TUTORIAL (Mid Week #1) Data Attributes and Visual Encoding

M1.1 Data Attributes

(To be done before Mid-week Session #1)

(1) Discuss for each data type listed below, if it has a **continuous** or **discrete** attribute:

(a) Number of patients in the ward
(b) The blood pressure of the patient
(c) The pulse rate of the patient (*state how you measured the pulse rate*)
(d) The emergency room waiting time displayed at the ER lobby (rounded to the nearest minute)
(e) The emergency room waiting time for each patient (rounded to the nearest minute)

(2) Discuss how the discrete or continuous nature of the data attributes can affect machine learning algorithms. You should consider the following perspectives:

(a) The benefits of each type of data attribute.
(b) The potential challenges of each type of data attribute.

(3) The passenger liner Titanic collided with an iceberg and sank on 5th April 1912. Out of the 2224 passengers, 1502 of them perished. The data file "*Titanic.csv*" shown in Table 1, contains information on 891 real passengers that were onboard the Titanic on that fateful night.

Which **NOIR** data scale best represents each data category (i.e. column) in Table 1?

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. O	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs.	female	38	1	0	PC 17599	71.28	C85	C
:	:	:	:	:	:	:	:	:	:	:	:
890	1	1	Behr, Mr. Karl	male	26	0	0	111369	30	C148	C
891	0	3	Dooley, Mr. Pa	male	32	0	0	370376	7.75		Q
Unique passenger identification no.	1=Survived 0=Perished	1=1st class 2=2nd class 3=3rd class	Name of passenger	Gender	Age in years	Sibling or spuse on board	Parents or children on board	Ticket number	Fare for each ticket	Cabin number	C=Cherbourg S=Southampton Q=Queenstown

Table 1 – Information in the *Titanic.csv* datafile and a summary of its interpretation. This datafile is downloadable at: https://github.com/rashida048/Datasets/blob/master/titanic_data.csv.

(4) Discuss how data features with **ordinal** and **nominal** data scale can be handled in machine learning algorithms that require numerical data (e.g. decision trees, linear regression, etc). Give the necessary Python code segment that can conditions these types of data features for the *Titanic.csv* dataset. (*Hint*: consider the use of techniques like One-Hot Encoding).

(5) How many data dimensions have been simultaneously visualised in the clustered bar (column) chart shown in Figure 1? Draw a possible table (partial table will do) for this dataset and state which **NOIR** data scale best represent each data dimension in your table.

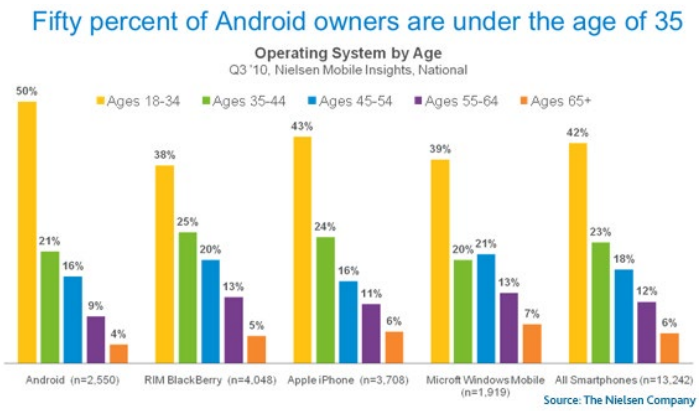

Figure 1 – Mobile OS by Age (From https://www.nielsen.com/us/en/insights/article/2010/mobile-snapshot-smartphones-now-28-of-u-s-cellphone-market/)

M1.2 Visual Encoding

(1) In the context of Jacque Bertin's visual marks and visual variables, describe what visual marks are employed in the visualisation shown in Figure 2 and what visual variable (perceptual channels) are employed to encode the different telecommunication traffic flow information. Discuss the potential limitations of the perceptual channels employed.

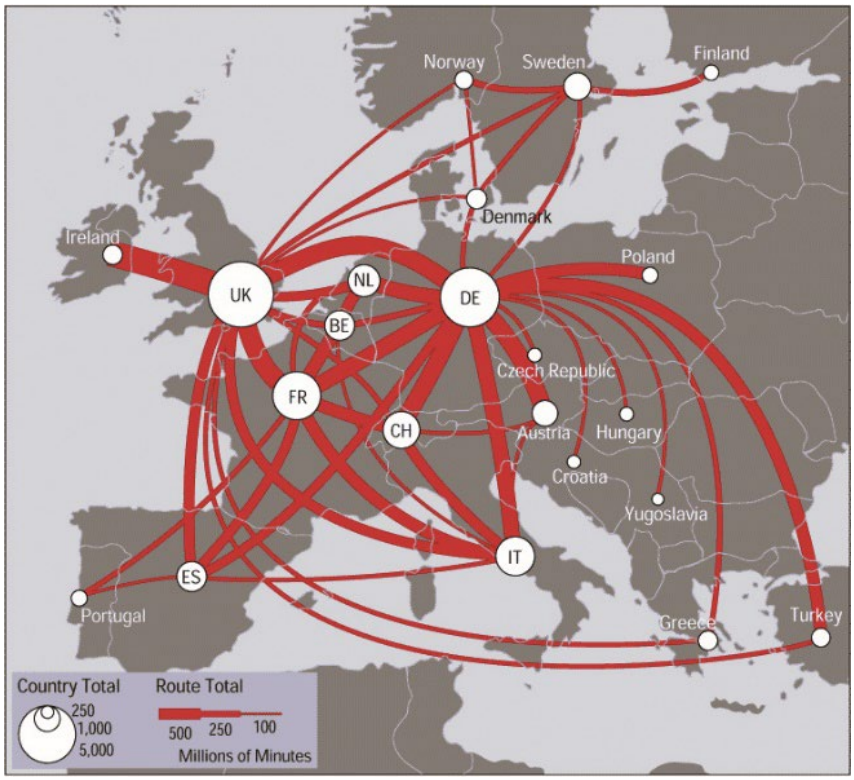

Figure 2 – Telecommunication Traffic Flow Map, © 2000 - TeleGeography, Inc.
Image taken from https://mappa.mundi.net/maps/maps_014/ . For more details on the visualisation, check out the link.

(2) Based on the suggested visual mark and your own visual encoding design, sketch a possible chart that will allow the effective visualisation of each of data tables shown in Figures 3(a) to (c). For each case, consider carefully the implication of the dataset size and the likely numeric range and scale of measure of the listed data dimensions, and how they may impact the visual effectiveness of your proposed chart. Discuss a possible way you can address each of this issue.

Line

Living Members from Class of '45

	Names	Gender	Age
1			
2			
3			
4			
5			
6			

(a)

Point

Orientation T-shirts - Comparing Halls

	No.	Gender	Weight	Height	Hall (A,B,C,D)	T-shirt Size
1						
2						
3						
:						
:						
700						

(b)

Area

COVID-19 Infection in 2020

	Dates	Total Cases	Warded	ICU
1				
2				
3				
:				
:				
365				

(c)

Figure 3 – Data table examples and the suggested visual mark that can use to design the chart.
Note the characteristics of the data description and the numeric values indicated in the table when coming up with your chart design.