Data Literacy: need and competencies

7th National Meet&Greet of Swiss Medical Librarians

September 2020

UNIVERSITÄT BERN

Building a digital toolbox for scientific data handling

Dr. Nuria Plattner

Dr. Michael Horn

UNIVERSITY LIBRARY BERN
www.unibe.ch/ub/sciencelibrary

Outline

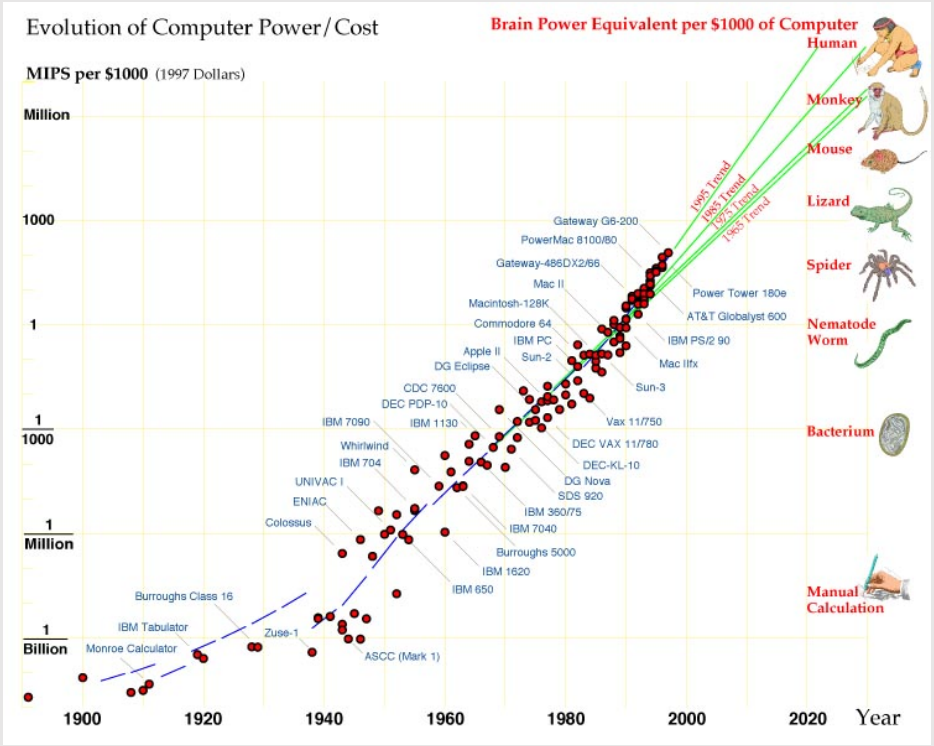
Building a digital toolbox for scientific data handling

UNIVERSITÄT BERN

- **Introduction: Science, data, increasing computer power and libraries**
- Digital toolbox: motivation and concepts
- Digital tool examples with Jupyter notebook demonstration
- Outlook: future directions and developments

Increase of computational ressources

Availability of computer power over time



- Computational resources have increased drastically in recent years
- The computing efficiency has also increased
- The easy availability of computer power is transforming society in various ways

H. Moravec "When will computer power match the human brain?"
Journal of Evolution and Technology 1, (1998)

Data, more data and big data

Handling general and scientific data

- The fast increase of computer usage and computer power produces ever larger amounts of data
- Large data collections may contain redundant or defective information
- Handling large datasets requires computational tools for automated data analysis

Symbolic picture: drowning in data

Transformation of Science

The impact of data and digitalization

- Digitalization is not only a revolution for libraries, but also for scientific research and education[1]
- In many research areas digitalization and the availability and collection of large datasets has transformed the research process
- Libraries can play an important, partially new role within this process

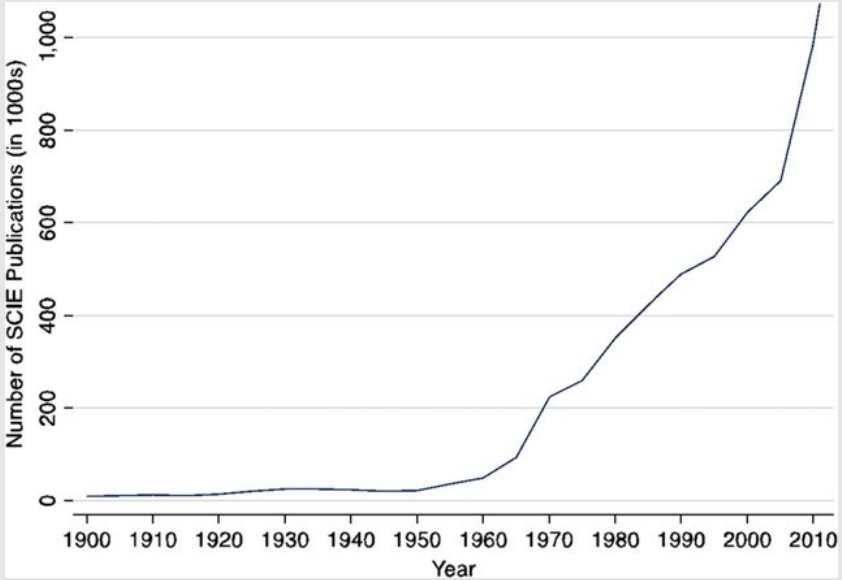
Symbolic picture: Digitalization (pixfuel.com)

[1] https://www.oecd.org/going-digital/digitalisation-of-STI-summary.pdf

Digital transformation of Science

Increasing amounts of data and articles

- Scientific Paper and data publication is easier, faster, and possible on more platforms
- Experimental data is easier to record digitally at high resolution, store and share
- Computer simulations and data analysis can handle and produce more data
- Efficient data handling required



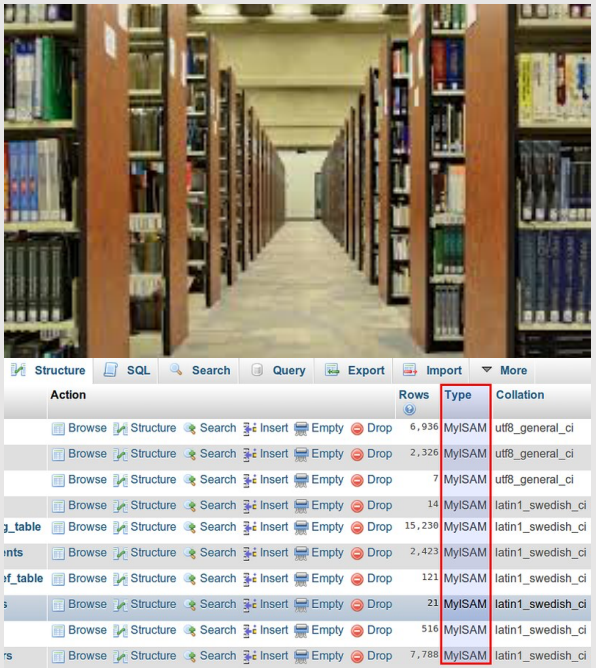
Increase of scientific publication output over time

@ https://academia.stackexchange.com/questions/126980/global-number-of-publications-over-time

The library as a data provider

Data handling required

- Traditional form of providing data: books, magazines
- Newer forms: E-Books, E-Papers, databases
- New trend: Research data sharing platforms
- In all cases adequate data handling is required
- Help with handling data partially included in library services

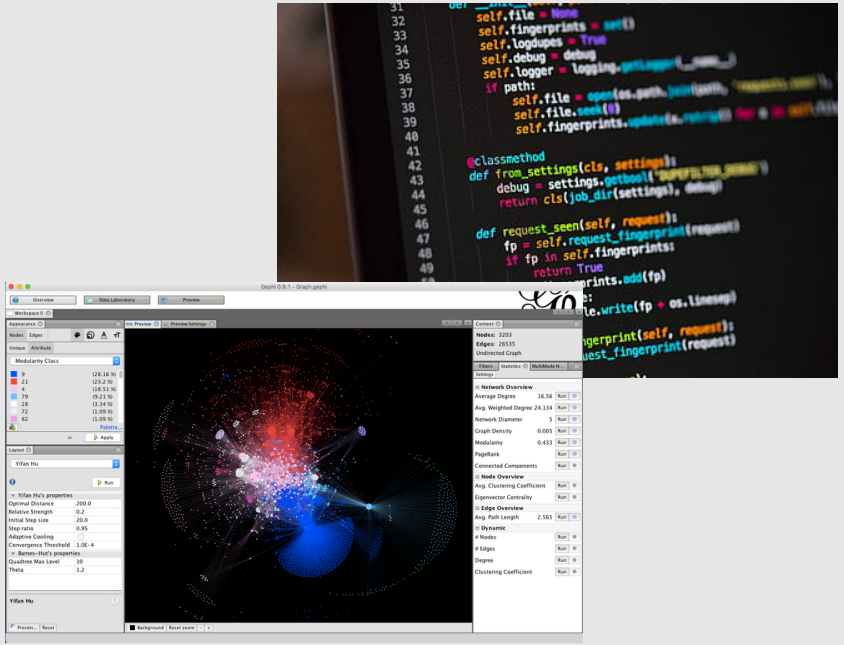


Bookshelfs and Database

Data handling

Software and code

- Specialized software for handling various types of data exists
- In some cases, easy to use software with graphical user interface is available and affordable
- Larger amounts of data or more specialized analyses require automation
- For this task, simple computer code building blocks can be used

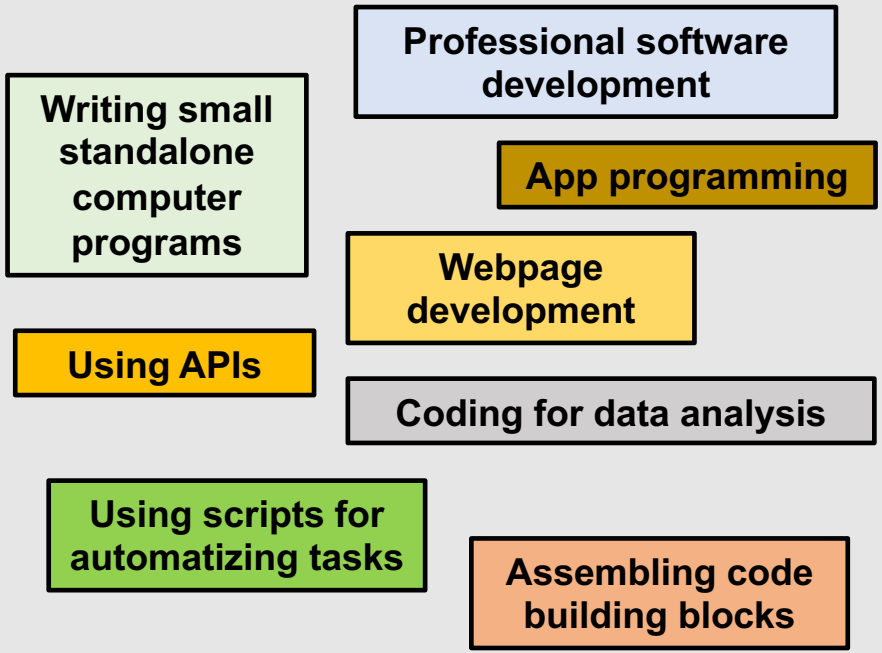


Code vs. Software

Using software vs. using code

How difficult is it to use code?

- Psychological and technical barriers: high initial barrier for using code instead of software
- Easier to transfer knowledge to new tasks if code is used
- Better technical understanding of data handling process, data structures formatting issues etc.
- Code easier to document and check reproducibility



What does coding mean?

Using software vs. using code

When is automation required?

- Large dataset difficult to handle without automation
- Coding more efficient if effort/time needed to code is smaller than time needed to manually (by clicking, copy-pasting…) edit data
- Effort depends on initial skill level; initial learning curve shallow

Symbolic picture learning curves; source Wikipedia

Outline

Building a digital toolbox for scientific data handling

UNIVERSITÄT BERN

- Introduction: Science, data, increasing computer power and libraries
- **Digital toolbox: motivation and concepts**
- Digital tool examples with Jupyter notebook demonstration
- Outlook: future directions and developments

Digital toolbox concept

Code building blocks

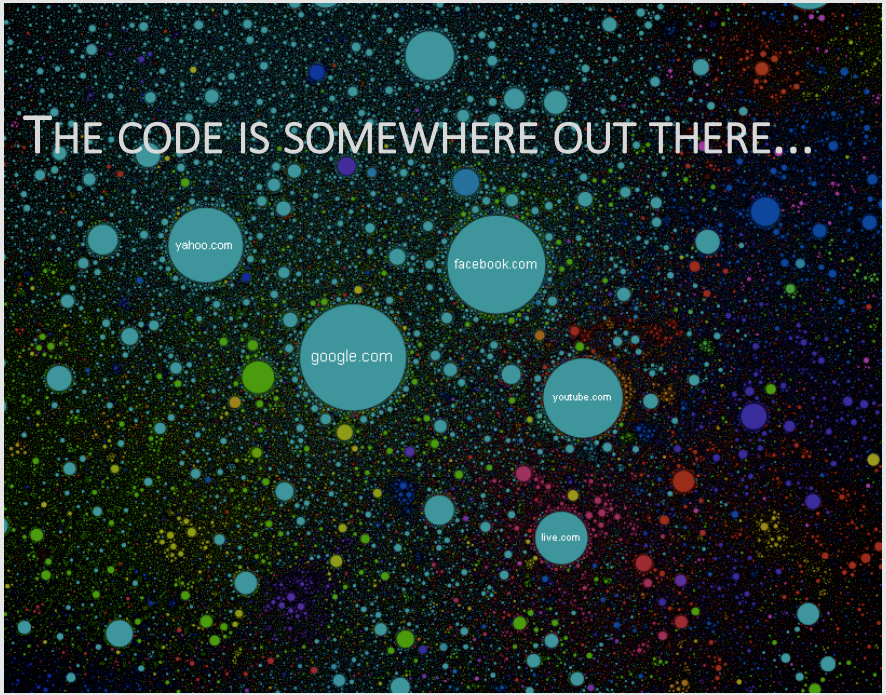
- Assemble code from simple building blocks
- Generate examples for various tasks
- Document code examples
- Demonstrate how to use code libraries
- Use simple representative problems for demonstration

Symbolic picture: building blocks pikrepo.com

Digital toolbox concept

Code available on the internet

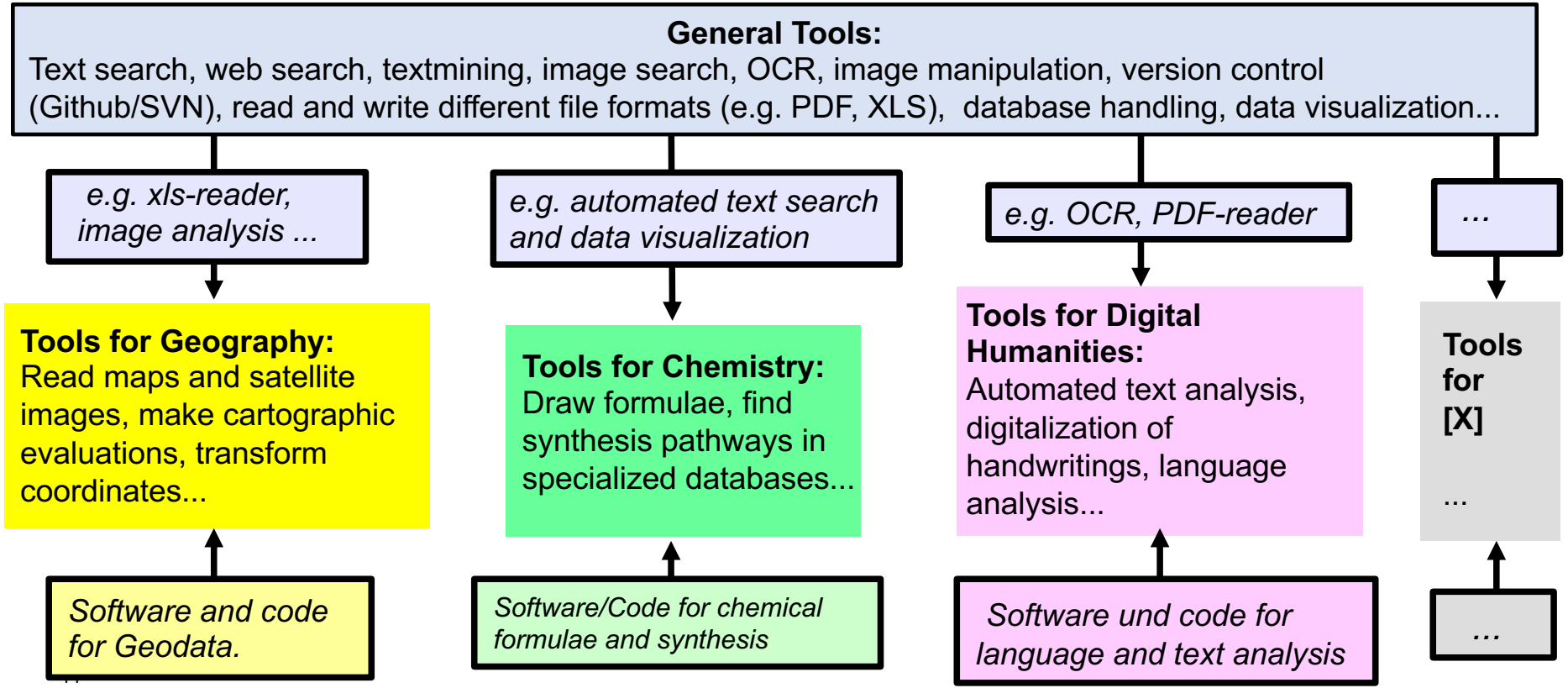
- Coding needed for handling data
- Huge amount of code libraries, API and code building blocks available on the internet
- ToDo: basic coding skills for using this building blocks
- ToDo: test and assemble available building blocks for specific data handling tasks



Background: internet-map.net

Digital toolbox

Schematic overview



Different needs for different research areas

Familiarity with coding unevenly distributed

- Different needs depending on how familiar researchers and students of different areas are with specialized software and code
- Larges need will be for areas with large amounts of data but little exposure to coding
- Science becomes more data-intensive in general

Example: scientists drowning in {COVID}-19 papers[1]

[1] J. Brainard, Science, May 2020
https://doi.org/10.1126/science.abc7839

Python vs. other programming languages

Versatility for scientific data handling

Python programming language

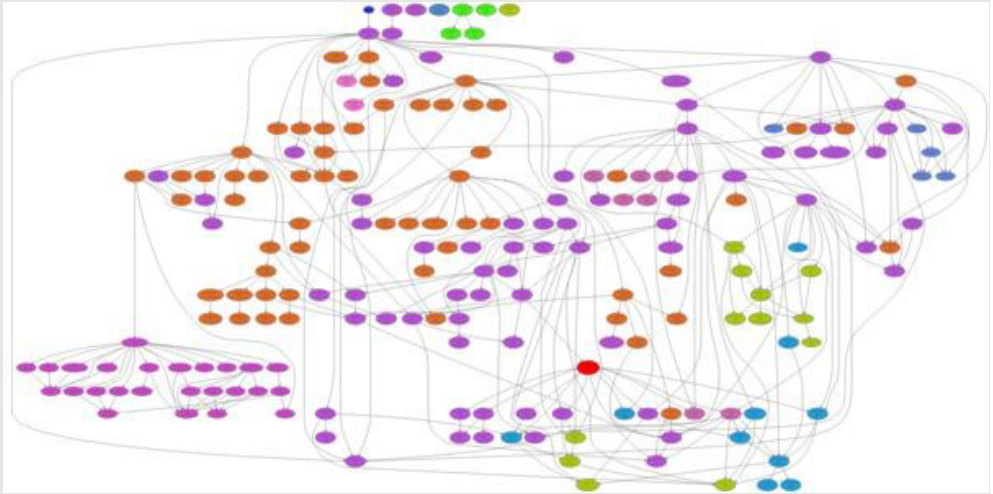
- Interpreted programming language with modules for numerical operations available precompiled in C++
- Code building blocks: python modules for various tasks available and easy to combine
- Many tools specialized for various types of scientific data exist
- Requirement: package manager in order to combine modules and control versions and dependencies

Python package managers

Managing dependencies and code versions

UNIVERSITÄT BERN

- Code dependencies and code versions need to be managed
- Dependencies and required versions quickly grow into a complex network



Dependency graph; more information:

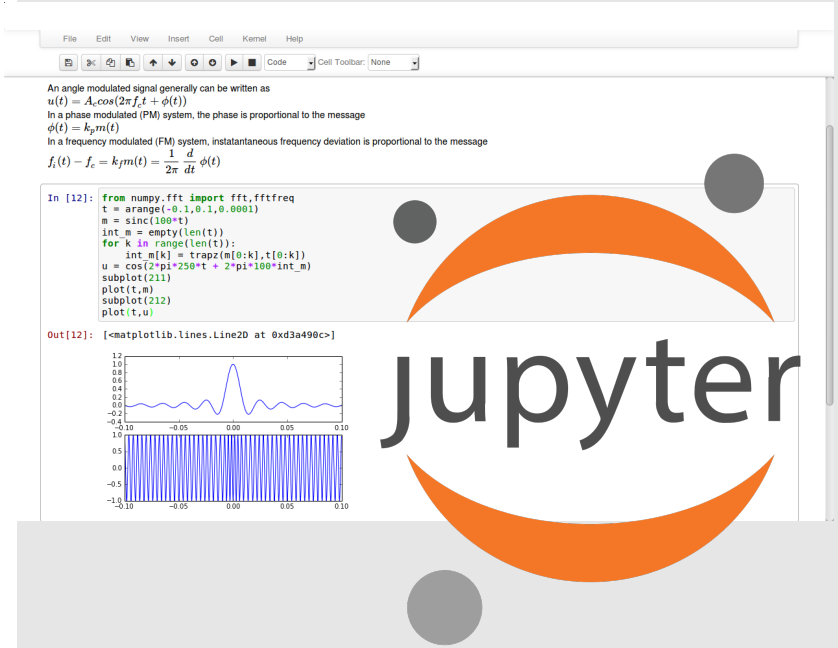
https://www.freecodecamp.org/news/code-dependencies-are-the-devil-35ed28b556d/

ANACONDA®

Python package managers, useful link:
https://docs.conda.io/en/latest/miniconda.html

Jupyter notebooks and Python scripts

Combine code, graphics and documentation



@ https://jupyter.org

- For building the toolbox, code, documentation and application examples are required
- Jupyter notebooks provide a browser-based platform where code can be executed blockwise
- Documentation can be inserted between code blocks
- Graphical output can be shown directly in the notebook

Outline

Building a digital toolbox for scientific data handling

UNIVERSITÄT BERN

- Introduction: Science, data, increasing computer power and libraries
- Digital toolbox: motivation and concepts
- **Digital tool examples with Jupyter notebook demonstration**
- Outlook: future directions and developments

Digital tool examples: Geotools

Visualization of data on map

UNIVERSITÄT BERN

@ https://github.com/ubnpl/pytools/tree/master/geo_data

Digital tool examples: Data visualization

Graphical representation of statistical data

UNIVERSITÄT BERN

https://github.com/ubnpl/pytools/tree/master/data_visualization

Outline

Building a digital toolbox for scientific data handling

- Introduction: Science, data, increasing computer power and libraries
- Digital toolbox: motivation and concepts
- Digital tool examples with Jupyter notebook demonstration
- **Outlook: future directions and developments**

Initial repository as starting point

Growing collection of digital tools

UNIVERSITÄT BERN

Preliminary Github repository with initial collection of tools: https://github.com/ubnpl/pytools

Work in progress

Can be used as starting point to learn how to use simple code building blocks

New tools can be suggested or added directly

Collection will grow over time and hopefully more contributors will join

Digital tool extensions

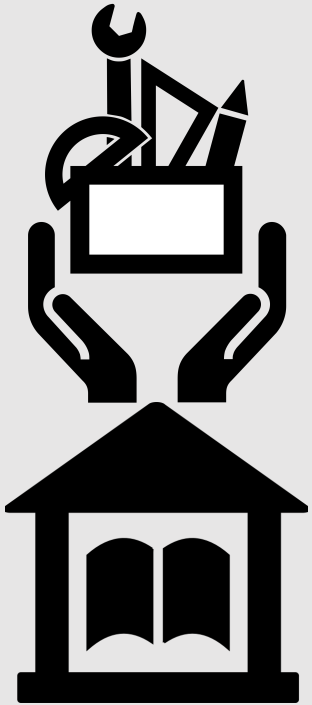
Include more programming languages

Digital tools in R under construction by Kathi Woitas

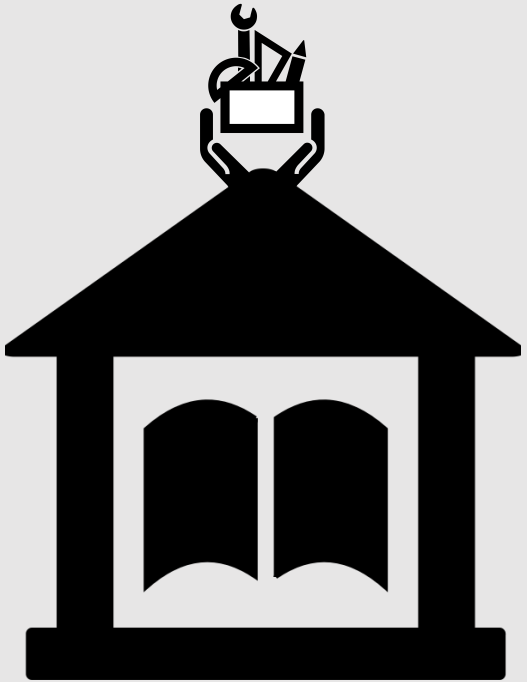
https://github.com/k-woitas/rtools

Toolbox development and adaptation

Adaptation to subject-specific needs



- Depending on subject-specific need, it is probable that more development effort will be invested in certain topics
- Complementary tools to software currently used in different areas
- General tools can also be used for library purposes



Toolbox potential advantages

Potential for more open-source tools

- Specialized software often commercialized
- In some cases simple code can be a viable alternative
- In addition, open-source replacements for commercial software can be advertised
- Complements open science strategies

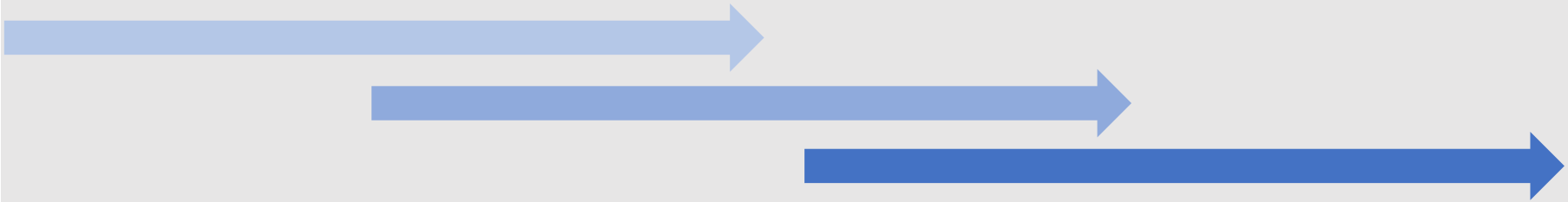
Open science with open source tools

Outlook - future directions

Development and usage of toolbox

UNIVERSITÄT BERN

- Initial phase: collecting tools and getting familiar with tools for different areas, find more people interested in participating
- 2nd phase: Introduction to tools within the scope of existing courses, e.g. coffee lectures or scientific information search courses
- 3rd phase: specific tool development upon request and workshop for data handling in specific areas



Discussion / Conclusions

Building a digital Toolbox for scientific data handling

UNIVERSITÄT BERN

- *Increasing amounts of scientific data require automation of data handling*
- *Small code building blocks can be assembled in order to carry out various data handling tasks*
- *Initial toolbox under construction in Python using Jupyter notebooks*
- *More tools will be added over time and adapted to subject-specific needs*

Thanks for your attention

Questions?

UNIVERSITÄT BERN

Building a digital toolbox for scientific data handling

Dr. Nuria Plattner
Dr. Michael Horn

UNIVERSITY LIBRARY BERN
www.unibe.ch/ub/sciencelibrary