OTTAWA 2023

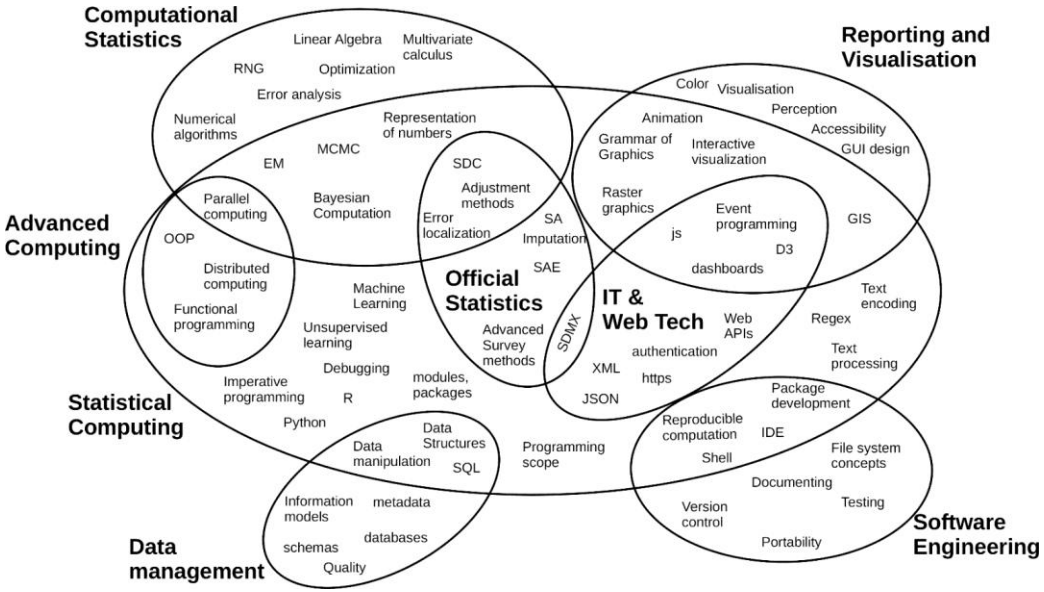
64TH WORLD STATISTICS CONGRESS

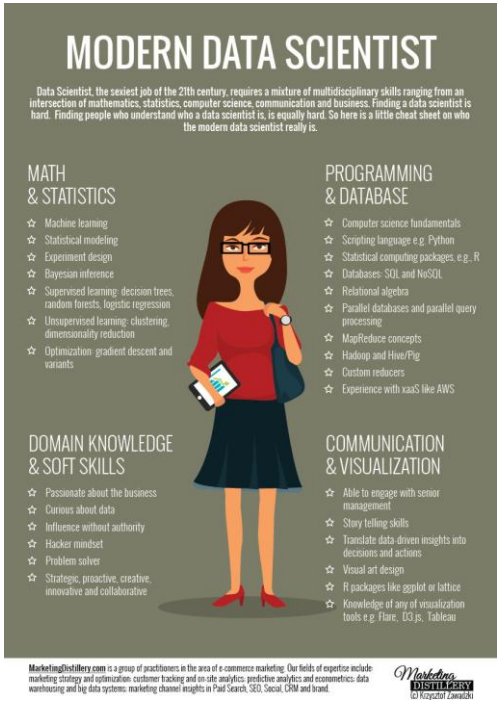
IPS 96 - Computing In The Modern Statistical Office

DevOps for the statistical production process?

Alexander Kowarik
Statistics Austria
18 July 2023

(Computing in the statistical office I)





- Loo, Mark. (2021). Computing in the statistical office. Statistical Journal of the IAOS. 37. 1023-1036. 10.3233/SJI-210862.
- http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

(Computing in the statistical office II)

Tasks are split up between IT, methodologist and subject matter expert

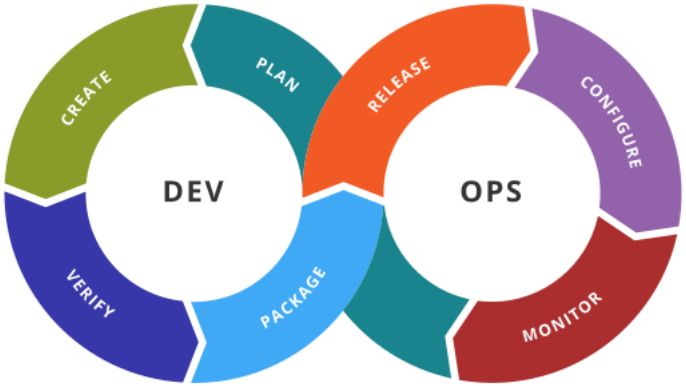
- IT Development skills
- Mathematical, numerical and statistical skills
- Expert knowledge in the field

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

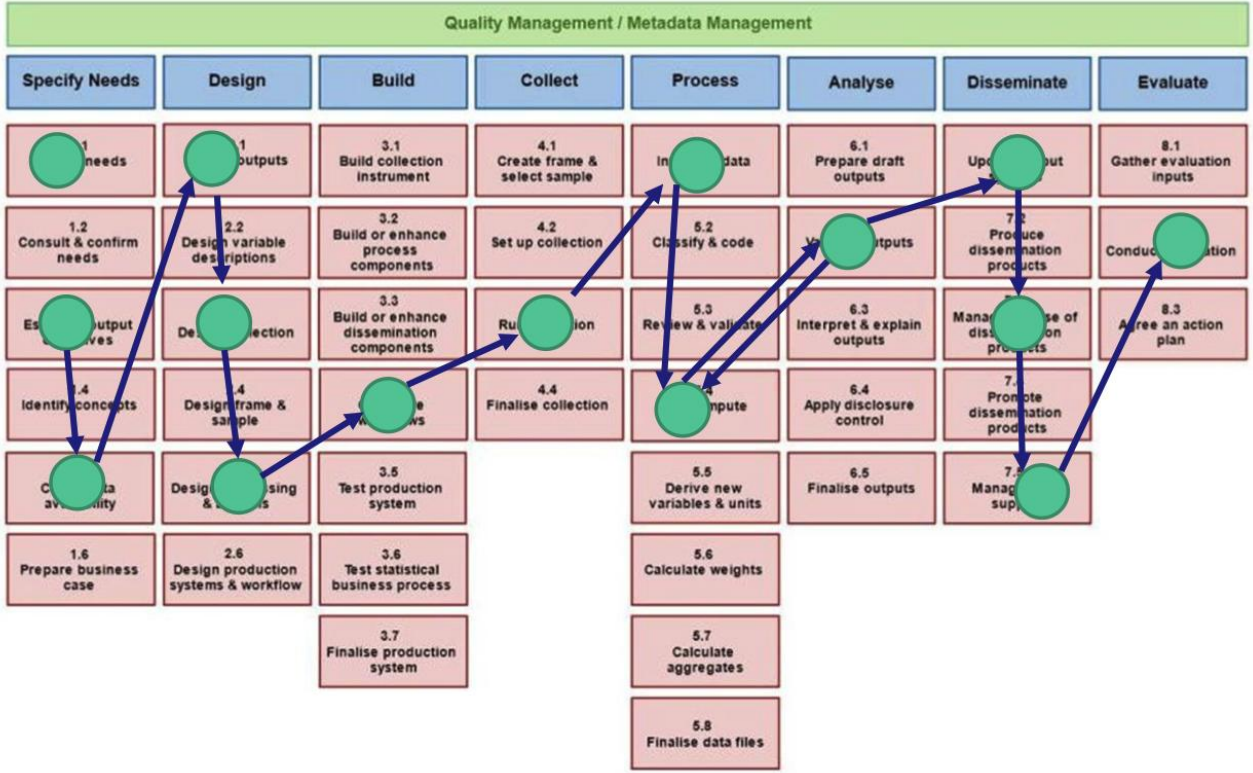
Learning from DevOps practices in system development

(development and operations)

A few ideas can be “translated” to the statistical production process



Statistical Production is (also) not linear



GSBPM is not designed to be linear.

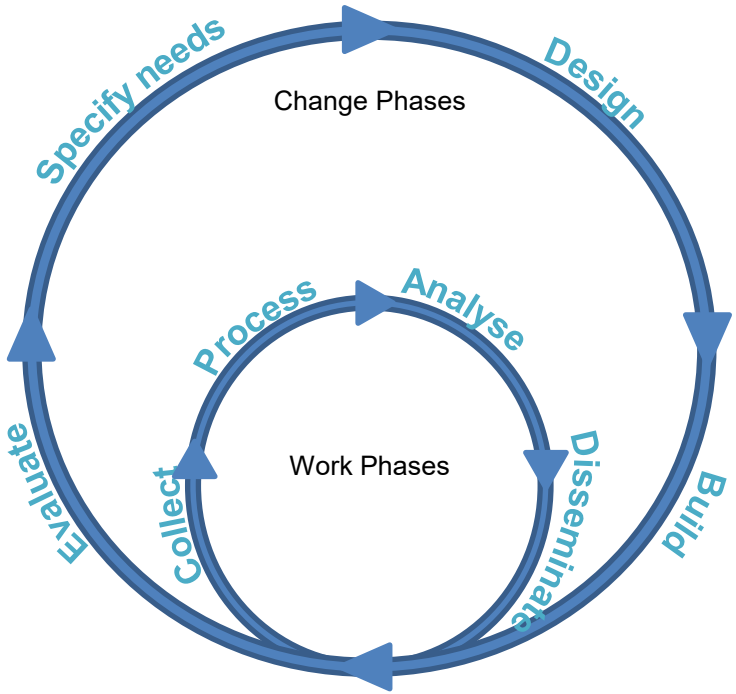
Process

Phases

Sub-processes

https://unece.org/statistics/events/MWW2020

GSPBM phases have different intensities and are cyclic

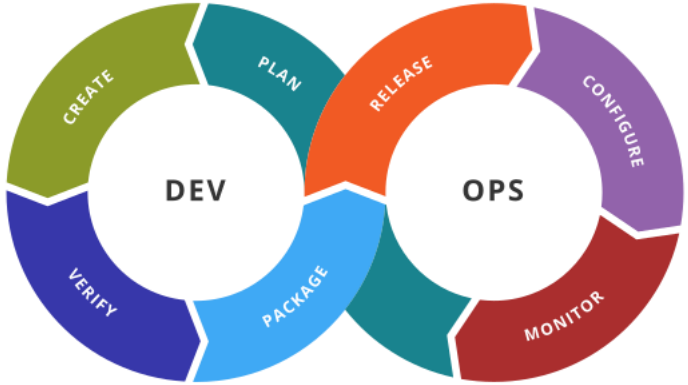


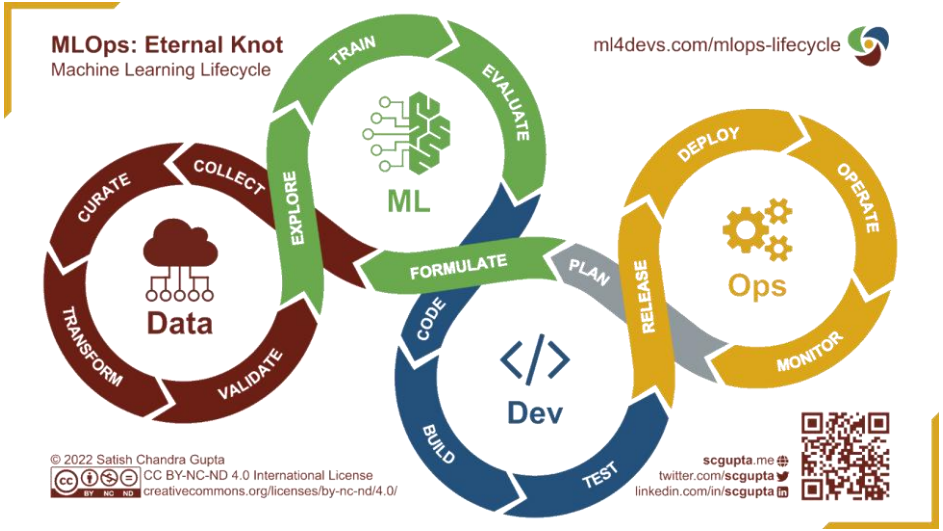
https://statswiki.unece.org/display/GSBPM/Clickable+GSBPM+v5.1

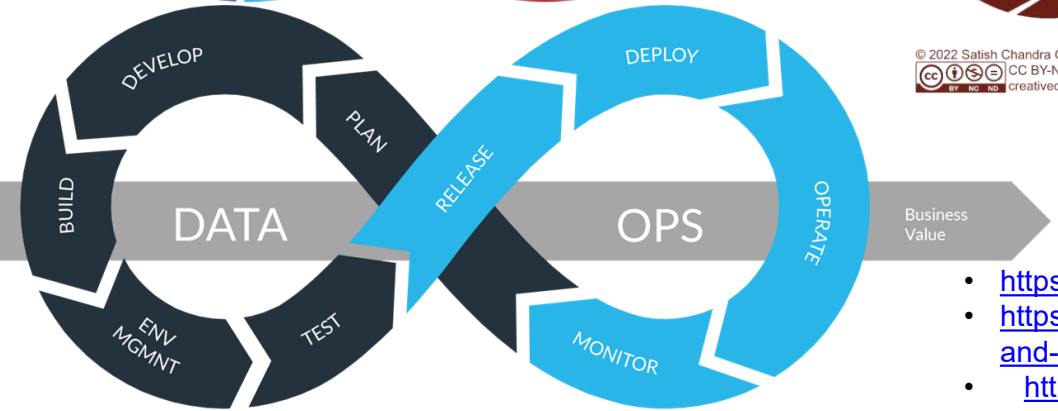
In GSBPM, there are some phases which are undertaken quickly and frequently – the Work Phases.

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

Which X_Ops you look at might not matter

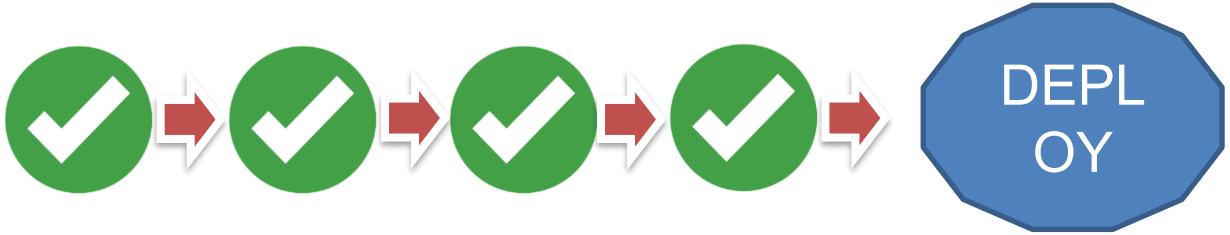






- https://en.wikipedia.org/wiki/DevOps_toolchain
- https://www.snowflake.com/blog/the-rise-of-dataops-governance-and-agility-with-truedataops/
- https://www.ml4devs.com/articles/mlops-machine-learning-life-cycle/

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.
- Integrated teams: statisticians, subject matter experts and developers

Be Reproducible

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

Building pipelines for deployments is crucial

- What tools do you need?
  - Code based statistical software, e.g. R 😀, Python 🙂, SAS 😐
  - Version control -> probably some GIT
  - CI/CD Pipeline tool, e.g. Gitlab CI, Github Actions, Jenkins, Airflow, …
  - Deployment facilities - Storage to deploy artifacts, e.g. R packages to a CRAN-like mirror, web applications to a web server, AI models as APIs, object storage for general purpose, etc.

Our workflows currently depend on R/RStudio

- Posit Workbench -> RStudio IDE
- Posit Connect -> Hosting APIs, shiny Apps, (scheduled) Reports
- Jenkins connects these two

Pipelines can have many different outputs

- APIs:
  - Models, e.g. a text-to-code classification model
  - …
- Building blocks to be used by several other pipelines, e.g. R-packages
- Data Pipelines – “Classical” ETL kind of process step
- “Products”: Visualizations, Dashboards, Reports

OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

THANK YOU. JUST TWO MORE THINGS.

Looking for specific modules? Look at the **Awesome list**

a ***community approach*** to knowledge:

How:

- Using the ***awesome concept*** on GitHub
- A ***public*** list which started ***simple*** and continues to ***grow***
- Clear and simple ***criteria***
- ***awesomeofficialstatistics.org***

Ten Bosch, van der Loo, Kowarik 2020 "The awesome list of official statistics software: 100… and counting"

Look at http://awesomeofficialstatistics.org

What is the awesome list?

Curated list of software for official statistics

awesome

www.awesomeofficialstatistics.org

Search or jump to... / Pulls Issues Marketplace Explore

SNStatComp / awesome-official-statistics-software

Unwatch 30 | Unstar 161 | Fork 41

Code | Issues | Pull requests 1 | Actions | Projects | Wiki | Security | Insights

Social interactions

master | Go to file | Add file | Code

Awesome official statistics software awesome

An awesome list of open source statistical software packages useful for creating and accessing official statistics.

An item on this list is awesome because

1. it is free, open source, and available for download and
2. it is confirmed to be used in the production of official statistics by at least one institute or it provides access to official statistics publications.

Criteria

We prefer packages that are easy to install and use, have at least one stable version, and are actively maintained. Contributions are welcome.

An awesome list of statistical software for creating and accessing official statistics

official-statistics gsbpm

Contributors 15

+ 4 contributors

Working together

Contributions

Awesome contributions are welcome, here are ways to do it:

- The GitHub way: send us a pull request to add directly to this list.
- Add an item to the issue tracker issue tracker. (you need a GH account)
- Send an e-mail to `mark dot vanderloo at gmail dot com` or `olav dot ten berch at gmail dot com` or tweet @markvdloo

License

This work is licensed under a Creative Commons Attribution 4.0 International License.

Open license

Join the Use of R in Official Statistics 2023

www.urosconf.org
- Call open until 8 Sept.

The 11th International Conference The Use of R in Official Statistics

uRos2023

Romanian National Institute of Statistics
Ecological University of Bucharest

12-14 December 2023