Tikrit University
Computer Science Dept.

Master Degree
Lecture -1-

Associate Professor

Dr. Eng. Zaidoon.T.AL-Qaysi

College of Computer Science and Mathematics

(2023-2024)

- **Data science** is the art and science of acquiring knowledge through data.

- **Data science** is a collection of techniques used to extract value from data. It has become an essential tool for any organization that collects, stores, and processes data as part of its operations. Data science techniques rely on finding useful patterns, connections, and relationships within data.

- **Data science** is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics. It is an interdisciplinary field that extracts value from data. In the context of how data science is used today, it relies heavily on machine learning and is sometimes called data mining. Examples of data science user cases are: recommendation engines that can recommend movies for a particular user, a fraud alert model that detects fraudulent credit card transactions, find customers who will most likely churn next month, or predict revenue for the next quarter.

- **Data science** starts with data, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables. Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset. The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI).

- **Data science** involves inference and iteration of many different hypotheses. One of the key aspects of data science is the process of generalization of patterns from a dataset. The generalization should be valid, not just for the dataset used to observe the pattern, but also for new unseen data.

- **Data Science tasks:** Data Science-related activities are broadly classified into predictive and descriptive tasks. The former deals with novel inferences based on acquired knowledge and the latter describes the inherent patterns hidden inside data.

    1. **Predictive tasks** apply supervised Machine Learning to predict the future by learning from past experiences. Examples of predictive analysis are classification, regression, and deviation detection.

    2. **Descriptive Data Science:** Descriptive analysis is also termed exploratory data analysis. Unlike predictive models, descriptive models avoid inference, but analyze historical or past data at hand and present the data in a more interpretable way such that it may better convey the underlying configuration of the data elements. Leveraging powerful visualization techniques usually enhances descriptive analysis for better interpretation of data. A descriptive task may act as the starting point to prepare data for further downstream analysis.

- **Diagnostic Data Science:** The diagnostic analysis builds on the outcomes of descriptive analysis to investigate the root causes of a problem. It includes processes such as data discovery, data mining, drilling down, and computing data correlations to identify potential sources of data anomalies. Probability theory, regression analysis, and time-series analysis are key tools for diagnostic analysis. For example, Amazon can drill down the sales and profit numbers to various product categories to identify why sales have gone down drastically in a certain span of time in certain markets.

- **Prescriptive Data Science**: This is the most recent addition to analysis tasks. A prescriptive model suggests guidelines or a follow-up course of action to achieve a goal. It uses an understanding of what has happened, why it has happened, and what might happen to suggest the best possible alternatives. Unlike predictive analysis, prescriptive analysis does not need to be perfect but suggests the “best” possible way toward the future.

Google Maps is an excellent prescriptive model that suggests the best route from the current location to the destination on the fly, considering traffic conditions and the shortest route.

- **Data Science objectives:** Data Science aims to achieve four basic objectives: (i) to extract interesting patterns in data without human intervention, (ii) to predict the most likely future outcomes based on past data, (iii) to create actionable knowledge to achieve certain goals, and (iv) to focus on how to handle voluminous data in achieving the previous three objectives.

- **Understanding data science begins with three basic areas:**
  1. Math/statistics: This is the use of equations and formulas to perform analysis
  2. Computer programming: This is the ability to use code to create outcomes on the computer
  3. Domain knowledge: This refers to understanding the problem domain (medicine, finance, social science, and so on).

- A **data model** refers to an organized and formal relationship between elements of data, usually meant to simulate a real-world phenomenon.
- A **program**, a set of instructions to a computer, transforms input signals into output signals using predetermined rules and relationships.

- **Data mining** is the process of finding relationships between elements of data. Data mining is the part of data science where we try to find relationships between variables (think spawn-recruit model).

- **Machine learning** can either be considered a sub-field or one of the tools of artificial intelligence is providing machines with the capability of learning from experience. Experience for machines comes in the form of data. Data that is used to teach machines is called training data. Machine learning turns the traditional programing model upside down (Figure 1). Speaking of data models, we will concern ourselves with the following two basic types of data models:

  1. **Probabilistic model:** This refers to using probability to find a relationship between elements that includes a degree of randomness.
  2. **Statistical model:** This refers to taking advantage of statistical theorems to formalize relationships between data elements in a (usually) simple mathematical formula.

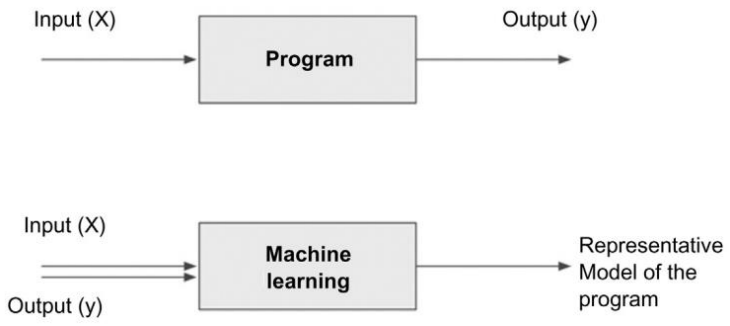

Figure 1: Traditional program and machine learning.

Modeling is a process in which a representative abstraction is built from the observed dataset. For example, based on credit score, income level, and requested loan amount, a model can be developed to determine the interest rate of a loan. For this task, previously known observational data including credit score, income level, loan amount, and interest rate are needed. **Figure 2** shows the process of generating a model. Once the representative model is created, it can be used to predict the value of the interest rate, based on all the input variables.

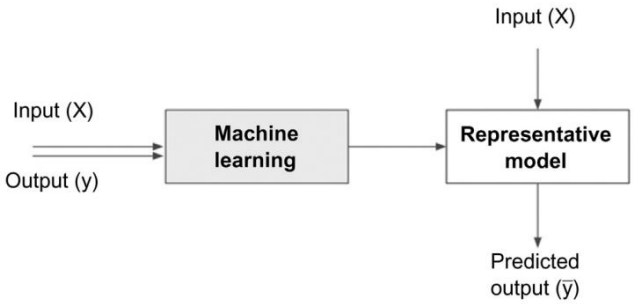

Figure 2: Data science models

- In the center of the AI revolution, **deep learning technology** is a set of machine learning techniques that learn multiple layers of neural networks for supporting prediction, classification, clustering, and data generation tasks. The success of deep learning comes from:
  1. **Data:** Large amounts of rich data, especially in images and natural language texts, become available for training deep learning models.
  2. **Algorithms:** Efficient neural network methods have been proposed and enhanced by many researchers in recent years.
  3. **Hardware:** Advances in parallel computing, especially graphic process units (GPUs), have enabled a fast and affordable computing engine for deep learning workload.
  4. **Software:** Scalable and easy-to-use programming frameworks have been developed and released via open source projects to the public. Most of them, including TensorFlow and Pytorch, have strong support from the technology industry.

- **Descriptive statistics:** Computing mean, standard deviation, correlation, and other descriptive statistics, quantify the aggregate structure of a dataset. This is essential information for understanding any dataset in order to understand the structure of the data and the relationships within the dataset. They are used in the exploration stage of the data science process.

- **Exploratory visualization:** The process of expressing data in visual coordinates enables users to find patterns and relationships in the data and to comprehend large datasets. Similar to descriptive statistics, they are integral in the pre- and post-processing steps in data science.

- **Data engineering:** Data engineering is the process of sourcing, organizing, assembling, storing, and distributing data for effective analysis and usage. Database engineering, distributed storage, and computing frameworks (e.g., Apache Hadoop, Spark, Kafka), parallel computing, extraction transformation and loading processing, and data warehousing constitute data engineering techniques. Data engineering helps source and prepare for data science learning algorithms.

- **Business intelligence:** Business intelligence helps organizations consume data effectively. It helps query the ad hoc data without the need to write the technical query command or use dashboards or visualizations to communicate the facts and trends. Business intelligence specializes in the secure delivery of information to right roles and the distribution of information at scale. Historical trends are usually reported, but in combination with data science, both the past and the predicted future data can be combined. BI can hold and distribute the results of data science.

- Of course, most of us have learned these terms and realize that **"data is the new oil**." The most important task that organizations and businesses have employed in the last decade to utilize their data and understand and employ this information is for making better informed decisions. In fact, with big developments in technology, a successful environment has been created around fields such as machine learning, artificial intelligence, and deep learning. Researchers, engineers, and data scientists have created frameworks, tools, techniques, algorithms, and methodologies to achieve intelligent systems and models that can automate tasks, detect anomalies, perform complex analyses, and predict events.

- **Data science problems** can also be classified into tasks such as: classification, regression, association analysis, clustering, anomaly detection, recommendation engines, feature selection, time series forecasting, deep learning, and text mining (Figure 3).

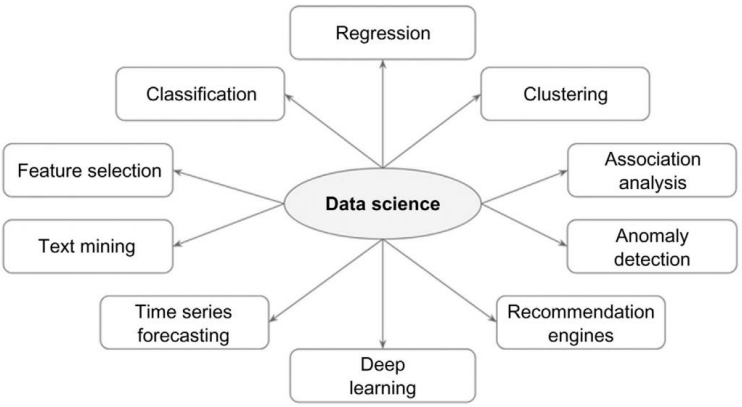

Figure 3: Data science tasks

- The methodical discovery of useful relationships and patterns in data is enabled by a set of iterative activities collectively known as the data science process. The standard **data science process** involves:
  1. Understanding the problem.
  2. Preparing the data samples.
  3. Developing the model.
  4. Applying the model on a dataset to see how the model may work in the real world.
  5. Deploying and maintaining the models.

- Understanding how the data is **collected, stored, transformed, reported,** and **used** is essential to the data **science process**. This part of the step surveys all the data available to answer the business question and narrows down the new data that need to be sourced. There are quite a range of factors to consider: quality of the data, quantity of data, availability of data, gaps in data, and does lack of data compel the practitioner to change the business question, etc. The objective of this step is to come up with a dataset to answer the business question through the data science process. It is critical to recognize that an inferred model is only as good as the data used to create it.

The data science process presented in **Figure 4**, is a generic set of steps that is problem, algorithm, and, data science tool agnostic. The fundamental objective of any process that involves data science is to address the analysis question. The problem at hand could be a segmentation of customers, a prediction of climate patterns, or a simple data exploration. The learning algorithm used to solve the business question could be a decision tree, an artificial neural network, or a scatterplot. The software tool to develop and implement the data science algorithm used could be custom coding, RapidMiner, R, Weka, SAS, Oracle Data Miner, Python, etc.

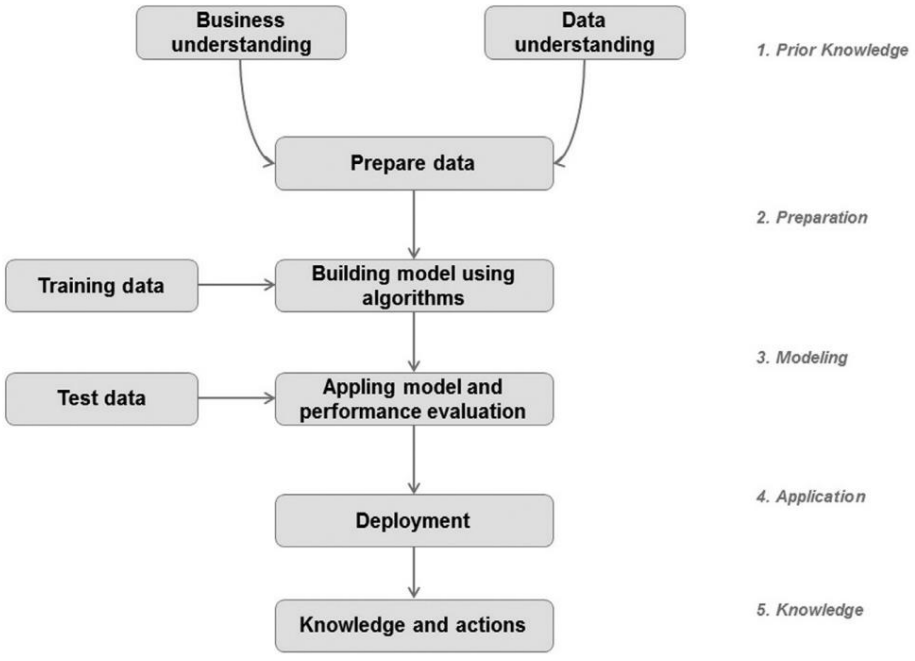

Figure 4: Data Science Process

Prior knowledge refers to information that is already known about a subject. The prior knowledge step in the data science process helps to define what problem is being solved, how it fits in the business context, and what data is needed in order to solve the problem.

- Some of the terminology used in the data science process are:
  - A **dataset** (example set) is a collection of data with a defined structure. Table 1 shows a dataset. It has a well-defined structure with 10 rows and 3 columns along with the column headers. This structure is also sometimes referred to as a "data frame".
  - A **data point** (record, object or example) is a single instance in the dataset. Each row in Table 1 is a data point. Each instance contains the same structure as the dataset.
  - An **attribute** (feature, input, dimension, variable, or predictor) is a single property of the dataset. Each column in Table 1 is an attribute. Attributes can be numeric, categorical, date-time, text, or Boolean data types. In this example, both the credit score and the interest rate are numeric attributes.
  - A **label** (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes. In Table 1, the interest rate is the output variable.
  - **Identifiers** are special attributes that are used for locating or providing context to individual records. For example, common attributes like names, account numbers, and employee ID numbers are identifier attributes. Identifiers are often used as lookup keys to join multiple datasets. They bear no information that is suitable for building data science models and should, thus, be excluded for the actual modeling step. In Table 1, the attribute ID is the identifier.

Table .1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

- Depending on its nature, data is stored in various formats such as
  1. CSV (Comma-Separated Values).
  2. TSV (Tab-Separated Values).
  3. XML (eXtensible Markup Language).
  4. RSS (Really Simple Syndication).
  5. JSON (JavaScript Object Notation)

- Data Sources:
  1. **Healthcare Data:** Among all healthcare technologies, electronic health records (EHRs) had vast adoption and a huge impact on healthcare delivery in recent years. One important benefit of EHRs is to capture all the patient encounters with rich multi-modality data. Healthcare data include both structured and unstructured information. Structured data include various medical codes for diagnoses and procedures, lab results, and medication information. Unstructured data contain (1) clinical notes as text, (2) medical imaging data such as X-rays, echocardiogram, and magnetic resonance imaging (MRI), and (3) time-series data such as the electrocardiogram (ECG) and electroencephalogram (EEG). Beyond the data collected during clinical visits, patient self-generated/reported data start to grow thanks to wearable sensors' increasing use.

  2. **Social media Data:** Social media has become a gold mine for collecting data to analyze for research or marketing purposes. This is facilitated by the Application Programming Interface (API) that social media companies provide to researchers and developers. Think of the API as a set of rules and methods for asking and sending data. For various data-related needs (e.g., retrieving a user's profile picture), one could send API requests to a particular social media service. This is typically a programmatic call that results in that service sending a response in a structured data format, such as an XML. The Facebook Graph API is a commonly used example.7 These APIs can be used by any individual or organization to collect and use this data to accomplish a variety of tasks, such as developing new socially impactful applications, research on human information behavior, and monitoring the aftermath of natural calamities, etc.

  3. **Multimodal Data:** We are living in a world where more and more devices exist – from light bulbs to cars – and are getting connected to the Internet, creating an emerging trend of the Internet of Things (IoT). These devices are generating and using much data, but not all of which are "traditional" types (numbers, text). When dealing with such contexts, we may need to collect and explore multimodal (different forms) and multimedia (different media) data such as images, music and other sounds, gestures, body posture, and the use of space. Once the sources are identified, the next thing to consider is the kind of data that can be extracted from those sources.

- Google has developed a dedicated dataset search engine, **Google Dataset** Search that helps researchers search and download publicly available databases. Google claims that its Dataset Search engine has indexed about 25 million datasets, and one may obtain useful information about them. To locate these datasets in the search results, Google Dataset Search leverages schema.org and other metadata standards.

- **Open Data:** The idea behind open data is that some data should be freely available in a public domain that can be used by anyone as they wish, without restrictions from copyright, patents, or other mechanisms of control. Following is the list of principles associated with open data as observed in the policy document:

1. **Public.** Agencies must adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions.
2. **Accessible.** Open data are made available in convenient, modifiable, and open formats that can be retrieved, downloaded, indexed, and searched. Formats should be machinereadable (i.e., data are reasonably structured to allow automated processing). Open data structures do not discriminate against any person or group of persons and should be made available to the widest range of users for the widest range of purposes, often by providing the data in multiple formats for consumption. To the extent permitted by law, these formats should be non-proprietary, publicly available, and no restrictions should be placed on their use.

3. **Described.** Open data are described fully so that consumers of the data have sufficient information to understand their strengths, weaknesses, analytical limitations, and security requirements, as well as how to process them. This involves the use of robust, granular metadata (i.e., fields or elements that describe data), thorough documentation of data elements, data dictionaries, and, if applicable, additional descriptions of the purpose of the collection, the population of interest, the characteristics of the sample, and the method of data collection.

4. **Reusable.** Open data are made available under an open license that places no restrictions on their use.

5. **Complete.** Open data are published in primary forms (i.e., as collected at the source), with the finest possible level of granularity that is practicable and permitted by law and other requirements. Derived or aggregate open data should also be published but must reference the primary data.

6. **Timely.** Open data are made available as quickly as necessary to preserve the value of the data. Frequency of release should account for key audiences and downstream needs.

7. **Managed Post-Release.** A point of contact must be designated to assist with data use and to respond to complaints about adherence to these open data requirements.\

- Example of Open Data Sources
  1. **UCI ML Repository**: The machine-learning repository at the University of California, Irvine (UCI), is a great place to find open-source and cost-free datasets for machinelearning experiments, covering a wide range of domains from biology to particle physics. It houses 622 datasets for machine-learning research. The datasets contain attributes that are of Categorical, Integer, and Real types. The datasets have been curated and cleaned. The repository includes only tabular data, primarily for classification tasks. This is limiting for anyone interested in natural language, computer vision, and other types of data.

2. **Kaggle:** This platform, recently acquired by Google, publishes datasets for machinelearning experiments. It also offers GPU-integrated notebooks to solve Data Science challenges. At present, it houses approximately 77k datasets. Kaggle fosters advancements in machine learning through open community competitions. The datasets on Kaggle are divided into three categories: general-public datasets, private-competition datasets, and public-competition datasets. Except for the private datasets, all others are free to download. A majority of datasets are based on real-life use cases.

3. **Awesome Public Datasets:** This is a public repository of datasets covering 30 topics and tasks in diverse domains hosted in GitHub. The large datasets can be used to perform BigData-related tasks. The quality and uniformity of the datasets are high.

4. **NCBI:** This is an open-access large collection of databases for exploratory molecularbiology data analysis provided by the National Center for Biotechnology Information (NCBI) in the USA. Databases are grouped into six categories: Literature, Health, Genomes, Genes, Proteins, and Chemicals. It is a rich source of sequences, annotations, metadata, and other data related to genes and genomes.

5. **SNAP:** The Stanford Network Analysis Platform (SNAP) offers large network datasets. It covers large graphs derived from 80+ categories of domains, such as social networks, citation networks, web graphs, online communities, and online reviews. Libraries implementing more than 140 graph algorithms are available for download in SNAP. These algorithms can effectively process the characteristics and metadata of nodes and edges, manipulate massive graphs, compute structural properties, produce regular and random graphs, and generate nodes and edges.

6. **MNIST:** The Modified National Institute of Standards and Technology (MNIST) dataset is a large collection of hand-written digits, popular among deep-learning beginners and the image-processing community. The MNIST database contains 60 000 training images and 10 000 testing images. Extended MNIST (EMNIST) is a more refined database with 28×28 pixel-sized images. EMNIST is a dataset with balanced and unbalanced classes. MNIST is a very popular dataset for machine-learning experiments.

7. **VoxCeleb Speech Corpus:** This is a large collection of audiovisual data for speaker recognition. It includes more than one million real-world utterances from more than 6000 speakers. The dataset is available in two parts, VoxCeleb1 and VoxCeleb2. In VoxCeleb1, there are over 100 000 utterances for 1251 celebrities; in VoxCeleb2, there are over 1 million utterances for more than 6000 celebrities taken from YouTube videos. Different development and test sets are segregated with different speakers.

Data Science Applications

- **Healthcare:** Recent advances in Data Science are a blessing to healthcare and allied sectors. Data Science has positively and extensively impacted upon these application areas, in turn helping mankind significantly. Health informatics and smart biomedical devices are pushing medical and health sciences to the next level. Precision medicine is likely to be a game changer in extending the human life span. Computer-vision and biomedical technologies are making quick and accurate disease diagnosis ubiquitous. Even a handheld smartphone may be able to continuously monitor the health metrics of a person and generate smart early alarms when things go wrong.

- **Computational Biology:** With the availability of high-throughput omits data, it is now possible to understand the genetic causes of many terminal diseases. Computational biology and bioinformatics mine massive amounts of genomic and proteomic data to better understand the causes of diseases. Once the causes are identified with precision, it becomes possible to develop appropriate drug molecules for treatment. On average, traditional drug development requires more than 14 years of effort, which can now be reduced drastically due to effective Data Science techniques. Precision medicine is the future of drug technology, customizing drugs for the individual and avoiding the onesize-fits-all approach that does not always work.

- **Business:** The rise of Data Science was originally intended to benefit business sectors. With the need for business intelligence, the use of data analytics has gained momentum. Almost every business venture is investing significant resources in smart business decision making with Data Science. Analyzing customer purchasing behavior is a great challenge, and important for improving revenue and profit. Integration of heterogeneous data is an effective way to promote sales of products. Data-analysis experts apply statistical and visualization tools to understand the moods, desires, and wants of customers and suggests effective ways for business and product promotion and plans for progress and expansion.

- **Smart Devices:** A mobile device is no longer just a communication device, but rather a miniature multipurpose smart tool that enhances the lifestyle of the common man or woman. Apps installed in smart devices like smartphones or smart tablets can be used to understand a person well in regard to their choices, preferences, likes, and dislikes. Technology is becoming so personalized that installed apps in such devices will be able to predict a user's actions in advance and make appropriate recommendations. In the near future, smart devices will be able to monitor user health status, recommend doctors, book appointments, place orders for medicine, and remind users of medicine schedules. The Internet-of-Things (IoT) and speaking smart devices make our life easier. For example, a home automation system can monitor and control an entire home remotely with the help of a smartphone, starting from kitchen to home security.

- **Transportation:** Another important application area is smart transportation. Data Science is actively involved in making it possible to take impressive steps toward safe and secure driving. The driverless car will be a big leap into the future in the automobile 12 Fundamentals of Data Science sector. The ability to analyze fuel-consumption patterns and driving styles and monitor vehicle status makes it possible to create optimized individual driving experiences, spurring new designs for cars by manufacturers. A transportation company like

Uber uses Data Science for price optimization and offers better riding experiences to customers. In addition to the above, the list of applications is huge and is growing every day. There are other industry sectors like banking, finance, manufacturing, e-commerce, internet, gaming, and education that also use Data Science extensively.

- **Public Policy:** Simply put, public policy is the application of policies, regulations, and laws to the problems of society through the actions of government and agencies for the good of a citizenry. Many branches of social sciences (economics, political science, sociology, etc.) are foundational to the creation of public policy. Data science helps governments and agencies gain insights into citizen behaviors that affect the quality of public life, including traffic, public transportation, social welfare, community wellbeing, etc. This information, or data, can be used to develop plans that address the betterment of these areas. It has become easier than ever to obtain useful data about policies and regulations to analyze and create insights. The following open data repositories are examples:

(1) US government (https://www.data.gov/)
(2) City of Chicago (https://data.cityofchicago.org/)
(3) New York City (https://nycopendata.socrata.com/)

- **Politics:** Politics is a broad term for the process of electing officials who exercise the policies that govern a state. It includes the process of getting policies enacted and the action of the officials wielding the power to do so. For instance, data scientists analyzed former US President Obama's 2008 presidential campaign success with Internet-based campaign efforts. Data scientists have been quite successful in constructing the most accurate voter targeting models and increasing voter participation.11 In 2016, the campaign to elect Donald Trump was a brilliant example of the use of data science in social media to tailor individual messages to individual people. As Twitter has emerged as a major digital PR tool for politics over the last decade, studies12 analyzing the content of tweets from both candidates' (Trump and Hillary Clinton) Twitter handles as well as the content of their websites found significant difference in the emphasis on traits and issues, main content of tweet, main source of retweet, multimedia use, and the level of civility.

- **Urban Planning :** Many scientists and engineers have come to believe that the field of urban planning is ripe for a significant – and possibly disruptive – change in approach as a result of the new methods of data science. This belief is based on the number of new initiatives in "informatics" – the acquisition, integration, and analysis of data to understand and improve urban systems and quality of life. The Urban Center for Computation and Data (UrbanCCD), at the University of Chicago, traffics in such initiatives. The research center is using advanced computational methods to understand the rapid growth of cities. The center brings together scholars and scientists from the University of Chicago and Argonne National Laboratory18 with architects, city planners, and many others.

- **Education:** Technology will definitely have a large part to play in the future of education, but how exactly that happens is still an open question. There is a growing realization among educators and technology evangelists that we are heading toward more data-driven and personalized use of technology in education. And some of that is already happening. Students can improve their reading skills by reading short stories, taking a test every other week, and receiving graded papers from teachers. Also , students could learn to read through "a computerized software program," the computer constantly measuring and collecting data, linking to websites providing further assistance, and giving the student instant feedback. "At the end of the session," the teacher will receive an automated readout on [students in the class] summarizing their reading time, vocabulary knowledge, reading comprehension, and use of supplemental electronic resources. So, in essence, teachers of the future will be data scientists.

- **Libraries:** Data science is also frequently applied to libraries. Even though the role of data science in future libraries seems too rosy to be true, in reality it is nearer than you think. Imagine that Alice, a scientist conducting research on diabetes, asks Mark, a research librarian, to help her understand the research gap in previous literature. Armed with the digital technologies, Mark can automate literature reviews for any discipline by reducing ideas and results from thousands of articles into a cohesive bulleted list and then apply data science algorithms, such as network analysis, to visualize trends in emerging lines of research on similar topics. This will make Alice's job far easier than if she had to painstakingly read all the articles.