The Rise of Data Science, It's Impact on Privacy & Ethics

Ramakrishna Danda
Bellevue University
rdanda@my365.bellevue.edu

Courtney Klatt
Bellevue University
cklatt@my365.bellevue.edu

Abstract

Evolution of Data Science methods and its practical implementation in real world has played a major role in digital transformation, opportunities for achieving economic and social goals within a short span, compared to prior approaches and technologies. Thanks to new age compute and storage, return of investments on Data Science has multiplied many folds for enterprises & organizations and are increasingly paying high attention to digital footprints their customers are leaving behind intentionally and/or unintentionally. There, lie the biggest danger, where enterprises and organizations choose on how lenient they will be in preserving privacy of their customers, and ethical values that are pushed to brink in achieving their goals, by manipulating people's behavior for economic gains by organizations in control of such data applying algorithmic decision making process, ignoring human intricacies and biased analysis. Organizations have no clear understanding of the ethical use of data science in terms of standards and procedures for sourcing, analyzing, sharing, training leaderships, means and measures to control unethical behavior. Governments and societal bodies are responsible for regulating discrimination, making customers aware of their private data use, breaking free from individual monitoring and surveillance and giving back ownership rights to individuals.

Author Keywords
Data Science; Privacy; Ethics

ACM Classification Keywords
Data Science; Privacy; Ethics

Introduction

Data science and technology have been progressing rapidly, and along with these advancements; there have been new ways to integrate those learnings that are in finance, supply chain, policing, health specific fields, such as biology.

In Health care. Data science is extremely helpful and important in genetic research and opens new avenues for exploring and understanding every aspect of biology. Data science decreases the time it takes to make new discoveries that can save lives. Through data mining and machine learning, human evolution can be mimicked and the best solutions to genetic problems can be found (1).

Data science has many applications in biological research and is becoming an essential tool for geneticists. It is commonly used for analyzing genomes, which includes looking at gene expression, gene regulation, and specific sequences and their functions (3). Without data science, studies of these processes would take years to accomplish, but now they can be completed in just a few days.

However, with the rise of data scientists working with genetics comes a heightened concern about individual privacy. Genetic information includes family medical history, an individual's risk of getting a disease, information about diseases a person already has, and much more (2). While it is not legal to discriminate against or harass anyone because of their genetic information, this information can be widely shared and can be traced back to the individual.

Since data science has become integrated with genetic research, many privacy concerns have been raised. Genetics reveal a lot of information about an individual, which can leave them susceptible to losing their privacy. It is not just the genetic information on its own that makes an individual vulnerable, but also the combination of that genetic information and other available data, which can lead directly to a person's identity and other personal information (4). An individual's identity can be traced in many ways, including genealogy, searching within the meta-data, and using phenotypes (5). Despite concerns about privacy, scientists have a desire, and oftentimes a need, to share their data, in order to share their discoveries with the world. There are many different options for protecting this privacy. Everything from a universal waiver, encouraging researchers to keep their data private, and limiting which pieces of genetic information from an individual are shared together (4).

In current societies, we have reached a point in our understanding of data and acceptance of them in our daily lives. It is at the same time societies are expecting to safeguard, their data, that is individual data given to companies that provide services both voluntary and involuntary. Since, data is pushed through many Data Science methods, there is no stopping of collection. Individuals are asking and posing more appropriate questions on protecting the privacy of their data.

Complying with the rules

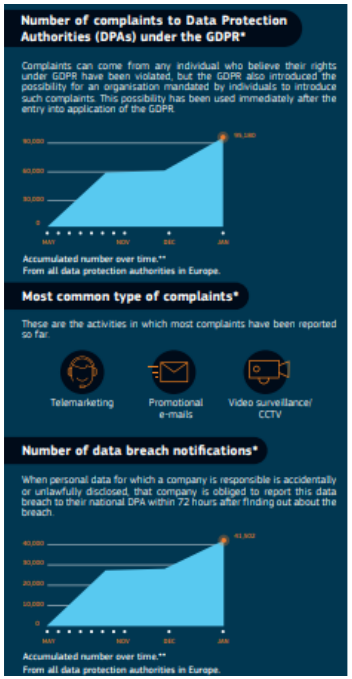


Figure above: EU commission published infographic on compliance with and enforcement of DGPR since May 2018 to Jan 2019.

Measures to protect privacy

It is social and legal responsibility to protect data. Currently there are laws protecting individual genetic information. Individuals cannot be harassed, discriminated against, fired, demoted, or retaliated against based on their genetic information (2). Furthermore, it is unlawful to gain or share genetic information about an individual, other than a few exceptions (2).

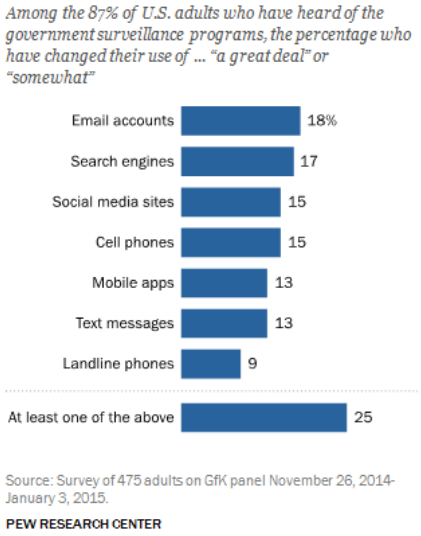
On general legal terms, there are laws protecting through the jurisdiction they are hosted in. Like

- US-UE safe Harbor
- General Data Protection Regulation (GDPR)
- Russian Federal Law on Personal data
- German Bundesdataenschutzgesetz

GDPR has been in effect since 2018 and has played a major role in defining the safeguard rules of data for European union and applies to non-European Union organizations that process data related to EU residents. This regulation has fine up to 20mil EUR or up to 4% of annual worldwide revenue.

Figure above shows fines assed on companies by GDPR for over a period time since its inception.

In United states, the people were divided on leaks of organizations and believe government should prosecute the leakers. And at the same time disapproving government surveillance programs on citizens. On the other note, government disclosure of surveillance prompted citizens to change the way they use technologies. It is also interesting for US citizens to do surveillance on other country citizens and not on themselves.



Of the most important part of the conversation is about individual citizens getting control of who get their information has played a major role on which laws to pass and not. And expressed confident in federal laws and government abilities to protect their data.

Role of Data Science teams

As a data science team, making calls on securing data, preserving privacy with ethical consideration is a tough call. This usually means navigating complex technical and political issues weather or not to use raw data and to what extent without possibility vulnerability to fraud attacks and hacks. This essential means, spending quite a bit of time on privacy and ethical values and still contribute to data science process.

Those times could be best spent in below order-

1. Not to use any sensitive data to begin with.
2. If needed to use. What would you do if you don't have it?
3. If you must collect it as answer 'yes' to 2, then ask is 'nice' to have or 'must' have?
4. Ask the ethical question for yourself if you are getting to bias answering 3.
5. Create models as base with least amount of sensitive and unethical ways of data.
6. Still models demand private data, then can we anonymous the PII data for e.g.: k-anonymity, homomorphic pseudonymization or synthetic data without retaining true values.
7. For a must use case of sensitive data, despite all the above steps, then ask questions.
    a. Will these be shared among other teams if so how to proceed.
    b. For ML building, look for privacy preserving models or training on anonymized data.
    c. Use aggregated results to make decision making
    d. And finally making sure to lockdown the model access to public use and stored in a secured place.

Privacy and Ethics should be in built in the model by *design*. And when in design mode keeping the following guidelines will help alleviate the pain of sources change dynamically.

- Regulations should be built into the data source that can be changed dynamically.
- Gate keepers controlling the policies and control data.
- Access management tools to regulate the flow of data. Data, essentially asking the owners of data to let them control the flow of data and to which purpose.
- Applicants are required to file periodic reports about the data usage and any adverse events.
- Trust-but-verify approach, where parties cannot download data without imposing restriction by appropriate privileges which are vetted out in advance.
- Should not execute certain type of queries that limit access to personal information and such queries are audited with system traceable to people that queried such records.

Health care has already done and or improving wide array of such reforms and responsibilities throughout the lifetime of the data by a method called PEER (Platform for Engaging Everyone Responsibly)

Conclusion:

As new technologies and techniques flow in the field or data science it is essentially the duties of all parties in the process to take the ownership and treat the data, as if their own data is at use. No amount of oversight will alleviate today privacy and ethical concerns, but a conscious effort with above listed process will pave steppingstone to Data Science excellence, trust and confidence to users and practitioners alike.

References

1. Genetic Algorithms and their Applications in Data Science. Retrieved August 11, 2019 from https://www.it4nextgen.com/genetic-algorithm/
2. Genetic Information Discrimination. Retrieved August 11, 2019 from https://www.eeoc.gov/laws/types/genetic.cfm
3. Computational Genomics and Data Science Program, Retrieved August 11, 2019 from https://www.genome.gov/Funded-Programs-Projects/Computational-Genomics-and-Data-Science-Program
4. Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection. Published: October 2, 2009. Retrieved August 11, 2019 from https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000665
5. Navigating the Data Privacy Maze: Tips for Data Scientists. Original 2 Aug 2018, Retrieved on 12 Aug 2019 https://towardsdatascience.com/navigating-the-data-privacy-maze-tips-for-data-scientists-c2f784969f29
6. Routes for breaching and protecting genetic privacy. Original published on 2014 May 8. Retrieved on 12 Aug 2019 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151119/
7. How can we mitigate ethical and privacy issues in data science? Original 2017 Oct 26, retrieved on 12 Aug 2019 https://www.siliconrepublic.com/enterprise/ethics-data-science-bias
8. Ethical Implications of Big Data Analytics. Original June 2016, Retrieved on 12 Aug 2019 https://www.researchgate.net/publication/308024119_ETHICAL_IMPLICATIONS_OF_BIG_DATA_AN
9. Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection, Original 2 Oct 2009, Retrieved on 10 Aug 2019 https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000665
10. Data Privacy & Data Science: The Next Generation of Data Experimentation. Original 12 April 2016. Retrieved on 10 Jul 2019 https://www.immuta.com/data-privacy-data-science-the-next-generation-of-data-experimentation/