Faculty of engineering - Shoubra
Benha University

Research Article / Research Project / Literature Review

in fulfillment of the requirements of

Department	**Engineering Mathematics and Physics**
Division	**----------**
Academic Year	**2019-2020 Preparatory**
Course name	**Computer**
Course code	**ECE001**

Title: -

BIG DATA

By:

	Name	Edu mail	B.N
1	**محمد نبيل السيد دياب**	**muhammed195903@feng.bu.edu.eg**	**811**

Approved by:

Examiners committee	Signature
Dr.Ahmed Bayoumi	
Dr.Shady Elmashad	
Dr.Abdelhamid Attaby	

Abstract

Big data is one of the hottest topics in the last few years in this research some of its main aspects and applications are covered as a first step to get the whole picture understood.

A simple website is created to demonstrate this research and uploaded to Github:

https://github.com/MNElsayed/html-project

and published through GitHub Pages:

https://mnelsayed.github.io/html-project/

Table of contents

website Source code and Screenshots

1. The first page is a simple introduction to Big data

```
<html><style>
body{width:75%; margin:auto; font-size :20px;}
</style>
<head>
<h1 style="text-align:center;" >BIG DATA</h1>
</head>
<body>
<ul>
<li>Introduction</li>
<li><a href="the Vs.html">The <b>V</b>s of Big Data</a></li>
<li><a href="applications.html">Big Data Applications</a></li>
<li><a href="Types.html">Types Of Big Data</a></li>
<li><a href="challenges.html">Big data challenges</a></li>
</ul>
<h2 >Introduction</h2>
<p>Big data is a combination of structured, semistructured and unstructured data collected by organizations
that can be mined for information and used in machine learning projects, predictive modeling and other
advanced analytics applications.
Systems that process and store big data have become a common component of data management architectures
in organizations.</p>
<img src="big-data.jpg">
<h3>Big data is often characterized by the 3<b>V</b>s:</h3>
<ul>
<li>the <b>V</b>olume of data in many environments</li>
<li>the <b>V</b>ariety of data types stored in big data systems</li>
<li>the <b>V</b>elocity at which the data is generated, collected and processed</li>
</ul>
<p>These characteristics were first identified by Doug Laney, then an analyst at Meta Group Inc., in 2001.
Gartner further popularized them after it acquired Meta Group in 2005. More recently,
several other Vs have been added to different descriptions of big data, including veracity, value and variability.
Although big data doesn't equate to any specific volume of data, big data deployments often involve
terabytes (TB), petabytes (PB) and even exabytes (EB) of data captured over time.</p>
<h2 >Importance of Big Data</h2>
<p>Companies use the big data accumulated in their systems to improve operations, provide better customer
service,
create personalized marketing campaigns based on specific customer preferences and, ultimately,
increase profitability. Businesses that utilize big data hold a potential competitive advantage over those
that don't since they're able to make faster and more informed business decisions, provided they use the data
effectively.
For example, big data can provide companies with valuable insights into their customers that can be used to
refine marketing campaigns and techniques in order to increase customer engagement and conversion rates. </p>
</body></html>
```

BIG DATA

Introduction

Big data is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications. Systems that process and store big data have become a common component of data management architectures in organizations.

Big data is often characterized by the 3Vs:

- the **V**olume of data in many environments
- the **V**ariety of data types stored in big data systems
- the **V**elocity at which the data is generated, collected and processed

These characteristics were first identified by Doug Laney, then an analyst at Meta Group Inc., in 2001. Gartner further popularized them after it acquired Meta Group in 2005. More recently, several other Vs have been added to different descriptions of big data, including veracity, value and variability. Although big data doesn't equate to any specific volume of data, big data deployments often involve terabytes (TB), petabytes (PB) and even exabytes (EB) of data captured over time.

Importance of Big Data

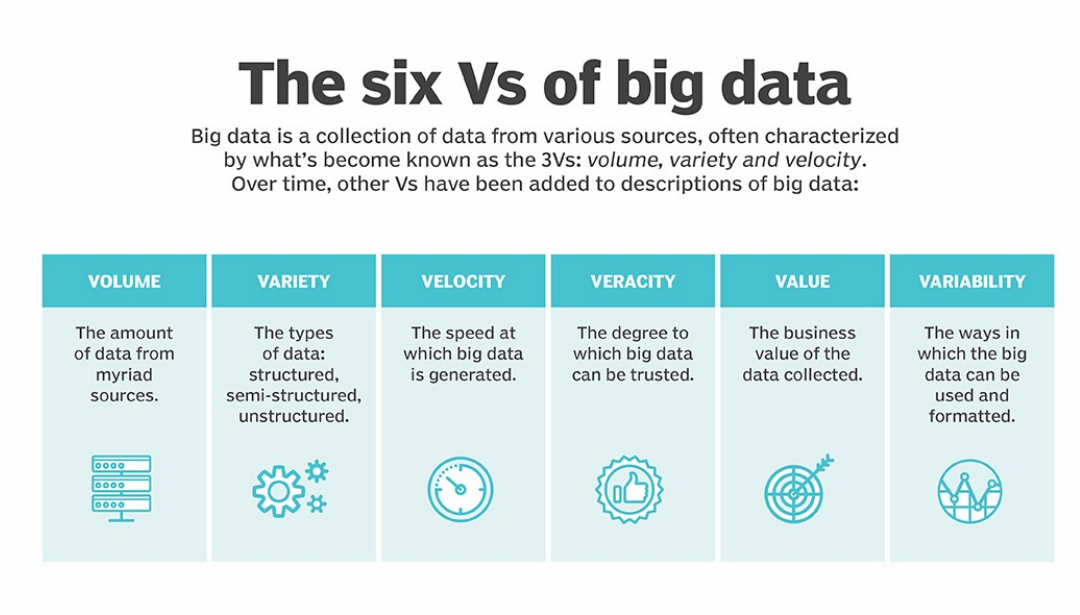
Companies use the big data accumulated in their systems to improve operations, provide better customer service, create personalized marketing campaigns based on specific customer preferences and, ultimately, increase profitability. Businesses that utilize big data hold a potential competitive advantage over those that don't since they're able to make faster and more informed business decisions, provided they use the data effectively. For example, big data can provide companies with valuable insights into their customers that can be used to refine marketing campaigns and techniques in order to increase customer engagement and conversion rates.

2. In The second page some of big data features are discussed.

```
<html><style>
body{width:75%; margin:auto; font-size :20px;}
</style>
<head><h1 style="text-align:center;" >BIG DATA</h1></head>
<body>
<ul>
<li><a href="index.html">Introduction</a></li>
<li>The <b>V</b>s of Big Data</li>
<li><a href="applications.html">Big Data Applications</a></li>
<li><a href="Types.html">Types Of Big Data</a></li>
<li><a href="challenges.html">Big data challenges</a></li>
</ul>
<h2>Breaking down the <b>V</b>s of big data</h2>
<img src="big_data_vs.jpg">
<p>
Volume is the most commonly cited characteristic of big data. A big data environment doesn't have to contain
a large amount of data, but most do because of the nature of the data being collected and stored in them.
Clickstreams, system logs and stream processing systems are among the sources that typically produce massive
volumes of big data on an ongoing basis.
</p>
<h3>Big data also encompasses a wide variety of data types, including the following:</h3>
<ul>
<li>structured data in databases and data warehouses based on Structured Query Language (SQL).</li>
<li>unstructured data, such as text and document files held in Hadoop clusters or NoSQL database systems.</li>
<li>semistructured data, such as web server logs or streaming data from sensors.</li>
</ul>
<p>
All of the various data types can be stored together in a data lake, which typically is based on Hadoop
or a cloud object storage service. In addition, big data applications often include multiple data sources
that may not otherwise be integrated. For example, a big data analytics project may attempt to gauge a
product's success and future sales by correlating past sales data, return data and online buyer review
data for that product.
</p>
<p>
Velocity refers to the speed at which big data is generated and must be processed and analyzed.
In many cases, sets of big data are updated on a real- or near-real-time basis, instead of the daily,
weekly or monthly updates made in many traditional data warehouses. Big data analytics applications ingest,
correlate and analyze the incoming data and then render an answer or result based on an overarching query.
This means data scientists and other data analysts must have a detailed understanding of the available
data and possess some sense of what answers they're looking for to make sure the information they get is
valid and up to date.
</p>
<p>
Managing data velocity is also important as big data analysis expands into fields like machine learning
and artificial intelligence (AI), where analytical processes automatically find patterns in the collected
data and use them to generate insights.
</p></body></html>
```

BIG DATA

Breaking down the Vs of big data



Volume is the most commonly cited characteristic of big data. A big data environment doesn't have to contain a large amount of data, but most do because of the nature of the data being collected and stored in them. Clickstreams, system logs and stream processing systems are among the sources that typically produce massive volumes of big data on an ongoing basis.

Big data also encompasses a wide variety of data types, including the following:

- structured data in databases and data warehouses based on Structured Query Language (SQL).
- unstructured data, such as text and document files held in Hadoop clusters or NoSQL database systems.
- semistructured data, such as web server logs or streaming data from sensors.

All of the various data types can be stored together in a data lake, which typically is based on Hadoop or a cloud object storage service. In addition, big data applications often include multiple data sources that may not otherwise be integrated. For example, a big data analytics project may attempt to gauge a product's success and future sales by correlating past sales data, return data and online buyer review data for that product.

Velocity refers to the speed at which big data is generated and must be processed and analyzed. In many cases, sets of big data are updated on a real- or near-real-time basis, instead of the daily, weekly or monthly updates made in many traditional data warehouses. Big data analytics applications ingest, correlate and analyze the incoming data and then render an answer or result based on an overarching query. This means data scientists and other data analysts must have a detailed understanding of the available data and possess some sense of what answers they're looking for to make sure the information they get is valid and up to date.

Managing data velocity is also important as big data analysis expands into fields like machine learning and artificial intelligence (AI), where analytical processes automatically find patterns in the collected data and use them to generate insights.

3. The third page discusses some of the biggest areas where big data has a huge impact.

```
<html>
	<style>
		body{width:75%; margin:auto; font-size :20px;}
	</style>
	<head>
		<h1 style="text-align:center;" >BIG DATA</h1>
	</head>
	<body>
		<ul>
			<li><a href="index.html">Introduction</a></li>
			<li><a href="the Vs.html">The <b>V</b>s of Big Data</a></li>
			<li>Big Data Applications</li>
			<li><a href="Types.html">Types Of Big Data</a></li>
			<li><a href="challenges.html">Big data challenges</a></li>
		</ul>

		<h2 >Big Data Applications</h2>

		<h2>Big Data in Retail</h2>
		<p>
			The retail industry is the one that faces the most fierce competition of
all. Retailers constantly hunt
			for ways that will give them a competitive edge over others. Customers
are the real king sounds legit
			for the retail industry in particular.
			For retailers to thrive in this competitive world, they need to understand
their customers in a better way.
			If they are aware of their customers' needs and how to fulfill those
needs in the best possible way,
			then they know everything.
		</p>
		<img src="big-data-in-retail.jpg">
		<h2 >Big Data in Healthcare</h2>
		<p>
			Big Data and healthcare are an ideal match. It complements the
healthcare industry better than anything ever will.
			The amount of data the healthcare industry has to deal with is
unimaginable.
			Gone are the days when healthcare practitioners were incapable of
harnessing this data. From finding a cure to
			cancer to detecting Ebola and much more, Big Data has got it all under
its belt and researchers have seen some
			life-saving outcomes through it.
		</p>
		<img src="big-data-in-healthcare.jpg">
	</body>
</html>
```

BIG DATA

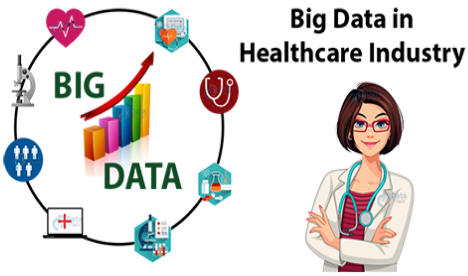
Big Data Applications

Big Data in Retail

The retail industry is the one that faces the most fierce competition of all. Retailers constantly hunt for ways that will give them a competitive edge over others. Customers are the real king sounds legit for the retail industry in particular. For retailers to thrive in this competitive world, they need to understand their customers in a better way. If they are aware of their customers' needs and how to fulfill those needs in the best possible way, then they know everything.

Big Data in Healthcare

Big Data and healthcare are an ideal match. It complements the healthcare industry better than anything ever will. The amount of data the healthcare industry has to deal with is unimaginable. Gone are the days when healthcare practitioners were incapable of harnessing this data. From finding a cure to cancer to detecting Ebola and much more, Big Data has got it all under its belt and researchers have seen some life-saving outcomes through it.



4. In the fourth page various types of Data are illustrated with examples

```
<html><style>
		body{width:75%; margin:auto; font-size :20px;}
		table, th, td {border: 1px solid black;border-collapse: collapse;}
	</style>
	<head><h1 style="text-align:center;" >BIG DATA</h1></head>
	<body><ul>
			<li><a href="index.html">Introduction</a></li>
			<li><a href="the Vs.html">The <b>V</b>s of Big Data</a></li>
			<li><a href="applications.html">Big Data Applications</a></li>
			<li>Types Of Big Data</li>
			<li><a href="challenges.html">Big data challenges</a></li>
		</ul>
		<h2 >Types Of Big Data</h2>
		<h3>Big Data could be found in three forms:</h3>
		<ol>
			<li><b>Structured</b></li>
			<li><b>Unstructured</b></li>
			<li><b>Semi-structured</b></li>
		</ol>
		<h3>Structured</h3>
		<p>Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes. </p><h4>Examples Of Structured Data</h4>
		<table>
			<tr><th>Employee_ID</th>
				<th>Employee_Name</th>
				<th>Gender</th>
				<th>Department</th></tr>
			<tr><td>2365</td>  <td>ahmed</td>  <td>Male</td>  <td>Finance</td>tr>
			<tr><td>3398</td>  <td>marwa</td>  <td>Female</td>  <td>Admin</td></tr>
			<tr><td>7465</td>	<td>Ali</td>	<td>Male</td>  <td>Admin</td></tr>
		</table>
		<h3 >Unstructured</h3>
		<p>Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.
		<b>The output returned by 'Google Search'</b>is a good example for Unstructured data</p>
		<img src="unstructured.jpg">
		<h3 >Semi-Structured</h3>
		<p>
			Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.
			<b>Personal data stored in an XML file</b> is Example Of Semi-structured Data<p>
		<img src="semi-structured.jpg">
</body></html>
```

BIG DATA

Types Of Big Data

Big Data could be found in three forms:

1. **Structured**
2. **Unstructured**
3. **Semi-structured**

Structured

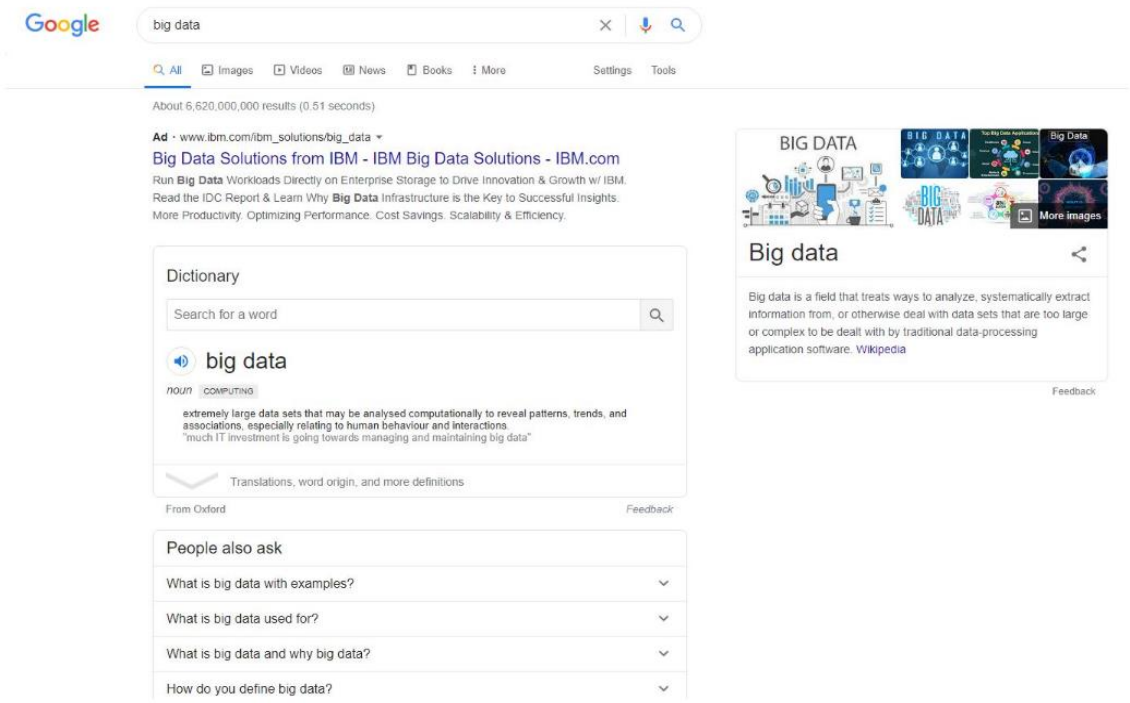
Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

Examples Of Structured Data

Employee_ID	Employee_Name	Gender	Department
2365	ahmed	Male	Finance
3398	marwa	Female	Admin
7465	Ali	Male	Admin

Unstructured

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format. **The output returned by 'Google Search'** is a good example for Unstructured data.



Semi-Structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file. **Personal data stored in an XML file** is Example Of Semi-structured Data

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

5. the fifth and final page is a brief about big data future.

```
<html>
	<style>
		body{width:75%; margin:auto; font-size :20px;}
	</style>
	<head><h1 style="text-align:center;" >BIG DATA</h1><head>
	<body>
		<ul>
			<li><a href="index.html"> Introduction </a> </li>
				<li><a href="the Vs.html">The <b>V</b>s of Big Data</a></li>
			<li><a href="applications.html">Big Data Applications</a></li>
			<li><a href="Types.html">Types Of Big Data</a></li>
			<li>Big data challenges</li>
		</ul>
		<h2 >Big data challenges</h2>
		<p>Besides the processing capacity and cost issues, designing a big data architecture is another common challenge for users. Big data systems must be tailored to an organization's particular needs, a DIY undertaking that requires IT teams and application developers to piece together a set of tools from all the available technologies. Deploying and managing big data systems also require new skills compared to the ones possessed by database administrators (DBAs) and developers focused on relational software. Both of those issues can be eased by using a managed cloud service, but IT managers need to keep a close eye on cloud usage to make sure costs don't get out of hand. Also, migrating on-premises data sets and processing workloads to the cloud is often a complex process for organizations. Making the data in big data systems accessible to data scientists and other analysts is also a challenge, especially in distributed environments that include a mix of different platforms and data stores. To help analysts find relevant data, IT and analytics teams are increasingly working to build data catalogs that incorporate metadata management and data lineage functions. Data quality and data governance also need to be priorities to ensure that sets of big data are clean, consistent and used properly. </p>
	</body>
</html>
```

BIG DATA

- Introduction
- The Vs of Big Data
- Big Data Applications
- Types Of Big Data
- Big data challenges

Big data challenges

Besides the processing capacity and cost issues, designing a big data architecture is another common challenge for users. Big data systems must be tailored to an organization's particular needs, a DIY undertaking that requires IT teams and application developers to piece together a set of tools from all the available technologies. Deploying and managing big data systems also require new skills compared to the ones possessed by database administrators (DBAs) and developers focused on relational software.Both of those issues can be eased by using a managed cloud service, but IT managers need to keep a close eye on cloud usage to make sure costs don't get out of hand. Also, migrating on-premises data sets and processing workloads to the cloud is often a complex process for organizations. Making the data in big data systems accessible to data scientists and other analysts is also a challenge, especially in distributed environments that include a mix of different platforms and data stores. To help analysts find relevant data, IT and analytics teams are increasingly working to build data catalogs that incorporate metadata management and data lineage functions. Data quality and data governance also need to be priorities to ensure that sets of big data are clean, consistent and used properly.

Results and discussion

Big data that term we have been hearing and still hearing about in almost all technological and none technological events.
It has simply affected almost all of our modern day to day life.
Big data has increased the demand of information management specialists so much so that Software AG, Oracle
Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole.

- ## Government

The use and adoption of big data within governmental processes allows efficiencies in terms of cost, productivity, and innovation, but does not come without its flaws. Data analysis often requires multiple parts of government (central and local) to work in collaboration and create new and innovative processes to deliver the desired outcome.
CRVS (civil registration and vital statistics) collects all certificates status from birth to death. CRVS is a source of big data for governments.

- ## International development

Research on the effective usage of information and communication technologies for development (also known as ICT4D) suggests that big data technology can make important contributions but also present unique challenges to International development. Advancements in big data analysis offer cost-effective opportunities to improve decision-making in critical development areas such as health care, employment, economic productivity, crime, security, and natural disaster and resource management. Additionally, user-generated data offers new opportunities to give the unheard a
voice. However, longstanding challenges for developing regions such as inadequate technological infrastructure and economic and human resource scarcity exacerbate existing concerns with big data such as privacy, imperfect methodology, and interoperability issues.

- **Healthcare**

Big data analytics has helped healthcare improve by providing personalized medicine and prescriptive analytics, clinical risk intervention and predictive analytics, waste and care variability reduction, automated external and internal reporting of patient data, standardized medical terms and patient registries and fragmented point solutions. Some areas of improvement are more aspirational than actually implemented. The level of data generated within healthcare systems is not trivial. With the added adoption of mHealth, eHealth and wearable technologies the volume of data will continue to increase. This includes electronic health record data, imaging data, patient generated data, sensor data, and other forms of difficult to process data. There is now an even greater need for such environments to pay greater attention to data and information quality. "Big data very often means 'dirty data' and the fraction of data inaccuracies increases with data volume growth." Human inspection at the big data scale is impossible and there is a desperate need in health service for intelligent tools for accuracy and believability control and handling of information missed. While extensive information in healthcare is now electronic, it fits under the big data umbrella as most is unstructured and difficult to use. The use of big data in healthcare has raised significant ethical challenges ranging from risks for individual rights, privacy and autonomy, to transparency and trust.
Big data in health research is particularly promising in terms of exploratory biomedical research, as data-driven analysis can move forward more quickly than hypothesis-driven research. Then, trends seen in data analysis can be tested in traditional, hypothesis-driven followup biological research and eventually clinical research.
A related application sub-area, that heavily relies on big data, within the healthcare field is that of computer-aided diagnosis in medicine. [69] One only needs to recall that, for instance, for epilepsy monitoring it is customary to create 5 to 10 GB of data daily. Similarly, a single uncompressed image of breast tomosynthesis averages 450 MB of data. These are just few of the many examples where computer-aided diagnosis uses big data. For this reason, big data has been recognized as one of the seven key challenges that computer-aided diagnosis systems need to overcome in order to reach the next level of performance.

- **Education**

A McKinsey Global Institute study found a shortage of 1.5 million highly trained data professionals and managers and a number of universities

including University of Tennessee and UC Berkeley, have created masters programs to meet this demand. Private boot camps have also developed programs to meet that demand, including free programs like The Data Incubator or paid programs like General Assembly. In the specific field of marketing, one of the problems stressed by Wedel and Kannan is that marketing has several sub domains (e.g., advertising, promotions, product development, branding) that all use different types of data. Because one-size-fits-all analytical solutions are not desirable, business schools should prepare marketing managers to have wide knowledge on all the different techniques used in these sub domains to get a big picture and work effectively with analysts.

- **Media**

To understand how the media utilizes big data, it is first necessary to provide some context into the mechanism used for media process. It has been suggested by Nick Couldry and Joseph Turow that practitioners in Media and Advertising approach big data as many actionable points of information about millions of individuals. The industry appears to be moving away from the traditional approach of using specific media environments such as newspapers, magazines, or television shows and instead taps into consumers with technologies that reach targeted people at optimal times in optimal locations. The ultimate aim is to serve or convey, a message or content that is (statistically speaking) in line with the consumer's mindset. For example, publishing environments are increasingly tailoring messages (advertisements) and content (articles) to appeal to consumers that have been exclusively gleaned through various data-mining activities.

- Targeting of consumers (for advertising by marketers)
- Data capture
- Data journalism: publishers and journalists use big data tools to provide unique and innovative insights and infographics.

- **Insurance**

Health insurance providers are collecting data on social "determinants of health" such as food and TV consumption, marital status, clothing size and purchasing habits, from which they make predictions on health costs, in order to spot health issues in their clients. It is controversial whether these predictions are currently being used for pricing.

- # Internet of Things (IoT)

Main article: Internet of Things
Further information: Edge computing
Big data and the IoT work in conjunction. Data extracted from IoT devices provides a mapping of device inter-connectivity. Such mappings have been used by the media industry, companies and governments to more accurately target their audience and increase media efficiency. IoT is also increasingly adopted as a means of gathering sensory data, and this sensory data has been used in medical, manufacturing and transportation contexts.
Kevin Ashton, digital innovation expert who is credited with coining the term, defines the Internet of Things in this quote: "If we had computers that knew everything there was to know about things—using data they gathered without any help from us—we would be able to track and count everything, and greatly reduce waste, loss, and cost. We would know when things needed replacing, repairing or recalling, and whether they were fresh or past their best."

- # Information technology

Especially since 2015, big data has come to prominence within business operations as a tool to help employees work more efficiently and streamline the collection and distribution of information technology (IT). The use of big data to resolve IT and data collection issues within an enterprise is called IT operations analytics (ITOA).[84] By applying big data principles into the concepts of machine intelligence and deep computing, IT departments can predict potential issues and move to provide solutions before the problems even happen.[84] In this time, ITOA businesses were also beginning to play a major role in systems management by offering platforms that brought individual data silos together and generated insights from the whole of the system rather than from isolated pockets of data.

Conclusions

In conclusion Big data has invaded almost every part of our lives. Developed economies increasingly use data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide, and between 1 billion and 2 billion people accessing the internet. Between 1990 and 2005, more than 1 billion people worldwide entered the middle class, which means more people became more literate, which in turn led to information growth. The world's effective capacity to exchange information through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007 and predictions put the amount of internet traffic at 667 exabytes annually by 2014. According to one estimate, one-third of the globally stored information is in the form of alphanumeric text and still image data, which is the format most useful for most big data applications. This also shows the potential of yet unused data (i.e. in the form of video and audio content).

While many vendors offer off-the-shelf solutions for big data, experts recommend the development of in-house solutions custom-tailored to solve the company's problem at hand if the company has sufficient technical capabilities.