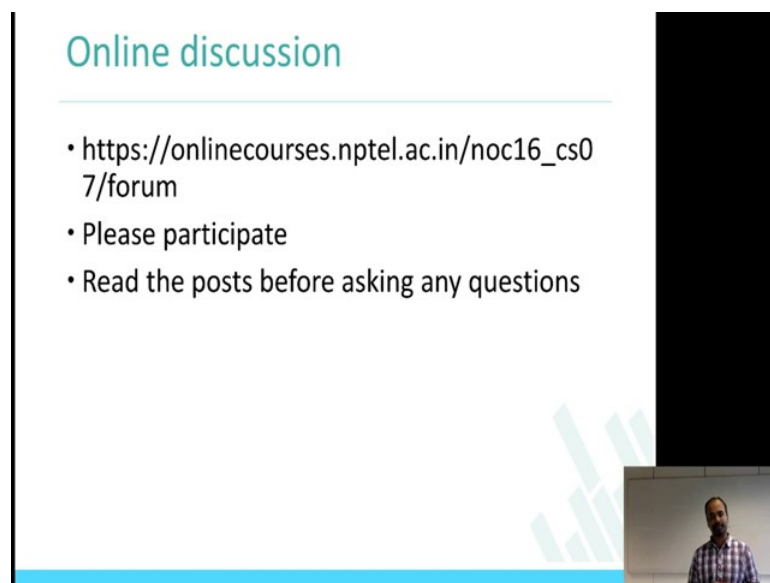**Privacy and Security in Online Social Networks**
Prof. Ponnurangam Kumaraguru ("PK")
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Week - 2.1
Lecture – 06
OSM APIs and tools for data collection

Welcome back to the course on Privacy and Security in Online Social Media, week 2.
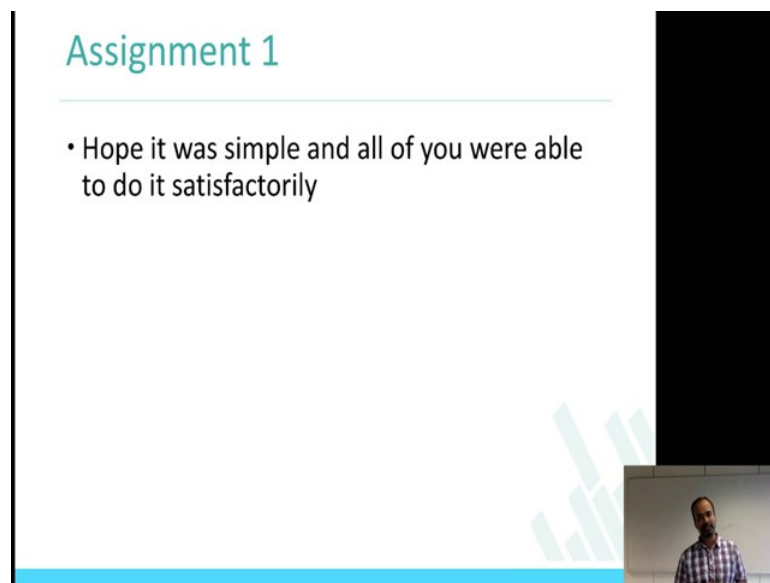
(Refer Slide Time: 00:16)

I hope you are participating in the online forum that we have in the course. I already see a lot of people asking questions, and trying to answer. My sincere request will be, please, please read the posts before you actually ask the question; that is, read the posts that have been already asked, the questions that have already been asked and the answers that has been already given, before asking the question. And, please participate also in the online forum, not just only asking questions; if you know the answers for the questions that others are asking, please try and answer them also.

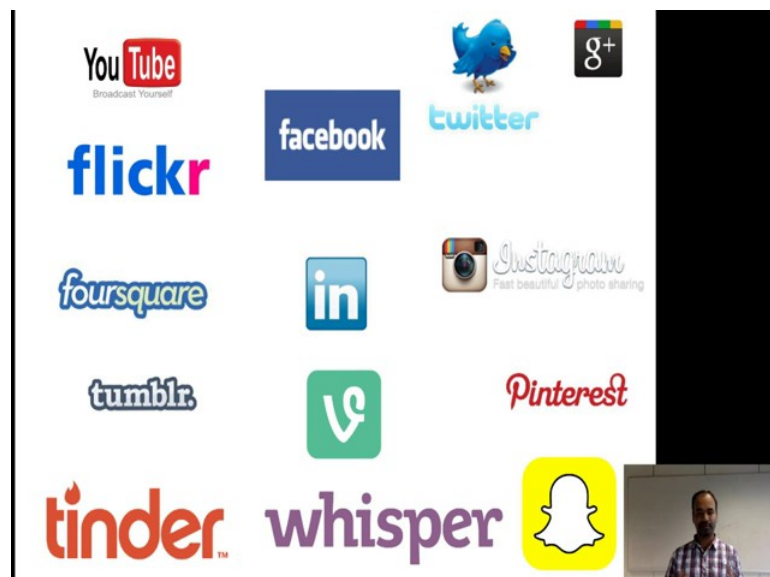I hope most of you got to see the assignment 1 that we had posted. So, I think, the weekend of the week 2 is the deadline for the assignment 1. Please try to work it out. The assignment 1 is actually pretty simple. We have just captured some questions from the slides that we did, and some from the tutorials.

So, let me just give you a quick summary of what we have seen until now and then, I will go ahead with the topics that I wanted to cover today. So, first, we saw what social media is; different types of social networks, different types of content that gets generated on our

social network; classical online social media services, and then some, which are more like ephemeral social networks and anonymous social networks.
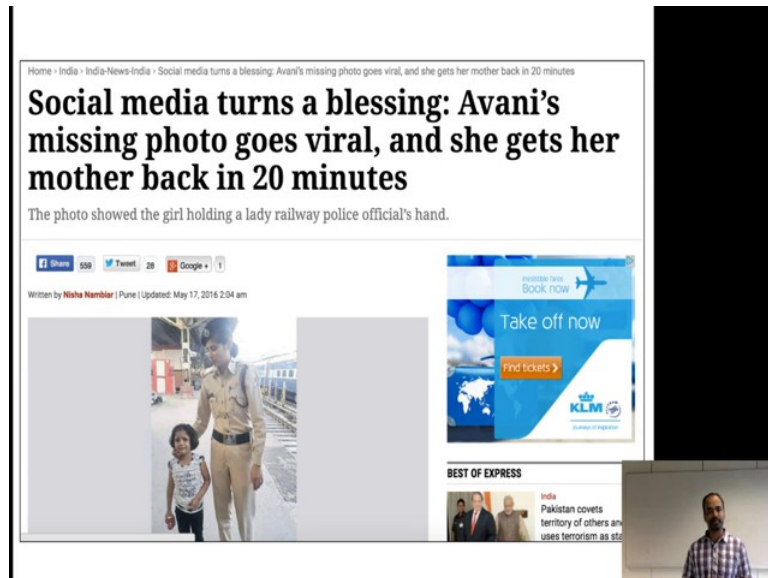
(Refer Slide Time: 01:41)

We also saw what online social media means in 60 seconds; so, how much of data is getting generated on social media in 60 seconds. We saw 400 hours of videos uploaded on YouTube, and 3.3 million posts are done on Facebook and things like that. This basically shows us that, large amount of content that are getting generated on online social media services.

(Refer Slide Time: 02:09)

We also saw what 4 or 5 V's of online social media are, they are volume, velocity, veracity, variety and value - those are the 5 V's of online social media.

(Refer Slide Time: 02:25)

And then, I looked at, I showed you some events, where online social media has played an important role in the real world and in the society also.
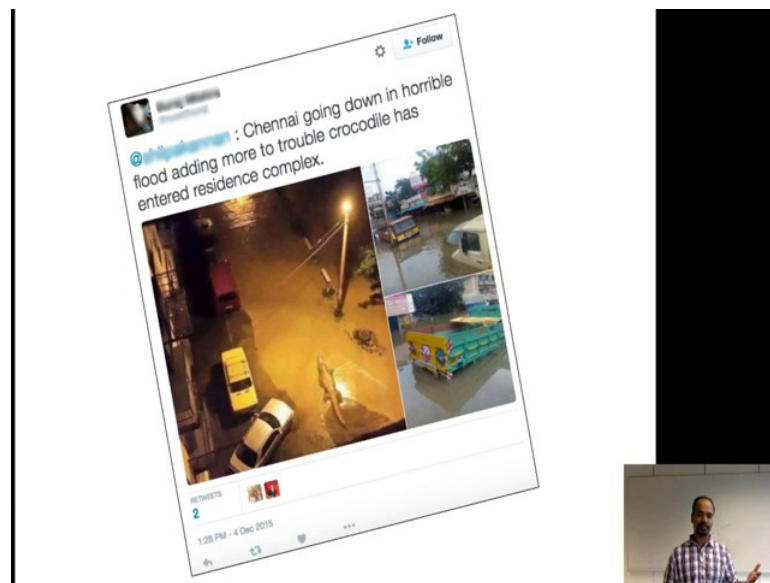
(Refer Slide Time: 02:37)

Telling you about different issues on online social media, for example, in this case, it is compromised account; an account was compromised, where the post said 'Two explosions in White House and Barrack Obama injured', and, there was, there was after

==effects== of this tweet. So, we looked at different issues that are happening on online social media; compromised account, fake content in this case; and image of a crocodile on the streets of Chennai, while Chennai floods was going on in December 2015, caused panic among citizens.
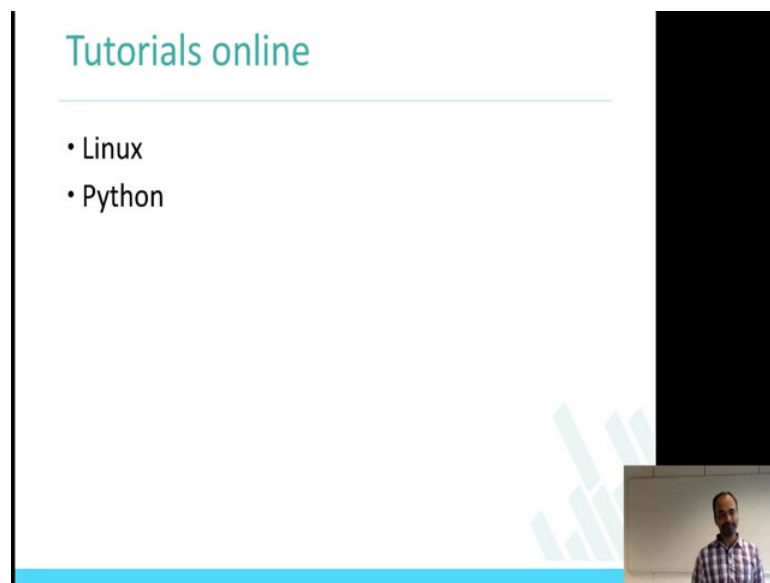
(Refer Slide Time: 02:59)

(Refer Slide Time: 03:11)

And, there are also people who lose jobs and others issues, because of the usage of online social media.
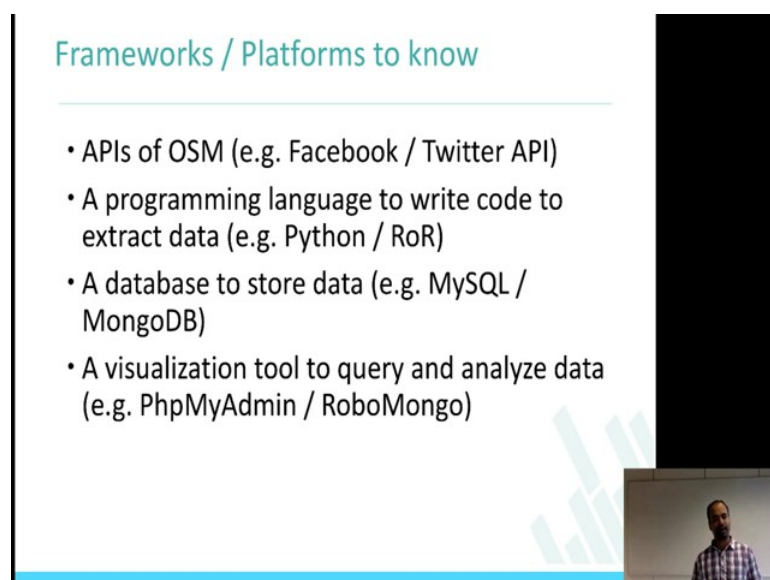
(Refer Slide Time: 03:19)

And, in the week 1, we also covered a little bit about Linux and python; hopefully, you are all set, in terms of using the platforms, because, I think, there were some questions about, 'can we use windows?' You should be able to use windows, and do programs on python, but it just said our support will be mostly on Linux. And, of course learning Linux will also be good for you.
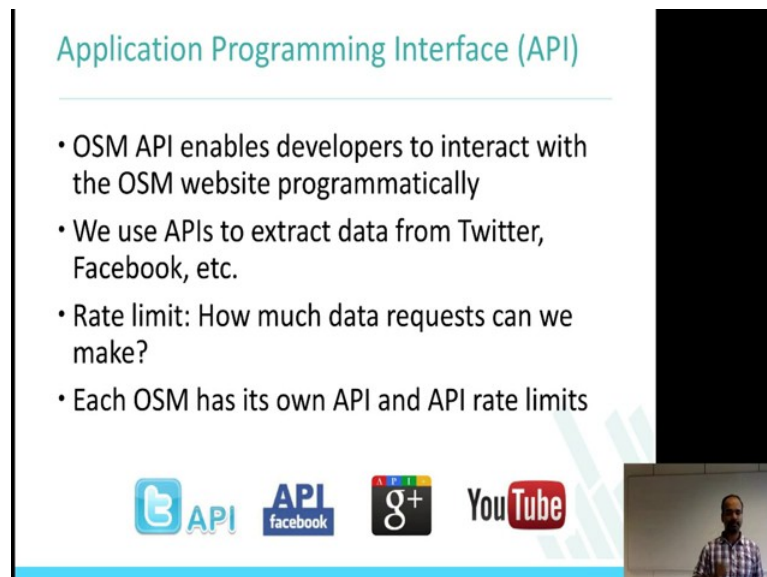
(Refer Slide Time: 03:44)

So, what I want to cover today is a couple of things; one, I wanted to actually look at different frameworks or platforms, that you would get to know while doing this course,

or in another terms, you should know while doing this course, and collecting data from online social media, analyzing and making inferences.

We will look at what an API is; different kinds of APIs that are available for Facebook and Twitter. Then, we will also look at programming language. There has been a tutorial on python. So, I will just quickly go over. In any case, my work for this week 2.1, about these topics, are only generally, to introduce and then, we will have a tutorials, which are specifically focused on some of them.

Then, we look at programming languages; and then, we will also look at a little bit of database, how this data is stored, what kind of format that the data is coming out; and a little bit about visualization tool.

(Refer Slide Time: 04:45)

First, API, which is Application Programming Interface; this basically enables you to interact with the online social media, programmatically, and collect data from there. What does this mean? This basically means that, you can actually have a tunnel that is from your program to the social media services, to collect data. It just creates a tunnel between your program and the online social media services, where you are going to ask some data and then, the social media service is going to respond with saying, here is the data that you asked for, right.

Particularly, in our case, we will actually look at APIs for Facebook and Twitter, which will help you to collect data from Facebook and Twitter. There is other APIs also; all other social media services or majority of the social media services provide you with an API. We can't cover everything in this course. So, we are going to start looking at only the most popular ones, or the ones that we can actually use for this course, which will help you to understand how APIs work, what data can be collected. So, you can actually do it for other social media services, yourself.

So, one of the important thing that you want to also keep in mind is that, about the rate limit, which is that in social media services when we want to collect data you cannot collect the data everything that is available on social, on these services. Because, I am sure, the companies do not want to give you all the data also. They have set it up, you know, by saying that, they have a rate limits for every social media service, and every piece of data that we want to collect from them. So, we will look at something in the tutorials about rate limits, particularly about each of the social media API , but I wanted to just give an idea about, there is going to be a rate limit, in terms of the data you can collect from these services.

(Refer Slide Time: 06:33)

Also then in python, since you have already done a tutorial on python, I will keep it really short. It is basically a programming language, that is used to collect data and is one

of the popular languages currently used in terms of writing API requests to the social media services.

And, it also has a lot of libraries for reading URLs, parsing data, interact with API, and understanding the JSON objects, and things like that.
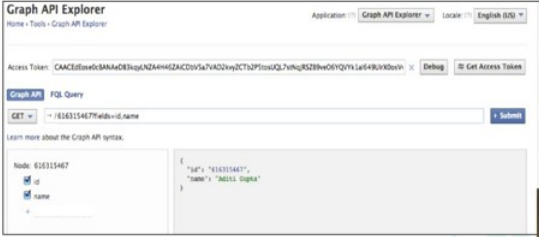
(Refer Slide Time: 07:00)

Data format, so particularly the API, when you send the request to Facebook saying that, 'please give me all the data about friends that PK has', or, about the date of birth of PK, or about my friends' network. So, what it is going to give you back is actually, it is going to give you in some format. One of the formats that it gives you is actually a JSON format, which we will see in brief what this format means and how we actually interpret the data that is coming back from Facebook, or Twitter. XML, which is also a format with some social media services give, or the JSON, is also a little bit like an XML, which is Extended Mark-up Language.
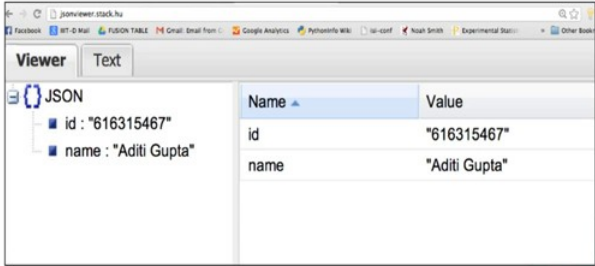
(Refer Slide Time: 07:49)

So, here is what a JSON means. JSON means, JavaScript Object Notation, which is a data that you get back from the social media services. So, here is an example that I have in this slide, which just shows you about the JSON object that is returned, when you are asking for id and name of a particular user. So, this is the Graph API Explorer, which you will see in the tutorial in more detail but, it is essentially a through by browser you can actually look at the data, look at the JSON objects of the Facebook data of yourself, or whatever the Facebook API allows, which we will be able to see through this graph API.

So, again, that we emphasize JSON is the JavaScript Object Notation, which is the way that the data is stored in Facebook, data is stored in twitter when you request through the API, for saying, 'give me this data about PK', it is returning the data in JSON format. It is basically the format that most social media services use today.

(Refer Slide Time: 08:53)

So, when you take the data from JSON, and when you want to interpret the data that is available in this JSON, data that is coming back from Facebook or Twitter, you can actually use JSON dot viewer dot stack dot hu. This is only for you to see visually, what data is coming back; you can take the data that is coming out of Facebook, copy paste it into this JSON viewer, and you will be able to see, what the fields are. When you look at the data that is coming back from Facebook, it is generally a block of data; it is just a lot of data that comes back. So, you can actually take it, and put it into the JSON viewer, to see what are the fields that it is actually giving you. We go through this slowly, when we do the tutorials.

And, of course when you collect the data, so first is API which is a way by which you want to collect the data, and the data is coming back in JSON. When you collect the data, you have to store it in some format, right. So, the format that majority of the times, the data is stored, is in MySQL. Basically, it is a relational database to store the data, and data is stored in rows and columns, and simple queries, you could use to get the data.

For example, in this case, I am just selecting user id, user screen name from the data that is being collected through Facebook, right. So, that helps meaning, again I am emphasizing that, this is not a course on MySQL itself; we will only look at some simple queries on how to look at the data that you have actually stored through the programs that you have written.

MongoDB is one of the popular ones, more recently we have started looking at and people are actually using this. So, MongoDB is another way by which the data is stored and the data that is collected from Facebook is actually stored.

So, again, let me emphasize which is API; then, there is programming language; then, there is MySQL database or MongoDB, which is data is coming through an API, collected and dumped into this MySQL or MongoDB. So, now, we also need a way by which to look at the data that is being stored. So, one of the ways you could use this

actually phpMyAdmin, which actually allows you to look at the data that you have in your own database.

(Refer Slide Time: 11:20)
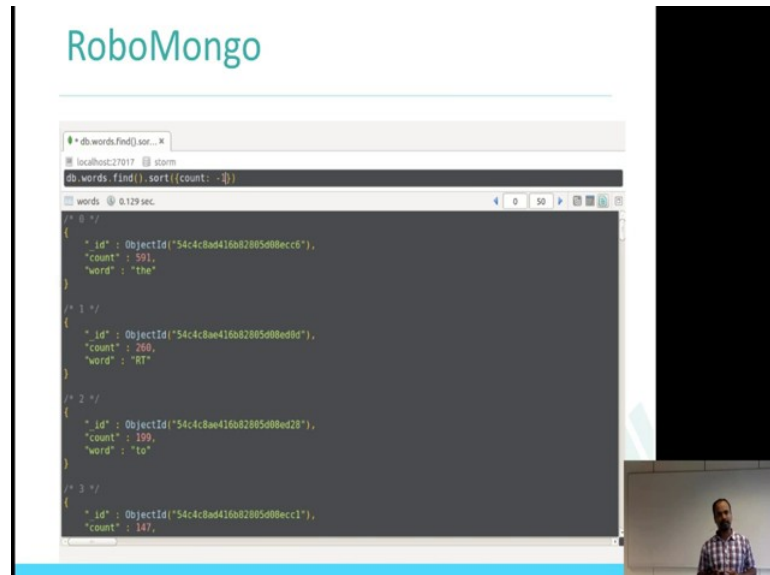
So MySQL phpMyAdmin can look at the data from MySQL, and RoboMongo will help you to look at the data from a MongoDB. So, essentially, these are the ways by which you can collect the data, store the data and look at the data that is available with you.

(Refer Slide Time: 11:38)
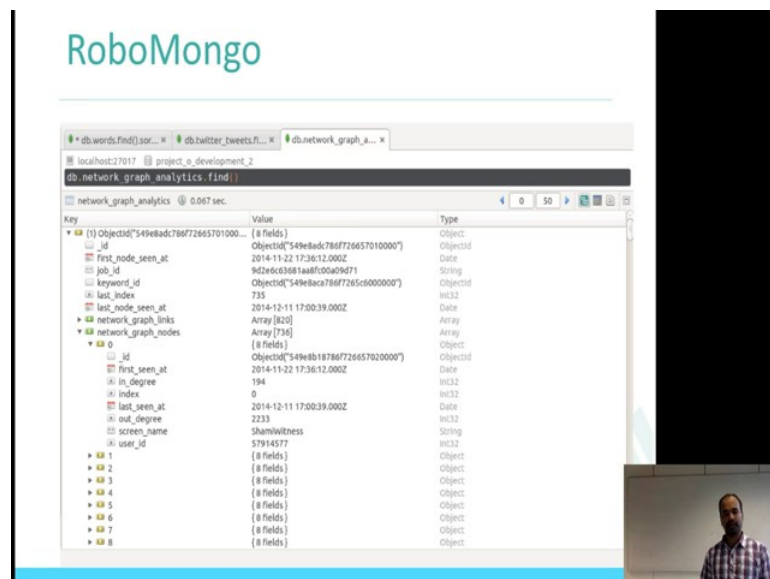
So, this is another view of RoboMongo, which shows you what are the different fields that are available; what data is stored in those fields.

(Refer Slide Time: 11:53)
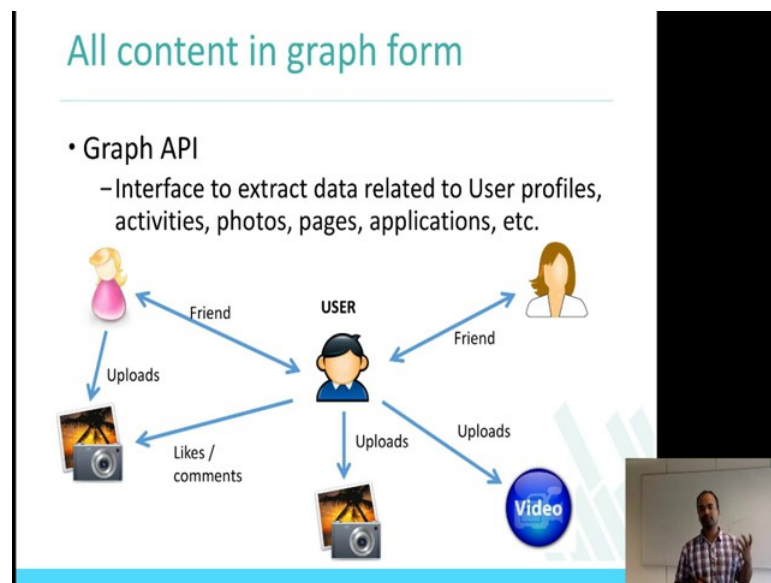
All content on Facebook is actually stored in a graph format; that is, user - the friends that I would have, the pictures that I upload, the videos that I upload, and the status updates that I do, everything is actually a node in the graph. And, every interaction, which is basically like the comments, likes and things like that, becomes edges in this graph. Facebook actually stores all interactions, of all data that they have within the graph format; that is why the API that they have is also called as a graph API.

(Refer Slide Time: 12:27)
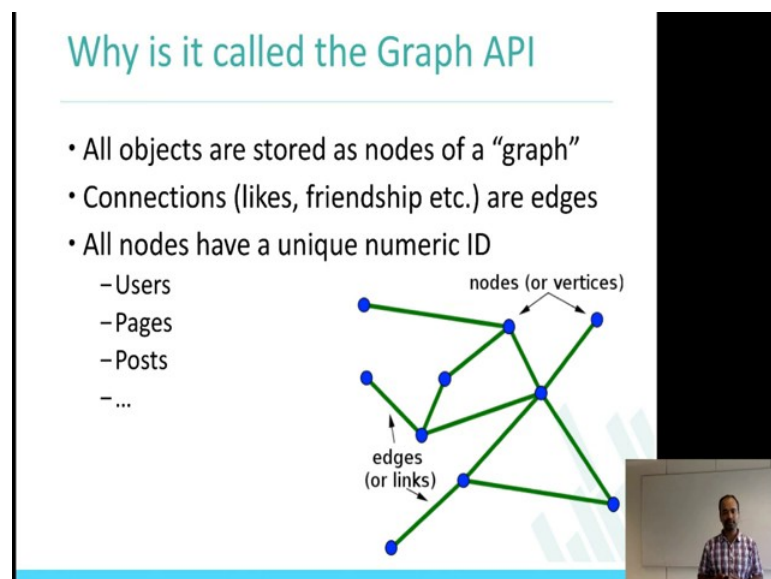
Here is the another view of the same message, which is, all objects are stored as nodes in the graph; connections like friends, friendships, likes are edges and all nodes have a unique numeric ID, which is users, pages and posts. And, we will be talking mostly about users; we shall later talk about also, pages, which is one of the ways by which content can be generated on Facebook.

(Refer Slide Time: 12:52)
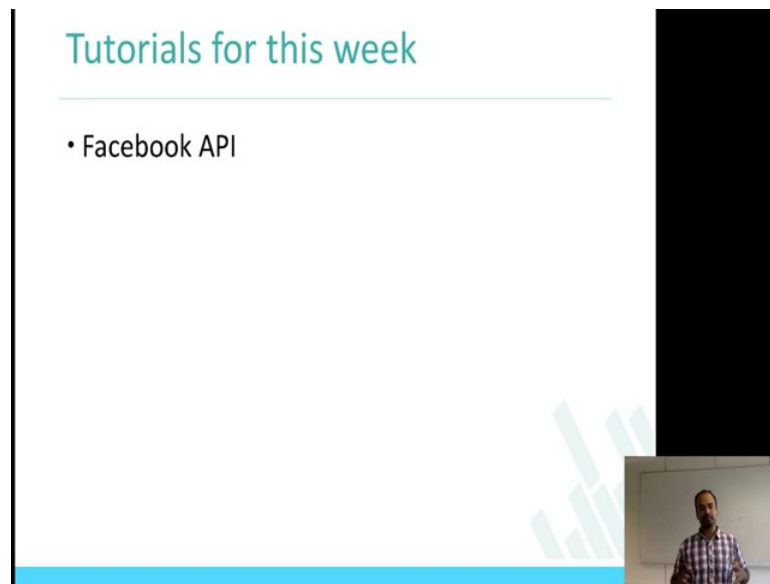
In tutorials this week, you will actually look at in detail about what a Facebook API is, how do you actually create the secret key, what kind of authentication that you will have to provide Facebook, in terms of collecting data, what data can be collected and things like that.