
Research Software Engineering

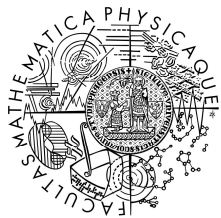
Mgr. Zdeněk Mašín, PhD

Institute of theoretical physics

Faculty of Mathematics and Physics

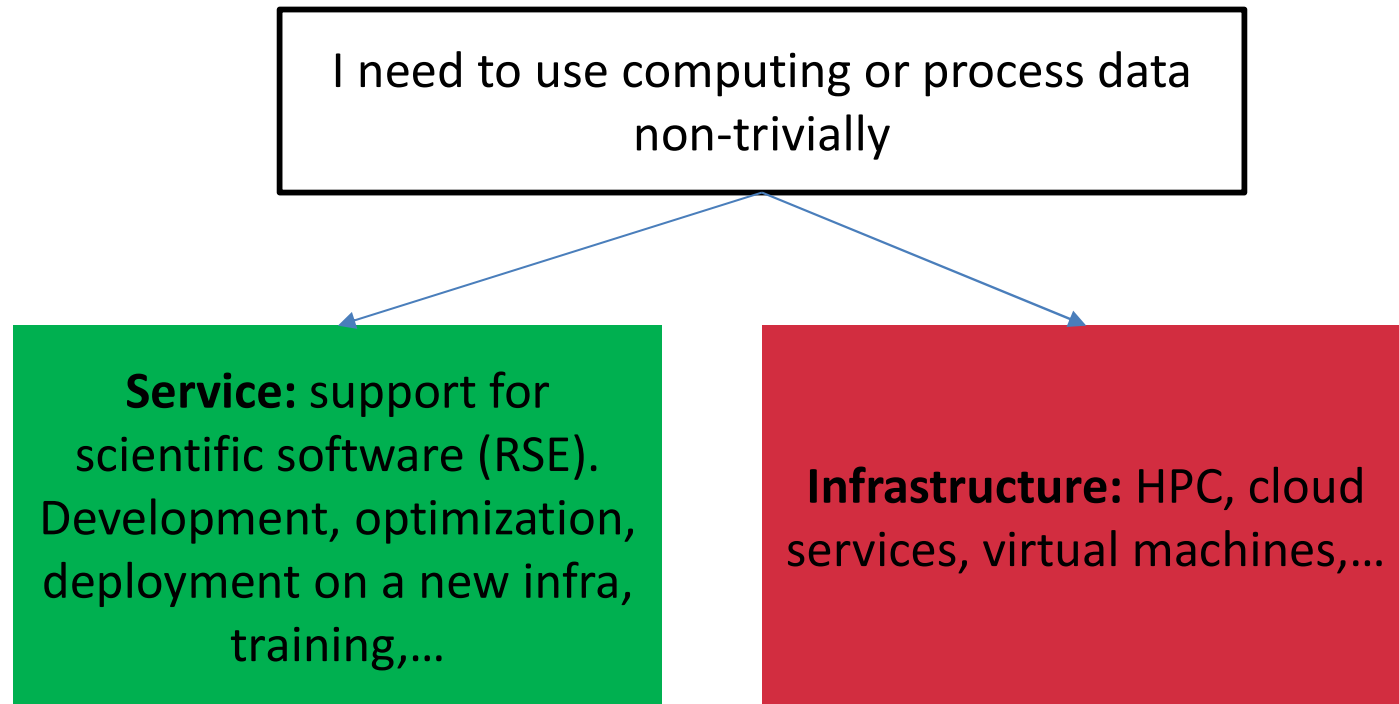
&

RSE team head



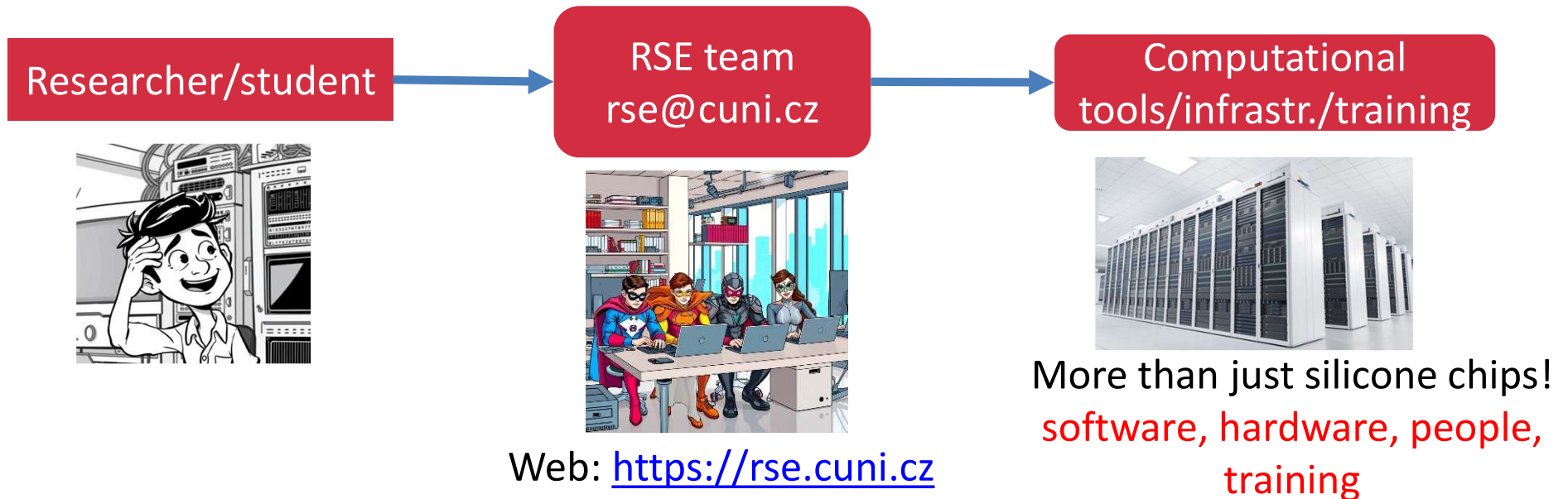
FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

What do I need as a researcher?



What is Research Software Engineering (RSE)?

We lower the barrier of researchers to computing tools and infrastructures





CÍL RSE

Vývoj a údržba výzkumného softwaru jsou nedílnou součástí akademické činnosti a vyžadují specializované odborné znalosti, čas a dlouhodobou kontinuitu.

Posláním [týmu Research Software Engineering \(RSE\)](#) je systematicky zajišťovat tuto činnost v úzké spolupráci s akademickými pracovníky. Tým RSE

Why are we doing this?

- **Researchers = domain experts**
- Evaluation: number and quality of publications
- Need to use software and/or develop their own
- (Research) students need training in computing but sometimes not available locally
- Need to ensure openness/reproducibility/reliability of data/computation
- Waste of time/resources if software not developed sustainably



Typical use cases

- I need to **automate** processing of medical imaging data (e.g. volumetry, segmentation).
- I need to use a **specialized infrastructure** for working with sensitive data (SensitiveCloud) but I need assistance setting it up.
- I am preparing a **research proposal** with a computing component, but I need to consult my ideas or requirements.
- I need **help designing** an online platform for collection of research data from clinicians.
- I need to **deploy my software** on different computing infrastructures.

How are we doing this?

- Combine expertise from different fields/faculties
- **Six RSE core team members**
- Merged with AI CUNI team (3 people)
- Requests send to rse@cuni.cz

Expertise

- Bioinformatics
- Computer Vision
- Software containerization/portability
- Cloud computing
- AI/LLM in biomedicine and humanities
- High Performance Computing

Infrastructure

- Computing cluster MFF
- JupyterHUB
- Kubernetes cluster (planned)
- <https://www.mff.cuni.cz/en/hpc-cluster/general-information>
- Ask for access/assistance

Who is doing it?

2nd Faculty of Medicine

Sára Veselá, Tomáš Preisler
Leaders: K. Fišer, J. Stuchlý

Faculty of Philosophy

Luděk Svoboda
Leaders: O. Tichý, M. Vacura (Center for
Digital Humanities)

Faculty of Mathematics and Physics

Martin Topinka, Jiří Eliášek
Leaders: Z. Mašín, J. Yaghob

Central Library

RSE coordinator: M. Basovník
+ AI team (3 members)

How to secure our support?

- Contact us at rse@cuni.cz
- “Short-term” support is for free

Project funding

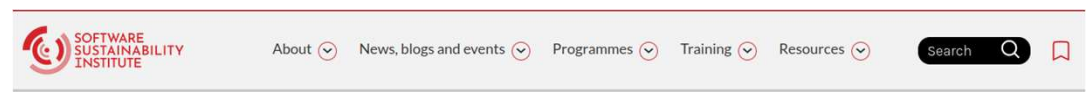
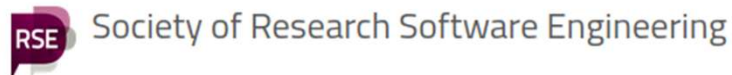
- Extensive long-term support: team member on grant
- Continuous low-intensity support: e.g. 5% FTE for technical staff support

Computing infrastructure for cluster

- Server procurement: efficient scheme currently under investigation

We're not alone

- RSE is widespread at top universities
- E.g. UCL Advanced Research Computing Dept. cca 120 people (<https://www.ucl.ac.uk/advanced-research-computing/>)
- We're part of the EU project EVERSE, have partners in Heidelberg, Southampton



<https://www.software.ac.uk/>

About us

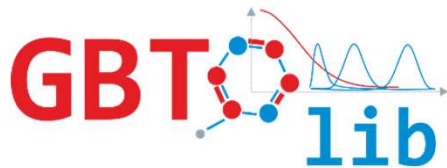
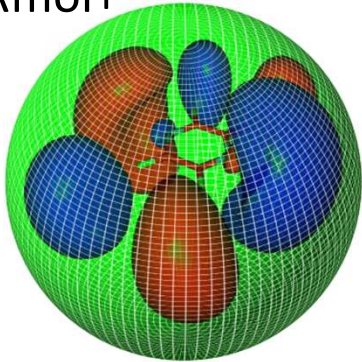
The Software Sustainability Institute is the first organisation in the world that was dedicated to improving software in research. We help people build better and more sustainable software to enable world-class research.

Read more >



Use case – code containerization

UKRmol+

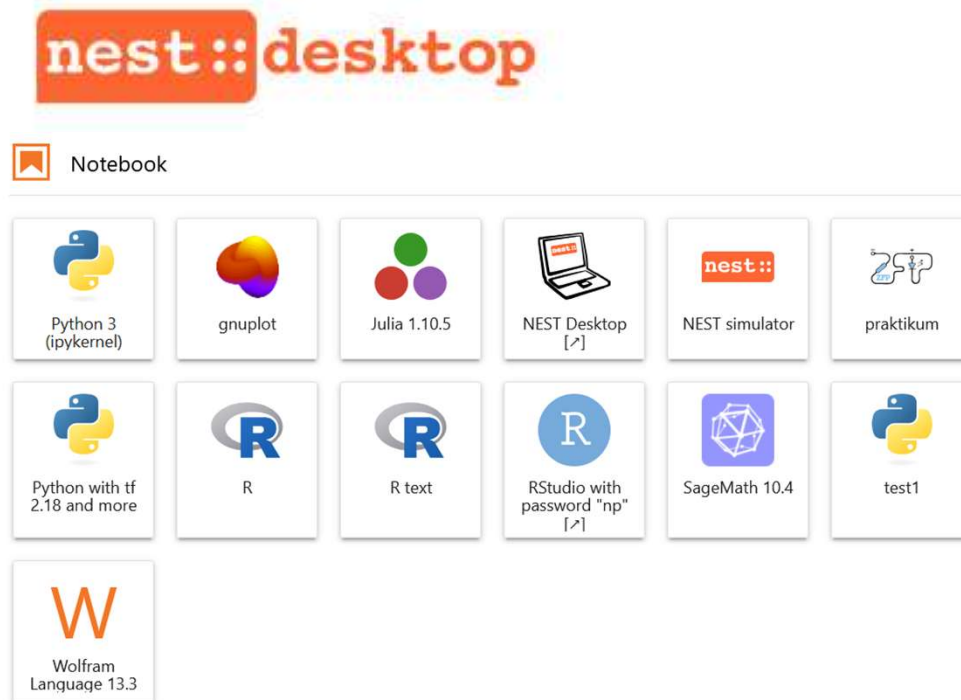


- Complex code developed over decades
- Complicated compilation: external libraries (LAPACK, ELPA, SCALAPACK, PETSc, SLEPc, MPI)
- Lots of queries from users
- New release of the code packaged with container

```
docker pull zdenekmasin/ukrmol_plus:3.3.0
```

- Conda environment available too: compilation on the host computer

Use case – teaching, code optimization



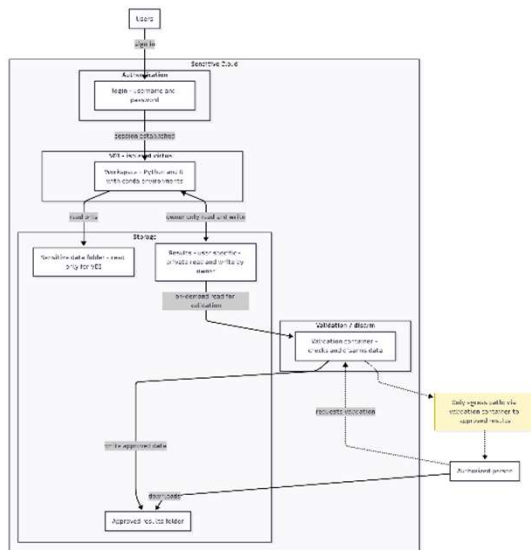
- SW for computational neuroscience
- Developed partially at MFF
- Needed for teaching
- Deployment at MFF JupyterHUB

- Follow-up project: optimization of a HPC code MOZAIC (python, C++, external components)

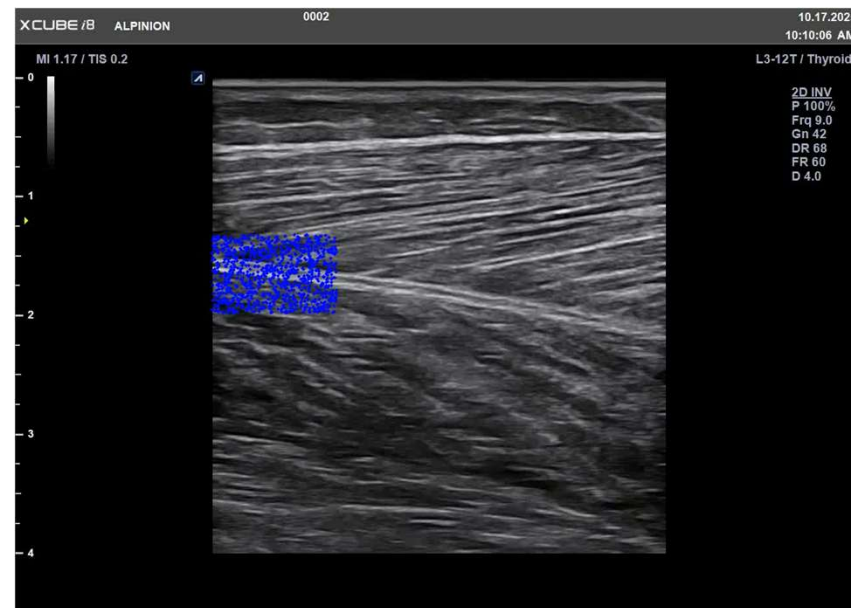
Snippets of ongoing projects

SensitiveCloud

- complex computing env. for faculty of law



Analysis of sonographic images and videos



Snippets of ongoing projects

AI medical translator/rewriter

- rewrite medical reports into plain czech language using LLM (knowledge base, embedding models, ...)

Clinical study

- goal: develop a web app for data collection and analysis

AI prompting and RAG

- AI for teaching/studying (patient simulation)
- **RAG** as resource manager
- **AI Sandbox** - available models on university's hardware



Snippets of ongoing projects

Bioinformatics – single cell RNA analysis for immunology research
- Implementation of complex data processing pipelines and toolsets callable from a single UI

1. Zarovnání sekvencí a zpracování dat (Alignment)

- BWA-MEM (v0.7.12): Zarovnání DNA sekvencí (DNA alignment).
- STAR (v2.7.11): Zarovnání RNA sekvencí (RNA alignment).
- Minimap2 (v1.0): Zarovnání dlouhých čtení (Iso-Seq).
- SAMtools (v1.10) & htslib (v1.10.2): Manipulace s BAM soubory.
- Trim Galore (v0.4.4): Odstranění adaptérů a kontrola kvality (trimming).
- bcftools (v1.10.2 / 1.19): Utility pro práci s variantami.
- bedtools (v2.25.0): Genomická aritmetika a operace s intervaly.

2. Volání variant (SNVs a Indely)

- Mutect2 (GATK v4.1.2.0)
- Strelka2 (v2.9.10)
- VarScan2 (v2.4.3)
- SomaticSniper (v1.0.5.0)
- MuSE (v1.0rc)
- Pindel (v0.2.5): Specificky pro detekci indelů.
- Annovar (v1.0): Anotace variant.

3. Analýza strukturních variant (SV) a CNV

What do you do?

High Performance Computing

→ Teaching

Code that requires a large number of computing cores/storage/time (dozens to thousands of cores)

Code that I need to run many times (e.g. data processing pipelines, exploration of parameter space of computational problems, etc.)

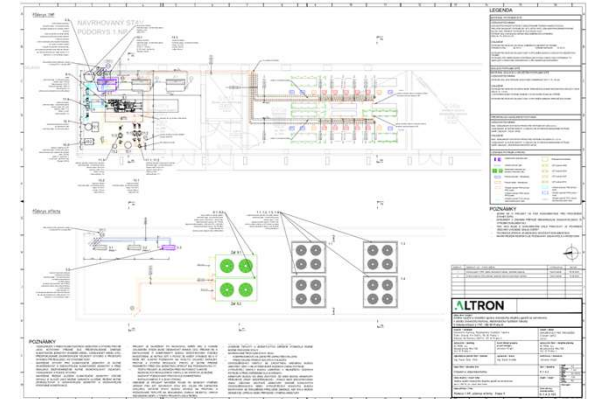
Scientific code development across disciplines

Data center Troja

500 kW total capacity:

- 250 kW using liquid cooling
- 250 kW using standard air cooling

- Under construction
- Completion: March 2026
- Purchase of new computing HW underway:
19 CPU servers, 8 GPU servers, 4 high-memory CPU servers, storage



Users of computing infrastructure

Natural and computer sciences

- Proficient in low-level programming
 - Need large computing power
 - Assistance needed in highly specialized optimization and adoption of new HW
- e.g. bioinformatics

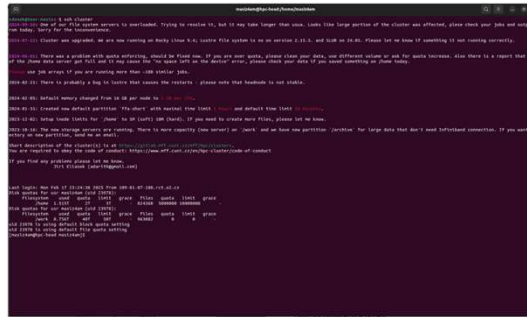
Biomedical disciplines

- Data sharing and infrastructure
 - Sensitive data
- e.g. LINDAT

e.g. SensitiveCloud

Humanities and social sciences

- Focus on data processing and visualization
- Sensitive data
- Scripting is the main tool
- Cloud applications
- Typically not (yet) large computing resources
- Assistance needed in setting up the computing environment



LINDAT/CLARIAH-CZ has been established as a unification of two research infrastructures, LINDAT/CLARIN and DARIAH-CZ. It is a unique research infrastructure, which deals primarily with language data but also with other digital resources and tools for their exploitation, maintenance and enhancement and offers them to research community, to industry for the development of applications and in specific cases, such as e.g. language culture, also directly to the public domain. LINDAT/CLARIAH-CZ is a joined, distributed Czech national node of the pan-European CLARIN ERIC (Common Language Resources and Technology Infrastructure) and DARIAH ERIC (Digital Research Infrastructure for the Arts and Humanities) networks. It consists of 11 top research organizations that are active in the domain of humanities and arts in the Czech Republic – in linguistics, history and historical bibliography, culture and science on culture, history of arts, philosophy, film culture, visual arts, musicology and history of music, ethnology, folklore, archaeology and also in some crossdisciplinary domains.

Centrum pro digital humanities

FF > Fakulta > Struktura a historie > Katedry, ústavy a další pracoviště > Centra a specializovaná pracoviště > Centrum pro digital humanities



Centrum pro digital humanities sdružuje především členky a členy akademické obce FF UK se zájmem o digital humanities.

kontaktní e-mail: cdh@ff.cuni.cz

Posláním CDH je podpora a koordinace pedagogické, vědecké, výzkumné a osvětové činnosti na fakultě zaměřené zejména na oblast digital humanities.

Jedná se například o:

- Rozvoj metodologie pro získávání, zpracovávání, analýzu, prezentaci a uchování dat v digitální podobě pro účely humanitních, společenských a uměnovědných oborů.
- Spolupráci s dalšími pracovišti podobného zaměření, vytváření podmínek pro tvorbu a sdílení dat, přípravu grantů a projektů, včetně publikace specifických typů výsledků výzkumu v oblasti digital humanities.
- Zprostředkování výsledků vědeckého výzkumu širší odborné i laické veřejnosti a snaha o uznání specifických typů výstupů digital humanities ve vědeckém i celospolečenském prostředí.

Členstvo CDH

Projekty

Kurzy

Odkazy a nástroje

Vnitřní organizace CDH:

CDH bylo založeno a řídí se Opatřením děkanky 28/2022 „[Statut Centra pro digital humanities](#)“.

Radu CDH tvoří:

[Ondřej Tichý](#) (předseda, ÚAJD), [Jindřich Marek](#) (ÚISK), [Pavel Vondříčka](#) (ÚČNK), [Anna Chromá](#) (ÚČJTK) a [Martina Vacková Reiterová](#) (OIS)

Tajemníkem CDH je: [Ondřej Fúsik](#)

Pokud máte nějaké otázky týkající se činnosti CDH či členství, neváhejte se na nás obrátit e-mailem: cdh@ff.cuni.cz.