Fighting Cancer with Rust

Enola Knezevic

1 Federated Information Systems, German Cancer Research Center
2 Complex Data Processing in Medical Informatics, University Medical Center Mannheim

enola.knezevic@dkfz-heidelberg.de
Biobanks and data stores

- biospecimens, such as serum, plasma, tissue samples
- data about those samples and pseudonymized data about their patient donors
- researchers need to find samples and data by search criteria e.g.
  - sample type
  - storage temperature
  - molecular markers (nucleotide changes, amino-acid changes)
  - patient’s diagnoses
  - therapies patients underwent
Federated search results (BBMRI-ERIC)

- central search
- only receives aggregated data – individual data never leaves the site
- obfuscated and rounded counts
Simplified architecture

Web Traffic
Encrypted
CQL/Beacon/...
Operations

Researcher
Lens (Web Browser)
Webserver
Lens Backend
Beam

Webserver
Beam
Focus
Blaze

Site A
DKFZ Operations
Configuration
Deployment
Monitoring

Site B
Site C

Laplace

Encrypted

DKFZ
Operations

Cancer Research Center
in the Helmholtz Association

UNIVERSITÄTSKLINIKUM
MANNHEIM
Projects

- GBN – German Biobank Node
- BBMRI-ERIC - Biobanking and Biomolecular Resources Research Infrastructure – European Research Infrastructure Consortium
- DKTK – German Cancer Consortium
- CCP – Clinical Communication Platform
- Cancer Core Europe
- ITCC P4 - Paediatric Preclinical Proof Of Concept Platform
European Federation for Cancer Images
Samply.Beam

• Distributed task broker designed for efficient communication across strict network environments present in medical informatics:
  • End-to-end encryption
  • Certificate management and validation
  • Only outbound connections
Samply.Focus

- Federated query dispatcher working with Samply.Beam
- CQL query generation to prevent CQL injections
- AST translation: EUCAIM: Chaimeleon, ProCAncer-I
- Running the query against the data stores, other applications
- Query result obfuscation using Samply.Laplace library
Differential privacy algorithms

• Promises:
  • Preserve privacy while keeping the data useful for research: publish aggregate data, withhold individual data
  • Resist differencing attacks trying to identify whether an individual’s data is in a certain database or not (e.g. by selecting a certain diagnosis, date of diagnosis, age, and gender of the patient)
  • **Offer similar level of privacy as having individual’s data removed from the database**
Why k-anonymity is not enough

• Medical data often anonymised and pseudonymised
• k-anonymity - all combinations of attributes are satisfied by at least k entries in the dataset
• “Cut-off" value of k not sufficient to ensure k-anonymity, especially if multiple sensitive criteria are involved
• Sensitive characteristic evenly distributed within a class - can be inferred
• Associating anonymised information with additional external information
Laplace distribution

The Laplace distribution is a continuous probability distribution that is often used in statistics. It is characterized by two parameters: \( \mu \) (the location parameter) and \( b \) (the scale parameter). The probability density function (PDF) of the Laplace distribution is given by:

\[
f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}
\]

where \( x \) is a random variable, \( \mu \) is the location parameter, and \( b \) is the scale parameter.

In the GeoGebra calculator suite, you can adjust the parameters \( \mu \) and \( b \) to visualize the Laplace distribution. The graph shows the behavior of the distribution for different values of \( \mu \) and \( b \).
Privacy – usability trade-off
Privacy – usability trade-off

$$b = \frac{\delta}{\varepsilon}$$

- $\delta$ – sensitivity
- $\varepsilon$ – privacy budget
Samply.Laplace

- differential privacy-inspired query result obfuscation
- Rust crate (& a Java library in case anyone prefers that)
- highly configurable:
  - sensitivity, privacy budget
  - values under 10: change all to 10, change all to 0, or obfuscate in the usual way
  - turn off obfuscation of zeroes
  - rounding step
  - obfuscation value cache – consistent results
Open-source

• Apache-2.0 license
• Feel free to use and contribute to our software