Free and Open Source Software - R

in our statistical office – why?

Alexander Kowarik
Methods Unit

Pisa, 20.9.2022

www.statistik.at

Outline

- Computing in the Statistical Production
- Learn from DevOps
- Open Source Tool Development

About myself:

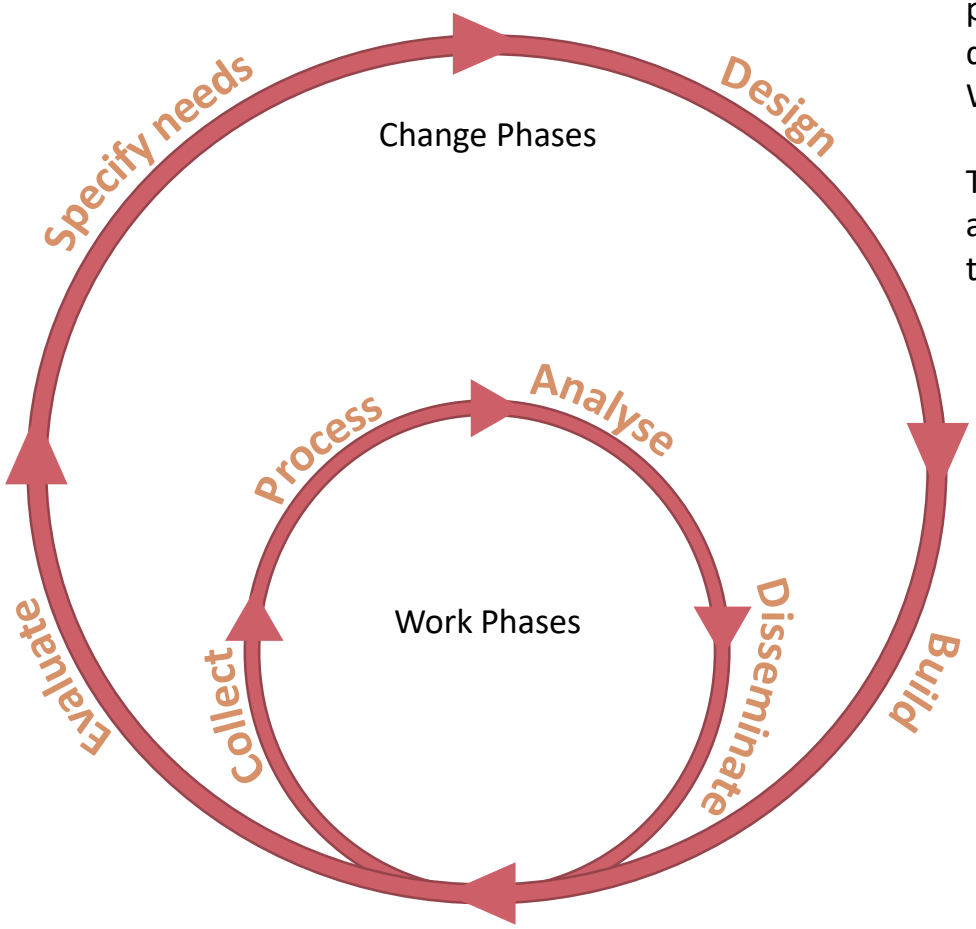
- Master and PhD in technical mathematics with focus on computational statistics
- Since 2008 in the methods unit @Statistics Austria
- Since 2014 head of the methods unit
- (Co-) author of several R packages

Computing in the statistical office II

- Tasks are split up between IT, methodologist and subject matter expert
  - IT Development skills
  - Mathematical, numerical and statistical skills
  - Expert knowledge in the field

- A task similar to Research Software Engineering: https://society-rse.org

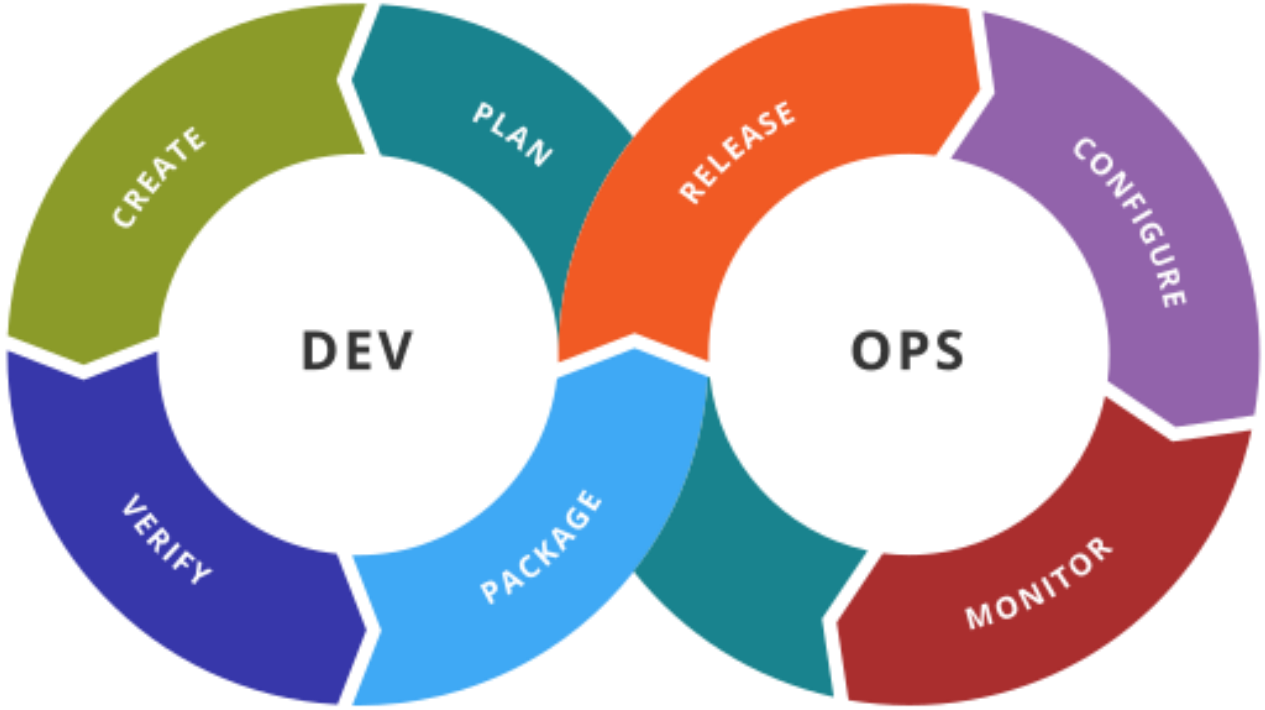
Statistical Production II



In GSBPM, there are some phases which are undertaken quickly and frequently – the Work Phases.

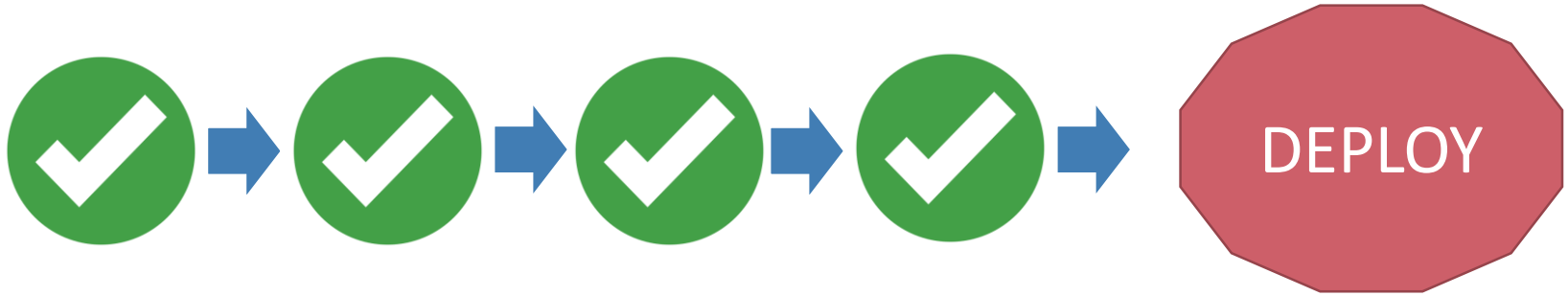
There are other phases which are undertaken less often - the Change Phases.

https://statswiki.unece.org/display/GSBPM/Clickable+GSBPM+v5.1



- Learning from DevOps (development and operations) practices in system development
- A few ideas can be “translated” to the statistical production process

Automate Everything (as much as possible)



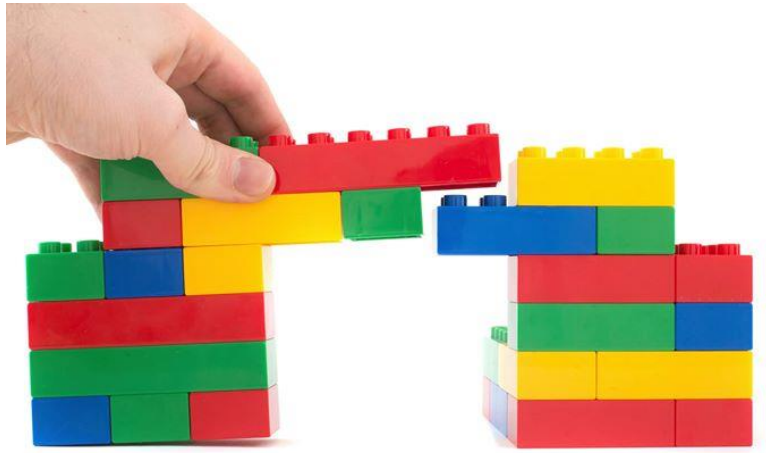
- Continuous Integration and Continuous Delivery (CI/CD) pipelines for data processing
- Quickly react to updated data or improved methods
- Data science, machine learning and statistical methods as automated steps in a streamlined process
- Infrastructure as code

Testing / Observing Quality Dimensions

- Unit testing of software components
- Plausibility checks for data (automated data editing)
- Quality controls/indicators throughout the process including traditional statistical quality dimensions
- Trigger manual intervention in case of „extreme quality events“

Foster Continuous Improvement

- Modularity allows to quickly integrate new methods, new data sources or new software tools.
- Integrate teams: statisticians, subject matter experts and developers



https://freshideen.com/trends/lego-spiele-im-begriff-einen-vorbildlichen-schritt-zu-tun.html

Be Reproducible

- Making the data generation process transparent
- Data versioning
- GUIs for data editing -> produce code

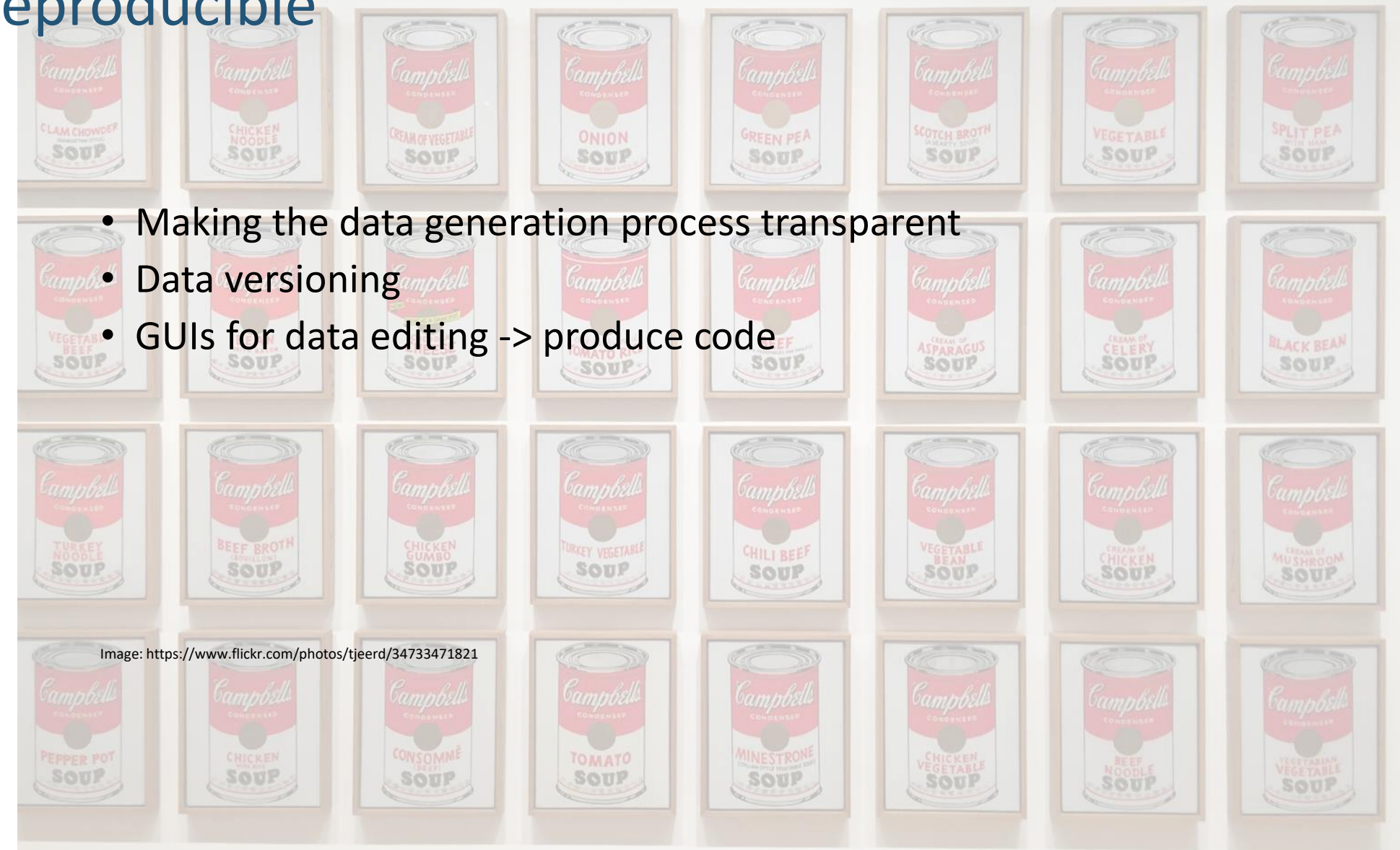
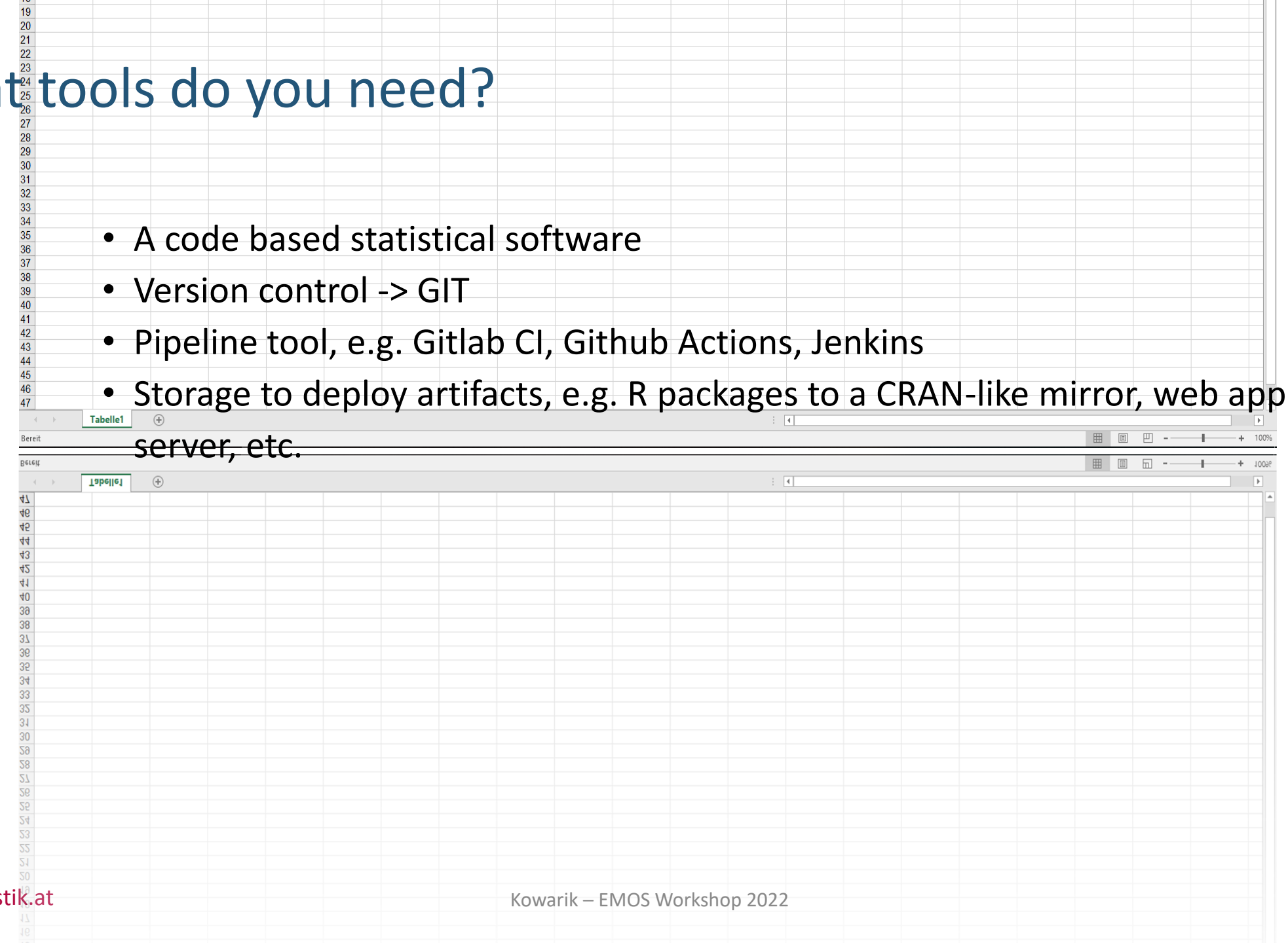


Image: https://www.flickr.com/photos/tjeerd/34733471821

What tools do you need?

- A code based statistical software
- Version control -> GIT
- Pipeline tool, e.g. Gitlab CI, Github Actions, Jenkins
- Storage to deploy artifacts, e.g. R packages to a CRAN-like mirror, web applications to a server, etc.



Collaborative developments

ESSNet projects to develop/maintain SDC software

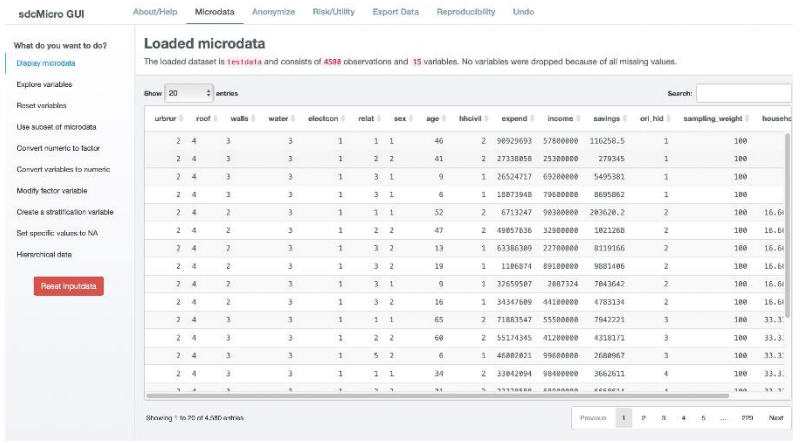
- **muArgus + tau Argus:** Initial development by CBS et al as closed software
- Later published as open source
- R packages **sdcMicro + sdcTable** developed initially in parallel
- Convergence of the developed methods

- **Now**: newly implemented method „Target Record Swapping“ as OS C++ library used by sdcMicro and muArgus

- Contributors from outside the project, worldwide usage of tools

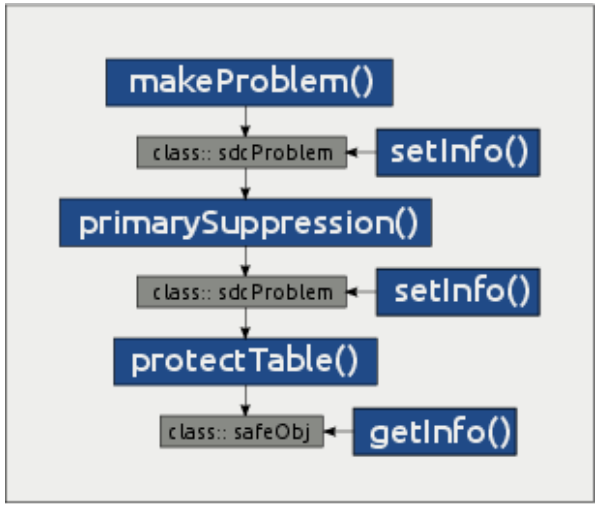
6.4 Apply disclosure control

- sdcMicro: https://sdctools.github.io/sdcMicro/
- Disclosure risk estimation
- Microaggregation, (correlated) noise , local suppression, …
- Shiny application with
  - reproducible output



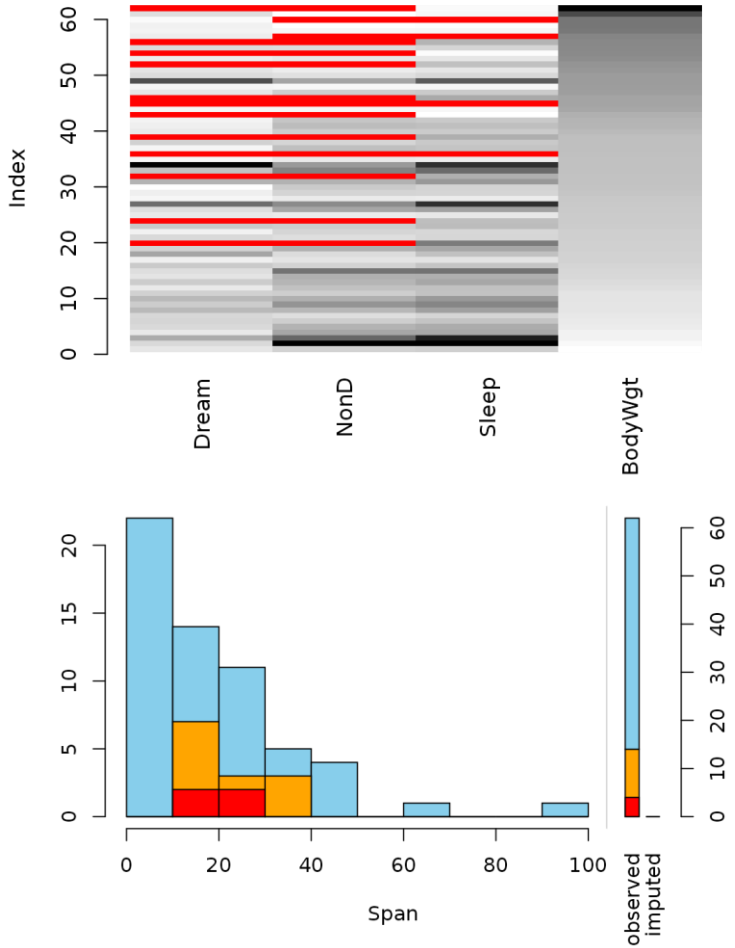
6.4 Apply disclosure control I

- sdcTable: https://sdctools.github.io/sdcTable/



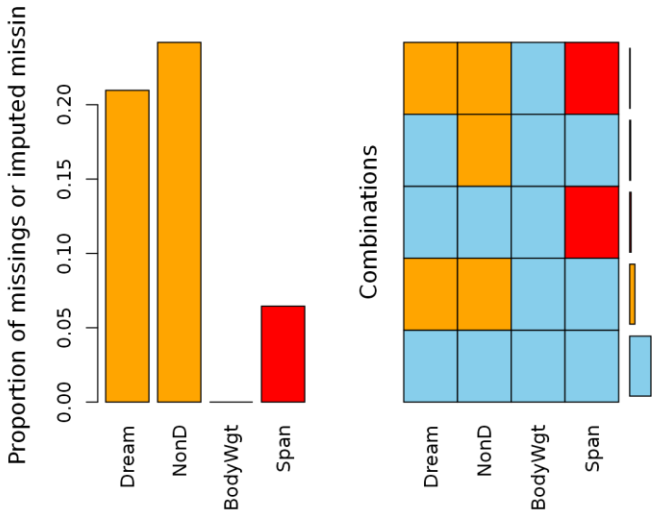
W	A	B	C	Total
X	20	50	10	80
Y	S	19	P	49
Z	S	32	S	61
Total	45	101	44	190

5.4 Edit and impute (6.2 validate outputs)



R package VIM:

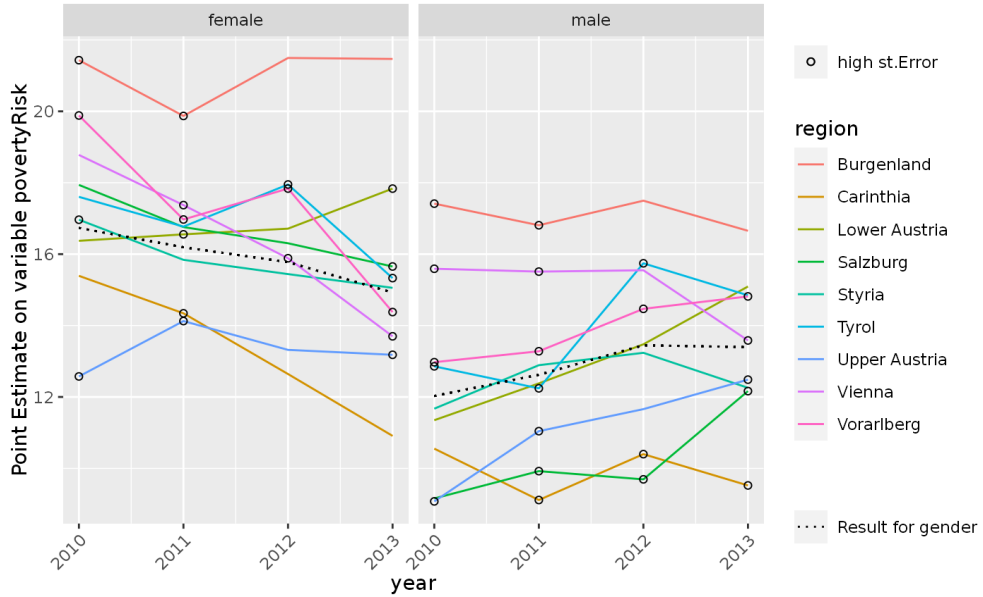
- Visualization and Imputation of Missing Values
- Imputation:
  - Donor-based: kNN, hotdeck, matchImpute
  - Model-based: irmi, regressionImp



5.6 Calculate weights (+ analyze 6.1 - 6.3)

- Calculate weights - calibration
- Draw bootstrap samples with calibration
- Estimate errors and confidence intervals
- Some simple visualizations to analysis results

- https://statistikat.github.io/surveysd



Open Source Tools – Use, Contribute, Modify

What is the awesome list?