Templatization of Analytics and Research Data Warehousing

Andriy Koval

2023-05-26

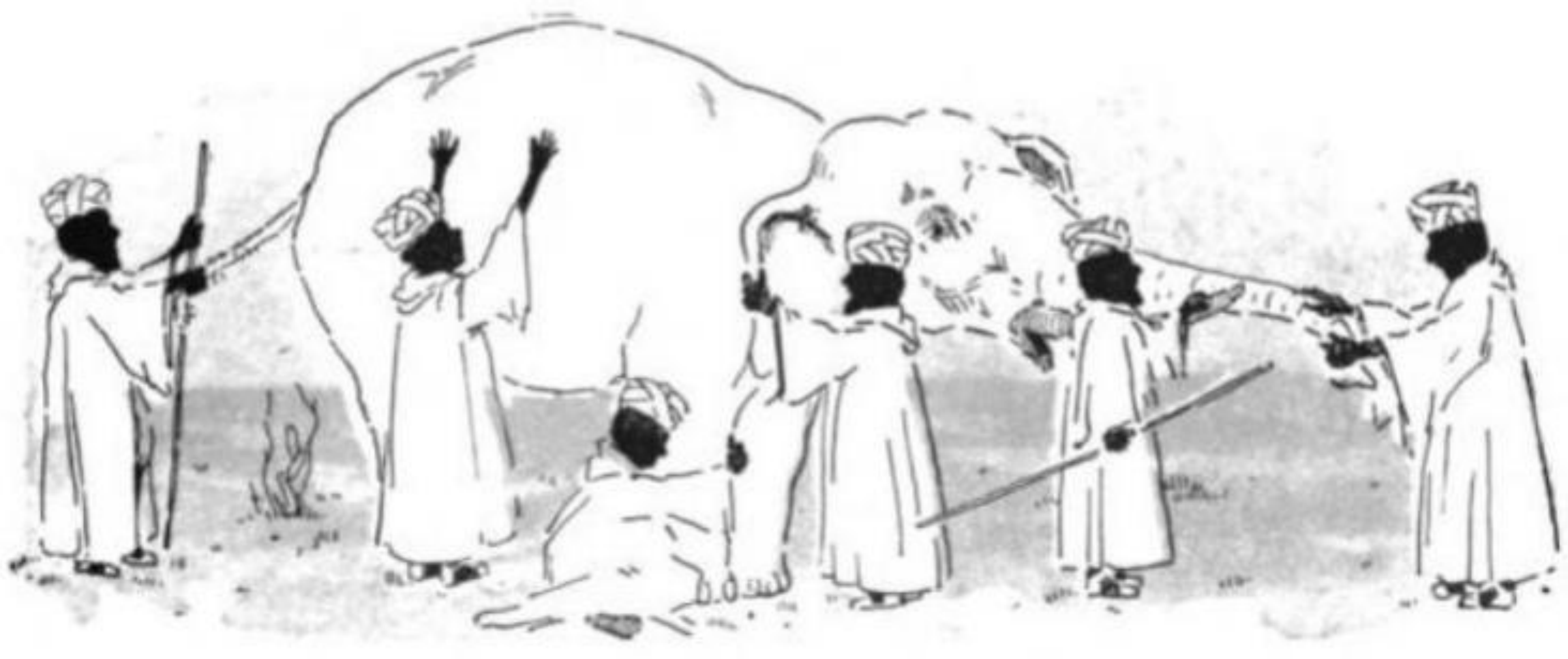
Plan for today

1. Motivation and Design principles of RDB
2. Threats to validity in the age of big data and cheap computing
3. Examples of progressively specific templates:
    - Quick Start Template – specific to GoA
    - R Analysis Skeleton - Generic
    - Generic Explorer – specific to RDB of SCSS

Motivation for RDB

Puzzle #1

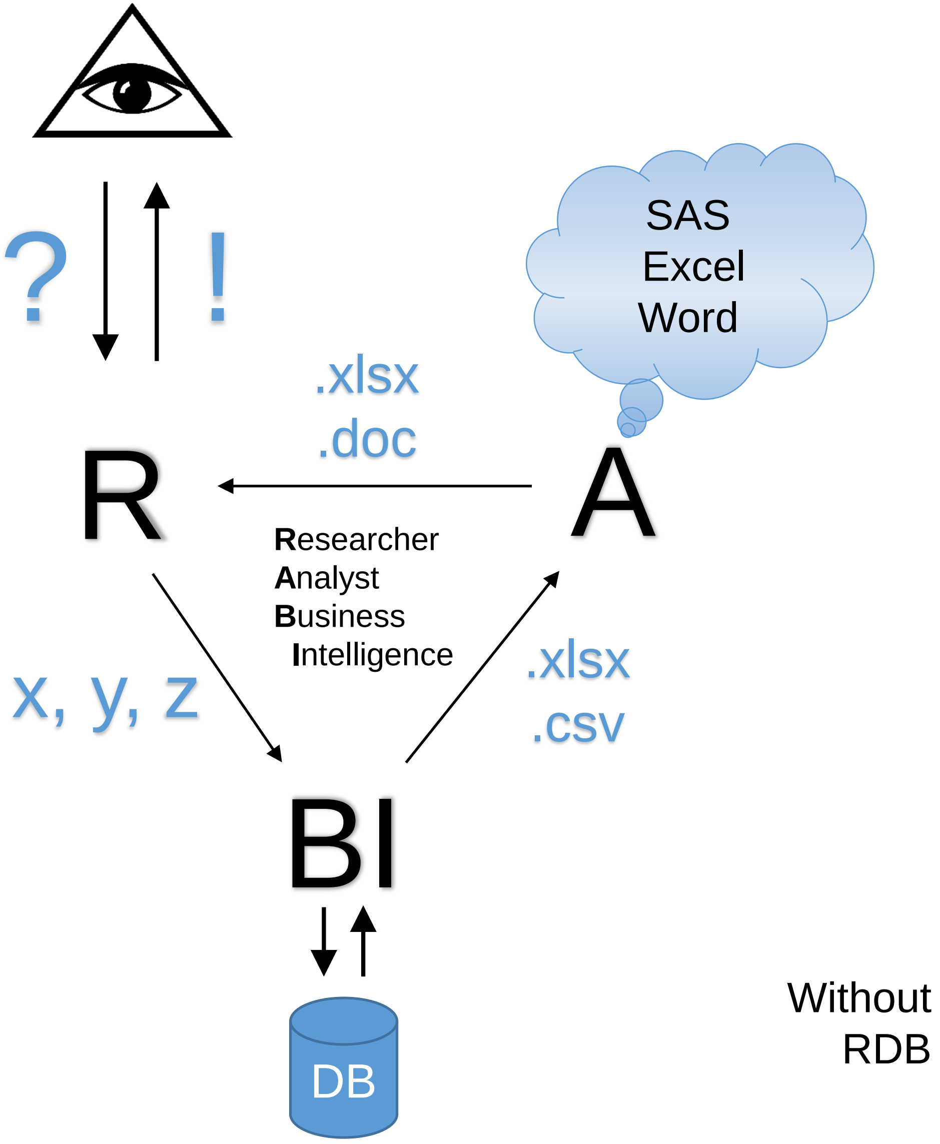
Why ~~do~~ can they disagree?



https://en.wikipedia.org/wiki/File:Blind_men_and_elephant3.jpg

Motivation for RDB

Traditional analytic cycle of an information request is laden with poor reproducibility and scalability



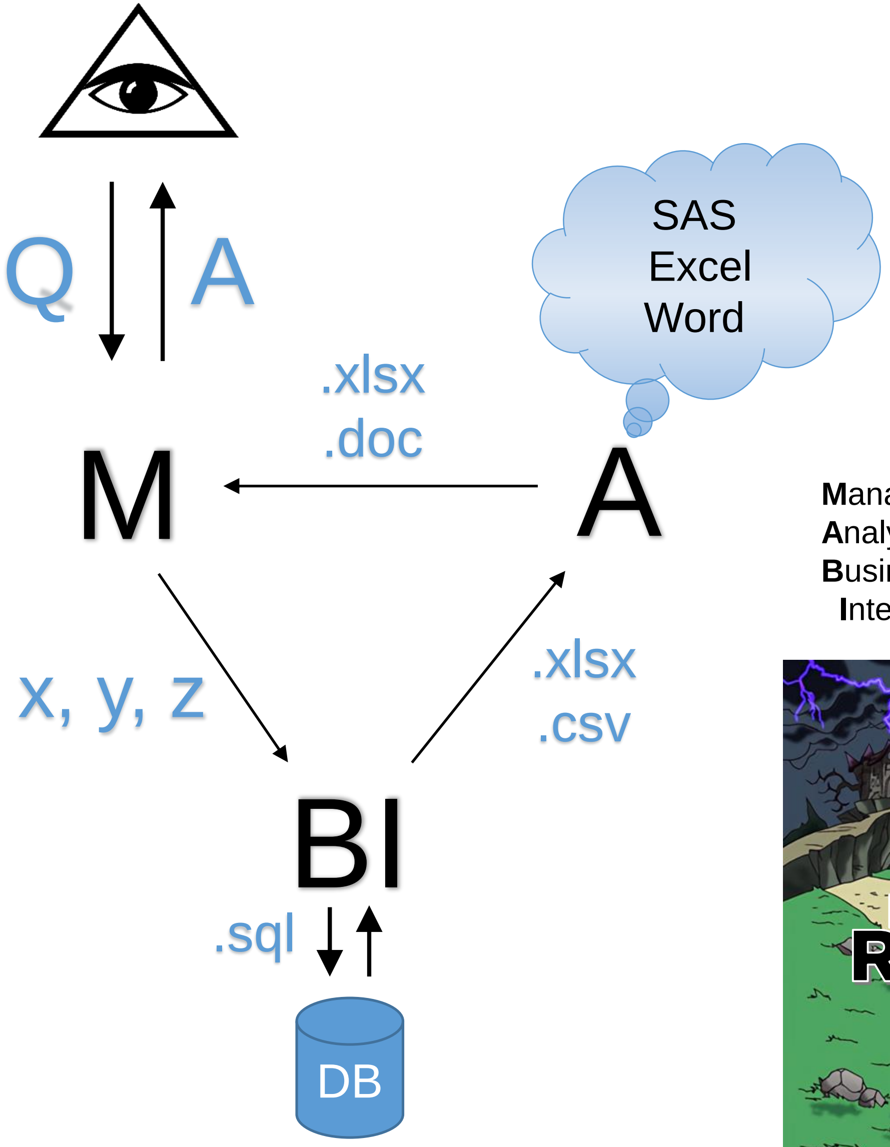
Problem:
no elephant to examine

Classification: Protected A

Motivation for RDB

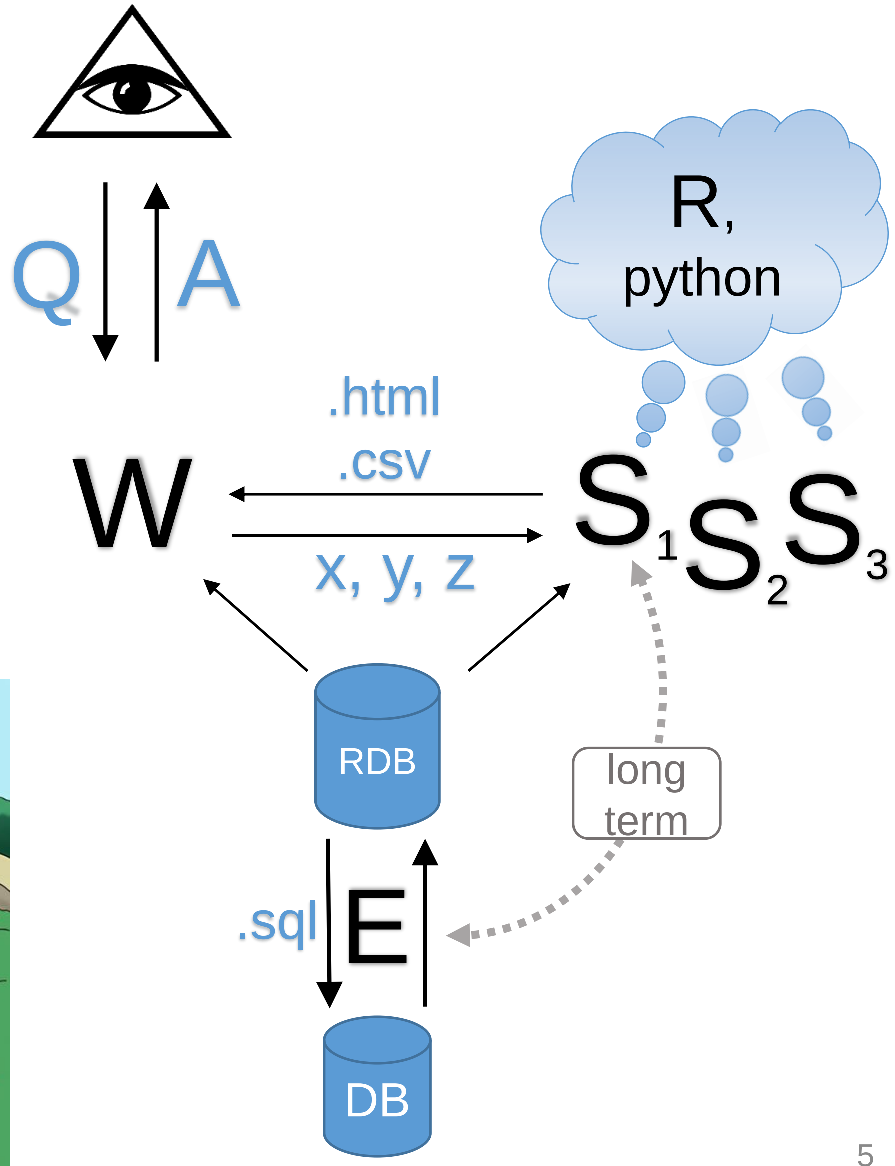
Traditional analytic cycle of an information request is laden with poor reproducibility and scalability

RDB-centered approach removes some workflow bottlenecks, but calls for new skills and roles.



Manager
Analyst
Business
Intelligence

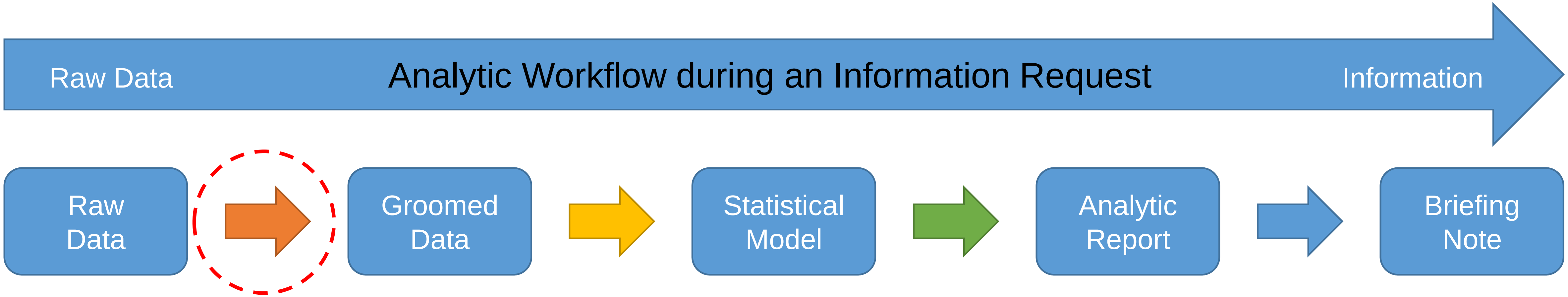
Writer
Scientist
Engineer



Motivation for RDB

Puzzle #2

Motivation for RDB



Raw Data

Analytic Workflow during an Information Request

Information

Raw Data

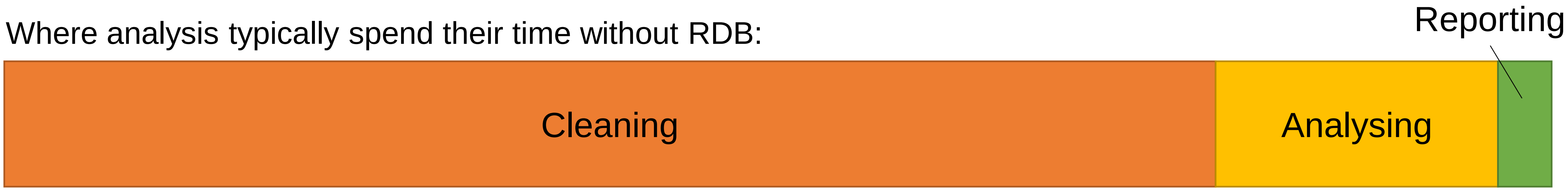
Groomed Data

Statistical Model

Analytic Report

Briefing Note

Where analysis typically spend their time without RDB:

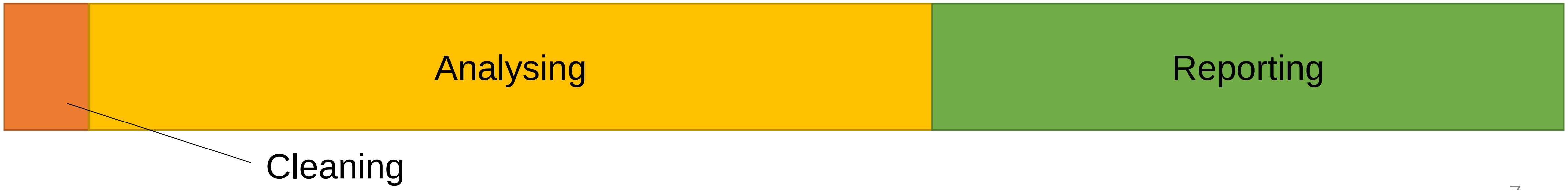


Reporting

Cleaning

Analysing

Goals of Research DB:

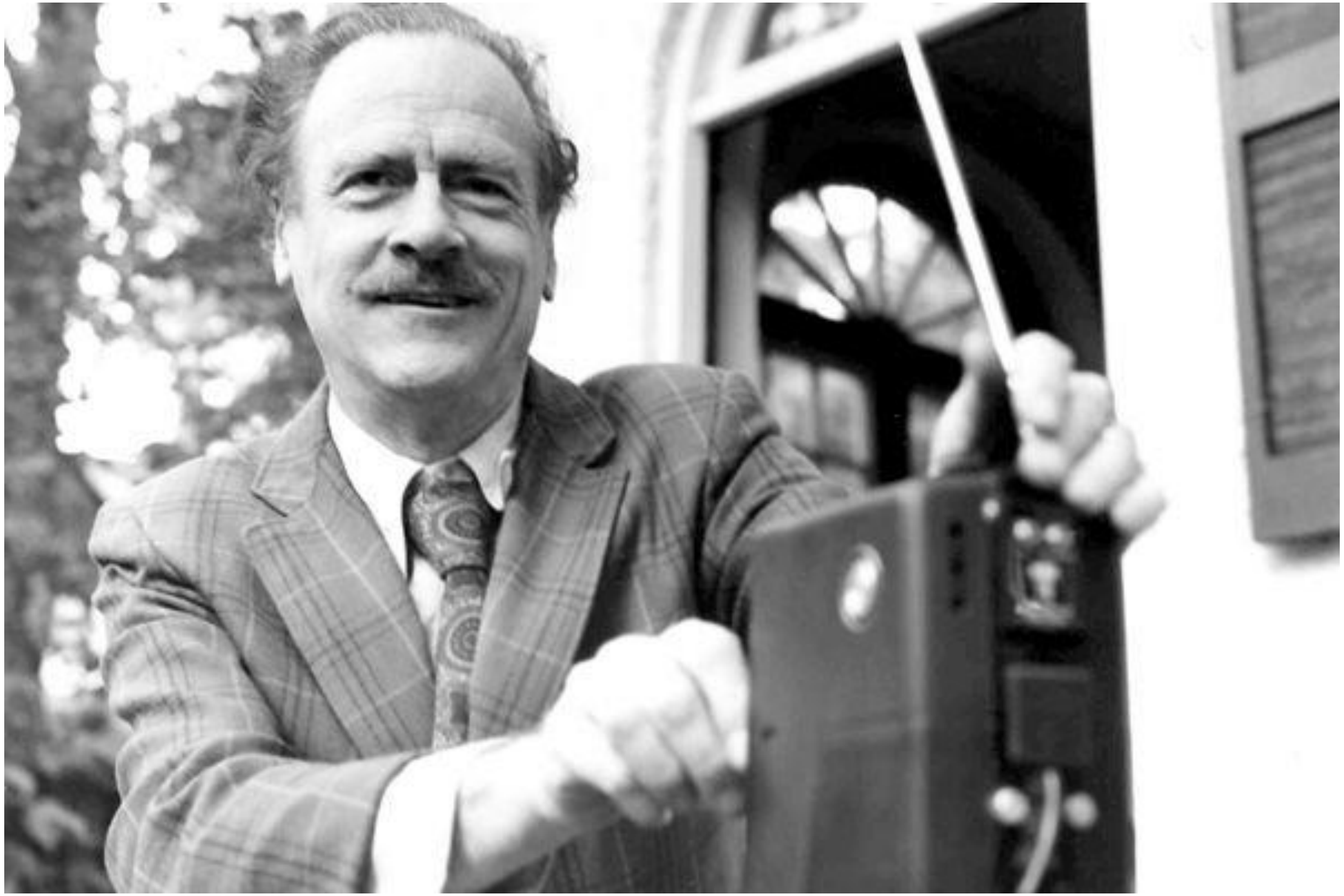


Analysing

Reporting

Cleaning

Puzzle #3

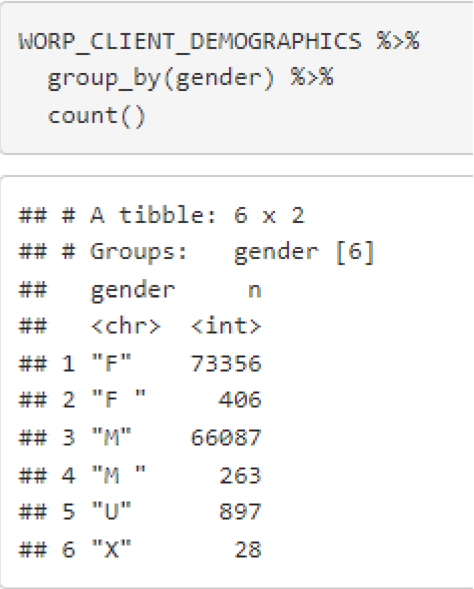


Macro-level:
The data is the analysis

IF YOU TREAT DATA PREPARATION AS MERELY AN INTERMEDIATE STEP
NSTRUCTO
imgflip.com

Micro-level: The script is the evidence

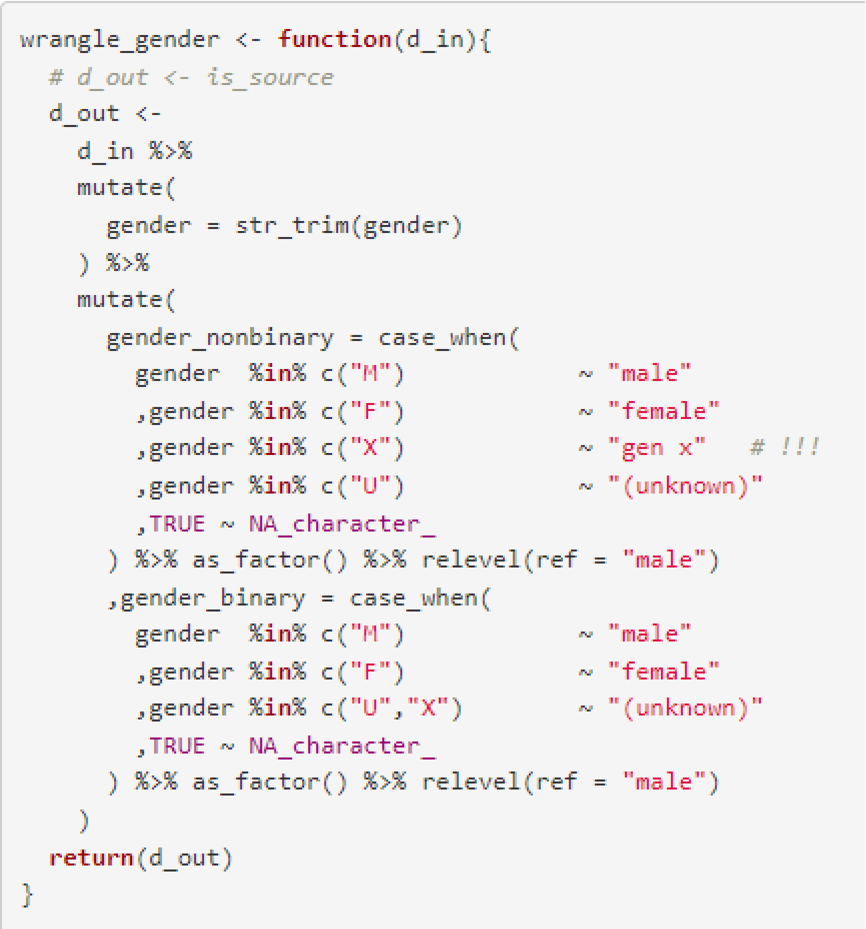
Exposition



```
WORP_CLIENT_DEMOGRAPHICS %>%
  group_by(gender) %>%
  count()
```

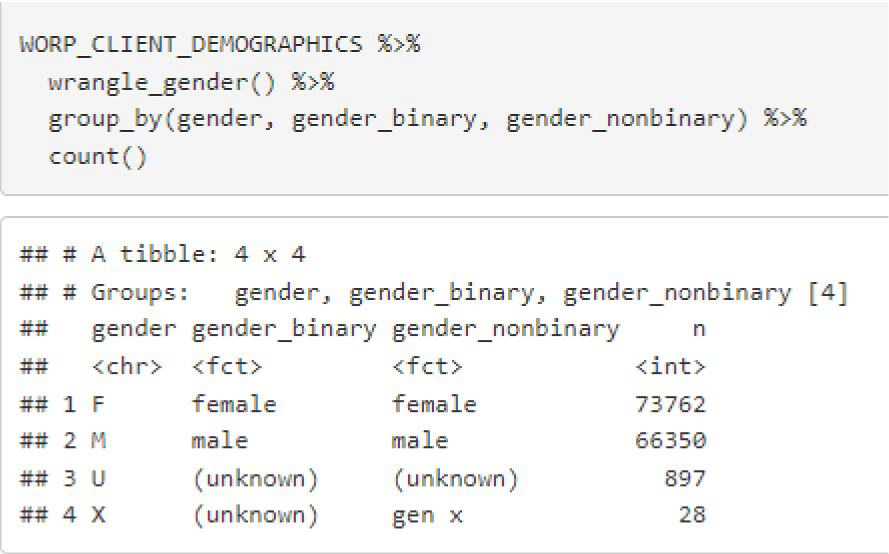
```
## # A tibble: 6 x 2
## # Groups:   gender [6]
##   gender     n
##   <chr>  <int>
## 1 "F"    73356
## 2 "F "     406
## 3 "M"    66087
## 4 "M "     263
## 5 "U"      897
## 6 "X"       28
```

Transformation



```
wrangle_gender <- function(d_in){
  # d_out <- is_source
  d_out <-
    d_in %>%
    mutate(
      gender = str_trim(gender)
    ) %>%
    mutate(
      gender_nonbinary = case_when(
        gender  %in% c("M")             ~ "male"
        ,gender %in% c("F")             ~ "female"
        ,gender %in% c("X")             ~ "gen x"   # !!!
        ,gender %in% c("U")             ~ "(unknown)"
        ,TRUE ~ NA_character_
      ) %>% as_factor() %>% relevel(ref = "male")
      ,gender_binary = case_when(
        gender  %in% c("M")             ~ "male"
        ,gender %in% c("F")             ~ "female"
        ,gender %in% c("U","X")         ~ "(unknown)"
        ,TRUE ~ NA_character_
      ) %>% as_factor() %>% relevel(ref = "male")
    )
  return(d_out)
}
```

Validation



```
WORP_CLIENT_DEMOGRAPHICS %>%
  wrangle_gender() %>%
  group_by(gender, gender_binary, gender_nonbinary) %>%
  count()
```

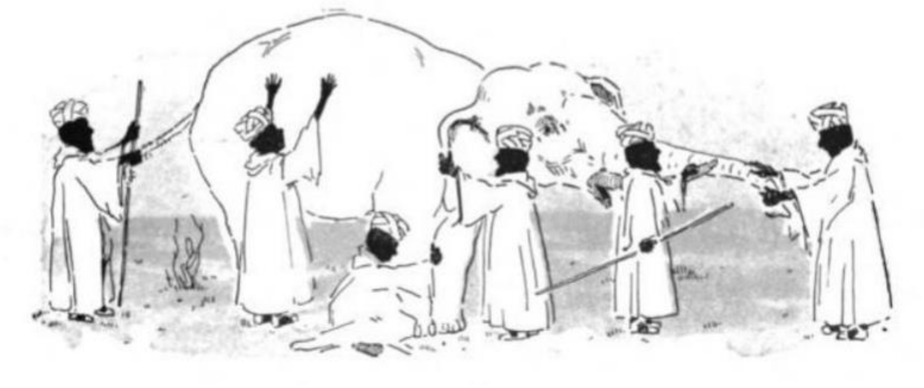
```
## # A tibble: 4 x 4
## # Groups:   gender, gender_binary, gender_nonbinary [4]
##   gender gender_binary gender_nonbinary     n
##   <chr>  <fct>         <fct>            <int>
## 1 F      female        female           73762
## 2 M      male          male             66350
## 3 U      (unknown)     (unknown)          897
## 4 X      (unknown)     gen x               28
```

QED

The construct “gender” now has an auditable operationalization

Take away points

- RDB is the elephant to disagree about
- RDB is the washing machine to liberate you
- RDB is the subject of the study



Design Principles of RDB

- Literate Programming
- Reproducibility
- Scalability
- Collaboration
- Transparency
- Version Control
- Interoperability
- Continuous Improvement

There are other considerations without which a thorough discussion of RDB would be incomplete, but we leave it for dedicated discussions.

- Data Quality
- Data Security
- Data Ethics
- Bias and Fairness
- Performance & Optimization
- Continuous Improvement

Design Principles of RDB

- **Literate Programming**
- Reproducibility
- Scalability
- Collaboration
- Transparency
- Version Control
- Interoperability
- Continuous Improvement

Analytical report is a script that can be executed as one program, generates visible output, and contains instructions for reproduction.

Literate Programming

- Code + Output + Annotation
- Readable by machines, understood by humans
- Donald Knuth (paper)

Data Science for Evidence-based decisions

- If we want to use the results of data analysis as evidence to support our views and decisions, we must demonstrate its chain of custody and address threats to validity
- Analysis Templatization as a response to new threats to validity emerging from big data and cheap computing
- Please download quick-start-template to start practical part of the session

Authoring formats

- .md
- .Rmd or .qmd
- .R

Plan for today

1. Motivation and Design principles of RDB
2. Threats to validity in the age of big data and cheap computing
3. Examples of progressively specific templates:
   - Quick Start Template – specific to GoA
   - R Analysis Skeleton - Generic
   - Generic Explorer – specific to RDB of SCSS