ETH zürich

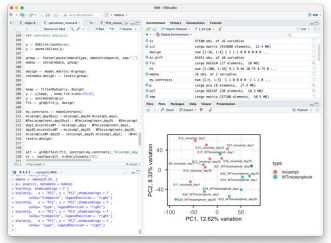
Scientific IT Services

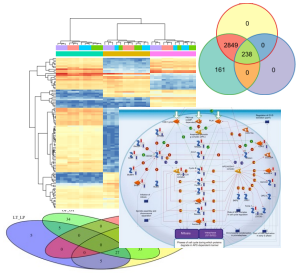
SIB days workshop

Best practices to support the reproducible research value chain in bioinformatics: from raw data to final publication

Caterina Barillari, Rostyslav Kuzyakiv, Michal Okoniewski

Bioinformatic research value chain





Findable Accessible Interoperable Reusable

added value

RRP

data stewardship

What else we do… other activities for biology and bioinformatics

High Performance Computing: Shareholder model

- Euler is financed (for the most part) by its users
- So far, over **186 (!)** research groups from almost all departments of ETH have invested in Euler
- These so-called **shareholders** receive a share of the cluster's resources (processors, memory, storage) proportional to their investment
  → fair share principle
- **Public share**: A small share of Euler financed by IT Services is open to all members of ETH; **guest users** can use limited resources (48 cores and 128 GB of memory)

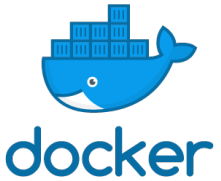
Scientific Software Development: We help ***you*** to develop software

We ensure long term support for your software!

Code custodianship.

Our developers can be part of your team.

GitLab CI



We help your team to improve!

We offer courses and consulting in many fields:

- Best practices of professional software development
- Version control using git
- Introduction to Python, Writing fast Python code.
- Tailored courses for your needs

The quality of your software reflects the quality of your research!

We review, package and improve your software previous to release.

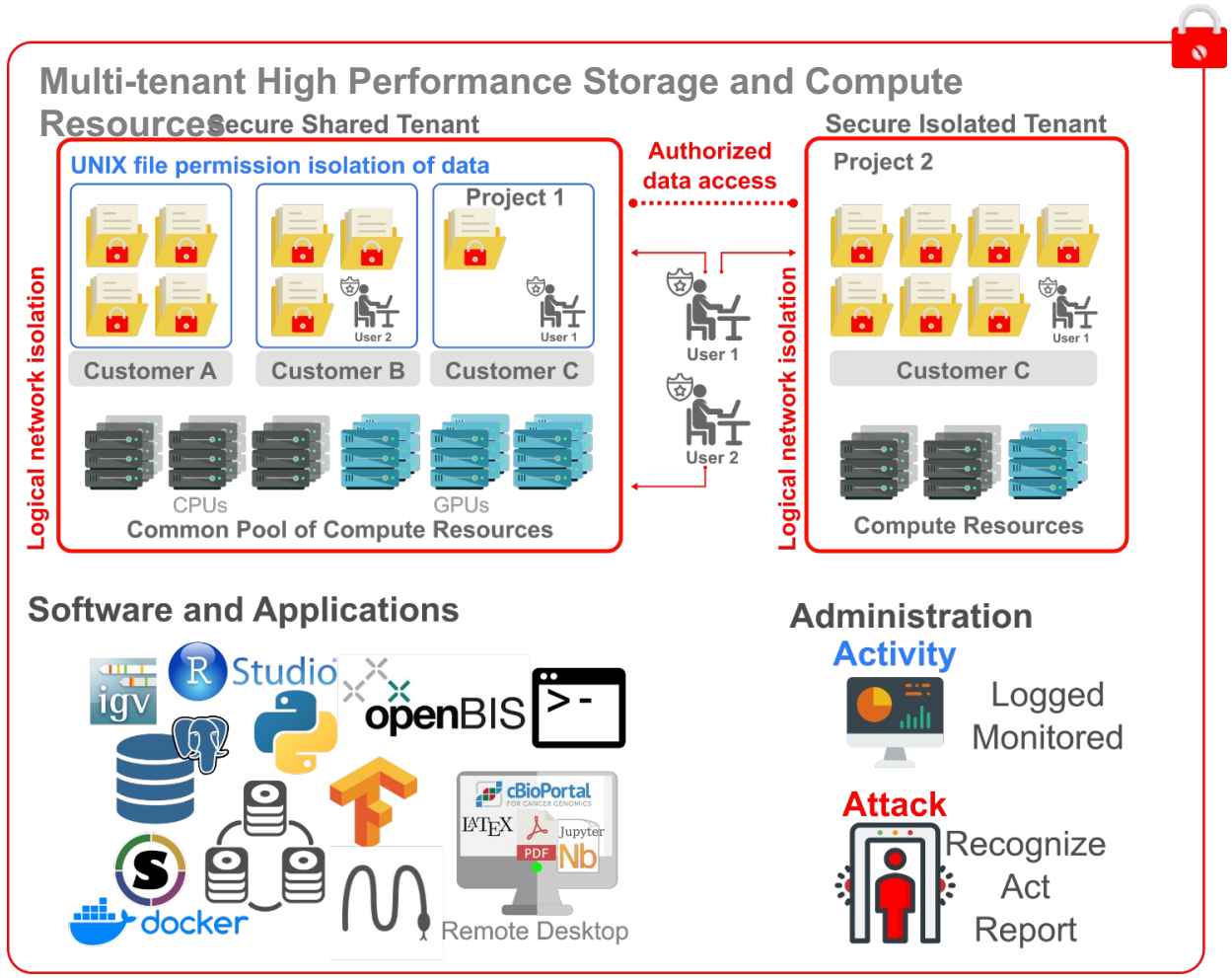
We teach you principles of professional software development.

Example support types in co-analysis mode

- Scaling and parallelization on the **computing cluster**
- **Code** profiling and optimization (R, Python, Bash…)
- Analysis of RNA-seq or other **–omics** data
- Reproducible analysis **workflows**, e.g. snakemake
- Non-typical analysis, e.g. **alternative splicing**, protein domains, regulatory elements
- Bioinformatic data **visualization**
- Making **external software** or code run
- Statistical consulting, e.g. **tuning statistical tests** for specific data
- Data mining in the results and public data, e.g. sequences searches
- Database use and data management consultancy
- Proposing and implementing novel data science applications

data stewardship

Leonhard Med: Secure scientific data & IT platform for research with confidential data



- powerful research IT platform to securely store, manage and process (e.g. bioinformatics, data science) **confidential research data**
- enables collaborative, large-scale and very diverse **biomedical research** (including academies and hospitals) at ETH Zurich
- part of the national **BioMedIT network** of secure data centers supporting projects in the SPHN and PHRT national programs

SPHN: Swiss Personalized Health Network
PHRT: Personalized Health and Related Technologies

Model development and training

Machine Learning support includes:

- Literature review to find appropriate models
- Implementation and test of models
- Proficient use of a broad spectrum of ML tools (e.g. TensorFlow, PyTorch, scikit-learn, etc.)

Support can range from consultancy to development of whole machine learning pipeline.

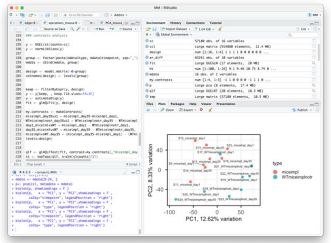
Consulting & Training

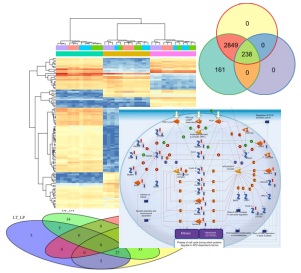
SIS hold / organize regular trainings and workshops in various areas of scientific computing / IT with a *hands-on focus* (complementary to other ETH courses)

Format: both on-site / online possible

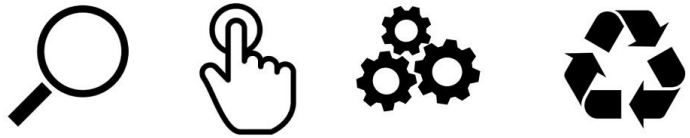
Courses	Duration
ETH Research Data Management Summer School	5 days
openBIS ELN and LIMS Trainings	Varying
Introduction to Machine Learning using Python	4 days
Writing fast(er) code in Python	4 days
Parallel programming with MPI/OpenMP	6 days
Scientific Visualization using Python	2 days
Introduction to Euler	Ca. ½ days
High Performance Computing for Genomic Applications	2 days
Workshop on best practices in Programming (git, unit testing, clean code)	2 days
Introduction to programming with Python	8-10 lectures

Bioinformatic research value chain





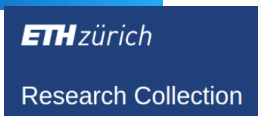
Findable Accessible Interoperable Reusable



added value

RRP

data stewardship



Discussion points

- Do you use similar solutions along your bioinformatic value chain?
- What else you do to ensure reproducibility of data, code and results?
- What are your other favourite tools to ensure FAIR-ness?
- What gaps you see in our value chain?
- What else you have in your bioinformatics value chain?
- What stages of the value chain you would like to see automated?
- Would your collaborators and users also use the automation?
- How much you are eager to learn or change your habits to ensure reproducibility?
- What other experience in this area you would like to share with the group?

ETH zürich

Best practices to support the reproducible research value chain in bioinformatics: from raw data to final publication

Caterina Barillari, Rostyslav Kuzyakiv, Michal Okoniewski

Thank you!