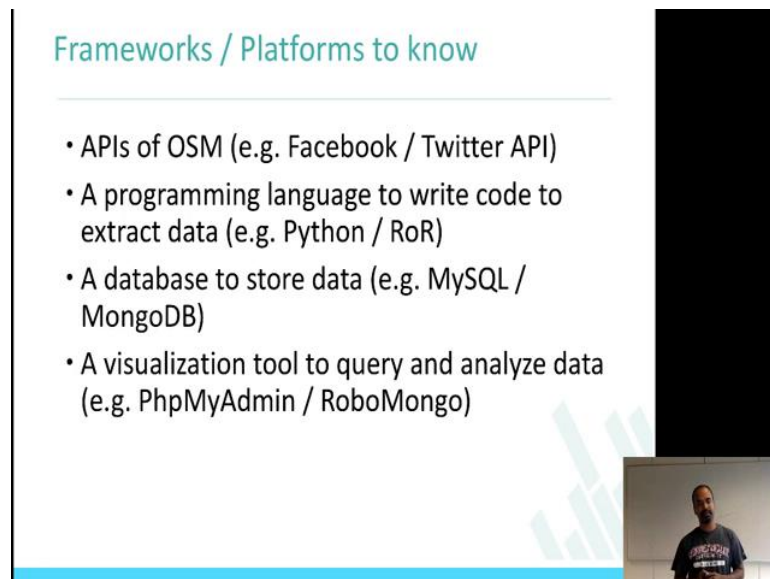


**Privacy and Security in Online Social Media**  
**Prof. Ponnuram Kumaraguru (“PK”)**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Madras**

**Week – 3.1**  
**Lecture – 10**  
**Misinformation on Social Media**

Welcome back to the course Privacy and Security in Online Social Media, this is week 3.  
Let me put you go over what we covered in week 2.

(Refer Slide Time: 00:20)



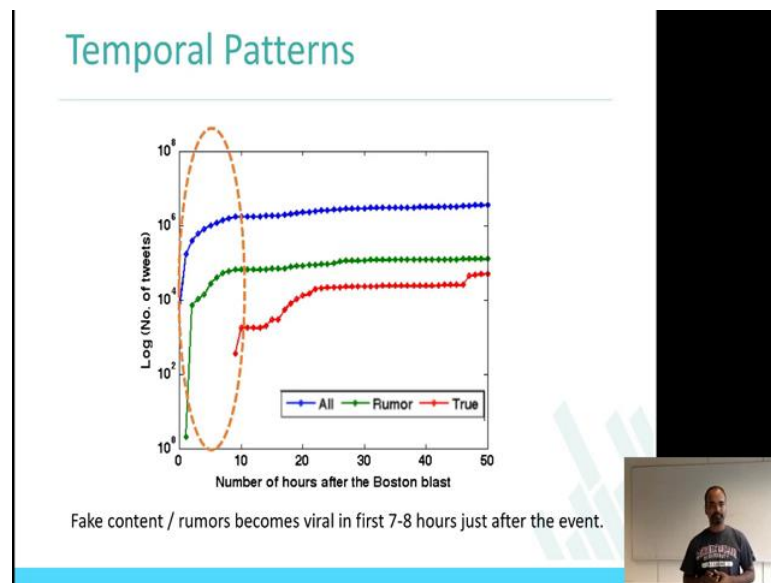
**Frameworks / Platforms to know**

- APIs of OSM (e.g. Facebook / Twitter API)
- A programming language to write code to extract data (e.g. Python / RoR)
- A database to store data (e.g. MySQL / MongoDB)
- A visualization tool to query and analyze data (e.g. PhpMyAdmin / RoboMongo)

The slide features a light blue header and footer. A small video inset in the bottom right corner shows Prof. Ponnuram Kumaraguru, a man with a beard wearing a blue t-shirt, standing in front of a whiteboard.

We **started** looking at what **an API is**, and then we looked at what Python Programming Languages, MongoDB, how the data is stored, how we can actually visualize the data using PhpMyAdmin or RoboMongo. I hope all of you have already done hands on exercises and practices with Facebook API and twitter API.

(Refer Slide Time: 00:45)



After that we looked at the topic Trust and Credibility, in that we looked at multiple events; through the events we looked at some concepts. Here is a slide that I used in week 2, where we showed that the truthful information is coming into the social media slower than the rumours, and there are multiple techniques by which you can actually attack this problem.

(Refer Slide Time: 01:10)

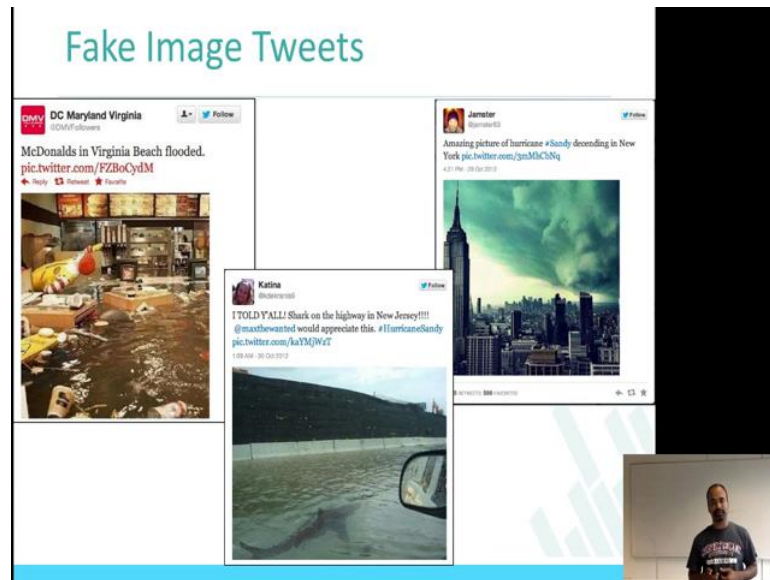
### Misinformation Tweets

The slide displays three tweets. The first tweet, from 'DC Maryland Virginia', shows a flooded McDonald's and is labeled 'FAKE' with a red circle. The second tweet, from 'AP The Associated Press', reports 'Breaking: Two Explosions in the White House and Barack Obama is injured' and is labeled 'RUMORS' with a blue box. The third tweet, from '@Twiggy\_Garcia', reports '#LondonRiots hearing reports that london zoo was broken into and a large amount of animals have escaped. Too far! Thats not cool :-('. A yellow diamond with a dollar sign is also present.

Tweet Source	Content	Label
DC Maryland Virginia	McDonalds in Virginia Beach flooded. pic.twitter.com/FZBoCydM	FAKE
AP The Associated Press	Breaking: Two Explosions in the White House and Barack Obama is injured	RUMORS
@Twiggy_Garcia	#LondonRiots hearing reports that london zoo was broken into and a large amount of animals have escaped. Too far! Thats not cool :-(	

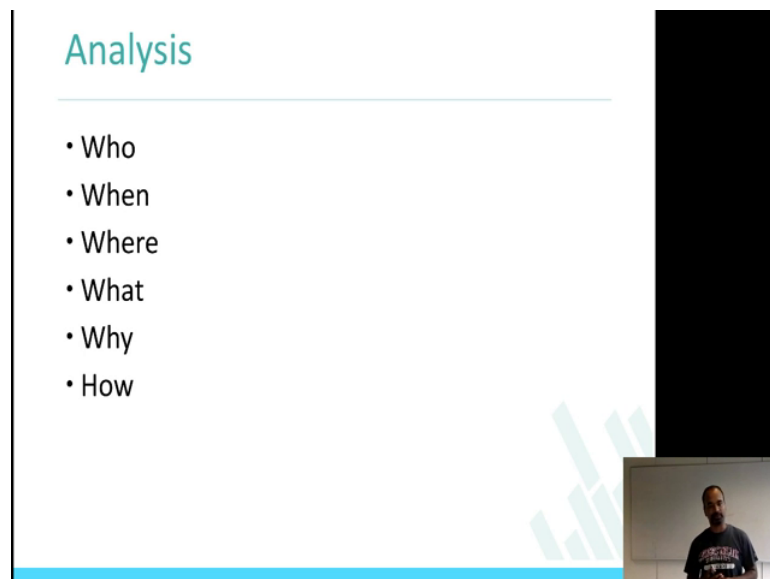
And I also showed you some examples about Misinformation; tweets that were being posted on social media and there have been multiple effects of it; fake content getting viral, some values on the stock market getting affected and of course rumours also.

(Refer Slide Time: 01:30)



More examples that I showed you in week 2, showing that there is a shock in the hurricane sandy where this picture got viral and it had effects on the public also.

(Refer Slide Time: 01:45)



Most **specifically** I was trying to give you an intuition about how, what analysis can be done. Particularly; who, when, where, what, why, and how. These are the kind of analysis that you should be interested in doing while looking at the social media content.

(Refer Slide Time: 02:04)


### Classification

User Features [F1]	Tweet Features [F2]
Number of Friends	Length of Tweet
Number of Followers	Number of Words
Follower-Friend Ratio	Contains Question Mark?
Number of times listed	Contains Exclamation Mark?
User has a URL	Number of Question Marks
User is a verified user	Number of Exclamation Marks
Age of user account	Contains Happy Emoticon
	Contains Sad Emoticon
	Contains First Order Pronoun
	Contains Second Order Pronoun
	Contains Third Order Pronoun
	Number of uppercase characters
	Number of negative sentiment words
	Number of positive sentiment words
	Number of mentions
	Number of hashtags
	Number of URLs
	Retweet count


And then, we later looked at different features that are available in tweets particularly user features and that tweet features and we tell **detail** about what these features mean, how **these** features can be put together to create a classifier which can look at tweet and then say that whether it is legitimate or fake tweet.

(Refer Slide Time: 02:24)

### Sample Fake Tweets



> 30,000 RTs



> 50,000 RTs

Some more examples, particularly this is from the Boston Marathon where I showed you that a tweet which said, RIP to the 8 year-old who died in Boston explosion was retweeted more than 30000 times and malicious user used this spread or **occurrence** of an events to actually spread the content and get victims to go to malicious URL's which share **malicious** information.

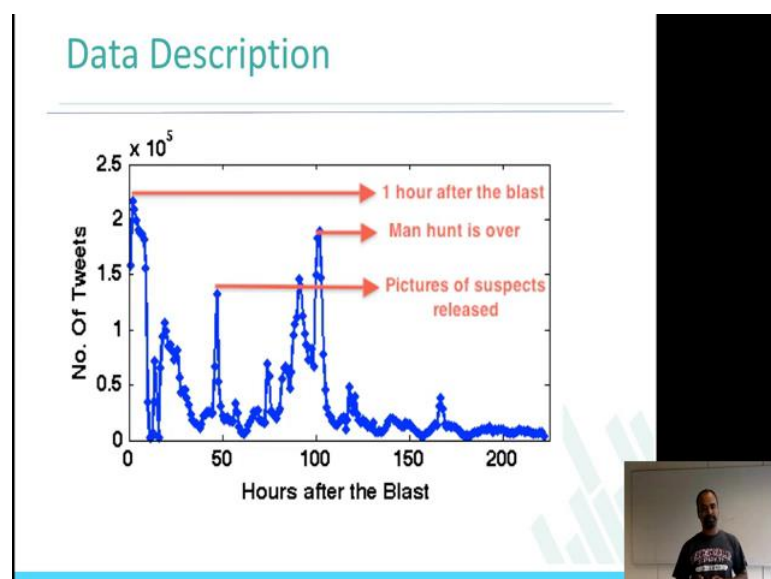
(Refer Slide Time: 02:51)

### Data Description

Total tweets	7,888,374
Total users	3,677,531
Tweets with URLs	3,420,228
Tweets with Geo-tag	62,629
Retweets	4,464,201
Replies	260,627
Time of the blast	Mon Apr 15 18:50 2013
Time of first tweet	Mon Apr 15 18:53 2013
Time of first image	Mon Apr 15 18:54 2013
Time of last tweet	Thu Apr 25 01:23 2013

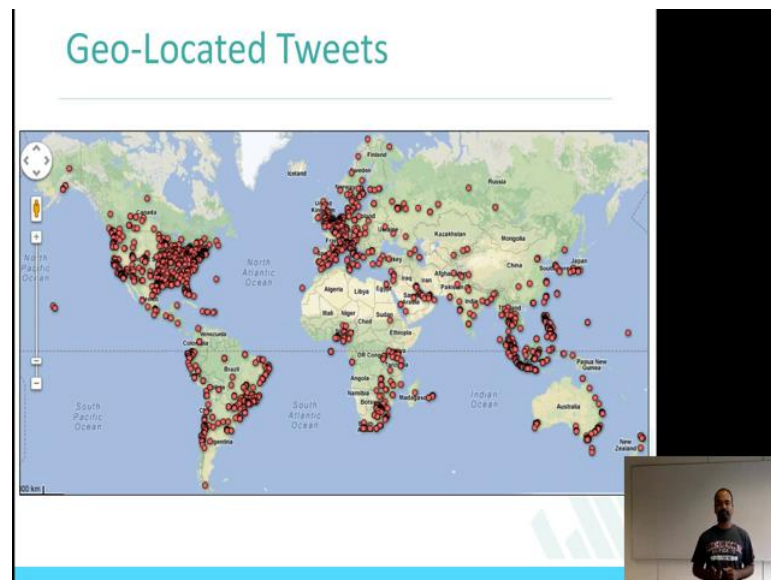
This is one slide when I talked about, how the data can be represented, what data has been collected for doing these kinds of analysis.

(Refer Slide Time: 03:02)



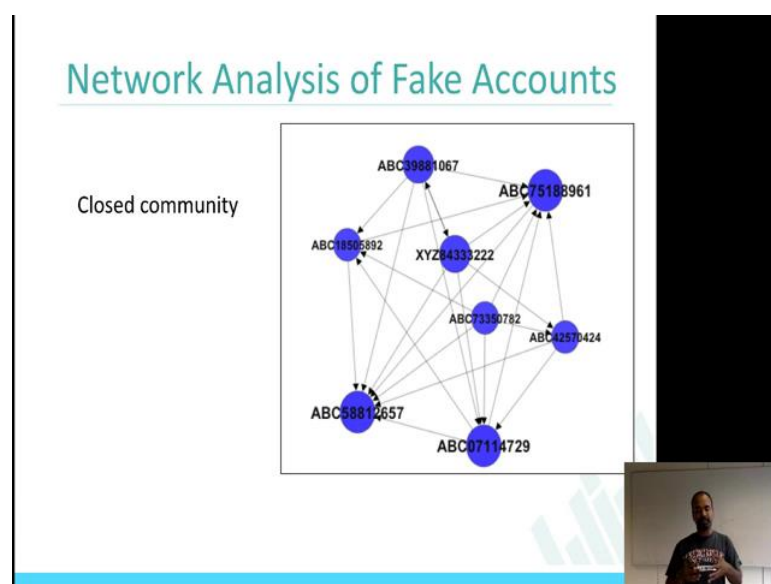
I also mentioned about the spikes in the social media data content that is generated on social media is very, very populated with the **actually** event that happens in the real world. This is one example where man hunt is over, a lot of people are talking about it and therefore **there** is spike in the tweets that are showing up.

(Refer Slide Time: 03:23)



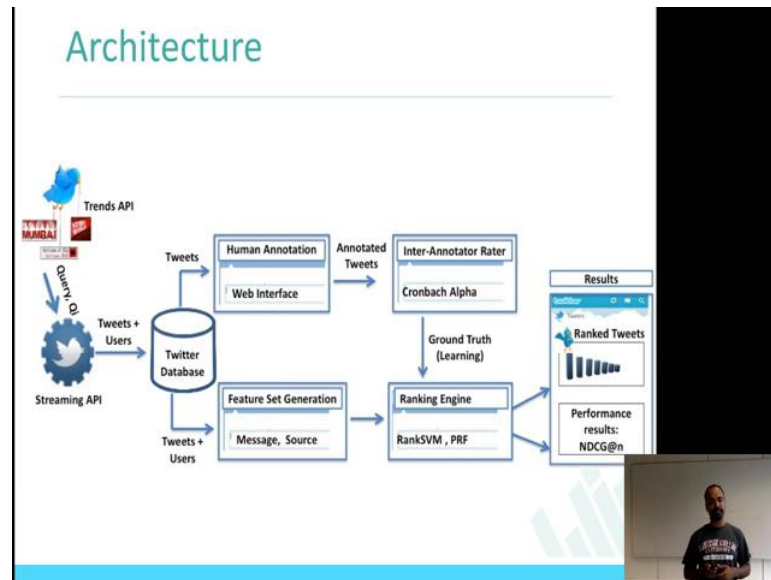
This is Geo-Located tweets where each **dot** is a tweet which **has** a geo tag attached with it and such kind of graphs can be helpful in saying where these tweets are coming from.

(Refer Slide Time: 03:36)



We also talked about how the community of users who are posting this fake content, who created fake accounts, how they are connected. Interestingly, they are all connected very closely and it is a closed community.

(Refer Slide Time: 03:53)



And I also walked you through a multiple architecture diagrams mentioning about how data is collected from social media, what kind of annotations and how do you actually verify with the data that you annotated is actually of high quality and, what kind of feature generations can be done, what is the model that we developed and what is the model that one can develop, and now what are the evaluation matrix to actually find out whether the technique that we have applied and the model that we have created is actually good.


So, this is an architecture that I tell in detail talking about each block and explaining all this block helps in creating some interesting solutions for the problems in the trust and credibility space.



(Refer Slide Time: 04:39)

## TweetCred

- Available as a Chrome Extension

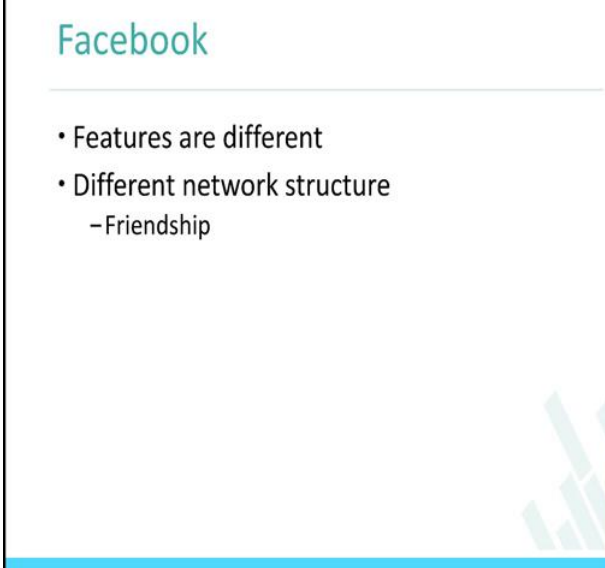


And then I showed you about plugin which is called TweetCred, I hope some of you have played around with the tweetcred plugin to find out how the tweets are evaluated and the value of  $x$  on 7 is presented to the users.

(Refer Slide Time: 04:58)

## Facebook

- Features are different
- Different network structure
  - Friendship



So, now what I want to actually cover is little bit about how one can actually take this understanding of twitter and then **apply into other** social networks, because this is a privacy and security in online social media course so I thought that it will be interesting to find out how these kind of techniques that we learnt from twitter can be acquiring into

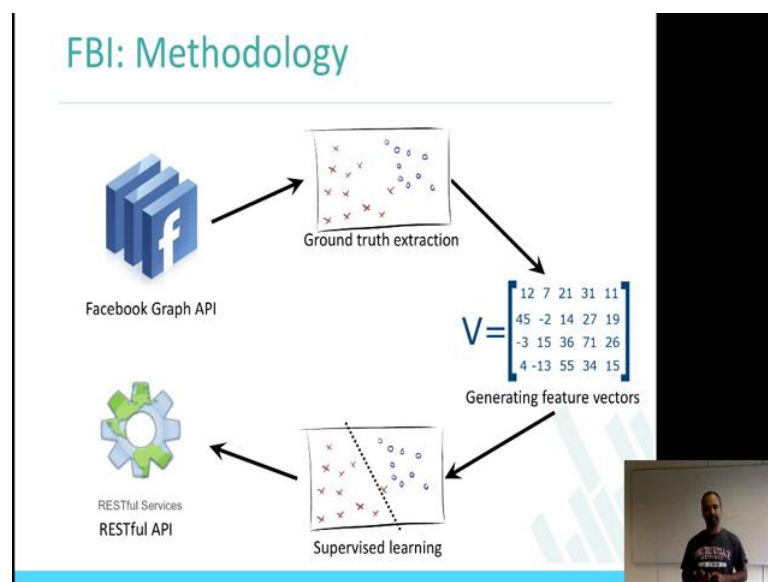


other social networks particularly I will talk about Facebook. If you think about it initially I talked about how Facebook and twitter are different in my week 1 lecture, where we said that Facebook is a **bi-directional** network and twitter is a unidirectional network and the structure itself is very different.

And, the features that are available in these two social networks to study are also different. In twitter it is followers and following and Facebook, it is actually friends and the information that these networks provide through API are also very different. And the structure of the networks have different, particularly I wanted to highlight this friendship **thing** in Facebook; the connections **are** more personal and if there is a post that shows up by your friend, **there is** some tendency that it is more likely to be truthful and then we you believe **that your friend's** post is actually more truthful than a random person's post.

So, that is the one of the differences between the Facebook and twitter network, particularly keeping this trust and credibility as the space of discussion. I wanted to highlight this difference, and now given this difference we should also look at how we can actually use the model that we have understood in twitter to **apply it into** Facebook.

(Refer Slide Time: 06:37)

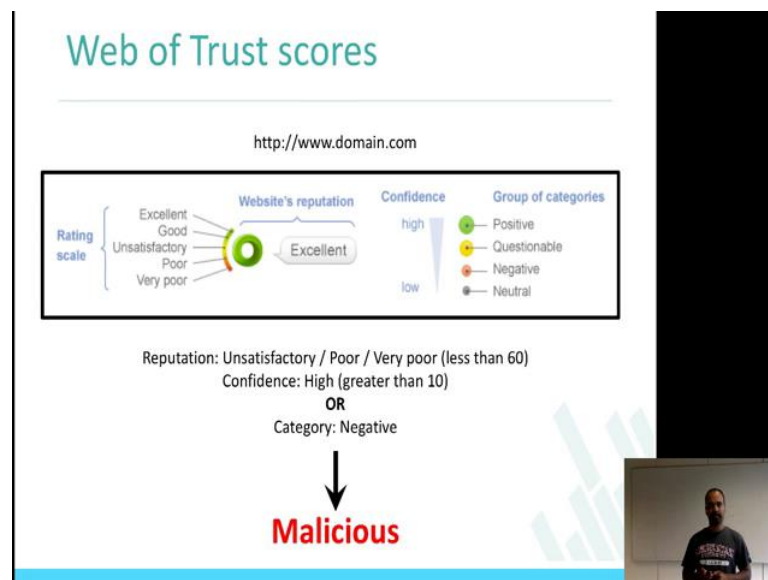


The architecture if you see it is almost the same, in this case it is just presented slightly differently. FBI: stands for Facebook Inspector, a similar tool that is like tweetcred which takes the Facebook post from Facebook **graph** API and then looks at the posts and

make some judgement on how whether these post are malicious or not, credible or not, trust worthy are not.

In this case it is the same architecture which takes the post, do some feature extraction, do some ground truth understanding of the post, then creates some feature vectors out of it, create a model out of it, in this case supervised learning model because we actually have data from the posts that we are collecting and then **create a** RESTful API through which you can actually find out whether this post is malicious or not. Same architecture, very similar to tweetcred so **I do not think** we should actually **spend a** lot of time in understanding more details of this. If there is any question please **feel free to ask in the forum** for sure.

(Refer Slide Time: 07:43)



So, one thing that was also mentioned in the tweetcred or the twitter trust and credibility slides is a Web of Trust that is called WOT. Then I thought I would just mention it briefly what does it mean. It basically takes a domain and produces an output which says that a value similar to tweetcred, similar to other services that you may have seen where input is a domain and the output is score, which you can use to say that whether it is a malicious domain or not. Then in the past also I mentioned about how long the domain has been registered, who registered the domain and things like that. These features can be used to make the **judgement**.

So web of trust basically gives you value of excellent, good, satisfactory, poor and very poor. If you give domain saying iit dot edu dot in, it will actually come back with the rating scale and a confidence scale. We use this in Facebook inspector because in Facebook inspector it is also going to look at URL as the feature or particularly the domain as a feature from the post that we are **analyzing**.

(Refer Slide Time: 08:57)

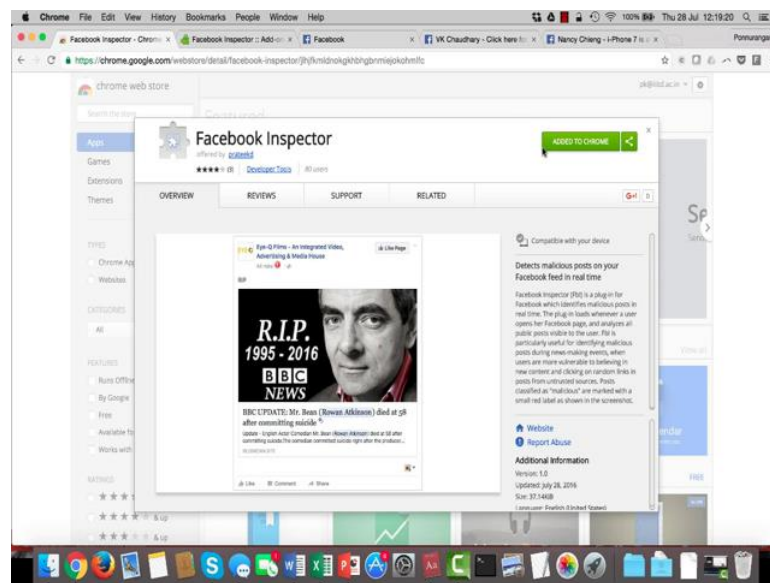


The slide is titled "Plugin" in a teal font. Below the title, there are two bullet points, each with a URL. The first URL is for the Chrome Web Store, and the second is for the Firefox Add-ons site. A small video inset in the bottom right corner shows a man in a dark shirt standing in front of a whiteboard.

- <https://chrome.google.com/webstore/detail/facebook-inspector/jlhjfkmlndnokgkhhbghbnmiejokohmlfc>
- <https://addons.mozilla.org/en-US/firefox/addon/fbi-facebook-inspector/>

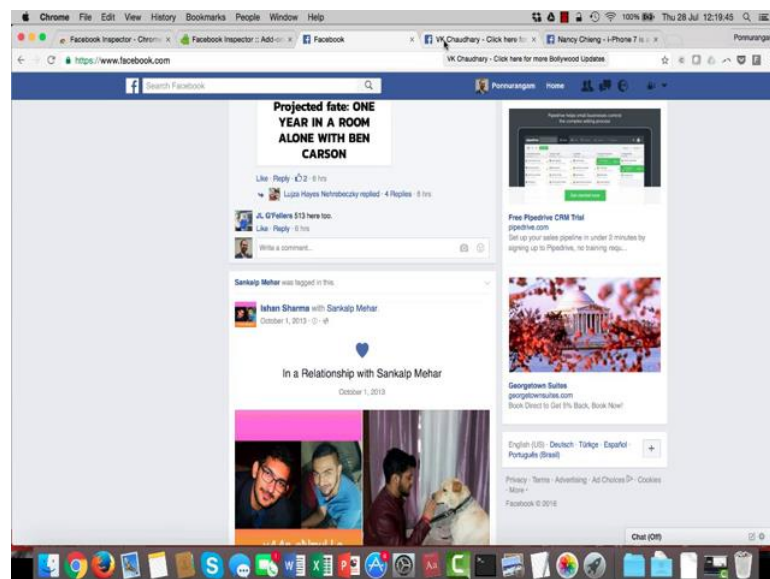
So, here is the pointer to the plugin. It will be interesting if you can actually download it and play around with it. These are links to the Chrome extension and to the Firefox. Let me just walk you through how this plugin works, what does it do, how is it different or how is it very similar to tweetcred.

(Refer Slide Time: 09:21)



This is Facebook inspector on the Chrome store, and basically you can add this to your browser, I mean I already have it on my Chrome otherwise it should say **add to** Chrome. When you add it, when you go to your Facebook timeline, you should be able to see some difference in the post that you are **seeing**.

(Refer Slide Time: 09:47)



For example here is my Facebook timeline for now and newsfeed, if you see there is no many, if you look at the post there is no annotations done in the posts that you can see on

my timeline. Whereas let me show you some examples where the Facebook inspector is actually showing you some information.

(Refer Slide Time: 10:09)



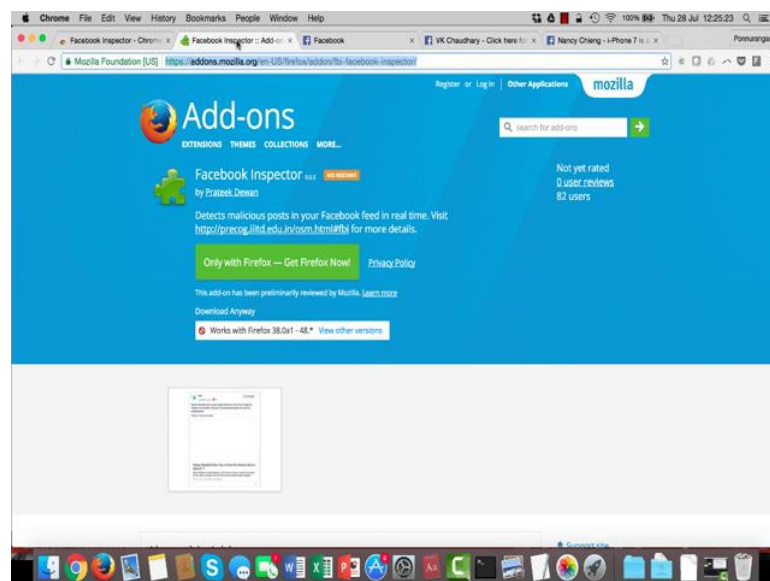
Here is an example of a post which is done by VK Choudhary and the post we just click here for Bollywood updates. Hema Malini congratulates Deepika. Is Deepika Padukone engaged? In this post if you see, there is some annotation done by the Facebook inspector, it says confidence is low, the decision that was made with the model that was generated is low and it is using features, click on the image for more details. If you are interested, you can actually click on the image, see for more details. Here **is another** example that I will show you.

(Refer Slide Time: 10:42)



In the first case it is probably a rumour, that is why it is actually finding out and saying Facebook inspector is producing this result with red mark. Here is an example which could be a **spam**, which is as of now we really do not know iPhone 7 designs and features that are available, but this post says about iPhone 7 is awesome and amazing which could have one. And this post is being annotated by Facebook inspector saying it is a malicious post. That is how Facebook inspector works.

(Refer Slide Time: 11:20)



And here is also a plugin that you can use if you are a Firefox user. Therefore, if a Facebook inspector is available as a Chrome browser plugin and as a Firefox plugin add on which you can use on your browser.

So, that is the way you could think about taking way and understandings from twitter, where we studied about how to build techniques using the features from twitter to create an understanding of whether the post is a credible or not, here I showed you about Facebook.