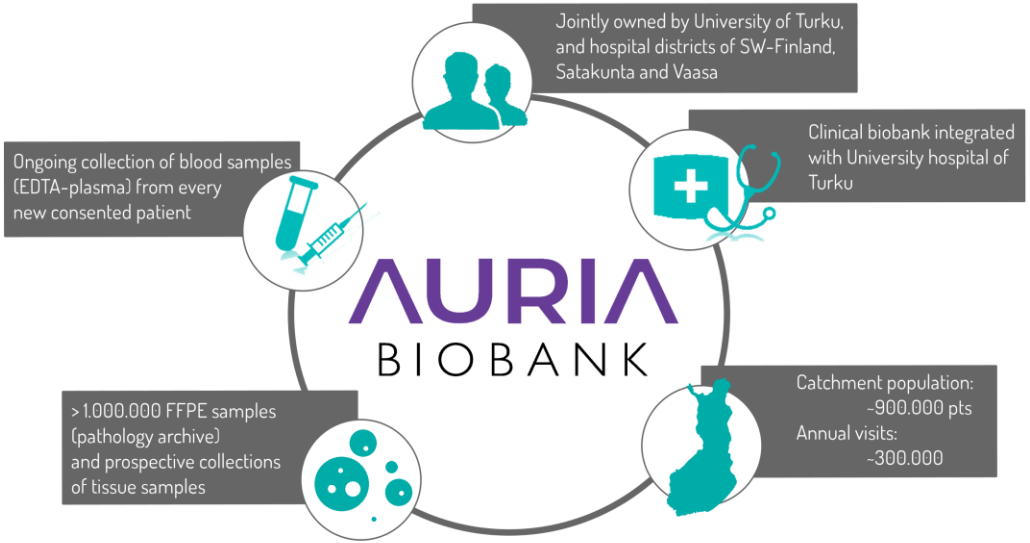
AURIA BIOBANK

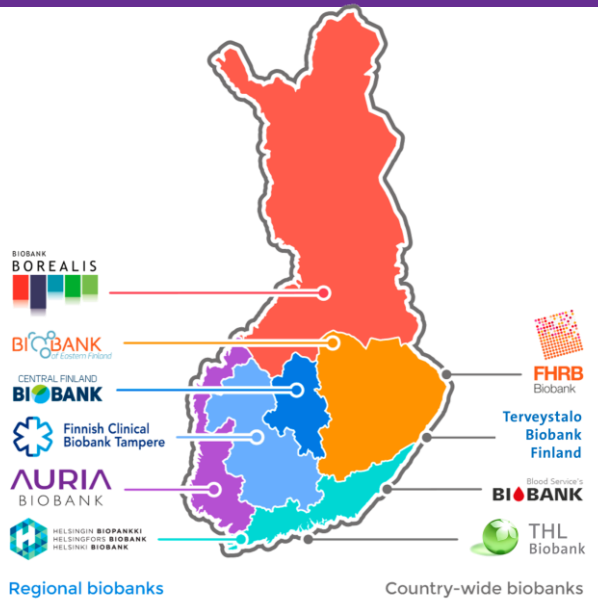
Health in 1s and 0s

Managing data and
data anonymisation

Mikko Tukiainen, PhD

Auria Biobank





Biobank act (1.9.2013)

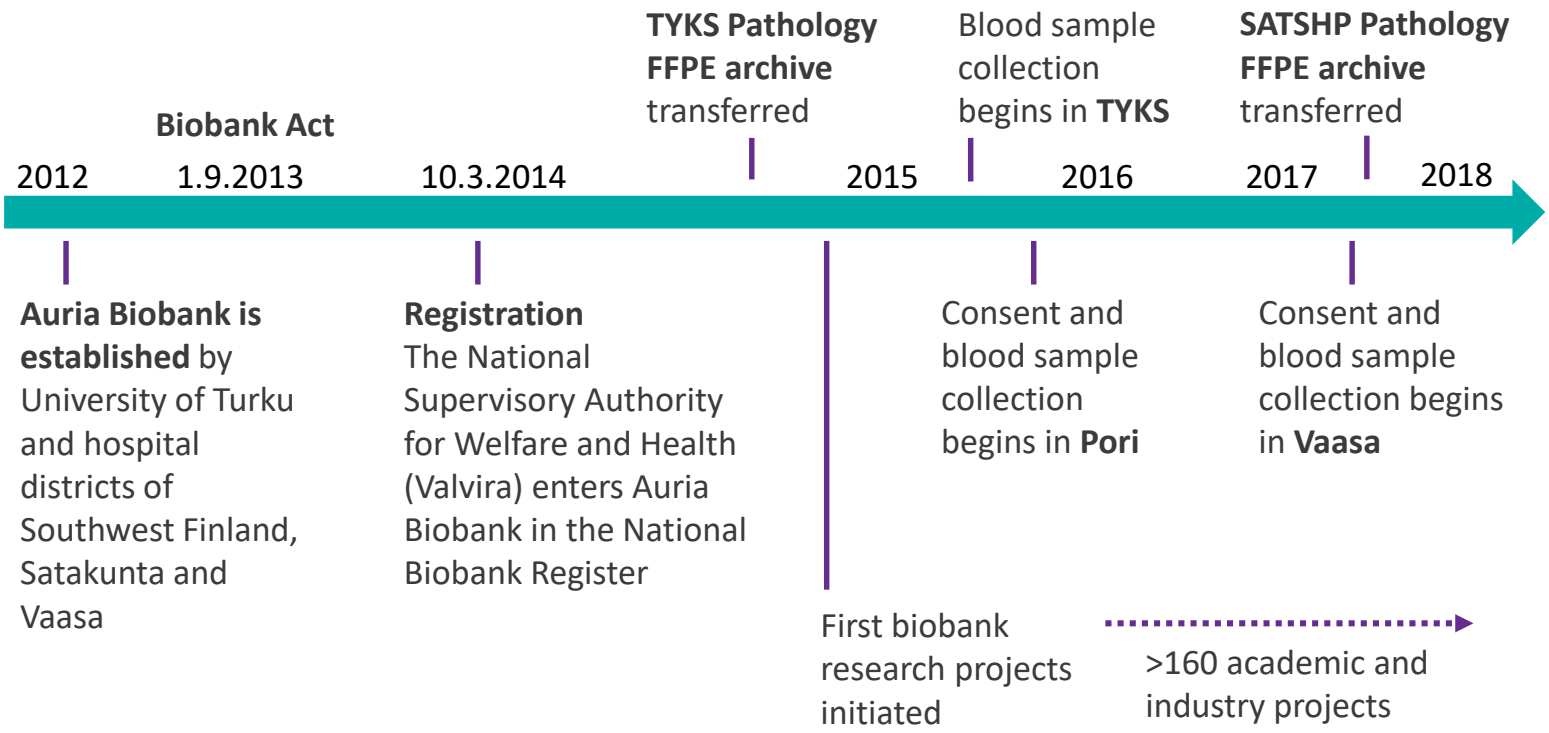
⇒ *Regulation* by national authorities (professionalism, quality standards)

⇒ *Protection* of donors' rights (informed consent, privacy protection, sample pseudonymisation, right to know in which projects samples have been used)

⇒ *Promotion* of research and R&D (broad consent, permission to link samples with information from hospital databases and national registries, all raw data from projects should return to biobank for future biobank research purposes)

AURIA BIOBANK

Timeline



Sample collection

- Samples from consented donors (see https://bit.ly/2zFIud4)
  - Taken during regular medical procedures
  - Opt-out possible anytime
- Goal is to get a blood sample from all consented donors (TYKSlab, SataDiag, Vaasa)
- Tissue samples in collaboration with surgery and pathology units

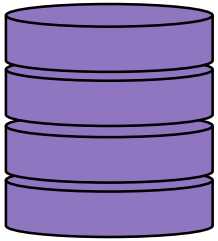
Information about the donor's identity is stored separately from the samples and related data

- Before moving the samples to biobank all personal information related to them are pseudonymised

AURIA BIOBANK

Real world clinical data linked to samples

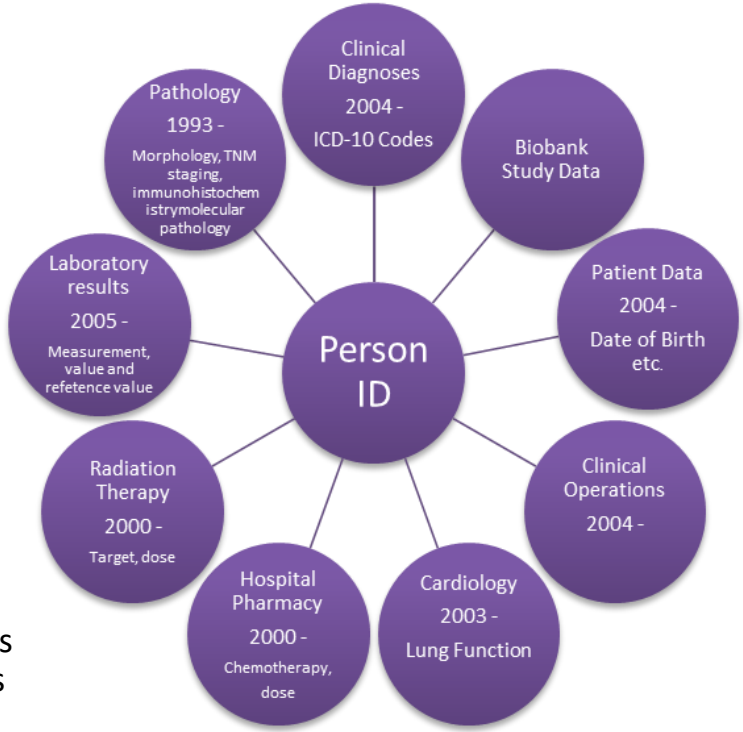
Hospital EHRs (in Data lake)



Clinical data

ICD-10
SNOMED
ATC
TNM

Biobank studies
National registers
Self-monitoring devices

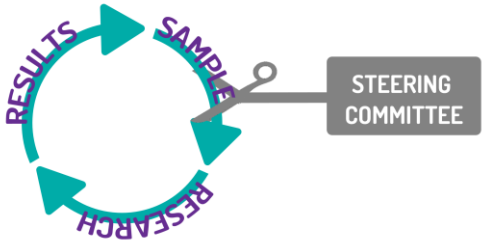


Provides the possibility to combine hospital databases (typically from year 2004 on) with biobank specimens

Longitudinal information on the course of disease, operations, response to treatments, outcome etc.

Data security

Valvira
Sosiaali- ja terveysalan lupa- ja valvontavirasto



Biobank is regulated by:
- Biobank Act
- Personal Data Act
- Act on the Openness of Government Activities
- Medical Research Act

Requests for samples and data are processed by Scientific Steering Committee
- in charge of the scientific evaluation of research projects involving the biobank

The biobank's own data registers are processed in accordance with the requirements concerning the data security of confidential information.

AURIA BIOBANK

Anonymous vs. pseudonymous

Anonymisation

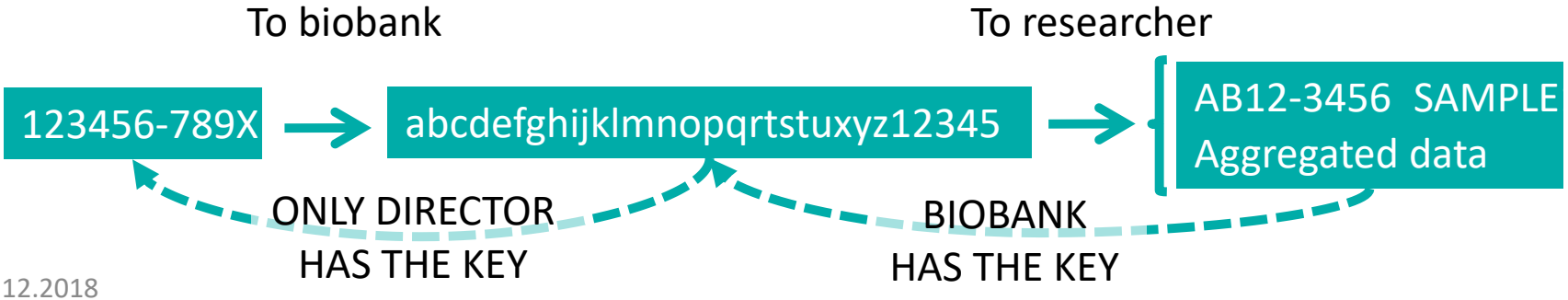
⇒ Destroys any way of identifying the data subject. It is irreversible

Pseudonymisation

⇒ Substitutes the identity of the data subject, meaning you need additional information to re-identify the data subject. It is reversible

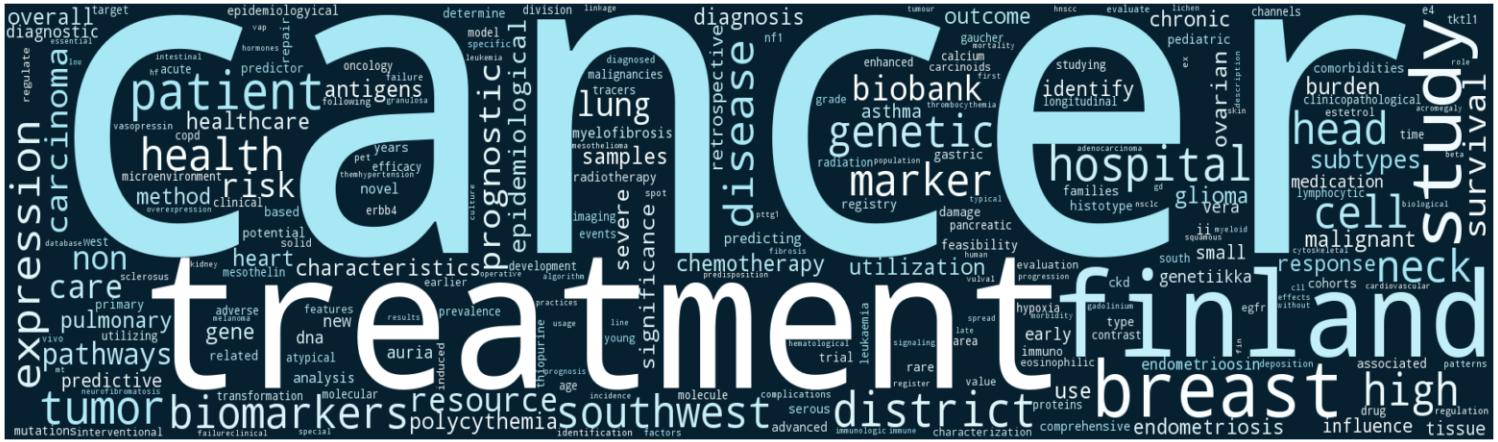
Biobank act

⇒ Donors' rights includes the right to know in which projects their samples have been used

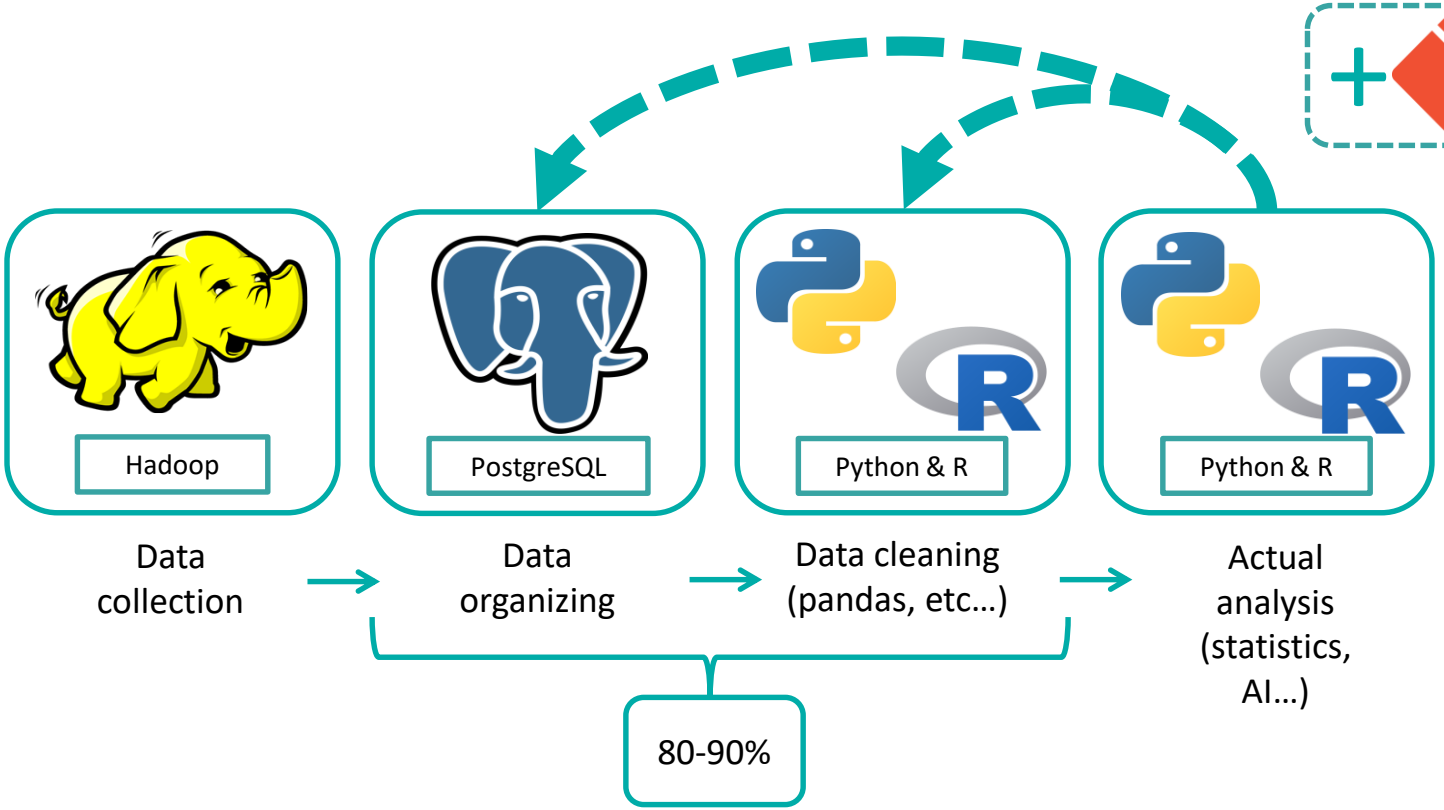


AURIA BIOBANK

Research projects


cancer
treatment
finland
study
patient
health
tumor
biomarkers
breast
genetic
hospital
marker
biobank
lung
disease
prognostic
epidemiological
expression
carcinoma
survival
cell
head
neck
high
southwest
district
resource
polycythemia
chemotherapy
utilization
characteristics
risk
method
healthcare
antigens
pathways
predictive
pulmonary
gene
heart
non
care
diagnosis
outcome
chronic
ovarian
glioma
malignant
subtypes
medication
burden
genetiikka
endometriosis
influence
tissue
overall
diagnostic
epidemiologyical
determine
division
severe
significance
retrospective
identify
studying
pediatric
comorbidities
clinicopathological
estetrol
lymphocytic
response
small
feasibility
pancreatic
damage
registry
families
histotype
vera
hypoxia
early
type
contrast
immuno
eosinophilic
endometrioosin
associated
cohorts
ckd
use
proteins
rare
trial
leukaemia
area
auria
dna
new
atypical
analysis
molecular
transformation
mutations
interventional
advanced
serous
molecule
complications
characterization
comprehensive
drug
regulation
patterns
deposition
egfr
gadolinium
hematological
signaling
register
value
age
incidence
prognosis
usage
practices
prevalence
results
features
adverse
melanoma
related
utilizing
vivo
primary
sclerosus
kidney
mesothelin
development
operative
algorithm
earlier
induced
thiopurine
line
young
spread
late
fin
cii
without
effects
vulval
human
progression
evaluation
predisposition
fibrosis
predicting
pttg1
nf1
tumour
hnscc
evaluate
lichen
channels
e4
tktl1
role
description
skin
acromegaly
beta
time
biological
cytoskeletal
squamous
myeloid
adenocarcinoma
nsclc
typical
gd
first
thrombocythemia
grade
radiation
population
gastric
calcium
carcinoids
enhanced
mortality
gaucher
longitudinal
asthma
tracers
diagnosed
malignancies
myelofibrosis
mesothelioma
samples
radiotherapy
spot
imaging
events
culture
model
specific
leukemia
linkage
division
determine
epidemiologyical
vap
repair
hormones
intestinal
acute
predictor
oncology
failure
granulosa
following
vasopressin
copd
microenvironment
overexpression
clinical
based
potential
solid
years
pet
efficacy
themhypertension
novel
erbb4
west
database
non
care
sclerosus
primary
pulmonary
utilizing
vivo
pathways
predictive
neurofibromatosis
transformation
molecular
mutations
interventional
failureclinical
special
identification
factors
immunologic
immune
characterization
comprehensive
endometriosis
influence
tissue

Data analysis pipeline



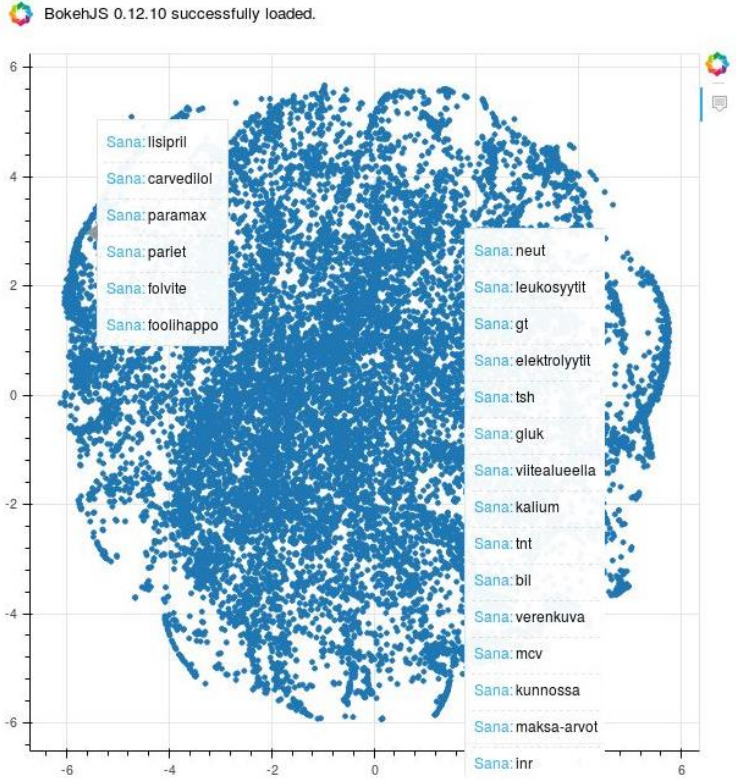
EXAMPLE: Text mining

- Vast amount of the interesting data is not in a structural form
  - Cancer TNM-classificatios
  - Some mutation statuses
  - Smoking status

→ Doctors' dictations

- Regular expressions (regex)
  - See e.g. https://bit.ly/2BPz9AP
  - Algorithms for determining the smoking status (https://bit.ly/2IjIzJZ)

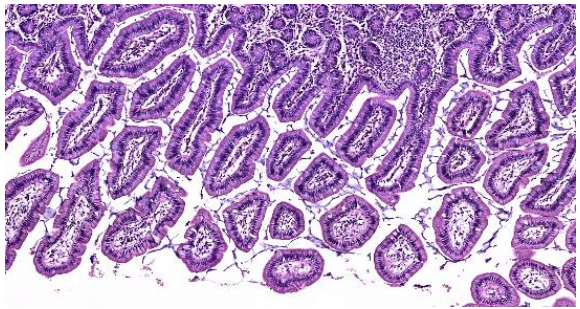
EXAMPLE2: Algorithms for text mining




```
In [173]: foo.most_similar('rinta')

Out[173]: [('munuainen', 0.806662917137146),
           ('ovario', 0.7986398935317993),
           ('lohko', 0.7956081628799438),
           ('nilkka', 0.7721655368804932),
           ('nänni', 0.767995297908783),
           ('pohje', 0.7642505764961243),
           ('alaraaja', 0.754327654838562),
           ('lonkka', 0.7512207627296448),
           ('kammio', 0.746738076210022),
           ('jalka', 0.7437646389007568)]

In [157]: foo.most_similar('kuumetta')

Out[157]: [('lämpöilyä', 0.9187069535255432),
           ('kuumeilua', 0.9126585721969604),
           ('yskää', 0.8458117842674255),
           ('flunssaa', 0.837439239025116),
           ('hengenahdistusta', 0.8253785967826843),
           ('vastetta', 0.8127169013023376),
           ('rintakipua', 0.8104287981987),
           ('laihtumista', 0.8087208271026611),
           ('rintakipuja', 0.8083760142326355),
           ('nuhaa', 0.8018956780433655)]

In [150]: foo.most_similar('kaarinan')

Out[150]: [('naantalin', 0.9719816446304321),
           ('laitilan', 0.9714298844337463),
           ('paraisten', 0.9669349789619446),
           ('maskun', 0.965885579586029),
           ('someron', 0.9613572955131531),
           ('kemiönsaaren', 0.9591622352600098),
           ('pöytyän', 0.9563663005828857),
           ('raision', 0.9524632096290588),
           ('porin', 0.9510113596916199),
           ('mynämäen', 0.949148416519165)]
```

AURIA BIOBANK

Benefits for the sample donors



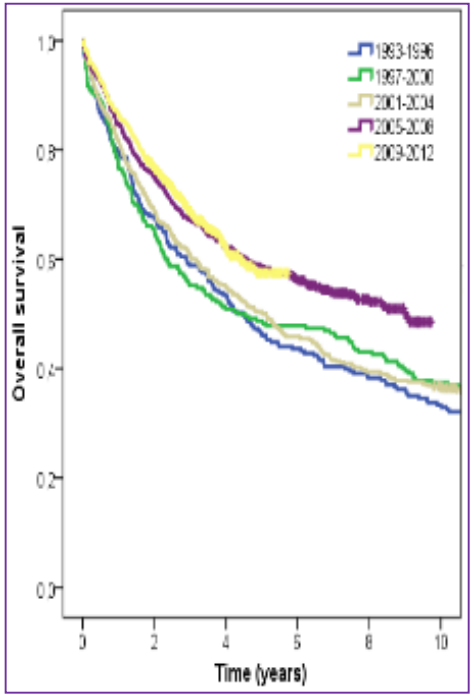
- Possibility to support research in the area of donor's illness
- Research information can be used in donors' own treatment => **Personalized medicine**
- The outcome of the research can help other people

Biobank supporting diagnostics and treatment

Major improvement in survival of CRC patients was observed in 2004, at the time of

- centralization of rectal cancer surgery
- introduction of multidisciplinary teams
- higher number of lymph nodes examined
- implementation of preoperative radiotherapy in rectal cancer
- the use of adjuvant chemotherapy in stage III CRC became also slightly more frequent

The clinical presentation of CRC has remained essentially the same between 2001 and 2012.



The observed changes have resulted in improved survival in CRC and a marked decrease of non-operable rectal cancer.

See: "Trends in presentation, treatment and survival of 1777 patients with colorectal cancer over a decade: a Biobank study". Heervä E et al. Acta Oncol. 2017 Dec 23:1-8.