Chap3#4: Anonymization and Randomization based approaches #1

February 27, 2023

Devesh C Jinwala,
Professor, SVNIT and Adjunct Prof., CSE, IIT Jammu
Department of Computer Science and Engineering,
Sardar Vallabhhai National Institute of Technology, SURAT

Chap 2: ML Applications in Security: Topics to study

- Privacy Preservation, What is Privacy? Data Privacy. Machine Learning in Privacy Preservation: Four Main stakes to Privacy preservation in ML. Two principle approaches: (a) Augmenting the ML techniques with the conventional approaches in the domain of privacy preservation to achieve privacy viz. Homomorphic Encryption(HE Algorithms and the associated mathematics), Secret Multi-party Computations, Zero Knowledge Proofs, Anonymization techniques (e.g.)k-Anonymity, l-Diversity) Perturbation techniques (e.g. differential privacy) (b) ML-specific approaches like Federated Learning OR Ensemble Learning. Ethical issues and Law for data / process privacy : GDPR, Alexa, other relevant applications [6 hours]

Reviewing the theme of ML Paradigms for Privacy Preservation

Four Main stakes to Privacy preservation

There are four main stakes to privacy preservation in general:

- Privacy of the input data, input queries , web search queries
- Privacy of the computations
- Privacy of the output data, web search query results
- Data Privacy General Regulations, Data protection strategies, processes and principles

Four Main stakes to Privacy preservation

There are four main stakes to privacy preservation in general:

- Privacy of the input data, input queries , web search queries
- Privacy of the computations
- Privacy of the output data, web search query results
- Data Privacy General Regulations, Data protection strategies, processes and principles

We examine one of these viz. Privacy of Computations in greater detail shortly hereafter seeing main stakes to Privacy preservation in ML

Four Main stakes to Privacy preservation in ML

There are four main stakes to privacy preservation in general:

- Privacy of the input data
  - the assurance that other parties, including the model developer, will not be able to see a user's input data
- Privacy of the output data
  - the assurance that the output of a model is only accessible to the client whose data is being inferred upon.
- Privacy of the model
  - rhe assurance that a hostile party will not be able to steal the model
- Data privacy in training
  - the assurance that a malicious party will not reverse-engineer the training data - although gathering information about training data and model is more difficult than that for the data.

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....
    - cryptographic approaches like

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....
  - cryptographic approaches like
    - homomorphic encryption

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....
    - cryptographic approaches like
        - homomorphic encryption
        - secure multi-party computing,

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....
  - cryptographic approaches like
    - homomorphic encryption
    - secure multi-party computing,
    - Zero knowledge proofs

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....
    - cryptographic approaches like
        - homomorphic encryption
        - secure multi-party computing,
        - Zero knowledge proofs
    - perturbation techniques like differential privacy

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....
  - cryptographic approaches like
    - homomorphic encryption
    - secure multi-party computing,
    - Zero knowledge proofs
  - perturbation techniques like differential privacy
  - anonymization techniques like k-Anonymity and l-Diversity

Privacy Preserving Machine Learning: How to achieve?

The goal of privacy-preserving machine learning is

- to bridge the gap between privacy while receiving the benefits of machine learning.
- is a critical facilitator for the protection of acquired data and adhering to data privacy laws.

Privacy-preservation in ML

- is achieved by augmenting conventional ML with different strategies that protect data privacy, that include....
  - cryptographic approaches like
    - homomorphic encryption
    - secure multi-party computing,
    - Zero knowledge proofs
  - perturbation techniques like differential privacy
  - anonymization techniques like k-Anonymity and l-Diversity
  - ML-specific approaches like Federated Learning OR Ensemble Learning - the Privacy-Preserving Techniques - modifying the conventional ML training methods to keep user data private.

Augmenting ML for Privacy Preservation: Anonymization Methods

Anonymization Methods: Background

Anonymization method

- mainly applied to the databases, to preserve the privacy while mining the data.

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Anonymization Methods: Background

Anonymization method

- mainly applied to the databases, to preserve the privacy while mining the data.
- is useful when there is a data leak leading the violation of privacy....

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Anonymization Methods: Background

Anonymization method

- mainly applied to the databases, to preserve the privacy while mining the data.
- is useful when there is a data leak leading the violation of privacy....
- Let us look at an example.....

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Suppose the data that a hospital wishes to publish has the schema as follows

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital

Published Data

	Non-Sensitive Data			Sensitive Data
#	**Zip**	**Age**	**Nationality**	**Condition**
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Data leak!

#	**Name**	**Zip**	**Age**	**Nationality**
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Voter List

Figure: Data published but leaks

Src: B. Aditya Prakash, IITB and CMU

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Suppose the data that a hospital wishes to publish has the schema as follows
  - Attribute values which can uniquely identify an individual {zip-code, nationality, age } or/and {name} or/and {SSN}

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital

Published Data

	Non-Sensitive Data			Sensitive Data
#	**Zip**	**Age**	**Nationality**	**Condition**
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Data leak!

#	**Name**	**Zip**	**Age**	**Nationality**
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Voter List

Figure: Data published but leaks

Src: B. Aditya Prakash, IITB and CMU

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Suppose the data that a hospital wishes to publish has the schema as follows
  - Attribute values which can uniquely identify an individual {zip-code, nationality, age } or/and {name} or/and {SSN}
  - sensitive information corresponding to individuals {medical condition, salary, location }

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital

Published Data

	Non-Sensitive Data			Sensitive Data
#	**Zip**	**Age**	**Nationality**	**Condition**
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Data leak!

Voter List

#	Name	Zip	Age	Nationality
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Figure: Data published but leaks

Src: B. Aditya Prakash, IITB and CMU

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Suppose the data that a hospital wishes to publish has the schema as follows
  - Attribute values which can uniquely identify an individual {zip-code, nationality, age } or/and {name} or/and {SSN}
  - sensitive information corresponding to individuals {medical condition, salary, location }
- the aim is to prevent a situation where even if one removes the direct uniquely identifying attributes from a table, there are some fields that may still uniquely identify some individual.

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital

Published Data

	Non-Sensitive Data			Sensitive Data
#	**Zip**	**Age**	**Nationality**	**Condition**
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Data leak!

Voter List

#	Name	Zip	Age	Nationality
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Figure: Data published but leaks

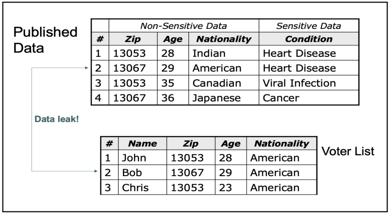
Src: B. Aditya Prakash, IITB and CMU

Anonymization method. Let us look at an example....

- Suppose the data that a hospital wishes to publish has the schema as follows
  - Attribute values which can uniquely identify an individual {zip-code, nationality, age } or/and {name} or/and {SSN}
  - sensitive information corresponding to individuals {medical condition, salary, location }
- the aim is to prevent a situation where even if one removes the direct uniquely identifying attributes from a table, there are some fields that may still uniquely identify some individual.
- The attacker can join them with other sources and identify individuals.

	Non-Sensitive Data			Sensitive Data	
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure: Data with a hospital



Published Data

	Non-Sensitive Data			Sensitive Data
#	**Zip**	**Age**	**Nationality**	**Condition**
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection
4	13067	36	Japanese	Cancer

Voter List

#	Name	Zip	Age	Nationality
1	John	13053	28	American
2	Bob	13067	29	American
3	Chris	13053	23	American

Figure: Data published but leaks

Src: B. Aditya Prakash, IITB and CMU

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Even if we remove the direct uniquely identifying attributes

	Non-Sensitive Data			Sensitive Data
#	Zip	Age	Nationality	Condition
...	...	...	...	...

$\underbrace{\text{Zip} \quad \text{Age} \quad \text{Nationality}}_{\text{Quasi-Identifiers}}$

Figure: Data with a hospital

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Even if we remove the direct uniquely identifying attributes
- There are some fields that may still uniquely identify some individual!

	Non-Sensitive Data			Sensitive Data
#	**Zip**	**Age**	**Nationality**	**Condition**
...	...	...	...	...

Quasi-Identifiers

Figure: Data with a hospital

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Even if we remove the direct uniquely identifying attributes
- There are some fields that may still uniquely identify some individual!
- The attacker can join them with other sources and identify individuals

	Non-Sensitive Data			Sensitive Data
#	**Zip**	**Age**	**Nationality**	**Condition**
...	...	...	...	...

Quasi-Identifiers

Figure: Data with a hospital

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Even if we remove the direct uniquely identifying attributes
- There are some fields that may still uniquely identify some individual!
- The attacker can join them with other sources and identify individuals

	Non-Sensitive Data			Sensitive Data
#	Zip	Age	Nationality	Condition
...	...	...	...	...

Quasi-Identifiers

Figure: Data with a hospital

Anonymization Methods: Background...

Anonymization method. Let us look at an example....

- Even if we remove the direct uniquely identifying attributes
- There are some fields that may still uniquely identify some individual!
- The attacker can join them with other sources and identify individuals

Hence the need for anonymization methods

	Non-Sensitive Data			Sensitive Data
#	Zip	Age	Nationality	Condition
...	...	...	...	...

Quasi-Identifiers

Figure: Data with a hospital

Src: B. Aditya Prakash, IITB and CMU

Anonymization Methods: Background

Anonymization method

- was first proposed by Sweeney in the paper referenced below.

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Anonymization Methods: Background

Anonymization method

- was first proposed by Sweeney in the paper referenced below.
- mainly applied to the databases, to preserve the privacy while mining the data.

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Anonymization method

- was first proposed by Sweeney in the paper referenced below.
- mainly applied to the databases, to preserve the privacy while mining the data.
- the focus is to change data in such a way that for each tuple in the resulting table there are atleast (k-1) other tuples with the same value for the quasi-identifier

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Anonymization method

- was first proposed by Sweeney in the paper referenced below.
- mainly applied to the databases, to preserve the privacy while mining the data.
- the focus is to change data in such a way that for each tuple in the resulting table there are atleast (k-1) other tuples with the same value for the quasi-identifier
- this is to prevent a situation where even if one removes the direct uniquely identifying attributes from a table, there are some fields that may still uniquely identify some individual.

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Anonymization method

- was first proposed by Sweeney in the paper referenced below.
- mainly applied to the databases, to preserve the privacy while mining the data.
- the focus is to change data in such a way that for each tuple in the resulting table there are atleast (k-1) other tuples with the same value for the quasi-identifier
- this is to prevent a situation where even if one removes the direct uniquely identifying attributes from a table, there are some fields that may still uniquely identify some individual.
- here, we have a 4-anonymized table

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

Figure: Data with a hospital

Src: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression – P. Samarati and L. Sweeney, 1998,
Latanya Sweeney, k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570

Techniques for Anonymization

Techniques

- Data Swapping

Techniques for Anonymization

Techniques

- Data Swapping
- Randomization

Techniques for Anonymization

Techniques

- Data Swapping
- Randomization
- Generalization

Techniques for Anonymization

Techniques

- Data Swapping
- Randomization
- Generalization
  - Replace the original value by a semantically consistent but less specific value

Techniques for Anonymization

Techniques

- Data Swapping
- Randomization
- Generalization
  - Replace the original value by a semantically consistent but less specific value
- Suppression

Techniques for Anonymization

Techniques

- Data Swapping
- Randomization
- Generalization
  - Replace the original value by a semantically consistent but less specific value
- Suppression
  - Data not released at all

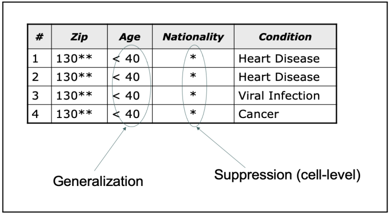
Techniques for Anonymization

Techniques

- Data Swapping
- Randomization
- Generalization
  - Replace the original value by a semantically consistent but less specific value
- Suppression
  - Data not released at all
  - Can be Cell-Level or (more commonly) Tuple-Level

Data Generalization

- is the process of creating a broader categorization of the data in a database,



#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Figure: Data Generilization/Suppresion

Anonymization Methods: Generalization and Suppression

Data Generalization

- is the process of creating a broader categorization of the data in a database,
- creating a more general picture of the trends or insights it provides.

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

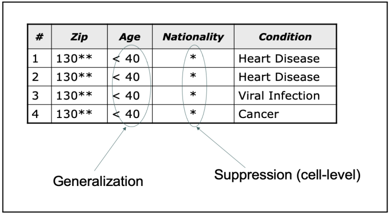
Generalization

Suppression (cell-level)

Figure: Data Generilization/Suppresion

Data Generalization

- is the process of creating a broader categorization of the data in a database,
- creating a more general picture of the trends or insights it provides.
- involves deliberately excluding some data to make them less identifiable.



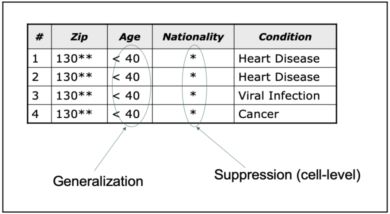
#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Figure: Data Generilization/Suppresion

Anonymization Methods: Generalization and Suppression

Data Generalization

- is the process of creating a broader categorization of the data in a database,
- creating a more general picture of the trends or insights it provides.
- involves deliberately excluding some data to make them less identifiable.
- here, data can be modified within a series of ranges with logical limits.



#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Figure: Data Generilization/Suppresion

Data Generalization

- is the process of creating a broader categorization of the data in a database,
- creating a more general picture of the trends or insights it provides.
- involves deliberately excluding some data to make them less identifiable.
- here, data can be modified within a series of ranges with logical limits.
- the result is a reduced granularity of the data, making it difficult or even impossible to retrieve the exact values associated with an individual.

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Generalization

Suppression (cell-level)

Figure: Data Generilization/Suppresion

Anonymization Methods: Generalization and Suppression

Data Generalization

- is the process of creating a broader categorization of the data in a database,
- creating a more general picture of the trends or insights it provides.
- involves deliberately excluding some data to make them less identifiable.
- here, data can be modified within a series of ranges with logical limits.
- the result is a reduced granularity of the data, making it difficult or even impossible to retrieve the exact values associated with an individual.

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

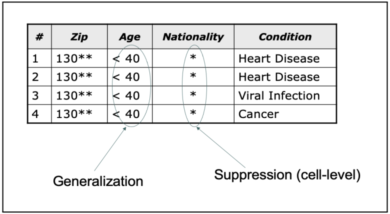
Generalization

Suppression (cell-level)

Figure: Data Generilization/Suppresion

Data Generalization

- is the process of creating a broader categorization of the data in a database,
- creating a more general picture of the trends or insights it provides.
- involves deliberately excluding some data to make them less identifiable.
- here, data can be modified within a series of ranges with logical limits.
- the result is a reduced granularity of the data, making it difficult or even impossible to retrieve the exact values associated with an individual.



#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Figure: Data Generilization/Suppresion

In Data Suppression certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'.

Src: Prof B. Aditya Prakash, IITB and CMU

Anonymization Methods: Generalization Hierarchies

- Data owner defines how values can be generalized

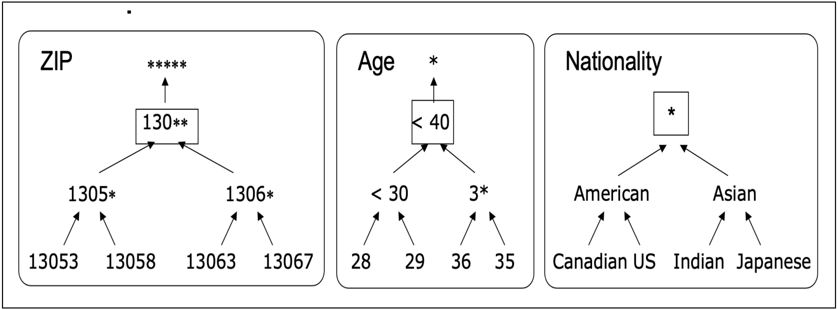


Figure: Data Generilization Hierarchies

Src: Prof B. Aditya Prakash, IITB and CMU

Anonymization Methods: Generalization Hierarchies

- Data owner defines how values can be generalized
- A table generalization is created by generalizing all values in a column to a specific level of generalization

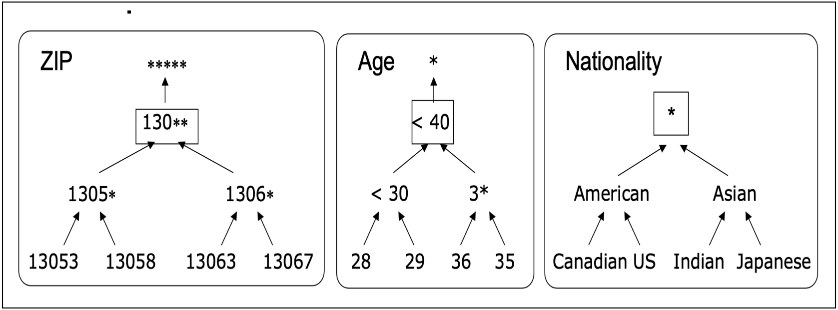


Figure: Data Generilization Hierarchies

Src: Prof B. Aditya Prakash, IITB and CMU

- There are many k-anonymizations – which one to pick?

#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13053	< 40	*	Viral Infection
3	13067	< 40	*	Heart Disease
4	13067	< 40	*	Cancer

2-minimal Generalizations

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Viral Infection
3	130**	3*	Asian	Heart Disease
4	130**	3*	Asian	Cancer

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Viral Infection
3	130**	< 40	*	Heart Disease
4	130**	< 40	*	Cancer

NOT a 2-minimal Generalization

Figure: K Minimum Generalization

Anonymization Methods: K-minimal Generalizations

- There are many k-anonymizations – which one to pick?
- Intuitively one that does not generalize the data more than needed (decrease in utility of the published dataset!)

#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13053	< 40	*	Viral Infection
3	13067	< 40	*	Heart Disease
4	13067	< 40	*	Cancer

2-minimal Generalizations

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Viral Infection
3	130**	3*	Asian	Heart Disease
4	130**	3*	Asian	Cancer

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Viral Infection
3	130**	< 40	*	Heart Disease
4	130**	< 40	*	Cancer

NOT a 2-minimal Generalization

Figure: K Minimum Generalization

Anonymization Methods: K-minimal Generalizations

- There are many k-anonymizations – which one to pick?
- Intuitively one that does not generalize the data more than needed (decrease in utility of the published dataset!)
- K-minimal generalization: A k-anonymized table that is not a generalization of another k-anonymized table

#	Zip	Age	Nationality	Condition
1	13053	< 40	*	Heart Disease
2	13053	< 40	*	Viral Infection
3	13067	< 40	*	Heart Disease
4	13067	< 40	*	Cancer

2-minimal Generalizations

#	Zip	Age	Nationality	Condition
1	130**	< 30	American	Heart Disease
2	130**	< 30	American	Viral Infection
3	130**	3*	Asian	Heart Disease
4	130**	3*	Asian	Cancer

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Viral Infection
3	130**	< 40	*	Heart Disease
4	130**	< 40	*	Cancer

NOT a 2-minimal Generalization

Figure: K Minimum Generalization

Using k-Anonymization

- To use k-anonymity to process a dataset so that it can be released with privacy protection, a data scientist must first examine the dataset and

Using k-Anonymization

- To use k-anonymity to process a dataset so that it can be released with privacy protection, a data scientist must first examine the dataset and
  - decide if each attribute (column) is an identifier (identifying), a non-identifier (not-identifying), or a quasi-identifier (somewhat identifying).

Using k-Anonymization

- To use k-anonymity to process a dataset so that it can be released with privacy protection, a data scientist must first examine the dataset and
  - decide if each attribute (column) is an identifier (identifying), a non-identifier (not-identifying), or a quasi-identifier (somewhat identifying).
  - Identifiers such as names are suppressed, non-identifying values are allowed to remain, and the quasi-identifiers need to be processed

Using k-Anonymization

- To use k-anonymity to process a dataset so that it can be released with privacy protection, a data scientist must first examine the dataset and
  - decide if each attribute (column) is an identifier (identifying), a non-identifier (not-identifying), or a quasi-identifier (somewhat identifying).
  - Identifiers such as names are suppressed, non-identifying values are allowed to remain, and the quasi-identifiers need to be processed
  - this should be such that every distinct combination of quasi-identifiers designates at least k records.

Using k-Anonymization

- To use k-anonymity to process a dataset so that it can be released with privacy protection, a data scientist must first examine the dataset and
  - decide if each attribute (column) is an identifier (identifying), a non-identifier (not-identifying), or a quasi-identifier (somewhat identifying).
  - Identifiers such as names are suppressed, non-identifying values are allowed to remain, and the quasi-identifiers need to be processed
  - this should be such that every distinct combination of quasi-identifiers designates at least k records.
- Limitation: K-anonymity alone does not provide full privacy. This can be seen from the next diagram

Using k-Anonymization

- To use k-anonymity to process a dataset so that it can be released with privacy protection, a data scientist must first examine the dataset and
  - decide if each attribute (column) is an identifier (identifying), a non-identifier (not-identifying), or a quasi-identifier (somewhat identifying).
  - Identifiers such as names are suppressed, non-identifying values are allowed to remain, and the quasi-identifiers need to be processed
  - this should be such that every distinct combination of quasi-identifiers designates at least k records.
- Limitation: K-anonymity alone does not provide full privacy. This can be seen from the next diagram

Using k-Anonymization

- To use k-anonymity to process a dataset so that it can be released with privacy protection, a data scientist must first examine the dataset and
  - decide if each attribute (column) is an identifier (identifying), a non-identifier (not-identifying), or a quasi-identifier (somewhat identifying).
  - Identifiers such as names are suppressed, non-identifying values are allowed to remain, and the quasi-identifiers need to be processed
  - this should be such that every distinct combination of quasi-identifiers designates at least k records.
- Limitation: K-anonymity alone does not provide full privacy. This can be seen from the next diagram

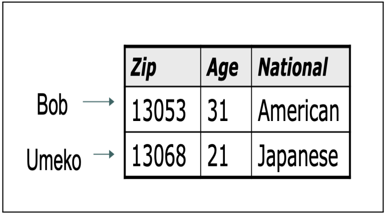


Figure: KAnonymity Attack

k-Anonymization Attack

Original Data →

	Non-Sensitive Data			Sensitive Data
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure: KAnonymity Attack

k-Anonymization Attack

4-anonymized Table

	Non-Sensitive Data			Sensitive Data
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*Cancer*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
8	*1485**	*> = 40*	*	*Viral Infection*
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Umeko Matches here (rows 1–4)

Bob Matches here (rows 9–12)

Figure: KAnonymity Attack

k-Anonymization Attack

4-anonymized Table

	Non-Sensitive Data			Sensitive Data
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130*			Heart Disease
3	130*			Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*Cancer*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
8	*1485**	*> = 40*	*	*Viral Infection*
9	130**	3*	*	Cancer
10				Cancer
11	130			Cancer
12	130**	3*	*	Cancer

Umeko Matches here

Umeko has Viral Infection!

Bob Matches here

Bob has Cancer!

Figure: KAnonymity Attack

k-Anonymization Limitation

- Basic Reasons for leak – Sensitive attributes lack diversity in values - Homogeneity Attack

k-Anonymization Limitation

- Basic Reasons for leak – Sensitive attributes lack diversity in values - Homogeneity Attack
- Attacker has additional background knowledge - Background knowledge Attack

k-Anonymization Limitation

- Basic Reasons for leak – Sensitive attributes lack diversity in values - Homogeneity Attack
- Attacker has additional background knowledge - Background knowledge Attack
- Hence a new solution has been proposed in-addition to k-anonymity – l-diversity