Security and Privacy

Privacy Preserving Data Publishing (PPDP)

Data Collection

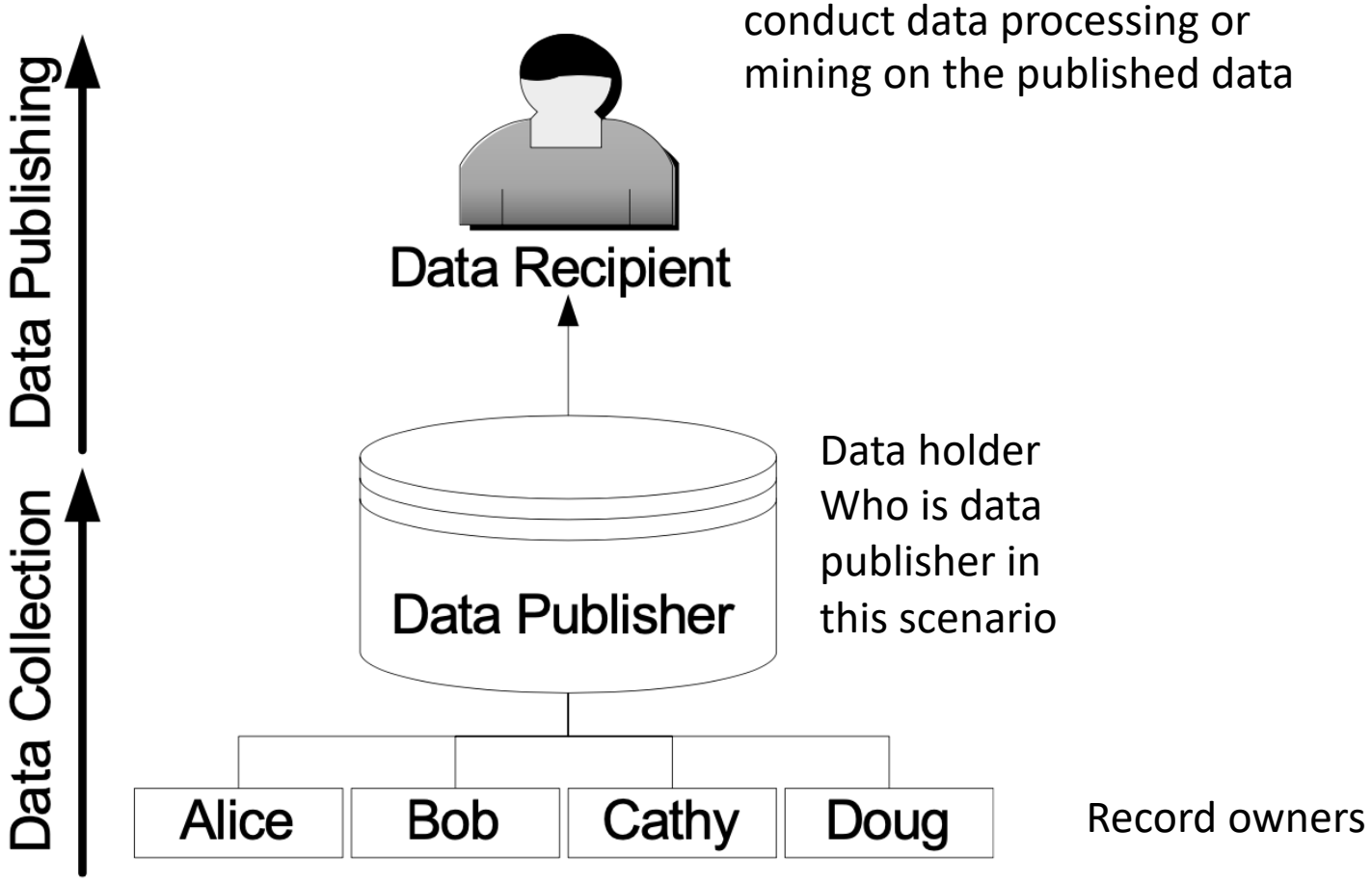
Data Visualization

Data Analysis or Data Mining

- **Data mining** is the process of **extracting useful, interesting, and previously unknown information** from large data sets.
- The success of data mining relies on:
  - The availability of **high quality data**
  - **Effective data sharing**
- Data sharing (publishing) driven by:
  - mutual benefits
  - regulations that require certain data to be published (e.g., medical center)

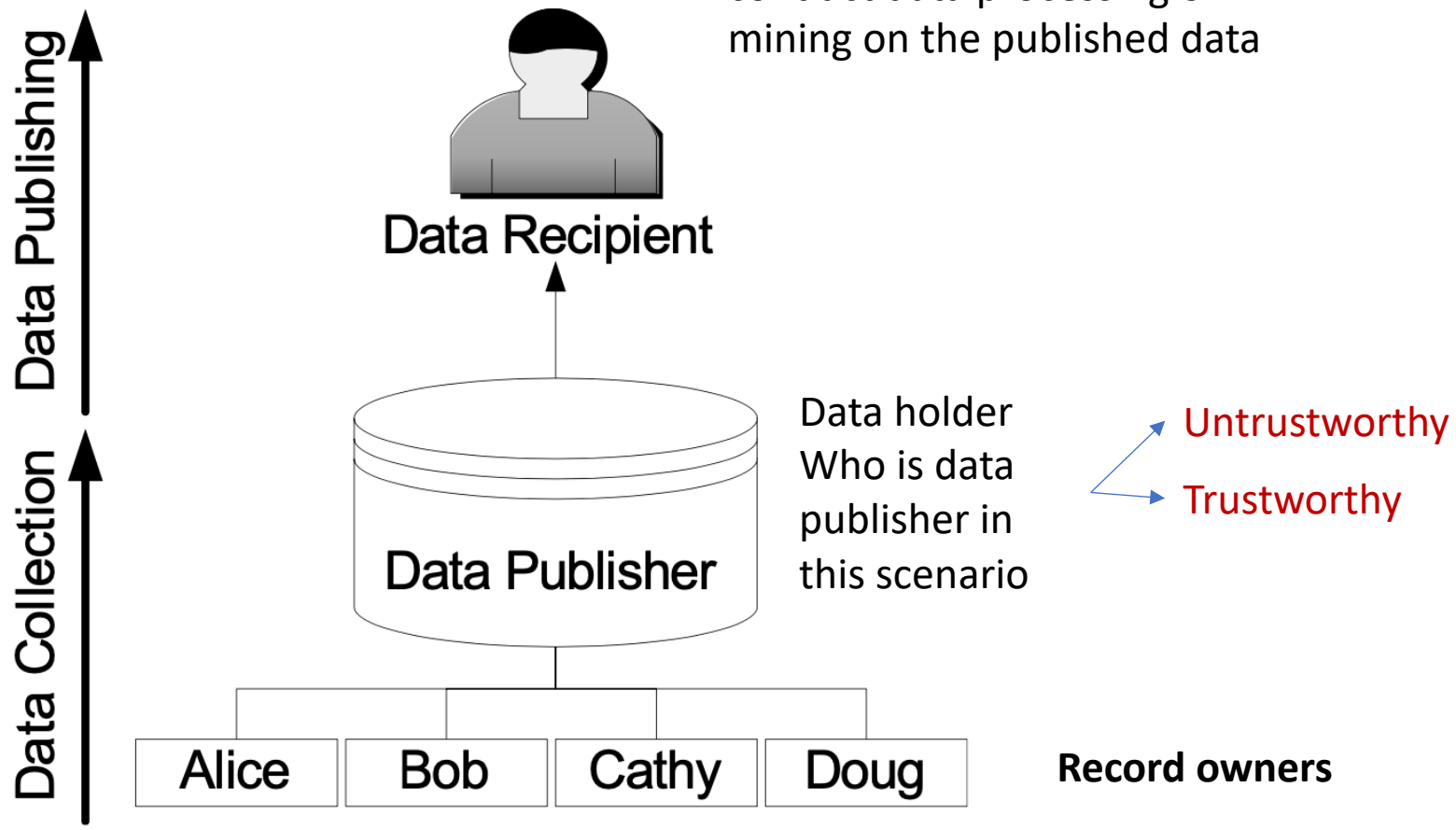
A typical scenario of data collection and publishing

- Data collection phase
  - data holder collects data from record owners
- Data publishing phase
  - data holder releases the collected data
- Example:
  - Hospital: data holder
  - Patients: record owners
  - Medical center: data recipient



A typical scenario of data collection and publishing

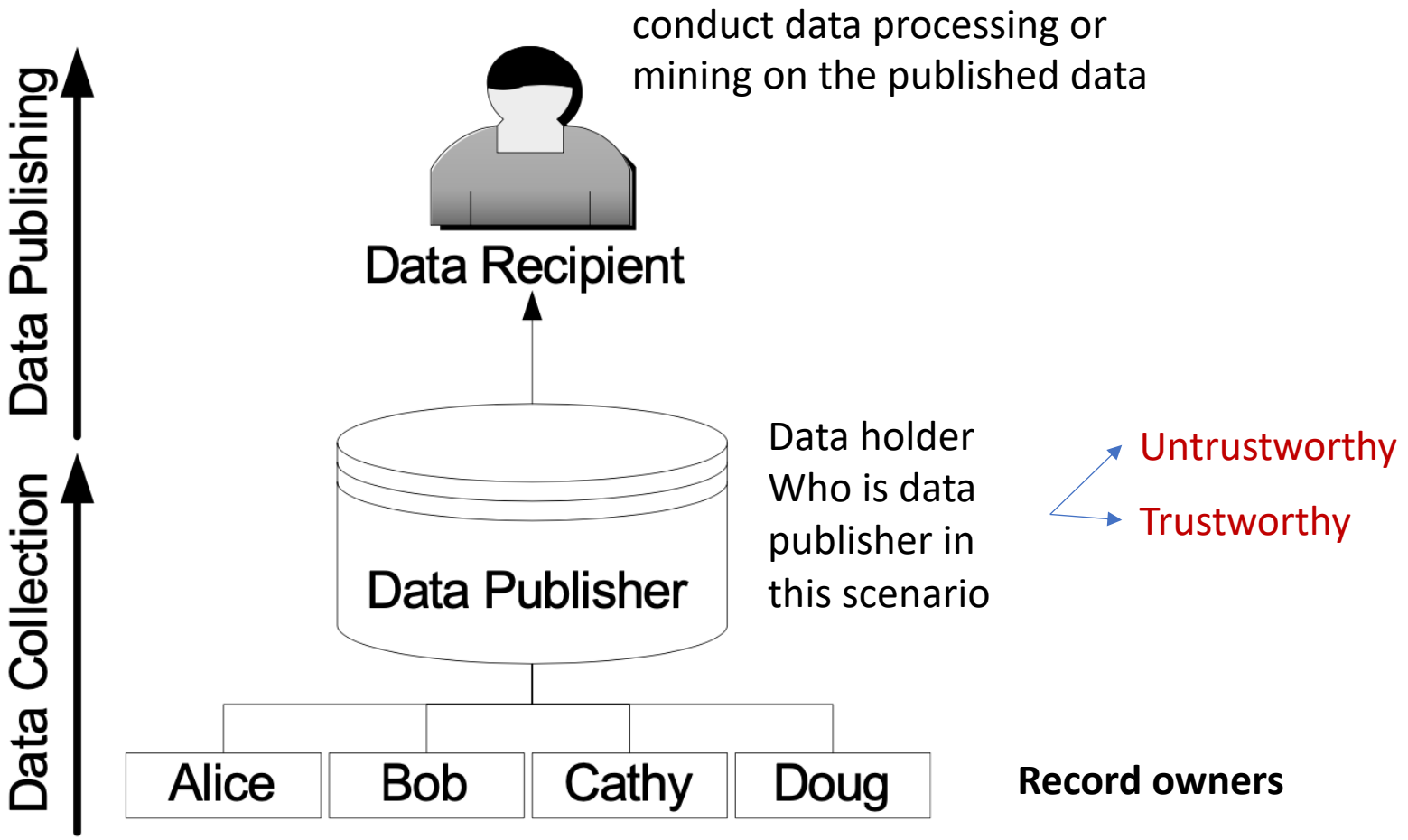
- Two models of data holders
  - **Untrustworthy**: may attempt to identify sensitive information from record owners
  - **Trustworthy**: use different techniques to collect records anonymously from their owners



trust is not transitive to the data recipient

A typical scenario of data collection and publishing

- two models of data holders
  - Untrustworthy: may attempt to identify sensitive information from record owners
  - Trustworthy: use different techniques to collect records anonymously from their owners



we assume the trusted model of data holders and consider privacy issues in the data publishing phase

Data Publishing Privacy

- **Objective:**

  Entities wish to release data collections either publicly or to third parties for data analysis without disclosing the ownership of the sensitive data

- Current privacy protection practice primarily relies on: **Policies**, **guidelines**, and **Agreements**
  - Policies and guidelines to **restrict the types of publishable data**
  - Agreements on **the use and storage** of sensitive data

How to Achieve Data Publishing Privacy?

Develop **methods** and **tools** for publishing data in a hostile environment so that the published **data remain practically useful** while **individual privacy is preserved**.

This undertaking is called **privacy-preserving data publishing (PPDP)**, which can be viewed as a **technical response** to **complement the privacy policies, guidelines and agreements**.

What is privacy preservation?

In 1977, Dalenius provided a very precise definition:

"Access to the published data should not enable the adversary to learn anything extra about any target victim compared to no access to the database, even with the **presence of any adversary's background knowledge** obtained from other sources. "

Such **absolute privacy protection** is almost impossible to achieve in real-life

PPDP - Terminologies

Data holder has a table of data records:

D (*Explicit Identifier*, *Quasi Identifier*, *Sensitive Attributes*, *Non-Sensitive Attributes*)

- Each data record includes:
  - Explicit Identifier
  - Quasi Identifier
  - Sensitive Attributes
  - Non-Sensitive Attributes

PPDP - Terminologies

- **Explicit Identifier or Personally identifiable information (PII):** a set of attributes, such as name, phone number, address, and social security number (SSN), containing information that explicitly/uniquely identify (PII) the record owner.

- **Quasi-identifier (QID)**: attribute (or set of attributes) that do not explicitly identify a user, but can be combined with data from other public sources to *de-anonymize* the owner of a record
  - ***5-digit ZIP code, birth date, gender*** uniquely identify 87% of the population in the U.S.

PPDP - Terminologies

- **Sensitive attribute**: individual-specific private/sensitive attributes that should not be publicly disclosed; may be also linked to identify individuals
  - E.g., diseases in medical records, salary, disability status

- **Non-Sensitive attribute:** contains all attributes that do not fall into the previous three categories

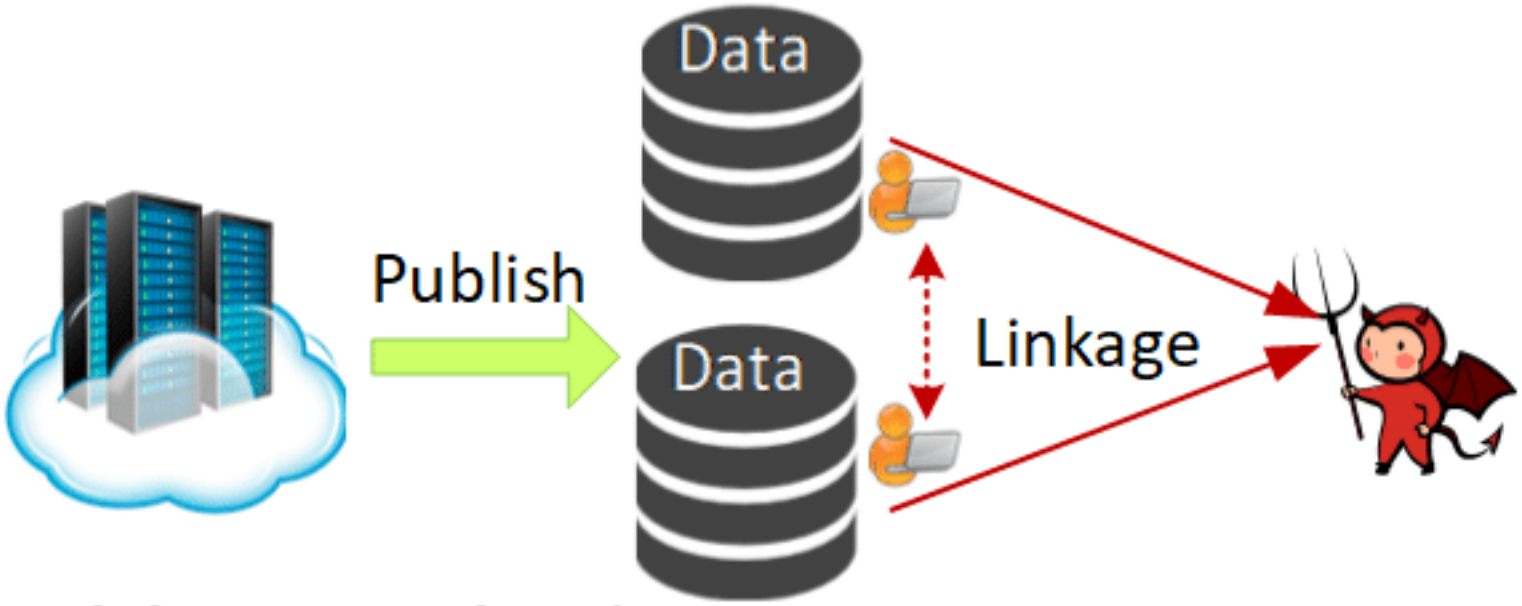
PPDP - Anonymization

- Anonymization: refers to the PPDP approach that seeks to hide the **identity and/or the sensitive data** of record owners

- First idea: remove personally identifiable information (PII)/explicit identifiers (a form of Anonymization):
  - Name, social security number, phone number, email address, …

Is it enough?

Linkage Attack



Linkage Attack: Adversary acquires private information by correlating multiple datasets

Linkage Attack

- Linkage attack occurs when an adversary is able to **link a record owne**r to:
  - a record in a published data table -> **record Linkage**
  - a sensitive attribute in a published data table -> **attribute Linkage**
  - the published data table itself -> **table linkage**

In all three types of linkages, we assume that the adversary **knows the QID of the victim**.

Linkage Attack

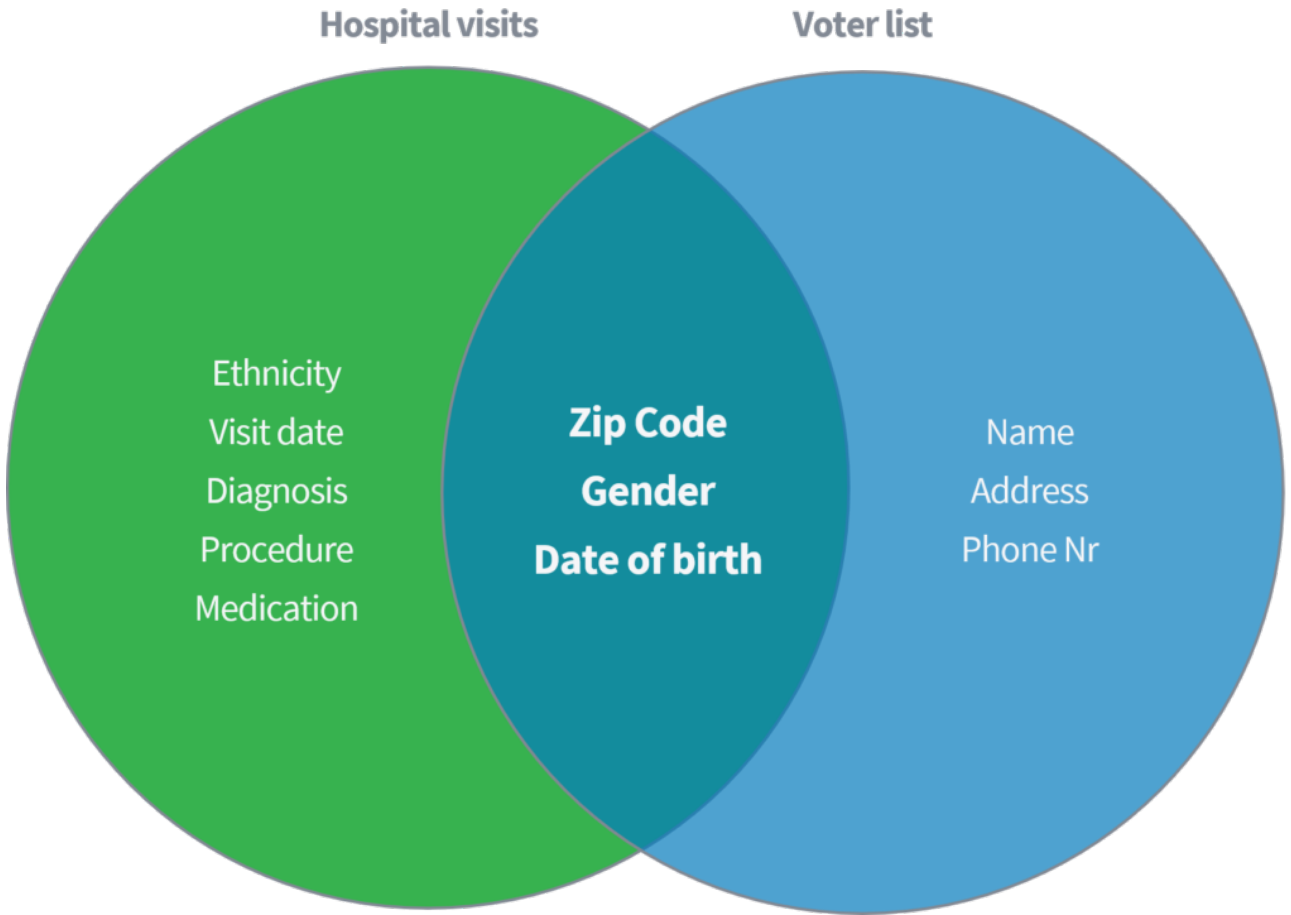
- In record and attribute linkages, we further assume that the adversary knows the **victim's record is in the released table**, and seeks to identify the victim's record and/or sensitive information from the table

- In table linkage, the attack **seeks to determine the presence or absence of the victim's record in the released table**.

A data table is considered to be privacy-preserving if it can effectively prevent the adversary from successfully performing these linkages.

Linkage Attack

- 87% of the US population uniquely identifiable by (5-digit ZIP code, birth date, gender)

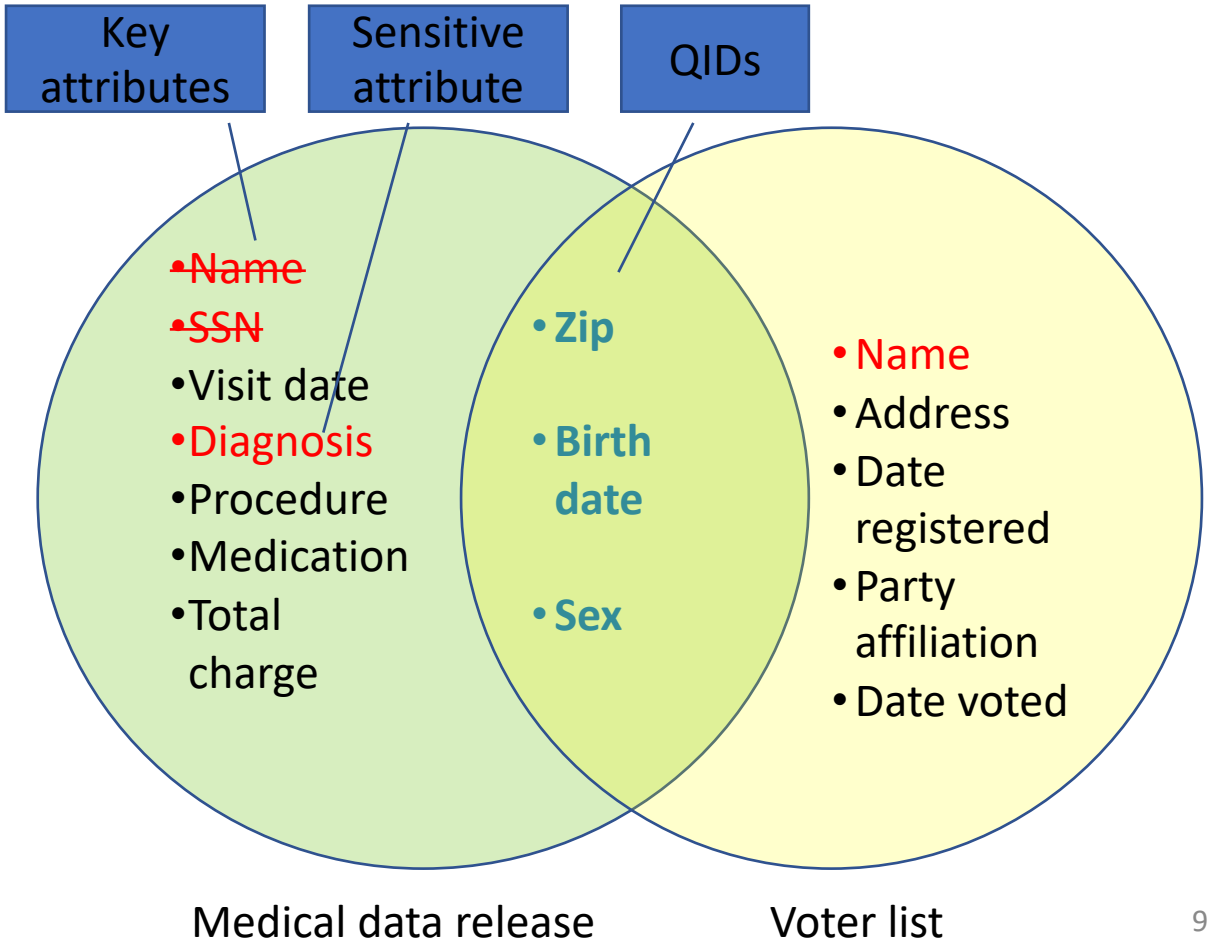


Latanya Sweeney's Attack (1997)

William Weld, former governor of the state of Massachusetts was **uniquely identified** using Zip, Birth date and Sex

To perform such linking attacks, the adversary needs **two pieces of prior knowledge**:

- the **victim's record** exist in the released data
- the **quasi-identifier** of the victim



Linkage Attack - Example

Medical Data in Data Holder/ data publisher

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

QID <Zipcode, Age, Sex>

Voter registration data in Data Recipient

Victim: Alice

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

QID = <47677, 29, F>

Linkage Attack - Example

Medical Records in Data Holder/ data publisher

ID	QID		SA
SSN	Zipcode	Age	Disease
012-345-6789	10598	24	HIV
823-627-9291	90210	37	Hepatitis C
987-654-3210	10547	26	HIV
382-827-8264	90345	38	Hepatitis C
847-872-7276	89119	36	Diabetes
422-061-0089	02139	25	HIV

QID = <Zipcode, Age>

Voter registration data in Data Recipient

Name	Zipcode	Age
Mary A.	90345	38
John S.	89119	36
Ann L.	02139	31
Jack M.	10562	57
Joy M.	10547	26
Victor B.	90345	46
Peter P.	01239	25
Diana X.	10598	24
William W.	90210	37
Sue G.	10547	26

QID = <10547, 26>

QID = <90210, 37>

Solution to Linkage Attack

To prevent linking attacks, the data holder publishes an anonymous **table T**

T (**QID′**, **Sensitive Attributes**, **Non-Sensitive Attributes**)

- **QID′** : is an *anonymous* version of the original **QID** obtained by applying ***anonymization operations*** to the attributes in **QID** in the original table D.

Anonymization operations hide some detailed information so that multiple records become indistinguishable with respect to **QID′**.

Solution to Linkage Attack

Anonymization operation: hide some detailed information so that multiple records become indistinguishable with respect to **QID'**.

Anonymization Problem: is to use a method to produce an anonymous **Table T** that **satisfies a given privacy requirement** determined and to **retain as much data utility as possible.**

Anonymization Operations

- Generalization
  - Replacement of a value for a more general one
- Suppression
  - Removal of some attribute values or records
- Anatomization
  - De-associates QIDs and sensitive attributes
- Permutation-Perturbation
  - Replacement of original data for synthetic values

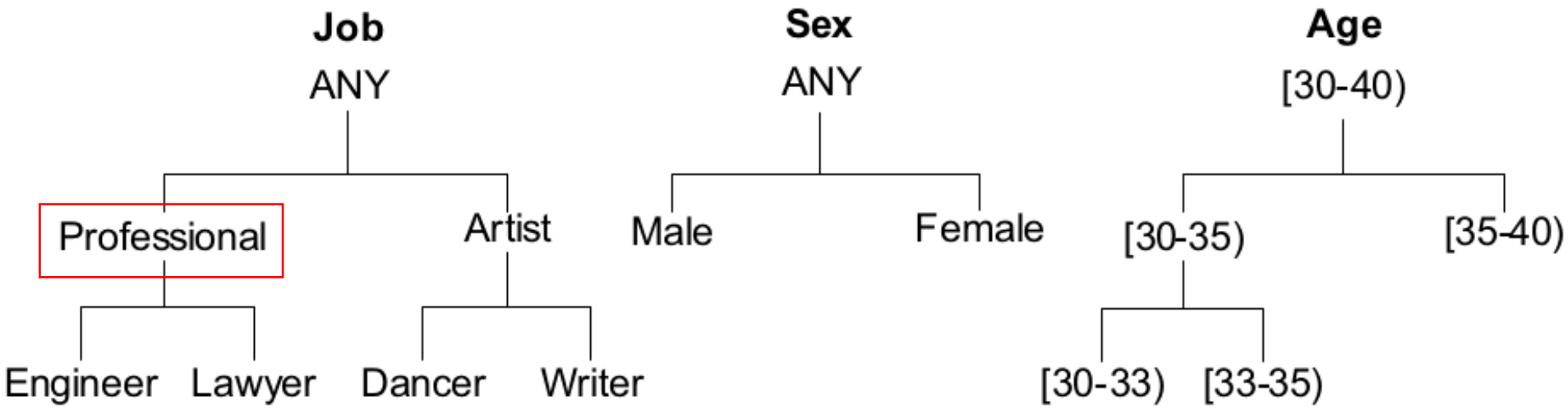
Generalization or Suppression operation hides some details in QID

Anatomization and permutation de-associate the correlation between QID and sensitive attributes by grouping or shuffling sensitive values in a QID group

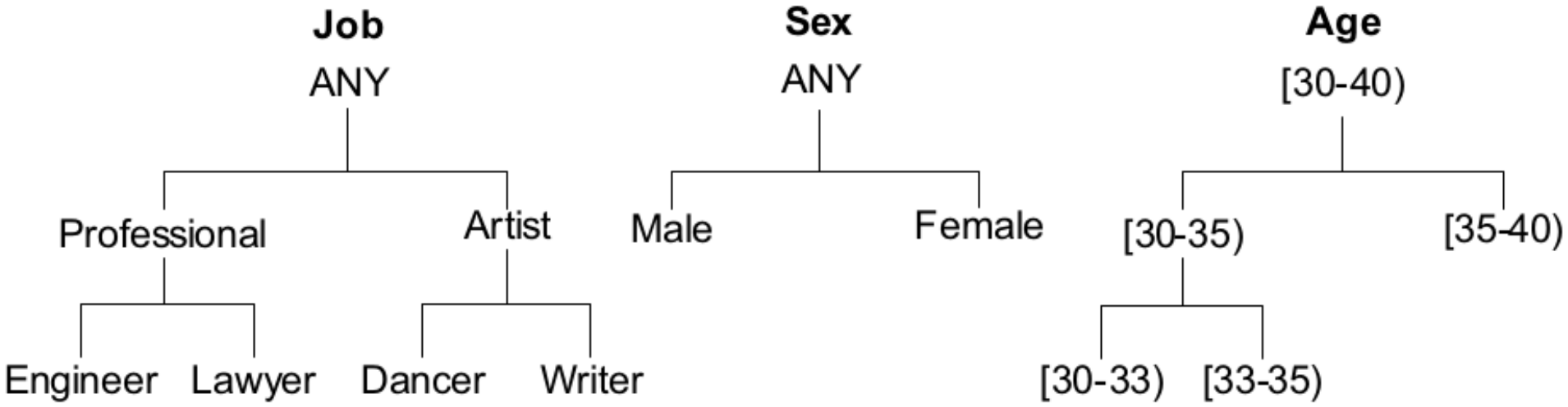
Generalization

- For a **categorical attribute**, a specific value can be replaced with a general value according to a given taxonomy.
- For a **numerical attribute**, exact values can be replaced with an interval that covers exact values.
- Example: Taxonomy trees for Job, Sex, Age

Professional is more general than the child nodes **Engineer** and **Lawyer**

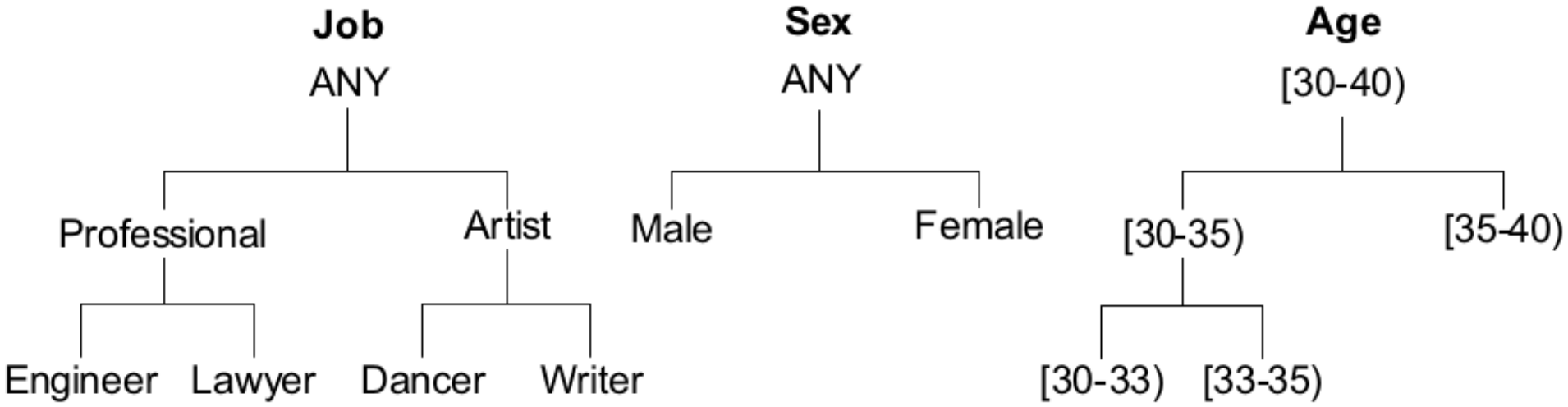


Generalization



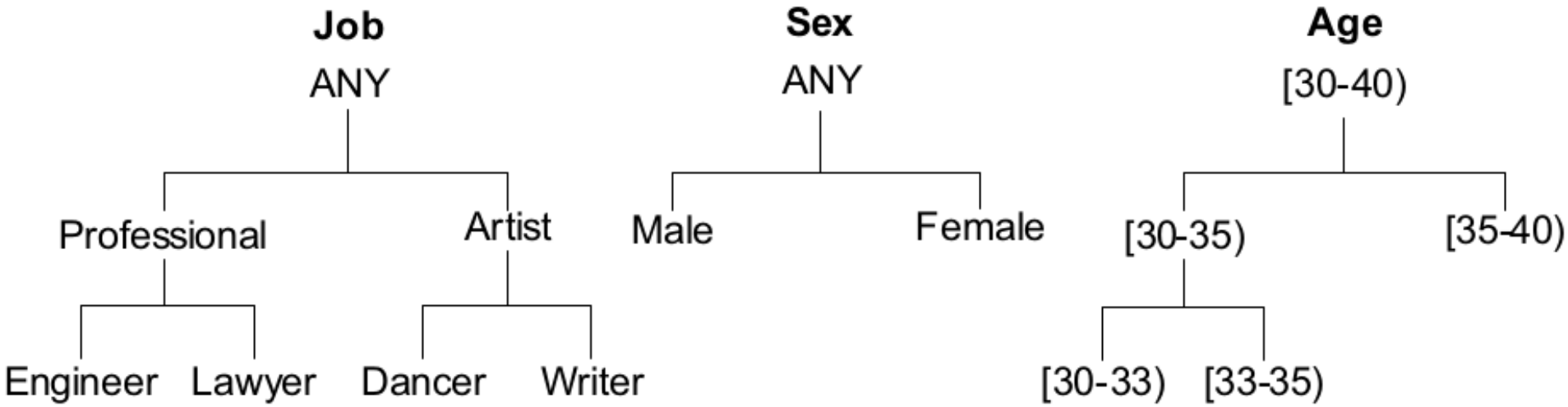
- A ***generalization*** replaces some values with a parent value in the taxonomy of an attribute.
- The reverse operation of generalization is called ***specialization***

Generalization schemes



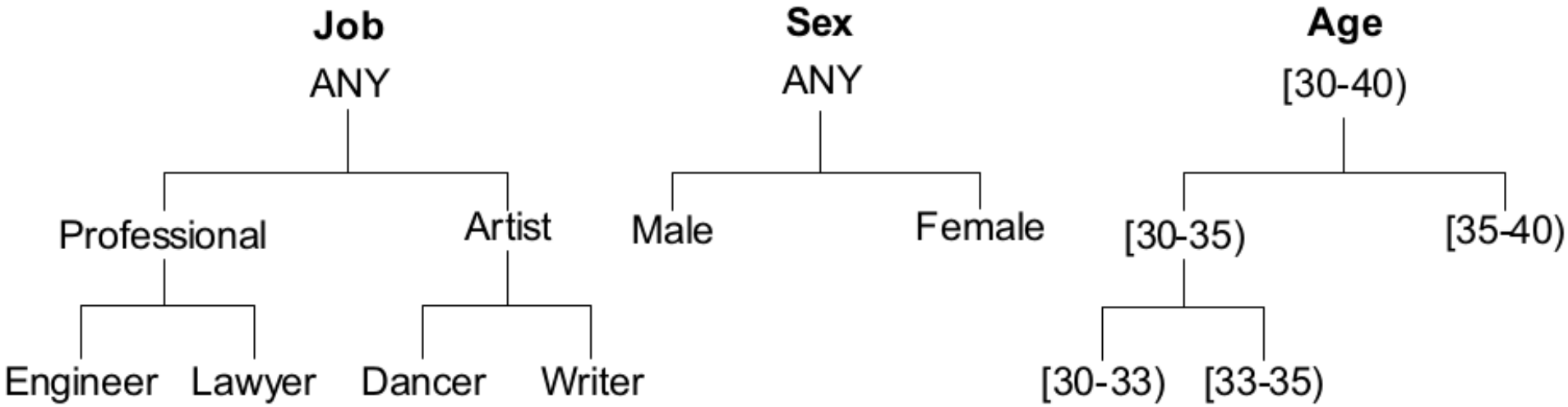
- **Full-Domain Generalization Scheme**: all values in an attribute are generalized to the same level of the taxonomy tree
- **Example:** if ***Lawyer*** and ***Engineer*** are generalized to ***Professional***, then it also requires generalizing ***Dancer*** and ***Writer*** to ***Artist***.

Generalization schemes



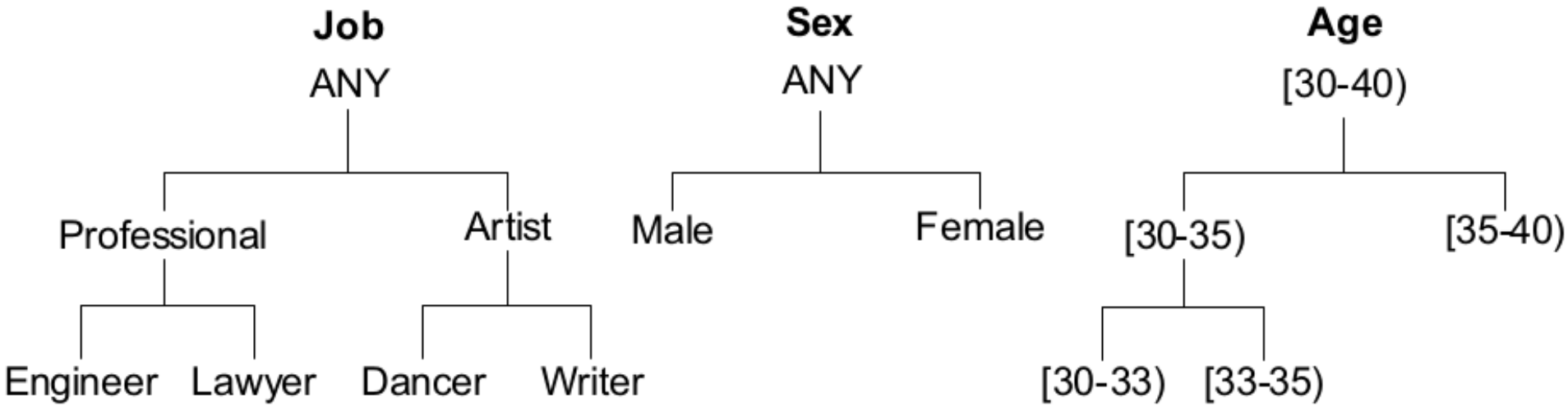
- **Local/Subtree Generalization Scheme**: different levels of generalization may be applied. At a non-leaf node, either all child values or none are generalized.
- **Example**: if ***Engineer*** is generalized to ***Professional***, the other child node, ***Lawyer***, to be generalized to ***Professional***, **but *Dancer*** and ***Writer***, which are child nodes of ***Artist***, can remain **ungeneralized**.

Generalization schemes



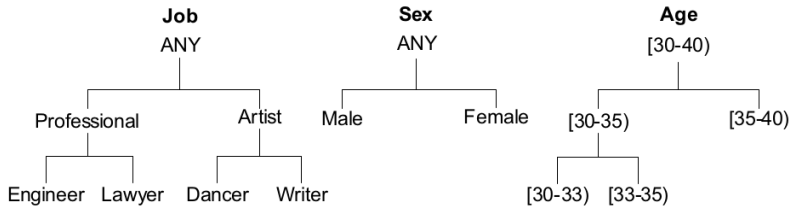
- **Sibling generalization scheme:** This scheme is similar to the subtree generalization, except that some siblings may remain ungeneralized.
- **Example**: if ***Engineer*** is generalized to ***Professional***, ***Lawyer*** can remain ungeneralized.

Generalization schemes



- **Cell generalization scheme:** some instances of a value (some records) may remain ungeneralized while other instances are generalized.
- **Example**: ***Engineer*** in the one record is generalized to ***Professional***, while the ***Engineer*** in another record can remain ungeneralized.

Generalization schemes



- **Multidimensional generalization scheme:**

Let D_i be the domain of an attribute A_i

single-dimensional generalization, for each attribute A_i in ***QID***:

$$f_i : D_{Ai} \rightarrow D'_{Ai}$$

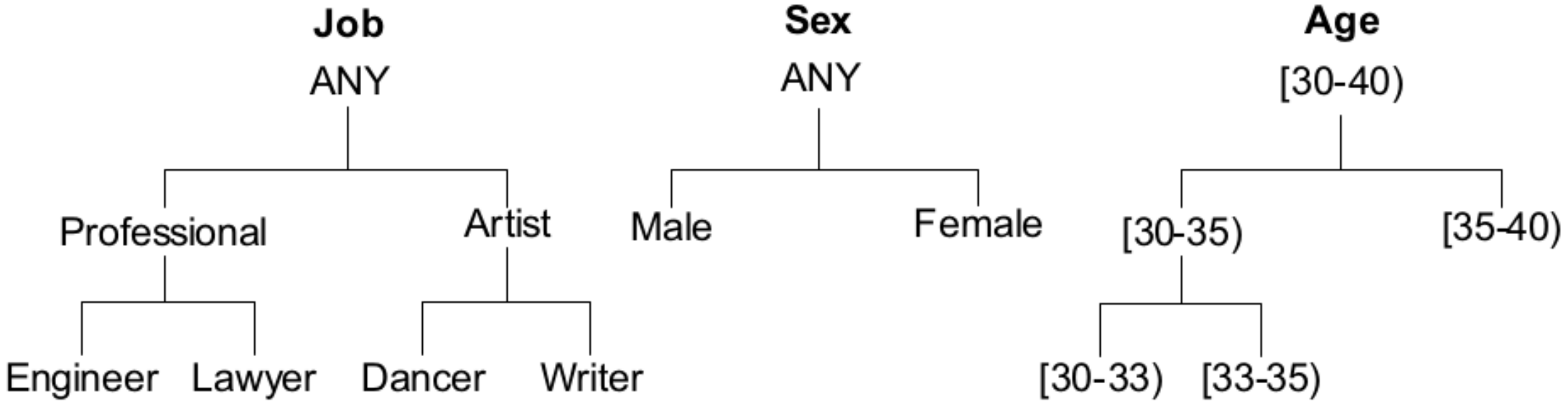
multidimensional generalization, for a **QID:**

$$f : D_{A1} \times \cdots \times D_{An} \rightarrow D' \ (D'_{A1} \times \cdots \times D'_{An})$$

Which is used to generalize $QID = \langle v_1, \ldots, v_n \rangle$ to QID′ = $\langle u_1, \ldots, u_n \rangle$

Where for every v_i, either $v_i = u_i$ or v_i is a child node of u_i in the taxonomy of A_i.

Generalization schemes



- **Multidimensional generalization scheme:**

Example: ⟨Engineer, Male⟩ can be generalized to ⟨Engineer, ANY Sex⟩ while ⟨Engineer, Female⟩ can be generalized to ⟨Professional, Female⟩.

Suppression

- Generalization often leads to too much loss of information
- An alternative could be **Suppression**: remove some attributes/records
- Suppression schemes:
  - **Record Suppression (Row-wise)**: refers to suppressing/removing an entire record (a row)
  - **Value Suppression (Column-wise)**: refers to suppressing/removing all values of an attribute in a table (a column)
  - **Cell Suppression** or **Local Suppression**: refers to suppressing/removing some instances of a given value in a table

Anatomization

- Does not modify the QID or the sensitive attribute, but de-associates the two

How does it work?

1. Data on QID and the data on the sensitive attribute we released in two separate tables:
   - **A quasi-identifier table** (**QIT**) contains the QID attributes
   - **A sensitive table** (**ST**) contains the sensitive attributes
2. Both **QIT** and **ST** will have one common attribute, **GroupID**

Anatomization

- If a group has l distinct sensitive values and each distinct value occurs exactly once in the group, then **the probability of linking a record to a sensitive value** by GroupID is **1/l**.

- This way, the attribute linkage attack can be distorted by increasing l.

Anatomization - Example

Original Data

Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

QID

Sensitive Attribute

intermediate QID-grouped table

	Age	Sex	Disease (sensitive)
1	[30 – 35)	Male	Hepatitis
	[30 – 35)	Male	Hepatitis
	[30 – 35)	Male	HIV
	[30 – 35)	Male	Hepatitis
	[30 – 35)	Male	HIV
	[30 – 35)	Male	HIV
2	[35 – 40)	Female	Flu
	[35 – 40)	Female	Flu
	[35 – 40)	Female	Heart
	[35 – 40)	Female	Heart

These groups are used to define GroupID in QIT and ST tables

Anatomization - Example

QIT Table

Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

ST Table

GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

Group 1 -> number of distinct values = 2

Group 2 -> number of distinct values = 2

The probability of linking a record to a sensitive value -> $1/l = 1/2 = 50\%$

Both QIT and ST satisfy the **privacy requirement with $l \leq 2$** because each group in QIT can be associated with two diseases in ST Table

Anatomization

- Advantage: data in both QIT and ST are unmodified

- Disadvantage: added number of tables
- Disadvantage: with the data published in two tables, it is unclear how standard data mining tools, such as classification, clustering, and association mining tools, can be applied to the published data, and new tools and algorithms need to be designed.

Perturbation

The general idea is to replace the **original data values** with some **synthetic data values** so that the **statistical information computed** from the perturbed data **does not differ significantly** from the statistical information computed from the original data.

Depending on the **degree of randomization**, the perturbed data records may or may not correspond to real-world record owners, so the adversary **cannot perform the sensitive linkage attacks** or recover sensitive information from the published data **with high confidence.**

Perturbation

- Additive Noise: replace the original sensitive value *s* with *s + r* where *r* is a **random value** drawn from some distribution
  - often used for hiding sensitive numerical data (e.g., salary)
- Privacy is measured by determining **how closely the original values of a modified attribute can be estimated**

Perturbation

- Data swapping: swap values of sensitive attributes among individual records while the swaps should maintain the frequency counts for statistical analysis

- Example: **rank swapping**

  - First rank the values of an attribute *A* in ascending order.
  - Then for each value $v \in A$, swap *v* with another value $u \in A$, where u is randomly chosen within a restricted range **p% of v**.
  - Rank swapping can better preserve statistical information than the random data swapping

Perturbation

- Synthetic data generation: generate synthetic data that retains useful statistical information
- Example: condensation
  - Condense the records into multiple groups
  - For each group, extract some statistical information, such as sum and covariance, that suffices to preserve the mean and correlations across the different attributes.
  - Then, based on the statistical information, for publication generate points for each group following the statistical characteristics of the group.

Random Perturbation

Disadvantage: published records are “synthetic”, so that they do not correspond to the real-world entities represented by the original data; therefore, individual records in the perturbed data are basically meaningless to the human recipients.

Privacy Models

- Privacy Models: Used for Privacy Assurance and Privacy Measurement

- **Syntactic models (k-anonymity, l-diversity, t-closeness, …)**
  - Specify syntactic conditions for releasing data
  - Strong assumptions about attack vectors

- Semantic models (differential privacy)
  - Use information on the characteristics of data itself to selectively add noise to the output
  - Much fewer assumptions about attackers

K-anonymity

- **Objective:** To prevent record linkage through QID

- **Definition**: If one record in the table has some value *QID*, at least *k-1* other records also have the value *QID*.

  - These at-least k records are from one **equivalence class**
  - The minimum **equivalence group** size on *QID* is at least *k*

- A table satisfying this requirement is called k-anonymous.

K-anonymity

- In a **k-anonymous table**, each record is indistinguishable from at least **k – 1** other records with respect to QID.

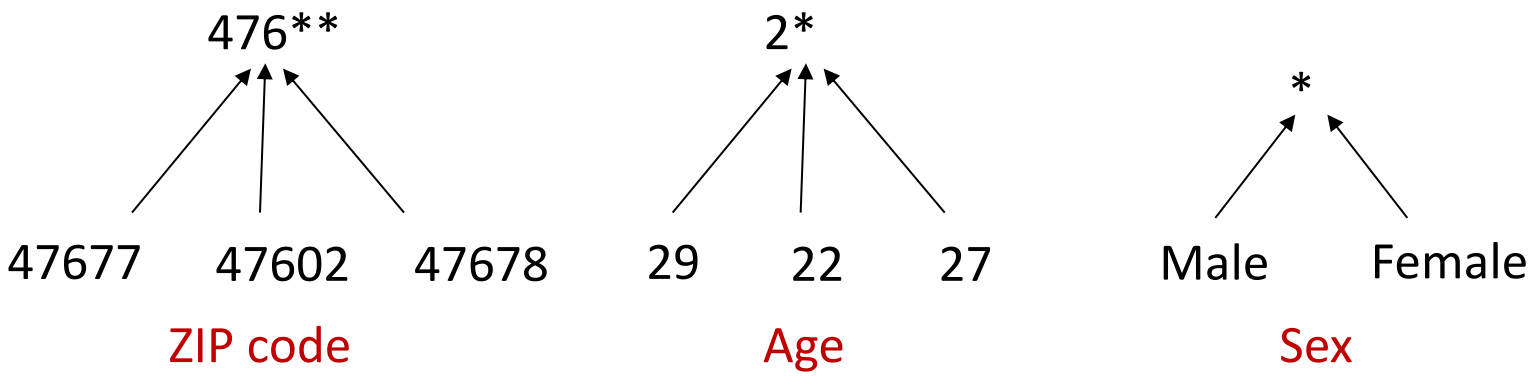
- Consequently, the **probability of linking** a victim to a specific record through QID is at most **1/k**.

- Example: you want to identify an individual based on his birth date and gender. There are k individuals with the same birth date and gender

K-anonymity by Generalization

- Generalization: replace quasi-identifiers with less specific, but semantically consistent values

- **To Achieve K-anonymity using generalization**: Continue replacing specific QID attributes with less specific values until get **k** identical values



K-anonymity Generalization Example

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Generalized table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

!!

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record
- Is that correct?

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

K-anonymity by Suppression

- Suppression:
  - Omit an entire record (row suppression)
    - Common with “outliers” (which are hard to anonymize)
  - Omit an attribute from all individuals (column suppression)
    - Common for explicit identifiers (e.g., name or SSN)

K-anonymity Example

SSN	Zipcode	Age	Disease
012-345-6789	10598	24	HIV
823-627-9231	90210	37	Hepatitis C
987-654-3210	10547	26	HIV
382-827-8264	90345	38	Hepatitis C
847-872-7276	89119	36	Diabetes
422-061-0089	02139	25	HIV

Example of a 3-anonymized version of above table

Row Index	Age	ZIP Code	Disease
1	[20, 30]	Northeastern US	HIV
2	[30, 40]	Western US	Hepatitis C
3	[20, 30]	Northeastern US	HIV
4	[30, 40]	Western US	Hepatitis C
5	[30, 40]	Western US	Diabetes
6	[20, 30]	Northeastern US	HIV

- SSN suppression
- Age and ZIP Code generalization

2 equivalence classes containing 3 records each

Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Cancer
476**	2*	Cancer
476**	2*	Cancer
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Heart Disease
476**	3*	Viral Infection
476**	3*	Viral Infection

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Umeko (japanese)	
Zipcode	***Age***
47653	31

?

Japanese have extremely low incidence of heart disease

L-diversity

- Idea: sensitive attributes must be "diverse" within each quasi-identifier equivalence class
- Principle: each equivalence class has at least L well-represented values

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Umeko	
Zipcode	***Age***
47653	31

A 3-diversity patient table

Zipcode	Age	Disease
4767*	<40	Cancer
4767*	<40	Cancer
4765*	<40	Cancer
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
4767*	<40	Heart Disease
4765*	<40	Heart Disease
4767*	<40	Viral Infection
4765*	<40	Viral Infection

Limitations of L-Diversity

- Example: sensitive attribute is HIV+ **(1%)** or HIV- (99%)
  - Very different degrees of sensitivity!

- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be **at most** 10000***1%=100** (100 HIV+ in the data) equivalence classes

C1	HIV-
C1	HIV-
C1	HIV-
C2	HIV+
C2	HIV-
C2	HIV-
C2	HIV-
C3	HIV+
C3	HIV-
C4	HIV-
C4	HIV-
C4	HIV+

L-diversity Skewness Attack

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
  - Diverse, but potentially violates privacy! Why?
  - **Because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population.**
- Problem: l-diversity does not differentiate:
  - Equivalence class 1: 49 HIV+ and 1 HIV-
  - Equivalence class 2: 1 HIV+ and 49 HIV-

C1	HIV-
C1	HIV-
C1	HIV-
C2	HIV+
C2	HIV-
C2	HIV-
C2	HIV-
C3	HIV+
C3	HIV-
C4	HIV-
C4	HIV-
C4	HIV+

l-diversity does not consider the overall distribution of sensitive values!

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	3K	Gastric Ulcer
476**	2*	4K	Gastritis
476**	2*	5K	Stomach Cancer
4790*	≥40	6K	Gastritis
4790*	≥40	11K	Flu
4790*	≥40	8K	Bronchitis
476**	3*	7K	Bronchitis
476**	3*	9K	Pneumonia
476**	3*	10K	Stomach Cancer

Conclusion

1. Bob's salary is in [3K,5K], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider the semantics of sensitive values!

T-Closeness

- Distribution of the sensitive values in each equivalence class must be "**close**" to the corresponding distribution in the original table
- **"Close"** means upper bounded by a threshold *t*

- ***Q*** = $(q_1, q_2, \ldots, q_m)$ – distribution of values for the SA in the original table
- ***P*** = $(p_1, p_2, \ldots, p_m)$ – distribution of the same attribute in an equivalence class

- This class satisfies t-closeness if: ***Distance*(*P*, *Q*) ≤ *t***

Back to Earlier Example

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	3K	Gastric Ulcer
476**	2*	4K	Gastritis
476**	2*	5K	Stomach Cancer
4790*	≥40	6K	Gastritis
4790*	≥40	11K	Flu
4790*	≥40	8K	Bronchitis
476**	3*	7K	Bronchitis
476**	3*	9K	Pneumonia
476**	3*	10K	Stomach Cancer

Conclusion

1. Bob's salary is in [3K,5K], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider the semantics of sensitive values!

Applying t-Closeness

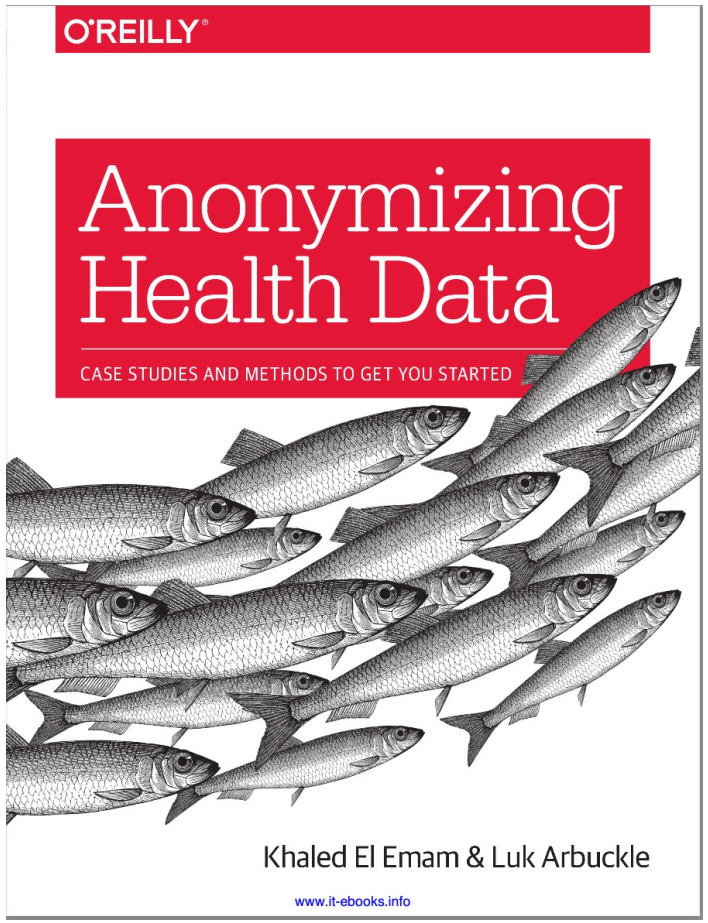
	Zipcode	Age	Salary	Disease
1	4767*	≤40	3K	Gastric Ulcer
3	4767*	≤40	5K	Stomach Cancer
8	4767*	≤40	9K	Pneumonia
4	4790*	≥40	6K	Gastritis
5	4790*	≥40	11K	Flu
6	4790*	≥40	8K	Bronchitis
2	4760*	≤40	4K	Gastritis
7	4760*	≤40	7K	Bronchitis
9	4760*	≤40	10K	Stomach Cancer

- No longer possible to infer Bob's:
  - Low salary
  - Stomach-related disease

Bob	
Zip	***Age***
47678	27

Bibliography

Chapter 2



Chapter 4