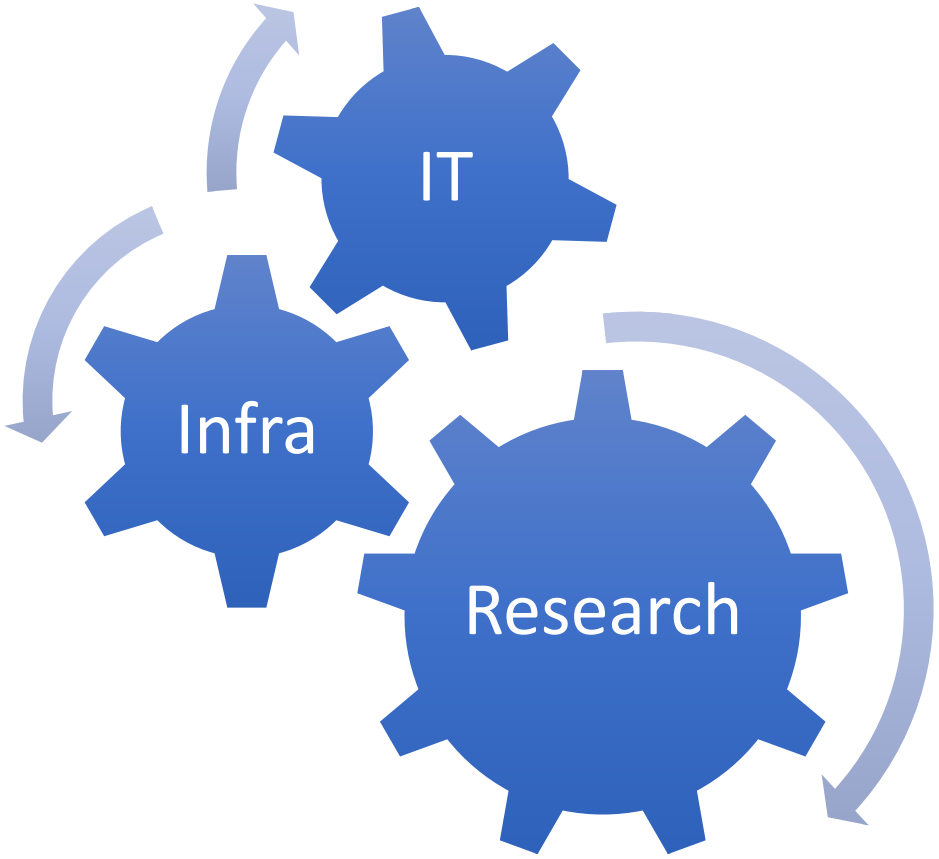
Academic excellence since 1632

EMIF

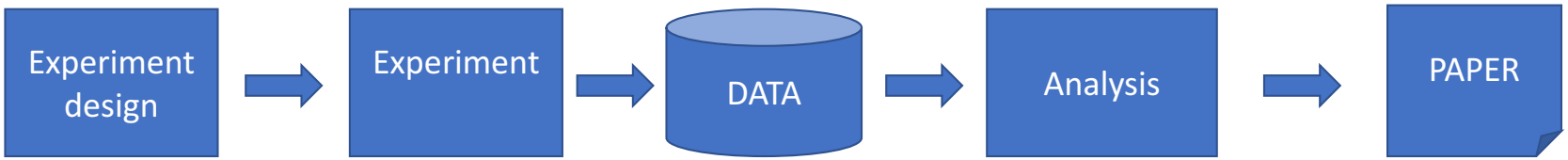
European Medica

Information

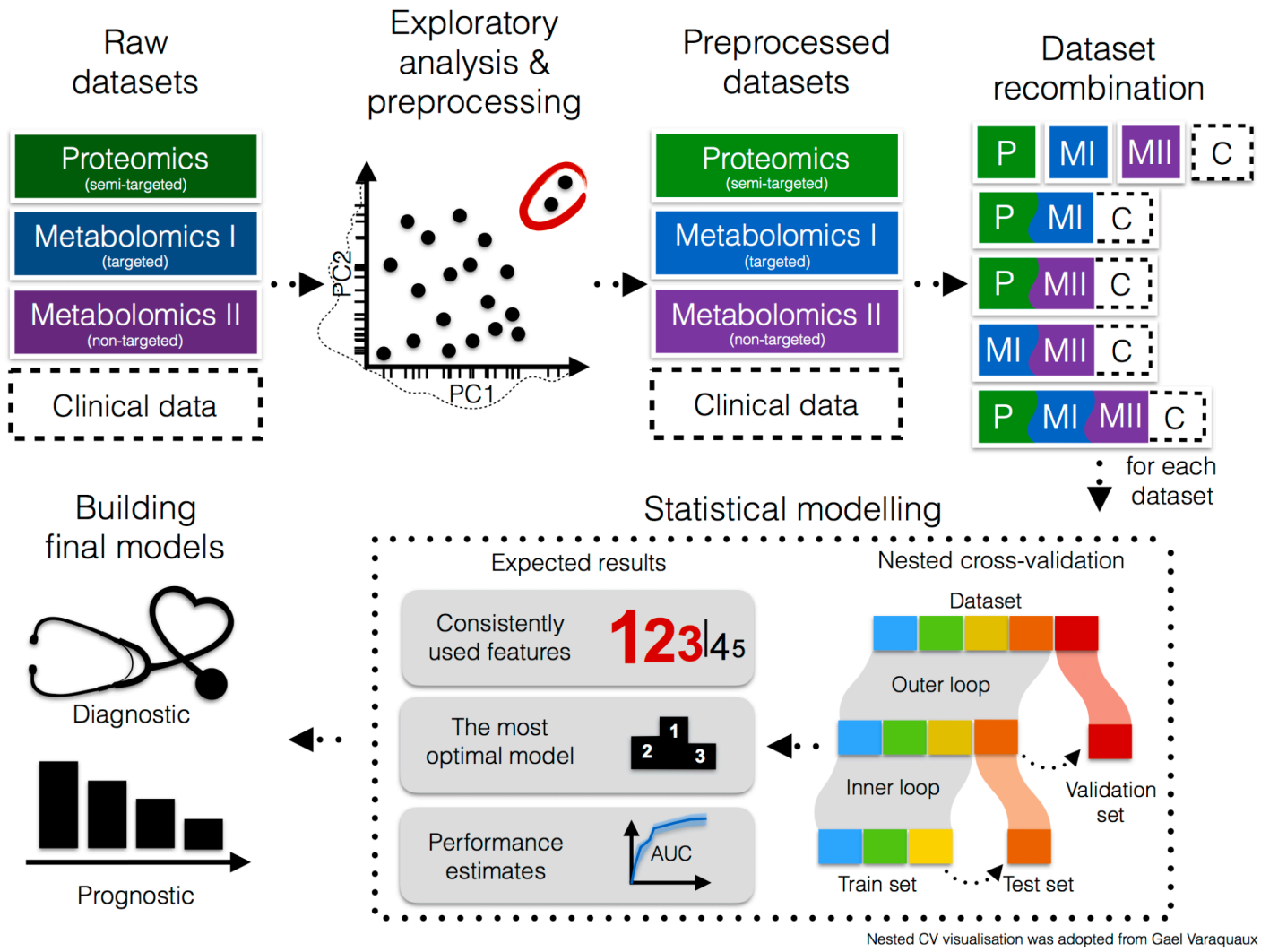
Framework


IT
Infra
Research

Research “workflow”



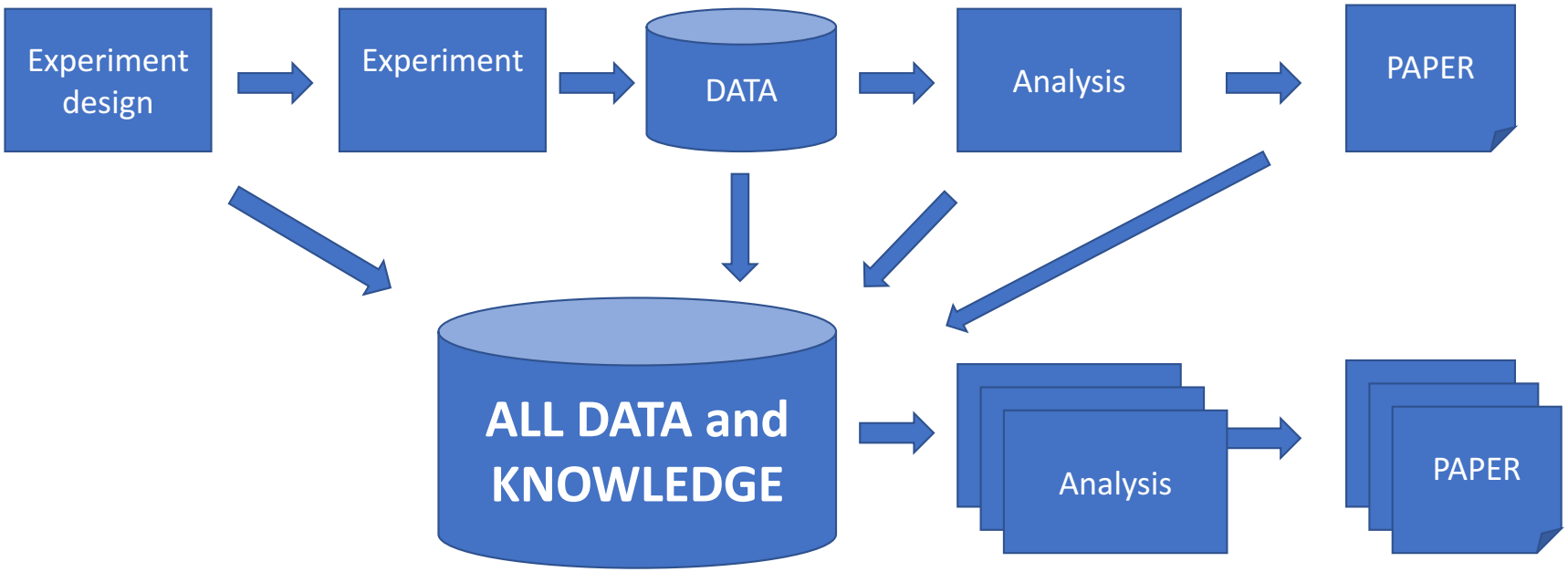
Annex 7: Strategic plan for data analysis within the BioEndoCar project

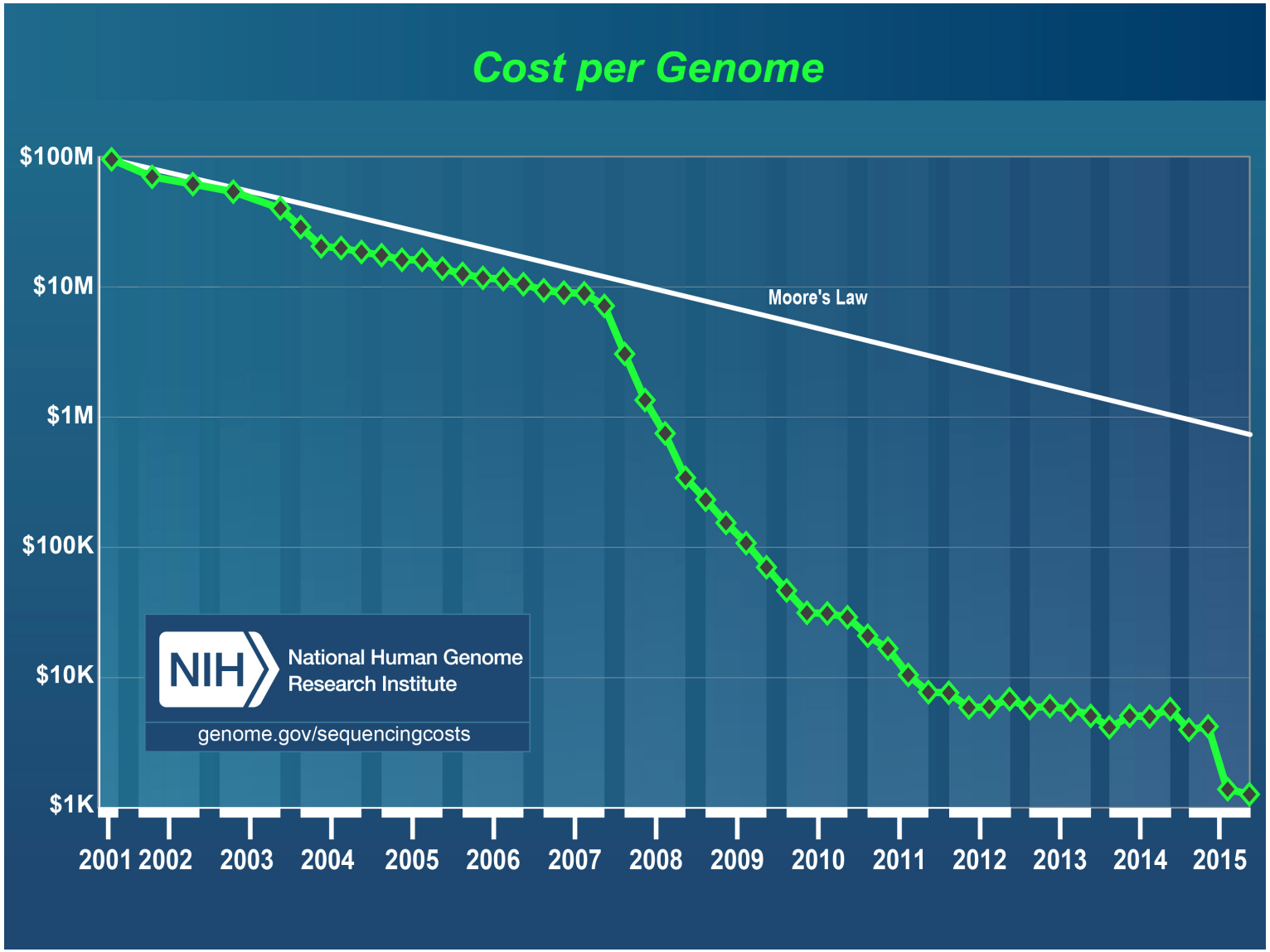


Nested CV visualisation was adopted from Gael Varaquaux

Hypothesis-driven
vs
Hypothesis-free(?)

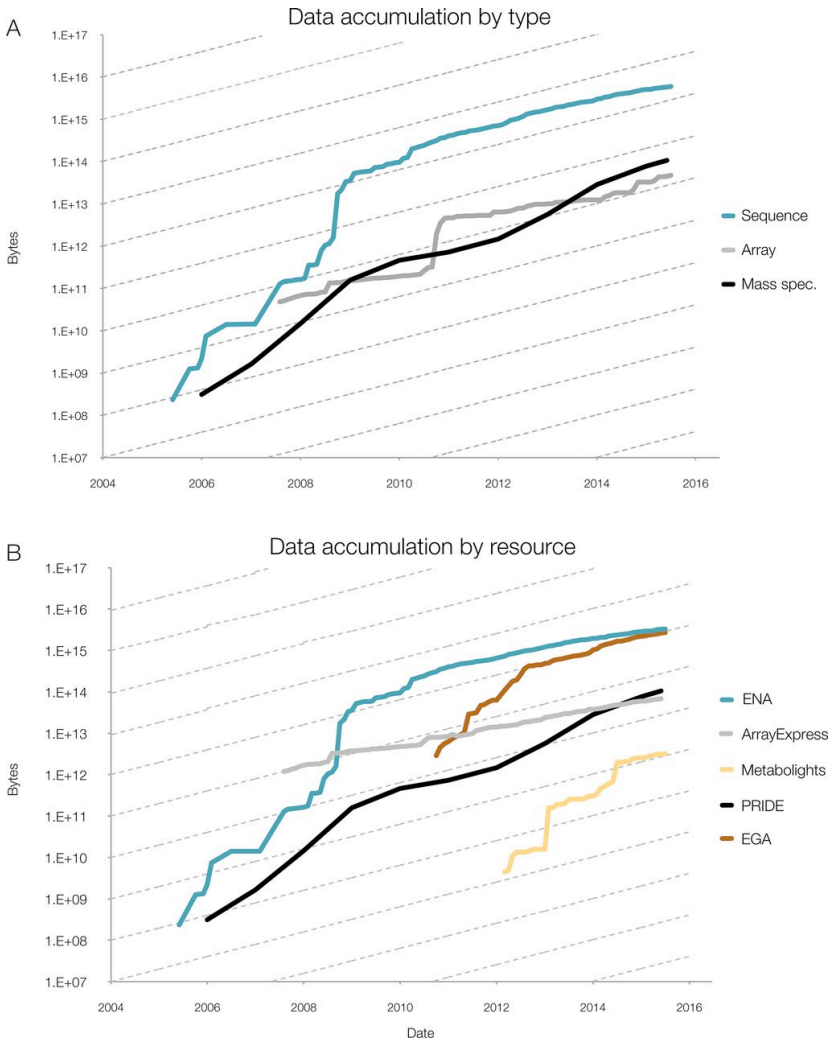
Research “workflow”




Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

SOLID 5500

Illumina HiSeq2000



Data growth in the life sciences

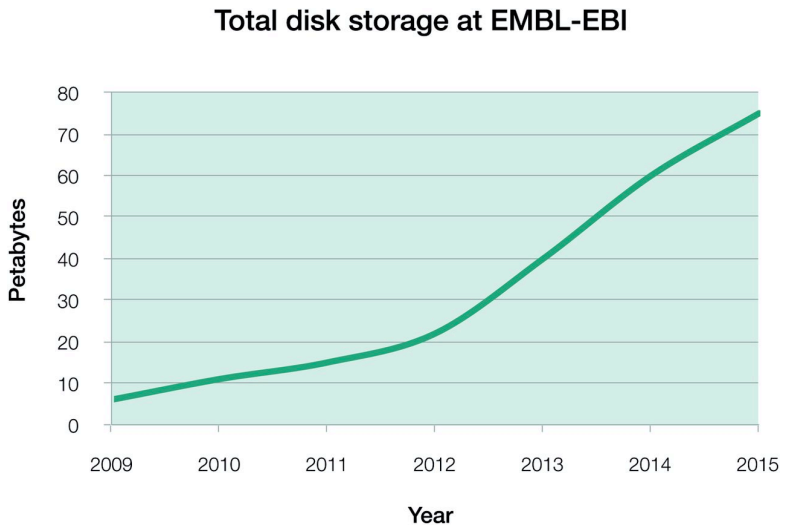
Data growth at EMBL-EBI

Source: Charles E. Cook et al. *Nucl. Acids Res.* 2016;44:D20-D26

The data challenge: Data growth

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady

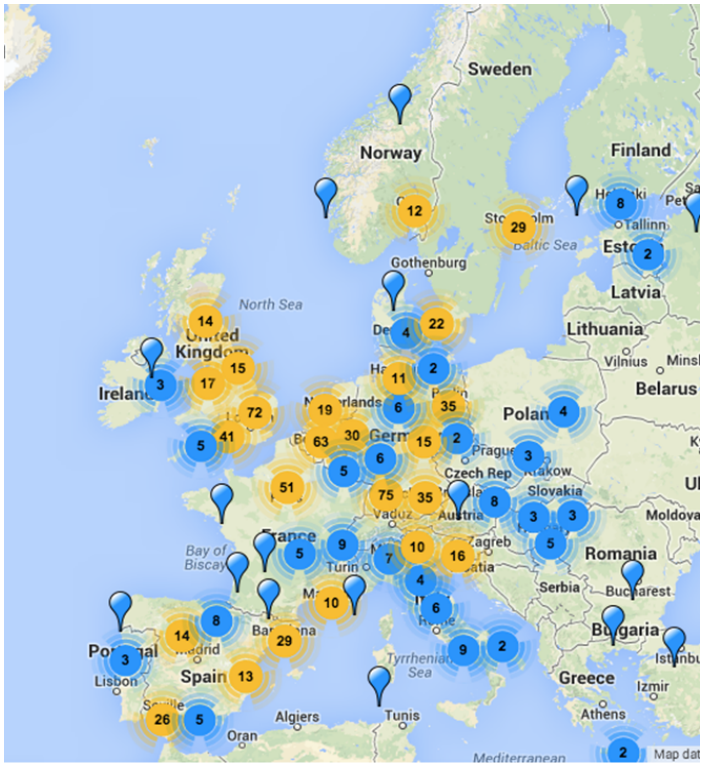
- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this decade



Source: Charles E. Cook at al. *Nucl. Acids Res. 2016; 44: D20-D26*

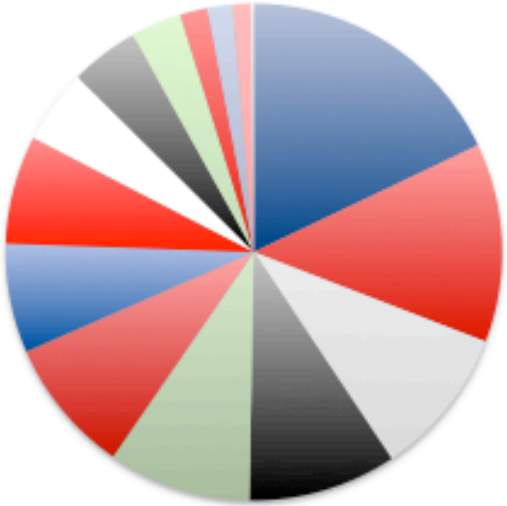
The data challenge: Geographic spread

- Data production sites increasing across Europe
- European Illumina seq sales up 20% 2014



Source: http://omicsmaps.com

Data resources in life science



- Genomics Databases (non-vertebrate) (17.9%)
- Protein sequence databases (12.9%)
- Human Genes and Diseases (9.8%)
- Structure Databases (9.7%)
- Metabolic and Signaling Pathways (9.3%)
- Nucleotide Sequence Databases (8.8%)
- Human and other Vertebrate Genomes (7.1%)
- Plant databases (7.1%)
- RNA sequence databases (4.9%)
- Microarray and other Gene Expression Databases (4.5%)
- Other Molecular Biology Databases (3.3%)
- Immunological databases (1.8%)
- Organelle databases (1.6%)
- Proteomics Resources (1.2%)
- Cell biology (0.2%)

molecular biology data resources

~1800

Nucleic Acids Research annual Database Issue
and the NAR online Molecular Biology Database Collection in 2012.
MY Galperin, GR Cochrane – Nucleic Acids Research, 2011

We generate data faster than we can deposit it

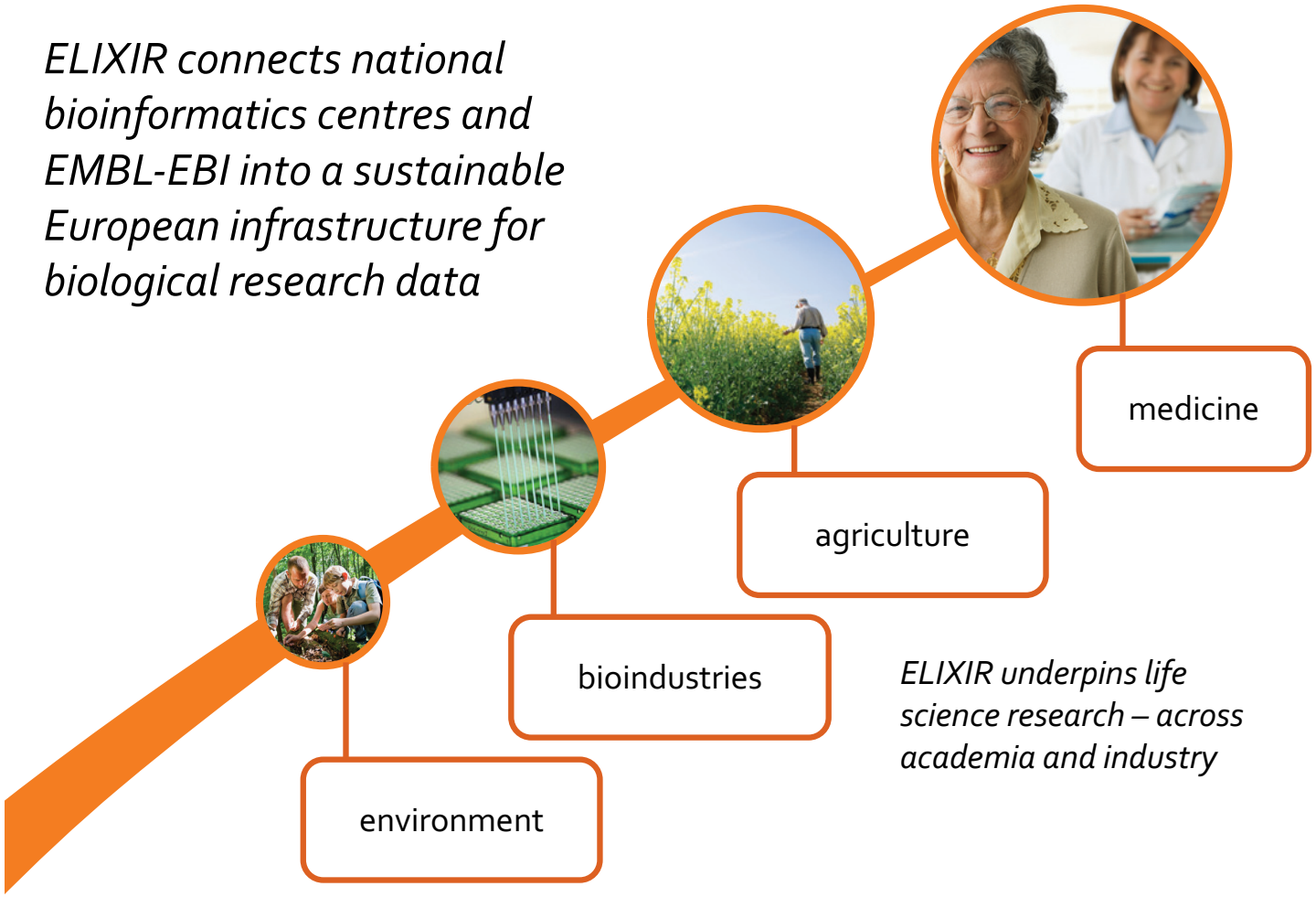
		Network file transfer rate
24 hours		**100 Mb**
	DNA sequencing ~100 GB	~5 hours
	Mass spectrometry ~4 TB	~4 days
	Microscopy ~4 TB	~4 days

ELIXIR

Safeguarding the results of life science research in Europe

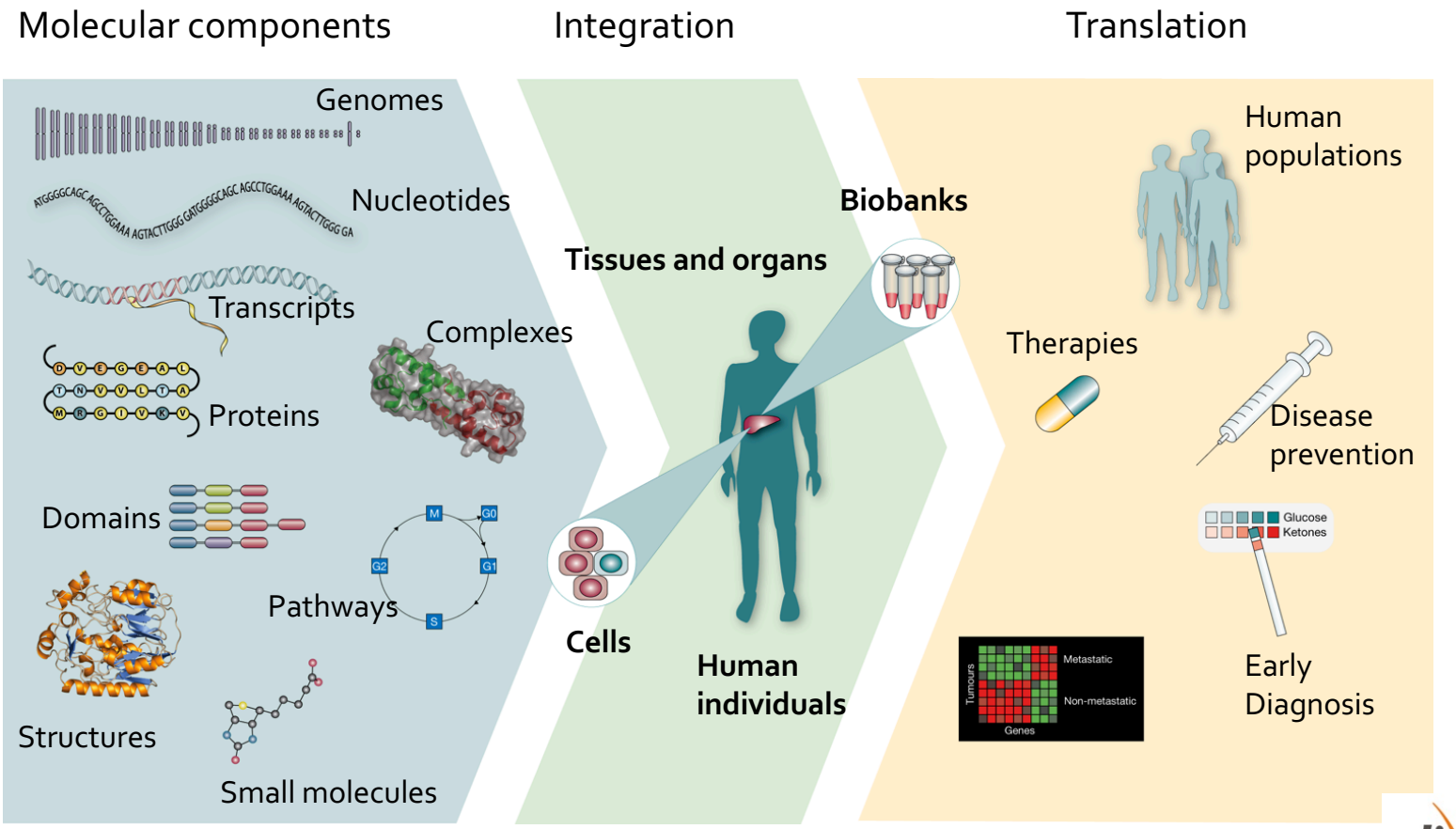
www.elixir-europe.org

ELIXIR connects national bioinformatics centres and EMBL-EBI into a sustainable European infrastructure for biological research data



ELIXIR underpins life science research – across academia and industry

From molecules to medicine



ELIXIR: European infrastructure for biological information

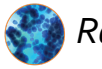
Data infrastructure for Europe's life-science research:

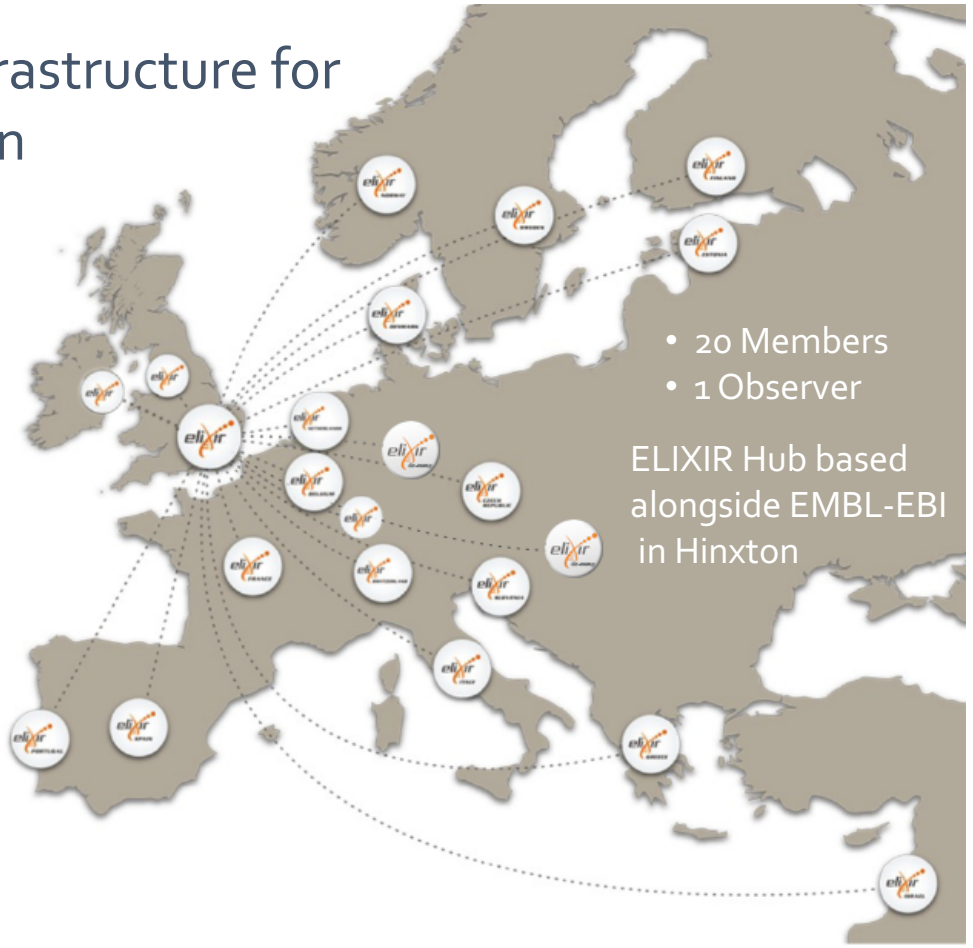
- *Data*
- *Interoperability*
- *Tools*
- *Compute*
- *Training*


Marine metagenomics


Crop and forest plants


Human data


Rare diseases

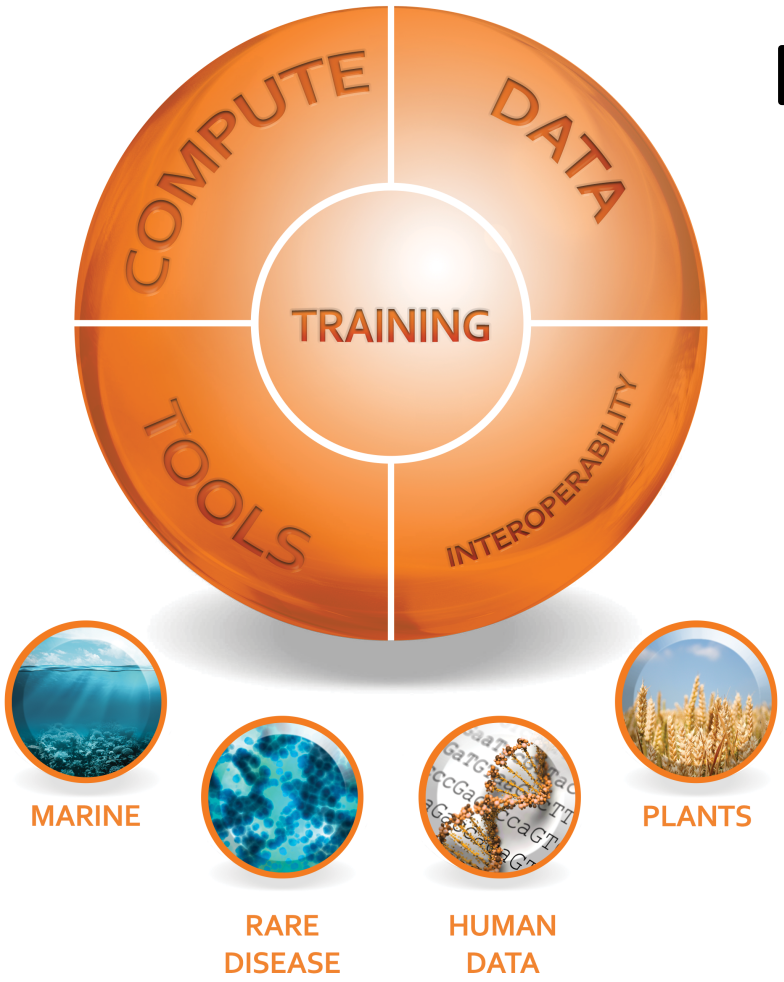


www.elixir-europe.org

@ELIXIREurope

ELIXIR Members

ELIXIR Observers



ELIXIR Structure

Five technical platforms for Compute, Data, Tools and Interoperability

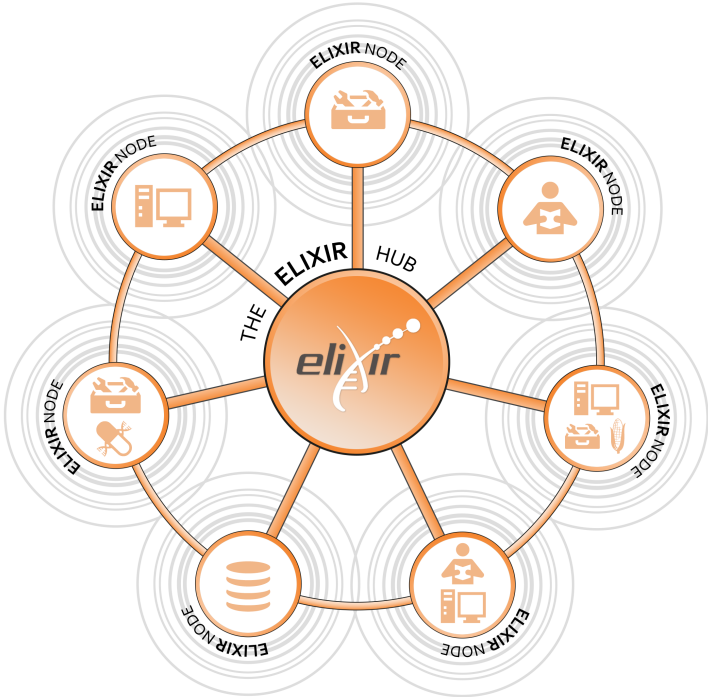
Complemented by four use cases for marine meta-genomics, rare diseases, human data and plants sciences

A distributed infrastructure to scale with the challenge

ELIXIR data infrastructure for Europe's life science research sector

ELIXIR Nodes build local bioinformatics capacity throughout Europe

ELIXIR Nodes build on national strengths and priorities



FAIR principles for data

Findable

Accessible

Interoperable

Reusable

Data Management Plan; Data management life-cycle

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

To be Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

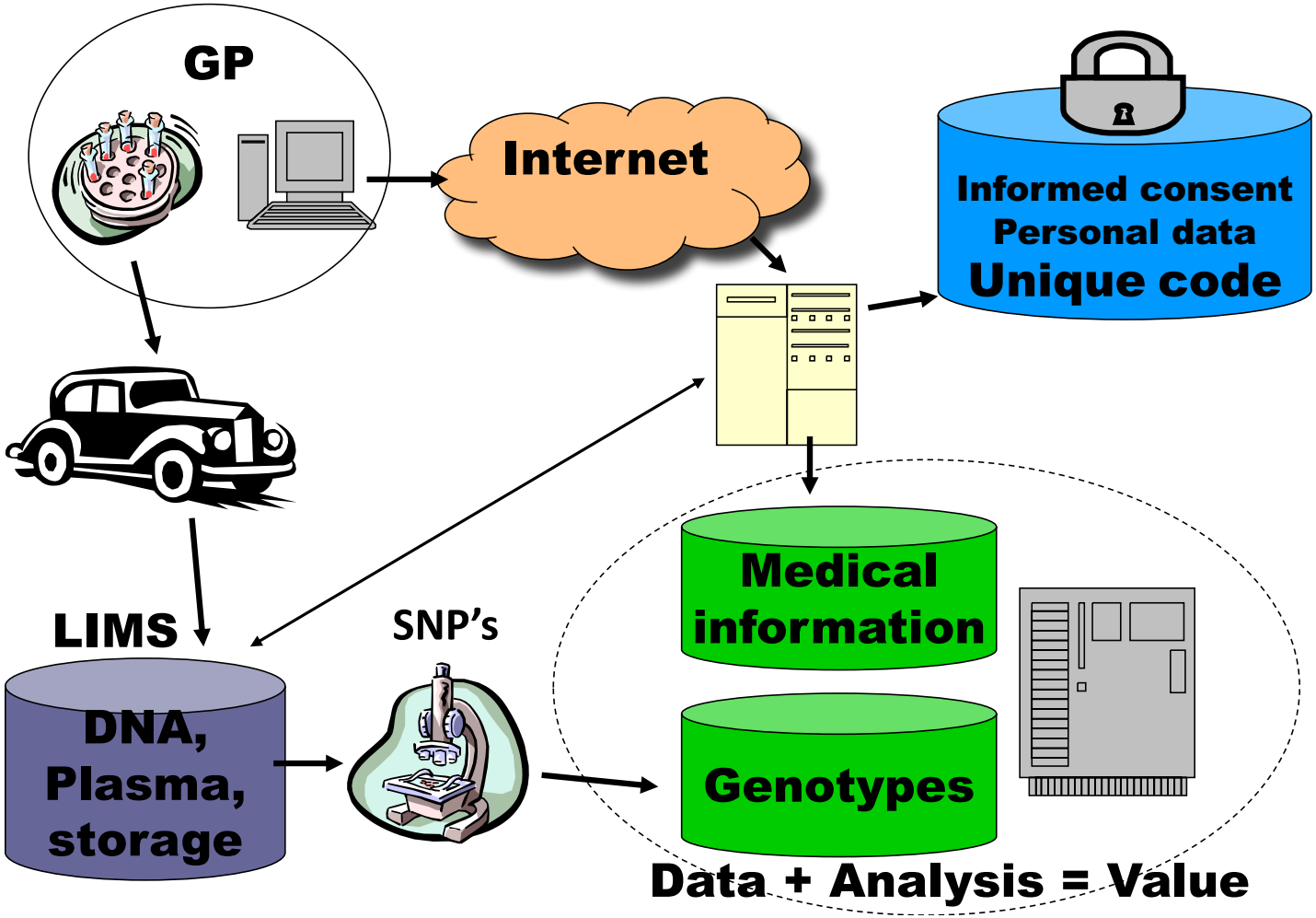
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

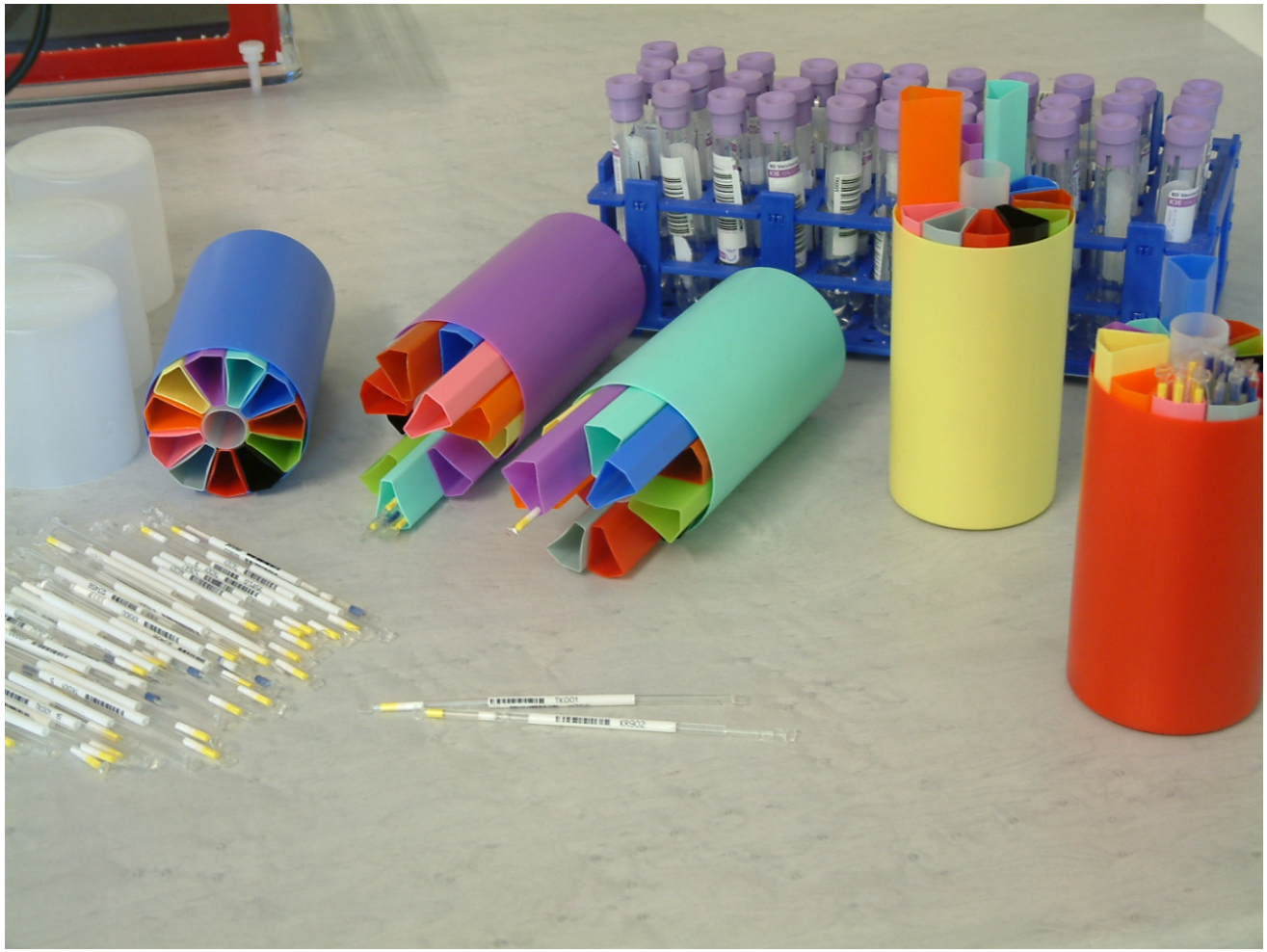
To be Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

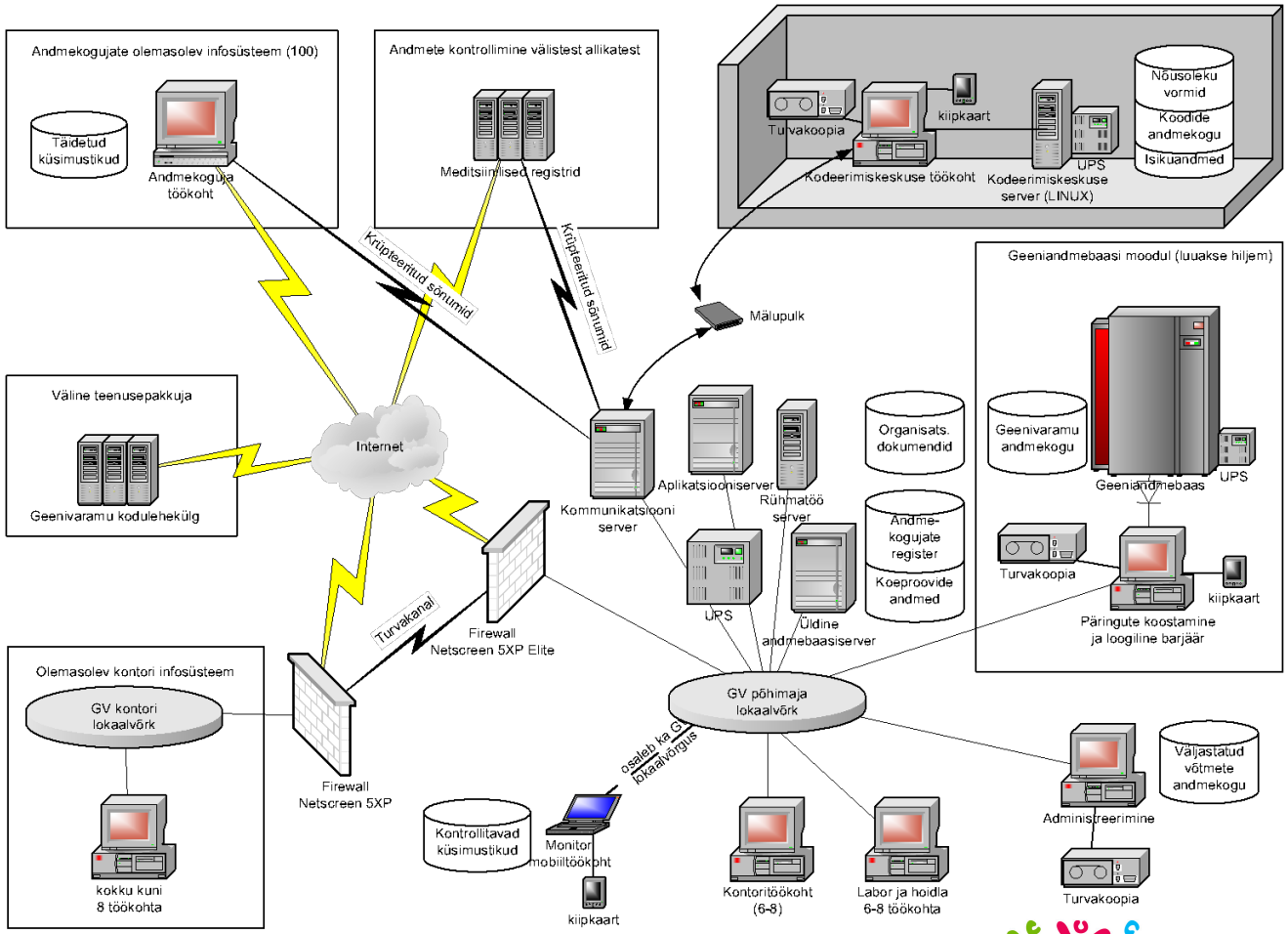
Rewind 18 years… Estonian Biobank

Process of data collection and handling





Data collections



eesti geenivaramu

EGeen

Company Introduction & Estonian Registries

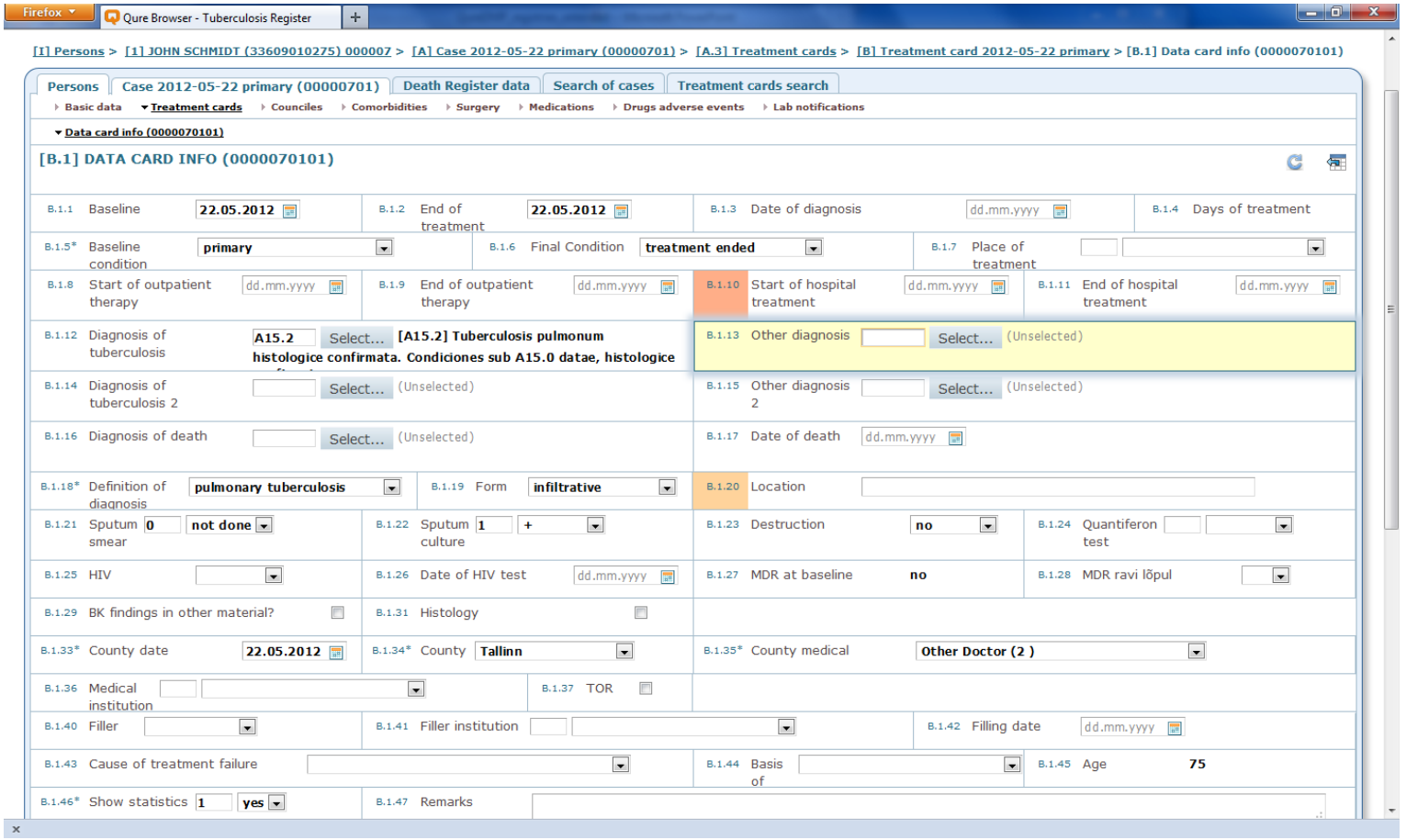
www.quretec.com

Qure Data Management Platform

The core product of Quretec, the Qure Data Management software Platform has been designed for collection, handling, and analysis of complex data like health registries, questionnaires, clinical case report forms and other rich metadata under high quality, security, and robustness requirements.

QureTEC
www.quretec.com

Qure Browser – a web based data entry user interface



QureTEC
www.quretec.com

Estonian Health Registries on Qure Data Management Platform

Tervise Arengu Instituut
National Institute for Health Development

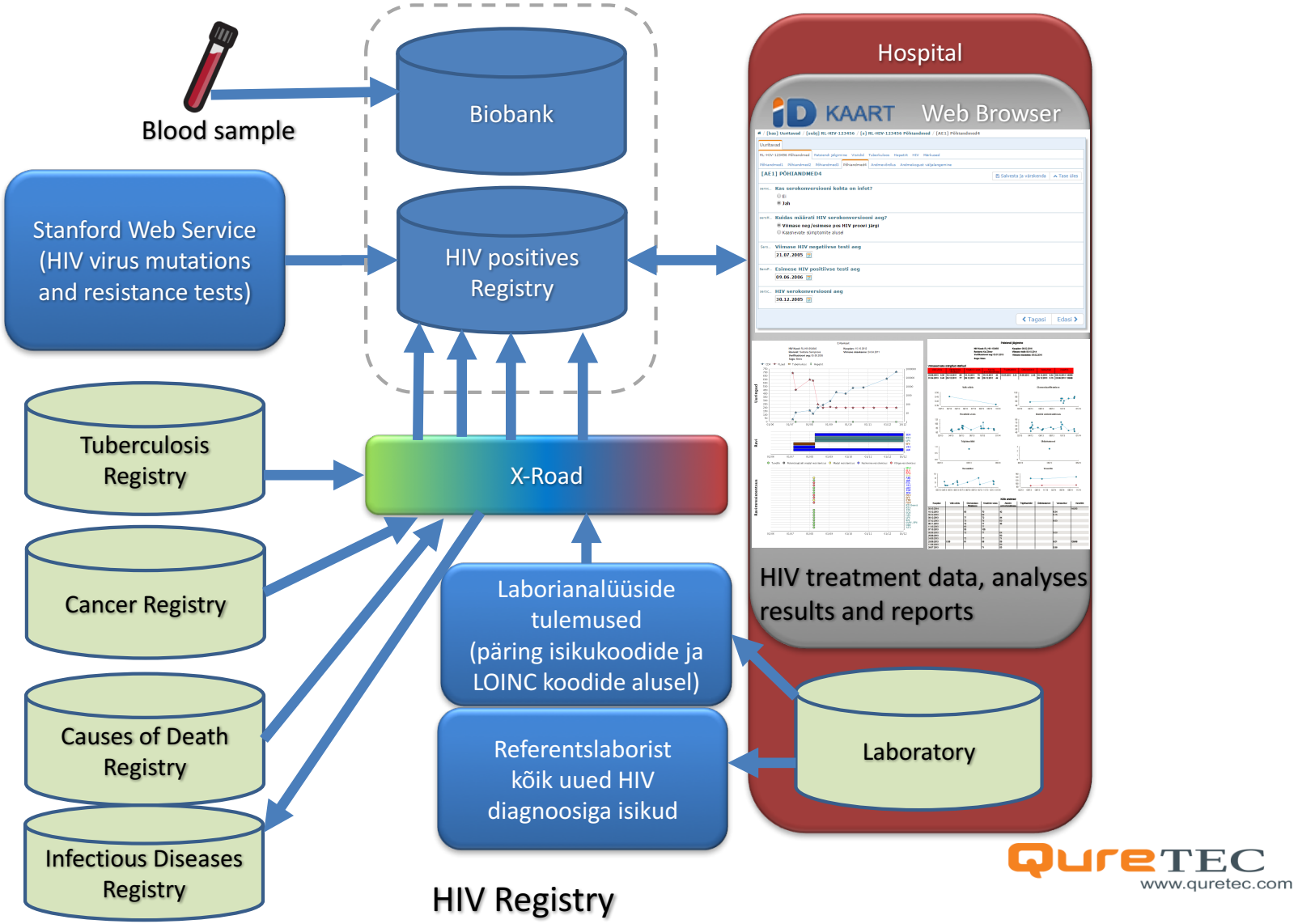
UNIVERSITY OF TARTU

estonian genome center

- HIV registry
- Cancer registry
- Tuberculosis registry
- Causes of death registry
- Medical birth registry
- Abortion registry
- Drug treatment database
- Infectious diseases registry

- Estonian Genome Center Biobank
- North Estonia Medical Centre hospital registries
  - Breast Cancer
  - Colon Cancer
  - Pulmonary Arterial Hypertension

QureTEC
www.quretec.com


Blood sample
Biobank
Stanford Web Service (HIV virus mutations and resistance tests)
HIV positives Registry
Tuberculosis Registry
Cancer Registry
Causes of Death Registry
Infectious Diseases Registry
X-Road
Laborianalüüside tulemused (päring isikukoodide ja LOINC koodide alusel)
Referentslaborist kõik uued HIV diagnoosiga isikud
Hospital
KAART Web Browser
HIV treatment data, analyses results and reports
Laboratory
HIV Registry
QureTEC
www.quretec.com

QureTEC

- Health registries (Cancer, HIV, Tuberculosis and other registries)
- Biobanks (Estonian Genome Center)
- Clinical Trials
  - EDC
  - Data Management
  - Statistical Analysis
  - Medical Writing

QureCLINICAL

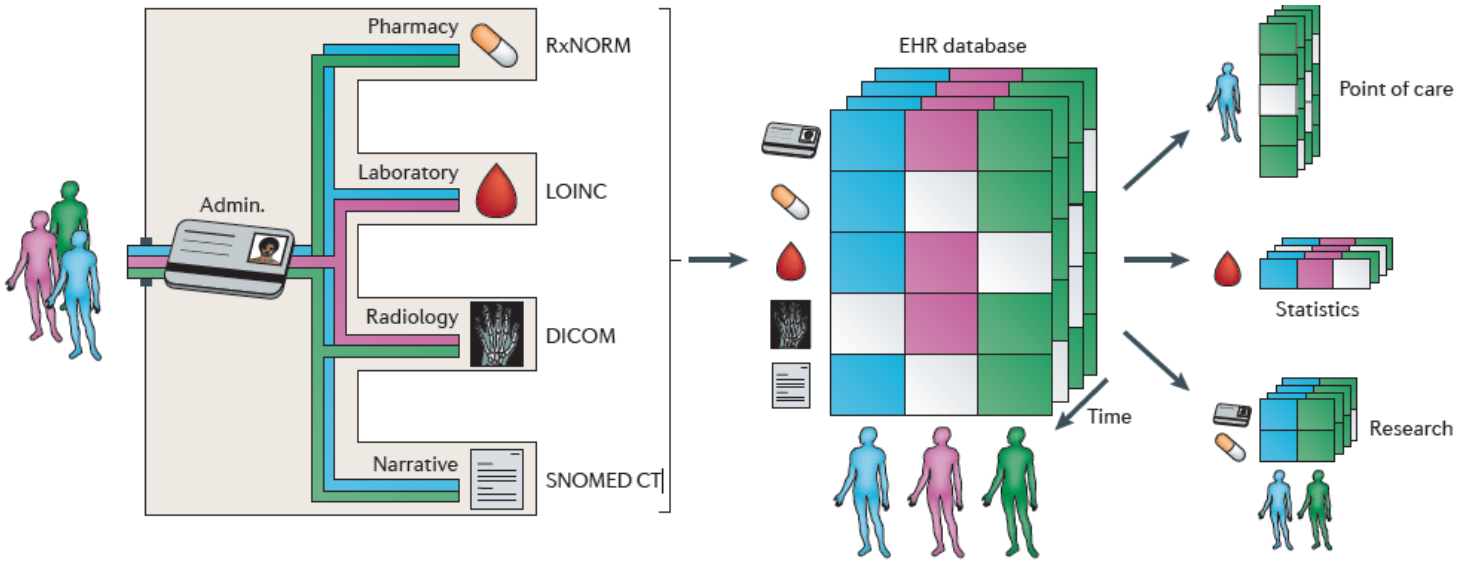
- Clinical Trials
  - EDC, Data Management, Statistical Analysis, Medical Writing
- 5 active studies at the moment
- Clinical Trial phases I-III
- >100 sites
- >400 active EDC users
- >1000 subjects eCRFs
- Rheumatology, Gynecology, Oncology, Pediatrics

QureTEC
www.quretec.com

Näiteid toetatud uuringutest:

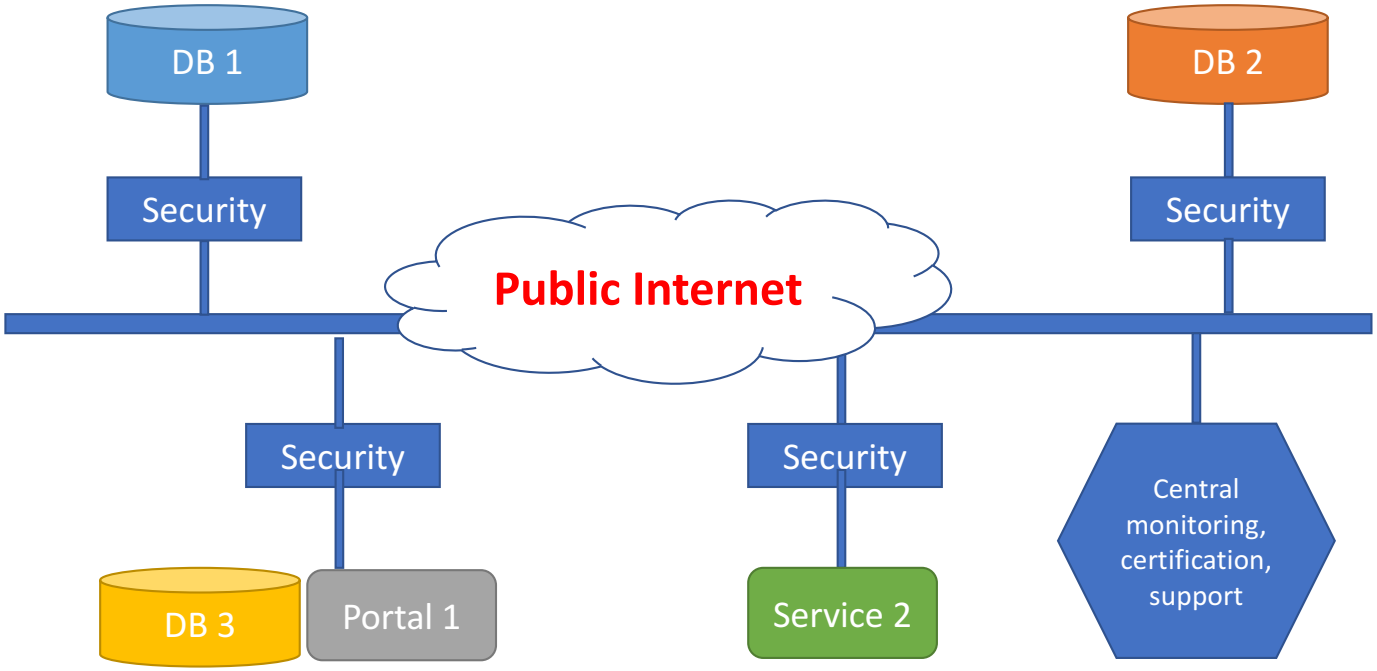
CF111 - 2014-2015 uuring Drospirenone farmakokineetika ja -dünaamika hindamiseks rinnapiimas
H-2011/07-UA - 2013-2014 uuring Herbeprot-P efektiivsuse ja ohutuse hingamiseks troofilise jalahaavandiga patsientidel.
DICL001 - 2014-2017 uuring Diclofenac geeli bioekvivalentsuse määramiseks Voltaren geeli suhtes põlve osteoartroosiga patsientidel SPIRO - 2014-2015 - uuring Spironolaktooni farmakokineetika hindamiseks lastel
HPV-EU-001 - 2015-... uuring HPV vaktsiini turvalisuse ja efektiivsuse määramiseks HPV 16/18 positiivsetel patsientidel
MDP-CLIN-001 - 2015-... uuring Cisplatini ja radioteraapia turvalisuse hindamiseks pea- ja kaelavähi patsientidel
ARC 2015-... uuring Cefepime farmakokineetika ja -dünaamika hindamiseks lastel
EFIT - 2016-... uuring rTMSi (repetitve transcranial magnetic stimulation) efektiivsuse hindamiseks insuldijärgsetel patsientidel
ETVAX - 2017-... uuring ETEC vaktsiini turvalisuse, efektiivsuse ja immunogeensuse hindamiseks Lääne-Aafrikat külastavatel patsientidel

Digitaalne terviselugu ja selle potentsiaal

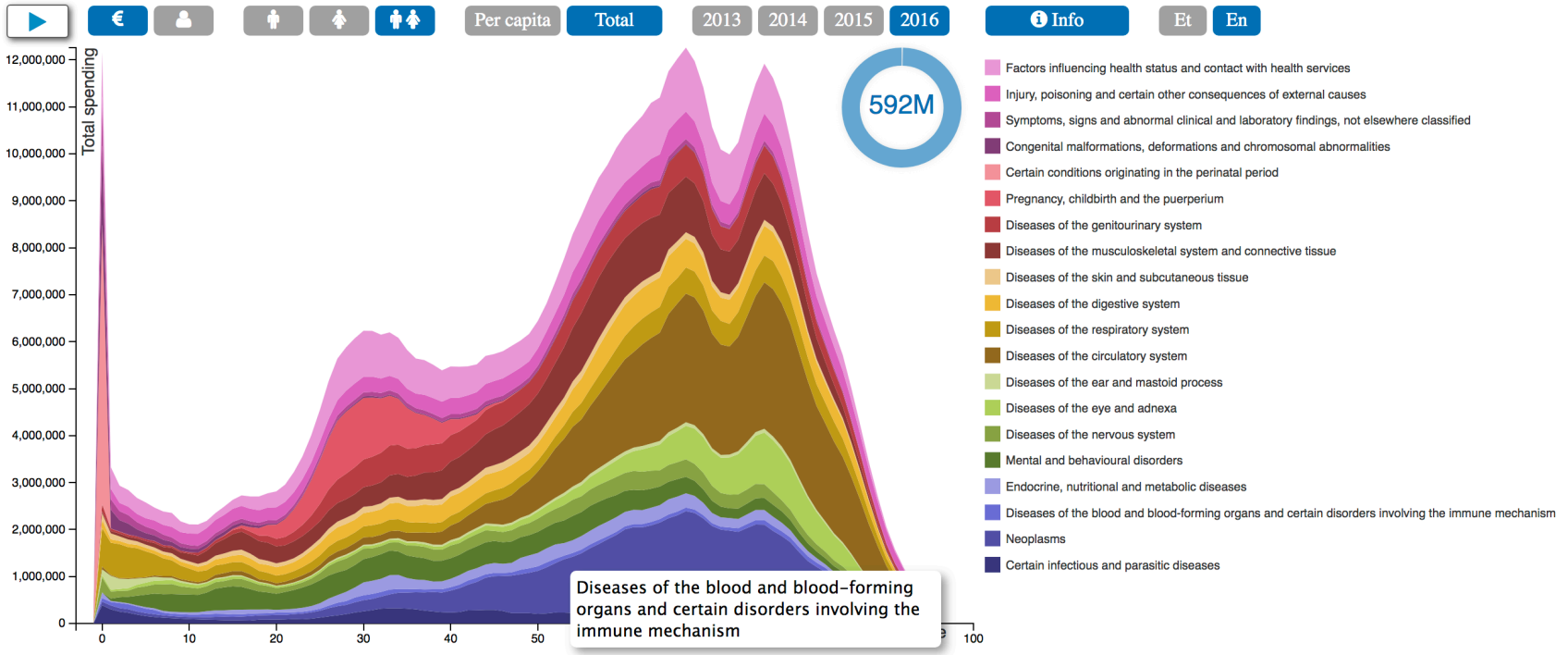


Peter B. Jensen, Lars J. Jensen and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 13, 395-405.

Linked databases - X-Road common bus



Overview of the Estonian specialized medical care spending and patient counts by age and diagnosis



Source: Estonian Health Insurance Fund

Estonian Health Insurance Fund

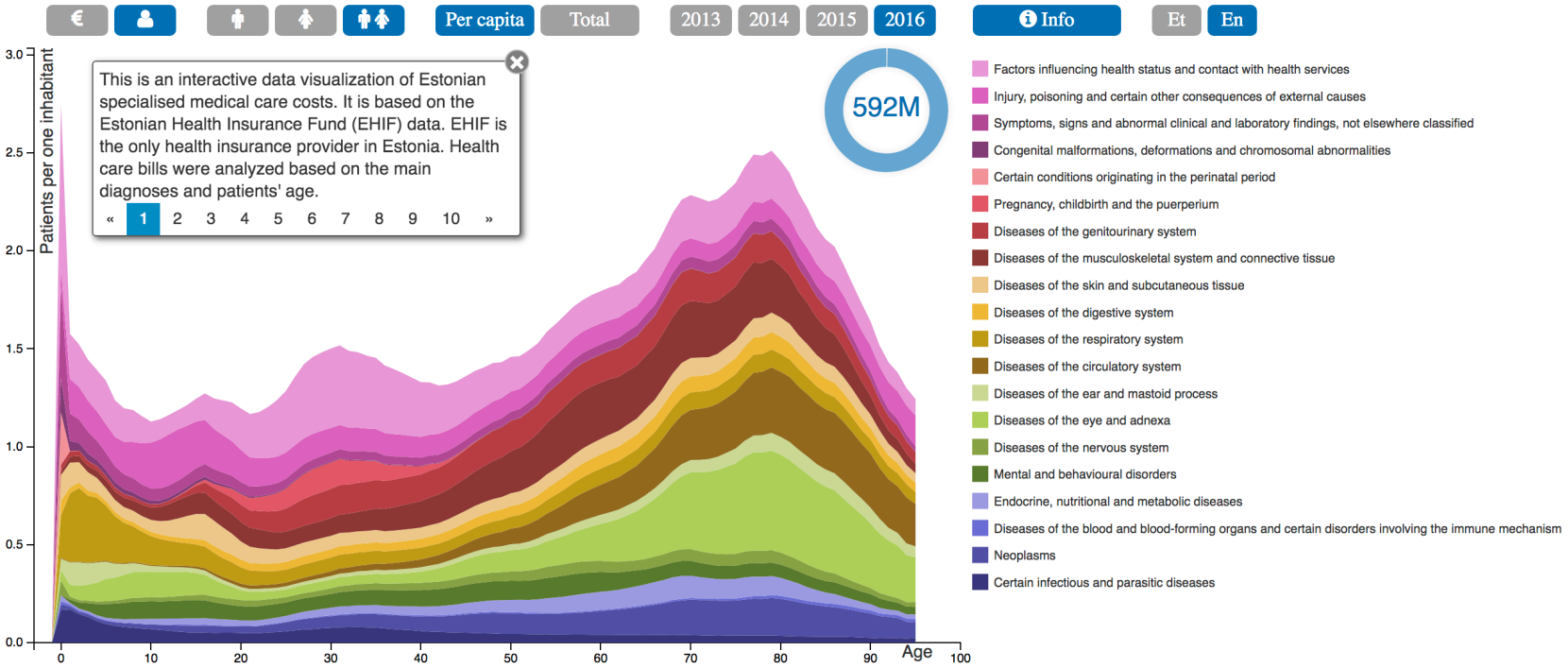
STACC Software Technology and Applications Competence Center

Euroopa Liit Ühtekuuluvusfond

Eesti tuleviku heaks

Tartu Ülikool Universitas Tartuensis 1632

Overview of the Estonian specialized medical care spending and patient counts by age and diagnosis



Source: Estonian Health Insurance Fund

Estonian Health Insurance Fund

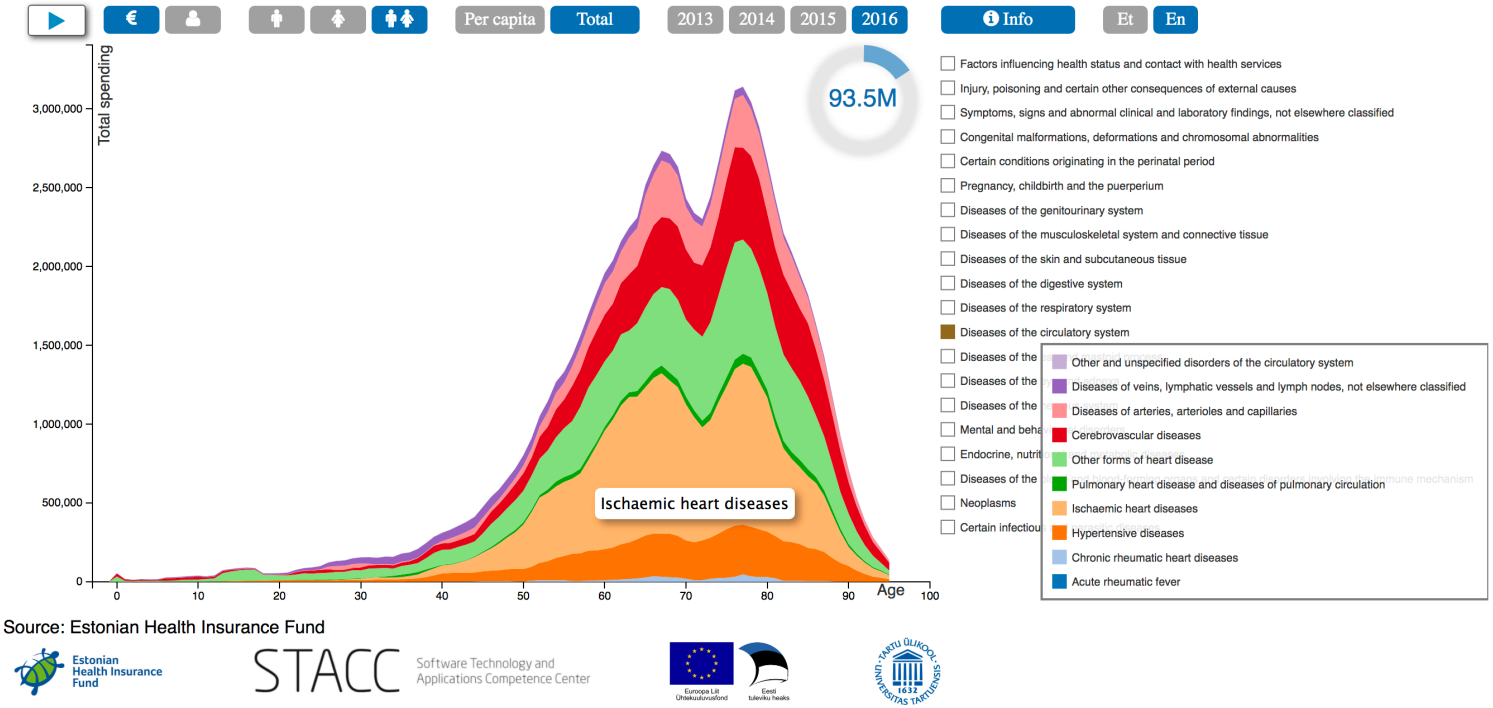
STACC Software Technology and Applications Competence Center

Euroopa Liit Ühtekuuluvusfond

Eesti tuleviku heaks

Tartu Ülikool Universitas Tartuensis 1632

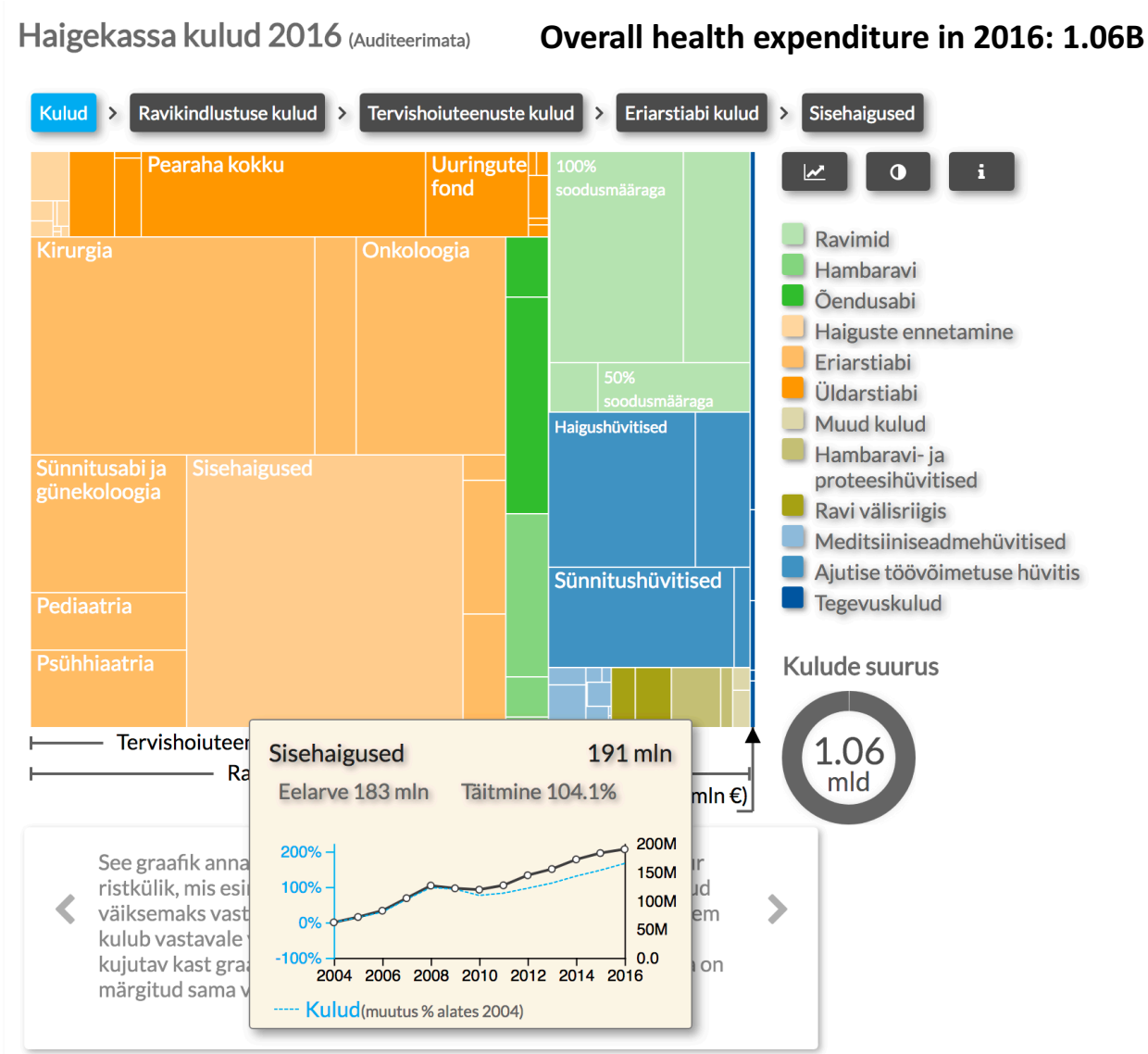
Overview of the Estonian specialized medical care spending and patient counts by age and diagnosis



Source: Estonian Health Insurance Fund

Estonian Health Insurance Fund

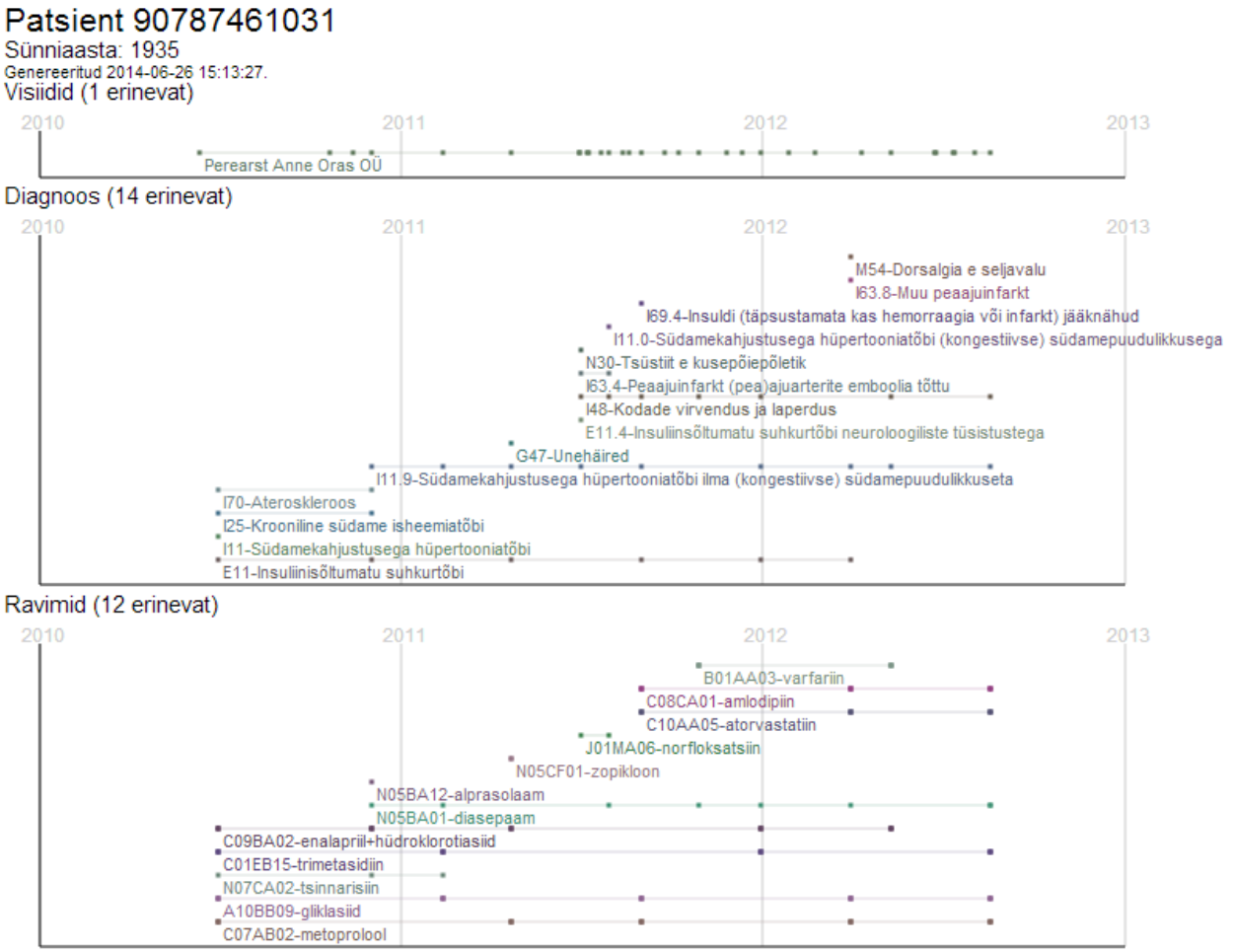
STACC Software Technology and Applications Competence Center


Haigekassa kulud 2016 (Auditeerimata)
Overall health expenditure in 2016: 1.06B
Kulud
Ravikindlustuse kulud
Tervishoiuteenuste kulud
Eriarstiabi kulud
Sisehaigused
Pearaha kokku
Uuringute fond
100% soodusmääraga
Kirurgia
Onkoloogia
50% soodusmääraga
Haigushüvitised
Sünnitusabi ja günekoloogia
Sisehaigused
Sünnitushüvitised
Pediaatria
Psühhiaatria
Ravimid
Hambaravi
Õendusabi
Haiguste ennetamine
Eriarstiabi
Üldarstiabi
Muud kulud
Hambaravi- ja proteesihüvitised
Ravi välisriigis
Meditsiiniseadmehüvitised
Ajutise töövõimetuse hüvitis
Tegevuskulud
Kulude suurus
1.06 mld
Sisehaigused 191 mln
Eelarve 183 mln Täitmine 104.1%
200%
100%
0%
-100%
200M
150M
100M
50M
0.0
2004 2006 2008 2010 2012 2014 2016
Kulud (muutus % alates 2004)

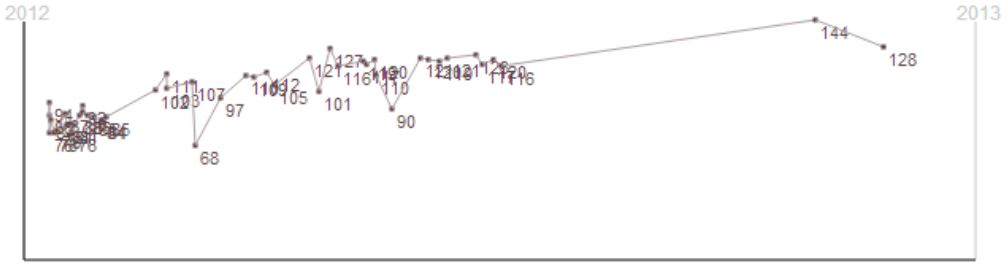
Nr of documents in E-Health DB 1.53 M individuals (06.03.2017)

Document type	Nr. of documents in E-health
Outpatient notes	15,323,163
Inpatient discharge summaries	1,643,296
Development assessment notices	34,308
Immunization side effects	6
Immunization notes	490,624
Growth notes	130,448
Ambulance Cards	340,274
Home nurse reports	5,634
Examination reports	122,945
Advisory notices	105,224
Pointers to “pictures” in PACS	2,866,152
Referrals	991,790
Referral responses	6,902,871

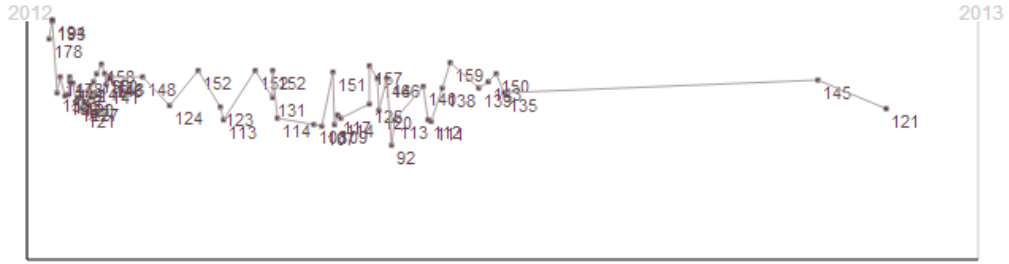
Tehnoloogia: andmete visualiseerimine



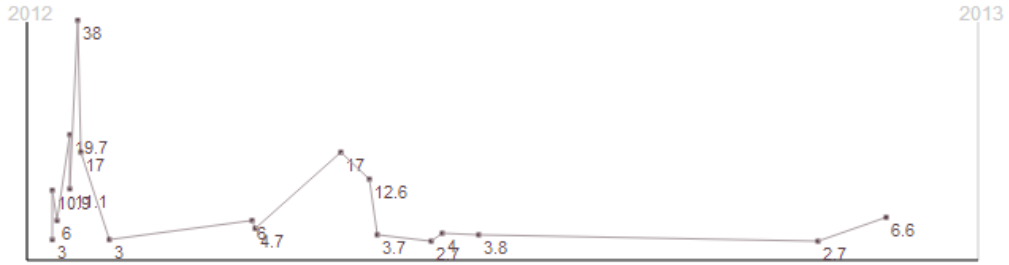
HEMOGLOBIIN



KREATIINIIN



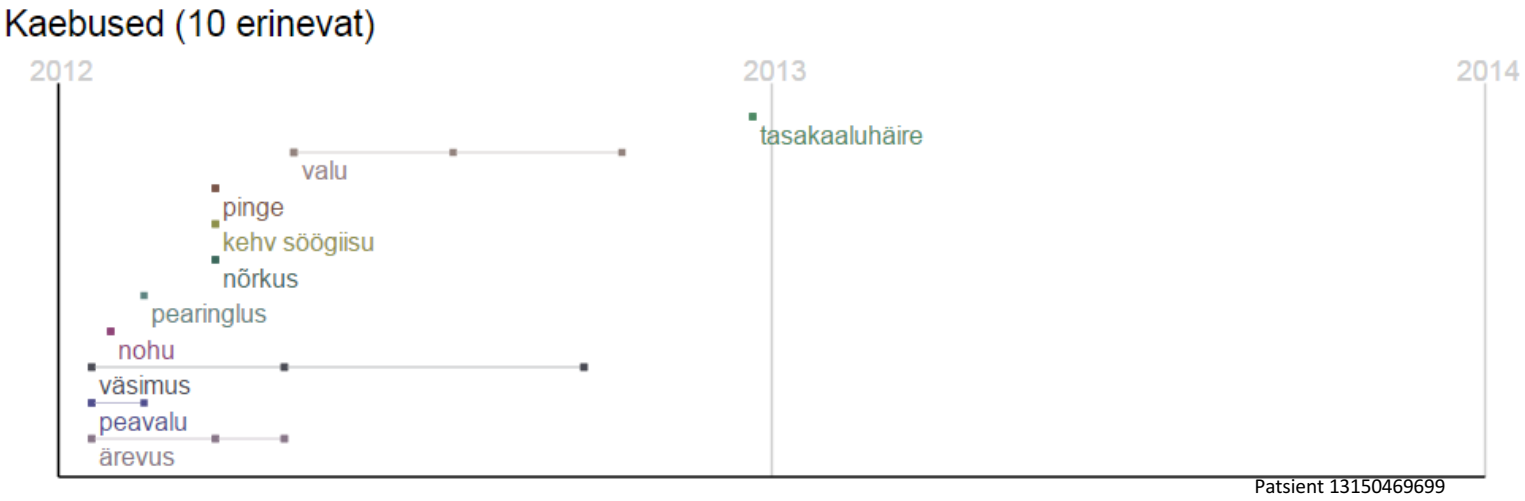
VERESUHKUR

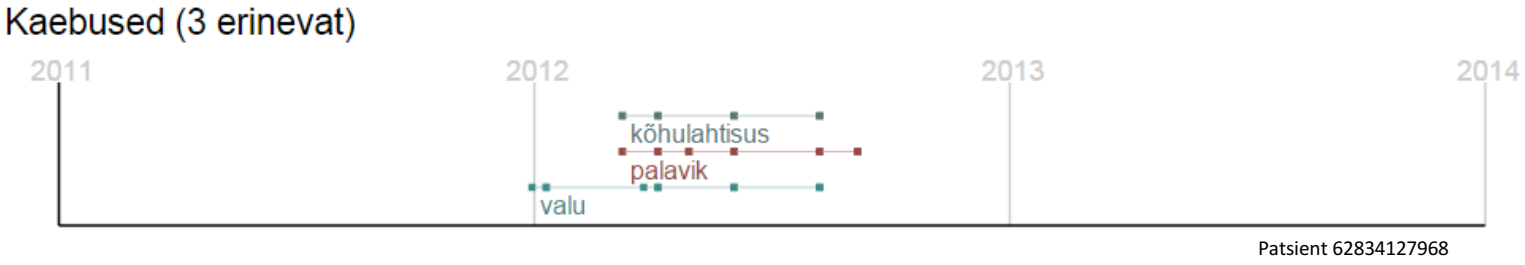


Patsient 16422115157

Tekstiosast eraldatud andmed!

Näiteid kaebuste automaatsest eraldamisest ja visualiseerimisest





Vabatekstides on veel kasulikku informatsiooni

Suunatud PA-lt , probleemiks kõhuvalu, mis ca aasta vältel olid 2 x nädalas, aga aprilli algul äge haigus seedehäiretega.

Sellest ajast kaebab iga päev, rohkem hommikul ärgates või enne und. Iiveldust, oksendust sel ajal ei ole.

Iste tavaliselt regulaarne, eile-täna iste vedel.
Anamneesis ka peavalud.

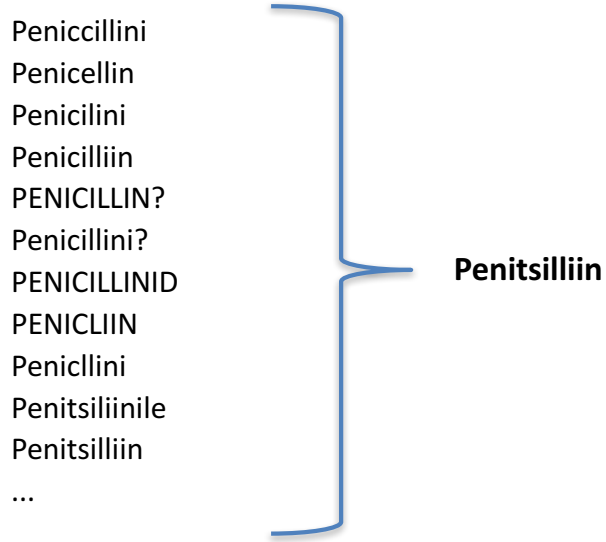
Saadetud gastroskoopiasse.

Kaebused:
- Kõhuvalu
- Seedehäire
- Iiveldus (neg)
- Oksendamine (neg)
- Iste vedel
- Peavalu

Uuringud:
- gastroskoopia

Meditsiiniterminoloogia ühtlustamise vajadus

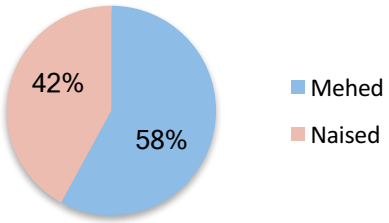
- Arstide keelekasutust iseloomistab variatiivsus
- Vabatekstiväljade kasulikkuse tõstmiseks on vaja see variatiivsus kirjeldada
- Näide epikriisi allergiablokist:

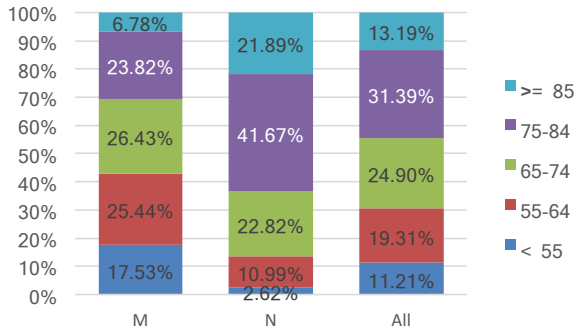


Tehnoloogia ja kompetentsi võimekuse test (võrdlus müokardi-infarktiregistriga)

E-TERVIS

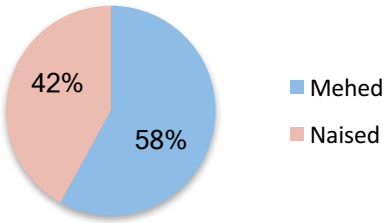
- 2847 haigusjuhtu
- 2754 erinevat patsienti

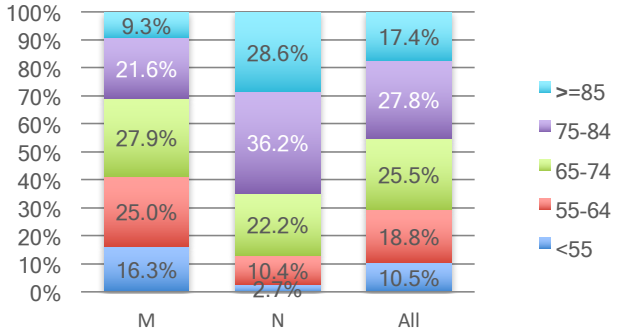




Müokardi-infarktiregister

- 2879 infarktijuhtu

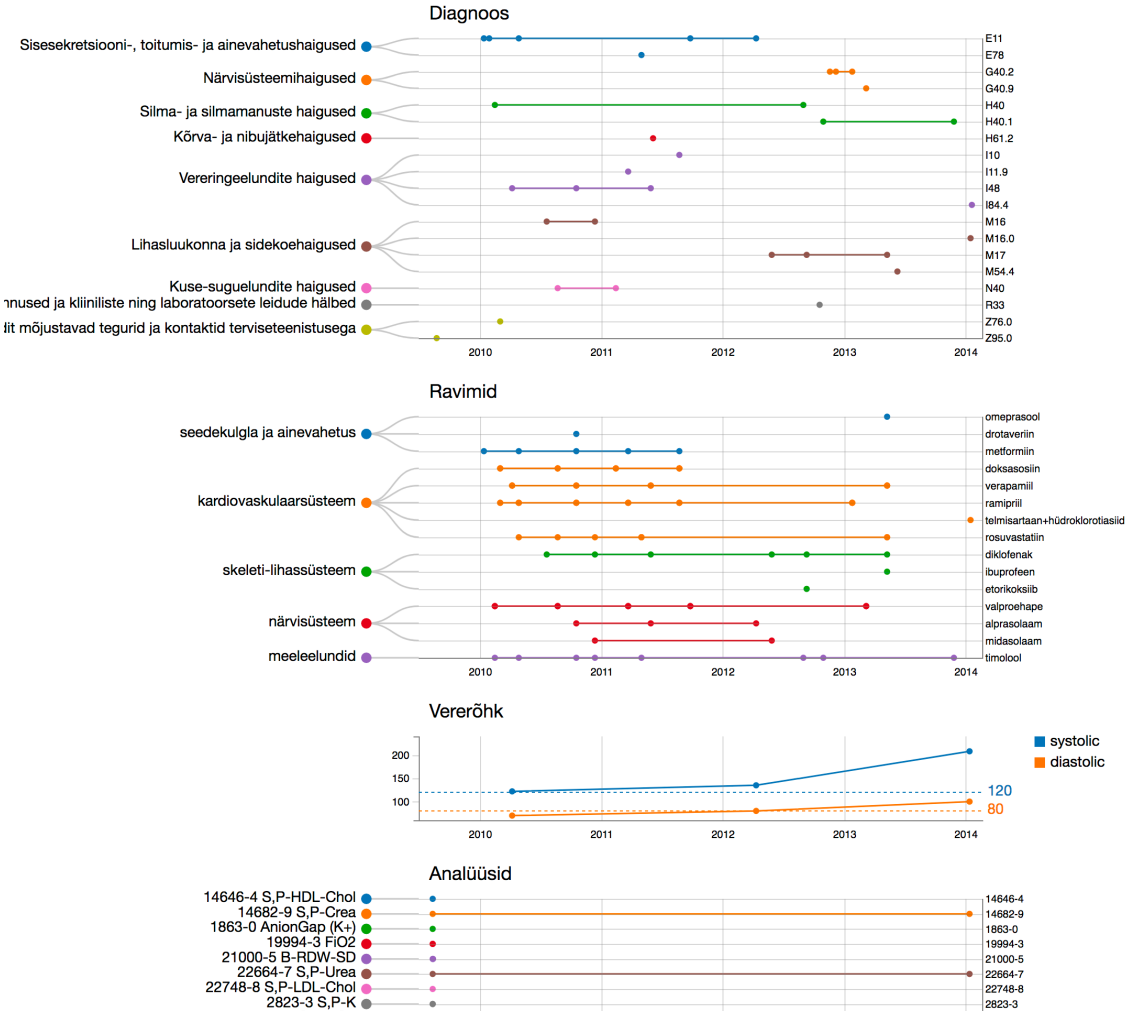




Tehnoloogia, mis suudab epikriisidest kätte saada praktiliselt kõik infarktijuhtumid!

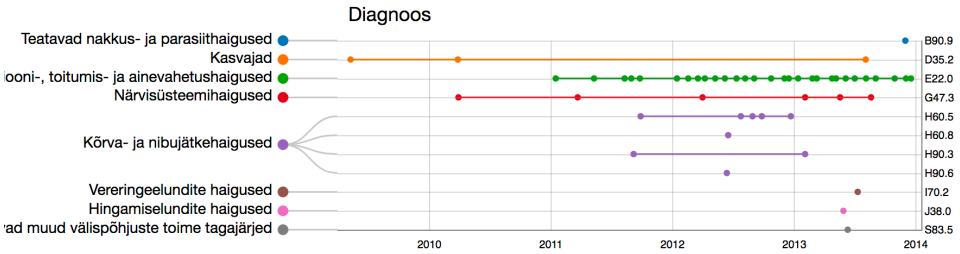
STACC: Health data analysis for >1M individuals

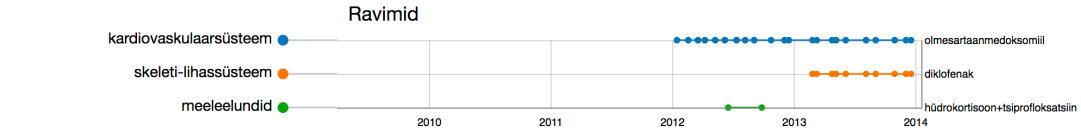
- **Text Mining** (developing and Estonian NLP)
- **Machine learning** – predicting context, semantic meaning, semantic similarities,
- **Information extraction** from unstructured data
- Disease **co-morbidities** and **trajectories**
- **Quality indicators**
- **Observational data vs registries (e.g. infarctions)**
- IMI – EMIF-Platform: EGCUT data and OMOP
- Linking Genomics data to EHR and the needed information architectures

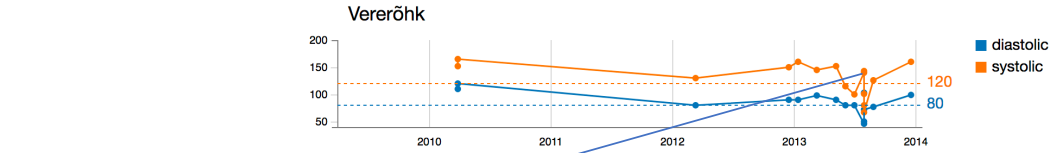

Diagnoos
Sisesekretsiooni-, toitumis- ja ainevahetushaigused
Närvisüsteemihaigused
Silma- ja silmamanuste haigused
Kõrva- ja nibujätkehaigused
Vereringeelundite haigused
Lihasluukonna ja sidekoehaigused
Kuse-suguelundite haigused
Ravimid
seedekulgla ja ainevahetus
kardiovaskulaarsüsteem
skeleti-lihassüsteem
närvisüsteem
meeleelundid
omeprasool
drotaveriin
metformiin
doksasosiin
verapamiil
ramipriil
telmisartaan+hüdroklorotiasiid
rosuvastatiin
diklofenak
ibuprofeen
etorikoksiib
valproehape
alprasolaam
midasolaam
timolool
Vererõhk
systolic
diastolic
120
80
Analüüsid
14646-4 S,P-HDL-Chol
14682-9 S,P-Crea
1863-0 AnionGap (K+)
19994-3 FiO2
21000-5 B-RDW-SD
22664-7 S,P-Urea
22748-8 S,P-LDL-Chol
2823-3 S,P-K
2010
2011
2012
2013
2014

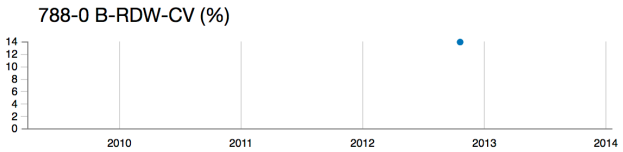
70905119764 M92

Soovitused puuduvad









Mida saab teaduslikus mõttes tuvastada:

- Statistiline juhtumite kokku loendamine (koos **tekstikaevega**)
- Haiguste **trajektoorid** – ajalised sõltuvused; ravi trajektoorid (näide)
- Haiguste **koos-esinemine** (co-morbidity) ja välistamine (riskid)
- Kvaliteedi-indikaatorid
- Kõrvalmõjud, kaebused, tulemused …

KOK (COPD) näitel “protsessid” (MSc töö)

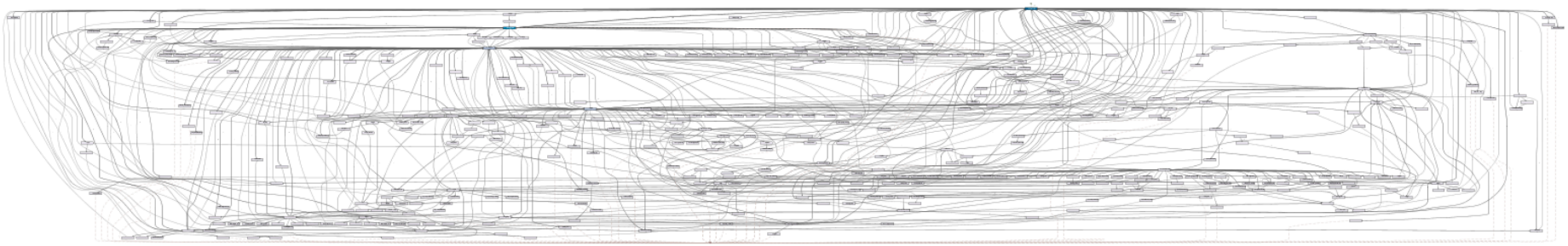


Figure 15: Figure showing the process model used in cluster 3 of diagnosis J44 with all events and edges displayed. This is what is called a “spaghetti model” in the field of process modeling.

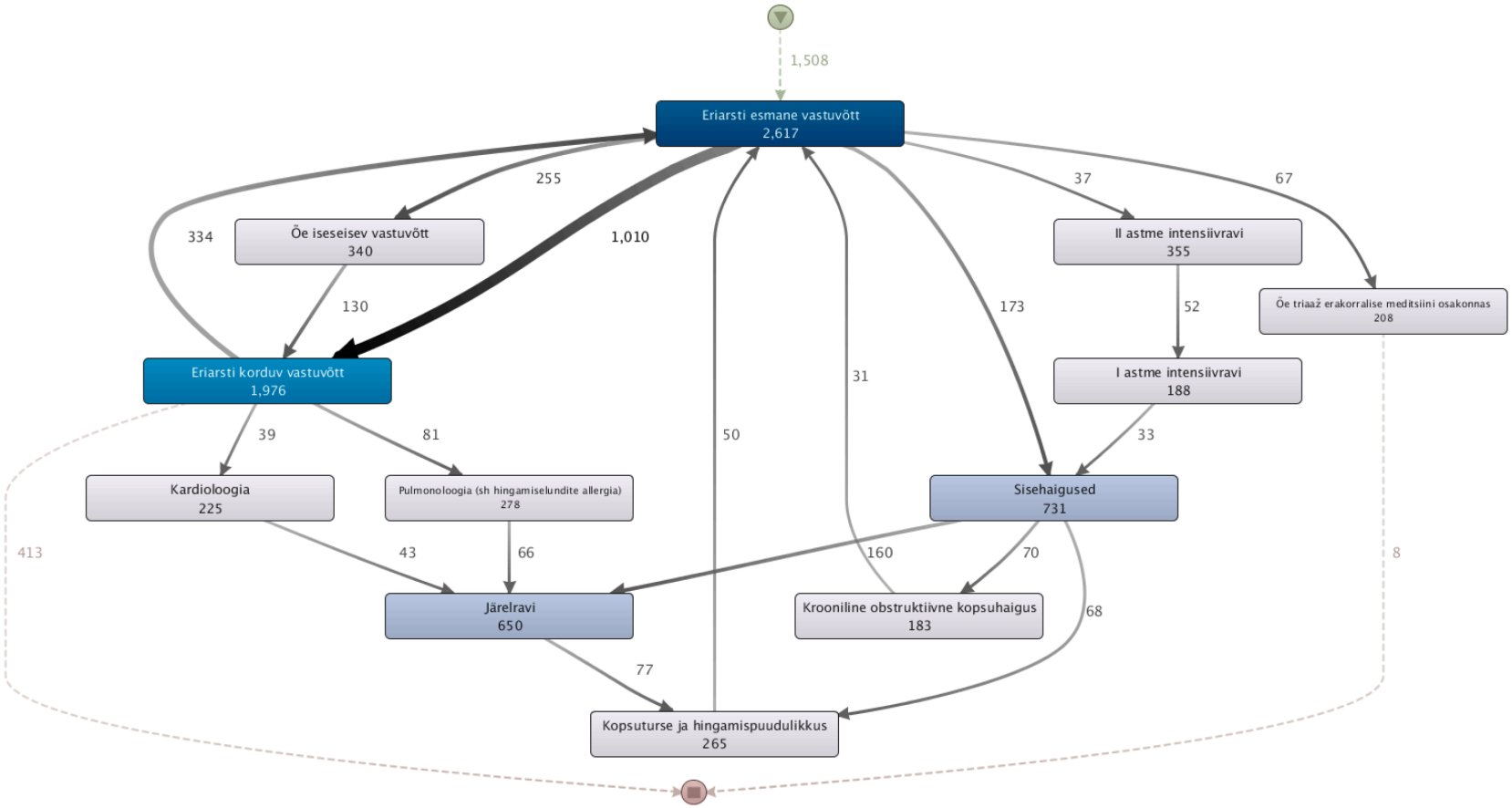


Figure 16: A process model found using Disco. Nodes limited to 2.6% edges to just the most significant ones.

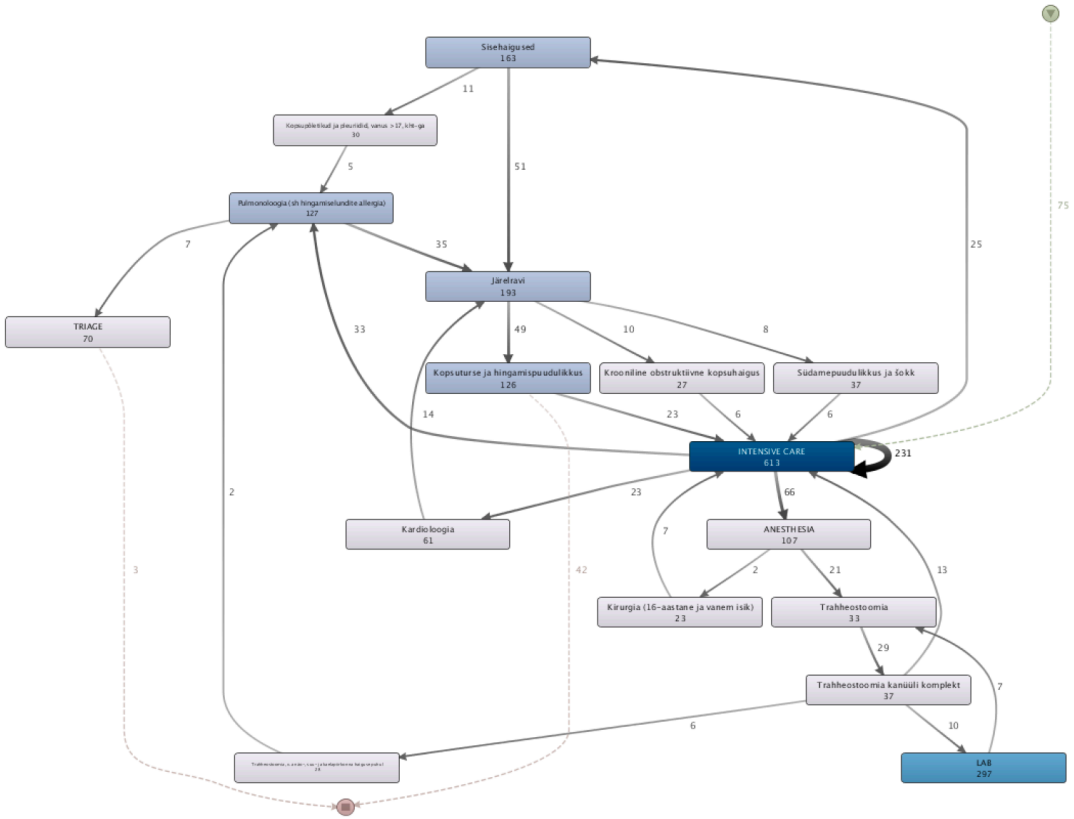
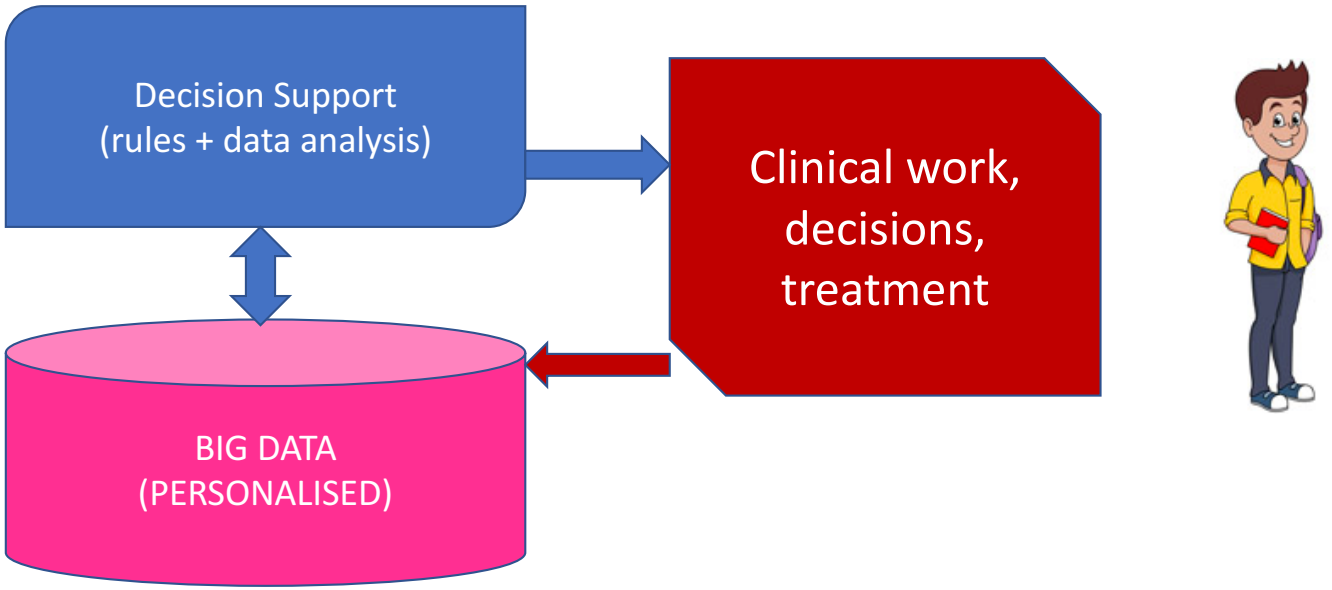
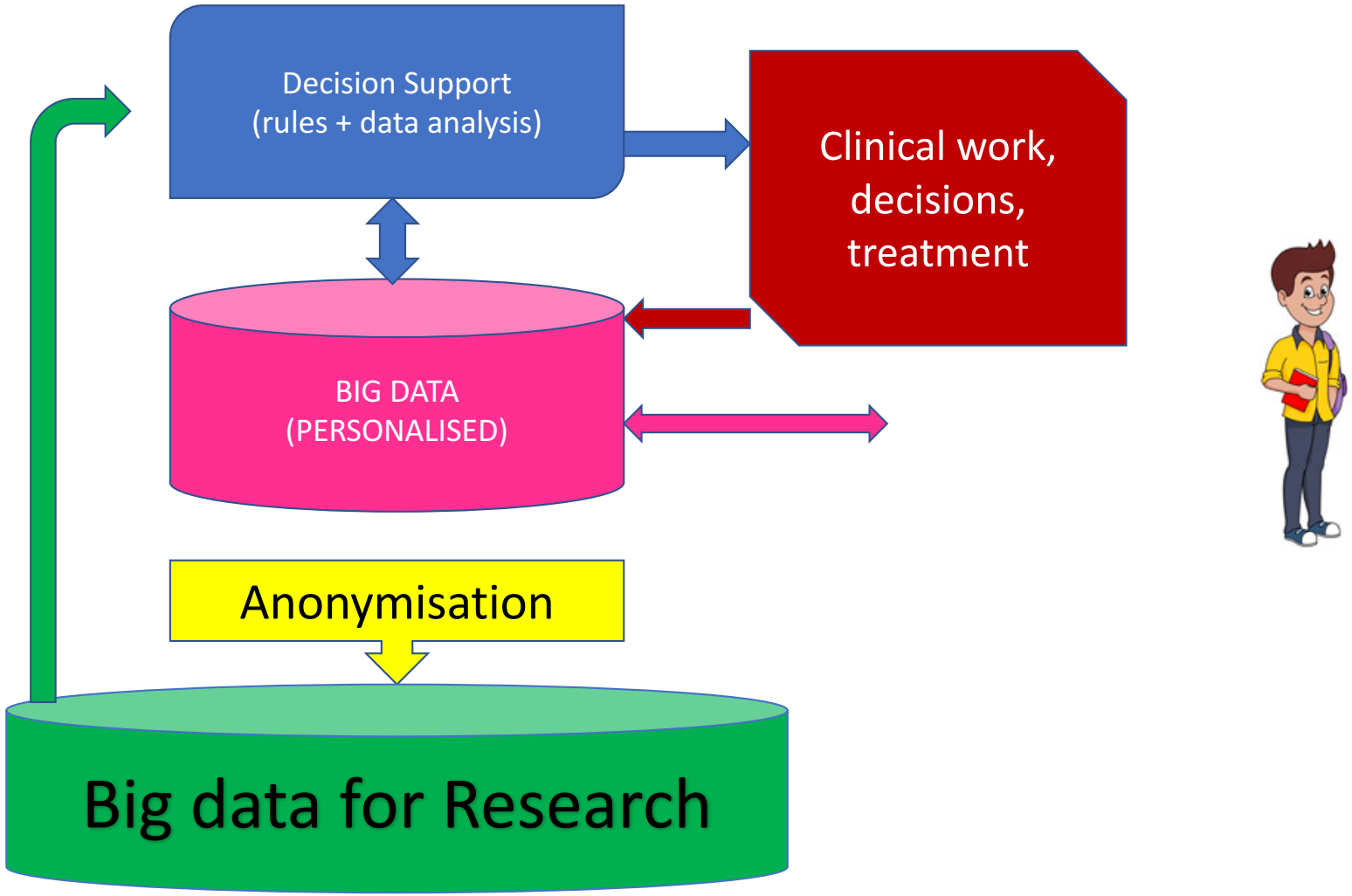


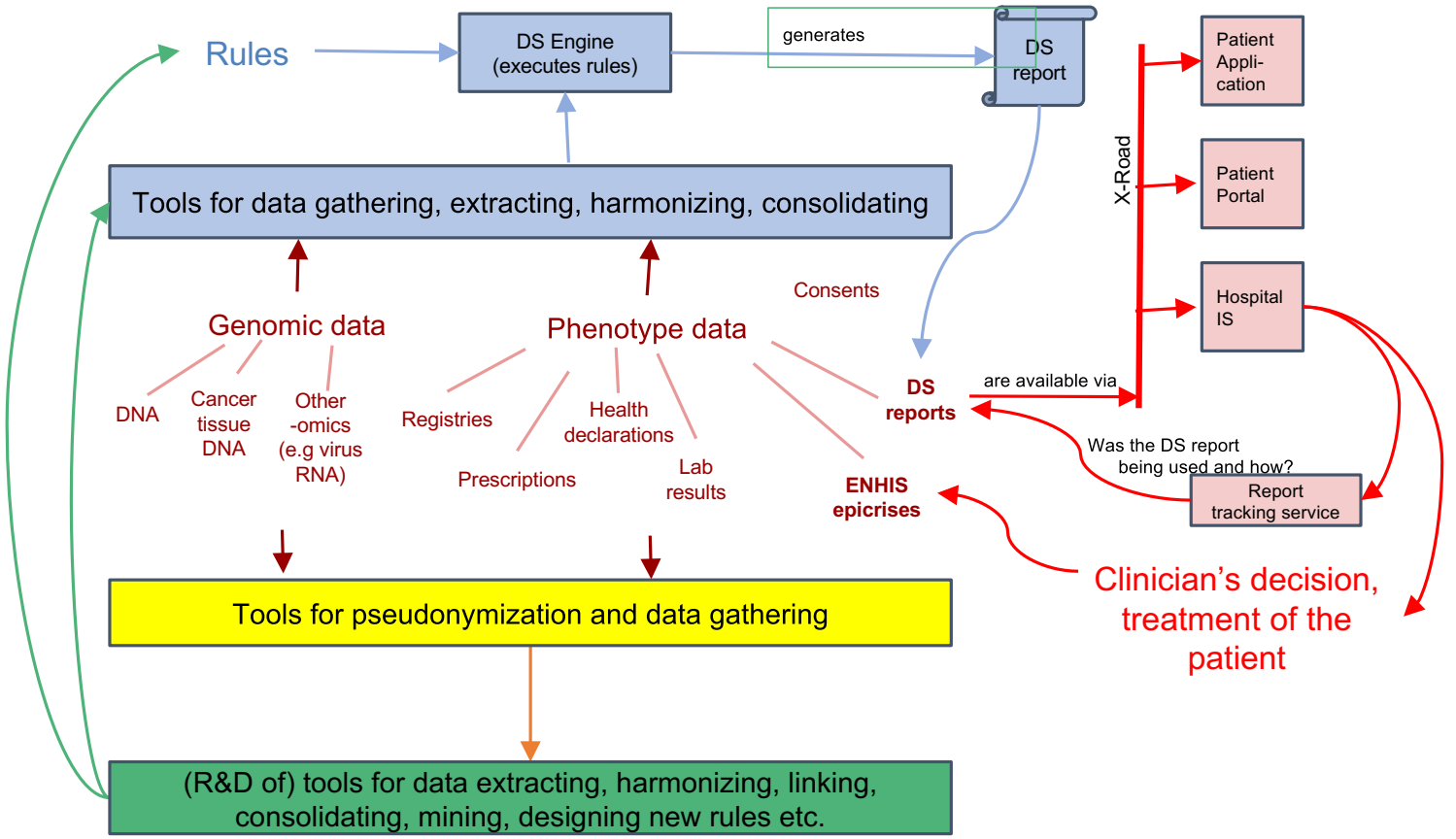
Figure 18: Clinical pathway for cluster 2 of diagnosis J44 on logs filtered with manually defined rules.

The pathways for this cluster are noticeably more complex, with more complications related to respiratory organs and heart. Interestingly, cancer seems to have much less importance in this cluster.


Decision Support
(rules + data analysis)
Clinical work,
decisions,
treatment
BIG DATA
(PERSONALISED)

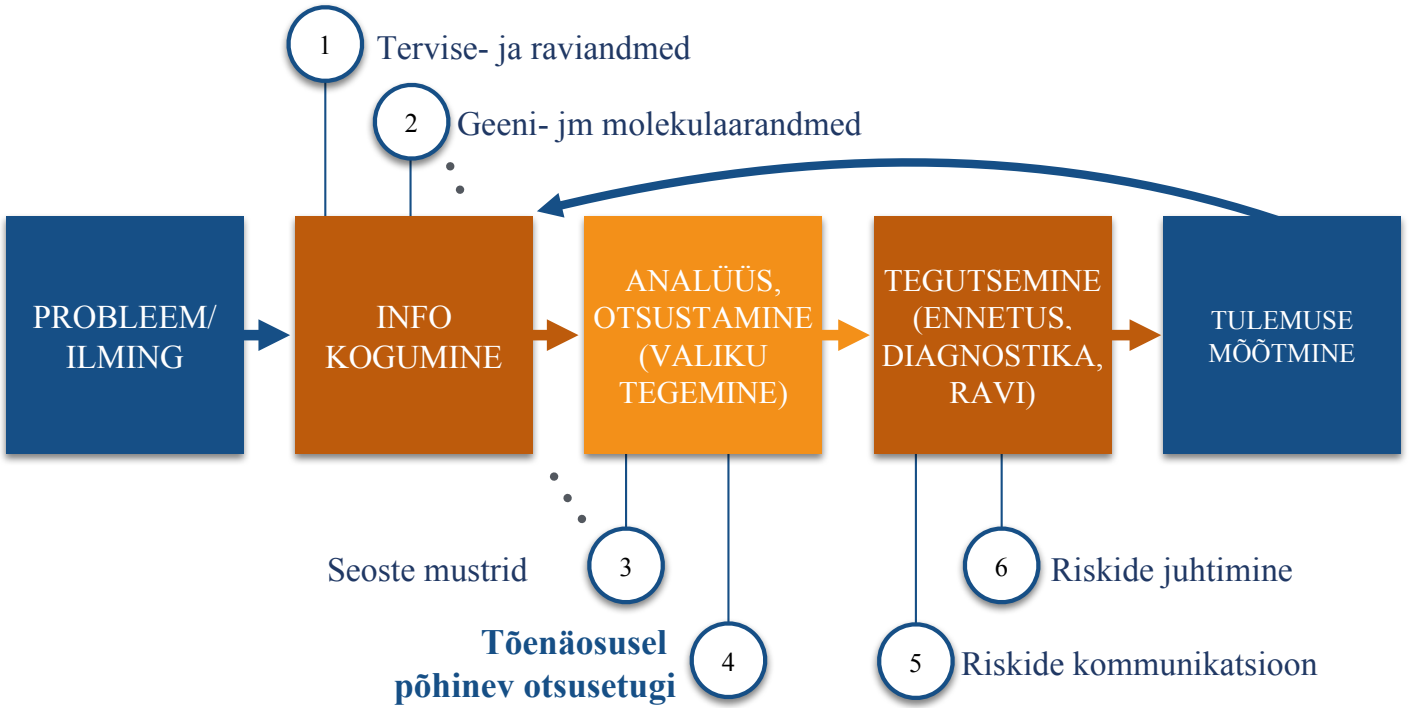

Decision Support
(rules + data analysis)
Clinical work,
decisions,
treatment
BIG DATA
(PERSONALISED)
Anonymisation
Big data for Research

What tools & data in each process



Personaalmeditsiin = 1 + 2 + 3 + 4 + 5 + 6

Personalised medicine



UNIVERSITY OF TARTU

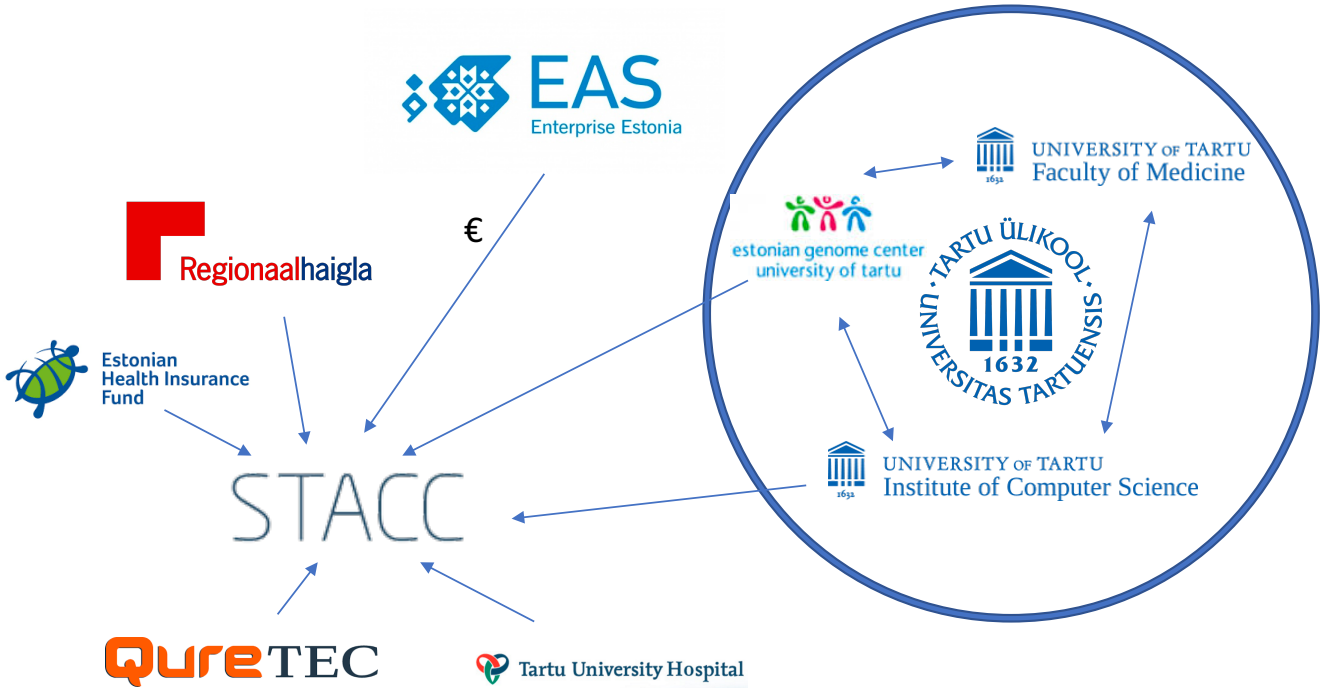
STACC Software Technology and Applications Competence Center

estonian genome center

BBMRI-ERIC
Biobanking and
BioMolecular resources
Research Infrastructure

eatris

Organisations:



STACC Software Technology and Applications Competence Center