Module INM433 – Visual Analytics

Lecture 01

Fundamentals of Visual Analytics

given by
prof. Gennady Andrienko and
prof. Natalia Andrienko
Content and objectives

• After defining what visual analytics is and why visualisation is important, the main principles of representing data visually will be introduced.

• Since the purpose of data visualisation is to support data analysis, we shall consider the general types of analysis tasks and the consequent requirements to analysis-supporting visual displays.

• After the lecture and following practical, you will be able to interpret the content of different types of data display and use various interactive operations designed to support data exploration.

• You will be acquainted with the software system V-Analytics, which will be used in the further study.
Definition of Visual Analytics
What did you learn earlier? (Module INM430 – Introduction to Data Science)

• Visualisation is important!

• Visualisation is used to
  • understand whether data contain what you need
  • uncover data imperfections (strange values, strange gaps, ...)
  • understand how to process the data to make them useful
  • check the results of data processing
  • compensate for limitations of descriptive statistics (recall Anscombe’s Quartet [http://en.wikipedia.org/wiki/Anscombe%27s_quartet]

• understand outcomes of statistical analysis, machine learning, other computational analysis methods

• Shortly: visualisation enables understanding!
Visualisation is essential for thinking

• Visualise = make perceptible to human’s mind

• “An estimated 50 percent of the brain's neurons are associated with vision. Visualisation <…> aims to put that neurological machinery to work.”

• “An abstractive grasp of structural features is the very basis of perception and the beginning of all cognition.”

⇒ The act of seeing already includes analytical activities (abstraction and feature extraction) and triggers all further mental processes.
Visual Analytics definition (reminder)

- The science of analytical reasoning* facilitated by interactive visual interfaces
- People use visual analytics tools and techniques to
  - synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
  - detect the expected and discover the unexpected
  - provide timely, defensible, and understandable assessments
  - communicate assessment effectively for action

*Analytical reasoning =
  data → information → knowledge → solution, decision, ...
  (interpreted data)

VA focuses on human reasoning

- VA deals with problems that cannot (yet) be solved algorithmically
  - ill-defined
  - involving incomplete and/or uncertain and/or conflicting data
- Human thinking is essential
  - pattern grasping, guessing, flexible use of previous knowledge and experience, novel approaches, trial and error
- Visualisation is essential for human thinking

- But: only human thinking is often insufficient
  - Cannot effectively cope with massive data amounts, high dimensionality, ...
  - May be too slow
Principle of Visual Analytics:

*Use the best of the humans and the computers*

**Computers**
- can store and process great amounts of data
- are very fast in searching information
- are very fast in data processing
- can interlink to extend their capacities
- can efficiently render high quality graphics

**Humans**
- are flexible and inventive, can deal with new situations and problems
- can solve problems that are hard to formalise
- can reasonably act in cases of incomplete and/or inconsistent information
- can simply *see* things that are hard to compute
Visual Analytics technology:

• combine visual and computational analysis methods

Goal: divide the labour between humans and computers so as to enable their synergistic work.
Visual Analytics

a summary (1)

• VA develops methods and technologies for solving complex problems requiring joint efforts of humans and computers
  • Ill-defined, hardly formalisable problems; incomplete and/or inconsistent data; new situations → humans
  • Massive data amounts, high dimensionality, rapid growth → computers
• VA may thus be defined as
  the science of human-computer data analysis, knowledge building, and problem solving
Visual Analytics

a summary (2)

• VA technology combines interactive visualisations with computational processing
  • transformations, database querying, data mining algorithms, statistics, geographical analysis methods, ...

• VA aims at effective division of labour between humans and machines
  • Visual representations are the most effective means to convey information to human’s mind and prompt human cognition and reasoning
  • Computational power must amplify humans’ inherent perceptual and cognitive capabilities
Questions?

Definition of Visual Analytics
Visualisation primitives
Components of a visual display

• Display space (or visual space): set of positions

• Visual marks: points, geometric shapes, symbols positioned in the display space

• Types of marks:
  • Points
  • Lines
  • Areas
  • Surfaces
  • Volumes

• Visual properties of marks: colour, shape, orientation, size, texture
Visual variables

- Position in the display space (x,y)
- Size
- Value, or degree of darkness
- Texture
- Colour (hue)
- Orientation
- Shape

Perceptual properties of visual variables
(Bertin: level of organisation)

• **Selection**: easy to locate marks with a given value of the visual variable and perceive them separately from others

• **Association**: several marks with the same value of this visual variable can be easily grouped (perceived all together) despite differences in other visual variables

• **Order**: the values of the variable can be put in an order, e.g., from small to big, from light to dark, ...

• **Quantity**: differences between two values of the variable can be interpreted numerically, e.g., how much bigger

• **Perceptual length**: number of different values of the variable that can be easily distinguished by an average human
Selectivity

• Can attention be focused on one value of the variable, excluding other variables and values?

• Find all letters on the left half of the image (position)
• Find all small letters (size)
• Find all red letters (colour)
• Find all letters ‘K’ (shape)

Easy to find ⇒ selective variable
Associativity

• Can marks with the same value of the visual variable be perceived simultaneously?

• Group all letters located on the left half of the image (position)
• Group all small letters (size)
• Group all red letters (colour)
• Group all letters ‘K’ (shape)

Easy to group ⇒ associative variable
Visual variable ‘position’

- Selection
- Association
- Order
- Quantity
- Length
“Retinal” visual variables

- Size
- Value
- Texture
- Colour
- Orientation
- Shape

Diagram showing associations and quantities for each visual variable.
Perceptual properties of visual variables

<table>
<thead>
<tr>
<th></th>
<th>Association</th>
<th>Selection</th>
<th>Order</th>
<th>Quantity</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>limited by display size</td>
</tr>
<tr>
<td>Size</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>5-7</td>
</tr>
<tr>
<td>Value (darkness)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>5-7</td>
</tr>
<tr>
<td>Texture</td>
<td>+</td>
<td>+</td>
<td>+/- *</td>
<td>-</td>
<td>5-7</td>
</tr>
<tr>
<td>Colour</td>
<td>+</td>
<td>+</td>
<td>-/+ **</td>
<td>-</td>
<td>7-8</td>
</tr>
<tr>
<td>Orientation</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Shape</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5-7</td>
</tr>
</tbody>
</table>

* + : grain density or size; - : grain shape
** + for parts of the spectrum (e.g., cold – hot)
The fundamental rule of visual mapping

• Visual mapping: the process of mapping data components to visual variables

• The fundamental rule: the perceptual properties of the visual variables must correspond to the properties of the data components they represent
Types and properties of data components

Types of values:
- Numeric
- Textual
  - Predefined values (e.g., codes)
  - Free text
- Spatial
  - Coordinates
  - Place names
  - Addresses
- Temporal
  - Other (image, video, audio, ...)

Scales of measurement*:
- Nominal (¬ order, ¬ distances)
  - gender, nationality, ...
- Ordinal (✓ order, ¬ distances)
  - evaluations: bad, fair, good, excellent
- Interval (✓ order, ✓ distances, ¬ ratios, ¬ meaningful zero)
  - temperature, time, ...
- Ratio (✓ order, ✓ distances, ✓ ratios, ✓ meaningful zero)
  - quantities, distances, durations, ...

<table>
<thead>
<tr>
<th>Name</th>
<th>Birth date</th>
<th>School grade</th>
<th>Address</th>
<th>Distance to school, m</th>
<th>Getting to school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>17/05/2005</td>
<td>3</td>
<td>12, Pine street</td>
<td>850</td>
<td>by bus</td>
</tr>
<tr>
<td>Julia</td>
<td>23/08/2004</td>
<td>4</td>
<td>9, Oak avenue</td>
<td>400</td>
<td>on foot</td>
</tr>
<tr>
<td>Paul</td>
<td>10/12/2005</td>
<td>2</td>
<td>56, Maple road</td>
<td>1500</td>
<td>by car</td>
</tr>
<tr>
<td>Mary</td>
<td>06/10/2003</td>
<td>5</td>
<td>71, Linden lane</td>
<td>900</td>
<td>on foot</td>
</tr>
</tbody>
</table>

Correspondence of visual variables to types and scales of data components

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
<th>Ordinal</th>
<th>Interval</th>
<th>Ratio</th>
<th>Spatial</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Size</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Value (darkness)</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Texture</td>
<td>+</td>
<td>+ *</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Colour</td>
<td>+</td>
<td>+ **</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Orientation</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Shape</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* grain density or size  ** parts of the spectrum (e.g., cold – hot)
Extensions and refinements of Bertin’s theory

Suitability for representing numbers


Questions?

Visualisation primitives
Data analysis tasks (briefly)
Data analysis tasks

• Data reflect some pieces of reality. We analyse data to learn something about the reality.

data → information → knowledge → solution, decision ...

• Analysis can be seen as a sequence of tasks.

• Each task is an attempt to obtain a certain kind of information or knowledge about the reality from the available data.

• It can be seen as finding an answer to some question about the reality.

• Tasks can be distinguished according to the kind of information or knowledge they seek to obtain.
Task types

- **Bertin**: There are as many possible question types as components in the data, i.e., each component may be a target of a question (task).

  - Who lives on Pine street? **When** was Paul born? **What** is Julia’s school grade? **Where** does Mary live? **How far** is the school from Peter’s home? **How** does Mary get to the school?

<table>
<thead>
<tr>
<th>Name</th>
<th>Birth date</th>
<th>School grade</th>
<th>Address</th>
<th>Distance to school, m</th>
<th>Getting to school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>17/05/2005</td>
<td>3</td>
<td>12, Pine street</td>
<td>850</td>
<td>by bus</td>
</tr>
<tr>
<td>Julia</td>
<td>23/08/2004</td>
<td>4</td>
<td>9, Oak avenue</td>
<td>400</td>
<td>on foot</td>
</tr>
<tr>
<td>Paul</td>
<td>10/12/2005</td>
<td>2</td>
<td>56, Maple road</td>
<td>1500</td>
<td>by car</td>
</tr>
<tr>
<td>Mary</td>
<td>06/10/2003</td>
<td>5</td>
<td>71, Linden lane</td>
<td>900</td>
<td>on foot</td>
</tr>
</tbody>
</table>
Task levels

**Bertin:**
- **Elementary**: question (task) addresses one or several *individual data items*, e.g., How do Peter and Julia get to the school?
- **Overall**: question (task) addresses the *whole set* of data items, e.g., What are the methods of getting to the school? Which method is the most frequent?
- **Intermediate**: question (task) addresses one or several *subsets of data items*, e.g., Which method of getting to school is more frequently used by girls?

**Synoptic tasks**: overall + intermediate
- Distinguishing property: a set or subset of data items is considered in its entirety; individual items are not of interest.
  ⇒ Synoptic tasks entail *abstraction* from individual items to sets
Semantic roles of data components

- **Reference**: What is described?

- **Characteristic**: What is known about it?

<table>
<thead>
<tr>
<th>Name</th>
<th>Birth date</th>
<th>School grade</th>
<th>Address</th>
<th>Distance to school, m</th>
<th>Getting to school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>17/05/2005</td>
<td>3</td>
<td>12, Pine street</td>
<td>850</td>
<td>by bus</td>
</tr>
<tr>
<td>Julia</td>
<td>23/08/2004</td>
<td>4</td>
<td>9, Oak avenue</td>
<td>400</td>
<td>on foot</td>
</tr>
<tr>
<td>Paul</td>
<td>10/12/2005</td>
<td>2</td>
<td>56, Maple road</td>
<td>1500</td>
<td>by car</td>
</tr>
<tr>
<td>Mary</td>
<td>06/10/2003</td>
<td>5</td>
<td>71, Linden lane</td>
<td>900</td>
<td>on foot</td>
</tr>
</tbody>
</table>

Referential component (referrer)

Characteristic components (attributes)
Data may have >1 referrers

<table>
<thead>
<tr>
<th>Year</th>
<th>Referrer 1: Time</th>
<th>Referrer 2: Place</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960</td>
<td>Alabama</td>
<td>3256740</td>
<td>39920</td>
</tr>
<tr>
<td>1960</td>
<td>Alaska</td>
<td>221607</td>
<td>3730</td>
</tr>
<tr>
<td>1960</td>
<td>Arizona</td>
<td>1302161</td>
<td>39243</td>
</tr>
<tr>
<td>1960</td>
<td>Arkansas</td>
<td>1795272</td>
<td>18472</td>
</tr>
<tr>
<td>1960</td>
<td>California</td>
<td>15717204</td>
<td>546069</td>
</tr>
<tr>
<td>1960</td>
<td>Colorado</td>
<td>1753947</td>
<td>38103</td>
</tr>
<tr>
<td>1960</td>
<td>Connecticut</td>
<td>2358234</td>
<td>29321</td>
</tr>
<tr>
<td>1960</td>
<td>Delaware</td>
<td>445292</td>
<td>9642</td>
</tr>
<tr>
<td>1960</td>
<td>District of Co</td>
<td>753956</td>
<td>20725</td>
</tr>
<tr>
<td>1960</td>
<td>Florida</td>
<td>4951560</td>
<td>133519</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Referrer 1: Time</th>
<th>Referrer 2: Place</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1972</td>
<td>West Virginia</td>
<td>1781000</td>
<td>25584</td>
</tr>
<tr>
<td>1972</td>
<td>Wisconsin</td>
<td>4520000</td>
<td>133382</td>
</tr>
<tr>
<td>1972</td>
<td>Wyoming</td>
<td>3450000</td>
<td>10461</td>
</tr>
<tr>
<td>1973</td>
<td>Alabama</td>
<td>3539000</td>
<td>91389</td>
</tr>
<tr>
<td>1973</td>
<td>Alaska</td>
<td>3300000</td>
<td>16313</td>
</tr>
<tr>
<td>1973</td>
<td>Arizona</td>
<td>2058000</td>
<td>137566</td>
</tr>
<tr>
<td>1973</td>
<td>Arkansas</td>
<td>2037000</td>
<td>56149</td>
</tr>
<tr>
<td>1973</td>
<td>California</td>
<td>20601000</td>
<td>1298872</td>
</tr>
<tr>
<td>1973</td>
<td>Colorado</td>
<td>2437000</td>
<td>133953</td>
</tr>
<tr>
<td>1973</td>
<td>Connecticut</td>
<td>3076000</td>
<td>112717</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Referrer 1: Time</th>
<th>Referrer 2: Place</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>Rhode Island</td>
<td>1048319</td>
<td>36444</td>
</tr>
<tr>
<td>2000</td>
<td>South Carolina</td>
<td>4012012</td>
<td>209482</td>
</tr>
<tr>
<td>2000</td>
<td>South Dakota</td>
<td>7548444</td>
<td>17511</td>
</tr>
<tr>
<td>2000</td>
<td>Tennessee</td>
<td>5669283</td>
<td>278218</td>
</tr>
<tr>
<td>2000</td>
<td>Texas</td>
<td>20851820</td>
<td>1033881</td>
</tr>
<tr>
<td>2000</td>
<td>Utah</td>
<td>2233169</td>
<td>99958</td>
</tr>
<tr>
<td>2000</td>
<td>Vermont</td>
<td>608887</td>
<td>13185</td>
</tr>
<tr>
<td>2000</td>
<td>Virginia</td>
<td>7078513</td>
<td>214348</td>
</tr>
<tr>
<td>2000</td>
<td>Washington</td>
<td>5894121</td>
<td>300932</td>
</tr>
<tr>
<td>2000</td>
<td>West Virginia</td>
<td>1808344</td>
<td>47067</td>
</tr>
<tr>
<td>2000</td>
<td>Wisconsin</td>
<td>5363675</td>
<td>172124</td>
</tr>
<tr>
<td>2000</td>
<td>Wyoming</td>
<td>4937923</td>
<td>16285</td>
</tr>
</tbody>
</table>
Data may have >1 referrers

Please note: data structure ≠ table structure!

It is not necessary that there is exactly one table column for each referrer.

1. Values from two or more table columns may specify the same reference in a complementary way, as in this example.
2. There may be no special column with references
   - Implicit references: when data describe objects with no identifiers assigned (e.g., lightning strokes), it is assumed that each table row refers to a different object.
3. Table columns may be associated with different references. E.g., the data from this table can be put in an equivalent table where columns correspond to different combinations (attribute, year). This does not change the structure of the data! See the next slide.
References are not always in columns

<table>
<thead>
<tr>
<th>Referrer 1: place</th>
<th>Referrer 2: time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbey</td>
<td>E05000026</td>
</tr>
<tr>
<td>Allerton</td>
<td>E05000027</td>
</tr>
<tr>
<td>Bexley</td>
<td>E05000028</td>
</tr>
<tr>
<td>Chadwell Heath</td>
<td>E05000029</td>
</tr>
<tr>
<td>Eastbrook</td>
<td>E05000030</td>
</tr>
<tr>
<td>Eastbury</td>
<td>E05000031</td>
</tr>
<tr>
<td>Gosforth</td>
<td>E05000032</td>
</tr>
<tr>
<td>Gomersby</td>
<td>E05000033</td>
</tr>
<tr>
<td>Heath</td>
<td>E05000034</td>
</tr>
<tr>
<td>London Bridge</td>
<td>E05000035</td>
</tr>
<tr>
<td>Mayesbrook</td>
<td>E05000036</td>
</tr>
<tr>
<td>Parishes</td>
<td>E05000037</td>
</tr>
<tr>
<td>River</td>
<td>E05000038</td>
</tr>
<tr>
<td>Thames</td>
<td>E05000039</td>
</tr>
<tr>
<td>Valentia</td>
<td>E05000040</td>
</tr>
<tr>
<td>Village</td>
<td>E05000041</td>
</tr>
<tr>
<td>Wharncliffe</td>
<td>E05000042</td>
</tr>
<tr>
<td>Brunswick Park</td>
<td>E05000043</td>
</tr>
<tr>
<td>Burnt Oak</td>
<td>E05000044</td>
</tr>
<tr>
<td>Chilworth</td>
<td>E05000045</td>
</tr>
<tr>
<td>Colindale</td>
<td>E05000046</td>
</tr>
<tr>
<td>Coppetts</td>
<td>E05000047</td>
</tr>
<tr>
<td>East Barnet</td>
<td>E05000048</td>
</tr>
<tr>
<td>East Finchley</td>
<td>E05000049</td>
</tr>
<tr>
<td>Edgeware</td>
<td>E05000050</td>
</tr>
<tr>
<td>Finchley Church</td>
<td>E05000051</td>
</tr>
<tr>
<td>Garden Suburb</td>
<td>E05000052</td>
</tr>
<tr>
<td>Golder Green</td>
<td>E05000053</td>
</tr>
<tr>
<td>Hale</td>
<td>E05000054</td>
</tr>
<tr>
<td>Hendon</td>
<td>E05000055</td>
</tr>
<tr>
<td>High Barnet</td>
<td>E05000056</td>
</tr>
<tr>
<td>Hill Hill</td>
<td>E05000057</td>
</tr>
<tr>
<td>Oakleigh</td>
<td>E05000058</td>
</tr>
</tbody>
</table>
Data components viewed as variables

• **Referrers** ~ independent variables

• **Attributes** ~ dependent variables

• Data represent (are generated by) a *function*  
  **Referrers** → **Attributes**
  
  • References (values of referrers): function’s inputs
  • Characteristics (values of attributes): function’s outputs

• Function’s *behaviour*:  
  how the outputs vary over the set of inputs → how the characteristics vary over the set of references
  
  • E.g., how the crime rates vary over space and time
Synoptic tasks ~
tasks addressing behaviours

• Describe the behaviour of one or more attributes

• Find subsets of references where attributes have particular behaviours

• Compare two or more behaviours (find similarities and differences)
  • Different attributes over the same set of references
  • Same attributes over different subsets of references

• Relate behaviours of two or more attributes
Classes of synoptic tasks with examples

• Describe the behaviour of one or more attributes
  • Describe the variation of the crime rates over time and space

• Find subsets of references where attributes have particular behaviours
  • Find time periods of decreasing burglary rates

• Compare two or more behaviours (find similarities and differences)
  • Different attributes over the same set of references
    • Compare the variations of the burglary and robbery rates
  • Same attributes over different subsets of references
    • Compare the temporal variations of the burglary rates in the eastern and western parts

• Relate behaviours of two or more attributes
  • Relate the crime rates to socio-demographic characteristics
Elementary tasks ~
tasks addressing individual values

- Major classes of elementary tasks:
  - Determine the characteristics (attribute values) corresponding to particular references
    - What were the crime rates in Clerkenwell in the last year?
  - Find references having particular attribute values
    - Find districts and years where burglary rates were >50
  - Compare two or more characteristics
    - Different attributes for the same reference
      - Compare the burglary and robbery rates in Clerkenwell in the last year
    - Same attributes for different references
      - Compare the burglary rates in Clerkenwell and Bunhill; in 2013 and 2014
Synoptic tasks for data with >1 referrers

• When data have 2 or more referential components, there are tasks that are synoptic with regard to one component and elementary with regard to the other component(s)
  ➢ “Semi-synoptic” tasks

• E.g. crime rates by districts and time
  • Describe the temporal variation of the crime rates in Clerkenwell – synoptic w.r.t. time and elementary w.r.t. space (the set of districts)
  • Describe the spatial distribution of the crime rates in 2012 – synoptic w.r.t. space and elementary w.r.t. time
Data analysis ~ system of tasks

• Data analysis consists of one or more **synoptic tasks**

• Complex synoptic tasks may be decomposed into simpler **synoptic tasks** (subtasks)
  
  • Describe the variation of different types of crime → describe the variation of each type of crime, compare the variations

• Synoptic tasks may need to be decomposed into semi-synoptic tasks
  
  • Describe the overall spatio-temporal variation of the crime rates → describe the temporal variation in each district + describe the spatial variation in each year

• Elementary tasks play subordinate role: help in fulfilling synoptic tasks
Analysis tasks and data visualisation

- **Purpose of data visualisation**: enable a human analyst to accomplish analysis tasks
  - The visualisation must suit the analysis tasks

- Synoptic tasks entail *abstraction* from individual data items to sets and behaviours
  - The visualisation should support *abstraction*
  - It should provide an *overall view* of the behaviour(s) of interest
    - Ideally, the analyst should be able to grasp the entire behaviour by a single sight
Bertin’s “image”

• The meaningful visual form perceptible in the minimum instant of vision is called the **image**.

• The most efficient visualisations are those in which any question, whatever its type or level, can be answered in a single instant of perception, in a **single image**.

A behaviour over time is represented by a single line.

A behaviour over space is represented by colour value variation that can be perceived as a single figure.
Betrin’s “image” (continued)

• An image can be composed using
  • two display dimensions
  • two display dimensions + single retinal variable

⇒ An image will not accommodate more than three data components

• For a given set of data components, a visualisation with fewer images is more efficient
Each diagram on the map requires an individual instance of perception, i.e., $N$ of images $= N$ of diagrams $= N$ of districts.

Each map requires one instance of perception, i.e., $N$ of images $= 2$. more efficient!
Support of synoptic and elementary tasks

• It is not always possible to construct a visualisation that is equally good for synoptic and elementary tasks

General approach:

• Visualisation should first and foremost support synoptic tasks
• Elementary tasks can be supported by interactive operations (to be considered later)
Questions?

Data analysis tasks
(briefly)
Types of visual display
Bertin’s “impositions”*

* Display types distinguished according to the way of utilising the plane dimensions

<table>
<thead>
<tr>
<th>Arrangement</th>
<th>Rectilinear</th>
<th>Circular</th>
<th>Orthogonal</th>
<th>Polar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagrams</strong></td>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
<td><img src="image" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Coordinate system</strong></td>
<td><img src="image" alt="Coordinate System" /></td>
<td><img src="image" alt="Coordinate System" /></td>
<td><img src="image" alt="Coordinate System" /></td>
<td><img src="image" alt="Coordinate System" /></td>
</tr>
<tr>
<td><strong>Networks</strong></td>
<td><img src="image" alt="Networks" /></td>
<td><img src="image" alt="Networks" /></td>
<td><img src="image" alt="Networks" /></td>
<td><img src="image" alt="Networks" /></td>
</tr>
<tr>
<td><strong>Layout</strong></td>
<td><img src="image" alt="Layout" /></td>
<td><img src="image" alt="Layout" /></td>
<td><img src="image" alt="Layout" /></td>
<td><img src="image" alt="Layout" /></td>
</tr>
<tr>
<td><strong>Maps</strong></td>
<td><img src="image" alt="Map" /></td>
<td><img src="image" alt="Map" /></td>
<td><img src="image" alt="Map" /></td>
<td><img src="image" alt="Map" /></td>
</tr>
<tr>
<td><strong>Isomorphism to physical space</strong></td>
<td><img src="image" alt="Isomorphism" /></td>
<td><img src="image" alt="Isomorphism" /></td>
<td><img src="image" alt="Isomorphism" /></td>
<td><img src="image" alt="Isomorphism" /></td>
</tr>
<tr>
<td><strong>Symbols</strong></td>
<td><img src="image" alt="Symbol" /></td>
<td><img src="image" alt="Symbol" /></td>
<td><img src="image" alt="Symbol" /></td>
<td><img src="image" alt="Symbol" /></td>
</tr>
<tr>
<td><strong>Arbitrary</strong></td>
<td><img src="image" alt="Arbitrary" /></td>
<td><img src="image" alt="Arbitrary" /></td>
<td><img src="image" alt="Arbitrary" /></td>
<td><img src="image" alt="Arbitrary" /></td>
</tr>
</tbody>
</table>

(a.k.a. Graphs)

(a.k.a. Charts)

(a.k.a. Icons)
# Common display types

<table>
<thead>
<tr>
<th>Display elements</th>
<th>Data components</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bar graph</strong></td>
<td><strong>Display</strong></td>
</tr>
<tr>
<td>(bar chart)</td>
<td>x- or y-position</td>
</tr>
<tr>
<td></td>
<td>bar size (length)</td>
</tr>
<tr>
<td><strong>Scatter plot</strong></td>
<td>marks (dots)</td>
</tr>
<tr>
<td></td>
<td>x-position</td>
</tr>
<tr>
<td></td>
<td>y-position</td>
</tr>
<tr>
<td><strong>Line graph</strong></td>
<td>x-position</td>
</tr>
<tr>
<td></td>
<td>y-position</td>
</tr>
<tr>
<td><strong>Gantt chart</strong></td>
<td>y-position</td>
</tr>
<tr>
<td></td>
<td>x-position</td>
</tr>
<tr>
<td></td>
<td>bar size (length)</td>
</tr>
<tr>
<td><strong>Map</strong></td>
<td>marks</td>
</tr>
<tr>
<td></td>
<td>(x,y)-position</td>
</tr>
<tr>
<td></td>
<td>other variables</td>
</tr>
</tbody>
</table>
One display may be insufficient

Techniques to visualise complex data

Display space division (small multiples)

Space embedding (e.g., diagram spaces within map space)

Complementary displays
Additional visual variable: display time (animation)

Has a limited value for analysis:
• Keeping previous states in the memory is hard
• It is therefore hard to compare states and detect changes
Here there is a bar diagram in each table column corresponding to a numeric attribute.

Table rows (y-positions) correspond to references.

The length of the darker grey bars represent numeric values.

The number of simultaneously shown references is quite limited.

Synoptic tasks are not supported.
“Condensed” bar diagram

All references and corresponding attribute values are represented; however, individual references and attribute values can only be accessed through interaction.

Synoptic tasks are still not well supported.
We can grasp the properties of value distribution within the sorted column by a single instance of vision.
Scatter plot

Represents the mutual behaviours of two numeric attributes ⇒ allows the analyst to relate these two attributes.

The marks form one or more figures (clouds) that can be perceived in a single instance of vision.

The relationship is inferred from the figure shape(s).
Scatterplot matrix

An application of the “small multiples” technique.

Allows comparison of mutual behaviours of different attribute pairs.
One curve (polygonal line) can represent the behaviour of a numeric attribute over an ordered set (sequence) of references, e.g., temporal. In a case of two referrers (e.g., temporal and non-temporal), multiple curves corresponding to different non-temporal references can be drawn in the same display space. This, however, creates display clutter.

Interactive selection of singular curves supports semi-synoptic tasks.
The horizontal dimension represents time. The horizontal positions and lengths of the bars show the existence times of objects, events, activities, or processes (here: life times of roe deer). Additionally, the bars can be painted to represent values of a time-variant attribute (here: number of neighbours).
When all bars can be seen simultaneously, the display enables an overall view of the temporal distribution of the referents’ existence times and can thereby support synoptic tasks. However, this is possible for a relatively small number of distinct references. The bars need to be sorted according to the existence times.
Values of a numeric attribute are represented by the visual variable ‘colour value’. Darker shades correspond to higher attribute values. The shades are used for painting areas or objects on the map. The map is perceived as a single image and thus supports synoptic tasks w.r.t. space.
Choropleth map: diverging colour scale

A diverging scale of colour values uses two distinct colour hues for representing positive and negative values or positive and negative differences from a chosen central value, such as the overall mean. Darker shades correspond to larger differences. The map is perceived as a single image and exposes spatial clusters of high and low values.
Map with proportional symbols (bars)
Diverging bars use orientation (upward and downward) and, complementarily, two distinct colour hues for showing positive and negative attribute values or differences from a selected central value, such as the overall mean.
Map with proportional symbols (circles)

Problem: too much visual clutter; requires zooming for discerning the symbols.
Map with proportional circles

The symbols are now discernible, but the overall view is lost.
Similarly to bars, circles of two distinct colour hues can show positive and negative values or differences (the orientation cannot be used in this case).
Maps with proportional symbols vs. choropleth maps

• Studies of human perception show that people can better estimate numeric values represented by symbol sizes than represented by colour shades.

⇒ It may seem that maps with proportional symbols are more effective than choropleth maps.

• However, a choropleth map has serious advantages:
  • provides a single image and thereby supports synoptic tasks;
  • is free from overlapping and clutter;
  • small differences in shades may be easier detectable than small differences in sizes.

• Accurate estimation of numeric values is only required for elementary tasks, which can be supported by interactive techniques.

• Both choropleth maps and proportional symbol maps become more expressive and effective when diverging scales are applied.
**Area cartograms**

Geographic regions are transformed so that their sizes become proportional to their population or some other demographic attribute, such as income or disease incidence.


Area cartograms: pros and cons

The principle of isomorphism of the display space to the physical space is violated.

+ Can be very impressive
− but only when you know well the true sizes and shapes

− Even well-known regions may be hardly recognisable due to the distortions

+ Can provide a background for visualising another attribute, as on a choropleth map, and enable the analyst to relate the two attributes.
− It is hard to relate any of them to the true geographic space.

⇒ It may be a good means for expressive communication of some message but not a tool for analysis.
Maps with diagrams (bar diagrams)
Diverging bars show the differences of the attribute values from the respective means. The bars showing negative differences are painted in less saturated colours.
- Too much visual clutter.
+ Still, areas with the prevalence of particular colours are detectable.
- The method is applicable to attributes that make together a meaningful sum. Here: population N by social status.
Map zooming reduces the clutter
The diagrams represent the variation of attribute values over time. This method can be used for data with 2 referrers: space and time. However, such a map does not enable an overall view and therefore does not support synoptic tasks. It also suffers from visual clutter, which complicates the analysis.
Map with “value flow” diagrams
Map with “value flow” diagrams

Only at this zoom level, the diagrams can be easily discerned.
The technique of diverging scale can also be applied to value flow diagrams. This can work well when the diagrams do not overlap.
Maps with diagrams: general notes

• In principle, can be used for representing the spatial variation of multiple numeric attributes or spatio-temporal variation of a single numeric attribute.

• However, diagram maps are prone to visual clutter, which complicates the perception and analysis.

⇒ In practice, diagram maps can be used when data refer to a small number of places (requiring a small number of diagrams).

• Even in absence of visual clutter, a diagram map does not provide a single image (cannot be perceived all at once) and therefore does not support synoptic tasks.
Parallel coordinates plot

not so common but popular in visualisation research community

Multiple parallel axes are used for multiple attributes. Attribute values are represented by positions on the respective axes. For each reference, there is a polygonal line connecting the positions on neighbouring axes representing the corresponding attribute values.
Parallel coordinates plot

not so common but popular in visualisation research community

A parallel coordinate plot in its basic form supports neither synoptic nor elementary tasks. Various enhancements have been suggested by visualisation researchers (consideration would require a separate lecture).
Common display types: general notes

• All display types have their limitations.

• Display types that can support synoptic tasks:
  ± Bar chart: for a relatively small number of references (all bars need to be visible simultaneously); sorting is essential
  ✓ Scatter plot
  ✓ Line graph (single curve or a few curves)
  ± Gantt chart: same as bar chart
  ✓ Choropleth map
  ± Maps with proportional symbols: for a relatively small number of spatial references (such that visual clutter can be avoided); diverging scales are helpful.

• Complex data may require multiple complementary displays, interactive techniques, and computational processing.
Questions?

Types of visual display
Displays of aggregated data
Aggregated vs. detailed data displays

• In detailed data displays (considered so far), marks correspond to individual references. This imposes limitations and creates problems for analysis:
  • Display size may be insufficient to include marks for all references ⇒ no overall view, no support to synoptic tasks.
  • Display may be cluttered and hardly readable.
  • Even when marks are not too numerous, some of them may overlap (e.g., in a scatter plot).

• In aggregated data displays, marks correspond to groups (subsets) of references.
  ✓ Groups may be arbitrarily large ⇒ displays are scalable to large amounts of data.
  ✓ Synoptic tasks can be well supported.
  - No support to elementary tasks.
Frequency histogram

Shows the distribution of numeric attribute values over the set of references. Each bar represents a group of references such that the corresponding attribute values lie within a certain interval, which is represented by the horizontal position of the bar. The height of the bar is proportional to the group size (number of references).
Frequency histogram

- The most frequent values
- Frequent values
- Outliers
Value distributions of several attributes can be compared using several frequency histograms. Here: the histograms represent the same attributes we tried to look at using bar diagrams (slide 50). The comparison of the distributions is easier with the histograms.
Like in a scatter plot, two dimensions represent two numeric attributes. The value ranges are divided into intervals, which creates a rectangular grid over the plot area. The sizes of the symbols within the grid cells are proportional to the numbers of references having the attribute values within the respective intervals.
What can a 2D histogram show us?

- Relationships (such as correlation) between the attributes
- Frequencies of value combinations

- For different value intervals of one attribute, how the values of the other attribute are distributed.
- Multiple 1D histograms that are easy to compare.
Each bar stands for one time step (generally, one value of referrer 1). The value range of the attribute is divided into intervals. The height of each segment is proportional to the number of values of referrer 2 for which the values in this time step lie within the respective interval. The intervals are represented by the segment colours.

This technique is applicable to data with two referrers: one linearly ordered (such as time) and one arbitrary. Note that a line graph with multiple curves may be unreadable due to line overlapping.
Quantile graph

¼ (25%) of the values

¼ (25%) of the values

¼ (25%) of the values

¼ (25%) of the values
Segmented bars and quantile graphs show how the value distribution w.r.t. one referrer (set of objects, places, etc.) varies over the range of linearly ordered values of another referrer (such as time).
The idea of representing value quantiles can also be applied on a parallel coordinate plot.
Aggregated data displays: general notes

• Aggregated data displays solve particular problems of detailed data displays: insufficient display size for available data, visual clutter, and overlapping of marks.

• Aggregated data displays can better support synoptic tasks but do not support elementary tasks.

• Interactive operations are necessary for both classes of data display.
  • To be considered next.
Questions?

Displays of aggregated data
Reading:

http://0-dx.doi.org.wam.city.ac.uk/10.1007/3-540-31190-4

Natalia and Gennady Andrienko

Exploratory Analysis of Spatial and Temporal Data
A Systematic Approach


Section 4.3
Visualisation in a Nutshell