Building Software Systems

Lecture 5.3
Privacy Issues in AI

SAURABH SRIVASTAVA
ASSISTANT PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
IIT (ISM) DHANBAD

What is Privacy?

Privacy is considered as the *“ability of an individual (or an organisation) to control what information about her (or them) gets exposed to the outside world”*

- The data could be personal information like birthday or PAN
- Or organisational information such as Sales Targets or Employee Remunerations

Consequently, a "breach of Privacy" is an event where some information about the individual (or the organisation) is "leaked" to a source that was not explicitly authorised

- For instance, to an eavesdropper or a rival firm

In some parts of the world, e.g. the European Union, a Privacy breach can fetch substantial fines

- Of the order of €20 million, or 4% annual global turnover – whichever is higher !!

Any organisation that works with user data is therefore liable to take measures to protect its users’ Privacy concerns

Privacy in AI-intensive Systems (1/2)

AI-intensive Systems are usually “data centric”

- Machine Learning techniques rely on substantial amount of data to produce accurate real-world models
- While synthetic data could be used in initial stages, often the training data is curated out of user data

What data to capture and store?

- Systems that interact with end-users have the choice of capturing and storing large amount of data
- This may include user’s personal information, browsing routine, preferences etc.
- Organisations may be tempted to store *as much data as possible*, but considering the risks associated to a Privacy breach, it may not be wise to do so

Privacy by Design [1]

- A set of seven principles to keep privacy concerns in the loop while designing a system
- The principles however are mostly theoretical and implementing them in practice is not straightforward

Privacy in AI-intensive Systems (2/2)

Collection Limitation and *Data Minimisation* [1]

- Nevertheless, the principles do provide useful hints to avoid common Privacy pitfalls
- The idea of *Collection Limitation* says *"the collection of personal information must be fair, lawful and limited to that which is necessary for the specified purposes"*
- *Data Minimisation* stresses that *"the collection of personally identifiable information should be kept to a strict minimum"*

Utilising user data while honouring Privacy concerns

- Machine Learning techniques attempt to *find correlations* among input attributes to guess the output
- Privacy preserving techniques attempt to *remove or obfuscate correlations* in data
- There is a *trade-off* between Privacy and "Utility" – some correlations must remain in data for it to be useful, while others must be removed to minimise risks in case of a Privacy breach

The Privacy vs Utility Trade-off

In a nutshell,

- Utility is about "finding correlations in data"
- Privacy is about "removing correlations in data"

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH	BS	350 - 500K

Name	Roll Number	Department	Program	Income
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH	BS	350 - 500K

This data can be used to identify financially weaker students

But Alice doesn't want this information to be public

Name	Roll Number	Department		ange
Bob	1003	ME		- 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH	BS	350 - 500K

The Privacy vs Utility Trade-off

In a nutshell,

- Utility is about "finding correlations in data"
- Privacy is about "removing correlations in data”

What are the ways to remove "correlations” here?

- Anonymise data

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH	BS	350 - 500K

Anonymised Data

Name	Roll Number	Department	Program	Income Range
P1	1003	ME	BT	50K - 100K
P2	1002	CSE	MS	>500K
P3	1004	PHY	MT	100K - 350K
P4	1005	CSE	PHD	50K - 100K
P5	1006	MTH	BS	350 - 500K

The Privacy vs Utility Trade-off

In a nutshell,

- Utility is about "finding correlations in data"
- Privacy is about "removing correlations in data”

What are the ways to remove "correlations” here?

- Anonymise data
- Add “noise” following the “same distribution”

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH	BS	350 - 500K

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K

Inclusion of Spurious Rows

Name	Roll Number	Department	Program	Income Range
Eve	1011	CHEM	PHD	200K - 300K
Grace	1013	ART	MS	300K - 350K
John	1004	PHY	MT	100K - 350K
Frank	1012	ENG	BS	150K - 200K
Mary	1005	CSE	PHD	50K - 100K
Hank	1014	LAW	BT	100K - 150K
Charlie	1010	BIO	MS	<50K
Bob	1003	ME	BT	50K - 100K
José	1006	MTH	BS	350K - 500K
Alice	1002	CSE	MS	>500K

The Privacy vs Utility Trade-off

In a nutshell,

- Utility is about "finding correlations in data"
- Privacy is about "removing correlations in data”

What are the ways to remove "correlations” here?

- Anonymise data
- Add “noise” following the “same distribution”
- Remove “sensitive” columns

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH	BS	350 - 500K

Department	Program	Income Range
ME	BT	50K - 100K
CSE	MS	>500K
PHY	MT	100K - 350K
CSE	PHD	50K - 100K
MTH	BS	350 - 500K

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH		

The correlation between individuals and their incomes has been removed

Department	Program	Income Range
ME	BT	50K - 100K
CSE	MS	>500K
PHY	MT	100K - 350K
CSE	PHD	50K - 100K
MTH	BS	350 - 500K

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
		MTH	BS	350 - 500K

But some utility of the data is also "lost" (e.g. selecting financially weaker students for "scholarships")

Department	Program	Income Range
ME	BT	50K - 100K
CSE	MS	>500K
PHY	MT	100K - 350K
CSE	PHD	50K - 100K
MTH	BS	350 - 500K

The Privacy vs Utility Trade-off

In a nutshell,

- Utility is about "finding correlations in data"
- Privacy is about "removing correlations in data”

What are the ways to remove "correlations” here?

- Anonymise data
- Add “noise” following the “same distribution”
- Remove “sensitive” columns

Irrespective of what options we choose, the data almost always uses "some utility"

So, there is a *trade-off* here, and we need to find a mid-way out of it !!

- Usually, the decision here lies with Lead Architect of the system
- The solution may be a part of the Solution Architecture itself (e.g., deciding upon what data attributes to use)

Identifiers vs Quasi-Identifiers

Identifiers

- Definition: Data that uniquely identifies an individual
- Examples: Aadhaar Number, Voter Id Card Number, Passport Number etc.
- Characteristics: Direct identifiers that can pinpoint an individual without additional data
- Privacy Approach: Typically removed or encrypted to prevent direct linkage to an individual

Quasi-Identifiers

- Definition: Data that does not uniquely identify an individual itself but can do so when combined with other data
- Examples: Date of Birth, PIN Code, Gender, Category
- Characteristics: Can indirectly identify individuals when linked with other quasi-identifiers or external data
- Privacy Approach: Generalized or obfuscated to prevent re-identification

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
José	1006	MTH	BS	350 - 500K

Department	Program	Income Range
ME	BT	50K - 100K
CSE	MS	>500K
PHY	MT	100K - 350K
CSE	PHD	50K - 100K
MTH	BS	350 - 500K

Name	Roll Number	Department	Program	Income Range
Bob	1003	ME	BT	50K - 100K
Alice	1002	CSE	MS	>500K
John	1004	PHY	MT	100K - 350K
Mary	1005	CSE	PHD	50K - 100K
		MTH	BS	350 - 500K

We removed the Identifiers here, but if you know that Mary is the only PhD scholar in CSE, she can be identified

Department	Program	Income Range
ME	BT	50K - 100K
CSE	MS	>500K
PHY	MT	100K - 350K
CSE	PHD	50K - 100K
MTH	BS	350 - 500K

Identifiers vs Quasi-Identifiers

Identifiers

- Definition: Data that uniquely identifies an individual
- Examples: Aadhaar Number, Voter Id Card Number, Passport Number etc.
- Characteristics: Direct identifiers that can pinpoint an individual without additional data
- Privacy Approach: Typically removed or encrypted to prevent direct linkage to an individual

Quasi-Identifiers

- Definition: Data that does not uniquely identify an individual itself but can do so when combined with other data
- Examples: Date of Birth, PIN Code, Gender, Category
- Characteristics: Can indirectly identify individuals when linked with other quasi-identifiers or external data
- Privacy Approach: Generalized or obfuscated to prevent re-identification

It is the Quasi-Identifiers, which are often subjected to sophisticated privacy-breach attacks

- **A concept that is often used to tackle these issues with Quasi-Identifiers is *k*-anonymity**

The concept of *k*-Anonymity

What Is *k*-Anonymity?

- A model that prevents the re-identification of individuals in a dataset by ensuring each record is indistinguishable from at least $k-1$ others
- For the last example, it would mean that irrespective of any background knowledge (e.g., knowledge that Mary is the only PhD in CSE), there are still at least *k* rows in the dataset, within which, Mary's row is hidden

Ways to achieve *k*-anonymity

- Generalisation – Involves creating broader categories to hide individual rows

Department	Program	Income Range
ME	BT	50K - 100K
CSE	MS	>500K
PHY	MT	100K - 350K
CSE	PHD	50K - 100K
MTH	BS	350 - 500K

Department	Program	Income Range
ME	UG	50K - 100K
CSE	PG	>500K
PHY	PG	100K - 350K
CSE	PG	50K - 100K
MTH	UG	350 - 500K

The concept of *k*-Anonymity

What Is *k*-Anonymity?

- A model that prevents the re-identification of individuals in a dataset by ensuring each record is indistinguishable from at least $k-1$ others
- For the last example, it would mean that irrespective of any background knowledge (e.g., knowledge that Mary is the only PhD in CSE), there are still at least *k* rows in the dataset, within which, Mary's row is hidden

Ways to achieve *k*-anonymity

- Generalisation – Involves creating broader categories to hide individual rows
- Suppression – Omit or remove rows where privacy risk is higher (e.g., remove Mary's row from the data)
- Data Manipulation – Add more rows with spurious data to hide the individuals at risk

However, achieving optimal *k*-anonymity for a given value of *k* is not that easy for a dataset

- It is because the problem is proven to be NP Hard (in simple terms, there is no efficient algorithm for it, yet !!)
- There are, however, some tools that can achieve a best-effort approach towards anonymity (check the Further Reading section)

Homework

Privacy Breaches are a huge threat to Individual’s privacy

- Go through some of the previous such incidents:
  https://www.ekransystem.com/en/blog/real-life-examples-insider-threat-caused-breaches
  https://www.upguard.com/blog/biggest-data-breaches-in-healthcare
  https://www.upguard.com/blog/biggest-data-breaches-australia

Further Reading

Have a look at this section of the Sensitive Data Protection tutorial by Google:

- https://cloud.google.com/sensitive-data-protection/docs/compute-k-anonymity

Some tools that you may check out:

- AIJack
- pyCANON
- Pynonymizer

Also have a look at this comics on *Federated Learning*

- https://federated.withgoogle.com/