UiO : Universitetet i Oslo

Research on human data ? How do I handle it ?

Gard Thomassen, PhD
Dept Head Research Computing Services
University Center for Information Technology (USIT)
University of Oslo

Outline

- Who are you, what are you doing ?
- (Sensitive) Data, laws and regulations
- TSD setup, solutions, demo, status and future
- Data handler agreement, what is this ?
- How to get on board

Privacy 1

Personopplysninger" (person information)= data that can be connected on one single person.

Personopplysningsloven § 2 1)

I. Direct identification:
   - data like name and social security

II. Indirect identification :
   - data that by combinations of information still makes you point at one single person.

Privacy 2

I. Psudo-anonymized data (de-identified)
   - data that can not be mapped to a single person unless a key is used, key must be stored apart from data and well guarded

II. Anonymized data:
   - data that by no means can be connected to a single person : THIS IS NOT PERSON DATA

Bildet er lånt fra:
http://www.gokaker.com/kaker/sjokoladekaker/slides/PICT2752.htmlxx

What is sensitive data?

Norway : Personal Data Act §2, point 8

- race/ethnic data, political opinion, philosophical and religious beliefs, the fact that a person has been suspected of, charged with, indicted for or convicted a criminal act, **health**, sex life and trade-union membership

Who has sensitive data

Almost everyone

What does privacy mean

- It is all about letting people be in charge of their own

Lånt fra www.datatilsynet.no

- Each person governs their own private data

Personvern ved UiO – hvilke regler setter rammer for våre handlinger?

- Personopplysningsloven med forskrift
- Helseforskningsloven (hfl)
  - (§ 2 3 ledd: I den utstrekning ikke annet følger av denne loven, gjelder personopplysningsloven med forskrifter som utfyllende bestemmelser.)
- Forvaltningslovens taushetsbestemmelse og ansettelsesavtale
- Interne regler og rutiner:
  - IT-reglementet
  - IT-sikkerhetshåndboken

EU regulations

- USA found not to be trusted as of 12/10-15
  - Safe harbour agreement is found invalid
- New EU/EEC regulations decided upon approx 1/1-16, will take effect in after 2 years

- And this is even tighter regulations that in Norway as of today

EU/EEC What is sensitive data

- personal data, revealing race or ethnic origin, political opinions, religion or beliefs, trade-union membership, as well as genetic data or data concerning health or sex life or criminal convictions or related security measures
- Health data: any information which relates to the physical or mental health of an individual, or to the provision of health services to the individual;
- Genetic data: all data, of whatever type, concerning the characteristics of an individual which are inherited or acquired during early prenatal development

EU /EEC Research on sensitive data

- With the consent of the data subject
- Processing is necessary for research purposes; under the conditions of Article 83 of the Regulation
- Research is defined as fundamental research, applied research, and privately funded research research taking into the Union's objective under Article 179(1) of the Treaty on the functioning of the European Union of achieving a European Research Area
- NB : High public interest projects... in addition

Doing (medical) research

- You need an “official” go-ahead for your project before starting
  - REK (Regional Committee for Medical and Health Research Ethics)
  - National Center Sociel Science Data Centre(NSD)
  - Or someone else like Norwegian Data Protection Authority
- Personal approval from humans involved (Samtykke”

Where to store your data ?

- UiO and rr-research networks are off course okay with anonymized data
- UiO and rr-research networks have been used extensively for psedo-anonymized data
  - Not a good practise
  - UiO aims to get this data into more secure systems
  - Key must not be stored in these networks
- UiO and rr-research networks are not okay for sensitive person information

Where to store your sensitive data

- UiO -> In TSD (Services for Sensitive Data)
- OUS -> Clinical system

- Cloud : Doable within EU/EEC, but lots and lots of paperwork, risk assessments, data-handler agreements etc etc.

UiO : Universitetet i Oslo

TSD

Pilot 2009 - 2012

TSD launch in Computerworld 16/5-14

It-helse

MR er veldig bra for å studere hjernen på personer fra fire fem års alderen. Vi må bruke mye regnekraft. Det kreves 60 timer per deltager.
ANDERS M FJELL, UIO

FOTO: AHLERT HYSING

SEKVENSERER: Kreftforskningen trenger avansert utstyr for DNA-sekvensering av vevsprøver. Professor Eivind Hovig med Hiseq 2500 som gjør sekvenseringen som generer så mye data at Colossosus er helt nødvendig. Over fem år trengs 800 TB med data.

Ny giv for livgivende forskning

Verdens første sikre tjeneste for sensitive data kan gi norske forskere muligheter som andre bare kan drømme om. Forskerne stod i kø for å fortelle om utfordringer de kan løse.

AHLERT HYSING

> De nye tjenestene og infrastrukturen de bygger på vil gi en høy grad av be-

Arkivering

Artikkelen startet på forrige side

Store behov

> Vi har veldig sensitive data med innebygget personvern som må tas vare på.
> Sissel Jor, OUS

Kontraster

Mat og helse

Colossus II

Colossus II er datamaskineriet for bearbeiding av sensitive data. Colossus var den tidligere eksperimentmaskinen basert på eldre teknologi.

Verdi:
3,5 millioner kroner

Beregning:
72 Megware prosesseringsnoder hver med to Intel Xeon E5 med 10 prosessorkjerner med 2,5 GHz klokke

Minne:
8 G trans/sek til 64 GB minne per node
Sammenkobling: 4 Mellanox svitsjer med 36 porter for Infiniband på 56 Gb/s

Ytelse:
25 TFLOPS

Minneberegning:
To Megware hver med 4 Intel Xeon E5 med 8 prosessorkjerner med 2,7 GHz klokke og 1 TB minne for gensekvensiering

Lagring:
Kun for bearbeidingen, 4 I/O-noder hver med 64 GB minne og 45 TB, 1 Metadatanode med SSD

Permanent lagring:
Anskaffes til hvert prosjekt, vil bli mange PB over tid

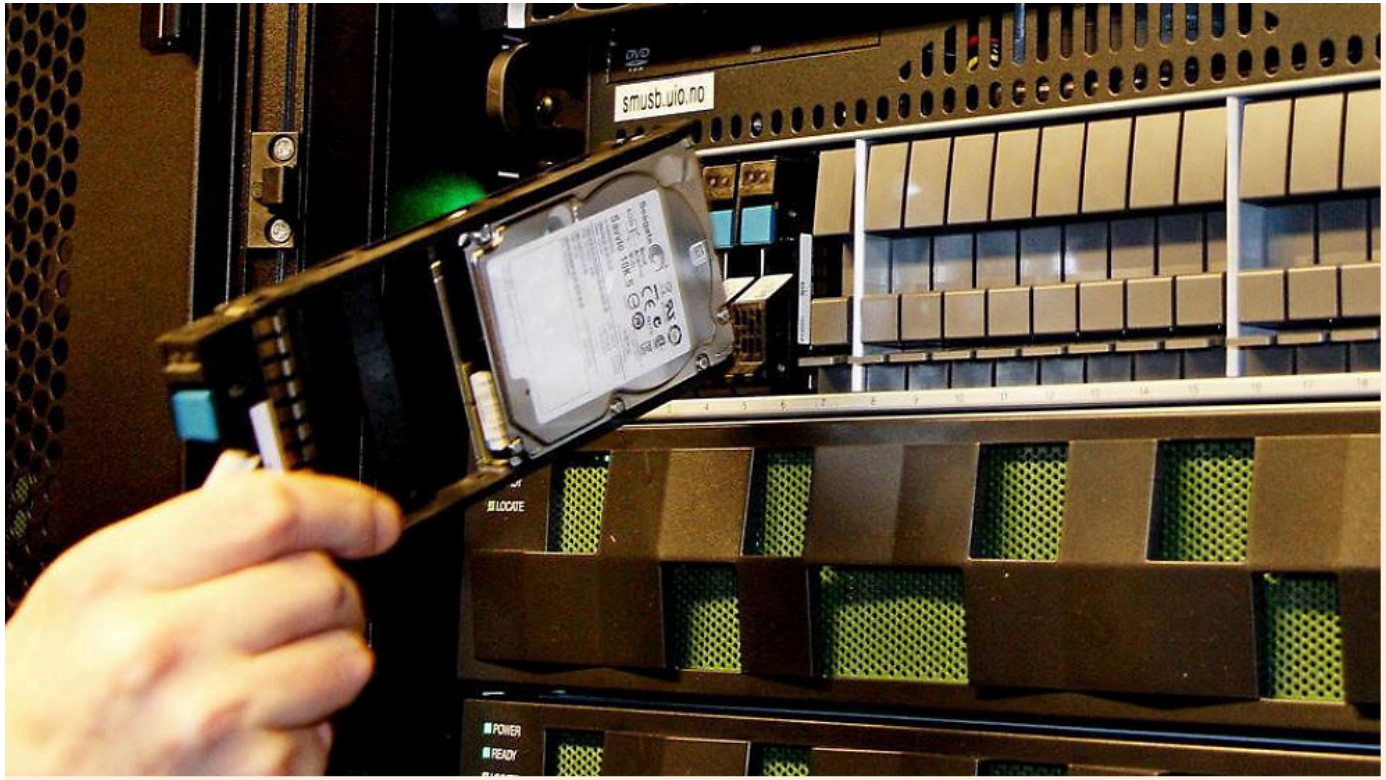
SIKRET: Forskningsgruppeleder Hans Eide og it-direktør Lars Oftedal har vært prosjektleiere for TSD. Mellom tive data. Maskin og lagring er sikret både med tilgang og fysisk. Dataene er anonymiserte.

Storskala dataanalyse

Startklar

> Sammenhengen mellom mat og helse er en komplisert problemstilling.

Teknisk ukeblad & e24, 5/5-14



SIKKERHET: En egen tjeneste ved Universitetet i Oslo skal håndtere store mengder forskningsdata som ikke skal komme på avveie. **FOTO:** ESPEN ZACHARIASSEN *TEKNISK UKEBLAD*

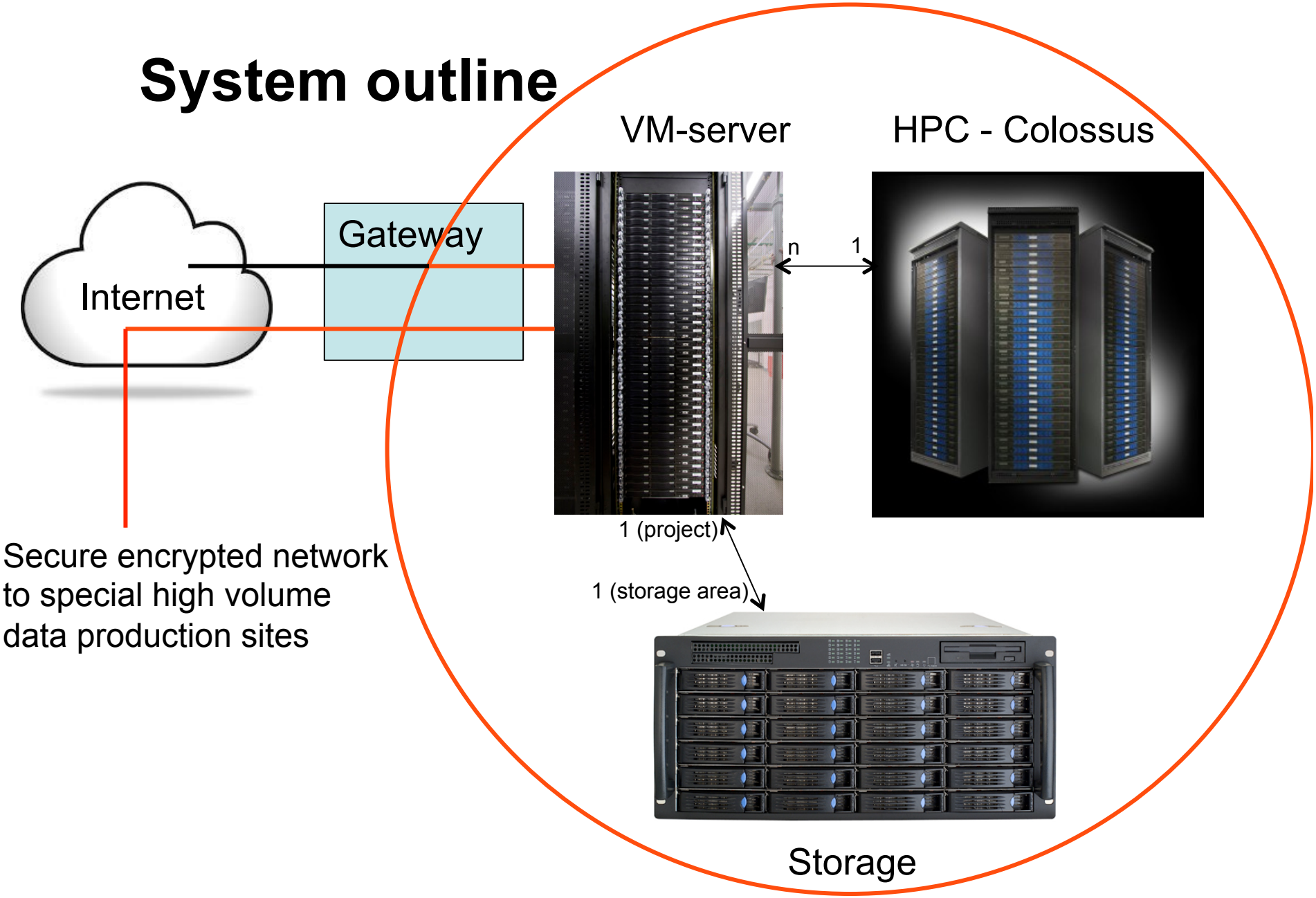
Norge får Nordens beste anlegg for hemmelige data

System requirements

- Security, isolation and access control as given by law
- Large storage capacity
- Multi tenant (multiple users)
- High performance computing (HPC) resource
- High bandwidth
- Easy to maintain and operate
- Easy to use and “practical” (also for audio and video)
- Some freedom within confined user space
- Accessible from *anywhere* through proper mechanisms
- A variety of software and public data-sources must be available
- Windows and Linux support (server/host-side)
- Data collection services
- Data sharing services

Setup, solutions and status

System outline



Secure encrypted network
to special high volume
data production sites

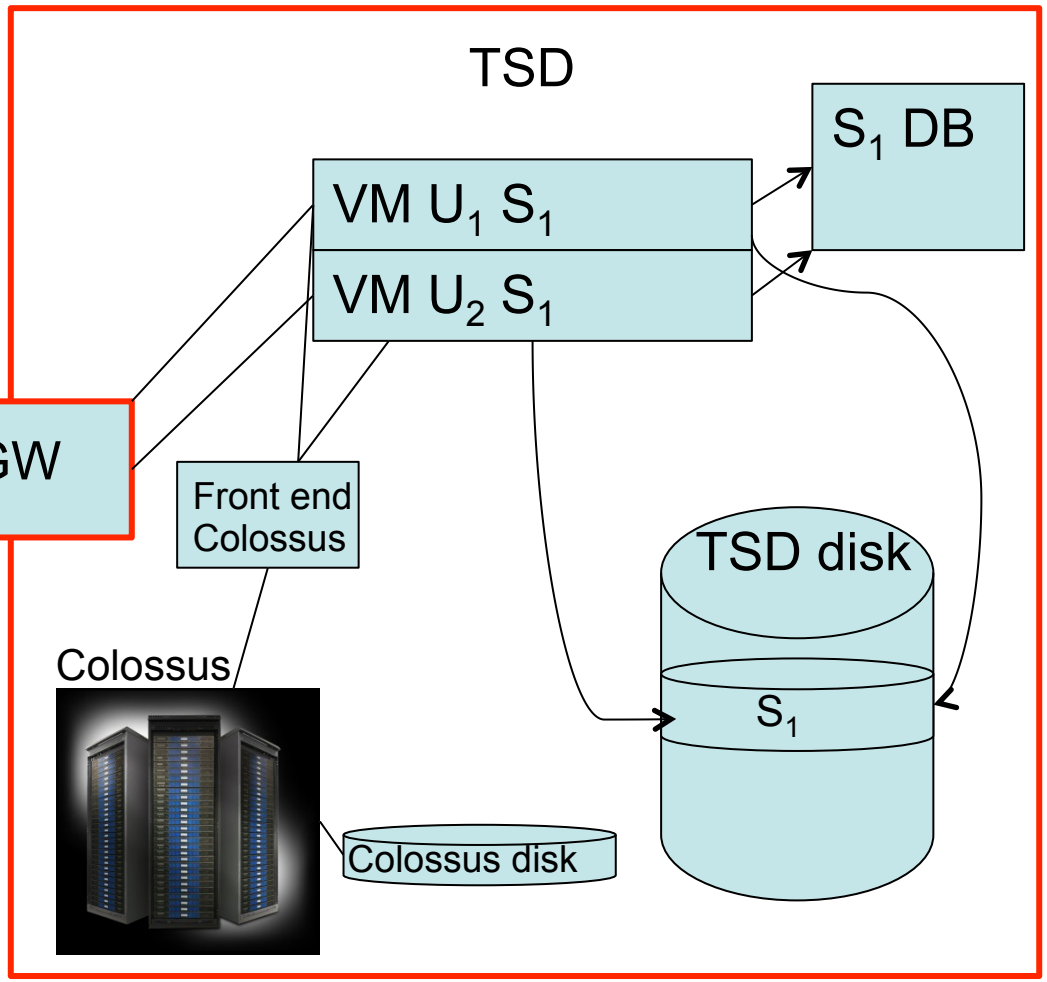
Using TSD

Libre Office
R
Module load
...

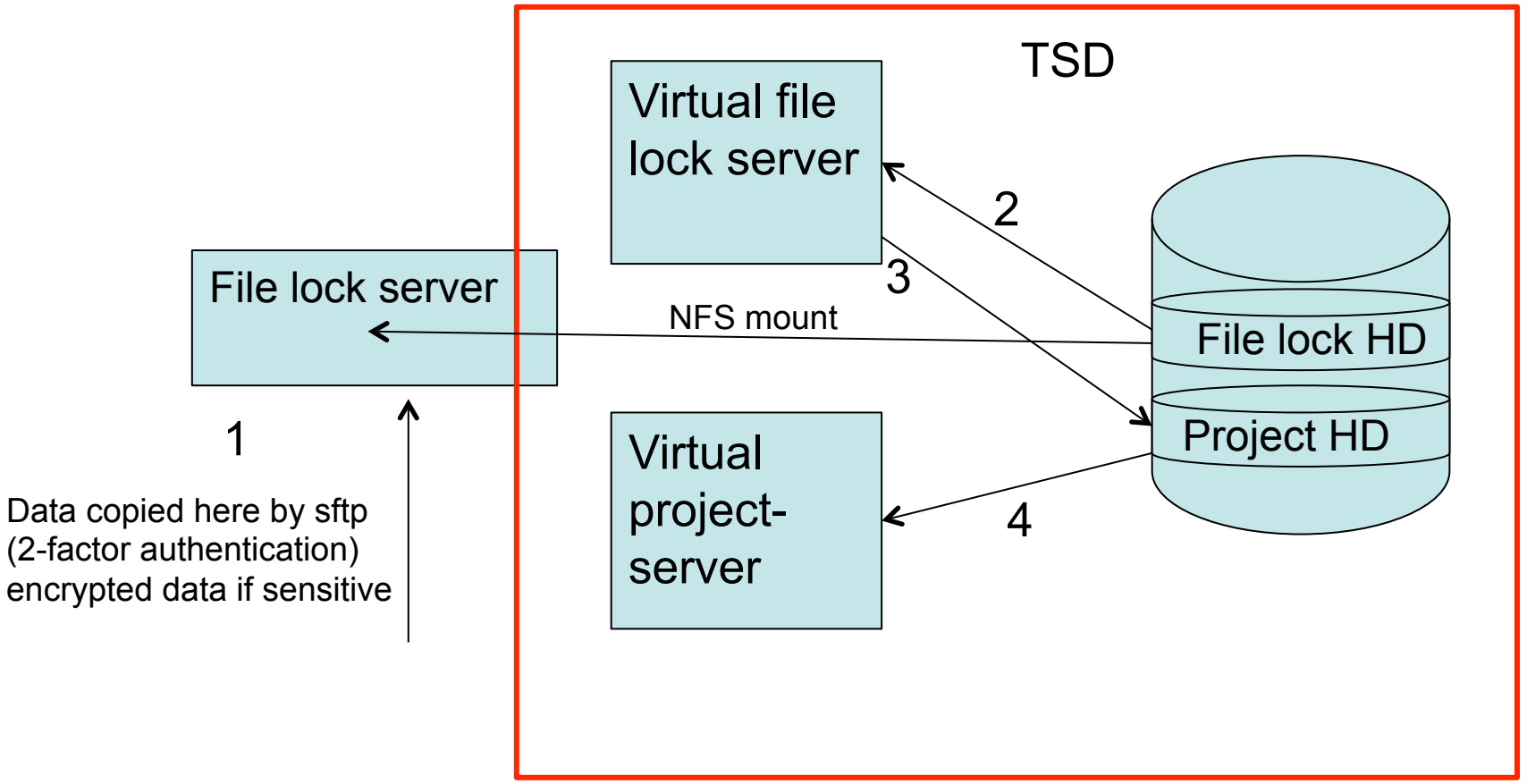
$User_1$ $Study_1$

$User_2$ $Study_1$

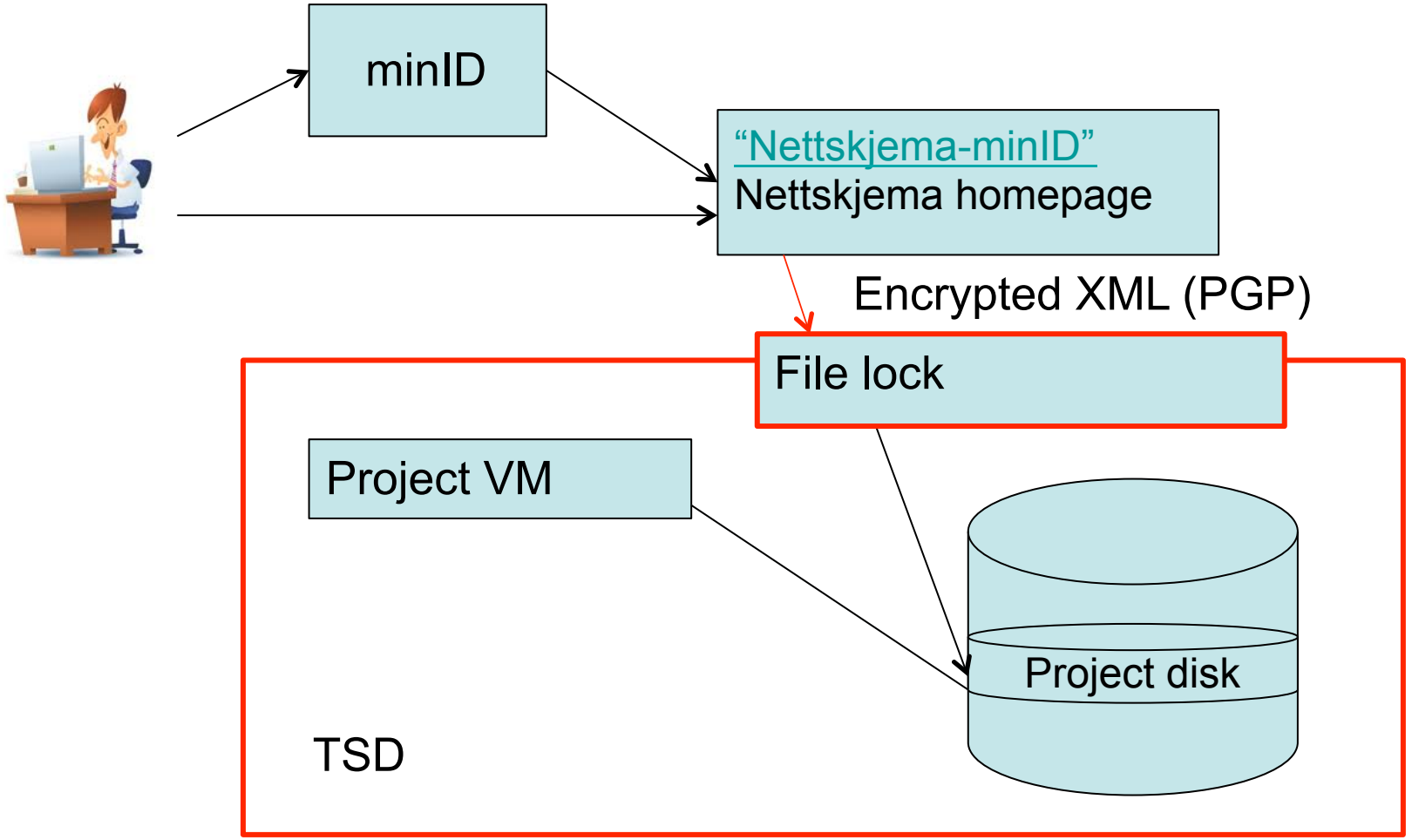
SPSS
Office
Stata
SAS
R
Matlab
....



Data import and export using TSD



Data collection using TSD



Security details

- OATH TOTP 2-factor authentication
  - Smart phones or programmable hardware tokens
- Import/export is under strict control
- No open connection to the internet
- All administration happens from the inside
- Strong separation between projects
- Hardened FreeBSD gateway and firewall
- Encrypted backup, one key per project
- Sys-admins are single users (traceability)
- Sys-admins have to use same authentication process

Homepage

http://www.uio.no/tjenester/it/forskning/sensitiv/

Projects

http://www.uio.no/tjenester/it/forskning/sensitiv/mer-om/kunder/

Demonstration

Login
Nettskjema

TSD status

- > 100 research projects
- > 350 users
- Secure storage (> 1 PiB on disk)
- Secure data analysis
- Linux or windows hosts (> 250 VMs)
- Secure import and export
- Web-based data harvesting
- HPC cluster (>1500 cores)
- Postgres DBs
- Video and sound display

Capabilities enabled by TSD

- Large scale NGS research on human genomes
- Large scale medical imaging studies
- Large scale studies with web-based data collection
- Off-site analysis of sensitive data
- Secure storage for verification of published research
- Electronic consent

Future of TSD - main topics

- Better user interfaces
- More software support
- Better and faster infrastructure
- National eInfrastructure investement in TSD

HPC resource – Colossus

- At present about 1500 cores (~30 TFLOPs)
- No project users are to log in on any nodes
- One global job daemon to control data integrity (to ensure project data separation)
- $SCRATCH exists on a per project pr job basis
- As similar to Abel (the non-sensitive HPC resource in Oslo) as possible
- Separate parallel file-system
- Huge-mem nodes and Infiniband interconnect

Data handler agreement

- Data handling responsible : The institution that is responsible for the research project, has the REK approval etc
- Data handler : Another institution, hired by the abovce institution, that is to host or in some other way handle and treat the data
- The data handling responsible institution is also responsible to get an agreement in place between the institutions so that the other institution guarantees data safety.

How to get on board

tsd-contact@usit.uio.no

tsd-drift@usit.uio.no

Main collaborators on TSD

Collaborators

- Norwegian Storage Infrastructure (NorStore)
- Norwegian Genetics Analysis Platform (GenAp)
- Norwegian Dietary Registry (Medical Faculty)
- Institute of Psychology (Faculty of Social Sciences)
- Norwegian Cancer Sequencing Consortium (NCGC)

Reference group

Oslo University Hospital, NorStore, Regional Ethical Committee, National Institute of Public Health, Norwegian Cancer Registry, Research Network at OUS, Elixir Norway, NCGC, GenAP, Institute of Psychology.

Thanks to

Project group / developers

- tsd-core@usit
- virt-core@usit
- storage-core@usit
- postgres-core@usit
- network-core@usit
- hpc-core@usit
- windows-core@usit
- unix-core@usit
- IT-security@usit

Administration / associated

- IT-dir Lars Oftedal
- Hans A. Eide
- Märtha Felton

norstore

genAP norwegian genetic analysis platform

UiO : Department of Psychology
Faculty of Social Sciences

UiO : Institute of Basic Medical Sciences
Faculty of Medicine

Norwegian Cancer Genomics Consortium
A national health service collaboration to establish and evaluate genome-based diagnostics for cancer therapy decisions

Nordic collaboration opportunities

- Laws are fairly similar (Norway very strict)
- Difficult to exchange sensitive data for research
- One should learn from each other as these systems demands very special IT-knowledge
- Services development and system-administration know-how is non-sensitive and may be shared
- Building TSD addressed many novel security questions in a University setting to be learnt from
- Large DBs/registeries of health data may enable very interesting research in the future
- TSD is involved in the NeIC-based Tryggve project
- We are happy to collaborate!