Show me the data you didn't consider!

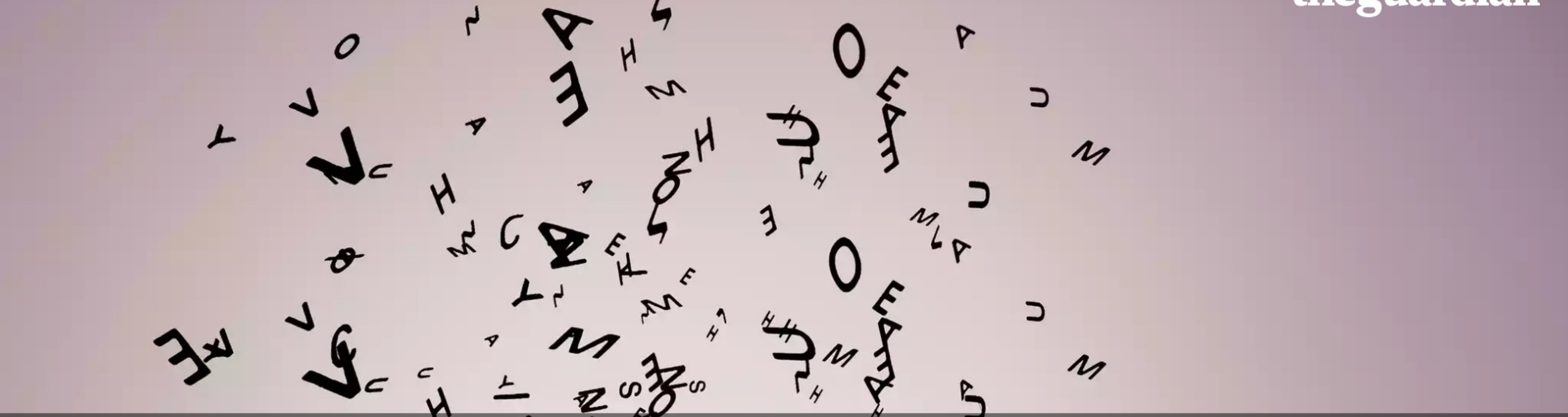
Reducing unknown unknowns in data science

Claus Thorn Ekstrøm
UCPH Biostatistics
ekstrom@sund.ku.dk

Ann Arbor R Users meeting, March 9th

Slides @ biostatistics.dk/talks/

theguardian



The long read

How statistics lost their power – and why we should fear what comes next

The ability of statistics to accurately represent the world is declining. In its wake, a new age of big data controlled by private companies is taking over – and putting democracy in peril

p-value

I CAN'T KEEP CALM

because of

SOCIOLOGY

mirror.co.uk

DAILY
Mirror
NEWSPAPER OF THE YEAR

HISTORIC 5AM EDITION

Referendum shock as Leave head for victory

WE'RE OUT

» Britain votes to quit the EU » Pound goes into freefall

THE
Sun

6AM LATEST

OUT 52%

IN 48%

BRITAIN'S BEST-SELLING PAPER

Britain waking to an EU exit

Leave claims win in huge poll

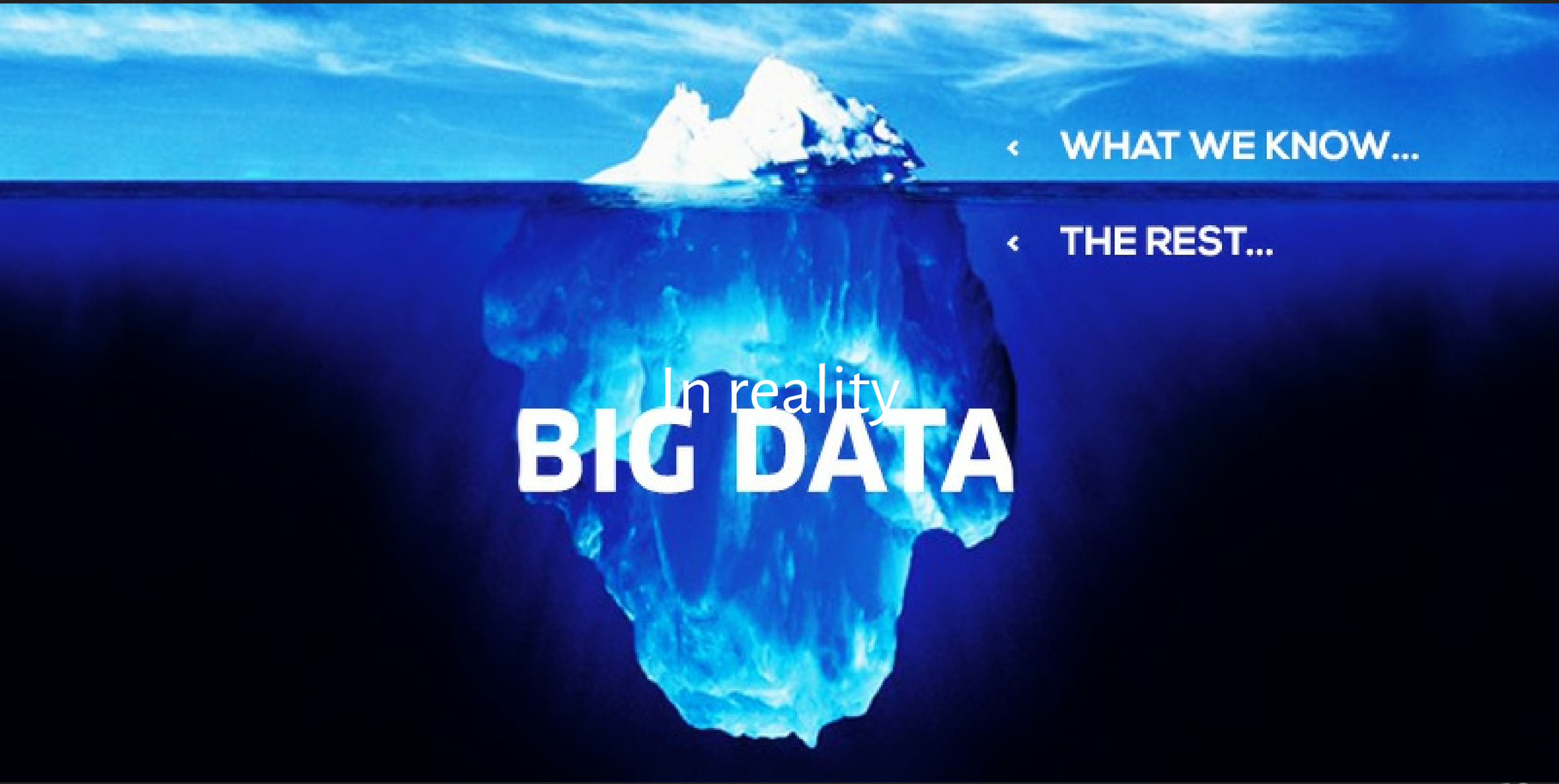
SEE EU LATER!


Trump on brink of
White House as
ocks Hillary
he world
edge poll
Standard
BILLIONAIRE PULLS OFF HUGE UPSET TO BECOME PRESIDENT
HE PLEDGES TO 'BIND WOUNDS OF DIVISION' AS CLINTON CONCEDES
PUTIN AND MAY SEND MESSAGES OF CONGRATULATION
Croydon tram crash horror: 'several dead'
TRUMP TRIUMPH SHOCKS WORLD
EPORTS AND ANALYSIS
COMMENT

Quiz

The life of a data scientist

> Data scientists, according to interviews and expert estimates, spend from **50 percent to 80 percent** of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.
>
> -- "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insight" - The New York Times, 2014


WHAT WE KNOW...
THE REST...
In reality
BIG DATA

Beyond the hype

> Many of the existing problems with small data are also applicable to big data. …
>
> The problems do not disappear because the data sizes becomes larger. They become **worse**.

State of the art?

> **Statistical analysis**
>
> All of the data were analyzed with data processing software and figures with Microsoft excel 2007.

-- Tayefe *et al*, Advances in Bioresearch, 2014

The RESCueH project

Timeline follow back

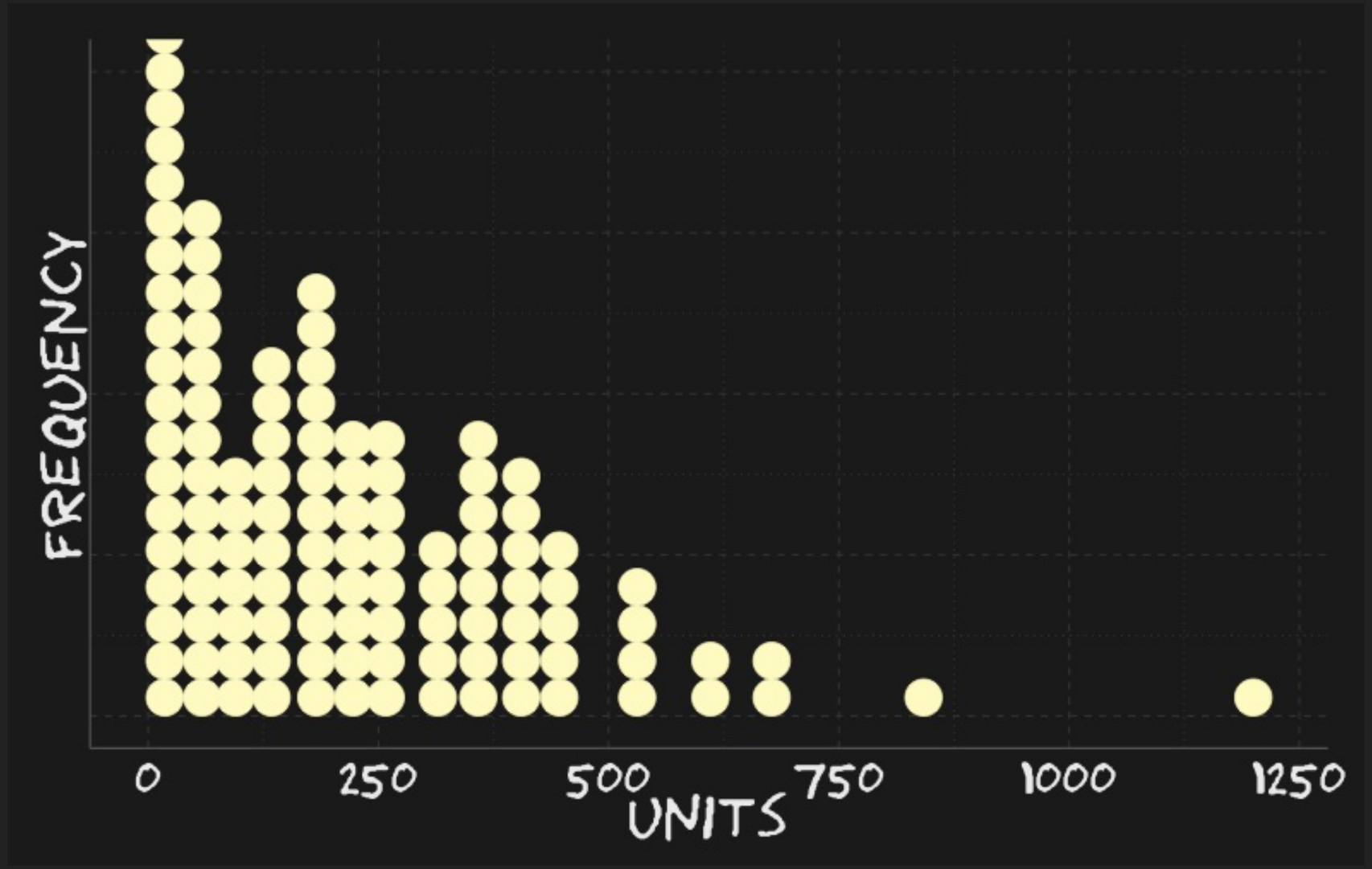
	day1	day2	day3
1	18	NA	NA
2	14	NA	99
3	20	17	40
4	23	14	17

Timeline follow back

```
  day1 day2 day3
1   18   NA   NA
2   14   NA   99
3   20   17   40
4   23   14   17
```

```
  day1 day2 day3
1   18   NA   NA
2   14   NA   99
3   20   17   40
4   23   14   17
5   10   24    2
6   19   88    8
```

Monthly Alcohol units



Reproducible research

What **didn't** we check?

Reproducible research

What **didn't** we check?

- Many studies cannot be replicated: time, money, unique
- New technologies increase data sizes
- Merge existing databases into megadatabases
- Genetic data for future analyses

Reproducible research

What **didn't** we check?

- Many studies cannot be replicated: time, money, unique
- New technologies increase data sizes
- Merge existing databases into megadatabases
- Genetic data for future analyses

You are your worst collaborator.

dataMaid

```
devtools::install_github("ekstroem/dataMaid")
library(dataMaid)
data(toyData)
clean(toyData)
```

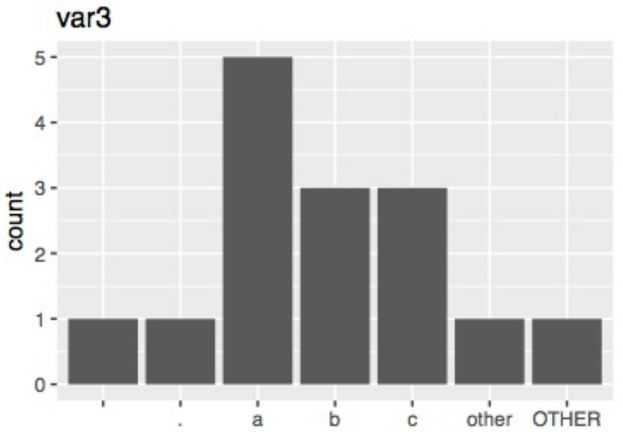
Documentation to be **read** and **evaluated** by a human.

See github.com/ekstroem/dataMaid for more info.

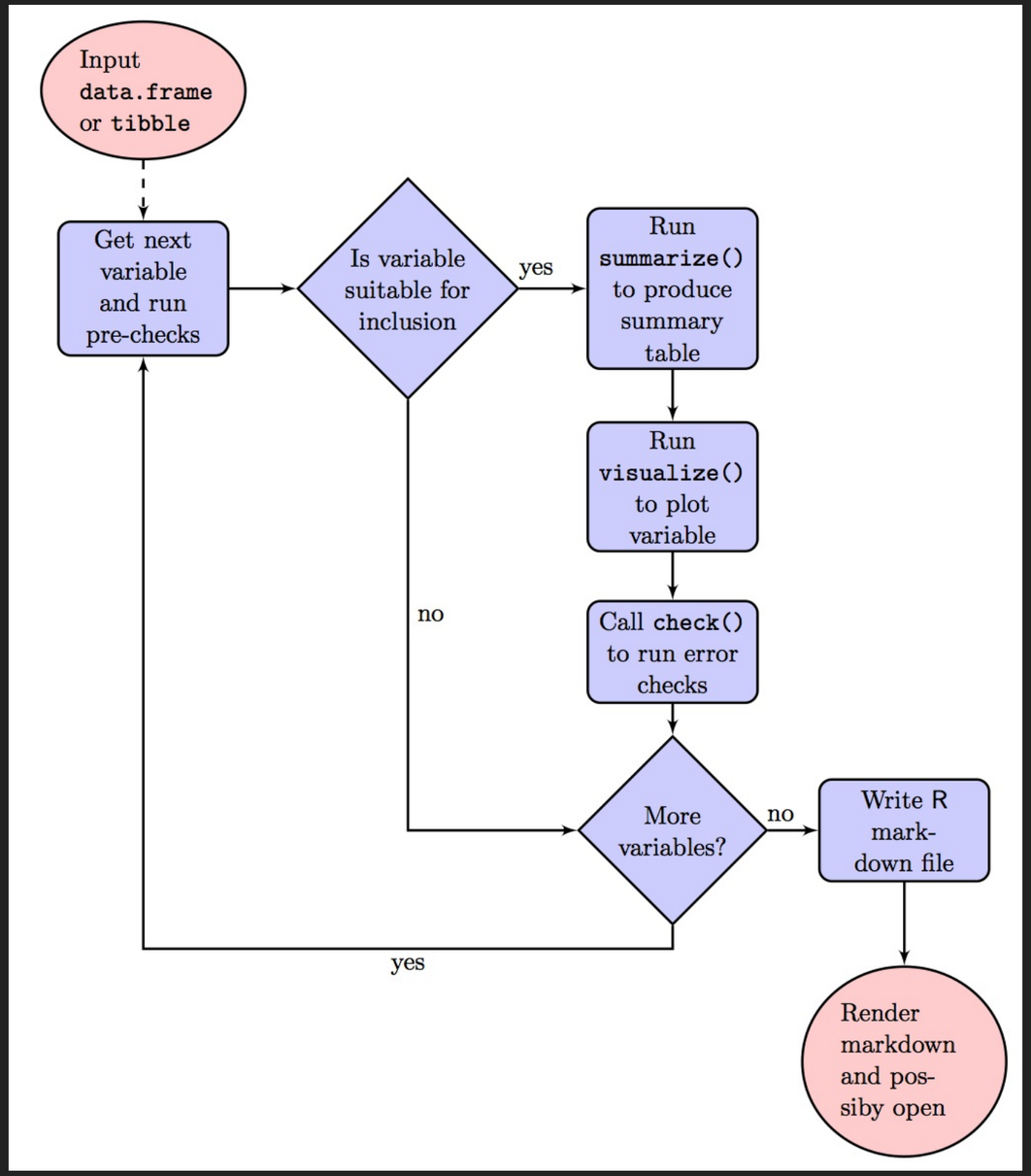
	character	factor	labelled	numeric	integer	logical	Date
Identify miscoded missing values	×	×	×	×	×		
Identify prefixed and suffixed whitespace	×	×	×				
Identify levels with < 6 obs.	×	×	×				
Identify case issues	×	×	×				
Identify misclassified numeric or integer variables	×	×	×				
Identify outliers				×	×		×

var3

Feature	Result
Variable type	factor
Number of missing obs.	0 (0 %)
Number of unique values	7
Mode	“a”



- The following suspected missing value codes enter as regular values: " ", ".".
- The following values appear with prefixed or suffixed white space: " ".
- Note that the following levels have at most five observations: " ", ".", "a", "b", "c", "other", "OTHER".
- Note that there might be case problems with the following levels: "other", "OTHER".


Input
data.frame
or tibble
Get next
variable
and run
pre-checks
Is variable
suitable for
inclusion
yes
Run
summarize()
to produce
summary
table
Run
visualize()
to plot
variable
Call check()
to run error
checks
no
More
variables?
no
Write R
mark-
down file
yes
Render
markdown
and pos-
siby open

Using dataMaid interactively

```
check(toyData$var2)  # Individual check
check(toyData$var2, numericChecks = "identifyMissing")
visualize(toyData$var2)
summarize(toyData$var2)
summarize(toyData$var2,
          numericSummaries = c("centralValue", "minMax"))
```

Extending dataMaid

```
isID <- function(v, nMax = NULL, ...) {
  out <- list(problem = FALSE, message = "")
  if (class(v) %in% setdiff(allClasses(),
                            c("logical", "Date"))) {
    v <- as.character(v)
    lengths <- c(nchar(v))
    if (all(lengths > 10) &
        length(unique(lengths)) == 1) {
      out$problem <- TRUE
      out$message <- "Warning: Seems to contain IDs."
    }  }
  out }
```

Adding the function

```
data("exampleData")
exampleData$names <- sapply(1:300,
  function(i) { paste0(sample(LETTERS, size=10),
                       collapse="") })
clean(exampleData,
  preChecks = c("isID"))
```

Communicate!