AIDA Data Hub

Services for Clinical Innovation in Medical Imaging Diagnostic AI.

National data infrastructure supporting the Analytic Imaging Diagnostic Arena (AIDA)
Hosted by LiU and the Center for Medical Image Science and Visualization (CMIV)
Funded by SciLifeLab Bioinformatics platform (NBIS)

230502 AIDA & AIDA Data Hub for Infralife

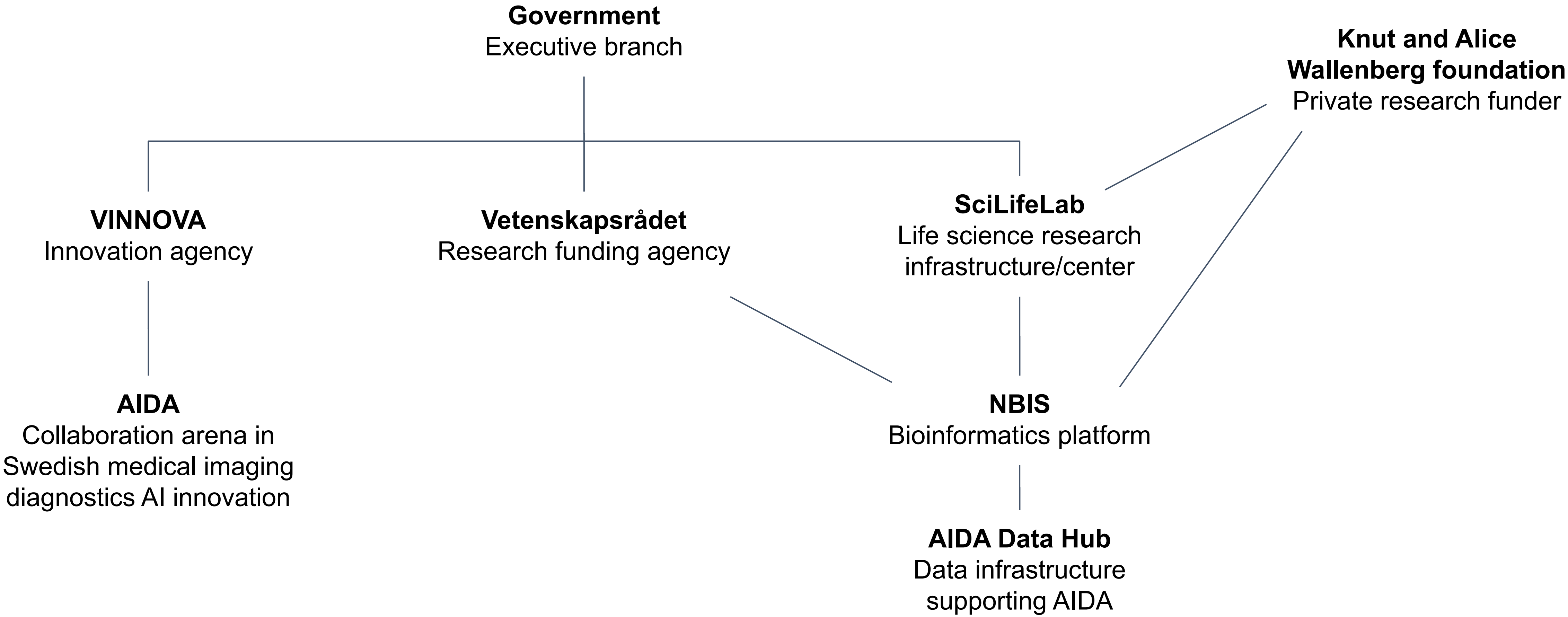
AIDA & AIDA Data Hub

AIDA Community - medtech4health.se/aida

National collaboration arena in AI research and innovation in medical imaging diagnostics.

AIDA Data Hub - datahub.aida.scilifelab.se

The data infrastructure supporting AIDA.



Government
Executive branch

Knut and Alice Wallenberg foundation
Private research funder

VINNOVA
Innovation agency

Vetenskapsrådet
Research funding agency

SciLifeLab
Life science research infrastructure/center

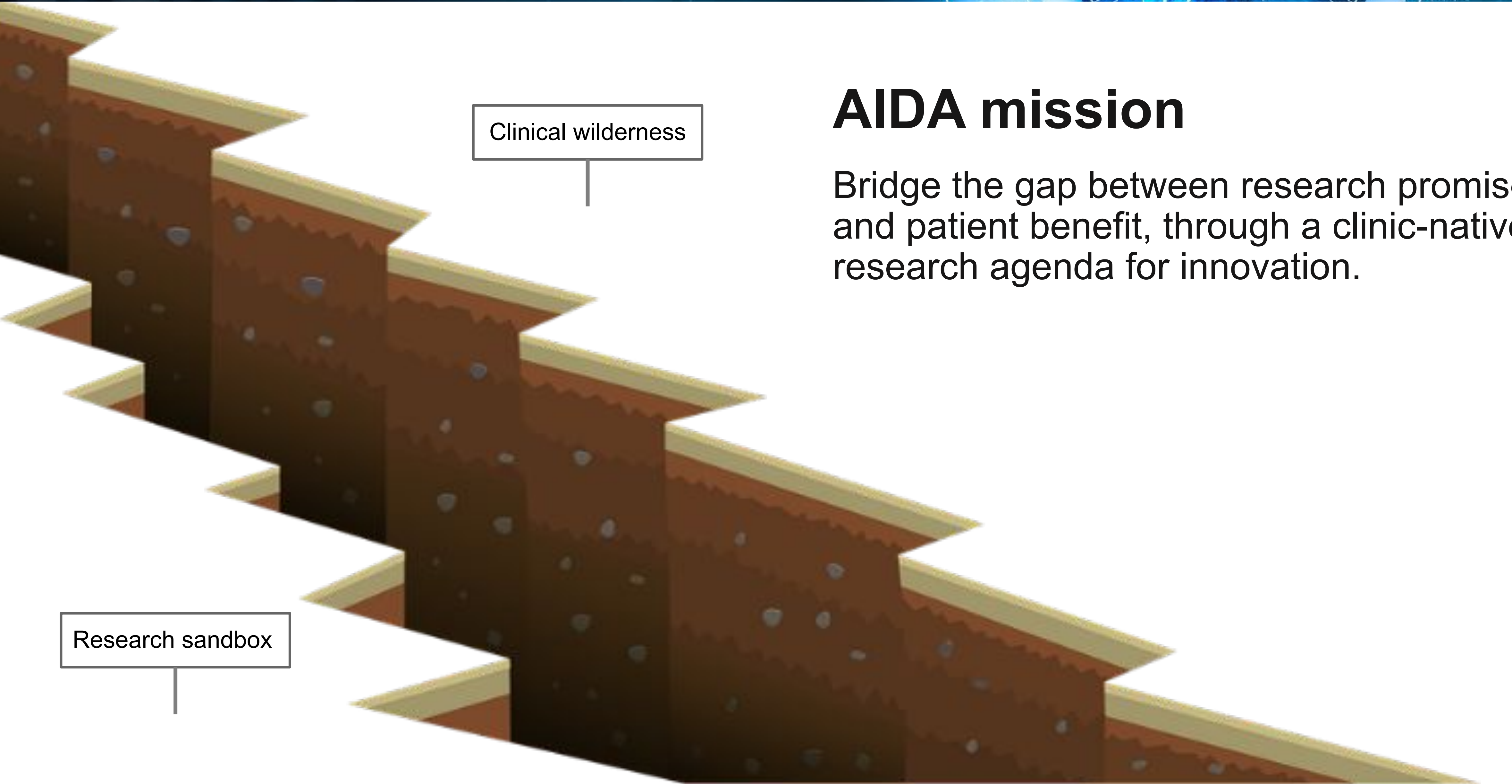
AIDA
Collaboration arena in Swedish medical imaging diagnostics AI innovation

NBIS
Bioinformatics platform

AIDA Data Hub
Data infrastructure supporting AIDA

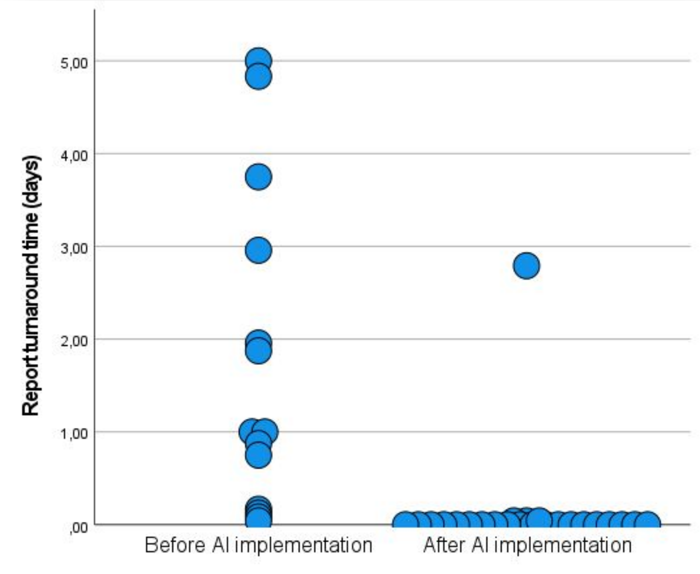
AIDA mission

Bridge the gap between research promise and patient benefit, through a clinic-native research agenda for innovation.

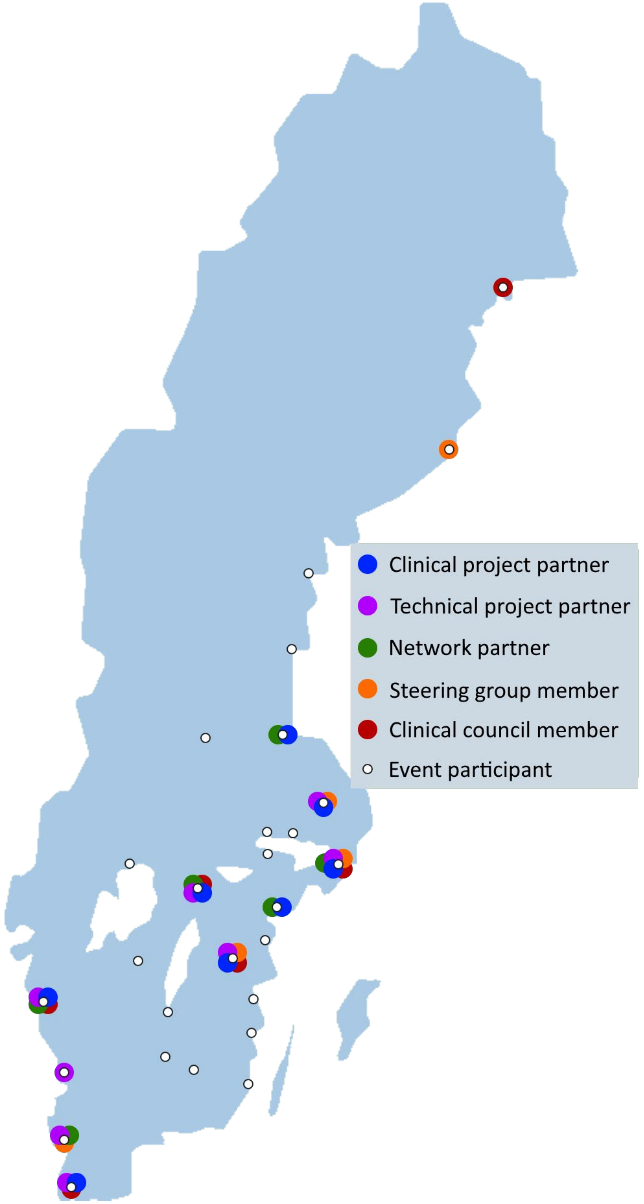


Success story: Region Halland

- Participated in AI course
- Participated in AI showcase event
- Interest in pulmonary embolism tool
- Started clinical evaluation and implementation
- 2022: Patient benefit achieved



Wiklund et al., Incidental pulmonary... , Euro Radiol. 2022

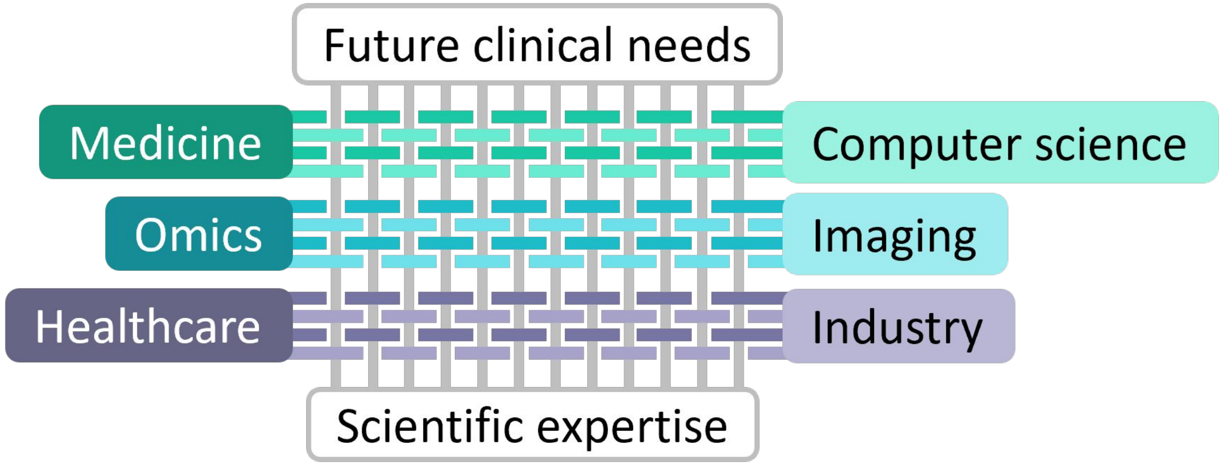

● Clinical project partner
● Technical project partner
● Network partner
● Steering group member
● Clinical council member
○ Event participant

AIDA Community

Publicly funded collaboration arena for AI innovation in medical imaging diagnostics.

- ~50 partners, ~60 projects
- Academia, industry and healthcare
- Research & innovation projects
- Fellowships & Clinical evaluations
- Incubator for AI validation
- Training

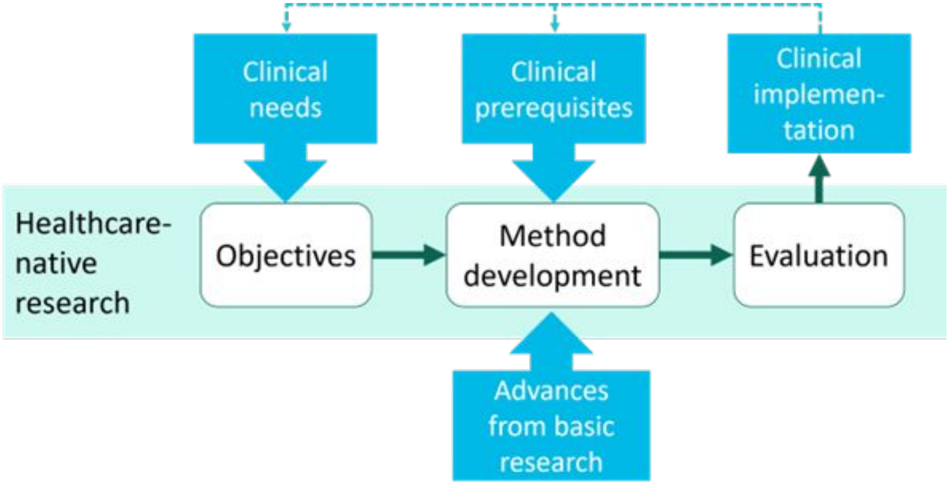
AIDA community principles



Multi-cross-disciplinary life science innovation with a healthcare-native agenda.

- Data-driven methods, bringing medicine and computer science together.
- Bridging diagnostic specialties, where findings in one diagnostic silo need to be reviewed in the context of the others.
- Intersectoriality, where tight collaborations between academia and industry is needed for arriving at healthcare improvements.

AIDA healthcare-native agenda



Clinical needs

Clinical prerequisites

Clinical implemen-tation

Healthcare-native research

Objectives

Method development

Evaluation

Advances from basic research

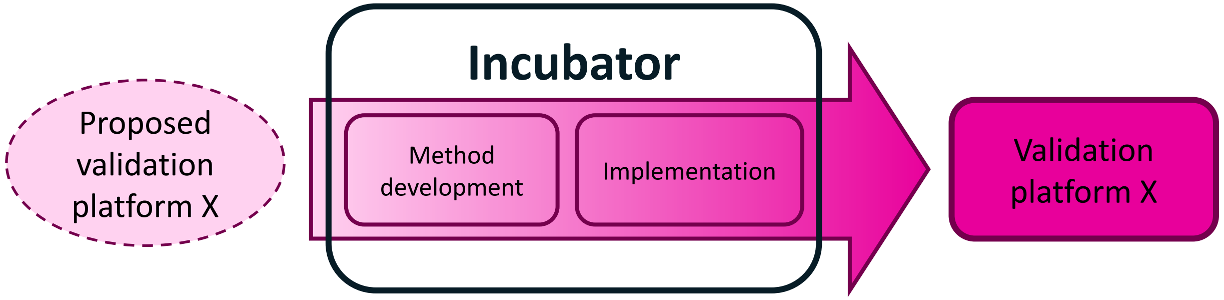
- Proximity to clinical reality permeates all of the research agenda.
- Research objectives are directly gathered from clinical needs, current or distinctly identified future ones.
- Tight interaction with clinical counterparts throughout the research studies.

I.e: AIDA activities are not waterfalls that start from identified scientific knowledge gaps and only eventually considering possible clinical implementations.

VAI.B

Validation
Platform for AI
in Breast Imaging

Example:
AIDA VAI AI validation incubator



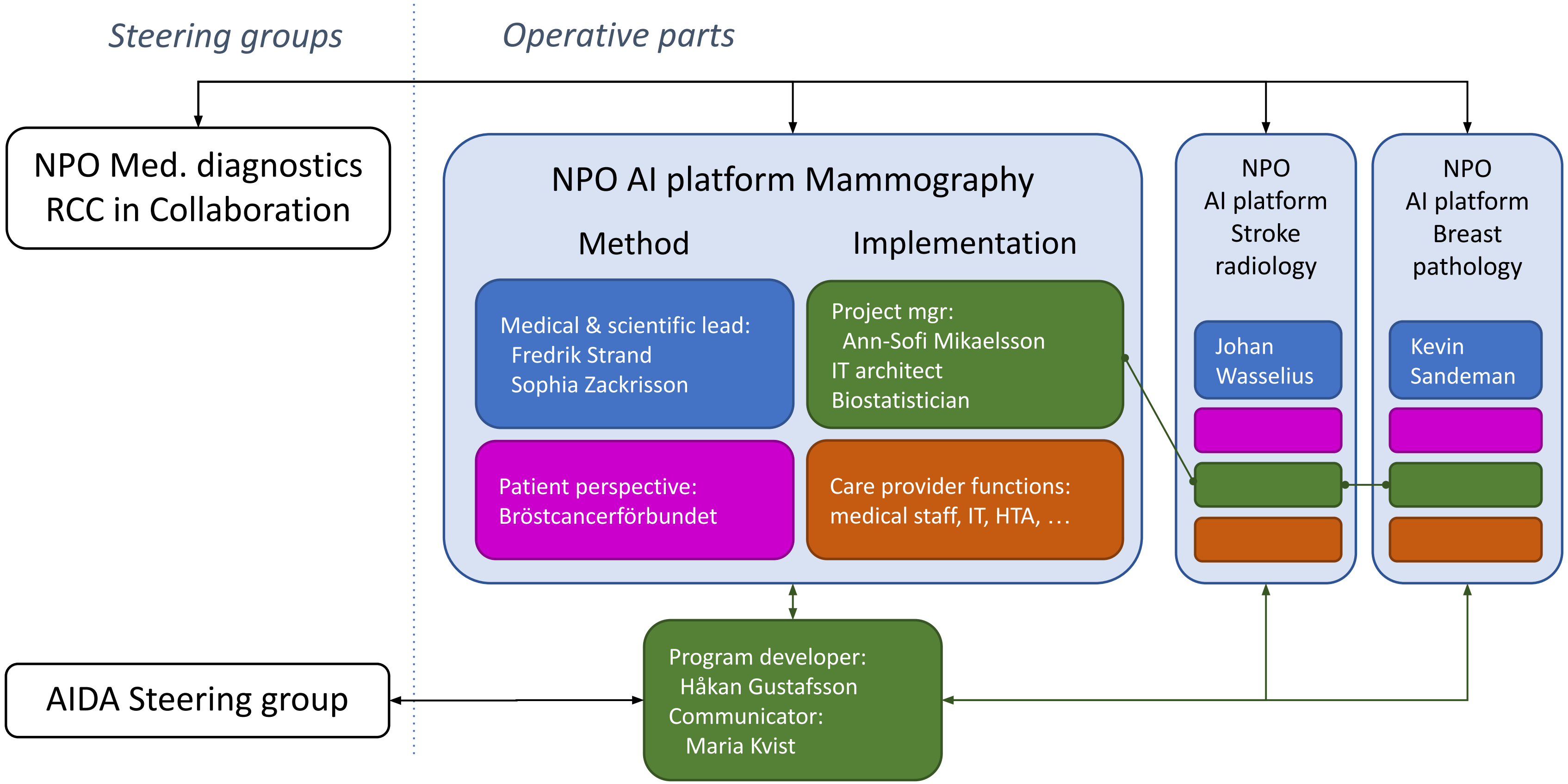
AIDA helps healthcare set up national topical AI validation platforms for:

1. VAI-B for Mammography
2. VAI-S for Stroke radiology
3. VAI-P for Breast pathology

Clinics can evaluate existing AI tools in private using data from their own patients.

AI vendors can reach all clinics in Sweden with a single deployment, and can get real world performance benchmarks.

Example: AIDA VAI organization



Steering groups

Operative parts

NPO Med. diagnostics RCC in Collaboration

NPO AI platform Mammography

Method

Implementation

Medical & scientific lead:
Fredrik Strand
Sophia Zackrisson

Project mgr:
Ann-Sofi Mikaelsson
IT architect
Biostatistician

Patient perspective:
Bröstcancerförbundet

Care provider functions:
medical staff, IT, HTA, …

NPO AI platform Stroke radiology

Johan Wasselius

NPO AI platform Breast pathology

Kevin Sandeman

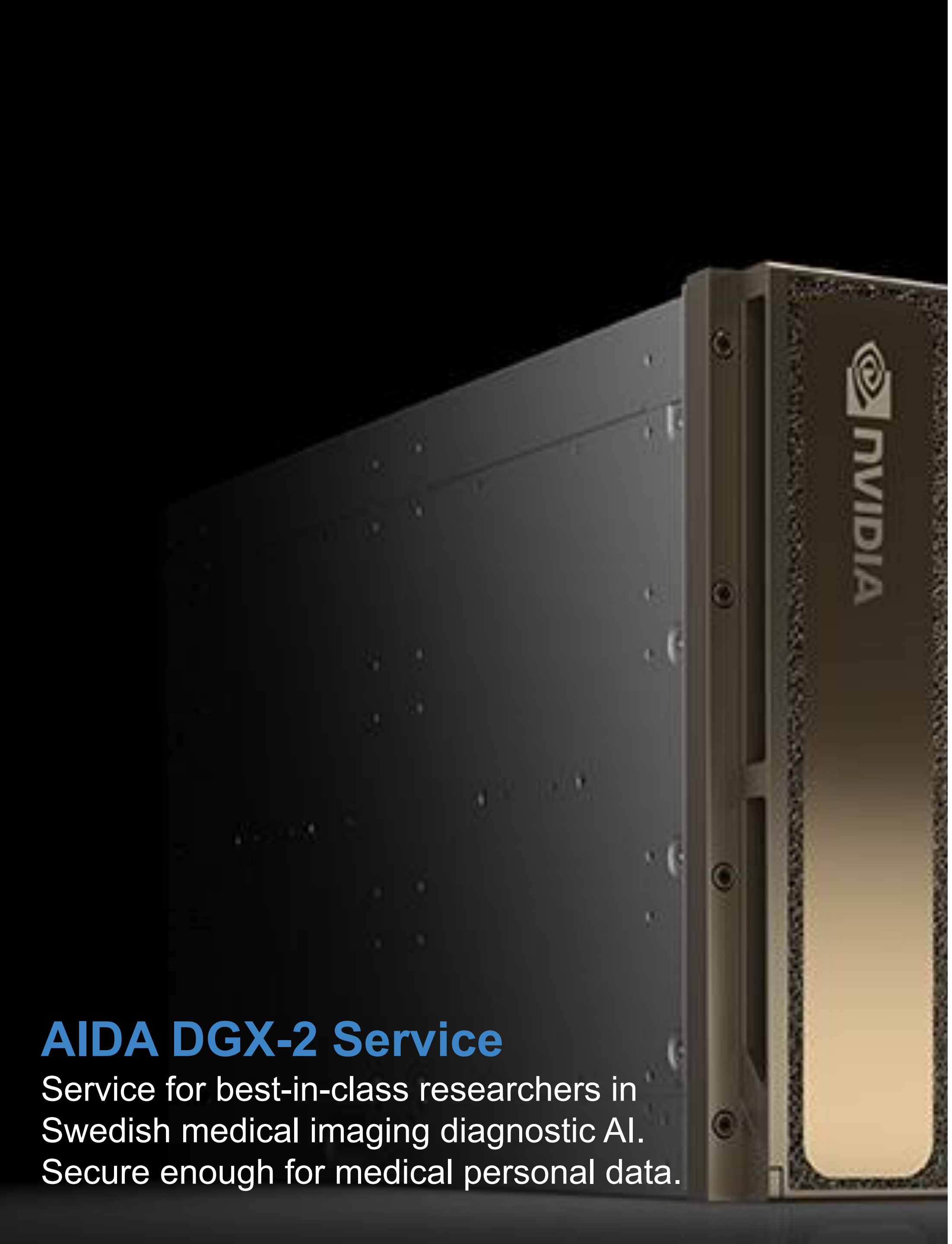
AIDA Steering group

Program developer:
Håkan Gustafsson
Communicator:
Maria Kvist

AIDA Data Hub

Data infrastructure supporting AIDA with:

- Compute systems for AI training on sensitive personal medical data.
- Data sharing & support.
- Cover costs for extraction of prioritized clinical data for research.
- Ethics and legal policy support.
- System design expertise.
- **New**: AI development expertise.



AIDA DGX-2 Service

Service for best-in-class researchers in Swedish medical imaging diagnostic AI. Secure enough for medical personal data.

Secure AI training systems

Utilization: 96%

Set up at CMIV in collaboration with Nvidia.

Hosting VINNOVA funded SCAPIS data lab, where AI researchers can securely process SCAPIS data for research.

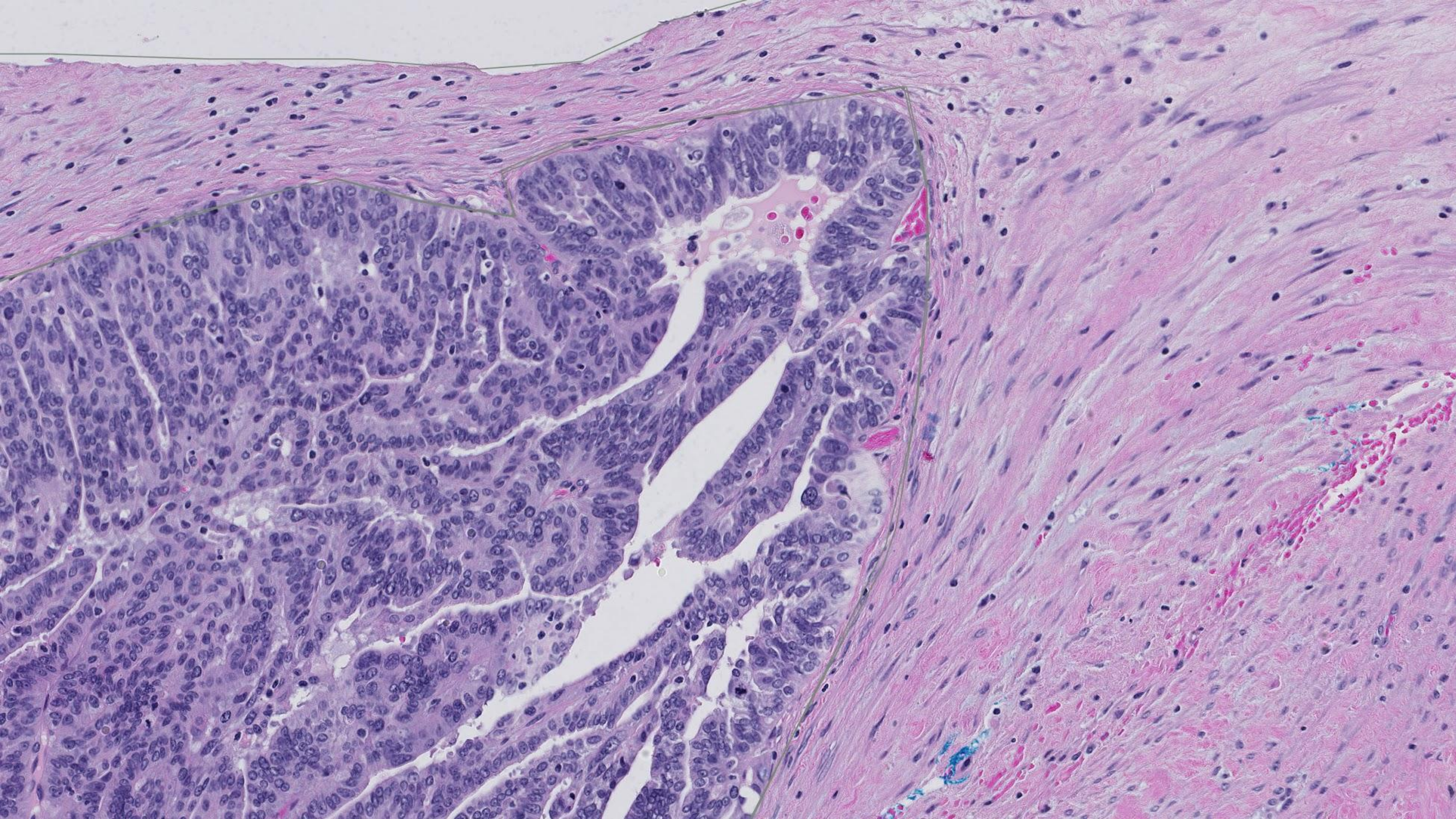
Expansion planned 2023q3, in collaboration with RÖ, DDLS and Berzelius, contributing to implementation of DDLS and EUCAIM data service platforms, and the upcoming Linköping Health Data Spaces.

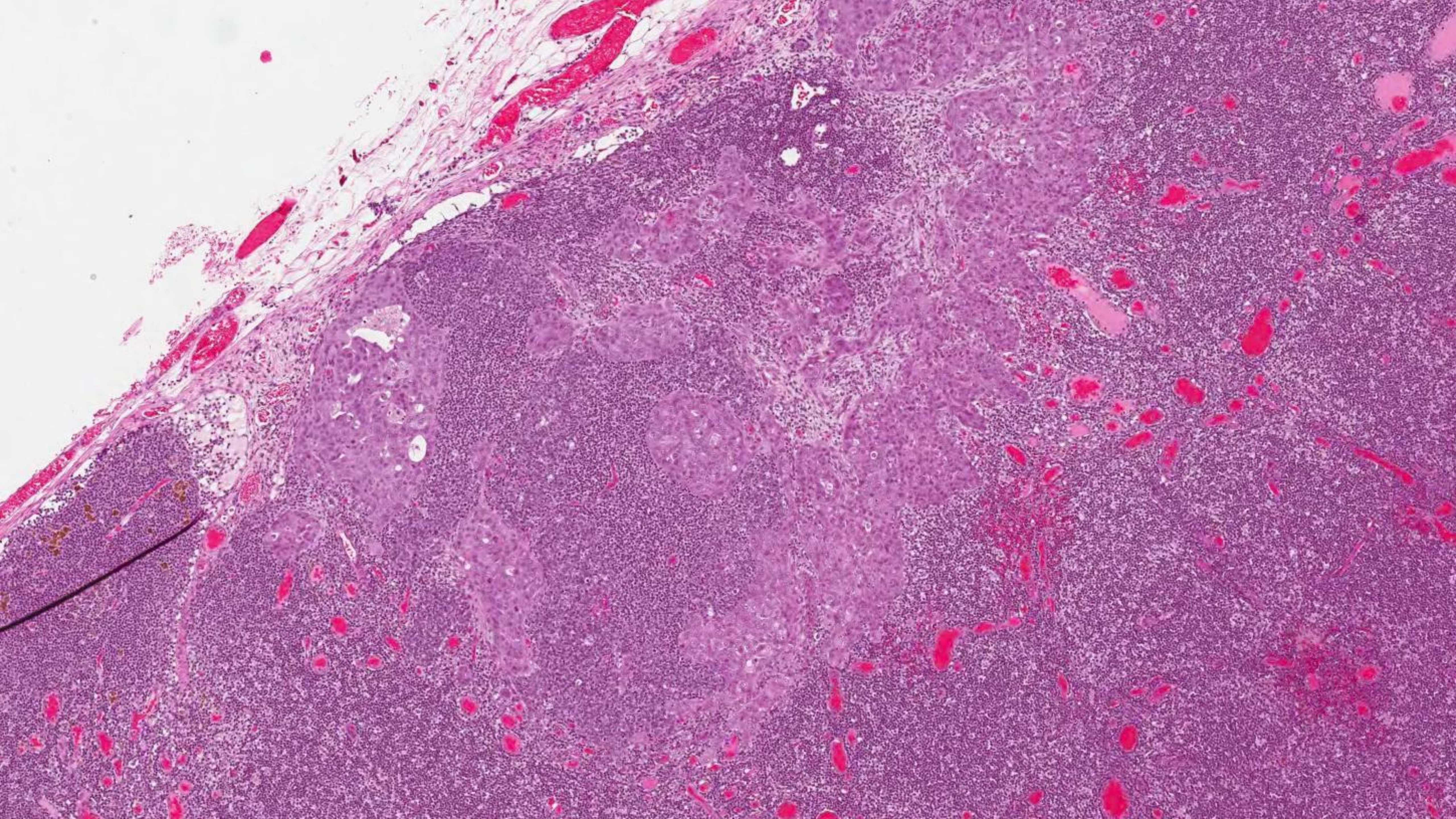
Data in

Metrics:

- Datasets: 20   12.3TB
- Modalities: 5
- Organs: 13

	Datasets	Scans	Annotations	Size
Total	**20**	**32081**	**39093**	**12.3TB**
Annotated	11	4190	38401	1.80TB
Pathology	9	11881	34020	10.74TB
Radiology	11	20200	5073	1.56TB

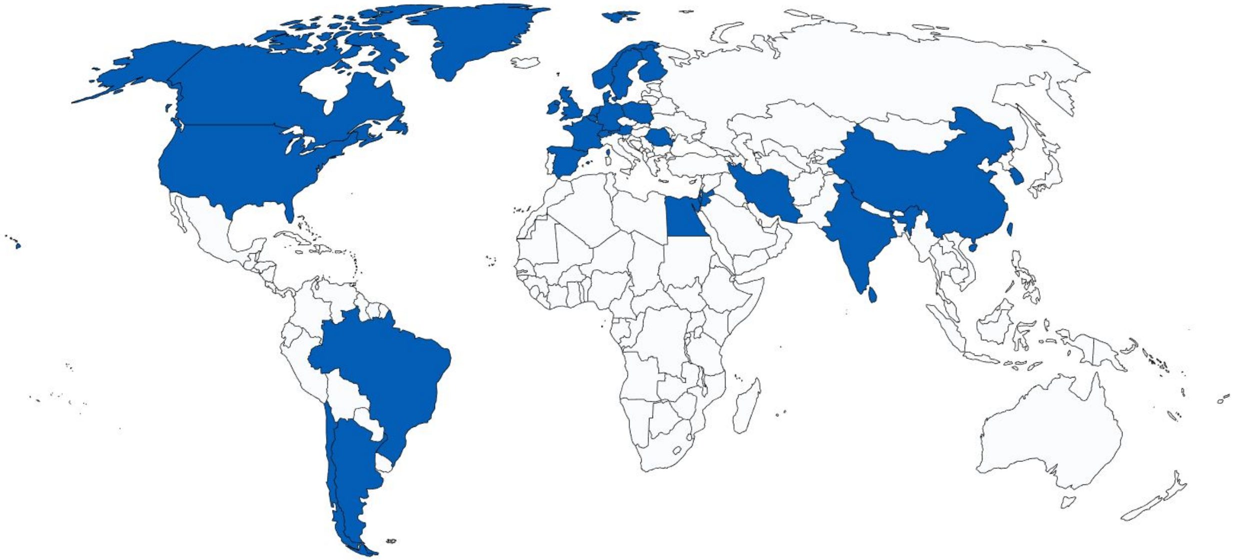




Data sharing worldwide

Metrics:

- Countries: 29
- External sharing events: 153



Policy support

AIDA Data Sharing Policy

Comprehensive resource describing best practices in handling and sharing medical imaging data for research in Sweden and similar countries.

Concrete guidelines and examples, with references to original sources in law.

Key insights have been published in Nature Scientific Data (OpenAccess).

Using Clinical Imaging Data for Research

Common practice in Sweden and similar countries, 1-paragraph summary:

The common practice is that caregivers disclose data to research institutions for specific activities described in approved ethical review applications, to be carried out under appropriate technical and organizational protective measures and supervised by a named competent researcher. The research institution is then data controller and copyright holder for the disclosed data, and is responsible for ensuring that data is processed and shared only as described in the approved ethical review application, with data processing agreements, pseudonymization, anonymization and licensing as tools, and with an obligation to store relevant data for 10 years after last use for purposes of research validation.

AI development expertise

Advanced user support and training to the AIDA community.

- Provide a core resource with deep technical expertise
- Support junior researchers
- Reduce startup latencies
- Facilitate knowledge transfer

Application expertise

Establish new support function.

Advanced user support and training to the AIDA community, in medical imaging diagnostics AI research and innovation.

Focus on projects with clear connection to the broader SciLifeLab aims, including precision medicine and multi-omics.

Cooperate with similar functions and development units at NBIS and BIIF.

Tryggve

Nordic collaboration on Sensitive personal data for research

Joel Hedlund, Executive manager, Senior advisor

neic

NORDIC E-INFRASTRUCTURE COLLABORATION

Federated EGA

Federated EGA strives to support the discovery of and secure access to human data globally, while respecting national data protection regulations, with the goal of accelerating disease research and understanding and improving human health.

SCAPIS Data Lab

Working with SCAPIS to make all imaging data available to approved research groups as read-only datasets through AIDA Data Hub (~100 TB).

Storage system has been extended to allow commencing upload.

RÖ-LiU Linköping Health Data Spaces

Double data lake systems for primary and secondary use of health data

AIDA Data Hub is the current LiU data lake

Bigpicture

Petabyte platform for European digital pathology AI

AIDA Data Hub leading repository infrastructure development, which is carried out in collaboration with sensitive data teams at the NBIS Systems Development unit and CSC.fi.

First three clinical datasets received, large scale archive operations start Mar 2023.

EUCAIM

EUCAIM Federated infrastructure for cancer imaging data

AIDA Data Hub contributing data collaboration workspaces for use in EUCAIM with cancer imaging data based on Bigpicture Federated node technologies.

Collaboration with sensitive data teams at the NBIS Systems Development unit.

Thank you!

AIDA Data Hub
Services for Clinical Innovation in Medical Imaging Diagnostic AI

National data infrastructure supporting the Analytic Imaging Diagnostic Arena (AIDA)
Hosted by LiU and the Center for Medical Image Science and Visualization (CMIV)
Part of SciLifeLab Bioinformatics platform (NBIS)

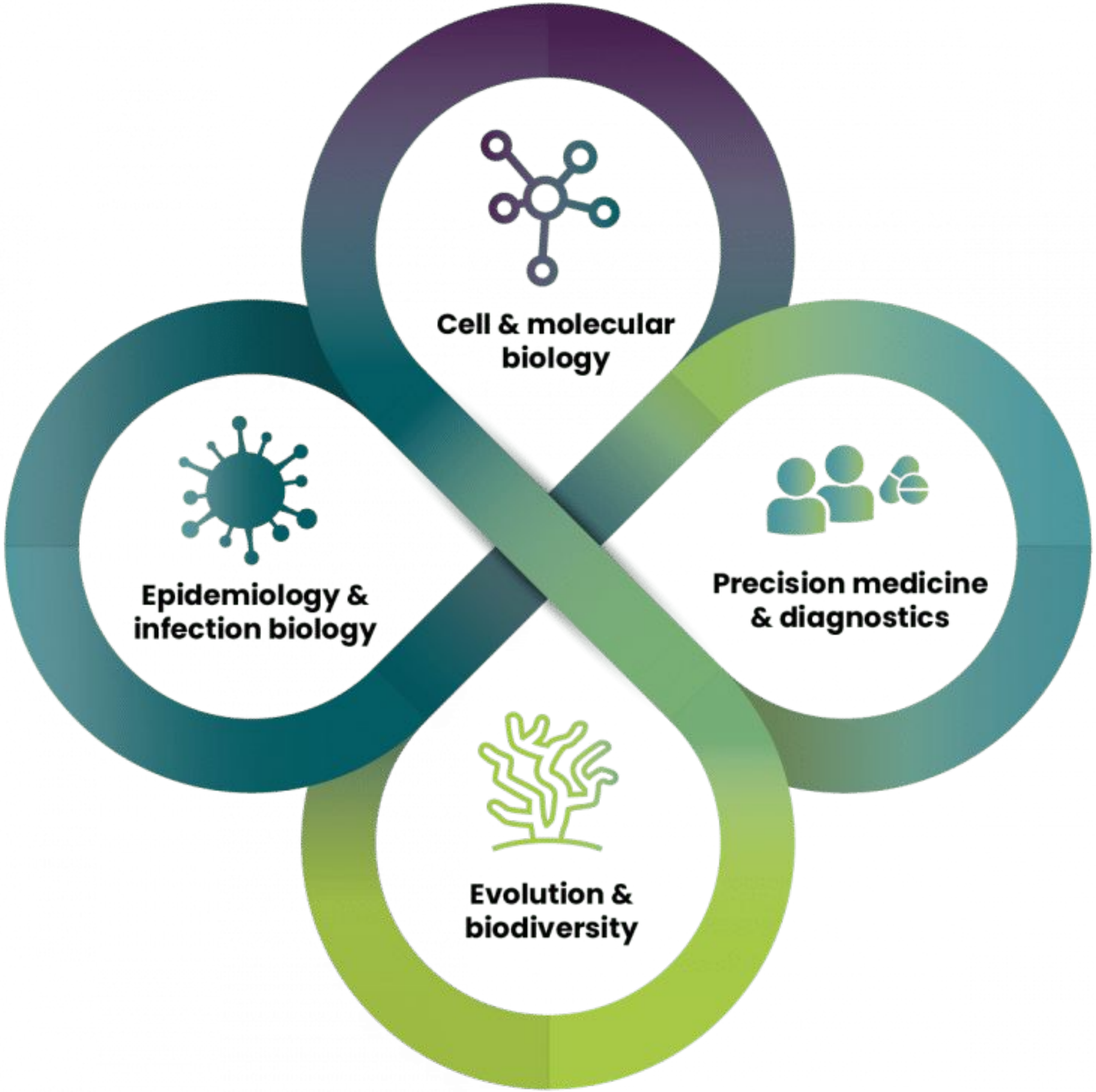
Extra slides in case of questions...

AIDA Data Hub supporting

Data-Driven Life Science

Increasing access
to clinical data for research

Engaging in data platform
and policy development


Cell & molecular biology
Epidemiology & infection biology
Precision medicine & diagnostics
Evolution & biodiversity

NVIDIA

AIDA DGX-2 Service

Service for best-in-class researchers in
Swedish medical imaging diagnostic AI.
Secure enough for medical personal data.

Design Vision

Extremely Powerful and Completely Safe

Design Vision

Extremely Powerful and Completely Safe

AIDA DGX-2 Service for Personal Data

AIDA is collaborating with Nvidia to offer up a DGX-2 machine learning system set up at CMIV as a service for leading edge researchers in Swedish medical imaging diagnostics AI.

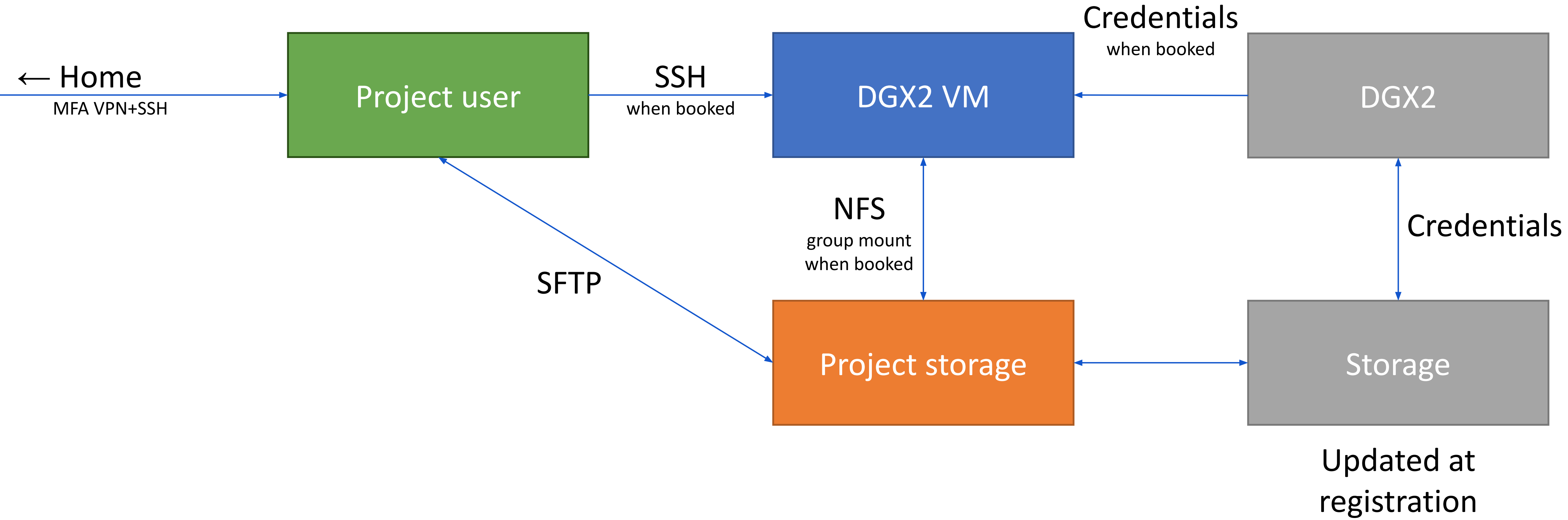
Establishment of this service was carried out in a phased approach, where full use of the system was provided to expert users from day 1, and further functionality, stability and guarantees were added in successive phases.

Establishment is now complete. The service has entered stable operations, and has been validated secure enough for processing sensitive personal data.

User model

- Service provided to PI under DPA, who can delegate full authorities.
- Full capabilities available to experts.
  - Persistent project storage.
  - Private virtual machines with powerful GPUs, where you are root.
- Booking via booking sheet, contact aida-compute or chat for practicalities.
- Work with your own data (AIDA Data Hub datasets available on request).
- Outgoing connections only to approved destinations per project.
- Login with MFA VPN + SSH pubkey

Design



Color: Fun things for you!

Gray: Stuff the system administrator has to deal with.

How to book time on the DGX-2

- Fill in an excel sheet.

How to book time on the DGX-2

- Go to the booking sheet. Read instructions (or keep listening :-)
- Find your name in the group information list on the right.
- You have a GPU budget. Past bookings don't count.

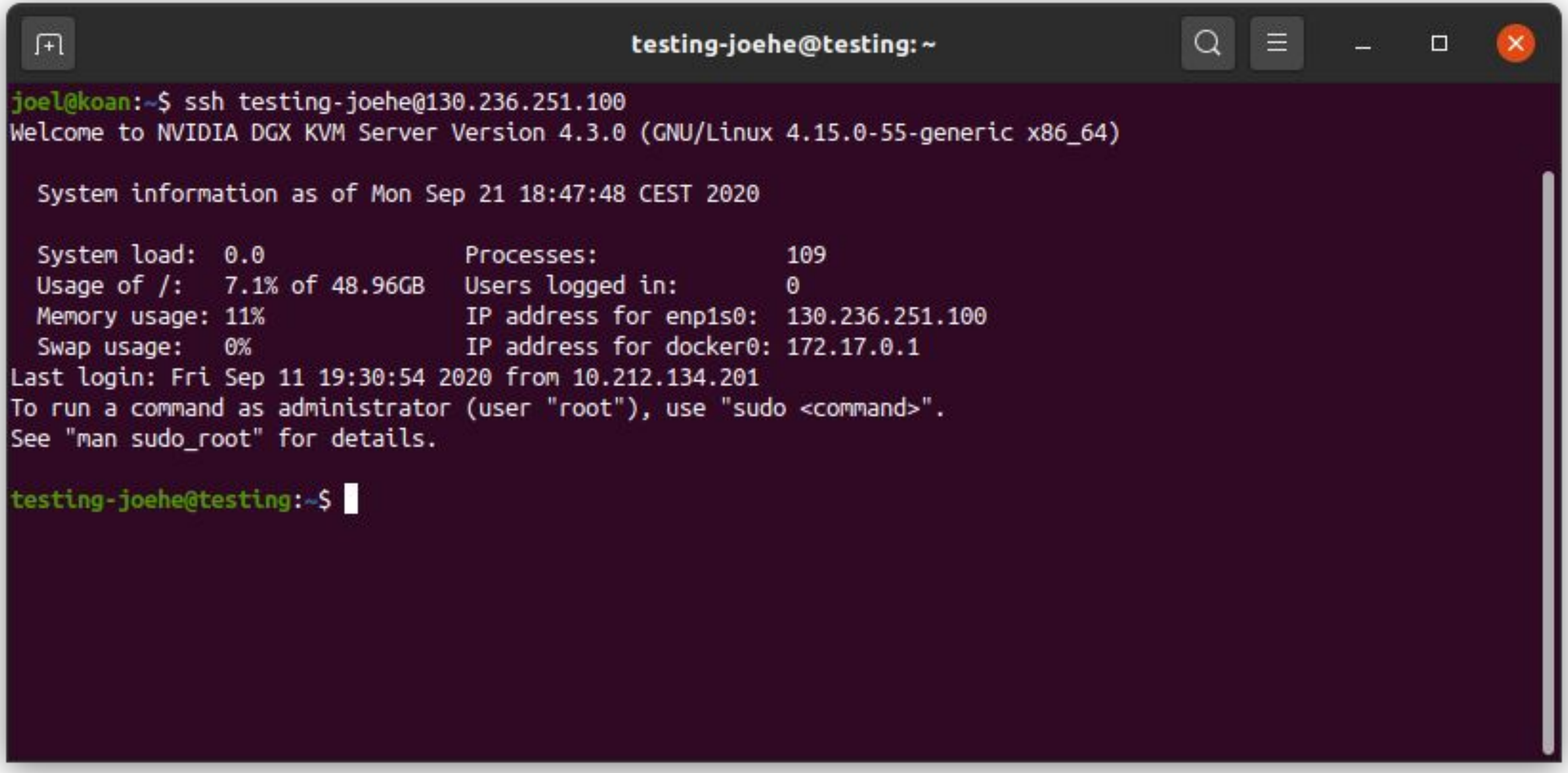
Project:	32 GPU weeks.
Fellowship:	16 GPU weeks.
Network partner:	8 GPU weeks.

- You have a storage quota (ask and you shall receive, if available)
- Put your name into an empty slot. Notice your used figure goes up. If it turns red, you booked too much; kindly remove some.
- Your booking ends Monday 09:00, and starts as soon as possible (<12:00).
- If you want to "Drop in" let me know! (Nb nobody ever wanted to "drop in")

How to use the DGX-2

- Get accounts.
- Log in to VPN with password and TOTP token.
- Log in to VM with SSH (pubkey).

Tada!



```
joel@koan:~$ ssh testing-joehe@130.236.251.100
Welcome to NVIDIA DGX KVM Server Version 4.3.0 (GNU/Linux 4.15.0-55-generic x86_64)

  System information as of Mon Sep 21 18:47:48 CEST 2020

  System load:  0.0               Processes:              109
  Usage of /:   7.1% of 48.96GB   Users logged in:        0
  Memory usage: 11%               IP address for enp1s0:  130.236.251.100
  Swap usage:   0%                IP address for docker0: 172.17.0.1
Last login: Fri Sep 11 19:30:54 2020 from 10.212.134.201
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

testing-joehe@testing:~$ 
```

Storage

- **/proj** - Very fast private persistent Project Storage, available through multi-10Gbit/s NFS mount on VMs, or through SFTP.
- **/raid** - Very ***very*** fast private local NVMe RAID array, available only on the VM. Non-persistent; data goes away when the VM goes away.

When reading data from /proj, VMs save a copy in /raid/cache/... . Next time the same data is read, it is read from the (faster) cache instead. This means that if you work exclusively in /proj you will get the benefits of both: persistency and speed.

You can use SFTP to transfer data to/from /proj before/during/after your booking, without affecting ongoing computations.

Storage

Safe. Power cord. Lightning bolt.

Storage

/proj FSC cache /raid

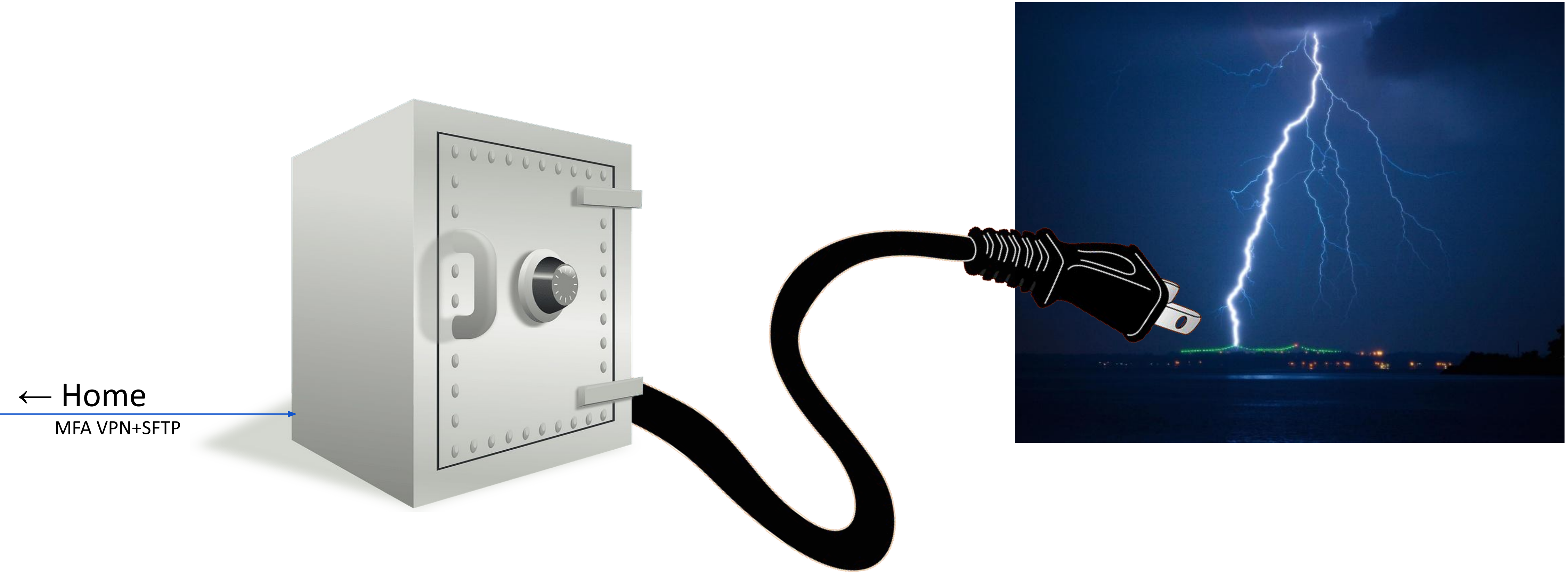
Storage

- **/proj** - Very fast private persistent Project Storage, available through multi-10Gbit/s NFS mount on VMs, or through SFTP.
- **/raid** - Very ***very*** fast private local NVMe RAID array, available only on the VM. Non-persistent; data goes away when the VM goes away.

When reading data from /proj, VMs save a copy in /raid/cache/... . Next time the same data is read, it is read from the (faster) cache instead. This means that if you work exclusively in /proj you will get the benefits of both: persistency and speed.

You can use SFTP to transfer data to/from /proj before/during/after your booking, without affecting ongoing computations.

Storage



← Home

MFA VPN+SFTP

/proj

FSC cache

/raid

Networking

- Outgoing connections are disallowed, except to approved destinations.
  - NTP and apt always allowed.
- Large-data-volume destinations can be approved based on feasibility: IP:PORT is easiest, while distributed services over cdn are harder.

- For small-data-volume traffic, we ask users to prefer using own VPN connection (see wiki for help).

This means:
Everyone follows their own policies, and AIDA does not have to implement them.

What should the next system be?

Data Sharing Policy

AIDA Data Sharing Policy v1.3.0?

Continuously updated as needs for guidance are identified.

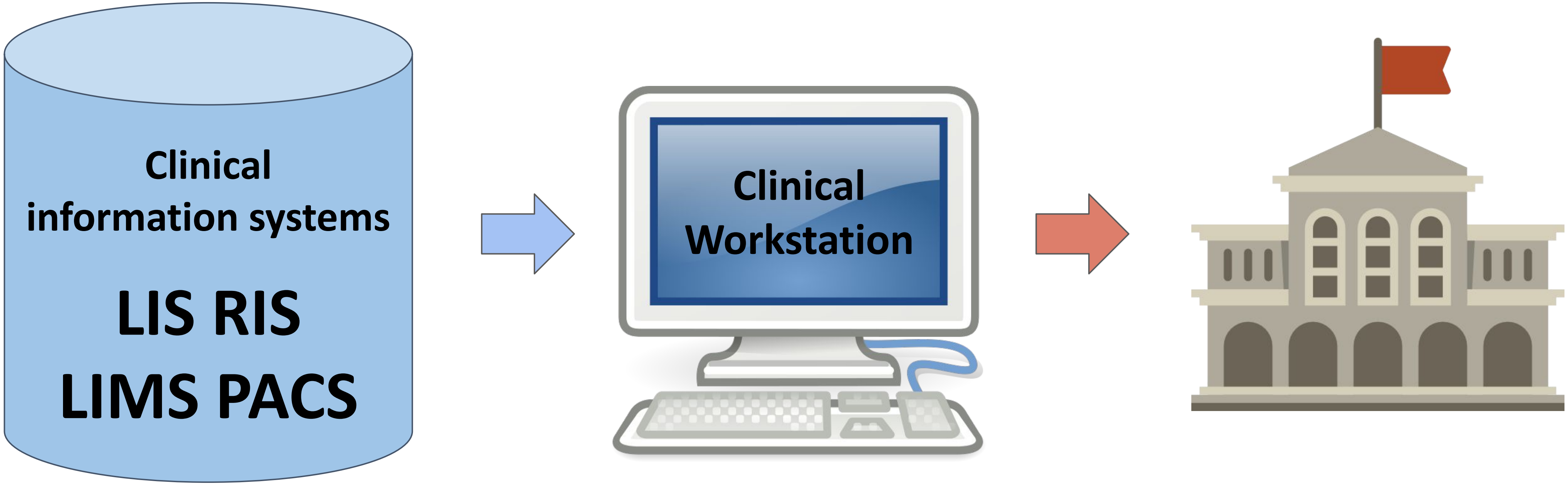
Add guidance for clinical data extraction and pseudonymization?

Add support sharing personal data?

Add guidance for
data extraction and anonymization?

How do I do this in practice???

Export and anonymize clinical data



Clinical information systems

LIS RIS

LIMS PACS

Clinical Workstation

1. Comply with policies
2. Define selection & parameters
3. Extract

4. Pseudonymize / anonymize
5. Verify results
6. Encrypt

7. Agree on terms of use
8. Transfer data
   using agreed method

Pseudonymization strategy

Ensure adequate pseudonymization/anonymization for each individual extraction.

Project <X> Pseudonymization Strategy				
Parameter	Source	Key	Data	Pseudonymize by
PIN	LIS	Yes	No	Delete.
Age	LIS	Yes	Yes	Stratify: 0-5, 6-10, 11-15, etc...
Date of request	PACS	Yes	Yes	Include year only. Must have >10 instances of diagnosis + anatomical site per year.
Diagnosis	LIS	No	Yes	-
...	...	...	...	...

Example terms of access

- Formal data request must include <all relevant information>.
- May only be used in ethically approved research.
- Same data may be disclosed to other research projects and purposes.
- Must have agreement in place to cover costs for work with data extraction.
- Must be processed in agreement with <all laws and regulations>.
- ...

Example request information

"Formal data request must include...", for example:

- Name of study (eg title of ethical review application)
- Ethics approval, registration number, attachments, ...
- Description of data, and parameters
  - Selection criteria: Time interval, examination type, sex, age interval, tissue type, ...
  - Parameters: Age, Diagnosis, Images, Resolution, ...
- Suggested pseudonymization / anonymization strategy
- Description of data sharing
  - "Data will be shared for research validation, and for further ethical and legal research."
- ...

Example modes of transfer

1. Through a FAIR Open Science data repository, such as AIDA Data Hub, for increased impact and citability, and to facilitate more and wider research.
2. Through a data transfer service. For example yours or the recipient's. There are existing research infrastructure services for this, such as provided by the AIDA Data Hub.
3. Send an encrypted hard disk. This is done less and less.

Tryggve

Nordic collaboration on Sensitive personal data for research

Joel Hedlund, Executive manager, Senior advisor

neic

NORDIC E-INFRASTRUCTURE COLLABORATION

Tryggve

Secure e-Infrastructure Services supporting Cross-Border Genomic and Register Studies

Slides: https://goo.gl/ru9b4y

Joel Hedlund
neic.no/tryggve Scientific manager

neic.no Nordic e-Infrastructure Collaboration
nbis.se National Bioinformatics Infrastructure Sweden
www.nsc.liu.se National Supercomputer Centre

NordForsk neic elixir DENMARK elixir FINLAND elixir NORWAY elixir SWEDEN

Nordic Register Genomics in Psychiatry - Overview of *Tryggve2*

Lu Yi, PhD. lu.yi@ki.se

Patrick Sullivan, Prof.

Psychiatric Genomics Institute,

Karolinska Institutet

Schizophrenia Basics

- Delusions & hallucinations, no known cause (minimum duration 6 months)
- Massive
  - **Morbidity**: top 10 in world
  - **Mortality**: life expectancy 10-15 years less
  - **Costs** (personal/familial/societal): $US 1.4M/life
- Intractable to extensive scientific study
- Subtle processes

Schizophrenia Genetics

A major clue, from generations of past work.

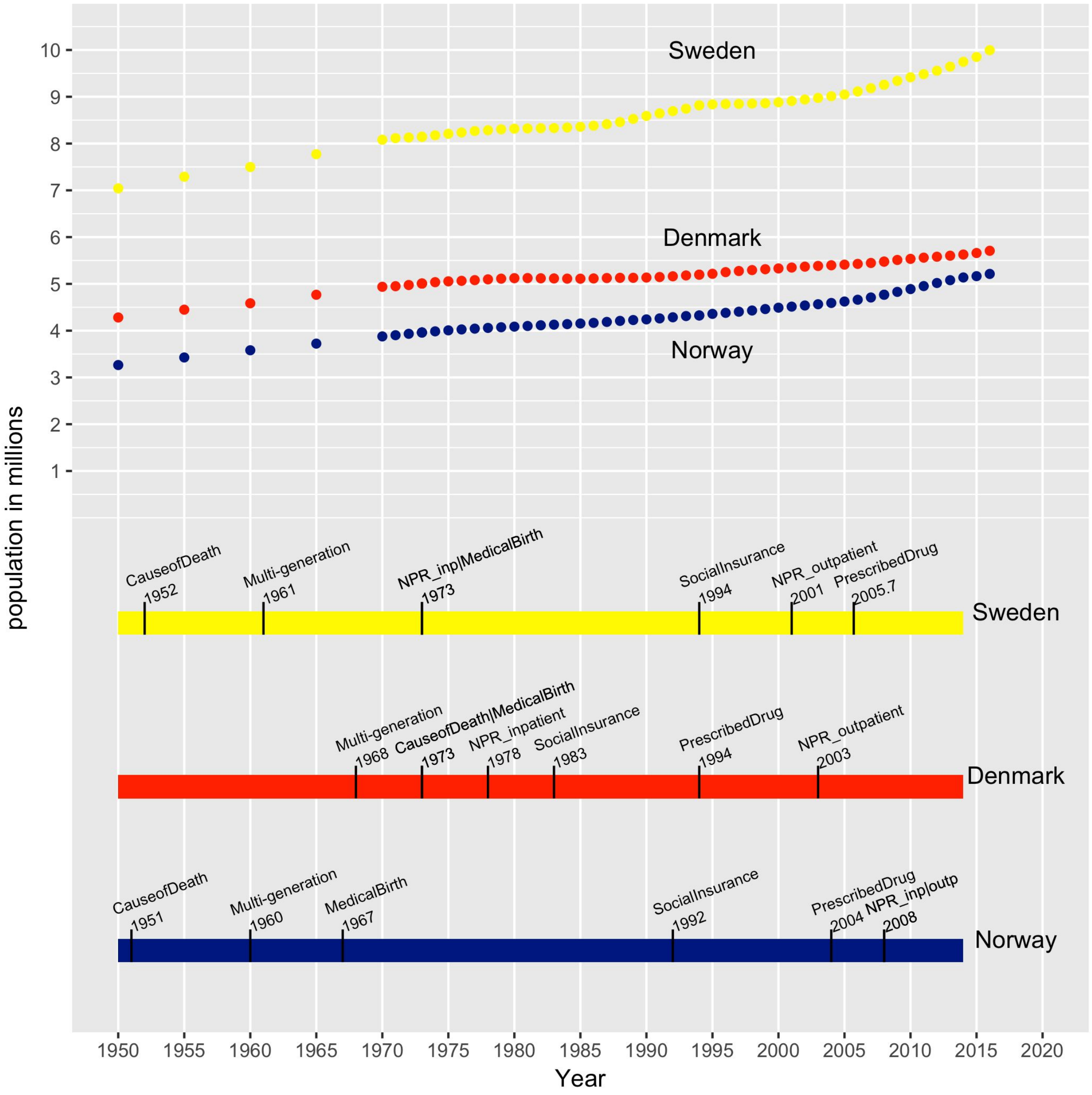
Probabilistic **not** deterministic:

- Family history, 10x increase (*but* 1% à 10%)
- MZ twins, risk to co-twin ~50%
- Heritability ~ 80%

No convincing single gene causes.

Nordic registers

- In-/out-patient register
- Prescription drug register
- Medical birth register
- Multi-generation register
- Social insurance register
- Cause of Deaths register



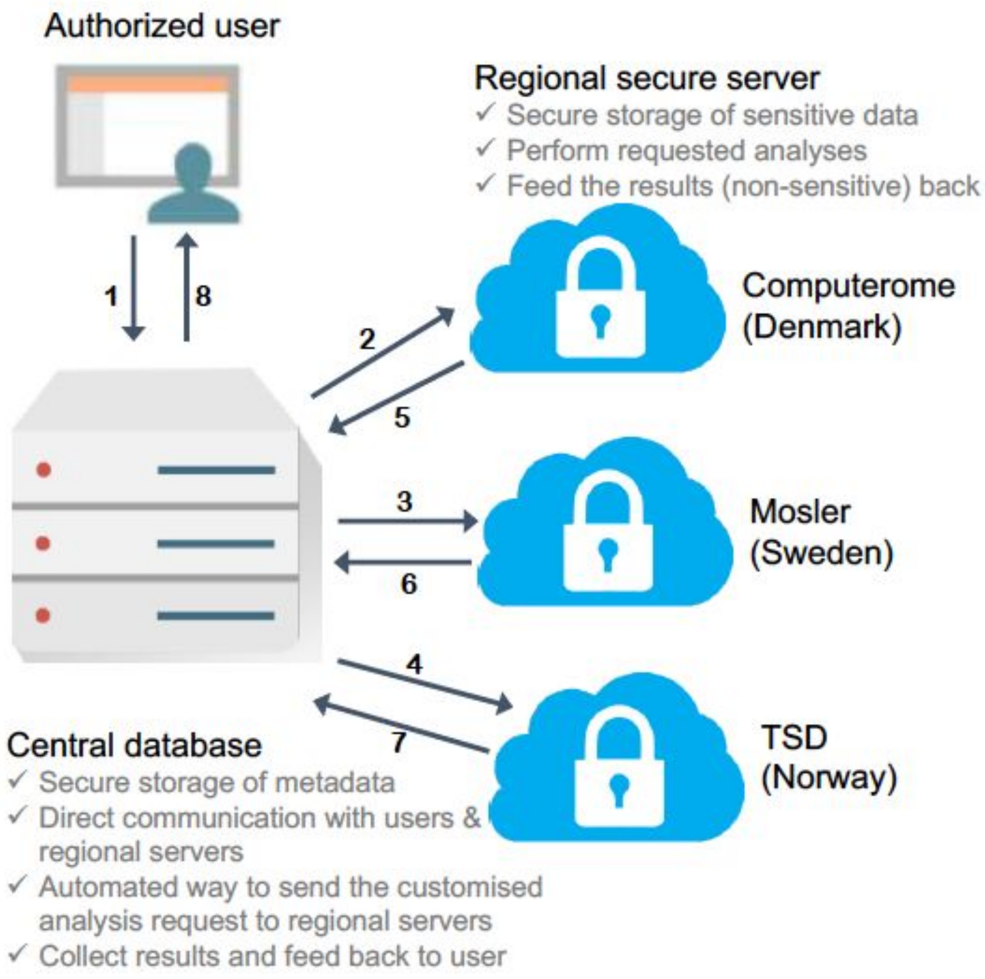
Study Ns

Descriptor	**Denmark**	**Norway**	**Sweden**	**TOTAL**
Vital statistics: Q4 2017				
– total population	5,781,190	5,295,619	10,120,242	21,197,051
– births	61,397	56,633	115,416	233,446
– foreign born (%)	0.085	0.141	0.185	0.137
Register analyses		MoBa		
– lifetime Schizophrenia (SCZ)	36,676	9,002	29,072	74,750
– lifetime Major depression (MD)	75,771	87,540	595,743	683,283
– lifetime Postpartum depression (PPD)	50,176	8,572	93,960	152,708
– lifetime Eating disorders in females (ED)	21,816	4,857	34,238	60,911
Microarray data: Q2 2018				
– Ns with GWAS	89,273	2,850	183,966	276,089
– SCZ cases	5,247	800	4,924	10,971
– MD cases	25,431	0	5,059	30,490
– PPD cases	1,600	0	1,381	2,981
– ED cases	5,114	0	4,118	9,232
Microarray data: Q4 2021				
– Ns with GWAS	425,000	386,000	300,000	1,111,000
– SCZ cases	9,622	2,240	12,000	23,862
– MD cases	45,701	11,750	10,000	67,451
– PPD cases	3,600	1,000	2,881	7,481

Tryggve2

A federated system that enables data sharing and analysis in a ***secure, streamlined &*** ***intelligent*** *way*

#2-7
Distributed compute solution via singularity container



Regional secure server
- ✓ Secure storage of sensitive data
- ✓ Perform requested analyses
- ✓ Feed the results (non-sensitive) back

Central database
- ✓ Secure storage of metadata
- ✓ Direct communication with users & regional servers
- ✓ Automated way to send the customised analysis request to regional servers
- ✓ Collect results and feed back to user

Note: ePouta (Finland) is another secure server. However, Finnish data will not be included in this project, thus not shown in the figure.

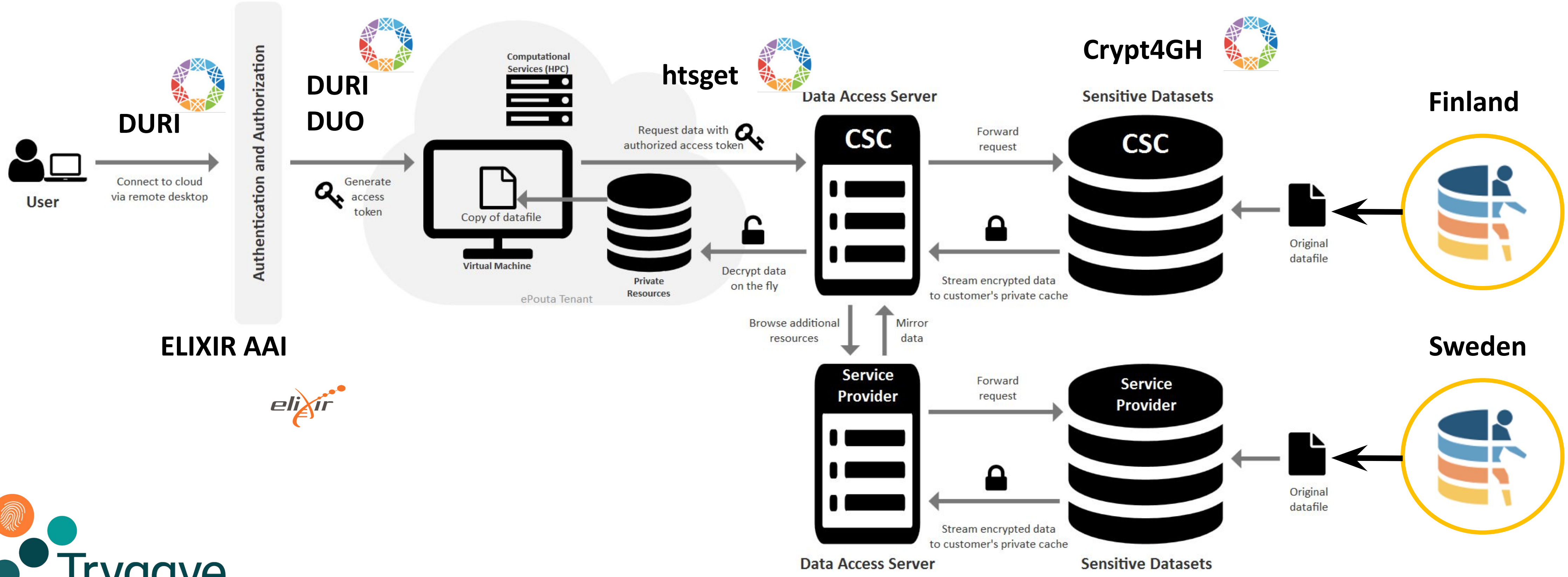
Slides: goo.gl/ru9b4y

Website & Contact: neic.no/tryggve/

GDPR-compliant tech stack

5. Services Federated EGA, Beacon, SD-Desktop...	Secure and private access to data and services.
4. Access requests REMS	Transparent electronic handling of ethical/legal processes for data access.
3. Grants REMS	Certifies what data resources you can legally access.
2. Membership Life Science AAI Perun	Certifies what research groups you belong to.
1. Identity Life Science AAI	Guarantees that you are you.

Federated EGA technical solution and standards compliance



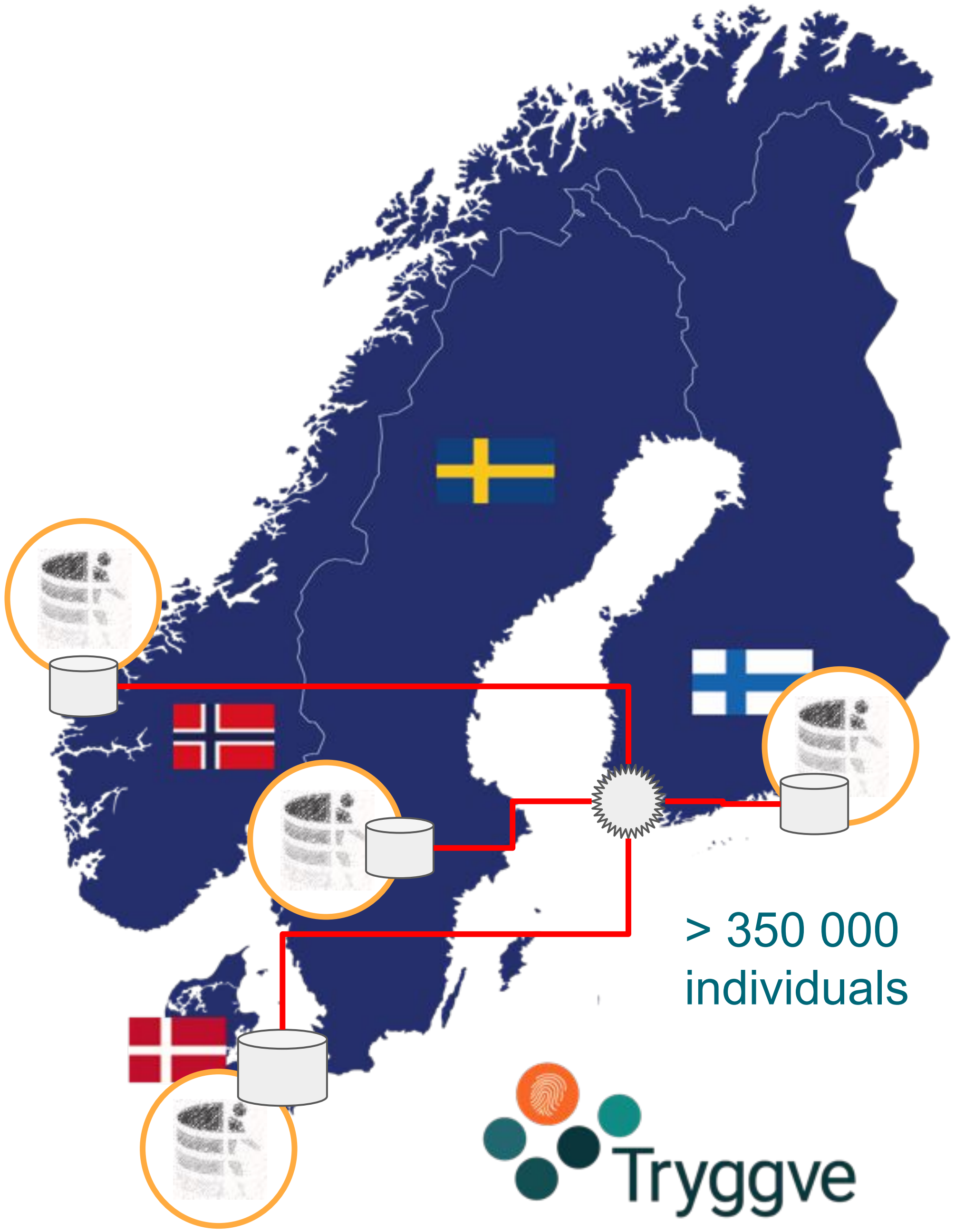
Tryggve

Nordic Twin Study on Cancer

- Twin research on heritable and familial risk in prostate, breast, ovarian and colon cancers
- Cohort constructed by linking the population-based twin registries of Denmark, Finland, Norway and Sweden to their country-specific national cancer and cause-of-death registries. Genomic data also collected from the samples.
- A shared sensitive data processing environment required for method development and data harmonization
- Tryggve use case in progress

NorTwinCan
Nordic Twin Study of Cancer

http://nortwincan.org



> 350 000 individuals

Tryggve