



Central European Institute of Technology
BRNO | CZECH REPUBLIC

David Potěšil
Kristína Gömöryová

Anna Schneiderová
Pavel Hruška
Ranjani Ganji

2019-11-20

Processing of proteomics data in KNIME

MUNI



Workshop outline

- morning session – theoretical part
 - 10:00 – 10:15 Opening and introduction
 - 10:15 – 10:45 Software container running KNIME
 - 10:45 – 11:00 Coffee break
 - 11:00 – 11:30 Introduction to KNIME
 - 11:30 – 11:45 Coffee break
 - 11:45 – 12:30 Practical applications, our KNIME metanodes
 - 12:30 – 13:30 Lunch break, visit of our laboratories for interested people

Workshop outline

- afternoon session – practical part
 - 13:30 – 16:30 KNIMing and coffee breaks
 - 16:30 – 18:00 Discussion

Workshop organization remarks

- morning session
 - presentation slides will be provided
 - keep your questions to the end of each presentation
- afternoon session
 - virtual workspaces provided to you by us will be kept for you till 31st January 2020
 - changes in the network settings will be necessary though (different IP::port settings)
 - let us know if you would need prolongment
 - data and workflows presented will be provided

Workshop organization remarks

- everyone has working WiFi connection?
- who does not have working environment accessible and wants to have one?

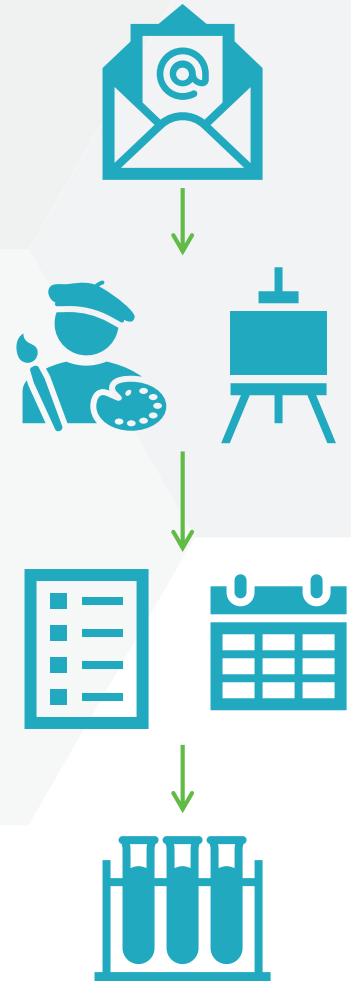
Workshop outline

- morning session – theoretical part
 - 10:00 – 10:15 Opening and introduction
 - **10:15 – 10:45** **Software container running KNIME**
 - 10:45 – 11:00 Coffee break
 - 11:00 – 11:30 Introduction to KNIME
 - 11:30 – 11:45 Coffee break
 - 11:45 – 12:30 Practical applications, our KNIME metanodes
 - 12:30 – 13:30 Lunch break, visit of our laboratories for interested people

0) General proteomics study

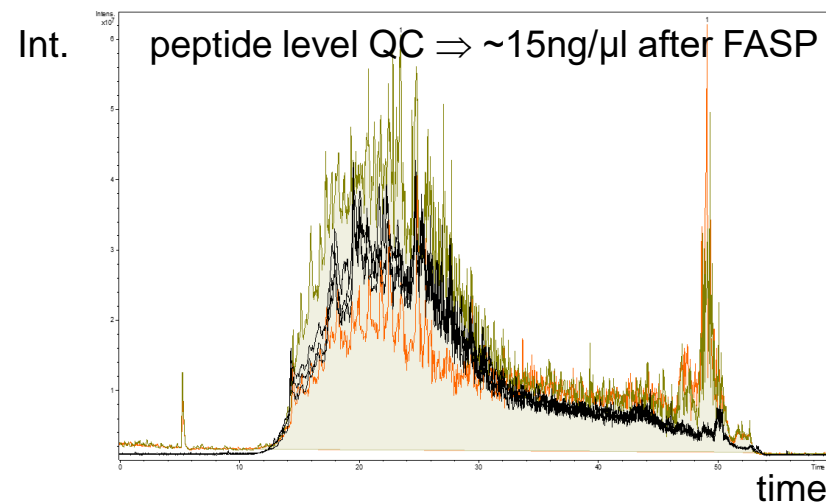
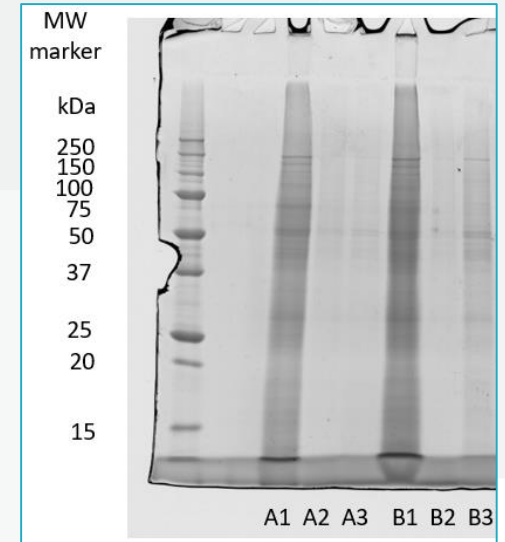
General proteomics study in our core facility

- **email** in our inbox
 - “Hello, we are studying very interesting **protein A** and would love to know **how it will look like if we will get rid of it in our cell line.**
- **discussion** about the study, **experiment design** specification
 - two types of cell line
 - **WT** – wild type
 - **KO** – protein A knocked out
 - **4 replicates**
 - prepared all in parallel
- samples **processing plan** on our side, **instrument booking**
- samples **preparation on the customer side**



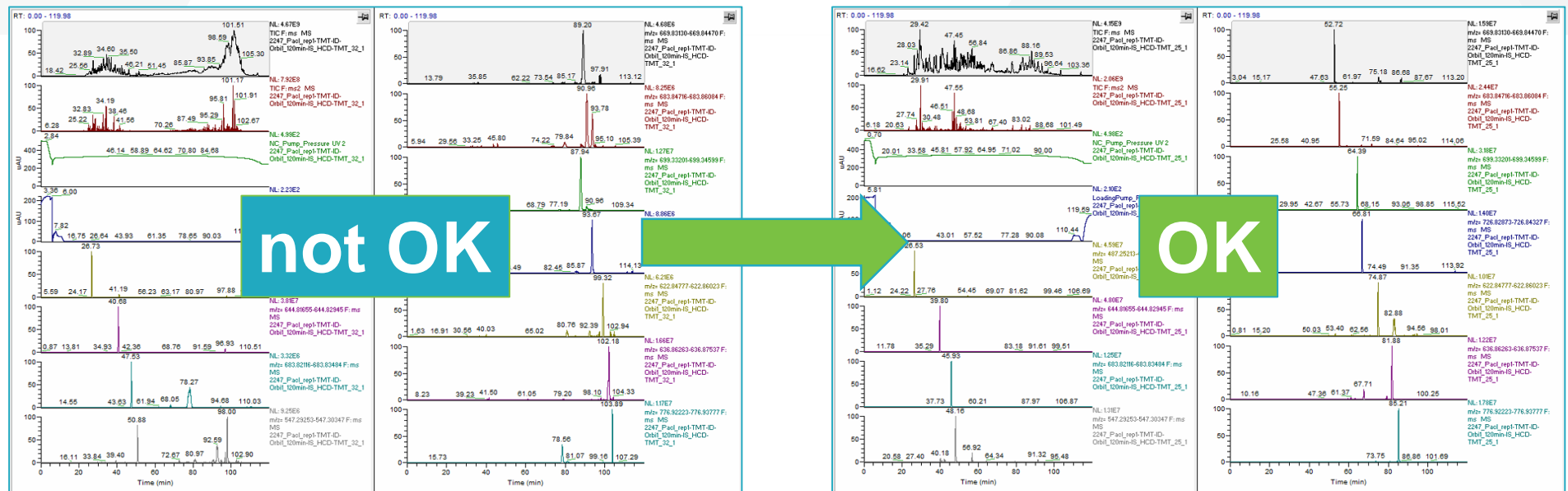
General proteomics study in our core facility

- samples **processing in our laboratory**
 - e.g. cells lysis (proteins solubilization)
 - **protein level quality control (QC)** step (1D SDS-PAGE; semi quant.)
 - FASP (proteins \Rightarrow peptides)
 - **peptide level QC** (LC-UV-MS; semi quant.)
 - peptides transfer into the LC-MS vial prior the LC-MS/MS measurement



General proteomics study in our core facility

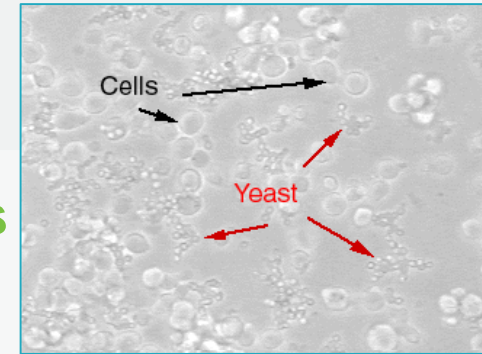
- final **LC-MS analyses** of the resulting peptide mixtures
 - potential issues due to samples matrix and or technical problems with the used instruments
 - e.g. residual detergents affecting the peptides LC separation, “dirty” mass spectrometer
 - iRT injected together with each sample, checked for retention and intensity profiles
- partial samples reprocessing might be sometimes necessary with another round of the LC-MS analyses



General proteomics study in our core facility

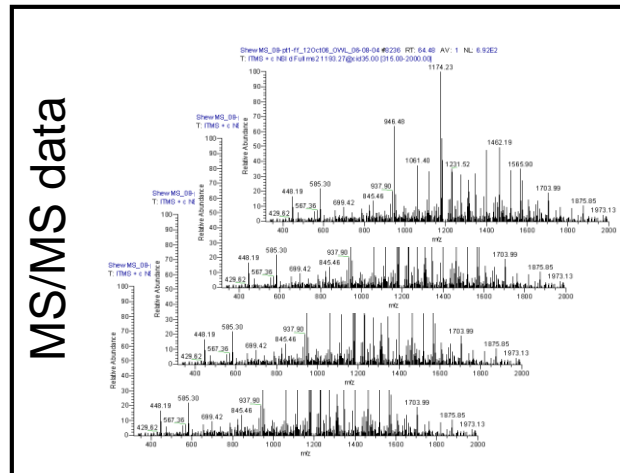
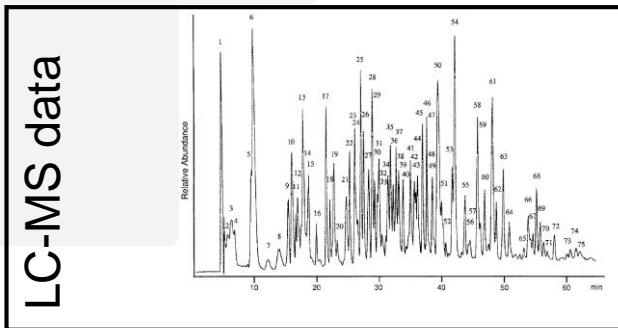
- **QC database searches**

- potential issues with **protein type contaminations**
 - e.g. FBS, bacterial contamination
 - QC database searches to check on any potential samples contamination
- potential issues with **not expected peptide or protein level modifications**
 - e.g. partially digested sample (not specific peptides)
 - QC database searches to check presence of not expected modifications
- adjustment of the final database search conditions if needed
- **sometimes critical issue** \Rightarrow new samples preparation



General proteomics study in our core facility

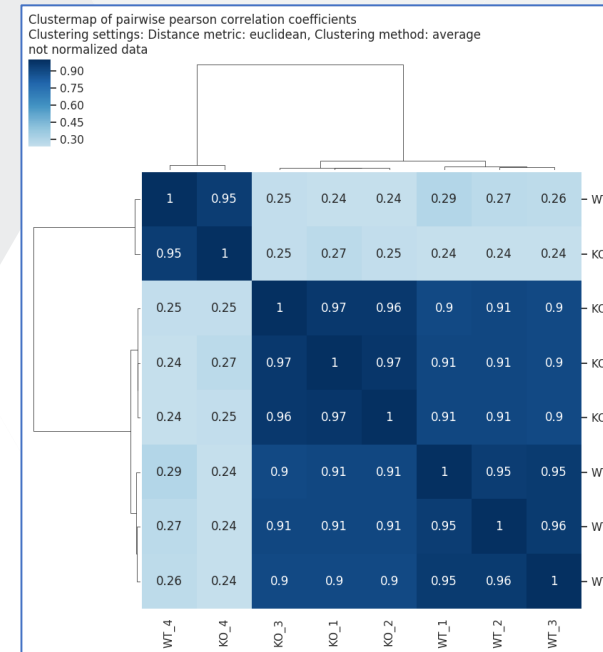
- final **database searches** and **protein quantification**
 - peptides identification, proteins list generation and peptides and proteins quantification
 - including LC-MS signal, not MS/MS based quant., e.g. MBR
 - more and more complex approaches
 - potential issues** affecting the results worth to check
 - varying or generally too low identification rates
 - separation specific issues (varying peptides LC-MS peaks widths and or retention times)



PG	Accession	Protein intensity					
		A1	A2	A3	B1	B2	B3
1	P12345	20	18	19	12	15	13
2	P23456	28	24	23	0	0	0
3	P34567	18	17	15	16	19	18
4	P45678	0	0	0	0	0	0
5	P56789	0	0	0	20	24	20
...

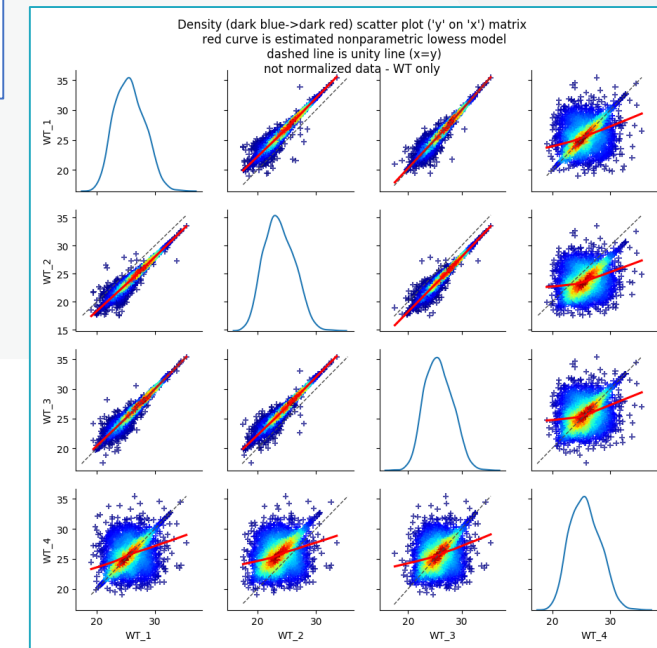
General proteomics study in our core facility

- proteins **quantification data preprocessing**
 - proteins list filtering
 - transformation, normalization
 - missing data imputation
- + another level of QC steps (cluster analysis, correlation plots inspection, ...)
 - get to know the data prior statistics...**
- potential issues** observed/affecting the results
 - not expected design “features” – **paired design, batch effects**
 - outlying sample replicates**



cluster analysis

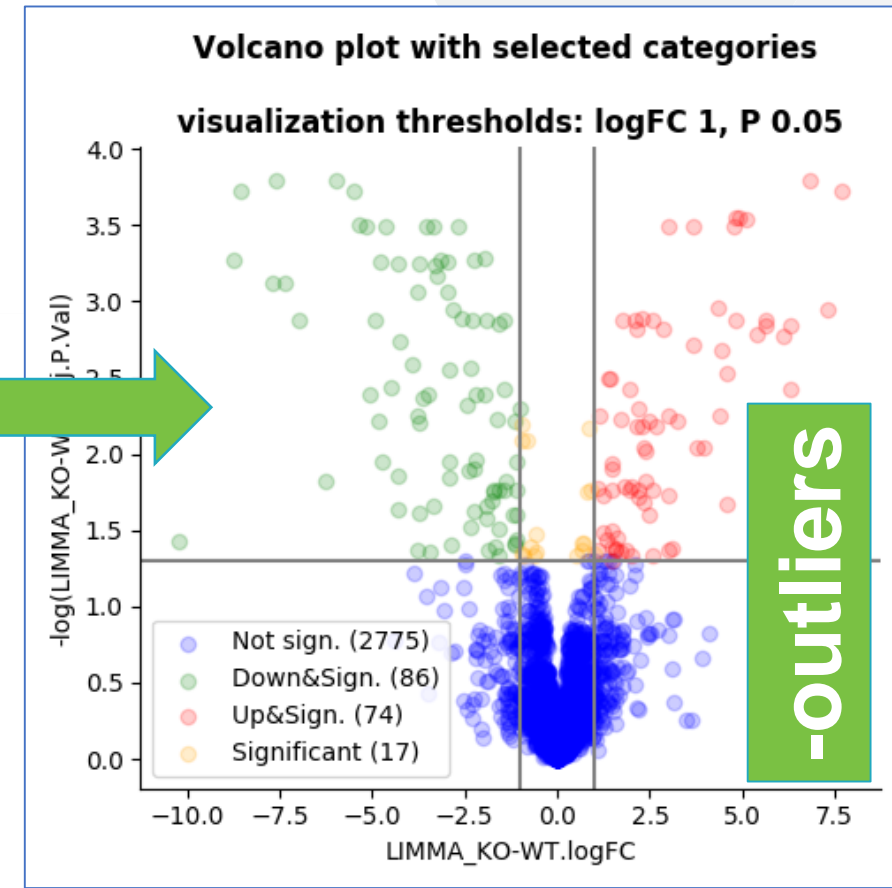
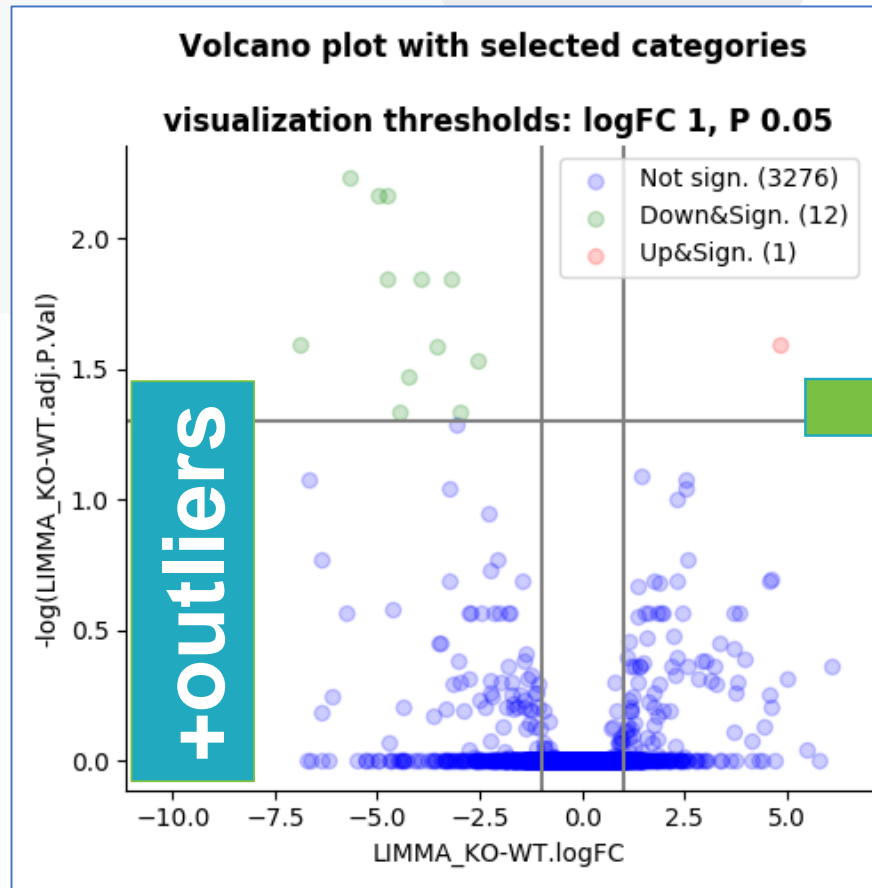
samples protein
intensities correlations



General proteomics study in our core facility

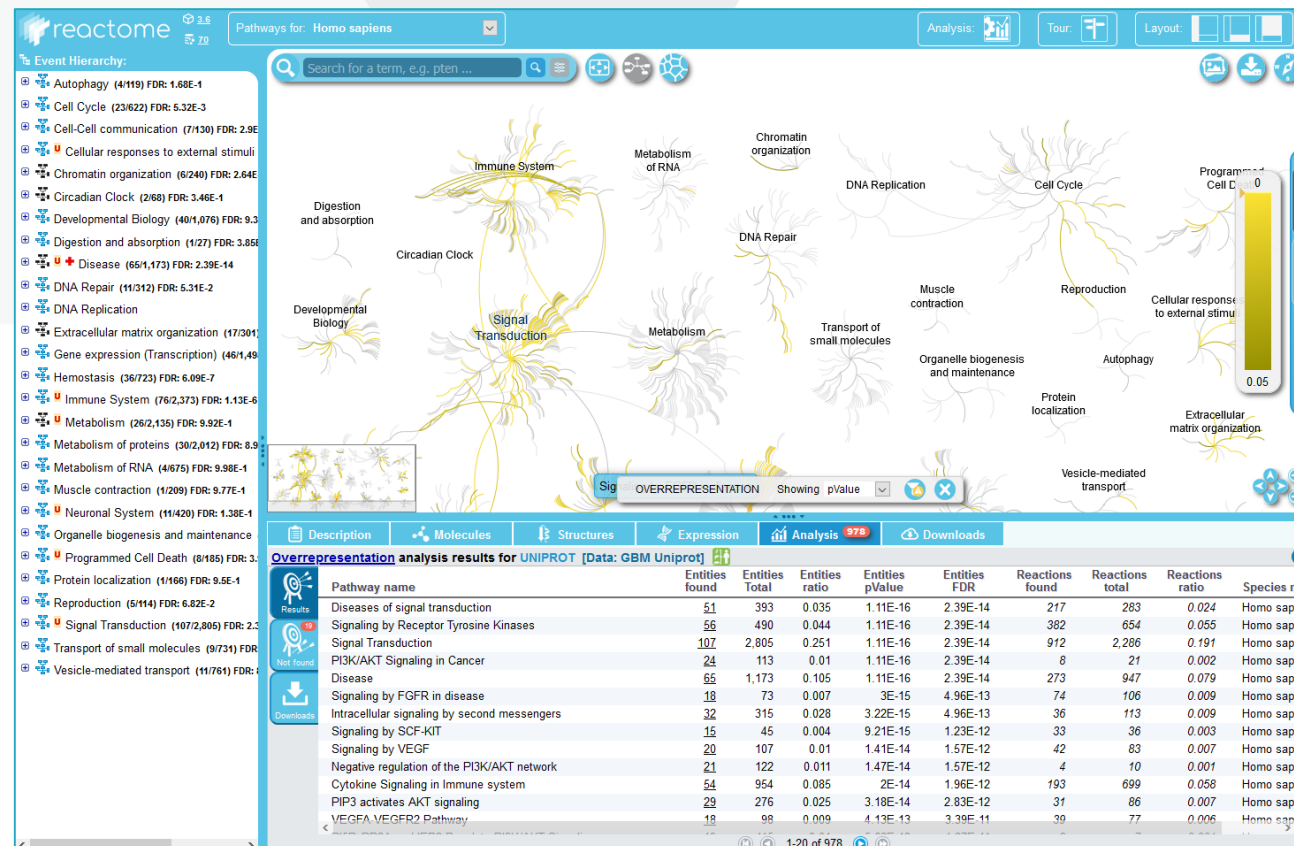
- **statistical data analysis**

- e.g. LIMMA test (moderated t-test)
- importance of proper **experimental design**, **outliers exclusion**, ...



General proteomics study in our core facility

- **initial** proteins list **interpretation**
 - using whole protein list or just candidate proteins list
 - e.g. pathway analysis using Reactome



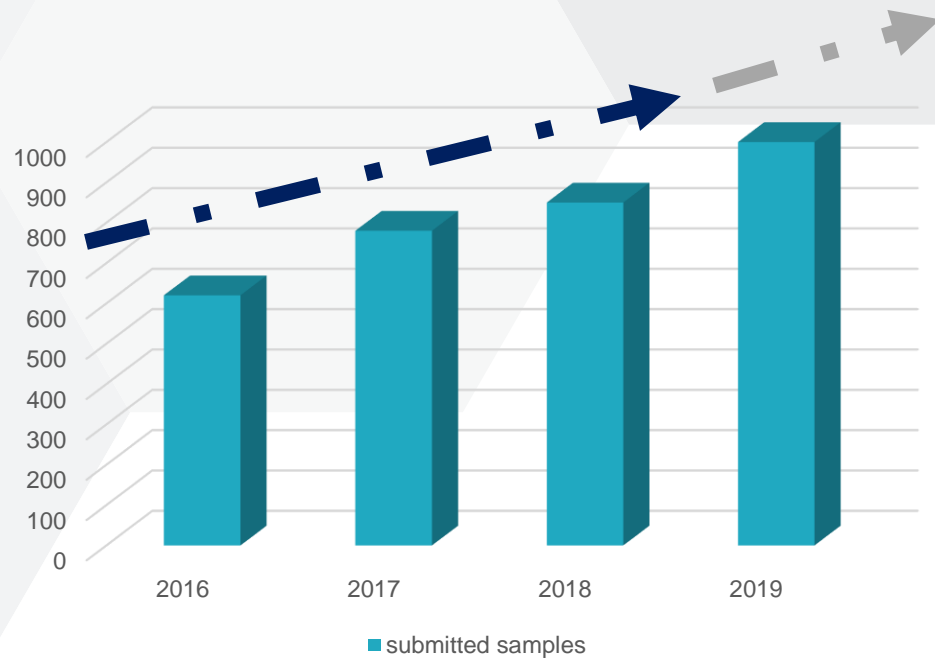
General proteomics study in our core facility

- study **results discussion** with the customer
- study **results verification**
 - mainly on the customer side, with our assistance if needed
 - MS data **re-evaluation and finalization** sometimes needed
 - including e.g. MS data database re-searching
 - combination with other omics data
 - combining with another set of samples prepared
 - **even several years long process...**
- study **publication**



General proteomics study in our core facility

Number of LC-MS/MS samples (04/2016 – 11/2019)

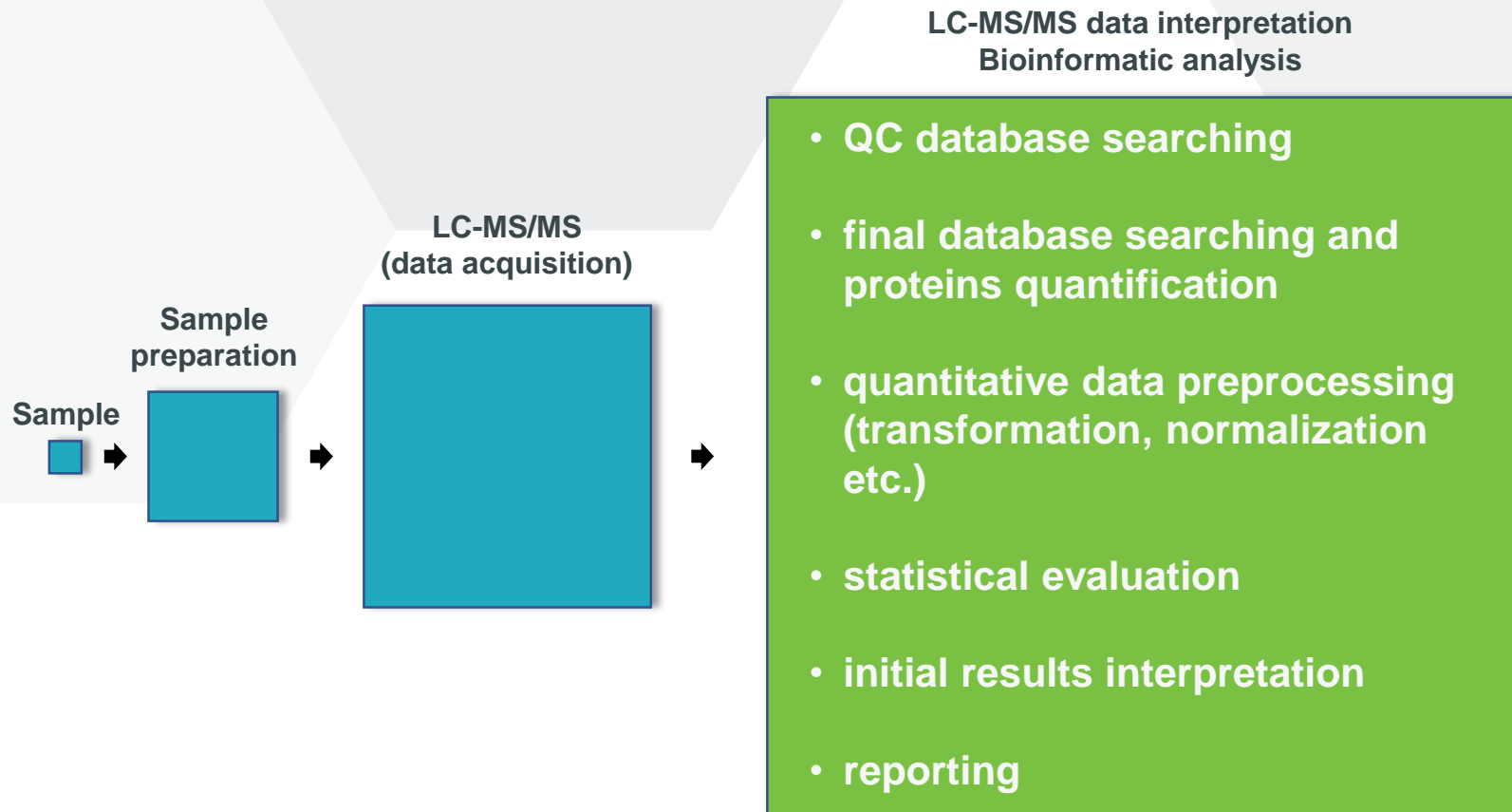


CIISB project - 10 Core facilities



funding of our CF

General proteomics study in our core facility - overview



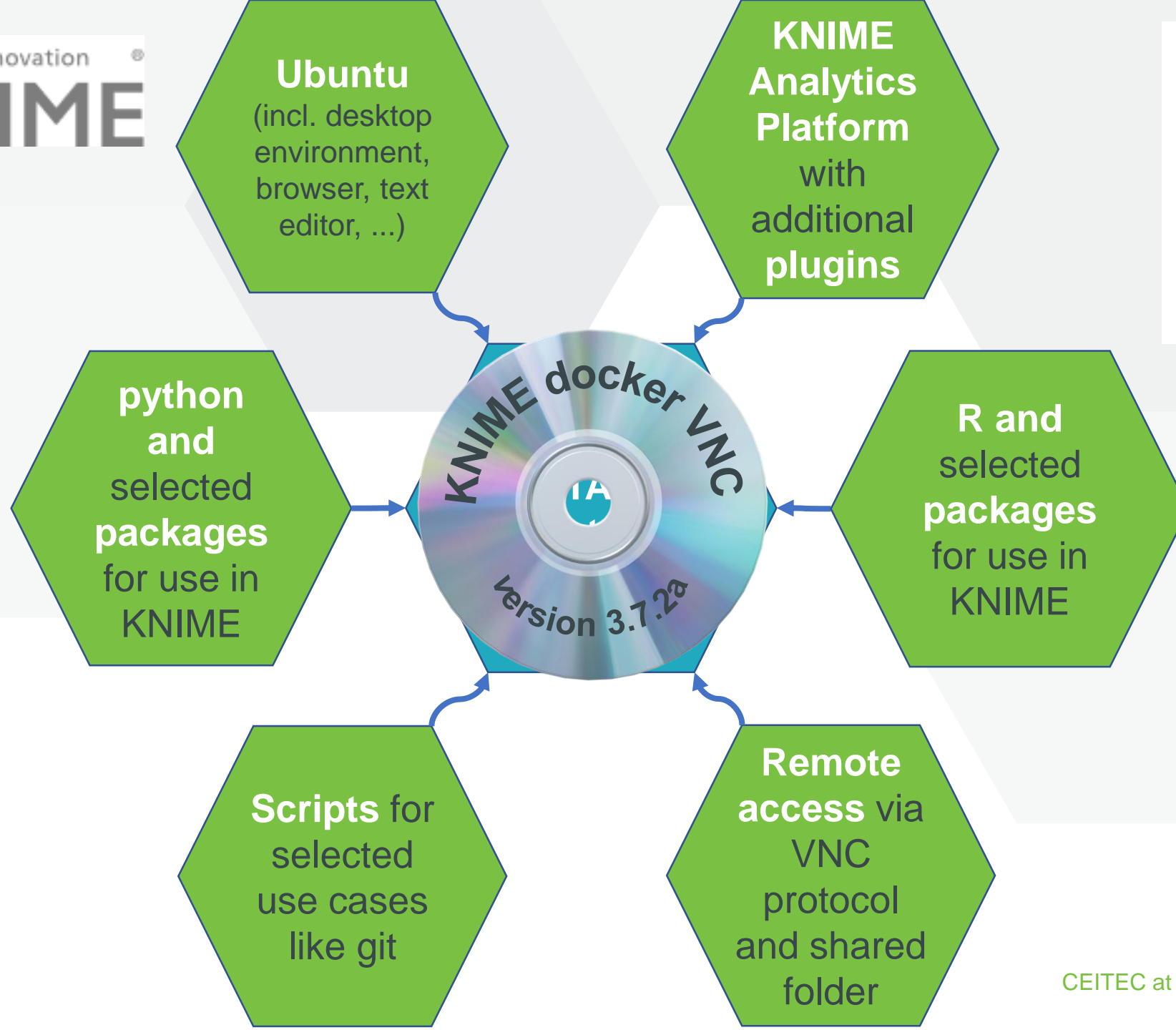
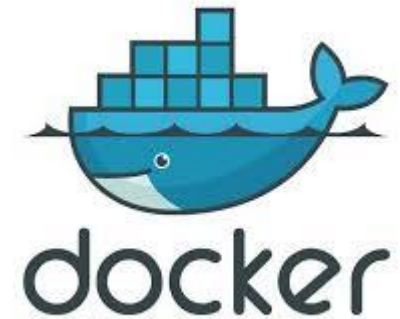
1) Software container running KNIME

Reasons to create such environment?

- more and more **complex** data processing and visualization **steps needed and used**
- **combining individual steps** into **complex pipelines**
 - **flexibility** of the processing ways necessary
 - one approach not generally applicable for all datasets, even though with similar concept
 - multiple settings tested, benchmarked and considered
- to have **easy to be used environment** without any scripting necessary
- to have **reproducible environment**, yet with up to date techniques, i.e. **versioning** needed
 - use identical environment on many installations covering also e.g. R, python specific versions
 - **older versions easily accessible** if needed even after years
- **revisiting the old pipelines** needed (publication, candidates verification, ...)
 - detailed **documentation** of the used tools/steps with **concrete settings**
 - single processing step settings details having major effect on the results!
 - including all the scripts used
- be able to **reproduce/reuse the older pipelines** on other data – **Don't Repeat Yourself**
- use, support and build our processing pipelines on **free and open-source** tools

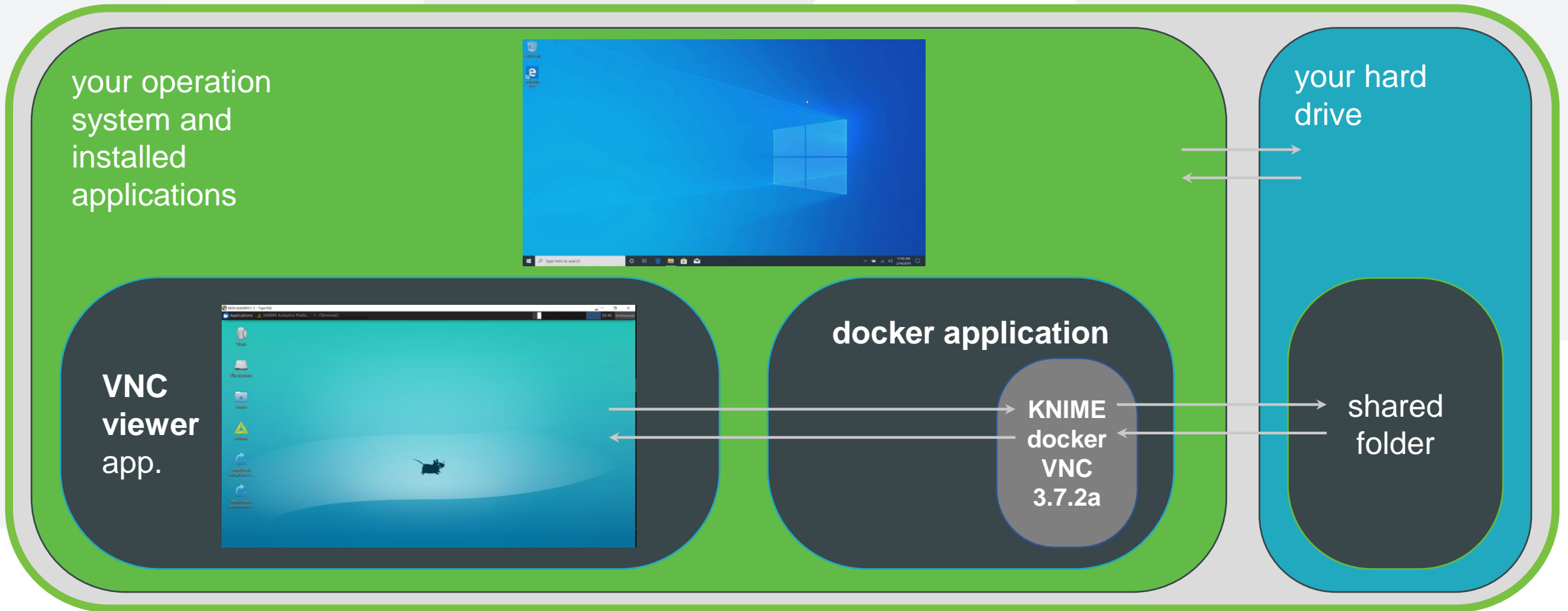
Potential tools/solutions for data processing

- selected tools/solutions
 - Proteome Discoverer – commercial, ready-to-use solution, limited functionality
 - MaxQuant + Perseus – free, but closed source (black boxes)
 - R studio/R console (DEP, MSnbase)
 - Jupyter notebooks – using python, R, or other scripting language
 - Galaxy (<https://galaxyproject.org>)
 - great community
 - less clear and user-friendly interface
 - harder to get new features in and script on the go



Running software container scheme – local version

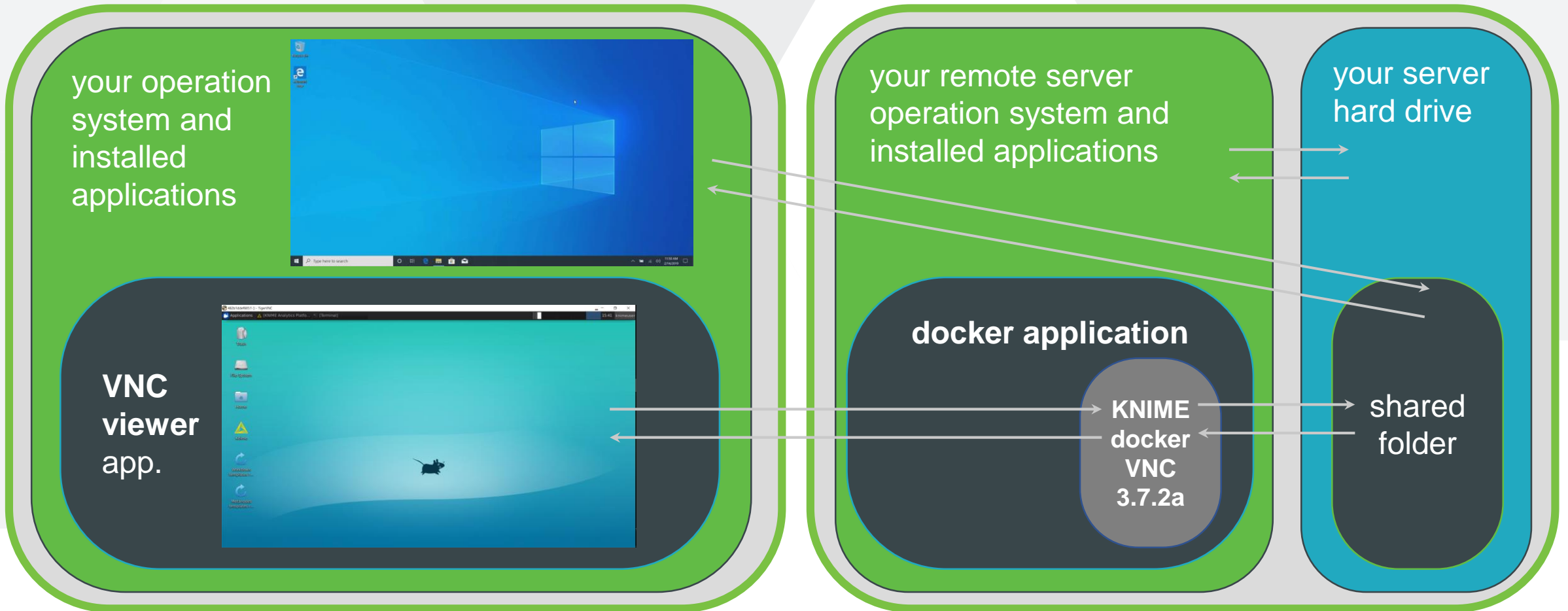
your PC (Windows, Linux) or Mac



Running software container scheme – server version

your PC (Windows, Linux) or Mac

your remote server



Requirements to run container – run locally

- hardware
 - **supporting virtualization** (you may need to enable it in the bios)
 - desktop or server; Mac hardware 2010 or newer
- operation system
 - **Linux** (64-bit, kernel ≥ 3.10)
 - **Mac** (macOS ≥ 10.12)
 - **Windows** (64-bit, Win 10, Pro, Enterprise or Education, Build 15063 or later)
- software
 - **docker application** installed
 - **VNC viewer** (TigerVNC viewer recommended)
 - **VirtualBox** can **NOT** be installed
 - **folder for sharing** with the environment



Requirements to run container – access to server

- hardware
 - **supporting virtualization** (you may need to enable it in the bios)
 - desktop or server; Mac hardware 2010 or newer
- operation system
 - **Linux** (64-bit, kernel ≥ 3.10)
 - **Mac** (macOS ≥ 10.12)
 - **Windows** (64-bit, Win 10, Pro, Enterprise or Education, Build 15063 or later)
- software
 - **docker application** installed
 - **VNC viewer** (TigerVNC viewer recommended)
 - **VirtualBox** can **NOT** be installed
 - **folder for sharing** with the environment – access only

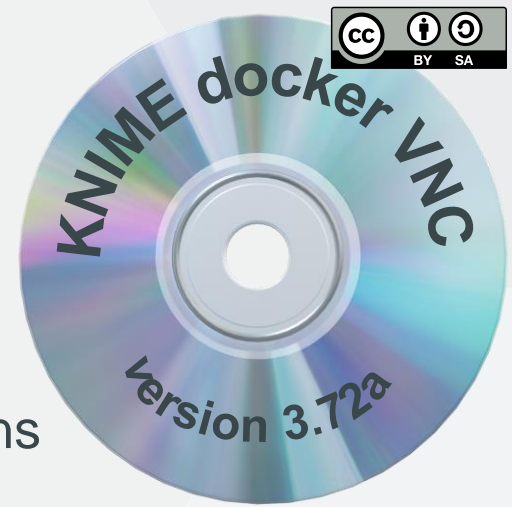


Software container availability

- GitHub repository with docker files and associated files + scripts
 - to know how the environment has been build and what components it contains
 - https://github.com/OmicsWorkflows/KNIME_docker_vnc



- prebuild, ready to be used docker images on Docker Hub
 - to use the environment directly (info on how to use it is on the GitHub repository)
 - <https://hub.docker.com/r/cfprot/knime>



Thank you
for your
attention

Workshop outline

- morning session – theoretical part
 - 10:00 – 10:15 Opening and introduction
 - 10:15 – 10:45 Software container running KNIME
 - **10:45 – 11:00 Coffee break**
 - 11:00 – 11:30 Introduction to KNIME
 - 11:30 – 11:45 Coffee break
 - 11:45 – 12:30 Practical applications, our KNIME metanodes
 - 12:30 – 13:30 Lunch break, visit of our laboratories for interested people