École polytechnique fédérale de Lausanne

Assessment of ethical data awareness of EPFL students

Björn Gudmundsson, Mariam Hakobyan, Maxime Quentin Pisa, Rémi Edouard Coudert

Supervisor: Roland Tormey

A Business report for How People Learn class

in the

Department of Social and Human Sciences (SHS) Program

February 7, 2020

Executive summary

0.1 The Purpose and the method of the report

Data processing and privacy are hot topics in the modern day technological landscape. EPFL claims that engineers and scientists that graduate have been well informed on how to make morally and ethically complex decision as part of their education. In this report we aim to answer whether EPFL students tend to process things on a higher moral level and if that capability is something acquired as part of their studies at EPFL.

We conducted a questionnaire aimed at gathering information on how the students process a potentially morally questionable situation. There were 2 different variations of the questionnaire that tested varying levels of severity with regards to public concern and the distribution between the two variants was around 50 percent per variant. The questionnaires were distributed during breaks during lectures for first year, second year and third year students.

0.2 Findings and conclusions

Our findings suggest that when students are presented with a severe enough situation there does not seem to be a statistically significant difference from the first year students to their second and third year counterparts while if the situation could be considered significantly less severe there was a statistically significant difference in whether they decided to give away their data between first, second and third years. This seemed to indicate that the students seem to become more deontological in their later years of study. We were unable however to prove that there seemed to be statistically significant difference between the moral reasoning between the years.

0.3 Recommendations

One of the primary problems we had while conducting the research was the uneven distribution of students between years. The first year students had a good representation while the second and third year students had a very poor representation with a much smaller sample size. The distribution of answers was also very skewed for all years making it hard to make a conclusive statement about their thought process with such a heavily skewed distribution leaning towards one side of the problem. Our recommendation is therefore to try and find a bigger sample size for all of the year groups to get a better representation of each group.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The modern age of Big data and machine learning has opened up a new world of possible issues regarding morality and ethics that is completely unprecedented in all of human history. In the last few years the news have been littered with articles about data breaches, data leaks and misuse of private data by large corporations, most dominantly large tech companies, all over the world. The most evident examples being the Cambridge Analytica scandal and the reports of how russian programmers have been meddling in the politics of other nations such as the presidential elections in the United States and Brexit elections in the UK. This is a consequence of a much more severe problem in the sense that the general public does not seem to be aware of what and how much of their data they are giving away to these organizations.

There are some efforts being made to fight back against the power that these companies hold over users data and how they process it, evident by the recent reports of fines on facebook on their data processing[Reference here] and the new data protections regulations(the GDPR) passed by the European union that went into effect last year. These changes are a step forward but the biggest changes need to happen on the individual level by raising awareness and making sure that future developers are more aware of what they are doing.

1.1 Purpose of this study

EPFL claims that their students are properly trained during their studies to process morally complex situations that they may encounter later on in their careers and that this education is of higher quality than their counterparts who have not received education from EPFL or something of similar pedigree. This research is part of the course called How People learn and aims to answer the questions: "Is a student that has attended EPFL for their bachelors study more data aware than when they first enrolled at EPFL?"

There is a big emphasis on Data Science at EPFL and students are exposed to it as soon as their second year of studies. A good data scientist should be someone that is also aware of ethical issues that arise when handling data and can detect when they should not use a sensitive dataset. We would expect EPFL courses to teach to some extend to students to be critical with the data sets they are asked to process and it would seem reasonable to think that the higher the level of studies, the more aware students are of these issues.

Data Analysis project can sometimes be big in terms of findings and some project are

critical to nowadays' big society issues. In the medical field in particular, the privacy issues are critical but so are the potential findings. For example, some Machine Learning algorithm was recently able to beat a panel of experts in the field of oncology at detecting cancers early looking at MRI images and it looks like in general ML could have a huge impact in health care [1]. We would also like to know how do students respond to the potential of social good output with regards to the sensitivity of the data.

1.2 Hypothesis on the axis

As described above, the purpose of this study is to test students awareness and behavior in data ethical issues. To that end 2 hypothesis have been defined:

- H_0: EPFL students are educated enough to identify data ethical issues, no matter the settings of the problem.
- H_0: The awareness of data ethical issues are increasing with regards to the evolution of EPFL bachelor students.

The variables that are considered for these hypothesis are:

- Project urgency and severity.
- Number of years studied at EPFL.

1.3 The scope of this project

This project aims to investigate whether students at EPFL are data aware. By data aware we mean

- Aware of the context Whether students comprehend the differences in the context of the data source
- Data processing If students are aware of the differences in how their data can be processed
- Impact How much students take into consideration the impact that a given dataset could have.

What this report does not answer since it was outside the scope of the project is for example to name a few.

- Data contents Whether students are aware that some data may be more sensitive than other data and how much that is taken into consideration
- Scenario Whether the scenario of a given data processing research i.e. differences between a university research versus a company funded research.

1.4 Overview of the Report

This report will be covering how the project was structured, how we tackled this problem and what our results indicate. We will begin by discussing the methodology of our research, what we decided to do and decided not to when undergoing this research and why we made those decisions. Then we will move onto our conclusions and interpretations of our findings followed by a more detailed look into at it and why we drew these conclusions.

Chapter 2

Methodology

2.1 The defining issues test

This research was heavily inspired by the research done on the defining issues test 1 and 2(DIT) developed by James Rest [2]. Our methodology drew heavy inspiration from the DIT in the sense that we made a questionnaire that heavily resembles the one developed in the DIT and intends to answer similar, but modified versions of the questions raised by it to be more data awareness centric.

2.2 The contextual design of the case study

Our study was primarily targeted towards information and computer science and communication science students at the bachelors level. The reason for picking these students were primarily to have a more homogeneous sample. By homogeneous we mean in the sense that they are likely to have similar backgrounds in terms of education and interests. We decided to not go with masters students as well for a multitude of reasons. The primary one being that the population of masters students is significantly more diverse than the bachelors students and are much more likely to have received their bachelors education at a different university at EPFL, making our abilities to claim the effects that EPFL had on the students much more limited. Another reason for this choice is that students are much more likely to have taken a break from studies between their masters and bachelors education meaning they are much more likely to have received exposure to similar kinds of situations.

To check the hypothesis H_0, two variations of problems have been considered with different health severity levels. We used a "sudden human death" virus and one that causes an acne-like skin rash. These are described in the variations of the questionnaire B.1 and B.2.

To empathize on the sensitivity and the privacy of the data, our fictional data set comes from full DNA sequencing (with a huge potential of personal information retrieval such as physical characteristics or likeliness of genetic diseases), which was collected with the consent of patients but for a previous research and not the current one. We clearly stated the dilemma such that student can identify it.

2.3 Constructing the questionnaire

When constructing the questionnaire we set out a list of requirements and criteria that we wanted to test out. The questionnaire consists of a scenario that has a potentially morally questionable situation regarding data processing. The students were asked whether they would use the given data set or not. Then the students were asked to answer a set of follow up questions that ask how the students rated certain criteria in terms of importance on a scale from 1 to 5 with 1 being strongly disagree and 5 being strongly agree.

When creating the scenario for the students there were a few things that the scenario wanted to represent. Primarily, the scenario was supposed to expose the students to a situation where the data collection was in a possible moral gray-zone and both sides of the argument had valid reasoning. In order for the students to be able to relate better to the scenario we wanted the students to have a surrogate in the situation that could help them envision them in the surrogate's shoes. To accomplish this we gave the data scientist in the scenario a name(Dominique) and tried to make the data scientist appear as authentic of a person as possible. To avoid introducing gender bias as a potential variable in our results, we gave the data scientist a gender-neutral name to try and avoid this problem. By advice from our supervisor we gave the data scientist the name Dominique that could be interpreted as both a traditionally masculine and feminine name.

2.3.1 Conducting the experiments

We distributed our two variations of the questionnaire to three different classes, in 1st, 2nd and 3rd years of Bachelor respectively.

Since there are less students as the study level increases and that it also depends on optional courses and sections, the number of persons studied is not the same for each year (n = 111, 24 and 32 respectively).

We conducted our experiment in the span of two weeks, in relevant courses right before the beginning of lectures. The data was then processed by hand.

Chapter 3

Conclusions

3.0.1 Conclusive Statement

Our research suggests that *Bachelor students at Computer and Communication sciences are not educated enough to identify data ethical issues in the defined settings of the problem.*

However, we have noticed that *students tend to not have the same decision making process regarding moral and ethical reasoning in the later years of their studies compared to their first year counterparts.*

Students seem to become more deontological in their decision making process as they build up more experience and exposure to data science and evaluate questionable data processing on higher level of moral reasoning than freshman students at EPFL.

This statement holds for both variations, but when the students were faced with a situation that was presented to be severe, then the students seemed to give less weight to their deontological views and became more utilitarian and weighed the end results more heavily than the process of getting into the results.

3.0.2 Results

Discussion of the first hypothesis

The experiment results have shown that *the majority of students answered that the surrogate should work on the project instead of rejecting, given that there is a data privacy issue.*((fig 3.1).

From 168 surveyed students 143 of them (85%) thought that Dominic should accept the project, whereas 23 of them (13.7%) rejected it, considering the data privacy issues. This unbalanced result encapsulates a huge message in it which will be discussed afterwards, and at the same time makes the data analysis hard to process because of the small sample of 'rejected' category.

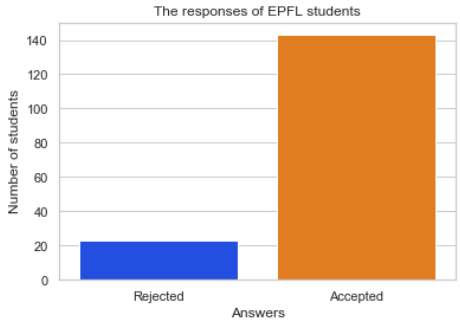


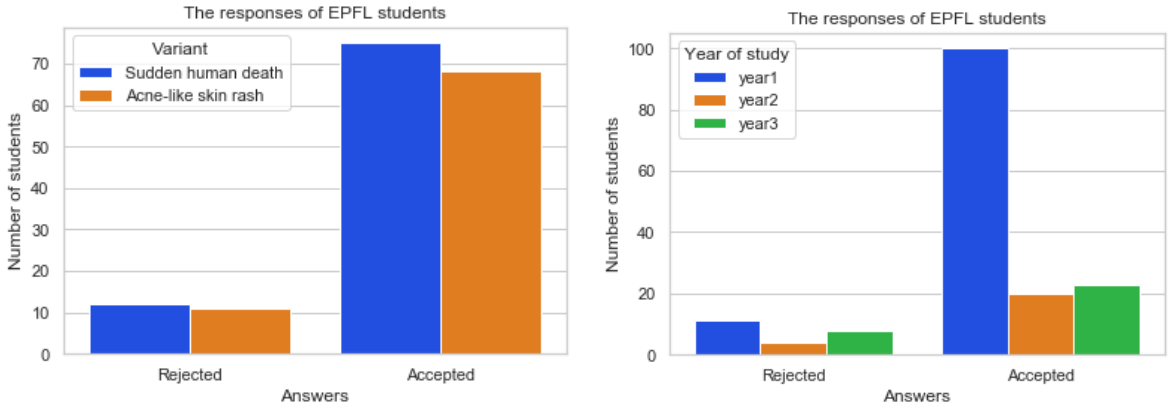
Figure 3.1: *The responses of EPFL students*

These results don't change significantly depending on the variation of the problem ('Sudden Human death' or 'Acne-like skin rash'). See Figure 3.2a. Which makes us think that no matter the severity of the introduced problem, students preferred Dominic to accept the project and work with the data without any consent.

We used a Chi-squared test to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories or groups.

Overall, the significance between variations of the scenario did not affect much on the students' decision. The chi-square statistic is 0.0006, with p-value 0.980542. So the variation group difference is not significant at $p < 0.05$.

This leads to the conclusion that the severity of the disease and thus the potential social good output doesn't affect students' final decision.



(a) **(b)**

Figure 3.2: *Students responses grouped by variations and years*

Regarding to differences of the year study(freshman, sophomore, senior) groups, the chi-squared statistic was 5.3162. The p-value is 0.070081. The result is not significant at $p < .05$. This points out the fact that the final decision is almost the same for different generations of students.

With this two aforementioned chi-squared significance difference statistics we reject the first hypothesis, which states that *EPFL students are educated enough to identify data ethical issues, no matter the settings of the problem.*

While testing the above hypothesis we have found out significant difference on the groups of freshman, sophomore and senior students who read the 'Acne-like skin rash' scenario. The chi-square statistic is 6.475 and p-value is 0.039262. The result is significant at $p < .05$. This suggests that the students' responses to "Acne-like skin rash" scenario changed throughout the years.

Besides the binary answer demanding question(either rejection or acceptance) there have been posed other questions in the questionnaire to see the trajectory of the factors that affected on students decisions. Some of the questions have been discussed in terms of their importance and results.

- **Q1.The urgency of preventing the virus and preventing epidemics** This question was primarily designed to see how important the necessity of making sure that people are not in danger or can be helped is for students. For this question, in regards to our hypothesis, we would expect that the students that accept to use the data set would rate this option highly important, while those that reject would lean towards thinking that it was not as important. Our findings show that it is indeed true that students think this is important when handing out the data set, but students who gave negative answer to the binary question consider that this is not as important as the possible ethical concerns. This is shown by the fact that a bigger percentage seemed to have disagreed with this alternative in their later years.

	Accepted			Rejected		
Q1	Agree	Neutral	Disagree	Agree	Neutral	Disagree
First Year	100	0	0	54.54	18.18	27.27
Second Year	100	0	0	100	0	0
Third Year	100	0	0	75	0	25

Figure 3.3: *Q1.The urgency of preventing the virus and preventing epidemics (in %)*

- **Q2.Detecting the disease is a useful contribution to health research and global health care**. This question highlights the change in decision making process for senior, sophomore and freshman students. This factor is taken more into consideration as the student progress in their studies for both decisions(accepting/ rejecting), with an increase of the proportion of students who agreed and a decrease of the proportion of students who disagreed throughout the years.

	Accepted			Rejected		
Q2	Agree	Neutral	Disagree	Agree	Neutral	Disagree
First Year	75	17	8	27.27	36.36	36.36
Second Year	80	15	5	50	25	25
Third Year	86.95	8.69	4.34	75	12.5	12.5

Figure 3.4: *Q2.Detecting the disease is a useful contribution to health research and global health care(in %)*

- **Q4.The patients rights of being informed about their own data usage in other research.** In this question, we expected the students to consider this factor heavily when rejecting the data set and that in general the number of students rating this option heavily would go up. This question tries to measure if students think that the rights of people to have autonomy over their own data is important. We see the expected change from freshman to sophomore, as a lot more people agreed, and for those that rejected this factor it seemed to be an important point to them. The results for the senior students are not consistent with our hypothesis and show a less general trend than for the first or second year students. Instead of a split opinion or a general consensus it seems to be more skewed and hard to assess what kind of trend was from second to third year or from first year to third year.

	Accepted			Rejected		
Q4	Agree	Neutral	Disagree	Agree	Neutral	Disagree
First Year	33	35	32	81.81	18.18	0
Second Year	55	25	20	100	0	0
Third Year	34.78	17.39	47.82	87.5	0	12.5

Figure 3.5: *Q4.The patients rights of being informed about their own data usage in other research (in %)*

- **Q7.The possible moral/ethical issues regarding the data usage.** This question is asked specifically for those students who might reject the project because of the data ethical issues. The results prove our beliefs. Those students who rejected, mostly considered data ethical factor in their decision, and the proportion of students increases throughout the years.But as we saw from 3.2a the proportion of rejections is very small compared to acceptances, we cannot make any sure statement for the whole image of EPFL students. Whereas for those students who accepted the project, the ethical issues haven't mostly affected on their decision and it has not changed through the years.

	Accepted			Rejected		
Q7	Agree	Neutral	Disagree	Agree	Neutral	Disagree
First Year	31	32	37	81.81	9.09	9.09
Second Year	50	30	20	100	0	0
Third Year	30.43	26.08	43.47	87.5	0	12.5

Figure 3.6: *Q7.The possible moral/ethical issues regarding the data usage (in %)*

- **Q8.How the opportunity to help people in need outweigh the potential privacy/ethical concerns.** This question is asked for those students who may accept the project because of social goodness. Those students who accepted the project mostly considered this factor in their decision and very small amount of students didn't consider this factor in their decision. The results are strange for different years of students who rejected the project, first year students mostly considered this factor even if they rejected the question, second year were neutral, and third year students mostly didn't consider this factor in their decision. This is showing a huge transition in decision making for different years of students, but again because of the lack of the number of rejected students solid inferences are not possible to make.

	Accepted			Rejected		
Q8	Agree	Neutral	Disagree	Agree	Neutral	Disagree
First Year	85	11	4	63.63	18.18	18.18
Second Year	85	10	5	25	50	25
Third Year	82.6	13.04	4.34	12.5	25	62.5

Figure 3.7: *Q8.How the opportunity to help people in need outweigh the potential privacy/ethical concerns(in %)*

- **Q9. The source context of the data set** In this question we were assuming that this would increase in importance for students as they progressed through their studies as they should become more aware and this should be of high importance if a student were to reject the data set. For students who accepted the study, this factor increases throughout the years. This shows that even tho they valued the greater good more than the right to privacy, they still considered this factor strongly. As for the students who rejected the study the results are extremely noisy due to the small amount of rejections and no conclusion can really be drawn.

	Accepted			Rejected		
Q9	Agree	Neutral	Disagree	Agree	Neutral	Disagree
First Year	31	44	25	54.54	9.09	36.36
Second Year	50	20	30	0	50	50
Third Year	43.47	30.43	26.08	50	37.5	12.5

Figure 3.8: *Q9. The source context of the data set*

3.0.3 Summary of scored questions

In general, there seems to be a trend in decision making process in students as they move forward with their studies at EPFL, as they think of the situations with more ethical considerations than the first year students did. We do not consider these changes to be significant for multiple reasons. The primary reason is the size of the data sets that we were working with and the very uneven distribution of answers that we received. Trying to assess when a student would not accept to release their data in this scenario is very hard for this data set since for each year or variant there are handful of students that answered no and the vast majority answered yes. For the analysis of the rejected students a single outlier or "irregular" input can drastically skew the output.

Second Hypothesis

To test the next hypothesis which relates to the data awareness positive correlation with the evolution of students throughout the years, questions were designed in the case study to validate some factors affecting differently on students final decision.

A full table of correlations between Questions and Answers can be found in Table A.1 and Table A.2. The first one corresponds to differences in correlations between variations (no matter the year of study) whereas the second one corresponds to the difference between years of study (no matter the variation). We can point out several interesting points here.

For ease of notation we'll refer as $A.X[Y, Z; W]$ for "Table $A.X$, line Y and column Z for variation/year W".

In these tables, we took the correlations between all questions and the Answer. We coded "Yes" for the answer as 1 and 0 for "No".

Table A.1

1. A.1[A,Q1;1] and A.1[A,Q1;2] are both high (0.58 and 0.44)
2. A.1[A,Q2;1] and A.1[A,Q2;2] are quite different (0.49 and 0.02)
3. A.1[A,Q8;1] and A.1[A,Q8;2] are quite different (0.46 and 0.33)
4. A.1[A,Q4;1] and A.1[A,Q4;2] are quite different (-0.3 and -0.5)
5. A.1[A,Q3;1] and A.1[A,Q3;2] are quite different (-0.18 and -0.5)
6. A.1[A,Q7;1] and A.1[A,Q7;2] are quite different (-0.15 and -0.48)

7. A.1[Q1,Q2;1] and A.1[Q1;Q2] are quite different (0.44 and -0.06)
8. A.1[Q3,Q4;1] and A.1[Q3,Q4;2] are quite different (0.12 and 0.45)
9. A.1[Q3,Q7;1] and A.1[Q3,Q7;2] are quite different (0.12 and 0.5)
10. A.1[Q3,Q9;1] and A.1[Q3,Q9;2] are quite different (-0.02 and 0.34)
11. A.1[Q4,Q7;1] and A.1[Q4,Q7;2] are both high (0.6 and 0.6)
12. A.1[Q4,Q9;1] and A.1[Q4,Q9;2] are quite different (0.05 and 0.25)

As could be expected, we have (1) which suggests that the urgency of the virus is taken into account a lot, and a bit less for variation 2.

All the big differences in $(2, 3, 4, 5, 6, 7, 8, 9, 10, 12)$ suggest that the variation of the questionnaire impacts the decision making of students, even if in general the answer will likely be the same. This is interesting as we do expected differences, but we thought it would impact the overall answer more.

Some correlations are very interesting.

For example (6) shows that the privacy concerns are highly (negatively) correlated with the answer in variation 2 but is almost 0 in variation 1. This might suggests that students that answered "No" in variation 2 thought a lot more about privacy concerns than their counterparts.

(10) shows that the (absence of) guarantee of success is correlated with concerns over the source of the data in variation 2 but is almost 0 in variation 1. This might suggests as well that students that answered "No" in variation 2 thought a lot more about privacy concerns than their counterparts.

Table A.2

1. A.2[A,Q2;1] and A.2[A,Q2;2] are quite different than A.2[A,Q2;3] (0.31, 0.37 and 0.14)
2. A.2[A,Q7;1] and A.2[A,Q7;2] are quite different than A.2[A,Q7;3] (-0.34, -0.41 and -0.18)
3. A.2[A,Q8;1], A.2[A,Q8;2] and A.2[A,Q8;3] are increasing (0.18, 0.42 and 0.7)
4. A.2[Q1,Q3;1] and A.2[Q1,Q3;2] are quite different than A.2[Q1,Q3;3] (-0.13, 0.16 and -0.38)
5. A.2[Q1,Q5;1] and A.2[Q1,Q5;2] are quite different than A.2[Q1,Q5;3] (0.25, 0.26 and 0.05)
6. A.2[Q1,Q8;1] and A.2[Q1,Q8;2] are quite different than A.2[Q1,Q8;3] (0.36, 0.2 and 0.54)
7. A.2[Q10,Q3;1] and A.2[Q10,Q3;2] are quite different than A.2[Q10,Q3;3] (0.20, 0.29 and 0.48)

8. A.2[Q2,Q7;1] and A.2[Q2,Q7;2] are quite different than A.2[Q2,Q7;3] (-0.31, -0.40 and 0.35)
9. A.2[Q2,Q8;1] and A.2[Q2,Q8;2] are quite different than A.2[Q2,Q8;3] (0.22, 0.19 and 0.01)
10. A.2[Q3,Q9;1] is quite different than A.2[Q3,Q9;2] and A.2[Q3,Q9;3] (0.21, -0.05 and 0.09)
11. A.2[Q4,Q5;1] and A.2[Q4,Q5;2] are quite different than A.2[Q4,Q5;3] (-0.09, -0.19 and 0.28)
12. A.2[Q4,Q6;1] and A.2[Q4,Q6;2] are quite different than A.2[Q4,Q6;3] (0, -0.06 and 0.4)
13. A.2[Q5,Q6;1] is quite different than A.2[Q5,Q6;2] and A.2[Q5,Q6;3] (0.21, 0.57 and 0.49)
14. A.2[Q6,Q9;1] is quite different than A.2[Q5,Q9;2] and A.2[Q6,Q9;3] (0.09, 0.24 and 0.3)
15. A.2[Q7,Q8;1] is quite different than A.2[Q7,Q8;2] and A.2[Q7,Q8;3] (-0.08, -0.38 and -0.24)
16. A.2[Q7,Q9;1] and A.2[Q7,Q9;2] are quite different than A.2[Q7,Q9;3] (0.28, 0.04 and 0.44)

The main interesting thing here is (3) : It looks like the higher the level of study, the higher the correlation between the answer and the amount of consideration given to the social output vs privacy issue tradeoff. This suggests that the more student advance in their study, the more they will think about the underlying data source and concerns over it (even though they still accept the project in the end overall). $(6, 8, 9)$ seem to confirm this.

Furthemore, (13) seems to point that first year students weigh in more Dominic's career perspective than 2nd and 3rd year students

3.0.4 Closing statement

Our findings shows a small change in decision making throughout the years at EPFL but nothing indicates that there is sufficient education done through courses. This trend could simply come from students getting older or from more exposition to data sensitive issue during courses.

For future studies we recommend being more careful with the way we setup the scales in the questionnaire. We proposed 5 options : from strongly disagree to strongly agree. This gave us great information but we ended up grouping the strongly categories with their normal counter parts for analysis. This was done because we realized that these categories represented the same answer and that weighting them differently would not really make sense. We think that proposing more choice is great but restraining itself to not ask more than what will ultimately be analyzed it better because it removes useless information for the person taking the questionnaires.

Bibliography

[1] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[2] J. R. Rest, D. Narvaez, S. J. Thoma, and M. J. Bebeau, "Dit2: Devising and testing a revised instrument of moral judgment." *Journal of educational psychology*, vol. 91, no. 4, p. 644, 1999.

Appendices

Appendix A

Correlations between answers

Variation		Answer	Q1	Q10	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
1	Answer	1.000000	0.580146	-0.172080	0.492596	-0.185525	-0.302977	0.198339	-0.174105	-0.150357	0.462635	0.136310
	Q1	0.580146	1.000000	0.010933	0.448475	-0.059725	-0.283951	0.259913	-0.265329	-0.164395	0.620902	-0.003933
	Q10	-0.172080	0.010933	1.000000	-0.095581	0.181807	0.071571	0.211622	0.334744	0.203199	-0.127513	0.103410
	Q2	0.492596	0.448475	-0.095581	1.000000	0.087911	-0.233328	0.162054	0.065695	-0.340084	0.292766	-0.057145
	Q3	-0.185525	-0.059725	0.181807	0.087911	1.000000	0.120566	0.110710	0.382780	0.121390	-0.129064	-0.021559
	Q4	-0.302977	-0.283951	0.071571	-0.233328	0.120566	1.000000	-0.098875	0.136699	0.619177	-0.336910	0.057419
	Q5	0.198339	0.259913	0.211622	0.162054	0.110710	-0.098875	1.000000	0.256748	-0.116375	0.181249	0.079979
	Q6	-0.174105	-0.265329	0.334744	0.065695	0.382780	0.136699	0.256748	1.000000	0.142391	-0.232060	0.226009
	Q7	-0.150357	-0.164395	0.203199	-0.340084	0.121390	0.619177	-0.116375	0.142391	1.000000	-0.258807	0.214485
	Q8	0.462635	0.620902	-0.127513	0.292766	-0.129064	-0.336910	0.181249	-0.232060	-0.258807	1.000000	-0.107314
	Q9	0.136310	-0.003933	0.103410	-0.057145	-0.021559	0.057419	0.079979	0.226009	0.214485	-0.107314	1.000000
2	Answer	1.000000	0.448250	0.056623	0.025587	-0.505544	-0.520772	0.043626	0.178501	-0.480274	0.336545	-0.184481
	Q1	0.448250	1.000000	-0.095017	-0.062569	-0.340178	-0.305133	0.087942	0.086704	-0.231757	0.221070	-0.023792
	Q10	0.056623	-0.095017	1.000000	-0.069174	0.376242	0.101888	0.165107	0.484937	0.184136	-0.182085	0.211979
	Q2	0.025587	-0.062569	-0.069174	1.000000	0.097563	0.068498	0.252765	0.167867	-0.033583	0.046641	-0.065856
	Q3	-0.505544	-0.340178	0.376242	0.097563	1.000000	0.459533	0.022923	0.073037	0.500106	-0.184039	0.343380
	Q4	-0.520772	-0.305133	0.101888	0.068498	0.459533	1.000000	0.012364	0.025791	0.620764	-0.206322	0.257695
	Q5	0.043626	0.087942	0.165107	0.252765	0.022923	0.012364	1.000000	0.376052	0.005281	-0.056093	0.113062
	Q6	0.178501	0.086704	0.484937	0.167867	0.073037	0.025791	0.376052	1.000000	-0.057014	-0.051737	0.034203
	Q7	-0.480274	-0.231757	0.184136	-0.033583	0.500106	0.620764	0.005281	-0.057014	1.000000	-0.072071	0.363041
	Q8	0.336545	0.221070	-0.182085	0.046641	-0.184039	-0.206322	-0.056093	-0.051737	-0.072071	1.000000	-0.241159
	Q9	-0.184481	-0.023792	0.211979	-0.065856	0.343380	0.257695	0.113062	0.034203	0.363041	-0.241159	1.000000

Table A.1: *Correlation between answers for the two different variations*

Year		Answer	Q1	Q10	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
1	Answer	1.000000	6.096661e-01	0.001797	0.317792	-0.361546	-3.823425e-01	0.196759	-0.056826	-0.340280	0.182589	-0.052265
	Q1	0.609666	1.000000e+00	0.040448	0.287838	-0.136100	-3.137446e-01	0.250529	-0.144527	-0.237426	0.363429	0.016823
	Q10	0.001797	4.044779e-02	1.000000	-0.151251	0.207741	1.480592e-01	0.153545	0.402369	0.252634	-0.100148	0.129883
	Q2	0.317792	2.878384e-01	-0.151251	1.000000	0.042048	-1.540241e-01	0.202195	0.040544	-0.310883	0.221804	-0.067716
	Q3	-0.361546	-1.360995e-01	0.207741	0.042048	1.000000	2.833792e-01	-0.006446	0.238899	0.243539	-0.036889	0.214616
	Q4	-0.382342	-3.137446e-01	0.148059	-0.154024	0.283379	1.000000e+00	-0.094004	0.003184	0.615066	-0.164781	0.171413
	Q5	0.196759	2.505287e-01	0.153545	0.202195	-0.006446	-9.400403e-02	1.000000	0.211084	-0.054767	0.150561	0.093730
	Q6	-0.056826	-1.445269e-01	0.402369	0.040544	0.238899	3.183721e-03	0.211084	1.000000	-0.010806	-0.156899	0.084627
	Q7	-0.340280	-2.374258e-01	0.252634	-0.310883	0.243539	6.150661e-01	-0.054767	-0.010806	1.000000	-0.080378	0.280175
	Q8	0.182589	3.634291e-01	-0.100148	0.221804	-0.036889	-1.647806e-01	0.150561	-0.156899	-0.080378	1.000000	-0.175441
	Q9	-0.052265	1.682263e-02	0.129883	-0.067716	0.214616	1.714128e-01	0.093730	0.084627	0.280175	-0.175441	1.000000
2	Answer	1.000000	-1.690309e-01	0.064718	0.370810	-0.146536	-4.382299e-01	0.218604	0.103510	-0.412096	0.428952	0.277884
	Q1	-0.169031	1.000000e+00	-0.027349	-0.056980	0.168881	6.326083e-18	0.266867	-0.233285	-0.287714	0.204909	-0.284297
	Q10	0.064718	-2.734855e-02	1.000000	0.141808	0.299596	1.090830e-02	0.341903	0.502425	0.176835	-0.238384	0.229535
	Q2	0.370810	-5.698029e-02	0.141808	1.000000	0.076341	-2.613636e-01	0.294765	0.314038	-0.404673	0.194895	0.083814
	Q3	-0.146536	1.688808e-01	0.299596	0.076341	1.000000	1.975897e-01	0.402845	0.330937	0.406957	-0.520497	-0.053579
	Q4	-0.438230	6.326083e-18	0.010908	-0.261364	0.197590	1.000000e+00	-0.196510	-0.069786	0.374474	-0.276626	-0.345117
	Q5	0.218604	2.668668e-01	0.341903	0.294765	0.402845	-1.965102e-01	1.000000	0.578262	-0.306814	-0.183464	0.197159
	Q6	0.103510	-2.332847e-01	0.502425	0.314038	0.330937	-6.978632e-02	0.578262	1.000000	0.111277	-0.231656	0.242220
	Q7	-0.412096	-2.877137e-01	0.176835	-0.404673	0.406957	3.744737e-01	-0.306814	0.111277	1.000000	-0.389294	0.045858
	Q8	0.428952	2.049087e-01	-0.238384	0.194895	-0.520497	-2.766256e-01	-0.183464	-0.231656	-0.389294	1.000000	-0.072283
	Q9	0.277884	-2.842975e-01	0.229535	0.083814	-0.053579	-3.451166e-01	0.197159	0.242220	0.045858	-0.072283	1.000000
3	Answer	1.000000	5.661988e-01	-0.348359	0.140909	-0.389143	-4.371401e-01	-0.200502	-0.031904	-0.187977	0.707174	-0.074355
	Q1	0.566199	1.000000e+00	-0.217162	0.150312	-0.384359	-3.723629e-01	0.058216	-0.067859	-0.104950	0.548268	0.091186
	Q10	-0.348359	-2.171621e-01	1.000000	0.031280	0.483908	1.860052e-03	0.132530	0.314350	0.089650	-0.262379	0.157606
	Q2	0.140909	1.503117e-01	0.031280	1.000000	0.245562	2.405425e-01	0.255027	0.222135	0.351170	0.013775	-0.273594
	Q3	-0.389143	-3.843592e-01	0.483908	0.245562	1.000000	3.064666e-01	0.087242	0.252165	0.408508	-0.182898	0.093128
	Q4	-0.437140	-3.723629e-01	0.001860	0.240542	0.306467	1.000000e+00	0.281035	0.407056	0.721045	-0.429289	0.352169
	Q5	-0.200502	5.821607e-02	0.132530	0.255027	0.087242	2.810346e-01	1.000000	0.495431	0.176159	-0.111918	0.086381
	Q6	-0.031904	-6.785921e-02	0.314350	0.222135	0.252165	4.070563e-01	0.495431	1.000000	0.264340	-0.044707	0.304757
	Q7	-0.187977	-1.049504e-01	0.089650	0.351170	0.408508	7.210451e-01	0.176159	0.264340	1.000000	-0.240545	0.443135
	Q8	0.707174	5.482677e-01	-0.262379	0.013775	-0.182898	-4.292893e-01	-0.111918	-0.044707	-0.240545	1.000000	-0.254019
	Q9	-0.074355	9.118635e-02	0.157606	-0.273594	0.093128	3.521689e-01	0.086381	0.304757	0.443135	-0.254019	1.000000

Table A.2: *Correlation between answers for the three different years of study*

Appendix B

Questionnaire

Case Study: Dominic in BioScience

Dominic is a data scientist in a bioengineering company. His supervisor assigned him the task of detecting the special type of virus which causes sudden human death. Detection will be based on complete DNA sequencing data collected from various patients.

This type of virus can be properly treated and prevented if the symptoms are identified early, otherwise, it may cause major epidemics. The task is assigned to Dominic because he is the only person with experience and knowledge in similar types of projects and has higher chances of successfully analyzing the reasons and factors of the virus and reporting significant solutions through data-driven methods.

Because the virus spreads very rapidly and a lot of people die, there is no time to collect new data. His supervisor asks him to use pre-existing data that was collected for a previous research project. While the participants in that other project agreed for their data to be collected and used in research, Dominic feels that the consent form is not very clear for data reusability in other research projects. Therefore he wonders whether the data should be used or not.

He is highly motivated to conduct the research, to prevent the spread of the virus and save lives. But the setting of the dataset puts him in the dilemma of conducting the research or not.

Questions

1. What should Dominic do?

Work on the project using the provided data set. ❑

Reject the project because of the provided data set. ❑

2. Please indicate how much you have considered the following factors in your final decision. Score them according to the scale:

5 = Strongly agree; 4 = agree; 3 = Neither agree nor disagree; 2 = Disagree; 1 = Strongly Disagree

1.	The urgency of preventing the virus and preventing epidemics.	(5)	(4)	(3)	(2)	(1)
2.	Detecting the disease is a useful contribution to health research and global health care.	(5)	(4)	(3)	(2)	(1)
3.	The fact that success was not guaranteed.	(5)	(4)	(3)	(2)	(1)
4.	The patients rights of being informed about their own data usage in other research.	(5)	(4)	(3)	(2)	(1)
5.	Dominic's motivation towards helping people.	(5)	(4)	(3)	(2)	(1)
6.	Dominic's motivation towards his career perspectives.	(5)	(4)	(3)	(2)	(1)
7.	The possible moral/ethical issues regarding the data usage.	(5)	(4)	(3)	(2)	(1)
8.	How the opportunity to help people in need outweigh the potential privacy/ethical concerns.	(5)	(4)	(3)	(2)	(1)
9.	The source context of the dataset.	(5)	(4)	(3)	(2)	(1)
10.	The importance of employee-employer responsibilities.	(5)	(4)	(3)	(2)	(1)

Figure B.1: *Variation 1 of the questionnaire used for our experiment*

Case Study: Dominic in BioScience

Dominic is a data scientist in bioengineering company. His supervisor assigned him the task of detecting the special type of a virus which causes an acne-like skin rash. Detection will be based on complete DNA sequencing data collected from various patients.

This type of disease can be properly treated and prevented if it is detected early, otherwise, it may cause major epidemics. The task is assigned to Dominic because he is the only person with experience and knowledge in similar types of projects and has higher chances of successfully analyzing the reasons and factors of the virus and reporting significant solutions through data-driven methods.

Because the virus spreads very rapidly and a lot of people get affected, there is no time to collect new data. His supervisor asks him to use pre-existing data that was collected for a previous research project. While the participants in that other project agreed for their data to be collected and used in research, Dominic feels that the consent form is not very clear for data reusability in other research projects. Therefore he wonders whether the data should be used or not.

Dominic is highly motivated to conduct the research, to prevent the spread of the virus and save people from infections. But the setting of the dataset puts him in the dilemma of conducting the research or not.

Questions

1. What should Dominic do?

Work on the project using the provided data set. ❑

Reject the project because of the provided data set. ❑

2. Please indicate how much your final decision is based on the following factors. Score them according to the scale:

5 = Strongly agree; 4 = agree; 3 = Neither agree nor disagree; 2 = Disagree; 1 = Strongly Disagree

1.	The urgency of preventing the virus and preventing epidemics.	(5)	(4)	(3)	(2)	(1)
2.	Detecting the disease is a useful contribution to health research and global health care.	(5)	(4)	(3)	(2)	(1)
3.	The fact that success was not guaranteed.	(5)	(4)	(3)	(2)	(1)
4.	The patients rights of being informed about their own data usage in other research.	(5)	(4)	(3)	(2)	(1)
5.	Dominic's motivation towards helping people.	(5)	(4)	(3)	(2)	(1)
6.	Dominic's motivation towards his career perspectives.	(5)	(4)	(3)	(2)	(1)
7.	The possible moral/ethical issues regarding the data usage.	(5)	(4)	(3)	(2)	(1)
8.	How the opportunity to help people in need outweigh the potential privacy/ethical concerns.	(5)	(4)	(3)	(2)	(1)
9.	The source context of the dataset.	(5)	(4)	(3)	(2)	(1)
10.	The importance of employee-employer responsibilities.	(5)	(4)	(3)	(2)	(1)

Figure B.2: *Variation 2 of the questionnaire used for our experiment*