GCC 2018, 14th German Conference on Chemoinformatics

November 11-13, 2018, Mainz/Germany

PhenoMeNal

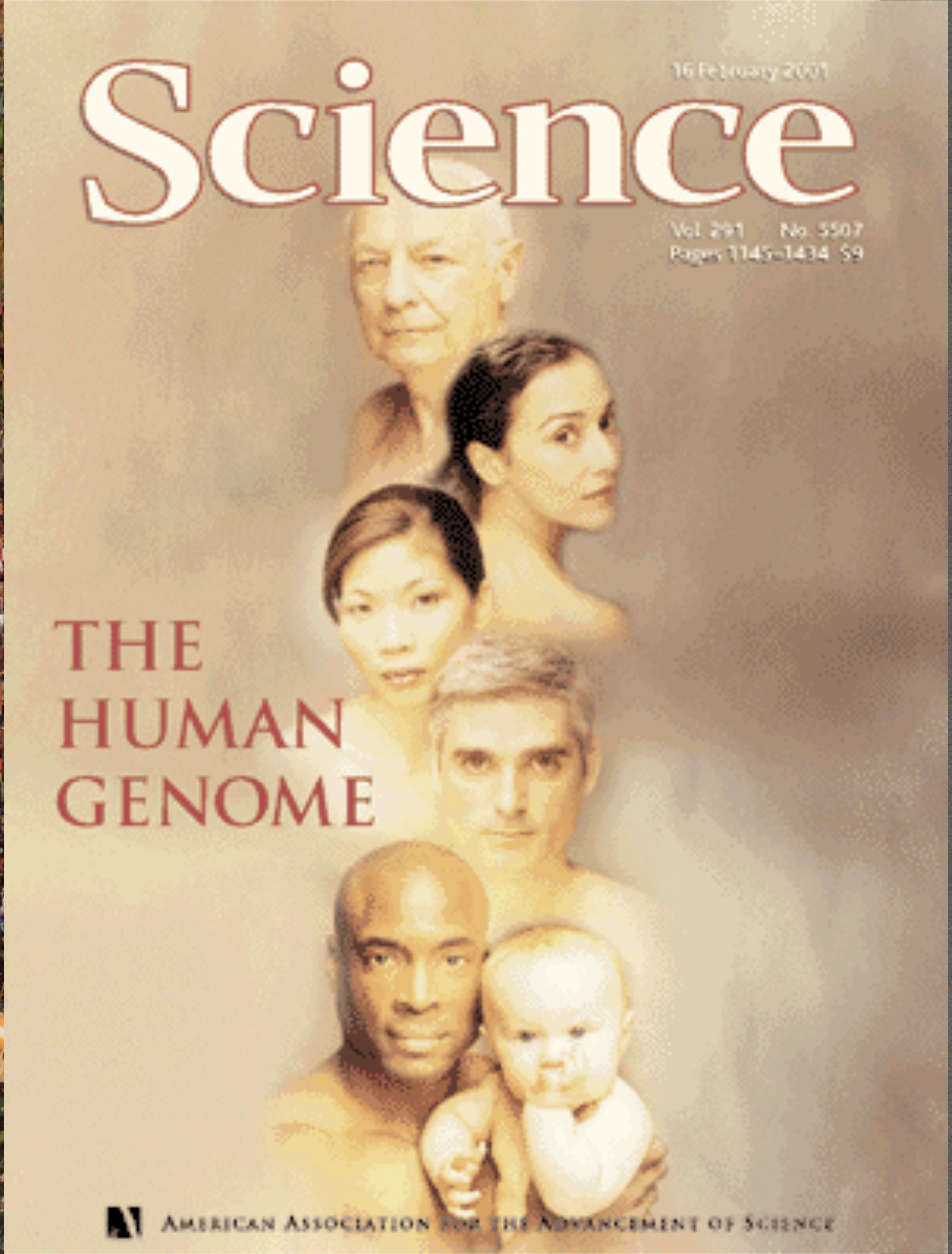
Interoperable and Scalable Metabolomics Data Analysis with Microservices

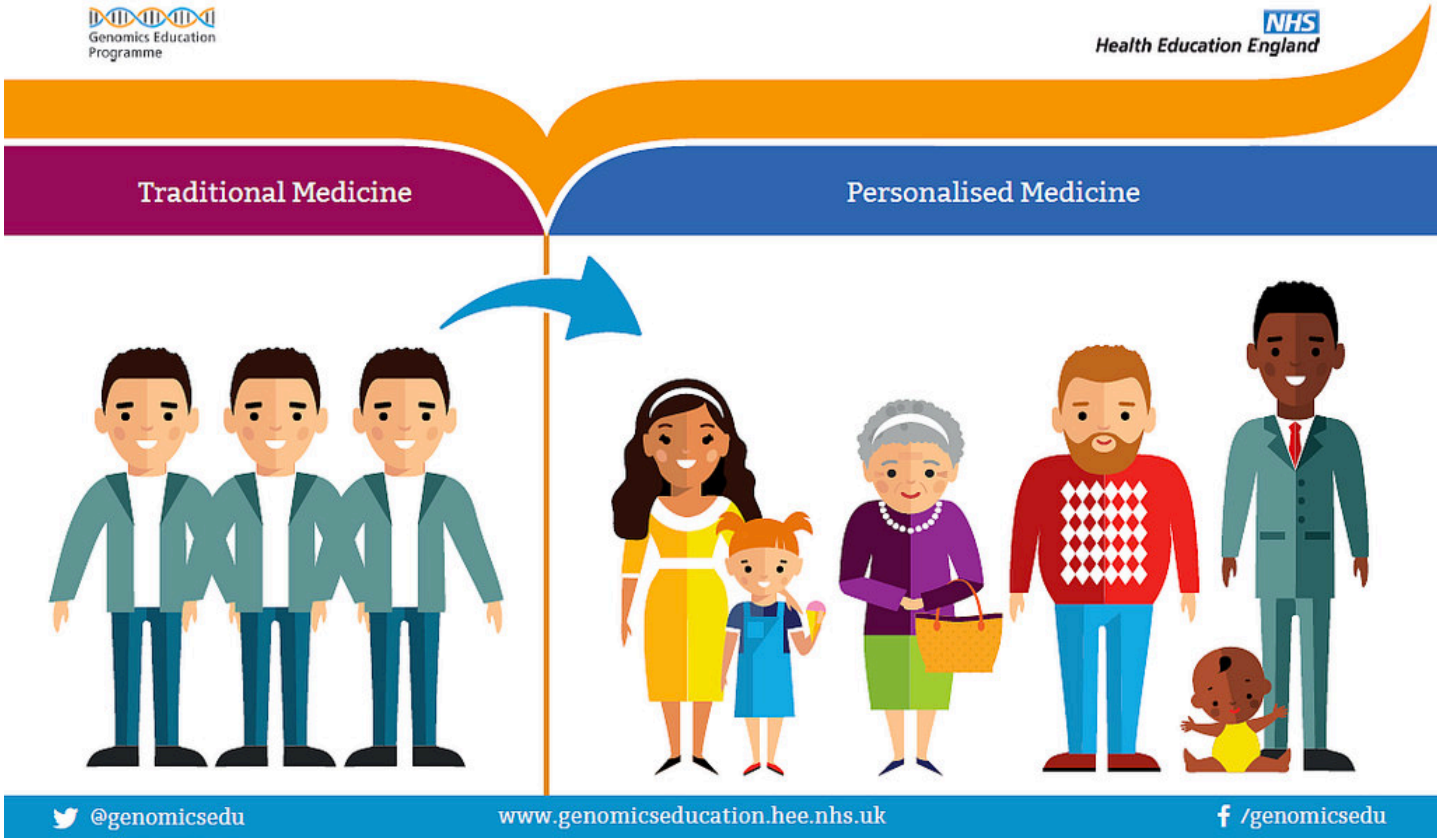
Christoph Steinbeck
and the
PhenoMeNal consortium

European Bioinformatics Institute, Hinxton
Friedrich-Schiller-University, Jena

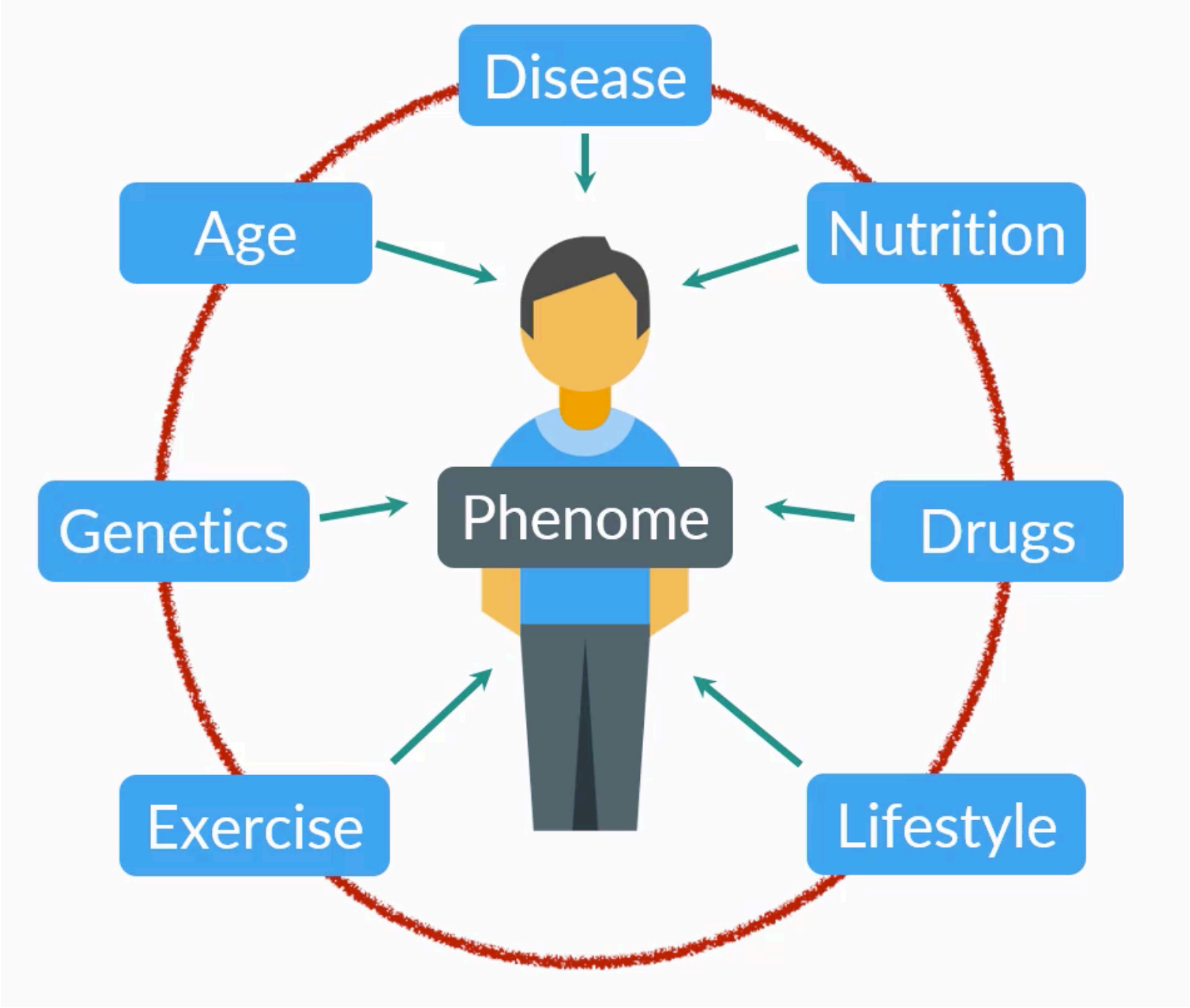
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

15 February 2001
nature
$10.00
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special


Science
16 February 2001
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
American Association for the Advancement of Science

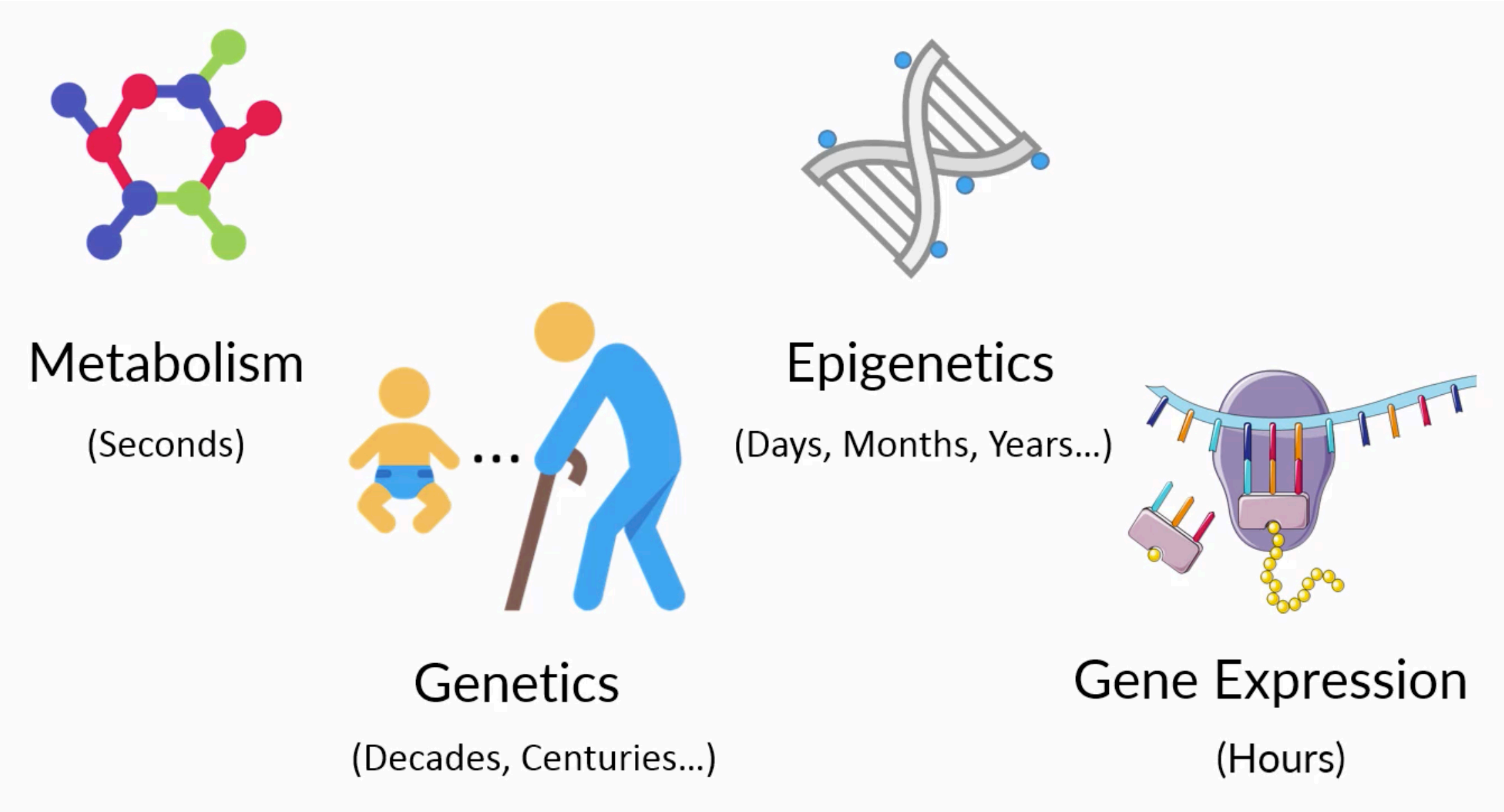

Genomics Education Programme
NHS
Health Education England
Traditional Medicine
Personalised Medicine
@genomicsedu
www.genomicseducation.hee.nhs.uk
/genomicsedu

Phenome - Exposome



FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Reaction times following external change



Metabolism

(Seconds)

Genetics

(Decades, Centuries...)

Epigenetics

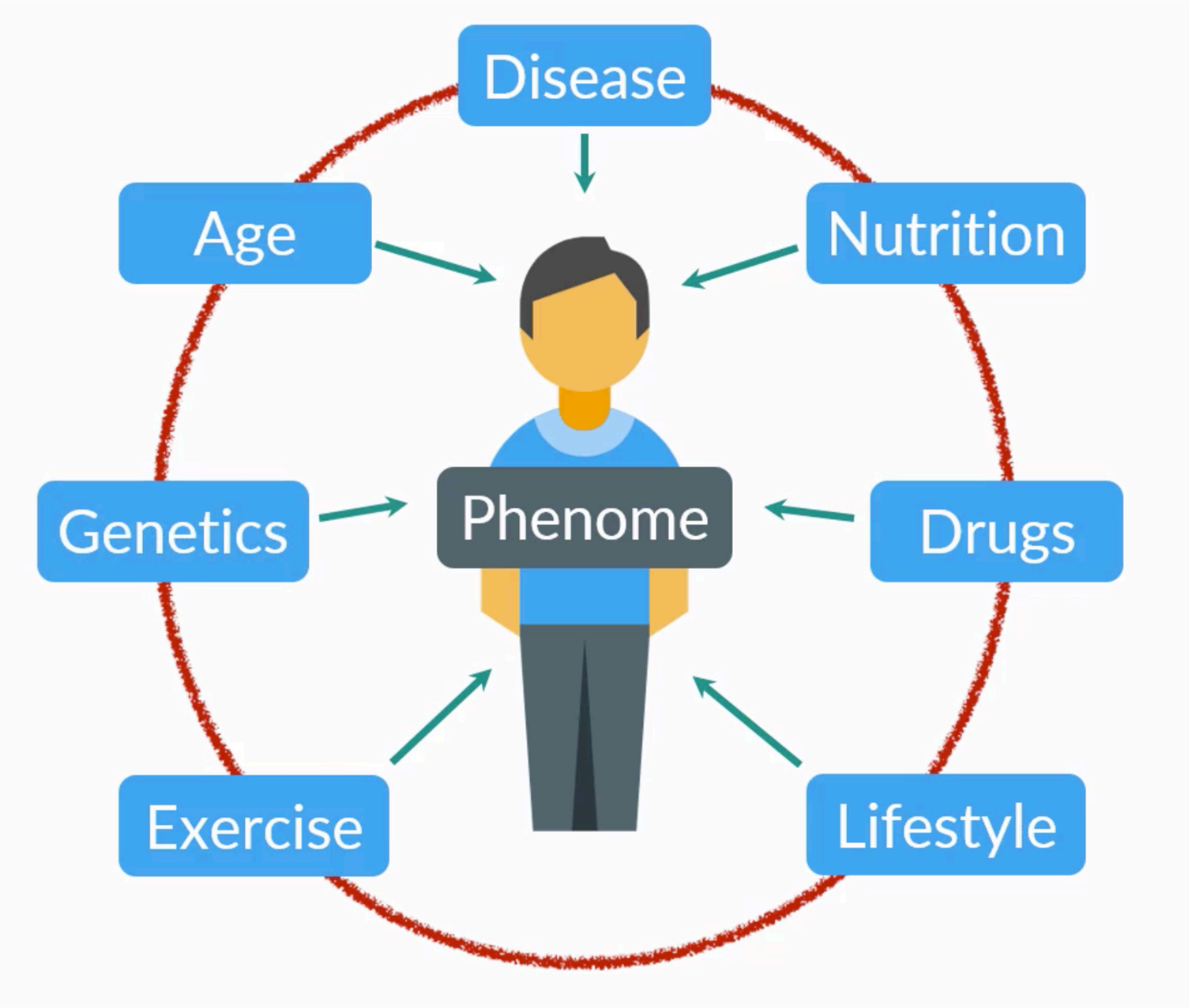
(Days, Months, Years...)

Gene Expression

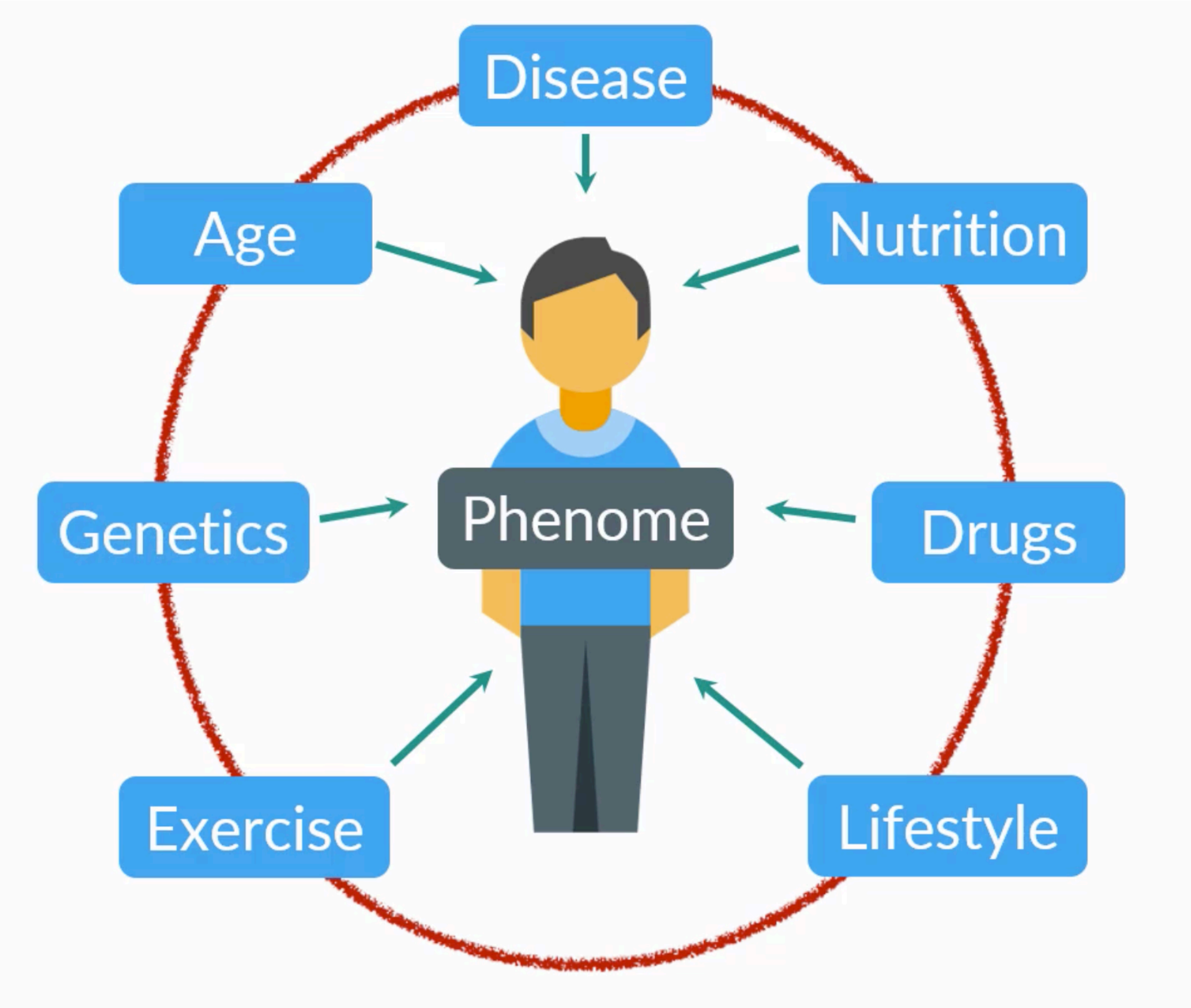
(Hours)

The Metabolome is an easily accessible and dynamically changing Molecular Phenotype

Phenome - Exposome



Phenome - Exposome

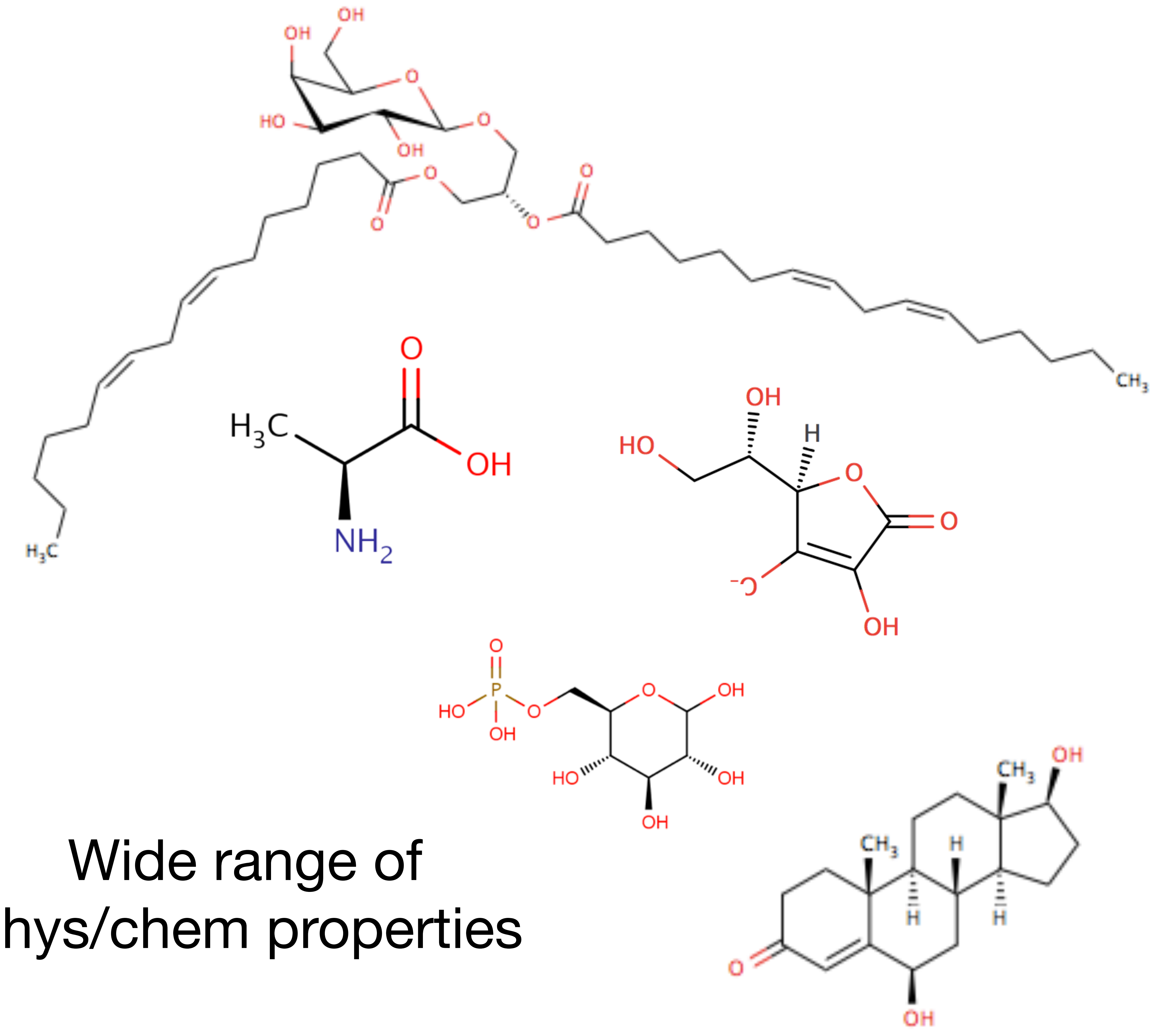


Big, well-annotated metabolomics **data** required to statistically link individual components of the **exposome** to effects in the **molecular phenotype**

Metabolomics

Metabolites: (Endogenous) small molecules in biological organisms

Measures **occurrence** and **concentrations** of many small molecules (**metabolites**) in an organism at once.



Wide range of phys/chem properties

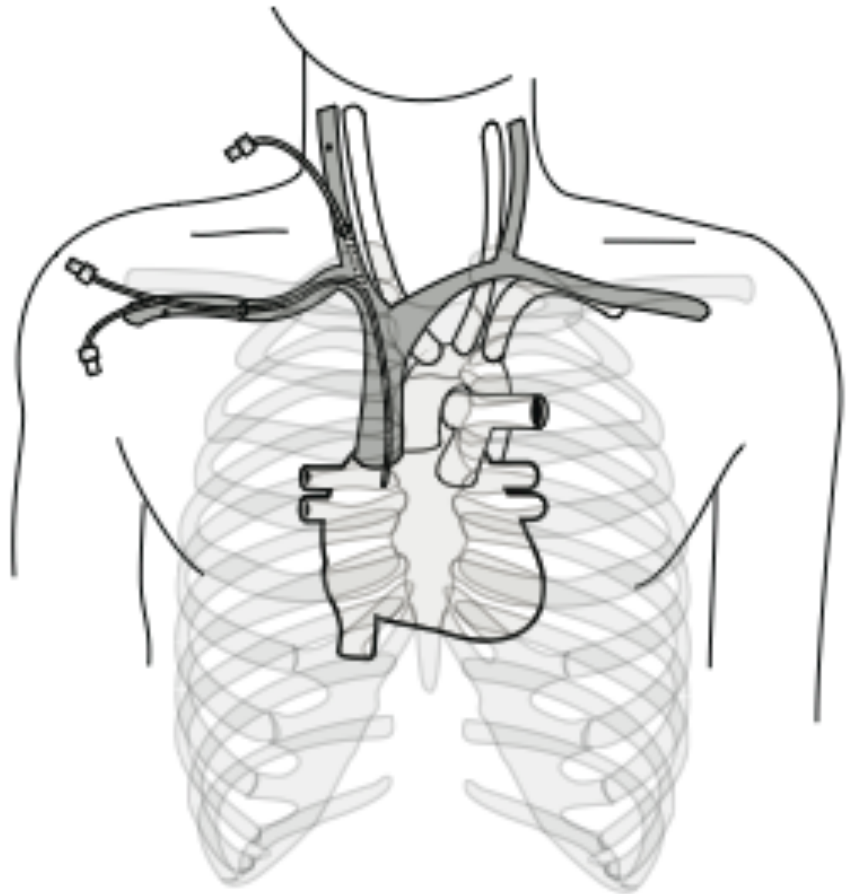
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Diagnostic fluids

Urine
(time-averaged data)

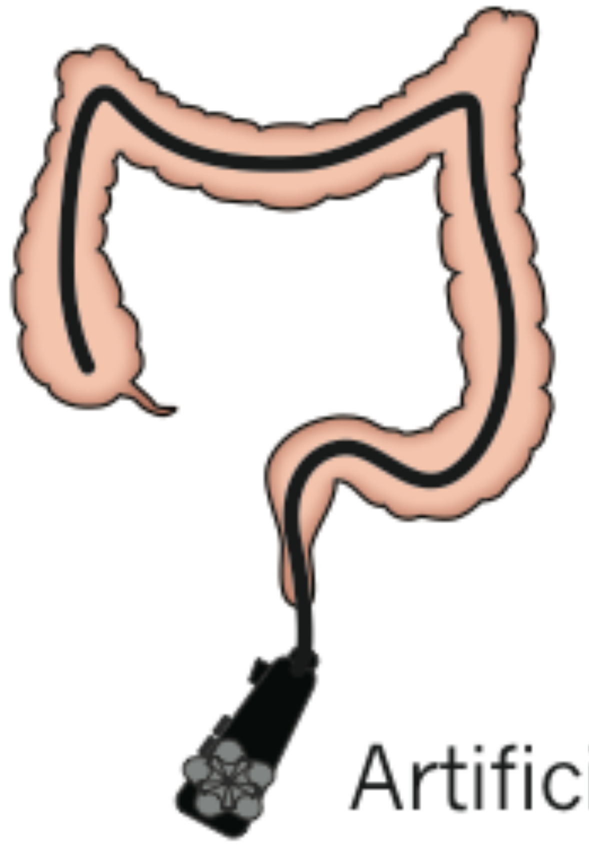
Plasma
(snap-shot data)

Other accessible analytical compartments



Specialized fluids and biopsies
(selected fluids)

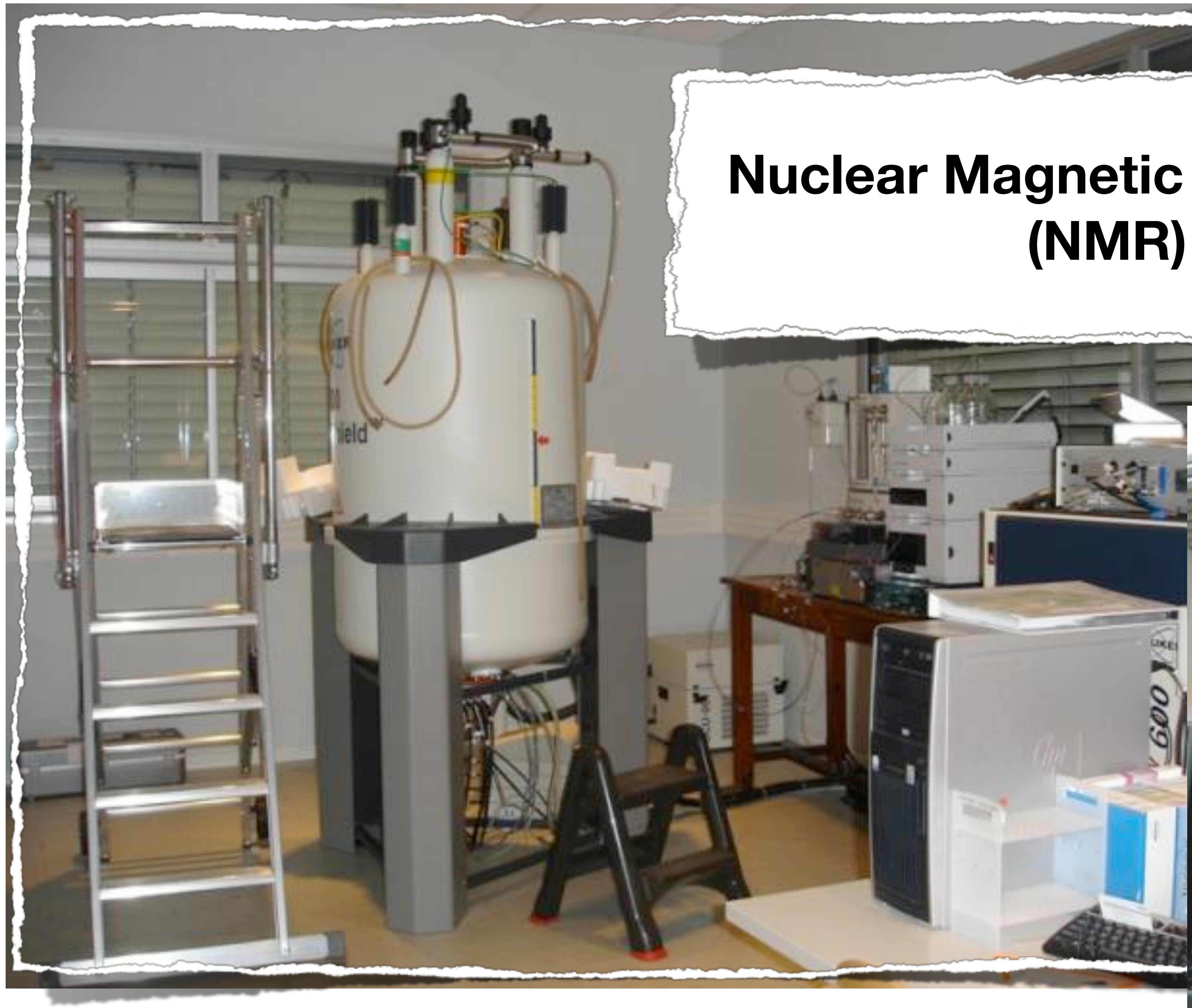
Pathological fluids



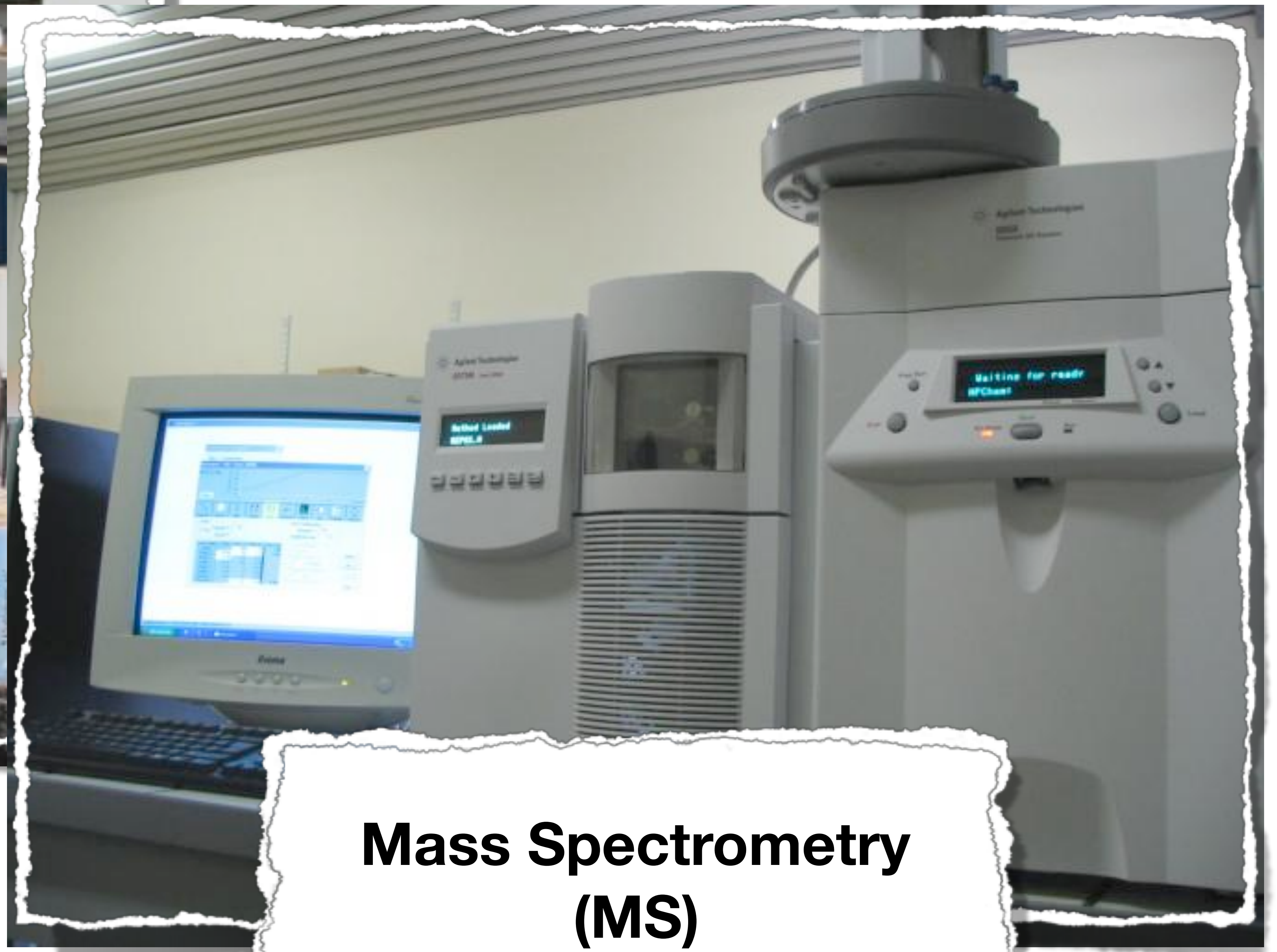
Artificial fluids

Nicholson et al., Nature, 491(7424), 384–392

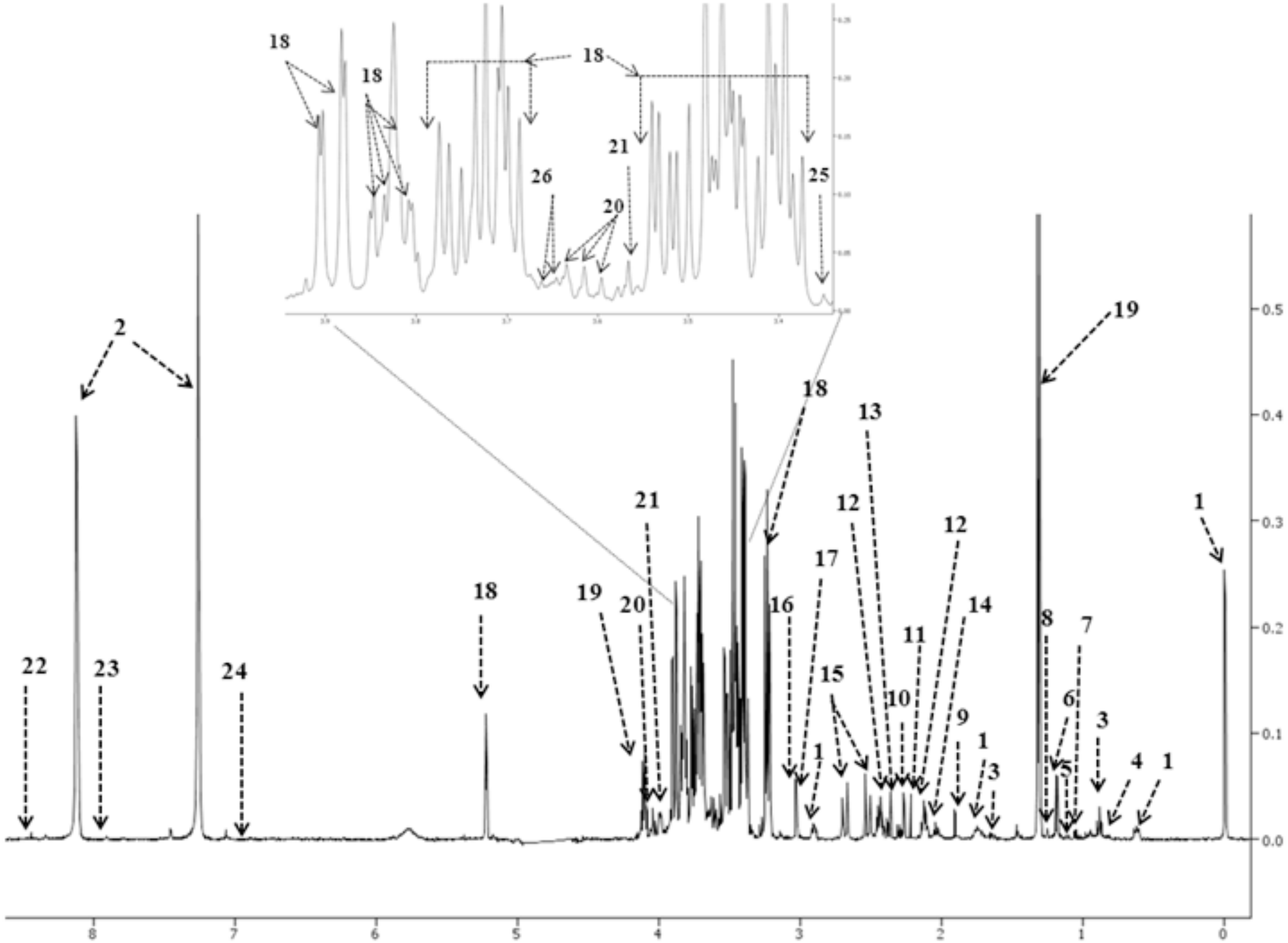
Metabolomics uses a wide-range of analytical techniques



Nuclear Magnetic Resonance (NMR)



Mass Spectrometry (MS)



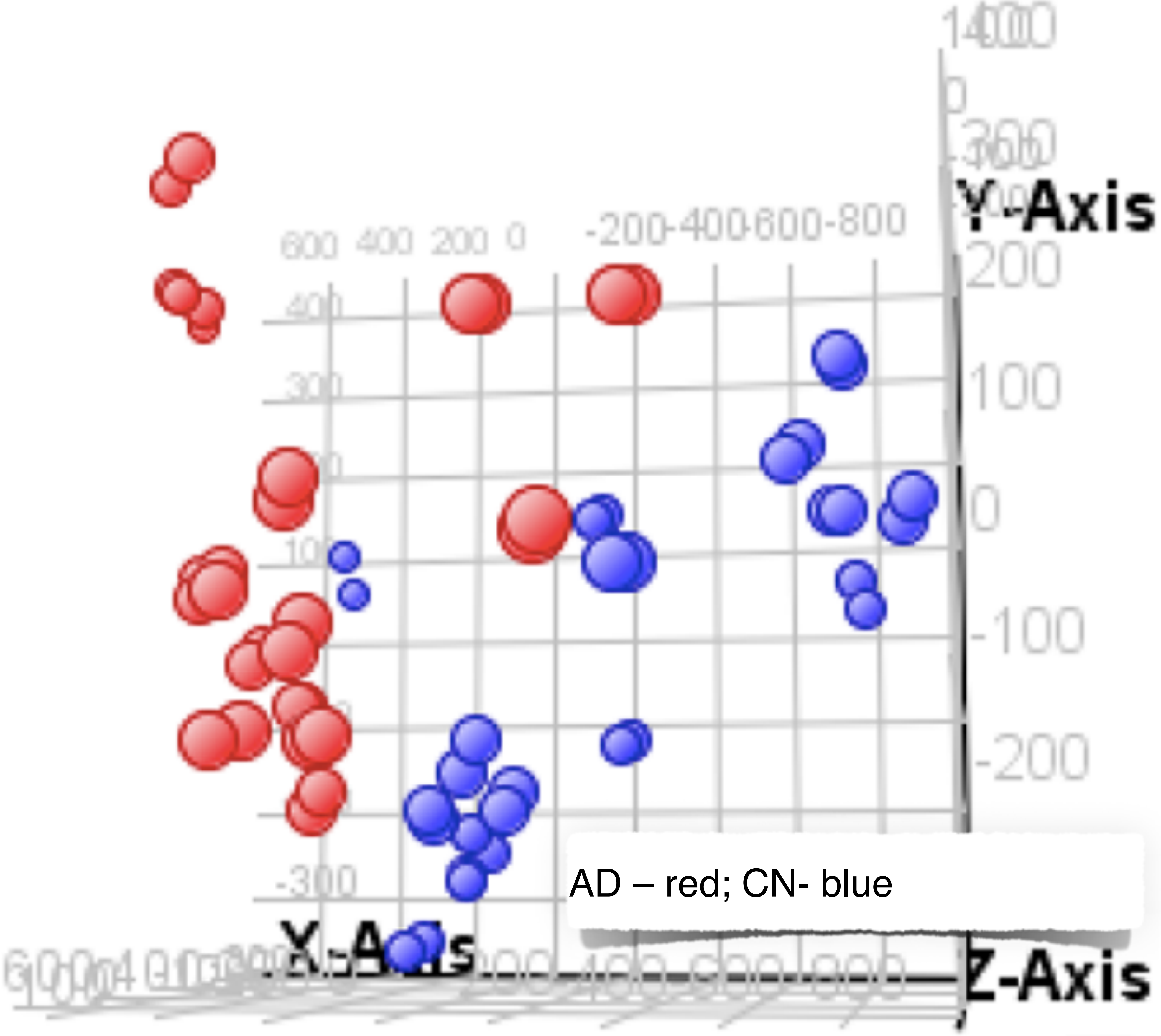
Typical 500 MHz ^{1}H-NMR spectrum of human cerebrospinal fluid. Numbers indicate the following metabolites:

1. DSS, 2. imidazole, 3. 2-hydroxybutyric acid, 4. 2-hydroxyisovaleric acid, 5. 2-oxoisovaleric acid, 6. 3-hydroxybutyric acid, 7. 3-hydroxyisobutyric acid, 8. 3-hydroxyisovaleric acid, 9. acetic acid, 10. acetoacetic acid, 11. acetone, 12. L-glutamine, 13. pyruvic acid, 14. L-glutamic acid, 15. citric acid, 16. creatinine, 17. creatine, 18. D-glucose, 19. lactic acid, 20. myo-inositol, 21. D-fructose, 22. formic acid, 23. L-histidine, 24. L-tyrosine, 25. methanol, 26. glycerol

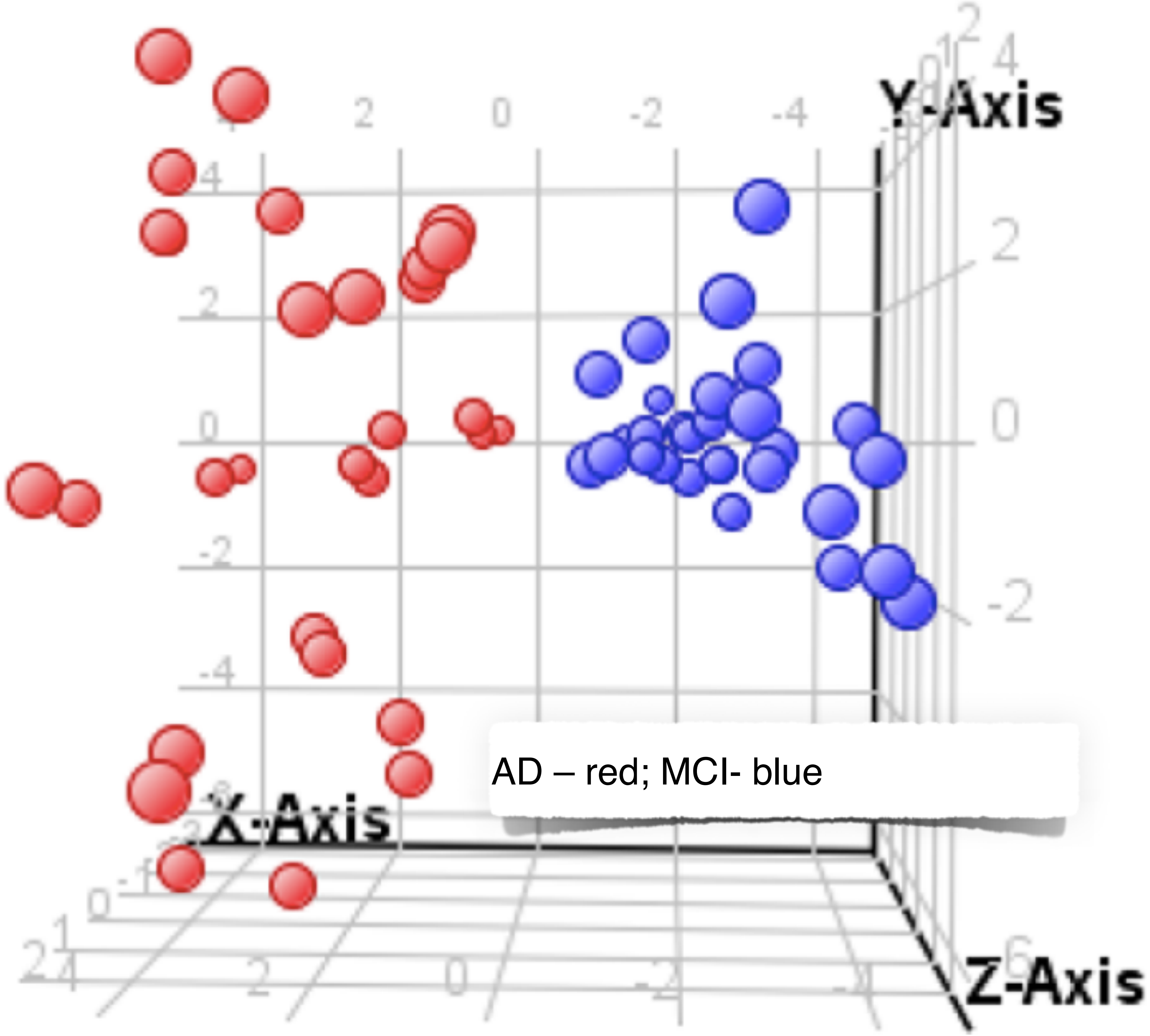
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

source: http://www.csfmetabolome.ca

Example: Untargeted Metabolomics of Cerebrospinal Fluid



AD – red; CN- blue



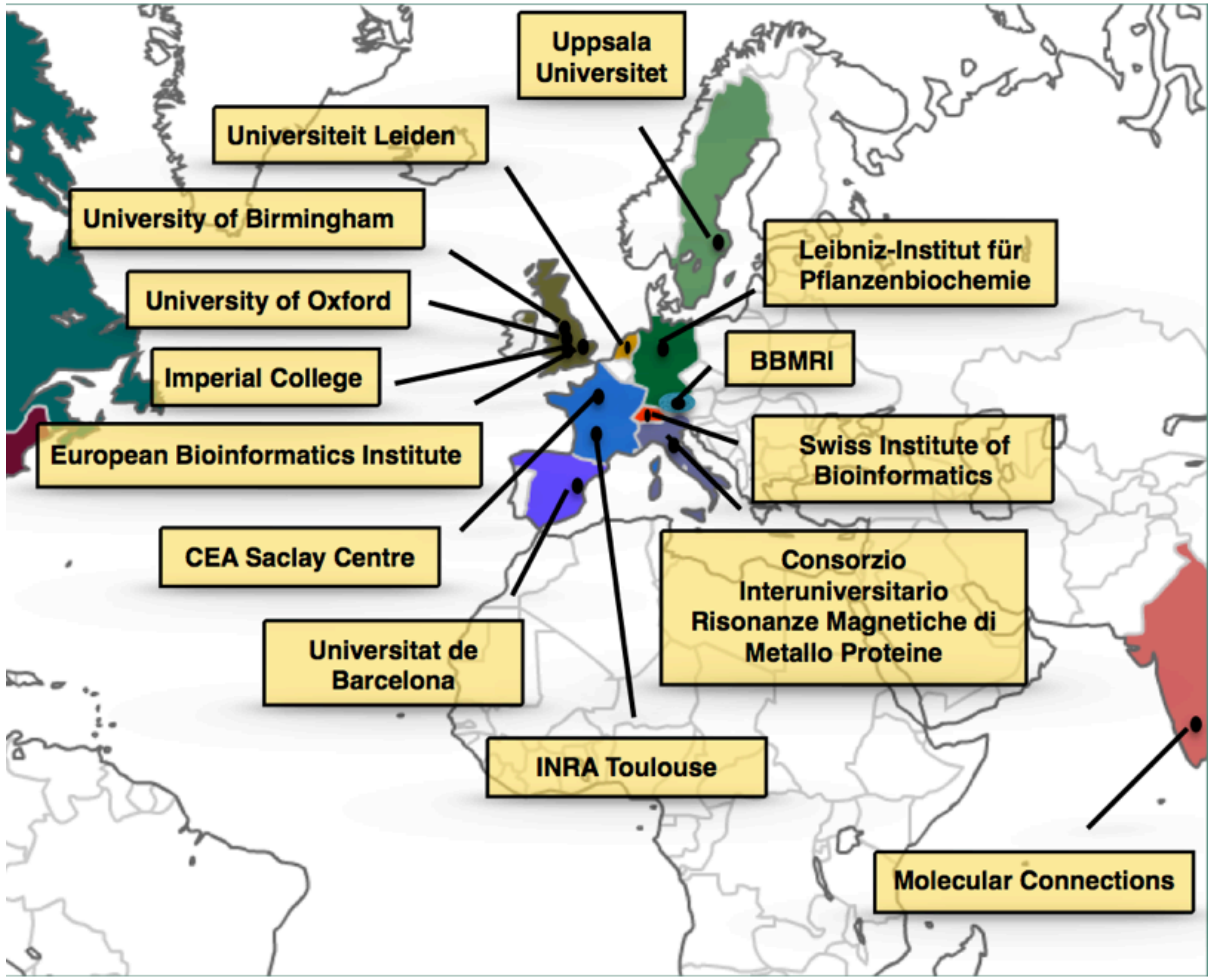
AD – red; MCI- blue

Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), Cognitive Normal (CN)

Trushina, et al. PLOS ONE 8, e63644 (2013).

PhenoMeNal

Large Scale Computing with Medical Metabolomics Data



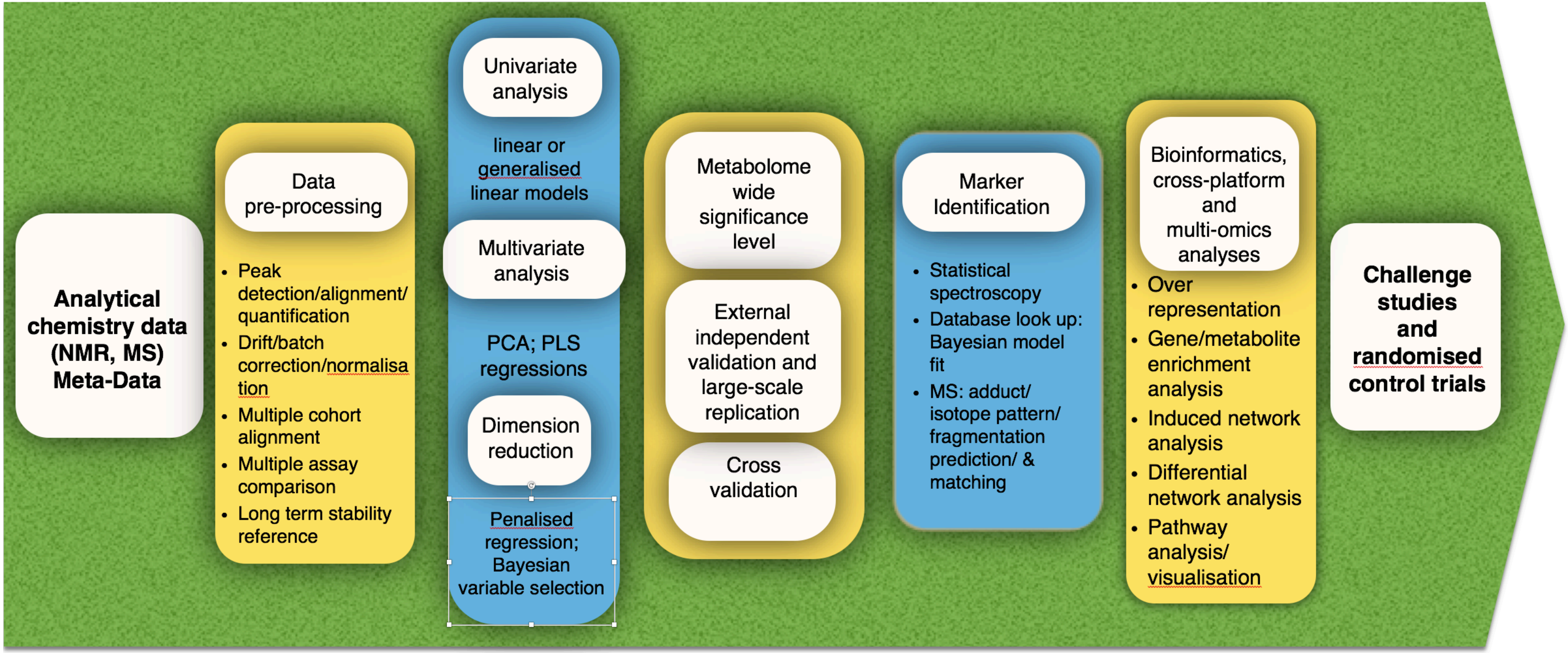
- H2020 e-infra
- 3 Years
- 13 Partners
- 8 Mio €
- 830 PM
- Kick-off 9/15

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

PhenoMeNal

An e-infrastructure for

Large Scale Computing with Medical Metabolomics Data



Analytical chemistry data (NMR, MS) Meta-Data

Data pre-processing

- Peak detection/alignment/ quantification
- Drift/batch correction/normalisation
- Multiple cohort alignment
- Multiple assay comparison
- Long term stability reference

Univariate analysis

linear or generalised linear models

Multivariate analysis

PCA; PLS regressions

Dimension reduction

Penalised regression; Bayesian variable selection

Metabolome wide significance level

External independent validation and large-scale replication

Cross validation

Marker Identification

- Statistical spectroscopy
- Database look up: Bayesian model fit
- MS: adduct/ isotope pattern/ fragmentation prediction/ & matching

Bioinformatics, cross-platform and multi-omics analyses

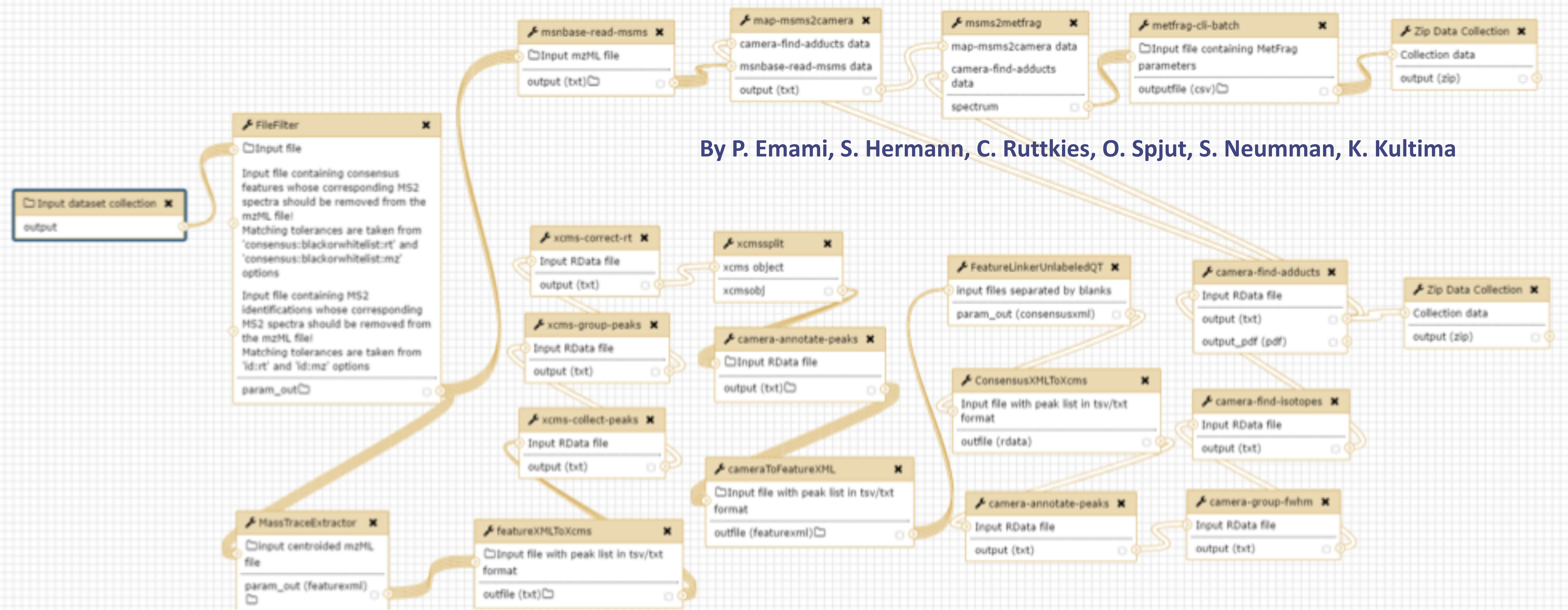
- Over representation
- Gene/metabolite enrichment analysis
- Induced network analysis
- Differential network analysis
- Pathway analysis/ visualisation

Challenge studies and randomised control trials

Microservices

- Architectural design pattern
- Independent, potentially distributed processes
- Language-independent interfaces.
- Services decoupled. Perform a small task (“Do one thing and do it well”).
- Individual service should be easy replaceable.

MS1/MS2 – XCMS-OpenMS



Input file containing consensus features whose corresponding MS2 spectra should be removed from the mzML file! Matching tolerances are taken from 'consensus:blackorwhitelist:rt' and 'consensus:blackorwhitelist:mz' options

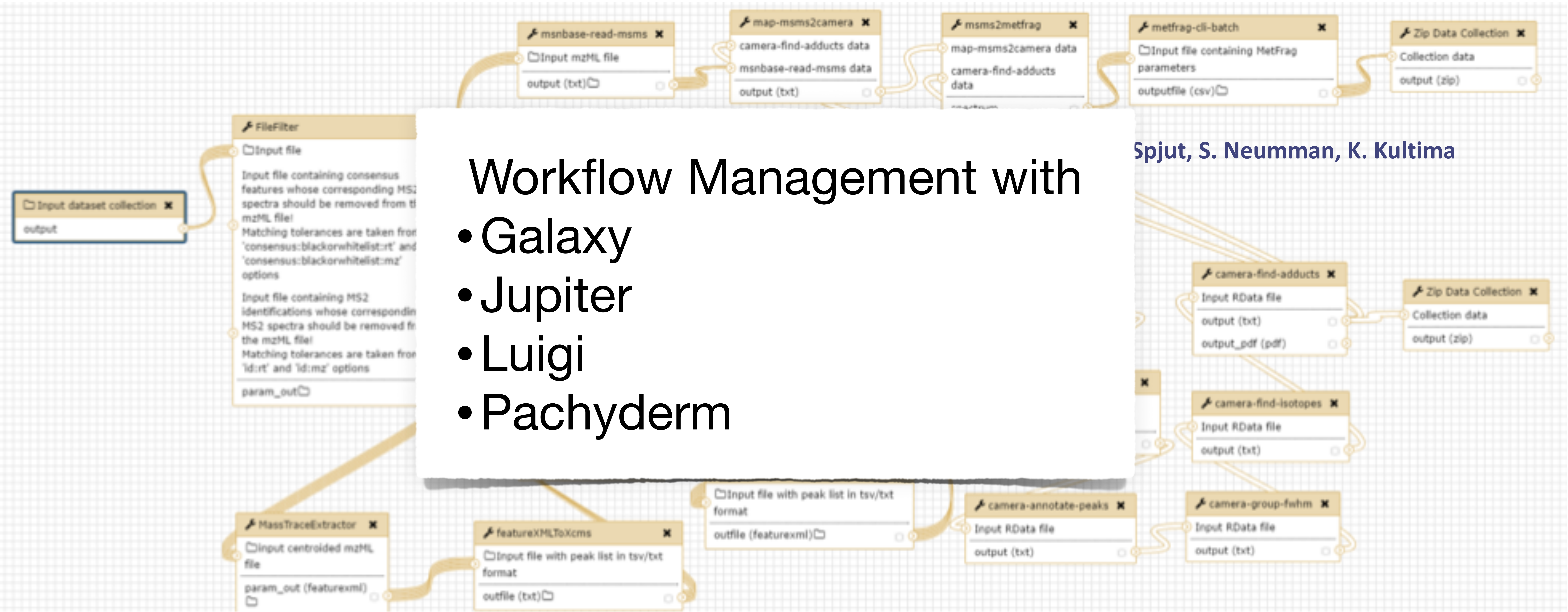
Input file containing MS2 identifications whose corresponding MS2 spectra should be removed from the mzML file! Matching tolerances are taken from 'id:rt' and 'id:mz' options

By P. Emami, S. Hermann, C. Ruttkies, O. Spjut, S. Neumman, K. Kultima

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

PhenoMeNal
Large-Scale Computing for Medical Metabolomics

MS1/MS2 – XCMS-OpenMS



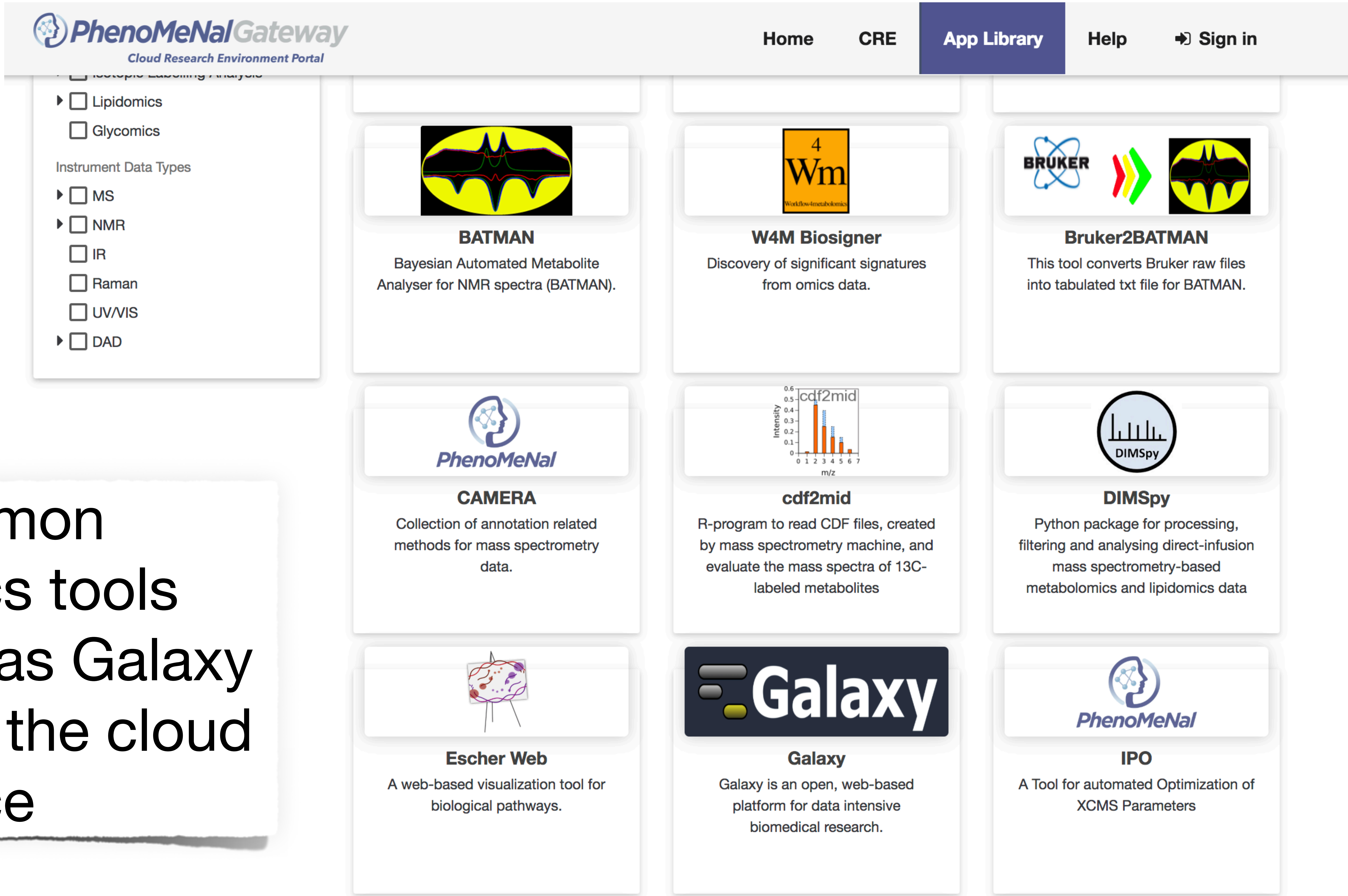
Spjut, S. Neumman, K. Kultima

Workflow Management with

- Galaxy
- Jupiter
- Luigi
- Pachyderm

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

PhenoMeNal
Large-Scale Computing for Medical Metabolomics



▸ □ Lipidomics
□ Glycomics

Instrument Data Types

▸ □ MS
▸ □ NMR
□ IR
□ Raman
□ UV/VIS
▸ □ DAD

BATMAN
Bayesian Automated Metabolite Analyser for NMR spectra (BATMAN).

W4M Biosigner
Discovery of significant signatures from omics data.

Bruker2BATMAN
This tool converts Bruker raw files into tabulated txt file for BATMAN.

CAMERA
Collection of annotation related methods for mass spectrometry data.

cdf2mid
R-program to read CDF files, created by mass spectrometry machine, and evaluate the mass spectra of 13C-labeled metabolites

DIMSpy
Python package for processing, filtering and analysing direct-infusion mass spectrometry-based metabolomics and lipidomics data

Escher Web
A web-based visualization tool for biological pathways.

Galaxy
Galaxy is an open, web-based platform for data intensive biomedical research.

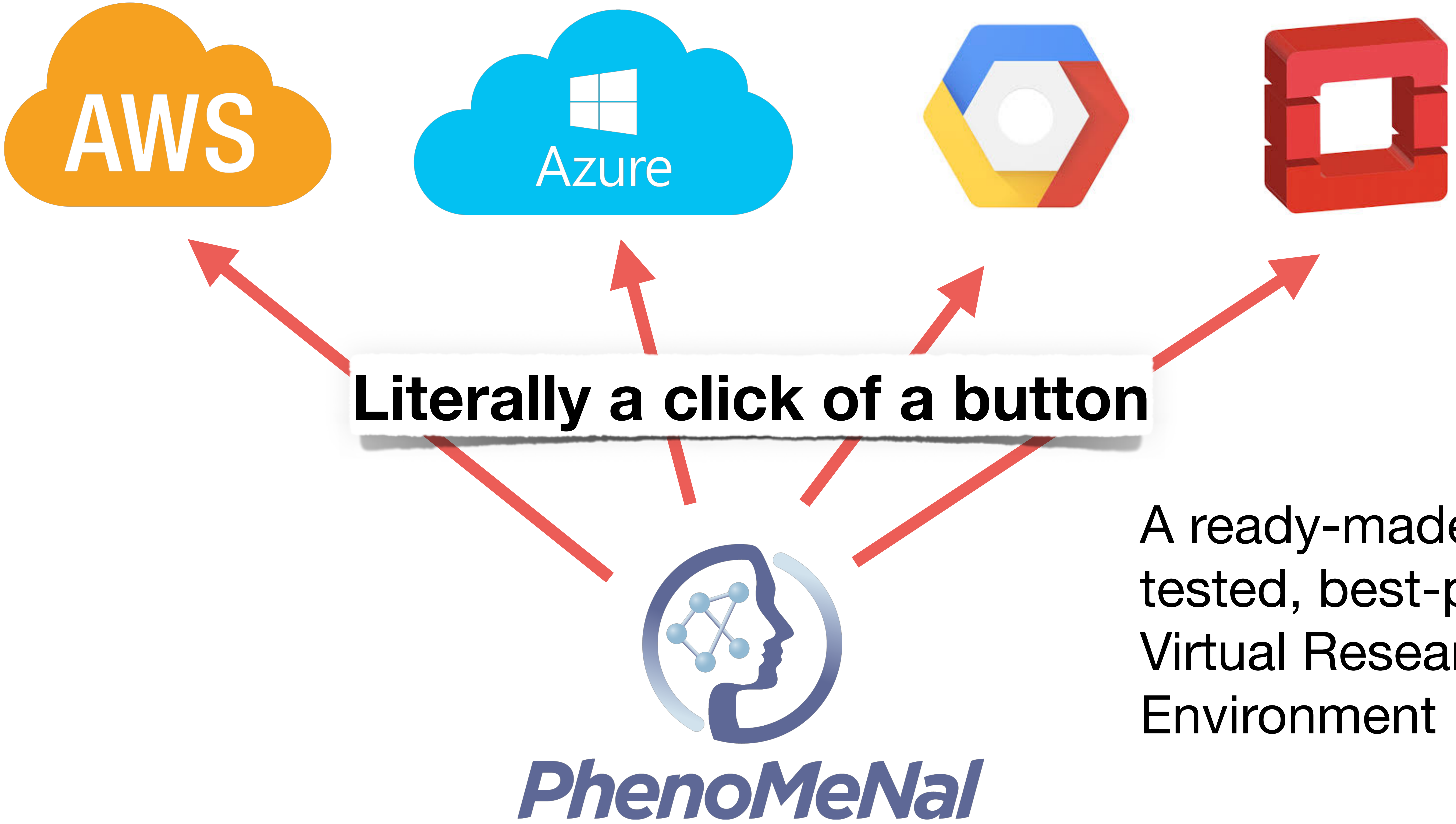
IPO
A Tool for automated Optimization of XCMS Parameters

100s of common metabolomics tools ready to run as Galaxy workflows in the cloud of your choice

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

PhenoMeNal
Large-Scale Computing for Medical Metabolomics

Large Scale Computing with Medical Metabolomics Data



Literally a click of a button

A ready-made, well-tested, best-practice, Virtual Research Environment

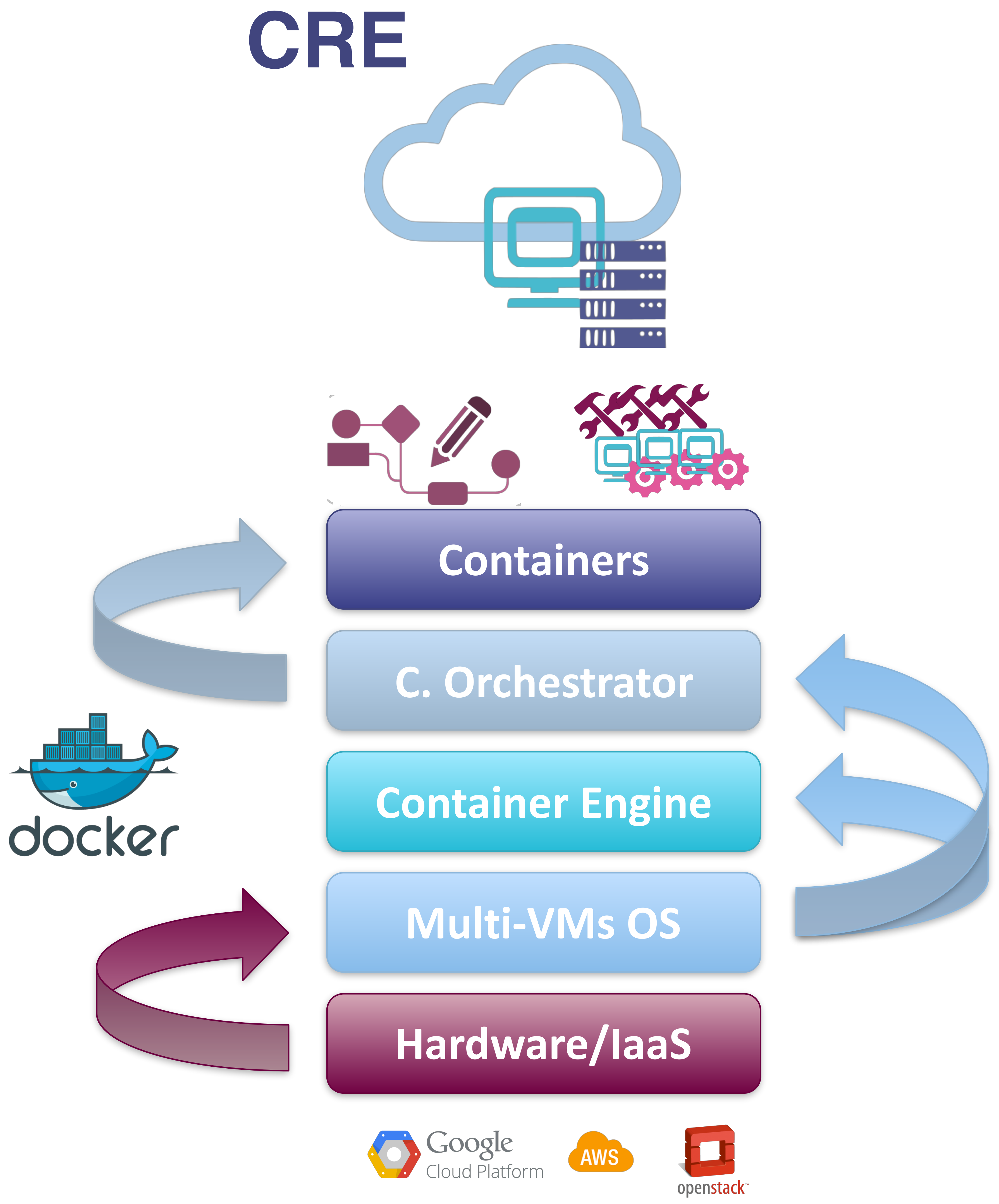
Managing large sets of machines in the cloud

Container Orchestration

kubernetes

Infrastructure Provisioning

TERRAFORM



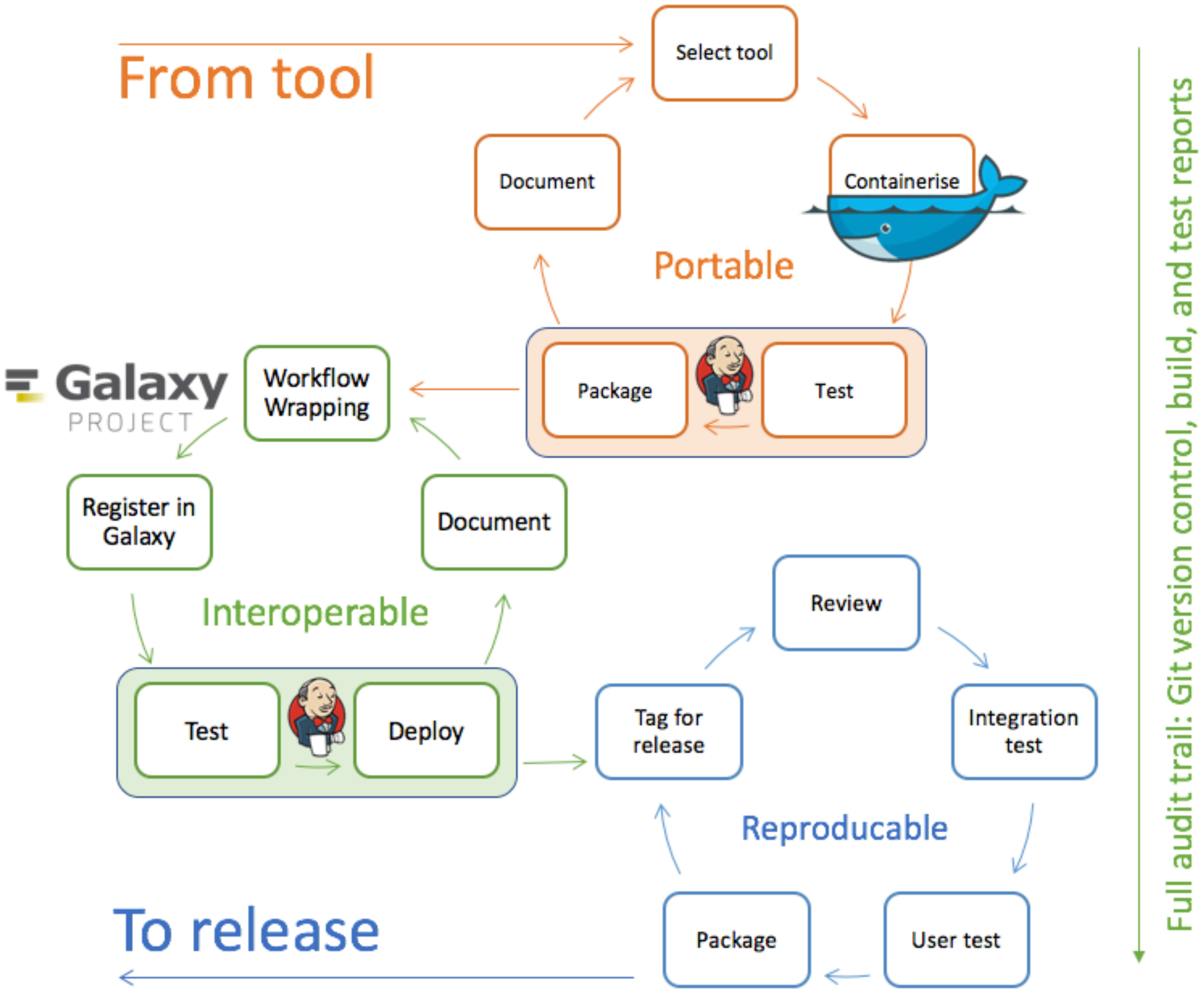
Software Provisioning

ANSIBLE

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

PhenoMeNal
Large-Scale Computing for Medical Metabolomics

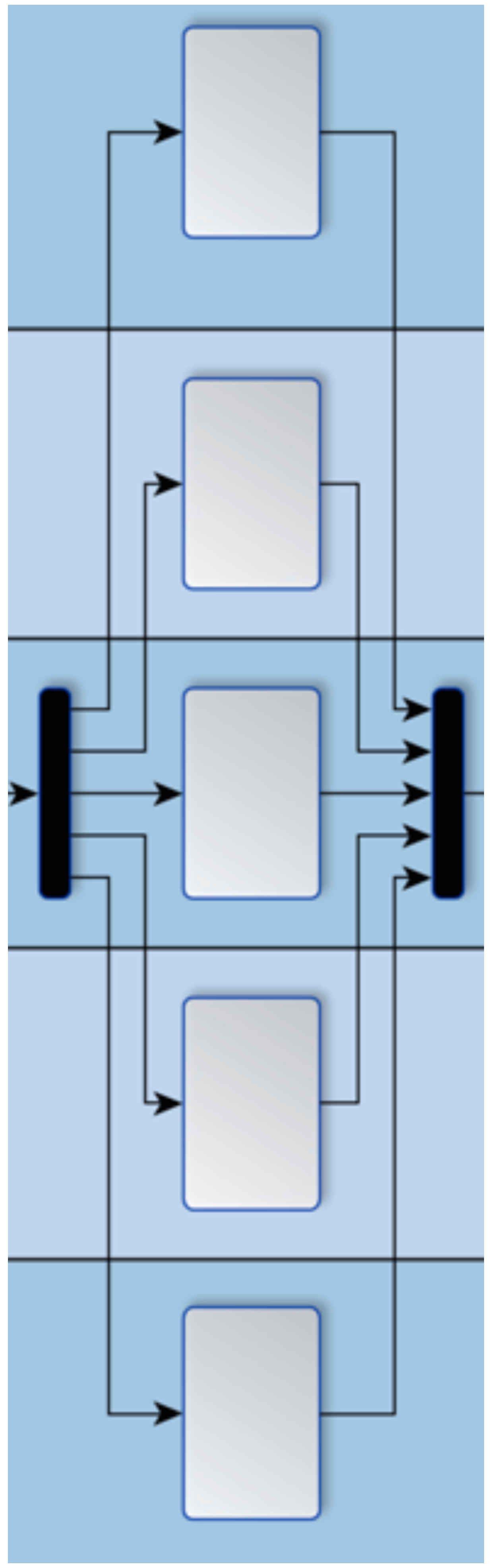
Generic Cloudification



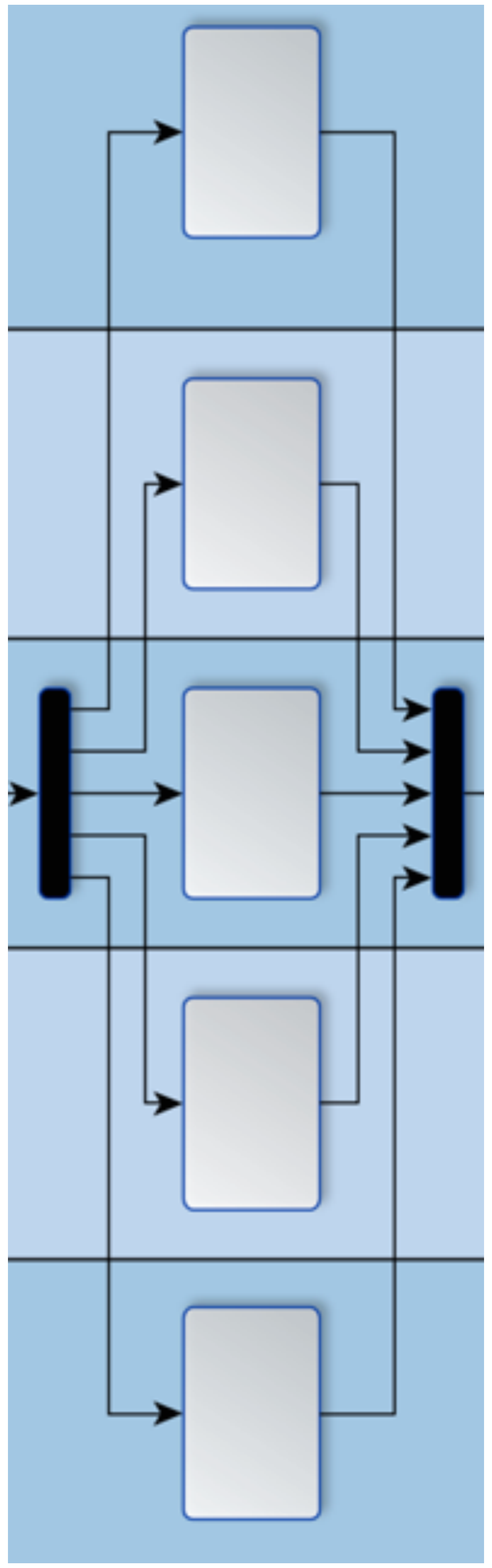
FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

PhenoMeNal
Large-Scale Computing for Medical Metabolomics

Economies of Scale

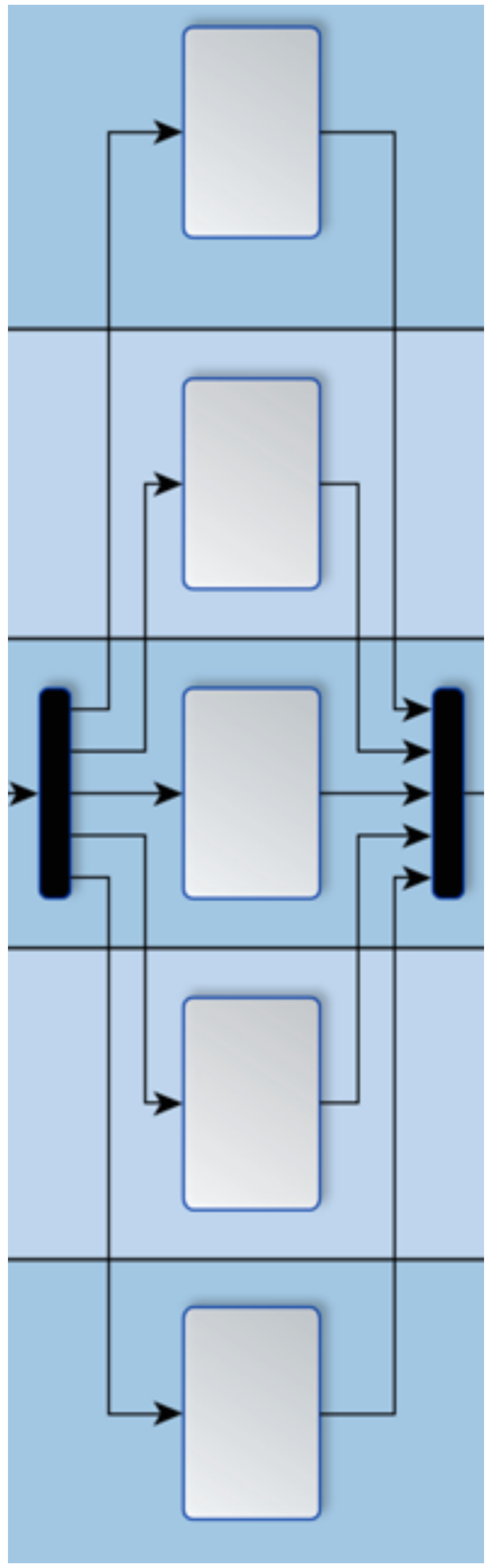


Economies of Scale



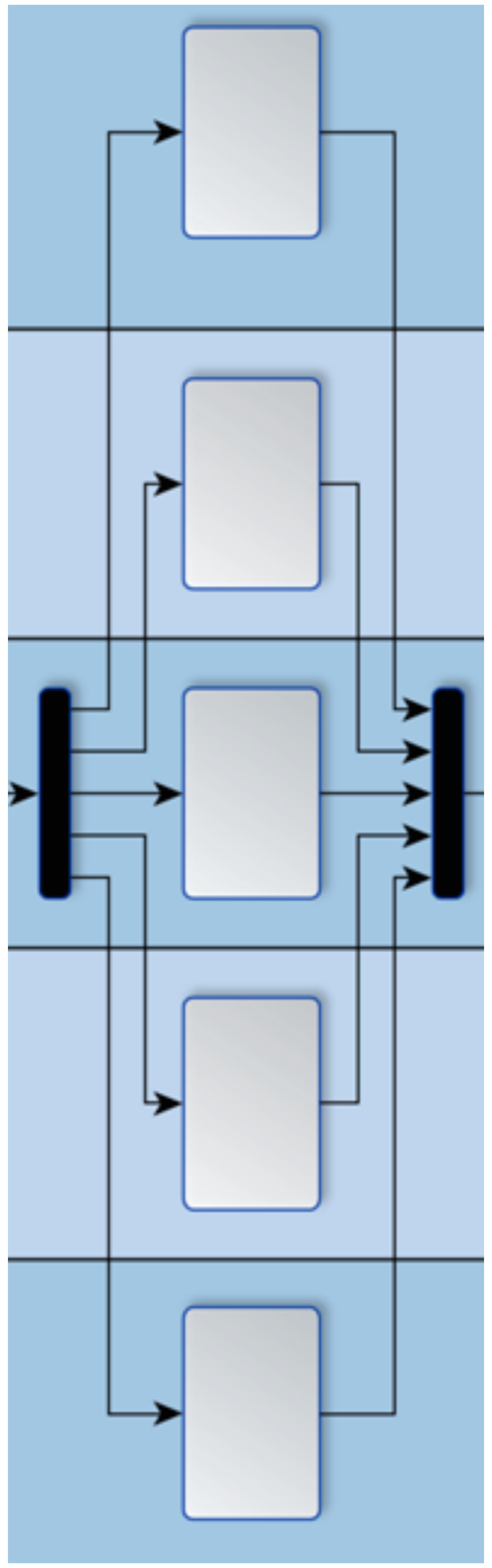
- Assume 24.000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.

Economies of Scale



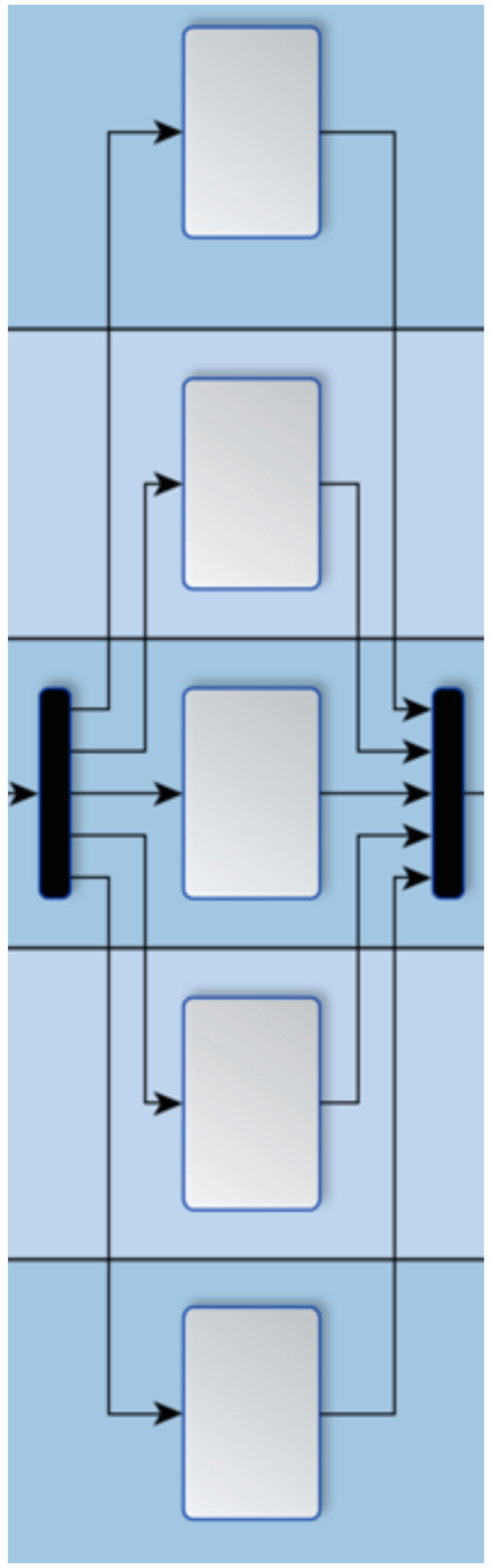
- Assume 24.000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.
- At $0.04 per core hour, plus 1TB storage, that‘s $16 in the cloud.

Economies of Scale



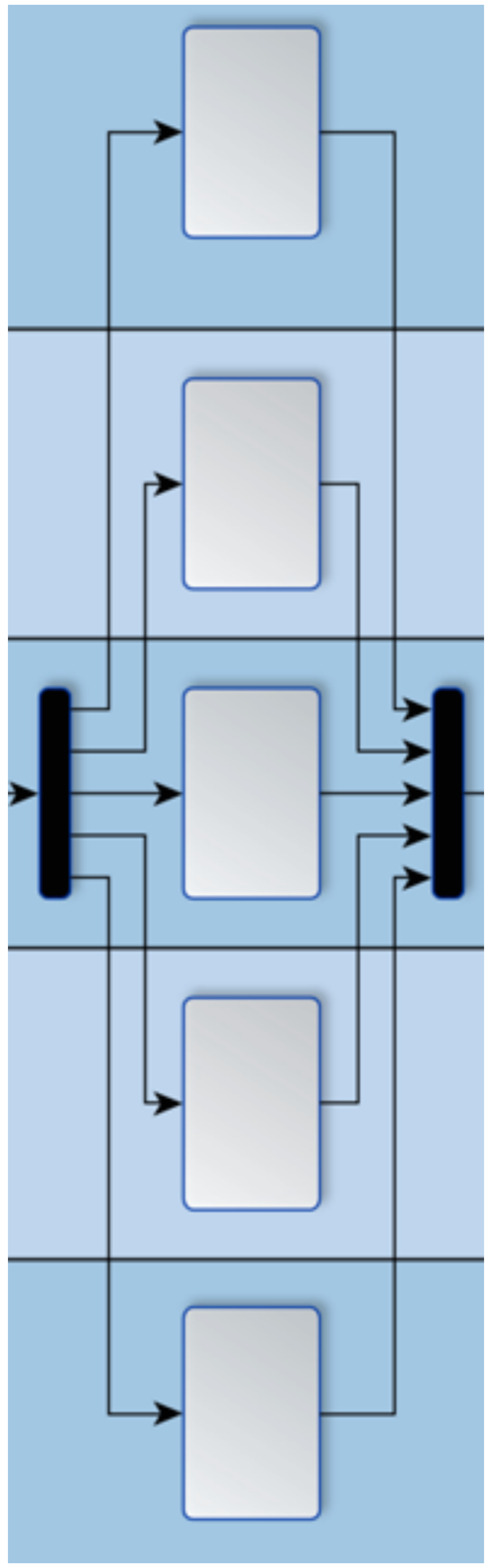
- Assume 24.000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.
- At $0.04 per core hour, plus 1TB storage, that‘s $16 in the cloud.
- It is also $16 if you rent 400 cloud servers and you‘d be done in 1 hour.

Economies of Scale



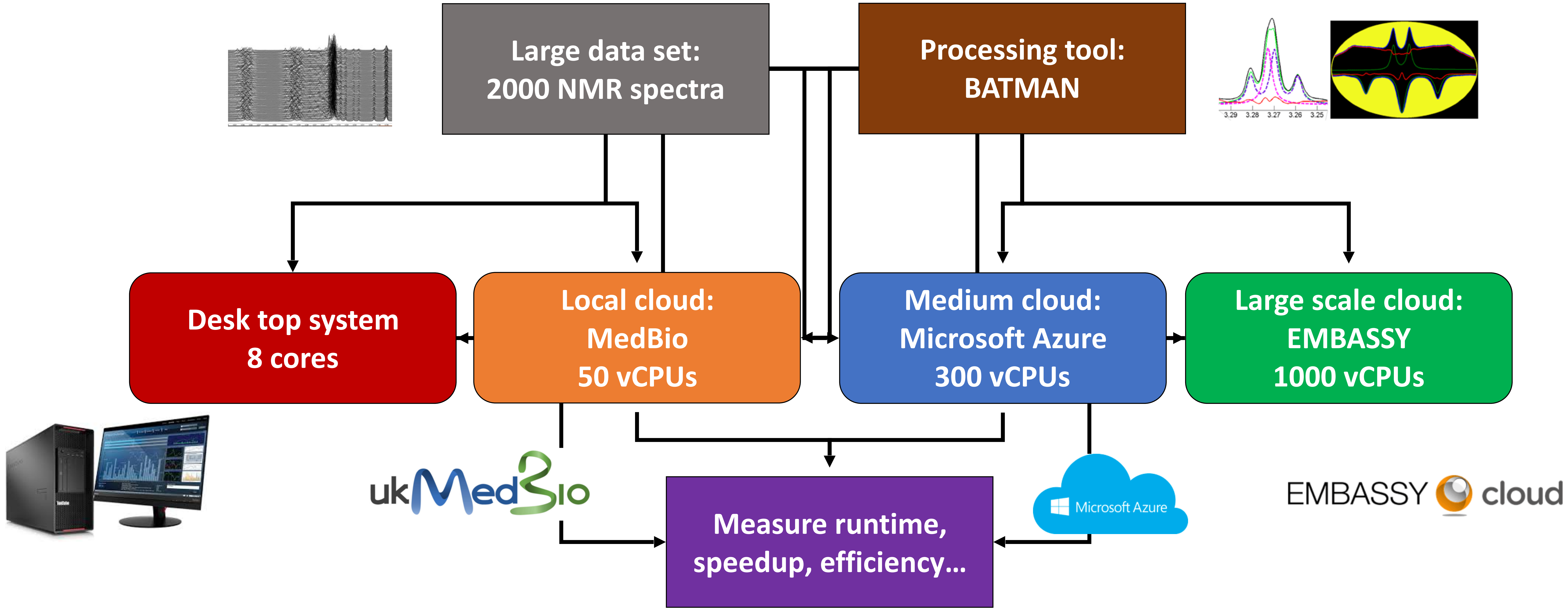
- Assume 24.000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.
- At $0.04 per core hour, plus 1TB storage, that‘s $16 in the cloud.
- It is also $16 if you rent 400 cloud servers and you‘d be done in 1 hour.
- It is also $16 if you rent 24.000 cloud servers, and be done in 1 minute (ignoring overhead …)

Economies of Scale



- Assume 24.000 samples. Each takes 1 minute to process. That is 400 CPU core hours, or 16 days, or 2 weeks.
- At $0.04 per core hour, plus 1TB storage, that‘s $16 in the cloud.
- It is also $16 if you rent 400 cloud servers and you‘d be done in 1 hour.
- It is also $16 if you rent 24.000 cloud servers, and be done in 1 minute (ignoring overhead …)

Testing scalability



Large data set:
2000 NMR spectra

Processing tool:
BATMAN

Desk top system
8 cores

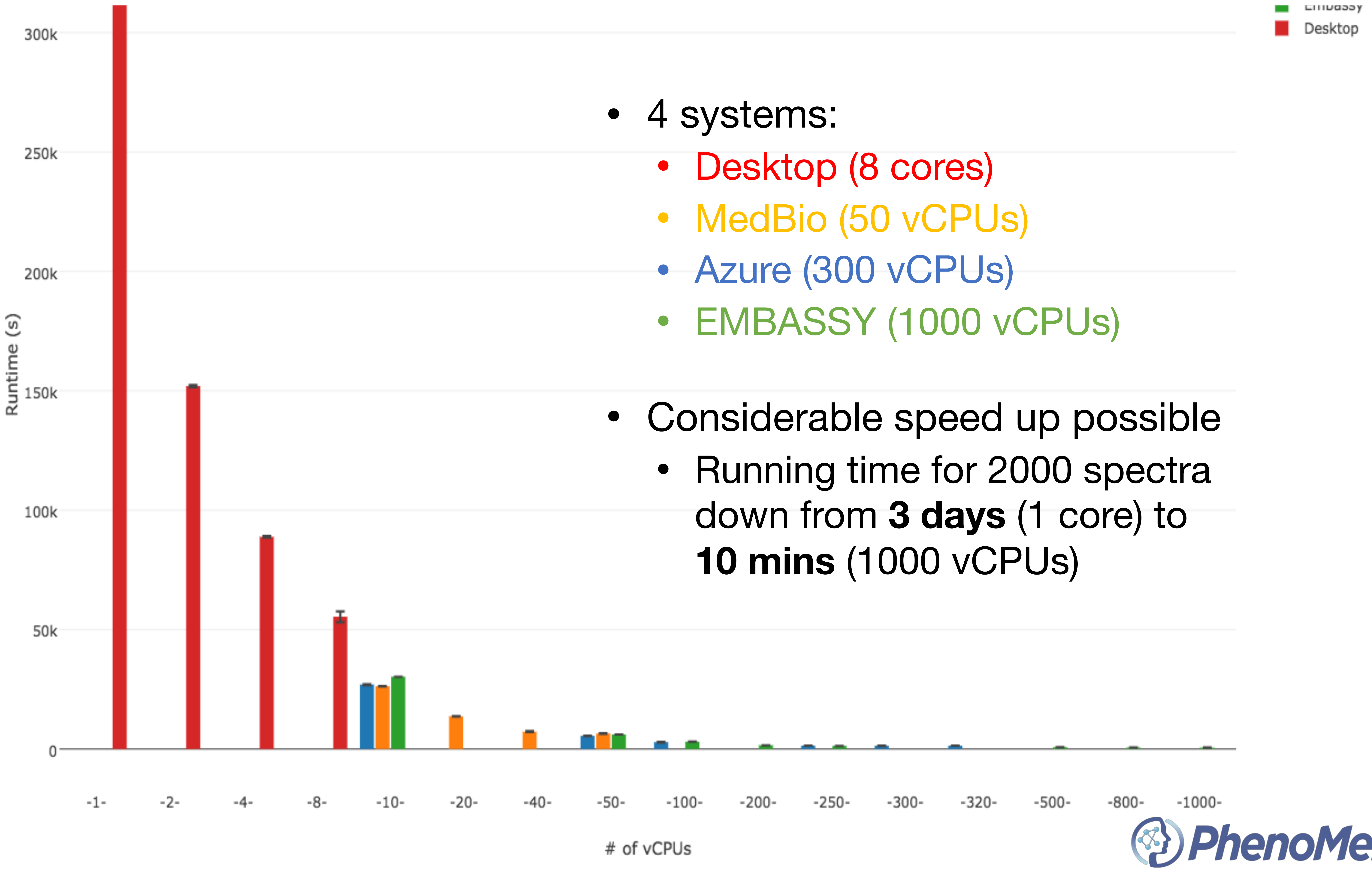
Local cloud:
MedBio
50 vCPUs

Medium cloud:
Microsoft Azure
300 vCPUs

Large scale cloud:
EMBASSY
1000 vCPUs

Measure runtime,
speedup, efficiency…

PhenoMeNal scalability – BATMAN NMR processing



- 4 systems:
  - Desktop (8 cores)
  - MedBio (50 vCPUs)
  - Azure (300 vCPUs)
  - EMBASSY (1000 vCPUs)
- Considerable speed up possible
  - Running time for 2000 spectra down from **3 days** (1 core) to **10 mins** (1000 vCPUs)

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

PhenoMeNal
Large-Scale Computing for Medical Metabolomics

Summary

- Large analytical chemistry data produced in medical metabolomics
- Computational analysis can take days or weeks on a single node
- **Industry standard tools** such as Kubernetes, Terraform and Ansible allow us to
    - seamlessly **deploy analysis pipelines**
        - composed of **microservices**
        - on **1000s of cloud machines**
    - at **low costs** and **no upfront investment**
- The free and open PhenoMeNal infrastructure encapsulates those **orchestration tools** together with **hundreds of tools in computational metabolomics**

Funding and Collaborators

European Commission
Grant # 654241

PhenoMeNal
Consortium

Netherlands Metabolomics Centre

4 Wm
Workflow4metabolomics

METABOHUB

EMBL-EBI

Imperial College London

isasoftwaresuite

MRC Human Nutrition Research
Improving health through nutrition research

UNIVERSITY OF CAMBRIDGE

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA