5Th Unit Data Science

Ethical Issue

I. 5 ethical questions in data science

5 ethical questions in data science, amid the growing concern of its ethical use by organisations.

Ethical Questions in Data Science

With the rapid growth in data science, there has been a growing concern around its ethical use by organizations. For example, concerns have arisen as:

- Data science algorithms are used to accept and deny bank loans and the insurance premiums payable for insurance. However, the question arises: What is the social cost of a wrong decision for a bank loan or insurance?
- Companies use data science to scan resumes and recommend the best candidate for a role. However, the question arises: What is the chance for a bias towards gender or age in the hiring algorithm if that algorithm is based on past data?
- Companies use cookies to monitor the online behaviour of individuals and advertise based on their browsing behaviour. However, the question arises: What if an individual views companies reading their behaviour as an intrusion of their privacy?
- Airlines use data science to decide on differential pricing for individuals based on their needs and rideshare companies (e.g., Uber) engage in surge pricing based on demands. However, the question arises: Is there a risk of these companies exploiting individuals beyond their means when they are in desperate need of their services?

As data science algorithms assist and replace human decision making, there are questions that every organisation should keep in mind. Some of the leading ethical concerns of harms by misuse of data include:

1. Unfair discrimination

The incorrect and unchecked use of data science can lead to unfair discrimination against individuals based on their gender, demographics and socio-economic conditions.

If you have really large data sets, you might not even realize that the data are slightly biased towards gender or whatever you're analyzing …. It might be that you've overstrained on those characteristics.'

2. Reinforcing human biases

Gartner ('Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence', 2020) predicts that by 2022, 85 percent of data science projects will deliver erroneous outcomes due to bias in data, algorithms or the teams responsible for managing them.

Data science algorithms use past data to predict future outcomes. Data are generated based on human decisions made in the past. Training the algorithm purely based on past data could lead to some of these biases being included in the algorithms.

Algorithms are also influenced by analysts' biases, as they may choose data and hypotheses that seem important to them.

3. Lack of transparency

Data science algorithms can sometimes be a black box where the model predicts an outcome but does not explain the rationale behind the result.

Numerous recent machine learning algorithms fall into this category. With black box solutions, it is not easy for a business to understand and explain the reason for a business decision.

As Andrews notes, 'Whether an AI system produces the right answer is not the only concern… Executives need to understand why it is effective and offer insights into its reasoning when it's not.'

4. Privacy

Data privacy has become a major focus in the past few years. Sensitive data are stored by various organisations and are subject to hacking and misuse.

During the 2016 United States presidential election, Cambridge Analytica, a data analytics firm that worked on Donald Trump's election campaign, used Facebook data to influence customers' behaviours in the US election.

Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in a major data breach. This incident highlighted ethical concerns related to the misuse of data.

There has been an increase in data breaches across the world. Rules and regulations, such as the General Data Protection Regulation (GDPR), have been introduced to monitor the way companies store and use sensitive data.

5. Consent and Power

Organisations are not transparent as to what data they collect, and use it to make decisions. Most web browsers and websites capture enormous amounts of user data even without their knowledge and consent.

For example, Google (Chrome and Gmail) and Facebook store individual browsing data and monetises it by selling insights from users' data for advertising.

The human side of analytics is the biggest challenge to implementing big data

II.A look back at data science

The term "Data Science" was created in the early 1960s to describe a new profession that would support the understanding and interpretation of the large amounts of data which was being amassed at the time.

(At the time, there was no way of predicting the truly massive amounts of data over the next fifty years.)

While Data Science is used in areas such as astronomy and medicine, it is also used in business to help make smarter decisions.

Data Science started with statistics and has evolved to include concepts/practices such as artificial intelligence, machine learning, and the Internet of Things, to name a few.

As more and more data has become available, first by way of recorded shopping behaviors and trends, businesses have been collecting and storing it in ever greater amounts. With the growth of the Internet, the Internet of Things, and the exponential growth of data volumes available to enterprises, there has been a flood of new information or big data.

A functional data scientist, as opposed to a general statistician, has a good understanding of software architecture and understands multiple programming languages.

They use the principles of Data Science, and all the related sub-fields and practices encompassed within Data Science, to gain deeper insight into the data assets under review.

From the 1960s to the Present

In 1962, John Tukey wrote a paper titled *The Future of Data Analysis* and described a shift in the world of statistics, saying, "… as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt…I have come to feel that my central interest is in data analysis…"

Tukey is referring to the merging of statistics and computers, when computers were first being used to solve mathematical problems and work with statistics, rather than doing the work by hand.

In 1974, Peter Naur authored the *Concise Survey of Computer Methods*, using the term "Data Science," repeatedly. Naur presented his own convoluted definition of the new concept:

"The usefulness of data and data processes derives from their application in building and handling models of reality."

In 2015, Bloomberg's Jack Clark, wrote that it had been a landmark year for artificial intelligence (AI). Within Google, the total of software projects using AI increased from "sporadic usage" to more than 2,700 projects over the year.

Data Science Today

In the past 30 years, Data Science has quietly grown to include businesses and organizations worldwide. It is now being used by governments, geneticists, engineers, and even astronomers. During its evolution, Data Science's use of big data was not simply a "scaling up" of the data, but included shifting to new systems for processing data and the ways data gets studied and analyzed.

Data Science has become an important part of business and academic research. Technically, this includes machine translation, robotics, speech recognition, the digital economy, and search engines. In terms of research areas, Data Science has expanded to include the biological sciences, health care, medical informatics, the humanities, and social sciences. Data Science now influences economics, governments, and business and finance.

III.. Next Generation of Data Scientists Needs to Develop

As reliance on data and analytics continues to expand across industries from agriculture to manufacturing, health care to financial services, it stands to reason that the next generation of data leaders will have far-reaching roles that impact strategy, decision-making, operations, and countless other functions.

To help prepare this new talent, I have developed a framework composed of four key areas of skills and capabilities that will help current and future data scientists hone their abilities to add maximum value to a business. This is done by ensuring that data science work is seen as important and indispensable by their business-function counterparts.

with greater understanding of what each area of business entails, today's data scientists and those entering this field can see how their knowledge and experiences stack up — and where they need more development.

1. Problem Spotting: Seeing the real issue

As they delve into analytics across the business, data leaders have a front row seat to nearly every operation and function. This provides them with a unique vantage point for both solving problems and identifying new ones. Management also found that people who rated check-in poorly had a lower rate of returning to the hotel.

Then an employee suggested they look at customer surveys that had been collected on a rolling basis. Some natural language text analytics teased out some themes — namely.

The Takeaway: Solving the problem that is in front of you can mean missing out on opportunities to help the business improve in other ways. Those who work with data often have access to deep, unique insights into numerous aspects of the business. To become adept at problem-spotting, data leaders need to embrace that big-picture view and gain deeper insights, with greater transparency around what matters most to business leaders. In this way, data leaders can add value by identifying problems that otherwise escape notice.

2. Problem Scoping: Gaining clarity and specificity

Once a problem has been spotted, the next step is determining its scope — that is, gaining clarity into the nature of the problem and how analytics can help solve it.

This is especially important if a business leader has approached the data team with a vague concern or challenge.

In my classes and my workshops, we practice scoping with an exercise. I assume the role of a product or strategy or marketing leader with a well-defined problem in my head. For instance, perhaps I manage customers, and want to be able to identify which customers are at risk of giving low **net promoter score** (NPS) ratings so that we can intervene and improve their experience..

It could be a pipeline issue, but we just don't have alignment. I think we're playing in the right sandboxes, now we just need to know the who and the why. Sound good?"

- What, precisely, is the problem we're trying to solve?
- What outcomes, if improved, would indicate that the problem has actually been solved?
- What data would ideally be available to solve the problem, and what data are actually available?
- How will the analysis lead to a solution?

The Takeaway: To excel at problem-scoping, data leaders need good communication skills to talk through the problem with the business leader to arrive at the requisite specificity that will enable data analytics tools and concepts to meaningfully contribute to the business. Only then can the problem be turned over to the data team for analysis.

3. Problem Shepherding: Getting updates, gathering feedback

Once the problem is identified and scoped out, many data analysts go into isolation and only emerge when they have found a solution. This approach is highly problematic. To be most effective, the process requires a great deal of information

This approach runs counter to how some data scientists prefer to work. Sometimes they get enamored with their models and their creative problem-solving techniques, and they can't wait for the big reveal. Surprising results often prompt people to start questioning the underlying data and methods.

However, by bringing the business team into decision-making along the way, they will buy into the results and commit their trust.

The Takeaway: Problem-shepherding sets up a process of providing regular updates and gathering feedback from the business team. Data scientists and team

leaders who are strong in this area are able to encourage and facilitate candid discussions that ensure the final deliverable hits the mark with the business team — with no surprises.

4. Solution Translating: Speaking in the language of the audience

At this point, we transition from problem to solution, the success of which depends on how well data leaders and their teams have executed on the first three steps. More than determining a final answer, the data team must also deliver a solution that's understandable and, therefore, actionable.

This isn't just about putting the data in a chart or another visual display. must be conveyed in language the business team can understand. One tool I've recommended is the two-page data analytics memo, which highlights the most important elements of the problem to be solved.

The two-page limit can avoid the temptation to go on and on about details of the data analysis and encourage focus on the recommendations being made and the evidence for them

The Takeaway: Solution translation requires data leaders to step back and consider how to make the most impact with their analyses and recommendations. By using simple language, while not compromising the complexity, data leaders who excel in this area can deliver the equivalent of an elevator speech to engage business leaders with compelling and understandable solutions.

Teaching Notes

For

DATA SCIENCE ESSENTIALS

R.Prabhakar Naidu (Ph.D)

Principal – MCA

"Privacy and Ethics in Data Science"

Privacy Privacy Reproducibility
Fair Information PracticesManaging sensitive dataAnonymizing sensitive dataRe-identifying datasetsReproducibilitySocietal value of data and data science

Privacy

The Rise of Privacy Concerns
Science:benefits of sharing clinical patient recordspatients shall control access to their recordspatients found to be altruistic:willing to grant access for purpose of advancing scienceGovernment:government and commercial use of data mining raises concerns about appropriate use of private citizen information,e.g., data collected for the purpose of airline passenger screening should not be used for the enforcement of other criminal lawsOpen Web:many users are happy to share private details on social websbut would be rightfully upset was this data used for other purposescontent is shared between networksnot very transparent to the userusers need to be reassured about appropriate use of their data

Private and Sensitive Data

Sensitive Data and Privacy
Data about individuals and organizations that should not be freely disseminated and publicizedHealthEducationFinanceDemographicCriminalLocationBehaviorDesire to limit the dissemination of sensitive dataLots of technology, but:Unclear requirementsUnclear behaviors

Sensitive Dataidentifying valuessensitive attribute

OECD's Eight Principles of Fair Information Practices [OECD 1980]
A framework for privacy protectionProtect useCollection for a purposeUse only for authorized purposeAccountability throughout these principlesYolanda Gil

Define questionsCollect/find dataStore dataExtract dataPre-process dataAnalyze dataPresent resultsPublish data

Define questions Collect/find data Publish data Present results
Store dataExtract dataPre-process dataAnalyze dataPresent resultsPublish dataInstitutional Review BoardProvisions for collection, storage, processing, and dissemination of sensitive data

Define questions Collect/find data Publish data Present results
Store dataExtract dataPre-process dataAnalyze dataPresent resultsPublish dataConsentState purpose/useDecent qualityAllow corrections

Define questions Collect/find data Publish data Present results
Store dataExtract dataPre-process dataAnalyze dataPresent resultsPublish dataPhysical safetyPersonnel trainingAccess controlEncryption

Define questions Collect/find data Publish data Present results
Store dataExtract dataPre-process dataAnalyze dataPresent resultsPublish dataLimit data use based on the purpose expressed in the original consentSecure data transmissionAnonymization

Anonymization Techniques
Replace identifiers with randomly-generated identifiersEg: “Jane Krakowski” -> “Patient6479”Abstraction: Replace values by rangesEg: Check-in date: 3/1/16 -> Check-in date: Spring 2016Eg: Replace zip code by stateCluster data points and replace individuals by their cluster centroidEg Ages: 21, 25, 28, 27, 18 -> 5 individuals with nominal age of 24Remove valuesEg: Omit birth date

Problems with Anonymization Techniques
Limited use for researchToo coarse-grainedRe-identificationRe-identification is often trivialE.g., anomymized list of students admitted showing undergraduate university and average GPARe-identification is possible with high certainty in many casesBy linking the anonymized dataset with other public data that is not anonymized

Examples of Re-Identification through Linking Data: (I) Medical Records
87% of the population can be uniquely identified based solely on birthdate, sex, and zip codeMost datasets even if anonymized contain this informationWilliam Weld was governor of Massachusetts at that timeand his medical records were in the GIC data. Governor Weld lived inCambridge Massachusetts. According to the Cambridge Voter list, six peoplehad his particular birth date; only three of them were men; and, he was theonly one in his 5-digit ZIP code

Examples of Re-Identification through Linking Data: (II) Opinions
Published anonymized data about reviewsPublic dataset contained reviews that were not anonymized and could be mapped based on the dateWilliam Weld was governor of Massachusetts at that timeand his medical records were in the GIC data. Governor Weld lived inCambridge Massachusetts. According to the Cambridge Voter list, six peoplehad his particular birth date; only three of them were men; and, he was theonly one in his 5-digit ZIP codeReview dateAnonymized Netflix dataNamed IMDB data

Examples of Re-Identification through Linking Data: (III) Behavior Patterns
Four spatiotemporal points are enough to uniquely re-identify 90% of individualsEven data sets that provide coarse information for all dimensions provide little anonymity

Addressing the Problems of Simple Anonymization Techniques
Provide guarantees that re-identification will not be possible within some boundsEg: can only map a given individual to a set of 50 individualsk-anonymizationl-diversityt-closenessDifferential privacy

Addressing Anomymization Problems: k-Anonymity
A dataset has k-anonymity if at least k individuals share the same identifying valuesk=2

Addressing Anomymization Problems: l-Diversity
A dataset has l-diversity if the individuals that share the same identifying values have at least l distinct values for the sensitive attributel=1

☐ **Addressing Anomymization Problems: t-Closeness**
A dataset has t-closeness if the individuals that share the same identifying values have values for the sensitive attribute that are within a threshold t of diversityThreshold is mathematically defined for the data

☐ Differential PrivacyOnly method that provides mathematical guarantees of anonymityMain problem addressed: Taking an individual I off a dataset reveals their sensitive attribute informationEg: retrieving aggregate data before removal, then retrieving aggregate data after removal, and then comparing the difference will give us the sensitive attribute of IMain idea: Differential privacy adds “noise” to the retrieval process so that such comparisons do not give us the actual sensitive attribute information“noise” is mathematically defined for the data

☐ **Privacy-Aware Workflows**
P1: No personal ID information canleave the data sourceP2: Sensitive data must be k-anonymizedDistributed workflow compliant with policiesAnonymizationAbstractionAnalysisLoc1 .. Loc nLoc3Loc2Centralized workflow not compliant with policiesAggregationAnalysisLoc1Loc2Loc3AggregationYolanda Gil

☐ **Summary: Threats to Privacy**
Privacy requirements are not well articulatedPeople want benefits in exchange for dataUnclear that we are able to limit collection and publicationUnique behavior of people (we don’t read legal contracts)Human error, not without consequencesMounds of sensitive data about individuals is readily available in the open webOpen web already contains sensitive information that should not be available and violates privacy actsLots of commercial data with personal information is for saleLimited understanding of anonymization and other privacy technologiesLinking to public datasets leads to re-identify individuals

☐ Reproducibility

☐ **Granting Access to Private Records: Health Information**
Anonymized information is often not useful for researchToo coarse grainedPrivate information has great valueTradeoff with quality of treatmentIncentivized through first access to new treatmentsAltruismGiving up privacy for pre-specified usesEg: for specific medical study, not for insurance purposes, not for employers, not for social studies

☐ There is zero privacy anyway, get over it Although you can upload your data using a pseudonym, there is no way to anonymously submit data. Statistically speaking it is really unlikely that your medical and genetic information matches that of someone else. By uploading you do not only disclose information about yourself, but also about your next kinship (parents and siblings), that shares half of a genome with you. Before uploading any genetic data you should make sure that those people approve of you doing so. This is especially important if you have monozygotic twin, who shares all of your genome!

☐ **Privacy Privacy Reproducibility**
Fair Information PracticesManaging sensitive dataAnonymizing sensitive dataRe-identifying datasetsReproducibilitySocietal value of data and data science