P I K

Potsdam Institute for Climate Impact Research

David Klein

How to save a scientist‘s career with data classes?

deRSE   Potsdam   5.6.2019

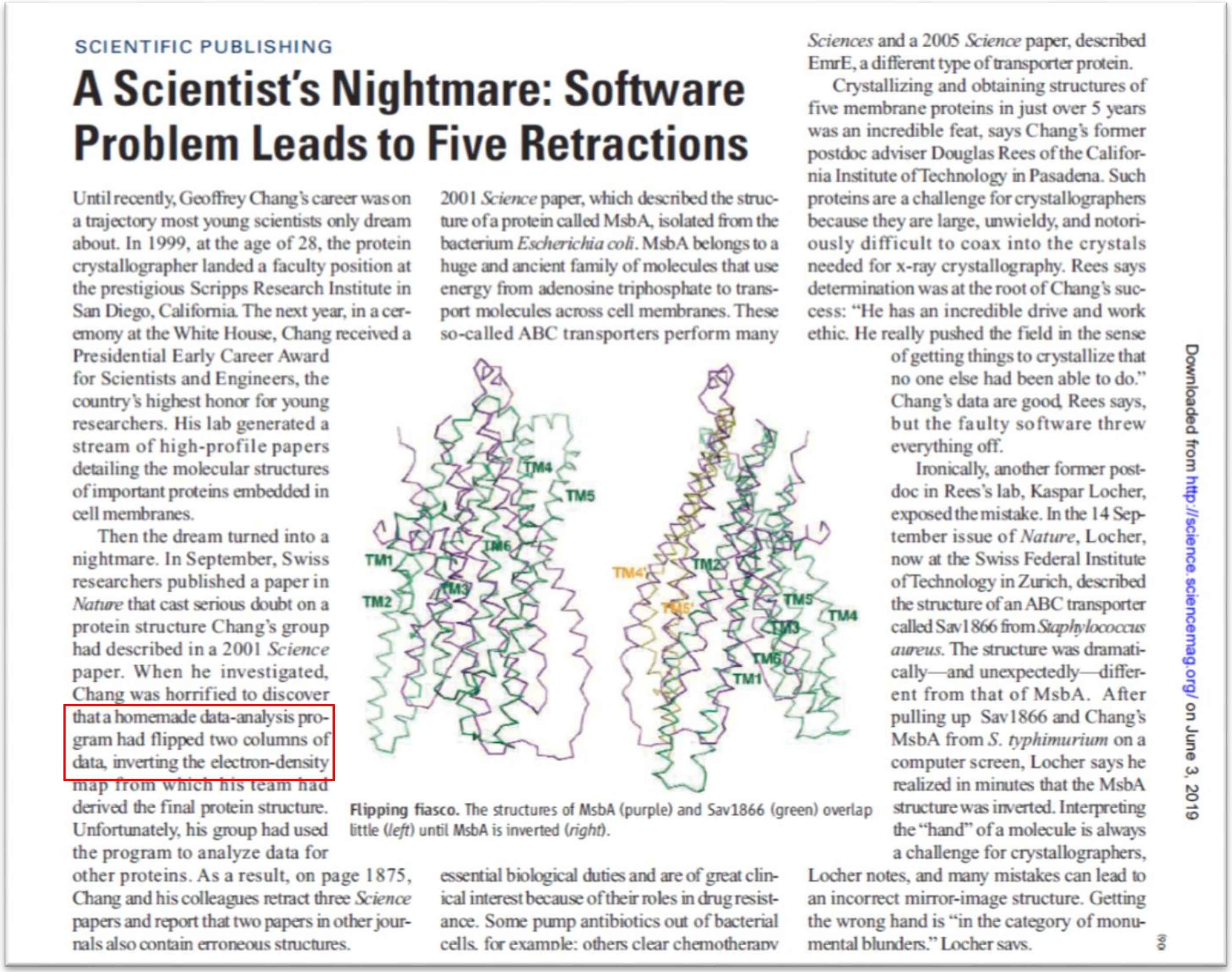
“I can’t replicate your results.”

https://www.theatlantic.com/business/archive/2013/04/forget-excel-this-was-reinhart-and-rogoffs-biggest-mistake/275088/

“I can’t replicate your results.”

“You made an Excel error.”

https://www.theatlantic.com/business/archive/2013/04/forget-excel-this-was-reinhart-and-rogoffs-biggest-mistake/275088/



TM4
TM5
TM6
TM1
TM2
TM3
TM4'
TM5'
TM2
TM5
TM3
TM4
TM6
TM1

SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was on a trajectory most young scientists only dream about. In 1999, at the age of 28, the protein crystallographer landed a faculty position at the prestigious Scripps Research Institute in San Diego, California. The next year, in a ceremony at the White House, Chang received a Presidential Early Career Award for Scientists and Engineers, the country's highest honor for young researchers. His lab generated a stream of high-profile papers detailing the molecular structures of important proteins embedded in cell membranes.

Then the dream turned into a nightmare. In September, Swiss researchers published a paper in *Nature* that cast serious doubt on a protein structure Chang's group had described in a 2001 *Science* paper. When he investigated, Chang was horrified to discover that a homemade data-analysis program had flipped two columns of data, inverting the electron-density map from which his team had derived the final protein structure. Unfortunately, his group had used the program to analyze data for other proteins. As a result, on page 1875, Chang and his colleagues retract three *Science* papers and report that two papers in other journals also contain erroneous structures.

2001 *Science* paper, which described the structure of a protein called MsbA, isolated from the bacterium *Escherichia coli*. MsbA belongs to a huge and ancient family of molecules that use energy from adenosine triphosphate to transport molecules across cell membranes. These so-called ABC transporters perform many essential biological duties and are of great clinical interest because of their roles in drug resistance. Some pump antibiotics out of bacterial cells, for example; others clear chemotherapy

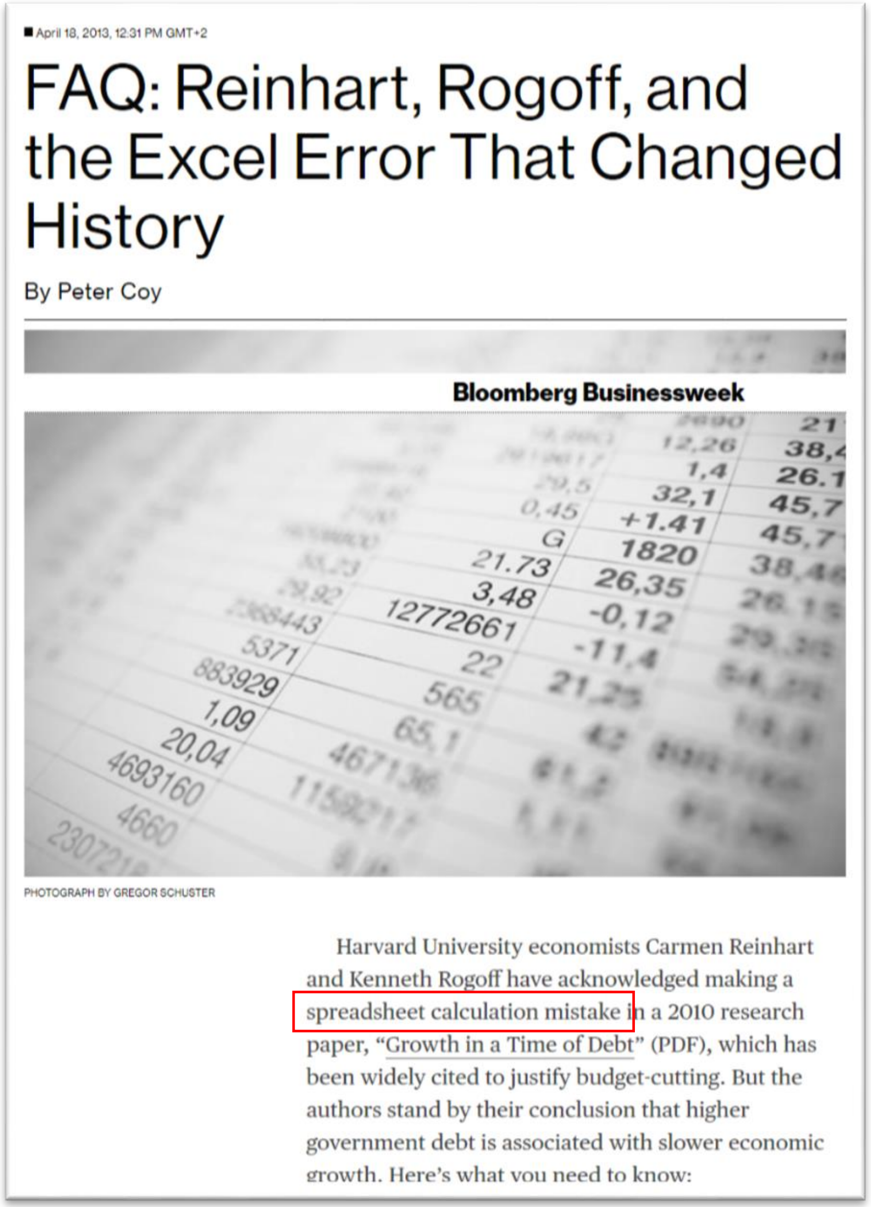
Flipping fiasco. The structures of MsbA (purple) and Sav1866 (green) overlap little (*left*) until MsbA is inverted (*right*).

Sciences and a 2005 *Science* paper, described EmrE, a different type of transporter protein.

Crystallizing and obtaining structures of five membrane proteins in just over 5 years was an incredible feat, says Chang's former postdoc adviser Douglas Rees of the California Institute of Technology in Pasadena. Such proteins are a challenge for crystallographers because they are large, unwieldy, and notoriously difficult to coax into the crystals needed for x-ray crystallography. Rees says determination was at the root of Chang's success: "He has an incredible drive and work ethic. He really pushed the field in the sense of getting things to crystallize that no one else had been able to do." Chang's data are good, Rees says, but the faulty software threw everything off.

Ironically, another former postdoc in Rees's lab, Kaspar Locher, exposed the mistake. In the 14 September issue of *Nature*, Locher, now at the Swiss Federal Institute of Technology in Zurich, described the structure of an ABC transporter called Sav1866 from *Staphylococcus aureus*. The structure was dramatically—and unexpectedly—different from that of MsbA. After pulling up Sav1866 and Chang's MsbA from *S. typhimurium* on a computer screen, Locher says he realized in minutes that the MsbA structure was inverted. Interpreting the "hand" of a molecule is always a challenge for crystallographers, Locher notes, and many mistakes can lead to an incorrect mirror-image structure. Getting the wrong hand is "in the category of monumental blunders." Locher says.

Downloaded from http://science.sciencemag.org/ on June 3, 2019



■ April 18, 2013, 12:31 PM GMT+2

FAQ: Reinhart, Rogoff, and the Excel Error That Changed History

By Peter Coy

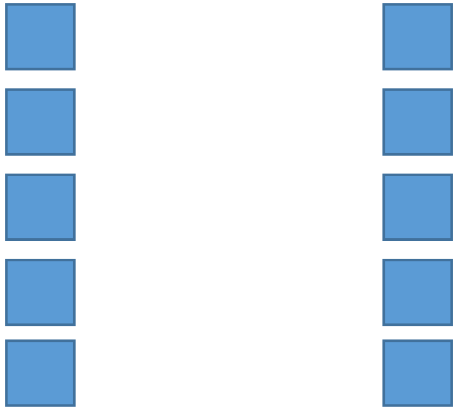
Bloomberg Businessweek

PHOTOGRAPH BY GREGOR SCHUSTER

Harvard University economists Carmen Reinhart and Kenneth Rogoff have acknowledged making a spreadsheet calculation mistake in a 2010 research paper, "Growth in a Time of Debt" (PDF), which has been widely cited to justify budget-cutting. But the authors stand by their conclusion that higher government debt is associated with slower economic growth. Here's what you need to know:

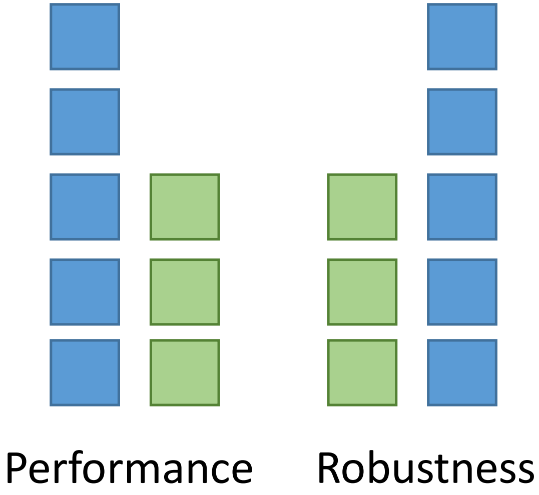
https://science.sciencemag.org/content/314/5807/1856
https://www.bloomberg.com/news/articles/2013-04-18/faq-reinhart-rogoff-and-the-excel-error-that-changed-history

Increase robustness of research software

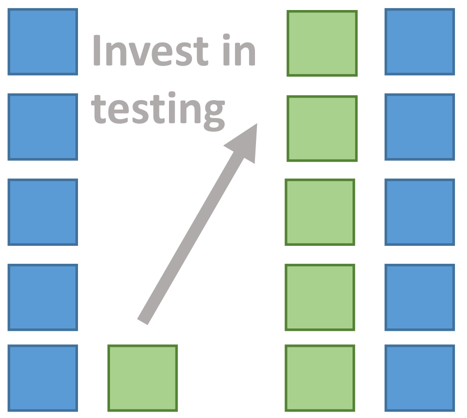


Performance Robustness

Increase robustness of research software

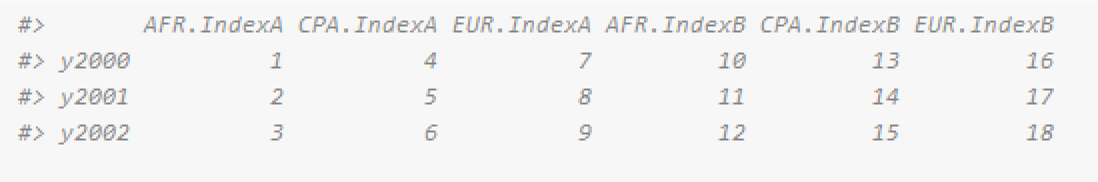


Increase robustness of research software

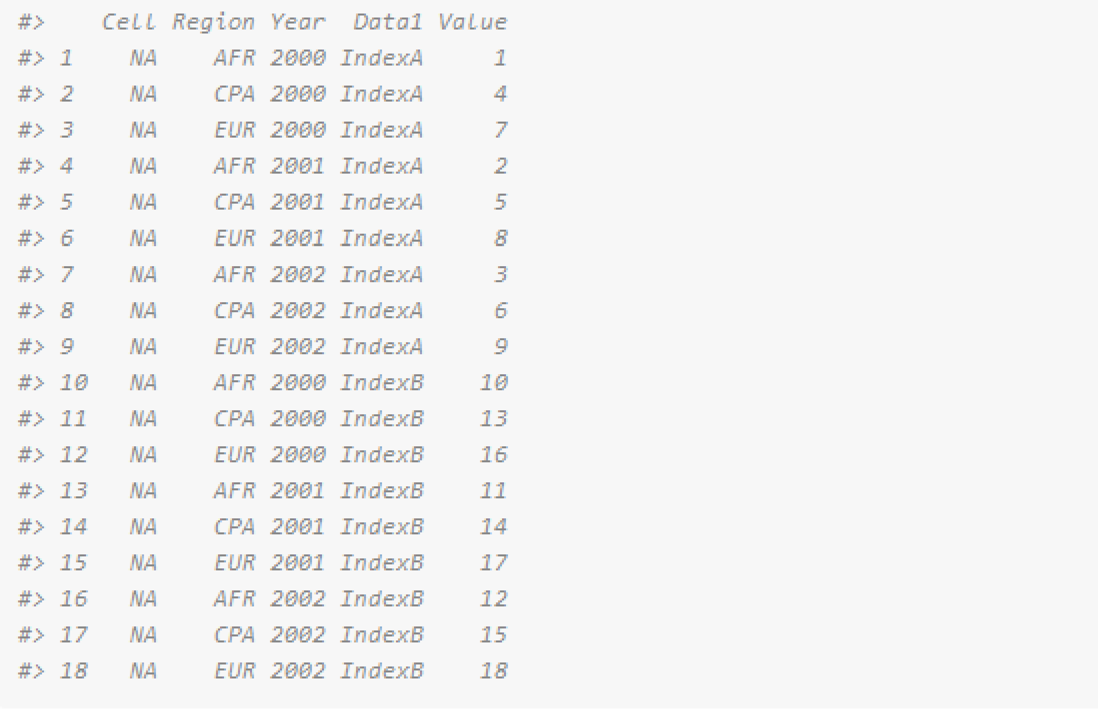


Performance Robustness

1. Standardized, generic structure



```
#>       AFR.IndexA CPA.IndexA EUR.IndexA AFR.IndexB CPA.IndexB EUR.IndexB
#> y2000          1          4          7         10         13         16
#> y2001          2          5          8         11         14         17
#> y2002          3          6          9         12         15         18
```



```
#>    Cell Region Year  Data1 Value
#> 1    NA    AFR 2000 IndexA     1
#> 2    NA    CPA 2000 IndexA     4
#> 3    NA    EUR 2000 IndexA     7
#> 4    NA    AFR 2001 IndexA     2
#> 5    NA    CPA 2001 IndexA     5
#> 6    NA    EUR 2001 IndexA     8
#> 7    NA    AFR 2002 IndexA     3
#> 8    NA    CPA 2002 IndexA     6
#> 9    NA    EUR 2002 IndexA     9
#> 10   NA    AFR 2000 IndexB    10
#> 11   NA    CPA 2000 IndexB    13
#> 12   NA    EUR 2000 IndexB    16
#> 13   NA    AFR 2001 IndexB    11
#> 14   NA    CPA 2001 IndexB    14
#> 15   NA    EUR 2001 IndexB    17
#> 16   NA    AFR 2002 IndexB    12
#> 17   NA    CPA 2002 IndexB    15
#> 18   NA    EUR 2002 IndexB    18
```

- Avoid mistakes by using a standardized data structure
- Various input formats are transferred into the same structure.
- Data class is flexible with regard to numbers of dimensions (columns)
- It detects dimensions in the raw data and adds them to the object (as columns)

2. Name matching

- Automatically matches entries when performing operations
- Columns can not be mixed up

GDP

region	period	value
EUR	2005	13386
LAM	2005	4984
USA	2005	12584
EUR	2010	14024
LAM	2010	5908
USA	2010	13291

Ordered by year

/

Population

region	period	value
EUR	2005	496
EUR	2010	505
LAM	2005	561
LAM	2010	597
USA	2005	296
USA	2010	309

Ordered by region

=

GDP per capita

region	period	value
EUR	2005	26.99
LAM	2005	8.88
USA	2005	42.51
EUR	2010	27.77
LAM	2010	9.90
USA	2010	43.01

3. Units

```
> a <- 2
> a <- set_units(a,"m")
> b <- 4
> b <- set_units(b,"km")

> a+b
4002
> units(a+b)
[m]

> time <- 5
> time <- set_units(time,"s")
> velo <- a/time

> velo
0.4
> units(velo)
[m*s^-1]

> units(velo) <- "km*h^-1"
> velo
1.44
```

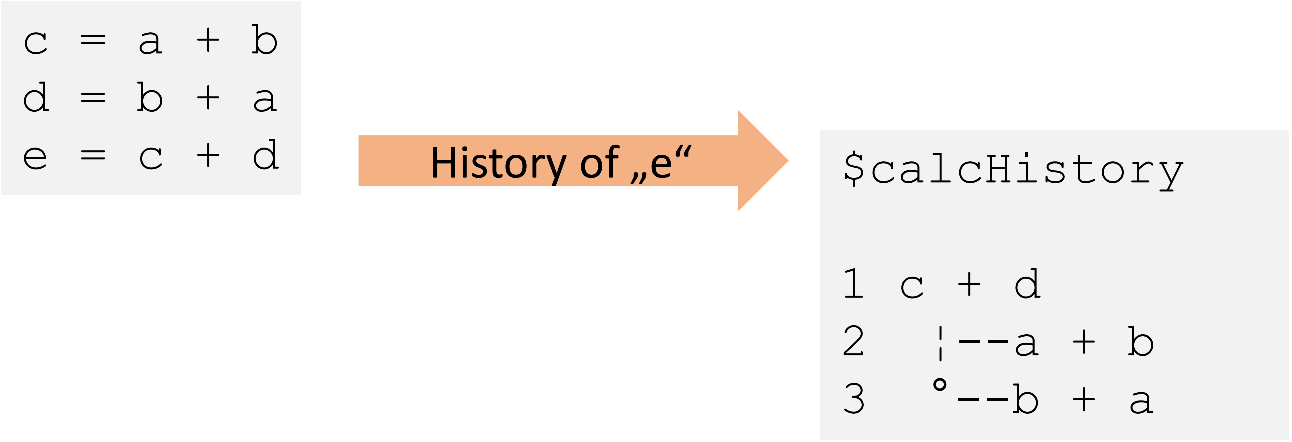
Manually changing the unit

Auto conversion of values

- Variables have units attached
- Applies conversion factors when combining variables
- Reports an error if conversion is unknown

4. History

- Automatically **logs operations** performed on the data
- Allows following the operations and **detecting mistakes ex post**
- even if routine that performed the operations is not available



```
c = a + b
d = b + a
e = c + d
```

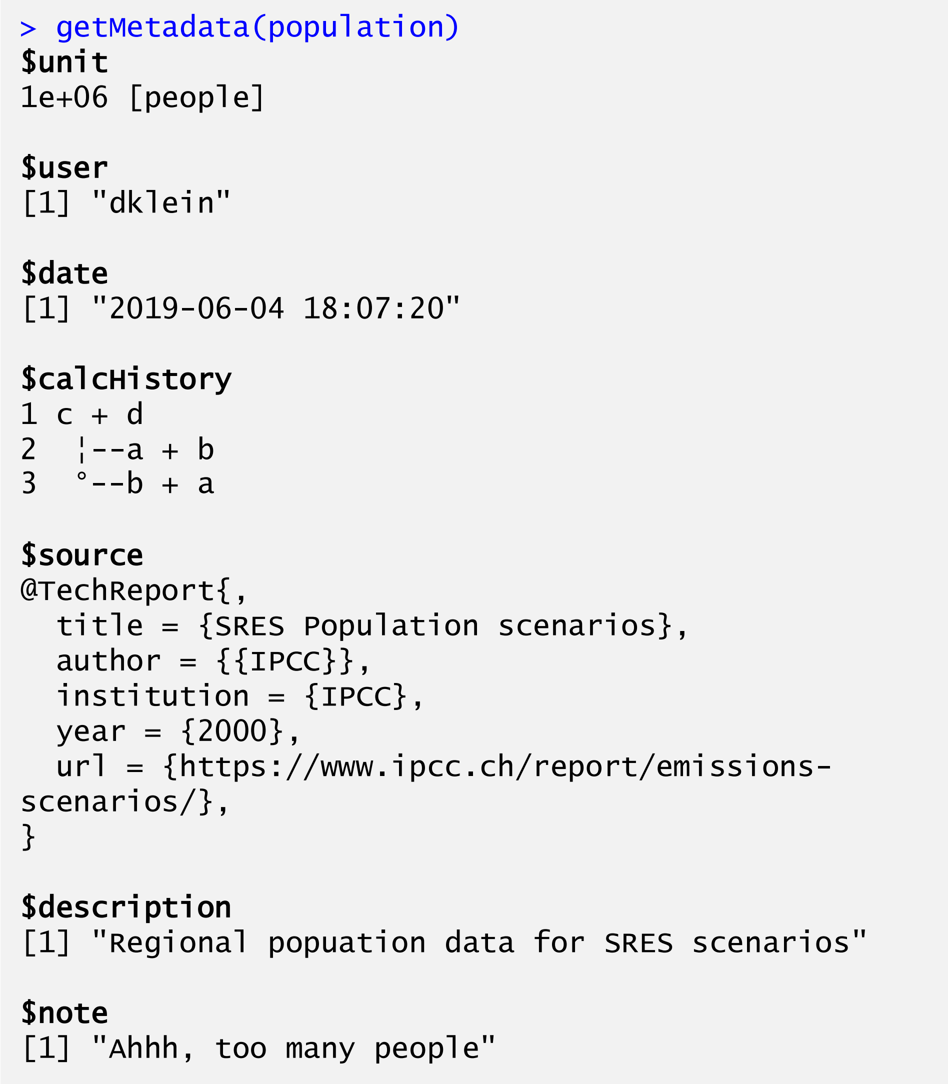
History of „e“

```
$calcHistory

1 c + d
2  ¦--a + b
3  °--b + a
```

4. History

```
$calcHistory

1  calcOutput("TauTotal")
2   °--toolAggregate(x = x$x, weight = x$weight, rel = reg_rel)
3       ¦--toolAggregate(x * weight, rel, from = from, to = to, dim = dim, partrel = partrel) * weight2
4       ¦   ¦--x * weight
5       ¦   ¦   °--readSource("Tau", "paper")
6       ¦   ¦       ¦--toolAggregate(tau, rel = iso_cell, weight = collapseNames(xref))
7       ¦   ¦       ¦   °--toolAggregate(x * weight, rel, from = from, to = to, dim = dim, partrel = partrel) * weight2
8       ¦   ¦       ¦       ¦--x * weight
9       ¦   ¦       ¦       °--1/(toolAggregate(weight, rel, from = from, to = to, dim = dim, partrel = partrel, verbosity = 10) + 10^-100)
10      ¦   ¦       °--toolAggregate(xref, rel = iso_cell)
11      ¦   °--1/(toolAggregate(weight, rel, from = from, to = to, dim = dim, partrel = partrel, verbosity = 10) + 10^-100)
12      ¦       °--readSource("Tau", "paper")
13      ¦           ¦--toolAggregate(tau, rel = iso_cell, weight = collapseNames(xref))
14      ¦           ¦   °--toolAggregate(x * weight, rel, from = from, to = to, dim = dim, partrel = partrel) * weight2
15      ¦           ¦       ¦--x * weight
16      ¦           ¦       °--1/(toolAggregate(weight, rel, from = from, to = to, dim = dim, partrel = partrel, verbosity = 10) + 10^-100)
17      ¦           °--toolAggregate(xref, rel = iso_cell)
18      °--readSource("Tau", "paper")
19          ¦--toolAggregate(tau, rel = iso_cell, weight = collapseNames(xref))
20          ¦   °--toolAggregate(x * weight, rel, from = from, to = to, dim = dim, partrel = partrel) * weight2
21          ¦       ¦--x * weight
22          ¦       °--1/(toolAggregate(weight, rel, from = from, to = to, dim = dim, partrel = partrel, verbosity = 10) + 10^-100)
23          °--toolAggregate(xref, rel = iso_cell)
```

5. Metadata: Documentation



```
> getMetadata(population)
$unit
1e+06 [people]

$user
[1] "dklein"

$date
[1] "2019-06-04 18:07:20"

$calcHistory
1 c + d
2  ¦--a + b
3  °--b + a

$source
@TechReport{,
  title = {SRES Population scenarios},
  author = {{IPCC}},
  institution = {IPCC},
  year = {2000},
  url = {https://www.ipcc.ch/report/emissions-
scenarios/},
}

$description
[1] "Regional popuation data for SRES scenarios"

$note
[1] "Ahhh, too many people"
```

- Documentation is attached to the data
- Helps user assessing whether it’s the data he/she expects

https://github.com/pik-piam/magclass

rse@pik-potsdam.de