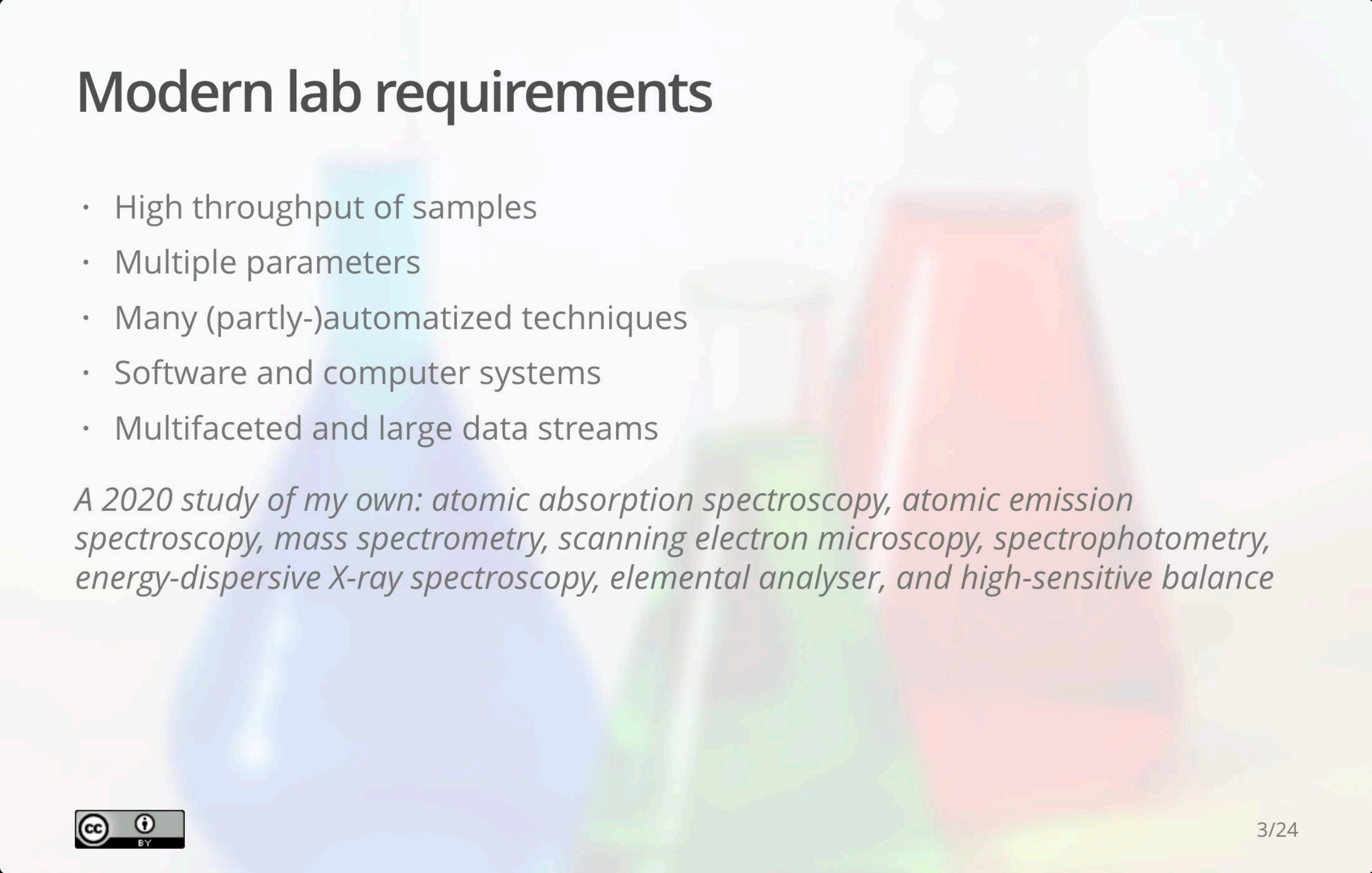
Utrecht University

Towards the FAIRification of lab-data

Integrated lab solutions for an open science lab

|Martin Schobben |Department of Earth Sciences |Utrecht University

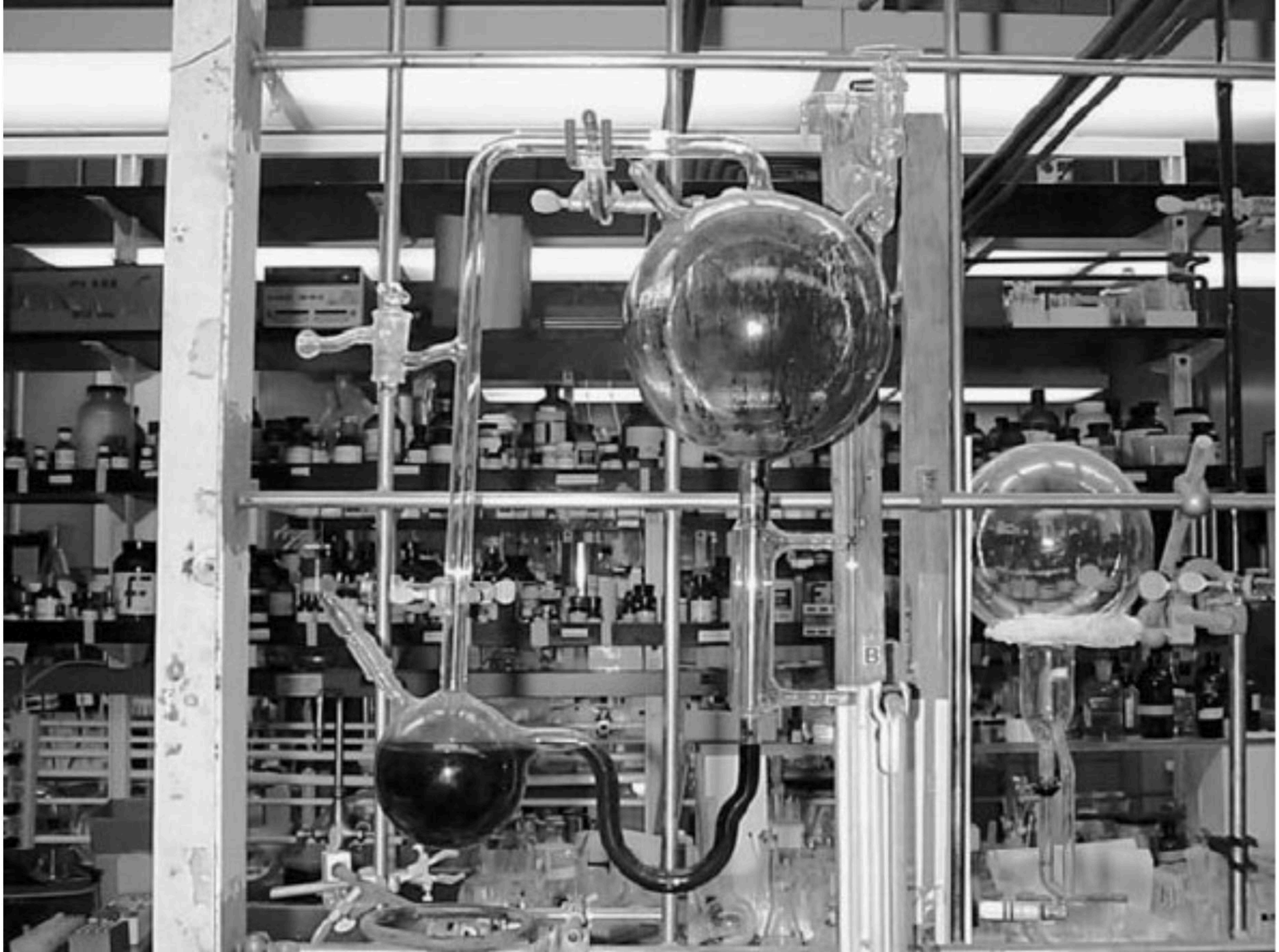
07 July, 2022



Modern lab requirements

- High throughput of samples
- Multiple parameters
- Many (partly-)automatized techniques
- Software and computer systems
- Multifaceted and large data streams

A 2020 study of my own: atomic absorption spectroscopy, atomic emission spectroscopy, mass spectrometry, scanning electron microscopy, spectrophotometry, energy-dispersive X-ray spectroscopy, elemental analyser, and high-sensitive balance



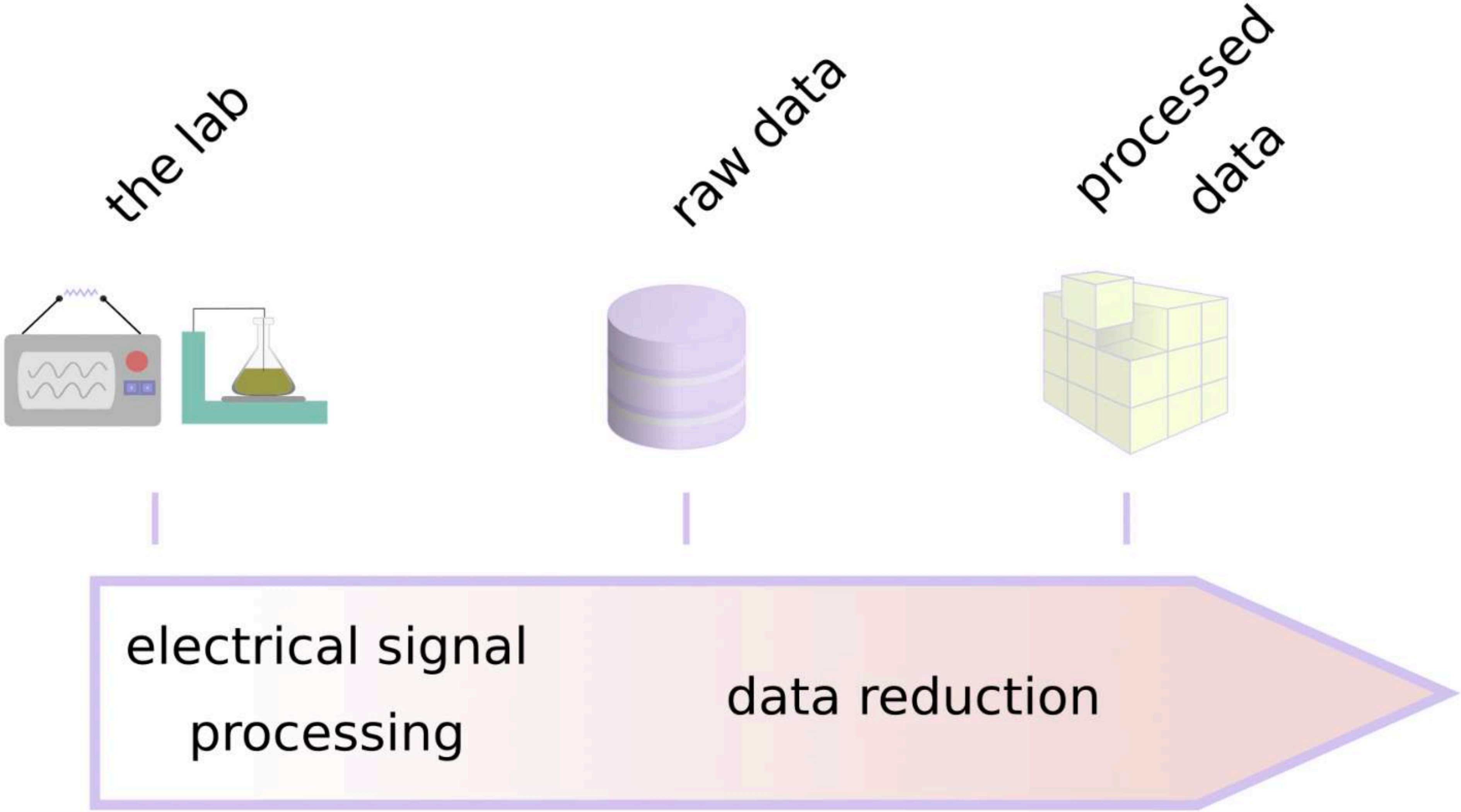
B

FEI
FEI
EVO 15

Taking command of lab-data

- Default commercial software (Cameca, Zeiss, Thermo Fisher)
- Prevent tracking data from source to publication
- Fragmented storage
- Monitoring and troubleshooting is reduced to current analysis

Vendor lock-in



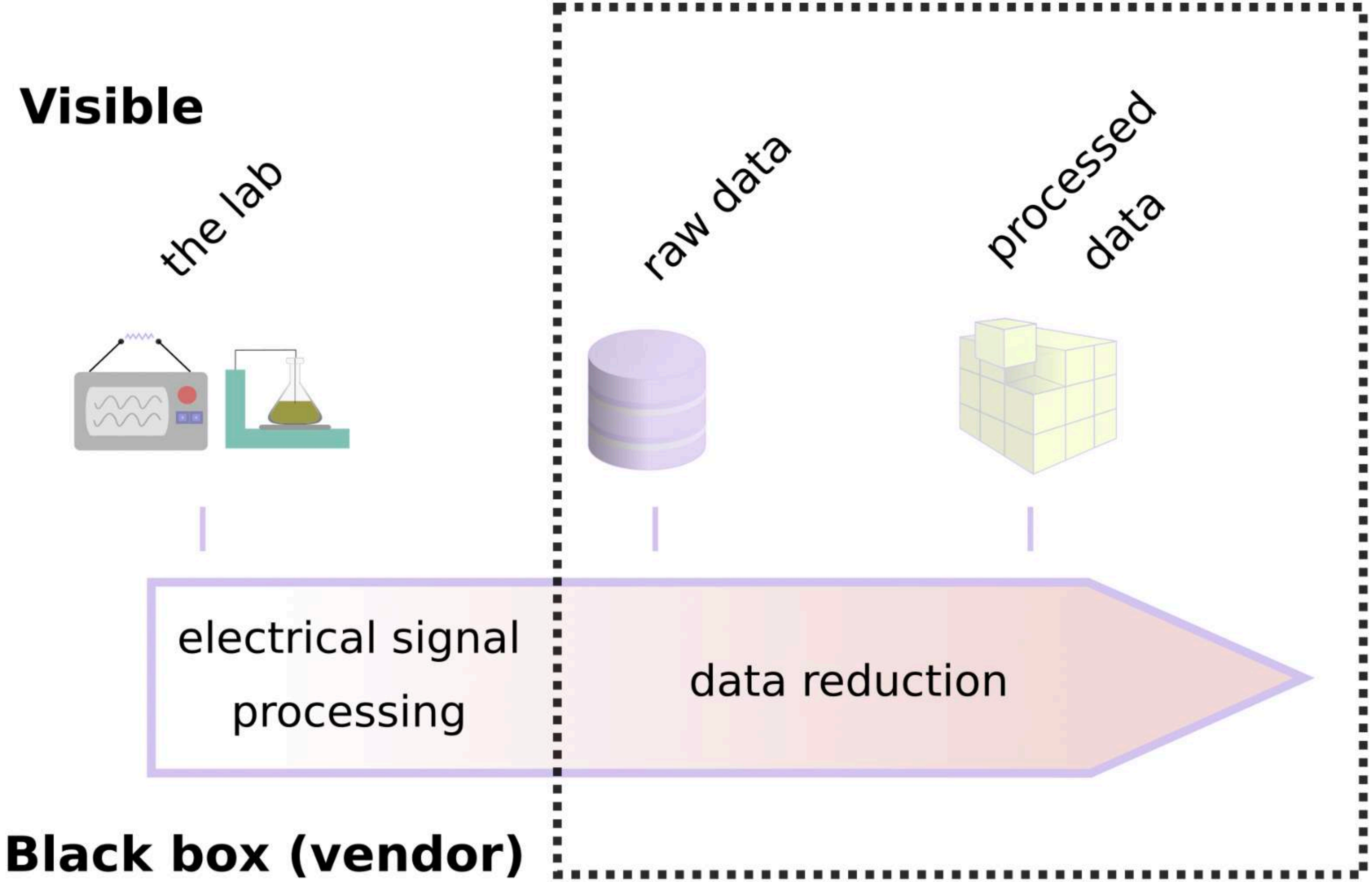
the lab

raw data

processed data

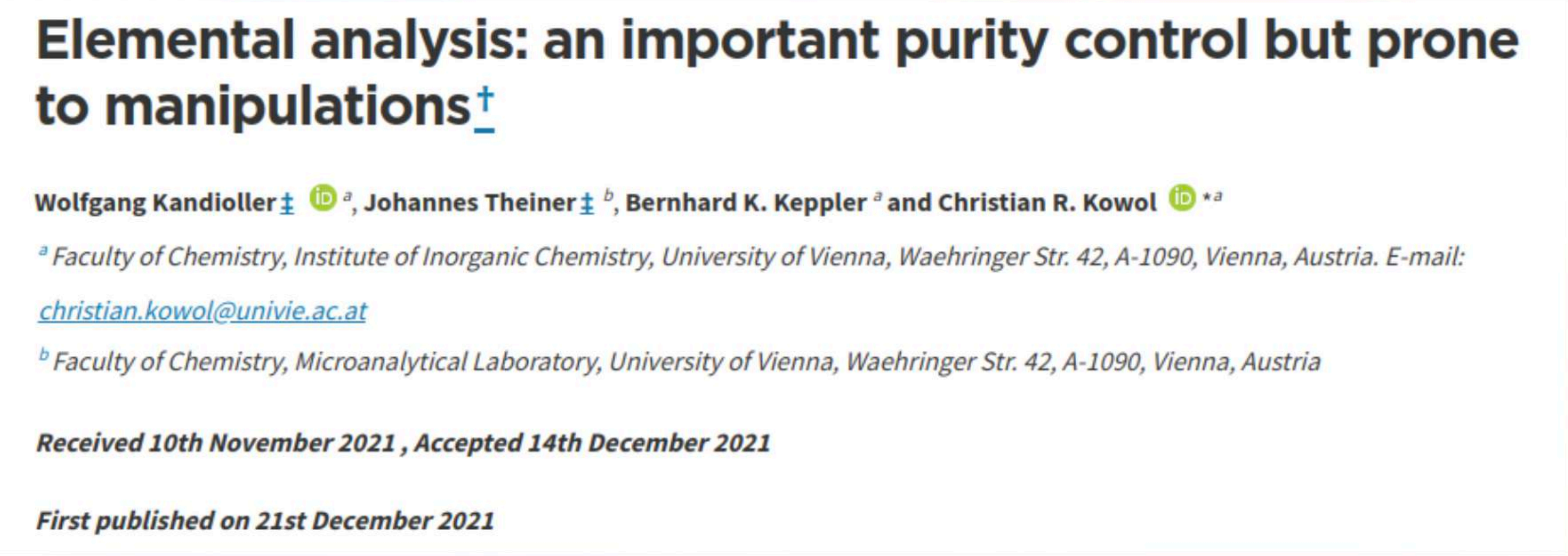
electrical signal processing

data reduction


Visible
the lab
raw data
processed data
electrical signal processing
data reduction
Black box (vendor)

Opening-up the black box of lab-data

- GUI based dashboards, wizards and dialogs that hide (part) of the transformations and calculations taking place
- The reviewer that wants to trace back the origin of data
- Old/defunct machinery



Elemental analysis: an important purity control but prone to manipulations†

Wolfgang Kandioller‡ [a], **Johannes Theiner‡** [b], **Bernhard K. Keppler** [a] **and Christian R. Kowol** *[a]

[a] *Faculty of Chemistry, Institute of Inorganic Chemistry, University of Vienna, Waehringer Str. 42, A-1090, Vienna, Austria. E-mail:* christian.kowol@univie.ac.at

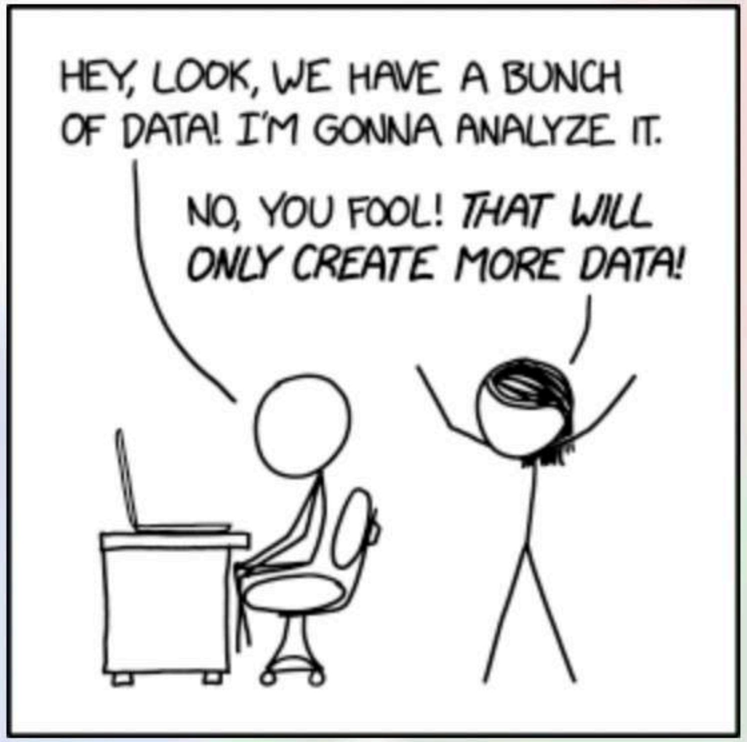
[b] *Faculty of Chemistry, Microanalytical Laboratory, University of Vienna, Waehringer Str. 42, A-1090, Vienna, Austria*

Received 10th November 2021 , Accepted 14th December 2021

First published on 21st December 2021

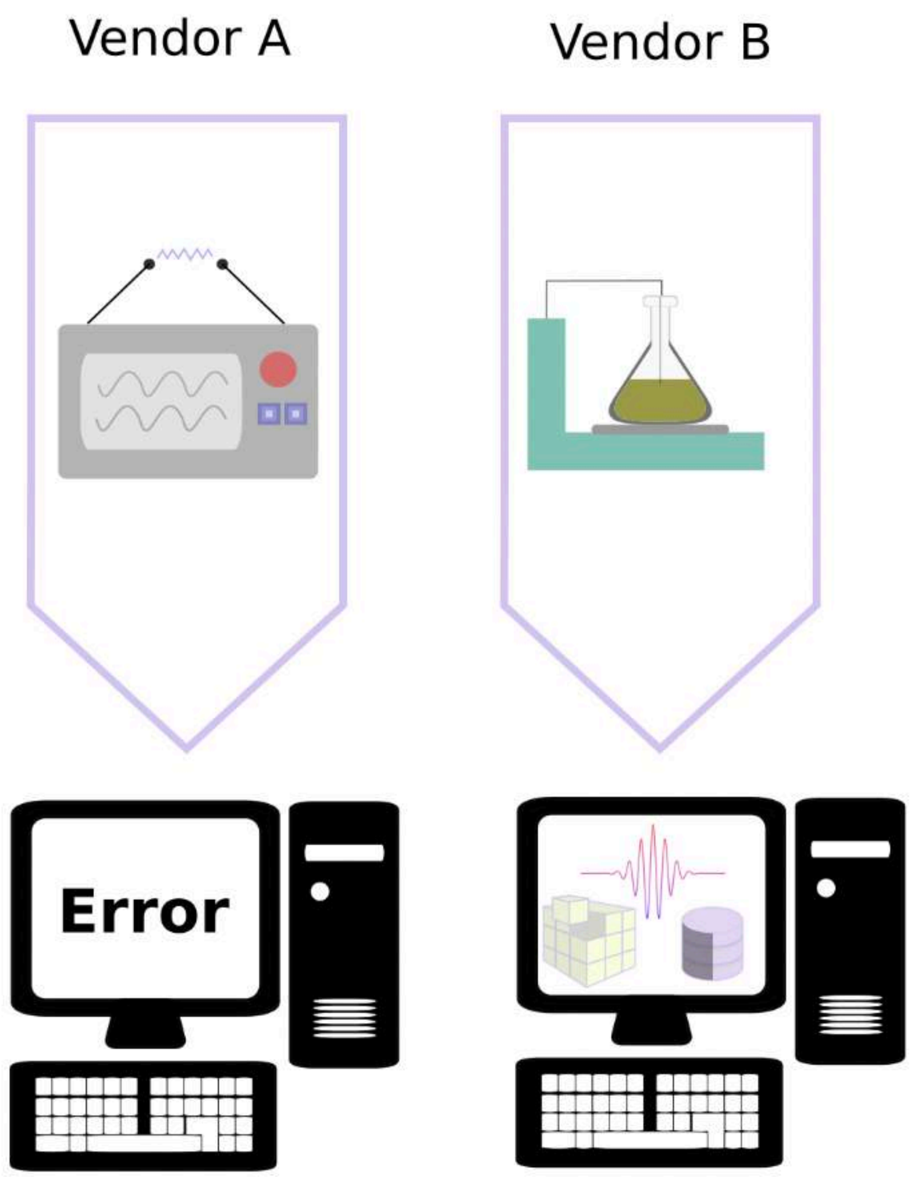
Is more data better?

- New innovations
- Inclusive science
- More transparent science (proof of final published values)


HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! *THAT WILL ONLY CREATE MORE DATA!*

xkcd.com

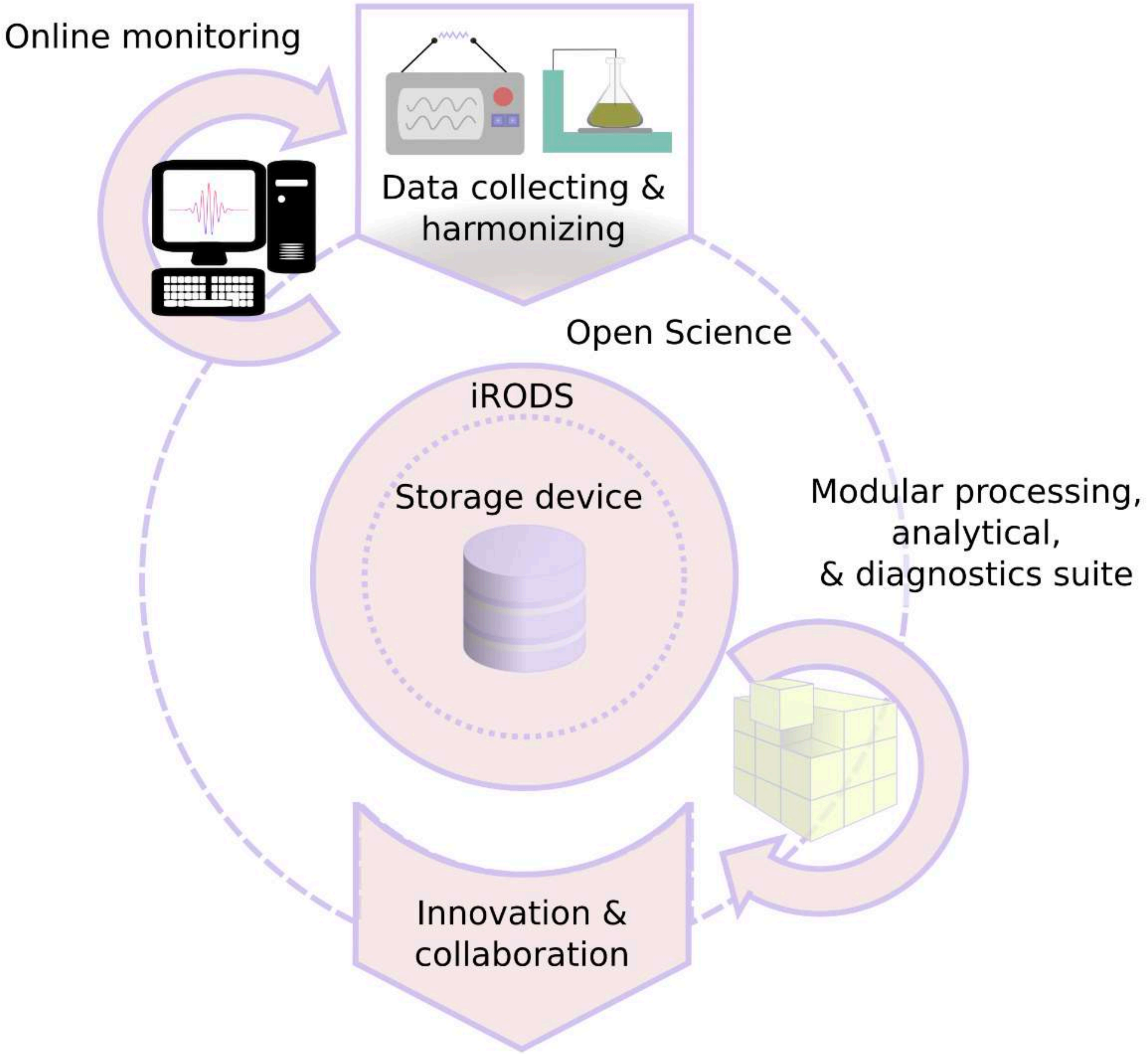
Unconnected Lab



Vendor A

Vendor B

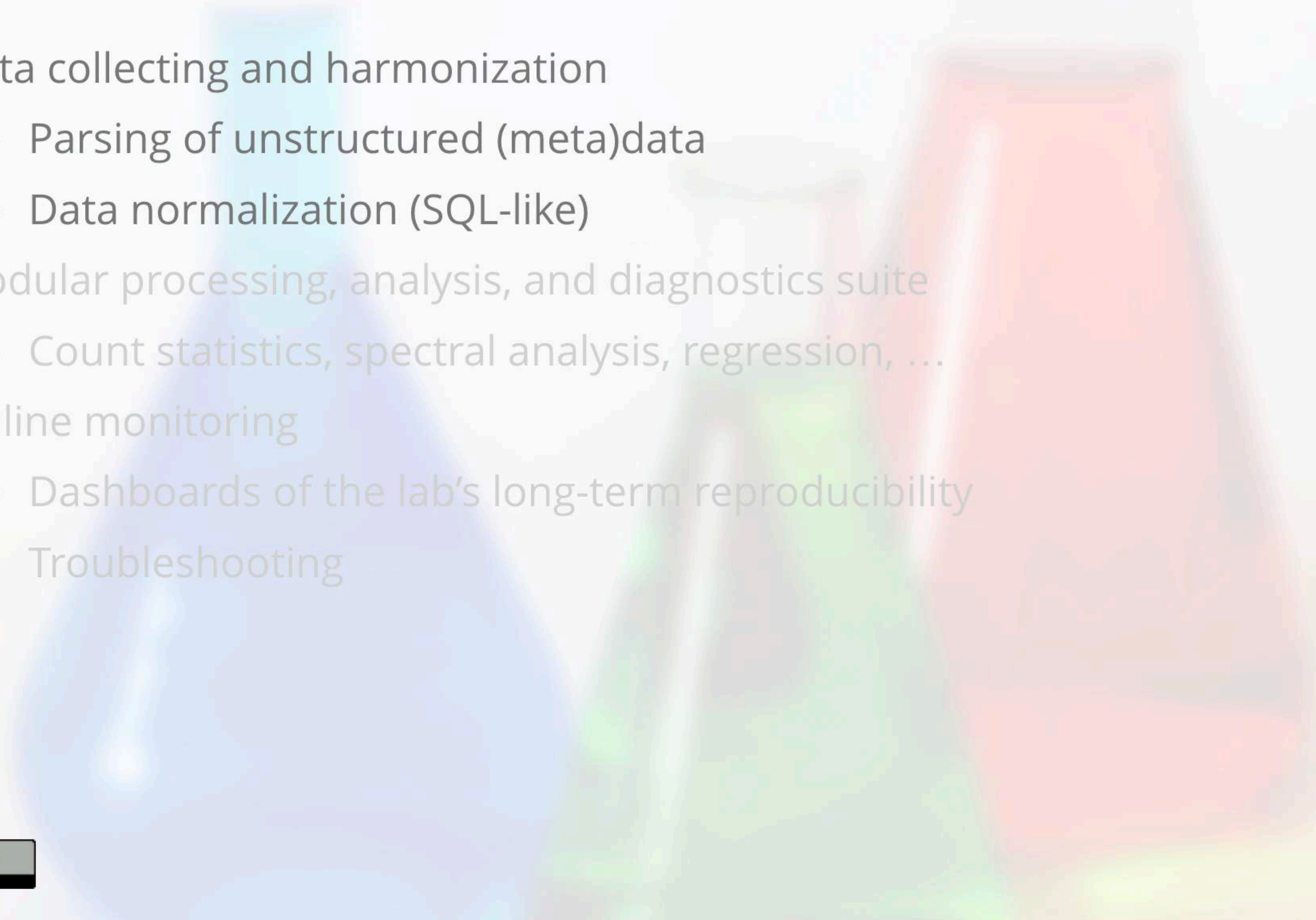
Integrated Lab



The integrated lab

- Data collecting and harmonization
  - Parsing of unstructured data
  - Data normalization (SQL-like)
- Modular processing, analysis, and diagnostics suite
  - Count statistics, spectral analysis, regression, …
- Online monitoring
  - Dashboards of the lab’s long-term reproducibility
  - Troubleshooting

The integrated lab



- Data collecting and harmonization
  - Parsing of unstructured (meta)data
  - Data normalization (SQL-like)
- Modular processing, analysis, and diagnostics suite
  - Count statistics, spectral analysis, regression, …
- Online monitoring
  - Dashboards of the lab’s long-term reproducibility
  - Troubleshooting

Data collecting and harmonization

Custom solutions

- Deciphering the vendor’s data-model is labor intensive
  - Multiple files
  - Many observations
  - Inconsistent syntax
  - Unstructured
- Accommodate vendor’s software/data-model updates

An universal solution?

Parsing lab-data

Text data (encoded or decoded)

```
2021-09-20  20:15                                Sample ID: MON-233

Peak Height Distribution: 210 V, EMHV: 2350 mV
Position: x=12um; y=2um; z=100um

  Time (s)  Count
  1         56
  2         60
  3         64
  4         64
  5         57
  6         59
  7         58
  8         58
  9         62
  10        54
```

Proposed solutions

Three possible solutions, which require varying degrees of human intervention:

1. A mechanism to aid the location of variables based on user input
2. A human-crafted (and adaptable) rule based system
3. A natural language processing approach involving self-supervised machine learning

The last two solutions would be preceded by a step entailing text normalization through tokenization.

iRODS and user accesibility

Integration with iRODS

- Sub-system for automated ingest
- Automated workflows
- Better collaboration

Accessibility

- Interfaces for R and Python (standalone usage)

Evan-Amos, 2011. A SanDisk Cruzer USB drive from 2011, with 4 GB of storage capacity Wikipedia

Implementation

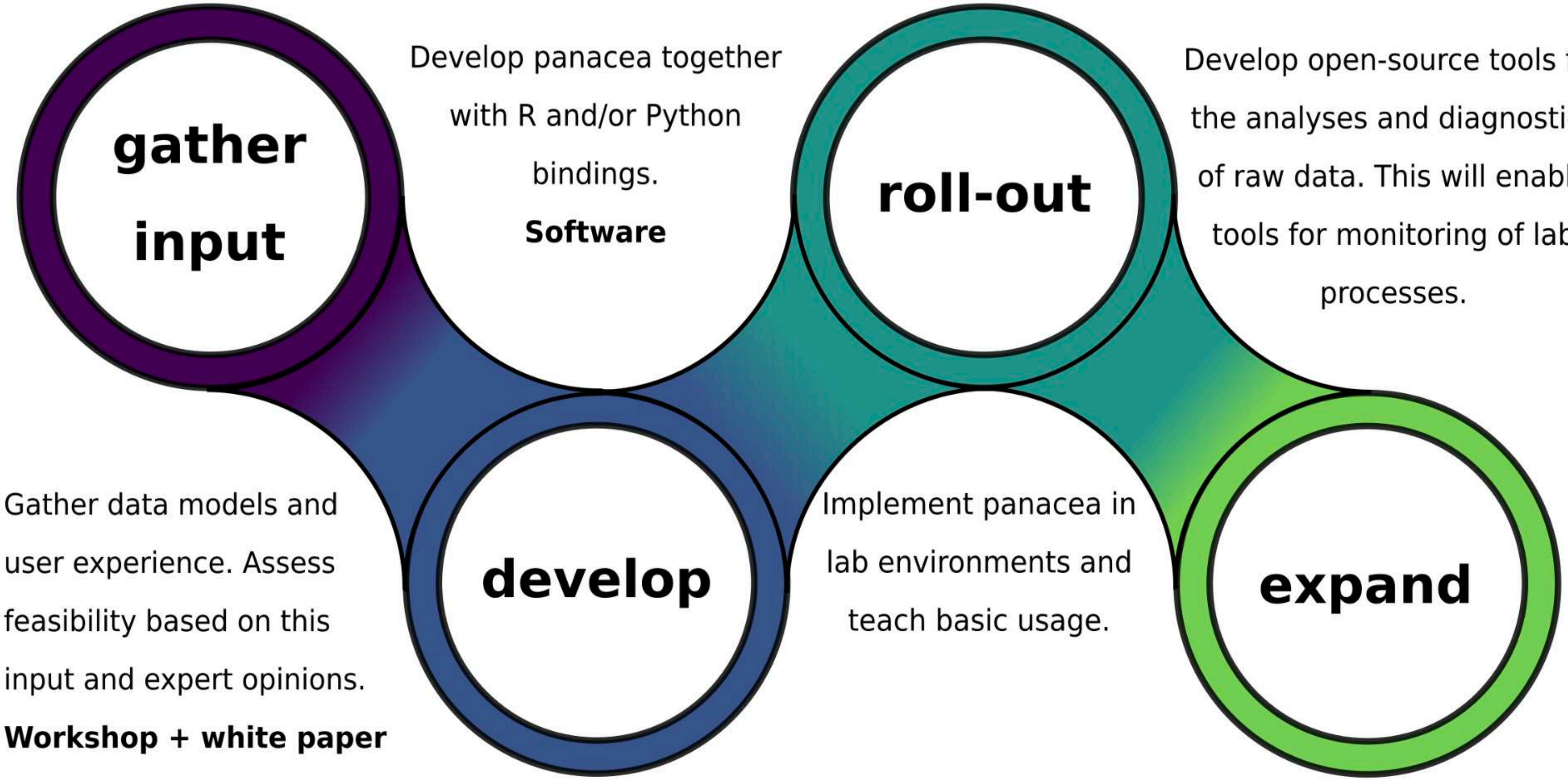
panacea: Portable ANalytical data Aggregation and Coordination for database Entry and Access

- C++ for optimal performance with large datasets
- R and python bindings for user-friendliness and standalone usage

auxiliary

- Updating *rirods* (*irods/irods_client_library_r_cpp*) to work with iRODS from R
- Restrictive and complex system requirements not ideal for R and C++ integration

Roadmap



Gather data models and user experience. Assess feasibility based on this input and expert opinions.

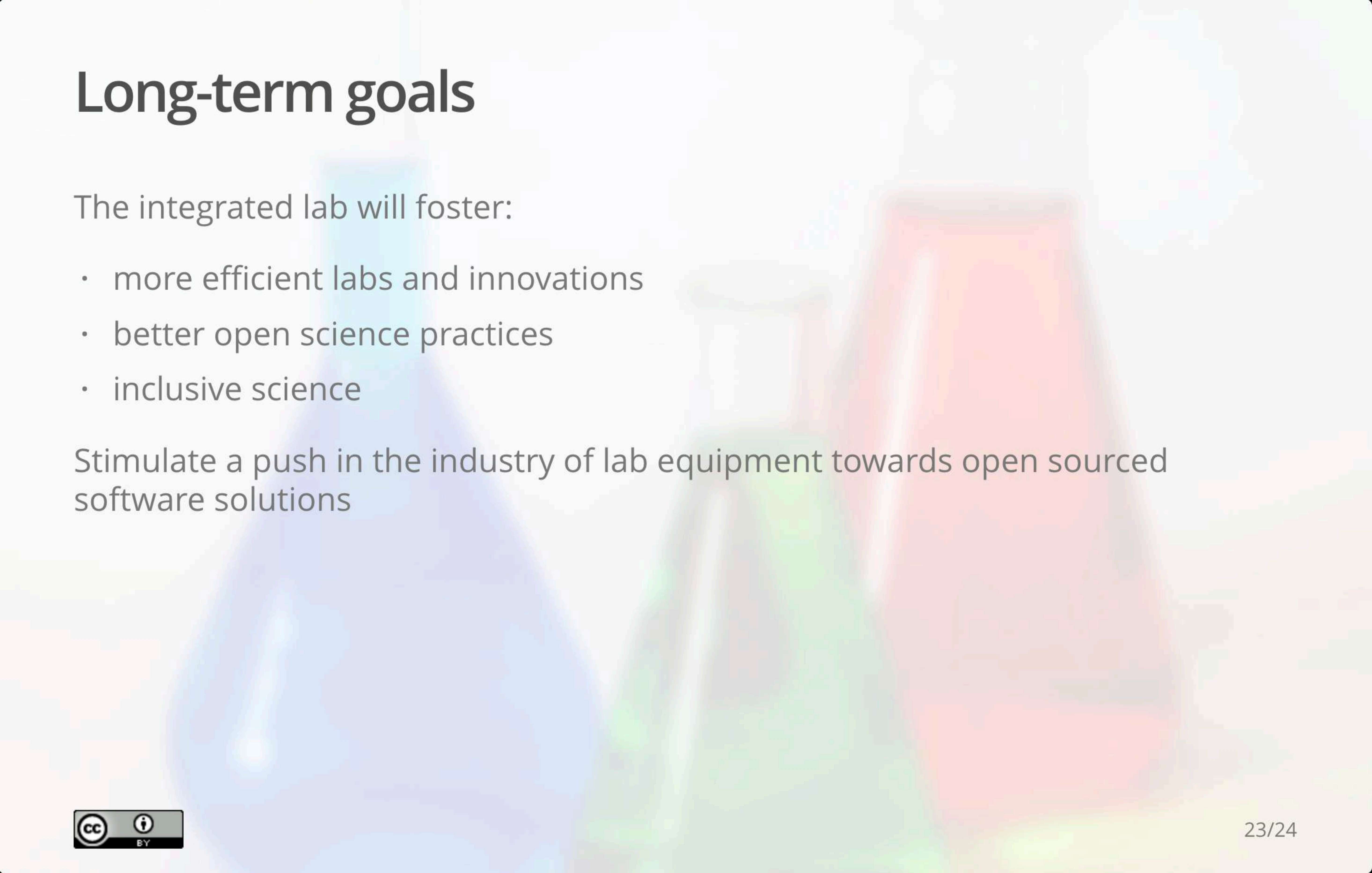
Workshop + white paper

Develop panacea together with R and/or Python bindings.

Software

Implement panacea in lab environments and teach basic usage.

Develop open-source tools for the analyses and diagnostics of raw data. This will enable tools for monitoring of lab processes.



Long-term goals

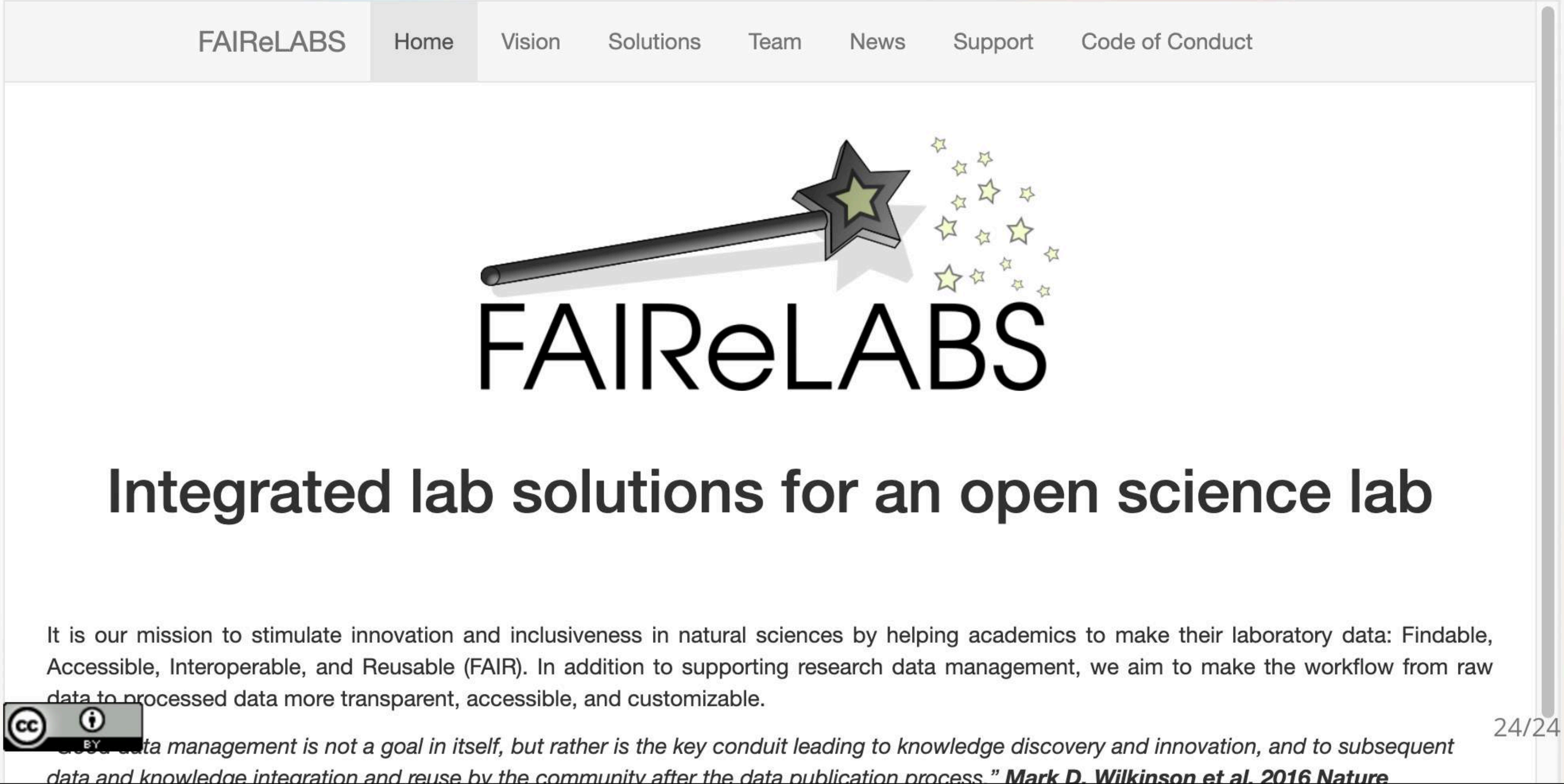
The integrated lab will foster:

- more efficient labs and innovations
- better open science practices
- inclusive science

Stimulate a push in the industry of lab equipment towards open sourced software solutions

FAIReLABS

Help us! https://fairelabs.github.io/webpage/



Integrated lab solutions for an open science lab

It is our mission to stimulate innovation and inclusiveness in natural sciences by helping academics to make their laboratory data: Findable, Accessible, Interoperable, and Reusable (FAIR). In addition to supporting research data management, we aim to make the workflow from raw data to processed data more transparent, accessible, and customizable.

"Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process." ***Mark D. Wilkinson et al. 2016 Nature***