Security and Privacy

Privacy Preserving Data Publishing (PPDP)

De-identification

- **De-identification** refers to the process of removing or masking personally identifiable information (PII) from a dataset
  - E.g., names, addresses, dates of birth, social security numbers, and other information that can be used to identify individuals.

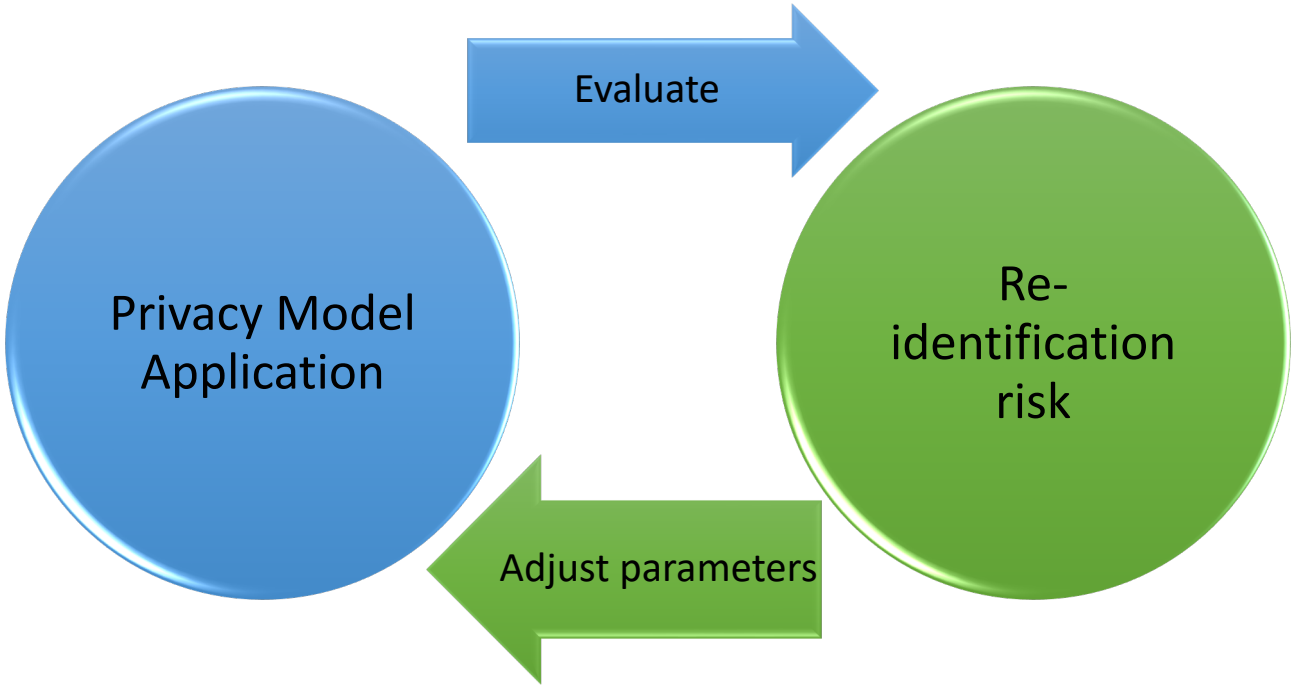
- The purpose of de-identification is to **protect individuals' privacy** and **reduce the risk** of **re-identification**, while still allowing the data to be used for research or other purposes.

Re-identification

- **Re-identification** refers to the process of **matching de-identified data** with **external information sources** in order to re-identify individuals whose identities were supposed to be protected by the de-identification process.

De-identification Process

- De-identification is often an iterative process



Re-identification Risk

- **Re-identification risk** is the likelihood that de-identified data can be re-identified and linked back to an individual or group of individuals.
- Re-identification risk arises when **there is still enough information in the de-identified dataset** to allow someone to re-identify individuals.

Re-identification Risk

- Having some way to **measure risk** allows us
  - to decide **whether it's too high**,
  - **how much de-identification needs to be applied** to a data set to reduce the risk

- Question: Is it possible to *quantify the risk of re-identification?*

Re-identification Risk

- To calculate the re-identification risk, it is required determining QIDs Distinction and Separation

  - **Distinction**: degree to which **variables** **make records distinct**

  - **Separation**: degree to which **combinations of variables** **separate the records**

Distinction - Definitions

- **Key/identifying key** is a subset of attributes that uniquely identifies each tuple/register in a table (e.g., name and address together)

- **α-distinct QID**: subset of attributes which becomes a **key** in the table after the removal of at most a **1−α fraction of tuples** in the original table (α = [0,1])

Quasi-Identifiers (QIDs) Distinction

- **Age** is a **0.6-distinct attribute** in this dataset. What does this mean?

- It means if we remove 1 - 0.6 = 0.4 fraction of data (**40%**) from the dataset, age becomes an **identifying key**

- **Let's verify …**
  - 40% x 5 records = 2 records
  - Let's remove 2 records with duplicate age value

	age	sex	state
2	30	Female	CA
4	20	Male	NY
5	40	Male	CA

Age became a Key

Quasi-Identifiers (QIDs) Distinction

- **Age** is a **0.6-distinct attribute** in this dataset. How do we calculate **α (0.6)?**
- Simply: as it has 3 distinct values in a total of 5 values (3/5 = 0.6)

	age	sex	state
1	20	Female	CA
2	30	Female	CA
3	40	Female	TX
4	20	Male	NY
5	40	Male	CA

QID = (age)

Set of all values = {**20**,**30**,**40**,**20**,**40**}, size = **5**

Set of distinct values = {**20**, **30**, **40**}, size = **3**

Distinction ratio = #distinct values / #all values

= **3/5**

= **0.6**

Quasi-Identifiers (QIDs) Distinction

- **(Sex, State)** is a **0.8-distinct attributes** in this dataset
- As it has 4 distinct values in a total of 5 values (4/5 = 0.8)

	age	sex	state
1	20	Female	CA
2	30	Female	CA
3	40	Female	TX
4	20	Male	NY
5	40	Male	CA

QID = (sex, state)

Set of all values = {**(F,CA)**,**(F,CA)**,**(F,TX)**,**(M,NY)**,**(M,CA)**}, size = **5**

Set of distinct values = {**(F,CA)**, **(F,TX)**, **(M,NY)**, **(M,CA)**}, size = **4**

Distinction ratio = #distinct values / #all values

= **4/5**

= **0.8**

Higher values indicate probable QID!

Separation definition

- Degree to which, **combinations of variables** separate the records

- A subset of attributes **separates** a pair of tuples **x** and **y** if **x** and **y** have different values on at least one attribute in the subset
  - We need to see **how many combinations of values** do we have
  - Then we need to **how many combinations have the same values in the tuples**

  - It somehow shows **how probable is to re-identify a record through linkage**

Quasi-Identifiers (QIDs) Separation

- **Age** is a **0.8-separation attribute** in this dataset
- As there are 10 distinct pairs of tuples and 8 pairs can be separated by age (8/10 = 0.8)

	age	sex	state
1	20	Female	CA
2	30	Female	CA
3	40	Female	TX
4	20	Male	NY
5	40	Male	CA

QID = (age)

Set of all tuples =

{(**20**,**30**),(**20**,**40**),(**20**,**20**),(**20**,**40**),(**30**,**40**), (**30**,**20**), (**30**,**40**), (**40**,**20**), (**40**,**40**), (**20**,**40**)}, size = **10**

Set of separated tuples = {(**20**,**30**),(**20**,**40**),(**20**,**40**),(**30**,**40**), (**30**,**20**), (**30**,**40**), (**40**,**20**), (**20**,**40**)}, size = **8**

(20,20) and (40,40) are removed from the first set

Separation ratio = #separated tuples / #all tuples

= **8/10**

= **0.8**

Quasi-Identifiers (QIDs) Separation

- **(Sex, State)** is a **0.9-seperation attributes** in this dataset
- As it has 9 distinct values in a total of 10 values (9/10 = 0.8)

	age	sex	state
1	20	Female	CA
2	30	Female	CA
3	40	Female	TX
4	20	Male	NY
5	40	Male	CA

QID = (sex, state)

Set of all tuples = {**((F,CA),(F,CA))**, **((F,CA),(F,TX))**, **((F,CA),(M,NY))**, **((F,CA),(M,CA))**, **((F,CA),(F,TX))**, **((F,CA),(M,NY))**, **((F,CA),(M,CA))**, **((F,TX),(M,NY))**, **((F,TX),(M,CA))**, **((M,NY),(M,CA))**}, size = **10**

Set of distinct values = {**((F,CA),(F,TX))**, **((F,CA),(M,NY))**, **((F,CA),(M,CA))**, **((F,CA),(F,TX))**, **((F,CA),(M,NY))**, **((F,CA),(M,CA))**, **((F,TX),(M,NY))**, **((F,TX),(M,CA))**, **((M,NY),(M,CA))**}, size = **9**

((F,CA),(F,CA)) is removed from the first set

Distinction ratio = #distinct values / #all values

= **9/10**

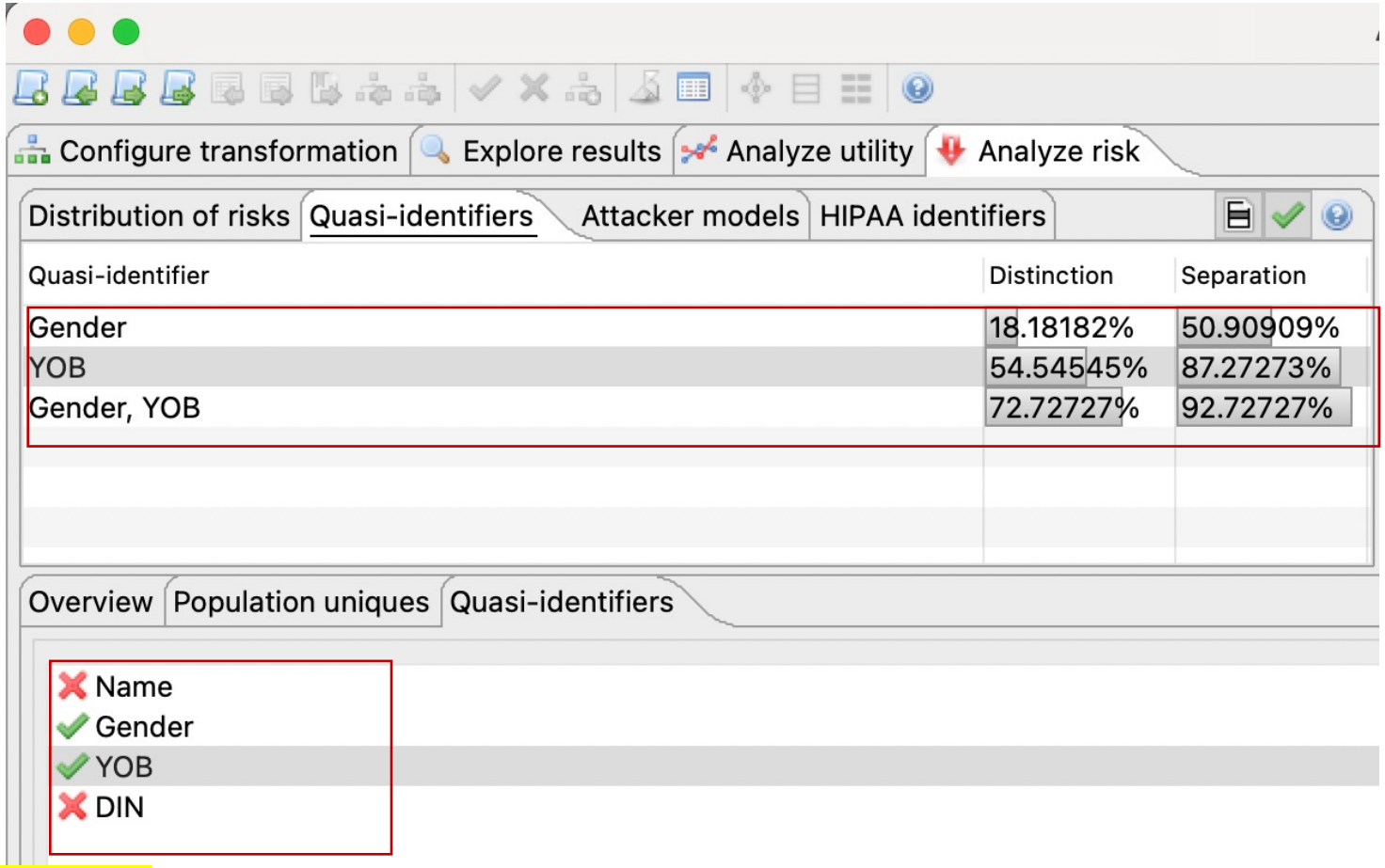
= **0.9**

QIDs Distinction and Separation in ARX

Dataset

Input data

	Name	Gender	YOB	DIN
1	Gill Stringer	F	1995	2046059
2	Freda Shields	F	1995	596612
3	John Smith	M	1979	2046059
4	Hercules Green	M	1979	2241497
5	Douglas Henry	M	1979	544981
6	Alice Smith	F	1987	392537
7	Beverly McCu...	F	1984	293512
8	Fred Thompson	M	1987	725765
9	Alan Patel	M	1982	716839
10	Bill Nash	M	1995	363766
11	Albert Blackwell	M	1998	544981



Higher values indicate probable QID!

Measuring Re-Identification Risk - Notations

- **D** = Original Dataset (full)
  - **N** - number of records in **D**
  - **K** - set of equivalence classes in **D** (distinct QIDs)
  - F_j - number of records in equivalence class ***j***
  - $N = \sum_{j \in K} F_j \quad (sum\ of\ all\ records\ of\ all\ equivalence\ classes)$

- **U** = Disclosed Dataset **U ⊆ D**
  - **n** - number of records in **U**
  - **k** - set of equivalence classes in **U** (distinct QIDs) **k ⊆ K**
  - f_j - number of records in equivalence class ***j***

Measuring Re-identification Risk - Assumptions

- What we really can do is just an estimate under **certain assumptions**
- The assumptions concern **data quality** and the **type of attack (attack target, attacker model)** that an adversary will likely launch on a data set.
  - **Data quality**: assuming ideal conditions about **data quality for the dataset** itself and **the information that an adversary** would use to attack the dataset

Data quality: refers to the degree to which data meets the requirements of its intended use, such as **accuracy**, **completeness**, **consistency**, and **timeliness**.

This assumption, although unrealistic, actually results in **conservative estimates of risk** (i.e., setting the risk **estimate a bit higher than it probably is**) because the better the data is, the more likely it is that an adversary will successfully re-identify someone.

Measuring Re-Identification Risk - Assumptions

- **Attack Target:** individual that is the target of re-identification
  - **Specific individual**
    - E.g., neighbor, co-worker, ex-spouse, relative, or famous person
  - **Individual selected at random**
    - E.g., Journalist wishes to embarrass or expose **data holder** (any record will do)
  - **As many individuals as possible**
    - E.g., Intruder wants to market a product to all of the individuals in the disclosed database

Measuring Re-Identification Risk - Assumptions

Attacker models:

- **Prosecutor scenario**
  - Targets **one** specific individual
  - Adversary knows whether target individual is in the dataset
- **Journalist scenario**
  - Targets **any** individual
  - Adversary selects a target at random because the re-identification of any record will achieve the purpose
- **Marketer scenario**
  - Targets **as many** individuals as possible
  - An attack is considered successful if a large portion of the records can be re-identified

How can the adversary know if the target is in the dataset?

- Three ways:
  - Dataset represents the whole population (U=D)
    - E.g., All citizens' records
  - Dataset is not a population registry but is a known sample from a population (e.g., teenagers)
  - Individuals in the dataset self-reveal that they are part of the sample
    - E.g., Public release of a clinical trials dataset: attack targets generally inform their family or friends that they are participating/participated in a trial.

Measuring Re-Identification Risk: Prosecutor Model

- Attacker aims to re-identify **Alice**
  - Data holder does not know in advance that Alice is the target
  - Must calculate risk for all equivalence classes

N = 11, **QID** = (**Gender**, **Year of Birth**)

Equivalence classes (**K**)

1. (Male, 1979) F_1 = 3
2. (Male, 1982) F_2 = 1
3. (Female, 1995) F_3 = 2
4. (Female, 1987) F_4 = 1
5. (Male, 1995) F_5 =1
6. (Male, 1998) F_6 =1
7. (Female, 1984) F_7 =1
8. (Male, 1987) F_8 =1

Original Database

IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
Name	Gender	Year of Birth	DIN
John Smith	Male	1979	2046059
Alan Patel	Male	1982	716839
Hercules Green	Male	1979	2241497
Gill Stringer	Female	1995	2046059
Alice Smith	Female	1987	392537
Bill Nash	Male	1995	363766
Albert Blackwell	Male	1998	544981
Beverly McCulsky	Female	1984	293512
Douglas Henry	Male	1979	544981
Freda Shields	Female	1995	596612
Fred Thompson	Male	1987	725765

Measuring Re-Identification Risk: Prosecutor Model

In general, $\mathbf{P_{Prosecutor}} = \frac{1}{F_j}$

Where F_j is size of the equivalence class J

- Lowest $\mathbf{P_{Prosecutor}}$ = 1/3 ≈ 0.33

For QID_1 = (Male, 1979), F_1 = 3

- Highest $\mathbf{P_{Prosecutor}}$ = 1/1 = 1

For QID_2 = (Male, 1982), F_2 = 1

- Average $\mathbf{P_{Prosecutor}}$ = size(K)/ N = **8/11 ≈ 0.73**

Original Database

IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
Name	Gender	Year of Birth	DIN
John Smith	Male	1979	2046059
Alan Patel	Male	1982	716839
Hercules Green	Male	1979	2241497
Gill Stringer	Female	1995	2046059
Alice Smith	Female	1987	392537
Bill Nash	Male	1995	363766
Albert Blackwell	Male	1998	544981
Beverly McCulsky	Female	1984	293512
Douglas Henry	Male	1979	544981
Freda Shields	Female	1995	596612
Fred Thompson	Male	1987	725765

Measuring Re-Identification Risk: Prosecutor Model

- After De-Identification (Anonymization)

Original Database

IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
Name	Gender	Year of Birth	DIN
John Smith	Male	1979	2046059
Alan Patel	Male	1982	716839
Hercules Green	Male	1979	2241497
Gill Stringer	Female	1995	2046059
Alice Smith	Female	1987	392537
Bill Nash	Male	1995	363766
Albert Blackwell	Male	1998	544981
Beverly McCulsky	Female	1984	293512
Douglas Henry	Male	1979	544981
Freda Shields	Female	1995	596612
Fred Thompson	Male	1987	725765

De-Identification

QUASI-IDENTIFIERS		
Gender	Decade of Birth	DIN
Male	1970-1979	2046059
Male	1980-1989	716839
Male	1970-1979	2241497
Female	1990-1999	2046059
Female	**1980-1989**	**392537**
Male	1990-1999	363766
Male	1990-1999	544981
Female	**1980-1989**	**293512**
Male	1970-1979	544981
Female	1990-1999	596612
Male	1980-1989	725765

Disclosed File

Measuring Re-Identification Risk: Prosecutor Model: After De-Identification (Anonymization)

- Data holder does not know in advance that the adversary will target Alice

- Thus, needs to compute the value of $P_{Prosecutor}$ for all of the equivalence classes

QUASI-IDENTIFIERS		
Gender	**Decade of Birth**	**DIN**
Male	1970-1979	2046059
Male	1980-1989	716839
Male	1970-1979	2241497
Female	1990-1999	2046059
Female	**1980-1989**	**392537**
Male	1990-1999	363766
Male	1990-1999	544981
Female	**1980-1989**	**293512**
Male	1970-1979	544981
Female	1990-1999	596612
Male	1980-1989	725765

Disclosed File

Measuring Re-Identification Risk: Prosecutor Model: After De-Identification (Anonymization)

N = 11, QID = (**Gender**, **Year of Birth**)

Equivalence classes (K)

1. (Male, 1970 - 1979) F_1 = 3, $P_{prosecutor}$ = **1/3 ≈ 0.33**
2. (Male, 1980 - 1989) F_2 = 2, $P_{prosecutor}$ = **1/2 ≈ 0.5**
3. (Female, 1990 - 1999) F_3 = 2, $P_{prosecutor}$ = **1/2 ≈ 0.5**
4. (Female, 1980 - 1989) F_4 = 2, $P_{prosecutor}$ = **1/2 ≈ 0.5**
5. (Male, 1990 - 1999) F_5 = 2, $P_{prosecutor}$ = **1/2 ≈ 0.5**

- Lowest $P_{Prosecutor}$ = **1/3 ≈ 0.33**

For QID_1 = (Male, 1970 - 1979), F_1 = 3

- Highest $P_{Prosecutor}$ = **1/2 = 0.50**

For QID_2 = (Male, 1982), F_2 = 2

Average $P_{Prosecutor}$ = size(K)/ N = **5/11 ≈ 0.45**

QUASI-IDENTIFIERS		
Gender	**Decade of Birth**	**DIN**
Male	1970-1979	2046059
Male	1980-1989	716839
Male	1970-1979	2241497
Female	1990-1999	2046059
Female	**1980-1989**	**392537**
Male	1990-1999	363766
Male	1990-1999	544981
Female	**1980-1989**	**293512**
Male	1970-1979	544981
Female	1990-1999	596612
Male	1980-1989	725765

Disclosed File

Measuring Re-Identification Risk: Prosecutor Model: After De-Identification (Anonymization)

- Attacker who aims to re-identify Alice, knows Alice YOB and Gender

  QID = (**Female**, **1987**)

  - Probability of re-identificationin prosecutor model:

$$F_{(Female,\ 1987)} = 2$$

$$P_{Prosecutor} = 1/2 = 0.5$$

QUASI-IDENTIFIERS		
Gender	**Decade of Birth**	**DIN**
Male	1970-1979	2046059
Male	1980-1989	716839
Male	1970-1979	2241497
Female	1990-1999	2046059
Female	**1980-1989**	**392537**
Male	1990-1999	363766
Male	1990-1999	544981
Female	**1980-1989**	**293512**
Male	1970-1979	544981
Female	1990-1999	596612
Male	1980-1989	725765

Disclosed File

Measuring Re-Identification Risk: Journalist Model

- Attacker targets any (arbitrary) individual

- Smart attacker: focus on smaller equivalence classes (highest probability of re-identification)

Measuring Re-Identification Risk: Journalist Model

- Focus on smaller equivalence classes:

$$P_{journalist} = \frac{1}{Min\ (F_j)}$$

where F_j *is the size of equivalence class j in* ***D***

Thus,

$P_{journalist}$ = Highest $P_{Prosecutor}$

Measuring Re-Identification Risk: Marketer Model

- Attacker wishes to re-identify as many records as possible

- Marketer risk =

  = expected proportion of records correctly re-identified

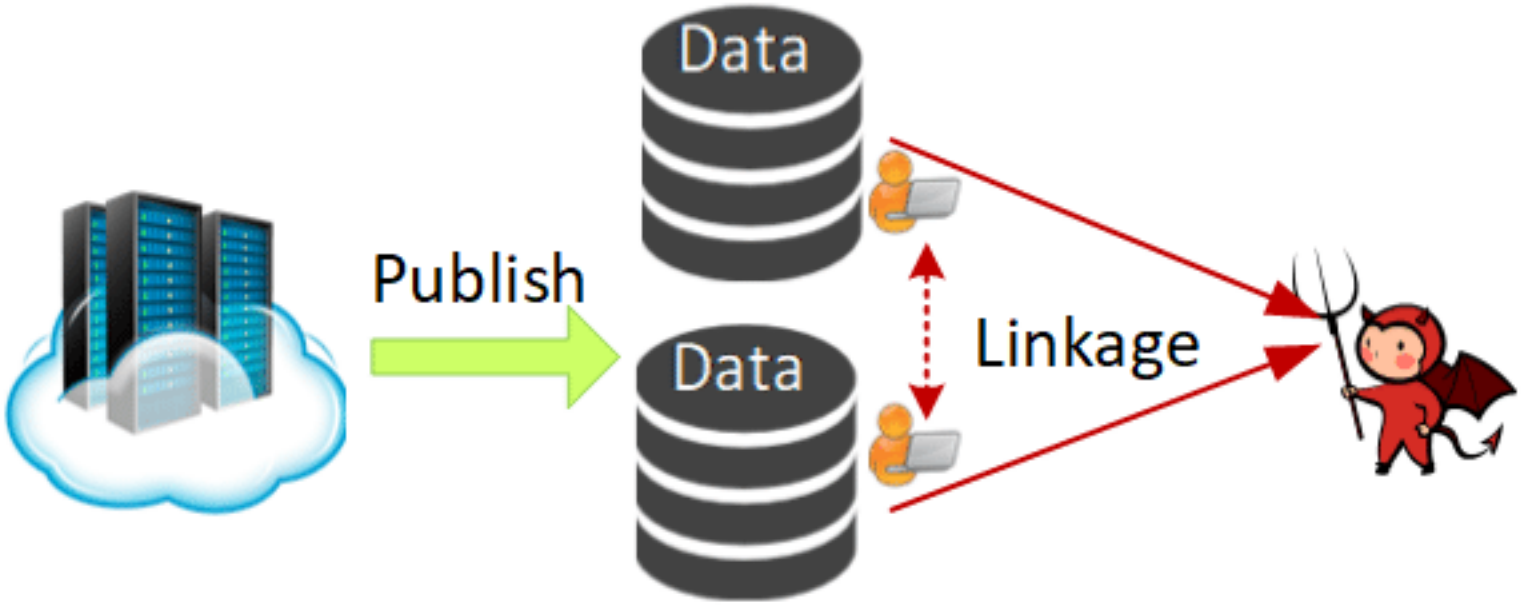
  = average probability of re-identification

- Thus,

$$p_{marketer} = \text{Average } P_{Prosecutor}$$

To remember ….

To remember …. Linkage Attack



Linkage Attack: Adversary acquires private information by correlating multiple datasets

To remember …. Linkage Attack

- Linkage attack occurs when an adversary is able to link a record owner 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.

Privacy Models

- We can broadly classify privacy models into two categories based on their attack principles.
  1. A privacy threat occurs when an adversary is able to link a record owner to a record in a published data table, to a sensitive attribute in a published data table, or to the published data table itself (e.g., **record linkage**, **attribute linkage**, and **table linkage**)

  2. The published table provide the adversary with **little additional information beyond the background knowledge**.
     - If the adversary has a large variation between the prior and posterior beliefs, we call it the **probabilistic attack.**

Privacy Models

Each Privacy Model is suitable to protect the data against a particular attack models

Privacy Model	**Attack Model**			
	Record linkage	Attribute linkage	Table linkage	Probabilistic attack
k-Anonymity [201, 217]	✓			
MultiR k-Anonymity [178]	✓			
ℓ-Diversity [162]	✓	✓		
Confidence Bounding [237]		✓		
(α, k)-Anonymity [246]	✓	✓		
(X, Y)-Privacy [236]	✓	✓		
(k, e)-Anonymity [269]		✓		
(ϵ, m)-Anonymity [152]		✓		
Personalized Privacy [250]		✓		
t-Closeness [153]		✓		✓
δ-Presence [176]			✓	
(c, t)-Isolation [46]	✓			✓
ϵ-Differential Privacy [74]			✓	✓
(d, γ)-Privacy [193]			✓	✓
Distributional Privacy [33]			✓	✓

To remember …. K-anonymity

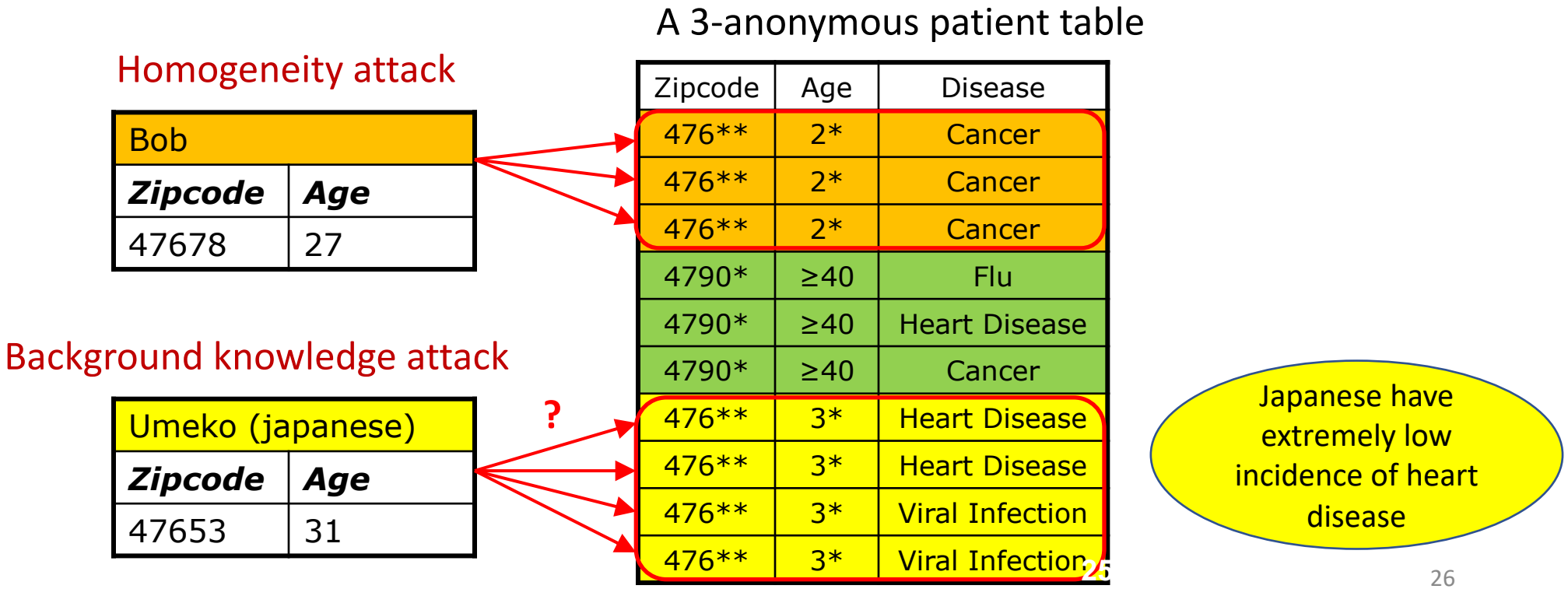
- In a k-anonymous table, each record is **indistinguishable** from **at least k – 1** other records with respect to **QID**.

- 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 in the table

To remember …. Attacks on k-Anonymity

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



k-Anonymity – One more limitation

- It has been shown in research that when the **number of QID attributes is large**, that is, when **the dimensionality of data is high**, most of the data **have to be suppressed** in order to achieve k-anonymity.

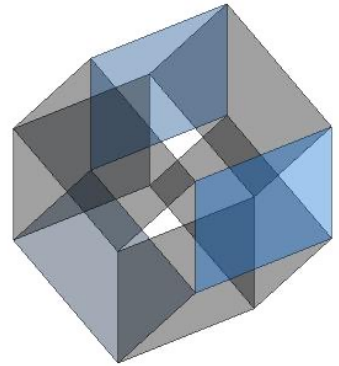
[Aggarwal VLDB '05]

- Applying k-anonymity on the high-dimensional data would significantly **degrade** the **data quality and utility**.

- This problem is known as the **curse of high-dimensionality** on **k-anonymity**

k-Anonymity & Curse of High-Dimensionality

- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
    - The closest neighbors may have many attributes that are **just too different to agregate**, leading to **gross generalizations** and rather **high loss of utility**

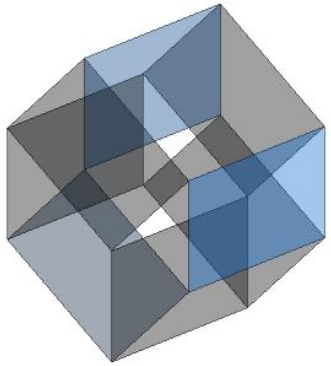


k-Anonymity & Curse of High-Dimensionality

- In many realistic scenario, **some sensitive attributes** may be known to adversary.

- The **boundary between quasi-identifiers and sensitive attributes** becomes unclear.
  - Example: An attribute such as **salary** may be both a quasi-identifier as well as a sensitive attribute

- In such cases, **most or all sensitive attributes may need to be included in the anonymization process** as quasi-identifiers.



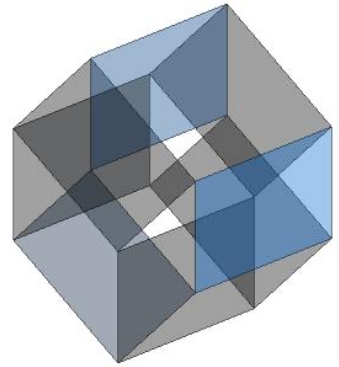
k-Anonymity & Curse of High-Dimensionality

- Example: Consider the following 2-dimensional records on (**Age, Salary**) => (**26, 94000**) and (**29, 97000**).

- Generalization solution: **age** is generalized to the range **25-30**, and **salary** is generalized to the range **90000-100000**,
  - two record cannot be distinguished from one another.

- What about when we have a large number of attributes? Can we find a generalization solution?
  - The problem of finding optimal k-anonymization in high dimensional datasets is **NP-hard**.

- In this case k-anonymized datasets are useless



LKC-Privacy Model - Assumption

- **Limited prior knowledge of adversary**
  - in real-life privacy attacks, it is very difficult for an adversary to acquire all the information in QID of a target victim
  - Thus, it is reasonable to assume that the adversary's prior knowledge is bounded by **at most *L* values of the QID attributes** of the target victim.

Adversary's prior knowledge does not exceed **L** values

LKC-Privacy Model - General Intuition

Objective of the privacy model:

- To ensure that every combination of values in $\mathrm{QID}' \subseteq \mathrm{QID}$ with maximum length **L** in the data table **T** is Shared by at least **K** records
  - Record-linkage probability <= 1/K

- To ensure that the confidence of inferring any sensitive values in **S** <= **C**
  - Attribute-linkage probability <= C

L, K, C are thresholds and S is a set of sensitive attributes

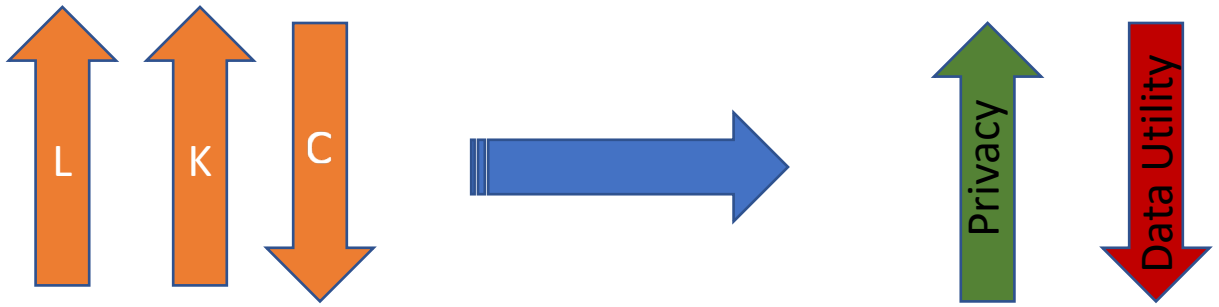
LKC-Privacy and High-dimensional Data

- Suitable for anonymizing high-dimensional data:
  - Requires only a subset of QID attributes to be shared by at least K records (relaxation of k-anonimity)

- General privacy model
  - Generalizes several traditional privacy models
    - **k-anonymity: special case of LKC-privacy with L = |QID|, K = k, and C = 100%**

LKC-Privacy Model

- Adjustable trade-off between privacy and utility:
  - Increasing L and K, or decreasing C improves privacy at the expense of data utility loss



LKC-Privacy Model

DEFINITION of ***LKC-privacy***: Let L be the maximum number of values of the prior knowledge. Let S be a set of sensitive values. A data table T satisfies *LKC-privacy* if and only if for any qid with $|qid| \leq L$,

1. $|T[qid]| \geq K$, where $K > 0$ is an integer anonymity threshold, where $T[qid]$ denotes the set of records containing qid in T, and

2. $conf(qid \rightarrow s) \leq C$ for any $s \in S$, where $0 < C \leq 1$ is a real number confidence threshold. ■

- The **data holder** specifies the thresholds L, K, and C.
- The maximum length **L** reflects the assumption of the **adversary's power**.

LKC-Privacy Model - Example

- **Purpose of releasing the dataset:**

classification analysis on the **class** attribute, ***Transfuse***, which has two values, Y or N, indicating whether or not the patient has received blood transfusion.

- **Sensitive data**: Only sensitive value in Surgery is **Transgender**

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Janitor	M	34	Y	Transgender
2	Doctor	M	58	N	Plastic
3	Mover	M	34	Y	Transgender
4	Lawyer	M	24	N	Vascular
5	Mover	M	58	N	Urology
6	Janitor	M	44	Y	Plastic
7	Doctor	M	24	N	Urology
8	Lawyer	F	58	N	Plastic
9	Doctor	F	44	N	Vascular
10	Carpenter	F	63	Y	Vascular
11	Technician	F	63	Y	Plastic

LKC-Privacy Model - Example

Possible Privacy threats:

- **Record linkage** to record #3 via qid = **⟨Mover,34⟩**
- **Attribute linkage via qid = ⟨M, 34⟩**

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Janitor	M	34	Y	Transgender
2	Doctor	M	58	N	Plastic
3	Mover	M	34	Y	Transgender
4	Lawyer	M	24	N	Vascular
5	Mover	M	58	N	Urology
6	Janitor	M	44	Y	Plastic
7	Doctor	M	24	N	Urology
8	Lawyer	F	58	N	Plastic
9	Doctor	F	44	N	Vascular
10	Carpenter	F	63	Y	Vascular
11	Technician	F	63	Y	Plastic

LKC-Privacy Model - Example

- Goal – transform a given dataset T into an anonymous version T' that:
  - Satisfies a given LKC-privacy requirement and
  - Preserves as much information as possible for the intended data analysis task

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Janitor	M	34	Y	Transgender
2	Doctor	M	58	N	Plastic
3	Mover	M	34	Y	Transgender
4	Lawyer	M	24	N	Vascular
5	Mover	M	58	N	Urology
6	Janitor	M	44	Y	Plastic
7	Doctor	M	24	N	Urology
8	Lawyer	F	58	N	Plastic
9	Doctor	F	44	N	Vascular
10	Carpenter	F	63	Y	Vascular
11	Technician	F	63	Y	Plastic

LKC-Privacy Parameters - Example

- Set goal Before Anonymization:
  - Every possible QID of max. length **L=2**

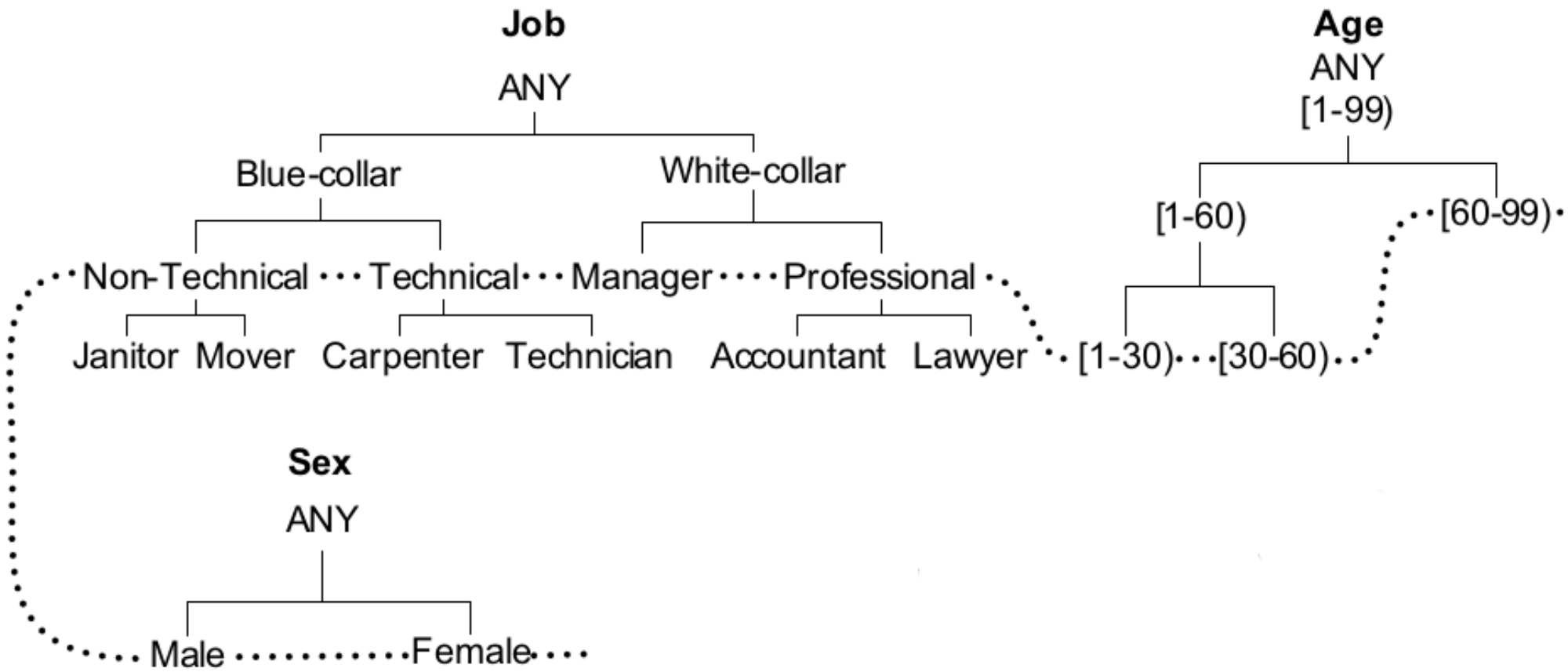
QID_1 = {Job, Sex}
QID_2 = {Job, Age}
QID_3 = {Sex, Age}

  - Shared by at least **K=2** records
  - Confidence in inferring Transgender <= 50% (**C = 50%**)

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Janitor	M	34	Y	Transgender
2	Doctor	M	58	N	Plastic
3	Mover	M	34	Y	Transgender
4	Lawyer	M	24	N	Vascular
5	Mover	M	58	N	Urology
6	Janitor	M	44	Y	Plastic
7	Doctor	M	24	N	Urology
8	Lawyer	F	58	N	Plastic
9	Doctor	F	44	N	Vascular
10	Carpenter	F	63	Y	Vascular
11	Technician	F	63	Y	Plastic

Privacy

LKC-Privacy Taxonomy Tree - Example



LKC-Privacy - Example

- Check the goal After Anonymization
  - Every possible QID of max. length L=2
  - Shared by at least K=2 records
  - Confidence in inferring Transgender <= 50% (C = 50%)

Table 6.2: Anonymous data ($L = 2$, $K = 2$, $C = 50\%$)

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Non-Technical	M	[30 – 60)	Y	Transgender
2	Professional	M	[30 – 60)	N	Plastic
3	Non-Technical	M	[30 – 60)	Y	Transgender
4	Professional	M	[1 – 30)	N	Vascular
5	Non-Technical	M	[30 – 60)	N	Urology
6	Non-Technical	M	[30 – 60)	Y	Plastic
7	Professional	M	[1 – 30)	N	Urology
8	Professional	F	[30 – 60)	N	Plastic
9	Professional	F	[30 – 60)	N	Vascular
10	Technical	F	[60 – 99)	Y	Vascular
11	Technical	F	[60 – 99)	Y	Plastic

We have to check for different combination of two attributes in QID

QID_1 = {Job, Sex}
QID_2 = {Job, Age}
QID_3 = {Sex, Age}

LKC-Privacy Versus K-anonymity

- K-anonymity (k=2) would require further generalization
  - <Professional, M, [30-60)> requires generalizing as it is unique in the dataset

  [1-30) and [30-60) => [1-60)

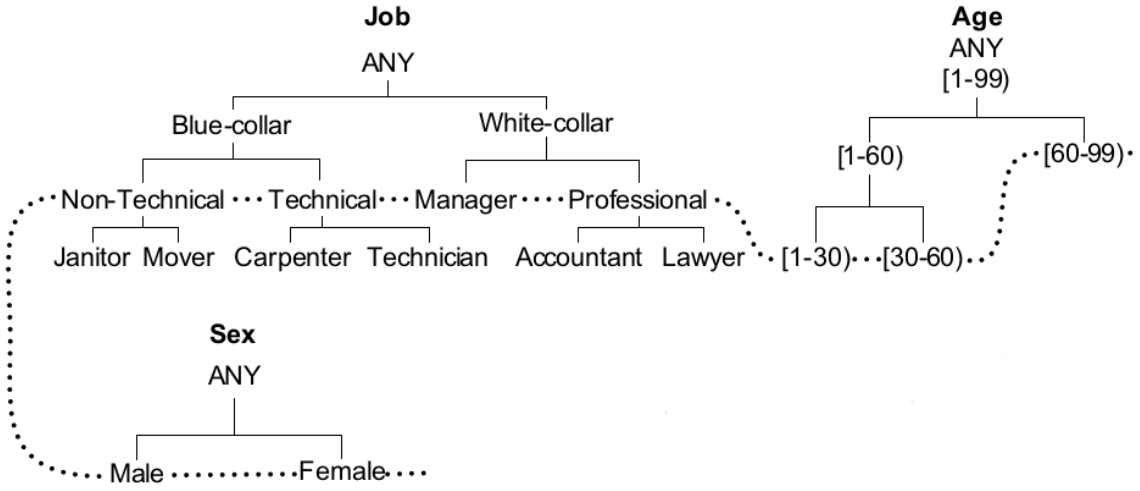
  - Higher utility loss

Table 6.2: Anonymous data ($L = 2$, $K = 2$, $C = 50\%$)

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Non-Technical	M	[30 – 60)	Y	Transgender
2	Professional	M	[30 – 60)	N	Plastic
3	Non-Technical	M	[30 – 60)	Y	Transgender
4	Professional	M	[1 – 30)	N	Vascular
5	Non-Technical	M	[30 – 60)	N	Urology
6	Non-Technical	M	[30 – 60)	Y	Plastic
7	Professional	M	[1 – 30)	N	Urology
8	Professional	F	[30 – 60)	N	Plastic
9	Professional	F	[30 – 60)	N	Vascular
10	Technical	F	[60 – 99)	Y	Vascular
11	Technical	F	[60 – 99)	Y	Plastic

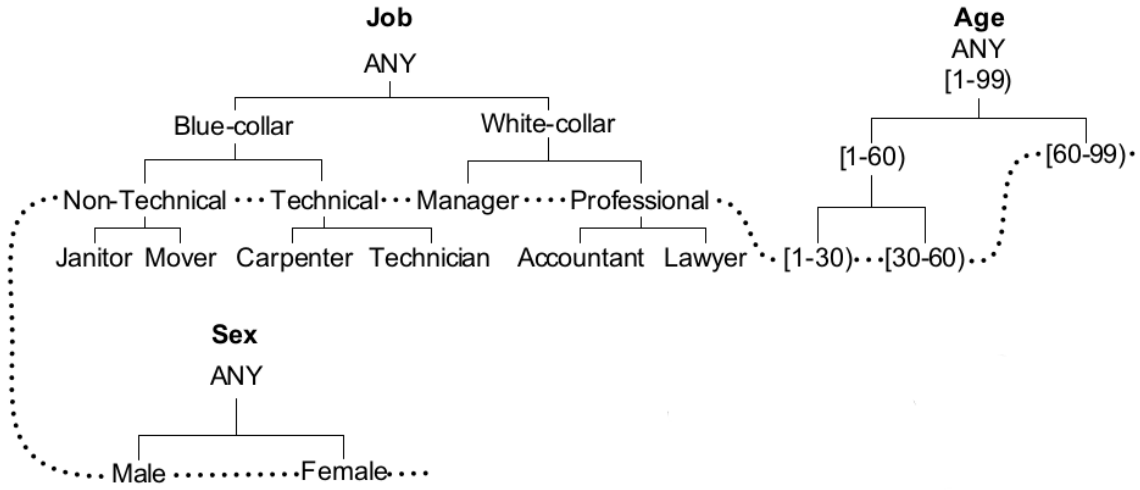
How to Achieve LKC-Privacy

- **Top-down specialization algorithm**
  - Achieves LKC-privacy by **subtree generalization**
- Idea: anonymize a table by **a sequence of specializations** starting from **the topmost general state** in which **each attribute has the topmost value of its taxonomy tree**



How to Achieve LKC-Privacy

- For **categorical attributes**: we assume that a taxonomy tree is specified for each categorical attribute in QID.
  - A leaf node represents a **domain value** and a parent node represents a **less specific value**.
- For **numerical attributes**: a taxonomy tree can be grown at runtime.
  - Each node represents an interval, and each non-leaf node has two child nodes representing some optimal binary split of the parent interval.



Problem Statement

- Objectives of data holder and adversary:
  - The data holder wants to **protect against linking** an individual to a **record** or **some sensitive** value in **T** through some subset of attributes called a quasi-identifier or **QID**, where **QID ⊆ {D1,...,Dm}.**
  - On the other hand, one recipient, who is an adversary, seeks to **identify the record or sensitive values** of some target victim **V** in **T**

Problem Statement

- Assumption: the adversary knows at most **L** values of **QID attributes** of the victim
- **qid** denotes such prior known values, where **|qid| ≤ L**
- Based on the prior knowledge **qid**, the adversary could identify **a group of records**, denoted by **T[qid]**
- **|T[qid]|** denotes the number of records in **T[qid]**

Example

- Suppose qid = ⟨Janitor,M⟩.

  T[qid] = {{ID=1,…},{ID=6,…}}

  and

  |T[qid]| = 2

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Janitor	M	34	Y	Transgender
2	Doctor	M	58	N	Plastic
3	Mover	M	34	Y	Transgender
4	Lawyer	M	24	N	Vascular
5	Mover	M	58	N	Urology
6	Janitor	M	44	Y	Plastic
7	Doctor	M	24	N	Urology
8	Lawyer	F	58	N	Plastic
9	Doctor	F	44	N	Vascular
10	Carpenter	F	63	Y	Vascular
11	Technician	F	63	Y	Plastic

High-Dimensional Top-down specialization algorithm (HDTDS)

- Suppose **a domain value d** has been generalized to **a value v** in a record.
- A specialization on **v** can be shown as**: v → child(v)**

where **child(v)** denotes **the set of child values of v**

Thus, specialization replaces the parent value **v** with a child value

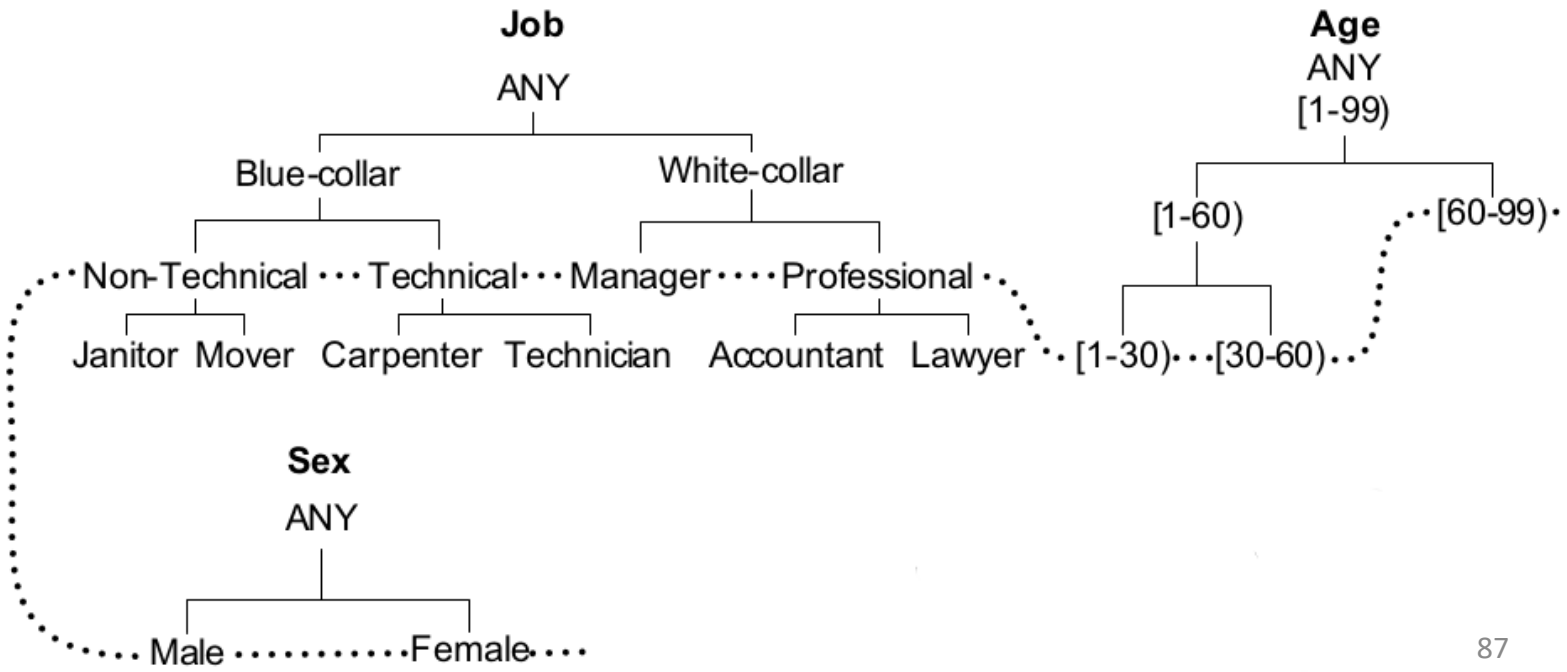
Validation: A specialization is valid if the specialization results in a table **satisfying the LKC-privacy requirement** after the specialization.

Performing a specialization: A specialization is performed only if it is valid.

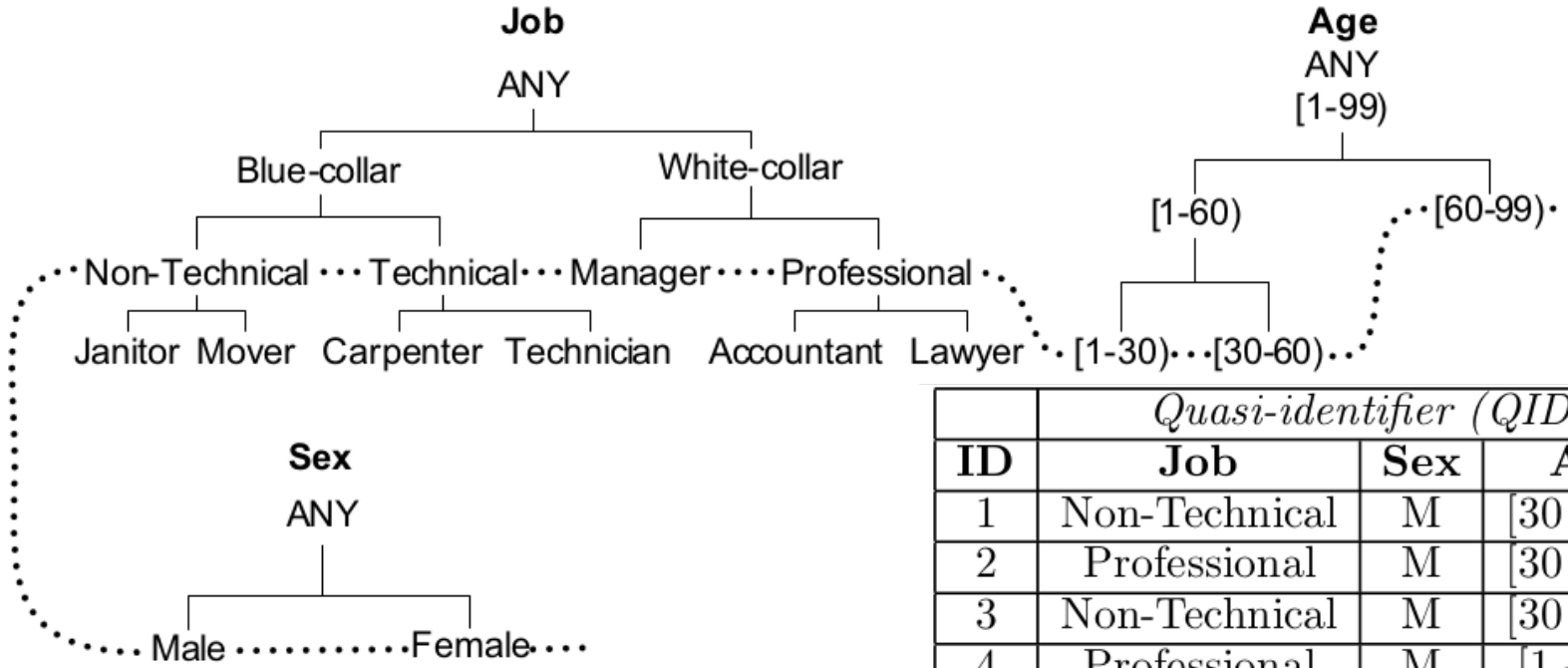
High-Dimensional Top-down specialization algorithm (HDTDS)

- The specialization process can be viewed as pushing the "**cut**" of each taxonomy tree downwards.
- A **cut** of the taxonomy tree for an attribute $\mathbf{D_i}$, denoted by $\mathbf{Cut_i}$, contains exactly one value on each root-to-leaf path.

Example: according to this cut, **attribute Job** can have one of the following values: *Non-technical, Technical, Manager, Professional*



High-Dimensional Top-down specialization algorithm (HDTDS)

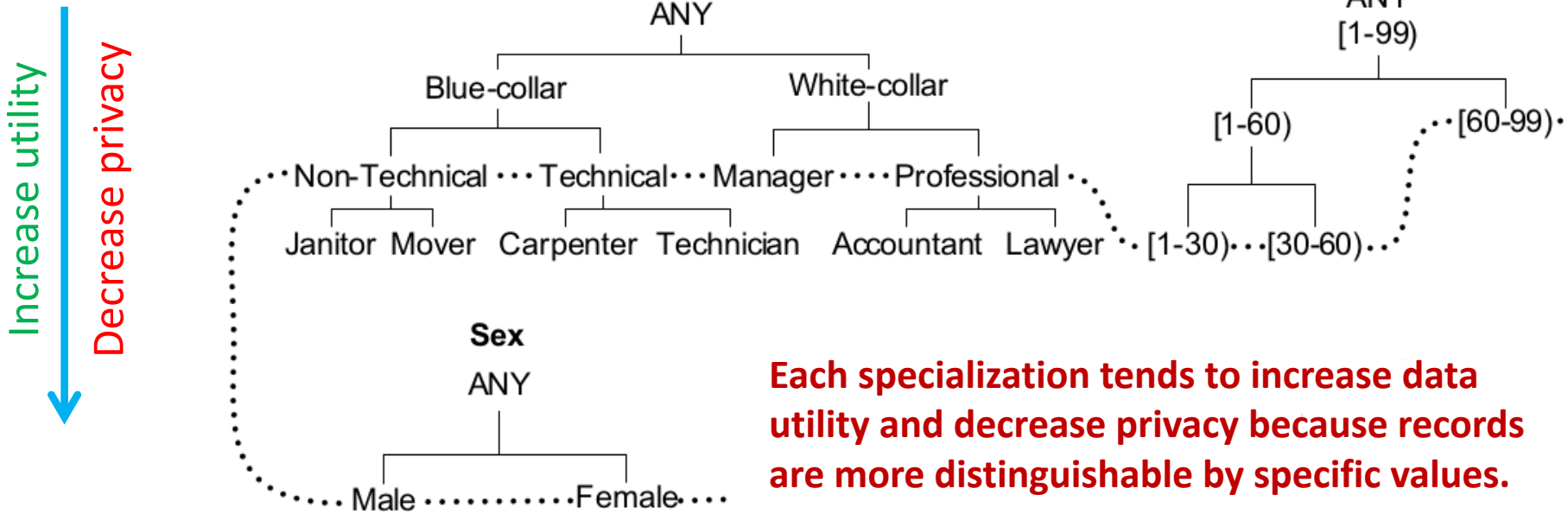


After applying a specialization based on the cut presented, data table will look like this.

	Quasi-identifier (QID)			*Class*	*Sensitive*
ID	**Job**	**Sex**	**Age**	**Transfuse**	**Surgery**
1	Non-Technical	M	$[30-60)$	Y	Transgender
2	Professional	M	$[30-60)$	N	Plastic
3	Non-Technical	M	$[30-60)$	Y	Transgender
4	Professional	M	$[1-30)$	N	Vascular
5	Non-Technical	M	$[30-60)$	N	Urology
6	Non-Technical	M	$[30-60)$	Y	Plastic
7	Professional	M	$[1-30)$	N	Urology
8	Professional	F	$[30-60)$	N	Plastic
9	Professional	F	$[30-60)$	N	Vascular
10	Technical	F	$[60-99)$	Y	Vascular
11	Technical	F	$[60-99)$	Y	Plastic

High-Dimensional Top-down specialization algorithm (HDTDS)

- The specialization starts from the **topmost cut** and pushes down the **cut** iteratively by specializing some value in the current cut **until violating the LKC-privacy requirement**.
- In other word: the specialization process pushes the cut downwards **until no valid specialization is possible.**



Each specialization tends to increase data utility and decrease privacy because records are more distinguishable by specific values.

High-Dimensional Top-down specialization algorithm (HDTDS)

Every value of QID in T

Algorithm 6.3.1 High-Dimensional Top-Down Specialization (HDTDS)

1: Initialize every value in T to the topmost value;
2: Initialize Cut_i to include the topmost value;
3: **while** some candidate $v \in \cup Cut_i$ is valid **do**
4: Find the $Best$ specialization from $\cup Cut_i$;
5: Perform $Best$ on T and update $\cup Cut_i$;
6: Update $Score(x)$ and validity for $x \in \cup Cut_i$;
7: **end while**;
8: Output T and $\cup Cut_i$.;

Candidate with highest score

It terminates when there are no more valid candidates in the cut (any further specialization would lead to a violation of the LKC-privacy requirement)

High-Dimensional Top-down specialization algorithm (HDTDS)

Algorithm 6.3.1 High-Dimensional Top-Down Specialization (HDTDS)

1: Initialize every value in T to the topmost value;
2: Initialize Cut_i to include the topmost value;
3: **while** some candidate $v \in \cup Cut_i$ is valid **do**
4: Find the $Best$ specialization from $\cup Cut_i$;
5: Perform $Best$ on T and update $\cup Cut_i$;
6: Update $Score(x)$ and validity for $x \in \cup Cut_i$;
7: **end while**;
8: Output T and $\cup Cut_i$.;

- **Anti-monotone Property**: if a generalized table violates LKC-privacy, it also violates LKC-privacy requirements after specialization
  - Guarantees sub-optimal solution

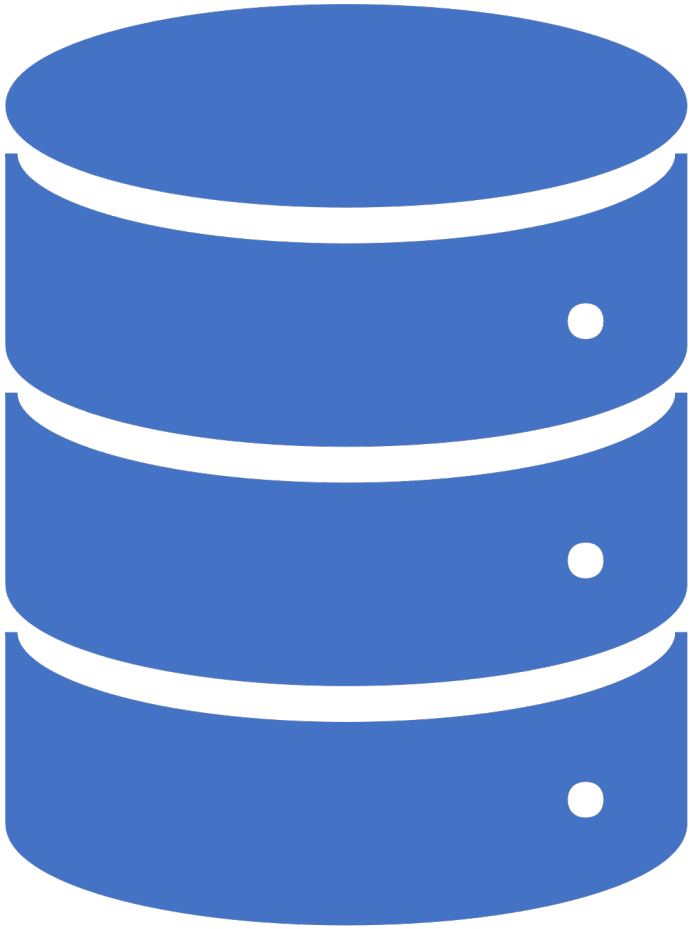
How to evaluate the “goodness” (score) of a specialization?

Since the table was released for data analysis, thus, **the Score for Classification Analysis** can be used to evaluate the score

For the requirement of classification analysis, **information gain** can be used, which denoted by:

$$InfoGain(v)$$

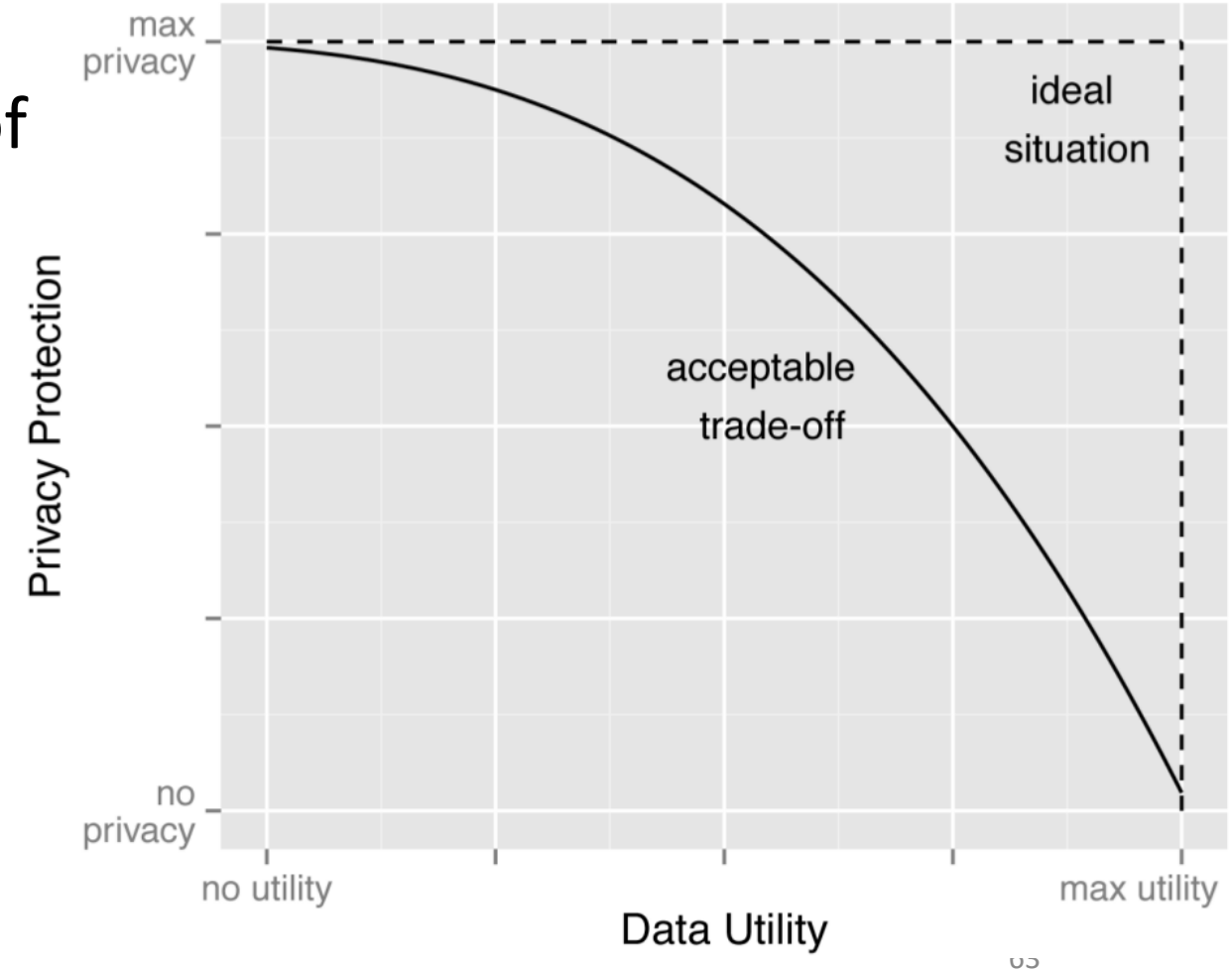
“Goodness” of a specialization on v



Data Utility

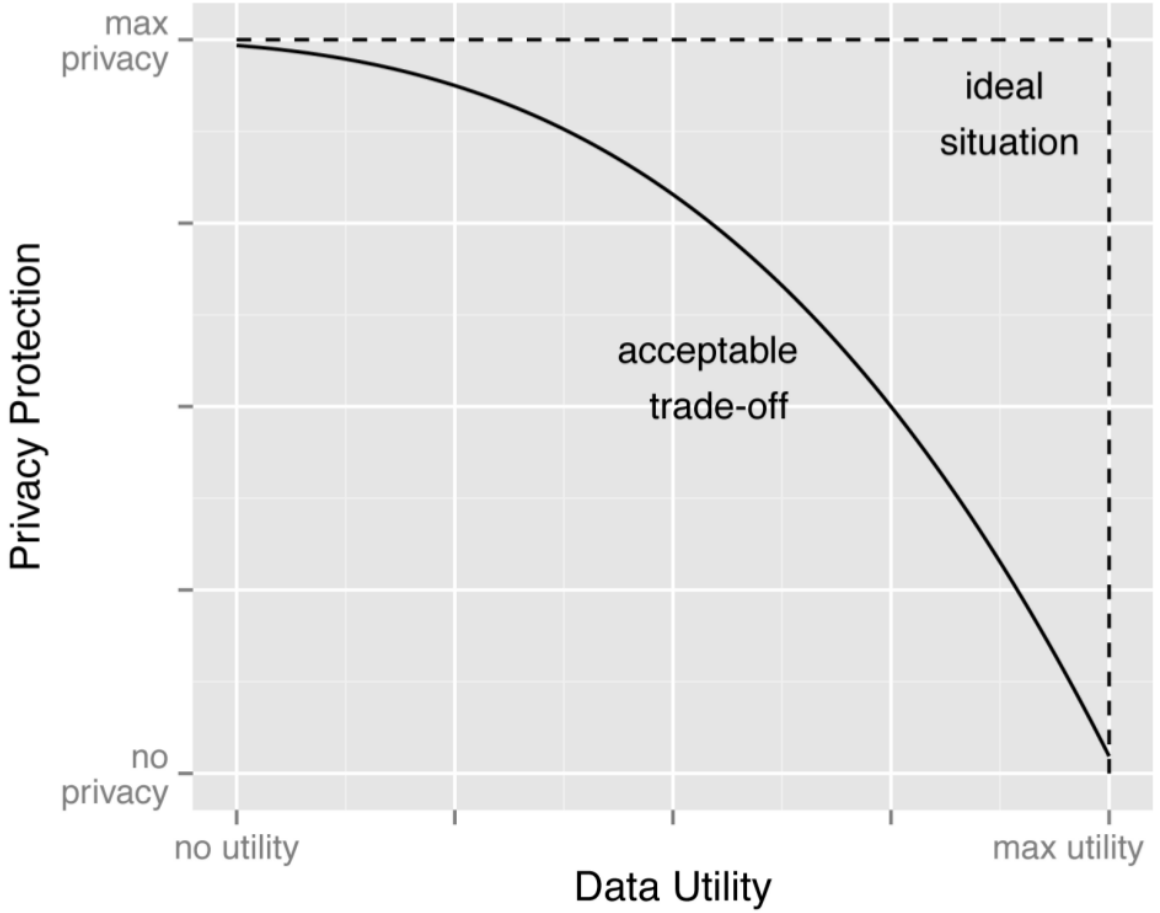
Utility versus Privacy

- Anonymization causes information loss, which can compromise utility of data
- **Objective:** Maximum privacy and Maximum usefulness (Ideal Situation)
- Impossible to achieve this objective



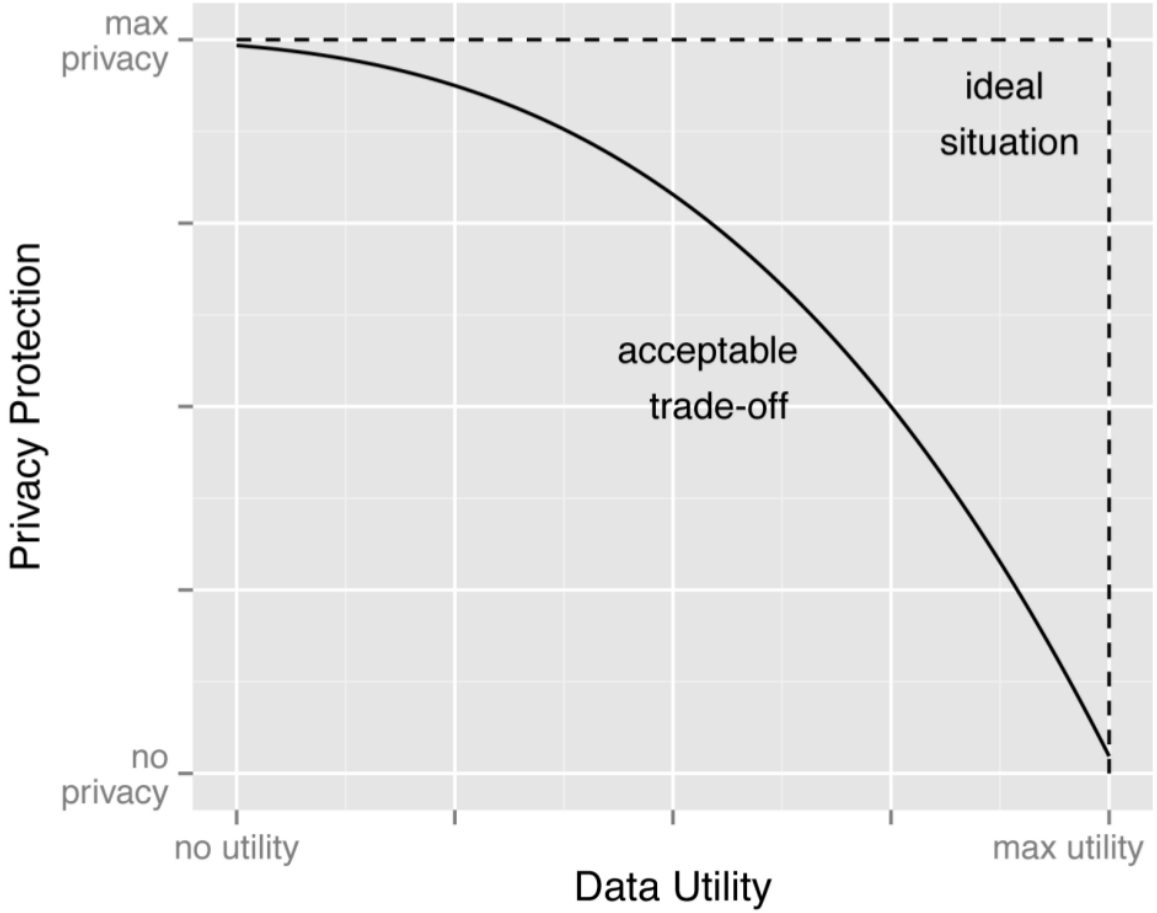
Utility versus Privacy

- **Maximum privacy** protection (i.e., zero risk) means very little to no information being released
- **De-identification** will always result in some **loss of information**, and hence a **reduction in data utility**
- Goal: **minimum loss of information** (acceptable data utility) and a **very small re-identification risk**



Utility versus Privacy

- Goal: minimum loss of information
  - so that the data can still be useful for data analysis
- At the same time, we want to make sure that the re-identification risk is very small
  - very small re-identification risk



Measuring Data Utility – Utility Metrics

- Precision
- Information Loss (ILoss)
- Discernibility
- Average Equivalence Class Size

Measuring Data Utility - Notation

- **D**: Domain
- **|D|** : number of registers
- N_a: number of attributes in **QID**

Attribute Value, AV_{ij}

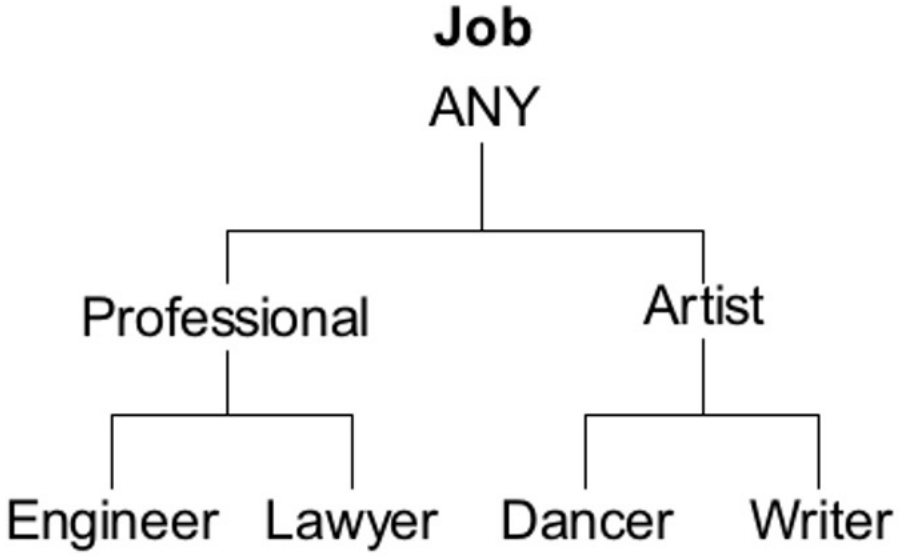
r_j : a register

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

A_i : an attribute

Utility Metrics: Precision/Minimal Distortion

- Measures data distortion
  - A penalty to every instance of generalized or suppressed attribute value (AV)
  - Example: generalizing 10 instances of **Engineer** to **Professional** causes 10 units of distortion, and further generalizing these instances to ANY Job causes another 10 units of distortion.
  -



Utility Metrics: Precision/Minimal Distortion

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D|} \frac{h_{ij}}{HGV}}{|D| * |N_a|} \in [0, 1]$$

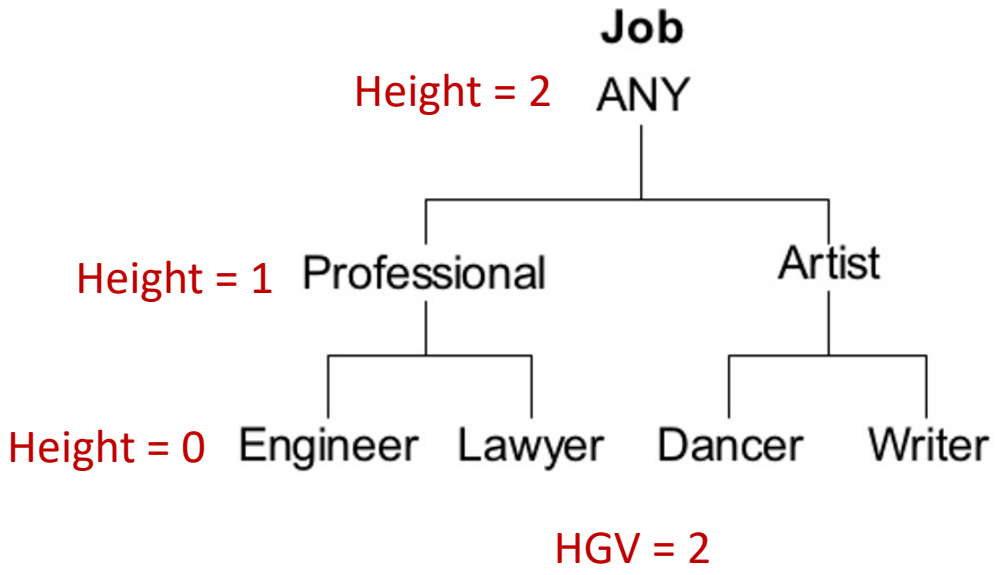
Height of AV_{ij}

Highest Generalization Value (max. hierarchy height)

Number of QID attributes

Number of registers in the table

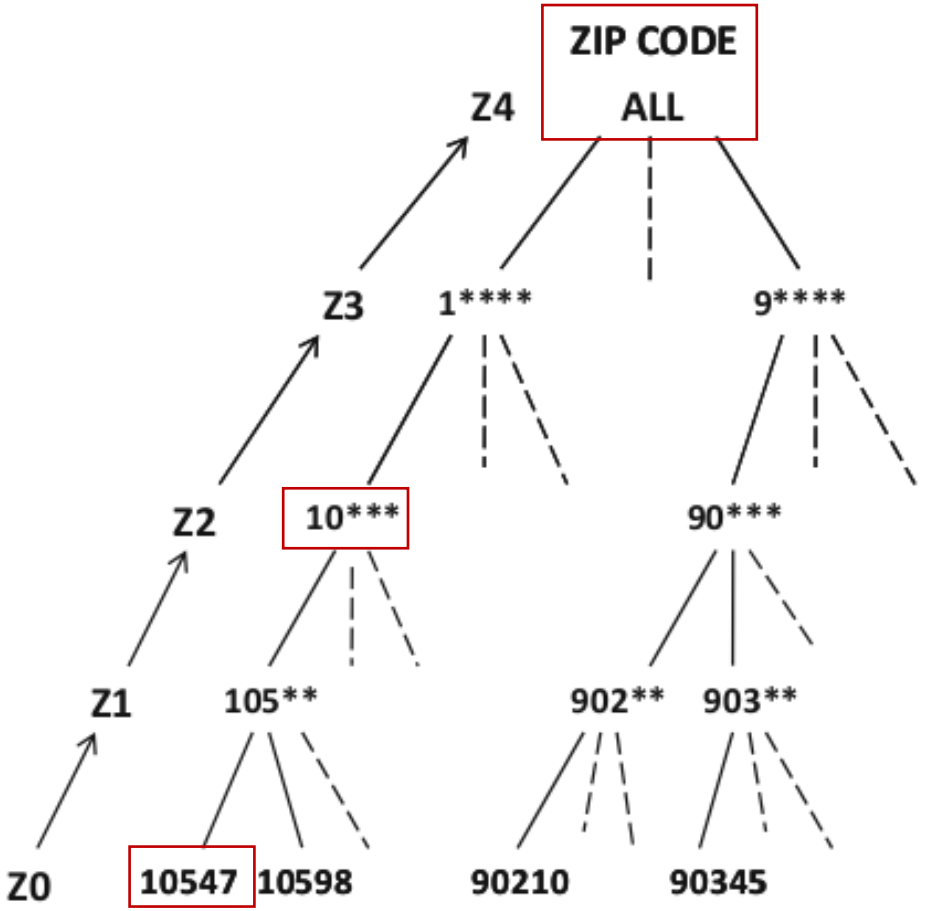
- For any attribute in QID (**N_a** times), check the height of each attribute value in all records we have in the dataset (**|D|** times) and
- Calculate **h_{ij}/HGV** (distortion for each single attribute)
- Sum all values
- Divide the result by the total number of attributes checked ($|D| * N_a$)



Utility Metrics: Precision/Minimal Distortion

$$Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D|} \frac{h_{ij}}{HGV}}{|D| * |N_a|} \in [0, 1]$$

$$Distortion\ of\ AVij_{.} = \frac{h_{ij}}{HGV}$$



HGV = 4

- For 10547: height = 0 -> $Distortion$ = 0/4 (no distortion)
- For 10***: height = 2 -> $Distortion$ = 2/4
- For Zip CODE ALL: height = 4 -> $Distortion$ = 4/4 (suppression, highest distortion)

Utility Metrics: Information Loss (ILoss)

- Ratio of leaf nodes that are generalized

$$\text{ILoss}(D) = \frac{\sum_{i=1}^{Na} W_i \times \text{ILoss}(A_i)}{|D| * |N_a|}, \text{where}$$

Weight of attribute A_i

Number of Leaf Nodes for A_{ij}

$$\text{I}Loss(A_i) = \sum_{j=1}^{|D|} \frac{LN(A_{ij}) - 1}{|D_{A_i}|}$$

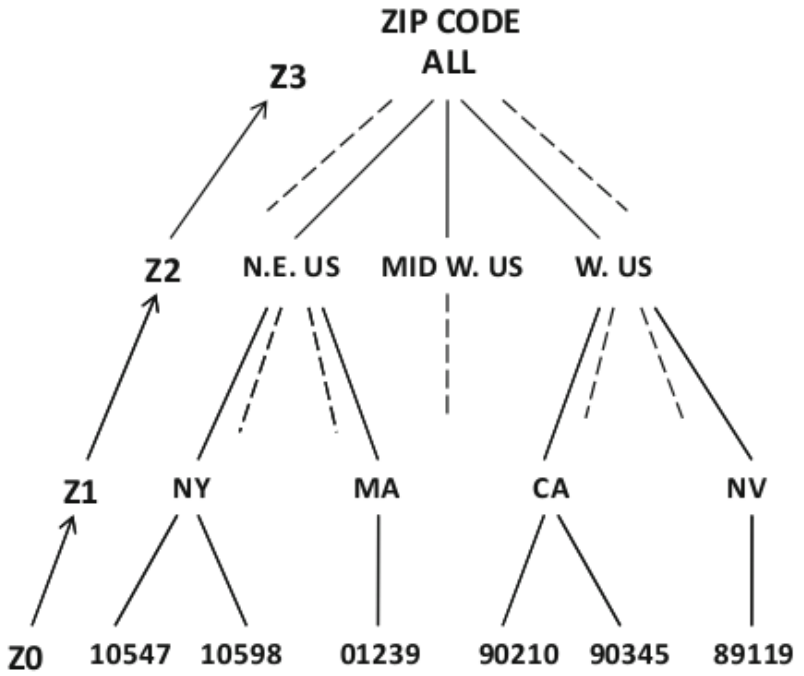
For each Attribute in QID

Number of domain values for attribute A_i

Utility Metrics: Information Loss (ILoss)

$$\mathrm{ILoss}(D) = \frac{\sum_{i=1}^{N_a} W_i \times \mathrm{ILoss}(A_i)}{|D| * |N_a|}, \text{where}$$

$$ILoss(A_i) = \sum_{j=1}^{|D|} \frac{LN(A_{ij}) - 1}{|D_{A_i}|}$$



$|D_{A_i}|$: Can be the total number of leaf nodes, or the total number of nodes in the hierarchy of attribute Ai (used for normalization)

- LN("NY") = 2
- LN("10547") = 1

A_i :ZIP CODE
A_{IJ}: NY, MA, 10547

Utility Metrics: Discernibility

- Addresses the notion of loss by charging a **penalty** to each record for being **indistinguishable** from other records with respect to QID
- If a record belongs to a $\mathbf{QID_i}$ group(equivalence class) of size E_i, the penalty for the record will be E_i.
- Thus, the penalty on an equivalence class (EC) of records is

$$E_i + E_i + \ldots E_i = E_i^2$$

for each record in the EC: E_i

$$\text{DM (D)} = \sum_{QID_i} E_i^2$$

Utility Metrics: Discernibility

Example:

- Penalty of:
  - EC = ([20-30], M, 60800***) -> $2^2 = 4$
  - EC = (>40, *, 60790***) -> $3^2 = 9$

DM (D) = 4 +9 = 13

Age	Gender	ID
[20-30]	M	60800***
[20-30]	M	60800***
>40	*	60790***
>40	*	60790***
>40	*	60790***

EC with large number of records have more information loss

Utility Metrics: Average Equivalence Class Size

- Another metric based on Equivalent Classes (EC) which is related to K-anonymity
- K-anonymity: If one record in the table has some value *QID*, at least *k-1* other records also have the value *QID*, Thus, the minimum equivalence class size on *QID* is at least ***k***

$$C_{AVG} = \underbrace{\left(\frac{\text{total \#registers}}{\text{total \#ECs}}\right)}_{\text{Average size}} \underbrace{/k}_{\text{Is used to normalize the value}}$$

By doing this normalization, the metric measures **how well a EC size** is **close to the optimal case** (i.e., according to K-anonymity which is the **smallest possible EC size, k**)

Utility Metrics: InfoGain

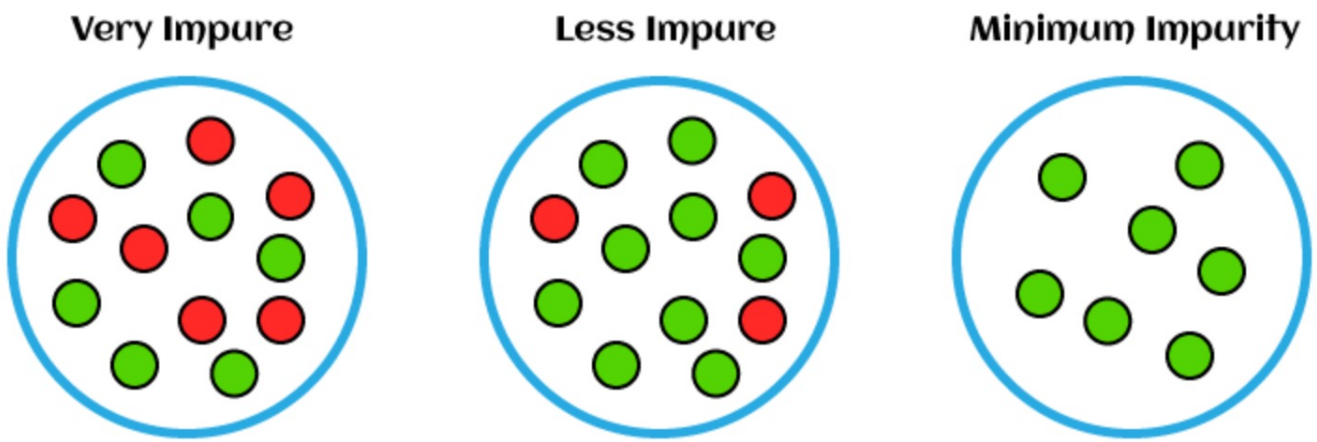
$$Score(v) = InfoGain(v) = E(T[v]) - \sum_c \frac{|T[c]|}{|T[v]|} E(T[c]),$$

$$\text{where } E(T[x]) \text{ is the } \textit{entropy} \text{ of } T[x]$$

T[x] is the set of records in T generalized to value x

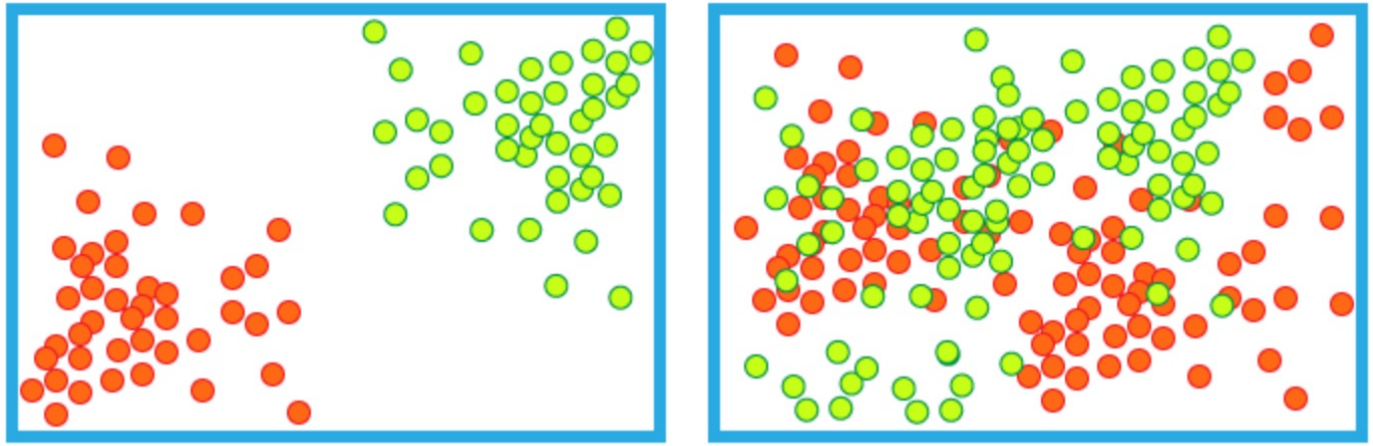
Utility Metrics: InfoGain

- What is Entropy?
- Entropy is defined as the **randomness** or measuring the **disorder/confusion** of the information
- In machine learning, entropy measures the **unpredictability** or **impurity** in the data.



Utility Metrics: InfoGain

- What is Entropy?
- if it is easier to **draw a valuable conclusion** from a piece of information, then **entropy will be lower** in Machine Learning
- if **entropy is higher**, then it will be **difficult to draw any conclusion** from that piece of information.



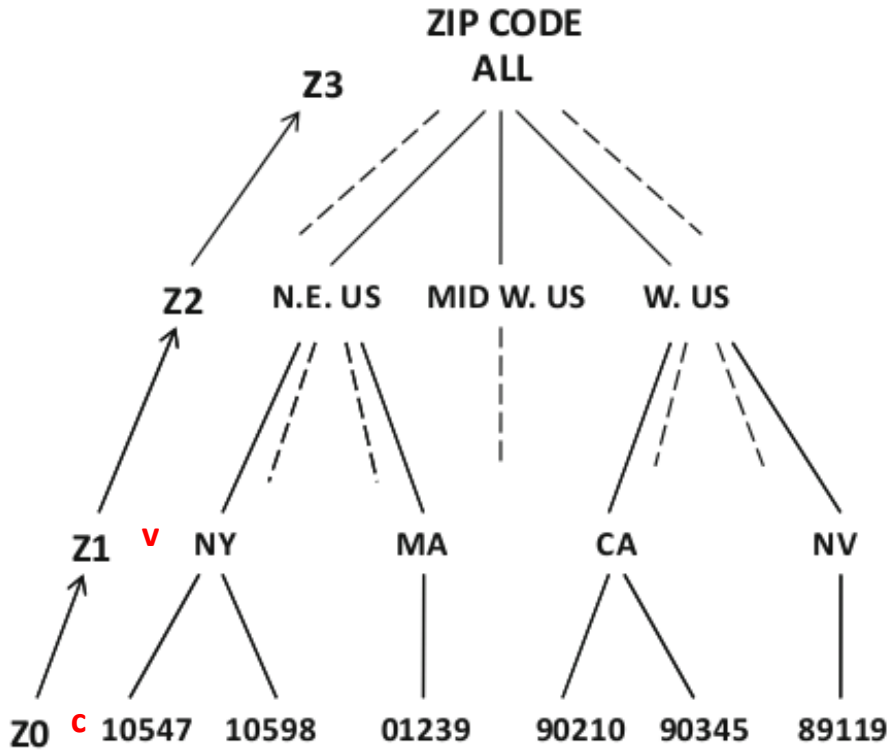
Utility Metrics: InfoGain

- Each specialization (specialization of v ⟶ c = child(v))
  - decrease of confusion (or entropy) -> increases utility -> Higher InfoGain

$$Score(v) = InfoGain(v) = E(T[v]) - \sum_{c} \frac{|T[c]|}{|T[v]|} E(T[c])$$

Entropy: confusion for generalizing to v (e.g., “NY”)

For each child c: Reduction on the confusion for specializing to c (e.g., “10547”)



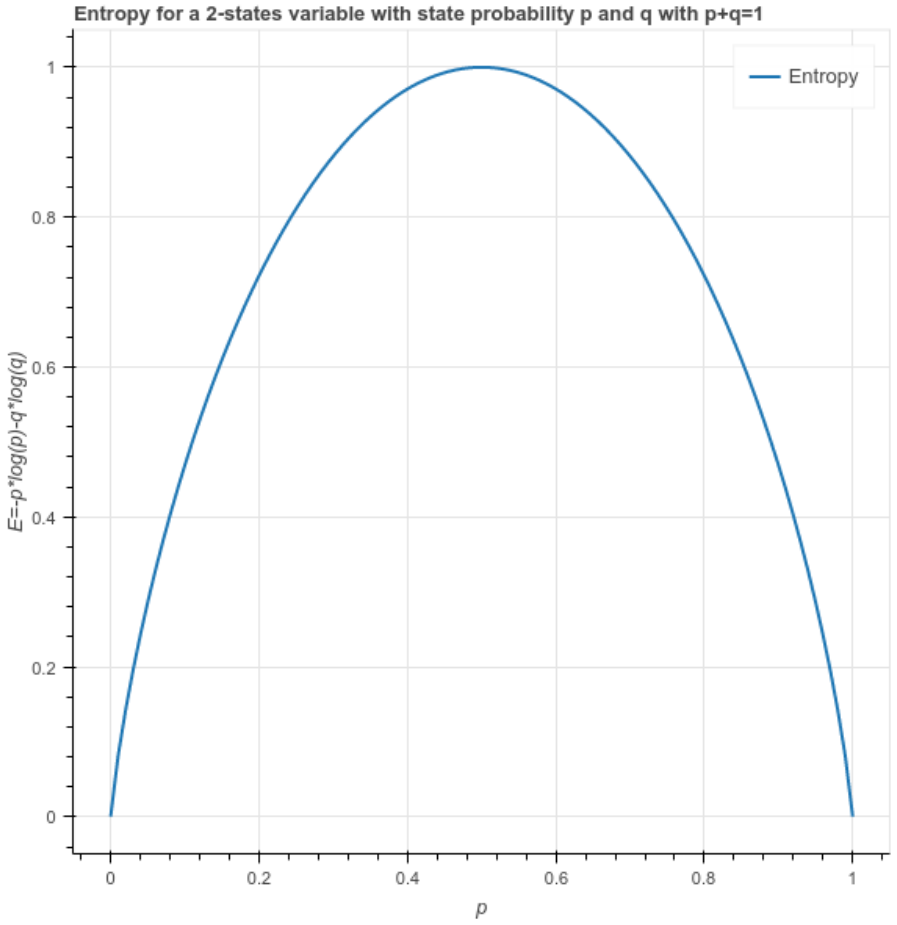
Utility Metrics: InfoGain

$$Score(v) = InfoGain(v) = E(T[v]) - \sum_{c} \frac{|T[c]|}{|T[v]|} E(T[c])$$

$$E(T[x]) = -\sum_{cls} \frac{freq(T[x], cls)}{|T[x]|} \times log_2 \frac{freq(T[x], cls)}{|T[x]|}$$

probability of a class

log to base 2 of the class



Entropy(x) measures the amount of uncertainty associated with x:

E(x) = 0 -> x is deterministic
E(x) = 1 -> x is uniformly distributed

Utility Metrics: InfoGain Example

$$E(T[x]) = -\sum_{cls} \frac{freq(T[x], cls)}{|T[x]|} \times log_2 \frac{freq(T[x], cls)}{|T[x]|}$$

Classes = {“cancer”, “no cancer”}

|T[“**F**”]| = **10**

Freq(T[“**F**”], “cancer”) = **10** -> Freq(T[“**F**”], “no cancer”) = **0**

E(T[“**F**”]) = - (**10** / **10** * log2(**10** / **10**) + **0** / **10** * log2(**0** / **10**)

0 0

E = 0 -> “F” is deterministic

Utility Metrics: InfoGain Example

$$E(T[x]) = -\sum_{cls} \frac{freq(T[x], cls)}{|T[x]|} \times log_2 \frac{freq(T[x], cls)}{|T[x]|}$$

Classes = {“cancer”, “no cancer”}

|T[“**F**”]| = **10**

Freq(T[“**F**”], “cancer”) = **5** -> Freq(T[“**F**”], “no cancer”) = **5**

E(T[“**F**”]) = - (**5** / **10** * log2(**5** / **10**) + **5** / **10** * log2(**5** / **10**)

-0.5 **-0.5**

E = 1 -> “**F**” is uniformly distributed

Privacy Models

- 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

Semantic Privacy Models

Differential Privacy

Two Scenarios

1. Adversary: “everyone has two feets”
   - Statistical DB shows that almost everyone has one left foot and one right foot
   - Increases knowledge of the adversary
   - But has privacy been compromised?
2. Adversary: “Ana is two inches taller than the average Portuguese women”
   - Statistical DB teaches avg height of Portuguese women
   - Adversary learns Ana’s height
   - Without access to DB, adversary learns much less

Differential Privacy

- Goals:
  - Released DB reveals “little” about any individual
    - Even if adversary knows (almost) everything about anyone else
  - Answer to any query is “probably indistinguishable” with or without a particular row in the DB
- Encourages participation in the dataset, increasing utility
- Key mechanism: data perturbation (adding noise)

DP Formal Definition

- D_1 and D_2: DBs that differ at most in 1 element (D_1 and D_2 are neighbors)
- f: is a query function over the DBs
- D: is the domain of all data
- M: is a randomized function/algorithm that adds noise to queries: M(D) = f(D) + noise
- $S \subseteq range(\mathrm{M})$, i.e., contained in the range of results from M

$$\Pr[\mathrm{M}(D_1) \in S] \leq e^{\varepsilon} \times \Pr[\mathrm{M}(D_2) \in S]$$

Difference between probabilities (distributions) of a query returning the same result in two neighbor datasets is bounded by ε

DP Formal Definition

$$\Pr[\mathrm{M}(D_1) \in S] \leq e^{\varepsilon} \times \Pr[\mathrm{M}(D_2) \in S]$$

For any neighbor DBs that differ in 1 registry,

the adversary will not be able to distinguish

D_1 and D_2 from the result of M

The inclusion/exclusion of an individual is probabilistically indistinguishable

DP: How much noise is needed?

- f: is a query function over the DBs
- M: is a randomized function/algorithm that adds noise to queries: M(D) = f(D) + noise
- **Amount of noise to add depends on the query f**

Sensitivity of query f

$$\Delta f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||$$

for all D_1, D_2 differing at most by 1 element

- Measures how much difference 1 individual makes in the output of the query

- Important for privacy: the higher Δf is, the more noise must be added

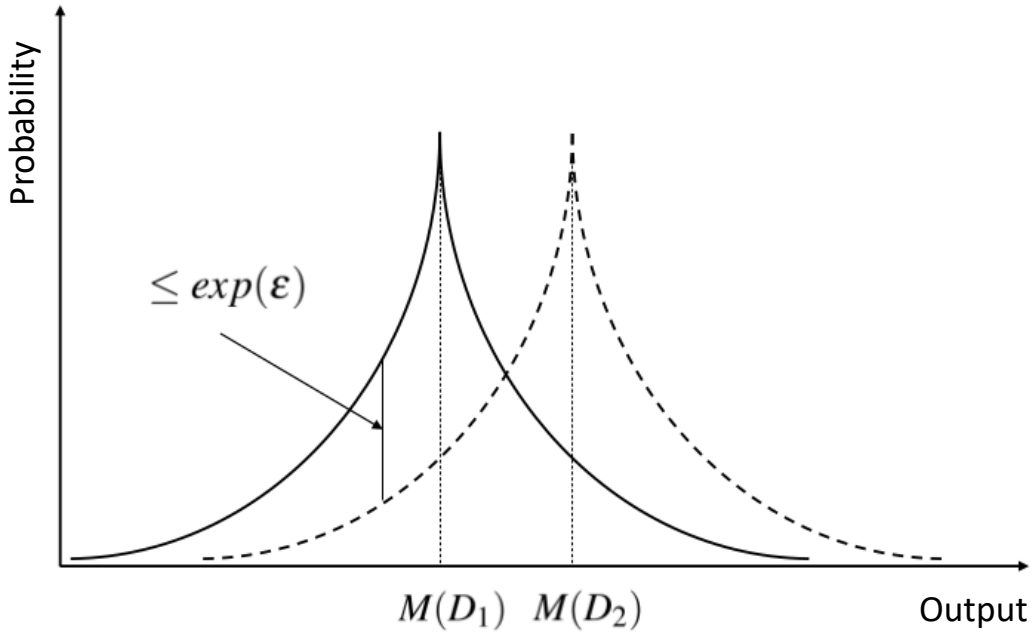
ε-DP Laplace Mechanism

$$M(D) = f(D) + \text{Lap}\left(0, {}^{\Delta f}/_{\varepsilon}\right)$$

noise

where $\mathbf{Lap}\left(\mathbf{0}, {}^{\boldsymbol{\Delta f}}/_{\boldsymbol{\varepsilon}}\right)$ is the Laplace distribution with mean **0** and scale ${}^{\boldsymbol{\Delta f}}/_{\boldsymbol{\varepsilon}}$

- Adds noise from Laplace distribution (a symmetric exponential distribution)



Proven to guarantee
ε-Differential Privacy

$$\Pr[\mathrm{M}(D_1) \in S] \leq e^{\varepsilon} \times \Pr[\mathrm{M}(D_2) \in S]$$

DP Practical - Example

$$D_i = D \; excluding \; \{\text{user } ID = i\}$$

ID	Name	# sold real estates
1	Roney	4
2	André	2
3	Leo	7
4	Bruno	1

4 neighboring datasets

ID	Name	#real estates
2	André	2
3	Leo	7
4	Bruno	1

$f(D_1) = 2 + 7 + 1 = 10$

ID	Name	#real estates
1	Roney	4
3	Leo	7
4	Bruno	1

$f(D_2) = 4 + 7 + 1 = 12$

ID	Name	#real estates
1	Roney	4
2	André	2
4	Bruno	1

$f(D_3) = 4 + 2 + 1 = 7$

ID	Name	#real estates
1	Roney	4
2	André	2
3	Leo	7

$f(D_4) = 4 + 2 + 7 = 13$

$f(D_i)$: sum of all real estates in D_i

$f(D) = 14$

DP Practical Example

- Sensitivity: $\Delta f = \max_{D,D_i} ||f(D) - f(D_i)||$
  - $\Delta f = f(\mathrm{D}) - f(D_3) = 14 - 7 = 7$
- Apply ε-DP Laplace mechanism with $\mathbf{Lap}(\mathbf{0}, {}^{7}/_{\varepsilon})$

ID	Name	# real estates
1	Roney	4
2	André	2
3	Leo	7
4	Bruno	1

How to select ε?

How to select ε?

- ε is a public parameter
  - Not a direct measure of privacy, but of the impact of data from 1 user in the dataset
  - Hard to estimate

- Smaller $\varepsilon \rightarrow$ more noise (increases privacy)

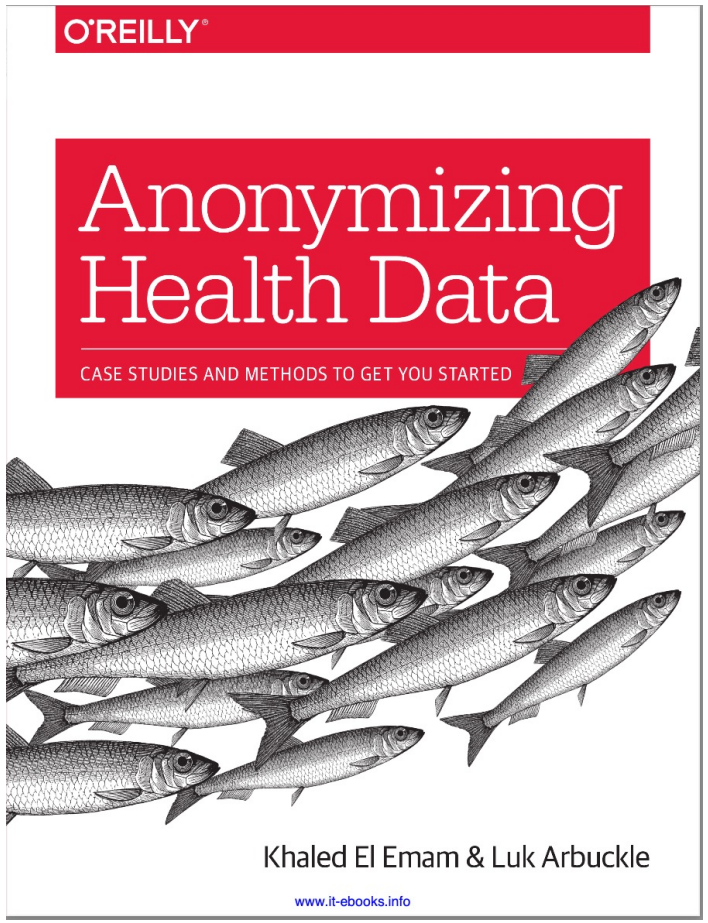
- Its selection is a social question
  - Common values: 0.01, 0.2, ln(2), ln(3)

PPDP References

1. [IPPDP] Introduction to Privacy-Preserving Data Publishing Concepts and Techniques, Fung et al., CRC 2011
2. [PPDM] Data Mining: The Textbook, Charu C. Aggarwal, Springer 2015
3. [GDPHI] Guide to the De-identification of Personal Health Information, Khaled El Emam, CRC 2013
4. Anonymizing Health Data, Khaled El Emam, Luk Arbuckle, O'Reilly, 2014
5. Privacy-Preserving Data Mining: Methods, Metrics and Applications, R. Mendes, J.P. Vilela, IEEE Access, 2017
6. k-Anonymity and Other Cluster-Based Methods slides, Vitaly Shmatikov, CS 380S
7. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, Li et al., ICDE 2007
8. l-Diversity: Privacy Beyond k-Anonymity, Machanavajjhala et al.,TKDD 2007
9. Preservação de Privacidade de Dados: Fundamentos, Técnicas e Aplicações, Felipe Timbó, Brito Javam Machado, 2017
10. A Firm Foundation for Private Data Analysis, Cynthia Dwork, CACM 2011
11. ARX - A Comprehensive Tool for Anonymizing Biomedical Data, Prasser et al.

Bibliography

Chapter 2



Chapter 6