A Perspective Approach on the Mathematics behind the Big Data

[1]*Dr. C.T.Suryanarayanachari, Vice-Principal and Lecturer in Mathematics*
[2]*Karimulla Sha Shaik, PG Faculty in Computer Science and Applications*
[1,2]*Silver Jubilee Govt. College (A), Kurnool, Andhra Pradesh*

ABSTRACT:

We live in a digital world, which generates a lot of data. New technologies produce an enormous amount of data. Mathematics plays an important role in the existing algorithms for data processing through techniques of statistical learning, signal analysis, distributed optimization, compress sensing etc. The amounts of data that are available and that are going to be available in the near future call for significant efforts in mathematics. These efforts are needed to make the data useful. According to the sixth edition of DOMO's research report: "Over 2.5 quintillion bytes of data are created every single day, and it's only going to grow from there. By 2020, it's estimated that 1.7MB of data will be created every second for every person on earth." Big Data refers to a data set that is so large and/or complex that it cannot be perceived, acquired, managed, and processed by traditional Information Technology (IT) and software/hardware tools within a tolerable time. The extraction of meaningful information from data is one of the main tasks of Statistics. In presence of big data the most part of the usual techniques for statistical analysis cannot easily be applied, since they are based on the simultaneous processing of the whole dataset. Sometimes data are "big" because of their high dimensionality and space-time structure (e.g. to satellite images, signals registered by sensors or antennas, etc.). In such cases suitable mathematical techniques for dimensionality reduction are needed both for data visualization and for their numerical treatment. Functional Statistics, that is a field in which a lot of research is concentrating nowadays, may help in facing this task. This paper describes a perspective approach on the Mathematics behind the Big Data.

Keywords: ***Mathematics, Distributed optimization, Big Data, Information Technology, Data visualization***

INTRODUCTION:

The availability of huge amounts of data is often considered as the fourth industrial revolution we are living right now. The increase in data accumulation allows us to tackle a wide range of social, economic, industrial and scientific challenges. But extracting meaningful knowledge from the available data is not a trivial task and represents a severe challenge for data analysts. Minimization of a cost function, based on large amount of data is

a typical problem in all big data areas from smart agriculture, energy efficiency, computational biology, high tech industries based on simulations in material design to social networks, challenge in policy decisions based on data, risk assessment in finance, security, natural disasters etc. The challenges in these areas, mathematically speaking are the design of algorithms that will be able to process huge amounts of data within a reasonable time span and with computer power that is widely available today.

In other contexts data are considered "big" because of their complexity or heterogeneity (e.g. data extracted from social networks with text mining, mixed to socioeconomic data for marketing purposes; or data highly interrelated which may be represented by complex graphs, like atoms and bounds in a protein, relationships between users of a social network, etc.). Sentiment analysis and Topological Data Analysis are new statistical fields of research, still under development, which may help to tackle the problem of analyzing such data.

A big effort has been made during these years, mainly by computer scientists, to find fast and scalable procedures that have become popular in presence of distributed architectures (e.g. the well-known MapReduce paradigm). Unfortunately, in many situations, such procedures cannot be applied to solve statistical problems in a distributed way, or they work under too many restrictive and thus unrealistic conditions. The deepening of the mathematical insight in this context may help to better understand the theoretical and applied power of the new algorithms and to extend them to more realistic cases.

HOW BIG IS BIG DATA?

- The smallest unit of measurement for data is a bit (b).
- A bit can be either 0 or 1 and is therefore not large enough to hold any data.
- A byte (B), which is 8 bits, is used as the fundamental unit of measurement for data.
- A byte can hold 2 8 = 256 different values, which is enough to represent the standard ASCII characters, such as letters, numbers and some basic symbols.
- Following the tradition of the metric system, terms to measure large quantities of data can be formed using SI prefixes as shown in the table below.
- These prefixes are often used for the multiple of bytes. For example, a kilobyte is 10^3 bytes since kilo means 10^3.

Prefix	Unit Name	Symbol	SI Meaning
Kilo	Kilobyte	KB	10^3
Mega	Megabyte	MB	10^6
Giga	Gigabyte	GB	10^9
Tera	Terabyte	TB	10^{12}
Peta	Petabyte	PB	10^{15}
Exa	Exabyte	EB	10^{18}
Zetta	Zettabyte	ZB	10^{21}
Yotta	Yottabyte	YB	10^{24}

DATA SCIENCE:

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is

a continuation of some of the data analysis fields such as statistics, data mining, machine learning and predictive analytics. [Wikipedia]

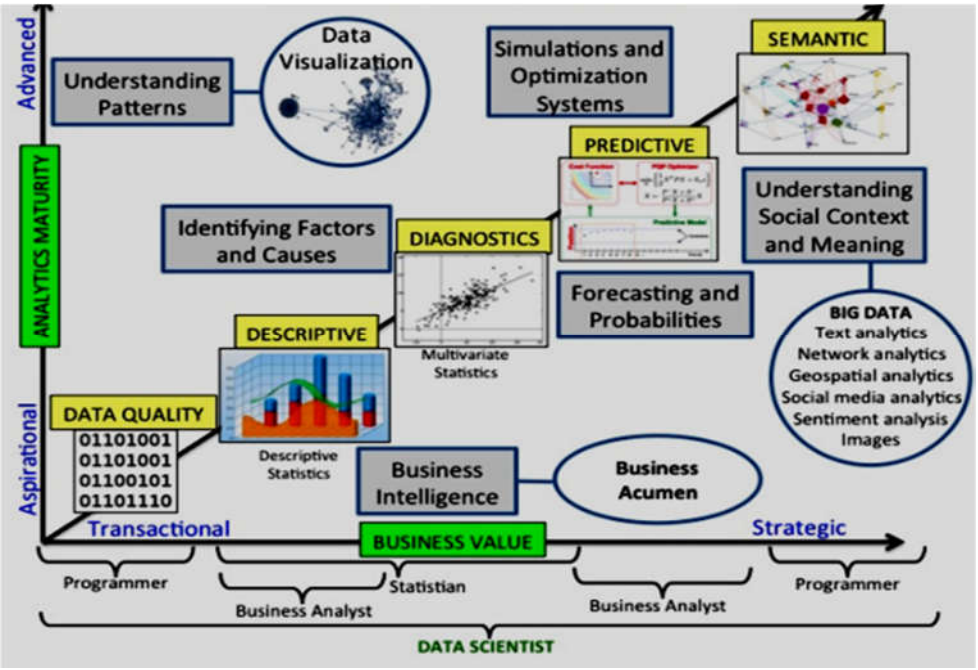

Fig.1. What is Data Science? [Source: www.datasciencecentral.com]

Data Scientist is the Sexiest Job of the 21st Century. Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions.

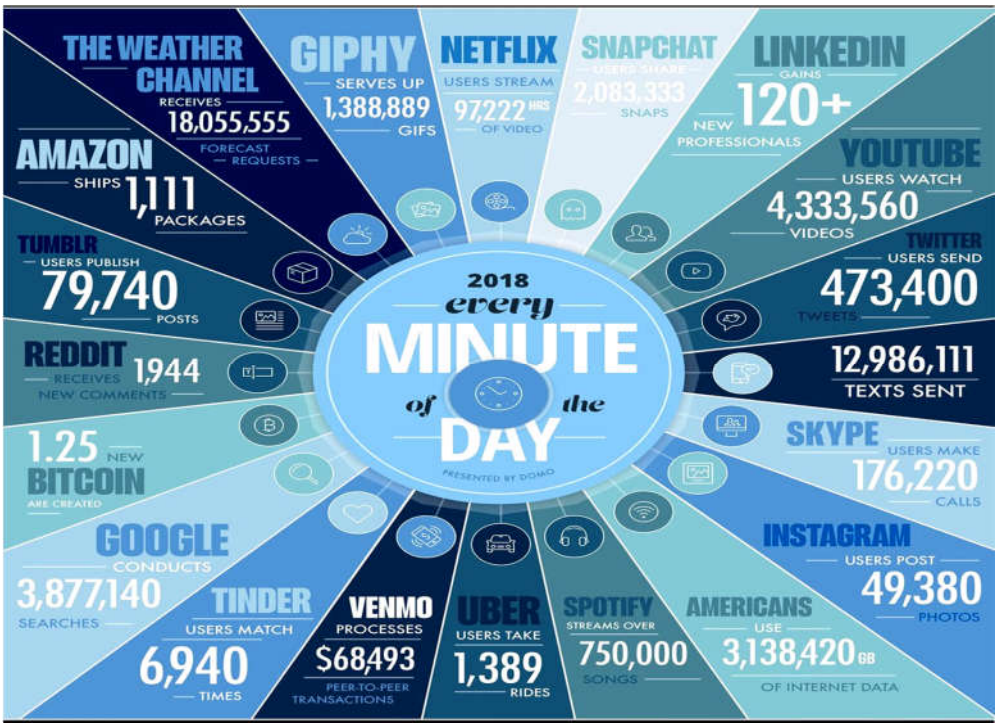

Fig. 2. How much data is generated every minute in 2018? [Source: www. domo.com]

The world's internet population is growing significantly year-over-year. In 2017, Internet usage reached 47% of the world's population and now represents 3.8 billion people.

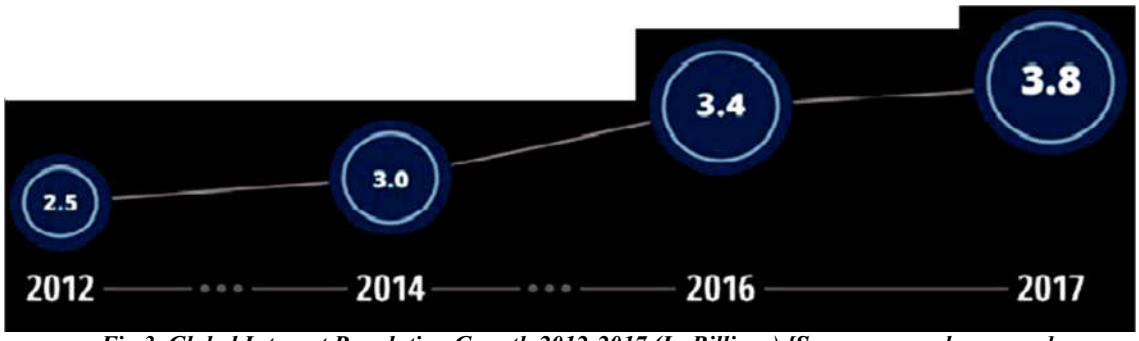

Fig.3. Global Internet Population Growth 2012-2017 (In Billions) [Source: www.domo.com]

HOW MUCH DATA IS CREATED ON THE INTERNET EACH DAY?

The Amount of Data Created Each Day on the Internet in 2017:

- In 2014, there were 2.4 billion internet users. That number grew to 3.4 billion by 2016, and in 2017, 300 million internet users were added — making a total of 3.8 billion internet users in 2017. This is a 42% increase in people using the internet in just three years!

Each Minute of Every Day, the following happens on the Internet:

- Social media is huge. Reports show that social media gains 840 new users each minute.
- Since 2013, the number of tweets each minute has increased 58% to more than 455,000 tweets per minute in 2017!
- YouTube usage more than tripled from 2014-2016 with users uploading 400 hours of new video each minute of every day! Now, in 2017, users are watching 4,146,600 videos every minute.
- Instagram users upload 46,740 million posts every minute!
- Since 2013, the number of Facebook posts shared each minute has increased 22%, from 2.5 million to 3 million posts per minute in 2016. This number has increased by more than 300 percent, from around 650,000 posts per minute in 2011!
- Every minute on Facebook, 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded.
- Facebook users also click the like button on more than four million posts every minute!
- 3,607,080 Google searches are conducted worldwide each minute of every day.
- Worldwide, 15,220,700 texts are sent every minute!
- Instagram users post 46,740 pictures every minute.

If we do some quick calculations, we can see the amount of data created on the internet each day. There are 1,440 minutes per day... so that means that there are approximate:

- 1,209,600 new data-producing social media users each day.

- 656 million tweets per day!
- More than four million hours of content uploaded to YouTube every day, with users watching 5.97 billion hours of YouTube videos each day.
- 67,305,600 Instagram posts uploaded each day,
- Facebook has 1.32 billion daily active users on average as of June 2017.
- 4.3 billion Facebook messages posted daily!
- 5.75 billion Facebook likes every day.
- 22 billion texts sent every day.
- 5.2 billion Daily Google Searches in 2017.

e-mail use continues to rise:

The Email Statistics Report 2017-2021 by the Radicati Group confirms this. 269 billion emails are sent daily in 2017, and this is expected to grow by 4.4% yearly to 319.6 billion in 2021.

Mobile Device Data:

- The amount of mobile data is also blowing up.
- At the start of 2014, mobile phones/tablets uploaded and downloaded around 2 exabytes (1 exabyte = 1 billion gigabytes) of data.
- At the start of 2017, data created on mobile devices quadrupled to over 8 exabytes.
- At the start of 2017, there were 3.394 billion mobile internet users. This means that in 2017, there are more mobile internet users than desktop internet users, with mobile being used to access 51.4% of web pages and desktop to access 43.4% (tablet is used for 4.9% and other devices for the remaining).
- Approximately 21.9 billion text messages are sent each day in 2017, compared to 18.7 billion in 2016 – a 17% increase in just one year.

Growth in Data generating Services:

The growth in data and the way it can be used is also changing the way business is being done and the services that organizations can offer due to an enhanced ability to produce, capture, and understand data.

- Amazon is dominating the marketplace – Amazon processes $373 MILLION in sales every day in 2017, compared to about 120 million Amazon sales in 2014
- By the end of 2016, Uber had 40 million monthly active users
- Venmo processes $74.7 million in transactions EVERY DAY

THE 42 V'S OF BIG DATA AND DATA SCIENCE:

1. *Vagueness:* The meaning of found data is often very unclear, regardless of how much data is available.
2. *Validity:* Rigor in the analysis (e.g., Target Shuffling) is essential for valid predictions.
3. *Valor:* In the face of big data, we must gamely tackle the big problems.
4. *Value:* Data science continues to provide ever-increasing value for users as more data becomes available and new techniques are developed.
5. *Vane:* Data science can point in the direction of correct decision making.
6. *Vanilla:* Even the simplest models, constructed with rigor, can provide value.
7. *Vantage:* Big data allows us a privileged view of complex systems.

8. *Variability:* Data science often models variable data sources. Models deployed into production can encounter especially wild data.
9. *Variety:* In data science, we work with many data formats (flat files, relational databases, graph networks) and varying levels of data completeness.
10. *Varifocal:* Big data and data science together allow us to see both the forest and the trees.
11. *Varmint:* As big data gets bigger, so can software bugs!
12. *Varnish:* How end-users interact with our work matters, and polish counts.
13. *Vastness:* With the advent of the internet of things, the "bigness" of big data is accelerating.
14. *Vaticination:* Predictive analytics provides the ability to forecast. (Of course, these forecasts can be more or less accurate depending on rigor and the complexity of the problem. The future is pesky and never conforms to our March Madness brackets.)
15. *Vault:* With many data science applications based on large and often sensitive data sets, data security is increasingly important.
16. *Veer:* With the rise of agile data science, we should be able to navigate the customer's needs and change directions quickly when called upon.
17. *Veil:* Data science provides the capability to peer behind the curtain and examine the effects of latent variables in the data.
18. *Velocity:* Not only is the volume of data ever increasing, but the rate of data generation (from the internet of things, social media, etc.) is increasing as well.
19. *Venue:* Data science work takes place in different locations and under different arrangements: Locally, on customer workstations, and in the cloud.
20. *Veracity:* Reproducibility is essential for accurate analysis.
21. *Verdict:* As an increasing number of people are affected by models' decisions, Veracity and Validity become ever more important.
22. *Versed:* Data scientists often need to know a little about a great many things: mathematics, statistics, programming, databases, etc.
23. *Version Control:* You're using it, right?
24. *Vet:* Data science allows us to vet our assumptions, augmenting intuition with evidence.
25. *Vexed:* Some of the excitement around data science is based on its potential to shed light on large, complicated problems.
26. *Viability:* It is difficult to build robust models, and it's harder still to build systems that will be viable in production.
27. *Vibrant:* A thriving data science community is vital, and it provides insights, ideas, and support in all of our endeavors.
28. *Victual:* Big data — the food that fuels data science.
29. *Viral:* How does data spread among other users and applications?
30. *Virtuosity:* If data scientists need to know a little about many things, we should also grow to know a lot about one thing.
31. *Viscosity:* Related to Velocity; how difficult is the data to work with?
32. *Visibility:* Data science provides visibility into complex big data problems.
33. *Visualization:* Often the only way customers interact with models.
34. *Vivify:* Data science has the potential to animate all manner of decision making and business processes, from advertising to fraud detection.
35. *Vocabulary:* Data science provides a vocabulary for addressing a variety of problems. Different modeling approaches tackle different problem domains, and different validation techniques harden these approaches in different applications.

36. *Vogue:* "Machine Learning" which becomes "Artificial Intelligence", which becomes...?
37. *Voice:* Data science provides the ability to speak with knowledge (though not all knowledge, of course) on a diverse range of topics.
38. *Volatility:* Especially in production systems, one has to prepare for data volatility. Data that should "never" be missing suddenly disappears, numbers suddenly contain characters!
39. *Volume:* More people use data-collecting devices as more devices become Internet-enabled. The volume of data is increasing at a staggering rate.
40. *Voodoo:* Data science and big data aren't voodoo, but how can we convince potential customers of data science's value to deliver results with real-world impact?
41. *Voyage:* May we always keep learning as we tackle the problems that data science provides.
42. *Vulpine:* Nate Silver would like you to be a fox, please.

HOW IS MATH USED IN TECHNOLOGY?

Math is used in several different ways in technology. For instance, the Internet is based on a form of math called binary code (this is what all computers work on).

- *Calculator:* It a calculator is used the most common tool used in math. It solves both complex and simple math. It uses whole, negative integers, positive integers, rational numbers, irrational numbers, and natural numbers.
- *Cell Phone:* It uses the binary system which are multiples of 2's, 0's and 1's. It has positive integers on the dial pad. It uses coordinates to locate the Satellite to receive & transmit other ends.
- *Speedometer:* The Speedometer is used to tell what speed a vehicle is going. It uses math because it takes the derivative of the odometer which measures the distance travelled by a vehicle.
- *Microwave:* Mathematics is crucial to the design of the microwave. The microwave generates electromagnetic radiation at microwave frequencies. Waves are created when electrons interact with the magnetic field created by the magnets. The specific wavelength of microwaves created within the oven is 12.2cm. To produce microwaves of 12.2cm, the strength of the magnet needs to be precise. It is calculated by starting with the Lorentz Force Law. $F=q(E+vxB)$

REQUIRED BACKGROUND IN MATHEMATICS:

What background knowledge would be helpful to know as a Data Scientist?

Most people give more emphasis on statistics (or computer science) over linear algebra and calculus since many of the machine learning algorithms are already implemented. While it is true that it is recommended to use already implemented algorithms, it is important to have a fundamental understanding of how the algorithms are implemented. This way you understand the assumptions in using the algorithm both quickly and correctly.

The required knowledge is broken down into the following sections:

- Pre-calculus and Mathematical Thinking
- Statistics: Stats is arguably the most important background knowledge required since machine learning is applied statistics. i.e. Descriptive Statistics for mean, median, range, standard deviation, variance, and exploratory data analysis.

- Linear Algebra: A lot of machine learning (ML) or applied statistics concepts are tied to linear algebra concepts. Some basic examples, PCA - eigenvalue, regression - matrix multiplication… As most ML techniques deal with high dimensional data, they are often times represented as matrices. Concepts such as singular value decomposition, projections, principal components, eigenvalue, eigenvectors, regression, matrix multiplication, matrix operations, matrix, inverse, solving differential equations using matrices are all important.
- Calculus: Calculus and differential equations is very important for mathematical modeling of systems and used in optimization.

CONCLUSION:

Big Data is everywhere as high volumes of varieties of valuable precise and uncertain data can be easily collected or generated at high velocity in various real-life applications. The explosive growth in web-based storage, management, processing, and accessibility of social, medical, scientific and engineering data has been driven by our need for fundamental understanding of the processes which produce this data. It is predicted that the volume of the produced data could reach 44 zettabytes in 2020. Mathematical methods are an important enabler in big data settings. Developments in big data settings not only require more computer speed and memory capacity, but also newly advanced mathematics.

REFERENCES:

1. http://www.science.unitn.it/~alonso/Orientamento/datascience.pdf
2. European Consortium for Mathematics in Industry - Mathematics for Big Data
3. https://dzone.com/articles/how-much-data-is-created-on-the-internet-each-day
4. https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/
5. D. Butler : A world where everyone has a robot: why 2040 could blow your mind Nature, 530 (7591) (2016)
6. https://www.elderresearch.com/blog/42-v-of-big-data
7. http://ksuweb.kennesaw.edu/~plaval/math4490/fall2017/mathsurvey_def.pdf
8. M. Chen, S. Mao, Y. Zhang, and V. C. Leung, Big data: related technologies, challenges and future prospects, Springer, 2014.