



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. Pseudo-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 Social 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.

TSD

Pilot 2009 - 2012

TSD launch in Computerworld 16/5-14

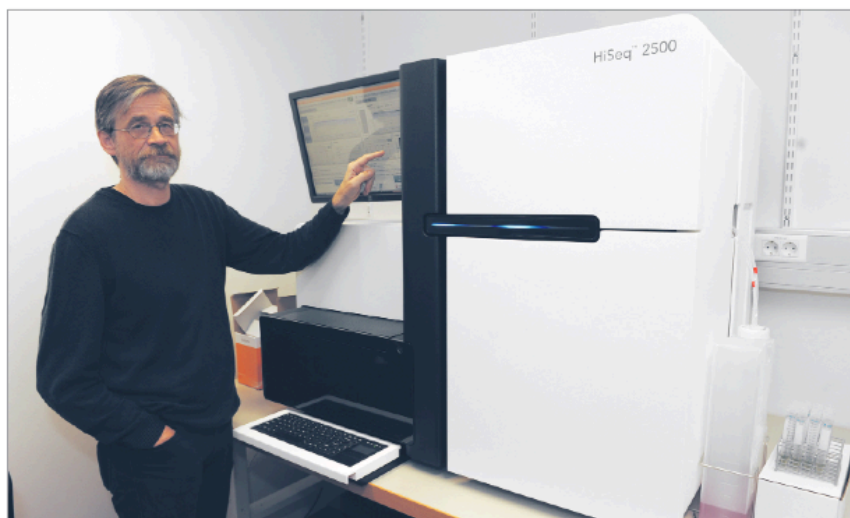


It-helse

COMPUTERWORLD NORGE • NR. 20 • FREDAG 16. MAI 2014

SIDE 29

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



SEKVENSERER: Kreforskningen trenger avansert utstyr for DNA-sekvensering av vevsprøver. Professor Eivind Hovig med HiSeq 2500 som gjør sekvenseringen som generer så mye data at Colossus 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.

er det tvisomt om noen andre land kan vise maken til. For siden 2009 har Universitetet i Oslo jobbet med lagringsutfordringer og sett behovet for å kunne forske på sensitive data.

— Universitetet i Oslo har veldig fokus på det brukerne vil ha. Med TSD tilbyr vi det forskerne har behov for, hvor vi tar hensyn til Helsepersonelloven og Personvernloven. Vi har fått mange henvendelser om forskning som sikrer personvernet og har aldri hatt så mange avtaler før vi er i gang, sier Lars Oftedal, IT-direktør ved USIT.

derfor oppstatt av at det må lages en ny modell for infrastruktur som også omfatter tjenester for lagring som kan videreføres selv om infrastrukturen forsvinner.

— Vi tilbyr alt fra drift og avansert brukerstøtte til spesialiserte tjenester for sensitive data, sier Andreas Jansen, prosjektleder for Norstore ved USIT på UIO.

Dataene lagres slik at det kan forskes på en sikker måte på datagrunnlaget. Hensikten er at kun betroede medarbeidere skal få tilgang til kun de dataene de skal jobbe med. For det er

I mange år har forskere hatt behov for å forske på sensitive data. Bare begrenset forskning har vært foretatt. Det skyldes delvis at det ikke krever veldig avansert analyse for å identifisere anonymiserte helsetilstand. Disse dataene må derfor være særlig godt beskyttet om de skal bli forsket på.

Arkivering
I forbindelse med forskning er arkivering av data vesentlig. Sikker arkivering er en forutsetning for TSD. Arkivering har først de senere årene fått betydning. Mye forskning har jobbet på data, men ikke tatt vare på resultatene for eventuelt ny bearbeidning. Eksempelvis er historiske verdata viktig for forskning, men ikke får morgendagens ver-

30 • Computerworld

It-helse

Nr. 20 • fredag 16. mai 2014

Artikkelen startet på forrige side

Universitetet i Oslos erfaring med Norstore er en viktig forutsetning for TSD.

På grunn av dataenes sensitive natur har prosjektleder Gard Thomassen vært i kontakt med avdelingsdirektør i Tilsyns og sikkerhetsavdelingen Helge Veum i Datailsynet. Henvendelsen søkte å oppnå en forståelse for kravene til den tekniske utrustningen med hensyn til bearbeidning av sensitive data.

Datailsynet fremhever at det er de enkelte forskningsprosjektene som må sørge for nødvendig sikring av de personsensitive dataene. Ett krav er at de skal holdes adskilt.

Større behov

Hjerne gjennom livet er en omfattende forskningsoppgave ved Psykologisk institutt ved Samfunnsvitenskapelig fakultet ved UIO. En forskningsgruppe på 20 personer under ledelse av professor Anders M. Fjell studerer livsløpsendringer, hjerne og kognition.

Utvalget er rundt 1.000 friske personer fra nyfødte til eldre som følges over tid. Forskningsgruppen studerer risikofaktorer, sykdom og skader. Det benyttes nevropsykologiske og kognitive tester. Magnetisk resonans brukes for avbildning av hjernen, MRI. Videre er det EEG/ERP, PET, CSF biomarkører, genetikk og hjerneskadestudier.

Colossus II

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

Verdt:
3,5 millioner kroner

Beregning:

12 Megaware prosesseringsnoder hver med to Intel Xeon E5 med 10 prosessor-kjerner med 2,5 GHz klokke

Minne:

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

Ytelse:

25 TFLOPS

Minneberegning:

To Megaware hver med 4 Intel Xeon E5 med 8 prosessor-kjerner med 2,7 GHz klokke og 1 TB minne for gensekvensering

Lagring:

Kun for bearbeidning, 4 iV-noder hver med 64 GB minne og 45 TB, 1 Metadata-node med SSD

Permanent lagring:

Anskaffes til hvert prosjekt, vil bli mange PB over tid

Forskningsgruppen har stort behov for TSD siden datagrunnlaget fra MRI for en person er på noe over 5 GB. Med PET og EEG/ERP øker datagrunnlaget med 3 GB. En grov vurdering for tusen personer tilsier dermed 10 TB.

— 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, sier professor Anders M. Fjell. Snittbildene av hjernen hos personer med forskjellig alder viser klare forandringer. Det gjøres avstandsmålinger mellom hvitt materie og grå materie i en tredimensjonal modell av hjernehalvkloden.

Kontraster

Kontrasten mellom hvitt og grå materie sier noe om myelin som er et fettaktivt stoff som isolerer nervefibrene slik at disse får sendt signalene bedre.

Håkon Grydelands doktoravhandling handler om hvordan Alzheimers sykdom kan oppdages tidligere ved å studere endringene i myelin.

Det er mye forskning som må gjøres for Alzheimers har bare mennesker, ikke dyr. Alle får den, bare de blir gamle nok. I USA regnes Alzheimers som den tredje største folksykdommen.

— Vi blårer opp hjernen og legger inn et koordinatsystem, forklarer Anders M. Fjell.

Tredimensjonale overflatemodeller benyttes. Vertex er møtepunktet for seks trekanter som får koordinatene x, y og z. To-tall blir det 150.000 trekanter. Hensikten er å måle endringer i baroktykkelsen over livsløpet. Det er normalt en ølg reduksjon.

— Kognitive evner endrer seg over livsløpet. Det er en kraftig reduksjon fra 20-årene, forklarer Anders M. Fjell.

Trening kan øke tykkelsen på hjernehalvkloden slik at vi truster bedre. Metoden hvor ting plasseres langs en kjent rute slik at det blir et mentalt kart bidrar bedre hukommelse. Andreas Engvik har ledet et studium som omhandler hvordan man kan lære seg opp for å umngå Alzheimers.

Mat og helse

— Det er kanskje den nyeste varianten for slanking. To dagers lite spising, fem dagers normalt kosthold. For aviser selger på kosthold og slanking. Høy kroppsmasse, BMI, henger sammen med Alzheimers sykdom.

Hva som er sunn mat krever omfattende dokumentasjon med behov for å gjøre befolkningsstudier gjennom mange år. — Sammenheng mellom mat og helse er en komplisert problemstilling. Det tar mange år å få sykdommer. Det er ikke mulig å følge mennesker i 30 år. Vi kan derfor bare se på indikatorer, men det er ikke det samme, sier professor Lene Frost Andersen, Institutt for medisinske basalfag ved Det medisinske fakultet ved UIO.



SIKKRETT: Forskningsgruppelider Hans Eide og IT-direktør Lars Oftedal har vært prosjektledere for TSD. Mellom tve data. Maskin og lagring er sikret både med tilgang og fysisk. Dataene er anonymiserte.

Siden det er personidentifiserbare data er TSD nødvendig for anonymisering og etterbehandling. Alle som er med på studien må knytte seg opp mot helsenorge.no og logge seg inn med Minid for å umngå jaks.

Derfor er samarbeidet med USIT (Universitetets senter for IT), både webseksjonen og TSD av stor betydning. I 2015 skal nettlesningen kunne brukes. Håpet er at forskningen skal bli lettere ved å kombinere datagrunnlaget med kartkodet data og biobanker.

Hvordan maten i ungdomstiden påvirker senere helse er av stor interesse, men det krever hyppig måling. Også ernæringsbehandling ved sykdom ønskes det mer data om. Spises det for lite grønnsaker eller for mye.

Storskala dataanalyse

— Det har vært en betydelig reduksjon i herte-karsykdommer og kreft, men ikke for psykiske lidelser, sier Martin Tveit, Norment, KG Jøhns senter for psykoneuroforskning OUS, UIO.

I 1996 var det for pasienter med schizofreni et gap i forventet levealder på 25 år i forhold til resten av befolkningen. I 2006 var den blitt redusert til 23 år.

Psykiatrien utfordrer er at diagnosen er basert på symptomer med lite

er målet å identifisere hjernefenotyper som forbinde gener med klinisk innleggelse. I Bergen fortas det forskning på hallusinasjoner. Hensikten er å analysere forlop og utfall.

For schizofreni er det 2,5 millioner genotyper per person hvilket tilsvarer 500 milliarder variabler. Behovet for bearbeidning av dataene er enormt. Det er fortsatt for lite kunnskap med hvor heterogenitet med et stort spørsmål om hvordan forskning skal oversettes til klinisk nytte.

Starkt

Professor Eivind Hovig ved OUS/UIO og Norsk kreftgenom konsortium sitter startklar og venter på Colossus. Oppgavene ligger klare. Heltet skal de bearbeides før 17. mai, men forskningen kan kanskje ikke starte før i slutten av måneden.

— Det er omtrent 27.000 krefttillfeller hvert år i Helse Sør-Øst. Dersom man seleveneriserte disse minner en gang og beholdt informasjonen om sekvensene, ville kravene til lagring og prosessering bli ganske krevende, sier Eivind Hovig.

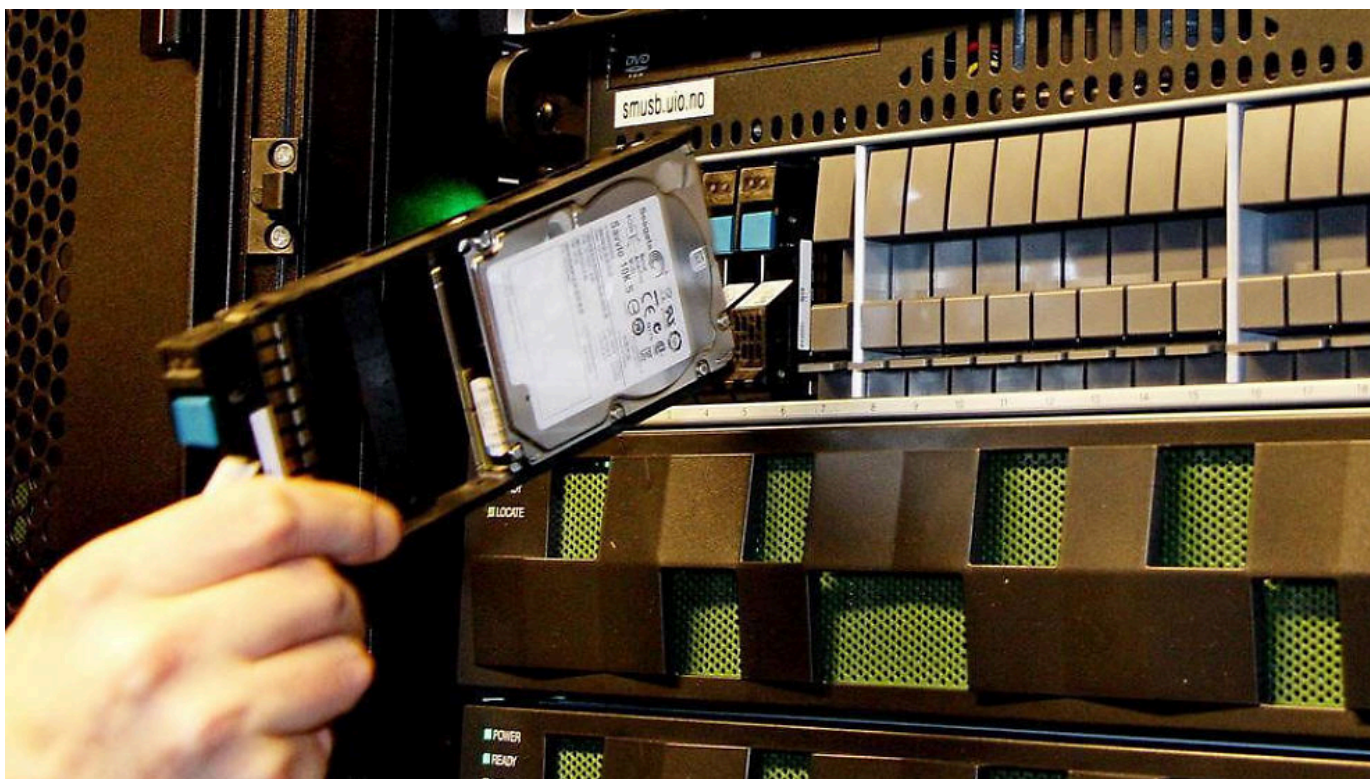
Genomanalyse av kreft har spesielle utfordringer siden det krever en tverrfaglig gruppe med kunnskap innen bioinformatikk, genoteknologi, kreftbiologi, onkologi og patologi. Konsekvensen er felles prosedyrer for

Sammenhengen mellom mat og helse er en komplisert problemstilling.

ANLETT HYSSING

Den femte mai åpnet Ole Petter Ottersen, rek-

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

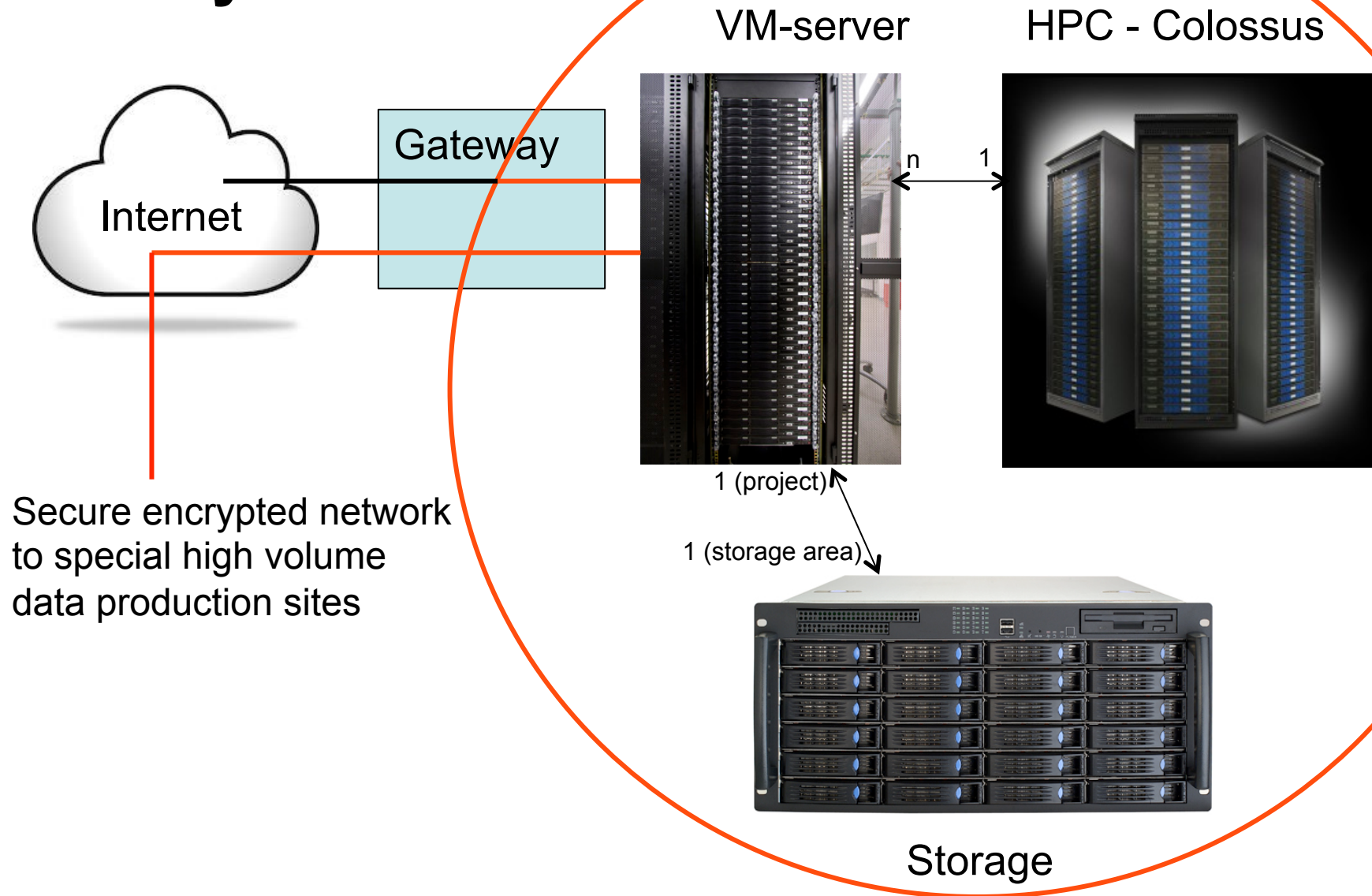
Norge får Nordens beste anlegg for hemmelige data

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



Using TSD

Libre Office
R
Module load
...



User₁ Study₁

User₂ Study₁

GW

SPSS
Office
Stata
SAS
R
Matlab
....



TSD

VM U₁ S₁

VM U₂ S₁

S₁ DB

Front end
Colossus

TSD disk

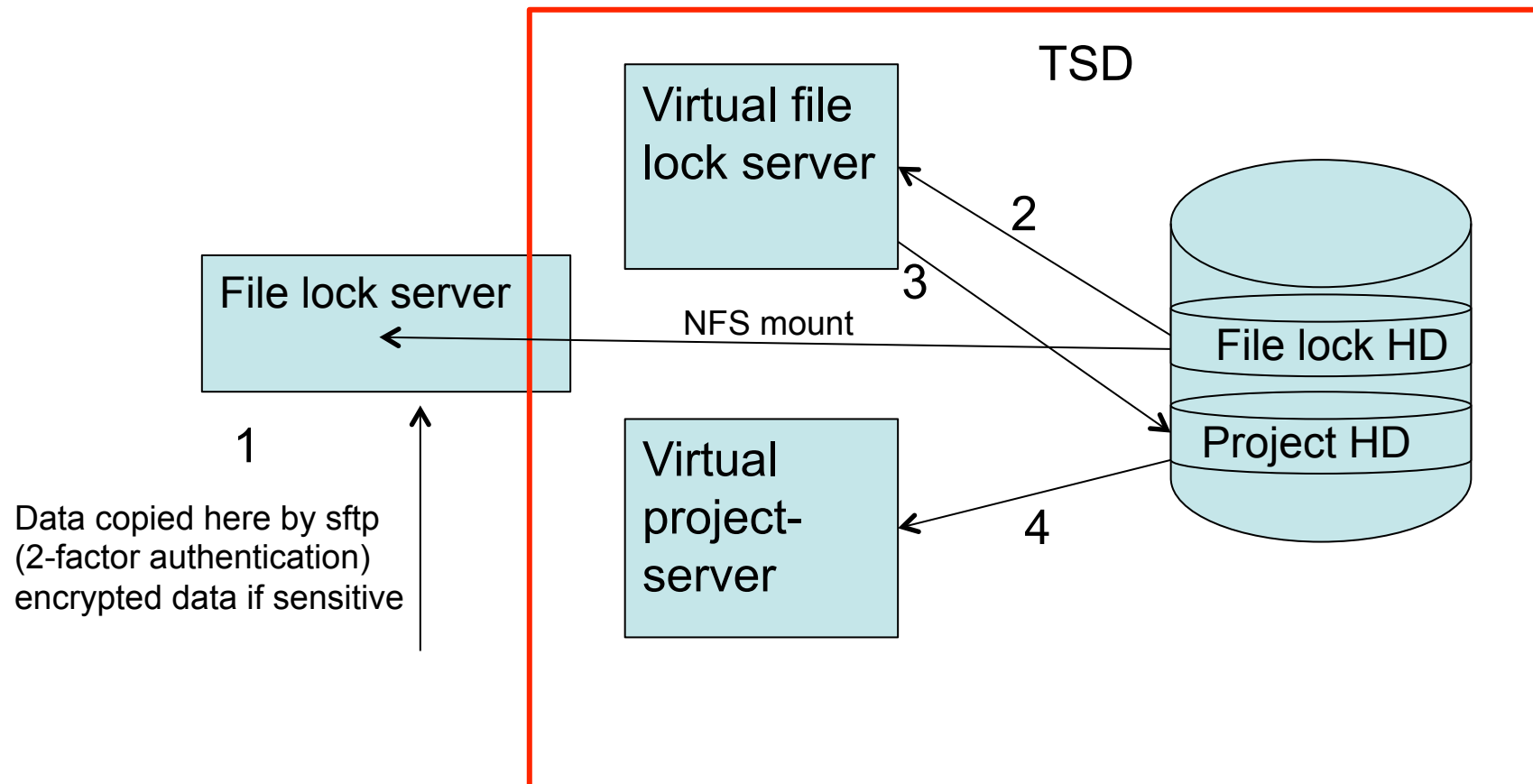
S₁

Colossus

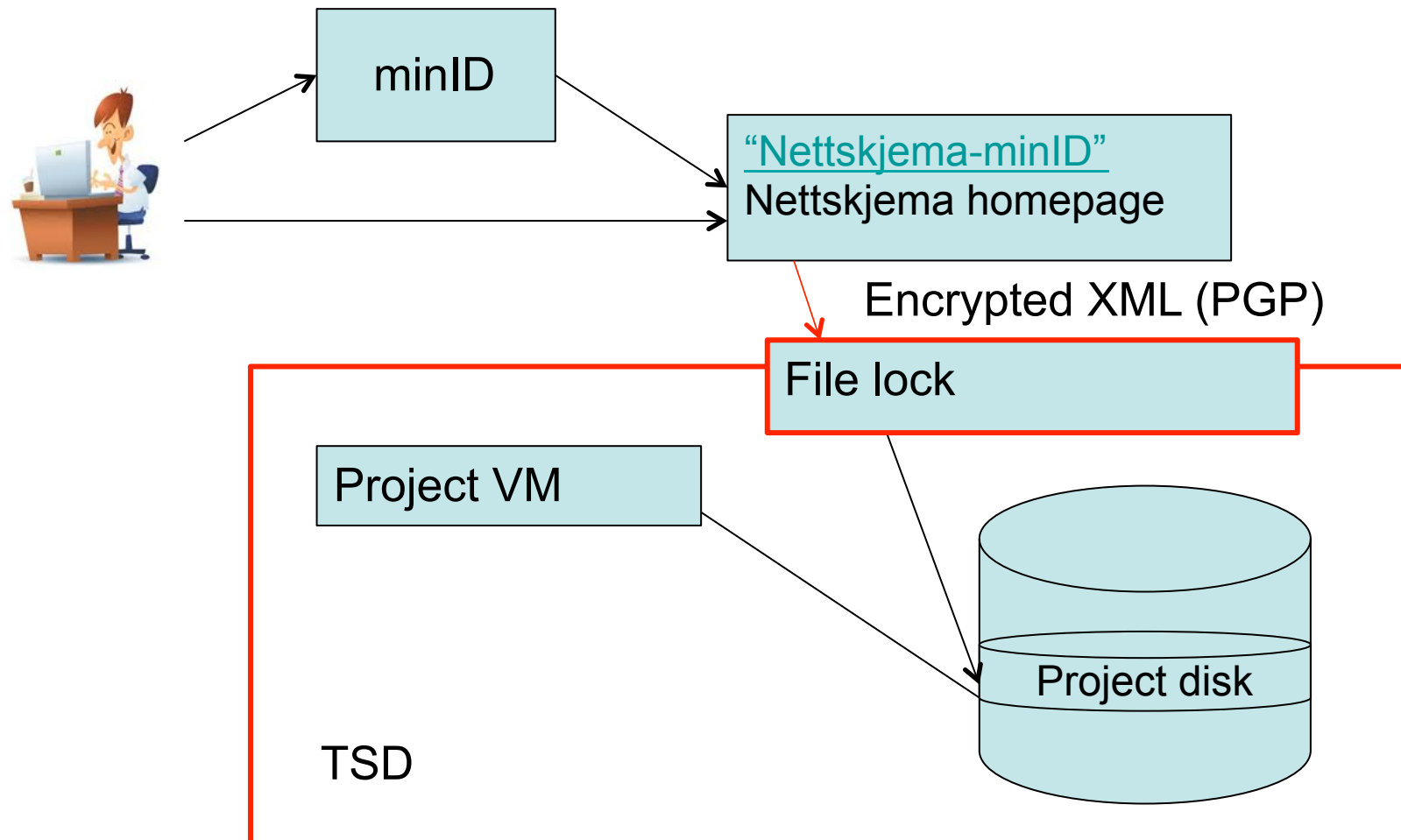


Colossus disk

Data import and export using TSD

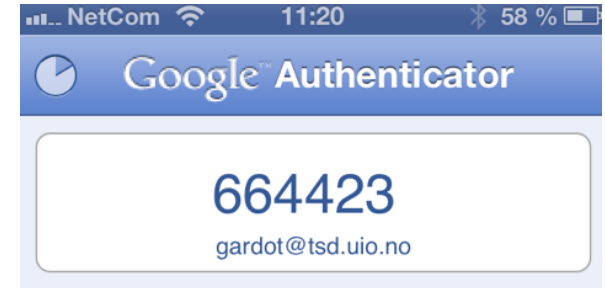


Data collection using TSD



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/](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 investment 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 above 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



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!