

Prepared by Dr. Wesam Ahmed

IT393
Business Intelligence

Lecture 9



Predictive Analytics : Web and Social Media Analytics

Lecture 9

IT393

Business Intelligence

Prepared by
Dr. Wesam Ahmed

01

Learning Objectives:

- 1.1 Web Mining Overview
- 1.2 Web mining can be viewed as consisting of three areas
- 1.3 Understand the Search Engines
- 1.4 Social Analytics
- 1.5 Tableau

02

Web mining

The growth of the Internet and its enabling technologies has made data creation, data collection, and data/information/opinion exchange easier. Delays in service, manufacturing, shipping, delivery, and customer inquiries are no longer private incidents and are accepted as necessary evils. Now, thanks to social media tools and technologies on the Internet, everybody knows everything. Successful companies are the ones who embrace these Internet technologies and use them for the betterment of their business processes so that they can better communicate with their customers, understand their needs and wants, and serve them thoroughly and expeditiously. Being customer focused and keeping customers happy has never been as important a concept for businesses as they are now, in this age of Internet and social media.

03

Web mining

The World Wide Web (or for short, Web) serves as an enormous repository of data and information on virtually everything one can conceive; business, personal, you name it, an abundant amount of it is there. The Web is perhaps the world's largest data and text repository, and the amount of information on the Web is growing rapidly. A lot of interesting information can be found online: whose home page is linked to which other pages, how many people have links to a specific Web page, and how a particular site is organized. In addition, each visitor to a Web site, each search on a search engine, each click on a link, and each transaction on an e-commerce site creates additional data.

04

Web mining

Although unstructured textual data in the form of Web pages coded in **HTML or XML** is the dominant content of the Web, the Web infrastructure also **contains hyperlink information (connections to other Web pages) and usage information (logs of visitors' interactions with Web sites)**, all of which provide **rich data for knowledge discovery. Analysis** of this information can help us make better use of **Web sites and also aid us in enhancing relationships and value for the visitors to our own Web sites.**

Web is not an easy undertaking by any means. **The Web also poses great challenges for effective and efficient knowledge discovery :**

05

Web mining

1. **The Web is too big for effective data mining.** The Web is so large and growing so rapidly that it is difficult to even quantify its size. Because of the sheer size of the Web, it is not feasible to set up a data warehouse to replicate, store, and integrate all of the data on the Web, making data collection and integration a challenge.
2. **The Web is too complex.** The complexity of a Web page is far greater than that of a page in a traditional text document collection. Web pages lack a unified structure. They contain far more authoring style and content variation than any set of books, articles, or other traditional text-based document.
3. **The Web is not specific to a domain.** The Web serves a broad diversity of communities and connects billions of workstations. Web users have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search that they perform.

06

Web mining

These challenges have prompted many research efforts to enhance the effectiveness and efficiency of discovering and using data assets on the Web. A number of index based Web search engines constantly search the Web and index Web pages under certain keywords. Using these search engines, an experienced user may be able to locate documents by providing a set of tightly constrained keywords or phrases. First, a topic of any breadth can easily contain hundreds or thousands of documents. This can lead to a large number of document entries returned by the search engine, many of which are marginally relevant to the topic. Second, many documents that are highly relevant to a topic may not contain the exact keywords defining them. **Web mining** is a prominent (and more challenging) approach that can be used to substantially enhance the power of Web search engines because Web mining can identify authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search engines.

07

Web mining

Web mining is the process of **discovering** intrinsic relationships (i.e., interesting and useful information) from **Web data**, which are expressed in the form of **textual, linkage, or usage information**.

Web mining is essentially the same as **data mining** that uses **data generated over the Web**. The goal is to **turn vast repositories of business transactions, customer interactions, and Web site usage data into actionable information** (i.e., knowledge) to promote **better decision making throughout the enterprise**. Because of the increased popularity of the term *analytics*, nowadays many have started to refer to **Web mining as Web analytics**. However, **these two terms are not the same**. Whereas **Web analytics is primarily Web site usage data focused**, Web mining is inclusive of **all data generated via the Internet including transaction, social, and usage data**.

08

Web mining

Web Analytics

Focus: Primarily on **website usage data**.

Purpose: To analyze **how users interact with a specific website** (e.g., clicks, page views, time on site).

Data Sources:

- Web server logs
- Google Analytics
- Session recordings, heatmaps

Goal: Improve **website performance, user experience, and conversion rates**.

09

Web mining

Web Mining

Focus: Broader; includes **all types of data from the web**.

Purpose: To discover patterns and insights from **large-scale web data**, not limited to just one site.

Data Sources:

- Web content (text, images, videos)
- Web structure (link analysis)
- Web usage (like analytics data)
- **Social media, e-commerce transactions**, forums, etc.

Goal: Extract knowledge for research, recommendation systems, market trends, etc.

10

Web mining

Where **Web analytics** aims to **describe what has happened on the Web site** (employing a predefined, metrics-driven descriptive analytics methodology), **Web mining** aims to **discover previously unknown patterns and relationships** (employing a novel predictive or prescriptive analytics methodology). From a big-picture perspective, **Web analytics can be considered to be a part of Web mining**. Figure 9.1 presents a simple taxonomy of Web mining, where it is divided into three main areas: **Web content mining, Web structure mining, and Web usage mining**. In the figure, the data sources used in these three main areas are also specified. Although these three areas are shown separately, **they are often used collectively and synergistically to address business problems and opportunities**.

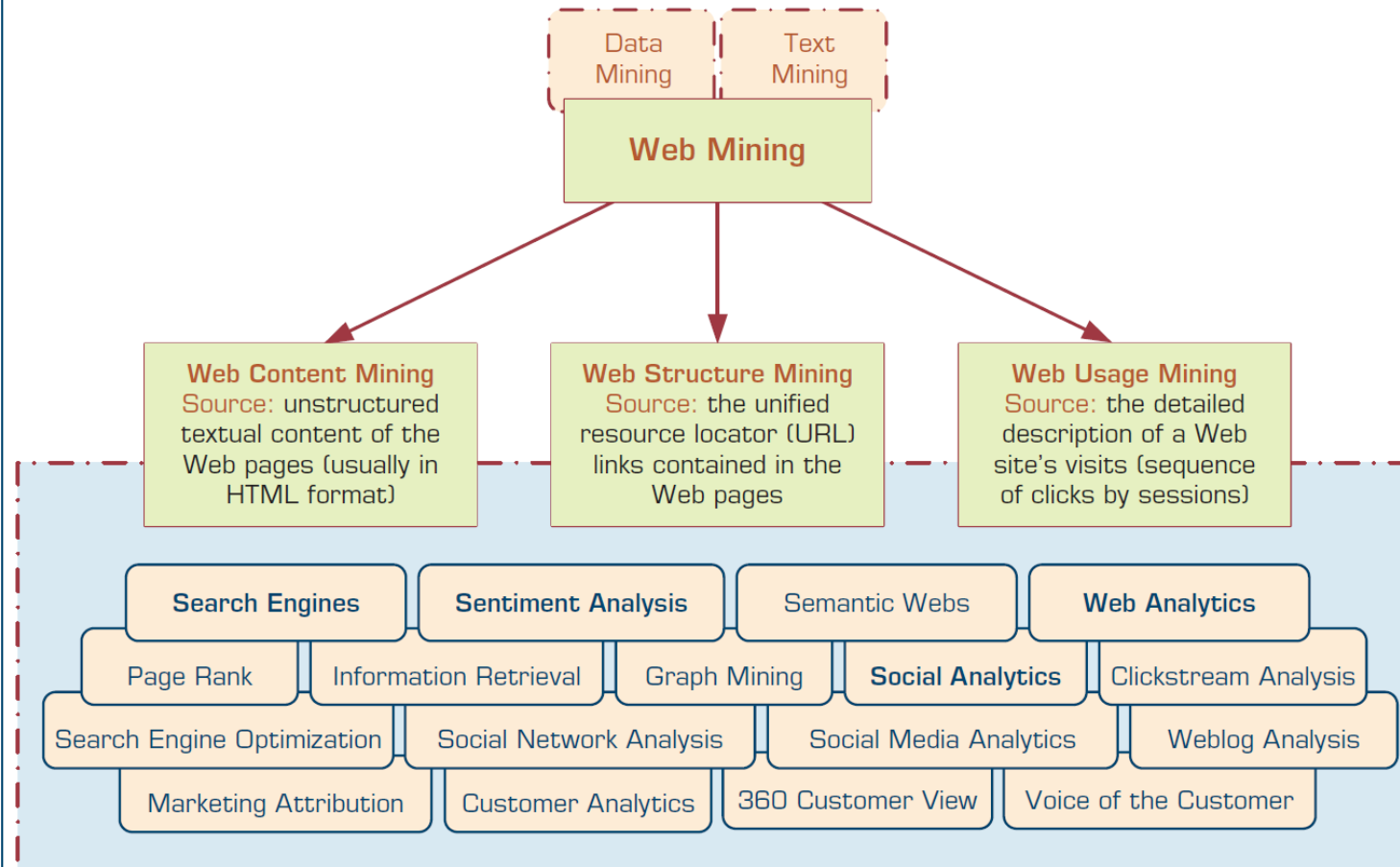


Figure 9.1 A Simple Taxonomy of Web Mining

12

Web mining

Web mining relies heavily on **data mining and text mining** and their enabling tools and techniques.

Web content mining refers to the **extraction of useful information from Web pages**. The documents may be extracted in **some machine-readable format** so that automated techniques **can extract some information from these Web pages**. Web crawlers (also called spiders) are used to **read through the content of a Web site automatically**. The information gathered may **include document characteristics similar to what is used in text mining**, but it may also **include additional concepts, such as the document hierarchy**. Such an automated process of **collecting and mining of Web content can be used for competitive intelligence** (collecting intelligence about competitors' products, services, and customers). It can also be used for **information/news/opinion collection and summarization, sentiment analysis, and automated data collection and structuring for predictive modeling**.

13

Web mining

As an illustrative example to using Web content mining as an automated data collection tool, consider the following. For more than 10 years now, two of the three authors of this book (Drs. Sharda and Delen) have been developing models to predict the financial success of Hollywood movies before their theatrical release. The data that they use for training of the models come from several Web sites, each having a different hierarchical page structure. Collecting a large set of variables on thousands of movies. (from the past several years) from these Web sites is a time-demanding, error-prone process. Therefore, they use Web content mining and spiders as an enabling technology to automatically collect, verify, validate (if the specific data item is available on more than one Web site, then the values are validated against each other and anomalies are captured and recorded), and store these values in a relational database. That way, they ensure the quality of the data while saving valuable time (days or weeks) in the process.

14

Web mining

Web content mining can also be used to enhance the results produced by search engines. In fact, search is perhaps the most prevailing application of Web content mining and Web structure mining. A search on the Web to obtain information on a specific topic (presented as a collection of keywords or a sentence) usually returns a few relevant, highquality Web pages and a larger number of unusable Web pages. Web structure mining is the process of extracting useful information from the links embedded in Web documents. It is used to identify authoritative pages and hubs, which are the cornerstones of the contemporary page-rank algorithms that are central to popular search engines such as Google and Yahoo! Just as links going to a Web page may indicate a site's popularity (or authority), links within the Web page (or the complete Web site) may indicate the depth of coverage of a specific topic.

15

Web mining

Analysis of links is very important in **understanding the interrelationships among large numbers of Web pages**, leading to a better understanding of a specific Web community, clan, or clique.

In this day and age, there is no denying the importance of Internet search engines. **As the size and complexity of the World Wide Web increase, finding what you want is becoming a complex and laborious process.** People use search engines for a variety of reasons. We use them to **learn about a product or service before committing to buy it** (including who else is selling it, what the prices are at different locations/sellers, **the common issues people are discussing about it, how satisfied previous buyers are, what other products or services might be better, etc.**), and to search for places to go, people to meet, things to do.

16

Web mining

a **search engine** is a software program that searches for documents (**Internet sites or files**), based on the keywords (**individual words, multiword terms, or a complete sentence**) users have provided, that have to do with the subject of their inquiry. **Search engines** are the workhorses of the Internet, **responding to billions of queries in hundreds of different languages every day.**

The overall goal of a search engine is to **return one or more documents/pages** (if more than one document/page applies, then a **ranked-order list is often provided**) that best match the user's query. The **two metrics** that are often used to evaluate search engines are **effectiveness (or quality—finding the right documents/pages)** and **efficiency (or speed—returning a response quickly)**. Better search engines are the **ones that excel in both at the same time.** Because search engines not only search, **but in fact find and return documents/pages**, perhaps a more appropriate name for them would have been **finding engines.**

17

Web mining

At the highest level, a search engine system is composed of two main cycles: a development cycle and a responding cycle (see the structure of a typical Internet search engine in Figure 9. 2). While one is interfacing with the World Wide Web, the other is interfacing with the user. One can think of the development cycle as a production process (manufacturing and inventorying documents/ pages) and the responding cycle as a retailing process (providing customers/users what they want).

1. Development Cycle

The two main components of the development cycle are the Web crawler and document indexer. The purpose of this cycle is to create a huge database of documents/pages organized and indexed based on their content and information value.

18

Web mining

1. Development Cycle

The reason for developing such a repository of documents/pages is quite obvious: Due to its sheer size and complexity, therefore, search engines “cache the Web” into their database and use the cached version of the Web for searching and finding. Once created, this database allows search engines to rapidly and accurately respond to user queries.

Web crawler starts with a list of URLs to visit, which are listed in the scheduler and often are called the seeds. These URLs may come from submissions made by Webmasters or, more often, they come from the internal hyperlinks of previously crawled documents/pages. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit (i.e., the scheduler).

19

Web mining

Because there are large volumes of Web pages, the crawler can only download a limited number of them within a given time; therefore, it may need to prioritize its downloads.

Document Indexer As the documents are found and fetched by the crawler, they are stored in a temporary staging area for the document indexer to grab and process. The document indexer is responsible for processing the documents (Web pages or document files) and placing them into the document database.

20

Web mining

2. Response Cycle: The two main components of the responding cycle are the **query analyzer and document matcher/ranker**.

The **query analyzer** is responsible for receiving a search request from the user (via the search engine's Web server interface) and converting it into a **standardized data structure**, so that it can be **easily queried/matched** against the entries in the document database. How the query analyzer does what it is supposed to do is quite similar to what the document indexer does . The query analyzer parses the search string into individual words/terms using a series of tasks. Then is converting a **query string** into the same structure so that it can be used to quickly locate most relevant documents/ pages.

21

Web mining

Document Matcher /Ranker This is where the structured query data is matched against the document database to find the most relevant documents/pages and also rank them in the order of relevance/importance. The proficiency of this step is perhaps the most important component when different search engines are compared to one another. Every search engine has its own algorithm that it uses to carry out this important step.

Search engine optimization (SEO) is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results. In general, the higher ranked on the search results page, and the more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users.

22

Web mining

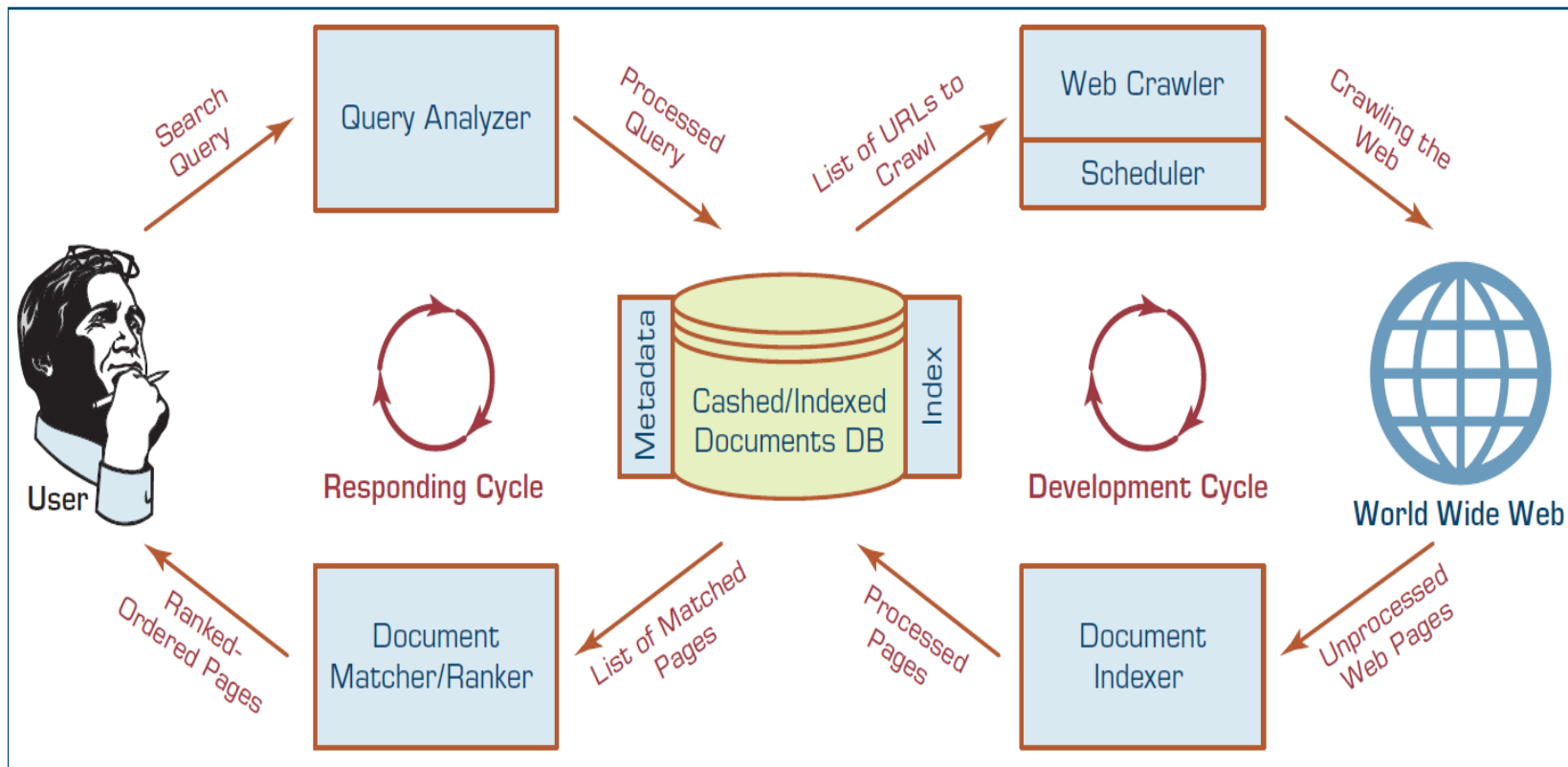


Figure 9. 2
Structure of a
Typical Internet
Search Engine

23

Web mining

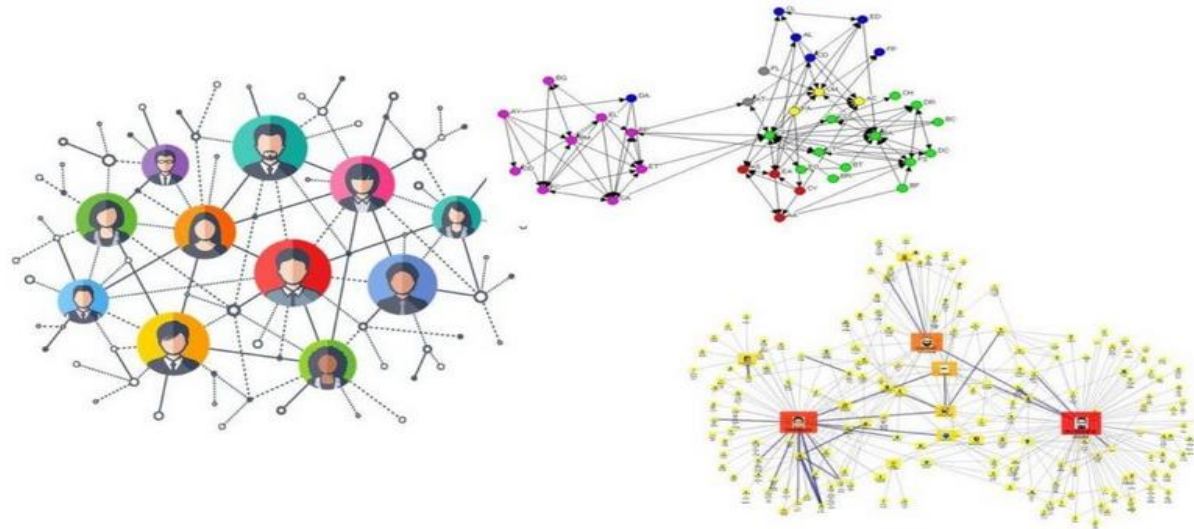
Here are the 15 most popular search engines as derived from eBizMBA Rank (ebizmba.com/articles/search-engines), which is a constantly updated average of each Web site's *Alexa* Global Traffic Rank, and U.S. Traffic Rank from both Compete and Quantcast.

Rank	Name	Estimated Unique Monthly Visitors
1	Google	1,600,000,000
2	Bing	400,000,000
3	Yahoo! Search	300,000,000
4	Ask	245,000,000
5	AOL Search	125,000,000
6	Wow	100,000,000
7	WebCrawler	65,000,000
8	MyWebSearch	60,000,000
9	Infospace	24,000,000
10	Info	13,500,000
11	DuckDuckGo	11,000,000
12	Contentko	10,500,000
13	Dogpile	7,500,000
14	Alhea	4,000,000
15	ixQuick	1,000,000

Top 15 Most Popular
Search Engines
(August 2016)

A **social network** is a social structure composed of **individuals/people** (or groups of individuals or organizations) linked to one another with some type of **connections/relationships**. The social network perspective provides a holistic approach to analyzing the structure and dynamics of social entities.

Social Network Analysis



25

Social Network Analysis

Social Network Analysis (SNA) is a method for mapping and measuring the relationships and interactions among various entities, which can include individuals, groups, and even digital elements like computers or URLs. It uses tools from graph theory, statistics, and data science to model and understand patterns of interaction, influence, communication, and behavior.

SNA is widely used in fields like:

- Marketing
- Public health
- Cybersecurity
- Political science
- Social media analysis

26

Social Network Analysis

Real-World Applications of Social Network Analysis (SNA):**1. Marketing & Influencer Identification**

- ◆ **Objective:** Find **key individuals who can influence the behavior of others**—especially in **social media** .
- ◆ **How SNA Helps:**
 - **Identify influencers** using centrality metrics .
 - Detect **information diffusion paths** (how product messages spread).
- ◆ **Example:** A cosmetics brand analyzes Instagram or TikTok networks to identify top influencers whose endorsements drive purchases.

27

Social Network Analysis

2. Public Health & Epidemiology

◆ Objective:

Track the **spread of infectious diseases** or behaviors (e.g., smoking, vaccination adoption) within communities.

◆ How SNA Helps:

- Identify **super-spreaders** in disease outbreaks.
- Model **contact tracing** networks to isolate risk zones.
- Understand **behavioral influence** (e.g., health beliefs spreading through family or peer groups).

◆ Example:

During COVID-19, SNA was used to identify and isolate hubs in transmission networks in hospitals or communities.

28

Social Network Analysis

For organizations, small or large, there is valuable insight hidden in all the user-generated content on social media sites. But how do you dig it out of dozens of review sites, thousands of blogs, millions of Facebook posts, and billions of tweets? Once you do that, how do you measure the impact of your efforts? These questions can be addressed by the analytics extension of the social media technologies. Once you decide on your goal for social media (what it is that you want to accomplish), there is a multitude of tools to help you get there. These analysis tools usually fall into three broad categories:

- **Descriptive analytics:** Uses simple statistics to identify activity characteristics and trends, such as how many followers you have, how many reviews were generated on Facebook, and which channels are being used most often.
- **Social network analysis:** Follows the links between friends, fans, and followers to identify connections of influence as well as the biggest sources of influence.
- **Advanced analytics:** Includes predictive analytics and text analytics that examine the content in online conversations to identify themes, sentiments, and connections that would not be revealed by casual surveillance.

Challenges :

1. **Data Collection and Quality:** Collecting comprehensive and reliable network data can be challenging, especially in large-scale networks. Ethical considerations regarding privacy and consent also need to be addressed.
2. **Complexity and Interpretation:** Analyzing complex networks and deriving meaningful insights can be challenging. Understanding causality and the dynamics of relationships requires advanced analysis techniques.

30

Business tools

What is Tableau?

- A data visualization and analytics solution that assists enterprises in making data-driven business decisions.
- It blends information from a wide range of sources to deliver actionable, real-time insights.
- It allows exploration of data via intuitive means such as drag-and-drop filtering and natural language queries, irrespective of skill levels.



31

Business tools

Why use Tableau?

- Powerful Visualization
- Explore any type of data sources
- Multi-Layered Security

Some key features

- Drag and Drop Visualizations
- Augmented Data Preparation
- Geographic Data Analysis
- Real-time Access & Role-Based Permission

32

Business tools

Installation of Tableau Public:

<https://www.tableau.com/products/desktop/download>

Tableau Desktop: Start your free 14-day trial

Almost there!

Complete the form to start your free trial. Our team will be in touch to help you make the most of your trial.

33

Business tools

Installation of Tableau Public:

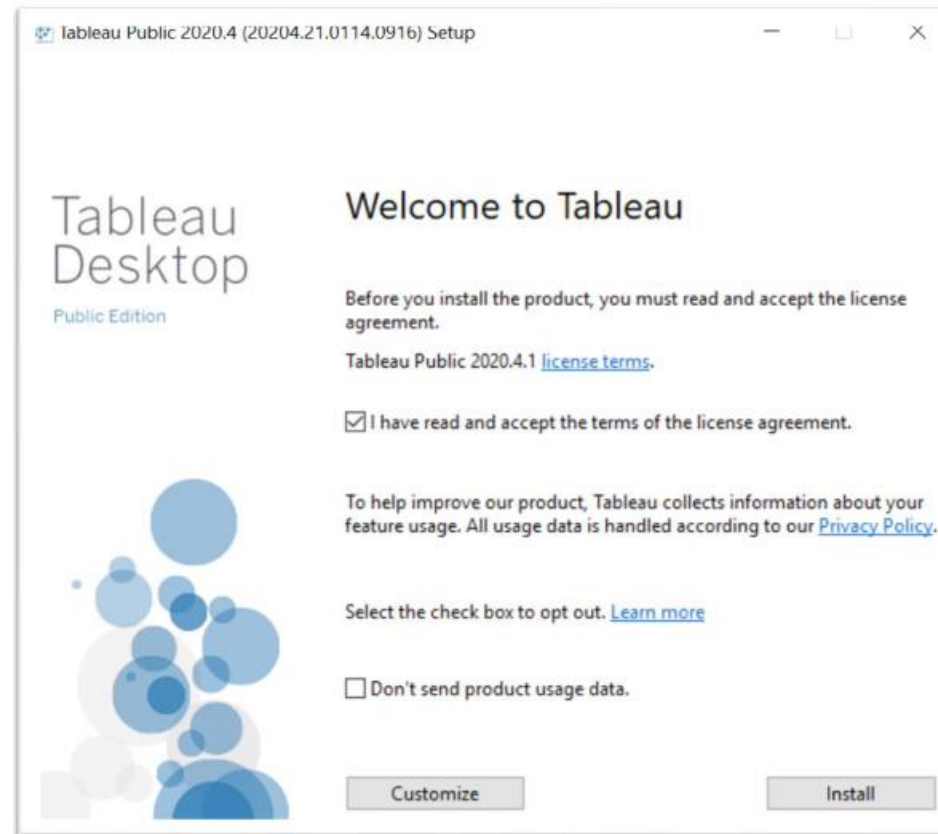
<https://www.tableau.com/products/public/download>

You'll be exploring in minutes

Connect to the data that matters most to you with Tableau's intuitive drag-and-drop interface to effortlessly create interactive graphs, stunning maps, and live dashboards in minutes. Save your work to your local desktop or publish your viz to your Tableau Public profile. Anyone can do it, it's that easy—and it's free.

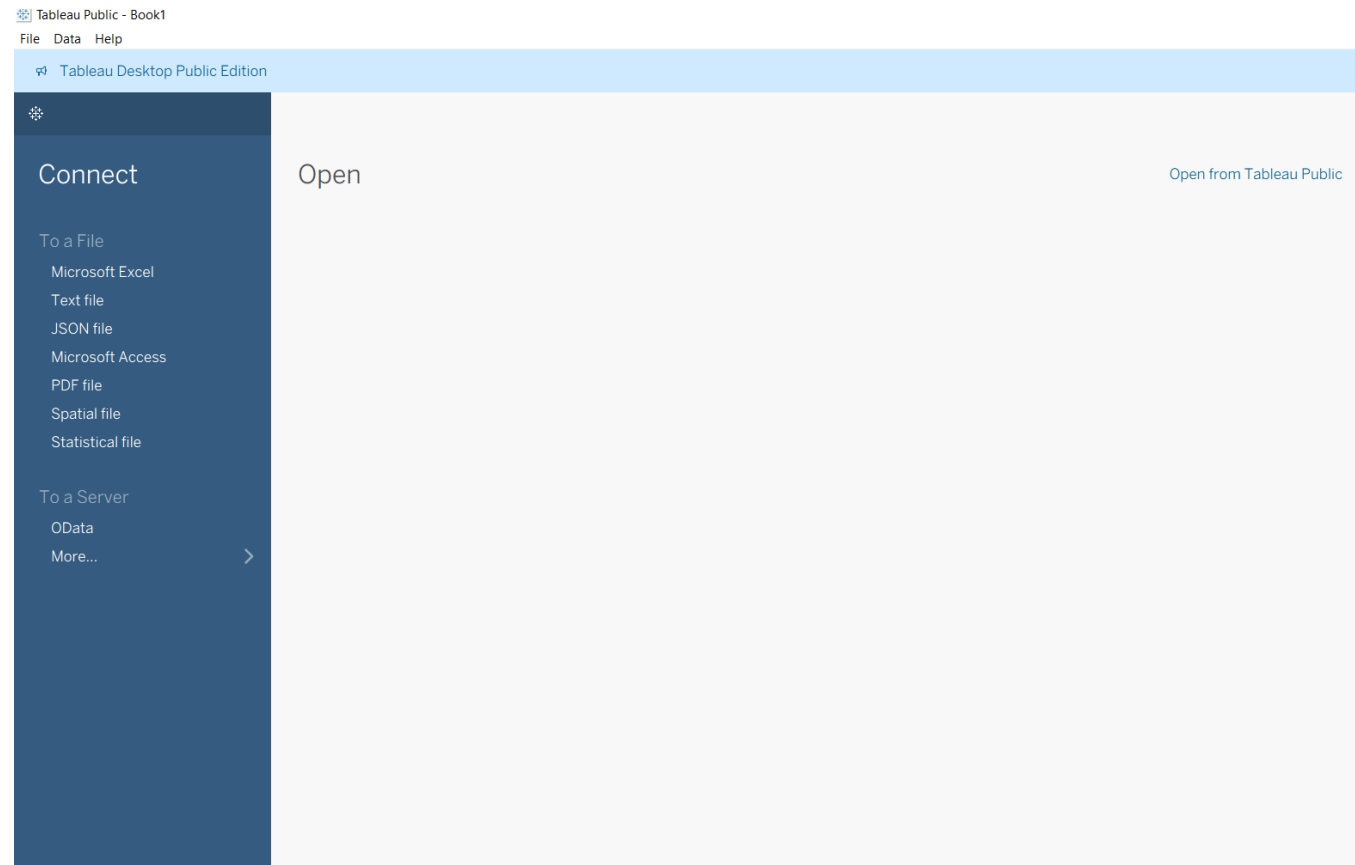
Almost there!

It only takes 15 seconds to fill out. If you're already registered, [sign in](#).



35

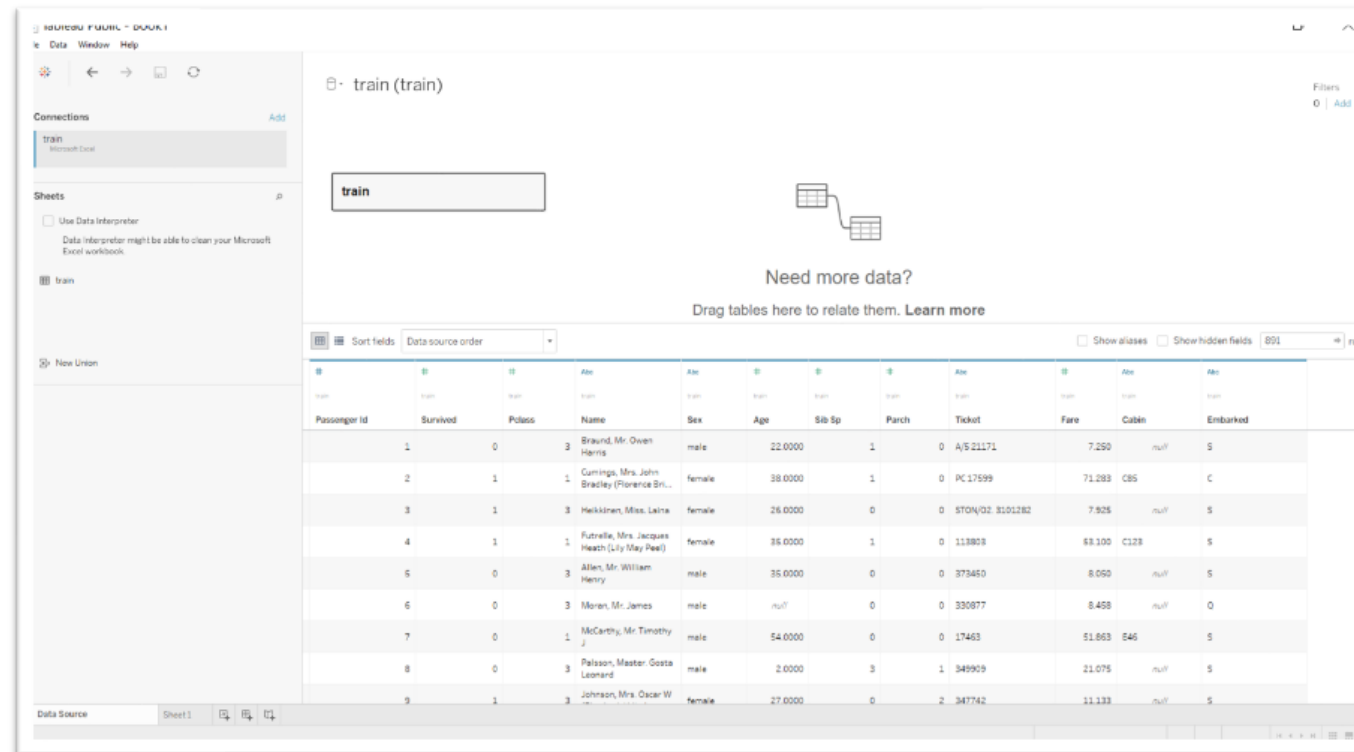
Business tools



36

Business tools

Tableau can connect to various types of data sources.



train (train)

train

Need more data?
Drag tables here to relate them. [Learn more](#)

Sort fields: Data source order

Show aliases Show hidden fields 891 rc

Passenger Id	Survived	Pclass	Name	Sex	Age	Sib Sp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0000	1	0	A/5 21171	7.250	null	S
2	1	1	Cummings, Mrs. John Bradley (Florence Briggs)	female	38.0000	1	0	PC 17599	71.283	C85	C
3	1	3	Hickson, Miss. Laina	female	26.0000	0	0	STON/O2 3101282	7.925	null	S
4	1	1	Patel, Mrs. Jacques Heath (Lily May Peel)	female	35.0000	1	0	113803	53.100	C128	S
5	0	3	Allen, Mr. William Henry	male	35.0000	0	0	373450	8.050	null	S
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.458	null	Q
7	0	1	McCarthy, Mr. Timothy J	male	54.0000	0	0	17463	51.863	846	S
8	0	3	Palsson, Master. Gosta Leonard	male	2.0000	3	1	349909	21.075	null	S
9	1	3	Johnson, Mrs. Oscar W	female	27.0000	0	2	347742	11.133	null	S

37

Summary

1. Web mining.
2. Web mining process.
3. Tableau



Thank You