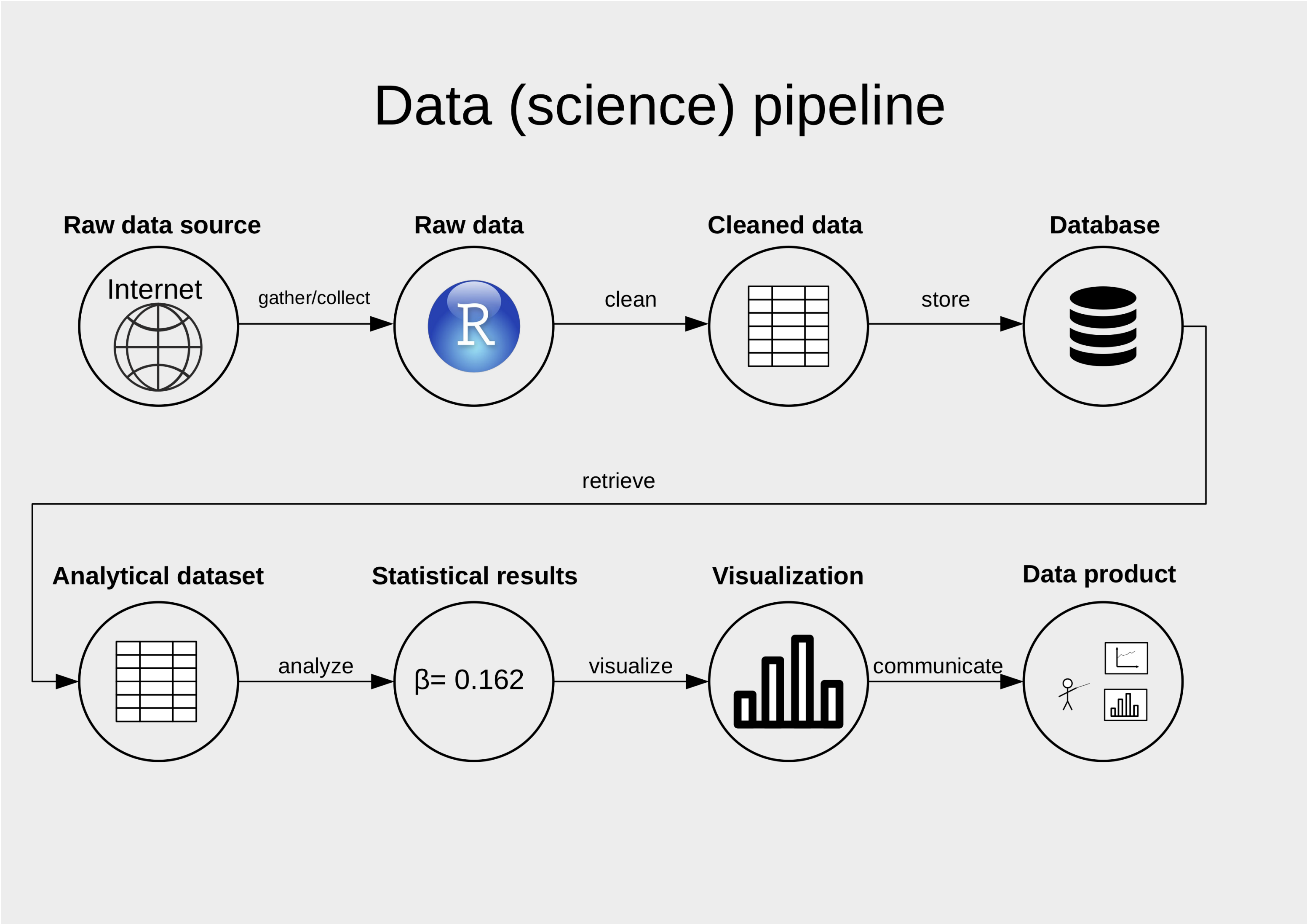
Basic Data Analysis with R

Prof. Dr. Ulrich Matter

Reminder

Send questions for the Q&A session (last lecture)

ulrich.matter@unisg.ch


Data (science) pipeline
Raw data source
Internet
gather/collect
Raw data
R
clean
Cleaned data
store
Database
retrieve
Analytical dataset
analyze
Statistical results
β= 0.162
visualize
Visualization
communicate
Data product

Data preparation/data cleaning

Goal of data preparation: Dataset is ready for analysis.

Key conditions:

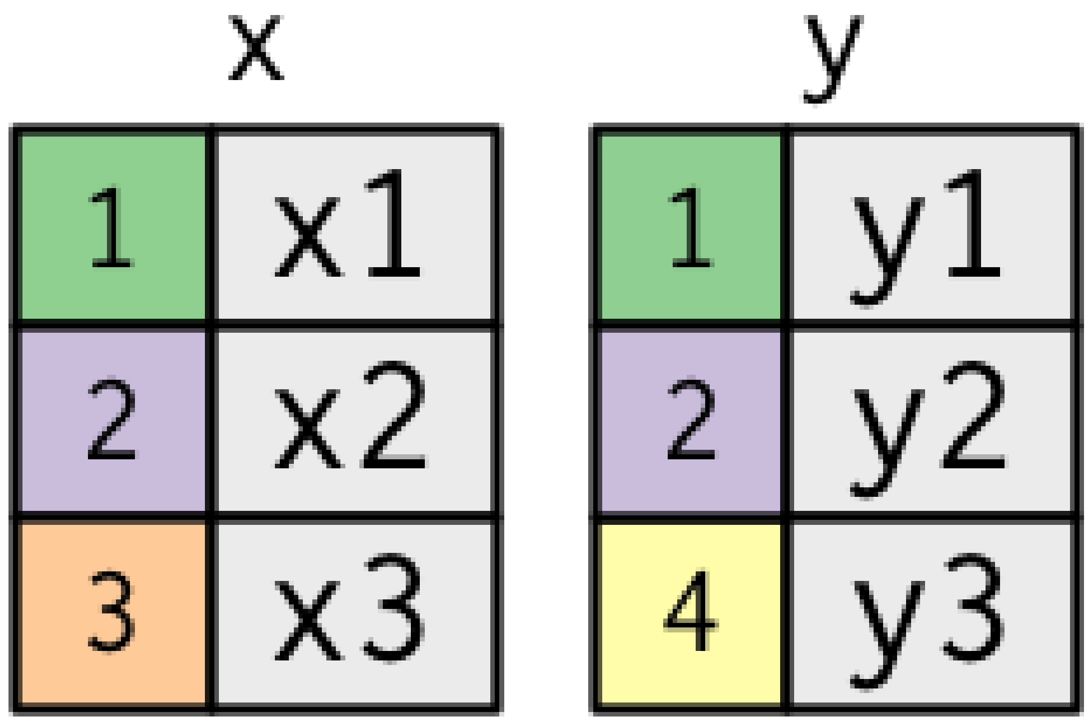
1. Data values are consistent/clean within each variable.
2. Variables are of proper data types.
3. Dataset is in ‘tidy’ (in long format)!

Merging (Joining) datasets

Combine data of two datasets in one dataset.

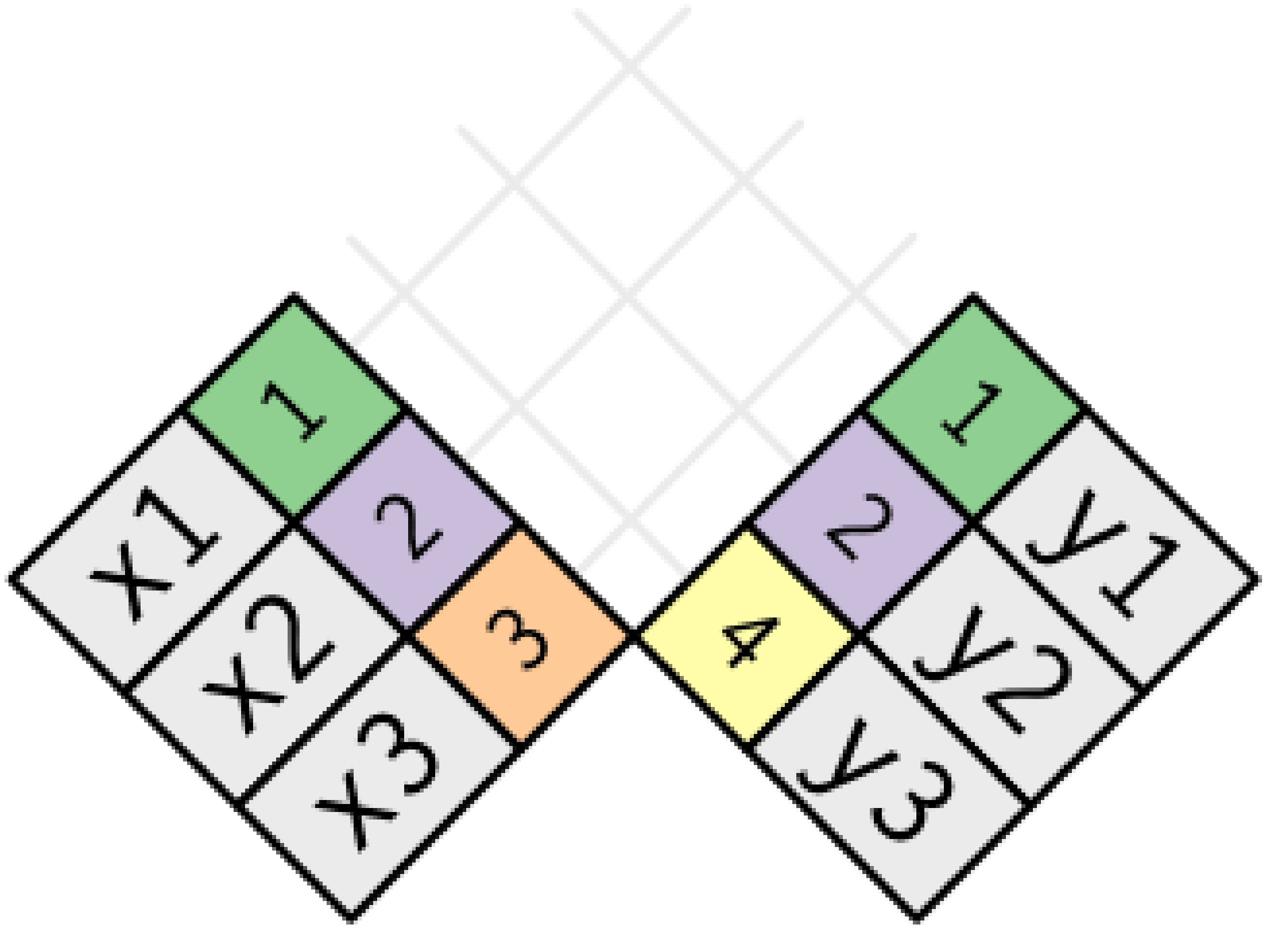
Needed: Unique identifiers for observations (‘keys’).

Merging (joining) datasets: concept



Join setup. Source: Wickham and Grolemund (2017), licensed under the Creative Commons Attribution-Share Alike 3.0 United States license.

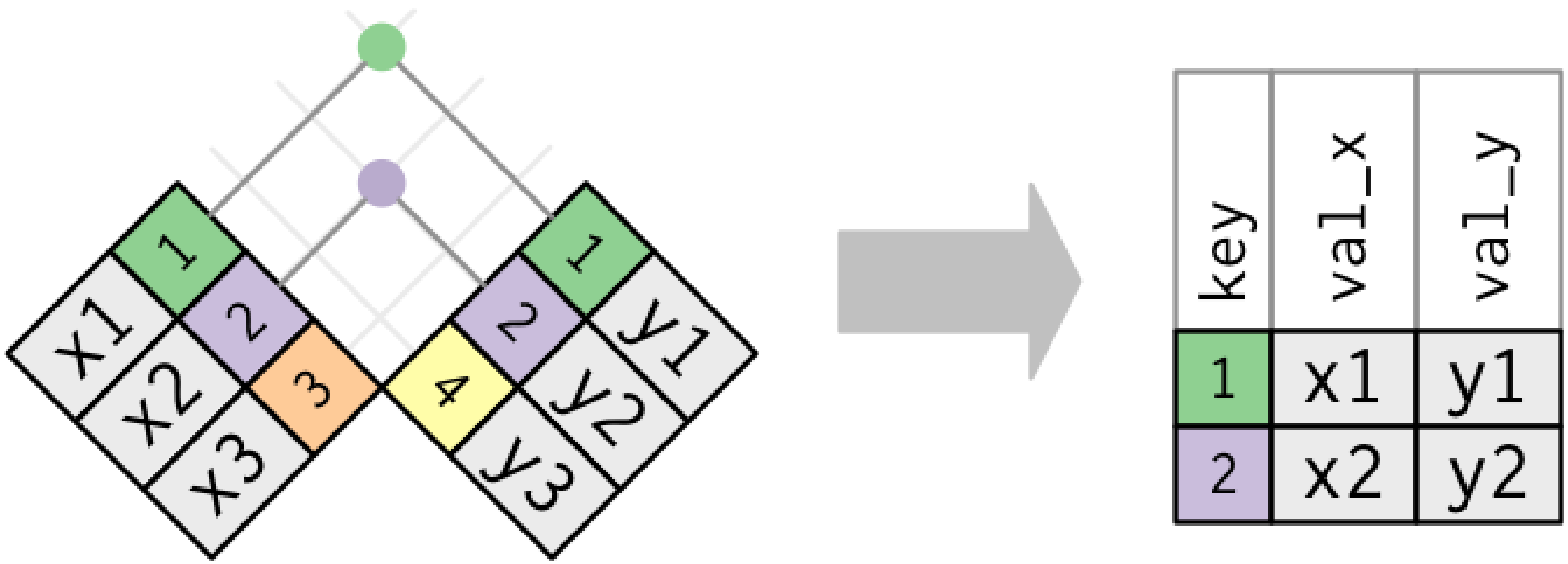
Merging (joining) datasets: concept



Join setup. Source: Wickham and Grolemund (2017), licensed under the Creative Commons Attribution-Share Alike 3.0 United States license.

Merging (joining) datasets: concept

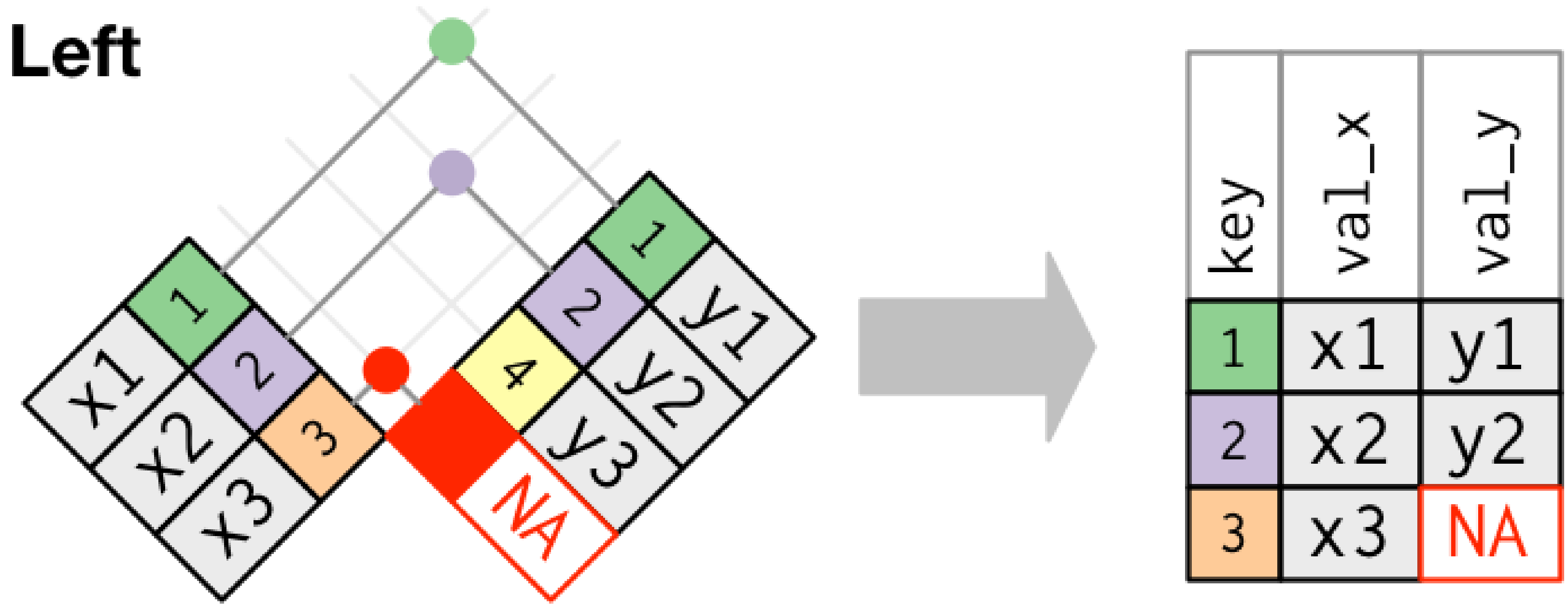
Merge: Inner join



Inner join. Source: Wickham and Grolemund (2017), licensed under the [Creative Commons Attribution-Share Alike 3.0 United States](#) license.

Merging (joining) datasets: concept

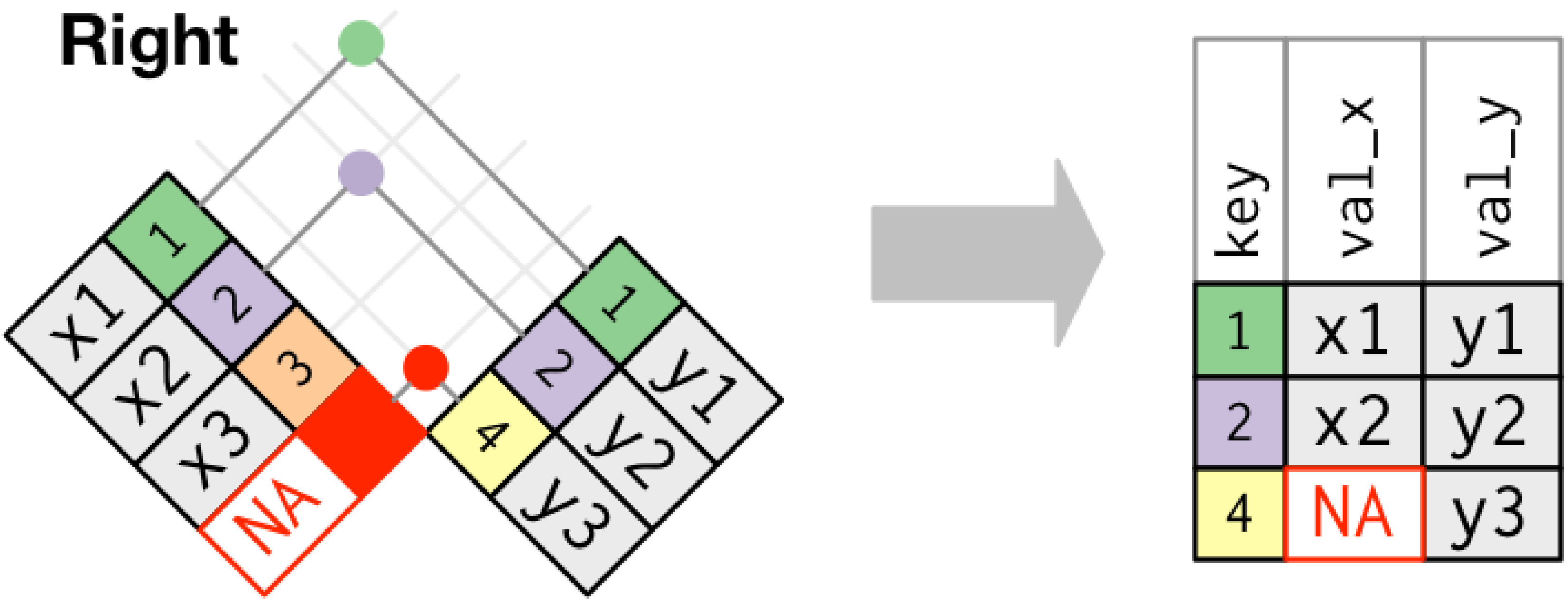
Merge all x: Left join



Outer join. Source: Wickham and Grolemund (2017), licensed under the Creative Commons Attribution-Share Alike 3.0 United States license.

Merging (joining) datasets: concept

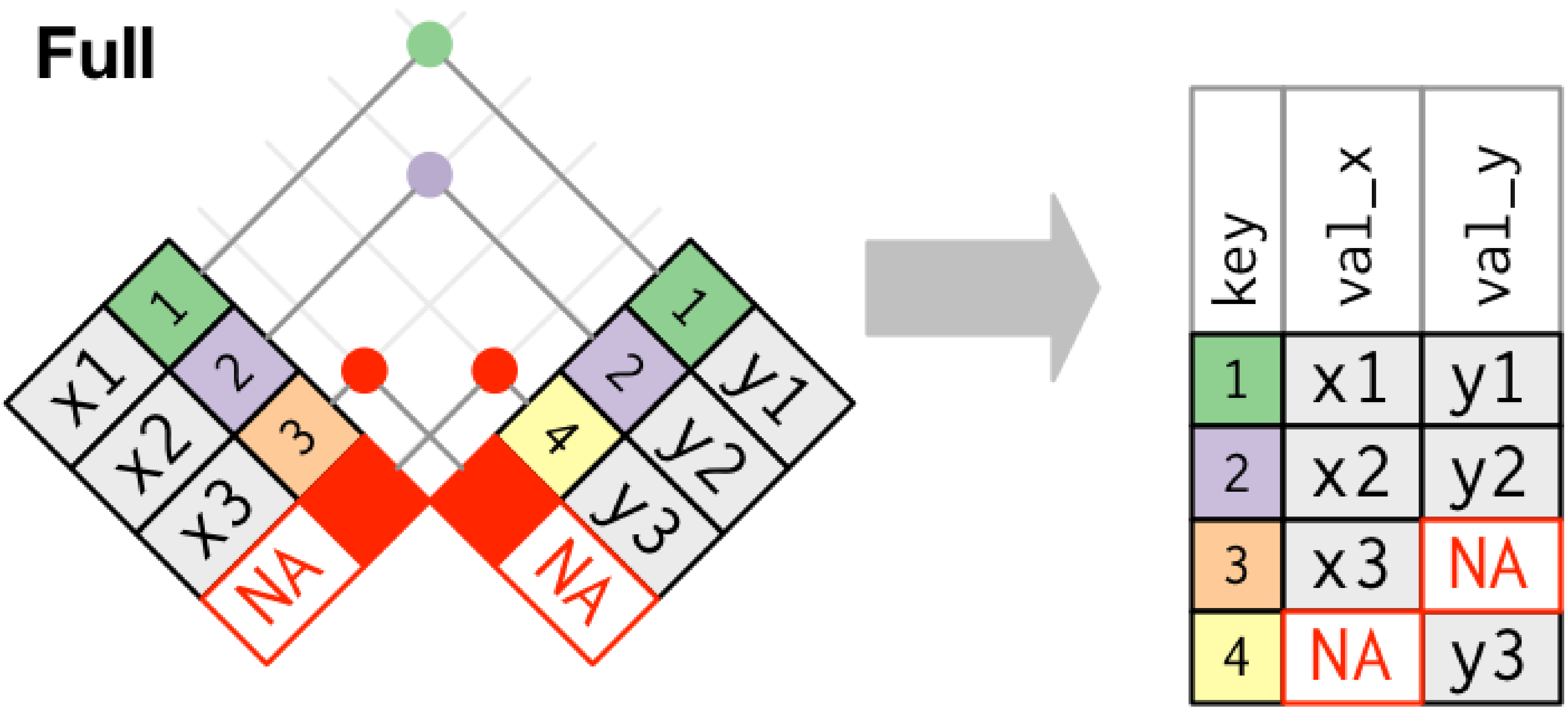
Merge all y: Right join



Outer join. Source: Wickham and Grolemund (2017), licensed under the [Creative Commons Attribution-Share Alike 3.0 United States](#) license.

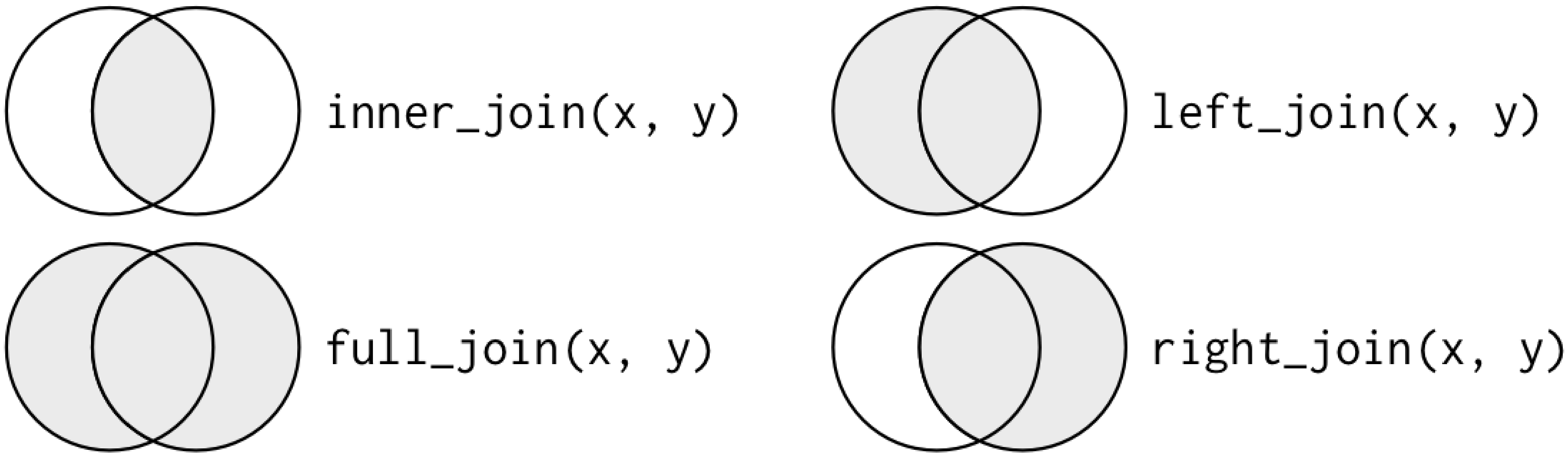
Merging (joining) datasets: concept

Merge all x and all y: Full join



Outer join. Source: Wickham and Grolemund (2017), licensed under the [Creative Commons Attribution-Share Alike 3.0 United States](#) license.

Merging (joining) datasets: concept



Join Venn Diagramm. Source: Wickham and Grolemund (2017), licensed under the [Creative Commons Attribution-Share Alike 3.0 United States](#) license.

Merging (joining) datasets: example

```r
# load packages
library(tidyverse)

# initiate data frame on persons personal spending
df_c <- data.frame(id = c(1:3,1:3),
                   money_spent= c(1000, 2000, 6000, 1500, 3000, 5500),
                   currency = c("CHF", "CHF", "USD", "EUR", "CHF", "USD"),
                   year=c(2017,2017,2017,2018,2018,2018))
df_c
```

```
##   id money_spent currency year
## 1  1        1000      CHF 2017
## 2  2        2000      CHF 2017
## 3  3        6000      USD 2017
## 4  1        1500      EUR 2018
## 5  2        3000      CHF 2018
## 6  3        5500      USD 2018
```

Merging (joining) datasets: example

```
# initiate data frame on persons' characteristics
df_p <- data.frame(id = 1:4,
                   first_name = c("Anna", "Betty", "Claire", "Diane"),
                   profession = c("Economist", "Data Scientist", "Data Scientist", "E
df_p
```

```
##   id first_name     profession
## 1  1       Anna      Economist
## 2  2      Betty Data Scientist
## 3  3     Claire Data Scientist
## 4  4      Diane      Economist
```

Merging (joining) datasets: example

```
df_merged <- merge(df_p, df_c, by="id")
df_merged
```

```
##   id first_name     profession money_spent currency year
## 1  1       Anna      Economist        1000      CHF 2017
## 2  1       Anna      Economist        1500      EUR 2018
## 3  2      Betty Data Scientist        2000      CHF 2017
## 4  2      Betty Data Scientist        3000      CHF 2018
## 5  3     Claire Data Scientist        6000      USD 2017
## 6  3     Claire Data Scientist        5500      USD 2018
```

Move to Nuvolos

Merging (joining) datasets: R

Overview by Wickham and Grolemund (2017):

dplyr (tidyverse)	base::merge
`inner_join(x, y)`	`merge(x, y)`
`left_join(x, y)`	`merge(x, y, all.x = TRUE)`
`right_join(x, y)`	`merge(x, y, all.y = TRUE)`,
`full_join(x, y)`	`merge(x, y, all.x = TRUE, all.y = TRUE)`

Data summaries

- First step of analysis.
- Get overview over dataset.
- Show key aspects of data.
    - Inform your own statistical analysis.
    - Inform audience (helps understand advanced analytics parts)

Data summaries: first steps

Quick overview: `summary()`

Cross-tabulation: `table()`

Data summaries and preparatory steps

Select subset of variables (e.g., for comparisons).

Filter the dataset (some observations not needed in **this** analysis).

Mutate the dataset: additional values needed

Select, filter, mutate in R (tidyverse)

```
select()
filter()
mutate()
```

Descriptive/aggregate statistics

Overview of key characteristics of main variables used in analysis.

Key characteristics:

- mean
- standard deviation
- No. of observations
- etc.

Aggregate statistics in R

1. Functions to compute statistics (e.g., `mean()`).
2. Functions to **apply** the statistics function to one or several columns in a tidy dataset.

   Including all values in a column.

   By group (observation categories, e.g. by location, year, etc.)

Aggregate statistics in R

`summarise()` (in `tidyverse`)

`group_by()` (in `tidyverse`)

`sapply()`, `apply()`, `lapply()`, etc. (in `base`)

Move to Nuvolos

Some vocabulary and notation

- **Dependent variable**: y_i.
- **Explanatory variable**: x_i.
- “All the rest”: u_{i} (the **‘residuals’** or the ‘error term’).

$y_{i}= \alpha + \beta x_{i} + u_{i}$.

Causality?

OLS Example: data

```r
# load the data
data(swiss)
# look at the description
?swiss
```

Research question

Do more years of schooling improve educational outcomes?

Approximate educational success with the variable `Education` and educational outcomes with the variable `Examination`.

Make use of the simple linear model to investigate whether more schooling improves educational outcomes (on average)?

Model specification

$Examination_{i}= \alpha + \beta Education_{i}$,

- Intuitive hypothesis: β is positive, indicating that a higher share of draftees with more years of schooling results in a higher share of draftees who reach the highest examination mark.
- **Problems**?

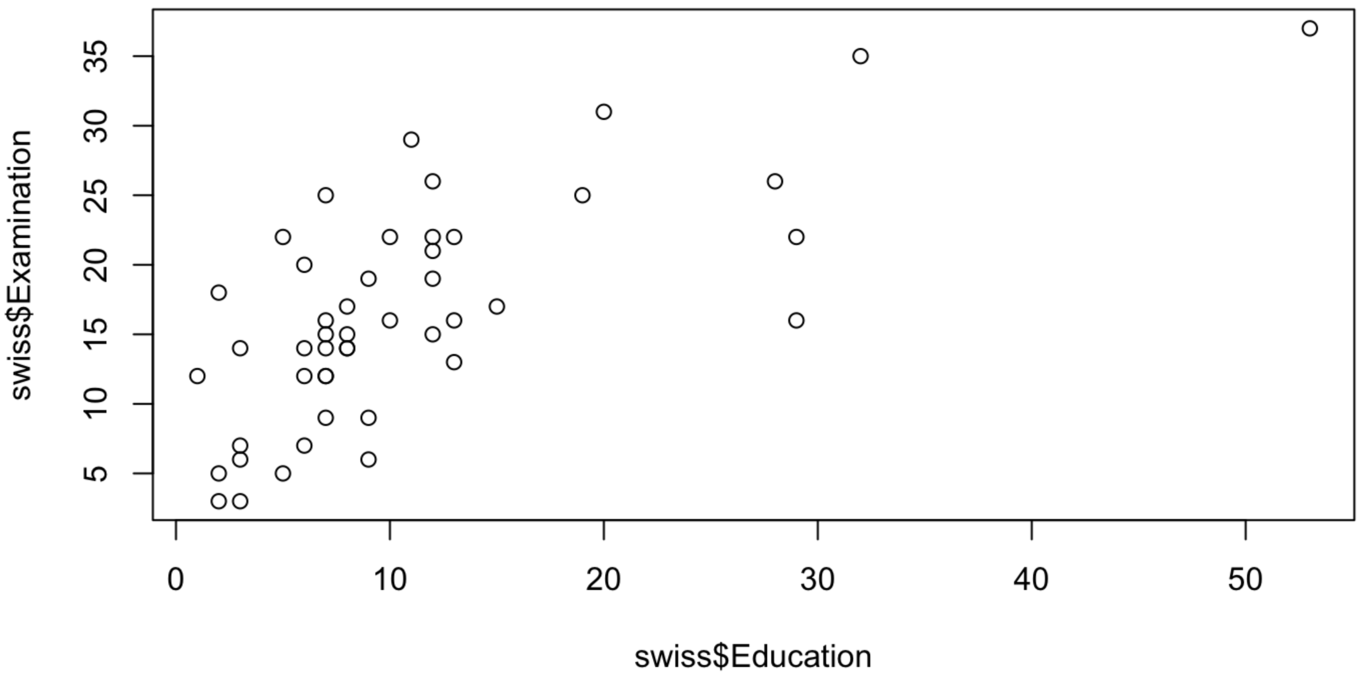
Model specification

To formally acknowledge that other factors might also play a role, we extend our model with the term u_{i}. For the moment, we thus subsume all other potentially relevant factors in that term:

$Examination_{i}= \alpha + \beta Education_{i} + u_{i}$.

Raw data

```
plot(swiss$Education, swiss$Examination)
```



Derivation and implementation of OLS estimator

From the model equation we easily see that these ‘differences’ between the predicted and the actual values of y are the remaining unexplained component u:

$y_{i}-\hat{\alpha}-\hat{\beta} x_i=u_i$.

Hence, we want to minimize the **sum of squared residuals (SSR)**: $\sum{u_i^2}=\sum{(y_{i}-\hat{\alpha}-\hat{\beta} x_i)^2}$. Using calculus, we define the two first order conditions:

$$\frac{\partial SSR}{\partial \hat{\alpha}}=\sum{-2(y_{i}-\hat{\alpha}-\hat{\beta} x_i)}=0$$

$$\frac{\partial SSR}{\partial \hat{\beta}}=\sum{-2x_i(y_{i}-\hat{\alpha}-\hat{\beta} x_i)}=0$$

Derivation and implementation of OLS estimator

The first condition is relatively easily solved by getting rid of the -2 and considering that $\sum{y_i}=N\bar{y}$: $\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$.

Derivation and implementation of OLS estimator

By plugging the solution for $\hat{\alpha}$ into the first order condition regarding $\hat{\beta}$ and again considering that $\sum{y_i}=N\bar{y}$, we get the solution for the slope coefficient estimator:

$\frac{\sum{x_{i}y_{i}}-N\bar{y}\bar{x}}{\sum{x_i^2}-N\bar{x}^2}$.

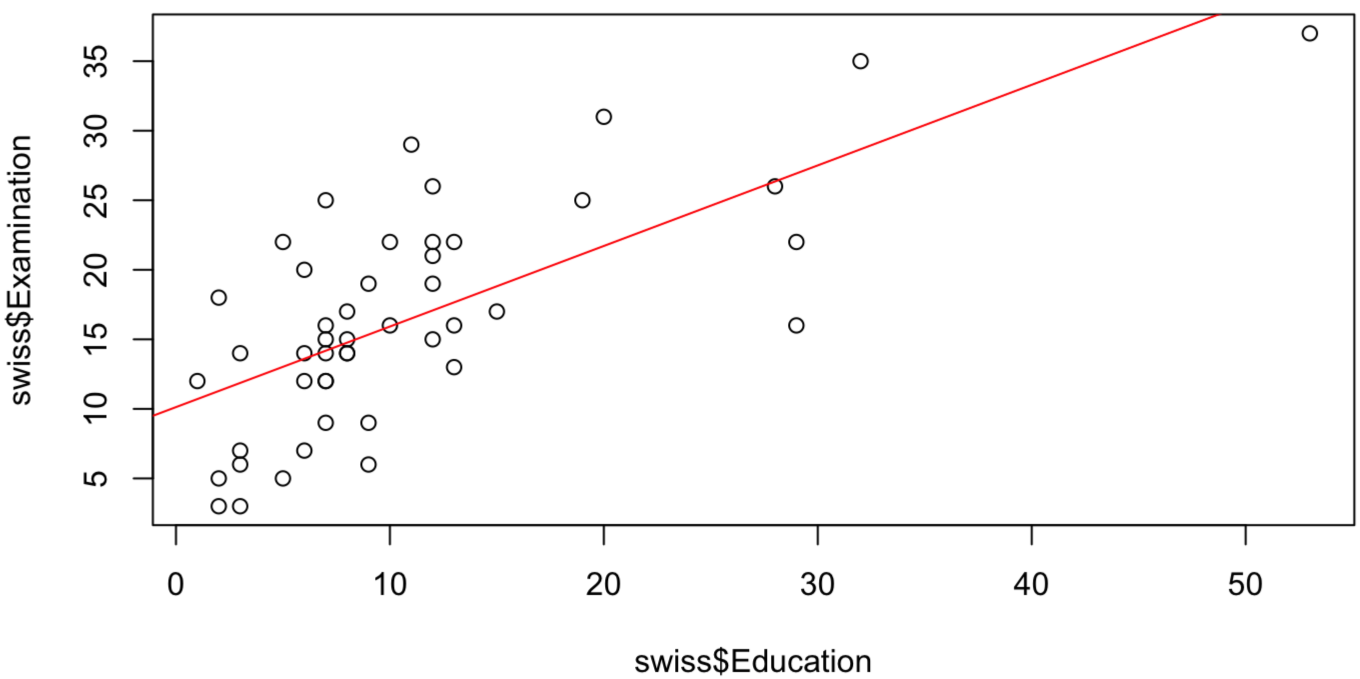
Implement OLS in R!

```
# implement the simple OLS estimator
# verify implementation with simulated data from above
# my_ols(y,x)
# should be very close to alpha=30 and beta=0.9
my_ols <-
  function(y,x) {
    N <- length(y)
    betahat <- (sum(y*x) - N*mean(x)*mean(y)) / (sum(x^2)-N*mean(x)^2)
    alphahat <- mean(y)-betahat*mean(x)

    return(list(alpha=alphahat,beta=betahat))
  }

# estimate effect of Education on Examination
estimates <- my_ols(swiss$Examination, swiss$Education)
estimates
```

```
## $alpha
## [1] 10.12748
##
## $beta
## [1] 0.5794737
```

Simple visualisation

```
plot(swiss$Education, swiss$Examination)
abline(estimates$alpha, estimates$beta, col="red")
```



Regression toolbox in R

```
estimates2 <- lm(Examination~Education, data=swiss)
estimates2
```

```
##
## Call:
## lm(formula = Examination ~ Education, data = swiss)
##
## Coefficients:
## (Intercept)    Education
##     10.1275       0.5795
```

With one additional line of code we can compute all the common statistics about the regression estimation:

```
summary(estimates2)
```

```
##
## Call:
## lm(formula = Examination ~ Education, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9322  -4.7633  -0.1838   3.8907  12.4983
##
```

References

Wickham, Hadley, and Garrett Grolemund. 2017. Sebastopol, CA: O'Reilly. http://r4ds.had.co.nz/.