Jan 8

课上PPT（更加详细一些）

M1.1 Data Attributes

Handling Ordinal and Nominal Data

- Machine learning algorithms generally require **numerical input**, but nominal and sometimes ordinal (e.g. hot, warm, cold) data are inherently categorical.
- Most algorithms (e.g., decision trees, linear regression, SVMs) **cannot process categorical variables** in their raw form.
- We **need to convert** these categorical variables **into a numerical format** that the algorithm can interpret and learn from.

Nominal	Nominal	Ordinal	Nominal	Nominal	Ratio	Ratio	Ratio	Nominal	Ratio	Nominal	Nominal
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

M1.1 Data Attributes

Handling Ordinal and Nominal Data

Nominal **What is One-Hot Encoding?**

a) **One-hot encoding**: A widely used techniques for encoding nominal data. It converts categorical values into **binary vectors** where each value in the category is represented as a binary feature (1 or 0).

b) **Dummy variables** are used to implement one-hot encoding (OHE).

Nominal

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

Goh Wooi Boon (Assoc Prof)

One-hot encoding

M1.1 Data Attributes

Handling Ordinal and Nominal Data

Nominal What is One-Hot Encoding?

a) **One-hot encoding**: A widely used techniques for encoding nominal data. It converts categorical values into **binary vectors** where each value in the category is represented as a binary feature (1 or 0).

b) **Dummy variables** are used to implement one-hot encoding (OHE).

Multicollinearity: When 2 or more features are highly correlated with each other. OHE's inclusion of all dummy variables creates perfect correlations, leading to multicollinearity.

Embarked_S	Embarked_C	Embarked_Q
1	0	0
0	1	0
:	:	:
0	1	0
0	0	1

Increased Dimensionality: The number of features (columns) increases with the number of categories, leading to sparse data and increased computational complexity.

Nominal

Embarked
S
C
:
C
Q

C=Cherbourg
S=Southampton
Q=Queenstown

Goh Wooi Boon (Assoc Prof)

M1.1 Data Attributes

Handling Ordinal and Nominal Data

Nominal How to resolve the multicollinearity problem arising from the dummy variable trap?

a) **Drop first**: Drop one dummy variable to avoid the multicollinearity issue. This can be done by setting `drop_first=True` in the pandas `get_dummies` function.

```
df = pd.get_dummies(df, columns=['Embarked'], drop_first=True)
```

Multicollinearity: When 2 or more features are highly correlated with each other. OHE's inclusion of all dummy variables creates perfect correlations, leading to multicollinearity.

Embarked_S	Embarked_C	Embarked_Q
1	0	0
0	1	0
:	:	:
0	1	0
0	0	1

One-hot encoding

Nominal

Embarked
S
C
:
C
Q

C=Cherbourg

Goh Wooi Boon (Assoc Prof)

M1.1 Data Attributes

Handling Ordinal and Nominal Data

Ordinal **What is Ordinal Encoding?**

a) **Ordinal encoding**: Encodes the categorical data into numeric forms understood by the algorithms, while **retaining the inherent order** in the ordinal data feature, which one-shot encoding will destroy.

Ordinal

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3						A/5 21171	7.25		S
2	1	1						PC 17599	71.28	C85	C
:	:	:						:	:	:	:
890	1	1						111369	30	C148	C
891	0	3						370376	7.75		Q
Unique passenger identification no.	1=Survived 0=Perished	1=1st class 2=2nd class 3=3rd class	passenger		years	spuse on board	children on board	Ticket number	Fare for each ticket	Cabin number	C=Cherbourg S=Southampton Q=Queenstown

No need for Ordinal Encoding:
Some ordinal features are already in numeric form, like the "**Pclass**" feature. There is no need for further encoding.

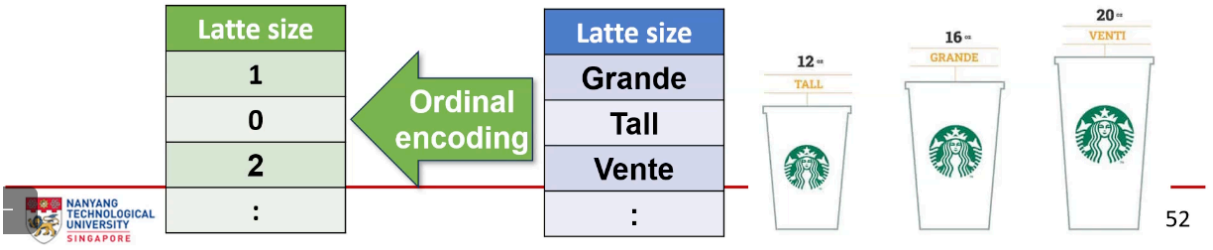
M1.1 Data Attributes

Handling Ordinal and Nominal Data

Ordinal **What is Ordinal Encoding?**

a) **Ordinal encoding**: Encodes the categorical data into numeric forms understood by the algorithms, while **retaining the inherent order** in the ordinal data feature, which one-shot encoding will destroy.

b) **Order mapping**: Ordinal encoding maps each unique category to an **increasing integer value** that matches the implicit increasing order in the categorical data.



M1.1 Data Attributes

Handling Ordinal and Nominal Data

Ordinal How to implement Ordinal Encoding?

a) **Ordinal encoder**: The scikit-learn's `OrdinalEncoder` is recommended for ML pipelines.

```
from sklearn.preprocessing import OrdinalEncoder

order = ['S', 'C', 'Q']
encoder = OrdinalEncoder(categories=[order])

# Create new "Eorder" column with ordinal data for order of port of embarkation
df['Eorder'] = encoder.fit_transform(df[['Embarked']])
```

Convert **Embarked** to port of call order

Ordinal

Embarked
S
C
:
C
Q

C=Cherbourg
S=Southampton
Q=Queenstown

M1.1 Data Attributes

Handling Ordinal and Nominal Data

Ordinal How to implement Ordinal Encoding?

a) **Ordinal encoder**: The scikit-learn's `OrdinalEncoder` is recommended for ML pipelines.

```
from sklearn.preprocessing import OrdinalEncoder

order = ['S', 'C', 'Q']
encoder = OrdinalEncoder(categorie

# Create new "Eorder" column with ordinal data for
df['Eorder'] = encoder.fit_transfo
```

Convert **Embarked** to port of call order

Ordinal

Eorder
0
1
:
1
2

S=0
C=1
Q=2

Ordinal encoding

Embarked
S
C
:
C
Q

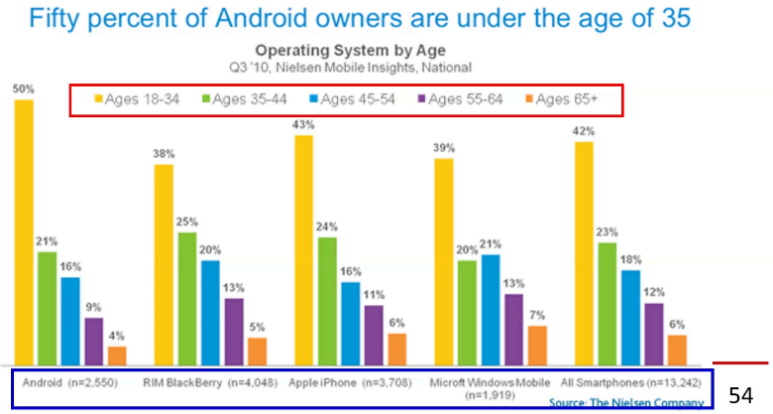
C=Cherbourg
S=Southampton
Q=Queenstown

M1.1 Data Attributes

(5) Data Dimensionality

- What data dimensions are simultaneously visualised?
- Name them and their respective NOIR scale?

1. Age group (e.g. 18-34, 35-44, etc) – **Ordinal scale**
2. Types of Mobile OS (e.g. Android, Blackberry, etc) – **Nominal scale**



1.1 Data Attributes

(5) Data Dimensionality

Draw a possible table for this dataset.

1. Age group (e.g. 18-34, 35-44, etc) – **Ordinal scale**
2. Types of Mobile OS (e.g. Android, Blackberry, etc) – **Nominal scale**
3. Population of users – **Ratio scale**

Mobile OS	Age Group	Population
Android	18 – 34	1275
Blackberry	18 – 34	1538
:	:	:
Android	35 – 44	536
Blackberry	35 – 44	1012
:	:	:

M1.2 Visual Encoding - Review

a) Visual Marks

- What are visual marks for item?

  Visual marks are basic graphic elements that can be used to represent information.

- List the three visual marks proposed by Jacque Bertin and describe the main differences in their basic property:
  - **The three visual marks for items listed by Bertin are points, lines and areas.**

  - **They differ in their spatial dimensions. Points are 0D, lines are 1D and areas are 2D.**

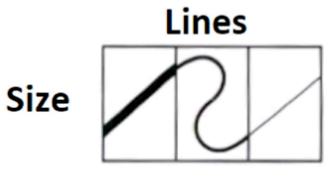
M1.2 Visual Encoding - Review

b) Visual Variables (Perceptual Channels)

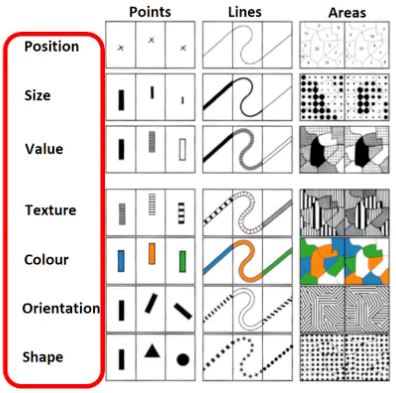
- Name the 7 perceptual channels (visual variables) proposed by Bertin and describe how they would alter the appearance of a line mark:

  Bertin listed 7 different visual variables (perceptual channels):

  - **Position**
  - **Size**
  - **Value**
  - **Texture**
  - **Colour**
  - **Orientation**
  - **Shape**



- **E.g. - the way the size perceptual channel will change the line mark is to make it thicker.**



M1.2 Visual Encoding

(1) Limitation of Perceptual Channels

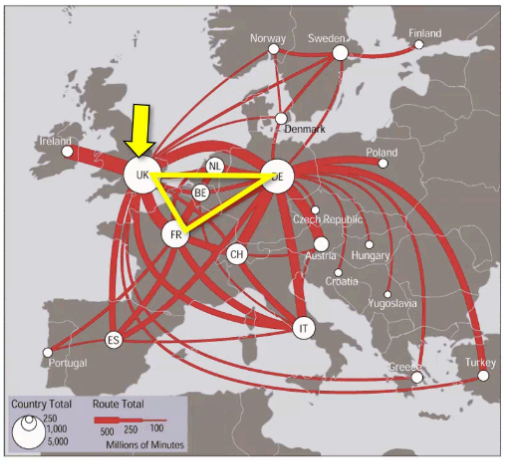
What other visual mark is employed here?

- **The point mark – show as white circular symbols, located on the capital city.**

What perceptual channel of this visual mark has been used to encode information?

- **The size of the point** (circle) **encodes the country's total annual outgoing traffic to all other countries.**

From the map, it can be seen that the United Kingdom (**UK**), Germany (**DE**), and France (**FR**) dominate traffic intra-European telecommunication flows, forming a powerful triangle at the heart of the European continent.



59

M1.2 Visual Encoding

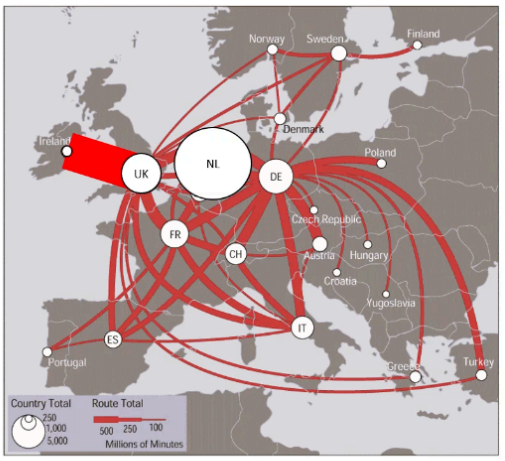
(1) Limitation of Perceptual Channels

Limitation of the perceptual channel used?

- **Changing line width only works for a fairly small number of steps.**
- **Width that are too thick will be perceived as a polygon area rather than a line mark.**
- Visual shows an effective use of linewidth, which can work well to show about **3 or 4 different values** for a data attribute.

What about the **point**'s perceptual channel?

- **Potential occlusion problem.**



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

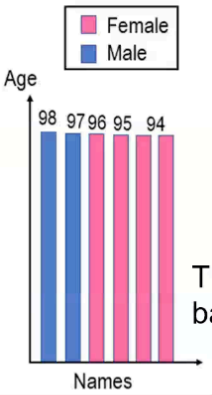
60

M1.2 Visual Encoding

(2) Designing the Chart

Draw chart (using **line** marker) to encode this table and what problems may arise?

- **We could use a bar chart.**
- **Length encodes the age and colour encodes the gender (male or female).**
- **Problem:** **The ages will be very large (seniors) and the difference in ages will be difficult to distinguish due to the need to maintain a zero baseline.**
- **Solution:** **Annotate the actual age for each bar**



Living Members from Class of '45

	Names	Gender	Age
1			
2			
3			
4			
5			
6			

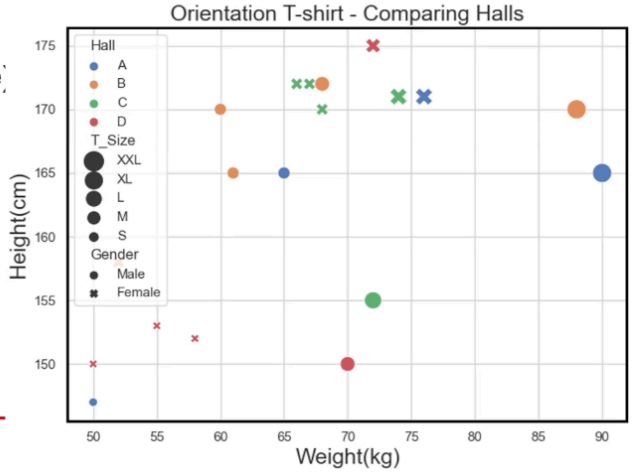
The likely shape of the bar chart given their ages

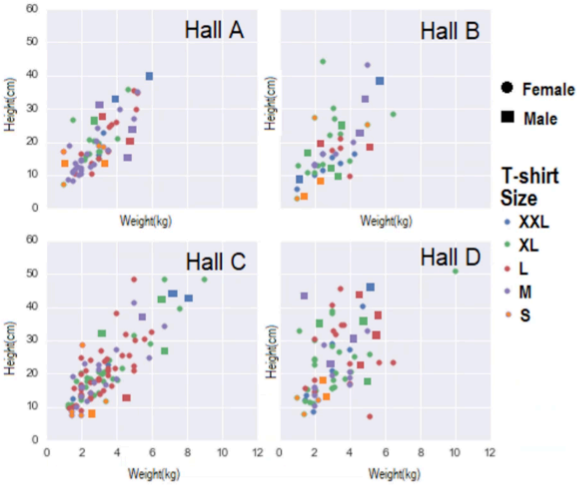
M1.2 Visual Encoding

(2) Designing the Chart

Draw chart (using **point** marker) to encode this table and what problems may arise?

- We could use a **scatter plot.**
- **Positions** encode height & weight (Quantitative)
  Colour encode **4 halls** (Nominal),
  Shape encode **gender** (Nominal),
  Size of shape encode **t-shirt sizes** (Ordinal).

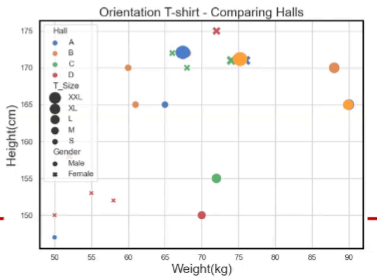




ıble and what problems may arise?

Orientation T-shirts - Comparing Halls

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



- **Solution:** Use trellis plots to sub-divide space to enable comparison across multiple plots (less points per plot).

M1.2 Visual Encoding

(2) Designing the Chart

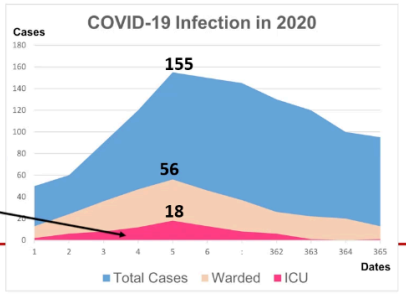
Draw chart (using **area** marker) to encode this table and what problems may arise?

COVID-19 Infection in 2020

	Dates	Total Cases	Warded	ICU
1				
2				
3				
:				
:				
365				

- We could use a **stacked area chart.**
- **Coloured areas** encode **the different values.**
- **Problem:** Hard to tell the exact number of cases, especially for the **Warded** category sandwiched between the two plots.
- **Solution:** Annotate with values, especially the peak values

Small ICU numbers makes it hard to see ICU category. How to address this?



Solution

M1.2 Visual Encoding

(2) Designing the Chart

Draw chart (using **area** marker) to encode this table and what problems may arise?

- We could use a **stacked area chart.**
- **Coloured areas** encode **the different va**
- **Problem:** Hard to tell the exact number of cases, especially for the **Warded** category sandwiched between the two plots.
- **Solution:** Annotate with values, especially the peak values

Small ICU numbers makes it hard to see ICU category. How to address this?

COVID-19 Infection in 2020

Dates	Total Cases	Warded	ICU
1