Anatomy of a New Data Science Course in Privacy, Ethics, and Security

Edwin M. Knorr
University of British Columbia
Department of Computer Science
Vancouver, BC Canada V6T 1Z4
knorr@cs.ubc.ca

Giulio Valentino Dalla Riva
University of British Columbia
Department of Statistics
Vancouver, BC Canada V6T 1Z4
giulio.riva@stat.ubc.ca

Orlin Vakarelov
Duke University
Department of Philosophy
Durham, NC USA 27708
o.vakarelov@duke.edu

ABSTRACT

This paper is an experience report describing a course created for the new Master of Data Science program at the University of British Columbia. The course is meant to give students an overview of important and relevant concepts in the security world with a natural bridge to privacy and ethics topics. We do not focus on traditional ethical theories in this course, but rather we explore *information ethics* and the relationship between human dignity and privacy. This course may be of interest to educators attending WCCCE because of its importance in an age of "Big Data"; the increasing and alarming number of real-world privacy, ethics, or security breaches/compromises; and the expectation of some employers that data scientists have an understanding of—and an appreciation for—appropriate management of personal data. We are aware that some schools do not offer a course on these topics at the undergraduate or graduate level; so, this paper might stimulate some ideas for initiating such a course.

CCS CONCEPTS

• **Social and professional topics~Computer science education** • *Social and professional topics~Privacy policies*

KEYWORDS

Computer security, privacy, ethics, data science, information ethics, human dignity, database privacy

ACM Reference Format:
Edwin M. Knorr, Giulio Valentino Dalla Riva, and Orlin Vakarelov. 2018. Anatomy of a New Data Science Course in Privacy, Ethics, and Security. In *WCCCE '18: Western Canadian Conference on Computing Education*, May 4-5, 2018, Victoria, BC, Canada. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3209635.3209640

1. INTRODUCTION

UBC began offering a 10-month Master of Data Science (MDS) specialization in September 2017 because of the need to produce more graduates who are literate in both Computer Science and Statistics [22]. Workplace demand is very strong for data scientists. The MDS program is intense, packing 24 single-credit modules, plus a 6-credit capstone/project course into its curriculum. (A typical UBC undergraduate or graduate course is 3 credits, over 4 months.) The MDS program is aimed at students who have an undergraduate degree, but in neither Computer Science nor Statistics.

The MDS program offers short, core courses in Statistics and Computer Science. Many of them directly or indirectly involve machine learning. One of the 24 required 1-credit courses is the subject of this paper. It is called DSCI 541: "Privacy, Ethics, and Security".

UBC has a large, diverse, multicultural population, including many international students; and this is also true of our MDS program. In our inaugural year (Fall 2016), we accepted about 24 students. In Fall 2017, we had about 400 applicants, competing for about 42 seats. In Fall 2018, we plan to approximately double the intake to 80, which we anticipate will be our steady-state enrollment and capacity for the next few years.

This 4-week course was run by an instructor who delivered the lectures (two 1.5-hour lectures per week), and a Teaching Fellow who delivered the labs (one 2-hour lab per week) and facilitated much of the discussion board. We were assisted by two Teaching Assistants who primarily helped to mark the labs. All of us held office hours. Two of the lectures were delivered by an instructor with a background in Information Ethics. All authors of this paper delivered both offerings of the course to date.

2. DESIGN OF COURSE

This course deals with three related topics: privacy, ethics, and security. Security is directly connected to privacy because security is what we need when trust is missing. In order to achieve privacy, it is necessary that appropriate security precautions be taken; otherwise, attackers can access data in an unauthorized way, leading to possible harm. Encryption is an important tool—both symmetric and public-key. Cryptographic hash functions also appeared multiple times in the course.

Ethics, in its normative reading, is a set of shared principles or values that we need, to live in a civil, just society. The Association for Computing Machinery is in the final stages of formalizing its first updates to its Code of Ethics in over 25 years [1][6]. Much has changed in the computing landscape during this time. Unfortunately, many schools hear very little about ethics since many schools do not include it in their required curriculum. The ACM is encouraging schools to include snippets of ethics in various courses, including introductory programming. For example, students can be instructed about plagiarism rules, right from Day 1—informing them that there are consequences for unethical behaviour, even if someone is just "helping" their friend.

WCCCE '18, May 4-5, 2018, Victoria, BC, Canada
© 2018 Association for Computing Machinery
ACM ISBN 978-1-4503-5805-7/18/05…$15.00
DOI: https://doi.org/10.1145/3209635.3209640

Ethics principles are not laws; but laws often use ethics as a foundation. As a default, computer scientists often use a utilitarian framework for ethical case studies (“deciding what is the right thing to do”) since it is claimed that utilitarianism brings the greatest good for the majority of society (e.g., [8]). However, within the fields of Information Ethics and Computer Ethics, it is well recognized that the utilitarian framework alone is inadequate for most ethical problems in the information age. Instead, we need a broader ethical framework that includes modern ideas about human rights, human dignity, and diversity of culture. We also need a framework that incorporates the latest research in the cognitive and social sciences and connects them to the new problems of the information society.

Compared to most computing courses in ethics and the social implications of computing (e.g., [13]), we take a more current approach. We don’t explore or contrast traditional ethical theories of Kantianism, act utilitarianism, rule utilitarianism, social contract theory, virtue ethics, subjective relativism, cultural relativism, divine command theory, or ethical egoism. Instead, we base the discussion on the latest research in *information ethics*. Two of the specialists in this field are Luciano Floridi [5] and Larry Lessig [11]. They demonstrate the complex nature of privacy and how human dignity and social relations form the foundations for privacy.

Europe, in particular, seems to be placing greater weight on users’ privacy. We discuss the European General Data Protection Regulation (GDPR) that is due for implementation in May 2018. We use it to contrast Europe’s expectations of privacy from North America’s. It has been mentioned that had Mark Zuckerberg begun working on Facebook in Cambridge, UK instead of Cambridge, Massachusetts, that Facebook may never have gotten off the ground. This begs the question of balance among technological innovation, entrepreneurship, and privacy.

A major aim of the course is to demonstrate to the students that, in the settings of computer science and data science, consideration of the ethics of privacy must take central place throughout the entire design process—ethics-by-design and privacy-by-design. Consistent with the broader scholarship in business ethics, we emphasize the duty of all data scientists to think about the ethical implications of their work. They must consider potential and foreseeable effects of their work on the greatest number of stakeholders—and this includes not only employers and customers, but society as a whole. To that end, we highlight a number of general approaches to protecting privacy in the context of information technologies: law, market solutions, social norms, and technological solutions.

We also study some of Michal Kosinski’s work that deals with the “end of privacy” debate [9]. One of the examples that Kosinski uses is that of predicting psychological traits from digital footprints, such as Facebook “likes” and other information that users volunteer, perhaps unknowingly. At the time of writing, there is a major privacy breach involving Cambridge Analytica and Facebook [14]. Cambridge Analytica collected and misused Facebook profiles to gain insight into the psychological traits of users and their friends, particularly in regard to voting patterns and voting influence.

In the course, we also explore implicit and explicit *bias* in data science applications, particularly with respect to diversity. In our most recent offering, students were invited to complete and reflect on the short, free, online course on “Feminist Quantitative Data Analysis” [10] which exposed students to some of the challenges involved in decision-making when females and other underrepresented minorities are part of the analysis. This ties in nicely with some of the students’ other courses in statistics and machine learning. The examples discussed in the short videos include smoking, mental health, poverty, housing, violence against women, the gender pay gap, and financial inclusion. They also considered different worldviews, confounders, mediators, and the individual in a group.

Besides these video clips, students do reading and preparation for their labs outside of class time. There are 4 two-hour labs. The required textbook for the course is Bruce Schneier’s classic *Secrets & Lies*, now in its 15th anniversary edition [18]. Bruce has a long track record in both security and privacy, and his books make for engaging reading, including plenty of examples that are of interest to data scientists (e.g., [16][17]). A recommended, but optional, textbook is Michael Goodrich and Roberto Tamassia’s *Introduction to Computer Security* [7]—one of the most readable and accessible books on security concepts, theory, and (gentle) mathematics that we’ve seen.

3. TOPICS

We have eight 1.5-hour lectures with the topics listed below. We’ve been able to cover almost all of the topics—albeit some quite superficially. In a one-credit course, it is impossible to cover all the topics in depth; hence, we use a best efforts basis, and provide students with the slides and references for future use. We are aware that we probably need to do some selective pruning of topics, given that this is only a one-credit course. The course topics are:

1. Types of sensitive data; privacy; assessing risks; anonymization, de-identification, and re-identification; *k*-anonymity; *l*-diversity; functional dependencies
2. Ethics, legalities, and privacy: issues and case studies; information ethics; what is legal vs. what is ethical; social good vs. individual harm
3. Human dignity; privacy vs. freedom of information; professional and industry codes of ethics; ethics boards; conflicts of interest; whistle-blowing
4. The notion of security as a process and a mindset; complexity and security; the weakest link in a chain; security as a moving target; security terminology with examples
5. Trust, cryptographic hash functions; symmetric and asymmetric encryption; public-key infrastructure (PKI)
6. Continuation of cryptography and security concepts; access control
7. Database security including SQL injections and permissions; Web security including cross-site scripting; cookies
8. Human factors including social engineering and “usable security”; backup and recovery; phishing; ransomware; logging; audit; Blockchain

Students downloaded the lecture slides, labs, and reading materials on GitHub Enterprise, hosted at UBC. The course discussion board was on Slack, and the class was quite active on it, offering reflections and examples of current and other high-profile security and privacy violations. It was great to see so many students actively interested in the material. It also provided us with some ideas for future case studies.

The activities in the 4 weekly labs are organized as follows:

1. A hands-on, coding lab about data anonymization, de-identification, and re-identification using an ad hoc synthetic dataset. Students estimate how much secondary information is needed to uniquely identify an individual. Students identify trade-offs between the usefulness of data and having “too much” anonymity.
2. Case studies in ethics. Students come prepared with their preliminary reading done, some notes written up, and preparation for in-lab discussions of 2-3 case studies.
3. Case studies in security and privacy, including real-world breaches. Again, students come prepared to the lab.
4. More case studies in security and privacy, including real-world security violations, privacy breaches, and other risks. Again, students come prepared to the lab.

4. COURSE-LEVEL LEARNING GOALS

This brief section lists the key outcomes for the course. By the end of the course, students should be able to:

1. Identify situations in which data is sensitive, assess the risks, and articulate a reasoned response.
2. Identify the pros and cons of situations in which data was collected for one purpose and later analyzed for other purposes.
3. Identify trade-offs in security and privacy.
4. Apply ethics principles to case studies. Consider privacy, human dignity, harm, the public good, legal issues, the role of ethics boards, and consent.
5. Implement good security and privacy practices in data storage, use, and reporting.
6. Explain why good security is not a product, but rather a process and a mindset.
7. Argue for why security is complex and difficult, and why perfect security may be unachievable.

5. LEARNING ACTIVITIES

This section explores the various learning activities for the course, and in particular, the labs. We give a rationale for why the learning activities were chosen.

5.1 Labs

Because the course is only one month long, we had no technical labs except for the first lab. The second lab was on ethics and privacy. The third and fourth labs involved case studies in privacy and ethics, with implicit security issues.

Before each lab, students had to prepare by doing some short readings or viewing some videos. To hold students accountable to us and their peers, and to make sure the in-lab group discussions were likely to be meaningful, we had students upload (to GitHub Enterprise) their notes about the readings or videos. This included some questions we asked them to answer before the lab took place.

The labs were conducted in a room whose configuration was 8 rows of 6 seats per row. This was not the best setting because it was very difficult to hear the group discussions, as we walked among the rows. We had to get very close to each group just to hear what was being said. We’ll need to carefully consider classroom logistics for next year, as the MDS program expands.

During the labs, we divided the class into groups of 3 for part of the lab time, and then we had a short class discussion. After this, we merged 2 groups into 1 for more discussion, followed by yet another class discussion. After the two-hour lab, students were asked to create a 1-2 page write-up of the part of the lab that was deemed to be the most interesting, controversial, or fostered the most discussion. (Sometimes, we voted on the topic to be written up.)

Lab 1 utilized material from Lecture 1, and gave a preview of concepts from later lectures. Students were asked to take a synthetic dataset of superheroes and private information about those fictitious characters, and attempt to de-identify data, so that their secret identities would not be leaked. This exercise was done in R. First, students had to determine the *quasi-identifiers* and the *sensitive attributes*. A quasi-identifier is a column (attribute) in a database or a spreadsheet that is not capable of identifying an individual by itself, but could be used in conjunction with other quasi-identifiers to uniquely identify a person. A sensitive attribute is an attribute that data subjects intend to keep to a strict level of privacy, such as, someone’s disease, in a medical scenario. In our case, the data subjects were the superheroes from which the data was collected.

Latanya Sweeney pointed out that 5 quasi-identifiers in typical, publicly-accessible, US demographic datasets could uniquely identify individuals [20]. For example, just knowing somebody’s birthdate, zip code, and gender, would likely be sufficient to uniquely identify 87% of the US population. Furthermore, about 50% of the population would likely be identified just based on birthdate, city/town/municipality, and gender.

Two key concepts for Lab 1 are k-anonymity and l-diversity. These are terms encountered in privacy preservation in databases, and relate somewhat in spirit to functional dependencies in relational database systems [20]. The term k-anonymity simply means that it is highly unlikely that any person can be identified because there will be at least $k - 1$ other people that share the same quasi-identifiers—and theoretically it should be (essentially) impossible to determine the unique identity of an individual. Of course, small subsets of the population might have privileged background information (e.g., a neighbor, close friend, or family member) of a given individual, so all guarantees are off. Some privacy discussions end pretty quickly when someone points out a fine-grained loophole or outlier; but, the anonymity measures we strive for are reasonableness of anonymity (i.e., as much as possible with no foreseeable risks), yet still being able to maintain some *utility* of the de-identified dataset [2]. After all, one of the goals of de-identified datasets is to make them usable for research purposes. This is consistent with US HIPAA principles in de-identifying data for research in health care.

To explain the second key concept, l-diversity, consider the sensitive attribute “disease” (e.g., flu, pneumonia, prostate cancer, HIV). The term l-diversity means that for a given group of individuals sharing the same quasi-identifiers, there would be at least l different diseases (sensitive attributes) for that group of individuals, and therefore you cannot determine which disease someone in that group had. To see why this is important, suppose all 5 patients in a group had hepatitis. This gives away the sensitive attribute for everyone in that group, and there is no privacy for this situation. As a second example, suppose 5 individuals had the same quasi-identifiers, and suppose 4 of those patients were diagnosed with the flu and the other was diagnosed with pancreatitis. In this case, $l = 2$. Here, there are only 2 diseases. Furthermore, if you knew that Bob was one of the 5 individuals, there is an 80% chance that he has the flu, and a 20% chance that he has pancreatitis. Thus, there is also the issue of *probabilistic inference*. We want the diseases to be *sufficiently different* so that there are l possible choices.

Students also worked with cryptographic hash functions, salting, generalization, and outliers in Lab 1. Students had to justify a balance between anonymity and usability.

The second part of Lab 1 asked students to work in pairs and attempt to re-identify each other's anonymized datasets, on the basis of functional dependencies between variables (the capacity of one (or a set of) variables to identify sensitive variables) and some synthetic leaked data not available during the anonymization effort. This second activity reinforced the concept that anonymity is not an absolute property of a dataset, but a property which depends on the environment, namely the insight of the de-anonymizer and the information available. The linkage of data is the problem.

Lab 2 dealt with ethics and privacy, focusing on enabling technologies and their side effects on privacy. We had students read short articles and watch video clips about:

- Europe's General Data Protection Regulation (GDPR) that is scheduled to go into effect in May 2018
- Michal Kosinski's work on "The End of Privacy" debate [9]
- Luciano Floridi's work on information ethics, including a focus on human dignity [5]
- Cambridge Analytica's work on *psychometrics* [12]

Interestingly, two days after our final exam, Cambridge Analytica and Facebook became the subjects of whistle-blowing and a major news story on privacy and ethics violations with respect to certain US elections [14]. The violations involved the scraping and downloading of profiles of 50 million Facebook users (i.e., hundreds of thousands of users and their Facebook friends). This resulted in a $36.4 billion one-day loss in Facebook's market capitalization (i.e., 6.77% drop in stock price) on the first business day immediately after the weekend story ran; and another 2.56% drop the following day—for about a $50 billion two-day loss.

Labs 3 dealt with privacy cases:

- Surveillance by cell phone
- Surveillance by listening devices (i.e., Amazon's Echo device) and Internet of Things sensors in the home: a murder took place in an Echo owner's home, and the owner was a suspect [3]
- Anonymity and the Netflix dataset [19]: A Netflix customer was identified by connecting the title, date, and approximate time of movie reviews on IMDB with the Netflix de-identified dataset—revealing very sensitive information about the rest of her choice of movies, causing personal harm.
- 23andMe and the sale of deidentified data [21]

Lab 4 dealt with additional privacy cases:

- Redlining
- Correct, but discriminatory, algorithms
- OkCupid's experiment in deliberately changing clients' compatibility scores [15]
- A privacy breach involving OkCupid client data—scraping and consolidating semi-private data for other than its intended purpose [23]
- Terms of Service

5.2 End-of-Term Essay

In the inaugural version of the course, we had two in-person quizzes: one at the middle of the course, and one at the end. In year 2, we opted to go with no mid-term quiz, and instead have both an end-of-term, in-person quiz; and an end-of-term essay due approximately a week after the last lecture. The essay (up to 4 pages) was a mini-capstone essay encompassing a variety of themes from the course, but students were able to pick their topic.

The rationale for trading a quiz with the essay is to shift some of the main cognitive effort from remembering key concepts to actually using the said concepts critically in the analysis of situated case studies.

5.3 End-of-Term Quiz

We provided students with some sample questions and answers, and a checklist itemizing the concepts that they were expected to know. We strongly recommended that they be able to define the terms and provide examples. There were few, if any, calculations.

About half of the questions on the end-of-term quiz were multiple-choice, and others were either short-answer or involved writing a few sentences. Most students completed the quiz in under 45 minutes, with about half of the students finishing in about 30-35 minutes. The quiz was administered on students' laptops, and submission was via push on GitHub Enterprise.

6. DISCUSSION

The impact of this course is not trivial to assess. The ultimate goal is to promote reflection on the ethical and security consequences of practices in data collection, analysis, and communication. It has been noted that ethical commitment (e.g., the decision to act based on ethical consideration) is not simply determined by ethical awareness (e.g., the identification of ethical issues) [4].

The reception and the impact of the course has been mixed, yet encouraging. Based on private communications with the first and second cohort of MDS students, it seems that the course is perceived as being more important once the students enter into industry, but it may feel disorienting or irrelevant while they are still studying. The awareness of social and ethical implications in the practice of data science is generally perceived as relevant by data scientists who have been working for at least a year.

The data science industry is often perceived as an industry which does not care about ethical consequences. Regardless of whether or not that perception is true, it seems to be a main factor demoting commitment. Some students, before entering the job market, expressed doubt that a hiring committee would consider a candidate's preparation in privacy, ethics, and security, during the hiring process. Given the current chronicles regarding some of the major competitors in the market, it is hard to deny a bit of truth to those doubts.

Security is a big field, and a moving target. Fortunately, the key concepts haven't changed in decades. However, hackers have gotten smarter and more mischievous—and privacy, ethical, and security breaches seem to be occurring with increasing frequency. The notion of trust is getting harder to quantify.

6.1 Sustainability

The course is highly sustainable, both in terms of technologies and case studies. The modular format of the lab activities (for Labs 2, 3, and 4) ensures the possibility of updating the case studies following the social relevance of new cases. We can continue to use the default labs, or we can selectively replace them with privacy, ethics, or security breaches from the news. The students themselves are frequently suggesting items, as we found out on Slack. So when a relevant story breaks, such as the Cambridge

Analytica and Facebook story, we can inject it into the course, and get engaging discussion both on Slack and in the classroom.

6.2 Scalability

The next cohort of about 80 students represents a challenge mostly for the class-wide discussion sessions. In particular, it is important to ensure that every student actively participates in the discussions, and that his or her diverse experiences and points of view are acknowledged, and encourage interaction. We are considering splitting the class into two or three lab sections next year to better facilitate the conversations.

7. FUTURE WORK

The direction is to keep the course together, and to work on integrating references to this course in other MDS courses, so as to make evident this course's relevance across the curriculum. Also, it may be interesting to promote more outward-facing discussions. So far, only a few students have published (in public posts, blogs, etc.) what they have been writing for this course; but, it may be good to expose their ideas to a more general audience. Finally, the perception or the reality of the data science industry as an *a-ethical* industry needs to be addressed if we want the course to have any impact.

REFERENCES

[1] ACM. 2018. ACM Code of Ethics and Professional Conduct: Draft 3. https://ethics.acm.org/wp-content/uploads/2017/12/diff3.pdf?32ddaa

[2] Jules J. Berman. 2013. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Morgan-Kaufmann (Elsevier).

[3] Sarah Buhr. 2016. "An Amazon Echo May Be the Key to Solving a Murder Case", *Tech Crunch*, December 27, 2016.

[4] Simon Critchley. 2013. *Infinitely Demanding: Ethics of Commitment, Politics of Resistance*. Verso.

[5] Luciano Floridi. 2010. *The Ethics of Information*. Oxford University Press.

[6] Don Gotterbarn, Amy Bruckman, Catherine Flick, Keith Miller, Marty J. Wolf. 2018. "ACM Code of Ethics: A Guide for Positive Action", *CACM*, 61(1), 121-128. https://cacm.acm.org/magazines/2018/1/223896-acm-code-of-ethics/fulltext

[7] Michael Goodrich and Roberto Tamassia. 2011. *Introduction to Computer Security*, Addison-Wesley (Pearson).

[8] H.V. Jagadish. 2017. "Data Science Ethics", Coursera Online Course (MOOC), University of Michigan.

[9] Michal Kosinski. 2017. "The End of Privacy". CeBIT 2017 Keynote Presentation, March 23, 2017, https://www.youtube.com/watch?v=NesTWiKfpD0

[10] Heather Krause. 2018. "Feminist Quantitative Data Analysis", Short video course, https://app.ruzuku.com/courses/25230/members/245899/edit.

[11] Lawrence Lessig. 2006. *Code: And Other Laws of Cyberspace, Version 2.0*. Second revised edition, Basic Books.

[12] Alexander Nix. 2016. "The Power of Big Data and Psychographics", Concordia Summit 2016, https://www.youtube.com/watch?v=n8Dd5aVXLCc

[13] Michael J. Quinn. 2016. *Ethics for the Information Age*. 7th edition, Pearson.

[14] Matthew Rosenberg, Nicholas Confessore, and Carole Cadwalladr. 2018. "How Trump Consultants Exploited the Facebook Data of Millions", *The New York Times*, March 17, 2018.

[15] Christian Rudder. 2014. "We Experiment on Human Beings! (So does everyone else.)" OkCupid Blog, https://theblog.okcupid.com/we-experiment-on-human-beings-5dd9fe280cd5

[16] Bruce Schneier. 2015. *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. Norton.

[17] Bruce Schneier. 2014. *Liars & Outliers: Enabling the Trust that Society Needs to Survive.* Wiley.

[18] Bruce Schneier. 2000 and 2015. *Secrets & Lies: Digital Security in a Networked World*. Wiley.

[19] Bruce Schneier. 2007. "Why 'Anonymous' Data Sometimes Isn't", *Wired*, December 12, 2007.

[20] Latanya Sweeney. 2002. "*k*-Anonymity: A Model for Protecting Privacy", *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570.

[21] Eric Topol. 2015. *The Patient Will See You Now: The Future of Medicine is in Your Hands*. Basic Books.

[22] UBC. 2018. UBC's Master of Data Science Program Web Site: https://masterdatascience.science.ubc.ca/

[23] Michael Zimmer. 2016. "OkCupid Study reveals the Perils of Big-Data Science", *Wired*, May 14, 2016.