Data Visualisation

Associate Professor Goh Wooi Boon

School of Computer Science and Engineering
Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

1

1

Chapter 2 – Data Attributes & Wrangling

Contents

- Basic Data Attributes
- Other Data Attributes
- Data Wrangling

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE   © A/P Goh Wooi Boon (SCSE/NTU)

2

2

Chapter 2.1 – Basic Data Attributes

Why the Need to Understand Data Attributes?

- One of the key part of the data visualisation design process concerns the understanding of different attributes of the data.

- The data type or its attribute influences:
  - The types of plot or chart you should use.
  - The visual attributes (e.g. colour, layout, etc) you should employ.
  - The operations or computations you can perform on the data (e.g. add, multiply, median, mean).
  - The types of data analysis you can undertake.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

3

3

Finding Meaning in Data

Data Models versus Conceptual Models

- **Data** models are **low-level description** of the data.
  - Examples: integers with its + and × operators.

- **Conceptual** models are **mental constructions** of the domain. It provides **semantic** (meaning) to the data model and supports **reasoning** about the data.

- Examples of data models vs. conceptual models:
  - 1D value (float) vs. temperature
  - 2D pair (float) vs. geographical location (longitude & latitude)
  - 3D vector (float) vs. point in space

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

4

4

Basic Data Attributes

Quantitative vs Qualitative

- A basic attribute of a data is whether it is qualitative or quantitative.

- Qualitative data is **descriptive** and **conceptual**. Such data are categorised based on observable traits and characteristics.

  - Examples: Gender (Male or Female), Color (**Red**, **Green**, **Blue**), Countries (Singapore, Malaysia, etc).

- Quantitative data can be **counted** or **measured** and is expressed as **numbers**.

  - Examples: Height & weight (in cm & kg), Light intensity (in Lumens), Population (in Singapore).

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

5

5

Basic Data Attributes

Discrete versus Continuous

- **Discrete** data only takes on **certain** pre-determined **values** and is usually obtained by **counting**.
  - Examples: Number of students in the class, goals scored in a match, annual profits of a company.

- Discrete data can be represented using tally charts or pie charts.

- **Continuous** data can take on **any value** and is usually obtained by **measuring**.
  - Examples: Height, weight, temperature, sound intensity (dB).

- Continuous data may change over time (e.g. my daily weight in May) and is best visualised using a **line graph** that can shows continuous data changes over time.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

6

6

Level of Measurement

NOIR Scale of Measure

- In 1946, the American psychologist Stanley Smith Stevens introduced a theory of **level of measurements**.

- He claimed all measurement in science was conducted using four different types of scales, namely **nominal** (N), **ordinal** (O), **interval** (I) and **ratio** (R).

- The proposed **scale of measure** provides an interesting way to categorised different types of data variables in ways that can help us choose appropriate statistical tests, visualisation techniques and data analysis methods.
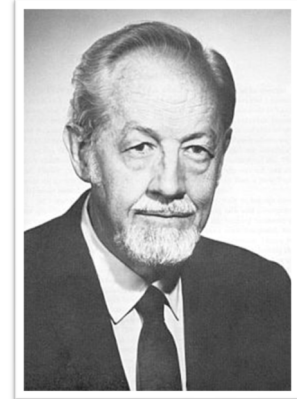
Stanley Stevens (4 Nov 1906 – 18 Jan 1973), founder of Harvard's Psycho-Acoustic Lab. Image from Neurotree.org

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

7

7

Level of Measurement

Nominal

- Nominal scale is used to **label qualitative** data variables, as such it has **no quantitative** value.
  - Examples: Gender (Male, Female), Hair colour (Brown, Black, Blonde, Red, White), Ice Cream Flavours (Vanilla, Chocolate, etc).

- Nominal scale is **mutually exclusive** (no overlaps) and carries **no numerical significant** or **ranked order**.

- Valid operations on such data include ($=$ and $\neq$).

- Since nominal data can be counted, typical visualisation techniques include bar and pie charts.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

8

8
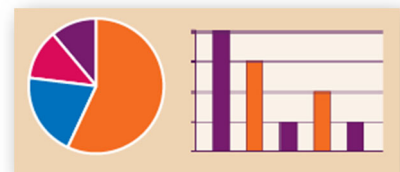
Level of Measurement

Ordinal

- In ordinal scale, the **order** of the data values is **important** and carries significance.
- However, the **differences** between each item have **no significance** or measure.
  - Examples: Grades (**A**, **B**, **C**, **D**, **F**), Star ratings (★ , ★★ , ★★★ , ★★★★ , ★★★★★).

- Ordinal scale typically measures **non-numeric** concepts like subjective rating (e.g. rating your lecturer or satisfaction level).

☹	☹	😐	🙂	😊
Very Unsatisfied	Unsatisfied	Neutral	Satisfied	Very Satisfied

5-point Likert Scale based on Emoticons

- Valid operations on such data include (= , ≠, < and >).

- Visualisation techniques such as bar and pie charts can also be used for ordinal scale data.

9

9
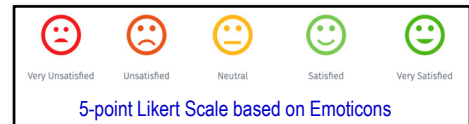
Level of Measurement

Interval

- Interval scale is a **numeric** scale (quantitative) in which the order and measurable **differences between values** are known.
  - Examples: Temperature (in °C or °F), Map locations (Spore, LAT 1.29, LONG 103.85), Dates (9-11-2001).

- Valid operations on such data include (= , ≠, < , > , + and −).
- With the ability to add values, we can now have **mean** and **standard deviations**.

- Interval scale does not have a "true zero" or absolute zero. Since interval data has **no true zero**, such data cannot be multiplied or divided. (e.g. 0°C is not "no temperature" but just another temperature).

- A common visualisation technique for the interval scale is the **histogram**.

10

10

Level of Measurement

Ratio

- The ratio scale is **numeric** and it informs **order** and **exact values** between units.

- This scale has an **absolute zero**. It therefore has properties like **proportion**.

  - Examples: Height and weight (in cm and kg), the population of a city and temperature (in Kelvin, K),

- Valid operations on such data include ($=$, $\neq$, $<$, $>$, $+$, $-$, $\times$ and $\div$).

- The ratio scale supports the use of a wide range of **inferential statistics**, such the **coefficient of variation** (CV=SD/Mean) that measures dispersion about the mean.

- Since proportions are valid in ratio scale, visualisation techniques like **bubble chart** can provide meaningful comparison between ratio scale variables.

Browser market share

11
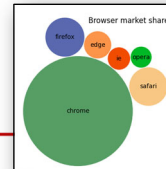
Review
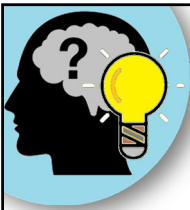
The NOIR Scale of Measure

- Which scale represents each of the following data attributes:

1) The different types of coffee beans -

2) Year coffee bean was harvested -

3) Weight of coffee beans -

4) Size of my latte {tall (12 ounces), grande (16), venti (20)} -

| Nominal | Ordinal | Interval | Ratio |

12

Think and Apply

Identifying the NOIR Scale of Measure

- The Excel table below shows results for a recent 30 minute online test. What scale of measure does each of the data attributes have.
- Which attribute is **discrete**, and which is **continuous** in nature?

	A	B	C	D	E	F	G
1	Student ID	Gender	Year of Birth	Attendance (no. of weeks)	Score (%)	Grade	Completion Time (sec)
2	1	M	2002	12	85.5	A+	1600.31
3	2	F	2004	2	20	F	600.13
4	3	F	2003	13	55.5	C+	1800.00
5	4	M	1999	10	73.5	B+	1800.00
6	5	M	2000	11	65	B	1500.23
7	6	F	2004	8	70	B+	1700.00
8	7	F	2005	13	90.5	A+	900.34

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

13

13

From Data Model to the NOIR Scale

An Example

- **Data model**: -42.55, 0, 3.98, 100

- **Conceptual model**: Temperature (°C)

- **Data Scales**:
  - Temperature values **(Interval)**
  - Solid vs. Liquid **(Nominal)** – derived attribute
  - Cold, Warm, Hot **(Ordinal)** – derived attribute

Most dense at 3.98°C
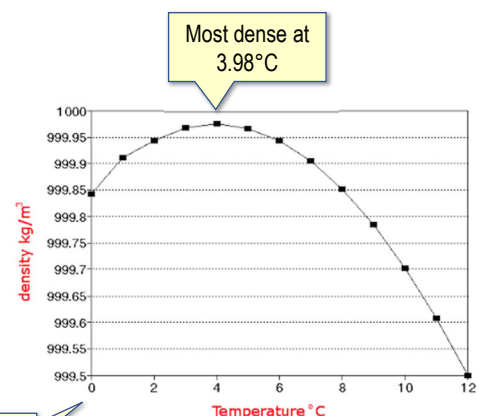
Turns Solid at 0°C

The Amazing Property of Water

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

14

14

Summary

Basic Data Attributes

- **Qualitative** data is described by observed categories and is mostly non-numeric. **Quantitative** data is expressed in numeric values and can be counted or measured.

- Countable **discrete** data only takes on pre-determine values, while measurable **continuous** data can take on any value.

- Steven's **level of measurement** or scale of measure is a widely adopted way of looking at the nature of the scale data is measured with.

  - **Nominal** scale used labelled qualitative measures with no specific order.
  - **Ordinal** scale is ordered but differences between measures have no significance.
  - **Interval** scale is numeric with measurable differences but has no absolute zero.
  - **Ratio** scale measures exact numeric values from an absolute zero reference.

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

15

15

Chapter 2.2 – Other Data Attributes

Contents

- Data Dimensionality

- Hierarchy

- Temporal

- Spatial
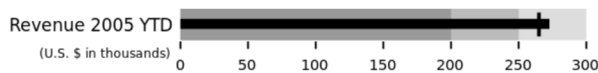
Data Dimensionality

Relational Data Model

- Structured data typically consist of **data observations** that are represented by rows and **data attributes** that are represented by columns.

- A relational data model represents data as a **table**.

- Each column can be called a **dimension** of the dataset and these columns or attributes define the relation.

- Each row in the relation is known as **tuple**. The relation below contains 3 tuples.

ID No.	Name	Gender	Age	Score
1	Long Kang Kin	M	17	50
2	See Peh Loh	F	21	236
3	Tentu Tepat	M	18	300

Data Dimensionality

0-Dimension

- This type of data consists of a **single value**.

- Such data can be visualised using **gauges** like thermometer graphs, speedometer dials and bullet graphs.

Bullet graphs [1]

Gauges, speedometer and thermometer graphs [2]

[1] Image Source – Bullet Graph - https://en.wikipedia.org/wiki/Bullet_graph
[2] Image Source – Gauges - https://apandre.wordpress.com/dataviews/dimensionality/

18

18

Data Dimensionality

1-Dimension

- This type of dataset consists of only a **single attribute** of observed data (1 column).

- For example, this could be a single day record of all NTU students entering canteen A and whether they wore glasses (**Yes**) or not (**No**).

- Such 1-D datasets can be visualised using a **pie chart**, for example.
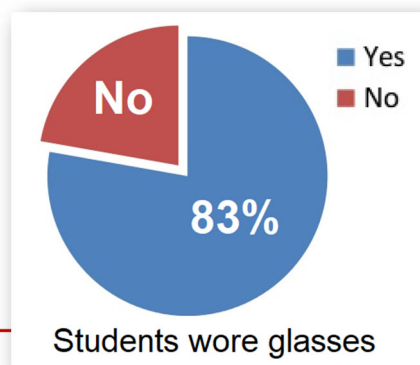
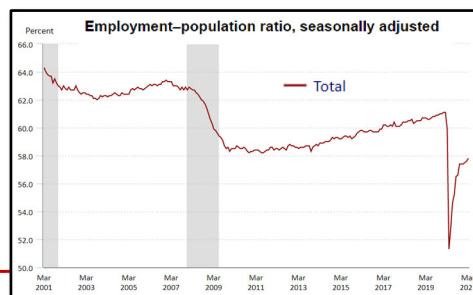| Wore Glasses |
|---|
| Yes |
| No |
| Yes |
| Yes |
| : |
| : |
| No |
| Yes |

19

19

Data Dimensionality

2-Dimension

- This type of dataset consists of data with **two different** but **related attributes**.
- The appropriate ways to visualise such datasets will depend on the nature of the **scales of measure** of these two attributes. For example, if one scale is **ratio** and the other **ordinal**, then a **line chart** can be used. Those with **ratio** and **nominal** scales, a **pie chart** would be more appropriate

[3] Image Source – Line Chart
https://www.bls.gov/charts/employment-situation/employment-population-ratio.htm

[4] Image Source – Pie Chart -
https://www.exportgenius.in/blog/singapore-imports-crude-petroleum-oil-singapore-import-data-483.php
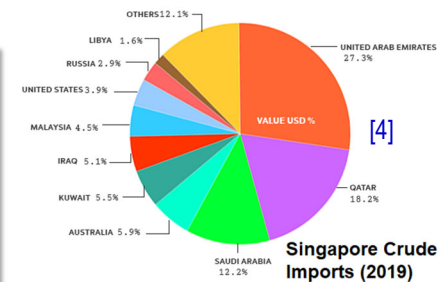
[3]

20

20
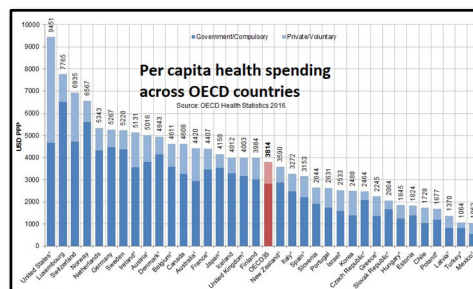
Data Dimensionality

3-Dimension

- As the data dimension increases to three, **other visual attributes** (e.g. colour, size, shape) may be needed to allow all dimensions to be visualised together.
- The **scale of measure** and nature of the **relationship** between dimensions will affect the choice of visualisation techniques.

[5] Stacked-column Chart from
https://search.oecd.org/fr/els/systemes-sante/graph-of-the-month.htm

[6] 2D Scatter Plot from
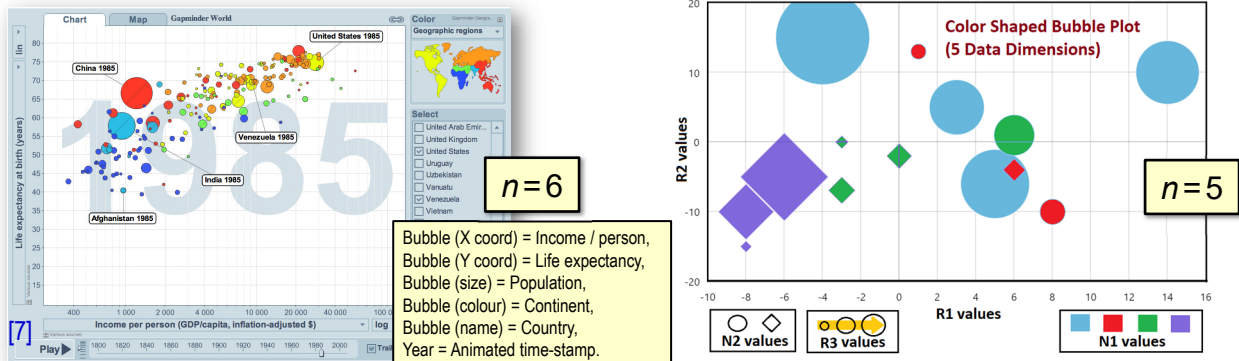https://huebler.blogspot.com/2008/10/child-labor.html
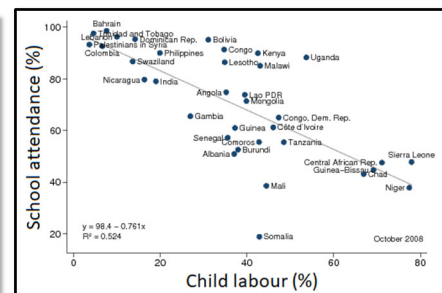
[5]

[6]

21

21

Data Dimensionality

n-Dimension

- Visualising *n* data dimensions simultaneously will require using a combination of visual attributes that matches the respective scale measure of each data dimension (e.g. colour & shape for nominal scale. Size, height, position for interval or ratio scale).


$n=6$

Bubble (X coord) = Income / person,
Bubble (Y coord) = Life expectancy,
Bubble (size) = Population,
Bubble (colour) = Continent,
Bubble (name) = Country,
Year = Animated time-stamp.

$n=5$

[7]

[7] Image Source – Motion Chart - http://www.gapminder.org/world/   [8] Video taken from BBC News at https://www.youtube.com/watch?v=jbkSRLYSojo

22

22

Think and Apply

Identifying the Dimensions

- How many data dimensions is simultaneously visualised in the Clustered Bar chart shown below?

- What would the table for this dataset look like?
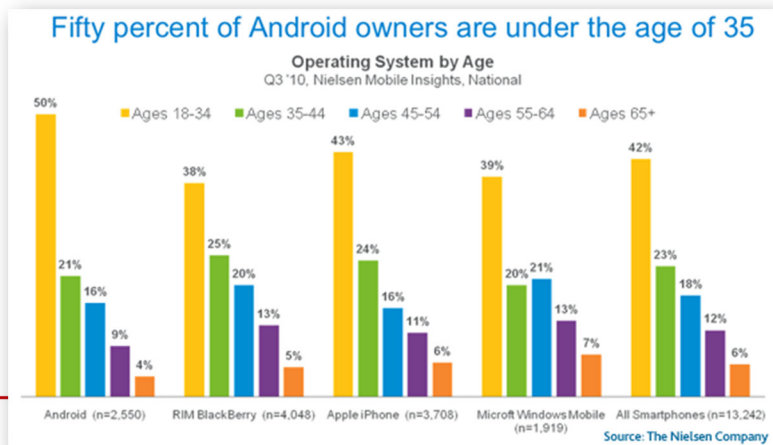
23

23

Hierarchy

Hierarchical Data

- Hierarchical data is a dataset in which each item of data defines a node in the **tree**, and each node may have a collection of other nodes as **child nodes**.

- The relationship between parent nodes and child nodes forms a **tree network** and the most basic method to visualise simple data hierarchy is a **Tree diagram**.

- Other visualisation methods that show how hierarchical data are ranked and ordered together in an organisation or system include the **Treemap**, **Sunburst diagram**, **Circle packing**, etc[9].
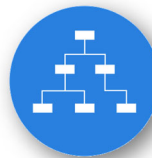
| Tree Diagram | Treemap | Sunburst Diagram | Circle Packing |

[9] For examples of hierarchical visualisations see - https://insightwhale.medium.com/how-to-show-hierarchy-with-data-visualization-526fb45ee4c2

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

24

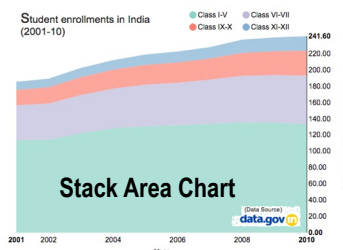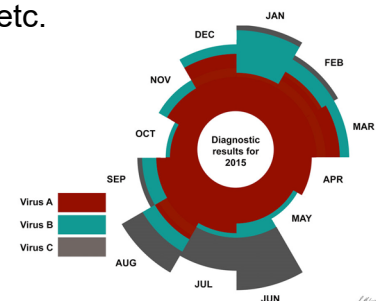# Temporal

Temporal Data

- Temporal datasets have data that represents a **state in time**. Its time dimension has measures that are usually **uniformly spaced** (from milliseconds to years).

- Visualisation methods for temporal data include line & bar charts, stacked area chart, scatter plot, polar area diagram (cyclical time series), etc.
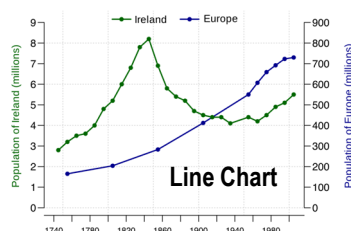
Line Chart

Stack Area Chart

Polar Area Diagrams

[10] For examples of hierarchical visualisations see - https://humansofdata.atlan.com/2016/11/visualizing-time-series-data/

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

25

Spatial

Geospatial Data

- Geospatial data has attributes related to a specific **location** on the **Earth's surface**.

- Geospatial data combines **location** information (e.g. coordinates on the earth), **attribute** information (e.g. attributes of the object, event, or phenomena concerned), and sometimes **temporal** information (the time or life span at which the location and attributes exist).

- Visualisation methods for geospatial data include Choropleth Map, Dot Density Map, Bubble Map, Heat Map, etc.

Choropleth Map

Dot Density Map

Heat Map

26

26
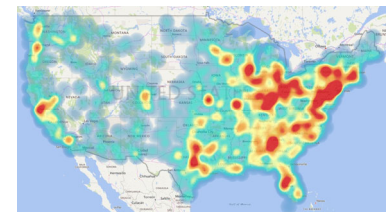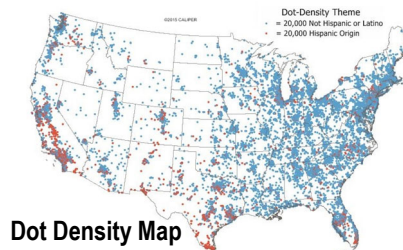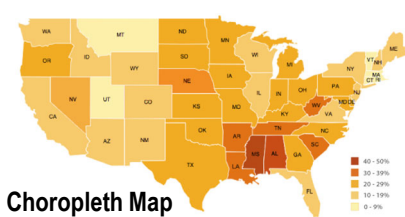
Other Data Attributes

- The tools and methods used to visualise data effectively depends on the **number of dimensions** being visually compared at the same time. And the dimension's **scale measure** influences the choice of visual attributes used in the chart.

- The nature of the **main characteristic** of the dataset will determine the type of visualisation method to employ.

  - **Hierarchical** data has significance in its organisational structures.
  - **Temporal** data has significance in the way it changes with time.
  - **Spatial** data shows its significance when associated with geographical maps.

27

27

References for Data Attributes

[1] Image Source – Bullet Graph - https://en.wikipedia.org/wiki/Bullet_graph

[2] Image Source – Gauges - https://apandre.wordpress.com/dataviews/dimensionality/

[3] Image Source – Line Chart https://www.bls.gov/charts/employment-situation/employment-population-ratio.htm

[4] Image Source – Pie Chart - https://www.exportgenius.in/blog/singapore-imports-crude-petroleum-oil-singapore-import-data-483.php

[5] Stacked-column Chart from https://search.oecd.org/fr/els/systemes-sante/graph-of-the-month.htm

[6] 2D Scatter Plot from   https://huebler.blogspot.com/2008/10/child-labor.html

[7] Image Source – Motion Chart - http://www.gapminder.org/world/

[8] Video taken from BBC News at https://www.youtube.com/watch?v=jbkSRLYSojo

[9] For examples of hierarchical visualisations see - https://insightwhale.medium.com/how-to-show-hierarchy-with-data-visualization-526fb45ee4c2

[10] For examples of hierarchical visualisations see - https://humansofdata.atlan.com/2016/11/visualizing-time-series-data/

[11] Gif image taken from - https://www.birdseyeviewgis.com/blog/2020/8/14/creating-a-covid-19-temporal-animation-with-qgis

Note: All online articles were accessed between May to June 2021

28