TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO SINH HOẠT HỌC THUẬT

Xây dựng bài giảng học phần

“Tiếng Anh ngành Khoa học Dữ liệu”

Người báo cáo: TS. Trần Thị Hòa

Bộ môn: Tin học – Trắc địa

HÀ NỘI, 06 -2024

MỤC LỤC

MỞ ĐẦU

Đối với ngành Khoa học dữ liệu, tiếng Anh đóng vai trò rất quan trọng bởi:

- Thứ nhất, tiêng Anh là ngôn ngữ chính của các tài liệu và các nguồn tài nguyên học tập;
- Thứ hai, các công cụ và phần mềm phổ biến trong Khoa học dữ liệu như Python, R, SQL, Hadoop, Spark và TensorFlow cùng các thư viên và framework như NumPy, hay Pandas đều sử dụng tiếng Anh là tài liệu hướng dẫn và trong cộng đồng hỗ trợ;
- Thứ ba, các công động trao đổi trực tuyến dành cho kỹ thuật viên và chuyên gia trong Khoa học dữ liệu hầu hết đều sử dụng tiếng Anh là ngôn ngữ trao đổi và thảo luận;
- Thứ tư, các bài báo hội thảo, hội nghi cũng như các blogs chia sẻ - hướng dẫn đều thường viết bằng tiếng Anh;
- Thứ năm, hầu hết trong các vị trí tuyển dụng của ngành này đều yêu cầu thành thạo tiếng Anh như một kỹ năng nghề nghiệp bắt buộc;
- Cuối cùng, hầu hết các công nghệ hướng tiếp cận mới trong ngành này đều được công bố và thảo luận đầu tiên bằng tiếng Anh.

Như vậy, tiếng Anh không chỉ là ngôn ngữ mà chính là công cụ thiết yếu để học tập, nghiên cứu và làm việc hiệu quả trong lĩnh vực Khoa học dữ liệu. Việc sinh viên được tiếp cận các thuật ngữ chuyên ngành ngay từ đầu là cần thiết để sinh viên có thể chủ động tìm tòi, tiếp cận với kiến thức, công nghệ và cơ hội nghề nghiệp một cách dễ dàng và rộng rãi hơn.

Bài giảng học phần “Tiếng Anh ngành Khoa học Dữ liệu” (English for Data Science) được thiết kế để đảm bảo sinh viên nắm được những thuật ngữ thông dụng trong ngành Khoa học Dữ liệu và đồng thời rèn luyện được kỹ năng đọc hiểu, kỹ năng thuyết trình và viết báo cáo bằng tiếng Anh học thuật.

.

PHẦN 1 – TÓM LƯỢC VỀ HỌC PHẦN

"English for Data Science" Syllabus

I.1. Course Introduction

English is a language that has been designated as an international language. English has been studied by more than millions of people around the world. The rapid development of data resulted in the demand for data analysts to improve their capabilities. In this era of Big Data, many companies need people who are experts in the field of data and the ability to process data and speak good language. This is where English language skills are needed. For this reason, in processing data, it is necessary to have good language skills, one of which is English.

I.2. Course Goals

- Build up vocabulary and terminology of data science;
- Differentiate those concepts and functions in data science;
- Reading skills: skimming, scanning and summarizing;
- Speaking skills: group presenting;
- Writing skills: short reports;
- Visualizing and computing skills via examples;
- Critical thinking and prob-solution thinking.

I. 3. General Course Evaluation

The evaluation progress is based on two criteria:

- Individual performance: in-class exercises, B1, B2 test, and the final test.
- Group performance: preparing and presenting reflected group work;

Otherwise, students' attendance will be evaluated as C1.

I.4. In-class Activities

"Active learning and Discussion"

Students work in groups preparing topics which will be discussed before the meeting time. Topics will be followed for each week as assigned below.

At meeting time, presenting groups will be selected randomly via "lucky number program", thus group number will be assigned at week 1. Instructor will give opening before the "showing time" begins. Finally, the instructor will summarize all discussed topics to highlight important ideas.

I.5. Main Topics

No	Topic	Hours	Achievement
1.	Data, Data Engineering and Data Science	4	- Definition of Data; - Data Cyle; -Concept of Data Engineering and Data Science; - Components of Data Science.
2.	Data Engineers and Data Scientists		- Definition; - Responsibilities and Necessary Skills for each position; - Opportunities and Average Salary.
3.	Data Collection and Dataset *19 Free Public Data Sets for Your First Data Science Project*).	1	- Type of Data; - Categorizing Data.
4.	Data Processing with exercise	4	- Definition; - 6 Steps of Data Processing; - Summary table exercise
5.	Data Analysis, Data Analyst and Business Analyst		- Definition; - Steps of Data Analysis; - Differentiate Data Analyst and Business Analyst.
6.	Data Wrangling with exercise *A Comprehensive Introduction to Data Wrangling*	2	- Definition; - 6 Steps of Data Wrangling; - Practicing Data Wrangling with Python.
7.	Data Modeling and Visualization with group exercise	4	- Definition; -Type of Data Visualization; - Hand on with different graphs/chart and group presenting.

8.	Big Data Moore’s Law	2	-Definition and Characteristics; -Reading text and reflection.
9.	Machine Learning and Artificial Intelligence *How to Become a Machine Learning Engineer*	2	- Definition; -Reading text and reflection.
10.	Common Algorithms of AI applied in Data Science commonly used algorithms	3	- Supervised Learning; - Unsupervised Learning. - Group presenting
11.	Application Program Interface	1	- Definition and Function; - Importance.
12.	Data Science programming languages with exercises: - Python; - R and Ruby; - SQL and Excel	6	- General formation; - Ranking; - Common uses; - Exercises.
13.	Open-Source Software and Library - Hadoop; - Pandas	2	- Reading text; - Exercises.
14.	B1 and B2 test	6	- Individual test;
15.	Extra exercises	10	- In-class exercises - Programming languages; - Internet Security
16.	Final test	NA	

**Additional exercises are presented at the Appendices*

PHẦN 2 – NỘI DUNG CHI TIẾT BÀI GIẢNG

II.1. Topic 1 and topic 2 – Data, Data Engineering and Data Science

What is Data Science?

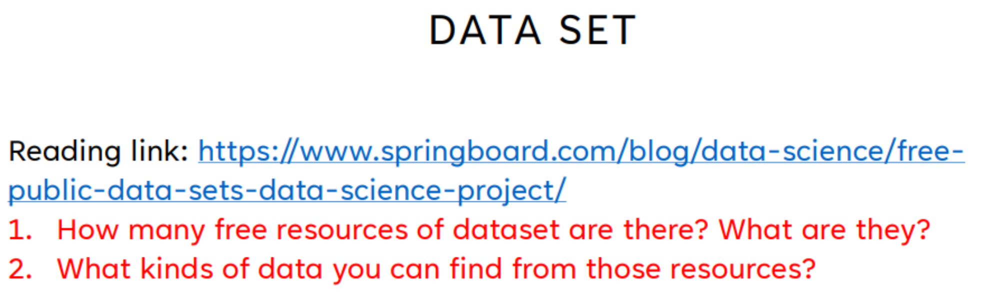
1. What is Data?
2. Data Science?
3. Data Engineering?
4. Data Scientists and Engineerers?

02/2023 English for Data Science 16

- Time consumption: 03 hours;
- Materials: 2 reading text with exercises (Reading text 1- You and your data; Reading text 2 – What is Data Science);
- Evaluation: understanding the context, discussion, question reflection and terminologies capturing.

II. 2. Topic 3 – Data Collection and Dataset:

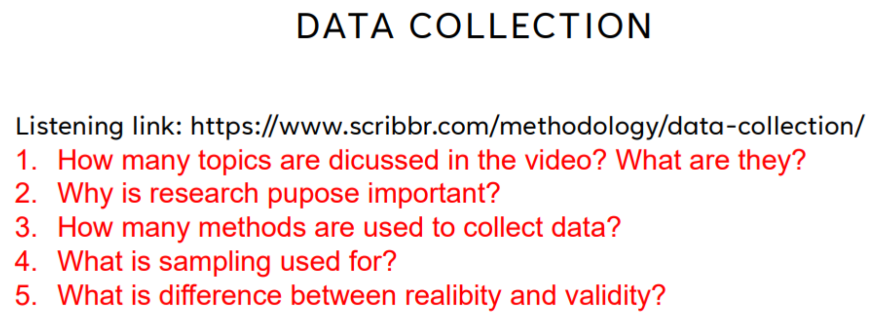
- Time consumption: 01 hours;
- Materials: Slides and audio- reading text;
- Evaluation: Question responses.



DATA SET

Reading link: https://www.springboard.com/blog/data-science/free-public-data-sets-data-science-project/

1. How many free resources of dataset are there? What are they?
2. What kinds of data you can find from those resources?



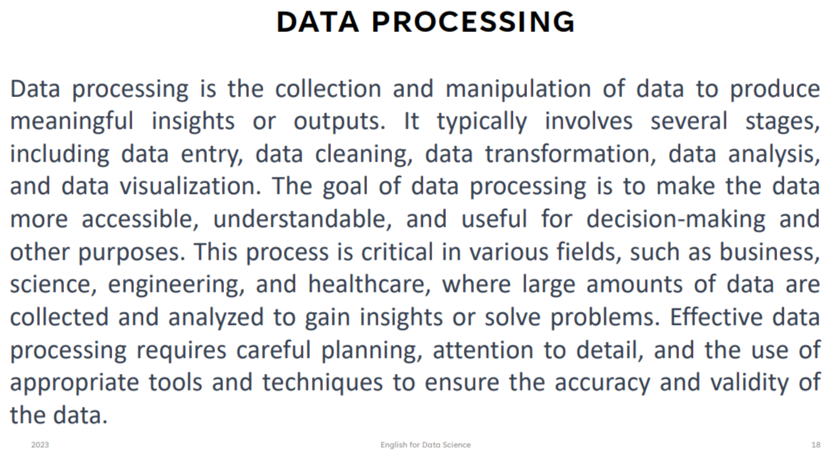
DATA COLLECTION

Listening link: https://www.scribbr.com/methodology/data-collection/

1. How many topics are dicussed in the video? What are they?
2. Why is research pupose important?
3. How many methods are used to collect data?
4. What is sampling used for?
5. What is difference between realibity and validity?

II. 3. Topic 4 and 5– Data Processing and Data Analysis

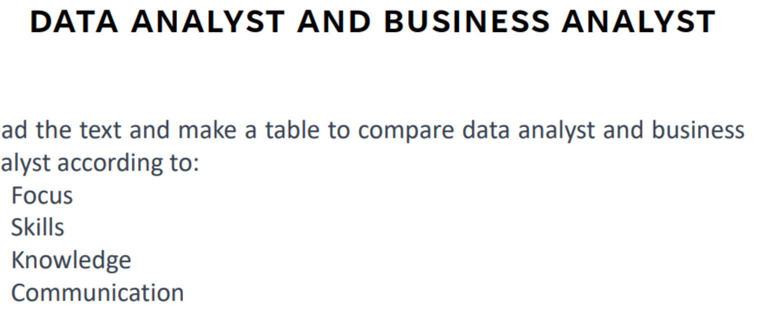
- Time consumption: 3 hours;
- Materials: Reading texts;
- Evaluation: Question Reponses, Mind Map Generation.



DATA PROCESSING

Data processing is the collection and manipulation of data to produce meaningful insights or outputs. It typically involves several stages, including data entry, data cleaning, data transformation, data analysis, and data visualization. The goal of data processing is to make the data more accessible, understandable, and useful for decision-making and other purposes. This process is critical in various fields, such as business, science, engineering, and healthcare, where large amounts of data are collected and analyzed to gain insights or solve problems. Effective data processing requires careful planning, attention to detail, and the use of appropriate tools and techniques to ensure the accuracy and validity of the data.

2023 English for Data Science 18



DATA ANALYST AND BUSINESS ANALYST

Read the text and make a table to compare data analyst and business analyst according to:

- Focus
- Skills
- Knowledge
- Communication

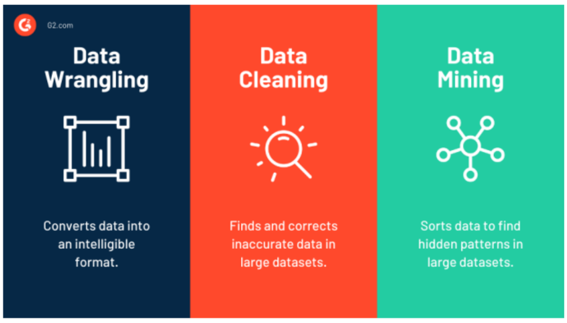
II.4. Topic 6 – Data Wrangling:

- Time consumption: 2 hours
- Materials: Reading texts, Code template;
- Evaluation: Exercise completion.

Data wrangling

https://www.g2.com/articles/data-wrangling

Data Wrangling is the process of gathering, collecting, and transforming **Raw data** into another format for better understanding, decision-making, accessing, and analysis in less time.


G2.com
Data Wrangling
Converts data into an intelligible format.
Data Cleaning
Finds and corrects inaccurate data in large datasets.
Data Mining
Sorts data to find hidden patterns in large datasets.

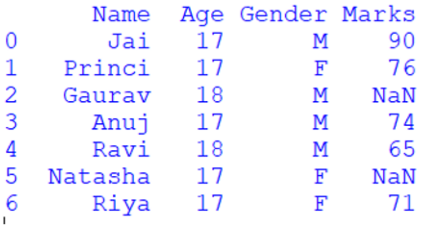
```
# Import pandas package
from IPython.display import display
import pandas as pd

# Assign data
data = {'Name': ['Jai', 'Princi', 'Gaurav',
            'Anuj', 'Ravi', 'Natasha', 'Riya'],
      'Age': [17, 17, 18, 17, 18, 17, 17],
      'Gender': ['M', 'F', 'M', 'M', 'M', 'F', 'F'],
      'Marks': [90, 76, 'NaN', 74, 65, 'NaN', 71]}

# Convert into DataFrame
df = pd.DataFrame(data)

# Display data
display(df)
```

Output is:



```
      Name  Age Gender Marks
0      Jai   17      M    90
1   Princi   17      F    76
2   Gaurav   18      M   NaN
3     Anuj   17      M    74
4     Ravi   18      M    65
5  Natasha   17      F   NaN
6     Riya   17      F    71
```

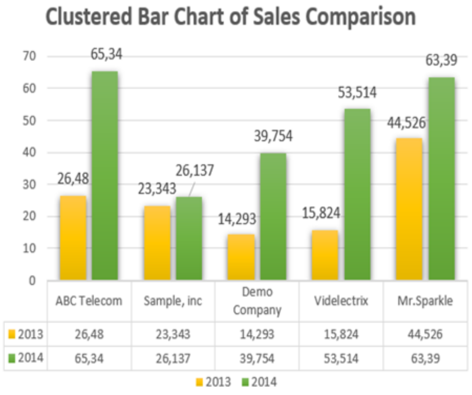
II.5. Topic 7 – Data Modeling and Data Visualization

- Time consumption: 4 hours
- Materials: Reading text, graph/chart categories;
- Evaluation: group report.

Data visualization

- Reading text: https://www.tableau.com/learn/articles/data-visualization
- Excercises:
- Advantages and disadvantages of data visualization;
- Different types of data visualization (examples)

Group members:
1.Nguyễn Hoài Thu
2.Bùi Mai Hương
3.Khổng Thị Hòa



	ABC Telecom	Sample, inc	Demo Company	Videlectrix	Mr.Sparkle
🟨 2013	26,48	23,343	14,293	15,824	44,526
🟩 2014	65,34	26,137	39,754	53,514	63,39

1.This chart is a cluster bar chart which is a type of statistical chart used to compare values of different groups through column bars arranged in clusters. This histogram is often used to illustrate comparisons between spatial groups and show the distribution of variables within each group.
The main use of cluster bar charts is to help viewers easily compare and analyze data between different groups visually. It provides an overview of differences between groups and helps users quickly draw important conclusions from the data.

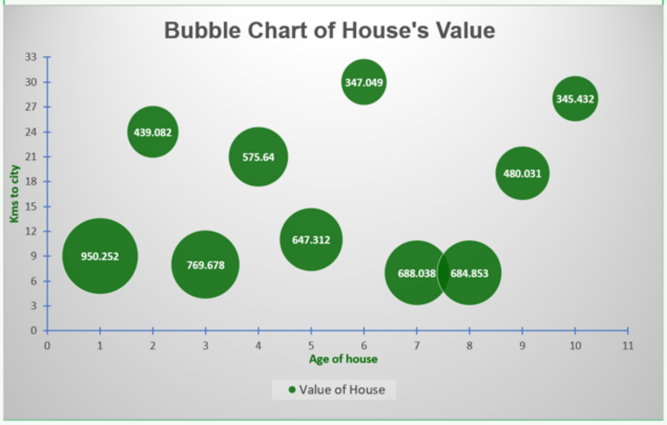
2. Chart comparing sales of a store in 2013 and 2014. Specifically, in 2013, ABC Telecom achieved sales of 26.48, Sample reached 23,343, Demo Company reached 14,293, Videlectrix reached 15,824 and Mr. Sparkle scored 44,526. Next in 2014, ABC Telecom achieved sales of 65.34, Sample reached 26,137, Demo Company reached 39,754, Videlectrix reached 53,514 and Mr.Sparkle reached 63.39

3.Comparison chart of sales of some stores from 2013 to 2014 / In 2013, MR.sparkle had the highest sales of 44,526 and demo company had the lowest sales of 14,293. In 2014, ABC Telecom had sales The highest was 65,340 and Sample.inc had the lowest sales of 26,137 .In general, the sales of all stores increased strongly. because market demand increased sharply and the stores' business strategies were effective.

4. Cluster bar charts have the following advantages:
- Show clear comparisons between groups or elements.
-Easy to read and understand, suitable for presenting market-related data or comparing figures.
-Provide an overview of data distribution.
However, cluster bar charts also have some disadvantages: Cannot display detailed data at the individual level. Cannot display multiple data dimensions at the same time effectively. It can easily lead to misunderstandings if you do not clearly understand how to read and interpret this chart.

Group members:
1. Lê Minh Đông - 2321050057
2. Lê Trần Minh Trang - 2321050044

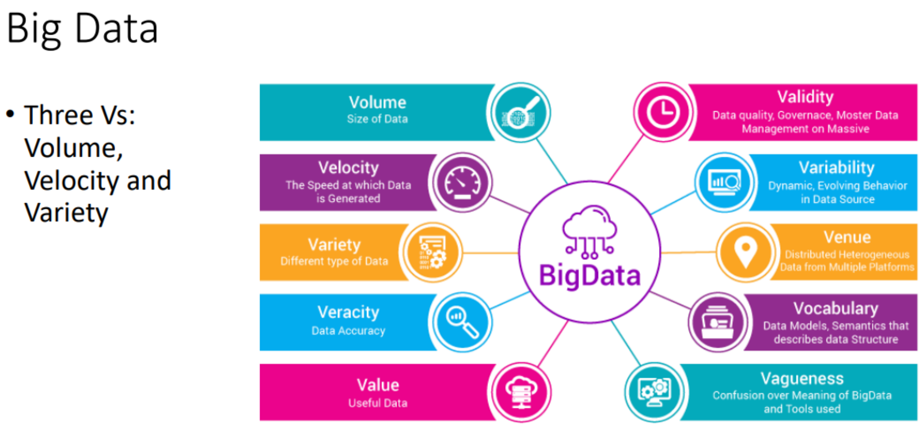
Exercise 6:



1. Introduce about the graph type (name and what is used for):
- **Graph Type**: Bubble Chart.
- **Purpose**: Bubble Chart is used to display and analyze relationships between three or more numeric variables.
2. Introduce about the data
- Each bubble on the chart represents a single data point.
  X-axis: Age of houses (measured in years).
- **Y-axis**: House's value (measured in currency, eg: dollars).
- **Bubbles**: Each bubble represents a specific house, with size reflecting its value. The size of the bubble represents a third variable, adding a dimension of comparison.
3. Some highlights of the data you want to report
- Eight houses shown, values range from **345,432** to **950,252**
- Higher value houses are located closer to the city and are younger in age.
4. Lesson learn: advantages and disadvantages of the graph type
- **Advantages**: Multi-dimensional Analysis, Visual Impact, Data Comparison.
- **Disadvantages**: Complexity for Audience, Overcrowding, Precision.

II.6. Topic 8 – Big Data

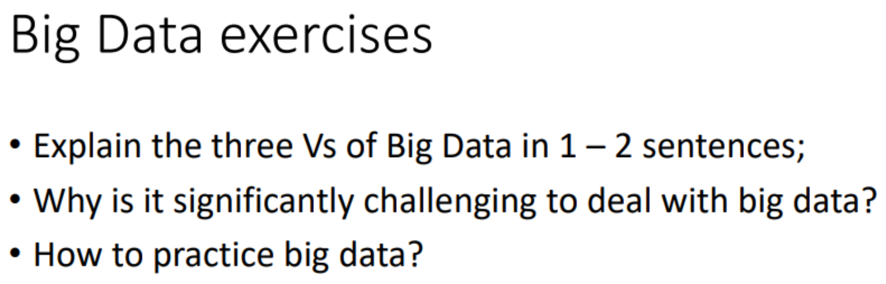
- Time consumption: 2 hours;
- Materials: Reading text;
- Evaluation: Question responses.



Big Data

- Three Vs: Volume, Velocity and Variety

Volume
Size of Data
Velocity
The Speed at which Data is Generated
Variety
Different type of Data
Veracity
Data Accuracy
Value
Useful Data
BigData
Validity
Data quality, Governace, Moster Data Management on Massive
Variability
Dynamic, Evolving Behavior in Data Source
Venue
Distributed Heterogeneous Data from Multiple Platforms
Vocabulary
Data Models, Semantics that describes data Structure
Vagueness
Confusion over Meaning of BigData and Tools used



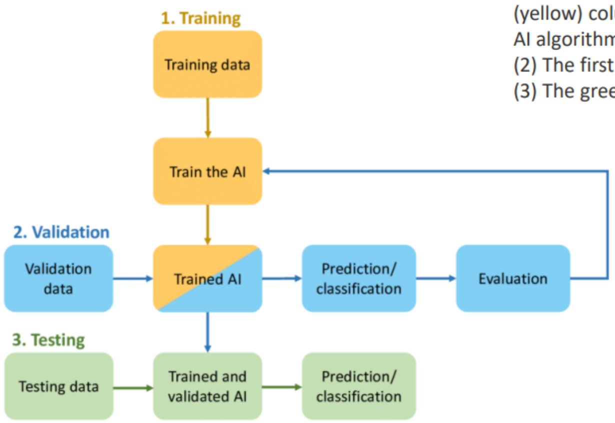
Big Data exercises

- Explain the three Vs of Big Data in 1 – 2 sentences;
- Why is it significantly challenging to deal with big data?
- How to practice big data?

II. 7. Topic 9 – Machine Learning and Artificial Intelligence

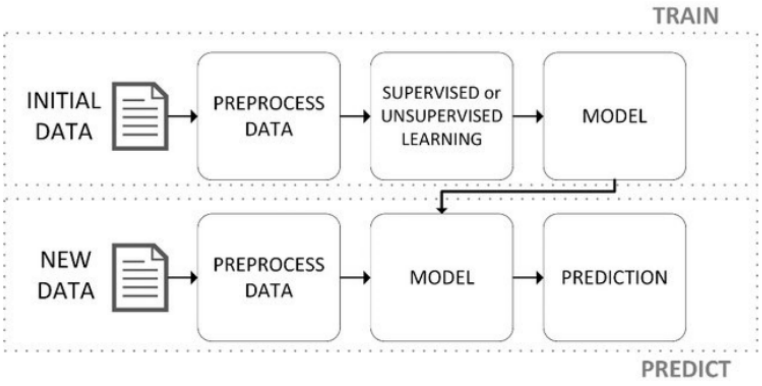
- Time consumption: 2 hours;
- Materials: Reading texts;
- Evaluation: Question responses, short writing.

A model of AI workflow



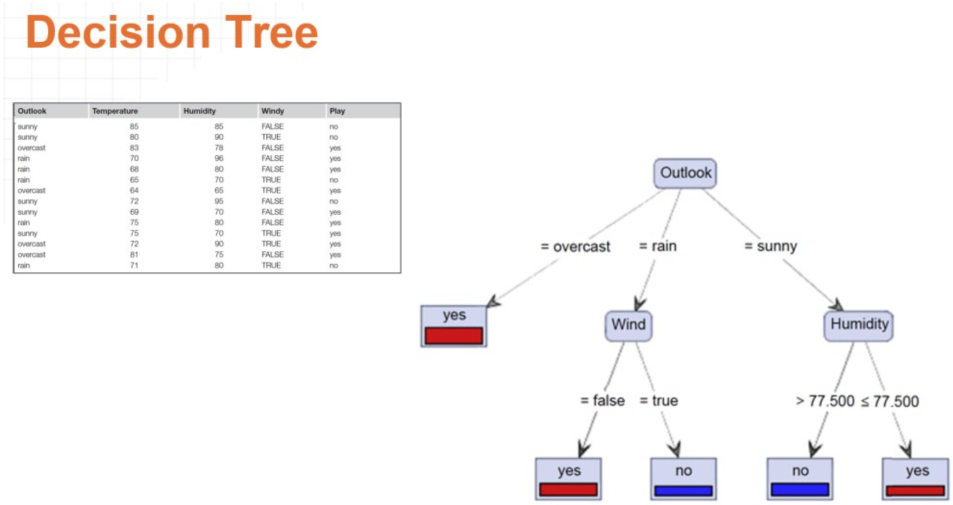
The workflow of AI based strategies. (1) The (yellow) column: the training phase to train the AI algorithm.
(2) The first row (blue): validate and refine data
(3) The green row: the application phase

Here is an example of a ML model



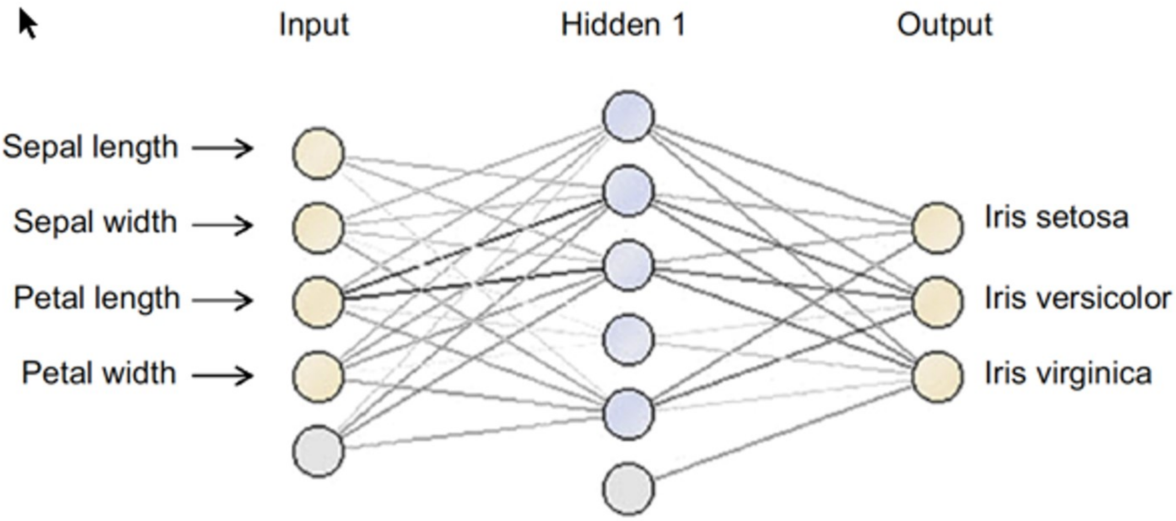
II.8. Topic 10 – AI algorithms

- Time consumption: 3 hours;
- Materials: prepared by students;
- Evaluation: upon student’s performance (ensuring cover definition, structure and application of that algorithms)
- Organized in 15 groups.



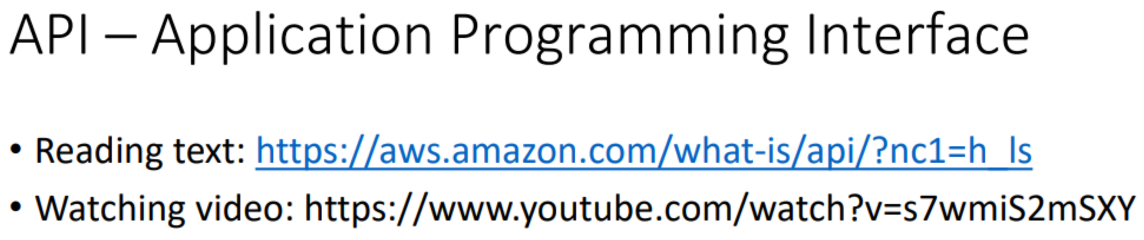
Decision Tree

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	78	FALSE	yes
rain	70	96	FALSE	yes
rain	68	80	FALSE	yes
rain	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rain	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rain	71	80	TRUE	no



II.9. Topic 11- Application Program Interface

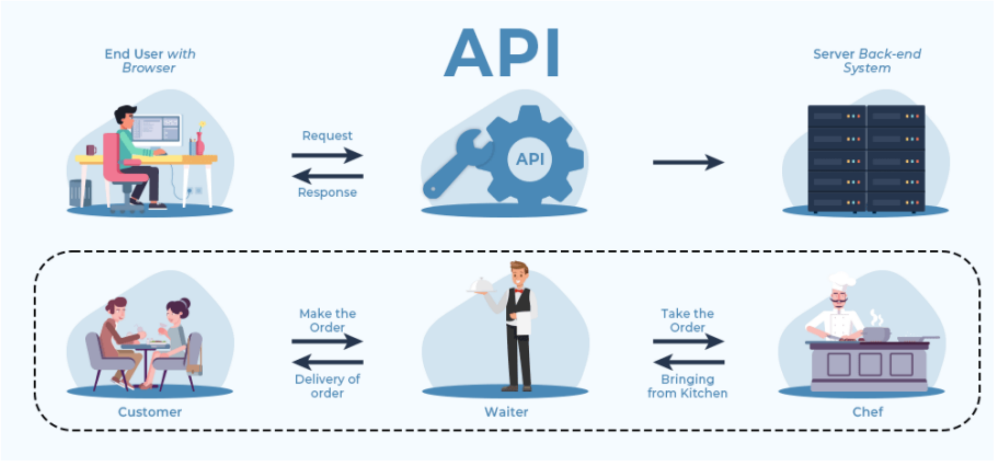
- Time consumption: 1 hours;
- Materials: Reading text, watching video.
- Evaluation: QA.



API – Application Programming Interface

- Reading text: https://aws.amazon.com/what-is/api/?nc1=h_ls
- Watching video: https://www.youtube.com/watch?v=s7wmiS2mSXY

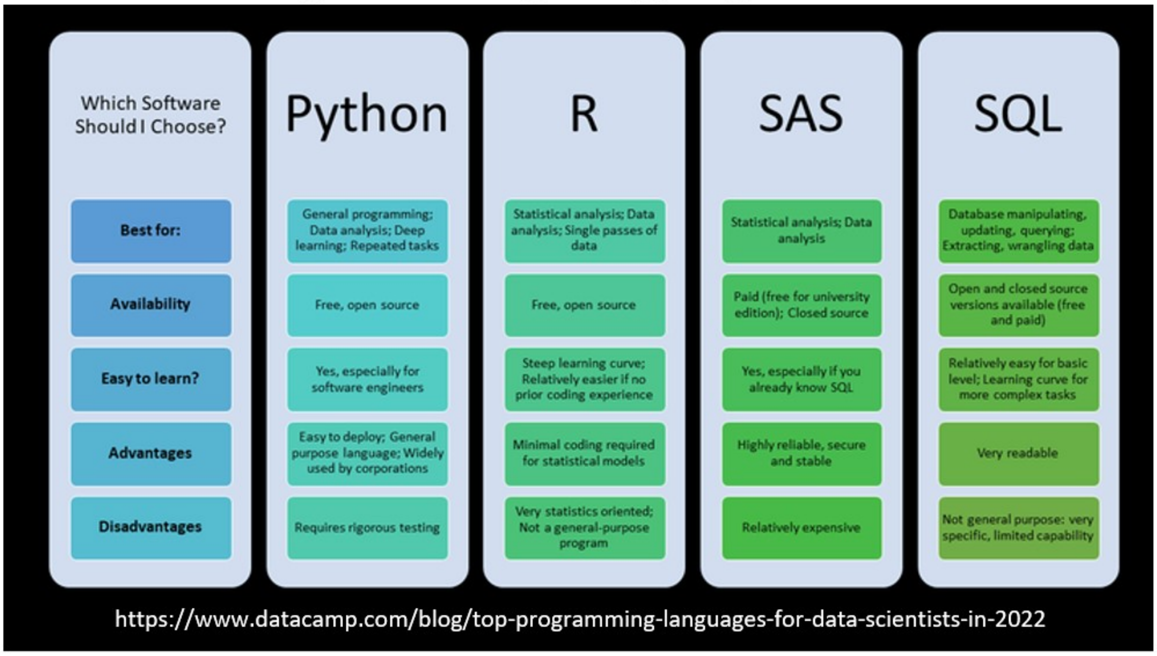
API in a short note



II.10. Topic 12- Data Science – Programming Languages

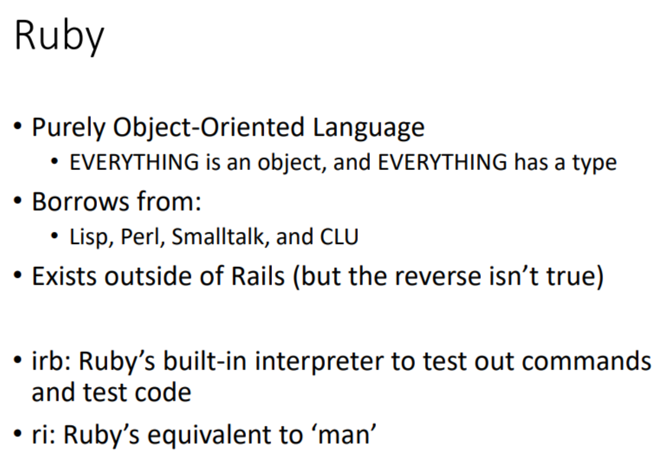
- Time consumptions: 6 hours;
- Materials: reading texts and exercise template;
- Evaluation: upon individual completion.

Programming languages for DS



Which Software Should I Choose?	Python	R	SAS	SQL
Best for:	General programming; Data analysis; Deep learning; Repeated tasks	Statistical analysis; Data analysis; Single passes of data	Statistical analysis; Data analysis	Database manipulating, updating, querying; Extracting, wrangling data
Availability	Free, open source	Free, open source	Paid (free for university edition); Closed source	Open and closed source versions available (free and paid)
Easy to learn?	Yes, especially for software engineers	Steep learning curve; Relatively easier if no prior coding experience	Yes, especially if you already know SQL	Relatively easy for basic level; Learning curve for more complex tasks
Advantages	Easy to deploy; General purpose language; Widely used by corporations	Minimal coding required for statistical models	Highly reliable, secure and stable	Very readable
Disadvantages	Requires rigorous testing	Very statistics oriented; Not a general-purpose program	Relatively expensive	Not general purpose: very specific, limited capability

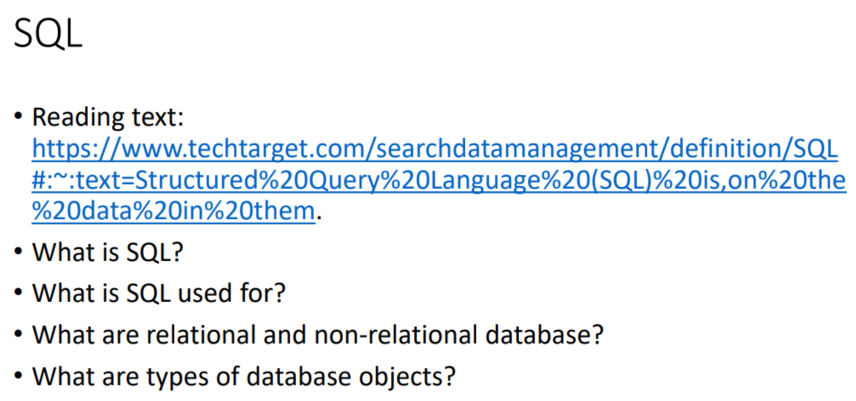
https://www.datacamp.com/blog/top-programming-languages-for-data-scientists-in-2022



Ruby

- Purely Object-Oriented Language
  - EVERYTHING is an object, and EVERYTHING has a type
- Borrows from:
  - Lisp, Perl, Smalltalk, and CLU
- Exists outside of Rails (but the reverse isn’t true)

- irb: Ruby’s built-in interpreter to test out commands and test code
- ri: Ruby’s equivalent to ‘man’

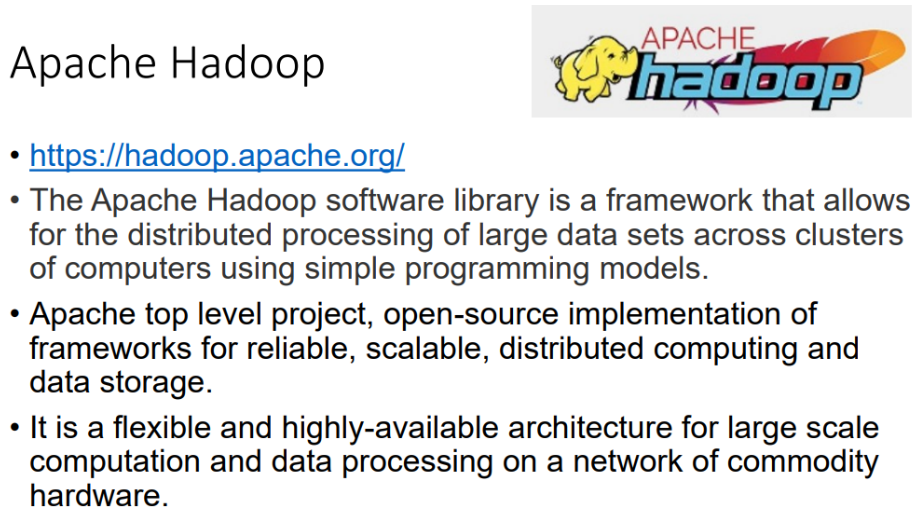


SQL

- Reading text: https://www.techtarget.com/searchdatamanagement/definition/SQL#:~:text=Structured%20Query%20Language%20(SQL)%20is,on%20the%20data%20in%20them.
- What is SQL?
- What is SQL used for?
- What are relational and non-relational database?
- What are types of database objects?

II.11. Topic 13 – Open sources

- Time consumption: 2 hours
- Materials: reading texts;
- Evaluation: upon individual performance

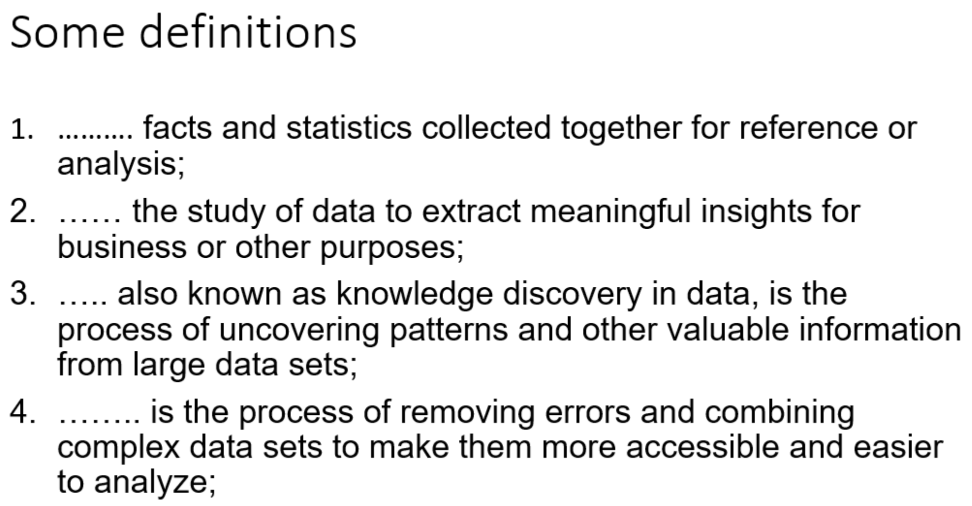


Apache Hadoop

- https://hadoop.apache.org/
- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- Apache top level project, open-source implementation of frameworks for reliable, scalable, distributed computing and data storage.
- It is a flexible and highly-available architecture for large scale computation and data processing on a network of commodity hardware.

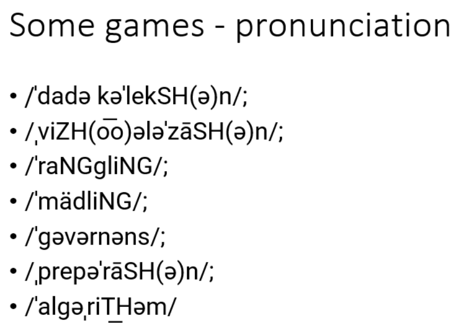
II. 12. Topic 14 to 16 – Tests and Additional exercises

- Time consumption: 16 hours;
- B1 and B2 test: individual tests;
- Additional exercises: practicing reading and writing skills.



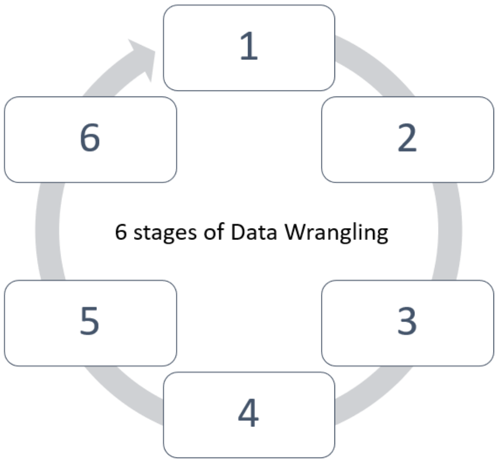
Some definitions

1. .... facts and statistics collected together for reference or analysis;
2. …… the study of data to extract meaningful insights for business or other purposes;
3. ….. also known as knowledge discovery in data, is the process of uncovering patterns and other valuable information from large data sets;
4. …….. is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze;



Some games - pronunciation

- /ˈdadə kəˈlekSH(ə)n/;
- /ˌviZH(o͞o)ələˈzāSH(ə)n/;
- /ˈraNGgliNG/;
- /ˈmädliNG/;
- /ˈɡəvərnəns/;
- /ˌprepəˈrāSH(ə)n/;
- /ˈalɡəˌriT͟Həm/



PHẦN 3 – PHỤ LỤC

III.1. You and Your data

Before reading

Do the preparation task first. Then read the article and do the exercises.

Preparation task: Match the definitions (a–h) with the vocabulary (1–8).

Vocabulary	Definition
1. …… data	a. directed at a particular person or group
2. …… to be aware of	b. permission to do something
3. …… consent	c. to risk having a harmful effect on something
4. …… to keep track / to track	d. to control an activity or process, especially with rules
5. …… a scandal	e. information, especially facts or numbers, that is collected for a future purpose
6. …… targeted	f. to study or record someone's behavior over time
7. …… to regulate	g. to have noticed or know about something
8. …… to compromise	h. a public feeling of shock and disapproval

Answers:

Reading text

YOU AND YOUR DATA

As the internet and digital technology become a bigger part of our lives, more of our data becomes publicly accessible, leading to questions about privacy. So, how do we interact with the growing digital world without compromising the security of our information and our right to privacy? Imagine that you want to learn a new language. You search 'Is German a difficult language?' on your phone. You click on a link and read an article with advice for learning German. There's a search function to find German courses, so you enter your city name. It asks you to activate location services to find courses near you. You click 'accept'. You then message a German friend to ask for her advice. When you look her up on social media, an advertisement for a book and an app called German for Beginners instantly pops up. Later the same day, while you're sending an email, you see an advert offering you a discount at a local language school. How did they know? The simple answer is online data. At all stages of your search, your devices, websites and applications were collecting data on your preferences and tracking your behavior online. 'They' have been following you.

Who uses our data and why?

In the past, it was easy for people to keep track of their personal information. Like their possessions, people's information existed mostly in physical form: on paper, kept in a folder, locked in a cupboard or an office. Today, our personal information can be collected and stored online, and it's accessible to more people than ever before. Many of us share our physical location, our travel plans, our political opinions, our shopping interests and our family photos online – as key services like ordering a takeaway meal, booking a plane, taking part in a poll or buying new clothes now take place online and require us to give out our data. Every search you make, service you use, message you send and item you buy is part of your 'digital footprint'. Companies and online platforms use this 'footprint' to track exactly what we are doing; from what links we click on to how much time we spend on a website. Based on your online activity, they can guess what you are interested in and what things you might want to buy. Knowing so much about you gives online platforms and companies a lot of power and a lot of money. By selling your data or providing targeted content, companies can turn your online activity into profit. This is the foundation of the growing industry of digital marketing.

Can you protect your data?

Yes … and no!

Some of the time our personal data is shared online with our consent. We post our birthday, our photographs and even our opinions online on social media. We know that this information is publicly accessible. However, our data often travels further than we realize, and can be used in ways that we did not intend. Certain news scandals about data breaches, where personal data has been lost, leaked or shared without consent, have recently made people much more aware of the potential dangers of sharing information online. So, can we do anything to protect our data? Or should we just accept that in fact nothing is 'free' and sharing our data is the price we have to pay for using many online services? As people are increasingly aware of and worried about data protection, governments and organizations are taking a more active role in protecting privacy. For example, the European Union passed the General Data Protection Law, which regulates how personal information is collected online. However, there is still much work to be done. As internet users, we should all have a say in how our data is used. It is important that we pay more attention to how data is acquired, where it is stored and how it is used. As the ways in which we use the internet continue to grow and change, we will need to stay informed and keep demanding new laws and regulations, and better information about how to protect ourselves. Safer Internet Day is an ideal time to find out more about this topic.

TASK 1

Are the sentences true or false?

1. Information about you is collected when you look at websites.	True False
2. Using different devices (for example, your phone and your laptop) makes it impossible for companies to track you.	True False
3. The train of information you leave online is called your ‘digital footprint’.	True False
4. Companies use your digital footprint to make money.	True False
5. This issue has not been in the news, so most people are completely unaware of it.	True False
6. European law on the protection of online data has changed.	True False
7. The writer thinks the new law has solved the problem.	True False
8. The article concludes by saying individuals should stay up to date and know how their information is used.	True False

TASK 2

Complete the sentences with the words.

aware compromise consent data
regulates scandal targeted track

1. Our devices, websites and applications collect ……………………………… about our online behavior.

2. Until recently, many people were not ……………………………… of how much of their personal information was collected and shared.

3. Information about products you are interested in is used to create ……………………………… advertising.

4. The news of how certain applications used people’s private information caused a ………………………………

5. People felt their information had been used for purposes that they had not agreed to, without their ………………………………

6. The General Data Protection Law ……………………………… how personal data is collected online.

7. When private information was stored physically, on paper, it was easier to keep ……………………………… of where your data went.

8. If you want to use many online apps and services, you still have to ……………………………… your right to privacy.

DISCUSSION

What do you do to protect your data?

III.2. What is Data Science?

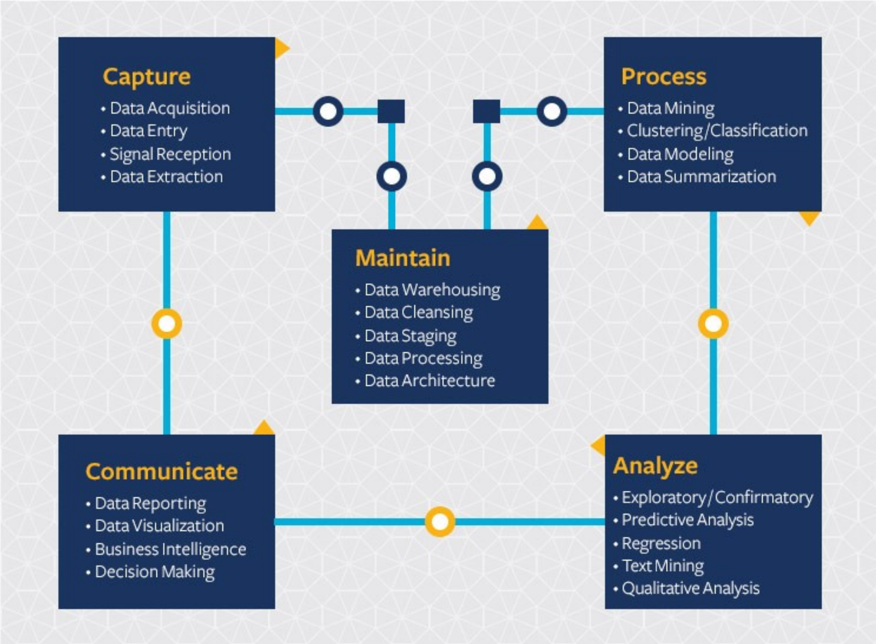
WHAT IS DATA SCIENCE?

In this section, you will learn about:

1. *Definition of Data Science, Data Scientists and Data Analysists (see Slides and readings);*
2. *The processes of Data Science;*
3. *Required skills to become Data engineer, Data Analyst or Data Scientist.*

Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. To uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.

The Data Science Life Cycle



The image represents the five stages of the data science life cycle: Capture, (data acquisition, data entry, signal reception, data extraction); Maintain (data warehousing, data cleansing, data staging, data processing, data architecture); Process (data mining, clustering/classification, data modeling, data summarization); Analyze (exploratory/confirmatory, predictive analysis, regression, text mining, qualitative analysis); Communicate (data reporting, data visualization, business intelligence, decision making).

The term “data scientist” was coined when companies first realized the need for data professionals skilled in organizing and analyzing massive amounts of data. Ten years after the widespread business adoption of the internet, Hal Varian, Google’s chief economist, first dean of the UC Berkeley School of Information (I School), and UC Berkeley emeritus professor of information sciences, business, and economics, predicted the importance of adapting to technology’s influence and reconfiguration of different industries.

“The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”
– Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics[1]

Today, effective data scientists masterfully identify relevant questions, collect data from a multitude of different data sources, organize the information, translate results into solutions, and communicate their findings in a way that positively affects business decisions. These skills are now required in almost all industries, which means data scientists have become increasingly valuable to companies.

What Does a Data Scientist Do?

Data scientists have become assets across the globe and are present in almost all organizations. These professionals are well-rounded, analytical individuals with high-level technical skills who can build complex quantitative algorithms to organize and synthesize large amounts of information used to answer questions and drive strategy in their organizations. They also have the communication and leadership experience to deliver tangible results to various stakeholders across an organization or business.

Data scientists are typically curious and result-oriented, with exceptional industry-specific knowledge and communication skills that allow them to explain highly technical results to their non-technical counterparts. They possess a strong quantitative background in statistics and linear algebra as well as programming knowledge with focuses in data warehousing, mining, and modeling to build and analyze algorithms.

They also use key technical tools and skills, including:

R
Python
Apache Hadoop
MapReduce
Apache Spark
NoSQL databases
Cloud computing
D3
Apache Pig
Tableau
iPython notebooks
GitHub

Why Become a Data Scientist?

As increasing amounts of data become more accessible, large tech companies are no longer the only ones in need of data scientists. There’s now a demand for qualified data science professionals across organizations, big and small.

With the power to shape decisions, solve real-world challenges, and make a meaningful impact in diverse sectors, data science professionals have the opportunity to pursue various career paths.

Where Do You Fit in Data Science?

Data is everywhere and expansive. Various terms related to mining, cleaning, analyzing, and interpreting data are often used interchangeably, but the roles typically involve different skill sets. The complexity of the data analyzed also differs.

Data Scientist

Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skills as well as the ability to mine, clean, and present data. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data. Data scientists also leverage machine learning techniques to model information and interpret results effectively, a skill that differentiates them from data analysts. Results are then synthesized and communicated to key stakeholders to drive strategic decision making in the organization.

Skills needed: Programming skills (SAS, R, Python), statistical and mathematical skills, storytelling and data visualization, Hadoop, SQL, machine learning

Data Analyst

Data analysts bridge the gap between data scientists and business analysts. They're provided with the questions that need answering from an organization and then organize and analyze data to find results that align with high-level business strategy. Data analysts are responsible for translating technical analysis to qualitative action items and effectively communicating their findings to diverse stakeholders.

Skills needed: Programming skills (SAS, R, Python), statistical and mathematical skills, data wrangling, data visualization

Data Engineer

Data engineers manage exponentially growing and rapidly changing data. They focus on developing, deploying, managing, and optimizing data pipelines and infrastructure to transform and transfer data to data scientists and data analysts for querying.

Skills needed: Programming languages (Java, Scala), NoSQL databases (MongoDB, Cassandra DB), frameworks (Apache Hadoop)

Extra Reading

Is There a Difference Between a Data Engineer and a Data Scientist?

There was a time when data scientists were expected to perform the role of data engineers. But as the field of data has grown and evolved, with data gathering and management becoming more complex and unwieldy, and organizations expecting more answers and insights from the data gathered, the job has been split into two.

Today, the main difference between these two data professionals is that data engineers build and maintain the systems and structures that store, extract, and organize data, while data scientists analyze that data to predict trends, glean business insights, and answer questions that are relevant to the organization.

Data Engineer vs. Data Scientist

Although there is overlap in the skills between data engineers and data scientists, and in the past data scientists were expected to perform some of the duties of data engineers, the two roles are distinctly separate and different.

Role and Responsibilities

It helps to think of data engineers and data scientists as having complementary roles. Data engineers build and optimize the systems that allow data scientists to do their job. Data scientists, meanwhile, find meaning in the troves of data that data engineers manage.

What Does a Data Engineer Do?

A data engineer is a data professional who prepares the data infrastructure for analysis. They are focused on the production readiness of raw data and elements such as formats, resilience, scaling, data storage, and security. Data engineers are tasked with designing, building, testing, integrating, managing, and optimizing data from a variety of sources. They also build the infrastructure and architectures that enable data generation.

Their primary focus is to build free-flowing data pipelines by combining a variety of big data technologies that enable real-time analytics. Data engineers also write complex queries to

What Does a Data Scientist Do?

Data scientists concentrate on finding new insights from the data that was prepared for them by data engineers. As part of their job, they conduct online experiments, develop hypotheses, and use their knowledge of statistics, data analytics, data visualization, and machine learning algorithms to identify trends and create forecasts for the business.

They also engage with business leaders to understand their specific needs and present complex findings, both verbally and visually, in a manner that can be followed by a general business audience.

Education and Requirements

Many data engineers and data scientists hold a bachelor’s degree in computer science or a related field such as mathematics, statistics, economics, or information technology. And while employers often look for candidates with advanced degrees, it is possible to land a role in data science or data engineering without a degree.

What Are the Requirements To Become a Data Engineer?

Data engineers usually hail from a software engineering background and are proficient in programming languages like Java, Python, SQL, and Scala. Alternatively, they might have a degree in mathematics or statistics that helps them apply different analytical approaches to solve business problems.

To get hired as a data engineer, most companies look for candidates with a bachelor’s degree in computer science, applied math, or information technology. Candidates may also be required to have a few data engineering certifications, like Google's Professional Data Engineer or IBM Certified Data Engineer. It also helps if they are experienced in building big data warehouses that can run some Extract, Transform, and Load, or ETL, on top of big data sets.

What Are the Requirements To Become a Data Scientist?

Data scientists are usually presented with large volumes of data without any particular business problems to solve. In this scenario, the data scientist will be expected to explore the data, formulate the right questions, and present their findings. This makes it essential for data scientists to have a broad knowledge of different techniques in big data infrastructures, data mining, machine learning algorithms, and statistics. As they also have to work with data sets that come in various forms to run their algorithms effectively and efficiently, they also need to be up-to-date with all the latest technologies.

Data scientists are expected to be proficient in programming languages such as SQL, Python, R, and Java, and be familiar with tools such as Hive, Hadoop, Cassandra, and MongoDB.

Data Scientist vs. Data Engineer Salary

For the analytical mind, both positions offer a highly rewarding and lucrative career.

What Does a Data Engineer Earn?

Data engineers' salaries depend on variables such as the type of role, relevant experience, and where the job is located. According to Glassdoor, the average salary for a data engineer is about $142,000 per year.

What Does a Data Scientist Earn?

Again, what data scientists earn also depends on the type of job, their skills, qualifications, and where it's located. According to Glassdoor, on average, a data scientist makes about $139,000 per year.

Career Paths

There is no one set path to becoming a data engineer or a data scientist, but below are some of the common ways in which people have navigated the field to get to their dream jobs.

What's a Typical Career Path for a Data Engineer?

Data engineering is not usually an entry-level role. Because of this, many data engineers get their start in software engineering or business intelligence/systems analytics — roles that give them exposure to the systems and infrastructure that are crucial to the field of data science.

Many data engineers take advantage of roles such as data architect, solutions architect, and database developer to perfect their data engineering skills, develop a deeper knowledge of data processing and cloud computing, and gain experience with ETL and data layers. Some may also work in data analytics to bolster their knowledge of what data analysts and data scientists need before transitioning into data engineering.

What's a Typical Career Path for a Data Scientist?

Many data scientists get their start in an entry-level data science role, whether through an internship or getting hired as a junior data scientist. These entry-level positions give new data scientists the opportunity to continue developing their technical skills and to work on projects assigned to them before they advance to designing their own experiments and solving more ambitious business problems.

Data analysts commonly pivot into data science roles either by teaching themselves the relevant data science skills or by enrolling in an online course or bootcamp.

Related Read: Data Analyst vs. Data Scientist: Salary, Skills, & Background

Can a Data Engineer Become a Data Scientist (or Vice Versa)?

The short answer is yes, data engineers can become data scientists and vice versa, with some additional training. The overlap in skills — from knowledge of programming languages to working with data pipelines — means that members of both professions are equipped with the foundational knowledge and vocabulary to have a relatively easy career transition. But, given that data engineers have a greater focus on the architecture and infrastructure that supports the work of data scientists, and data scientists are more concerned with developing and testing hypotheses through data, both professions would have to brush up on additional skills before that can make the leap.

Data Scientist vs. Data Engineer: Which Is Best for You?

Despite the overlap in skills between the two professions, data scientists and data engineers have different responsibilities, and the roles may be better suited to certain personality types.

Consider Being a Data Engineer if…

Data engineers deal mostly with the infrastructure and architecture that stores and organizes data. They are strong coders who like learning and using new technologies, enjoy discovering new ways to make software and systems more efficient, and thrive on helping an organization save time and resources. If you’re a tinkerer who’s always looking for ways to improve the things you build, find purpose in creating the supportive tools that help others do their job, and love playing with the latest tools and technologies, then data engineering might be the right career for you.

Consider Being a Data Scientist if…

Data scientists are analytical thinkers who are curious, aren’t afraid of asking questions, and live for putting their hypotheses to the test. Data scientists not only use data to make sense of things that have happened in the past, they also forecast trends and try to understand what might happen in the future. If you enjoy running advanced statistical analyses, writing machine learning algorithms, and using creativity to solve problems, then a career as a data scientist might be right for you.

EXERCISES

1. **Vocabulary Exercise: Define the following terms based on the text:**

a. Data Science Life Cycle

b. Data Scientist

c. Data Analyst

d. Data Engineer

e. Data Warehousing

f. Data Mining

g. Predictive Analysis

h. Machine Learning

i. Cloud Computing

j. NoSQL Databases

2. **Comprehension Questions: Answer the following questions based on the text:**

a. What are the five stages of the Data Science Life Cycle?

b. Who coined the term "data scientist" and when?

c. What are some key skills required for a Data Scientist?

d. What is the difference between a Data Scientist and a Data Engineer?

e. What are some tools and programming languages commonly used by Data Scientists?

3. **Critical Thinking:**

a. Do you think data scientists are essential in today's business landscape? Why or why not?

b. How do you think advancements in technology will impact the role of data scientists in the future?

c. Compare and contrast the roles of Data Scientists, Data Analysts, and Data Engineers. Why are these distinctions important?

d. In what industries do you think data science skills are most valuable? Provide examples and explanations.

4. **Role Comparison: Compare the responsibilities and skill sets of a Data Scientist and a Data Engineer. Identify at least three similarities and three differences between these roles.**

III.3. Data Processing and Data Visualization

Data processing

Data processing describes the collection and **transformation of raw data** into meaningful information.

Once processed, this information can be used for a variety of different purposes by everyone from **data scientists** to **business analysts**, C-suite decision-makers, and IT managers, to name just a few. Regardless of the end-user or their task, the ultimate goal of data processing always remains the same. It is about turning data into information.

Within the context of modern data analytics, much of the data processing lifecycle is automated using sophisticated hardware and algorithms. Often, this is the precursor to more in-depth and hands-on data analysis, where the information gleaned is further analyzed to extract more focused and actionable insights.

Data processing is a cycle, not a once-off

Since data is constantly evolving, updating, and changing, it’s important to understand that data processing is not a standalone task. Rather, it is an iterative cycle. The cycle is constantly repeated—every time data is updated, or whenever you want to carry out a new analysis. For this reason, data processing—even using machines to streamline things—takes an awful lot of time.

Important note

It's worth noting here that the term "data processing" is sometimes also used to describe individual steps in the overall process, as well as dedicated departments within large organizations whose function is to carry out data processing.

We're just mentioning this in case you come across these terms in your travels. But for the sake of this article, we'll stick to the first definition for now: data processing as a methodology.

Why is data processing important?

As we've already mentioned, data processing is important for transforming meaningless raw data into meaningful information for further analysis. But it has numerous other benefits, too. These include:

- **More effective storage:** Storing processed data in relational databases (as opposed to unstructured, text-heavy documents) makes them much easier to store, manipulate and explore using database tools like SQL.
- **Easier to produce reports:** Once a dataset is effectively processed, you can quickly create reports, dashboards, and other summaries of its characteristics.
- **Improved productivity:** By being easier to navigate, processed data saves users from having to heavily reprocess a dataset every time they want to use it.
- **Sensible housekeeping:** Data processing isn't a one-off task, but an ongoing cycle. Reprocessing helps maintain order and minimizes the number of errors or mistakes that creep into your data.
- **It's more accurate:** Regularly removing outliers, errors and unnecessary data points (and using clearly defined data models) increases the accuracy of your insights.

These are just a few of the reasons why data processing is important. While none of these should come as a big surprise, this hopefully illustrates just how many areas of business effective data processing can impact (beyond merely being used for data analytics tasks).

What is the data processing lifecycle? (step-by-step)

As we've already seen, data processing is an ongoing cycle, not a standalone task. In this section, we explore the different steps that make up this lifecycle. These include:

- Data collection
- Data preparation
- Data input
- Data processing
- Data output
- Data storage

Now let's look at each of these in a bit more detail.

Data collection

The first task in data processing is to collect raw data. Straightforward as this sounds, it requires careful planning.

A common saying in data analytics (particularly in **machine learning**) is “**garbage in, garbage out,**” which means the quality of your data directly impacts the quality of your insights. You must carefully map out which data you require, where you’ll collect them from, and ensure that the source (or sources) are reliable.

It doesn’t matter how much you process erroneous data, it won’t make it any more accurate!

Common sources of raw data include:

- stock market and financial data
- social media
- websites
- apps
- emails

Data preparation

Also referred to as **data cleaning** or **data wrangling**, data preparation involves tidying a raw dataset and introducing structure.

This might include removing unwanted observations, duplicates, and outliers, fixing structural and contradictory data errors, type conversion, and so on.

In reality, the exact tasks will differ depending on the nature of the data and how you intend to use them. For example, maybe you’ve collected housing data to compare house prices. If so, you might choose to remove any buyer information that isn’t directly relevant to the transaction.

Whatever your task, the ultimate goal of data preparation is to get a dataset into its best possible state before actively processing it.

To explore this step of the process in more detail, you can learn more about data cleaning in our guide.

Data input

Next up: data input. This is the process of converting raw (but tidied) data into a machine-readable format.

Once this is done, the data is then fed into a central processing unit (CPU). This could be a powerful computer with a custom-built, open-source big data architecture, or a piece of existing enterprise software.

While this step might seem straightforward, data input is as important as data collection. You should always use validated data at this stage to avoid inputting ‘garbage.’ Data input is commonly done electronically.

While it can also be carried out using scanners or manually (for smaller datasets) this is increasingly considered poor practice because **it allows human error to creep in**. It’s also increasingly impractical with today’s vast datasets.

Data processing

Once your data have been input into the appropriate system, they can be processed using a variety of different techniques (which we summarize in section three).

At this stage of the process, machine learning or **artificial intelligence algorithms** will make sense of the data, preparing to output useful information.

As a standalone task, the data processing step of the cycle can be quite time-consuming. Processing speed depends on things like your computing power, the complexity and size of the dataset, and other factors relating to the infrastructure you're using.

Data output

Once a dataset has been processed, the results can finally be delivered to the end-user.

The format will depend on the type of data you started with and/or your preferred medium. It could include written reports, videos, images, documents, graphs, tables, and more.

At this stage, the data will have been transformed and can no longer be considered raw data. Depending on the use case, the data may be used to create dashboards, to carry out exploratory data analysis, or it could be processed further to refine relevant details.

Data storage

The final step is data storage. This is where the output is either loaded back into the system it came from or imported into another system for future use.

Storage might be in the form of a CRM or a relational database that can be queried using tools like SQL or a graphical user interface. All this depends heavily on who the data is for (e.g. is it for data scientists or corporate business leaders?) and how they intend to use or access it.

In the future, this storage facility might also be used as a source of data for another data processing cycle.

Task: Draw out the circle of data processing with some highlighted notes of each step (name of step, responsibilities)

Data visualization

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Data visualization can be utilized for a variety of purposes, and it's important to note that is not only reserved for use by data teams. Management also leverages it to convey organizational structure and hierarchy while data analysts and data scientists use it to discover and explain patterns and trends. Harvard Business Review (link resides outside ibm.com) categorizes data visualization into four key purposes: idea generation, idea illustration, visual discovery, and everyday data visualization. We'll delve deeper into these below:

Idea generation

Data visualization is commonly used to spur idea generation across teams. They are frequently leveraged during brainstorming or Design Thinking sessions at the start of a project by supporting the collection of different perspectives and highlighting the common concerns of the collective. While these visualizations are usually unpolished and unrefined, they help set the foundation within the project to ensure that the team is aligned on the problem that they're looking to address for key stakeholders.

Idea illustration

Data visualization for idea illustration assists in conveying an idea, such as a tactic or process. It is commonly used in learning settings, such as tutorials, certification courses, centers of excellence, but it can also be used to represent organization structures or processes, facilitating communication between the right individuals for specific tasks. Project managers frequently use Gantt charts and waterfall charts to illustrate workflows. Data modeling also uses abstraction to represent and better understand data flow within an enterprise's information system, making it easier for developers, business analysts, data architects, and others to understand the relationships in a database or data warehouse.

Visual discovery

Visual discovery and every day data viz are more closely aligned with data teams. While visual discovery helps data analysts, data scientists, and other data professionals identify patterns and trends within a dataset, every day data viz supports the subsequent storytelling after a new insight has been found.

Data visualization

Data visualization is a critical step in the data science process, helping teams and individuals convey data more effectively to colleagues and decision makers. Teams that manage reporting systems typically leverage defined template views to monitor performance. However, data visualization isn't limited to performance dashboards. For example, while text mining an analyst may use a word cloud to to capture key concepts, trends, and hidden relationships within this unstructured data. Alternatively, they may utilize a graph structure to illustrate relationships between entities in a knowledge graph. There are a number of ways to represent different types of data, and it's important to remember that it is a skillset that should extend beyond your core analytics team.

Use this model selection framework to choose the most appropriate model while balancing your performance requirements with cost, risks and deployment needs.

Advantages and disadvantages of data visualization

Something as simple as presenting data in graphic format may seem to have no downsides. But sometimes data can be misrepresented or misinterpreted when placed in the wrong style of data visualization. When choosing to create a data visualization, it's best to keep both the advantages and disadvantages in mind.

Advantages

Our eyes are drawn to colors and patterns. We can quickly identify red from blue, and squares from circles. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

Some other advantages of data visualization include:

- Easily sharing information.
- Interactively explore opportunities.
- Visualize patterns and relationships.

Disadvantages

While there are many advantages, some of the disadvantages may seem less obvious. For example, when viewing a visualization with many different datapoints, it's easy to make an inaccurate assumption. Or sometimes the visualization is just designed wrong so that it's biased or confusing.

Some other disadvantages include:

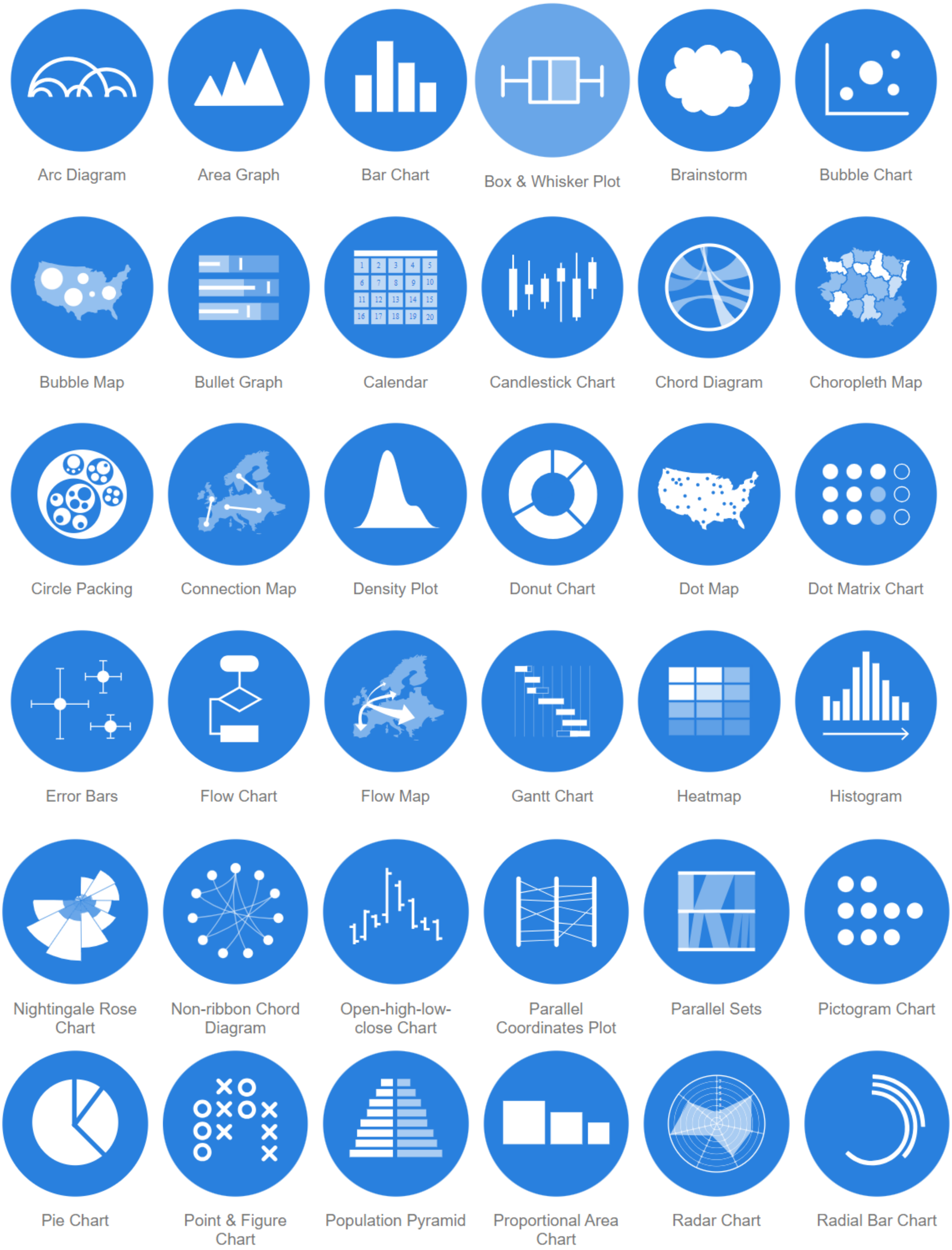
- Biased or inaccurate information.
- Correlation doesn't always mean causation.
- Core messages can get lost in translation.

Types of data visualizations

The earliest form of data visualization can be traced back the Egyptians in the pre-17th century, largely used to assist in navigation. As time progressed, people leveraged data visualizations for broader applications, such as in economic, social, health disciplines. Perhaps most notably, Edward Tufte published The Visual Display of Quantitative Information (link resides outside ibm.com), which illustrated that individuals could utilize data visualization to present data in a more effective manner. His book continues to stand the test of time, especially as companies turn to dashboards to report their performance metrics in real-time. Dashboards are effective data visualization tools for tracking and visualizing data from multiple data sources, providing visibility into the effects of specific behaviors by a team or an adjacent one on performance. Dashboards include common visualization techniques, such as:

- **Tables:** This consists of rows and columns used to compare variables. Tables can show a great deal of information in a structured way, but they can also overwhelm users that are simply looking for high-level trends.
- **Pie charts and stacked bar charts:** These graphs are divided into sections that represent parts of a whole. They provide a simple way to organize data and compare the size of each component to one other.
- **Line charts and area charts:** These visuals show change in one or more quantities by plotting a series of data points over time and are frequently used within predictive analytics. Line graphs utilize lines to demonstrate these changes while area charts connect data points with line segments, stacking variables on top of one another and using color to distinguish between variables.
- **Histograms:** This graph plots a distribution of numbers using a bar chart (with no spaces between the bars), representing the quantity of data that falls within a particular range. This visual makes it easy for an end user to identify outliers within a given dataset.
- **Scatter plots:** These visuals are beneficial in reveling the relationship between two variables, and they are commonly used within regression data analysis. However, these can sometimes be confused with bubble charts, which are used to visualize three variables via the x-axis, the y-axis, and the size of the bubble.
- **Heat maps:** These graphical representation displays are helpful in visualizing behavioral data by location. This can be a location on a map, or even a webpage.
- **Tree maps,** which display hierarchical data as a set of nested shapes, typically rectangles. Treemaps are great for comparing the proportions between categories via their area size.

EXERCISE IS BELOWED!


Arc Diagram
Area Graph
Bar Chart
Box & Whisker Plot
Brainstorm
Bubble Chart
Bubble Map
Bullet Graph
Calendar
Candlestick Chart
Chord Diagram
Choropleth Map
Circle Packing
Connection Map
Density Plot
Donut Chart
Dot Map
Dot Matrix Chart
Error Bars
Flow Chart
Flow Map
Gantt Chart
Heatmap
Histogram
Nightingale Rose Chart
Non-ribbon Chord Diagram
Open-high-low-close Chart
Parallel Coordinates Plot
Parallel Sets
Pictogram Chart
Pie Chart
Point & Figure Chart
Population Pyramid
Proportional Area Chart
Radar Chart
Radial Bar Chart

Radial Column Chart

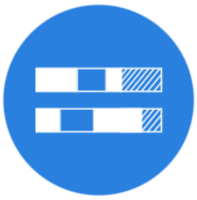
Sankey Diagram

Scatterplot

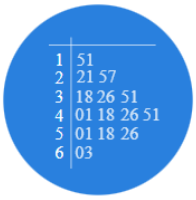
Span Chart

Spiral Plot

Stacked Area Graph



Stacked Bar Graph



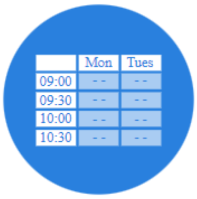
Stem & Leaf Plot

Stream Graph

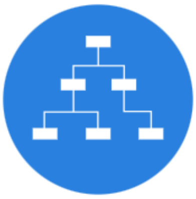
Sunburst Diagram

Tally Chart

Timeline



Timetable



Tree Diagram

Treemap

Venn Diagram

Violin Plot

Word Cloud

Task: Make a table to categorize those charts above in right type of data visualization

III.4. Programming Languages

Programming languages

Unfortunately, computers cannot understand ordinary spoken English or any other natural language. The only language they can understand directly is called machine code. This consists of the 1s and 0s (binary codes) that are processed by the CPU.

However, machine code as a means of communication is very difficult to write. For this reason, we use symbolic languages that are easier to understand. Then, by using a special program, these languages can be translated into machine code. For example, the so-called assembly languages use abbreviations such as ADD, SUB, MPY to represent instructions. These mnemonic codes are like labels easily associated with the items to which they refer.

Basic languages, where the program is similar to the machine code version, are known as low-level languages. In these languages, each instruction is equivalent to a single machine code instruction, and the program is converted into machine code by a special program called an assembler. These languages are still quite complex and restricted to particular machines.

To make the programs easier to write and to overcome the problem of intercommunication between different types of machines, higher-level languages were designed such as BASIC, COBOL, FORTRAN or PASCAL. These are all problem-oriented rather than machine-oriented. Programs written in one of these languages (known as source programs) are converted into a lower-level language by means of a compiler (generating the object program). On compilation, each statement in a high-level language is generally translated into many machine code instructions.

People communicate instructions to the computer in symbolic languages and the easier this communication can be made, the wider the application of computers will be. Scientists are already working on Artificial Intelligence and the next generation of computers may be able to understand human languages.

Ex. 1. Answer the following questions.

1. Do computers understand human languages?
2. What are the differences between low-level and high-level languages?
3. What is an assembler?
4. What is the function of compiler?
5. What do you understand by the terms source program and object program?
6. In the future, could computers be programmed in Spanish, French or Japanese?

Task 1. a) Look at the groups of words and decide what part of speech each word is. Then complete the sentences with the correct word.

Compile, compiler, compilation.

1. Programs written in a high-level language require …, or translation into machine code.
2. A … generates several low-level instructions for each source language statement.
3. Programmers usually … their programs to create an object program and diagnose possible errors.

Program, programmers, programming, programmable.

4. Most computer … make a plan of the program before they write it. This plan is called a flowchart.
5. A computer … is a set of instructions that tells the computer what to do.
6. Converting an algorithm into a sequence of instructions in a programming language is called …

Bug, debug, debugger, debugging.

7. New programs need … to make them work properly.
8. Any error or malfunction of a computer program is known as a …
9. The best compilers usually include an integrated … which detects syntax errors.

b) In the word debug the prefix de- is used. This prefix means 'to reverse an action'. Here are a few more examples:

Defrost, debrief, declassify, decode, decompose, decentralize.

Write down the base form of each verb. What do the verbs mean in your language? And what do the verbs with *de-* ***mean?***

c) Can you think of any more verbs with *de-* ***in English?***

Database

Task

a) Companies often use databases to store information about customers, suppliers and their own personnel. Try to answer these questions.

1. What is a database?
2. Which tasks can be performed by using a database? Make a list of possible applications.
3. "What do the terms mean in your language: File, record, field?

b) Here is part of an article about databases. First, read all the way through and underline the basic features of a database.

Basic features of database programs

With a **database** you can store, organize and retrieve a large collection of related information on computer. If you like, it is the electronic equivalent of an indexed filing cabinet. Let us look at some features and applications.

- Information is entered on a database via **fields.** Each field holds a separate piece of information, and the fields are collected together into **records.** For example, a record about an employee might consist of several fields which give their name, address, telephone number, age, salary and length of employment with the company. Records are grouped together into **files** which hold large amounts of information. Files can easily be updated: you can always change fields, add new records or delete old ones. With the right database software, you are able to keep track of stock, sales, market trends, orders, invoices and many more details that can make your company successful.
- Another feature of database programs is that you can automatically look up and find records containing particular information. You can also search on more than one field at a time. For example, if a managing director wanted to know all the customers that spend more than £7,000 per month, the program would search on the name field and the money field simultaneously.

A computer database is much faster to consult and update than a card index system. It occupies a lot less space, and records can be automatically sorted into numerical or alphabetical order using any field.

The best packages also include networking facilities, which add a new dimension of productivity to businesses. For example, managers of different departments can have direct access to a common database, which represents an enormous advantage. Thanks to security devices, you can share part of your files on a network and control who sees the information. Most aspects of the program can be protected by user-defined passwords. For example, if you wanted to share an employee's personal details, but not their commission, you could protect the commission field.

In short, a database manager helps you control the data you have at home, in the library or in your business.

Ex. 1. Now make a list of the words you don't understand. Can you guess their meaning? Compare your ideas with other students.

Ex. 2. Using the information in the text, complete these statements.

1. A database is used to … .
2. information is entered on a database via ... .
3. Each field holds ….
4. 'Updating' a file means ….
5. The advantages of a database program over a manual filing system are ….
6. Access to a common database can be protected by using ….

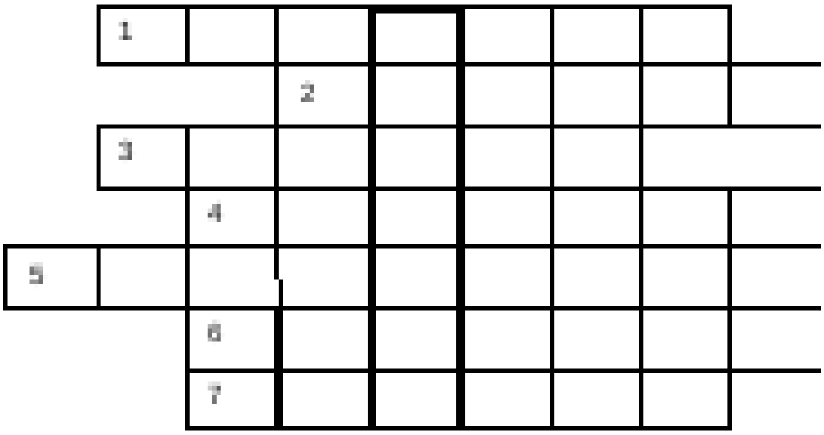
Ex. 3. Translate into English the words and expressions given in brackets.

1. (Hãy xem xét) at some features and applications.
2. Records (tập hợp) together into files.
3. Files can (dễ dàng) be updated; you can always (thay đổi) fields, (thêm) new records or (xóa/hủy) old ones.
4. You can find records, (kèm theo) particular information
5. You can also search on more than one field (cùng một lúc).

Task 1. a) Complete the sentences by using a term from the list. Then write the words in the crossword to find the hidden message.

Puzzle

database field layout merging record sorted updated



1. In order to personalize a standard letter you can use 'mail … ' (a technique which consists of combining a database with a document made with a word processor).
2. Records can be automatically … into any order.
3. You can decide how many fields you want to have on a … .
4. Files can easily be … by adding new information or deleting the old one.
5. A … program can be used to store, organize and retrieve information of any kind.
6. The … of the records can be designed by the user.
7. Each piece of information is given in a separate … .

b) Write the plural of these words.

1. slot
2. key
3. directory
4. businessman
5. fax
6. mouse
7. floppy
8. virus

A short description of Visual Basic

Visual Basic is a programming language and an environment (1) (develop) by Microsoft in 1990. It (2) (use) to create applications for Windows operating systems.

The name ‘BASIC’ (3) (stand) for Beginner’s All-purpose Symbolic Instruction Code. The original BASIC language (4) (create) in 1965 and adopted by many programmers and PC manufacturers because it was user-friendly and easy to learn.

The adjective ‘Visual’ refers to the technique used to create a graphical user interface. Instead of (5) (write) a lot of instructions to describe interface elements, you just (6) (add) pre-defined objects such as buttons and dialog boxes, which can (7) (choose) from a toolbox. It (8) (take) only a few minutes to create a Visual BASIC program. Using the mouse, you simply (9) (drag) and drop controls (e.g. option buttons, text boxes, icons, menu bars, etc.) into the required position, and then define their colour, size and behaviour.

Thanks to its object-oriented philosophy and interactive nature, Visual BASIC (10) (enable) the programmer to quickly create all sorts of applications from small system utilities to database programs and Internet server applications.

Task 1. Work in pairs.

a) Student A, complete the table by asking for information, like this:

1. What does ‘COBOL’ mean?
2. ‘COBOL’ stands for…
3. When was it developed?
4. In …
5. What’s it used for?
6. It’s used for …
7. What features has it got?
8. It is easy to use and it’s written in English. It can handle very large data files.

b) Answer your partner’s questions too.

Computer language	Date	Characteristics	Uses
COBOL (Common Business Oriented Language)	1958-1959	Easy to read. Able to handle very large files. Written in English.	Mainly used for business applications.

BASIC … .	…..	…	General purpose language. Used to teach programming.
Pascal (named after …)	1970-1973	Structured language with algorithmic features designed for fast execution of the object program. A fast compiler called Turbo Pascal was created in 1982-very popular.	…
LOGO	1969	…	Designed for use in schools to encourage children to experiment with programming.
SQL (…) Introduced by Oracle Corp.	1979	Supports distributed databases, which run on several computer systems. Allows various users on a LAN to access the same database at the same time.	…

c) Student B, complete the table by asking for information, like this:

1. What does 'COBOL' mean?
2. 'COBOL' stands for …
3. When was it developed?
4. In…
5. What's it used for?
6. It is used for…
7. What features has it got?
8. It is easy to use and it's written in English. It can handle very large data files.

d) Answer your partner's questions too

Computer language	Date	Characteristics	Uses
COBOL (Common Business Oriented Language)	1958-1959	Easy to read. Able to handle very large files. Written in English.	Mainly used for business applications.
… (Beginner's All-purpose	1964-1965	High-level programming language. Interactive.	…

Symbolic Instruction Code)		Easy to learn. Displays error messages that help users to correct mistakes. Has a large number of dialects.	
Pascal (named after the famous scientist Blaise Pascal)	1970-1973	…	General purpose. Often used in colleges and universities to teach programming.
LOGO	…	Easy to learn. Flexible- it can do maths, make lists, construct graphs, etc. Its drawing capabilities allow children to construct simple graphics programs.	…
Structured Query Language	…	…	A standard query language used for requesting information from a database. It allows users to specify search criteria in databases.

Task 2. Summarize everything you've learnt about these computing languages.

TEXT B

Task

a) The term 'Java' refers to three things:

- an island in Indonesia

-a cup of coffee, in American slang

-a language for Internet applications.

But what exactly is Java?

b) Can you guess the meaning of these words? Use the dictionary if necessary.

Applet, object-oriented, compiler, plug-in, real-time, download.

c) Read the text.

What is Java?

Java is a programming language developed by Sun Microsystems. When you see a web page that uses Java, a small program called 'applet' is done automatically. Java applets let you watch animated characters and moving text, play music and interact with information on the screen.

Characteristics of the Java language.

Java is an object-oriented language similar to C++, but it is more dynamic and simplified to eliminate possible programming errors.

A Java program is both compiled and interpreted. First the source code is compiled and converted into a format called bytecode, which can then be executed by a Java interpreter. Compiled Java code run on most computers, because there are Java interpreters, known as Java Virtual Machines.

Java is a multi-threaded. A Java program can have multiple threads (parts), i.e. many different things processing independently and continuously.

Why is Java cool?

Java lets you create moving images and animated drawings. You can also create graphical objects (e.g. bar charts, graphs, diagrams) and new ‘controls’ (e.g. buttons, check boxes, pushbuttons with special properties). A web page that uses Java can have inline sounds that play in real-time, music that plays in the background, cartoon style animations, real-time video and interactive games.

Alternatives to Java.

One alternative technology is ActiveX, the Microsoft product for including multimedia effects on web pages. Another competitor is Macromedia’s Shockwave, a plug-in that lets you animate pictures, add sound and even make interactive pages so that people can play games on websites.

Ex. 1. These statements about Java are all false. Correct them.

1. Java was invented by Microsoft.
2. Small applications written in Java are called ‘animations’.
3. With the interpreter, a program is first converted into Java bytecodes.
4. Java is not compatible with most computing platforms.
5. The Java language is single-threaded, one part executing at a time.
6. Java doesn’t let you watch animated characters on your webpages.
7. ActiveX and Shockwave are not real competitors for Java.

Ex. 2. Solve the anagrams in the right-hand column and match them with the words in the left-hand column to complete the phrases. The first one has been done for you.

1. high-level	a. mestnttae
2. machine	b. thirmacite
3. systems	c. peat
4. object	d. taporeor
5. linkage	e. omelud
6. magnetic	f. egguanal - language
7. binary	g. trodite
8. declaration	h. deco
9. comment	i. enil
10. relational	j. nituroe

Ex. 3. a) Match each word on the left with its partner to make a common technical term.

1. programming	a. browser

2. web	b. error
3. Java	c. code
4. multimedia	d. protection
5. source	e. format
6. virus	f. effects
7. compression	g. applet

b) Which verbs on the left are often found with nouns on the right?

1. to download	a. the Web
2. to play	b. a source program
3. to run	c. files
4. to browse	d. an application
5. to compile	e. data
6. to process	f. music

III.5. Internet Security

Internet issues reading material

TEXT A

Ex 1: *Trước khi đọc bài, trả lời những nội dung dưới đây*

a) Try to answer these questions

1. Is it technically possible for computer criminals to infiltrate into the Internet and steal sensitive information?
2. What is a hacker?
3. Can viruses enter your PC from the Internet?

b) Translate the sentences (1-4) into Vietnamese.

1. Web browsers warn you if the connection is not secure; they display a message when you try to send personal information to a server.
2. Private networks use a software and hardware mechanism, called a 'firewall', to block unauthorized traffic from the Internet.
3. You have to type your username and password to access a locked computer system or network.
4. An open padlock in Netscape Communicator indicates the page is not secure; a closed padlock indicates the page is encrypted (secure).

c) Read the text and do the exercises below.

Security and privacy on the Internet

There are a lot of benefits from an open system like the Internet, but we are also exposed to hackers who break into computer systems just for fun, as well as to steal information or propagate viruses. So how do you go about making online transactions secure?

Security on the Web

The question of security is crucial when sending confidential information such \
;/s credit card numbers. For example, consider the process of buying a book on the Web. You have to type your credit card number into an order form which passes from computer to computer on its way to the online bookstore. If one of the intermediary computers is infiltrated by hackers, your data can be copied. It is difficult to say how often this happens, but it's technically possible.

To avoid risks, you should set ail security alerts to high on your Web browser. Netscape Communicator and Internet Explorer display a lock when the Web page is secure and allow you to disable or delete “cookies”.

If you use online bank services, make sure your bank uses digital certificates. A popular security standard is SET (secure electronic transactions).

E-mail privacy

Similarly, as your e-mail message travels across the net, it is copied temporarily on many computers in between. This means it can be read, by unscrupulous people who illegally enter computer systems.

The only way to protect a message is to put it in a sort of 'envelope', that is, to encode it with some form of encryption. A system designed to send e-mail privately is Pretty Good Privacy, a freeware program written by Phil Zimmerman.

Network security

Private networks connected to the Internet can be attacked by intruders who attempt to take valuable information such as Social Security numbers, bank accounts or research and business reports.

To protect crucial data, companies hire security consultants who analyze the risks and provide security solutions. The most common methods of protection are passwords for access control, encryption and decryption systems, and firewalls.

Virus protection

Viruses can enter a PC through files from disks, the Internet or bulletin board systems. If you want to protect your system, don't open e-mail attachments from strangers and take care when downloading files from the Web. (Plain text e-mail alone can't pass a virus.)

Remember also to update your anti-virus software as often as possible, since new viruses are being created all the time.

Preventative tips

Don’t open email attachments from unknown people; always take note of the file extension.

Run and update antivirus programs, e.g. virus scanners.

Install a firewall, a program designed to prevent spyware from gaining access to the internal network.

Make backup copies of your files regularly.

Don’t accept files from high-risk sources.

Use a digital certificate, an electronic way of proving your identity, when you are doing business on the Internet. Avoid giving credit card numbers.

Don’t believe everything you read on the Net. Have a suspicious attitude toward its contents.

Help box

- **hacker**: a person who obtains unauthorized access to computer data
- **cookies**: small files used by Web servers to know if you have visited their site before
- **certificates**: files that identify users and Web servers on the net, like digital identification cards
- **encryption**: the process of encoding data so that unauthorized users can't read it
- **decryption**: the process of decoding encrypted data transmitted to you

Ex. 2. Find the answers to these questions.

1. Why is security so important on the Internet?
2. What security features are offered by Netscape Communicator and Internet Explorer?
3. What security standard is used by most banks to make online transactions secure?
4. How can we protect and keep our e-mail private?
5. What methods are used by companies to make internal networks secure?
6. Which ways can a virus enter a computer system?

Ex.3. Complete these sentences by using a term from the text. *Điền những chữ còn thiếu trong từ để thành câu hoàn chỉnh.*

1. Users have to enter a p … to gain access to a network.
2. You can download a lot of f … or public domain programs from the net.
3. Hundreds of h … break into computer systems every year.
4. A computer v … can infect your files and corrupt your hard disk.
5. The process of encoding data so that unauthorized users can't read the data is known as e … .
6. A f … is a device which allows limited access to an internal network from the Internet.
7. You can include an a … as part of your e-mail message.
8. This company uses d … techniques to decode (or decipher) secret data.

Ex. 4. Fill in the gaps in these security tips with words from the box.

digital, certificate, malware, virus, scanner, spyware, firewall, antivirus

1. Malicious software, (1) … , can be avoided by following some basic rules.

2. Internet users who like cybershopping should get a (2) … , an electronic identity card.
3. To prevent crackers from breaking into your internal network and obtaining your data, install a (3) … . It will protect you from (4) … .
4. If you have been hit by a (5) … , don't panic! Download a clean-up utility and always remember to use on (6) … program, for example, a virus (7) ….

TEXT B

Internet crime

The Internet provides a wide variety of opportunities for communication and development, but unfortunately it also has its dark side.

Crackers, or black-hat hackers, are computer criminals who use technology to perform a variety of crimes: virus propagation, fraud, intellectual property theft, etc.

Internet-based crimes include scam, email fraud to obtain money or valuables, and phishing, bank fraud, to get banking information such as passwords of Internet bank accounts or credit card details. Both crimes use emails of websites that look like those of real organizations.

Due to its anonymity, the Internet also provides the right environment for cyberstalking, online harassment or abuse, mainly in chat rooms or newsgroups.

Piracy, the illegal copying and distribution of copyrighted software, information, music and video files, is widespread.

But by far the most common type of crime involves malware.

Malware: viruses, worms, trojans and spyware

Malware (malicious software) is software created to damage or alter the computer data or its operations. These are the main types.

- Viruses arc programs that spread by attaching themselves to executable files or documents. When the infected program is run, the virus propagates to other files or programs on the computer. Some viruses are designed to work at a particular time or on a specific date, e.g. on Friday 13th. An email virus spreads by sending a copy of itself to everyone in an email address book.
- Worms are self-copying programs that have the capacity to move from one computer to another without human help, by exploiting security flaws in computer networks. Worms are self-contained and don't need to be attached to a document or program the way viruses do.
- Trojan horses are malicious programs disguised as innocent-looking files or embedded within legitimate software. Once they are activated, they may affect the computer in a variety of ways: some are just annoying, others are more ominous, creating a backdoor to the computer which can be used to collect stored data. They don't copy themselves or reproduce by infecting other files.

- Spyware, software designed to collect information from computers for commercial or criminal purposes, is another example of malicious software. It usually comes hidden in fake freeware or shareware applications downloadable from the Internet.

Ex 5. Identify the Internet crimes sentences (1-6) refer to. Then match them with the advice below (a-f). *Xác định những câu từ 1 – 6 liên quan đến loại tội phạm Internet nào và nối với a-f về những lời khuyên trong xử lý vấn đề đó.*

1. Crackers try to find a way to copy the latest game or computer program.
2. A study has revealed that half a million people will automatically open an email they believe to be from their bank and happily send off all their security details.
3. This software's danger is hidden behind an attractive appearance. That's why it is often wrapped in attractive packages promising photos of celebrities like Anna Kournikova or Jennifer Lopez.
4. There is a particular danger in Internet commerce and emails. Many people believe they have been offered a special gift only to find out later they have been deceived.
5. 'Nimda' spreads by sending infected emails and is also able to infect websites, so when a user visits a compromised website, the browser can infect the computer.
6. Every day, millions of children spend time in Internet chat rooms talking to strangers. But what many of them don't realize is that some of the surfers chatting with them may be sexual predators.

   a. People shouldn't buy cracked software or download music illegally from the Internet.
   b. Be suspicious of wonderful offers. Don't buy if you aren't sure.
   c. It's dangerous to give personal information to people you contact in chat rooms.
   d. Don’t open attachments from people you don't know even if the subject looks attractive.
   e. Scan your email and be careful about which websites you visit.
   f. Check with your bank before sending information.

TEXT C

Hackers!

Sept ‘70 John Draper, also known as Captain Crunch, discovers that the penny whistle offered in boxes of Cap'n Crunch breakfast cereal perfectly generates the 2,600 cycles per second (Hz) signal that AT&T used to control its phone network at the time. He starts to make free calls.

Aug ‘74 Kevin Mitnick, a legend among hackers, begins his career, hacking into banking networks and destroying data, altering credit reports of his enemies, and disconnecting the phone lines of celebrities. His most famous exploit – hacking into the North American Defense Command in Colorado Springs – inspired *War Games*, the 1983 movie.

Jul ’81 Ian Murphy, a 23-year-old known as Captain Zap on the networks, gains instant notoriety when he hacks into the White House and the Pentagon.

*Dec ’*87 IBM international network is paralysed by hacker's Christmas message.

Jul ’88 Union Bank of Switzerland 'almost' loses £32 million to hacker
criminals. Nicholas Whitely is arrested in connection with virus propagation.

Oct ’89 Fifteen-year-old hacker cracks US defense computer.

Nov ’90 Hong Kong introduces anti-hacking legislation.

Aug ’91 Israelis arrest 18-year-old for hacking foreign banking and credit card networks.

Dec ’92 Kevin Poulsen, known as 'Dark Dante' on the networks, is charged with stealing tasking orders relating to an Air Force military exercise. He is accused of theft of US national secrets and faces up to 10 years in jail.

Feb ‘97 German Chaos Computer Club shows on TV the way to electronically obtain money from bank accounts using a special program on the Web.

May ’99 Computer criminals propagate a lot of viruses through the Internet.

Feb ’00 A massive 'denial of service' attack is launched against websites like Yahoo, Amazon and eBay.

Aug ’01 'Code Red' computer worm infects many PCs through the Internet.

Ex 6. Answer the following questions.

1. Which hacking case inspired the film War Games?
2. Why was Nicholas Whitely arrested in 1988?
3. How old was the hacker that cracked the US defense computer in October 1989?
4. Who was known as 'Dark Dante' on the networks? What was he accused of?
5. Which computer club showed on TV a way to attack bank accounts?
6. What type of virus infected thousands of PCs in 2001?

---- THE END! -----

TÀI LIỆU THAM KHẢO

[1] https://academy.constructor.org/blog/data-science-terminology
[2] https://www.hariduskeskus.ee/opiobjektid/ingliseit