Data Analytics

Foundational Curricula:
Cluster 8: Data
Module 15: Data Analytics, Modeling and Reporting
Unit 1: Data Analytics
FC-C8M15U1

Curriculum Developers: Angelique Blake, Rachelle Blake, Pauliina Hulkkonen, Sonja Huotari, Milla Jauhiainen, Johanna Tolonen, and Alpo Värri

43/60

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Unit Objectives

- Define basic data analysis
- Describe the functions of a database and basic data collection techniques
- Explain the difference types of data storage including cloud, warehouse, server etc.
- Explain the use of multiple identifiers (i.e. MPI, medical record number, etc.) and unique identifiers (i.e. primary key)
- Explain what it means to anonymize data and identify techniques that can be used to do this
- Describe basic data mining and knowledge acquisition principles
- Describe basic data principles including data representation, data types, etc.
- Describe measurement sampling techniques
- Identify common health information/eHealth standardized data sets
- List the elements of a common clinical data set, explaining key data categories

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Basics of Data Analysis

- **Data analysis** is the process of inspecting, transforming and modeling data to discover useful information and suggest conclusions
- Data analysis includes some pre-processing, testing, combining models and algorithms, and finally making a decision or a conclusion based on the data set
- Important part of data analysis is the visualization. Curves, fittings and representations help to explain the results in a common way

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Example of Data Analysis process

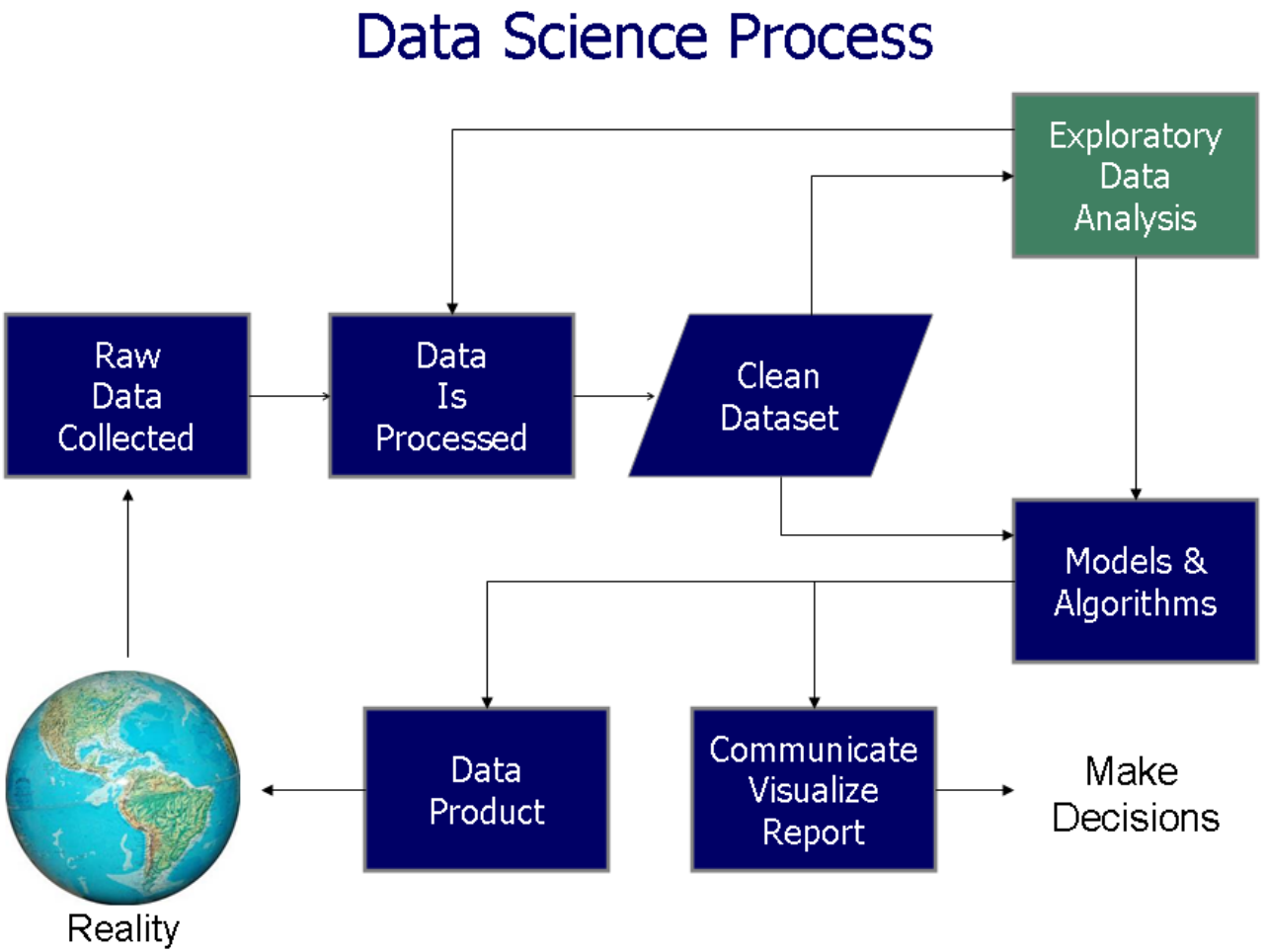


Figure by Farcaster (CC BY-SA 3.0)

This work is produced by the EU-US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552 EUUSEHEALTHWORK

Database and Data Collection

- **Database** means any collection of data or information that is specially organized for rapid search and retrieval by a computer. Databases are structured to facilitate the storage, retrieval, modification, and deletion of data in conjunction with various data-processing operations. (Encyclopedia Britannica)
- **Data collection** is the process of gathering the information. While different fields might have different methods for the collection procedure, some principles apply for all fields:
  - Ensuring accurate and appropriate data
  - Maintaining the integrity of the collection

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552 EUUSEHEALTHWORK

Data Storage Locations

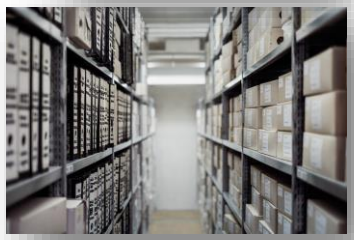
- Data can be stored in various ways, just as your personal documents are stored differently at your home (photos, books, music, movies, legal documents, prescriptions,..). Examples of ways of storing data:

Cloud services

- **Cloud services** are online storage formats for large amounts of data. The data is stored in logical pools and to multiple physical locations, thus retrieving old information is possible.

Warehouses

- In **Warehouses** data is stored from operational systems. Warehouse maintains a copy of information



Servers

- **Servers** are physical storage computers that are only meant for storing the data

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Multiple Identifiers & Unique Identifiers

- **Master patient index (MPI)** is a database for identifying patients across various departments. A unique identifier for each patient is produced from e.g. name, gender, date of birth, social security number, …
- **Medical record number** is a hospital specific database for discriminating the patient's medical history
- **Primary key** is the piece of information that does not change (e.g. ID). When reporting an address, it should be stored with e.g. social security number

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Anonymization and De-Identification of Data

- It is not intended to collect data with personal information, when conducting research or making public health analysis
- Personal information may cause biased results, or the information could be wrongly used
- Anonymizing and de-identification of data helps to protect the participants but also to conduct better research.

Techniques:

  - Patients are marked with ID's, ID-name information is stored only by the data collector, e.g. a hospital
  - Anonymization means the removal of personal information from the data
  - Also, data could be collected anonymously without ever collecting personal information (for example hand-written surveys)

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

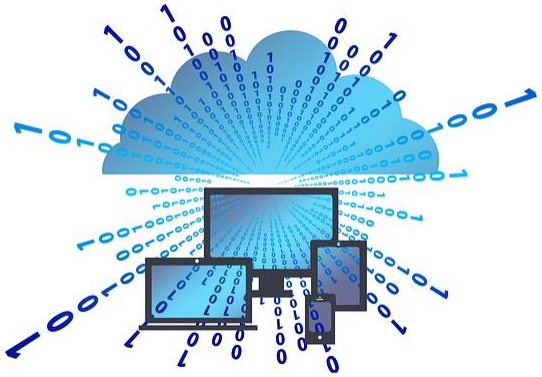
Data Mining and Knowledge Acquisition Principles

- **Data Mining** is the process of discovering useful patterns from large volumes of data. The field includes statistical testing as well as artificial intelligence with database management
- **Knowledge acquisition** includes changing the data into program readable format. For example changing the data from interview or surveys to tables, knowledge to computer program, ..

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Basics of Data Principles

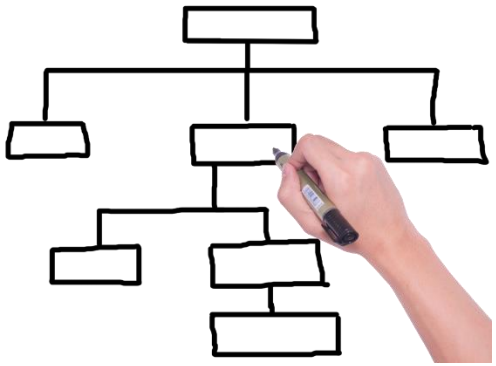
- The basic type of all data on a computer is just a sequence of logical values TRUE (1) or FALSE (0)
  - Different sequences of 0's and 1's can be stored to represent numbers, alphabets, special characters or other data types
  - Different document formats are created to store images, sound, text or other signals efficiently and without loosing information



This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Measurement Sampling

- To collect information about a specific topic, a sampling is needed to collect the data. One cannot possibly measure every “female smoker in the US”, but the researcher has to collect a population which represents the whole population at a given accuracy.
  - E.g. collect 100 female smokers and 100 female non-smokers to create comparison between the groups
- Sampling can be done in various ways
  - Random sampling – every member has an equal chance
  - Stratified sampling – population divided into subgroups and members are randomly selected from each group
  - Systematic sampling – uses a specific system to select members such as every 10th person on an alphabetized list
  - Cluster random sampling – divides the population into clusters, clusters are randomly selected and all members of the cluster selected are sampled



This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Common health information/eHealth standardized data sets: Patient Summary Dataset

PATIENT ADMINISTRATIVE DATA:

- Identification
- Personal Information
- Contact Information
- Insurance Information

CLINICAL DATA:

- Alerts: allergy, medical
- Medical history: vaccinations, problem list, surgical procedures, diagnoses, devices/implants, surgical procedures, treatment recommendations
- Medication summary: current medication list
- Social history
- Pregnancy history
- Physical findings: vital signs
- Diagnostic tests

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Common health information/eHealth standardized data sets: Patient Summary Dataset (cont'd)

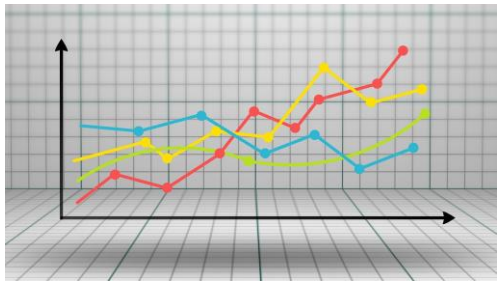
METADATA:

- Country
- Patient Summary
- Last Update
- Nature of Patient Summary
- Author Organization

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Healthcare Data Categories

- Demographic Data consists of elements that distinguish one patient from another, such as name, address and birth date.
- Socioeconomic Data consists of information of personal life and habits, such as marital status, religion and culture.
- Financial Data consists information of the payer, e.g. the name, address and telephone number of the patient's insurance company.
- Clinical Data consists of all of the medical data that have been recorded during the care process



Healthcare data is not only used by the caregivers, but also lawyers, researchers, hospital administration and others

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Unit Review Checklist

- ❑ Defined basic data analysis (GB16)
- ❑ Described the functions of a database and basic data collection techniques (GB05)
- ❑ Explained the difference types of data storage including cloud, warehouse, server etc. (GB06)
- ❑ Explained the use of multiple identifiers (i.e. MPI, medical record number, etc.) and unique identifiers (i.e. primary key) (GB08)
- ❑ Explained what it means to anonymize data and identify techniques that can be used to do this (GB09)
- ❑ Described basic data mining and knowledge acquisition principles (GB18)
- ❑ Described basic data principles including data representation, data types, etc. (GB19)
- ❑ Described measurement sampling techniques (GB19)
- ❑ Identified common health information/eHealth standardized data sets
- ❑ Listed the elements of a common clinical data set, explaining key data categories

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552 EUUSEHEALTHWORK

Unit Review Exercise/Activity

1. Identify different ways of storing data.
2. Give examples of following data types:
   a) Demographic Data
   b) Socioeconomic Data
   c) Financial Data
   d) Clinical Data

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Unit Exam

1. Which of these are methods to anonymize healthcare data?
   a) Personal data (such as name or social security number) are not collected at all, e.g. an anonymous hand-written survey
   b) Name, birth date etc are removed from the final data
   c) Personal IDs are substituted with identification code
   d) All of the above
2. Data analysis includes a decision making process based on the data set that has been analyzed.
   a) True
   b) False

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK

Unit Exam

3. Which of the sampling types is used when population is divided into subgroups and members are randomly selected from each group?
   a) Random sampling
   b) Stratified sampling
   c) Systematic sampling
   d) Cluster random sampling

This work is produced by the EU*US eHealth Work Project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 727552
EUUSEHEALTHWORK