Analytics –Data Discovery
QlikView

3rd-5th September 2014

KS Gopinath Narayan, IAAS
CIA, CFE, PMP
Pr. Director (IT Audit)
Office of the CAG of India
narayanKSG@cag.gov.in

Presentation Outline

- About Data Analytics
- Data Discovery through Data Visualisation
- Various types of Graphs and Charts
- Tools for Data Discovery
  - Microsoft Excel – Pivot and PowerPivot
  - QlikView / QlikSense
  - Tableau Desktop/Public
- QlikView

About Data Analytics

- *Analytics – Reduction of data to understandable findings.*

- *Data analytics is an analytical process by which **insights are extracted** from operational, financial, and other forms of electronic data.*
- ***….provide the "how?" and "why?" answers** to initial "what?" questions frequently found in the information initially extracted from the data.*

(KPMG-2013)

About Data Analytics

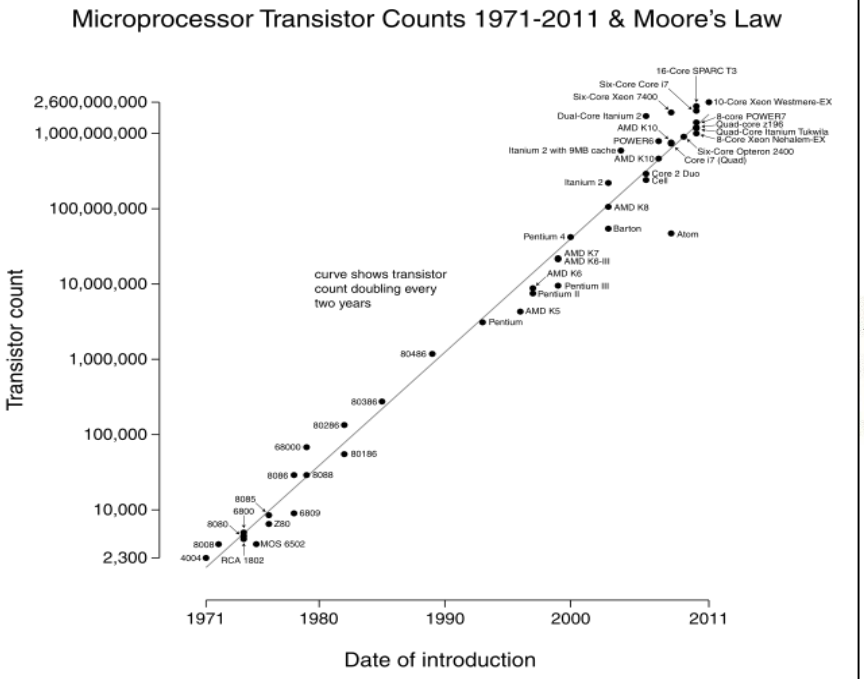
- Data Analytics driven by the Hype around Big Data Analytics
- **Big data:** "collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications"
- Massive growth in data volumes
- *Google CEO Eric Schmidt in 2010: "There were 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."*
- Growth Largely in unstructured data.

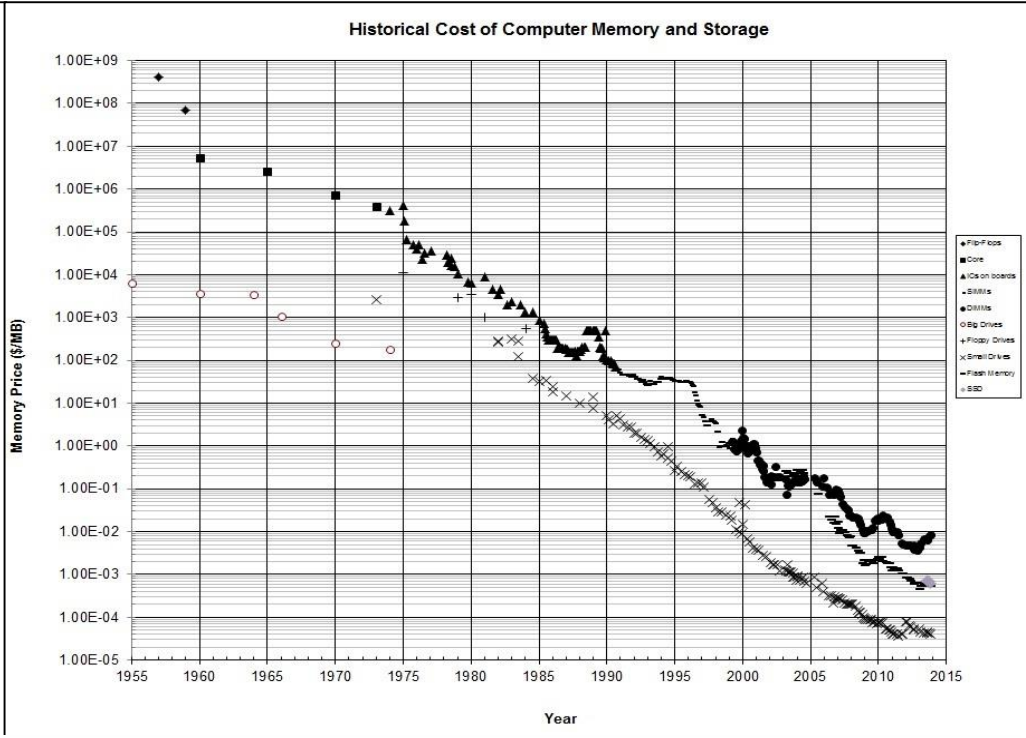
Data Size Primer

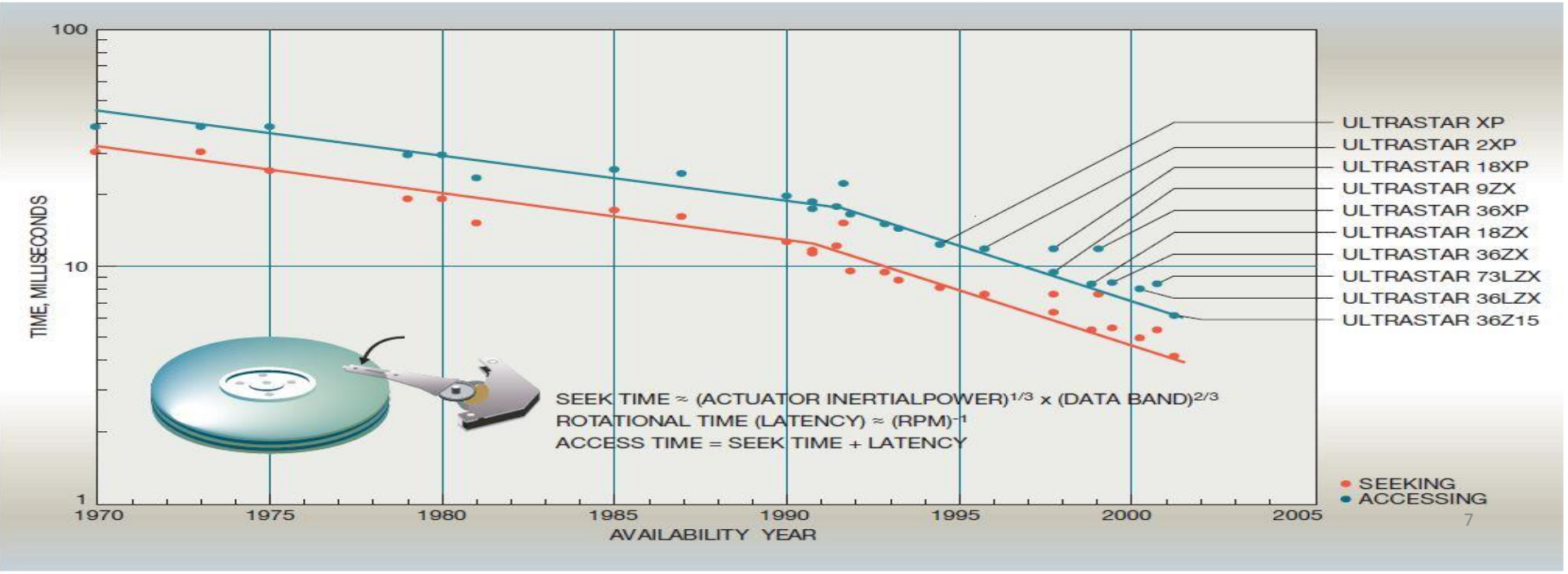
- KB = kilobyte, a paragraph of text
- MB= Megabyte, 10^6 Bytes, Complete works of Shakespeare = 5MB
- GB= Gigabyte, 10^9 Bytes, Accounts Transaction Data of one year 1-3 GB (States in India)
- TB=Terabyte, 10^{12} Bytes, Tweets created daily = 12+ TB, US Library of Congress=235 TB
- PB=Petabyte, 10^{15} Bytes, Data processed in a day by Google=24 PB
- EB=Exabyte, 10^{18} Bytes, Total data created in 2006 = 161 EB

Technological Changes Impacting Analytics

- Traditionally, Analytics capability through Business Intelligence (BI) Tools / Data Warehouse
  - Top down, IT modeled with reports, KPIs, slice and dice capability
- BI moving towards self-service delivery model. New tools/segments being created
- New opportunity in the form of In-Memory Analytics
- Advanced analytics capability now affordable. Powered by:
  - Increased computational power
  - Lower cost of RAM
  - 64 Bit computing

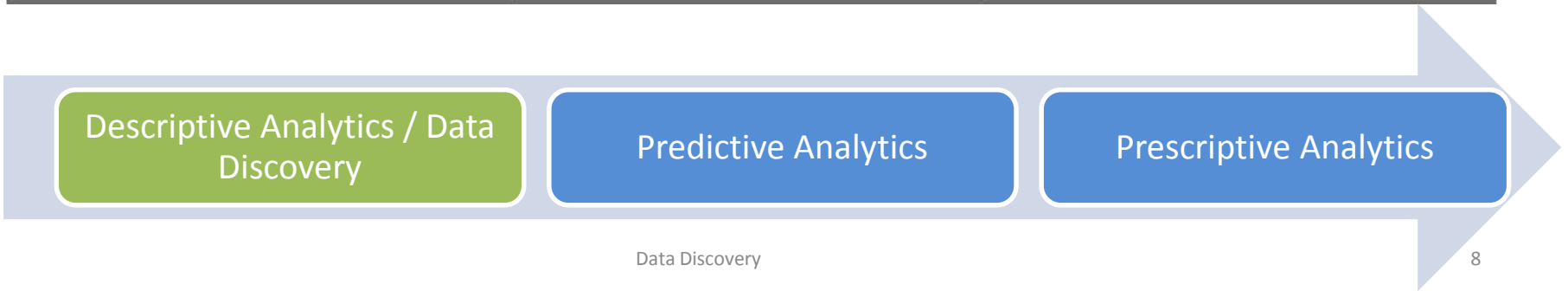

Microprocessor Transistor Counts 1971-2011 & Moore's Law
Transistor count
2,600,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
2,300
1971
1980
1990
2000
2011
Date of introduction
curve shows transistor count doubling every two years
16-Core SPARC T3
Six-Core Core i7
Six-Core Xeon 7400
10-Core Xeon Westmere-EX
Dual-Core Itanium 2
8-core POWER7
Quad-core z196
Quad-Core Itanium Tukwila
AMD K10
POWER6
8-Core Xeon Nehalem-EX
Itanium 2 with 9MB cache
Six-Core Opteron 2400
AMD K10
Core i7 (Quad)
Core 2 Duo
Cell
Itanium 2
AMD K8
Barton
Atom
Pentium 4
AMD K7
AMD K6-III
AMD K6
Pentium III
Pentium II
AMD K5
Pentium
80486
80386
80286
68000
80186
8086
8088
8085
6800
6809
8080
Z80
8008
MOS 6502
4004
RCA 1802


Historical Cost of Computer Memory and Storage
Memory Price ($/MB)
1.00E+09
1.00E+08
1.00E+07
1.00E+06
1.00E+05
1.00E+04
1.00E+03
1.00E+02
1.00E+01
1.00E+00
1.00E-01
1.00E-02
1.00E-03
1.00E-04
1.00E-05
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year
Flip-Flops
Core
ICs on boards
SIMMs
DIMMs
Big Drives
Floppy Drives
Small Drives
Flash Memory
SSD


TIME, MILLISECONDS
100
10
1
1970
1975
1980
1985
1990
1995
2000
2005
AVAILABILITY YEAR
SEEK TIME ≈ (ACTUATOR INERTIALPOWER)1/3 x (DATA BAND)2/3
ROTATIONAL TIME (LATENCY) ≈ (RPM)-1
ACCESS TIME = SEEK TIME + LATENCY
ULTRASTAR XP
ULTRASTAR 2XP
ULTRASTAR 18XP
ULTRASTAR 9ZX
ULTRASTAR 36XP
ULTRASTAR 18ZX
ULTRASTAR 36ZX
ULTRASTAR 73LZX
ULTRASTAR 36LZX
ULTRASTAR 36Z15
SEEKING
ACCESSING

Data Analytics Classification

Descriptive Analytics (Business Intelligence)	Predictive Analytics	Prescriptive Analytics
o What and when did it happen? o How much is impacted and how often does it happen? o What is the problem?	o What is likely to happen next? o What if these trends continue? o What if?	o What is the best answer? o What is the best outcome given uncertainty? o What are significantly differing and better choices?
Statistics	Data Mining Predictive Modeling Machine Learning Forecasting Simulation	Constraint-based optimization Multiobjective optimization Global optimization



CAATs and Audit- Why Data Discovery?

- Traditional CAATs - IDEA , ACL Software
- Some usage of MS Excel and MS Access and SQL
- Generalised Audit Software with transaction based analytics- Rule based or micro-level analytics
  - Data extraction and analysis entailing sorting, grouping, filtering, joining, sampling, irregularity testing-gap detection, Benford analysis
- Good for evaluating known conditions/ Compliance audit
- Lack Macro level Analytics capability- Understand the big picture to identify key areas to check for non-compliance
- Need for new generation Analytics tools to supplement Generalised Audit Software- Data Discovery/Visualisation

Need for Data Discovery – A quote

... there are known knowns; there are things we know that we know.
There are known unknowns; that is to say, there are things that we now know we don't know.
But there are also unknown unknowns – there are things we do not know we don't know.

Donald Rumsfeld: Secretary of Defense

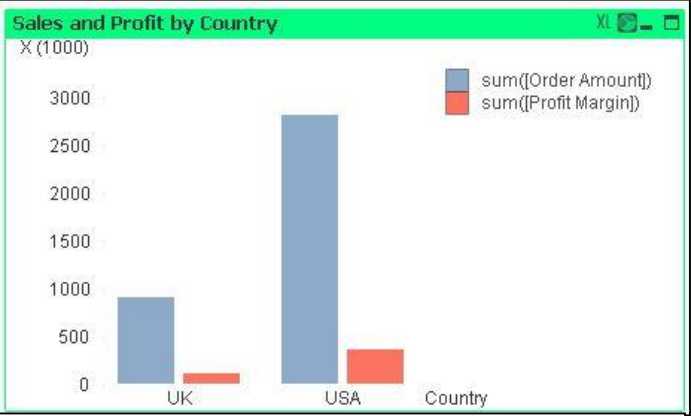
February 2002

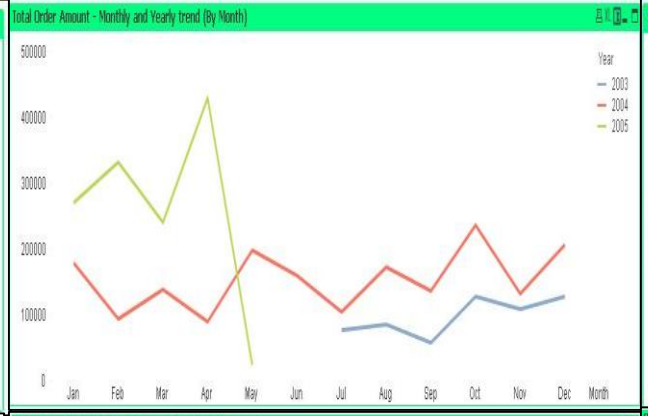
about the lack of evidence linking the government of Iraq with the supply of weapons of mass destruction to terrorist groups

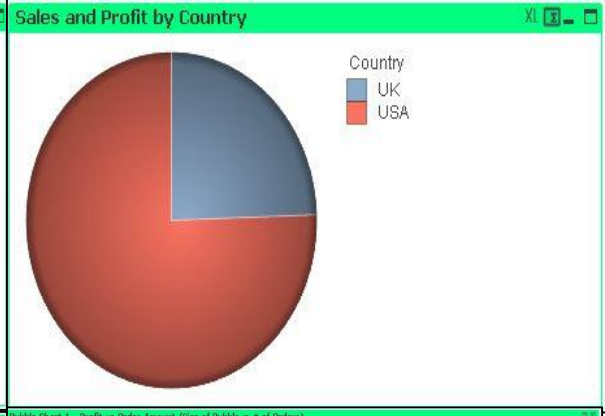
Why Visualisation?

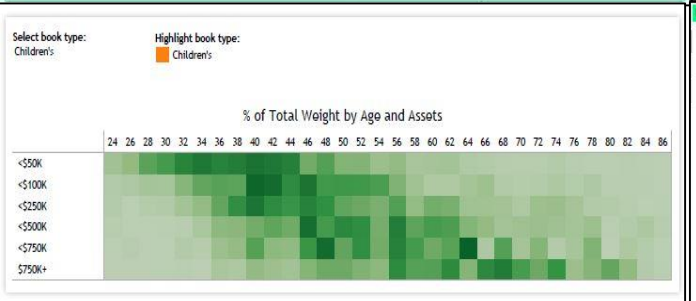
- Visual analysis aids analytical reasoning
- leverages the incredible capabilities and bandwidth of the visual system
- takes advantage of our brains' built-in "software" to identify patterns and communicate relationships and meaning
  - Identify trends and outliers, discover or search for interesting or specific data points in a larger field
- inspire new questions and further exploration

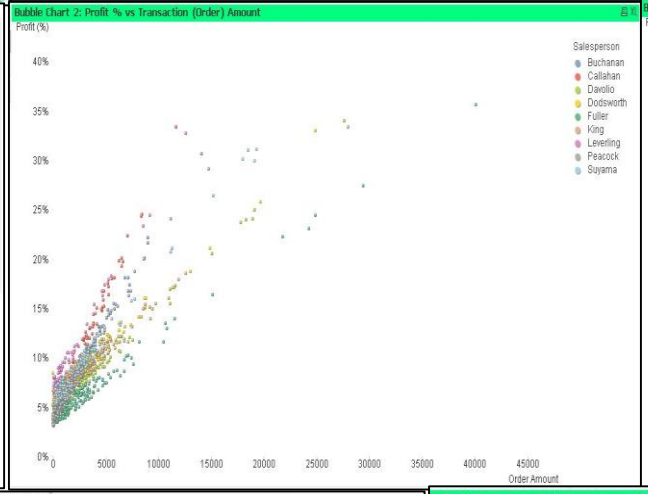
John Tukey - *"…the picture-examining eye is the best finder we have of the wholly unanticipated"*











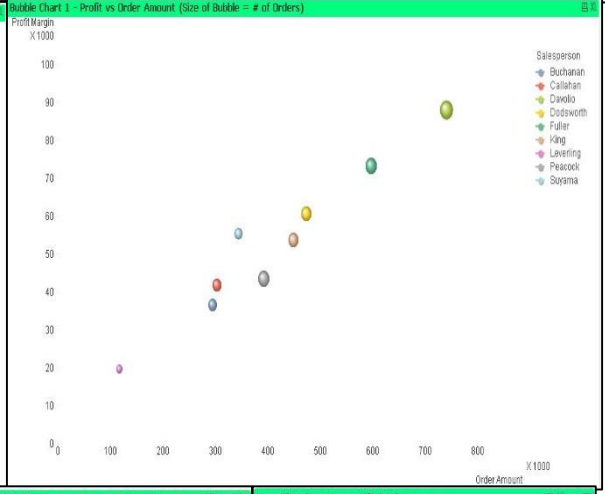
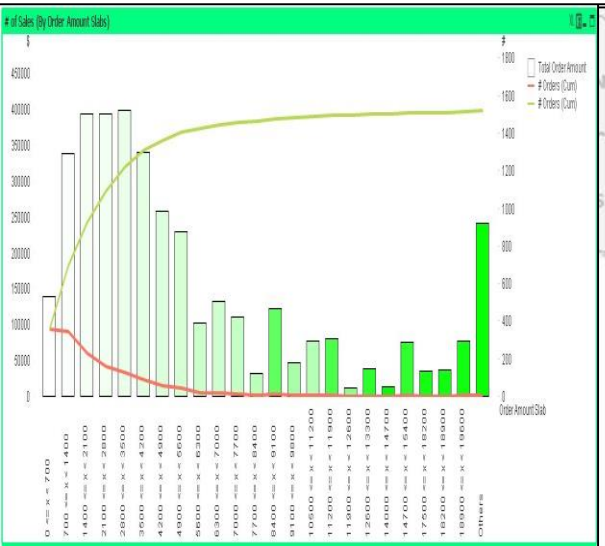
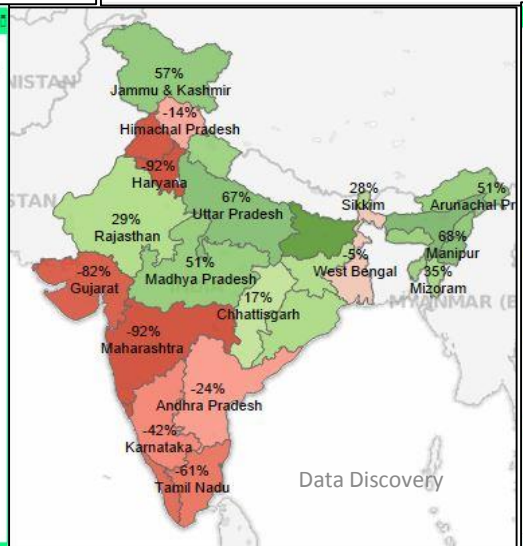
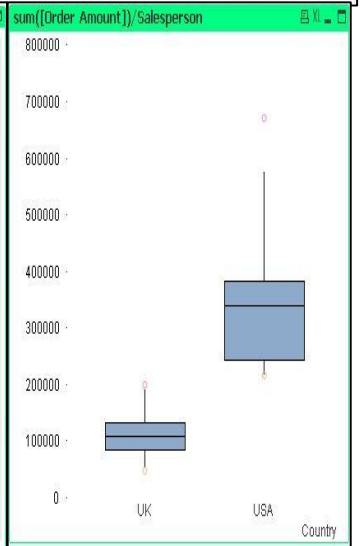


Figure 16: Who buys the most books?

In this market segmentation analysis, the heat map reveals a new campaign idea. High-income households of people in their sixties buy children's books. Perhaps it's time for a new grandparent-oriented campaign?







Anscombe's Quartet

- Four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.
- Each dataset consists of eleven (x,y) points.
- Constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Anscombe's Quartet Data

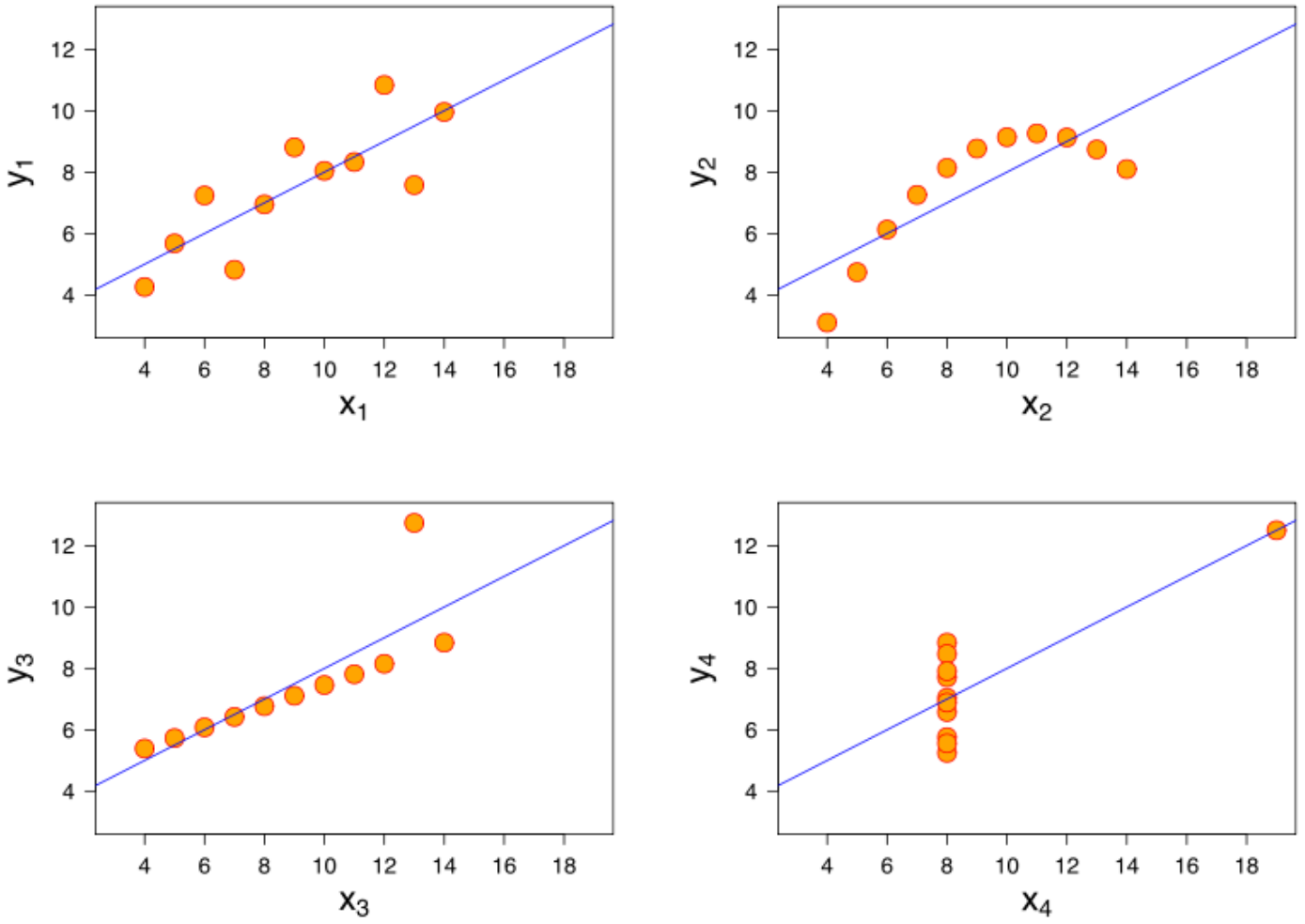
Anscombe's quartet

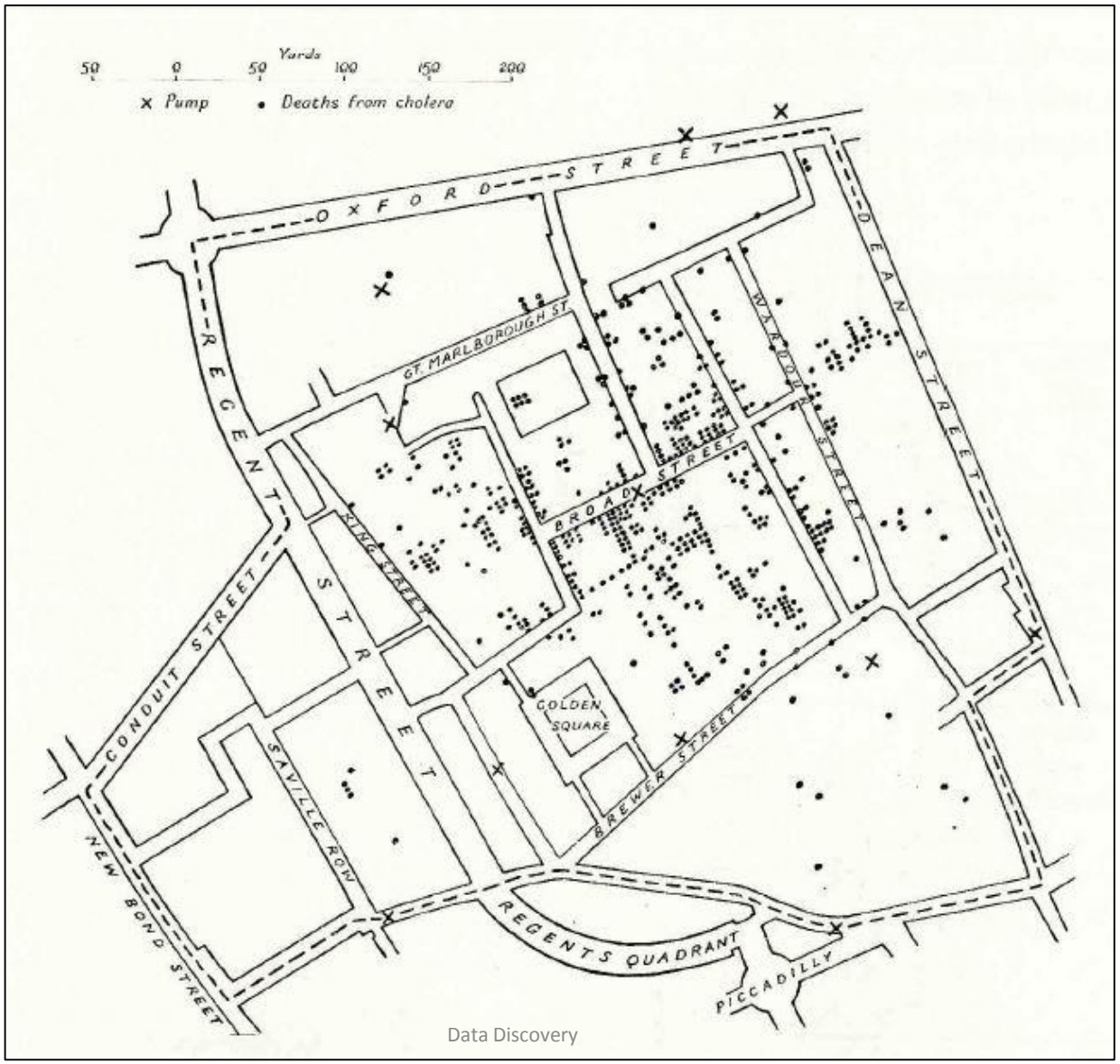
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

For all four datasets:

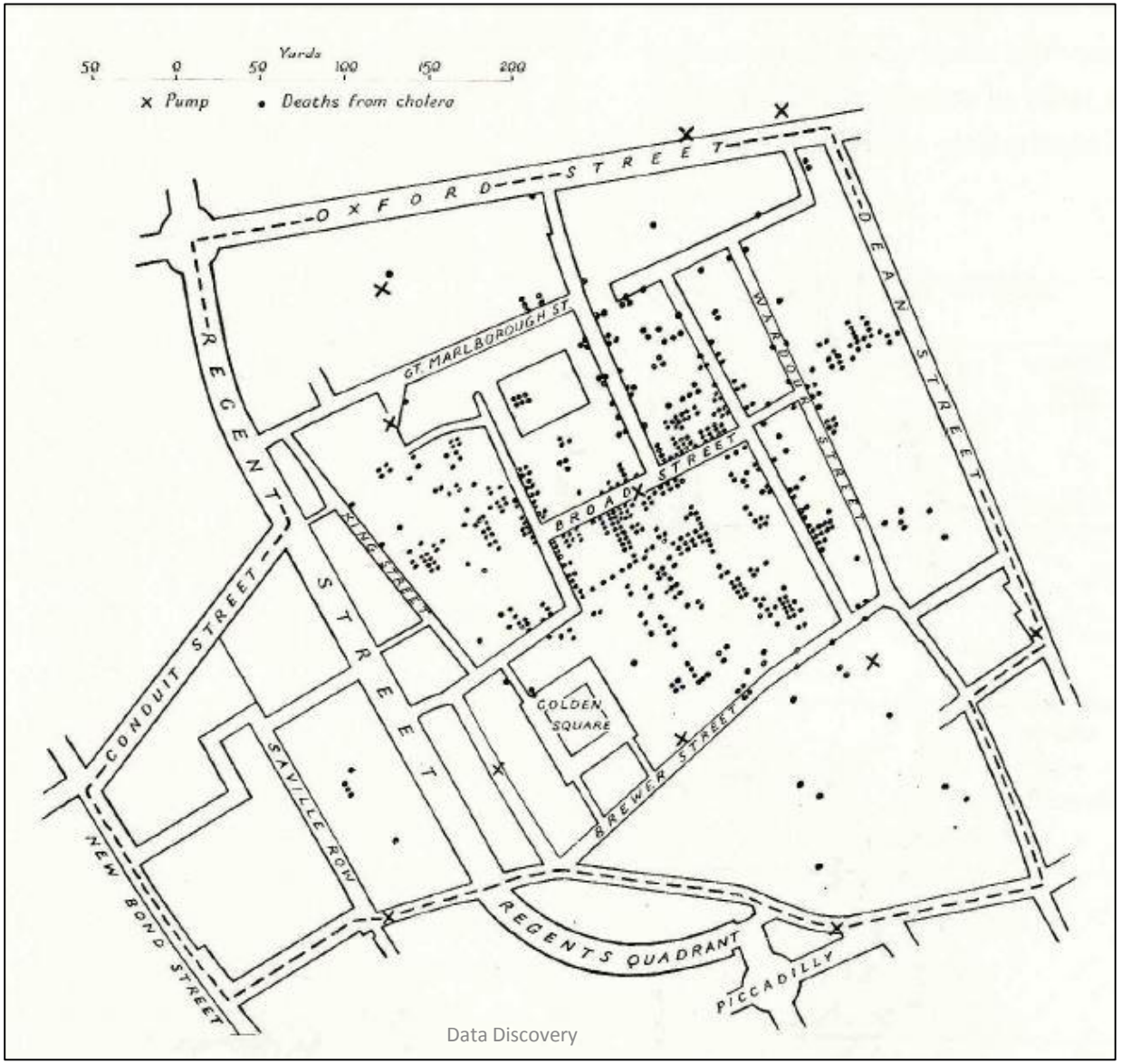
Property	Value
Mean of *x* in each case	9 (exact)
Variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

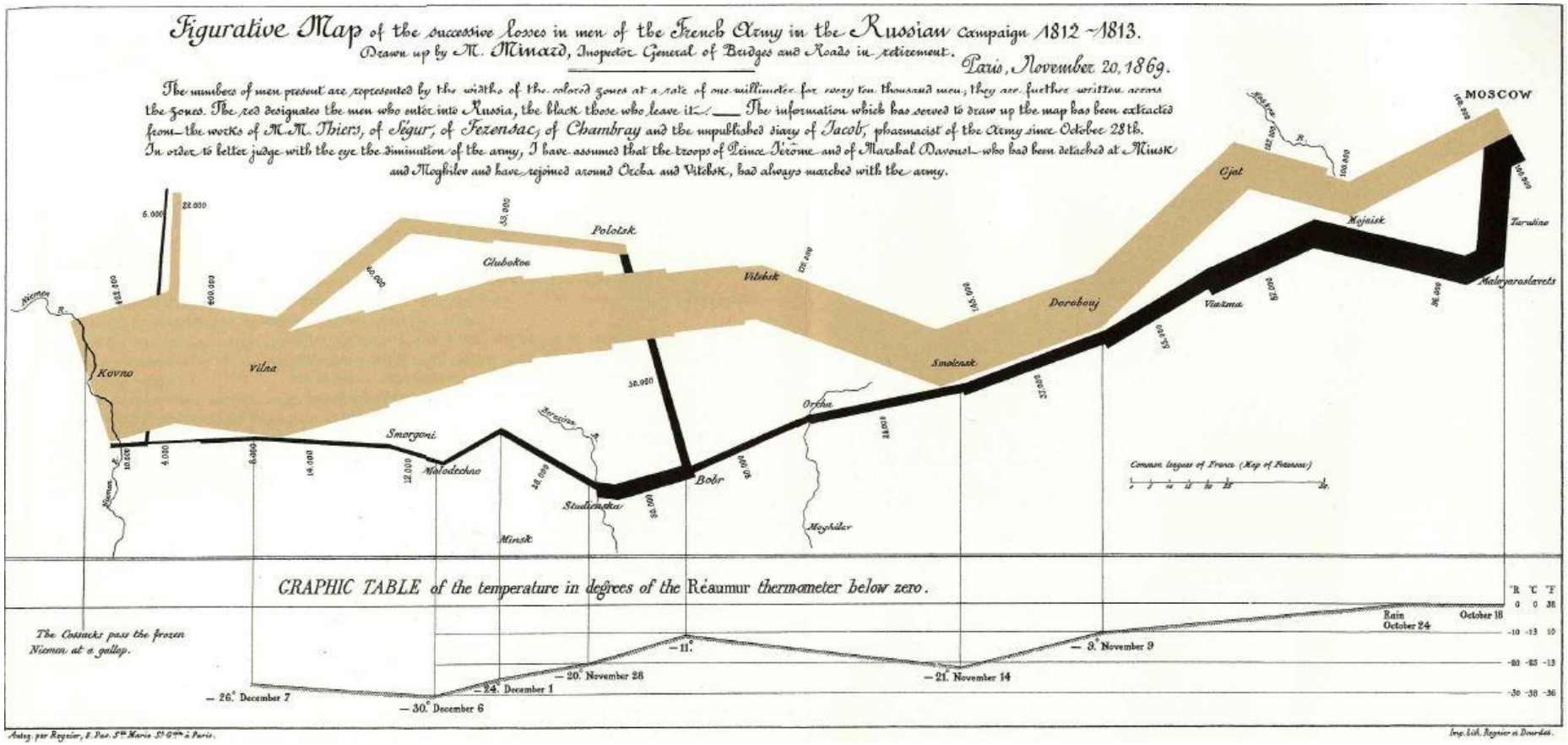
Anscombe's Quartet - Graphed



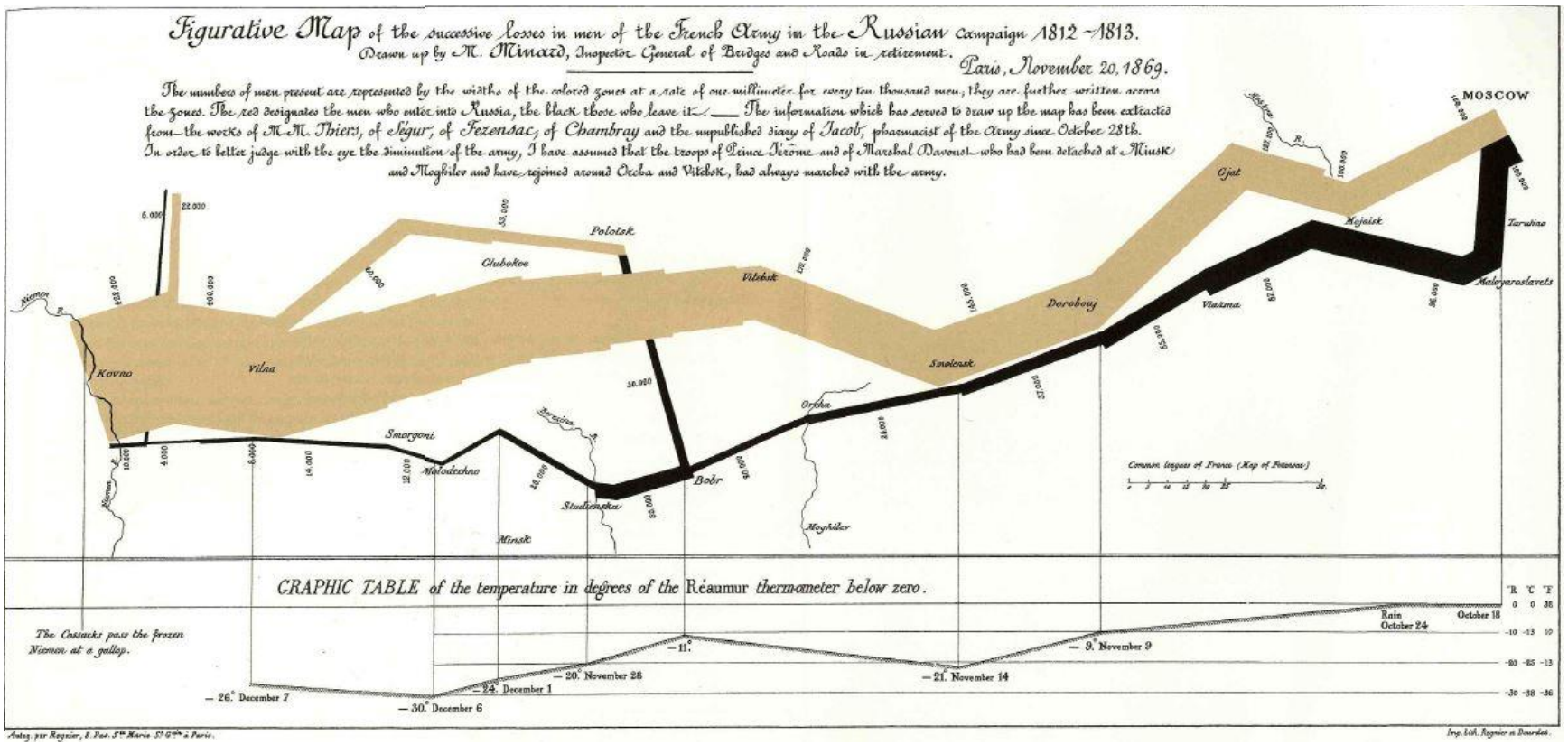

50 0 50 Yards 100 150 200
× Pump • Deaths from cholera
OXFORD STREET
REGENT STREET
GT. MARLBOROUGH ST.
KING STREET
BROAD STREET
WARDOUR STREET
DEAN STREET
CONDUIT STREET
GOLDEN SQUARE
BREWER STREET
SAVILLE ROW
NEW BOND STREET
REGENTS QUADRANT
PICCADILLY

Data Visualisation-1- John Snow's map of London Cholera Epidemic of 1854




Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 - 1813.
Drawn up by M. Minard, Inspector General of Bridges and Roads in retirement.
Paris, November 20, 1869.
The numbers of men present are represented by the widths of the colored zones at a rate of one millimetre for every ten thousand men; they are furthermore written across the zones. The red designates the men who enter into Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M.M. Thiers, of Ségur, of Fezensac, of Chambray and the unpublished diary of Jacob, pharmacist of the Army since October 28th.
In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jerome and of Marshal Davoust who had been detached at Minsk and Moghilev and have rejoined around Orcha and Vitebsk, had always marched with the army.
MOSCOW
Polotsk
Chlubokoe
Vitebsk
Dorobouj
Smolensk
Mojaisk
Taroutino
Malo-yaroslavets
Viazma
Gjat
Kovno
Vilna
Smorgoni
Molodechno
Minsk
Studienska
Bobr
Orcha
Moghilev
Niemen R.
Borizow
Common leagues of France (Map of Peutinger)
GRAPHIC TABLE of the temperature in degrees of the Réaumur thermometer below zero.
The Cossacks pass the frozen Niemen at a gallop.
— 26° December 7
— 30° December 6
— 24° December 1
— 20° November 26
— 11°
— 21° November 14
— 3° November 9
Rain October 24
October 18
R °C °F
0 0 32
-10 -13 10
-20 -25 -13
-30 -38 -36

Data Visualisation-2: Charles Minard's graph (1869) of French invasion of Russia in 1812



Visualisation Types/Purpose

1. Exploratory Visualization
2. Explanatory visualization

Purpose of Visualisation - to move information from point A to point B

- In Exploratory Visualisation:
  - A: Dataset
  - B: Designer's mind
- In Explanatory Visulaisation:
  - A: Designer's mind
  - B: Reader's mind

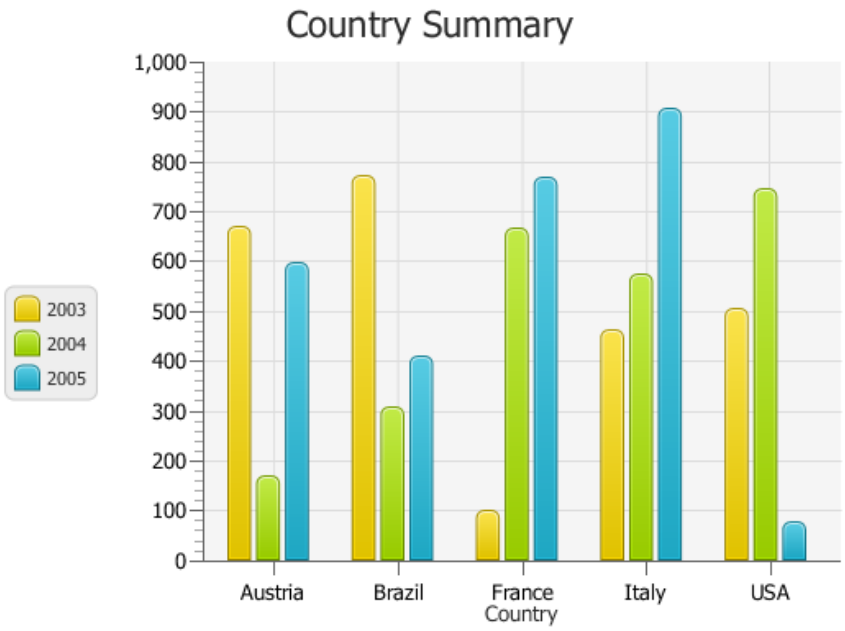
Charts and Graphs in Visualisation

1. Bar Chart
2. Line Chart
3. Pie Chart
4. Map Chart
5. Scatter Plot
6. Bubble Chart
7. Histogram Chart
8. Heat Chart
9. TreeMap
10. Box-and-whisker Plot

Bar Chart

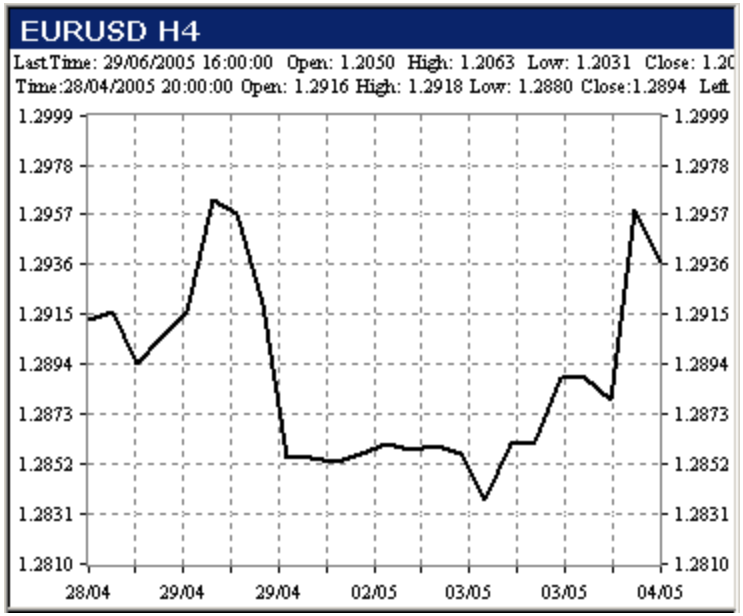
- Quick to compare information, revealing highs and lows at a glance.
- Effective with numerical data that splits nicely into different categories.

- When to use bar charts:
  - **Comparing data across categories**



Line Chart

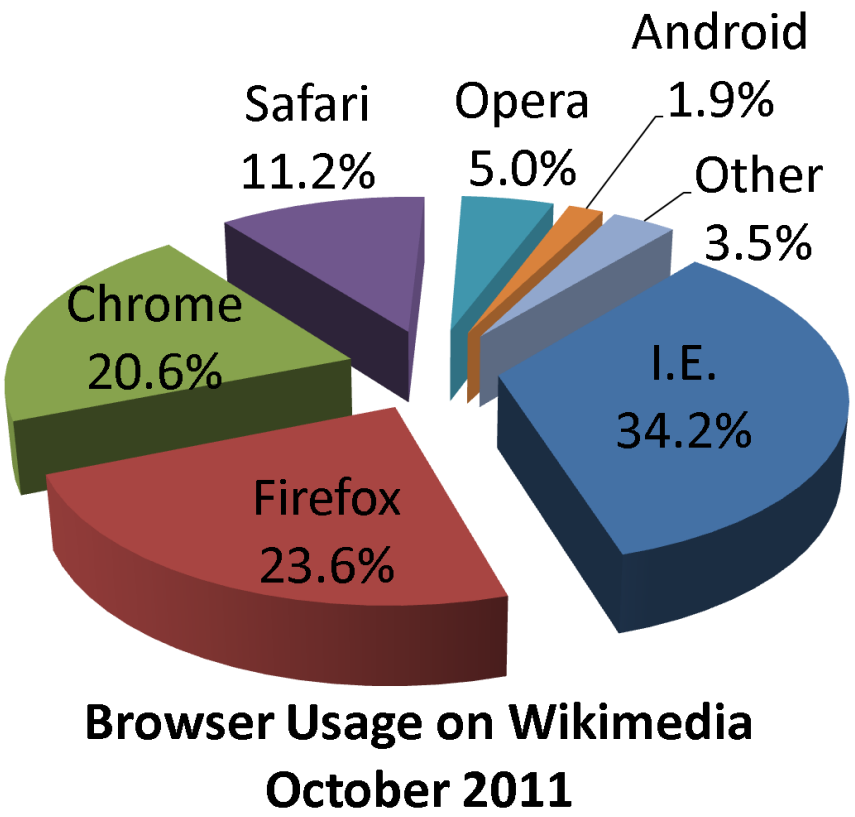
- Connect individual numeric data points.
- Simple way to visualize a sequence of values.
- When to use line charts:
  - **Viewing trends in data over time**



Pie Chart

- To show relative proportions – or percentages – of information
- the most commonly mis-used chart type
- Not good for comparing data

- When to use pie charts:
- **Showing proportions.**



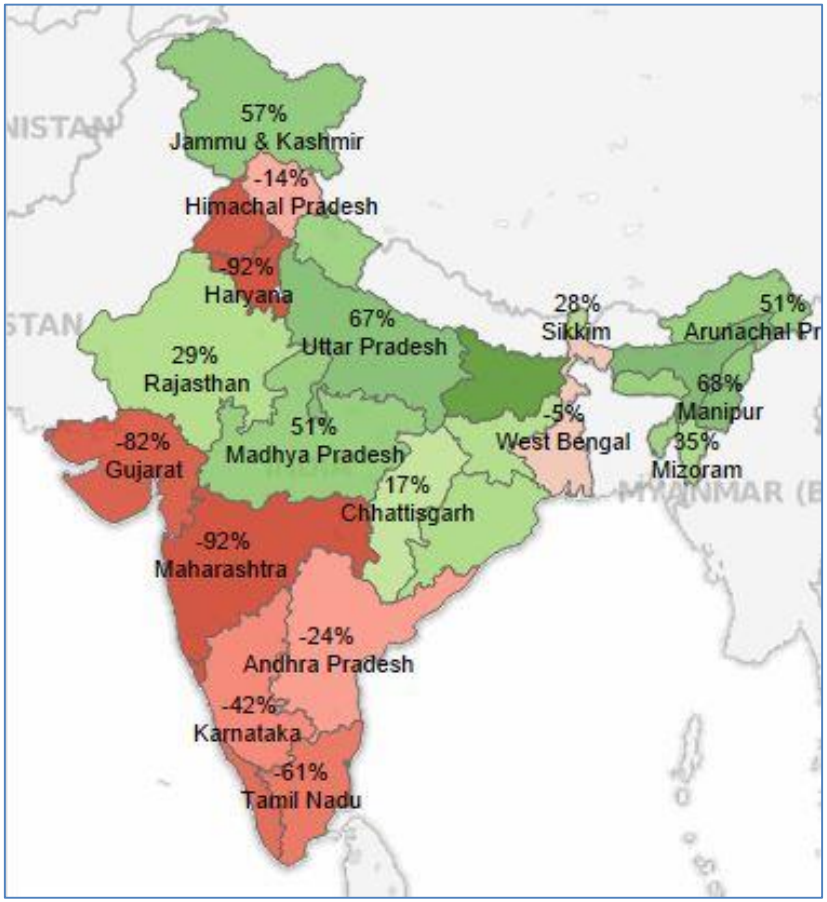
Browser Usage on Wikimedia October 2011

Map Chart

- New Chart Type
- When you have any kind of location data

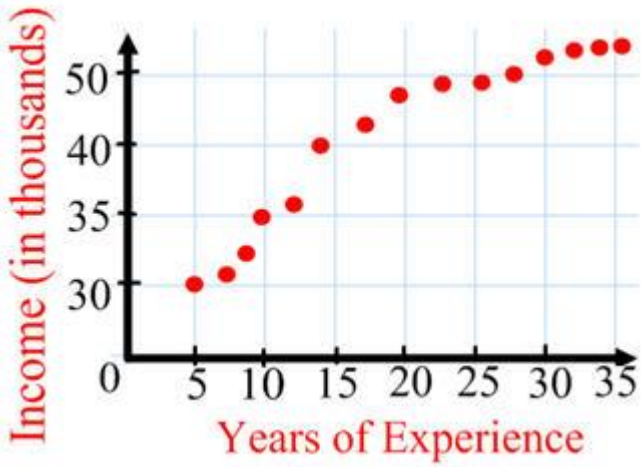
- When to use maps:
- **Showing geocoded data.**

Examples: Insurance claims by state, product export destinations by country etc.



Scatter Plot

- To see how different pieces of information relate
- effective way to get a sense of trends, concentrations and outliers – identify areas to focus further investigation
- When to use scatter plots: **Investigating the relationship between different variables**



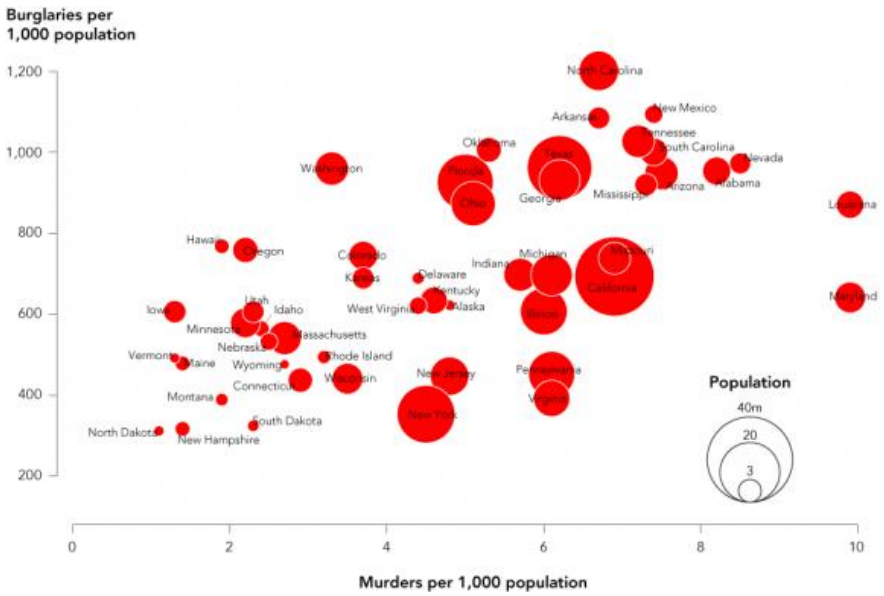
Bubble Chart

- A technique to accentuate data on scatter plots or maps
- Not a separate type of visualization
- varied size of circles provides meaning about the data

- When to use bubbles:

• Showing the concentration of data along two axes.

Examples: sales concentration by product and geography

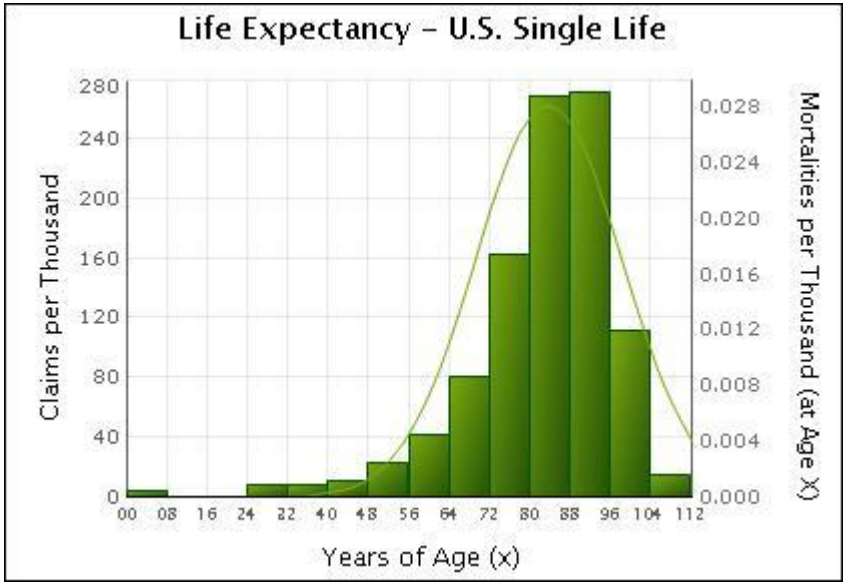


Histogram Chart

- To see how data distributed across groups
- To understand which categorization approach makes sense

- When to use histograms:

• **Understanding the distribution of your data.** Examples: Number of customers by company size, student performance on an exam



Heat Chart

- Excellent for comparing data across two categories using colour
- Shows where intersection of the categories is strongest and weakest

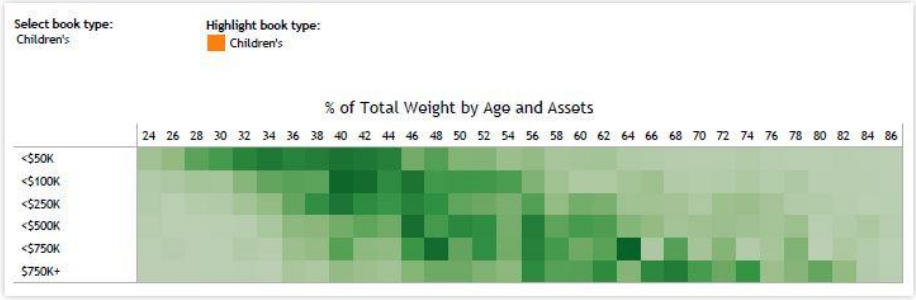


Figure 16: Who buys the most books?
In this market segmentation analysis, the heat map reveals a new campaign idea. High-income households of people in their sixties buy children's books. Perhaps it's time for a new grandparent-oriented campaign?

- When to use heat maps:

• **Showing the relationship between two factors.** Examples: product adoption across regions

TreeMap

- To see data at a glance and discover how the different pieces relate to the whole
- Uses a series of rectangles, nested within other rectangles, to show hierarchical data as a proportion to the whole

- When to use treemaps:

• Showing hierarchical data as a proportion of a whole:

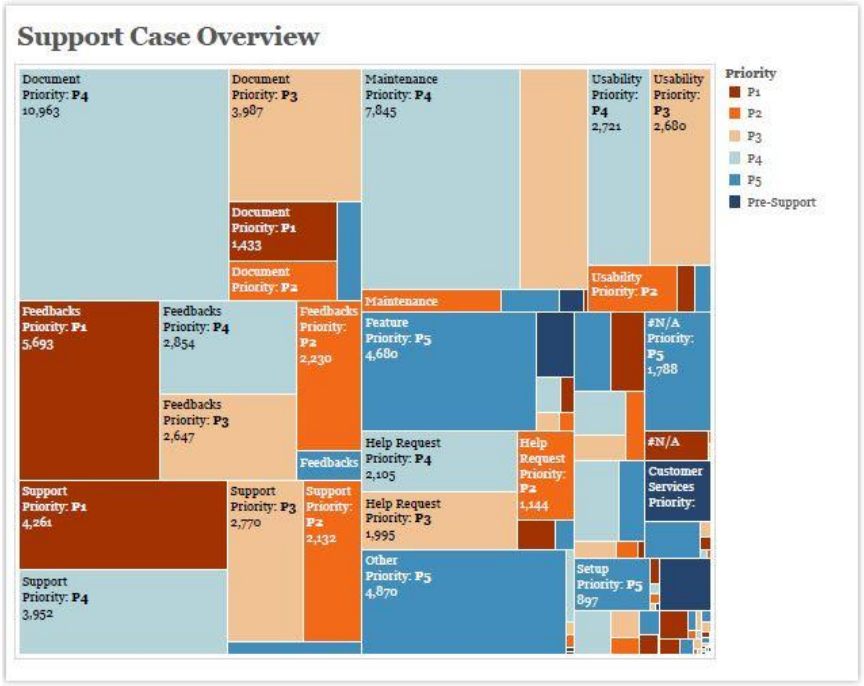


Figure 18: Support Cases at a Glance

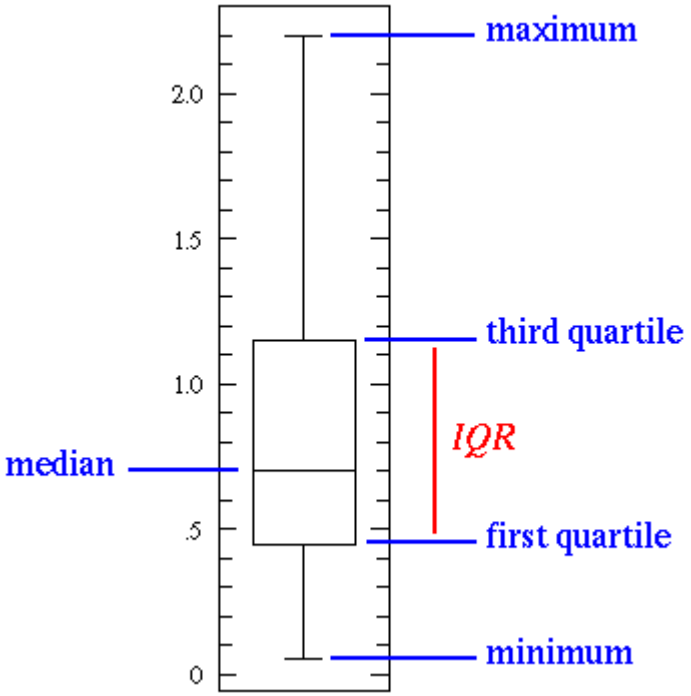
This treemap shows all of a company's support cases, broken by case type, and also priority level. You can see that Document, Feedback, Support and Maintenance make up the lion share of support cases. However, in Feedback and Support, P1 cases make up the most number of cases, whereas most other categories are dominated by relatively mild P4 cases.

Box-and-whisker Plot

- Shows distributions of data
- Box contains the median of the data along with the 1st and 3rd quartiles
- Whiskers typically represents data within 1.5 times the Inter-quartile Range (or Maximum/Minimum)

- When to use box-and-whisker plots:

• **Showing the distribution of a set of a data:** Examples: understanding data at a glance, seeing how data is skewed towards one end, identifying outliers



Analytics for Audit-Tool Fitment Criteria

Strong Data Discovery and Data Visualisation capabilities

Reasonably easy to learn- quick solution development time

Operate from Desktop/Laptop of the Auditors

Capable of scaling up to handle large volumes of data through a central server based model

Reasonably priced

Data Discovery/Visualisation Tools
(Selected by IAAD)

- **QlikView**:

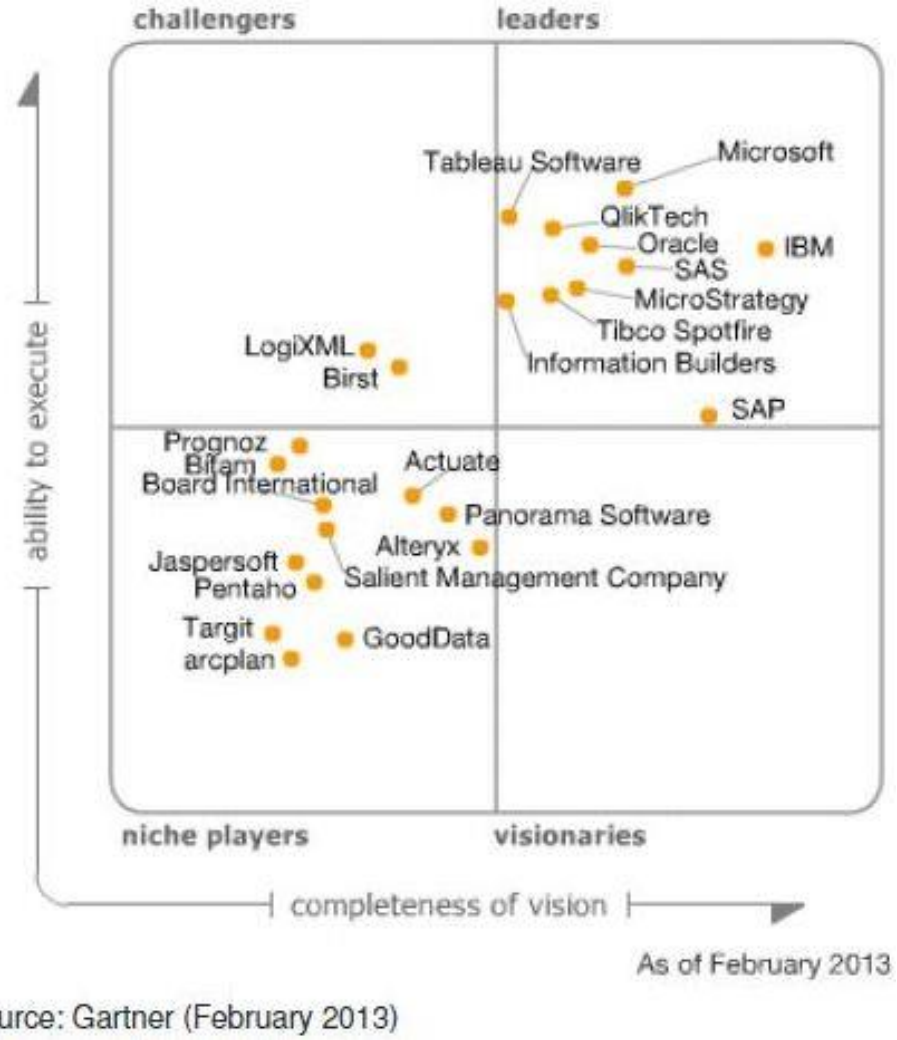
http://www.qlikview.com/

- **Tableau** :

http://www.tableausoftware.com/

- **Microsoft BI** :

(Pivot/PowerPivot)



Source: Gartner (February 2013)

Gartner's Magic Quadrant for Analytics - 2014

Business Intelligence & Analytics Platforms

Source: Gartner (February 2014)

Advanced Analytics Platforms

Source: Gartner (February 2014)

In Memory Analytics Performance

- Compared performance for data stratification.
- **Stratification** on Billed Units, summation on Fixed Charge and Energy Charge
- 2 hours 10 minutes in IDEA, 3 seconds in QlikView

Stratification of Billed Units

Bin	Fixed Charge	Energy Charge	# Records	% Fixed Charge	% Energy Charge	% Records
	3,359,422,230.14	**42,004,418,247.97**	**12,010,919**	**100.00%**	**100.00%**	**12010919**
0-200	771,127,002.94	2,044,158,177.75	5,973,425	22.95%	4.87%	5973425
200-400	396,319,867.56	3,366,764,728.93	2,873,430	11.80%	8.02%	2873430
400 and Above	2,191,975,359.64	36,593,495,341.29	3,164,064	65.25%	87.12%	3164064

THANK YOU