Fall 2021	**CSC 466:** Knowledge Discovery from Data	Alexander Dekhtyar

Ethics in Data Mining Project

Due: Friday, December 10, 2021, 11:59pm

Preface

This year, our coursework is structured largely around the technical aspects of Knowledge Discovery in Data. In class we talk about how to achieve specific goals, how to gain insight into the data, and how to answer important questions about it.

However, KDD also has what I call a **sinister** aspect. KDD enables humans to achieve things that were not achievable before. It allows us to understand things about other human beings and their behavior, that, at other times may have remained private. KDD methods can penetrate security and privacy or individuals, organizations, governments, etc...

The class has a learning objective that states that the class will make you think about the societal impacts of KDD technologies. In the past it has been achieved by proposing a semi-plausible[1] scenario and asking the teams to design a KDD system around it, or by asking the students to write a science fiction dystopia story featuring KDD technologies.

This quarter, the assignment is different.

Assignment

This project is **individual** assignment. You are welcome to discuss your work with others in the class (in fact, this is encouraged), but the final deliverable, your report, has to be your work.

Over the past 20 years, there have been a number of situations that data became available either fully publicly (i.e., anyone can access it), or to a select set of individuals, *that was either not indended for public consumption*, or *whose availability had far-reaching implications beyond the original intent of the data owners.*

Examples of this include a variety of data leaks, datasets of stolen bank transactions, password databases, celebrity photo archives, unearthed document caches, as well as public releases of datasets by the dataset owners themselves that allowed for privacy breaches.

In each such case, a data scientist or a machine learning engineer faces a choice. On one hand, the data may present interesting insight, and provide public with new information that may be important to the public discourse. On the other hand, the means by which some data had been obtained may be on the wrong side of the national and international law, the information contained in the datasets may be potentially harmful to individuals.

As part of this assignment each of you will select one such example of data that has been made availble. You will describe a variety of uses for such a dataset, and you will then take a position on whether the dataset should or should not be used, and will construct an argument in support of your position on ethical grounds. Specifically:

1. **Step 1.** Search the world wide web for information about a dataset that looks interesting to you. Select one dataset, and write a short, one-paragraph description of your choice of data, and the reasons why it is interresting to you. Submit the paragraph as a PDF file with your name and section number, by end of the day Firday, November 19.
2. **Step 2. The Machine Learning Engineer view.** Study the information available in the dataset of your choice. If the data is available publicly, you can study the data itself. If it is not available - find third-party descriptions of the data to give you an idea of what information is stored there. Based on your understanding of the information available in the dataset, determine what kinds of data analysis (statistical, machine learning, other KDD methods) could be performed, and what kinds of questions about this data/insight can be obtained.
3. **Step 3. The Data Ethicist.** Study the information about the responsible use of data (I will provide some resources, you can also find your own), and on societal implications of the use of data of questionable provenance. Identify the ethical challenges associated with the use of the dataset you selected for this assignment.
4. **Step 4. The Decision-maker.** Based on your analysis undertaken in Steps 2 and 3, determine whether the dataset you have selected should be used for data analysis; if yes: what type of data analysis it could be used for and what type of data analysis should be off-limits, and present an argument in support of your position.

Your final deliverable is a paper that combines all four steps into a single narrative. The narrative shall contain the following parts:

1. **Introduction.** A very brief and informal preview of the rest of the paper.

[1] Every single scenario proposed in the past is, at present, very feasible.

2. **Dataset description.** The description of the dataset, its provenance/history, its contents, and the history of its public use.

3. **Dataset Uses.** The **Step 2** narrative in which you describe the types of analysis this dataset can be used for.

4. **Ethical Challenges.** The **Step 3** narrative in which you describe the issues and challenges assoicated with the use of the dataset.

5. **Suggested Uses.** The **Step 4** narrative in which you stake out a personal position, and provide a supporting argument.

Please make sure to properly cite all your sources, including web pages on which data may be available, and any on-line articles and other resources you used in your work on this assignment.

Comments and Expectations

While this seems like a lot of work, I expect a moderately-sized paper (10-12 pages) *at best*.

The paper should be professionally typeset using word-processing software (use of LaTeX is encouraged, but not required. Google Docs or MS Word, or analagous word processing software sufficient).

The choice of the dataset, in many ways, determines how difficult it might be for you to both learn things about it, and stake out a clear position. Choose the dataset you are comfortable with studying and arguing about.

As a hint, the inherrent conflict you are trying to observe and resolve is that between the potentially questionable origins/provenance of the data and the public good that some data analysis may achieve, as well as the conflict between the abovementioned public good, and the potential for harm to specific individuals, organizations, causes, etc.

People who released data have been called heroes, and/or traitors, have become fugitives and/or have gone to jail. Some data may have been stolen by hacker groups hostile to (for example) the US. Some data remains in public domain but has the ability to cause harm. Some countries regulate who is entitled to an injunction against the use of data to harm them. All of these are potentially interesting considerations that you could take into account.

What to Submit

Proposal. Submit the PDF of your single paragraph dataset description including your name and section number by the end of the day **November 19 (Friday)**.

Both sections should submit as follows:

```
$handin dekhtyar 466-paragraph <file>
```

Paper. Submit your final report for this assignment by the end of the day, **December 10 (Friday)** The report should be submitted in PDF format and shall contain a title, your name, email address and section number. Both sections should submit as follows:

```
$handin dekhtyar 466-paper <file>
```

Good Luck