POLITECNICO DI MILANO
Master of Science in Computer Science and Engineering
Dipartimento di Elettronica, Informazione e Bioingegneria

[Title of the Thesis]

Internal supervisor: Prof. Viola Schiaffonati
Internal co-supervisor: Prof. Letizia Tanca
External supervisor: Prof. Pierre Senellart
External co-supervisor: Prof. Karine Gentelet

M.Sc. Thesis by:
Riccardo Corona, 927975

Academic Year 2020-2021

About this template

With this template I want to give you some input on how to structure your thesis if you develop your thesis with me in Politecnico di Milano. Next to the pure structure, which you should reuse and adapt to your own needs, the document also contains instructions on how to approach the different sections, the writing and, sometimes, even the work on your thesis project itself. Sometimes you will also find boxes like this one. These are meant to provide you with explanations and insights or hints that go beyond the mere structure of a thesis.

I hope this template will help you do the best thesis ever, if not in the World, at least in your life.

Florian Daniel
October 12, 2017

Disclaimer: Sometimes I may make statements that are general, if not over-generalized, personal considerations, or give hints on how to do work or research. Be aware that these are just my own opinions and by no way represent official statements by Politecnico di Milano or its community of professors. If something goes wrong with your thesis or presentation, you cannot refer to these statements as a defense. You are the final responsible of what goes into your thesis and what not.

Acknowledgements: The original template for this document was not created by me. I would love to acknowledge the real creator, but I actually do not know who it is. The template has been passed on to me by a former student, who also didn't know the exact origin of it. It was circulating among students. However, to the best of my knowledge at the time of writing, it seems that Marco D. Santambrogio and Matto Matteucci may have contributed at some point with considerations on structure and funny citations. Both were helpful and enjoyable when preparing this version of the template. I will be glad to add more precise acknowledgements if properly informed about the origins of this template.

Supervisors and co-supervisors

If the supervisor is internal to Politecnico di Milano (a professor or researcher), then on the first page use "Supervisor" plus the titles "Prof." and "Dr." for professors and researches, respectively. If the work was co-supervised by someone else, refer to him/her as the "Co-supervisor." If the work was supervised by someone external to Politecnico di Milano, use "External supervisor" for the external supervisor plus "Internal supervisor" for the internal supervisor that mandatorily must co-supervise the work with the external supervisor.

Optionally, here goes the dedication.

Abstract

The abstract is a small summary of the thesis. It tells the reader in few words (up to one/one and a half page of total text) everything he/she needs to understand:

- ☐ the *context* of the work (e.g., chatbots),
- ☐ the specific *problem* approached by the thesis (e.g., the development of personal bots by non-programmers),
- ☐ if applicable, clearly state the *research questions* you would like to answer (e.g., "is it possible to enable non-programmers to do X using A?"),
- ☐ the three/four *core aspects of the proposed solution* (e.g., use predefined rules, use machine learning, assisted development, etc.),
- ☐ the *concrete outputs* produced by the thesis (e.g., a state of the art analysis, a conceptual/mathematical model, an application, middleware or API, an empirical study with/without users, etc.), and
- ☐ the *findings and conclusions* that one can draw from the evaluation of the approach (e.g., that under some very specific conditions non-programmers are indeed able to implement own chatbots effectively using the proposed technique).

Checklists
Now and there I propose checklists with items, such as the one just above this box. They are meant for you to check if you included all the content that is relevant and that should be included, in order to make your text complete. When reading your thesis, I will look for all these items.

Writing style

This is a M.Sc. thesis. It's neither Facebook nor Twitter nor an email. This is going to be an official document with legal value that will decide on the final mark of your yearlong university career and perhaps even on your future work perspectives. So, you surely don't want to be judged badly because of grammar errors, flawed/wrong vocabulary or superficial layout and/or text structure. It is a must that what you write is always *correct* content- and language-wise (no false statements or claims, no language mistakes), *readable* (no sentences that cannot be understood) and targeted at the *average-skilled reader* (professors, but also your own colleagues).

Plagiarism

This is a M.Sc. thesis. It's neither Facebook nor Twitter nor an email. This is going to be an official document with legal value that will decide on the final mark of your yearlong university career and perhaps even on your future work perspectives – yes, I plagiarized myself here a little bit. So, you surely don't want to copy/paste material from scientific articles, online resources, books, and similar without adequately acknowledging the holders of the respective intellectual property rights. If you do so, it is a must that you properly *cite* each source where you take text or inspiration from. It is fine to do so – actually, citing someone is a compliment! – but it becomes a crime if the source is not cited. Not only M.Sc. titles but also Ph.D. titles have been withdrawn for fraudulent "reuse" of others' intellectual property. Be aware that Politecnico di Milano, like most higher educational institutions that issue university degrees or scientific publishers, may use specialized software to automatically detect plagiarism.

Sommario

Here goes the translation into Italian of the abstract. If the thesis is written in Italian, no translation into English is needed. Hence, one of the following must be checked:

☐ Thesis written in *English*, properly proofread translation needed

☐ Thesis written in *Italian*, no translation needed, chapter omitted

Acknowledgements

If you would like to thank somebody for given support, this is the right place to do so.

Contents

Chapter 1

Introduction

The introduction is one of the core chapters of your thesis. It expands what has already been said in the abstract with additional details on the content and contribution and on the structure of the thesis. It is meant to introduce the reader to the work he/she will be reading in the rest of the document and, most importantly, to get the reader curious about reading on, knowing more about your work.

1.1 Context: [topic]

This thesis is about describing the work you are doing in your final thesis project. You have been working on it for months, and nobody knows the work better than you do. This is great and exactly how things should be: by doing your thesis project you became an expert – if not *the* expert – in this specific field of research and/or technology.

But attention: being the expert is also dangerous when it comes to explaining others what you did and why you think you did a great work that deserves attention (I give it for granted that you work does so). There are only very few people around you (your supervisor and possible co-supervisor, some friends, maybe someone else) who are as expert as you are in this topic. So, if you start in a full-impact fashion to tell that you implemented an extraordinarily cool, new algorithm to solve X, or that you discovered this extremely surprising finding Y, or that you mathematically proofed that Z, etc. (you got it), your reader will not understand anything. Therefore, before talking about what you actually did, you need to introduce the reader to the context of your work, provide the necessary core definitions that are needed to understand the terminology you will be using in the rest of the thesis (if it's not standard IT terminology).

Therefore:

- ☐ Tell the *research area(s)* your work/project focuses on. If you are doing your thesis with me, likely candidates of research areas are Web Engineering, Data Science, Crowdsourcing, Service-Oriented Computing, Business Process Management.
- ☐ Tell possible *sub-areas* that are more specifically related to what you are doing. Again, if you are doing your thesis with me, likely candidates of sub-areas are chatbots, social knowledge extraction, business process matching/modeling, quality control in crowdsourcing, etc.
- ☐ Make the *heading* of your context section self-explaining by substituting "[topic]" in heading 1.1 with the sub-area most relevant to your work. It should read like "Context: quality control in crowdsourcing" or similar.
- ☐ If needed, introduce some *key definitions* (no need to introduce everything here, but be sure that the introduction does not use terminology the reader may not be familiar with). For instance, if you are working on chatbots, this is definitely a term that needs to be introduced here; it's not yet commonly known but it's crucial for the understanding of the rest of the thesis and introduction.
- ☐ Use *examples* to make definitions and ideas concrete and clear.
- ☐ Throughout, make *references* to the relevant literature.

Use of tenses and pronouns

Writing a thesis is writing a scientific document like scientific articles or research publications. There are two conventions that are usually applied in this kind of publications (admittedly, they may seem somewhat odd if not used to):

First, the most used tense is the *simple present*. The thesis is meant to describe a piece of work, from problem statement, to the conception of a solution, its implementation and evaluation. Yet, it's not a novel about your life, and it's not meant to provide a chronological story about what you did and didn't do. Content is presented in an order that is most effective to convey its message, not in time order. In this spirit, it's much more effective to say "in order to get result A, first we do X, then we do Y and then Z," instead of saying "in order to get result A, we did Y after having done X, then we went on doing Z." The order of actions, their interconnections, inputs and outputs already tell the dependency – if properly described. Most of the times, the most effective way to describe a solution or

methodology only becomes clear after trial and error. It's enough to explain the result, not how you got there chronologicallly.

Second, the *pronoun* used to talk about the own work is "our" (work). That is, it is custom to say "we" instead of "I," even if you are writing your thesis alone. However, don't forget about all the people that helped you get there: your supervisor, co-supervisor, colleagues, etc. This may sound strange at the beginning, but, at the other hand, using "I" too often risks to convey the impression that you are self-focused and egoistic, which is never good.

1.2 Scenario and Problem Statement

Now that the reader got the general context of your work and has an intuition of the problem you will be solving in the rest of the thesis, it's time to be clear about which *specific problems* your thesis project is going to solve. One way of doing so is by describing a *scenario* (a description of a real situation, with all its actors, roles, tasks, instruments, etc.) that provides evidence that there are one or more real problems right now that, with the current technology and understanding of the domain, are hard to solve or not solvable at all. If instead the problem(s) can be solved already, it should be evident from the scenario that this is possible only at a prohibiting cost or with unsatisfying guarantees on the quality of the result or not within useful time for the target user.

It's important that the scenario is written in such a way that the reader, after reading it, agrees with you that the problem you are focusing on is a relevant one, one that deserves being studied and solved. Consider that if you convince the reader here that your thesis is needed (after all, that's what this section is about), he/she will be very open to possible solutions and happy to see how you solve it. If instead you fail to convince the reader – let me be harsh – the whole rest of your thesis is useless in the eyes of the reader. This is the worst outcome you want.

Conclude this section by explicitly stating which of the problems evident in the scenario you are approaching. Don't raise false expectations! Never ever tell the reader there are five core problems and then solve only two of them in the thesis, without telling upfront that this is what you intended to do in the first place. As soon as you list problems, the reader wants to see a solution, unless you stop him/her immediately from thinking so by telling that out of the described problems you focus on a subset only, usually because this subset is already a huge research and development problem in its own.

In summary:

- ☐ Describe a *real scenario* that provides evidence of *real problems*.
- ☐ Convince the *reader* that the problems need to be solved.
- ☐ Use an *illustration* or *figure* to help the reader understand.
- ☐ If possible, provide *references* to literature that backs your assessment of the problem.
- ☐ Provide a clear *problem statement* that summarizes what came out of the scenario and your specific focus.

1.3 Methodology

Fixed the problem(s) you want to approach, you can approach it/them in thousands of different ways. Your way is just one of the thousands, and the reader may have (and very likely will have) a very different intuition of how to solve the problem(s) you just pointed out. So, clarify how you intend to proceed:

- ☐ Tell if you follow an existing *methodology* or not; if yes, name it and provide a reference to literature, if available. For example, Design Science [**?**] is a likely methodology to cite here.
- ☐ Tell which of the following *procedures*, *techniques*, *methods* you use in your work and for which purpose (put them also into the right order, so that their application or use makes immediate sense to the reader):
  - ☐ *Systematic literature review, survey*
  - ☐ *Statistical hypothesis formulation and testing*
  - ☐ *Software prototyping*
  - ☐ *Iterative development*
  - ☐ *Participatory design*
  - ☐ *Performance evaluation*
  - ☐ *Comparative studies*
  - ☐ *User studies*
  - ☐ *Expert interviews*
  - ☐ *Simulation/emulation*

- ☐ *Live experiments*
- ☐ *Case studies*
- ☐ *Mathematical theorem proofing*
- ☐ *Mathematical modeling*
- ☐ *Pseudocode*
- ☐ *Graphical modeling* (e.g., UML, ER)
- ☐ *Model-driven development*
- ☐ *Automatic code generation*
- ☐ ...

☐ Tell if you use some special *software instruments* that help you in your work. We are of course not talking about Word or Google Search. Perhaps you can tell that you used R for data analysis or some specific modeling instrument for automated code generation or simulation.

1.4 Contributions

Now that the reader knows what you want to solve and how you intend to proceed, you can anticipate the contributions your thesis makes to the state of the art. Attention, a thesis project may produce lots of different *outputs* (e.g., a software prototype, a set of registrations and transcripts of interviews, datasets collected during experiments) and *contributions* (e.g., a demonstration that some software solutions solves a given problem under well defined conditions, a formal proof that some property holds, empirical evidence that something works as expected). The former are all the artifacts produced throughout the work. The latter refer to *new knowledge* (if you are doing a full thesis) or the most important, *final output* (if you are doing a tesina). Sometimes, outputs and contributions overlap, but not necessarily.

Typical contributions are (multiple choices may apply to your thesis):

☐ A *systematic literature review* of the state of the art providing evidence for some argument

☐ The design of a *model* (mathematical, graphical, algebraic, etc.) describing how to solve a real world problem in a reusable fashion

☐ The drawing of *conclusions* (findings) from the analysis of a dataset describing some physical or virtual phenomenon

☐ The implementation of a *software prototype* solving a real world application problem

☐ The design of a *language* (textual, graphical) enabling others to solve own problems or to solve them easier

☐ *Formal proofs* of correctness, completeness or other properties of the proposed models or theorems

☐ *Objective evidence* from empirical studies (e.g., performance analyses or simluations) that demonstrate that the proposed prototype or solution works / works better than existing software or solutions that solve the same/similar problem(s)

☐ *Subjective evidence* from user studies or expert interviews backing the claims of viability of the proposed problem or solution/artifact

☐ A reasoned *argumentation*, e.g., based on a detailed case study, supporting the viability of the proposed problem or solution/artifact

Thesis vs. Tesina

Let me spend some words on the difference between these two. Before that, however, it is important to clarify the very purpose of your final project, be it a thesis or a tesina (a small thesis). The purpose of it is giving you the possibility to show that, after years of attending classes and giving exams, you are also able to *apply* the knowledge you acquired during your studies. In short, it's all about you showing that you are *mature*. Mature form a knowledge perspective, mature from an application perspective, mature from a work/teamwork perspective, mature from an ethical perspective.

It is common that a thesis project is not very well defined in its beginning and that even the supervisor does not really know how to approach a given problem or which problem to focus on in the first place. This may even be annoying to you, but attention: there is no intention behind it. Your supervisor is not withholding information from you to test you or to see if you get something. It's just the nature of real *problem solving*. If things were clear from the beginning, there wouldn't be any problem! Fledging out the problem and agreeing on a solution and methodology is a core part of you demonstrating your maturity – if not the most important one. *How* you proceed from the inception of the thesis idea to the final solution is as important as *what* you find and/or produce in the end.

This being said, a *thesis* in Politecnico di Milano usually requires you to make a contribution to the literature (the so-called state of the art). Making a contribution – from a science point of view – means creating new *knowledge*, that is, finding something that nobody knew before, demonstrating a property that nobody showed

before, improving the performance of a given system with a new algorithm, and similar. For a thesis, it is therefore not enough to produce a perfectly engineered solution. It is key that you also demonstrate, provide empirical evidence or proof that your solutions performs as claimed. Well, for a *tesina* this last demonstration is usually not required, and the focus is on the engineering of the solution. In addition, perhaps in the case of the tesina the solution to be engineered is also less complex then for a thesis, but this depends on the context and on how you want to measure complexity.

1.5 Structure of Thesis

Here you explain the structure of the thesis, so that the reader knows how to read it. Consider that not every reader wants to read through the whole thesis to find some specific information. Actually, only few will do so (your supervisor and co-supervisor, and the possible reviewer for sure). Many more will just leaf through it and look for specific types of information (e.g., the context of your work, your findings, how you implemented something, which technologies you used). It is your duty to accommodate them all. How? By telling them how your thesis is structured.

Therefore, in this section you provide a brief description (2-3 sentences) for *each* chapter that follows this introduction. Use an itemized or numbered list to structure the text, like this:

- ☐ Chapter 2 introduces the state of the art and...
- ☐ Chapter 3 provides...
- ☐ ...

Structuring text
Besides telling the reader how the content of your thesis is organized into chapters, it is important that you master some basic text structuring techniques. To organize your text there are lots of instruments you can use: chapters, sections, sub-sections, paragraphs, itemized lists, numbered lists, code examples, figures, images, screen shots, captions below figures, tables, and so on. Use them all! Don't write text without structure. Never.

Be aware that the structure of your text, that is, how you present your work, conveys a lot of information about how well you actually understand what you are writing about, how much you care about being clear and helping your reader understand, and how much value you give yourself to your own thesis. A well

structured presentation of content that the reader can understand and agree with is a huge plus in this respect. Text that lacks proper paragraphs, does not use lists where needed, etc. is a minus and also much harder to read (think about how much a well structured text can help you go back ten pages and find concepts you know you read about compared to a text that comes without an easy to memorize formatting and structure). When writing, think about some of your textbooks. Since you are doing an engineering degree, I'm sure these are textbooks that make exemplary use of the different formatting instruments available.

Chapter 2

Socio-Ethical Preliminaries

The aim of this chapter is to provide to the reader preliminary notions of ethical and sociological rather than technical nature.

Starting from the concept of *bias*, passing through *discrimination* and *human rights*, we will discuss about *equality*, *equity* and finally *fairness*, by providing definitions and relevant examples from the literature on these topics, with a focus on the data and computer systems perspective. It is important to emphasize that, despite the exhaustiveness of the definitions, these terms often have different meanings depending of the context of use, and there is a lot of debate on how to interpret them and eventually include all their dimensions in computer systems.

Because of the "dual nature" of this research, this chapter has to be seen as complementary to the next one, in which some technical bases will be provided, together with an overview on the tools adopted.

2.1 Bias

Although the word "**bias**" does not have an intrinsically negative meaning (it is informally used to indicate a deviation from neutrality), it is mostly adopted in contexts where it entails a moral and social dimension. As reported in [13]:

> *We use the term bias to refer to computer systems that* systematically *and* unfairly discriminate *against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate.* [13, p. 332]

Therefore, it is important to underline that unfair discrimination due to bias is strictly related to systematic and unfair outcome, where the word "systematic" is used with the meaning of "regular, which occurs methodically when certain conditions arise".

By following the classification provided in [13], we can distinguish three overarching categories of bias:

- **Preexisting bias**: it has its roots in social institutions, practices and attitudes. Preexisting bias may originate in the society at large or in subcultures and organizations (*societal bias*), but it is also intrinsic in the nature of every human being (*individual bias*), and can enter a computer system either voluntarily or implicitly and unconsciously, even in spite of the best intentions of the system designer. Furthermore, since preexisting bias is often related to historical discrimination of disadvantaged groups, it could lead to the introduction or the exacerbation of representation issues in the data. An example of preexisting (gender) bias is the one present in the society that leads to the development of educational software that overall appeals more to boys than girls [13].

- **Technical bias**: it arises from the resolution of issues in the technical design. Technical bias may originate from design choices, constraints and technological tools, or exacerbate preexisting bias. An example of technical bias, due to technical constraints, is the one of a monitor screen displaying the flight options most relevant to an airline customer: the screen dimension forces a piecemeal representation of the flights and therefore if the ranking algorithm systematically places certain flights on initial screens and other flights on later screens, it exhibits technical bias [13].

- **Emerging bias**: it emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. Emerging bias is strictly related to the specific context of use, and it is the most difficult to detect. An example of emerging bias, caused by a *mismatch between users and system design* due to *different values* (that is, originated when a computer system is used by a population with different values than those assumed in the design), is the one of an educational software embedded in a game situation that rewards individualistic and competitive strategies used by students with a cultural background that eschews competition and promotes collaboration [13].

For the purpose of this research, we will focus on preexisting bias (in particular, societal bias) and technical bias, but it is important to point out that emerging bias should not be underestimated in the long run, especially when it arises from a mismatch between users and system design due to different values, because society is in constant change and systems should be readjusted or reinvented in order to keep up with the present.

A significant example of preexisting (racial) bias, exacerbated by technical bias, is provided in [2]: a commercial tool called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was used in courts in the U.S. to automatically predict some categories of future crime to assist in bail and sentencing decisions. On average, the tool correctly predicted recidivism 61% of the time, but blacks were almost twice as likely as whites to be labeled a higher risk but not actually re-offend. The tool made the opposite mistake among whites: they were much more likely than blacks to be labeled lower risk but go on to commit other crimes.

Other examples, related to gender bias and technology, are given by [15] and [9]. The former is about a research conducted in 2015, in which a tool was used to simulate job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender. One experiment showed that Google displayed adverts for a career coaching service for "$200k+" executive jobs 1,852 times to the male group and only 318 times to the female group. The latter concerns another Big Tech company, Amazon, whose machine learning specialists, back in 2015, discovered that their new recruiting engine was not rating candidates in a gender-neutral way, because the system taught itself that male candidates were preferable by penalizing resumes that included the word "women's".

2.2 Discrimination

Bias can lead to **discrimination**, but what discrimination is and how it occurs is a controversial issue. As reported in [22], often the law, rather than providing a definition of discrimination, defines a list of attributes, called **protected attributes**, that cannot be used to take decisions in various settings. The list is non-exhaustive and includes characteristics such as race, sex, religion, or sexual orientation. Groups of people that are more likely to be discriminated against because of these attributes are therefore classified as *protected groups*.

Trying to elaborate a bit more, we can define discrimination as the result of either one or both the following:

- **Disparate treatment**, or *intentional discrimination*: the illegal practice of treating an entity, such as a job applicant, differently based on a protected attribute such as race, gender, age, religion, sexual orientation or national origin because of a discriminatory motive.

- **Disparate impact**, or *unintentional discrimination*: the result of structural disparate treatment, in which policies, practices, rules or other systems that appear to be neutral result in a disproportionate adverse impact on a protected group. Disparate impact is not based on a discriminatory motive and the discriminating agent is usually unaware of the discrimination.

Protected attributes are mentioned in the article 2 of the Universal Declaration of Human Rights (UDHR), which states:

> *Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.* [3]

Discrimination is therefore strictly related to *human rights*, together with the concept of *equality*, since "equal", "equally" and "equality" itself are recurring words in several articles of the Declaration.

2.3 Human Rights

For what concerns **human rights**, there is a lot of debate on how to incorporate them in computer systems by following a "human-rights-by-design" approach [21], in order to contrast the negative effects of the so called "dual-use technologies": products which may serve legitimate societal objectives but are also used to undermine human rights like freedom of expression or privacy. Of course, reaching this goal would require a culture shift and huge efforts from both national governments and businesses, which should design tools, technologies, and services to respect human rights by default, rather than permit abuse or exploitation of them. A similar concept is proposed in [26], where the authors sketch the contours of a comprehensive governance framework for ensuring AI systems to be ethical in their design, development and deployment, and not violate human rights. This framework should be effective in contrasting *ethics washing*: the practice of fabricating or exaggerating a company's interest in equitable AI systems that work for everyone, a sort of side door that companies use to substitute regulation with ethics.

For the purpose of this research, we can define human rights as "inalienable fundamental rights to which a person is inherently entitled simply because she or he is a human being" [23, p. 3]. A few examples are the rights to life and liberty, freedom from slavery and torture, freedom of opinion and expression, the rights to work and education, and the right to the pursuit of happiness. These norms are concerning every human being, regardless of sex, age, language, religion, ethnicity, or any other status.

2.4 Equality & Equity

Equality is generally intended as "an ideal of uniformity in treatment or status by those in a position to affect either" [5]. The concept of equality is often associated with discrimination mostly because of the article 7 of the UDHR, which states:

> *All are equal before the law and are entitled without any discrimination to equal protection of the law.* [3]

This principle is known as "equality before the law", and establishes that everyone must be treated equally under the law regardless of race, gender, color, ethnicity, religion, disability, or other characteristics, without privilege, discrimination or bias.

However, it is important to distinguish between two different political and social theories:

- **Equality of opportunity**:

  > *The idea that people ought to be able to compete on equal terms, or on a "level playing field", for advantaged offices and positions.* [19]

  This principle is based on the notion of *sameness*, where fairness is achieved through equal treatment regardless of people's needs. Equality of opportunity is usually simply referred as **equality**, and from now on we will adopt the same terminology for this research.

- **Equality of outcome**: the idea that people should have access to resources (possibly of a different nature and to a different extent) in order to be able to reach the same condition. This principle is based on the notion of *need*, where fairness is achieved by treating people differently depending on their endowments and necessities. Equality of outcome is also known as **equity**, and from now on we will adopt the same terminology for this research.

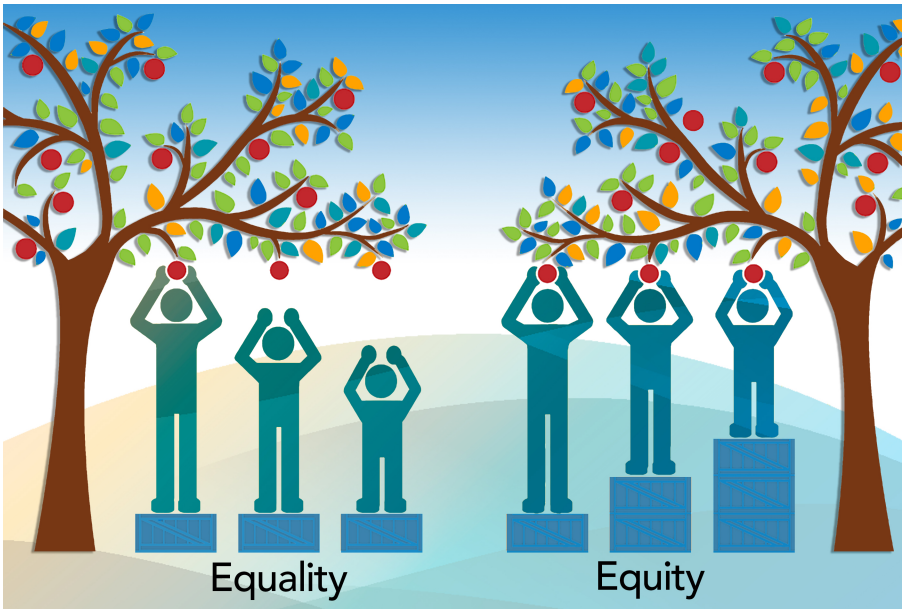

Figure 2.1: Visual example of the difference between equality and equity. ©2014, Saskatoon Health Region. Source: `https://www.nwhu.on.ca/ourservices/Pages/Equity-vs-Equality.aspx`

Figure 2.1 provides a simple example of the difference between equality and equity. Treating people equally, in this scenario, means to give everyone the same one box to reach the fruit, while treating people equitably means to give them as many boxes as they need to achieve the goal. It is important to notice that equity could require (and often requires) unequal treatment.

Moving on the data perspective, we can now extend (or restrict, depending on the point of view) the equality and equity concepts to data equality and data equity. Data equality usually refers to transparency of institutions and companies towards customers regarding the information collected on their account, whereas data equity is used in a different context. The authors of [17] distinguish between four different facets of data equity:

- **Representation equity**: bias may arise because of material deviations between the data and the world represented by the data, often with respect to historically disadvantaged and underrepresented groups. Even when dealing with contemporary data, disparities rooted in historical discrimination can lead to representation inequities and therefore to the introduction or the exacerbation of problems. For example, in the U.S. there has been a lot of discussion about racial disparities concerning COVID-19, regarding both availability of testing

(fewer test sites in minorities neighborhoods, historically poorer) and desire of individuals to be tested (black people more suspicious about the medical system, because of their history of unfair treatments) [17]. Another example, already mentioned in Section 2.1, is related to Amazon and a software developed by the company for screening candidates for employment: the software was trained on the already hired employees and since they were mostly males, females became underrepresented in the data and the software was much more likely to mark women as unsuitable for hiring.

- **Feature equity**: bias may arise because not all the features needed to represent a marginalized group of people and required for a particular analysis are available in the data, or because some of these features are voluntarily removed in the decision-making process. As an example, for a specific study involving transgender people, it may be important to distinguish between their birth name and their self-assigned name.

- **Access equity**: bias may arise because of a non-equitable and participatory access to data and data products across domains and levels of expertise due for instance to the opacity of data systems or the need to respect the privacy of data subjects. A classical example is the one of medical records: making them public could lead to the development of new techniques to eradicate diseases, but on the other side most people are very sensitive about sharing medical information because of the simplicity of re-identify anonymized data, and there are a lot of regulatory constraints on such sharing.

- **Outcome equity**: bias may arise because of a lack of monitoring and mitigation of unintended consequences for any group affected by the system after deployment, directly or indirectly (for example, contact tracing apps may facilitate stigma or harassment).

2.5 Fairness

As discussed in the previous section, both equality and equity aim to achieve **fairness**, despite the different approaches of the two theories, but what fairness really is is a widely debated topic. A very generic definition, taken from [11], depicts it as "the quality of treating people equally or in a way that is right or reasonable".

In the sociological context, fairness is often seen as a synonym of *justice*, and consequently **social justice** is fairness as it manifests in the society,

described by [4, p. 405] as "an ideal condition in which all members of a society have the same rights, protections, opportunities, obligations, and social benefits". Although the literature on this subject does not always agree on their number, we can delineate five interrelated principles of social justice, by following the classification provided in [1]:

- **Access to resources**: a just society should provide services and resources that are available to each different socioeconomic group, in order to give everyone an equal start in life.
- **Equity**: in unjust societies, there are always disenfranchised groups. These groups need to receive more support from the society than privileged ones, in order to move towards the same outcome.
- **Participation**: everyone in a just society, and not just small groups of individuals, should be able to participate in the decisional processes that affect their lives.
- **Diversity**: a just society should recognize the value of diversity and cultural differences, and develop ad-hoc policies with the aim of breaking down societal barriers.
- **Human rights**: a just society should ensure the protection of everyone's civil, political, economic, cultural, and social rights.

Moving back to the data and computer systems perspective, and recalling the aforementioned concepts of equality and equity, we can distinguish between two different concepts of fairness [12]:

- **Individual fairness**: any two individuals who are similar *with respect to a task* should receive similar outcomes. The similarity between individuals should be captured by an appropriate metric function, usually difficult to determine. For example, deciding whether or not to display a specific advertisement is a classification task, and the definition of individual fairness assumes the existance of a task-specific metric (e.g. the number of clicks made by the users) capable of determining, for any two individuals, how (dis)similar they are for the specific task. Individual fairness is strictly related to the idea of equality.
- **Group fairness** (also known as *statistical parity*): demographics of the individuals receiving any outcome - positive or negative - should be the same as demographics of the underlying population. For example, in the problem of predicting if hiring applicants, assuming to divide

them into groups according to their gender, this means the acceptance rates of the applicants from the groups must be equal regardless of the protected attribute. Group fairness equalizes outcomes across protected and non-protected groups, and is therefore strictly related to the idea of equity.

Despite individual and group fairness are not mutually exclusive in theory, in real life it is often hard to conciliate the two approaches. Furthermore, this categorization is not the only possible one: the authors of [24] collected and provided about twenty among the most prominent definitions of fairness, and applied each of them to a case study based on gender-related discrimination, in which the aim was to assign a credit score to people requesting a loan by using "Personal status and gender" as a protected attribute for the decision-making process, operated by a classifier (an algorithm that automatically orders or categorizes data into one or more of a set of "classes", in this case only "good credit score" and "bad credit score").

Among the others, a couple of peculiar definitions, often listed together with individual and group fairness, are the following:

- **Fairness through unawareness**: protected attributes are not used in the decision-making process, and therefore the subsequent decisions cannot rely on them. This "blind" approach relies on *impartiality* and is consistent with the disparate treatment principle, but removing features means losing information, and furthermore there could be features correlated to protected attributes that would not be removed, potentially introducing bias.

- **Counterfactual fairness**: a precise and non-technical definition is provided in [25]:

  > *A model is fair if for a particular individual or group its prediction in the real world is the same as that in the counterfactual world where the individual(s) had belonged to a different demographic group. However, an inherent limitation of counterfactual fairness is that it cannot be uniquely quantified from the observational data in certain situations, due to the unidentifiability of the counterfactual quantity.* [25, p. 1]

  To better clarify the concept, we could imagine a situation in which a software has the task of deciding whether or not to assign a promotion to the employees of a company by looking at their profile that includes,

among the other attributes, sex and race. The software is counterfactually fair if, for each individual, the outcome of the analysis is the same both in the case in which the real values of these attributes are used and in the case in which these values are replaced with others (counterfactuals).

The classifier resulted to be fair depending on the notion of fairness adopted, showing the impossibility of addressing fairness as a unique, broad and inseparable concept. This result is coherent with *Chouldechova's impossibility theorem* [7], which demonstrates, taking three definitions of fairness, the impossibility of satisfying all of them.

Chapter 3

Technical Preliminaries

3.1 Relational Database

When dealing with computer systems, one of the most basic notions, often inappropriately taken for granted, is the one of "**data**", definable as:

> *Information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer.* [10]

Therefore, a large amount of data stored in a computer in some organized manner is called a **database**. To be more precise, a database is "any collection of data, or information, that is specially organized for rapid search and retrieval by a computer" [6]; while the software that supports the management of these data is called a **database management system (DBMS)**.

The history of databases is deeply interconnected with the history of informatics itself, because the problem of how to store and retrieve information appeared as one of the initial challenges of computer creators. However, in the past few decades the rapid and enormous evolution of computer systems and databases led to the adoption and the development of the so called "data models". A **data model** is an abstract representation of an information system, which defines the data elements and the relationships between data elements. The aim of a data model is to give a clear and intuitive overview on how a system looks like, by providing a standardized description of its components, in such a way as to facilitate the understanding of the system itself and the possible integration with other systems.

Nowadays, the most widespread data model is the **relational model**, firstly proposed by E. F. Codd in [8]. The relational model represents a

database as a collection of relations, depicted as tables of values. Each row of the table is a collection of related data values, referring to a real-world entity or relationship between entities. Therefore, we can simply define a **relational database** as a digital database based on the relational model of data. To make it clearer, the following list provides the main terms used in this context, together with a concise explanation.

- **Table**, or **relation**: modeling of a real-world entity or of a relationship between real-world entities.
- **Row**, or **tuple**: single data record.
- **Column**, or **attribute**: property, or feature, of a relation.
- **Cardinality**: total number of tuples of a relation.
- **Degree**: total number of attributes of a relation.
- **Primary key**: attribute, or combination of attributes, that uniquely identifies a tuple among the others.
- **Domain**, or **data type**: set of values that a specific attribute can assume (for example, integer numbers or boolean values).
- **Database schema**, or simply **schema**: blueprint of the database that outlines the way its structure organizes data into tables.
- **Database instance**, or simply **instance**: set of tuples in which each tuple has the same number of attributes as one of the relations of the database schema. It specifies the actual content of the database.

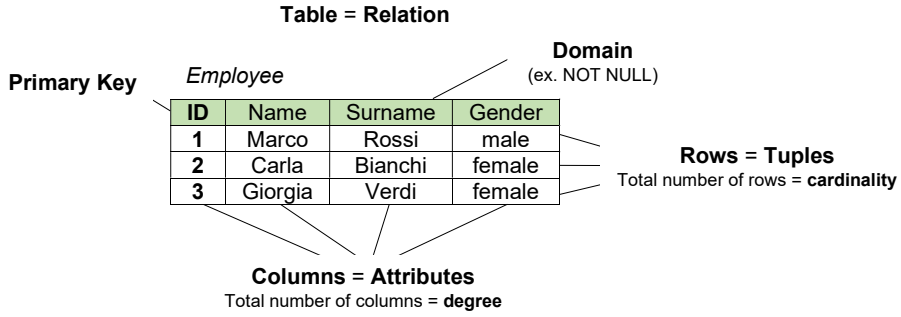

Figure 3.1: Relational model concepts in a trivial example. "Employee" is the name of the real-world entity of reference and therefore of the related table in the model.

Lastly, since this term will be often used in the subsequent sections and chapters, we define a **dataset** as a collection of data. More specifically, since our data are in a tabular format according to the relational model, a dataset simply corresponds to one or more database tables.

3.2 Data Science Pipeline

Because of the broadness of the concept, there is not a unique and precise definition of data management. In general, we can identify it as the process of acquiring, storing, organizing, and maintaining data created and collected by an organization. In [14], the author, referring to [18], classifies *data management*, together with *analytics*, as one of the two sub-processes to extract insights from data, while the overarching process is referred as **data science pipeline**. For the sake of clarity, since the term is the one used in [14], although it is not a concept strictly inherent to this research, we define big data as:

> *Large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.* [20]

However we preferred to adopt the name of "data science pipeline" instead of "big data pipeline", since we will not deal with big data, which are not a concept strictly inherent to this research.

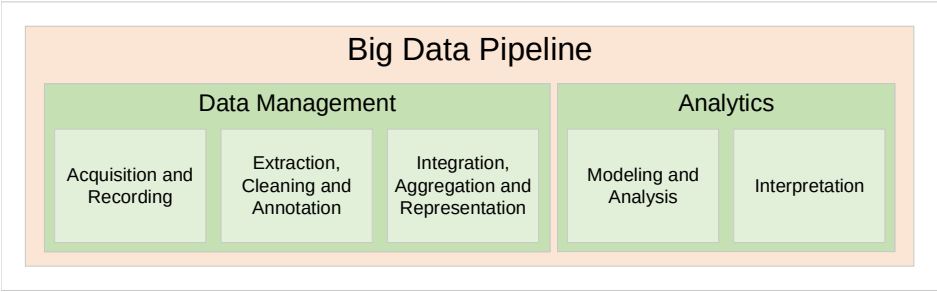

Figure 3.2: Data science pipeline. Image based on the one shown in [14].

Since fairness should be addressed in each phase of the data science pipeline, the subsequent list provides a concise explanation of the operations performed in each step, by following the classification proposed in [16], together with the main potential sources of bias.

- **Acquisition and recording**: data are recovered and captured. In this phase the introduction of bias could derive from some preliminary critical choices we have to deal with, concerning the availability of sources, the identification of who is represented by the data, the definition of what has been measured and of our duties to the people in the data (for example, we may owe them a certain degree of privacy).

- **Extraction, cleaning and annotation**: real data are most of the time messy and dirty, therefore we need to extract the relevant information and clean them, in order to express them in a structured form suitable for analysis. Unfortunately, data cleaning itself is based on assumptions, and wrong assumptions may lead to bias (for example, we may assume missing values in the data as missing at random, while there could be other, maybe ethical, reasons behind).

- **Integration, aggregation and representation**: data analysis often requires the collection of heterogeneous data from different sources, therefore we need to integrate them in order to guarantee syntactic and semantic coherence. Again, we have to rely on assumptions on the world, as for the case of data representation, in which a lot of choices are made in order to decide what to represent, potentially leading to bias (for example, in the context of sentiment analysis we may ascribe sentiment to labels, or we may decide to group age values instead of considering every single year).

- **Modeling and analysis**: before the actual analysis, an abstract model of the data is generated, in order to capture the essential components of the system and their interactions. However, the process of abstraction of concrete data in a conceptual standard model necessarily leads to the loss of information.

- **Interpretation**: a decision-maker, provided with the results of the analysis, has to interpret these results. This process usually requires to examine all the assumptions made and to retrace the analysis, and because of the complexity of the task and the problems that may arise from computer systems (bugs, errors), a human (and therefore impossibly perfectly fair) supervision is needed.

Bibliography

[1] Social Justice - Overview, History and Evolution, Five Principles, 2020. https://corporatefinanceinstitute.com/resources/knowledge/other/social-justice/.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] UN General Assembly et al. Universal Declaration of Human Rights. *UN General Assembly*, 302(2):14–25, 1948.

[4] Robert L Barker et al. The Social Work Dictionary. 2003.

[5] The Editors of Encyclopaedia Britannica. "Equality". *Encyclopedia Britannica*, 2009. https://www.britannica.com/topic/equality-human-rights. Accessed 7 June 2021.

[6] The Editors of Encyclopaedia Britannica. "Database". *Encyclopedia Britannica*, 2020. https://www.britannica.com/technology/database. Accessed 10 June 2021.

[7] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[8] Edgar F Codd. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6):377–387, 1970.

[9] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-thatshowed-bias-against-women-idUSKCN1MK08G.

[10] Cambridge Advanced Learner's Dictionary. "Data". *Cambridge University Press*, 2013. https://dictionary.cambridge.org/dictionary/english/data. Accessed 10 June 2021.

[11] Cambridge Advanced Learner's Dictionary. "Fairness". *Cambridge University Press*, 2013. https://dictionary.cambridge.org/dictionary/english/fairness. Accessed 15 June 2021.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[13] Batya Friedman and Helen Nissenbaum. *Bias in Computer Systems*. Routledge, 2017.

[14] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, 35(2):137–144, 2015.

[15] Samuel Gibbs. Women less likely to be shown ads for high-paid jobs on Google, study shows. *The Guardian*, 8(7), 2015. Available at: https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study.

[16] H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big Data and Its Technical Challenges. *Commun. ACM*, 57(7):86–94, 2014. Available at: https://doi.org/10.1145/2611567.

[17] H. V. Jagadish, Julia Stoyanovich, and Bill Howe. The Many Facets of Data Equity. In *24th International Conference on Extending Database Technology (EDBT)*, 2021.

[18] Alexandros Labrinidis and H. V. Jagadish. Challenges and Opportunities with Big Data. *Proceedings of the VLDB Endowment*, 5(12):2032–2033, 2012.

[19] Andy Mason. "Equal opportunity". *Encyclopedia Britannica*, 2019. https://www.britannica.com/topic/equal-opportunity. Accessed 7 June 2021.

[20] Steve Mills, Steve Lucas, Leo Irakliotis, Michael Rappa, Teresa Carlson, and Bill Perlowitz. Demystifying Big Data: A Practical Guide To

Transforming The Business of Government. *TechAmerica Foundation, Washington*, 2012.

[21] Jonathon Penney, Sarah McKune, Lex Gill, and Ronald J Deibert. Advancing Human-Rights-by-Design in the Dual-Use Technology Industry. *Journal of International Affairs*, 71(2):103–110, 2018.

[22] Teresa Scantamburlo, Andrew Charlesworth, and Nello Cristianini. Machine decisions and human consequences. *arXiv preprint arXiv:1811.06747*, 2018.

[23] M Sepuldeva, Th Van Banning, Gudrún Gudmundsdóttir, Christine Chamoun, and Willem JM Van Genugten. *Human Rights Reference Handbook*. University for Peace, 2010.

[24] Sahil Verma and Julia Rubin. Fairness Definitions Explained. In *2018 2018 ACM/IEEE International Workshop on Software Fairness*, pages 1–7. IEEE, 2018.

[25] Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[26] Karen Yeung, Andrew Howes, and Ganna Pogrebna. AI Governance by Human Rights-Centered Design, Deliberation and Oversight: An End to Ethics Washing. *The Oxford Handbook of Ethics of AI*, page 77, 2020.