"Etc."

What about all those 'peripheral' topics?

Our topics for today

'Data' handling is not just about storage, processing, and utilization! There is a constellation of 'issues' (topics/items) that surround the core, and these items are just as (if not, more) important [why?].

Here is what we'll be looking at:

- privacy
- security
- ethics
- trust
- compliance
- governance

Please care about the above! In the coming years, they will be become even more important, as data pipelines become rather commonplace. What are other examples of this?

PRIVACY

Background: 'Fair Information Practices', 1970s (FEP): The FIP efforts in organizations followed five tenants:

1. Openness.

2. Disclosure.

3. Secondary usage limits.

4. Correctability.

5. Security.

Today, e-commerce companies collect LOTS of info about customers - for immediate sales, and for analytics. Too much data collection makes them vulnerable to theft/leakage etc.

Simply put, our lives ARE NOT 'PRIVATE' ANYMORE!

1. Practically all web sites track us - e-commerce, social media... - and exchange data among themselves and brokers.

2. Your medical data belongs to - your hospital!

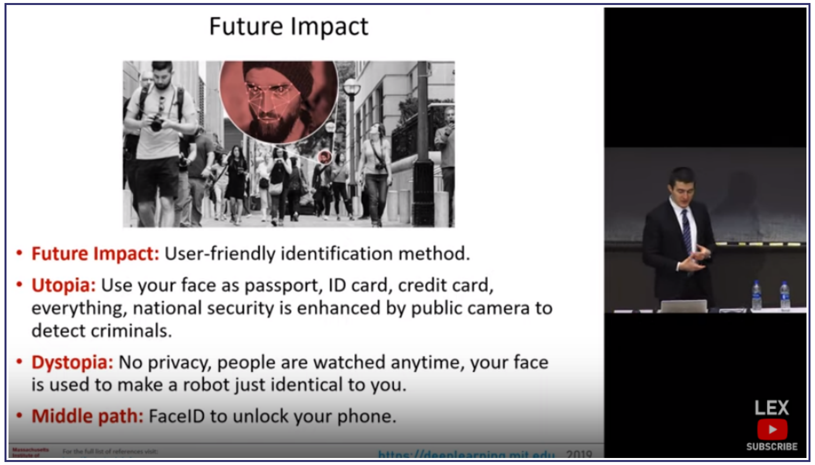
3. You are under surveillance, esp. if you are in China or Japan [for now].

4. Your search data is not private. Through data inference, dots can be connected...

5. 'They' know where you've been! Eg. malls, public security cameras, your own phone - all know where you are...

The notion of "privacy" needs to be redefined?!

Our digital conveniences have a flip side:



Future Impact

- **Future Impact:** User-friendly identification method.
- **Utopia:** Use your face as passport, ID card, credit card, everything, national security is enhanced by public camera to detect criminals.
- **Dystopia:** No privacy, people are watched anytime, your face is used to make a robot just identical to you.
- **Middle path:** FaceID to unlock your phone.

https://deeplearning.mit.edu 2019

SECURITY

Unfortunately, data breaches are commonplace, and involve theft/exposure of value, identity...

Novel fronts: attacks on IoT [devices, data], including cars.

ETHICS

'Ethical' use of data involves multiple aspects!

First, there's the potential for 'rogue' AI. Eg. autonomous drones, and robot soldiers can act with bias, or equally badly, without bias and without morality.

Then there is the issue of fairness. This plays out for ex, in face recognition. IBM is contributing a dataset to help mitigate this.

A gov't can unfairly target protected groups...

Data, via ML, can be used to generate fake news, eg. fakevideo. This is a specific form of 'disinformation'. Disinformation was defined in Great Soviet Encyclopedia (1952) as "false information with the intention to deceive public opinion". Question: what, then, is 'misinformation'?

Here is one way to reduce bias, at the algorithm level.

Another angle to approach fairness is to improve the interpretability of ML.

You can learn more, here.

TRUST

How (much) can we (individuals) TRUST organizations (business, government, non-profit...) to RESPONSIBLY use **our** data?

Trust is proportional to transparency, value delivery, consequence acceptance. You can learn more, here.

COMPLIANCE

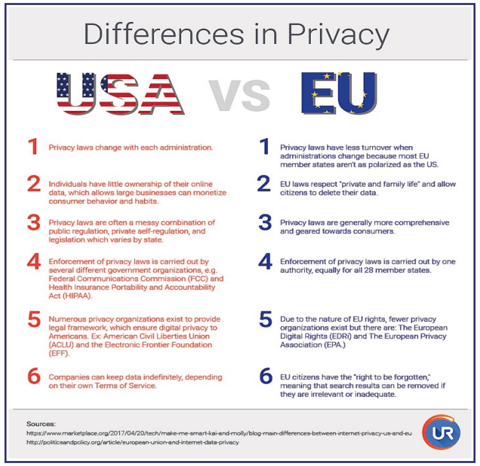
Compliance is a LEGAL/REGULATORY issue - what laws be passed, to help citizens have/gain control over their data? Aside: what is the difference between a law, regulation, and policy?

In the EU, there's GDPR [General Data Protection Regulation].

Effective 5/25/18, the EU has a SINGLE set of privacy guidelines for its member countries and citizens, called GDPR, which requires businesses to protect the "personal data and privacy of EU citizens for transactions that occur within the EU."

Here is more on GDPR.

Interestingly (or not so), the US has a maze of regulations when it comes to digital privacy ("patch quilt protections"). To be fair, so did Europe, pre-GDPR.



Differences in Privacy

USA vs EU

USA	EU
1 Privacy laws change with each administration.	1 Privacy laws have less turnover when administrations change because most EU member states aren't as polarized as the US.
2 Individuals have little ownership of their online data, which allows large businesses can monetize consumer behavior and habits.	2 EU laws respect "private and family life" and allow citizens to delete their data.
3 Privacy laws are often a messy combination of public regulation, private self-regulation, and legislation which varies by state.	3 Privacy laws are generally more comprehensive and geared towards consumers.
4 Enforcement of privacy laws is carried out by several different government organizations, e.g. Federal Communications Commission (FCC) and Health Insurance Portability and Accountability Act (HIPAA).	4 Enforcement of privacy laws is carried out by one authority, equally for all 28 member states.
5 Numerous privacy organizations exist to provide legal framework, which ensure digital privacy to Americans. Ex: American Civil Liberties Union (ACLU) and the Electronic Frontier Foundation (EFF).	5 Due to the nature of EU rights, fewer privacy organizations exist but there are: The European Digital Rights (EDRi) and The European Privacy Association (EPA.)
6 Companies can keep data indefinitely, depending on their own Terms of Service.	6 EU citizens have the "right to be forgotten," meaning that search results can be removed if they are irrelevant or inadequate.

Sources:
https://www.marketplace.org/2017/04/20/tech/make-me-smart-kai-and-molly/blog-main-differences-between-internet-privacy-us-and-eu
http://politicsandpolicy.org/article/european-union-and-internet-data-privacy

GOVERNANCE

What are the RULES, related to the data we use?

The goal is to tease apart, learn about, and explore the connections between the following (data-related items): governance, curation, stewardship, MDM, provenance, metadata, security, privacy.

As you know, the purpose of capturing and storing data, is to process and benefit from it - this involves the use of statistics, data mining and machine learning.

But, that is not all there is to it! What about policies, procedures, rules, guidelines, practices... regarding the collection, storage and use of data?

Data Curation ['raw' => 'curated' datasets]

Regardless of collection procedures, analysis and usage, the ONE prime characteristic of data is QUALITY ('GIGO').

'Data curation' refers to set of processes and technologies ("methods and tools") that are focused on maintaining high-quality data in an organization, for the purposes of:

- visibility
- accessibility
- interoperability/heterogeneity
- reuse
- repurpose
- transparency
- ...

Any (which means ALL!) data-driven organization/s need(s) a 'data curation infrastructure' that supports curation practices and software.

Curation: key elements

Below are important points to keep in mind, while undertaking a data curation effort:

- the type of benefit derived from curated data, depends on the type of organization utilizing the curated datasets [eg. industrial R&D vs government vs new media companies]
- curation can be stimulated via incentives [that help justify the COST of curating]
- economic impact of curation can also help justify its need
- facilitating human-data interaction helps with curating - needs ways for non-technical users to handle data [eg. via natural language interfaces, semantic searching, viz, summarizing, transforming...] - building METADATA is a crucial step towards this
- large-scale curation efforts need to be hybrid between automated and human-involved efforts (curation by demonstration ['CBD'], crowdsourcing platforms, integration with enterprise data...]
- data curators need permission to access data they are curating; they need to be able to assign permissions (digital rights) to end-users of curated data; curators also need 'provenance' (data trail) info in order to determine curation specifics
- standards-based data models and representations (eg. ontology modeling using OWL) is necessary, for curation to include third-party/crowdsourced etc. data

Here is a description, by Craigin et. al. [Cragin, M., Heidorn, P., Palmer, C. L.; Smith, Linda C., An Educational Program on Data Curation, ALA Science & Technology Section Conference, (2007)]: "Data curation is the active and on-going management of data through its lifecycle of interest and usefulness; ... curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time".

Data Governance (DG)

From Wikipedia: 'Data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data'.

In other words, governance == curation?

NOT REALLY: As per the DAMA International Data Management Book of Knowledge, "Data Governance (DG) is defined as the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets." In other words, Governance is about POLICIES, which can be seen as complementary to Curation.

So an organization would have Governance **policies** in place, which would aid in Curation's producing **customized business data.**

Data Provenance

Provenance ~= "lineage".

'Data provenance documents the inputs, entities, systems, and processes that influence data of interest, in effect providing a historical record of the data and its origins.'

Provenance has to do with origins, while lineage has to do with tracing data's 'journey' to the current point of usage.

Provenance/lineage is a form of metadata that needs to be added to data, during curation.

Provenance helps establish TRUST (or lack thereof) in data. With scientific data for example, this is crucial [incorrect/invalid data would lead of the acceptance of incorrect hypotheses!].

Here is an interesting list of provenance-related issues related to scientific research (just fyi).

Lineage could have life or death consequences.

MDM [Master Data Management]

MDM is the management of "master data" (similar to a master key): In business, master data management (MDM) is a method used to define and manage the critical data of an organization to provide, with data integration, a single point of reference. [Wikipedia]

The idea is to maintain a single (meaningful, accurate, complete, timely) reference for data that is shared - this is done in order to maintain consistency. The alternative (to replicate such data for each request) would be highly problematic - would lead to inconsistency, errors, wasted disk space, increased network traffic, etc.

Here is more on MDM.

Governance, Security, Privacy - a TRINITY!

Security breaches are almost 'normal' - Yahoo, Home Depot, Facebook, Uber, Equifax... what is going on?

Governance is not being followed properly - policies for handling data and accountability, culture of/training in handling data, (pro)actively managing (sensitive) data - these are missing.

Privacy and security breaches are very costly, literally - lost revenue in the form of customer attrition, fines (eg. levied by SEC), lawsuit awards... These losses are monumental, compared to investing in technologies and policies that guard against breaches!

Note that maintaining security and privacy both involve minimizing RISK - something that every business ought to be concerned about.

Here is more on governance, as it relates to training ML models.

Security vs Privacy..

Data security/protection has to do with (preventing) UNAUTHORIZED access to data. Data privacy on the other hand, has to do with (limiting) AUTHORIZED access to data - related, but not identical!

Protection/security is a technical issue, related to protecting servers, encrypting data, restricting access (eg via passwords or biometrics), etc; privacy compliance on the other hand is a legal issue.

Also: data needs to be protectable first, before privacy can be ensured!

Discussion

Gov't surveillance - China

Social Credit - China

Ethics, and legalities, of fakevideo/fakeaudio/fakeimage/faketext... these are all weapons of 'information warfare'. Here is a related talk, and another.

COVID-19: privacy issues - location

COVID-19: provenance

COVID-19: disinformation

COVID-19: discrimination

COVID-19: data breach

...?