

$\pi$

# Lecture 0: Introduction

**CSE 4261: DATA ANALYTICS** Theory 3 credits, Lab 0.75 credits

**Prof. Dr. Shamim Akhter**, Professor, CSE, AUST

<https://www.linkedin.com/in/prof-dr-md-shamim-akhter-a3a96666/>

[shamimakhter.cse@aust.edu](mailto:shamimakhter.cse@aust.edu)

**Room Number: 7A01/E**

**FALL 2023**

**CLASS CODE: af3wt3u**

<https://classroom.google.com/c/NjYwMjM4NjE2NDI0?cjc=af3wt3u>

	08-09.15	09.20-10.35	10.40-11.55	12.00-01.15	01.20-02.35	02.40-03.55	04.00-05.15	05.20-06.40
Sunday								
Monday			CSE 4261 Shamim 7A04	CSE 4261 Shamim 7A04				
Tuesday			CSE 4261 Shamim 7A04	CSE 4261 Shamim 7A04				
Wednesday					CSE 4262 (Shamim, Zahid) 7B03			
Thursday								



**AISIP-Applied Intelligent  
System and Information  
Processing.**



## SHORT BIOGRAPHY



**Dr. Md. Shamim Akhter**  
Professor,  
Computer Science and  
Engineering

**IEEE Senior Member,**  
Member, IEEE Computer Society Technical  
Committee-Parallel Processing  
Member, IEEE Computer Society Technical  
Committee-Intelligent Informatics  
Former JSPS Post Doctoral Fellow, Japan  
Former Scholar, Yoshida Foundation, Japan  
Ph: 01795716777/02-55091805 Ext: 7081  
<https://www.linkedin.com/in/dr-md-shamim-akhter-a3a96666>

**Prof. Dr. Md. Shamim Akhter** received a B.S. in computer science from the American International University of Bangladesh, in 2001, an M.S in computer science and information management from the Asian Institute of Technology, Thailand, in 2005, and a Ph.D. degree in information processing from Tokyo Institute of Technology, Japan, in 2009.

Since 2001, He has been working as a faculty in Computer science, computer science and engineering, Information Processing, and the related fields of studies.

He started his teaching at AIUB as a lecturer, then in 2005, he became an assistant professor. In addition, he served as the head of the graduate program at AIUB. He worked at EWU from 2014 to 2019 as an assistant and associate professor. From 2019 May, he worked as a full-time professor. He also worked as a contract faculty in TRU, CANADA for 8 months.

As a researcher, he has more than 5 years of experience. He worked as a JSPS postdoctoral fellow in NII, Japan, as a research associate in AIT Thailand, and GCOE project in TITECH Japan.

He is the author of a book and more than 60 articles. His research interests include Machine Intelligence, Intelligent Algorithms, Evolutionary Algorithms, and Parallelization Models. He has been a member of IEEE since 2008 and a senior member since 2012.

During his 24-year career, he worked on the program committee, as a reviewer, as a session chair, and as an organizer for over twenty IEEE conferences. He participated in several R10 regional award distribution events, seminars, and workshops as a volunteer.

# Research Group

## Applied Intelligent System and Information Processing (AISIP)

works directly under the supervision of  
**Prof. Dr. Md. Shamim Akhter**

Topics: Artificial Intelligence, Data Science and Machine learning technologies will apply to [Industry 4.0 Applications](#)

$\pi$



**AISIP-Applied Intelligent System and Information Processing.**



## Course Outcomes

**After the successful completion of this course, students are expected to be able to:**

Sl. No.	COs	POs	Bloom's Taxonomy		
			C	A	P
1	Incorporate the basic principles of data analytics to transform the unstructured into structured data	1	2		
2	Apply cutting-edge technologies and advanced analytics to solve complex data science problems	4	3		
3	Analyze the categorical, numerical, or graphical data representations for effective inferences	2	4		

Week	Topics	Teaching-Learning Strategy	Assessment Strategy	Corresponding COs
1.	Introduction to Data Analytics and its Life Cycle: an overview of the current challenges of data and applications of data analytics life cycle of data analytics including discovery, data preparation, model planning, model building, communication results, and operationalization.	<ul style="list-style-type: none"> <li>- Lecture</li> <li>- Brainstorming Session</li> <li>-Think-Pair-Share (TPS)</li> </ul>	<ul style="list-style-type: none"> <li>- Final Exam</li> <li>- Quiz 1</li> <li>-Class Performance</li> </ul>	1
2.	Review of Basic Data Analytic Methods: data pre-processing and transformation: small and large data sets; missing data and dealing with it; noisy data and sampling; techniques for data cleaning, reshaping, merging, and transforming. Data correlation, and similarity measures.	<ul style="list-style-type: none"> <li>- Lecture</li> <li>- Brainstorming Session</li> <li>- Hand Notes</li> <li>-Think-Pair-Share (TPS)</li> </ul>	<ul style="list-style-type: none"> <li>- Final Exam</li> <li>- Quiz 1</li> <li>-Class Performance</li> </ul>	1
3.	Sampling and Statistical Interference: Sample distribution of mean, proportion, t-distribution, chi-square, and f-distribution. Statistical Interferences including point estimation, unbiasedness, consistency, sufficiency, and efficiency; interval estimation; determine of sample size. Testing of hypothesis and Chi-Square tests.	<ul style="list-style-type: none"> <li>- Lecture</li> <li>- Brainstorming Session</li> <li>- Video Demonstration</li> </ul>	<ul style="list-style-type: none"> <li>- Final Exam</li> <li>- Quiz 1</li> <li>-Class Performance</li> </ul>	1

4.	Advanced Analytical Theory and Methods: Regression: Linear Regression Analysis-least squares estimation, coefficient of determination, standard error, use case for slope analysis. Logistic Regression Analysis- Model description, Deviance and the Pseudo-R <sup>2</sup> , Deviance and the Log-Likelihood Ratio Test, Receiver Operating Characteristic (ROC) Curve, Additional Regression Models-Ridge and Lasso.	- Lecture - Brainstorming Session - Video Demonstration - Think-Pair-Share (TPS)	- Final Exam - Quiz 2 -Class Performance	2,3		
5.	Factor Analysis: the centroid method, the principal components method, and the maximum likelihood method. Discriminant Analysis: Fisher's Method and <u>Mahalanobis' Method</u> .	- Lecture - Brainstorming Session - Hand Notes	- Final Exam -Class Performance Quiz 2	2,3		
6.	Advanced Analytical Theory and Methods: Association Rules: <u>Apriori</u> algorithm, Support, Confidence, Lift, Leverage, Use case and application.	- Lecture - Brainstorming Session - Case Study	- Final Exam -Class Performance Quiz 2	2,3		
Mic		Classifications. Decision Tree Analysis-classification trees and regression trees, Formulation of trees using entropy and information gain, detecting significant splits. Naive Bayes analysis with Bayes' Theorem.	- Brainstorming Session - Case Study -Video Demonstration	- Quiz 3 -Class Performance		
8.	Advanced Analytical Theory and Methods: Clustering: Overview of clustering, K-means clustering methods, determining the number of clusters, and evaluating clusters with inter-cluster and intra-cluster heuristics.					
9.	Advanced Analytical Theory and Methods:	10. Advanced Analytical Theory and Methods: Time Series Analysis: Overview of Time Series Analysis, Box-Jenkins Methodology; ARMA vs ARIMA model in terms of Autocorrelation Function (ACF), Autoregressive Models, Moving Average Models. Comparing Fitted Time Series Models- AIC (Akaike Information Criterion), <u>AICc</u> (Akaike Information Criterion, corrected) and BIC (Bayesian Information Criterion).	- Lecture - Brainstorming Session -Think-Pair-Share (TPS) -Case Study	- Final Exam - Quiz3 -Class Performance -Assignment1	2,3	

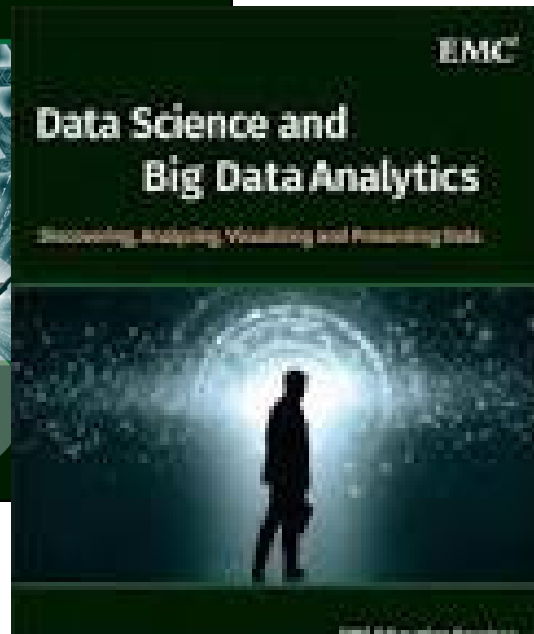
11.	Data Visualization: Visual data exploration and analytics: data visualization techniques and their applicability in data analytics, visual data.	<ul style="list-style-type: none"><li>- Lecture</li><li>- Brainstorming Session</li><li>- Hand Notes</li></ul>	<ul style="list-style-type: none"><li>- Final Exam</li><li>-Class Performance Quiz3</li></ul>	3
12.	Introduction to Knowledge Discovery and Data Mining Techniques: basic data mining and related concepts including information retrieval, data warehousing, and dimensional modeling; Data Cube and OLAP Analytical processing: an overview of Data Cube architecture, dimension hierarchies, OLAP vs OLTP, ROLAP and MOLAP.	<ul style="list-style-type: none"><li>- Lecture</li><li>- Brainstorming Session</li><li>- Think-Pair-Share (TPS)</li><li>-Question-Answer Session</li><li>- Video Demonstration</li></ul>	<ul style="list-style-type: none"><li>- Final Exam</li><li>-Class Performance</li></ul>	3
13.	Big Data Analytics-describes basics of Hadoop architecture and its components, Hadoop distributed file system (HDFS), NoSQL data stores, MapReduce paradigm, Hive, HiveQL and Pig architecture. It introduces Spark architecture features, software stack components and their functions.	<ul style="list-style-type: none"><li>- Lecture</li><li>- Brainstorming Session</li><li>-Think-Pair-Share (TPS)</li><li>-Case Study</li></ul>	<ul style="list-style-type: none"><li>Final Exam</li><li>-Class Performance</li></ul>	2,3
14.	Review Classes	<ul style="list-style-type: none"><li>- Lecture</li><li>- Brainstorming Session</li><li>- Think-Pair-Share (TPS)</li><li>-Question-Answer Session</li></ul>	<ul style="list-style-type: none"><li>- Final Exam</li><li>-Class Performance</li></ul>	1, 2, 3



## Text Book: Data Science and Big Data Analytics



David Dietrich and et. al., 2015, Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, John Wiley & Sons, Inc. Publisher.





# Overview of Data Management

## Technology and Application

**Prof. Dr. Shamim Akhter**

Professor, Dept. of CSE  
Ahsanullah University of Science and Technology

$\pi$

## Technology trends and underlying technologies

### Industry-agnostic trends



#### 1 Next-level process automation...

Industrial IoT<sup>1</sup>  
Robots/cobots<sup>2</sup>/RPA<sup>3</sup>



#### ... and process virtualization

Digital twins  
3-D/4-D printing



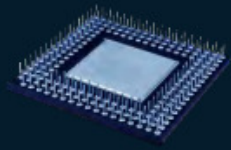
#### 2 Future of connectivity

5G and IoT connectivity



#### 3 Distributed infrastructure

Cloud and edge computing



#### 4 Next-generation computing

Quantum computing  
Neuromorphic chips (ASICs<sup>4</sup>)



Alexa

#### 5 Applied AI

Computer vision, natural-language processing, and speech technology



#### 6 Future of programming

Software 2.0

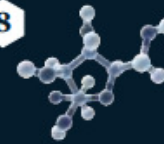


#### 7 Trust architecture

Zero-trust security  
Blockchain

### Industry-specific trends

8



#### Bio Revolution

Biomolecules/"-omics"/  
biosystems

Biomachines/biocomputing/augmentation

9



#### Next-generation materials

Nanomaterials, graphene and 2-D materials, molybdenum disulfide nanoparticles

10



#### Future of clean technologies

Nuclear fusion

Smart distribution/metering

Battery/battery storage

Carbon-neutral energy generation

$\pi$ 

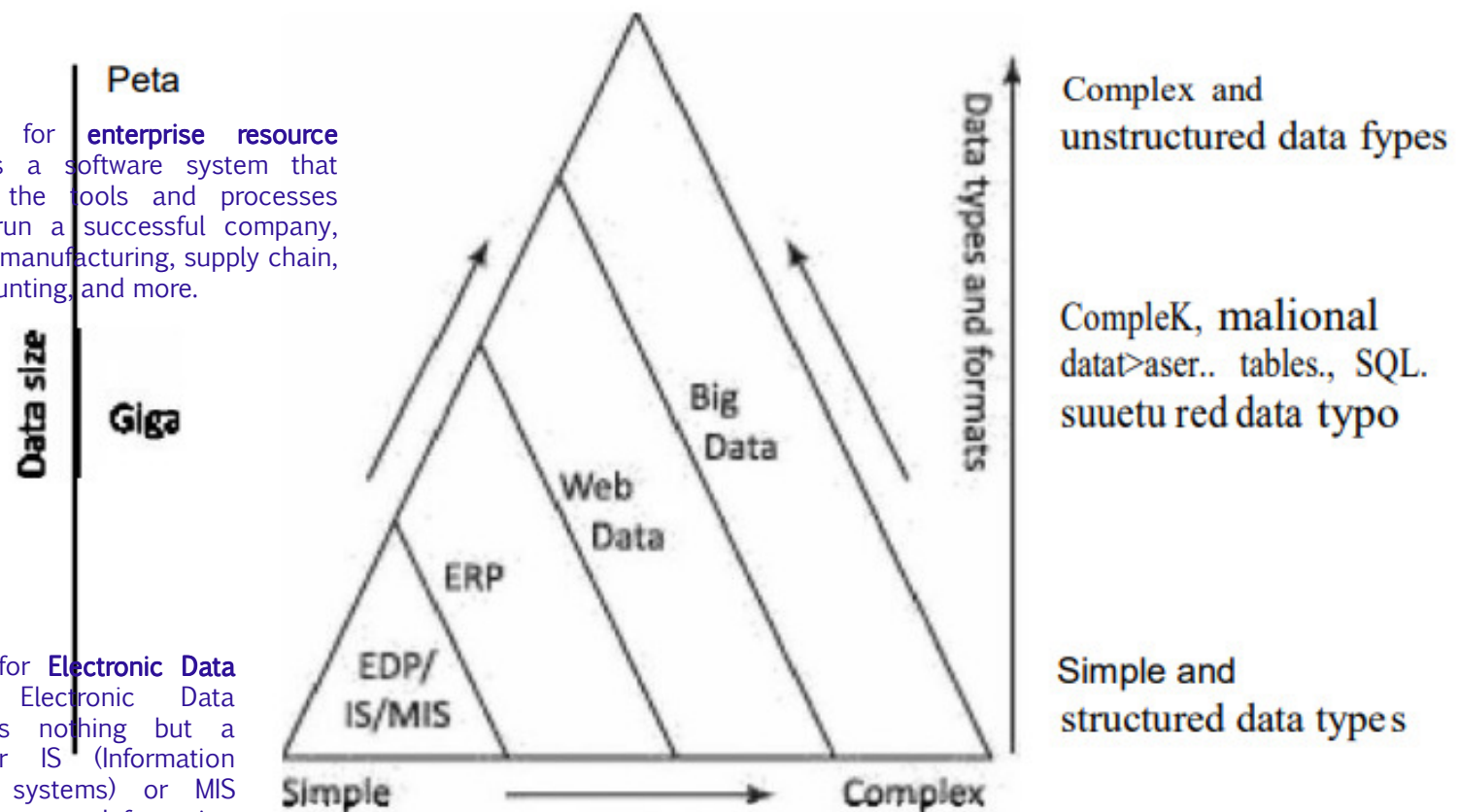
# Communication Trends



# Evolution of Big Data and their characteristics

ERP stands for **enterprise resource planning**. It's a software system that includes all the tools and processes required to run a successful company, including HR, manufacturing, supply chain, finance, accounting, and more.

EDP stands for **Electronic Data Processing**. Electronic Data Processing is nothing but a synonym for IS (Information Services or systems) or MIS (Management Information Services or systems).





## Volume

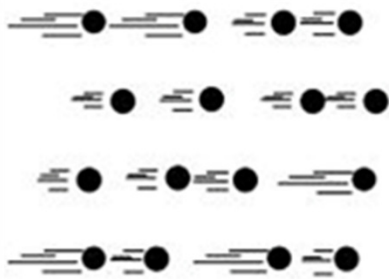


### Data at Rest

1000GB  
Terabytes to  
Exabytes of existing  
data to process  
1000PB  
1000000000 GB



## Velocity

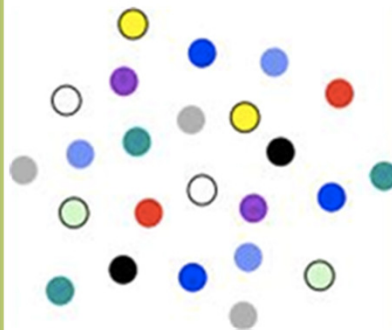


### Data in Motion

Streaming data,  
requiring milliseconds  
to seconds to respond



## Variety

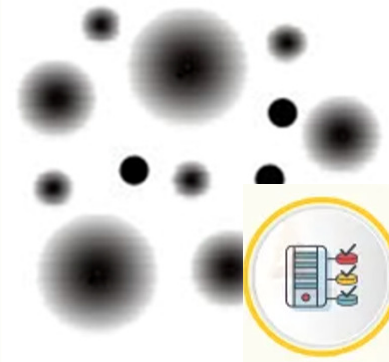


### Data in Many Forms

Structured,  
unstructured, text,  
multimedia,...



## Veracity



### Data in Doubt

Uncertainty due to  
data inconsistency &  
incompleteness,  
ambiguities, latency,  
deception, model  
approximations

## Value



### Data into Money

Business models can  
be associated to the  
data



“The goal is to turn data into information, and information into insight.”

– By Carly Fiorina, ex-CEO of Hewlett-Packard.

Every day, 500+ terabytes of fresh data are absorbed into the Facebook systems. This information is mostly gathered through photo and video uploads, message exchanges, and the posting of comments, among other things.



In 30 minutes of flying time, a single Jet engine may create 10+ gigabytes of data. With thousands of flights every day, the amount of data generated can amount to several Petabytes.



Every day, the Fresh York Stock Exchange creates around a terabyte of new trading data.





# Current Challenges with Data(Big Data)

## › Volume:

- Big Data refers to a massive amount (volume) of information
- The magnitude of data plays a critical role in determining its worth.
- In 2016, worldwide mobile traffic was predicted to be 6.2 Exabytes (6.2 billion GB) per month. Furthermore, by 2020, we had about 40000 ExaBytes of data.
- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain approximately 2.5 petabytes of data.
- In today's technological world data is generated from various sources in different formats.
  - › Data formats are in the form of Word, excel documents, PDFs, and media content such as images, videos, etc. are produced at a great pace.
- It is becoming challenging for enterprises to store and process data
  - › using the conventional methods of business intelligence and analytics.
  - › need to implement modern business intelligence tools to effectively capture, store, and process such huge amounts of data in real life.

# Current Challenges with Data(Big Data)

## › Velocity:

- The term "velocity" refers to the rapid collection of data.
- Data comes in at a high rate from machines, networks, social media, mobile phones, and other sources in Big Data velocity.
- Now, this data **needs to be captured as close to real-time as possible**, so that the right data can be available at the right time.
- For making timely and accurate business decisions the speed at which data can be accessed matters the most.
- **Data sampling** can assist in dealing with issues such as 'velocity.'
- For instance, **Google receives more than 3.5 billion queries every day. In addition, the number of Facebook users is growing at a rate of around 22% every year.**

# Current Challenges with Data(Big Data)

## › Variety:

- The volume and velocity of data add value to an organization or business, but the diverse data types collected from varied data sources are also an important factor of Big data.
- Big data is generally classified as structured, semi-structured, or unstructured data.

Unstructured data

The university has 5600 students.  
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.  
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

- ✓ Around 80% of the data produced globally including the videos, photos, mobile data, social media content, is unstructured in nature.
- ✓ Decoding the human genome originally took 10 years to process, but now with the help of Big data it can be achieved in one week
- ✓ A 10% increase in data accessibility by a Fortune 1000 company would give that company approximately \$65 million more in annual net income.

# Quasi-structured data

- › Textual data with inconsistent data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats).
- › Quasi-structured data is a common phenomenon that bears closer examination. Consider the following example. A user attends the EMC World conference and subsequently runs a Google search online to find information related to EMC and Data Science. This would produce a URL such as <https://www.google.com/#q=EMC+data+science> and a list of results

Structured data include text documents, PDFs, images, and video.

Examples of semi-structured data include XML data files that are self-describing and defined by an xml schema.

"Quasi" Structured data is defined as data that has no inherent structure and is usually stored as different types of files.



# Current Challenges with Data(Big Data)

## › Veracity/Validity:

- Bad data can create weak analysis, bad interpretation, wrong decision, faulty execution, no learning.
- It is important to check the validity of the data before proceeding with further analysis.
- Questions like Can you trust the data that you have collected? Is the data reliable enough? , etc. need to be entertained. More than 5 billion people are calling, texting, browsing and tweeting on mobile phones worldwide.
- Bad data or poor quality data costs organizations as much as 10-20 % of their revenue.
- Poor data across businesses and the government costs the U.S. economy \$3.1 trillion dollars a year.

# Current Challenges with Data(Big Data)

## › Value:

- Today data is being produced in large volumes. And just collecting the produced data is of no use. Instead, we have to look for data from which business insights can be generated which adds “value” to the company. So we can say that the **Value is the most important V of all the 5 V's.**
- **Data analytics** helps to derive useful insights from the collected data. These insights, in turn, **add value to the decision-making process.**
- Now, how to make sure that the value of Big data is considerable and worth investing time and effort into?
  - › It can be done by conducting a **cost vs benefit analysis**. By calculating the **total cost of processing Big data and comparing it with the ROI**, business insights are expected to be generated. Using these companies can effectively decide whether Big data analytics adds any value to their business or not.
- **According to McKinsey, a retailer using Big Data to its fullest potential could increase its operating margin by more than 60%.**

- › Give three(3) examples of the machine-generated data.
- › Examples of machine-generated data are:
  1. Data from computer systems: Logs, weblogs, security/surveillance systems, videos/images, etc.
  2. Data from fixed sensors: Home automation, weather sensors, pollution sensors, traffic sensors, etc.
  3. Mobile sensors (tracking) and location data.



$\pi$

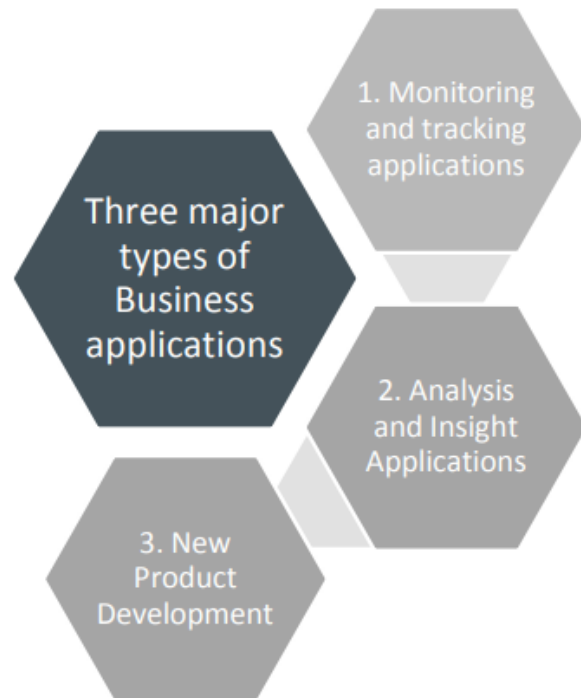
Think of a manufacturing and retail marketing company, such as LEGO Toys. How does such a toy company optimize the services offered, products, and schedules, devise ways, and use Big Data processing and storing for predictions using analytics?

› SOLUTION

- › Assume that a retail and marketing company of toys uses several Big Data sources, such as
  - i. machine-generated data from sensors (RFID readers) at the toy packaging,
  - ii. transaction data of the sales stored as web data for automated reordering by the retail stores and
  - iii. tweets, Facebook posts, e-mails, messages, and web data for messages and reports.
- › The company uses Big Data to understand the toys and themes in the present day that are **popularly demanded by children, predicting the future types and demands.**
  - The company using such predictive analytics, optimizes the product mix and manufacturing processes of toys.
  - The company optimizes the services to retailers by maintaining toy supply schedules. The company sends messages to retailers and children using social media on the arrival of new and popular toys.

# Application of Data Analytics/Big Data

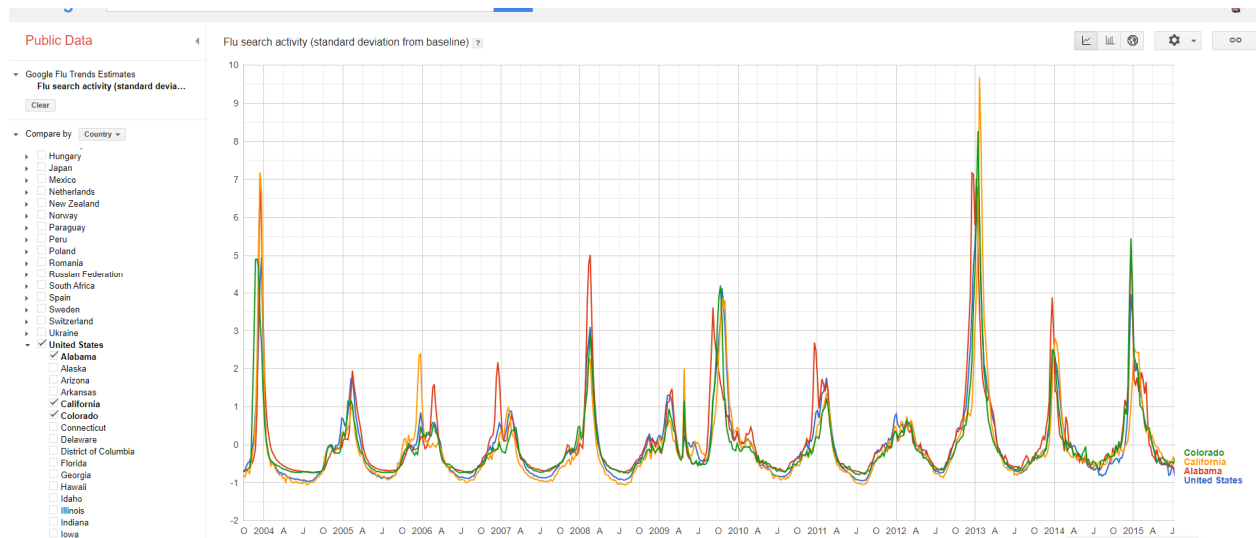
## Monitoring and tracking application



### • Public health monitoring: Google Flu Trends

The US government is encouraging all healthcare stakeholders to establish a national platform for **interoperability and data-sharing** standards. This would enable the secondary use of health data, which would **advance BIG DATA analytics and personalized holistic precision medicine**.

[https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac\\_](https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_)

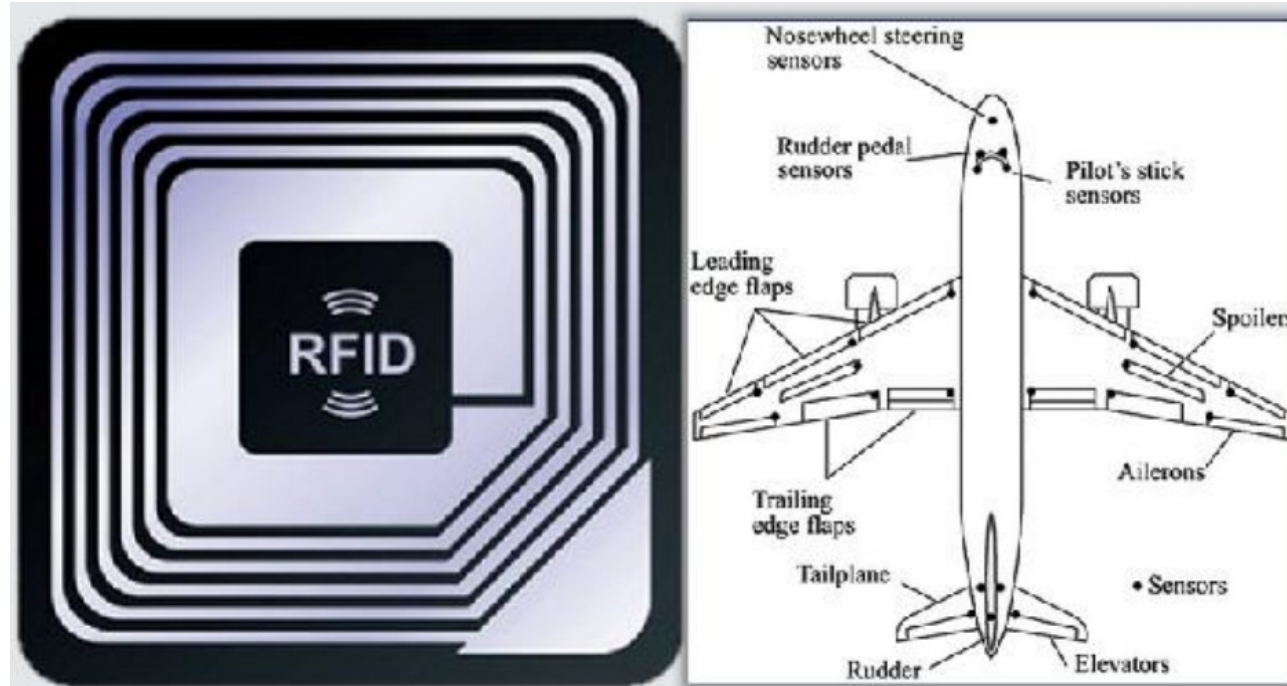


## › Consumer Sentiment Monitoring



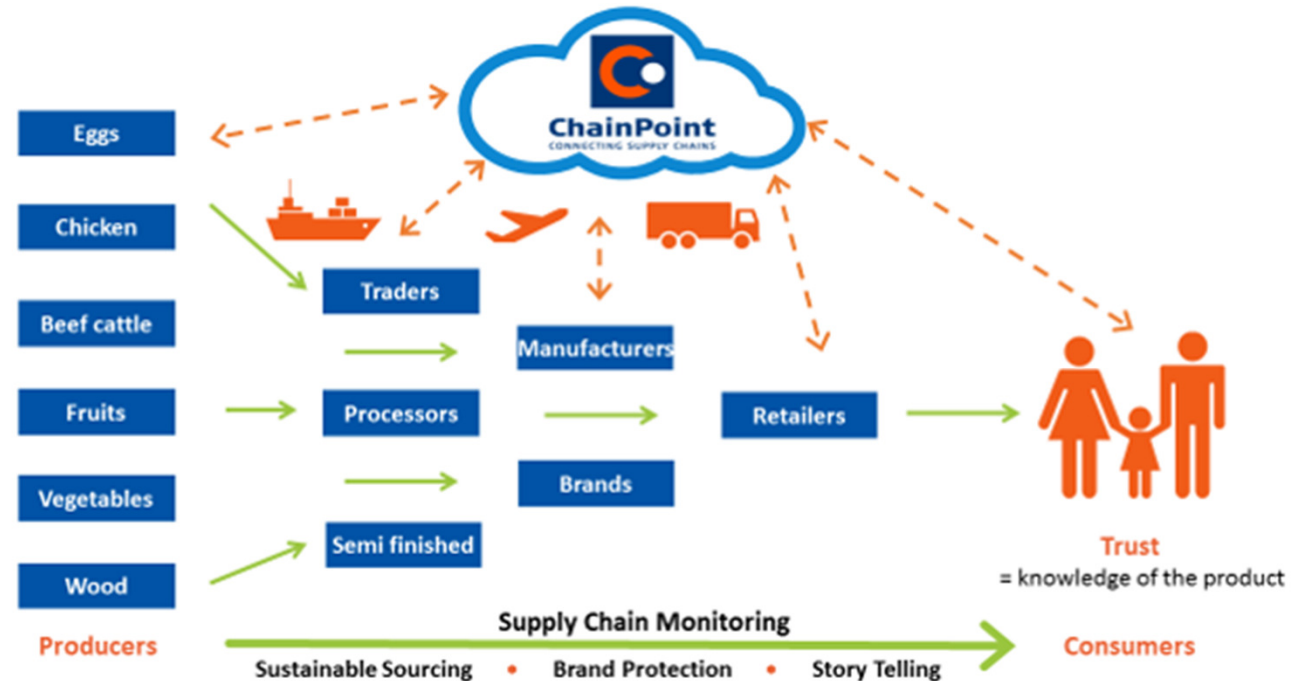
- **Social media** has become more **powerful than advertising**. Many good companies have moved the bulk of their advertising budgets from traditional media to social media.
- They have set up **Big Data listening platforms**, where social media data streams (including tweets, and Facebook posts, and blog posts) are **filtered and analyzed for certain keywords or sentiments**, by certain demographics and regions. Actionable information from this analysis is delivered to marketing professionals for appropriate action, especially when the product is new to the market.

## › Asset Tracking



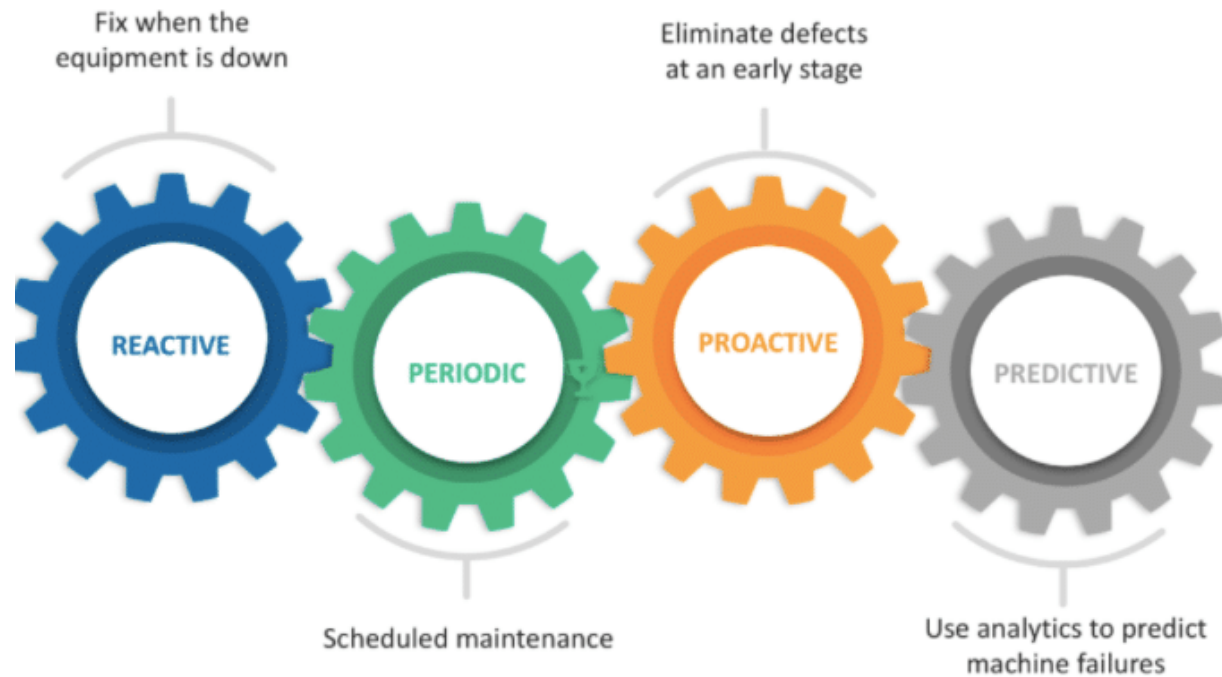
- Theft by shoppers and employees is a major source of loss of revenue for retailers. All valuable items in the store can be assigned RFID tags, and the gates of the store can be equipped with RF readers. This can help secure the products, and reduce leakage(theft) from the store.
- Airplanes are one of the heaviest users of sensors which track every aspect of the performance of every part of the plane. The data can be displayed on the dashboard as well as stored for later detailed analysis. Working with communicating devices, these sensors can produce a torrent of data.

## › Supply chain monitoring



- All containers on ships communicate their status and location using RFID tags.
- Thus retailers and their suppliers can gain real-time visibility to the inventory throughout the global supply chain. Retailers can know exactly where the items are in the warehouse, and so can bring them into the store at the right time.
- This is particularly relevant for seasonal items that must be sold on time, or else they will be sold at a discount. With item-level RFID tags, retailers also gain full visibility of each item and can serve their customers better.

## › Preventive machine maintenance



All machines, including cars and computers, do tend to fail sometimes. This is because one or more of their components may cease to function. As a preventive measure, precious equipment could be equipped with sensors. The continuous stream of data from the sensors could be monitored and analyzed to forecast the status of key components, and thus, monitor the overall machine's health. Preventive maintenance can, thus, reduce the cost of downtime.

## Analysis and Insight Applications-next generation of big data apps

### › Predictive Policing

The notion of predictive policing was created by the Los Angeles Police Department. The LAPD collaborated with UC Berkeley academics to examine its massive database of 13 million crimes spanning 80 years and forecast the likelihood of particular sorts of crimes occurring at specific times and in specific areas. They pinpointed crime hotspots of certain categories, at specific times, and in specific areas. They identified crime hotspots where crimes have happened and were likely to occur in the future.



By aligning the police car patrol schedule with the model's predictions, the LAPD could reduce crime by 12 percent to 26 percent for different categories of crime.



## Analysis and Insight Applications-next generation of big data apps

### › Winning political elections

The US president, Barack Obama was the first major political candidate to use big data in a significant way, in the 2008 elections. He is the first big data president. His campaign gathered data about millions of people, including supporters. They invented the mechanism to obtain small campaign contributions from millions of supporters. They created **personal profiles of millions of supporters** and what they had done and could do for the campaign. **Data was used to determine undecided voters who could be converted to their side.** They provided the phone numbers of these undecided voters to the volunteers.

Senator Bernie Sanders used the same big data playbook to build an effective national political machine powered entirely by small donors. Election analyst, Nate Silver, created sophisticated predictive models using inputs from many political polls and surveys to win Pundits to successfully predict the winner of the US elections. **Nate was, however, unsuccessful in predicting Donald Trump's rise and ultimate victory and that shows the limits of big data.**



## Analysis and Insight Applications-next generation of big data apps

### › Personal health

IBM's **Watson system** is a big data analytics engine that ingests and digests all the medical information in the world and then applies it intelligently to an individual situation.

**Watson can provide a detailed and accurate medical diagnosis using current symptoms, patient history, medical history environmental trends, and other parameters.** Similar products might be offered as an APP to licensed doctors, and even individuals, to improve productivity and accuracy in health care.



## New Product Development

### › Location-based retail promotion

A retailer or a third-party advertiser, can target **customers with specific promotions and coupons based on location data obtained through the Global positioning system (GPS)** the time of day, the presence of stores nearby, and mapping it to the consumer preference data available from social media databases. Advertisements and offers can be delivered through mobile apps, SMS and email. These are examples of mobile apps.



# New Product Development

## > Recommendation service

- E-commerce has been a fast-growing industry in the last couple of decades. A variety of products are sold and shared over the internet.
- Web users' browsing and purchase history on e-commerce sites are utilized to learn about their preferences and needs and to advertise relevant product and pricing offers in real-time. Amazon uses a personalized recommendation engine system to suggest new additional products to consumers based on the affinities of various products.
- Netflix also uses a recommendation engine to suggest entertainment options to its users. Big data is valuable across all industries.

