Week 5 Lecture 1

Slide 1

Hi, I'm Dr Nigel Houlden, this is week 5, lecture 1 of data visualisation.

Slide 2

So far we have encountered visualization few times and we have seen its for example some definitions and application about it. An widely formal accepted definition is coming from 1987 National Science foundation F Panel by McCormick, DeFanti, Brown as follow:

"Visualization is a method of computing. It transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations. Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights. In many fields it is already revolutionizing the way scientists do science.

Visualization embraces both image understanding and image synthesis. That is, visualization is a tool both for interpreting image data fed into a computer, and for generating images from complex multi-dimensional data sets. It studies those mechanisms in humans and computers which allow them in concert to perceive, use and communicate visual information.

An estimated 50 percent of the brain's neurons are associated with vision. Visualization in scientific computing aims to put that neurological machinery to work."

For example this image is a featured plot, i.e. it is a volume plot of the logarithm of gas/dust density in an Enzo star and galaxy simulation. Regions of high density are white while less dense regions are more blue and also more transparent. The data used to make this image were provided by Tom Abel Ph.D. and Matthew Turk of the Kavli Institute for Particle Astrophysics and Cosmology and visulaised by a tool called Vis it

Slide 3

Visualization – Background:

Visualization is very old, back 1000 years ago it was often an intuitive step to make something clearer such as a map

Traditional BI primarily utilizes descriptive and diagnostic analytics to provide information on historical and current events. It is not "intelligent" because it only provides answers to correctly formulated questions. Correctly formulating questions requires an understanding of business problems and issues and of the data itself. BI reports on different KPIs through:

• ad-hoc reports

• dashboards

Ad-hoc reporting is a process that involves manually processing data to produce custom made reports, as shown in the Figure, in which OLAP and OLTP data sources can be used by BI tools for both ad-hoc reporting and dashboards.

The focus of an ad-hoc report is usually on a specific area of the business, such as its marketing or supply chain management. The generated custom reports are detailed and often tabular in nature.

Slide 4

Data set sizes are ever-increasing making a graphical approach necessary. Classical (easy) approaches known from business graphics (Excel, etc.) to dashboards.

Dashboards provided a holistic view of key business areas. The information displayed on dashboards is generated at periodic intervals in realtime or near-realtime. The presentation of data on dashboards is graphical in nature, using bar charts, pie charts and gauges, as shown **Figure where** tools use both OLAP and OLTP to display the information on dashboards.

Slide 5

As previously explained, data warehouses and data marts contain consolidated and validated information about enterprise-wide business entities. Traditional BI cannot function effectively without data marts because they contain the optimized and segregated data that BI requires for reporting purposes.

Without data marts, data needs to be extracted from the data warehouse via an ETL(Extract, transform, load) process on an ad-hoc basis whenever a query needs to be run. This increases the time and effort to execute queries and generate reports.

Traditional BI uses data warehouses and data marts for visualisation and data analysis because they allow complex data analysis queries with multiple joins and aggregations to be issued.
While traditional BI analyses generally focus on individual business processes, Big Data BI analyses focus on multiple business processes simultaneously. This helps reveal patterns and anomalies across a broader scope within the enterprise. It also leads to data discovery by identifying insights and information that may have been previously absent or unknown.
Big Data BI requires the analysis of unstructured, semi-structured and structured data residing in the enterprise data warehouse.
This acts as a uniform and central repository of structured, semi-structured and unstructured data that can provide Big Data BI tools with all of the required data. This eliminates the need for Big Data BI tools to have to connect to multiple data sources to retrieve or access data. Figure shows transformation to next-generation data warehouse establishes a standardized data access layer across a range of data sources for visulaisation.

Slide 6

Data visualization is a technique whereby analytical results are graphically communicated using elements like charts, maps, data grids, infographics and alerts. Graphically representing data can make it easier to understand reports, view trends and identify patterns.
Traditional data visualization provides mostly static charts and graphs in reports and dashboards, whereas contemporary data visualization tools are interactive and can provide

both summarized and detailed views of data. They are designed to help people who lack statistical and/or mathematical skills to better understand analytical results without having to resort to spreadsheets.
Traditional data visualization tools query data from relational databases, OLAP systems, data warehouses and spreadsheets to present both descriptive and diagnostic analytics results.

Slide 7

Big Data solutions require data visualization tools that can seamlessly connect to structured, semi-structured and unstructured data sources and are further capable of handling millions of data records. Data visualization tools for Big Data solutions generally use in-memory analytical technologies that reduce the latency normally attributed to traditional, disk-based data visualization tools.

Advanced data visualization tools for Big Data solutions incorporate predictive and prescriptive data analytics and data transformation features. These tools eliminate the need for data pre-processing methods, such as ETL. The tools also provide the ability to directly connect to structured, semi-structured and unstructured data sources. As part of Big Data solutions, advanced data visualization tools can join structured and unstructured data that is kept in memory for fast data access. Queries and statistical formulas can then be applied as part of various data analysis tasks for viewing data in a user-friendly format, such as on a dashboard.

Common features of visualization tools used in Big Data:

- *Aggregation* – provides a holistic and summarized view of data across multiple contexts
- *Drill-down* – enables a detailed view of the data of interest by focusing in on a data subset from the summarized view
- *Filtering* – helps focus on a particular set of data by filtering away the data that is not of immediate interest
- *Roll-up* – groups data across multiple categories to show subtotals and totals
- *What-if analysis* – enables multiple outcomes to be visualized by enabling related factors to be dynamically changed.

Slide 8

Visualization become its own scientific discipline since ~1987 and First visualization conference held in 1990

Visual analysis is a form of data analysis that involves the graphic representation of data to enable or enhance its visual perception. Based on the premise that humans can understand and draw conclusions from graphics more quickly than from text, visual analysis acts as a discovery tool in the field of Big Data.

The objective is to use graphic representations to develop a deeper understanding of the data being analyzed. Specifically, it helps identify and highlight hidden patterns, correlations and anomalies. Visual analysis is also directly related to exploratory data analysis as it encourages the formulation of questions from different angles. the following types of visual analysis are described next :

• Heat Maps

• Time Series Plots

• Network Graphs

• Spatial Data Mapping

Slide 9

Heat Maps

Heat maps are an effective visual analysis technique for expressing patterns, data compositions via part-whole relations and geographic distributions of data. They also facilitate the identification of areas of interest and the discovery of extreme (high/low) values within a dataset.

For example, in order to identify the top- and worst-selling regions for ice cream sales, the ice cream sales data is plotted using a heat map. Green is used to highlight the best performing regions, while red is used to highlight worst performing regions. The heat map itself is a visual, color-coded representation of data values. Each value is given a colour according to its type or the range that it falls under.

For example, a heat map may assign the values of 0–3 to the colour red, 4–6 to amber and 7–10 to green.

A heat map can be in the form of a chart or a map. A chart represents a matrix of values in which each cell is color-coded according to the value, as shown in left Figure .

This chart heat map depicts the sales of three divisions within a company over a period of six months.

It can also represent hierarchical values by using color-coded nested rectangles.

In right Figure a map represents a geographic measure by which different regions are color-coded or shaded according to a certain theme. Instead of colouring or shading the whole region, the map may be superimposed by a layer made up of collections of coloured/shaded points relating to various regions, or coloured/shaded shapes representing various regions. In right figure shows A heat map of the Sea Surface Temperature.

Sample questions can include:

- *How can I visually identify any patterns related to carbon emissions across a large number of cities around the world?*
- *How can I see if there are any patterns of different types of cancers in relation to different ethnicities?*
- *How can I analyze soccer players according to their strengths and weaknesses?*

Slide 10

Time series plots allow the analysis of data that is recorded over periodic intervals of time. This type of analysis makes use of time series, which is a time-ordered collection of values recorded over regular time intervals. An example is a time series that contains sales figures that are recorded at the end of each month.

Time series analysis helps to uncover patterns within data that are time- dependent. Once identified, the pattern can be extrapolated for future predictions. For example, to identify seasonal sales patterns, monthly ice cream sales figures are plotted as a time series, which further helps to forecast sales figures for the next season.

Time series analyses are usually used for forecasting by identifying long-term trends, seasonal periodic patterns and irregular short-term variations in the dataset. Unlike other types of analyses, time series analysis always includes time as a comparison variable, and the data collected is always time dependent.

A time series plot is generally expressed using a line chart, with time plotted on the x-axis and the recorded data value plotted on the y-axis, as shown in Figure (A line chart depicts a sales time series from 1990 to 1996).

The time series presented in Figure spans seven years. The evenly spaced peaks toward the end of each year show seasonal periodic patterns, for example Christmas sales. The dotted red circles represent short-term irregular variations. The blue line shows an upward trend, indicating an increase in sales.

Sample questions can include:

- *How much yield should the farmer expect based on historical yield data?*
- *What is the expected increase in population in the next 5 years?*
- *Is the current decrease in sales a one-off occurrence or does it occur regularly?*

Slide 11

Within the context of visual analysis, a network graph depicts an interconnected collection of entities. An entity can be a person, a group, or some other business domain object such as a product. Entities may be connected with one another directly or indirectly. Some connections may only be one-way, so that traversal in the reverse direction is not possible.

Network analysis is a technique that focuses on analyzing relationships between entities within the network. It involves plotting entities as nodes and connections as edges between nodes. There are specialized variations of network analysis, including:

• route optimization
• social network analysis
• spread prediction, such as the spread of a contagious disease

The following is a simple example based on ice cream sales for the application of network analysis for route optimization.

Some ice cream store managers are complaining about the time it takes for delivery trucks to drive between the central warehouse and stores in remote areas. On hotter days, ice cream delivered from the central warehouse to the remote stores melts and cannot be sold. Network analysis is used to find the shortest routes between the central warehouse and the remote stores in order to minimize the durations of deliveries.

Figure shows an example of a social network graph.

Consider the social network graph in Figure for a simple example of social network analysis:

• John has many friends, whereas Alice only has one friend.
• The results of a social network analysis reveal that Alice will most likely befriend John and Katie, since they have a common friend named Oliver.

Sample questions may include:

• *How can I identify influencers within a large group of users?*
• *Are two individuals related to each other via a long chain of ancestry?*
• *How can I identify interaction patterns among a very large number of protein-to protein interactions?*

Slide 12

Spatial or geospatial data is commonly used to identify the geographic location of individual entities that can then be mapped. Spatial data analysis is focused on analysing locationbased data in order to find different geographic relationships and patterns between entities.

Spatial data is manipulated through a Geographic Information System (GIS) that plots spatial data on a map generally using its longitude and latitude coordinates.

The GIS provides tooling that enables interactive exploration of the spatial data, for example measuring the distance between two points, or defining a region around a point as a circle with a defined distance-based radius.

With the ever-increasing availability of location-based data, such as sensor and social media data, spatial data can be analysed to gain location insights.

For example, as part of a corporate expansion, more ice cream stores are planned to open. There is a requirement that no two stores can be within a distance of 5 kilometres of each other to prevent the stores from competing with each other. Spatial data is used to plot existing store locations and to then identify optimal locations for new stores at least 5 kilometres away from existing stores.

Applications of spatial data analysis include operations and logistic optimization, environmental sciences and infrastructure planning. Data used as input for spatial data analysis can either contain exact locations, such as longitude and latitude, or the information required to calculate locations, such as zip codes or IP addresses.
Furthermore, spatial data analysis can be used to determine the number of entities that fall within a certain radius of another entity. For example, a supermarket is using spatial analysis for targeted marketing, as shown in Figure Locations are extracted from the users' social media messages, and personalized offers are delivered in real-time based on the proximity of the user to the store.
Therefore, it shows Spatial data analysis can be used for targeted marketing.

Sample questions can include:

- *How many houses will be affected due to a road widening project?*
- *How far do customers have to commute in order to get to a supermarket?*
- *Where are the high and low concentrations of a particular mineral based on readings taken from a number of sample locations within an area?*

Slide 13

Edward Tufte was a pioneer of data visualisation, meaning how data is presented as graphical content for important use.
These are Tufte's six principles of Graphical Integrity.

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
2. Clear, detailed, and thorough labelling should be used to defeat graphical distortion and ambiguity.
3. Show data variation not design variation.

Slide 14

4. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data
5. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.
6. Graphics must not quote data out of context

Slide 15

It is easier than you think to portray wrong information in graphics by mistake,

Graphical Ethics is very important and about Graphical integrity,

Don't lie with images!

Don’t distort with images!

Show data in proper context!

As a measure Graphical integrity Lie factor is defined as the size of the effect shown over the size of the effect in data

Slide 16

Be careful of Lying Visually by mistake or Intentionally It can happen if Horizontal spacing are inconsistent! Or shows False perspective!

Just be careful when using.