Course Assessment

What are the Graded Components?

To be done Individually

	Weightage
• **Project (Video presentation)**	35%
• **Critique the presentations**	15%
• **Mid-semester quiz**	25%
• **Final quiz**	25%
Total	100%

Assignments

lockdownbrowser
(bring notebook)

- Assignment instructions and report templates can be found in the **Assignment** folder in NTULearn.

Project

Putting Skills and Knowledge into Creative Practice

- Use a case study example (dataset) of your choice.
- Investigate, explore, design and develop appropriate data visualisation to share about the hidden insights within the data.

Components

- **Presentation** - short video presentation (≈ 5 mins max) using appropriate data visualisation techniques to deliver visual narrative of data insights.
- **Summary Report** – to highlight project objectives, novel contributions, technical challenges and contributions (2-page limit max, inclusive of references).

Submission Deadline — Weeks #11

Brief Course Outline

What Will You Learn?

- Data Attributes and Wrangling
- Visual Encoding
- Basic Plots and Charts
- Visual Perception
- Designing the Visualisation
- Interactive Visualisation

Mid-semester Quiz Coverage — Covered by A/P Goh Wooi Boon

- Exploratory Data Analysis
- Geospatial Visualisation
- Visualisation of Abstract Data
- Visualisation of Scientific Data

Final Quiz Coverage — Covered by Ast/P Wang Yong

Critique The Presentations

Having A Critical Eye

- We learn by looking at how other people have done good and effective visualisation of their dataset.
- Being critical and **constructive** show our mastery of the subject.
- Each student will be given **two random presentations** to critique. A written report* (with both +ve and –ve criticisms) is to be submitted for each presentation. **One** will be **randomly** selected for **grading**.
- The anonymous critiques will be made available to the respective presenters for you to **learn from your peers**.

* (4-page limit, inclusive of references and figures).

Submission Deadline — Weeks #13

Mid-semester Quiz

Gauge what you have learnt in the 1st half

- A mid-semester quiz will be given after the recess week to evaluate how well you have grasped the concepts taught in the **first half** of the semester (excluding Interactive Visualisation).
- Quiz will be done in week #9.
- Duration will be 1.0 hours.
- Format of quiz to be communicated at a later date.
- **Checkpoint** - Your general grade for the mid-semester quiz will be revealed to you to help you gauge your understanding so far.

Quiz Schedule — Weeks #9

Final Quiz

Wrapping up all that you have learnt

- A final quiz will be given at the end of the course to evaluate how well you have understood the concepts taught in the **2nd half** of the semester (including Interactive Visualisation).
- Quiz will be done in week #13.
- Duration will be 1 hour.
- Format of the quiz will be communicated to you at a later date.

Last Modified (GWB)
Ver1.3- 11Jul24

	12 to 16 Aug		19 to 23 Aug		26 Aug to 30 Aug		02 to 06 Sep		09 to 13 Sep		16 to 20 Sep		23 to 27 Sep	
	Week #1		Week #2		Week #3		Week #4		Week #5		Week #6		Week #7	
Session	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)
Lecture Session on Friday @ LT19A		Lecture #1 Briefing, Introduction, Data Attributes & Wrangling		Lecture #2 Visual Encoding, Visualisation Tools		Lecture #3 Basic Plots - Comparison and Composition Plots		Lecture #4 Basic Plots - Relation and Distribution Plots		Lecture #5 Visual Perception, Gestalt Principles		Lecture #6 Colour Perception, Designing the Visualisation		Lecture #7 Interactive Visualisation
Instructor	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wooi Boon	Wang Yong
Tutorial on Tuesday @LT19A			Tutorial #1 Data Attributes & Representation		Tutorial #2 Visual Encoding		Tutorial #3 Basic Plots and Charts (Part 1)		Tutorial #3 Basic Plots and Charts (Part 2)		Tutorial #4 Visual Perception (Part 1) Demo - PPT Video		Tutorial #4 Visual Perception (Part 2)	
Questions			Q1.2 - Q1.2		Q2.1 - Q2.4		Q3.1 - Q3.3		Q3.4 - Q3.5		Q4.1 - Q4.4		Q4.5 - Q4.6	
Assignment Deadlines														

Recess Week	07 to 11 Oct		14 to 18 Oct		21 to 25 Oct		28 Oct to 01 Nov		04 to 08 Nov		11 to 15 Nov		Date / Week No.
	Week #8		Week #9		Week #10		Week #11		Week #12		Week #13		
	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Tue (3.30pm - 5.20)	Fri (2.30pm - 4.20)	Session
		Lecture #8 Exploratory Data Analysis	Course Assessment Mid Semester Quiz (1.0 hour)		Lecture #9 Geospatial Visualisation		Lecture #10 Abstract Visualisation		Lecture #11 Scientific Visualisation			Course Assessment Final Quiz (1 hour)	Lecture Session @ LT19A
	Wang Yong	Wang Yong	Wooi Boon	Wang Yong	Wang Yong	Wang Yong	Wang Yong	Wang Yong	Wang Yong	Wang Yong	Wang Yong	Wang Yong	Instructors
	Tutorial #5 Interactive Data Visualisation			Tutorial #6 Exploratory Data Analysis		Tutorial #7 Geospatial Visualisation		Tutorial #8 Abstract Data Visualisation		Tutorial #9 Scientific Data Visualisation			Tutorial Topic @LT19A
			Mid-Semester Quiz									Final Quiz	Questions
								Project+Report Friday 1 Nov 12 noon				Critique 1&2 Friday 15 Nov 12 noon	Assignment Deadlines

Data Visualisation

Associate Professor Goh Wooi Boon
College of Computing and Data Science
Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 1 – Introduction

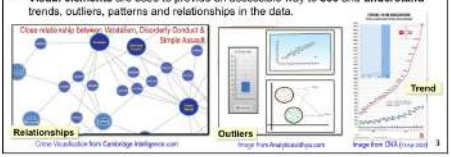
Contents

- What is Data Visualisation?
- Importance of Data Visualisation
- What will You Learn in this Course

Introduction

What is Data Visualisation?

- Data visualization is the **representation** of **data** or **information** in a graph, chart, map or other **visual format**.
- **Visual elements** are used to provide an accessible way to **see** and **understand** trends, outliers, patterns and relationships in the data.



Crime Visualisation from Cambridge Intelligence.com — Image from Analyticsvidhya.com — Image from DNA (Hong Kong)

Importance of Data Visualisation

The visual creature

- About half the human brain is directly or indirectly devoted to **visual processing**.
- Humans are much better at remembering **pictures** than words[1]. Students given **visuals** to go along with lectures have also been shown to have **better retention** than students listening to lectures without visual aids[2].
- The McGurk effect seems to also suggest that the **visual** sensory modality can **dominate** the auditory modality.

Image from Wikipedia

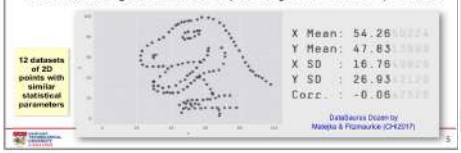
BBC video on McGurk Effect - https://youtu.be/G9NvVNnDmE

[1] Neural correlates of the episodic encoding of pictures & words - https://www.pnas.org/content/99/8/5523.full
[2] Wu, D. C. & McDaniel, M.A. (2012) – Enhancing learning during lecture - https://www.sciencedirect.com/science/article/abs/pii/S0191491X11001050

Importance of Data Visualisation

Blinded by statistics

- Basic statistical analysis alone may not provide useful insights into how datasets can have **differing characteristics** despite having **similar statistical** parameters.



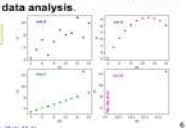
Importance of Data Visualisation

Blinded by statistics

- Basic statistical analysis alone may not provide useful insights into how datasets can have **differing characteristics** despite having **similar statistical** parameters.
- Statistician Francis Anscombe[3] used the **Anscombe's Quartet**[3] of four datasets to show why **visualisation** is a **crucial part of data analysis**.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
1	10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
2	8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
3	13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
4	9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
5	11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
6	14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
7	6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
8	4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
9	12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
10	7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
11	5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Correlation Coefficient	0.816		0.816		0.816		0.816	
Std Dev.	3.16	2.03	3.16	2.03	3.16	2.03	3.16	2.03

Anscombe's Quartet

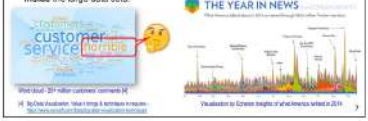


[3] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21

Importance of Data Visualisation

Seeing insights in big data

- The amount of data being generated is increasing rapidly. Gathering **meaningful insights** from such **big data** is not possible without good data visualisation.
- Good visualization allows users to easily identify relationships and patterns **hidden inside** the large data sets.



Word cloud - 35+ million customer comments [4]

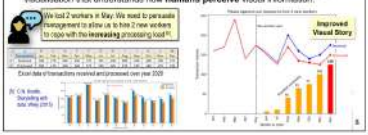
Visualisation by Echelon Insights of what America talked in 2019

[4] Big Data Visualization: Value it brings & techniques it requires - https://www.omnisci.com/blog/big-data-visualization-techniques

Importance of Data Visualisation

Telling the story

- Data visualisation can help **communicate the message** embedded in the data.
- Telling a visual story that conveys the **intended** message requires designing clear visualisation that understands how **humans perceive** visual information.



Excel data of transactions received and processed over year 2020

[X] C.N. Knafic, Storytelling with data, Wiley (2015)

Visualising Datasets

Is There An Ideal Way?

- There are many ways to visualise a dataset.
- Choosing the best way to visualise the dataset will depend on:
  - The data **types** (e.g. its scale of measure, spatial (geodata), temporal, etc)
  - The **context** and **relationship** within the data (e.g. highest value, part of a whole)
  - The **purpose** of the visualisation (e.g. identify trends, compare values, show composition, communicate a specific message)

[X] Chart from: https://www.analyticsvidhya.com/blog/2020/07/guide-to-exploratory-data-analysis/

Visualising Datasets

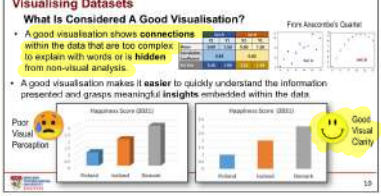
What Is Considered A Good Visualisation?

- A good visualisation shows **connections** within the data that are too complex to explain with words or is **hidden** from non-visual analysis.

	Mean	Std Dev
x	54.26	16.76
y	47.83	26.93
Correlation	-0.06	

From Anscombe's Quartet

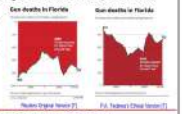
- A good visualisation makes it **easier** to quickly understand the information presented and grasps meaningful **insights** embedded within the data.



Ethical Visualisation

Staying True

- Ethics is the set of **moral principles** that guides a person's **decision** and **actions**.
- In data visualisation, there are many opportunities to **manipulate** and **distort** the message a person perceives using misleading or untruthful data representation.
- **Ethical conduct** is an important part of **your responsibility** when you design and present visual information.
- Your **professional integrity** requires you to be mindful of how your audience will interpret the information you present in your visualisation.
- So, communicate **clearly & truthfully.**



Reuters Original Version [7] — Fox Tampa's Cheat Version [7]

[7] Businessinsider.com, This Chart Shows An Alarming Rise In Florida Gun Deaths After 'Stand Your Ground' Was Enacted - https://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2

Welcome to Data Visualisation

What will you be learning in this course?

- **The Person** – the theory of human visual perception and its implications in designing effective data visualisation.
- **The Process** – from preparing the data to choosing and implementing the right visualisation solution for displaying and exploring various types of data.
- **The Practice** – solve data visualisation problems through design and code development for different application domains.
- **The Purpose** – learn to design visualisation narratives that effectively communicate your intended purpose and facilitate data analysis.

Summary

Introduction

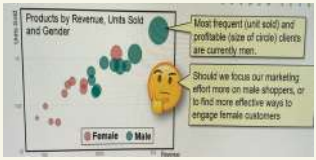
- Data visualisation is the transformation of data into a **visual form** so that **meaningful information** embedded within it can be made **more accessible**.
- Data visualisation is important to any disciplines or individual who **deals with data** and the need to **extract** or **share insights** from these data.
- Effective data visualisation serves two main purposes. It allows us to **explore** the dataset to gain useful insights and it allows us to **communicate** effectively what the data is telling us.

Data Visualisation

References for Introduction

[1] Neural correlates of the episodic encoding of pictures & words - https://www.pnas.org/content/99/8/5523.full
[2] Wu, D. C. & McDaniel, M.A. (2012) – Enhancing learning during lecture - https://www.sciencedirect.com/science/article/abs/pii/S0191491X11001050
[3] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21
[4] Big Data Visualization: Value it brings & techniques it requires - https://www.omnisci.com/blog/big-data-visualization-techniques
[5] C.N. Knafic, Storytelling with data, Wiley (2015)
[6] Chart from: https://www.analyticsvidhya.com/blog/2020/07/guide-to-exploratory-data-analysis/
[7] Businessinsider.com, This Chart Shows An Alarming Rise In Florida Gun Deaths After 'Stand Your Ground' Was Enacted - https://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2

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



Chapter 2 – Data Attributes & Wrangling

Contents

- Basic Data Attributes
- Other Data Attributes
- Data Wrangling → if missing data set

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.

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

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).

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.

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 Psychoacoustic Lab. Image from Newview.org

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

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).
- Valid operations on such data include (=, ≠, < and >).
- Visualisation techniques such as bar and pie charts can also be used for ordinal scale data.

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**.

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), κ_{∞}
- Valid operations on such data include (=, ≠, <, >, +, −, * and ÷).
- 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.

Review

The NOIR Scale of Measure

- Which scale represents each of the following data attributes:

1) The different types of coffee beans - N
2) Year coffee bean was harvested - I
3) Weight of coffee beans - R
4) Size of my latte (all 12 ounces), grande (16), venti (20)) - O

Nominal | Ordinal | Interval | Ratio

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+	1600.11
3	2	F	2004	2	30	F	800.13
4	3	F	2003	13	55.5	C+	1800.00
5	4	M	1999	10	72.5	B+	1800.00
6	5	M	2000	11	65	B	1500.22
7	6	F	2004	4	50	B-	1700.00
8	7	F	2005	13	90.5	A+	900.34

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 back at 0°C

The Amazing Property of Water

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.

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	Lung 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]

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

83% No

Students wore glasses

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.

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.

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).

- treemap
- Sunburst
- Circle Packing

- Polar area diagram
- Geoplot

n=6

n=5

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?

Fifty percent of Android owners are under the age of 35

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[4].

Tree Diagram | Treemap | Sunburst Diagram | Circle Packing

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

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

Summary

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.

References for Data Attributes

[1] Image Source – Bullet Graph – https://en.wikipedia.org/wiki/Bullet_graph
[2] Image Source – Gauges – https://apexcharts.org/gallery/radial-charts/multi/
[3] Image Source – Line Chart https://www.bls.gov/mlr/articles/employment-situation-july.htm
[4] Image Source – Pie Chart – https://www.economics.mit.edu/grad/sdm/spring-2011/project-ideas-and-data/singapore-import-data-492.png
[5] Stacked column Chart from https://search.cpd.org/?q=child%20labor#/?q=child%20labor&f%5B0%5D%5B0%5D=dataset%3A%22Global%20Human%20Capital%20Project%22&f%5B0%5D%5B1%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Sub-Saharan%20Africa%22&f%5B0%5D%5B2%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Latin%20America%20and%20the%20Caribbean%22&f%5B0%5D%5B3%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20East%20Asia%20and%20Pacific%22&f%5B0%5D%5B4%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Europe%20and%20Central%20Asia%22&f%5B0%5D%5B5%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Middle%20East%20and%20North%20Africa%22&f%5B0%5D%5B6%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20South%20Asia%22&f%5B0%5D%5B7%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Oceania%22&f%5B0%5D%5B8%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Canada%20and%20United%20States%22&f%5B0%5D%5B9%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%22&f%5B0%5D%5B10%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%22&f%5B0%5D%5B11%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%22&f%5B0%5D%5B12%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%22&f%5B0%5D%5B13%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Female%22&f%5B0%5D%5B14%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Male%22&f%5B0%5D%5B15%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Adjusted%20for%20Inequality%22&f%5B0%5D%5B16%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Adjusted%20for%20Inequality%22&f%5B0%5D%5B17%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Female%20-%20Female%22&f%5B0%5D%5B18%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Male%20-%20Male%22&f%5B0%5D%5B19%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Adjusted%20for%20Inequality%20-%20Female%22&f%5B0%5D%5B20%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Adjusted%20for%20Inequality%20-%20Male%22&f%5B0%5D%5B21%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B22%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B23%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B24%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B25%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B26%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B27%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B28%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B29%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B30%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B31%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B32%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B33%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B34%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B35%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B36%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B37%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B38%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B39%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B40%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B41%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B42%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B43%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B44%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B45%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B46%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B47%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B48%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B49%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B50%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B51%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B52%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B53%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B54%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B55%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B56%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B57%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B58%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B59%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B60%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B61%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B62%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B63%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B64%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B65%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B66%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B67%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B68%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B69%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B70%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B71%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B72%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B73%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B74%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B75%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B76%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B77%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B78%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B79%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B80%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B81%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B82%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B83%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B84%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B85%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B86%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B87%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B88%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B89%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B90%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B91%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B92%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B93%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B94%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B95%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B96%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B97%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B98%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B99%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B100%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B101%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B102%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B103%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B104%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B105%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B106%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B107%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B108%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B109%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B110%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B111%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B112%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B113%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B114%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B115%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B116%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B117%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B118%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B119%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B120%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B121%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B122%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B123%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B124%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&f%5B0%5D%5B125%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Male%22&f%5B0%5D%5B126%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Female%22&f%5B0%5D%5B127%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Female%20-%20Female%22&f%5B0%5D%5B128%5D=dataset%3A%22Global%20Human%20Capital%20Project%20-%20Global%20Average%20-%20Adjusted%20for%20Inequality%20-%20Male%20-%20Male%22&

Chapter 2.3 – Data Wrangling

Contents

- What is and Why the need for Data Wrangling
- Pandas and Dataframes
- Combining Dataframes
- Reshaping Dataframes
- Extracting Subsets
- Sorting Values and Index
- Handling Missing Values

Chapter 2.3 – Data Wrangling

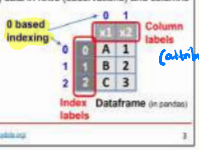
What is and Why the Need for Data Wrangling?

- Data wrangling is the process of **cleaning**, **structuring** and **enhancing** raw data into a format that is more appropriate and readily used for analysis & visualisation.
- The data available is often not in a **suitable form** for use with data visualisation tools. Some examples include:
  - Data needed are stored in **different files**. They need to be retrieved and combined into a single table for analysis.
  - Data available is extensive. Only a **subset is needed** for analysis.
  - Data in the table is **not organised** in the manner suitable for analysis.
  - Data in the table has **missing** or **outlier values** that can affect the analysis.
  - **New enhanced data** (e.g computed mean) must be entered in the table for analysis.

Data Wrangling - Getting Started

Pandas and Dataframes

- **Pandas** is a powerful and easy to use **open-source** data analysis and manipulation **library**, built on top of the **Python** programming language.
- **Dataframes** are used in pandas for storing data in rows (observations) and columns (variables or dimensions).
  - **Columns** are dimension of observations (variables) and are given a **label** (e.g. x1).
  - The row labels are referred to as **index** and has default labels of (0, 1, 2..) but can be renamed to any label.
  - Dataframes uses **0 based indexing**.



Combining Dataframes

Merge

- Data from two dataframes can be combined into a dataframe in many ways using the **merge operator**.
- Two dataframes can be **merged** using qualifiers like **left**, **right**, **inner** and **outer**.
- **Missing values** in the merged dataframe are given the **NaN** (Not a number) values.

Combining Dataframes

Concatenate

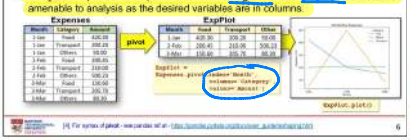
- Data from 2 dataframes (or the same frame itself) can be appended using **concat**.
- The dataframes can be appended based on **rows** or **columns** (use **axis=1**)
- **Missing values** in the concatenated dataframe are given the **NaN** values.

use pd.Concat([adf,bdf],ignore_index=True)

Reshaping Dataframes

Pivot

- Data reshaping **rearranges** the form of the data **without changing** the content of the dataset so as to facilitate the next stage of analysis (e.g. plotting a graph).
- The **pivot** operator reshapes a table in the **long form to a wide form** that is more amenable to analysis as the desired variables are in columns.

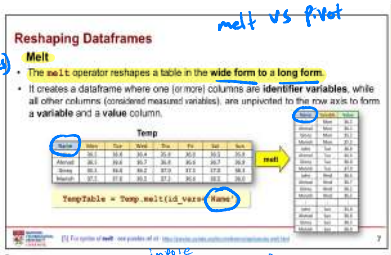


melt vs Pivot

Reshaping Dataframes

Melt

- The **melt** operator reshapes a table in the **wide form to a long form**.
- It creates a dataframe where one (or more) columns are **identifier variables**, while all other columns (considered measured variables), are unpivoted to the row axis to form a **variable** and a **value** column.

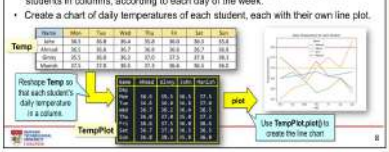


pd.Categorical (input order of axis)

Think and Apply

Reshaping Dataframe with Pivot and Melt

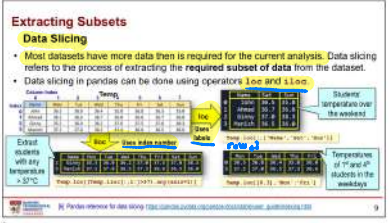
- The table in the dataset *"Daily Temperature.csv"* list the temperatures of four students in columns, according to each day of the week.
- Create a chart of daily temperatures of each student, each with their own line plot.



Extracting Subsets

Data Slicing

- Most datasets have more data then is required for the current analysis. Data slicing refers to the process of extracting the **required subset of data** from the dataset.
- Data slicing in pandas can be done using operators **loc** and **iloc**.



Rearranging Data

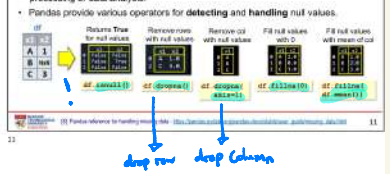
Sorting values and index

- Data in a table may need to be sorted based on one or more selected **attributes** to facilitate further processing (e.g. plotting an ascending bar chart)
- Pandas provides the **sort_values** and **sort_index** operators for this purpose.

Handling Missing Values

Pandas operators for NaN values

- Raw datasets and reshaped dataframes often contain missing or **null values** (NaN). These null values need to be accounted for and replaced/removed before further processing or data analysis.
- Pandas provide various operators for **detecting** and **handling** null values.

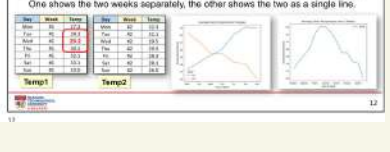


drop row drop column

Think and Apply

Data Wrangling Exercise

- Two temperature tables for week #1 and #2 are stored in **two different datafiles**.
- **Missing value** needs to be **interpolated** from its nearest data points.
- **Combine** and **reshape** the two dataframes to produce **two different line plots**. One shows the two weeks separately, the other shows the two as a single line.



What is Data Interpolation?

Last Updated : 07 Feb, 2024

The use of real-world data in machine learning tasks can sometimes present challenges due to missing values or incomplete datasets. Such data can lead to inaccurate predictions, and ignoring the missing values can cause bias in model training and distort the original distribution of the data. As machine learning algorithms are not designed to handle missing data, it is important to either remove the missing values or fill the missing positions with other data. One way to fill in missing data values is through a process called Data Interpolation. In this blog, we will be thoroughly discussing about data interpolation.

What is Data Interpolation?

Data interpolation stands out as a crucial technique in data preprocessing, serving the purpose of estimating unknown values within the range of known data points. This method utilizes the existing data points to infer and fill in missing or unknown values in a dataset. The significance lies in its ability to replace missing values with predicted ones, enhancing the completeness and reliability of the dataset. In essence, data interpolation acts as a systematic tool, leveraging the available information to bridge gaps and provide a more comprehensive view of the data.

What data interpolation does is that it maps out an estimate between the known values such that the values present in between the known values that are absent could be replaced with the prediction through the estimated map. The assumption in the data interpolation process is that the changes between different points is continuous and smooth in nature. Through this assumption the data interpolation process is able to predict the unknown missing values.

Difference Between Interpolation and Extrapolation:

As data interpolation is a useful technique in estimating the data value in the range of the present data, data extrapolation helps in predicting the value outside of the range of data that is given to us. Interpolation and Extrapolation are sort of similar techniques which are used to handle data, both the processes serve different purposes have different merits of their use and different demerits. Here we will be discussing the key differences between these two processes.

Need of Data Interpolation

Data Interpolation have different uses in applications related to data science, especially in the field of data analysis and scientific research. It can help in enhancing our understanding of the data which is present, and it can elaborate the trends seen in the data in a better manner, let's see some of the most important reasons why data interpolation is performed:

- **Missing Data Handling:** Data Interpolation helps in handling missing data points in the dataset. This is an essential step to be followed in order to fill up the gap in the data which occurs due measurement errors, malfunctioning equipment or some other reason.
- **Smoothing Noisy Data:** Data Interpolation helps in smoothing out the dataset by reducing the impact of random fluctuations and create a clear-cut representation of underlying trends in the data. The sudden spikes in the dataset could be removed with the help of data interpolation, we use the interpolation method to interpolate the point where the spike in the data is present.
- **Creating Continuous Dataset:** Continuous distribution of dataset could be created with the help of data interpolation. the specified data interpolation method could be used to find out the data point intermediate data points in between the given data points such that a continuous dataset is created.
- **Enhancing Visualization:** As we have seen data interpolation provides a smoother and continuous dataset which becomes easy to visualize and understand, the reduction of noise and the enhancement of trends in the dataset increases the aesthetics of the visualizations, ultimately increasing the ability of visualizations to extract useful information.

	Interpolation	Extrapolation
Definition	Interpolation predicts the values which are present within the known data range.	Extrapolation estimates the values which are present outside of the known data range.
Uncertainity	Interpolation is considered to be a more reliable process as it depends on the observed data	Extrapolation involves high risk. It assumes that the observed trends in the data continues outside of the known data range.
Dependence on Data	Interpolation directly depends on the known data points.	Extrapolation depends on the assumption that the trends in the known data continues outside the data range.
Accuracy	Interpolation is often more accurate than extrapolation since it is based on the existing data.	Extrapolation is less accurate in nature as it assumes the data trends to continue for values out of range.
Example	Measurement of the atmospheric pressure is present at 1km above the ground and 2km above the ground and atmospheric pressure at 1.5 km from the ground is required.	Measurement of the atmospheric pressure is given at 1km above the ground and 2km above the ground and atmospheric pressure at 3 km from the ground is required.

Summary

Data Wrangling

- Data wrangling is an important step in the process of data visualisation as it **prepares** that data into a **format** that can be **readily plotted** with the various visualisation routines.
- Data created or loaded from various stored sources need to be **combined** into a suitable single data structure called a **dataframe**.
- Dataframes often need to be **reshaped** into the **appropriate format** so that it is suitable for used with different visualisation routines.
- Data in dataframes may also need to be **sorted** into a specific order for display.
- **Missing data** need to be **filled in** or **removed** in order to ensure they do not **interfere** with the visualisation process.

References – Data Attributes and Wrangling

[1] Pandas – Python Data Analysis Library - https://pandas.pydata.org/
[2] For syntax of merge - see Data Wrangling with Pandas Cheat Sheet - https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf
[3] For syntax of concat - see pandas ref at - https://pandas.pydata.org/docs/reference/generated/pandas.concat.html
[4] For syntax of pivot - see pandas ref at - https://pandas.pydata.org/docs/reference/ ; https://pandas.pydata.org/docs/reference/generated/pandas.melt.html
[5] For syntax of melt - see pandas ref at - https://pandas.pydata.org/docs/reference/generated/pandas.melt.html
[6] Pandas reference for data slicing https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html
[7] Refer pandas sorting - https://www.python.org/wiki/sorting/ & https://pandas.pydata.org/docs/reference/generated/pandas.DataFrame.sort_values.html
[8] Pandas reference for handling missing data - https://pandas.pydata.org/docs/reference/generated/pandas.isnull.html

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

Chapter 3 – Visual Encoding

Contents

- Visual Encoding
- Deconstructions

Chapter 3.1 – Visual Encoding

Contents

- Visual Variables
- Visual Marks
- Visual Encoding Examples

The Data Visualisation Process

An Overview

- Data visualisation is the transformation of data into an image using an appropriate visual encoding strategies.

Month	Category	Amount
1-Jan	Food	428.38
1-Jan	Transport	306.20
1-Jan	Others	50.00
2-Feb	Food	306.40
2-Feb	Transport	210.00
2-Feb	Others	300.10
3-Mar	Food	200.40
3-Mar	Transport	301.70
3-Mar	Others	80.00

Visual Encoding

Mapping from data to image

Data
Data & Conceptual Models

Image
Visual Marks and Channels (Variables)
Lines **Colour**

4

Visual Variables

Bertin's Visual Variables

- Jacque Bertin, a French cartographer proposed set of **visual variables** in 1967 that can be used to construct map symbols and other graphical techniques.

> "*A graphic is not only a drawing; it is a responsibility, sometimes a weighty one, in decision-making.*"
>
> – Jacque Bertin
> French Cartographer (1918-2010)

Image from Wikipedia

[1] Bertin's quote from: J. Bertin, graphics and graphic information processing (1981), https://www.google.com/search?q=j+bertin+graphics+and+graphic+information+processing&ie=utf-8&oe=utf-8&aq=t&rls=com.microsoft:en-gb:ie&rlz=1I7ADBF_en-GB#sclient=psy-ab&hl=en-GB&q=%22A+graphic+is+not+only+a+drawing%2C+it+is+a+responsibility%2C+sometimes+a+weighty+one%2C+in+decision-making%2C%22+-+Jacque+Bertin&aq=f&aqi=&aqi=&aqi=

Chapter 3.2 – Deconstruction

Contents

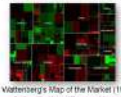
- Playfair's Import/Export (Line Charts)
- Wattenberg's Map of the Market (Tree Map)
- Fry's Basketball Salary versus Performance (Slope Chart)

Deconstruction

Seeing How Others Do It

- The process of deconstruction allows us to **analyse** how others have visually encoded the information they wanted to convey in their dataset.
- Deconstruction is a great way to **learn how to construct** your own visualisation, especially if the examples analysed have been effective in telling their visual story.

Playfair's Import/Export (1786)
A good old pre-digital chart

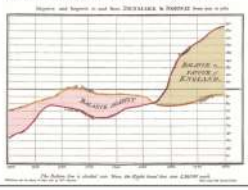


Wattenberg's Map of the Market (1998)
A visualisation still current in use
checked: https://finviz.com/map.ashx

Fry's Salary vs Performance (2014)
Visualisation for US Basketball fans
checked: https://fathom.info/salaryper/

Deconstruction - Example #1

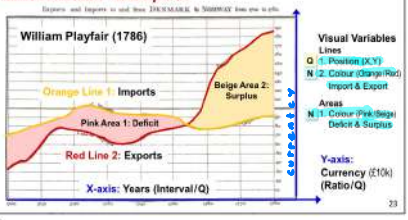
William Playfair's Line Charts (1786)



- William Playfair's Commercial and Political Atlas was published in 1786
- Playfair examined the imports and exports between Britain and various countries.
- To illustrate these trade relationships, Playfair created the first ever line charts that show the change in trade over time.

[6] Image can be found at https://upload.wikimedia.org/wikipedia/commons/5/52/Playfair_TradeBalance-3.jpg

Deconstruction - Example #1



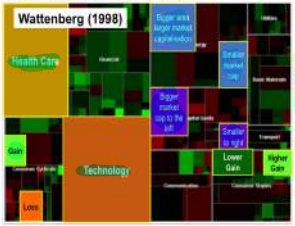
Deconstruction - Example #2

Martin Wattenberg's Map of the Market (1998)

- Martin Wattenberg & co. created one of the 1st web-based visualisations, Map of the Market.
- It displayed live stock market data for more than a decade, starting in 1998, while Martin was at SmartMoney.com.
- Map of the Market uses a modified **Tree Map** algo that creates squarish tiles.

[7] Map of the Market details can be found at http://www.kasdim.com/marketmap.html

Deconstruction - Example #2



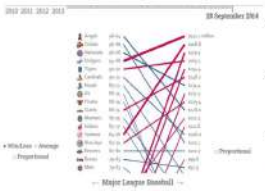
Visual Variables

Areas (Rectangle)

Q 1. Size (Market Cap)
N 2. Position (Market Sector)
O 3. Position (Market Cap)
N 4. Colour Hue
(Green – gain)
(Red – loss)
Q 5. Colour Saturation
(Bright – higher gain)
(Dark – lower gain)
(Bright – higher loss)
(Dark – lower loss)

Deconstruction - Example #3

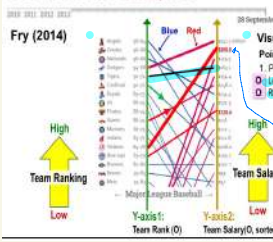
Ben Fry's Slope Chart (2010-2014)



- Ben Fry is a visualisation expert and principal of Fathom, a design and software consultancy in Boston, USA.
- He designed Salary vs. Performance to visualised if baseball teams are spending their money well over the season.
- The visualisation uses a slope chart.

[8] Fry's website for Salary vs Performance https://fathom.info/salaryper/

Deconstruction - Example #3



Think and Apply

Visualising Relationships

- Joshua Keating & Chris Kirk (@Slate.com) created the **Middle East Friendship Chart** to visualise the complex relationships between different players in Syria's civil war.

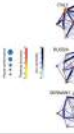
- Draw the table for the relational data model.
- What visual marks and variables are used here?
- Redundant encoding?
- Colour choices?
- Effective?

[9] Middle East Friendship Chart at http://www.slate.com/blogs/the_slate/2013/10/10/syria_civil_war_alliances_explained.html

Think and Apply

Visualising Performance in Team Games

- Duch et al. analysed three Euro 2008 matches of the Spanish team using a visualisation technique from social network analysis.



- What types of data variables are visually represented in the image.
- What perceptual channels were used to encode their values?
- Colour choices? Can be better?

[10] Visualisation described in Krougan http://krouge.com/blog/2008/06/euro-2008-social-network-analysis.html
Duch et al.'s visualisation described in the paper: http://journals.aps.org/pre/abstract/10.1103/PhysRevLett.95.548702

Summary

Deconstruction

- Data visualisation is both an **art** and a **science**.
- Designing **informative** and **visually appealing** visualisations requires us to learn from many other visualisation experts.
- We can do this by **deconstructing** their good pieces of work.
- We observe how they make use of the various visual marks and variables to **encode useful information** embedded in the datasets.
- We need to do these observation with a **critical eye,** to see what has **worked well** and what has not and **can be improved**.

References for Visual Encoding and Deconstruction

[1] Bertin's quote from: J. Bertin, graphics and graphic information processing (1981)
[2] T. Munzner, Visualization Analysis & Design, CRC Press (2015)
[3] J. Mackinlay, Automating the Design of Graphical Presentation of Relational Information (1986), http://www.tableau.com/sites/default/files/publications/assets/p110-mackinlay.pdf
[4] C. Ware, Information Visualization: Perception for Design, 3rd Edition, Morgan Kaufmann (2013).
[5] William Playfair's image can be found at: https://upload.wikimedia.org/wikipedia/commons/5/52/Playfair_TradeBalance-3.jpg
[6] William Playfair's image can be found at: https://upload.wikimedia.org/wikipedia/commons/5/52/Playfair_TradeBalance-3.jpg
[7] Martin Wattenberg's Map of the Market details can be found at: http://www.kasdim.com/marketmap.html
[8] Ben Fry's website for Salary vs Performance: https://fathom.info/salaryper/
[9] Keating & Kirk's Middle East Friendship Chart at: http://www.slate.com/blogs/the_slate/2013/10/10/syria_civil_war_alliances_explained.html
[10] Duch et al.'s visualisation described in the paper: http://journals.aps.org/pre/abstract/10.1103/PhysRevLett.95.548702

Chapter 4 –Visualisation Tools

Contents

- Types Of Visualisation Tools
- Choosing The Visualisation Tools
- Data Visualisation Tools Used In This Course
- Other Python Data Visualisation Libraries

Types Of Visualisation Tools

"Like a Child in a Candy Store"

- There is a huge variety of tools available to refine and query your data, and to create, edit, and display your visualisations.
- The types of tools for building visualisations fall into four major categories[1]:
  - Basic productivity applications: Microsoft Excel, LibreOffice, Google Sheet
  - Visualisation software: Paraview, ArcGIS, D3.js
  - Business intelligence tools: tableau, Power BI, Qlik QlikView
  - Developer based packages: D3.js - Javascript; bokeh, seaborn, matplotlib - Python-based

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

3

Choosing The Visualisation Tool

How do I Choose the Right One?

- When given a free choice, it is important to select the tool that is most suitable to support the nature and purpose of your data visualisation work and workflow.
- Consider the following aspects in the evaluation[1]:
  - Sharing
  - Output
  - Display types
  - Data Exploration
  - Interoperability
  - Simplicity
  - Persistence

Image taken from [1]

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

4

Choosing The Visualisation Tool

How do I Choose the Right One?

- **Output** – Are you publishing the visualizations to the web, PDF documents or embedding them into other applications? The destination of your visualization will dictate tool choice. e.g. not all tools have good web interactive support[1].
- **Display types** – What types of visualizations do you intend to create? e.g. maps, networks, and text-based visualizations may not be supported by some tool[1].
- **Interoperability** – How easily can you connect to other data sources? For instance, does the software allow you to import diverse file types, such as .xls, CSV, .txt, or allow you to link to your databases?[1].
- **Data exploration** – Do you need a tool to explore and present your data visually, or just to present your data visualization visuals? e.g. features such as visual querying may not be supported by some tools[1].

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

5

Choosing The Visualisation Tool

How do I Choose the Right One?

- **Simplicity** – Are you looking to create charts and graphs quickly? How much time can you afford to learn how to use the tool as some may be powerful but require a steep learning curve, even to build a simple bar chart[1].
- **Sharing** –Can others view and edit your visualization and analysis? If your workflow is essentially collaborative, then choose a tool that enables collaboration during the design, prototyping and creation process[1].
- **Persistence** – Do you think that you will have to go back and revise the visualizations you create? If so, it is important to choose a tool that is well supported and will be around for a while[1].

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

6

Data Visualisation Tools Used In This Course

Data Visualisation Using Python

- This data visualization course is specifically designed for students with basic computer science background, as such programming skills are assumed.
- The Python-based tools are supported by numerous **visualisation** and **statistical** libraries that will facilitate both **explanatory** visualisation and **exploratory** analysis.

matplotlib, seaborn – Static Visualisation	pandas – Data Manipulation; bokeh – Interactive Visualisation	SciPy, statsmodels – Statistical Analysis

7

Data Visualisation Tools Used In This Course

Other Data Visualisation Tools

- This course focuses on the principles and practice of good **human-centric** data visualisation **design** and not the visualisation tool per se.
- As such, students are free to use any other data visualisation tools they are familiar with, as long as it can support the required deliverables in the course.

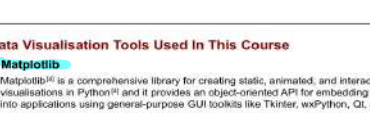
Microsoft Excel – Widely available tool for quick non-programming charts	**D3.js** – Works well with web standards like HTML, CSS and SVG	**tableau, Power BI** – These are commercial tools. Free version may be limited.

8

Data Visualisation Tools Used In This Course

Pandas

- Pandas is the Python data analysis library[2].
- It is a powerful, flexible and easy to use open source **data analysis** and **manipulation** tool, built on top of the Python programming language.
- Pandas can also support a variety of basic charts and plots[3] such as:



Line plots | Stacked bar plots | Box plots

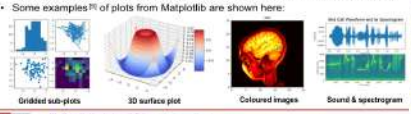
[2] Pandas – Python Data Analysis Library - https://pandas.pydata.org/
[3] Pandas – Chart Visualization - https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html

9

Data Visualisation Tools Used In This Course

Matplotlib

- Matplotlib[4] is a comprehensive library for creating static, animated, and interactive visualisations in Python[4] and it provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, etc.
- Some examples[5] of plots from Matplotlib are shown here:



Gridded sub-plots | 3D surface plot | Coloured images | Sound & spectrogram

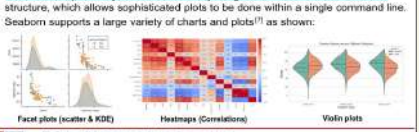
[4] Matplotlib - Visualization with Python - https://matplotlib.org/
[5] Matplotlib – Example plots in Matplotlib - https://matplotlib.org/stable/tutorials/introductory/sample_plots.html

10

Data Visualisation Tools Used In This Course

Seaborn

- Seaborn[6] is a Python library predominantly used for making **statistical graphics.**
- It is built on top of **Matplotlib** and is closely integrated with **Pandas**'s dataframe structure, which allows sophisticated plots to be done within a single command line.
- Seaborn supports a large variety of charts and plots[7] as shown:



Facet plots (scatter & KDE) | Heatmaps (Correlations) | Violin plots

[6] Seaborn – Statistical data visualisation - https://seaborn.pydata.org/
[7] Seaborn – Example Gallery - https://seaborn.pydata.org/examples/index.html

11

Data Visualisation Tools Used In This Course

Bokeh

- Bokeh[8] is a Python library for creating **interactive** visualisations for web browsers.
- Bokeh can produce elegant visualisation (including dashboards) that are interactive and works with large or streaming datasets.

Embedding widgets like slider & menus with Bokeh | Example Dashboard create using Bokeh[9]

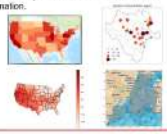
[8] Bokeh - https://bokeh.org/
[9] Will Koehrsen – Data Visualization with Bokeh in Python – https://towardsdatascience.com/data-visualization-with-bokeh-in-python-part-one-getting-started-a11655ef47f8

12

Other Python Data Visualisation Libraries

Visualising Geospatial Data

- The are many useful visualisation libraries in Python that allow us to visualise geographical and location-based information.
- **GeoPandas**[10] – extends the datatypes used by Pandas to allow spatial operations on geometric types. Its goal is to make working with geospatial data in Python easier.
- **Geoplot**[11] – is a high-level Python geospatial plotting library. It is built on top of matplotlib with the specific purpose of doing cartographic plots.



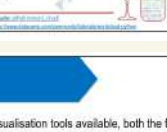
[10] GeoPandas - https://geopandas.org/
[11] Geoplot: geospatial data visualization - https://residentmario.github.io/geoplot/index.html

13

Other Python Data Visualisation Libraries

Visualising Textual Data

- The are other visualisation libraries in Python that allow us to visualise various properties in text-based data.
- **WordCloud**[12] – is a text visualization library that can analyse the frequency of words and correlates the size and opacity of a word to its frequency within a body of text.
- It has an interesting feature that allows an image mask to define the shape over which the word cloud is populated. This can serve as a visual context to the document being analysed[13].



[12] WordCloud for Python documentation - https://amueller.github.io/word_cloud/
[13] Duong Vu, Generating WordClouds in Python - https://www.dataquest.io/blog/generating-wordclouds-in-python/

14

Summary

Visualisation Tools

- There are numerous types of visualisation tools available, both the free open-source variety and sophisticated commercial tools.
- Choosing a suitable tool depends on the purpose of the visualisation. Factors related to the **design** (e.g. collaborative, fast turnaround) and **deployment** environments (e.g. interactive, web-based) should also be considered when making the choice.
- This course has adopted open source tools based on the **Python programming** environment as it supports many different visualisation libraries with good support for data wrangling and statistical analysis.

15

References for Visualisation Tools

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/
[2] Pandas – Python Data Analysis Library - https://pandas.pydata.org/
[3] Pandas – Chart Visualization - https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html
[4] Matplotlib: Visualization with Python - https://matplotlib.org/
[5] Matplotlib – Sample plots in Matplotlib - https://matplotlib.org/stable/tutorials/introductory/sample_plots.html
[6] Seaborn – Statistical data visualisation - https://seaborn.pydata.org/
[7] Seaborn – Example Gallery - https://seaborn.pydata.org/examples/index.html
[8] Bokeh - https://bokeh.org/
[9] Will Koehrsen – Data Visualization with Bokeh in Python - https://towardsdatascience.com/data-visualization-with-bokeh-in-python-part-one-getting-started-a11655ef47f8
[10] GeoPandas - https://geopandas.org/
[11] Geoplot: geospatial data visualization - https://residentmario.github.io/geoplot/index.html
[12] WordCloud for Python documentation - https://amueller.github.io/word_cloud/
[13] Duong Vu, Generating WordClouds in Python - https://www.dataquest.io/blog/generating-wordclouds-in-python/

Note: All online articles were accessed between Jul – Aug 2021

16

Chapter 5 – Basic Plots and Charts

Contents

- Comparison Plots
- Composition Plots
- Relation Plots
- Distribution Plots

+ Smoothing

© A/P Goh Wooi Boon (SCSE/NTU)

Chapter 5.1 – Visualising Comparison in Data

Contents

- Basic Comparison Plots
- Line Charts
- Bar Charts

© A/P Goh Wooi Boon (SCSE/NTU)

3

Basic Comparison Plots

Draw the Line or Raise the Bar?

- Line charts should be use for **time series** (i.e. chronological) or when the x-axis has an **interval scale** (e.g. distances from a point, size of an object). While bar charts should be use when the x-axis have **nominal** (i.e. unordered) scale.[1].
- Line charts should be used to show **trends** or pattern changes but **not precise quantitative values**. Bar charts could be used if specific measurements are to be **compared** at each **discrete intervals** along the x-axis.

Profits in 2019 | Profits in 2019 | Last 3 Years | Comparing Last 3 Yrs

[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

Dept = nominal vs discrete.

4

Line Charts

Axes Scaling and Aspect Ratio

- The aspect ratio of a graph is an important factor for judging **rate of change**.
- The "**banking to 45°**" rule[2] can be used to improve the visually encoding accuracy of changing trends.
- The axes scales or aspect ratio are adjusted such that the arithmetic average of all line orientations is nearly +45° or -45° (depending on the the direction of the trend).

Voltage change | Voltage change | Voltage change

too steep

Too flat

Time (s)

Taken from: [2] W.S. Cleveland, The Elements of Graphing Data, Hobart Press (1994)

5

Can plot horizontal

Bar Charts

Vertical or Horizontal Bars

- Avoid using vertical text. If the x-axis categories have **long text** descriptions, consider using **horizontal bars**.
- The nature of the **conceptual data**[1] can also help decide which bar orientation better **scaffold the interpretation** of the data. (e.g. height - vertical, distance - horizontal)

Long text descriptions

3 Tallest Building | 3 Longest Snakes

[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

6

Zero Baseline

Zero Heroes

- Some argue that zero baseline must be used for **all plots**[3], whether bar or line charts to maintain visual "truthfulness".
- **Bar charts** should normally have a zero baseline as data is encoded using length and truncating the axis is naturally misleading.
- There is some flexibility with **line** charts, especially when the changing trends encoded in the **angle cannot be discern** if zero-baseline is imposed.

better with zero baseline

! failed here

Visually exaggerated annual profit growth | Actual profits plotted with zero base

Up trend not visible when zero base is imposed but becomes apparent when relaxed

[3] C Shelton, Bar charts should always start at zero. But what about line charts? (2018), http://www.chadsheltonblog.com/2018/08/bar-charts-should-always-start-at-zero.html

7

13

Ethical Visualisation

Don't Exaggerate

- Head of **Dept B** is concerned about the lack of growth in his product line compared to the huge growth in **Dept C**, which is likely to surpass Dept B soon. He has to make a yearly review presentation to the board and started working on his chart...

Department Profits After 2018 | Dept W's Increase in Profits since 2020

Has ethical visualisation been violated here?

Not zero base

Changing profits of departments B and C in the last 4 years | The chart that the Head of Dept B presented the board of directors

8

Visualising General Trends

Smoothing the Series

- Smoothing line or bar charts can reveal the more general or long-term trends in the data. Smoothing is particularly useful for noisy time series data.
- However, care must be taken to ensure that such smoothing does not affect the integrity of what the data represents[1].

COVID-19 new cases in Singapore (Mar-Aug'20)

Daily fluctuations & spikes in cases makes it hard to see the general trend

Mean smoothing over 7 days centred about each daily sample

[1] Jon Peltier, When you use a smoothed line chart, your data is not affected. It's misrepresented!

9

Think and Apply

When Smoothies are a Bad Idea

- Smoothing the data series distorts the plot and create misleading visual trends.

bad smoothing = change the meaning

Year	Dept A	Dept B	Dept C
2013	4.0	3.0	2.1
2014	3.0	2.0	2.2
2015	2.5	1.6	1.0
2016	3.5	1.4	3.4
2017	7.0	1.0	7.0
2018	8.0	1.2	3.0
2019	9.4	1.0	3.0
2020	9.4	1.0	3.0

Profits in last 7 years (2014)

Profits of three departments over 7 years

Profit Variations Over 7 years | Smoothing Can Distort Actual Trends

Line Chart showing exact values | Line plots smoothed using Excel "smoothed line" feature

10

Data Smoothing

Smoothing Data Using Pandas

- Data smoothing[1] is done over a sliding window of a specified **width** . A **wider** window that covers more data points will result in **more smoothing**.

WS = Window Size

```
Smooth = Data.rolling(
    window = WS
    center = CN
    win_type = WT).mean()
```

Window size

WS = 7 | WS = 21

[1] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html

11

Data Smoothing

Smoothing Data Using Pandas

- Data smoothing[1] is done over a sliding window of a specified **width** . A **wider** window that covers more data points will result in **more smoothing**.
- The window can be **centred** or to the **edge** of the data point during smoothing and can take different "**shapes**"[2] to reduce the influences of data further away in the window when computing each **mean**.

```
Smooth = Data.rolling(
    window = WS
    center = CN
    win_type = WT).mean()
```

Centering (default is right edge)

CN = True | CN = None

to front

[1] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html
[2] Window types for win_type - https://docs.scipy.org/doc/scipy/reference/signal.windows.html#module-scipy.signal.windows

12

Data Smoothing

Smoothing Data Using Pandas

- Data smoothing[1] is done over a sliding window of a specified **width** . A **wider** window that covers more data points will result in **more smoothing**.
- The window can be **centred** or to the **edge** of the data point during smoothing and can take different "**shapes**"[2] to reduce the influences of data further away in the window when computing each **mean**.

```
Smooth = Data.rolling(
    window = WS
    center = CN
    win_type = WT).mean()
```

Window shape (default is equal weight)

WT = None | WT = 'cosine'

to every data is equal

[1] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html
[2] Window types for win_type - https://docs.scipy.org/doc/scipy/reference/signal.windows.html#module-scipy.signal.windows

13

Summary

Comparison Plots

- **Line** and **bar** charts are two basic plots for comparing data values.
- **Bar charts** are useful for **comparing values** between variables, especially those having **nominal** scales of measure.
- **Line charts** are effective in observing **changing trends** in the data, especially time series data.
- **Data smoothing** is sometimes needed in order to observe **trend** changes that occur over a **longer time** duration or sampling span.
- However, be careful that smoothing does not **distort** the true interpretation of the data values.

1

References for Comparison Plots

[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

[2] W.S. Cleveland, The Elements of Graphing Data, Hobart Press (1994)

[3] C. Shelton, Bar charts should always start at zero. But what about line charts? (2018), http://www.chadsheltonblog.com/2018/08/bar-charts-should-always-start-at-zero.html

[4] Jon Peltier, When you use a smoothed line chart, your data is not affected. It's misrepresented! https://www.peltiertech.com/2011/12/when-you-use-smoothed-line-chart-your-data/

[5] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html

[6] Window types for win_type - https://docs.scipy.org/doc/scipy/reference/signal.windows.html#module-scipy.signal.windows

! need to know when to use line chart/bar chart in vid presentation

Chapter 5.2 – Visualising Composition in Data

Contents

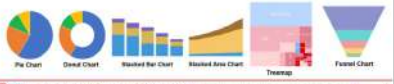
- Basic Composition Plots
- Pie Charts
- Stacked Bar Charts
- Stacked Area Charts

© A/P Goh Wooi Boon (SCSE/NTU)

Basic Composition Plots

Part-to-Whole Analysis

- Composition plots help visualise how individual parts comprise a whole.
- **Pie and donut charts** – are used for visualising compositions that are **static**.
- **Stacked charts** – are for composition that are **changing** with **another dimension** like time or category. Stacked **bar** or stacked **area** charts are common examples.
- Some other charts for viewing composition include the **treemap** and **funnel** chart.

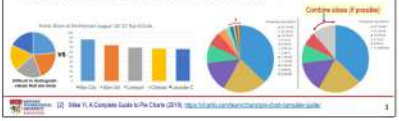


[1] How to Choose Charts to Show Data Composition (2016)

Pie Charts

When Should I Not Have A Pie?

- Pie charts are not suitable when the slices have **little variations**. Bar charts provide more accurate **visualisation of values** if they are deemed important and of interest.
- Pie charts are not suitable when there are **too many slices**. Consider if some of the smaller slices can be combined into an 'Other' category[3].



[2] Mike Y, A Complete Guide to Pie Charts (2019)

Pie Charts

What Type Of Pie Should I Have?

- It is generally good practice to keep your pie charts **simple** for visual clarity.
- **Exploding** can help **highlight** slice of interest but use this carefully as such gaps can distorts part-to-whole comparison[2]. Annotating the slices will help in this case.
- **Ordering slices** (e.g. largest to smallest slice) usually aid easier understand, especially if categories do not have inherent ordering to comply with.

[2] Mike Y, A Complete Guide to Pie Charts (2019)

Ethical Visualisation

Be Visually Truthful

- You are participating in a week-long election for the position of President of the Students' Union. You decided to put in some last-minute effort to get people to vote for you using this visual pitch...

3D pie chart visual to canvas for last-minute votes

This is the true picture that a plain pie chart would show.

Remember, people are not fools and they are voting for your honesty as well.

Pie Charts

Can I Have a Doughnut Instead?

- A doughnut (donut) chart is basically a pie chart with a central circle removed and there are **no significant differences** in readability between the two types of plots.
- The central area of a donut chart can be used for **additional information**.
- It can also show more than one set of data to allow **comparisons** over two different compositions[2].

[2] Mike Y, A Complete Guide to Pie Charts (2019)

Stacked Bar Charts

Visualising Multiple Compositions

- Stacked bar charts show relative decomposition of each primary bar (e.g. Persons) based on the levels of a **second categorical variable** (e.g. Expense type)[2].
- The stacked sub-bars of each secondary variable is **coloured** based on the category of the parts that make up the whole bar.
- If the primary bars are drawn across a time series (e.g. years), they express how the composition is changing over time.

[2] Mike Y, A Complete Guide to Stacked Bar Charts (2019)

Stacked Bar Charts

Ordering and Colouring the Bars

- **Ordering** stacked bar charts can help make the chart **easier to comprehend**.
- The rule of thumb is to order the bars from largest to smallest unless there is an intrinsic order in the primary bar (e.g. this is a time-varying stacked bar chart).
- The choice of colour palette to assign to each categorical level should match the variable type: a qualitative palette for purely categorical variables, and sequential, or diverging for variables with a meaningful order[2].

[2] Mike Y, A Complete Guide to Stacked Bar Charts (2019)

Stacked Bar Charts

Visualising Percentages

- In the percentage stacked bar chart, each primary bar is scaled to the **same height** with each **sub-bar** encoding its **percentage** contribution to the primary whole.
- A uniform height makes it easier to visually compare the **percentage contributions** of each sub-bar category to the whole. But this removes the ability to visualise the actual values of each of the sub-bar category[3].

[2] Mike Y, A Complete Guide to Stacked Bar Charts (2019)

Stacked Area Charts

→ do Comparison accross different plot

Visualising Composition Trends

- A **stacked area chart** can be considered if the x-axis is **interval** or **ratio** scale.
- It is created by plotting lines (and filling its area below) one at a time. The height of the most recently plotted line serves as a **moving baseline** for the next line plot[4].
- Height of the topmost line encodes the total sum across all composition categories.

[4] Mike Y, A Complete Guide to Stacked Area Charts (2019)

Stacked Area Charts

To Stack or Not to Stack

- Stacked area chart are useful for tracking the total value and analysing the varying breakdown of the parts making up this total[4].
- Only the exact values of the overall **total** and **bottommost** categories are easy to gauge as intermediate categories are plotted against varying baselines.
- If exact values are required for all categories, consider using **multiple line charts**.

[4] Mike Y, A Complete Guide to Stacked Area Charts (2019)

Think and Apply

Singapore's Virus Situation since 29 Apr 2021

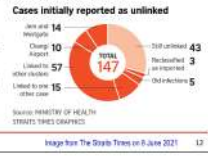
- Why do you think both line and bar charts are used?
- Would a stack area chart be better than using three line charts?
- What story do you think ST was trying to communicate with the line marker at May 16?

Image from The Straits Times on 8 June 2021

Think and Apply

Singapore's Virus Situation since 29 Apr 2021

- Why was a donut chart used to visualise the unlinked cases?
- Should we replace the donut chart with a pie chart instead?
- Should each slice of the donut be encoded with a different colour? Is there a reason why ST Graphics used the colour scheme?



Source: MINISTRY OF HEALTH
STRAITS TIMES GRAPHICS

Image from The Straits Times on 8 June 2021

Summary

Composition Plots

- Composition plots help us visualise the **individual parts** comprising the **whole**.
- Pie and donut charts are effective for visualising **small number** of individual parts, especially if the **variations** in the composition is **distinct** and **precision** in reading the values is not important.
- Stacked bar charts are effective in visualising **multiple compositions**.
  + choice of color
- Stacked area chart can be used to visualise **composition trend** changes and when **precise** values at every interval is not critical to the **analysis**.
  → comparison
  → important category put at bottom

References for Composition Plots

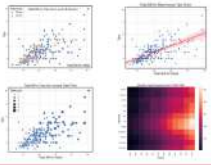
[1] How to Choose Charts to Show Data Composition (2016)
[2] Mike Y, A Complete Guide to Pie Charts (2019)
[3] Mike Y, A Complete Guide to Stacked Bar Charts (2019)
[4] Mike Y, A Complete Guide to Stacked Area Charts (2019)

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

Chapter 5.3 – Visualising Relationships in Data

Contents

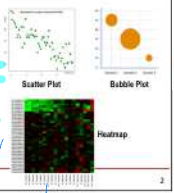
- Basic Relation Plots
- Scatter Pots
- Bubble Plots
- Heatmaps



Basic Relation Plots

Seeing Relationship Among Variables

- Relation plots help show relationships or patterns among variables.
- **Scatter and Bubble plots** – Scatter plots are useful to visualise correlation between two variables for one or more groups. The value of an additional third variable can be encoded by the dot size in a bubble plot.
- **Heatmaps** – When one or more of these variables are non-continuous or non-numeric, a heatmap can be used. Relationships between categories are visualised through a grid of values encoded in different colours.

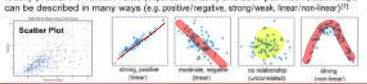


discrete values

Scatter Plots

Seeing Relationship in the Dots

- Scatter plots use the 2-D position of dots to represent values of two different numeric variables. The collection of these scattered dots can reveal **correlational relationships** or **patterns** between these two variables.
- Given a horizontal value (independent variable), a highly **correlated** relationship can readily predict the vertical value (dependent variable) and such pairwise relationships can be described in many ways (e.g. positive/negative, strong/weak, linear/non-linear)[1].

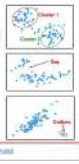


[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot

Scatter Plots

Seeing Groupings in the Dots

- Scatter plots can show interesting grouping patterns of dots[1].
- **Clusters** – are distinctive groupings in the dots forming various discernible shapes. Some clusters are circular, while others may be elongated.
- **Empty regions or gaps** – are areas that have a distinctively lack of dots. These unexpected gaps can be informative as they reveal incompatible value pairs between the two variables of interest.
- **Outliers** – are dots that are far off and do not appear to belong to any cluster. These data points could be erroneously recorded or they could be rarely occurring events of special interest.



[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot

Scatter Plots

Seeing Trends in the Dots

- Trend lines can be drawn to show the mathematically **best fit** to the data.
- Seaborn's `regplot()`[2] method is used to fit a **linear regression model** to the data. The shaded area about the line is the **confidence interval**.

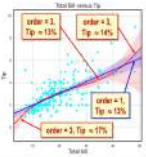
- Lines of higher polynomial order can be used if such fit makes sense to the data.

[2] Seaborn, Plot data and a linear regression model fit., https://seaborn.pydata.org/generated/seaborn.regplot.html

Scatter Plots

When a Polynomial May Not Fit

- The polynomial regression model with order=1 provides a constant estimation of a 13% tip irrespective of the cost of the meal.
- A polynomial model of order=3 predicts a varying tip, which about 17% with meals < $10, drops to about 13% when meal cost starts increasing and then goes up again to about 14% as meals gets even more expensive.
- It seems unlikely that a single polynomial can **model all the data** (irrespective of increasing total bill), especially if the data is noisy and there may exist outliers at the two extremities.

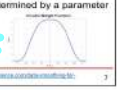


Data from "tips.csv"

Scatter Plots

Using LOWESS

- LOcally Weighted Scatterplot Smoothing (LOWESS) was developed by William Cleveland in 1979 at AT&T Bell Labs.
- Lowess is a **non-parametric** technique that makes no assumption about the model that should fit the relationship between the independent and dependent variable[3].
- Lowess is computed **piecewise** (i.e. point by point) as a series of **least squares linear regression** about the point using a nearest-neighbour subset of the data.
- This subset of neighbouring points used at each point is determined by a parameter (e.g. `frac` in statsmodels's version of `lowess`) that creates a window size that is a faction of all points in the dataset.
- In each regression, points within the nearest-neighbours window are **weighted** by their proximity, usually using the tri-cubic weight function given by $w(x)=(1-|d|^3)^3$.



[3] J.B. Mislinski, Data Smoothing in Data Science Visualization (2021), https://towardsdatascience.com/data-smoothing-in-data-science-visualization-fb4a4f0ddc4c#ccc0-1-BE7F3S0S0C0

Scatter Plots

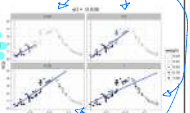
How Does LOWESS Work?

- The `lowess` algorithms moves from **left to right** doing **piecewise** concatenation of individual least-squares linear regression using a **weighted** fraction of all nearest data points to the current point being computed.
- The Statsmodels Python module provides a useful implementation of `lowess`[4]. The size of the neighbourhood is given by `frac`, a float value between 0 to 1.
- It can be imported in Spyder with the following:

```
from statsmodels.nonparametric.smoothers_lowess import lowess
```

small neighbourhood

large



Animation by David Robinson (2017) http://varianceexplained.org/files/loess.html

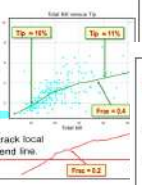
[4] Statsmodels Python library reference for its lowess method, https://www.statsmodels.org/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html

Scatter Plots

Appying LOWESS

- Lowess with `frac`=0.4 was applied to the tips dataset.

```
ls = lowess(Tips, TotalBill, frac = 0.4)
ax = sns.lineplot(x=ls[:,0], y=ls[:,1])
```

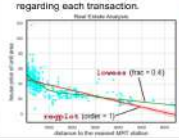


- The results show that a tip of about 16% is given for meals < $20 but as the cost of the meal increases, the tip starts to drop to about 11%.
- By using a sizable portion (but not all) of the dataset, lowess reduces the **influence of outliers** but can still track the **changing trend** over the entire total bill.
- Reducing the fraction (`frac`) of data points used will track local variations better but reduces the smoothness of the trend line.

Think and Apply

What Affects Property Prices?

- A dataset of 414 property transactions that list various property-related data regarding each transaction.
  - Unit area price of the property
  - Age of the property
  - Distance from a MRT station
  - Number of convenient stores nearby



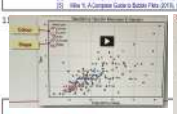
- Better property price nearer MRT station?
- Is older cheaper?
- Does more convenience have a price?
- What really makes the difference?

Bubble Plots

Handling Additional Data Dimensions

- Additional **categorical** variables (e.g. countries, gender) can be added to the scatter plot by using points with **different nominal visual attributes** (e.g. colour or shapes)[1].
- However, if the relationship of a **third** additional variable to be depicted is **ordinal** (i.e. ordered) or **quantitative** (i.e. numeric), then the **size the dots** can be used to encode this quantitative value in a **Bubble Plot**[5].

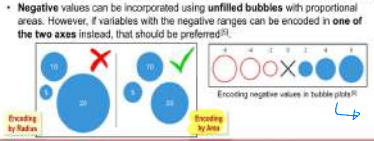
[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot
[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/



Bubble Plots

Encoding Quantitative Values to Bubble Size

- For **proportional** visual comparison, the bubble's **area** (not radius) is scaled to the value[5]. A dot with value $2n$, is only $(\sqrt{2})$=1.41 times the radius of a dot of value n.
- **Negative** values can be incorporated using **unfilled bubbles** with proportional areas. However, if variables with the negative ranges can be encoded in **one of the two axes** instead, that should be preferred[5].



better?

[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/

Bubble Plots

Length versus Area

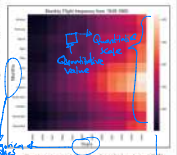
- When the **differences** in magnitudes between categories are **large**, **area** encoding make better comparative visuals then value encoding using length[6].

[6] Nathan Yau, Data Points, Visualization that Means Something, Wiley (2013)

Heatmaps

Seeing Relationships using Grids

- Heatmap depict values for a main variable of interest in **coloured rectangles** across two axis variables[7].
- Any patterns in the values for one or both of these axes can be observed by how the cell **colour changes** across each axis.
- These two axis variables are either **categorical** or if quantitative, they are binned into **discrete ranges** to form the desired number of cells.



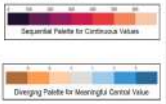
[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/
[8] Seaborn, Plot rectangular data as a color-encoded matrix, https://seaborn.pydata.org/generated/seaborn.heatmap.html

put colour attribute legend

Heatmaps

Choosing the Colours

- Since colour is a key feature of heatmaps, an **appropriate colour palette** should be chosen to match the nature of the primary variable of interest[7].
- A **sequential** colour ramp between value and colour is often used. Lighter colours map to larger values & darker colours to smaller values (or vice versa).



- If the values have a meaningful zero point, a **diverging** colour palette would add better visual clarity.

- The colour distribution **legend** must be included in the heatmap to allow associated values of the colours to be interpreted.

[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/

Summary

Relation Plots

- Relation plots help us visualise **relationships** or **patterns** in the data.
- **Scatter plots** uses 2D position of dots to visualise **spatial relationships** (e.g. clustering, strong correlation, outliers, etc.) among data with **two** interval or ratio scale variables.
- **Linear regression** and **LOWESS** allow us to extract best fit trend lines within these scattering of dots.
- **Bubble plots** extend scatter plots ability to handle one more **quantitative dimension**, but discrimination of area variations is limited.
- **Heatmaps** allow us to visualised relationships over a **coloured grid**.

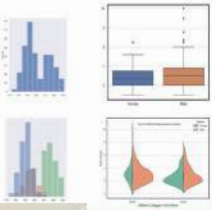
References for Relation Plots

[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/
[2] Seaborn, Plot data and a linear regression model fit., https://seaborn.pydata.org/generated/seaborn.regplot.html
[3] J.B. Mislinski, Data Smoothing in Data Science Visualization (2021), https://towardsdatascience.com/data-smoothing-in-data-science-visualization-fb4a4f0ddc4c#ccc0-1-BE7F3S0S0C0
[4] Statsmodels Python library reference for its lowess method, https://www.statsmodels.org/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html
[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/
[6] Nathan Yau, Data Points, Visualization that Means Something, Wiley (2013)
[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/
[8] Seaborn, Plot rectangular data as a color-encoded matrix, https://seaborn.pydata.org/generated/seaborn.heatmap.html

Chapter 5.4 – Visualising Distribution in Data

Contents

- Basic Distribution Plots
- Histogram
- Box Plots
- Violin Plots



1

Basic Distribution Plots

Visualising Distribution in the Data

- Distribution plots gives a visual insight into how the data values are distributed within the dataset.

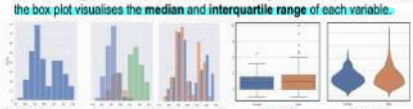
Question: Based of the scores, which test do you think w better at testing the level of understand

Answer: The score distribution that is more spread out.

Basic Distribution Plots

Visualising Distribution in the Data

- Distribution plots gives a visual insight into how the data **values are distributed** within the dataset.
- **Histograms** – are often used to visualise the distribution of a **single variable**. Multiple variables[1] are possible but may suffer from occlusion and clutter.
- **Box & Violin Plots** – are useful for exploring the distribution in **multiple variables** simultaneously. The violin plot visualises the **density curve** of the variable, whereas the box plot visualises the **median** and **interquartile range** of each variable.



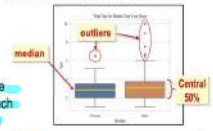
[1] Visualizing distributions of data (2019). https://seaborn.pydata.org/tutorial/distributions.html

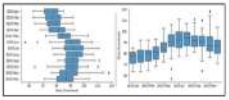
2

Box Plots

Visualising Statistical Properties of Multiple Variables

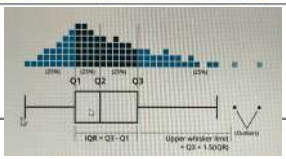
- Box plots (whisker plot) uses boxes and lines to depict the distributions of one or more numeric variables.
- The **box limits** show the range of the **central 50%** of the data, with a central line marking its **median**. Lines extend from each box to capture the **range** of the remaining data, with dots placed past the line edges to indicate outliers.
- Box plots allow multiple variables to be visualised either vertically or horizontally[3].





[3] Mike Yi, A Complete Guide to Box Plot (2019). https://chartio.com/learn/charts/box-plot-complete-guide/

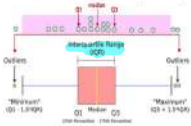
5



Box Plots

Constructing the Box and Whiskers

- Box plots display data variable's quartiles (values that divide the dataset into equal fourths).
- The 1st quartile (**Q1**) is greater than 25% of the data. The 2nd quartile (**Q2**), the **median**, sits in the middle, dividing the data in half. The 3rd quartile (**Q3**) is larger than 75% of the data. The box boundaries and its centre line mark the locations of Q1, Q2 and Q3[4].



- The left and right whiskers are **1.5 times** the interquartile range (IQR) from Q1 and Q3 respectively, but not exceeding data points at the extremities. All points outside these limits are considered **outliers** and are plotted individually horizontally[4].

[4] M. Galarnyk, Understanding Boxplots (2018). https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

6

Think and Apply

What Affects Academic Performance?

- Analyse the math, reading and writing scores of 1000 students with different backgrounds, gender and preparedness. Looking at various score distributions, answer these questions:
  - What subjects do the male students do better in?
  - Which subject was the most difficult?
  - Did taking the prep course help improve scores?
  - Do ethnicity and parent's educational background influence performance?

8

Summary

Distribution Plots

- Distribution plots allow us to visualise the way the data values in the dataset is **distributed** across the permissible range.
- The **histogram** is a bar chart that shows the distribution of data values. Its shape can **reveal much more** (e.g. skewed, bimodal, etc) than what typical statistical measures like mean and variance can inform.
- **Box plots** is useful for visualising the **statistical** properties of **multiple variables** simultaneously.
- **Violin plots** allow the **kernel density curves** of multiple variables to be visualised and compared in a **pairwise** manner.

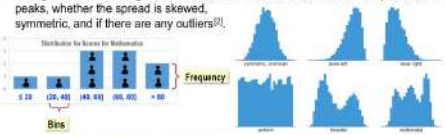
10

25%

Histograms

Visualising Data Distribution

- Histograms plot the distribution of a variable's numeric values as a series of bars. Each bar covers a range of numeric values (**bin**) and its **height** encodes the **frequency** of data points that have values within the corresponding bin.
- Histograms reveals the general distributional features of variables, their positions of peaks, whether the spread is skewed, symmetric, and if there are any outliers[2].



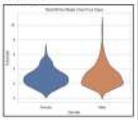
[2] Mike Yi, A Complete Guide to Histograms (2019). https://chartio.com/learn/charts/histogram-complete-guide/

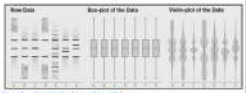
3

Violin Plots

Visualising Kernel Density Curves of Multiple Variables

- Violin plots visualise distributions of numeric data values using kernel density curves, whose widths encode the approximate frequency of data points in each region[5].
- Distribution comparison between multiple groups can be done as the peaks, valleys, and tails in of their respective density curves can be contrasted by the side-by-side plots.
- Datasets can be modified in a way that the quartiles do not change, but the shape of the distribution differs dramatically[6]. In such cases, violin plots give **more insights** into the **data distribution** than box plots.





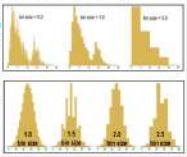
[5] Mike Yi, A Complete Guide to Violin Plots (2019). https://chartio.com/learn/charts/violin-plot-complete-guide/
[6] J. Matejka, G. Fitzmaurice, Same Stats, Different Graphs (CHI'17). https://www.autodesk.com/research/publications/same-stats-different-graphs

7

Histograms

Selecting Number of Bins and Boundaries

- The bin count and their boundaries for tallying data points must be chosen properly as it can have an effect of the **shape** of the plot and how it is interpreted[2].
- **Too many** bins may result in a **noisy** plot that is difficult to discern general data distribution. **Too few** bins may **reduce details** needed to discern patterns in the data (e.g. a double peak distribution).
- **Bin sizes** should match the **type of values** the data can have. Fractional bin sizes (e.g. 1.5) may not be suitable for integer variables as not all bins can take the same numbers of valid data values, leading to "bumpy" looking plots[2].



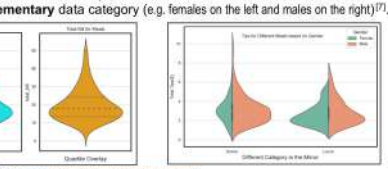
[2] Mike Yi, A Complete Guide to Histograms (2019). https://chartio.com/learn/charts/histogram-complete-guide/

4

Violin Plots

Getting the Best of Two Worlds

- Violin plots have additional flexibility compared to box plots. They can also be overlayed with additional information such as the **box plot** or **quartile lines**.
- Each **mirrored half** of the violin plot can be used to encode the distribution of another **complementary** data category (e.g. females on the left and males on the right)[7].



[7] Seaborn, Violin Plots. https://seaborn.pydata.org/generated/seaborn.violinplot.html

8

References for Distribution Plots

[1] Visualizing distributions of data (2019). https://seaborn.pydata.org/tutorial/distributions.html

[2] Mike Yi, A Complete Guide to Histograms (2019). https://chartio.com/learn/charts/histogram-complete-guide/

[3] Mike Yi, A Complete Guide to Box Plot (2019). https://chartio.com/learn/charts/box-plot-complete-guide/

[4] M. Galarnyk, Understanding Boxplots (2018). https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

[5] Mike Yi, A Complete Guide to Violin Plots (2019). https://chartio.com/learn/charts/violin-plot-complete-guide/

[6] J. Matejka, G. Fitzmaurice, Same Stats, Different Graphs (CHI'17). https://www.autodesk.com/research/publications/same-stats-different-graphs

[7] Seaborn, Violin Plots. https://seaborn.pydata.org/generated/seaborn.violinplot.html

Chapter 6 – Visual Perception

Contents

- Human Visual Perception
- Gestalt Principles in Data Visualisation
- Colour Perception

!

Chapter 6.1 – Human Visual Perception

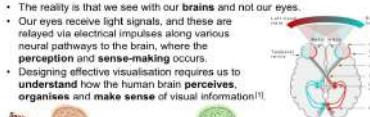
Contents

- The Visual Brain
- Estimating Magnitude
- Estimating Change
- Estimating Rate of Change
- Pre-attentive Visual Processing
- Multiple Visual Attributes
- Visual Background Perception

The Visual Brain

The Picture Tells the Story

- The reality is that we see with our **brains** and not our eyes.
- Our eyes receive light signals, and these are relayed via electrical impulses along various neural pathways to the brain, where the **perception** and **sense-making** occurs.
- Designing effective visualisation requires us to **understand** how the human brain **perceives**, **organises** and **make sense** of visual information[1].



The Visual Brain

The Memory Model

- Human visual processing is supported by **iconic memory (sensory memory)**, **short-term memory (working memory)** and **long-term memory**, each having different roles and characteristics[2].

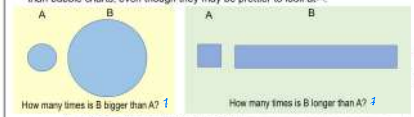
example that making 10 in 2 seen

see triangle

Estimating Magnitude

Big, Bigger, Biggest

- The human visual system is much better at **estimating magnitude** based on visual **length** than visual **area**.
- For this reason, bar charts are much better in presenting accurate visual information than bubble charts, even though they may be prettier to look at[3].

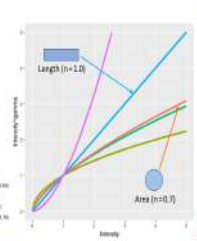


Estimating Magnitude

Stevens's Power Law

- Humans perceive different visual channels (e.g. length, area, brightness, etc) with **different** levels of **accuracy**[4].
- Our perceived sensation (***S***) of a physical intensity (***I***) is characterised by the psychophysical **power law** of **Stevens**, where the exponent (***n***) depends on the exact sensory modality of the measured ***I***.

$$S = I^n$$

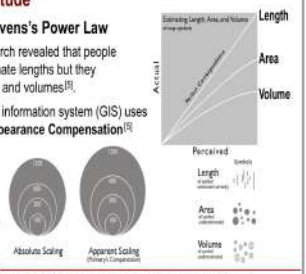


Estimating Magnitude

Compare colour chart

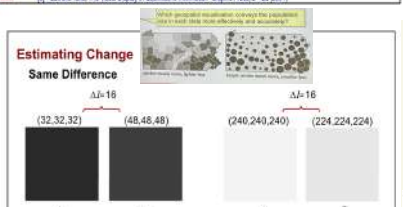
Implication of Stevens's Power Law

- Psychophysical research revealed that people tend to correctly estimate lengths but they **underestimate** areas and volumes[5].
- ArcGIS, a geographic information system (GIS) uses James Flannery's **Appearance Compensation**[5] method to scale up proportional circles (i.e. the larger the circle, the more the scaling).
- **Note:** Edward Tufte[6] only supports absolute scaling for faithfulness to the data.



Estimating Change

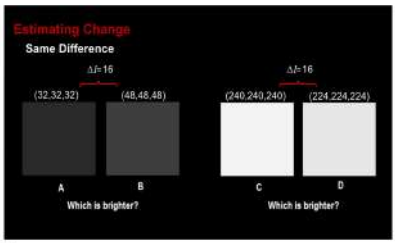
Same Difference



take note of issues of contrast

Estimating Change

Same Difference



Estimating Change

Same Difference

- The **Just Noticeable Difference** (JND) is the minimum amount by which stimulus intensity must be changed in order to produce a noticeable variation in sensory experience.
- Ernst Weber[7] observed that the size of the **difference threshold** appeared to be related to the initial **stimulus magnitude**.
- JND is govern Weber's Law and is given by

$$\frac{\Delta I}{I} = k$$

Addition to intensity Δ required for the change to be perceived → ΔI; Current intensity of stimulus → I; k ← Weber constant

Ernst Heinrich Weber (24 Jun 1795 – 26 Jan 1878)
German physician & early pioneer of Experimental Psychology

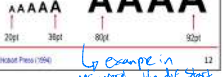
Estimating Change

Implication of Weber's Law

- According to Weber's Law, the higher the intensity (or length), the larger the disparity required for us to sense the change.
- Visualisation designs that takes such perceptual behaviour into account will facilitate more accurate visual comparison (e.g. framing long bar values to make their small difference more apparent)[8].
- Weber's law also applies to choosing **shape** or **font sizes**. As the shapes or fonts get larger, the absolute difference must be made larger to allow changes to be noticeable.

using an equal-length to see difference (compensating for Weber's law)

Each Arial font "A" increases by 4 points

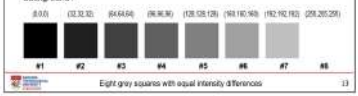


example in MS word, the diff start from 9 pt, but then become 8 pt diff in font size select

Think and Apply

Encoding Ordinal Data Using Grey Values

- Which two grey squares will be most difficult to tell apart with a white background?
- What if the background is black? What background colour can maximise the discriminability of all these grey levels if their values cannot be changed?
- How would you change the grey values to improve their discriminability in white background?



Eight grey squares with equal intensity differences

Estimating Change

Using Visual Reference Grids

- Weber's Law explains why visual reference grids enhance pattern perception.
- Without the grid, estimation of lengths with small percentage differences is difficult[9].
- The grids shorten the base lengths that are being compared, making it easier to compare highs, lows, and steady state behaviour[9].
- Graphs can be compared by superimposing them, but this only works for limited numbers of plots before clutter and line differentiation becomes problematic.

Estimating Change

Proper Layering of Visual Reference Grids

- If visual reference grids are not layered correctly, they can be distracting and make the actual data difficult to visualise.

Competing signal & background in two different electrocardiogram trace lines. The prominent gridlines & poor contrast makes the trace of the left more difficult to read than the one on the right.

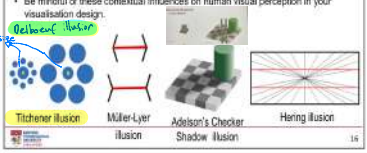
- Heer and Bostock[8] crowdsourced experiments on an acceptable luminance contrast settings for visual reference elements such as gridlines showed a safe **Alpha** setting of about **20%** (Alpha 0% = Total transparent, 100% = Opaque).

Estimating Magnitude

Context Matters

- Our ability to estimate the magnitude of visual attributes (e.g. length, size, colour, parallelism, etc) can be influenced by the context in which it is visualised.
- Be mindful of these contextual influences on human visual perception in your visualisation design.

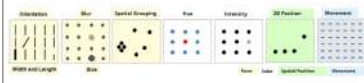
Delboeuf illusion



Preattentive Visual Processing

Standing Out

- Preattentive processing occurs at the **early stage** of visual perception and is tuned to **rapidly detect** a specific set of visual attributes at a **sub-conscious** level[10].
- The **sequential attentive** processing used to find a specific visual target is **slower**.



Some examples of the preattentive attributes of human visual perception

!! use in Video

Preattentive Visual Processing

Exploiting Preattentive Processing

- Preattentive visual attributes can be used to **highlight** (i.e. make it pop out) particular data points of interest by making them **distinct** on a particular feature channel (e.g. coloured when the rest are grey scaled) or made more distinct by an **appropriate** redundant combination of **multiple preattentive attributes** (e.g. colour and size)[10].

Preattentive Visual Processing

Conjunction Search

- Data points can be encoded with multiple visual attributes. A visual search for an object with more than one attribute (e.g. green ellipse) is a **conjunction search**. Conjunction searches are generally **not preattentive**[11].
- However, there are conjunction of some attributes that support preattentive search (e.g. position and colour or luminance polarity and shape)[11].

Search is slow with the conjunction of shape and colour | Search is fast with conjunction of position & colour (left green object) | Luminance polarity with targets (white circles) lighter & darker than a grey backdrop supports preattentive conjunction | The same colour and shape encoding has no preattentive conjunction search property

[11] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012).

Gestalt Psychology

The Whole is Other than the Sum of Parts

- Gestalt psychology describes various principles or laws that govern how the human visual system forms **patterns** and **visual groupings** to simplify the interpretation of visual stimuli received.
- These Gestalt principles explain the **pattern-seeking** behaviour of human visual perception and should therefore be used in visualisation to help **communicate visual information** more effectively.
- Some Gestalt principles explored include:
  - Proximity
  - Connectedness
  - Figure & Ground
  - Similarity
  - Continuity
  - Simplicity
  - Symmetry

Similarity

Exploiting Similarity

- Elements that share the **same visual attributes** will be perceived as being **related**. Consider carefully if the bar charts should have different hues or a uniform hue.
- Dissimilarity can be used to do the reverse and **highlight** a particular **exception** or **focal point** in the visualisation by giving it a distinctly different visual attribute.

Ordinal categories (e.g. month in a year) should have uniform colour as ranking has increases cognitive load and disrupt clarity of changing trend[2] | Wallet month in 2020 made to stand out in a different colour[4]

[2] William Craig, How to Make Data Visualization Better with Gestalt Laws.

Think and Apply

Telling It Like It Is – Effective Visual Story

- The table shows the average levels of an imaginary hormone Vitalis in a population based on age group, gender and BMI. The story you want to tell is that the only group with increasing levels of Vitalis as they age are females with BMI < 25.
- How would you design a chart to make this **story stand out** while providing all the information shown in the table.

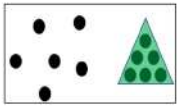
	Levels of the hormone Vitalis			
	Males		Females	
Body Mass Index (BMI)	Under 60 years	60 years or over	Under 60 years	60 years or over
Under 25	255	230	380	550
25 or over	440	325	720	500

TUT

Proximity

Closer is Together

- Principle of **proximity** states that visual elements **close to** each other are perceived as **belonging** to the same group.



Proximity of shapes creates a larger unit entity

Think and Apply

Improving Visualisation using Gestalt Principles

- How can applying Gestalt principles improve the visualisation of the changing company profits over the years and the changing trend within each department?

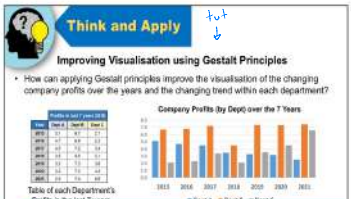
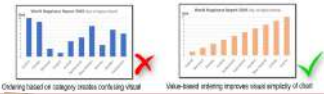


Table of each Department's Profits in the last 7 years | Visualisation of Company's Changing Profits

Simplicity

Simplicity is the Ultimate Sophistication

- Principle of simplicity states that visual elements are always perceived in the **easiest way possible**.
- **Ordering for simplicity** - visual information (e.g. values in bar charts) should be sorted and ordered (if possible) to increase its visual simplicity, thus easing interpretation.



Ordering based on category creates confusing visual | Value-based ordering improves visual simplicity of chart

Summary

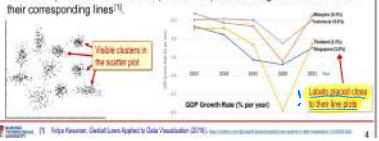
Human Visual Perception

- Human visual perception is **complex** because visual interpretation takes place in the **brain** and is **influenced** by many **factors** (e.g. our memories) besides the visual stimulus entering our eyes.
- Characteristics of the different visual channels based on Stevens's power law and Weber's law should influence the way we design **visual encoding** for data visualisation.
- Understanding **preattentive** visual processing can help us exploit its characteristics in designing visuals that can capture people's **attention** and avoid designing visuals that are **not effective** in communicating **useful meaning** in the data patterns.

Proximity

Closer is Together

- Principle of **proximity** states that visual elements **close to** each other are perceived as **belonging** to the same group.
- Clusters in scatter plots are identified based on the proximity of the dots.
- Labels are placed close to their respective line plots to strengthen association to their corresponding lines[1].



[1] Yelya Kasumar, Gestalt Laws Applied to Data Visualization (2016).

Simplicity

Data-Ink Ratio

- Edward Tufte proposed the concept of **data-ink ratio** for data visualisation[8].
- This ratio measures the proportion of **ink** that is used to present **actual data** compared to the **total amount of ink** (or pixels) used in the entire visual.
- The goal in good data visualisation is to **maximise** the data-ink ratio.

Changes to Excel charts over the years to maximise data-ink ratio

[8] Edward Tufte, The Visual Display of Quantitative Information. Graphic Press, 2nd Ed (2001).
[9] Carlos Moore, Step 10 Visualisation.

References for Visual Encoding

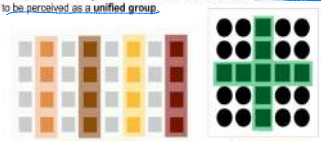
[1] Stephen Few, Data Visualization for Human Perception - https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception
[2] R.C. Atkinson, R.M. Shiffrin, Human memory: A proposed system and its control processes, In K. Spence (Ed.), The psychology of learning and motivation (Vol. 2). Oxford, England: Academic Press (1968).
[3] Steve Wexler, Why the f**k do we see so many bar charts? (2021). https://www.linkedin.com/pulse/why-%-do-we-see-so-many-bar-charts-steve-wexler/
[4] T. Munzner, Visualization Analysis & Design, CRC Press (2015).
[5] John Krygier, Perceptual Scaling of Map Symbols - https://making-maps.net/2017/06/28/perceptual-scaling-of-map-symbols/
[6] Edward Tufte, The Visual Display of Quantitative Information. Graphic Press, 2nd Ed (2001).
[7] Weber's Law - https://www.britannica.com/science/Webers-law
[8] W.S. Cleveland, The Elements of Graphing Data, Hobart Press (1994).
[9] J. Heer, M. Bostock, Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design (CHI 2010) - http://www.perceptualedge.com/articles/ieee/visual_perception_mechanical_turk.pdf
[10] Stephen Few, Information Dashboard Design, O'Reilly Media (2006).
[11] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012).

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

Similarity

Birds of a Feather Flock Together

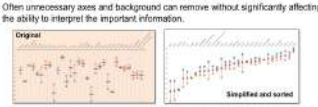
- Principle of **similarity** states that visual elements that are **similar** to each other tend to be perceived as a **unified group**.



Simplicity

Data-Ink Ratio Maximisation

- Data-ink ratio can be maximised by considering what can be **removed** from the chart while **keeping the information** that needs to be communicated.
- Often unnecessary axes and background can remove without significantly affecting the ability to interpret the important information.



An example[11] of improving visual simplicity by maximising data-ink ratio and ordering the data

[11] Carlos Moore, Step 10 Visualisation.

Chapter 6.2 – Gestalt Principles in Visualisation

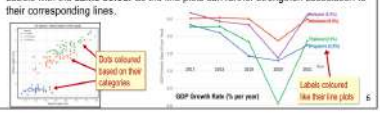
Contents

- Gestalt Psychology
- Proximity
- Similarity
- Simplicity
- Connectedness
- Enclosure
- Symmetry
- Figure-Ground

Similarity

Birds of a Feather Flock Together

- Principle of **similarity** states that visual elements that are **similar** to each other tend to be perceived as a **unified group**.
- Similarity-based perceptual grouping can be influenced by various visual features, including shape, colour, orientation, size or even motion.
- **Colour** can be used in scatter plots to identified data belonging to similar categories.
- Labels with the **same colour** as the line plots can further strengthen association to their corresponding lines.



Think and Apply

Making it Easy for Others

- How would you redesign these charts to help the readers answer the quiz questions more easily?

Connectedness

Close Links Bond

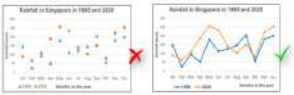
- Principle of connectedness states that visual elements that are **connected** to each other are perceived as a **single unit**.

13

Connectedness

Close Links Bond

- Principle of connectedness states that visual elements that are **connected** to each other are perceived as a **single unit**.
- Related data points should be connected with appropriate lines to allow fast visual recognition of their grouping and allow changing trends to be readily perceived[8].



Lines in graphs create clear visual connection between related points that even similar colour and shape attributes cannot evoke.

[8] B. Wong, Gestalt Principles (Part 1), Nature Methods, Vol 7, pp. 863 (2010); https://www.nature.com/articles/nmeth1113-863

Gestalt Principles in Data Visualisation

Continuation .....

- The Gestalt principles explain the **pattern-seeking** behaviour of human visual perception and it perceive visual groupings to simplify the interpretation of visual stimuli.
- Some Gestalt principles explored last week include:

• Proximity	• Connectedness	• Symmetry
• Similarity	• Continuity	• Figure & Ground
• Simplicity	• Enclosure	

15

Enclosure

Feeling Boxed In

- Principle of enclosure states that visual elements are seen as belonging to a group if they are **enclosed** in a way that creates a **boundary** or **border around** them .

16

Enclosure

Feeling Boxed In

- Principle of enclosure states that visual elements are seen as belonging to a group if they are **enclosed** in a way that creates a **boundary** or **border around** them .

- Enclosure can be useful in organising a complex interface, like on the instrument dashboard of a 747 plane[7].
- Square and rectangular frames group similar categories of controls and display together.

[7] A. Rutledge, Gestalt Principles of Perception - 3- Proximity, Uniform Connectedness, and Good Continuation (2009) - https://www.aru-rutledge.com/gestalt-principles-3.html

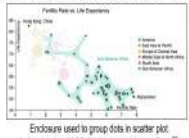
17

Enclosure

Feeling Boxed In

- Principle of enclosure states that visual elements are seen as belonging to a group if they are **enclosed** in a way that creates a **boundary** or **border around** them .
- Enclosure can be used to **highlight** and **group** specific data in the plot or chart.

Transparent enclosure used to highlight data points during the inter-monsoon period



Enclosure used to group dots in scatter plot belonging to a highlighted category. Image from [8]

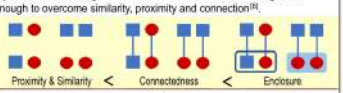
[8] DataAscent, Visualisation Examples - https://dataascent.github.io/

18

Enclosure

Dominant Gestalt Principles

- The relative strength each Gestalt principle exerts on perceptual grouping is not the same and there are many situations where different principles will compete to provide a different sense of visual grouping.
- Connectedness through the lines create a much stronger sense of grouping than either proximity or similarity[8].
- Grouping by enclosure resulting in elements bounded in a common region is powerful enough to overcome similarity, proximity and connection[8].



[8] B. Wong, Gestalt Principles (Part 1), Nature Methods, Vol 7, pp. 863 (2010); https://www.nature.com/articles/nmeth1113-863

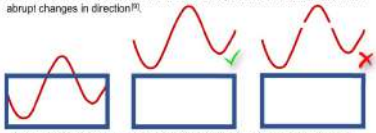
19

Continuity

we see it as 1 line rather than changes

It Cannot Change So Quickly

- Principle of continuity states that we are more likely to construct visual entities out of visual elements that are smooth and continuous, rather than ones that contain abrupt changes in direction[8].



Our minds tend to perceive a line as continuous and continuing in its established direction.

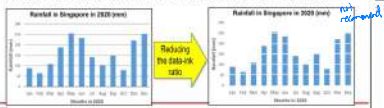
[8] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

20

Continuity

It Cannot Change So Quickly

- Principle of continuity states that we are more likely to construct visual entities out of visual elements that are smooth and continuous, rather than ones that contain abrupt changes in direction[8].
- This sense of continuity can be exploited to reduce the data-ink ratio of bar charts by removing the axes and plotting the axes tick lines within the bars themselves[3].



not recommend

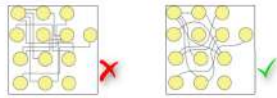
[3] Edward Tufte, The Visual Display of Quantitative Information: Graphic Press, 2nd Ed (2001)

21

Continuity

We Follow Smooth Lines Better

- The principle of good continuity can be applied to the problem of drawing diagrams consisting of networks of **nodes** and the **links** between them.
- The human visual system finds it easier to identify the sources and destinations of **connecting lines** if they are **smooth** and **continuous**[8].



Where possible, smooth lines should be used to link nodes in a network graph

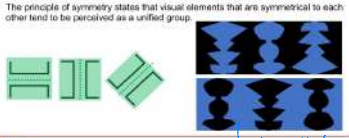
[8] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

22

Symmetry

Seeing Both Sides

- The principle of symmetry states that visual elements that are symmetrical to each other tend to be perceived as a unified group.



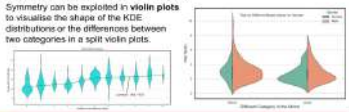
we see shape and background

23

Symmetry

Seeing Both Sides

- The principle of symmetry states that visual elements that are symmetrical to each other tend to be perceived as a unified group.
- Symmetry can be exploited in **violin plots** to visualise the shape of the KDE distributions or the differences between two categories in a split violin plots.

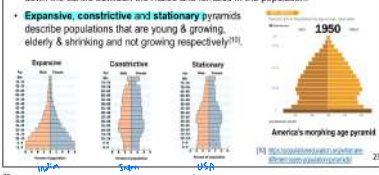


24

Symmetry

Population Pyramids

- These are graphs that show the distribution of ages across a population divided down the centre between the males and females in the population.
- **Expansive, constrictive and stationary** pyramids describe populations that are young & growing, elderly & shrinking and not growing respectively[10].



India Expansive
China Constrictive
USA Stationary

[10] https://population.sciencehistory.org/what-are-the-main-types-population-pyramids

25

Figure-Ground

Standing Out from the Crowd

- The figure-ground principle states that people instinctively perceive objects as either being in the **foreground** or the **background**. They either **stand out** prominently in the front (the figure) or **recede** into the back (the ground)

Image from Wikipedia

Image from Cinema Underworld poster series by Simon C. Page

Kiss of the Sea by Octavio Ocampo
https://www.saatchiart.com/en/artwork/Paintings-Contemporary-Oil-Kiss-of-the-Sea/1000000-1000000/1000000-1000000/1000000-1000000

Chapter 6.3 – Colour

Contents

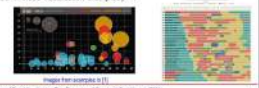
- Colour Perception
- Additive and Subtractive Colour Systems
- Colour Models
- Data Legibility of Colour
- Inclusive Use of Colour
- Colour in Data Visualisation

© AP Gin Wool Boon (SCSE/NTU)

Colour Perception

Living Colours

- The ability to perceive colours has been crucial for the **survival** of early hunters and gathers as it helps them spot lurking dangers, look and identify edible food.
- The ecological role of colour perception in humans suggests that it can also play an important role in the design of effective information visualisation (see the effective use of colour in these visualisation examples)[1].



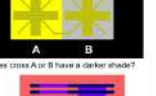
Images from examples in [1]

[1] Ross Crooks, The Power of Data Visualisation Plus Examples of Good and Bad Visuals (2021) - https://blog.hubspot.com/marketing/data-visualization-examples

Colour Perception

What is Colour?

- The perception of **colour** in humans is due to the stimulation of **photoreceptor** cells (in particular cone cells) by electromagnetic radiation in the visible spectrum, which can be either **reflected** of an object surface or **emitted** from a light source.
- However, the colour we eventually perceive is created in our **brain**. For example, it is able to maintain the colour of an object in a relatively stable manner despite changes in illumination due to shadows[2].
- It is also for this reason the visual attribute of colour is **complex** and its perception can be influenced by numerous factors such as the background, other adjacent colours, etc.



Photoreceptor Cells in the Retina. Image from American Academy of Ophthalmology

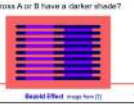
[2] Scientific American, Illusory Color & the Brain (2006) - https://www.scientificamerican.com/article/illusory-color-and-the-brain/

Colour Perception

Colour Appearance

- The **colour contrast** of the **background** can change the way we perceived the actual colour of an object.
- The **Bezold effect** was first reported by Wilhelm von Bezold, a German professor of meteorology.
- Bezold discovered that a colour may appear different depending on its relation to **adjacent colours**. It happens when small areas of colour are interspersed.

Does cross A or B have a darker shade?



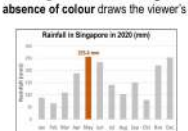
Bezold Effect. Image from [1]

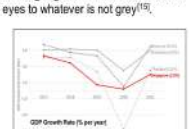
[1] L. Giovanni and H. Kettler, Is Seeing Really Believing? (2010) - https://slideplayer.com/slide/6128/#/slide/Is-Seeing-Really-Believing-Lucia-Giovanni-and-Hannah-Kettler-Group-Presentation-Bezold-Effect-Illusion-Color-Perception-Brain-Eye-Vision-Optical-Illusions-Psychology-Neuroscience-Science-Biology-Medicine-Health-Science-Technology-Engineering-Math-Science-Technology-Engineering-Math-Science-Technology-Engineering-Math

Colour in Data Visualization

Colour Allows us to Differentiate

- Colour can be used to **draw attention** to a data point of interest.
- An effective way to enhance this effect is to employ **grey** as the primary colour for visualising the data and a **strong colour** to **highlight** the data of interest. The **absence of colour** draws the viewer's eyes to whatever is not grey[15].





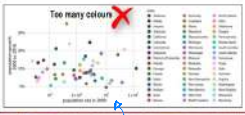
[15]. Conner Rothschild, Color in Data Visualization: Less How, More Why – https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-34f814a3c4d8

18

Colour in Data Visualization

Colour Allows us to Explore

- Colour can be used as a tool to **explore** how a variety of **different categories** of data points that are mapped to **different colours** related to each other.
- Effective exploration requires the **colour count** to be kept **small** so patterns in the data groups can emerge without having to constantly reference the colour legend[15].



Less How, More Why – https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-34f814a3c4d8

19

19

Colour in Data Visualization

Colour Allows us to Explore

- Another common use of colour for exploratory purposes is to show data progressing across a **sequential colour gradient** (e.g. low to high, bad to good, cold to warm)[15].
- Colour shading in **Choropleth maps** facilitate the exploration of **regional variation** in severity or intensity of a particular variable[15].

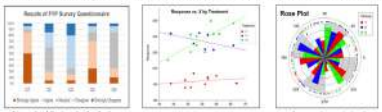
[15]. Conner Rothschild, Color in Data Visualization: Less How, More Why – https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-34f814a3c4d8

29

Colour in Data Visualization

Colour Allows us to Compare

- As colour allows us to easily distinguish between data points, it can be used in aiding **visual comparison** between the characteristics of **different data groups.**
- For effective comparison, the colours should be **distinct** and their numbers limited.



Comparing Likert scale distributions across different questions | Comparing slopes and intersections of data groups[16] | Comparative analysis of circular or angular data[16]

[16] NCSS 2021, Data Analysis & Graphics - https://www.ncss.com/software/ncss/ncss-plots-and-graphs/

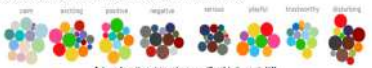
21

21

Colour in Data Visualization

Colour Allows us to Convey Emotions

- Colours have association with **emotions** and if used effectively, it can make the message embedded in the visualisation more **readily accessible**.
- Studies have shown that if the visualisation goal is to create a **calm** affect, **light, cool, pastel** colours should be use. If the goal is to create a **disturbing** affect, then **dark colours**, especially **reds**, are better[17].
- However, it is good to be mindful that the types of emotions associated with different colours do have **cultural** and **contextual** variations.



Colours from the palettes chosen as "Best" in the study [18]

[17]. L. Bartram et al., Affective Color in Visualization (CHI 2017) - https://research.tableau.com/sites/default/files/Affective%20Color%20CHI%202017.pdf

23

22

Colour in Data Visualization

Colour Allows us to Convey Emotions

- Contrasting combination of calm and dull colours like grey with strong disturbing colours like dark red tones can make viewers feel the story that is being told[18].
- In this example, the grey colour is used to map out the background information and context. The main message (i.e. prevalence of mental health disorder) is presented in alarming **red**.
- This **affective colour contrast** works as a visual punch and **amplifies** the critical facts[19].

[18] Olga Taubiks, Color in Data Visualization – The role of color in data storytelling - https://www.olgataubiks.com/data-for-excuses-visualizations

Visualisation by Arravendhanath D

23

23

Think and Apply

Can You Do Better?

- Comment of the chart type and colours used for the purpose of **comparing** teaching performances of Prof A, B and C?
- Can you do better? Redesign it.

Likert Rating	Prof A	Prof B	Prof C
Strongly Agree	55%	5%	69%
Agree	29%	23%	15%
Neutral	8%	40%	5%
Disagree	6%	26%	4%
Strongly Disagree	2%	6%	7%

Table of each teaching score percentages

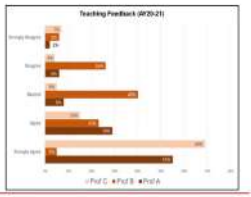


Chart Comparing Teaching Performances

24

24

Summary

Colour Perception

- Colour is a very **important visual attribute** in data visualisation and should be given **consideration** at the very start of your design.
- Colour must be used judiciously and **its purpose** in your visualisation design must be **clear** and **effective**.
- Choose an appropriate **colour palette** that fits the **characteristics** and **scale** of your data type; and if relevant, try to **blend** it with the visual **story** and the **emotional** theme you want to communicate.
- Make your visuals as **inclusive** as possible, bearing in mind that not everyone perceives colour in the same manner.

25

25

References for Colour Perception

[1]. Rosa Crooks, The Power of Data Visualization Plus Examples of Good and Bad Visuals (2021) - https://blog.hubspot.com/marketing/data-visualization-examples
[2]. Scientific American, Mystery Color in the Brain (2008) - https://www.scientificamerican.com/article/illusion-project-explains-the-brain-2008-06/
[3]. L. Giovanni and H. Kellen, In Seeing Reality (Believing?) (2015) - https://publishablecore.org/2015/01/science-and-technology-in-seeing-reality-believing/
[4]. Tim Mouw, Additive vs. Subtractive Color Models - https://www.pnb.com/blog/additive-subtractive-color-models
[5]. Andy Kirk, Data Visualisation, A Handbook for Data Driven Design, SAGE Publications (2016).
[6]. Understanding Hue, Saturation & Lightness (HSL) for Photo Retouching - https://purple11.com/blog/hue-saturation-lightness/
[7]. Axis Maps, Using Colors on Maps - https://www.axismaps.com/guide/using-colors-on-maps
[8]. Carmen Chan, Choosing Color Palettes in Display! - https://www.displayr.com/choosing-color-palettes-in-display/
[9]. Barry Ritholtz, Executive Pay by the Numbers - https://ritholtz.com/2011/07/executive-pay-by-the-numbers/
[10] Adam E McCann, Beatles Analysis - https://public.tableau.com/app/profile/adam.e.mccann/viz/BeatlesAnalysis/BeatlesAnalysis
[11] Claus O. Wilke, Fundamentals of Data Visualization, Color Pitfalls - https://clauswilke.com/dataviz/color-pitfalls.html
[12] Seaborn, General principles for using color in plots - https://seaborn.pydata.org/tutorial/color_palettes.html
[13] Robyn Collingbo, How to Design for Color Blindness (2017) - https://uxpublic.com/blog/how-to-design-for-color-blindness/
[14] Ian Tuckiso, Color blindness: how to design an accessible user interface (2018) - https://uxdesign.cc/color-blindness-in-user-interfaces-66c2733318f58
[15] Conner Rothschild, Color in Data Visualization: Less How, More Why - https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-34f814a3c4d8
[16] NCSS 2021, Data Analysis & Graphics - https://www.ncss.com/software/ncss/ncss-plots-and-graphs/
[17] L. Bartram et al., Affective Color in Visualization (CHI 2017) - https://research.tableau.com/sites/default/files/Affective%20Color%20CHI%202017.pdf
[18] Olga Taubiks, Color in Data Visualization – The role of color in data storytelling - https://www.olgataubiks.com/data-for-excuses-visualizations

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

26

Chapter 7 – Designing The Visualisation

Contents

- The Visualisation Design Process
- Know The Purpose
- Know Your Audience
- Know The Right Chart Type
- Selecting The Chart
- Psychological Principles of Effective Graphics

The Visualisation Design Process

Getting Started

- Data visualisation has both **aesthetic** and **scientific** design elements.
- The aesthetic elements are more **subjective** and can be strongly biased by the designer's **artistic preferences**.
- However, the **scientific** aspects of presenting numbers and measurements has a more **logical sequence of steps** that we can follow[1],[2], which are essentially:

Know the Purpose → Know the Audience → Know the Right Chart Type

[1] S.M. Kosslyn: Graph Design for the Eye and Mind, Oxford University Press (2006)
[2] Michael Schermann - A Reader on Data Visualisation Patterns - https://michaelsch.github.io/data_viz_reader/patterns.html

3

Know The Purpose

What Is It For?

- First, find out if the **purpose**[2] of the visualisation is **exploratory** or **explanatory**[1].
- **Exploratory** visualisation is designed help us explore and analyse the data to **derive insights** and **test hypotheses**.
- When we want to share our specific findings to an audience, **explanatory** visualisation is used to **communicate** the **specific story** embedded in the data.
- These two purposes are seldom effectively served by the same visualisation design.

Exploratory chart of Bitcoin prices from CryptoCompare. Taken from [4]

Explanatory chart of the stock market performance during 2016 US Presidential election night. Taken from [4]

[3] R. Kosara, Reflecting on the Design Criteria for Exploratory Visualization (2016) - http://eagereyes.org/criticism/theyfeelfine-CHI2010.html
[4] Data Viz 101: Explanatory vs. Exploratory Charts - https://www.andrewfb.com/data-viz-2-explanatory-vs-exploratory-charts/

4

Know Your Audience

Who Is It For?

- It is important to know the audience the visualisation is targeted for as factors like their **cultural** background, **expertise** level, etc can influence the design of an effective and meaningful visualisation[2].
- For example, for a Chinese audience, the colour red represents a positive event (e.g. increasing profits), while western audiences tend to view green as representing a positive trends.
- **Discipline variations** too can be significant. Visualisation charts the financial community is familiar with are usually different from that of scientific communities, and these differences can even vary between the various scientific & engineering disciplines.

[3] Image extracted from R. Allison, Colors represent different things, in different cultures (2017) - https://blogs.ieee.org/spectrum/author/rob-allison/

[5] Hubspot Marketing: 5 Steps to Create an Effective Data Visualization (2017) - https://www.highcharts.com/blog/news/5-steps-to-create-an-effective-data-visualisation/

5

Know The Right Chart Type

Which Should I Use?

- While it is possible to visualise the data using several different chart types, it is important to choose the type that can best communicate the story **clearly** and **accurate** by separating the background noise from the **intended message**[1].
- There are four basic chart types:
  - **Comparison chart** - e.g. a line chart displaying the profits of 3 departments over 7 years.
  - **Composition chart** - e.g. a pie chart displaying the market share of telcos in Singapore.
  - **Relationship chart** - e.g. a scatter plot of the relationship between tips given and cost of meal.
  - **Distribution chart** - e.g. a column histogram displaying the distribution of exam grades.

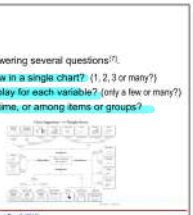
[1] Jens Gulba, Data Visualization – How to Pick the Right Chart Type? (2016) - https://eazybi.com/blog/data_visualization_and_chart_types

6

Selecting The Chart

Which Should I Use?

- A suitable chart can be determined by answering several questions[7].
  - How many variables do you want to show in a single chart? - (1, 2, 3 or many?)
  - How many items (data points) will you display for each variable? (only a few or many?)
  - Will you display values over a period of time, or among items or groups?
- The chart selection diagram by Andrew Abela can then be used to pick the right chart for the data type being visualise.



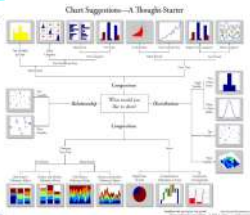
[7] Jens Gulba, Data Visualization – How to Pick the Right Chart Type? (2016) - https://eazybi.com/blog/data_visualization_and_chart_types

7

Selecting The Chart

Which Should I Use?

- Andrew Abela's Chart Chooser in Colour[8].



[8] Andrew Abela's Chart Chooser - https://qusonte.files.wordpress.com/2010/02/chartchoosercolor.jpg

8

Psychological Principles of Effective Graphics

The Eight-Fold Way

- Stephen Kosslyn proposed eight very useful principles to guide the design of **effective data visualisation**[1].
- This principles are rooted in his findings about how humans **perceive** and **comprehend visual information**.
- The eight principles are organised into three groups, with each group addressing different goals of:
  - **Connect** with your audience.
  - **Direct** the audience attention through the visual.
  - **Promote** understanding and memory retention.

[1] Stephen M. Kosslyn: Graph Design for the Eye and Mind, Oxford University Press (2006)

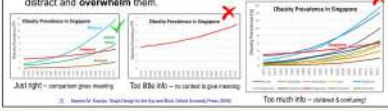
9

1. Connect with audience

Principle of Relevance

The Goldilocks Principle

- Communication is most effective when neither too much nor too little information is presented[1].
- Decide on **exactly what message** you want to convey. Then use only the **relevant data** that will help you communicate this message.
- Presenting too little may **puzzle** the audience, while presenting too much will distract and **overwhelm** them.



Just right – comparison gives meaning | Too little info – no context to give meaning | Too much info – cluttered & confusing!

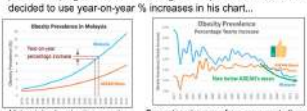
[1] Stephen M. Kosslyn: Graph Design for the Eye and Mind, Oxford University Press (2006)

11

Ethical Visualisation

Don't Hide Behind Percentages

- The Director of Weight Loss is concern about the effectiveness of his obesity prevention programmes and has to give an update to the cabinet next week. He decided to use year-on-year % increases in his chart...



Malaysia's alarming increase in obesity prevalence | Percentage increase from one year to the next (smoothing on moving average data, 3 samples)

Note: Obesity dataset available at https://ourworldindata.org/grapher/share-of-adults-defined-as-obese

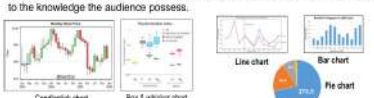
11

2. Connect with audience

Principle of Appropriate Knowledge

Know Your Audience

- Communication requires prior knowledge of relevant concepts, jargon and symbols[1].
- **Know** you audience and their familiarity with the content and charts being shared. Good visuals present new information but it should not be so new that it is unrelated to the knowledge the audience possess.



Candlestick chart (used in financial visualisation) | Box & whisker chart (scientific visualisation) | The common & familiar type of charts

[1] Stephen M. Kosslyn: Graph Design for the Eye and Mind, Oxford University Press (2006)

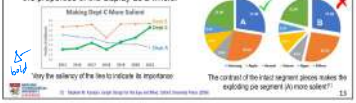
12

3. Direct and hold attention

Principle of Salience

Attention to the Important

- Attention is drawn to large perceptible differences[1].
- The most visually prominent aspects of a visual will **draw attention** to them, for this reason they should embed the **most important** information being conveyed.
- All visual properties are relative, so what counts as "visually striking" depends on the properties of the display as a whole.



Very the saliency of the line to indicate its importance | The contrast of the intact segment pieces makes the exploding pie segment (A) more salient?

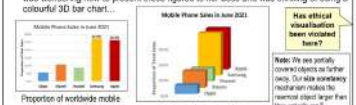
[1] Stephen M. Kosslyn: Graph Design for the Eye and Mind, Oxford University Press (2006)

13

Ethical Visualisation

Don't Hide Behind Another 3D Bar

- Apple mobile sales has fallen behind slightly in June 2021. An Apple sales executive was wondering how to present these figures to her boss and was thinking of using a colourful 3D bar chart...



Proportion of worldwide mobile phone sales in June 2021 | The 3D occluding bar chart

Note: We are partially covered objects as farther away. Our size constancy mechanism makes the nearest object larger than they actually are[1].

Note: Mobile sales dataset available at https://gs.statcounter.com/vendor-market-share/mobile

14

3. Direct and hold attention

Principle of Discriminability

Maintain Good Contrast

- Two properties must differ by a large enough proportion or they will not be distinguished[1].
- This principle relates to Weber's law on just noticeable difference. Be mindful your design and colour choices allow viewers to discriminate between variables and between the variable and its background.

Combine strong distinctive colours & shapes to improve discriminability | Small pie segments in A cannot be discriminated. Design B allows them to be visualise.

[1] Stephen M. Kosslyn: Graph Design for the Eye and Mind, Oxford University Press (2006)

15

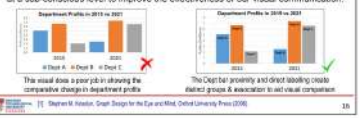
Change to 12 O'Clock view

4. Promote understanding and memory

Principle of Perceptual Organisation

Making Visual Sense

- People automatically group elements into units, which they then attend to and remember[1].
- We should exploit the various Gestalt laws that govern how we group visual object at a sub-conscious level to improve the effectiveness of our visual communication.



This visual does a poor job in showing the comparative change in department profits | The Dept bar proximity and direct labelling create distinct groups & association to aid visual comparison

[1] Stephen M. Kosslyn: Graph Design for the Eye and Mind, Oxford University Press (2006)

16

Ethical Visualisation

Scales That Mislead

- The profits of **Dept A** has been steadily declining over the years and is currently the lowest. There is a critical profit review with the board to decide its fate going forward. The head of Dept A is working on the presentation charts....

Profits of Dept A is currently the lowest of the three departments | The log_{10} scale applied to the $M profits hides the big profit differentials

Note: The logarithmic scale has a tendency to compress the distances between large numbers making seen large differences insignificant[1].

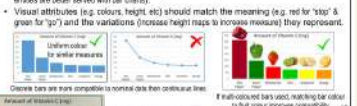
17

4. Promote understanding and memory

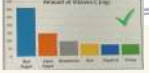
Principle of Compatibility

Saying The Same Thing

- A message is easier to understand if its form is compatible with its meaning[1].
- Compatibility dictates that we choose the right type of chart for the intended purposes (e.g. trends and continuous changes are best depicted with line charts, whilst discrete entities are better served with bar charts).
- Visual attributes (e.g. colours, height, etc) should match the meaning (e.g. red for "stop" & green for "go") and the variations (increase height maps to increase measure) they represent.



Discrete bars are more compatible to nominal data than continuous lines | If multi-coloured bars used, matching bar colour to fruit colour improves compatibility

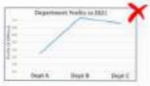


18

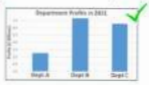
Principle of Information Changes

Meaningful Changes

- **People expect changes in properties to carry information**[1].
- Changes in the visual (e.g. changing colour, texture, line height & width, etc) will be interpreted by viewer as conveying information .
- Similarly, viewers expect any information that should be conveyed in the visual will be indicated by some visible change (e.g. transition from current to projected profits in line chart should be indicated by a different colour or pattern).



If entities don't vary continuously, use bar instead of line



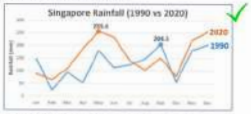
Projected profits & recession year clearly indicated

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

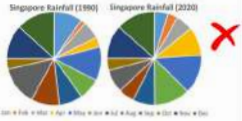
Principle of Capacity Limitation

Too Much Means Only A Little

- **People have a limited capacity to retain and process information. A message will not be understood if too much information must be retained or processed**[1].
- We can only maintain a limited amount of information in mind at any one time. Visuals should not require the viewer to hold more than about **four perceptual groups** in mind at once.



Visualising peak months with minimal cognitive load



Too many colours to compare!

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

Summary

Designing The Visualisation

- Effective visualisation design requires you to know the **purpose** and the **audience** you are designing for before you begin selecting what **suitable chart** to use.
- During the data visualisation design process, you need to be mindful of how humans **perceive** and **comprehend visual information**.
- Employ the eight psychological principles proposed by Kosslyn to help **connect** with the audience, **direct** their attention to your main points and facilitate their **understanding** and **retention** of your intended message.

References for Designing the Visualisation

[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

[2] Michael Schermann - A Reader on Data Visualization, Patterns - https://mschermann.github.io/data_viz_reader/patterns.html

[3] R. Kosara, Reflecting on the Design Criteria for Explanatory Visualizations (2016) - https://research.wildapricot.org/wp-content/uploads/kosara-GHPSV-2016.pdf

[4] DataViz Tip #21: Explanatory vs. Exploratory Charts - https://www.smcharts.com/datavizitip-21-explanatory-vs-exploratory-charts/

[5] Mustapha Mekhalna, 5 Steps to Create an Effective Data Visualization (2017) - https://www.highcharts.com/blog/post/5-steps-to-create-an-effective-data-visualization/

[6] Image extracted from R. Allison, Colors represent different things, in different cultures (2017) - https://blogs.sas.com/content/iagdm/2017/06/29/colors-in-different-cultures/

[7] Jonis Gulbis, Data Visualization - How to Pick the Right Chart Type? (2016) - https://supply.com/blog/data_visualization_and_chart_types/

[8] Andrew Abela's Chart Chooser - https://andrewfabela.files.wordpress.com/2011/02/chartchoosersmcolor.jpg

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

Part 2

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

Chapter 8.1
Design Graph to Enlighten: Principles and Best Practices

Outline

- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization

Prof. Edward Tufte from Yale University

- Other Best Practices

3

Outline

- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization

Prof. Edward Tufte from Yale University

- Other Best Practices

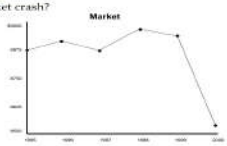
Examples of Misleading Visualizations

- Big discount or benefits?

5

Examples of Misleading Visualizations

- Is it a market crash?



6

Graphical Integrity

- **Principle: Graphics must not quote data out of context**
  - Showing the entire scale

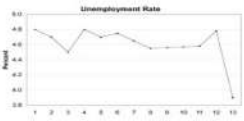
If data visualization does not start at zero, the viewers can be tricked into seeing false information.

7

Graphical Integrity

- **Principle: Graphics must not quote data out of context**
  - Showing the entire scale

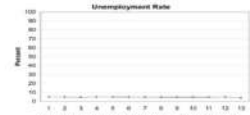
Examples of Misleading Visualizations



Unemployment Rate: A misleading visualization

9

Examples of Misleading Visualizations



Unemployment Rate: A visualization telling the truth

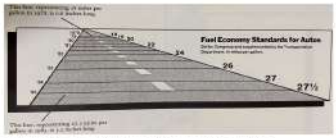
Graphical Integrity

- **Principle: Visual attribute value should be directly proportional to data attribute value**
  - Visual attributes: the visual encoding channels used to encode the underlying data, e.g., bar height, color density, circle radius, etc.
- Lie factor = $\frac{\text{size of effect shown in graphic}}{\text{sizes of effect in data}}$
- Truth = 1.0

11

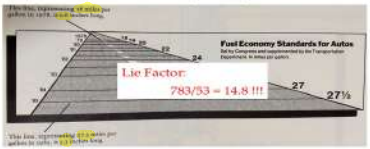
Fuel Economy Standards for Autos

- Lie factor = ?



Fuel Economy Standards for Autos

- Size of effect in data: (27.5 – 18.0)/18.0 = 53%
- Size of effect shown in graphic: (5.3 – 0.6)/0.6 = 783%



13

$\frac{27.5-18}{18} = 0.527$ $\quad \frac{5.3-0.6}{0.6} = 7.83$

$\frac{7.83}{0.527} = 14.84$

Graphical Integrity

- **Principle: The number of information-carrying dimensions (i.e., visual encoding channels) should not exceed the number of dimensions in the data.**

For example, avoid using two (or three) varying dimensions to show one-dimensional data

The visualization uses the height and width simultaneously of the bottle to represent the crude oil price, which is problematic:

- Ambiguity in perception: Without an explicit indication, it is difficult for viewers to know which visual channel is used to encode the data.
- Space inefficiency: it results in the usage of a larger space.

Height & Width Encoding

better

15

Graphical Integrity

- **Principle: The number of information-carrying dimensions (i.e., visual encoding channels) should not exceed the number of dimensions in the data.**

For example, avoid using two (or three) varying dimensions to show one-dimensional data

Questions:

- By reading the visualization itself, are you able to accurately identify which visual channel is used to encode the underlying data?
- How should we fix the issue?

Use only one visual encoding channel (e.g., only width or height)!

Rankings of Visual Encoding Effectiveness

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat Color Hue Texture	Length	Shape
Connection	Angle	Length
Containment	Slope	Angle
Shape	Area (Size)	Slope
	Volume	Area
	Shape	Volume

17

Outline

- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization

Prof. Edward Tufte from Yale University

Data-ink Ratio Maximization

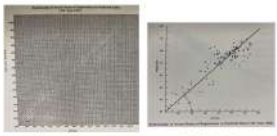
- Data-ink is the non-erasable ink used for the presentation of data
- Reduce the non-data-ink (a.k.a., chartjunk)
  - Remove unnecessary non-data-ink
  - Emphasize data-ink
- Enhance the data-ink

$$\frac{\text{Data Ink}}{\text{Total Ink}} = \text{Data Ink Ratio}$$

19

Data-ink Ratio Maximization

- For the following two visualizations that shows exactly the same data, which one is more clear for you to understand the underlying data?

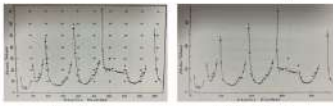


Low data-ink ratio — High data-ink ratio

23

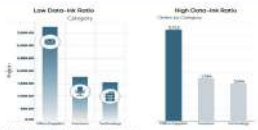
Data-ink Ratio Maximization

- For the following two visualizations that shows exactly the same data, which one is more clear for you to understand the underlying data?



Low data-ink ratio — High data-ink ratio

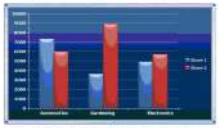
Data-ink Ratio Maximization



It is not necessary to put the icons on the bar chart, as it does not bring more information.

Data-ink Ratio Maximization: Avoid Chartjunk

- Avoid using unnecessary color shading for the bar



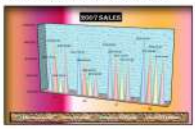
- The color shading here is useless and can actually be misleading for viewers!
- The background color is also meaningless.

23

Data-ink Ratio Maximization: Avoid Chartjunk

- Avoid colorful or wallpaper background

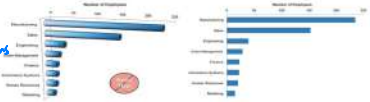
avoid unlikely 3D Rep



- The colorful background is also meaningless.

Data-ink Ratio Maximization; Avoid Chartjunk

- Avoid using 3D effects in graphics



- The 3D effect makes it harder for viewers to perceive the underlying data accurately.

25

Practical Guides for Using Color in Charts

- Rule 1: If you want different objects of the same color in a graph to look the same, make sure that the background- the color that surrounds them - is consistent.

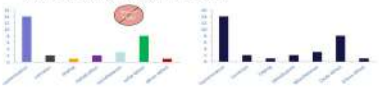
bad

Practical Guides for Using Color in Charts

- Rule 2: If you want objects in a graph to be easily seen, use a background colour that constracts sufficiently with the object

Practical Guides for Using Color in Charts

- Rule 3: Use different colors only when they correspond to different meanings in the data

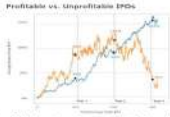


It is not necessary to use different colors for different bars.

29

Practical Guides for Using Color in Charts

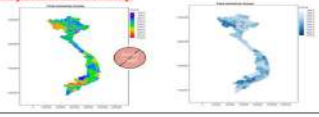
- Rule 4: Use soft, natural colours to display most information and bright and/or dark colours to highlight information that requires greater attention



! The grids and labels are in soft gray, while the line charts are in bright yellow and blue.

Practical Guides for Using Color in Charts

- Rule 5: When using colour to encode a sequential range of quantitative values, stick with a single hue and vary intensity from pale colours for low values to increasingly darker and brighter colours for high values, i.e., use sequential colormap



31

Outline

- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization
- Other Best Practices

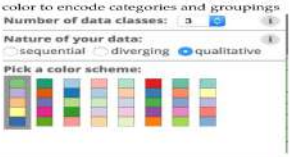
Prof. Edward Tufte from Yale University

Colormap Design

- Colormap: a mapping between data values and colors
- Two types of colormaps
  - Categorical (aka., qualitative)
  - Ordered (aka., quantitative): sequential + diverging
- A very useful tool: Colorbrewer2 http://colorbrewer2.org/

Categorical Colormaps

- Use color to encode categories and groupings



33

Rules for Categorical Colormaps

- The maximum number of colors which can be used is usually between 6 to 12
- The 12 recommended colors

1. Red	2. Green	3. Yellow
4. Blue	5. Black	6. White
7. Pink	8. Cyan	9. Gray
10 Orange	11. Brown	12. Purple

Ordered Colormaps

- **Sequential Colormap**
  - Ramp in luminance, possibly also hue.
  - Typically higher values map to darker colors.

- **Diverging Colormap**
  - Useful when data has a meaningful "midpoint".
  - Use neutral color (e.g., gray) for midpoint.
  - Use two saturated colors for endpoints.

- **Limit number of steps in color to 3-9**

  Why? avoid simultaneous contrast; need to hold mappings in memory

35

Ordered Colormaps

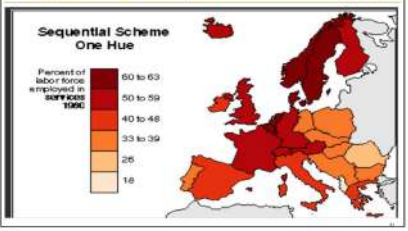
- **Sequential Colormap**
  - Ramp in luminance, possibly also hue.
  - Typically higher values map to darker colors.

- **Diverging Colormap**
  - Useful when data has a meaningful "midpoint".
  - Use neutral color (e.g., gray) for midpoint.
  - Use two saturated colors for endpoints.

- **Limit number of steps in color to 3-9**

  Avoid simultaneous contrast; Need to hold mappings in memory

Sequential Colormap



37

Practical Guides for Using Color in Charts

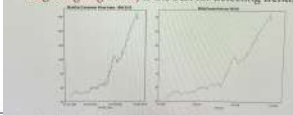
- Rule 7: Pay attention to the meaning of different colors

Example: Logo color
- Red = Adventure. Energy. (Coca Cola)
- Pink = Compassion. Nurturing. (Breast Cancer Research Foundation; Barbie)
- Orange = Cheerful. Warm. Optimistic. (Amazon)
- Yellow = Happy. Bright. Sunny. (McDonalds)
- Green = Health and Healing. Environment. Organic. (Whole Foods Market; Starbucks)
- Blue = Honesty. Trust. Dependability. (Facebook; Pfizer)
- …

41

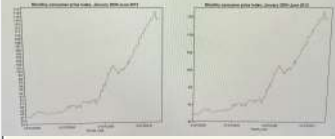
What is the Best Aspect-ration of a Two-variable Visualization?

- Should not manipulate the aspect ratio to intentionally exaggerate or downplay the rate of change.
- The general consensus seems to be that an aspect ratio that keeps the line segments around 45 degrees going up (or -45 degrees going down) is the best for detecting trends.

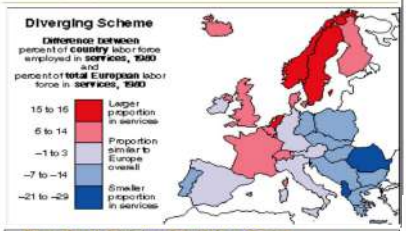


How many tick marks should you use?

- There is no exact number that works best in all circumstances, and the size of the graph is a factor that must considered: the longer the scale line, the more tick marks it should contain.



Diverging Colormap



Practical Guides for Using Color in Charts

- Rule 7: Pay attention to the meaning of different colors

Use colors that viewers are familiar with
- "red" for hot, dangers, cautions, & warnings
  >> watch out for culture conflicts (e.g., red in Asian cultures)
  >> watch out for differences of conventions: the meaning of red and green in different stock markets (China vs. USA)

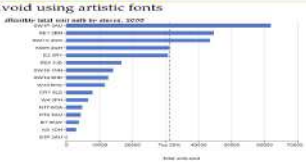
Practical Guides for Using Color in Charts

- Rule 6: Respect the **colorblind and** be colorblind friendly.

39

Graph Typography

- Avoid using artistic fonts



45

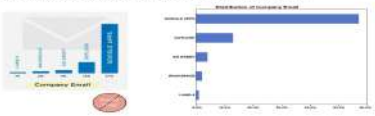
Practical Guides for Using Color in Charts

- Rule 7: Pay attention to the meaning of different colors

Example: Logo color

Graph Labeling

- Orientation of label should be reader friendly



Information Graphics and Data Visualization blogs

- Junk Charts (http://junkcharts.typepad.com/junk_charts/)
- Perceptual Edge (http://www.perceptualedge.com/blog/)
- Statistical Graphics and more (http://www.theusrus.de/blog/)
- EagerEyes (http://eagereyes.org/)
- Visualizing Economics (http://visualizingeconomics.com/)
- The Global Sociology Blog (http://globalsociology.com/)

51

Practical Guides for Using Color in Charts

- Rule 7: Pay attention to the meaning of different colors

Example: Logo color
- Red = Adventure. Energy. (Coca Cola)
- Pink = Compassion. Nurturing. (Breast Cancer Research Foundation; Barbie)
- Orange = Cheerful. Warm. Optimistic. (Amazon)
- Yellow = Happy. Bright. Sunny. (McDonalds)
- Green = Health and Healing. Environment. Organic. (Whole Foods Market; Starbucks)
- Blue = Honesty. Trust. Dependability. (Facebook; Pfizer)
- …

Reference

- Tufte, Edward (2nd Edition) The Visual Display of Quantitative Information, Graphics Press LLC, Connecticut, USA
- Yau, Nathan (2013) Data Points: Visualization that means something, John Wiley & Sons, Inc.

49

Practical Guides for Using Color in Charts

- Rule 7: Pay attention to the meaning of different colors

Use colors that viewers are familiar with
- "red" for hot, dangers, cautions, & warnings
  >> watch out for culture conflicts (e.g., red in Asian cultures)
  >> watch out for differences of conventions: the meaning of red and green in different stock markets (China vs. USA)

Reference

- Few, Stephen (2012) (2nd edition) **Show Me the Numbers: Designing Tables and Graphs to Enlighten**, Analytics Press, Oakland, USA
- Robbins, Naomi B. (2005) **Creating More Effective Graphs**, John Wiley & Sons, New Jersey, USA
- Wong, Dona M. (2010) **The Wall Street Journal Guide to Information Graphics**, W. W. Norton & Company, Inc. New York

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

Summary

- Interaction is an essential part of data visualization, which provides users more flexibility and control to explore the data of their interest.
- Without interaction, a visualization technique or system becomes a static image or autonomously animated images, limiting the usefulness of data visualization with the inclusion of dataset size and dimensions.
- You are recommended to enable necessary interactions in your own visualization approaches or systems!

Interesting Data Visualizations

- https://pudding.cool/projects/gallery/?sort=-is-of-the-day&type=is-of-the-day
- https://bost.ocke.org/links/
- https://archive.nytimes.com/www.nytimes.com/interactive/2013/04/04/business/global/asia-map.html

Chapter 8.2
Interactive Visualization

Outline

- What is Interactive Data Visualization?
- Why is Interaction Useful for Data Visualization?
- Seven Types of General Interactions

What is Interactive Data Visualization?

- Visible Change Under Your Control

- Unlike static charts, interactive data visualisation enables the user to further explore the dataset **by interactively manipulating** and **changing** the visual representation of all or a selected subset of the data using appropriate interaction techniques and input devices.

Why is Interaction Useful for Data Visualization?

- It is an essential requirement of data visualization

- Ben Shneiderman's mantra:

"Overview first, zoom and filter, then details-on-demand"

http://www.cs.umd.edu/users/ben/

Why is Interactive Data Visualization Useful?

- **Some benefits**

- Interactive visualisation is useful when the dataset is **large** and is embedded with **complex data insights** which have yet to be verified or uncovered.

- **Identify relationships** more effectively – by selectively focusing on specific metrics so that non-obvious cause-and-effect relationships within the data can be seen.

- **Identify trends and patterns faster** – by having the ability to visualise complex data at varying scales and in decluttered views by using only a selective data subsets.

- **Telling data stories** more convincingly – by presenting a sequenced and focused narrative through the ability to zoom in/out, highlight relevant information, filter, and change the parameters to present multiple viewpoints of the data.

Image from: Dennis Kul, Interactive Data Visualisation Definition - https://www.researchgate.net/publication/228644164_Interactive_Data_Visualisation_Definition

7 General Categories of Interactions

- *Select*: mark something as interesting
- *Explore*: show me something else
- *Reconfigure*: show me a different arrangement
- *Encode*: show me a different representation
- *Abstract/Elaborate*: show me more or less details
- *Filter*: show me something conditionally
- *Connect*: show me related items

Yi, J. Stasko, et al. "Toward a deeper understanding of the role of interaction in information visualization." TVCG, 2007.

7

Select

"Mark something as interesting"

- Mark items of interest to keep track
- Seems to often work as a preceding action to subsequent operations

Examples

- Select a landmark in Google Map
- Select a bar in a histogram
- Select a set of points by brushing

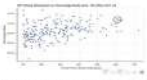


Image from: Leon D'Angio @ Pierophilia.com - Interactive Data Visualisation in Python With Bokeh - https://medium.com/pierophilia/interactive-data-visualisation-in-python-with-bokeh

Selection Functions

- ***Pointer*** selects a single object in a plot.
- ***Drag-box*** selects a rectangular region in a box.
- ***Brush*** is a generalisation of the drag-box. Once a rectangular region is defined, the brush allows users to move that region across a plot and thus dynamically change the selected subset.
- ***Slicer*** is an axis-parallel selection tool, which selects a range along an axis, where the end-point of the interval can be modified dynamically.
- ***Lasso*** allows users to define an arbitrary contiguous shape to select data.

Explore

"Show me something different"

- Exploration enables users to examine a different subset of data
- Exploration overcomes the limitation of display size

Examples

- Panning in Google Earth
- Direct Walking in Visual Thesaurus

Explore: More Examples

2D Navigation: MatrixWave [1]

3D Navigation: Interactive visualization of scanned mummies [2]

1 Zhao et al. MatrixWave: Visual comparison of event sequence data. SCHOLICS 2013. https://www.researchgate.net/publication/259644146_Matrixwave_Visualization_of_event_sequence_data_for_pattern_discovery
2 Vitenberg et al. "Interactive visualization of 3D scanned mummies at public venues." ACM Communications, 2010.

Encode

"Show me a different representation"

- Change visual appearances

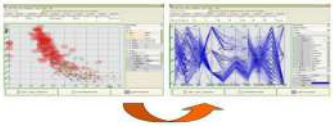
Examples

- Change color encoding
- Change size
- Change orientation
- Change font
- Change shape

https://github.com/d3/d3/wiki/Gallery

Encode

- Example: Select a different representation



Abstract/Elaborate

"Show me more or less detail"

- Adjust the level of abstraction (overview and details)

Examples

- Drill-down in Treemap
  https://observablehq.com/@d3/zoomable-treemap
- Zooming

Image from: SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations. https://www.researchgate.net/publication/317153128_Smartadp_Visualizing_Citywide_Billboard_Features

15

Filter

"Show me something conditionally"

- Change the set of data items being presented based on some specific conditions

Examples

- Attribute based filtering
- Dynamic query

Image from: MapTiss: Interactive https://www.researchgate.net/publication/279142854

Connect

"Show me related items"

- Highlight associations and relationships
- Show hidden data items that are relevant to a specified item

Examples

- Linking the relevant items in SmartAdp and Vizsfer

Image from: SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations. https://www.researchgate.net/publication/317153128_Smartadp_Visualizing_Citywide_Billboard_Features

17

Reconfigure

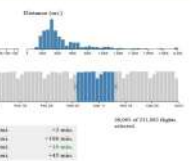
"Show me a different arrangement"

- Reconfiguring provides different perspectives by changing the spatial arrangement of representation

Examples

- Sorting data with respect to a particular attribute
- Changing the attributes in a scatter plot
- Reposition the visual elements

Connect: Linked Multiple Views (Demo)



Chapter 9.1
Exploratory Data Analysis

not assumption

- non parametric
  - ↳ Ranks/ordinal data
  - ↳ skewed data
  - ↳ Counts / freq

Outline

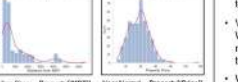
- What is Exploratory Data Analysis?
- Correlation Analysis
  - Measurement of Correlation
  - Univariate Analysis
  - Bivariate Analysis
  - Confirmatory Analysis
  - Multivariate Analysis

Univariate Analysis

Is My Distribution Normal Enough?

→ data is cont (interval / ratio) normal dist

- In order to use **parametric statistical methods** (e.g. Student's t-test), our data must have a **normal distribution**, otherwise nonparametric statistical methods must be used.
- The **normality** of the data or how close they conform to a Gaussian distribution can be tested using visual methods (e.g. **statsmodels's** `qqplot()`[4]).
- Numerical evaluations can be done using **statistical methods**[5] such as:
  - Shapiro-Wilk test
  - D'Agostino's K^2 test
  - Anderson-Darling test



Positive Skew – **Property['MRT']** Near Normal – **Property['Price']**

[4] Statsmodels Q-Q plot function – https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html
[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2019) – https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

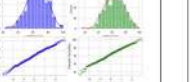
Student's t-test → statistical test to determine if there is a significant diff between the means of 2 groups

method of fam → Student's t-Test Analysis of Var, Person Corr Linear reg

Confirmatory Analysis

Check For Normality

- In order to decide if we should use a parametric or non-parametric statistical test, we should check the **normality** of the two distributions we are comparing.
- We can do a normality check using an appropriate statistical tests[6] or use the **Q-Q plot**[6].
- Visually, both the Reading & Writing score distributions of male students seem to conform to a Gaussian distribution.
- We can do further numerical statistical test to verify this.



Histogram/KDE plots of Reading & Writing scores and their Q-Q plots

[6] Statsmodels Q-Q plot function – https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html
[7] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2019) – https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

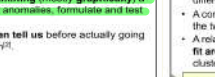
Statistical Analysis Alone is Not Enough

A Review

- Basic statistical analysis alone may not provide useful insights into how datasets can have **differing characteristics** despite having **similar statistical** parameters.
- Francis Anscombe[1] used the **Anscombe's Quartet**[1] of four datasets to show why **visualisation** is a **crucial part of data analysis**.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
1	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
2	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
3	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
4	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
5	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
6	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
7	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
8	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
9	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
10	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
11	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std Dev	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Anscombe's Quartet



[1] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21

Univariate Analysis

Visualising Normality

- A popular plot for checking normality is the **quantile-quantile plot** or **Q-Q plot**[8].
- The plot generates its own sample of the ideal Gaussian distribution and divides these samples into groups (e.g. 5) called **quantiles**. Each data point in our sample is **paired** with a similar member in the idealised distribution and plotted on the same cumulative distribution[9].
- **Perfect normality** will see all the data points fall along the 45° **red** line (i.e. matching ideal distribution).
- Dots seen deviating from the line shows deviation from the expected normal distribution.

Positive Skew – **Property['MRT']** Near Normal – **Property['Price']**

[8] Frank Wilcoxon, Q-Q Plots Explained (2020) – https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0
[9] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2019) – https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

Confirmatory Analysis

Statistical Test for Normality

- The **Shapiro-Wilk** test evaluates a sample of data and quantifies how likely the data sample is drawn from a Gaussian distribution[5]. Python **scipy.stats** library provides several functions for normality statistical tests, including the Shapiro-Wilks.

```
from scipy.stats import shapiro                # import the Shapiro-Wilks test
S,p = shapiro(Dist)                            # run the Shapiro-Wilks test
if (p > 0.05):                                 # check p value
  print('Distribution is normal')              # distribution is normal if p > 0.05
```

- Both Reading & Writing distributions meet the normality confidence level of 0.05.
- We can thus look at using a parametric test (e.g. the t-test or Z-test)

```
Distribution | male , reading score | mean = 65.47
Shapiro-Wilk test = | Stats: 0.99 , p = 0.9558 |
Distribution | male , reading score | is normal

Distribution | male , writing score | mean = 63.32
Shapiro-Wilk test = | Stats: 0.99 , p = 0.3848 |
Distribution | male , writing score | is normal
```

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2019) – https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

What is Exploratory Data Analysis?

- Exploratory Data Analysis (EDA) is a process of **examining** (mostly **graphically**) a dataset to **discover** patterns and relationships, spot anomalies, formulate and test **initial assumptions** using statistical measures[2].
- The primary aim of EDA is to **examine what data can tell us** before actually going through formal modelling and hypothesis formulation[2].
- EDA is usually an iterative process that involves:
  1. **Asking questions** about the available data.
  2. **Construct** appropriate data **visualisations** to answer the questions.
  3. **Evaluate** and **inspect** the answers, then derive further questions.

Repeat

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

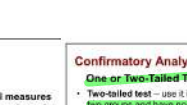
Bivariate Analysis

How Are These Two Variables Related?

- Bivariate analysis is used to find out if there is a **relationship** between two different variables in the dataset.
- A common plot used for bivariate analysis is the two-axes **scatter plot**.
- A relationship exist if the data points seem to **fit around a line** or a **curve**. The **tighter** the clustering, the **stronger** the relationship.

Did students getting high Math scores also got high Writing scores?

Yes, there is a reasonably strong positive correlation between Math and Writing scores



Writing vs Math Scores in *Student_performance.csv*

Confirmatory Analysis

Z-test or Student's t-test?

- If there is **more than 30 data samples**, the Z-test can be used, otherwise the t-test should be used with small sample sizes[7].
- Both the Z-test and t-test requires the data to be **normally distributed** and data points are **independent** from each other (i.e. one data point does not affect another data point). Strict normality requirements can be relaxed if the sample size is large.
- The Python **statsmodels** library provides both these statistical tests.

```
from statsmodels.stats.weightstats import ttest_ind   # t-test function
from statsmodels.stats.weightstats import ztest       # Z-test function
```

- Given that the sample size of male students is 482, the Z-test can be used here.

[7] Yugesh Agrawal, Hypothesis testing in Machine learning using Python (2019) – https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-2f9ec0a4f7bbe

Correlation Analysis

Are There Relationships in the Dataset?

- Datasets often have **different categories** (i.e. columns) containing many **measured values** describing a particular phenomena, event or situation.
- Since these values are collected from the **same event**, there is a possibility that they are **related** to one another[2].
- **Correlation** is the statistical technique that examines these relationships and describes how strongly different categories of measures or values are related.
- Correlation answers questions such as:
  - How does one variable change with respect to another?
  - If it does change, to what degree or strength is this change related?

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

Bivariate Analysis

What If One Variable is Categorical?

- Bivariate analysis using a **box plot** allows us to examine the **statistical measures** of one quantitative variable along with the relationship between multiple values of a **categorical variable** in the dataset.
- The **central lines** in the box plot allow us to **compare the median** values within the categories (e.g. female & male).
- The nature of the data distribution (i.e. normal or skewed) can be inferred from the **relative position** of the **central line** within the box.

Do female students do better in writing?

They seem to have a higher median score.

Box plots of Writing scores of Male & Female students in *Student_performance.csv*

Confirmatory Analysis

One or Two-Tailed Test?

- **Two-tailed test** – use it if you want to determine if there is a difference between two groups and have no concern about the direction of this difference. Example:

Null hypothesis (H_0)	$\mu_{Reading} = \mu_{Writing}$	(No difference in performance)
Alternate hypothesis (H_a)	$\mu_{Reading} \neq \mu_{Writing}$	(There is a difference in performance)

- **One-tailed test** – use it if you want to determine if there is a difference between groups in a specific direction. In our case, the Reading distribution seems to be visually higher than Writing one. We should therefore use:

Alternate hypothesis (H_a)	$\mu_{Reading} > \mu_{Writing}$	(Reading performance better than Writing)

- **Note**: One-tailed test has more statistical power than a two-tailed test at the same significance level. Results are more likely to be significant for a one-tailed test if there is indeed a difference between the groups in the direction predicted[8].

[8] Should you use a one-tailed test or a two-tailed test for your data analysis? – https://www.statisticsworldwide.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/

Measure Of Correlation

Pearson Correlation Coefficient

- Correlation measures how variables **change in tandem**, either in the same or in opposite directions, and also the magnitude or **strength** of these related changes between the variables.
- The **Pearson correlation coefficient** ($\rho_{X,Y}$), with a value between -1 and +1, measures the correlation between two variables X and Y. It is given by:

where: cov is the covariance
σ_X and σ_Y are the standard deviation of X and Y respectively

$$\rho_{X,Y} = \frac{cov(X,Y)}{(\sigma_X \sigma_Y)} \quad \text{Eqn. (1)}$$

Strong Correlation (> 0.7) [5] Image from Wikipedia, Pearson Correlation Coefficient – https://en.wikipedia.org/wiki/Pearson_correlation_coefficient Strong Correlation (< -0.7)

Bivariate Analysis

Confirmatory Analysis

- Visual analysis is useful for spotting relationships or patterns in data. However, we should check our assumption using appropriate statistical measures and **confirm** if an observation is **statistically significant**.
- The following confirmatory analysis steps could be taken:
  - Formulate a hypothesis.
  - Check normality of data distributions.
  - Select appropriate statistical test (e.g parametric or non-parametric).
  - Compute the test statistics to determine the **p value** (typically $p < 0.05$ is considered significant)

Are the differences between the reading & writing scores of the boys significant?

KDE plots of Reading & Writing scores of Male students in *Student_performance.csv*

Confirmatory Analysis

Statistical Significance

- **Z-test results** – Applying one-tailed ('larger') Z-test[8] on the **Reading & Writing** data:

```
Z,p = ztest(Reading, Writing, alternative='larger')   # call Z-test function
```

Z score	p value
Z-test = (2.393229395263579 , 0.008358097872215826)	

- **Rejecting null hypothesis** – the threshold at which is generally considered safe to reject the null hypothesis is **a p value of ($p < 0.05$). This means there is a less than** 5% chance that the observed data is due to chance.
- **Statistically Significant** – since the **p** value computed by the Z-test is way smaller than 0.05, we can safely say that the observed mean Reading score being higher than the mean Writing score for Male students is statistically significant.

[8] Statsmodels Ztest function – https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html

Univariate Analysis

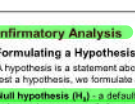
One Variable At A Time

- Univariate analysis involve the analysis of **one variable at a time** in the dataset. As such, it is not meant to find relationships between variables.
- **The main purpose of univariate analysis is to:**
  - find patterns in the variable such as its **central tendency** (e.g mean, median, mode)
  - understand its dispersion (e.g range, variance, maximum & minimum quartile, etc)
  - detect presence of outliers.

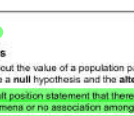
```
Property['Price'].describe()
count    414.000000
mean      37.980193
std       13.606159
min        7.600000
25%       27.700000
50%       38.450000
75%       46.600000
max       117.500000
Name: Price, dtype: float64

Property['Price']
Mean = 37.98
Median = 38.45
```

Apply `describe()` on category **Price** in *Property.csv*

`mean()` and `median()` values of Price



Histogram and Kernel Density Curve



Box Plot (Observing Outliers)

Confirmatory Analysis

Formulating a Hypothesis

- A hypothesis is a statement about the value of a population parameter. In order to test a hypothesis, we formulate a **null** hypothesis and the **alternate** hypothesis.
- **Null hypothesis (H_0)** - a default position statement that there is no relationship between two measured phenomena or no association among groups[10].
- **Alternate hypothesis (H_a)** - a statement that is contrary to the null hypothesis. It is usually the hypothesis being tested since we see a pattern, think it is true and want to find evidence to reject the null hypothesis and replace it with the alternate.

Null hypothesis (H_0)	$\mu_{Reading} = \mu_{Writing}$	(No difference in performance)
Alternate hypothesis (H_a)	$\mu_{Reading} > \mu_{Writing}$	(Reading scores better than writing)

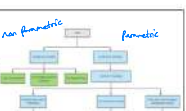
[10] Yugesh Agrawal, Hypothesis testing in Machine learning using Python (2019) – https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-2f9ec0a4f7bbe

Confirmatory Analysis

Selecting The Statistical Test

non parametric parametric

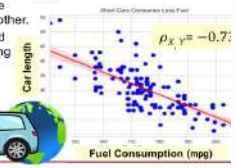
- Selecting an appropriate statistical test for your confirmatory analysis requires consideration of various factors such:
  - Continuous or discrete data.
  - Number of samples to compare.
  - Sample size.
  - Parametric or non-parametric test.
- A helpful flowchart in[10] shows some statistical tests that can be use based on the nature of the data and test requirements.



[10] Jaqueline Singh, Statistical Tests with Python (2020) – https://github.com/jaqueline-singh/statistical-tests-with-python/blob/master/Stats.ipynb

Correlation Does Not Imply Causation

- Remember the phrase "**Correlation does not imply causation**"[2].
- Just because two things are observed to be correlated does not mean one causes the other.
- Do not form conclusions **too quickly** based on correlation. **Invest time** to find underlying factors in the data that really influence the relationships between parameters



Strong negative correlation found between the length of the car and its fuel efficiency.
Save the Planet,
Drive Short Cars now!

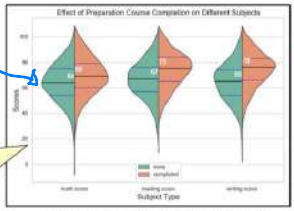
2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

Multivariate Analysis

Playing With Many Violins

- Multivariate analysis using **violin plots** allows us to compare the **probability density curve** of multiple quantitative variables for a **categorical variable pair**.
- The **central lines** in the violin plot allow us to **compare** the **median** values within the categories (e.g. those who completes test preparation course & those who did not).
- In Seaborn's `violinplot()`, the `hue` parameter can be assigned to a categorical variable with two values. Setting `split=True`, then compares them side-by-side.

Students do better in all subject when completing test preparation.

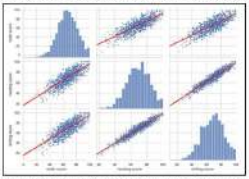


Violin plots showing effects of prep course completion on Math, Reading & Writing scores

Multivariate Analysis

Analysis of Three or More Variables

- Multivariate analysis allows us to look at the **correlation of three** or **more** variables at a time.
- A common way of visualising multivariate data is to use a **matrix** of multiple **scatter plots**.
- The convenient and powerful Seaborn `pairplot()` function provides this feature[11].
- Setting the `kind` parameter to `kind='reg'` will show the linear regression lines.



Analysing Math, Reading & Writing scores in *Student performance.csv*

[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

Summary

Exploratory Data Analysis

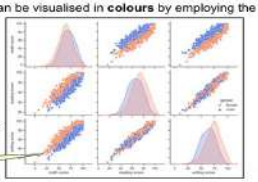
- Correlation analysis allows us to visualise **relationships** within the dataset.
- Data visualisation techniques for this purpose include histogram & kernel density plots (univariate), scatter plot (bivariate), box & violin plots (bivariate & multivariate), paired plot & heatmap (multivariate).
- Correlation between variable pairs can be quantified using the **Pearson correlation coefficient** and this values can be effectively visualised using appropriate **colour palettes** in heatmaps and paired plots.
- Visual observation of relationships and patterns should be confirmed with appropriate **statistical analysis** based on the **nature** and **number** of data distributions being compared.

Multivariate Analysis

Adding In Categorical Variables

- An additional **categorical variable** can be visualised in **colours** by employing the `hue` parameter in `pairplot()`[11].
- By assigning the hue parameter to a specific categorical variable (e.g. `hue='gender'`), each type in the category is assign a different colour from a default or specified palette.
- Such multivariate analysis allows interesting relationships and patterns to be observed in the different categorical types.

Boys do better in math than writing



Math, Reading & Writing scores of Female & Male students

[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

References for Correlation Analysis

[1] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21
[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)
[3] Image from Wikipedia, Pearson Correlation Coefficient - https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
[4] Statsmodels Q-Q plot function - https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html
[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) – https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/
[6] Paras Varshney, Q-Q Plots Explained (2020) - https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0
[7] Yogesh Agrawal, Hypothesis testing in Machine learning using Python (2019) – https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce
[8] Should you use a one-tailed test or a two-tailed test for your data analysis? - https://www.statisticssolutions.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/
[9] Statsmodels ztest function - https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html
[10] Jagandeep Singh, Statistical Tests with Python (2020) - https://python.plainenglish.io/statistical-tests-with-python-880251e9b572
[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html
[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html
[13] Bibor Szabo, How to Create a Seaborn Correlation Heatmap in Python? (2020) – https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

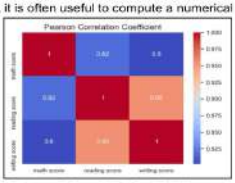
Note: All online articles were accessed between June to July 2021

Multivariate Analysis

Getting the Correlation Values

- Instead of just visualising the correlation, it is often useful to compute a numerical measure for each correlation pair.
- **Pandas** `corr()` function and **Seaborn**'s `heatmap()` can compute and display the **pairwise** linear (Pearson) correlation coefficients for all the columns in the dataframe with **numerical** values[12].

```
sns.heatmap(
  Data.corr(method='pearson'),  # Pearson coeff
  annot = True,                 # show coeff values
  cmap = 'coolwarm'))           # diverging palette
```



Correlation coefficients of Math, Reading & Writing scores

[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

Acknowledgement

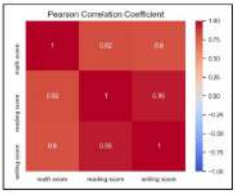
The slides of this lecture is adapted from the course material by Prof. Goh Wooi Boon, NTU.

Multivariate Analysis

Getting the Correlation Values

- Instead of just visualising the correlation, it is often useful to compute a numerical measure for each correlation pair.
- **Pandas** `corr()` function and **Seaborn**'s `heatmap()` can compute and display the **pairwise** linear (Pearson) correlation coefficients for all the columns in the dataframe with **numerical** values[12].

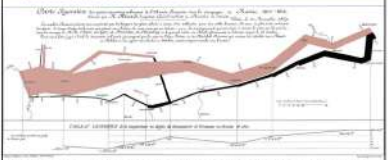
```
sns.heatmap(
  Data.corr(method='pearson'),  # Pearson coeff
  annot=True,                   # show coeff values
  vmin=-1, vmax=1,              # min-max scale
  cmap='coolwarm'))             # diverging palette
```



[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

Chapter 9.2
Time-series Data Visualization

A Famous Example: Napoleon's Retreat from Moscow

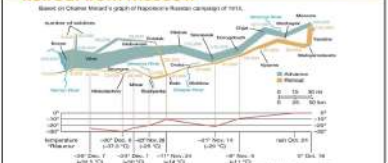


Charles Minard's 1869 chart showing the number of men in Napoleon's 1812 Russian campaign army; their movements, as well as the temperature they encountered on the return path.

3

A Famous Example: Napoleon's Retreat from Moscow

Based on Charles Minard's graph of Napoleon's Russian campaign of 1812.



Charles Minard's 1869 chart showing the number of men in Napoleon's 1812 Russian campaign army; their movements, as well as the temperature they encountered on the return path.

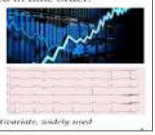
Outline

- Time-series data
- Time representation
- Time-series patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

5

Time-series Data

- Time series data, also referred to as time-stamped data, is a sequence of data points indexed in time order.
- Examples
  - Stock price
  - Temperature
  - Economic indicators
  - Patient health evolution metrics (e.g., ECG data)
  - Sensor data



Large volume, high dimensional, multivariate, widely used

Time Representation

- Taxonomy

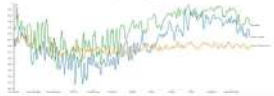
There are mainly five categories of time representations used in time-series data visualizations:

Bednarz, Matthew, et al. "Timelines revisited: A design space and considerations for expressive storytelling." TVCG 23.9 (2016): 2018-2034.

7

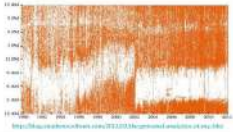
Linear Time

- Present time data as a 2D line graph
  - Time on x-axis
  - The other variable on y-axis
  - The most common way to represent a timeline



Linear Time

- Personal Data Visualization, *by* Stephen Wolfram
  - The x-axis and y-axis represent the year and daily timeslot respectively, and each dot represent an email
  - It shows the time of over 1 million emails since 1989



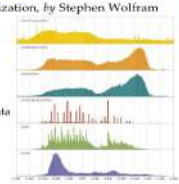
http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/

9

Linear Time

- Personal Data Visualization, *by* Stephen Wolfram

Visualization of other personal time-series data



http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/

Radial Time

- Data distributed along the spiral
- It is appropriate for presenting the periodic nature of time, such as natural cycles of people, weather systems, etc.

Example: Nightingale Rose Charts:

- It shows the causes of mortality in the army in the East and was created by Florence Nightingale
- It indicates the annual rate of mortality per 1,000 in each month that occurred from preventable diseases (in blue), wounds (in red), and other causes (in **black**).

https://en.wikipedia.org/wiki/File:Nightingale-mortality.jpg

11

Radial Time

Example: Historic creative's daily routines

https://pbs.twimg.com/media/CJ0aOSVWIAAEnbW.jpg

Grid Time

- Grid representations such as the standard Month-Week-Day calendar are ubiquitous.
- However, grid representations are only appropriate for timelines that reflect these granularities of time.

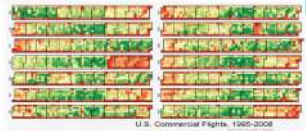
Example: Dow Jones stock price from 2006 to 2009

http://chartsnotebook.wordpress.com/2011/01/monthly-chart/

13

Grid Time

Example: US Commercial Flights



U.S. Commercial Flights, 1995-2008

Source: http://infosthetics.com/archives/2011/01/visualizing_us.html

Spiral Time

- Spiral representations are dense, space-filling shapes that radiate to or from a centre point.
- Spiral timelines are aesthetically appealing and appropriate for presenting many events within a single dense display

Example: Ross Spiral Curriculum

Ross spiral curriculum, 2015. https://spiral.com/spiralled

15

Arbitrary Time

- Arbitrary timeline representations may appear visually similar to connected scatterplots and time curves
- Like spiral representations, arbitrary representations can be useful for presenting many events within a single dense display and improve the memorability of a time-series data

Example: Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data

http://www.aviz.fr/~bbach/timecurves

Arbitrary Time

Demo Video of Time Curves

http://www.aviz.fr/~bbach/timecurves

17

Arbitrary Time

Example: Arab Spring Timeline

http://www.abc-journalism.com/infographics/events/arab-spring-timeline

Time Representation

- Taxonomy

- Among the above four types, linear time representation is more popular

Bednarz, Matthew, et al. "Timelines revisited: A design space and considerations for expressive storytelling." TVCG 23.9 (2016): 2018-2034.

19

What will you learn from this lesson?

- Time-series data
- Time representation
- Time-series data patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

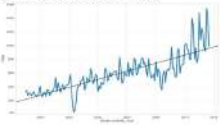
Time-series data patterns

- Trend
- Rate of change
- Co-variation
- Cycles
- Exceptions

21

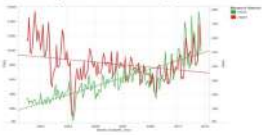
Time-series Patterns: Trend

- The overall or general direction of change in a series of time-series values is called the trend, which is often displayed in a graph as a trend line.



Time-series Patterns: Trend

- Example: Visitors from China is in increasing trend and visitor from Japan is in decreasing trend



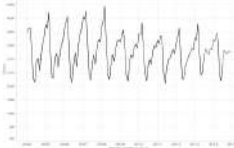
23

Time-series Patterns: Rate of change

- The percentage difference between one value to the next value.

Time-series Patterns: Cycles

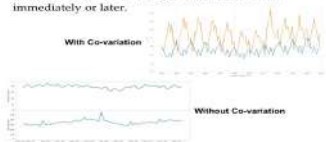
- Cycles are patterns that repeat at regular intervals.



25

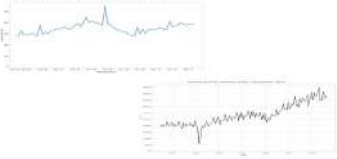
Time-series Patterns: Co-variation

- When two time series relate to one another so that changes in one are reflected as changes in the other, either immediately or later.



Time-series Patterns: Exceptions

- Values that fall outside the norm.



27

What will you learn from this lesson?

- Time-series data
- Time representation
- Time-series data patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

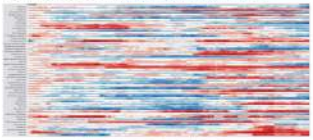
Massive Time-series Data

- Classic line chart
  - When there are massive time-series data, serious visual clutters will occur

29

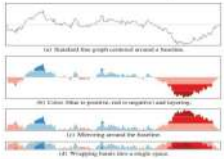
Massive Time-series Data

- Horizon graph is designed to mitigate the visual clutter issues when there are massive time-series data



Massive Time-series Data

- The Basic Construct of a Horizon Graph

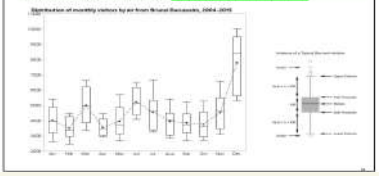


reduce space

31

Distribution Over Time

- Possible Visualization: boxplots with a trend line



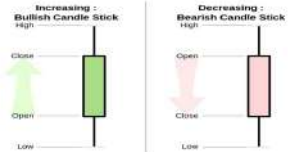
Distribution Over Time

- Possible Visualization: Candlestick Charts

33

Distribution Over Time

- Interpreting Candlestick Chart



Distribution Over Time

- Candlestick in D3.js

35

Part-Whole Relationships

- To show values that combine to form a whole comparison to one another and the whole change through time

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
East	54,102	70,172	30,027	33,209	33,031	33,248	31,074	32,273	34,119	112,264	90,607	37,400
Central	125,092	80,051	37,045	52,988	107,159	38,248	19,067	19,401	30,298	12,261	40,801	35,269
West	57,364	31,742	33,600	30,412	125,284	116,814	90,000	49,271	30,595	93,371	23,114	99,000
South	31,018	27,236	49,149	38,800	33,000	38,208	37,000	33,000	33,000	33,000	33,000	33,000
Total	$267,576	$209,201	$149,821	$155,409	$298,474	$226,518	$177,141	$133,945	$127,012	$217,896	$188,522	$205,669

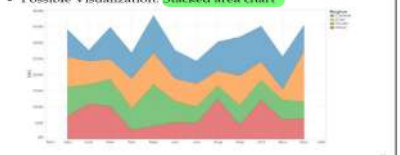
Part-Whole Relationships

- Possible Visualization: stacked bar chart

37

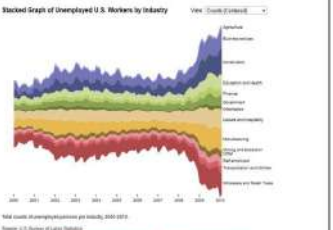
Part-Whole Relationships

- Possible Visualization: Stacked area chart



Part-Whole Relationships

- Possible Visualization: Stream Graph (Stacked Graph)



http://bmsc.washington.edu/~jhorvitz/ecoviz/time/stack.html
http://www.d3-graph-gallery.com/streamgraph

39

Animation: Visualizing the Change Over Time

- Example: Animated bubble plot

Source: http://www.gapminder.org/tools/#chart-type=bubble&dsh=1

43

Reference

Visualization of Time-Oriented Data

Aigner, W., Miksch, S., Schumann, H., Tominski, C.Springer

47

What will you learn from this lesson?

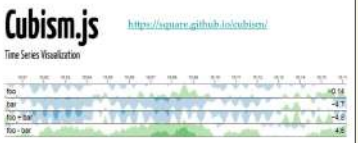
- Time-series data
- Time representation
- Time-series data patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

More Research on Time-series Data Visualization

Source: http://www.timeviz.net/

Visualizing the Change Over Time

- Animation + Interactivity: Horizon Graph using d3.js



https://square.github.io/cubism/

Tutorial: http://bost.ocks.org/mike/cubism/intro/

Summary

- Time representation
  Linear, radial, grid, spiral, arbitrary
- Time-series data patterns
  Trend, rate of change, co-variation, cycles, exceptions
- Interactive techniques for time-series data visualization

45

Reference

- Introduction to Cycle Plots (http://www.perceptualedge.com/articles/guests/intro_to_cycle_plots.pdf)
- Slopegraph
  - Slopegraphs for comparing gradients: Slopegraph theory and practice (http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0003nk)
  - In praise of slopegraphs (http://visualizingdata.com/index.php/2013/12/in-praise-of-slopegraphs/)
  - Displaying Change Between Two Points in Time (http://www.perceptualedge.com/articles/visual_business_intelligence/dl_ednaying_change_between_two_points_in_time.pdf)
  - Edward Tufte's "Slopegraphs" (http://charliepark.org/slopegraphs/)
  - A Slopegraph Update (http://charliepark.org/a-slopegraph-update/)

49

Interactivity

- Select: mark something as interesting
- Explore: show me something else
- Reconfigure: show me a different arrangement
- Encode: show me a different representation
- Abstract/Elaborate: show me more or less details
- Filter: show me something conditionally
- Connect: show me related items

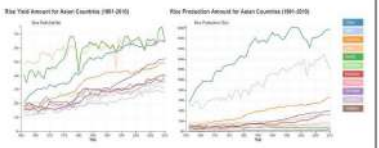
41

Reference

- Slopegraphs are your Friend. Tables secretly hate you. (http://makeapowerfulpoint.com/2014/01/3/slopegraphs-are-your-friend-tables-secretly-hate-you/)
- Communicating changes in rank over time: bumps charts and slopegraphs (http://datatodisplay.com/blog/chart-design/communicating-changes-rank-time/http://datatodisplay.com/blog/chart-design/communicating-changes-rank-time/)
- Tableau Tip Tuesday: How to Create Slopegraphs (http://vizwiz.blogspot.sg/2015/01/slopegraphs.html)
- How to make a slope chart in Tableau (http://graysanscalcite.com/tableau/how-to-make-a-slope-chart-in-tableau/)
- Slopegraphs in Tableau (http://datanomicud.com/2013/12/slopegraphs-in-tableau/)

Interactive Line Graph

- Example: Filtering



Link: http://tinyurl.com/mfgqxw

Questions?

Thank You!

Reference

- Horizon graph
  - The Development of the Horizon Graph (http://www.sciencevis.com/Vis08_Workshop/DVD/Reijner_submission.pdf)
  - Time on the Horizon (http://www.perceptualedge.com/articles/visual_business_intelligence/time_on_the_horizon.pdf)
  - Horizon Graphs Revisited (http://www.perceptualedge.com/blog/?p=390)
  - Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations (http://vis.stanford.edu/files/2009-TimeSeries-CHI.pdf)
  - Interactive Horizon Graphs (http://hal.archives-ouvertes.fr/docs/00/78/13/90/PDF/JHG.pdf)
  - Horizon charts in Tableau (http://www.tableau.com/learn/tutorials/on-demand/horizon-charts)

51

Reference

- Animated Bubble Plot (Motion Chart)
  - Wiki (http://en.wikipedia.org/wiki/Motion_chart)
  - Visualizing Change: An Introduction to Time-Series Analysis (http://www.perceptualedge.com/articles/visual_business_intelligence/visualizing_change.pdf)

Chapter 10
Geospatial Data Visualization

Outline

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

Geospatial Data

- Geospatial data (a.k.a, location-based data or geographic data) refers to data and information that has explicit or implicit association with a location relative to Earth.

XCOORD	YCOORD	NAME
103.8589	1.2940	Livania (Marina Bay Sands)
103.8574	1.2947	Singapore Pools Suntec City
103.8469	1.2835	Ya Shine
103.8483	1.2829	Singapore Turf Club China Square
103.8521	1.2936	Tay Sui Lan Agency
103.8492	1.2861	7-Eleven Circular Rd
103.8440	1.2856	Singapore Pools People's Park Centre
103.8464	1.2850	Fang Yuan Agency Enterprise

SUBZONE_C	TOTAL_POP	YOUNG
AMSZ01	6590	890
AMSZ02	30890	8980
AMSZ03	28850	8410
AMSZ04	26160	5280
AMSZ05	39190	2520
AMSZ06	24550	3050
AMSZ07	4860	1030
AMSZ08	8370	1360

Geospatial Data

- Geospatial data (a.k.a, location-based data or geographic data) refers to data and information that has explicit or implicit association with a location relative to Earth.
- Geospatial data can be sourced from telematics devices, Global Positioning System (GPS) data, geospatial satellite imagery, Internet of Things, geotagging, etc.

Geospatial Data Example: Location-based Services

- Geospatial data is highly relevant to our everyday life
- Geospatial data is also generated by mobile devices and sensors everyday

Google Maps | Yelp | Tripadvisor

Outline

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

What is Map?

- "The map speaks across the barriers of language; it is sometimes claimed as the language of geography."

— Carl Ortwin Sauer, 1956

What is Map?

- A map is a model of real world depicted by a collection of cartographic symbols or/and visual abstraction.

9

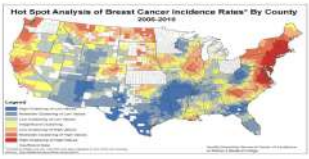
What is Map?

- An interface between geospatial data and map users.

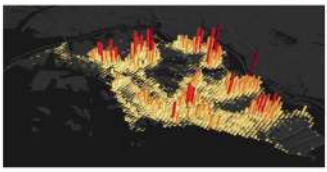
Map

11

Map

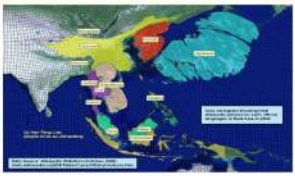


Map

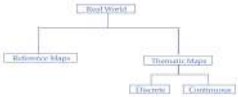


13

Map



Taxonomy of Maps



Slocum, T. A., McMaster, R. B., Kessler, F. C., and Howard, H. H. (2008) *Thematic Cartography and Geovisualization, 3rd Edition*

15

Reference Maps

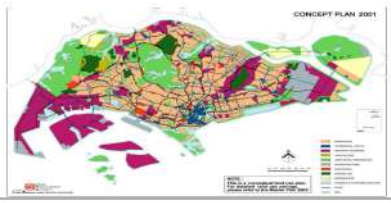
- A.k.a., General-Purpose Map
- Emphasize the location of spatial phenomena, such buildings, roads, vegetation, rivers, etc.
- Some popular reference maps: topographical map and internet map such as Google map

Thematic Maps

- A.k.a., Statistical Map
- A thematic map is used to display the spatial pattern of a theme or attribute
- Starting from 19th century
- Some popular thematic maps: Choropleth map, Cartogram map, etc.

17

Thematic Map: Land Use Map



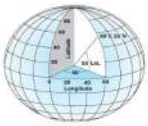
Outline

- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

19

Geographic Coordinate Systems

- Define the locations on the earth using a 3D spherical surface



Latitude-Longitude

Map Projection

- Represent the **surface** of a sphere or other 3D body on a plane
- Necessary for creating maps

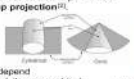
Wikipedia, Map projection.
https://en.wikipedia.org/wiki/Map_projection

21

Map Projection

From a Sphere to a Plane

- Hundreds of different map projections have been devised since the early Babylonian stone-inscribed world map[1] which was created in early 600 BC.
- It is generally acknowledged that most maps fall into **three** broad families of **map projection**[2].
- These map projections are based on the **developable surfaces** of the **cylinder, cone** and **plane**.
- Which surface is best will depend on the **region** to be mapped, its geographical **extent** and the desired **geometric properties** of the map[2].



An ancient Babylonian map of the Mesopotamian world (700-600 BC)
Image from [3]

[1] World History Encyclopedia, Babylonian Map of the World - https://www.worldhistory.org/image/3/babylonian-map-of-the-world/
[2] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html

23

Map Projection

Cylindrical Projection

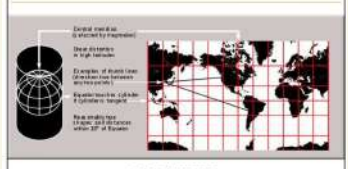
- Cylindrical projection is produced by wrapping a cylinder around a globe (Earth). The map projection is the image of the globe projected onto the cylindrical surface, which is then unwrapped into a flat surface.
- Cylindrical projections can be either **equal-area** (preserves area but distorts shape), **conformal** (preserves angle/shape but distorts area and distance), or **equidistant** (preserve distance).

Mercator conformal cylindrical projection
Image from [1]

Equal-area cylindrical projection
Image from [2]

[1] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html
[2] Wolframa, Mercator Projection - https://mathworld.wolfram.com/MercatorProjection.html

Example: Mercator Projection



A cylindrical map projection

Example: Mercator Projection

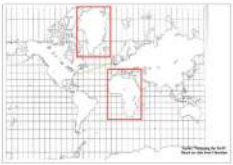
Distorted Area Size

Wikipedia, Mercator projection.
https://en.wikipedia.org/wiki/Mercator_projection

25

Example: Mercator Projection

Distorted Area Size



Example: Mercator Projection

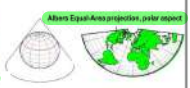
https://www.youtube.com/watch?v=6L_VpCbLuIs
https://www.youtube.com/watch?v=AHD7O3Q2J_

27

Map Projection

Conic Projection

- A conic projection is derived from the projection of the globe onto a **cone** placed over it. In the normal aspect, the apex of the cone lies on the polar axis of the Earth.
- Conic projections often achieve less distortion at mid and high latitudes than cylindrical projections[1].



Albers Equal-Area projection, polar aspect
Image from [1]

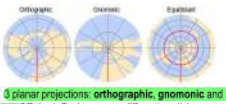
2 conical projections examples: **Albers equal area** (preserves area), **equidistant** (preserves distance) and **conformal** (preserves shape) [2]

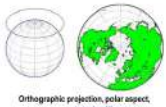
[1] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html
[2] ConicalProjections - https://mathscitools.com/gallery/conicalprojections/

Map Projection

Azimuthal Projection

- An azimuthal projection is a projection of the globe onto a plane.
- Most azimuthal projections are not suitable for displaying the entire Earth in one view, but give a sense of the spherical globe[1].





Orthographic projection, polar aspect,
Image from [1]

3 planar projections: **orthographic, gnomonic** and **equidistant**. Each covers a different spatial range (with the latter covering both northern and southern hemispheres) and each preserves a unique set of spatial properties.) [2]

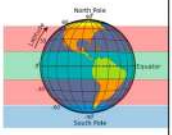
[1] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html
[2] Plane Projections - https://mathscitools.com/gallery/planeprojections/

29

Map Projection

Which Projection Should I Choose for a Region?

- The **latitude rule**[1] is a good rule of thumb in making a decision on which map projection to use for a given geographical region.
- To map tropical regions (e.g. Singapore, India), use **cylindrical** projection (e.g. Mercator).
- To map middle latitude (e.g. USA, China, New Zealand), use **conic** projection (e.g. Albers Equal-Area, Equidistant Conic).
- To map polar regions (e.g. North Pole), use **azimuthal** projection (e.g. Azimuthal Equidistant).



[1] Jochen Albrecht (Hunter College, City University of New York, CUNY), Choosing a projection - http://www.geo.hunter.cuny.edu/cartopy/2013/lectures/lecture-cartopy-1/powerpoint/Cartopy-Projection.ppt

Map Projection

What is the Purpose of the Map?

- The **purpose** of your geospatial analysis also influences the map projection choice. Map projections are not 100% accurate. Each projection distorts the surface of the Earth in some ways, while retaining some useful properties.
- The **equal-area** projections (e.g. "Lambert Cylindrical Equal Area", "Albers Equal Area Conic") **preserve area**[1]. This is a good choice, if you want to **calculate the area** of a country or city or plot a **dot density map** for each region.
- The **equidistant** projections (e.g. "Azimuthal Equidistant projection") **preserve distance**[2] and would be suitable if you want to analyse flight distances or shipping routes.
- The **conformal** projection is useful when the map's main purpose involves **measuring angles**, showing accurate **local directions**, or representing the **shapes** of features or **contour lines** (e.g. in weather maps, civil engineering & military maps)[3].

[1] Jochen Albrecht (Hunter College, City University of New York, CUNY), Choosing a projection - http://www.geo.hunter.cuny.edu/cartopy/2013/lectures/lecture-cartopy-1/powerpoint/Cartopy-Projection.ppt
[2] Alexis Cook, Coordinate Reference Systems - https://www.l-gages.com/wp-content/uploads/2016/07/Coordinate-Systems.pdf

31

Demo of More Map Projections

- https://observablehq.com/@d3/projection-transitions

Projection Transitions

Outline

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

33

Raster Data Model

- All raster formats are basically the same
  - Cells organized in a matrix of rows and columns
  - Content is more important than format: data or picture?

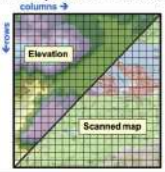
columns →

Raster data
- Elevation
- Landuse codes
- Population density

Good for analysis
- Slope from elevation

Good for mapping
- Thematic layers
- Derivative products (like shaded relief)

← rows



Raster pictures
- Scanned maps
- Satellite images (unclassified)
- Photos of buildings

Good for mapping
- Backgrounds

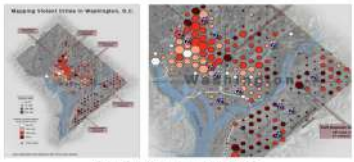
Good for attributes
- Picture of house

Bad for analysis

37

Point

Symbol Map



Cooper Thomas, Violent Crime in Washington, D.C.
https://visual.ly/community/infographics/geography/violent-crime-washington-dc

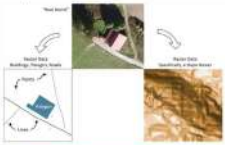
41

Symbol Map

The New York Times, How Democrats Took the House.
https://www.nytimes.com/interactive/2018/11/02/us/politics/how-democrats-took-the-house.html

Basic Geospatial Data Models

- Vector data model represent features as discrete **points, lines and polygons**, and is widely used in cartography, and network analysis
- Raster data model represent the continuous data as rectangular matrix of square cells (i.e., pixels), and is useful as background maps and for spatial analysis

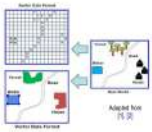


Data model definitions: https://www.fundwatch.org/wp-content/2020/01/28/Create/GeoDatabase/Geodatabase1.html

Geospatial Data File Formats

Basic types

- There are a variety of geospatial data file formats that are widely used by existing Geographical Information Systems (GIS). The two basic types are:
- **Raster GIS file** – These files contain raster data that are made up of **pixels** (i.e. grid cells of data values). They are usually regularly spaced and square but need not always be so[1].
- **Vector GIS file** – These files contain **vector graphics** that comprise of data that describes individual points (e.g. landmarks), vertices of polygons (e.g. boundaries), lines (e.g. roads or paths) and other geometric entities[2].



38

[1] What are the Raster Data Formats in GIS and Remote Sensing (2020) - https://www.geoinfo.com/2020/raster-data-formats-in-gis.html
[2] What is Vector Data Model of GIS (2020) - https://www.gisandtech.com/2020/04/vector-data-model-of-gis.html

Symbol Map

Rural West Initiative, Mapping Journalism's Voyage West.
https://web.stanford.edu/group/mapshed/cgi-bin/drupal/visualizations/ac_newspapers

43

Vector Data Model

- Used to represent **points, lines, and areas**
- All are represented using coordinates
  - One per **point**
  - Lines as **polylines**
    - Straight lines between points
  - Areas as **polygons**
    - Straight lines between points, connecting back to the start
    - Point locations recorded as coordinates

35

Outline

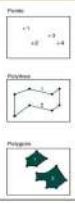
- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

39

Point Distribution Map

BBC, Every death on every road in Great Britain 1999-2010.
https://science.blog.co.uk/news/ids/13977/224

Vector Data Model



Points	Point number	(x,y) coordinates
	1	(0,4)
	2	(3,2)
	3	(5,3)
	4	(6,0)

Polylines	Polyline number	(x,y) coordinates
	1	(1,6) (2,6) (2,2) (7,3)
	2	(4,1) (5,5) (6,5) (7,2)

Polygons	Polygon number	(x,y) coordinates
	1	(0,4) (2,5) (3,5) (4,3) (3,0) (1,0)
	2	(5,2) (5,3) (4,5) (5,4) (6,5) (7,4) (8,1) (6,0) (5,2)

Point Data

- Points scattered in the geographic space, with longitude and latitude information, optionally with weights
  - Landmarks on the map
  - Restaurants of your neighborhood
- Marking the data on the map would be the most direct and simplest way to visualize point data
- Point-based visualizations are widely used

Outline

- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

45

Line Data

- Line data comprise locations and the paths between them
- Each line has a length property, namely the distance between two locations

Line

Air Traffic Network

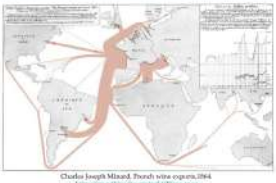
Martin Grandjean, Connected World: Untangling the Air Traffic Network.
http://www.martingrandjean.ch/connected-world-air-traffic-network/

47

Facebook Friend Relationship

Facebook, 2010.
https://blog.facebook.com/blog/

Bundling Lines



Charles Joseph Minard, French wine exports,1864.
http://en.wikipedia.org/wiki/Flow_map

49

Flow Map

Linking the immigrations from California to other states with arrows

Visualized with Flow Map

Buchin, Kevin, Bettina Speckmann, and Kevin Verbeek. "Flow map layout via spiral trees." IEEE transactions on visualization and computer graphics 17.12 (2011): 2536-2544.

Outline

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

51

Region

Selling Real Estate: Men vs. Women

Trulia. Is real estate a man's or woman's world.
https://bit.ly/2KBNLG

The World Gas Prices

Gasoline Prices Around the World
https://www.bloomberg.com/graphics/gas-prices/#20484,United-States,USD,g

53

Area

Choropleth Map

- A thematic map in which a set of pre-defined areas (e.g., states or countries) are shaded or patterned in proportion to the measurement of the statistical variable within different regions
- Also known as, filled map
- Assume that the data is evenly distributed in the area
- Often represent the scales of data with **colors**

Choropleth Map

- Colour scheme tool: ColorBrewer

http://colorbrewer2.org

55

Choropleth Map

The New York Times, 2016 Presidential Election Results.
https://www.nytimes.com/elections/2016/results/president

Choropleth Map Can Be Misleading

- The map fails to take account of **the population distribution**
  - It fails to allow for the fact that the population of the red states is on average significantly lower than that of the blue ones. The blue may be small in area, but they represent a large number of voters, which is what matters in an election
- When data is accumulated at the place where display space is less, there comes a mismatch, and vice versa
- This mismatch would very possibly **mislead** the audience

57

Choropleth Map Can Be Misleading

Mark Newman, 2016 Maps of the 2016 US presidential election results.
http://www-personal.umich.edu/~mejn/election/2016/

Area

Cartogram Map

- A map in which some thematic mapping variable – **such as travel time, population, or Gross National Product** – is substituted for land area or distance.

Mark Newman, 2016 Maps of the 2016 US presidential election results.
http://www-personal.umich.edu/~mejn/election/2016/

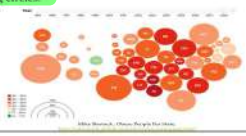
59

Cartogram Map

Mark Newman, 2016 Maps of the 2016 US presidential election results.
http://www-personal.umich.edu/~mejn/election/2016/

Other Maps: Dorling Cartogram

- Cartograms distort the shape of geographic regions so that the area directly encodes a data variable. A common example is to resize countries proportional to population or GDP.
- **Dorling cartogram** represents each geographic region as non-overlapping circles.



Mike Bostock, Obama Bangla Bar Graph.

61

Outline

- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

Geospatial Datasets

OpenStreetMap
https://www.openstreetmap.org/

HK Government Datasets
https://data.gov.hk

63

Geospatial Datasets

- Google/Baidu API
- OpenStreetMap (http://www.openstreetmap.org)
- HK government dataset (http://data.gov.hk)
- NYC Open Data (https://opendata.cityofnewyork.us)
- Data.gov (https://www.data.gov)

64

Common Geospatial Data Visualization Packages

- JavaScript
  - D3: d3.geo (https://d3js.org/d3-geo)
  - Leaflet (https://leafletjs.com/)
  - …

- Python
  - Geoplot (https://residentmario.github.io/geoplot)
  - GeoPandas (https://geopandas.org/en/stable/)
  - …

65

Summary

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

Summary

- Geospatial data refers to data and information that has explicit or implicit association with a location, and has been widely seen in various applications
- Map projection converts the **surface** of a sphere or other 3D body on a 2D plane. Each map projection method is a trade-off of different criteria and there is no optimal map projection for all situations
- Geospatial data models include vector and raster data models
- Geospatial data visualization techniques mainly cover three types: Point data visualization, line data visualization and region data visualization

67

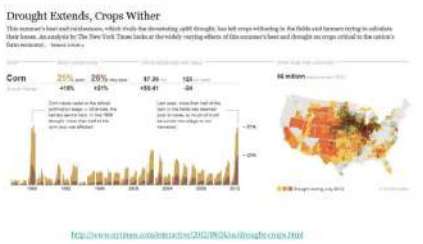
Interesting examples of cartogram

- Cartogram of the world's languages
  http://th-mayer.de/cartogram/
- China Still Dominates, but Some Manufacturers Look Elsewhere
  https://archive.nytimes.com/www.nytimes.com/interactive/2013/04/08/business/global/asia-map.html
- Dorling Cartogram of Obesity in the United States (1995 to 2008)

https://homes.cs.washington.edu/~jheer/files/zoo/ex/maps/cartogram.html

69

Example: HTML5 + JavaScript based GeoVisual Analytics



http://www.nytimes.com/interactive/2012/08/24/us/drought-crops.html

70

d3.js for Visualizing Geographical Data

- d3.geo (Geography)
  (https://github.com/mbostock/d3/wiki/Geo)

71

D3 Tutorial

- Let's Make a Map
  https://bost.ocks.org/mike/map/
- How to Make Choropleth Maps in D3
  https://rockcontent.com/blog/how-to-make-choropleth-maps-in-d3/
- A Map to Perfection: Using D3.js to Make Beautiful Web Maps
  https://www.toptal.com/javascript/a-map-to-perfection-using-d3-js-to-make-beautiful-web-maps

Chapter 11.1
Hierarchical Data (Tree) Visualization

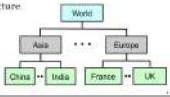
Outline

- What is hierarchical data?
- What are the hierarchical data visualization (a.k.a., Tree Visualization) methods?
  - Node-link (tree) diagrams
  - Indented trees
  - Layered (adjacency) diagrams
  - Enclosure diagrams

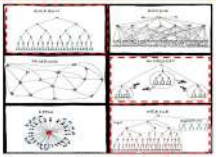
3

Hierarchical Data

- Hierarchical data represent the **hierarchical relationship** between objects/entities, where **the attributes** associated with objects/entities are often considered as well.
- As an example, **countries** (e.g. China, India, France, UK, etc) are a subset of continents (e.g. Asia, Europe, etc).
- More Examples:
  - Hierarchical organizational structure
  - Information organization
    - File directories
    - Development of species
  - Logical connections
    - Decision tree

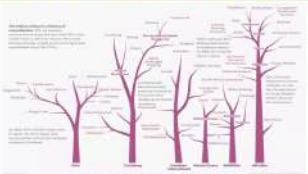


Hierarchical Data: Organizational Chart of IT Companies



Hierarchical data are intrinsically in a **tree-like** structure.

Hierarchical Data: Beer Family Tree



7

Hierarchical Data Visualization

- Node-link (tree) diagrams
- Indented trees
- Layered (adjacency) diagrams
- Enclosure diagrams
  - Treemap
  - Circle packing layout

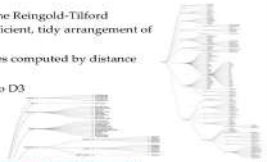
Node-link Tree Diagrams

- Nodes are distributed in space, connected by straight or curved lines
- Typical approach is to use 2D space to break apart breadth and depth
- Often, space is used to communicate hierarchical orientation

9

Node-link Tree Diagrams

- Tidy Tree
  - Implements the Reingold-Tilford algorithm for efficient, tidy arrangement of layered nodes
  - Depth of nodes computed by distance from the root
  - Integrated into D3



Node-link Tree Diagrams

- Radial Tidy Tree
  - A radial layout is based on a polar rather than Cartesian coordinates
  - It has a pleasing aesthetic, and uses space more efficiently
  - Integrated into D3

11

Node-link Tree Diagrams

- Reingold-Tilford algorithm for drawing node-link diagrams
  - Bottom-up recursive approach: Repeatedly divide space by leaf count
  - For each parent, make sure subtrees are drawn
  - Make smarter use of space
    - + Maximize density and symmetry
    - + Clearly encode depth level
    - + No edge crossings
    - + Pack subtrees as closely as possible
    - + Centers parent over subtrees

Indented Trees

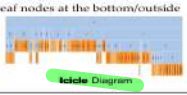
- Indentation is used to show parent-child relationships
- It is easy to implement and intuitive to understand
- But much scrolling is often needed in big data browsing and may lose the context

The directory structure of Flare package

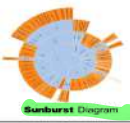
13

Layered (adjacency) Diagrams

- They are space-filling variants of node-link diagrams, such as **icicle diagram** and **Sunburst diagram**
- Nodes drawn as solid areas (arcs or bars)
- Placement relative to adjacent nodes reveals place in hierarchy
  - Root node at the top/center
  - Leaf nodes at the bottom/outside



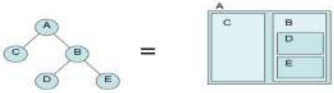
Icicle Diagram



Sunburst Diagram

Enclosure Diagrams or Treemaps

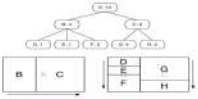
- Encodes tree structure using *spatial enclosure*
  - Enclosure indicates hierarchy
- Benefits:
  - Provides single view of entire tree
  - Easier to spot small / large nodes



15

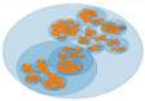
Enclosure Diagrams or Treemaps

- Assume that each leaf node has a size-related attribute (e.g., size of the file); size of parent nodes is the sum of its children's sizes
- Recursively slice and dice a rectangle in different directions to represent a (potentially large) hierarchy



Enclosure Diagrams or Treemaps

Enclosure diagram (treemap)



Circle-packing layout

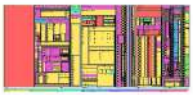
Question: Is it easy for you to visual identify the ***depth*** of the tree?

17

Father of Treemap

- Treemap was first introduced in 1991
- It was first used to visualize storage usage in a hard disk
- For more historical development of treemap: http://www.cs.umd.edu/hcil/treemap-history/

Ben Shneiderman



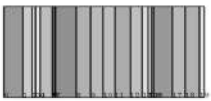
Slice-and-dice Treemap

- Ben Shneiderman proposed the slice-and-dice Treemap in 1991
- Layout method:
  - Recursive subdivision
  - Root is the largest rectangle
  - Size of child nodes are assigned according to their weights
  - Subdivide children nodes

19

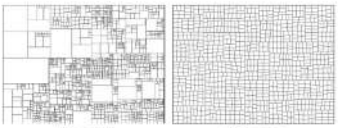
Problems of Slice-and-dice Treemap

- Simple slice-and-dice subdivision method results in thin and long rectangles
- It would be difficult to interact with nodes



Squarified Treemap

- Optimize the placement of nodes within a level to make them as square as possible

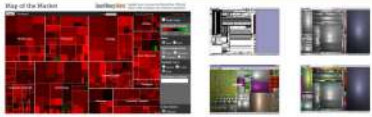


(a) File system (b) Organization

Source: http://www.cs.umd.edu/~ben/papers/Bederson-Quantum-Treemaps-and-Bubblemaps-UIST2001.pdf

21

Treemap Applications



Stock Market Disk Browser

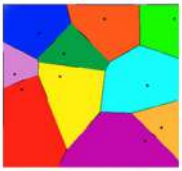
Business Sale Data Analytics

Source: http://www.smartmoney.com/map-of-the-market/?tab=SM_legend_more
http://www.treemap.com/solutions/demos/selection.html

25

Variant of Treemap: Voronoi Treemap

- Voronoi treemaps visualize hierarchical data by recursively partitioning convex polygons using weighted centroidal Voronoi diagrams



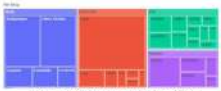
https://en.wikipedia.org/wiki/Voronoi_diagram

Treemap Implementation with Plotly

- The **treemap** method from the Python package **Plotly** can be used.

```
import plotly.express as px                    # import Plotly Express
L1 = 'Species'                                 # top hierarchy – Species
L2 = 'Animal Name'                             # next hierarchy – Animal Name
Fig = px.treemap(Data, path=[px.Constant('Pet Shop'), L1, L2], # call treemap
                 values='No. in Shop')         # map rectangle area to no. in pet shop
```

L1	L2	Area
Species	Animal Name	No. in Shop
Birds	Zebra Finches	30
Birds	Canaries	10
Mammals	Dogs	100
Fish	Goldfish	12
Reptiles	Monitor	2



Treemap of Animal Inventory (No. in Shop) encoded in the Area

Plotly Treemap: https://plotly.com/python/treemaps/

Variant of Treemap: Voronoi Treemap

- Example: Use animated Voronoi Treemap to summarize 35 Years of World's Economy

https://howmuch.net/articles/world-economy-as-a-voronoi-animation

23

Summary

- Hierarchical data represent the **hierarchical relationship** between objects/entities, where **the attributes** associated with objects/entities are often considered as well.

- We have introduced four major types of **hierarchical data visualization** (a.k.a., **tree visualization**) techniques
  - Node-link (tree) diagrams
  - Indented trees
  - Layered (adjacency) diagrams
  - Enclosure diagrams

Variant of Treemap: Voronoi Treemap

- Voronoi treemap in D3

https://github.com/fkenel01/voronoi-treemap

24

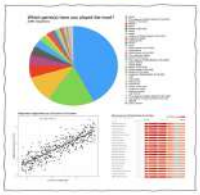
Chapter 11.2
Graph Visualization

Outline

- What is a graph/network?
  - Real world networks
  - Common graph vocabulary
- Graph Visualization Methods
  - Node-link diagram
  - Adjacency matrix
  - Others
- Packages and tools for graph visualization

3

What is a Graph?



Graphs

Common charts that represent data are often referred to as "graphs", or "graph visualizations"

- Bar charts
- Line charts
- Pie charts
- Etc.

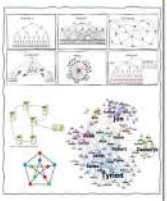
Clarification: for this lecture, when I refer to graphs, I do not mean the type of charts shown on the left.

What is a Graph?

Graphs

Let's instead talk about graphs, networks, & trees in the mathematical sense: a model for representing items and the relationships between those items

- Social / friendship networks
- Computer networks
- Energy or transportation grids
- Organizational structures
- Etc.



Why do we care about visualizing graphs?

5

Networks in Real World

- Telecommunication network

Networks in Real World

- Social network

7

Common Graph Vocabulary

9

Common Graph Vocabulary

Common Graph Vocabulary

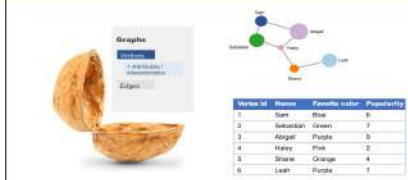
Vertex Id	Name	Favorite color	Popularity
1	Sam	Blue	8
2	Sebastian	Green	7
3	Abigail	Purple	6
4	Haley	Pink	2
5	Shane	Orange	4
6	Leah	Purple	7

11

Common Graph Vocabulary

Vertex Id	Name	Favorite color	Popularity
1	Sam	Blue	8
2	Sebastian	Green	7
3	Abigail	Purple	6
4	Haley	Pink	2
5	Shane	Orange	4
6	Leah	Purple	7

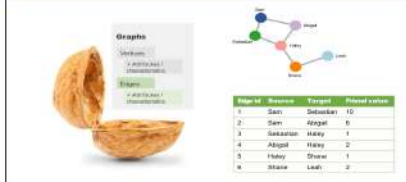
Common Graph Vocabulary



Vertex Id	Name	Favorite color	Popularity
1	Sam	Blue	8
2	Sebastian	Green	7
3	Abigail	Purple	6
4	Haley	Pink	2
5	Shane	Orange	4
6	Leah	Purple	7

13

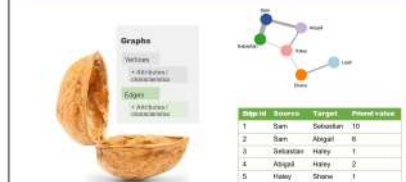
Common Graph Vocabulary



Edge Id	Source	Target	Friend value
1	Sam	Sebastian	10
2	Sam	Abigail	6
3	Sebastian	Haley	1
4	Abigail	Haley	2
5	Haley	Shane	1
6	Shane	Leah	2

14

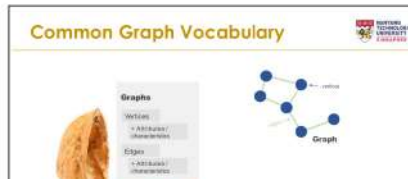
Common Graph Vocabulary



Edge Id	Source	Target	Friend value
1	Sam	Sebastian	10
2	Sam	Abigail	6
3	Sebastian	Haley	1
4	Abigail	Haley	2
5	Haley	Shane	1
6	Shane	Leah	2

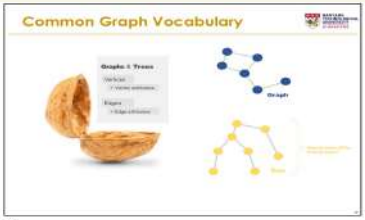
15

Common Graph Vocabulary



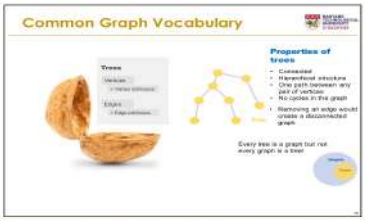
Questions?

Common Graph Vocabulary

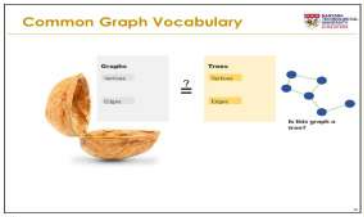


17

Common Graph Vocabulary

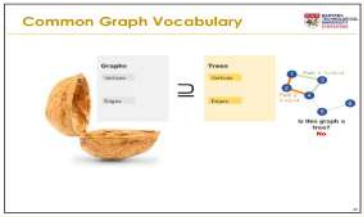


Common Graph Vocabulary

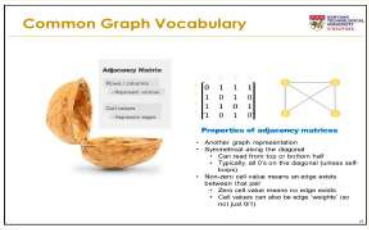


19

Common Graph Vocabulary

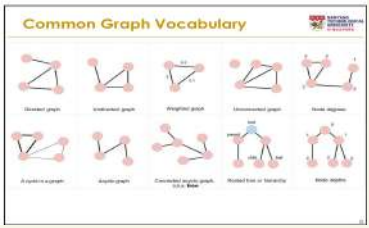


Common Graph Vocabulary



21

Common Graph Vocabulary



Outline

- What is a graph/network?
  - Real world networks
  - Common graph vocabulary
- Graph Visualization Methods
  - Node-link diagram
  - Adjacency matrix
  - Others
- Packages and tools for graph visualization

23

Graph Visualization Methods

- Graph visualization (a.k.a, network visualization) is concerned with visual representations of graph or network data
- Effective graph visualization reveals graph structures and help users understand and analyze the network data

Enron Email Network

https://medium.com/analytics-vidhya/social-network-analysis-measures/

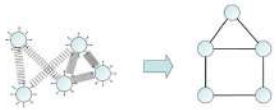
Graph Visualization Methods

- Node-link diagram
  - Force-directed layout
  - MDS layout
- Adjacency matrix
- Others
  - Hybrid layout
  - Graph simplification methods

25

Node-link diagram: Force-Directed Layout

- What about graphs without an intrinsic order?
- Physical model:
  - edge → spring
  - node → mass point



Node-link diagram: Force-Directed Layout

- It assumes that there is a spring between each pair of nodes, which leads to attractive force and repulsive force as follows:

$$f_a(d) = d^2/k$$
$$f_r(d) = -k^2/d$$

where d is the distance between the two nodes and k is a constant.

- With the effects of **attractive force and repulsive force**, nodes far away will be dragged near and nodes that are overlapped will be pushed away, and finally reach a stable balance after **iterations.**

T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. Software-Practice and Experience. 21(11):1129–1164, 1991.

27

Node-link diagram: Force-Directed Layout

- Starting positions: random or initial **configuration**
- Loop:
  - Compute **the attractive and repulsive** forces for every pair of nodes
  - Accumulate the **force** (vector) for every node
  - Update the position of each node **step by step** according to their forces
- The loop stops until the layout is "good enough"

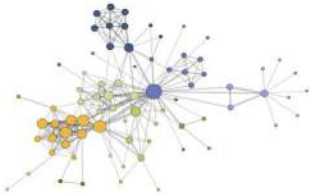
Results of Force-directed Layout

Force-directed graph with elliptic forces

https://d.ocks.org/jpenna/bb6f0f0c1f2a5d5dc4a14969f237

29

Results of Force-directed Layout



Character relations in Les Miserables

http://bio.stanford.edu/jhee/files/zoo/

https://homes.cs.washington.edu/~jhee/files/zoo/cs_networks/force.html

Pros and Cons of Force-directed Layout

- Pros
  - Very flexible for any type of graphs
  - Forces can be customized
  - Easy to implement
- Cons
  - **Local optimal**
  - **Initial configuration** is important
  - **Computation complexity** of iterative algorithm
- Extensions
  - Barnes-Hut quadtree decomposition
  - FADE, GRIP, FMS, FM^3, GVA

try it interactively at https://observablehq.com/@bl3/force-directed-graph

31

Graph Visualization Methods

- Node-link diagram
  - Force-directed layout
  - MDS layout
- Adjacency matrix
- Others
  - Hybrid layout
  - Graph simplification methods

Node-link diagram: MDS (Multidimensional Scaling) Layout

- Focus on addressing the limitations of force-directed layout
- Dimension reduction
  - Keep the consistency of relative distance between nodes
- MDS is a global optimal method
- Optimal function $||x_i - x_j|| \approx d_{ij}$, where d_{ij} is the graph-theoretical distance between Node i and j, and x_i and x_j correspond to their coordinates in the 2D plane

33

Solving Optimal Function

- Stress majorization
  - $\text{Stress}(X) = \sum_{i,j} w_{ij}(d_{ij} - ||x_i - x_j||)^2$
  - d_{ij} is the graph-theoretical distance between the i-th node and the j-th node
  - $w_{ij} = d_{ij}^q$, usually $q = -2$

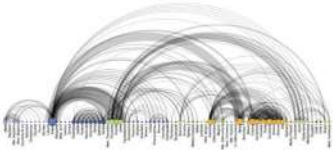
Other Presentations of Node-link Diagram

- Orthogonal Diagram
  - UML diagram

- Nested ordered
  - Recursively applying nested layout
  - For intrinsic ordered topology
- Arc Diagram

35

Arc Diagram



Character relations in Les Miserables

http://bl.ocks.org/mbostock/raw/4284103/

https://fanteam.cs.washington.edu/~theory/files/assets/interactive.html

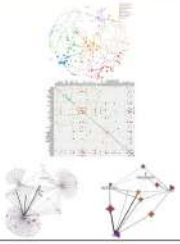
Node-Link Diagram Summary

- Pros:
  - Intuitive visual interpretation
  - Good representation of topology, clusters and paths
  - Flexible, many variants
- Cons:
  - Almost for all algorithms, time complexity is a bit high ~ $O(n^2)$
  - Not so good for dense graphs (especially edge cluttered graphs)

37

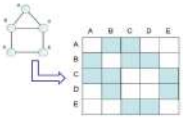
Graph Visualization Methods

- Node-link diagram
  - Force-directed layout
  - MDS layout
- Adjacency matrix
- Others
  - Hybrid layout



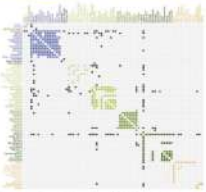
Adjacency Matrix

- $N \times N$ matrix, representing relations among N objects
- Position (i,j) represents the relation between the i-th object and the j-th object
  - Weight
  - Direction
- Related issues
  - Ordering
  - Path finding



39

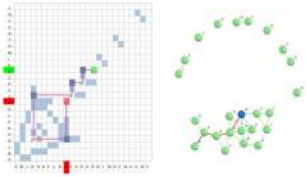
Good Ordering of Adjacency Matrix



Appropriate node ordering is critical for visually identifying the meaningful graph patterns

https://homes.cs.washington.edu/~theory/files/assets/interactive.html

Path Visualization for Adjacency Matrix

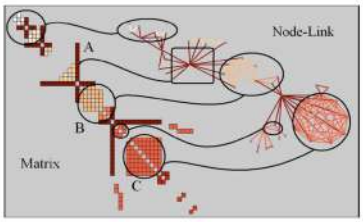


Left: Connecting matrix entries and the diagonal.
Right: Path {N,D,S,P,R} is highlighted in the Node-link Diagram.

Compared with node-link diagrams, it is **hard to perform path-finding tasks** in adjacency matrix.

41

Recognizing the Patterns of Matrix



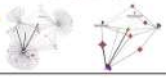
Adjacency Matrix Summary

- Pros
  - No edge crossing, good for edge cluttered graph
  - Good visual scalability
  - Good presentation of graph pattern
- Cons
  - Visualization is abstract to understand
  - Difficult to follow a transitive relation path

43

Graph Visualization Methods

- Node-link diagram
  - Force-directed layout
  - MDS layout
- Adjacency matrix
- Others
  - Hybrid layout
  - Graph simplification methods



Hybrid Layout

- Adjacency matrix can handle complex and dense edge relations, but suffers from insufficient space usage when there are many nodes
- Node-link diagram can handle relatively more nodes, but can suffer from serious visual clutters when there are dense edge relations
- What if there are a large number of nodes, and some of them have dense edge relations?

-- Combine node-link diagram with matrix

Fig. 1: NodeTrix: Representation of the largest component of the Info-Vis Co-authorship Network

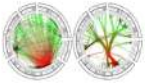
Nathalie Henry, et al. NodeTrix: A Hybrid Visualization of Social Networks. TVCG 2007

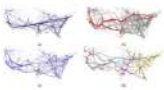
45

Graph Simplification Methods: - Edge Bundling

- Edge bundling approaches are designed to handle the dense edges between nodes, reducing visual clutters and helping users identify the overall links between nodes

Original Graph Visualization



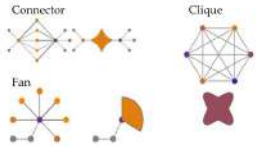


[1] D Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. TVCG 2006

[2] Zhou, Hong, et al. "Edge bundling in information visualization." Tsinghua Science and Technology 19.2 (2013): 115-135.

Graph Simplification Methods: - Network Motifs

- Network motifs aims to represent representative subgraphs of a large graph as a series of meaningful motifs to simplify the visualization of large graphs

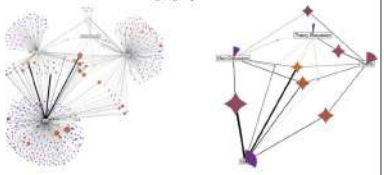


Dunne and Shneiderman, Motif Simplification: Improving Network Visualization Readability with Fan, Connector, and Clique Glyphs. CHI2013

47

Graph Simplification Methods: - Network Motifs

- Network motifs aims to represent representative subgraphs of a large graph as a series of meaningful motifs to simplify the visualization of large graphs



Dunne and Shneiderman, Motif Simplification: Improving Network Visualization Readability with Fan, Connector, and Clique Glyphs. CHI2013

Graph Visualization Tools

- Visualization packages

Alchemy.js
…

Sigma.js
Cytoscape.js
VivaGraphJS
…

D3.js
Pixi.js
…

More and more flexible

49

Graph Visualization Tools

- Softwares
  - Gephi
  - Cytoscape
  - Palantir
  - visone
  - …

50

Graph Visualization Tools

https://gephi.org/

51

Summary

- Graph visualizations are widely used
- Graph visualization methods
  - Node-link diagram
    - Familiar, but problematic for dense graphs
  - Matrix
    - Abstract, hard to follow paths
  - Graph simplification can help
    - Not always possible, and not always appropriate
- Take-home message: no best solution; graph visualization is still under active research!!!

Chapter 12
High-dimensional Data Visualization

Outline

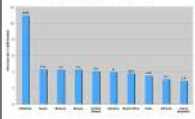
- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

Data Dimension

- Dimension (Number of attributes)
  - One Dimension
  - Two Dimension
  - Three Dimension
  - High Dimension

1-D Data

Eaten or not

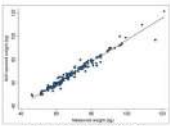


Top ten divorce rates

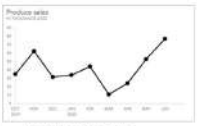
5

2-D Data

2-D plots

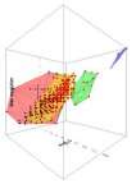


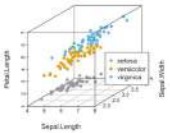
Self-reported weights vs. measured weight



Produces sales vs. time

3-D Data





3D data can be represented by the 3D coordinates of each dot in a 3D scatterplot.

[1] Elmqvist et al. "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation." IEEE transactions on Visualization and Computer Graphics 14.6 (2008): 1539-1148.
[2] http://www.r-bloggers.com/3d-scatterplot-of-iris-data-using-rstudio-and-plot3d/visualization

7

High-dimensional Data

- How to visualize high-dimensional data in visual space (2-D or 3-D) ?

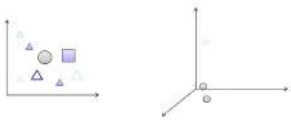
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
6.8	0.31	0.02	19.5	0.026	30	114	0.99268	3.23	0.44	12.4	6	white
5.9	0.225	0.4	4.2	0.03	29	138	0.989	3.57	0.36	12.8	8	white
6.2	0.17	0.36	1.8	0.029	95	161	0.99359	3.65	0.69	12	7	white
6.2	0.271	0.25	5.1	0.047	64	157	0.99608	3.42	0.44	9	5	white
6.6	0.46	0.3	2.8	0.034	31	117	0.98876	3.48	0.39	13.1	6	white
6.4	0.32	0.39	4.3	0.03	31	127	0.99004	3.46	0.36	12.8	8	white
6.6	0.34	0.09	5.3	0.038	52	87	0.99022	3.41	0.4	12.2	7	white
6.3	0.19	0.24	0.85	0.031	80	170	0.99133	3.24	0.41	9	5	white
6.6	0.32	0.35	2.1	0.054	8	93	0.9934	3.3	0.55	13.1	6	red
6.6	0.445	0	5.4	0.050	11	136	0.99424	3.19	0.53	10.2	5	white
6.7	0.6	0.17	2.3	0.058	17	146	0.9952	3.85	0.6	12.5	6	red
6.7	0.67	0.09	1	0.05	5	9	0.99122	3.3	0.54	13.6	7	white
6.7	0.405	0.38	3.9	0.036	53	140	0.98946	3.27	0.85	14	7	white
6.7	0.195	0.29	1	0.042	25	81	0.9908	3.35	0.44	11.3	6	white
6.7	0.61	0.19	5.2	0.038	60	148	0.99212	3.47	0.46	10.5	5	white

Wine Dataset

Source: UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Wine

Simple Solutions

- Add more channel on 2-D or 3-D plots
  - Shape/fill style/color/size of points



9

Simple Solutions

- Multiple coordinated views: present some attributes of objects in a view

[1] Andrew et al. "Leveraging Interaction History for Intelligent Configuration of Multiple Coordinated Views in Visualization Tools LIVEL: Logging Interactive Visualizations & Visualizing Interaction Logs (2016).
[2] https://github.com/livelogs/livelogs

High-Dimensional Data Visualization

- More solutions?

11

One Concrete Example – Show Me Your Visualization Designs?

- Suppose you are given a dataset describing 100 students' profile (anonymized) and academic information, which includes the following attributes:
  - Age
  - Height
  - Weight
  - Travel time to campus
  - Math score
  - Physics score
  - Visual analytics score
  - English language score
  - Algorithm design score
  - Python programming score
  - Art design score
  - Chemistry score

One Concrete Example – Show Me Your Visualization Designs?

Task 1: Show the correlation between any two attributes?

Task 2: Show the overall distribution of all these attributes of all students and help viewers visually identify the similarities between students

- Age
- Height
- Weight
- Travel time to campus
- Math score
- Physics score
- Visual analytics score
- English language score
- Algorithm design score
- Python programming score
- Art design score
- Chemistry score

13

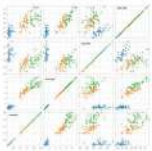
Outline

- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

Scatter-plot Matrix

- 2-D plot for each dimension pair
- Display correlations between dimensions
- The number of 2-D plots is proportional to square of dimensions

Implemented by D3.js

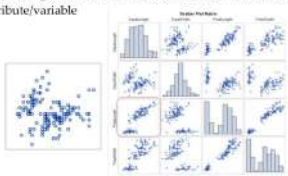


Iris Dataset: https://observablehq.com/@d3/brushable-scatterplot-matrix

15

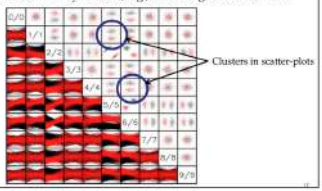
Scatter-plot Matrix

- Scatter-plot matrix can be further enhanced:
  - It can be combined with dimension reduction based projection
  - The diagonal cell can be used to show the distribution of each attribute/variable



Scatter-plot Matrix – Insight Discovery

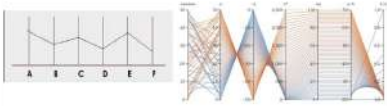
- Scatter-plot matrix can be further enhanced:
  - The cells of the upper triangle and bottom triangle can be used to serve for different analysis tasks, e.g., clustering and correlation



17

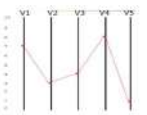
Parallel Coordinates

- It was proposed by Prof. Alfred Inselberg in 1985 for high-dimensional geometry
- Parallel axes
- Data points represented by lines



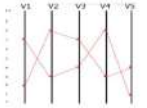
Inselberg, Alfred. "The plane with parallel coordinates." The visual computer 1. no. 2 (1985): 69-91.

Parallel Coordinates

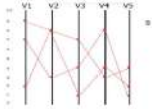


19

Parallel Coordinates

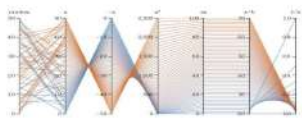


Parallel Coordinates



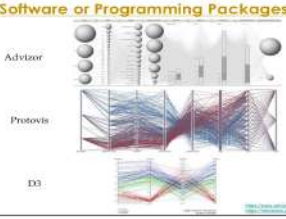
21

Correlations



- The parallel coordinate illustrates the correlation between adjacent variables
- The order of the axis or variables really matters, as appropriate orders can reveal the correlation between adjacent axes more clearly.

Parallel Coordinates in Visualization Software or Programming Packages



23

Parallel Coordinates + Interactions

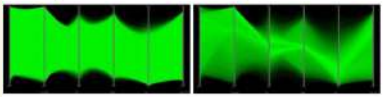
- Selection
- Brush and Filter

Multivariate Analysis Using Parallel Coordinates. Stephen Frey. September 12, 2008. http://www.stevenmfrey.com/multivariate-analysis/

Any Other Variants of Parallel Coordinates?

Transparent Parallel Coordinates

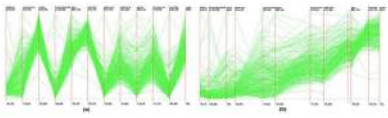
- When there are a huge number of lines in the parallel coordinates, visual clutters occur and it looks messy. A good way to mitigate it is to adjust the opacity of the lines, for example, make those lines half-transparent.



Chad Jones et al. "An Integrated Exploration Approach to Visualizing Multivariate Particle Data. " Computing in Science & Engineering (2009).

Parallel Coordinates: Re-ordering Axes to Reduce Visual Clutter

- By varying the dimension order in a display, it is possible to **reduce clutter** without reducing information content or modifying the data in any way.

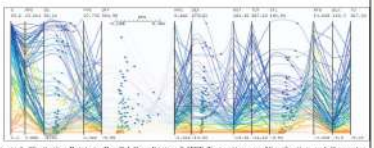


Peng, Wei , M. O. Ward et al. "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Re-ordering." IEEE Symposium on Information Visualization (2005).

27

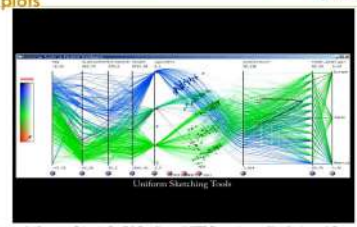
Parallel Coordinates with Scatter-plots

- Augment parallel coordinates with scatterplots to facilitate **data selection** and **data clustering**



Yuan et al. "Scattering Points in Parallel Coordinates. " IEEE Transactions on Visualization and Computer Graphics (2009).

Parallel Coordinates with Scatter-plots



Yuan et al. "Scattering Points in Parallel Coordinates. " IEEE Transactions on Visualization and Computer Graphics (2009).

29

Outline

- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

Glyph-based Methods

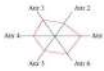
- Radar Chart (a.k.a.: star plot)
- Chernoff Faces

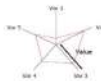
31

Radar Chart

Star Plot

- Space variables around a circle
- Encode values on "spokes"
- Each data point is now a shape

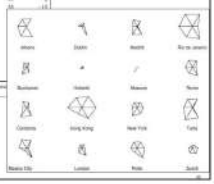




Multivariate Data: Climatic Variables

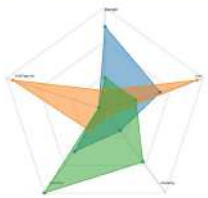
City	Precip. average	Temp. average	Temp. max average	Temp. min average	Record max	Record min
Athens	37	17	21	13	42	−1
Bucharest	59	11	16	5	49	−23
Canberra	62	12	19	6	42	−10
Dublin	74	10	12	8	28	−7
Helsinki	63	5	8	1	31	−36
Hong Kong	218	23	25	21	37	2
London	75	10	13	5	35	−13
Madrid	45	13	20	7	40	−10
Mexico City	63	17	23	11	32	−3
Moscow	59	4	8	1	33	−42
New York	116	12	17	8	40	−18
Paris	170	14	18	10	34	−2
Rio de Janeiro	109	25	30	20	43	7
Rome	80	15	20	11	37	−3
Tunis	44	18	23	13	46	−1
Zürich	107	9	12	6	35	−20

Table 4.1 Annual climatic values in Celsius of some world cities. Values from http://www.weatherbase.com.



33

Star Plot in D3.js



Source: http://gizeos.ch/radar-chart-d3/ and http://bl.ocks.org/nbremer/6506614

Chernoff Faces

- Use face encodings to represent different attributes

Facial Variables

Spinelli, Joseph G., and Yu Zhou. "Mapping quality of life with chernoff faces." In *Proceedings of 24th ESRI International User Conference*, 2004.

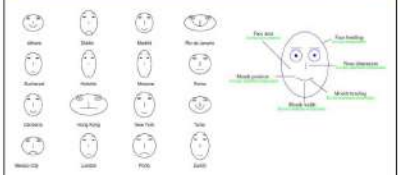
35

Multivariate Data: Climatic Variables

City	Precip. average	Temp. average	Temp. max average	Temp. min average	Record max	Record min
Athens	37	17	21	13	42	−1
Bucharest	59	11	16	5	49	−23
Canberra	62	12	19	6	42	−10
Dublin	74	10	12	8	28	−7
Helsinki	63	5	8	1	31	−36
Hong Kong	218	23	25	21	37	2
London	75	10	13	5	35	−13
Madrid	45	13	20	7	40	−10
Mexico City	63	17	23	11	32	−3
Moscow	59	4	8	1	33	−42
New York	116	12	17	8	40	−18
Paris	170	14	18	10	34	−2
Rio de Janeiro	109	25	30	20	43	7
Rome	80	15	20	11	37	−3
Tunis	44	18	23	13	46	−1
Zürich	107	9	12	6	35	−20

Table 4.1 Annual climatic values in Celsius of some world cities. Values from http://www.weatherbase.com.

Chernoff Faces



Source: The Use of Faces to Represent Points in k-Dimensional Space Graphically (http://www.jstor.org/stable/2286277) and A Critique of Chernoff Faces (http://cmgresearch.org/articles/chernoff-faces)

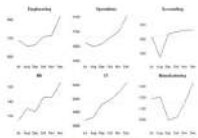
37

Outline

- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

Small Multiples

- A series or grid of small similar graphics or charts for an easy exploration and comparison of multiple dimensions



https://en.wikipedia.org/wiki/Small_multiple

39

Small Multiples

- Example: Use small multiples to visualize state-level support or opposition on the school vouchers across different races and in-come levels in the united states.

http://andrewgelman.com/2009/07/hard_soft_for_hi/

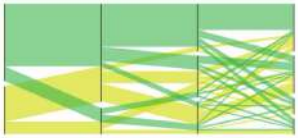
Outline

- What is high-dimensional data
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

41

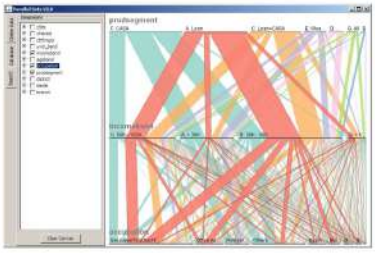
Parallel Sets

- It is mainly designed for visualizing high-dimensional categorical data
- It shows data frequencies instead of the individual data points



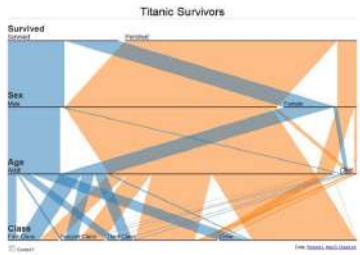
https://datavizcatalogue.com/methods/parallel_sets.html

Exploring High-dimensional Categorical Data with Parallel Sets



43

Parallel Sets – d3.js

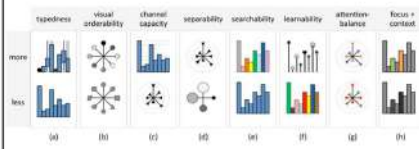


Questions?

Thank You!

45

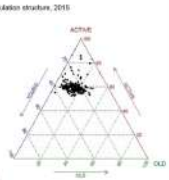
A Survey on Glyph-based Visualization



Borgo R, Kehrer J et al. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications, Eurographics (STARs) (2013)

TERNARY PLOT

- Ternary plots are a way of displaying the distribution and variability of three-part compositional data.
- Its display is a triangle with three components. Each side represents one of the three components.



Reference: https://en.wikipedia.org/wiki/Ternary_plot

47

Dimensionality Reduction

- Project the high-dimensional data onto a lower-dimensional subspace using linear or non-linear transformations.
- Projection preserves important relations (e.g., no information loss, data discrimination).

$$x = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{pmatrix} \rightarrow reduce\ dimensionality \rightarrow \hat{x} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{pmatrix} (K \ll N)$$

Dimensionality Reduction Methods

- Linear methods:
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
- Nonlinear methods:
  - *ISOMAP
  - *Local Linear Embedding (LLE)

49

Data Visualisation

Associate Professor Goh Wooi Boon

College of Computing and Data Science

Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 1 – Introduction

Contents

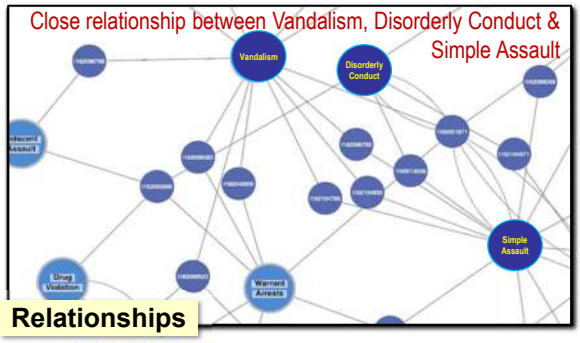
- What is Data Visualisation?
- Importance of Data Visualisation
- What will You Learn in this Course

© A/P Goh Wooi Boon (CCDS/NTU)

Introduction

What is Data Visualisation?

- Data visualization is the **representation** of **data** or **information** in a graph, chart, map or other **visual format**.
- **Visual elements** are used to provide an accessible way to **see** and **understand** trends, outliers, patterns and relationships in the data.



Crime Visualisation from Cambridge Intelligence.com

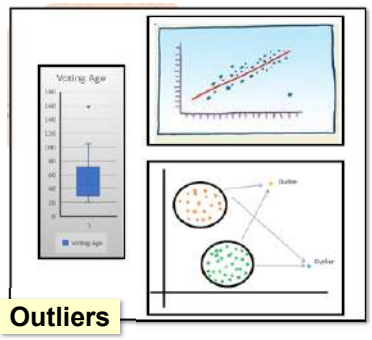


Image from Analyticsvidhya.com

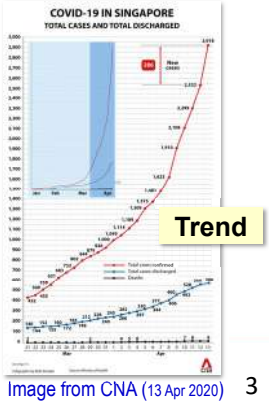


Image from CNA (13 Apr 2020)

3

3

Importance of Data Visualisation

The visual creature

- About half the human brain is directly or indirectly devoted to **visual processing**.
- Humans are much better at remembering **pictures** than words[1]. Students given **visuals** to go along with lectures have also been shown to have **better retention** than students listening to lectures without visual aids[2].
- The McGurk effect seems to also suggest that the **visual** sensory modality can **dominate** the auditory modality.

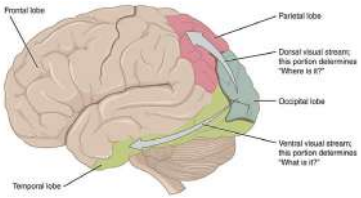


Image from Wikimedia

BBC video on McGurk Effect - https://youtu.be/G-IN8vWm3m0

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Neural correlates of the episodic encoding of pictures & words - https://www.pnas.org/content/95/5/2703.full

[2] Bui, D.C. & McDaniel, M.A. (2012) – Enhancing learning during lecture ... https://www.sciencedirect.com/science/article/abs/pii/S2211368115000200

4

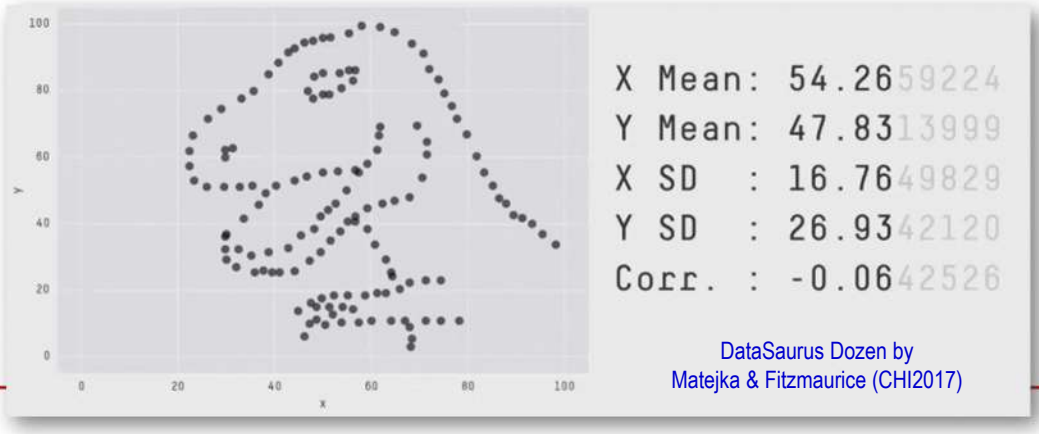
4

Importance of Data Visualisation

Blinded by statistics

- Basic statistical analysis alone may not provide useful insights into how datasets can have **differing characteristics** despite having **similar statistical** parameters.

12 datasets of 2D points with similar statistical parameters



DataSaurus Dozen by Matejka & Fitzmaurice (CHI2017)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

5

5

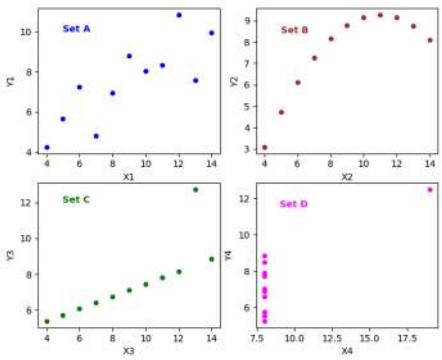
Importance of Data Visualisation

Blinded by statistics

- Basic statistical analysis alone may not provide useful insights into how datasets can have **differing characteristics** despite having **similar statistical** parameters.
- Statistician Francis Anscombe[3] used the **Anscombe's Quartet**[3] of four datasets to show why **visualisation** is a **crucial part of data analysis**.

	Set A		Set B		Set C		Set D	
	X1	Y1	X2	Y2	X3	Y3	X4	Y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Correlation Coefficient	0.82		0.82		0.82		0.82	
Std Dev	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94

Anscombe's Quartet



[3] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21

6

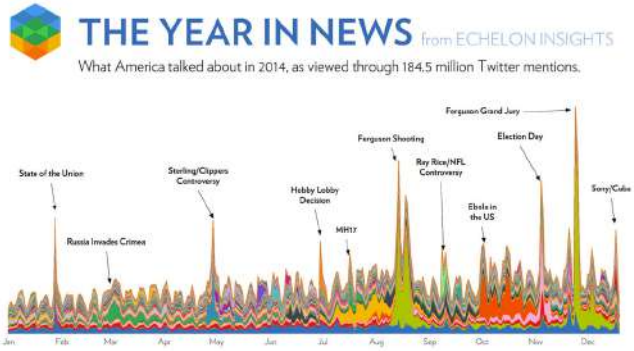
6

Importance of Data Visualisation

Seeing insights in big data

- The amount of data being generated is increasing rapidly. Gathering **meaningful insights** from such **big data** is not possible without good data visualisation.
- Good visualization allows users to easily identify relationships and patterns **hidden inside** the large data sets.

Word cloud - 20+ million customers' comments [4]



Visualisation by Echelon Insights of what America twitted in 2014

[4] Big Data Visualization. Value it brings & techniques in requires - https://www.scnsoft.com/blog/big-data-visualization-techniques

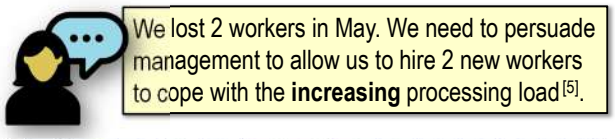
7

7

Importance of Data Visualisation

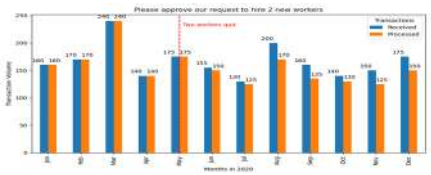
Telling the story

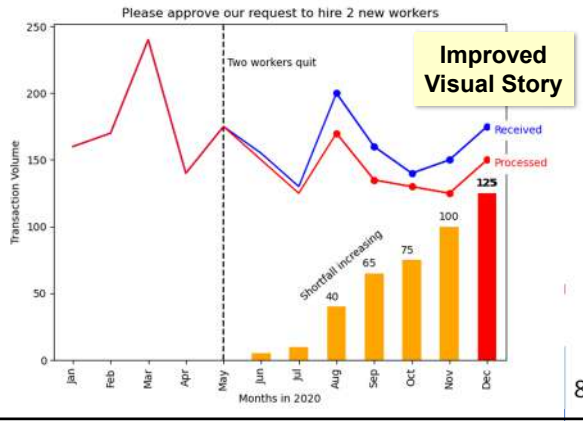
- Data visualisation can help **communicate the message** embedded in the data.
- Telling a visual story that conveys the **intended** message requires designing clear visualisation that understands how **humans perceive** visual information.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Transactions	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2	Received	160	170	240	140	175	155	130	200	160	140	150	175
3	Processed	160	170	240	140	175	150	125	170	135	130	125	150

Excel data of transactions received and processed over year 2020





[5] C.N. Knaflic, Storytelling with data, Wiley (2015)

8

8

Visualising Datasets

Is There An Ideal Way?

- There are many ways to visualise a dataset.
- Choosing the best way to visualise the dataset will depend on:
  - The data **types** (e.g. its scale of measure, spatial (geodata), temporal, etc)
  - The **context** and **relationship** within the data (e.g. highest value, part of a whole)
  - The **purpose** of the visualisation (e.g. identify trends, compare values, show composition, communicate a specific message)

[6]

Chart from: https://www.optimizesmart.com/what-type-of-chart-to-use-to-compare-data-in-excel/

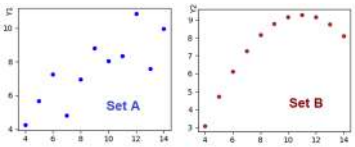
Visualising Datasets

What Is Considered A Good Visualisation?

- A good visualisation shows **connections** within the data that are too complex to explain with words or is **hidden** from non-visual analysis.

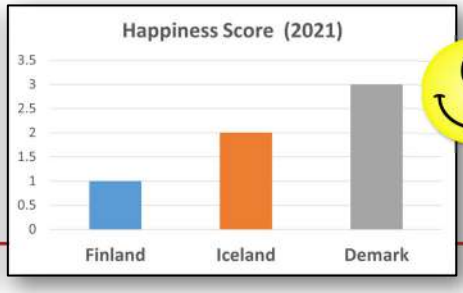
	Set A		Set B	
	X1	Y1	X2	Y2
Mean	9.00	7.50	9.00	7.50
Correlation Coefficient	0.82		0.82	
Std Dev	3.16	1.94	3.16	1.94

From Anscombe's Quartet



- A good visualisation makes it **easier** to quickly understand the information presented and grasps meaningful **insights** embedded within the data.

Poor Visual Perception

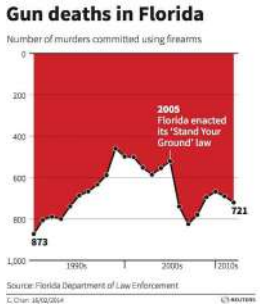


Good Visual Clarity

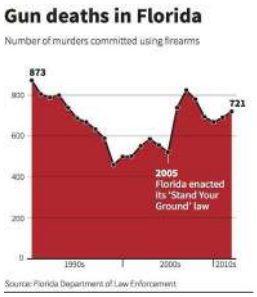
Ethical Visualisation

Staying True

- Ethics is the set of **moral principles** that guides a person's **decision** and **actions**.
- In data visualisation, there are many opportunities to **manipulate** and **distort** the message a person perceives using misleading or untruthful data representation.
- **Ethical conduct** is an important part of **your responsibility** when you design and present visual information.
- Your **professional integrity** requires you to be mindful of how your audience will interpret the information you present in your visualisation.
- So, communicate **clearly** & **truthfully.**



Reuters Original Version [7]



P.A. Fedewa's Ethical Version [7]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[7] BusinessInsider.com, This Chart Shows An Alarming Rise In Florida Gun Deaths After 'Stand Your Ground' Was Enacted - https://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2

11

11

Welcome to Data Visualisation

What will you be learning in this course?

The Person – the theory of human visual perception and its implications in designing effective data visualisation.

The Process – from preparing the data to choosing and implementing the right visualisation solution for displaying and exploring various types of data.

The Practice – solve data visualisation problems through design and code development for different application domains.

The Purpose – learn to design visualisation narratives that effectively communicate your intended purpose and facilitate data analysis.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

12

12

Summary

Introduction

- Data visualisation is the transformation of data into a **visual form** so that **meaningful information** embedded within it can be made **more accessible**.
- Data visualisation is important to any disciplines or individual who **deals with data** and the need to **extract** or **share insights** from these data.
- Effective data visualisation serves two main purposes. It allows us to **explore** the dataset to gain useful insights and it allows us to **communicate** effectively what the data is telling us.

References for Introduction

[1] Neural correlates of the episodic encoding of pictures & words - https://www.pnas.org/content/95/5/2703.full

[2] Bui, D.C. & McDaniel, M.A. (2012) – Enhancing learning during lecture... https://www.sciencedirect.com/science/article/abs/pii/S2211368115000200

[3] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21

[4] Big Data Visualization. Value it brings & techniques in requires - https://www.scnsoft.com/blog/big-data-visualization-techniques

[5] C.N. Knaflic, Storytelling with data, Wiley (2015)

[6] Chart from: https://www.optimizesmart.com/what-type-of-chart-to-use-to-compare-data-in-excel/

[7] BusinessInsider.com, This Chart Shows An Alarming Rise In Florida Gun Deaths After 'Stand Your Ground' Was Enacted - https://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2

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

Data Visualisation

Associate Professor Goh Wooi Boon

School of Computer Science and Engineering

Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 2 – Data Attributes & Wrangling

Contents

- Basic Data Attributes
- Other Data Attributes
- Data Wrangling

© A/P Goh Wooi Boon (SCSE/NTU)

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.

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

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

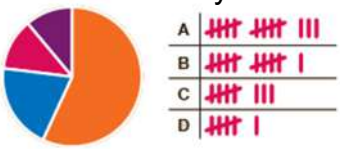
5

5

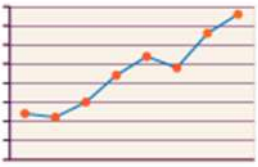
Basic Data Attributes

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

Level of Measurement

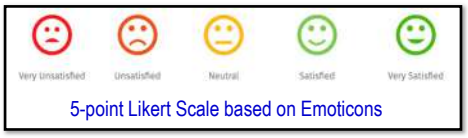
Nominal

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

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).



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.

Level of Measurement

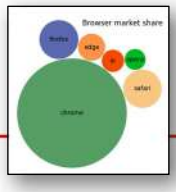
Interval

- 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**.

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 (= , ≠, < , > , + , – , × and ÷).
- 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.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

11

11

Review

The NOIR Scale of Measure

- Which scale represents each of the following data attributes:

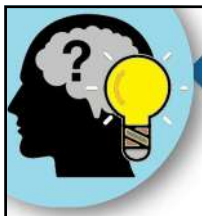
1) The different types of coffee beans -
2) Year coffee bean was harvested -
3) Weight of coffee beans -
4) Size of my latte {tall (12 ounces), grande (16), venti (20)} -

Nominal **Ordinal** **Interval** **Ratio**

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

12

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

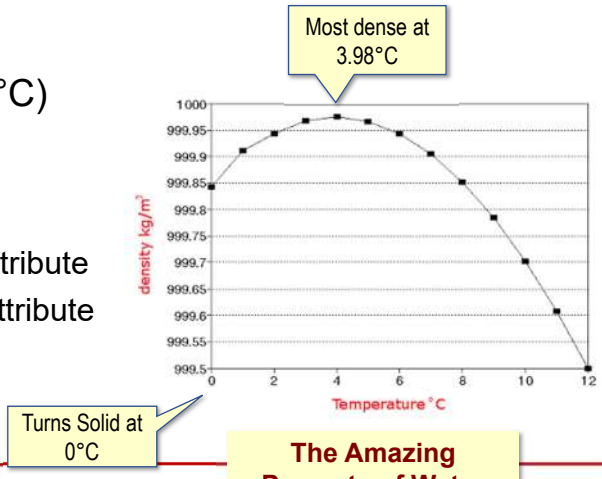
13

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



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.

Chapter 2.2 – Other Data Attributes

Contents

- Data Dimensionality
- Hierarchy
- Temporal
- Spatial

© A/P Goh Wooi Boon (SCSE/NTU)

Data Dimensionality

Relational Data Model

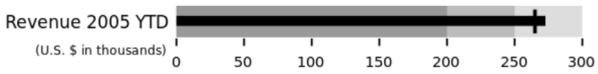
- Structured data typically consist of **data observations** that are represented by rows and **data attributes** that are represented by columns.
- A relational data model represents data as a **table**.
- Each column can be called a **dimension** of the dataset and these columns or attributes define 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

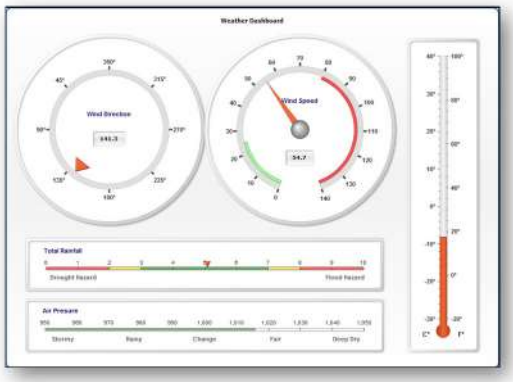
Data Dimensionality

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/

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

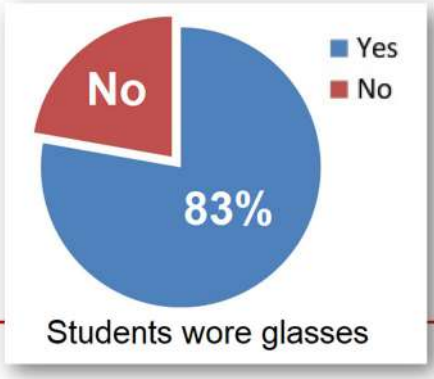
18

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

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

19

19

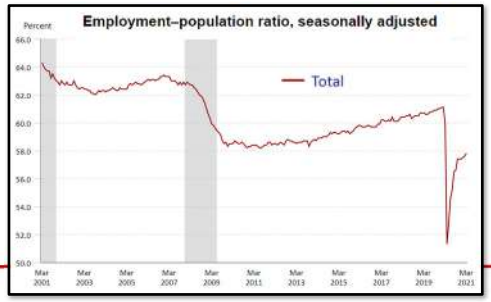
Data Dimensionality

2-Dimension

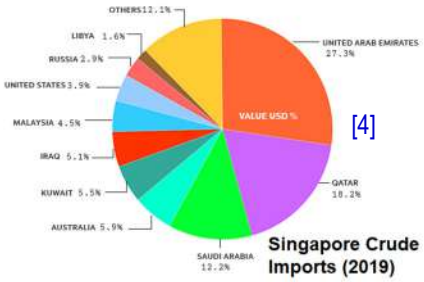
- This type of dataset consists of data with **two different** but **related attributes**.
- The appropriate ways to visualise such datasets will depend on the nature of the **scales of measure** of these two attributes. For example, if one scale is **ratio** and the other **ordinal**, then a **line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



[4]

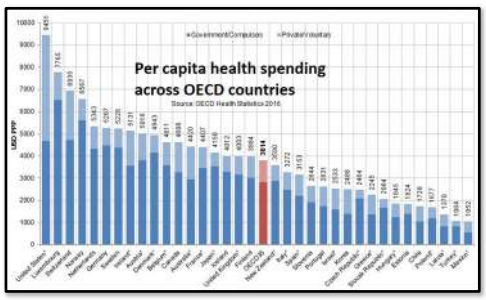
Data Dimensionality

3-Dimension

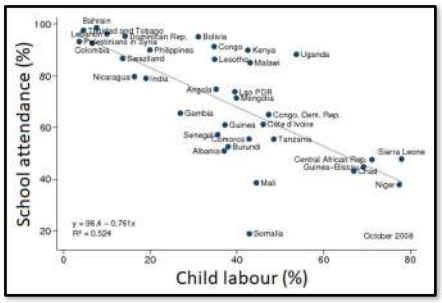
- As the data dimension increases to three, **other visual attributes** (e.g. colour, size, shape) may be needed to allow all dimensions to be visualised together.
- The **scale of measure** and nature of the **relationship** between dimensions will affect the choice of visualisation techniques.

[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]

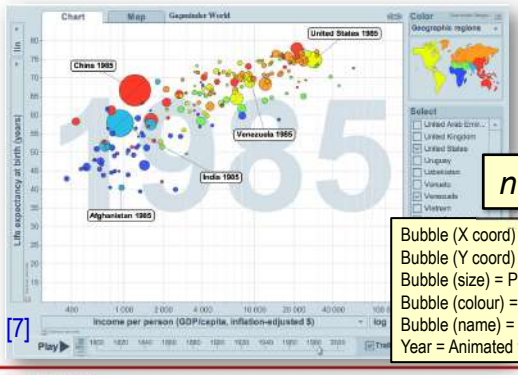


[6]

Data Dimensionality

n-Dimension

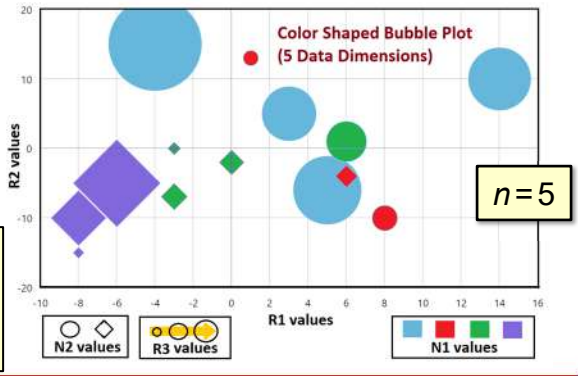
- Visualising *n* data dimensions simultaneously will require using a combination of visual attributes that matches the respective scale measure of each data dimension (e.g. colour & shape for nominal scale. Size, height, position for interval or ratio scale).



[7]

n = 6

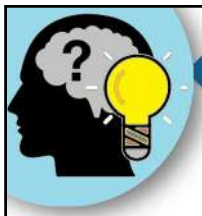
Bubble (X coord) = Income / person,
Bubble (Y coord) = Life expectancy,
Bubble (size) = Population,
Bubble (colour) = Continent,
Bubble (name) = Country,
Year = Animated time-stamp.



n = 5

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

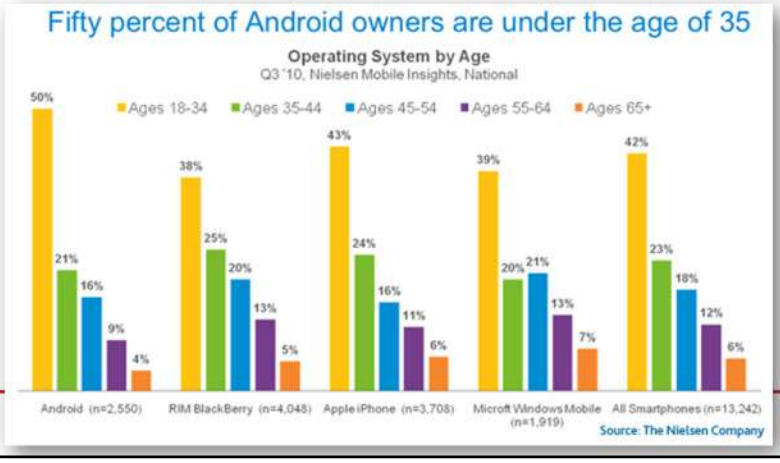
[8] Video taken from BBC News at https://www.youtube.com/watch?v=jbkSRLYSojo



Think and Apply

Identifying the Dimensions

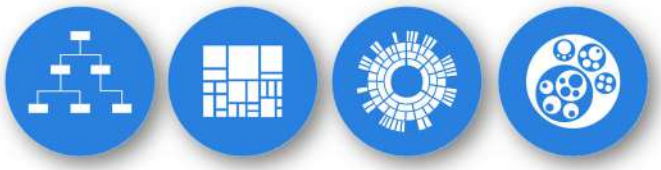
- How many data dimensions is simultaneously visualised in the Clustered Bar chart shown below?
- What would the table for this dataset look like?



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].



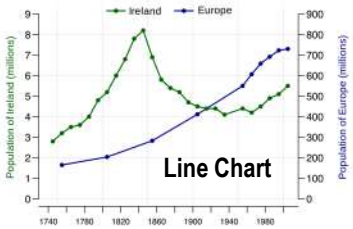
Tree Diagram | Treemap | Sunburst Diagram | Circle Packing

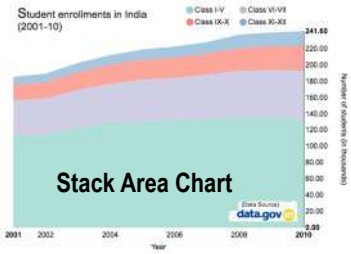
[9] For examples of hierarchical visualisations see - https://insightwhale.medium.com/how-to-show-hierarchy-with-data-visualization-526fb45ee4c2

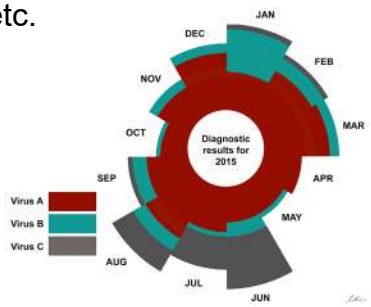
Temporal

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.







Polar Area Diagrams

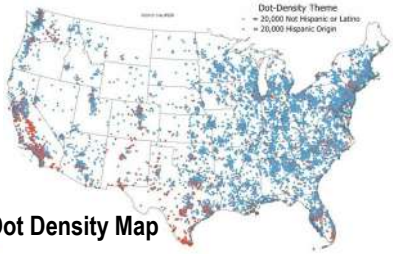
[10] For examples of hierarchical visualisations see - https://humansofdata.atlan.com/2016/11/visualizing-time-series-data/

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

Summary

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.

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

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

Chapter 2.3 – Data Wrangling

Contents

- What is and Why the need for Data Wrangling
- Pandas and Dataframes
- Combining Dataframes
- Reshaping Dataframes
- Extracting Subsets
- Sorting Values and Index
- Handling Missing Values

© A/P Goh Wooi Boon (SCSE/NTU)

Chapter 2.3 – Data Wrangling

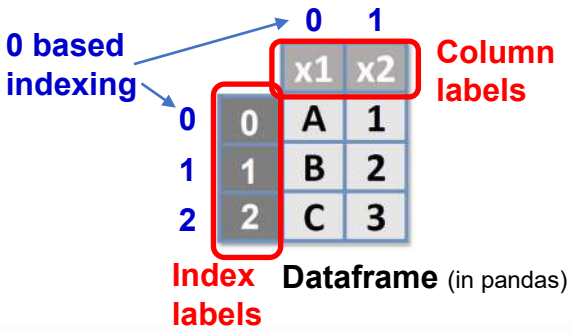
What is and Why the Need for Data Wrangling?

- Data wrangling is the process of **cleaning**, **structuring** and **enhancing** raw data into a format that is more appropriate and readily used for analysis & visualisation.
- The data available is often not in a **suitable form** for use with data visualisation tools. Some examples include:
  - Data needed are stored in **different files**. They need to be retrieved and combined into a single table for analysis.
  - Data available is extensive. Only a **subset is needed** for analysis.
  - Data in the table is **not organised** in the manner suitable for analysis.
  - Data in the table has **missing** or **outlier values** that can affect the analysis.
  - **New enhanced data** (e.g computed mean) must be entered in the table for analysis.

Data Wrangling - Getting Started

Pandas and Dataframes

- **Pandas** is a powerful and easy to use **open-source** data analysis and manipulation **library**, built on top of the **Python** programming language.
- **Dataframes** are used in pandas for storing data in rows (observations) and columns (variables or dimensions).
  - **Columns** are dimension of observations (variables) and are given a **label** (e.g. **x1**).
  - The row labels are referred to as **index** and has default labels of (0, 1, 2..) but can be renamed to any label.
  - Dataframes uses **0 based indexing**.

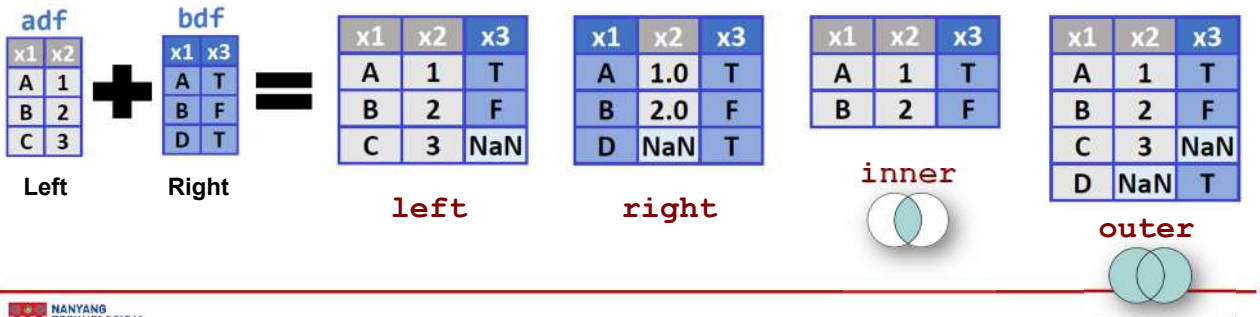


[1] Pandas – Python Data Analysis Library - https://pandas.pydata.org/

Combining Dataframes

Merge

- Data from two dataframes can be combined into a dataframe in many ways using the `merge` operator.
- Two dataframes can be **merged** using qualifiers like **left**, **right**, **inner** and **outer**.
- **Missing values** in the merged dataframe are given the **NaN** (Not a number) values.

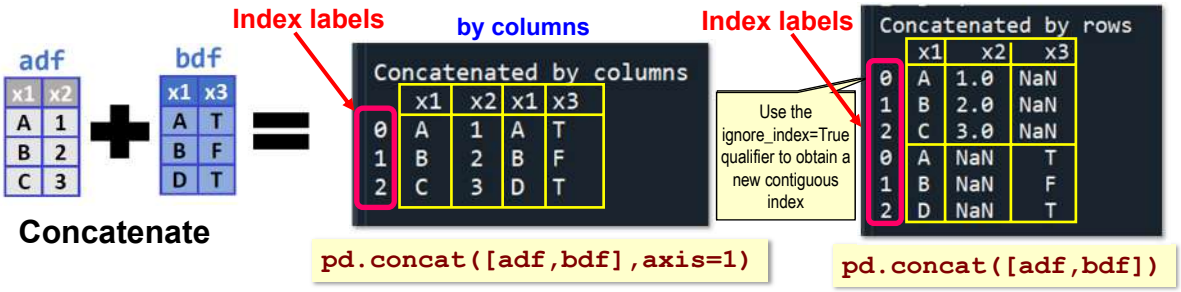


[2] For syntax of **merge** - see Data Wrangling with Pandas (Cheat Sheet) - https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

Combining Dataframes

Concatenate

- Data from 2 dataframes (or the same frame itself) can be appended using `concat`.
- The dataframes can be appended based on **rows** or **columns** (use `axis=1`)
- **Missing values** in the concatenated dataframe are given the **NaN** values.



[3] For syntax of **concat**- see pandas ref at - https://pandas.pydata.org/docs/reference/api/pandas.concat.html

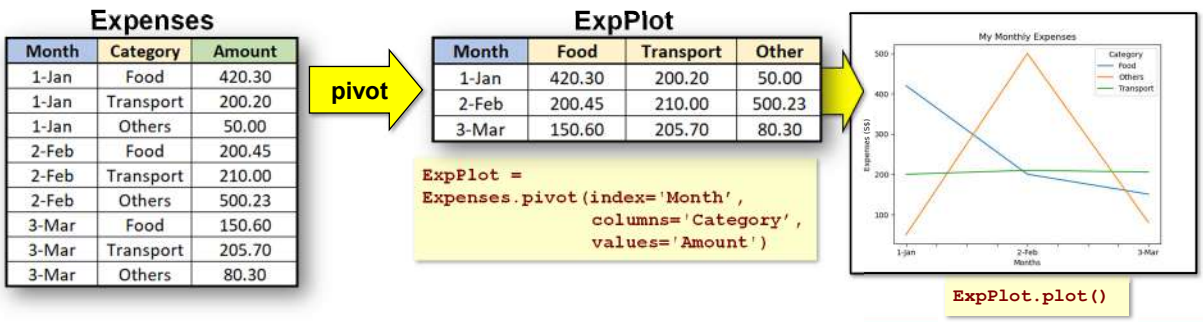
5

5

Reshaping Dataframes

Pivot

- Data reshaping **rearranges** the form of the data **without changing** the content of the dataset so as to facilitate the next stage of analysis (e.g. plotting a graph).
- The `pivot` operator reshapes a table in the **long form to** a **wide form** that is more amenable to analysis as the desired variables are in columns.




```
ExpPlot =
Expenses.pivot(index='Month',
               columns='Category',
               values='Amount')
```

[4] For syntax of **pivot** - see pandas ref at - https://pandas.pydata.org/docs/user_guide/reshaping.html

6

6

Reshaping Dataframes

Melt

- The `melt` operator reshapes a table in the **wide form to** a **long form**.
- It creates a dataframe where one (or more) columns are **identifier variables**, while all other columns (considered measured variables), are unpivoted to the row axis to form a **variable** and a **value** column.

Temp

Name	Mon	Tue	Wed	Thu	Fri	Sat	Sun
John	36.5	36.8	36.4	35.8	36.0	36.5	35.8
Ahmad	36.5	36.6	36.7	36.8	36.6	36.7	36.8
Ginny	35.5	36.0	36.2	37.0	37.5	37.0	38.3
Manish	37.5	37.0	36.5	37.3	36.6	36.5	36.0

```
TempTable = Temp.melt(id_vars='Name')
```

melt

Name	Variable	Value
John	Mon	36.5
Ahmad	Mon	36.5
Ginny	Mon	35.5
Manish	Mon	37.5
John	Tue	36.8
Ahmad	Tue	36.6
Ginny	Tue	36.0
Manish	Tue	37.0
John	Wed	36.4
Ahmad	Wed	36.7
Ginny	Wed	36.2
Manish	Wed	36.5
:	:	:
John	Sun	35.8
Ahmad	Sun	36.8
Ginny	Sun	38.3
Manish	Sun	36.0

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[5] For syntax of **melt** - see pandas ref at - https://pandas.pydata.org/docs/reference/api/pandas.melt.html

7

7

Think and Apply

Reshaping Dataframe with Pivot and Melt

- The table in the dataset "*Daily Temperature.csv*" list the temperatures of four students in columns, according to each day of the week.
- Create a chart of daily temperatures of each student, each with their own line plot.

Temp

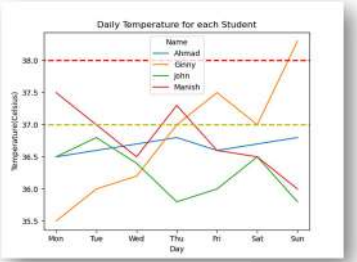
Name	Mon	Tue	Wed	Thu	Fri	Sat	Sun
John	36.5	36.8	36.4	35.8	36.0	36.5	35.8
Ahmad	36.5	36.6	36.7	36.8	36.6	36.7	36.8
Ginny	35.5	36.0	36.2	37.0	37.5	37.0	38.3
Manish	37.5	37.0	36.5	37.3	36.6	36.5	36.0

Reshape **Temp** so that each student's daily temperature in a column.

TempPlot

Name	Ahmad	Ginny	John	Manish
Day				
Mon	36.5	35.5	36.5	37.5
Tue	36.6	36.0	36.8	37.0
Wed	36.7	36.2	36.4	36.5
Thu	36.8	37.0	35.8	37.3
Fri	36.6	37.5	36.0	36.6
Sat	36.7	37.0	36.5	36.5
Sun	36.8	38.3	35.8	36.0

plot



Use **TempPlot.plot()** to create the line chart

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

8

8

Extracting Subsets

Data Slicing

- Most datasets have more data then is required for the current analysis. Data slicing refers to the process of extracting the **required subset of data** from the dataset.
- Data slicing in pandas can be done using operators `loc` and `iloc`.

Column Index 0 1 2 3 4 5 6 7

Temp

Index	Name	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	John	36.5	36.8	36.4	35.8	36.0	36.5	35.8
1	Ahmad	36.5	36.6	36.7	36.8	36.6	36.7	36.8
2	Ginny	35.5	36.0	36.2	37.0	37.5	37.0	38.3
3	Manish	37.5	37.0	36.5	37.3	36.6	36.5	36.0

loc — Uses labels

	Name	Sat	Sun
0	John	36.5	35.8
1	Ahmad	36.7	36.8
2	Ginny	37.0	38.3
3	Manish	36.5	36.0

Students' temperature over the weekend

```
Temp.loc[:,['Name','Sat','Sun']]
```

iloc — Uses index number

Extract students with any temperature > 37°C

	Name	Mon	Tue	Wed	Thu	Fri	Sat	Sun
2	Ginny	35.5	36.0	36.2	37.0	37.5	37.0	38.3
3	Manish	37.5	37.0	36.5	37.3	36.6	36.5	36.0

```
Temp.loc[(Temp.iloc[:,1:]>37).any(axis=1)]
```

	Mon	Tue	Wed	Thu	Fri
0	36.5	36.8	36.4	35.8	36.0
3	37.5	37.0	36.5	37.3	36.6

```
Temp.loc[[0,3],'Mon':'Fri']
```

Temperatures of 1st and 4th students in the weekdays

[6] Pandas reference for data slicing- https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html

Rearranging Data

Sorting values and index

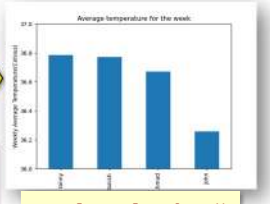
- Data in a table may need to be sorted based on one or more selected **attributes** to facilitate further processing (e.g. plotting an ascending bar chart)
- Pandas provides the `sort_values` and `sort_index` operators for this purpose.

Temp

Index	Name	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	John	36.5	36.8	36.4	35.8	36.0	36.5	35.8
1	Ahmad	36.5	36.6	36.7	36.8	36.6	36.7	36.8
2	Ginny	35.5	36.0	36.2	37.0	37.5	37.0	38.3
3	Manish	37.5	37.0	36.5	37.3	36.6	36.5	36.0

Get mean & sort

Ginny	36.785714
Manish	36.771429
Ahmad	36.671429
John	36.257143




```
BarPlot.plot.bar()
```

sort

	Name	Mon	Tue	Wed	Thu	Fri	Sat	Sun
3	Manish	37.5	37.0	36.5	37.3	36.6	36.5	36.0
0	John	36.5	36.8	36.4	35.8	36.0	36.5	35.8
2	Ginny	35.5	36.0	36.2	37.0	37.5	37.0	38.3
1	Ahmad	36.5	36.6	36.7	36.8	36.6	36.7	36.8

```
Temp.sort_index()
```

Sort index back to original order

```
Temp.sort_values(by='Name',ascending=False)
```

Sort Name in descending order

Get mean over numeric columns then sort in descending order

```
BarPlot = Temp.mean(axis=1)
:
BarPlot.sort_values(ascending=False)
```


[7] Ref to pandas sorting - https://realpython.com/pandas-sort-python/ & https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sort_values.html

Handling Missing Values

Pandas operators for NaN values

- Raw datasets and reshaped dataframes often contain missing or **null values** (NaN). These null values need to be accounted for and replaced/removed before further processing or data analysis.
- Pandas provide various operators for **detecting** and **handling** null values.

df

x1	x2
A	1
B	NaN
C	3

Returns **True** for null values

	x1	x2
0	False	False
1	False	True
2	False	False

```
df.isnull()
```

Remove rows with null values

	x1	x2
0	A	1.0
2	C	3.0

```
df.dropna()
```

Remove col with null values

	x1
0	A
1	B
2	C

```
df.dropna(
    axis=1)
```

Fill null values with 0

	x1	x2
0	A	1.0
1	B	0.0
2	C	3.0

```
df.fillna(0)
```

Fill null values with mean of col

	x1	x2
0	A	1.0
1	B	2.0
2	C	3.0

```
df.fillna(
df.mean())
```

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[8] Pandas reference for handling missing data - https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

11

11

Think and Apply

Data Wrangling Exercise

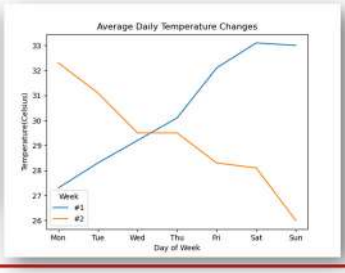
- Two temperature tables for week #1 and #2 are stored in **two different datafiles**.
- **Missing value** needs to be interpolated from its nearest data points.
- **Combine** and **reshape** the two dataframes to produce **two different line plots**. One shows the two weeks separately, the other shows the two as a single line.

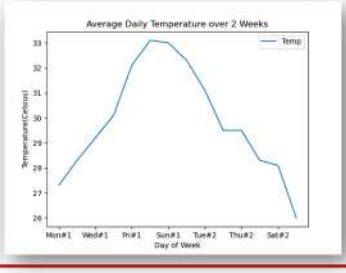
Day	Week	Temp
Mon	#1	27.3
Tue	#1	28.3
Wed	#1	29.2
Thu	#1	30.1
Fri	#1	32.1
Sat	#1	33.1
Sun	#1	33.0

Temp1

Day	Week	Temp
Mon	#2	32.3
Tue	#2	31.1
Wed	#2	29.5
Thu	#2	29.5
Fri	#2	28.3
Sat	#2	28.1
Sun	#2	26.0

Temp2





NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

12

12

Summary

Data Wrangling

- Data wrangling is an important step in the process of data visualisation as it **prepares** that data into a **format** that can be **readily plotted** with the various visualisation routines.
- Data created or loaded from various stored sources need to be **combined** into a suitable single data structure called a **dataframe**.
- Dataframes often need to be **reshaped** into the **appropriate format** so that it is suitable for used with different visualisation routines.
- Data in dataframes may also need to be **sorted** into a specific order for display.
- **Missing data** need to be **filled in** or **removed** in order to ensure they do not **interfere** with the visualisation process.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

13

13

References – Data Attributes and Wrangling

[1] Pandas – Python Data Analysis Library - https://pandas.pydata.org/

[2] For syntax of merge - see Data Wrangling with Pandas (Cheat Sheet) - https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

[3] For syntax of concat- see pandas ref at - https://pandas.pydata.org/docs/reference/api/pandas.concat.html

[4] For syntax of pivot - see pandas ref at - https://pandas.pydata.org/docs/user_guide/reshaping.html

[5] For syntax of melt - see pandas ref at - https://pandas.pydata.org/docs/reference/api/pandas.melt.html

[6] Pandas reference for data slicing- https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html

[7] Ref to pandas sorting - https://realpython.com/pandas-sort-python/ & https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sort_values.html

[8] Pandas reference for handling missing data - https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

14

14

Data Visualisation

Associate Professor Goh Wooi Boon
School of Computer Science and Engineering
Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 3 – Visual Encoding

Contents

- Visual Encoding
- Deconstructions

© A/P Goh Wooi Boon (SCSE/NTU)

Chapter 3.1 – Visual Encoding

Contents

- Visual Variables
- Visual Marks
- Visual Encoding Examples

© A/P Goh Wooi Boon (SCSE/NTU)

The Data Visualisation Process

An Overview

- Data visualisation is the transformation of data into an image using an appropriate visual encoding strategies.

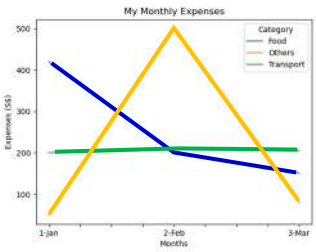
Month	Category	Amount
1-Jan	Food	420.30
1-Jan	Transport	200.20
1-Jan	Others	50.00
2-Feb	Food	200.45
2-Feb	Transport	210.00
2-Feb	Others	500.23
3-Mar	Food	150.60
3-Mar	Transport	205.70
3-Mar	Others	80.30

Data

Data & Conceptual Models

Visual Encoding

Mapping from data to image



Image

Visual Marks and Channels (Variables)

Lines **Colour**

Visual Variables

Bertin's Visual Variables

- Jacque Bertin, a French cartographer proposed set of **visual variables** in 1967 that can be used to construct map symbols and other graphical techniques.

"A graphic is not only a drawing; it is a responsibility, sometimes a weighty one, in decision-making."

- Jacque Bertin
French Cartographer (1918-2010)

Image from Wikipedia

[1] Bertin's quote from: J. Bertin, graphics and graphic information processing (1981), https://books.google.com.sg/books?id=csqX_xnm4tcC&printsec=copyright&redir_esc=y#v=onepage&q&f=false

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

5

5

Visual Variables

Bertin's Visual Variables

- Jacque Bertin, a French cartographer proposed set of **visual variables** in 1967 that can be used to construct map symbols and other graphical techniques.
- Bertin identified **seven** main categories of visual variables:
  - position
  - size
  - shape
  - value (intensity)
  - colour
  - orientation
  - texture

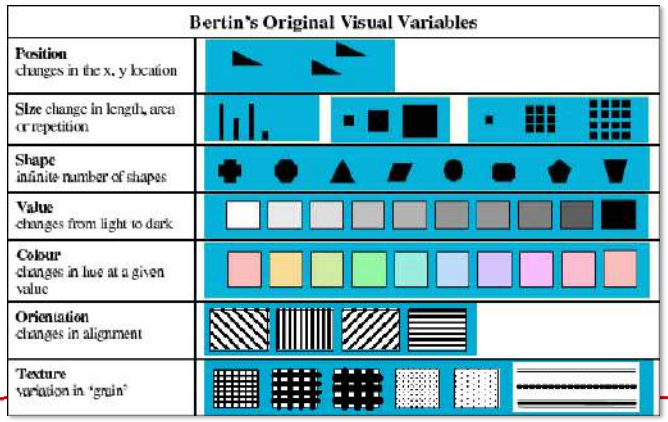


Image from https://infovis-wiki.net/wiki/Visual_Variables

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

6

6

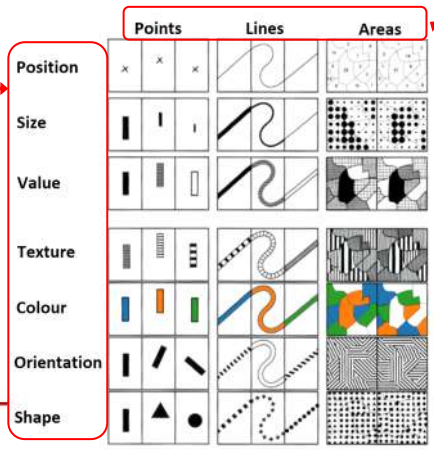
Visual Variables

Marks and Channels

- Basic graphic elements (**visual marks**) like points, lines and areas can then be modified with different **perceptual channels** (visual variables) to encode information.

Channels (Visual variables)
Control the appearance of the visual marks

Recap: What's their difference?
Marks are basic geometric elements that depicts items or links. **Channels** control their appearance.



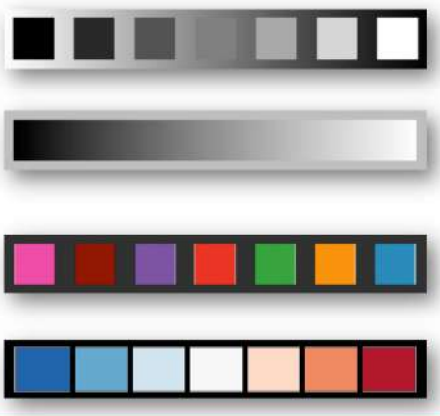
Visual marks
Basic graphical elements in an image

Image adapted from https://graphworkflow.com/retinal/
Bertin's taxonomy of visual encodings from The Semiology of Graphics

Visual Encoding

Encoding Effectiveness of Perceptual Channels

- The effectiveness in encoding different **NOIR** data scales depends on the channel.
- **Value** (intensity) is perceived as **ordered** and can encoded **ordinal** scale when suitably **discretised**.
- **Value**, when **continuously varying**, can encode **interval/ratio** scales but **lacks precision** due to our poor brightness discrimination ability.
- **Hue** (colour) is normally perceived as **unordered** and can only encode **nominal** scale variables.
- **Hue** and **value,** suitably combined, can encode **ordered** variables with a meaningful **mid-point**.



Visual Variables

Bertin's Levels of Organisation

- Bertin's levels of organisation illustrates the degree of effectiveness in encoding the various **NOIR** data scales.

Nominal
Ordinal
Quantitative

Note 1: **Interval** and **Ratio** is classified as a **Quantitative** scale.

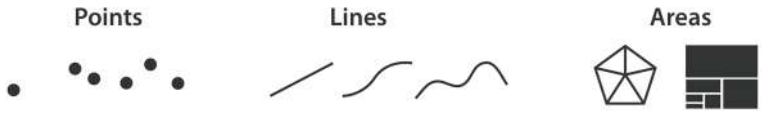
Note 2: **Size** indicates **degree** of encoding effectiveness

Position	**N**	**O**	**Q**
Size	**N**	**O**	**Q**
Value	**N**	**O**	**Q**
Texture	**N**	**o**	
Colour	**N**		
Orientation	**N**		
Shape	**N**		

Visual Marks

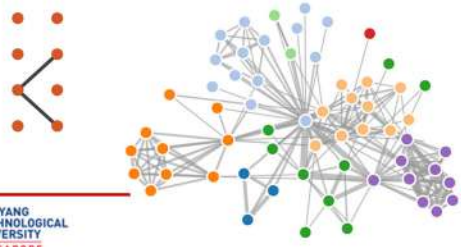
Types of Marks

- **Item marks** – are basic geometric objects that depict **items** on a graph[2].

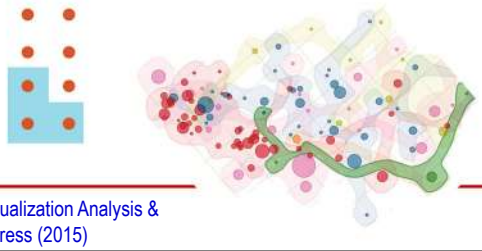


- **Link marks** – show relationship between items[2].

Connection - shows pairwise relationships



Containment - shows group relationships



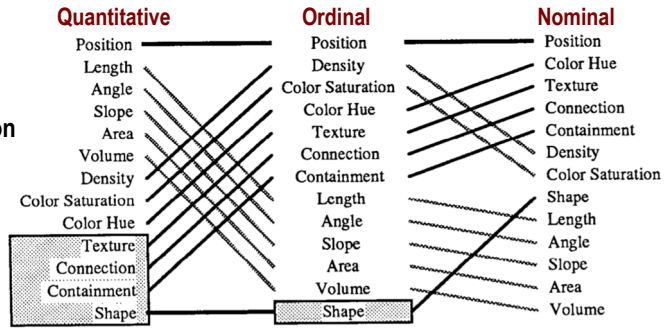
[2] T. Munzner, Visualization Analysis & Design, CRC Press (2015)

Visual Encoding

Mackinlay's Visual Variables

- **Jock Mackinlay**[3] expanded Bertin's variables and proposed an **effectiveness ranking** based on the NOIR data scales.

Note:
Effectiveness is measured by how readily the **information conveyed** by a visualisation is **perceived** by the intended audience.



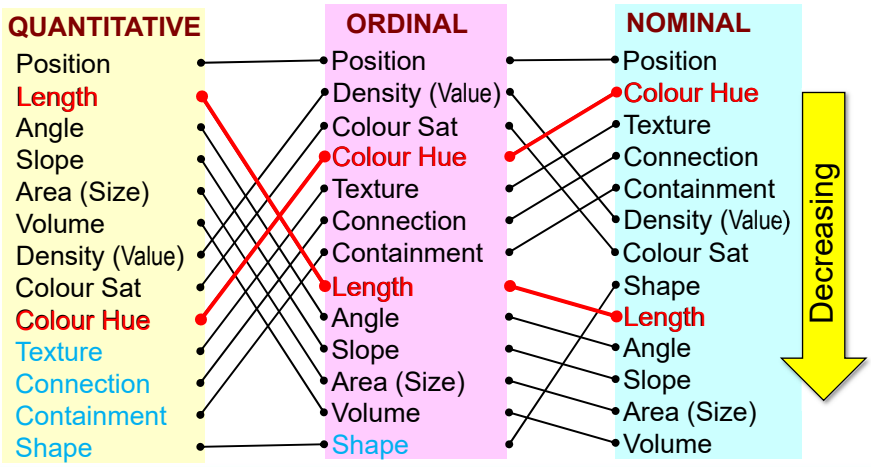
Ranking of perceptual tasks. Tasks shown in the gray boxes are not relevant to these types of data.

Jock D. Mackinlay
Information Visualisation
Expert and Vice President of
R&D at Tableau Software
(Image from Twitter)

[3] J. Mackinlay, Automating the Design of Graphical Presentation of Relational Information (1986), https://research.tableau.com/sites/default/files/p110-mackinlay.pdf

Visual Encoding

Mackinlay's Effectiveness Ranking[3]



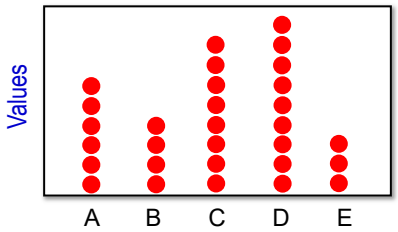
[3] J. Mackinlay, Automating the Design of Graphical Presentation of Relational Information (1986), https://research.tableau.com/sites/default/files/p110-mackinlay.pdf

Visual Encoding Examples

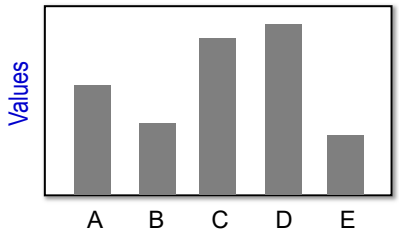
Univariate Data

- In visual encoding of univariate datasets, common visual marks used are **lines** (e.g. line and bar charts) and **points** (e.g. dot plot).

	A	B	C	D	E	Variables
1						Values



Dot Plot (Points)



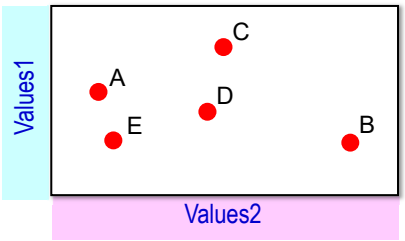
Bar Chart (Lines)

Visual Encoding Examples

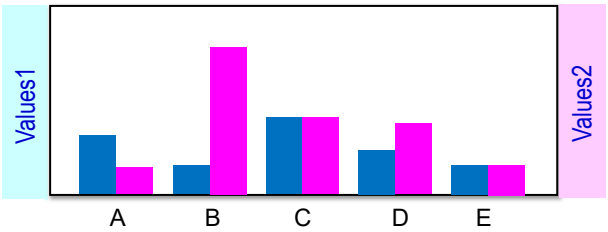
Bivariate Data

- In visual encoding of bivariate datasets, the most common visual mark used are **points** in 2D **scatter plot**. **Line** mark may be used, such as with a **dual Y-axis bar charts**.

	A	B	C	D	E	Variables
1						Values1
2						Values2



2D Scatter Plot (Points)

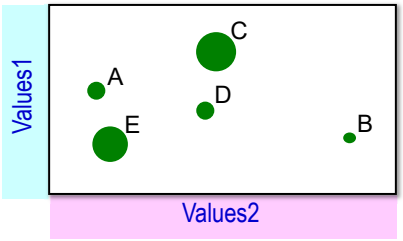


Dual-axis Bar Chart (Lines)

Visual Encoding Examples

Trivariate Data

- Trivariate datasets are commonly encoded using **points**. The two axes of a **2D scatter plot** can encode two quantitative (Q) data type and **area** of the dots can encode the third ordinal or Q data type.



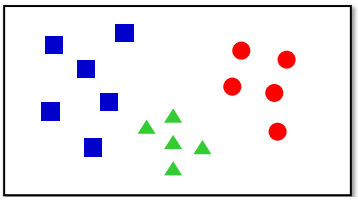
2D Scatter Plot
(with size of dots representing Values3)

	A	B	C	D	E	Variables
1						Values1
2						Values2
3						Values3

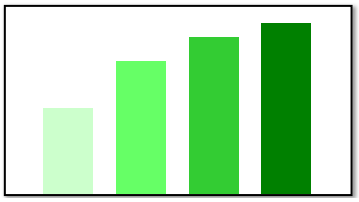
Visual Encoding Examples

Redundant Encoding

- Redundant encoding is the use of **more than one graphical** or visual variable (e.g. color + shape) to encode/represent **one variable of data**.
- Redundancy allows **more perceptual channels** to get the same information into our brain. This helps information to be perceived **faster**, **easier** and more **accurately**[4]. It should be considered if there are leftover unused visual variables.



2D Scatter Plot Example
(Redundant **Colour** + **Shape**)



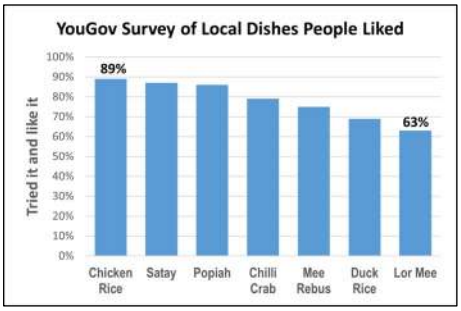
Bar Chart Example
(Redundant **Length** + **Value**)

[4] C. Ware, Information Visualization: Perception for Design, 3rd Edition, Morgan Kaufmann (2013),

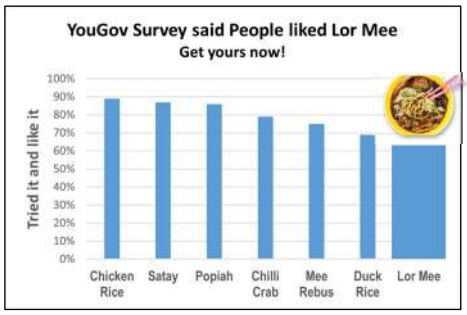
Ethical Visualisation

Use fair and consistent visual encoding

- Your stall specialises in Lor Mee and you saw the results of the recent YouGov local food survey. You decided to share these results with your potential customers.



The YouGov survey results of selected local dishes



Your visualisation of the YouGov food survey

Has ethical visualisation been violated here?

Note: Height and width are integral perceptual dimensions. People tend to see the size of the bar as representing the data value[5].

Note: Food survey figures from: https://sg.yougov.com/en-sg/news/2019/08/13/classic-singaporean-cuisine-ranked-singaporeans/

[5] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

Think and Apply

Visual Encoding Exercise

- Sketch a possible chart that uses an appropriate visual encoding design for each of the data tables. Consider the given data attributes carefully in your encoding design.

Living Members from Class of '45

	Names	Gender	Age
1			
2			
3			
4			
5			
6			

Visual mark – **lines**

Orientation T-shirts - Comparing Halls

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

Visual mark – **points**

COVID-19 Infection in 2020

	Dates	Total Cases	Warded	ICU
1				
2				
3				
:				
:				
365				

Visual mark – **areas**

Summary

Visual Encoding

- Data visualisation is the transformation of **data** into an **image** that is **informative** and conveys **useful insights** embedded within the data.
- **Visual encoding** is the process of **mapping** these data into that image.
- Based on the attributes of the various dimensions of the dataset, appropriate combinations of visual **marks** and **variables** (channels) are used to encode the data.
- Effective visual encoding requires a good understanding of the characteristics and limitations of **human visual perception**.

Chapter 3.2 – Deconstruction

Contents

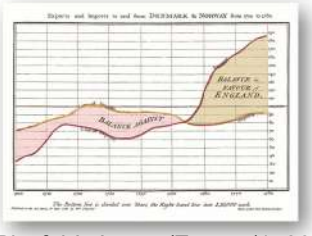
- Playfair's Import/Export (Line Charts)
- Wattenberg's Map of the Market (Tree Map)
- Fry's Basketball Salary versus Performance (Slope Chart)

© A/P Goh Wooi Boon (SCSE/NTU)

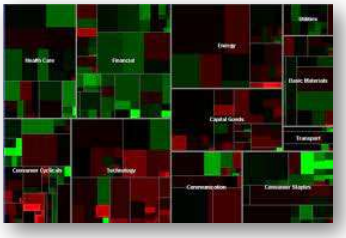
Deconstruction

Seeing How Others Do It

- The process of deconstruction allows us to **analyse** how others have visually encoded the information they wanted to convey in their dataset.
- Deconstruction is a great way to **learn how to construct** your own visualisation, especially if the examples analysed have been effective in telling their visual story.

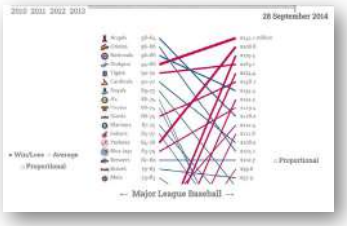


Playfair's Import/Export (1786)
A good old pre-digital chart



Wattenberg's Map of the Market (1998)
A visualisation still current in use

checkout: https://finviz.com/map.ashx

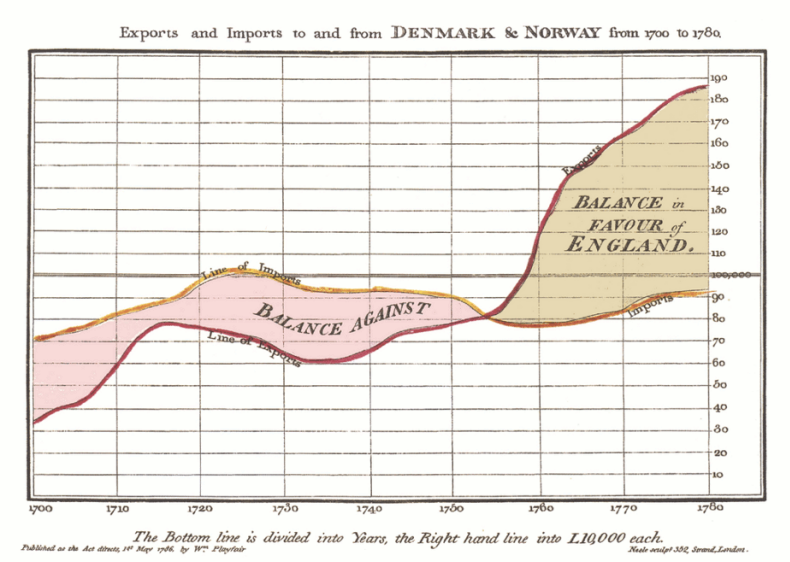


Fry's Salary vs Performance (2014)
Visualisation for US Basketball fans

checkout: https://fathom.info/salaryper/

Deconstruction - Example #1

William Playfair's Line Charts (1786)



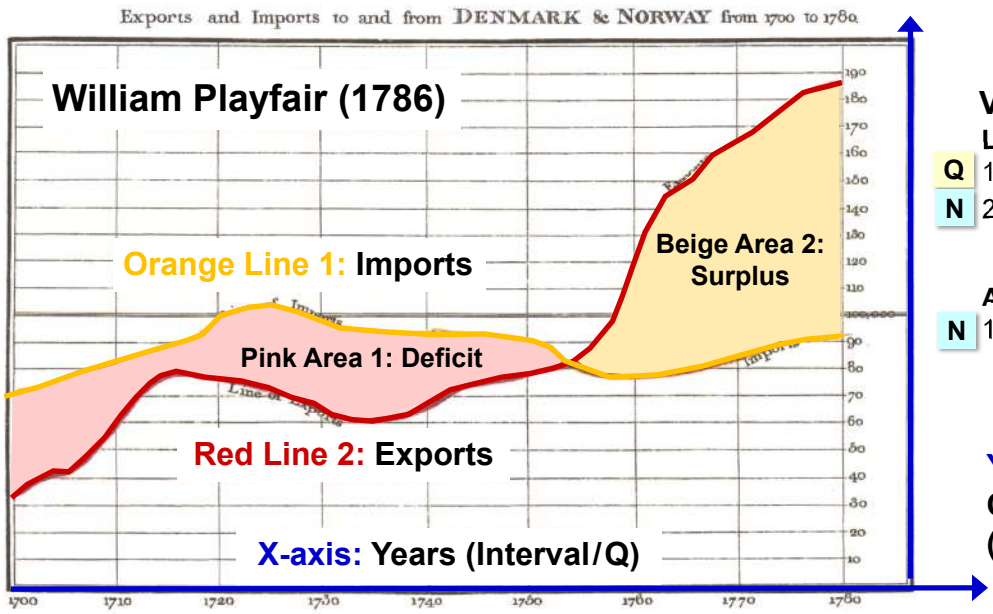
- William Playfair's Commercial and Political Atlas was published in 1786
- Playfair examined the imports and exports between Britain and various countries.
- To illustrate these trade relationships, Playfair created the first ever line charts that show the change in trade over time.

[6] Image can be found at: https://upload.wikimedia.org/wikipedia/commons/5/52/Playfair_TimeSeries-2.png

22

22

Deconstruction - Example #1



Visual Variables

Lines

Q 1. Position (X,Y)

N 2. Colour (Orange/Red) Import & Export

Areas

N 1. Colour (Pink/Beige) Deficit & Surplus

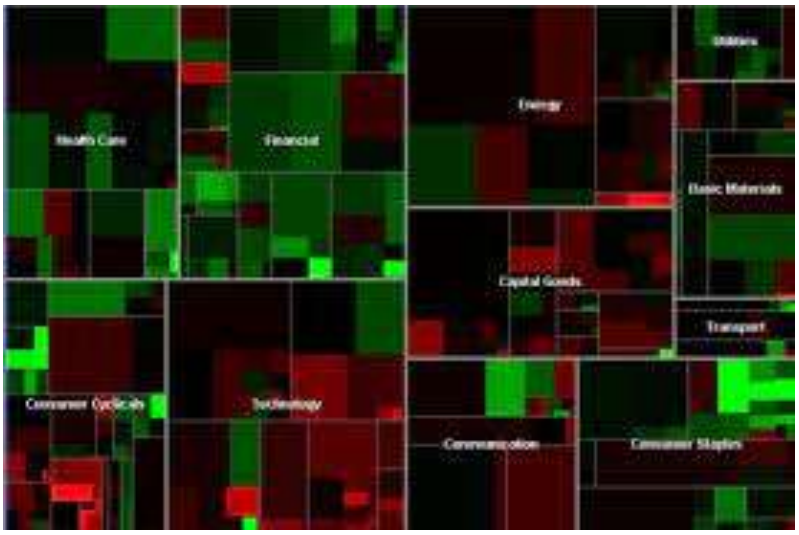
Y-axis: Currency (£10k) **(Ratio/Q)**

23

23

Deconstruction - Example #2

Martin Wattenberg's Map of the Market (1998)



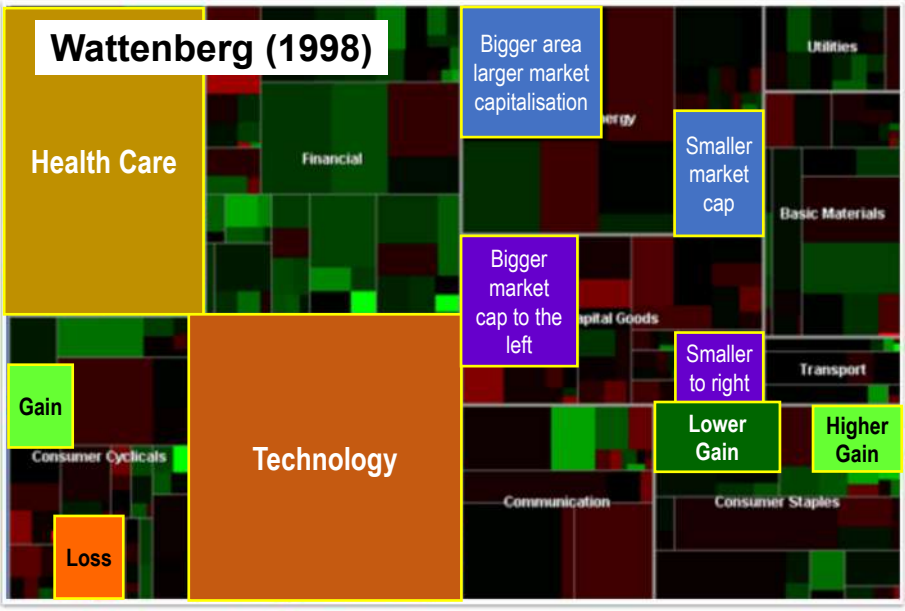
- Martin Wattenberg & co. created one of the 1st web-based visualisations, Map of the Market.
- It displayed live stock market data for more than a decade, starting in 1998, while Martin was at SmartMoney.com.
- Map of the Market uses a modified **Tree Map** algo that creates squarish tiles.

[7] Map of the Market details can be found at: http://www.bewitched.com/marketmap.html

24

24

Deconstruction - Example #2



Visual Variables

Areas (Rectangle)

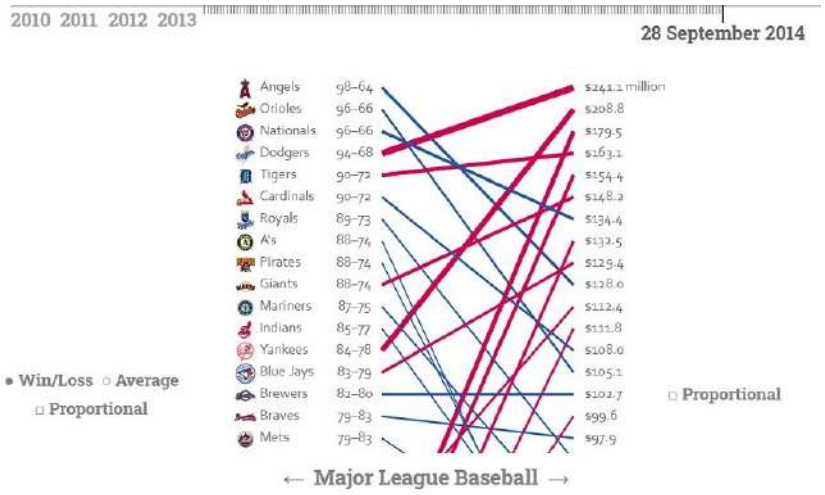
- Q 1. Size (Market Cap)
- N 2. Position (Market Sector)
- Q 3. Position (Market Cap)
- N 4. Colour Hue
  (Green - gain)
  (Red – loss)
- Q 5. Colour Saturation
  (Bright – higher gain)
  (Dark – lower gain)
  (Bright – higher loss)
  (Dark – lower loss)

25

25

Deconstruction - Example #3

Ben Fry's Slope Chart (2010-2014)



- Ben Fry is a visualisation expert and principal of Fathom, a design and software consultancy in Boston, USA.
- He designed Salary vs. Performance to visualised if baseball teams are spending their money well over the season.
- The visualisation uses a slope chart.

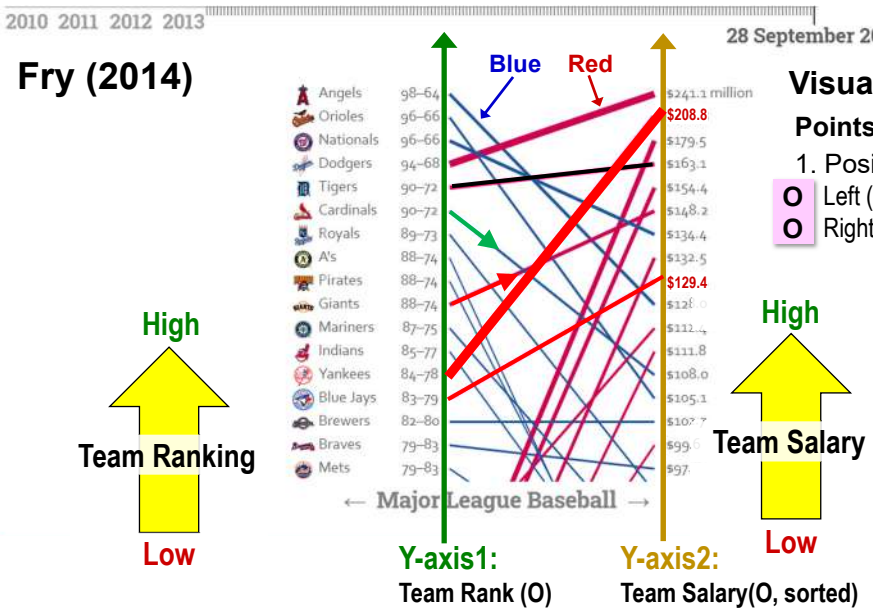
[8] Fry's website for Salary vs Performance: https://benfry.com/salaryper/

26

26

Deconstruction - Example #3

Fry (2014)



Visual Variables

Points

1. Position (Y1, Y2)
   - O Left (Ranking)
   - O Right (Salaries)

Links

2. Position (Team - Salary)
3. Slope
   - O ↘ (Good value)
   - O ↗ (Poor value)
4. Colour Hue
   - N (Blue +ve value) (Red -ve value)
5. Size (thickness)
   - Q (Salary)

27

27

Think and Apply

Visualising Relationships

- Joshua Keating & Chris Kirk (@Slate.com) created the **Middle East Friendship Chart** to visualise the complex relationships between different players in Syria's civil war.

- Draw the table for the relational data model.
- What visual marks and variables are used here?
- Redundant encoding?
- Colour choices?
- Effective?

[9] Find Middle East Friendship Chart at: http://www.slate.com/blogs/the_slatest/2015/10/06/syrian_conflict_relationships_explained.html

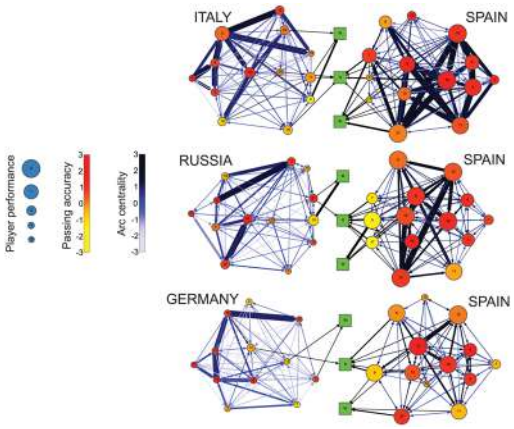
28

28

Think and Apply

Visualising Performance in Team Games

- Duch et al. analysed three Euro 2008 matches of the Spanish team using a visualisation technique from social network analysis.



- What types of data variables are visually represented in the image.
- What perceptual channels were used to encode their values?
- Colour choices? Can be better?

[10] Visualisation described in this paper: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010937

29

29

Summary

Deconstruction

- Data visualisation is both an **art** and a **science**.
- Designing **informative** and **visually appealing** visualisations requires us to learn from many other visualisation experts.
- We can do this by **deconstructing** their good pieces of work.
- We observe how they make use of the various visual marks and variables to **encode useful information** embedded in the datasets.
- We need to do these observation with a **critical eye,** to see what has **worked well** and what has not and **can be improved**.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

30

30

References for Visual Encoding and Deconstruction

[1] Bertin's quote from: J. Bertin, graphics and graphic information processing (1981) - https://books.google.com.sg/books?id=csqX_xnm4tcC&printsec=copyright&redir_esc=y#v=onepage&q&f=false

[2] T. Munzner, Visualization Analysis & Design, CRC Press (2015)

[3] J. Mackinlay, Automating the Design of Graphical Presentation of Relational Information (1986), https://research.tableau.com/sites/default/files/p110-mackinlay.pdf

[4] C. Ware, Information Visualization: Perception for Design, 3rd Edition, Morgan Kaufmann (2013),

[5] William Playfair's image can be found at: https://upload.wikimedia.org/wikipedia/commons/5/52/Playfair_TimeSeries-2.png

[6] William Playfair's image can be found at: https://upload.wikimedia.org/wikipedia/commons/5/52/Playfair_TimeSeries-2.png

[7] Martin Wattenberg's Map of the Market details can be found at: http://www.bewitched.com/marketmap.html

[8] Ben Fry's website for Salary vs Performance: https://benfry.com/salaryper

[9] Keating & Kirk's Middle East Friendship Chart at: http://www.slate.com/blogs/the_slatest/2015/10/06/syrian_conflict_relationships_explained.html

[10] Duch et al.'s visualisation described in this paper: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010937

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

31

31

Data Visualisation

Associate Professor Goh Wooi Boon

School of Computer Science and Engineering

Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 4 –Visualisation Tools

Contents

- Types Of Visualisation Tools
- Choosing The Visualisation Tools
- Data Visualisation Tools Used In This Course
- Other Python Data Visualisation Libraries

© A/P Goh Wooi Boon (SCSE/NTU)

Types Of Visualisation Tools

"Like a Child in a Candy Store"

- There is a huge variety of tools available to refine and query your data, and to create, edit, and display your visualisations.
- The types of tools for building visualisations fall into four major categories[1]:
- Basic productivity applications — Microsoft Excel, LibreOffice, Google Sheet
- Visualisation software — Proteopedia, ArcGIS, 3DSlicer
- Business intelligence tools — tableau, Power BI, QilkView
- Developer based packages — D3.js - Javascript, bokeh, seaborn, matplotlib, Python-based

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

3

3

Choosing The Visualisation Tool

How do I Choose the Right One?

- When given a free choice, it is important to select the tool that is most suitable to support the nature and purpose of your data visualisation work and workflow.
- Consider the following aspects in the evaluation[1]:

- **Sharing**
- **Output**
- **Display types**
- **Data Exploration**
- **Interoperability**
- **Simplicity**
- **Persistence**

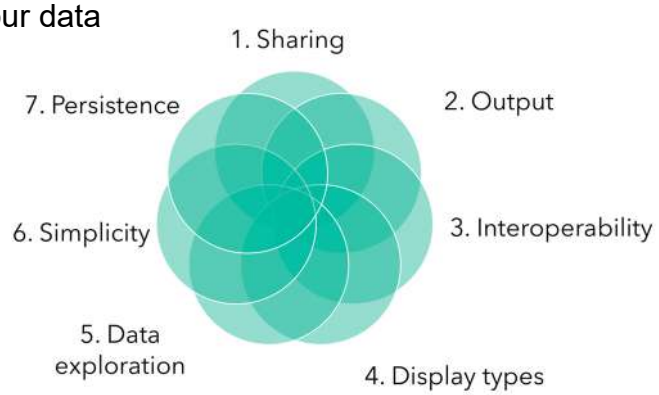


Image taken from [1]

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

4

4

Choosing The Visualisation Tool

How do I Choose the Right One?

- **Output** – Are you publishing the visualizations to the web, PDF documents or embedding them into other applications? The destination of your visualization will dictate tool choice. e.g. not all tools have good web interactive support[1].
- **Display types** – What types of visualizations do you intend to create? e.g. maps, networks, and text-based visualizations may not be supported by some tool[1].
- **Interoperability** – How easily can you connect to other data sources? For instance, does the software allow you to import diverse file types, such as .xls, CSV, .txt, or allow you to link to your databases?[1].
- **Data exploration** – Do you need a tool to explore and present your data visually, or just to present your data visualization visuals? e.g. features such as visual querying may not be supported by some tools[1].

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

5

5

Choosing The Visualisation Tool

How do I Choose the Right One?

- **Simplicity** – Are you looking to create charts and graphs quickly? How much time can you afford to learn how to use the tool as some may be powerful but require a steep learning curve, even to build a simple bar chart[1].
- **Sharing** –Can others view and edit your visualization and analysis? If your workflow is essentially collaborative, then choose a tool that enables collaboration during the design, prototyping and creation process[1].
- **Persistence** – Do you think that you will have to go back and revise the visualizations you create? If so, it is important to choose a tool that is well supported and will be around for a while[1].

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools - http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

6

6

Data Visualisation Tools Used In This Course

Data Visualisation Using Python

- This data visualization course is specifically designed for students with basic computer science background, as such programming skills are assumed.
- The Python-based tools are supported by numerous **visualisation** and **statistical** libraries that will facilitate both **explanatory** visualisation and **exploratory** analysis.

matplotlib
seaborn
Static Visualisation

pandas
Data Manipulation

bokeh
Interactive Visualisation

SciPy
statsmodels
Statistical Analysis

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

7

7

Data Visualisation Tools Used In This Course

Other Data Visualisation Tools

- This course focuses on the principles and practice of good **human-centric** data visualisation **design** and not the visualisation tool per se.
- As such, students are free to use any other data visualisation tools they are familiar with, as long as it can support the required deliverables in the course.

Microsoft Excel
Widely available tool for quick non-programming charts

D3.js
Works well with web standards like HTML, CSS and SVG

tableau
Power BI
These are commercial tools. Free version may be limited.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

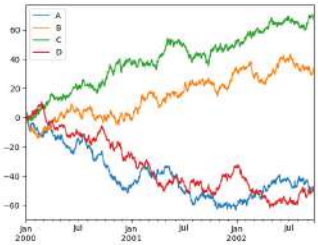
8

8

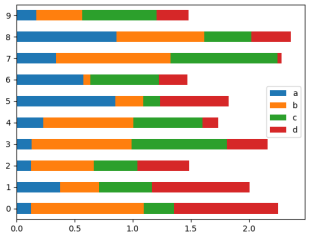
Data Visualisation Tools Used In This Course

Pandas

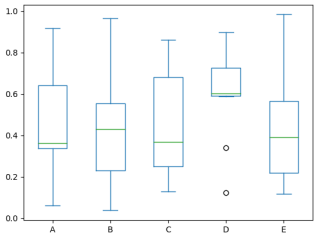
- Pandas is the Python data analysis library[2].
- It is a powerful, flexible and easy to use open source **data analysis** and **manipulation** tool, built on top of the Python programming language.
- Pandas can also support a variety of basic charts and plots[3] such as:



Line plots



Stacked bar plots



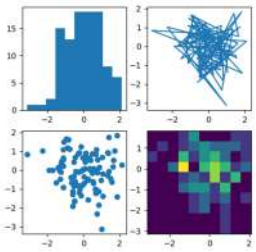
Box plots

[2] Pandas – Python Data Anaysis Library - https://pandas.pydata.org/
[3] Pandas – Chart Visualization - https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html

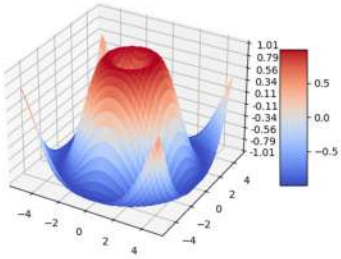
Data Visualisation Tools Used In This Course

Matplotlib

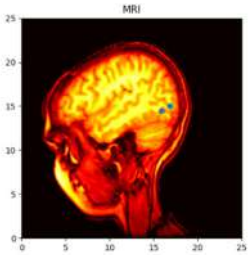
- Matplotlib[4] is a comprehensive library for creating static, animated, and interactive visualisations in Python[4] and it provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, etc.
- Some examples[5] of plots from Matplotlib are shown here:



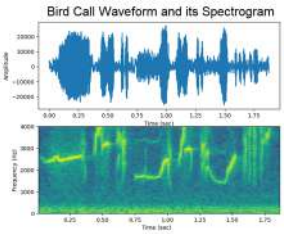
Gridded sub-plots



3D surface plot



Coloured images



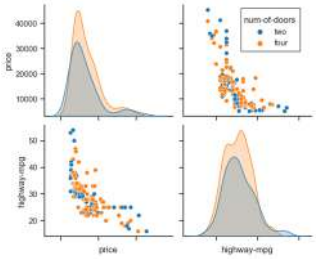
Sound & spectrogram

[4] Matplotlib: Visualization with Python - https://matplotlib.org/
[5] Matplotlib – Sample plots in Matplotlib - https://matplotlib.org/stable/tutorials/introductory/sample_plots.html

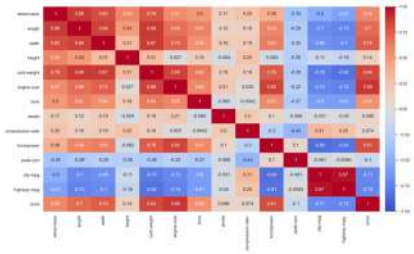
Data Visualisation Tools Used In This Course

Seaborn

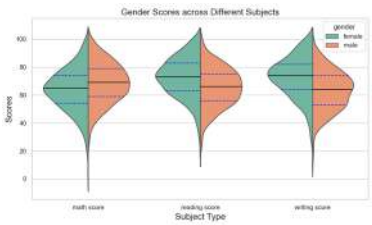
- Seaborn[6] is a Python library predominantly used for making **statistical graphics**.
- It is built on top of **Matplotlib** and is closely integrated with **Pandas**'s dataframe structure, which allows sophisticated plots to be done within a single command line.
- Seaborn supports a large variety of charts and plots[7] as shown:



Facet plots (scatter & KDE)



Heatmaps (Correlations)



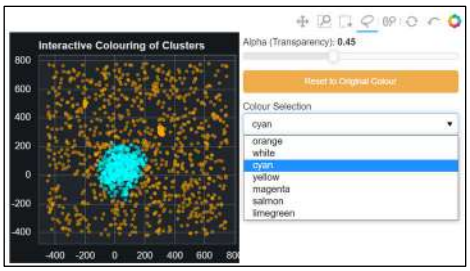
Violin plots

[6] Seaborn – Statistical data visualization - https://seaborn.pydata.org/
[7] Seaborn – Example Gallery - https://seaborn.pydata.org/examples/index.html

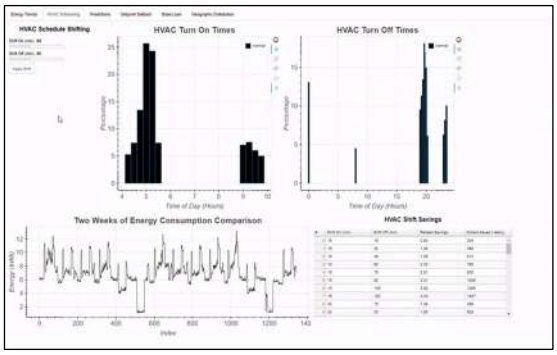
Data Visualisation Tools Used In This Course

Bokeh

- Bokeh[8] is a Python library for creating **interactive** visualisations for web browsers.
- Bokeh can produce elegant visualisation (including dashboards) that are interactive and works with large or streaming datasets.



Embedding widgets like slider & menus with Bokeh



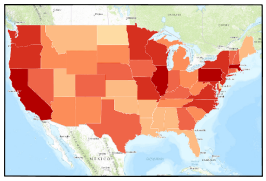
Example Dashboard create using Bokeh[9]

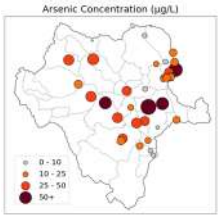
[8] Bokeh - https://bokeh.org/
[9] Will Koehrsen – Data Visualization with Bokeh in Python – https://towardsdatascience.com/data-visualization-with-bokeh-in-python-part-one-getting-started-a11655a467d4

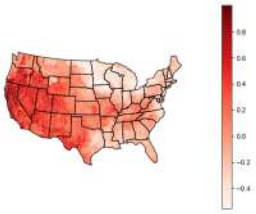
Other Python Data Visualisation Libraries

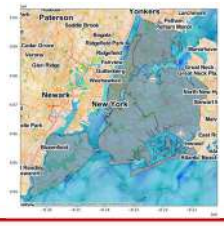
Visualising Geospatial Data

- The are many useful visualisation libraries in Python that allow us to visualise geographical and location-based information.
- **GeoPandas**[10] - extends the datatypes used by Pandas to allow spatial operations on geometric types. Its goal is to make working with geospatial data in Python easier.
- **Geoplot**[11] – is a high-level Python geospatial plotting library. It is built on top of matplotlib with the specific purpose of doing cartographic plots.







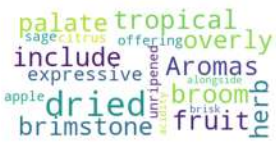


[10] GeoPandas - https://geopandas.org/
[11] Geoplot: geospatial data visualization – https://residentmario.github.io/geoplot/index.html

Other Python Data Visualisation Libraries

Visualising Textual Data

- The are other visualisation libraries in Python that allow us to visualise various properties in text-based data.
- **WordCloud**[12] – is a text visualization library that can analyse the frequency of words and correlates the size and opacity of a word to its frequency within a body of text.
- It has an interesting feature that allows an image mask to define the shape over which the word cloud is populated. This can serve as a visual context to the document being analysed[13].



[12] WordCloud for Python documentation – https://amueller.github.io/word_cloud/
[13] Duong Vu, Generating WordClouds in Python – https://www.datacamp.com/community/tutorials/wordcloud-python

Summary

Visualisation Tools

- There are numerous types of visualisation tools available, both the free open-source variety and sophisticated commercial tools.
- Choosing a suitable tool depends on the purpose of the visualisation. Factors related to the **design** (e.g. collaborative, fast turnaround) and **deployment** environments (e.g. interactive, web-based) should also be considered when making the choice.
- This course has adopted open source tools based on the **Python programming** environment as it supports many different visualisation libraries with good support for data wrangling and statistical analysis.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

15

15

References for Visualisation Tools

[1] Kristen Sosulski, Data Visualization Software: Tips for picking the right tools – http://www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

[2] Pandas – Python Data Anaysis Library - https://pandas.pydata.org/

[3] Pandas – Chart Visualization - https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html

[4] Matplotlib: Visualization with Python - https://matplotlib.org/

[5] Matplotlib – Sample plots in Matplotlib - https://matplotlib.org/stable/tutorials/introductory/sample_plots.html

[6] Seaborn – Statistical data visualization - https://seaborn.pydata.org/

[7] Seaborn – Example Gallery - https://seaborn.pydata.org/examples/index.html

[8] Bokeh - https://bokeh.org/

[9] Will Koehrsen – Data Visualization with Bokeh in Python – https://towardsdatascience.com/data-visualization-with-bokeh-in-python-part-one-getting-started-a11655a467d4

[10] GeoPandas - https://geopandas.org/

[11] Geoplot: geospatial data visualization – https://residentmario.github.io/geoplot/index.html

[12] WordCloud for Python documentation – https://amueller.github.io/word_cloud/

[13] Duong Vu, Generating WordClouds in Python – https://www.datacamp.com/community/tutorials/wordcloud-python

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Note: All online articles were accessed between Jul – Aug 2021

16

16

Data Visualisation

Associate Professor Goh Wooi Boon

School of Computer Science and Engineering

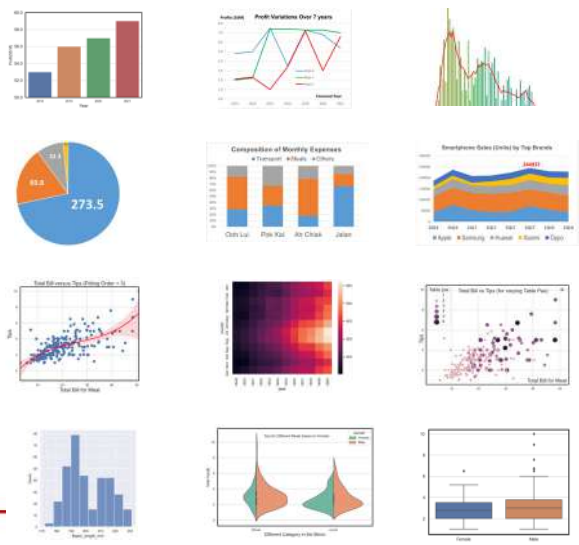
Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 5 – Basic Plots and Charts

Contents

- Comparison Plots
- Composition Plots
- Relation Plots
- Distribution Plots

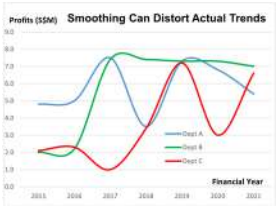


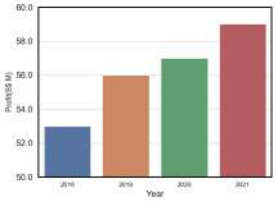
© A/P Goh Wooi Boon (SCSE/NTU)

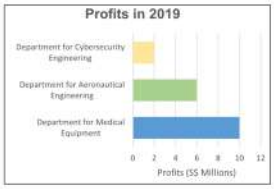
Chapter 5.1 – Visualising Comparison in Data

Contents

- Basic Comparison Plots
- Line Charts
- Bar Charts





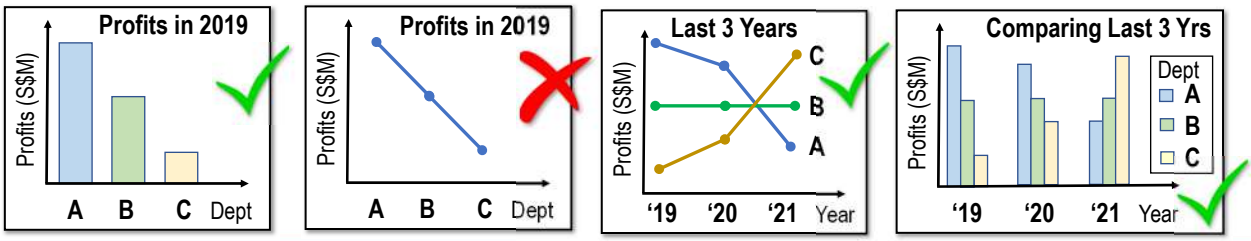


© A/P Goh Wooi Boon (SCSE/NTU)

Basic Comparison Plots

Draw the Line or Raise the Bar?

- Line charts should be use for **time series** (i.e. chronological) or when the x-axis has an **interval scale** (e.g. distances from a point, size of an object). While bar charts should be use when the x-axis have **nominal** (i.e. unordered) scale[1].
- Line charts should be used to show **trends** or pattern changes but **not precise quantitative values**. Bar charts could be used if specific measurements are to be **compared** at each **discrete intervals** along the x-axis.

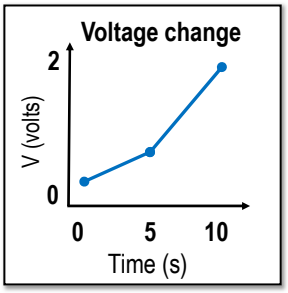


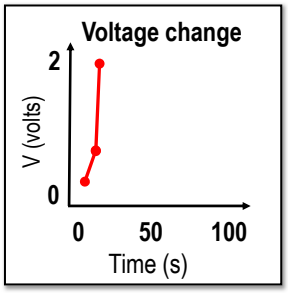
[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

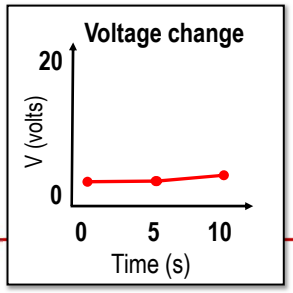
Line Charts

Axes Scaling and Aspect Ratio

- The aspect ratio of a graph is an important factor for judging **rate of change**.
- The "**banking to 45°**" rule[2] can be used to improve the visually encoding accuracy of changing trends.
- The axes scales or aspect ratio are adjusted such that the arithmetic average of all line orientations is nearly +45° or -45° (depending on the the direction of the trend).







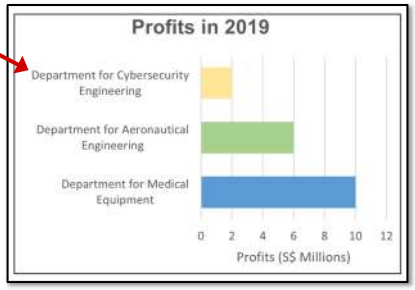
Taken from:
[2] W.S. Cleveland, The Elements of Graphing Data,, Hobart Press (1994)

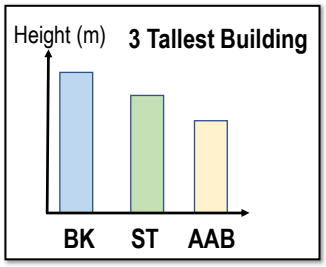
Bar Charts

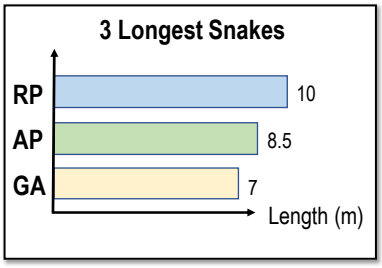
Vertical or Horizontal Bars

- Avoid using vertical text. If the x-axis categories have **long text** descriptions, consider using **horizontal bars**.

Long text descriptions







- The nature of the **conceptual data**[1] can also help decide which bar orientation better **scaffold the interpretation** of the data. (e.g. height - vertical, distance - horizontal) .

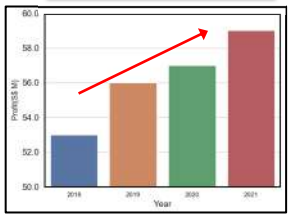
[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

Zero Baseline

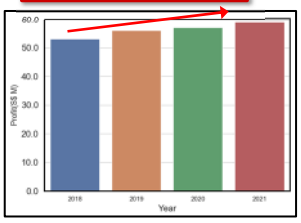
Zero Heroes

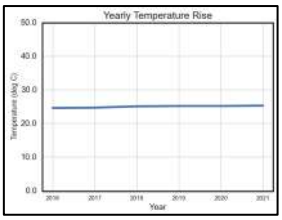
- Some argue that zero baseline must be used for **all plots**[3], whether bar or line charts to maintain visual "truthfulness".
- **Bar charts** should normally have a zero baseline as data is encoded using length and truncating the axis is naturally misleading.
- There is some flexibility with **line** charts, especially when the changing trends encoded in the **angle cannot be discern** if zero-baseline is imposed.

Visually exaggerated annual profit growth

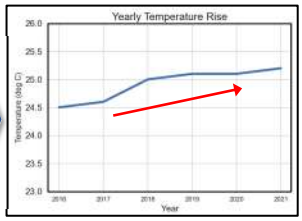


Actual profits plotted with zero base





Up trend not visible when zero base is imposed but becomes apparent when relaxed



7

[3] C. Skelton, Bar charts should always start at zero. But what about line charts? (2018), http://www.chadskelton.com/2018/06/bar-charts-should-always-start-at-zero.html

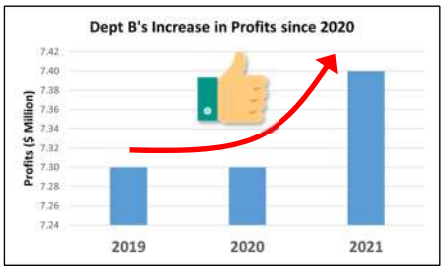
7

Ethical Visualisation

Don't Exaggerate

- Head of **Dept B** is concerned about the lack of growth in his product line compared to the huge growth in **Dept C**, which is likely to surpass Dept B soon. He has to make a yearly review presentation to the board and started working on his chart...

Changing profits of departments B and C in the last 4 years



The chart that the Head of Dept B presented the board of directors

Has ethical visualisation been violated here?

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

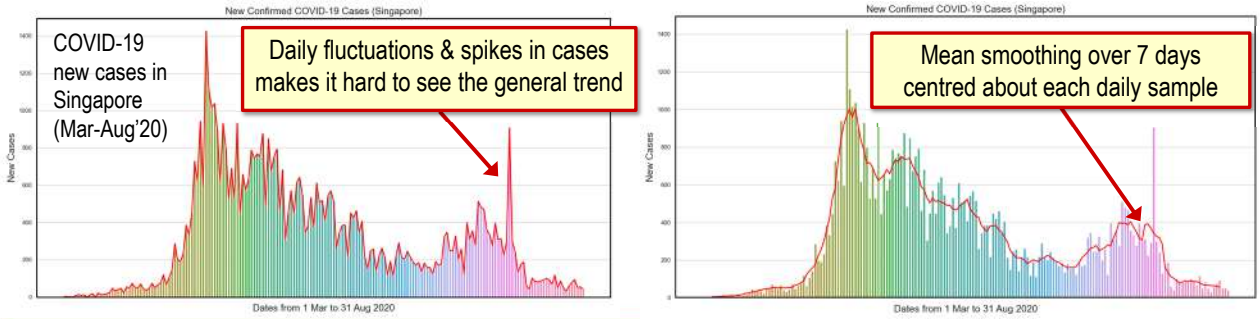
8

8

Visualising General Trends

Smoothing the Series

- Smoothing line or bar charts can reveal the more general or long-term trends in the data. Smoothing is particularly useful for noisy time series data.
- However, care must be taken to ensure that such smoothing does not affect the integrity of what the data represents[4].



[4] Jon Peltier, When you use a smoothed line chart, your data is not affected, it's misrepresented! https://www.vizwiz.com/2011/12/when-you-use-smoothed-line-chart-your.html

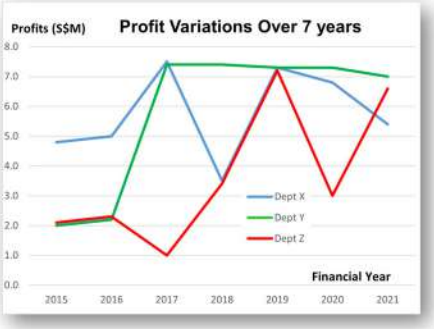
Think and Apply

When Smoothies are a Bad Idea

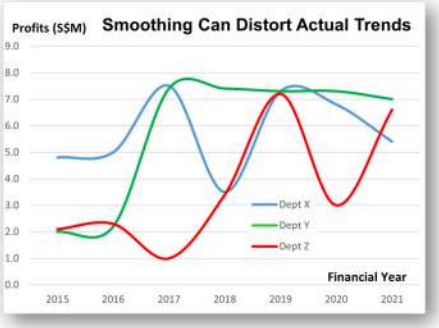
- Smoothing the data series distorts the plot and create misleading visual trends.

	Profits in last 7 years (S$ M)		
Year	Dept X	Dept Y	Dept Z
2015	4.8	2.0	2.1
2016	5.0	2.2	2.3
2017	7.5	7.4	1.0
2018	3.5	7.4	3.4
2019	7.3	7.3	7.2
2020	6.8	7.3	3.0
2021	5.4	7.0	6.6

Profits of three departments over 7 years



Line Chart showing exact values



Line plots smoothed using Excel "smoothed line" feature

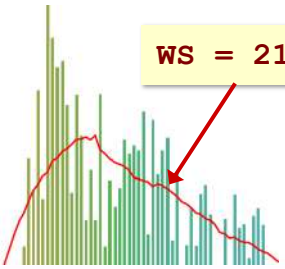
Data Smoothing

Smoothing Data Using Pandas

- Data smoothing[5] is done over a sliding window of a specified **width** . A **wider** window that covers more data points will result in **more smoothing**.

```
Smooth = Data.rolling(
        window = WS
        center = CW,
        win_type = WT).mean()
```

Window size



[5] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html

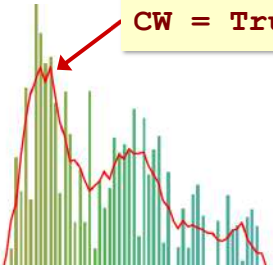
Data Smoothing

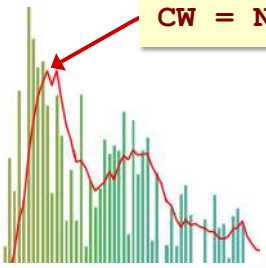
Smoothing Data Using Pandas

- Data smoothing[5] is done over a sliding window of a specified **width** . A **wider** window that covers more data points will result in **more smoothing**.
- The window can be **centred** or to the **edge** of the data point during smoothing and can take different "**shapes**"[6] to reduce the influences of data further away in the window when computing each **mean**.

```
Smooth = Data.rolling(
        window = WS
        center = CW,
        win_type = WT).mean()
```

Centering (default is right edge)





[5] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html
[6] Window types for win_type- https://docs.scipy.org/doc/scipy/reference/signal.windows.html#module-scipy.signal.windows

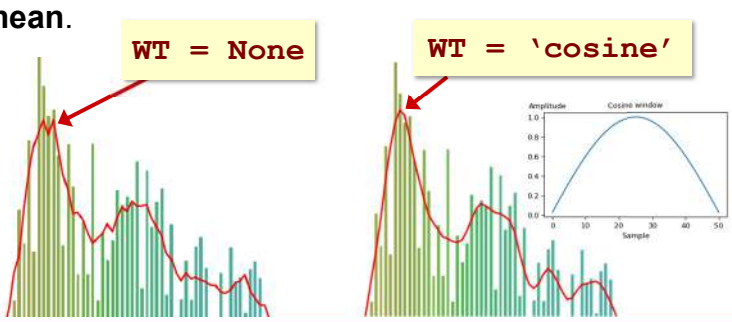
Data Smoothing

Smoothing Data Using Pandas

- Data smoothing[5] is done over a sliding window of a specified **width** . A **wider** window that covers more data points will result in **more smoothing**.
- The window can be **centred** or to the **edge** of the data point during smoothing and can take different "**shapes**"[6] to reduce the influences of data further away in the window when computing each **mean**.

```
Smooth = Data.rolling(
        window = WS
        center = CW,
        win_type = WT).mean()
```

Window shape
(default is equal weight)



[5] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html
[6] Window types for win_type- https://docs.scipy.org/doc/scipy/reference/signal.windows.html#module-scipy.signal.windows

Summary

Comparison Plots

- **Line** and **bar** charts are two basic plots for comparing data values.
- **Bar charts** are useful for **comparing values** between variables, especially those having **nominal** scales of measure.
- **Line charts** are effective in observing **changing trends** in the data, especially time series data.
- **Data smoothing** is sometimes needed in order to observe **trend** changes that occur over a **longer time** duration or sampling span.
- However, be careful that smoothing does not **distort** the true interpretation of the data values.

References for Comparison Plots

[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

[2] W.S. Cleveland, The Elements of Graphing Data,, Hobart Press (1994)

[3] C. Skelton, Bar charts should always start at zero. But what about line charts? (2018), http://www.chadskelton.com/2018/06/bar-charts-should-always-start-at-zero.html

[4] Jon Peltier, When you use a smoothed line chart, your data is not affected, it's misrepresented! https://www.vizwiz.com/2011/12/when-you-use-smoothed-line-chart-your.html

[5] Pandas, Dataframe.Rolling reference - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html

[6] Window types for win_type- https://docs.scipy.org/doc/scipy/reference/signal.windows.html#module-scipy.signal.windows

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

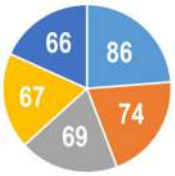
15

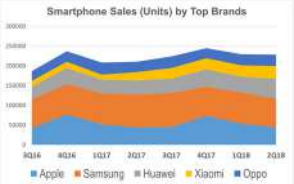
15

Chapter 5.2 – Visualising Composition in Data

Contents

- Basic Composition Plots
- Pie Charts
- Stacked Bar Charts
- Stacked Area Charts



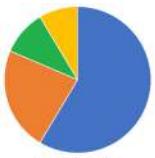


© A/P Goh Wooi Boon (SCSE/NTU)

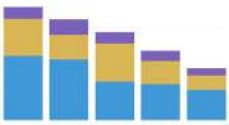
Basic Composition Plots

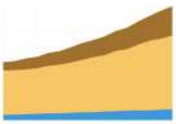
Part-to-Whole Analysis

- Composition plots help visualise how individual parts comprise a whole.
- **Pie and donut charts** – are used for visualising compositions that are **static**.
- **Stacked charts** – are for composition that are **changing** with **another dimension** like time or category. Stacked **bar** or stacked **area** charts are common examples.
- Some other charts for viewing composition include the **treemap** and **funnel** chart.


Pie Chart

Donut Chart


Stacked Bar Chart


Stacked Area Chart

Treemap

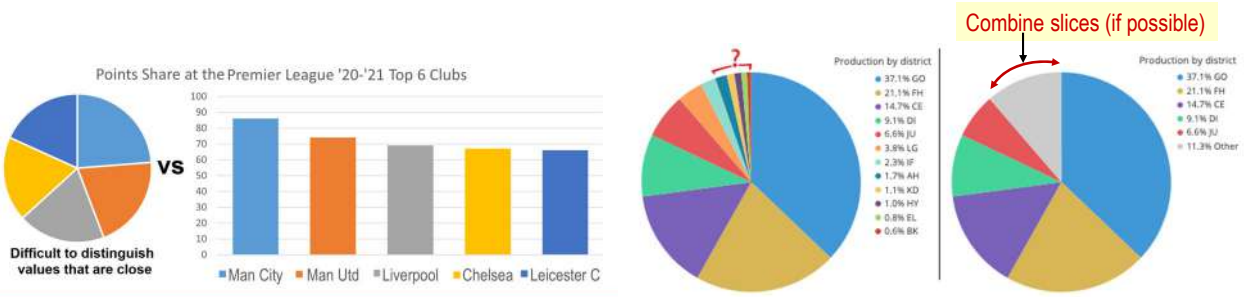
Funnel Chart

[1] How to Choose Charts to Show Data Composition (2019), https://www.webdatarocks.com/blog/how-to-choose-charts-to-show-data-composition/

Pie Charts

When Should I Not Have A Pie?

- Pie charts are not suitable when the slices have **little variations**. **Bar** charts provide more accurate **visualisation of values** if they are deemed important and of interest.
- Pie charts are not suitable when there are **too many slices**. Consider if some of the smaller slices can be combined into an 'Other' category[2].

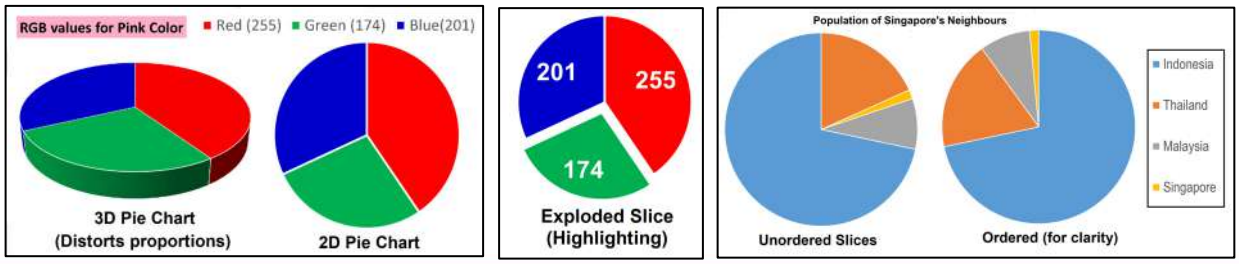


[2] Mike Yi, A Complete Guide to Pie Charts (2019), https://chartio.com/learn/charts/pie-chart-complete-guide/

Pie Charts

What Type Of Pie Should I Have?

- It is generally good practice to keep your pie charts **simple** for visual clarity.
- **Exploding** can help **highlight** slice of interest but use this carefully as such gaps can distorts part-to-whole comparison[2]. Annotating the slices will help in this case.
- **Ordering slices** (e.g. largest to smallest slice) usually aid easier understand, especially if categories do not have inherent ordering to comply with.



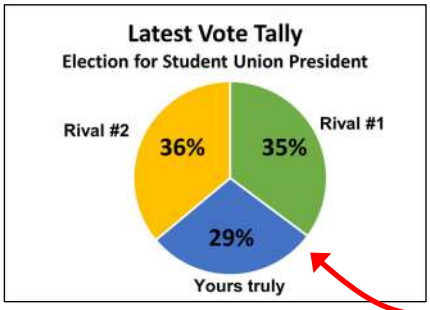
[2] Mike Yi, A Complete Guide to Pie Charts (2019), https://chartio.com/learn/charts/pie-chart-complete-guide/

Ethical Visualisation

Be Visually Truthful

- You are participating in a week-long election for the position of President of the Students' Union. You decided to put in some last-minute effort to get people to vote for you using this visual pitch...

3D pie chart visual to canvas for last minute votes



This is the true picture that a plain pie chart would show.

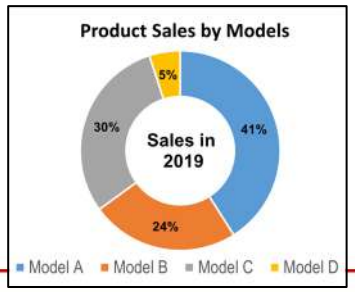
Has ethical visualisation been violated here?

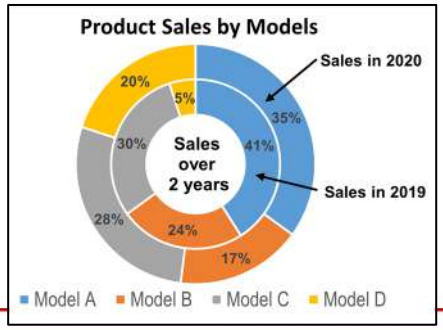
Remember, people are not fools and they are voting for your honesty as well

Pie Charts

Can I Have a Doughnut Instead?

- A doughnut (donut) chart is basically a pie chart with a central circle removed and there are **no significant differences** in readability between the two types of plots.
- The central area of a donut chart can be used for **additional information**.
- It can also show more than one set of data to allow **comparisons** over two different compositions[2].



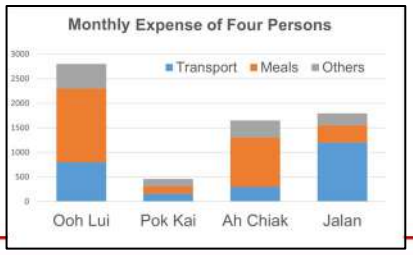


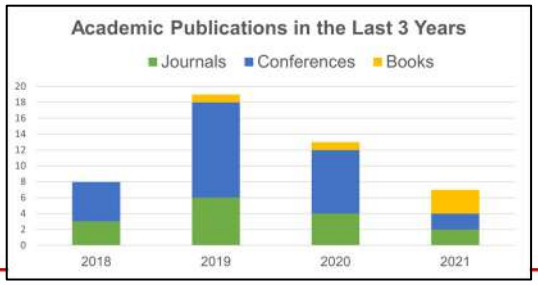
[2] Mike Yi, A Complete Guide to Pie Charts (2019), https://chartio.com/learn/charts/pie-chart-complete-guide/

Stacked Bar Charts

Visualising Multiple Compositions

- Stacked bar charts show relative decomposition of each primary bar (e.g. Persons) based on the levels of a **second categorical variable** (e.g. Expense type)[3].
- The stacked sub-bars of each secondary variable is **coloured** based on the category of the parts that make up the whole bar.
- If the primary bars are drawn across a time series (e.g. years), they express how the composition is changing over time.



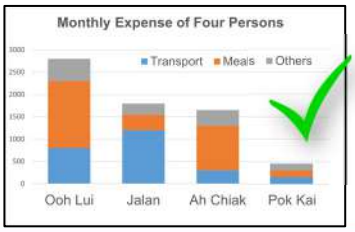


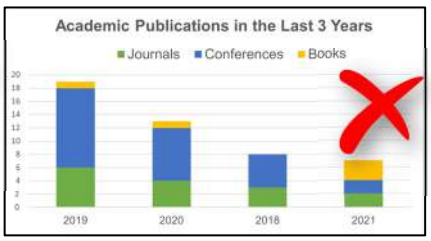
[3] Mike Yi, A Complete Guide to Stacked Bar Charts (2019), https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/

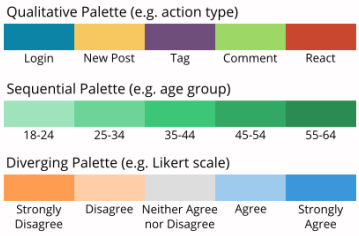
Stacked Bar Charts

Ordering and Colouring the Bars

- **Ordering** stacked bar charts can help make the chart **easier to comprehend**.
- The rule of thumb is to order the bars from largest to smallest unless there is an intrinsic order in the primary bar (e.g. this is a time-varying stacked bar chart).
- The choice of colour palette to assign to each categorical level should match the variable type: a qualitative palette for purely categorical variables, and sequential, or diverging for variables with a meaningful order[3].





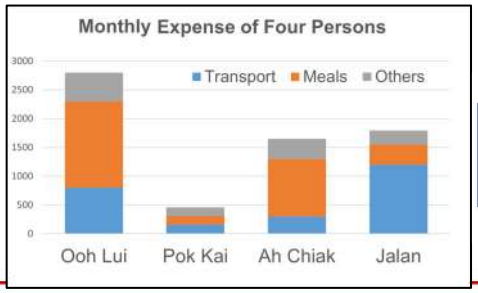


[3] Mike Yi, A Complete Guide to Stacked Bar Charts (2019), https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/

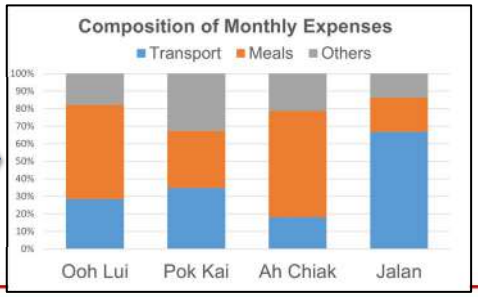
Stacked Bar Charts

Visualising Percentages

- In the percentage stacked bar chart, each primary bar is scaled to the **same height** with each **sub-bar** encoding its **percentage** contribution to the primary whole.
- A uniform height makes it easier to visually compare the **percentage contributions** of each sub-bar category to the whole. But this removes the ability to visualise the actual values of each of the sub-bar category)[3].



Percentage Stacked Bar Chart



[3] Mike Yi, A Complete Guide to Stacked Bar Charts (2019), https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/

Stacked Area Charts

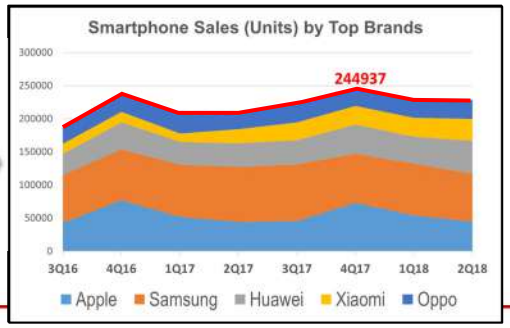
Visualising Composition Trends

- A **stacked area chart** can be considered if the x-axis is **interval** or **ratio** scale.
- It is created by plotting lines (and filling its area below) one at a time. The height of the most recently plotted line serves as a **moving baseline** for the next line plot[4].
- Height of the topmost line encodes the total sum across all composition categories.

Quarter	Apple	Samsung	Huawei	Xiaomi	Oppo	TOTAL
3Q16	43001	71734	32490	14926	24591	186742
4Q16	77039	76783	40804	15751	26705	237082
1Q17	51993	78776	34181	12707	30922	208579
2Q17	44315	82855	35964	21179	26093	210406
3Q17	45442	85605	36502	26853	29449	223851
4Q17	73175	74027	43887	28188	25660	244937
1Q18	54059	78565	40426	28498	28173	229721
2Q18	44715	72336	49847	32826	28511	228235

[2] Data from M.Dobler & T.Grossmann, The Data Visualization Workshop, Packt Publishing (2020) - https://github.com/PacktWorkshops/The-Data-Visualization-Workshop

Stacked Area Chart

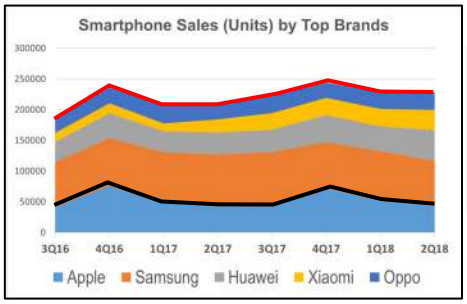


[4] Mike Yi, A Complete Guide to Stacked Area Charts (2019), https://chartio.com/learn/charts/area-chart-complete-guide/

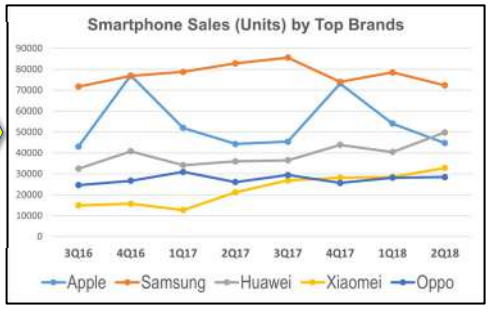
Stacked Area Charts

To Stack or Not to Stack

- Stacked area chart are useful for tracking the total value and analysing the varying breakdown of the parts making up this total[4].
- Only the exact values of the overall **total** and **bottommost** categories are easy to gauge as intermediate categories are plotted against varying baselines.
- If exact values are required for all categories, consider using **multiple line charts**.



Use Line Charts to Visualise Exact Values



[4] Mike Yi, A Complete Guide to Stacked Area Charts (2019), https://chartio.com/learn/charts/area-chart-complete-guide/

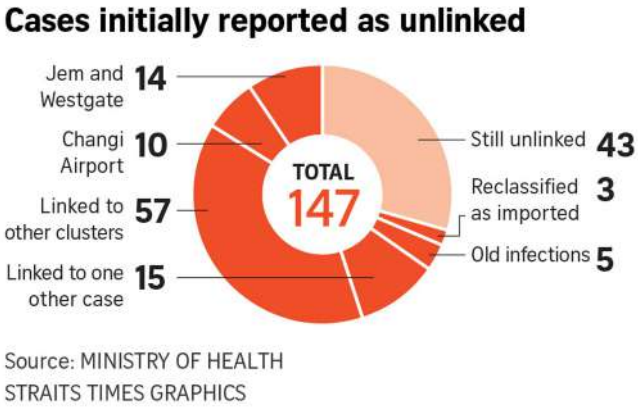
11

11

Think and Apply

Singapore's Virus Situation since 29 Apr 2021

- Why was a donut chart used to visualise the unlinked cases?
- Should we replace the donut chart with a pie chart instead?
- Should each slice of the donut be encoded with a different colour? Is there a reason why ST Graphics used the colour scheme?



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Image from The Straits Times on 8 June 2021

12

12

Summary

Composition Plots

- Composition plots help us visualise the **individual parts** comprising the **whole**.
- Pie and donut charts are effective for visualising **small number** of individual parts, especially if the **variations** in the composition is **distinct** and **precision** in reading the values is not important.
- Stacked bar charts are effective in visualising **multiple compositions**.
- Stacked area chart can be used to visualise **composition trend** changes and when **precise** values at every interval is not critical to the analysis.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

13

13

References for Composition Plots

[1] How to Choose Charts to Show Data Composition (2019), https://www.webdatarocks.com/blog/how-to-choose-charts-to-show-data-composition

[2] Mike Yi, A Complete Guide to Pie Charts (2019), https://chartio.com/learn/charts/pie-chart-complete-guide/

[3] Mike Yi, A Complete Guide to Stacked Bar Charts (2019), https://chartio.com/learn/charts/stacked-bar-chart-complete-guide/

[4] Mike Yi, A Complete Guide to Stacked Area Charts (2019), https://chartio.com/learn/charts/area-chart-complete-guide/

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

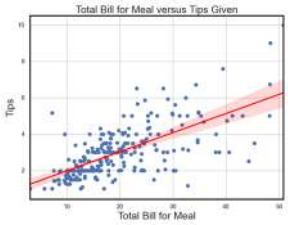
14

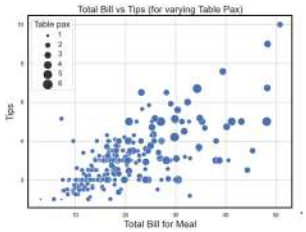
14

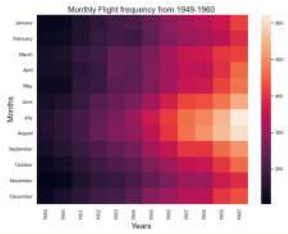
Chapter 5.3 – Visualising Relationships in Data

Contents

- Basic Relation Plots
- Scatter Pots
- Bubble Plots
- Heatmaps



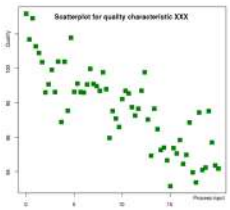




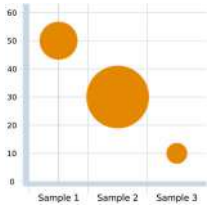
Basic Relation Plots

Seeing Relationship Among Variables

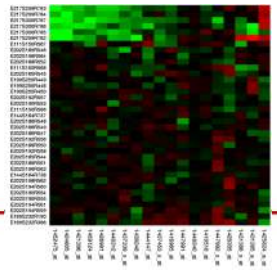
- Relation plots help show relationships or patterns among variables.
- **Scatter and Bubble plots** – Scatter plots are useful to visualise correlation between two variables for one or more groups. The value of an additional third variable can be encoded by the dot size in a bubble plot.
- **Heatmaps** – When one or more of these variables are non-continuous or non-numeric, a heatmap can be used. Relationships between categories are visualised through a grid of values encoded in different colours.



Scatter Plot



Bubble Plot

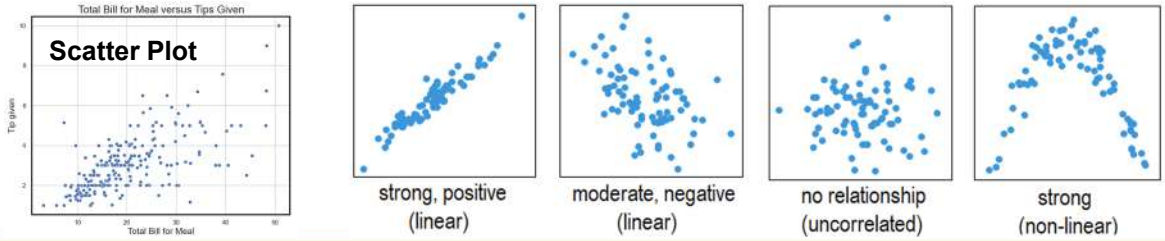


Heatmap

Scatter Plots

Seeing Relationship in the Dots

- Scatter plots use the 2-D position of dots to represent values of two different numeric variables. The collection of these scattered dots can reveal **correlational relationships** or **patterns** between these two variables.
- Given a horizontal value (independent variable), a highly **correlated** relationship can readily predict the vertical value (dependent variable) and such pairwise relationships can be described in many ways (e.g. positive/negative, strong/weak, linear/non-linear)[1]

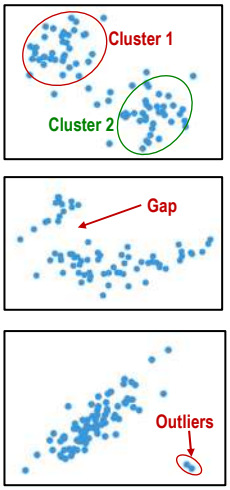


[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/

Scatter Plots

Seeing Groupings in the Dots

- Scatter plots can show interesting grouping patterns of dots[1].
- **Clusters** – are distinctive groupings in the dots forming various discernible shapes. Some clusters are circular, while others may be elongated.
- **Empty regions or gaps** – are areas that have a distinctively lack of dots. These unexpected gaps can be informative as they reveal incompatible value pairs between the two variables of interest.
- **Outliers** – are dots that are far off and do not appear to belong to any cluster. These data points could be erroneously recorded or they could be rarely occurring events of special interest.

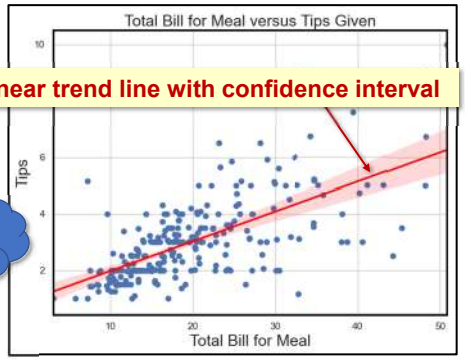


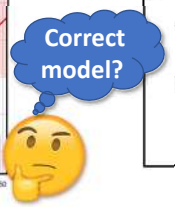
[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/

Scatter Plots

Seeing Trends in the Dots

- Trend lines can be drawn to show the mathematically **best fit** to the data.
- Seaborn's `regplot()`[2] method is used to fit a **linear regression model** to the data. The shaded area about the line is the **confidence interval**.





- Lines of higher polynominal order can be used if such fit makes sense to the data.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[2] Seaborn, Plot data and a linear regression model fit , https://seaborn.pydata.org/generated/seaborn.regplot.html

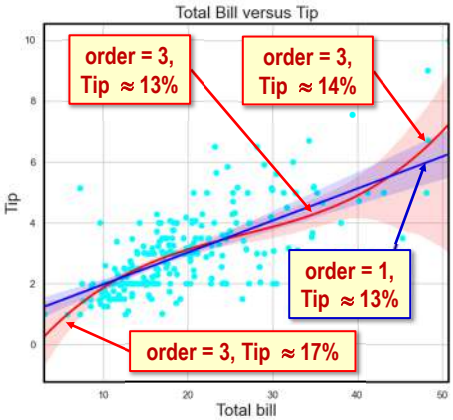
5

5

Scatter Plots

When a Polynomial May Not Fit

- The polynomial regression model with order=1 provides a constant estimation of a 13% tip irrespective of the cost of the meal.
- A polynomial model of order=3 predicts a varying tip, which about 17% with meals < $10, drops to about 13% when meal cost starts increasing and then goes up again to about 14% as meals gets even more expensive.
- It seems unlikely that a single polynomial can **model all the data** (irrespective of increasing total bill), especially if the data is noisy and there may exist outliers at the two extremities.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Data from "*tips.csv*"

6

6

Scatter Plots

Using LOWESS

- LOcally Weighted Scatterplot Smoothing (LOWESS) was developed by William Cleveland in 1979 at AT&T Bell Labs.
- Lowess is a **non-parametric** technique that makes no assumption about the model that should fit the relationship between the independent and dependent variable[3].
- Lowess is computed **piecewise** (i.e. point by point) as a series of **least squares linear regression** about the point using a nearest-neighbour subset of the data.
- This subset of neighbouring points used at each point is determined by a parameter (e.g. `frac` in statsmodels's version of `lowess`) that creates a window size that is a faction of all points in the dataset.
- In each regression, points within the nearest-neighbours window are **weighted** by their proximity, usually using the tri-cubic weight function given by $w(x)=(1-|d|^3)^3$.

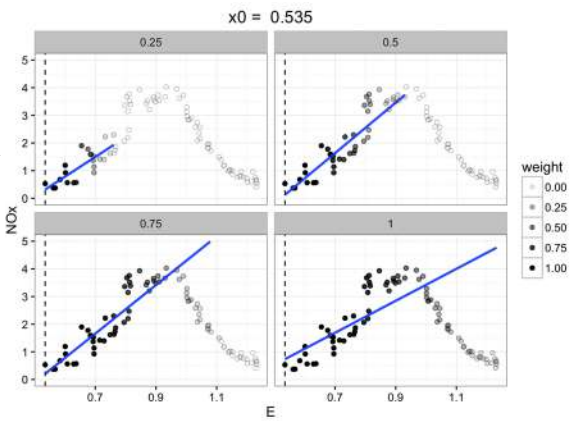
[3] J.B. Mailman, Data Smoothing in Data Science Visualization (2021), https://towardsdatascience.com/data-smoothing-for-data-science-visualization-the-goldilocks-trio-part-1-867765050615

Scatter Plots

How Does LOWESS Work?

- The `lowess` algorithms moves from **left to right** doing **piecewise** concatenation of individual least-squares linear regression using a **weighted** fraction of all nearest data points to the current point being computed.
- The Statsmodels Python module provides a useful implementation of `lowess`[4]. The size of the neighbourhood is given by `frac`, a float value between 0 to 1.
- It can be imported in Spyder with the following:

```
from statsmodels.nonparametric.smoothers_lowess import lowess
```



Animation by David Robinson from http://varianceexplained.org/files/loess.html

[4] Statsmodels Python library reference for its lowess method, https://www.statsmodels.org/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html

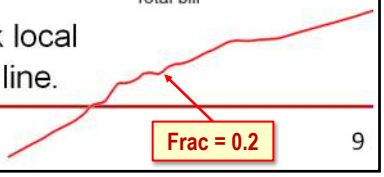
Scatter Plots

Appyling LOWESS

- Lowess with **`frac`**=0.4 was applied to the tips dataset.

```
LS = lowess(Tips, TotalBill, frac = 0.4)
ax = sns.lineplot(x=LS[:,0], y=LS[:,1])
```

- The results show that a tip of about 16% is given for meals < $20 but as the cost of the meal increases, the tip starts to drop to about 11%.
- By using a sizable portion (but not all) of the dataset, lowess reduces the **influence of outliers** but can still track the **changing trend** over the entire total bill.
- Reducing the fraction (**`frac`**) of data points used will track local variations better but reduces the smoothness of the trend line.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

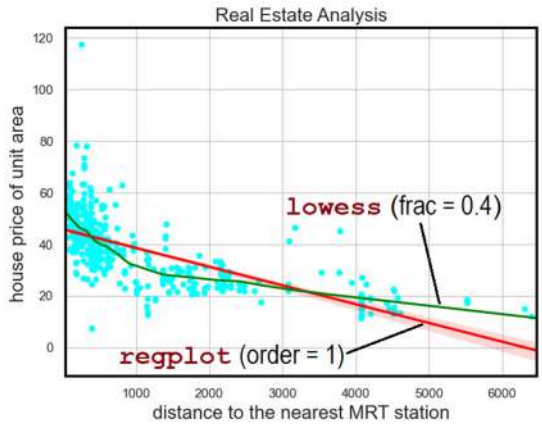
9

9

Think and Apply

What Affects Property Prices?

- A dataset of 414 property transactions that list various property-related data regarding each transaction.



- Unit area price of the property
- Age of the property
- Distance from a MRT station
- Number of convenient stores nearby

- Better property price nearer MRT station?
- Is older cheaper?
- Does more convenience have a price?
- What really makes the difference?

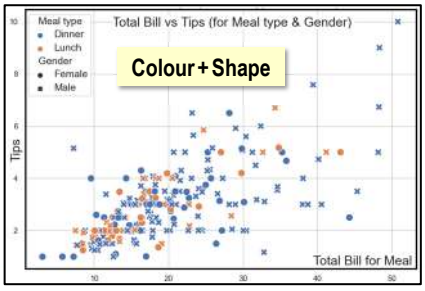
10

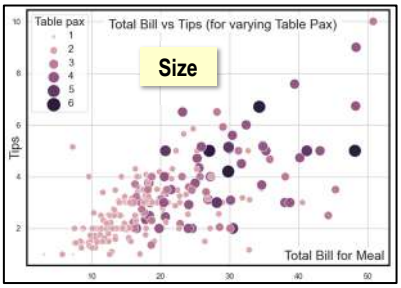
10

Bubble Plots

Handling Additional Data Dimensions

- Additional **categorical** variables (e.g. countries, gender) can be added to the scatter plot by using points with **different nominal visual attributes** (e.g. colour or shapes)[1].
- However, if the relationship of a **third** additional variable to be depicted is **ordinal** (i.e. ordered) or **quantitative** (i.e. numeric), then the **size the dots** can be used to encode this quantitative value in a **Bubble Plot**[5].



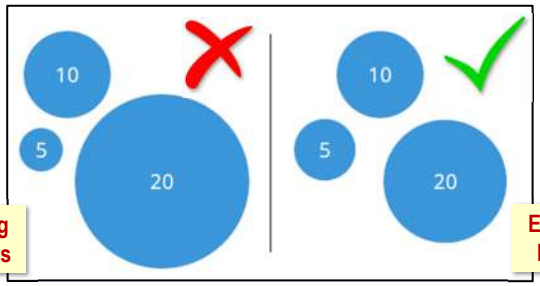


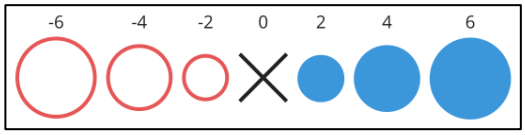
[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/
[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/

Bubble Plots

Encoding Quantitative Values to Bubble Size

- For **proportional** visual comparison, the bubble's **area** (not radius) is scaled to the value[5]. A dot with value *2n*, is only $(\sqrt{2})=1.41$ times the radius of a dot of value *n*.
- **Negative** values can be incorporated using **unfilled bubbles** with proportional areas. However, if variables with the negative ranges can be encoded in **one of the two axes** instead, that should be preferred[5].





Encoding negative values in bubble plots[5]

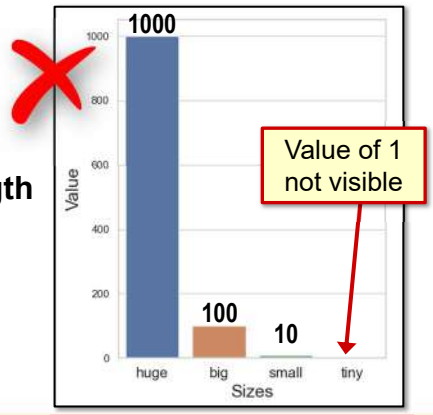
[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/

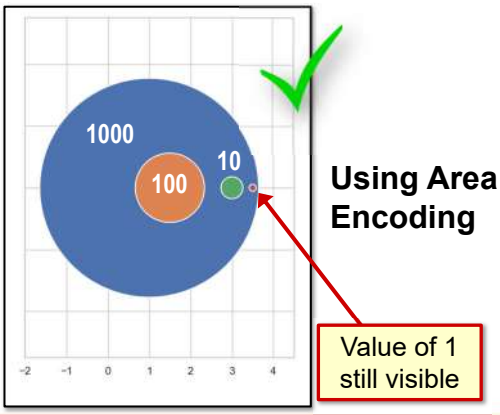
Bubble Plots

Length versus Area

- When the **differences** in magnitudes between categories are **large**, **area** encoding make better comparative visuals then value encoding using length[6].

Using Length Encoding





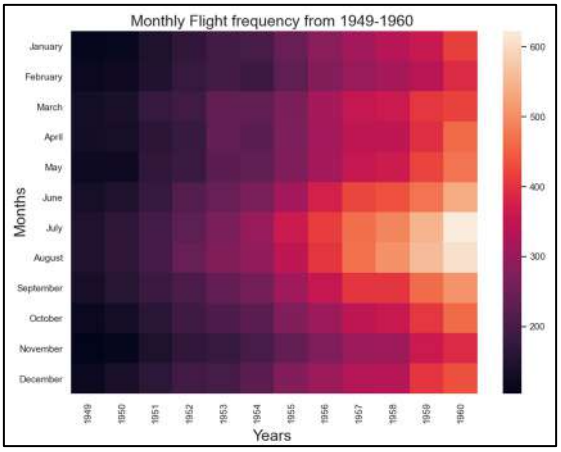
Using Area Encoding

[6] Nathan Yau, Data Points, Visualisation that Means Something, Wiley (2013)

Heatmaps

Seeing Relationships using Grids

- Heatmap depict values for a main variable of interest in **coloured rectangles** across two axis variables[7].
- Any patterns in the values for one or both of these axes can be observed by how the cell **colour changes** across each axis.
- These two axis variables are either **categorical** or if quantitative, they are binned into **discrete ranges** to form the desired number of cells.



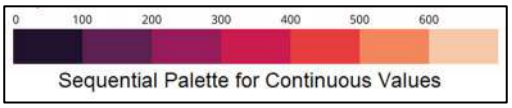
Heatmap plot using Seaborn's **heatmap()**[8]

[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/
[8] Seaborn, Plot rectangular data as a color-encoded matrix, https://seaborn.pydata.org/generated/seaborn.heatmap.html

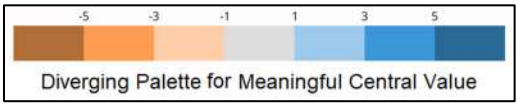
Heatmaps

Choosing the Colours

- Since colour is a key feature of heatmaps, an **appropriate colour palette** should be chosen to match the nature of the primary variable of interest[7].
- A **sequential** colour ramp between value and colour is often used. Lighter colours map to larger values & darker colours to smaller values (or vice versa).



- If the values have a meaningful zero point, a **diverging** colour palette would add better visual clarity.



- The colour distribution **legend** must be included in the heatmap to allow associated values of the colours to be interpreted.

[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/

Summary

Relation Plots

- Relation plots help us visualise **relationships** or **patterns** in the data.
- **Scatter plots** uses 2D position of dots to visualise **spatial relationships** (e.g. clustering, strong correlation, outliers, etc.) among data with **two** interval or ratio scale variables.
- **Linear regression** and **LOWESS** allow us to extract best fit trend lines within these scattering of dots.
- **Bubble plots** extend scatter plots ability to handle one more **quantitative dimension**, but discrimination of area variations is limited.
- **Heatmaps** allow us to visualised relationships over a **coloured grid**.

References for Relation Plots

[1] Mike Yi, A Complete Guide to Scatter Plots (2019), https://chartio.com/learn/charts/what-is-a-scatter-plot/

[2] Seaborn, Plot data and a linear regression model fit , https://seaborn.pydata.org/generated/seaborn.regplot.html

[3] J.B. Mailman, Data Smoothing in Data Science Visualization (2021), https://towardsdatascience.com/data-smoothing-for-data-science-visualization-the-goldilocks-trio-part-1-867765050615

[4] Statsmodels Python library reference for its lowess method, https://www.statsmodels.org/devel/generated/statsmodels.nonparametric.smoothers_lowess.lowess.html

[5] Mike Yi, A Complete Guide to Bubble Plots (2019), https://chartio.com/learn/charts/bubble-chart-complete-guide/

[6] Nathan Yau, Data Points, Visualisation that Means Something, Wiley (2013)

[7] Mike Yi, A Complete Guide to Heatmaps (2019), https://chartio.com/learn/charts/heatmap-complete-guide/

[8] Seaborn, Plot rectangular data as a color-encoded matrix, https://seaborn.pydata.org/generated/seaborn.heatmap.html

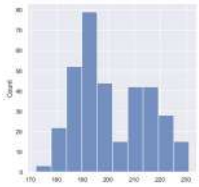
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

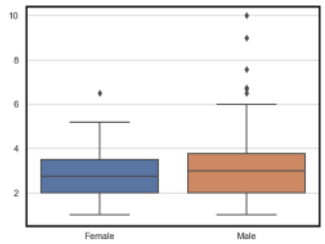
Note: All online articles were accessed between May to June 2021

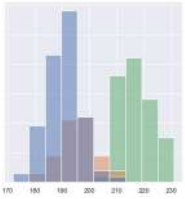
Chapter 5.4 – Visualising Distribution in Data

Contents

- Basic Distribution Plots
- Histogram
- Box Plots
- Violin Plots





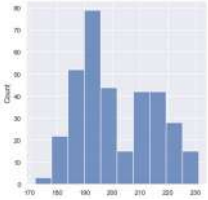


© A/P Goh Wooi Boon (SCSE/NTU)

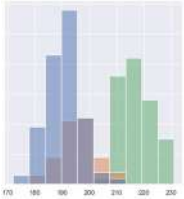
Basic Distribution Plots

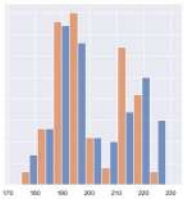
Visualising Distribution in the Data

- Distribution plots gives a visual insight into how the data **values are distributed** within the dataset.
- **Histograms** – are often used to visualise the distribution of a **single variable**. Multiple variables[1] are possible but may suffer from occlusion and clutter.
- **Box & Violin Plots** – are useful for exploring the distribution in **multiple variables** simultaneously. The violin plot visualises the **density curve** of the variable, whereas the box plot visualises the **median** and **interquartile range** of each variable.

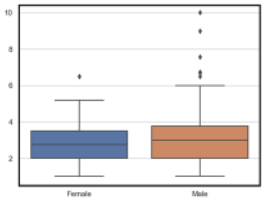


Standard Histogram

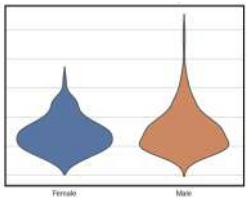




Histogram with Multiple Categories



Box Plot



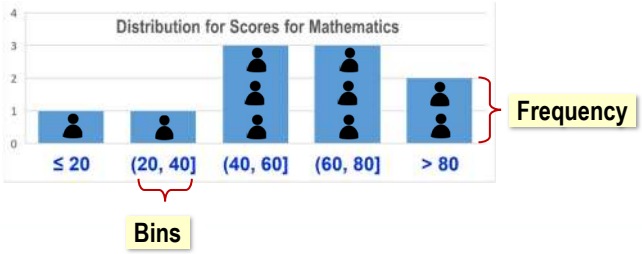
Violin Plot

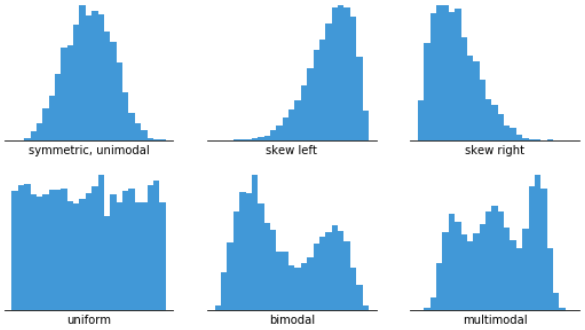
[1] Visualizing distributions of data (2019), https://seaborn.pydata.org/tutorial/distributions.html

Histograms

Visualising Data Distribution

- Histograms plot the distribution of a variable's numeric values as a series of bars. Each bar covers a range of numeric values (**bin**) and its **height** encodes the **frequency** of data points that have values within the corresponding bin.
- Histograms reveals the general distributional features of variables, their positions of peaks, whether the spread is skewed, symmetric, and if there are any outliers[2].





[2] Mike Yi, A Complete Guide to Histograms (2019), https://chartio.com/learn/charts/histogram-complete-guide/

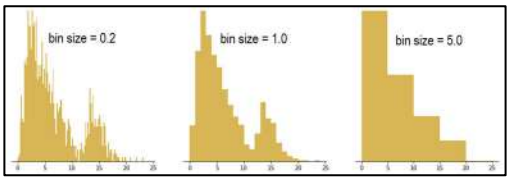
3

3

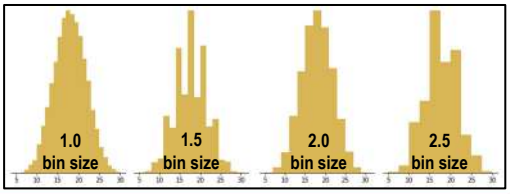
Histograms

Selecting Number of Bins and Boundaries

- The bin count and their boundaries for tallying data points must be chosen properly as it can have an effect of the **shape** of the plot and how it is interpreted[2].
- **Too many** bins may result in a **noisy** plot that is difficult to discern general data distribution. **Too few** bins may **reduce details** needed to discern patterns in the data (e.g. a double peak distribution).



- **Bin sizes** should match the **type of values** the data can have. Fractional bin sizes (e.g. 1.5) may not be suitable for integer variables as not all bins can take the same numbers of valid data values, leading to "bumpy" looking plots[2].



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[2] Mike Yi, A Complete Guide to Histograms (2019), https://chartio.com/learn/charts/histogram-complete-guide/

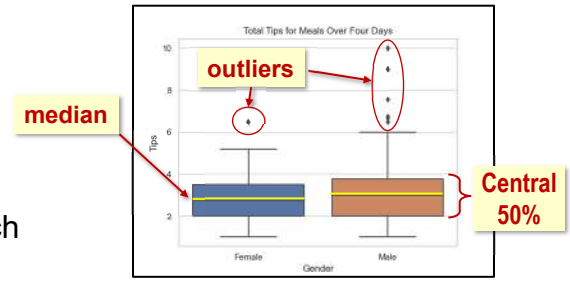
4

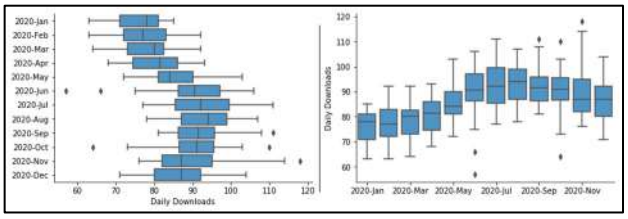
4

Box Plots

Visualising Statistical Properties of Multiple Variables

- Box plots (whisker plot) uses boxes and lines to depict the distributions of one or more numeric variables.
- The **box limits** show the range of the **central 50%** of the data, with a central line marking its **median**. Lines extend from each box to capture the **range** of the remaining data, with dots placed past the line edges to indicate outliers.
- Box plots allow multiple variables to be visualised either vertically or horizontally[3].





[3] Mike Yi, A Complete Guide to Box Plot (2019), https://chartio.com/learn/charts/box-plot-complete-guide/

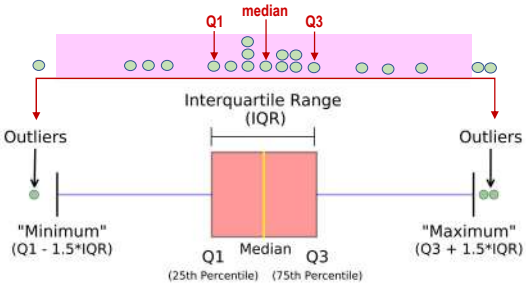
5

5

Box Plots

Constructing the Box and Whiskers

- Box plots display data variable's quartiles (values that divide the dataset into equal fourths).
- The 1st quartile (**Q1**) is greater than 25% of the data. The 2nd quartile (**Q2**), the **median**, sits in the middle, dividing the data in half. The 3rd quartile (**Q3**) is larger than 75% of the data. The box boundaries and its centre line mark the locations of Q1, Q2 and Q3[4].



- The left and right whiskers are **1.5 times** the interquartile range (IQR) from Q1 and Q3 respectively, but not exceeding data points at the extremities. All points outside these limits are considered **outliers** and are plotted individually horizontally[4].

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[4] M. Galarnyk, Understanding Boxplots (2018), https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

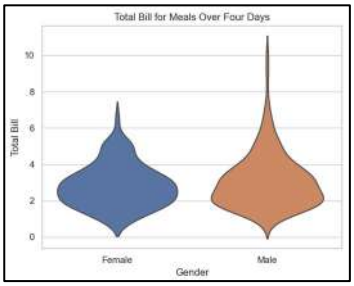
6

6

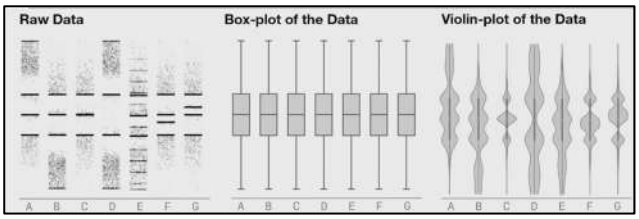
Violin Plots

Visualising Kernel Density Curves of Multiple Variables

- Violin plots visualise distributions of numeric data values using kernel density curves, whose widths encode the approximate frequency of data points in each region[5].
- Distribution comparison between multiple groups can be done as the peaks, valleys, and tails in of their respective density curves can be contrasted by the side-by-side plots.



- Datasets can be modified in a way that the quartiles do not change, but the shape of the distribution differs dramatically[6]. In such cases, violin plots give **more insights** into the **data distribution** than box plots.



[5] Mike Yi, A Complete Guide to Violin Plots (2019), https://chartio.com/learn/charts/violin-plot-complete-guide/

[6] J. Matejka, G. Fitzmaurice, Same Stats, Different Graphs (CHI'17), https://www.autodesk.com/research/publications/same-stats-different-graphs

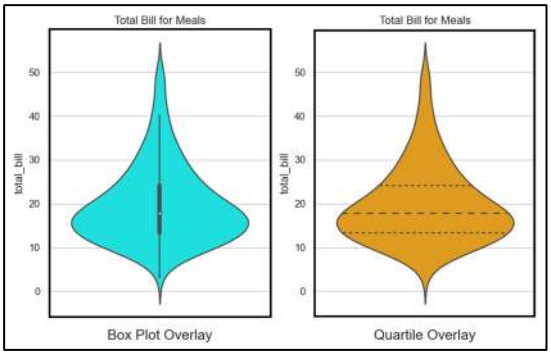
7

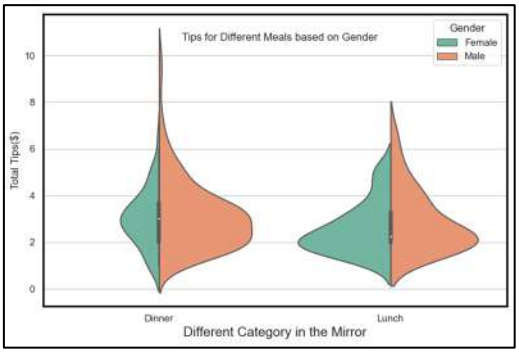
7

Violin Plots

Getting the Best of Two Worlds

- Violin plots have additional flexibility compared to box plots. They can also be overlayed with additional information such as the **box plot** or **quartile lines**.
- Each **mirrored half** of the violin plot can be used to encode the distribution of another **complementary** data category (e.g. females on the left and males on the right)[7].

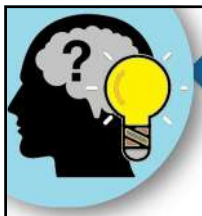




[7] Seaborn, Violin Plots, https://seaborn.pydata.org/generated/seaborn.violinplot.html

8

8

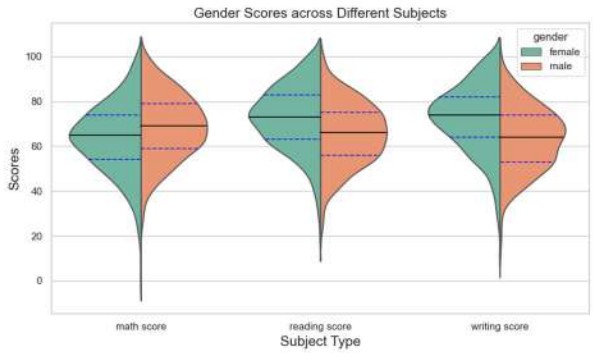


Think and Apply

What Affects Academic Performance?

- Analyse the math, reading and writing scores of 1000 students with different backgrounds, gender and preparedness. Looking at various score distributions, answer these questions:

  - What subjects do the male students do better in?
  - Which subject was the most difficult?
  - Did taking the prep course help improve scores?
  - Do ethnicity and parent's educational background influence performance?



Summary

Distribution Plots

- Distribution plots allow us to visualise the way the data values in the dataset is **distributed** across the permissible range.
- The **histogram** is a bar chart that shows the distribution of data values. Its shape can **reveal much more** (e.g. skewed, bimodal, etc) than what typical statistical measures like mean and variance can inform.
- **Box plots** is useful for visualising the **statistical** properties of **multiple variables** simultaneously.
- **Violin plots** allow the **kernel density curves** of multiple variables to be visualised and compared in a **pairwise** manner.

References for Distribution Plots

[1] Visualizing distributions of data (2019), https://seaborn.pydata.org/tutorial/distributions.html

[2] Mike Yi, A Complete Guide to Histograms (2019), https://chartio.com/learn/charts/histogram-complete-guide/

[3] Mike Yi, A Complete Guide to Box Plot (2019), https://chartio.com/learn/charts/box-plot-complete-guide/

[4] M. Galarnyk, Understanding Boxplots (2018), https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

[5] Mike Yi, A Complete Guide to Violin Plots (2019), https://chartio.com/learn/charts/violin-plot-complete-guide/

[6] J. Matejka, G. Fitzmaurice, Same Stats, Different Graphs (CHI'17), https://www.autodesk.com/research/publications/same-stats-different-graphs

[7] Seaborn, Violin Plots, https://seaborn.pydata.org/generated/seaborn.violinplot.html

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

Data Visualisation

Associate Professor Goh Wooi Boon
School of Computer Science and Engineering
Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 6 – Visual Perception

Contents

- Human Visual Perception
- Gestalt Principles in Data Visualisation
- Colour Perception

© A/P Goh Wooi Boon (SCSE/NTU)

Chapter 6.1 – Human Visual Perception

Contents

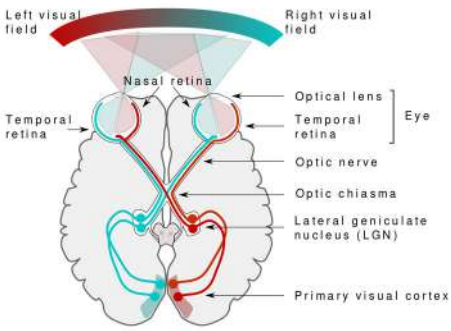
- The Visual Brain
- Estimating Magnitude
- Estimating Change
- Estimating Rate of Change
- Pre-attentive Visual Processing
- Multiple Visual Attributes
- Visual Background Perception

© A/P Goh Wooi Boon (SCSE/NTU)

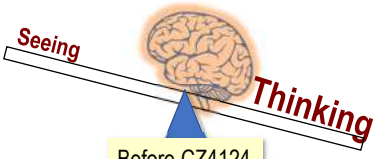
The Visual Brain

The Picture Tells the Story

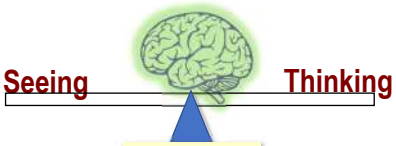
- The reality is that we see with our **brains** and not our eyes.
- Our eyes receive light signals, and these are relayed via electrical impulses along various neural pathways to the brain, where the **perception** and **sense-making** occurs.
- Designing effective visualisation requires us to **understand** how the human brain **perceives**, **organises** and **make sense** of visual information[1].



Visual System
image from Wikipedia



Before CZ4124



After CZ4124

[1] Stephen Few, Data Visualization for Human Perception - https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception

The Visual Brain

The Memory Model

- Human visual processing is supported by **iconic** memory (sensory memory), **short-term** memory (working memory) and **long-term** memory; each having different roles and characteristics[2].

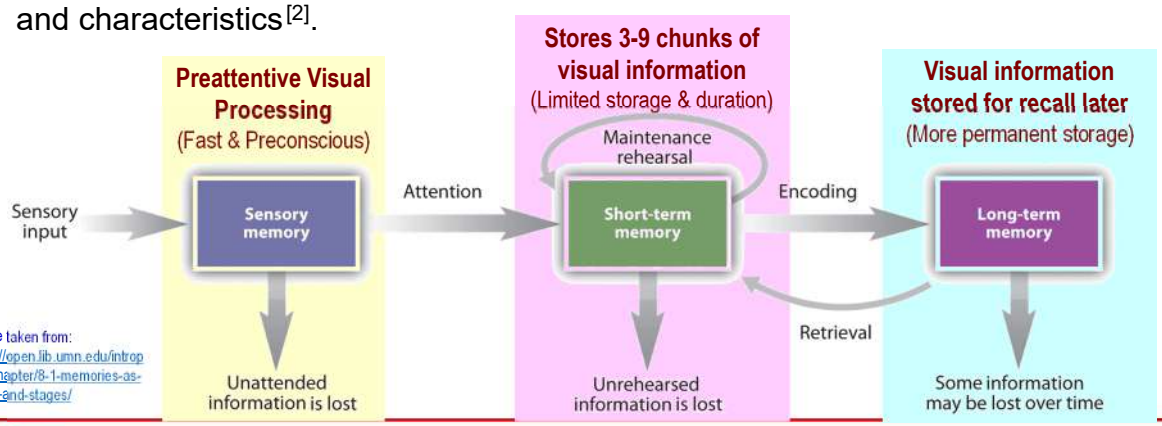


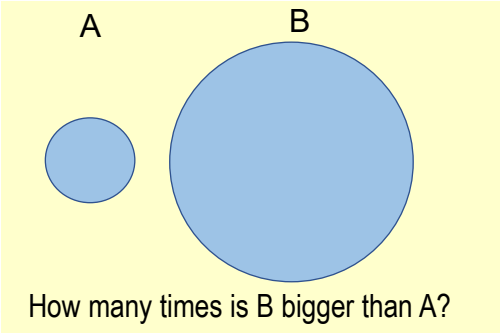
Image taken from: https://open.lib.umn.edu/intropsyc/chapter/8-1-memories-as-types-and-stages/

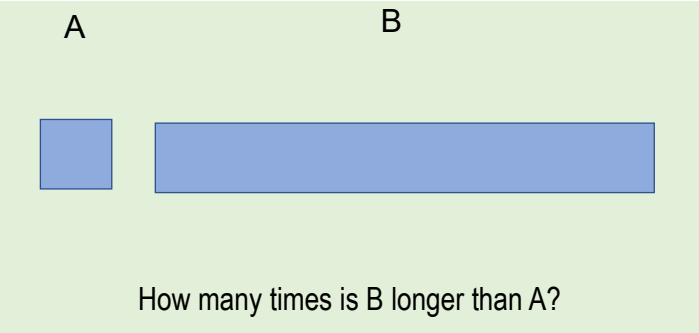
[2] R.C. Atkinson, R.M. Shiffrin, Human memory: A proposed system and its control processes. in K. Spence (Ed.), The psychology of learning and motivation (Vol. 2). Oxford, England: Academic Press (1968).

Estimating Magnitude

Big, Bigger, Biggest

- The human visual system is much better at **estimating magnitude** based on visual **length** than visual **area**.
- For this reason, bar charts are much better in presenting accurate visual information than bubble charts, even though they may be prettier to look at[3].





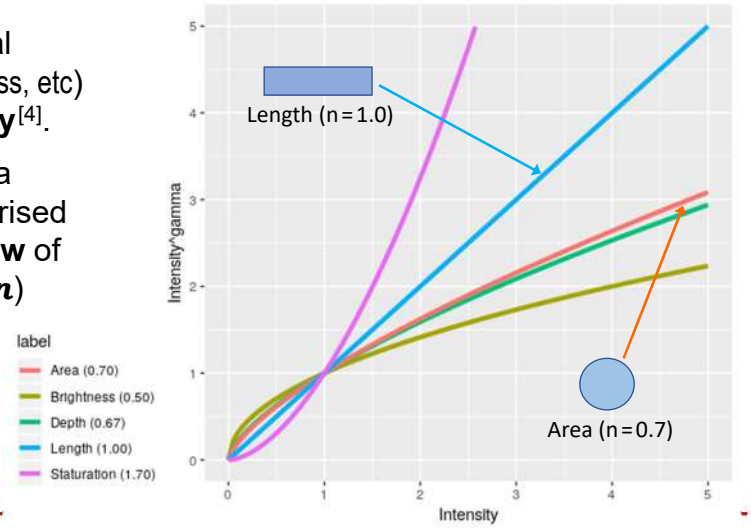
[3] Steve Wexler, Why the f**k do we see so many bar charts? (2021), https://www.linkedin.com/pulse/why-fk-do-we-see-so-many-bar-charts-steve-wexler

Estimating Magnitude

Stevens's Power Law

- Humans perceive different visual channels (e.g. length, area, brightness, etc) with **different** levels of **accuracy**[4].
- Our perceived sensation (***S***) of a physical intensity (***I***) is characterised by the psychophysical **power law** of **Stevens**, where the exponent (***n***) depends on the exact sensory modality of the measured ***I***.

$$S = I^n$$



[4] T. Munzner, Visualization Analysis & Design, CRC Press (2015)

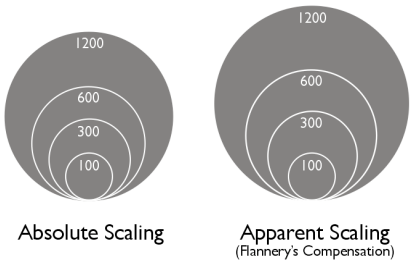
7

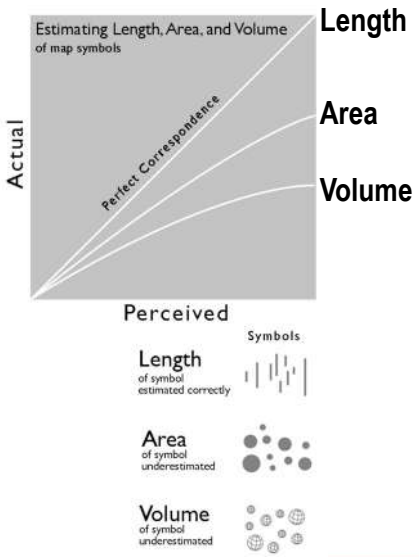
7

Estimating Magnitude

Implication of Stevens's Power Law

- Psychophysical research revealed that people tend to correctly estimate lengths but they **underestimate** areas and volumes[5].
- ArcGIS, a geographic information system (GIS) uses James Flannery's **Appearance Compensation**[5] method to scale up proportional circles (i.e. the larger the circle, the more the scaling).
- **Note:** Edward Tufte[6] only supports absolute scaling for faithfulness to the data.





[5] John Krygler, Perceptual Scaling of Map Symbols - https://makingmaps.net/2007/08/28/perceptual-scaling-of-map-symbols/

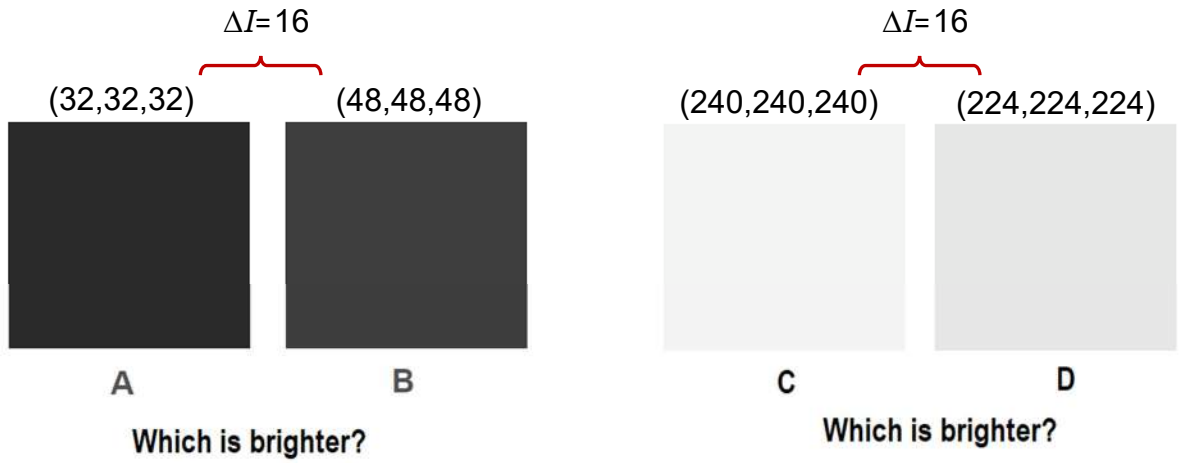
[6] Edward Tufte, The Visual Display of Quantitative Information Graphic Press, 2nd Ed (2001)

8

8

Estimating Change

Same Difference

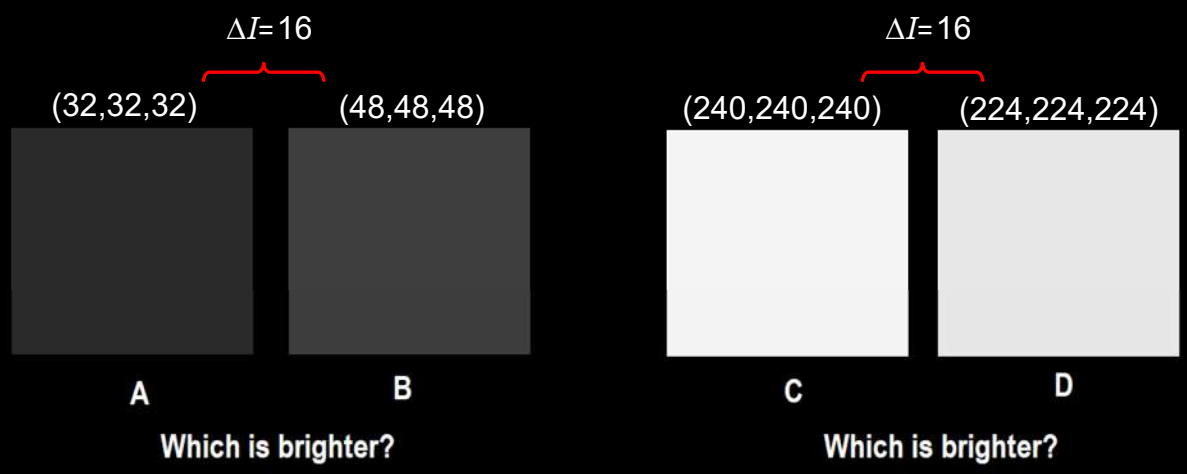


9

9

Estimating Change

Same Difference



10

Estimating Change

Same Difference

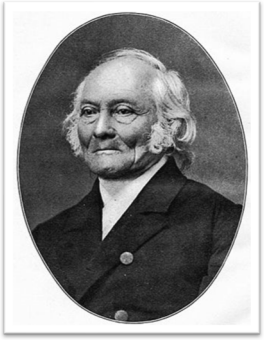
- The **Just Noticeable Difference** (JND) is the minimum amount by which stimulus intensity must be changed in order to produce a noticeable variation in sensory experience.
- Ernst Weber[7] observed that the size of the **difference threshold** appeared to be related to the initial **stimulus magnitude**.
- JND is govern Weber's Law and is given by

Addition to intensity I required for the change to be perceived

Current intensity of stimulus

$$\frac{\Delta I}{I} = k$$

Weber constant



Ernst Heinrich Weber (24 Jun 1795 – 26 Jan 1878), German physician & early pioneer of Experimental Psychology

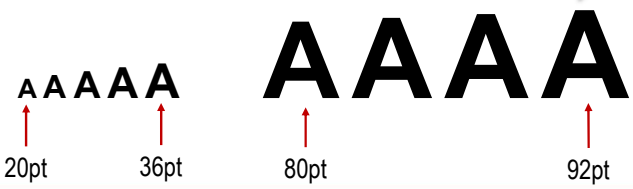
[7] Weber's Law - https://www.britannica.com/science/Webers-law

Estimating Change

Implication of Weber's Law

- According to Weber's Law, the higher the intensity (or length), the larger the disparity required for us to sense the change.
- Visualisation designs that takes such perceptual behaviour into account will facilitate more accurate visual comparison (e.g. framing long bar values to make their small difference more apparent)[8].
- Weber's law also applies to choosing **shape** or **font sizes**. As the shapes or fonts get larger, the absolute difference must be made larger to allow changes to be noticeable.

Each Arial font "**A**" increases by 4 points

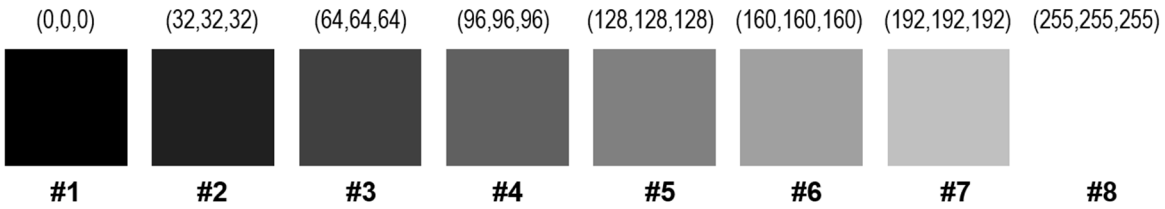


[8] W.S. Cleveland, The Elements of Graphing Data, Hobart Press (1994)

Think and Apply

Encoding Ordinal Data Using Grey Values

- Which two grey squares will be most difficult to tell apart with a white background?
- What if the background is black? What background colour can maximise the discriminability of all these grey levels if their values cannot be changed?
- How would you change the grey values to improve their discriminability in white background?

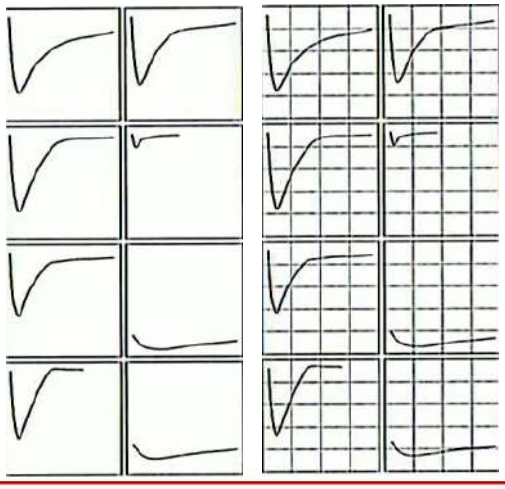


Eight grey squares with equal intensity differences

Estimating Change

Using Visual Reference Grids

- Weber's Law explains why visual reference grids enhance pattern perception.
- Without the grid, estimation of lengths with small percentage differences is difficult[8].
- The grids shorten the base lengths that are being compared, making it easier to compare highs, lows, and steady state behaviour[8].
- Graphs can be compared by superimposing them, but this only works for limited numbers of plots before clutter and line differentiation becomes problematic.

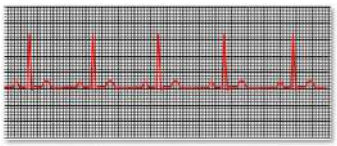


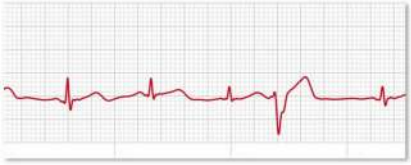
[8] W.S. Cleveland, The Elements of Graphing Data, Hobart Press (1994)

Estimating Change

Proper Layering of Visual Reference Grids

- If visual reference grids are not layered correctly, they can be distracting and make the actual data difficult to visualise.





Competing signal & background in two different electrocardiogram trace lines. The prominent gridlines & poor contrast makes the trace of the left more difficult to read than the one on the right.

- Heer and Bostock[9] crowdsourced experiments on an acceptable luminance contrast settings for visual reference elements such as gridlines showed a safe **Alpha** setting of about **20%** (Alpha 0% = Total transparent, 100% = Opaque).

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[9] J. Heer, M. Bostock, Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design (CHI 2010) - http://www.pensivepuffin.com/dwmcphd/syllabi/insc547_wi13/papers/crowd/heer-crowdsourcing-CHI10.pdf

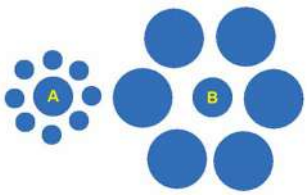
15

15

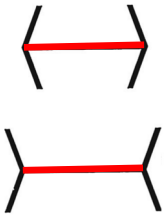
Estimating Magnitude

Context Matters

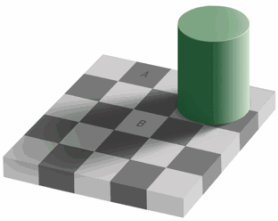
- Our ability to estimate the magnitude of visual attributes (e.g. length, size, colour, parallelism, etc) can be influenced by the context in which it is visualised.
- Be mindful of these contextual influences on human visual perception in your visualisation design.



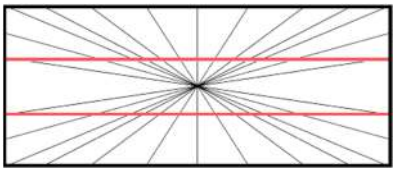
Titchener illusion



Müller-Lyer illusion



Adelson's Checker Shadow illusion



Hering illusion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

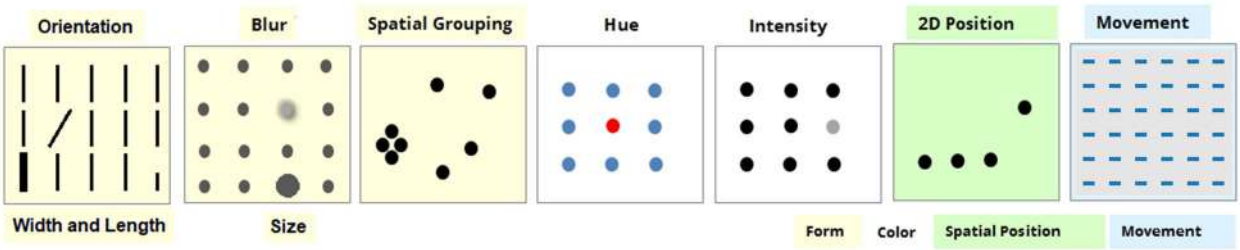
16

16

Preattentive Visual Processing

Standing Out

- Preattentive processing occurs at the **early stage** of visual perception and is tuned to **rapidly detect** a specific set of visual attributes at a **sub-conscious** level[10].
- The **sequential attentive** processing used to find a specific visual target is **slower**.



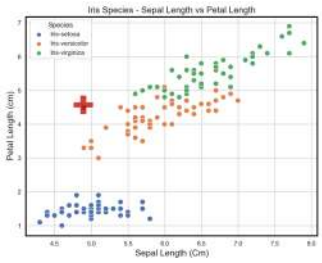
Some examples of the preattentive attributes of human visual perception

[10] Stephen Few, Information Dashboard Design, O'Reilly Media (2006)

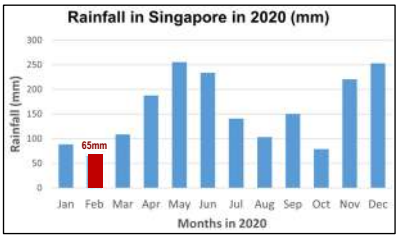
Preattentive Visual Processing

Exploiting Preattentive Processing

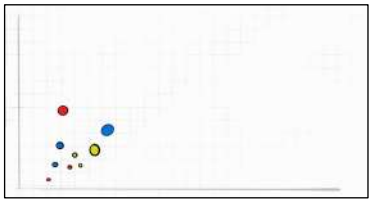
- Preattentive visual attributes can be used to **highlight** (i.e. make it pop out) particular data points of interest by making them **distinct** on a particular feature channel (e.g. coloured when the rest are grey scaled) or made more distinct by an **appropriate** redundant combination of **multiple preattentive attributes** (e.g. colour and size)[11].



Distinct colour,shape & size to highlight interesting outlier



Distinct colour bar (with annotation) to denote driest month



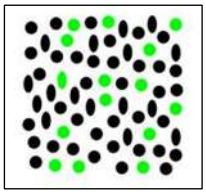
Flickering animation highlights data of interest

[11] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

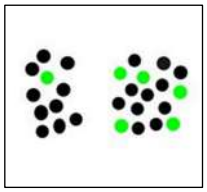
Preattentive Visual Processing

Conjunction Search

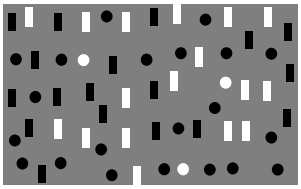
- Data points can be encoded with multiple visual attributes. A visual search for an object with more than one attribute (e.g. green ellipse) is a **conjunction search**. Conjunction searches are generally **not preattentive**[11]. .
- However, there are conjunction of some attributes that support preattentive search (e.g. position and colour or luminance polarity and shape)[11].



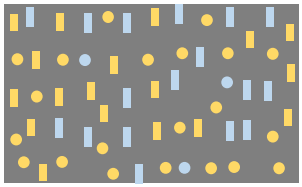
Search is slow with the conjunction of shape and colour



Search is fast with conjunction of position & colour (left green object)



Luminance polarity with targets (white circles) lighter & darker than a grey backdrop supports preattentive conjunction



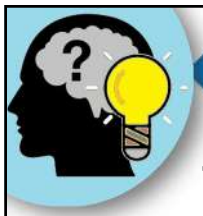
The same colour and shape encoding has no preattentive conjunction search property

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[11] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

19

19



Think and Apply

Telling It Like It Is – Effective Visual Story

- The table shows the average levels of an imaginary hormone Vitalis in a population based on age group, gender and BMI. The story you want to tell is that the only group with increasing levels of Vitalis as they age are females with BMI < 25.
- How would you design a chart to make this **story stand out** while providing all the information shown in the table.

	Levels of the hormone Vitalis			
	Males		Females	
Body Mass Index (BMI)	Under 60 years	60 years or over	Under 60 years	60 years or over
Under 25	255	230	380	550
25 or over	440	325	720	500

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

20

20

Summary

Human Visual Perception

- Human visual perception is **complex** because visual interpretation takes place in the **brain** and is **influenced** by many **factors** (e.g. our memories) besides the visual stimulus entering our eyes.
- Characteristics of the different visual channels based on Stevens's power law and Weber's law should influence the way we design **visual encoding** for data visualisation.
- Understanding **preattentive** visual processing can help us exploit its characteristics in designing visuals that can capture people's **attention** and avoid designing visuals that are **not effective** in communicating **useful meaning** in the data patterns.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

21

21

References for Visual Encoding

[1] Stephen Few, Data Visualization for Human Perception - https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception

[2] R.C. Atkinson, R.M. Shiffrin, Human memory: A proposed system and its control processes.,in K. Spence (Ed.), The psychology of learning and motivation (Vol. 2). Oxford, England: Academic Press (1968).

[3] Steve Wexler, Why the f**k do we see so many bar charts? (2021), https://www.linkedin.com/pulse/why-fk-do-we-see-so-many-bar-charts-steve-wexler

[4] T. Munzner, Visualization Analysis & Design, CRC Press (2015)

[5] John Krygler, Perceptual Scaling of Map Symbols - https://makingmaps.net/2007/08/28/perceptual-scaling-of-map-symbols/

[6] Edward Tufte, The Visual Display of Quantitative Information Graphic Press, 2nd Ed (2001)

[7] Weber's Law - https://www.britannica.com/science/Webers-law

[8] W.S. Cleveland, The Elements of Graphing Data, Hobart Press (1994)

[9] J. Heer, M. Bostock, Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design (CHI 2010) - http://www.pensivepuffin.com/dwmcphd/syllabi/insc547_wi13/papers/crowd/heer-crowdsourcing-CHI10.pdf

[10] Stephen Few, Information Dashboard Design, O'Reilly Media (2006)

[11] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

22

22

Chapter 6.2 – Gestalt Principles in Visualisation

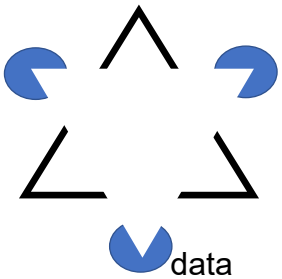
Contents

- Gestalt Psychology
- Proximity
- Similarity
- Simplicity
- Connectedness
- Enclosure
- Symmetry
- Figure-Ground

© A/P Goh Wooi Boon (SCSE/NTU)

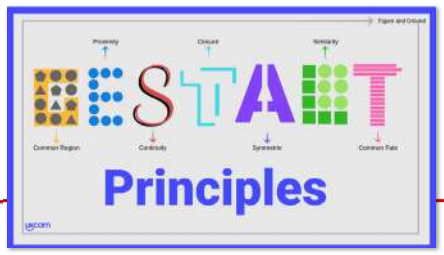
Gestalt Psychology

The Whole is Other than the Sum of Parts



- Gestalt psychology describes various principles or laws that govern how the human visual system forms **patterns** and **visual groupings** to simplify the interpretation of visual stimuli received.
- These Gestalt principles explain the **pattern-seeking** behaviour of human visual perception and should therefore be used in visualisation to help **communicate visual information** more effectively.
- Some Gestalt principles explored include:

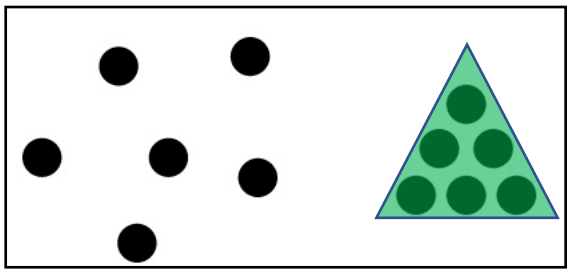
  - **Proximity**
  - **Similarity**
  - **Simplicity**
  - **Connectedness**
  - **Continuity**
  - **Symmetry**
  - **Figure & Ground**



Proximity

Closer is Together

- Principle of **proximity** states that visual elements **close to** each other are perceived as **belonging** to the same group.

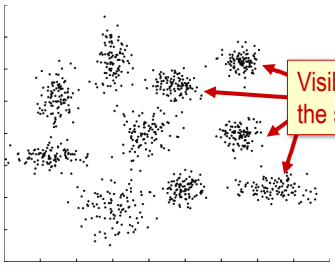


Proximity of shapes creates a larger unit entity

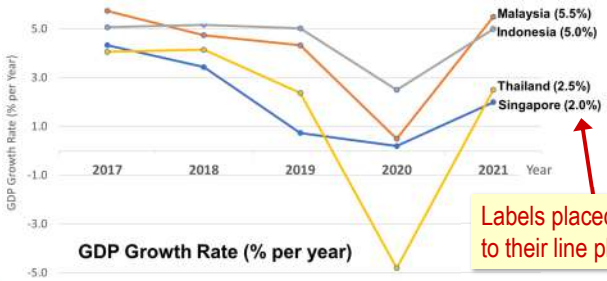
Proximity

Closer is Together

- Principle of **proximity** states that visual elements **close to** each other are perceived as **belonging** to the same group.
- Clusters in scatter plots are identified based on the proximity of the dots.
- Labels are placed close to their respective line plots to strengthen association to their corresponding lines[1].



Visible clusters in the scatter plot



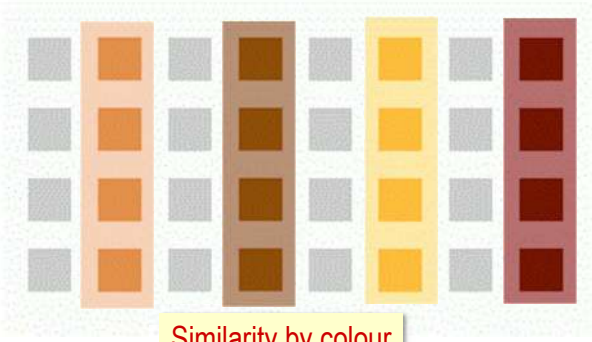
Labels placed close to their line plots

[1] Vidya Kesavan, Gestalt Laws Applied to Data Visualization (2018), https://medium.com/@vidya83.kesavan/gestalt-laws-applied-to-data-visualization-1e18385b7b00

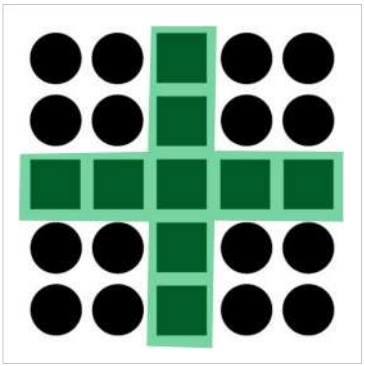
Similarity

Birds of a Feather Flock Together

- Principle of **similarity** states that visual elements that are **similar** to each other tend to be perceived as a **unified group**.



Similarity by colour



Similarity by shape

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

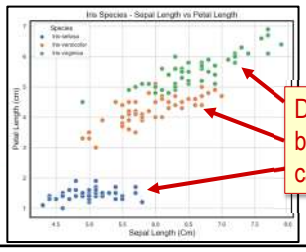
5

5

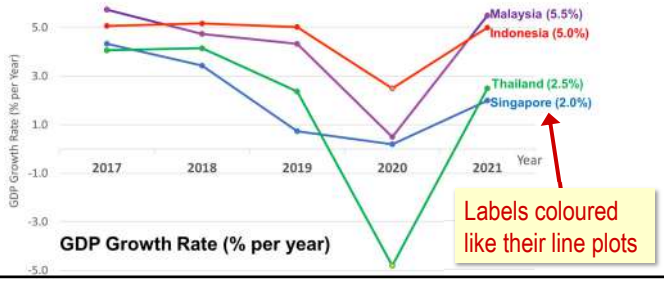
Similarity

Birds of a Feather Flock Together

- Principle of **similarity** states that visual elements that are **similar** to each other tend to be perceived as a **unified group**.
- Similarity-based perceptual grouping can be influenced by various visual features, including shape, colour, orientation, size or even motion.
- **Colour** can be used in scatter plots to identified data belonging to similar categories.
- Labels with the **same colour** as the line plots can further strengthen association to their corresponding lines.



Dots coloured based on their categories



Labels coloured like their line plots

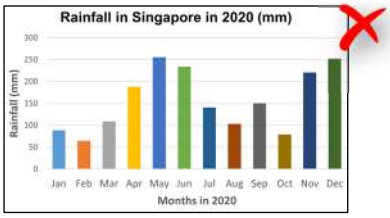
6

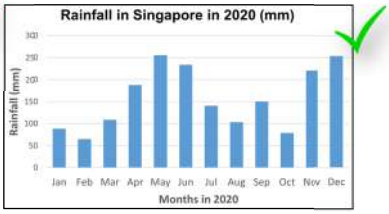
6

Similarity

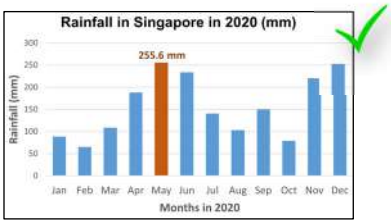
Exploiting Similarity

- Elements that share the **same visual attributes** will be perceived as being **related**. Consider carefully if the bar charts should have different hues or a uniform hue.
- Dissimilarity can be used to do the reverse and **highlight** a particular **exception** or **focal point** in the visualisation by giving it a distinctly different visual attribute.





Ordinal categories (e.g. months in a year) should have uniform colour as varying hue increases cognitive load and disrupt clarity of changing trend [2]



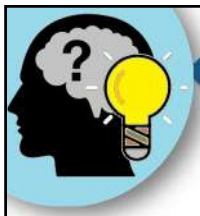
Wettest month in 2020 made to stand out in a different colour [2]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[2] William Craig, How to Make Data Visualization Better with Gestalt Laws, https://www.webfx.com/blog/web-design/data-visualization-gestalt-laws/

7

7



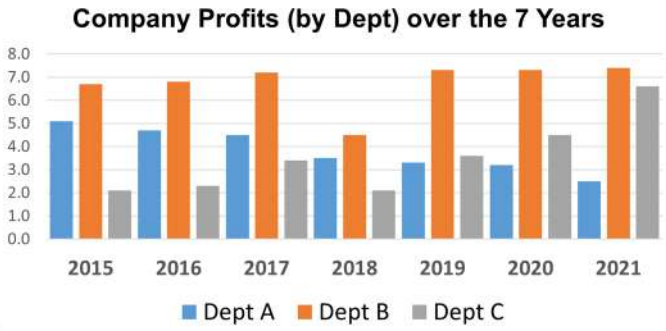
Think and Apply

Improving Visualisation using Gestalt Principles

- How can applying Gestalt principles improve the visualisation of the changing company profits over the years and the changing trend within each department?

	Profits in last 7 years (S$ M)		
Year	Dept A	Dept B	Dept C
2015	5.1	6.7	2.1
2016	4.7	6.8	2.3
2017	4.5	7.2	3.4
2018	3.5	4.5	2.1
2019	3.3	7.3	3.6
2020	3.2	7.3	4.5
2021	2.5	7.4	6.6

Table of each Department's Profits in the last 7 years



Visualisation of Company's Changing Profits

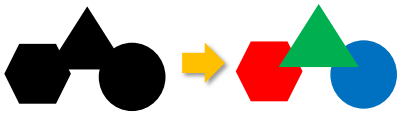
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

8

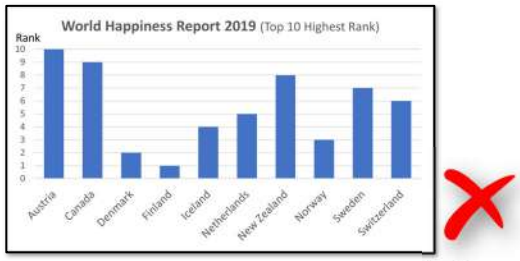
8

Simplicity

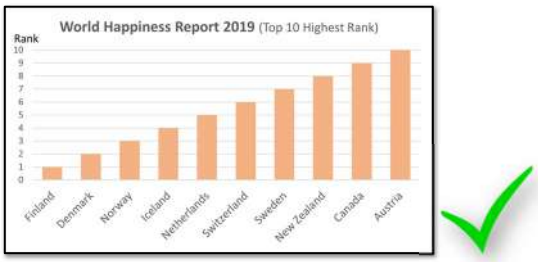
Simplicity is the Ultimate Sophistication



- Principle of simplicity states that visual elements are always perceived in the **easiest way possible**.
- **Ordering for simplicity** - visual information (e.g. values in bar charts) should be sorted and ordered (if possible) to increase its visual simplicity, thus easing interpretation.



Ordering based on category creates confusing visual



Value-based ordering improves visual simplicity of chart

Simplicity

Data-Ink Ratio

- Edward Tufte proposed the concept of **data-ink ratio** for data visualisation[3].
- This ratio measures the proportion of **ink** that is used to present **actual data** compared to the **total amount of ink** (or pixels) used in the entire visual.
- The goal in good data visualisation is to **maximise** the data-ink ratio.

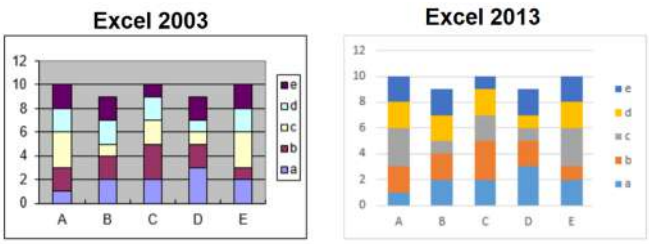


Image adapted from [4]

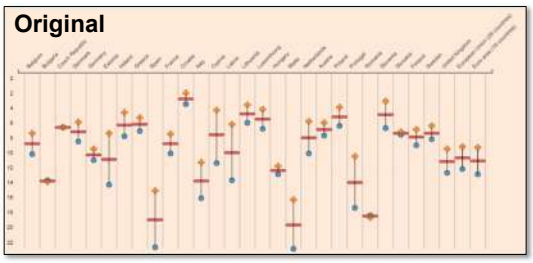
Changes to Excel charts over the years to maximise data-ink ratio

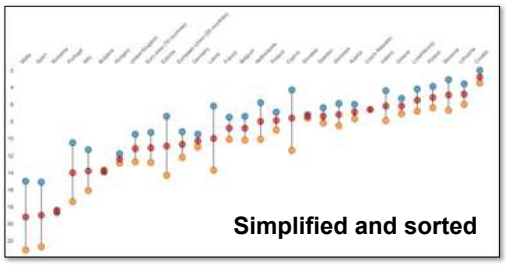
[3] Edward Tufte, The Visual Display of Quantitative Information Graphic Press, 2nd Ed (2001)
[4] Carlos Moura, Step 10 Visualisation - https://composite-indicators.jrc.ec.europa.eu/sites/default/files/COIN%202018%20Step%209%20Visualization.pdf

Simplicity

Data-Ink Ratio Maximisation

- Data-ink ratio can be maximised by considering what can be **removed** from the chart while **keeping the information** that needs to be communicated.
- Often unnecessary axes and background can remove without significantly affecting the ability to interpret the important information.





An example[4] of improving visual simplicity by maximising data-ink ratio and ordering the data

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[4] Carlos Moura, Step 10 Visualisation - https://composite-indicators.jrc.ec.europa.eu/sites/default/files/COIN%202018%20Step%209%20Visualization.pdf

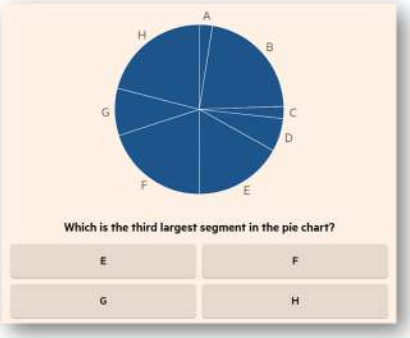
11

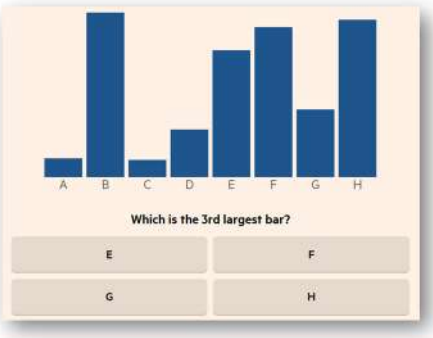
11

Think and Apply

Making it Easy for Others

- How would you redesign these charts to help the readers answer the quiz questions more easily?





NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[5] Visuals for Quiz Questions taken from : https://ig.ft.com/science-of-charts/

12

12

Connectedness

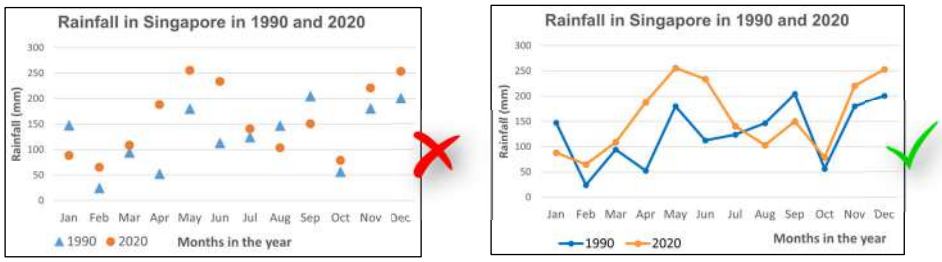
Close Links Bond

- Principle of connectedness states that visual elements that are **connected** to each other are perceived as a **single unit**.

Connectedness

Close Links Bond

- Principle of connectedness states that visual elements that are **connected** to each other are perceived as a **single unit**.
- Related data points should be connected with appropriate lines to allow fast visual recognition of their grouping and allow changing trends to be readily perceived[6].



Lines in graphs create clear visual connection between related points that even similar colour and shape attributes cannot evoke.

[6] B. Wong, Gestalt Principles (Part 1), Nature Methods, Vol 7, pp. 863 (2010)), https://www.nature.com/articles/nmeth1110-863

Gestalt Principles in Data Visualisation

Continuation ……

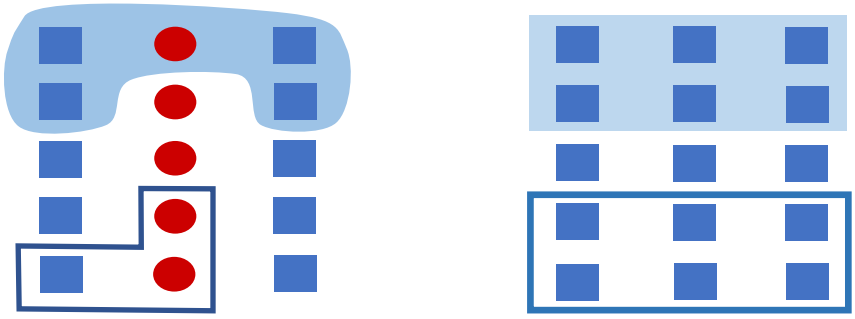
- The Gestalt principles explain the **pattern-seeking** behaviour of human visual perception and it perceive visual groupings to simplify the interpretation of visual stimuli.
- Some Gestalt principles explored last week include:

• Proximity	• Connectedness	• Symmetry
• Similarity	• Continuity	• Figure & Ground
• Simplicity	• Enclosure	

Enclosure

Feeling Boxed In

- Principle of enclosure states that visual elements are seen as belonging to a group if they are **enclosed** in a way that creates a **boundary** or **border around** them .



Enclosure

Feeling Boxed In

- Principle of enclosure states that visual elements are seen as belonging to a group if they are **enclosed** in a way that creates a **boundary** or **border around** them .

- Enclosure can be useful in organising a complex interface, like on the instrument dashboard of a 747 plane[7].
- Square and rectangular frames group similar categories of controls and display together.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[7] A. Rutledge, Gestalt Principles of Perception - 3: Proximity, Uniform Connectedness, and Good Continuation (2009) - https://www.andyrutledge.com/gestalt-principles-3.html

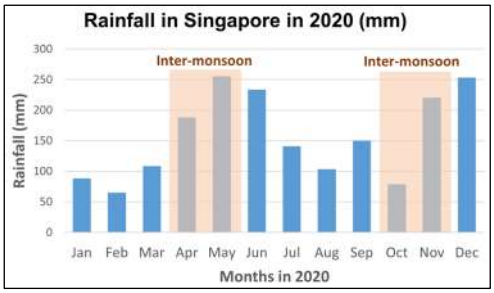
17

17

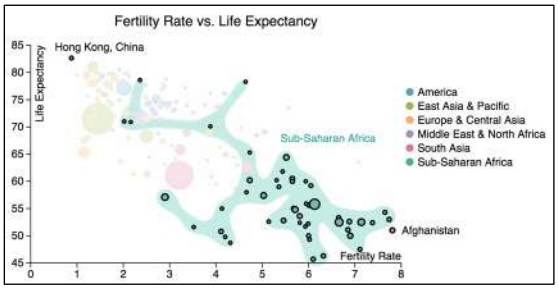
Enclosure

Feeling Boxed In

- Principle of enclosure states that visual elements are seen as belonging to a group if they are **enclosed** in a way that creates a **boundary** or **border around** them .
- Enclosure can be used to **highlight** and **group** specific data in the plot or chart.



Transparent enclosure used to highlight data points during the inter-monsoon period



Enclosure used to group dots in scatter plot belonging to a highlighted category. Image from[8]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[8] ChartAccent, Visualisation Examples - https://chartaccent.github.io/

18

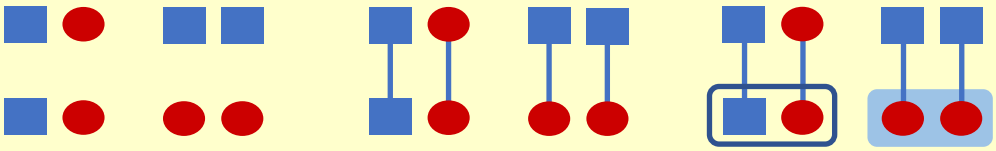
18

Enclosure

Dominant Gestalt Principles

- The relative strength each Gestalt principle exerts on perceptual grouping is not the same and there are many situations where different principles will compete to provide a different sense of visual grouping.
- Connectedness through the lines create a much stronger sense of grouping than either proximity or similarity[6].
- Grouping by enclosure resulting in elements bounded in a common region is powerful enough to overcome similarity, proximity and connection[6].

Relative perceptual grouping strength



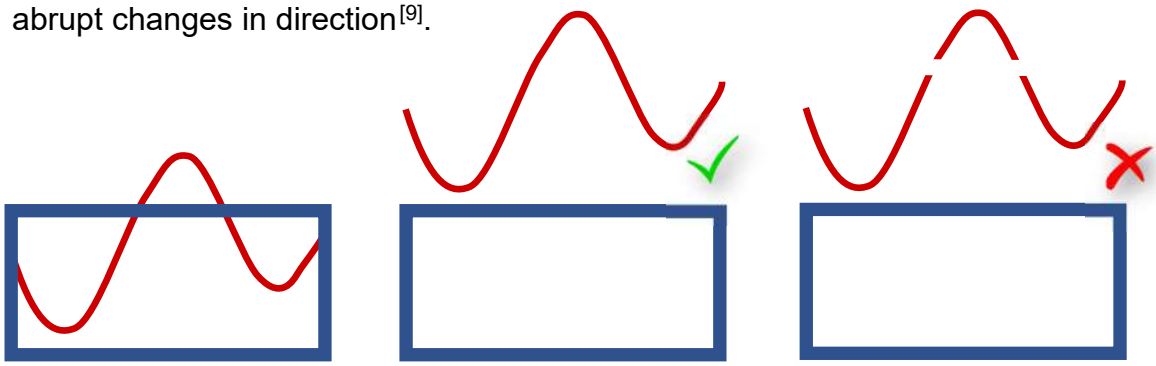
Proximity & Similarity < Connectedness < Enclosure

[6] B. Wong, Gestalt Principles (Part 1), Nature Methods, Vol 7, pp. 863 (2010)), https://www.nature.com/articles/nmeth1110-863

Continuity

It Cannot Change So Quickly

- Principle of continuity states that we are more likely to construct visual entities out of visual elements that are smooth and continuous, rather than ones that contain abrupt changes in direction[9].



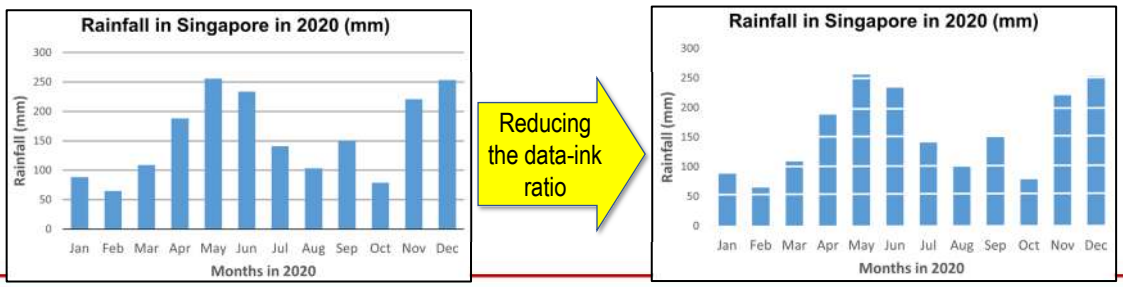
Our minds tend to perceive a line as continuous and continuing in its established direction.

[9] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

Continuity

It Cannot Change So Quickly

- Principle of continuity states that we are more likely to construct visual entities out of visual elements that are smooth and continuous, rather than ones that contain abrupt changes in direction[9].
- This sense of continuity can be exploited to reduce the data-ink ratio of bar charts by removing the axes and plotting the axes tick lines within the bars themselves[3].

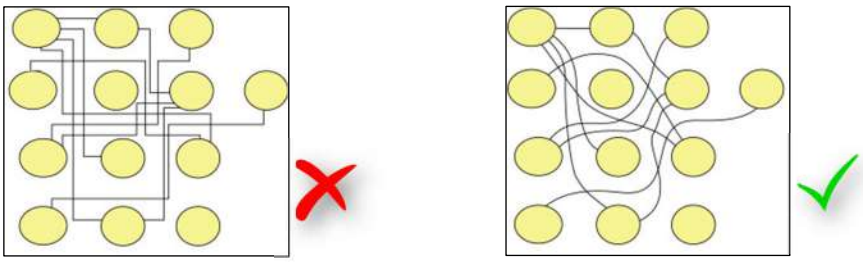


[3] Edward Tufte, The Visual Display of Quantitative Information Graphic Press, 2nd Ed (2001)

Continuity

We Follow Smooth Lines Better

- The principle of good continuity can be applied to the problem of drawing diagrams consisting of networks of **nodes** and the **links** between them.
- The human visual system finds it easier to identify the sources and destinations of **connecting lines** if they are **smooth** and **continuous**[9].



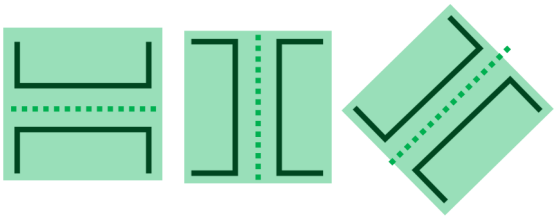
Where possible, smooth lines should be used to link nodes in a network graph

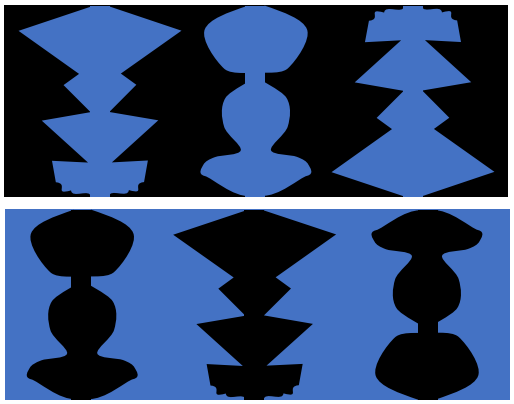
[9] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

Symmetry

Seeing Both Sides

- The principle of symmetry states that visual elements that are symmetrical to each other tend to be perceived as a unified group.





NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

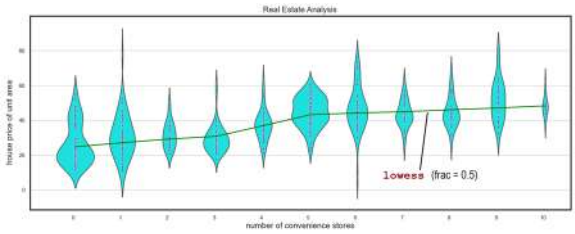
23

23

Symmetry

Seeing Both Sides

- The principle of symmetry states that visual elements that are symmetrical to each other tend to be perceived as a unified group.
- Symmetry can be exploited in **violin plots** to visualise the shape of the KDE distributions or the differences between two categories in a split violin plots.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Distributions visualised in the symmetry

Distribution variations in dissymmetry

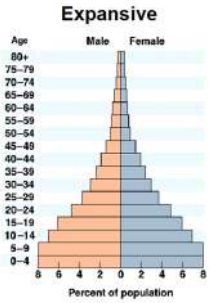
24

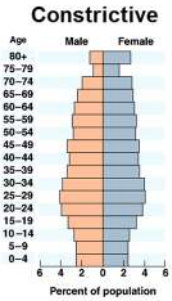
24

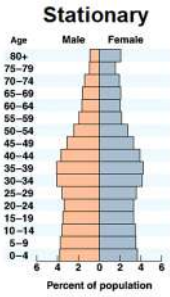
Symmetry

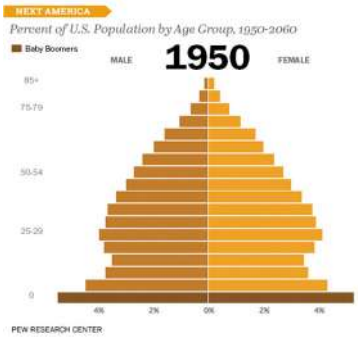
Population Pyramids

- These are graphs that show the distribution of ages across a population divided down the centre between the males and females in the population.
- **Expansive**, **constrictive** and **stationary** pyramids describe populations that are young & growing, elderly & shrinking and not growing respectively[10].









America's morphing age pyramid

[10] https://populationeducation.org/what-are-different-types-population-pyramids/

Figure-Ground

Standing Out from the Crowd

- The figure-ground principle states that people instinctively perceive objects as either being in the **foreground** or the **background**. They either **stand out** prominently in the front (the figure) or **recede** into the back (the ground)

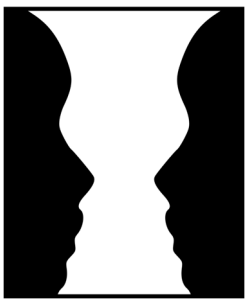


Image from Wikipedia

Image from Criminal Underworld poster series by Simon C. Page

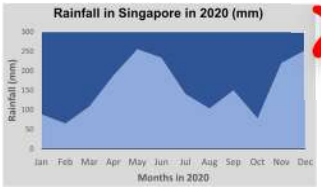
Kiss of the Sea by Octavio Ocampo
https://www.wikiart.org/en/octavio-ocampo/kiss-of-the-sea

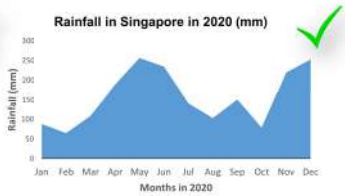
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Figure-Ground

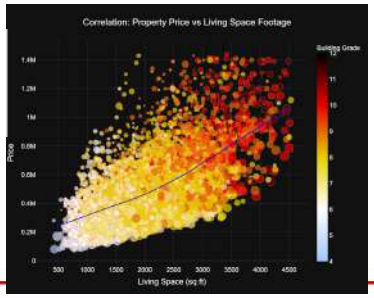
Standing Out from the Crowd

- The figure-ground principle states that people instinctively perceive objects as either being in the **foreground** or the **background**. They either **stand out** prominently in the front (the figure) or **recede** into the back (the ground)
- Information to be communicated must be made to move to the foreground by providing enough **contrast** between the foreground and background to make charts and graphs more legible and distinctive.





Use strong contrast to make foreground chart stand out from background



Good contrast Image from [11]

[11] E.V. Kazakova, The Psychology behind Data Visualization Techniques (2021)- https://towardsdatascience.com/the-psychology-behind-data-visualization-techniques-68ef12865720

Summary

Gestalt Principles in Data Visualisation

- Understanding how the human visual system constantly seeks out visual **patterns**, **groupings** and **meaningful interpretation** can help us design more **accessible** visualisation.
- **Proximity**, **similarity**, **connectedness** and **continuity** help us quickly **associate** related visual elements.
- **Simplicity** and **symmetry** facilitate the exploration and identification of **visual patterns**.
- **Enclosure** and good **figure-ground** separation helps highlight what is important to **focus on**.

References for Visual Encoding

[1] Vidya Kesavan, Gestalt Laws Applied to Data Visualization (2018) – https://medium.com/@vidya83.kesavan/gestalt-laws-applied-to-data-visualization-1e18385b7b00

[2] William Craig, How to Make Data Visualization Better with Gestalt Laws, https://www.webfx.com/blog/web-design/data-visualization-gestalt-laws/

[3] Edward Tufte, The Visual Display of Quantitative Information Graphic Press, 2nd Ed (2001)

[4] Carlos Moura, Step 10 Visualisation - https://composite-indicators.jrc.ec.europa.eu/sites/default/files/COIN%202018%20Step%209%20Visualization.pdf

[5] Visuals for Quiz Questions taken from : https://ig.ft.com/science-of-charts/

[6] B. Wong, Gestalt Principles (Part 1), Nature Methods, Vol 7, pp. 863 (2010)), https://www.nature.com/articles/nmeth1110-863

[7] A. Rutledge, Gestalt Principles of Perception - 3: Proximity, Uniform Connectedness, and Good Continuation (2009) – https://www.andyrutledge.com/gestalt-principles-3.html

[8] ChartAccent, Visualisation Examples - https://chartaccent.github.io/

[9] Colin Ware, Information Visualization, Morgan Kaufmann, 3rd Ed (2012)

[10] https://populationeducation.org/what-are-different-types-population-pyramids/

[11] E.V. Kazakova, The Psychology behind Data Visualization Techniques (2021) – https://towardsdatascience.com/the-psychology-behind-data-visualization-techniques-68ef12865720

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

Chapter 6.3 – Colour

Contents

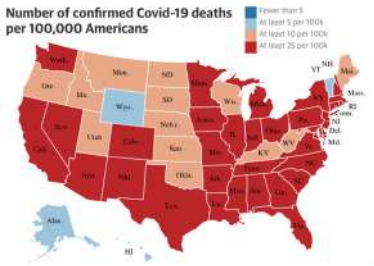
- Colour Perception
- Additive and Subtractive Colour Systems
- Colour Models
- Data Legibility of Colour
- Inclusive Use of Colour
- Colour in Data Visualisation

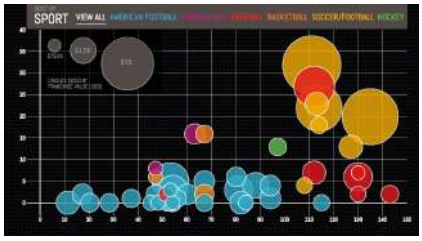
© A/P Goh Wooi Boon (SCSE/NTU)

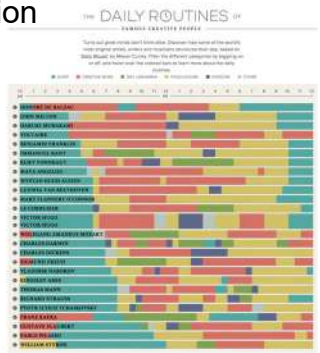
Colour Perception

Living Colours

- The ability to perceive colours has been crucial for the **survival** of early hunters and gathers as it helps them spot lurking dangers, look and identify edible food.
- The ecological role of colour perception in humans suggests that it can also play an important role in the design of effective information visualisation (see the effective use of colour in these visualisation examples)[1].







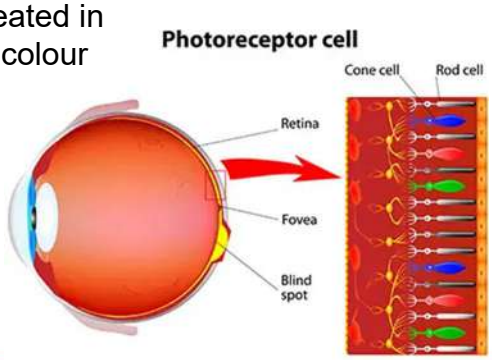
Images from examples in [1]

[1] Ross Crooks, The Power of Data Visualization Plus Examples of Good and Bad Visuals (2021) - https://blog.hubspot.com/marketing/great-data-visualization-examples

Colour Perception

What is Colour?

- The perception of **colour** in humans is due to the stimulation of **photoreceptor** cells (in particular cone cells) by electromagnetic radiation in the visible spectrum, which can be either **reflected** of an object surface or **emitted** from a light source.
- However, the colour we eventually perceive is created in our **brain**. For example, it is able to maintain the colour of an object in a relatively stable manner despite changes in illumination due to shadows[2].
- It is also for this reason the visual attribute of colour is **complex** and its perception can be influenced by numerous factors such as the background, other adjacent colours, etc.



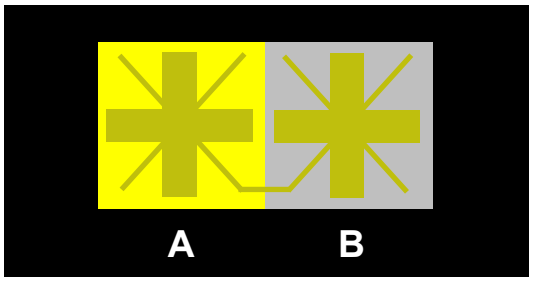
Photoreceptor Cells in the Retina
Image from American Academy of Ophthalmology

[2] Scientific American, Illusory Color & the Brain (2008) - https://www.scientificamerican.com/article/illusory-color-andamp-the-brain-2008-05/

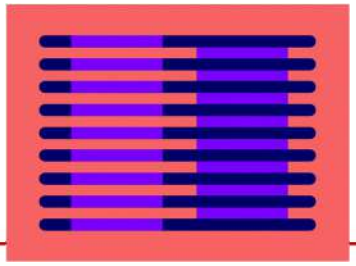
Colour Perception

Colour Appearance

- The **colour contrast** of the **background** can change the way we perceived the actual colour of an object.
- The **Bezold effect** was first reported by Wilhelm von Bezold, a German professor of meteorology.
- Bezold discovered that a colour may appear different depending on its relation to **adjacent colours**. It happens when small areas of colour are interspersed.



Does cross A or B have a darker shade?



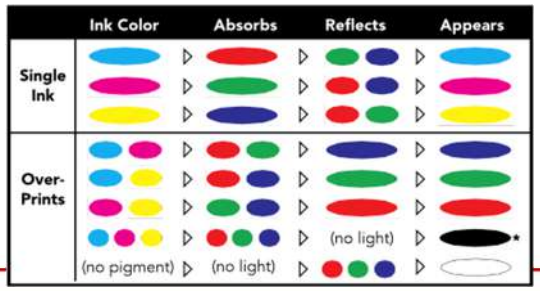
Bezold Effect image from [3]

[3] L. Giovanni and H. Kellen, Is Seeing Really Believing? (2019)- https://dailydolphin.hbcsd.org/6151/science-and-technology/is-seeing-really-believing/

Additive and Subtractive Colour Systems

Display versus Print

- **Additive** and **subtractive** colour systems use a **different** but **small** combination of **primary colours** to produce a whole gamut of colours.
- A display monitor uses the **additive** colour system because they are **emissive devices**. It starts with darkness and add red, green & blue light to create the spectrum of colours[4].
- Printers render colours on paper and must work with **reflected light**. They thus employ the opposing **subtractive primaries** of cyan, magenta & yellow[4].
- In subtractive colour printing, a 4th black colour (K, which stands for key) is added to make four-colour printing (**CMYK**).



[4] Tim Mouw, Additive vs. Subtractive Color Models - https://www.xrite.com/blog/additive-subtractive-color-models

Additive & Subtractive Colour Systems, Images from [4]

Colour Models

HSL and HSV Colour Models

- The RGB colour model is the way displays (and some drawing applications) specify colours, but it is not realistic to think in RGB terms when considering colour choices for visualisation design[5].
- A more accessible colour model for data visualisation design is the **HSL** (Hue, Saturation, Lightness)[5] or the **HSV** (Hue, Saturation, Value) models.
- The 3 dimensions combine to form a **cylindrical-coordinate** colour representation of the RGB colour model.

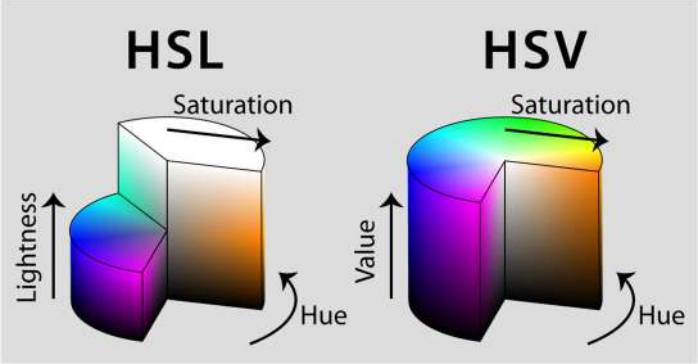


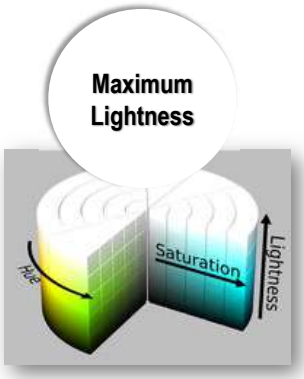
Image taken from Wikipedia

[5] Andy Kirk, Data Visualisation, A Handbook for Data Driven Design, SAGE Publications (2016).

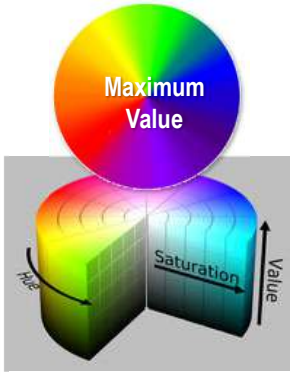
Colour Models

HSL versus HSV

- The **HSL** colour model mimics the way different paints mix together to create colour in the real world, with the **lightness** dimension resembling the varying amounts of black or white paint in the mixture.
- The **HSV** model depict how colours appear as the **value** representing the amount of light falling on it varies.
- The colour with maximum lightness in **HSL** is **pure white**, but a colour with maximum value in **HSV** is **pure hue**, as if the coloured object is illuminated by the brightest white light.



HSL colour model



HSV colour model

Image taken from Wikipedia

Colour Models

HSL Colour Model

- **Hue(H)** is considered true colour. When a **colour** is **described** or labelled, it is most commonly referred to by its hue (e.g. red, blue, yellow, orange)[5].
- **Saturation(S)** is a **scale** that defines the purity or colourfulness of a hue. From intense **pure colour** (high saturation) to a no-colour state of **grey** (low saturation)[5].
- **Lightness(L)** is a scale that defines the contrast of a single hue from **dark** to **light**. It is a scale of light tints (adding white) through to dark shades (adding black)[5].
- **H**,**S** or **L** and its combination can be varied depending on the visualisation design.

Image taken from [6]

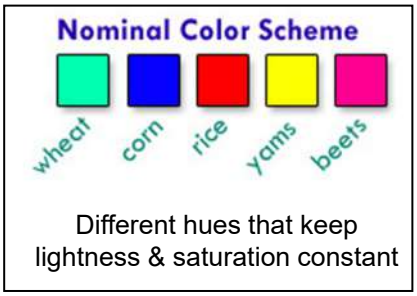
[5] Andy Kirk, Data Visualisation, A Handbook for Data Driven Design, SAGE Publications (2016).
[6] Understanding Hue, Saturation & Lightness (HSL) for Photo Retouching - https://purple11.com/basics/hue-saturation-lightness/

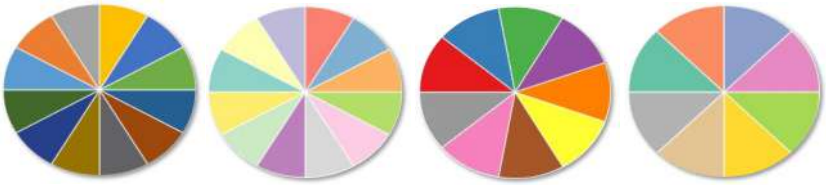
Data Legibility of Colour

Nominal Scale

- Data legibility concerns the use of colour attributes to **encode data** values in charts.
- The colour selection for data with **nominal** scale should be to classify different **categorical** values that are qualitative and have no particular order.
- The goal in the colour scheme is to create visible **distinctions** between each category, while facilitating **efficient** and **accurate** discernment of the categories.

Image taken from [7]





Default Colours | Light Colours | Strong Colours | Colourblind Safe

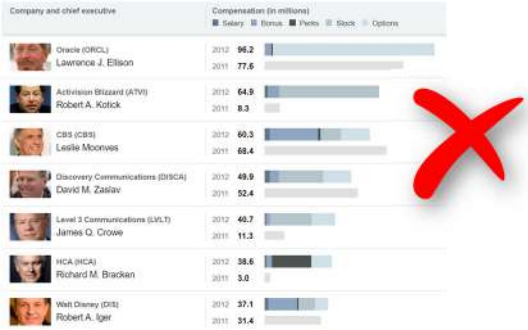
Palette Images taken from [8]

[7] Axis Maps, Using Colors on Maps - https://www.axismaps.com/guide/using-colors-on-maps
[8] Carmen Chan, Choosing Color Palettes in Displayr - https://www.displayr.com/choosing-color-palettes-in-displayr/

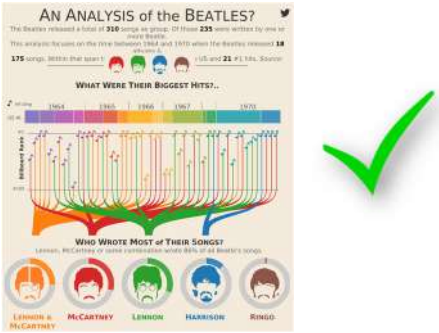
Data Legibility of Colour

Nominal Scale – Creating Contrast

- Creating **contrast** is the main aim of representing nominal data. Varying the **hue** is the most effective means of achieving this goal. Variation in colour tone (**saturation**) may also be considered but varying **lightness** will not be sufficiently discernible[5].



Images taken from [9]



Images taken from [10]

[9] Barry Ritholtz, Executive Pay by the Numbers - https://ritholtz.com/2013/07/executive-pay-by-the-numbers/
[10] Adam E McCann, Beatles Analysis - https://public.tableau.com/app/profile/adam.e.mccann/viz/BeatlesAnalysis/BeatlesAnalysis

Data Legibility of Colour

Nominal Scale – Too Many Colours

- As category count increases, the ability to maintain clear colour differentiation diminishes. The rule of thumb is that no more than 12 categories can be comfortably distinguished based on different colours[5].
- Additional visual attributes like texture or patterns can be added to create further visible distinction but consider its use carefully as the visual may look cluttered and confusing.

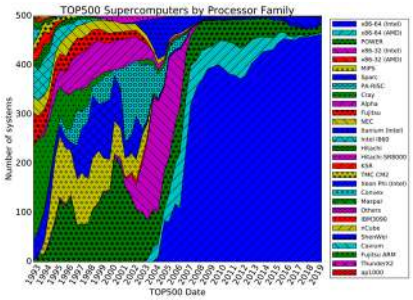
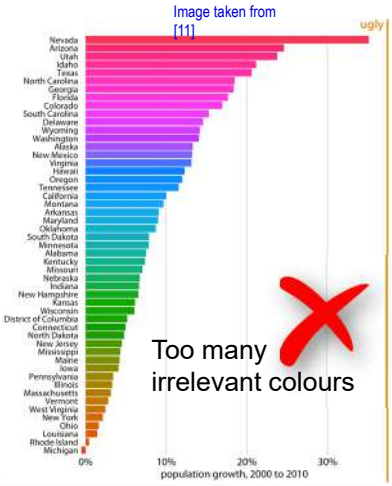


Image taken from https://en.wikipedia.org/wiki/Infographic

Image taken from [11]



Too many irrelevant colours

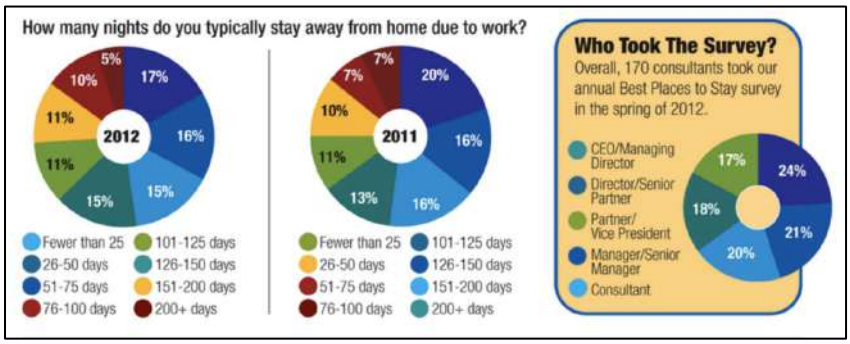
[5] Andy Kirk, Data Visualisation, A Handbook for Data Driven Design, SAGE Publications (2016).
[11] Claus O. Wilke, Fundamentals of Data Visualization, Color Pitfalls - https://clauswilke.com/dataviz/color-pitfalls.html

Think and Apply

Less Colour But More Clarity

- List down what you think are poorly designed aspects of the visualisation that compromised the **accessibility** of the information presented.
- Can you redesign an improved version with no constraints?
- Can you redesign an improved version with only one colour?

(excluding black for text & borders, white for background)



A poorly designed visual that was featured in https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8

Data Legibility of Colour

Ordinal Scale

- Ordinal data has categories that have a natural sense of **ordering** or **hierarchy**.
- The colour scheme aims to create a **visible distinction** between these categories as well as imply a **sense of order** or magnitude through the colour variations.
- The colour dimensions used to achieve this is usually a variation of the **saturation** or **lightness** or both[5].
- The **sequential** colours are for a unidirectional ordinal scale. A **diverging** palette is used for dual direction ordering (e.g. Likert scale survey of positive & negative sentiments).

Sequential
("`Greys`")

Sequential
("`Blues`")

Diverging
("`vlag`")[12]

[5] Andy Kirk, Data Visualisation, A Handbook for Data Driven Design, SAGE Publications (2016).
[12] Seaborn, General principles for using color in plots - https://seaborn.pydata.org/tutorial/color_palettes.html

13

13

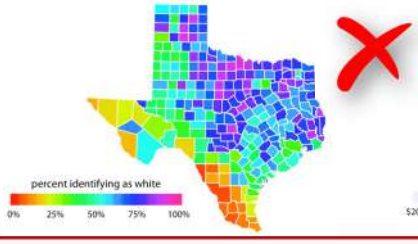
Data Legibility of Colour

Interval and Ratio (Quantitative) Scale

- With **quantitative** scale data, the goal is to use colour variations to distinguish the **relative values** or **magnitude** of the data variable.
- To improve the ease of reading value ranges from their associate colour shade, the data values are divided up into **discrete** classification or 'bins'[5].
- Avoid colour variations using **hue alone** (rainbow scale) as there is little sense of order in hue. Variations in **colour saturation** generally create more intuitive visuals[11].

Continuous Palette

Discrete Palette



[5] Andy Kirk, Data Visualisation, A Handbook for Data Driven Design, SAGE Publications (2016).
[11] Claus O. Wilke, Fundamentals of Data Visualization, Color Pitfalls - https://clauswilke.com/dataviz/color-pitfalls.html

14

14

Inclusive Use of Colour

Colour Blindness

- Colour blindness or colour deficiency affects around 1 in 12 men and 1 in 200 women worldwide[13]. The most common type is the red/green colour blindness (e.g. people with Protanopia or Deuteranopia have problem seeing red or green respectively).
- Data visualisation designers should be **mindful** of how colours are used so that charts created are **inclusive** and meaningful to as large an audience as possible.

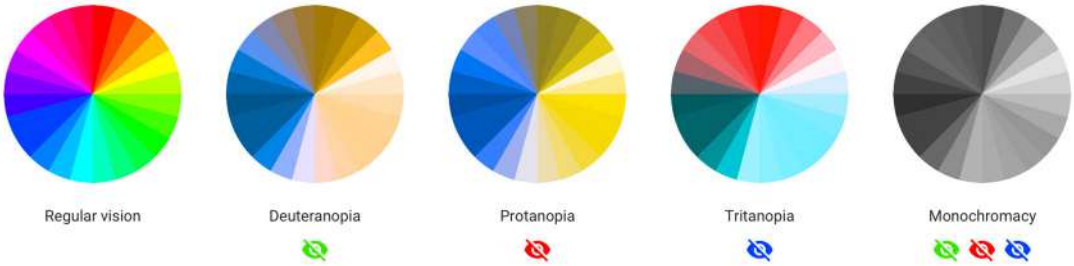


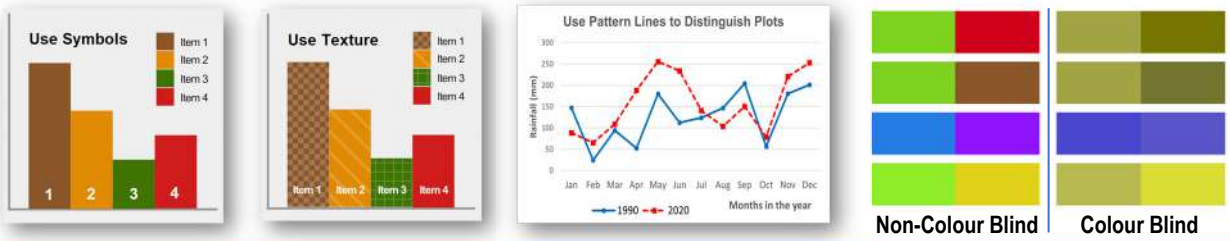
Image taken from [14]

[13] Robyn Collindge, How to Design for Color Blindness (2017) - https://usabilla.com/blog/how-to-design-for-color-blindness/
[14] Ian Tuchkov, Color blindness: how to design an accessible user interface (2018) - https://uxdesign.cc/color-blindness-in-user-interfaces-66c27331b858

Inclusive Use of Colour

Designing for Colour Blindness

- Data interpretation should not be based solely on colour. Both **colour** and **symbols** (e.g. text annotations) can be employed in a redundant manner for better clarity[13].
- Different **textures** or **patterns** (e.g. dotted and dashed lines) can also be added to the coloured areas or lines for different data variables[13].
- Avoid using too **many colours** (limit to 2 to 3) and avoid bad **colour combinations** (e.g. Green & Red, Green & Brown, Blue & Purple, Green & Blue, Blue & Grey, Green & Yellow)[13].



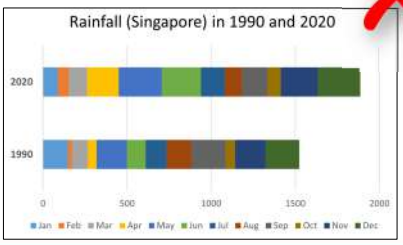
[13] Robyn Collindge, How to Design for Color Blindness (2017) - https://usabilla.com/blog/how-to-design-for-color-blindness/

Colour in Data Visualization

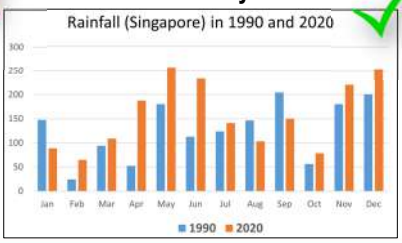
Colour – Friend or Foe?

- Colour is too often seen as a tool to make pretty charts when it should be used primarily to **inform** and **communicate** the essential point or the story being told.
- Too often we asked **how** we should use colour in our visualisation when we should be asking **why** we are using colour[15].
- It is an error to **overused** colours when the same message can be said with fewer.

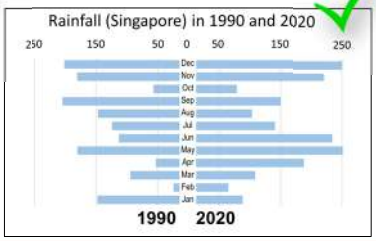
Poor use of colours



Only two colours



Just one colour

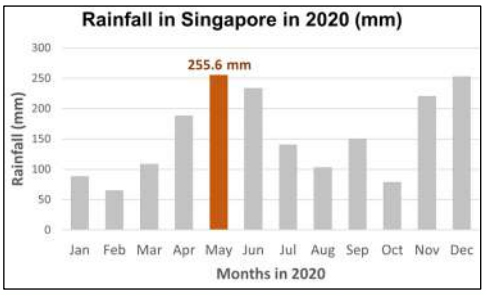


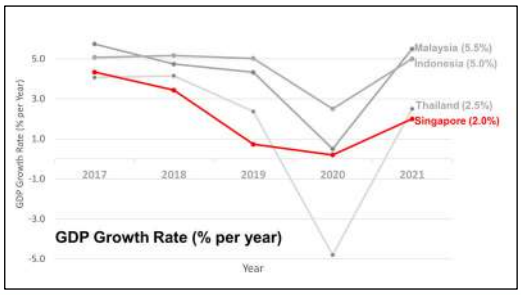
[15] Connor Rothschild, Color in Data Visualization: Less How, More Why – https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8

Colour in Data Visualization

Colour Allows us to Differentiate

- Colour can be used to **draw attention** to a data point of interest.
- An effective way to enhance this effect is to employ **grey** as the primary colour for visualising the data and a **strong colour** to **highlight** the data of interest. The **absence of colour** draws the viewer's eyes to whatever is not grey[15].



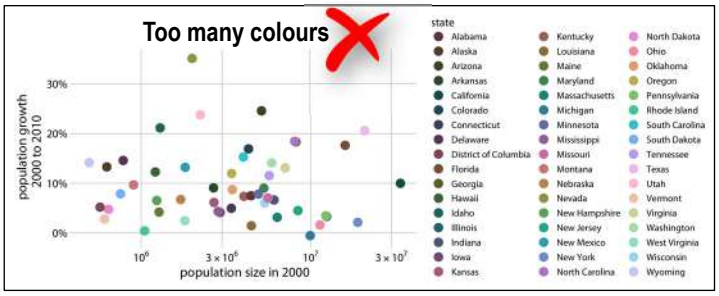


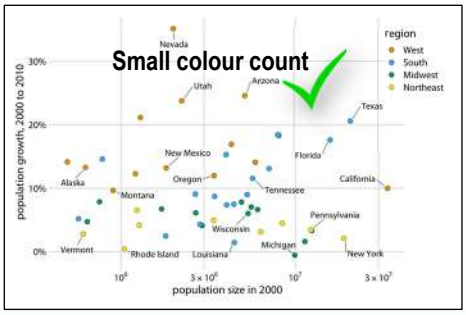
[15] Connor Rothschild, Color in Data Visualization: Less How, More Why – https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8

Colour in Data Visualization

Colour Allows us to Explore

- Colour can be used as a tool to **explore** how a variety of **different categories** of data points that are mapped to **different colours** related to each other.
- Effective exploration requires the **colour count** to be kept **small** so patterns in the data groups can emerge without having to constantly reference the colour legend[15].



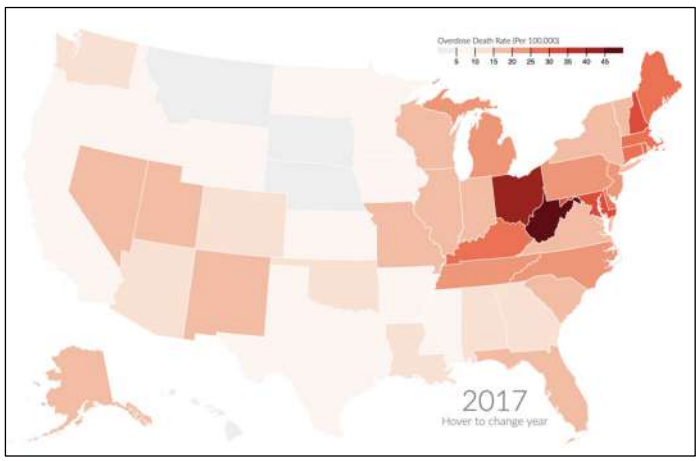


[15] Connor Rothschild, Color in Data Visualization: Less How, More Why – https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8

Colour in Data Visualization

Colour Allows us to Explore

- Another common use of colour for exploratory purposes is to show data progressing across a **sequential colour gradient** (e.g. low to high, bad to good, cold to warm)[15].
- Colour shading in **Choropleth maps** facilitate the exploration of **regional variation** in severity or intensity of a particular variable[15].

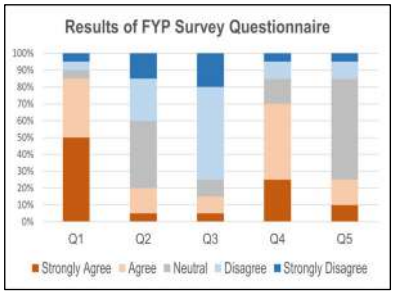


[15] Connor Rothschild, Color in Data Visualization: Less How, More Why – https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8

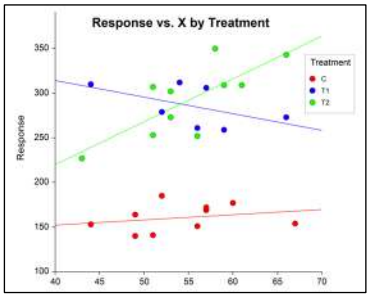
Colour in Data Visualization

Colour Allows us to Compare

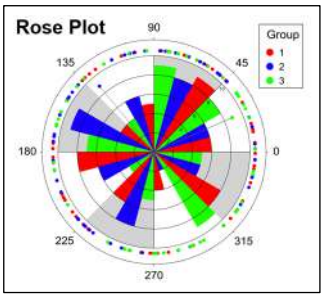
- As colour allows us to easily distinguish between data points, it can be used in aiding **visual comparison** between the characteristics of **different data groups.**
- For effective comparison, the colours should be **distinct** and their numbers limited.



Comparing Likert scale distributions across different questions



Comparing slopes and intersections of data groups [16]



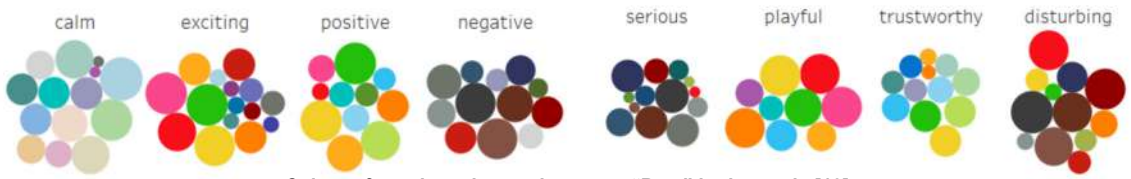
Comparative analysis of circular of angular data [16]

[16] NCSS 2021, Data Analysis & Graphics - https://www.ncss.com/software/ncss/ncss-plots-and-graphs/

Colour in Data Visualization

Colour Allows us to Convey Emotions

- Colours have association with **emotions** and if used effectively, it can make the message embedded in the visualisation more **readily accessible**.
- Studies have shown that if the visualisation goal is to create a **calm** affect**, light, cool, pastel** colours should be use. If the goal is to create a **disturbing** affect, then **dark colours**, especially **reds**, are better[17].
- However, it is good to be mindful that the types of emotions associated with different colours do have **cultural** and **contextual** variations.



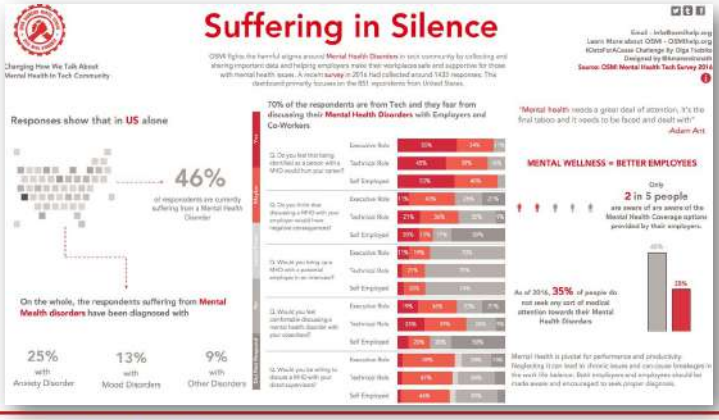
Colours from the palettes chosen as "Best" in the study [18]

[17] L. Bartram et al., Affective Color in Visualization (CHI 2017) - https://research.tableau.com/sites/default/files/Affective%20Color%20CHI%202017.pdf

Colour in Data Visualization

Colour Allows us to Convey Emotions

- Contrasting combination of calm and dull colours like grey with strong disturbing colours like dark red tones can make viewers feel the story that is being told[19].
- In this example, the **grey** colour is used to map out the background information and context. The main message (i.e. prevalence of mental health disorder) is presented in alarming **red**.
- This **affective colour contrast** works as a visual punch and **amplifies** the critical facts[19].



[18] Olga Tsubiks, Color in Data Visualization - The role of color in data storytelling - https://www.olgatsubiks.com/data-for-a-cause-visualizations

Visualisation by Amarendranath D

Think and Apply

Can You Do Better?

- Comment of the chart type and colours used for the purpose of **comparing** teaching performances of Prof A, B and C?
- Can you do better? Redesign it.

Likert Rating	Prof A	Prof B	Prof C
Strongly Agree	55%	5%	69%
Agree	29%	23%	15%
Neutral	8%	40%	5%
Disagree	6%	26%	4%
Strongly Disagree	2%	6%	7%

Table of each teaching score percentages

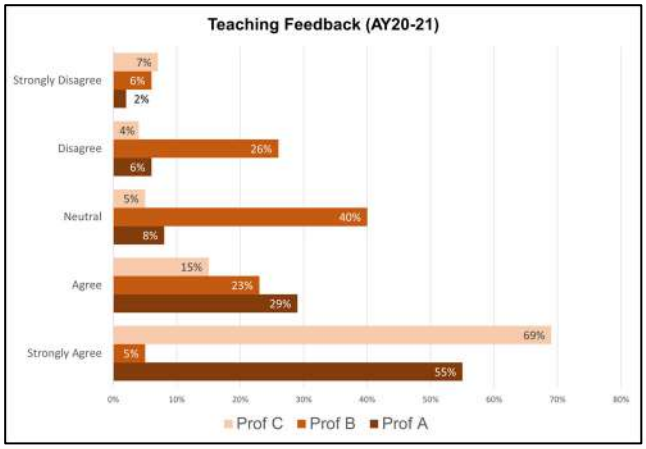


Chart Comparing Teaching Performances

Summary

Colour Perception

- Colour is a very **important visual attribute** in data visualisation and should be given **consideration** at the very start of your design.
- Colour must be used judiciously and **its purpose** in your visualisation design must be **clear** and **effective**.
- Choose an appropriate **colour palette** that fits the **characteristics** and **scale** of your data type; and if relevant, try to **blend** it with the visual **story** and the **emotional** theme you want to communicate.
- Make your visuals as **inclusive** as possible, bearing in mind that not everyone perceives colour in the same manner.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

25

25

References for Colour Perception

[1] Ross Crooks, The Power of Data Visualization Plus Examples of Good and Bad Visuals (2021) - https://blog.hubspot.com/marketing/great-data-visualization-examples
[2] Scientific American, Illusory Color & the Brain (2008) - https://www.scientificamerican.com/article/illusory-color-andamp-the-brain-2008-05/
[3] L. Giovanni and H. Kellen, Is Seeing Really Believing? (2019)- https://dailydolphin.hbcsd.org/6151/science-and-technology/is-seeing-really-believing/
[4] Tim Mouw, Additive vs. Subtractive Color Models - https://www.xrite.com/blog/additive-subtractive-color-models
[5] Andy Kirk, Data Visualisation, A Handbook for Data Driven Design, SAGE Publications (2016).
[6] Understanding Hue, Saturation & Lightness (HSL) for Photo Retouching - https://purple11.com/basics/hue-saturation-lightness/
[7] Axis Maps, Using Colors on Maps - https://www.axismaps.com/guide/using-colors-on-maps
[8] Carmen Chan, Choosing Color Palettes in Displayr - https://www.displayr.com/choosing-color-palettes-in-displayr
[9] Barry Ritholtz, Executive Pay by the Numbers - https://ritholtz.com/2013/07/executive-pay-by-the-numbers/
[10] Adam E McCann, Beatles Analysis - https://public.tableau.com/app/profile/adam.e.mccann/viz/BeatlesAnalysis/BeatlesAnalysis
[11] Claus O. Wilke, Fundamentals of Data Visualization, Color Pitfalls - https://clauswilke.com/dataviz/color-pitfalls.html
[12] Seaborn, General principles for using color in plots - https://seaborn.pydata.org/tutorial/color_palettes.html
[13] Robyn Collindge, How to Design for Color Blindness (2017) - https://usabilla.com/blog/how-to-design-for-color-blindness/
[14] Ian Tuchkov, Color blindness: how to design an accessible user interface (2018) - https://uxdesign.cc/color-blindness-in-user-interfaces-66c27331b858
[15] Connor Rothschild, Color in Data Visualization: Less How, More Why - https://towardsdatascience.com/color-in-data-visualization-less-how-more-why-348514a3c4d8
[16] NCSS 2021, Data Analysis & Graphics - https://www.ncss.com/software/ncss/ncss-plots-and-graphs/
[17] L. Bartram et al., Affective Color in Visualization (CHI 2017) - https://research.tableau.com/sites/default/files/Affective%20Color%20CHI%202017.pdf
[18] Olga Tsubiks, Color in Data Visualization - The role of color in data storytelling - https://www.olgatsubiks.com/data-for-a-cause-visualizations

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

26

26

Data Visualisation

Associate Professor Goh Wooi Boon
School of Computer Science and Engineering
Nanyang Technological University

email: *aswbgoh@ntu.edu.sg*

Chapter 7 – Designing The Visualisation

Contents

- The Visualisation Design Process
- Know The Purpose
- Know Your Audience
- Know The Right Chart Type
- Selecting The Chart
- Psychological Principles of Effective Graphics

© A/P Goh Wooi Boon (SCSE/NTU)

The Visualisation Design Process

Getting Started

- Data visualisation has both **aesthetic** and **scientific** design elements.
- The aesthetic elements are more **subjective** and can be strongly biased by the designer's **artistic preferences**.
- However, the **scientific** aspects of presenting numbers and measurements has a more **logical sequence of steps** that we can follow[1],[2], which are essentially:

Know the Purpose → **Know the Audience** → **Know the Right Chart Type**

[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)
[2] Michael Schermann - A Reader on Data Visualization, Patterns - https://mschermann.github.io/data_viz_reader/patterns.html

Know The Purpose

What Is It For?

- First, find out if the **purpose**[3] of the visualisation is **exploratory** or **explanatory**[1].
- **Exploratory** visualisation is designed help us explore and analyse the data to **derive insights** and **test hypotheses**.
- When we want to share our specific findings to an audience, **explanatory** visualisation is used to **communicate** the **specific story** embedded in the data.
- These two purposes are seldom effectively served by the same visualisation design.

Exploratory chart of Bitcoin prices from CryptoCompare.

Taken from [4]

Explanatory chart of the stock market performance during 2016 US Presidential election night.

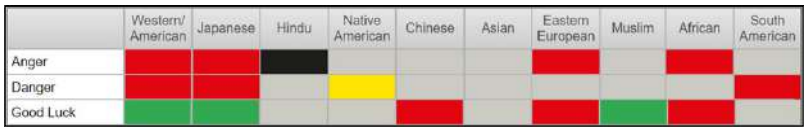
Taken from [4]

[3] R. Kosara, Reflecting on the Design Criteria for Explanatory Visualizations (2016) - https://research.tableau.com/sites/default/files/Kosara-C4PGV-2016.pdf
[4] DataViz Tip #21: Explanatory vs. Exploratory Charts - https://www.amcharts.com/dataviz-tip-21-explanatory-vs-exploratory-charts/

Know Your Audience

Who Is It For?

- It is important to know the audience the visualisation is targeted for as factors like their **cultural** background, **expertise** level, etc can influence the design of an effective and meaningful visualisation[5].
- For example, for a Chinese audience, the colour red represents a positive event (e.g. increasing profits), while western audiences tend to view green as representing a positive trends.
- **Discipline variations** too can be significant. Visualisation charts the financial community is familiar with are usually different from that of scientific communities, and these differences can even vary between the various scientific & engineering disciplines.



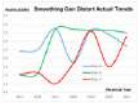
[6] Image extracted from R. Allison, Colors represent different things, in different cultures (2017) - https://blogs.sas.com/content/sastraining/2017/06/29/colors-in-different-cultures/

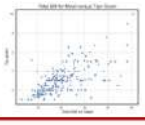
[5] Mustapha Mekhatria, 5 Steps to Create an Effective Data Visualization (2017) - https://www.highcharts.com/blog/post/5-steps-to-create-an-effective-data-visualization/

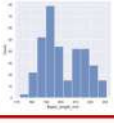
Know The Right Chart Type

Which Should I Use?

- While it is possible to visualise the data using several different chart types, it is important to choose the type that can best communicate the story **clearly** and **accurate** by separating the background noise from the **intended message**[7].
- There are four basic chart types:
  - **Comparison chart** - e.g. a line chart displaying the profits of 3 departments over 7 years.
  - **Composition chart** - e.g. a pie chart displaying the market share of telcos in Singapore.
  - **Relationship chart** - e.g. a scatter plot of the relationship between tips given and cost of meal.
  - **Distribution chart** - e.g. a column histogram displaying the distribution of exam grades.





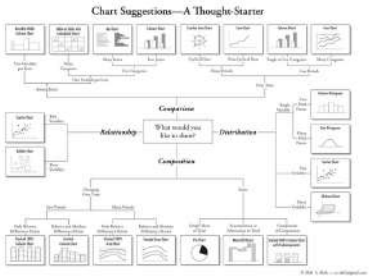


[7] Jānis Gulbis, Data Visualization – How to Pick the Right Chart Type? (2016) - https://eazybi.com/blog/data_visualization_and_chart_types/

Selecting The Chart

Which Should I Use?

- A suitable chart can be determined by answering several questions[7].
  - How many variables do you want to show in a single chart? (1, 2, 3 or many?)
  - How many items (data points) will you display for each variable? (only a few or many?)
  - Will you display values over a period of time, or among items or groups?
- The chart selection diagram by Andrew Abela can then be used to pick the right chart for the data type being visualise.

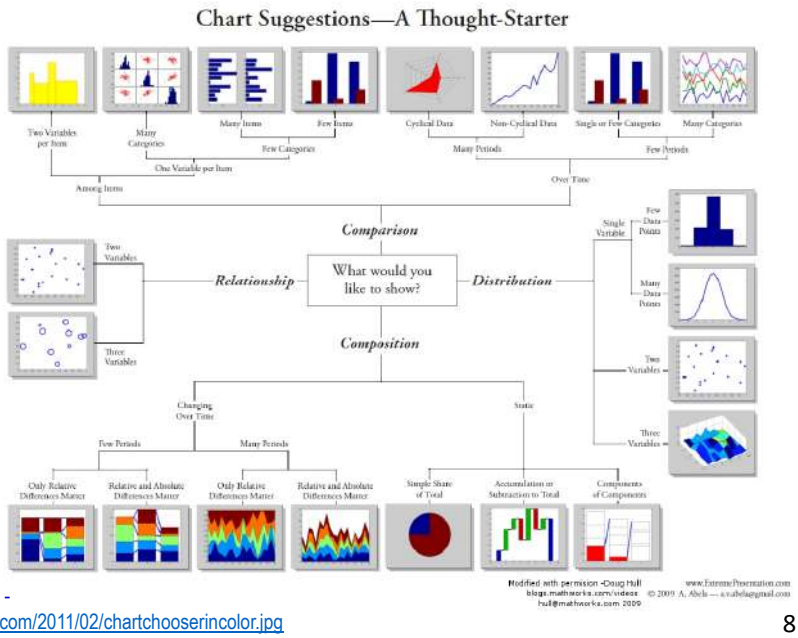


[7] Jānis Gulbis, Data Visualization – How to Pick the Right Chart Type? (2016) - https://eazybi.com/blog/data_visualization_and_chart_types/

Selecting The Chart

Which Should I Use?

- Andrew Abela's Chart Chooser in Colour[8].

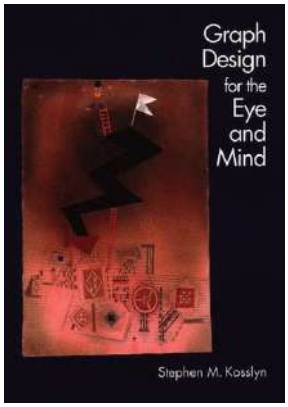


[8] Andrew Abela's Chart Chooser - https://apandre.files.wordpress.com/2011/02/chartchooserincolor.jpg

Psychological Principles of Effective Graphics

The Eight-Fold Way

- Stephen Kosslyn proposed eight very useful principles to guide the design of **effective data visualisation**[1].
- This principles are rooted in his findings about how humans **perceive** and **comprehend visual information**.
- The eight principles are organised into three groups, with each group addressing different goals of:
  - **Connect** with your audience.
  - **Direct** the audience attention through the visual.
  - **Promote** understanding and memory retention.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

9

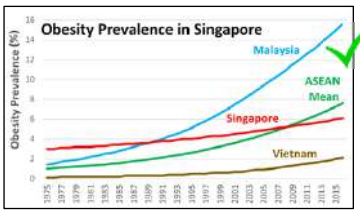
9

1. Connect with audience

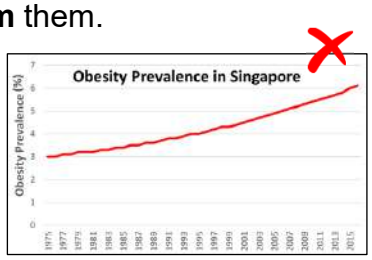
Principle of Relevance

The Goldilocks Principle

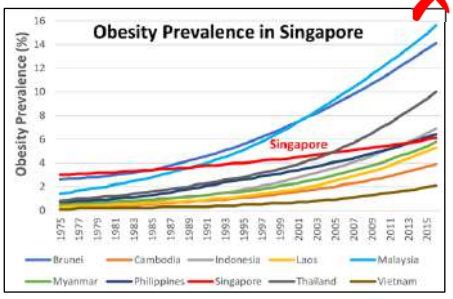
- **Communication is most effective when neither too much nor too little information is presented**[1].
- Decide on **exactly what message** you want to convey. Then use only the **relevant data** that will help you communicate this message.
- Presenting too little may **puzzle** the audience, while presenting too much will distract and **overwhelm** them.



Just right – comparison gives meaning



Too little info – no context to give meaning



Too much info – cluttered & confusing!

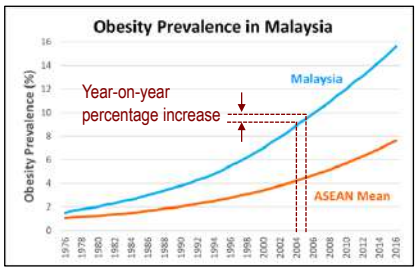
[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

10

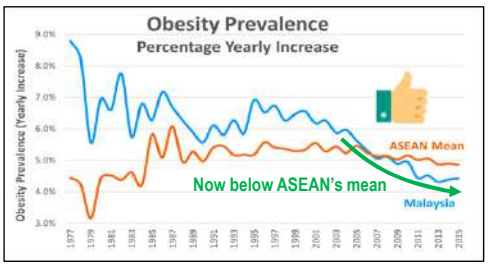
Ethical Visualisation

Don't Hide Behind Percentages

- The Director of Weight Loss is concern about the effectiveness of his obesity prevention programmes and has to give an update to the cabinet next week. He decided to use year-on-year % increases in his chart...



Malaysia's alarming increase in obesity prevalence



Percentage increase from one year to the next
(smoothing on moving average data, 3 samples)

Has ethical visualisation been violated here?

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Note: Obesity dataset available at: https://ourworldindata.org/grapher/share-of-adults-defined-as-obese

11

11

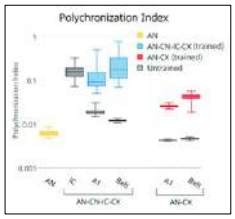
2. Connect with audience

Principle of Appropriate Knowledge

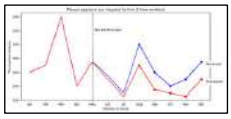
Know Your Audience

- **Communication requires prior knowledge of relevant concepts, jargon and symbols**[1].
- **Know** you audience and their familiarity with the content and charts being shared. Good visuals present new information but it should not be so new that it is unrelated to the knowledge the audience possess.

Candlestick chart
(used in financial visualisation)

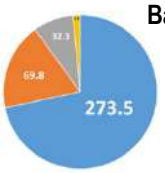


Box & whisker chart
(scientific visualisation)



Line chart

Bar chart



Pie chart

The common & familiar type of charts

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

12

12

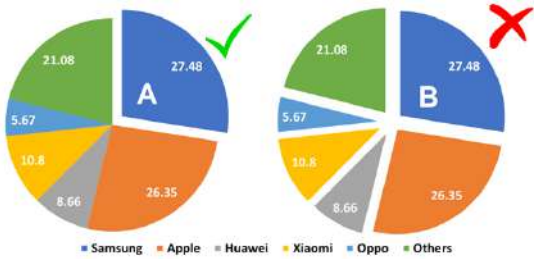
3. Direct and hold attention

Principle of Salience

Attention to the Important

- **Attention is drawn to large perceptible differences**[1].
- The most visually prominent aspects of a visual will **draw attention** to them, for this reason they should embed the **most important** information being conveyed.
- All visual properties are relative, so what counts as "visually striking" depends on the properties of the display as a whole.

Vary the saliency of the line to indicate its importance



The contrast of the intact segment pieces makes the exploding pie segment (A) more salient[1]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

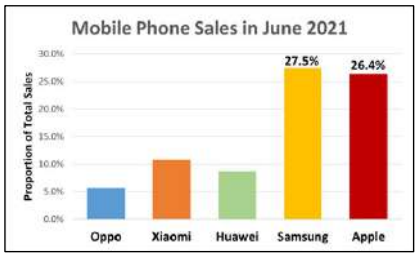
13

13

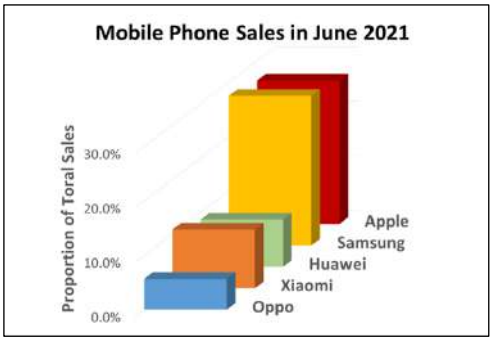
Ethical Visualisation

Don't Hide Behind Another 3D Bar

- Apple mobile sales has fallen behind slightly in June 2021. An Apple sales executive was wondering how to present these figures to her boss and was thinking of using a colourful 3D bar chart…



Proportion of worldwide mobile phone sales in June 2021



The 3D occluding bar chart

Has ethical visualisation been violated here?

Note: We see partially covered objects as further away. Our **size constancy** mechanism makes the rearmost object larger than they actually are[1].

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Note: Mobile sales dataset available at: https://gs.statcounter.com/vendor-market-share/mobile

14

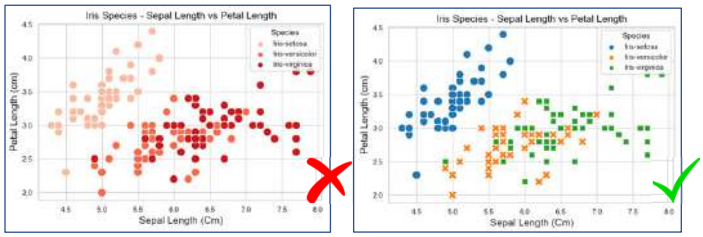
14

4. Direct and hold attention

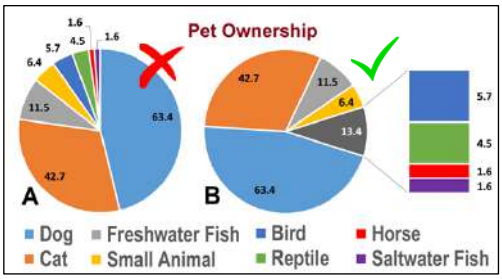
Principle of Discriminability

Maintain Good Contrast

- **Two properties must differ by a large enough proportion or they will not be distinguished**[1].
- This principle relates to Weber's law on just noticeable difference. Be mindful your design and colour choices allow viewers to discriminate between variables and between the variable and its background.



Combine strong distinctive colours & shapes to improve discriminability



Small pie segments in A cannot be discriminated. Design B allows them to be visualise.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

15

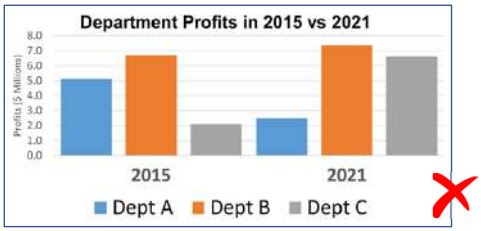
15

5. Direct and hold attention

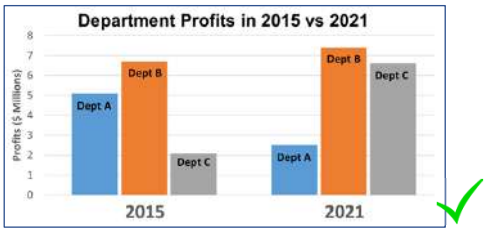
Principle of Perceptual Organisation

Making Visual Sense

- **People automatically group elements into units, which they then attend to and remember**[1].
- We should exploit the various Gestalt laws that govern how we group visual object at a sub-conscious level to improve the effectiveness of our visual communication.



This visual does a poor job in showing the comparative change in department profits



The Dept bar proximity and direct labelling create distinct groups & association to aid visual comparison

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

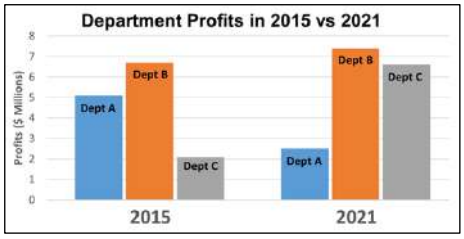
16

16

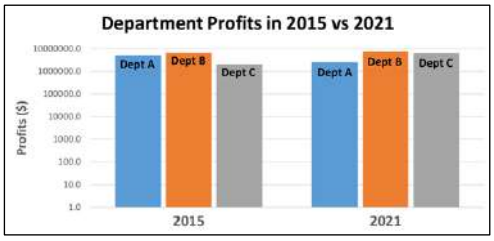
Ethical Visualisation

Scales That Mislead

- The profits of **Dept A** has been steadily declining over the years and is currently the lowest. There is a critical profit review with the board to decide its fate going forward. The head of Dept A is working on the presentation charts….



Profits of Dept A is currently the lowest of the three departments



The $\log_{10}$ scale applied to the $M profits hides the big profit differentials

Has ethical visualisation been violated here?

Note: The logarithmic scale has a tendency to compress the distances between large numbers making even large differences insignificant[1].

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

17

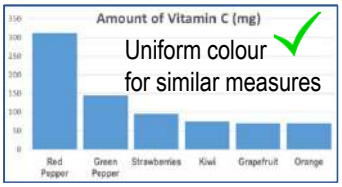
17

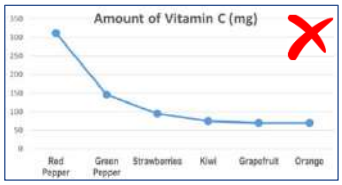
6. Promote understanding and memory

Principle of Compatibility

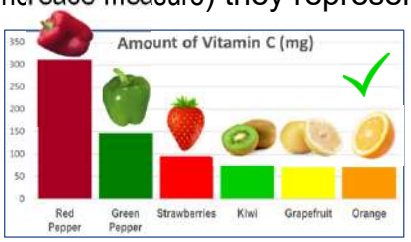
Saying The Same Thing

- **A message is easier to understand if its form is compatible with its meaning**[1].
- Compatibility dictates that we choose the right type of chart for the intended purposes (e.g. trends and continuous changes are best depicted with line charts, whilst discrete entities are better served with bar charts).
- Visual attributes (e.g. colours, height, etc) should match the meaning (e.g. red for "stop" & green for "go") and the variations (increase height maps to increase measure) they represent.





Discrete bars are more compatible to nominal data then continuous lines



If multi-coloured bars used, matching bar colour to fruit colour improves compatibility

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

18

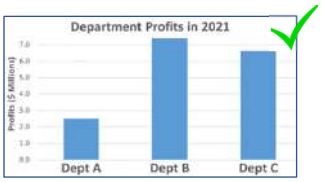
18

7. Promote understanding and memory

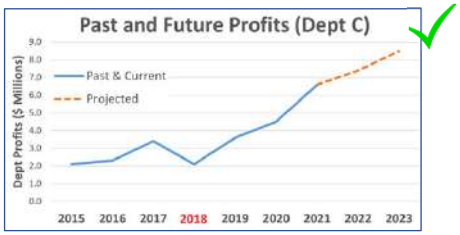
Principle of Information Changes

Meaningful Changes

- **People expect changes in properties to carry information**[1].
- Changes in the visual (e.g. changing colour, texture, line height & width, etc) will be interpreted by viewer as conveying information .
- Similarly, viewers expect any information that should be conveyed in the visual will be indicated by some visible change (e.g. transition from current to projected profits in line chart should be indicated by a different colour or pattern).



If entities don't vary continuously, use bar instead of line



Projected profits & recession year clearly indicated

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

19

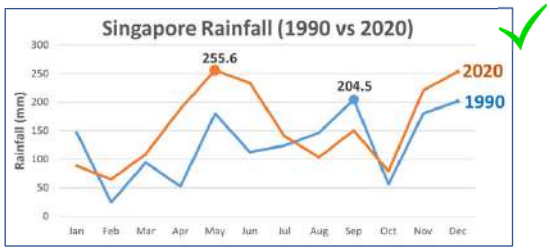
19

8. Promote understanding and memory

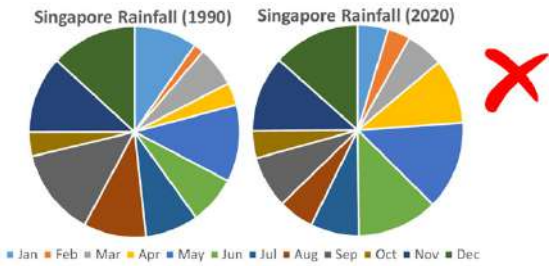
Principle of Capacity Limitation

Too Much Means Only A Little

- **People have a limited capacity to retain and process information. A message will not be understood if too much information must be retained or processed**[1].
- We can only maintain a limited amount of information in mind at any one time. Visuals should not require the viewer to hold more than about **four perceptual groups** in mind at once.



Visualising peak months with minimal cognitive load



Too many colours to compare!

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[1] Stephen M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

20

20

Summary

Designing The Visualisation

- Effective visualisation design requires you to know the **purpose** and the **audience** you are designing for before you begin selecting what **suitable chart** to use.
- During the data visualisation design process, you need to be mindful of how humans **perceive** and **comprehend visual information**.
- Employ the eight psychological principles proposed by Kosslyn to help **connect** with the audience, **direct** their attention to your main points and facilitate their **understanding** and **retention** of your intended message.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

21

21

References for Designing the Visualisation

[1] S.M. Kosslyn, Graph Design for the Eye and Mind, Oxford University Press (2006)

[2] Michael Schermann - A Reader on Data Visualization, Patterns - https://mschermann.github.io/data_viz_reader/patterns.html

[3] R. Kosara, Reflecting on the Design Criteria for Explanatory Visualizations (2016) - https://research.tableau.com/sites/default/files/Kosara-C4PGV-2016.pdf

[4] DataViz Tip #21: Explanatory vs. Exploratory Charts - https://www.amcharts.com/dataviz-tip-21-explanatory-vs-exploratory-charts/

[5] Mustapha Mekhatria, 5 Steps to Create an Effective Data Visualization (2017) - https://www.highcharts.com/blog/post/5-steps-to-create-an-effective-data-visualization/

[6] Image extracted from R. Allison, Colors represent different things, in different cultures (2017) - https://blogs.sas.com/content/sastraining/2017/06/29/colors-in-different-cultures/

[7] Jānis Gulbis, Data Visualization – How to Pick the Right Chart Type? (2016) - https://eazybi.com/blog/data_visualization_and_chart_types/

[8] Andrew Abela's Chart Chooser - https://apandre.files.wordpress.com/2011/02/chartchooserincolor.jpg

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

22

22

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 8.1
Design Graph to Enlighten: Principles and Best Practices

2

Outline

- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization

Prof. Edward Tufte from Yale University

- Other Best Practices

3

3

Outline

- Visualization Principles by Prof. Edward Tufte
  - **Graphical Integrity**
  - Data-ink Ratio Maximization

Prof. Edward Tufte from Yale University

- Other Best Practices

4

4

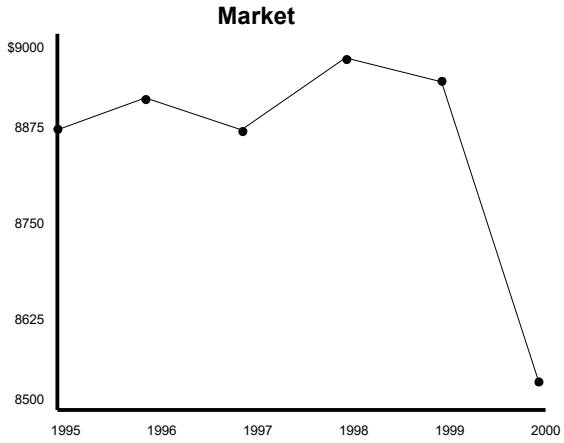
Examples of Misleading Visualizations

- Big discount or benefits?

5

Examples of Misleading Visualizations

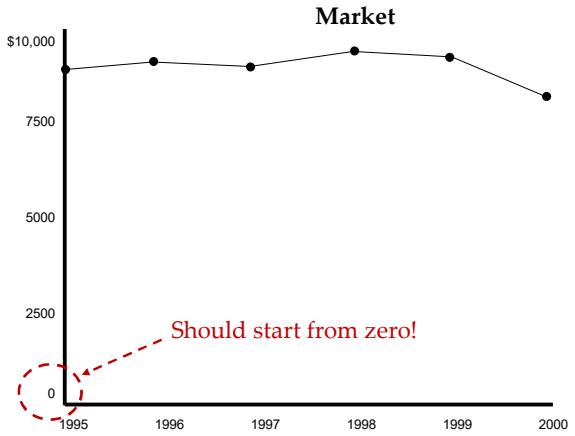
- Is it a market crash?



6

Graphical Integrity

- **Principle: Graphics must not quote data out of context**
  - Showing the entire scale



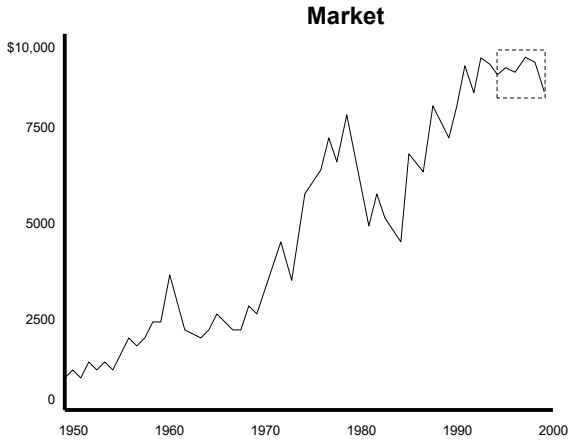
If data visualization does not start at zero, the viewer can be tricked into seeing false information.

7

7

Graphical Integrity

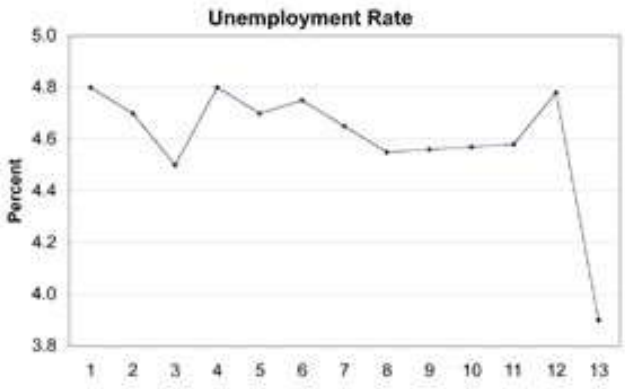
- **Principle: Graphics must not quote data out of context**
  - Showing the entire scale



8

8

Examples of Misleading Visualizations

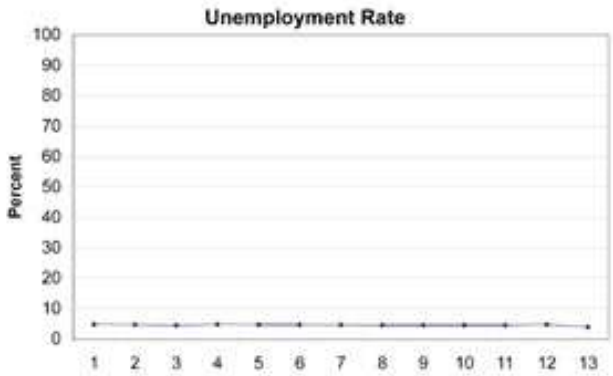


Unemployment Rate: A misleading visualization

9

9

Examples of Misleading Visualizations



Unemployment Rate: A visualization telling the truth

10

10

Graphical Integrity

- **Principle: Visual attribute value should be directly proportional to data attribute value**

  - Visual attributes: the visual encoding channels used to encode the underlying data, e.g., bar height, color density, circle radius, etc.

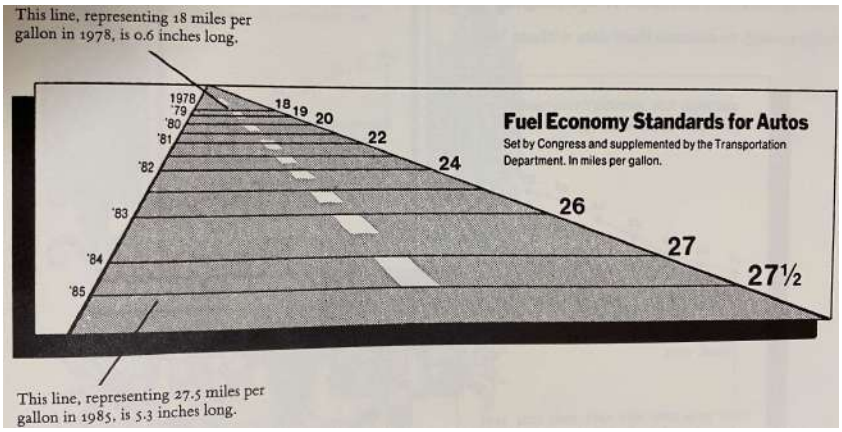
- Lie factor = $\frac{\textit{size of effect shown in graphic}}{\textit{sizes of effect in data}}$
- Truth = 1.0

11

11

Fuel Economy Standards for Autos

- Lie factor = ?



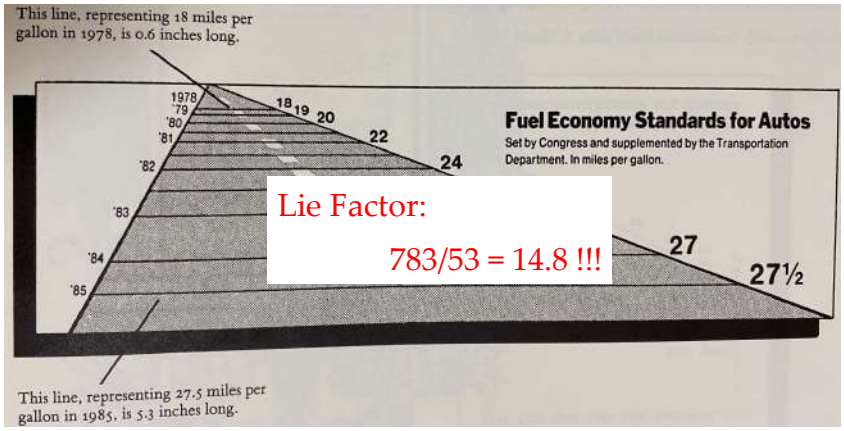
Tufte, Edward R. *The visual display of quantitative information*. Vol. 2. No. 9. Cheshire, CT: Graphics press, 1983.

12

12

Fuel Economy Standards for Autos

- Size of effect in data:
  (27.5 – 18.0)/18.0 = 53%
- Size of effect shown in graphic:
  (5.3 – 0.6)/0.6 = 783%



Tufte, Edward R. *The visual display of quantitative information*. Vol. 2. No. 9. Cheshire, CT: Graphics press, 1983.

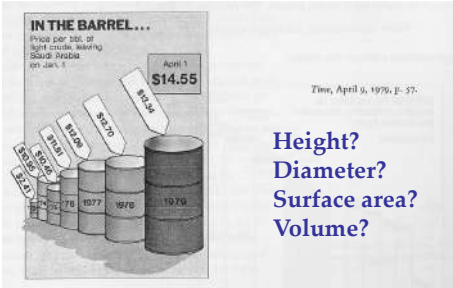
13

13

Graphical Integrity

- **Principle: The number of information-carrying dimensions (i.e., visual encoding channels) should not exceed the number of dimensions in the data.**

For example, avoid using two (or three) varying dimensions to show one-dimensional data



The visualization uses the height and width simultaneously of the bottle to represent the crude oil price, which is problematic:

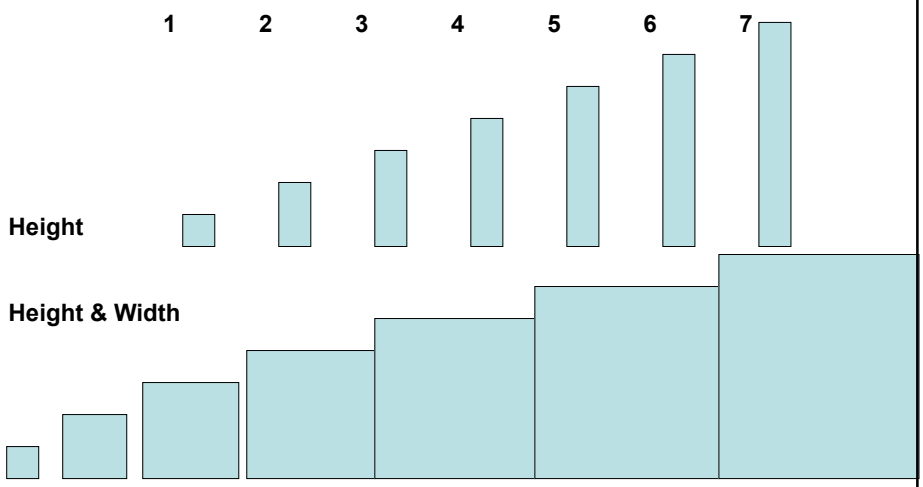
- Ambiguity in perception: Without an explicit indication, it is difficult for viewers to know which visual channel is used to encode the data.
- Space inefficiency: it results in the usage of a larger space.

Tufte, Edward R. *The visual display of quantitative information*. Vol. 2. No. 9. Cheshire, CT: Graphics press, 1983.

14

14

Height & Width Encoding



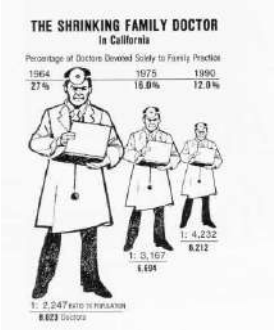
15

15

Graphical Integrity

- **Principle: The number of information-carrying dimensions (i.e., visual encoding channels) should not exceed the number of dimensions in the data.**

For example, avoid using two (or three) varying dimensions to show one-dimensional data



Questions:

- By reading the visualization itself, are you able to accurately identify which visual channel is used to encode the underlying data?
- How should we fix the issue?

Use only one visual encoding channel (e.g., only width or height)!

Tufte, Edward R. *The visual display of quantitative information*. Vol. 2. No. 9. Cheshire, CT: Graphics press, 1983.

16

16

Rankings of Visual Encoding Effectiveness

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat Color Hue Texture	Length	Shape
Connection	Angle	Length
Containment	Slope	Angle
Shape	Area (Size)	Slope
	Volume	Area
	Shape	Volume

Mackinlay, Jock. "Automating the design of graphical presentations of relational information." *Acm Transactions On Graphics (Tog)* 5.2 (1986): 110-141.

17

17

Outline

- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization

Prof. Edward Tufte from Yale University

- Other Best Practices

18

18

Data-ink Ratio Maximization

- Data-ink is the non-erasable ink used for the presentation of data

- Reduce the non-data-ink (a.k.a., chartjunk)
  - Remove unnecessary non-data-ink
  - Emphasize data-ink

- Enhance the data-ink

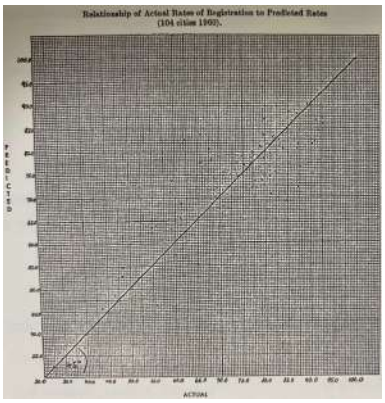
$$\frac{\text{Data Ink}}{\text{Total Ink}} = \text{Data Ink Ratio}$$

19

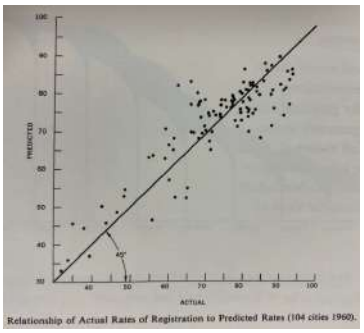
19

Data-ink Ratio Maximization

- For the following two visualizations that shows exactly the same data, which one is more clear for you to understand the underlying data?



Low data-ink ratio



High data-ink ratio

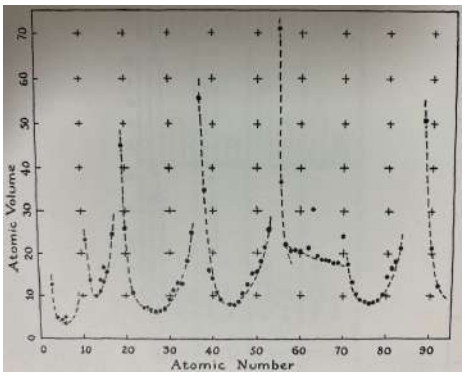
Tufte, Edward R. *The visual display of quantitative information*. Vol. 2. No. 9. Cheshire, CT: Graphics press, 1983.

20

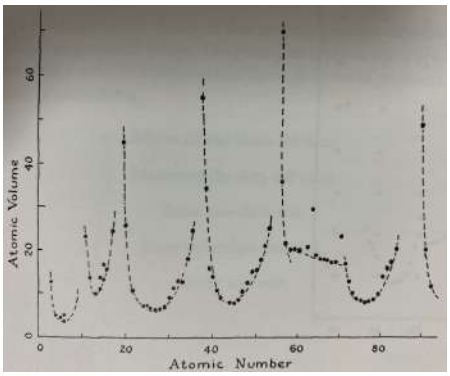
20

Data-ink Ratio Maximization

- For the following two visualizations that shows exactly the same data, which one is more clear for you to understand the underlying data?



Low data-ink ratio



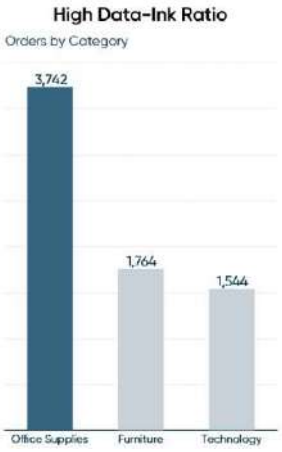
High data-ink ratio

Tufte, Edward R. *The visual display of quantitative information*. Vol. 2. No. 9. Cheshire, CT: Graphics press, 1983.

21

21

Data-ink Ratio Maximization



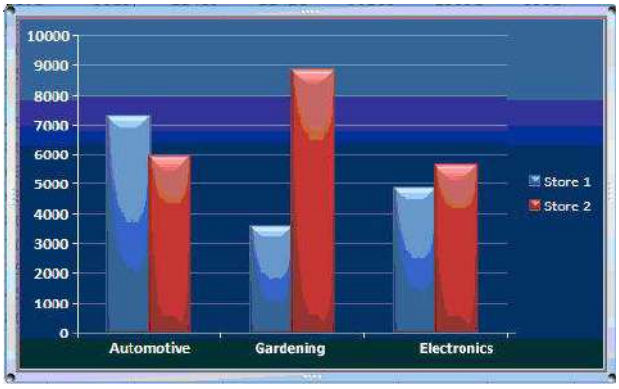
It is not necessary to put the icons on the bar chart, as it does not bring more information.

22

22

Data-ink Ratio Maximization: Avoid Chartjunk

- Avoid using unnecessary color shading for the bar



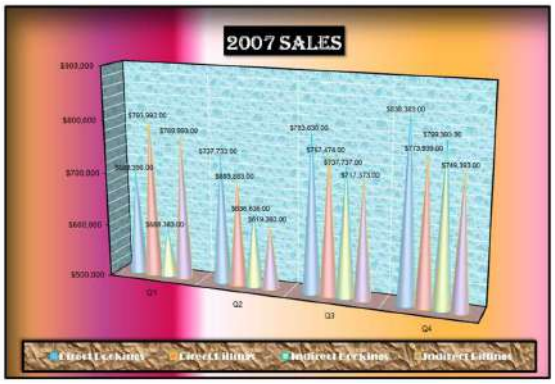
- The color shading here is useless and can actually be misleading for viewers!
- The background color is also meaningless.

23

23

Data-ink Ratio Maximization: Avoid Chartjunk

- Avoid colorful or wallpaper background



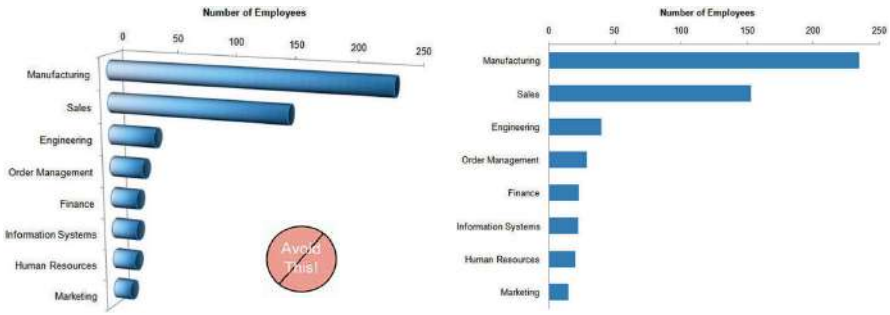
- The colorful background is also meaningless.

24

24

Data-ink Ratio Maximization: Avoid Chartjunk

- Avoid using 3D effects in graphics



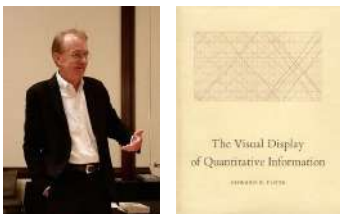
- The 3D effect makes it harder for viewers to perceive the underlying data accurately.

25

25

Outline

- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization



Prof. Edward Tufte from Yale University

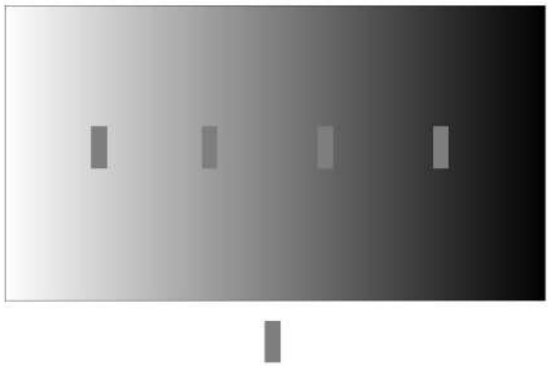
- Other Best Practices

26

26

Practical Guides for Using Color in Charts

- Rule 1: If you want different objects of the same color in a graph to look the same, make sure that the background- the color that surrounds them – is consistent.

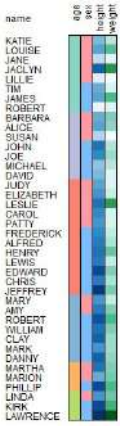


27

27

Practical Guides for Using Color in Charts

- Rule 2: If you want objects in a graph to be easily seen, use a background colour that constracts sufficiently with the object



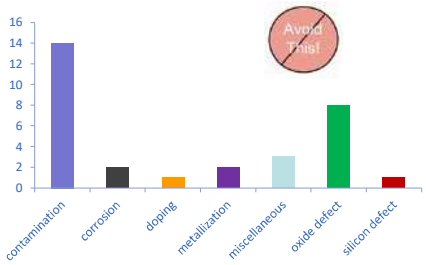
	name	age	sex	height	weight	mean weight by age
1	KATIE	12	F	59	96	99.000
2	LOUISE	12	F	61	123	99.000
3	JANE	12	F	55	74	99.000
4	JACLYN	12	F	66	145	99.000
5	LILLIE	12	F	52	64	99.000
6	TIM	12	M	60	84	99.000
7	JAMES	12	M	61	128	99.000
8	ROBERT	12	M	51	79	99.000
9	BARBARA	13	F	60	112	94.714
10	ALICE	13	F	61	107	94.714
11	SUSAN	13	F	56	67	94.714
12	JOHN	13	M	66	98	94.714
13	JOE	13	M	63	106	94.714
14	MICHAEL	13	M	58	95	94.714
15	DAVID	13	M	59	79	94.714
16	JUDY	14	F	61	81	100.833
17	ELIZABETH	14	F	62	91	100.833
18	LESLIE	14	F	65	142	100.833
19	CAROL	14	F	63	84	100.833
20	PATTY	14	F	62	85	100.833
21	FREDERICK	14	M	63	93	100.833
22	ALFRED	14	M	64	99	100.833
23	HENRY	14	M	66	119	100.833
24	LEWIS	14	M	64	92	100.833
25	EDWARD	14	M	66	112	100.833
26	CHRIS	14	M	64	99	100.833
27	JEFFREY	14	M	66	113	100.833
28	MARY	15	F	62	92	106.286
29	AMY	15	F	64	112	106.286
30	ROBERT	15	M	67	128	106.286
31	WILLIAM	15	M	66	111	106.286

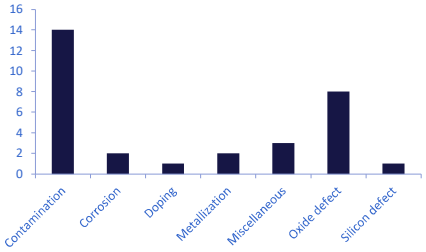
28

28

Practical Guides for Using Color in Charts

- Rule 3: Use different colors only when they correspond to different meanings in the data





It is not necessary to use different colors for different bars.

29

29

Practical Guides for Using Color in Charts

- Rule 4: Use soft, natural colours to display most information and bright and/or dark colours to highlight information that requires greater attention

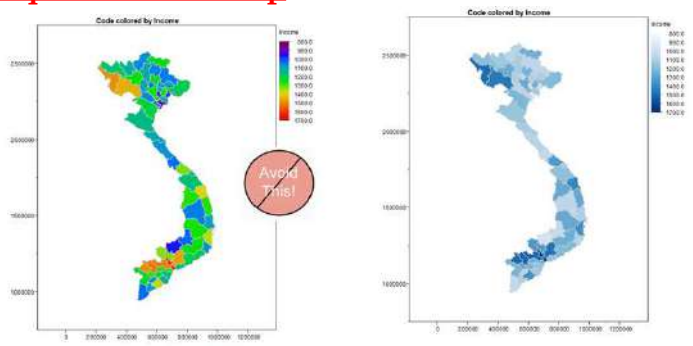
The grids and labels are in soft gray, while the line charts are in bright yellow and blue.

30

30

Practical Guides for Using Color in Charts

- Rule 5: When using colour to encode a sequential range of quantitative values, stick with a single hue and vary intensity from pale colours for low values to increasingly darker and brighter colours for high values, i.e., use **sequential colormap**

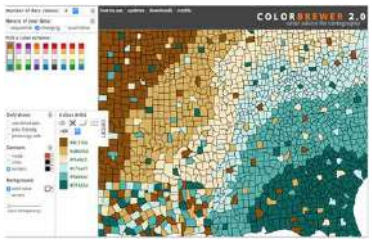


31

31

Colormap Design

- Colormap: a mapping between data values and colors
- Two types of colormaps
  - Categorical (aka., qualitative)
  - Ordered (aka., quantitative): sequential + diverging
- A very useful tool: Colorbrewer2 http://colorbrewer2.org/

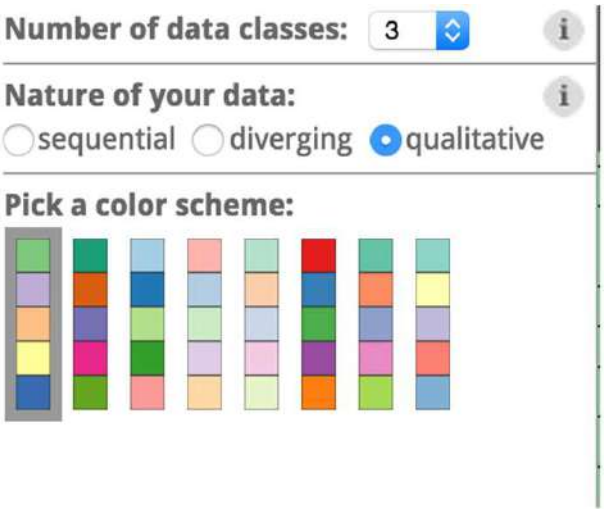


32

32

Categorical Colormaps

- Use color to encode categories and groupings



33

33

Rules for Categorical Colormaps

- The maximum number of colors which can be used is usually between 6 to 12
- The 12 recommended colors

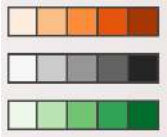
1. Red 2. Green 3. Yellow
4. Blue 5. Black 6. White
7. Pink 8. Cyan 9. Gray
10 Orange 11. Brown 12. Purple

34

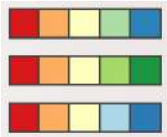
34

Ordered Colormaps

- **Sequential Colormap**
  - Ramp in luminance, possibly also hue.
  - Typically higher values map to darker colors.



- **Diverging Colormap**
  - Useful when data has a meaningful "midpoint".
  - Use neutral color (e.g., gray) for midpoint.
  - Use two saturated colors for endpoints.



- **Limit number of steps in color to 3-9**

  Why?

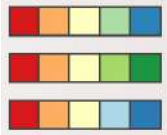
35

35

Ordered Colormaps

- **Sequential Colormap**
  - Ramp in luminance, possibly also hue.
  - Typically higher values map to darker colors.

- **Diverging Colormap**
  - Useful when data has a meaningful "midpoint".
  - Use neutral color (e.g., gray) for midpoint.
  - Use two saturated colors for endpoints.



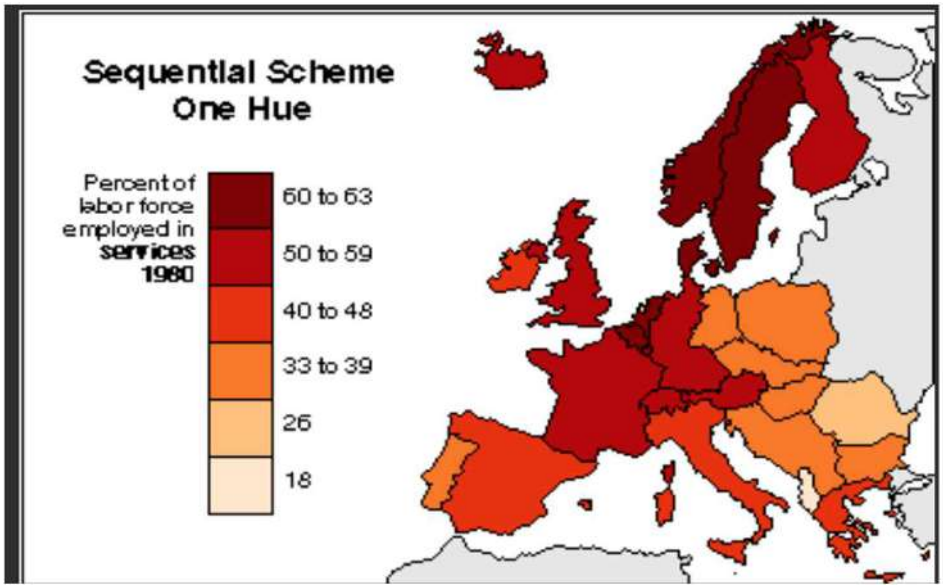
- **Limit number of steps in color to 3-9**

  Avoid simultaneous contrast; Need to hold mappings in memory

36

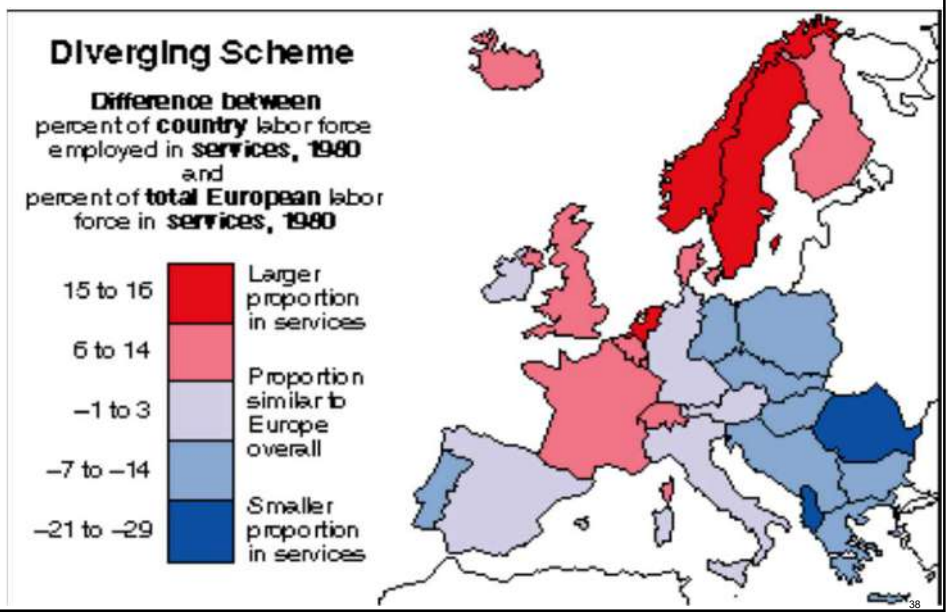
36

Sequential Colormap



37

Diverging Colormap



38

Practical Guides for Using Color in Charts

- Rule 6: Respect the **color blind and** be colorblind friendly.

39

39

Practical Guides for Using Color in Charts

- Rule 7: Pay attention to the meaning of different colors
  Example: Logo color

40

40

Practical Guides for Using Color in Charts

- Rule 7: Pay attention to the meaning of different colors

Example: Logo color

- Red = Adventure. Energy. (Coca Cola)
- Pink = Compassion. Nurturing. (Breast Cancer Research Foundation; Barbie)
- Orange = Cheerful. Warm. Optimistic. (Amazon)
- Yellow = Happy. Bright. Sunny. (McDonalds)
- Green = Health and Healing. Environment. Organic. (Whole Foods Market; Starbucks)
- Blue = Honesty. Trust. Dependability. (Facebook; Pfizer)
- …

41

41

Practical Guides for Using Color in Charts

- Rule 7: Pay attention to the meaning of different colors

Use colors that viewers are familiar with

- "red" for hot, dangers, cautions, & warnings

>> watch out for culture conflicts (e.g., red in Asian cultures)

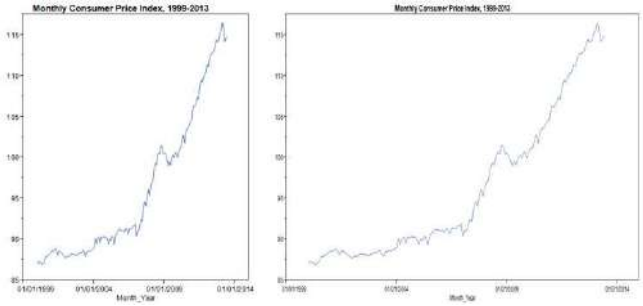
>> watch out for differences of conventions: the meaning of red and green in different stock markets (China vs. USA)

42

42

What is the Best Aspect-ration of a Two-variable Visualization?

- Should not manipulate the aspect ratio to intentionally exaggerate or downplay the rate of change.
- The general consensus seems to be that an aspect ratio that keeps the line segments around 45 degrees going up (or -45 degrees going down) is the best for detecting trends.



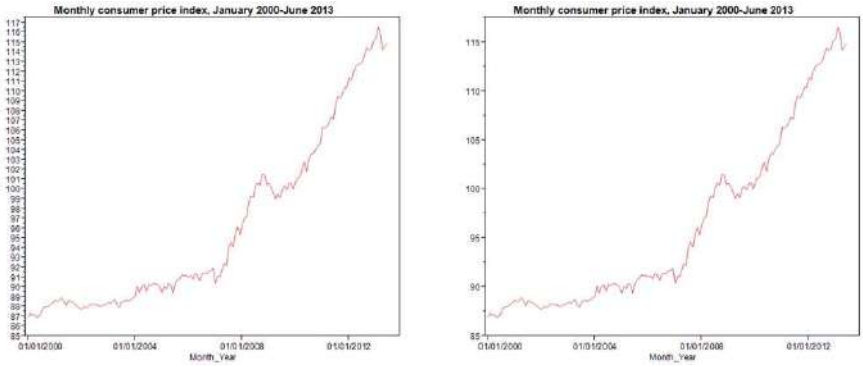
Heer, Jeffrey, and Maneesh Agrawala. "Multi-scale banking to 45 degrees." *IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006): 701-708.

43

43

How many tick marks should you use?

- There is no exact number that works best in all circumstances, and the size of the graph is a factor that must considered: the longer the scale line, the more tick marks it should contain.

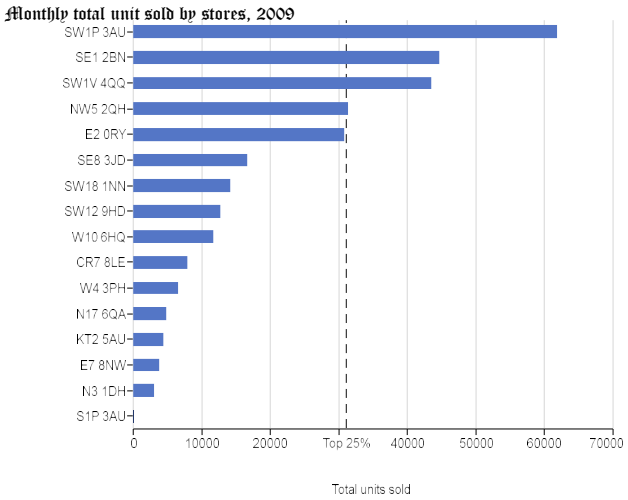


44

44

Graph Typography

- Avoid using artistic fonts

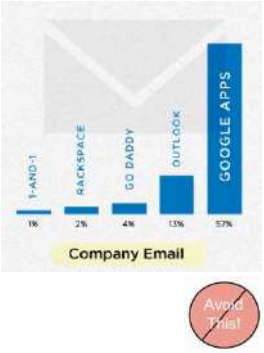


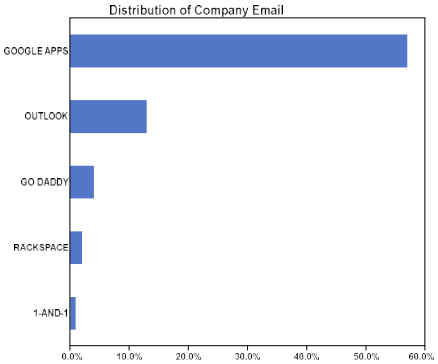
45

45

Graph Labeling

- Orientation of label should be reader friendly





46

46

Recap

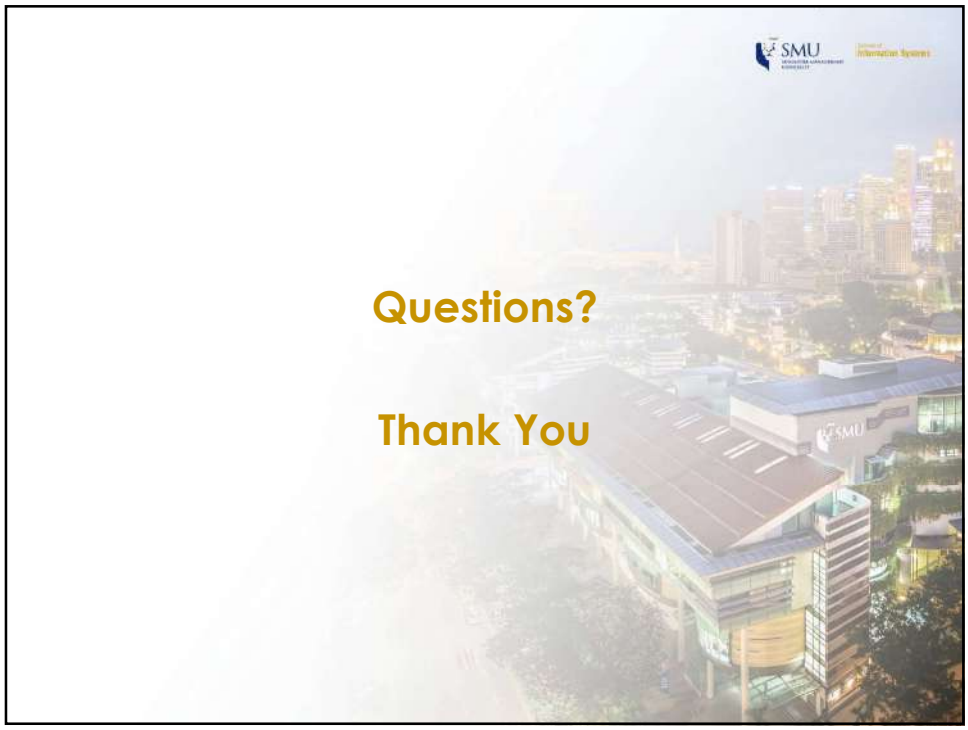
- Visualization Principles by Prof. Edward Tufte
  - Graphical Integrity
  - Data-ink Ratio Maximization

Prof. Edward Tufte from Yale University

- Other Best Practices

47

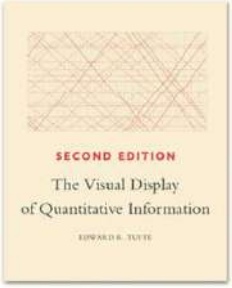
47

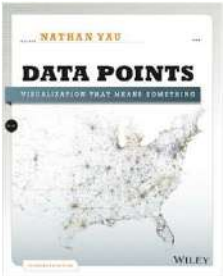


48

Reference

- Tufte, Edward (2nd Edition) The Visual Display of Quantitative Information, Graphics Press LLC, Connecticut, USA
- Yau, Nathan (2013) Data Points: Visualization that means something, John Wiley & Sons, Inc.



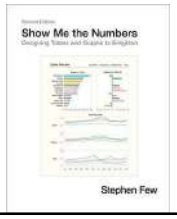


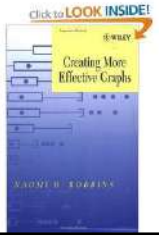
49

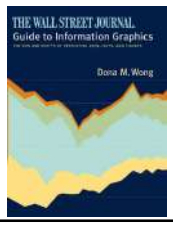
49

Reference

- Few, Stephen (2012) (2nd edition) **Show Me the Numbers: Designing Tables and Graphs to Enlighten**, Analytics Press, Oakland, USA
- Robbins, Naomi B. (2005) **Creating More Effective Graphs**, John Wiley & Sons, New Jersey, USA
- Wong, Dona M. (2010) **The Wall Street Journal Guide to Information Graphics**, W. W. Norton & Company, Inc. New York







50

50

Information Graphics and Data Visualization blogs

- Junk Charts (http://junkcharts.typepad.com/junk_charts/)
- Perceptual Edge (http://www.perceptualedge.com/blog/)
- Statistical Graphics and more (http://www.theusrus.de/blog/)
- EagerEyes (http://eagereyes.org/)
- Visualizing Economics (http://visualizingeconomics.com/)
- The Global Sociology Blog (http://globalsociology.com/)

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 8.2
Interactive Visualization

2

Outline

- What is Interactive Data Visualization?
- Why is Interaction Useful for Data Visualization?
- Seven Types of General Interactions

3

3

What is Interactive Data Visualization?

- Visible Change Under Your Control

- Unlike static charts, interactive data visualisation enables the user to further explore the dataset **by interactively manipulating** and **changing** the visual representation of all or a selected subset of the data using appropriate interaction techniques and input devices.

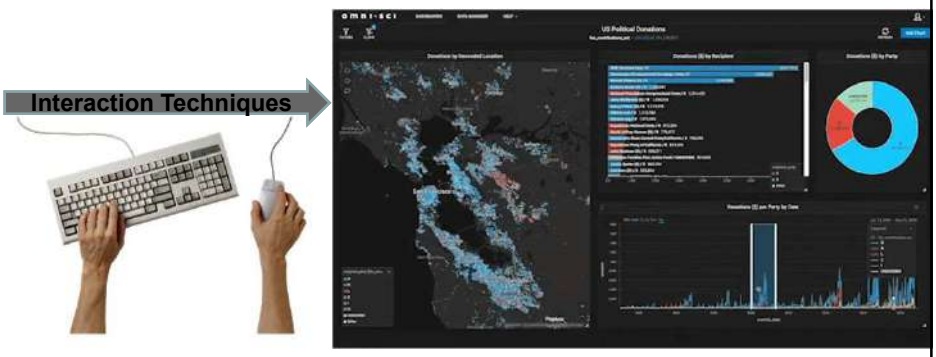


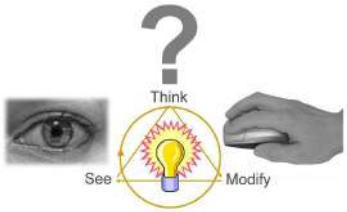
Image from: OmniSci, Interactive Data Visualization Definition - https://www.omnisci.com/technical-glossary/interactive-data-visualization

4

4

Why is Interaction Useful for Data Visualization?

- It is an essential requirement of data visualization



- Ben Shneiderman's mantra:

  "*Overview first, zoom and filter, then details-on-demand*"

http://www.cs.umd.edu/users/ben/

5

5

Why is Interactive Data Visualization Useful?

- Some benefits

- Interactive visualisation is useful when the dataset is **large** and is embedded with **complex data insights** which have yet to be verified or uncovered.

- **Identify relationships** more effectively - by selectively focusing on specific metrics so that non-obvious cause-and-effect relationships within the data can be seen.

-**Identify trends** and **patterns** faster – by having the ability to visualise complex data at varying scales and in decluttered views by using only a selective data subsets.

- **Telling data stories** more convincingly – by presenting a sequenced and focused narrative through the ability to zoom in/out, highlight relevant information, filter, and change the parameters to present multiple viewpoints of the data.

Image from: OmniSci, Interactive Data Visualization Definition - https://www.omnisci.com/technical-glossary/interactive-data-visualization

6

6

7 General Categories of Interactions

- *Select:* mark something as interesting
- *Explore:* show me something else
- *Reconfigure:* show me a different arrangement
- *Encode:* show me a different representation
- *Abstract/Elaborate:* show me more or less details
- *Filter*: show me something conditionally
- *Connect*: show me related items

Yi, Ji Soo, et al. "Toward a deeper understanding of the role of interaction in information visualization." TVCG, 2007.

7

7

Select

"Mark something as interesting"

- Mark items of interest to keep track
- Seems to often work as a preceding action to subsequent operations

Examples

- Select a landmark in Google Map
- Select a bar in a histogram
- Select a set of points by brushing

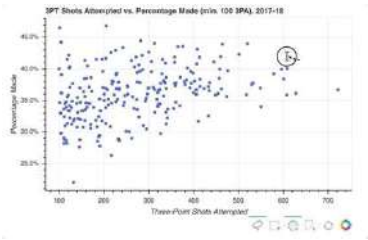


Image from: Leon D'Angio @ Realphython.com - Interactive Data Visualization in Python With Bokeh - https://realpython.com/python-data-visualization-bokeh/#selecting-data-points

8

8

Selection Functions

- ***Pointer*** selects a single object in a plot.
- ***Drag-box*** selects a rectangular region in a box.
- ***Brush*** is a generalisation of the drag-box. Once a rectangular region is defined, the brush allows users to move that region across a plot and thus dynamically change the selected subset.
- ***Slicer*** is an axis-parallel selection tool, which selects a range along an axis, where the end-point of the interval can be modified dynamically.
- ***Lasso*** allows users to define an arbitrary contiguous shape to select data.

9

9

Explore

"Show me something different"

- Exploration enables users to examine a different subset of data
- Exploration overcomes the limitation of display size

Examples

- Panning in Google Earth
- Direct Walking in Visual Thesaurus

Visual Thesaurus: https://www.visualthesaurus.com/app/view

10

10

Explore: More Examples

2D Navigation: MatrixWave [1]

3D Navigation: Interactive visualization of scanned mummies [2]

[1] Zhao et al. MatrixWave: Visual comparison of event sequence data. ACM HFCS, 2015. https://www.youtube.com/watch?v=HranxBG9F_w
[2] Ynnerman, et al. "Interactive visualization of 3D scanned mummies at public venues." ACM Communications, 2016. https://www.youtube.com/watch?v=tl_sJuA2LWg

11

11

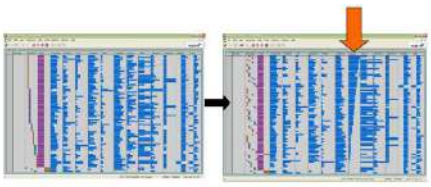
Reconfigure

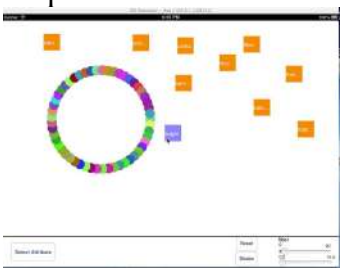
"Show me a different arrangement"

- Reconfiguring provides different perspectives by changing the spatial arrangement of representation

Examples

- Sorting data with respect to a particular attribute
- Changing the attributes in a scatter plot
- Reposition the visual elements





Yi, et al, Dust & Magnet: multivariate information visualization using a magnet metaphor. IEEE https://www.youtube.com/watch?v=laGJ4v7DEU0

12

12

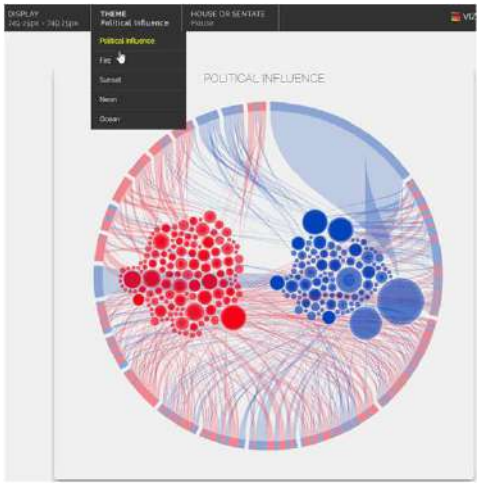
Encode

"Show me a different representation"

- Change visual appearances

Examples

- Change color encoding
- Change size
- Change orientation
- Change font
- Change shape



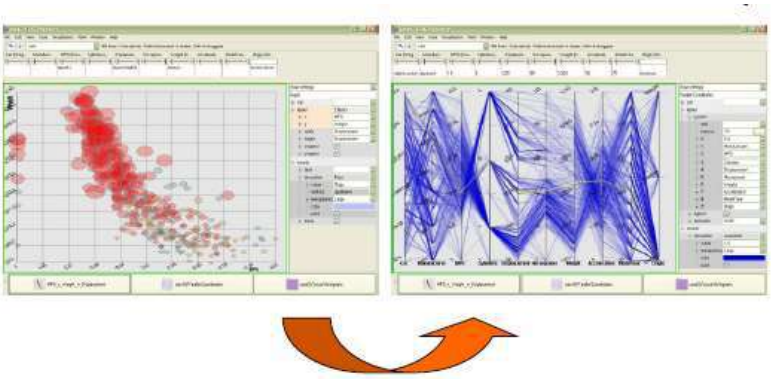
https://github.com/d3/d3/wiki/Gallery

13

13

Encode

- Example: Select a different representation



14

14

Abstract/Elaborate

"Show me more or less detail"

- Adjust the level of abstraction (overview and details)

Examples

- Drill-down in Treemap

  https://observablehq.com/@d3/zoomable-treemap

- Zooming

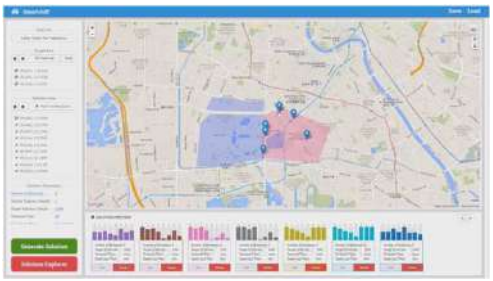


Image from: SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations. https://www.youtube.com/watch?v=7vkYubflVuo

15

15

Filter

"Show me something conditionally"

- Change the set of data items being presented based on some specific conditions

Examples

- Attribute based filtering
- Dynamic query

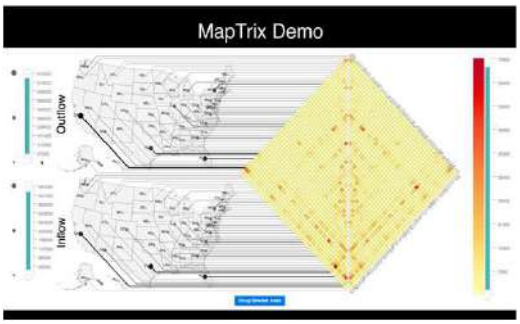


Image from: MapTrix Interactions. https://www.youtube.com/watch?v=kc1PP4J9BH8

16

16

Connect

"Show me related items"

- Highlight associations and relationships
- Show hidden data items that are relevant to a specified item

Examples

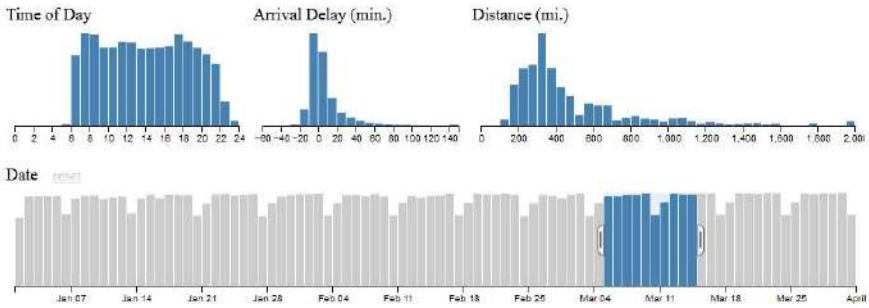
- Linking the relevant items in SmartAdp and *Vizster*

Image from: SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations. https://www.youtube.com/watch?v=7vkYubflVuo
Vizster: Visualizing Online Social Networks. https://www.youtube.com/watch?v=UxsACr2d-iA

17

17

Connect: Linked Multiple Views (Demo)



March 14, 2001

04:38 PM	ONT	SMF	389 mi.	+3 min.
11:50 PM	MDW	BWI	611 mi.	+186 min.
11:50 PM	MCI	MDW	405 mi.	−10 min.
11:49 PM	STL	MDW	251 mi.	+45 min.

26,093 of 231,083 flights selected.

Demo Link:
https://square.github.io/crossfilter/

18

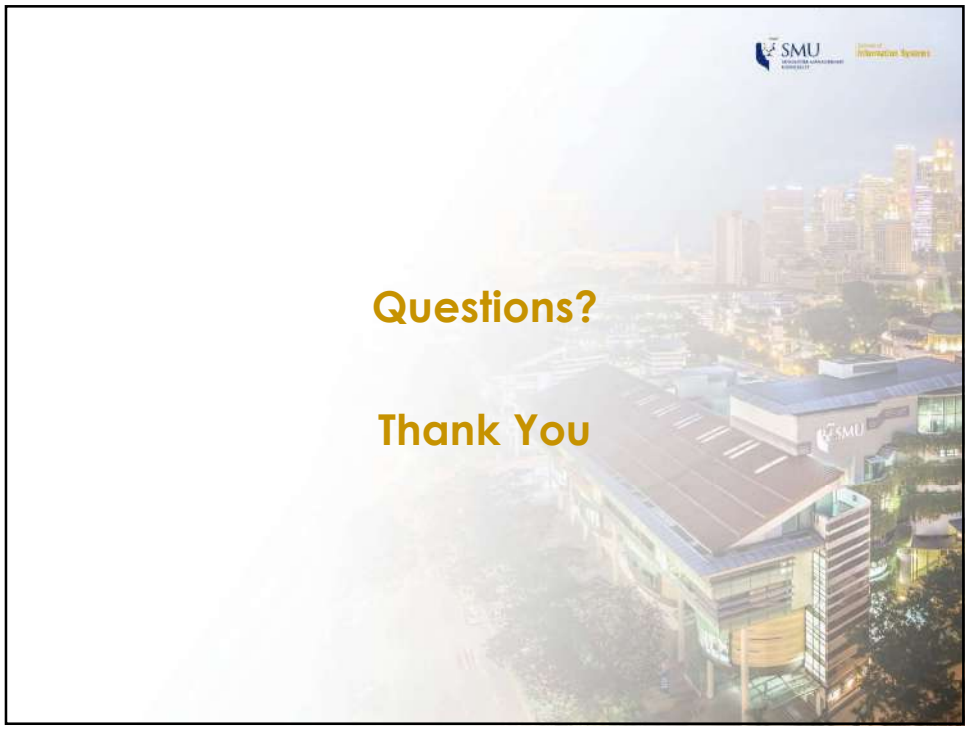
18

Summary

- Interaction is an essential part of data visualization, which provides users more flexibility and control to explore the data of their interest.
- Without interaction, a visualization technique or system becomes a static image or autonomously animated images, limiting the usefulness of data visualization with the increase of dataset size and dimensions.
- You are recommended to enable necessary interactions in your own visualization approaches or systems!

19

19



20

Interesting Data Visualizations

- https://public.tableau.com/en-us/gallery/?tab=viz-of-the-day&type=viz-of-the-day
- https://bost.ocks.org/mike/
- https://archive.nytimes.com/www.nytimes.com/interactive/2013/04/08/business/global/asia-map.html

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 9.1
Exploratory Data Analysis

2

Outline

- What is Exploratory Data Analysis?
- Correlation Analysis
  - Measurement of Correlation
  - Univariate Analysis
  - Bivariate Analysis
  - Confirmatory Analysis
  - Multivariate Analysis

3

3

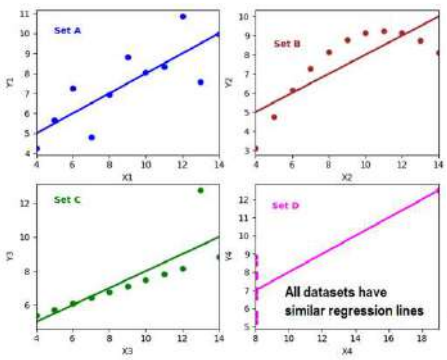
Statistical Analysis Alone is Not Enough

A Review

- Basic statistical analysis alone may not provide useful insights into how datasets can have **differing characteristics** despite having **similar statistical** parameters.
- Francis Anscombe[1] used the **Anscombe's Quartet**[1] of four datasets to show why **visualisation** is a **crucial part of data analysis**.

	Set A		Set B		Set C		Set D	
	X1	Y1	X2	Y2	X3	Y3	X4	Y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Correlation Coefficient	0.82		0.82		0.82		0.82	
Std Dev	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94

Anscombe's Quartet



[1] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21

4

4

What is Exploratory Data Analysis?

- Exploratory Data Analysis (EDA) is a process of **examining** (mostly **graphically**) a dataset to **discover** patterns and relationships, spot anomalies, formulate and test **initial assumptions** using statistical measures[2].
- The primary aim of EDA is to **examine what data can tell us** before actually going through formal modelling and hypothesis formulation[2].
- EDA is usually an iterative process that involves:

1. **Asking questions** about the available data.
2. **Construct** appropriate data **visualisations** to answer the questions.
3. **Evaluate** and **inspect** the answers, then derive further questions.

Repeat

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

Correlation Analysis

Are There Relationships in the Dataset?

- Datasets often have **different categories** (i.e. columns) containing many **measured values** describing a particular phenomena, event or situation.
- Since these values are collected from the **same event**, there is a possibility that they are **related** to one another[2].
- **Correlation** is the statistical technique that examines these relationships and describes how strongly different categories of measures or values are related.
- Correlation answers questions such as:
  - How does one variable change with respect to another?
  - If it does change, to what degree or strength is this change related?

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

Measure Of Correlation

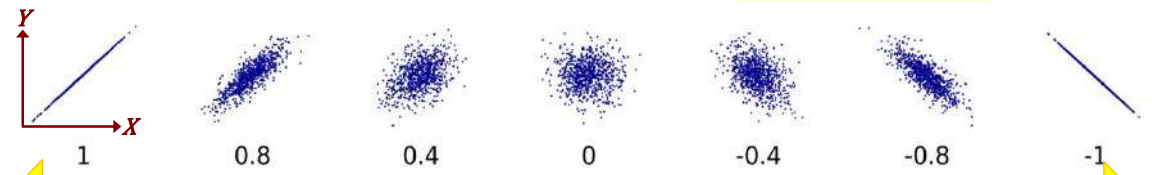
Pearson Correlation Coefficient

- Correlation measures how variables **change in tandem**, either in the same or in opposite directions, and also the magnitude or **strength** of these related changes between the variables.
- The **Pearson correlation coefficient** (ρ_{xy}), with a value between -1 and +1, measures the correlation between two variables X and Y. It is given by:

where: cov is the covariance

σ_X and σ_Y are the standard deviation of X and Y respectively.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \qquad \text{Eqn. (1)}$$



Strong Correlation (> 0.7)

[3] Image from Wikipedia, Pearson Correlation Coefficient - https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Strong Correlation (< -0.7)

7

7

Univariate Analysis

One Variable At A Time

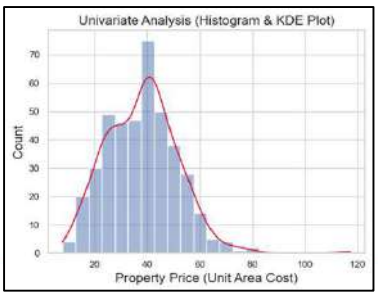
- Univariate analysis involve the analysis of **one variable at a time** in the dataset. As such, it is not meant to find relationships between variables.
- The main purpose of univariate analysis is to:
  - find patterns in the variable such as its **central tendency** (e.g mean, median, mode)
  - understand its dispersion (e.g range, variance, maximum & minimum quartile, etc)
  - detect presence of outliers.

```
Property[Price].decsribe()
count    414.000000
mean      37.980193
std       13.606488
min        7.600000
25%       27.700000
50%       38.450000
75%       46.600000
max      117.500000
Name: Price, dtype: float64
```

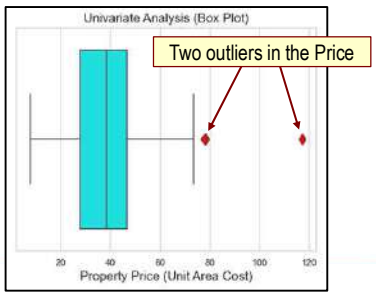
Apply `describe()` on category **Price** in *Property.csv*

```
Property[Price]
Mean = 37.98
Median = 38.45
```

`mean()` and `median()` values of Price



Histogram and Kernel Density Curve



Box Plot (Observing Outliers)

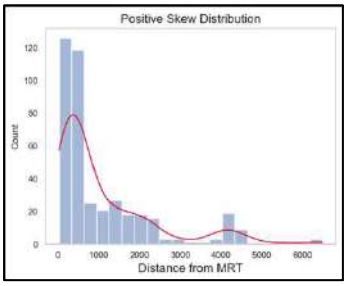
8

8

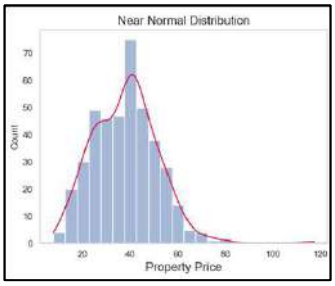
Univariate Analysis

Is My Distribution Normal Enough?

- In order to use **parametric statistical methods** (e.g. Student's t-test), our data must have a **normal distribution**, otherwise nonparametric statistical methods must be used.
- The **normality** of the data or how close they conform to a Gaussian distribution can be tested using visual methods (e.g. **statsmodels's** `qqplot()`)[4].
- Numerical evaluations can be done using **statistical methods**[5] such as:
  - Shapiro-Wilk test
  - D'Agostino's K^2 test
  - Anderson-Darling test



Positive Skew – **Property['MRT']**



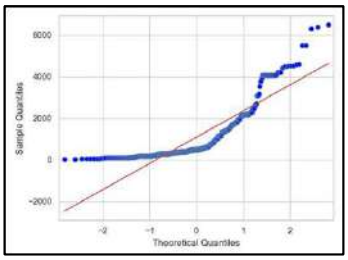
Near Normal – **Property['Price']**

[4] Statsmodels Q-Q plot function - https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html
[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

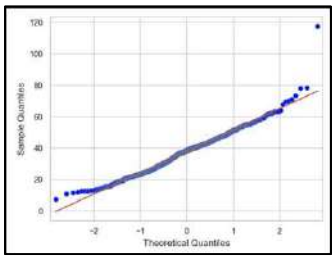
Univariate Analysis

Visualising Normality

- A popular plot for checking normality is the **quantile-quantile plot** or Q-Q plot[6].
- The plot generates its own sample of the ideal Gaussian distribution and divides these samples into groups (e.g. 5) called **quantiles**. Each data point in our sample is **paired** with a similar member in the idealised distribution and plotted on the same cumulative distribution[5].
- **Perfect normality** will see all the data points fall along the 45° **red** line (i.e. matching ideal distribution).
- Dots seen deviating from the line shows deviation from the expected normal distribution.



Positive Skew – **Property['MRT']**



Near Normal – **Property['Price']**

[6] Paras Varshney, Q-Q Plots Explained (2020) - https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0
[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

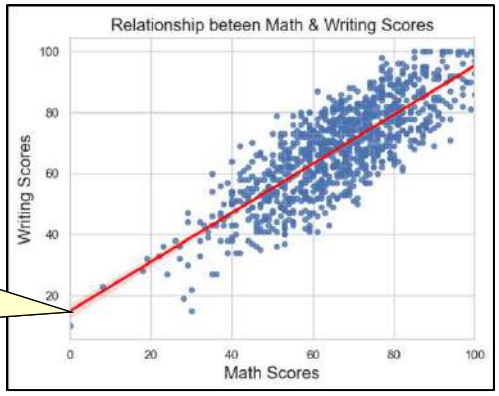
Bivariate Analysis

How Are These Two Variables Related?

- Bivariate analysis is used to find out if there is a **relationship** between two different variables in the dataset.
- A common plot used for bivariate analysis is the two-axes **scatter plot**.
- A relationship exist if the data points seem to **fit around** a **line** or a **curve**. The **tighter** the clustering, the **stronger** the relationship.

Did students getting high Math scores also got high Writing scores?

Yes, there is a reasonably strong positive correlation between Math and Writing scores



Writing vs Math Scores in *Student performance.csv*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

11

11

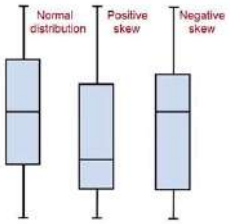
Bivariate Analysis

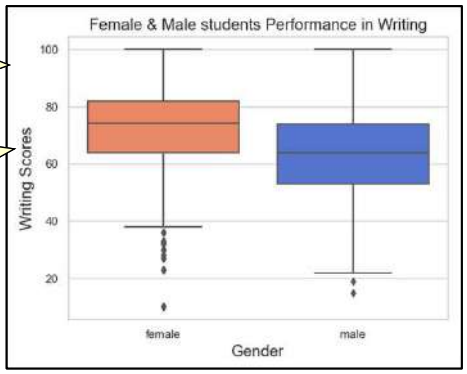
What If One Variable is Categorical?

- Bivariate analysis using a **box plot** allows us to examine the **statistical measures** of one quantitative variable along with the relationship between multiple values of a **categorical variable** in the dataset.
- The **central lines** in the box plot allow us to **compare** the **median** values within the categories (e.g. female & male).
- The nature of the data distribution (i.e. normal or skewed) can be inferred from the **relative position** of the **central line** within the box.

Do female students do better in writing?

They seem to have a higher median score.





Box plots of Writing scores of Male & Female students in *Student performance.csv*

12

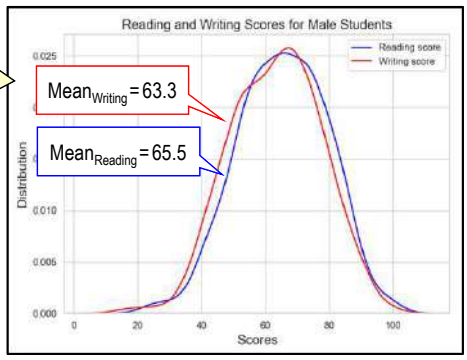
12

Bivariate Analysis

Confirmatory Analysis

- Visual analysis is useful for spotting relationships or patterns in data. However, we should check our assumption using appropriate statistical measures and **confirm** if an observation is **statistically significant**.
- The following confirmatory analysis steps could be taken:
  - Formulate a hypothesis.
  - Check normality of data distributions.
  - Select appropriate statistical test (e.g parametric or non-parametric).
  - Compute the test statistics to determine the **p** value (typically **p** < 0.05 is considered significant)

Are the differences between the reading & writing scores of the boys significant?



KDE plots of Reading & Writing scores of Male students in *Student performance.csv*

13

13

Confirmatory Analysis

Formulating a Hypothesis

- A hypothesis is a statement about the value of a population parameter. In order to test a hypothesis, we formulate a **null** hypothesis and the **alternate** hypothesis.
- **Null hypothesis (H_0)** - a default position statement that there is no relationship between two measured phenomena or no association among groups[7].
- **Alternate hypothesis (H_a)** - a statement that is contrary to the null hypothesis. It is usually the hypothesis being tested since we see a pattern, think it is true and want to find evidence to reject the null hypothesis and replace it with the alternate.

Null hypothesis (H_0)	$\mu_{Reading} = \mu_{Writing}$	(No difference in performance)
Alternate hypothesis (H_a)	$\mu_{Reading} > \mu_{Writing}$	(Reading scores better than writing)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[7] Yogesh Agrawal, Hypothesis testing in Machine learning using Python (2019) - https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

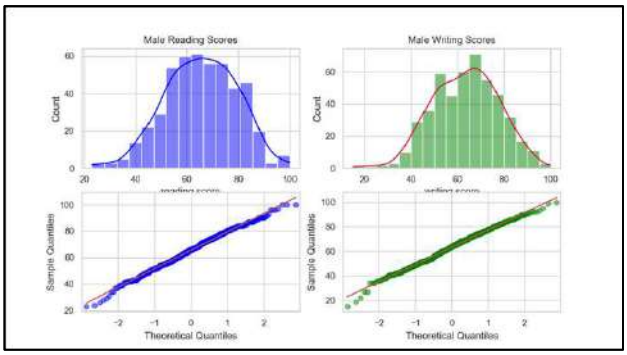
14

14

Confirmatory Analysis

Check For Normality

- In order to decide if we should use a parametric or non-parametric statistical test, we should check the **normality** of the two distributions we are comparing.
- We can do a normality check using an appropriate statistical tests[5] or use the **Q-Q plot**[4].
- Visually, both the Reading & Writing score distributions of male students seem to conform to a Gaussian distribution.
- We can do further numerical statistical test to verify this.



Histogram/KDE plots of Reading & Writing scores and their Q-Q plots

[4] Statsmodels Q-Q plot function - https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html
[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

Confirmatory Analysis

Statistical Test for Normality

- The **Shapiro-Wilk** test evaluates a sample of data and quantifies how likely the data sample is drawn from a Gaussian distribution[5]. Python **scipy.stats** library provides several functions for normality statistical tests, including the Shapiro-Wilks.

```
from scipy.stats import shapiro          # import the Shapiro-Wilks test
S,p = shapiro(Dist)                      # run the Shapiro-Wilks test
if (p > 0.05):                           # check p value
  print('Distribution is normal')        # distribution is normal if p > 0.05
```

- Both Reading & Writing distributions meet the normality confidence level of 0.05.
- We can thus look at using a parametric test (e.g. the t-test or Z-test)

```
Distribution ( male , reading score ) mean = 65.47
Shapiro-Wilk test = ( Stats= 0.99 , p = 0.08968 )
Distribution ( male , reading score ) is normal

Distribution ( male , writing score ) mean = 63.31
Shapiro-Wilk test = ( Stats= 0.99 , p = 0.10408 )
Distribution ( male , writing score ) is normal
```

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) - https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

Confirmatory Analysis

Z-test or Student's t-test?

- If there is **more than 30 data samples**, the Z-test can be used, otherwise the t-test should be used with small sample sizes[7].
- Both the Z-test and t-test requires the data to be **normally distributed** and data points are **independent** from each other (i.e. one data point does not affect another data point). Strict normality requirements can be relaxed if the sample size is large.
- The Python **statsmodels** library provides both these statistical tests.

```
from statsmodels.stats.weightstats import ttest_ind   # t-test function
from statsmodels.stats.weightstats import ztest       # Z-test function
```

- Given that the sample size of male students is 482, the Z-test can be used here.

[7] Yogesh Agrawal, Hypothesis testing in Machine learning using Python (2019) - https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

Confirmatory Analysis

One or Two-Tailed Test?

- **Two-tailed test** – use it if you want to determine if there is a difference between two groups and have no concern about the direction of this difference. Example:

Null hypothesis ($\mathbf{H_0}$)	$\mu_{Reading} = \mu_{Writing}$	(No difference in performance)
Alternate hypothesis ($\mathbf{H_a}$)	$\mu_{Reading} \neq \mu_{Writing}$	(There is a difference in performance)

- **One-tailed test** – use it if you want to determine if there is a difference between groups in a **specific direction**. In our case, the Reading distribution seems to be visually higher than Writing one. We should therefore use:

Alternate hypothesis ($\mathbf{H_a}$)	$\mu_{Reading} > \mu_{Writing}$	(Reading performance better than Writing)

- **Note:** One-tailed test has more statistical power than a two-tailed test at the same significance level. Results are more likely to be significant for a one-tailed test if there is indeed a difference between the groups in the direction predicted[8].

[8] Should you use a one-tailed test or a two-tailed test for your data analysis? - https://www.statisticssolutions.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/

Confirmatory Analysis

Statistical Significance

- **Z-test results** – Applying one-tailed ('larger') Z-test[9] on the **Reading** & **Writing** data:

```
Z,p = ztest(Reading, Writing, alternative='larger')   # call Z-test function
```

Z score | p value

```
Z-test = ( 2.393229395263579 , 0.008350397872215826 )
```

- **Rejecting null hypothesis** – the threshold at which is generally considered safe to reject the null hypothesis is a ***p*** value of (***p*** < 0.05). This means there is a less than 5% chance that the observed data is due to chance.
- **Statistically Significant** – since the ***p*** value computed by the Z-test is way smaller than 0.05, we can safely say that the observed mean Reading score being higher than the mean Writing score for Male students is statistically significant.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

[9] Statsmodels ztest fucntion - https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html

19

19

Confirmatory Analysis

Selecting The Statistical Test

- Selecting an appropriate statistical test for your confirmatory analysis requires consideration of various factors such:
  - Continuous or discrete data.
  - Number of samples to compare.
  - Sample size.
  - Parametric or non-parametric test.
- A helpful flowchart in[10] shows some statistical tests that can be use based on the nature of the data and test requirements.

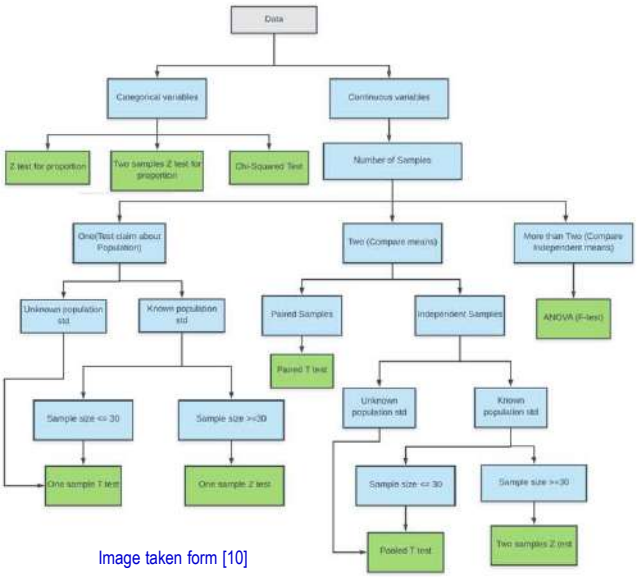


Image taken form [10]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

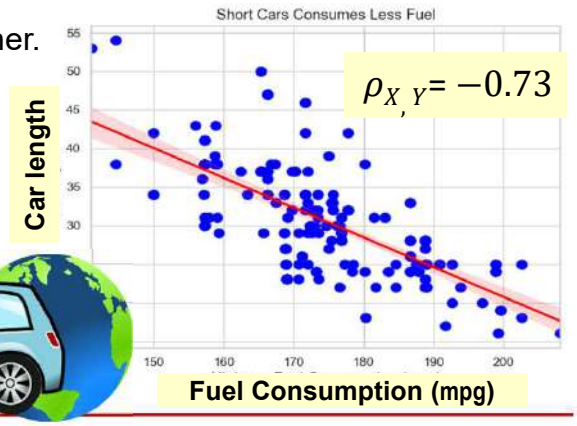
[10] Jagandeep Singh, Statistical Tests with Python (2020) - https://python.plainenglish.io/statistical-tests-with-python-880251e9b572

20

20

Correlation Does Not Imply Causation

- Remember the phrase "**Correlation does not imply causation**"[2].
- Just because two things are observed to be correlated does not mean one causes the other.
- Do not form conclusions **too quickly** based on correlation. **Invest time** to find underlying factors in the data that really influence the relationships between parameters



Strong negative correlation found between the length of the car and its fuel efficiency.
Save the Planet,
Drive Short Cars now!

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

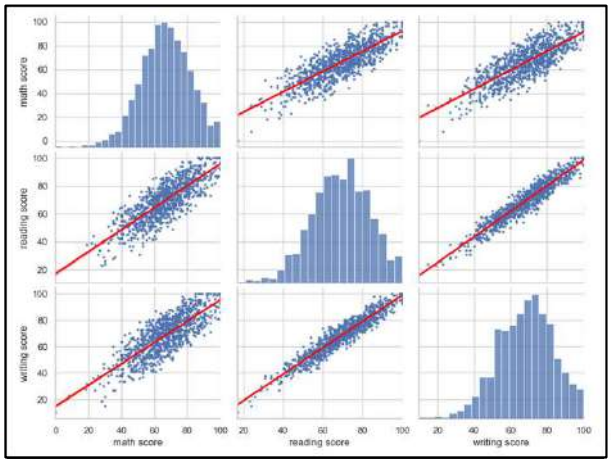
21

21

Multivariate Analysis

Analysis of Three or More Variables

- Multivariate analysis allows us to look at the **correlation of three** or **more** variables at a time.
- A common way of visualising multivariate data is to use a **matrix** of multiple **scatter plots**.
- The convenient and powerful Seaborn `pairplot()` function provides this feature[11].
- Setting the `kind` parameter to `kind='reg'` will show the linear regression lines.



Analysing Math, Reading & Writing scores in *Student performance.csv*

[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

22

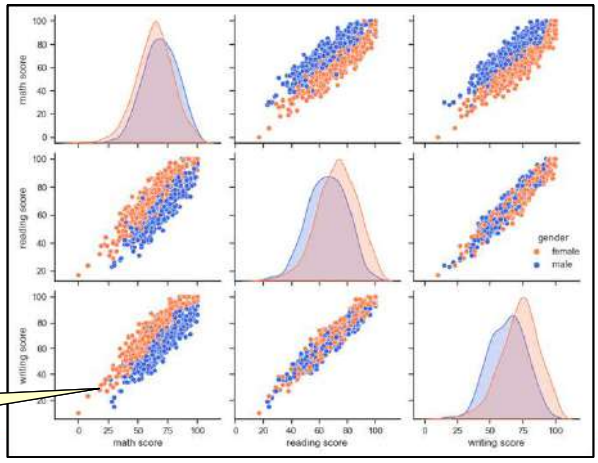
22

Multivariate Analysis

Adding In Categorical Variables

- An additional **categorical variable** can be visualised in **colours** by employing the `hue` parameter in `pairplot()`[11].
- By assigning the hue parameter to a specific categorical variable (e.g. `hue='gender'`), each type in the category is assign a different colour from a default or specified palette.
- Such multivariate analysis allows interesting relationships and patterns to be observed in the different categorical types.

Boys do better in math than writing



Math, Reading & Writing scores of Female & Male students

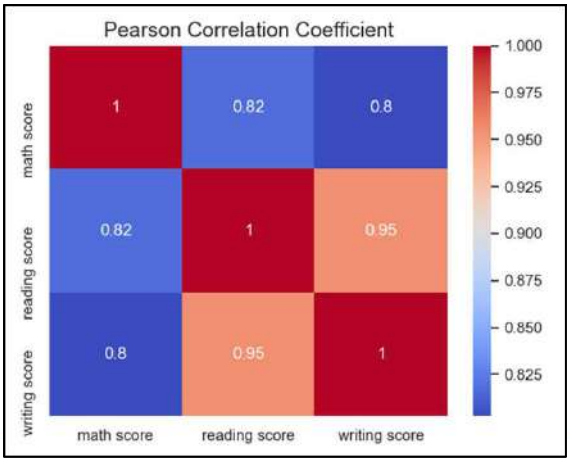
[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

Multivariate Analysis

Getting the Correlation Values

- Instead of just visualising the correlation, it is often useful to compute a numerical measure for each correlation pair.
- **Pandas** `corr()` function and **Seaborn**'s `heatmap()` can compute and display the **pairwise** linear (Pearson) correlation coefficients for all the columns in the dataframe with **numerical** values[12].

```
sns.heatmap(
  Data.corr(method='pearson'),   # Pearson coeff
  annot = True,                  # show coeff values
  cmap = 'coolwarm'))            # diverging palette
```



Correlation coefficients of Math, Reading & Writing scores

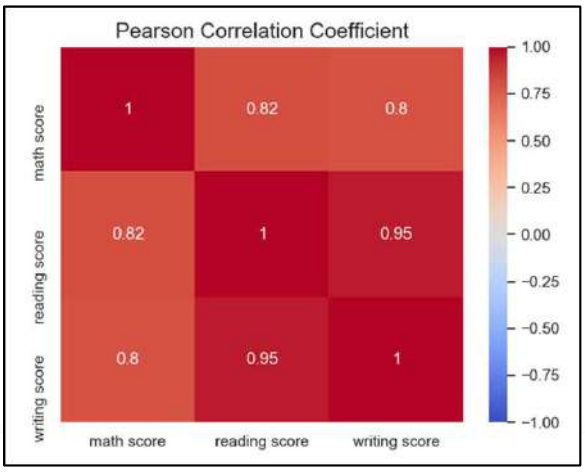
[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

Multivariate Analysis

Getting the Correlation Values

- Instead of just visualising the correlation, it is often useful to compute a numerical measure for each correlation pair.
- **Pandas** `corr()` function and **Seaborn**'s `heatmap()` can compute and display the **pairwise** linear (Pearson) correlation coefficients for all the columns in the dataframe with **numerical** values[12].

```
sns.heatmap(
  Data.corr(method='pearson'),   # Pearson coeff
  annot=True,                    # show coeff values
  vmin=-1, vmax=1,               # min-max scale
  cmap='coolwarm'))              # diverging palette
```



[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

25

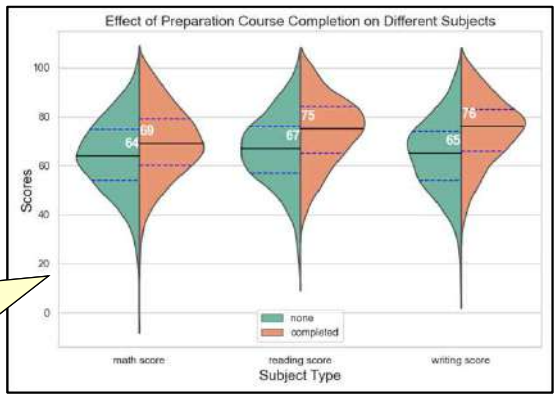
25

Multivariate Analysis

Playing With Many Violins

- Multivariate analysis using **violin plots** allows us to compare the **probability density curve** of multiple quantitative variables for a **categorical variable pair**.
- The **central lines** in the violin plot allow us to **compare** the **median** values within the categories (e.g. those who completes test preparation course & those who did not).
- In Seaborn's `violinplot()`, the `hue` parameter can be assigned to a categorical variable with two values. Setting `split=True`, then compares them side-by-side.

Students do better in all subject when completing test preparation.



Violin plots showing effects of prep course completion on Math, Reading & Writing scores

26

26

Summary

Exploratory Data Analysis

- Correlation analysis allows us to visualise **relationships** within the dataset.
- Data visualisation techniques for this purpose include histogram & kernel density plots (univariate), scatter plot (bivariate), box & violin plots (bivariate & multivariate), paired plot & heatmap (multivariate).
- Correlation between variable pairs can be quantified using the **Pearson correlation coefficient** and this values can be effectively visualised using appropriate **colour palettes** in heatmaps and paired plots.
- Visual observation of relationships and patterns should be confirmed with appropriate **statistical analysis** based on the **nature** and **number** of data distributions being compared.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

27

27

References for Correlation Analysis

[1] Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21

[2] S.M. Mukhiya & U. Ahmed, Hands-on Exploratory Data Analysis with Python. Packt Publishing (2020)

[3] Image from Wikipedia, Pearson Correlation Coefficient - https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[4] Statsmodels Q-Q plot function - https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html

[5] Jason Brownlee, A Gentle Introduction to Normality Tests in Python (2018) – https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

[6] Paras Varshney, Q-Q Plots Explained (2020) - https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

[7] Yogesh Agrawal, Hypothesis testing in Machine learning using Python (2019) – https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce

[8] Should you use a one-tailed test or a two-tailed test for your data analysis? - https://www.statisticssolutions.com/should-you-use-a-one-tailed-test-or-a-two-tailed-test-for-your-data-analysis/

[9] Statsmodels ztest fucntion - https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html

[10] Jagandeep Singh, Statistical Tests with Python (2020) - https://python.plainenglish.io/statistical-tests-with-python-880251e9b572

[11] Seaborn pairplot() documentation - https://seaborn.pydata.org/generated/seaborn.pairplot.html

[12] Pandas dataframe.corr documentation - https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html

[13] Bibor Szabo, How to Create a Seaborn Correlation Heatmap in Python? (2020) – https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

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

28

28

Acknowledgement

The slides of this lecture is adapted from the course material by Prof. Goh Wooi Boon, NTU.

29

SC4024/CZ4124
Data Visualization

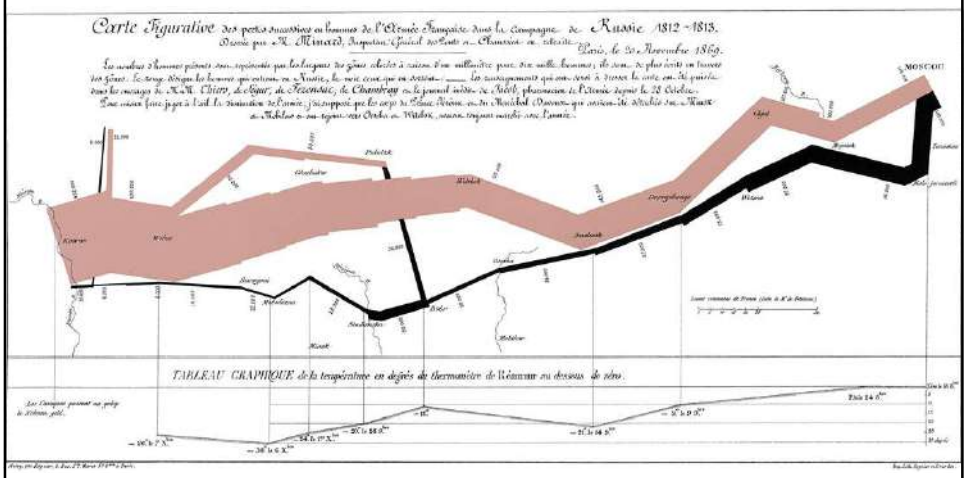
Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 9.2
Time-series Data Visualization

2

A Famous Example: Napoleon's Retreat from Moscow



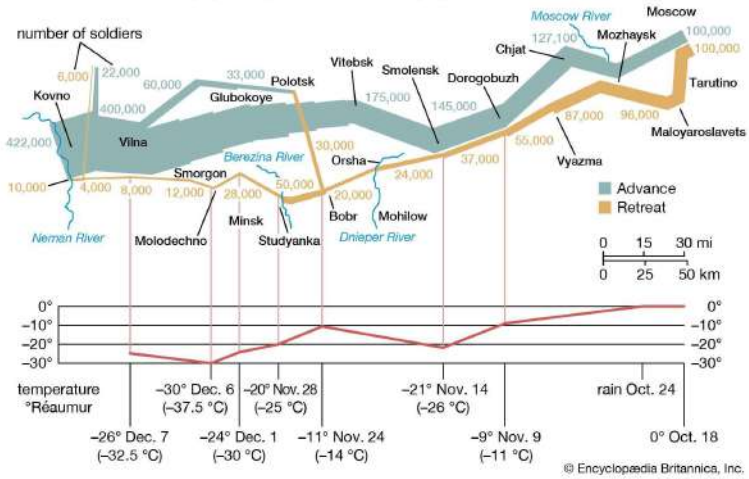
Charles Minard's 1869 chart showing the number of men in Napoleon's 1812 Russian campaign army, their movements, as well as the temperature they encountered on the return path.

3

3

A Famous Example: Napoleon's Retreat from Moscow

Based on Charles Minard's graph of Napoleon's Russian campaign of 1812.



Charles Minard's 1869 chart showing the number of men in Napoleon's 1812 Russian campaign army, their movements, as well as the temperature they encountered on the return path.

4

4

Outline

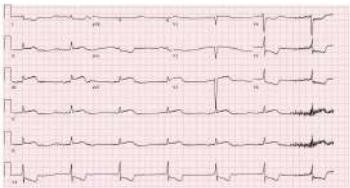
- Time-series data
- Time representation
- Time-series patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

5

5

Time-series Data

- Time series data, also referred to as time-stamped data, is a sequence of data points indexed in time order.
- Examples
  - Stock price
  - Temperature
  - Economic indicators
  - Patient health evolution metrics (e.g., ECG data)
  - Sensor data



Large volume, high dimensional, multivariate, widely used

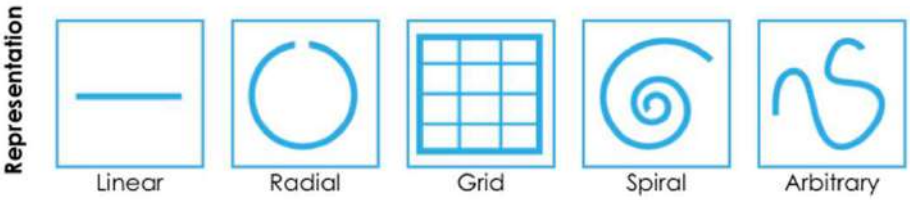
6

6

Time Representation

- Taxonomy

There are mainly five categories of time representations used in time-series data visualizations:



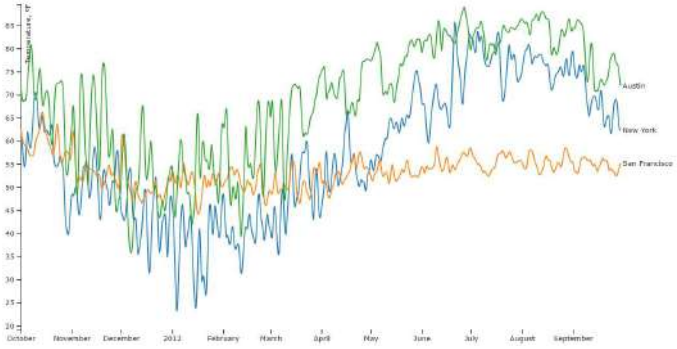
Brehmer, Matthew, et al. "Timelines revisited: A design space and considerations for expressive storytelling." TVCG 23.9 (2016): 2151-2164.

7

7

Linear Time

- Present time data as a 2D line graph
  - Time on x-axis
  - The other variable on y-axis
  - The most common way to represent a timeline

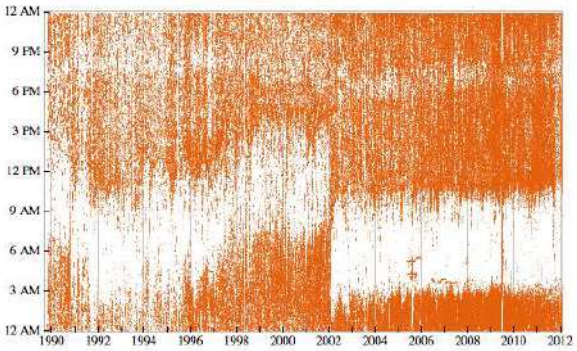


8

8

Linear Time

- Personal Data Visualization, *by* Stephen Wolfram
  - The x-axis and y-axis represent the year and daily timeslot respectively, and each dot represent an email
  - It shows the time of over 1 million emails since 1989



http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/

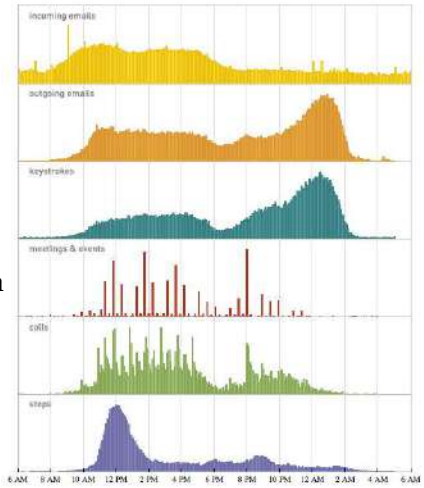
9

9

Linear Time

- Personal Data Visualization, *by* Stephen Wolfram

Visualization of other personal time-series data



http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/

10

10

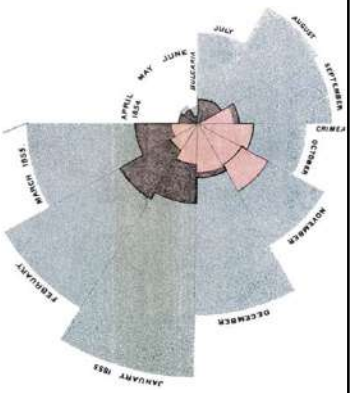
Radial Time

- Data distributed along the spiral
- It is appropriate for presenting the periodic nature of time, such as natural cycles of people, weather systems, etc.

Example: Nightingale Rose Charts:

- It shows the causes of mortality in the army in the East and was created by Florence Nightingale

- It indicates the annual rate of mortality per 1,000 in each month that occurred from preventable diseases (in blue), wounds (in red), and other causes (in **black**).



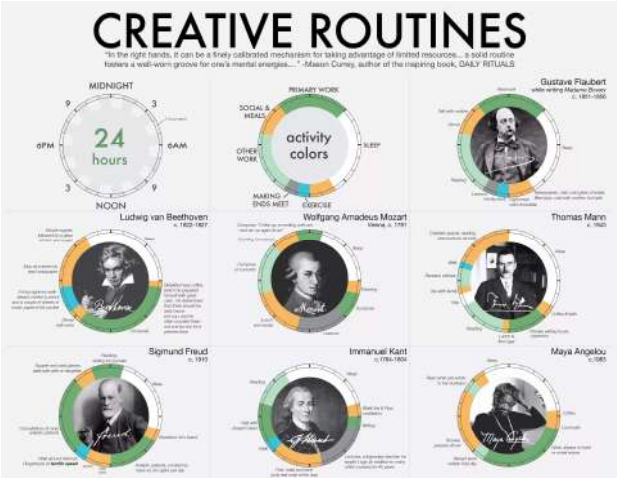
https://en.m.wikipedia.org/wiki/File:Nightingale-mortality.jpg

11

11

Radial Time

Example: Historic creative's daily routines



https://infowetrust.com/project/routines

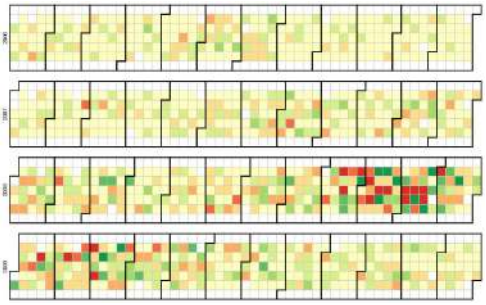
12

12

Grid Time

- Grid representations such as the standard Month-Week-Day calendar are ubiquitous.
- However, grid representations are only appropriate for timelines that reflect these granularities of time.

Example: Dow Jones stock price from 2006 to 2009



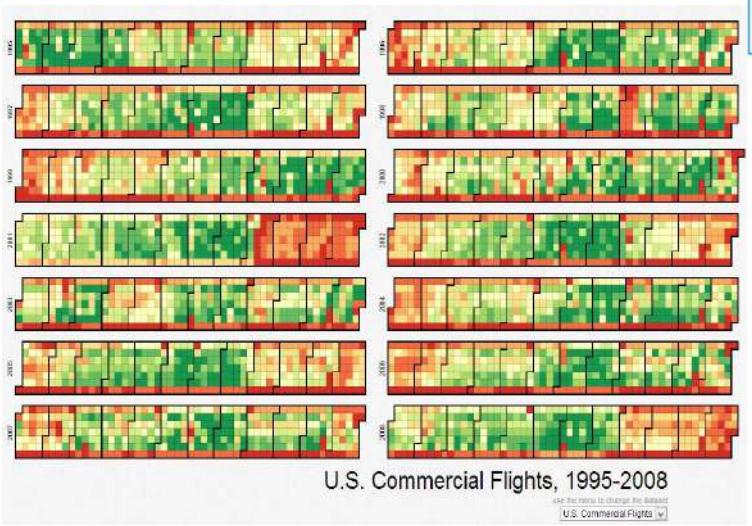
http://mbostock.github.io/d3/talk/20111018/calendar.html

13

13

Grid Time

Example: US Commercial Flights



Source: http://mbostock.github.io/d3/talk/20111018/calendar.html

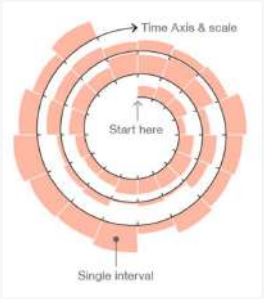
14

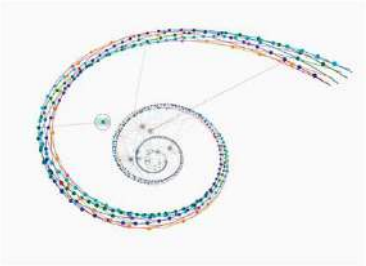
14

Spiral Time

- Spiral representations are dense, space-filling shapes that radiate to or from a centre point.
- Spiral timelines are aesthetically appealing and appropriate for presenting many events within a single dense display

Example: Ross Spiral Curriculum





Ross spiral curriculum, 2015. https://spiral.ross.org/spiral

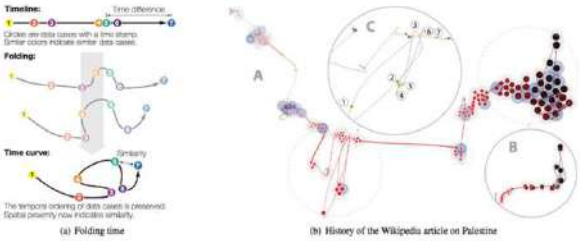
15

15

Arbitrary Time

- Arbitrary timeline representations may appear visually similar to connected scatterplots and time curves
- Like spiral representations, arbitrary representations can be useful for presenting many events within a single dense display and improve the memorability of a time-series data

Example: Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data



http://www.aviz.fr/~bbach/timecurves/

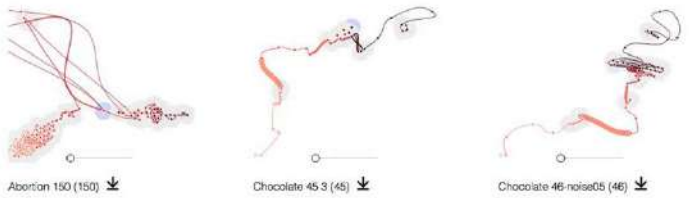
16

16

Arbitrary Time

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Demo Video of Time Curves



http://www.aviz.fr/~bbach/timecurves/

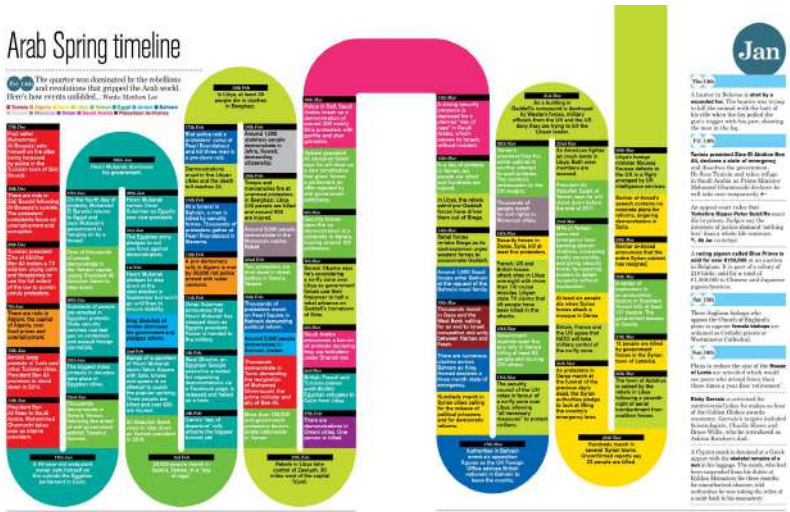
17

17

Arbitrary Time

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Example: Arab Spring Timeline



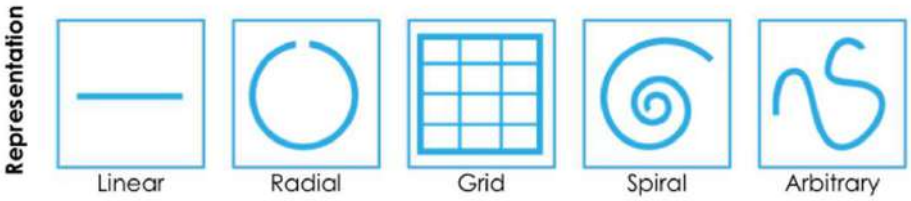
https://www.slow-journalism.com/infographics/society/arab-spring-timeline

18

18

Time Representation

- Taxonomy



- Among the above four types, linear time representation is more popular

Brehmer, Matthew, et al. "Timelines revisited: A design space and considerations for expressive storytelling." TVCG 23.9 (2016): 2151-2164.

19

19

What will you learn from this lesson?

- Time-series data
- Time representation
- Time-series data patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

20

20

Time-series data patterns

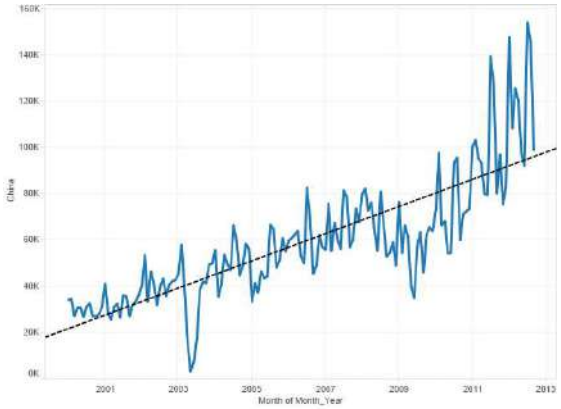
- Trend
- Rate of change
- Co-variation
- Cycles
- Exceptions

21

21

Time-series Patterns: Trend

- The overall or general direction of change in a series of time-series values is called the trend, which is often displayed in a graph as a trend line.

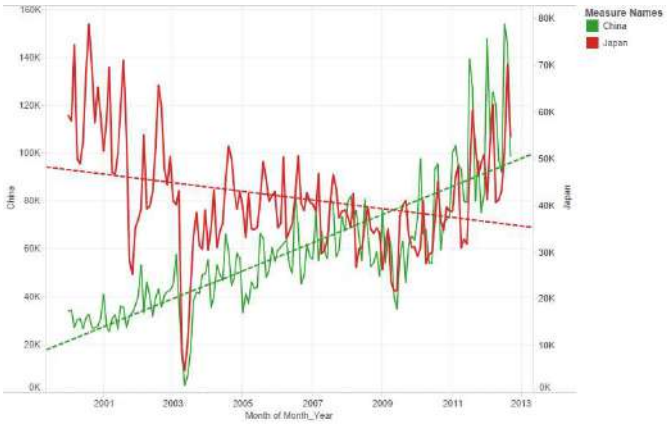


22

22

Time-series Patterns: Trend

- Example: Visitors from China is in increasing trend and visitor from Japan is in decreasing trend

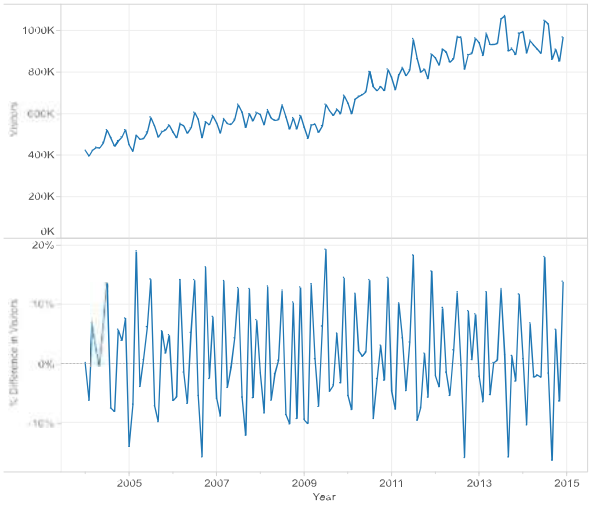


23

23

Time-series Patterns: Rate of change

- The percentage difference between one value to the next value.

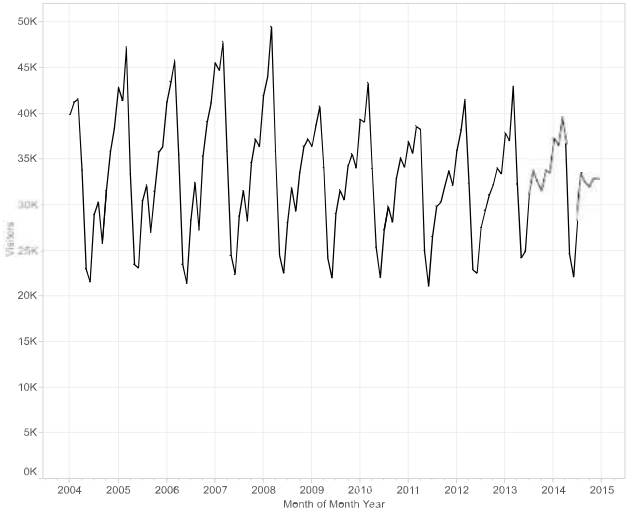


24

24

Time-series Patterns: Cycles

- Cycles are patterns that repeat at regular intervals.



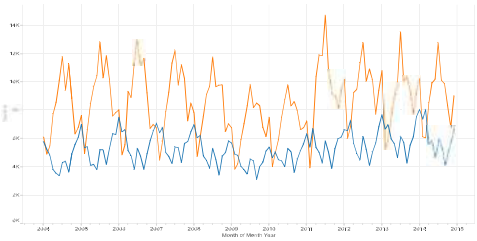
25

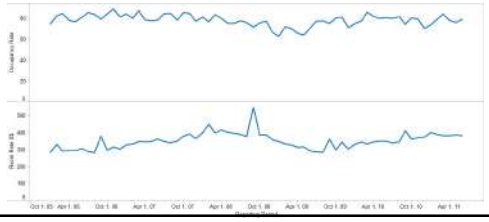
25

Time-series Patterns: Co-variation

- When two time series relate to one another so that changes in one are reflected as changes in the other, either immediately or later.

With Co-variation





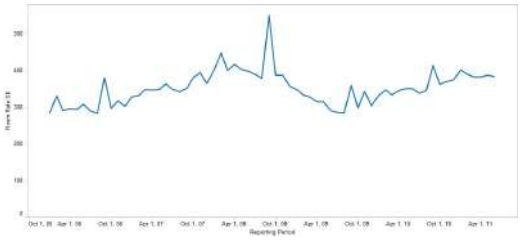
Without Co-variation

26

26

Time-series Patterns: Exceptions

- Values that fall outside the norm.



27

27

What will you learn from this lesson?

- Time-series data
- Time representation
- Time-series data patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

28

28

Massive Time-series Data

- Classic line chart

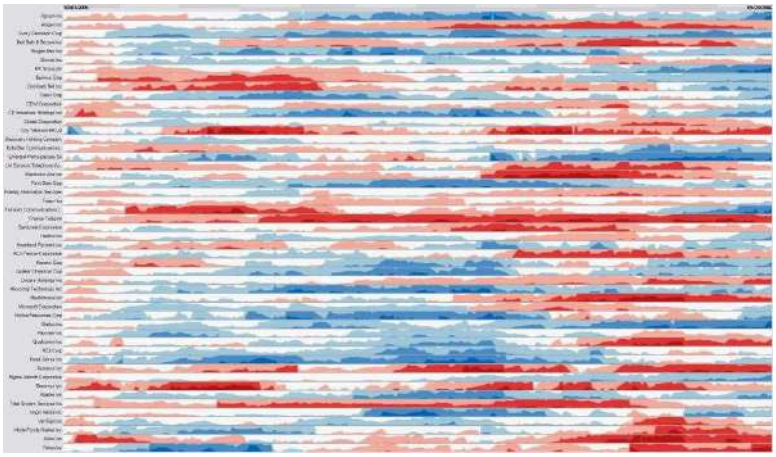
  - When there are massive time-series data, serious visual clutters will occur

29

29

Massive Time-series Data

- Horizon graph is designed to mitigate the visual clutter issues when there are massive time-series data



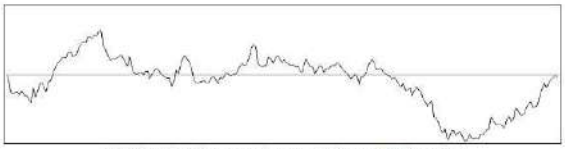
https://www.perceptualedge.com/blog/?p=390

30

30

Massive Time-series Data

- The Basic Construct of a Horizon Graph



(a) Standard line graph centered around a baseline.

(b) Color (blue is positive, red is negative) and layering.

(c) Mirroring around the baseline.

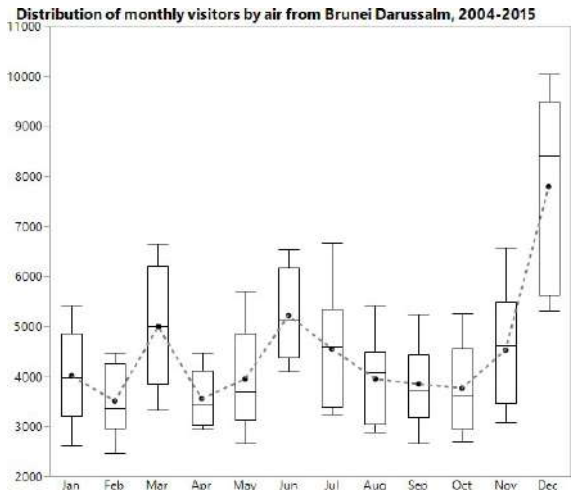
(d) Wrapping bands into a single space.

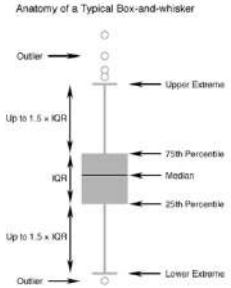
31

31

Distribution Over Time

- Possible Visualization: Boxplots with a trend line





32

32

Distribution Over Time

- Possible Visualization: Candlestick Charts

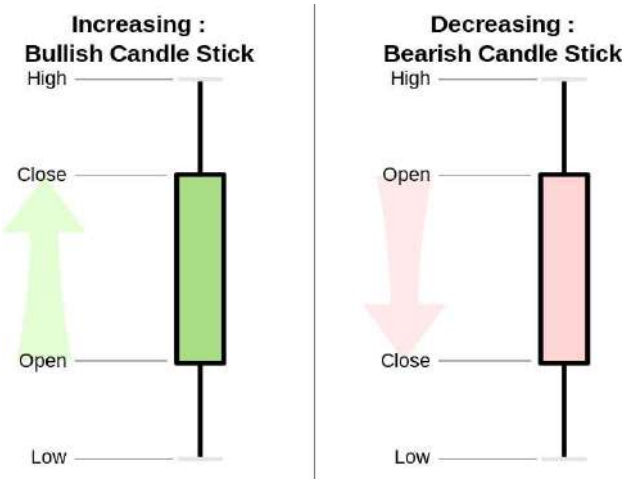
Source: https://patternswizard.com/how-to-read-stock-charts/

33

33

Distribution Over Time

- Interpreting Candlestick Chart



34

34

Distribution Over Time

- Candlestick in D3.js

Source: https://observablehq.com/@d3/candlestick-chart/2

35

35

Part-Whole Relationships

- To show values that combine to form a whole comparison to one another and the whole change through time

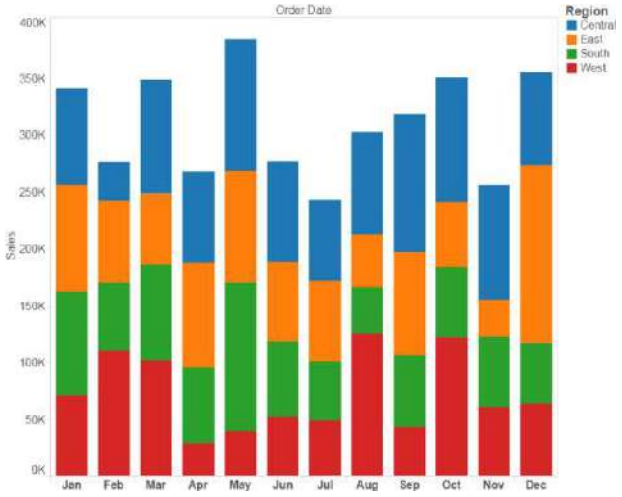
Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
East	64,162	70,172	93,657	55,056	63,631	53,040	51,974	63,272	54,110	112,284	69,697	27,437
Central	125,392	80,851	87,645	52,996	107,159	80,349	19,907	70,431	92,568	72,961	49,881	85,084
West	67,364	63,742	83,856	92,412	120,284	116,618	66,692	49,671	92,920	65,971	31,516	99,000
South	94,572	63,234	79,491	68,963	65,868	58,659	97,101	126,879	73,240	102,589	73,044	177,943
Total	$351,490	$277,998	$344,649	$269,425	$356,942	$308,666	$235,674	$310,254	$312,839	$353,805	$224,139	$389,465

36

36

Part-Whole Relationships

- Possible Visualization: stacked bar chart

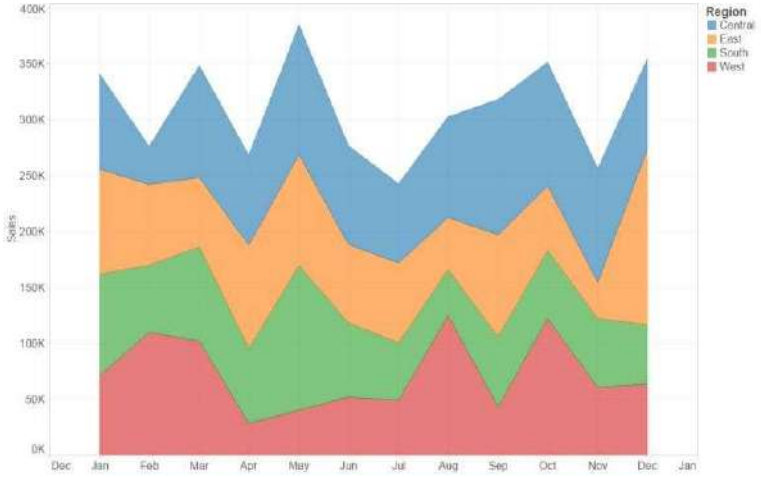


37

37

Part-Whole Relationships

- Possible Visualization: Stacked area chart

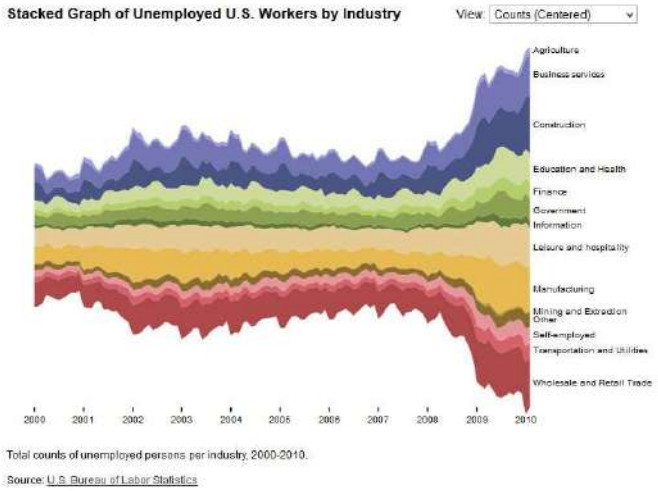


38

38

Part-Whole Relationships

- Possible Visualization: Stream Graph (Stacked Graph)



https://homes.cs.washington.edu/~jheer/files/zoo/ex/time/stack.html
https://www.d3-graph-gallery.com/streamgraph

39

39

What will you learn from this lesson?

- Time-series data
- Time representation
- Time-series data patterns
- Time-series data visualization with special requirements
- Interactive techniques for time-series data visualization

40

40

Interactivity

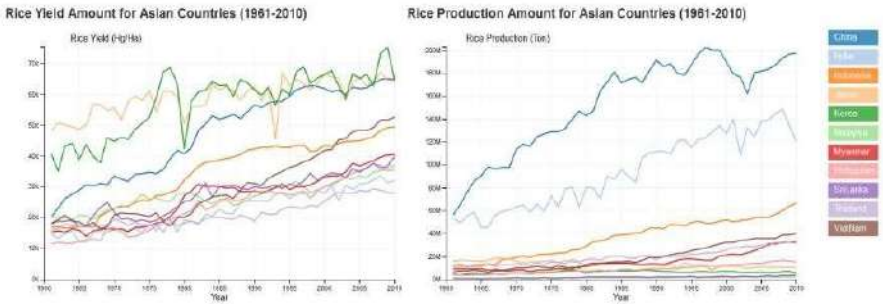
- Select: mark something as interesting
- Explore: show me something else
- Reconfigure: show me a different arrangement
- Encode: show me a different representation
- Abstract/Elaborate: show me more or less details
- Filter: show me something conditionally
- Connect: show me related items

41

41

Interactive Line Graph

- Example: Filtering



Link: http://tinyurl.com/ntfqocw

42

42

Animation: Visualizing the Change Over Time

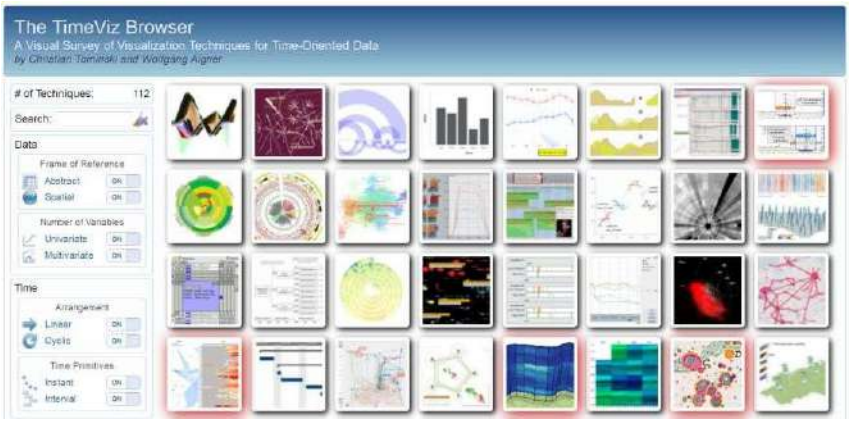
- Example: Animated bubble plot

Source: https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1

43

43

More Research on Time-series Data Visualization



Source: http://survey.timeviz.net/

44

44

Summary

- Time representation
  Linear, radial, grid, spiral, arbitrary
- Time-series data patterns
  Trend, rate of change, co-variation, cycles, exceptions
- Interactive techniques for time-series data visualization

45

45

Questions?

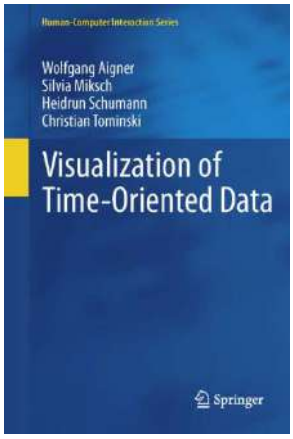
Thank You!

46

Reference

Visualization of Time-Oriented Data

Aigner, W., Miksch, S., Schumann, H., Tominski, C.Springer



47

47

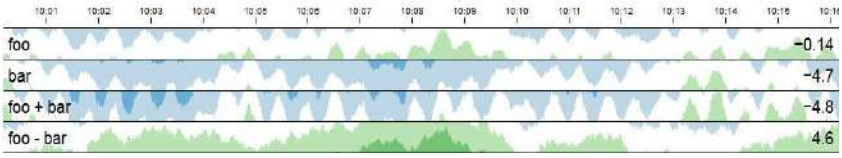
Visualizing the Change Over Time

- Animation + Interactivity: Horizon Graph using d3.js

Cubism.js

Time Series Visualization

https://square.github.io/cubism/



Tutorial: http://bost.ocks.org/mike/cubism/intro/#0

48

48

Reference

- Introduction to Cycle Plots (http://www.perceptualedge.com/articles/guests/intro_to_cycle_plots.pdf)
- Slopegraph
  - Slopegraphs for comparing gradients: Slopegraph theory and practice (http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0003nk)
  - In praise of slopegraphs (http://visualisingdata.com/index.php/2013/12/in-praise-of-_slopegraphs/)
  - Displaying Change Between Two Points in Time (http://www.perceptualedge.com/articles/visual_business_intelligence/di_splaying_change_between_two_points_in_time.pdf)
  - Edward Tufte's "Slopegraphs" (http://charliepark.org/slopegraphs/)
  - A Slopegraph Update (http://charliepark.org/a-slopegraph-update/)

49

49

Reference

- Slopegraphs are your Friend. Tables secretly hate you. (http://makeapowerfulpoint.com/2014/01/30/slopegraphs-are-your-friend-_tables-secretly-hate-you/)
- Communicating changes in rank over time: bumps charts and slopegraphs (http://datatodisplay.com/blog/chart-design/communicating-changes-rank-time/http://datatodisplay.com/blog/chart-design/communicating-changes-_rank-time/)
- Tableau Tip Tuesday: How to Create Slopegraphs (http://vizwiz.blogspot.sg/2015/01/slopegraphs.html)
- How to make a slope chart in Tableau (http://gravyanecdote.com/tableau/how-to-make-a-slope-chart-in-tableau/)
- Slopegraphs in Tableau (http://dataremixed.com/2013/12/slopegraphs-in-_tableau/)

50

50

Reference

- Horizon graph
  - The Development of the Horizon Graph (http://www.stonesc.com/Vis08_Workshop/DVD/Reijner_submission.pdf)
  - Time on the Horizon (http://www.perceptualedge.com/articles/visual_business_intelligence/time_on_the_horizon.pdf)
  - Horizon Graphs Revisited (http://www.perceptualedge.com/blog/?p=390)
  - Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations (http://vis.stanford.edu/files/2009-TimeSeries-CHI.pdf)
  - Interactive Horizon Graphs (http://hal.archives-ouvertes.fr/docs/00/78/13/90/PDF/IHG.pdf)
  - Horizon charts in Tableau (http://www.tableau.com/learn/tutorials/on-demand/horizon-charts)

51

51

Reference

- Animated Bubble Plot (Motion Chart)
  - Wiki (http://en.wikipedia.org/wiki/Motion_chart)
  - Visualizing Change: An Introduction in Time-Series Analysis (http://www.perceptualedge.com/articles/visual_business_intelligence/visualizing_change.pdf)

52

52

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 10
Geospatial Data Visualization

2

Outline

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

3

3

Geospatial Data

- Geospatial data (a.k.a, location-based data or geographic data) refers to data and information that has explicit or implicit association with a location relative to Earth.

XCOORD	YCOORD	NAME
103.8589	1.2840	Livewire (Marina Bay Sands)
103.8574	1.2947	Singapore Pools Suntec City
103.8469	1.2835	Kis Store
103.8493	1.2829	Singapore Turf Club China Square
103.8521	1.2836	Tay Sui Lan Agency
103.8492	1.2861	7-Eleven Circular Rd
103.8440	1.2858	Singapore Pools People's Park Centre
103.8464	1.2850	Feng Yuen Agency Enterprise

SUBZONE_C	TOTAL_POP	YOUNG
AMSZ01	4990	850
AMSZ02	30390	3950
AMSZ03	28650	3410
AMSZ04	24160	3260
AMSZ05	19190	2620
AMSZ06	24550	3050
AMSZ07	6860	1030
AMSZ08	8370	1200

4

4

Geospatial Data

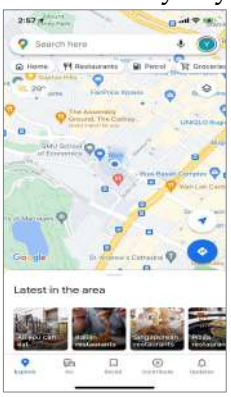
- Geospatial data (a.k.a, location-based data or geographic data) refers to data and information that has explicit or implicit association with a location relative to Earth.
- Geospatial data can be sourced from telematics devices, Global Positioning System (GPS) data, geospatial satellite imagery, Internet of Things, geotagging, etc.

5

5

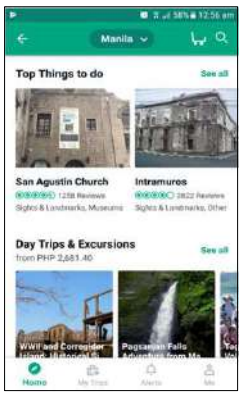
Geospatial Data Example: Location-based Services

- Geospatial data is highly relevant to our everyday life
- Geospatial data is also generated by mobile devices and sensors everyday



Google Maps

Yelp



Tripadvisor

6

6

Outline

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

7

7

What is Map?

- "The map speaks across the barriers of language; it is sometimes claimed as the language of geography. "

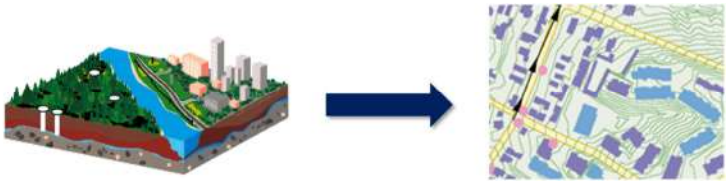
-- Carl Ortwin Sauer, 1956

8

8

What is Map?

- A map is a model of real world depicted by a collection of cartographic symbols or/and visual abstraction.

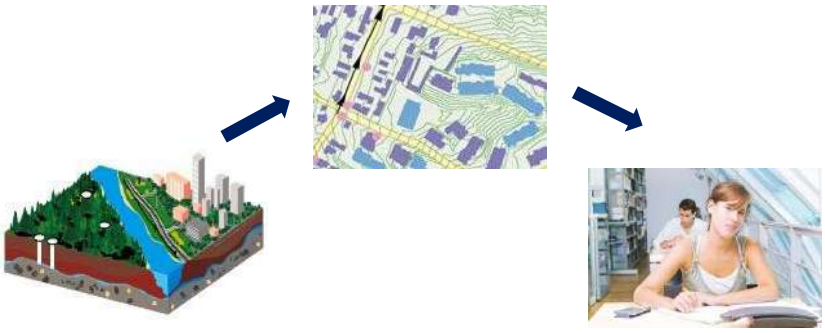


9

9

What is Map?

- An interface between geospatial data and map users.



10

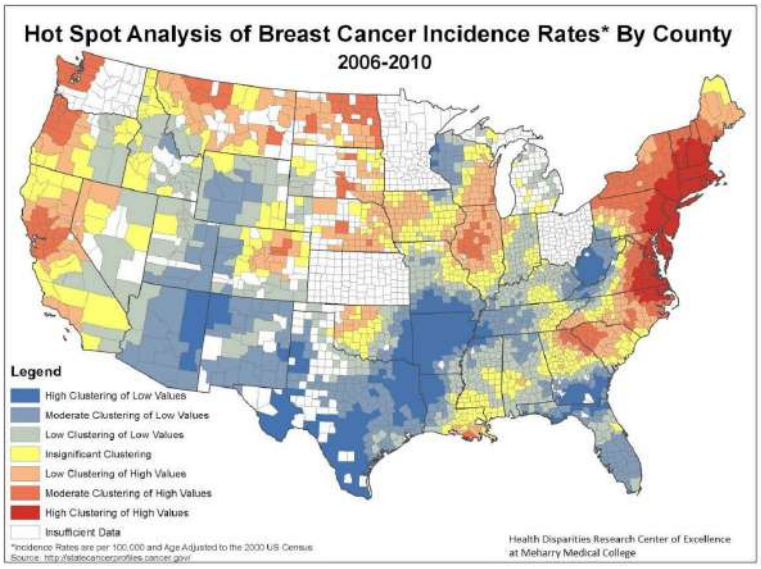
10

Map

11

11

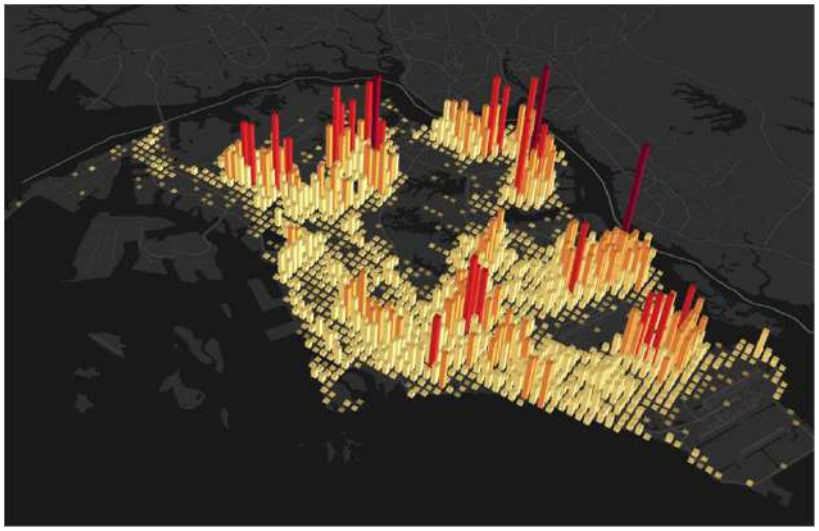
Map



12

12

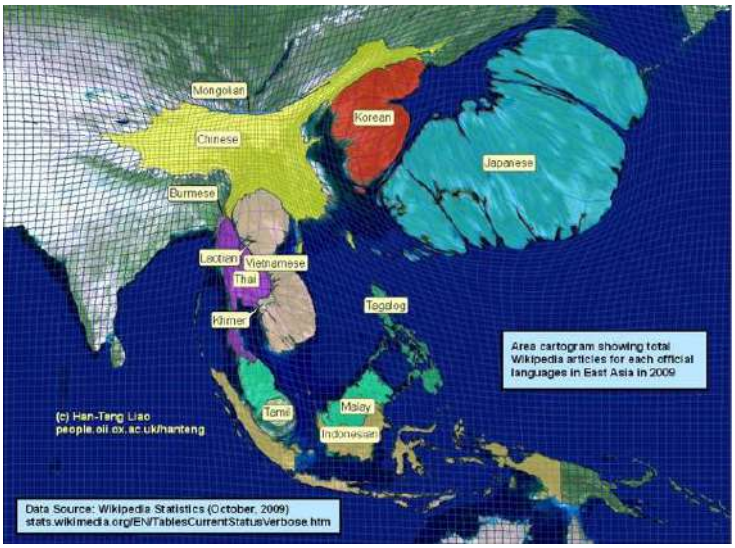
Map



13

13

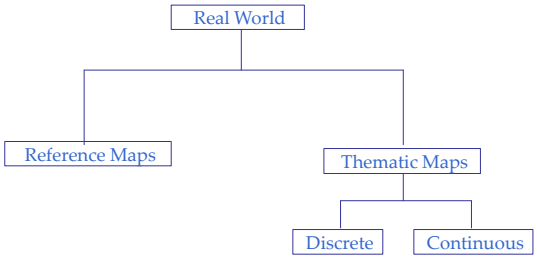
Map



14

14

Taxonomy of Maps



Slocum, T. A., McMaster, R. B., Kessler, F. C., and Howard, H. H. (2008) *Thematic Cartography and Geovisualization, 3rd Edition.*

15

15

Reference Maps

- A.k.a., General-Purpose Map
- Emphasize the location of spatial phenomena, such buildings, roads, vegetation, rivers, etc.
- Some of popular reference maps: topographical map and internet map such as Google map

(Source: US Geological Survey)

16

16

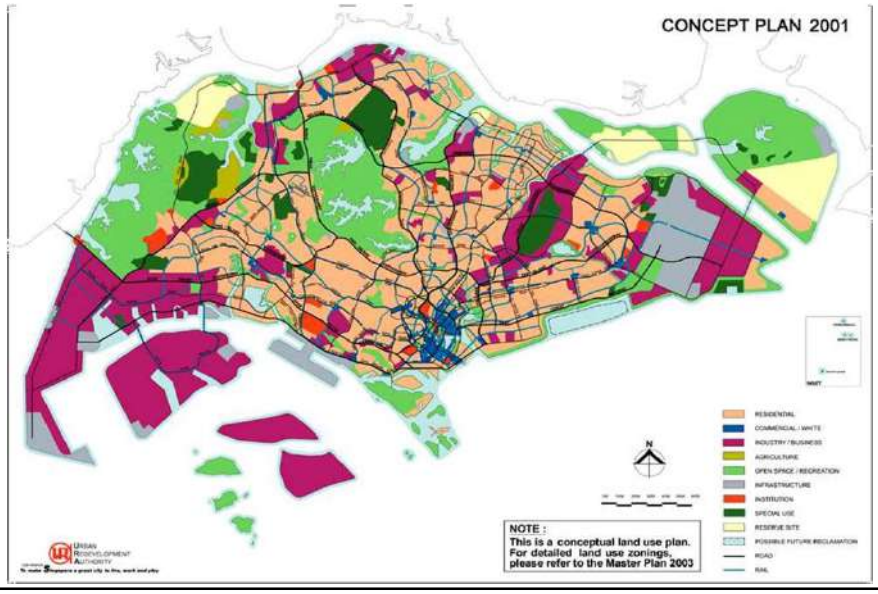
Thematic Maps

- A.k.a., Statistical Map
- A thematic map is used to display the spatial pattern of **a theme or attribute**
- Starting from 19th century
- Some popular thematic maps: Choropleth map, Cartogram map, etc.

17

17

Thematic Map: Land Use Map



18

18

Outline

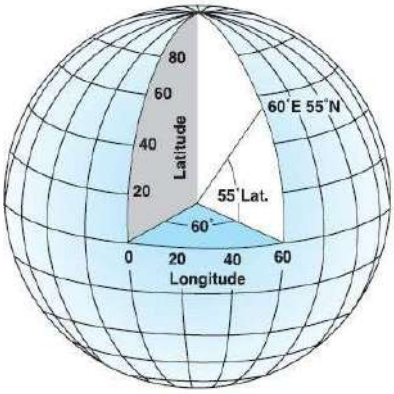
- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

19

19

Geographic Coordinate Systems

- Define the locations on the earth using a 3D spherical surface



Latitude-Longitude

https://desktop.arcgis.com/en/arcmap/10.3/guide-books/map-projections/about-geospatial-coordinate-systems.htm
https://www.youtube.com/watch?v=2PllX2YOAHo

20

20

Map Projection

- Represent the **surface** of a sphere or other 3D body on a plane
- Necessary for creating maps

Wikipedia, Map projection.
https://en.wikipedia.org/wiki/Map_projection

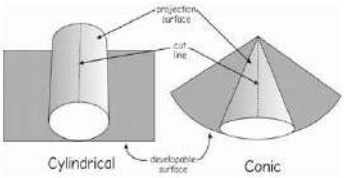
21

21

Map Projection

From a Sphere to a Plane

- Hundreds of different map projections have been devised since the early Babylonian stone-inscribed world map[1] which was created in early 600 BC.
- It is generally acknowledged that most maps fall into **three** broad families of **map projection**[2].
- These map projections are based on the **developable surfaces** of the **cylinder**, **cone** and **plane**.



- Which surface is best will depend on the **region** to be mapped, its geographical **extent** and the desired **geometric properties** of the map[2].

An ancient Babylonian map of the Mesopotamian world (700-500 BC)
Image from [2]

[1] World History Encyclopedia, Babylonian Map of the World - https://www.worldhistory.org/image/526/babylonian-map-of-the-world/
[2] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html

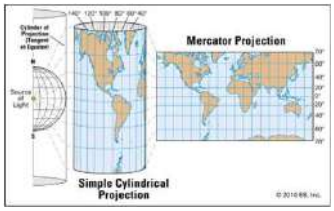
22

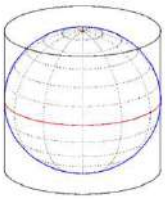
22

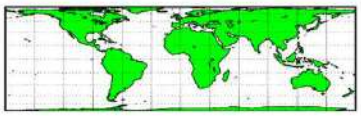
Map Projection

Cylindrical Projection

- Cylindrical projection is produced by wrapping a cylinder around a globe (Earth). The map projection is the image of the globe projected onto the cylindrical surface, which is then unwrapped into a flat surface.
- Cylindrical projections can be either **equal-area** (preserves area but distorts shape), **conformal** (preserves angle/shape but distorts area and distance), or equidistant (preserve distance).







Mercator conformal cylindrical projection

Image from [1]

Equal-area cylindrical projection

Image from [2]

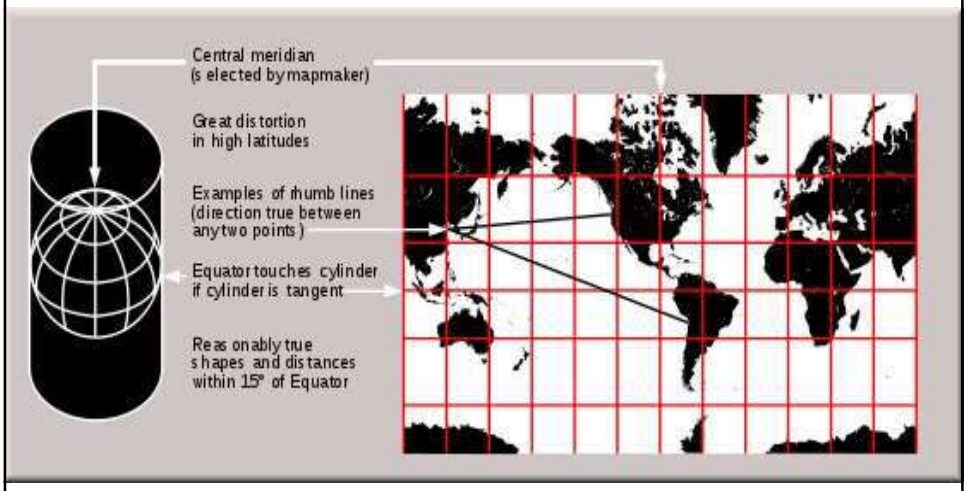
[1] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html

[2] Britannica, Mercator Projections - https://www.britannica.com/science/Mercator-projection

23

23

Example: Mercator Projection



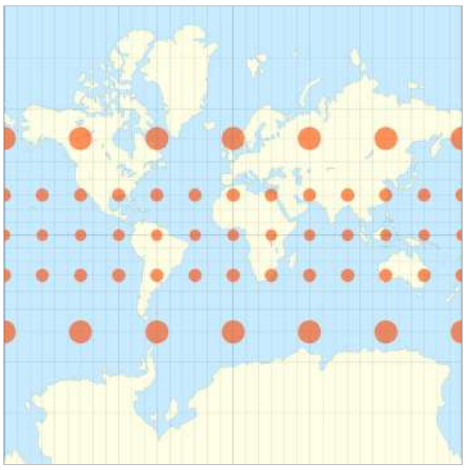
A cylindrical map projection

24

24

Example: Mercator Projection

Distorted Area Size



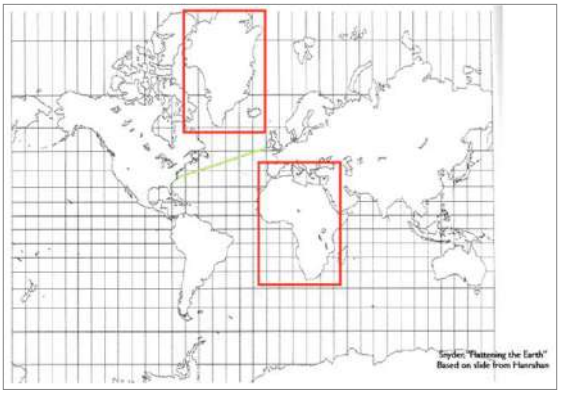
Wikipedia, Mercator projection.
https://en.wikipedia.org/wiki/Mercator_projection

25

25

Example: Mercator Projection

Distorted Area Size



26

26

Example: Mercator Projection

https://www.youtube.com/watch?v=8I_VpC6IuJs
https://www.youtube.com/watch?v=kIID5FDi2JQ

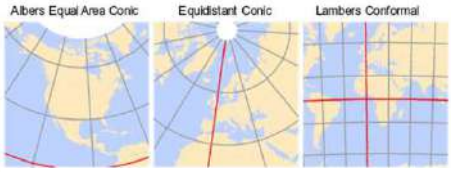
27

27

Map Projection

Conic Projection

- A conic projection is derived from the projection of the globe onto a **cone** placed over it. In the normal aspect, the apex of the cone lies on the polar axis of the Earth.
- Conic projections often achieve less distortion at mid and high latitudes than cylindrical projections[1].



3 conical projections examples: **Albers equal area** (preserves area), **equidistant** (preserves distance) and **conformal** (preserves shape) [2]

Albers Equal-Area projection, polar aspect

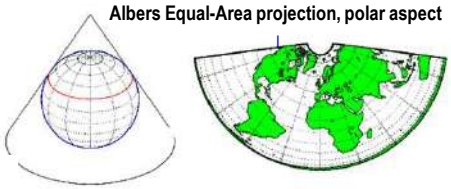


Image from [1]

[1] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html
[2] Conical Projections - https://mgimond.github.io/Spatial/chp09-0.html

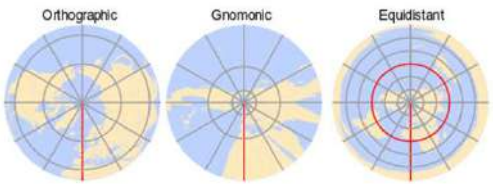
28

28

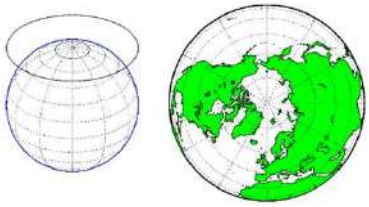
Map Projection

Azimuthal Projection

- An azimuthal projection is a projection of the globe onto a plane.
- Most azimuthal projections are not suitable for displaying the entire Earth in one view, but give a sense of the spherical globe[1].



3 planar projections: **orthographic**, **gnomonic** and **equidistant**. Each covers a different spatial range (with the latter covering both northern and southern hemispheres) and each preserves a unique set of spatial properties.) [2]



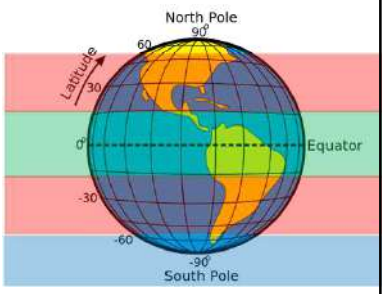
Orthographic projection, polar aspect,
Image from [1]

[1] Mathworks, The Three Main Families of Map Projections - https://www.mathworks.com/help/map/the-three-main-families-of-map-projections.html
[2] Planar Projections - https://mgimond.github.io/Spatial/chp09-0.html

Map Projection

Which Projection Should I Choose for a Region?

- The **latitude rule**[1] is a good rule of thumb in making a decision on which map projection to use for a given geographical region.
- To map **tropical regions** (e.g. Singapore, India), use **cylindrical** projection (e.g. Mercator).
- To map **middle latitude** (e.g. USA, China, New Zealand), use **conic** projection (e.g. Albers Equal-Area, Equidistant Conic).
- To map **polar regions** (e.g. North Pole), use **azimuthal** projection (e.g. Azimuthal Equidistant).



[1] Jochen Albrecht (Hunter College, City University of New York, CUNY), Choosing a projection - http://www.geo.hunter.cuny.edu/~jochen/gtech201/lectures/lec6concepts/map%20coordinate%20systems/how%20to%20choose%20a%20projection.htm

Map Projection

What is the Purpose of the Map?

- The **purpose** of your geospatial analysis also influences the map projection choice. Map projections are not 100% accurate. Each projection distorts the surface of the Earth in some ways, while retaining some useful properties.
- The **equal-area** projections (e.g. "Lambert Cylindrical Equal Area", "Albers Equal Area Conic") **preserve area**[1]. This is a good choice, if you want to **calculate the area** of a country or city or plot a **dot density map** for each region.
- The **equidistant** projections (e.g. "Azimuthal Equidistant projection") **preserve distance**[2] and would be suitable if you want to analyse flight distances or shipping routes.
- The **conformal** projection is useful when the map's main purpose involves **measuring angles**, showing accurate **local directions**, or representing the **shapes** of features or **contour lines** (e.g. in weather maps, civil engineering & military maps)[1].

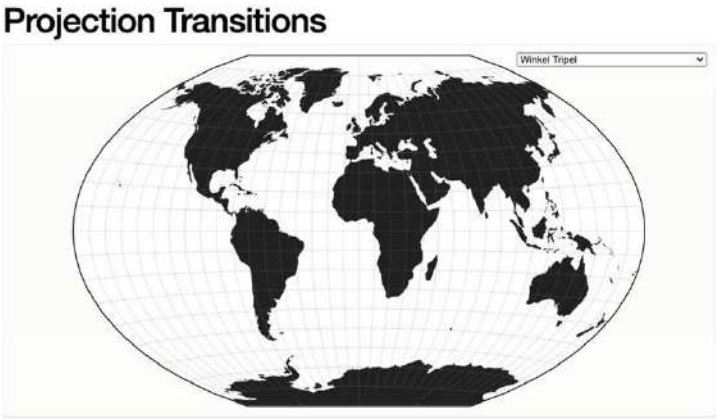
[1] Jochen Albrecht (Hunter College, City University of New York, CUNY), Choosing a projection - http://www.geo.hunter.cuny.edu/~jochen/gtech201/lectures/lec6concepts/map%20coordinate%20systems/how%20to%20choose%20a%20projection.htm
[2] Alexis Cook, Coordinate Reference Systems- https://www.kaggle.com/alexisbcook/coordinate-reference-systems

31

31

Demo of More Map Projections

- https://observablehq.com/@d3/projection-transitions



32

32

Outline

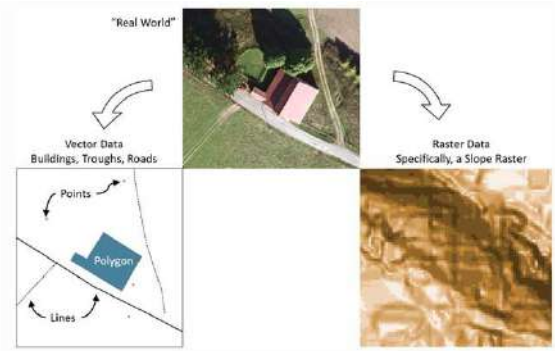
- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

33

33

Basic Geospatial Data Models

- Vector data model represent features as discrete **points, lines and polygons**, and is widely used in cartography, and network analysis
- Raster data model represent the continuous data as rectangular matrix of square cells (i.e., pixels), and is useful as background maps and for spatial analysis



Data model definitions: https://gsp.humboldt.edu/olm/Lessons/GIS/04%20CreatingSpatialData/VectorDataModels1.html

34

34

Vector Data Model

- Used to represent **points, lines, and areas**

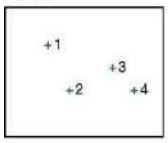
- All are represented using coordinates
  - One per **point**
  - Lines as **polylines**
    - Straight lines between points
  - Areas as **polygons**
    - Straight lines between points, connecting back to the start
    - Point locations recorded as coordinates

35

35

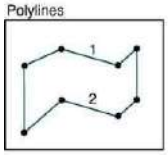
Vector Data Model

Points



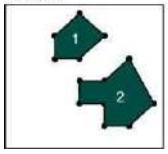
Point number	(x,y) coordinates
1	(2,4)
2	(3,2)
3	(5,3)
4	(6,2)

Polylines



Polyline number	(x,y) coordinates
1	(1,5) (3,6) (6,5) (7,6)
2	(1,1) (3,3) (6,2) (7,3)

Polygons



Polygon number	(x,y) coordinates
1	(2,4) (2,5) (3,6) (4,5) (3,4) (2,4)
2	(3,2) (3,3) (4,3) (5,4) (6,2) (5,1) (4,1) (4,2) (3,2)

36

36

Raster Data Model

- All raster formats are basically the same
  - Cells organized in a matrix of rows and columns
  - Content is more important than format: data or picture?

Raster data

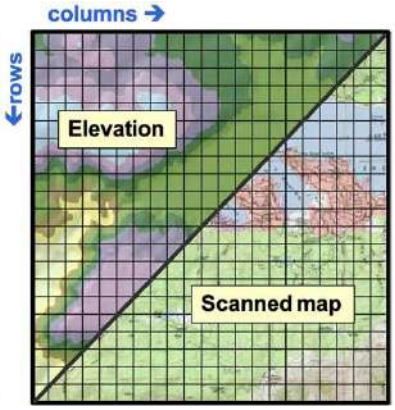
- Elevation
- Landuse codes
- Population density

Good for analysis

- Slope from elevation

Good for mapping

- Thematic layers
- Derivative products (like shaded relief)



Raster pictures

- Scanned maps
- Satellite images (unclassified)
- Photos of buildings

Good for mapping

- Backgrounds

Good for attributes

- Picture of house

Bad for analysis

37

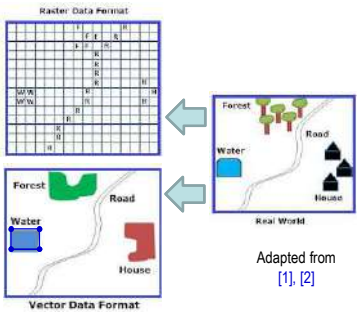
37

Geospatial Data File Formats

Basic types

- There are a variety of geospatial data file formats that are widely used by existing Geographical Information Systems (GIS). The two basic types are:

- **Raster GIS file** – These files contain raster data that are made up of **pixels** (i.e. grid cells of data values). They are usually regularly spaced and square but need not always be so[1].
- **Vector GIS file** – These files contain **vector graphics** that comprise of data that describes individual points (e.g. landmarks), vertices of polygons (e.g. boundaries), lines (e.g. roads or paths) and other geometric entities[2].



Adapted from [1], [2]

38

[1] What are the Raster Data Formats in GIS and Remote Sensing (2020) - https://www.gisoutlook.com/2020/04/raster-data-formats-in-gis.html
[2] What is Vector Data Model of GIS (2020) - https://www.gisoutlook.com/2020/04/vector-data-model-of-gis.html

38

Outline

- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

39

39

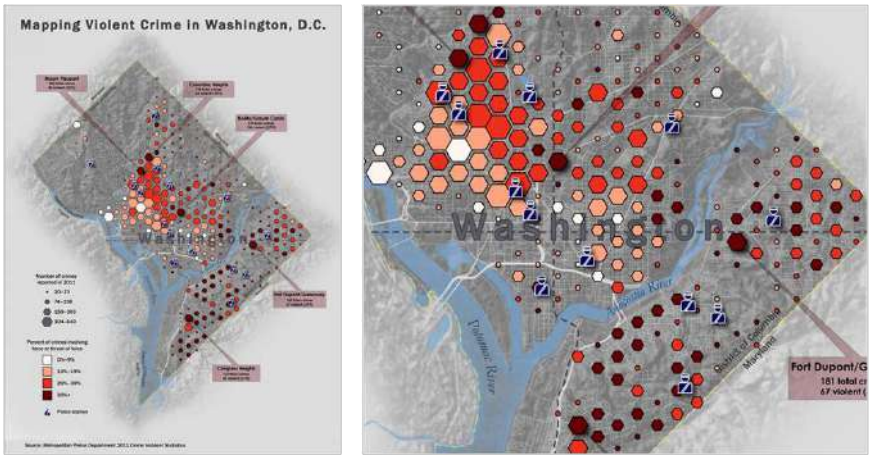
Point Data

- Points scattered in the geographic space, with longitude and latitude information, optionally with weights
  - Landmarks on the map
  - Restaurants of your neighborhood
- Marking the data on the map would be the most direct and simplest way to visualize point data
- Point-based visualizations are widely used

40

40

Symbol Map

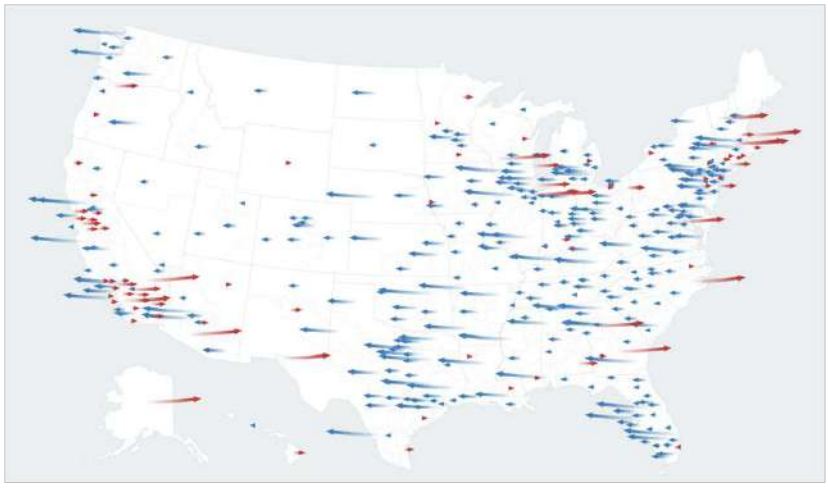


Cooper Thomas, Violent Crime in Washington, D.C.
https://visual.ly/community/infographic/geography/violent-crime-washington-dc

41

41

Symbol Map

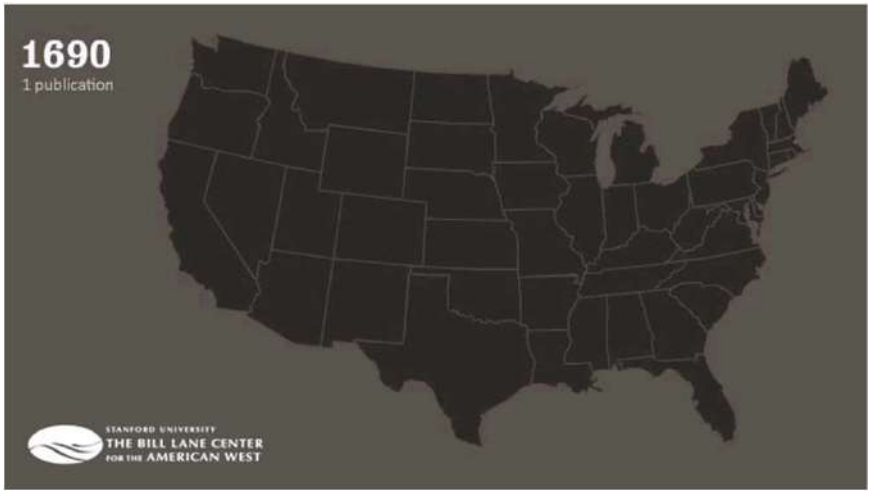


The New York Times, How Democrats Took the House.
https://www.nytimes.com/interactive/2018/11/07/us/politics/how-democrats-took-the-house.html

42

42

Symbol Map



Rural West Initiative, Mapping Journalism's Voyage West.
https://web.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers

43

43

Point Distribution Map

BBC, Every death on every road in Great Britain 1999-2010.
https://www.bbc.co.uk/news/uk-15975724

44

44

Outline

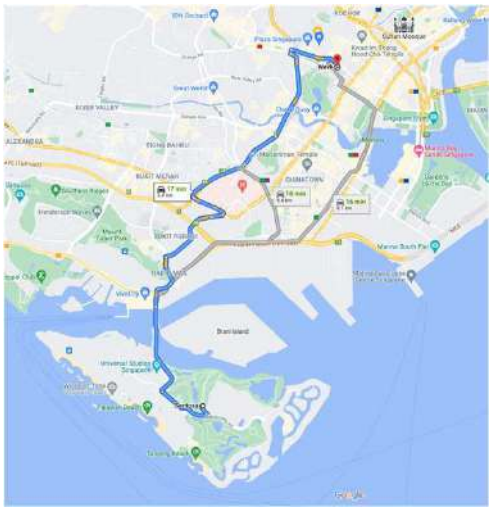
- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

45

45

Line Data

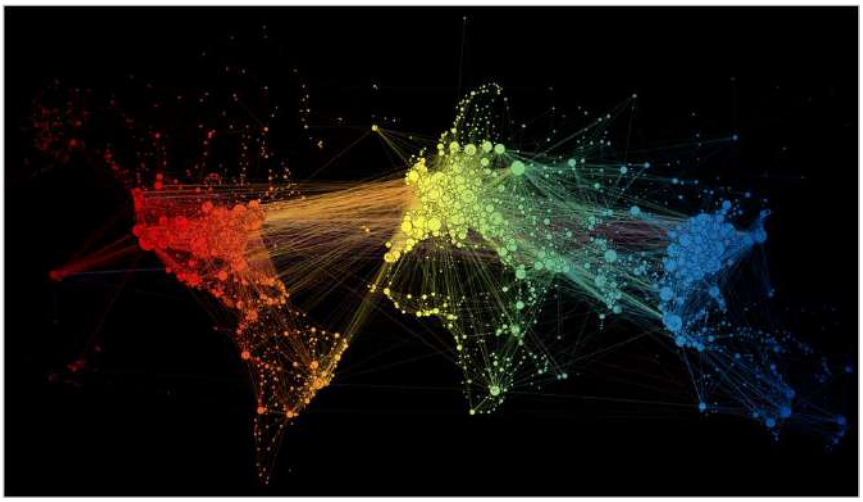
- Line data comprise locations and the paths between them
- Each line has a length property, namely the distance between two locations



46

46

Air Traffic Network



Martin Grandjean, Connected World: Untangling the Air Traffic Network.
http://www.martingrandjean.ch/connected-world-air-traffic-network/

47

47

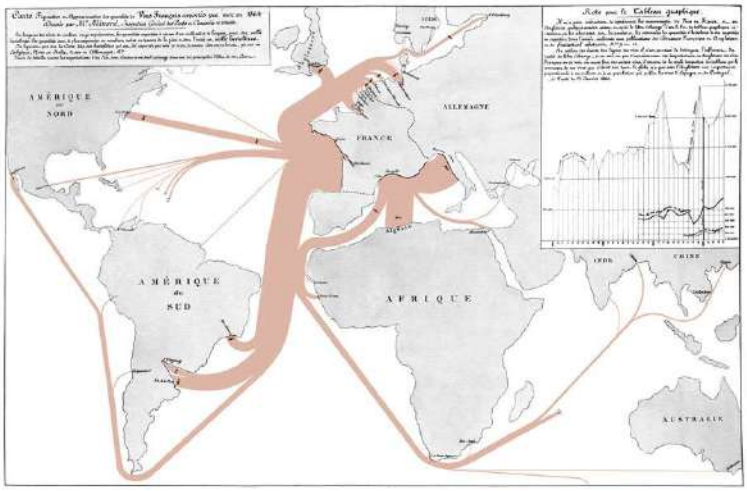
Facebook Friend Relationship

Facebook, 2013.
https://bit.ly/2P7WZep

48

48

Bundling Lines

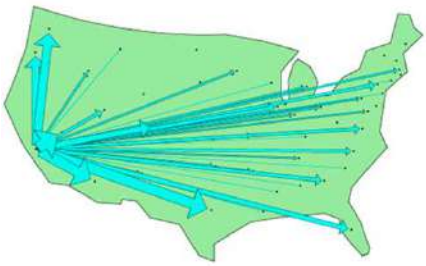


Charles Joseph Minard, French wine exports,1864.
https://en.wikipedia.org/wiki/Flow_map

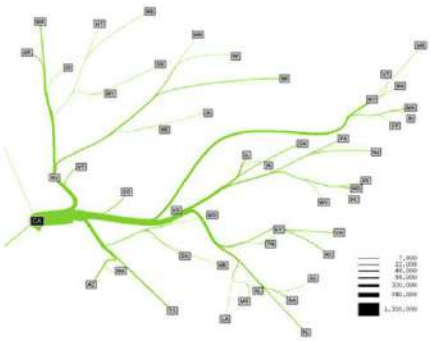
49

49

Flow Map



Linking the immigrations from California to other states with arrows



Visualized with Flow Map

Buchin, Kevin, Bettina Speckmann, and Kevin Verbeek. "Flow map layout via spiral trees." IEEE transactions on visualization and computer graphics 17.12 (2011): 2536-2544.

50

50

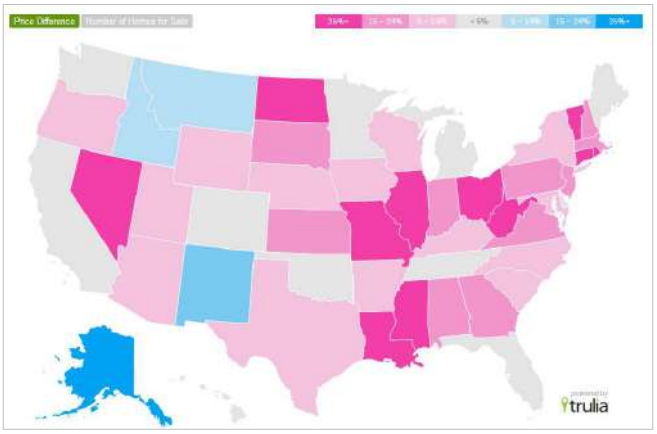
Outline

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

51

51

Selling Real Estate: Men vs. Women

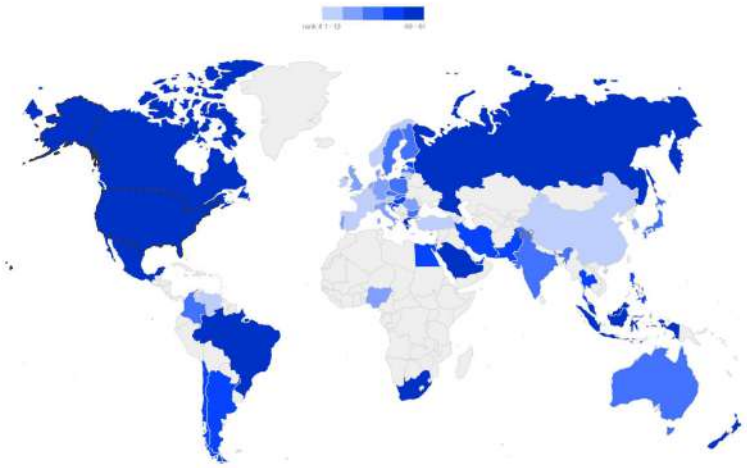


Trulia, Is real estate a man's or woman's world.
https://bit.ly/2KIh9LE

52

52

The World Gas Prices



Gasoline Prices Around the World
https://www.bloomberg.com/graphics/gas-prices/#20183:United-States:USD:g

53

53

Choropleth Map

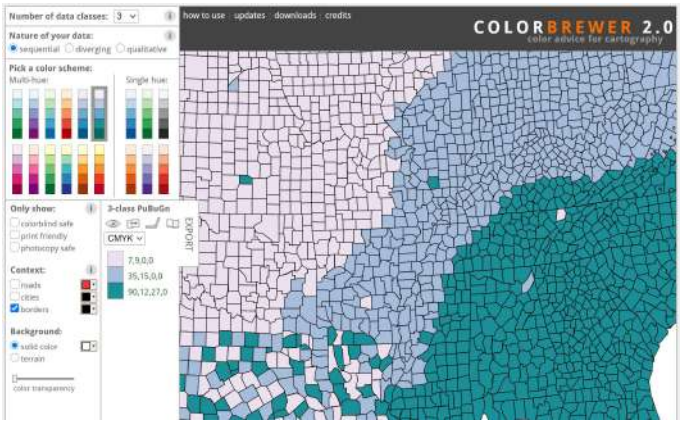
- A thematic map in which a set of pre-defined areas (e.g., states or countries) are shaded or patterned in proportion to the measurement of the statistical variable within different regions
- Also known as, filled map
- Assume that the data is evenly distributed in the area
- Often represent the scales of data with **colors**

54

54

Choropleth Map

- Colour scheme tool: ColorBrewer

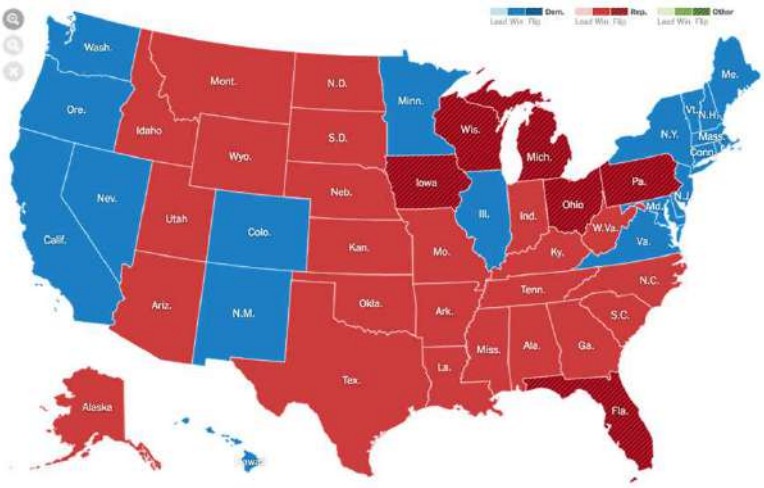


http://colorbrewer2.org

55

55

Choropleth Map



The New York Times, 2016 Presidential Election Results.
https://www.nytimes.com/elections/2016/results/president

56

56

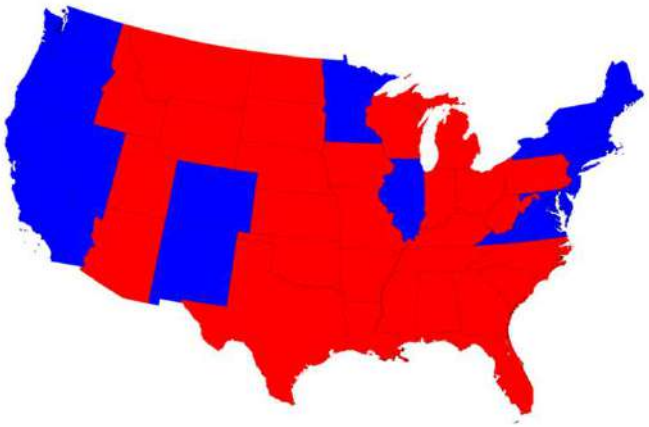
Choropleth Map Can Be Misleading

- The map fails to take account of **the population distribution**
  - It fails to allow for the fact that the population of the red states is on average significantly lower than that of the blue ones. The blue may be small in area, but they represent a large number of voters, which is what matters in an election
- When data is accumulated at the place where display space is less, there comes a mismatch, and vice versa
- This mismatch would very possibly **mislead** the audience

57

57

Choropleth Map Can Be Misleading



Mark Newman, 2016 Maps of the 2016 US presidential election results.
http://www-personal.umich.edu/~mejn/election/2016/

58

58

Cartogram Map

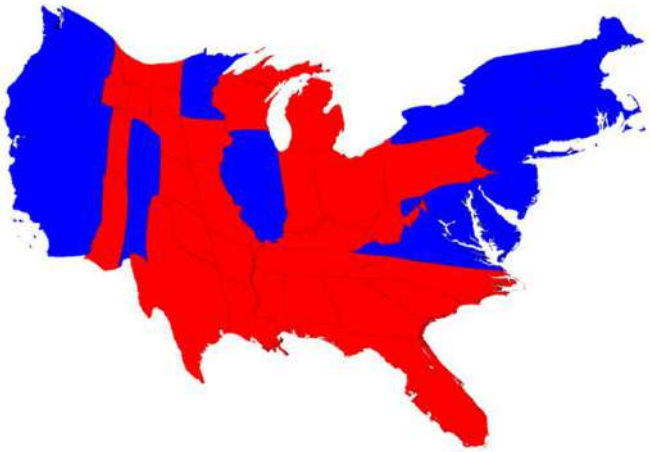
- A map in which some thematic mapping variable – **such as travel time, population, or Gross National Product** – is substituted for land area or distance.

Mark Newman, 2016 Maps of the 2016 US presidential election results.
http://www-personal.umich.edu/~mejn/election/2016/

59

59

Cartogram Map



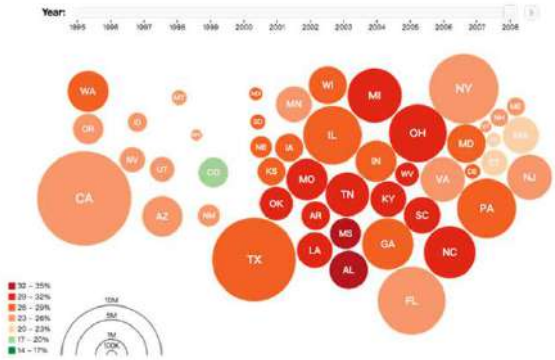
Mark Newman, 2016 Maps of the 2016 US presidential election results.
http://www-personal.umich.edu/~mejn/election/2016/

60

60

Other Maps: Dorling Cartogram

- Cartograms distort the shape of geographic regions so that the area directly encodes a data variable. A common example is to resize countries proportional to population or GDP.
- **Dorling cartogram** represents each geographic region as non-overlapping circles.



Mike Bostock , Obese People Per State.
http://mbostock.github.io/protovis/ex/cartogram.html

61

61

Outline

- What is geospatial data
- What is map?
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

62

62

Geospatial Datasets

OpenStreetMap
https://www.openstreetmap.org/

HK Government Datasets
https://data.gov.hk

63

63

Geospatial Datasets

- Google/Baidu API
- OpenStreetMap (http://www.openstreetmap.org)
- HK government dataset (http://data.gov.hk)
- NYC Open Data (https://opendata.cityofnewyork.us)
- Data.gov (https://www.data.gov)

64

64

Common Geospatial Data Visualization Packages

- JavaScript
  - D3: d3.geo (https://d3js.org/d3-geo)
  - Leaflet (https://leafletjs.com/)
  - …

- Python
  - Geoplot (https://residentmario.github.io/geoplot)
  - GeoPandas (https://geopandas.org/en/stable/)
  - …

65

65

Summary

- What is geospatial data
- What is map
- Map projection
- Visualization of geospatial data
  - Point data visualization
  - Line data visualization
  - Region data visualization
- Geospatial datasets

66

66

Summary

- Geospatial data refers to data and information that has explicit or implicit association with a location, and has been widely seen in various applications
- Map projection converts the **surface** of a sphere or other 3D body on a 2D plane. Each map projection method is a trade-off of different criteria and there is no optimal map projection for all situations
- Geospatial data models include vector and raster data models
- Geospatial data visualization techniques mainly cover three types: Point data visualization, line data visualization and region data visualization

67

67

Questions?

Thank You!

68

Interesting examples of cartogram

- Cartogram of the world's languages
  http://th-mayer.de/cartogram/

- China Still Dominates, but Some Manufacturers Look Elsewhere
  https://archive.nytimes.com/www.nytimes.com/interactive/2013/04/08/business/global/asia-map.html

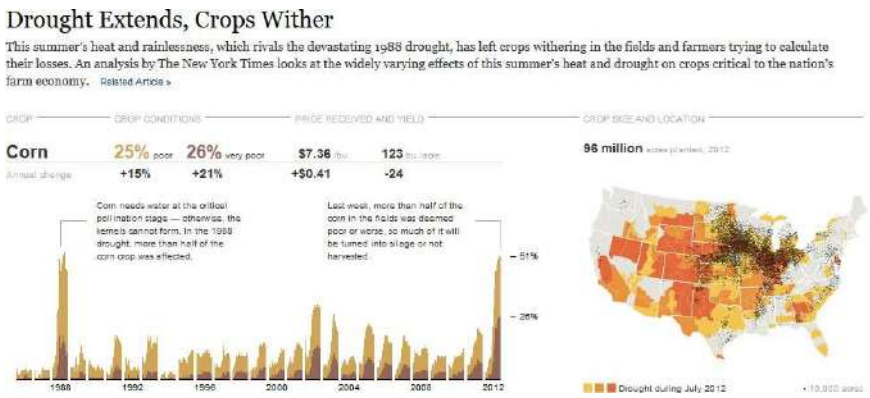
- Dorling Cartogram of Obesity in the United States (1995 to 2008)

https://homes.cs.washington.edu/~jheer/files/zoo/ex/maps/cartogram.html

69

69

Example: HTML5 + JavaScript based GeoVisual Analytics



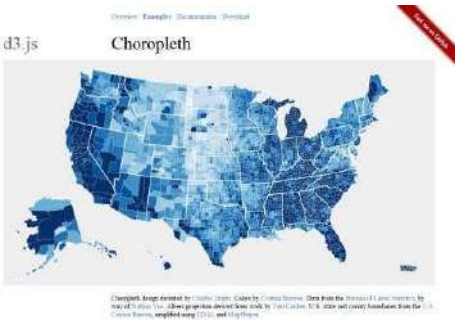
http://www.nytimes.com/interactive/2012/08/24/us/drought-crops.html

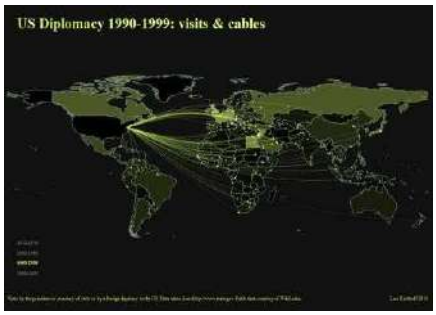
70

70

d3.js for Visualizing Geographical Data

- d3.geo (Geography) (https://github.com/mbostock/d3/wiki/Geo)





71

71

D3 Tutorial

- Let's Make a Map
  https://bost.ocks.org/mike/map/
- How to Make Choropleth Maps in D3
  https://rockcontent.com/blog/how-to-make-choropleth-maps-in-d3/
- A Map to Perfection: Using D3.js to Make Beautiful Web Maps
  https://www.toptal.com/javascript/a-map-to-perfection-using-d3-js-to-make-beautiful-web-maps

72

72

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 11.1
Hierarchical Data (Tree) Visualization

2

Outline

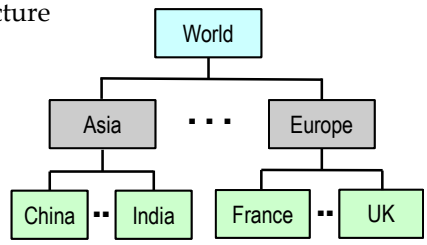
- What is hierarchical data?
- What are the hierarchical data visualization (a.k.a., Tree Visualization) methods?
  - Node-link (tree) diagrams
  - Indented trees
  - Layered (adjacency) diagrams
  - Enclosure diagrams

3

3

Hierarchical Data

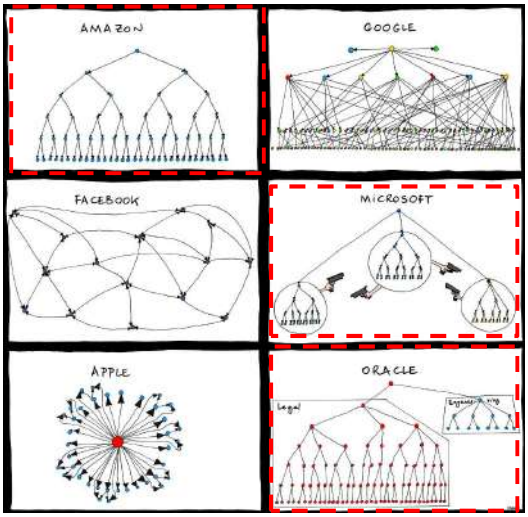
- Hierarchical data represent the **hierarchical relationship** between objects/entities, where **the attributes** associated with objects/entities are often considered as well.
- As an example, **countries** (e.g. China, India, France, UK, etc) are a subset of continents (e.g. Asia, Europe, etc).
- More Examples:
  - Hierarchical organizational structure
  - Information organization
    - File directories
    - Development of species
  - Logical connections
    - Decision tree



4

4

Hierarchical Data: Organizational Chart of IT Companies



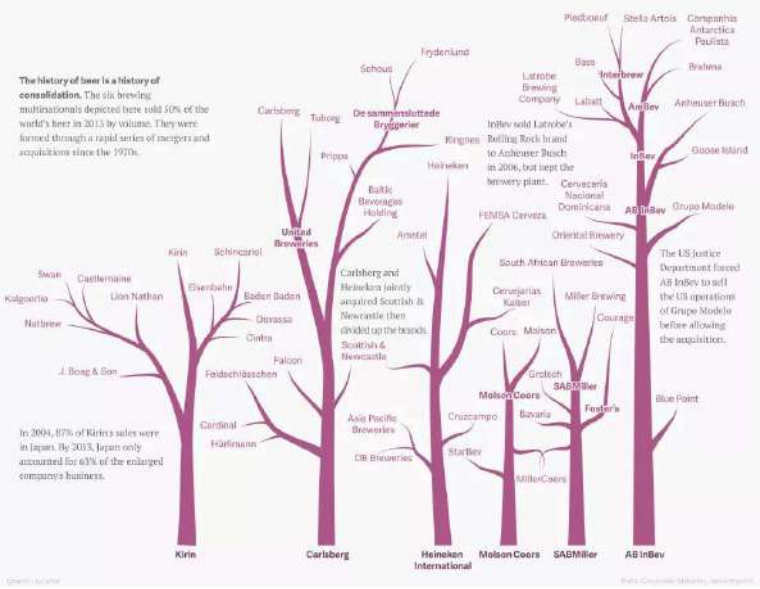
Hierarchical data are intrinsically in a **tree-like** structure.

https://ritholtz.com/2013/07/organizational-charts-of-amazon-apple-facebook-microsoft/

5

5

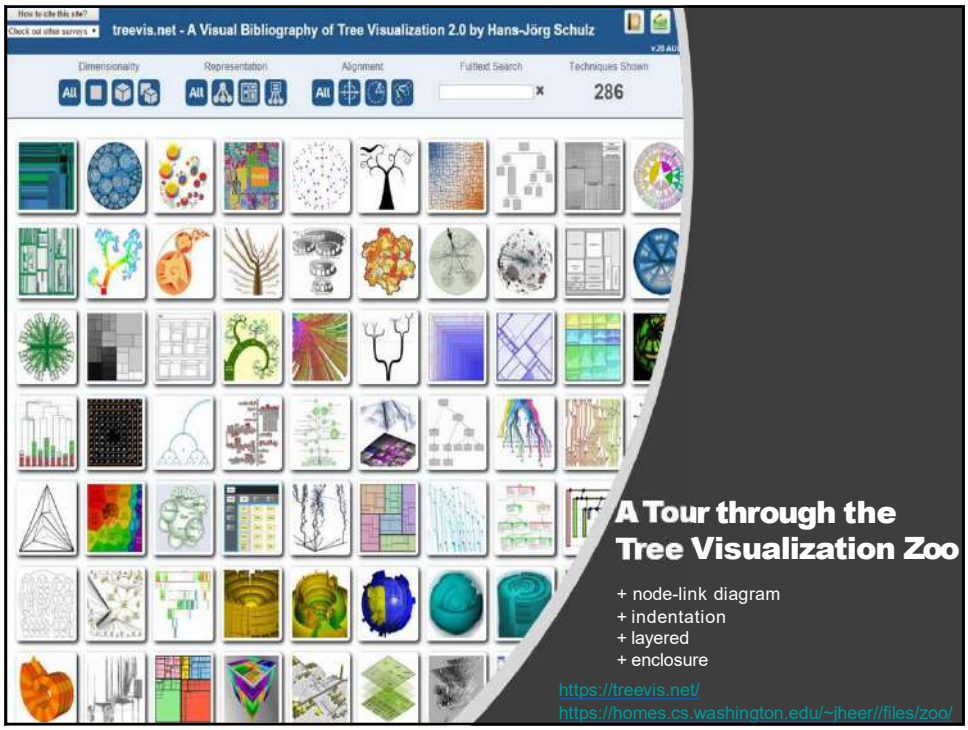
Hierarchical Data: Beer Family Tree



https://flowingdata.com/2014/10/02/beer-family-tree/

6

6



7

Hierarchical Data Visualization

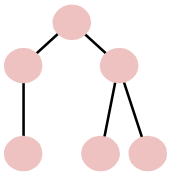
- Node-link (tree) diagrams
- Indented trees
- Layered (adjacency) diagrams
- Enclosure diagrams
  - Treemap
  - Circle packing layout

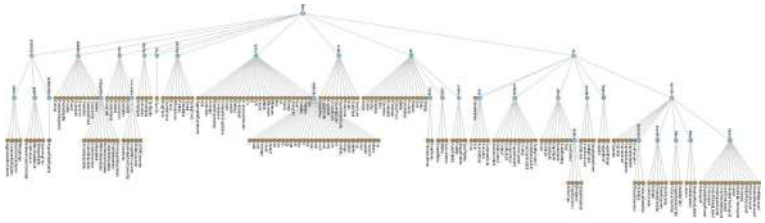
8

8

Node-link Tree Diagrams

- Nodes are distributed in space, connected by straight or curved lines
- Typical approach is to use 2D space to break apart breadth and depth
- Often, space is used to communicate hierarchical orientation





9

9

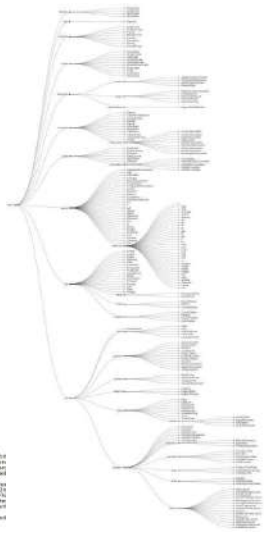
Node-link Tree Diagrams

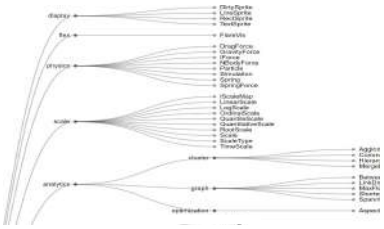
- Tidy Tree

  - Implements the Reingold-Tilford algorithm for efficient, tidy arrangement of layered nodes

  - Depth of nodes computed by distance from the root

  - Integrated into D3





https://observablehq.com/@d3/tree-component

10

10

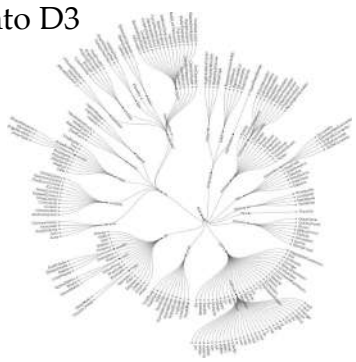
Node-link Tree Diagrams

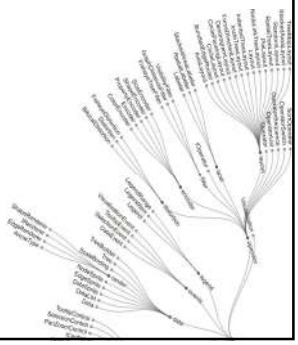
- Radial Tidy Tree

  - A radial layout is based on a polar rather than Cartesian coordinates

  - It has a pleasing aesthetic, and uses space more efficiently

  - Integrated into D3





https://observablehq.com/@d3/radial-tree-component

11

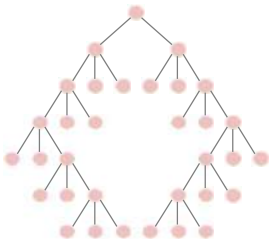
Node-link Tree Diagrams

- Reingold-Tilford algorithm for drawing node-link diagrams

  - Bottom-up recursive approach: Repeatedly divide space by leaf count

  - For each parent, make sure subtrees are drawn

  - Make smarter use of space
    + Maximize density and symmetry
    + Clearly encode depth level
    + No edge crossings
    + Pack subtrees as closely as possible
    + Centers parent over subtrees



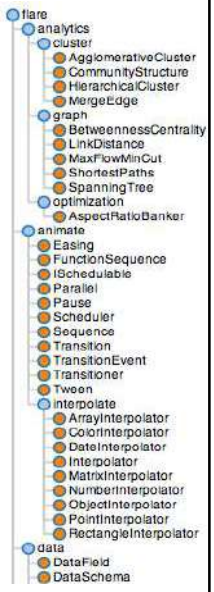
https://williamyaoh.com/posts/2023-04-22-drawing-trees-functionally.html

12

12

Indented Trees

- Indentation is used to show parent-child relationships
- It is easy to implement and intuitive to understand
- But much scrolling is often needed in big data browsing and may lose the context



The directory structure of Flare package

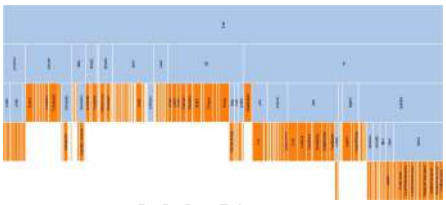
http://hci.stanford.edu/jheer/files/zoo/

13

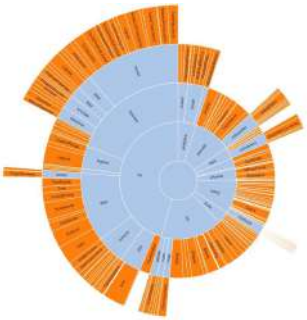
13

Layered (adjacency) Diagrams

- They are space-filling variants of node-link diagrams, such as **icicle diagram** and **Sunburst diagram**
- Nodes drawn as solid areas (arcs or bars)
- Placement relative to adjacent nodes reveals place in hierarchy
  - Root node at the top/center
  - Leaf nodes at the bottom/outside



Icicle Diagram



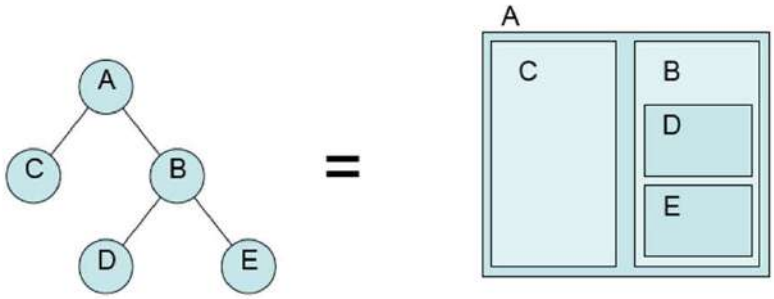
Sunburst Diagram

14

14

Enclosure Diagrams or Treemaps

- Encodes tree structure using *spatial enclosure*
  - Enclosure indicates hierarchy
- Benefits:
  - Provides single view of entire tree
  - Easier to spot small / large nodes

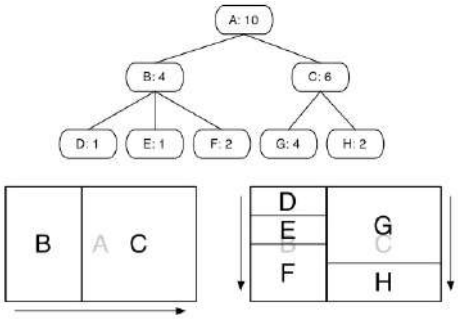


15

15

Enclosure Diagrams or Treemaps

- Assume that each leaf node has a size-related attribute (e.g., size of the file); size of parent nodes is the sum of its children's sizes
- Recursively slice and dice a rectangle in different directions to represent a (potentially large) hierarchy

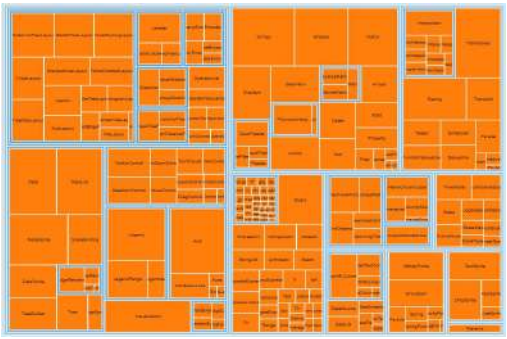


Source: http://eagereyes.org/Techniques/Treemaps.html

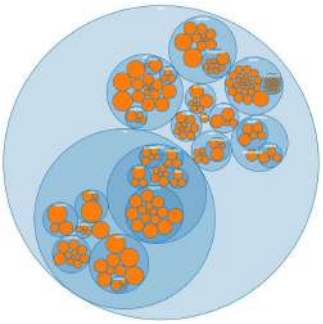
16

16

Enclosure Diagrams or Treemaps



Enclosure diagram (treemap)



Circle-packing layout

Question: Is it easy for you to visual identify the ***depth*** of the tree?

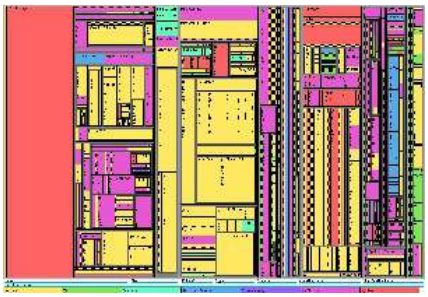
17

17

Father of Treemap

- Treemap was first introduced in 1991
- It was first used to visualize storage usage in a hard disk
- For more historical development of treemap: http://www.cs.umd.edu/hcil/treemap-history/

Ben Shneiderman



18

18

Slice-and-dice Treemap

- Ben Shneiderman proposed the slice-and-dice Treemap in 1991
- Layout method:
  - Recursive subdivision
  - Root is the largest rectangle
  - Size of child nodes are assigned according to their weights
  - Subdivide children nodes

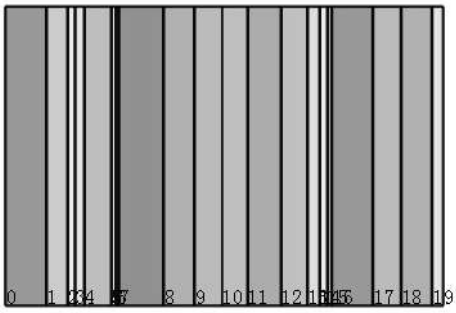
Tree visualization with treemaps: a 2-d space-filling approach. Shneiderman, B. March 1991. *ACM Transactions on Graphics*, vol. 11, 1 (Jan. 1992) 92-99. HCIL-91-03, CS-TR-2645, CAR-TR-548

19

19

Problems of Slice-and-dice Treemap

- Simple slice-and-dice subdivision method results in thin and long rectangles
- It would be difficult to interact with nodes

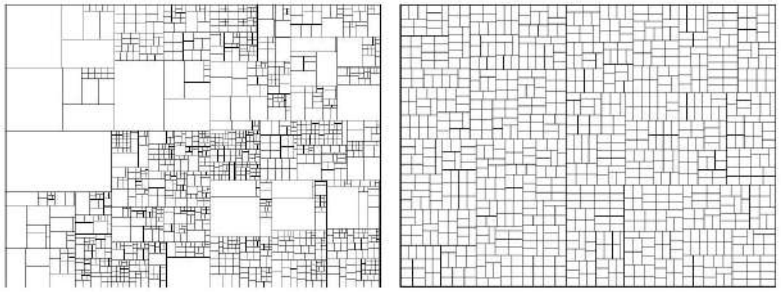


20

20

Squarified Treemap

- Optimize the placement of nodes within a level to make them as square as possible



(a) File system (b) Organization

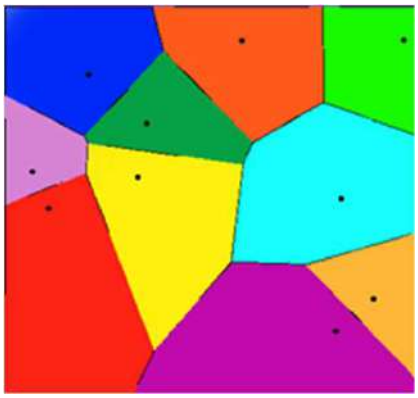
Source: http://www.win.tue.nl/~vanwijk/stm.pdf

21

21

Variant of Treemap: Voronoi Treemap

- Voronoi treemaps visualize hierarchical data by recursively partitioning convex polygons using weighted centroidal Voronoi diagrams



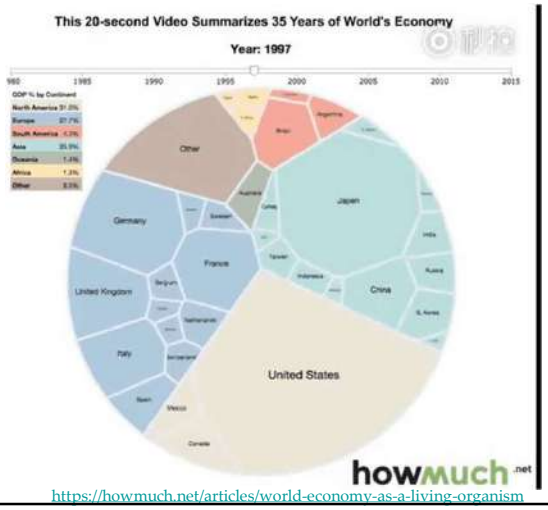
https://en.wikipedia.org/wiki/Voronoi_diagram

22

22

Variant of Treemap: Voronoi Treemap

- Example: Use animated Voronoi Treemap to summarize 35 Years of World's Economy



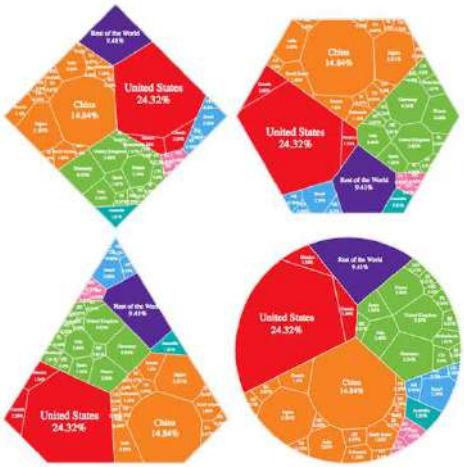
https://howmuch.net/articles/world-economy-as-a-living-organism

23

23

Variant of Treemap: Voronoi Treemap

- Voronoi treemap in D3



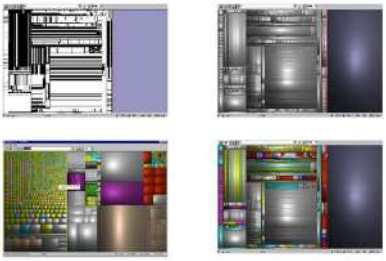
https://github.com/Kcnarf/d3-voronoi-treemap

24

24

Treemap Applications

Stock Market



Disk Browser

Business Sale
Data Analytics

Source: http://www.smartmoney.com/map-of-the-market/?link=SM_topnav_invest
http://www.hivegroup.com/solutions/demos/salesforce.html

25

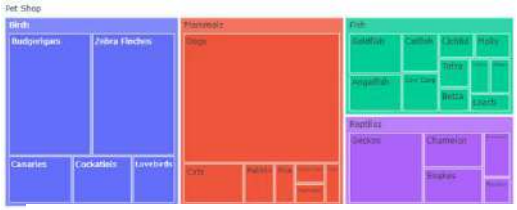
25

Treemap Implementation with Plotly

- The **treemap** method from the Python package **Plotly** can be used.

```
import plotly.express as px                # import Plotly Express
L1 = 'Species'                             # top hierarchy – Species
L2 = 'Animal Name'                         # next hierarchy – Animal Name
Fig = px.treemap(Data, path=[px.Constant('Pet Shop'), L1, L2], # call treemap
                 values='No. in Shop')  # map rectangle area to no. in pet shop
```

L1	L2	Area
Species	Animal Name	No. in Shop
Birds	Zebra Finches	50
Birds	Canaries	15
:	:	
Mammals	Dogs	100
:		
Fish	Goldfish	12
:		
Reptiles	Monitor	2



Treemap of Animal Inventory (No. in Shop) encoded in the Area

Plotly, Treemap - https://plotly.com/python/treemaps/

26

26

Summary

- Hierarchical data represent the **hierarchical relationship** between objects/entities, where **the attributes** associated with objects/entities are often considered as well.

- We have introduced four major types of **hierarchical data visualization** (a.k.a., **tree visualization**) techniques
  - Node-link (tree) diagrams
  - Indented trees
  - Layered (adjacency) diagrams
  - Enclosure diagrams

27

27

Questions?

Thank You!

28

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 11.2
Graph Visualization

2

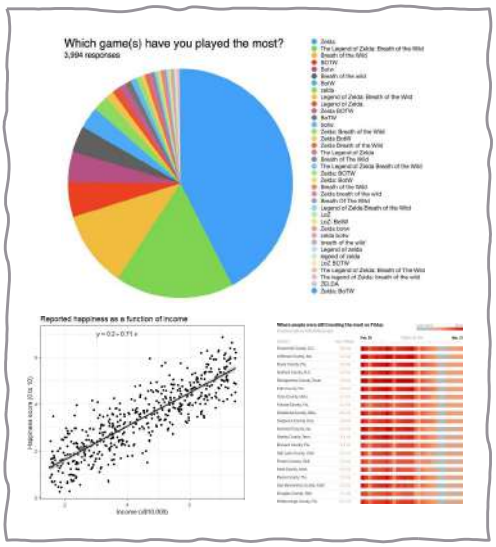
Outline

- What is a graph/network?
  - Real world networks
  - Common graph vocabulary
- Graph Visualization Methods
  - Node-link diagram
  - Adjacency matrix
  - Others
- Packages and tools for graph visualization

3

3

What is a Graph?



Graphs

Common charts that represent data are often referred to as "graphs", or "graph visualizations"

- Bar charts
- Line charts
- Pie charts
- Etc.

Clarification: for this lecture, when I refer to graphs, I do *not* mean the type of charts shown on the left.

4

4

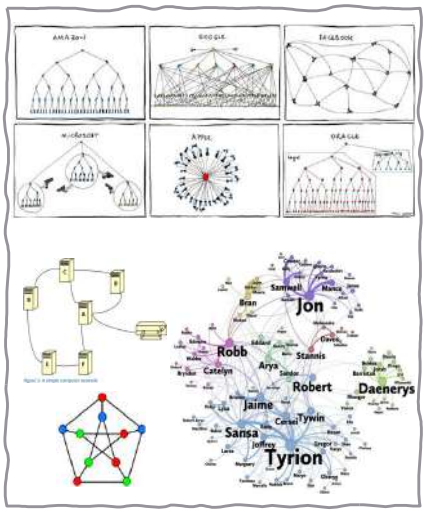
What is a Graph?

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Graphs

Let's instead talk about graphs, networks, & trees in the mathematical sense: a model for representing items and the relationships between those items

- Social / friendship networks
- Computer networks
- Energy or transportation grids
- Organizational structures
- Etc.



Why do we care about visualizing graphs?

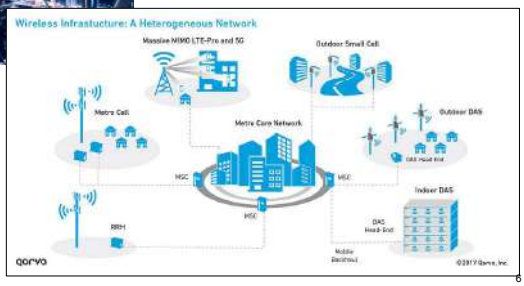
5

5

Networks in Real World

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

- Telecommunication network



6

6

Networks in Real World

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

- Social network

7

7

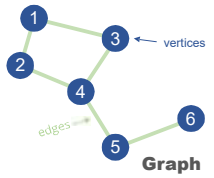
8

Common Graph Vocabulary

Graph G = {V, E}

V => collection of **vertices**

E => set of **edges** consisting of vertex pairs



Graph
(represented as a node-link diagram)

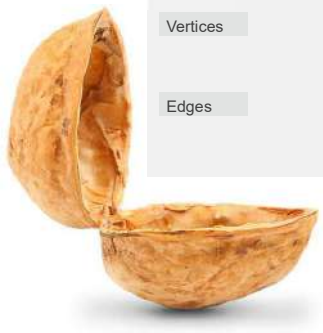
G = {V, E}

V = {1, 2, 3, 4, 5, 6}
E = {(1,2), (1,3), (2,4), (3,4), (4,5), (5,6)}

9

9

Common Graph Vocabulary



Graphs

Vertices

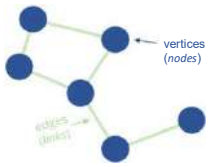
Edges

aka

Networks

Nodes

Links

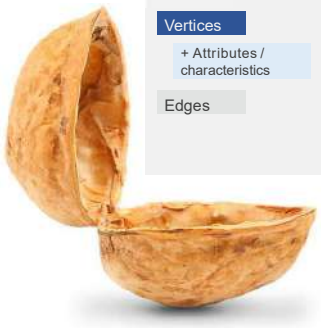


Graph (*network*)

10

10

Common Graph Vocabulary

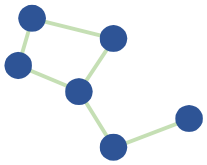


Graphs

Vertices

+ Attributes / characteristics

Edges



Vertex id	Name	Favorite color	Popularity
1	Sam	Blue	6
2	Sebastian	Green	7
3	Abigail	Purple	8
4	Haley	Pink	2
5	Shane	Orange	4
6	Leah	Purple	7

11

11

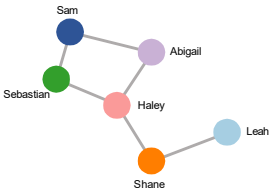
Common Graph Vocabulary

Graphs

Vertices

+ Attributes / characteristics

Edges

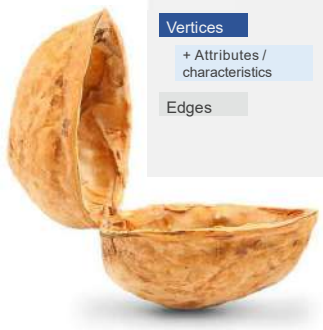


Vertex id	Name	Favorite color	Popularity
1	Sam	Blue	6
2	Sebastian	Green	7
3	Abigail	Purple	9
4	Haley	Pink	2
5	Shane	Orange	4
6	Leah	Purple	7

12

12

Common Graph Vocabulary

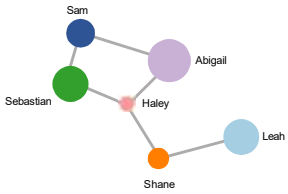


Graphs

Vertices

+ Attributes / characteristics

Edges

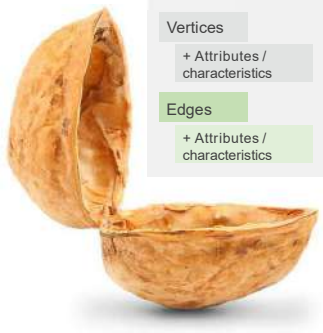


Vertex id	Name	Favorite color	Popularity
1	Sam	Blue	6
2	Sebastian	Green	7
3	Abigail	Purple	9
4	Haley	Pink	2
5	Shane	Orange	4
6	Leah	Purple	7

13

13

Common Graph Vocabulary



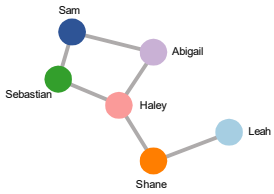
Graphs

Vertices

+ Attributes / characteristics

Edges

+ Attributes / characteristics

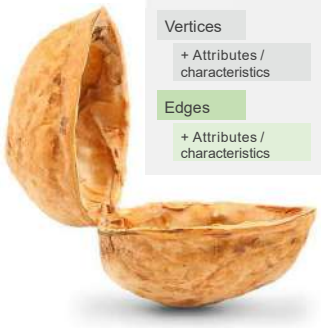


Edge id	Source	Target	Friend value
1	Sam	Sebastian	10
2	Sam	Abigail	6
3	Sebastian	Haley	1
4	Abigail	Haley	2
5	Haley	Shane	1
6	Shane	Leah	2

14

14

Common Graph Vocabulary



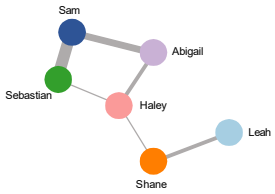
Graphs

Vertices

+ Attributes / characteristics

Edges

+ Attributes / characteristics

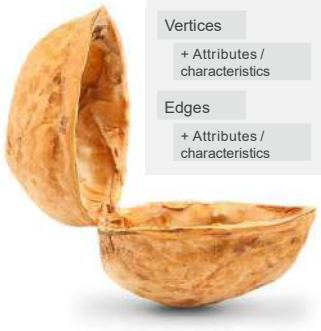


Edge id	Source	Target	Friend value
1	Sam	Sebastian	10
2	Sam	Abigail	6
3	Sebastian	Haley	1
4	Abigail	Haley	2
5	Haley	Shane	1
6	Shane	Leah	2

15

15

Common Graph Vocabulary



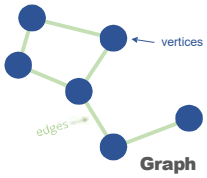
Graphs

Vertices

+ Attributes / characteristics

Edges

+ Attributes / characteristics

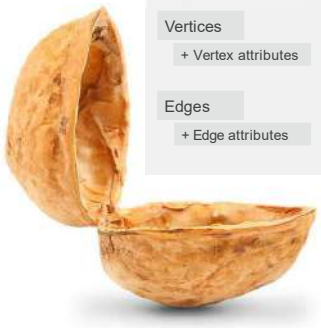


Questions?

16

16

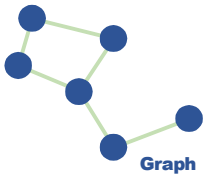
Common Graph Vocabulary



Graphs & Trees

Vertices
+ Vertex attributes

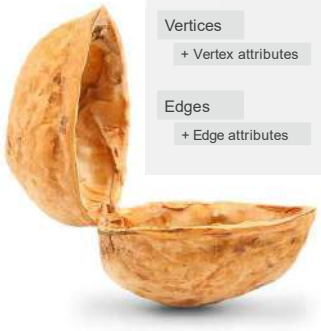
Edges
+ Edge attributes



17

17

Common Graph Vocabulary



Trees

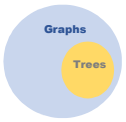
Vertices
+ Vertex attributes

Edges
+ Edge attributes

Properties of trees

- Connected
- Hierarchical structure
- One path between any pair of vertices
- No cycles in the graph
- Removing an edge would create a disconnected graph

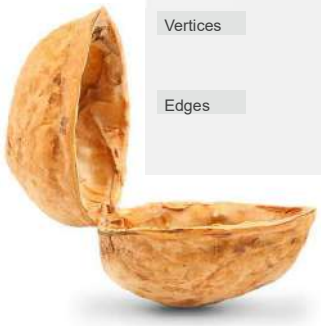
Every tree is a graph but not every graph is a tree!



18

18

Common Graph Vocabulary



Graphs

Vertices

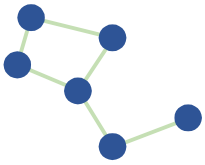
Edges

$\stackrel{?}{=}$

Trees

Vertices

Edges

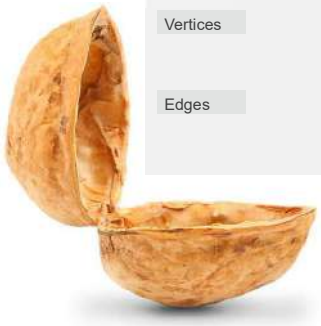


Is this graph a tree?

19

19

Common Graph Vocabulary



Graphs

Vertices

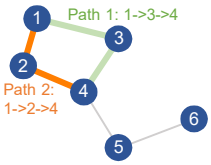
Edges

$\supseteq$

Trees

Vertices

Edges

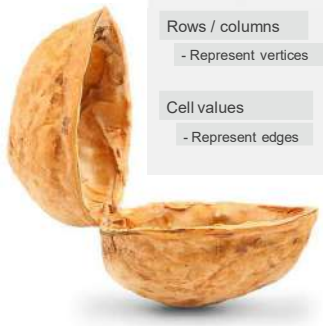


Is this graph a tree?
No

20

20

Common Graph Vocabulary



Adjacency Matrix

Rows / columns
- Represent vertices

Cell values
- Represent edges

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 1 & 1 \\ 2 & 1 & 0 & 1 & 0 \\ 3 & 1 & 1 & 0 & 1 \\ 4 & 1 & 0 & 1 & 0 \end{array}$$

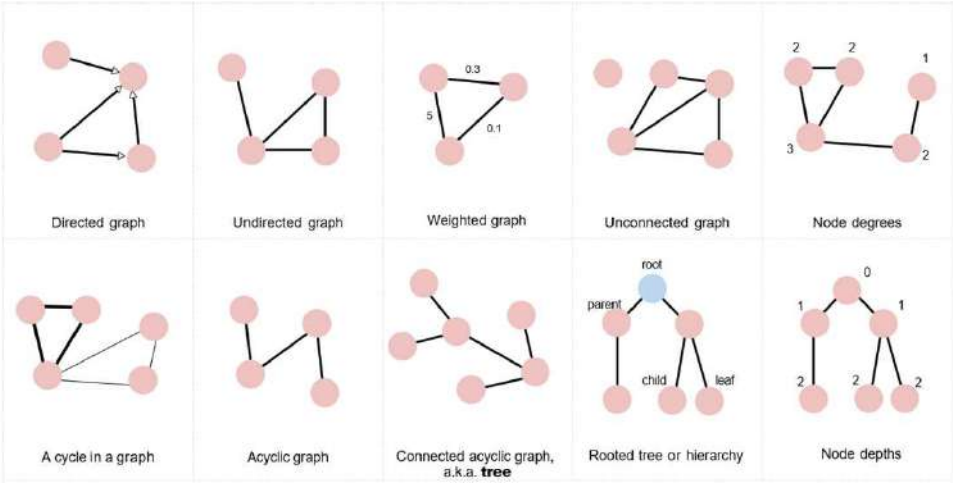
Properties of adjacency matrices

- Another graph representation
- Symmetrical along the diagonal
  - Can read from top or bottom half
  - Typically, all 0's on the diagonal (unless self-loops)
- Non-zero cell value means an edge exists between that pair
  - Zero cell value means no edge exists
  - Cell values can also be edge 'weights' (so not just 0/1)

21

21

Common Graph Vocabulary



22

22

Outline

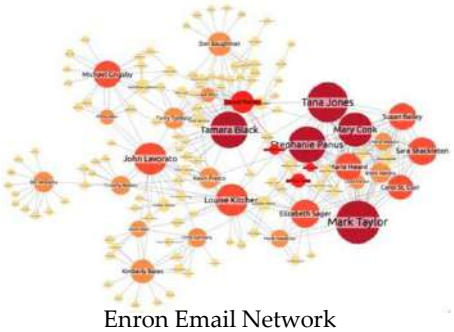
- What is a graph/network?
  - Real world networks
  - Common graph vocabulary
- Graph Visualization Methods
  - Node-link diagram
  - Adjacency matrix
  - Others
- Packages and tools for graph visualization

23

23

Graph Visualization Methods

- Graph visualization (a.k.a, network visualization) is concerned with visual representations of graph or network data
- Effective graph visualization reveals graph structures and help users understand and analyze the network data



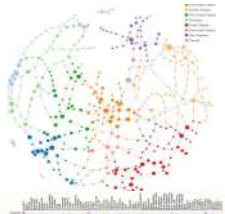
Enron Email Network
https://cambridge-intelligence.com/using-social-network-analysis-measures/

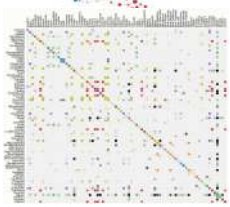
24

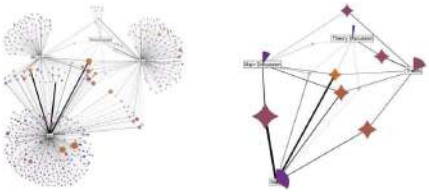
24

Graph Visualization Methods

- Node-link diagram
  - Force-directed layout
  - MDS layout
- Adjacency matrix
- Others
  - Hybrid layout
  - Graph simplification methods





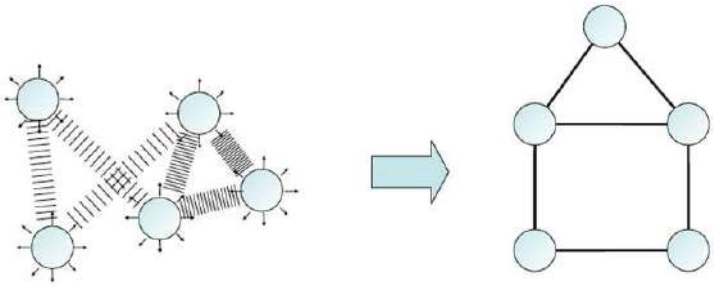


25

25

Node-link diagram: Force-Directed Layout

- What about graphs without an intrinsic order?
- Physical model:
  - edge → spring
  - node → mass point



26

26

Node-link diagram: Force-Directed Layout

- It assumes that there is a spring between each pair of nodes, which leads to attractive force and repulsive force as follows:

$$f_a(d) = d^2/k$$

$$f_r(d) = -k^2/d$$

where d is the distance between the two nodes and k is a constant.

- With the effects of **attractive force and repulsive force**, nodes far away will be dragged near and nodes that are overlapped will be pushed away, and finally reach a stable balance after **iterations.**

T. M. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. Software: Practice and Experience, 21(11):1129–1164, 1991.

27

27

Node-link diagram: Force-Directed Layout

- Starting positions: random or initial **configuration**
- Loop:
  - Compute **the attractive and repulsive** forces for every pair of nodes
  - Accumulate the **force** (vector) for every node
  - Update the position of each node **step by step** according to their forces
- The loop stops until the layout is "good enough"

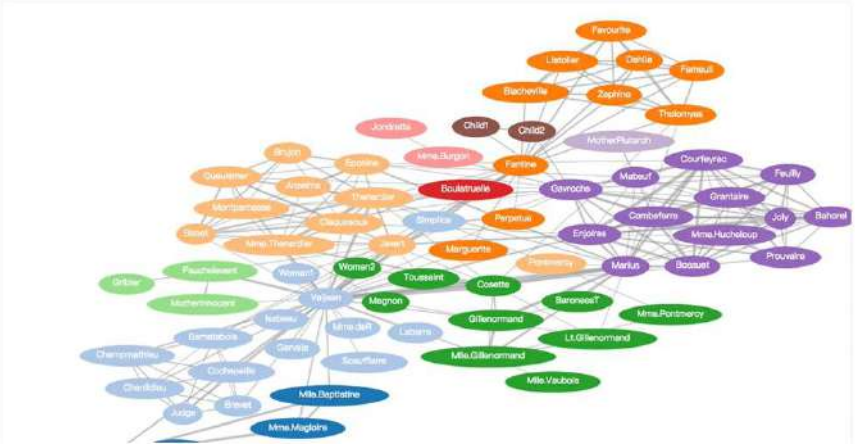
28

28

Results of Force-directed Layout

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Force-directed graph with elliptic forces



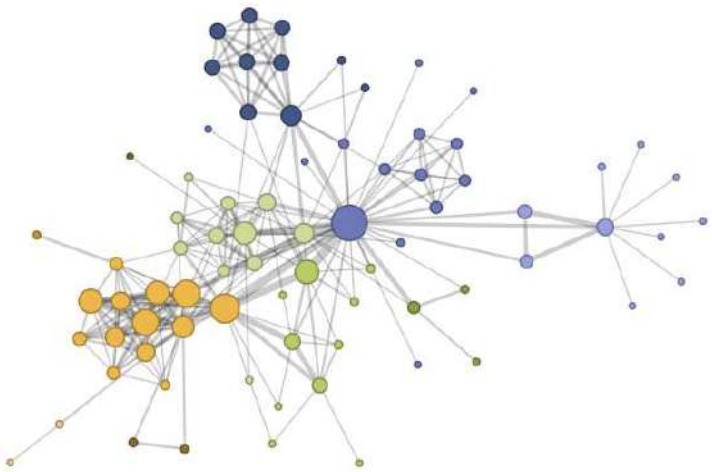
https://bl.ocks.org/jpurma/6dd2081cf25a5d2dfcdcab1a4868f237

29

29

Results of Force-directed Layout

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE



Character relations in Les Miserables

http://hci.stanford.edu/jheer/files/zoo/

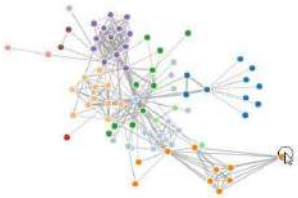
https://homes.cs.washington.edu/~jheer//files/zoo/ex/networks/force.html

30

30

Pros and Cons of Force-directed Layout

- Pros
  - Very flexible for any type of graphs
  - Forces can be customized
  - Easy to implement



- Cons
  - **Local optimal**
  - **Initial configuration** is important
  - **Computation complexity** of iterative algorithm
- Extensions
  - Barnes-Hut quadtree decomposition
  - FADE、GRIP、FMS、FM^3、GVA

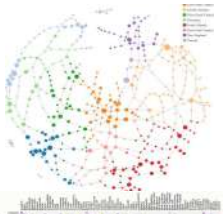
try it interactively at https://observablehq.com/@d3/force-directed-graph

31

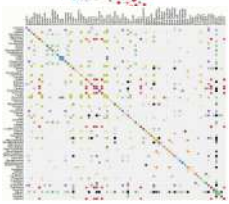
31

Graph Visualization Methods

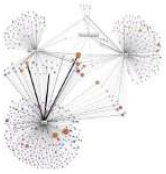
- Node-link diagram
  - Force-directed layout
  - **MDS layout**

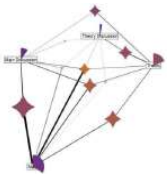


- Adjacency matrix



- Others
  - Hybrid layout
  - Graph simplification methods





32

32

Node-link diagram: MDS (Multidimensional Scaling) Layout

- Focus on addressing the limitations of force-directed layout
- Dimension reduction
  - Keep the consistency of relative distance between nodes
- MDS is a global optimal method
- Optimal function $||x_i - x_j|| \approx d_{\mathrm{ij}}$, where d_{ij} is the graph-theoretical distance between Node i and j, and x_i and x_j correspond to their coordinates in the 2D plane

33

33

Solving Optimal Function

- Stress majorization
  - $\mathrm{Stress(X)} = \sum_{i,j} w_{ij}(d_{ij} - ||x_i - x_j||)^2$
  - d_{ij} is the graph-theoretical distance between the i-th node and the j-th node
  - $w_{ij} = d_{ij}^q$, usually $\mathrm{q} = -2$

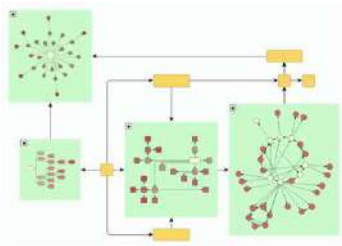
34

34

Other Presentations of Node-link Diagram

- Orthogonal Diagram
  - UML diagram

- Nested ordered
  - Recursively applying nested layout
  - For intrinsic ordered topology

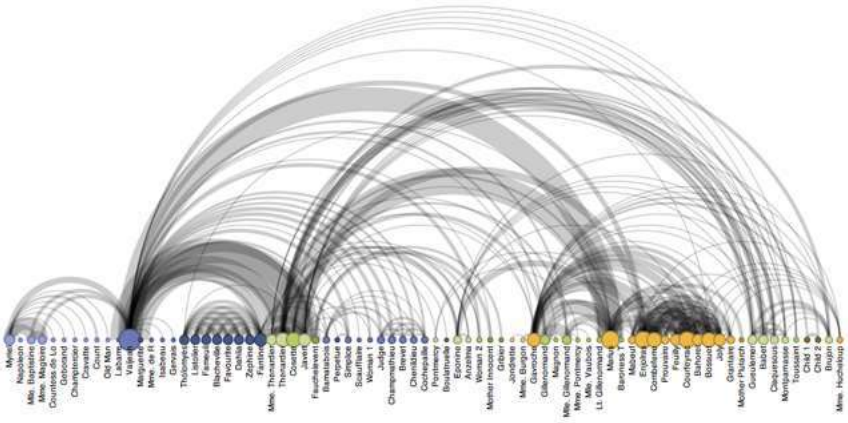


- Arc Diagram

35

35

Arc Diagram



Character relations in Les Miserables

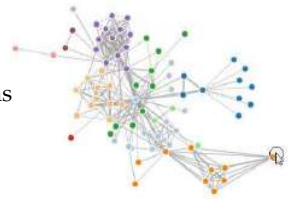
http://hci.stanford.edu/jheer/files/zoo/
https://homes.cs.washington.edu/~jheer//files/zoo/ex/networks/arc.html

36

36

Node-Link Diagram Summary

- Pros:
  - Intuitive visual interpretation
  - Good representation of topology, clusters and paths
  - Flexible, many variants



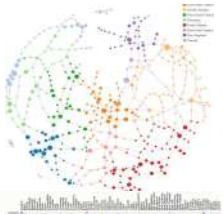
- Cons:
  - Almost for all algorithms, time complexity is a bit high ~ $O(n^2)$
  - Not so good for dense graphs (especially edge cluttered graphs)

37

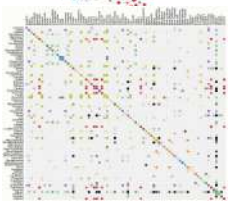
37

Graph Visualization Methods

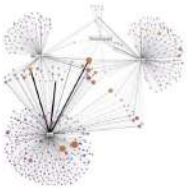
- Node-link diagram
  - Force-directed layout
  - MDS layout

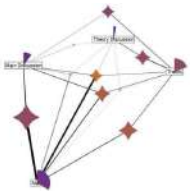


- **Adjacency matrix**



- Others
  - Hybrid layout



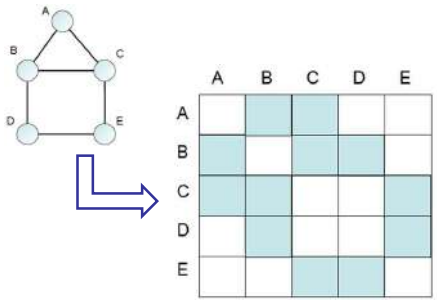


38

38

Adjacency Matrix

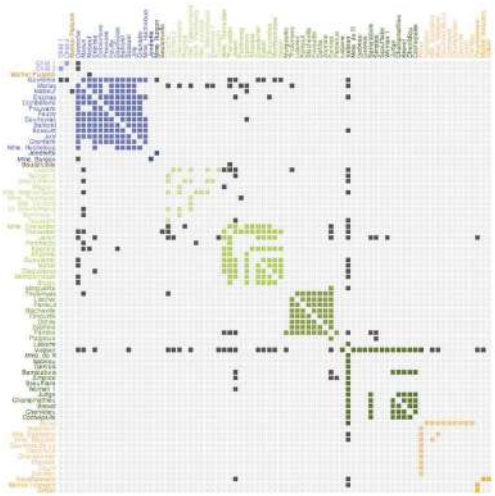
- $N \times N$ matrix, representing relations among N objects
- Position *(i,j)* represents the relation between the *i*-th object and the *j*-th object
  - Weight
  - Direction
- Related issues
  - Ordering
  - Path finding



39

39

Good Ordering of Adjacency Matrix



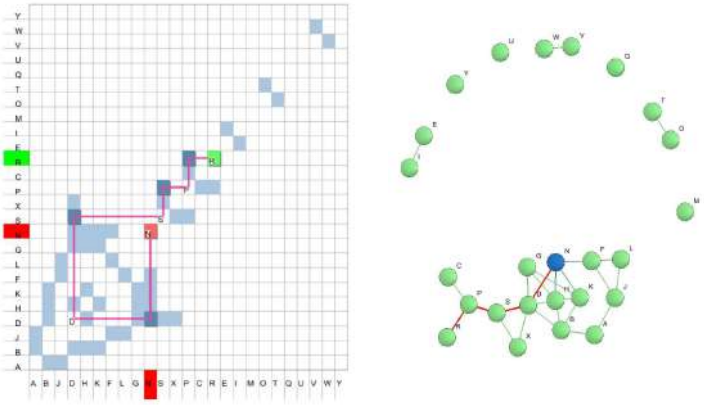
Appropriate node ordering is critical for visually identifying the meaningful graph patterns

https://homes.cs.washington.edu/~jheer/files/zoo/ex/networks/matrix.html

40

40

Path Visualization for Adjacency Matrix



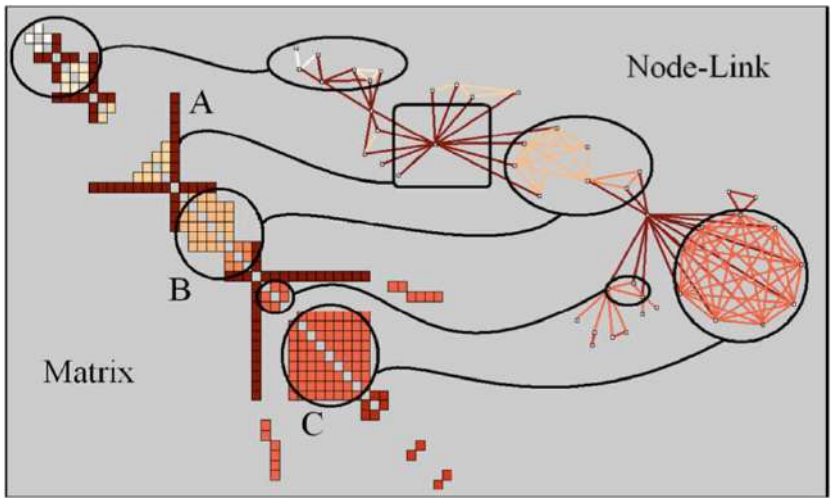
Left: Connecting matrix entries and the diagonal.
Right: Path {N;D;S;P;R} is highlighted in the Node-link Diagram.

Compared with node-link diagrams, it is **hard to perform path-finding tasks** in adjacency matrix.

41

41

Recognizing the Patterns of Matrix



42

42

Adjacency Matrix Summary

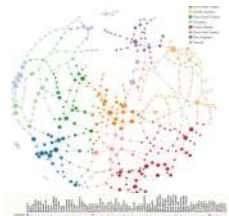
- Pros:
  - No edge crossing, good for edge cluttered graph
  - Good visual scalability
  - Good presentation of graph pattern
- Cons:
  - Visualization is abstract to understand
  - Difficult to follow a transitive relation path

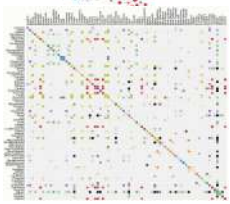
43

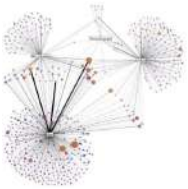
43

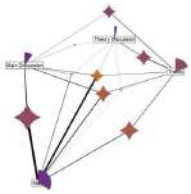
Graph Visualization Methods

- Node-link diagram
  - Force-directed layout
  - MDS layout
- Adjacency matrix
- Others
  - **Hybrid layout**
  - **Graph simplification methods**









44

44

Hybrid Layout

- Adjacency matrix can handle complex and dense edge relations, but suffers from insufficient space usage when there are many nodes
- Node-link diagram can handle relatively more nodes, but can suffer from serious visual clutters when there are dense edge relations
- What if there are a large number of nodes, and some of them have dense edge relations?

-- Combine node-link diagram with matrix

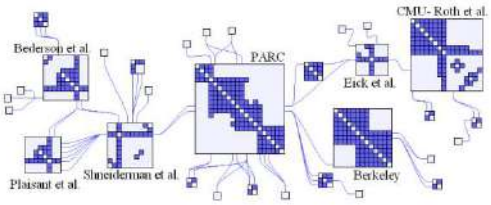


Fig. 1: NodeTrix Representation of the largest component of the Info-Vis Co-authorship Network

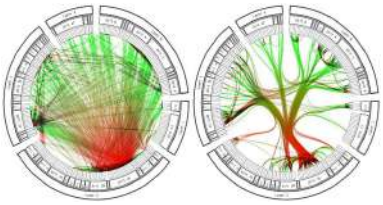
Nathalie Henry, et al. NodeTrix: A Hybrid Visualization of Social Networks. TVCG 2007.

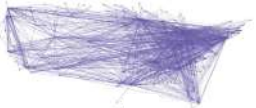
45

45

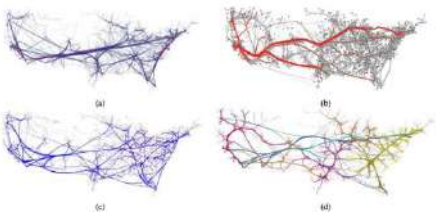
Graph Simplification Methods: - Edge Bundling

- Edge bundling approaches are designed to handle the dense edges between nodes, reducing visual clutters and helping users identify the overall links between nodes





Original Graph Visualization



[1] D. Holten, Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data, TVCG 2006.
[2] Zhou, Hong, et al. "Edge bundling in information visualization." *Tsinghua Science and Technology* 18.2 (2013): 145-156.

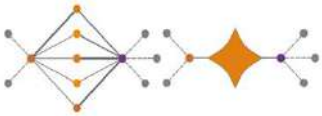
46

46

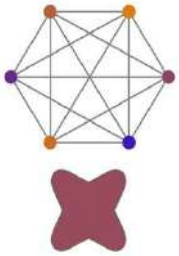
Graph Simplification Methods:
- Network Motifs

- Network motifs aims to represent representative subgraphs of a large graph as a series of meaningful motifs to simplify the visualization of large graphs

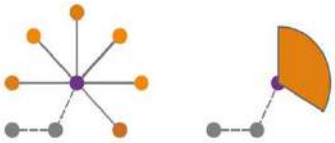
Connector



Clique



Fan

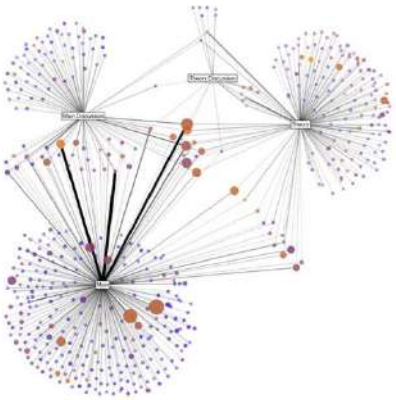


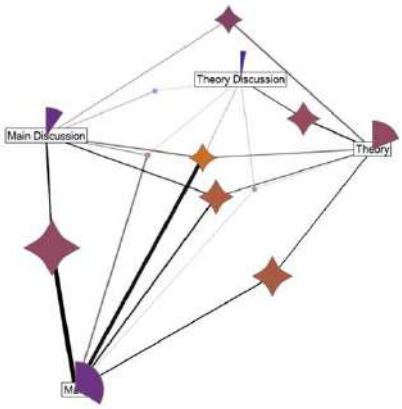
Dunne and Shneiderman, Motif Simplification: Improving Network Visualization Readability with Fan, Connector, and Clique Glyphs. CHI2013 47

47

Graph Simplification Methods:
- Network Motifs

- Network motifs aims to represent representative subgraphs of a large graph as a series of meaningful motifs to simplify the visualization of large graphs



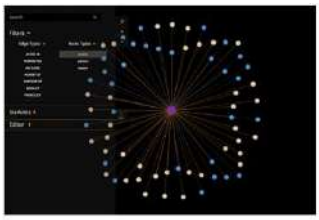


Dunne and Shneiderman, Motif Simplification: Improving Network Visualization Readability with Fan, Connector, and Clique Glyphs. CHI2013 48

48

Graph Visualization Tools

- Visualization packages



Alchemy.js
…

Sigma.js
Cytoscape.js
VivaGraphJS
…

D3.js
Pixi.js
…

More and more flexible

49

49

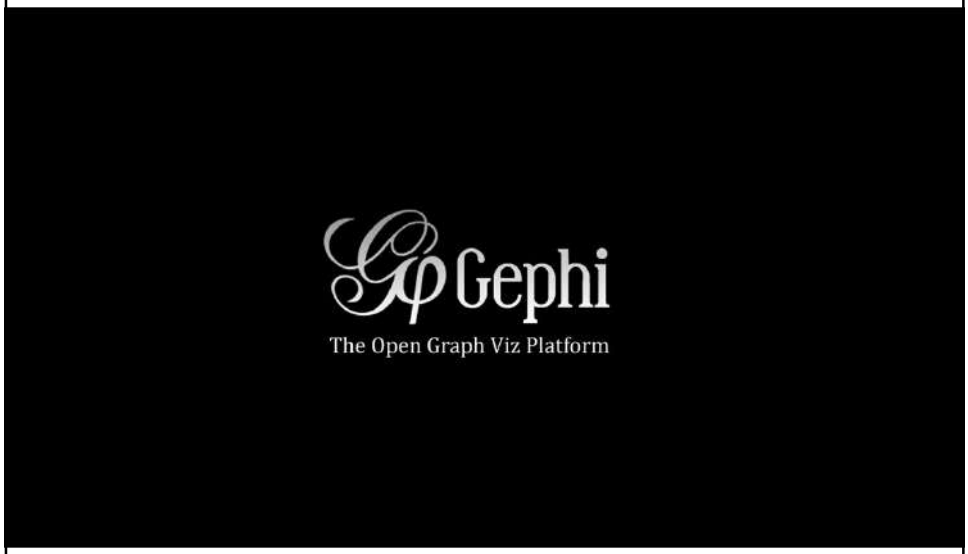
Graph Visualization Tools

- Softwares
  - Gephi
  - Cytoscape
  - Palantir
  - visone
  - …

50

50

Graph Visualization Tools



https://gephi.org/

51

51

Summary

- Graph visualizations are widely used
- Graph visualization methods
  - Node-link diagram
    - Familiar, but problematic for dense graphs
  - Matrix
    - Abstract, hard to follow paths
  - Graph simplification can help
    - Not always possible, and not always appropriate
- Take-home message: no best solution; graph visualization is still under active research!!!

52

52

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Questions?

Thank You!

53

SC4024/CZ4124
Data Visualization

Assistant Professor WANG Yong
yong-wang@ntu.edu.sg
CCDS, Nanyang Technological University

1

Chapter 12
High-dimensional Data Visualization

2

Outline

- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

3

3

Data Dimension

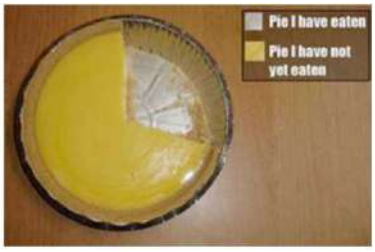
- Dimension (Number of attributes)
  - One Dimension
  - Two Dimension
  - Three Dimension
  - High Dimension

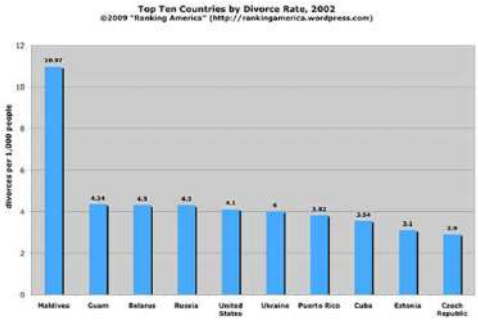
4

4

1-D Data

Eaten or not





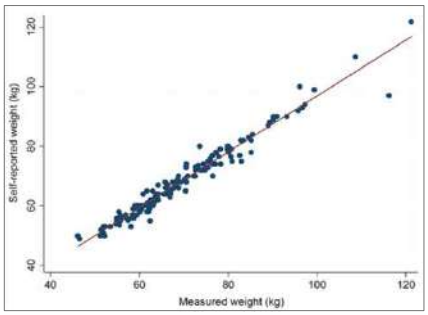
Top ten divorce rates

5

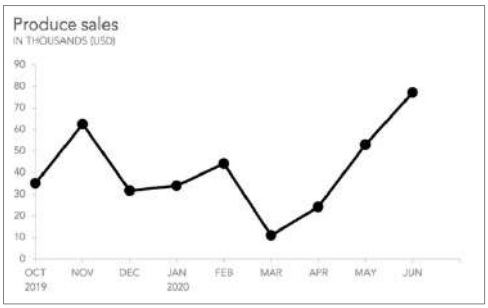
5

2-D Data

2-D plots



Self-reported weights *vs.* measured weight

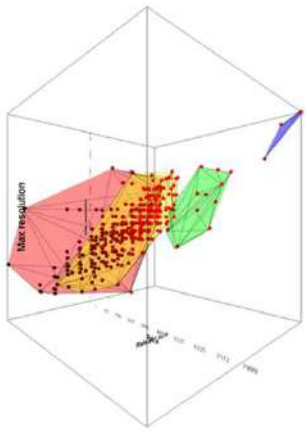


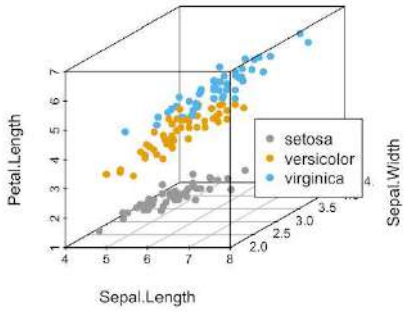
Produces sales *vs.* time

6

6

3-D Data





3D data can be represented by the 3D coordinates of each dot in a 3D scatterplot.

[1] Elmqvist et al. "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation." IEEE transactions on Visualization and Computer Graphics 14.6 (2008): 1539-1148.
[2] http://www.sthda.com/english/wiki/scatterplot3d-3d-graphics-r-software-and-data-visualization

7

7

High-dimensional Data

- How to visualize high-dimensional data in visual space (2-D or 3-D) ?

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
3.8	0.31	0.02	11.1	0.036	20	114	0.99248	3.75	0.44	12.4	6	white
3.9	0.225	0.4	4.2	0.03	29	118	0.989	3.57	0.36	12.8	8	white
4.2	0.17	0.36	1.8	0.029	93	161	0.98999	3.65	0.89	12	7	white
4.2	0.215	0.23	5.1	0.041	64	157	0.99688	3.42	0.44	8	3	white
4.4	0.46	0.1	2.8	0.024	31	111	0.98816	3.48	0.34	13.1	6	white
4.4	0.32	0.39	4.3	0.03	31	127	0.98904	3.46	0.36	12.8	8	white
4.4	0.54	0.09	5.1	0.038	52	97	0.99022	3.41	0.4	12.2	7	white
4.5	0.19	0.21	0.95	0.033	89	159	0.99332	3.34	0.42	8	5	white
4.6	0.52	0.15	2.1	0.054	8	65	0.9934	3.9	0.56	13.1	4	red
4.6	0.445	0	1.4	0.053	11	178	0.99426	3.79	0.55	10.2	5	white
4.7	0.6	0.17	2.3	0.058	17	106	0.9932	3.85	0.6	12.9	6	red
4.7	0.67	0.09	1	0.02	5	9	0.98722	3.3	0.34	13.6	5	white
4.7	0.455	0.18	1.9	0.036	33	106	0.98746	3.21	0.83	14	7	white
4.7	0.145	0.29	1	0.042	35	90	0.9908	3.76	0.49	11.3	6	white
4.7	0.335	0.14	1.3	0.036	69	168	0.99212	3.47	0.46	10.5	5	white

Wine Dataset

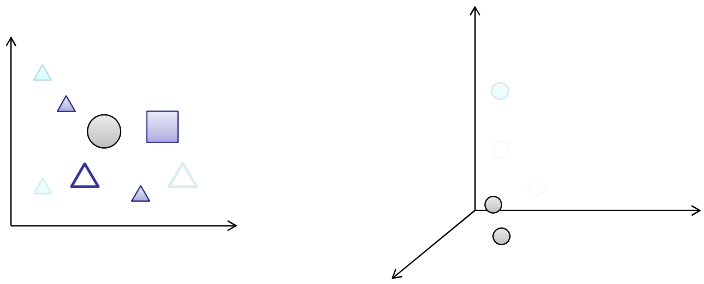
Source: UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Wine

8

8

Simple Solutions

- Add more channel on 2-D or 3-D plots
  - Shape/fill style/color/size of points



9

9

Simple Solutions

- Multiple coordinated views: present some attributes of objects in a view

[1] Andrew et al. "Leveraging Interaction History for Intelligent Configuration of Multiple Coordinated Views in Visualization Tools LIVVIL: Logging Interactive Visualizations & Visualizing Interaction Logs (2016).
[2] https://square.github.io/crossfilter/

10

10

High-Dimensional Data Visualization

- More solutions?

11

11

One Concrete Example – Show Me Your Visualization Designs?

- Suppose you are given a dataset describing 100 students' profile (anonymized) and academic information, which includes the following attributes:
  - Age
  - Height
  - Weight
  - Travel time to campus
  - Math score
  - Physics score
  - Visual analytics score
  - English language score
  - Algorithm design score
  - Python programming score
  - Art design score
  - Chemistry score

12

12

One Concrete Example – Show Me Your Visualization Designs?

Task 1: Show the correlation between any two attributes?

Task 2: Show the overall distribution of all these attributes of all students and help viewers visually identify the similarities between students

- Age
- Height
- Weight
- Travel time to campus
- Math score
- Physics score
- Visual analytics score
- English language score
- Algorithm design score
- Python programming score
- Art design score
- Chemistry score

13

13

Outline

- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

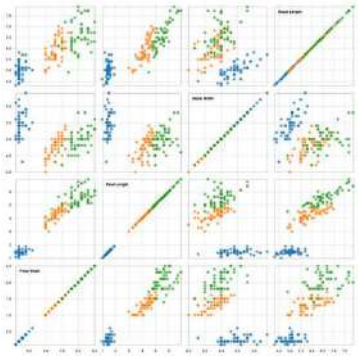
14

14

Scatter-plot Matrix

- 2-D plot for each dimension pair
- Display correlations between dimensions
- The number of 2-D plots is proportional to square of dimensions

Implemented by D3.js



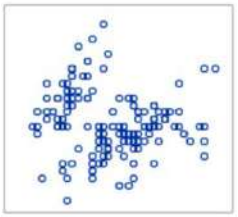
Iris Dataset: https://observablehq.com/@d3/brushable-scatterplot-matrix

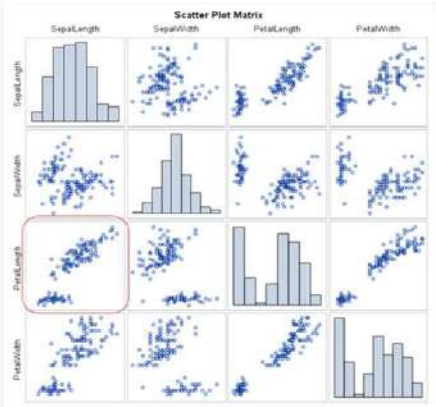
15

15

Scatter-plot Matrix

- Scatter-plot matrix can be further enhanced:
  - It can be combined with dimension reduction based projection
  - The diagonal cell can be used to show the distribution of each attribute/variable





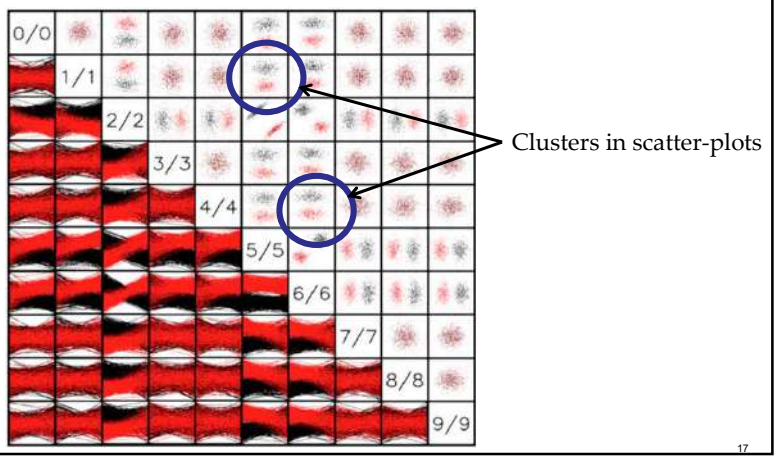
16

16

Scatter-plot Matrix – Insight Discovery

- Scatter-plot matrix can be further enhanced:

  - The cells of the upper triangle and bottom triangle can be used to serve for different analysis tasks, e.g., clustering and correlation

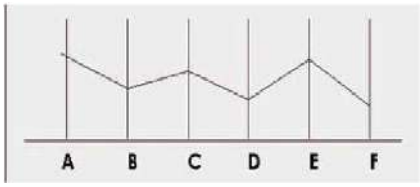


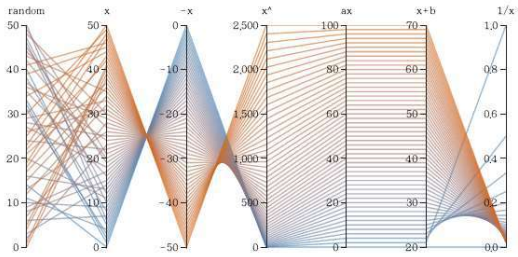
17

17

Parallel Coordinates

- It was proposed by Prof. Alfred Inselberg in 1985 for high-dimensional geometry
- Parallel axes
- Data points represented by lines



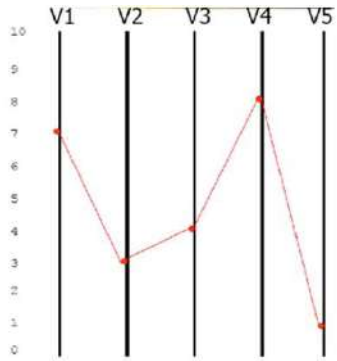


Inselberg, Alfred. "The plane with parallel coordinates." *The visual computer* 1, no. 2 (1985): 69-91.

18

18

Parallel Coordinates



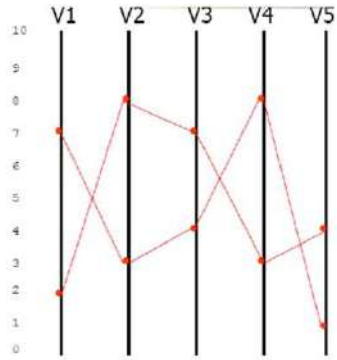
D1

7 3 4 8 1

19

19

Parallel Coordinates



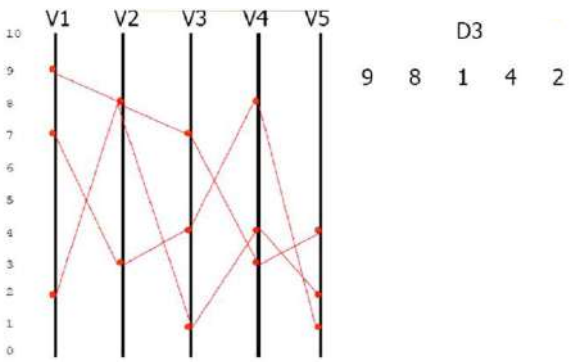
D2

2 7 6 3 4

20

20

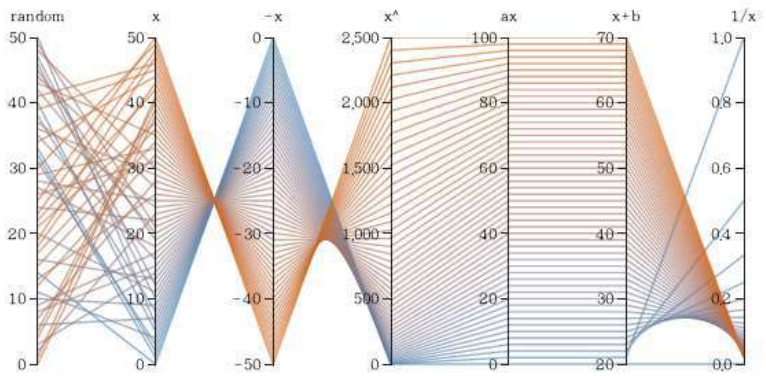
Parallel Coordinates



21

21

Correlations



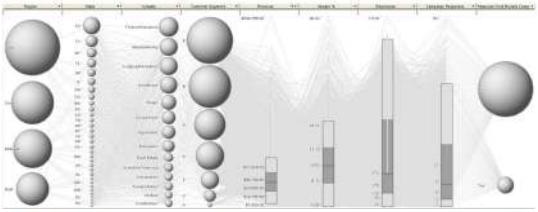
- The parallel coordinate illustrates the correlation between adjacent variables
- The order of the axis or variables really matters, as appropriate orders can reveal the correlation between adjacent axes more clearly

22

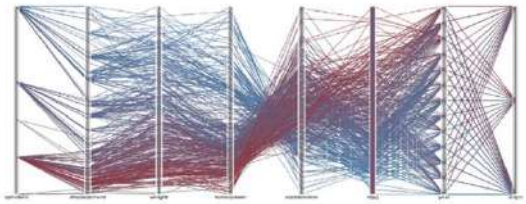
22

Parallel Coordinates in Visualization Software or Programming Packages

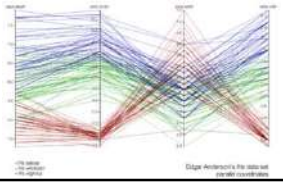
Advizor



Protovis



D3



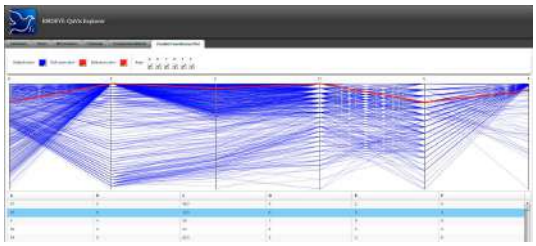
https://www.advizorsolutions.com/#home-demo
https://mbostock.github.io/protovis/

23

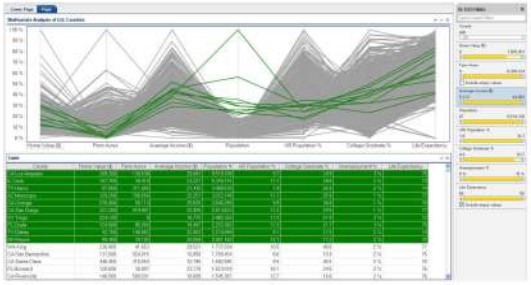
23

Parallel Coordinates + Interactions

- Selection



- Brush and Filter



Multivariate Analysis Using Parallel Coordinates. Stephen Few September 12, 2006
https://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf

24

24

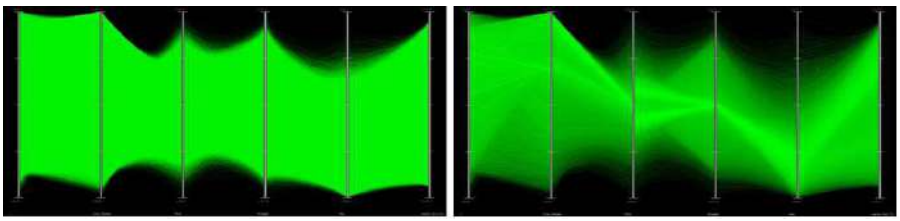
Any Other Variants of Parallel Coordinates?

25

25

Transparent Parallel Coordinates

- When there are a huge number of lines in the parallel coordinates, visual clutters occur and it looks messy. A good way to mitigate it is to adjust the opacity of the lines, for example, make those lines half-transparent.



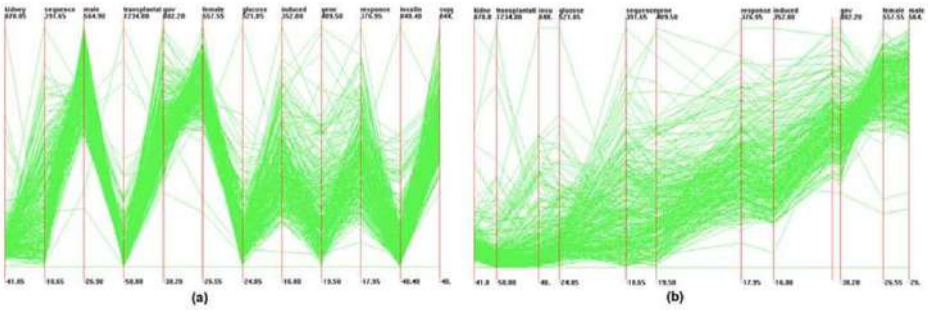
Chad Jones et al. "An Integrated Exploration Approach to Visualizing Multivariate Particle Data. " Computing in Science & Engineering (2008).

26

26

Parallel Coordinates: Re-ordering Axes to Reduce Visual Clutter

- By varying the dimension order in a display, it is possible to **reduce clutter** without reducing information content or modifying the data in any way.



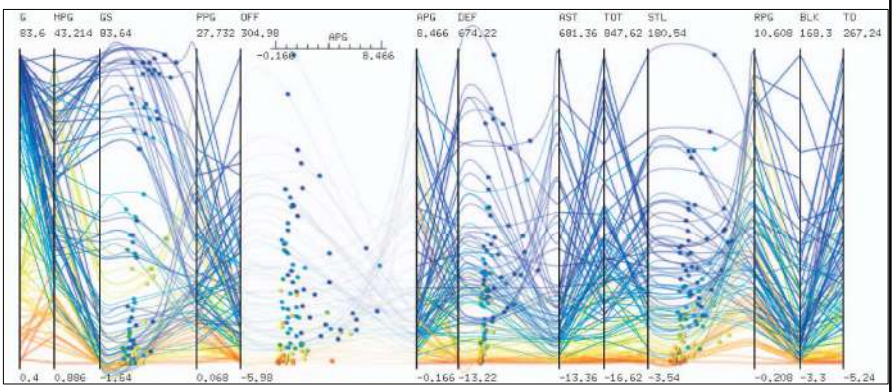
Peng, Wei , M. O. Ward et al. "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Re-ordering." IEEE Symposium on Information Visualization (2005).

27

27

Parallel Coordinates with Scatter-plots

- Augment parallel coordinates with scatterplots to facilitate **data selection** and **data clustering**

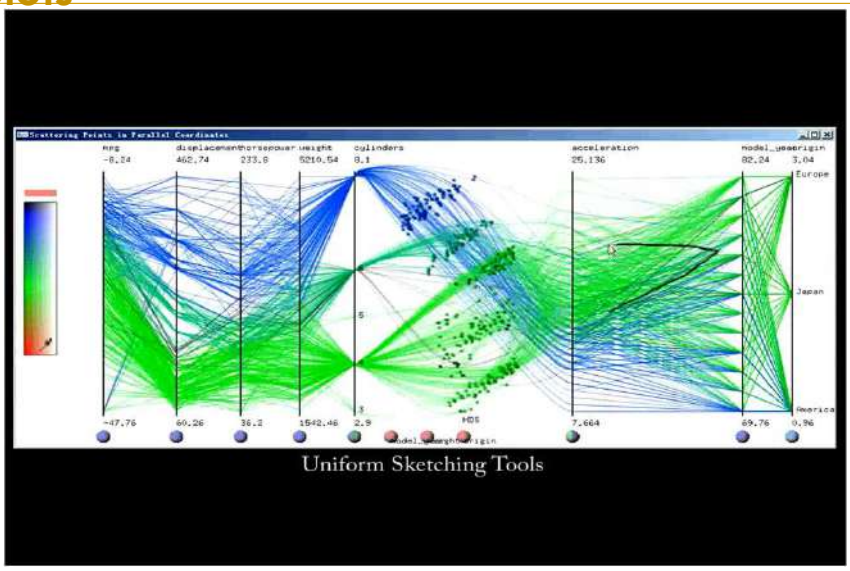


Yuan et al. "Scattering Points in Parallel Coordinates. " IEEE Transactions on Visualization and Computer Graphics (2009).

28

28

Parallel Coordinates with Scatter-plots



Yuan et al. "Scattering Points in Parallel Coordinates. " IEEE Transactions on Visualization and Computer Graphics (2009).

29

29

Outline

- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

30

30

Glyph-based Methods

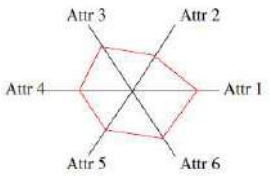
- Radar Chart (a.k.a.: star plot)
- Chernoff Faces

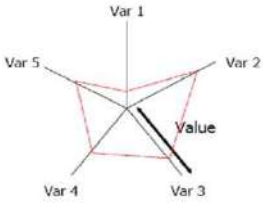
31

31

Star Plot

- Space variables around a circle
- Encode values on "spokes"
- Each data point is now a shape





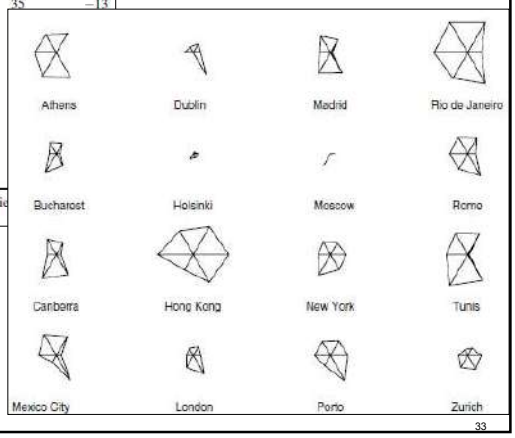
32

32

Multivariate Data: Climatic Variables

City	Precip. average	Temp. average	Temp. max average	Temp. min average	Record max	Record min
Athens	37	17	21	13	42	−3
Bucharest	58	11	16	5	49	−23
Canberra	62	12	19	6	42	−10
Dublin	74	10	12	6	28	−7
Helsinki	63	5	8	1	31	−36
Hong Kong	218	23	25	21	37	2
London	75	10	13	5	35	−13
Madrid	45	13	20	7		
Mexico City	63	17	23	11		
Moscow	59	4	8	1		
New York	118	12	17	8		
Porto	126	14	18	10		
Rio de Janeiro	109	25	30	20		
Rome	80	15	20	11		
Tunis	44	18	23	13		
Zurich	107	9	12	6		

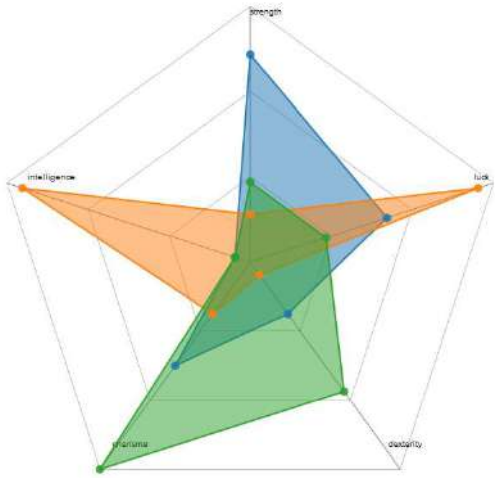
Table 4.1 Annual climatic values in Celsius of some world citie
http://www.weatherbase.com.



33

33

Star Plot in D3.js



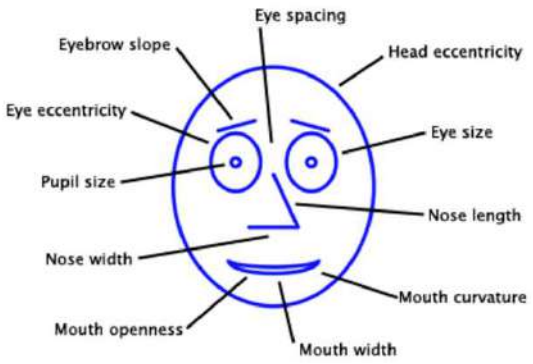
Source: http://graves.cl/radar-chart-d3/ and http://bl.ocks.org/nbremer/6506614

34

34

Chernoff Faces

- Use face encodings to represent different attributes



Facial Variables

Spinelli, Joseph G., and Yu Zhou. "Mapping quality of life with chernoff faces." In *Proceedings of 24th ESRI International User Conference*.. 2004.

35

35

Multivariate Data: Climatic Variables

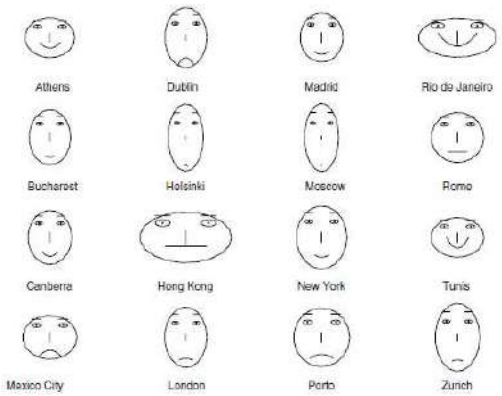
City	Precip. average	Temp. average	Temp. max average	Temp. min average	Record max	Record min
Athens	37	17	21	13	42	−3
Bucharest	58	11	16	5	49	−23
Canberra	62	12	19	6	42	−10
Dublin	74	10	12	6	28	−7
Helsinki	63	5	8	1	31	−36
Hong Kong	218	23	25	21	37	2
London	75	10	13	5	35	−13
Madrid	45	13	20	7	40	−10
Mexico City	63	17	23	11	32	−3
Moscow	59	4	8	1	35	−42
New York	118	12	17	8	40	−18
Porto	126	14	18	10	34	−2
Rio de Janeiro	109	25	30	20	43	7
Rome	80	15	20	11	37	−7
Tunis	44	18	23	13	46	−1
Zurich	107	9	12	6	35	−20

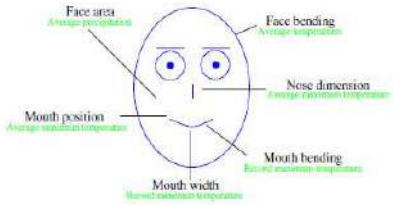
Table 4.1 Annual climatic values in Celsius of some world cities. Values from http://www.weatherbase.com.

36

36

Chernoff Faces





Source: The Use of Faces to Represent Points in k-Dimensional Space Graphically (http://www.jstor.org/stable/2284077) and A Critique of Chernoff Faces (http://eagereyes.org/criticism/chernoff-faces)

37

37

Outline

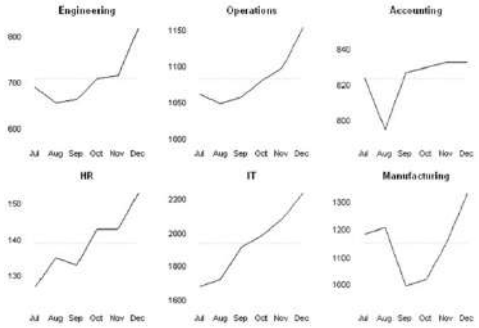
- What is high-dimensional data?
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

38

38

Small Multiples

- A series or grid of small similar graphics or charts for an easy exploration and comparison of multiple dimensions



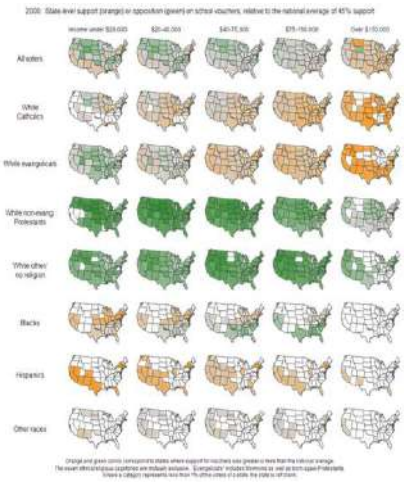
https://en.wikipedia.org/wiki/Small_multiple

39

39

Small Multiples

- Example: Use small multiples to visualize state-level support or opposition on the school vouchers across different races and in-come levels in the united states.



http://andrewgelman.com/2009/07/hard_sell_for_b/

40

40

Outline

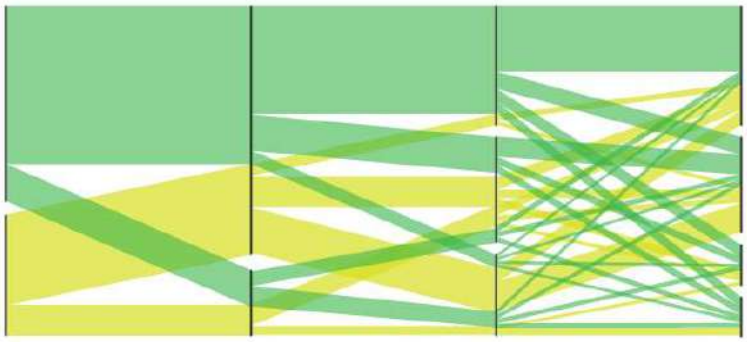
- What is high-dimensional data
- High-dimensional data visualization approaches
  - Scatter-plot Matrix
  - Parallel Coordinates
  - Glyph-based Methods
  - Small Multiples
  - Parallel Sets

41

41

Parallel Sets

- It is mainly designed for visualizing high-dimensional **categorical** data
- It shows data frequencies instead of the individual data points

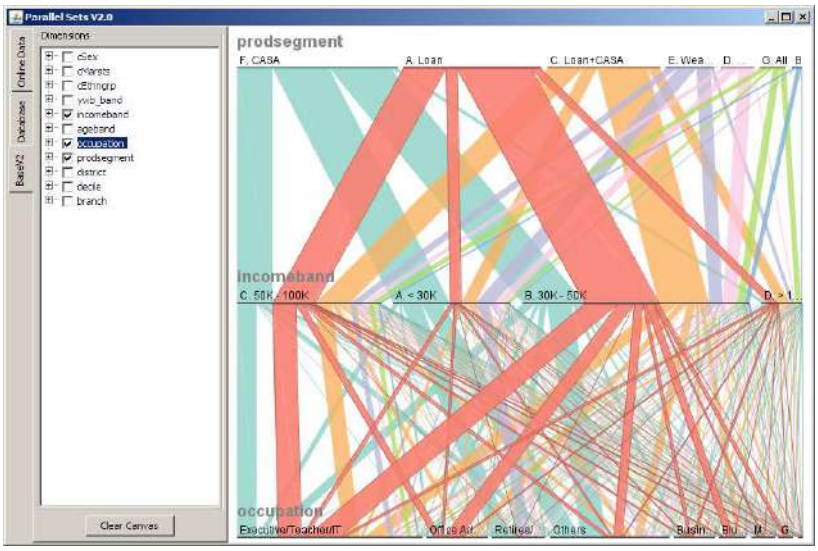


https://datavizcatalogue.com/methods/parallel_sets.html

42

42

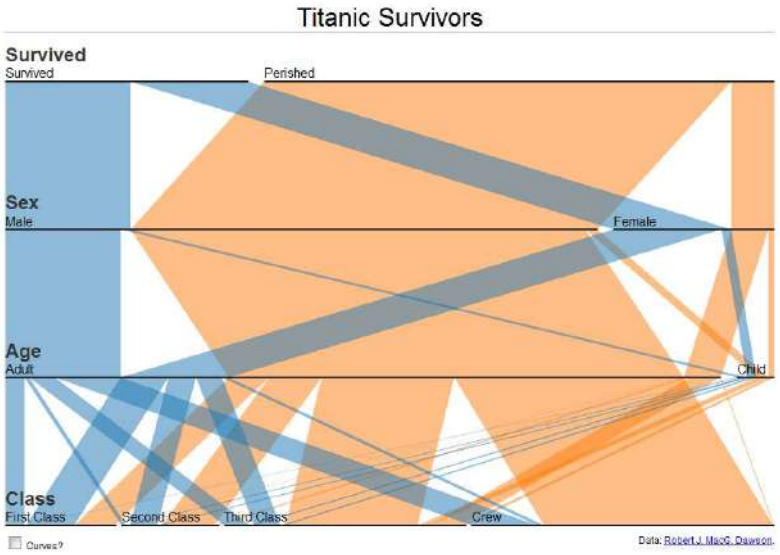
Exploring High-dimensional Categorical Data with Parallel Sets



43

43

Parallel Sets – d3.js



44

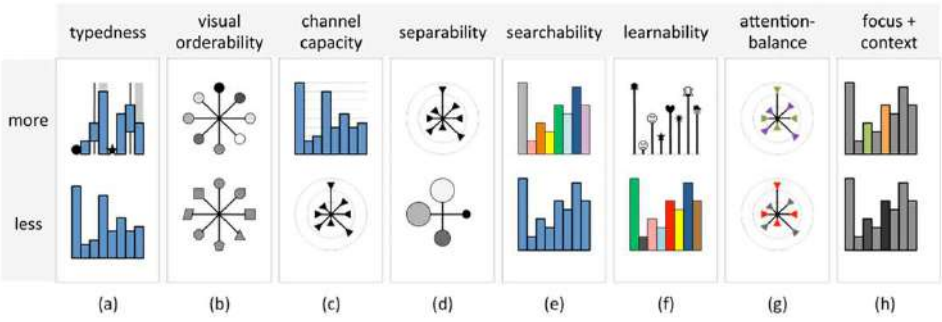
44

Questions?

Thank You!

45

A Survey on Glyph-based Visualization



Borgo R, Kehrer J et al. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. Eurographics (STARs) (2013)

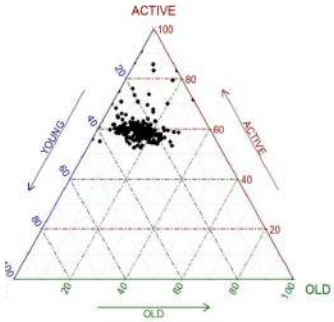
46

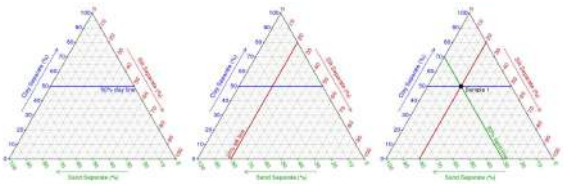
46

TERNARY PLOT

- Ternary plots are a way of displaying the distribution and variability of three-part compositional data.
- Its display is a triangle with three components. Each side represents one of the three components.

Population structure, 2015





Reference: https://en.wikipedia.org/wiki/Ternary_plot

47

47

Dimensionality Reduction

- Project the high-dimensional data onto a lower-dimensional subspace using linear or non-linear transformations.
- Projection preserves important relations (e.g., no information loss, data discrimination).

$$x = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_N \end{pmatrix} \rightarrow reduce\ dimensionality \rightarrow \hat{x} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{pmatrix} (K \ll N)$$

48

48

Dimensionality Reduction Methods

- Linear methods:
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
- Nonlinear methods:
  - * ISOMAP
  - * Local Linear Embedding (LLE)