

Introduction to Data Analytics
Prof. Nandan Sudarsanam and Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 07

Lecture – 39

What is Big Data? A Small Introduction

Hello and welcome to this module on big data analytics. So, what I would do in the next two modules, is trying to introduce you to the problems of big data analytics. We thought of looking at what is big data. Of course, I cannot hope to cover all the aspects of big data analytics in a couple of modules in this course. The whole idea is to give you a very brief or as I say a small introduction to big data analytics.

(Refer Slide Time: 00:42)

The slide is titled "Big Data?" and features a logo on the left and the "RISE" logo on the right. It contains three rows of content:

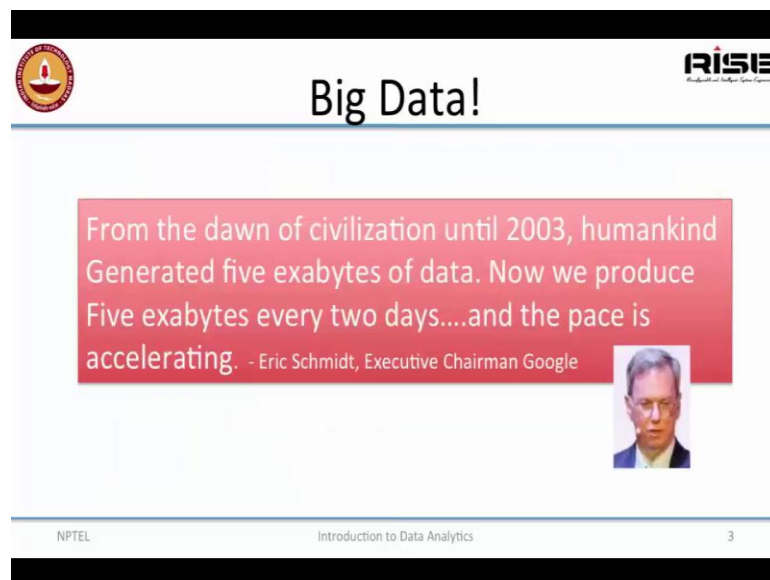
Database Management Systems	Organizing data; Ease of access; Answer aggregation queries
Data Mining	Detect patterns in data; Ideas from machine learning and statistics
?	What Next?

At the bottom of the slide, it says "NPTEL Introduction to Data Analytics 2".

So, when people started talking about data, organizing data, and managing data they initially started with data base management systems. Beta offered of rational data bases and other forms of data management systems. So, the goal here was is of excess, and to be able to answer simple aggregation select queries; essentially trying to do some form of analytics of the data. But while looking at very specific, very static, and run on the data. So, the volumes of data were not as large as we talk about now a days.

And then from data base management systems, when the data became much larger that you could not make sense out of the raw data itself; and people started talking about data mining or data analytic systems. Where the goal was to detect? Patterns in the data and people used lot of ideas from machine learning and applied statistics in order to do this kind of data mining, are to understand the data better. So, what was happen now a days? So, this whole system is evolved to a point where you cannot just hope to run your machine learning algorithms directly on the row data. But you have to worry about the data management aspects of it; as well as the analytics aspects of it. So, that is essentially the crux of what people call big data in now a days. The fact that the data is so large, that you cannot look at analytics divide of data management.

(Refer Slide Time: 02:19)



The slide features a black header bar with a circular logo on the left and the text 'RISE' on the right. Below the header, the title 'Big Data!' is centered in a large, bold, black font. A red rectangular box contains the following text in white: 'From the dawn of civilization until 2003, humankind Generated five exabytes of data. Now we produce Five exabytes every two days....and the pace is accelerating. - Eric Schmidt, Executive Chairman Google'. To the right of this text is a small portrait of Eric Schmidt. At the bottom of the slide, there is a footer bar with 'NPTEL' on the left, 'Introduction to Data Analytics' in the center, and the number '3' on the right.

So, give you a feel of what this big data, it is court from rich man oppose the executive chairman at Google. So, from the dawn of civilization until 2003, human kind generated five exabytes of data. Now we produce five exabytes of data every 2 days, and the pace is accelerating. So, if you can just think about it, what we have produce for 1000s of years since till 2003, we are able to produce that kind of data every 2 days. That is a volume of data that humans are producing and obviously, we cannot make sense out of this data under some kind of organization and analytics goes handle with us.

(Refer Slide Time: 03:02)

The slide features a header with a circular logo on the left and the RISE logo on the right. The title 'Challenges – The 4 V's' is centered. Below the title is a bulleted list of five items. The footer contains the text 'NPTEL Introduction to Data Analytics' and the number '4'.

- Volume
- Velocity
- Variety
- Veracity
- *Value?*

So, one of the main challenges that come about, when we are talking about data of this scale. So, this is the some out of other the coming at shape that view of big data analytic, but one that is popular. So, I thought I should present it as part of this model. So, the challenges are essentially in capitulated under the 4V's of big data: that is volume, velocity, variety, veracity. And I would like to add a 5th V to it; is it called value, which is essentially equal of all of data analytics in subsets. So, I will briefly introduce you to each of these ways and another may be the end of the first module.

(Refer Slide Time: 03:50)

The slide features a header with a circular logo on the left and the RISE logo on the right. The title 'Volume' is centered. Below the title is a bulleted list of two items. Surrounding the list are several logos for digital services. The footer contains the text 'NPTEL Introduction to Data Analytics' and the number '5'.

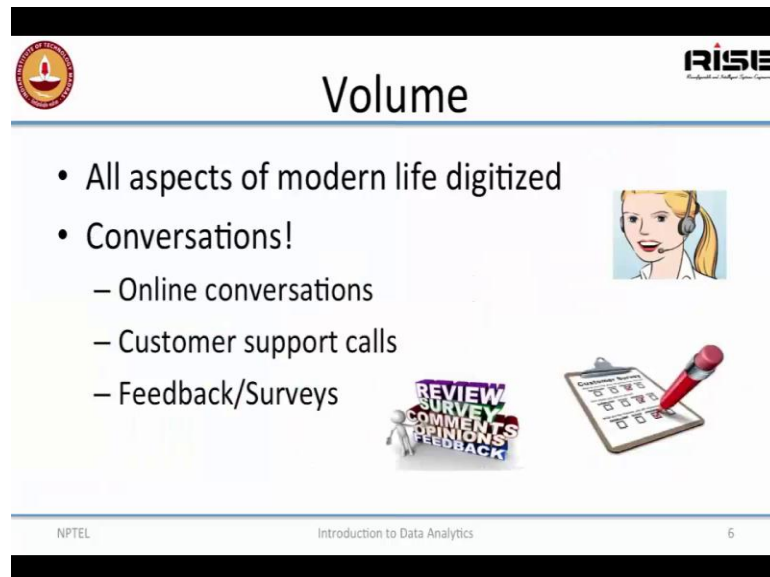
- All aspects of modern life digitized
- Social Activity

Logos shown: WhatsApp, Facebook, YouTube, Picasa Web Albums, flickr, YAHOO!

So, why you said that we are able to generate such huge volumes of data. So, one of the reasons is that all aspects of modern life has been digitized, all aspects of modern life has

been digitized. And no where is it more a parent than in your social activity. So, we have Facebook; lot of peoples friend significant portion of their breaking hours on Facebook and possibly significant portion of the sleeping hours reviewing about Facebook. And then we have such a volumes of data that is probably Youtube, and WhatsApp and twitter and flickr and all the web sharing totals at. So, if you think about it at almost every aspect of your life is now being recorded somewhere or the other.

(Refer Slide Time: 04:38)



The slide features a black header bar with a circular logo on the left and the 'RISE' logo on the right. The main title 'Volume' is centered in a large, dark font. Below the title, there are two bullet points: 'All aspects of modern life digitized' and 'Conversations!'. Under 'Conversations!', there are three sub-points: 'Online conversations', 'Customer support calls', and 'Feedback/Surveys'. To the right of the sub-points are three icons: a woman in a headset, a 3D blocky graphic with the words 'REVIEW SURVEY COMMENTS OPINIONS FEEDBACK', and a clipboard with a red pen. At the bottom, there is a footer bar with 'NPTEL' on the left, 'Introduction to Data Analytics' in the center, and the number '6' on the right.

Volume

- All aspects of modern life digitized
- Conversations!
 - Online conversations
 - Customer support calls
 - Feedback/Surveys

NPTEL Introduction to Data Analytics 6

And the next thing is just not the social activities that you perform social media, any kind of interactions you are having. This with very high probability is being recorded; for example, phone conversations. We call costumer support calls they are being recorded any feedback and surveys; that will have full fill out of the form anywhere as being that preserve for posterity, any kind of online conversation, chats or sub transcript and other things they are also recorded.

(Refer Slide Time: 05:10)

The slide features a title 'Volume' at the top center. On the left is a circular logo with a lamp. On the right is the 'RISE' logo. Below the title is a list of bullet points: 'All aspects of modern life digitized', 'Multi-media', and 'Growing at an amazing rate'. The third point has four sub-points: 'Instagram: More than one billion photographs', 'Youtube: 100 hours of new videos are uploaded to the site every minute', 'Over 1 billion unique users visit each month', and 'Facebook: 30 billion pieces of content every day'. To the right of the text are logos for Flickr, Instagram (with the tagline 'Fast beautiful photo sharing'), YouTube, and Picasa (with 'Web Albums'). At the bottom, there is a footer with 'NPTEL', 'Introduction to Data Analytics', and the number '7'.

And so, it seems like everything that are working at a, it is actually being recorded at some point of other. Another aspect is the increasing availability of cheap story and cheaper band with that loves us to share multimedia content. At rates and volumes that is never imagine possible even 5 years back. So, here are few examples that I took from the internet. So, Instagram is growing at more than one billion photographs. In Instagram has more than one and over one billion unique users visit Youtube each month. at Facebook adds 30 billion pieces of content every day and likewise you could take any of these online portals and you can come up these main boggling numbers, that tell you how much data is being shared on these portals.

(Refer Slide Time: 06:12)

The slide features a title 'Volume' at the top center. On the left is a circular logo with a lamp. On the right is the 'RISE' logo. Below the title is a list of bullet points: 'Biological Data'. This point has two sub-points: 'The three gigabases (3×10^9 base pairs) of the human genome can be sequenced in a few days.' and 'Protein data banks have 10s of thousands of structures amounting to several terabytes of data'. The second sub-point has a further sub-point: 'Data is being generated at a tremendous rate.'. The background of the slide is a faint image of a DNA double helix. At the bottom, there is a footer with 'NPTEL', 'Introduction to Data Analytics', and the number '8'.

So, apart from these the aspects of everyday life and regenerating this data, there is being significant advance in technology in other area; for example, in biological data. So, 3 gigabases of a human genome can now be sequenced in a few days. This is like 3×10^9 base pairs which is significant volume of data, but then this is for one human genome. And you can essentially get your genome sequence for few thousand dollars in the US and so it should be made the access full much lower price range. And so the data has being generated at every enormous unbelievable main boggling space. Protein data banks have 10s of thousands of structures amounting to several terabytes of data and again people are complaint add to this experimentation.

(Refer Slide Time: 07:11)

The slide features a header with a logo on the left and the word "Volume" in the center. Below the header, there is a bulleted list under the heading "Biological Data". The list includes two main points: one about sequencing the human genome and another about protein data banks. A red box contains a quote from Donald Knuth. At the bottom, there is a footer with "NPTEL", "Introduction to Data Analytics", and the number "9".

Volume

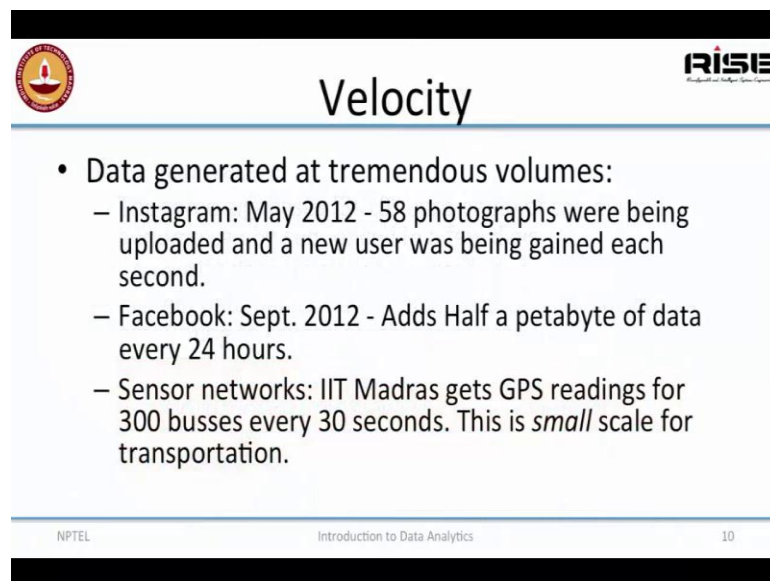
- **Biological Data**
 - The three gigabases (3×10^9 base pairs) of the human genome can be sequenced in a few days.
 - Data is being generated at a tremendous rate.
 - Protein data banks have 10s of thousands of structures amounting to several terabytes of data

Knuth on Computational Biology: "I think the most exciting computer research now is partly in robotics, and partly in applications to biochemistry... Biology is so digital, and incredibly complicated, but incredibly useful... Biology easily has 500 years of exciting problems to work on, it's at that level."

NPTEL Introduction to Data Analytics 9

So, to put this in prospective; Donald Knuth one of the fathers of computing algorithms and so on so forth. Has this to say about computational biology. I think the most exiting computer research, now is partly in robotics and partly in applications to bio chemistry. Biology is so digital, incredibly complicated, but incredibly useful. Biology easily has 500 years of exiting problems to work on, it is at that level. And not only is it in the data analytics appear which is most relevant to us, but the variety of other area has to well including data storage and also in computing hardware and access technologies; biology is challenging not regular science at levels never seen before.

(Refer Slide Time: 08:08)



The slide features the IIT Madras logo on the left and the RISE logo on the right. The title 'Velocity' is centered at the top. Below the title, there is a bulleted list of data generation statistics. At the bottom, there is a footer with 'NPTEL', 'Introduction to Data Analytics', and the number '10'.

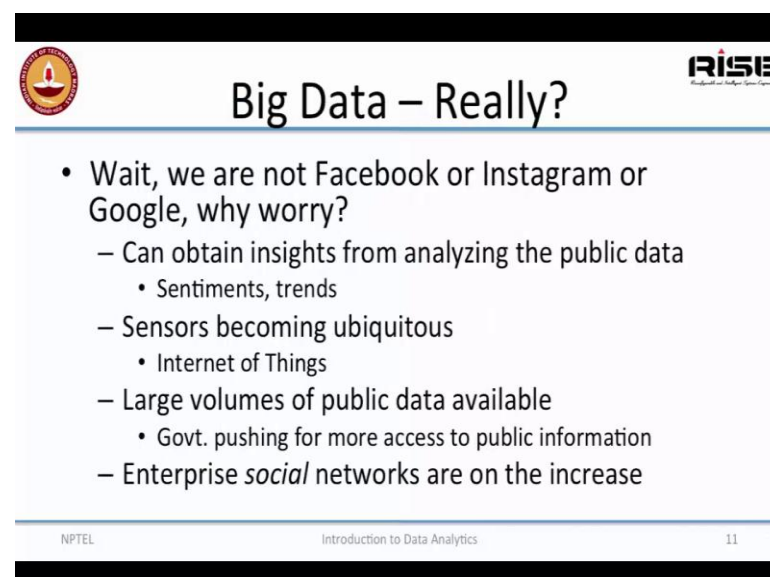
Velocity

- Data generated at tremendous volumes:
 - Instagram: May 2012 - 58 photographs were being uploaded and a new user was being gained each second.
 - Facebook: Sept. 2012 - Adds Half a petabyte of data every 24 hours.
 - Sensor networks: IIT Madras gets GPS readings for 300 busses every 30 seconds. This is *small* scale for transportation.

NPTEL Introduction to Data Analytics 10

So, this is about volume but, what about that rate at which this data is being accumulated. So, data is generated at tremendous volumes, but tremendous rates. Instagram in may 2012, 58 photographs were being uploaded and a new user was being gained each second. In Facebook adds half a petabyte of data every 24 hours. and then that is not even talk about sensor networks; which are generating data most of the continuous rate; for example, at IIT Madras we get GPS readings from about 300 buses every 30 seconds. And if you think about cities city white transportation, this is really a small scale. And if you are able to instrument all the buses that run in Chennai city we can get data at a much higher velocity.

(Refer Slide Time: 08:58)



The slide features the IIT Madras logo on the left and the RISE logo on the right. The title 'Big Data – Really?' is centered at the top. Below the title, there is a bulleted list of questions and trends related to big data. At the bottom, there is a footer with 'NPTEL', 'Introduction to Data Analytics', and the number '11'.

Big Data – Really?

- Wait, we are not Facebook or Instagram or Google, why worry?
 - Can obtain insights from analyzing the public data
 - Sentiments, trends
 - Sensors becoming ubiquitous
 - Internet of Things
 - Large volumes of public data available
 - Govt. pushing for more access to public information
 - Enterprise *social* networks are on the increase

NPTEL Introduction to Data Analytics 11

Just hold on one minute, why were we worried about all these big data. So, we are not really, I mean not everyone of this is not going to work for Facebook or Instagram or Google. So, tremendous amount of data is already available in the public data. Either through in Facebook APIs or twitter APIs or through Government data that will be made available on open forums; and we can obtain significant insights from analyzing this public data. And you could think about sentiments, you could think about trends and the variety of things. Is the kind of insights you can derive is limited by your imagination and access to the data; and there is the significant amount of public data available, that we do not really have to work for Google or Facebook to think about big data.

The second thing is that sensors are becoming ubiquitous. So, a lot of a everyday I think that be c or being instrumented, and then raw sensors, and buildings, there are sensors on bridges, and other kinds of infrastructure are there. And ubiquitous of internet of things, this essentially the whole idea is to connect different kinds of equipment through sensors on low power wireless networks. So, once you have these kinds of data being generated by sensors, then you do not have any depth of volume or velocity of data. And so you really need to develop technique like this; that love us to work with data. Can I say all ready mentioned and large volumes of public data is already available. Government is pushing for more access to public information.

So, the all of these factors really makes sense for us to look at big data and we do not have to work for such large company; is even though the examples I gave you through make you appreciate the volumes that we have to look at, where drawn from this kinds of online social media companies.

(Refer Slide Time: 11:06)

The slide is titled "Variety" and features the RISE logo in the top right corner. It contains a bulleted list:

- Traditional
 - RDBMS: Structured data with well defined fields
- Now?
 - Unstructured text

Below the list is a screenshot of a website titled "Biological Macromolecular Resource". The website content includes:

- Learn Featured Molecules**: A section with a "Full Description" and a "List View of Articles by Title Date Category".
- Structural View of Biology**: A section with a "Molecule of the Month" titled "Ebola Virus Proteins". The text describes the genome of Ebola virus, its membrane structure, and the role of glycoproteins. It also mentions "Full Article".
- Protein Structure Initiative Featured System**: A section titled "Deciphering Microbial DUFs". The text discusses the goals of the PSI researchers, such as understanding the structure of proteins in the gut and bacterium, and the importance of determining the structure of all of their component proteins.

The website footer includes "NPTEL", "Introduction to Data Analytics", and the number "12".

So, the next V, I am talk about a little bit this variety. So, traditionally we talk about relational data bases, we talk about structure data with very well defined fields; AB of a few types in all could have strange, you could have numbers, you could have categorical variables things like that. But now with big data, data coming from the different kinds of sources, you have many variations of this. So, the first and very widely available source of data is unstructured text. So, we can get things from the web we can get scientific articles, news paper clippings, variety of sources review unstructured text. We keep these are not really marked into sink, this is the heading, this is the most important keyword here and these are all the attributes of this key word. I mean so you do not get this kind of organize data, essentially you have to just read free from text and from that from some kind of a representation which you can then use for your data analytics form.

(Refer Slide Time: 12:05)

The slide is titled "Variety" and features the logos of IIT Bombay and RISE. It lists "Traditional" data as RDBMS with structured fields and "Now?" data as unstructured text and multi-media. An aerial photograph of a busy city street with many cars is shown.

- Traditional
 - RDBMS: Structured data with well defined fields
- Now?
 - Unstructured text
 - Multi-media

NPTEL Introduction to Data Analytics 13

Second source of information now a days is multi media. As I am mentioning you get pictures, you get videos, and you get variety of different sources of videos and pictures, and as well as audio song clippings, and so on so forth on the internet now. And so, you have to come up with technology for analyzing all of this, especially at a scale.

(Refer Slide Time: 12:30)

The slide is titled "Variety" and features the logos of IIT Bombay and RISE. It lists "Traditional" data as RDBMS with structured fields and "Now?" data as unstructured text, multi-media, sensor data, scientific data, and linked data. A network diagram with many interconnected nodes is shown.

- Traditional
 - RDBMS: Structured data with well defined fields
- Now?
 - Unstructured text
 - Multi-media
 - Sensor data
 - Scientific data
 - Linked data

NPTEL Introduction to Data Analytics 17

And mention about sense of data and again this is going to look like unstructured data. So, it is going to be company wise time varying signals that we would be measuring from these sensors. And how do you even record these, how when you make sense out of this data? That is the other main challenge and then you have scientific data. And scientific data comes in variety of different forms. We are taking about medical data, it

could come in terms of reading some instruments, it could be black and white images, it could be thermal images or here whatever seeing here is snap shot from a micro array data used in computational biology a lot. And so, making sense of these kind of data its requires not only new computing techniques, but also significant understanding of the domain in which your operating, and therefore we have to work about close being collaboration with the this scientist; who are handling this kind of data. And data analytics cannot be performed in escalation here.

So, lastly I would like to talk about link data; and this is essentially data that comes with some kind of structure, but not the kind of structure that we are normally used to. this is data and that comes with the network structure. So, we talking, you can talk about the entity; let us look at the small close up of this. So, each of this notes here this very complex graph and each of this note here is essentially an entity; and the links tell you how of these entities connected. So, such large volumes of link data which give you relationship between entities, already available in the public domain. And apart from that many sources of data that you generate now a days, have this kind of links structure as social group. So, now the question is how would you mind such large volumes of big data? So, variety that is tremendous lot of variety and each of these come with its own challenge and quite often significant requirement for domain.

(Refer Slide Time: 14:36)

The slide features a header with a circular logo on the left and the 'RISE' logo on the right. The title 'Veracity' is centered below the header. The main content is a bulleted list of data types and their associated challenges. At the bottom, there is a footer with 'NPTEL', 'Introduction to Data Analytics', and the number '18'.

- Social media data
 - Falsification!
 - Hard to disambiguate
- Sensor data
 - Noisy
 - Missing Values
- Biological data
 - Unsure data
- Verification hard due to scale
 - Current technology: Internal consistency checks and manual verification!

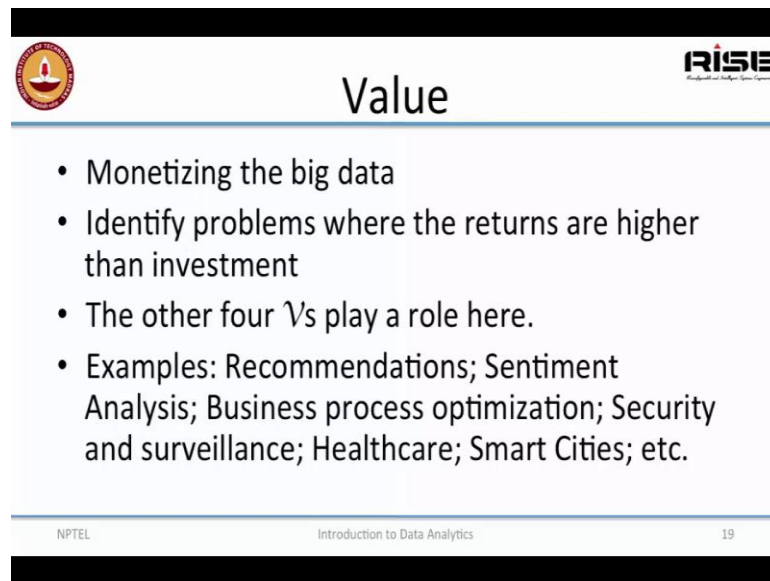
So, on the next V, we look at this veracity. One of biggest problems in social media data is figuring out if the data that is given to you is true or not. So, the falsification data I am pretty sure that many of you have actually given false information to create email ids and

other things online, and this being cases of people maintaining multiple profile on Facebook; and so, that selectively share information etc. So, it is hard to disambiguate when you are giving false and when you are giving true information. So, the significant mode of resources of spend by many of these companies in order to verify the information this provided to that. Or if you take sense of data, data could be noisy, that could be missing values. So, how do you account for all of this? And it is how do you trust the data that you are getting from the sensor is it; or we sure that the sensor is not malfunctioning. So, how do you account for that? Maybe there are few sensors; you can do that manually, but suppose you are talking about million of sensors diploid over high rise building.

Then the question of manual verification is moved it is cannot do that. And talk about how biological data is one of the biggest sources of future problems for us, but even though there are huge volumes of data available more often than not. These are estimated interactions, estimated data; and again you are not sure about the veracity of the data. not sure with that lets if you say protein interact with another protein, you are not 100 percent sure interaction happen. You can exhale 80percent confidence, I am sure this interaction happens. So, in such cases how would you handle this kind of unsure data? So, this is just example. So, if we take any source of data coming from any kind of domains will always find that, there are issues of noise and trust worthiness in the data.

So, the verification the people still work with this kind of data already. We can, some of you if you already worked or working in some kind of a data analytics company you know that; you have to work with this kind of data already. But then verification becomes hard due to the scale that we are talking about. So, current technology people typically end up doing some kind of internal consistency checks and some amount of manual verification but, we need something more scalable to over count this a big data scale we are talking about.

(Refer Slide Time: 17:12)



The slide features a black header bar at the top. On the left is the NPTEL logo, and on the right is the RISE logo. The title 'Value' is centered in a large, dark font. Below the title is a blue horizontal line. The main content is a list of four bullet points. At the bottom, there is a footer bar with 'NPTEL' on the left, 'Introduction to Data Analytics' in the center, and '19' on the right.

- Monetizing the big data
- Identify problems where the returns are higher than investment
- The other four Vs play a role here.
- Examples: Recommendations; Sentiment Analysis; Business process optimization; Security and surveillance; Healthcare; Smart Cities; etc.

And at the last view, I wanted to mention very briefly this whole idea of value. So, I have all this big data; how do I monetize the big data? So, we have to think about problems where substance are returns or higher than the investment. And now, the investment is no long at trivial because we really have to put in the lots of resources in order to just manage this data at the scale; and then drawn analytics on top of it. So, all the 4 V's play a role here, that is so basically have to figure out what is the right balance interns of the effort, that you put in and the kind of monetization that we can do.

So some example which people actually try out or on recommendation sent able better recommend assistance for different domains. Looking at sentiment analysis trying to get a sense of what people feel about new products; may be movies or what do people feel about the particular candidates in election. So, we are looking at that and then other examples include business process optimization, surveillance, healthcare, smart cities; that many many domains in which people have trying to find out monetization for big data.

So, in the next module we will look at some of the challenges in running data analytics at the scale that we have talked about today. That brings us to end of this module.