Scalable Storage and Processing of Privacy-Sensitive Data: More Than Just a Dream?

Assoc. Prof. RNDr. Petr Holub, Ph.D.
IT & Data Protection Manager @ BBMRI-ERIC
CIO of CS IT

TERENA Networking Conference, Prague, 2016–06–15
Medical Research and Biobanking

- Medical research
  - aims at improving treatments, drugs, or public health strategies,
  - employs medicine, natural sciences, computer science, sociology, etc.,
  - is subject to various regulatory frameworks.
Medical Research and Biobanking

- Citizen (Research Participant, donor, patient)
- Clinician
- Industry (pharma, ...)
- Biobank
- Researcher (academic, industrial)
- Healthcare registries
- Public health policy maker
- Healthcare policies
- Cryo

- Provides sample/data/expertise/services
- Returns data
- Eligible data
- Returns data/samples
- Provides samples/data
- Retrieves data/samples
- Clinician
- Treatments
- Drugs
- Research results commercialization
- Clinical trials
- Commercialization

Holub P. • Scalable Privacy-Sensitive Data Processing • TNC2016, Prague
Human Data in Medical Research

- Citizen (research participant)
- Samples
- Social networks
- Information from wearables
- Healthcare
- Health-related registries
- Civic
- Environmental data
- Non-human biological data
- Exposure data
- Health and lifestyle data
- Data from samples, metadata
- Data given directly, metadata
- Lifestyle data
- Biobanks (medical research data depots in general)
Human Data in Medical Research

- Types of privacy-sensitive data (GDPR-based)
  - Anonymized, (de-identified)
  - Personal data:
    - Directly identifying
    - Pseudonymized
  - Privacy-sensitive data used in research
    - Minimize risk of revealing persons
    - Pseudonymized or de-identified/anonymized data
  - Privacy by design
Human Data in Medical Research

Expected data volumes and needed processing capacity in 2025:

<table>
<thead>
<tr>
<th>Data Phase</th>
<th>Astronomy</th>
<th>Twitter</th>
<th>YouTube</th>
<th>Genomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquisition</td>
<td>25 zetta-bytes/year</td>
<td>0.5–15 billion tweets/year</td>
<td>500–900 million hours/year</td>
<td>1 zetta-bases/year</td>
</tr>
<tr>
<td>Storage</td>
<td>1 EB/year</td>
<td>1–17 PB/year</td>
<td>1–2 EB/year</td>
<td>2–40 EB/year</td>
</tr>
<tr>
<td>Analysis</td>
<td>In situ data reduction</td>
<td>Topic and sentiment mining</td>
<td>Limited requirements</td>
<td>Heterogeneous data and analysis</td>
</tr>
<tr>
<td></td>
<td>Real-time processing</td>
<td>Metadata analysis</td>
<td></td>
<td>Variant calling, ~2 trillion central processing unit (CPU) hours</td>
</tr>
<tr>
<td></td>
<td>Massive volumes</td>
<td></td>
<td></td>
<td>All-pairs genome alignments, ~10,000 trillion CPU hours</td>
</tr>
<tr>
<td>Distribution</td>
<td>Dedicated lines from antennae to server (600 TB/s)</td>
<td>Small units of distribution</td>
<td>Major component of modern user’s bandwidth (10 MB/s)</td>
<td>Many small (10 MB/s) and fewer massive (10 TB/s) data movement</td>
</tr>
</tbody>
</table>

do:10.1371/journal.pbio.1002195.t001

Other big data in medicine – also as time series:
- omics data: proteomics, metabolomics, ...
- imaging data
- lifestyle and clinical data
Is it effective to store all the data?

- depends on purpose...
- human ability to generate data always exceeds our capacity to store and curate the data:
  - >1 PB for 1 patient?

One way to store genomics data...

... do analysis only once needed.
Some challenges of medical research

- Low hanging fruit has been mostly collected.
- Development of new drugs/therapies is increasingly more difficult.

Published Genome-Wide Associations through 12/2013
Published GWA at \( p \leq 5 \times 10^{-8} \) for 17 trait categories

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/
Some challenges of medical research

» Severe reproducibility problems.

Fig 1. Studies reporting the prevalence of irreproducibility. Source: Begley and Ellis [6], Prinz et al. [7], Vasilevsky [8], Hartshorne and Schachner [5], and Glasziou et al. [9].

doi:10.1371/journal.pbio.1002165.g001
What is a biobank?

- Repositories of material (incl. data) for medicine and medical research.
- **Biosamples,**
- **accompanying data:** clinical, phenotypes, lifestyle, ...
- **data generated** from samples: imaging, omics, ...
- **expertise:** data interpretation and integration, ...
- **services:** sample & data hosting, analysis of samples, ...

biobanks := samples + data + expertise + services;
What is BBMRI-ERIC

- European Research Infrastructure Consortium to facilitate access to high-quality biobanks and biomolecular resources to advance (bio)medical research
  - legal entity on European level,
  - est. 3 December 2013.

- Hierarchical distributed architecture

- 516 biobanks
- estimated
  >60,000,000 samples
**Biobank Use Cases: Prostate Cancer**

- 10,000 men/y in Sweden
- 2,500 deaths yearly
- 80,000 prevalent cancers

→ 300M€ yearly healthcare direct costs
→ substantial burden on society

- today’s screening with prostate-specific antigen (PSA) is not optimal
  - wrong people get screened
  - 2 of 3 positive diagnoses incorrect
  - 30-40% of all aggressive cancers missed

→ PSA has poor specificity and sensitivity

*With courtesy of Prof. Henrik Grönberg, KI*
Biobank Use Cases: Prostate Cancer

- STHLM3 is collaboration between Stockholm County Council and BBMRI.se

  - Goal:
    - develop a new prostate cancer test that can replace PSA

  - Conditions:
    - identify at least as many aggressive cancers as PSA
    - significantly reduce numbers of biopsies
    - demonstrate clear health economy

  - Method:
    - combine many existing biomarkers

---

With courtesy of Prof. Henrik Grönberg, KI

STHLM3 Model $= f$

- Protein markers
- 254 Genetic markers
- Clinical data

Risk factors
- Age
- Family history
- Previous biopsy

Prostate Exam
- Prostate volume*
- DRE*

* Only measured on biopsied men
Biobank Use Cases: Prostate Cancer

**Extent of study:**
- 145,905 randomly invited to the study
  - 11,130 men were recruited to the STHLM3 training cohort
  - 113,082 men aged 50–69 years were invited to participate in the STHLM3 validation cohort and 1,263 men were excluded because of previous prostate cancer diagnosis.

**Results:**
- reduced the number of biopsies by 32%
- reduced the number of benign biopsies by 44%
- Of the 603 high-risk cancers identified by the STHLM3 model, 124 (21%) were identified in the PSA range 1–3 ng/mL.

*With courtesy of Prof. Henrik Grönberg, KI*
Biobank Use Cases: Prostate Cancer

Results:

Custom-made plasma biomarker chip

Custom-made SNP chip

Fully automated
High-throughput

With courtesy of Prof. Henrik Grönberg, KI

Henrik Grönberg shows the two computer-chips used to analyse the protein markers and genetic markers that together comprise the STHLM 3 prostate cancer test. Credit: Stefan Zimmerman.
Biobank Use Cases: Multiple Sclerosis

- ≈ 17,000 MS cases in Sweden
- New therapy 2006, Tysabri™
  - monoclonal antibody
  - very effective, very expensive
  - serious side effects for some (polyoma viral cross-reaction)
- Population biobanks in Nordic countries
  - many patients have their blood stored in the biobanks
- Screen relevant biobank for polyoma virus
  - save big drug cost 21,000€/patient/yr
  - match a good drug to the right patients
  - save some lives, improve others

* T Olsson et al
Scalable Processing of Sensitive Data?

▶ Scalability with respect to:
  ■ volumes of stored data,
  ■ processing capacity,
  ■ institutional boundaries.

▶ Compliant with
  ■ mandatory processes related to privacy-sensitive data,
  ■ regulatory frameworks (privacy-sensitive data, medicine,...)
Scalable Processing of Sensitive Data?

- Participant
  - Broad/narrow informed consent
  - Data
  - Samples
  - Metadata repository
  - Custodian
  - Computing & storage
  - (private clouds)
  - Publish

- Researcher
  - Generates
  - Access (MTA/DTA)
  - Reviews

- Project
  - Ethical board
  - Reviews

- Trust
  - Compliant?
  - Information loss!

- Privacy
  - Aggregated or anonymized data

- Computing data storage

Holub P. • Scalable Privacy-Sensitive Data Processing • TNC2016, Prague
Scalable Processing of Sensitive Data?

- Challenges:
  - dealing with liability of data “owners” and “processors”,
  - assessing compliance of informed consent and projects,
    - multi-tenancy aware access control
  - assurance of identities of people accessing/processing data,
  - liability of storage and processing infrastructure providers,
  - designing scalable algorithms to many interesting problems.
Scalable Processing of Sensitive Data?

- Solutions are a mix of
  - theoretical approaches,
  - practical technologies,
  - organizational and regulatory frameworks.

- Needed components are
  - algorithms,
  - available and maintained tools,
  - capacities of infrastructures.
Select Computer Sci&Tech Challenges

▶ Protecting privacy-sensitive data
  ■ close relation to regulatory frameworks
    ● upcoming GDPR and its local implementations
  ■ explore compliance of privacy enhancing technologies to regulatory frameworks
    ● needs to be translated to understandable risk evaluation schemes for privacy non-experts: e.g., \( \varepsilon \)-differential privacy is not easy to understand
    ● e.g., definition of Code of Conduct assumed by GDPR Article 38
  ■ investigate privacy enhancing technologies for almost non-anonymizable data
    ● genomics
    ● but other types too: imaging data, *-omics, etc.
Select Computer Sci&Tech Challenges

- Structured data extraction
  - clinical records contain some structured data, but unstructured free text is major source of information,
  - lifestyle information mining from social network, fitness services, etc.
  - language, cultural, and domain specifics,
  - natural language processing, deep learning, ...

<table>
<thead>
<tr>
<th>Boundary detection</th>
<th>… … ]</th>
<th>[ Fx of obesity but no fx of coronary artery diseases. ]</th>
<th>[ … … ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalization</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-of-speech tagging</td>
<td>NN IN NN CC DT NN IN JJ NN NNS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow parsing</td>
<td>NP PP NP</td>
<td>NN</td>
<td>NP</td>
</tr>
<tr>
<td>Entity recognition</td>
<td>Obesity Disease or disorder</td>
<td>UMLS ID: C0028754 Status: family history Negated: no</td>
<td>Coronary artery disease Disease or disorder</td>
</tr>
<tr>
<td></td>
<td>Coronary artery Anatomy</td>
<td>UMLS ID: C0205042</td>
<td></td>
</tr>
</tbody>
</table>

Source: doi:10.1038/nrg3208
Figure 4 | Diabetes and cardiovascular disease trajectory clusters. (a) Diabetes cluster showing progression from non-insulin-dependent to insulin-dependent diabetes. Retinal disorders are key diagnoses marking progression to worse conditions. (b) Cardiovascular cluster. A key finding is that gout is a central diagnosis in the cardiovascular cluster, supporting evidence that gout is important to progression of cardiovascular diseases in a keystone manner.

Extraction example on 6,200,000 patients – cardiovascular disease trajectories.

Source: doi:10.1038/ncomms5022

Holub P. • Scalable Privacy-Sensitive Data Processing • TNC2016, Prague
Select Computer Sci&Tech Challenges

▶ Data compression
  - for different data types: imaging, *-omics data, continuous life(style) measurements,
  - addressing “fitness for purpose”
    - e.g., need for efficient indexing and search capabilities,
  - subject to standardization – for interoperability
    - e.g., MPEG is now looking into genomics data compression.

Genomics data compression example.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Size [MB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM</td>
<td>220,000</td>
</tr>
<tr>
<td>BAM</td>
<td>100,000</td>
</tr>
<tr>
<td>CRAM</td>
<td>50,000</td>
</tr>
<tr>
<td>text genome</td>
<td>1,200</td>
</tr>
<tr>
<td>VCF</td>
<td>135</td>
</tr>
<tr>
<td>VINT+DELTA+DBSNIP+KMER</td>
<td>4</td>
</tr>
</tbody>
</table>

... versus DNA “on the paper”:
Select Medical Informatics and Bioinformatics Challenges I

- Data integration
  - development of common ontologies,
  - development of translations/mappings,
  - healthcare systems interfaces are particularly tricky: bound to national standards (e.g., ICD-10 vs. SNOMED CT).

- Standardization of interfaces to health care systems
Select Medical Informatics and Bioinformatics Challenges II

▶ Domain-specific image analysis – digital pathology
  ■ biomarker identification in ≥1 Gpix imagery
▶ Domain-specific large data visualization
Challenges for (e)Infrastructures I

- Federated identity management and authorization
  - suitable for accessing very privacy sensitive data
  - definition of acceptable levels of assurance for
    - identity vetting,
    - authentication instance,
  - practical implementation in infrastructures
    - government-backed – eIDAS,
    - academic infrastructures,
    - interconnecting with commercial research,
    - dealing with “homeless” users.

- Adoption of user-centric privacy models in the future?

- Multi-tenancy supporting authorization frameworks.
Challenges for (e)Infrastructures II

- Support for informed consent vs. project proposal compliance assessment
  - semi-automation or full automation,
  - flexibility vs. efficiency,
  - deterministic behavior vs. freedom of decision of ethical review boards,
  - GA4GH ADA-M
    https://genomicsandhealth.org/working-groups/our-work/automatable-discovery-and-access

- Support for informed consent withdrawal
  - traceability vs. privacy protection.
Challenges for (e)Infrastructures III

▶ Liability and guarantees of infrastructure providers

Cloud computing example
What if virtual machine host gets hacked into?

▶ Development of commonly accepted best practices

Cloud computing example
How to build “private clouds” using 3rd party resources? Will ISO 27018 become the commonly accepted sufficient standard?

Common guidelines for which purposes public cloud resources can be used.
Challenges for (e)Infrastructures IV

- Developing common provenance/traceability schemes may be modeled using provenance graphs:
  - referring to both physical and virtual objects,
  - privacy protection when identifying humans.
- Requires persistent identifiers with support for:
  - arbitrary datasets and their subsets,
  - supporting datasets evolving in time.
Challenges for (e)Infrastructures V

We need to identify also humans within traceability information:
- to deal with incidental finding,
- to support of informed consent withdrawal.
- Organized decomposition of traceability information compliant with privacy-protection regulations.

What is incidental finding?

You decide to give data for research purposes. As a part of the research, it is discovered that your (then anonymized) data contains a high-risk mutation that can be mitigated.

Whether & to whom & how do you communicate this?
Conclusions I

- Medical research is needed in Europe, despite privacy regulations :)
- We need to bridge the gap between academic research and practically available technologies.
- **De-fragmentation** should be the motto
  - regulations and guidelines
  - interaction between citizens and scientists
  - interoperability: data semantics, coding & compression, ...
- Development of scalable algorithms and tools
- Capacity building suitable for privacy sensitive data
Conclusions II

- BBMRI-ERIC is working on substantial bits
  - quality: biosamples & data
  - findability: biosamples & data
  - interoperability: data
  - partially also scalable data processing
  - working with regulatory authorities, policy makers, etc.

... but scalable privacy-sensitive data processing is a big bite for any infrastructure alone and collaboration of many will be needed in order to transform it into reality!
Thank you for your attention!

Q/?A!

http://www.bbmri-eric.eu/
petr.holub@bbmri-eric.eu