UNIT-V

Data Science & Ethical Issues - Privacy, Security & Ethics:

Security & Privacy Concerns r growing as data becomes
more & more accessible the collection & aggregation of massive
quantities of heterogeneous data r now possible. Large scale
data sharing is becoming routine among scientists, clinicians,
businesses, govt agencies & citizens. However, the tools & technolo
-gies that r being developed to manage these massive data
sets r often not designed to integrate sufficient security &
privacy measures, in part bcoz we lack sufficient training &
a fundamental understanding of how to provide large scale
data security & privacy.

We also lack adequate policies to ensure compliance with
current approaches to security & privacy. Futhermore, existing
technological approaches to security & privacy r increasing being
breached, thus demanding the frequent checking & demanding
of current approaches to prevent data leakage. While the
aggregation of such data presents a security concern in itself,
another concern is that these rich databases r being shared
with other entities, both private & public.

In many cases, potentially sensitive data r in the hands of
private Companies. For some of these companies such as
Google, Facebook & Instagram, the data abt their users is
of their main assests, & already a part of the product that
they sell. But even if this is not the company's main
business, the ability to protect the privacy & data of its
customers & users will represent a major risk. Potential
private issues, data protection, & privacy related risks should
therefore be addressed early in any big data project.

Data Scientists use big data to find out abt our shopping
preferences, health status, sleep cycles, moving patterns,
online consumption & friendships. In some cases, such inf
is individualized. Removing elements that allow data to be
connected to one individual is, however, just one feature
of anonymization. Therefore, regardless of the data being
anonymous in the sense of being not individualized,
groups r often becoming more transparent.

In order to protect the user privacy, best practices in the prevention & detection of abuse by continuous monitoring must be implemented. privacy preserving analytics is open area of research that can help minimize the success malicious actress from the dataset. However there are few practical soln's at the moment.

Differential privacy is a good first step towards privacy preservation. Differential privacy defines a formal model of privacy that can be implemented & proven secure at the cost of adding computational overhead & noisy results to data analytics results. perhaps the current definition of differential privacy is too conservative, & a new, more practical definition might address some of costs associated with the implementation of this principle.

Despite privacy challenges, the utilization of big data could also have huge benefits for society at large. By using demographic & mobility data, we can obtain key insights into human behaviour, including traffic patterns, crime trends, crisis responses, & social unrest. These, in turn, can be used by business & policy makers to create better, safer & more efficient societies.

2) A Look back at Data Science:

Data Science could be defined simply as what datascientists do, as we did earlier when we walked abt profiles of data scientists. In fact, before Rachel taught the data science course at Columbia, she wrote up a list of all the things data scientists do & didn't want to show it to anyone bcoz it was overwhelming & disorganized. That list become the raw material of the profiles. So, the list ~~list~~ is

- Exploratory data Analysis → Visualization
- Dash-boards & metrics → find business insights.
- Data driven decisions making
- Data Engineering / Big data.
- Get the data themselves.
- Build data pipelines
- Build products instead of describing existing product usage.

→ Hack → patent writing → Detective work.
→ predict future behavior / Performance
→ write up findings in reports, presentations & journals.
→ programming (R, C, Java etc)
→ Conditional prob → optimization
→ Alg's, statistical models & machine learning.
→ Tell & interpret stories → Ask good Questions.
→ ~~Ask good questions~~ investigation → Research
→ Make inferences from data → Build data products.
→ find ways to do data processing, munging & analysis at scale.
→ Interact with domain experts
→ Design & analyze experiments
→ find correlation in data & try to establish causality.

Let's define data science beyond a set of best practices used in tech companies. Now consider data science to be beyond tech companies to include all other domains: neuroscience, health analytics, eDiscovery, Computational social sciences, digital humanities, genomics, policy to encompass the space of all problems that could possibly be solved with data using a set of best practices

Data science happens both in industry & in academia, i.e., where & what domain data science happens in is not the issue - rather, defining it as a "problem space" with a corresponding "sol^n space" in alg's & code & data is the key.

Data science is a set of best practices used in tech companies, working within a broad space of problems that could be solved with data, possibly even at times deserving the name science. Even so, it's sometimes nothing more than pure hype, which we need to hyp guard against & avoid adding to.

3. Next-Generation Data Scientists:

Ideally the generation of data scientists in training r seeking to do more than become technically proficient & land a comfy salary in a nice city - although those things would be nice. we'd like to encourage the next-gen data scientists to become problem solvers & question askers, to think deeply abt appropriate design & process, & to use data responsibly & make the world better, not worse

i) Being Problem Solvers:

First, let's discuss the technical skills. Next-gen data scientists should strive to have a variety of hard skills including Coding, Statistics, machine learning, visualization, comm, & math. Also, a solid foundation in writing code, & Coding Practices such as Pair Programming, Code reviews, debugging, & version control r incredibly valuable.

Another caution: many people go straight from a dataset to applying a fancy alg. But there's a huge space of important stuff in between. It's easy to run a piece of code that predicts & classifies, & to declare victory when the alg converges. That's not the hard part. The hard part is doing it well & making sure the results are correct & interpretable.

ii) Cultivating soft Skills!

Tons of People can implement K-Nearest Neighbor (KNN), & many do it badly. In fact, almost everyone starts out doing it badly. What matters isn't where u start out, it's where u go from there. It's imp that one cultivates good habits & that one remains open to continuous learning.

Some habits of mind that we believe might help solve problems r persistence, thinking abt thinking, thinking flexibly, striving for accuracy & listening with empathy.

Let's frame this somewhat differently: in education in traditional settings, we focus on answers. But what we probably should focus on, & atleast emphasize more strongly, is how students behave when they don't know the answer. We need to have qualities that help us find the answer.

~~Basic~~ Speaking of this issue, have you ever wondered why People don't say "I don't know" when they don't know something? This is partly explained thru an unconscious bias called the Dunning-Kruger effect.

Basically, people who r bad at something have no idea that they r bad at it & overestimate their confidence. People who r super good at something underestimate their mastery of it. Actual competition may weaken self confidence. Keep this in mind & try not to over & underestimate ur abilities - give yourself reality checks by making sure u can code what u speak & by chatting with other data scientists about approaches

(iii) Being Question Askers!

People tend to overfit their models. It's human nature to want ur baby to be awesome, & u could be working on it for months, so yes, ur feelings can become pretty maternal.

It's also human nature to underestimate the bad news & blame other people for bad news, bcoz from the parent's perspective, nothing one's own baby has done & is capable of is bad, unless someone else somehow made them do it. How do we work against this human tendency?

Ideally, we'd like data scientists to merit the word "scientist", so they act as someone who tests hypotheses & welcomes challenges & alternative theories. If someone thinks they can do better, then let them try, & agree on an evaluation method beforehand. Try to make things objective.

Get used to going thru a standard list of critical steps: Does it have to be this way? How can I measure this? what is the appropriate alg & why? How will I evaluate this? Do I really have the skills to do this? If not, how can I learn them? who can I work with? who can I ask? And possibly the most important: how will it impact the real world?

Second, get used to asking other people questions. When u approach a problem a person posing a question, start with the assumption that u r smart, & don't assume the person you're talking to know more or less than u do. u r not trying to prove anything — u r trying to find out truth. Be curious like a child, not worried abt appearing stupid. Ask for clarification around notation, terminology, & process: where did this data come from? How will it be used? why is this the right data to use? what data r we ignoring, & does it have more features? who is going to do what? How will we work together?

(iv) Being an Ethical Data Scientist:

You all r not just nerds sitting in the corner. u have increasingly important ethical questions to consider while u at work.

we now have tons of data on market & human behavio
data scientists, we bring not just a set of machine lea
tools, but also our humanity, to interpret & find meaning i
data & make ethical, data-driven decisions.

Keep in mind that the data generated by user behavio bec
the building ~~product~~ blocks of data products, which simultane
r used by users & influence user behavior

Much is made about predicting the future, predicting the present & exploring causal relationships from observed data.

The next logical concept then is: models & alg's r not only capable of predicting the future, but also of causing the future. That's what we can look forward to, in the best of cases, & what we should fear in the worst.

V) Career Advice:

Lot's of people ask us whether they should become data scientists, so we're pretty used to it. we often start out the advice session with 2 questions:

1. what r u optimizing for?
→ u want money, bcoz u need some min to live at the standard of living & u might want to lot.
→ Maybe u value time for with loved ones & friends.
→ what r ur goals? what do u want achieve? Are u interested in becoming famous?

2. What constrains (limitation) r u under:

There might be external factors, outside of ur control, like u might need to live on certain areas with ur family. Consider also money & time constraints, whether u need to think abt vacation & maternity leave policies

Don't be painted into a corner, but consider how to promote the +ve aspects of urself: ur education, ur strengths & weaknesses, & the things u can & cannot change abt urself.

on the other hand, remember that whatever u decide to do is not permanent, so don't feel too anxious abt it. u can always do something else later - people change jobs at all time. on the other hand, life is short, so always try ~~not~~ to be moving in the rt direction - optimize for what u care abt & don't get stagnant.

Scanned by TapScanner