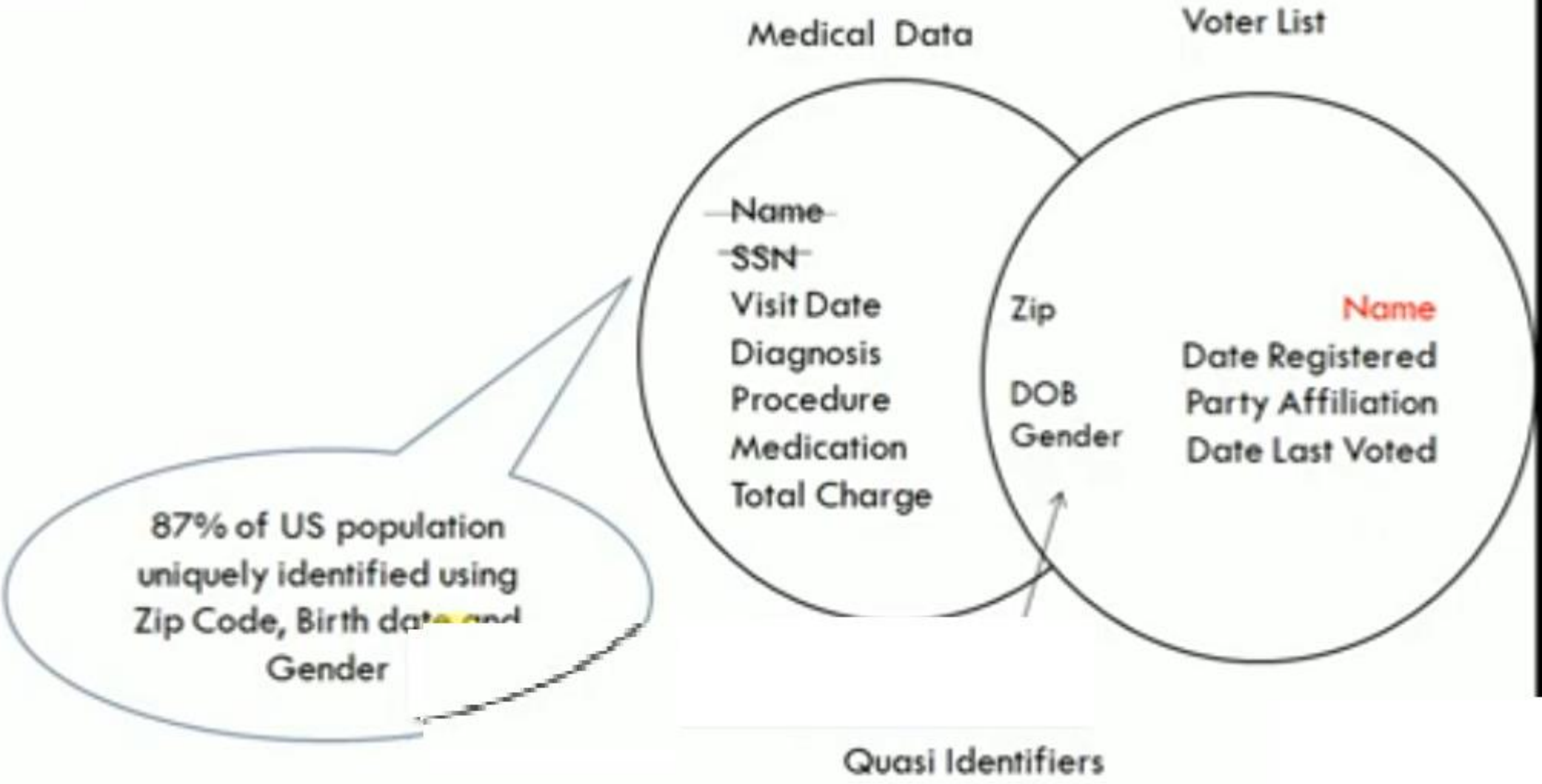
k-anonymity

Here, k-anonymity comes into play. k-anonymity means, that an individual's quasi identifiers have to be equivalent to at least k-1 other individuals.

Activate Windows
Go to Settings to activate Windows

Re-identification Attack



Re-identification Attack Example

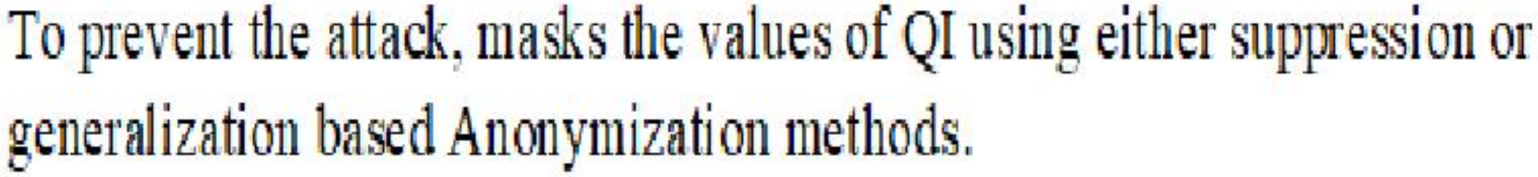
Quasi Identifier (QI)

- □ **Quasi Identifier (QI): Set of attributes included in private table, also externally available and therefore exploitable for linking.**

  OR

  QI is a minimal set of attributes that is used to uniquely identify individuals. Attack is mainly using Quasi Identifier.

- □ Attacks may be re-identification or linking attack.

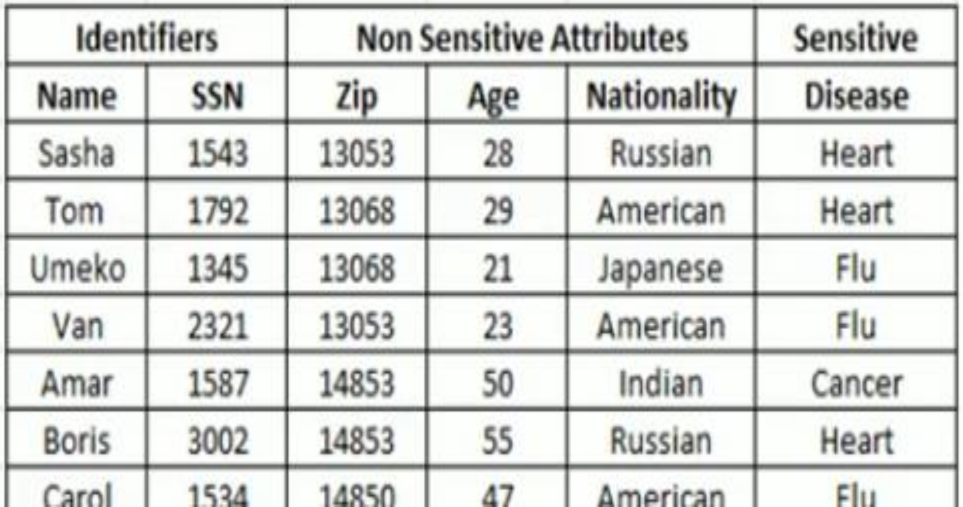


To prevent the attack, masks the values of QI using either suppression or generalization based Anonymization methods.

K-Anonymity

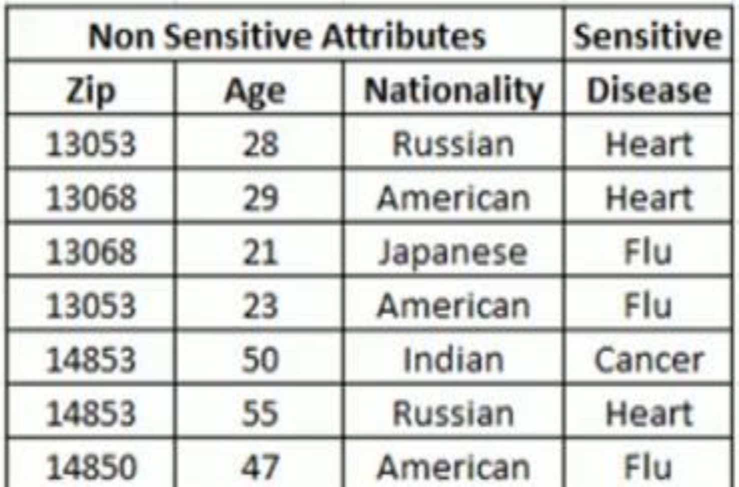
- K-anonymity is a key concept that was introduced to address the risk of re-identification of anonymized data through linkage to other datasets.

Table III Complete Table



Identifiers		Non Sensitive Attributes			Sensitive
Name	SSN	Zip	Age	Nationality	Disease
Sasha	1543	13053	28	Russian	Heart
Tom	1792	13068	29	American	Heart
Umeko	1345	13068	21	Japanese	Flu
Van	2321	13053	23	American	Flu
Amar	1587	14853	50	Indian	Cancer
Boris	3002	14853	55	Russian	Heart
Carol	1534	14850	47	American	Flu

Table IV After removing Identifiers



Non Sensitive Attributes			Sensitive
Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu

Privacy Preserving: Suppression

- Quasi identifier in previous table is {Zip code, Age, Nationality}.
- So we have to anonymize these QI.
- We suppress the QI values with ‘*’.
- Age=3* means that Age in the range [30-39].
- Nationality is suppressed by using *

Privacy Preserving: Generalization

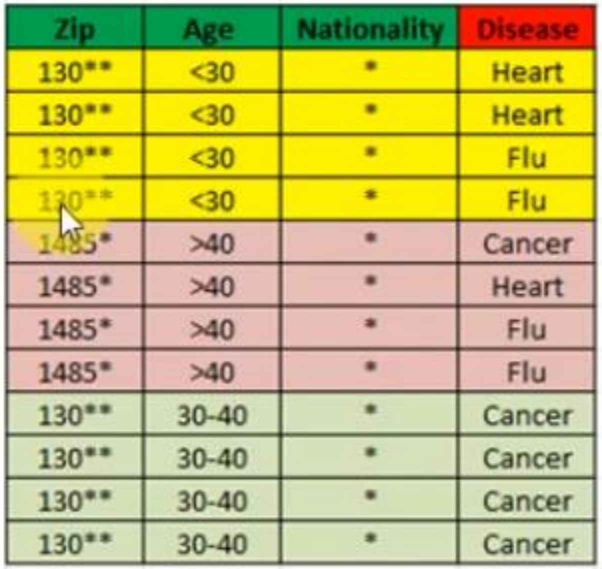
- In Generalization based anonymization method, a specific value is replaced with more general value.

Ex. Age 27 is replaced by <30

K-Anonymity Example

- □ Take quasi identifier attributes and coarsens them such that every tuple in the table shares its quasi identifier value with at least k-1 other values in the table.

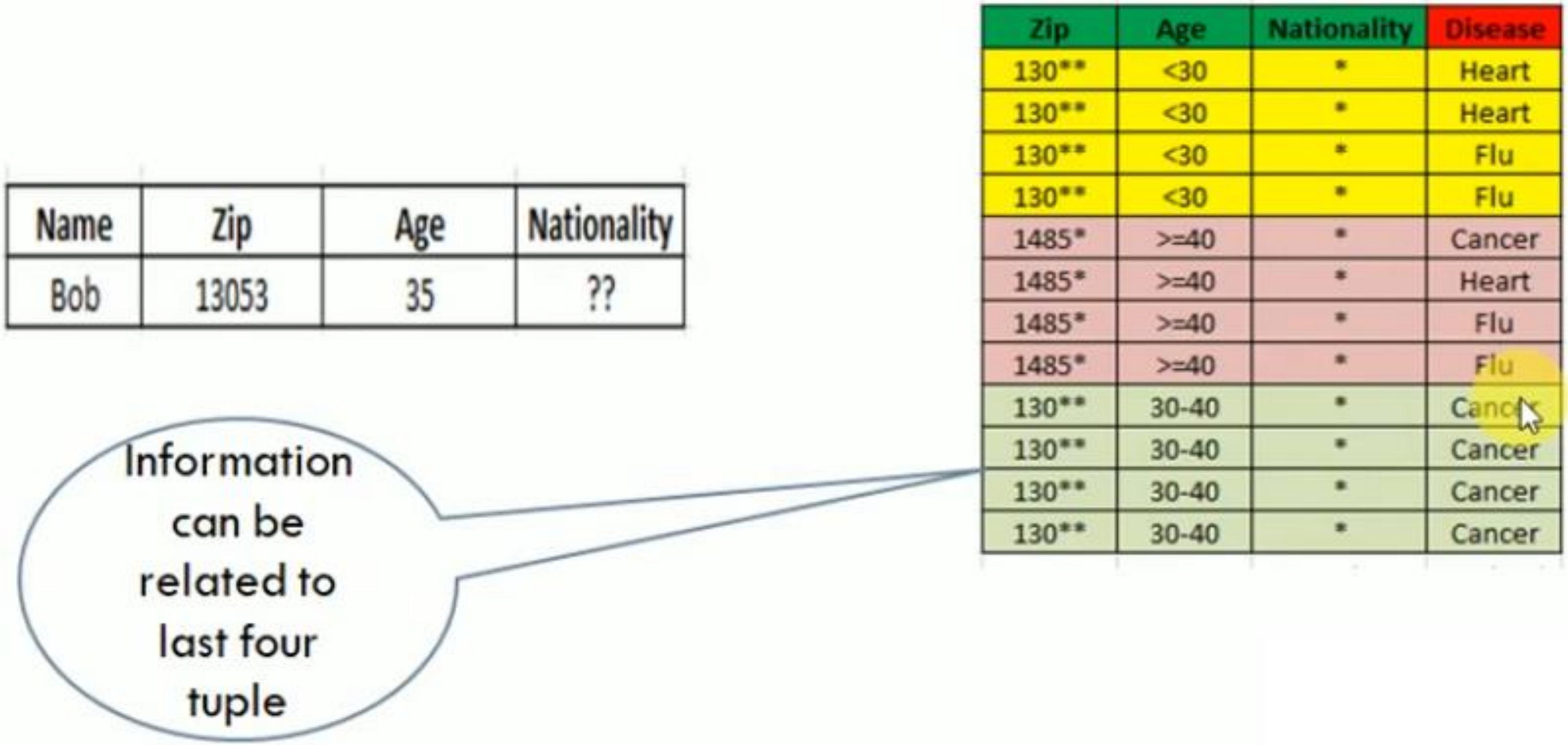
Table V 4-Anonymous Table



Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Does K-anonymity guarantee sufficient privacy?

Homogeneity Attack:



Name	Zip	Age	Nationality
Bob	13053	35	??

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>=40	*	Cancer
1485*	>=40	*	Heart
1485*	>=40	*	Flu
1485*	>=40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

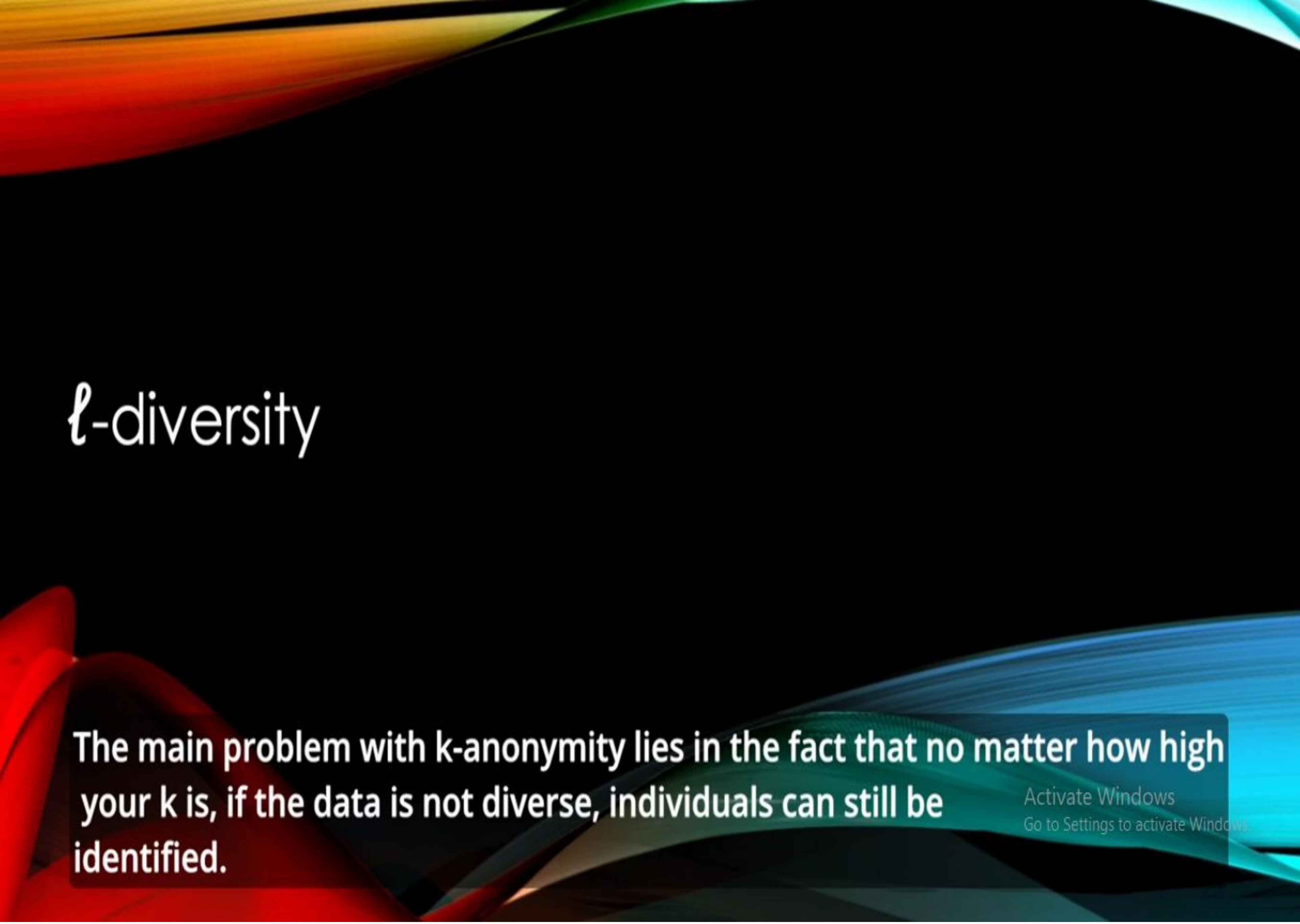
Information can be related to last four tuple

Figure-4 Homogeneity Attack

ℓ-diversity

l-diversity extends on the concept of k-anonymity and addresses some privacy issues that remain after k-anonymity is applied to protect a database from attacks.

Activate Windows
Go to Settings to activate Windows.



ℓ-diversity

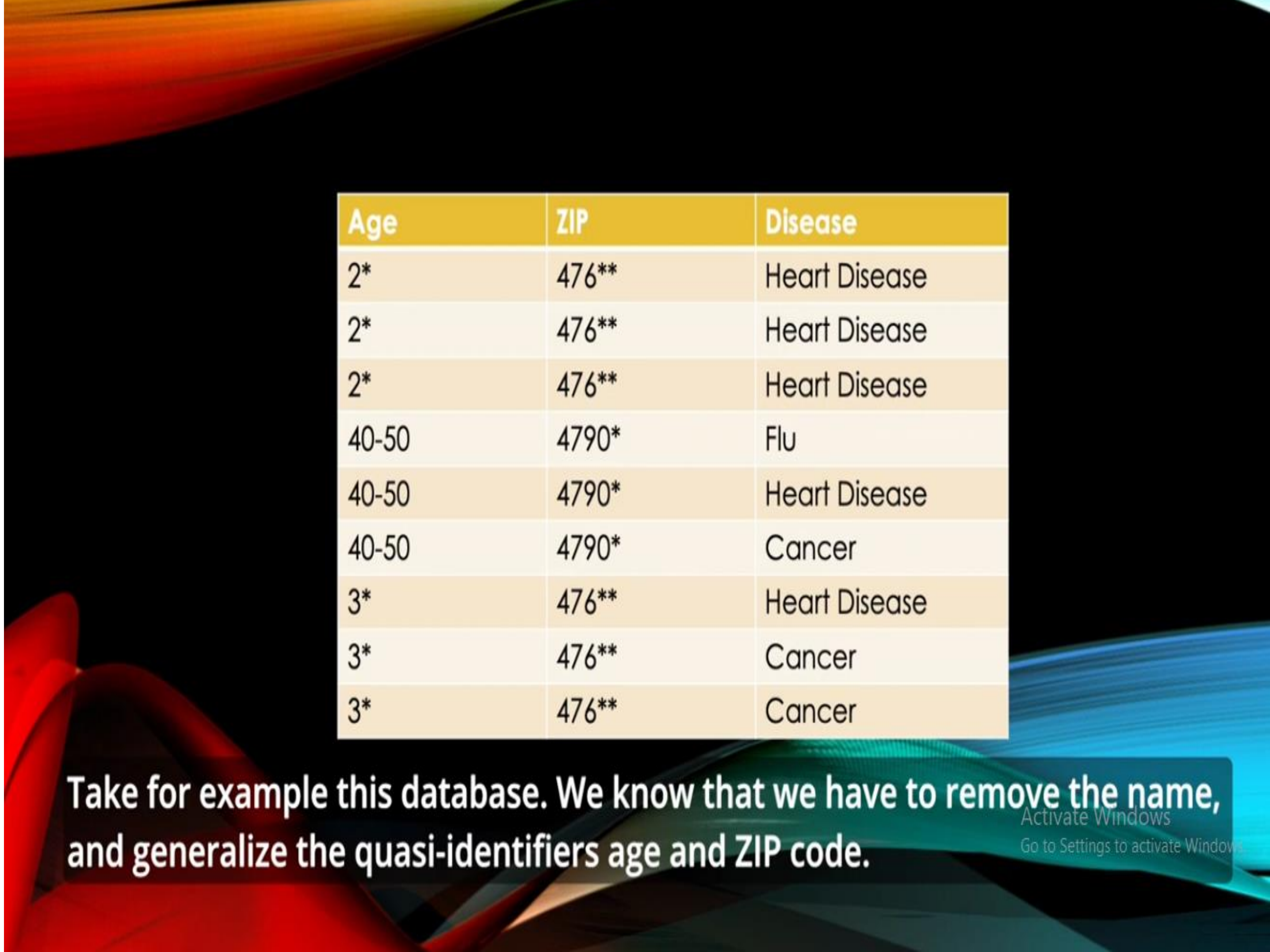
The main problem with k-anonymity lies in the fact that no matter how high your k is, if the data is not diverse, individuals can still be identified.

Activate Windows
Go to Settings to activate Windows

Name	Age	ZIP	Disease
Alice	29	47677	Heart Disease
Bob	22	47602	Heart Disease
Charly	27	47678	Heart Disease
Dave	43	47905	Flu
Eve	52	47909	Heart Disease
Ferris	47	47906	Cancer
George	30	47605	Heart Disease
Harvey	36	47673	Cancer
Iris	32	47607	Cancer

Take for example this database. We know that we have to remove the name, and generalize the quasi-identifiers age and ZIP code.

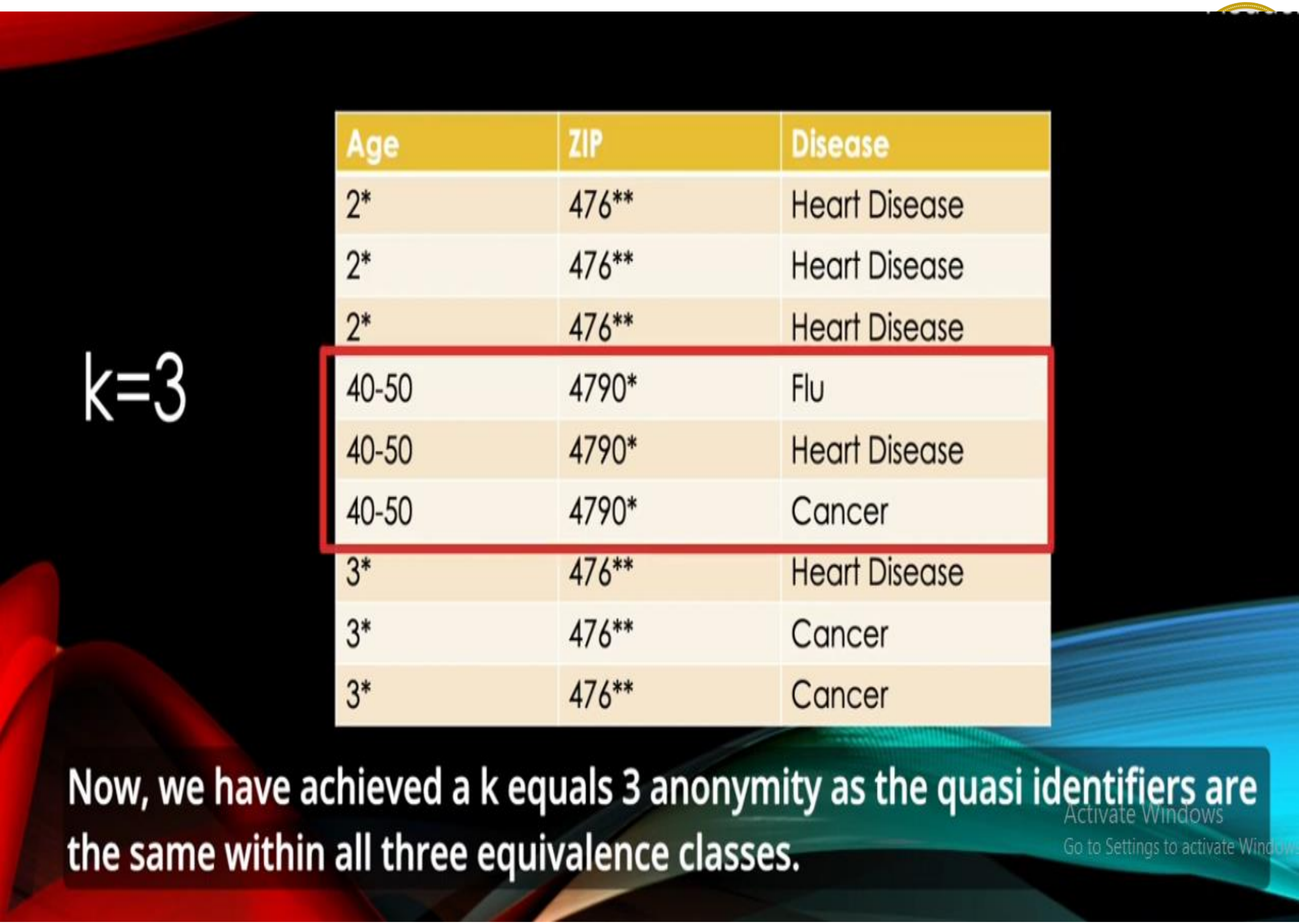
Activate Windows
Go to Settings to activate Windows.



Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

Take for example this database. We know that we have to remove the name, and generalize the quasi-identifiers age and ZIP code.

Activate Windows
Go to Settings to activate Windows.

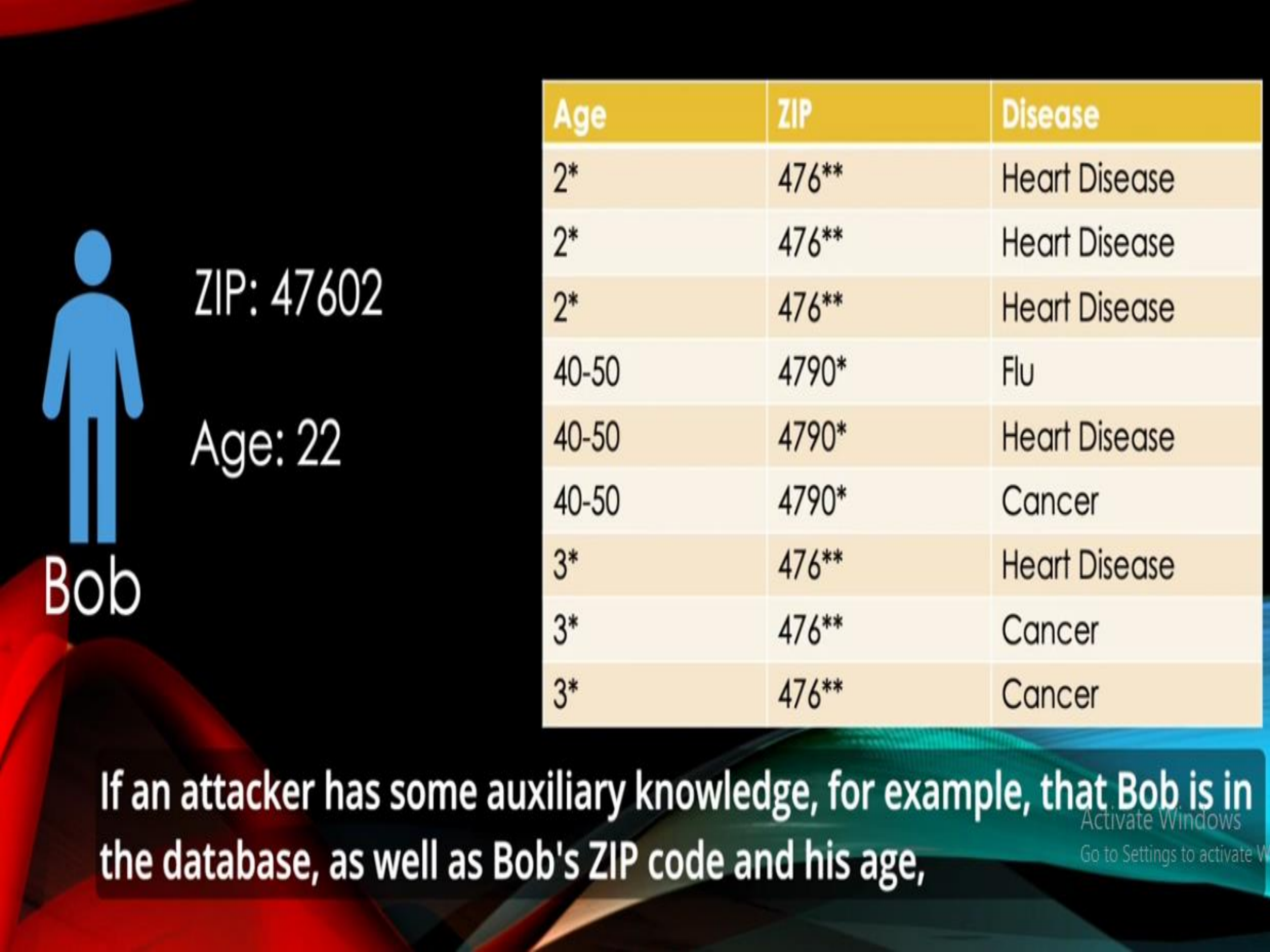


k=3

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

Now, we have achieved a k equals 3 anonymity as the quasi identifiers are the same within all three equivalence classes.

Activate Windows
Go to Settings to activate Windows



ZIP: 47602

Age: 22

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

If an attacker has some auxiliary knowledge, for example, that Bob is in the database, as well as Bob's ZIP code and his age,

Activate Windows
Go to Settings to activate W

ZIP: 47602

Age: 22

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

the attacker still knows that he has a heart disease because it doesn't matter which one of those three entries Bob is, all three have the same disease.

Activate Windows
Go to Settings to activate Windows

ℓ-diversity

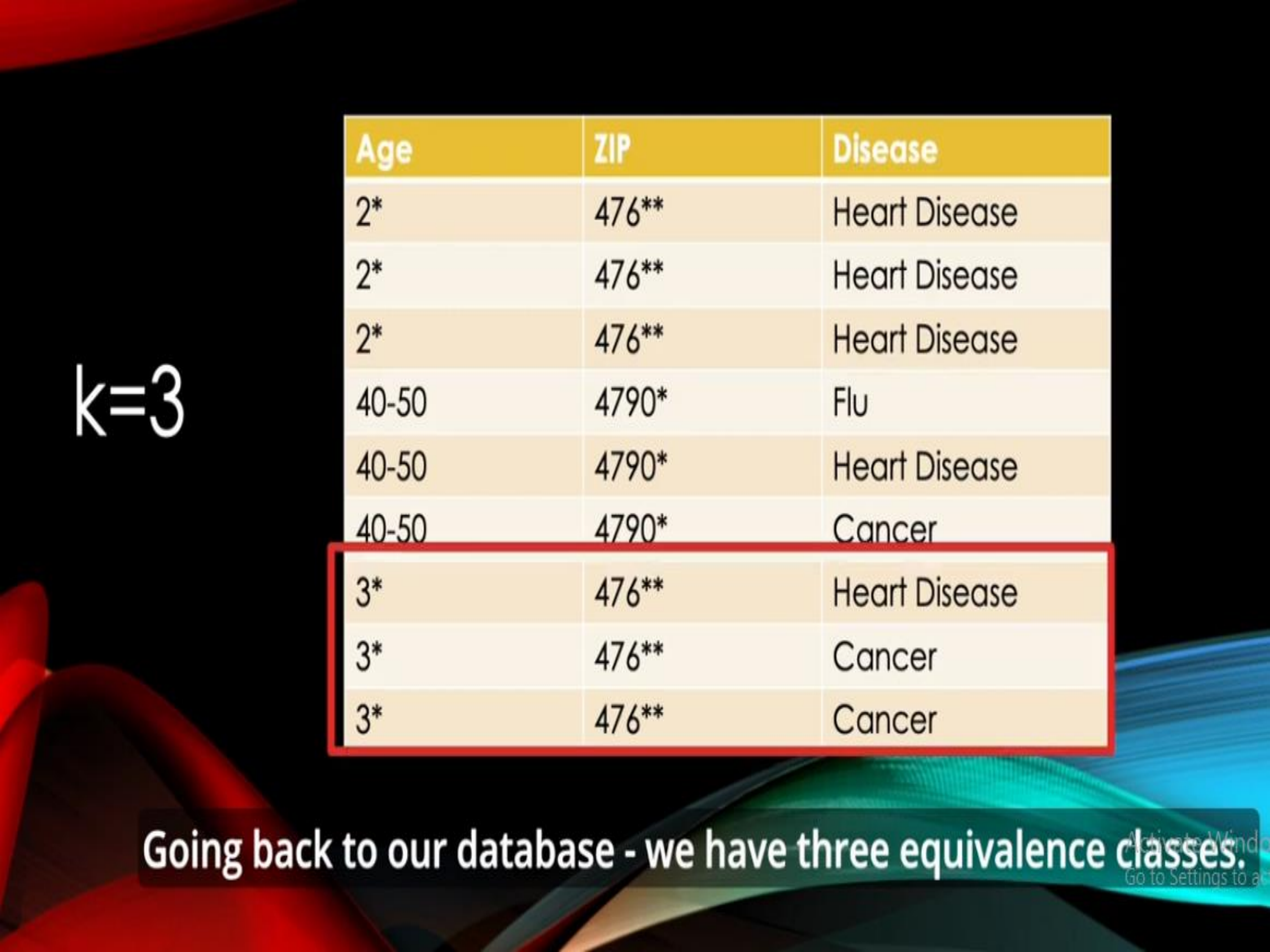
l-Diversity as a concept has been introduced to tackle this problem.

Activate Windows
Go to Settings to activate W

ℓ-diversity

It's definition is that there should be at least L well represented different values in the sensitive attribute field within each equivalence class.

Activate Windows
Go to Settings to activate Windo



k=3

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

Going back to our database - we have three equivalence classes.

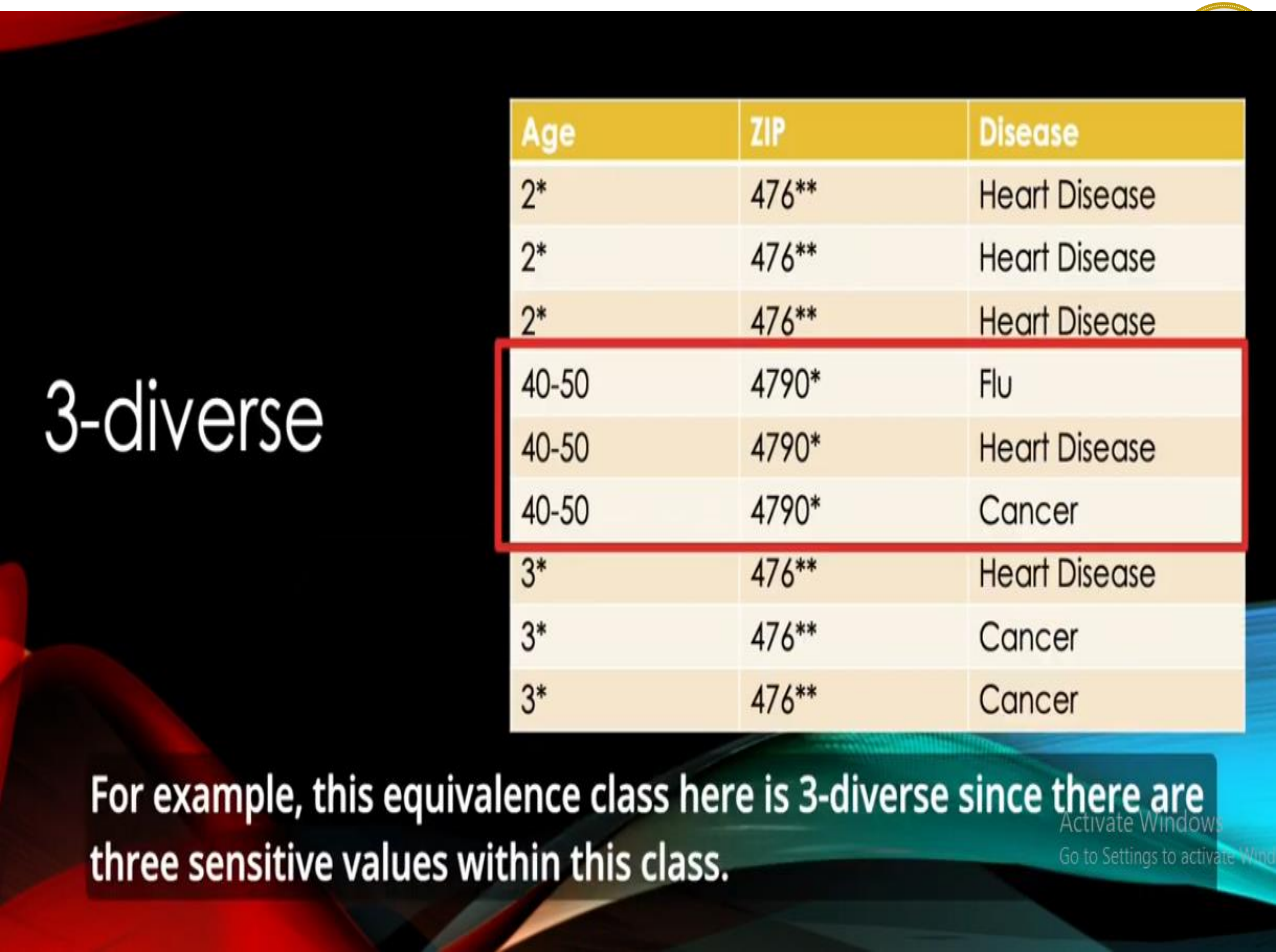
Activate Windo
Go to Settings to a

k=3

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

If we want to achieve, for example, an L equals 2 diversity, we have to have at least two different values within each equivalence class.

Activate Windows
Go to Settings to activate

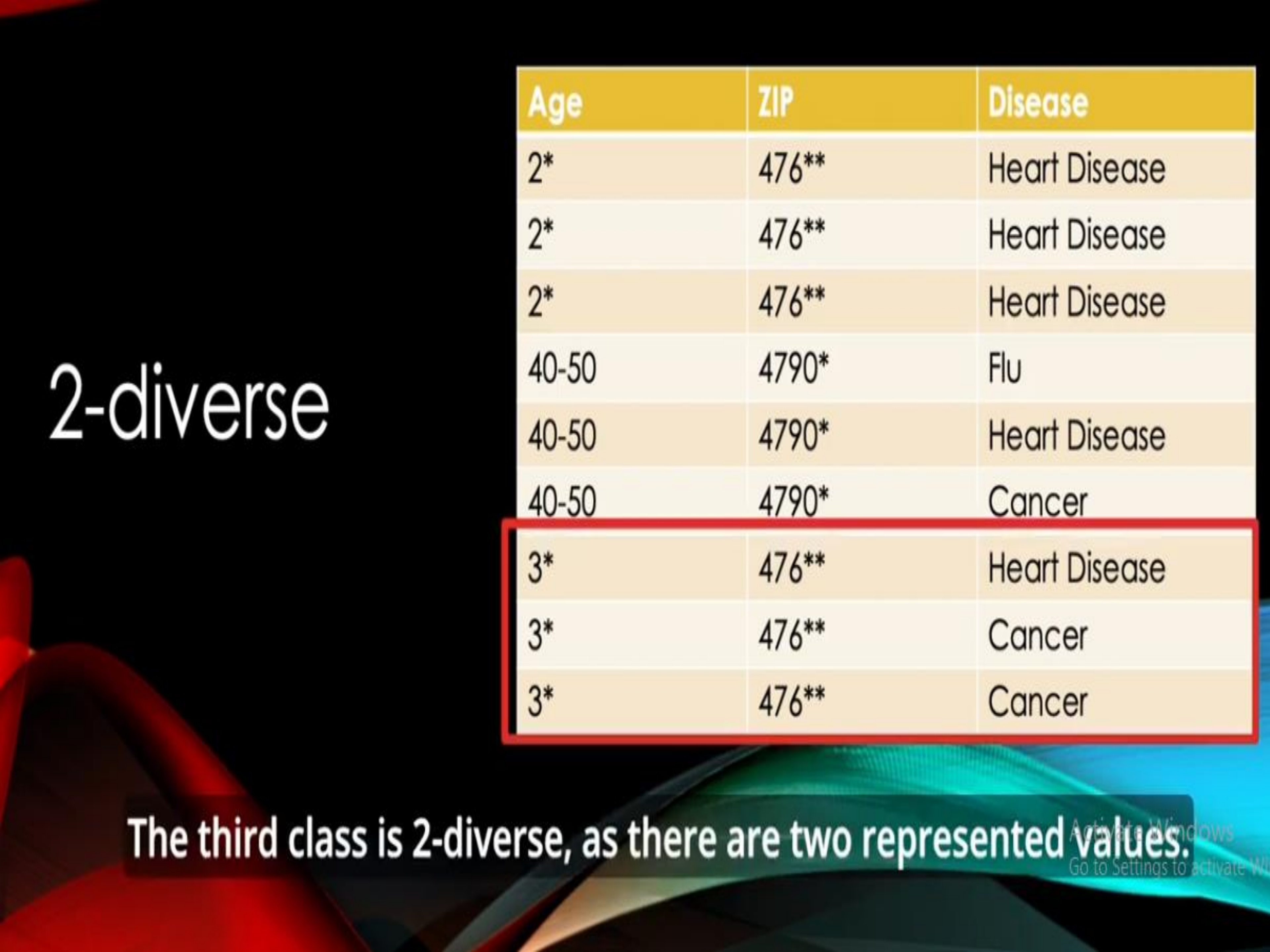


3-diverse

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

For example, this equivalence class here is 3-diverse since there are three sensitive values within this class.

Activate Windows
Go to Settings to activate Wind

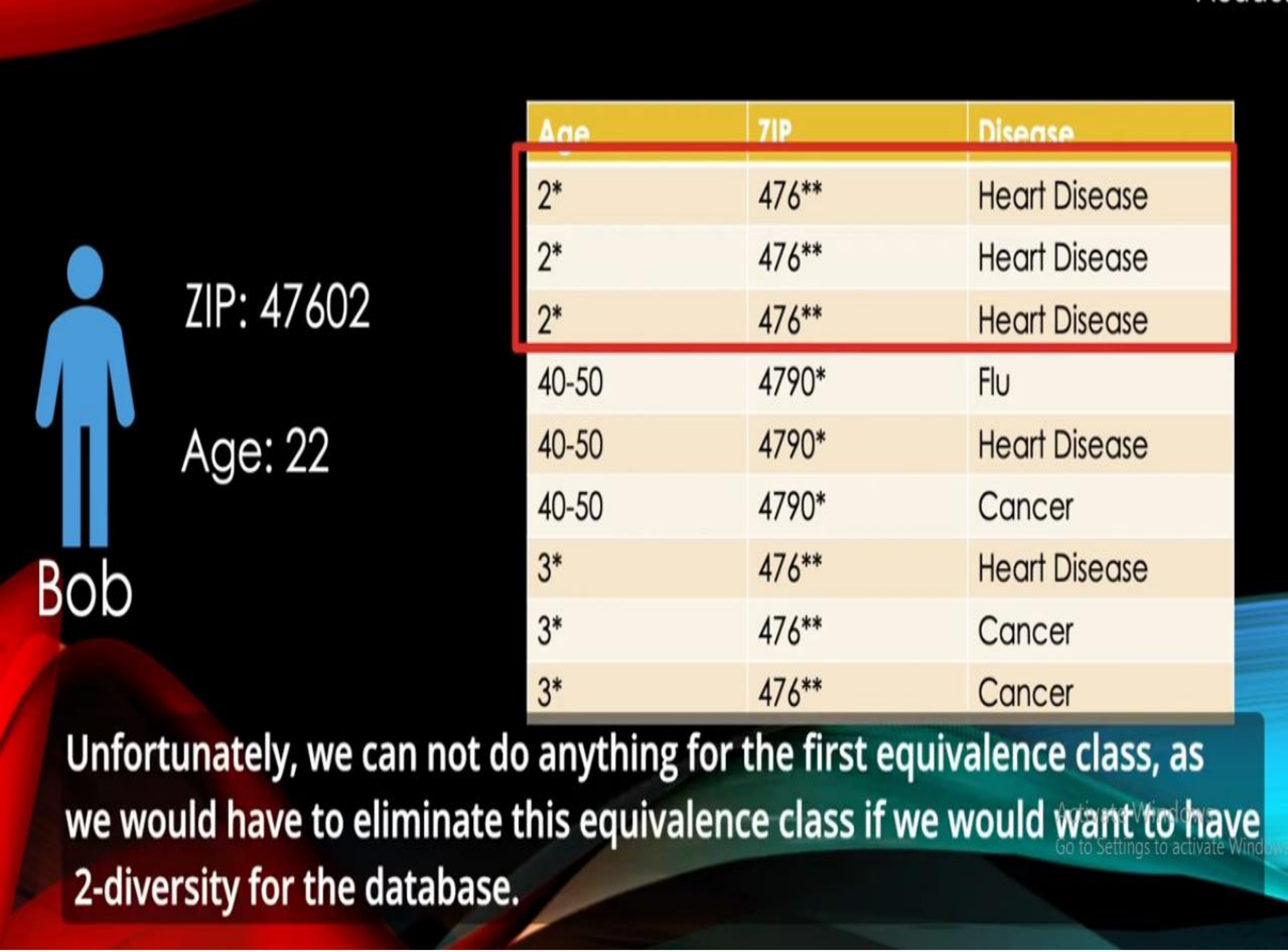


2-diverse

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

The third class is 2-diverse, as there are two represented values.

Activate Windows
Go to Settings to activate Wi



Bob

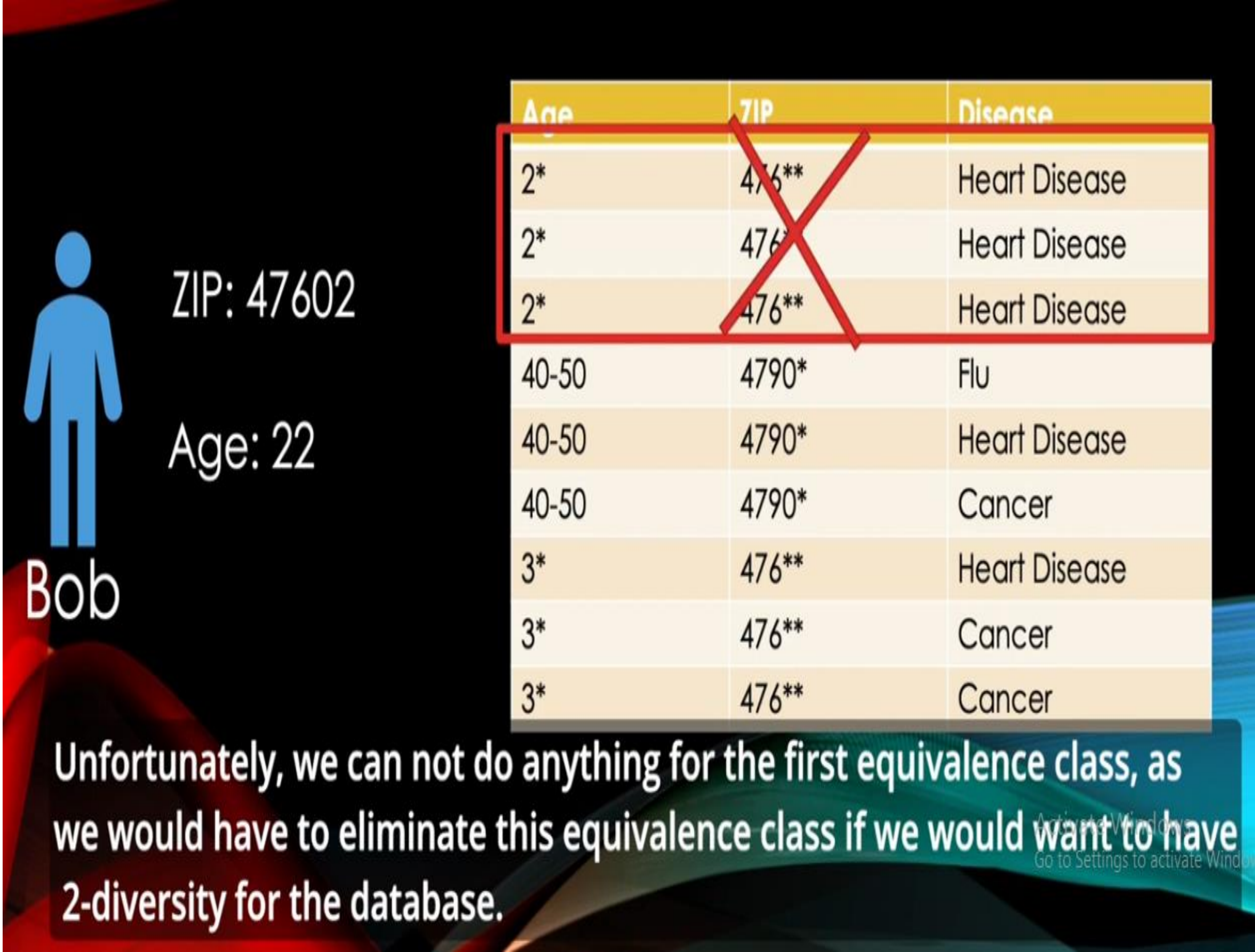
ZIP: 47602

Age: 22

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

Unfortunately, we can not do anything for the first equivalence class, as we would have to eliminate this equivalence class if we would want to have 2-diversity for the database.

Activate Windows
Go to Settings to activate Windows.



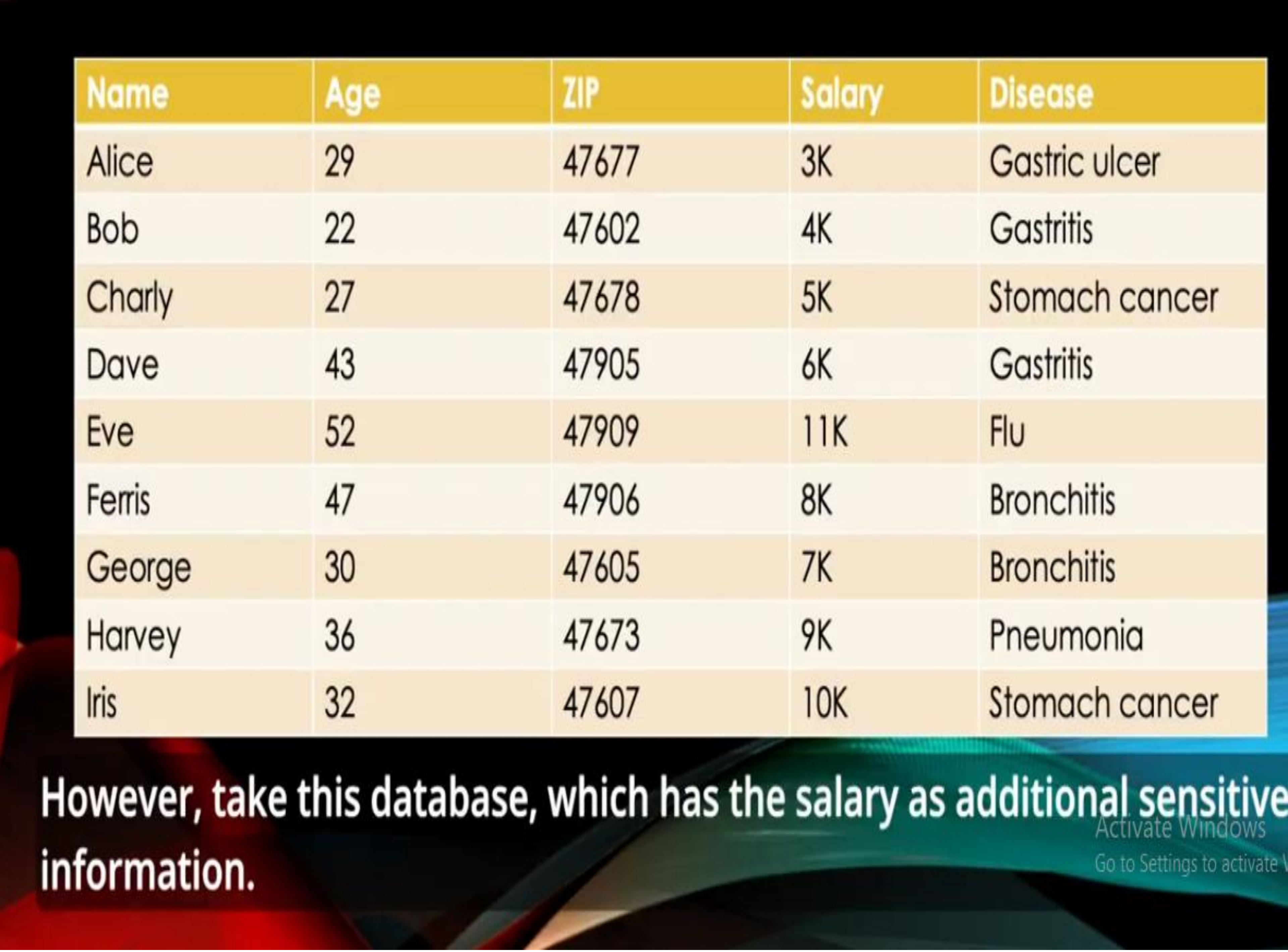
ZIP: 47602

Age: 22

Age	ZIP	Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
2*	476**	Heart Disease
40-50	4790*	Flu
40-50	4790*	Heart Disease
40-50	4790*	Cancer
3*	476**	Heart Disease
3*	476**	Cancer
3*	476**	Cancer

Unfortunately, we can not do anything for the first equivalence class, as we would have to eliminate this equivalence class if we would want to have 2-diversity for the database.

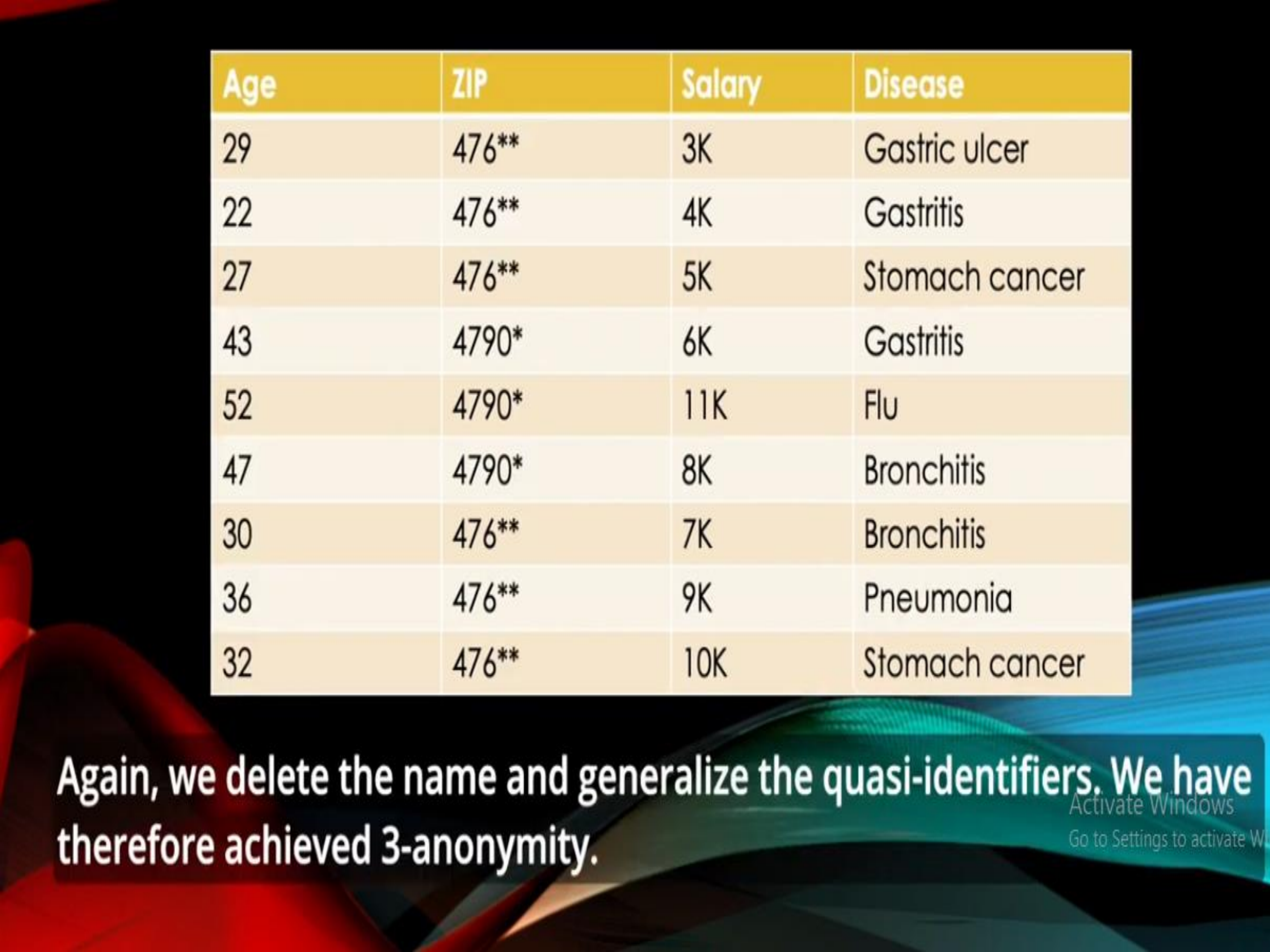
Activate Windows
Go to Settings to activate Windo



Name	Age	ZIP	Salary	Disease
Alice	29	47677	3K	Gastric ulcer
Bob	22	47602	4K	Gastritis
Charly	27	47678	5K	Stomach cancer
Dave	43	47905	6K	Gastritis
Eve	52	47909	11K	Flu
Ferris	47	47906	8K	Bronchitis
George	30	47605	7K	Bronchitis
Harvey	36	47673	9K	Pneumonia
Iris	32	47607	10K	Stomach cancer

However, take this database, which has the salary as additional sensitive information.

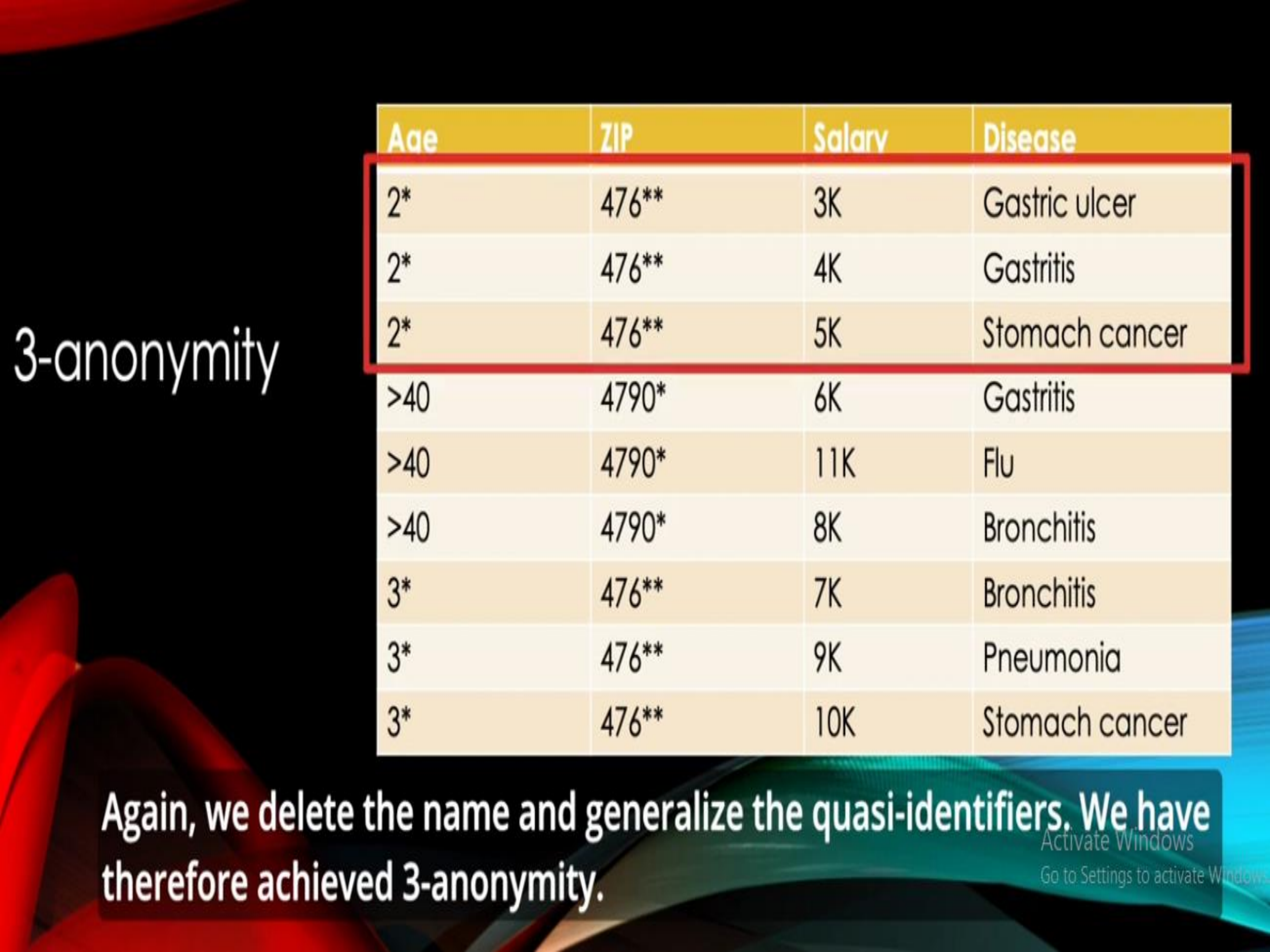
Activate Windows
Go to Settings to activate W



Age	ZIP	Salary	Disease
29	476**	3K	Gastric ulcer
22	476**	4K	Gastritis
27	476**	5K	Stomach cancer
43	4790*	6K	Gastritis
52	4790*	11K	Flu
47	4790*	8K	Bronchitis
30	476**	7K	Bronchitis
36	476**	9K	Pneumonia
32	476**	10K	Stomach cancer

Again, we delete the name and generalize the quasi-identifiers. We have therefore achieved 3-anonymity.

Activate Windows
Go to Settings to activate W



3-anonymity

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

Again, we delete the name and generalize the quasi-identifiers. We have therefore achieved 3-anonymity.

Activate Windows
Go to Settings to activate Windows.

3-diverse

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

We also achieved 3-diversity, as all equivalence classes have three distinct values in both sensitive value fields. Is Bob now safe?

Activate Windows
Go to Settings to activate Windo

ZIP: 47602

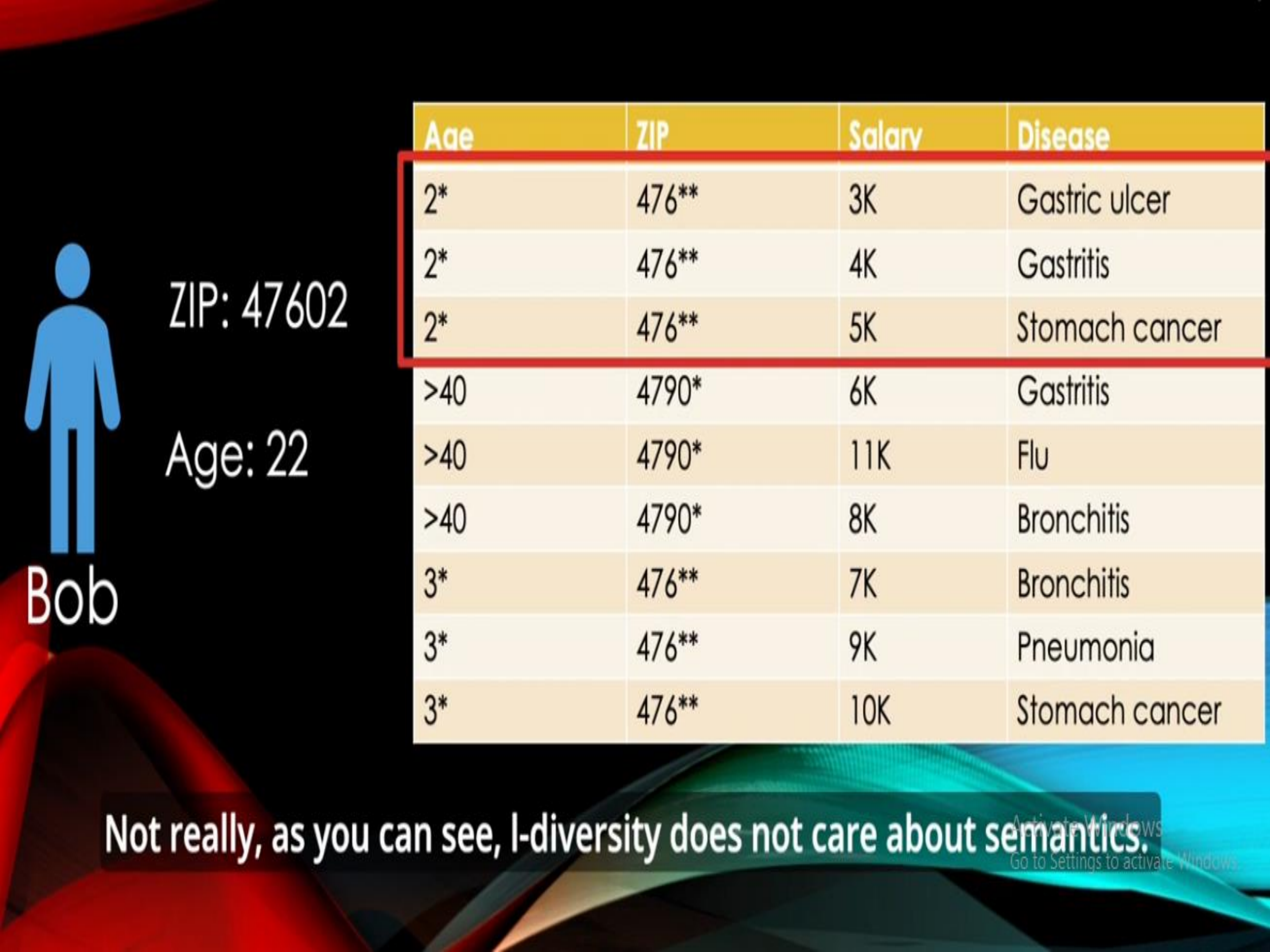
Age: 22

Bob

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

We also achieved 3-diversity, as all equivalence classes have three distinct values in both sensitive value fields. Is Bob now safe?

Activate Windows
Go to Settings to activate Windows.



ZIP: 47602

Age: 22

Bob

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

Not really, as you can see, l-diversity does not care about semantics.

Activate Windows
Go to Settings to activate Windows.

ZIP: 47602

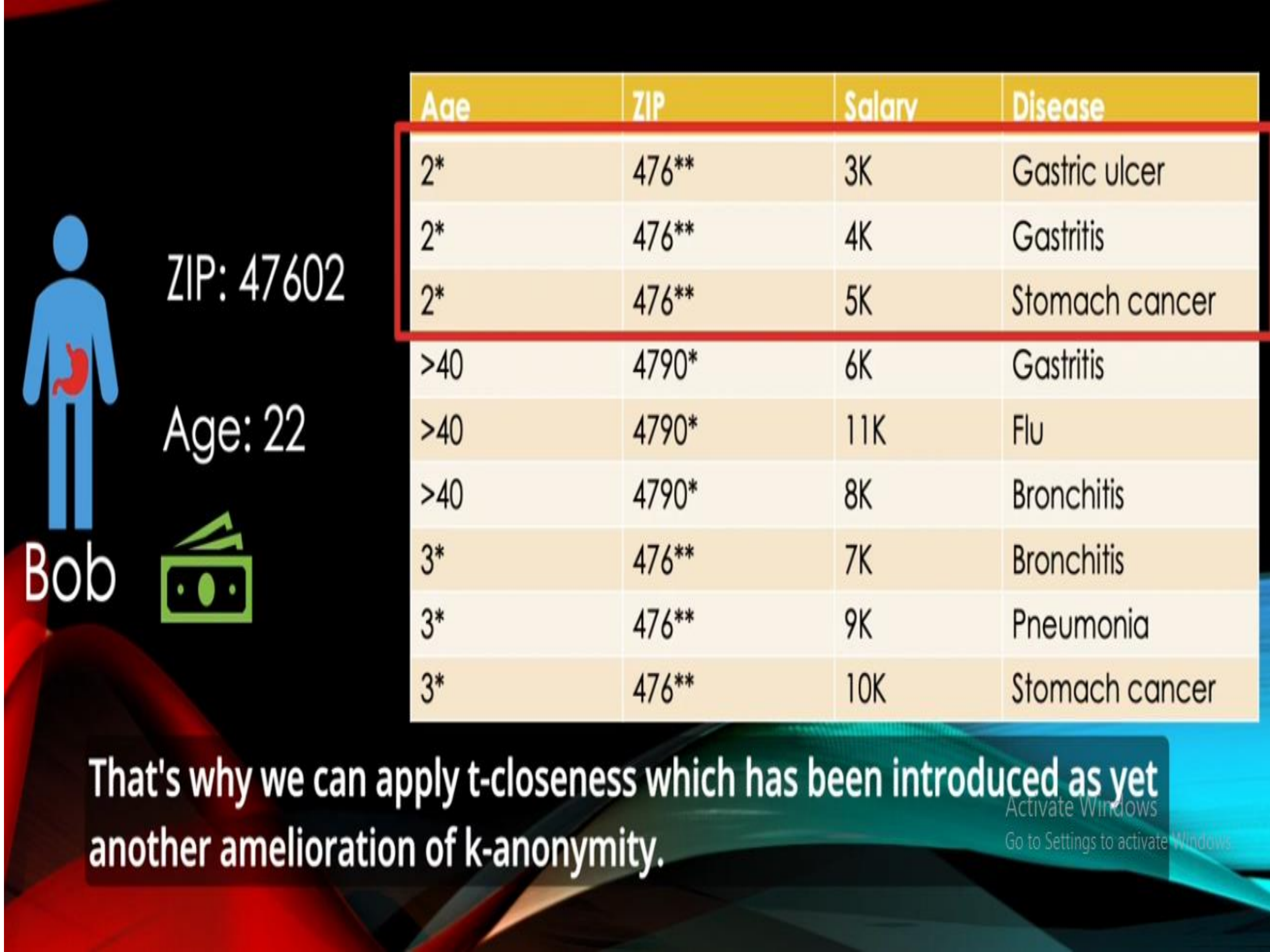
Age: 22

Bob

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

That means, that we might not know which exactly is Bob's disease, but we do know that he has problems with his stomach and that his salary is comparatively low.

Activate Windows
Go to Settings to activate Windows.



Bob

ZIP: 47602

Age: 22

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

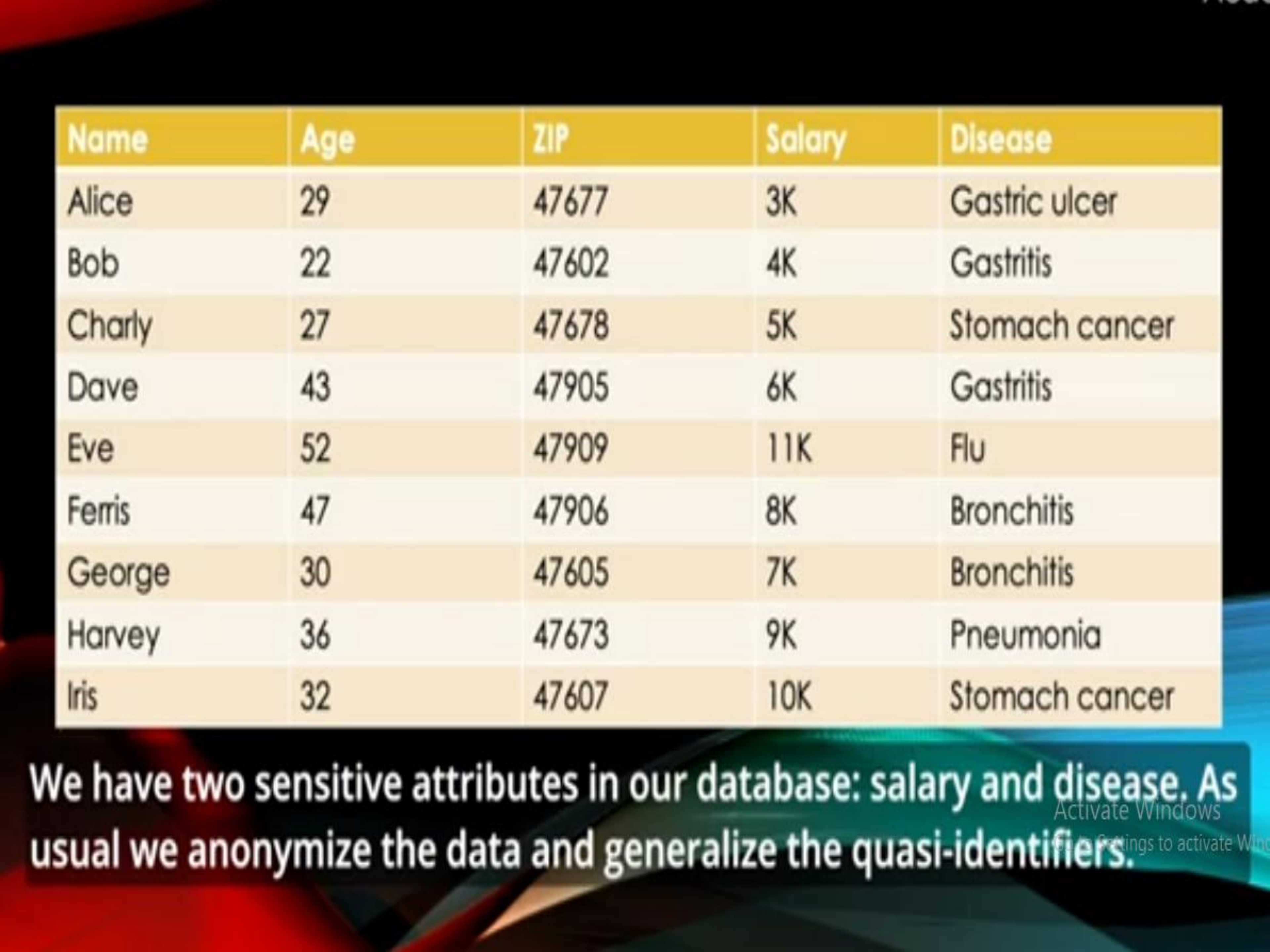
That's why we can apply t-closeness which has been introduced as yet another amelioration of k-anonymity.

Activate Windows
Go to Settings to activate Windows

t-closeness

t-closeness further extends on the concept of k-anonymity by measuring the distance of the distribution of sensitive values between equivalence classes and the original database.

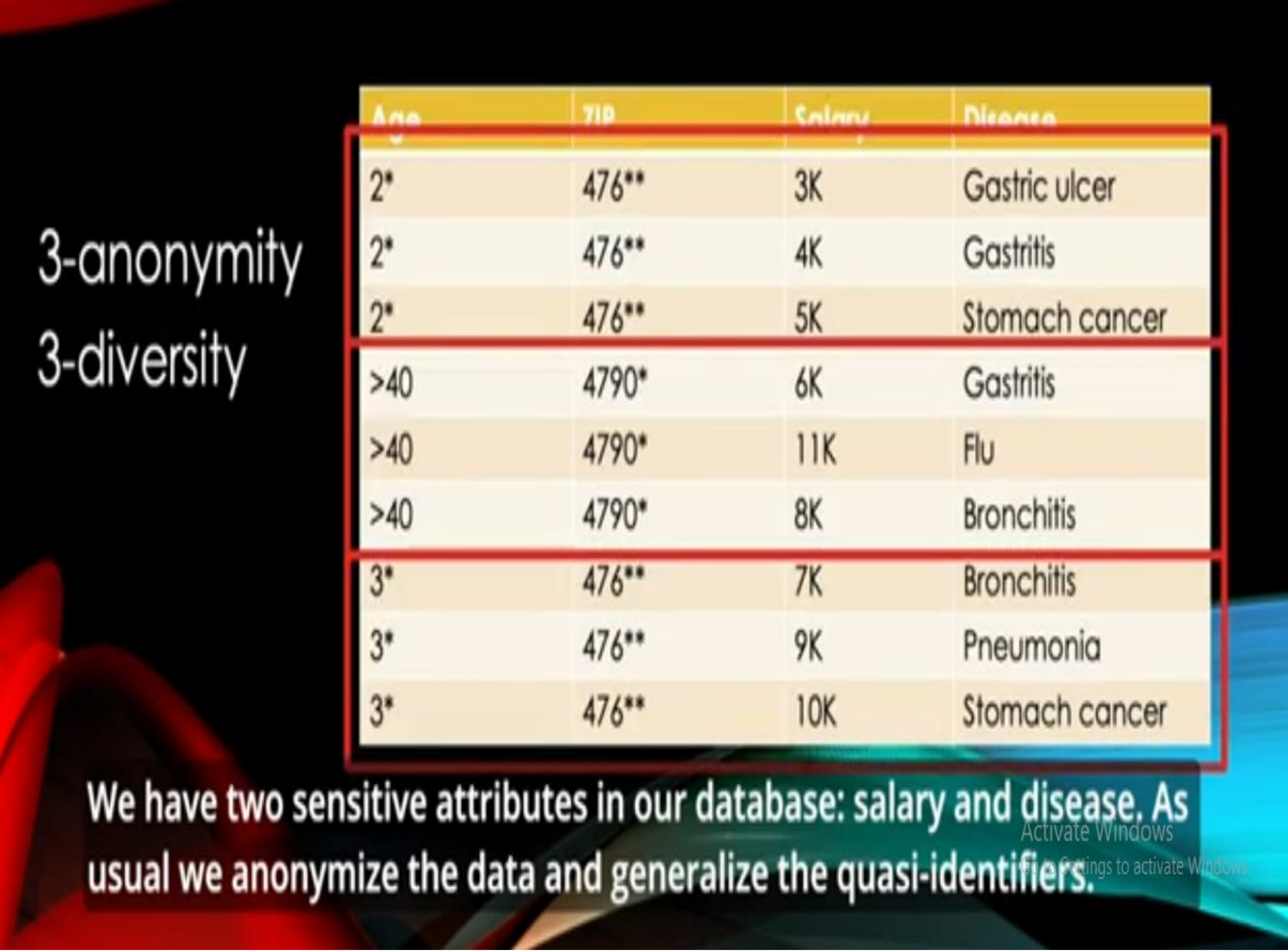
Activate Windows
Go to Settings to activate Windows



Name	Age	ZIP	Salary	Disease
Alice	29	47677	3K	Gastric ulcer
Bob	22	47602	4K	Gastritis
Charly	27	47678	5K	Stomach cancer
Dave	43	47905	6K	Gastritis
Eve	52	47909	11K	Flu
Ferris	47	47906	8K	Bronchitis
George	30	47605	7K	Bronchitis
Harvey	36	47673	9K	Pneumonia
Iris	32	47607	10K	Stomach cancer

We have two sensitive attributes in our database: salary and disease. As usual we anonymize the data and generalize the quasi-identifiers.

Activate Windows
Go to Settings to activate Win

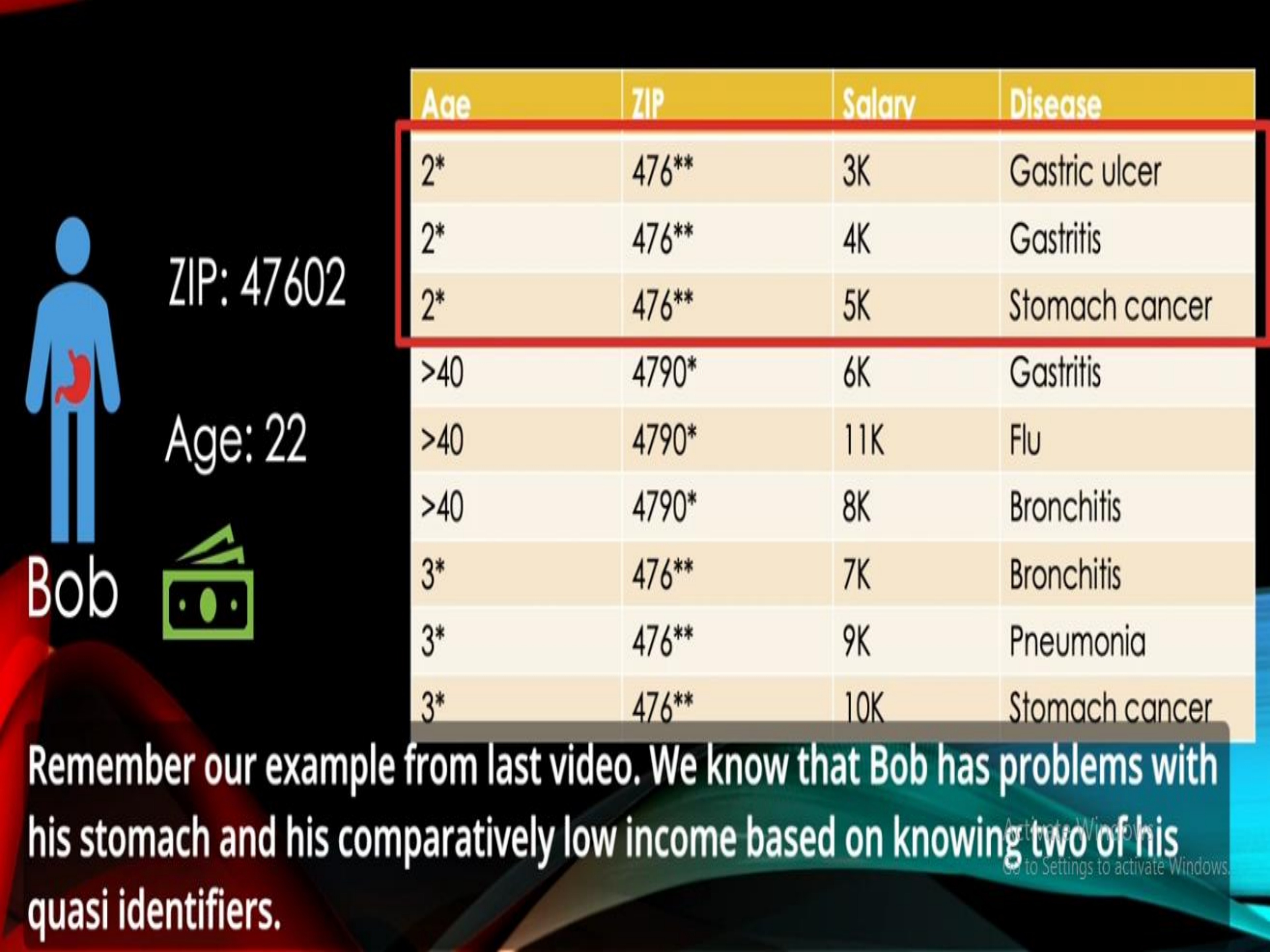


3-anonymity
3-diversity

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

We have two sensitive attributes in our database: salary and disease. As usual we anonymize the data and generalize the quasi-identifiers.

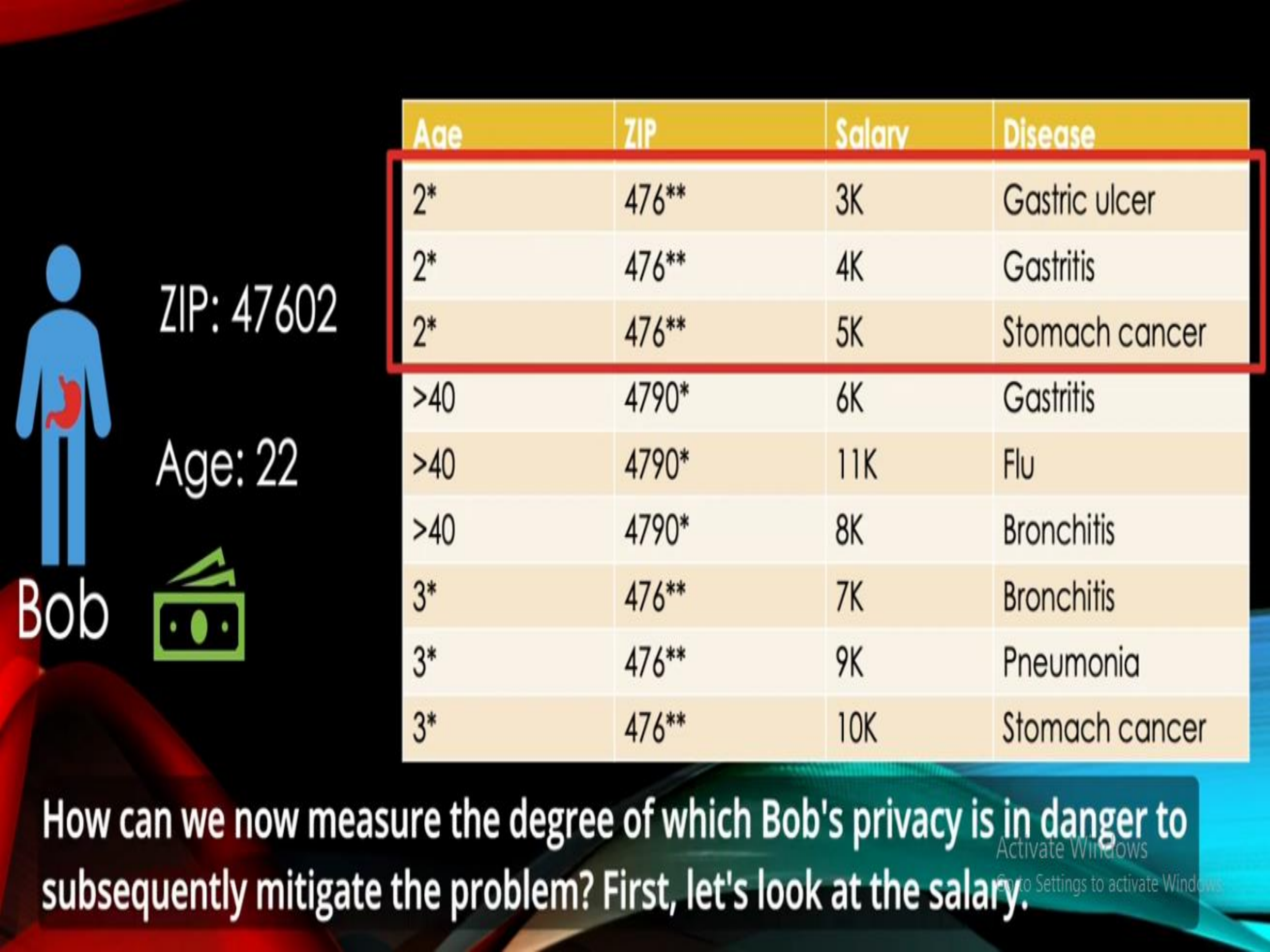
Activate Windows
Go to Settings to activate Windows.



Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

Remember our example from last video. We know that Bob has problems with his stomach and his comparatively low income based on knowing two of his quasi identifiers.

Activate Windows
Go to Settings to activate Windows.



ZIP: 47602

Age: 22

Bob

Age	ZIP	Salary	Disease
2*	476**	3K	Gastric ulcer
2*	476**	4K	Gastritis
2*	476**	5K	Stomach cancer
>40	4790*	6K	Gastritis
>40	4790*	11K	Flu
>40	4790*	8K	Bronchitis
3*	476**	7K	Bronchitis
3*	476**	9K	Pneumonia
3*	476**	10K	Stomach cancer

How can we now measure the degree of which Bob's privacy is in danger to subsequently mitigate the problem? First, let's look at the salary.

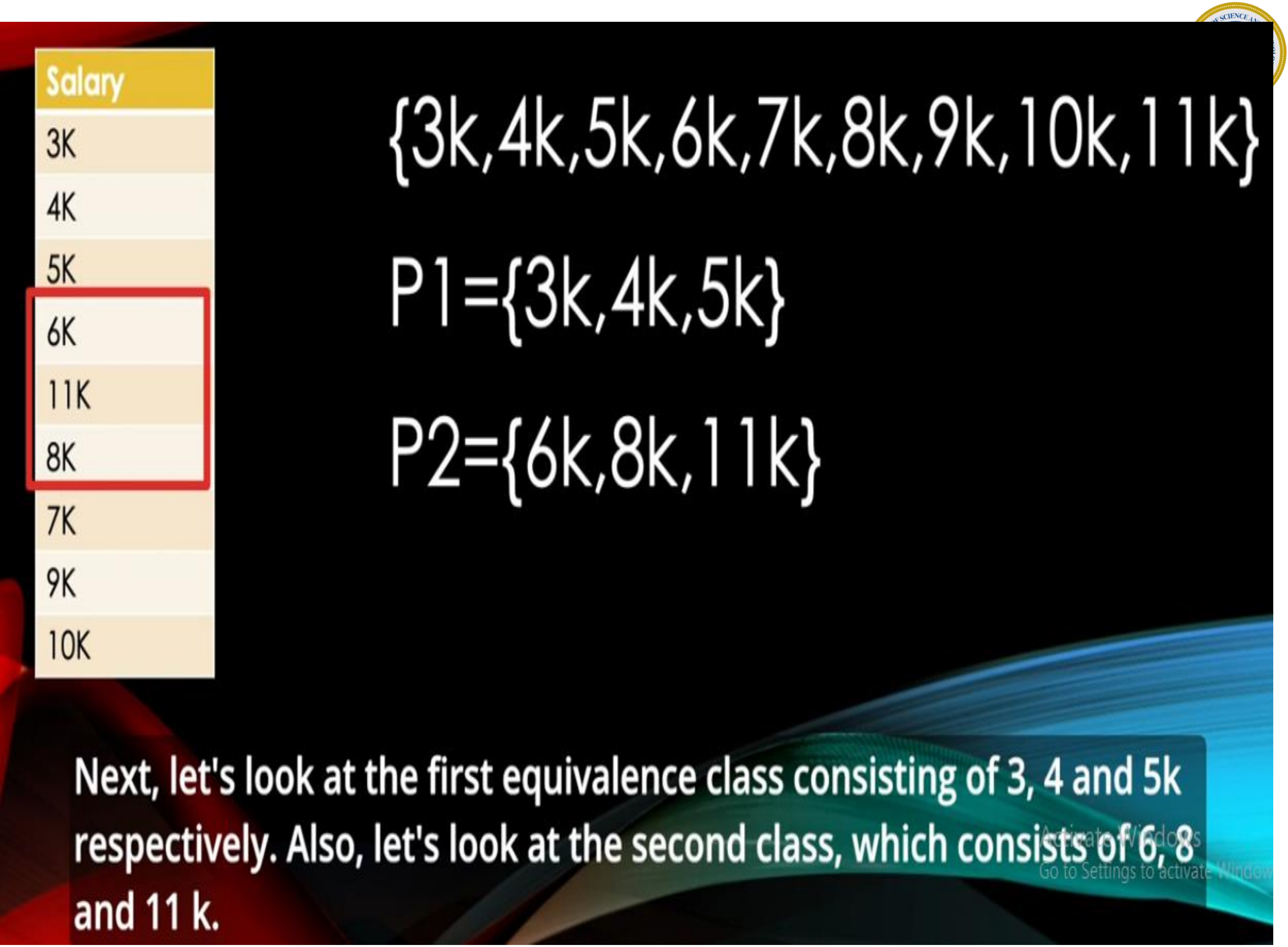
Activate Windows
Go to Settings to activate Windows.

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K

{3k,4k,5k,6k,7k,8k,9k,10k,11k}

Salary is numeric data and as such a bit easier to calculate distances between distributions. We see here the original distribution of the salary, meaning all salaries within the original database.

Activate Windows
Go to Settings to activate Windows.



Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K

{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

P2={6k,8k,11k}

Next, let's look at the first equivalence class consisting of 3, 4 and 5k respectively. Also, let's look at the second class, which consists of 6, 8 and 11 k.

Activate Windows
Go to Settings to activate Windows

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K

{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

P2={6k,8k,11k}

?

Which of these two distributions is now closer to the original one? Let's find out.

Activate Windows
Go to Settings to activate Windows.

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K

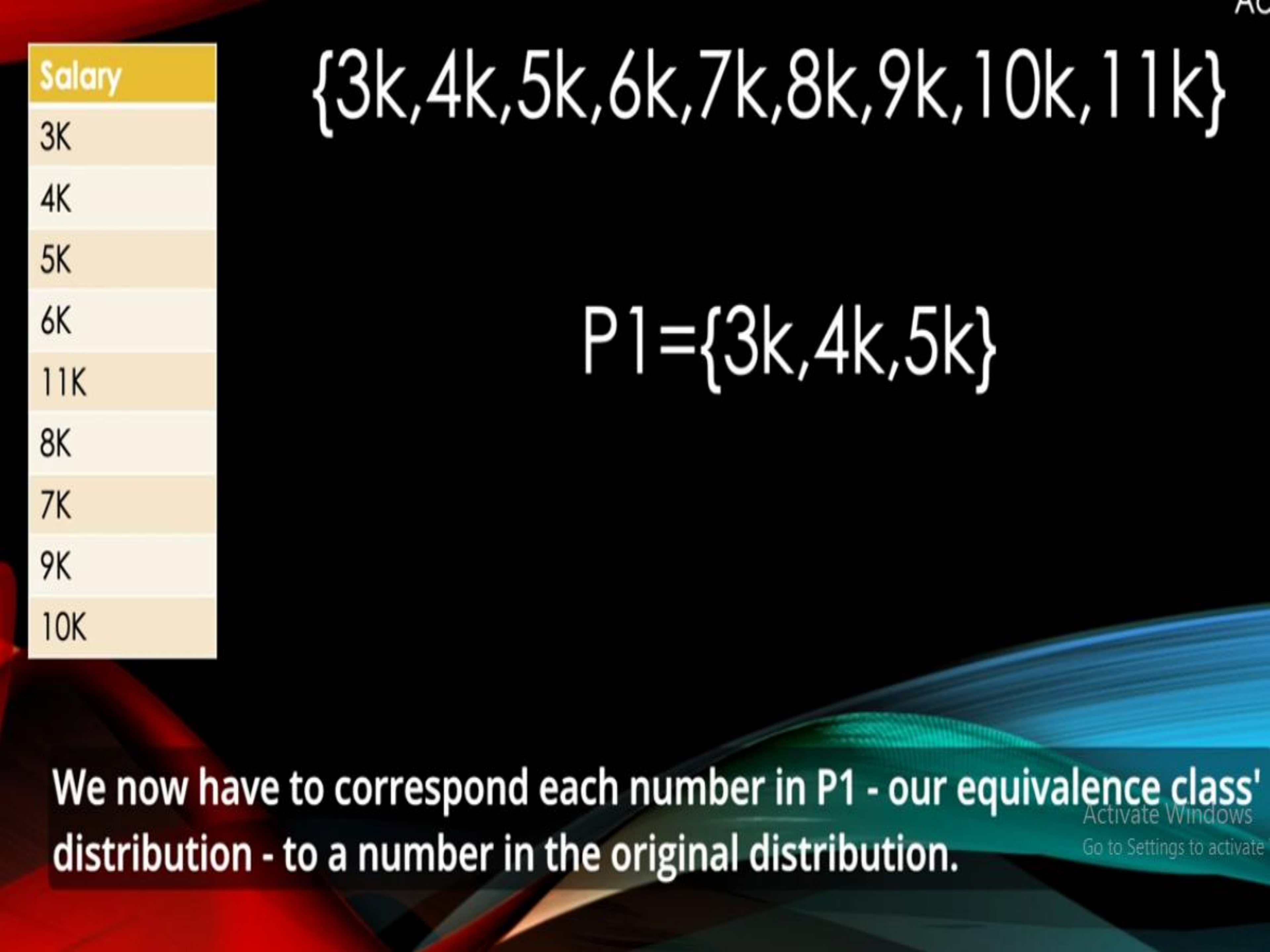
{3k,4k,5k,6k,7k,8k,9k,10k,11k}

Ordered distance D_O

$$D_O = \frac{|i-j|}{n-1}$$

Going back to our example, the earth mover's distance for ordered data, such as numerical data we have in our case, is calculated as i minus j, meaning two objects of two different distributions, divided by the total number of objects in the original distribution minus 1.

Activate Windows
Go to Settings to activate Windows.



Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K

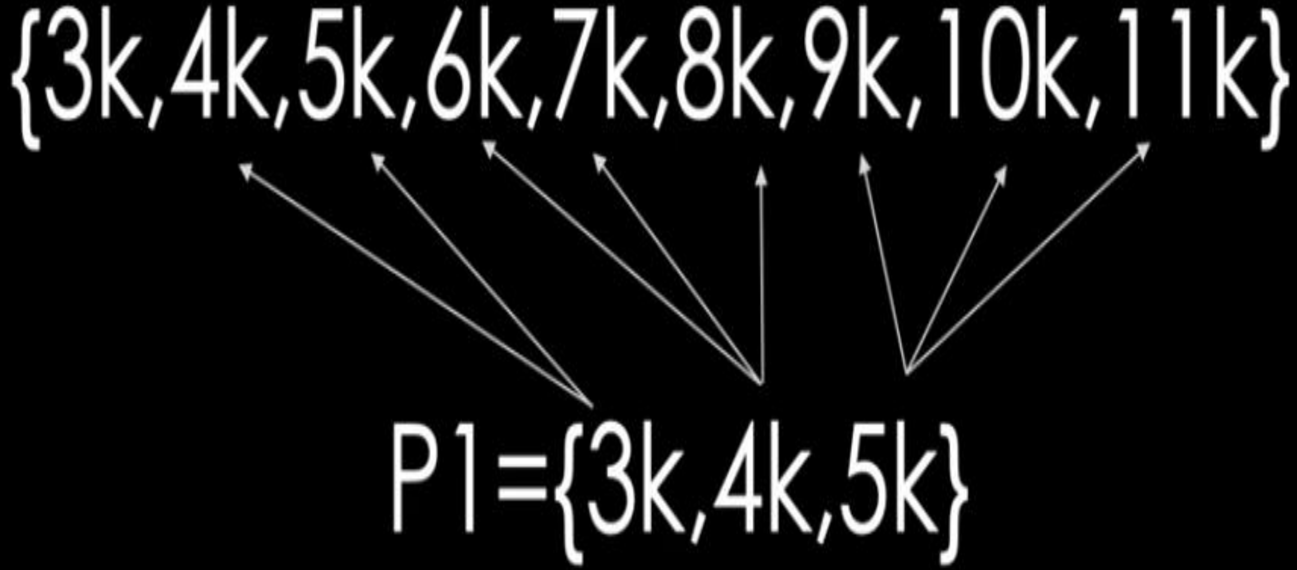
{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

We now have to correspond each number in P1 - our equivalence class' distribution - to a number in the original distribution.

Activate Windows
Go to Settings to activate

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K



{3k,4k,5k,6k,7k,8k,9k,10k,11k}

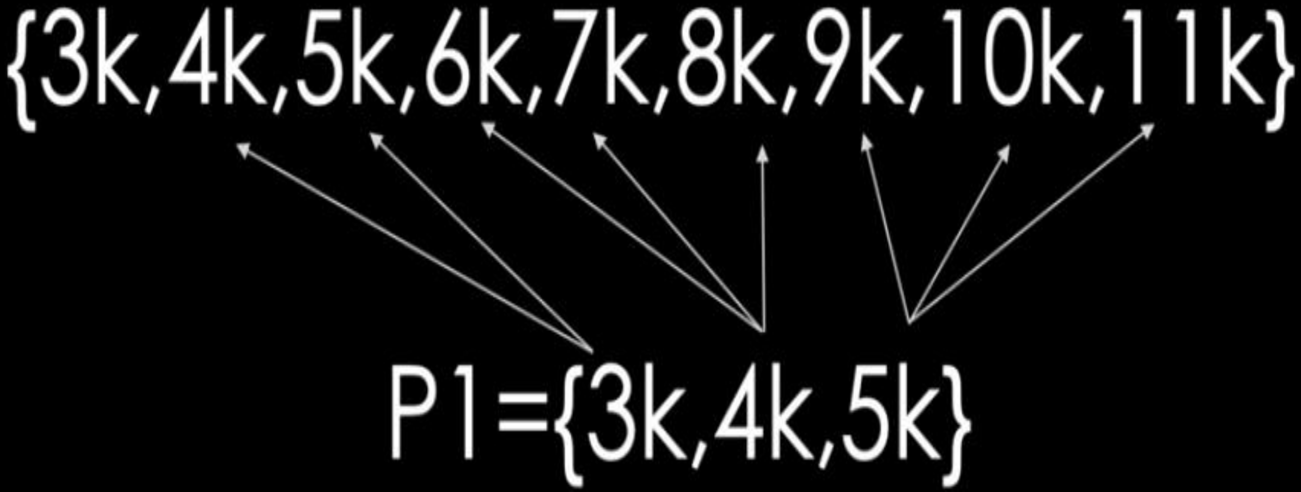
P1={3k,4k,5k}

We can do this at random, but we want to minimize this distance in order to find out the effort required. One minimal mapping is shown here.

Activate Windows
Go to Settings to activate Windows

$$D_O = \frac{|i-j|}{n-1}$$

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K



{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

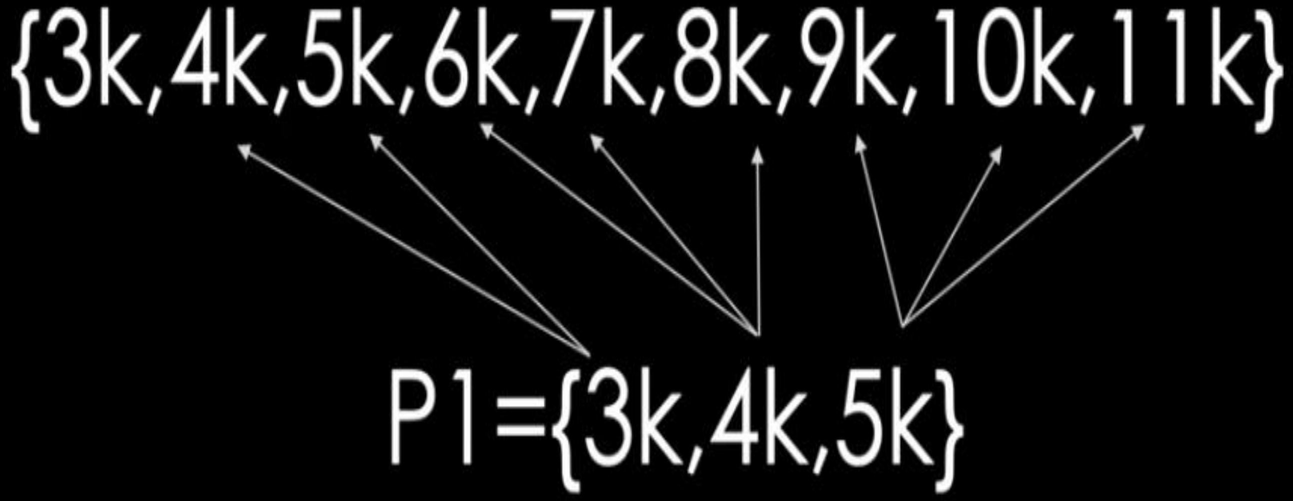
11-5 + 10-5 + 9-5 + 8-4 + 7-4 + 6-4 + 5-3 + 4-3 = 27

Let's start the calculation. We subtract the elements we mapped from each other and arrive at 27.

Activate Windows
Go to Settings to activate Windows

$$D_O = \frac{|i-j|}{n-1}$$

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K



{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

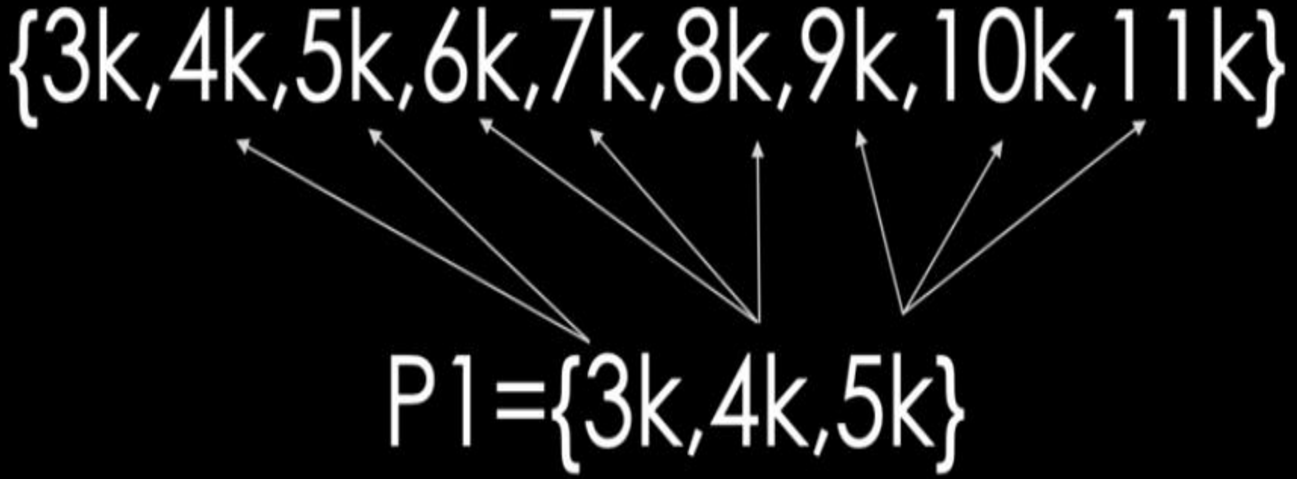
27 / 8 = 3.375

Next, we divide this number by 8 which is the total number of elements minus one.

Activate Windows
Go to Settings to activate Windows.

$$D_O = \frac{|i-j|}{n-1}$$

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K



{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

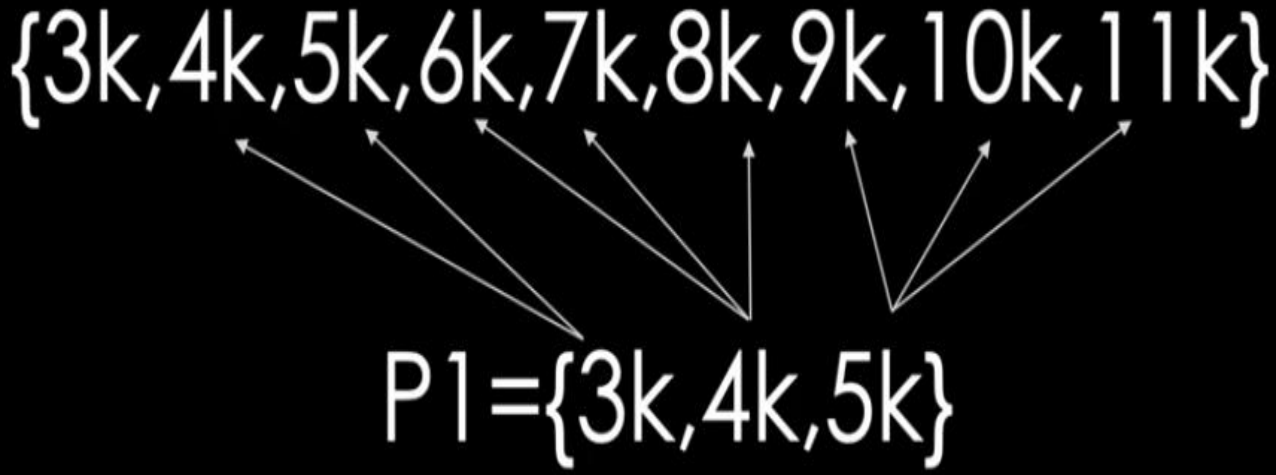
3.375/9 = 0.375

We divide the result again by 9, because we only have to move one ninth probability mass, as each element has the same probability of appearing in the distribution: one ninth.

Activate Windows
Go to Settings to activate Windows

$$D_O = \frac{|i-j|}{n-1}$$

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K



{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

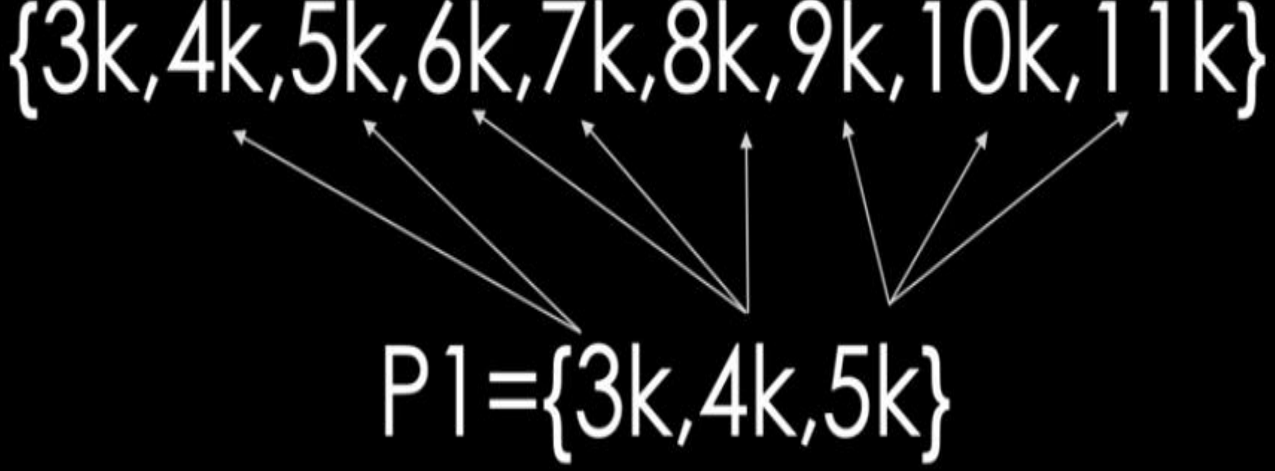
3.375/9 = 0.375

We therefore denote .375 as optimal mass flow between equivalence class P1 and the original distribution. Calculating the same way we reach an optimal mass flow for P2 as 0.167 which is less than P1, meaning the distribution is closer to the original's.

Activate Windows
Go to Settings to activate Windows

$$D_O = \frac{|i-j|}{n-1}$$

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K



{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k}

0.375 = optimal mass flow

We therefore denote .375 as optimal mass flow between equivalence class P1 and the original distribution. Calculating the same way we reach an optimal mass flow for P2 as 0.167 which is less than P1, meaning the distribution is closer to the original's.

Activate Windows
Go to Settings to activate Windows

Salary
3K
4K
5K
6K
11K
8K
7K
9K
10K

{3k,4k,5k,6k,7k,8k,9k,10k,11k}

P1={3k,4k,5k} D[P1,Q]=0.375

P2={6k,8k,11k} D[P2,Q]=0.167

We therefore denote .375 as optimal mass flow between equivalence class P1 and the original distribution. Calculating the same way we reach an optimal mass flow for P2 as 0.167 which is less than P1, meaning the distribution is closer to the original's.

Activate Windows
Go to Settings to activate Windows