# Remarks

**22:** state of the art dla aplikacji sieciowych

**23:** Dlaczego aplikacja webowa?

Silesian
University
of Technology


**SILESIAN UNIVERSITY OF TECHNOLOGY**

**FACULTY OF AUTOMATIC CONTROL, ELECTRONICS AND COMPUTER SCIENCE**

**PROGRAMME: INFORMATICS**


Final Project

# Data anonymisation web platform


author: Szymon Pluta

supervisor: Krzysztof Simiński, DSc PhD


Gliwice, January 2022

# Contents

# Abstract

The text of the abstract should be copied into a respective field in the APD system. The Abstract with keywords should not exceed one page.

**Keywords:** 2-5 keywords, separated by commas

# Chapter 1

# Introduction

Technological advancement being observed in the past years fundamentally changed the relevance of data in today's digitalized world. Information became an innovation stimulus in the area of research and development. The quantity of data that organizations produce, process, store and share is at a continuous growth. An enormous amount of 1.8 zettabytes ($1.8 \cdot 10^{21}$ bytes) of new data was produced only in the 2011, and every two consecutive years this number is doubling [41]. After decades of observed technological advancement and innovation, the global internet traffic finally entered the zettabyte-era, as it had reached a magnitude of one zettabyte in 2016, and in the calender month being as early as September [9].

The vast quantities of processed information allowed for brand new research fields such as data science or big data analytics to form, which are used by organizations to derive new insights in a way that was previously impossible. Organizations collect and process the data to enhance the services they provide to the customers through statistical analysis or newly developed computer science processes including data mining and machine learning. The utility of delivered services is increased at a lower cost and improved efficiency through the insights extracted from the collected information about how the services are consumed [4].

Następne do zrobienia, w uproszczeniu i w podanej kolejności:

- Jak widać, dane są używane wszędzie. Przeciętna osoba sobie nawet nie zdaje sprawy.
- Istota prywatności danych, ochrona danych osobowych.

- Prawo wolno reaguje na uregulowanie ochrony danych osobowych: GDPR dopiero w 2018. Prawo różnie działa w różnych krajach.
- W czasach narastających danych, dane muszą być zanominizowane, bo...
- ...ale:

Wniosek: generyczny sposób anonimizacji - nie ma takiego.

- Mogą istnieć dobre algorytmy anonimizacji, ale są optymalale jedynie w określonych kontekstach
- **Cel pracy**: stworzenie generycznego rozwiązania do anonimizacji baz danych w dowolnej formie danych (nieważne co dane reprezentują).
- Kontroler danych zna swoje dane i sam dostosowuje optymalny sposób anonimizacji.
- Rozwiązanie może być dostarczone jako free-access (research), B2C lub nawet B2B.
- Zakres pracy, opis rozdziałów - na samym końcu, po zrobieniu innych rozdziałów.

- introduction into the problem domain
- settling of the problem in the domain
- objective of the thesis
- scope of the thesis
- short description of chapters
- clear description of contribution of the thesis's author – in case of more authors table with enumeration of contribution of authors

# Chapter 2

# Problem analysis

## 2.1 Data explosion

### 2.1.1 Technology advancement

The continuous and rapid exponential growth of data being collected globally is further excited by improvements to the overall population's accessibility to the digital technology [6]. Cisco Systems estimates within its annual report [8] that 66 percent of the world population will have an access to the web by the 2023, compared to 51 percent in 2018, whereas the number of devices that are connected to the web will reach a staggering value of three times as many as the entire population size – demonstrating a total of 60 percent expansion when compared to 2018. Even the area of mobile connection, which was established long ago, is still sustaining a growth – by 2023, a mobile connectivity will be a privilege for 70 percent of the world's population, compared to 66 percent in 2018. The global average mobile network speeds will be tripled through a rapid increase from 13.2 Mbps in 2018 to 43.9 Mbps in 2023.

### 2.1.2 Data analytics

Raw representation form of the data is not interpretable until it is put under a context and processed into practical information. Acquiring relevant insights and conclusions from the information can be achieved through a wide term of analytics,

which encompasses the actions needed to be performed to produce new information, including: analysis, filtering, processing, grouping and contextualizing the information. Newly discovered knowledge is infered from the produced information. Apart from the processes, analytics also includes the technologies, methodologies and algorithms to use and could be divided into descriptive analytics, diagnostic analytics, prescriptive analytics and predictive analytics [6].

### 2.1.3 Big data

Big data analytics deals with the difficulties of managing the observed exponentially increasing collected volumes of data. Its purpose is not only to handle the processing and analysis of the data through specialized software tools and frameworks, but also to handle the means on how this enormous amount of data is collected and stored in the first place. It is in its nature that big data is all about massive volumes of information that require specialized hardware infrastructure to store it [6].

Services of enterprise organizations are running on all the collected data which can take various forms such as database entries, metrics, logs or outgoing messages. New data streaming technologies working at a large scale needed to be engineered to handle the continuous flow of data between systems and databases. An example of such technology includes Apache Kafka which generates even more than a trillion of messages per day for individual large enterprise organizations taking advantage of it [24][33].

This only proves that big data deals not only with massive volumes – it has also to deal with the high velocities of data generation, which is yet another characteristic of data [6]. According to DOMO report published back in 2020, 90% of world's data was generated just in the preceding two years, and on average every person in the world created 1.7 megabytes of data per second – which yields 2.5 quintillion ($2.5 \cdot 10^{18}$) bytes of new data each day [19].

The momentum of the immense big data interest growth among organizations is not fading away yet, as more and more new businesses and researches are drawn to this subject. The benefits of big data especially concern scientific organizations and large enterprises of which financial domain and IT industry are the common

consumers [21]. The organizations find the interest in information analytics for remarkably diversified reasons. It is recognized as a field that will entirely alter all parts of civilization such as businesses or society as a whole [30].

### Applications

Various types of organizations collect data to take advantage from the insights derived out of the data. The big data analytics applications' impact can be observed already today in a broad spectrum of domains.

Leading technology companies, such as Google and Facebook, to name a few, sell anonymised collected user data access to their partner advertisers [30]. This is legally possible as the information that was anonymised, i.e., deidentified in a way that it is no longer bound to an individual, may flow from a system to a system.

Large Hadron Collider (LHC) located in CERN, being the largest physical experiment, annually produce approximately 30 petabytes of data. LHC takes an advantage of light sensors that monitor the collisions of hundreds of millions of particles accelerated nearly to the speed of light. The collisions create enormous amount of data to be processed by computer algorithms in the hope of discovering new particles, e.g., a Nobel prize awarding discovery of Higgs boson had taken place in 2012 [1].

Enterprise stores such as Amazon or SAP Commerce Cloud collect the information regarding the way of how the visitors browse and interact with these stores. Collected information may involve the behavioral data related to customer engagement, such as the pages we visit, event clicks, or the way we scroll the page. The insights derived from the collected information enable making future improvements of these services – for example by improving the digital marketing or performance improvements based on the metrics [28]. The customer experience is also improved as based on the collected data the advertisements or item recommendations can be taylored to the specific user's preferences. The recommendation engine may also attempt to match your profile data to people of similiar profile to provide better recommendations. Services attempt to analyze the behavioural patterns such as time of day we browse the store or what circumstances caused our last visit to finish. Even the details such as the exact neighbourhood location we live in, com-

bined with its estimated wealth, organizations may attempt to guess our potential income level [30]. These data analytics are performed to improve the possibility of customer buying yet another item.

## 2.2 Data privacy

The privacy endangerment is inherent to big data and it is its major drawback. Our personal data we continuously give away to third parties is the big data fuel.

As the technology evolves, concerns relating to the privacy of our personal information should also grow – and for a relevant reason. We tend not to give a second thought to whom the data is shared, how it may be used and in what kind of circumstances. We don't wonder how our own data may be exploited to alter our thinking, decisions, or even ideas – whether in an ethical manner or not.

Although the use of our personal information existed ever since the very first census was created, data privacy is a relatively new concept, as it did not exist prior to the global adoption of the internet. Granted that our data was used by the researchers even before the digitalized era of the internet, the motives for that usage were not commercial [38]. Nowadays, the data has most certainly become an asset – a resource like any others, and a rather precious one.

It is argued that the data privacy should be centered only around the data usages that have the potential to be a privacy breach. On the other hand, it is argued that merely a collection of the data is already a privacy harm. The information that was collected is endangered by many threats, including data misuse, data breach, data leak or even authorities accesses without legal obligations. Anonymisation is the best method to mitigate conflicts raised by big data with respect to data privacy and data protection [40].

Luckily the law had finally caught up to the circumstances of the increasing data usages and the associated risks. New European regulations were adopted in 2018 in the form of General Data Protection Regulation (GDPR) to protect our data privacy in a refined fashion. The data subjects, i.e., the individuals represented by the data [4], now have a better control over their personal information – we are now entitled to know what information concerning us is being processed and for what purpose. We are also entitled to withdraw at any time the consent for

the processing of our data. In case of violations we have the right to complain to authorities and seek justice against both the data controllers, i.e., the entities that determine the intention and means of processing the personal information, and the data processors, i.e., the entities that process the personal information on behalf of the data controllers [42].

An overall awareness of the data privacy significance had improved in the society, and the means to achieve the data privacy through data protection had also improved as organizations needed to adopt to the new situation by applying enhanced measures to their protection of data – anonymisation and pseudonymisation being the notable examples of such measures.

### 2.2.1 Authorization to share data

Majority of data privacy regulations are based on a consent of an individual, i.e., it is lawful to process and use the information for secondary purposes only if an individual explicitly acknowledge their consent for that [44]. This may appear easier said than done due to the unobvious difficulties data controllers face when trying to obtain such a broad authorization consent that will take into account all possible secondary purpose usages.

Consider a patient entering medical facility for an ordinary appointment. The patient would likely find it unusual, disturbing or even shocking if upon his entrance to the facility he was to receive an overwhelming form that included dozens of independent consent authorization requests. The consents could give the impression of being seemingly unrelated to his visit in the first place, e.g., a constent to share the data with researchers of an university located on another continent. In the end that could destroy the data subject's trust – in this case patient's trust.

This theoretical scenario may not easily be implemented in the real world counterpart, as it could be even impossible to know or predict all possible secondary purpose usages in the first place, and consent based authorization is all about knowing the usages.

Consider a newly discovered purpose to process personal information of an already existing database. Getting a consent after the data had already been collected, i.e., backwards in time, would be impossible to accomplish as the data

controller would need to contact potentially hundreds of thousands of people for their explicit consent. New purpose can be discovered years after the data collection.

Having that in mind, no consent is required when processing the data that is already anonymised. A data that was stripped from personal identifying or identifiable information data can be used in any way and can be shared with third parties without previously agreed consent. Data controllers now face a realistic to solve problem of information anonymisation rather than an unrealistic problem of consents collection [20].

It is worth mentioning that there exists cases which are defined under Article 6 of General Data Protection Regulation (GDPR) [15] when consent is not required to process the data, e.g., if the processing is required to defend the data subject's interests or in order not to break the compliance with legal obligations as a data controller. GDPR is a new European law that replaces the preceeding Data Protection Directive regulation adopted by European Community in 1995 [44].

### 2.2.2 General Data Protection Regulation

One of the primary objectives of GDPR is personal data privacy protection which is a fundamental right and freedom of people as defined under the Recital 1 of GDPR [11] and the Charter of Fundamental Rights of the European Union [10]. Newly discovered challenges for the protection of personal data arise from the ongoing globalization and quick development of digital technologies. This in turn vastly had increased both the scope of the gathering of the data, and the sharing of thereof. General Data Protection Regulation (GDPR) is a data protection law that came into force on May 25, 2018 to addresses these data privacy related issues in a strict manner[13].

Compliance with GDPR law is critical for organizations in the view of significant administrative fines they face. Violations of the data processor and data controller obligations defined in GPDR are subject to costly penalties that are imposed by European authorities. Non compliance with technical rules imply a penalty of 2 percent of the total annual turnover of the previous year, or €10 milion, whichever one is higher, whereas non compliance with basic data pro-

tection principles imply an even higher penalty of 4 percent of the total annual turnover of the previous year, or €20 million, whichever one is higher [14][31].

Implementation of this law had immediately increased the significance of data anonymisation as an information sanitization process in today's world [34]. Anonymisation being a specific form of data masking suddenly became more relevant in today's world for the reason that the strict regulations, and therefore administrative fines, defined in GDPR do not apply to the anonymised information. Data protection principles covered throughout GDPR concern only the processing of information that is already identified to a natural person, or that is identifiable to a natural person, i.e., an individual is yet to be identified. Given the fact that anonymised information is by definition not relating to a person, hence it can be completely exempted from falling under the GDPR requirements, which apply only to personal data, as stated under Recital 26 [12]:

> The principles of data protection should apply to any information concerning an identified or identifiable natural person. [...] The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

GDPR distinguishes personal data, anonymised data and pseudonymised data as distinct variations of data. The information that had gone merely through pseudonymisation process would still fall under the regulations of GDPR, due to the existing relevant possibility of future re-identification of the data subject, whereas in case of the anonymised data, such re-identification is by definition either impossible or extremely impractical, and the anonymisation is irreversible by definition. Anonymised data is completely exempted from being governed by GDPR. Nevertheless, pseudonymisation is still one of many possibilities for the data controllers and data processors to be GDPR compliant [44]. The point of anonymisation in the context of GDPR is to be completely exempted from being governed by this regulation.

## 2.3 Data anonymisation

The data released by organizations exclude identity related information such as names, addresses, telephone numbers or credit card numbers. Personal data is stripped from disseminated data sets through anonymisation to protect the anonymity of the individuals, i.e., data subjects [36].

### 2.3.1 Background

Consider a collection of medical data concerning patients' clinical information. Processing and sharing of medical data volumes is crucial for the evolution of world's healthcare services. Medical researchers and doctors take an advantage of the collected data sets to improve their comprehension of diseases and explore new possibilities to treat these diseases, and hence both the overall capability to treat the diseases and the general efficiency of health services are improved. At last it is the patients who benefit from the research conducted on their data since the services they are offered with continuously improve. Nevertheless, it is known that medical data is exceptionally sensitive by its nature due to the details that include e.g., patient data, laboratory tests results, diagnosis details, prescribed medications and history of diseases [26].

Having understood how sensitive by its essence is the patient information and the vital needs to share this data, this is where data anonymisation plays an indeed crucial role. It would be impossible to disseminate patient information without prior anonymization of thereof.

### 2.3.2 Definition

Anonymisation is a statistical disclosure control method of a particular importance. The ultimate anonymisation goal is to de-identify the data by pruning the personal information in such a way that the relation between data subject and corresponding records is blurred.

The data anonymisation is considered to be an effective one only if both of the following criterias hold true [27]:

- performed anonymisation operations are irreversible

- data subject re-identification is impossible or impractical

### 2.3.3 Data classification

The prerequisite for anonymisation is to understand the context of what does the data represent and by whom – in addition to how – will it be processed [13].

The commonly adopted approach when attempting to anonymise the data is to first classify the data attributes as direct identifiers (i.e., personal identifying confidential attributes unique to an individual, such as address or tax identification number), indirect identifiers (i.e., so called quasi-identifiers that when combined can discover the individual's identity – such as sex or date of birth) and non identifiers [18][20].

It is context dependent to determine which attributes are quasi-identifiers – and finally which techniques need to be combined to achieve anonymisation. Virtually, even the most unexpected attributes could be quasi-identifiers. Ignoring such fields when preparing for anonymisation may eventually lead to future re-identification of the data subjects, as successfully demonstrated, e.g., when breaking the anonymity of the Netflix Prize dataset back in 2006 [3]. Since then significant advancements in the field of de-identification were yielded – and in various domains [5].

On the other hand, too much anonymisation of quasi-identifiers strips the data out of utility.

### 2.3.4 Utility trade-off

The organizations need to compromise between the utility and privacy when releasing the data. As anonymisation level increases when more and more restrictive techniques are applied, the utility of the same data decreases.

Consider the two theoretical boundary cases of either releasing data in its original form which maximizes the utility at the cost of discarding the protection of the data subjects' privacy altogether, and the second case of not releasing any data at all which completely preserves the privacy [45].

Clearly the data must not be released in its original state – strict GDPR that imposes large administrative fines [14] exists to mitigate such violations. On the other hand, data that completely suppresses all the information has no value.

The data controllers therefore need to strike a balance between the two cases. An appropriate anonymisation strategy maximizing the utility without endangering the privacy must be adopted.

### 2.3.5 Pseudonymisation

What is pseudonymization? How does it differ?

### 2.3.6 Deidentification

Nawet jeśli dane personall identifiable information PII są usunięte, to można połączyć osobe z pomocą istniejących nieusuniętych danych np. płeć oraz external informacjach np. social media. Komputery są coraz szybsze, źródeł danych coraz więcej - napastnikom coraz łatwiej.

Przykłady takich ataków (Netflix Prize, IMDB). https://www.cs.princeton.edu/ arvindn/publ anonymization-retrospective.pdf https://arxiv.org/abs/1512.08546

// todo Conclusion: Anonymisation = false sense of security! All the known examples of this type of identification are from the research world — no commercial or malicious uses have yet come to light — but they prove that anonymization is not an absolute protection.

Why you can't really anonymize the data:

https://web.archive.org/web/20140109052803/http://strata.oreilly.com/2011/05/anonymize data-limits.html

## 2.4 Data masking techniques

It is not possible to predetermine – in a general fashion and without an extra context – the data masking techniques that need to be combined together in order to achieve an actual anonymisation of the previously identifiable information. Instead, a context consisting of:

- exact representation form of the data being processed *(What is the data?)*
- data processors who use the data *(Who will use the data?)*

- processing purpose, e.g., research objective *(How will the data be used?)*

is always required when considering an effective way to anonymise data under that context [20].

### 2.4.1 Suppression

Suppression is the strongest anonymisation technique which completely removes all the values associated with the given attribute, hence rendering the complete protection of the data and therefore enforcing the best data subjects privacy. It is guaranteed that no further implication attacks can be performed against this data [36]. Nevertheless, suppression being the strongest privacy preserving method implies the highest (i.e., worst) data utility loss as a inseparable consequence – the ultimate cost of complete anonymisation that is to be paid.

Consider the raw data of the three attributes presented in the table 2.1 to be put under the process of suppression.

Table 2.1: Data prior to suppression.

| Sex | PIN codes | Phone number |
|---|---|---|
| F | 3248 | 1212 – 345345 |
| F | 8090 | 4000 – 303030 |
| M | 1337 | 5191 – 915100 |

Exemplary results of performed suppression are depicted in the table 2.2. All of the three attributes values were suppressed to the value of the suppression token supplied by the data controller.

Table 2.2: Data after suppression.

| Sex | PIN codes | Phone number |
|---|---|---|
| F/M | #### | 3000 – 123123 |
| F/M | #### | 3000 – 123123 |
| F/M | #### | 3000 – 123123 |

Phone number suppression deserves a brief discussion. Consider a context in which an attacker successfully obtained the information of exactly one individual

data subject of his interest. Consider that the attacker decided to call the anonymised phone number. This number could be suppressed in a way to specifically protect against this attack and the attacker could call someone intentionally most definitely did not, i.e., data protection officer or authorities. In this scenario it could be trivial to trace the call and find the potential attacker. As demonstrated, suppressed could be used in unobvious ways.

Suppression is typically applied on the column level of an attribute, however it is also applicable to the individual records. When compared to plain generalisation, suppression yields a higher information loss and can be though of as a particular case of generalisation [26].

The designed software offers column level suppression with a user specified value (i.e., suppression token). The values are transformed to this user specified value using a simple algorithm shown in 2.5.

```
1 procedure suppression(values, token)
2   for (value in values)
3     value := token
4   return values
```

Figure 2.1: Suppression pseudocode.

### 2.4.2 Generalisation

Generalisation is the second most common non-perturbative anonymisation technique [18]. This technique processes the raw data by aggregating and substituting the initial values of a given attribute to the values that are more general [2]. The process of generalisation constitutes a conversion of any value to a more general scope.

The proposed software solution includes two generalisation strategies, namely:

- based on the size of distributions
- based on the number of distributions

Consider the raw data of the three attributes presented in the table 2.3 to be put under the process of generalisation.

Table 2.3: Data prior to generalisation.

| Age | Salary | Location |
|---|---|---|
| 27 | 36 000 | Poland |
| 52 | 54 000 | Canada |
| 30 | 180 000 | Poland |
| 68 | 128 000 | Switzerland |

Results shown in the table 2.4 include exemplary possible values after being processed by the generalisation method. As depicted in the table, all three attributes have been grouped into broader value ranges, hence undergoing the generalisation, yet every column was generalised with a different generalisation strategy.

Table 2.4: Generalised data.

| Age | Salary | Location |
|---|---|---|
| 26 – 30 | 1 – 60000 | Europe |
| 51 – 55 | 1 – 60000 | North America |
| 26 – 30 | 120001 – 180000 | Europe |
| 66 – 70 | 120001 – 180000 | Europe |

Age column was generalised with a strategy based on the size of distribution. In this case the size of distribution was defined as 5, as every interval aggregates five distinct increasing integer values.

On the other hand, salary data was generalised with a strategy based on the number of distributions – the values were aggregated to three even distributions.

Finally, the generalisation does not necessarily concern only numerical values as observed in the generalisation of location attribute. To better understand this type of generalisation, consider that the age attribute could be generalised into the values of e.g., *young adult, adult, senior*. This type of generalisation is not supported by the developed software, however an illustration of existence of such method is necessary. The fundamental assumption of the developed software is to be generic, i.e., data context-agnostic, and this generalisation method is too specific in its essence to be implemented in a generic fashion.

Both supported generalisation strategies were further enhanced with customizations. The data processor is allowed to predefine minimum and maximum bound-

ary values for the generalised intervals. The computed starting value for the distribution concerning lowest interval is either minimum raw value or user predefined minimum value, whichever is lower. The maximal value for the highest interval is computed in an analogical way.

```
1  procedure generalise(values, configuration)
2    computedMin := min(values.min(), configuration.userMin())
3    computedMax := max(values.max(), configuration.userMax())
4    distributionSize := configuration.distributionSize()
5
6    // Phase 1: Generate empty intervals.
7    intervals := <Interval, Values>
8    i := computedMin
9    while (i < computedMax)
10     interval := asEmptyInterval(i, i + distributionSize)
11     insert interval into intervals
12     i += distributionSize
13
14   // Phase 2: Populate intervals.
15   for (value in values)
16     for (interval in intervals)
17       if (value is within interval)
18         interval := get interval from intervals
19         add value to interval
20   return intervals
```

Figure 2.2: Pseudocode for generalisation based on distribution size.

The generalisation strategies can be summarized with the pseudocodes shown in 2.2 and 2.3.

Generalisation may be applied on a global level or local level [36]. In terms of database related vocabulary, as this paper primarily concerns an anonymisation of databases, we say generalisation on a column or record level, respectively.

### 2.4.3 Perturbation

Perturbation is a data masking technique which encompasses the process of swapping the original values with artificial ones in a way that the statistical factors of the data are similar to the initial data. Perturbation is typically achieved by adding noise to the information so that the values are slightly different [23].

```
procedure generalise(values, configuration)
  computedMin := min(values.min(), configuration.userMin())
  computedMax := max(values.max(), configuration.userMax())
  numberOfDistributions = configuration.numberOfDistributions()

  // Phase 1: Generate empty intervals.
  intervals := <Interval, Values>
  i := computedMin

  distributionSize := (computedMax - computedMin) / numberOfDistributions
  j := 0
  while (j < numberOfDistributions)
    interval := asEmptyInterval(i, i + distributionSize)
    insert interval into intervals
    i += distributionSize
    j += 1

  // Phase 2: Populate intervals.
  for (value in values)
    for (interval in intervals)
      if (value is within interval)
        interval := get interval from intervals
        add value to interval
  return intervals
```

Figure 2.3: Pseudocode for generalisation based on number of distributions.

Consider weight and height data shown in the table 2.5.

Table 2.5: Data prior to perturbation.

| Height | Weight |
|---|---|
| 166 | 58 |
| 170 | 66 |
| 194 | 91 |

The table 2.6 shows the data after it was perturbed using the two different perturbation methods that are supported by the engineered software.

The supported methods include adding the noise based on the strategies of:

- fixed value
- percentage value

Table 2.6: Perturbed data.

| Height | Weight |
|---|---|
| 169 | 57 |
| 168 | 67 |
| 193 | 91 |

Height attribute values could have been perturbed using the former (i.e., fixed noise of ±3), whereas weight attribute values could have been perturbed using the latter (i.e., percentage value of ±5%). In either case, the linkage between individual records is blurred with this technique.

```
1 procedure perturbation(values, noise)
2   for (value in values)
3     random := pick random from [value - noise, value + noise] interval
4     value := random
5     value := fit value within boundaries
6   return values
```

Figure 2.4: Pseudocode for perturbation based on fixed noise.

Similarly to generalisation, both perturbation techniques allow the user to explicitly specify the minimum and maximum accepted boundary values.

```
1 procedure perturbation(values, noise)
2   for (value in values)
3     coeff := pick random from [1 - noise, 1 + noise] interval
4     value := value * coeff
5     value := fit value within boundaries
6   return values
```

Figure 2.5: Pseudocode for perturbation based on percentage noise.

Perturbation technique works the best with continuous values - it is an effective method for de-identification of quasi-identifiers such as dates and numbers [39]. Scientifically advanced perturbation approaches are further divided into linear and non-linear perturbation models [37].

### 2.4.4 Pattern masking

### 2.4.5 Randomisation

#### Column shuffle

#### Row shuffle

### 2.4.6 Artificial data

#### Random data

### 2.4.7 Shortening

### 2.4.8 Hashing

### 2.4.9 Encryption

### 2.4.10 Tokenisation

## 2.5 Existing solutions

List of existing solutions.

## 2.6 Platform

Why web platform, state of the art for web.

- definicja anonimizacji
- rodzaje animonimizacji (od razu odnośniki do literatury (bibliografia))
- metody, narzędzia
- problem analysis
- state of the art, problem statement
- literature research (all sources in the thesis have to be referenced [**bib:article**, **bib:book**, **bib:conference**, **bib:internet**])

- description of existing solutions (also scientific ones, if the problem is scientifically researched), algorithms, location of the thesis in the scientific domain

[state of the art dla aplikacji sieciowych]

# Chapter 3

# Design specification

[Dlaczego aplikacja webowa?]

## 3.1 Requirements engineering

### 3.1.1 Functional requirements

### 3.1.2 Non-functional requirements

## 3.2 System engineering

### 3.2.1 Use cases

### 3.2.2 REST API

## 3.3 Software tools and technologies

The following listings are presented to give an overview of the designed system in terms of tools and technologies involved to engineer it. This is a brief summary – should a particular technology receive a greater detail of explanation, e.g., exemplary usage, it will be described in another dedicated unit.

### 3.3.1 Tools

This section encompasses technologies needed to develop, build, and maintain the system.

#### Docker

Docker is a virtualization software used to develop, run and manage applications in the form of containers taking advantage of Linux kernel features namespaces and cgroups. It consists of three components: the runtime, the daemon engine and the orchaestrator.

To run the production environment of anonymisation platform, only prior installation of Docker on the host system is required, because containerization technologies like Docker enable a separation of the software from the infrastructure the software is running on [17][35].

#### Apache Maven

Apache Maven is an open source that automates the build process of foremostly Java based projects. Primary Maven objectives include ensuring an easy build process and an uniformly established build system which is achieved through Maven's build lifecycle (i.e., predefined build steps, e.g., *validate*, *compile*, *test* or *install*) [25].

Anonymisation platform uses this tool to build the backend server. Maven is put into service in one of the two ways – depending on the selected environment – namely:

- internally from the Docker container where Maven can be considered as an abstraction that the system administrator does not have be aware of
- on the host system itself

Maven is also used to run the tests. Gradle could be a great alternative to Maven, the choice being preference based in usual scenarios.

### Node.js

Node.js is a Javascript runtime engine encapsulating V8 engine developed by Google which implements ECMAScript [29]. While node.js may function as a dedicated backend server, in the case of anonymisation platform it is used merely to build the client application with yarn.

### yarn

Yet another resource negotiator (yarn for short) is a package manager developed by Facebook, Google and others [32], whose primary purpose is – similarly to *npm* – installation and management of dependencies used by Node.js runtime environment based projects. Yarn also focuses on ensuring high level of security, performance and reliability [22].

While the purpose of Maven is to build and run the backend server, yarn's purpose is to build and run the frontend client application. Consequently, yarn is used in the same way as Maven – either internally in Docker or on the host system – depending on the chosen environment.

### nginx

Nginx is an open-source that works as a reverse proxy, web server and HTTP load balancer with the first one being the primary function for anonymisation platform. Reverse proxies improve resiliency, security, scalability and performance of the software systems. This is particularly relevant when designing systems that scale horizontally [16].

### git

It's been a while since most developers had already started using git for development, which is the leading software for version control, however not everyone knows it was originally invented by Linus Torvalds specifically for Linux kernel development needs [43]. It is clearly necessary to use git during software development due to the essential needs to publish, stash or revert the code changes [7].

- Docker

- Maven
- Yarn
- Nodejs
- Nginx
- IntelliJ and Webstorm
- Git
- GithubActions
- SonarQube
- ESLint and Prettier

### 3.3.2 Server

- Java 17
- Spring Framework (Boot: Web, Data, Security, Test, Validation, Mail)
- Hibernate + Hibernate types55
- PostgreSQL Server
- PostgreSQL Client
- Junit5
- Testcontainers
- Rest assured
- Swagger
- Other libraries: commons lang3, bouncycastle, jjwt, ztexec

### 3.3.3 Client

- Typescript
- React
- React router
- React query
- React hook form
- Yup
- Material ui
- Axios

## 3.4 Development methodology

- functional and nonfunctional requirements
- use cases (UML diagrams)
- description of tools
- methodology of design and implementation

# Chapter 4

# External specification

- hardware and software requirements
- installation procedure
- activation procedure
- types of users
- user manual
- system administration
- security issues
- example of usage
- working scenarios (with screenshots or output files)

## 4.1 Environments

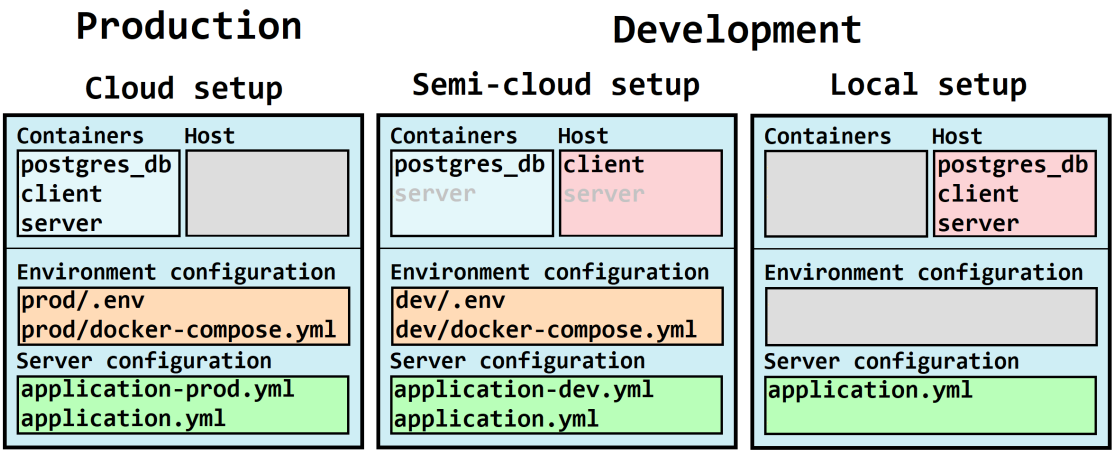


Figure 4.1: Supported environments


Figure 4.2: Figure caption (below the figure).

# Chapter 5

# Internal specification

- concept of the system
- system architecture
- description of data structures (and data bases)
- components, modules, libraries, resume of important classes (if used)
- resume of important algorithms (if used)
- details of implementation of selected parts
- applied design patterns
- UML diagrams

Use special environment for inline code, eg **descriptor** or **descriptor_gaussian**. Longer parts of code put in the figure environment, eg. code in Fig. 5.1. Very long listings–move to an appendix.

```
1  class descriptor_gaussian : virtual public descriptor
2  {
3      protected:
4          /** core of the gaussian fuzzy set */
5          double _mean;
6          /** fuzzyfication of the gaussian fuzzy set */
7          double _stddev;
8
9      public:
10         /** @param mean core of the set
11             @param stddev standard deviation */
12         descriptor_gaussian (double mean, double stddev);
13         descriptor_gaussian (const descriptor_gaussian & w
               );
14         virtual ~descriptor_gaussian();
15         virtual descriptor * clone () const;
16
17         /** The method elaborates membership to the
               gaussian fuzzy set. */
18         virtual double getMembership (double x) const;
19
20 };
```

Figure 5.1: The **descriptor_gaussian** class.

# Chapter 6

# Verification and validation

- testing paradigm (eg V model)
- test cases, testing scope (full / partial)
- detected and fixed bugs
- results of experiments (optional)

# Chapter 7

# Conclusions

```
1  public class ColumnShuffleFacade extends AnonymisationFacade {
2
3    AnonymisationService<ColumnShuffle> anonymisationService = new
     ↪  ColumnShuffleService();
4
5    @Override
6    List<Pair<String, String>> getAnonymisedRows(ColumnOperations oper,
     ↪  List<Pair<String, String>> rows) {
7      ColumnShuffle columnShuffle = oper.getColumnShuffle();
8      if (columnShuffle == null) {
9        return null;
10     }
11     return anonymisationService.anonymise(rows, columnShuffle);
12   }
13
14   @Override
15   protected AnonymisationService<ColumnShuffle> getAnonymisationService() {
16     return anonymisationService;
17   }
18 }
```

Figure 7.1: Java code gets easily verbose.

- achieved results with regard to objectives of the thesis and requirements
- path of further development (eg functional extension . . . )
- encountered difficulties and problems

Table 7.1: A caption of a table is **above** it.

| | method | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | alg. 3 | | | alg. 4, $\gamma = 2$ | |
| $\zeta$ | alg. 1 | alg. 2 | $\alpha = 1.5$ | $\alpha = 2$ | $\alpha = 3$ | $\beta = 0.1$ | $\beta = -0.1$ |
| 0 | 8.3250 | 1.45305 | 7.5791 | 14.8517 | 20.0028 | 1.16396 | 1.1365 |
| 5 | 0.6111 | 2.27126 | 6.9952 | 13.8560 | 18.6064 | 1.18659 | 1.1630 |
| 10 | 11.6126 | 2.69218 | 6.2520 | 12.5202 | 16.8278 | 1.23180 | 1.2045 |
| 15 | 0.5665 | 2.95046 | 5.7753 | 11.4588 | 15.4837 | 1.25131 | 1.2614 |
| 20 | 15.8728 | 3.07225 | 5.3071 | 10.3935 | 13.8738 | 1.25307 | 1.2217 |
| 25 | 0.9791 | 3.19034 | 5.4575 | 9.9533 | 13.0721 | 1.27104 | 1.2640 |
| 30 | 2.0228 | 3.27474 | 5.7461 | 9.7164 | 12.2637 | 1.33404 | 1.3209 |
| 35 | 13.4210 | 3.36086 | 6.6735 | 10.0442 | 12.0270 | 1.35385 | 1.3059 |
| 40 | 13.2226 | 3.36420 | 7.7248 | 10.4495 | 12.0379 | 1.34919 | 1.2768 |
| 45 | 12.8445 | 3.47436 | 8.5539 | 10.8552 | 12.2773 | 1.42303 | 1.4362 |
| 50 | 12.9245 | 3.58228 | 9.2702 | 11.2183 | 12.3990 | 1.40922 | 1.3724 |

# Bibliography


[1] H. Abramowicz et al. 'Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC'. In: *Physics Letter B* 716.1 (2012), pp. 1–29.

[2] Jens Albrecht, Marc Fiedler and Tim Kiefer. 'A Rule-Based Approach to Local Anonymization for Exclusivity Handling in Statistical Databases'. In: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference*. Ed. by Josep Domingo-Ferrer and Mirjana Pejić-Bach. Springer Publishing, 2016, pp. 81–93. ISBN: 978-3319453804.

[3] Jens Albrecht, Marc Fiedler and Tim Kiefer. 'Robust De-anonymization of Large Sparse Datasets'. In: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference*. Ed. by Josep Domingo-Ferrer and Mirjana Pejić-Bach. Springer Publishing, 2016, pp. 81–93. ISBN: 978-3319453804.

[4] Luk Arbuckle and Khaled El Emam. *Building an Anonymization Pipeline: Creating Safe Data*. O'Reilly Media, Inc, 2020. ISBN: 978-1492053439.

[5] Narayanan Arvind and Shmatikov Vitaly. *Robust de-anonymization of large sparse datasets: a decade later*. URL: `https://www.cs.princeton.edu/~arvindn/publications/de-anonymization-retrospective.pdf` (visited on 21/01/2022).

[6] Arshdeep Bahga and Vijay Madisetti. *Big Data Science and Analytics: A Hands-On Approach*. Vpt, 2016. ISBN: 978-0996025539.

[7] Scott Chacon and Ben Straub. *Pro Git*. Apress, 2014. ISBN: 978-1484200773.

[8] Inc. Cisco Systems. *Cisco Annual Internet Report (2018–2023)*. Tech. rep. URL: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf (visited on 21/01/2022).

[9] Inc. Cisco Systems. *The Zettabyte Era Officially Begins*. URL: https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-that (visited on 21/01/2022).

[10] European Comission. *Charter of Fundamental Rights of the European Union*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=EN (visited on 21/01/2022).

[11] European Comission. *General Data Protection Regulation - Recital 1*. URL: https://gdpr-info.eu/recitals/no-1 (visited on 21/01/2022).

[12] European Comission. *General Data Protection Regulation - Recital 26*. URL: https://www.privacy-regulation.eu/en/recital-26-GDPR.htm (visited on 21/01/2022).

[13] European Comission. *General Data Protection Regulation - Recital 6*. URL: https://gdpr-info.eu/recitals/no-6 (visited on 21/01/2022).

[14] European Comission. *General Data Protection Regulation – Art. 83 – General conditions for imposing administrative fines*. URL: https://gdpr-info.eu/art-83-gdpr/ (visited on 21/01/2022).

[15] European Comission. *General Data Protection Regulation – Art. 83 – Lawfulness of processing*. URL: https://gdpr-info.eu/art-6-gdpr/ (visited on 21/01/2022).

[16] Derek DeJonghe. *NGINX Cookbook: Advanced Recipes for High-Performance Load Balancing*. O'Reilly Media, Inc., 2020. ISBN: 978-1492078487.

[17] Inc. Docker. *Docker overview*. URL: https://docs.docker.com/get-started/overview (visited on 21/01/2022).

[18] Josep Domingo-Ferrer and Jordi Soria-Comas. 'Anonymization in the Time of Big Data'. In: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference*. Ed. by Josep Domingo-Ferrer and Mirjana Pejić-Bach. Springer Publishing, 2016, pp. 57–68. ISBN: 978-3319453804.

[19] Inc. Domo. *Data Never Sleeps 6.0*. Tech. rep. URL: `https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf` (visited on 21/01/2022).

[20] Khaled El Emam and Luk Arbuckle. *Anonymizing Health Data: Case Studies and Methods to Get You Started*. O'Reilly Media, Inc, 2020. ISBN: 978-1449363079.

[21] Can Eyupoglu et al. 'An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques'. In: *Entropy* 20.5 (2018), p. 373.

[22] Facebook. *Yarn – Introduction*. URL: `https://yarnpkg.com/getting-started` (visited on 21/01/2022).

[23] Carl Folkesson. 'Anonymization of directory-structured sensitive data'. MA thesis. Linköping University, 2019.

[24] Apache Software Foundation. *Apache Kafka – uses*. URL: `https://kafka.apache.org/uses` (visited on 21/01/2022).

[25] Apache Software Foundation. *Maven – Introduction*. URL: `https://maven.apache.org/what-is-maven.html` (visited on 21/01/2022).

[26] Aris Gkoulalas-Divanis and Grigorios Loukides. *Anonymization of Electronic Medical Records to Support Clinical Analysis*. Springer Publishing, 2012. ISBN: 978-1461456698.

[27] Richie Koch. *Data anonymization and GDPR compliance: the case of Taxa 4×35*. URL: `https://gdpr.eu/data-anonymization-taxa-4x35` (visited on 21/01/2022).

[28] Chris Lee. *Integrating Web Analytics with your SAP Commerce Cloud Storefront - An Overview*. URL: `https://www.sap.com/cxworks/article/2589633935/integrating_web_analytics_with_your_sap_commerce_cloud_storefront_an_overview` (visited on 21/01/2022).

[29] Google LLC. *V8 JavaScript Engine*. URL: https://v8.dev/docs (visited on 21/01/2022).

[30] Bernard Marr. *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. John Wiley and Sons Ltd, 2016. ISBN: 978-1119231387.

[31] Andrew Moore. *The GDPR and Managing Data Risk*. John Wiley and Sons, Ltd., 2018. ISBN: 978-1119487746.

[32] Christoph Nakazawa, Sebastian McKenzie and Jamie Kyle. *Yarn: A new package manager for JavaScript*. URL: https://engineering.fb.com/2016/10/11/web/yarn-a-new-package-manager-for-javascript (visited on 21/01/2022).

[33] Neha Narkhede, Gwen Shapira and Todd Palino. *Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale*. O'Reilly Media, Inc., 2017. ISBN: 978-1491936160.

[34] Petter Ødegård. 'Data Anonymization for Research'. MA thesis. Norwegian University of Science and Technology, 2019.

[35] Nigel Poulton. *Docker Deep Dive: Zero to Docker in a single book*. Independently published, 2016. ISBN: 978-1521822807.

[36] Simi M. S., Sankara Nayaki K. and Sudheep Elayidom M. 'An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity'. In: *IOP Conference Series: Materials Science and Engineering* 225.1 (2017).

[37] Rathindra Sarathy and Krish Muralidhar. 'Perturbation Methods for Protecting Numerical Data: Evolution and Evaluation'. In: ed. by Ranajit Chakraborty, C.R. Rao and Pranab Sen. Vol. 28. Handbook of Statistics. Elsevier, 2012, pp. 513–531. DOI: https://doi.org/10.1016/B978-0-44-451875-0.00019-1. URL: https://www.sciencedirect.com/science/article/pii/B9780444518750000191.

[38] Sanjay Sharma. *Data Privacy and GDPR Handbook*. John Wiley and Sons Ltd, 2019. ISBN: 978-1119594246.

[39] Personal Data Protection Commission of Singapore. *Guide to Basic Data Anonymisation Techniques*. 2018. URL: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pd (visited on 21/01/2022).

[40] Jordi Soria-Comas and Josep Domingo-Ferrer. 'Big Data Privacy: Challenges to Privacy Principles and Model'. In: *Data Science and Engineering* 1.1 (2015), pp. 21–28.

[41] Colin Tankard. *Big data security*. URL: https://www.sciencedirect.com/science/article/abs/pii/S1353485812700636 (visited on 21/01/2022).

[42] IT Governance Privacy Team. *EU General Data Protection Regulation (GDPR): An Implementation and Compliance Guide*. IT Governance Publishing Ltd, 2020. ISBN: 978-1787782495.

[43] Linus Torvalds. *Tech Talk: Linus Torvalds on git*. 2007. URL: https://www.youtube.com/watch?v=4XpnKHJAok8 (visited on 21/01/2022).

[44] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing, 2017. ISBN: 978-3319579580.

[45] Sherali Zeadally and Mohamad Badra. *Privacy in a Digital, Networked World: Technologies, Implications and Solutions*. Springer Publishing, 2015. ISBN: 978-3319084718.

# Appendices

# Index of abbreviations and symbols

| | |
|---|---|
| DNA | deoxyribonucleic acid |
| MVC | model–view–controller |
| $N$ | cardinality of data set |
| $\mu$ | membership function of a fuzzy set |
| $\mathbb{E}$ | set of edges of a graph |
| $\mathcal{L}$ | Laplace transformation |

# Listings

(Put long listings in the appendix.)

```
1  partition fcm_possibilistic::doPartition
2                                  (const dataset & ds)
3  {
4     try
5     {
6        if (_nClusters < 1)
7           throw std::string ("unknown number of clusters"
               );
8        if (_nIterations < 1 and _epsilon < 0)
9           throw std::string ("You should set a maximal
               number of iteration or minimal difference --
               epsilon.");
10       if (_nIterations > 0 and _epsilon > 0)
11          throw std::string ("Both number of iterations
               and minimal epsilon set -- you should set
               either number of iterations or minimal
               epsilon.");
12
13       auto mX = ds.getMatrix();
14       std::size_t nAttr = ds.getNumberOfAttributes();
15       std::size_t nX    = ds.getNumberOfData();
16       std::vector<std::vector<double>> mV;
17       mU = std::vector<std::vector<double>> (_nClusters)
```

```
        ;
18      for (auto & u : mU)
19          u = std::vector<double> (nX);
20      randomise(mU);
21      normaliseByColumns(mU);
22      calculateEtas(_nClusters, nX, ds);
23      if (_nIterations > 0)
24      {
25          for (int iter = 0; iter < _nIterations; iter++)
26          {
27              mV = calculateClusterCentres(mU, mX);
28              mU = modifyPartitionMatrix (mV, mX);
29          }
30      }
31      else if (_epsilon > 0)
32      {
33          double frob;
34          do
35          {
36              mV = calculateClusterCentres(mU, mX);
37              auto mUnew = modifyPartitionMatrix (mV, mX);
38
39              frob = Frobenius_norm_of_difference (mU,
                    mUnew);
40              mU = mUnew;
41          } while (frob > _epsilon);
42      }
43      mV = calculateClusterCentres(mU, mX);
44      std::vector<std::vector<double>> mS =
            calculateClusterFuzzification(mU, mV, mX);
45
46      partition part;
47      for (int c = 0; c < _nClusters; c++)
```

```
48        {
49            cluster cl;
50            for (std::size_t a = 0; a < nAttr; a++)
51            {
52                descriptor_gaussian d (mV[c][a], mS[c][a]);
53                cl.addDescriptor(d);
54            }
55            part.addCluster(cl);
56        }
57        return part;
58    }
59    catch (my_exception & ex)
60    {
61        throw my_exception (__FILE__, __FUNCTION__,
              __LINE__, ex.what());
62    }
63    catch (std::exception & ex)
64    {
65        throw my_exceptionn (__FILE__, __FUNCTION__,
              __LINE__, ex.what());
66    }
67    catch (std::string & ex)
68    {
69        throw my_exception (__FILE__, __FUNCTION__,
              __LINE__, ex);
70    }
71    catch (...)
72    {
73        throw my_exception (__FILE__, __FUNCTION__,
              __LINE__, "unknown␣expection");
74    }
75 }
```

# List of additional files in electronic submission (if applicable)

Additional files uploaded to the system include:

- source code of the application,
- test data,
- a video file showing how software or hardware developed for thesis is used,
- etc.

# List of Figures

# List of Tables