Source of Privacy Problem

- Even if we do not publish the identities of individuals, there are some (non-sensitive) fields that may *uniquely* identify some individuals
  - These attributes form the *quasi identifier*

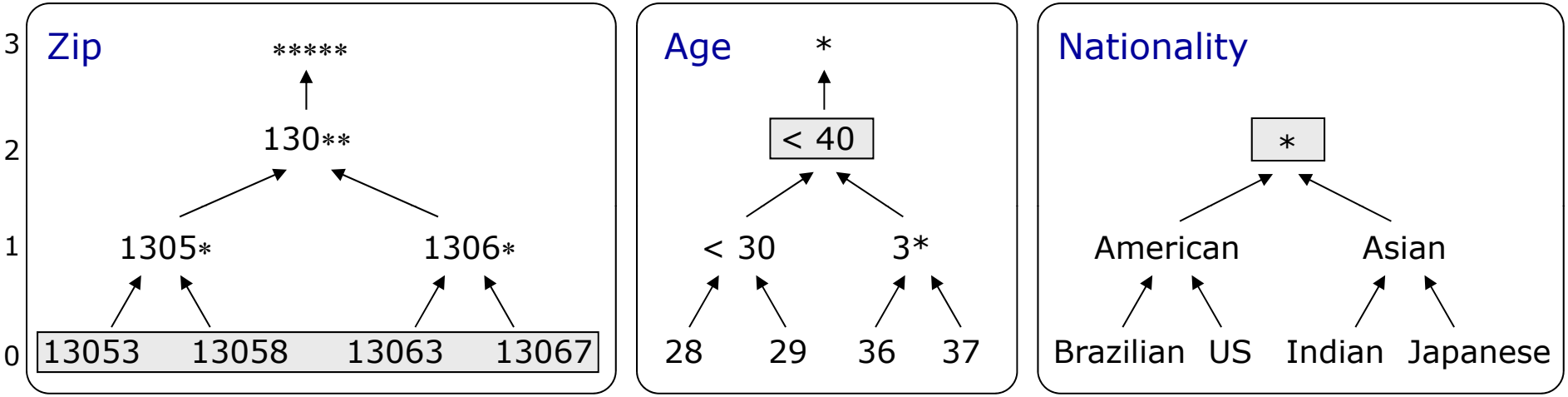
	Non-Sensitive Data (Quasi identifier)			*Other non-sensitive attributes (non-QI)*	*Sensitive Data*
#	***Zip***	***Age***	***Nationality***	***…***	***Condition***
…	…	…	…		…

Quasi Identifier

- The attacker can use them to *join* with other sources and identify the individuals

How to k-anonymize a dataset? Generalization Hierarchies

- Generalization Hierarchies: Data owner defines how values can be generalized



- Table Generalization: A table generalization is created by generalizing all values in a column to a specific level of generalization

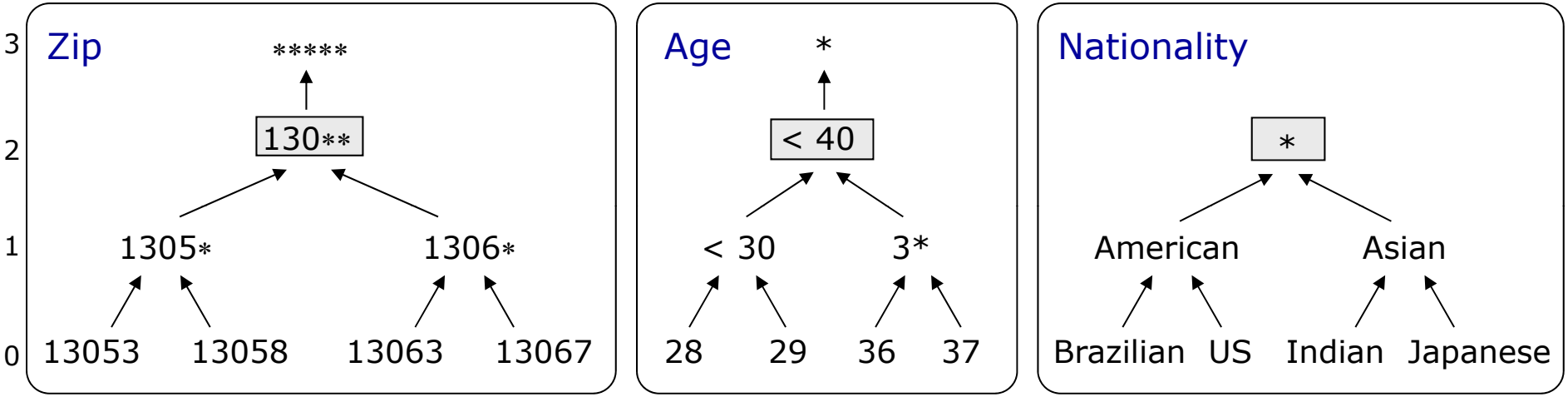
e.g.

2-anonymization

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

Generalization Hierarchies

- Generalization Hierarchies: Data owner defines how values can be generalized



- Table Generalization: A table generalization is created by generalizing all values in a column to a specific level of generalization

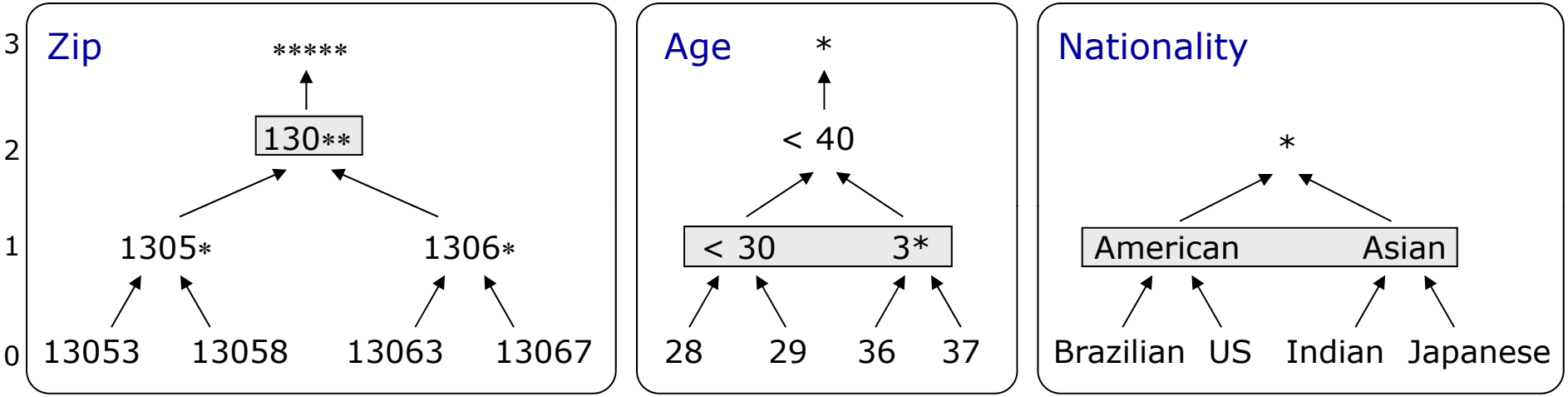
e.g.

2-anonymization

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

Generalization Hierarchies

- Generalization Hierarchies: Data owner defines how values can be generalized



- Table Generalization: A table generalization is created by generalizing all values in a column to a specific level of generalization

e.g.

2-anonymization

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

k-minimal Generalizations

- There are *many* k-anonymizations. Which to pick?
  The ones that do not generalize the data more than needed

k-minimal Generalization: A k-anonymization that is not a generalization of another k-anonymization

e.g.

✓ *2-minimal Generalization*

#	Zip	Age	Nationality	
1	13053	< 40	*	He
2	13067	< 40	*	He
3	13053	< 40	*	Ca
4	13067	< 40	*	Ca

✓ *2-minimal Generalization*

#	Zip	Age	Nationality	
1	130**	< 30	American	He
2	130**	< 30	American	He
3	130**	3*	Asian	Ca
4	130**	3*	Asian	Ca

#	Zip	Age	Nationality	
1	130**	< 40	*	H
2	130**	< 40	*	H
3	130**	< 40	*	C
4	130**	< 40	*	C

✘ *Non-minimal 2-anonymization*

k-Anonymity Attack Example

Original Data

	Quasi-Identifier			*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

The attacker knows:

- About quasi-identifiers:

Umeko		
Zip	**Age**	**National**
13068	21	Japanese

Bob		
Zip	**Age**	**National**
13053	31	American

- Other background knowledge:

  Japanese have low incidence of heart disease

k-Anonymity Attack Example

4-anonymization

	Quasi-Identifiers			*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

k-Anonymity Attack Example

4-anonymization

	Quasi-Identifiers			*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**

Bob		
Zip	***Age***	***National***
13053	31	American

Bob has Cancer!

k-Anonymity Attack Example

4-anonymization

	Quasi-Identifiers			*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

Zip	***Age***	***National***
13068	21	Japanese

Umeko has Viral Infection!

Data Leak !

Bob

Zip	***Age***	***National***
13053	31	American

Bob has Cancer!

Return a k-anonymization with the additional property that: For each distinct value of the quasi-identifier there exists ***l*** *different values for the sensitive attributes (i.e.,* ***l****-diversed)*

3-diversified

	Quasi-Identifiers			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	1305*	<= 40	*	Heart Disease
2	1306*	<= 40	*	~~Heart Disease~~
3	1306*	<= 40	*	Viral Infection
4	1305*	<= 40	*	Viral Infection
5	1485*	>= 40	*	Cancer
6	1485*	>= 40	*	Heart Disease
7	1485*	>= 40	*	Viral Infection
8	1485*	>= 40	*	Viral Infection
9	1305*	<= 40	*	Cancer
10	1305*	<= 40	*	Cancer
11	1306*	<= 40	*	Cancer
12	1306*	<= 40	*	Cancer

Attack does not work!

Umeko

Zip	***Age***	***National***
13068	21	Japanese

Umeko has Viral Infection or Cancer

Bob

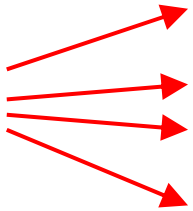
Zip	***Age***	***National***
13053	31	American

Bob has Viral Infection or Cancer or Heart Disease

What *l*-diversity guarantees

- From an *l*-diverse generalized table, an adversary (without any prior knowledge) can infer the sensitive value of each individual with confidence at most 1/***l***

Name	Age	Sex	Zipcode
Bob	23	M	11000



A 2-diverse generalized table

Age	Sex	Zipcode	Disease
[21, 60]	M	[10001, 60000]	pneumonia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	dyspepsia
[21, 60]	M	[10001, 60000]	pneumonia
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	gastritis
[61, 70]	F	[10001, 60000]	flu
[61, 70]	F	[10001, 60000]	bronchitis

Limitations of *l*-Diversity

l-diversity is insufficient to prevent attribute disclosure.

Similarity Attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relative low.
2. Bob has some stomach-related disease.

l-diversity does not consider semantic meanings of sensitive values

New notion of privacy needed to factor in the data distribution – t-closeness!