LECTURE 1: INTRODUCTION TO OPEN SOURCE TOOLS AND TERMS FOR DATA SCIENCE

1.1 WHAT ARE OPEN-SOURCE TOOLS AND WHY ARE THEY IMPORTANT FOR DATA SCIENCE

Open-source tools are tools that are available for free online thus you don't need to buy them

Pros of Open-source Software in Data Analytics

Data analytics can be expensive. However, open-source software saves you money while getting the job done. With its low cost, flexibility, and vibrant community, open-source software has become a game-changer in data analytics. Here we'll explore the pros of open-source software in data analytics and show you how it can help your business thrive

1. Free to use and distribute

Using open-source software for data analytics has several benefits. The fact that it is free to use, and share is undoubtedly a benefit. This implies that no license is required to utilize it, and it is perfect for individuals or small enterprises that might not have the funds for commercial software.

2. Secure and reliable

Compared to proprietary software, open-source software is frequently safer and more trustworthy. This is because open-source software is often created by a community of developers capable of swiftly identifying and correcting security flaws. Furthermore, open-source software is more likely to be subjected to independent security assessments than proprietary software. As a result, it can give more security and dependability for data analytics.

3. Flexible

Open-source software is more adaptable and flexible than proprietary software, so it's an excellent choice for data analytics. Modularity and a permissive license make it easy to extend and customize, and this flexibility is significant for data analytics, which is constantly changing and evolving. Open-source software can easily be adapted to new technologies and trends, whereas proprietary software often requires considerable investment to keep up with the latest changes.

4. Sustainable

Due to its greater adaptability and customizability, open-source software is more sustainable. The open-source community may collaborate to provide the necessary updates or improvements to data analytics. Also, by working together, the programme is made to be trustworthy and safe. The open-source community is also vibrant and helpful, with various online forums where users may post queries and seek assistance from other users.

5. Built on the work of others

Open-source software is based on the contributions of others. This implies that developers may improve their products by building on the work of others. This also means that customers may report any issues with the programme and get them resolved as soon as possible. Also, as the community contributes new features, it is constantly evolving.

6. Fast and Innovative

Since it instantly enables developers to exchange concepts and code, open-source development is swift and inventive. This makes it possible for more individuals to participate in the project, leading to a better result. Also, due to its increased user testing, open-source software is frequently more dependable.

7. Publicly Available

The public availability of source code is one advantage of open-source software. This entails that anybody may review and edit the code, which can result in high-calibre software. Moreover, it enables developer cooperation to enhance the code jointly.

Cons of Open-source Software in Data Analytics

Given the benefits of this type of software, it's essential to consider the possible drawbacks. For one, it can be prone to security issues, making it difficult to use. Also, finding good recommendations and reviews can be challenging, implying you might not get the best quality software.

1. Security risks

Due to the open-source nature of the code, open-source software is frequently more susceptible to security risks. This makes it simpler for hackers to identify and use vulnerabilities in the system.

2. Complex installation

Another drawback is that open-source software might be difficult to install and configure. This is often the result of a shortage of developers and documentation.

3. Limited Functionality

In general, open-source software provides less functionality than commercial versions. This might be an issue if you want special features that the open-source version does not offer.

4. Lack of support

The absence of assistance is one of the open-source software's main drawbacks. To locate a solution, you will need to rely on internet discussion boards and local assistance.

5. Lack of updates

Developers of open-source software don't always release updates regularly. This can result in flaws and security holes that are never repaired.

1.2 INTRODUCTION TO DATA SCIENCE TERMS

What is data science?

It is the collection of data and processing it statistically to get information that can be used to aid in decision making for business intelligence

It is the collection of raw facts and processing the facts statistically to get information that can be used to aid in decision making for business intelligence

Why data science

The field of data science emerged from having a lot of data online due to use of internet. The data was too much and there was need to make good use of the data to assist in decision making process.

It is important for business intelligence and making business decisions

What is big data?

When data is too much or voluminous, it is of many data types (such as text, integers, etc) and it is changing very frequently the we call this Big data

Big data: refers to large amount that is of many data types and changes frequently (this is the characteristic of online data)

What is data analytics?

Data analytics: the process of analyzing big data with an aim of extracting meaning from the data by observing trends.

Data analytics (DA) is the process of examining data sets in order to find trends and draw conclusions about the information they contain.

What is data mining?

Refers to a process that is used to turn raw data into meaningful data. Data mining is based on research, so many organizations follow the data mining process to transform data into useful information. It helps the organizations build more innovative strategies, increase sales, generate revenue, and grow a business by cost reduction

What is the difference between data analysis, data analytics and Data Mining?

Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight. Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions

.

Data mining is a step in the process of data analytics. Data Analytics is the umbrella which deals with every step in the pipeline of any data-driven model.

https://www.javatpoint.com/data-mining-vs-data-analysis

Hence

- **data analysis comes before analytics**
- **data mining is a step in data analytics**
- **data analytics utilizes data analysis to make business decisions**

Data Mining	Data Analysis
Data mining is a process of extracting useful information, patterns, and trends from raw data.	Data analysis is a method that can be used to investigate, analyze, and demonstrate data to find useful information.
The data mining output gives the data pattern.	The data analysis output is a verified hypothesis or insights based on the data.
It includes the intersection of databases, machine learning, and statistics.	It requires expertise in computer science, mathematics, statistics, AI.
It is also called KDD.	It is of various types - text analytics, predictive analysis, data mining, etc.
It is responsible for extracting useful patterns and trends in data.	It is responsible for developing models, testing, and proposing hypotheses using analytical methods.
The best example of a data mining application is in the E-commerce sector, where websites display options of those who purchased and viewed the specific product.	The best example of data analysis is the study of the census.

Data mining can be undertaken by a single specialist with excellent technological skills. With the right software, they are able to collect the data ready for further analysis. At this stage, a larger team simply isn't required. From here, a data mining specialist will usually report their findings to the client, leaving the next steps in someone else's hands.

However, when it comes to data analytics, a team of specialists may be needed. They need to assess the data, figure out patterns, and draw conclusions. They may use machine learning or prognostication analytics to help with the processing, but this still has a human element involved

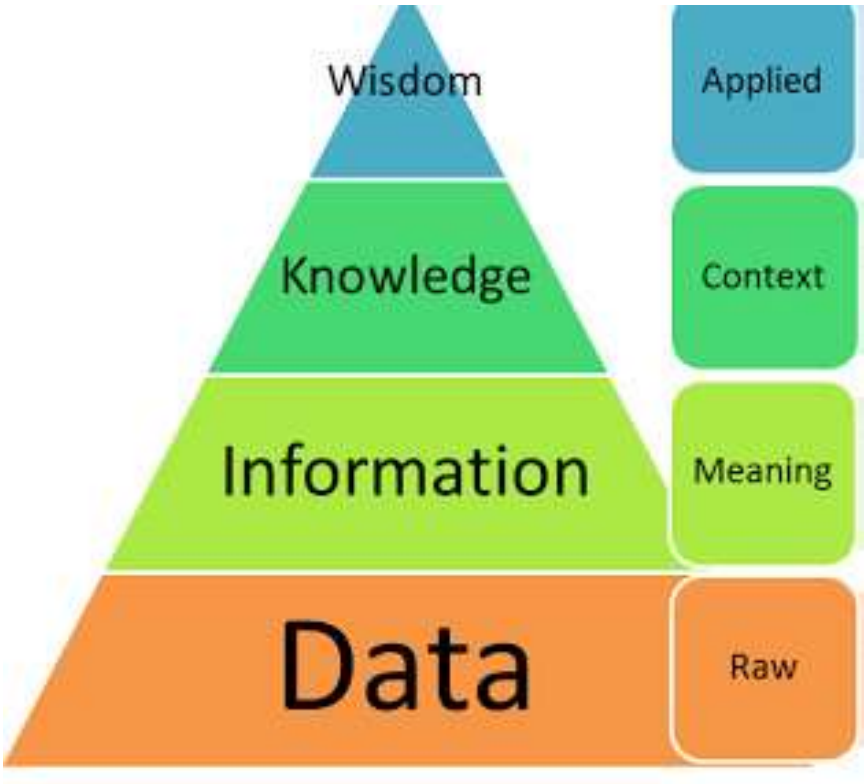
- KDD means Knowledge discovery from databases this is **data mining**

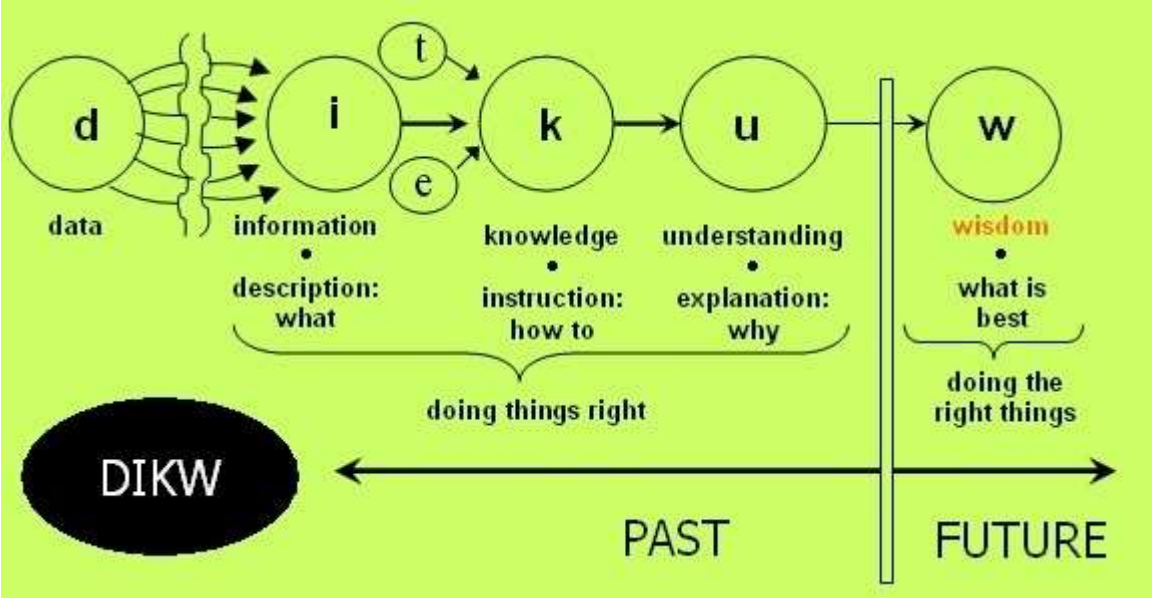
What is data Visualization?

Data visualization is the representation of information in the form of a chart, diagram, picture, etc.

Relationship between data, information, knowledge and wisdom

The DIKW pyramid





Data is conceived of as numbers, symbols or signs, representing stimuli or signals etc. i.e. raw facts. **Information** is defined as organized data that are endowed with meaning and purpose. **Knowledge** is a fluid mix of framed experience, values, contextual information, expert insight and grounded intuition that provides an environment and framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations it often becomes embedded not only in documents and repositories but also in organizational routines, processes, practices and norms. **Wisdom** is the ability to increase effectiveness. Wisdom adds value, which requires the mental function that we call judgment. The ethical and aesthetic values that this implies are inherent to the actor and are unique and personal.

In summary **Wisdom** is the ability to apply the **knowledge** acquired from **information** while information id obtained from organized and interpreted raw **data**