Edinburgh Napier UNIVERSITY

Data Analytics
SET10109

A Visual Analysis Process

Natalie Kerracher
(n.kerracher@napier.ac.uk)

Edinburgh Napier
UNIVERSITY

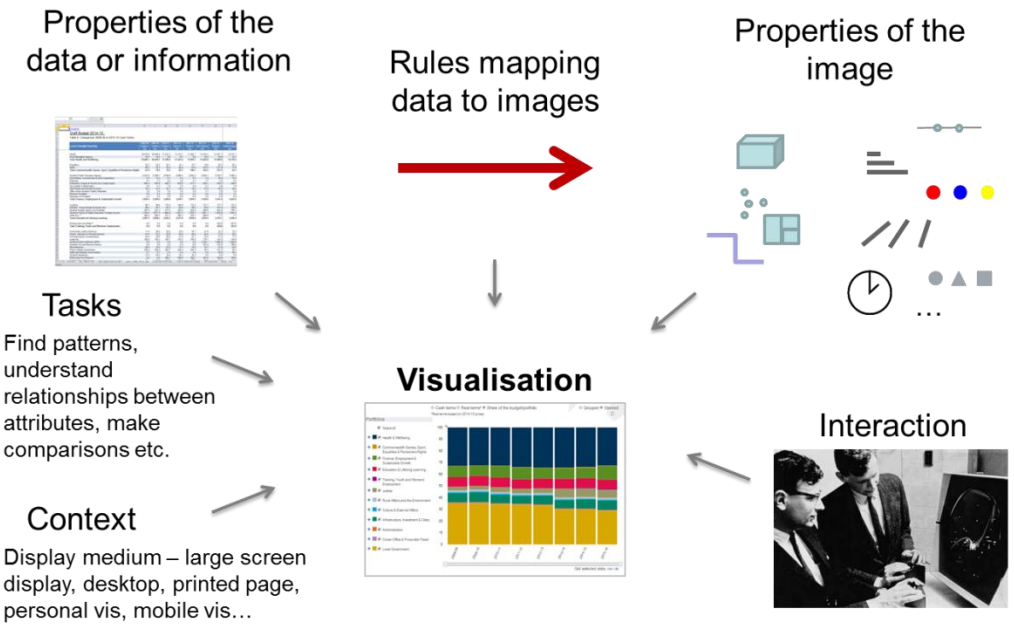
The

PROCESS OF VISUALIZATION

In this lecture…

- We will consider how we might apply what we've learned over in last seven lectures to the process of analysing an unfamiliar dataset

- There will also be a recap lecture in week 13



For reference:
what we've covered week by week

- Week 6:
  - Lecture 1: Introduction to visualisation
  - Lecture 2: Visual approaches for understanding data
- Week 9:
  - Lectures 3 & 4: Visualisation design principles
- Week 10:
  - Lectures 5 & 6: Interaction
- Week 11:
  - Lectures 7: Graphs and time
  - Lecture 8: Recap and design critique exercise

Edinburgh Napier UNIVERSITY

How would you carry out a visual analysis?

- There is no formal methodology of which I am aware for carrying out a visual analysis
- Related to Exploratory Data Analysis (EDA)

- This section will discuss two methodologies.
  - Berthold et al. 2010 – The course textbook
  - Chittaro 2006 – *ACM Computer*

Edinburgh Napier UNIVERSITY

When is a visualisation not appropriate?

- To read an individual value (tables are better)
- To compare two values (again tables are better)

- To answer a specific (statistical) question
  (an appropriate statistical test is more appropriate)

Edinburgh Napier UNIVERSITY

Exploratory Data Analysis

Deductive Method

Model -> Data -> Analysis -> Conclusion

Inductive Method (Exploratory Data Analysis)

Data -> Analysis -> Model -> Conclusions

Edinburgh Napier UNIVERSITY

Exploratory Data Analysis

- When to use
  - Some information is known about the data-set
  - Good contextual (background) information
  - When you want to examine the data to learn more about it (browsing to make discoveries/gain insight)
  - To narrow down the interesting parts of the data for closer inspection (when browsing a large data set)
- Use visual methods to explore the data.
  - And non-visual methods
- Search for hypotheses

Edinburgh Napier UNIVERSITY

Guidance on exploring an unfamiliar dataset

1. Initial question
2. Assess data
    1. Is the data appropriate (type, context, meaning)?
    2. Are the raw values appropriate?
    3. Is the data ready (see data preparation)?
    4. Does the data need reformulated?
    5. Is the initial question appropriate (reformulate goto step 1)
3. Construct Visualisation
    1. Does the visualisation answer the question (reformulate goto step 3)?
    2. Does it lead to further questions (new question from step 1)?

4. Conclusions

Process of Visualization

Edinburgh Napier UNIVERSITY

Chittaro 2006

1. Mapping
2. Selection
3. Presentation
4. Interactivity
5. Human factors
6. Evaluation

Berthold et al. 2010

1. Project Understanding
2. Data Understanding
3. Data preparation
4. Modelling
5. Evaluation
6. Deployment

Process of Visualization

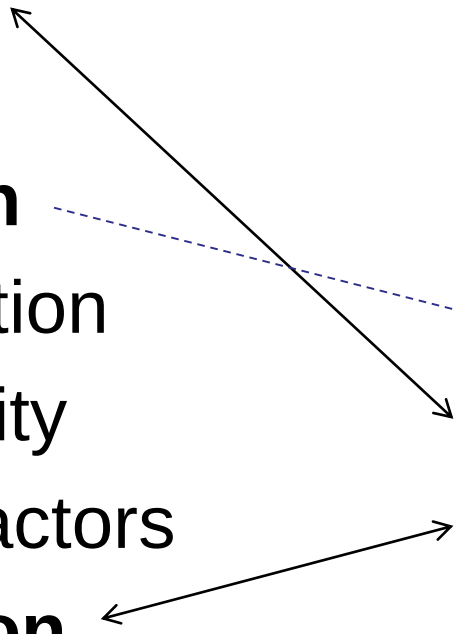
Edinburgh Napier UNIVERSITY

Chittaro 2006

1. **Mapping**
2. **Selection**
3. Presentation
4. Interactivity
5. Human factors
6. **Evaluation**

Berthold et al. 2010

1. Project Understanding
2. Data Understanding
3. **Data preparation**
4. **Modelling**
5. **Evaluation**
6. Deployment



Process of Visualization

Edinburgh Napier UNIVERSITY

Exploratory Data Analysis

Chittaro 2006

1. **Mapping**
2. **Selection**
3. Presentation
4. Interactivity
5. Human factors
6. **Evaluation**

Berthold et al. 2010

1. Project Understanding
2. **Data Understanding**
3. Data preparation
4. **Modelling**
5. **Evaluation**
6. Deployment

Edinburgh Napier UNIVERSITY

Complete Process of Visualization

1. Project Understanding
2. Data Understanding
3. Data preparation
4. Selection
5. Modelling
6. Human factors
7. Evaluation
8. Deployment

We are here!

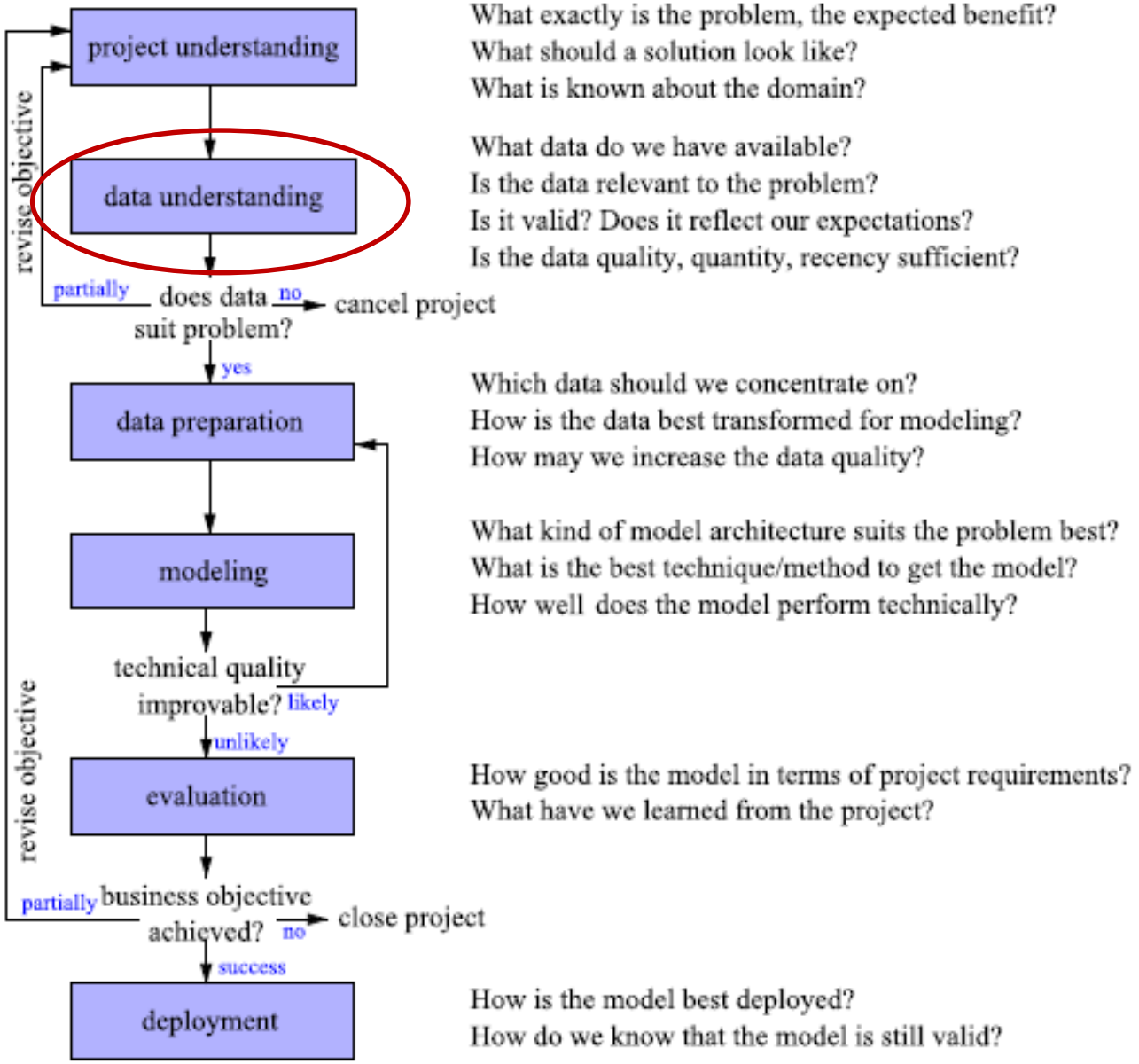


Fig. 1.1 Overview of the CRISP-DM process together with typical questions to be asked in the respective phases

Figure from Berthold et al. (2010), p9

Edinburgh Napier UNIVERSITY

Data Understanding (Recap)

First stage of a data analysis

Aims

- Understand the dataset
  - What is the range of the
  - Wat are the limitations of the data?
- Check the data
  - Is the data consistent with expectations?
- Understand relationships in the dataset
  - What links (if any) exist between items in the dataset

Revise Lecture 2.

Consider your data

Edinburgh Napier UNIVERSITY

- In lecture 2 we looked at data understanding [chapter 4 Berthold et al. (2011)]

- Berthold et al. give the following checklist for data understanding:
  - Determine the quality of the data
  - Find outliers
  - Detect and examine missing values
  - Discover new, or confirm expected dependencies or correlations between attributes
  - Check specific application dependent assumptions (e.g. the attribute follows a normal distribution)
  - Compare statistics with expected behaviour

- In particular, they note two “must dos”:
  - Check the distributions for each attribute
    (unexpected properties like outliers, correct domains, correct medians)
  - Check correlations or dependencies between pairs of attributes

From compendium slides for Guide to Intelligent Data Analysis, Springer 2011. c Michael R. Berthold, Christian Borgelt, Frank Höppner, Frank Klawonn and Iris Adä.(slides 44 & 45)

Edinburgh Napier UNIVERSITY

Finding Patterns/Problems (recap)

Finding outliers:

- Single variable: Histograms, Frequency bar charts, box plots
- 2 or more Variables: Scatterplots, Box Plot
- Many: Multi-dimensional scaling or PCA (see Berthold 2010)

Finding clusters

- 2 or more Variables: Scatterplots
- Many: Multi-dimensional scaling or PCA (see Berthold 2010)

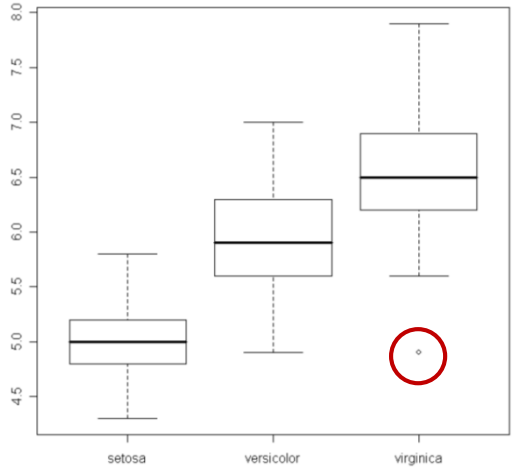
Finding relationships

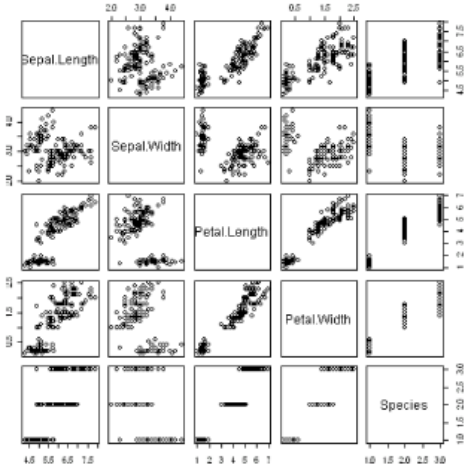
- 2 variables: Scatterplots
- Many: Matrix of Scatterplots

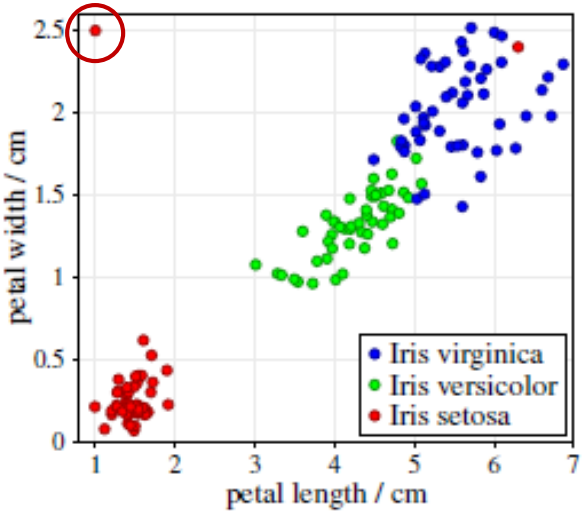
Visualisation for understanding data

We can use visualisation as a tool to help us understand our data:

- To understand the basic **structure** and **shape** of the data
- To discover **patterns** (e.g. correlations and dependencies)
- To spot peculiar **deviations** in the data set, such as outliers

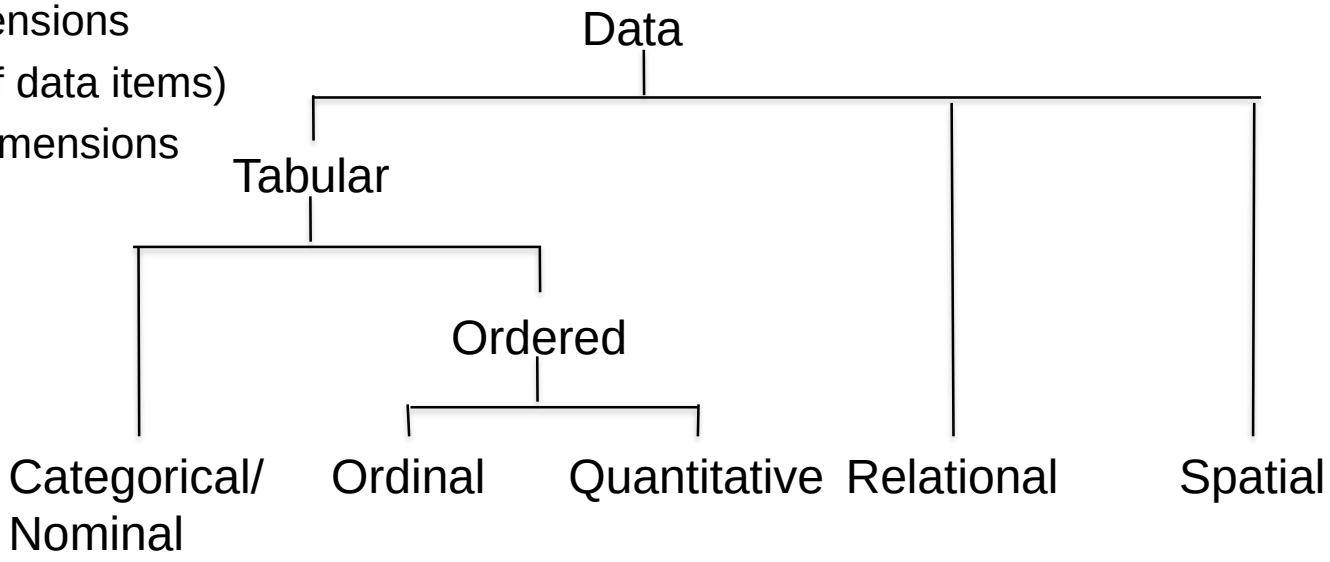






Look at the raw data

- Before we can decide how to visualise data, we need to know what sort of data we are dealing with [covered in lectures 2 and 3&4]
- Some considerations:
  - Type: tabular (categorical, ordinal, numeric); relational; spatial; temporal)
  - Continuous v discrete
  - Number of dimensions
  - Size (number of data items)
  - Cardinality of dimensions



Coursework tip: spend some time looking at the data; consider its size and number of dimensions: decide the data type, cardinality of dimensions etc. for each attribute,

Edinburgh Napier UNIVERSITY

Berthold et al.'s 'Must Do' 1: Check the distributions for each attribute

- Look at each attribute in turn
- Check for unexpected properties (outliers, correct domains, correct medians)
- Here we can use visualisation to:
  - understand the basic **structure** and **shape** of the data
  - spot peculiar **deviations** in the data set

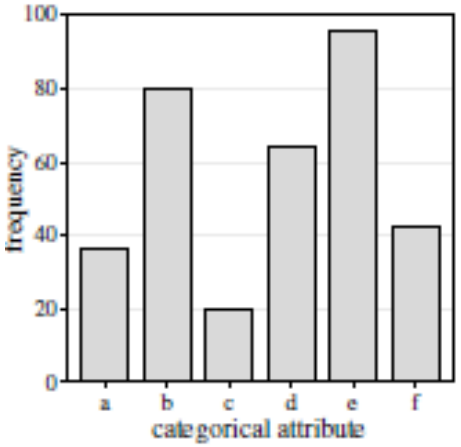
- In lecture 2, we looked at ways to do this based on data type and number of dimensions involved

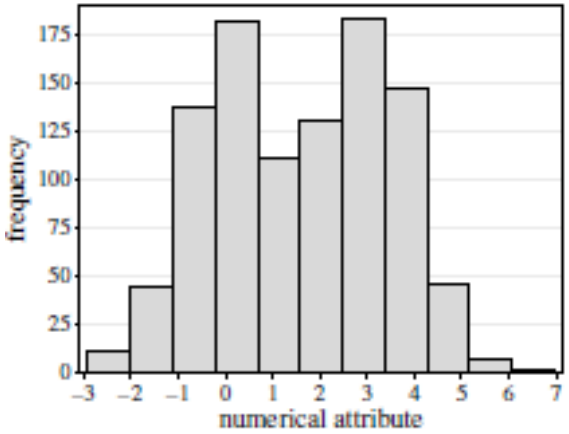
(NB rest of discussion based on tabular data)

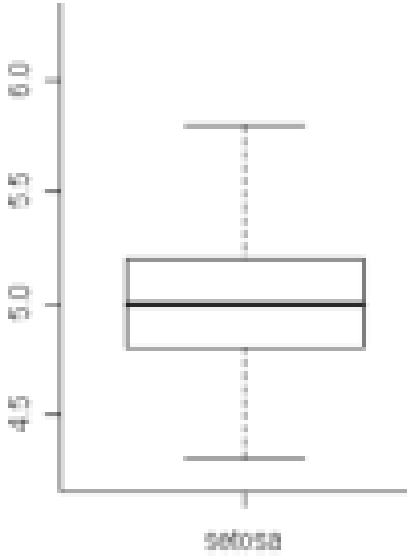
Berthold et al.'s 'Must Do' 1: Check the distributions for each attribute

Edinburgh Napier UNIVERSITY

Possibilities for visualising a single dimension:







Bar chart:
frequency of values of a **categorical** attribute

Histogram:
frequency distribution of a **numerical** attribute

Boxplot:
distribution of **numerical** attribute, its central value, and variability.

Images from compendium slides for Guide to Intelligent Data Analysis, Springer 2011. c Michael R. Berthold, Christian Borgelt, Frank Höppner, Frank Klawonn and Iris Adä.(slides 9,10, 16)

Process of Visualization

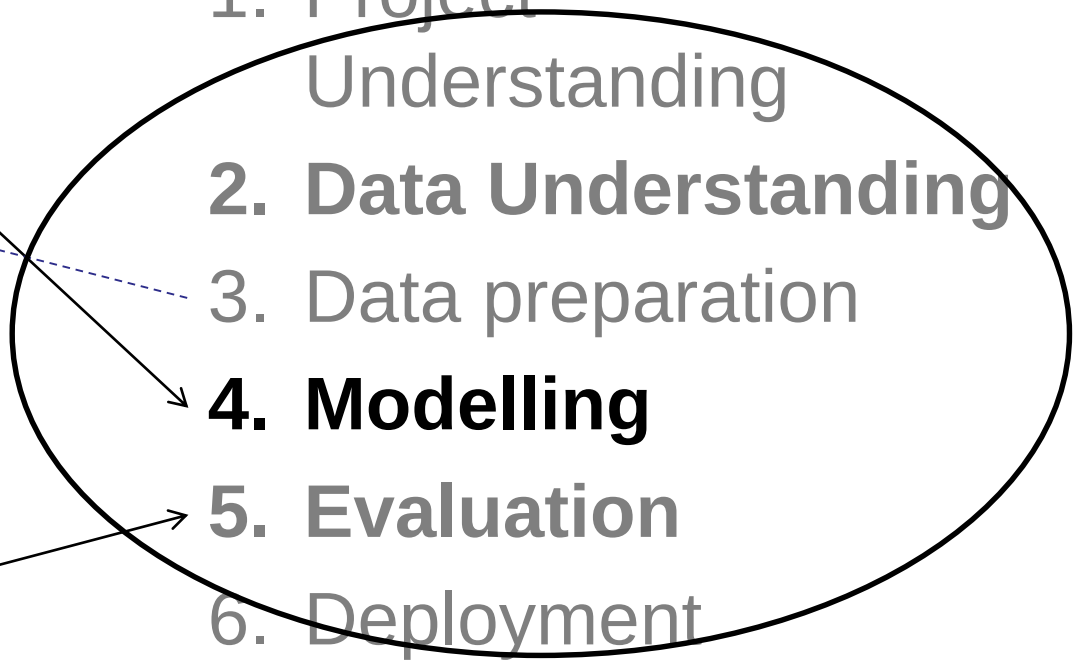
Edinburgh Napier UNIVERSITY

Chittaro 2006

1. **Mapping**
2. **Selection**
3. Presentation
4. Interactivity
5. Human factors
6. **Evaluation**

Berthold et al. 2010

1. Project Understanding
2. **Data Understanding**
3. Data preparation
4. **Modelling**
5. **Evaluation**
6. Deployment



Edinburgh Napier
UNIVERSITY

THINK OF A QUESTION…

Think of a question…

Edinburgh Napier UNIVERSITY

- Once we've considered the data in terms of its type and the individual attributes, we can ask some more interesting, relevant questions
- These are likely to concern relationships between attributes and between data items
- Visualisation can help us to discover patterns and relationships (e.g. correlations and dependencies, trends) and make comparisons

Example course work questions:

- Under which condition is it likely to have fatal accident? (correlation/dependency)
- The number of which kind of accident has decreased from year 2000 to year 2005? (temporal trend)
- What is the relationship between weather conditions and the severity of accidents on different types of road (relationship between multiple attributes)

Berthold et al.’s ‘Must Do’ 2: Check correlations or dependencies between pairs of attributes

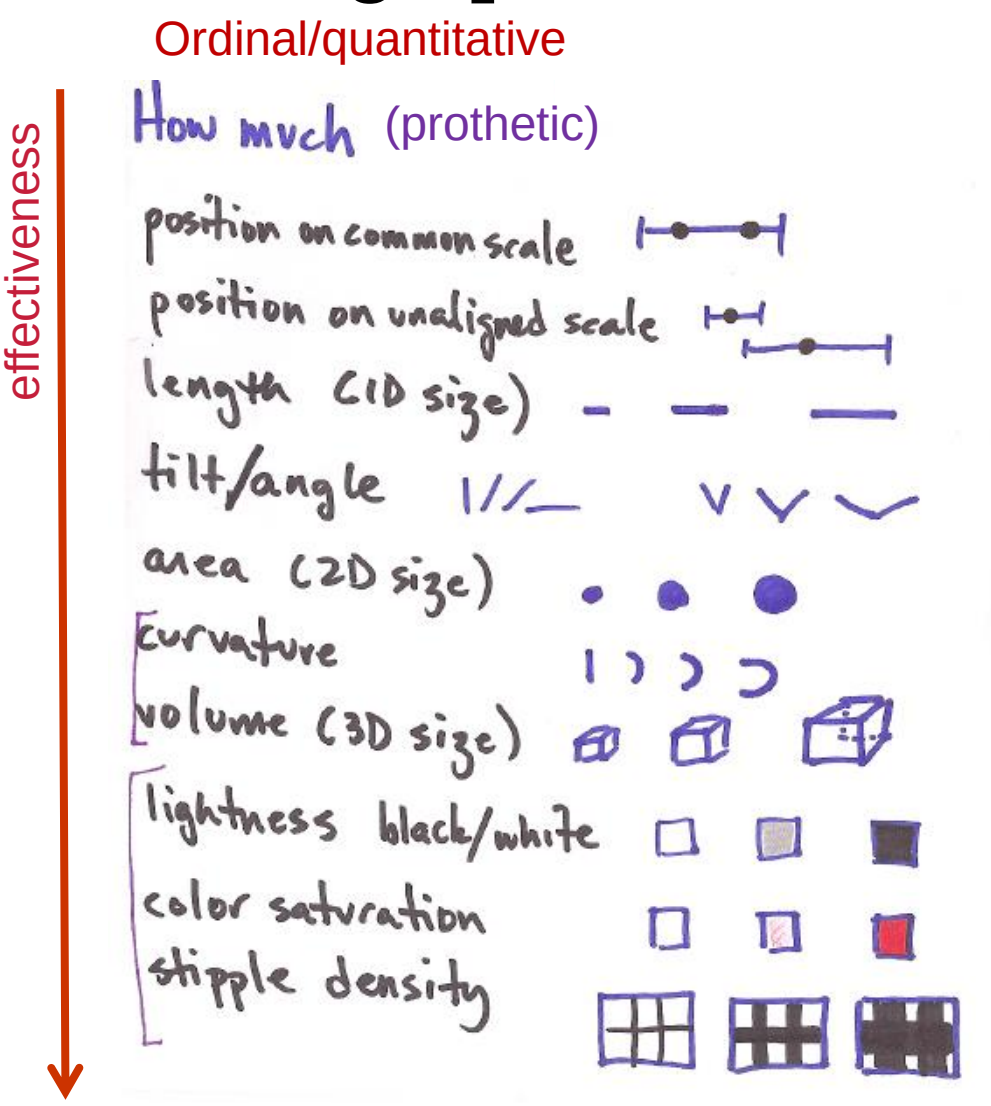
- Consider the attributes in relation to one another
- In lecture 2 we looked at layouts which can show multiple attributes
- In lectures 5 & 6 we looked at multiple views and interaction techniques to co-ordinate/link them

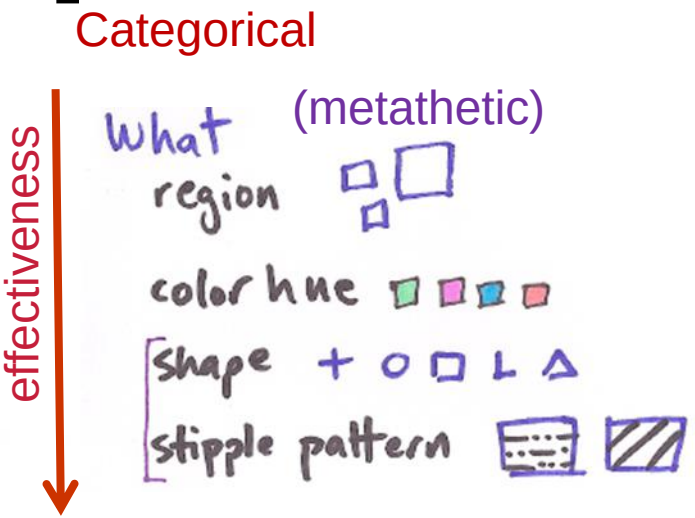
Edinburgh Napier
UNIVERSITY

SELECT APPROPRIATE VISUAL ENCODINGS

Select the most appropriate encodings [lectures 3 & 4]

Edinburgh Napier UNIVERSITY





Effectiveness principle: encode the most important information in the most "effective" way (accurate, discriminable, separable)

Spatial position (layout) is most salient –use this to encode the most important data to answer your question

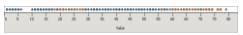
-

Source: Munzner (2014)

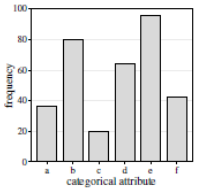
Some Common Layouts for Tabular Data

[Lecture 2]

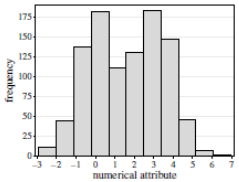
1D



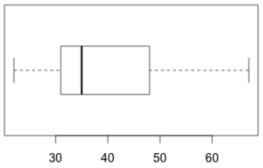
Scatter Line



Bar Chart

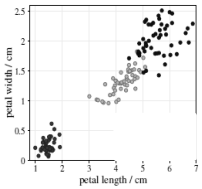


Histogram

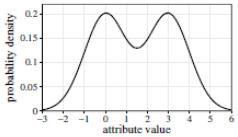


Boxplot

2D

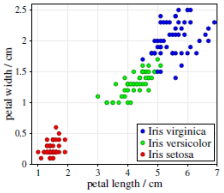


Scatter Plot



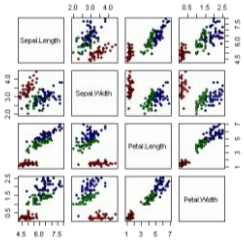
Line Chart

3D

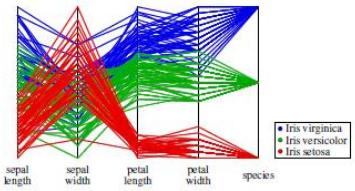


Scatter Plot +
1 non-spatial encoding

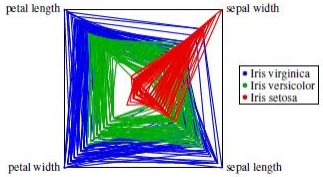
ND (3+)



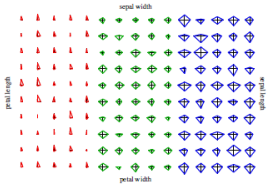
Scatter Plot Matrix



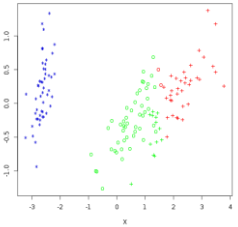
Parallel Co-ordinates



Radar Plot



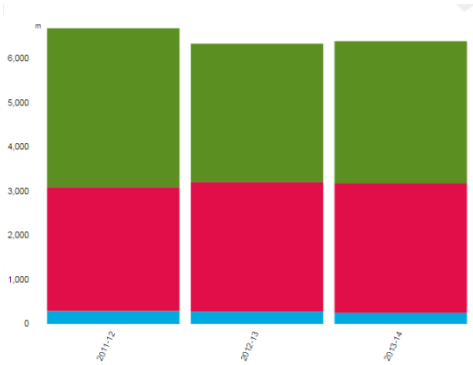
Star Plot



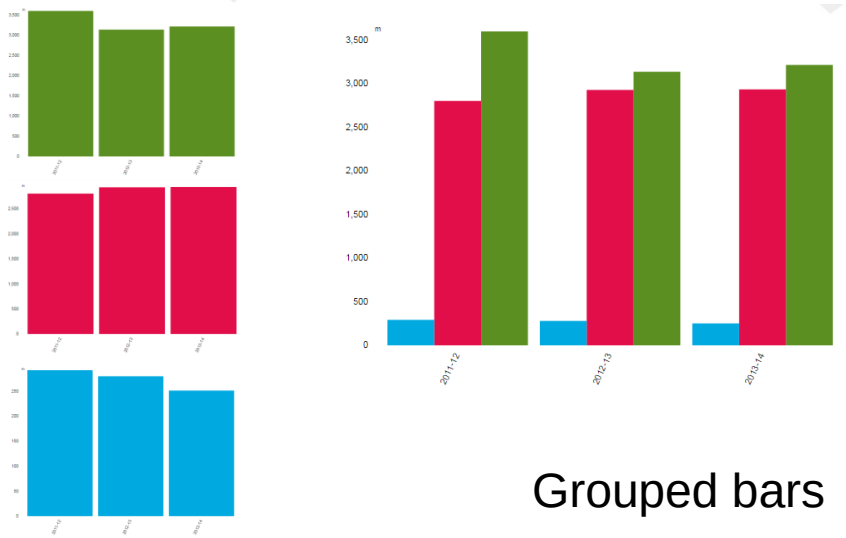
MDS

Multiple categorical attributes – some options

Edinburgh Napier UNIVERSITY

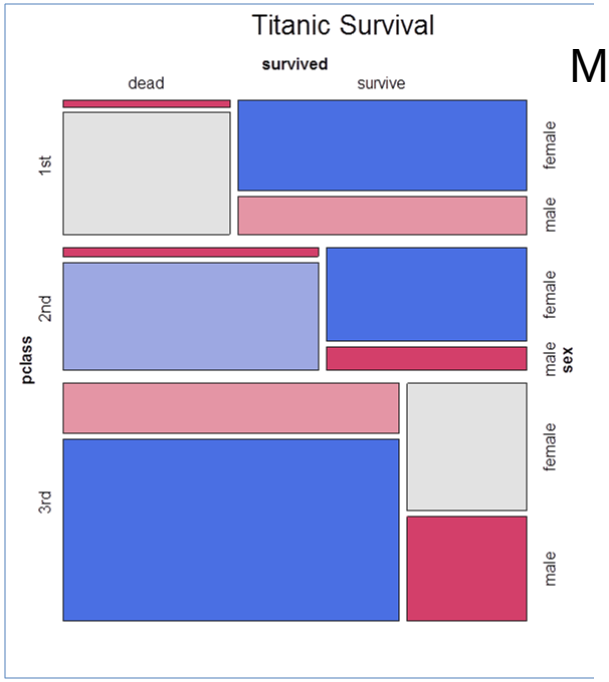


Stacked bars

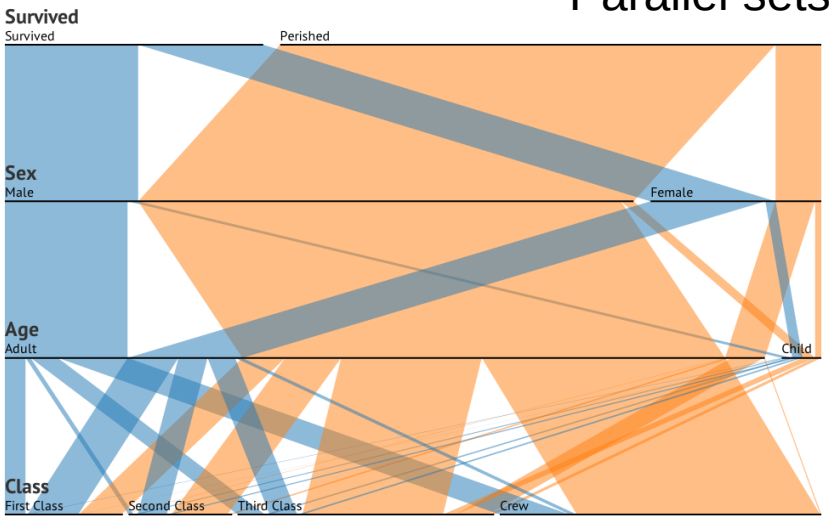


Grouped bars

Multiple/separate plots



Mosaic plot

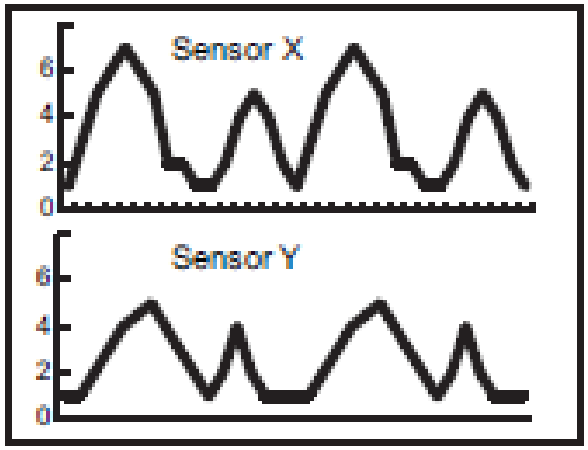


Parallel sets

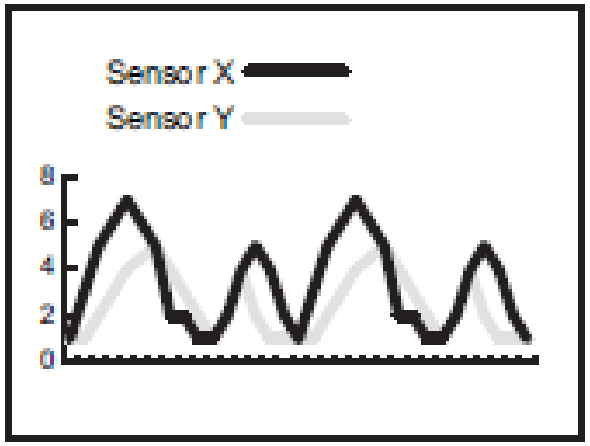
Parallel sets see: https://www.jasondavies.com/parallel-sets/

Visual comparison

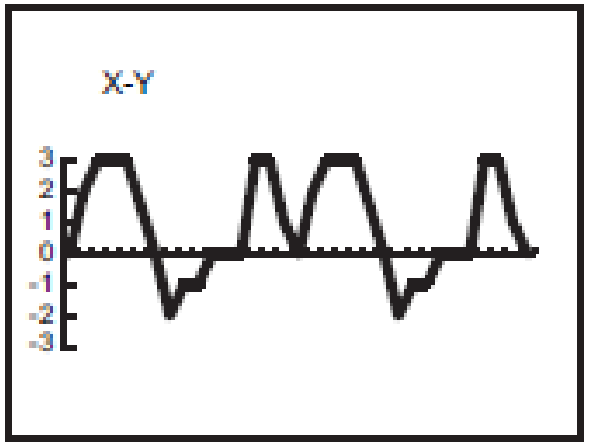
- Three basic categories of design for visual comparison (hybrids also possible):



a) Juxtaposition



b) Superposition



c) Explicit Encoding: Difference

Uses…

viewer's memory

visual system

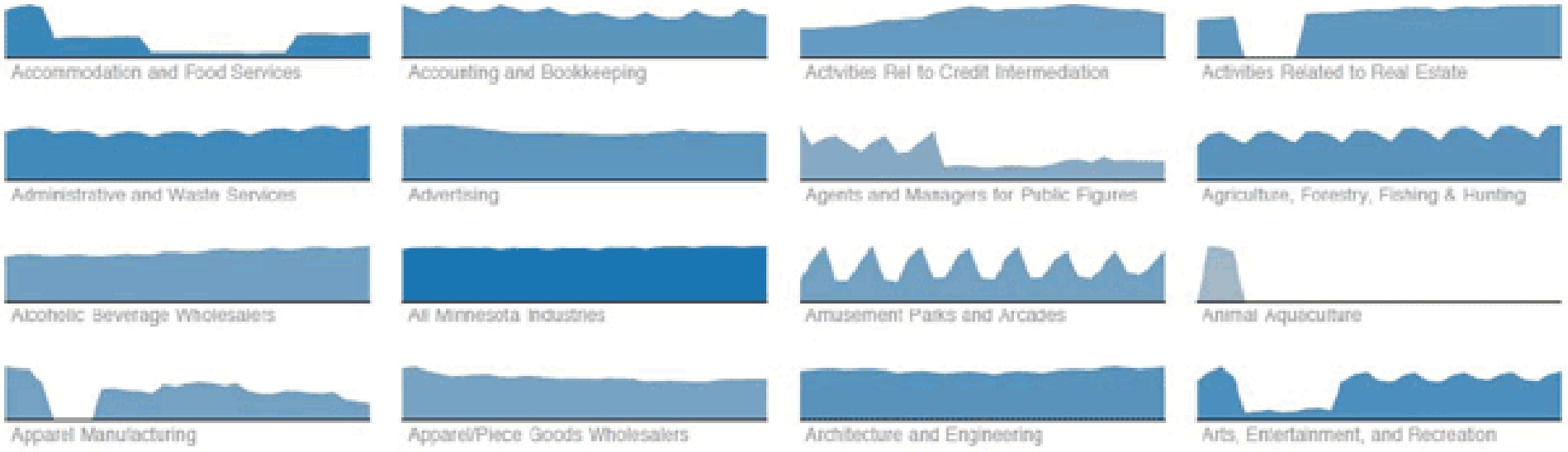
computation

(Gleicher et al., 2011)

…to determine relationships

Multiple views

- Small multiples (Tufte)
- Enable comparison across variables
- Discussed in lectures 5 & 6

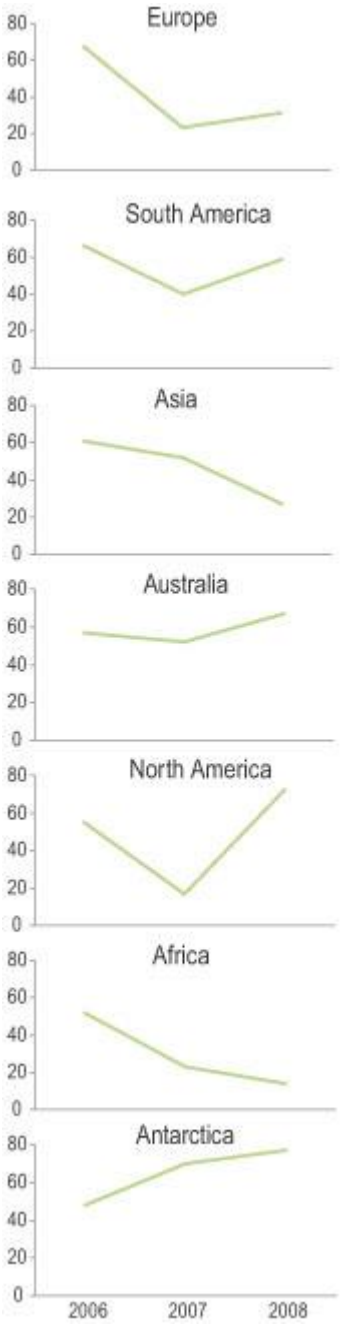


Heer and Shneiderman (2012) figure 11: employment figures by economic sector

Multiple views

Small multiples example - timeslices

Moderate to extreme drought in the US from 1899 to 2012, The New York Times
http://www.nytimes.com/interactive/2012/07/20/us/drought-footprint.html



Edinburgh Napier UNIVERSITY

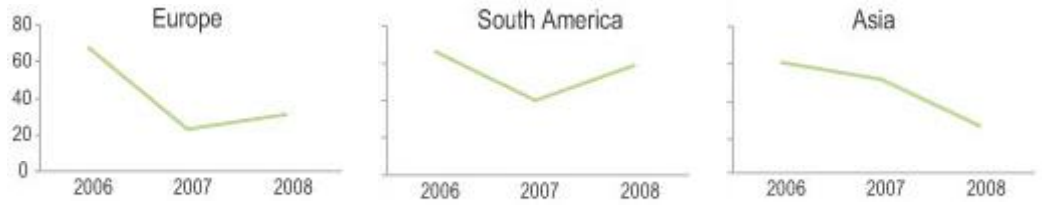
Juxtaposition and importance ordering

Example:

Small multiples arranged …

…vertically (aligned on x axis)
For comparison of patterns of change over time

…horizontally (aligned on y axis)
For comparison of magnitudes



Source: http://www.perceptualedge.com/example11.php

Multiple co-ordinated views

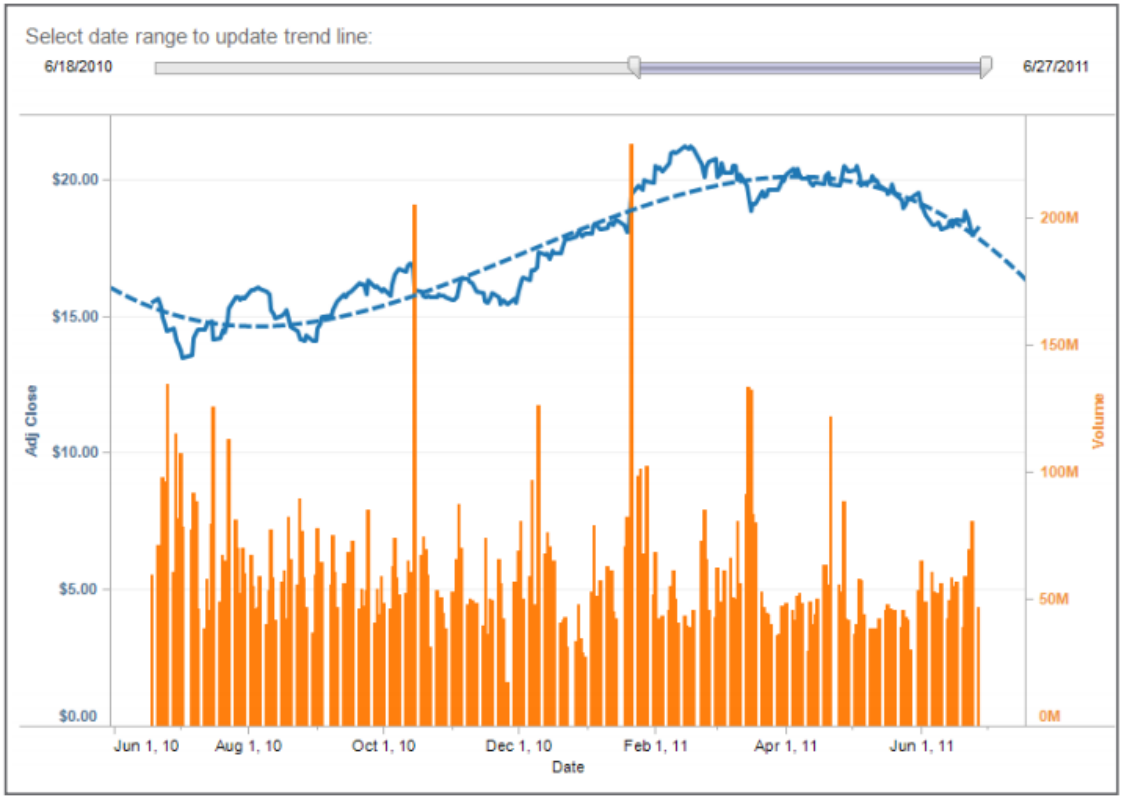
- Interactive changes made in one visualization are automatically reflected in the other visualizations
- Interactive linking techniques assist exploration across views
  - Brushing and linking – selecting in one view highlights (or hides/filters) corresponding data in other views
  - Linked navigation - co-ordinated scroll or zoom across views
- Allows analysts to see how patterns in one view project onto the others.

Selecting high-income players (top-right plot) shows little dependence on career length or fielding ability, but correlates with hitting performance (middle right plot).

Heer & Shneiderman (2012) baseball statistics example (images from Figure 13)

Combining views using superposition

Edinburgh Napier UNIVERSITY



Also possible to combine different plot types in the same space.

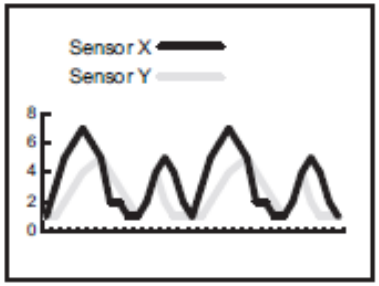
Line chart shows stock price over time

Bar chart shows volume sold per day

Allows us to see the relationship between two different attributes

Shows there were two significant events: one resulting in a sell-off and the other a gain for shareholders.

Source: Tableau (2012) Which chart or graph is right for you? Figure 5.
http://www.tableausoftware.com/learn/whitepapers/which-chart-or-graph-is-right-for-you



b) Superposition

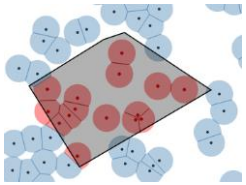
Edinburgh Napier
UNIVERSITY

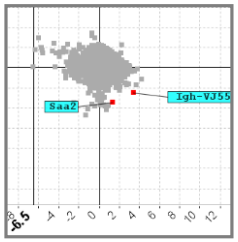
ALTER THE EXISTING OR CONSTRUCT A NEW VISUALISATION

Why interaction is important

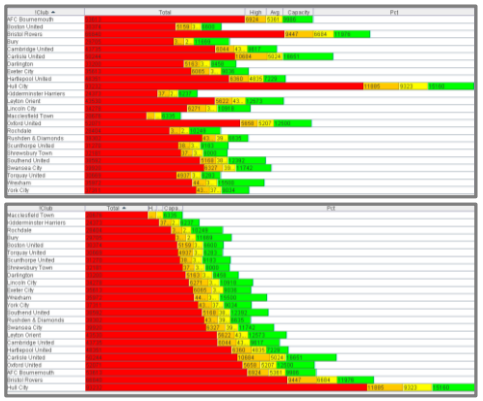
Edinburgh Napier UNIVERSITY

- Full understanding of large and complex datasets require more than one view of the data - interaction techniques allows us to:
  - Avoid “incomprehensible clutter” of showing everything simultaneously
  - Construct, link, and navigate between different views of the data, supporting the iterative process of exploration
  - Manipulate views to reveal patterns, possibly at different levels of granularity
  - Investigate hypotheses - ask “what-is” and “what-if” questions
  - Investigate relationships within the data, between items and between attributes

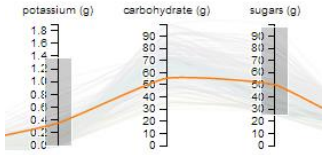


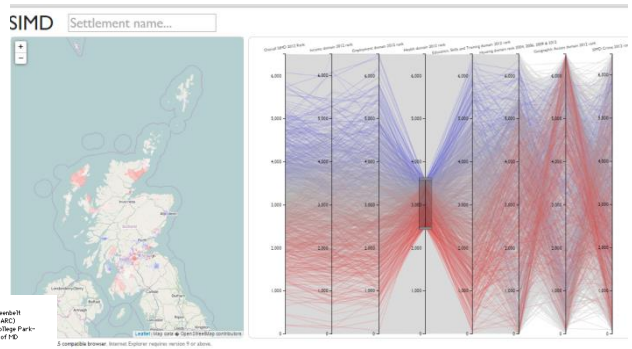


select

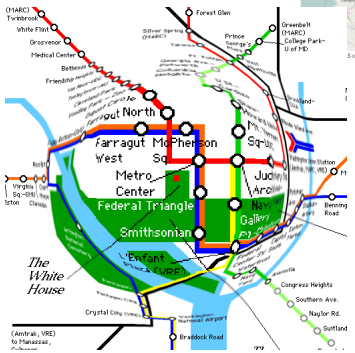


sort





filter, co-ordinate



focus + context

Edinburgh Napier UNIVERSITY

Alter the existing visualisation

Shneiderman's information seeking mantra:

Overview first, zoom and filter, then details on demand

(Shneiderman, 1996)

Interaction techniques [lectures 5&6]:

- **Navigate** (pan, zoom, focus+ context to maintain orientation) – look at a different portion of the data, in more or less detail
- **Filter** – show only (or highlight) data items that meet some condition
- **Aggregate** - show an abstracted view of the data to reveal structure/patterns at a higher level
- **Sort** – reorder items in the display to help surface patterns
- **Select** – an item to follow it in a changing display, show links, carry out further actions
- **Brushing and linking** – link multiple views through co-ordinated highlighting
- **Details on demand** – access individual values for specific data items

Coursework reminder…

Edinburgh Napier UNIVERSITY

2b. Use Tableau to visualise and explore the data set and discover interesting patterns and features in the data. Your tasks are:

(i) Before beginning the analysis, **formulate an initial question** that would be interesting to ask of the data. For example, you might ask “what is the relationship between weather conditions and the severity of accidents on different types of road?” This question should be included in your report. (1%)

(ii) Use Tableau to **gain a preliminary understanding of the data** e.g. check the distributions of each attribute and look for correlations/dependencies between pairs of attributes. You should include in the report:

a. Which layouts and visualisation encodings you used to carry out this stage of analysis.

b. Any interesting findings made.

c. **Any revisions you make to your initial question** in light of this step. (6%)

(iii) Sketch at least three **possible visualisation solutions** that could potentially be used to answer your (revised) question. Be sure to annotate your sketches with any interaction techniques required. Select one of these visual solutions to implement in Tableau. Include in your report:

a. A copy of your sketches (as an appendix).

b. A brief discussion of the layouts, encodings, and interaction techniques, considered in your sketches.

c. A discussion of your reasons for selecting your chosen visual solution over the others. (12%)

(iv) Implement your chosen visual solution using Tableau and use it to answer your question. Include in your report a discussion of any interesting discoveries that you were able to make and how your visual solution helped you to make these discoveries. Include screenshots of the visualisation as evidence of your findings (6%)

Link to sketching website

Roberts, Jonathan, Chris Headleand, and Panagiotis Ritsos. "Sketching designs using the Five Design-Sheet methodology." (2015).

Website: http://fds.design/

Sketching designs using the Five Design-Sheet methodology

Jonathan C. Roberts, *Member, IEEE*, Chris Headleand *Member, IEEE*, and Panagiotis D. Ritsos, *Member, IEEE*

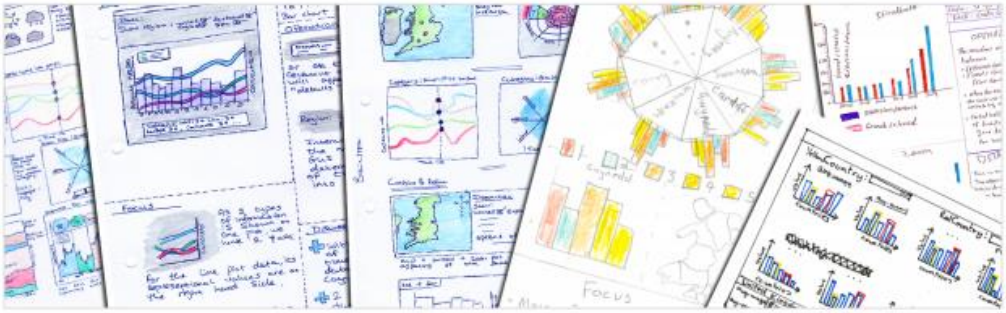


Fig. 1: The FdS consists of five sheets, each sheet with five parts. It is designed to help users consider alternative visualization designs; plan designs and create a realization sheet that can then be coded. Pictures of students' FdS sheets are shown.

Abstract— Sketching designs has been shown to be a useful way of planning and considering alternative solutions. The use of lo-fidelity prototyping, especially paper-based sketching, can save time, money and converge to better solutions more quickly. However, this design process is often viewed to be too informal. Consequently users do not know how to manage their thoughts and ideas (to first think divergently, to then finally converge on a suitable solution). We present the Five Design Sheet (FdS) methodology. The methodology enables users to create information visualization interfaces through lo-fidelity methods. Users sketch and plan their ideas, helping them express different possibilities, think through these ideas to consider their potential effectiveness as solutions to the task (sheet 1); they create three principle designs (sheets 2,3 and 4); before converging on a final realization design that can then be implemented (sheet 5). In this article, we present (i) a review of the use of sketching as a planning method for visualization and the benefits of sketching, (ii) a detailed description of the Five Design Sheet (FdS) methodology, and (iii) an evaluation of the FdS using the System Usability Scale, along with a case-study of its use in industry and experience of its use in teaching.

Index Terms—Lo-fidelity prototyping, User-centred design, Sketching for visualization, Ideation

Edinburgh Napier UNIVERSITY

IMPORTANT!!

WHAT YOU SHOULD HAVE LEARNED DURING THIS PART OF THE MODULE…

Edinburgh Napier UNIVERSITY

What you should take away from this section of the module…

- What visualisation is; why we would use it; when its use is appropriate and not appropriate.
- The differences between using visualisation for analysis and for presentation (in terms of audience, the purpose/goal of the visualisation, the techniques used).
- The properties of data that we need to consider when creating a visualisation.
- An awareness of the range of visual encodings (marks, channels, layouts) available when creating a visualisation.
- How to choose an appropriate visual encoding for the data: the considerations that we need to make when mapping data to a visual representation.
- Why interaction is important, an awareness of a variety of interaction techniques, and why and when to use them.

Edinburgh Napier
UNIVERSITY

QUESTIONS?

Edinburgh Napier UNIVERSITY

Some useful resources

- Glasgow University's STEPS project – basic overview of data types and charts http://www.stats.gla.ac.uk/steps/glossary/presenting_data.html
- Tableau whitepaper: Which chart or graph is right for you? http://www.tableausoftware.com/learn/whitepapers/which-chart-or-graph-is-right-for-you explains which chart to use when, and useful tips on combining chart types.
- Tableau online help for building each chart type: http://onlinehelp.tableausoftware.com/v8.0/pro/online/en-us/help.htm#dataview_examples.html%3FTocPath%3DExamples%7C_____0
- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Commun. ACM*, *53*(6), 59-67. Available at http://queue.acm.org/detail.cfm?id=1805128