Herausforderungen durch wissenschaftliche Software Entwicklung

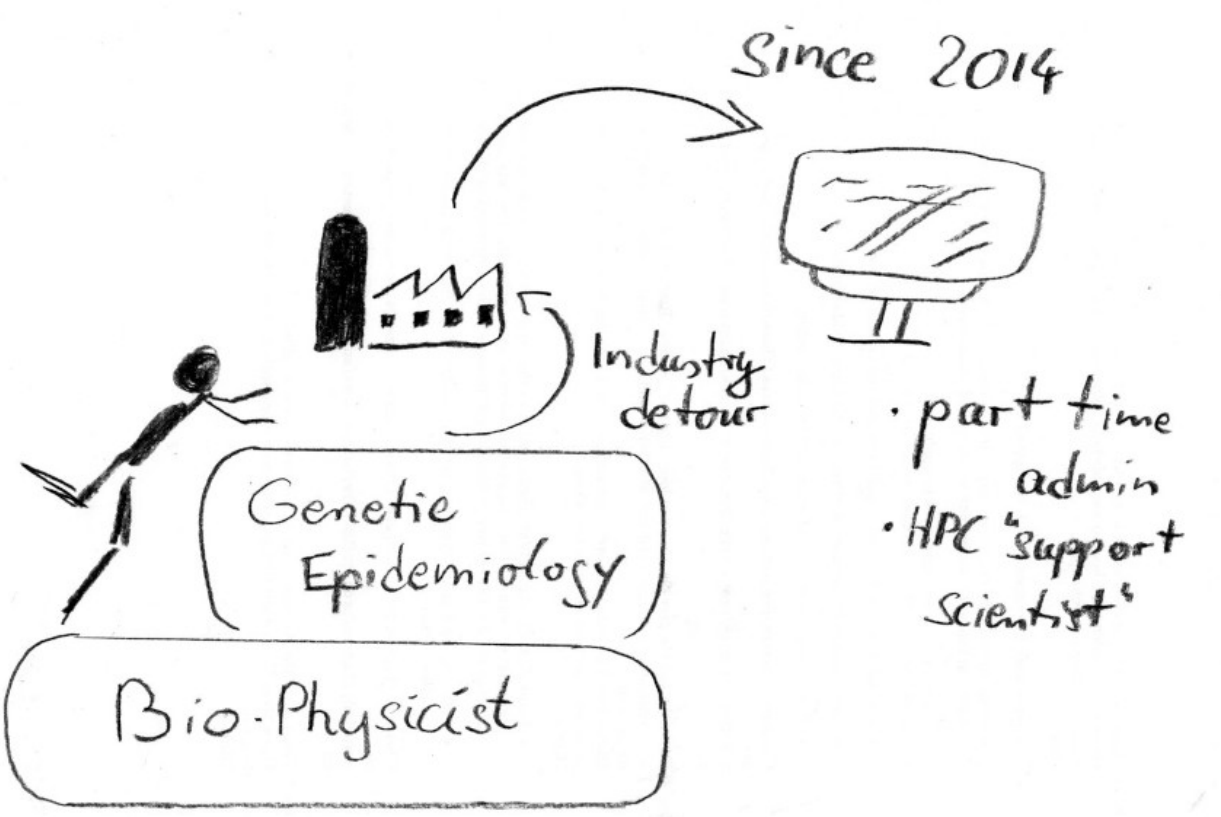
Eine persönliche Einschätzung

Christian Meesters

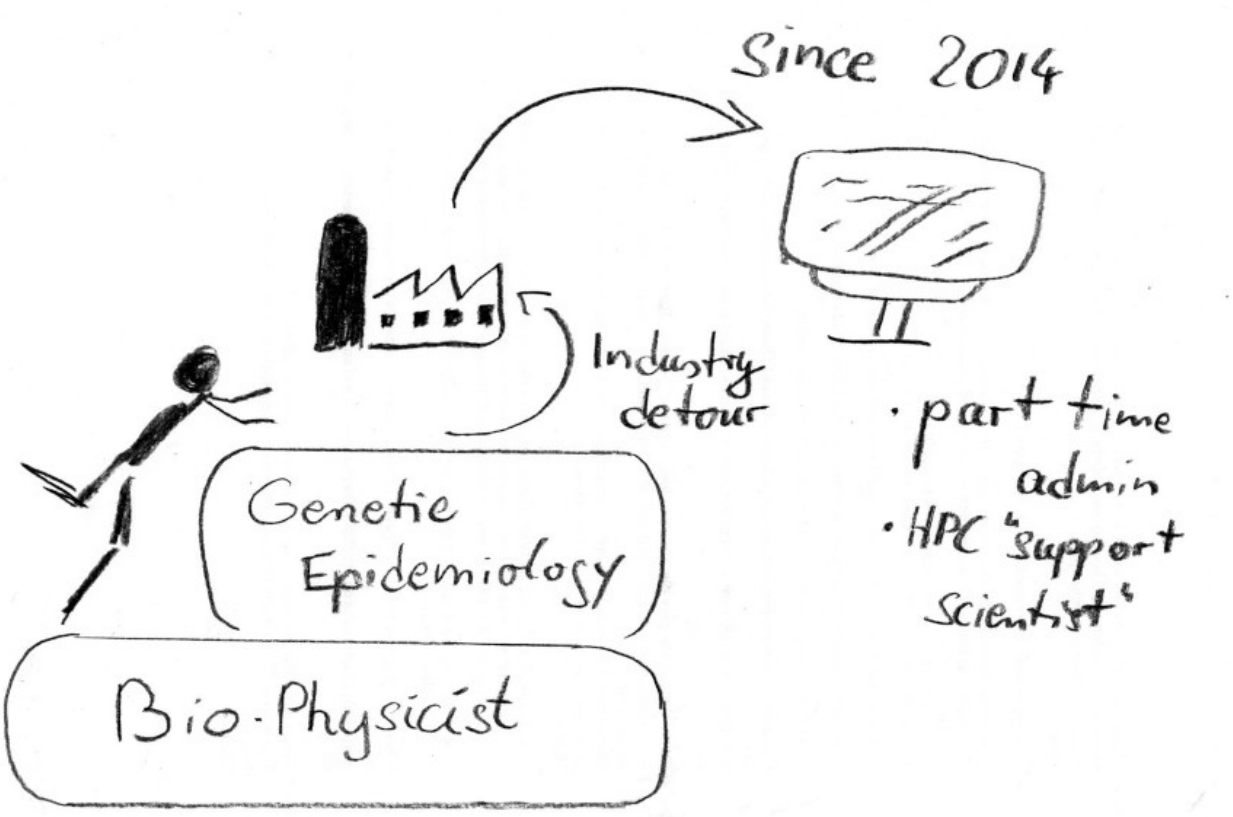
4. Oktober 2022

CC BY NC

Was qualifiziert den Sprecher über FLOSS zu sprechen?



Was qualifiziert den Sprecher über FLOSS zu sprechen?



Außerdem Scienceblogger:

Ein Blog über

- Wissenschaft
- Wissenschaftspolitik
- Software und Wissenschaft
- Gott und die Welt

Bedeutung von FLOSS in der Wissenschaft I

Wir haben bereits Tonnen von Material zur Einschätzung der Bedeutung von FLOSS für wissenschaftliches Rechnen: Blog für Facharbeit für Blog für Facharbeit für Meinungsartikel in "Nature" für Facharbeit für Facharbeit für ...

Bedeutung von FLOSS in der Wissenschaft I

Wir haben bereits Tonnen von Material zur Einschätzung der Bedeutung von FLOSS für wissenschaftliches Rechnen: Blog für Facharbeit für Blog für Facharbeit für Meinungsartikel in "Nature" für Facharbeit für Facharbeit für ...

Ergo:

Viele Leute haben sich schon Gedanken über diese Frage gemacht!

Bedeutung von FLOSS in der Wissenschaft I

Wir haben bereits Tonnen von Material zur Einschätzung der Bedeutung von FLOSS für wissenschaftliches Rechnen: Blog für Facharbeit für Blog für Facharbeit für Meinungsartikel in "Nature" für Facharbeit für Facharbeit für ...

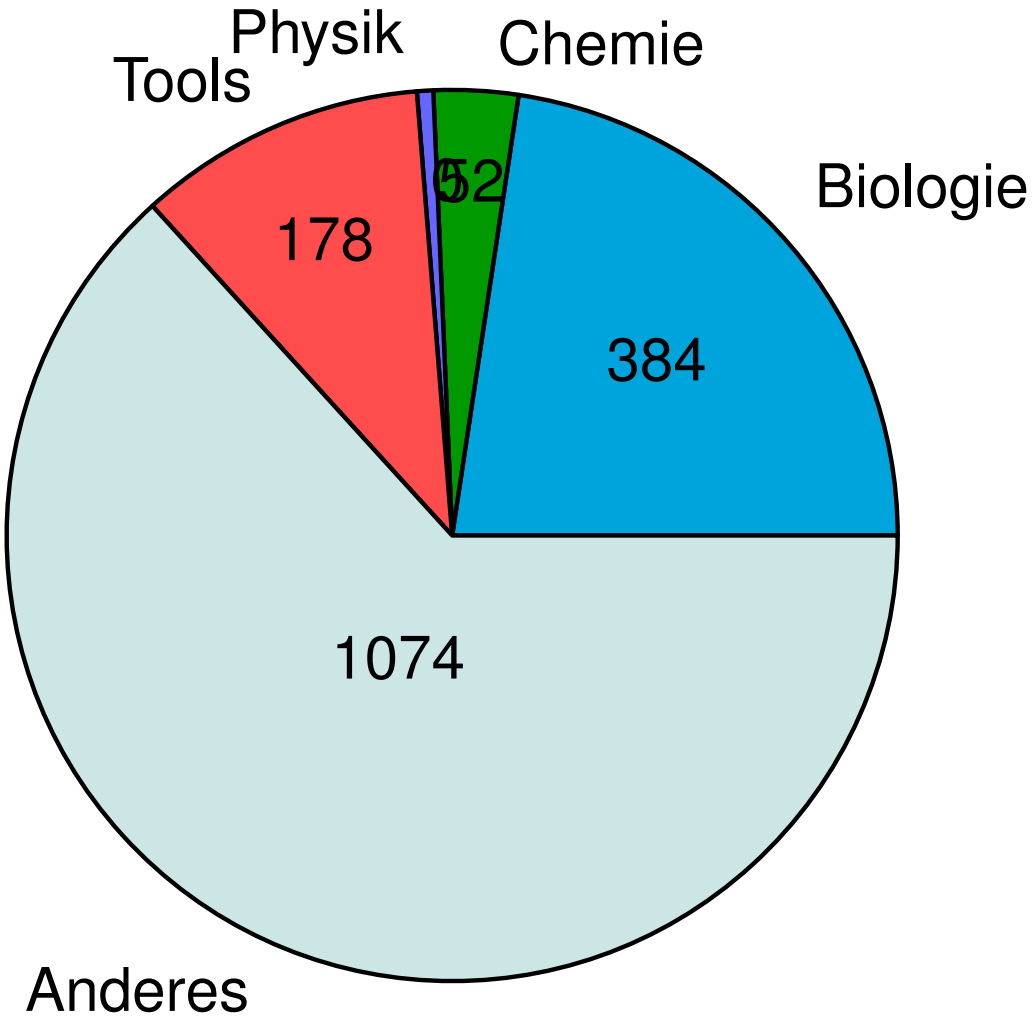
Ergo:

Viele Leute haben sich schon Gedanken über diese Frage gemacht!

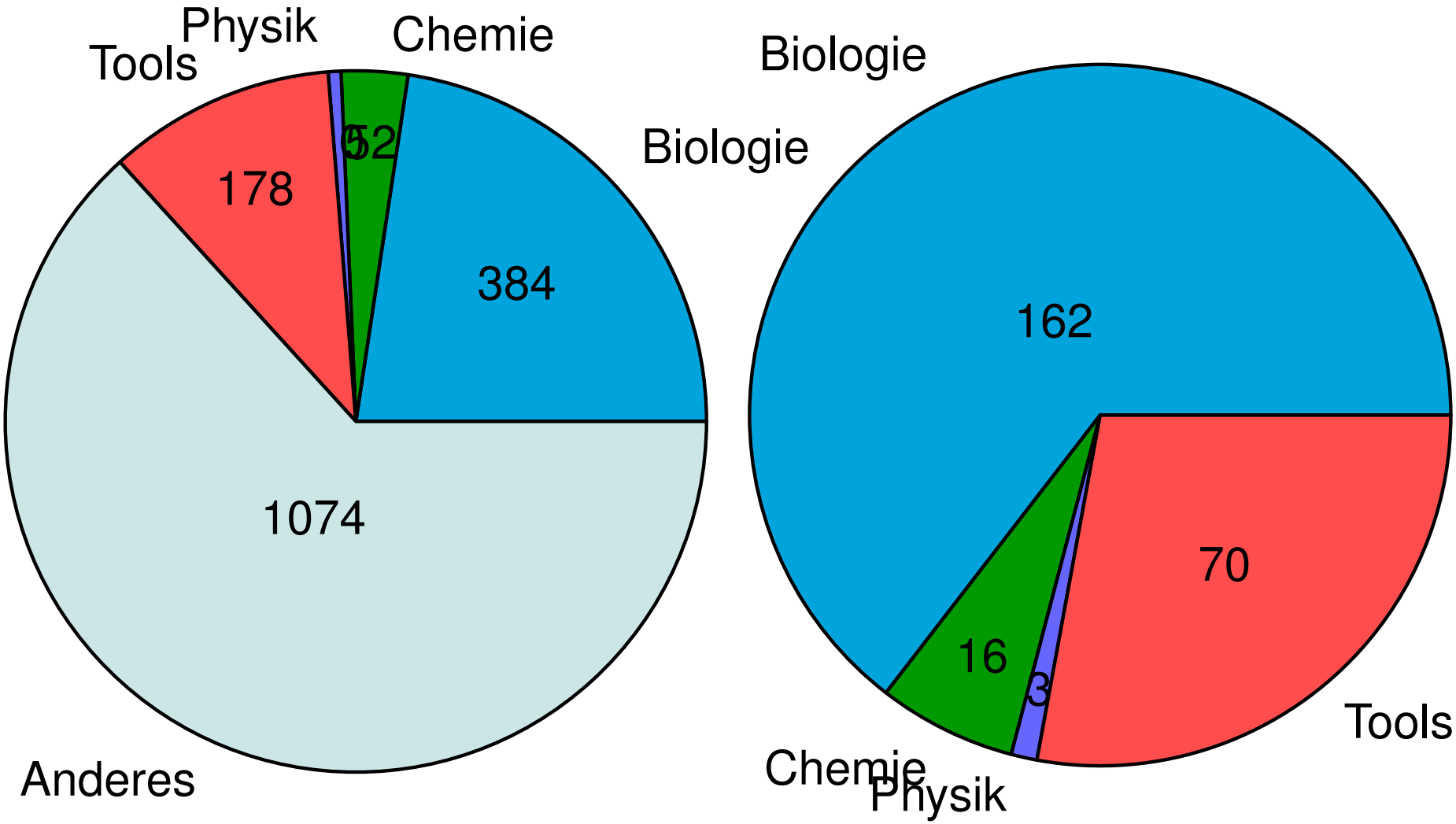
Systematisches Material fehlt oft

Bei Interesse an einer Facharbeit, meine Adresse (`meesters@uni-mainz.de`) weitergeben.

Bedeutung von FLOSS in der Wissenschaft II



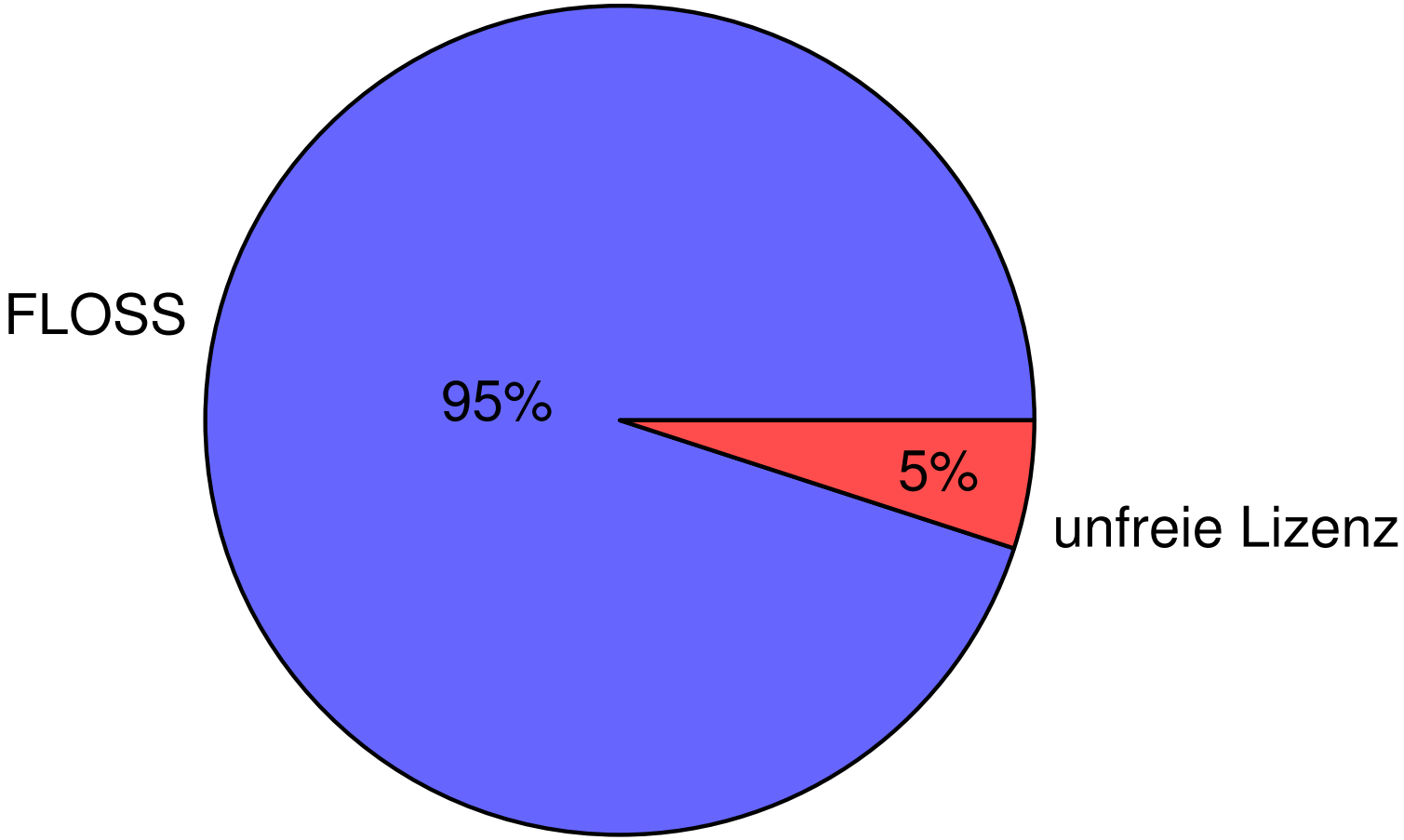
Bedeutung von FLOSS in der Wissenschaft II



† Zahlen hängen von Größe und Heterogenität der Nutzerbasis ab. Physiker bauen Software oft selbst. – Fun fact: Physik-Code bedingt mehr als 85% der CPU-Stunden

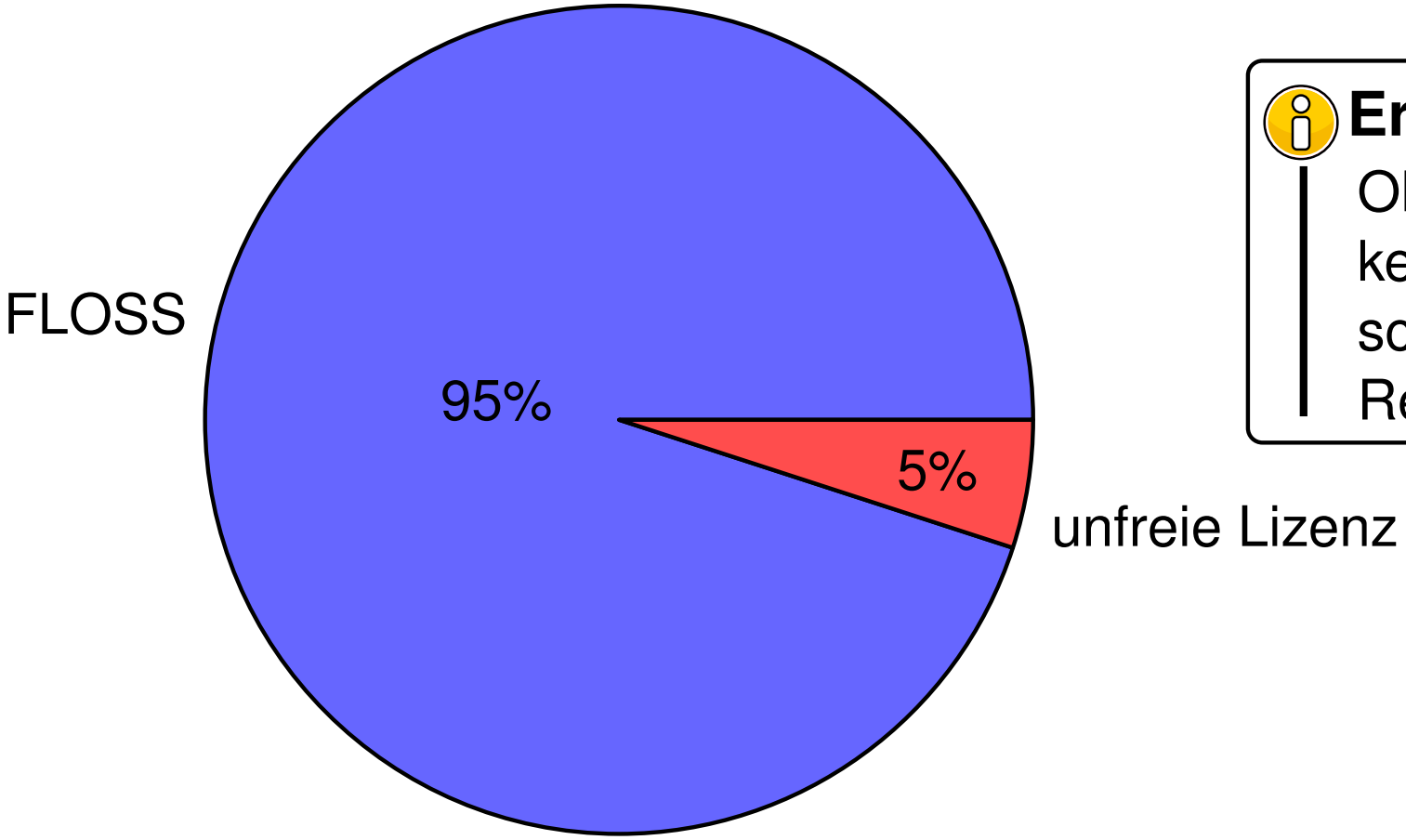
Bedeutung von FLOSS in der Wissenschaft - III

Und der FLOSS-Anteil?†



Bedeutung von FLOSS in der Wissenschaft - III

Und der FLOSS-Anteil?†



Ergo:
Ohne FLOSS kein wissenschaftliches Rechnen.

† Zahlen mögen von Einrichtung zu Einrichtung unterschiedlich ein - die Größenordnung wird im Mittel gleich bleiben.

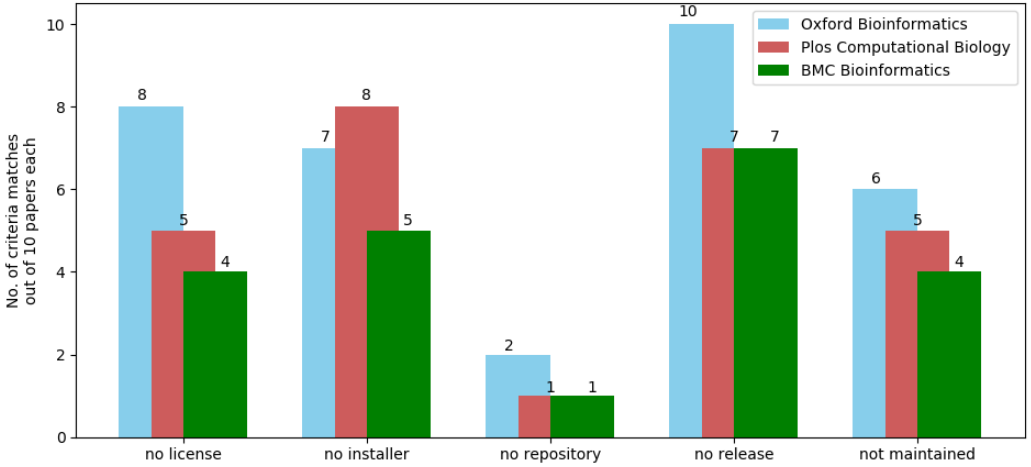
Wie steht es um die Qualität?

Hintergrund

Quantität ist nicht alles. Die meiste Software, die Verwendung findet ist gut gewartet. Aber:

- Gerade Bioinformatik-Software erfordert viel Arbeit der Paket-Maintainer.
- Die technologischen Entwicklungen erfordern immer neue Software.

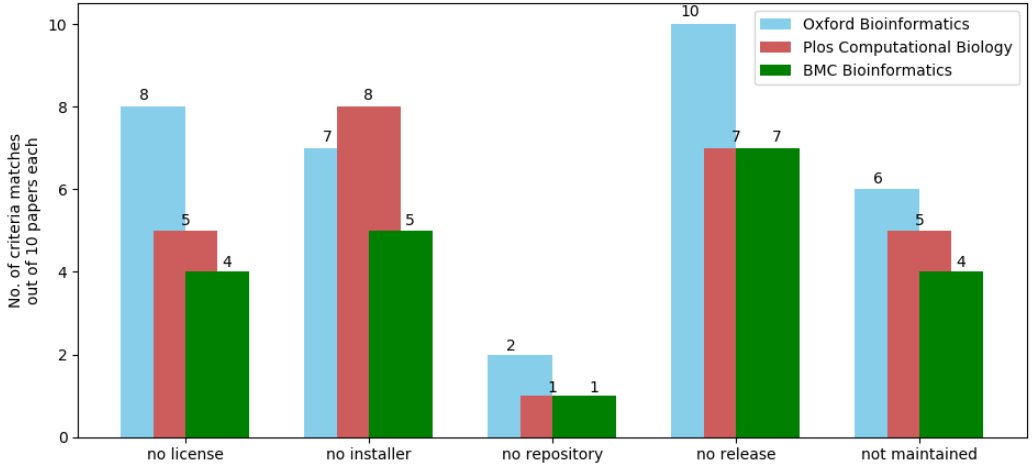
Probe aufs Exempel



aus einem Vergleichsartikel - je 10 Softwarepublikationen aus den angegebenen Zeitschriften wurden untersucht.

“nicht gewartet”, hier: letzte Änderung $> \frac{1}{2}$ Jahr her; teilweise ist ohnehin keine “Entwicklung” feststellbar, sondern nur *ein* Upload.

Probe aufs Exempel



aus einem Vergleichsartikel - je 10 Softwarepublikationen aus den angegebenen Zeitschriften wurden untersucht.

“nicht gewartet”, hier: letzte Änderung $> \frac{1}{2}$ Jahr her; teilweise ist ohnehin keine “Entwicklung” feststellbar, sondern nur *ein* Upload.

Ergo:

- systematische Evalution ist nie erfolgt,
- doch um die Qualität steht es wahrscheinlich nicht gut.

Konsequenz: Miserable Wissenschaft I

Aus einer Publikation über das Feststellen genetischer "Verschiebungen" (*genetic disorder*):

WeichenZhou Update README.md `90b206a` on Jun 24, 2021 **27** commits

Methylation.pipelines	updates April 27	2 years ago
RNA.design.pipelines	Updates Apr 24th	2 years ago
lib	Updates Apr 24th	2 years ago
.DS_Store	updates April 27	2 years ago
MEI_ONTreads_alignment.py	updated run command example	2 years ago
Nano-Pal.AluYa.sh	Feb 1st	2 years ago
Nano-Pal.AluYb.sh	Feb 1st	2 years ago
Nano-Pal.LINE.sh	Feb 1st	2 years ago
Nano-Pal.SVA_E.sh	Feb 1st	2 years ago
Nano-Pal.SVA_F.sh	Feb 1st	2 years ago
README.md	Update README.md	16 months ago

- nur 27 commits
- 5 shell Skripte die dasselbe auf verschiedenen Daten ausführen?!

Konsequenz: Miserable Wissenschaft I

Aus einer Publikation über das Feststellen genetischer "Verschiebungen" (*genetic disorder*):
Schaut man in solch ein Skript, so stellt man fest, ...

```
echo "Enter the your input reference GRCh38.fa file: "
read REF
#REF='/nfs/turbo/dcmb-brainsom/.../GRCh38_BSM.fa'
```

..., dass es mehre händische Inputs verlangt (und gerne auch mal ignoriert).

Konsequenz: Miserable Wissenschaft I

Aus einer Publikation über das Feststellen genetischer "Verschiebungen" (*genetic disorder*):
Ein Blick in ein anderes Skript verrät uns . . .

```
/home/crmumm/software/nanopolish/nanopolish index \
        -d /nfs/boylelab_turbo/nanopore_data/MEI/LINE/20191
        -d /nfs/boylelab_turbo/nanopore_data/MEI/LINE/20191
        -d /nfs/boylelab_turbo/nanopore_data/MEI/LINE/20191
        ...
<snip>
```

. . . , dass der Autor seine Software selber installiert (unversioniert?) und das Labor eigenes Equipment betreibt. – Toll!

? Die Pointe?

So was wird in "*Nature*" publiziert. Nebenan in "*Science*" finden wir u. a. ein leeres Repo! – Keine Einzelfälle!

Konsequenz: Miserable Wissenschaft I

Aus einer Publikation über das Feststellen genetischer "Verschiebungen" (*genetic disorder*):

Die Pointe?

So was wird in "*Nature*" publiziert. Nebenan in "*Science*" finden wir u. a. ein leeres Repo! – Keine Einzelfälle!

Ergo:

- Unkritischer Einsatz von FLOSS und Einsatz von "Hobbiysten" stellen einen deftigen Beitrag zur Reproduzierbarkeitskrise dar.
- Ihre Lösungen ("Pipelines") sind nicht einsetzbar.

Review to Rescue?

Journals wie *BMC Bioinformatics* offerieren Code Review – fakultativ und nach Einreichen der Veröffentlichung:

Piloting code peer review with Code Ocean

The Editor of *BMC Bioinformatics* is delighted to announce a 6 month pilot in partnership with Code Ocean, a computational reproducibility platform with a mission to make code more readily executable and discoverable.

During the author opt-in trial, *BMC Bioinformatics* and Code Ocean will be working together to facilitate easier peer review of source code and also code citation and sharing.

Authors submitting software articles to the journal may be invited to take part and, if they agree, will be guided though the process of creating fully functional "code capsules" that reviewers can access and run as part of the peer review process. And, once published, the authors can share their code and data, both of which will be citable via a Digital Object Identifier (DOI).

We hope that working with Code Ocean will provide many benefits to authors, reviewers and readers by improving the peer review experience and facilitating sharing of code that is reproducible and useful.

Finale

Beobachtungen:

- FLOSS ist für wissenschaftliches Rechnen unverzichtbar.
- Gute Software (meist: C/C++) setzt sich durch.
- Schlechte Software dominiert quantitativ.
- Publizierte Lösungen finden Anwender - und schaden oft der wissenschaftlichen Qualität.

Schlußfolgerungen:

- Nur gezeigt für Bioinformatik, aber auch zutreffend für Physik und Chemie: FLOSS in der Wissenschaft steckt in der Qualitätskrise!
- Wissenschaftliche Zeitschriften brauchen Regeln für den Review (Checklisten)!
- Nicht nur, aber auch in Deutschland, müssen wir auf Qualität in der Betreuung achten!