

Ethics for Data Scientists

Paul Pidou

EURECOM

Overview

Scope of the class

- **This class is about:**

- ▶ The social responsibility of Data Scientists
- ▶ Data & Model validity
- ▶ The need for a Code of Ethics for Data Scientists

- **This class is NOT about:**

- ▶ Laws: legality doesn't imply ethicality
- ▶ Philosophy: we will not dive into complex notions of ethics

In a nutshell, this class is a simple framework for thinking about the societal impact of a Data Scientist's job

Introduction

Need for Codes of Ethics & regulations through history

● Birth of modern journalism ethics

- ▶ As modern journalism takes over, its role in a democracy has been heavily debated
- ▶ Journalists have a role of mediator between the general public and policy-making elites
- ▶ A Code of Ethics of the profession appeared and evolved over time

● Industrial Revolution

- ▶ First industrial revolution started in the late 17th century
- ▶ Second industrial revolution followed in the mid-19th century
- ▶ Increased use of coal, petroleum & chemicals
- ▶ First Clean Air Act (US) was enacted in 1955
- ▶ Kyoto Protocol was signed in 1997 and is effective since 2005

Codes of Ethics and **regulations** appear after domains emerged and evolved with them

Information Revolution

- **Most of the data in the world today has been created in the last few years alone**
 - ▶ It is a major shift in the relationship we have with our environment
- **It is now easier to collect data than ever before**
 - ▶ People are the producers and/or the consumers of those data
 - ▶ Data are fundamentally about people
- **Who owns the data?** (e.g. if I take a picture of you)
- **Sense of responsibilities: call for a *Hippocratic Oath***

What are Ethics?

- **Ethics are not about laws**

- ▶ It is about what we, as society, think is right
- ▶ Laws are somehow used to enforce ethical behavior

Ethics in a nutshell

Ethics are our shared social values. Those values lead to shared rules that we all agree to follow because of the resulting benefits.

Illustration with the *Tragedy of the Commons*

Economic theory of a situation within a shared-resource system where individual users acting independently according to their own self-interest behave contrary to the common good of all users by depleting or spoiling that resource through their collective action.

Informed consent, Privacy & Anonymity

Informed Consent

Definition

Informed consent is a legal procedure to ensure that a patient, client and research participants are aware of all the potential risks and costs involved in a treatment or procedure.

- **Common to accept multiple pages of *Terms of Service* when you use a new service**
 - ▶ Useful from a legal point of view
 - ▶ All-or-nothing choice, no granularity generally
 - ▶ It is really a choice in the first place?
 - ★ e.g. Is owning a cellphone still a choice today?

Facebook's mood experiment

- In 2014, **Facebook** conducted an experiment on 689,003 users without informing them that it was manipulating their *News Feed*
 - ▶ It was surely legal but was it ethical?

Original paper: *Experimental evidence of massive-scale emotional contagion through social networks*

Significance

We show, via a massive ($N = 689,003$) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. [...]

Data Usage Policies & Restrictions

- **Prospective data collection vs Retrospective data analysis**

- ▶ In the first case you are able to inform that will be the usage of the data when you collect them
- ▶ In the second case you are performing analysis on already collected data
- ▶ Repurposing is often necessary (e.g. medical data)

- **Data usage limitations are necessary**

- ▶ Usage of already collected data is not fundamentally bad
- ▶ Your opinion on how your data can be used can change over time
- ▶ More granularity on how one can use your data should be the norm

Privacy

- **Privacy is a basic need**

- ▶ Even for people who have "nothing to hide"
- ▶ Privacy is **not** complete non-disclosure
- ▶ Privacy is choosing with whom you want to share your data

Definition

Privacy is the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively

- **Modern privacy risks**

- ▶ Easy to collect and store data, potentially forever
- ▶ Easy to aggregate & correlate data
 - ★ Business of *Data Brokers*

Target & the pregnant teenager

- **Target, a US retailer, is performing analytics on customer purchases**
 - ▶ For instance, Target is able to figure out if you have a baby on the way by looking at your purchases
 - ▶ In order to keep the future parents as loyal customers it sends them coupons for baby items
 - ▶ In 2012, an article from Forbes tells the story of a father learning that his teen daughter was pregnant through Target coupons
 - ▶ Can be seen as an intrusion into privacy
 - ▶ Aware of this issue, Target now sends mixed ads
 - ★ The idea is to pretend it knows less than it actually does so as not to create unease

N.B.: If you are interested in this type of analysis, you are encouraged to read Chapters 6 & 9 of *Mining of Massive Datasets*

Right to privacy

- **We are used to privacy by trust**

- ▶ We know with whom we are sharing your data (e.g. a doctor)
- ▶ Exchange of information is symmetric in this case
 - ★ In the sense that you know what one can do with your data

- **We need to have privacy by design**

- ▶ Your data is exchanged between lots of players now
- ▶ Exchange of information is therefore asymmetric
- ▶ We have to ensure that privacy is appropriately respected and managed
 - ★ There is the *Right to be forgotten* in the EU

Anonymity

Definition

Anonymity is the quality or state of being anonymous, i.e. not being named or identified

- **Anonymity seems simple to put in place**
 - ▶ But it can be easy to trace back individuals
 - ★ De-identification has limited value
 - ★ Netflix Prize example
 - ▶ Anonymity is virtually impossible given so many data streams
- **As for privacy we need to have anonymity by design**
 - ▶ License data to trusted parties
 - ▶ Enforce anonymity by contracts and/or professional standards

The Netflix Prize example

The Netflix Prize

It was an open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.

- In 2007, Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas, published a paper called *Robust De-anonymization of Large Sparse Datasets*
- They successfully de-anonymized some of the Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database (IMDb)

The Netflix Prize example (cont'd)

- Their idea was fairly simple: when a Netflix user finishes watching a movie, he is likely to rate it on IMDb
 - ▶ The timestamps and the ratings between the two websites are likely to be close
 - ▶ Thanks to IMDb accounts they were able to find back some of the individuals behind the Netflix ones
- You might wonder: *but what is the issue? People purposely displayed on IMDb the movies they watched right?*
 - ▶ The issue is that they displayed *some* of the movies they watched on Netflix but not *all* of them
 - ▶ We are back to **privacy** issues where users' data are displayed and used without their fully **informed consent**

Data & Models validity

Data validity

- **Data is the most important component**

- ▶ Your model could be the best in the world, but if your data is unreliable then so will the result
- ▶ Bad data/information leads to bad decisions
- ▶ Unfortunately we often don't have access to the data we require

- **Learn more about the field**

- ▶ Data does **not** prevent you to study the field you are working on
- ▶ It enables feature selection

Simple example: if you try to predict the probability that an individual has a beard you will discover that taller people are more keen to have beards. As men are taller than women on average the result is obvious i.e. the size alone is not a good feature in this context.

Data validity (cont'd)

- **Before to start building a model you have to explore the data**

- ▶ Are the data clean?
 - ★ Are the continuous values in the range expected?
 - ★ Are empty fields always marked in the same way?
- ▶ Are the data really representative of what you want to model?
 - ★ If I had to collect data from scratch for a given problem what will I collect?

- **Data errors**

- ▶ Errors can be of form and substance
- ▶ Take for granted that there are always form errors in the data
 - ★ Form errors: both schema errors and improbable values e.g. due to damaged sensor
 - ★ A cleaning pipeline to avoid those errors is necessary
- ▶ Skewed data

Skewed data problem

- **Likely in classification problems**

- ▶ Classes are possibly not evenly distributed
- ▶ Leading models to favor some classes over others
 - ★ In 2015, Google launched Google Photos that, in Google's words, can automatically tag and label your photos
 - ★ Google Photos app identified photos of black people as gorillas

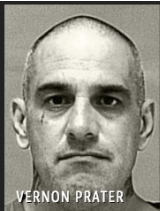
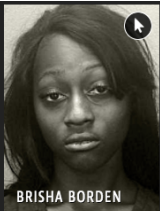
- **Data are likely biased**

- ▶ Data are embedding biases of people who generated and/or collected the data
- ▶ Data are embedding also societal biases
 - ★ Racial biases
 - ★ Gender biases

Racial biases

- **Risk assessment software has become increasingly common in courtrooms in the US**

- ▶ The system is giving a score to criminals which is supposed to represent the likelihood they will commit a future crime
- ▶ Those scores are given to judges during criminal sentencing
- ▶ Those score are heavily biased against black people
 - ★ Skin color is not one of the features
 - ★ Pay attention to features that can serve as proxy

 <p>VERNON PRATER</p>	 <p>BRISHA BORDEN</p>	<p>VERNON PRATER</p> <hr/> <p>Prior Offenses 2 armed robberies, 1 attempted armed robbery</p> <hr/> <p>Subsequent Offenses 1 grand theft</p>	<p>BRISHA BORDEN</p> <hr/> <p>Prior Offenses 4 juvenile misdemeanors</p> <hr/> <p>Subsequent Offenses None</p>
<p>LOW RISK 3</p>	<p>HIGH RISK 8</p>	<p>LOW RISK 3</p>	<p>HIGH RISK 8</p>

Gender biases

- **In Natural Language Processing, one of the most popular systems of word representation is *word2vec***
 - ▶ Trained on hundreds of thousands of articles taken from Google News, *word2vec* has been tested to show that it represents the English vocabulary very well
 - ▶ For example, the difference between the vectors "man" and "king" is equal to the difference between the vectors "woman" and "queen"
 - ★ Noted as "man: king; woman: queen"
 - ▶ The issue is that some of the relations are questionable
 - ★ "father: doctor; mother: nurse"
 - ★ "man: computer programmer; woman: homemaker"
 - ★ "she: he; sewing: carpentry"
 - ★ "she: he; nude: shirtless"

N.B.: See the paper *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*

Some tips

- **Always explore the data**

- ▶ Look for unlikely values and/or classes
- ▶ Establish whether the data follows some distribution
- ▶ Observe how classes are distributed
- ▶ Ask yourself how those data were produced in the first place

- **Set up a cleaning pipeline**

- ▶ Analyze the repartition of the damaged data accros the different classes
- ▶ Look if some features can be used as proxy to infer missing ones
- ▶ Rearrange the data to have evenly distributed classes
- ▶ Normalize (rescale) the data if needed
- ▶ Get rid off the too damaged samples

- **Study the field and survey the common issues related to it**

Model validity

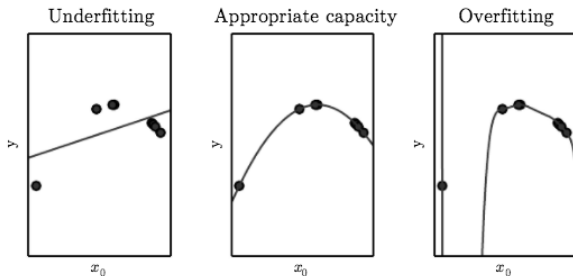
● Model design

- ▶ Even if you have perfect inputs, perfect data, there are many reasons why the model could be incorrect
- ▶ Most machine learning just estimates parameters to fit a predetermined model
- ▶ **No Free Lunch theorem:** there is no one ML algorithm that works best all the time, you have to try many of them

● Importance of the model's capacity

- ▶ A model's capacity is its ability to fit a wide variety of functions
- ▶ Models with low capacity may struggle to fit the training set
- ▶ Models with high capacity can overfit by memorizing properties of the training set that do not serve them well on the test set

Model's capacity



From left to right: a linear function, a quadratic function and a polynomial of degree 9

N.B.: Example extracted from *Deep Learning*, Chapter 5 - Goodfellow

Some tips

- **Keep your model simple**

- ▶ Start from a simple model and use it as baseline
- ▶ Generally, the fewer parameters that you have to tune the better

- **Use cross validation**

- ▶ Use K-Fold cross validation
- ▶ According to your dataset you can use Leave One Out Cross Validation (LOOCV)

- **Use regularization**

- ▶ Add regularizations terms (such as L1, L2, AIC, BIC, MDL or a probabilistic prior) to the objective function

- **Pay attention to the objective function you are using**

- ▶ It is very important as it will be used to optimize the model

Don't conclude too fast

Simpson's paradox

Simpson's paradox, or the Yule-Simpson effect, is a paradox in probability and statistics, in which a trend appears in different groups of data but disappears or reverses when these groups are combined

- **UC Berkeley gender bias example**

- ▶ Study of gender bias among graduate school admissions to University of California, Berkeley
- ▶ Here is the admission figures for the fall of 1973:

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

UC Berkeley gender bias example (cont'd)

It seems biased in favor of men but when examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas only four were significantly biased against women. Here is the the data from the six largest departments:

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

The research paper by Bickel et al concluded that women tended to apply to competitive departments (low rates of admission) whereas men tended to apply to less-competitive departments

Managing change

Campbell's law

It is an adage developed by Donald T. Campbell: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

- In other words as stated by the Goodhart's law: "When a measure becomes a target, it ceases to be a good measure."
- **Things change over time**
 - ▶ Your model once put in place can modified the environment from which the initial data came from
 - ▶ People behavior can change over time and therefore your model can become less relevant
 - ▶ You have to monitor the model's efficiency over time

Societal Impact

Algorithmic fairness

- **We saw that data can be biased**

- ▶ If they are not representative of the population
- ▶ If the past population is not representative of the future population
- ▶ If confounding processes that lead to correlations are just flukes

- **Those data can lead to an algorithmic vicious cycle**

- ▶ A bias raises in the algorithm because of the data it is trained on
- ▶ Once put in place the algorithm perpetuates the bias
- ▶ Let's say that a company with only 10% women employees develops a hiring algorithm

The company has a "boys' club culture" that makes it difficult for women to succeed. The training algorithm is trained on current data and based on current employee success, it scores the women candidates lower. The algorithm, albeit fairly representing that happens today, perpetuates a gender bias.

Correct but misleading results

● Tell a story

- ▶ In your role as a Data Scientist, once you conducted an analysis you are expected to present your results
- ▶ Visualization is a powerful tool but it can be misleading
- ▶ You often have to aggregate the results to present them in a more concise way
- ▶ Your inner biases can lead to highlight specific parts of the results, leading to a specific interpretation of the results

● Diversity suppression

- ▶ Happens when the criteria that an algorithm is using have been tuned to fit the majority
- ▶ The algorithm ends up discriminating
- ▶ Don't rely on accuracy alone: see F-score, type I and type II errors and the tradeoff between precision and recall.

Unexpected discrimination

- **In 2012, the city of Boston put in place a project called "Street Bump"**
 - ▶ The idea is to use the accelerometer and the GPS of citizens' smartphones to detect potholes around the city
 - ▶ It is a very good way to use technology to improve citizens' life
 - ▶ Unfortunately, the issue is that it is only working in the areas where people are rich enough to afford a car and a smartphone
 - ▶ The poorer neighborhoods might be underserved
 - ▶ To avoid this, city employees in city vehicles worked to compensate by driving in poorer neighborhoods
- **Always step back and think about that could be the consequences of the technology you are putting in place**
 - ▶ This notion of evaluating the impact is far removed from the every day concerns of data scientists

Echo chambers / Filter bubbles



ted.com/talks/eli_pariser_beware_online_filter_bubbles

Conclusion

Conclusion

- **We have to accept that algorithms can have biases**

- ▶ Due to the data they are based on
- ▶ Due to people designing the algorithm
- ▶ Due to the interpretation of the results

- **We have to accept that we have inner biases**

- ▶ It is only by being aware of it that we can mitigate it
- ▶ If you have still a doubt about it I invite you to read *The Hidden Brain* from Shankar Vedantam

- **A simple Code of Ethics**

- ▶ Inform people exactly how you intend to use their data
- ▶ Own the outcomes
 - ★ Is it a valid data analysis?
 - ★ Is it a fair data analysis?
 - ★ What are the societal consequences?

References

References

- MichiganX: DS101x Data Science Ethics
- Experimental evidence of massive-scale emotional contagion through social networks
- How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did
- Robust De-anonymization of Large Sparse Datasets
- Machine bias: Risk assessments in criminal sentencing
- How Vector Space Mathematics Reveals the Hidden Sexism in Language
- Sex Bias in Graduate Admissions: Data from Berkeley
- Potholes and Big Data: Crowdsourcing Our Way to Better Government
- Eli Pariser: Beware online "filter bubbles"