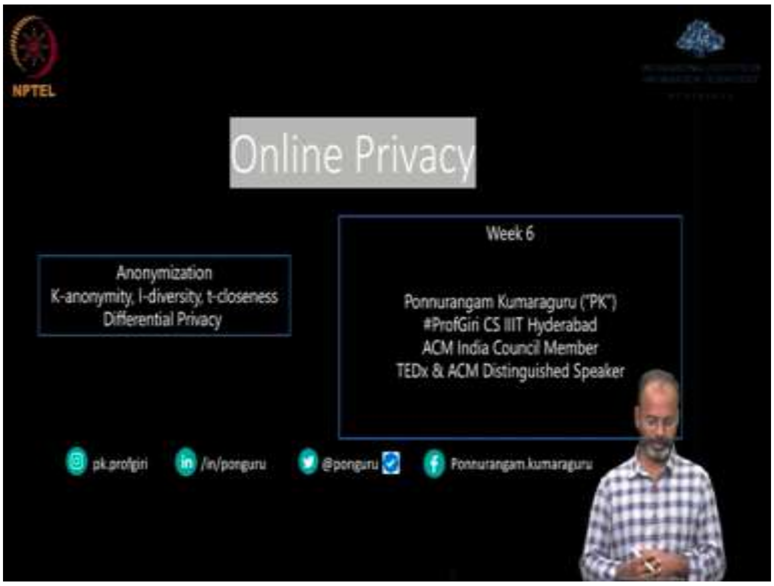
Online Privacy
Professor Ponnurangam Kumaraguru PK
Indian Institute of Technology, Hyderabad
Anonymization techniques and Differential Privacy

Welcome back NPTEL students for week 6, this is almost the middle of the semester for NPTEL course.

(Refer Slide Time: 0:27)



So, we are going to look at a very very important and exciting topic in that sense called anonymization. We will look at importance of anonymization, why do we need anonymization, different methods to do anonymization called K-anonymity, l-diversity, t-closeness and differential privacy, that is what we will cover for week 6.

(Refer Slide Time: 0:53)

So, just to do a quick recap given that it is the middle of the semester. So, what have we covered until now? So, we start off with what is privacy these are basically the text that I took from the first slide that I have something like this and I put it together here. So, we started the semester with what is privacy, why study privacy, fair information practices, OECD guidelines, FTC guidelines.

And then to privacy we saw a paper, Brandon’s paper, contextual integrity, Helen's paper, Helen's work on it, then we looked at what is privacy policy, what are the different components of privacy policy, then privacy enhancing technologies, privacy invasive technologies, social media privacy a little bit of Facebook privacy settings all of that.

Then identity resolution, privacy nudges, cookies and last week we saw cookies and ethics and institutional review board, so that is the topics that we have covered until now. It is quite a lot of topics, but I hope that you are kind of getting a hang of all these topics, how does this privacy come together, what is the complication in actually studying privacy also it is a heavily a multi-disciplinary problem, multi-disciplinary topic, it needs an understanding of different aspects of how things are being done.

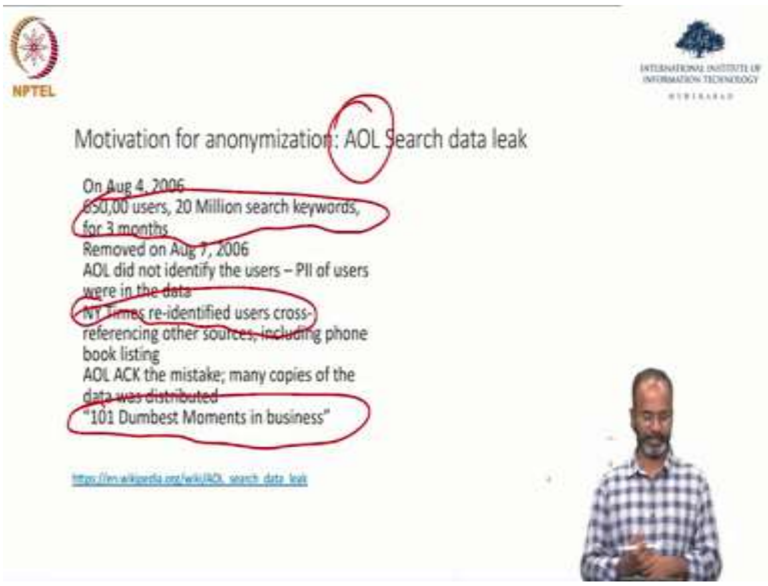
And as always if you have any questions for any topics that we are covering in any of these weeks, feel free to post it on a mailing list. Again, I will try and do some sessions for students to just join and then ask, clarify questions and I am also thinking of actually doing some interactions with the students if you have, if you are doing the projects that I said or if you are doing the activity and you wanted some discussion we can do that too.

(Refer Slide Time: 3:11)

So, anonymization, the word anonymization, I am sure in dictionary meaning you can understand that anonymization is to find a way by which you can suppress some data and share it with people as and when you are sharing. What is the motivation of anonymization becoming an important topic, important topic at different levels in terms of methods that you can do, in terms of companies getting worried about it, in terms of the institutions looking at it as an important topic of research all of that.

So, here is one very sort of say important event that happened which helped anonymization also to become more and more popular, more popular is this AOL search data leak, I have put the link for you to get more details there but here is what happened in AOL search data leak.

(Refer Slide Time: 4:11)



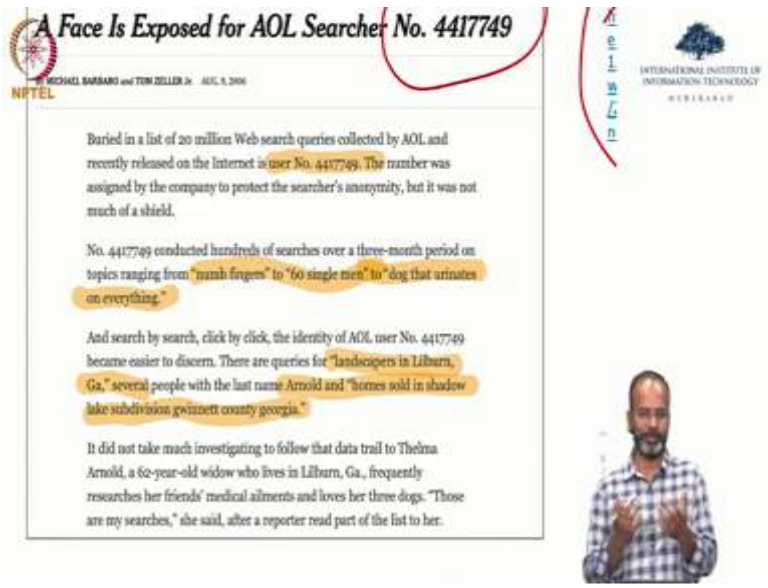
In august 2006, 650,000 users and 20 million search keywords for 3 months AOL released, AOL is this American Online Company which is the, which is the company which used to actually have, give internet access in the initial days and they also had other services in which search was one of them and they shared 650 users data, 20 million search records for 3 months, they made this public on August 4th.

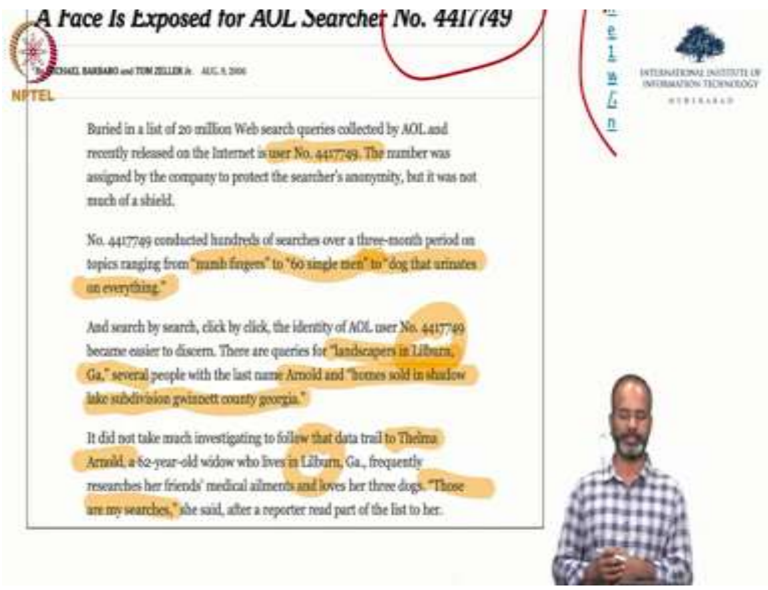
On August 7, 2006 the data was taken down, within three days lots of things happened and I will tell you some of the things that happened, so which will actually motivate the problem of why anonymization is important. AOL did not identify the users, when they made the data public they did not have the data, they did not have details of let us take users like PK on it where somebody could identify that this was PK search results.

Pi of users where in the data, personally identifiable information was in the data, so there was some data that was there which could be used to re-identify users. New York times re-identify users, cross-referencing other sources including phone book listing, this was an important article, in the next slide I have the link to the article also which is the first article which talked about re-identifying people from this AOL data that was made public.

So, when AOL got to know about this that users could be re-identified from the data, they took the data down but interestingly many copies of the data was already circulated, I am sure if you search for the data now you would get some copies of this data lying around somewhere on the internet. And this particular AOL search data that was made public is actually listed as one of the 101 dumbest moments in business ever. So, that is the level of impact that this AOL search data that was made public had.

(Refer Slide Time: 6:23)





And by now you would have realized what ended up happening, there is some data about searches what all search that PK did, what all search you did is in this database, by using external information you could actually re-identify people in the search, in the database, that is what happened.

This is the New York Times article, that is a link to the article, the article is “A face is exposed for AOL search”, this is the search user from the database and they were able to make this user data public and they identified this user and the article was written about this user itself because the user also consented into talking about the data. So, some parts of it would be interesting for us to know from the article also but what kind of thing that came out.

Buried in a list of 20 million web searches, queries collected by AOL and recently released on the Internet is user force, user number 4417749. The number was assigned by the company to protect the searches anonymity, but it was not much as of a shield. So, what did AOL tried to do, AOL basically removed PK’s name and then they replaced it with this number, hoping that, that number may not be re-identified as a PK.

“Number x conducted hundreds of searches over a three-month period of topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything,” just the different types of searches that the user had done. And search by search, click by click the identity of AOL user became easier to discern. There are queries of “Landscapes in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision location Georgia.” Just giving out details of what are the searches that was done.

It did not take much investigation to follow that data trail of Thelma Arnold, that is the user 4417749 that they re-identified. Arnold a 62-year-old widow who lives in Lilburn, you can see the connection of Lilburn here and Lilburn here; frequently researches her friends medical ailments and loves her three dogs. “Those are my searches,” I think that is the reason why this user was made public is because the user also consented into talking about her. I let you read the whole article, I just pulled up some parts of it for conversation in the class.

(Refer Slide Time: 9:21)

In the privacy of her four bedroom home, Ms Arnold searched for the answers to scores of life’s question big and small. How could she buy “school supplies for Iraq children”? What is the “safest place to live”? What is “the best season to visit Italy”? And of course that is her with a dog. Her searches are a catalogue of intentions, curiosity, anxieties and questions.

And you would also realize, we have been, I mean the first week of the class we saw social dilemma and great hack all that, which is clearly this search results that we are talking about here can be re-identified who you are and if Facebook's and Twitter's knows about it, they are able to use this better for your search results, for your recommendation all of personalization, everything. And this is 2006 please keep in mind.

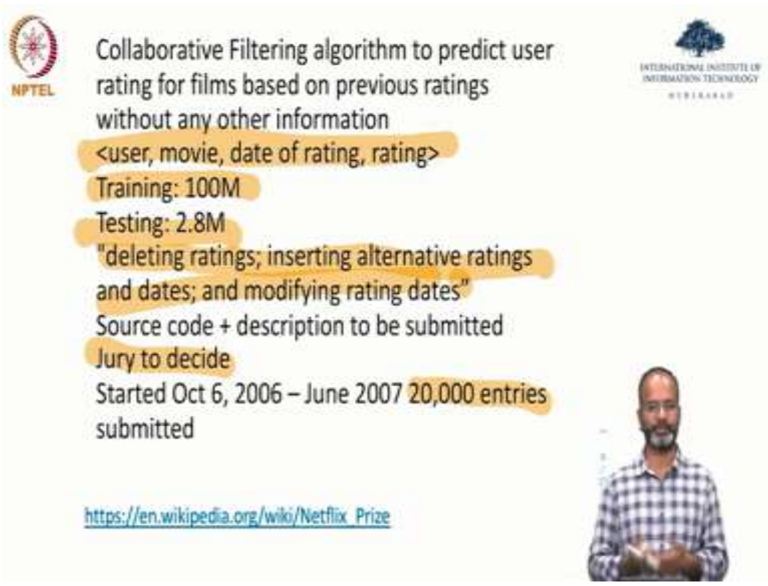
There was a day in May, for example, when she typed “termites,” then “tea for good health,” then “mature living” all within few hours. Her queries mirror million mirror millions of those captured in a AOL’s databases which revealed the concerns of expectant mothers. So, this is basically arguing that what other data that AOL search results had, expectant mothers, cancer patients, college students and music lovers.

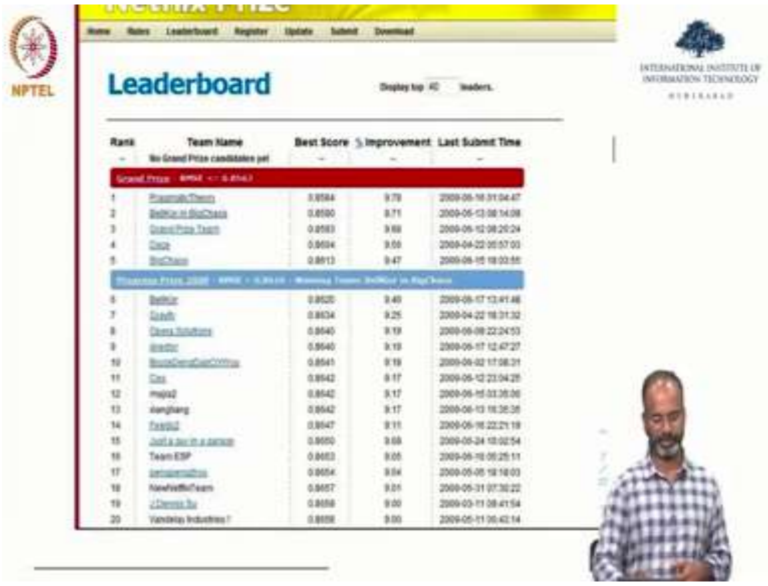
User number 2178 searches for “food to avoid when breastfeeding.” Number this seeks guidance on calorie counting, number x searches for the songs time after time. So, essentially they are just arguing about different users but they re-identified only Thelma, so therefore that is the most of the information is about Thelma. So, that just gives you a sense of what happened with AOL.

(Refer Slide Time: 11:34)

Here is another example, this one was with the Netflix prize. So, I am sure all of you are watching Netflix, the goal here was Netflix said that look we have a recommendation system for ratings of the users, using the ratings of the users we can figure out what recommendations to provide. Can you come and tell us what better recommendations can we actually do?

(Refer Slide Time: 12:05)





Again, pointer I have put for the Netflix Prize but here are some details. So, this is collaborative filtering, so the goal was collaborative filtering algorithm to predict user rating for films based on previous ratings without any other information, the goal was I could give you user rating of, users rating of movies, of the movies that I watched in the past, so can you tell us that what rating the user would give for this particular movie.

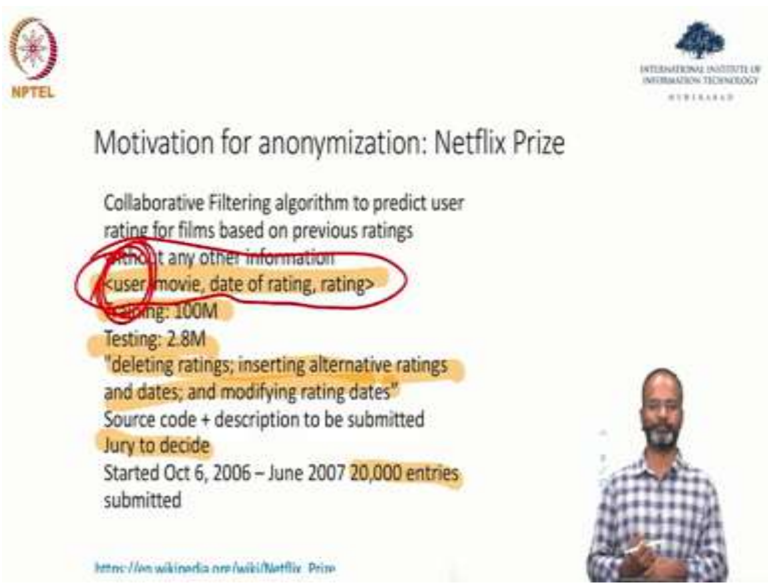
User, movie, date of rating and rating, that is what was shared. So, this this is the key here what data was shared. And this was an open challenge, this was, Netflix was saying that look we are doing but we wanted our filtering process, our recommendation process to be better please come and help us. Actually in in a sense these are very good methods for finding out new solutions, in another word called as also mass collaboration you can think of it, challenges, all that.

What did they do? They gave training data for 100 million, testing data for 2.8 million rows, and deleting ratings, inserting alternative ratings and dates and modified rating dates, when they shared the data this is what they did, they deleted some ratings, inserted alternative ratings, let us take PK saw a movie, movie rating he gave it as 3, but they changed it to 5 and the date of the rating that was let us take November 2021, they changed it to let us take October 2021 and modifying rating date also.

Source code plus description should be submitted and then they set it up in a nice way that they could actually, I mean today I think you do it on Kaggle, there are many platforms that you could do this but again remember this was 2006 again. So, source code and description to be submitted and there was a jury that they had, the jury would decide who was the winner and they were actually having these kind of leaderboards to show who was doing well in the metrics that they had, interesting, very interesting for Netflix to do all this.

Started 2006 and until 2007, June 2007, 20,000 entries were submitted and everybody got interested about this and then this was I think a million dollar price. So, there was money attached to it so people are actually trying to find a better mechanism for this rating.

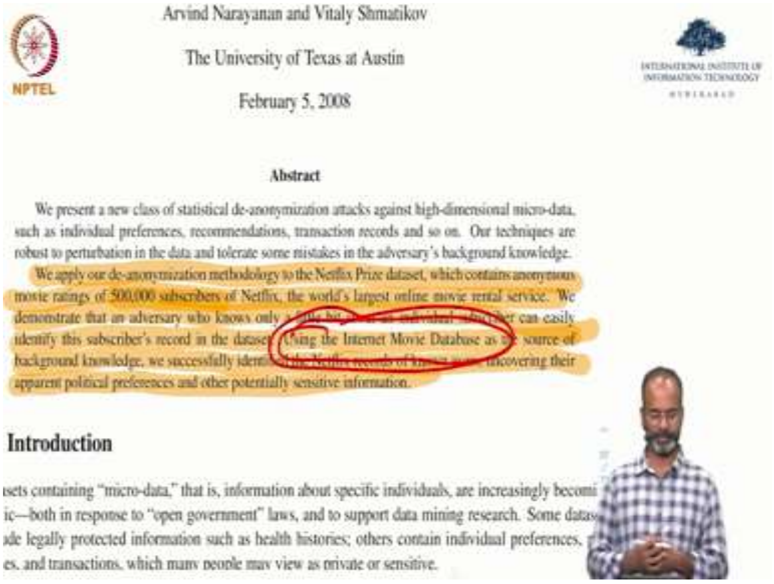
(Refer Slide Time: 14:56)



What happened was, so Arvid Narayanan now at Princeton and earlier at university of Texas at Austin. What he said was look this data is publicly available, can we just go find out who are in this data. Why? Because the user, the data that is shared is this, user movie rating, so can you just re-identify this particular user who said, who is the user, who is giving rating.

Same problem as AOL, where re-identification of the user was happening that is what they tried. So, the paper that they wrote the paper is public I will show you the paper in a second. Robust De-anonymization of large data set, how to break anonymity of the Netflix Prize data set. This paper became very popular and Arvind is doing, Arvind is continuing to do work in the space of privacy, cryptocurrency, all of that.

(Refer Slide Time: 16:07)


Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Introduction

sets containing "micro-data," that is, information about specific individuals, are increasingly becomi
ic—both in response to "open government" laws, and to support data mining research. Some datas
de legally protected information such as health histories; others contain individual preferences,
es, and transactions, which many people may view as private or sensitive.

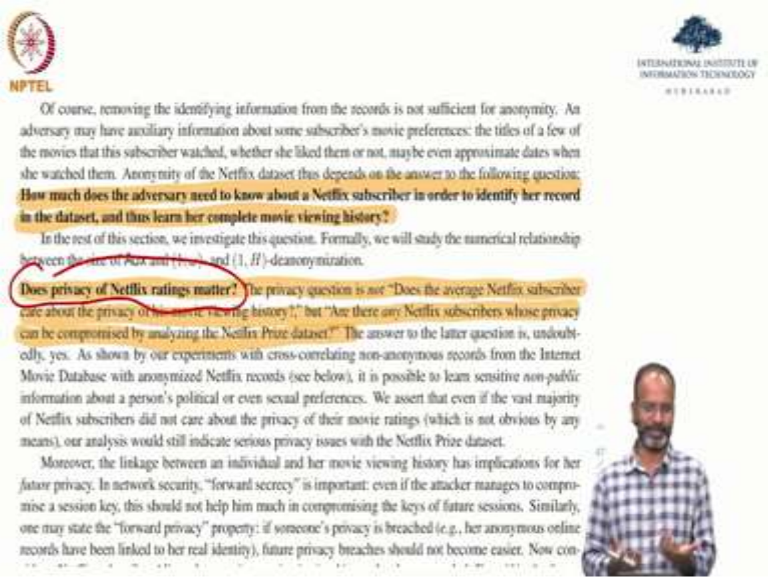
So, this is the paper on Netflix Prize, so I am going to make this sheet one pdf file with all the papers public, so you should be able to get it. So, this is the paper. So, this is what they did. We apply our de-anonymization methodology to the Netflix Prize data which contains anonymous movie ratings so 50,000 subscribers, 500,000 subscribers of Netflix the world's largest online movie rentals service, I think that statement is still true, Netflix is probably the largest rental service right now too.

We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify the subscriber's record in the data set using the internet movie database as a source of background knowledge we successfully identified Netflix records of known users uncovering their apparent political preferences and other potential sensitive information.

So, this is the IMDB database that you may be aware of but if not please go take a look at it, this is all movies ratings that are available with the cast, with the storyline, with reviews of users, everything, so IMDB and the data is also public that you can actually analyse it. So, they used IMDB data with the Netflix publicly made data and then re-identified users, political affiliations and sensitive information that is what they did.

Today I am sure when you think about it, it is probably possible because we have seen these kind of attacks that has happened, we have seen users re-identified in other platforms and context also, but again going back to 2006 it was pretty novel. So, that is the paper I just highlighted some parts that we can, I will show it.

(Refer Slide Time: 18:00)



So, the question that they were interested in asking in looking at this Netflix data was how much does the user advisory need to know about a Netflix subscriber in order to identify a record in the data set and thus learn her complete movie viewing history. So, the goal is that how much, so for example, meaning we should actually do an activity.

For example, one of the activity that I like doing in my class, the class that I teach at IIIT, Hyderabad is that I would ask them to go find my cell number for example, my cell number online, let us do this activity, why do not you also try finding out my cell number, finding out my date of birth, finding out any of these kinds of Aadhaar Card, PAN card number, any sensitive information about me and if you find it please send it to me.

Hopefully you will send it only to me and not to the mailing list. So, it will be, why is this interesting, because you will actually be able to use supplement some information and know that you know I am a faculty at IIIT, Hyderabad, now that you know that I also teach Computer Science, my area of interest is privacy, I live in Hyderabad.

Warlier I used to live in Delhi, all this information that is available you could actually use to find more information about me, so that is the question, how much do you need, how much of this information is necessary to go find out that it is actually PK, instead of some professor in

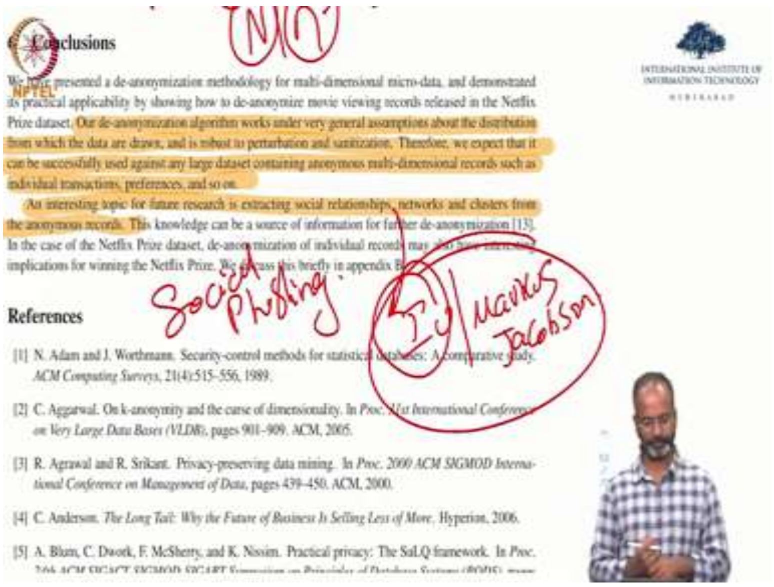
Hyderabad let us take. Thus privacy also, actually this one was also an interesting question because does movie rating even matter?

Meaning I kind of rate some movies, I kind of rate some hotels that I visit, does not even matter with in terms of re-identifying me and how does it affect. The privacy question, does privacy of Netflix ratings matter, the privacy question is not does the average Netflix subscriber care about privacy of his movie viewing history, but are there any Netflix subscribers whose privacy can be compromised by analyzing the Netflix privacy data set.

So, essentially the question is can you actually find out some user's details from this data set that is made public which can be, which can be pretty damaging for that particular user. When we look at k-anonymity I will show you some interesting things that Latonya did, where she kind of got hold of governor's personally identifiable information, sensitive information which helped to actually highlight the problem also.

Because I think when you look at, when you want to make these kind of topics more accessible to people, finding out information about popular celebrities, popular people will help actually people understand than the generic people, the general citizen to understand the problem also better, because otherwise it is an academic exercise. So, one or two more lines at the end.

(Refer Slide Time: 21:30)



So, our de-anonymization algorithm works under very general assumptions about the distribution from which the data are drawn. One of the other critical things that you want to also keep in mind is about this distribution that of the population. So, if you have taken statistics, some quickly statistics this is population, this is sample.

Can you actually derive a sample size from a distribution which is representing the population, population could be 40 percent female, 60 percent male, or employees of a company, 30 percent undergraduates, 40 percent postgraduates and 20 percent Ph.Ds., if that distribution is there if you do a study of the company can you get the samples which are very similar to the population itself, so that is the question, that is the point that is made here.

Our de-anonymization algorithms works under very general assumptions about the distribution from which the data are drawn and is robust to perturbation and sanitized, sanitization, perturbation, sanitization and methods by which you can actually change so to say cells in the data and make the data more anonymous. Therefore, we expect that it can be successfully used against any large data set containing anonymous multi-dimensional records such as individual transactions, preferences and so on.

An interesting topic for future research is extracting social relationships networks and clusters from de-anonymous records. I think this is a very very interesting point, this paper makes and I made a note that this would connect to Indiana university study which a paper called Social Phishing, just do a search for title Social Phishing and Marcus Jacobson, you will get a paper which looked as.

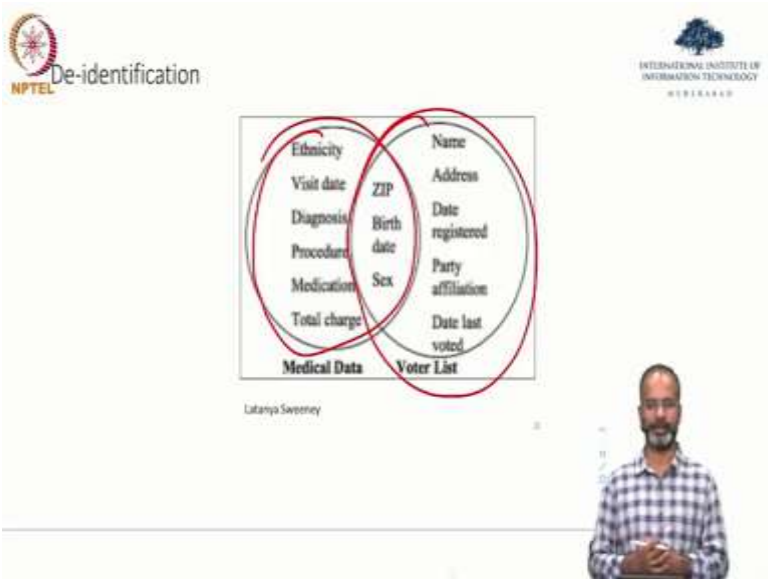
So, this is the paper that we looked at even in the IRB last week's content where we talked about a study, where they collected social information and then send out an email to a participant as though it is coming from somebody they are connected, but in the ethical IRB section I showed you one slide which was this complicated data collection that they did an email sent out, that is the study I am referring back again here.

Social Phishing title, Indiana University and Marcus Jacobson. So, they, this is saying that future research is extracting social relationships and seeing how vulnerable they are, how anonymous people can be in this. Social phishing actually does it to show that if you send out phishing emails as though it is coming from people that I am already connected with, there is a high probability that I will actually click on the links. So, that is about that is about Netflix price privacy concern. These are just motivation for doing anonymization.

(Refer Slide Time: 24:54)

Different methods for anonymization, k-anonymity, L-diversity, T-closeness, differential privacy, there are many many methods for anonymization I am sure if you look at the literature there will be many many techniques. I will focus on some of them because these are also slightly more important and has gained a lot of attention over a period of time, so these may be important for you to know the spectrum of what are the anonymization methods.

(Refer Slide Time: 25:27)

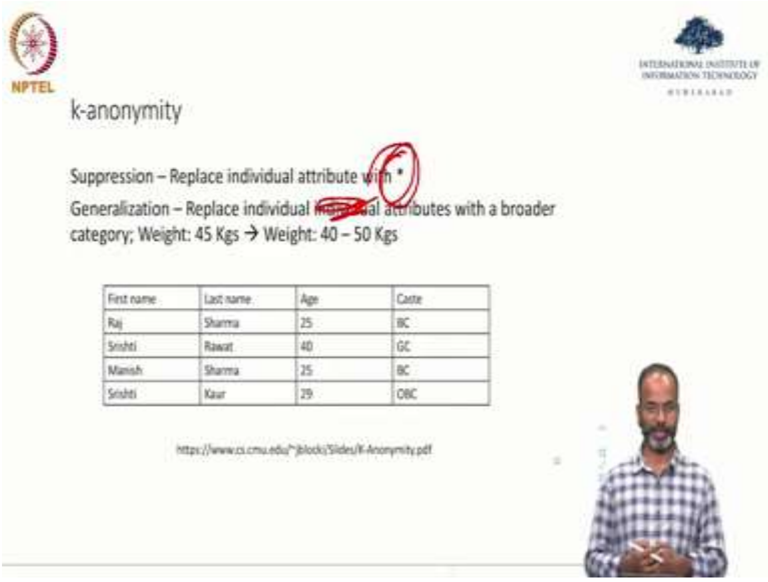


De-identification, first one is this whole idea done by Latanya Sweeney, where she took medical records and she took voter records, put them together and de-identified people. Medical records had all these details, ethnicity, visit day, diagnosis procedure, medication total charge, ZIP, birth date and gender.

Name, address, date registered, party affiliation, date last voted, ZIP, birth date and gender was again available in this voter id list. And both of these data she was able to collect it either publicly or sending a request to government agencies saying that I would like this data and they gave this data for Latanya to do this analysis.

Again if there is of any interest to you, you should figure out, you should try and understand how this could be done in the Indian context itself, look at how you can put some data together that is either publicly available or you can curate it from online sources and put them together to re-identify people.

(Refer Slide Time: 26:41)

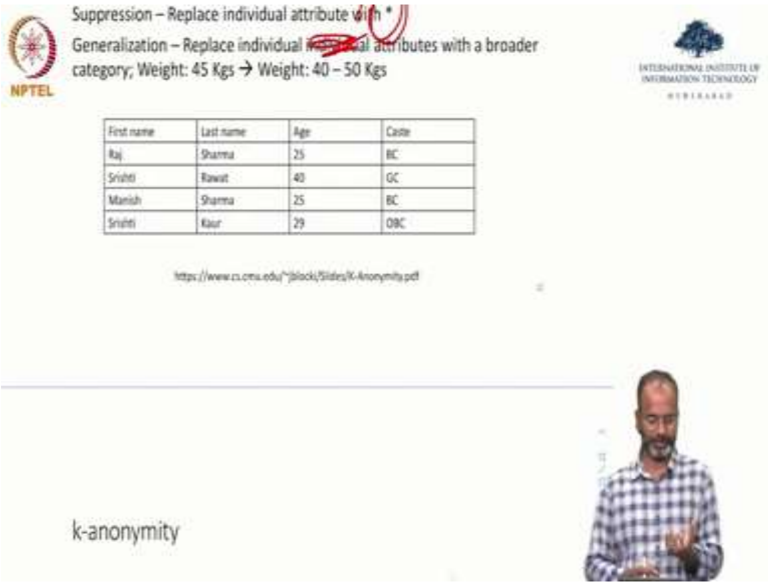


K-anonymity was a phenomenal work at that point in time because this was the first method which looked at how to actually anonymize the data set when you are making the data public. And the goal for anonymization is that you want to share the data for I mean I think from week 1 I have been motivating this, you want to share some data more publicly and more publicly or to a analytics company or to a start-up.

You want to make sure that they do not infer anything from the data that you are giving it to them or at least they do not identify users. One method is suppression, replace individual attribute with a star, meaning this could be anything, star is just an example. The other method is generalization which is replacing individual attributes, individual attributes with the broader category which is if let us take if one's weight is 45 kgs instead of suppressing it with an asterisk you make it that the weight is between 40 and 50 kgs.

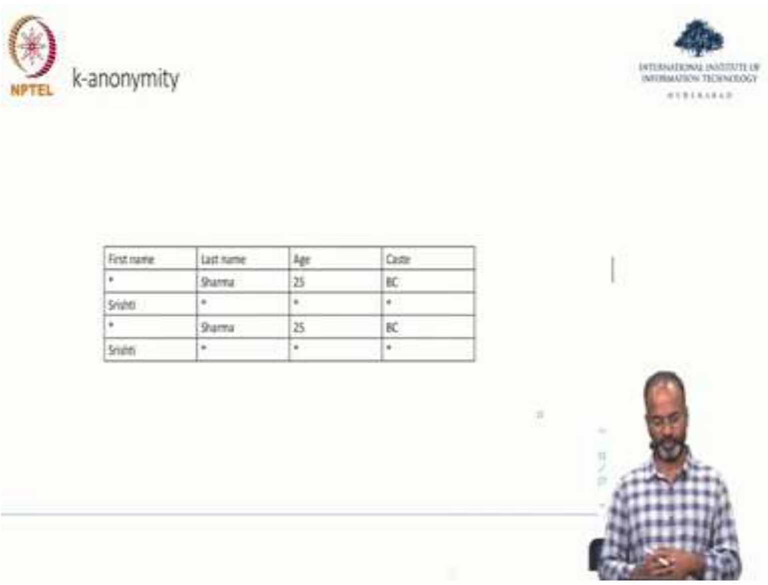
So, now whoever gets access to the data they cannot really end for that what exactly my weight is, they would be able to only infer saying it is between 40 and 50. Here is one table which gives you example of what suppression, what generalization you can actually do in this data.

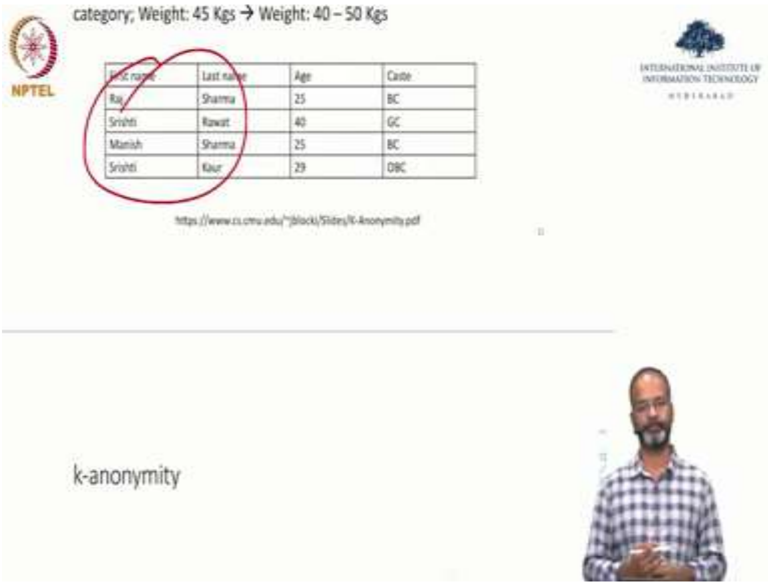
(Refer Slide Time: 28:14)



So, if you look at the data that you can that is a table which is first name, last name, age and caste is the four columns that are in the data, I just took the caste just to make it more sensitive.

(Refer Slide Time: 28:27)

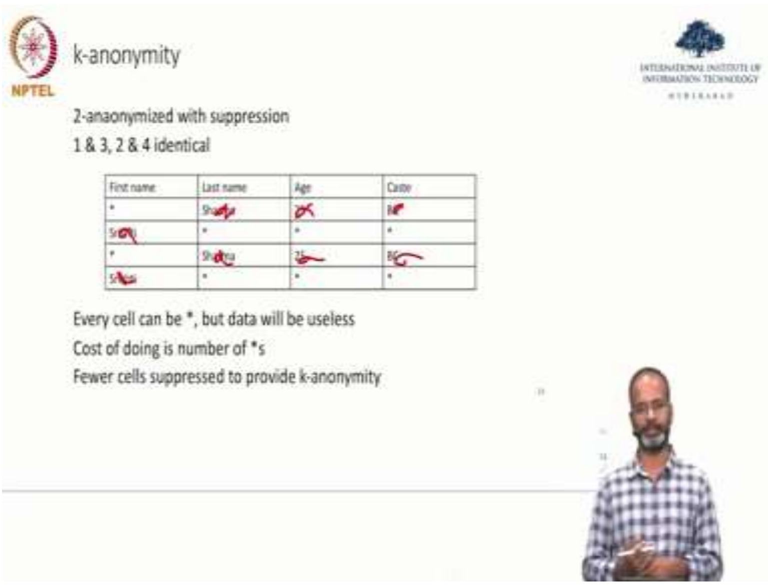




So, here if you see if you want to do suppression you can say that first name for the first row I will suppress the first name, for the second row I will suppress the last name and age and caste, third row I will suppress the first name, for the fourth row I will surprise last name, age and caste again.

What does this do? This just makes it the row 1 and row 3 to be exactly the same, and row 2 and row 4 to be exactly the same. So, now the point is that if you get access to this data, you just will not be able to identify which is actually Raj Sharma versus which is Manish Sharma. Similarly, which is Shristi Rawat versus which is Shristi Kaur. So, that is the idea here, if you are getting the idea suppression metric method.

(Refer Slide Time: 29:25)

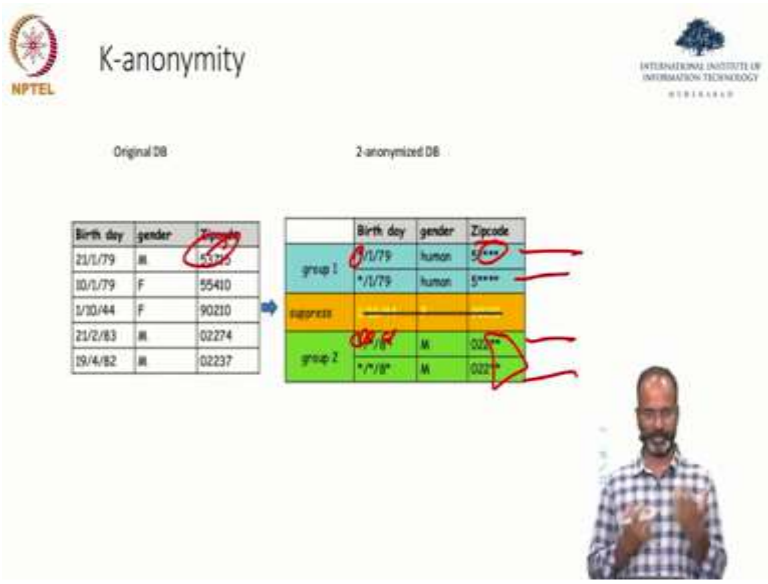


So, the idea here is that it is called two anonymized with suppression which is two rows are very identical, row 1 and row 3, row 2 and row 4 are identical, every cell can be suppressed with an asterisk but data will be useless, I mean one of the arguments that you can make is that look why do we even worry about finding these methods, just make asterisks for all of them. Making asterisks for all of them the problem is that okay good nobody will be identified from this but the problem is that you just cannot do anything with the data also.

Cost of doing is a number of stars that you have to put. The cost of actually doing this anonymization is about just getting the number of asterisks and the suppression, how many do you have to do, so that is the cost, let us take if you have a million rows and if you have to do this for like 10 columns how many places do you have to put asterisks and how do you decide which asterisks to put, where to put the asterisks.

Fewer cells suppressed to provide k-anonymity. The goal for k-anonymity was look I want to reduce, I want to use the data, I want to make the least number of stars in the cells but the data should be very useful for anybody whoever is accessing, using the data, that is the k-anonymity goal.

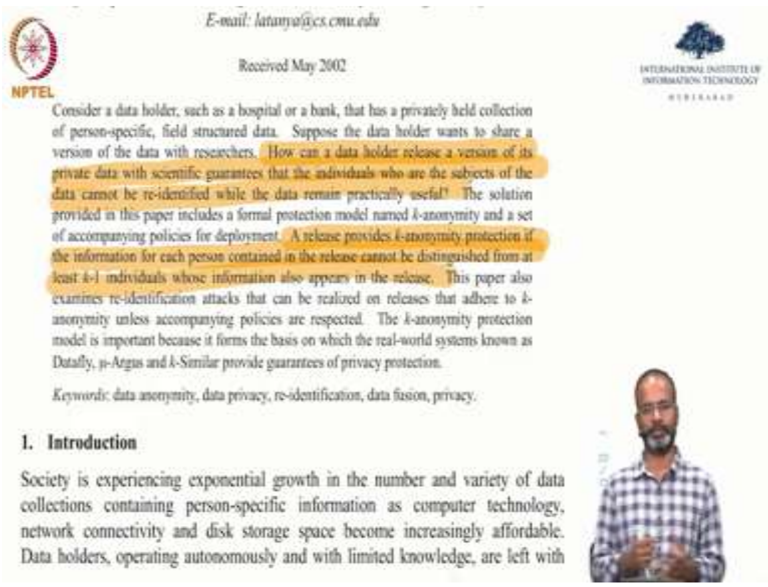
(Refer Slide Time: 31:06)



Here is another example of two anonymized data itself so again left in the row it says birth date, birthday, gender and zip code very similar to what Latanya had, now you can actually suppress one row in this which is get rid of this row, third row and then make the group one and group two which is by making only the date of birth date and the first two rows as star and the zip code last four digits in the zip code as stars again.

So, the ui zipcode is a five digit number, so it is suppressed here with four digits, this one, and then in group two it is suppressed as these three only the last two digits are suppressed here and now you will not be able to re-identify people, differentiate between these two users and differentiate between these two users, that is what k-anonymiti's goal was. Hope that is sinking in.

(Refer Slide Time: 32:30)

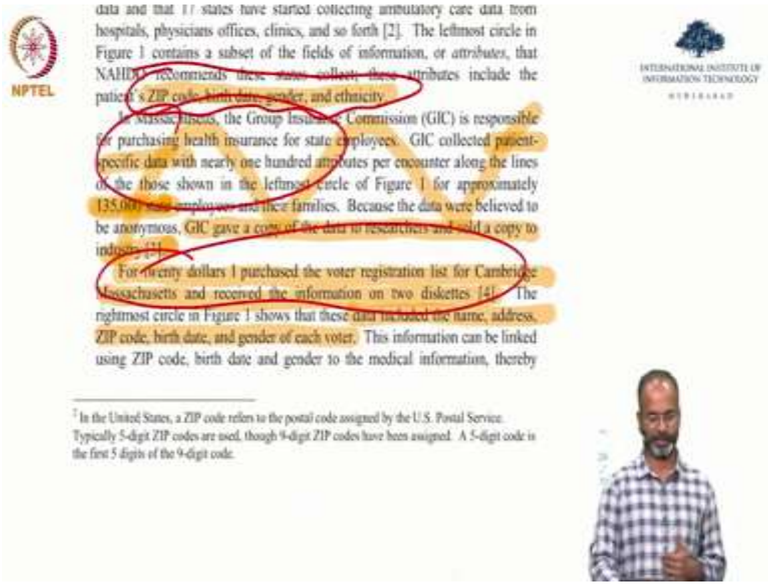


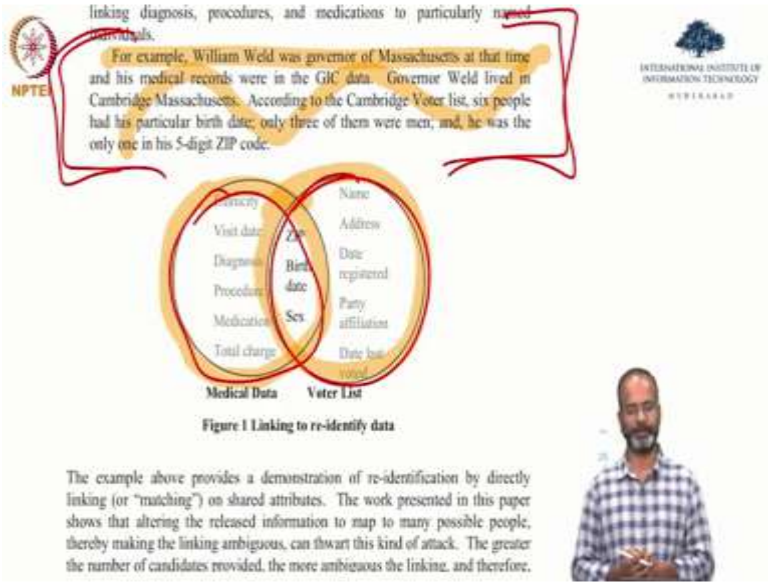
Let us look at what k anonymity paper was quickly. So, this is the k anonymity paper that Latanya had and some of it is we will see, we have already seen in the slides. So, this paper talks about k-anonymity a model for protecting privacy. How can a data holder release a version of its private data with a scientific guarantees that the individuals who are the subjects of this data cannot be re-identified, while the data remain practically useful, that is the goal.

A release provides k anonymity protection if the information of each person contained in this release cannot be distinguished from at least k minus one individuals whose information also appears in this release which is the two anonymized where at least two rows are very similar which is two anonymized meaning the second part of the sentence if you see, two anonymization protection if the information for each person contained.

And the release cannot be distinguished from k minus 1 which is 1, individuals whose information also appears in the release. In the Sharma example that I gave in the group one, group two example that I showed, you will not be able to identify, distinguish between these two users in the raw that is the idea for two anonymization. If that was three anonymization, similarly three rows would be there, you will not be able to identify, distinguish between three rows in the data.

(Refer Slide Time: 33:59)





So, some more details of what she actually ended up doing which is actually pretty exciting method that you followed, it will be nice for you also to think about if you can redo, try some things in Indian context. So, this one reads as re-identification by linking a national association of health data organizations reported that 37 states in the US has legislative mandates to collect hospital level data and 17 states have started collecting ambulatory care data from hospitals, physician offices, clinics and so forth.

So, essentially giving a background saying that look there are states which have to collect some of these health records. And the health record contains zip code, birthday, gender, ethnicity also. In Massachusetts the group insurance commission is responsible for purchasing health insurance for state employees, GIC collected patient specific data with nearly 100 attributes per encounter along the lines of those shown in the left most circle of the figure one which is what I showed you in the slide also for approximately 135,000 state employees and their families.

Because the data were believed to be anonymous GIC gave a copy of the data to researchers and sold a copy to industry also. So, this is basically the circle that is on the on the left hand side which is what GIC provided. For 20 dollars I which is Latanya purchased the water registration list from Cambridge and received the information in two disks, those days disks were the only ways data was shared, so she got two disks, actually you can also do this even these days.

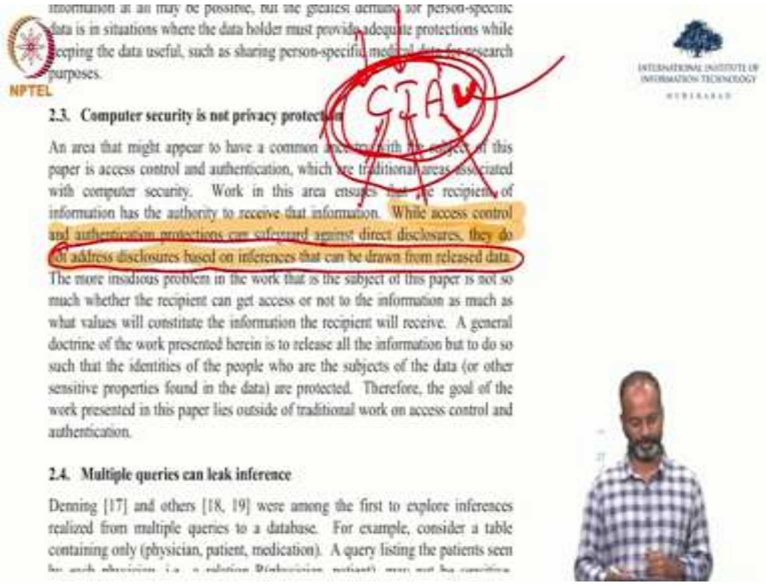
Now, I did a couple of data like this from few cities in the US by sending some request and actually collecting this data, looking at the data. The most circle which is how I have already shown in figure 1 shows that the data included the name, address, zip code, birth date and gender of each vote. That is this one. How did Latanya's work became again it is an academic work at MIT, it is a Ph.D. level work, she was just doing this research, how did this work became very attractive to others is because of this paragraph, this paragraph here.

For example, William Weld was governor of MA at that time and his medical records were in the GIC data because he was a governor, because he was so if you look at it this GIC it is saying that the government employees data has to be part of the data that is GIC has. Governor well lived in Cambridge, according to the Cambridge voter list, six people had his particular birth date and only three of them were men.

And he was the only one in the five, in his five digit zip code, zip code again like our pin code, people live in all these pin codes, the argument that is made here is that the governor who is male and whose Cambridge voter list, he was part of Cambridge voter list and people had his birthday, six people had his particular birth date and only three of them are men and he was the only one living in the five digit zip code, the other could be in Cambridge living but in other zip codes of Cambridge.

You can think about whatever city you are from, the pin codes are very different, meaning few kilometres, few areas are distributed into this pin codes, if the pin code is different you will not be able to identify that person again. If you are the only one living in that pin code for example if you are the only one, if you have to re-identify a person like this, if you are the only one student or if you are the only female student taking this class sitting in let us take Chennai, Kolathur with the zip code of Ananagar, some are Tinagars, some zip code it is very evident that it is just you, that is what was done here.

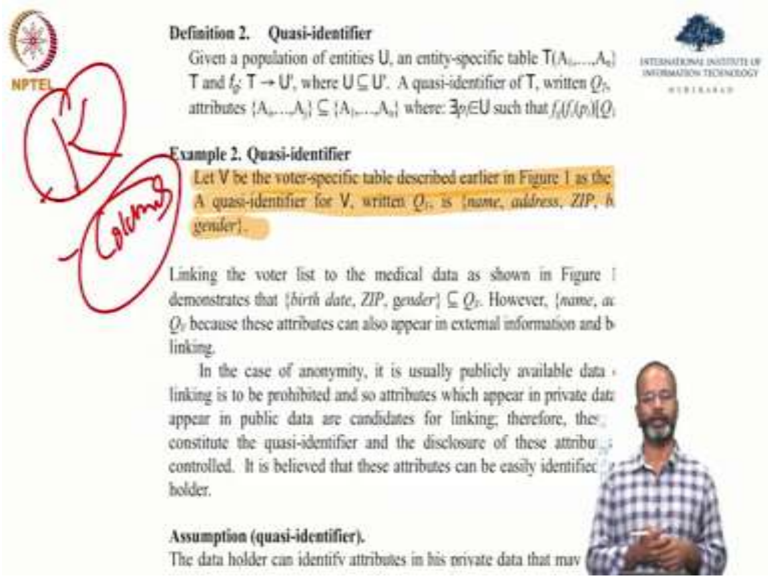
(Refer Slide Time: 38:54)



So, I think this paper is slightly philosophical also, this part is more like computer security is not a privacy protection. It argues that while access control and authentication protections can safeguard against direct disclosures, they do not address disclosures based on inferences that can be drawn from the release data, the argument here Latanya is making is that difference between what a security is and what a privacy is.

And I wrote the CIAU is how security is taught in fundamental security classes, confidentiality, integrity and availability and then use is usability which is what security is generally talked about, whereas privacy is look the data is shared, confidentiality is provided, integrity is there, data was given, data was received the same way it was actually sent. And availability is there, the data is provided if you need at any given point in time, but you know what data can be broken, data can be actually re-de-identified just because we can use outside information to de-identify users.

(Refer Slide Time: 40:06)

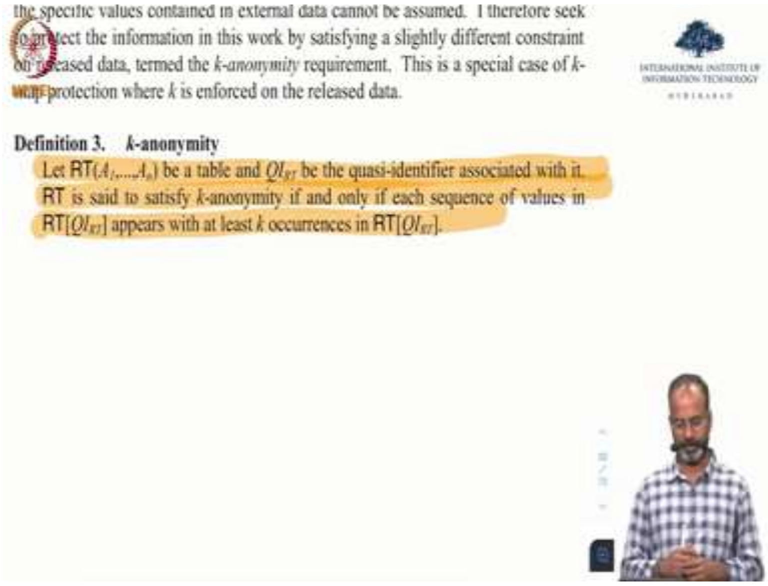


So, one critical thing, two critical things I think when you have to do, when you have to implement k-anonymity in a data just think about it, you are working on let us take a company, where the company wants to release data or you are from your college and your college wants to actually release the data of all the marks that the students got, for some research projects, for some companies to do some analysis and give it.

You have to identify two things which is what anonymity do you want to do which is this k value, which is this k value. Second you have to identify is what columns do you want to suppress, that is what this quasi identifier is which is to say that look let B be the voter specific table described in earlier in figure 1 as the voters list.

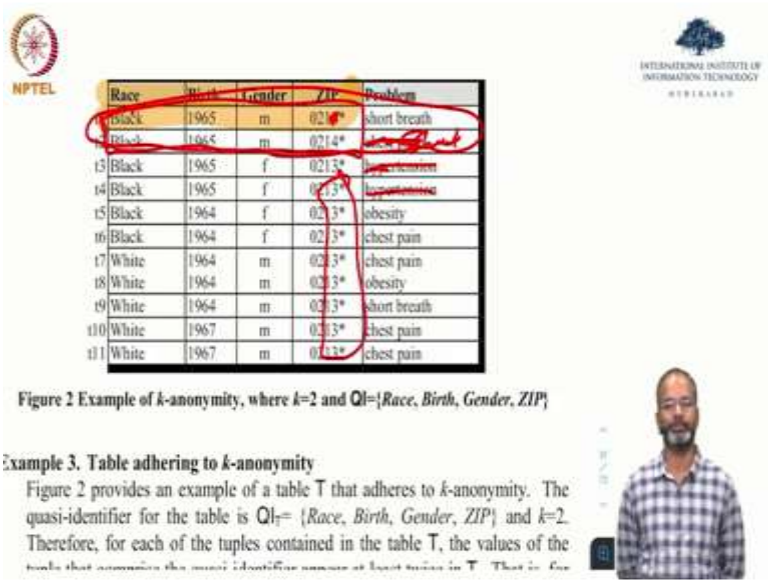
A quasi identifier for v, written q v is name address, zip code, birth date and gender which is the columns that you want to actually anonymize are the quasi identifiers that you want to actually look at. The more the quasi identifier is the more you want to keep the k, the more the cost of anonymization is. So, that is how you decide which column to anonymize and what anonymity do you want to keep, how many people are you with being re-identified in the data, what level of protection are you planning to give with the data.

(Refer Slide Time: 41:39)



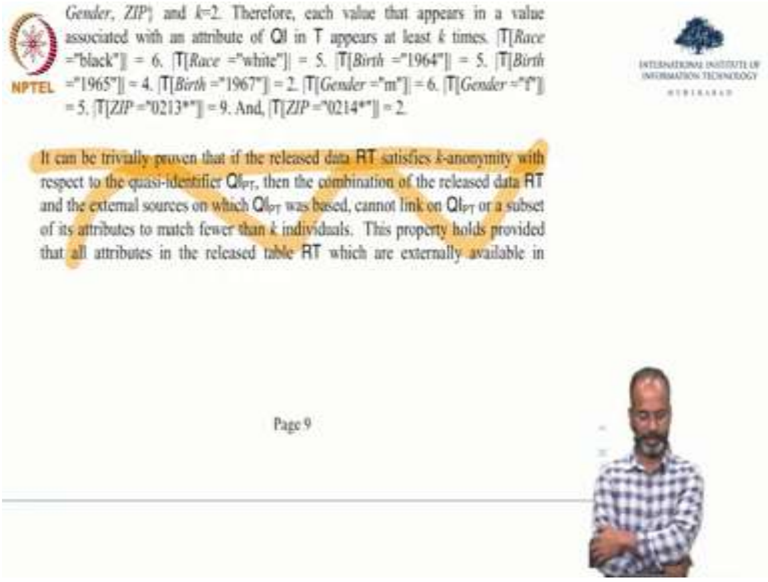
So, that is what is defined here. Let RT be a table and QI RT be the quasi identifier associated with it, RT is set to satisfy k anonymity if and only if each sequence of values in RT, QI RT appears with at least k occurrences which I have already shown you examples of two.

(Refer Slide Time: 42:02)



Just formalizing the whole thing, this is the example of k anonymity again, where k is equal to 2 and the columns are raised, birthday, gender, zip and the problem is on the last column, if you look at it you will not be able to identify people from this table.

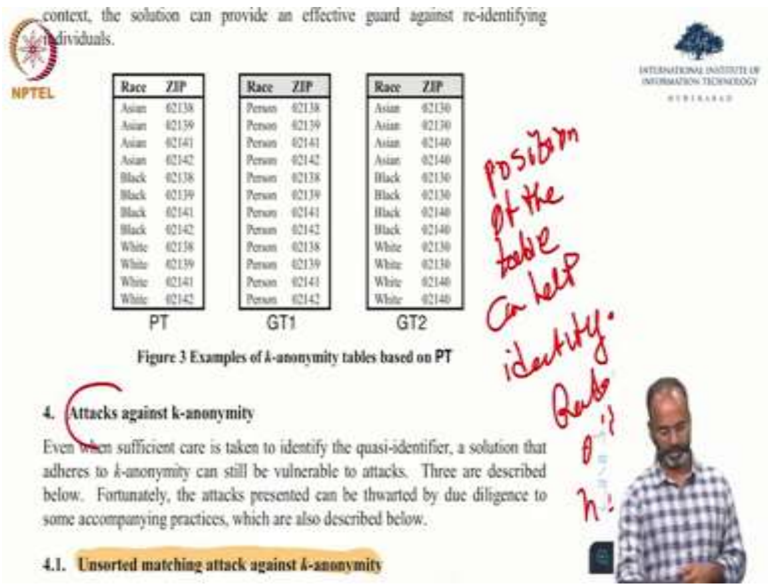
(Refer Slide Time: 42:23)



So, here is a argument that it can be trivially proven that if released data RT satisfies k anonymity with respect to quasi identifies. Again, keep in mind the key is this quasi identifier, because I think that is what controls, because if you pick the right quasi identifiers then the anonymity can be very very powerful.

QI PT then the combination of the release data RT and the external sources on which QI PT was based cannot link to QI, link on QI PT or a subset of its attributes to match fewer than k individuals. This property holds provider that all attributes in the released table RT which are externally available in combination blah blah blah. So, essentially it is arguing that you want to make sure that you are carefully identifying the quasi-identifier.

(Refer Slide Time: 43:17)



Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

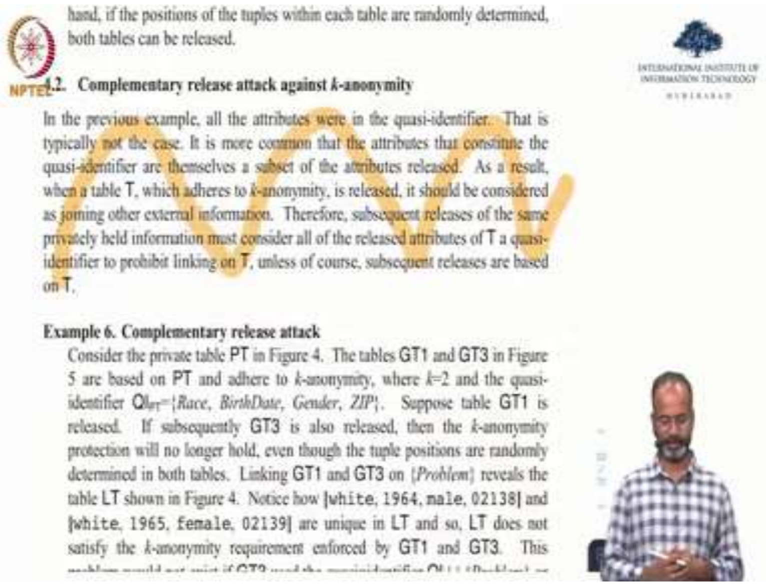
Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

So, you also want to think about how k anonymity can be at least the paper argues about how k anonymity can be, attacks can be against k anonymity. So, one of the attacks, let us go through the attack. So, one attack is unsorted matching attacks against k anonymity. So, one of this is position of the table can help identify also.

So, this is if the rows in the table are in particular order and the columns of the tables are provided in a particular order and there is also this temporal attack that that k anonymity also talks about which is you release the data now and you release the data sometime later, if the k anonymity is not kept in mind what was the data released before, it could have some concerns, so it could be re-identified, it could be used to re-identify data and the k anonymized dataset also.

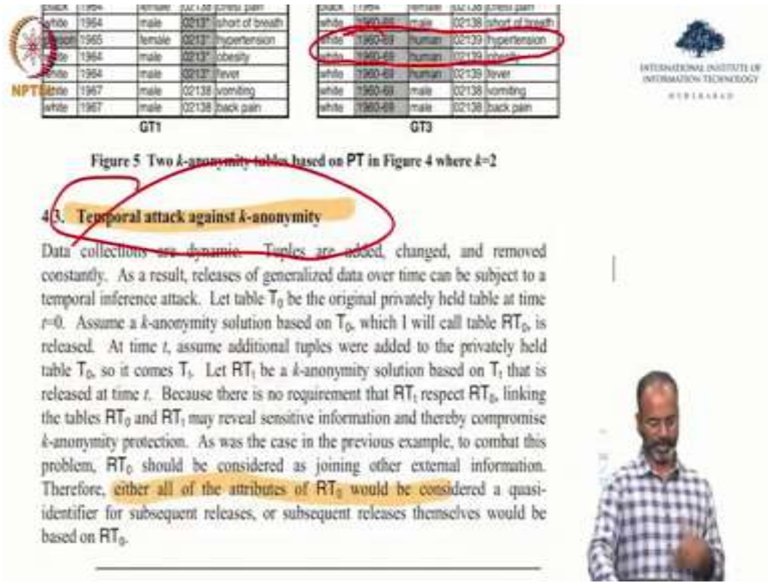
(Refer Slide Time: 44:20)



Yeah, this is complementary in the previous release attack against k anonymity which is in the previous example all the attributes were in the quasi identifier that is typically not the case, it is more common that the attributes that constitute the quasi-identifier are themselves a subset of attributes of released as a result when a table T which adheres to k anonymity is released it should be considered as joining other external information.

Therefore, subsequent releases of the same privately held information must consider all the released attributes of T quasi-identifier to prohibit linking of T unless of course subsequent releases are based on T. Essentially if the data set was released with the four column, columns is quasi identifier, you want to continue using those four columns and beyond as quasi identifies when you release data future, in future.

(Refer Slide Time: 45:17)



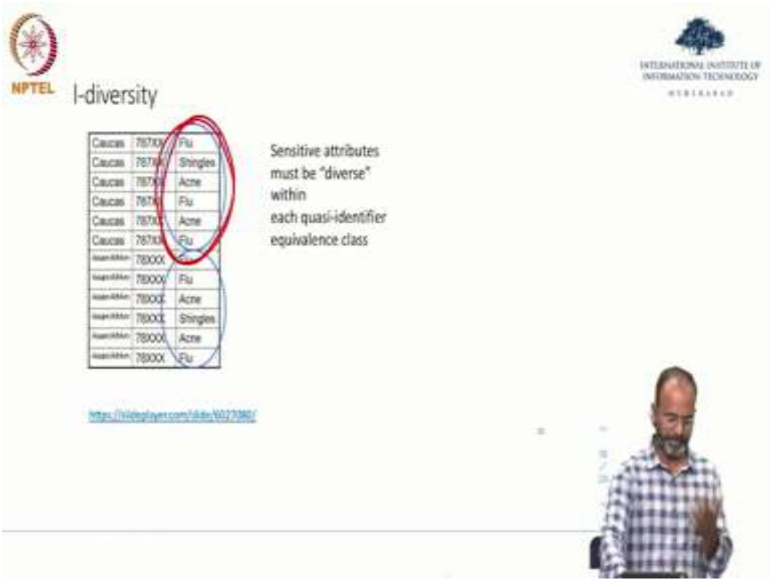
Temporal attack which builds on that also, so that is what k anonymity is. Let us go back to the slides.

(Refer Slide Time: 45:27)

Lack of diversity in sensitive attributes. So, here are the three limitations of the k anonymity methods itself. Lack of diversity in sensitive data, the columns if you see problem in the paper, if the columns are not very diverse, if the values in the cells are not diverse then there is a problem, users can be de-identified.

Background knowledge, supplement knowledge which is I know that you live in Annanagar makes me, I know that you live in Kachi Bowli in Hyderabad can be used to de-identify people in the rows also. Subsequent release of the same data set I just told you that a future release of the data set should be made sure that the earlier data quasi identifiers are kept in mind while the data is made public. So, those are the limitations of k anonymity. Let us continue on the other methods that I mentioned before, L diversity, T closeness and differential privacy.

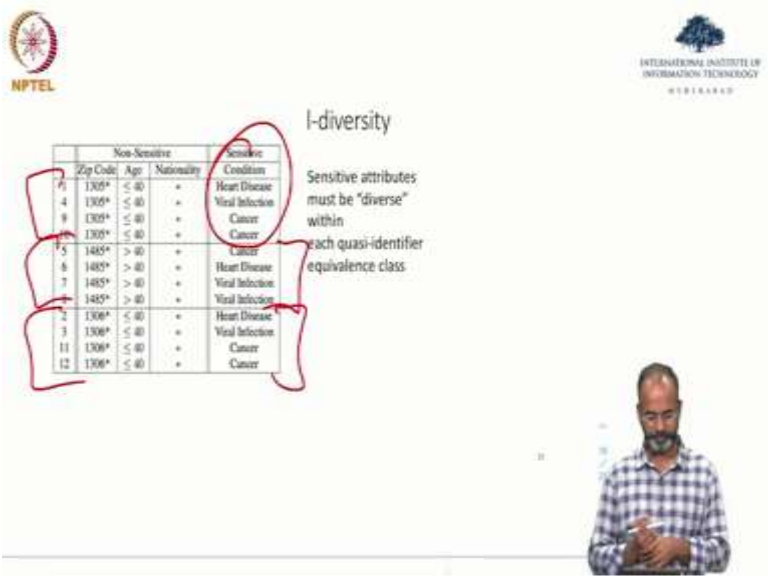
(Refer Slide Time: 46:42)



So, this is L diversity, the idea here is that if you remember the quasi identifier and k anonymity and the last column being problem we saw a column, where the diversity in that particular column was actually lesser. So, the idea that L diversity was arguing is that sensitive attributes must be diverse within the each quasi-identifier equivalent cell. What does that mean?

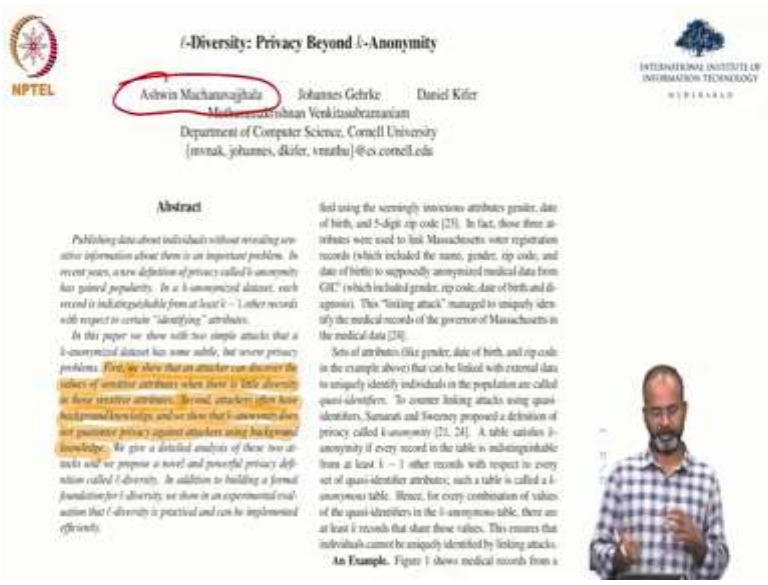
That means, in this quasi-identifier class which is the disease flu, shingles, acne, there is diversity there that is what makes the L diversity better, k anonymity this was not the case. We will see more details, I am going to give you more details, we will look at the paper also to get you more understanding of this.

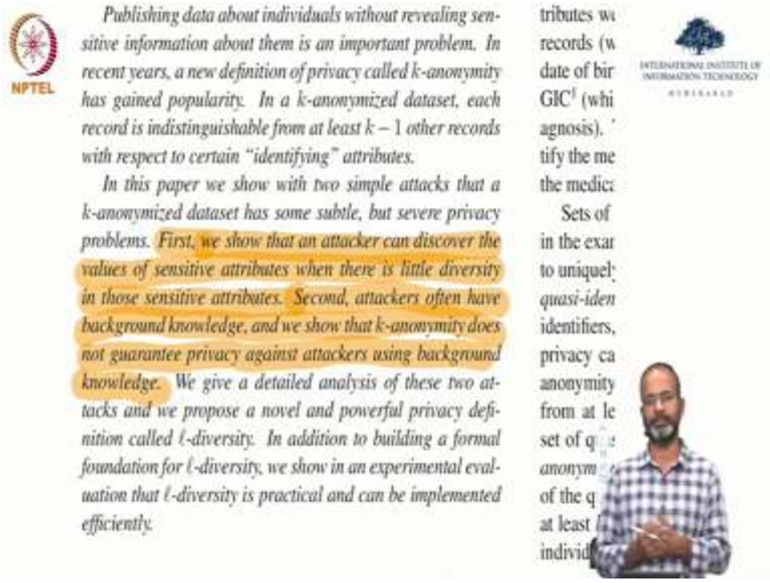
(Refer Slide Time: 47:38)



So, another example here which if you just look at it here, it is heart disease, viral infection, cancer. So, every class if you think, if you take this as one class, each class, every class is diverse enough that identifying the, identifying the rows in them would be actually much harder, that is L diversity.

(Refer Slide Time: 48:03)



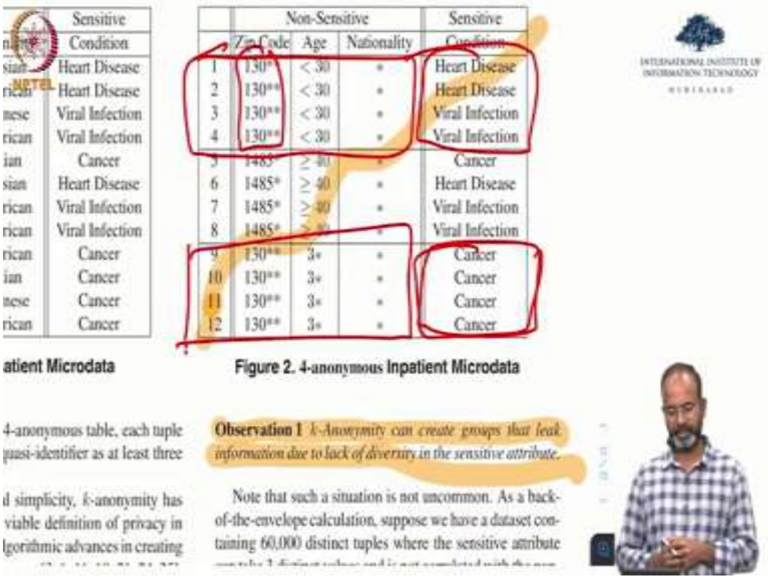


So, this is the paper for L diversity let us look at the paper, so that will give you more details of the algorithm, of the methods, everything. So, that is L diversity in the paper L diversity, privacy beyond k anonymity and then the algorithms that I am walking you through also is built temporally, first k anonymity came and then L diversity as therefore the papers are also kind of arguing about the prior methods.

What is, what was the goal and what did this paper show? First we showed that an attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. Examples of heart disease, cancer, flu, all that. Second, attackers often have background knowledge and we show that k anonymity does not guarantee privacy against attackers using background knowledge.

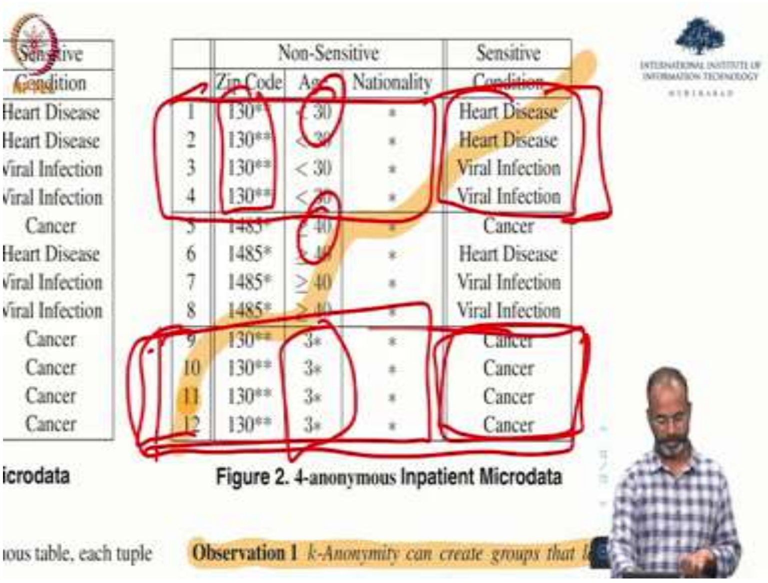
This background knowledge is what I said earlier also that look I know that you live in this zip code, I know that you are a male, I know that you must be aged between 40 and 45. Let us say if my home was next to you or if you are in my class all that information is background knowledge additional information which can be used to actually do the attacks.

(Refer Slide Time: 49:35)



So, here is the example again the same thing that I used in the slide. So, this is 4k anonymous in inpatient micro data which is in k anonymity if you see the diversity here is pretty low that is the argument that this paper is making, look I think we need to have more diversity in the sensitive column which will make it more harder to de-anonymize the data.

(Refer Slide Time: 50:07)



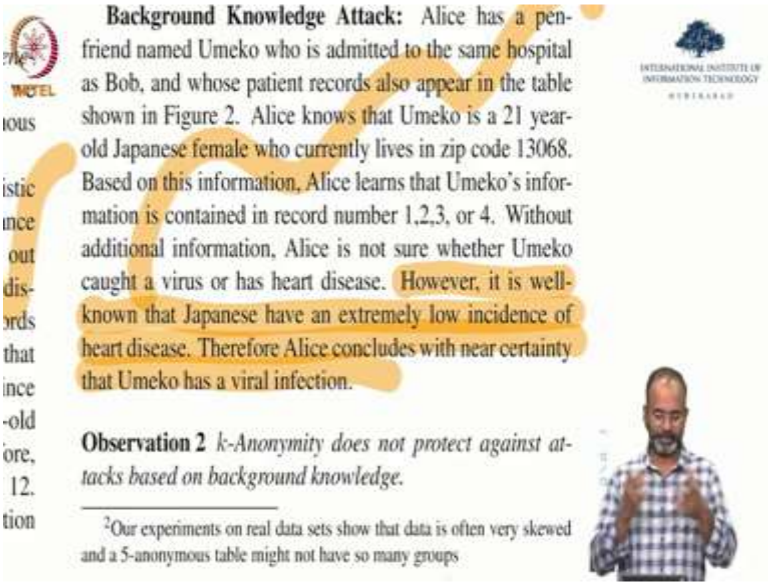
Some interesting attacks that if let us go through it in slightly detail. So, Alice and Bob are antagonistic neighbours. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice, so please keep a watch the background knowledge all this will come. Alice sets out to discover what disease Bob is suffering from.

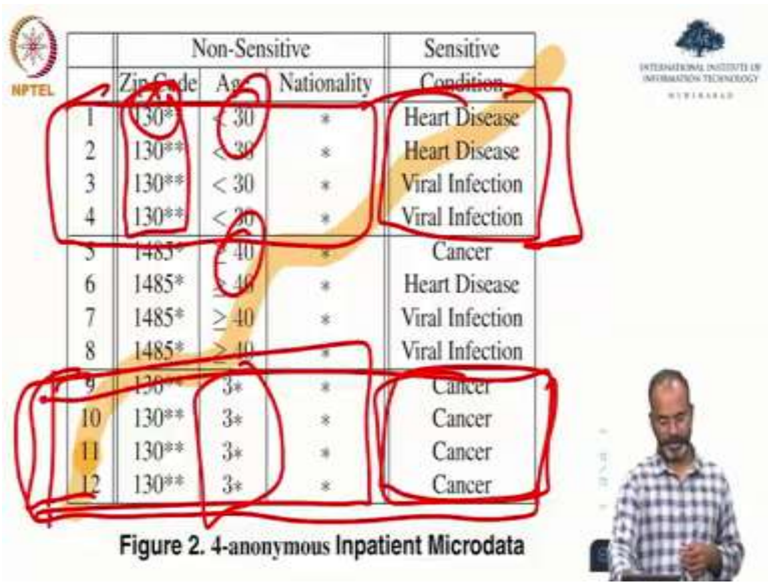
Alice discovers the four anonymous table of current inpatient records which is what I showed you right now, which is this that is the table. Again, you will have access to the papers so you are welcome you can actually take it as leisurely as you want to look at the details of the paper. And so she knows that one of the records in this table contains Bob's data.

Since Alice is Bob's neighbor, she knows that Bob is a 31 year old American male who lives in the zip code. Therefore, Alice know that Bob's record, his record number is 9, 10, 11 or 12. So, that is the, that is the idea that she can figure out because she lives in the zip code x and he is actually 31, so this one is less than 30, this one is greater than 40, this is within the age group that she knows that he is 31, so he has to be in this class of patients that is the inference that she is making.

Now, all those patients have the same medical conditions cancer. So, Alice concludes that the Bob has cancers.

(Refer Slide Time: 51:51)





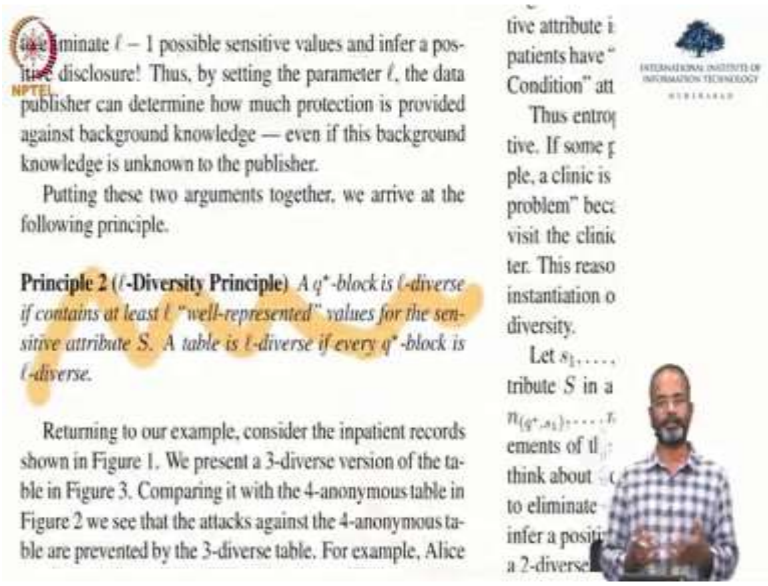
	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Here is the next attack which is background knowledge attack. Alice has a pen friend named Umeko who is admitted to the hospital as Bob and whose patient records also appear on the table shown in figure 2. Alice knows that Umeko is a 21 year old Japanese female who currently lives in the zip code 13068, based on this information Alice infers that Umeko's information is contained in the record 1, 2, 3 and 4, without additional information Alice is not sure whether Umeko caught a virus or a heart disease.

So, that is the same table, zip code is here and heart disease and viral infection. However, it is well known that Japanese have an extremely low incidence of heart disease. Therefore, Alice concludes with mere certainty that Umeko has a viral infection. That is the kind of attacks that you can do in terms of k anonymity which is what the paper is arguing about. We also saw in the k anonymity paper itself different kinds of attacks that Latanya mentioned what are possible.

(Refer Slide Time: 53:09)

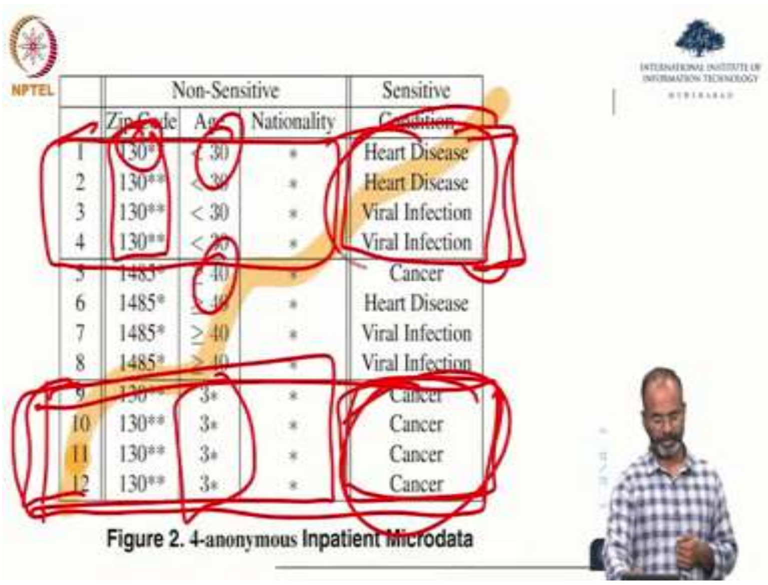


...minate $\ell - 1$ possible sensitive values and infer a positive disclosure! Thus, by setting the parameter ℓ, the data publisher can determine how much protection is provided against background knowledge — even if this background knowledge is unknown to the publisher.

Putting these two arguments together, we arrive at the following principle.

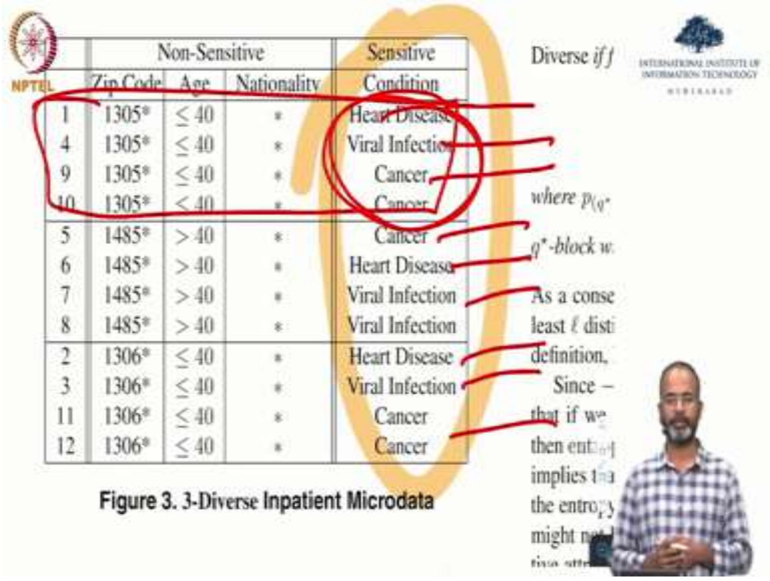
Principle 2 (ℓ-Diversity Principle) *A q^*-block is ℓ-diverse if contains at least ℓ "well-represented" values for the sensitive attribute S. A table is ℓ-diverse if every q^*-block is ℓ-diverse.*

Returning to our example, consider the inpatient records shown in Figure 1. We present a 3-diverse version of the table in Figure 3. Comparing it with the 4-anonymous table in Figure 2 we see that the attacks against the 4-anonymous table are prevented by the 3-diverse table. For example, Alice



	Non-Sensitive			Sensitive
	Zip Code	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

Figure 2. 4-anonymous Inpatient Microdata



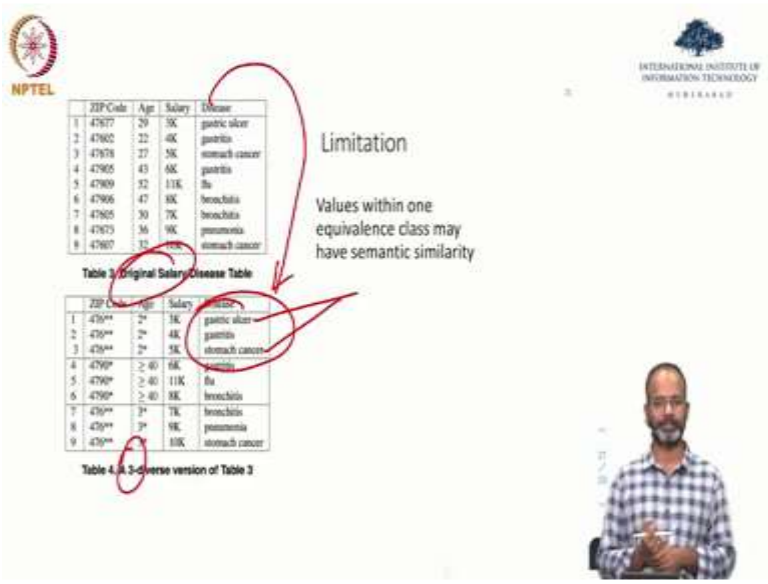
So, what is L diversity arguing that it is better than k anonymity. L diversity principle is a q block is l diverse which is block as in the class is l diverse, if contains at least l well represented values for the sensitive attribute, yes, a table l diverse if every q block is l diverse. So, essentially they would show that this example which is every row, every column in the sensitive would be different, would have three categories here, viral, heart disease and cancer, cancer, heart disease, viral, heart, viral, cancer.

So, if you go back to the earlier table that kind of an attack Alice knowing that Bob is in this table and she could infer that he has cancer is just not possible and the same example with the Umeko also that heart disease and viral infection is not possible. Because there is another sensitive information, the probability is slightly lower that is all. I hope that is making sense in terms of what is the expectation of l diversity is, how l diversity works.

(Refer Slide Time: 54:38)

So, the paper meaning you are welcome to take a look at the paper, go in details of evaluation, how they can actually empirically they show that l diversity is more stronger than k anonymity.

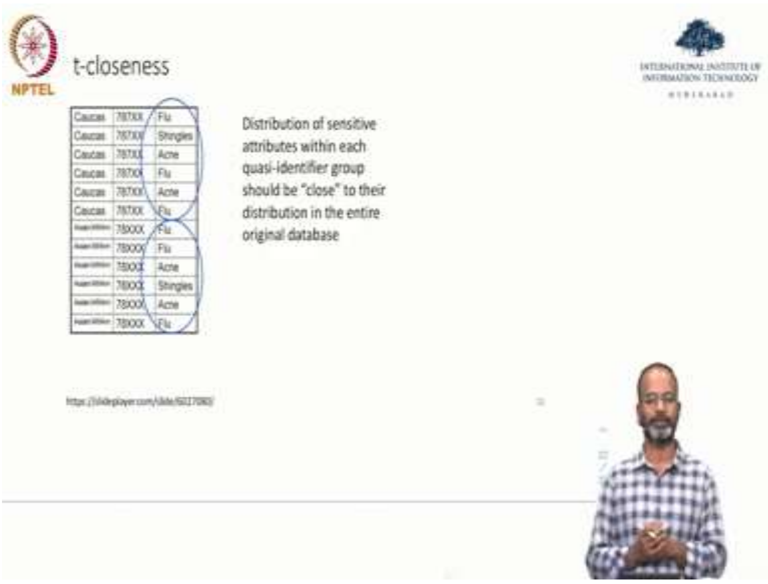
(Refer Slide Time: 54:49)



So, now if you look at l diversity itself you can pause the video for a second and think about what are the limitations of l diversity itself. So, if you look at one of the limitations values within one equivalence class may have semantic similarity, even though the diversity may be there but there is semantic similarity between the values is the concern that T closeness researchers argued. What is that?

So, if you look at let us take this one, so this is the original table again this will show up in the paper but this is the original table and anonymous table is this with respect to l diversity. So, this is three diverse that is the original salary disease table. If you look at this, gastric ulcer and stomach cancer are both semantically similar to something relevant to gastric problems. And therefore, an attacker can identify that let us take if a patient, if a friend is in this data he or she could be re-identified that is the problem that T closeness was arguing for.

(Refer Slide Time: 56:08)

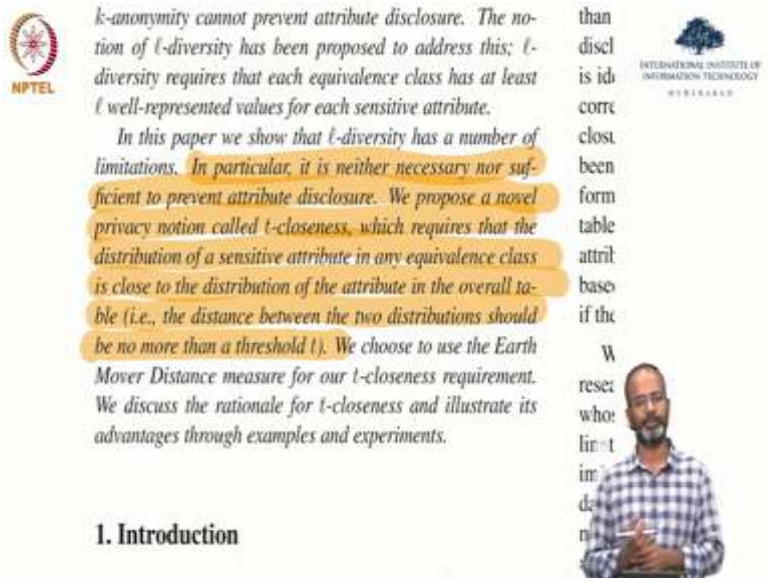


Distribution of sensitive attributes within each quasi-identified group should be close to their distribution in the entire original database. So, the arguing, the argument that the T closeness was making said look the diversity that we see in the table, the quasi-identified group should be close to the distribution the entire table itself, not just only the class of was that we are looking at. We will see again in the paper.

(Refer Slide Time: 56:36)

So, that is the paper which is T closeness privacy again because of temporal they are they titled as T closeness privacy beyond k anonymity and l diversity. These are all phenomenally influential papers, meaning if any of you are interested in this topic which is anonymization, feel free to read the paper and come back and we can discuss the paper in detail too.

(Refer Slide Time: 57:04)



Here is what the paper is so we can actually look at the paper quickly, some parts of the paper to generate some interest in you. So, this is the T closeness paper. In particular it is neither necessary nor sufficient to prevent attribute disclosure; we propose a novel privacy notion called t closeness which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in overall table.

So, any class should be similar close to entire table that is the distance between the two distributions should be no more than a threshold t, they would use the earth movers distance but at some distance metric you can actually keep and you can say that the distance between these two which is only the class and the table itself should be a small value threshold t.

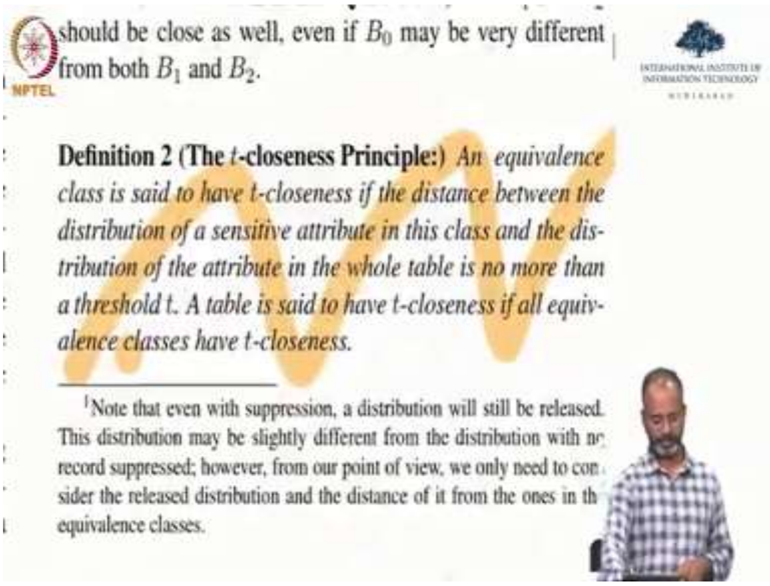
(Refer Slide Time: 58:08)

So, these are examples again to argue that the k anonymity, l diversity does not work.

(Refer Slide Time: 58:19)

So, this is again an example where they showed that semantically similar and therefore does it fails.

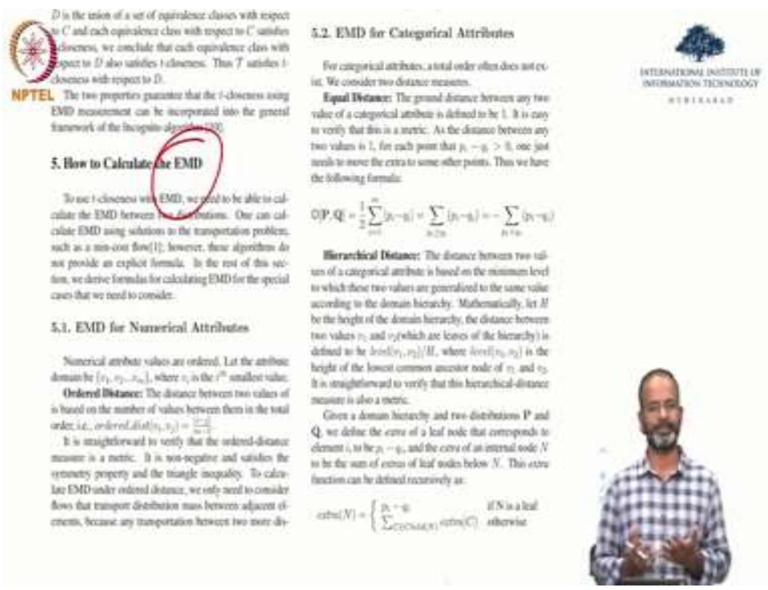
(Refer Slide Time: 58:29)



So, here is the t closeness principle. An equivalence class is said to have t closeness if the distance between the distribution of the sensitive attribute in this class and the distance of the attribute in the whole table is no more than a threshold t, a table is said to have t closeness if all equivalence class of t closeness.

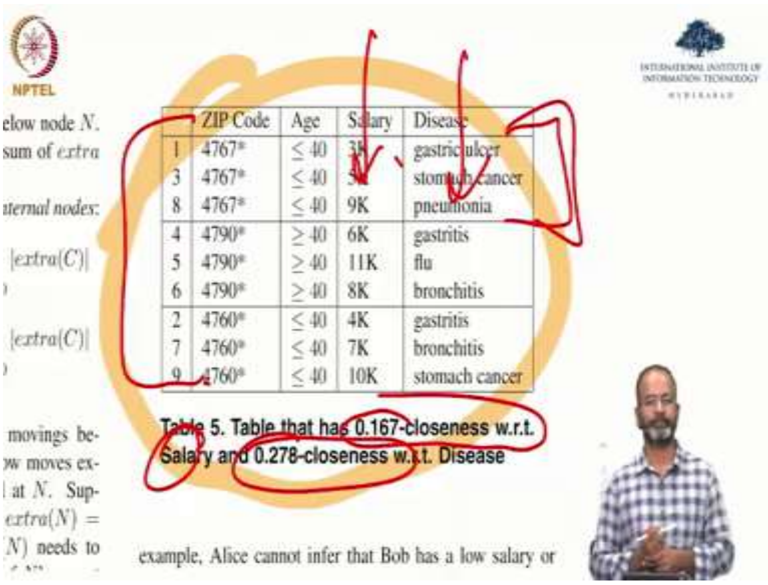
The goal here is to measure the t closeness that is what they would show how t closeness is measured but it is basically a distance metric that you can keep, you can have any distance metric that you are aware of and use that saying that the distance between the class and the table is small.

(Refer Slide Time: 59:14)



So, they would do earth movers distance, earth movers distance is a distance where the initial idea was if you had to put the thrash, gets accumulated if you were to move one thrash so to say hill which is of some height, some shape to a different shape and a height, what is the distance it takes? So, there are also Levenstein distance a Gyro Winkler distance, there are many distance measures which you can actually use for the data that you have.

(Refer Slide Time: 59:56)



	ZIP Code	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

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

example, Alice cannot infer that Bob has a low salary or

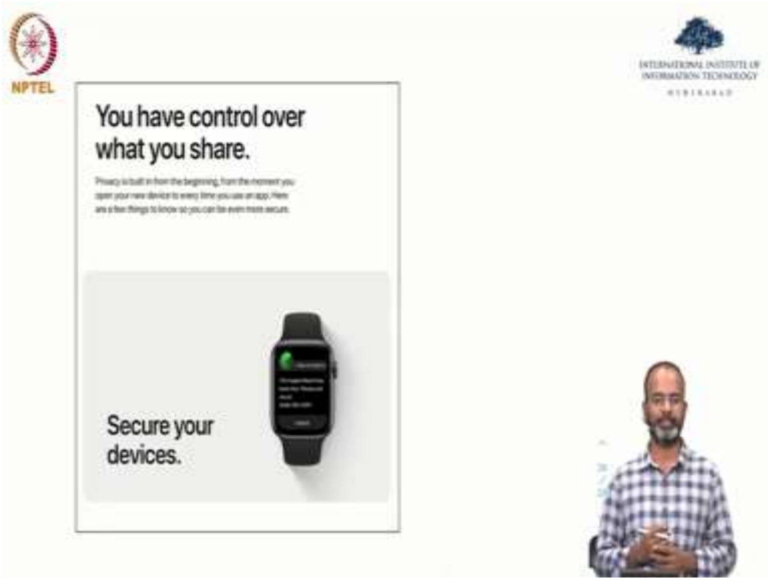
So, here is what the, so the paper deals with the how the distance metrics you can measure but here is one example that they argue that look it is the same table, there they are saying the table that has t closeness of this with respect to salary and ex-closeness of with the disease for the for this table and therefore it is more anonymous compared to the table that was given for k anonymity and l diversity.

And again the distance metric should be, could be measured as look what is the come up with a metric for this and how different is this from the whole table that is the metric you want to measure.

(Refer Slide Time: 60:57)

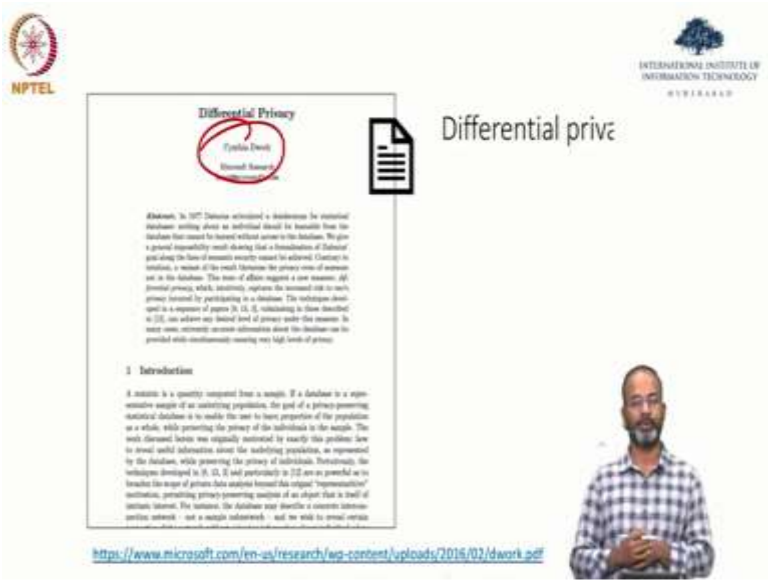
So again this paper goes on to detail of how this metric is measured, they do some experiments to argue that this t closeness method is more powerful than k anonymity and l diversity. So, hope you understood that. So, from the let us go back to the slides.

(Refer Slide Time: 61:30)



So, we saw k anonymity, l diversity and t closeness. Now, we will see something more, something that probably is relatable also. So, if you have heard about apple using differential privacy, differential privacy is the next concept. The apple products have now implemented differential privacy and many other products, many other platforms are also trying to find ways to implement differential privacy in it. So that is the idea differential privacy.

(Refer Slide Time: 61:56)



So, this method again, so this this method was developed by Cynthia Dwark, Cynthia Dwark has a very short I think this video is about 16, 17 minutes. What I recommend doing is take a look at this video, like you have seen the videos of social dilemma all of that, watch this video and come back and let us discuss if you have any questions or something and I have also pointed the paper.

So, the idea again is that there is a formal mechanism by which they have said that the anonymization is stronger than the earlier methods. And this has become differential privacy has become extremely popular because the implementation has been going on in some of the products and popular services have started using it.

(Refer Slide Time: 62:52)

So what we covered for week 6 that is the content for week 6, week 6 is slightly dense, I am going to keep it a little shorter for this week because we can also do some discussion around papers that you read, the videos that you watch. So, what we covered, why anonymize, AOL, data leak, Netflix and methods for anonymization, we saw all these four methods.

(Refer Slide Time: 63:22)

So, again thanks for attending this, listening to this lecture and if you have any questions feel free to drop it on the mailing list, I hope that the mailing list will help in answering any of the questions that you may have.