

Data Literacy: need and competencies

7th National Meet&Greet of Swiss Medical Librarians

September 2020

u^b

b
**UNIVERSITÄT
BERN**

Building a digital toolbox for scientific data handling

Dr. Nuria Plattner

Dr. Michael Horn

Outline

u^b

b
UNIVERSITÄT
BERN

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

Increase of computational resources

u^b

Availability of computer power over time

b
UNIVERSITÄT
BERN

Evolution of Computer Power/Cost

MIPS per \$1000 (1997 Dollars)

Million

1000

1

1/1000

1/Million

1/Billion

1900

1920

1940

1960

1980

2000

2020

Year

Brain Power Equivalent per \$1000 of Computer

Human



Monkey



Mouse



Lizard



Spider



Nematode



Worm



Bacterium



Manual Calculation

- 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

u^b

Handling general and scientific data

b
UNIVERSITÄT
BERN

- 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

u^b

The impact of data and digitalization

b
UNIVERSITÄT
BERN

- 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>

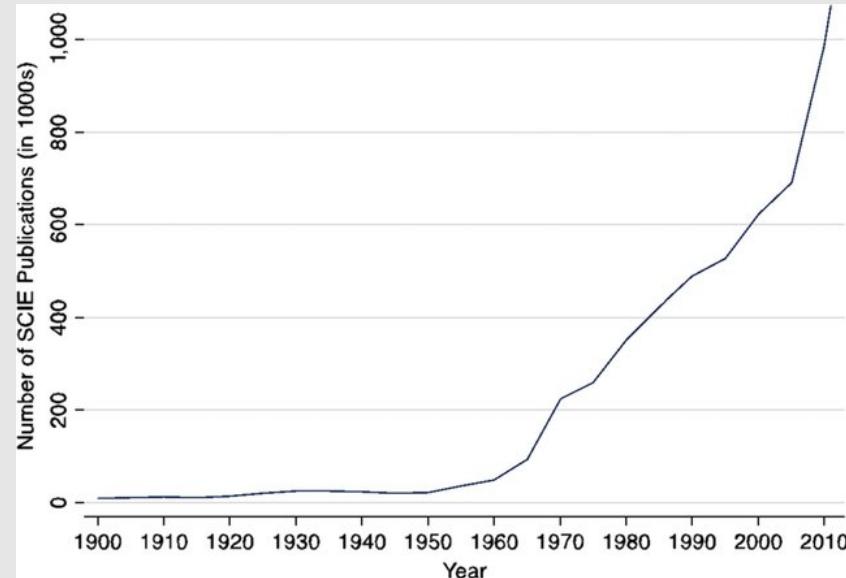
Digital transformation of Science

u^b

Increasing amounts of data and articles

b
UNIVERSITÄT
BERN

- 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

u^b

b
UNIVERSITÄT
BERN

- 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



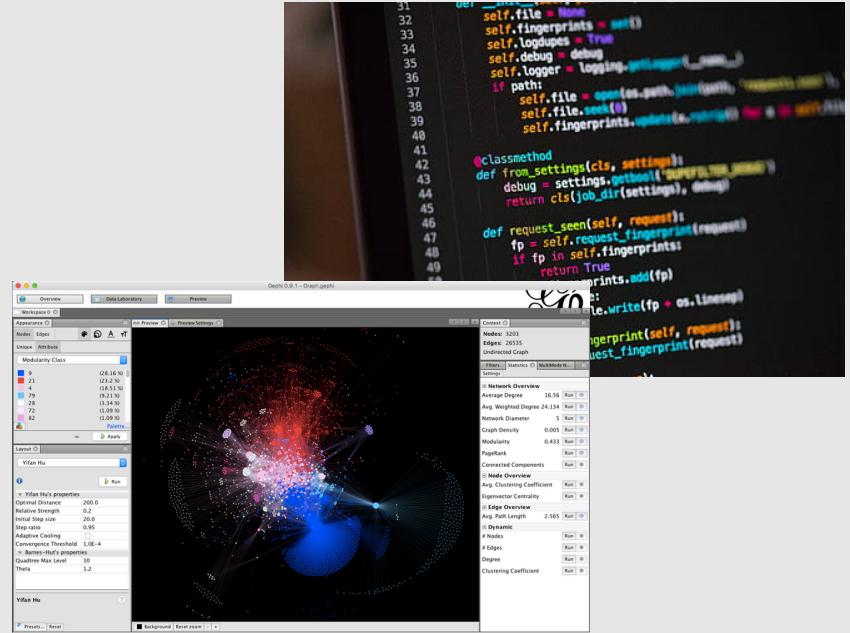
Action	Rows	Type	Collation
Browse	6,936	MyISAM	utf8_general_ci
Browse	2,326	MyISAM	utf8_general_ci
Browse	7	MyISAM	utf8_general_ci
Browse	14	MyISAM	latin1_swedish_ci
g_table	15,230	MyISAM	latin1_swedish_ci
nts	2,423	MyISAM	latin1_swedish_ci
of_table	121	MyISAM	latin1_swedish_ci
s	21	MyISAM	latin1_swedish_ci
rs	516	MyISAM	latin1_swedish_ci
rs	7,788	MyISAM	latin1_swedish_ci

Bookshelves 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

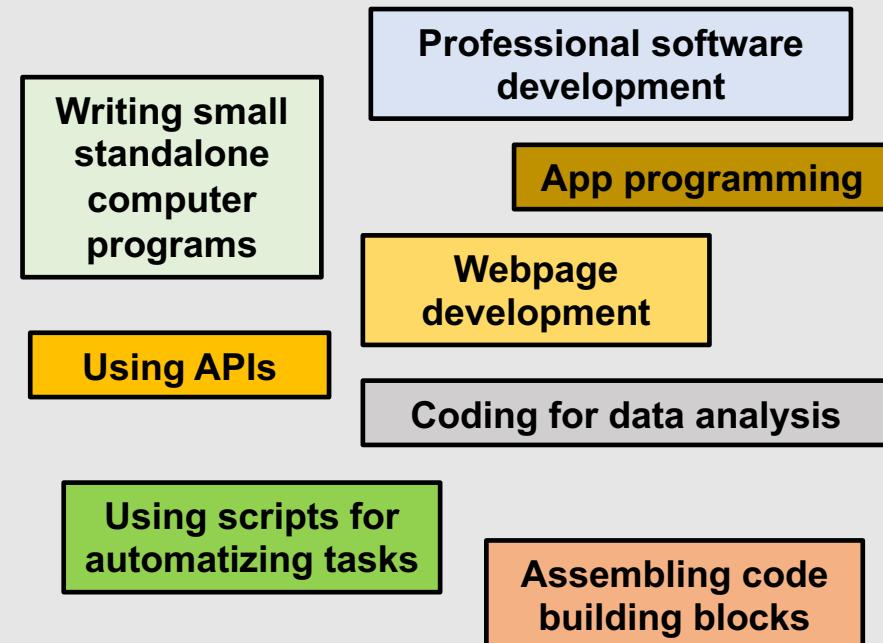


Code vs. Software

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?

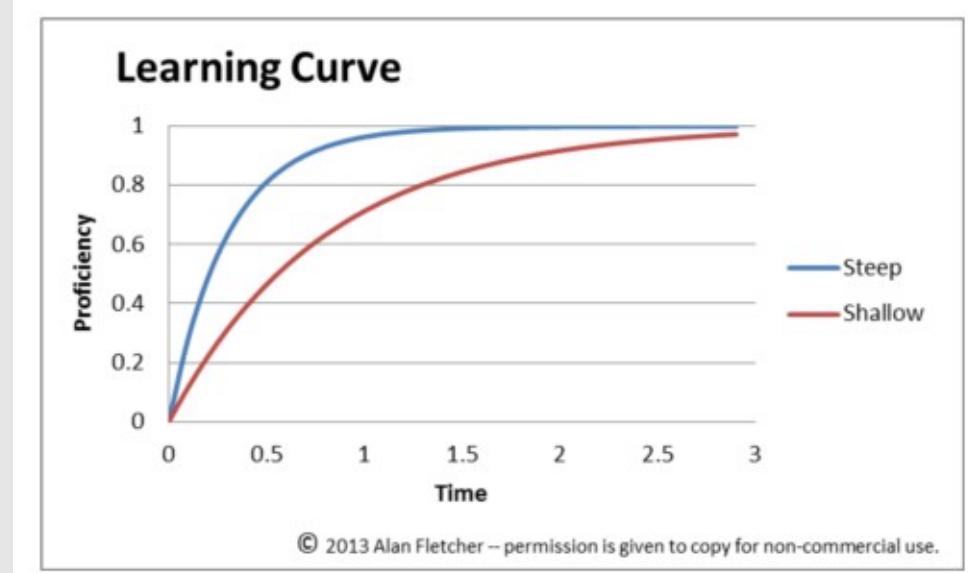
Using software vs. using code

When is automation required?

u^b

b
UNIVERSITÄT
BERN

- 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

u^b

b
UNIVERSITÄT
BERN

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

Digital toolbox concept

Code building blocks

u^b

b
UNIVERSITÄT
BERN

- 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*

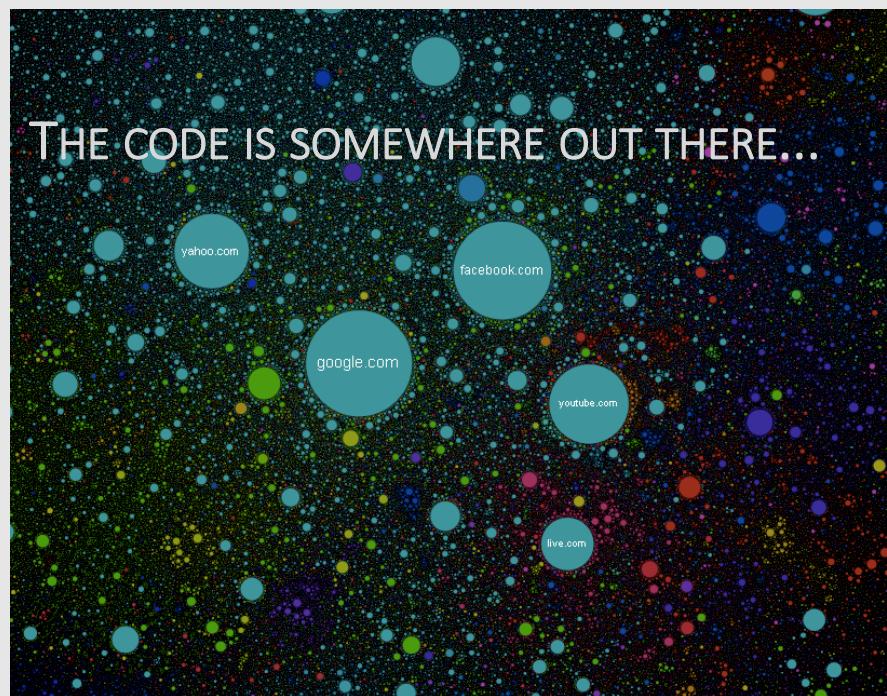
Digital toolbox concept

u^b

Code available on the internet

b
UNIVERSITÄT
BERN

- 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

Schematic overview

General Tools:

Text search, web search, textmining, image search, OCR, image manipulation, version control (Github/SVN), read and write different file formats (e.g. PDF, XLS), database handling, data visualization...

e.g. *xls-reader, image analysis ...*

e.g. *automated text search and data visualization*

e.g. *OCR, PDF-reader*

...

Tools for Geography:
Read maps and satellite images, make cartographic evaluations, transform coordinates...

Software and code for Geodata.

Tools for Chemistry:
Draw formulae, find synthesis pathways in specialized databases...

Software/Code for chemical formulae and synthesis

Tools for Digital Humanities:

Automated text analysis, digitalization of handwritings, language analysis...

Software und code for language and text analysis

Tools for [X]
...

...

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
- Largest 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>](https://doi.org/10.1126/science.abc7839)

Python vs. other programming languages

u^b

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

b
UNIVERSITÄT
BERN

Python package managers

Managing dependencies and code versions

u^b

b
UNIVERSITÄT
BERN

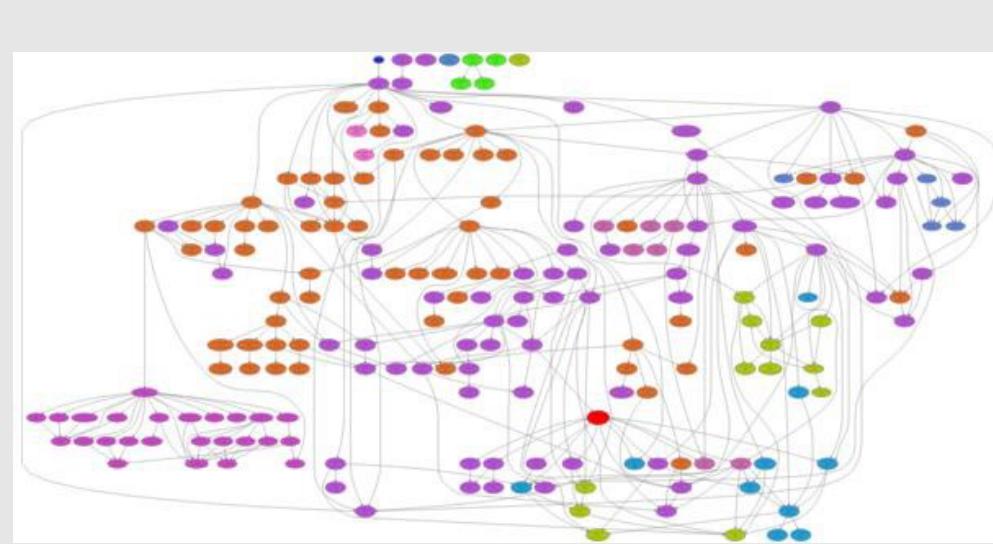
- Code dependencies and code versions need to be managed
- Dependencies and required versions quickly grow into a complex network



ANACONDA®

Python package managers, useful link:

<https://docs.conda.io/en/latest/miniconda.html>

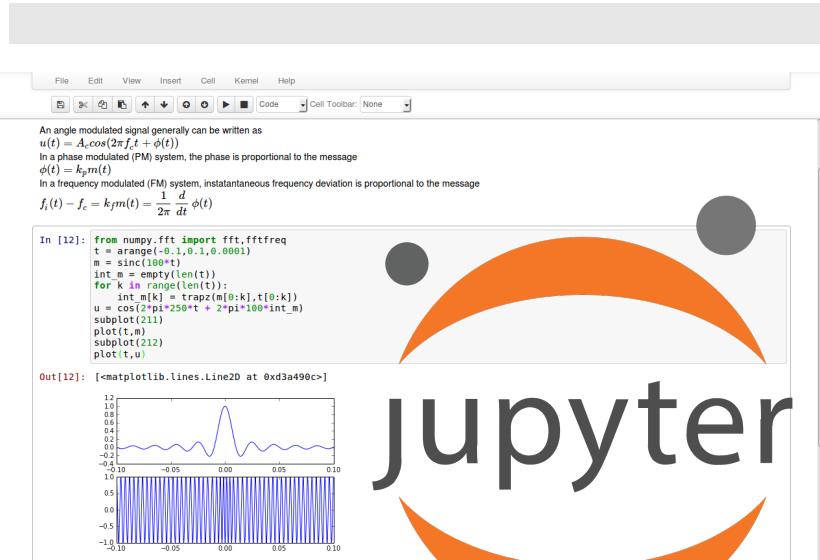


Dependency graph; more information:

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

Jupyter notebooks and Python scripts

Combine code, graphics and documentation



A screenshot of a Jupyter notebook interface. The menu bar includes File, Edit, View, Insert, Cell, Kernel, and Help. The toolbar includes icons for file operations like Open, Save, and Run, along with a Cell Toolbar dropdown set to None.

The notebook contains the following text and code:

```
File Edit View Insert Cell Kernel Help
File Edit View Insert Cell Kernel Help
A angle modulated signal generally can be written as
u(t) = Accos(2πfct + φ(t))
In a phase modulated (PM) system, the phase is proportional to the message
φ(t) = kpm(t)
In a frequency modulated (FM) system, instantaneous frequency deviation is proportional to the message
fi(t) - fc = kfm(t) =  $\frac{1}{2\pi} \frac{d}{dt} \phi(t)$ 
```

```
In [12]: from numpy.fft import fft,fftfreq
arrange(-0.1,0.1,0.0001)
m = arange(100*t)
int_m = empty(len(t))
for k in range(len(t)):
    int_m[k] = trapz(m[0:k],t[0:k])
u = cos(2*pi*250*t + 2*pi*100*int_m)
subplot(111)
plot(t,m)
subplot(212)
plot(t,u)
```

```
Out[12]: <matplotlib.lines.Line2D at 0xd3a490c>
```

Two plots are shown: a line plot of the message signal m over time t, and a line plot of the amplitude u over time t. Below the plots is a spectrogram showing power spectral density versus time and frequency.

- 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



<https://jupyter.org>

Outline

u^b

b
UNIVERSITÄT
BERN

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

Digital tool examples: Geotools

u^b

Visualization of data on map

b
UNIVERSITÄT
BERN



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

Digital tool examples: Data visualization

Graphical representation of statistical data



b
UNIVERSITÄT
BERN



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

Outline

u^b

b
UNIVERSITÄT
BERN

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

u^b

Growing collection of digital tools

b
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



b
UNIVERSITÄT
BERN



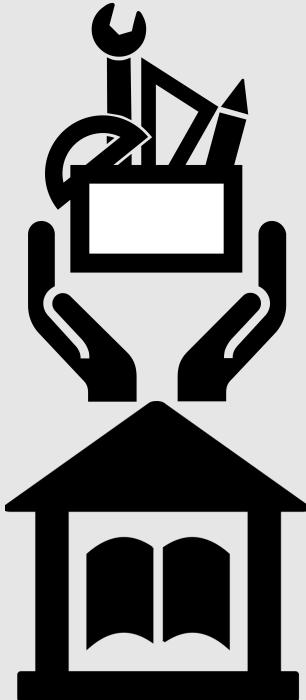
Digital tools in R under construction by Kathi Woitas



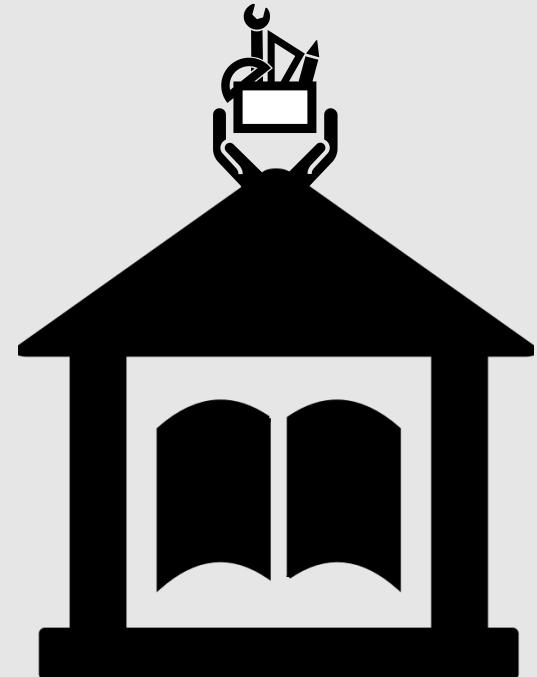
<https://github.com/k-woitas/rtools>

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

u^b

Potential for more open-source tools

b
UNIVERSITÄT
BERN

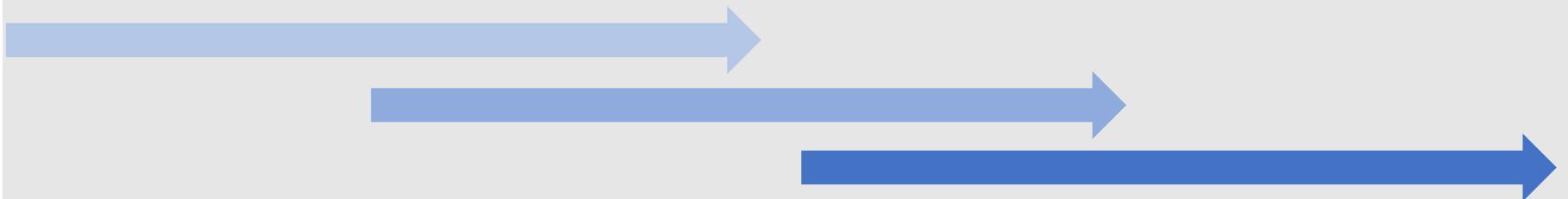
- Specialized software often commercialized
- In some cases simple code can be a viable alternative
- In addition, open-source replacements for commercial software can be advertised
- Complements open science strategies



Open science with open source tools

Development and usage of toolbox

- 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

u^b

Building a digital Toolbox for scientific data handling

b
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?

u^b

b
UNIVERSITÄT
BERN

Building a digital toolbox for scientific data handling

Dr. Nuria Plattner

Dr. Michael Horn

UNIVERSITY LIBRARY BERN

www.unibe.ch/ub/scielibrary