UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Interactive Data Exploration

Lars Ailo Bongo
Dept. of Computer Science & Center for Bioinformatics,
University of Tromsø, Norway

http://bdps.cs.uit.no

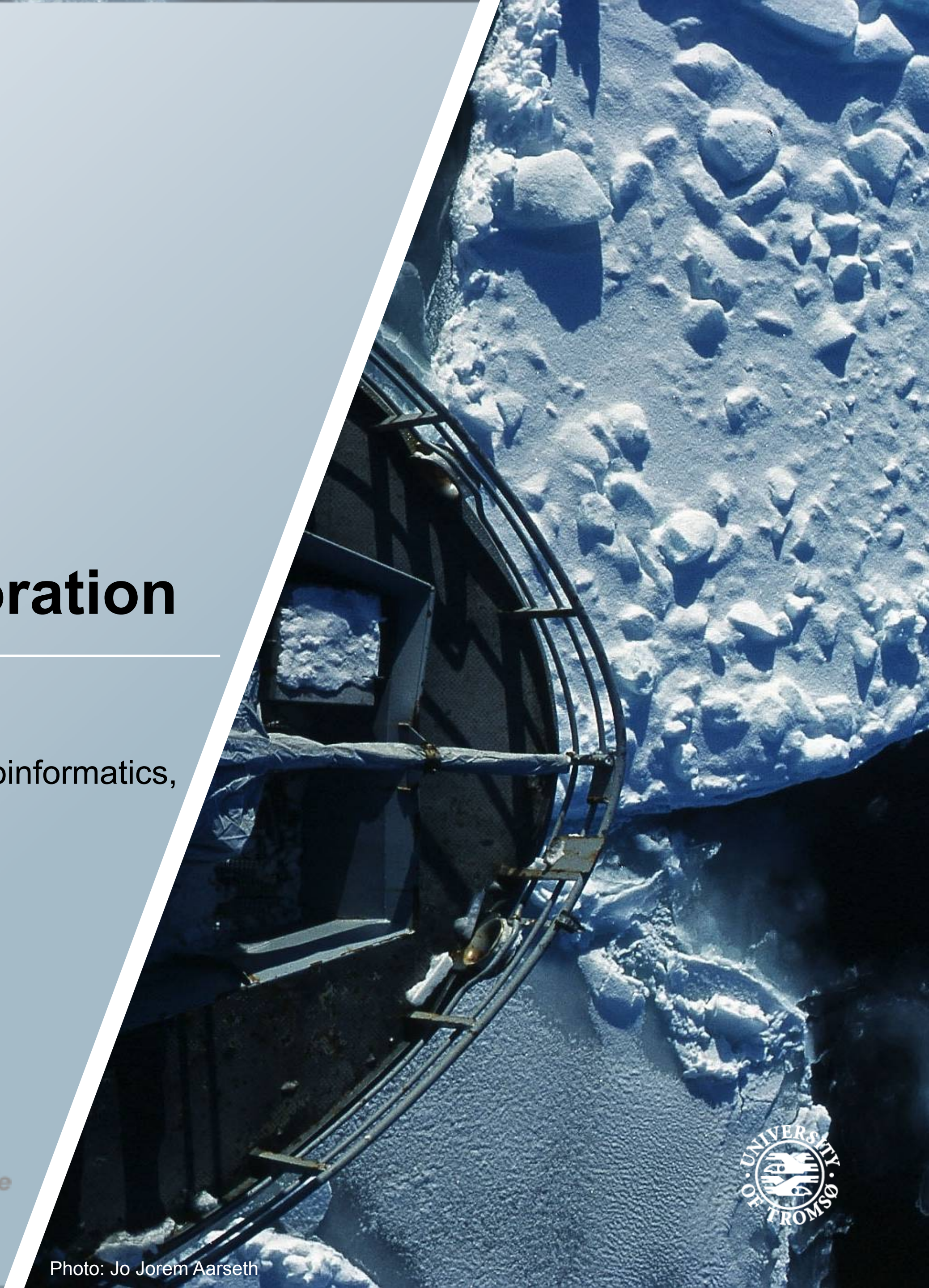
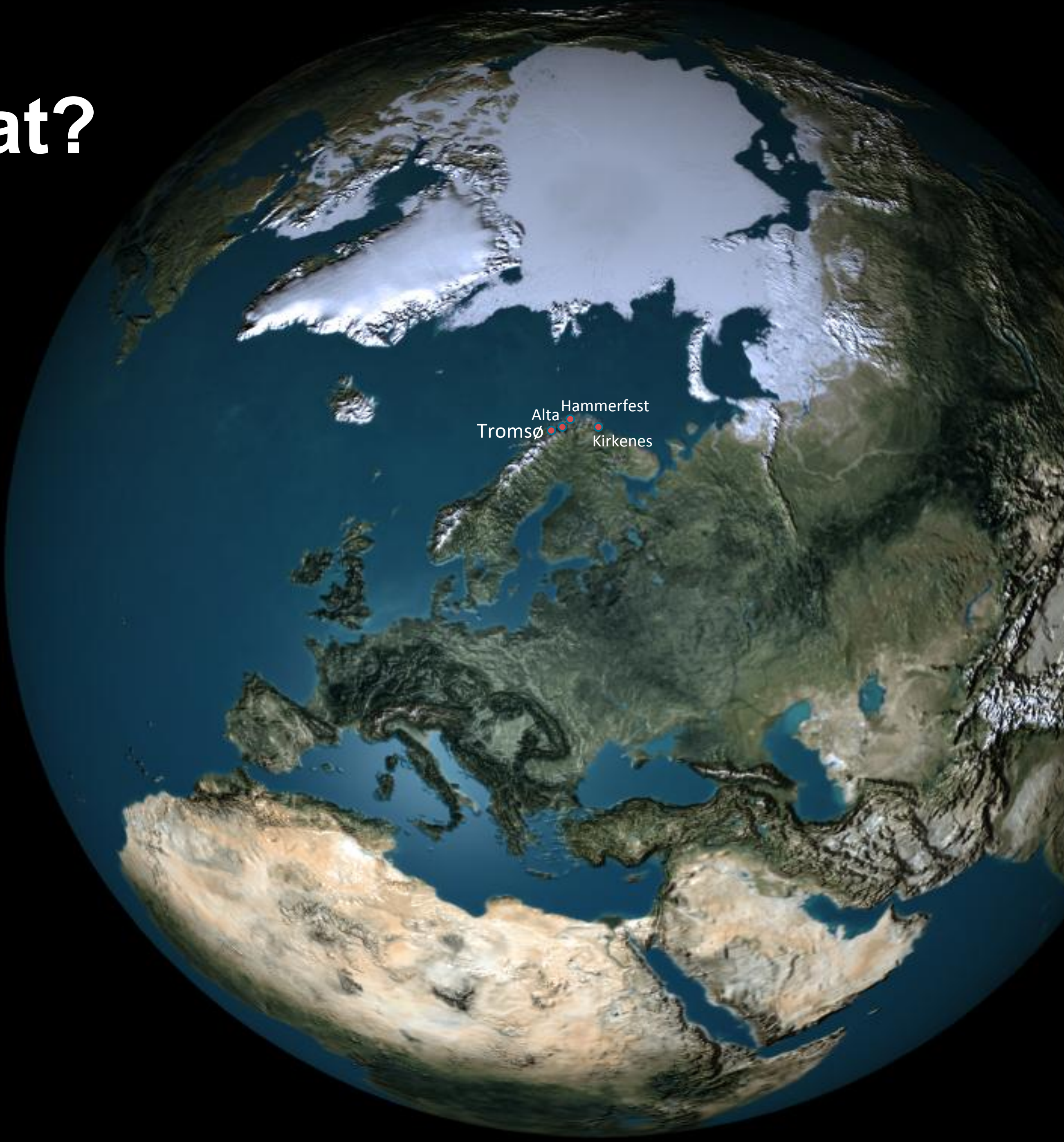
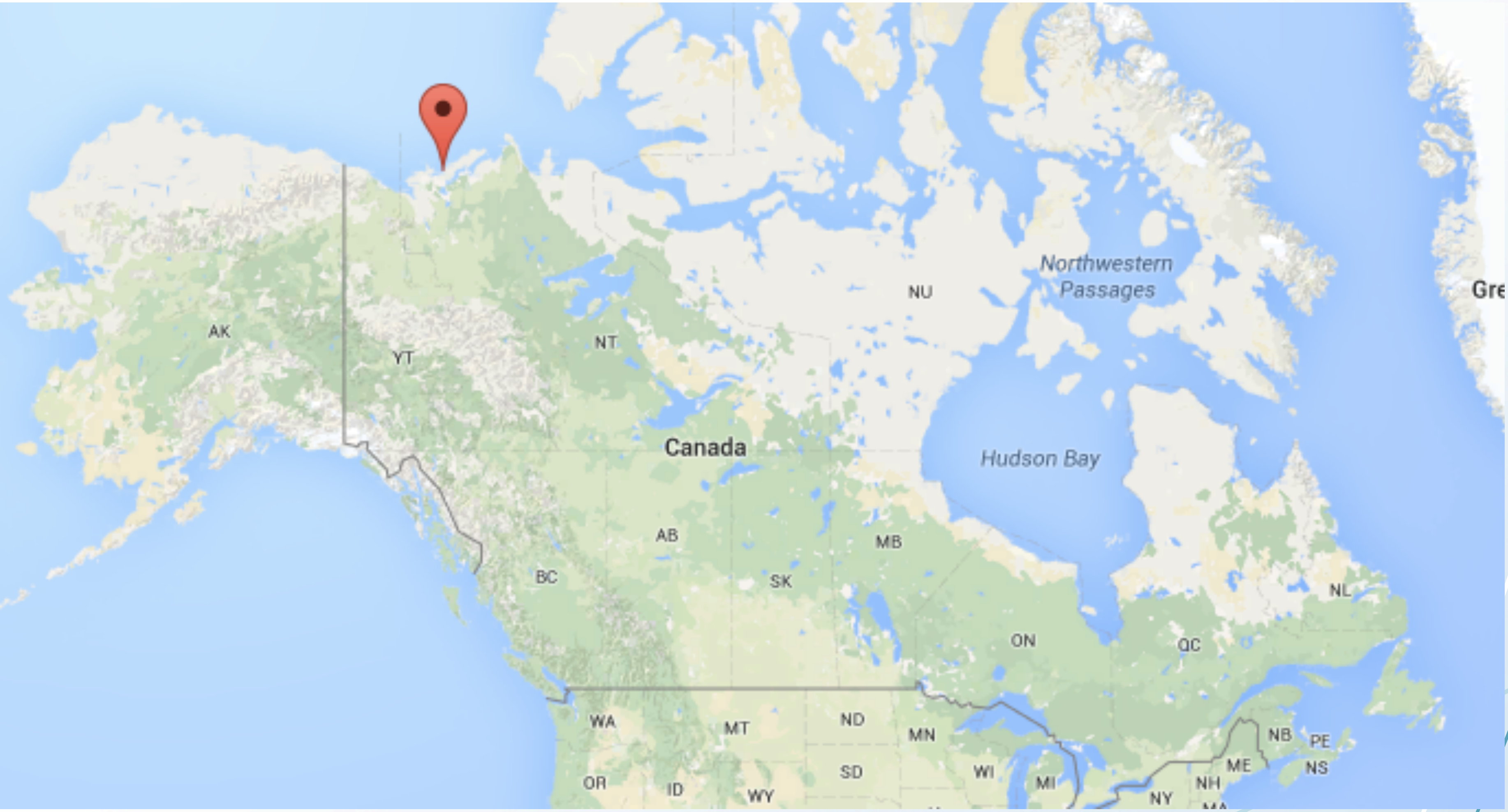


Photo: Jo Jorem Aarseth

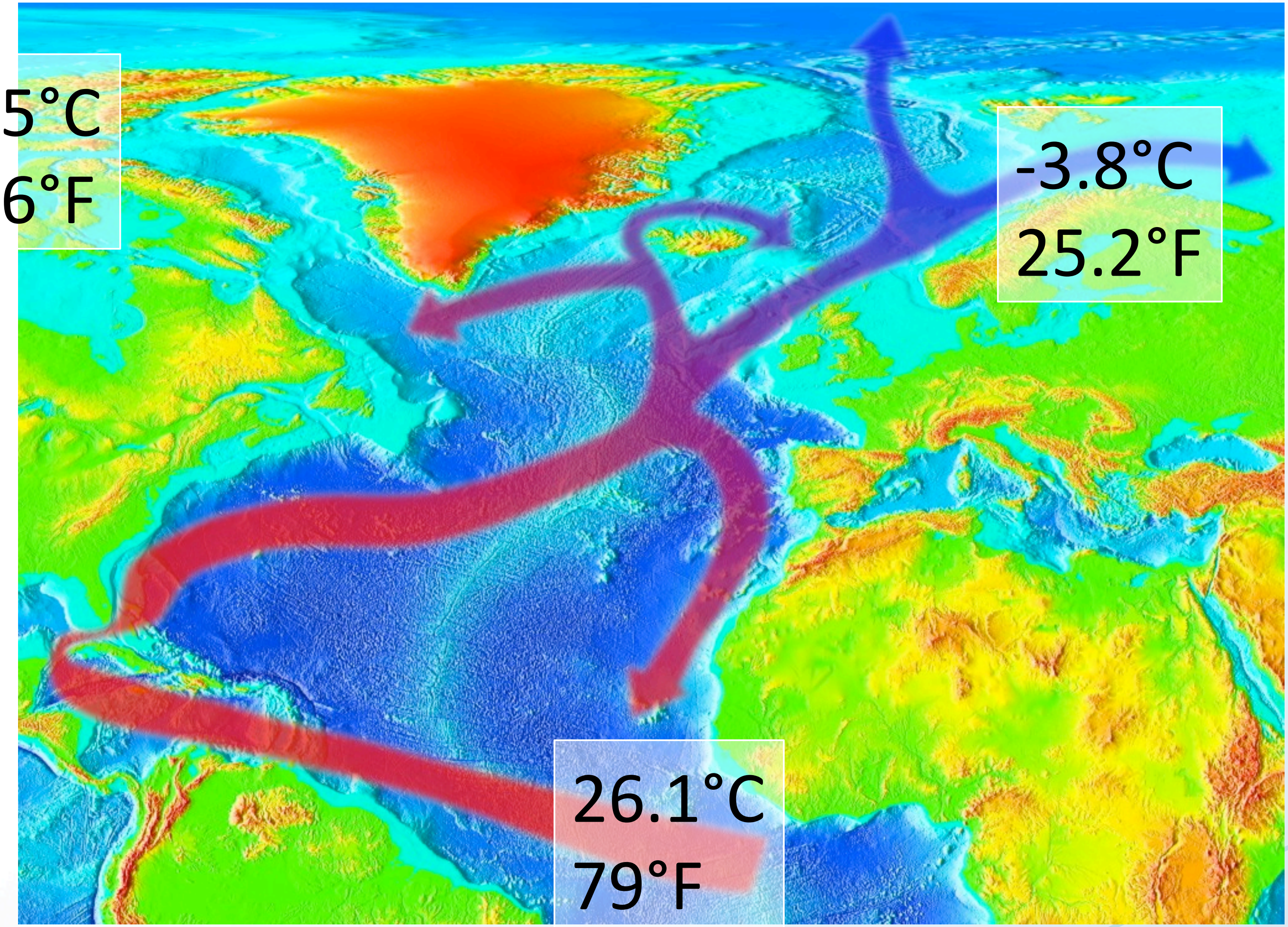
Where’s that?



Tuktoyaktuk (69° North)



Isn’t cold?



-26.5°C
-15.6°F

-3.8°C
25.2°F

26.1°C
79°F

Outline

- Visualization tool
- Data preprocessing
- Backend systems for data analysis
- Big data management and processing
- Distributed compute and storage resources

- Or, what the programmers in the lab are doing

Background

- Norwegian Woman and Cancer (NOWAC)
  …as described by Vanessa
- Big prospective cohort study
  - Questionnaires from 170 000 women
  - Blood samples from 50.000 women
  - Tumor tissues
- Integrated functional analysis
  - Questioners
  - Blood
  - Tumor tissues
  - Register data

NOWAC Datasets

Name	Case-Control Samples
AROS	80
Hospital case-control (CC1, CC2, CC3)	248
Postdiagnostic	434
Prospective breast	719
Prospective ovarian	95
Prospective endometrial	84
Stress	48
SUM	1708 Case-Controls

NOWAC

- Initial analyses done
  - But still more to discover
- Data analysis lessons learned
  - Analyses should be run agnostics without prior hypotheses
  - Use existing biological knowledge for testing and understanding
  - We lack the tools for such data exploration

Interactive Data Exploration

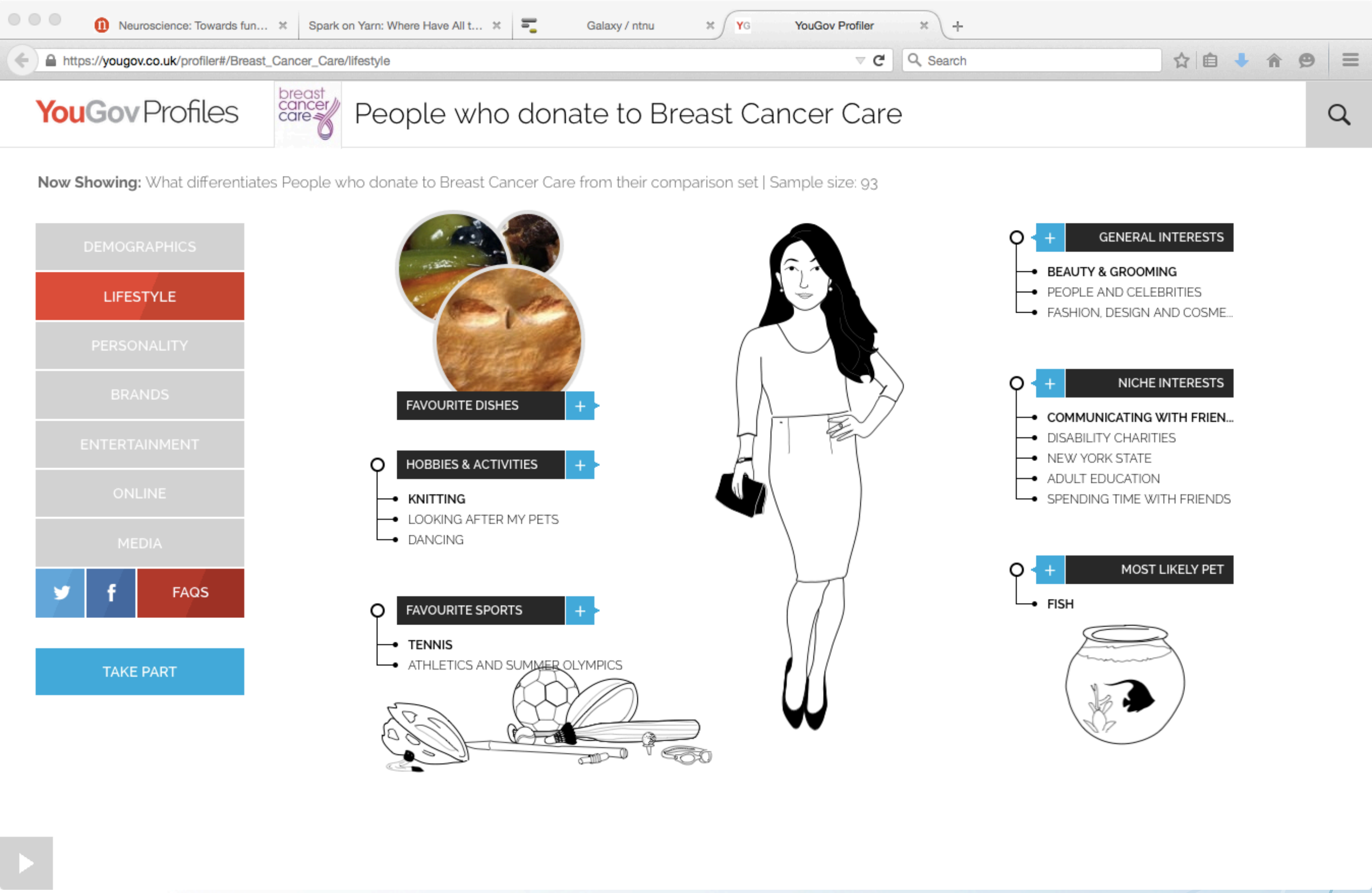
- Data exploration = “play with data”
  - No prior hypothesis
- Interactive
  - Human - computer
  - Short response times (seconds or milliseconds)
- Computers helps by making predictions
- Combined with (proper) hypothesis testing

tableau®

Qlik®

pentaho

TIBCO® Spotfire®


Neuroscience: Towards fun…
Spark on Yarn: Where Have All t…
Galaxy / ntnu
YouGov Profiler
https://yougov.co.uk/profiler#/Breast_Cancer_Care/lifestyle
Search
YouGov Profiles
breast cancer care

People who donate to Breast Cancer Care

Now Showing: What differentiates People who donate to Breast Cancer Care from their comparison set | Sample size: 93

FAVOURITE DISHES +

HOBBIES & ACTIVITIES +
- KNITTING
- LOOKING AFTER MY PETS
- DANCING

FAVOURITE SPORTS +
- TENNIS
- ATHLETICS AND SUMMER OLYMPICS

+ GENERAL INTERESTS
- BEAUTY & GROOMING
- PEOPLE AND CELEBRITIES
- FASHION, DESIGN AND COSME...

+ NICHE INTERESTS
- COMMUNICATING WITH FRIEN...
- DISABILITY CHARITIES
- NEW YORK STATE
- ADULT EDUCATION
- SPENDING TIME WITH FRIENDS

+ MOST LIKELY PET
- FISH

Interactive Data Exploration Requirements

- Human experts for data analysis
- Interactive user interface
- Analysis methods and models
- Data management and backend processing
- Compute and storage resources

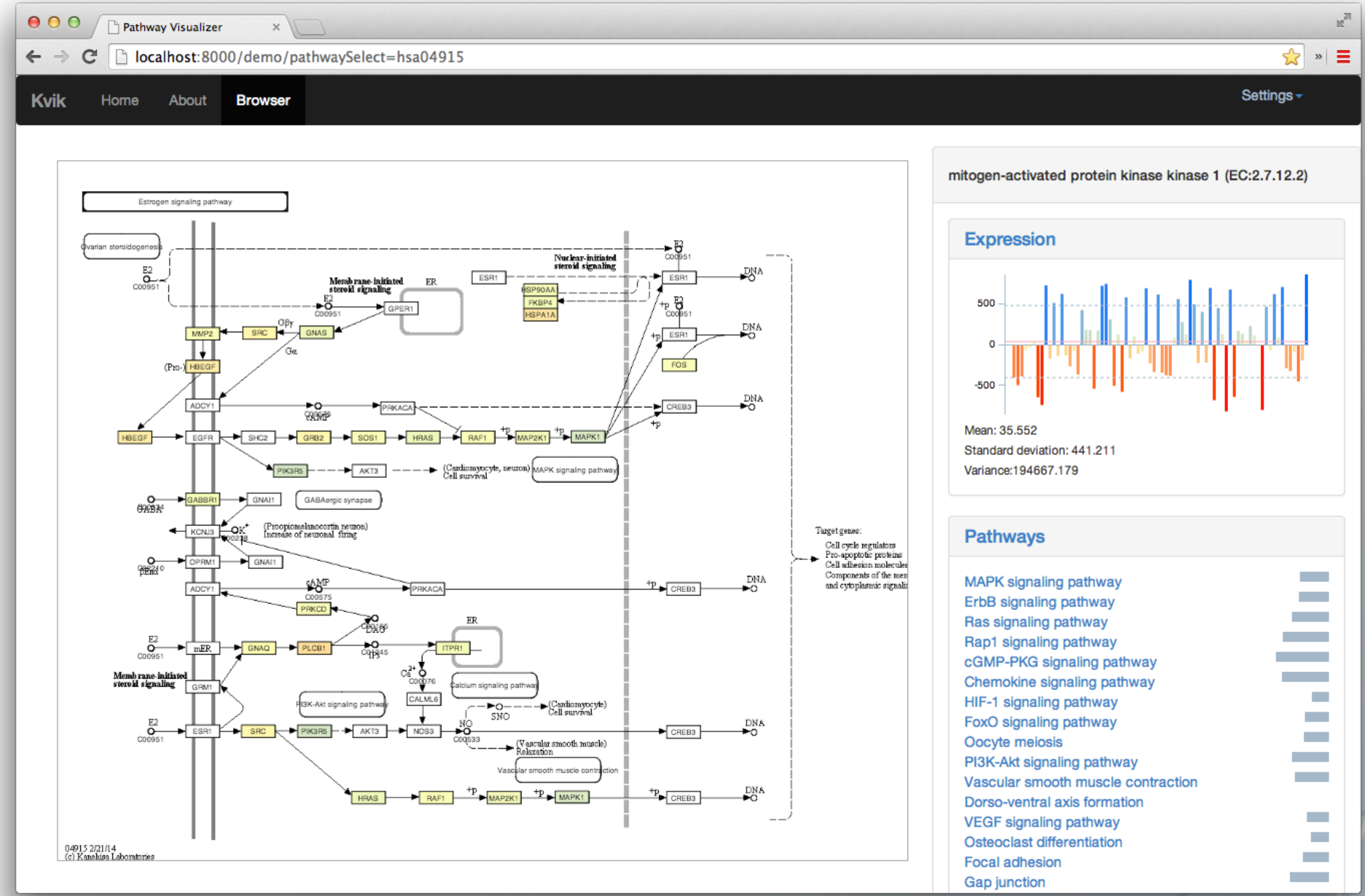
Interactive Visual User Interface

- Visualization tool to map NOWAC case-control gene expressions to known biology
- Existing visualization tools not flexible enough
- Developed with NOWAC data analysts
- Approach:
  - One specialized tool per analysis project
  - Framework that makes it easy to implement tools

Requirements: Solutions

- Flexible: 3-tier architecture and R based backend
- Interactive performance: good implementation
- Scalable: parallel or distributed backend
- Familiar visualizations: KEGG pathways
- Easy-to-use: web app
- Secure data storage: backend runs on secure server

Kvik– NOWAC Data Exploration



mitogen-activated protein kinase kinase 1 (EC:2.7.12.2)

Expression

Mean: 35.552

Standard deviation: 441.211

Variance:194667.179

Pathways

- MAPK signaling pathway
- ErbB signaling pathway
- Ras signaling pathway
- Rap1 signaling pathway
- cGMP-PKG signaling pathway
- Chemokine signaling pathway
- HIF-1 signaling pathway
- FoxO signaling pathway
- Oocyte meiosis
- PI3K-Akt signaling pathway
- Vascular smooth muscle contraction
- Dorso-ventral axis formation
- VEGF signaling pathway
- Osteoclast differentiation
- Focal adhesion
- Gap junction

Kvik – NOWAC Data Exploration

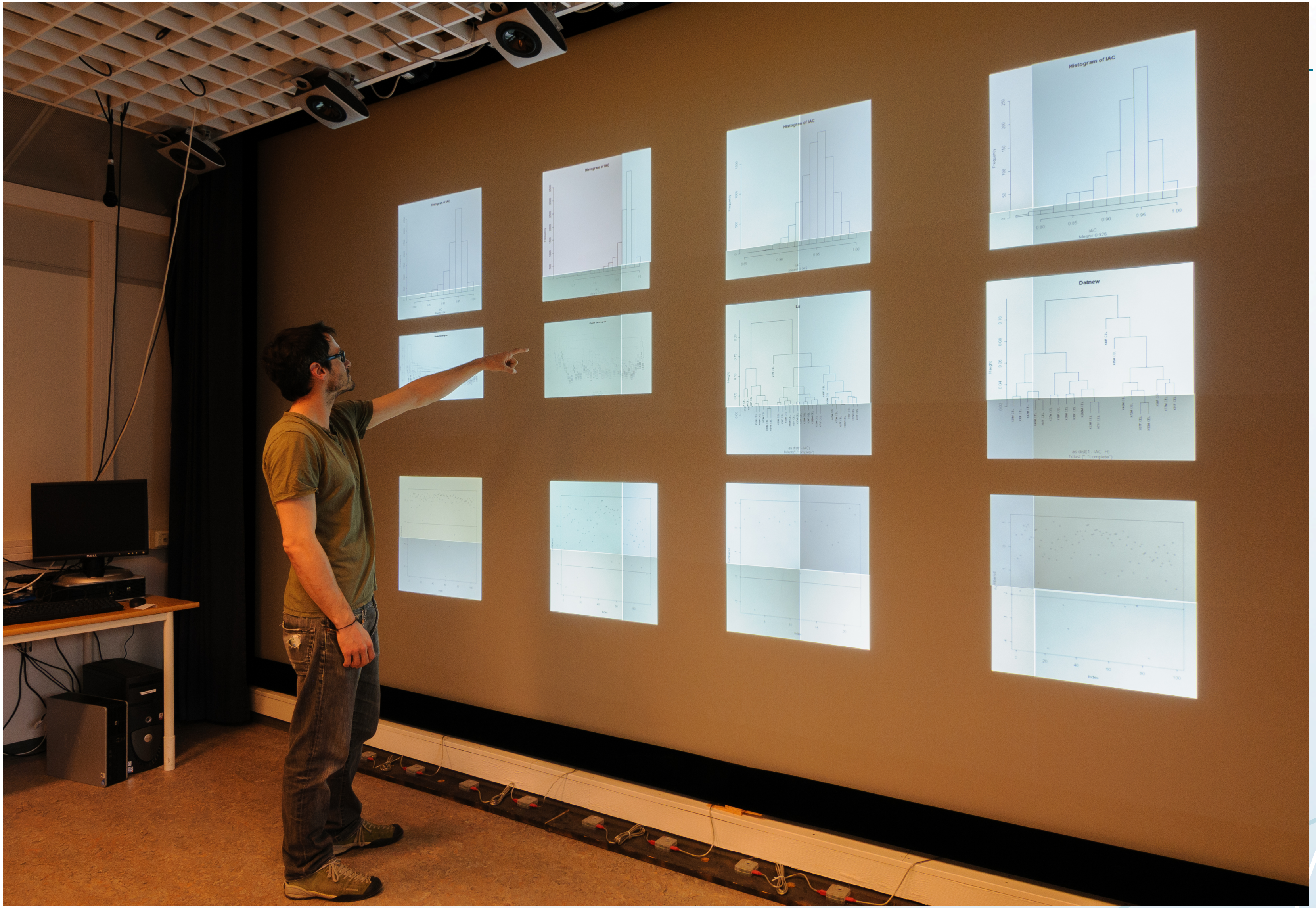
- Currently used for NOWAC data exploration
- Publically available and open-sourced:
    - kvik.cs.uit.no
    - github.com/fjukstad/kvik
    - Docker containers
    - Ongoing work
- Bjørn Fjukstad (PhD student)

Data Cleaning Toolchain

- Data cleaning important, but time consuming (and boring)
- Good data cleaning tools for textual and tabular data
  - Not suited for scientific data cleaning
- Approach
  - R scripts generates images with visualizations
  - Interactively group and sort images
  - Compare related images

TRIFACTA

Mr. Clean – Data Cleaning



Mr. Clean – Data Cleaning

- Gesture based interaction with many visualizations
  - Use case: NOWAC outlier removal
  - Use case: computer vision algorithm development
- Availability:
  - github.com/UniversityofTromso/mrclean
  - youtu.be/NFUDsPQRwqE
  - Proc. of VISSOFT'14
- Giacomo Tartari & Einar Holsbø (PhD student)

Kvik Backend

- Backend for executing data analysis methods
  - Machine learning algorithms
  - Computationally demanding
  - Must be very fast
  - Using a Supercomputer
- Einar Holsbø (PhD Student)

Dataset Size

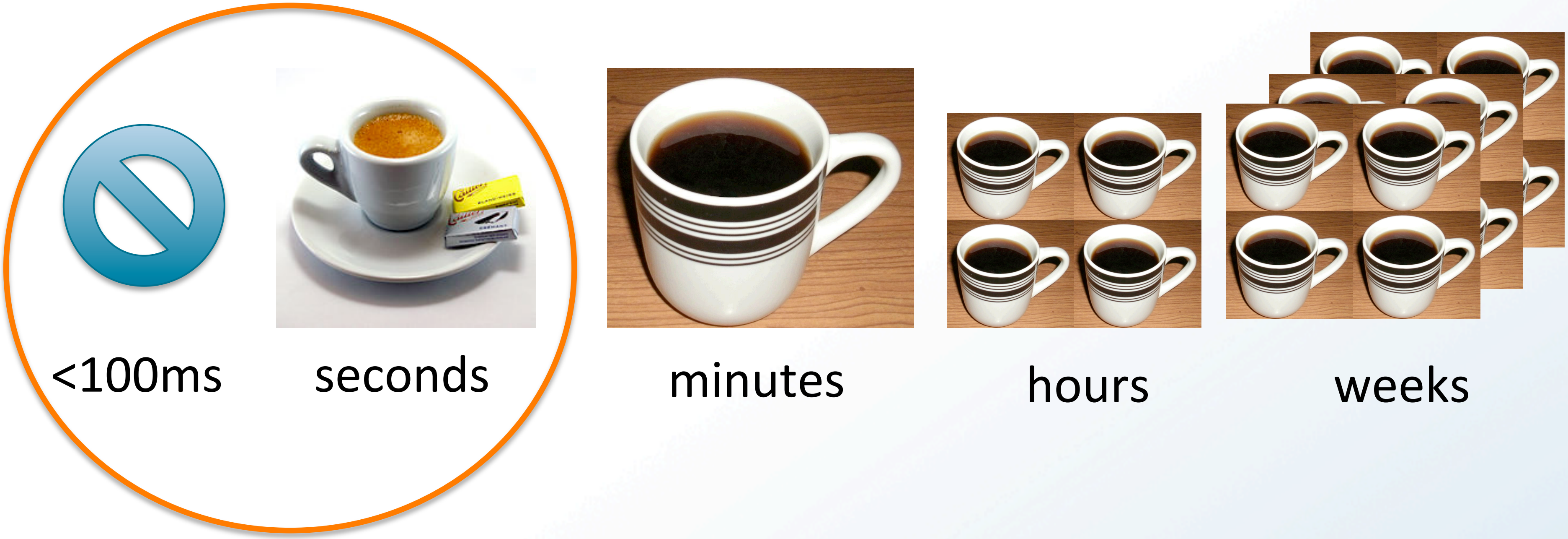
< 4GB

< 512GB

TBs

PBs

Computation Time



<100ms

seconds

minutes

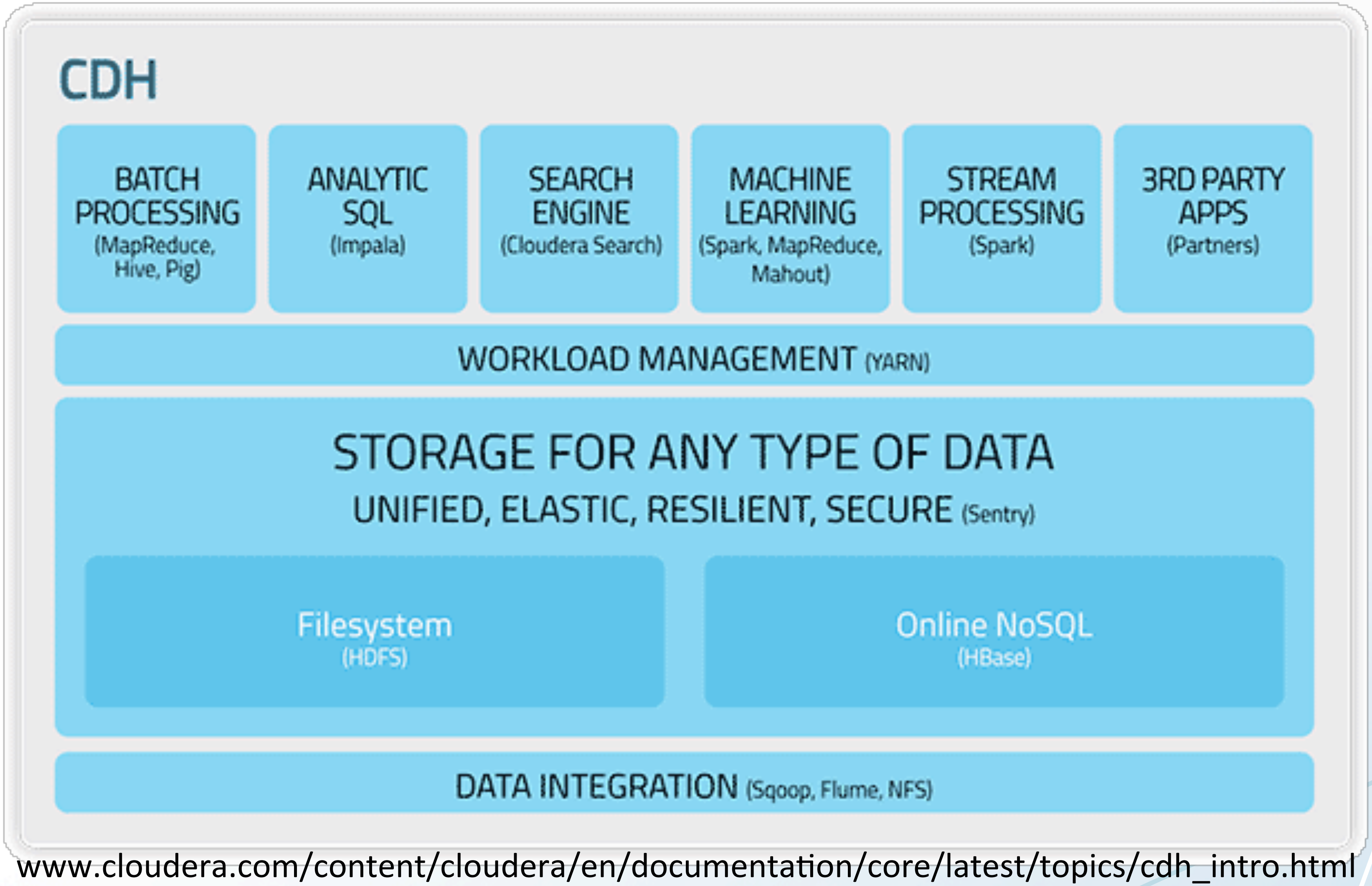
hours

weeks

Optimizations

- Assuming we start with an R or Matlab implementation
- Algorithm parameter tuning
- C++/ Java / … implementation
- Data structure optimization
- Multi-threaded parallelization (single machine)
- Distributed parallelization (multiple-machines)

= Complex Software Stack


CDH
BATCH PROCESSING (MapReduce, Hive, Pig)
ANALYTIC SQL (Impala)
SEARCH ENGINE (Cloudera Search)
MACHINE LEARNING (Spark, MapReduce, Mahout)
STREAM PROCESSING (Spark)
3RD PARTY APPS (Partners)
WORKLOAD MANAGEMENT (YARN)
STORAGE FOR ANY TYPE OF DATA
UNIFIED, ELASTIC, RESILIENT, SECURE (Sentry)
Filesystem (HDFS)
Online NoSQL (HBase)
DATA INTEGRATION (Sqoop, Flume, NFS)

www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cdh_intro.html

Meta-database Management

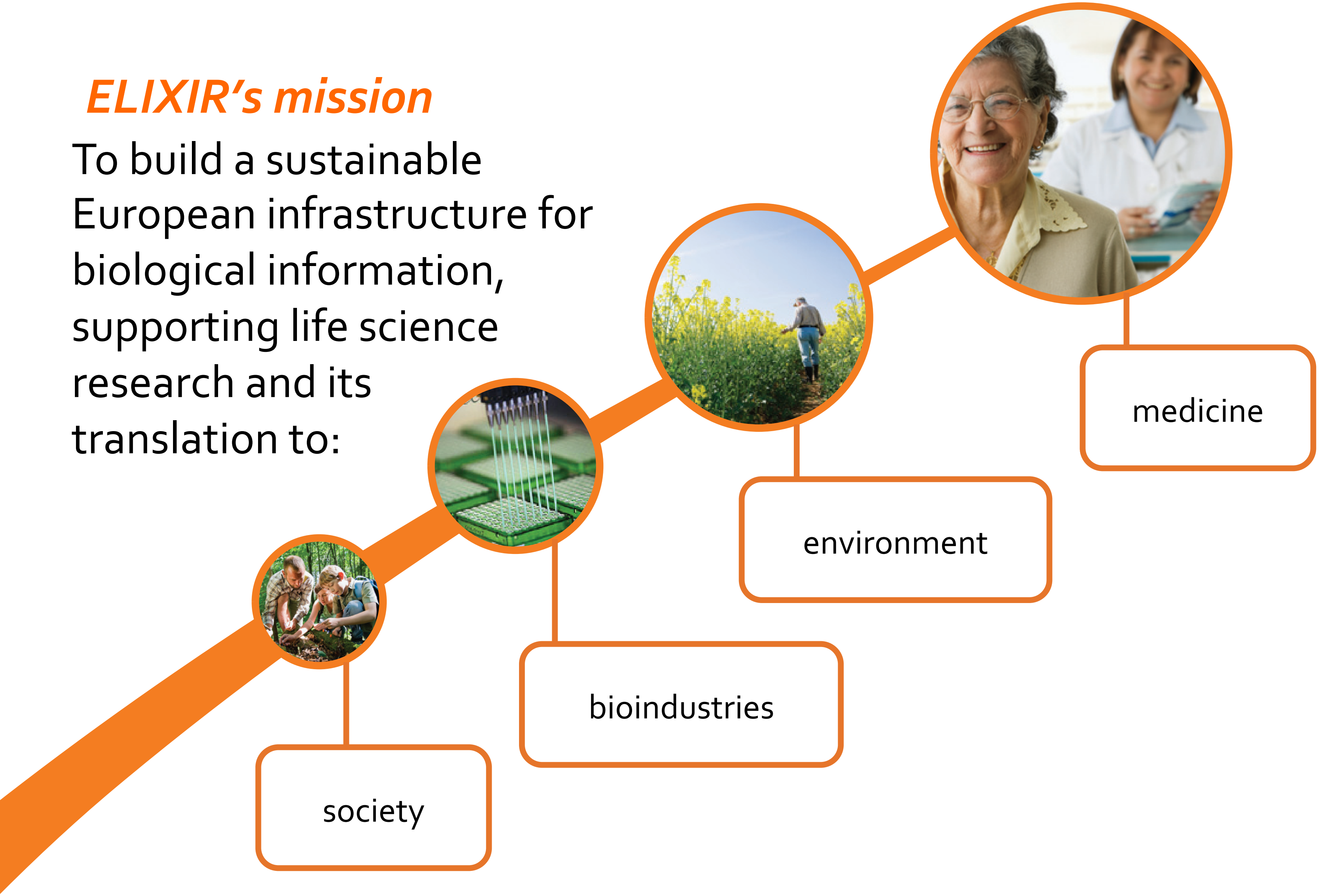
- How to use state-of-the-art data-intensive computing systems for biological data processing?
- Approach:
  - Scalable incremental updates
  - Unmodified data analysis tools
  - Integrated with Galaxy
  - Utilize data-intensive computing systems
- Edvard Pedersen (PhD student)
  - github.com/EdvardPedersen/GeStore
  - In Proc. of. EurPar'13, CIBB'14, PDP'15

Infrastructure

- Compute and storage resources
  - Systems for data management, parallel execution, accounting, data transfer, data integration, security…
  - Per lab? Per university? National?

ELIXIR's mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:



society

bioindustries

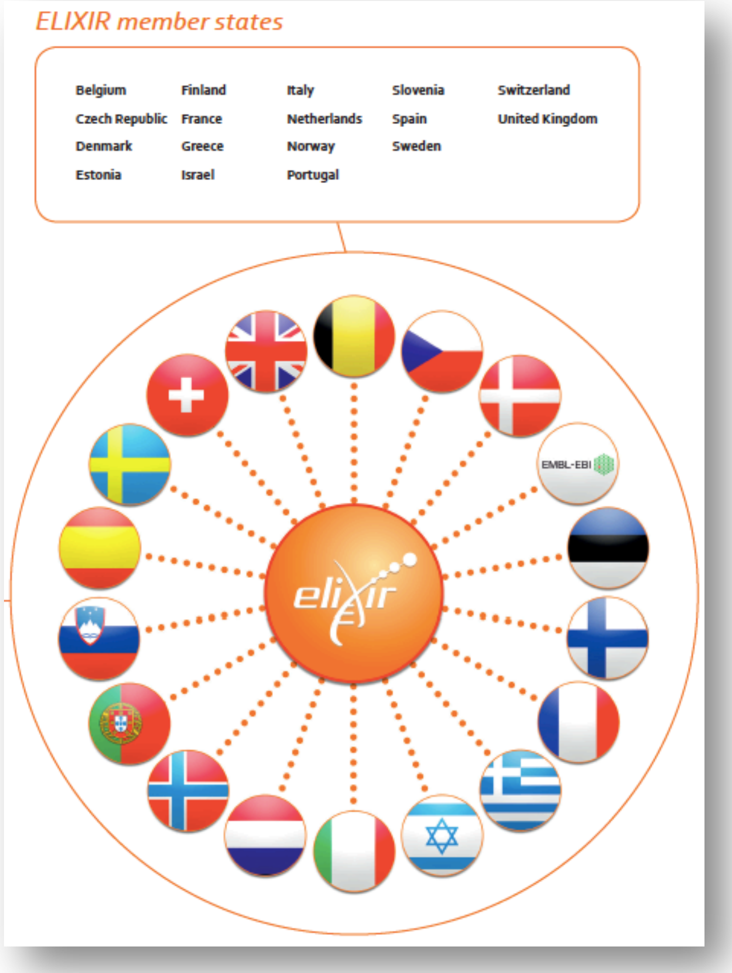
environment

medicine

ELIXIR Consortium Agreement (ECA)

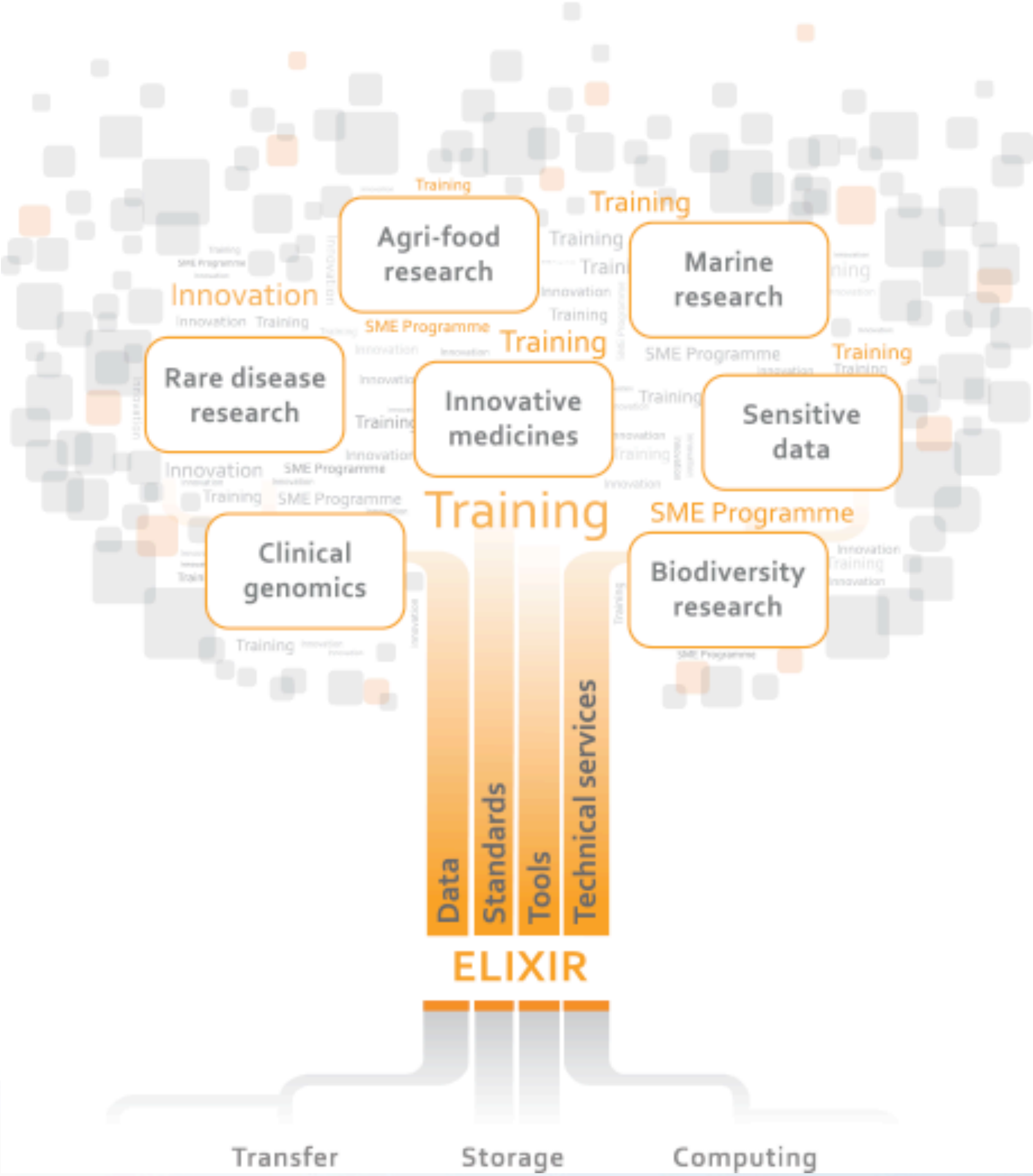
- 17 countries plus EMBL have signed the Memorandum of Understanding (MoU)

- 12 Countries has signing the ELIXIR Consortium Agreement (ECA)



ELIXIR member states

Belgium
Czech Republic
Denmark
Estonia
Finland
France
Greece
Israel
Italy
Netherlands
Norway
Portugal
Slovenia
Spain
Sweden
Switzerland
United Kingdom



Agri-food research

Marine research

Rare disease research

Innovative medicines

Sensitive data

Clinical genomics

Biodiversity research

Data

Standards

Tools

Technical services

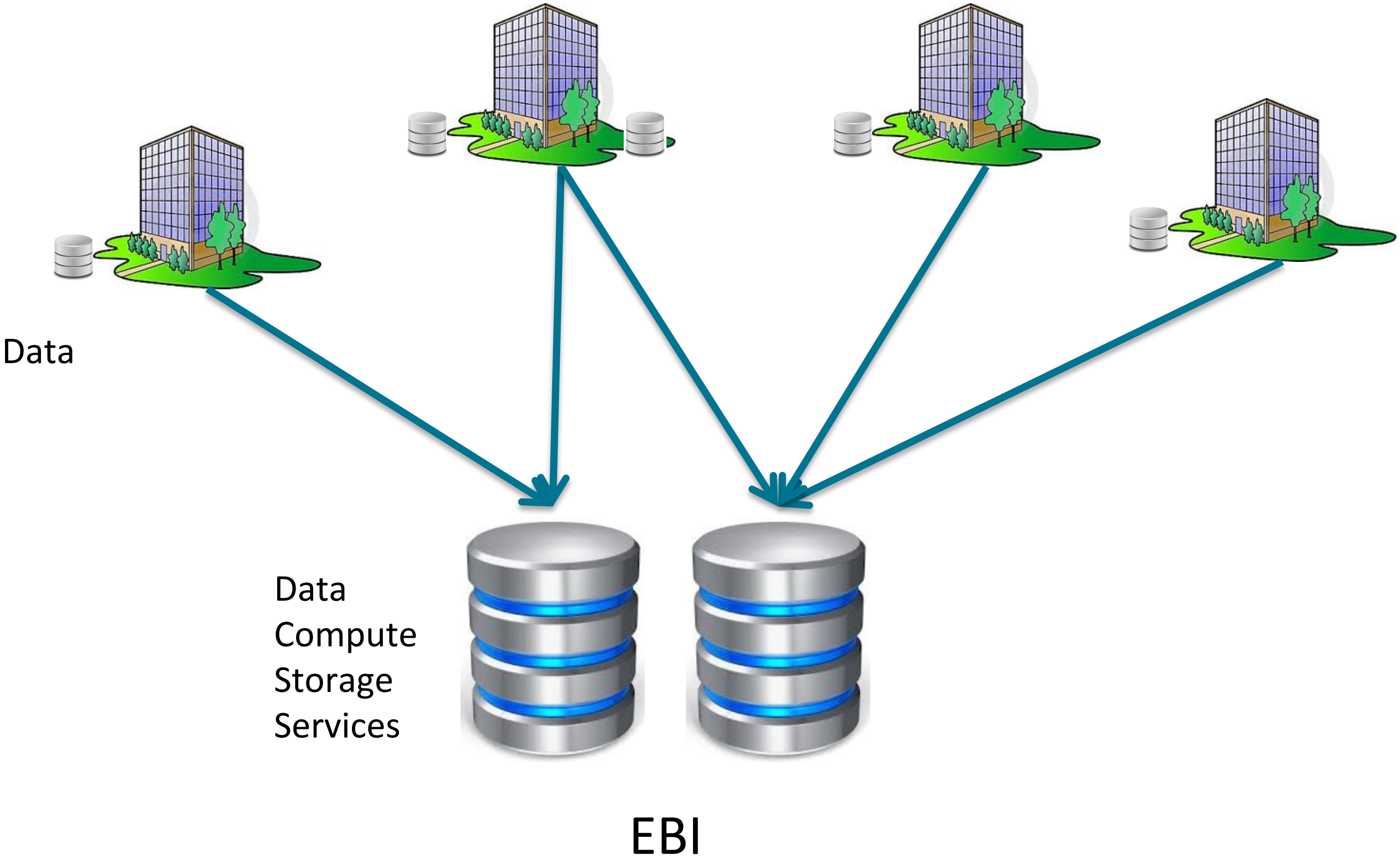
ELIXIR

Transfer

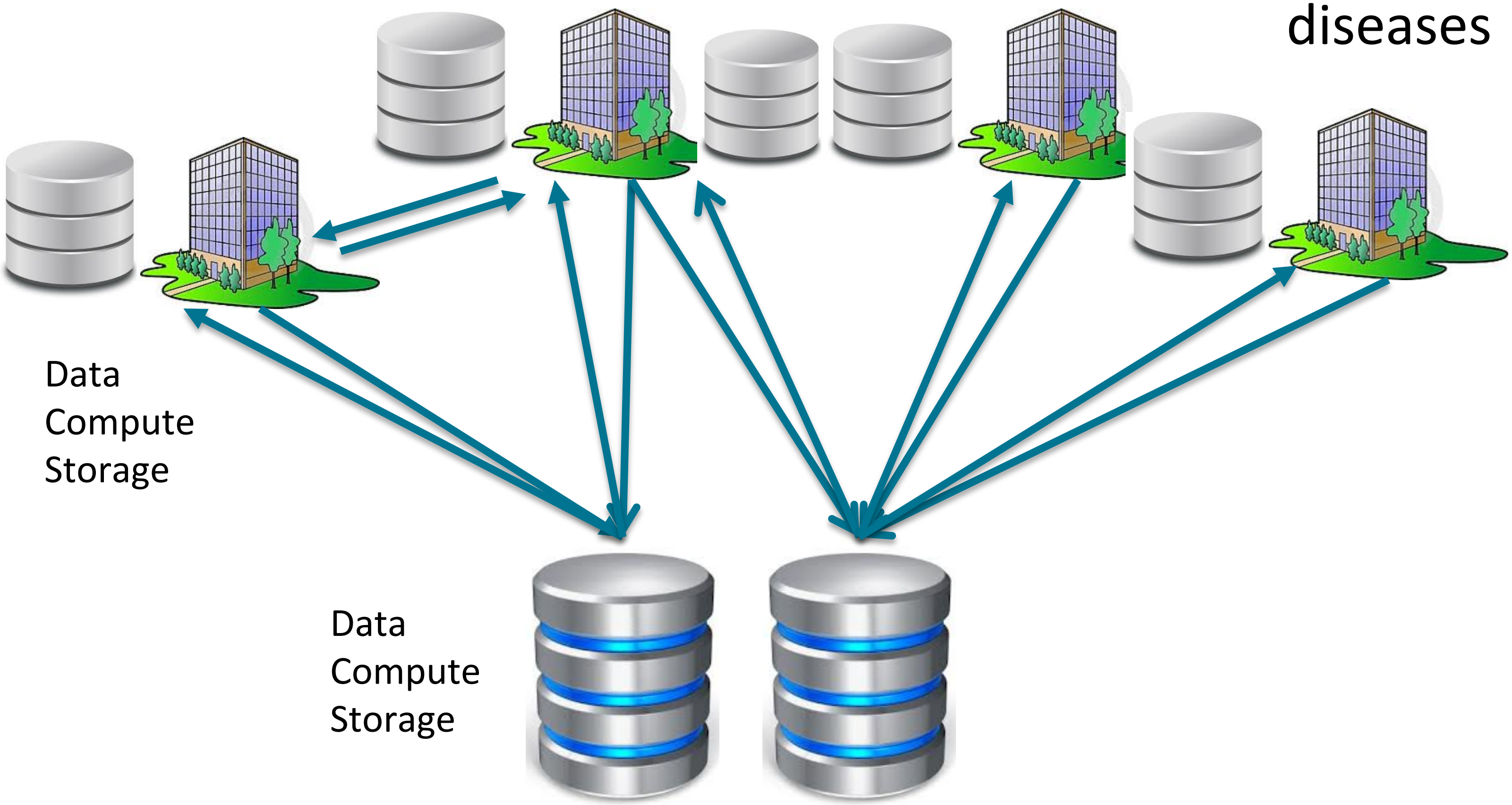
Storage

Computing

Current situation



Future



Marine metagenomics

Clinical genomics

Rare diseases

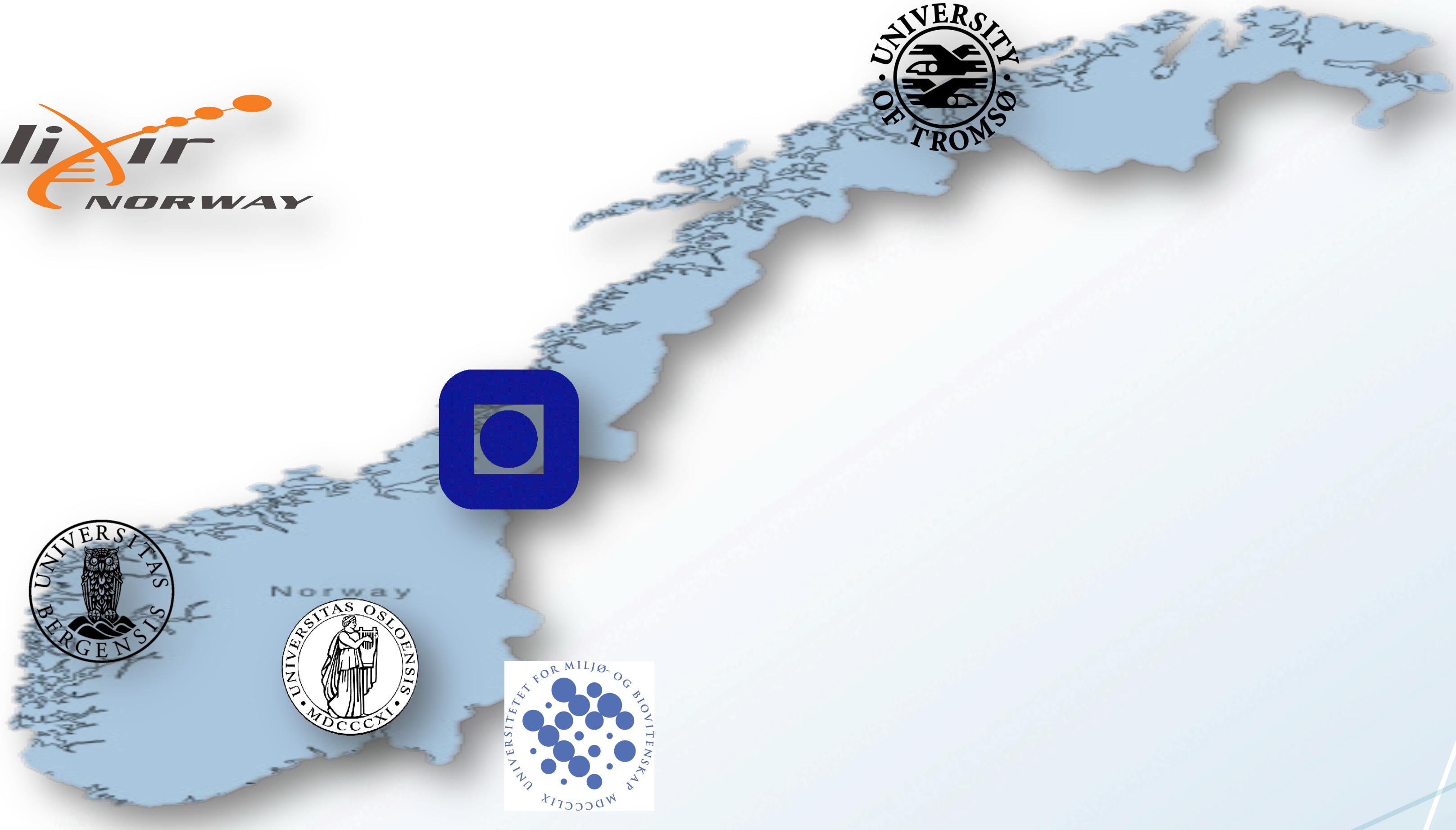
Data
Compute
Storage

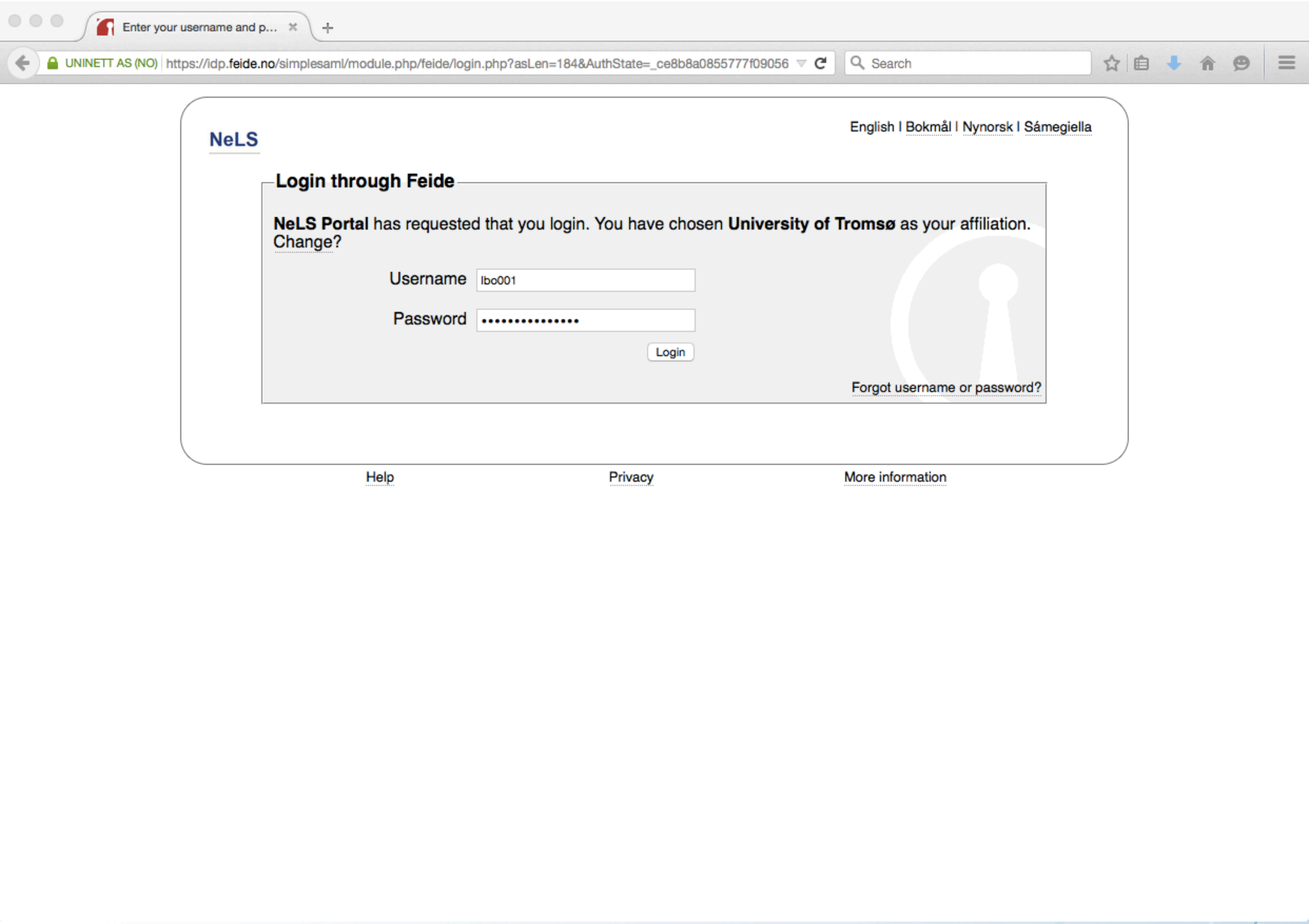
Data
Compute
Storage

EBI

ELIXIR-NO

elixir NORWAY


UNIVERSITY OF TROMSØ
UNIVERSITAS BERGENSIS
Norway
UNIVERSITAS OSLOENSIS MDCCCXI
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP MDCCCLIX



Enter your username and p…
UNINETT AS (NO)
Search

https://idp.feide.no/simplesaml/module.php/feide/login.php?asLen=184&AuthState=_ce8b8a0855777f09056

English | Bokmål | Nynorsk | Sámegiella

NeLS

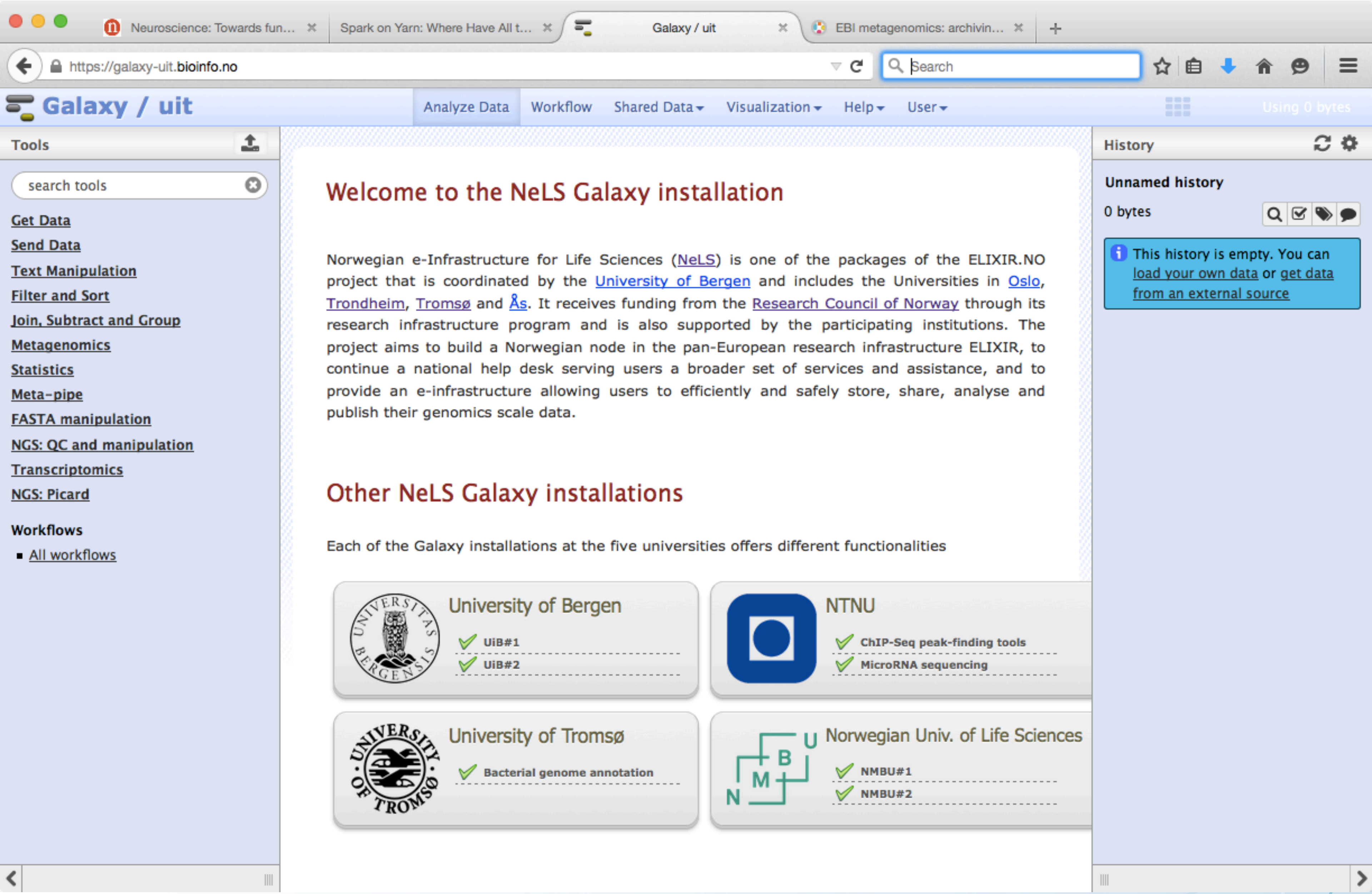
Login through Feide

NeLS Portal has requested that you login. You have chosen **University of Tromsø** as your affiliation. Change?

Username: lbo001

Password: •••••••••••••••

Login
Forgot username or password?
Help
Privacy
More information



This history is empty. You can load your own data or get data from an external source

Welcome to the NeLS Galaxy installation

Norwegian e-Infrastructure for Life Sciences (NeLS) is one of the packages of the ELIXIR.NO project that is coordinated by the University of Bergen and includes the Universities in Oslo, Trondheim, Tromsø and Ås. It receives funding from the Research Council of Norway through its research infrastructure program and is also supported by the participating institutions. The project aims to build a Norwegian node in the pan-European research infrastructure ELIXIR, to continue a national help desk serving users a broader set of services and assistance, and to provide an e-infrastructure allowing users to efficiently and safely store, share, analyse and publish their genomics scale data.

Other NeLS Galaxy installations

Each of the Galaxy installations at the five universities offers different functionalities

University of Bergen
- ✓ UiB#1
- ✓ UiB#2

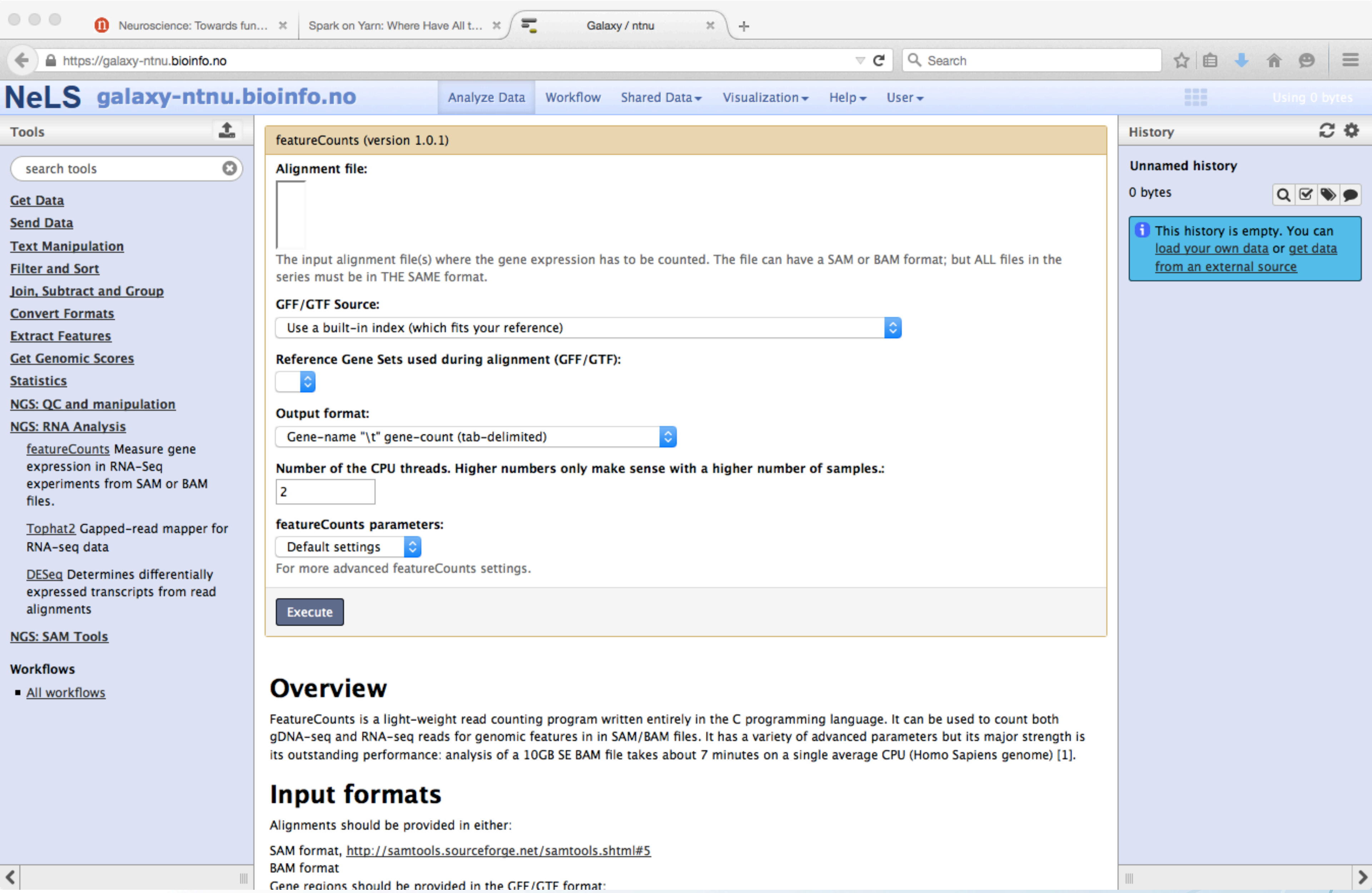
NTNU
- ✓ ChIP-Seq peak-finding tools
- ✓ MicroRNA sequencing

University of Tromsø
- ✓ Bacterial genome annotation

Norwegian Univ. of Life Sciences
- ✓ NMBU#1
- ✓ NMBU#2

Data Transfer

- Through web-browser in Galaxy
- Or, *scp* to Elixir-NO storage system
  - Over high-bandwidth networks
  - Data available in all Galaxy instances



featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files.

Tophat2 Gapped-read mapper for RNA-seq data

DESeq Determines differentially expressed transcripts from read alignments

featureCounts (version 1.0.1)

Alignment file:

The input alignment file(s) where the gene expression has to be counted. The file can have a SAM or BAM format; but ALL files in the series must be in THE SAME format.

GFF/GTF Source: Use a built-in index (which fits your reference)

Reference Gene Sets used during alignment (GFF/GTF):

Output format: Gene-name "\t" gene-count (tab-delimited)

Number of the CPU threads. Higher numbers only make sense with a higher number of samples.: 2

featureCounts parameters: Default settings

For more advanced featureCounts settings.

Overview

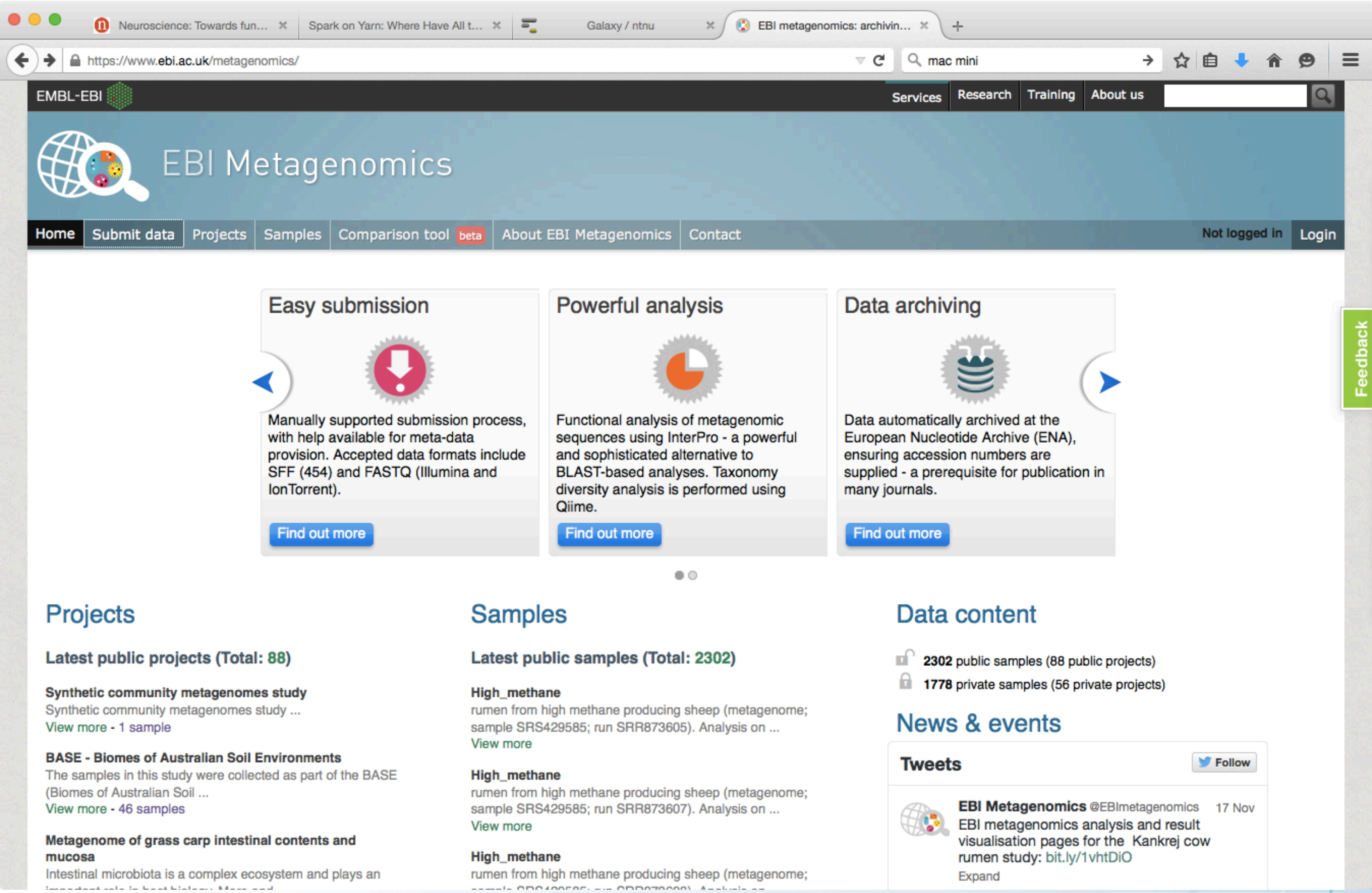
FeatureCounts is a light-weight read counting program written entirely in the C programming language. It can be used to count both gDNA-seq and RNA-seq reads for genomic features in in SAM/BAM files. It has a variety of advanced parameters but its major strength is its outstanding performance: analysis of a 10GB SE BAM file takes about 7 minutes on a single average CPU (Homo Sapiens genome) [1].

Input formats

Alignments should be provided in either:

SAM format, http://samtools.sourceforge.net/samtools.shtml#5
BAM format
Gene regions should be provided in the GFF/GTF format:

This history is empty. You can load your own data or get data from an external source



EBI Metagenomics

Easy submission

Manually supported submission process, with help available for meta-data provision. Accepted data formats include SFF (454) and FASTQ (Illumina and IonTorrent).

Powerful analysis

Functional analysis of metagenomic sequences using InterPro - a powerful and sophisticated alternative to BLAST-based analyses. Taxonomy diversity analysis is performed using Qiime.

Data archiving

Data automatically archived at the European Nucleotide Archive (ENA), ensuring accession numbers are supplied - a prerequisite for publication in many journals.

Projects

Latest public projects (Total: 88)

Synthetic community metagenomes study
Synthetic community metagenomes study ...
View more - 1 sample

BASE - Biomes of Australian Soil Environments
The samples in this study were collected as part of the BASE (Biomes of Australian Soil ...
View more - 46 samples

Metagenome of grass carp intestinal contents and mucosa
Intestinal microbiota is a complex ecosystem and plays an

Samples

Latest public samples (Total: 2302)

High_methane
rumen from high methane producing sheep (metagenome; sample SRS429585; run SRR873605). Analysis on ...
View more

High_methane
rumen from high methane producing sheep (metagenome; sample SRS429585; run SRR873607). Analysis on ...
View more

High_methane
rumen from high methane producing sheep (metagenome;

Data content

2302 public samples (88 public projects)

1778 private samples (56 private projects)

News & events

Tweets

EBI Metagenomics @EBImetagenomics 17 Nov
EBI metagenomics analysis and result visualisation pages for the Kankrej cow rumen study: bit.ly/1vhtDiO

Summary

- Many labs implement their own visualizations tools
  - Does anybody else use them?
  - Kvik: framework for implementing visualization tools
- Software stack for biological data processing is small
  - Are our analysis tools less complex then mobile apps?
  - Mr. Clean: data-cleaning tool
- Data analysis is run on platforms built for batch processing
  - Why are data-intensive computing systems not used?
  - GeStore: big data management
- Many labs maintain their own compute and storage resources
  - Is this reliable? Is this efficient?
  - Elixir: distributed infrastructure

Collaborators

NOWAC

- Eiliv Lund
- Bjørn Fjukstad
- Einar Holsbø
- Kenneth Knudsen
- Karina Olsen
- Mie Jareid
- Hege Bøvelstad
- Nicolle Mode
- Etienne Birmelé (Université Paris Descartes)
- Lars & Marit Holden (Norsk Regnesentral)

ELIXIR

- Nils Peder Willassen
- Edvard Pedersen
- Inge Alexander Raknes
- Ida Jaklin Johansen
- Erik Hjerde
- Espen M. Robertsen
- Roy Dragseth
- Rob Finn (EBI)