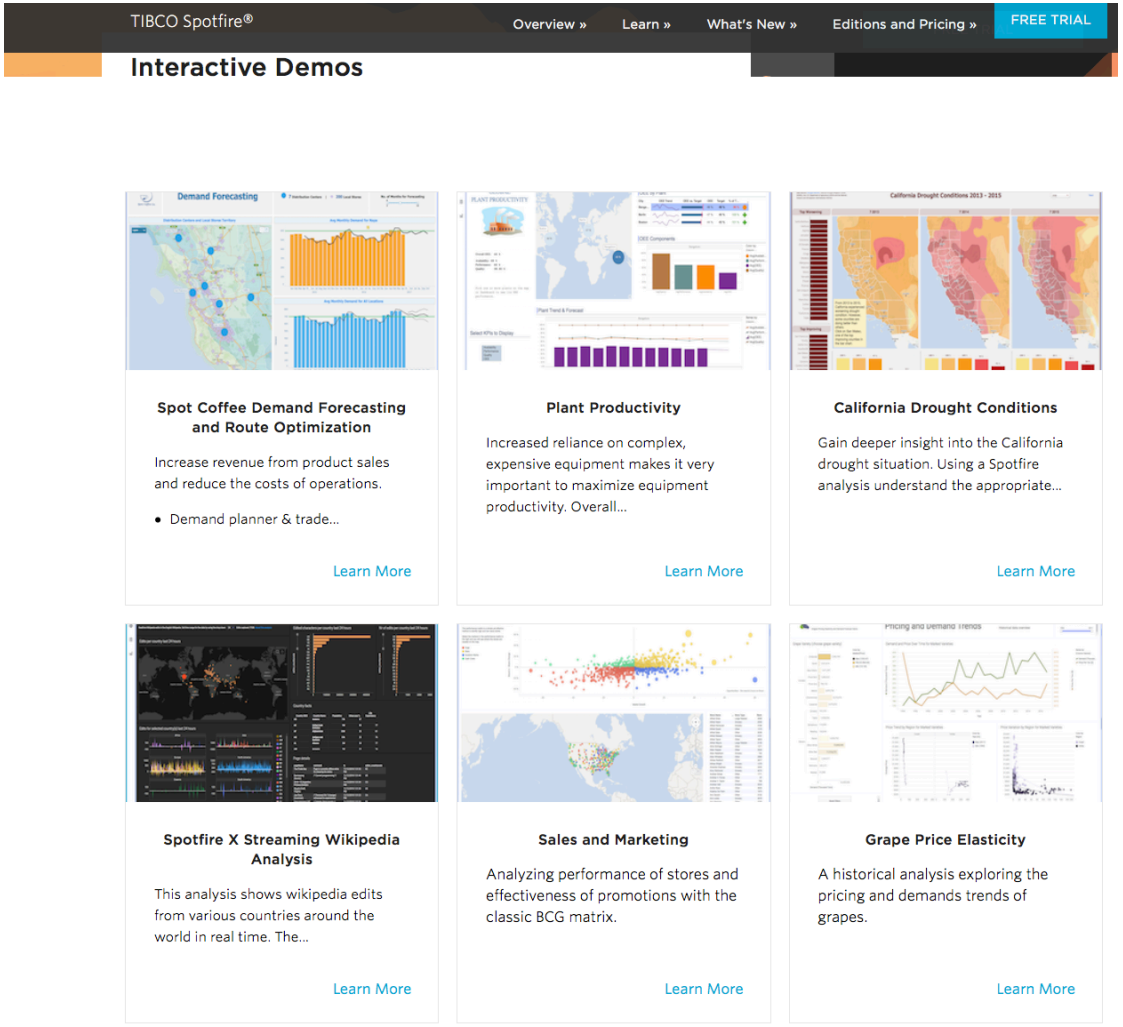
FIT5147 Data Exploration & Visualisation

Kim Marriott

Syllabus

Week	Lecture material	Lab/Tute
. 1	Visual analytics; Tools for data exploration	Intro to Tableau; R; D3
. 2	Visualisation of tabular data	Advanced graphics with R
. 3	Analysis of trends & patterns in tabular data	Interactive graphics with R
. 4	Data maps;Tools for creating data maps	Data maps with R
. 5	Spatial analytics	MapBox; Data Exploration Project feedback
. 6	Network data analysis & visualisation	Relational data and text and text analytics with R
. 7	Textual data analysis & visualisation	Data Exploration Project Feedback
Break		
. 8	Visualisation design methodology	Five design sheet visualisation design methodology
. 9	Human visual system	Introduction to D3
. 10	Visual communication	More D3;Data VisProject Feedback
. 11	Interactive data visualisation	Data Vis Project Presentations
. 12	History and future of data visualisation	Five design sheet visualisation design methodology

Data Exploration & Visualisation



https://www.tibco.com/products/tibco-spotfire/learn/demos

Group Discussion 1

Break into groups of about 6 people and try and answer the following questions

- What are the different uses of data visualisation in data science? Give examples.
- What are the benefits of data visualisation?

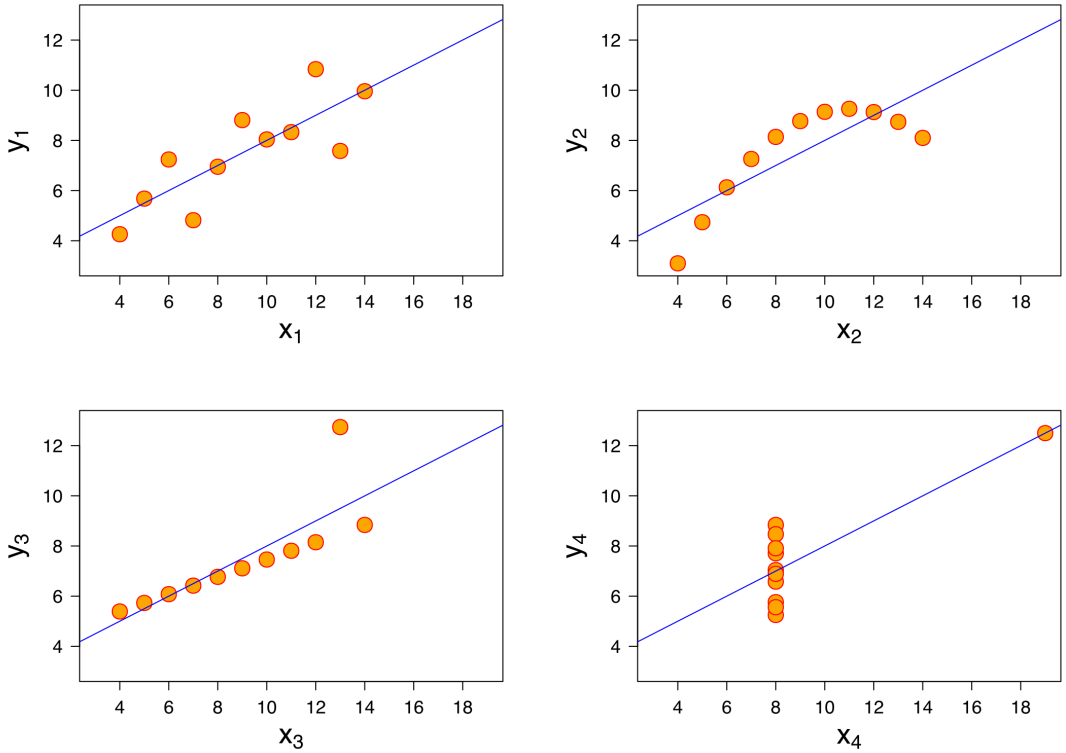
Role of Data Visualisation

Data visualisation used in Data Science for

- Data checking and cleaning
- Exploration and discovery
- Presentation and communication of results

Interactive computer visualisations are one of the best ways we have of dealing with big data.

Why Visualisation?

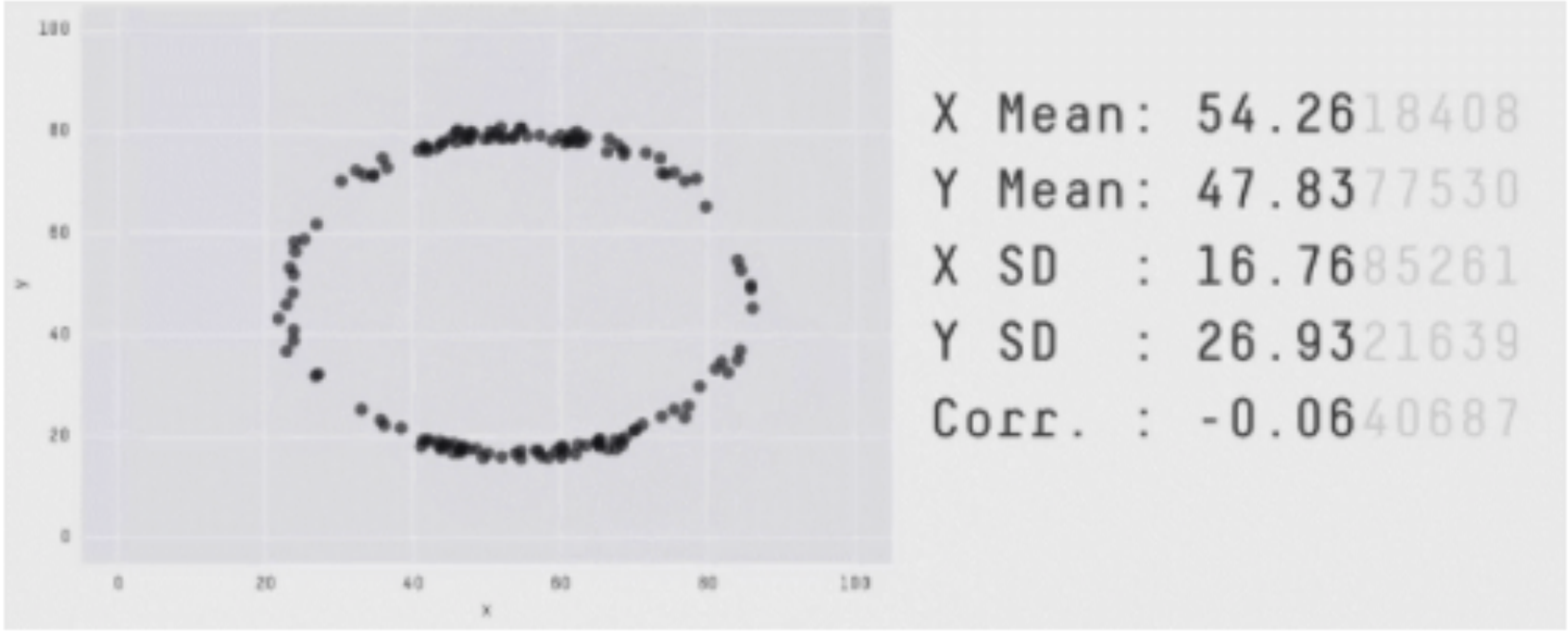


Anscombe's Data Quartet

Visual representations contain much more information than a few summary statistics

Human visual system is extremely effective, parallel processing

Check Out Datasuarus

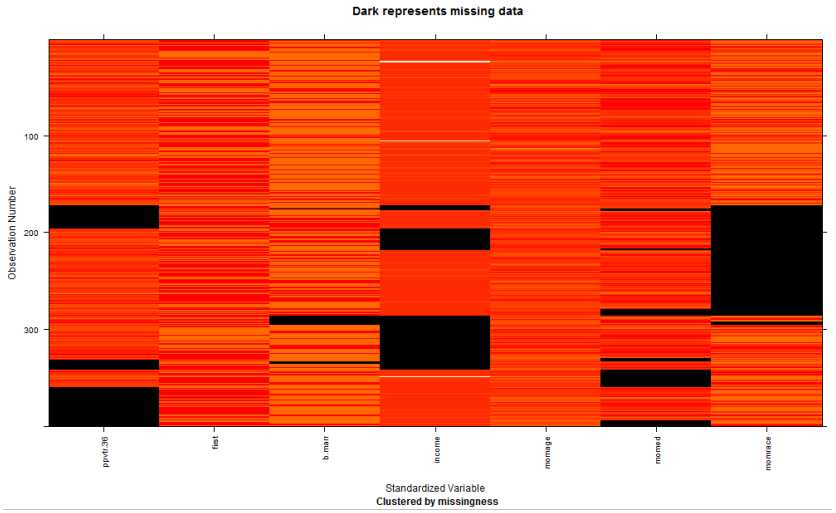


https://blog.revolutionanalytics.com/2017/05/the-datasaurus-dozen.html

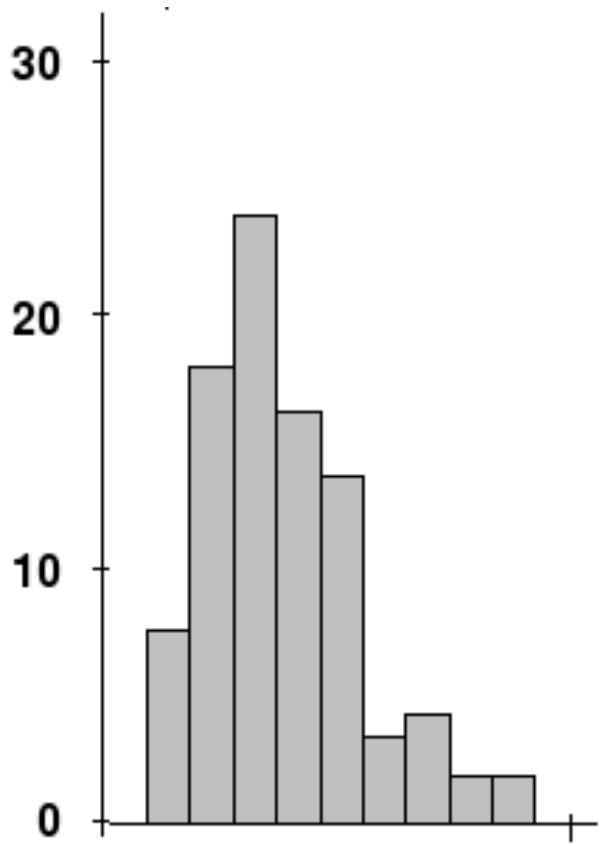
Data Checking & Cleaning

Check each attribute for obvious errors

- Look at some random records.
- Determine the number of missing values and invalid values (NaNs), number of special values like 0.
- Determine the number of distinct values and whether they really are distinct.
- Check formats for dates, that they are in comparable time zones.
- Plot latitude and longitude on a map to check they are sensible.
- Check text for strange characters or encoding



Get a Feel for the Data Distribution



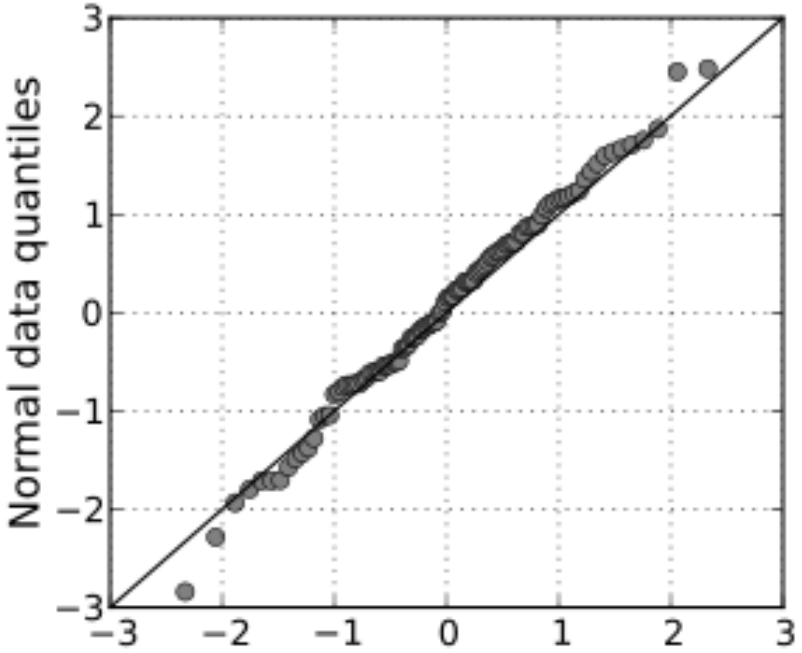
Compute the mean, median and quartiles for the data. Look at a box plot of these.

Plot the frequency distribution of values. This might be with a histogram or density plot. You should play around with the choice of bin width as this smooths the data.

Check for symmetry (skewness) and the flatness/spikiness of the distribution (kurtosis).

Look at the outliers and check whether they should be thrown away (trimmed) rounded up or down (Winsorised).

Check Assumptions



Look at the frequency distribution to see if it meets assumptions for statistical tests (e.g. normality)

Can use a **Q-Q** plot to check if two distributions are similar

- Plots the quantiles of one distribution against those of the other distribution.
- If the points lie along the y=x line then the two distributions are the same.
- If they lie along a straight line then they are the same up to scaling and translation.

Exploration and Discovery

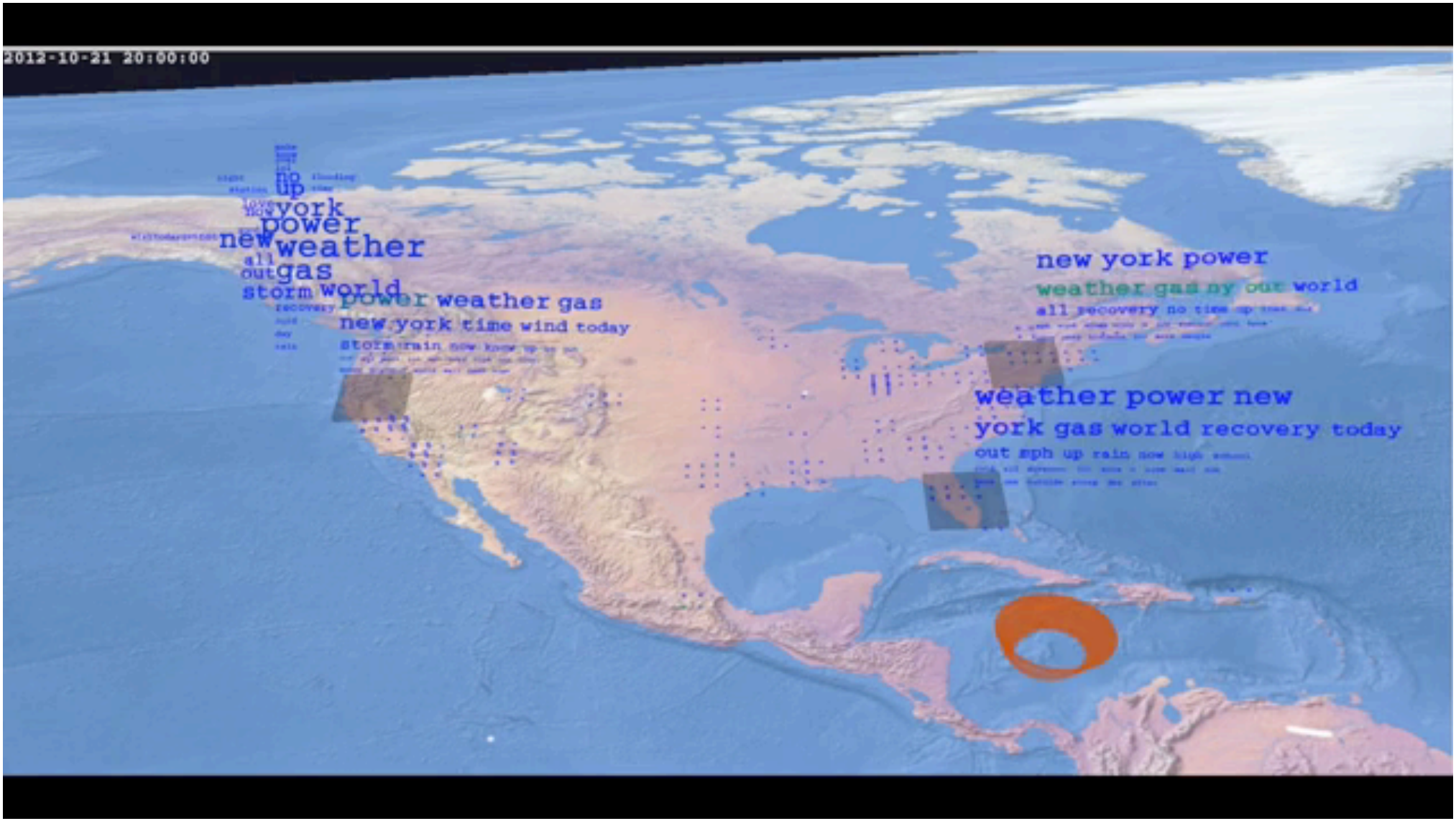
The greatest value of a picture is when it forces us to notice what we never expected to see. — J. W. Tukey. Exploratory Data Analysis, 1977

The heart of data science is exploring the data and discovering patterns and trends.

- Visualisation plays a core role in this process. It reveals possible connections and patterns that can then be confirmed (or not) using other kinds of analysis
- Visualisation is key in understanding data with a spatial component
- Exploration is inherently incremental.
- May require fusing, transforming or obtaining new data

Hilary Mason, one of the world's leading data scientists, says that when she gets a new data set, she starts by making a dozen or more scatter plots, trying to get a sense of what might be interesting.

Looking for Patterns



Group Discussion II

Break into groups of about 6 people and try and answer the following questions

- When should you use statistical/computational analytics and when should you use visualisation?
- Should you combine them? Why?

Human-in-the-Loop Analytics

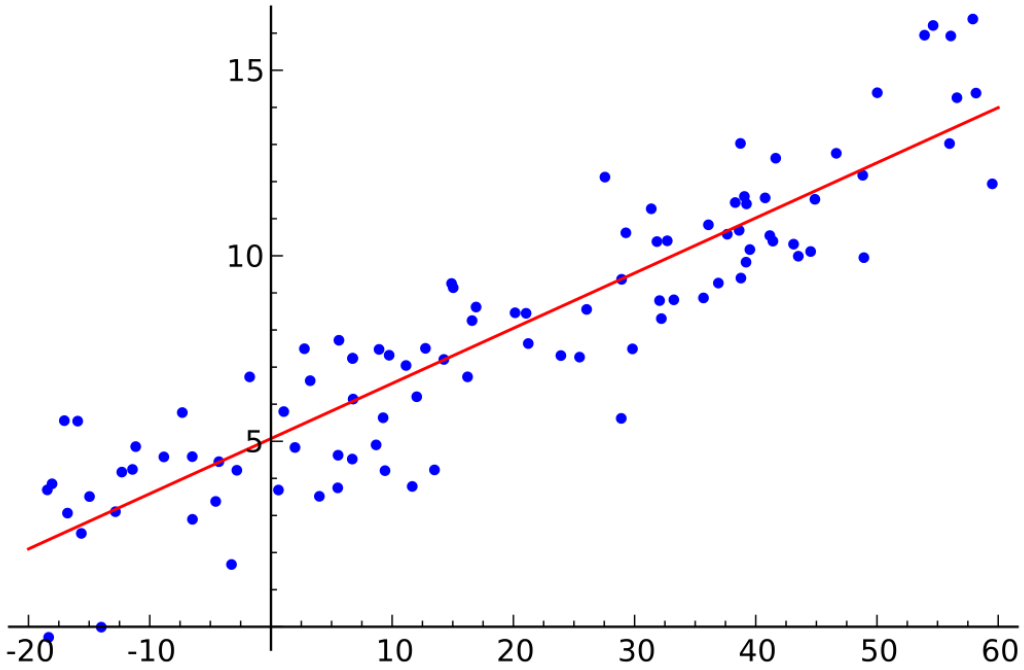
If you know exactly what you are looking for then you can write a computer program and simply automate the analysis

Most times you don't.

You need the Human-in-the-Loop Analytics

- Look at the data,
- Make some tentative hypothesis,
- Run appropriate analytics and visualise the results.
- Repeat this until you have found what you need.

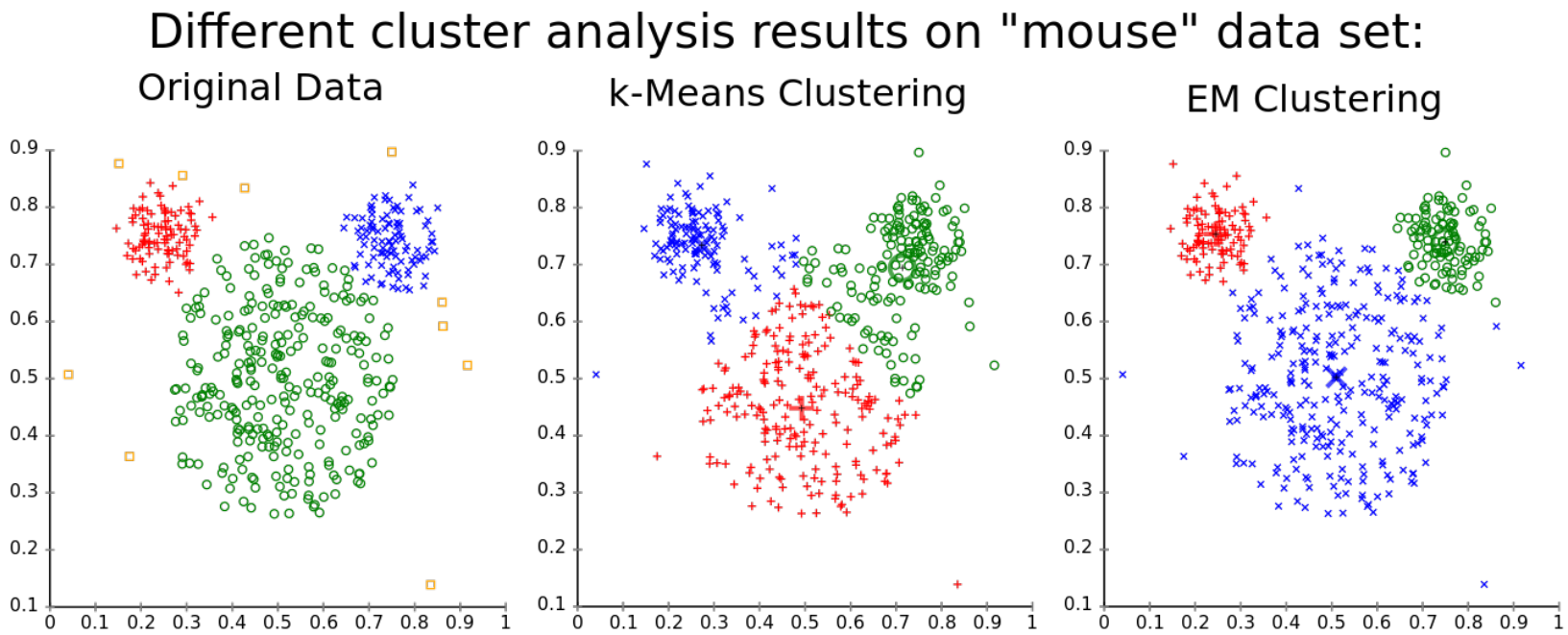
Model Fitting



Model fitting is an important part of data exploration.

Visualisation allows you to understand the results, limitations and check that the model makes sense

Model Fitting



Model fitting is an important part of data exploration.

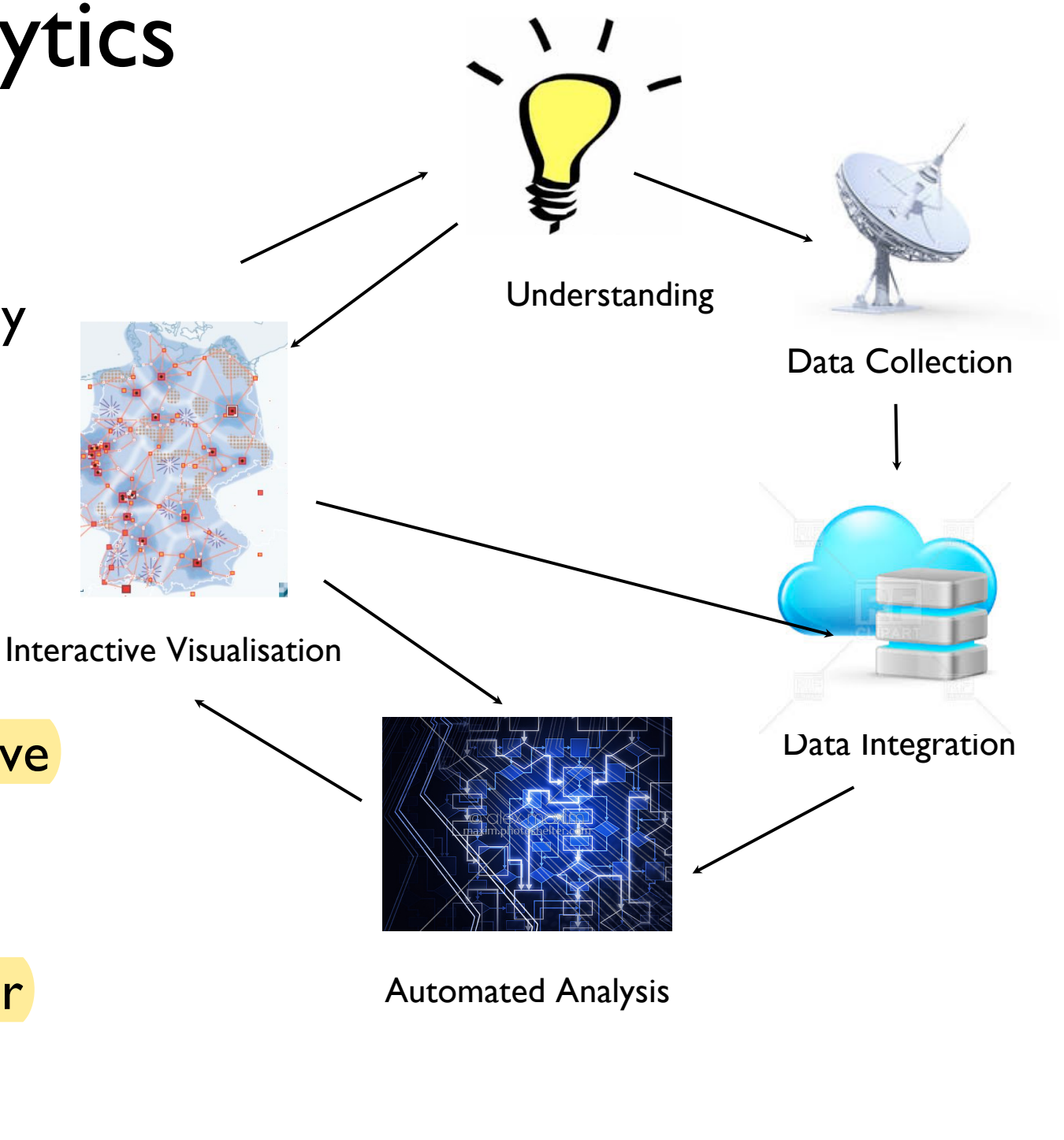
Visualisation allows you to understand the results, limitations and check that the model makes sense

Visual Analytics

Visual analytics is a recent focus in data visualisation started by James Thomas and Kristin Cook early 2000s.

It is "the science of analytical reasoning facilitated by interactive visual interfaces."

Aim is to "detect the expected and discover the unexpected."

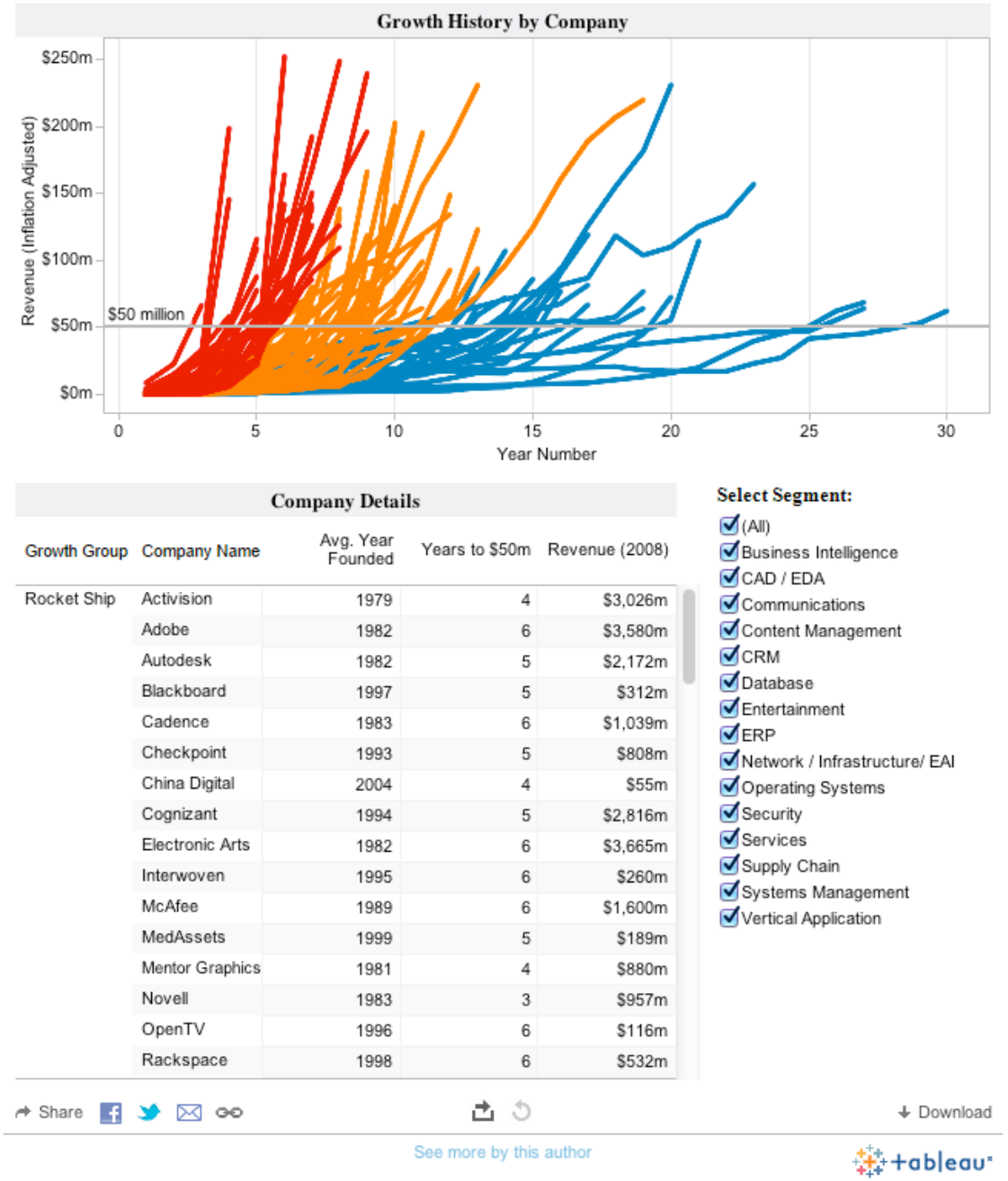


Visual Analytics

Original focus of visual analytics was US security.

Now widely used in business intelligence and sciences for data analysis.

Tools such as Spotfire, QLIK Sense and Tableau allow business analysts to visualise business data and easily share visualisations.



Communication of Results

Data scientists often need to communicate their results to stakeholders who are not data scientists: managers, policy makers, students or the general public.

Such visualisations require considerable time to prepare

- Designed to communicate a particular message or narrative.
- Production values are very high: often touched up using graphics editing tools like Adobe Illustrator or Inkscape.

Once presentation graphics were static, printed on glossy paper or shown in PowerPoint presentations.

Now often interactive and published on the Web.

Communication of Results

There is increasing sophistication in the data visualisations produced by magazines and newspapers.

For instance The New York Times and Washington Post employs data visualisation experts in both its graphics department and in a group dedicated to online interactive graphics.

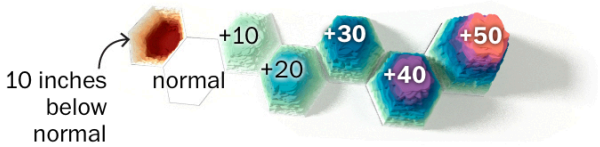
Take a look at some Nathan Yau's books or his Flowing Data web site https://flowingdata.com

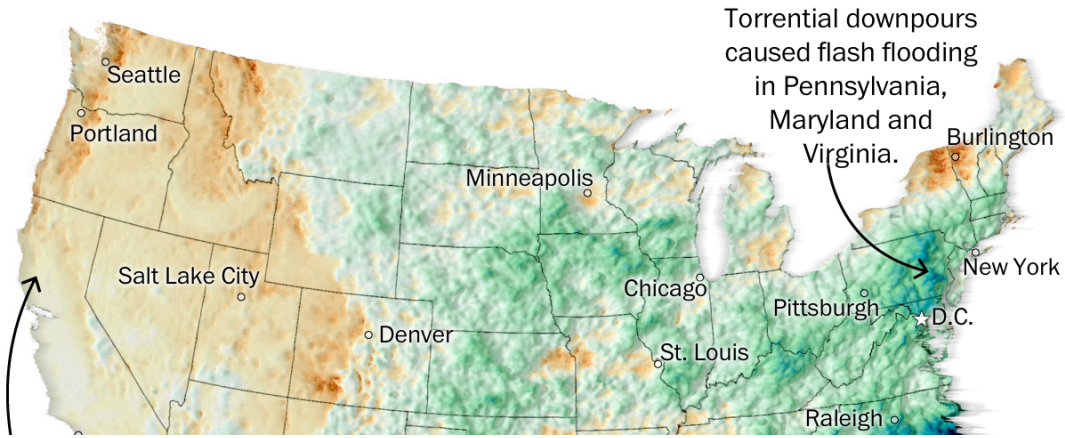
Summer rainfall: Opposite extremes split the nation

It's been a wet one, unless, of course, it's been a dry one. For much of the country, this summer's precipitation has been a little unusual.

Precipitation departure from normal

Terrain height and color show precipitation anomaly, June 25 to Sept. 25





https://www.washingtonpost.com/graphics/2018/national/summer-rain/?utm_term=.215da7c5224a

Visual Analytics Tools

A wide variety of data exploration and visualisation tools are used in data science

Programming languages

- R
- Python
- JavaScript with the D3 visualisation library

Generic Visual Analytics Tools

- Tableau
- Spotfire

Application Specific Visual Analytics Tools

- National Map
- Scaffold Hunter

Summary

Visualisation is used for three important purposes in data science:

- initial data checking and cleaning,
- exploration and discovery, and
- presentation and communication of results.

Visualisation is *effective* because the human visual system allows parallel perception of large amounts of information.

Visual analytics is the name given to the combination of interactive visualisation with statistics, data mining and other kinds of analytics.

Tutorial Activities

5.1 Activity: Exploring & Visualising Data with Tableau Public

5.2 Activity: Exploring & Visualising Data using R

5.3 Activity: Exploring & Visualising Data using D3

Assessment

Programming Exercise1: Tableau [Due end of Week 2]

Choice of Project for Data Exploration Project [Due end of Week 2]

Think about a Hall of Fame/Shame Entry

Initial Project Proposal

This should consist of a short text email containing

- Project title
- 1-3 questions you wish to answer
- Data source(s) you plan to use to answer these questions
- Brief description of the data in each data source (kind of data: tabular, spatial, network, textual or other, number of records, URL)

1. Project title: Causes of serious bicycle accidents

2. 1-3 questions you wish to answer

Q1) What are the most common kinds of serious bicycle accidents

Q2) How do lighting conditions affect these

3. Data source(s) you plan to use to answer these questions

A) ACT Road Cyclist Crashes, since 2012, which have been reported by the Police or the Public through the AFP Crash Report Form.

B) Canberra's sunrise and sunset times for 2018

A will allow me to answer Q1 at least for the ACT, while the combination of A and B will allow me to answer Q2

4. Brief description of the data in each data source (kind of data: tabular, spatial, network, textual or other, number of records, URL)

A) Tabular data: 1K rows x 11 columns It has both spatial and temporal attributes as well as some simple text (https://www.data.act.gov.au/Justice-Safety-and-Emergency/Cyclist-Crashes/n2kg-qkwj)

B) Tabular data in HTML: ~400 rows and 11 columns (http://members.iinet.net.au/~jacob/risesetcan.html)

Programming Assignment 1 (5%)

Task: Explore coral bleaching data for 8 sites in the Great Barrier Reef using Tableau. Determine

Q1) In which years and for which kinds of coral bleaching is the worst

Q2) How the location of the site affects bleaching on the different kinds of coral.

Your job is to

- Read the data into Tableau Public: careful you may need to reformat it.
- Check the data for entry errors
- Explore the data to answer the two questions
- Write a 4-5 page report. This should detailing what you did and what you found (see instructions on Moodle).