Methods for

Federated Data Analysis

Ruduan B.F. Plug

Supervisors:
Prof. Dr. Mirjam van Reisen
Prof. Dr. Mark Musen

MASTER THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Stanford Center for Biomedical Informatics Research (BMIR)

August 19, 2022

Abstract

Through the framework of data federation, coalesced analytics and modelling can be performed on distributed data without exposing the underlying data. These data are collected, stored and processed at the location where data are being produced, instead of being reposited in a centralized database. Under FAIR principles, standards are defined under which these data can be discovered and accessed as formalized access patterns. These standards provide the principal framework for interoperability and reusability of data, which are essential to federation of heterogeneous data sources.

In this research we build upon these fundamental concepts by expanding the framework of federated data for distributed data sources, which ensures retention of data ownership and provides safeguards for data security and privacy. Central to this is the architecture that incorporates distributed information systems as a composite computing model, which poses unique engineering challenges and inquiry in terms of reliability of aggregates, data quality and convergence of semantics. These challenges form the fundamental groundwork for statistical and computational methods with applications to federated data.

Research Questions

I. *What are the ways in which we can utilize existing statistical inference techniques in order to extend these analyses over heterogeneous federated data?*

II. *How can we utilize and enhance the graph properties of semantic data to enable interoperability over heterogeneous data sources?*

This research thesis has been produced in conjunction with one published conference paper and one accepted journal article.

Plug, Ruduan et al. "FAIR and GDPR Compliant Population Health Data Generation, Processing and Analytics.". SWAT4HCLS (2022). ***[published]***

Plug, Ruduan et al. "Terminology for a FAIR Framework for the Virus Outbreak Data Network-Africa.". Data Intelligence, MIT Press (2022). ***[accepted]***

Foreword

First of all, I would like to express my sincere appreciation and admiration towards my supervisors Prof. Dr. Mirjam van Reisen and Prof. Dr. Mark Musen for giving me the opportunity to perform this research and for providing me support, guidance and many learning opportunities on both a professional and personal level.

Mirjam, thank you for believing in me and introducing me to the VODAN family and the GAIC research group. I have been constantly inspired by the amazing, visionary work and have been constantly challenged by the diverse perspectives being presented. Throughout the year the research group has worked many late nights, through some incredibly challenging times and even conflict did not stop people from believing in the shared vision. Sometimes I do not know how you do it, but time after time you show true leadership, empathy and compassion for those around you.

Mark, thank you for giving me the opportunity to work with the CEDAR group and for letting me dive deep in to the technical infrastructure together with John Graybeal and Marcos Romero, both whose countless efforts and early calls I genuinely appreciate. Without your support, it wouldn't have been possible for us to make such progress in such a short span of time. I also want to thank you for giving me the opportunity to participate like a real peer over the summer with the grant proposal. I sincerely hope to work with your team again in the foreseeable future.

My deepest appreciation and admiration goes out to the entire VODAN family, I have learned so much from you and there are truly too many of you that I have learned from to mention. I want to especially mention Prof. Dr. Francisca Oladipo for her leadership and I want to congratulate her on her new position as Vice-Chancellor at Thomas Adewumi University. In addition

I want to thank the Vice-Chancellor of Kampala International University, Prof. Dr. Mouhammad Mpezamihigo for his support to all of us and it was wonderful to meet you in person. In addition my appreciation go out to all the technical staff members, in particular Mariam, Putu, Aliya, Samson, Getu, Oluwole and Ezra with whom I have worked countless hours and who feel like true colleagues throughout this journey.

Finally, I want to thank my parents from the bottom of my heart. Without their support I wouldn't have been able to get this far. As a first generation student in my family, attending a university has always seemed like an insurmountable challenge. With their support I was able to take on this challenge, and even though I met quite some difficulties along the way, I was able to pull through and show my true potential. Thank you for always standing by me and believing in the best of me.

Contents

1 Introduction

> *Simplicity is a great virtue but it requires hard work to achieve it and education to appreciate it.*
>
> **— Edsger W. Dijkstra**
> *On the nature of Computing Science*

Many of the methods we use in empirical scientific research are based on analysing granular data. These data form the basis from which evidence can be derived in order to evaluate a theoretical model of the real world. There are many settings however in which exposing such granular data is not desirable or permittable [1], often supported by legislative measures and ethical guidelines [2].

This is a considerable problem, as direct access to these data is a prerequisite for application of any frequentist statistics. These granular data can be used to make assumptions, construct models and test hypotheses. The most parsimonious and commonly used approach in practice is that of distributed data [3]. We illustrate this approach in Figure 1.1. In this approach we directly take data from multiple instances and append these data together. Then we can continue just as we would perform any other analysis to approach our research problem.

This approaches provides several advantages, such as being able to test subsets of these different data sources or the ability to compare data sources. This is viable approach when the data sources are public domain, all inside the same legislative area or are all owned by one party. However, this is not always the case.

One of the possible alternatives for this is to use synthetic data [4], which acts as a surrogate to the original data. There are many options to generate these data based on the type and complexity, which can range from sampling a distribution to deep leaning techniques such as generative diffusion models.

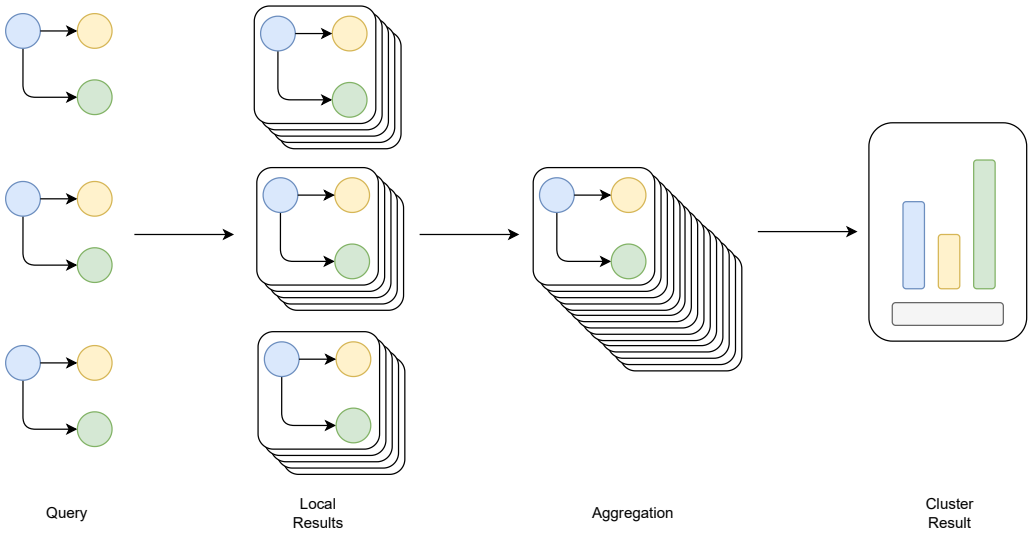


Fig. 1.1.: A representation of a distributed data approach to research queries

However what these techniques have in common is that they are no longer directly sampled from the original data, and as such information is lost and questions are raised around the reliability of such methods when used to make inferences [5]. Instead, we consider how we could devise a technique that can use the original data to answer research questions, without actually directly using these data ourselves. This brings us to the concept of federated computation, which proposes to relegate all initial computations to the most localized level [6]. This can range from simple aggregation to fitting statistical models and co-training machine learning models. In this research we will focus on the application towards statistical models, which is purposefully chosen as we seek to support analyses that discover causality.

We can then further ask ourselves how this technique may be applied to a multitude of data sources. This is advantageous as this may increase availability of data, increase numerical stability of our models and ultimately reduce bias and improve reliability [7]. Instead of sending our compute requests to a single instance, we can also sent these out to a cluster of data sources that all reposit their data locally as we showcase in Figure 1.2. These compute requests are handled locally as well, and from each instance only permissible results of these computes are served out for aggregation or comparison.

This raises several challenges. The first challenge that presents itself is the interoperability of the different data sources. We must be sure that the results are comparable if we want to perform reliable studies. Another

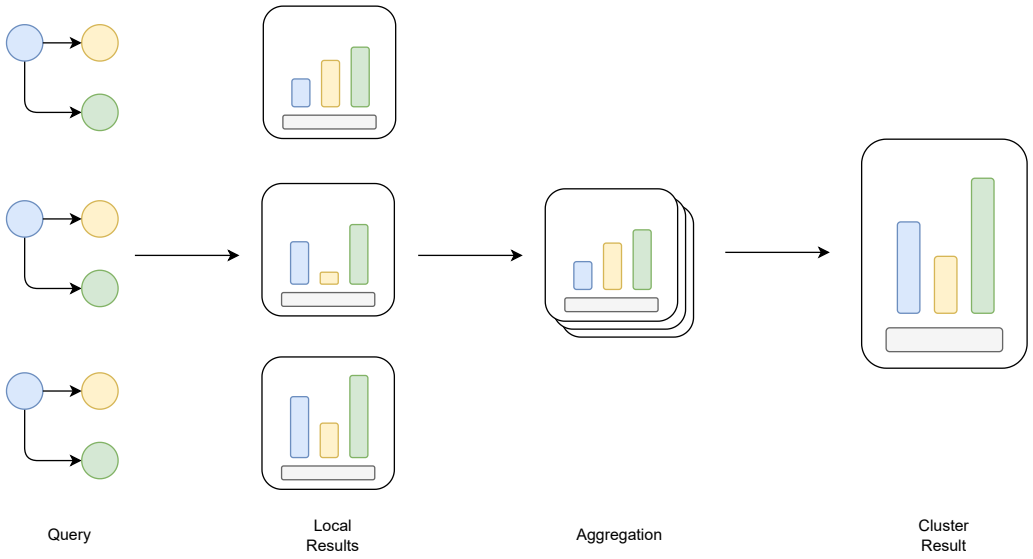


Fig. 1.2.: A representation of a federated data approach to research queries

challenge is the ability to aggregate or compare models. A single statistic can be easily aggregated over any number of instances, but an actual model may require model-specific techniques [8]. Both of these aspects cannot be achieved through analytical approaches alone. It becomes evident that in order to develop the methodology around this broad technique, we cannot simply rely on statistical or computational techniques alone. Instead, we propose to use computational techniques to harmonize data across different sources through semantics, and use composite analysis techniques to perform statistical inferences on harmonized data.

1.1 Research Approach

The focus of the research is based around incorporating multi-disciplinary methodologies, where we combine and optimize approaches from different fields including statistical science, computer science and semasiology in order to develop an overarching understanding and approach. Thus the main breadth of application lies on consolidating the proven methodological techniques from these fields by selecting the appropriate approaches and defining the methods and conditions on which these can be used in unison.

We identify a research gap as the lack of a rigorous framework to connect current techniques that support federated data. Current techniques within this field require strong assumptions to be met, such as assumed interoperability,

and do not present a method that incorporates the required techniques from the point of data generation to the actual analysis [9]. For bringing federated data into practice this is very important, as enabling the effective use of such practice requires a design and architecture that supports federated data techniques from the ground up.

We seek to build upon a knowledge framework where we can cross the thread all the way from the method that is used to generate data, as the initial starting point, to the point where data is analysed as the operational process. To study this and be able to achieve this outcome, we will be performing a methodological study, as in we will be laying a theoretical foundation based on existing and proven works to build a new combined methodological approach.

This methodological study will consist of generalizing techniques from the statistical sciences to be applicable to computational methods. In addition, we will study how we can incorporate semantic embeddings within our computational approach to support routine computations that are interoperable across selected instances.

1.2 Related Work

The primary context that we are interested in is that of federated data and architecture [6], which has seen rapid developments in the recent years [7] when the concept of federated learning was developed. This also brings many new challenges and opportunities, which are summarized in the paper *'Advances and Open Problems in Federated Learning'* by Kairouz et al. [9]. This paper specifically relates to some of the open challenges in interoperability, variability and reliability.

Since federated data provides various advantages in terms of privacy and security, it has seen application in various domains. Applications and related challenges are identified in the domains of health [10, 11], industry [12], automotive [13], smart cities [14] and the internet of things (IoT) [15].

What all of these review papers have in common is that they all address the same general theme that Kairouz addressed, with specific nuances for each domain. For example IoT sees a very dispersed network of data sources where

the process itself cannot be easily modified, so this requires a very up-front approach on assuring that devices are interoperable. Industry is however more flexible, but sees greater risks for data exposure. The domain of health is on the extreme end of security and privacy requirements, where any exposure of data could lead to right violations and serious consequences.

The general framework that has been proposed to approach the problem of heterogeneous data is the use of FAIR [16], specifically the application of FAIR data points [17]. A standardized approach for interoperability is a commonly stated necessity across the review papers that we have identified. This framework can be combined with methods for semantic interoperability, which is discussed in the health domain specifically by Gansel et al. [18]. By properly defining and structuring semantic metadata [19] to form knowledge graphs we can perform automated reasoning [20] to deal with interoperability over heterogeneous data and cross referencing [21] various sources to analyse concerns around reliability.

1.3 Contributions

In this manuscript we address two distinct research questions that are fundamentally linked to one another. These are as follows:

Research Questions

I. *What are the ways in which we can utilize existing statistical inference techniques in order to extend these analyses over heterogeneous federated data?*

II. *How can we utilize and enhance the graph properties of semantic data to enable interoperability over heterogeneous data sources?*

To prelude these research questions, we have provided a wide coverage of background material on the practical and sociological consequences of increasing usage and reliance on data. Specifically, we have provided original context on the advent of the semantic web, a deep dive in to the philosophical background of data ownership, the importance of embedding meaning in to data and the advent of federated data. This provides context for the two broad research questions we have covered.

In the first methodological part of this manuscript we covered the inquiry towards the statistical techniques that can be utilized heterogeneous federated data sources. Specifically, we have looked at techniques that generalize well over a wide variety of possible situations. This is what we have considered an inclusive approach, which reinforces the strength of federated data approaches as these may provide a good avenue to perform research across regional and legislative boundaries while adhering to compliance and ethical standards.

Specifically, our contributions in this regard are expanding the technique of multiple ψ distributions as composite distributions for federated analysis. These composite distributions are in essence piece-wise functions of which the parts are approximated as partial distributions using generalized distributions with Dirac-δ composition. In addition, we have shown a way to apply this generalized technique, which is challenging to perform traditional hypothesis tests with, with the Szymkiewicz-Simpson measure to formulate a generalized method as an application of federated analytics which can be utilized to detect differences between complex, aggregated distributions such as composite distributions.

In the latter half, we have expanded this framework of federated analysis by going from the assumption of interoperability to the actual practice of making data inherently interoperable. The technique which we have proposed incorporates the usage of the FAIR data guidelines with techniques from semantics to harmonize different data specifications towards a universal query format. We have shown the link between FAIR data as graph data and the link to fundamental computer science, such that we can perform graph algorithms to discover interoperability opportunities.

This demonstrably enables the use of automated methods to reliably and efficiently identify viable data sources, which may aid in answering research questions without the extra overhead and potential for information loss involved in pre-processing or transforming the original data. In addition, this technique allows each locale to preserve privacy and data ownership, which is an essential property of federated data techniques. Finally, we define a localized data architecture that incorporates the aspects of FAIR data and services as federated data points, which bring the potential of federated analysis in to practice by automating the generation of FAIR-based data and enabling secure, privacy-preserving queries towards these federated data.

Background

2

> *Data is a precious thing and will last longer than the systems themselves.*
>
> **— Tim Berners-Lee**
> *A Framework for Web Science*

Since the 21st century, data has played a pivotal role in social and economic development across the globe. We are now in what is ubiquitously considered the information age. Information that we formalize as knowledge is a product refined from the data that we collect and store. In essence, data can be viewed as a resource [22] that may be harnessed to produce value. These data can range from traditional scientific measurements and samples, to more abstract data formats such as images, audio and written text.

Most of the data we produce however are not being refined or harnessed to produce value as refining data has an opportunity cost attached [23]. The sheer amount of data produced means that we need to be selective in the data that we process further. Sometimes that is for a an evident reason, for example use of sensitive data outside of the direct intended purposes exposes risks to privacy and security.

Many of our daily interactions and transactions interact with existing data, and they in turn bring into motion the production of new data. At the global scale, these data are large in volume, diverse in scope and are being produced at increasing granularity [24]. Data at such scale brings many opportunities to extract information and produce knowledge, but also poses new risks and challenges for society to deal with. These challenges can come in many shapes, such as technological limitations, regulatory compliance, privacy risks, security concerns and lack of interoperability [25].

The core problems we are looking at in this research pertain to issues related to the veracity and granularity of data used to model phenomena. First, we will tackle some of the core challenges and developments that will set the conditions for the use case of a federated data ecosystem.

2.1 The Semantic Web and the Challenge of Scale

The vast majority of data is exchanged in real time through the internet, the globally linked network of computational entities that serve out data and web services. The communication between these entities is made possible using a uniform syntax specification by the W3C [26], the machine readable protocols that we know as the Hypertext Transfer Protocol (`http`) and the encrypted variant known as Hypertext Transfer Protocol Secure (`https`), which operate over the Transmission Control Protocol/Internet Protocol (`TCP/IP`). These protocols determine the way a request is propagated and handled through the world wide web, which is principally designed to connect systems together through various hierarchies of routing hardware.

However, with the increasing scale of the internet and increasing veracity of data and services, the limitations of this paradigm are increasingly apparent [27]. While these existing protocols standardize connectivity between entities and the logical exchange of data, the data and services themselves that these entities provide and exchange do not necessarily adhere to any standard. And while we have commonly defined file types, for instance a `.csv` file indicating a table of comma separated values or `.png` referring to a pixel matrix encoding standard, there isn't any standardized way to convey meaning about or relations between these data.

The visionary behind the world wide web as we know it today, Tim Berners-Lee, has authored many of the standards that are foundational to networking and data exchange. However, his idea of the world wide web didn't end with the machine readable syntax that is used ubiquitously to this day. Before the internet was conceptualized, there was the idea of the semantic web [28]. It can very much be seen as an old solution to a very relevant new problem, that of the exploding volume and veracity of data.

The semantic web is a framework that operates at a higher level than physical infrastructure, which instead fundamentally links together the data that is ultimately transported over the infrastructure layer. We consider that all transactions and computes performed across the internet are ultimately comprised of data exchange, whether this is requesting a web page, querying a database or interacting with a web application. If we link together meaning-

ful relations between these data that can be traversed by machine readable standards, we can implement the core principle of the semantic web.

According to Tim Berners-Lee, in practice the semantic web isn't a fundamentally different concept from the current world wide web:

> *" The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."*

In essence, the main challenges pertains to rigorously defining the way we give meaning to data to produce information, which can in turn be linked together to form knowledge. One of the primary challenges in bringing these definitions in to practice this is that while it is relatively easy to get people to agree on standards for communication by embedding them in to our technology, i.e. using a certain structure around your message, it is much more challenging to get people to agree on standardizing the content of the message itself [29].

One of the ways that this problem is approached is by standardization of the way that data standards are developed. This models the way that ontologies are developed, by defining terms within controlled vocabularies and semantics through linkages, you can formalize knowledge as a graph structure [30]. These structures can then be written using a default syntax for knowledge, such as the Resource Description Framework [31] and the eventual development of the Web Ontology Language [32]. Such developments do not take away from potential disagreements on metadata specifications, semantics and use of specific terminology, but at least provide a language description whereby knowledge representations can be made machine readable.

The importance of these developments cannot be understated with the growing demand towards services that make use of data [25]. Most of these data pertain to properties and activities of humans, data which is frequently replicated across services, e.g. personal details, and containing sensitive information. With the exponential growth of the amount of data stored across services, it is increasingly hard to keep track of replicates of personal data. Ensuring interoperability between an increasingly expanding amount of different data sources takes increasing effort to the point where traditional data processing techniques are rapidly becoming infeasible.

2.2 On Ownership and Protection of Data

The idea of property, which may be further subdivided into public, collective and private property, is a concept that dates back to ancient philosophy and is still ever relevant when dealing with data. The ancient Greek philosopher Aristotle underlined the importance of private property ownership in the work Politics, where he argues that shared property leads to an increase in disputes, and that shared property is more likely neglected when compared to private property. This is at odds with the argument of Plato in his work Republic, whom argues that the commons in society are only served well when the state properly controls and allocates property for all to benefit, which is a distinct argument against private property ownership.

The 19th century author on economic theory, William Forster Lloyd, recognized the underlying dilemma of the commons [33]. What was privately owned was well-maintained, but benefitted the few, while that which benefitted the many, was often neglected in pursuit of endeavours more beneficial to oneself. The problem of communal neglect is one of the primary reasons why such services have been centrally controlled by a state entity throughout both human history and across different cultures [34, 35].

These two perspectives are often at odds and are still relevant to this day, but with technological and civic advancement we have new opportunities to leverage the best of both perspectives: to ensure property ownership for individuals while allowing for common benefit within society with lesser centralized control. This idea is fully embraced in the idea of social entrepreneurship [36], a type of business where the private exchange of goods or services leads to benefit even to third parties with no external costs. The collective of such organizations act as the communal services once critiqued by Aristotle, and seen as inviable by Plato, yet are increasingly thriving within modern society.

The 21st century Nobel prize winner Ostrom demonstrated this collective, decentralized self-regulation in regard to common resources in her book Governing the Commons [37]. It is clear that top down governance is not a necessity to establish beneficial commons, and that private property ownership is not the only way for individuals to pertain interest in sustaining goods and resources.

The idea of personal data ownership starts at the civic level, as personal property is a fundamental principle that is shared ubiquitously across human societies. Consequentially, as ownership is such a fundamental principle, it has also been an important consideration for data even prior to the advent of the world wide web in 1983. For instance, considerations over data protection date back to the 1960s and the first law pertaining the protection of private digital data was conceptualized in 1970 in the German state of Hesse [38].

The historic progress of ideas on ownership over physical property draws parallels to current thoughts on digital data. Where we initially adhered to a centralized approach to data management, where responsibility over data was primarily held by those storing the data and offering only value to those holding those data, we are seeing a move towards ideas that support a decentralized model that distributes these responsibilities, and provides more opportunities to enable individual contributions to the commons through federation without exposing that individual to risks or costs that may have traditionally existed within a centralized model, one of these technological developments is that of federated data [7].

Many of the quintessential services and interactions that we rely on in our day to day life, depend on complex networks of digital infrastructures to function. Consequentially, the internet has become a fundamental part our lives in the information era, with over 4.66 billion people globally actively using the internet on a day-to-day basis [39]. The digital space is so integrated in our daily lives that a part of our identity, our persona, exists on the internet and is connected to a wealth of data that describes us, our society and the world around us.

It is that data, that links us to the digital space, that is both the biggest weakness and the greatest strength that we currently possess as a society. Data itself can be seen as the new metaphorical gold of the digital age [22], a resource that can be so valuable when used properly, that global organizations are investing billions into their data infrastructure and analytics operations just to harness a fraction of what is being produced through online interactions. In contrast, these data also pertain a wealth of sensitive, private information that could do irreconcilable damage if these were to be exposed to the public domain.

As such, it is of utmost importance that data pertaining to individuals are handled with care. But how do we guarantee that this is the case? How do we set and conform to standards for data security, and how do we even know, as an individual, what kind of data is produced or shared about us? With digital data processing being virtually ubiquitous for access to online services, and the increasing complexity of the internet, these questions are increasingly hard to answer, and consequentially, it is also harder to contain spread of sensitive information after a data breach has occurred.

There are ample examples available over the dangers of data breaches, which are thoroughly examined in the 2021 Data Breach Investigation Report conducted by Widup et al. on behalf of Verizon [40]. Evidence suggests that the vast majority of data breach incidents have an economic nature, and are primarily conducted through social engineering, misuse of privileged access and physical data theft. In addition, further evidence suggests that approximately 1 out of 5 data breaches are not discovered until months after they have occurred, in which harm to individuals may already occurred without the possibility to take precautionary action. It is clear that still to this day, there is a lack of transparency and control over personal data flow, which leads to collective losses to society as a whole. The European Commission reported that the total cost of cyber-crime, of which the majority of the reported incidents pertain ransom or theft of centralized data, had a total negative impact of €5.5 trillion to the global economy [41].

To bring this into perspective, the Thomson Reuters Foundation has determined that the total global investments required to meet the Paris Climate Agreement by 2030 are estimated to be $5 trillion on an annual basis [42]. It is for this reason that we should not, under any circumstances, underestimate the importance of increasing the resilience of our digital infrastructure. With the right technology and practices, we not only make our digital space more sustainable, but this will also allow global economies to allocate more resources to meet the sustainable development goals that will impact all future generations to come. Evidently, if we wish to achieve a sustainable ecosystem for a secure digital identity, a paradigm shift in data processing is required. The General Data Protection Regulation (GDPR) established in 2016 by the European Commission provides a strong basis for a new framework on sustainable data management, but to truly tackle this problem we require not only legislative action, but also technological safeguards on our private data.

2.3 The Decoupling of Data from Meaning

The main mechanism for data to provide value in the real world is through recombination and refinement towards application, but data without inherent context or meaning cannot be used towards these purposes. Data without meaning, i.e. without any form of annotation or semantic embedding, are in essence not that much different from structured noise. What provides value is the combination of data and the metadata descriptors that allows us to understand what these data mean, bring these data in to practice and to place these data in context with other data.

This issue presents itself more commonly than one may at first realize, for instance take a study where a large array of sensors produce data, which are collected for further analysis. Within this study, the data has a clearly defined meaning to the researchers and they can specifically apply it towards their research goals. However, if this data were to be presented as an independent data set, this meaning would be lost. Adding baseline metadata such as column names and time stamps would be insufficient to be able to retrieve the real meaning and context of such data [19]. This in turn limits the ability of further processing this data in to information, and thus limiting knowledge generation.

Another example can be found in the commonly used technique of data scraping, where through an automated process data are extracted from the web according to a specific rule set. The data that is being scraped is essentially a subset of the total data that is being transmitted, for example a line containing a specific keyword from a text, the outward links on a page or an image from a specific cell.

With methods that are commonly use to bundle together scraped data, not only is the surrounding context lost, but typically the lineage and provenance used to generate these data are not properly recorded [43]. This means that there may be less confidence in any inferences from these data and that there is no way to verify subsequent analyses when interpreted as being publicly sourced.

In the book Understanding Variation [44] by Dr. Donald Wheeler, one of the key texts on data process control, he highlights the importance of embedding meaning with data:

> *"No data have meaning apart from their context."*

This statement does not come from mere conjecture, but is based on the foundations of the work of Dr. Walter A. Shewhart, whom laid the groundwork for the field of statistical quality control (SQC). Within SQC he formulated two of the core principles that signify the importance of context [45], which can be generalized to any process that is centered around the application of data to make observations, predictions or decisions.

1. Data should always be presented in such a way that preserves the evidence in the data for all of the predictions that might be made from these data.

2. Whenever an average, range, or histogram is used to summarize data, the summary should not mislead the user into taking any action that the user would not take if the data were presented in a time series.

Shewhart's first rule is very centered around the concept of data provenance and lineage, while the second rule focuses on the representation of the meaning of data. The practical meaning of data is not just isolated to a factual statement of purpose, i.e. through metadata or a descriptor, but may also vary based on the observer.

2.4 The Federated Data Methodology

The concept of federated data dates back to 1985 with a publication by Heimbigner & McLeod on a new type of distributed architecture that could be used to construct a decentralized information system that samples from multiple sources with different permissiveness [46]. This was introduced as a federated database, which features a virtualisation layer that compresses data from multiple autonomous sources down into a single repository that uses a federated dictionary that indexes the external sources. The concept

of using a single dictionary to manage a scalable cluster can be leveraged to coalesce and query data from disparate data sources [47].

In 2016 Konečný et al. published the paper "*Federated Learning: Strategies for Improving Communication Efficiency*" [7], which introduced the concept of constructing models that use a high volume of data locally, and transferring those models instead of data to perform distributed analyses. Initially this paper focused on applications where bandwidth was limited, as this technique could essentially compress training data by using a trained model as a representation of the data.

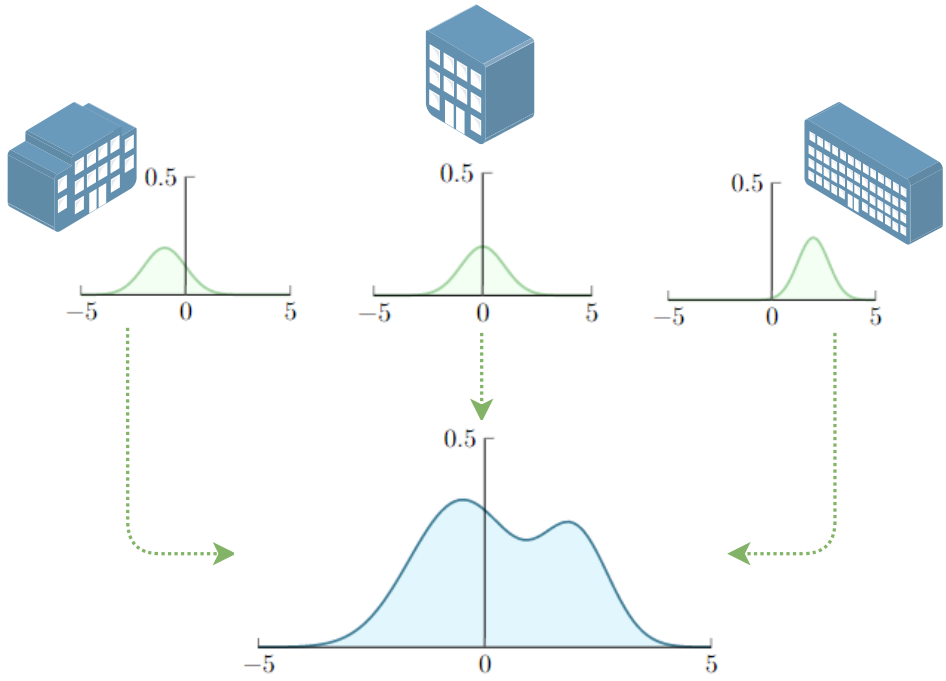


Fig. 2.1.: A representation of a federated data cluster producing a remote composite distribution model from individual local distributions

With the computational field increasingly focused on distributed computing, this sparked renewed interest in the idea of federated data. Such applications typically work on decentralized data that are standardized to a single, uniform data format. Building models locally, and then unifying those models from uniform data formats, can be used to coalesce models without exposing the underlying data. We show an example of this in Figure 2.1.

However, with computation across different heterogeneous data sources comes a significant overhead in data processing. Recent applications in federated learning, such as multi-institutional collaboration [48] have noted

the extensive pre-processing that was required to standardize the data sources before remote computational tasks could be performed.

The problem of data standardization and interoperability is not new, and has been noted by Berners-Lee with the inception of the semantic web [49]. This challenge is just not limited to the machine-readability of underlying data, but central to this issue is the semantic interoperability of associated metadata elements [18]. A crucial element of the semantic web is the use of persistent and globally unique identifiers for resources [50], which allows for unambiguous querying for specific data. This consequentially means that any given semantic concept expressed through data is linked through a unique identifier.

Pivotal to federated learning is decentralization of data, where the model is remotely trained without access to data. This enables secure and privacy-preserving computation [51]. The strong requirements on data standardization for federated learning poses serious challenges when there is no direct data access, which requires standardization at time of data generation or data processing. The usage of semantics provides a key component to harmonize data from different data sources through computationally unique, persistent identifiers. Within research domains, this can be implemented by using domain-specific ontologies [52, 53], based on dynamic controlled vocabularies [54] that form unique semantic identifiers. Data that is generated based on the same ontology, even if their implemented structure is heterogeneous, can be converged towards a uniform format by using the properties of node class equivalence and graph isomorphism.

Part I

Statistical Methods

> *The greatest value of a picture is when it forces us to notice what we never expected to see.*
>
> **— John W. Tukey**
> *Exploratory Data Analysis*

With the sheer scale of modern day data production, it is attractive to leverage volume to approach problems and answer questions. However, this also brings many problems. Data can have quality issues, may be biased, may not fully explain a phenomena and is typically sampled from from a limited frame. Many of the statistical methods that have already been developed can be used to identify and approach some of these contemporary issues.

Specifically, we are interested in methods that apply to data from distributed sources; because the current data landscape isn't just large in volume, but also spread across many different sources. If we want to leverage and combine multiple sources, especially when preserving privacy and data ownership, we have to take extra precaution to ensure the data is appropriate in structure and quality to be able to provide reliable answers to our queries in scientific investigation. In this section we show the equivalence between mixture and composite distributions, formulate a generalized model fitting technique for composites and apply these techniques to construct composite hypotheses which can be used to perform analyses.

3 Composite Distributions

The first step of any data-centric methodology is to build a quantitative hypothesis. This is typically structured around an existing operational process or a hypothetical scenario. The hypothesis generalizes and specifies a model of compounded factors within the process or scenario space and implicitly places a decision boundary, essentially discretizing from the model. In order to evaluate the hypothesis, we need to control for factors that we specify as experimental variables and gather experimental outcomes.

From the chosen control, the null and alternative hypotheses follow and between them lies the decision boundary, which is known as the critical value. Typical hypotheses are bounded to a binary decision, for instance you may have a hypothesis to test for different outcomes for a treatment protocol. However, hypotheses ultimately rely on the underlying distribution. If the exact distribution is unknown or of some composite type, we need to apply different techniques in order to test a hypothesis.

3.1 Definitions

The process of hypothesis testing involves building the evidence required to either reject or fail to reject a null hypothesis. Evidence is in the form of data that has been sampled from a (specific) population $\mathcal{P}$. However, when performing analyses on empirical data we should not only be looking at the data themselves, but also consider the context surrounding the data. The context is defined by the population from which data has been sampled, the sampling methodology that was utilized and the subsequent application of the data to answer hypotheses.

Whenever we sample data, we are dealing with many unknown factors. This is why the sampling process behalves like a stochastic process [55], where each time we sample a data point from some population we get a different result. In the totality of the process of hypothesis, sampling and modelling we are concerned with four distinct statistical spaces:

Composite Set Spaces

- **Hypothesis space** $\mathcal{H}$
  The set of all possible hypotheses $H_i \in \mathcal{H}$ within the chosen experimental frame that are congruent across the partial distributions.

- **Sample space** Ω
  The set of all possible sampled support points Ω belonging to any individual partial distribution $s_i \in S_j$ of the composite distribution in an experiment.

- **Feature space** $\mathcal{F}$
  The set of all possible feature properties $f_i \in \mathcal{F}$, often weighted by some weight π_i that can be derived from the sample space.

- **Parameter space** Θ
  The set of parameters θ that a stochastic process can functionally map, e.g. n-dimensional to 2-dimensional mapping $f : \theta_1 ... \theta_n \rightarrow \mathbb{R}^2$.

At the very essence the hypothesis space forms the set of questions that we can ask in our experiment and the sampling space contains all the possible outcomes. To answer these questions using our outcomes, we need to know how our experimental parameters effect the outcome and how specific features [56] of the outcomes are mapped on to the hypothesis.

A data complex generating process from which data points can be derived can be defined as a collection of random variables $X_1, X_2, ... X_n$, which together follow some composite distribution. Let ψ be some unknown distribution from the distribution family Ψ and $\{\theta_1, ... \theta_n\} \in \Theta_\Psi$ be the parameterization of ψ, we can define a single random variable generating points X as the projection $X : \Omega_\psi \rightarrow \mathbb{R}$.

Given that X follows some parameterized distribution as $X \sim \psi(\theta_1, ... \theta_n)$, we let the composite be a collection of n-dimensional points derived from the individual random variables and their parameterization as column vectors $(X_1^T, ... X_n^T)$ assuming $X_1 ... X_n$ are independent and identically distributed random variables.

3.2 Gaussian Mixture Distributions

Take a data generating process over $\Psi_{\mathcal{N}}$ as a mixture of Gaussian distributions $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and $X_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$ we define the Gaussian probability density function $\forall x \in X, \mathcal{N}_X(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$. Given parameterization sets $\mu_i = \{-1, 0, 2\}$ and $\sigma_i^2 = \{1, 1, 0.75\}$ we get the distribution shown in Figure 3.1.

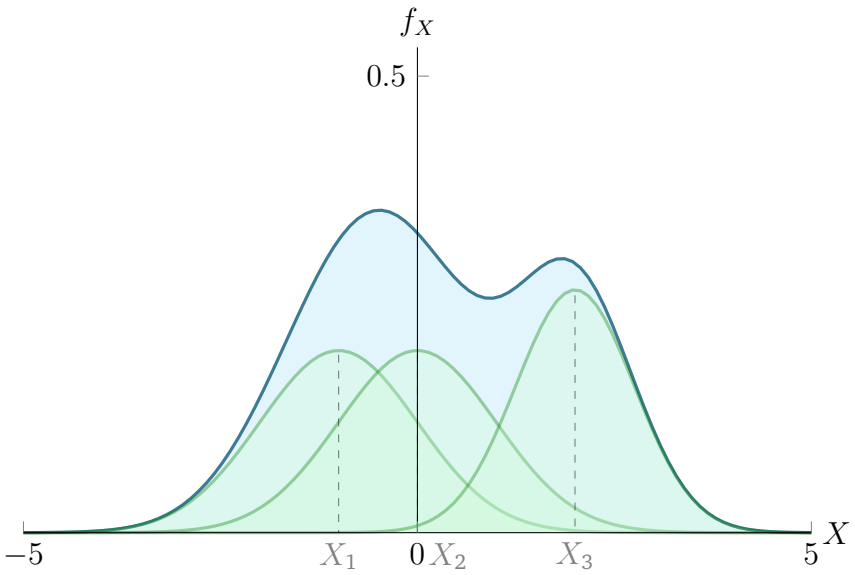


Fig. 3.1.: The composite point frequency distribution of X_1, X_2 and X_3

In the above figure we see the specified random variables X_1, X_2 and X_3 indicated in green together form a complex composite distribution as shown in blue. In this example, the resulting distribution is a bimodal Gaussian mixture [57]. This composite indicates the likelihood distribution of the points that we can sample when randomly choosing from X_1 through X_3, with equal probabilities for all distributions from the mixture if we draw from the composite.

In this case the composite distribution with even weights would be defined as $X_{\mathcal{C}} = \frac{1}{3}\mathcal{N}(-1, 1^2) + \frac{1}{3}\mathcal{N}(0, 1^2) + \frac{1}{3}\mathcal{N}(2, 0.75^2)$ such that total probability $\int_{-\infty}^{\infty} X_{\mathcal{C}} = 1$, which can be used as a sampling distribution to simulate the composite of the three individual distributions. It is evident that this definition only holds if the probabilities are independently distributed. Note however, that this is distinctly different from joint distribution sampling, which samples from the i.i.d. random variables at the same time as n-

dimensional set data points. Sampling from a composite only returns a single data point as all distributions reside on the same dimension and are aggregated from distinct partial distributions.

3.3 Generalized Composite Distributions

As the law of total probability states that the total probability over the entire sampling distribution needs to be one, we can define a generalized rule that holds for all distributions: $\forall\psi : \int \psi = 1$. When we consider composite distributions, we get mixtures of partial distributions $X \sim \psi$ that already adhere to the law of total probability to be valid probability distribution. It then follows that we can weigh each of the partial distributions such that the summation of all the partials once again equals 1.

For any possible composite distribution we can then define for the probability density function as the sum of partial distributions.

$$p(x;\Theta) = \sum_{i=1}^{n} \pi_i \psi_i(x \mid \Theta_i) \tag{3.1}$$

For this equation the law of total probability holds as the sum of weights $\pi_1 ... \pi_n$ that $\sum_{i=1}^{n} \pi_i = 1$. In reference to the composite distribution, the weights indicate the expected probability that a random sample from the composite comes from a specific partial distribution.

If a distribution is generated from a sub-sample of data points from some unknown distribution, then the known discrete set of data points form the support [58]. The support of each partial distribution, by means of clustering, in contrast to the total support forms the probabilistic weighting π_i of ψ_i assuming some data point $x \in X$ and distribution parameterization Θ_i. The support plays an important role in measuring and estimating the reliability and confidence of data sources when evaluating composite models from distributed data.

By formulating the support as the probability of distribution i being drawn from, where i is one of the n individual partial distributions, we can formu-

late the link between mixture distributions and composite distributions as follows. First we take the composite probability density function in $\forall\psi_i \in \Psi$ with parameters Θ_i then we get the following composite probability density function.

$$p_c(x;\Theta)\begin{cases}\pi_1\psi_1(x \mid \Theta_1)\\ \quad\vdots\\ \pi_n\psi_n(x \mid \Theta_n)\end{cases} \tag{3.2}$$

Now we normalize $\pi_1 ... \pi_n$ to the support ratio between the partial s_ψ and the total distribution S_ψ, such that composite $p_\Theta(x)$ follows the law of total probability. We draw a selection parameter ϕ from a uniform distribution as $\phi \sim \mathcal{U}(0,1)$ to perform the selection.

$$p_c(x,\phi;\Theta)\begin{cases}\frac{s_i}{S_i}\psi_1(x \mid \Theta_1), & 0 \leq \phi_1 < \phi_2\\ \quad\vdots \quad , & \quad\vdots\\ \frac{s_i}{S_i}\psi_n(x \mid \Theta_n), & \phi_{n-1} \leq \phi_n \leq 1\end{cases} \tag{3.3}$$

Under the assumption that for a composite distribution the ratio $\frac{s_i}{S_i}$ is equivalent to the selection weights π_i, requiring the distributions to be i.i.d., we can assume that $p_c(x,\phi;\Theta)$ is equivalent to $p(x;\Theta)$. Thus any mixture distribution where selection is performed on the magnitude of its cardinality ratio is congruent to a composite distribution.

4 Fitting ψ-partials to Composite Distributions

To build a composite model we need to estimate the number of partial distributions n, the type of distributions $\psi_1 ... \psi_n$, the parameterizations of the individual distributions $\Theta_1 ... \Theta_n$ and their weights $\pi_1 ... \pi_n$. However in a typical situation we can only estimate the complete composite distribution when sampling, the partial distributions that form the composite are unknown [59]. If we assume a composite is a Gaussian mixture, we can approach this by using an expectation–maximization (EM) algorithm [60, 61].

4.1 The EM Algorithm

With the EM algorithm we perform an iterative optimization procedure in order to find a set of distributions Ψ and their parameters Θ that result in a locally maximum likelihood towards a sample of data from the composite distribution. [62]. First we define the density mass function as the slope of the probability curve. We assume we draw samples x from a continuous random variable with sampling distribution $X : \Omega \to \mathbb{R}$ such that for any specific point $\mathbb{P}(X = x) = \epsilon$ and $\int_{x_0}^{x_1} \mathbb{P}(X) = 1$ where the closed interval is bounded by x_0 and x_1.

Then we have the parameterized probability density function by taking the derivative over the interval $\mathbb{D} = [x_0, x_1]$ by using $\psi(x, \Theta) = \lim_{\delta x \to 0} \frac{\mathbb{P}(X \leq x_0) - \mathbb{P}(X \leq x_0 + \delta x)}{\delta x}$. Since for a continuous function any singular $\mathbb{P}(X = x) = 0$, we find $\psi(x, \Theta) = \frac{\delta \mathbb{P}(X \leq x)}{\delta x}$.

From this we can find the expectation of any continuous ψ parameterized with Θ as $\mathbb{E}[\psi|\Theta] = \int x\psi(x\ \theta_0 ... \theta_n)\delta x$. For the likelihood $\mathcal{L}$ of a mixture of n distributions and parameterisations we use the probability distribution we have previously defined $p(x)_\Theta = \sum_{i=1}^{n} \pi_i \psi_i(x \mid \Theta_i)$, then $\mathcal{L}(\Theta_1, ..., \Theta_n) = p(x_1, ..., x_n; \Theta_1, ..., \Theta_n)$. Assuming that the distributions are i.i.d we can use

the product rule such that $\mathcal{L}(\Theta_1, ..., \Theta_n) = \prod_{i=1}^{n} p\left(x_i; \Theta_1, ..., \Theta_n\right) = \prod_{i=1}^{n} p_{\Theta}\left(x_i\right)$, which we can substitute back to get the likelihood function.

$$\mathcal{L}_{\Theta} = \prod_{i=1}^{n} \sum_{j=1}^{k} \left(\pi_j \psi_j(x_i \mid \Theta_j)\right) \tag{4.1}$$

Using a log-transform we can eliminate the product, which will provide the same solution space as the log-transform is monotone a transformation. In other words, the $\arg\max_{\Theta} \mathcal{L}_{\Theta}$ provides the same parametric solutions for Θ as $\arg\max_{\Theta} \log \mathcal{L}_{\Theta}$. Then we can use the transform to get the log-likelihood that we can optimize.

$$\mathbb{L}_{\Theta} = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} \left(\pi_j \psi_j(x_i \mid \Theta_j)\right)\right] \tag{4.2}$$

Since we have a double summation, we are essentially constructing a diagonal matrix of n data points and k partial distributions. One of the ways to approach this is to computationally limit the problem by introducing the Heaviside step function $\mathcal{H}$ [63, 64], for which $\mathcal{H}_{\psi}(x_i) = 1$ if $x_i \in \psi_j$ and $\mathcal{H}_{\psi}(x_i) = 0$ if $x_i \notin \psi_j$. This gives us the composite log-likelihood function resulting in a likelihood matrix, which we can iteratively solve using EM [65] by randomly initializing π_i and then instead estimating whether $x_i \in \phi_j$ using its expectation under the assumption of some random weight π_j given to each partial.

$$\mathbb{L}_{\Theta,\mathcal{H}} \begin{cases} \sum_{i=1}^{n} \log\left[\pi_1 \psi_1(x_i \mid \Theta_1)\right], & \forall j \neq 0 \; \mathcal{H}_1(x_i) = 1 \wedge \mathcal{H}_j(x_i) = 0 \\ \vdots \quad , & \vdots \\ \sum_{i=1}^{n} \log\left[\pi_k \psi_k(x_i \mid \Theta_k)\right], & \forall j \neq k \; \mathcal{H}_k(x_i) = 1 \wedge \mathcal{H}_j(x_i) = 0 \end{cases} \tag{4.3}$$

Now we can again use the mixture composite distribution equivalence to perform a back substitution to get the sum of the logs of the mixture distribution with the step function method.

$$\mathbb{L}_{\Theta,\mathcal{H}} = \sum_{i=1}^{n}\sum_{j=1}^{k}\Big(\mathcal{H}_j(x_i)\log[\pi_j\psi_j(x_i \mid \Theta_j)]\Big) \tag{4.4}$$

Now all that remains is to define the E- and M-fuctions for our EM algorithm. First, for the E-step we look to maximize the expected value of our optimization value, in this case we want to maximize the expected value of $\mathcal{H}$. In a practical sense this means that we want to find the configuration in which points x_i have a high likelihood to belong in partial distributions ψ_1 through ψ_k. This means we define the $\mathcal{H}$ expectation in our equation as follows.

$$\mathbb{L}_{\Theta,\mathcal{H}} = \sum_{i=1}^{n}\sum_{j=1}^{k}\Big(\mathbb{E}[\mathcal{H}_j(x_i) \mid \Theta_1,...,\Theta_n]\log[\pi_j\psi_j(x_i \mid \Theta_j)]\Big) \tag{4.5}$$

Then the expectation to be maximized can be isolated using Bayes [62].

$$\mathbb{E}[\mathcal{H}_j(x_i) \mid \Theta_1,...,\Theta_n] = \frac{\pi_j\psi_j(x_i;\Theta_j)}{\sum_{l=1}^{k}\hat{\pi}_l\psi_l(x_i;\Theta_l)} \tag{4.6}$$

Given the expectation, we can maximize the expectation in the M-step using Equation 4.6.

$$\hat{\pi}_j, \hat{\Theta}_j \leftarrow \arg\max_{\pi,\Theta}\mathbb{E}[\mathcal{H}_j(x_i) \mid \Theta_1,...,\Theta_n] \tag{4.7}$$

Now we can use maximum likelihood estimation to iteratively optimize the fit of some other candidate k-composite distribution to the observed data $x_1...x_n$ by finding the weights that maximize the expectation and then using the new weights and parameters to advance to the next iteration [66]. The optimization is constrained to the sum of weights equating to 1. In Figure 4.1 we show an example of two fits, one using a single Gaussian to the left and one showing a better fit with two Gaussians.

This also reveals a limitation of this method, as the amount of partial distributions k has to be defined at the start of the optimisation phase. We could always find a very close fit if we use enough distributions as the fourier transform of a Gaussian is a Gaussian in itself [67, 68]. However, doing

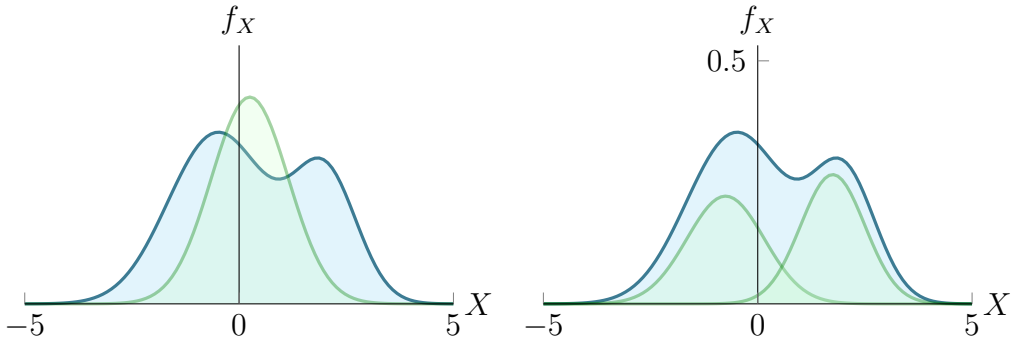


Fig. 4.1.: Two possible fit configurations to model partial distributions of X_C using $k = 1$ and $k = 2$ respectively

so means significantly overfitting on the sample data and will not produce a generalizable model. Instead, k can be selected by setting some fitting boundary, i.e. a $\alpha = 0.05$ coverage over the composite, optimizing models using $k = 1 \ldots k = k_{max}$ and then selecting the most parsimonious model that meets the coverage requirement.

Dirac δ-Composition

5

A more generalized way to treat composite distributions is to view it as a generalized transformation function over the number set $\mathbb{R}$. This can be achieved by using the Dirac δ-function to build δ-Composite Distributions. The Dirac δ-function is special in that it performs a linear mapping of any continuous function over a vector space, such as a distribution over a sample space, to an associated field of values at the zero point of the function in $\mathbb{R}$ [69].

This technique is also known as the normalisation of given state vectors, which may simplify the evaluation of generalized distribution functions in a composite. First we will evaluate the Diract δ-function, then we will show how this can be utilized to formulate any ψ to generalize the composition of distributions from partials [70].

5.1 The Dirac δ-Function

Remember that we defined the Heaviside step function $\mathcal{H}$ as the composite function that maps a real valued domain to the discrete set $\{0, 1\}$.

$$\mathcal{H}(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases} \tag{5.1}$$

We can then expand this function for any pivot point a with a transpose.

$$\mathcal{H}(x-a) = \begin{cases} 0, & (x-a) \leq 0 \\ 1, & (x-a) > 0 \end{cases} = \begin{cases} 0, & x \leq a \\ 1, & x > a \end{cases} \tag{5.2}$$

This forms a piece-wise function that can be formalized as the mapping function $\mathcal{H} : \mathbb{R} \to \{0, 1\}$ for any value of a. In Figure 5.1 we show the

behaviour of the step function at $a = 0$, any change to a transposes the stepping point on the x-axis.

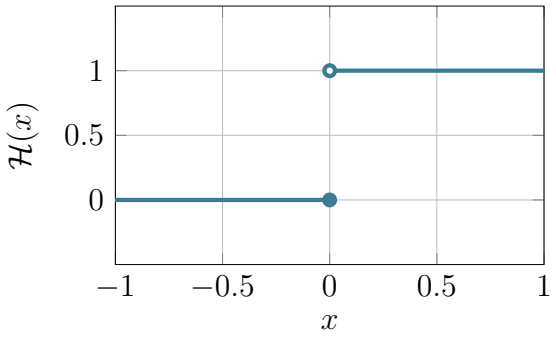


Fig. 5.1.: Heaviside Step Function

However, as this function is not symmetric, a common convention is used called the half-maximum convention. Following this convention, any stepping function's value on the break-point is the average between the upper bound and the lower bound. We show this in Figure 5.2.

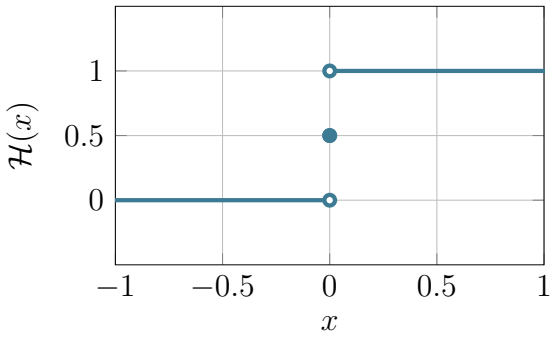


Fig. 5.2.: Heaviside Step Function using the Half-Maximum Convention

This piece-wise function and its a-expansion, where $x = a$ gives the mean of $\mathcal{H}$ can be defined as follows.

$$\mathcal{H}(x-a) = \begin{cases} 0, & x < a \\ 0.5, & x = a \\ 1, & x > a \end{cases} \tag{5.3}$$

From this function we can also derive a continuous step function $\mathcal{H}_a$, which instead of using the mean, connects the lower and upper bound together through a linear function which translates both ends $\frac{1}{2}a$ from the centre.

$$\mathcal{H}_a(x) = \begin{cases} 0, & x < -\frac{1}{2}a \\ \frac{1}{a}(x + \frac{1}{2}a), & x = -\frac{1}{2}a \leq x \leq \frac{1}{2}a \\ 1, & x > \frac{1}{2}a \end{cases} \tag{5.4}$$

In Figure 5.3 we demonstrate this using $a = 1$, which results in a linear step section between $x = -\frac{1}{2}$ and $x = \frac{1}{2}$.

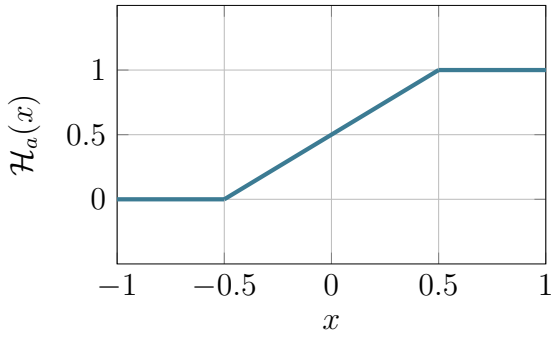


Fig. 5.3.: Continuous Step Function using a Linear Interpolation

However, while this new step function is continuous, if we look at the derivative we now find discontinuous points at $-\frac{1}{2}a$ and $\frac{1}{2}a$ respectively. We take the derivative of $\mathcal{H}_a$ and show.

$$\frac{\delta}{\delta x}\mathcal{H}_a(x) = \begin{cases} 0, & |x| < \frac{1}{2}a \\ \frac{1}{a}, & |x| \geq \frac{1}{2}a \end{cases} \tag{5.5}$$

And this gives us the the illustration in Figure 5.4, we can clearly see here that there is a discontinuity for which the derivative is not a continuous function.

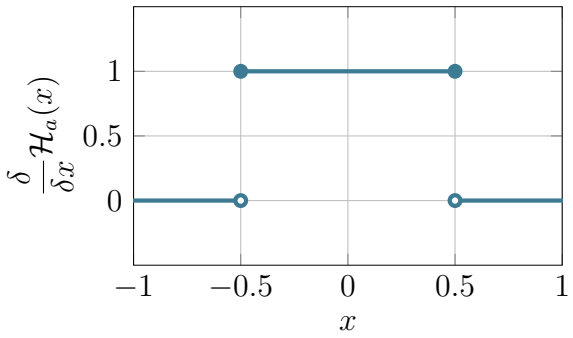


Fig. 5.4.: Derivative of the Continuous Step Function using a Linear Interpolation

We observe here that the size of the interval in which $\frac{\delta}{\delta x}\mathcal{H}_a(x) = 1$ depends on the magnitude of a. We also further note here that when $a = 0$, we get the original Heaviside step function $\mathcal{H}(x)$ as it follows $\mathcal{H}(x) = \lim_{a\to 0} \frac{\delta}{\delta x}\mathcal{H}_a(x)$. Thus if we take the limit of a to zero on the derivative of $\mathcal{H}_a(x)$, we get the derivative of $\mathcal{H}(x)$. We can now define the Dirac-δ function as the derivative of the Heaviside step function.

$$\delta(x) = \frac{\delta}{\delta x}\mathcal{H}(x) \tag{5.6}$$

The magnitude of $\delta(x)$ is inversely proportional to the size of a, thus when $a \to 0$, $\delta(x) \to \infty$, while it is 0 at all other x. This is a unique property as this allows $\delta(x)$ to map to any pivot x_0 of any continuous function, hence $\delta(x)$ allowing to form generalizable functions. For any possible mapping function $f : \mathbb{R} \to \mathbb{R}$ we can integrate $f(x)$ with respect to pivot x_0 to get the magnitude of $f(x_0)$.

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_0)dx = f(x_0) \tag{5.7}$$

For probabilities this has a very interesting property. Since the integral over any valid probability density function must be 1 according to the law of total probability, we can estimate any point probability over f using the Dirac-δ function. The standard Dirac-δ is visualized in Figure 5.5.

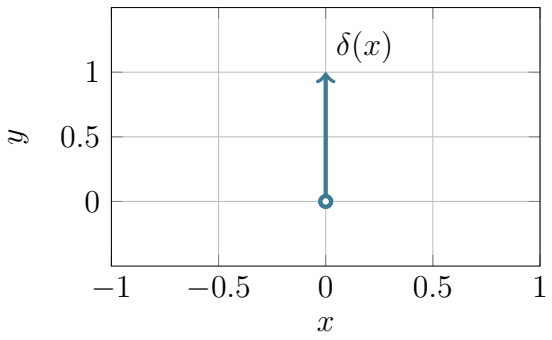


Fig. 5.5.: Visualisation of the Dirac-δ function

5.2 Composing Generalized ψ-Functions

With the definition of the Dirac-δ function we can define generalized functions that are a mixture of discrete and continous variables [70]. Since probability density functions are $\mathbb{R} \to \mathbb{R}$ mappings, we can thus use the Dirac-δ function to generate generalized ψ-functions. First let us define the expected value μ_ψ of any continuous distribution ψ as the integral of each x times the probability p at x.

$$\begin{aligned} X &\sim \psi(\Theta) \\ \mu_X = \mathbb{E}[X] &= \int_{-\infty}^{\infty} x\psi(x;\Theta_\psi)\delta x \end{aligned} \tag{5.8}$$

Now remember from Equation 5.7 that for any pivot x_0 we can map to $f(x_0)$ by integrating with the Dirac-δ function. For any continuous function, $x - x_0 \to 0$, such that $\delta(x - x_0) \to 1$. Thus for continuous functions the $\delta(x)$ term disappears from the integral. However, for discrete random variables this is not the case.

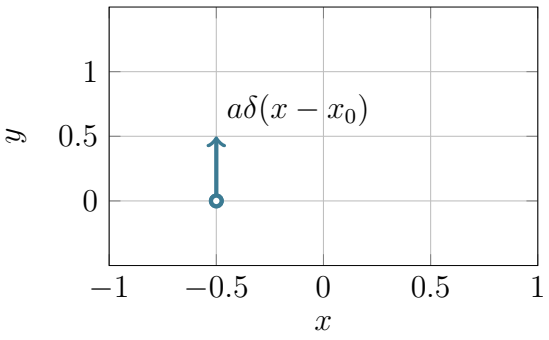


Fig. 5.6.: Dirac-δ function with pivot $x_0 = -0.5$ and magnitude $a = 0.5$

Derivation

To show this let us first step back to using the Heaviside function from Equation 5.1. Say now we have the delta $x - x_0$, then if $x > x_0$, then $\mathcal{H}(x - x_0) = 1$. For any discrete distribution we can define the probability density function as the derivative of the cumulative density function Ψ as $\frac{\delta\Psi(x;\Theta_\psi)}{\delta x}$. Since for the cumulative density function, we want every x_i be the sum of all $x \leq x_i$, we can sum up all the probabilities by defining for any

$x \leq x_i$ that $\mathcal{H}(x \leq x_i) = 1$ by using the delta trick on the probability mass function p for any $x \in X$.

$$\begin{aligned} X &\sim \Psi(\Theta_\psi) \\ \Psi(x; \Theta_\psi) &= \sum_{x_i \in X} p(x_i \mid \Theta_\psi) \mathcal{H}(x - x_i) \end{aligned} \tag{5.9}$$

Here we note that $\frac{\delta}{\delta x}\mathcal{H}(x) = \delta(x)$. Thus we then take the derivative of the cumulative distribution function as described in Equation 5.9.

$$\begin{aligned} \psi(x; \Theta_\psi) = \frac{\delta \Psi(x; \Theta_\psi)}{\delta x} &= \sum_{x_i \in X} p(x_i \mid \Theta_\psi) \frac{\delta}{\delta x} \mathcal{H}(x - x_i) \\ \psi(x; \Theta_\psi) &= \sum_{x_i \in X} p(x_i \mid \Theta_\psi) \delta(x - x_i) \end{aligned} \tag{5.10}$$

Evaluating ψ-Composites

6

Previously we have addressed sampling data as evidence for testing hypothesis. In this chapter we will further address this topic with techniques for fitted ψ-distributions. A typical case of hypothesis testing is to test whether a sample of limited data points conforms to some metric, instead we propose a technique where we computationally fit aggregated composite models and then computationally integrate against the ψ-function. In addition, we will cover how to perform composite hypothesis tests against multiple distributions and tests against composite distributions.

6.1 Numeric Overlap Method

There are various ways to perform hypothesis tests through inferential statistics, such as the student's t-test, paired t-test, f-test, chi-square test, the Wilcoxon test and ANOVA to name some of the most commonly used tests [71]. However, what all these tests have in common is that they either assume that the data follows some normalized distribution with set parameters or they may require direct access to the data to perform the test in the case of non-parametric testing.

An alternative is to use a model-agnostic hypothesis test, for which we propose to use the integral overlap $\mathcal{O}$ [72] as a computational method to test for model equivalence and monitor composite model inter-rater agreement on a specified variable.

To test hypotheses $H \in \mathcal{H}$ using the $\mathcal{O}$-method we use the process of computational inference. To do so we can select any k distributions to compare, which may also include the uniform distribution $\mathcal{U}$, which allows for a generalization of tests for μ. This test statistic is in essence a derivative descriptor over a set of evidence that the hypothesis is to be tested on. For instance, we can use $\mathcal{U}$ to estimate the critical boundary for any significance level α.

For instance take a standard normal as shown in Figure 6.1, giving us a random variable $X \sim \mathcal{N}(0, 1)$. The two-tailed critical region as shown can be

derived by taking the union between the distribution of X and forming the union over $\mathcal{U}(a, b)$, where a and b are the upper and lower boundary that form an area of $1 - \alpha$. In this case we are evaluating a two-tail measure, so with the law of total probability we can take the total probability 1 and subtract the cumulative probability distribution up until $x = 1$ to get the distribution of the possible boundary values where the surface area equals to $1 - \alpha$. Since there are many possible solutions for this, the solutions themselves occupy a distribution derivative of ψ.

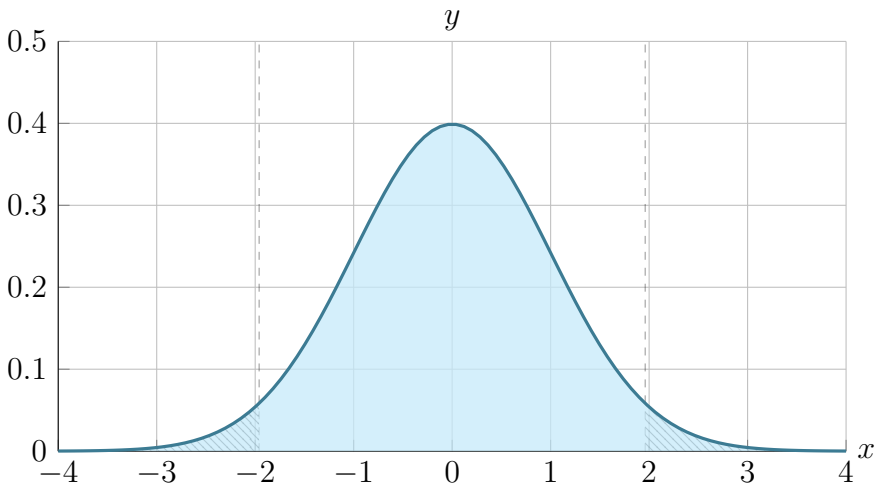


Fig. 6.1.: $\mathcal{N}(0, 1)$ using $\mathcal{U}(a, b)$ overlap with critical values for significance level $\alpha = 0.05$ shown in the shaded area in intervals $[-\infty, a]$ and $[b, \infty]$

Since this example is symmetrical, both sides occupy exactly 0.025 of the area under the curve. To evaluate this simple example computationally, we first define the probability density function f for this curve for the random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ as $f : X \to \mathbb{R}$, see Equation 6.1.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{6.1}$$

Now we consider the cumulative distribution function F_X, which is the integral over the probability distribution function f_X defined as $F_X(x) = \int_{-\infty}^{x} f_X(x)\delta x$. For the Gaussian case this can be evaluated as the following equation.

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2\right] \delta z \tag{6.2}$$

We can solve Equation 6.2 computationally with methods such as Newton-Cotes [73] or Simpson's rule [74] by approximating from the root $x_0 = \alpha$ on the right hand side of F_X. Since this example is symmetric, we can trivially find the solution $F_X(\frac{\alpha}{2}) \approx 1.96$. For distributions that are not symmetrical around μ, we can only use this method to find a right- or left-tail solution. A two-tail solution requires evaluating a sample of all possible combinations $(x, \alpha - x)$ where $\forall x, x \in [0, \alpha]$, which gives us a distribution over all possible solutions.

The primary type of tests we are interested in is tests for the overlap coefficient $\mathcal{O}$, which can determine the agreement or similarity between ψ-partials or Ψ composite distributions. We show an example in Figure 6.2, where we have marked the intersection area between the two distributions. The more both distributions overlap, the more similar the distributions are.

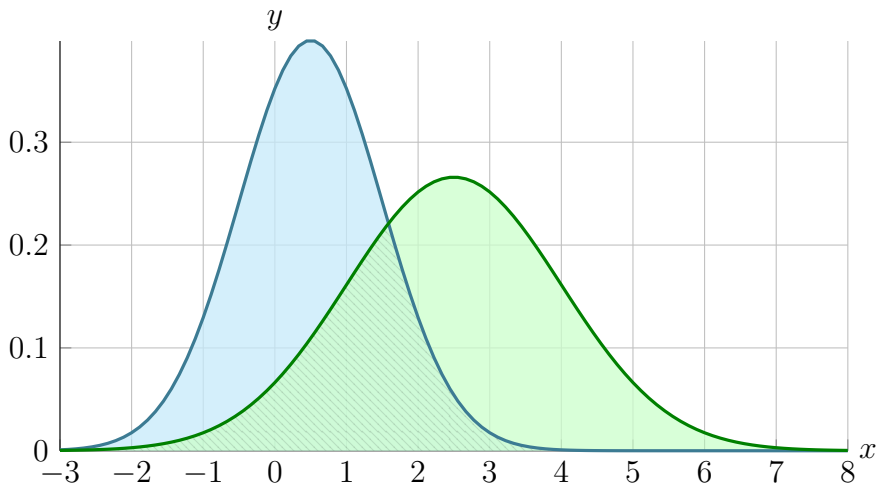


Fig. 6.2.: $\mathcal{N}(0.5, 1^2)$ in blue and $\mathcal{N}(2.5, 1.5^2)$ in green with shaded overlap

We define $\mathcal{O}$ as the ratio between the surface area of the intersection and the surface area of the union. This is an expansion of the Szymkiewicz-Simpson measure [75] to the continuous case in $\mathbb{R}$. Since it holds that for any distribution $\lim_{x\to\infty} F(x) = \int_{\infty}^{\infty}(x)dx = 1$, we can simplify this expression to $\mathcal{O}(\psi_1, \psi_2) = \frac{G(x)}{2-G(x)}$ where $g(x)$ defines the overlap curve between ψ_1 and ψ_2 and $G(x) = \int_{-\infty}^{\infty} g(x)\delta x$.

To evaluate the overlap, we may use the Newton-Raphson method [76] to find all intersections between ψ_1 and ψ_2. Given the intersections, we can construct a composite distribution function with $k + 1$ piecewise components where k is the number of intersecting points. In Figure 6.2 we have one

intersecting point computationally approximated at $x_0 \approx 1.587$. This gives us a new composite probability density function as follows.

$$g(x) = \begin{cases} \mathcal{N}(0.5, 1^2), & x \leq 1.587 \\ \mathcal{N}(2.5, 1.5^2), & x > 1.587 \end{cases} \tag{6.3}$$

We can then define the surface area function as a sum of integrals defined in Equation 6.4. Note that for $G(x) = \int_{-\infty}^{\infty} g(x)\delta x \leq 1$, as $g(x)$ maximizes under $\psi_1 = \psi_2$ to total probability. For disjoint distributions we find $G(x) \to 0$, which are maximally dissimilar distributions.

$$G(x) = \int_{-\infty}^{\infty} g(x)\delta x = \int_{-\infty}^{1.587} \mathcal{N}(0.5, 1^2)\delta x + \int_{1.587}^{\infty} \mathcal{N}(2.5, 1.5^2)\delta x \tag{6.4}$$

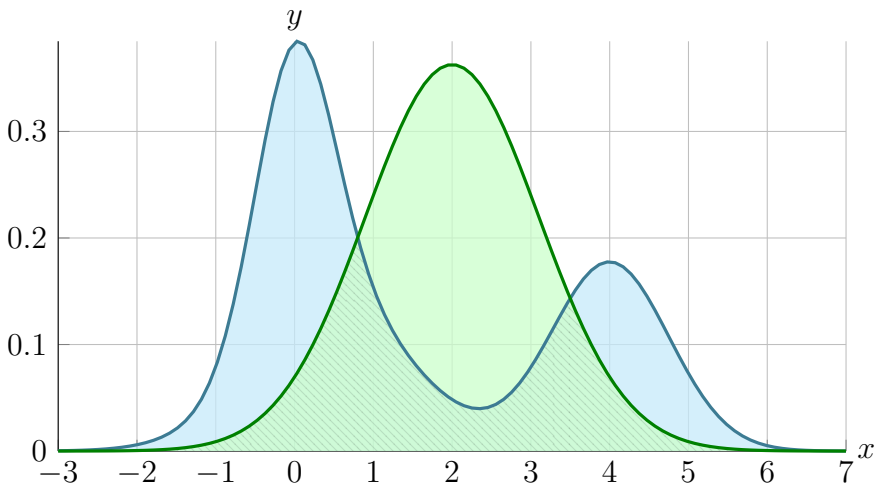


Fig. 6.3.: Composite distribution Ψ in blue and Gaussian distribution $\mathcal{N}(2, 1.1^2)$ in green with marked overlap

We can utilize this method measure $\mathcal{O}$ between two distributions. In Figure 6.3 we show for instance a composite that evaluates to $G(x) = \int_{-\infty}^{x_0} \mathcal{N}(2, 1.1^2)\delta x + \int_{x_0}^{x_1} \Psi(x, \Theta_\Psi)\delta x + \int_{x_1}^{\infty} \mathcal{N}(2, 1.1^2)\delta x$. In essence, we can generalize this to the set summation in Equation 6.5.

$$G(x) = \sum_{x_a, x_b \in X_{\text{int}}} \int_{x_a}^{x_b} \psi_i(x; \Theta_\psi)\delta x \tag{6.5}$$

Where X_{int} is the set that contains the lower and upper limits and all the intersection points as $X_{\text{int}} = \{-\infty, x_0, ...x_k, \infty\}$. The ψ_i that is used for each integral is based on the slope, for any interval $[a, b]$ if $\frac{\delta}{\delta x}\psi_i(x; \Theta_\psi) < 0$, then ψ_i is used to evaluate that interval. Naturally, for any intersection where $\frac{\delta}{\delta x}\psi_i(x; \Theta_\psi) > 0$, then the ψ that is not ψ_i is used as the surface for the previous interval. By summing each of these intervals together, the overlapping area between complex curves can be efficiently estimated.

Now for any given hypothesis we can evaluate $1 - \mathcal{O}(\psi_1, \psi_2) < \alpha$ as a method to test overlap hypotheses for generalized ψ-distributions. Note that overlap is a much more general estimation of similarity than t-tests, and also provides a very powerful testing criteria for time series or geographic data by comparing the evaluation of the ψ-distribution at different levels for t.

Part II

Computational Methods

> *Language is a process of free creation; its laws and principles are fixed, but the manner in which the principles of generation are used is free and infinitely varied.*
>
> **— Noam Chomsky**
> *Language and Freedom*

One of the most critical issues in any complex organisation or research programme is managing research data. Even the most advanced analytical techniques ultimately rely on the availability and quality of data. While there are many ways to deal with this issue, such as developing data pipelines and using linked object stores, the data themselves are rarely interoperable with each other. This results in data rarely being re-used, especially across different organisations or beyond the borders of nations.

The first step in tackling this issue is to formalize the method for data re-use, namely by embedding semantics as metadata within the data generating process using standardized ontologies. With this method, we can demonstrate that using graph algorithms we can make composite data, such as federated data, from multiple distributed data sources. In addition, we will conceptualise and showcase a generalized federated data architecture based on FAIR which has formed the basis of the computational solution in place at over a dozen physical sites.

7 Knowledge Graphs

Data is at the core of empirical research, so it is of utmost important that the generation, storage and handling of data are met with high standards in terms of reliability and accuracy. This includes being able to trace the provenance on how the data was generated, storing data in such a way that it is universally accessible under pre-defined conditions and that data is handled in a reliable and secure manner.

The typical research data that you would come across are in the form of tabular data or object files [77]. Generally, these are relatively compact and simple to manage within the scope of a single research project. More complex research programmes might use domain-specific storage solutions or databases to store scientific data. However, no matter the technology used, if data cannot be unambiguously understood and processed, then it is of no use outside of the direct sphere of the research.

As we have discussed in the introduction, this is a significant issue both within any large data ecosystem. Enormous amounts of data are essentially inaccessible, either because they are not properly indexed or because there is no sensible way to process or use the data. For applications in federated data, where we only have access to metadata and aggregated models or metrics [7], it is essential that the meaning, described through formalized semantics, is embedded within data for unambiguous interpretation and parsing. Without a formally described way on how to interpret and place data within context, there isn't any way to transform these data into information or knowledge that can provide value to operations or research.

7.1 Graph Representations

At the basis of formalizing interpretability in data is a fundamentally different data format, graph data. Graph data differs from tabular data in that graphs have a structure and ordering, with linkages between nodes that represent data classes and data points. This allows graph data to store knowledge: the combination of factual representations and the interactions between different

facts [78]. Tabular data is typically described through columnar metadata, which includes column names, measures and different properties which is stored within a tabular data store or file as data themselves. This concept is considered data as metadata.

Metadata within graph data takes on a unique role, as the metadata of a graph are classes that uniquely describe properties of the data. These properties are associated with nodes and linkages, which describe either the class of a node or a relation between two classes within a graph. Linkages can be unidirectional or bidirectional, which also means that in order to evaluate graph data, they need to be traversed in order to find specific combinations of results [20].

A regular directed graph is defined as a pair of sets $\mathcal{G}_2 = (N, L)$, where N are the nodes in the graph and L are the links connecting different nodes together represented as pairs of nodes in $L \subseteq N \times N$. For example if $N = \{x_1, x_2\}$ then we can represent the directed link $x_2 \rightarrow x_1$ as the 2-tuple in the link set $\mathcal{L} = \{(x_2, x_1)\}$. In Figure 7.1 we showcase an example graph. However, if we want to embed different types of relationships between nodes using the 2-tuple definition of a graph is insufficient.

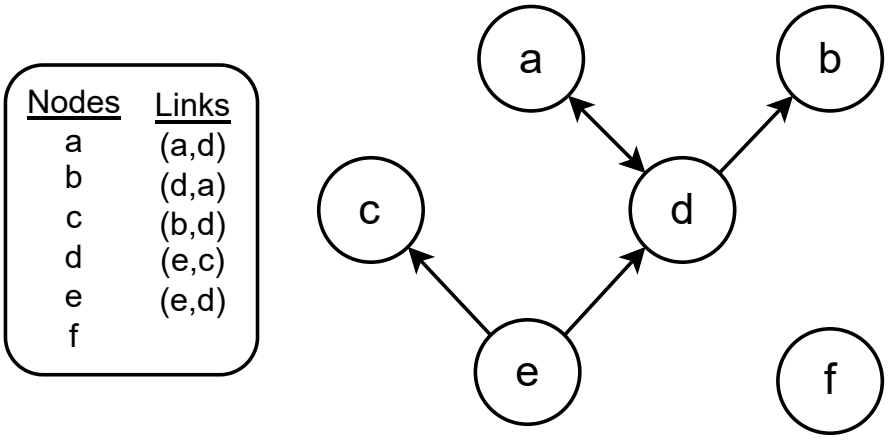


Fig. 7.1.: An example of a directed 2-tuple graph

For a knowledge graph we consider the 4-tuple definition, which is specified as $\mathcal{G}_4 = (N, L, R, f)$. In addition to the nodes and linkages, we add a set of relationships classes R. While nodes classes are unique, a relationship class may be repeated within the same graph. The mapping function f in this case generates the set $\forall l \in L, \exists r \in R, f : r \rightarrow l$. If we apply this to our previous example, we can for instance say $(x_2, x_1) \rightarrow$ "constraints" if we want to indicate that x_2 is constraining variable x_1.

The formalisation of classes, attributes, logic, axioms and the relationships between classes are considered ontologies, and as such you could argue that $\mathcal{G}_4$ is also an ontology. In essence N and R are controlled vocabularies, while L are relationships and f is a logical mapping function. Using an ontology, we can create instances of the ontology or perform ontology matching, the latter which we will cover in the next chapter.

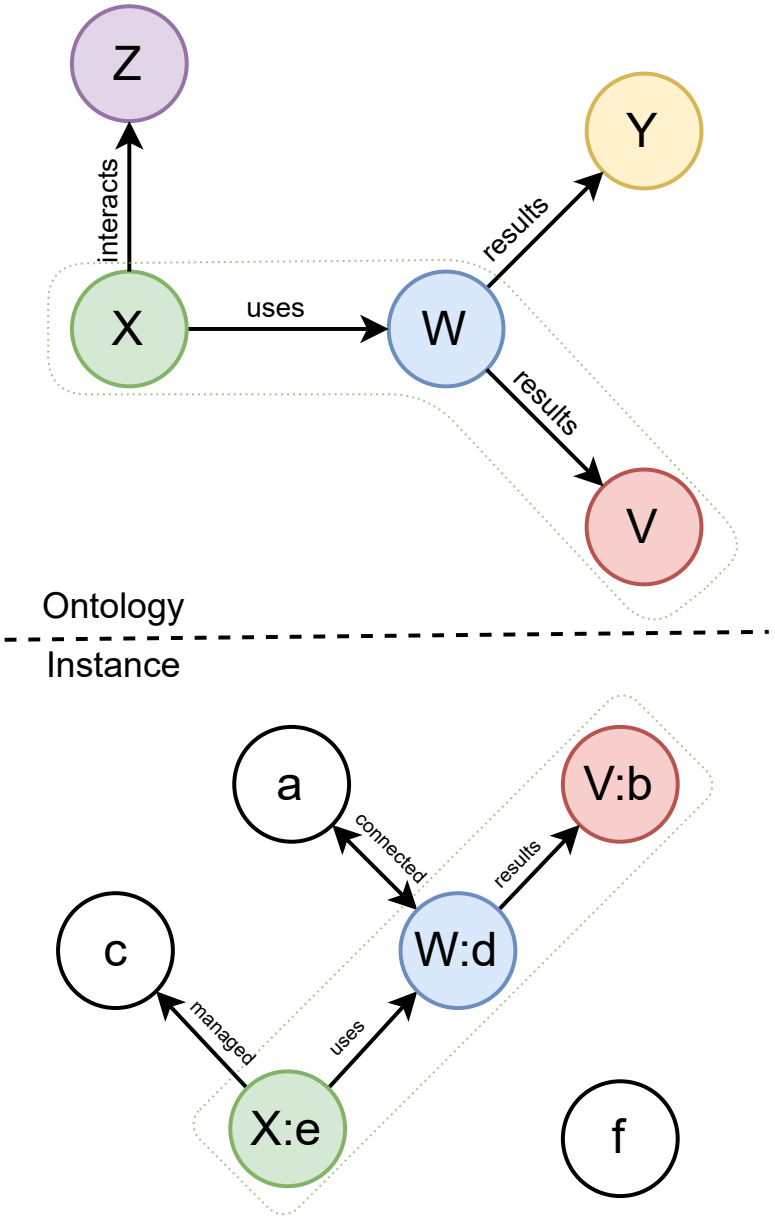


Fig. 7.2.: An ontology represented as a graph and an instance of the ontology

In Figure 7.2 we show a possible instance created from an ontology. Here edges b, d and e are instances of the ontology vocabulary classes X, W, V. Using ontologies to produce knowledge graphs for specific use cases, i.e. by

associating subclasses with data points or additional attributes, provides an essential property for interoperability of data. If we consider federated data sources, we can only traverse metadata. Thus if we can devise algorithms that operate over an ontology, agnostic of any underlying data points or instancing, we can be assured that this algorithm will run on any instance of the ontology [79].

This property of knowledge graphs can be compared to the rows and columns of a table. The node classes in an ontology supporting a knowledge graph are similar to the attributes in a table, they describe all the data instances in a column. In the context of knowledge graphs provide the meaning an explain ability that allows us to potentially combine, compare or interact with different data sources that are similar in meaning. The knowledge graph instance forms the equivalent of the row in a table, as it stores data points that are associated with the unique node classes.

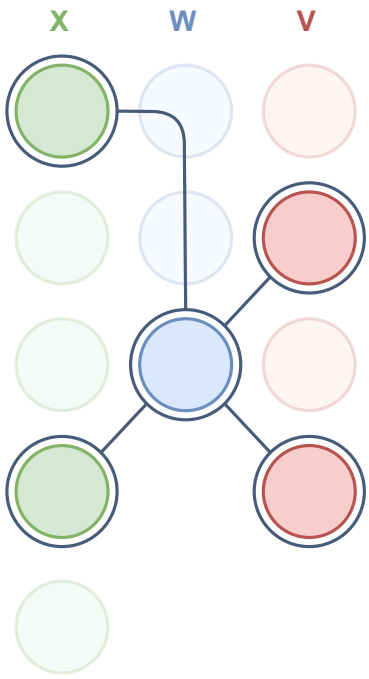


Fig. 7.3.: Representation of knowledge discovery across instances of an ontology

One of the powerful properties of knowledge graphs is the ability to perform knowledge discovery [80]. With this technique we can apply graph traversal to find meta-relationships between different instances of the same ontology. Expert knowledge from the ontology can be embedded in operational models, because we can learn from linked data and can can also build models that

utilize these causal relationships between nodes [81]. In addition, we can even express the type of the causality through the classification of the link.

In Figure 7.3 we provide a simple representation of knowledge discovery. This can even be applied to incomplete subsets of the ontology, for example we can discover a meta-relationship of multiple instances of a class resolving to similar instance values of another class [21]. In the case of federated data, we will not be able to see these individual instances of data, so instead we can make inferences over the resulting composite data. For example using the ψ-composite overlap technique we discussed in the previous part, we can discover the rate of agreement within a certain class and then analyse the rate of agreement with directly adjacent classes. This is an important property for the reliability of a model, as an unreliable super-class will propagate through its sub-classes.

7.2 Attribute Grammars

Graphs are abstract concepts that cannot be directly used in computational methods without a proper syntax. One of the ways to represent a graph is through a series of statements or rules, that represent a formalisation of the production function f we have introduced in the previous section. To be able to read and understand sequences of semantic embeddings, we do not only need understanding of the individual components, but also how those components may be formed and the rules governing combinations of components [82]. These rules are expressed as a formal grammar, which provide the production rules for a language. The way graph syntax can be expressed is through a formal grammar.

A formal grammar G is much like a knowledge graph $\mathcal{G}$ expressed as a 4-tuple. We define the formal grammar as $G_4 = (N, \Sigma, P, S)$, where N are non-terminal terms, Σ are terminal terms and $N \cap \Sigma = \varnothing$. P represents the production rules, which can be in the shape $\alpha \rightarrow \beta$ and define the different variations of possible syntax that can be generated over the total set of symbols $N \cup \Sigma$. $S \in N$ is the start symbol, which together with rules around S in P form the entry point for generating a graph syntax. Take for instance a knowledge graph, if we want to produce a sensible graph we always start with a node belonging to some class. Then we can produce other nodes with potential edges between them. We cannot produce edges that are not

connected to anything. A formal grammar can express these conditions for structuring the syntax of a valid and machine readable graph.

A basic grammar G_4 that can be used to produce directed graphs over a set of classes $x_0, ..., x_1$ can be produced by formalizing the production rules.

$$\begin{aligned} N &= \{S, P, R, N, L\} \\ \Sigma &= \{\mathcal{X}, >, (,)\} \\ P &= \{S \rightarrow PS, S \rightarrow P, P \rightarrow (R), R \rightarrow N, \\ &\quad R \rightarrow N > N, N \rightarrow \mathcal{X}\} \\ S &= S \end{aligned} \tag{7.1}$$

Where we assume that x is the variable for the node symbols. This is what we consider a context-free grammar, as in the production rules none of the left-hand terms have more than one term, e.g. they do not consider their surrounding context. This will result in a series of p-separated production rules that formulate the structure of a graph. For example the graph in Figure 7.1 can be produced by the grammar in Equation 7.1 using $\mathcal{X} = \{a, b, c, d, e, f\}$ as:

$$\begin{aligned} &S \rightarrow PS \rightarrow (R)S \rightarrow (N > N)S \rightarrow \\ &(\mathcal{X} > \mathcal{X})S \rightarrow (a > d)S \rightarrow ... \rightarrow \\ &(a > d) \\ &(d > a) \\ &(b > d) \\ &(e > c) \\ &(e > d) \\ &(f) \end{aligned} \tag{7.2}$$

Which are the exact production rules required to reproduce the directed graph of Figure 7.1. However, as we have seen in this context we have to redefine $\mathcal{X}$ to contain the definite symbols a, b, c, d, e and f. In this case this set is our controlled vocabulary. In addition, we also cannot define classes

to linkages without running into the same issue of having to redefine our grammar for every graph. The method that may be used to deal with this issue are attribute grammars.

An attribute grammar is an extension to a formal G_4 grammar, which allows embedding of semantics attributes such as classes, templates, instances and typed values within a grammar without fundamentally changing the context-free grammar that is responsible for the base syntax [83]. This extension is given by allowing any term $x \in N \cup \Sigma$ to have attributes associated with them, denoted by using a dot-specification. For instance, we could now say that $\mathcal{X}$ has the attribute *instance* associated with it, then we can define in the grammar the semantic rule $P_{\mathcal{X}} : \mathcal{X}.\text{instance} := \text{“}a\text{”}$ if we want to assign the string a to the *instance* attribute of $\mathcal{X}$.

Naturally, we want to perform dynamic assignments. This can be done through semantic rules that are embedded within the production rules P, which provide a meaningful relationship to the syntax being produced. In addition, because of this feature, we can generalize the grammatical rules significantly by referring to the superclass and producing sub-classes through attribute assignment. The most basic example of this is an inheritance assignment rule, which can be formulated as:

$$\text{Superclass} \rightarrow \text{Subclass} \quad [\text{Superclass.type} = \text{Subclass.type}] \tag{7.3}$$

This rule for instance says that any expression containing Superclass can be subject to a production rule that turns it into a Subclass while retaining the same type attribute. In essence, this semantic rule within production is what describes the semantic classification of relationship linkages that we discussed in the previous section.

This can also be applied to perform logic or arithmetic, for instance we can embed formulas in the graph that describe knowledge that has been defined in the ontology. We can provide a simple example which uses the length, width and height attributes of a class to generate a new associated volumetric class.

$$\text{Object} \rightarrow \text{Object} > \text{Volume} \quad [\,\text{Volume.value} := \text{Object.length} * \text{Object.width} * \text{Object.height}\,] \tag{7.4}$$

If we now return to the set of rules we defined in Equation 7.1 we can redefine these rules with the attribute grammar addition to embed semantics. In this case our issue was that there was no way for us to define a finite set of terminals without changing the grammar itself. Now, we propose to use the attribute grammar to define the value of the terminal rather than embedding it as a term. In addition, our link production rules can now embed rules to set the type of relationship class between two nodes.

$$\begin{aligned} N &= \{S, R, N, L\} \\ \Sigma &= \{\mathcal{X}, >, (,)\} \\ P = \{S &\rightarrow (R)S \quad [R.\text{instances \&}= S.\text{instances}], \\ R &\rightarrow N_i \quad [N_i.\text{instance} := S.\text{pop}[.\text{instances}_i]], \\ R &\rightarrow N_i > N_j \quad [N_i \neq N_j; >.\text{relation} := R.\text{instance}_{i,j}], \\ N &\rightarrow \mathcal{X} \quad [\mathcal{X}.\text{instance} := N.\text{instance}]\} \\ S &= S \end{aligned} \tag{7.5}$$

Now if we instantiate S with our class instance list such that S.instances = $\{a, ..., f\}$ and S.relations = $\mathcal{R}$, we can build or validate a semantic graph according to the attribute grammar production rule set.

8 Graph Interoperability

One of the primary use cases of embedding semantics in graphs is that we can apply automated reasoning methods to these graphs, such as converging different data sources for interoperability and embedding semantic graphs in graph-based models. This has great significance in methods for federated learning and analysis, as we are often dealing with disparate data sources.

Ontologies form meaningful standards on which data sources can be based or mapped, which provide avenues to make parts, denoted as sub-graphs, of the complete data interoperable. Data at different federated instances do not necessarily need to entirely match, as we can perform analyses on these sub-graphs to find meaningful relationships or patterns that combine expert knowledge from an ontology.

In the previous section we have discussed that graph data can either be produced from an ontology, or an ontology can be mapped onto an existing data source. In this section we show a method that enables us to utilize these mappings to make heterogeneous data from different restricted sources interoperable.

8.1 Semantic Convergence

Data across repositories and studies originate from data generating processes. These processes can vary from empirical research data to qualitative surveys, which results in data that can present itself in a wide range of formats, according to different standards and may differ in meaning depending on the domain of application. This makes re-use of data especially challenging, which is an issue within the federated data methodology.

The very essence of federated data is to support continuous re-use of data without direct data access, which can only be supported if the meaning of data is known ahead of time. The proposed way to deal with this is by embedding semantics in the accessible metadata that describe the format of the actual data instances. If these standards exist over many different

repositories within a federated data cluster, then there is the opportunity to apply the ontology matching technique.

To perform this, we must define a graph algorithm that can perform ontology matching. There are various possible techniques that focus on different aspects of graphs, such as maximally-matchable edges [84], maximum-cardinality matching [85], bipartite matching [86] or using a heuristic like the Hosoya index [87]. However, as our techniques are aimed to maximize the generalisability of our techniques across repositories, we decide to go for a very generalistic approach with subgraph isomorphism matching.

Subgraph Isomorphism

The subgraph isomorphism detection algorithm works by comparing two $\mathcal{G}_4$ graphs, G_1 and G_2. It then finds whether G_1 contains any subgraph that is isomorphically congruent with any subgraph in G_2. In other words, G_1 is subgraph isomorphic with G_2 if there exists a bijective mapping between G_1 and G_2. For this to hold it must also hold that for the bijective mapping $f : N_{G_1} \rightarrow N_{G_2}$ such that for $\exists x \exists y : \text{adj}(x, y) \cap \text{adj}(f(x), f(y)) \neq \varnothing$. We describe this procedure in Algorithm 2 using a double adjacency queue.

Algorithm 1 Baseline Subgraph Isomorphism

```
1:  M := ∅
2:  for all n ∈ N_G1 do
3:    if n ∈ N_G2 then
4:      P := n
5:      Q1 ← N_G1[n]
6:      Q2 ← N_G2[n]
7:      for all q ∈ Q1 do
8:        if q ∈ Q2 then
9:          P ← P ∪ q
10:         Q1 ← Q1 ∪ N_G1[Q1.pop(q)]
11:         Q2 ← Q1 ∪ N_G2[Q2.pop(q)]
12:       end if
13:     end for
14:     if |P| > 1 then
15:       M ← M ∪ P
16:     else
17:       DISCARD P
18:     end if
19:   end if
20: end for
21: return M
```

While this algorithm will find all possible subgraph isomorphisms between G_1 and G_2, the algorithm itself scales very poorly when applied to larger graphs since it needs to perform breath-first search through both graphs with a double queue. . For a complete search this is unavoidable, as finding subgraph isomorphisms is a NP-complete problem [88]. However, we can modify the problem in such a way that we can find a more optimised solution using the properties of our semantic data.

We propose to leverage selected pattern queries in source ontologies to match towards any n graphs to match semantic feature $\mathcal{F}$. By matching a feature, either as a full or partial match described in a binary matrix, we can circumvent the dual traversal complexity. Instead, we use the source ontology as a feature mapping that can then be transferred to other graphs. Any graph that matches a feature in the source ontology, is (partially) interoperable with another graph that matches the same feature. In Figure 8.1 we show a feature from a source ontology being matched with two distinct graphs.

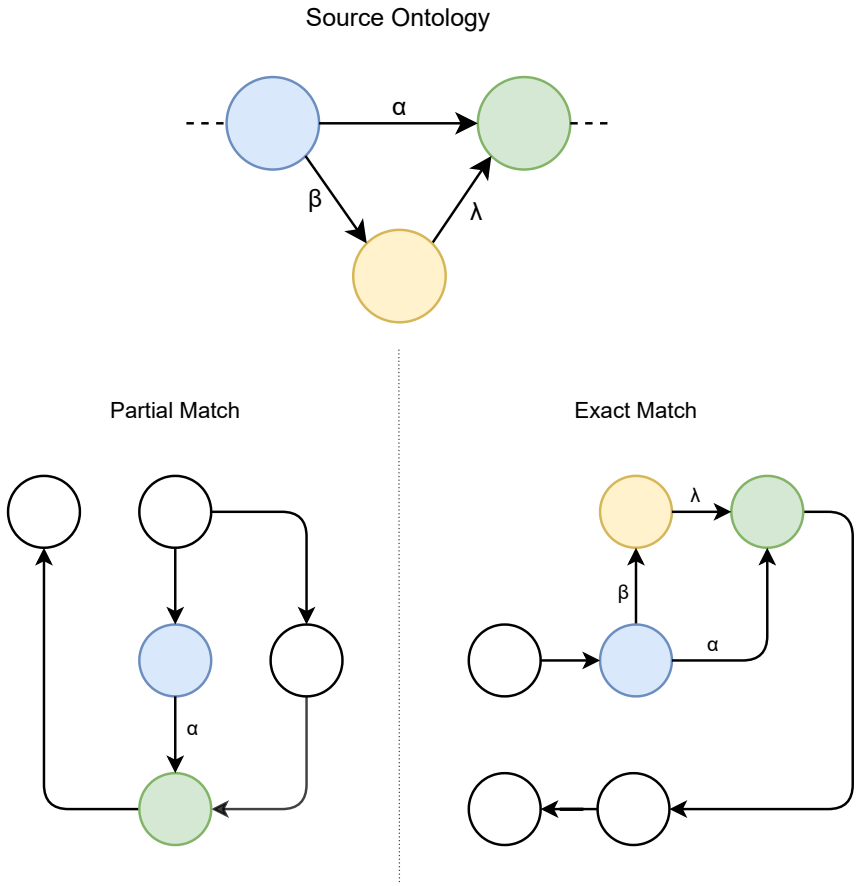


Fig. 8.1.: A schematic for partial and full ontology mapping from a source ontology towards heterogeneous target knowledge graphs

In essence, instead of searching for any possible isomorphism between two graphs, we search for specifically indexed features in individual graphs. The availability of such k-features can be recorded in a binary index matrix $\mathcal{I}$ of size $n \times k$, which can be used to find similar features in any n graphs.

Algorithm 2 Ontology Matching Subgraph Isomorphism

$M := \varnothing$
$T := \varnothing$
$\mathcal{F} := \mathcal{G}_{\mathcal{F}}$
for all $N_{\mathcal{F}} \in \mathcal{F}$ **do**
  if $n \in N \notin T$ **then**
    $T \leftarrow T \cup n$
  end if
end for
for all $n \in T$ **do**
  $z \leftarrow z \; ? \; z : n$
  if $n \in G$ **and** ($n \in N_G[z]$ **or** $|M| < 1$) **then**
    $z \leftarrow z \cup n$
    $M \leftarrow P \cup T.\text{pop}(n)$
  end if
end for
if $|M| > 1$ **then**
  return M
else
  return $\varnothing$
end if

In this instance we only need one graph traversal, as we only have to parse the source ontology and then check if our adjacency list contains any accessible matching elements. Upon finding a match, we update our ontology accordingly, otherwise we move on to the next element. Since we only have to consider adjacency for the source ontology feature, we do not have to continuously update our adjacency graph, but instead we iterate over all matches to see if they are adjacent. This allows us to split the algorithm in two individual loops, instead of using an inefficient nested loop. We can continue this process until the entire adjacency list is empty and return any matching subgraphs.

9 Federated Data Architecture

In order to bring the statistical and computational techniques we have discussed in the previous chapters in to action within a federated data framework, we need to look at the bigger picture. Data and their resulting models do not live in isolation, but in the sphere of data generating processes and their context, sensory equipment or data capturing technology, data pipelines and databases that store data. In addition, societal effects and legislature can also have a strong impact on data, as we have seen with data protection acts, biases in data, reliability issues or even the potential for fraud [25].

These are grand challenges when dealing with federated data, which essentially aims to build a single virtualised repository from multiple physical sites and storage solutions. In comparison to cloud-based techniques, various levels of access may be provided across sites and direct data access is restricted [6]. Instead, metrics and models are calculated on-site, parsed to the virtualised entity and then built into a composite model spanning all participating sites.

While this provides strong safeguards for privacy, security and data ownership, it also brings up questions in terms of reliability and quality of data [89, 90]. In the section covering statistical techniques we already covered how to work with composite distributions and how to measure large deviations from the norm, but we must also look at the process itself in order to ensure that analyses resulting from federated data are reliable. This means that within a federated data architecture, care should be taken that the complete data lineage and provenance is recorded and available for audit.

In order to address the concerns around data access, ownership, interoperability, reliability and usage we will use the framework of FAIR data and services. We will show that this framework is suitable for usage in a federated setting and we will illustrate the general structure of FAIR data points as federated data repositories to design a scalable federated research infrastructure.

9.1 FAIR Data

In 2016 Wilkinson et al. [16] developed a set of principles named the FAIR data principles, which aim to improve data re-use and support proper research data management and stewardship. One of the most fundamental issues in research is that data is not commonly re-used after a study has been concluded, while these data are not only costly to acquire, but may also in some cases not be acquired again. Furthermore, re-use is data is vital for the scientific process, repeat studies depend on accessibility of the original data sources and require provenance to replicate the experimental setup.

FAIR Principles

- **Findable**
  Data are properly indexed and have a universally unique, persistent identifier associated with them.
- **Accessible**
  Access and permission control surrounding the data are properly described, data are accessed through a universal, open communication protocol.
- **Interoperable**
  Data is properly described with semantic metadata and follows open standards for machine readable data representations.
- **Reusable**
  Data are reposited in a findable, accessible and interoperable format with full provenance, data lineage and licensing.

More than ever, the advancement of scientific research relies on empirical data. The FAIR principles guide the practice of stewardship and standards that aim to increase data reuse. The more data that is available, the more potential there is to derive value from data and the easier it is to verify scientific findings to a high degree of reliability.

At the forefront of this is the development of research infrastructure, such as the tooling and systems that enable us to produce, leverage and manage FAIR data. This poses significantly more challenging than regular data ecosystems, since some of the key issues that FAIR addresses were typically avoided for a

reason. Such developments are time consuming, often requiring significant overhead, and also require significant expertise to properly implement. For this reason data stewardship is of increasing importance, as it would be infeasible to expect researchers of every discipline to be aware of the exact conditions and procedures on which FAIR data rely.

The FAIR principles ultimately follow the fundamentals of the semantic web, the envisioned semantic layer on top of the regular internet protocol that drives the web. At the very essence, FAIR data from this perspective is an extension to semantic data, with specific additions to improve indexing and accessibility that is relevant when used across different research programmes.

9.2 Data Localization

When dealing with the process of localizing the creation and management of metadata we will need to consider not only the intricacies of the specific domain for which we engineer, but also the data governance framework of the locale in which we are operating. One of the central contexts within data governance is data ownership as noted by Janssen et al. [91], which may pertain to the legal aspects of possession of, responsibility over and rights to a specific element or set of data. In practice it is often challenging to determine exactly who the owner of a piece of data is, as also demonstrated by Al-Khouri [92]. The further data is removed from the origin, the more challenging the question of data ownership becomes.

A crucial element that is required in order to determine, and document, data ownership is data provenance provided by rich metadata. These are the metadata that provide specifications regarding to the lineage of data through the lifespan emerging from specification, to data generation and ultimately removal or archiving. How these metadata are structured within a specific domain is a subject of contemporary research [93]. However, there are aspects of metadata provenance, as common data elements, that are used across all domains, which are the provenance we are interested in when looking at data localization. Provenance in such regard is also cited as a key element to data reuse [94], providing both the legal and semantic framework from which these data originated through a localized context.

In order to reconcile the need for localized data ownership, especially when handling sensitive data across national borders, with the FAIR principles, the concept of the FAIR Data Point (FDP) arose [17]. By utilization of FAIR metadata engineering during the data creation process, and repositing these data in locally managed storage points, strong safeguards are provided both in terms of provenance and data ownership [95] that can be applied within federated data as federated data repositories.

Ensuring that data is FAIR at point of creation provides significant technical challenges, but also provides significant advantages over making data FAIR post hoc [96]. To address the technological challenges, we propose to use the CEDAR ecosystem [97] as a technology to perform ad-hoc generation of FAIR data. Using an ad-hoc process, which is based around the use of community-specified base ontologies, will provide significant advantages when tackling data interoperability and reusability challenges, and provides a strong baseline from which to further develop a FDP ecosystem.

9.3 Federated Data Repositories

The concept of the FAIR Data Point (FDP) as a federated data repository is centered around the practical implementation of the FAIR principles as first described by Wilkinson et al. [16], which describes a data management framework which enhances data interoperability and reuse. However, implementing these strategies in existing data, a process that is named FAIRification, is both time consuming and often not feasible due to lack of provenance [96] or due to ambiguity present within the data.

The manual implementation of FAIR principles and curation of metadata within a study may be an expensive, laborious and ambiguous process due to the lack of a specific technological implementation to support these methods [98]. Without direct benefits for the principal investigators of the study, there is little incentive to truly leverage the benefits of FAIR. Despite FAIR being a requirement for an increasing amount of scientific grants and funding opportunities, in order to improve reusability of scientific data [99], reusability remains limited if there is a lack of common technological or semantic standards in order to make FAIR metadata interoperable and machine actionable.

The FDP provides the much needed technological framework in order to support bringing the FAIR principles into practice. This also allows for the introduction of the concept of shared ontologies, such as the gene ontology by Ashburner et al. [100], into the framework of FAIR data and services. In order to ensure interoperability and reusability of data [101], it is not only needed that community-based metadata standards are developed as indicated by Wilkinson et al. [16], but these domain metadata standards as ontologies also need to be dynamic through the use of ontology services to support metadata specification and curation.

The importance of dynamic, shared ontologies also signifies for the importance of these ontologies to be FAIR [102]. As ontologies may change over time, it is important that provenance over such changes is retained to ensure machine interoperability of past data for reuse beyond the scope of ever changing scientific vocabularies. In Figure 9.1 we provide a diagram that illustrates the process in which FAIR data can be generated for data reuse through FAIR metadata templates, supported by dynamic ontology services. The environment in which both metadata templates are curated, and data is generated and stored, forms the baseline of the FDP.

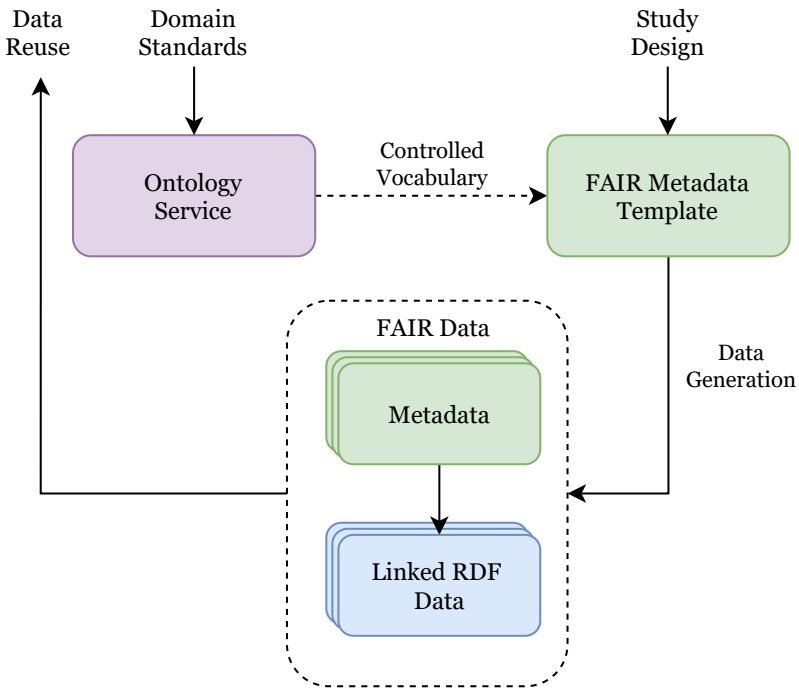


Fig. 9.1.: The process for data reuse through FDP based metadata templating.

Central to developing the FDP as a federated data repository is the semantic linkage. An ontology service provides the central provenance over all controlled vocabularies, terminologies and semantic linkages from which

localized FAIR metadata templates can be produced. Domain experts contribute to the ontology service by implementing domain standards, while those who want to perform a study can leverage this by building forms or data pipelines around these specifications.

Once data is being generated through a FAIR metadata template, it is considered intrinsically interoperable with other data that has been generated in accordance with the same template. In addition, since FAIR provides a specification for machine readability and accessability, data can easily be coalesced, levering the a priori knowledge from the semantic embeddings.

These type of data are typically formatted using the Resource Description Format (RDF), which is a machine readable syntax that can be used to formally describe graph data [31]. Just like directed labelled graphs, RDF is centered around the concept of triples. This consists of a source node, a semantic link descriptor and the target node.

As we have shown in our discussion of attribute grammars, production rules in this format can be used to describe any possible semantic graph. Attributes can be associated to nodes and links by pointing the descriptor to a uniform resource identifier (URI) referencing to a comprehensive attribute specification [50]. The graph can then be pared by utilizing a parameter specification for attribute values, or attribute values can even be encoded within the graph itself without making modifications to the base syntax.

Baseline queries are performed using either customized APIs or a RDF querying language such as SPARQL. Utilizing CEDAR, a standard REST API endpoint has been implemented which can query data in accordance to a semantic specification. In a federated framework, this querying is done by another internal microservice that then processes the data. External queries are translated to the format that an internal data repositing service can understand, and they are validated against an authentication service before being parsed. In **??** we show an abstract version of a querying routine.

Any incoming query according to the specification is accompanied with a required API key, indicating the level of access or granularity that can be returned. Data access policies do not just relate to credentials, but also to a legal rule-set that can differ between geographies. The backend services communicate directly with the data repository, query data internally, process

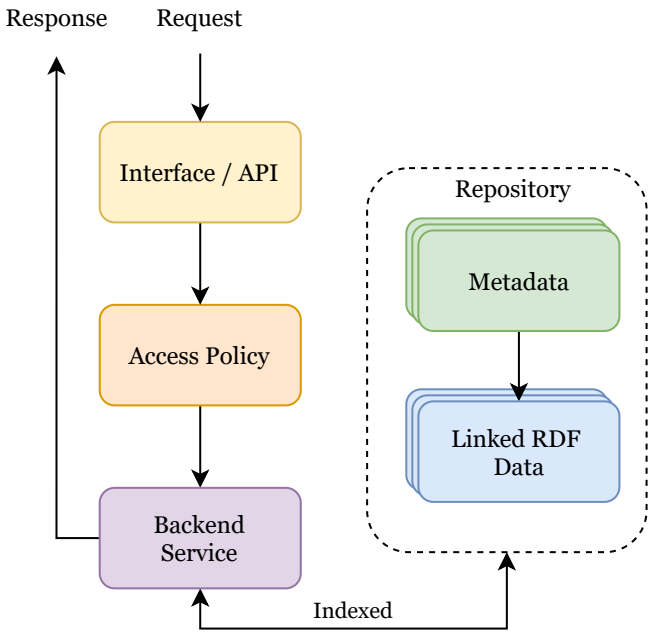


Fig. 9.2.: The basic framework of a FAIR Data Point.

it according to the specific needs and requirements, and then returns a response in the form of a metric, model or synthetic data set. Typically the range of internal queries is pre-specified, and custom query input is validated through an audit process.

In appendices A and B we present a complete diagram on the process level of repositing and sharing of federated data, furthermore in appendices C and D we provide a deeper level technical specification for each of these internal processes respectively. This implements a complete FDP as a federated data source, which can enable FAIR data production and secure federated data queries across FDP instances.

10 Discussion

> *The Web as I envisaged it, we have not seen it yet. The future is still so much bigger than the past.*
>
> **— Tim Berners-Lee**

Throughout this research we have discussed the challenges surrounding contemporary data use and proposed federated data as a potential solution for some of these problems. There we also noted that embedding semantics in data can provide some very powerful properties, which are just as viable outside use of the federated data methodology. To this extent we have covered two main branches of research, namely the analytical branch and the computational branch.

At the begin of this research we posed two primary research questions, which we have covered over the two main sections in this research. Below we will cover our conclusions in regard to these questions.

I. *What are the ways in which we can utilize existing statistical inference techniques in order to extend these analyses over heterogeneous federated data?*

In order to answer this question we will look at the analytical section covering statistical methods. Therein we developed the technique of composite statistics as a branch from generalized mixture models, δ-function and ψ-composition. We recognize that in most federated use case, we will come across piece-wise distributions that are challenging to work with using standard statistical analyses.

With composite statistical techniques, we can estimate generalized distribution functions for any set of partial distributions originating from a multitude of different sources, which can then be analysed and approximated with general statistical techniques. In particular, we have shown a technique called the overlap method which can be applied reliably to any possible distribution across a federated data cluster without prior assumptions.

II. *How can we utilize and enhance the graph properties of semantic data to enable interoperability over heterogeneous data sources?*

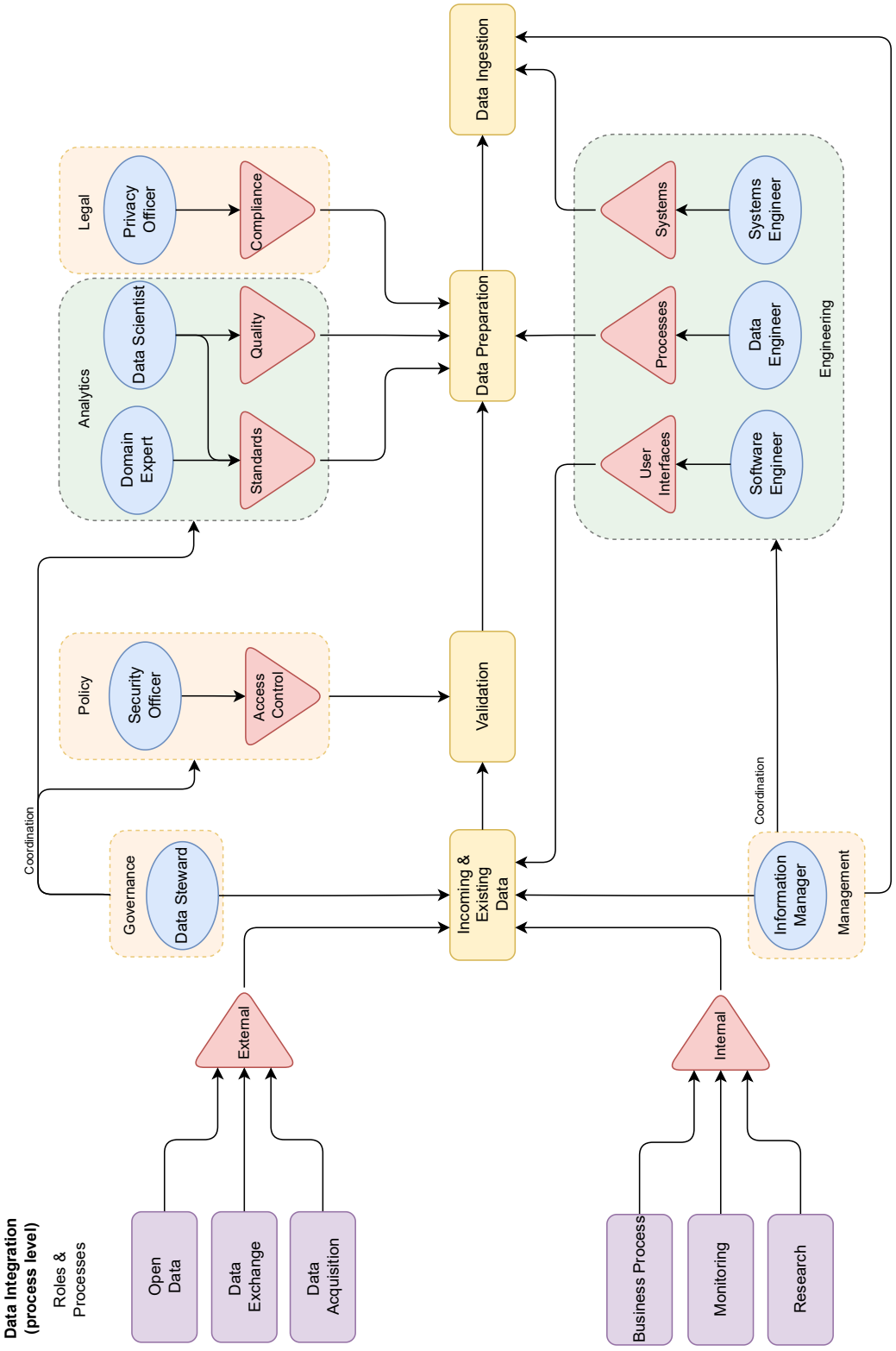
In the section on computational methods we covered methods that describe data formatted as knowledge graphs that can be syntactically generated through attribute grammars. These knowledge graphs can provide avenues for semantic convergence, the process of finding similar meaning in disparate data from federated sources, that is vastly different in content.

Here we proposed an ontology matching method for subgraph isomorphism, where we optimise knowledge discovery based on features $\mathcal{F}$ that are contained within a source ontology. Finally, we designed a federated data architecture based on the FAIR data principles and FAIR data points, which focusses on providing interoperability across a scalable architecture while ensuring data ownership through complete data localisation.

Bringing this all together we recognize the need for the development of techniques and methodologies that specifically support the federated data framework. By developing methods that embed semantics in data a priori as FAIR data, a lot of potential issues can be avoided while providing opportunities to leverage an increase in interoperability of data across sources. At the same time, existing concepts such as the semantic web closely match in principal application towards a uniform and scalable technological fabric and may provide a key ingredient in having a more widespread application of the federated data methodology.

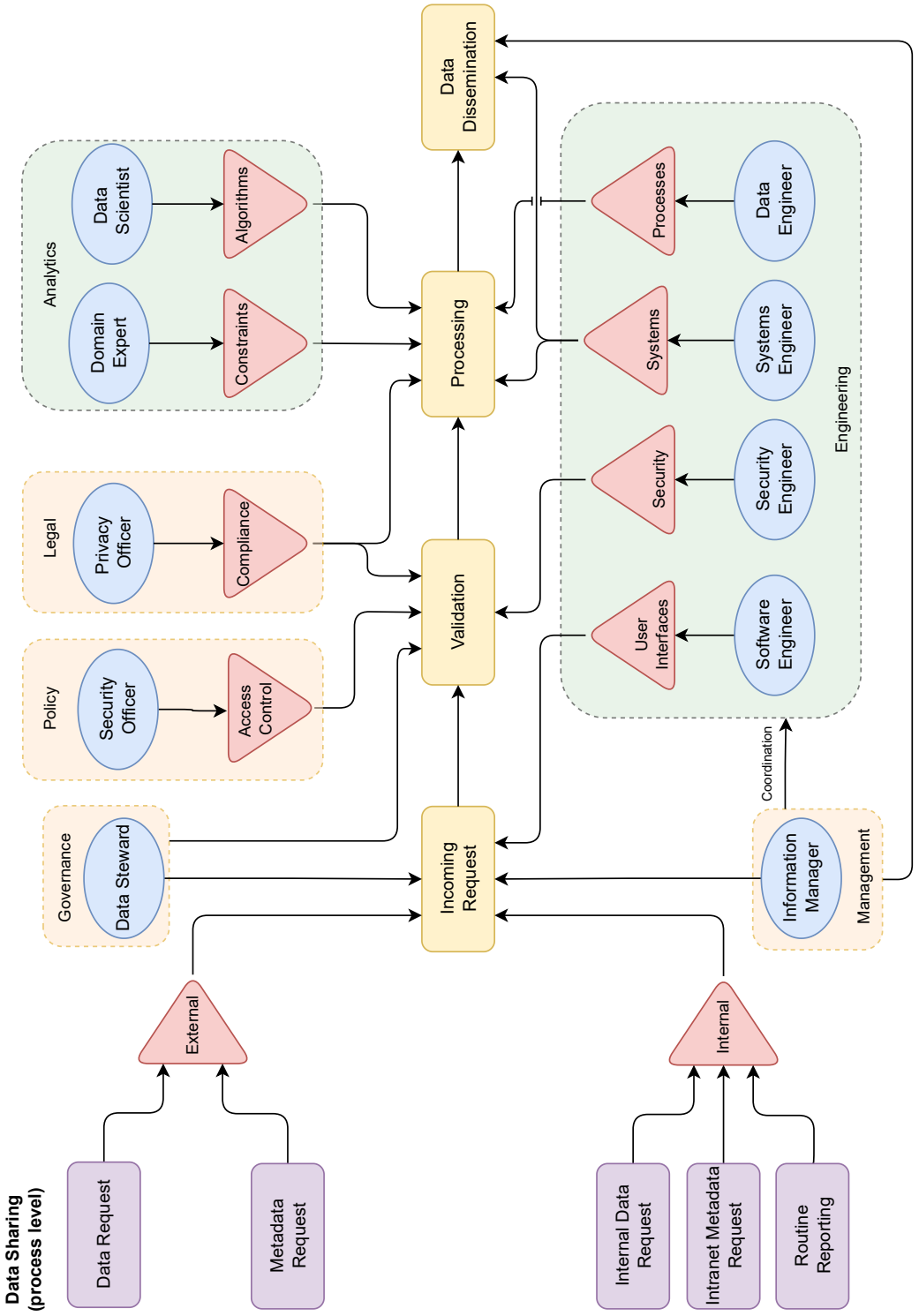
Appendix

A


Data Integration
(process level)
Roles &
Processes
Open Data
Data Exchange
Data Acquisition
Business Process
Monitoring
Research
External
Internal
Incoming & Existing Data
Data Ingestion
Data Preparation
Validation
Coordination
Governance
Data Steward
Policy
Security Officer
Access Control
Legal
Privacy Officer
Compliance
Analytics
Data Scientist
Domain Expert
Quality
Standards
Engineering
Systems
Processes
User Interfaces
Systems Engineer
Data Engineer
Software Engineer
Coordination
Management
Information Manager

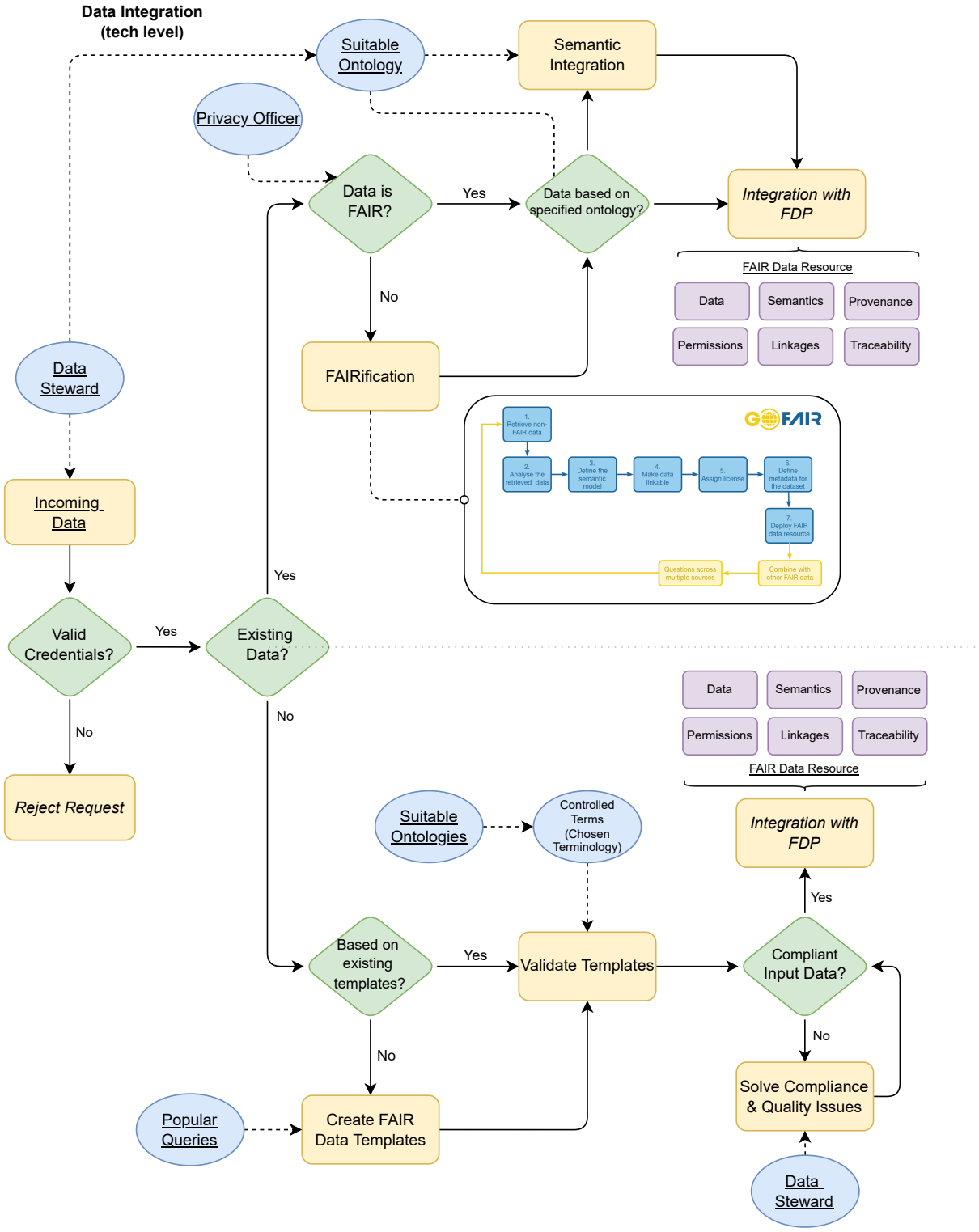
Appendix

B


Data Sharing
(process level)
Data Request
Metadata Request
External
Internal Data Request
Intranet Metadata Request
Routine Reporting
Internal
Incoming Request
Governance
Data Steward
Policy
Security Officer
Access Control
Legal
Privacy Officer
Compliance
Analytics
Data Scientist
Algorithms
Domain Expert
Constraints
Validation
Processing
Data Dissemination
Engineering
Processes
Data Engineer
Systems
Systems Engineer
Security
Security Engineer
User Interfaces
Software Engineer
Coordination
Information Manager
Management

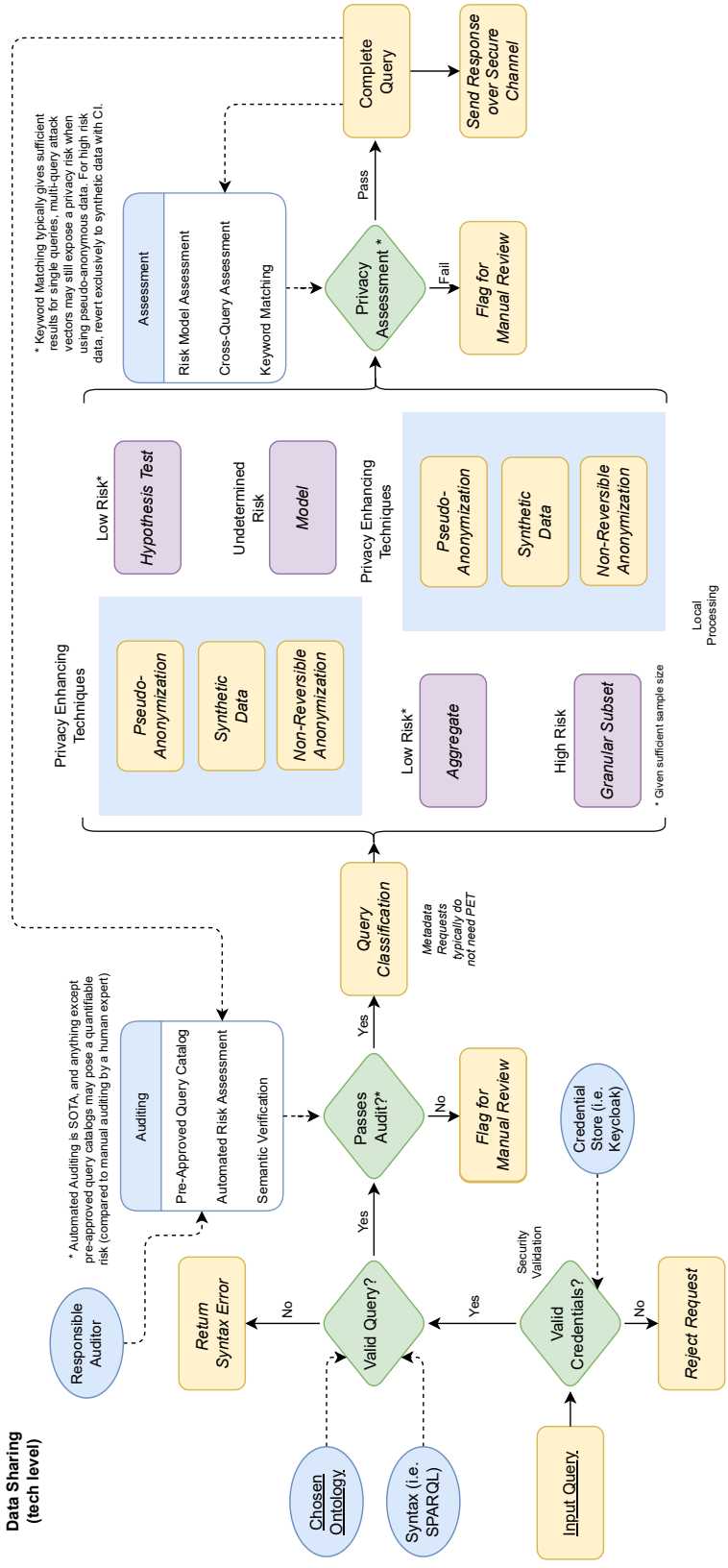
Appendix

C


Data Integration
(tech level)
Suitable Ontology
Semantic Integration
Privacy Officer
Data is FAIR?
Yes
Data based on specified ontology?
Integration with FDP
FAIR Data Resource
Data
Semantics
Provenance
Permissions
Linkages
Traceability
No
FAIRification
Data Steward
Incoming Data
Valid Credentials?
Yes
Existing Data?
Yes
No
No
Reject Request
Suitable Ontologies
Controlled Terms (Chosen Terminology)
Based on existing templates?
Yes
Validate Templates
Compliant Input Data?
Yes
No
Solve Compliance & Quality Issues
Popular Queries
Create FAIR Data Templates
Data Steward

Appendix

D


Data Sharing
(tech level)
Chosen Ontology
Syntax (i.e. SPARQL)
Input Query
Valid Query?
No
Return Syntax Error
Yes
Valid Credentials?
No
Reject Request
Yes
Security Validation
Credential Store (i.e. Keycloak)
Responsible Auditor
* Automated Auditing is SOTA, and anything except pre-approved query catalogs may pose a quantifiable risk (compared to manual auditing by a human expert)
Auditing
Pre-Approved Query Catalog
Automated Risk Assessment
Semantic Verification
Passes Audit?*
No
Flag for Manual Review
Yes
Query Classification
Metadata Requests typically do not need PET
Local Processing
Privacy Enhancing Techniques
Pseudo-Anonymization
Synthetic Data
Non-Reversible Anonymization
Low Risk*
Aggregate
High Risk
Granular Subset
* Given sufficient sample size
Low Risk*
Hypothesis Test
Undetermined Risk
Model
Privacy Enhancing Techniques
Pseudo-Anonymization
Synthetic Data
Non-Reversible Anonymization
* Keyword Matching typically gives sufficient results for single queries, multi-query attack vectors may still expose a privacy risk when using pseudo-anonymous data. For high risk data, revert exclusively to synthetic data with CI.
Assessment
Risk Model Assessment
Cross-Query Assessment
Keyword Matching
Privacy Assessment *
Fail
Flag for Manual Review
Pass
Complete Query
Send Response over Secure Channel

References

E

Literature

[1] Julia Ingrid Lane and Claudia L. Schur. „Balancing access to health data and privacy: a review of the issues and approaches for the future.“ In: *Health services research* 45 5 Pt 2 (2010), pp. 1456–67 (cit. on p. 1).

[2] Christina Tikkinen-Piri, Anna Rohunen, and Jouni Markkula. „EU General Data Protection Regulation: Changes and implications for personal data collecting companies“. In: *Comput. Law Secur. Rev.* 34 (2018), pp. 134–153 (cit. on p. 1).

[3] Haruna Isah, Tariq Abughofa, Sazia Mahfuz, et al. „A Survey of Distributed Data Stream Processing Frameworks“. In: *IEEE Access* 7 (2019), pp. 154300–154316 (cit. on p. 1).

[4] Henry Surendra and S MohanH. „A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing“. In: *International Journal of Scientific & Technology Research* 6 (2017), pp. 95–101 (cit. on p. 1).

[5] Aman Gupta, Deepak L. Bhatt, and Anubha Pandey. „Transitioning from Real to Synthetic data: Quantifying the bias in model“. In: *ArXiv* abs/2105.04144 (2021) (cit. on p. 2).

[6] Yuri Demchenko, Canh Ngo, Cees T. A. M. de Laat, and Craig A. Lee. „Federated Access Control in Heterogeneous Intercloud Environment: Basic Models and Architecture Patterns“. In: *2014 IEEE International Conference on Cloud Engineering* (2014), pp. 439–445 (cit. on pp. 2, 4, 51).

[7] Jakub Konecný, H. B. McMahan, Felix X. Yu, et al. „Federated Learning: Strategies for Improving Communication Efficiency“. In: *ArXiv* abs/1610.05492 (2016) (cit. on pp. 2, 4, 11, 15, 39).

[8] Mikhail Khodak, Renbo Tu, Tian Li, et al. „Federated Hyperparameter Tuning: Challenges, Baselines, and Connections to Weight-Sharing“. In: *ArXiv* abs/2106.04502 (2021) (cit. on p. 3).

[9] Peter Kairouz, H. B. McMahan, Brendan Avent, et al. „Advances and Open Problems in Federated Learning“. In: *ArXiv* abs/1912.04977 (2021) (cit. on p. 4).

[10] Nicola Rieke, Jonny Hancox, Wenqi Li, et al. „The future of digital health with federated learning“. In: *NPJ Digital Medicine* 3 (2020) (cit. on p. 4).

[11] Guodong Long, Tao Shen, Yue Tan, et al. „Federated Learning for Privacy-Preserving Open Innovation Future on Digital Health“. In: *ArXiv* abs/2108.10761 (2021) (cit. on p. 4).

[12] Stefano Savazzi, Monica Nicoli, Mehdi Bennis, Sanaz Kianoush, and Luca Barbieri. „Opportunities of Federated Learning in Connected, Cooperative, and Automated Industrial Systems“. In: *IEEE Communications Magazine* 59 (2021), pp. 16–21 (cit. on p. 4).

[13] Jason Posner, Lewis Tseng, Moayad Aloqaily, and Yaser Jararweh. „Federated Learning in Vehicular Networks: Opportunities and Solutions“. In: *IEEE Network* 35 (2021), pp. 152–159 (cit. on p. 4).

[14] Zhaohua Zheng, Yize Zhou, Yilong Sun, et al. „Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges“. In: *Connect. Sci.* 34 (2022), pp. 1–28 (cit. on p. 4).

[15] Latif Ullah Khan, Walid Saad, Zhu Han, and Choong Seon Hong. „Dispersed Federated Learning: Vision, Taxonomy, and Future Directions“. In: *IEEE Wireless Communications* 28 (2021), pp. 192–198 (cit. on p. 4).

[16] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. „The FAIR Guiding Principles for scientific data management and stewardship“. In: *Scientific Data* 3 (2016) (cit. on pp. 5, 52, 54, 55).

[17] Luiz Olavo Bonino da Silva Santos, Mark Wilkinson, Arnold Kuzniar, et al. „FAIR Data Points Supporting Big Data Interoperability“. In: Sept. 2016, pp. 270–279. ISBN: 9781847040442 (cit. on pp. 5, 54).

[18] Xavier Gansel, M. Arul Mary, and Alex van Belkum. „Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review“. In: *European Journal of Clinical Microbiology & Infectious Diseases* 38 (2019), pp. 1023–1034 (cit. on pp. 5, 16).

[19] Jung ran Park and Yuji Tosaka. „Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms“. In: *Cataloging & Classification Quarterly* 48 (2010), pp. 696 –715 (cit. on pp. 5, 13).

[20] Xiaojun Chen, Shengbin Jia, and Yang Xiang. „A review: Knowledge reasoning over knowledge graph“. In: *Expert Syst. Appl.* 141 (2020) (cit. on pp. 5, 40).

[21] Wanjun Zhong, Jingjing Xu, Duyu Tang, et al. „Reasoning Over Semantic-Level Graph for Fact Checking“. In: *ACL*. 2020 (cit. on pp. 5, 43).

[22] Barend Mons. „Data Stewardship for Open Science: Implementing FAIR Principles“. In: 2018 (cit. on pp. 7, 11).

[23] Wendy C. Y. Li, Nirei Makoto, and Yamana Kazufumi. „Value of Data: There's No Such Thing as a Free Lunch in the Digital Economy“. In: 2019 (cit. on p. 7).

[24] J. Manyika. „Big data: The next frontier for innovation, competition, and productivity“. In: 2011 (cit. on p. 7).

[25] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. „Critical analysis of Big Data challenges and analytical methods“. In: *Journal of Business Research* 70 (2017), pp. 263–286 (cit. on pp. 7, 9, 51).

[26] M. Nottingham, Roy T. Fielding, and Julian Reschke. „Hypertext Transfer Protocol (HTTP): Semantics and Content“. In: 2018 (cit. on p. 8).

[27] Sareh Aghaei, Mohammad Ali Nematbakhsh, and Hadi Khosravi Farsani. „Evolution of the World Wide Web: From Web 1.0 to Web 4.0“. In: *International Journal of Web & Semantic Technology* 3 (2012), pp. 1–10 (cit. on p. 8).

[28] Tim Berners-Lee, James A. Hendler, and Ora Lassila. „The Semantic Web“. In: 2001 (cit. on p. 8).

[29] Amelie Gyrard, Martin Serrano, and Ghislain Auguste Atemezing. „Semantic web methodologies, best practices and ontology engineering applied to Internet of Things“. In: *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)* (2015), pp. 412–417 (cit. on p. 9).

[30] Aidan Hogan, Eva Blomqvist, Michael Cochez, et al. „Knowledge Graphs“. In: *ACM Computing Surveys (CSUR)* 54 (2021), pp. 1 –37 (cit. on p. 9).

[31] Dan Brickley and Ramanathan V. Guha. „Resource Description Framework (RDF) Model and Syntax Specification“. In: 2002 (cit. on pp. 9, 56).

[32] Aidan Hogan. „Web Ontology Language“. In: *The Web of Data* (2020) (cit. on p. 9).

[33] Garrett James Hardin. „The Tragedy of the Commons“. In: *Journal of Natural Resources Policy Research* 1 (1968), pp. 243 –253 (cit. on p. 10).

[34] Haas J. „From Leaders to Rulers, Cultural Evolution and Political Centralization“. In: *Fundamental Issues in Archaeology* (2001) (cit. on p. 10).

[35] Paul B. Roscoe, Christopher Boehm, Carol R. Ember, et al. „Practice and Political Centralisation: A New Approach to Political Evolution“. In: *Current Anthropology* 34 (1993), pp. 111 –140 (cit. on p. 10).

[36] Maryam Darabi, Hoseinali Soltani, Kamran Nazari, and Mostafa Emami. „Social entrepreneurship: A critical review of the concept“. In: 2012 (cit. on p. 10).

[37] E. Ostrom. „Governing the commons“. In: 1990 (cit. on p. 10).

[38] Adekemi Omotubora and Subhajit Basu. „Next Generation Privacy“. In: *Information & Communications Technology Law* 29 (Jan. 2020) (cit. on p. 11).

[39] J. Johnson. „Global digital population as of January 2021“. In: 2021 (cit. on p. 11).

[40] Widup Suzanne, Pinto Alex, Hylender David, Bassett, and Gabriel Langlois. „2021 Data Breach Investigations Report“. In: 2021 (cit. on p. 12).

[41] European Commission. „The EU's Cybersecurity Strategy for the Digital Decade“. In: Dec. 2021 (cit. on p. 12).

[42] B.L Yi. „World needs $5 trillion in annual climate finance by 2030 for rapid action“. In: Oct. 2021 (cit. on p. 12).

[43] Philipp Meschenmoser, Norman Meuschke, Manuel Hotz, and Bela Gipp. „Scraping Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction“. In: *D Lib Mag.* 22 (2016) (cit. on p. 13).

[44] Donald Wheeler. „Understanding Variation: The Key to Managing Chaos“. In: *SPC Press* (1993) (cit. on p. 14).

[45] Walter A. Shewhart. „Statistical method from the viewpoint of quality control“. In: 1939 (cit. on p. 14).

[46] Dennis Heimbigner and Dennis McLeod. „A federated architecture for information management“. In: *ACM Trans. Inf. Syst.* 3 (1985), pp. 253–278 (cit. on p. 14).

[47] André Freitas, Edward Curry, João Gabriel Oliveira, and Seán O'Riain. „Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends“. In: *IEEE Internet Computing* 16 (2012), pp. 24–33 (cit. on p. 15).

[48] Micah J. Sheller, Brandon Edwards, G. Anthony Reina, et al. „Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data“. In: *Scientific Reports* 10 (2020) (cit. on p. 15).

[49] George A. Kuck. „Tim Berners-Lee's Semantic Web". In: *SA Journal of Information Management* 6 (2004) (cit. on p. 16).

[50] Tim Berners-Lee, Roy T. Fielding, and Larry Masinter. „Uniform Resource Identifiers (URI): Generic Syntax". In: *RFC* 2396 (1998), pp. 1–40 (cit. on pp. 16, 56).

[51] Xuefei Yin, Yanming Zhu, and Jiankun Hu. „A Comprehensive Survey of Privacy-preserving Federated Learning". In: *ACM Computing Surveys (CSUR)* 54 (2021), pp. 1 –36 (cit. on p. 16).

[52] Thomas R. Gruber. „Toward principles for the design of ontologies used for knowledge sharing?" In: *Int. J. Hum. Comput. Stud.* 43 (1995), pp. 907–928 (cit. on p. 16).

[53] Michael Uschold and Michael Grüninger. „Ontologies: principles, methods and applications". In: *The Knowledge Engineering Review* 11 (1996), pp. 93 –136 (cit. on p. 16).

[54] Patrick Golden, R. Shaw, and Michael K. Buckland. „Decentralized coordination of controlled vocabularies". In: *ASIST*. 2014 (cit. on p. 16).

[55] Masanobu Taniguchi and Yoshihide Kakizawa. „Elements of Stochastic Processes". In: *Diagnostic Methods in Time Series* (2021) (cit. on p. 18).

[56] Wittawat Jitkrittum, Zoltán Szabó, Kacper P. Chwialkowski, and Arthur Gretton. „Interpretable Distribution Features with Maximum Testing Power". In: *NIPS*. 2016 (cit. on p. 19).

[57] Z. Zhang, J. Wang, C. Jiang, and Z. L. Huang. „A new uncertainty propagation method considering multimodal probability density functions". In: *Structural and Multidisciplinary Optimization* (2019), pp. 1–17 (cit. on p. 20).

[58] Yuan Gao, Weidong Liu, Hansheng Wang, et al. „A review of distributed statistical inference". In: *Statistical Theory and Related Fields* 6 (2021), pp. 89 –99 (cit. on p. 21).

[59] Partha Deb. „Finite Mixture Models". In: *Encyclopedia of Autism Spectrum Disorders* (2021) (cit. on p. 23).

[60] Timothy L. Bailey and Charles Peter Elkan. „Fitting a Mixture Model By Expectation Maximization To Discover Motifs In Biopolymer". In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 2 (1994), pp. 28–36 (cit. on p. 23).

[61] Marc Bocquet, Julien Brajard, Alberto Carrassi, and Laurent Bertino. „Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization“. In: *Foundations of Data Science* (2020) (cit. on p. 23).

[62] Benyamin Ghojogh, Aydin Ghojogh, Mark Crowley, and Fakhri Karray. „Fitting A Mixture Distribution to Data: Tutorial“. In: *ArXiv* abs/1901.06708 (2019) (cit. on pp. 23, 25).

[63] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001 (cit. on p. 24).

[64] Weihong Zhang and Ying Zhou. „Feature-driven optimization method and applications“. In: Jan. 2021, pp. 157–240. ISBN: 9780128213308 (cit. on p. 24).

[65] Gyemin Lee and Clayton D. Scott. „EM algorithms for multivariate Gaussian mixture models with truncated and censored data“. In: *Comput. Stat. Data Anal.* 56 (2012), pp. 2816–2829 (cit. on p. 24).

[66] Simon N. Wood. „Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models“. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (2011) (cit. on p. 25).

[67] François Dubeau and Samir Mashoubi. „Fourier transform of generalized Gaussian distribution for real valued parameter“. In: *Advances and Applications in Mathematical Sciences* 9 (Jan. 2011) (cit. on p. 25).

[68] Paulo Rodrigues and Gilson Giraldi. „Fourier Analysis and q-Gaussian Functions: Analytical and Numerical Results“. In: *Theoretical and Applied Informatics* 27 (May 2016) (cit. on p. 25).

[69] Zaiyong Feng, Ling ya Ye, and Yi Zhang. „On the Fractional Derivative of Dirac Delta Function and Its Application“. In: *Advances in Mathematical Physics* (2020) (cit. on p. 27).

[70] Bamdad Hosseini, Nilima Nigam, and John M. Stockie. „On regularizations of the Dirac delta distribution“. In: *J. Comput. Phys.* 305 (2016), pp. 423–447 (cit. on pp. 27, 31).

[71] Priya Ranganathan. „An Introduction to Statistics: Choosing the Correct Statistical Test“. In: *Indian Journal of Critical Care Medicine : Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine* 25 (2021), S184 –S186 (cit. on p. 33).

[72] Henry F. Inman and Edwin L. Bradley. „The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities“. In: *Communications in Statistics-theory and Methods* 18 (1989), pp. 3851–3874 (cit. on p. 33).

[73] A H nzokem. „Numerical solution of a Gamma - integral equation using a higher order composite Newton-Cotes formulas“. In: *Journal of Physics: Conference Series* 2084 (2021) (cit. on p. 35).

[74] Slavko Simić and Bandar Bin-Mohsin. „Simpson's Rule and Hermite–Hadamard Inequality for Non-Convex Functions“. In: 2020 (cit. on p. 35).

[75] Ana Belén Ramos-Guajardo, Gil González-Rodríguez, and Ana Colubi. „Testing the degree of overlap for the expected value of random intervals“. In: *Int. J. Approx. Reason.* 119 (2020), pp. 1–19 (cit. on p. 35).

[76] A. Torres-Hernandez and Fernando Brambila-Paz. „Fractional Newton-Raphson Method“. In: *Applied Mathematics and Sciences An International Journal (MathSJ)* (2021) (cit. on p. 35).

[77] Andrew M. Cox and Eddy Verbaan. „Exploring Research Data Management“. In: 2018 (cit. on p. 39).

[78] Shaoxiong Ji, Shirui Pan, E. Cambria, Pekka Marttinen, and Philip S. Yu. „A Survey on Knowledge Graphs: Representation, Acquisition, and Applications“. In: *IEEE Transactions on Neural Networks and Learning Systems* 33 (2022), pp. 494–514 (cit. on p. 40).

[79] Miguel Ángel Rodríguez-García and R. Hoehndorf. „Inferring ontology graph structures using OWL reasoning“. In: *BMC Bioinformatics* 19 (2017) (cit. on p. 42).

[80] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. „From Data Mining to Knowledge Discovery in Databases“. In: *AI Mag.* 17 (1996), pp. 37–54 (cit. on p. 42).

[81] Meghamala Sinha and Stephen A. Ramsey. „Using a General Prior Knowledge Graph to Improve Data-Driven Causal Network Learning“. In: *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. 2021 (cit. on p. 43).

[82] Fernando C Pereira. „Formal grammar and information theory: together again?“ In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358 (2000), pp. 1239 –1253 (cit. on p. 43).

[83] Pierre Deransart and Martin Jourdan. „Attribute Grammars and their Applications“. In: *Lecture Notes in Computer Science*. 1990 (cit. on p. 45).

[84] Tamir Tassa. „Finding all maximally-matchable edges in a bipartite graph“. In: *Theor. Comput. Sci.* 423 (2012), pp. 50–58 (cit. on p. 48).

[85] George B. Mertzios, André Nichterlein, and Rolf Niedermeier. „Linear-Time Algorithm for Maximum-Cardinality Matching on Cocomparability Graphs“. In: *SIAM J. Discret. Math.* 32 (2018), pp. 2820–2835 (cit. on p. 48).

[86] Jan van den Brand, Yin Tat Lee, Danupon Nanongkai, et al. „Bipartite Matching in Nearly-linear Time on Moderately Dense Graphs“. In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)* (2020), pp. 919–930 (cit. on p. 48).

[87] Jia bao Liu, Jing Zhao, Jie Min, and Jinde Cao. „The Hosoya Index of Graphs formed by a Fractal Graph“. In: *Fractals* 27 (2019), p. 1950135 (cit. on p. 48).

[88] Peter Damaschke. „Induced Subgraph Isomorphism for Cographs is NP-Complete“. In: *WG*. 1990 (cit. on p. 49).

[89] Q. Li, Zeyi Wen, and Bingsheng He. „A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection“. In: *ArXiv* abs/1907.09693 (2021) (cit. on p. 51).

[90] Viraaji Mothukuri, Reza Meimandi Parizi, Seyedamin Pouriyeh, et al. „A survey on security and privacy of federated learning“. In: *Future Gener. Comput. Syst.* 115 (2021), pp. 619–640 (cit. on p. 51).

[91] Marijn Janssen, Paul Brous, Elsa Estevez, Luís Soares Barbosa, and Tomasz Janowski. „Data governance: Organizing data for trustworthy Artificial Intelligence“. In: *Gov. Inf. Q.* 37 (2020), p. 101493 (cit. on p. 53).

[92] Ali M. Al-Khouri. „Data Ownership: Who Owns 'My Data'?“ In: *International Journal of Management and Information Technology* 2 (2012), pp. 1–8 (cit. on p. 53).

[93] Peter Buneman and Wang Chiew Tan. „Data Provenance: What next?“ In: *SIGMOD Rec.* 47 (2019), pp. 5–16 (cit. on p. 53).

[94] Paul T. Groth, Helena Cousijn, Tim Clark, and Carole A. Goble. „FAIR Data Reuse – the Path through Data Citation“. In: *Data Intelligence* 2 (2020), pp. 78–86 (cit. on p. 53).

[95] Mirjam van Reisen, Mia Stokmans, Mariam Basajja, et al. „Towards the Tipping Point for FAIR Implementation“. In: *Data Intelligence* 2 (2020), pp. 264–275 (cit. on p. 54).

[96] Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, et al. „A Generic Workflow for the Data FAIRification Process“. In: *Data Intelligence* 2 (2020), pp. 56–65 (cit. on p. 54).

[97] RS Gonçalves, MJ O'Connor, M Martınez-Romero, et al. *The CEDAR Workbench: an ontology-assisted environment for authoring metadata that describe scientific experiments. Semant Web ISWC. 2017; 10588: 103–110* (cit. on p. 54).

[98] Annika Jacobsen, Ricardo de Miranda Azevedo, Nick S. Juty, et al. „FAIR Principles: Interpretations and Implementation Considerations“. In: *Data Intelligence* 2 (2020), pp. 10–29 (cit. on p. 54).

[99] Margreet Bloemers and Annalisa Montesanti. „The FAIR Funding Model: Providing a Framework for Research Funders to Drive the Transition toward FAIR Data Management and Stewardship Practices“. In: *Data Intelligence* 2 (2020), pp. 171–180 (cit. on p. 54).

[100] Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. „Gene Ontology: tool for the unification of biology“. In: *Nature Genetics* 25 (2000), pp. 25–29 (cit. on p. 55).

[101] Giancarlo Guizzardi. „Ontology, Ontologies and the "I" of FAIR“. In: *Data Intelligence* 2 (2020), pp. 181–191 (cit. on p. 55).

[102] María Poveda-Villalón, Paola Espinoza-Arias, Daniel Garijo, and Óscar Corcho. „Coming to Terms with FAIR Ontologies“. In: *EKAW*. 2020 (cit. on p. 55).