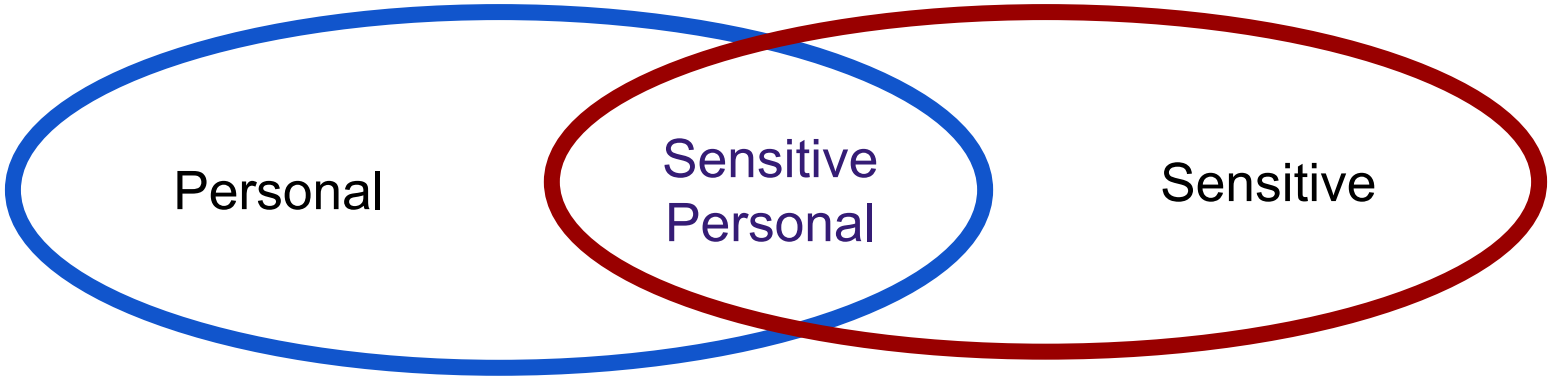
Bianca:
legal and administrative aspects of sensitive data

Marcus Lundberg
Lars Eklund

UPPSALA
UNIVERSITET

SND
Svensk nationell datatjänst

What is sensitive and/or personal data?



Name
Food preference
Width of left nostril

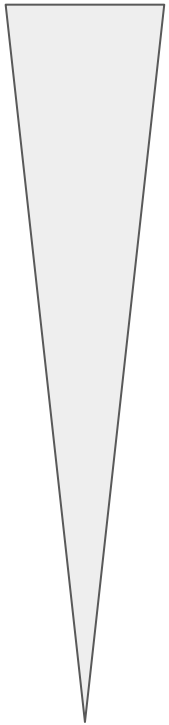
Ethnic origin
Political opinion
Religious beliefs
Trade union membership
Health
Sex life
Genetic data
Biometric data

IP from private industry

Sensitive environmental data, e.g. protected species

National security

Where to turn for guidance



General advice on sensitive data questions: SND:s domain specialists (us)

Ethical questions: Trusted Persons at your faculty, Etikprövningsmyndigheten (Swedish Ethical Review Authority)

Legal questions: contact your Data Protection Officer, legal department, and/or security department

Technical and practical questions on and around Bianca/UPPMAX: UPPMAX support

Technical and operational security

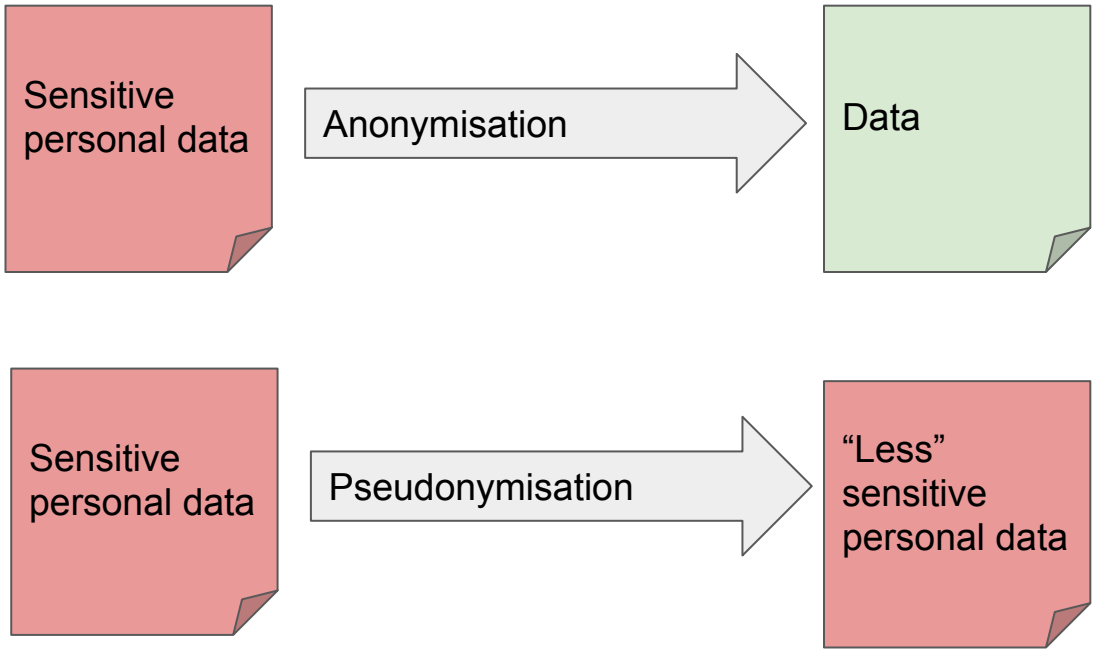
"Lätt att göra rätt"

Risk = harm x probability

Technical security measures	**Operational security measures**
Strong authentication	Classify all your data
Encrypted storage and transfer	Anonymise or pseudonymise
Limited physical access	Routines for access privileges

What is the weakest link?

Anonymisation and pseudonymisation



Sensitive sequence data

Simple but hard: whole genomes of living individuals = non-anonymisable sensitive personal data

More complex:

- Protein sequences?
- RNAseq?
- Transcriptomics?
- SNPs?
- Cancer genomes?
- Gut microbiome?

Sensitive sequence data

Challenges include:

- Sheer size of data, often requiring high technical sophistication
- Complex context with many actors, each with scientific, technical, legal considerations

Bianca (now funded by NAISS) a good analysis platform.

FEGA (https://fega.nbis.se) a good repository

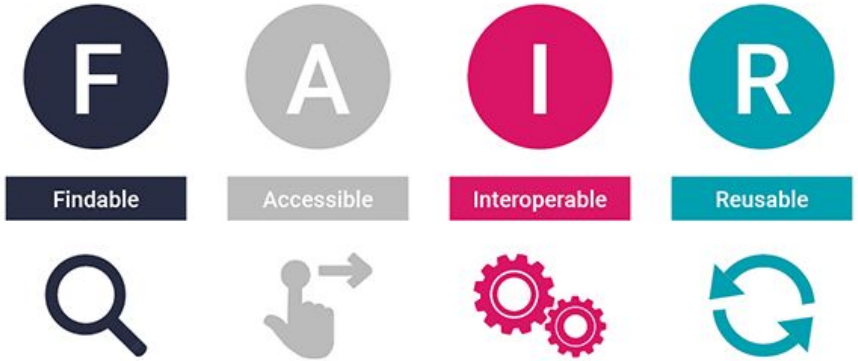
SND DORIS publishes e.g. SWEGEN

FAIR and sensitive data

Not openly available BUT:

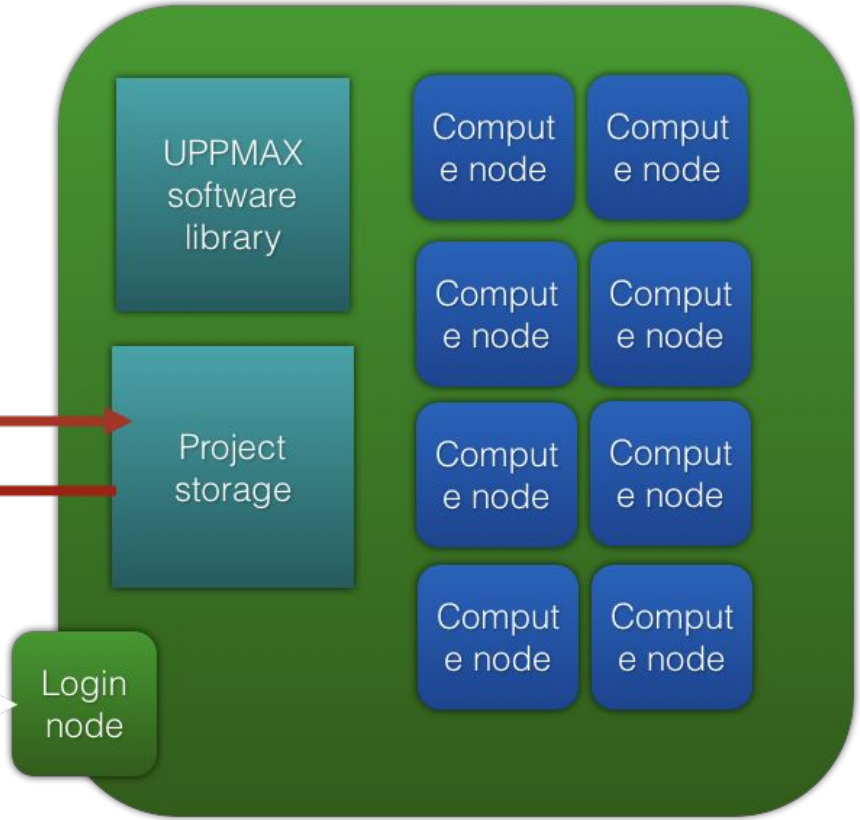
- Open metadata — publish with clear access conditions

- Can (usually) share for valid purpose, after consideration by data controller



Bianca: A locked-down cluster

- Data transfers logged
- No internet access
- Usable
- **Full** software stack
- Performant



Security mechanisms on Bianca, at UPPMAX

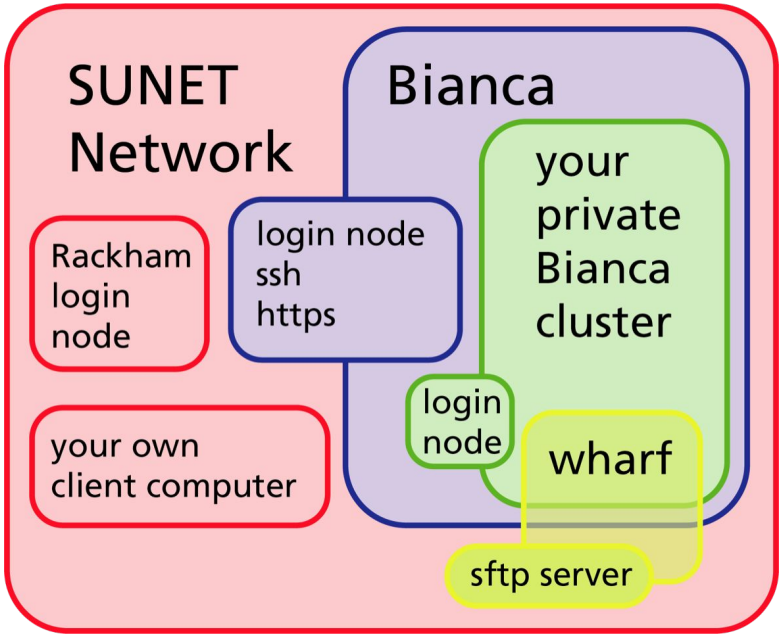
Logical separation of user projects

Data exports limited to a logged “wharf”

Principal investigator responsible for project membership == access privilege

Locked racks, strict routines for hall access, encrypted backups

…



Other sensitive data on Bianca

Health registries

Medical journals

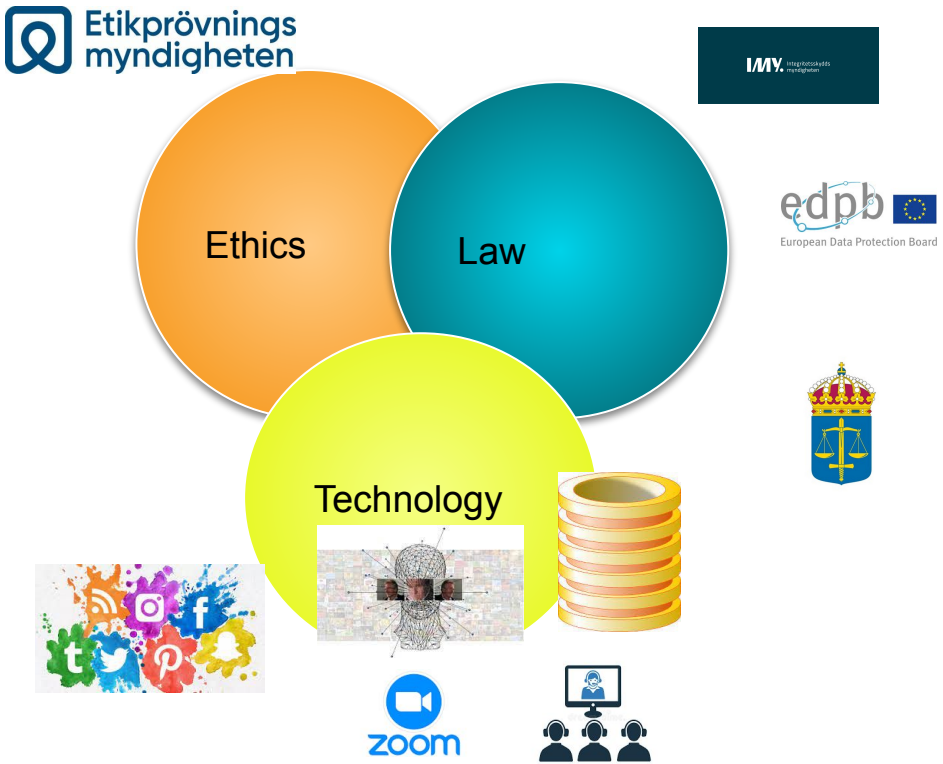
Interviews

Spatial data from cellphones

Guiding questions:

Is it personal data?

What is the harm/risk?



Reflection exercise (10 min)

5 min: Consider your sensitive research data and write down the following:

1. How you store and handle data outside of Bianca
    a. Before analysis on Bianca
    b. After analysis on Bianca
2. How you treat data inside Bianca (wharf, /proj, /proj/nobackup, home dir, etc)
3. How you handle data transfers to/from Bianca

5 min: Go through your list and reflect on which steps are the weakest in terms of information security

Quiz (10 min)

Mentimeter: 6606 7026

Applying for a project

1. Get a Data Processing Agreement (if applicable)
2. Estimate your needs:
    a. How much data (e.g. samples x size per sample x analysis for expansion)
    b. How much compute (e.g. 1 kch/month per TB)
3. Create proposal in SUPR:
    a. < 20 TB: NAISS SENS Small
    b. > 20 TB: NAISS SENS Medium
4. Describe your needs well
    a. State that you have sensitive data
    b. State that you have a Data Processing Agreement (if applicable)
5. Most proposals handled during last week of month.

Proposal evaluation

- **Motivation** (detailed itemisation of the data that covers the entire allocation of the project, and reasoning that motivates why described data needs to be present concurrently)
- Track record (previous history in SUPR)
- Scientific productivity (from Activity Reports, if applicable)
- FAIR
- Data formats, e.g. proper compression
- Use of nobackup (describe what needs backup etc)
- Use of central reference databases in /sw/data (https://www.uppmax.uu.se/resources/databases/)
- Time plan
- Plan for data after analysis, including a destination for it
- Whether you follow the recommendation of keeping your own copy of the data elsewhere

Project management

Proxy — a project manager who isn’t the PI

Project extension — email Support a few weeks before expiration

One project or many? — Let data, ethical consent, and practicality rule

I need more core-hours and terabytes! — Okay, but first make sure you actually do

The need for contextual metadata?

Do we need to be able to describe:

How the data may be sensitive?

Or

How the data may be open?

It’s not black and white so what defines the level of sensitivity?

Where are the Risks?

SND
Svensk nationell datatjänst