Practice 12-16

Data anonymization and security measures

Data anonymization vs. pseudonymization

- **Anonymization** - a process to render personal data nonpersonal
  - Complex
  - Extremely important for collection, management, and sharing of data
- **Pseudonymization** - data has been removed of identifying markers
- Difference - the degree of protection provided and whether re-identification is possible

According to the General Data Protection Regulation (GDPR) (European Parliament and Council 2016) Art. 4, pseudonymization is defined as: “the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.”

Potential risks of re-identification

- Privacy Violations
- Identity Theft and Fraud
- Discrimination and Profiling
- Reputational Harm
- Erosion of Trust
- Legal and Regulatory Ramifications
- Ethical Concerns

Data Anonymization Techniques

- Data redaction
- Data nulling
- Data masking
- Pseudonymization
- Generalization
- Data swapping
- Data perturbation
- Data encryption
- Hashing
- Bucketing
- Tokenization
- Synthetic data generation

Praktiniai sprendimai (1)

- PII – Personally Identifiable Information
- Registracija/Mokami:
  - https://www.k2view.com/solutions/data-masking-tools/
  - https://console.gretel.cloud
  - https://www.broadcom.com/products/software/app-dev/test-data-manager
  - https://www.ibm.com/products/infosphere-optim-data-privacy

Praktiniai sprendimai (2)

- SPIDER (Secure Privacy-preserving Identity management in Distributed Environments for Research)
  - https://eu-rd-platform.jrc.ec.europa.eu/spider/
- ARX
  - https://arx.deidentifier.org/
  - https://github.com/arx-deidentifier/arx/tree/master/src/example/org/deidentifier/arx/examples
- sdcMicro (R Package)
  - https://cran.r-project.org/web/packages/sdcMicro/
- Faker
  - https://faker.readthedocs.io/
- https://github.com/JohnSnowLabs/spark-nlp-workshop/tree/master/tutorials/Certification_Trainings
- Amnesia - https://amnesia.openaire.eu/download.html

Duomenų rinkiniai

- Įvairūs:
  - https://www.kaggle.com/datasets
  - https://archive.ics.uci.edu/

- Syntheticheath.github.io/synthea

Užduotis

- Savo baigiamojo darbo tema surasti duomenų rinkinius
- Dauguma duomenų bus jau sutvarkyti, todėl turite apgalvoti scenarijus, jei to nebūtų (kaip užtikrinsite anonimiškumą, ką naudosite ir pan.)
- Tinkamus duomenų rinkinius iš karto dėti į MBD ir juos aprašyti t.y. koks formatas, kokios reikšmės, kiek jų ir t.t.