Federated statistics and machine learning.
Synthetic data

Liina Kamm
Methods for privacy-preserving data analysis

- Data export using pseudonymisation
- Data export using anonymisation
- Federated statistics and machine learning
- Using synthetic data
- Statistical analysis using secure multi-party computation
Today on PETs

- Federated statistics
- Federated machine learning
- Synthetic data generation and using synthetic data
- Decentralised contact tracing (HOIA app)
# Federated statistics

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Graphs](image.png)
# Data synthesis

<table>
<thead>
<tr>
<th>code</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYP_01</td>
<td></td>
<td>27</td>
<td>N</td>
<td>165</td>
<td>75</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>HYP_02</td>
<td></td>
<td>25</td>
<td>M</td>
<td>165</td>
<td>85</td>
<td>162</td>
<td>132</td>
<td>88</td>
</tr>
<tr>
<td>HYP_03</td>
<td></td>
<td>27</td>
<td>N</td>
<td>192</td>
<td>90</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Data synthesis**

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>M</td>
<td>165</td>
<td>80</td>
<td>156</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>27</td>
<td>N</td>
<td>165</td>
<td>80</td>
<td>110</td>
<td>133</td>
<td>88</td>
</tr>
<tr>
<td>28</td>
<td>N</td>
<td>195</td>
<td>95</td>
<td>132</td>
<td>142</td>
<td>92</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

28.04.2021
Federated database system

- Multiple autonomous database systems mapped into a single database
- Similar to the X-Road in Estonia
- eesti.ee state portal where you can query data about yourself
  - What analyses have been carried out at the doctor’s and what were the results?
  - What is your census information?
  - How many boats do you own?
  - Who has queried data about you in these databases?
Horisontal and vertical fragmentation/partitioning

### Horizontally partitioned dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>cm</th>
<th>kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>49403136525</td>
<td>27</td>
<td>165</td>
<td>75</td>
</tr>
<tr>
<td>39507082997</td>
<td>25</td>
<td>165</td>
<td>85</td>
</tr>
<tr>
<td>49401223576</td>
<td>27</td>
<td>194</td>
<td>90</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### Vertically partitioned dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>cm</th>
<th>kg</th>
</tr>
</thead>
<tbody>
<tr>
<td>49403136525</td>
<td>27</td>
<td>165</td>
<td>75</td>
</tr>
<tr>
<td>39507082997</td>
<td>25</td>
<td>165</td>
<td>85</td>
</tr>
<tr>
<td>49401223576</td>
<td>27</td>
<td>194</td>
<td>90</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

28.04.2021
Examples of federated databases

- Estonian state database
- Different genome banks in different countries
- Android predictive keyboard
Federated statistics

- Data analysis where as much of the computation as possible is done in the separate databases and the results are aggregated at the analyst
- Does not need a federated database system
- Can be done in case of vertically or horizontally partitioned data
- Adheres to the data minimization principle of GDPR
Does federated statistics solve all our problems?

- Mean? \[ \bar{x} = \frac{\sum_{i=1}^{N} (x_i)}{N} \]

- Standard deviation?
\[ s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \]

- Linear regression?

- What if you have a really proprietary algorithm you are not willing to show?
Federated statistics in practice: DataSHIELD

- [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276062/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276062/)
- Analysis is taken to the data – not the data to the analysis
- Makes co-analysis of data from several studies possible when
  - Governance restrictions prohibit the release of some of the required data
  - A research group is vulnerable to loss of intellectual property (do not wish to hand over the physical data)
  - The physical size of the data precludes direct transfer to a new site for analysis.
- Works on horizontally and vertically partitioned data, and single-site data
DataSHIELD architecture

- Central analysis computer (AC)
- Several data computers (DCs) that store the data to be co-analysed
- Data sets are analysed simultaneously, in parallel
- Non-disclosive summary statistics and commands transmitted back and forth between the DCs and the AC.
- Uses a modified R statistical environment linked to a database deployed in each DC
- Analysis is controlled through a standard R environment at the AC
- Open source software
Federated (machine) learning

• Central analysis computer chooses a model to be trained and sends it to data computers
• Data computers train the model locally and send the results back
• Central analysis computer combines the results
• Requires the models to be combinable
Different settings of FML

- Equal data computers
- Edge computing
- Combined setting
Federated learning: TensorFlow Federated

- [https://www.tensorflow.org/federated/get_started](https://www.tensorflow.org/federated/get_started)
- Analyst can perform basic tasks, such as federated training or evaluation, without having to study the details of federated learning algorithms
- Analyst has to be familiar with TensorFlow, code has to be serializable as a TensorFlow graph
- Local on-device aggregation (task for the analyst)
- Federated cross-device aggregation (task for TFF)
Federated learning: Android predictive keyboard

- Mobile phones can collaboratively learn a shared prediction model and keep all the training data on device
- Device downloads current model
- Improves it by learning from data on your phone
- Summarizes the changes as a small focused update
- Only the update is sent to the cloud (encrypted communication)
- There it is averaged with other user updates to improve the shared model.
- No individual updates are stored in the cloud
Android Predictive Keyboard

- Smarter models
- Lower latency
- Less power consumption
- Ensures privacy
- The improved model on your phone can be used immediately
Technical challenges

• Highly iterative algorithms like Stochastic Gradient Descent (SGD) require a huge dataset and low-latency, high-throughput connections
• Here data is distributed across millions of devices in a highly uneven fashion
• High-latency, low-throughput connections, available for training only intermittently
Google’s Federated Averaging algorithm

• Can train deep networks using 10-100x less communication compared to a naively federated version of SGD

• Use the powerful processors in modern mobile devices to compute higher quality updates than simple gradient steps.

• Fewer iterations of high-quality updates to produce a good model, training can use much less communication

• Compressed updates reduce upload communication costs

• On device training uses a miniature version of TensorFlow.

• Training happens only when the device is idle, plugged in, and on a free wireless connection, so there is no impact on the phone's performance.
Secure aggregation protocol


- A coordinating server can only decrypt the average update if 100s or 1000s of users have participated
- No individual phone's update can be inspected before averaging
- Federated Averaging is designed so the coordinating server only needs the average update
- Protocol is general and can be applied to other problems as well
Does federated statistics solve our problems?

- Model inversion
  - An attacker, given the model and some demographic information about a person, can predict the person’s genetic markers

- How would you counteract this?
  - Secure multi-party computation
  - Differential privacy
Further reading: federated learning

- DataSHIELD: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276062/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276062/)
- TensorFlow Federated: [https://www.tensorflow.org/federated/get_started](https://www.tensorflow.org/federated/get_started)
Synthetic data has been generated from real data and has the same statistical properties as real data.

Synthetic data is not real data.

Types of data and synthetic data

Data can be
- structured (relational databases),
- unstructured text (docors’ notes, transcripts),
- images, videos, audio, virtual environments.

Data can be synthesised
- based on real datasets,
- not based on real data,
- using a hybrid of the two.
Synthesis from real data

1. Build model capturing in the data the
   • distributions and
   • structure (multivariate relationships and interactions).
2. Sample or generate the data from the model.
3. Measure the utility of the generated dataset.
This person does not exist

- https://www.thispersondoesnotexist.com
Synthesis without real data

Created by using existing models, e.g.,
• statistical model of a process,
• simulations (gaming engines do this);

or analyst’s background knowledge, e.g.,
• knowledge of how a financial market behaves,
• knowledge of the statistical distribution of some data.
Examples of data synthesis without real data

Pictures from: [https://tung.github.io/nethack-www/how-to-play-online.html](https://tung.github.io/nethack-www/how-to-play-online.html)

28.04.2021
Utility

Utility measures how similar the generated dataset is to the original.

- High utility important for accuracy (ML models)
- Low utility acceptable for testing throughput

Utility can be measured by comparing analysis results from real and synthetic data.
Benefits of synthetic data

- Access to data for secondary use is problematic
- Consent from users is not practical and can introduce bias
- Open source datasets might not match the problem too well
- Data synthesis without real data when modelling something completely new
- Synthesising data labels
- Trying out models to see whether to apply for data access
Data synthesis methods (based on real data)

• Data imputation methods
  – Need to build a model for the attribute of interest
  – Need to know how the data will be used

• Statistical machine learning methods (e.g., decision trees)
  – Can capture distributions and complex relationships among variables

• Deep learning can capture even more from the data so the results are quite good proxies for real data

• Iterative proportional fitting
Data synthesis steps

Real data → Data synthesis → Synthetic data

Utility assessment → Utility report

Privacy assessment → Assurance report
Examples

• Building the initial models for machine learning when access to data cannot be obtained.
  Example: Simulacrum Artificial patient-like cancer data to help researchers gain insights. [https://simulacrum.healthdatainsight.org.uk](https://simulacrum.healthdatainsight.org.uk)

• Hackathons, open competitions, closed competitions
  Example: Vivli-Microsoft Data Challenge: How can teams safeguard participant privacy and minimize privacy loss while maintaining the scientific analytic value of the data, for smaller data sets or rare disease data set that are more highly identifiable? [https://vivli.org/datathon/](https://vivli.org/datathon/)
Synthesis vs other PETs

Pseudonymisation
• Data are still considered personal information
• Could still be identifiable

Anonymisation (de-identification)
• Data utility becomes questionable

Secure multi-party computation
• Complex to understand, set up and run
• Requires multiple parties
• Performance could be a bottleneck
Further reading


• Simulacrum Artificial patient-like cancer data to help researchers gain insights. https://simulacrum.healthdatainsight.org.uk

• Vivli-Microsoft Data Challenge. https://vivli.org/datathon/
Decentralised contact tracing

• Download and install an app which will save your data
• When you test positive for a highly infectious disease (e.g., COVID-19), you can use the app to recount where you have been
• Inform doctors
• Inform people who might also be infected
Thank you!

Cybernetica AS
Mäealuse 2/1, 12618 Tallinn, Estonia

info@cyber.ee  |  cyber.ee

@cyberetica    @Cybernetica    @CyberneticaAS