Lecture 2.0: Data & Python & Sequence

Course: BMEG3105, Data Analytics in Personalized Genomics and Precision Medicine

Professor: Yu Li

Scriber: Asset Yermukhanbet

Brief Summary

Modules	**Key Points**
Recap from lecture on Wednesday	- What data can be used in healthcare?
Data: What and How?	- What is Data? - What are some existing Data Types? - Summary
Python 101: Introduction to the Programming Language	- What is programming? - What is Python? - What is a package (e.g. NumPy)?

Module I

Recap from Lecture on Wednesday

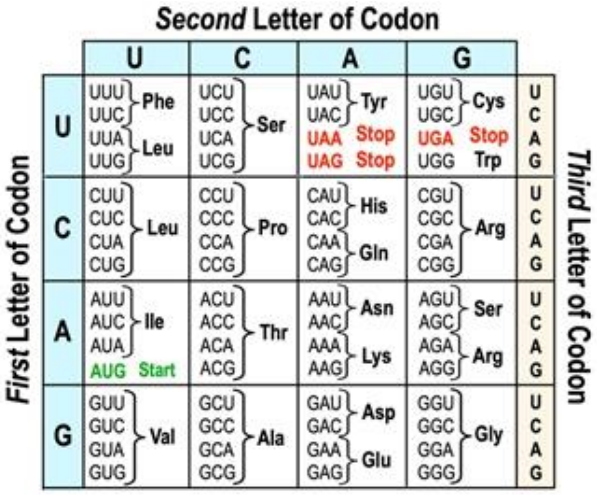
What data can be used in healthcare? Also Known As, What data we use to analyze patients?	Any data that might vary from patient to patient and doesn't not appear exactly the same in every circumstance to two different beings. From microscopic level, such as, gene sequence or protein chain, to macroscopic level, such as family tree or environment/climate exposure. All of these data examples are important in analyzing a patient.

Module II

Data: What and How?

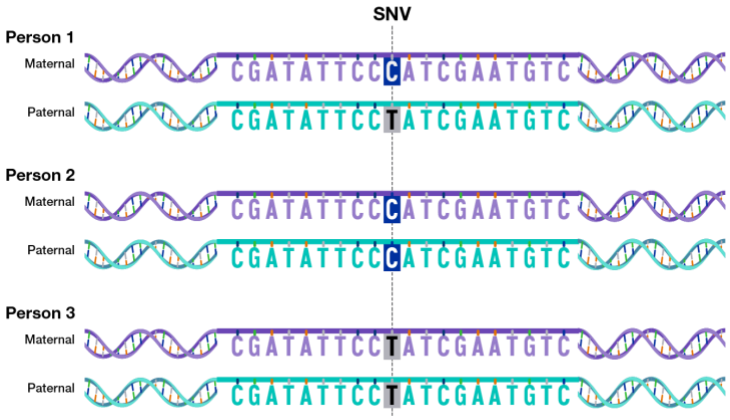
What is Data?	**Data** is an information, especially facts or numbers, collected to be examined and considered and used to help **decision-making process**, or information in an electronic form that can be stored and used by a computer. [Cambridge Dictionary] By definition, data is **contextually dependent** since its functionality lies in facilitating a *decision-making* process. Such process could be utilized in various fields, including personalized genomics and precision medicine; various health-related data (e.g., family tree, blood type, environment, etc.) could be used to infer some important patterns to yield a conclusion about the patient. Because data we see/examine/collect appears in various forms, there exists **several categories** that classify their types/forms.
What are the data types?	**Sequential Data** Sequence is a set of data items that are placed **in a specific order.** Usually, the order of the items provides **meaning/context** for the sequence. Sequential data is a type of data that is arranged in a specific order, where the position of each element plays a significant role in determining the purpose/meaning of the overall data. Changing the position of at least on element in sequential data implies changing the whole meaning of data itself. For example, we all know the importance of protein-contained diet due to its important functionality within our body. But do we know that protein, or to be precise, protein chains, consist of a sequential data? Amino acids consist of codons or nucleotide triplets (e.g. GCA, GAA, etc.). Each triplet has a

specific order to signal what amino acid to add to the protein chain.



Source: Pearson, General Biology

Other examples include DNA sequences, that determine a specific train, and WhatsApp text messages, where a combination of words signal the meaning of the message.



Source: Genome.gov, Human Genomic Variation

Data Matrix

From a linear algebra class, we all know that 2D matrix is a rectangle that consists of **n rows and m columns**.

Data matrix is a systematized **collection of records** (or can also be understood as an entity), which consists of a fixed set of attributes (or properties). The best way to visualize data matrix is to imagine a table in Microsoft Excel where each column represents an attribute/property and each row represents a separate record of something/someone.

For example,

Name	**Height(m)**	**Weight(kg)**	**Age**
Samantha	1.60	50	19
Alex	2.10	90	21
Leo	1.70	71	20

This is a data matrix that represents a collection of 3 records of people who disclosed their information; specifically, their height, weight, and age. *Please note that above table represents made-up people with made-up data records. Any resemblance is coincidental.*

One interesting property of data matrix is its flexibility to be adjusted in terms of order. If we shuffle one entire row OR one entire column all at once, the meaning of the data will NOT change. Say, if we swap height and weight in terms of the order as well as Samantha and Alex's orders, we will NOT change the meaning of the data.

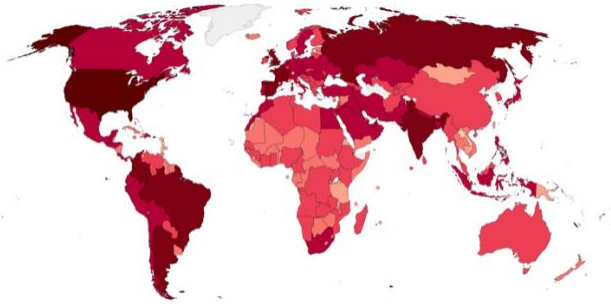
For reference,

Name	**Weight(kg)**	**Height(m)**	**Age**
Alex	2.10	90	21
Samantha	1.60	50	19
Leo	1.70	71	20

Spatial Data

Spatial data is a type of data where space/location/geographical attributes reveal the context of

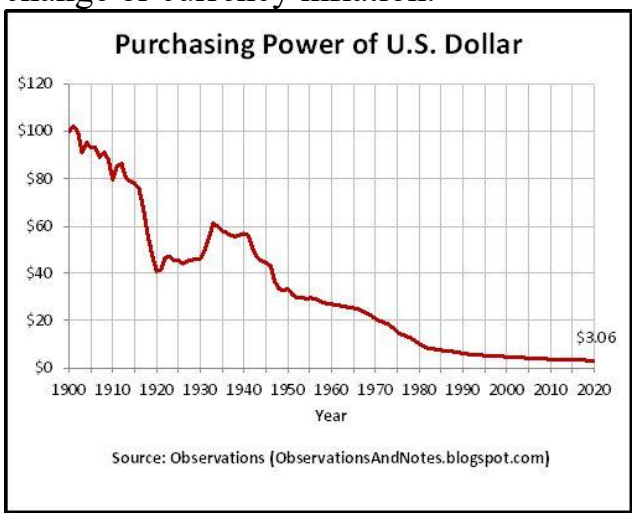
the data. The space can be understood in various forms, be it the world map or a location of items in the picture.



Source: Mapping the Coronavirus Outbreak Across the World

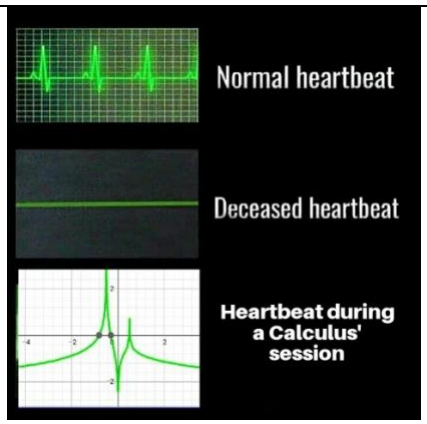
Temporal Data

Temporal Data is a type of data that is recorded in accordance with time. Time plays a significant role in understanding the data.
In finance, temporal data can be in the form of stock prices change or currency inflation.



Source: Observations, The Decrease of Purchasing Power of Dollar Since 1900

In healthcare, heartbeat is captured in terms of time frames. Thus, it could be classified as a temporal data which can analyze whether there are any anomalies with the way a patient's heart is functioning.



Source: A meme of some month ago, but always true from Reddit

---

Graph or Networks

Graph and networks are abstract mathematical concepts/structures that represent a relationship between objects/nodes. The relationship allows to set the context of the data itself.

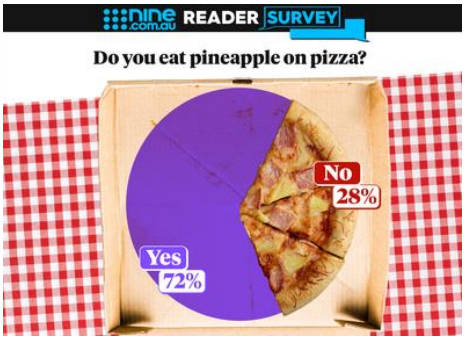
Network is particularly important in public health, where they track an interaction made between people during pandemic/outbreak of the disease.

Source: The Kardashian-Jenner Family Tree

Text

Text is something we see every day. Even this scribe is written using texts. In linguistic terms, text is an object used by humans to convey a specific information to its readers.

Text data can be understood as any data that contains text: one of the most popular ones in data analytics is survey.



Sources:

1. Twitter Thread
2. Pineapple Pizza Survey

Multi-modality Data

It is a type of data that integrated multiple elements of multiple data types (in other words, a mix of different data types).
Most complex data we know of nowadays is multi-modal. For example, our health record in the hospital contains data matrix in the forms of blood pressure, x-ray images, and prescriptions from the doctor.

Source: Medhealth Outlook

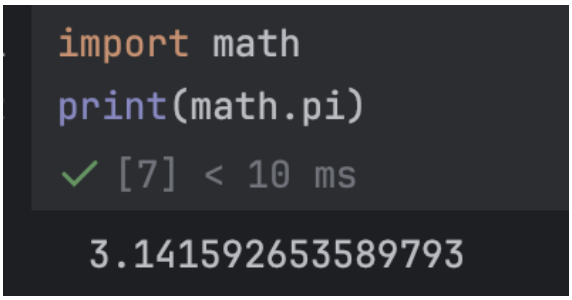
	Unknown Data Type Some data can be of ambiguous type since it has not been disclosed to us nor visualized in any forms; or the data it is trying to represent is hard to concisely structure in a data set. 
Summary	As we can see, there are various data types and lots of data examples. But data is very important in yielding decision, especially in a very rapidly evolving field, like healthcare.

Module III

Python 101: Introduction to the Programming Language

What is programming?	The activity or job of writing **computer programs** [Cambridge Dictionary]  But what is a computer program? It is a **sequence or set of instructions** in a programming language for a computer to execute. [Wikipedia] In other words, programming language is one of the tools to tell computer/machine on WHAT to do. In human-to-human interaction, we use human language to ask for a favor or give instructions. In human-to-computer interaction, we use computer/programming language to do the similar activity.

	Thus, it makes sense that each programming language has its own syntax rules, concepts, and grammar. In English, we can't say I BMEG3105 love because English follows SVO* sentence structure, or in other words because it just sounds weird… *SVO – Subject Verb Object structure Likewise, in Python, for example, you can't write ✕ "I love BMEG3105" print( ) It won't compile if you do so ! The correct way: ✓ print("I love BMEG3105")
What is Python. [Hint: it is not a snake 🐍]	Python is an interpreted high-level general-purpose programming language ⓘ **Interpreted*** - python gets compiled line-by-line instead of compiling the code all at once. **High-level*** - it uses syntax close to human language, instead of machine language which contains 1s and 0s **General-purpose*** - a programming language can be used to build/execute software platforms for various domains and purposes. Python, similar to other programming languages, uses several built-in tools that optimizes human work. For example, if we want to calculate a mean of an array of length 10'000, it will take us a while to input numbers into the calculator; yet, in

	Python, we can simply use built-in function called **numpy.mean**
What is numpy? What is package?	Package (or sometimes used interchangeably with Library) is a pre-formed collection of modules*/functions/statements. Modules are a set of functions we may include in our code. Package allows to collect those modules/functions to serve a certain purpose. ⓘ module is a unit file of code that contains functions, data, and relationship between data; the module is reusable. For example, we know number pi is precise and unchangeable anywhere. So instead of keep declaring int pi = 3.14…  We have a module called math that contains the constant variable pi that we may use of for our calculations. In case of NumPy (since it is considered a fundamental package of Python), it is used to provide users with the set of functions/modules utilized for **scientific computation** in multidimensional arrays and matrices, or simply a big chunk of data! Several functions we can make use of with the help of NumPy

Name	Purpose
Mean Numpy.mean( )	Calculate the mean value of the array of numbers
Standard Deviation Numpy.std( )	Calculate the standard deviation of the array of numbers
Median Numpy.median( )	Show the medial value in the array of numbers
Max Numpy.max( )	Show the max value in the array of numbers

Example:

```
import numpy as np
age = [18, 20, 34, 50, 8, 14, 32, 35, 20, 19]

#this is a comment in Python
#age is a list that represents our data sample
#now let's try to find mean, median, standard
# deviation, max values
print("This is a mean: ", np.mean(age))
print("This is a std: ", np.std(age))
print("This is a median: ", np.median(age))
print("This is a max: ", np.max(age))


✓ [5] < 10 ms
 This is a mean:  25.0
 This is a std:  11.832159566199232
 This is a median:  20.0
 This is a max:  50
```

ⓘ [ ] represents the LIST. If we use ( ) instead, we will be using TUPLE, another data structure that has slightly different properties. However, LIST [ ] is commonly used for the purpose of computation

Please note that to use numpy function, we first need to **import** it using code:

import numpy

	or import numpy as np → this changes the naming of numpy library into “np” Please also note that when calling a function, such as median, we first write the name of the package in which the specified function belongs to → numpy.mean ( ) or np.mean( ) because otherwise “mean” by itself is not a built-in module of Python* *unlike, for example, print( )

Reference Materials:

Yu Li. (September 2025). Data & Python & Sequence. The Chinese University of Hong Kong

[https://www.dropbox.com/scl/fi/rgvi35f0f0y24emcl1kl3/lec2-data-and-python.pdf?rlkey=q61bg4a9p4ejfo0mxjh7oviyy&e=1&dl=0]