Introduction to Data Analytics: Course Intro

Professor Ian Nabney

ian.nabney@bristol.ac.uk

Overview

- Structure of the unit
  - Information Visualisation (Prof. Ian Nabney): how to make sense of digital data by distilling and displaying it using an appropriate visualization technique. This unit introduces the science of information visualisation, covering topics such as data and task abstraction, visual thinking and how to arrange information for visualisation. The lectures will be accompanied by lab exercises using the interactive visualisation platform Tableau.
  - Text Analytics (Dr. Edwin Simpson): how to process natural language to extract information and identify patterns and trends, e.g., through sentiment analysis. This unit introduces natural language processing using rule-based and statistical machine learning methods for tasks such as text classification, sequence tagging, clustering, dependency parsing and relation extraction. Lectures will be accompanied by lab exercises using Python and Jupyter notebook.
- Why visualise data?
- How we can visualise data

Week	Asynchronous Content	Lectorial/Discussion	Lab Class
1	Information visualisation: overview of InfoVis and Data Abstraction	Data abstraction	Introduction to Tableau
2	Information visualisation: task abstraction and marks	Visual queries	Basic Visualisation in Tableau
3	Information visualisation: arranging tabular and non-tabular data	Visual search	Graphs in Tableau
4	Text analytics: introducing natural language processing, regular expressions, preprocessing	Rule-based text processing and preprocessing	Regular expressions and tokenisation with Python
5	Text analytics: classification and sentiment analysis with naïve Bayes and logistic regression	Text classification and sentiment analysis	Implementing sentiment classifiers with Python
-	READING WEEK		
6	Text analytics: document clustering, topic modelling, and vector representations of meaning	Text clustering	Latent Dirichlet allocation with Sklearn or Gensim
7	Text analytics: sequence labelling, HMMs, Conditional Random Fields (CRFs)	Sequence labelling	Sequence taggers in Python with HMMLearn and CRFStuie
8	Text analytics: named entity recognition, dependency parsing	Information extraction and syntax parsing	Sequence taggers (continued)
9	Text analytics: relation extraction	Information extraction continued	Coursework
-	SPRING VACATION		
10	Information visualisation: reduction and mapping marks	Visual thinking	Coursework
11	Information visualisation: info vis summary and conclusions	Reflection and evaluation	None (Bank Holiday)

What is Visualisation?

- Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively. Goal of visualisation is to present data in a human-readable way.
- Visualisation is an important tool for developing a better understanding of large complex datasets. It is particularly helpful for users who are not specialists in data modelling.
  - Detection of outliers.
  - Clustering and segmentation.
  - Aid to feature selection.
  - Feedback on results of analysis: seeing what you are doing.
- Two aspects: information visualisation and data projection (latter is covered in Advanced Data Analytics.

Information Visualisation Principles

- The design space for information visualisation is huge
- There is no definite ‘unique right way’ to visualise data (it is an unsupervised task).
- Older texts/methods tended to fall back on simple criteria, and even aesthetics (see Tufte, in himself a considerable step forward).
- While these matter, they do not provide a systematic way to design information visualisations.
- To address this we need to
  - abstract the task to get to core requirements
  - understand the principles of visual perception for analytics
  - validate and assess the effectiveness of the methods we use

Rules of Thumb

- No unjustified 3D
- No unjustified 2D
- Eyes beat memory
- Resolution over Immersion
- Overview first; zoom and filter; details on demand
- Responsiveness is Required
- Get it Right in Black and White
- Function First, Form Next

Visualisation Methods: Word Clouds

GIVEN
COUNTRY
YET
MIND
FIRST
MOMENT
TODAY
LIFE
EVERY
PERHAPS
SMALL
SURVIVE
HAPPEN
WORDS
FELT
TIME
MUST
WASH
NOW
MINE
ROAD
FACE
NEITHER
RE
UNHAPPINESS
LIVE
WATCH
EVEN
KEY
BELONGS
POWER
HAPPENED
WAR
FUTURE
ALMOST
GOT
STRENGTH
MAN
YEARS
EQUALLY
FACT
LEAST
OLD
HUNGER
RAINING
POWERFUL
WINDY
INFINITELY
GOES
HUMAN
ALSO
DIFFERENT
BETTER
OTHERS
QUESTION
FORGET
ANOTHER
DUTY
EXIST
AWAY
FIND
GERMANS
AMONG
PERFECT
PAVEL
LIKE
SEE
MEN
CLEAN
UNDERSTAND
LAGER
SHAME
COMES
WANT
ASK
NOTHING
RETURN
EXAMPLE
LATER
CURIOUS
STILL
MAN'S
AWARE
FEEL
WEAK
BEYOND
AUSCHWITZ
STATE
JUST
SOMETHING
FLUE
REALITY
KNOW
WITHOUT
IGNORANCE
END
ASKED
REALLY
CERTAIN
ONE
LIES
EVERYTHING
GUILT
ENEMIES
ONE'S
FAR
NEVER
ANIMALS
WAY
MATTER
CHEMISTRY
IMPURITIES
BEHIND
BRAIN
LISTEN
BELIEVE
DIRTY
LANGUAGE
FUNDAMENTAL
HAPPENS
WORLD
NOW
THINGS
MAKE
MADE
SECOND
MEMORY
ACT
STRONG
PAIN
GOOD
TWO
AROUND
THINK
KUHN
YOUNG
DEATH
HAPPINESS
AIR
EXPERIENCE
TURN
THROW
OUTSIDE
BOTTOM
CAUSE
MORAL
MANY
NAME
DUST
SOON

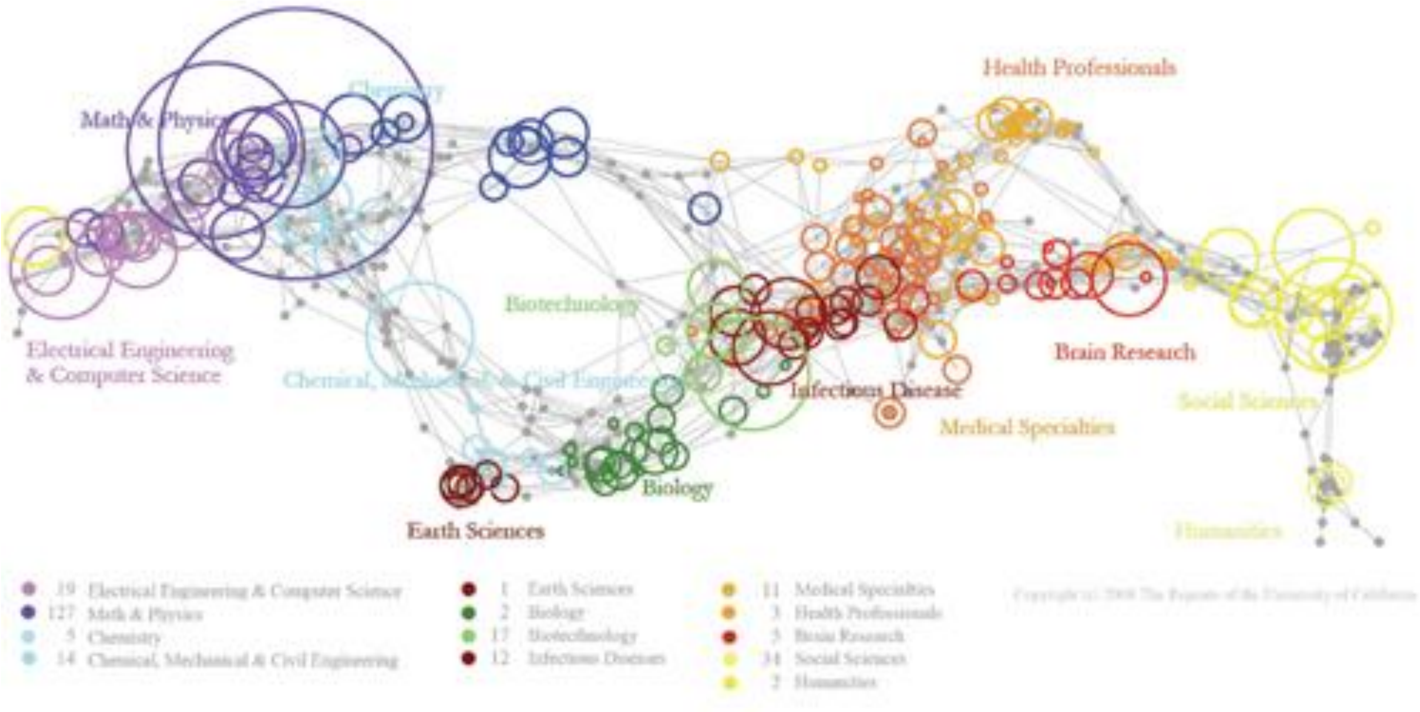
Word Cloud (www.wordle.net)

- Single words, the importance of each is shown with font size or colour.
- When used as website navigation aids, the terms are hyperlinked to items associated with the tag.

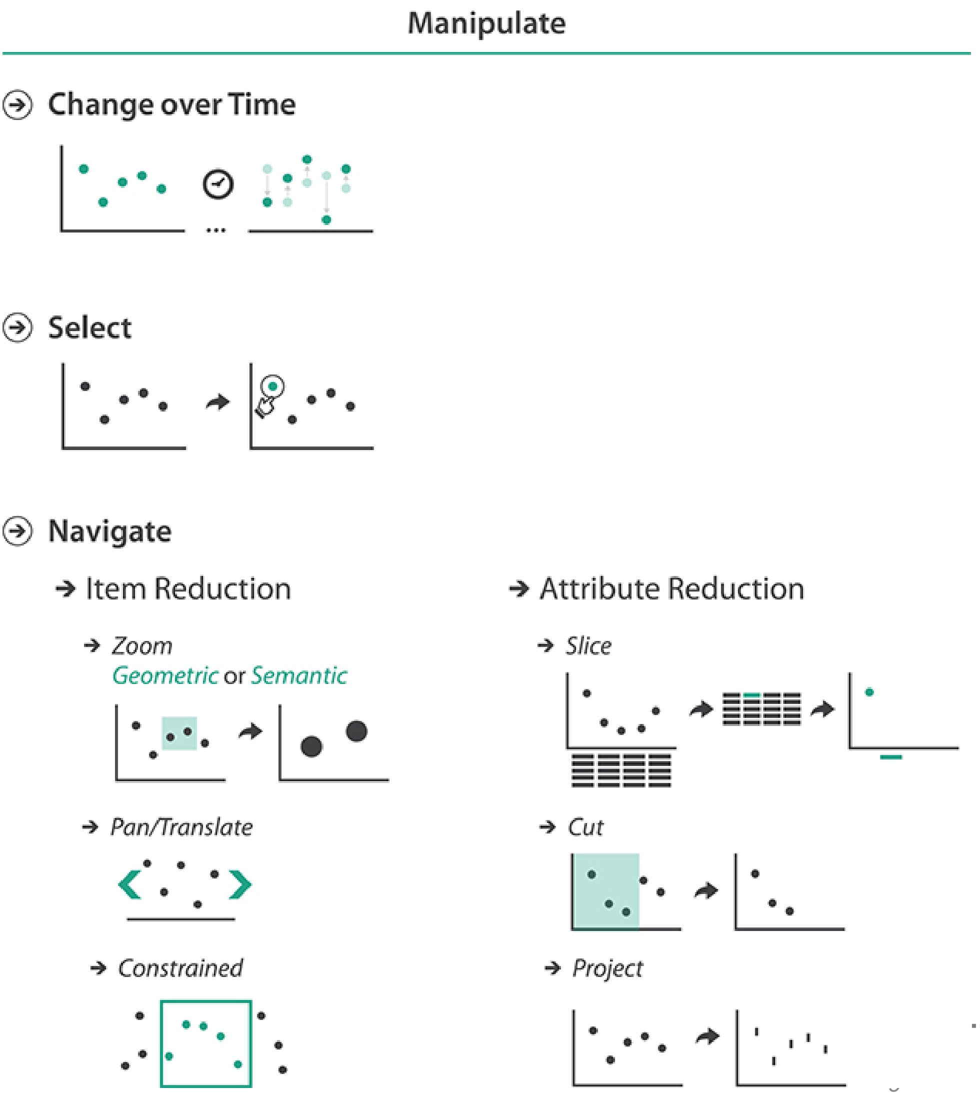
Visualisation Methods: Network Analysis

27/01/2022

UCSD Map of Science (Zoss and Borner)



Changing the View



- Ability to interact with a graphic is a key benefit of computer over print
- Adds functionality, but must be done with care to get the full benefit
- Also relates to managing multiple linked views and reducing items and attributes

Final thoughts

- This is a timely course – it has been a pandemic illustrated by graphics of many different types
- Just as the population has learned about infection models (and the dreaded ‘R’), it has also learned the value of log scales on graphs
- A wide range of topics:
    - Heat map of hospital admissions https://twitter.com/jburnmurdoch/status/1322233070137806850/photo/1
    - Change in acceptance of vaccination in France over time https://twitter.com/coulmont/status/1353255871552352256
    - False information: https://time.graphics/line/455000

Conclusions

- We need to understand the vast quantities of data that surround us; visualisation and machine learning can help us in that task.
- Models can be used to uncover the hidden meanings of data.
- Visual analytics is a powerful tool that provides insight to non-specialists.
- It is a multivariate, multi-skilled, collaborative effort.
- It can be beautiful – and fun! https://richardbrath.wordpress.com/2021/10/31/58-ways-to-visualize-alice-in-wonderland/