Classification of Digital Data

Digital data is classified into the following categories:

- Structured data
- Semi-structured data
- Unstructured data

Classification of Digital Data

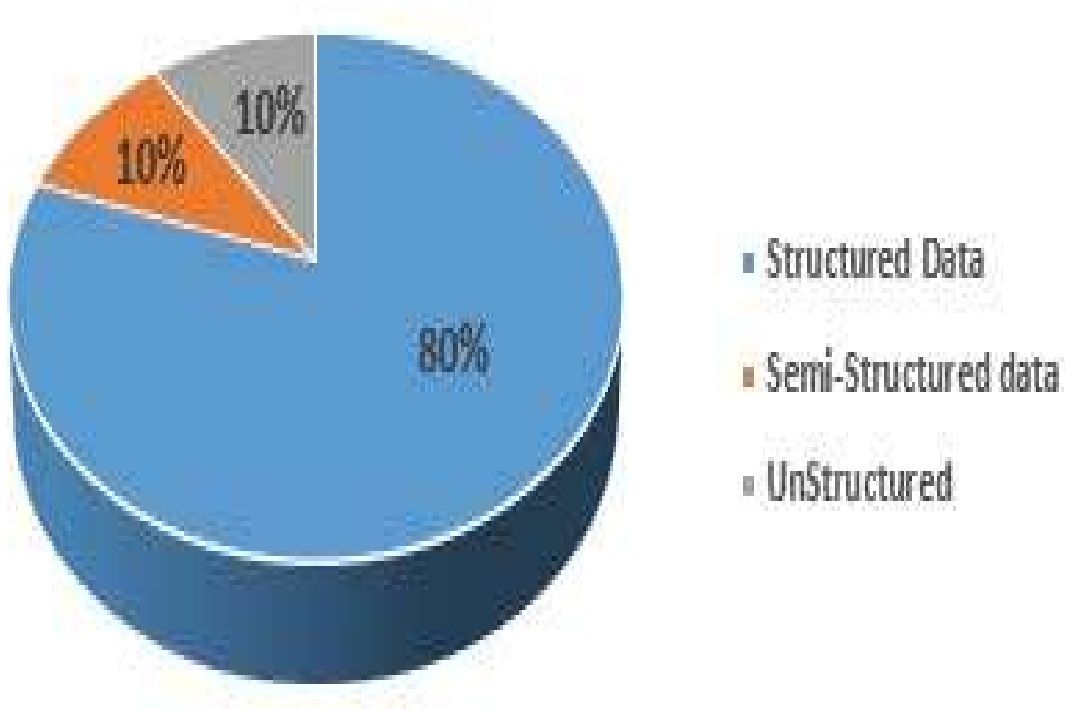
- **Unstructured data:**
  - This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.
  - About 80-90% data of an organization is in this for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email etc.

Classification of Digital Data..

- **Semi-structured data:** This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program;
- for example, en XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.
- **Structured data:** This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes their objects. Data stored in databases is an example of structured data.

Approximate Percentage Distribution of Digital Data

- Approximate percentage distribution of digital data



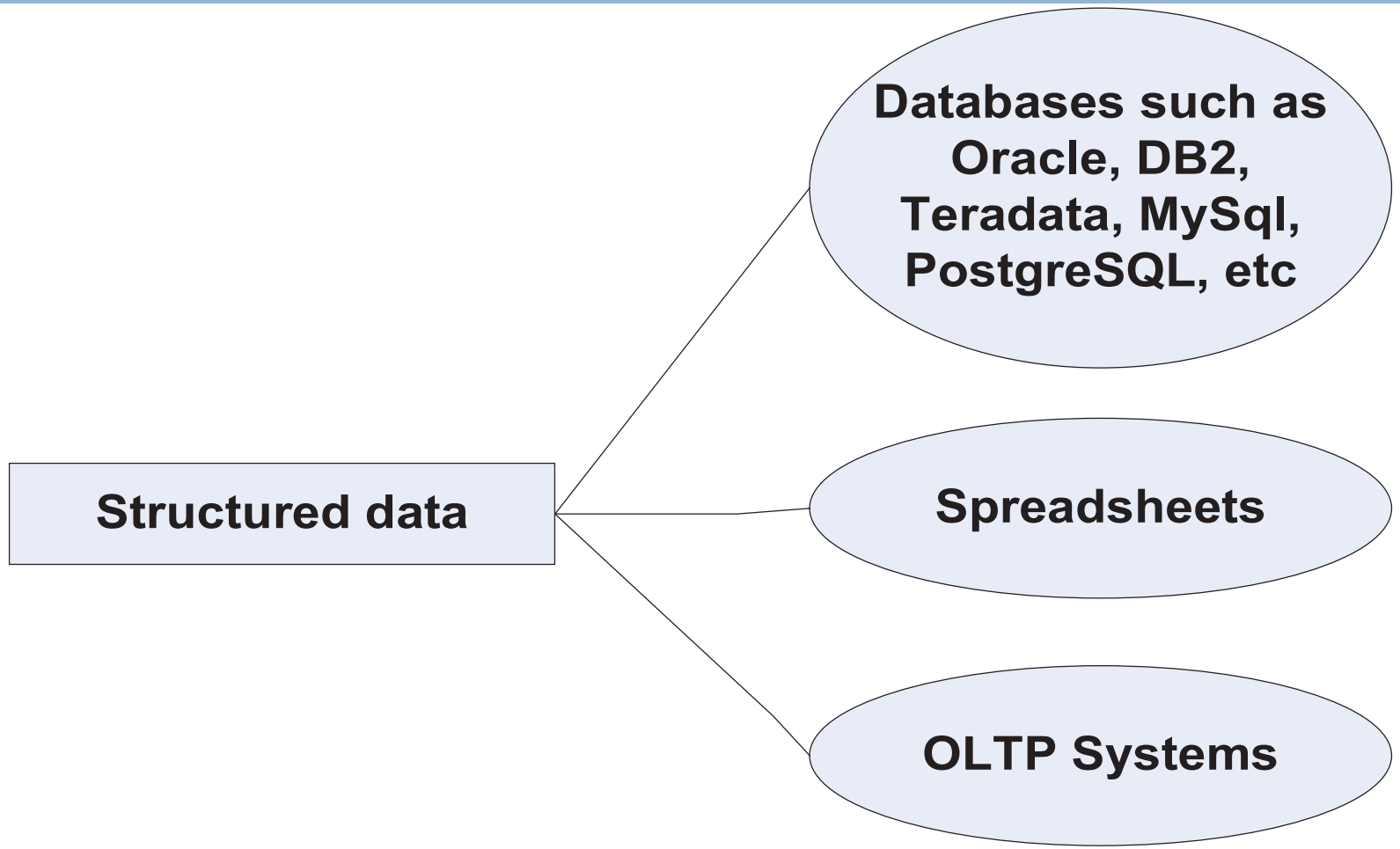
Structured Data

- This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program.
- Relationships exist between entities of data, such as classes and their objects.
- Data stored in databases is an example of structured data.

Sources of Structured Data

- If your data is highly structured, one can look at leveraging any of the available RDBMS
- [Oracle Corp. — Oracle, IBM — DB2, Microsoft — Microsoft SQL Server, EMC — Greenplum, Teradata — Teradata, MySQL (open source), PostgreSQL (advanced open source) etc.] to house it.
- These databases are typically used to hold transaction/operational data generated and collected by day-to-day business activities. In other words, the data of the On-Line Transaction Processing (OLTP) systems are generally quite structured.

Sources of Structured Data



Ease of Working with Structured Data

The ease is with respect to the following:

- **Insert/update/delete:** The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.
- **Security:** How does one ensure the security of information? There are available check encryption and tokenization solutions to warrant the security of information throughout its lifecycle.
- Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.

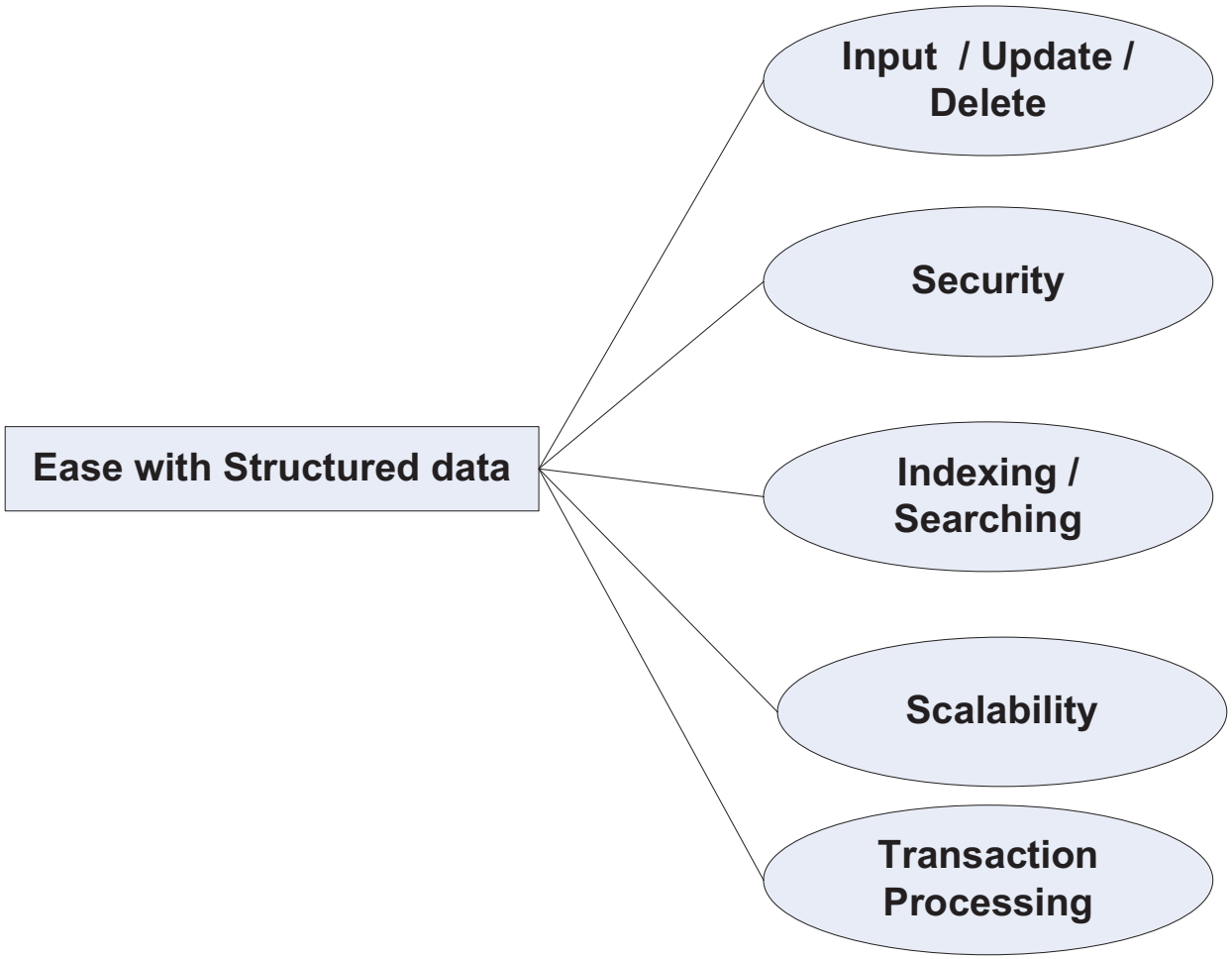
Ease of Working with Structured Data

- **Indexing:** An index is a data structure that speeds up the data retrieval operations (primarily the SELECT DML statement) at the cost of additional writes and storage space, but the benefits that ensue in search operation are worth the additional writes and storage space.
- **Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server *(increasing the primary and secondary or peripheral storage capacity, processing capacity of the processor, etc.)*.

Ease of Working with Structured Data

- **Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction.
  - ***Atomicity:*** A transaction is atomic, means that either it happens in its entirety or none of it at all.
  - ***Consistency:*** The database moves from one consistent state to another consistent state. In other words, if the same piece of information is stored at two or more places, they are in complete agreement.
  - ***Isolation:*** The resource allocation to the transaction happens such that the transaction gets the impression that it is the only transaction happening in isolation.
  - ***Durability:*** All changes made to the database during a transaction are permanent and that accounts for the durability of the transaction.

Ease with Structured Data



Semi-structured Data

- This is the data which does not conform to a data model but has some structure.
- However, it is not in a form which can be used easily by a computer program.
- Example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

Semi-structured Data

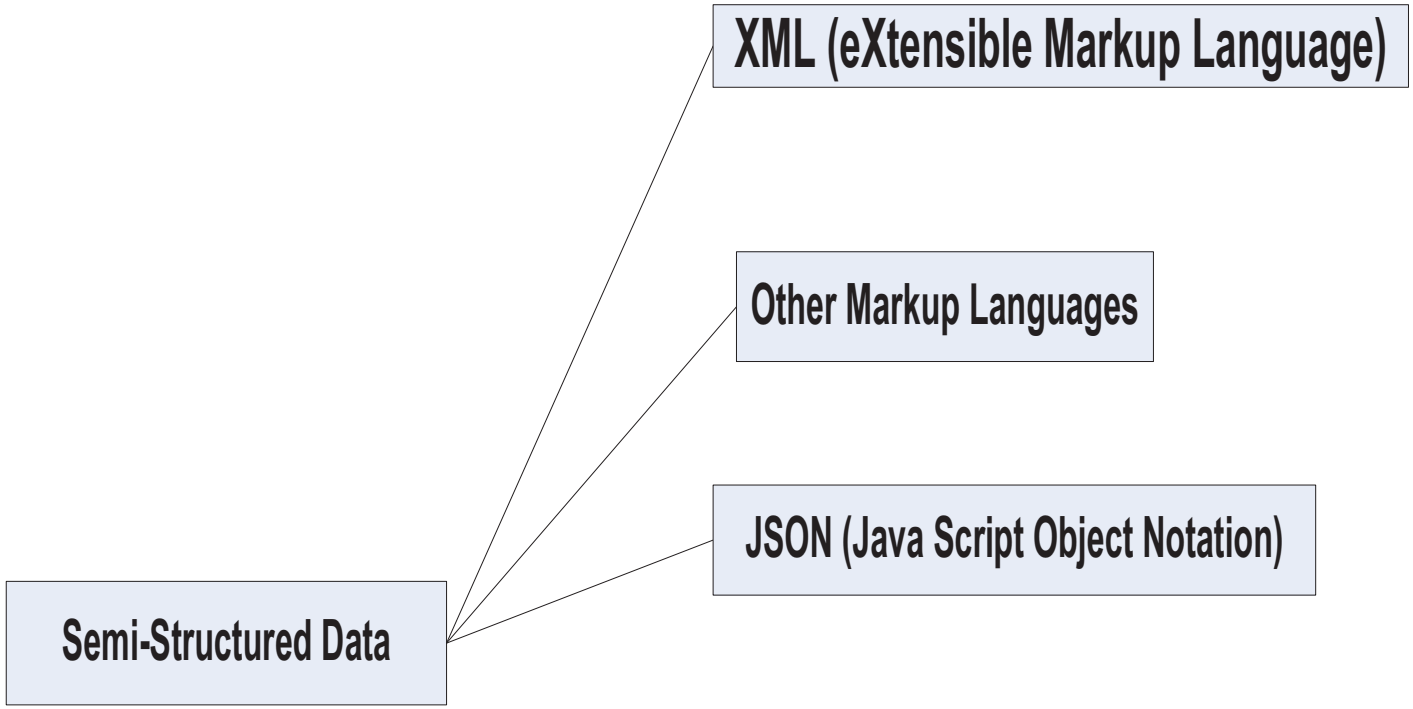
It has the following features:

- It does not conform to the data models that one typically associates with relational databases or any other form of data tables.
- It uses tags to segregate semantic elements.
- Tags are also used to enforce hierarchies of records and fields within data.
- There is no separation between the data and the schema.
- The amount of structure used is dictated by the purpose at hand.
- In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes.
- And if at all, they have the same set of attributes, the order of attributes may not be similar and for all practical purposes it is not important as well.

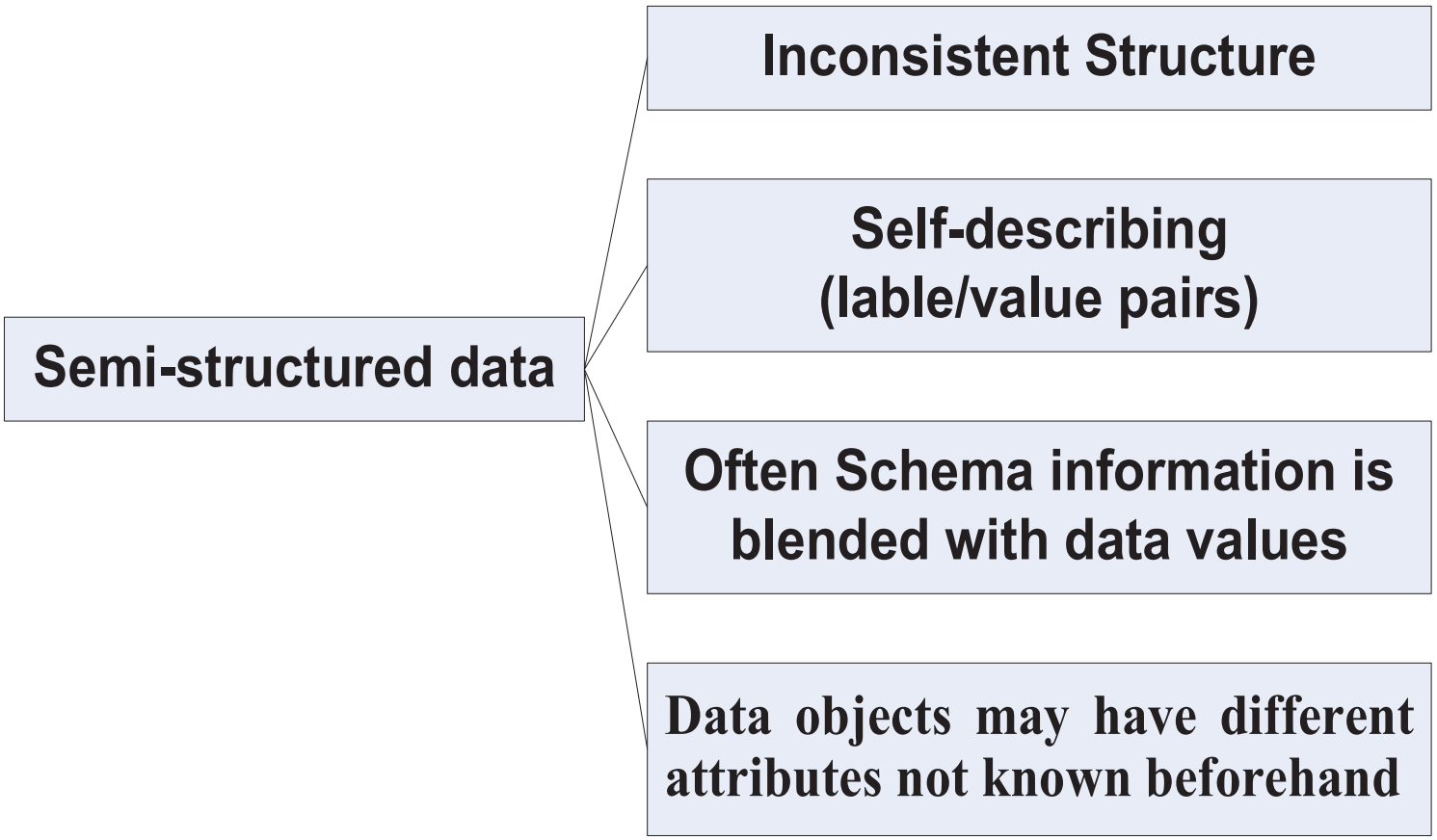
Sources of Semi-structured Data

- Amongst the sources for semi-structured data, the front runners are “XML” and “JSON”.
- **XML:** eXtensible Markup Language (XML) is hugely popularized by web services developed utilizing the Simple Object Access Protocol (SOAP) principles.

Sources of Semi-structured Data



Characteristics of Semi-structured Data



Sources of Semi-structured Data

- **JSON:** Java Script Object Notation (JSON) is used to transmit data between a server and a web application.
- JSON is popularized by web services developed utilizing the Representational State Transfer (REST) - an architecture style for creating scalable web services.
- MongoDB (open-source, distributed, NoSQL, documented-oriented database) and Couchbase (originally known as Membase, open-source, distributed, NoSQL, document-oriented database) store data natively in JSON format.

Sources of Semi-structured Data

An example of HTML is as follows:

```
<HTML>
    <HEAD>
    <TITLE>Place your title here</TITLE>
    </HEAD>
<BODY BGCOLOR="FFFFFF">
     <CENTER><IMG SRC="clouds.jpg" ALIGN="BOTTOM"x/CENTER>
     <HR>          <a href="http://bigdatauniversity.com">Link Name</a>
     <HI>this is a Header</HI>
     <H2>this is a sub Header</H2>
     Send me mail at <a href="mailto:support@yourcompany.com"> support@yourcompany.com</a>.
     <P>a new paragraph!
     <PxB>a new paragraph!</B>
     <BRxBxI>this is a new sentence without a paragraph break, in bold italics.</Ix/B>
     <HR>
</BODY>
</HTML>
```

Sources of Semi-structured Data

Sample JSON document

{

_id:9,

BookTitle: “Fundamentals of Business Analytics”,

AuthorName: “Seema Acharya”,

Publisher: “Wiley India”,

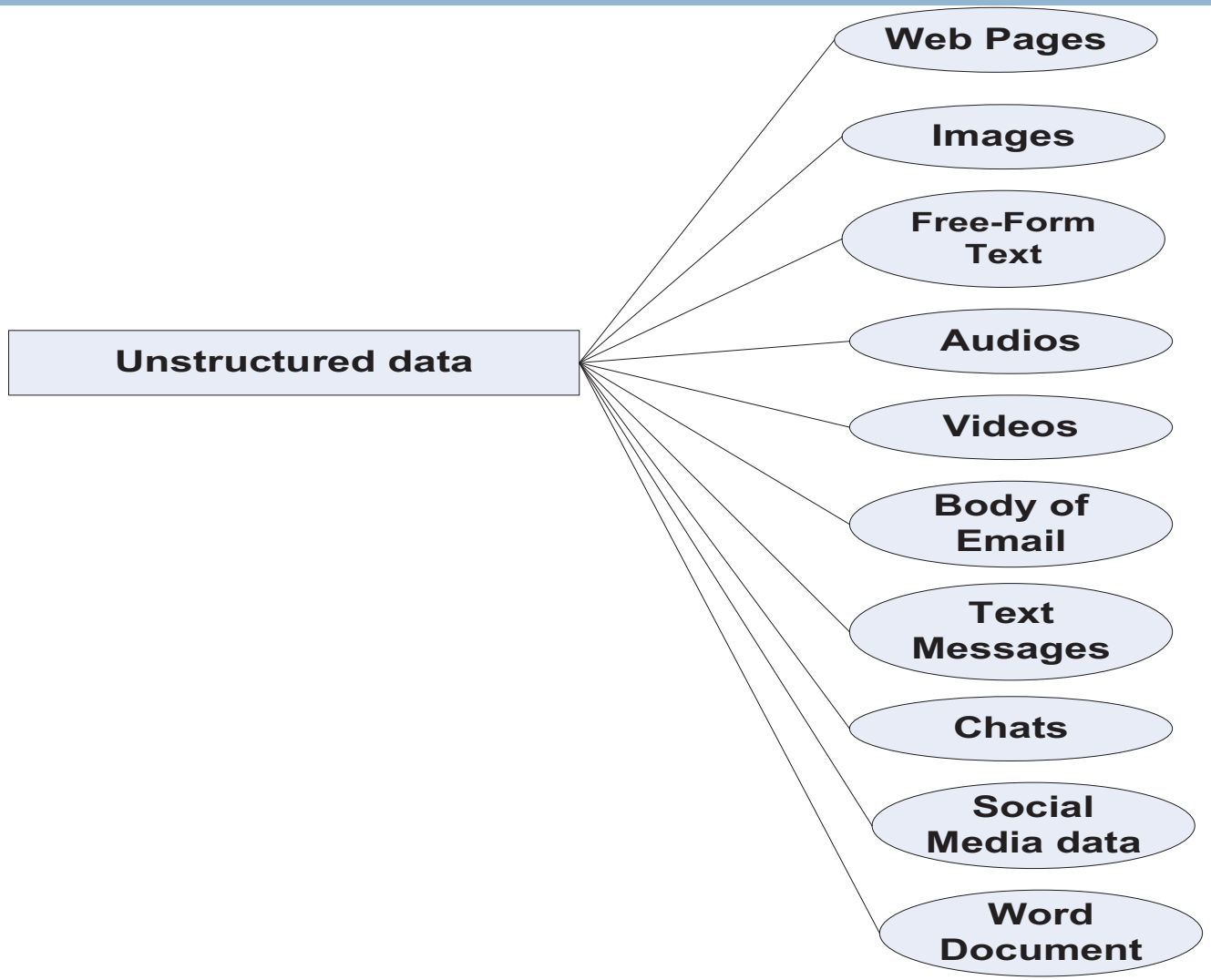
YearofPublication: “2011”

}

Unstructured Data

- This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.
- About 80–90% data of an organization is in this format.
- **Example:** memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

Sources of Unstructured Data



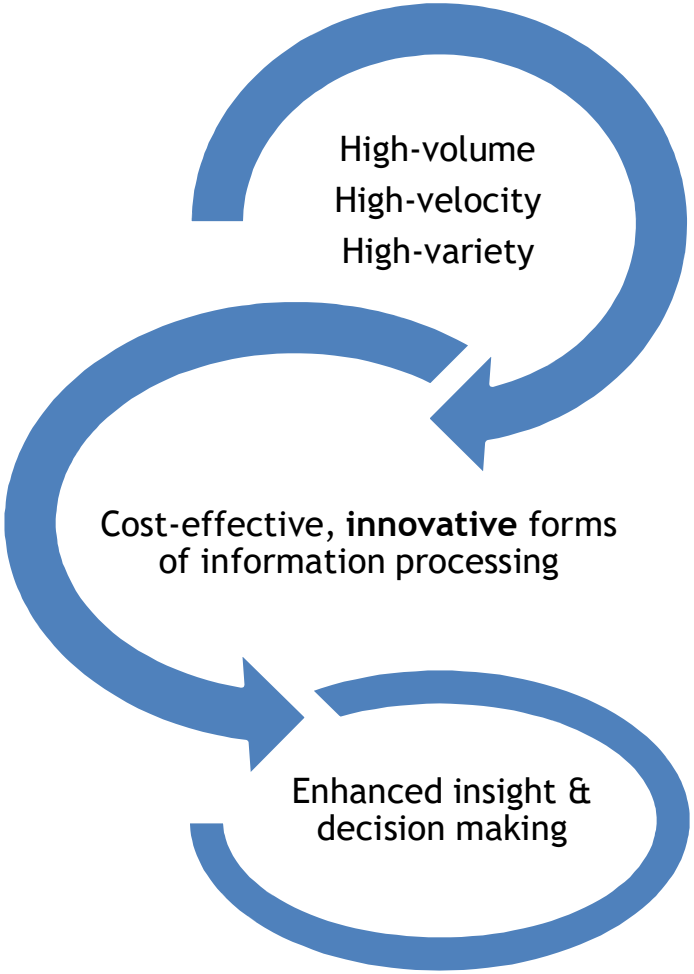
Introduction to Big Data

Agenda

- Definition of Big Data
  - ❖ Volume
  - ❖ Velocity
  - ❖ Variety
- Challenges of Big Data
- Other Characteristics of Data Which are Not Definitional Traits of Big Data
- Why Big Data?
- Traditional Business Intelligence (BI) versus Big Data
  - ❖ A Typical Data Warehouse Environment
  - ❖ A Typical Hadoop Environment
  - ❖ Coexistence of Big Data and Data Warehouse

Definition of Big Data

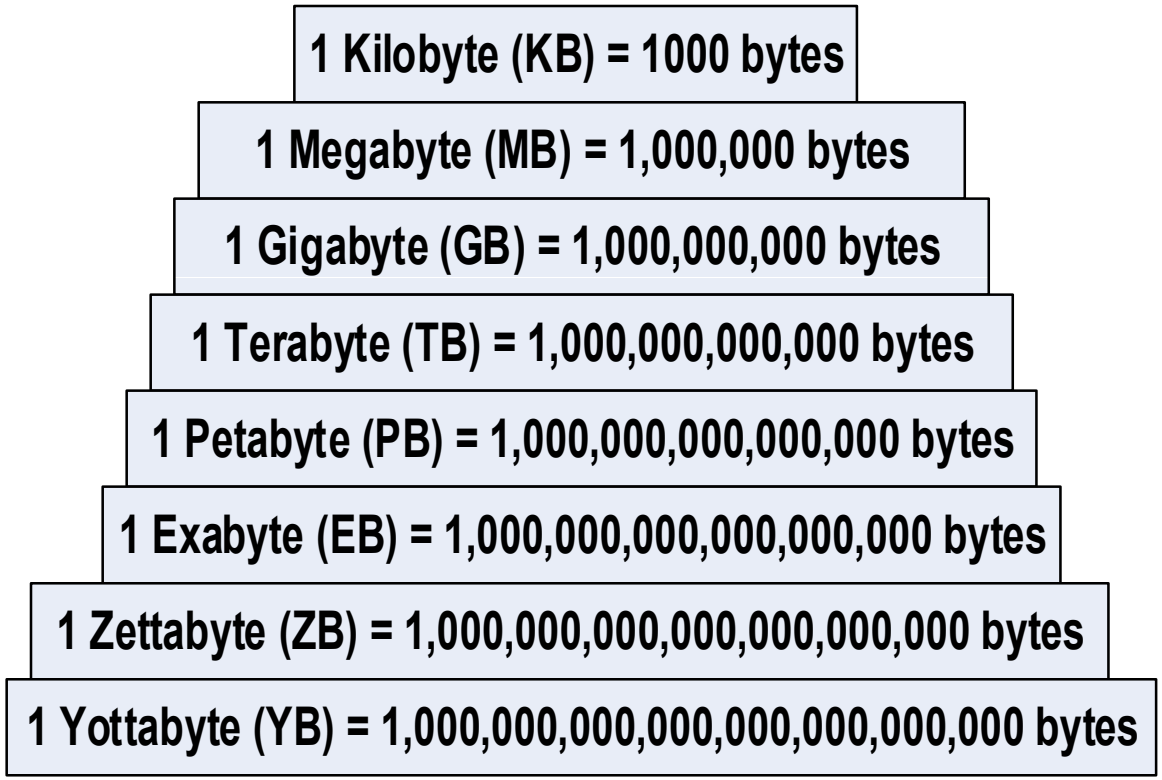
Definition of Big Data



Big Data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.

Source: Gartner IT Glossary

Volume - A Mountain of Data



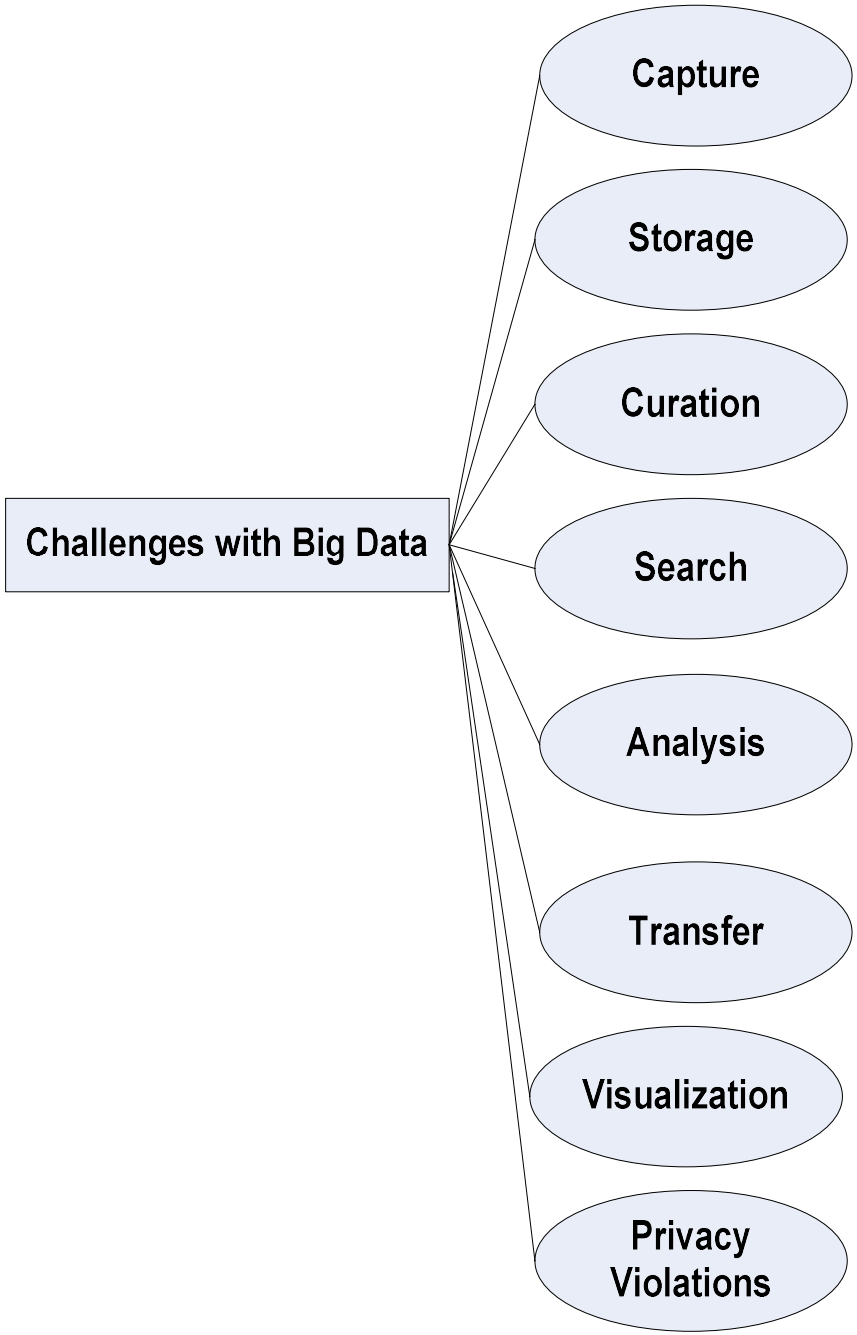
Velocity

Batch → Periodic → Near real time → Real-time processing

Variety

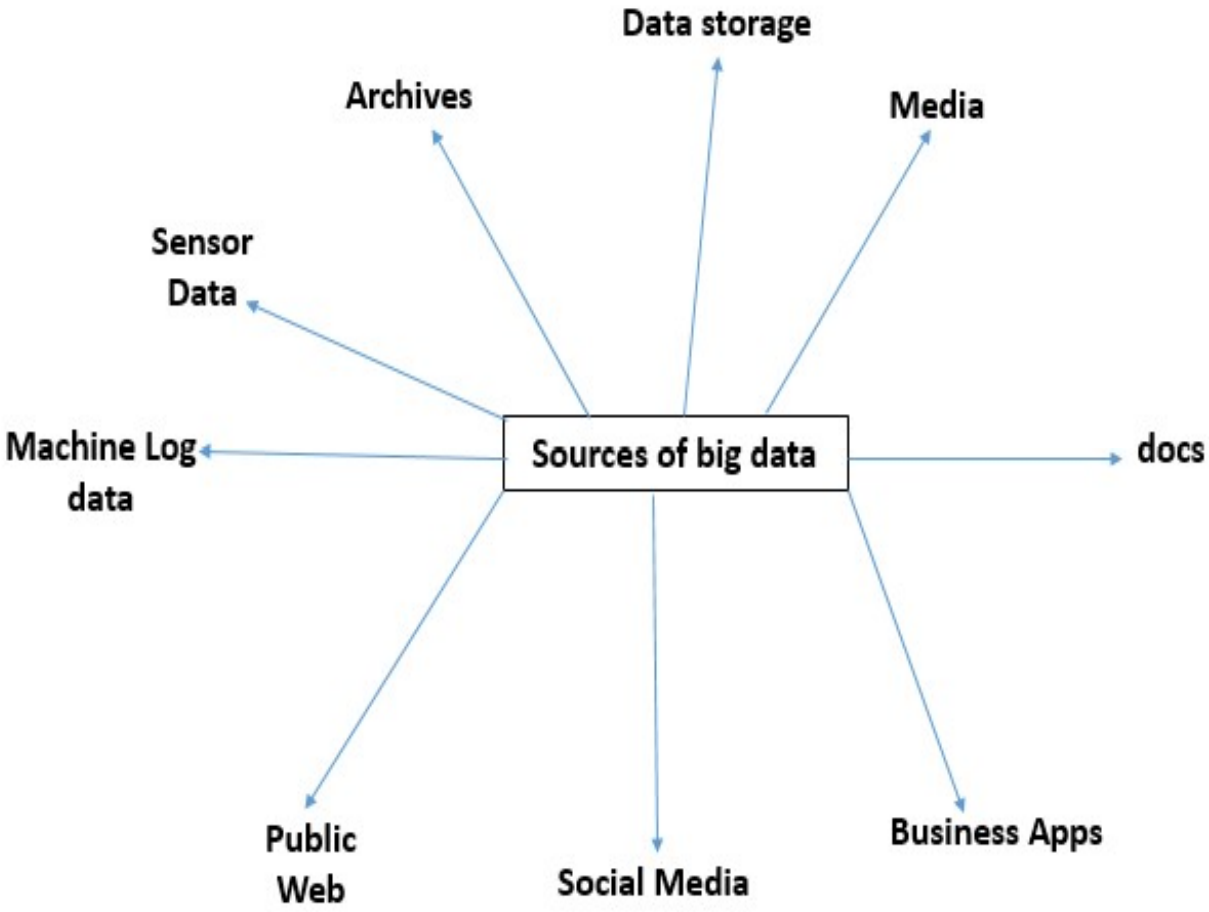
- **Structured data:** example: traditional transaction processing systems and RDBMS, etc.
- **Semi-structured data:** example: Hyper Text Markup Language (HTML), eXtensible Markup Language (XML).
- **Unstructured data:** example: unstructured text documents, audio, video, email, photos, PDFs, social media, etc.

Challenges with Big Data



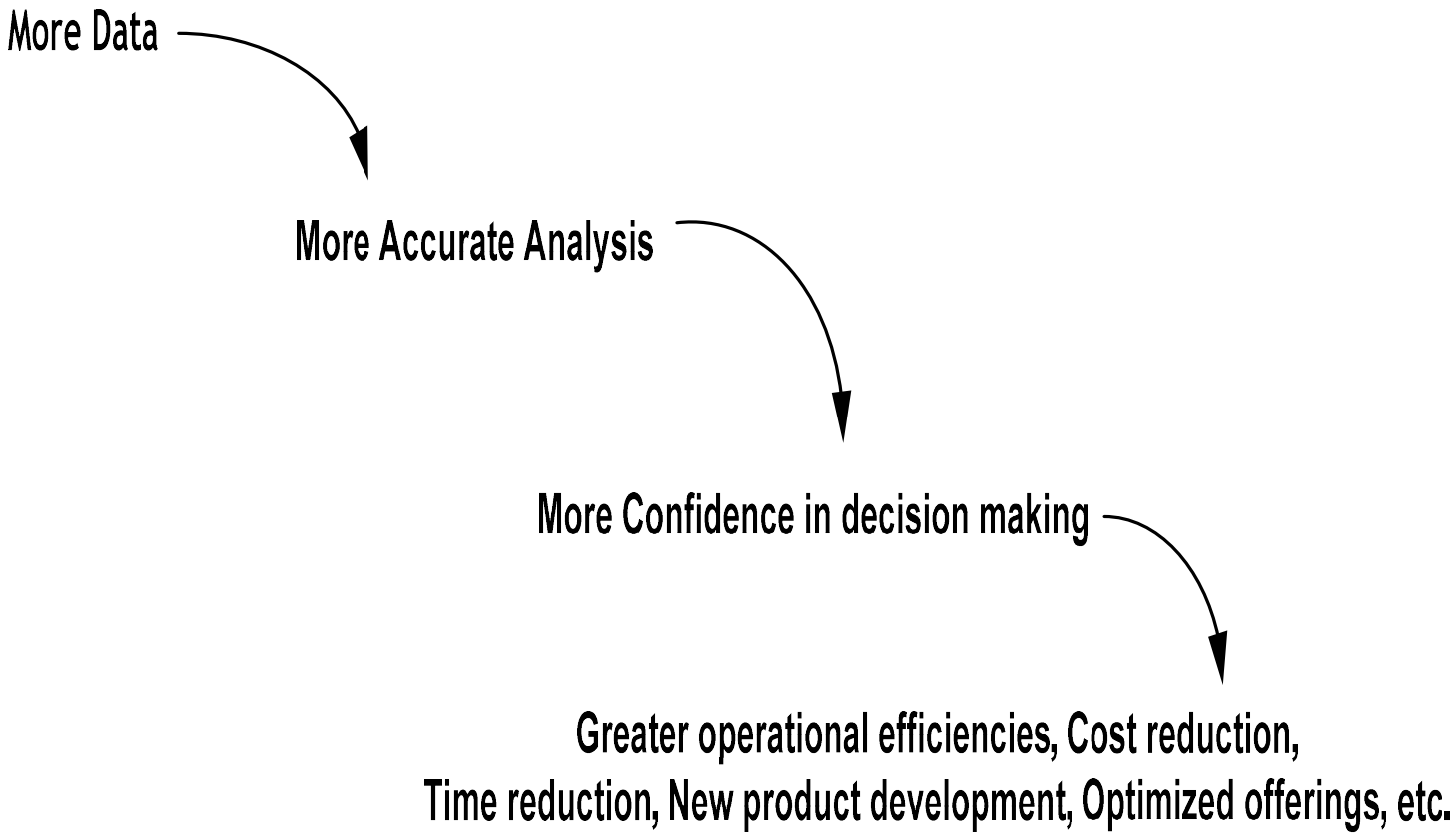
Sources of Big Data

Sources of Big Data



Why Big Data?

Why Big Data?


More Data
More Accurate Analysis
More Confidence in decision making
Greater operational efficiencies, Cost reduction,
Time reduction, New product development, Optimized offerings, etc.

1

Introduction to Big Data Analytics

Key Concepts

Big Data overview
State of the practice in analytics
Business Intelligence versus Data Science
Key roles for the new Big Data ecosystem
The Data Scientist
Examples of Big Data analytics

Much has been written about Big Data and the need for advanced analytics within industry, academia, and government. Availability of new data sources and the rise of more complex analytical opportunities have created a need to rethink existing data architectures to enable analytics that take advantage of Big Data. In addition, significant debate exists about what Big Data is and what kinds of skills are required to make best use of it. This chapter explains several key concepts to clarify what is meant by Big Data, why advanced analytics are needed, how Data Science differs from Business Intelligence (BI), and what new roles are needed for the new Big Data ecosystem.

1.1 Big Data Overview

Data is created constantly, and at an ever-increasing rate. Mobile phones, social media, imaging technologies to determine a medical diagnosis—all these and more create new data, and that must be stored somewhere for some purpose. Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time. Merely keeping up with this huge influx of data is difficult, but substantially more challenging is analyzing vast amounts of it, especially when it does not conform to traditional notions of data structure, to identify meaningful patterns and extract useful information. These challenges of the data deluge present the opportunity to transform business, government, science, and everyday life.

Several industries have led the way in developing their ability to gather and exploit data:

- Credit card companies monitor every purchase their customers make and can identify fraudulent purchases with a high degree of accuracy using rules derived by processing billions of transactions.
- Mobile phone companies analyze subscribers' calling patterns to determine, for example, whether a caller's frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.
- For companies such as LinkedIn and Facebook, data itself is their primary product. The valuations of these companies are heavily derived from the data they gather and host, which contains more and more intrinsic value as the data grows.

Three attributes stand out as defining Big Data characteristics:

- **Huge volume of data:** Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.
- **Complexity of data types and structures:** Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.
- **Speed of new data creation and growth:** Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

Although the volume of Big Data tends to attract the most attention, generally the variety and velocity of the data provide a more apt definition of Big Data. (Big Data is sometimes described as having 3 Vs: volume, variety, and velocity.) Due to its size or structure, Big Data cannot be efficiently analyzed using only traditional databases or methods. Big Data problems require new tools and technologies to store, manage, and realize the business benefit. These new tools and technologies enable creation, manipulation, and

management of large datasets and the storage environments that house them. Another definition of Big Data comes from the McKinsey Global report from 2011:

> ***Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.***
>
> McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition, and Productivity [1]

McKinsey's definition of Big Data implies that organizations will need new data architectures and analytic sandboxes, new tools, new analytical methods, and an integration of multiple skills into the new role of the data scientist, which will be discussed in Section 1.3. Figure 1-1 highlights several sources of the Big Data deluge.

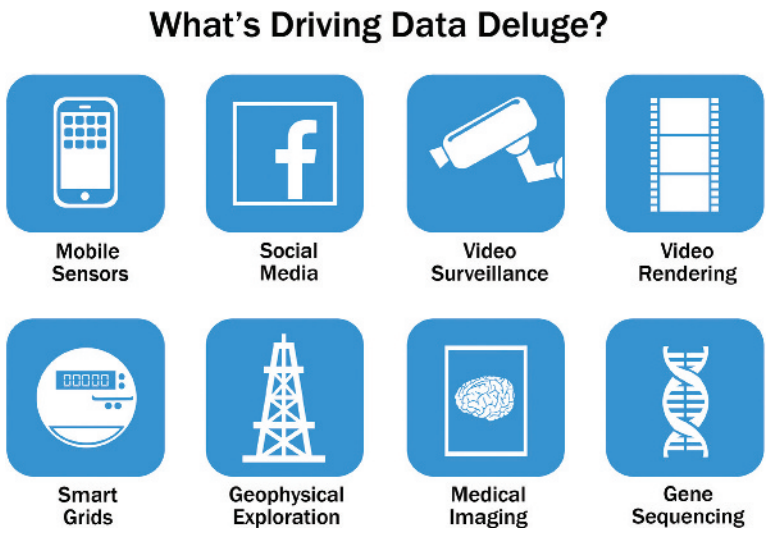


Figure 1-1 *What's driving the data deluge*

The rate of data creation is accelerating, driven by many of the items in Figure 1-1.

Social media and genetic sequencing are among the fastest-growing sources of Big Data and examples of untraditional sources of data being used for analysis.

For example, in 2012 Facebook users posted 700 status updates per second worldwide, which can be leveraged to deduce latent interests or political views of users and show relevant ads. For instance, an update in which a woman changes her relationship status from "single" to "engaged" would trigger ads on bridal dresses, wedding planning, or name-changing services.

Facebook can also construct social graphs to analyze which users are connected to each other as an interconnected network. In March 2013, Facebook released a new feature called "Graph Search," enabling users and developers to search social graphs for people with similar interests, hobbies, and shared locations.

Another example comes from genomics. Genetic sequencing and human genome mapping provide a detailed understanding of genetic makeup and lineage. The health care industry is looking toward these advances to help predict which illnesses a person is likely to get in his lifetime and take steps to avoid these maladies or reduce their impact through the use of personalized medicine and treatment. Such tests also highlight typical responses to different medications and pharmaceutical drugs, heightening risk awareness of specific drug treatments.

While data has grown, the cost to perform this work has fallen dramatically. The cost to sequence one human genome has fallen from $100 million in 2001 to $10,000 in 2011, and the cost continues to drop. Now, websites such as 23andme (Figure 1-2) offer genotyping for less than $100. Although genotyping analyzes only a fraction of a genome and does not provide as much granularity as genetic sequencing, it does point to the fact that data and complex analysis is becoming more prevalent and less expensive to deploy.

FIGURE 1-2 *Examples of what can be learned through genotyping, from 23andme.com*

As illustrated by the examples of social media and genetic sequencing, individuals and organizations both derive benefits from analysis of ever-larger and more complex datasets that require increasingly powerful analytical capabilities.

1.1.1 Data Structures

Big data can come in multiple forms, including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings. Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze. [2] Distributed computing environments and massively parallel processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data.

With this in mind, this section takes a closer look at data structures.

Figure 1-3 shows four types of data structures, with 80–90% of future data growth coming from non-structured data types. [2] Though different, the four are commonly mixed. For example, a classic Relational Database Management System (RDBMS) may store call logs for a software support call center. The RDBMS may store characteristics of the support calls as typical structured data, with attributes such as time stamps, machine type, problem type, and operating system. In addition, the system will likely have unstructured, quasi- or semi-structured data, such as free-form call log information taken from an e-mail ticket of the problem, customer chat history, or transcript of a phone call describing the technical problem and the solution or audio file of the phone call conversation. Many insights could be extracted from the unstructured, quasi- or semi-structured data in the call center data.

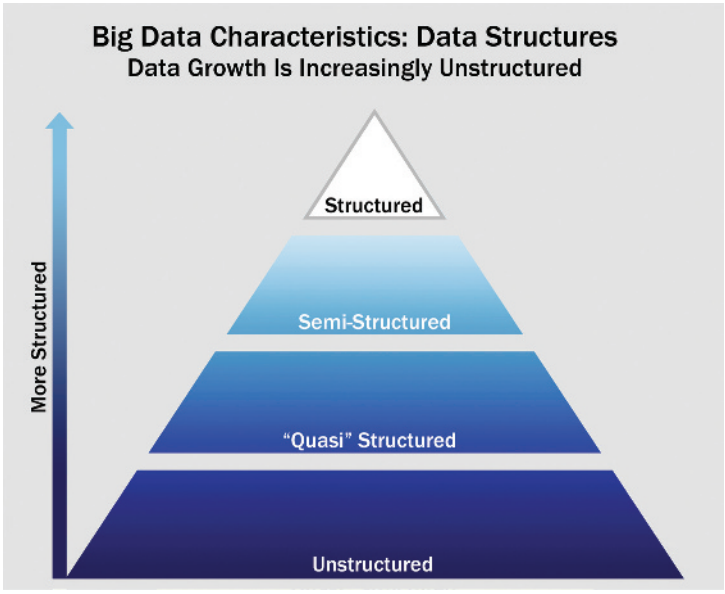


Figure 1-3 *Big Data Growth is increasingly unstructured*

Although analyzing structured data tends to be the most familiar technique, a different technique is required to meet the challenges to analyze semi-structured data (shown as XML), quasi-structured (shown as a clickstream), and unstructured data.

Here are examples of how each of the four main types of data structures may look.

- **Structured data:** Data containing a defined data type, format, and structure (that is, transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spreadsheets). See Figure 1-4.

SUMMER FOOD SERVICE PROGRAM 1]

(Data as of August 01, 2011)

Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

Figure 1-4 *Example of structured data*

- **Semi-structured data:** Textual data files with a discernible pattern that enables parsing (such as Extensible Markup Language [XML] data files that are self-describing and defined by an XML schema). See Figure 1-5.
- **Quasi-structured data:** Textual data with erratic 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). See Figure 1-6.
- **Unstructured data:** Data that has no inherent structure, which may include text documents, PDFs, images, and video. See Figure 1-7.

Quasi-structured data is a common phenomenon that bears closer scrutiny. 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, such as in the first graphic of Figure 1-5.

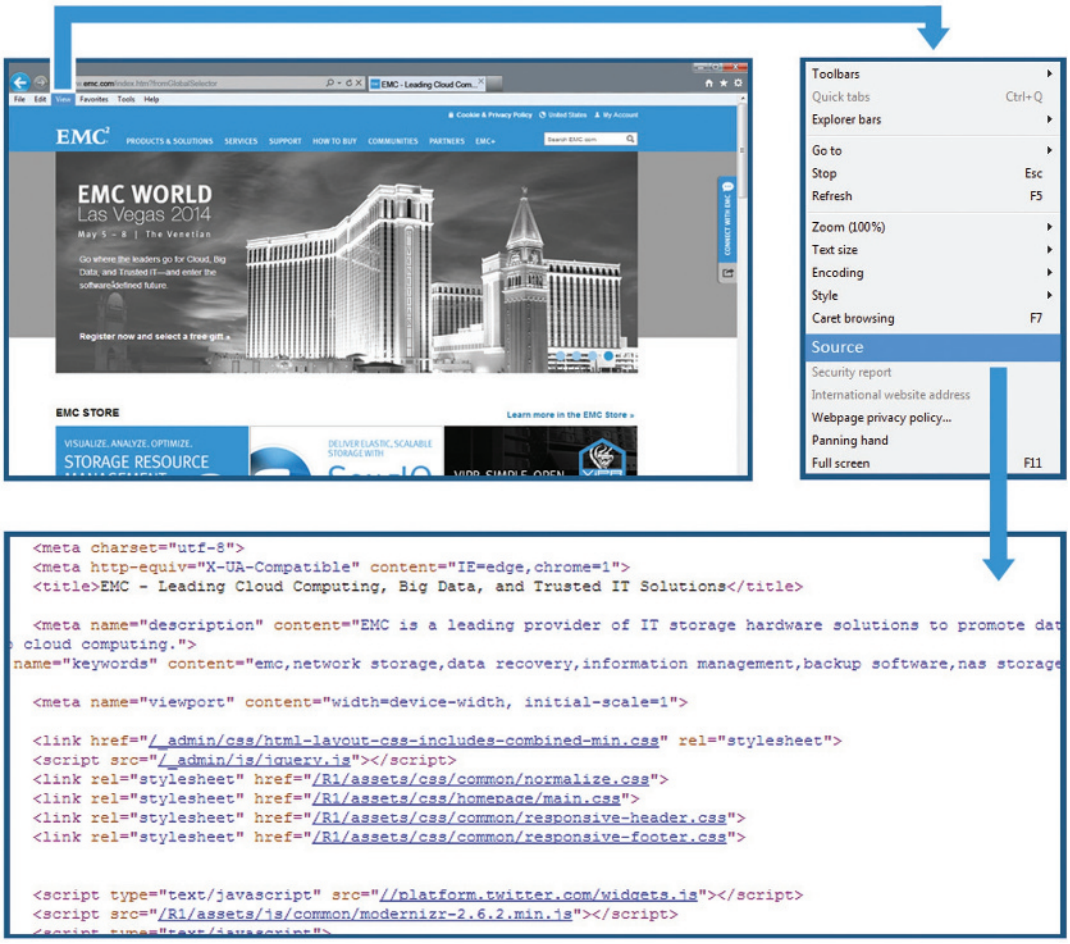


FIGURE 1-5 *Example of semi-structured data*

After doing this search, the user may choose the second link, to read more about the headline "Data Scientist—EMC Education, Training, and Certification." This brings the user to an `emc.com` site focused on this topic and a new URL, `https://education.emc.com/guest/campaign/data_science`

.aspx, that displays the page shown as (2) in Figure 1-6. Arriving at this site, the user may decide to click to learn more about the process of becoming certified in data science. The user chooses a link toward the top of the page on Certifications, bringing the user to a new URL: `https://education.emc.com/guest/certification/framework/stf/data_science.aspx`, which is (3) in Figure 1-6.

Visiting these three websites adds three URLs to the log files monitoring the user's computer or network use. These three URLs are:

```
https://www.google.com/#q=EMC+data+science
https://education.emc.com/guest/campaign/data_science.aspx
https://education.emc.com/guest/certification/framework/stf/data_
science.aspx
```

FIGURE 1-6 *Example of EMC Data Science search results*

Figure 1-7 *Example of unstructured data: video about Antarctica expedition [3]*

This set of three URLs reflects the websites and actions taken to find Data Science information related to EMC. Together, this comprises a *clickstream* that can be parsed and mined by data scientists to discover usage patterns and uncover relationships among clicks and areas of interest on a website or group of sites.

The four data types described in this chapter are sometimes generalized into two groups: structured and unstructured data. Big Data describes new kinds of data with which most organizations may not be used to working. With this in mind, the next section discusses common technology architectures from the standpoint of someone wanting to analyze Big Data.

1.1.2 Analyst Perspective on Data Repositories

The introduction of spreadsheets enabled business users to create simple logic on data structured in rows and columns and create their own analyses of business problems. Database administrator training is not required to create spreadsheets: They can be set up to do many things quickly and independently of information technology (IT) groups. Spreadsheets are easy to share, and end users have control over the logic involved. However, their proliferation can result in "many versions of the truth." In other words, it can be challenging to determine if a particular user has the most relevant version of a spreadsheet, with the most current data and logic in it. Moreover, if a laptop is lost or a file becomes corrupted, the data and logic within the spreadsheet could be lost. This is an ongoing challenge because spreadsheet programs such as Microsoft Excel still run on many computers worldwide. With the proliferation of data islands (or spreadmarts), the need to centralize the data is more pressing than ever.

As data needs grew, so did more scalable data warehousing solutions. These technologies enabled data to be managed centrally, providing benefits of security, failover, and a single repository where users

could rely on getting an "official" source of data for financial reporting or other mission-critical tasks. This structure also enabled the creation of OLAP cubes and BI analytical tools, which provided quick access to a set of dimensions within an RDBMS. More advanced features enabled performance of in-depth analytical techniques such as regressions and neural networks. Enterprise Data Warehouses (EDWs) are critical for reporting and BI tasks and solve many of the problems that proliferating spreadsheets introduce, such as which of multiple versions of a spreadsheet is correct. EDWs—and a good BI strategy—provide direct data feeds from sources that are centrally managed, backed up, and secured.

Despite the benefits of EDWs and BI, these systems tend to restrict the flexibility needed to perform robust or exploratory data analysis. With the EDW model, data is managed and controlled by IT groups and database administrators (DBAs), and data analysts must depend on IT for access and changes to the data schemas. This imposes longer lead times for analysts to get data; most of the time is spent waiting for approvals rather than starting meaningful work. Additionally, many times the EDW rules restrict analysts from building datasets. Consequently, it is common for additional systems to emerge containing critical data for constructing analytic datasets, managed locally by power users. IT groups generally dislike existence of data sources outside of their control because, unlike an EDW, these datasets are not managed, secured, or backed up. From an analyst perspective, EDW and BI solve problems related to data accuracy and availability. However, EDW and BI introduce new problems related to flexibility and agility, which were less pronounced when dealing with spreadsheets.

A solution to this problem is the analytic sandbox, which attempts to resolve the conflict for analysts and data scientists with EDW and more formally managed corporate data. In this model, the IT group may still manage the analytic sandboxes, but they will be purposefully designed to enable robust analytics, while being centrally managed and secured. These sandboxes, often referred to as *workspaces,* are designed to enable teams to explore many datasets in a controlled fashion and are not typically used for enterprise-level financial reporting and sales dashboards.

Many times, analytic sandboxes enable high-performance computing using in-database processing—the analytics occur within the database itself. The idea is that performance of the analysis will be better if the analytics are run in the database itself, rather than bringing the data to an analytical tool that resides somewhere else. In-database analytics, discussed further in Chapter 11, "Advanced Analytics—Technology and Tools: In-Database Analytics," creates relationships to multiple data sources within an organization and saves time spent creating these data feeds on an individual basis. In-database processing for deep analytics enables faster turnaround time for developing and executing new analytic models, while reducing, though not eliminating, the cost associated with data stored in local, "shadow" file systems. In addition, rather than the typical structured data in the EDW, analytic sandboxes can house a greater variety of data, such as raw data, textual data, and other kinds of unstructured data, without interfering with critical production databases. Table 1-1 summarizes the characteristics of the data repositories mentioned in this section.

TABLE 1-1 *Types of Data Repositories, from an Analyst Perspective*

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.

Data Warehouses	Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

There are several things to consider with Big Data Analytics projects to ensure the approach fits with the desired goals. Due to the characteristics of Big Data, these projects lend themselves to decision support for high-value, strategic decision making with high processing complexity. The analytic techniques used in this context need to be iterative and flexible, due to the high volume of data and its complexity. Performing rapid and complex analysis requires high throughput network connections and a consideration for the acceptable amount of latency. For instance, developing a real-time product recommender for a website imposes greater system demands than developing a near-real-time recommender, which may still provide acceptable performance, have slightly greater latency, and may be cheaper to deploy. These considerations require a different approach to thinking about analytics challenges, which will be explored further in the next section.

1.2 State of the Practice in Analytics

Current business problems provide many opportunities for organizations to become more analytical and data driven, as shown in Table 1-2.

Table 1-2 *Business Drivers for Advanced Analytics*

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

Table 1-2 outlines four categories of common business problems that organizations contend with where they have an opportunity to leverage advanced analytics to create competitive advantage. Rather than only performing standard reporting on these areas, organizations can apply advanced analytical techniques to optimize processes and derive more value from these common tasks. The first three examples do not represent new problems. Organizations have been trying to reduce customer churn, increase sales, and cross-sell customers for many years. What is new is the opportunity to fuse advanced analytical techniques with Big Data to produce more impactful analyses for these traditional problems. The last example portrays emerging regulatory requirements. Many compliance and regulatory laws have been in existence for decades, but additional requirements are added every year, which represent additional complexity and data requirements for organizations. Laws related to anti-money laundering (AML) and fraud prevention require advanced analytical techniques to comply with and manage properly.

1.2.1 BI Versus Data Science

The four business drivers shown in Table 1-2 require a variety of analytical techniques to address them properly. Although much is written generally about analytics, it is important to distinguish between BI and Data Science. As shown in Figure 1-8, there are several ways to compare these groups of analytical techniques.

One way to evaluate the type of analysis being performed is to examine the time horizon and the kind of analytical approaches being used. BI tends to provide reports, dashboards, and queries on business questions for the current period or in the past. BI systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior quarter or year. These questions tend to be closed-ended and explain current or past behavior, typically by aggregating historical data and grouping it in some way. BI provides hindsight and some insight and generally answers questions related to "when" and "where" events occurred.

By comparison, Data Science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analyzing the present and enabling informed decisions about the future. Rather than aggregating historical data to look at how many of a given product sold in the previous quarter, a team may employ Data Science techniques such as time series analysis, further discussed in Chapter 8, "Advanced Analytical Theory and Methods: Time Series Analysis," to forecast future product sales and revenue more accurately than extending a simple trend line. In addition, Data Science tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions. This approach provides insight into current activity and foresight into future events, while generally focusing on questions related to "how" and "why" events occur.

Where BI problems tend to require highly structured data organized in rows and columns for accurate reporting, Data Science projects tend to use many types of data sources, including large or unconventional datasets. Depending on an organization's goals, it may choose to embark on a BI project if it is doing reporting, creating dashboards, or performing simple visualizations, or it may choose Data Science projects if it needs to do a more sophisticated analysis with disaggregated or varied datasets.

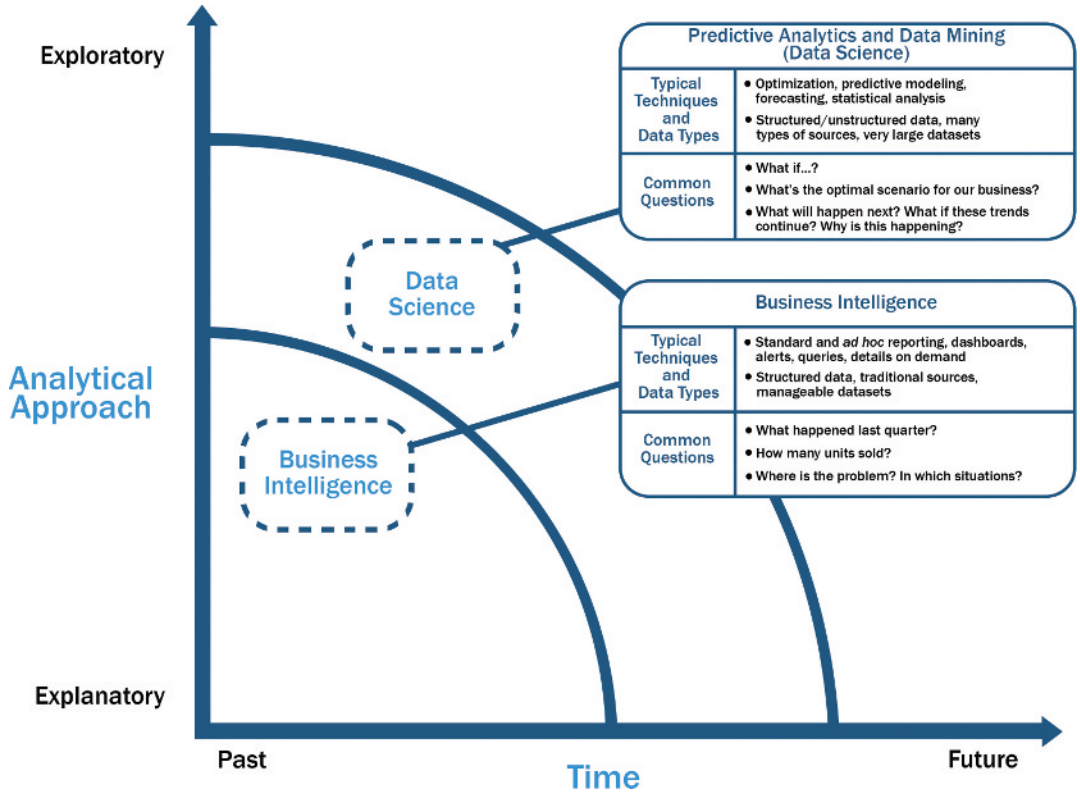


FIGURE 1-8 *Comparing BI with Data Science*

1.2.2 Current Analytical Architecture

As described earlier, Data Science projects need workspaces that are purpose-built for experimenting with data, with flexible and agile data architectures. Most organizations still have data warehouses that provide excellent support for traditional reporting and simple data analysis activities but unfortunately have a more difficult time supporting more robust analyses. This section examines a typical analytical data architecture that may exist within an organization.

Figure 1-9 shows a typical data architecture and several of the challenges it presents to data scientists and others trying to do advanced analytics. This section examines the data flow to the Data Scientist and how this individual fits into the process of getting data to analyze on projects.

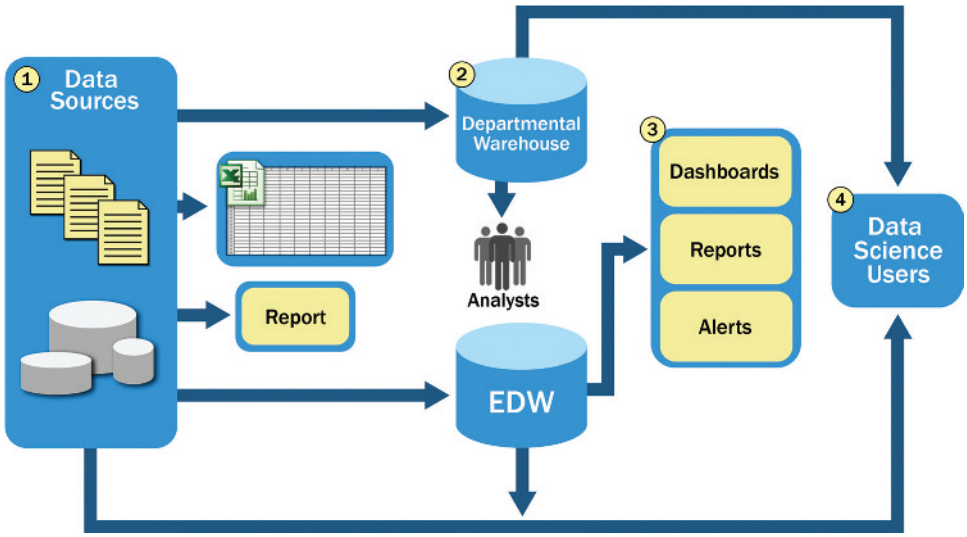


FIGURE 1-9 *Typical analytic architecture*

1. For data sources to be loaded into the data warehouse, data needs to be well understood, structured, and normalized with the appropriate data type definitions. Although this kind of centralization enables security, backup, and failover of highly critical data, it also means that data typically must go through significant preprocessing and checkpoints before it can enter this sort of controlled environment, which does not lend itself to data exploration and iterative analytics.
2. As a result of this level of control on the EDW, additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. These local data marts may not have the same constraints for security and structure as the main EDW and allow users to do some level of more in-depth analysis. However, these one-off systems reside in isolation, often are not synchronized or integrated with other data stores, and may not be backed up.
3. Once in the data warehouse, data is read by additional applications across the enterprise for BI and reporting purposes. These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.
4. At the end of this workflow, analysts get data provisioned for their downstream analytics. Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools. Many times these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset. Because these analyses are based on data extracts, they reside in a separate location, and the results of the analysis—and any insights on the quality of the data or anomalies—rarely are fed back into the main data repository.

Because new data sources slowly accumulate in the EDW due to the rigorous validation and data structuring process, data is slow to move into the EDW, and the data schema is slow to change.

Departmental data warehouses may have been originally designed for a specific purpose and set of business needs, but over time evolved to house more and more data, some of which may be forced into existing schemas to enable BI and the creation of OLAP cubes for analysis and reporting. Although the EDW achieves the objective of reporting and sometimes the creation of dashboards, EDWs generally limit the ability of analysts to iterate on the data in a separate nonproduction environment where they can conduct in-depth analytics or perform analysis on unstructured data.

The typical data architectures just described are designed for storing and processing mission-critical data, supporting enterprise applications, and enabling corporate reporting activities. Although reports and dashboards are still important for organizations, most traditional data architectures inhibit data exploration and more sophisticated analysis. Moreover, traditional data architectures have several additional implications for data scientists.

- High-value data is hard to reach and leverage, and predictive analytics and data mining activities are last in line for data. Because the EDWs are designed for central data management and reporting, those wanting data for analysis are generally prioritized after operational processes.
- Data moves in batches from EDW to local analytical tools. This workflow means that data scientists are limited to performing in-memory analytics (such as with R, SAS, SPSS, or Excel), which will restrict the size of the datasets they can use. As such, analysis may be subject to constraints of sampling, which can skew model accuracy.
- Data Science projects will remain isolated and ad hoc, rather than centrally managed. The implication of this isolation is that the organization can never harness the power of advanced analytics in a scalable way, and Data Science projects will exist as nonstandard initiatives, which are frequently not aligned with corporate business goals or strategy.

All these symptoms of the traditional data architecture result in a slow "time-to-insight" and lower business impact than could be achieved if the data were more readily accessible and supported by an environment that promoted advanced analytics. As stated earlier, one solution to this problem is to introduce analytic sandboxes to enable data scientists to perform advanced analytics in a controlled and sanctioned way. Meanwhile, the current Data Warehousing solutions continue offering reporting and BI services to support management and mission-critical operations.

1.2.3 Drivers of Big Data

To better understand the market drivers related to Big Data, it is helpful to first understand some past history of data stores and the kinds of repositories and tools to manage these data stores.

As shown in Figure 1-10, in the 1990s the volume of information was often measured in terabytes. Most organizations analyzed structured data in rows and columns and used relational databases and data warehouses to manage large stores of enterprise information. The following decade saw a proliferation of different kinds of data sources—mainly productivity and publishing tools such as content management repositories and networked attached storage systems—to manage this kind of information, and the data began to increase in size and started to be measured at petabyte scales. In the 2010s, the information that organizations try to manage has broadened to include many other kinds of data. In this era, everyone and everything is leaving a digital footprint. Figure 1-10 shows a summary perspective on sources of Big Data generated by new applications and the scale and growth rate of the data. These applications, which generate data volumes that can be measured in exabyte scale, provide opportunities for new analytics and driving new value for organizations. The data now comes from multiple sources, such as these:

- Medical information, such as genomic sequencing and diagnostic imaging
- Photos and video footage uploaded to the World Wide Web
- Video surveillance, such as the thousands of video cameras spread across a city
- Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones
- Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures
- Nontraditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing

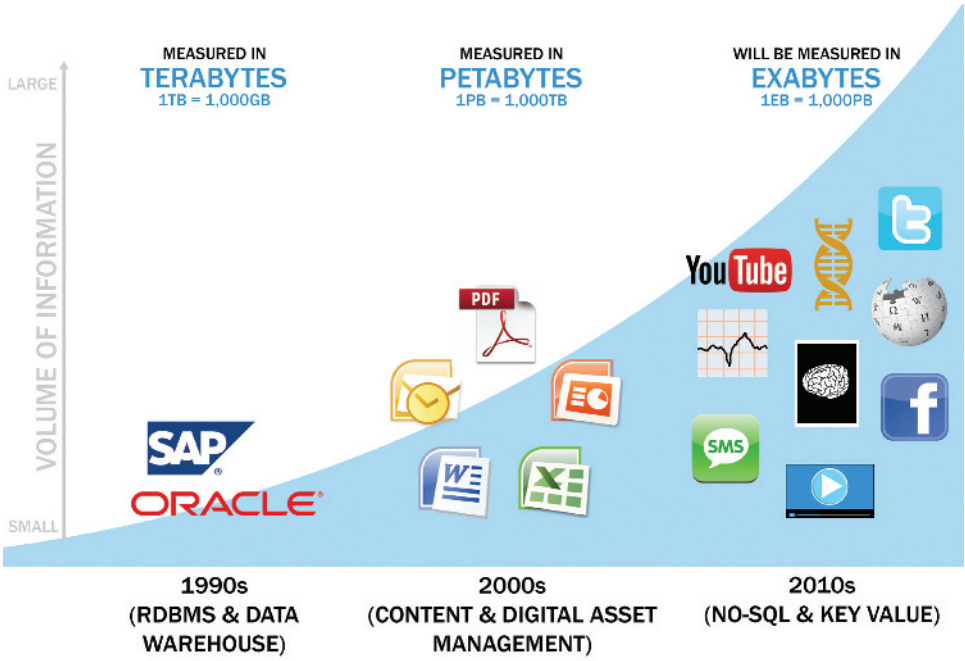


FIGURE 1-10 *Data evolution and the rise of Big Data sources*

The Big Data trend is generating an enormous amount of information from many new sources. This data deluge requires advanced analytics and new market players to take advantage of these opportunities and new market dynamics, which will be discussed in the following section.

1.2.4 Emerging Big Data Ecosystem and a New Approach to Analytics

Organizations and data collectors are realizing that the data they can gather from individuals contains intrinsic value and, as a result, a new economy is emerging. As this new digital economy continues to

evolve, the market sees the introduction of data vendors and data cleaners that use crowdsourcing (such as Mechanical Turk and GalaxyZoo) to test the outcomes of machine learning techniques. Other vendors offer added value by repackaging open source tools in a simpler way and bringing the tools to market. Vendors such as Cloudera, Hortonworks, and Pivotal have provided this value-add for the open source framework Hadoop.

As the new ecosystem takes shape, there are four main groups of players within this interconnected web. These are shown in Figure 1-11.

- **Data devices** [shown in the (1) section of Figure 1-11] and the "Sensornet" gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new data created, an additional petabyte of data is created about that data. [2]
    - For example, consider someone playing an online video game through a PC, game console, or smartphone. In this case, the video game provider captures data about the skill and levels attained by the player. Intelligent systems monitor and log how and when the user plays the game. As a consequence, the game provider can fine-tune the difficulty of the game, suggest other related games that would most likely interest the user, and offer additional equipment and enhancements for the character based on the user's age, gender, and interests. This information may get stored locally or uploaded to the game provider's cloud to analyze the gaming habits and opportunities for upsell and cross-sell, and identify archetypical profiles of specific kinds of users.
    - Smartphones provide another rich source of data. In addition to messaging and basic phone usage, they store and transmit data about Internet usage, SMS usage, and real-time location. This metadata can be used for analyzing traffic patterns by scanning the density of smartphones in locations to track the speed of cars or the relative traffic congestion on busy roads. In this way, GPS devices in cars can give drivers real-time updates and offer alternative routes to avoid traffic delays.
    - Retail shopping loyalty cards record not just the amount an individual spends, but the locations of stores that person visits, the kinds of products purchased, the stores where goods are purchased most often, and the combinations of products purchased together. Collecting this data provides insights into shopping and travel habits and the likelihood of successful advertisement targeting for certain types of retail promotions.
- **Data collectors** [the blue ovals, identified as (2) within Figure 1-11] include sample entities that collect data from the device and users.
    - Data results from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is willing to pay for premium TV content
    - Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can gauge which products get the most foot traffic using geospatial data collected from the RFID chips
- **Data aggregators** (the dark gray ovals in Figure 1-11, marked as (3)) make sense of the data collected from the various entities from the "SensorNet" or the "Internet of Things." These organizations compile data from the devices and usage patterns collected by government agencies, retail stores,

and websites. In turn, they can choose to transform and package the data as products to sell to list brokers, who may want to generate marketing lists of people who may be good targets for specific ad campaigns.

- **Data users and buyers** are denoted by (4) in Figure 1-11. These groups directly benefit from the data collected and aggregated by others within the data value chain.
    - Retail banks, acting as a data buyer, may want to know which customers have the highest likelihood to apply for a second mortgage or a home equity line of credit. To provide input for this analysis, retail banks may purchase data from a data aggregator. This kind of data may include demographic information about people living in specific locations; people who appear to have a specific level of debt, yet still have solid credit scores (or other characteristics such as paying bills on time and having savings accounts) that can be used to infer credit worthiness; and those who are searching the web for information about paying off debts or doing home remodeling projects. Obtaining data from these various sources and aggregators will enable a more targeted marketing campaign, which would have been more challenging before Big Data due to the lack of information or high-performing technologies.
    - Using technologies such as Hadoop to perform natural language processing on unstructured, textual data from social media websites, users can gauge the reaction to events such as presidential campaigns. People may, for example, want to determine public sentiments toward a candidate by analyzing related blogs and online comments. Similarly, data users may want to track and prepare for natural disasters by identifying which areas a hurricane affects first and how it moves, based on which geographic areas are tweeting about it or discussing it via social media.

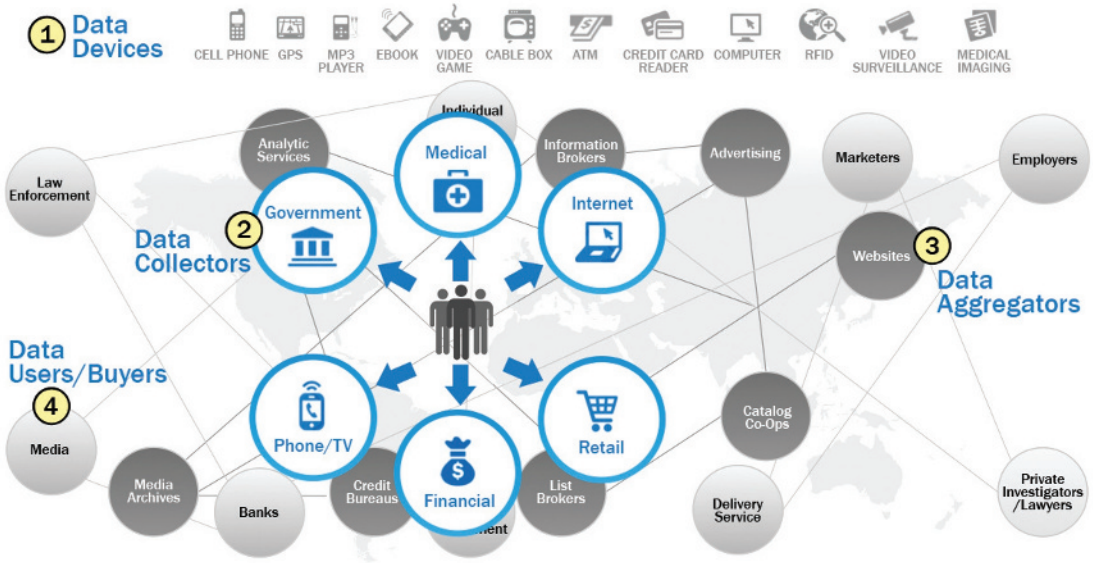


FIGURE 1-11 *Emerging Big Data ecosystem*

As illustrated by this emerging Big Data ecosystem, the kinds of data and the related market dynamics vary greatly. These datasets can include sensor data, text, structured datasets, and social media. With this in mind, it is worth recalling that these datasets will not work well within traditional EDWs, which were architected to streamline reporting and dashboards and be centrally managed. Instead, Big Data problems and projects require different approaches to succeed.

Analysts need to partner with IT and DBAs to get the data they need within an analytic sandbox. A typical analytical sandbox contains raw data, aggregated data, and data with multiple kinds of structure. The sandbox enables robust exploration of data and requires a savvy user to leverage and take advantage of data in the sandbox environment.

1.3 Key Roles for the New Big Data Ecosystem

As explained in the context of the Big Data ecosystem in Section 1.2.4, new players have emerged to curate, store, produce, clean, and transact data. In addition, the need for applying more advanced analytical techniques to increasingly complex business problems has driven the emergence of new roles, new technology platforms, and new analytical methods. This section explores the new roles that address these needs, and subsequent chapters explore some of the analytical methods and technology platforms.

The Big Data ecosystem demands three categories of roles, as shown in Figure 1-12. These roles were described in the McKinsey Global study on Big Data, from May 2011 [1].

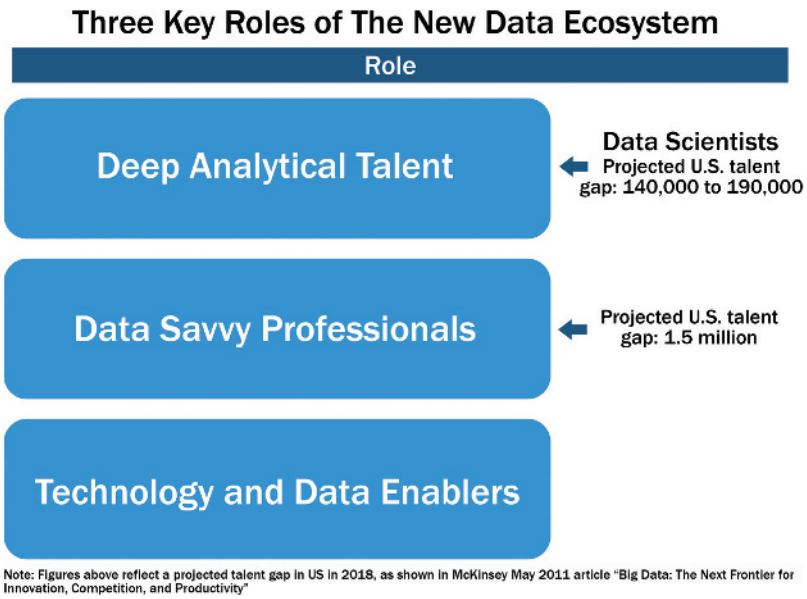


Figure 1-12 *Key roles of the new Big Data ecosystem*

The first group—Deep Analytical Talent— is technically savvy, with strong analytical skills. Members possess a combination of skills to handle raw, unstructured data and to apply complex analytical techniques at

massive scales. This group has advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning. To do their jobs, members need access to a robust analytic sandbox or workspace where they can perform large-scale analytical data experiments. Examples of current professions fitting into this group include statisticians, economists, mathematicians, and the new role of the Data Scientist.

The McKinsey study forecasts that by the year 2018, the United States will have a talent gap of 140,000–190,000 people with deep analytical talent. This does not represent the number of people needed with deep analytical talent; rather, this range represents the difference between what will be available in the workforce compared with what will be needed. In addition, these estimates only reflect forecasted talent shortages in the United States; the number would be much larger on a global basis.

The second group—Data Savvy Professionals—has less technical depth but has a basic knowledge of statistics or machine learning and can define key questions that can be answered using advanced analytics. These people tend to have a base knowledge of working with data, or an appreciation for some of the work being performed by data scientists and others with deep analytical talent. Examples of data savvy professionals include financial analysts, market research analysts, life scientists, operations managers, and business and functional managers.

The McKinsey study forecasts the projected U.S. talent gap for this group to be 1.5 million people by the year 2018. At a high level, this means for every Data Scientist profile needed, the gap will be ten times as large for Data Savvy Professionals. Moving toward becoming a data savvy professional is a critical step in broadening the perspective of managers, directors, and leaders, as this provides an idea of the kinds of questions that can be solved with data.

The third category of people mentioned in the study is Technology and Data Enablers. This group represents people providing technical expertise to support analytical projects, such as provisioning and administrating analytical sandboxes, and managing large-scale data architectures that enable widespread analytics within companies and other organizations. This role requires skills related to computer engineering, programming, and database administration.

These three groups must work together closely to solve complex Big Data challenges. Most organizations are familiar with people in the latter two groups mentioned, but the first group, Deep Analytical Talent, tends to be the newest role for most and the least understood. For simplicity, this discussion focuses on the emerging role of the Data Scientist. It describes the kinds of activities that role performs and provides a more detailed view of the skills needed to fulfill that role.

There are three recurring sets of activities that data scientists perform:

- **Reframe business challenges as analytics challenges.** Specifically, this is a skill to diagnose business problems, consider the core of a given problem, and determine which kinds of candidate analytical methods can be applied to solve it. This concept is explored further in Chapter 2, "Data Analytics Lifecycle."
- **Design, implement, and deploy statistical models and data mining techniques on Big Data.** This set of activities is mainly what people think about when they consider the role of the Data Scientist:

namely, applying complex or advanced analytical methods to a variety of business problems using data. Chapter 3 through Chapter 11 of this book introduces the reader to many of the most popular analytical techniques and tools in this area.

- **Develop insights that lead to actionable recommendations.** It is critical to note that applying advanced methods to data problems does not necessarily drive new business value. Instead, it is important to learn how to draw insights out of the data and communicate them effectively. Chapter 12, "The Endgame, or Putting It All Together," has a brief overview of techniques for doing this.

Data scientists are generally thought of as having five main sets of skills and behavioral characteristics, as shown in Figure 1-13:

- **Quantitative skill:** such as mathematics or statistics
- **Technical aptitude:** namely, software engineering, machine learning, and programming skills
- **Skeptical mind-set and critical thinking:** It is important that data scientists can examine their work critically rather than in a one-sided way.
- **Curious and creative**: Data scientists are passionate about data and finding creative ways to solve problems and portray information.
- **Communicative and collaborative:** Data scientists must be able to articulate the business value in a clear way and collaboratively work with other groups, including project sponsors and key stakeholders.

FIGURE 1-13 *Profile of a Data Scientist*

Data scientists are generally comfortable using this blend of skills to acquire, manage, analyze, and visualize data and tell compelling stories about it. The next section includes examples of what Data Science teams have created to drive new value or innovation with Big Data.

1.4 Examples of Big Data Analytics

After describing the emerging Big Data ecosystem and new roles needed to support its growth, this section provides three examples of Big Data Analytics in different areas: retail, IT infrastructure, and social media.

As mentioned earlier, Big Data presents many opportunities to improve sales and marketing analytics. An example of this is the U.S. retailer Target. Charles Duhigg's book *The Power of Habit* [4] discusses how Target used Big Data and advanced analytical methods to drive new revenue. After analyzing consumer-purchasing behavior, Target's statisticians determined that the retailer made a great deal of money from three main life-event situations.

- Marriage, when people tend to buy many new products
- Divorce, when people buy new products and change their spending habits
- Pregnancy, when people have many new things to buy and have an urgency to buy them

Target determined that the most lucrative of these life-events is the third situation: pregnancy. Using data collected from shoppers, Target was able to identify this fact and predict which of its shoppers were pregnant. In one case, Target knew a female shopper was pregnant even before her family knew [5]. This kind of knowledge allowed Target to offer specific coupons and incentives to their pregnant shoppers. In fact, Target could not only determine if a shopper was pregnant, but in which month of pregnancy a shopper may be. This enabled Target to manage its inventory, knowing that there would be demand for specific products and it would likely vary by month over the coming nine- to ten-month cycles.

Hadoop [6] represents another example of Big Data innovation on the IT infrastructure. Apache Hadoop is an open source framework that allows companies to process vast amounts of information in a highly parallelized way. Hadoop represents a specific implementation of the MapReduce paradigm and was designed by Doug Cutting and Mike Cafarella in 2005 to use data with varying structures. It is an ideal technical framework for many Big Data projects, which rely on large or unwieldy datasets with unconventional data structures. One of the main benefits of Hadoop is that it employs a distributed file system, meaning it can use a distributed cluster of servers and commodity hardware to process large amounts of data. Some of the most common examples of Hadoop implementations are in the social media space, where Hadoop can manage transactions, give textual updates, and develop social graphs among millions of users. Twitter and Facebook generate massive amounts of unstructured data and use Hadoop and its ecosystem of tools to manage this high volume. Hadoop and its ecosystem are covered in Chapter 10, "Advanced Analytics—Technology and Tools: MapReduce and Hadoop."

Finally, social media represents a tremendous opportunity to leverage social and professional interactions to derive new insights. LinkedIn exemplifies a company in which data itself is the product. Early on, LinkedIn founder Reid Hoffman saw the opportunity to create a social network for working professionals.

As of 2014, LinkedIn has more than 250 million user accounts and has added many additional features and data-related products, such as recruiting, job seeker tools, advertising, and InMaps, which show a social graph of a user's professional network. Figure 1-14 is an example of an InMap visualization that enables a LinkedIn user to get a broader view of the interconnectedness of his contacts and understand how he knows most of them.

Figure 1-14 *Data visualization of a user's social network using InMaps*

Summary

Big Data comes from myriad sources, including social media, sensors, the Internet of Things, video surveillance, and many sources of data that may not have been considered data even a few years ago. As businesses struggle to keep up with changing market requirements, some companies are finding creative ways to apply Big Data to their growing business needs and increasingly complex problems. As organizations evolve their processes and see the opportunities that Big Data can provide, they try to move beyond traditional BI activities, such as using data to populate reports and dashboards, and move toward Data Science- driven projects that attempt to answer more open-ended and complex questions.

However, exploiting the opportunities that Big Data presents requires new data architectures, including analytic sandboxes, new ways of working, and people with new skill sets. These drivers are causing organizations to set up analytic sandboxes and build Data Science teams. Although some organizations are fortunate to have data scientists, most are not, because there is a growing talent gap that makes finding and hiring data scientists in a timely manner difficult. Still, organizations such as those in web retail, health care, genomics, new IT infrastructures, and social media are beginning to take advantage of Big Data and apply it in creative and novel ways.

Exercises

1. What are the three characteristics of Big Data, and what are the main considerations in processing Big Data?
2. What is an analytic sandbox, and why is it important?
3. Explain the differences between BI and Data Science.
4. Describe the challenges of the current analytical architecture for data scientists.
5. What are the key skill sets and behavioral characteristics of a data scientist?

Bibliography

[1] C. B. B. D. Manyika, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011.

[2] D. R. John Gantz, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC, 2013.

[3] `http://www.willisresilience.com/emc-datalab` [Online].

[4] C. Duhigg, *The Power of Habit: Why We Do What We Do in Life and Business*, New York: Random House, 2012.

[5] K. Hill, "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did," Forbes, February 2012.

[6] `http://hadoop.apache.org` [Online].

7.1 INTRODUCTION

MY TIP
Parameter ⇔ Population
Statistic ⇔ Sample

In the previous three chapters, you have learned a lot about probability distributions, such as the binomial and normal distributions. The shape of the normal distribution is determined by its mean μ and its standard deviation σ, whereas the shape of the binomial distribution is determined by p. These numerical descriptive measures—called **parameters**—are needed to calculate the probability of observing sample results.

In practical situations, you may be able to decide which *type* of probability distribution to use as a model, but the values of the *parameters* that specify its *exact form* are unknown. Here are two examples:

- A pollster is sure that the responses to his "agree/disagree" questions will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and standard deviation σ of the yields are unknown.

In these cases, you must rely on the *sample* to learn about these parameters. The proportion of those who "agree" in the pollster's sample provides information about the actual value of p. The mean and standard deviation of the agronomist's sample approximate the actual values of μ and σ. If you want the sample to provide *reliable information* about the population, however, you must select your sample in a certain way!

7.2 SAMPLING PLANS AND EXPERIMENTAL DESIGNS

The way a sample is selected is called the **sampling plan** or **experimental design** and determines the quantity of information in the sample. Knowing the sampling plan used in a particular situation will often allow you to measure the reliability or goodness of your inference.

Simple random sampling is a commonly used sampling plan in which every sample of size n has the same chance of being selected. For example, suppose you want to select a sample of size $n = 2$ from a population containing $N = 4$ objects. If the four objects are identified by the symbols x_1, x_2, x_3, and x_4, there are six distinct pairs that could be selected, as listed in Table 7.1. If the sample of $n = 2$ observations is selected so that each of these six samples has the same chance of selection, given by 1/6, then the resulting sample is called a **simple random sample,** or just a **random sample.**

TABLE 7.1 Ways of Selecting a Sample of Size 2 from 4 Objects

Sample	Observations in Sample
1	x_1, x_2
2	x_1, x_3
3	x_1, x_4
4	x_2, x_3
5	x_2, x_4
6	x_3, x_4

Definition If a sample of n elements is selected from a population of N elements using a sampling plan in which each of the possible samples has the same chance of selection, then the sampling is said to be **random** and the resulting sample is a **simple random sample.**

Perfect random sampling is difficult to achieve in practice. If the size of the population N is small, you might write each of N numbers on a poker chip, mix the chips, and select a sample of n chips. The numbers that you select correspond to the n measurements that appear in the sample. Since this method is not always very practical, a simpler and more reliable method uses **random numbers**—digits generated so that the values 0 to 9 occur randomly and with equal frequency. These numbers can be generated by computer or may even be available on your scientific calculator. Alternatively, Table 10 in Appendix I is a table of random numbers that you can use to select a *random sample.*

EXAMPLE 7.1

A computer database at a downtown law firm contains files for $N = 1000$ clients. The firm wants to select $n = 5$ files for review. Select a simple random sample of 5 files from this database.

Solution You must first label each file with a number from 1 to 1000. Perhaps the files are stored alphabetically, and the computer has already assigned a number to each. Then generate a sequence of 10 three-digit random numbers. If you are using Table 10 of Appendix I, select a random starting point and use a portion of the table similar to the one shown in Table 7.2. The random starting point ensures that you will not use the same sequence over and over again. The first three digits of Table 7.2 indicate the number of the first file to be reviewed. The random number 001 corresponds to file #1, and the last file, #1000, corresponds to the random number 000. Using Table 7.2, you would choose the five files numbered 155, 450, 32, 882, and 350 for review. Alternately, you might choose to read across the lines, and choose files 155, 350, 989, 450 and 369 for review.

TABLE 7.2 Portion of a Table of Random Numbers

15574	35026	98924
45045	36933	28630
03225	78812	50856
88292	26053	21121

The situation described in Example 7.1 is called an **observational study** because the data already existed before you decided to *observe* or describe their characteristics. Most sample surveys, in which information is gathered with a questionnaire, fall into this category. Computer databases make it possible to assign identification numbers to each element even when the population is large and to select a simple random sample. You must be careful when conducting a *sample survey,* however, to watch for these frequently occurring problems:

- **Nonresponse:** You have carefully selected your random sample and sent out your questionnaires, but only 50% of those surveyed return their questionnaires. Are the responses you received still representative of the entire population, or are they **biased** because only those people who were particularly opinionated about the subject chose to respond?

program.[5] Here is a question from each poll, along with the responses of the sampled Americans:

Space Exploration

CNN/USA Today/Gallup Poll. Dec. 5–7, 2003. Nationwide:

"Would you favor or oppose a new U.S. space program that would send astronauts to the moon?" Form A (N = 510, MoE ± 5)

	Favor	Oppose	No Opinion
	%	%	%
12/03	53	45	2

"Would you favor or oppose the U.S. government spending billions of dollars to send astronauts to the moon?" Form B (N = 494, MoE ± 5)

	Favor	Oppose	No Opinion
	%	%	%
12/03	31	67	2

a. Read the two poll questions. Which of the two wordings is more unbiased? Explain.

b. Look at the responses for the two different polls. How would you explain the large differences in the percentages either favoring or opposing the new program?

7.14 Ask America A nationwide policy survey titled "Ask America" was sent by the National Republican Congressional Committee to voters in the Forty-fourth Congressional District, asking for opinions on a variety of political issues.[6] Here are some questions from the survey:

- In recent years has the federal government grown more or less intrusive in your personal and business affairs?
- Is President Bush right in trying to rein in the size and scope of the federal government against the wishes of the big government Democrats?
- Do you believe the death penalty is a deterrrent to crime?
- Do you agree that the obstructionist Democrats should not be allowed to gain control of the U.S. Congress in the upcoming elections?

Comment on the effect of wording bias on the responses gathered using this survey.

7.3 STATISTICS AND SAMPLING DISTRIBUTIONS

When you select a random sample from a population, the numerical descriptive measures you calculate from the sample are called **statistics.** These statistics vary or change for each different random sample you select; that is, they are *random variables.* The probability distributions for statistics are called **sampling distributions** because, in repeated sampling, they provide this information:

- What values of the statistic can occur
- How often each value occurs

Definition The **sampling distribution of a statistic** is the probability distribution for the possible values of the statistic that results when random samples of size n are repeatedly drawn from the population.

There are three ways to find the sampling distribution of a statistic:

1. Derive the distribution *mathematically* using the laws of probability.
2. Use a *simulation* to approximate the distribution. That is, draw a large number of samples of size n, calculating the value of the statistic for each sample, and tabulate the results in a relative frequency histogram. When the number of

samples is large, the histogram will be very close to the theoretical sampling distribution.

3. Use *statistical theorems* to derive exact or approximate sampling distributions.

The next example demonstrates how to derive the sampling distributions of two statistics for a very small population.

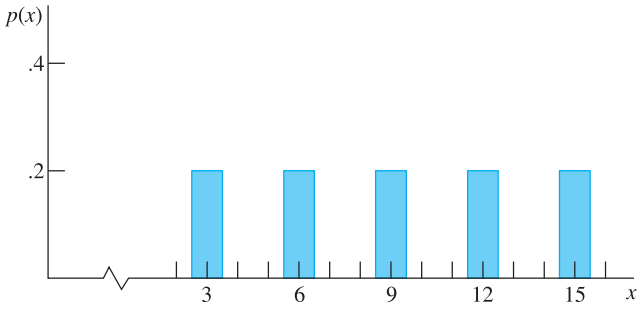
EXAMPLE 7.3

A population consists of $N = 5$ numbers: 3, 6, 9, 12, 15. If a random sample of size $n = 3$ is selected without replacement, find the sampling distributions for the sample mean $\bar{x}$ and the sample median m.

Solution You are sampling from the population shown in Figure 7.1. It contains five distinct numbers and each is equally likely, with probability $p(x) = 1/5$. You can easily find the population mean and median as

$$\mu = \frac{3 + 6 + 9 + 12 + 15}{5} = 9 \quad \text{and} \quad M = 9$$

FIGURE 7.1
Probability histogram for the $N = 5$ population values in Example 7.3



TIP
Sampling distributions can be either discrete or continuous.

There are 10 possible random samples of size $n = 3$ and each is equally likely, with probability 1/10. These samples, along with the calculated values of $\bar{x}$ and m for each, are listed in Table 7.3. You will notice that some values of $\bar{x}$ are more likely than others because they occur in more than one sample. For example,

$$P(\bar{x} = 8) = \frac{2}{10} = .2 \quad \text{and} \quad P(m = 6) = \frac{3}{10} = .3$$

The values in Table 7.3 are tabulated, and the sampling distributions for $\bar{x}$ and m are shown in Table 7.4 and Figure 7.2.

Since the population of $N = 5$ values is symmetric about the value $x = 9$, both the *population mean* and the *median* equal 9. It would seem reasonable, therefore, to consider using either $\bar{x}$ or m as a possible estimator of $M = \mu = 9$. Which estimator would you choose? From Table 7.3, you see that, in using m as an estimator, you would be in error by $9 - 6 = 3$ with probability .3 or by $9 - 12 = -3$ with probability .3. That is, the error in estimation using m would be 3 with probability .6. In using $\bar{x}$, however, an error of 3 would occur with probability only .2. On these grounds alone, you may wish to use $\bar{x}$ as an estimator in preference to m.

TABLE 7.3 Values of $\bar{x}$ and m for Simple Random Sampling when $n = 3$ and $N = 5$

Sample	Sample Values	$\bar{x}$	m
1	3, 6, 9	6	6
2	3, 6, 12	7	6
3	3, 6, 15	8	6
4	3, 9, 12	8	9
5	3, 9, 15	9	9
6	3, 12, 15	10	12
7	6, 9, 12	9	9
8	6, 9, 15	10	9
9	6, 12, 15	11	12
10	9, 12, 15	12	12

TABLE 7.4 Sampling Distributions for (a) the Sample Mean and (b) the Sample Median

(a) $\bar{x}$	$p(\bar{x})$	(b) m	$p(m)$
6	.1	6	.3
7	.1	9	.4
8	.2	12	.3
9	.2		
10	.2		
11	.1		
12	.1		

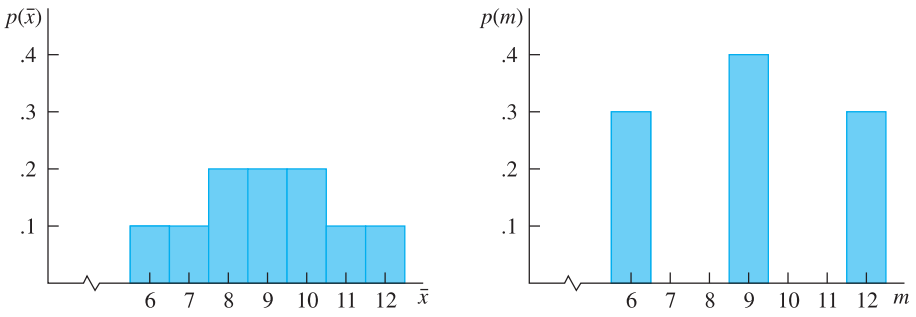


FIGURE 7.2 Probability histograms for the sampling distributions of the sample mean, $\bar{x}$, and the sample median, m, in Example 7.3

MY TIP

Almost every statistic has a mean and a standard deviation (or *standard error*) describing its center and spread.

It was not too difficult to derive these sampling distributions in Example 7.3 because the number of elements in the population was very small. When this is not the case, you may need to use one of these methods:

- Use a simulation to approximate the sampling distribution empirically.
- Rely on statistical theorems and theoretical results.

One important statistical theorem that describes the sampling distribution of statistics that are sums or averages is presented in the next section.

8.1 WHERE WE'VE BEEN

The first seven chapters of this book have given you the building blocks you will need to understand statistical inference and how it can be applied in practical situations. The first three chapters were concerned with using descriptive statistics, both graphical and numerical, to describe and interpret sets of measurements. In the next three chapters, you learned about probability and probability distributions—the basic tools used to describe *populations* of measurements. The binomial and the normal distributions were emphasized as important for practical applications. The seventh chapter provided the link between probability and statistical inference. Many statistics are either sums or averages calculated from sample measurements. The Central Limit Theorem states that, even if the sampled populations are not normal, the sampling distributions of these *statistics* will be approximately normal when the sample size n is large. These statistics are the tools you use for *inferential statistics*—making inferences about a population using information contained in a sample.

8.2 WHERE WE'RE GOING—STATISTICAL INFERENCE

Inference—specifically, decision making and prediction—is centuries old and plays a very important role in most peoples' lives. Here are some applications:

- The government needs to predict short- and long-term interest rates.
- A broker wants to forecast the behavior of the stock market.
- A metallurgist wants to decide whether a new type of steel is more resistant to high temperatures than the current type.
- A consumer wants to estimate the selling price of her house before putting it on the market.

MY TIP
Parameter ⇔ Population
Statistic ⇔ Sample

There are many ways to make these decisions or predictions, some subjective and some more objective in nature. How good will your predictions or decisions be? Although you may feel that your own built-in decision-making ability is quite good, experience suggests that this may not be the case. It is the job of the mathematical statistician to provide methods of statistical inference making that are better and more reliable than just subjective guesses.

Statistical inference is concerned with making decisions or predictions about **parameters**—the numerical descriptive measures that characterize a population. Three parameters you encountered in earlier chapters are the population mean μ, the population standard deviation σ, and the binomial proportion p. In statistical inference, a practical problem is restated in the framework of a population with a specific parameter of interest. For example, the metallurgist could measure the *average* coefficients of expansion for both types of steel and then compare their values.

Methods for making inferences about population parameters fall into one of two categories:

- **Estimation:** Estimating or predicting the value of the parameter
- **Hypothesis testing:** Making a decision about the value of a parameter based on some preconceived idea about what its value might be

EXAMPLE 8.1

The circuits in computers and other electronics equipment consist of one or more printed circuit boards (PCB), and computers are often repaired by simply replacing one or more defective PCBs. In an attempt to find the proper setting of a plating process applied to one side of a PCB, a production supervisor might *estimate* the average thickness of copper plating on PCBs using samples from several days of operation. Since he has no knowledge of the average thickness μ before observing the production process, his is an *estimation* problem.

EXAMPLE 8.2

The supervisor in Example 8.1 is told by the plant owner that the thickness of the copper plating must not be less than .001 inch in order for the process to be in control. To decide whether or not the process is in control, the supervisor might formulate a test. He could *hypothesize* that the process is in control—that is, assume that the average thickness of the copper plating is .001 or greater—and use samples from several days of operation to decide whether or not his hypothesis is correct. The supervisor's decision-making approach is called a *test of hypothesis.*

Which method of inference should be used? That is, should the parameter be estimated, or should you test a hypothesis concerning its value? The answer is dictated by the practical question posed and is often determined by personal preference. Since both estimation and tests of hypotheses are used frequently in scientific literature, we include both methods in this and the next chapter.

A statistical problem, which involves planning, analysis, and inference making, is incomplete without a measure of the **goodness of the inference.** That is, how accurate or reliable is the method you have used? If a stockbroker predicts that the price of a stock will be $80 next Monday, will you be willing to take action to buy or sell your stock without knowing how reliable her prediction is? Will the prediction be within $1, $2, or $10 of the actual price next Monday? Statistical procedures are important because they provide two types of information:

- Methods for making the inference
- A numerical measure of the goodness or reliability of the inference

8.3 TYPES OF ESTIMATORS

To estimate the value of a population parameter, you can use information from the sample in the form of an **estimator.** Estimators are calculated using information from the sample observations, and hence, by definition they are also *statistics.*

Definition An **estimator** is a rule, usually expressed as a formula, that tells us how to calculate an estimate based on information in the sample.

Estimators are used in two different ways:

- **Point estimation:** Based on sample data, a single number is calculated to estimate the population parameter. The rule or formula that describes this calculation is called the **point estimator,** and the resulting number is called a **point estimate.**

- **Interval estimation:** Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie. The rule or formula that describes this calculation is called the **interval estimator,** and the resulting pair of numbers is called an **interval estimate** or **confidence interval.**

EXAMPLE 8.3

A veterinarian wants to estimate the average weight gain per month of 4-month-old golden retriever pups that have been placed on a lamb and rice diet. The *population* consists of the weight gains per month of all 4-month-old golden retriever pups that are given this particular diet. The veterinarian wants to estimate the unknown parameter μ, the average monthly weight gain for this *hypothetical* population. One possible *estimator* based on sample data is the sample mean, $\bar{x} = \Sigma x_i/n$. It could be used in the form of a single number or *point estimate*—for instance, 3.8 pounds—or you could use an *interval estimate* and estimate that the average weight gain will be between 2.7 and 4.9 pounds.

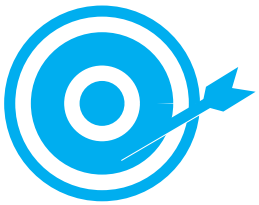
Both point and interval estimation procedures use information provided by the sampling distribution of the specific estimator you have chosen to use. We will begin by discussing *point estimation* and its use in estimating population means and proportions.

8.4 POINT ESTIMATION

MY TIP

parameter = target's bull's-eye

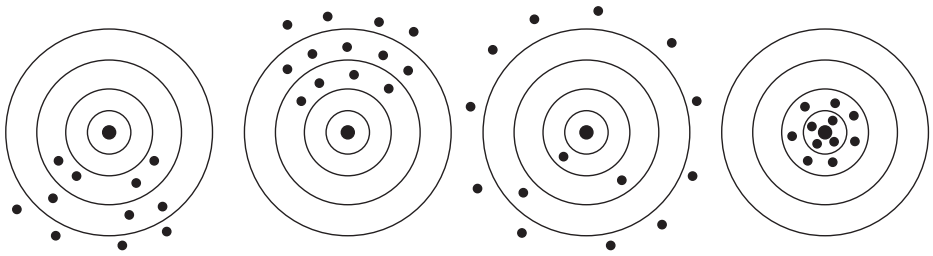
estimator = bullet or arrow



In a practical situation, there may be several statistics that could be used as point estimators for a population parameter. To decide which of several choices is best, you need to know how the estimator behaves in repeated sampling, described by its *sampling distribution.*

By way of analogy, think of firing a revolver at a target. The parameter of interest is the bull's-eye, at which you are firing bullets. Each bullet represents a single sample estimate, fired by the revolver, which represents the estimator. Suppose your friend fires a single shot and hits the bull's-eye. Can you conclude that he is an excellent shot? Would you stand next to the target while he fires a second shot? Probably not, because you have no measure of how well he performs in repeated trials. Does he always hit the bull's-eye, or is he consistently too high or too low? Do his shots cluster closely around the target, or do they consistently miss the target by a wide margin? Figure 8.1 shows several target configurations. Which target would you pick as belonging to the best shot?

FIGURE 8.1
Which marksman is best?



Resampling
Bootstraping
K fold cross validation

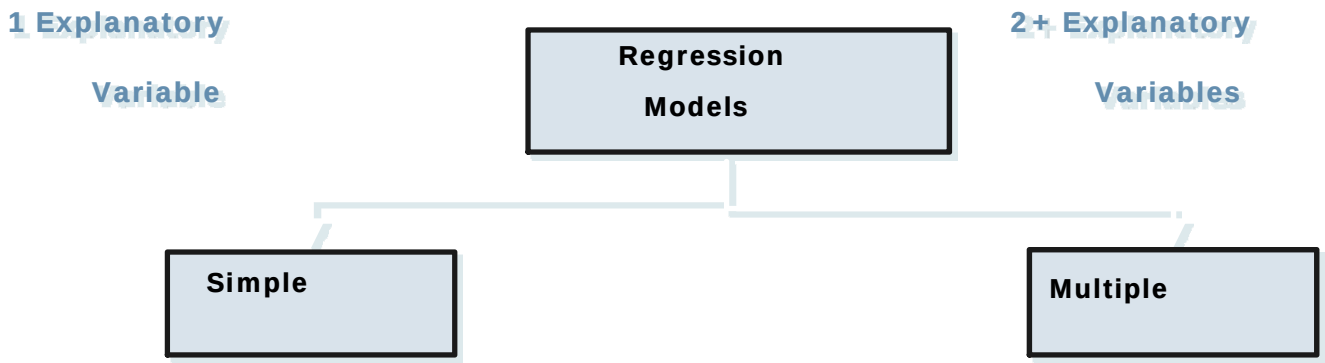
Linear Regression

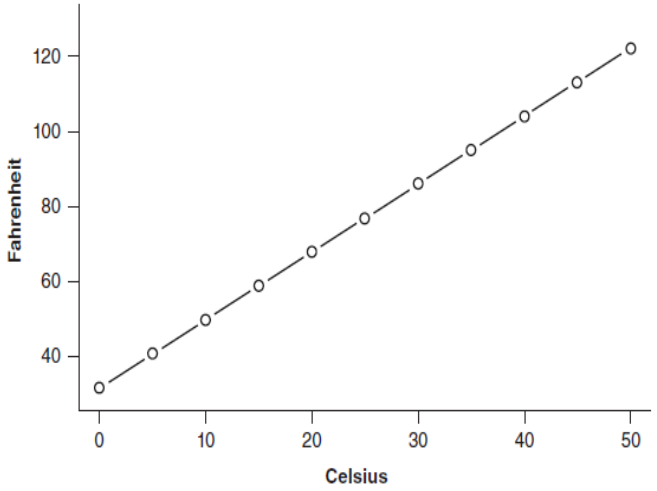
Introduction to Supervised Learning and Regression

- Regression analysis falls under supervised machine learning.
- The system tries to predict a value for an input based on previous information.
- Characteristics of regression:
  - The responses obtained from the model are always quantitative in nature.
  - The model can be constructed only if past data is available.

Simple Linear Regression

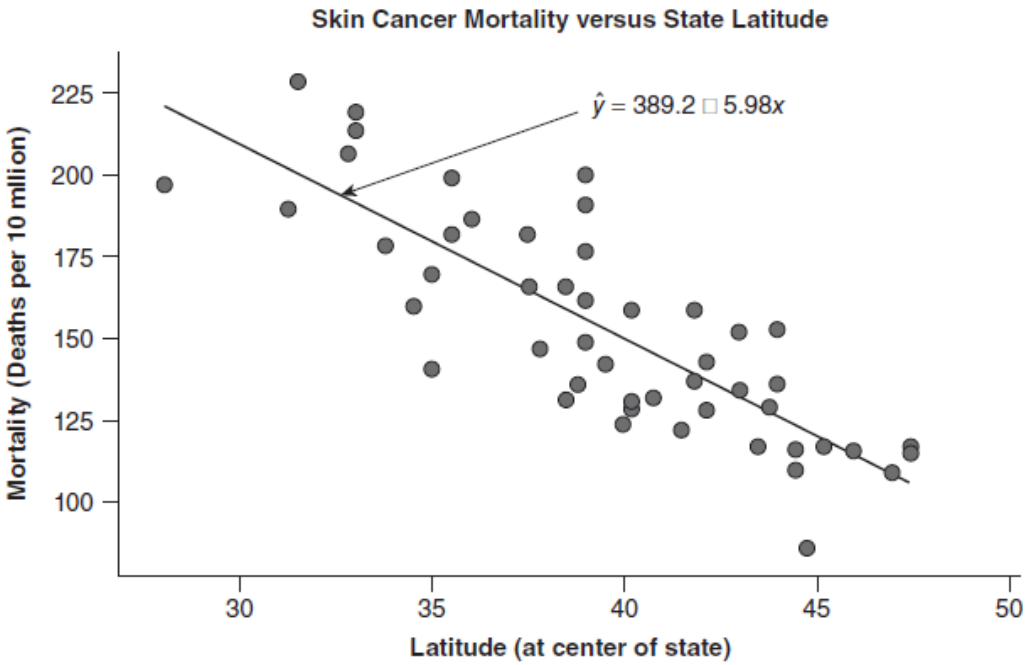
- Is a statistical method that allows us to summarize and study the relationship between two continuous (quantitative) variables.
  - Between predictor, explanatory, or independent variable and response, outcome, or dependent variable.
- Statistical relationships are different from deterministic relationships.





Graph showing the deterministic relationship between Fahrenheit and Celsius scales.

The plot of mortality due to skin cancer versus latitude.



Linear Regression

- A linear regression line has an equation of the form

$$y = a + bx$$

where x is the explanatory variable, y is the dependent variable, b is the slope of the line, and a is the intercept

Steps to Establish a Linear Regression

- Carry out an experiment of gathering a sample of observed values
- Create a relationship model.
- Find the coefficients from the model created and establish the mathematical equation using these.
- Compute the residual error or residual.
- Use the model for prediction.

Evaluation of Model Estimators

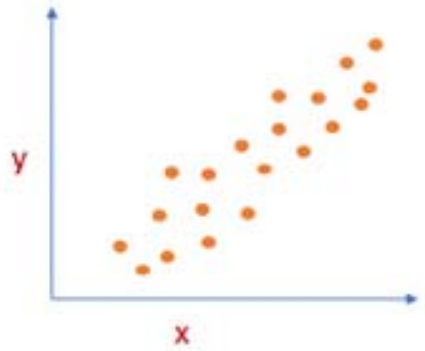
- Once the model is established, you need to confirm whether the model is good enough to make predictions
- Various metrics are used
  - Karl Pearson's Coefficient of Correlation

Karl Pearson's Coefficient of Correlation

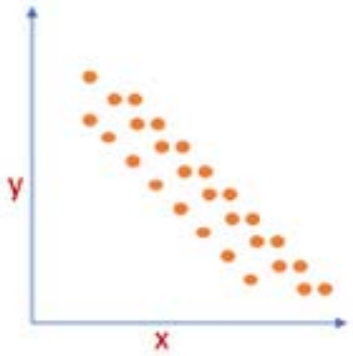
- Karl Pearson's coefficient of correlation is a helpful statistical formula that quantifies the strength between two variables.
- This coefficient value helps in determining how strong that relationship is between the two variables.

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

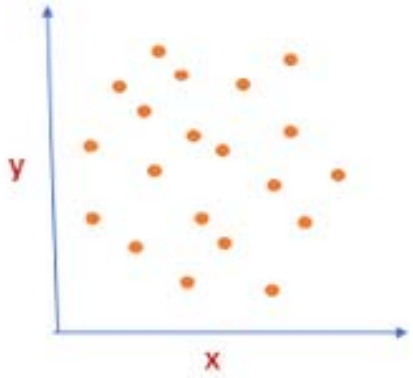
where *x* and *y* are variables and *N* is the number of instances



Positive Correlation



Negative Correlation



No Correlation

Example 1: calculate correlation coefficient for the following data:

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

X	Y	X^2	Y^2	XY
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
$\sum X = 36$	$\sum Y = 60$	$\sum X^2 = 266$	$\sum Y^2 = 706$	$\sum(XY) = 293$

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

$$r = \frac{6 \times 293 - 36 \times 60}{\sqrt{6 \times 266 - 36^2}\ \sqrt{6 \times 706 - 60^2}}$$

$$= \frac{1758 - 2160}{\sqrt{1590 - 1296}\ \sqrt{4236 - 3600}}$$

$$= \frac{-402}{17.32 \times 25.22}$$

$$= \frac{-402}{436.81}$$

$$= -0.920$$

Probable error of coefficient of Correlation ‘r’

- Probable error of the coefficient of correlation is a statistical measure which measures reliability and dependability of the value of coefficient of correlation.
- If probable error is added to or subtracted from the coefficient of correlation
  - it would give two such limits within which we can reasonably expect the value of coefficient of correlation to vary.

$$\text{Probable error of 'r'} = \frac{0.6745(1 - r^2)}{\sqrt{n}}$$

Correlation Based on Karl Pearson

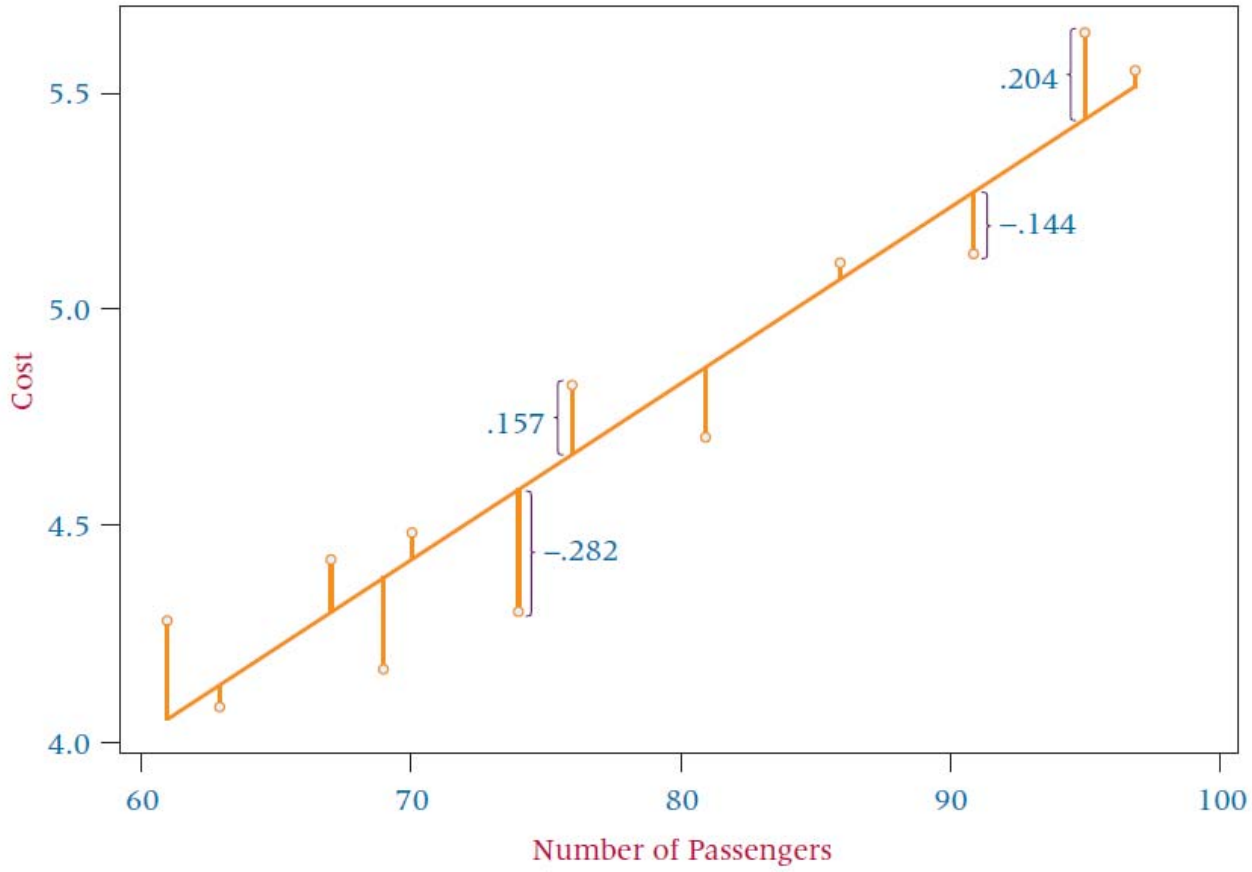
- It has a value between +1 and −1
- 1 is total positive linear correlation
- 0 is no linear correlation
- −1 is total negative linear correlation
- Limitations of Pearson coefficients
  - It always assumes linear relationship.
  - Interpreting the value of *r* is difficult.
  - It is time consuming.

Residual Analysis

- Each difference between the actual y values and the predicted y values is the error of the regression line at a given point, y-y^ and is called residual.
- Uses of Residuals are:

  To locate Outliers

  To test the assumptions of the Regression Model


5.5
5.0
4.5
4.0
Cost
60
70
80
90
100
Number of Passengers
.204
−.144
.157
−.282

1. What is data analytics ?

1. Data analytics is the science of analyzing raw data in order to make conclusions about that information.
2. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things.
3. Data analytics techniques can help in finding the trends and metrics that would be used to optimize processes to increase the overall efficiency of a business or system.
4. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
5. For example, manufacturing companies often record the runtime, downtime, and work queue for various machines and then analyze the data to better plan the workloads so the machines operate closer to peak capacity.

2. Source of data (or Big Data).

Three primary sources of Big Data are :

1. **Social data :**
   a. Social data comes from the likes, tweets & retweets, comments, video uploads, and general media that are uploaded and shared via social media platforms.
   b. This kind of data provides invaluable insights into consumer behaviour and sentiment and can be enormously influential in marketing analytics.
   c. The public web is another good source of social data, and tools like Google trends can be used to good effect to increase the volume of big data.
2. **Machine data :**
   a. Machine data is defined as information which is generated by industrial equipment, sensors that are installed in machinery, and even web logs which track user behaviour.
   b. This type of data is expected to grow exponentially as the internet of things grows ever more pervasive and expands around the world.
   c. Sensors such as medical devices, smart meters, road cameras, satellites, games and the rapidly growing Internet of Things will deliver high velocity, value, volume and variety of data in the very near future.
3. **Transactional data :**
   a. Transactional data is generated from all the daily transactions that take place both online and offline.

b. Invoices, payment orders, storage records, delivery receipts are characterized as transactional data.

3. Classification of databig data.

1. **Unstructured data** :
   a. Unstructured data is the rawest form of data.
   b. Data that has no inherent structure, which may include text documents, PDFs, images, and video.
   c. This data is often stored in a repository of files.
2. **Structured data** :
   a. Structured data is tabular data (rows and columns) which are very well defined.
   b. Data containing a defined data type, format, and structure, which may include transaction data, traditional RDBMS, CSV files, and simple spreadsheets.
3. **Semi-structured data:**
   a. Textual data files with a distinct pattern that enables parsing such as Extensible Markup Language [XML) data files or JSON.
   b. A consistent format is defined however the structure is not very strict.
   c. Semi-structured data are often stored as files.

Differentiate between structured, semi-structured and unstructured data.

Answer

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table.	It is based on XML/ RDF.	It is based on character and binary data.
Transaction management	Matured transaction and various concurrency techniques.	Transaction is adapted from DBMS.	No transaction management and no concurrency.
Flexibility	Itis schema dependent and less flexible.	It is more flexible than structured data but less than flexible than unstructured data.	It very flexible and there is absence of schema.
Scalability	It is very difficult to scale database schema.	It is more scalable than structured data.	It is very scalable.
Query performance	Structured query allow complex joining.	Queries over anonymous nodes are possible.	Only textual query are possible.

4. Characteristics of Big Data.

Big Data is characterized into four dimensions :

1. **Volume:**
   a. Volume is concerned about scale of data *i.e.,* the volume of the data at which it is growing.
   b. The volume of data is growing rapidly, due to several applications of business, social, web and scientific explorations.
2. **Velocity:**
   a. The speed at which data is increasing thus demanding analysis of streaming data.
   b. The velocity is due to growing speed of business intelligence applications such as trading, transaction of telecom and banking domain, growing number of internet connections with the increased usage of internet etc.
3. **Variety:** It depicts different forms of data to use for analysis such as structured, semi structured and unstructured.
4. **Veracity:**
   a. Veracity is concerned with uncertainty or inaccuracy of the data. b. In many cases the data will be inaccurate hence filtering and selecting the data which is actually needed is a complicated task.
   c. A lot of statistical and analytical process has to go for data cleansing for choosing intrinsic data for decision making.

Que 1. Write short note on big data platform.

Answer

1. Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.
2. It is an enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure/ environment.
3. Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities.
4. It also supports custom development, querying and integration with other systems.
5. The primary benefit behind a big data platform is to reduce the complexity of multiple vendors/ solutions into a one cohesive solution.
6. Big data platform are also delivered through cloud where the provider provides an all inclusive big data solutions and services.

Que 2 Why there is need of data analytics ?

Answer

Need of data analytics:

1. It optimizes the business performance.
2. It helps to make better decisions.
3. It helps to analyze customers trends and solutions.

Que 3. What are the steps involved in data analysis ?

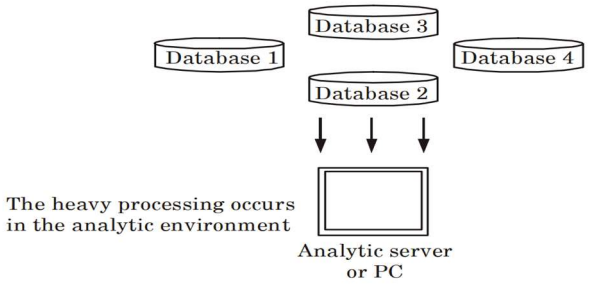
Answer

Steps involved in data analysis are :

1. **Determine the data :**
   a. The first step is to determine the data requirements or how the data is grouped.
   b. Data may be separated by age, demographic, income, or gender.
   c. Data values may be numerical or be divided by category.
2. **Collection of data :**
   a. The second step in data analytics is the process of collecting it.
   b. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
3. **Organization of data :**
   a. Third step is to organize the data.
   b. Once the data is collected, it must be organized so it can be analyzed.
   c. Organization may take place on a spreadsheet or other form of software that can take statistical data.
4. **Cleaning of data :**
   a. In fourth step, the data is then cleaned up before analysis.
   b. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete.
   c. This step helps correct any errors before it goes on to a data analyst to be analyzed.

5. Evolution of analytics scalability.

In analytic scalability, we have to pull the data together in a separate analytics environment and then start performing analysis.



2. Analysts do the merge operation on the data sets which contain rows and columns.
3. The columns represent information about the customers such as name, spending level, or status.
4. In merge or join, two or more data sets are combined together. They are typically merged / joined so that specific rows of one data set or table are combined with specific rows of another.
5. Analysts also do data preparation. Data preparation is made up of joins, aggregations, derivations, and transformations. In this process, they pull data from various sources and merge it all together to create the variables required for an analysis.
6. Massively Parallel Processing (MPP) system is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data.
7. An MPP database breaks the data into independent pieces managed by independent storage and central processing unit (CPU) resources.

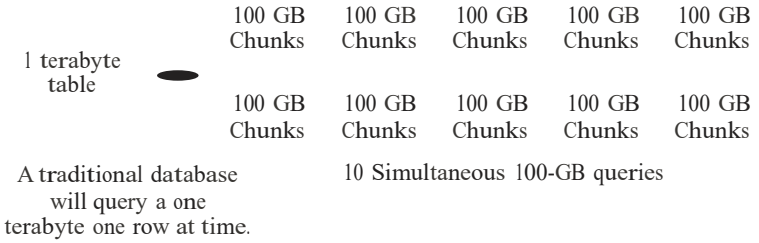


Fig. 1.10.1. Massively Parallel Processing system data storage.

8. MPP systems build in redundancy to make recovery easy.
9. MPP systems have resource management tools :
   a. Manage the CPU and disk space
   b. Query optimizer

6. Evolution of analytic process.

1. With increased level of scalability, it needs to update analytic processes to take advantage of it.
2. This can be achieved with the use of analytical sandboxes to provide analytic professionals with a scalable environment to build advanced analytics processes.
3. One of the uses of MPP database system is to facilitate the building and deployment of advanced analytic processes.
4. An analytic sandbox is the mechanism to utilize an enterprise data warehouse.
5. If used appropriately, an analytic sandbox can be one of the primary drivers of value in the world of big data.

Analytical sandbox :

1. An analytic sandbox provides a set of resources with which in-depth analysis can be done to answer critical business questions.
2. An analytic sandbox is ideal for data exploration, development of analytical processes, proof of concepts, and prototyping.
3. Once things progress into ongoing, user-managed processes or production processes, then the sandbox should not be involved.
4. A sandbox is going to be leveraged by a fairly small set of users.
5. There will be data created within the sandbox that is segregated from the production database.
6. Sandbox users will also be allowed to load data of their own for brief time periods as part of a project, even if that data is not part of the official enterprise data model.

Que 1 Explain modern data analytic tools.

Answer

Modern data analytic tools :

1. **Apache Hadoop** :
   a. Apache Hadoop, a big data analytics tool which is a Java based free software framework.
   b. It helps in effective storage of huge amount of data in a storage place known as a cluster.
   c. It runs in parallel on a cluster and also has ability to process huge data across all nodes in it.
   d. There is a storage system in Hadoop popularly known as the Hadoop Distributed File System (HDFS), which helps to splits the large volume of data and distribute across many nodes present in a cluster.
2. **KNIME:**

a. KNIME analytics platform is one of the leading open solutions for data-driven innovation.

b. This tool helps in discovering the potential and hidden in a huge volume of data, it also performs mine for fresh insights, or predicts the new futures.

3. **OpenRefine** :

a. OneRefine tool is one of the efficient tools to work on the messy and large volume of data.

b. It includes cleansing data, transforming that data from one format another.

c. It helps to explore large data sets easily.

4. **Orange** :

a. Orange is famous open-source data visualization and helps in data analysis for beginner and as well to the expert.

b. This tool provides interactive workflows with a large toolbox option to create the same which helps in analysis and visualizing of data.

5. **RapidMiner** :

a. RapidMiner tool operates using visual programming and also it is much capable of manipulating, analyzing and modeling the data.

b. RapidMiner tools make data science teams easier and productive by using an open-source platform for all their jobs like machine learning, data preparation, and model deployment.

6. **R-programming** :

a. R is a free open source software programming language and a software environment for statistical computing and graphics.

b. It is used by data miners for developing statistical software and data analysis.

c. It has become a highly popular tool for big data in recent years.

7. **Datawrapper** :

a. It is an online data visualization tool for making interactive charts.

b. It uses data file in a csv, pdf or excel format.

c. Datawrapper generate visualization in the form of bar, line, map etc. It can be embedded into any other website as well.

8. **Tableau** :

a. Tableau is another popular big data tool. It is simple and very intuitive to use.

b. It communicates the insights of the data through data visualization.

c. Through Tableau, an analyst can check a hypothesis and explore the data before starting to work on it extensively.

4. **Freedom** : Analytic professionals can reduce focus on the administration of systems and production processes by shifting those maintenance tasks to IT.
5. **Speed** : Massive speed improvement will be realized with the move to parallel processing. This also enables rapid iteration and the ability to "fail fast" and take more risks to innovate.

Que 2. Explain the application of data analytics.

Answer

Application of data analytics :

1. **Security:** Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas.
2. **Transportation** :
   a. Data analytics can be used to revolutionize transportation.
   b. It can be used especially in areas where we need to transport a large number of people to a specific area and require seamless transportation.
3. **Risk detection** :
   a. Many organizations were struggling under debt, and they wanted a solution to problem of fraud.
   b. They already had enough customer data in their hands, and so,
   c. They used 'divide and conquer' policy with the data, analyzing recent expenditure, profiles, and any other important information to understand any probability of a customer defaulting.
4. **Delivery** :
   a. Several top logistic companies are using data analysis to examine collected data and improve their overall efficiency.
   b. Using data analytics applications, the companies were able to find the best shipping routes, delivery time, as well as the most cost• efficient transport means.
5. **Fast internet allocation** :
   a. While it might seem that allocating fast internet in every area makes a city 'Smart', in reality, it is more important to engage in smart allocation. This smart allocation would mean understanding how bandwidth is being used in specific areas and for the right cause.
   b. It is also important to shift the data allocation based on timing and priority. It is assumed that financial and commercial areas require the most bandwidth during weekdays, while residential areas

require it during the weekends. But the situation is much more complex. Data analytics can solve it.

c. For example, using applications of data analysis, a community can draw the attention of high-tech industries and in such cases; higher bandwidth will be required in such areas.

6. **Internet searching** :
   a. When we use Google, we are using one of their many data analytics applications employed by the company.
   b. Most search engines like Google, Bing, Yahoo, AOL etc., use data analytics. These search engines use different algorithms to deliver the best result for a search query.
7. **Digital advertisement** :
   a. Data analytics has revolutionized digital advertising.
   b. Digital billboards in cities as well as banners on websites, that is, most of the advertisement sources nowadays use data analytics using data algorithms.

Que 3. What are the different types of Big Data analytics ?

Answer

Different types of Big Data analytics :

1. **Descriptive analytics** :
   a. It uses data aggregation and data mining to provide insight into the past.
   b. Descriptive analytics describe or summarize raw data and make it interpretable by humans.
2. **Predictive analytics** :
   a. It uses statistical models and forecasts techniques to understand the future.
   b. Predictive analytics provides companies with actionable insights based on data. It provides estimates about the likelihood of a future outcome.
3. **Prescriptive analytics** :
   a. It uses optimization and simulation algorithms to advice on possible outcomes.
   b. It allows users to "prescribe" a number of different possible actions and guide them towards a solution.
4. **Diagnostic analytics** :
   a. It is used to determine why something happened in the past.
   b. It is characterized by techniques such as drill-down, data discovery, data mining and correlations.
   c. Diagnostic analytics takes a deeper look at data to understand the root causes of the events.

.

7. Regression modeling.

1. Regression models are widely used in analytics, in general being among the most easy to understand and interpret type of analytics techniques.
2. Regression techniques allow the identification and estimation of possible relationships between a pattern or variable of interest, and factors that influence that pattern.
3. For example, a company may be interested in understanding the effectiveness of its marketing strategies.
4. A regression model can be used to understand and quantify which of its marketing activities actually drive sales, and to what extent.
5. Regression models are built to understand historical data and relationships to assess effectiveness, as in the marketing effectiveness models.
6. Regression techniques are used across a range of industries, including financial services, retail, telecom, pharmaceuticals, and medicine.

Types of regression analysis techniques

1. **Linear regression** : Linear regressions assumes that there is a linear relationship between the predictors (or the factors) and the target variable.
2. **Non-linear regression** : Non-linear regression allows modeling of non-linear relationships.
3. **Logistic regression** : Logistic regression is useful when our target variable is binomial (accept or reject).
4. **Time series regression** : Time series regressions is used to forecast future behavior of variables based on historical time ordered data.

7.1 Linear regression models.

1. We consider the modelling between the dependent and one independent variable. When there is only one independent variable in the regression model, the model is generally termed as a linear regression model.
2. Consider a simple linear regression model
$$y = \beta_0 + \beta_1 X + \varepsilon$$
Where,
y is termed as the dependent or study variable and X is termed as the independent or explanatory variable.
The terms β_0 and β_1 are the parameters of the model. The parameter β_0 is termed as an intercept term, and the parameter β_1 is termed as the slope parameter.
3. These parameters are usually called as regression coefficients. The unobservable error component accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y.
4. There can be several reasons for such difference, such as the effect of all deleted variables in the model, variables may be qualitative, inherent randomness in the observations etc.
5. We assume that ε is observed as independent and identically distributed random variable with mean zero and constant variance σ^2 and assume that ε is normally distributed.
6. The independent variables are viewed as controlled by the experimenter, so it is considered as non-stochastic whereas y is viewed as a random variable with
$$E(y) = \beta_0 + \beta_1 X \text{ and } Var(y) = \sigma^2.$$
7. Sometimes X can also be a random variable. In such a case, instead of the sample mean and sample variance of y, we consider the conditional mean of y given $X = x$ as
$$E(y \mid x) = \beta_0 + \beta_1 x$$
and the conditional variance of y given $X = x$ as
$$Var(y \mid x) = \sigma^2$$
8. When the values of β_0, β_1, and σ^2 are known, the model is completely described. The parameters β_0, β_1 and σ^2 arc generally unknown in practice and ε is unobserved. The determination of the statistical model $y = \beta_0 + \beta_1 X + \varepsilon$ depends on the determination (*i.e.* estimation) of β_0, β_1, and σ^2. In order to know the values of these parameters, n pairs of observations $(x_1, y_i)(i = 1, \ldots, n)$ on (X, y) are observed/collected and are used to determine these unknown parameters.

7.2 Multivariate analysis.

1. Multivariate analysis (MVA) is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time.
2. These variables are nothing but prototypes of real time situations, products and services or decision making involving more than one variable.
3. MVA is used to address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important.
4. Multiple regression analysis refers to a set of techniques for studying the straight-line relationships among two or more variables.
5. Multiple regression estimates the β's in the equation

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_p x_{pj} + \varepsilon_j$$

Where, the x's are the independent variables. y is the dependent variable. The subscript j represents the observation (row) number. The β's are the unknown regression coefficients. Their estimates are represented by b's. Each β represents the original unknown (population) parameter, while b is an estimate of this β. The ε_j is the error (residual) of observation j.

6. Regression problem is solved by least squares. In least squares method regression analysis, the b's are selected so as to minimize the sum of the squared residuals. This set of b's is not necessarily the set we want, since they may be distorted by outliers points that are not representative of the data. Robust regression, an alternative to least squares, seeks to reduce the influence of outliers.
7. Multiple regression analysis studies the relationship between a dependent (response) variable and p independent variables (predictors, regressors).
8. The sample multiple regression equation is

$$\hat{y}_j = b_0 + b_1 x_{2_j} + \ldots + b_p x_{p_j}$$

10. If $p = 1$, the model is called simple linear regression. The intercept, b_0, is the point at which the regression plane intersects the Y axis. The b_i are the slopes of the regression plane in the direction of x_i. These coefficients are called the partial-regression coefficients. Each partial regression coefficient represents the net effect the i^{th} variable has on the dependent variable, holding the remaining x's in the equation constant

Bayesian network.

1. Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations.
2. *A* Bayesian network is a directed acyclic graph in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable.
3. Bayesian networks aim to model conditional dependence by representing edges in a directed graph.

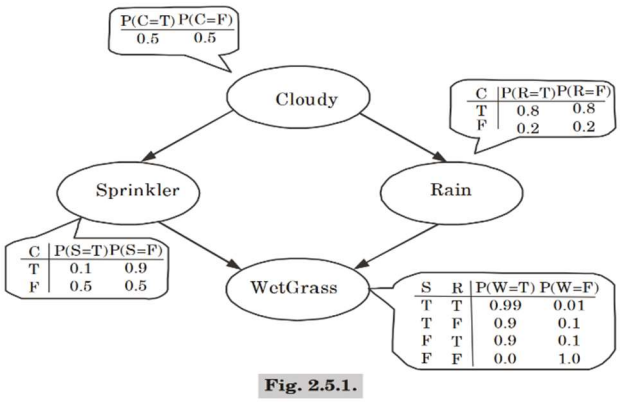


Fig. 2.5.1.

Through these relationships, one can efficiently conduct inference on the random variables in the graph through the use of factors.

4. Using the relationships specified by our Bayesian network, we can obtain a compact, factorized representation of the joint probability distribution by taking advantage of conditional independence.
5. Formally, if an edge (*A*, B) exists in the graph connecting random variables *A* and B, it means that P(B | A) is a factor in the joint probability distribution, so we must know *P*(*B* | *A*) for all values of *B* and *A* in order to conduct inference.
6. In the Fig. 2.5.1, since Rain has an edge going into WetGrass, it means that P(WetGrass | Rain) will be a factor, whose probability values are specified next to the WetGrass node in a conditional probability table.
7. Bayesian networks satisfy the Markov property, which states that a node is conditionally independent of its non-descendants given its parents. In the given example, this means that

   P(Sprinkler | Cloudy, Rain) = P(Sprinkler | Cloudy)

   Since Sprinkler is conditionally independent of its non-descendant, Rain, given Cloudy.