Report of

Expert Lecture

on

" Big Data and Crowdsourcing"

by

Professor Anirban Dasgupta

organized by

GTU PG SCHOOL

at

GTU PG SCHOOL, BISAG, Gandhinagar

25TH March-2015

Gujarat Technological University running its four Master of Engineering courses at GTU PG SCHOOL BISAG, Gandhinagar.

1. M.E. in Computer Engineering (IT Systems And Network Security)
2. M.E. in Computer Engineering (Wireless & Mobile Computing)
3. M.E. in Computer Engineering (High Performance Computing)
4. M.E. in Electronics & Communication (VLSI & Embedded Systems Design)

With the guidance of our honorable Vice Chancellor Dr. Akshai Aggrawal sir, we are frequently arranging expert lectures for our institute students for their course.

We had arrange expert lectured of Professor Anirban Dasgupta (Ph.D. From Cornell University-2005) on 25th March-2015, who is a Faculty Member at IIT Gandhinagar. His area of interest is Algorithms for large scale data, Social Networks, Computational Social Science, Crowdsourcing, and Machine Learning.

Report of the Technical Session

5Vs of Big Data

Big Data is a big thing. It will change our world completely and is not a passing fad that will go away. To understand the phenomenon that is big data, it is often described using five Vs: Volume, Velocity, Variety, Veracity and Value

I thought it might be worth just reiterating what these five Vs are, in plain and simple language:

1. **Volume:**

   It refers to the vast amounts of data generated every second. Just think of all the emails, twitter messages, photos, video clips, sensor data etc. we produce and share every second. We are not talking Terabytes but Zettabytes or Brontobytes. On Facebook alone we send 10 billion messages per day, click the "like' button 4.5 billion times and upload 350 million new pictures each and every day. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute! This increasingly makes data sets too large to store and analyses using traditional database technology. With big data technology we can now store and use these data sets with the help of distributed systems, where parts of the data is stored in different locations and brought together by software.

2. **Velocity:**

   It refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds, the speed at which credit card transactions are checked for fraudulent activities, or the milliseconds it takes trading systems to analyses social media networks to pick up signals that trigger decisions to buy or sell shares. Big data technology allows us now to analyses the data while it is being generated, without ever putting it into databases.

3. **Variety:**

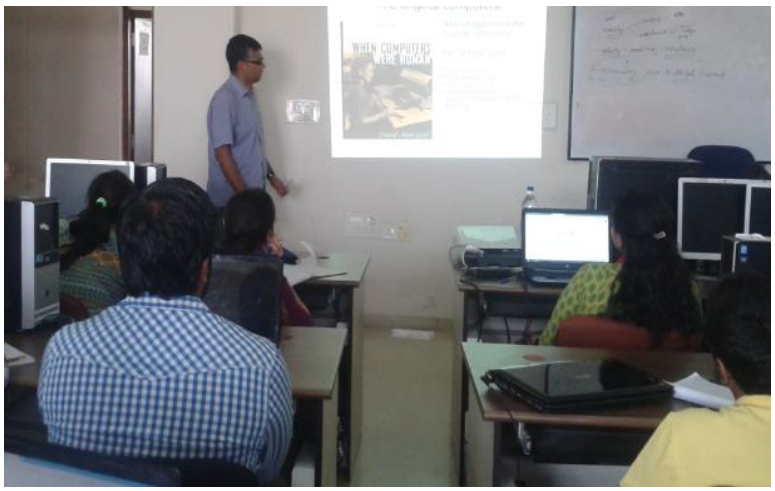
   It refers to the different types of data we can now use. In the past we focused on structured data that neatly fits into tables or relational databases, such as financial data (e.g. sales by product or region). In fact, 80% of the world's data is now unstructured, and therefore can't easily be put into tables (think of photos, video sequences or social media updates). With big data technology we can now harness differed types of data (structured and unstructured) including messages, social media conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data.

4. **Veracity:**

   It refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but big data and analytics Technology now allows us to work with these type of data. The volumes often make up for the lack of quality or accuracy.

5. **Value**:

   Then there is another V to take into account when looking at Big Data: Value! It is all well and good having access to big data but unless we can turn it into value it is useless. So you can safely argue that 'value' is the most important V of Big Data. It is important that businesses make a business case for any attempt to collect and leverage big data. It is so easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of costs and benefits.



Crowdsourcing

Crowdsourcing is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers. While this definition from Merriam Webster is valid, a more specific definition is heavily debated. The process of crowdsourcing is often used to subdivide tedious work and has occurred successfully offline−see the examples below. It combines the efforts of numerous self-identified volunteers or part-time workers, where each contributor of their own initiative adds a small portion to the greater result. The term "crowdsourcing" is a portmanteau of "crowd" and "outsourcing"; it is distinguished from outsourcing in that the work comes from an undefined public rather than being commissioned from a specific, named group.
A few examples based on crowdsourcing are as follows:

ESP Game

The ESP Game is a human-based computation game developed to address the problem of creating Difficult metadata. The idea behind the game is to use the computational power of humans to perform a task that computers cannot (originally, image recognition) by packaging the task as a game.

DARPA Challenge

The 2009 DARPA Network Challenge was a prize competition for exploring the roles the Internet and Social networking play in the real-time communications, wide-area collaborations, and practical actions required to solve broad-scope, time-critical problems. The competition was sponsored by the Defense Advanced Research Projects Agency (DARPA), a research organization of the United States Department of Defense. The challenge was designed to help the military generate ideas for working under a range of circumstances, such as natural disasters. The US Congress authorized DARPA to award cash prizes to further DARPA's mission to sponsor revolutionary, high-payoff research that bridges the gap between fundamental discoveries and their use for national security. In the competition, teams had to locate ten red balloons placed around the United States and then report their findings to DARPA. Due to the distributed nature of the contest, many teams used online resources, such as social media sites, to gather information or to recruit people that would look for balloons. Teams often had to deal with false submissions, and so they needed to come up with ways to validate and confirm reported sightings. The contest was concluded in under nine hours, much less than expected by DARPA, and had many implications with regards to the power of online social networking and crowdsourcing in general.

Amazon Mechanical Turk

The Amazon Mechanical Turk (MTurk) is a crowdsourcing Internet marketplace that enables individuals and businesses (known as Requesters) to coordinate the use of human intelligence to perform tasks that computers are currently unable to do. It is one of the sites of Amazon Web Services. The Requesters are able to post tasks known as HITs (Human Intelligence Tasks), such as choosing the best among several photographs of a storefront, writing product descriptions, or identifying performers on music CDs. Workers (called Providers in Mechanical Turk's Terms of Service, or, more colloquially, Turkers) can then browse among existing tasks and complete them for a monetary payment set by the Requester. To place HITs, the requesting programs use an open application programming interface (API), or the more limited MTurk Requester site. Requesters are restricted to US-based entities.

Top coder

Top Coder is a company which administers contests in computer programming. Top Coder hosts Fortnightly online competitive programming competitions—known as SRMs or "single round Matches"—as well as weekly competitions in design and development. The work in design and Development produces useful software which is licensed for profit by Top Coder. Competitors involved in the creation of these components are paid royalties based on these sales. The software resulting from algorithm competitions—and the less-frequent marathon matches—is not usually directly useful, but sponsor companies sometimes provide money to pay the victors. Statistics (including an overall "rating" for each developer) are tracked over time for competitors in each category.

Zomato

Zomato is an online restaurant search and discovery service providing information on home delivery, Dining-out, cafés and nightlife in cities of India and 21 other countries. The site has an Alexa rank of 1,369 in the world and 117 in India as of February 2015.