Data editing with editrules and deducorrect

Mark van der Loo and Edwin de Jonge

amst-R-dam user meeting 02.04.2012

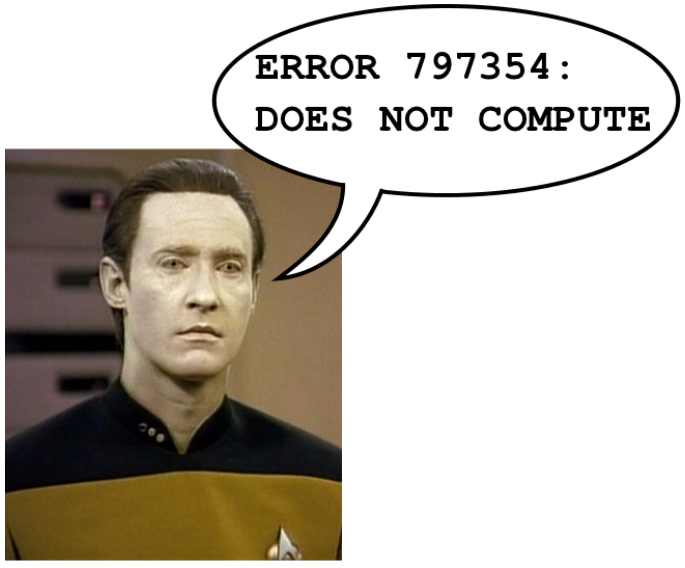
www.markvanderloo.eu

Newest versions available end of april 2012

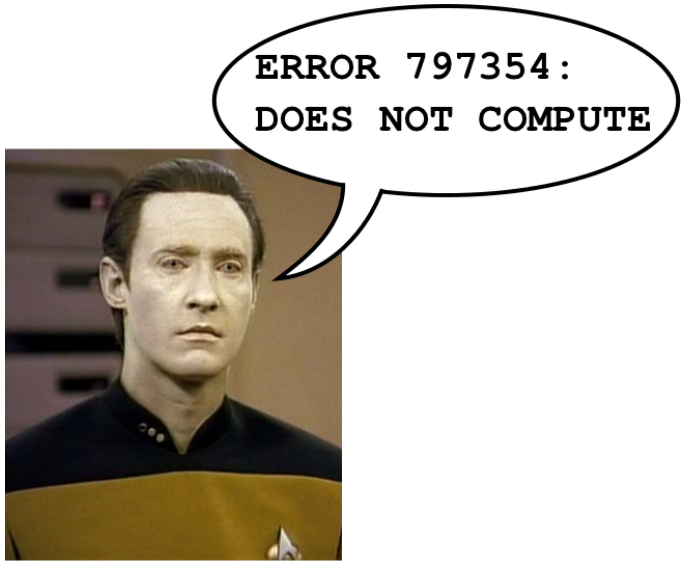
Typesetting and graphics: LaTeX, BEAMER, Ti*k*Z, R/Sweave

cost	profit	turnover
12	342	8

cost	profit	turnover
12	342	8



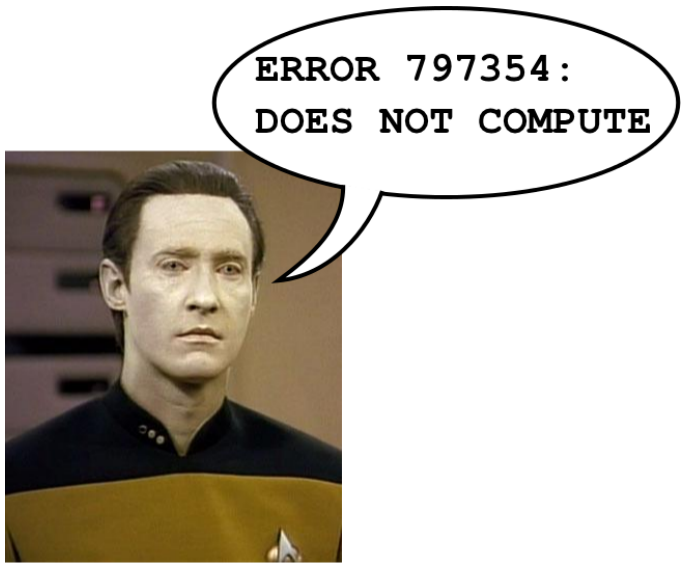
cost	profit	turnover
12	342	8



Expected:

$$\text{cost} + \text{profit} = \text{turnover}$$

cost	profit	turnover
12	342	8



Expected:

$$\text{cost} + \text{profit} = \text{turnover}$$

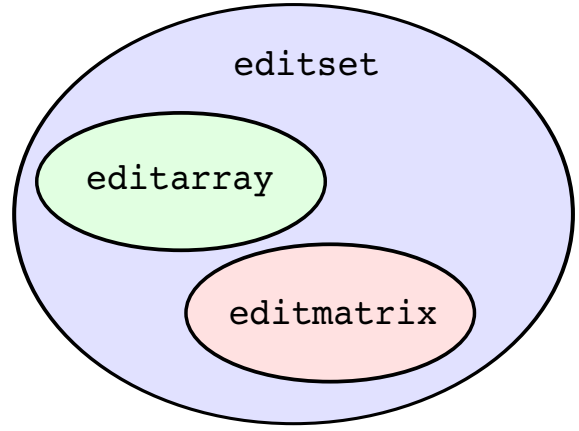
- Typical: $>$ 100 variables, several 100 rules, *e.g.* positivity, ratio's, account balances, conditional restrictions.

The editrules package.

- Read and write restrictions (edits) from/to text
- Edit manipulation
  - Check for consistencies
  - Graphical analyses.
  - Derive rules, substitute values.
- Data checks
  - Detect, summarize and visualize errors.
  - Localize erroneous fields, visualize and summarize.

Three types of rules

Pure numeric (linear)	`editmatrix`
Pure categorical	`editarray`
Mixed/conditional (all)	`editset`

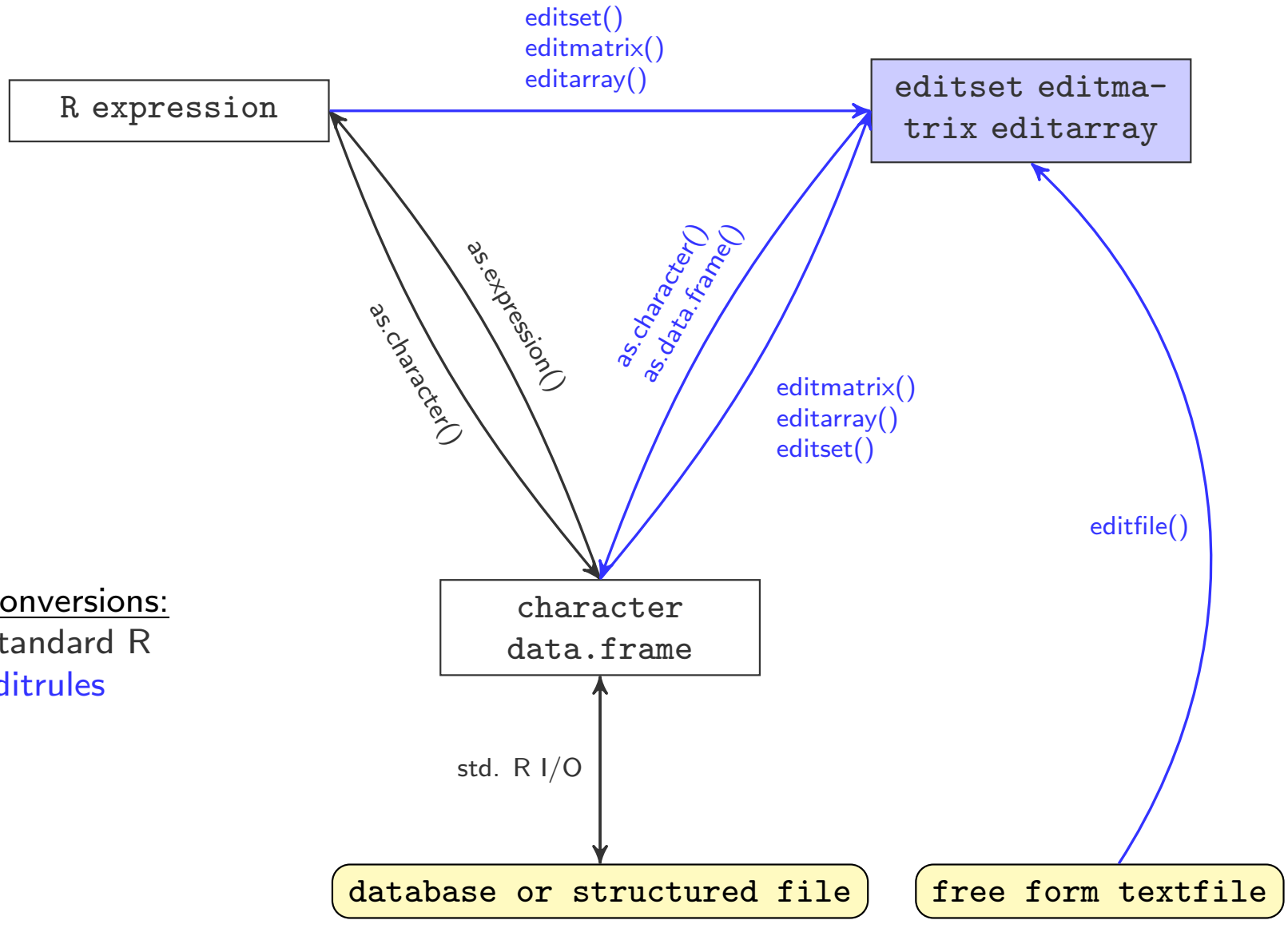


Example: read from file

```
# numerical rules
x + y == z
x >= 0
y >= 0

# categorical rules
A %in% letters[1:3]
B %in% letters[6:8]
if ( A %in% c('b','c') ) B %in% c('f','h')

# mixed rules
if ( x > 0 ) y > 0
if ( A %in% "a" ) 2*x > y
```

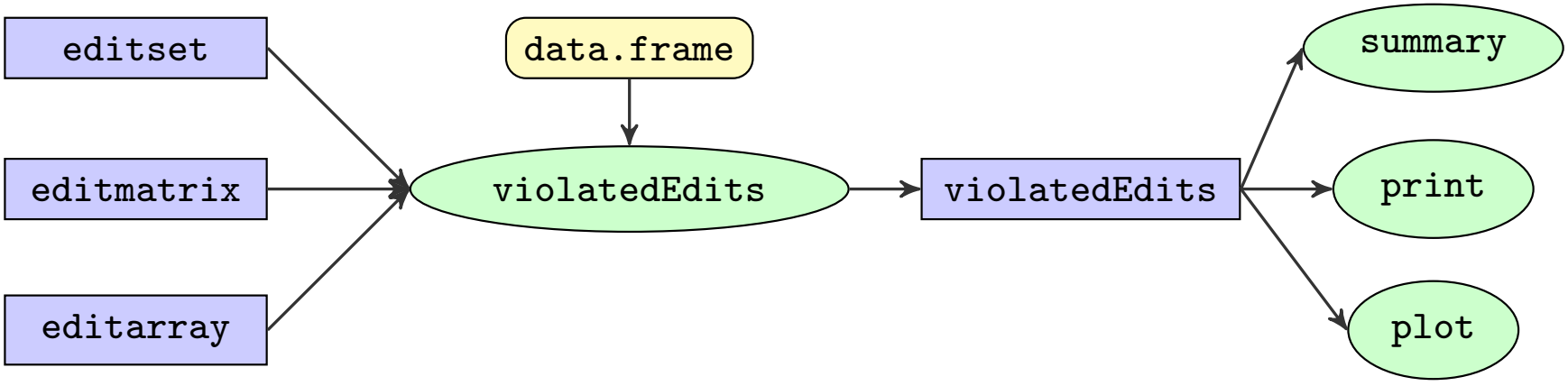
```
> editfile("edits.txt")

Data model:
dat1 : A %in% c('a', 'b', 'c')
dat2 : B %in% c('f', 'g', 'h')

Edit set:
num1 : x + y == z
num2 : 0 <= x
num3 : 0 <= y
cat4 : if( A %in% c('b', 'c') ) B != 'g'
mix5 : if( 0 < x ) y > 0
mix6 : if( A == 'a' ) 2*x > y
```

Storage and conversion

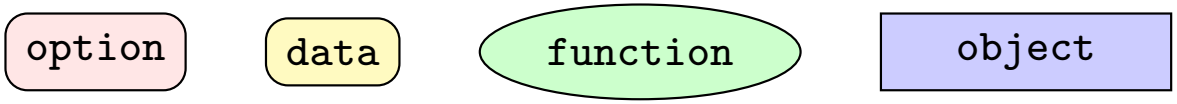


Conversions:
Standard R
editrules

Check data: violatedEdits



Legend

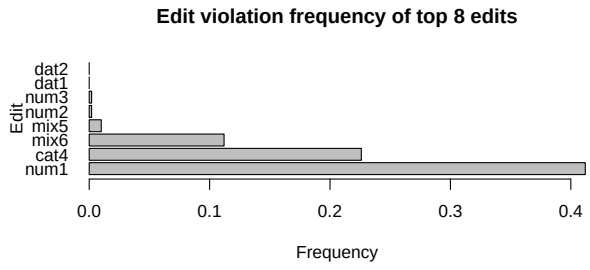


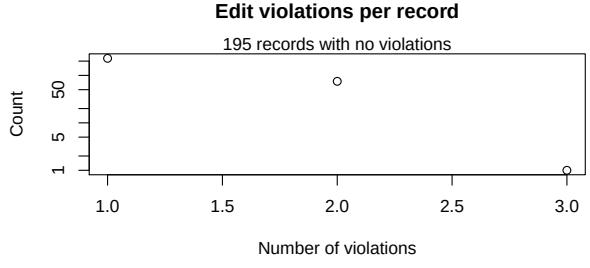
Example: violatedEdits

```
# mydata.csv
"x";"y";"A";"B";"z"
5;11;"c";"g";16
45;0;"a";"f";46
12;7;"c";"f";19
```

```
> dat <- read.csv2("mydata.csv",comment.char="#")
> E <- editfile("edits.txt")
> ve <- violatedEdits(E,dat)
> summary(ve)

Edit violations, 500 observations, 0 completely missing (0%):

 editname freq   rel
     num1  206 41.2%
     cat4  113 22.6%
     mix6   56 11.2%
     mix5    5    1%
     num2    1  0.2%
     num3    1  0.2%

Edit violations per record:

 errors freq   rel
      0  195   39%
      1  229 45.8%
      2   75   15%
      3    1  0.2%
```

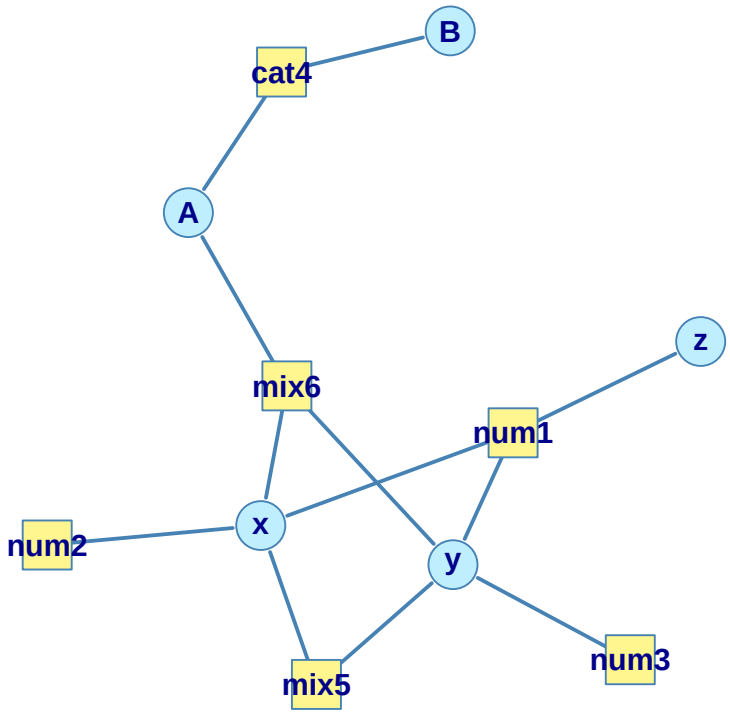
```
> plot(ve)
```





Visualisation of edit rules

```
> plot(E)
```



Error localization

Problem

Find the least number of (weighted) fields that can be adapted such that no rule is violated anymore.

Difficult because...

- Take non-violated restrictions into account
- Take implied rules into account.

Error localization

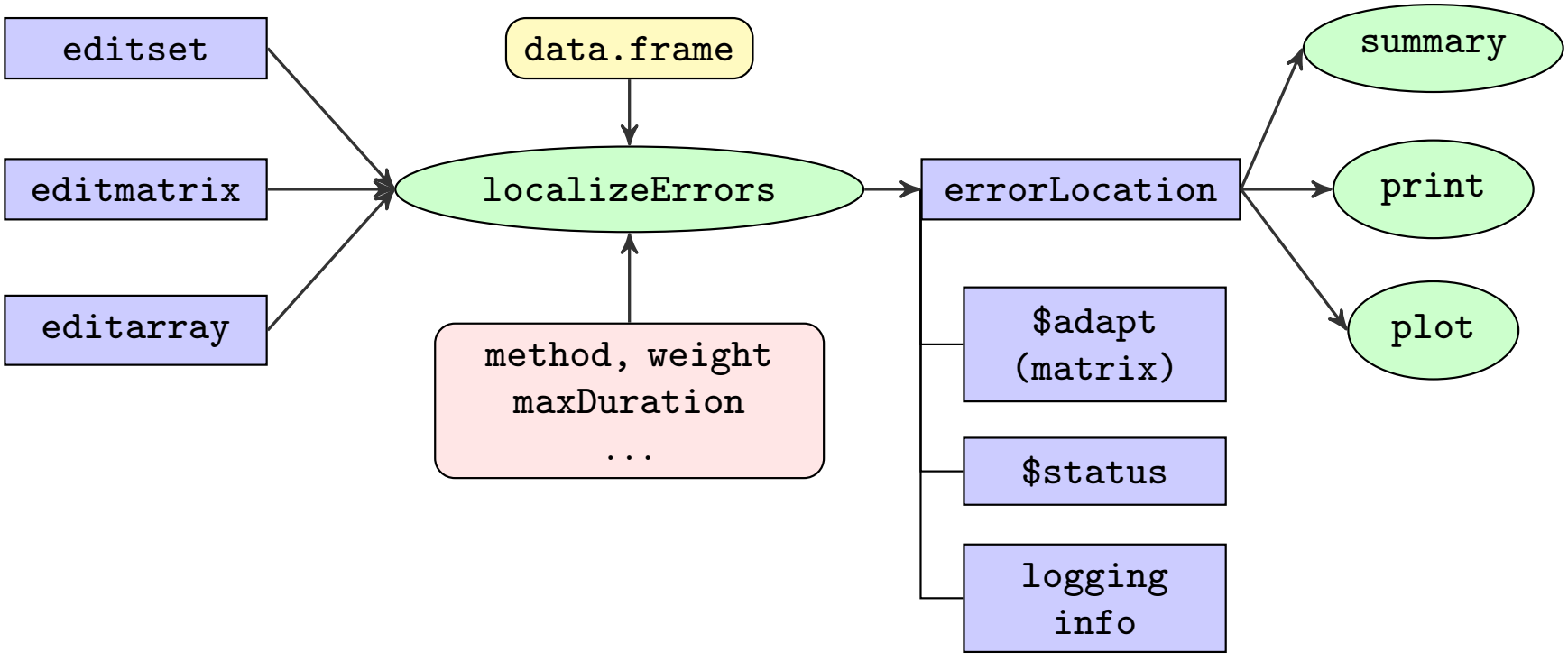
Problem

Find the least number of (weighted) fields that can be adapted such that no rule is violated anymore.

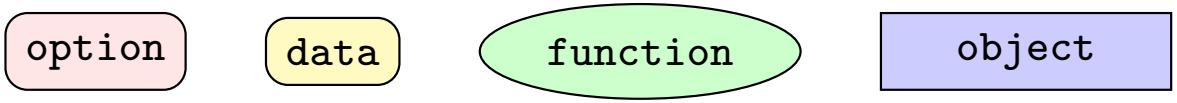
Solutions

- Branch-and-bound Algorithm
  - \+ Finds all equivalent solutions
  - \+ Highly controllable
  - – May be slow
- Translate to mixed-integer programming (MIP) problem
  - \+ Fast
  - \+ Returns location and possible values
  - – Numerical stability issues
  - – Only returns one solution

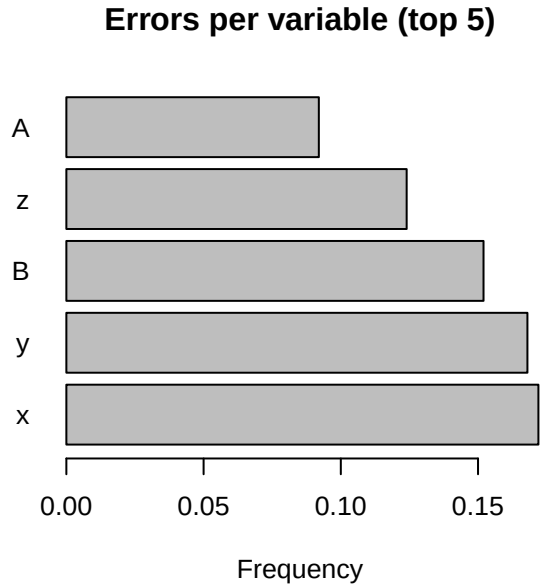
Error localization with `localizeErrors`

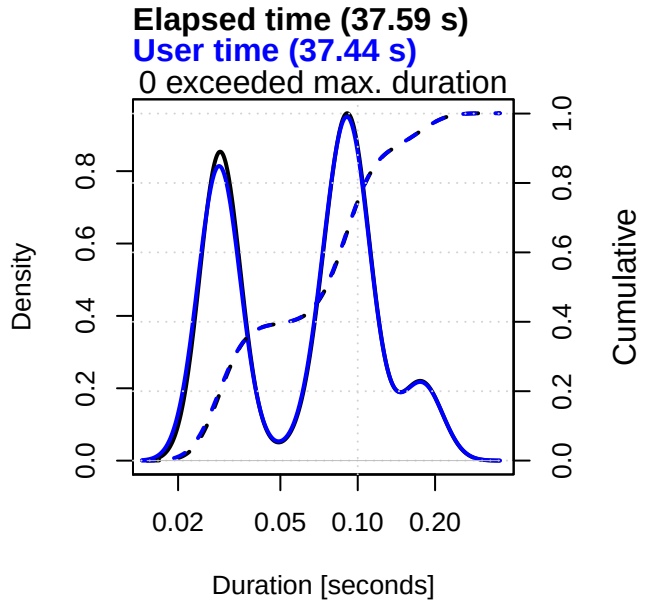


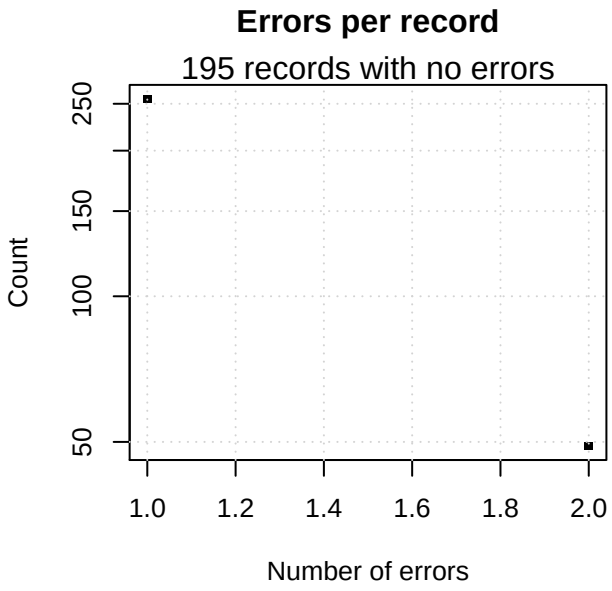
Legend

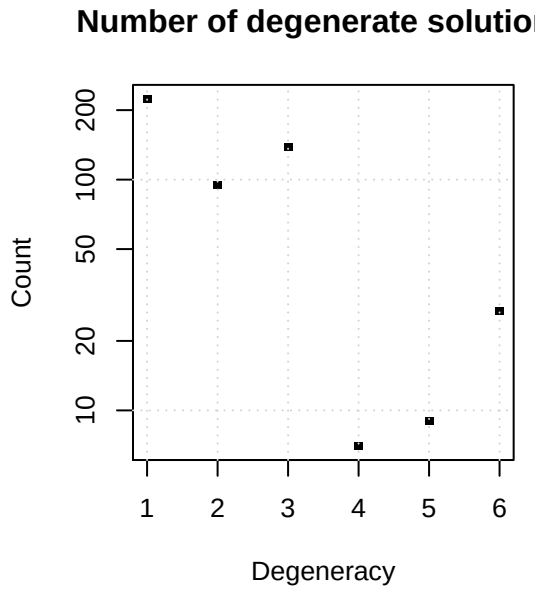


Example: `localizeErrors` with Branch and Bound

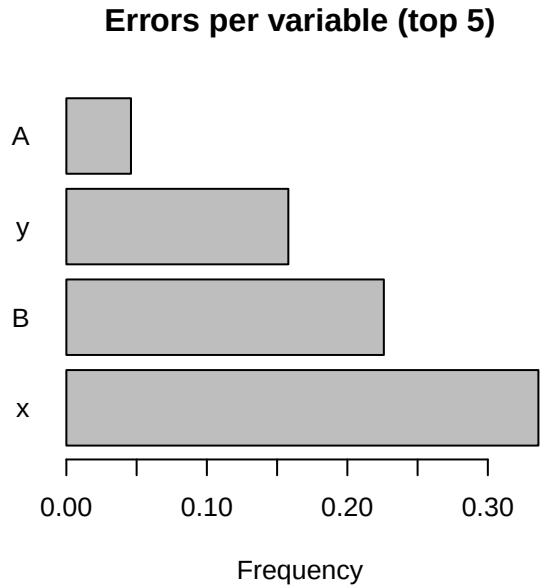


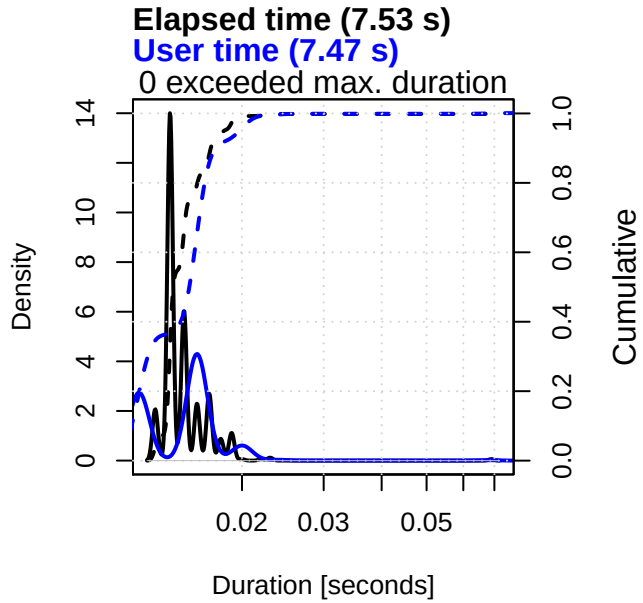


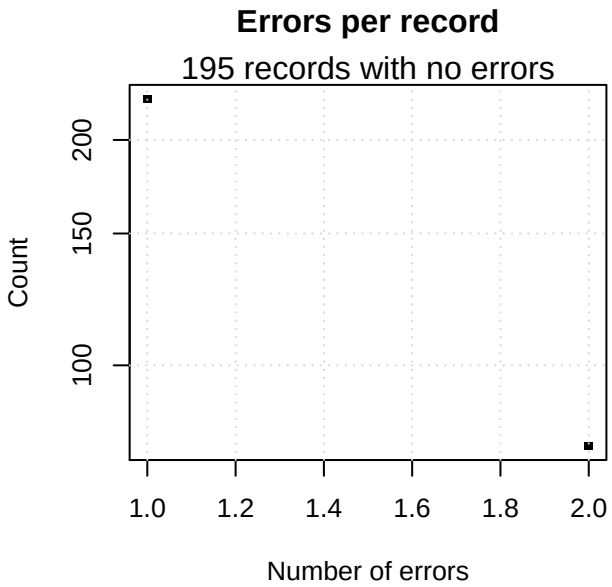




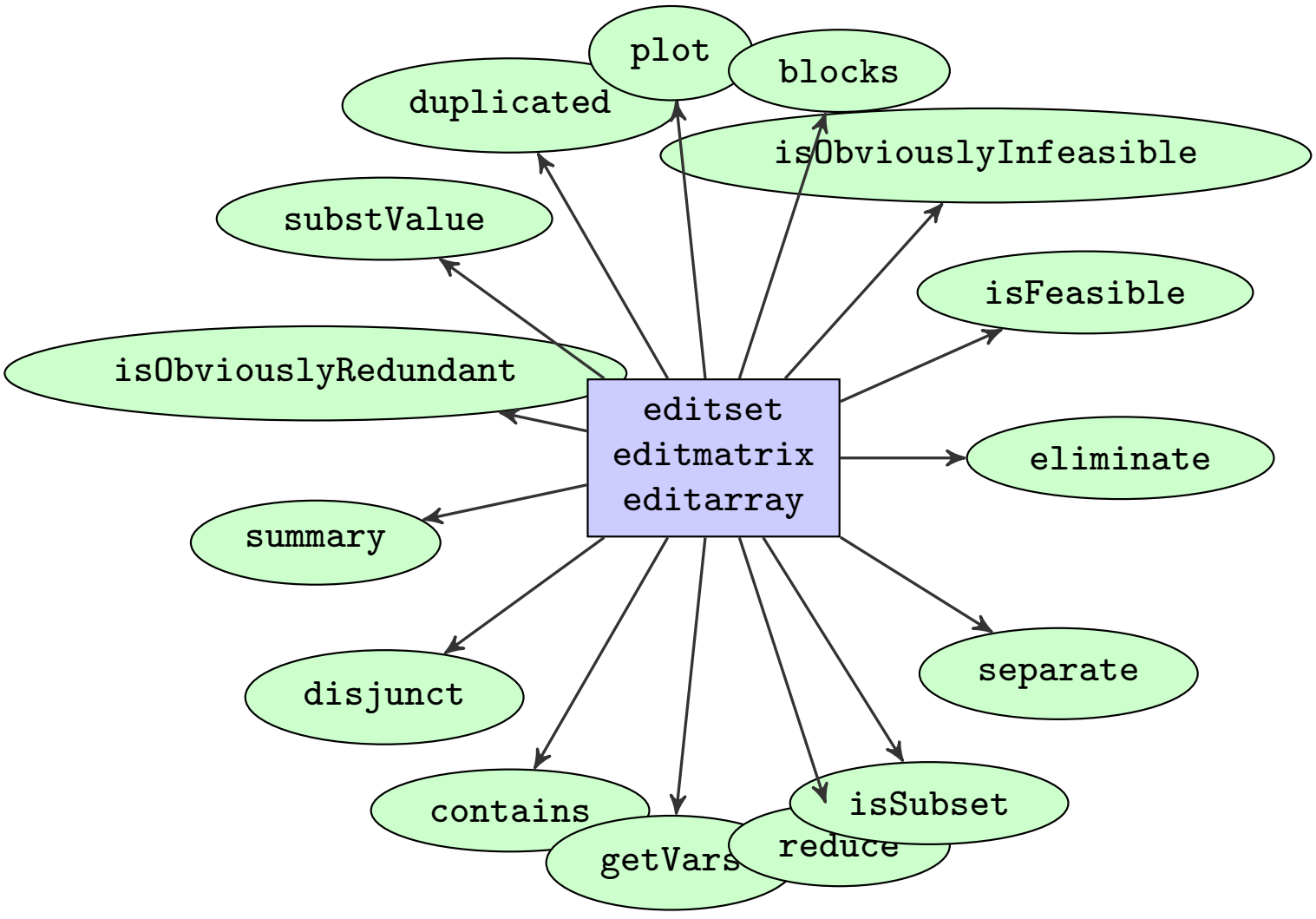
Example: `localizeErrors` using MIP







Manipulating edits



Deducorrect

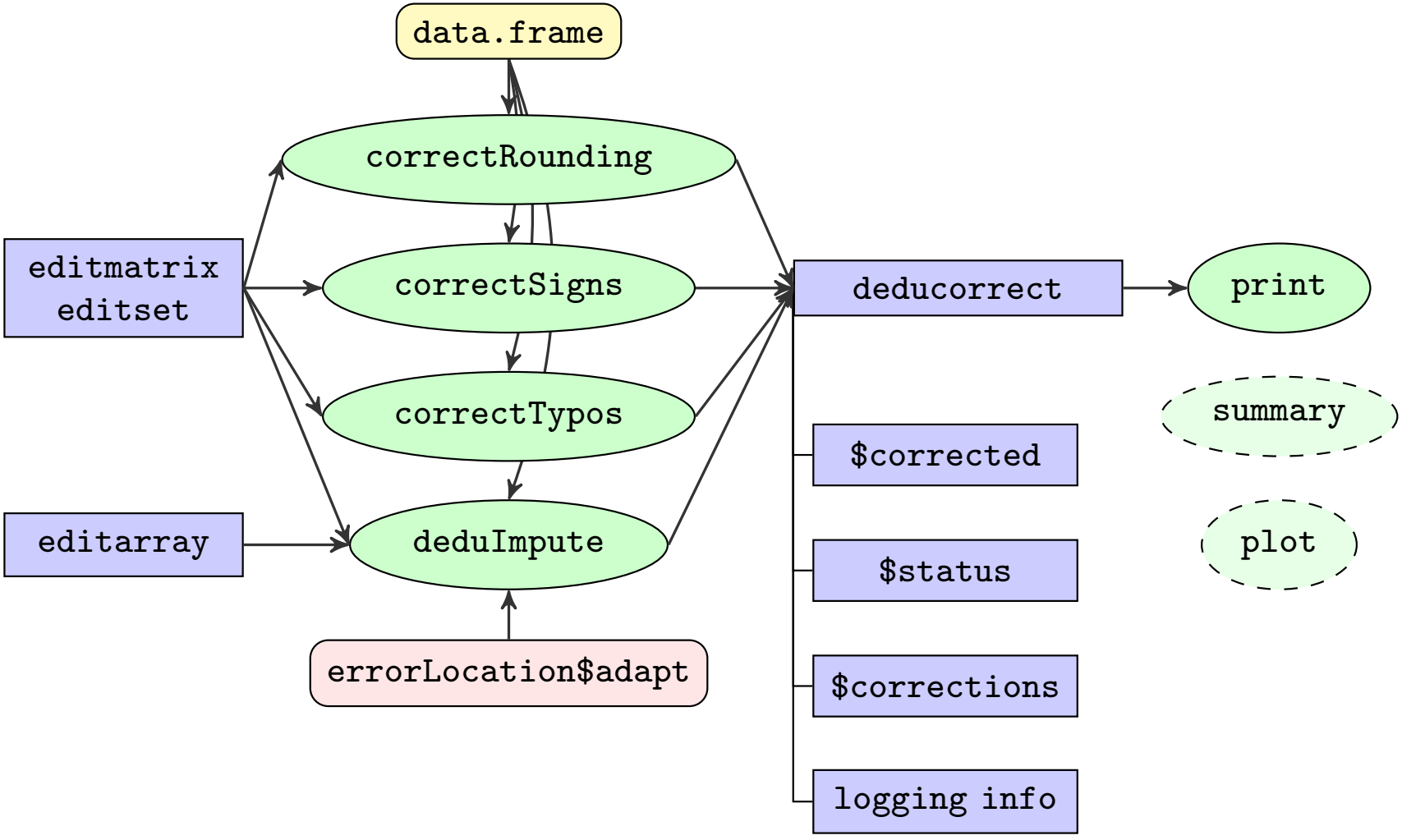
A uniform interface for deductive data correction methods:

- Solve typos
- Solve rounding errors
- Solve sign errors, variable swaps
- Deductive imputatition

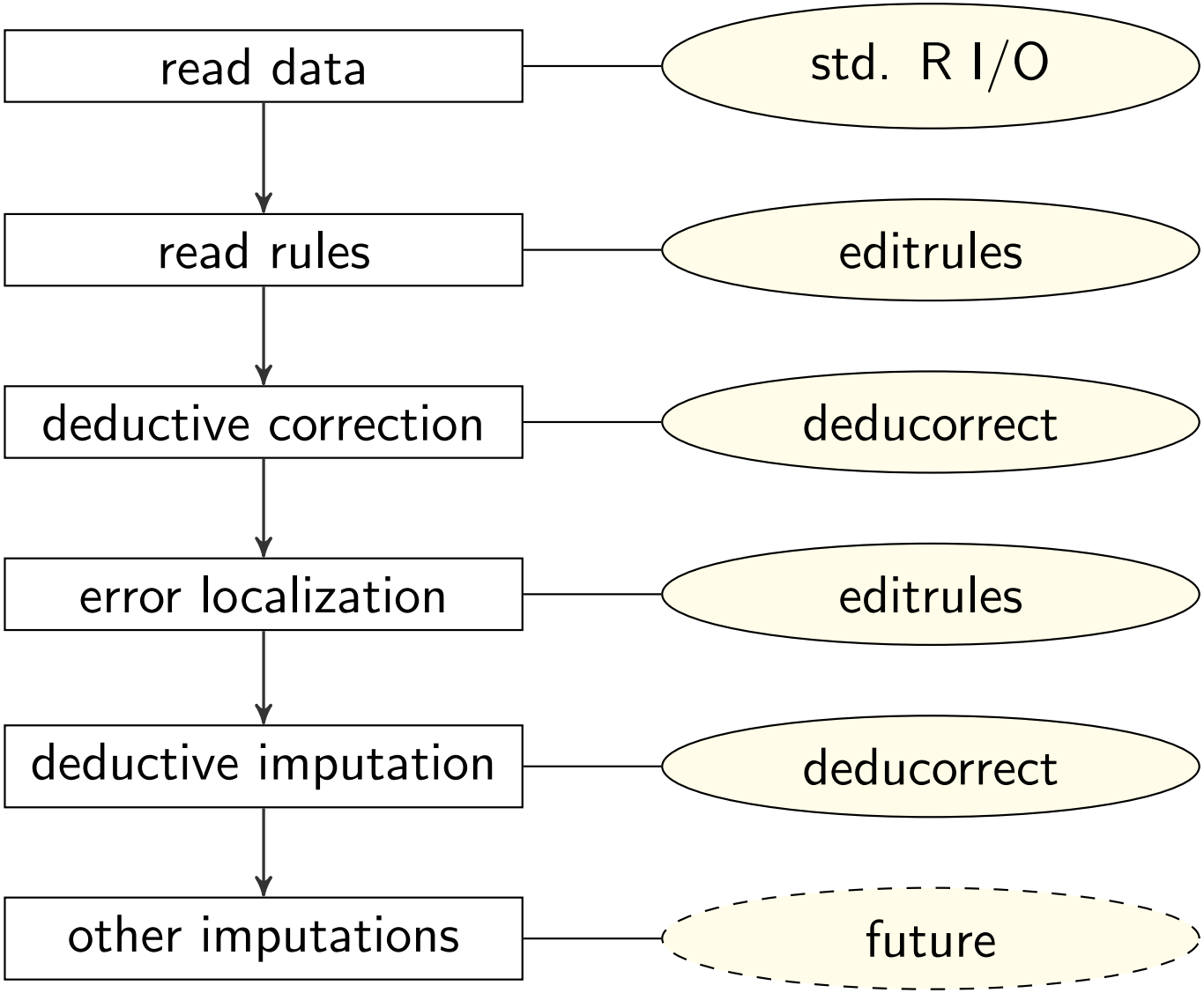
Main idea:

- Use values in erroneous record to trace the correct value
- Use rules to derive possible solutions

Interface correction methods

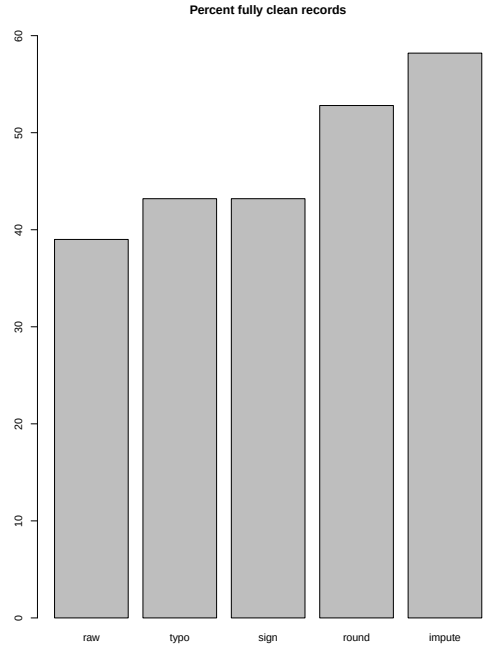


Workflow


read data
std. R I/O
read rules
editrules
deductive correction
deducorrect
error localization
editrules
deductive imputation
deducorrect
other imputations
future

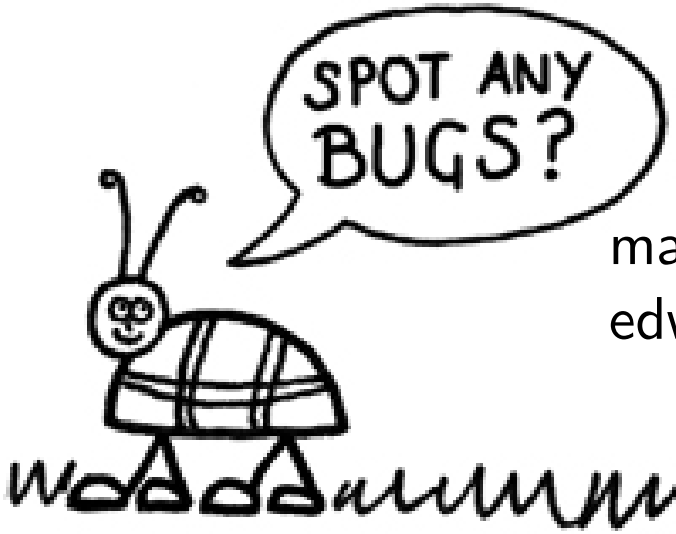
Effects on random 500 × 5 dataset example

```
### DATA CLEANING IN 7 STATEMENTS

### read data and rules
dat <- read.csv2("mydata.csv",comment.char="#")
E <- editfile("edits.txt")

### deductive correction
dat1 <- correctTypos(E, dat)
dat2 <- correctSigns(E, dat1$corrected)
dat3 <- correctRounding(E, dat2$corrected)

### localize errors
el <- localizeErrors(E, dat3$corrected)

### deductive imputation
dat4 <- deduImpute(E, dat3$corrected,
    adapt=el$adapt)
```



Treatment	Fully clean records	Nr of violations
nothing	195 (39.0%)	382
correctTypos	216 (43.2%)	345
correctSigns	216 (43.2%)	345
correctRounding	264 (52.8%)	272
errorLocalizer	—	
deduImpute	291 (58.2%)	239

End



mark.vanderloo@gmail.com
edwindjonge@gmail.com

- De Jonge, E. and Van der Loo, M. (2011) Manipulation of linear edits and error localization with the editrules package. Discussion paper 201120, Statistics Netherlands The Hague/Heerlen
- De Jonge, E. and Van der Loo, M. (2012) Error localization as a mixed-integer problem in editrules. (fortcoming).
- Scholtus, S. (2008) Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data. Discussion paper 08015, Statistics Netherlands The Hague/Heerlen
- Scholtus, S. (2009) Automatic correction of simple typing errors in numerical data with balance edits. Discussion paper 09046, Statistics Netherlands The Hague/Heerlen
- Van der Loo, M. and De Jonge (2011) Manipulation of categorical data edits and error localization with the editrules package Discussion paper 201129, Statistics Netherlands, The Hague/Heerlen.
- Van der Loo, M. (2012) Variable elimination and edit generation with a flavour of semigroup algebra (submitted).
- Van der Loo, M. and De Jonge ,E. (2011) Deductive imputation with the deducorrect package. Discussion paper 201126, Statistics Netherlands The Hague/Heerlen
- Van der Loo, M. De Jonge, E. and Scholtus, S. (2011) Correction of rounding, typing and sign errors with the deducorrect package. Discussion paper 201119, Statistics Netherlands The Hague/Heerlen.
- Van der Loo, M. and De Jonge, E. (2012) Manipulation of conditional restrictions and error localization with the editrules package (forthcoming).