MANCHESTER 1824

The University of Manchester

Analysing Unstructured Data

MiSET 2014
11/06/2014

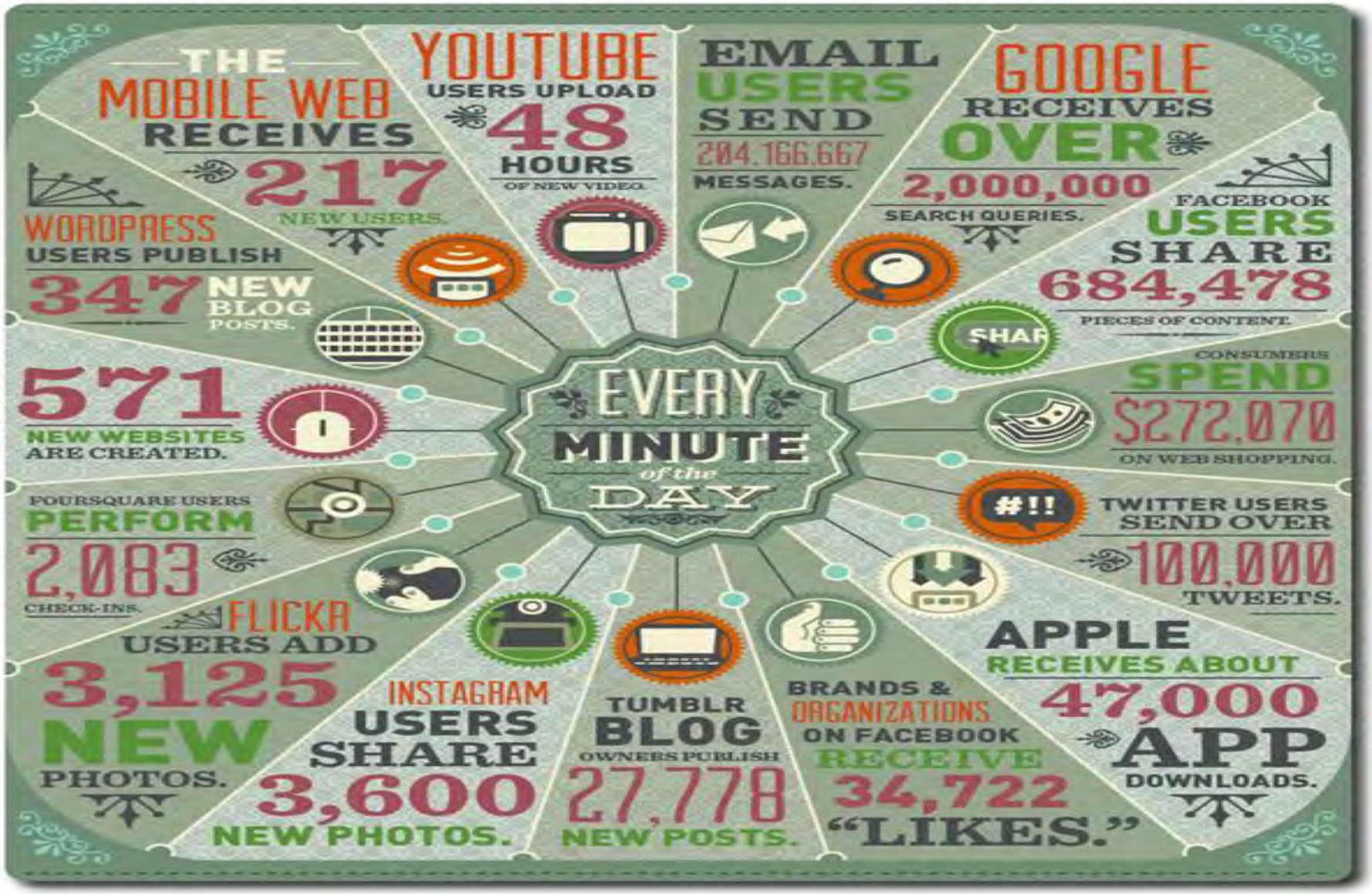
Dr Abdullah Gök

Manchester Institute of Innovation Research, University of Manchester

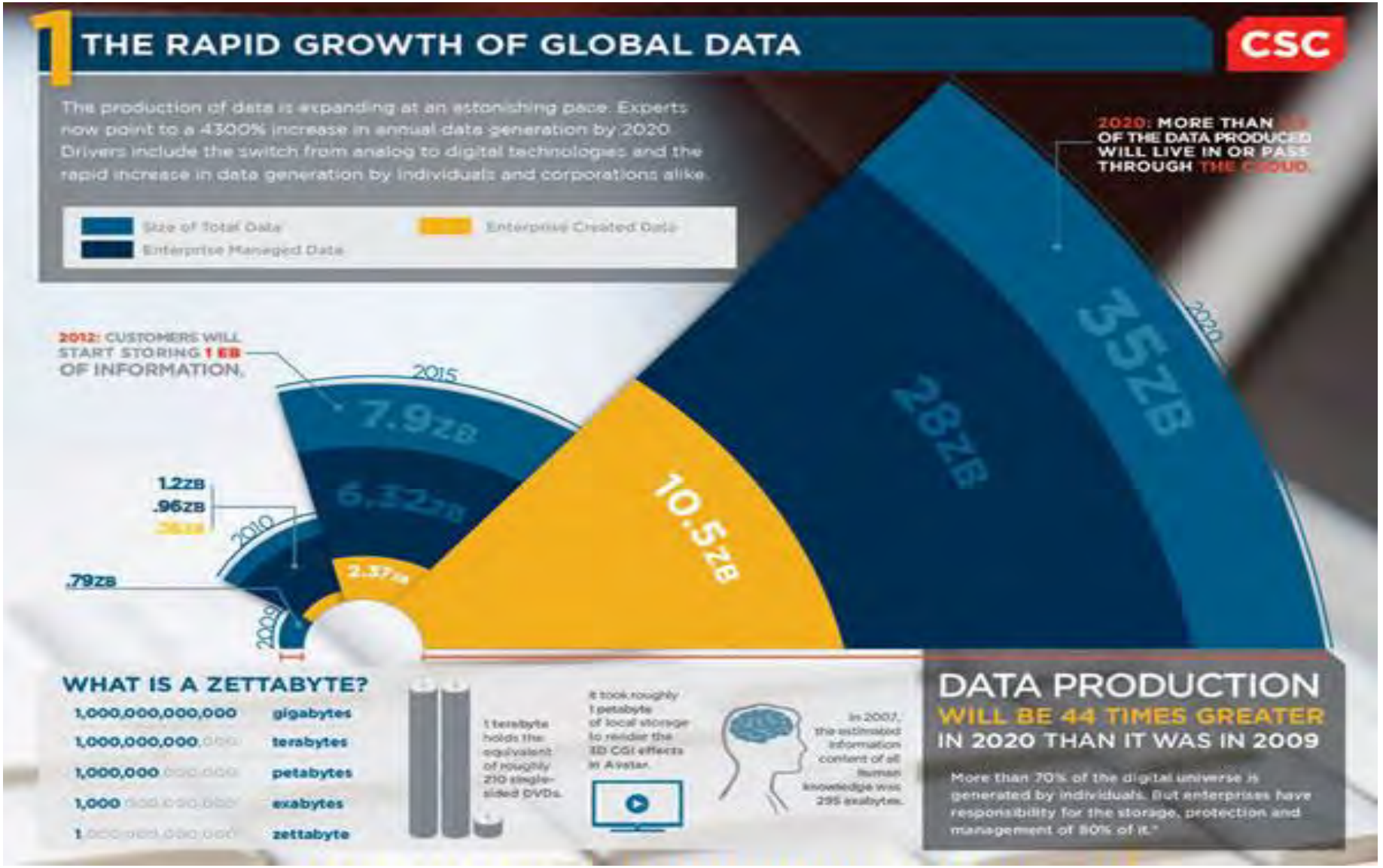
Outline

- Big Data Trends
- Types of Data and Research
- Seven Areas of Text Mining with examples from innovation studies
- Web Mining
- Conclusions

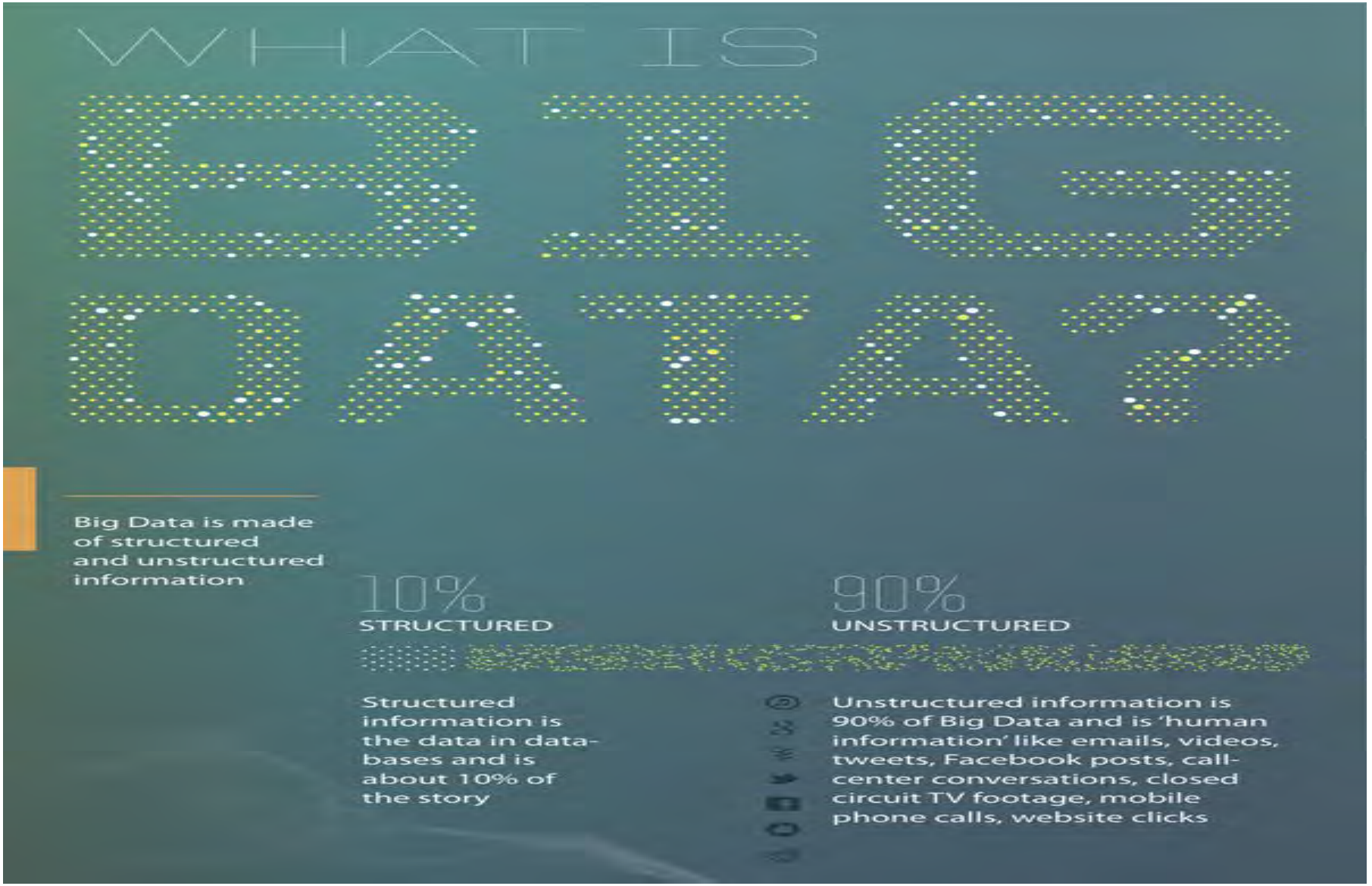
Big data in every minute



How big is the data becoming?



Big Data is Unstructured



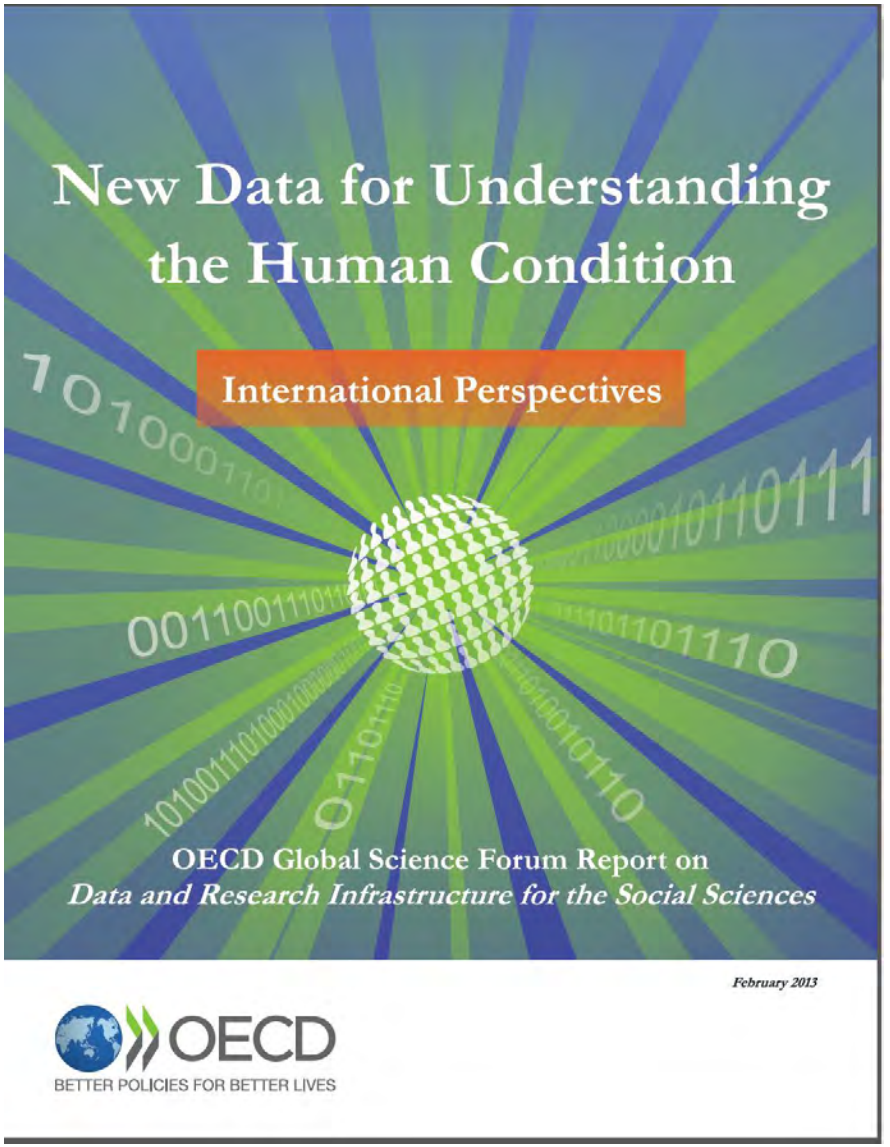
What is Big Data?

- Is the term "big data" a misnomer (Boyd and Crawford)?
- "Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse" (Manyika et al., 2011) + very similar RCUK definition
- Original 3V Definition (Laney, 2001):
  - Volume: increased amount of data: transactional, unstructured, machine-to-machine. storage is no longer a big issue.
  - Velocity: speed and recency of data is accelerating tremendously
  - Variety: unstructured data is dominating
- SAS (2012) additions:
  - Variability: periodic peaks
  - Complexity: multiple sources

Rise of Big Data in Research

- US: Obama administration invested $200 billion on big data
- Developments in the UK
  - One of the "Eight Great Technologies"
  - Alan Turing Institute
  - Elsevier and UCL's joint "UCL Big Data Institute"
  - Ministerial approval given for Life Study which will be the UK's largest national birth cohort study collecting data on up to 100,000 babies born in the UK

Big Data and Social Science



Types of Data

- Structure:
  - Structured: there is a schema (i.e. there are variables and standardised choices)
  - Unstructured: there is no schema (i.e. based on text, pictures, video etc.)
  - Semi-structured: there are both structured and unstructured elements
- Collection Methods:
  - Obtrusive: researcher actively collects and influences data (research subjects are aware of data collection)
  - Unobtrusive: data is indirect (research subjects are not aware of data collection)

Obtrusive versus Unobtrusive Data

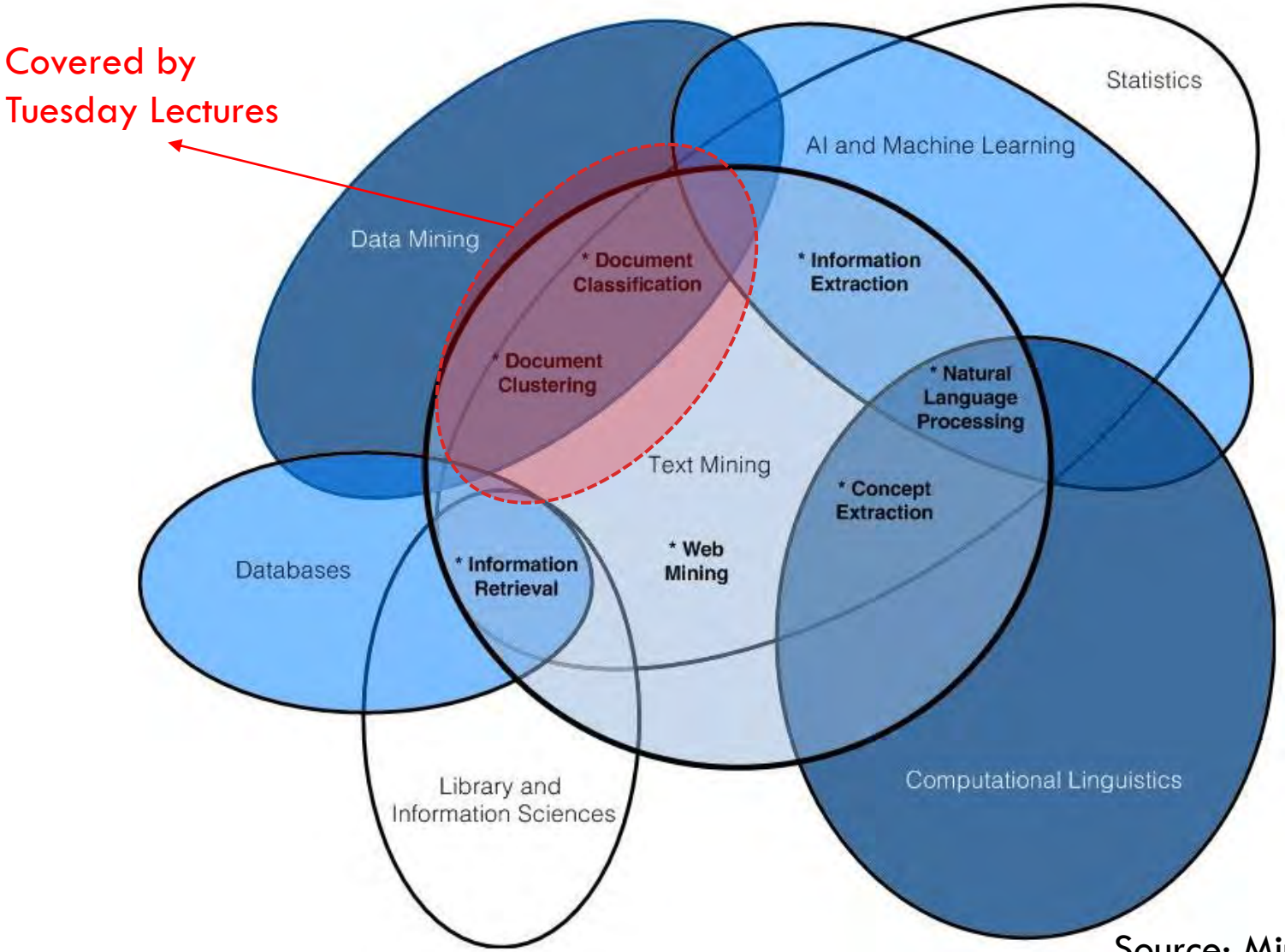
- Advantages:
  - No observation effect/reactivity/Hawtorne effect (particularly strategic answering)
  - Easier and cost effective to collect
  - Easier to adjust
  - Easier to replicate
  - Easier to study evolution over time
- Disadvantages:
  - Validity issues (and therefore more difficult to interpret)
  - Data availability dictates the scope
  - Isolated from the context

Recap on Big Data

- Data is increasing in volume, velocity and variety
- New data sources become available, especially on human interaction
- Unobtrusive data provides advantages
- Big data is mostly unstructured
- Big Question:

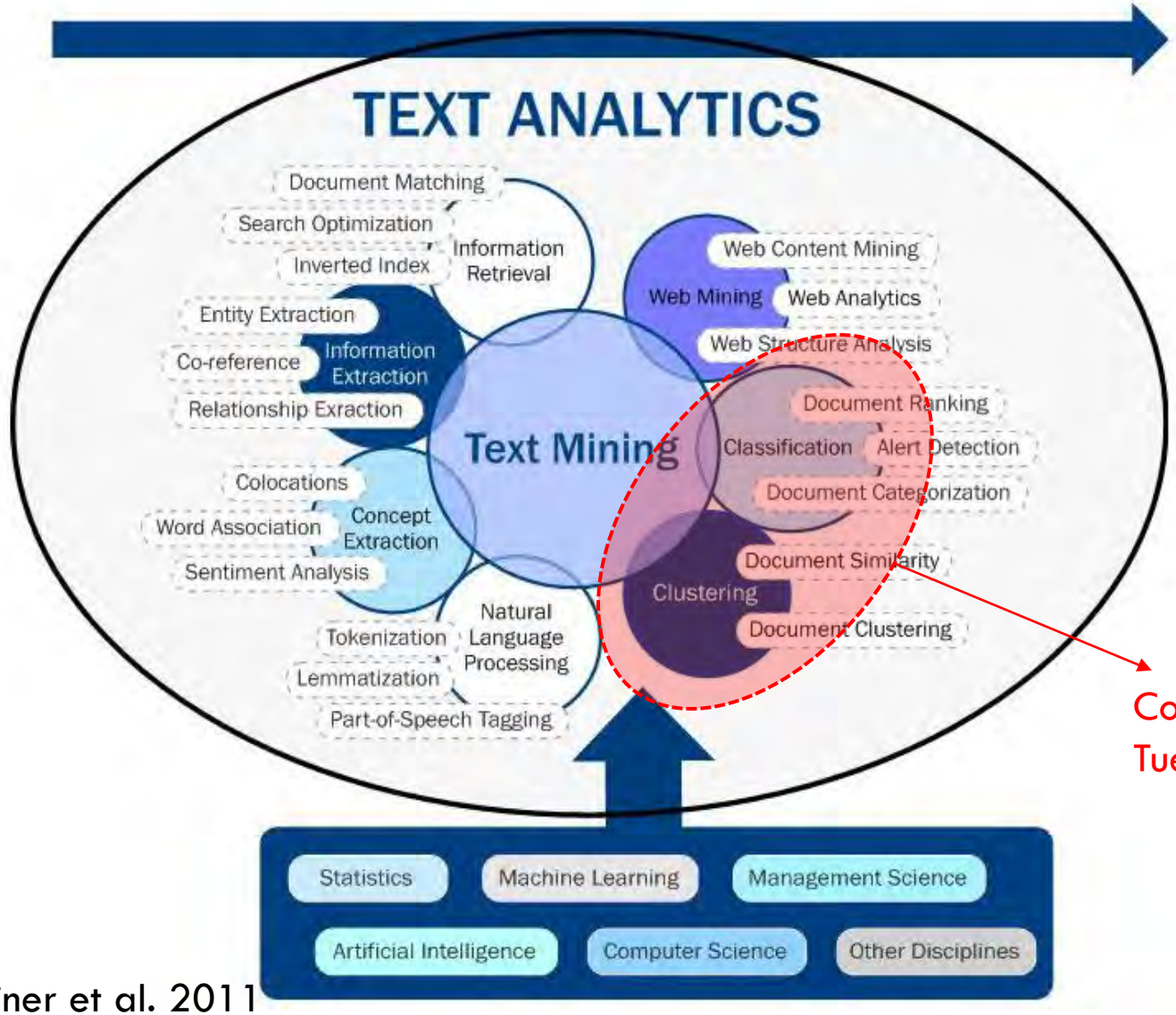
How to analyse unstructured data?

Seven Areas of Text Mining



Source: Miner et al. 2011

Seven Areas of Text Mining



Covered by Tuesday Lectures

Source: Miner et al. 2011

Natural Language Processing

- Foundation of text mining: computational linguistics
- Low level language processing
- Usual procedure
  - Choose the scope of the text to be processed (documents, paragraphs, etc.).
  - Tokenize: Break text into discrete words called tokens
  - Remove stopwords
  - Stem: Remove pre fixes and suffixes to normalize words (car, cars, car's, cars' >> car)
  - Normalize spelling: (analyse, analyse >> analyse)
  - Detect sentence boundaries: Mark the ends of sentences
  - Normalize case: Convert the text to either all lower or all upper case
  - Create a vector matrix

Information Extraction

- Syntax versus semantics
  - Bag of words approach
  - Semantic approach
- Part of speech tagging:
  - NOUN, VERB, etc.
- Named Entity Recognition:
  - Persons, Organisations, Locations, Time, Quantity, Monetary values, percentages, etc.
- Strategies for Entity Extraction
  - Rule based approaches
  - Statistical approaches

Information Extraction: Example

- See Shapira et al. (2013)
- Problem: how to define industries and identify firms within industries
- Conventional solution: Use self-declared industrial classification
- Issue:
  - It is too rigid: it does not work on emerging industries
  - Self declaration of single value is often misleading
  - It often does not reflect on a variety of activities firms do simultaneously
- Context: Defining green goods industry and locating firms within it
- Solution: Extract industrial information from the firm trade description information

Information Extraction: Example

SIC & trade description

Trade description :
Design, manufacture and sale of M2G, a boiler effiency tecnology, which is proven to reduce energy consumption on commercial boilers by up to 35%.

UK SIC (2007) Codes :
Primary Code :
25300 - Manufacture of steam generators, except central heating hot water boilers

All Codes :
25300 - Manufacture of steam generators, except central heating hot water boilers
71121 - Engineering design activities for industrial process and production

US SIC Codes [derived from UK SIC (2007) Codes]:
Primary Code :
3433 - Heating equipment, except electric and warm air furnaces

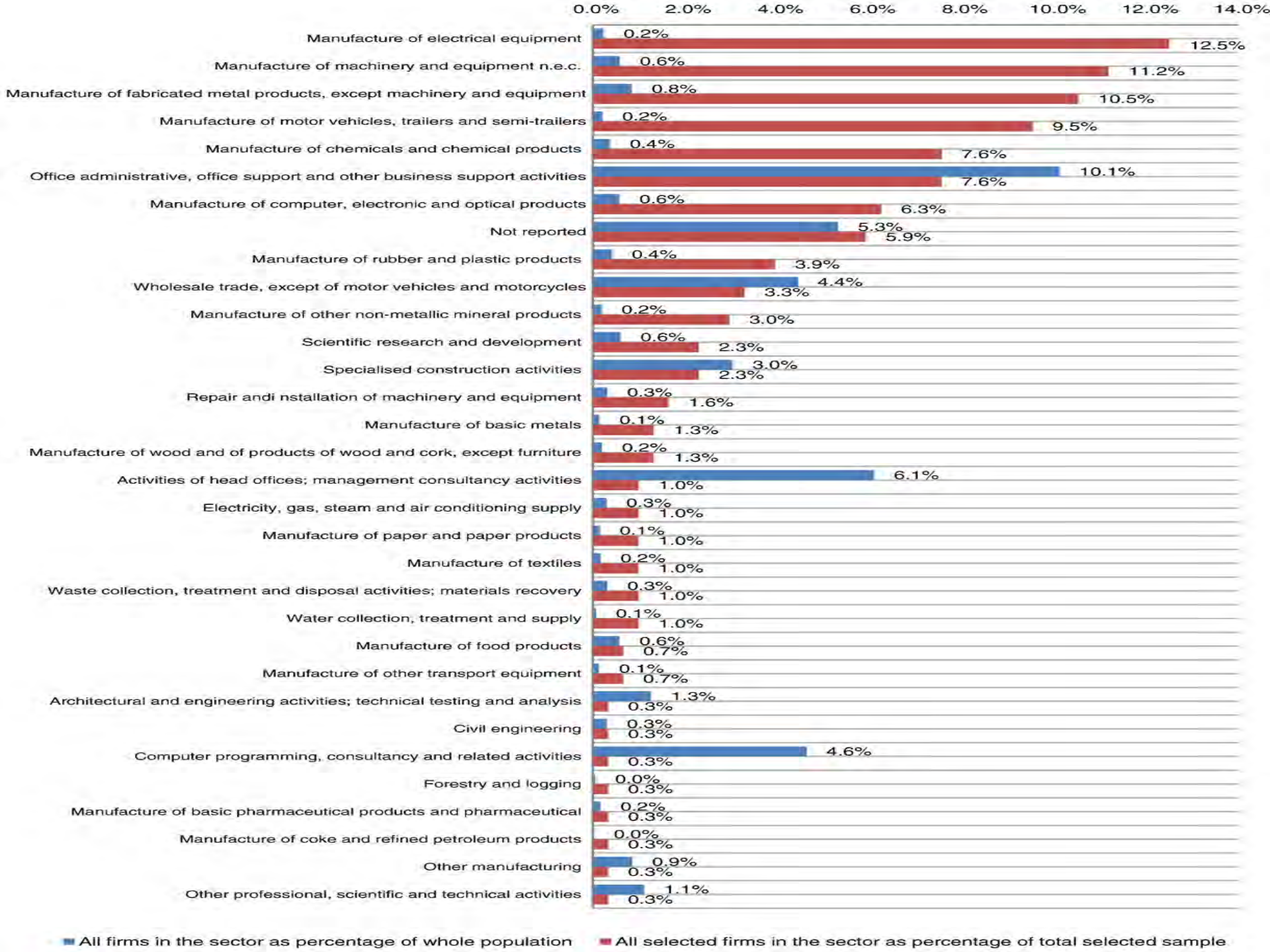
All Codes :
3433 - Heating equipment, except electric and warm air furnaces
8711 - Engineering services

Full overview

The company, formerly known as Sabien Technology Group Ltd., is a UK-based manufacturer of energy saving systems with emphasis on the reduction of carbon emissions. Founded in March 2004, the company has its headquarters in Watford, United Kingdom.

The company specializes in providing proven and commercially viable technology to reduce carbon emissions and energy usage for private and public organisations. The company offers energy efficient products, including M2G and M3G, to blue chip companies across the UK, Hong Kong, China, Italy and Ireland. M2G is a micro processor based product which helps gas & oil reduction required for commercial heating. Companies using M2G include O2, LloydsTSB, NHS Trust, KPMG, Bank of England, and Radisson Hotel Group. For commercial air conditioning, M3G is the solution with zero maintenance, reducing energy and carbon emissions by up to 50 percent. Some of the companies using M3G are Sony, Campbell Soup, Tesco, McDonald's, HSBC Bank - London, Scottish & Newcastle Breweries, Starbucks, Mitsubushi, and Kentucky Fried Chicken.

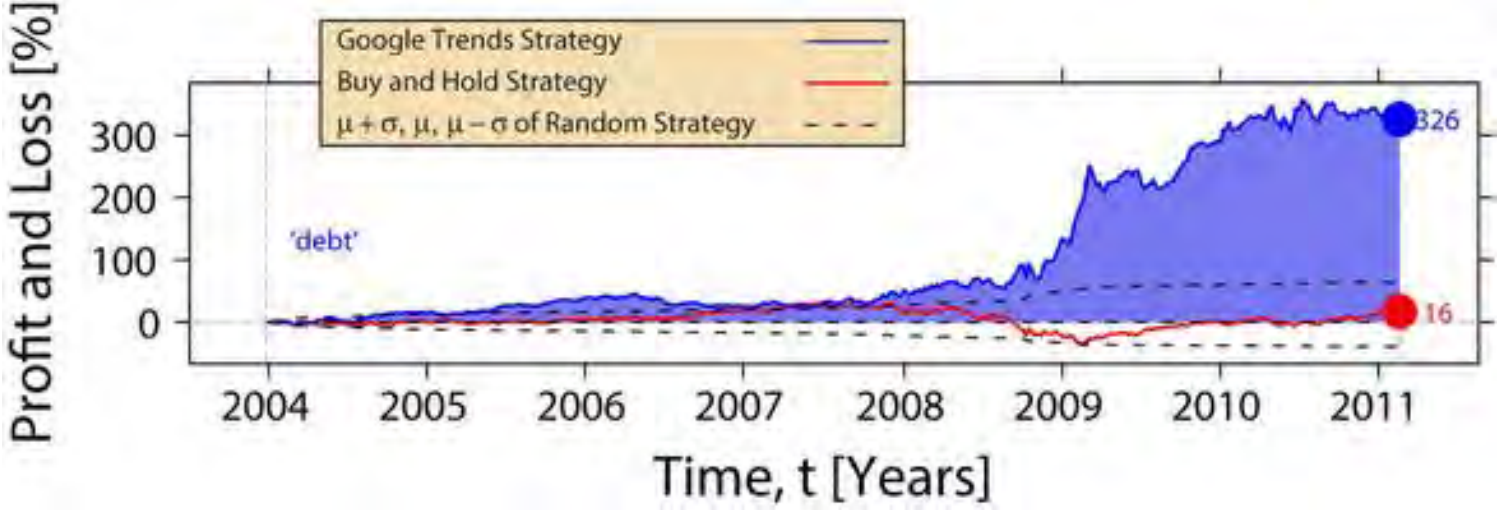
The company's products are Carbon Trust approved. The products also qualify for the Enhanced Capital Allowance scheme, which enables a business to claim 100 percent first-year capital allowances on their spending on qualifying plant and machinery. The company aims to assure customers of its continuous commitment to demonstrable environmental management standards and assist them to use products in an environmentally sensitive way.


0.0%
2.0%
4.0%
6.0%
8.0%
10.0%
12.0%
14.0%
Manufacture of electrical equipment 0.2% 12.5%
Manufacture of machinery and equipment n.e.c. 0.6% 11.2%
Manufacture of fabricated metal products, except machinery and equipment 0.8% 10.5%
Manufacture of motor vehicles, trailers and semi-trailers 0.2% 9.5%
Manufacture of chemicals and chemical products 0.4% 7.6%
Office administrative, office support and other business support activities 10.1% 7.6%
Manufacture of computer, electronic and optical products 0.6% 6.3%
Not reported 5.3% 5.9%
Manufacture of rubber and plastic products 0.4% 3.9%
Wholesale trade, except of motor vehicles and motorcycles 4.4% 3.3%
Manufacture of other non-metallic mineral products 0.2% 3.0%
Scientific research and development 0.6% 2.3%
Specialised construction activities 3.0% 2.3%
Repair andi nstallation of machinery and equipment 0.3% 1.6%
Manufacture of basic metals 0.1% 1.3%
Manufacture of wood and of products of wood and cork, except furniture 0.2% 1.3%
Activities of head offices; management consultancy activities 6.1% 1.0%
Electricity, gas, steam and air conditioning supply 0.3% 1.0%
Manufacture of paper and paper products 0.1% 1.0%
Manufacture of textiles 0.2% 1.0%
Waste collection, treatment and disposal activities; materials recovery 0.3% 1.0%
Water collection, treatment and supply 0.1% 1.0%
Manufacture of food products 0.6% 0.7%
Manufacture of other transport equipment 0.1% 0.7%
Architectural and engineering activities; technical testing and analysis 1.3% 0.3%
Civil engineering 0.3% 0.3%
Computer programming, consultancy and related activities 4.6% 0.3%
Forestry and logging 0.0% 0.3%
Manufacture of basic pharmaceutical products and pharmaceutical 0.2% 0.3%
Manufacture of coke and refined petroleum products 0.0% 0.3%
Other manufacturing 0.9% 0.3%
Other professional, scientific and technical activities 1.1% 0.3%
All firms in the sector as percentage of whole population
All selected firms in the sector as percentage of total selected sample

Information Retrieval

- Indexing, searching, and retrieving documents from large text databases
- Examples:
  - Google Trends: http://www.google.co.uk/trends/explore#q=%2Fm%2F03p5rs%2C%20%2Fm%2F0416x13%2C%20Nobel%20Physics&cmpt=q
  - Google Ngram Viewer https://books.google.com/ngrams
    - 30 million fiction and non-fiction books published since AD 1500
    - Natural Language Processing ready

- Examples in Social Science Research:
  - Preis et al. (2013) Quantifying Trading Behavior in Financial Markets Using *Google Trends*. Scientific Reports
  - Preis et al. (2012) Quantifying the Advantage of Looking Forward. Scientific Reports
  - Moat et al. (2013) Quantifying *Wikipedia* Usage Patterns Before Stock Market Moves. Scientific Reports.



Concept Extraction

(Reuters) - Research In Motion Ltd said on Tuesday its subscriber base has risen to 80 million from the 78 million it reported earlier this year, surprising many on Wall Street and sending its shares up more than 3 percent.

Most analysts had expected RIM, for the first time in its history, to begin losing subscribers in the recently completed quarter as it has rapidly lost market share in North America to Apple's snazzier iPhone and Samsung's Galaxy devices.

Document Summarisation:

Research In Motion subscriber base has risen to 80 million sending its shares up more than 3 percent. Most analysts had expected RIM, for the first time in its history, to begin losing subscribers.

Theme/Concept Extraction:

- Themes: Subscriber base, Shares, Market share
- Concepts: Technology, Business, Smart phones

Sentiment Analysis:

Entity	Type	Sentiment
RIM	Entity: Organisation	Positive
Apple	Entity: Organisation	Positive
Samsung	Entity: Organisation	Neutral
Smart Phones	Concept	Neutral
Subscriber Base	Theme	Positive
Shares	Theme	Positive
Market Share	Theme	Positive

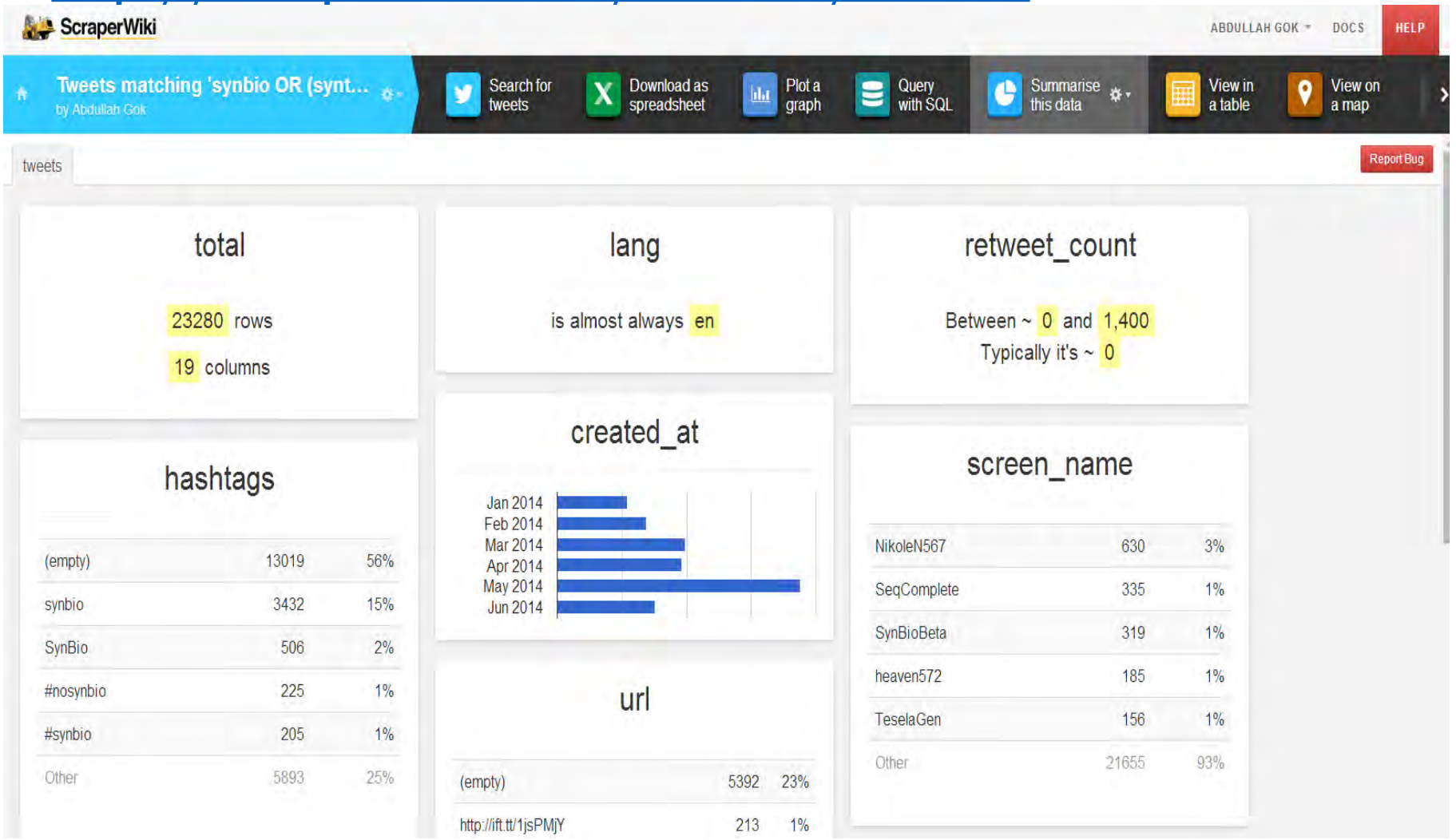
Example adapted from Datasciencecentral.com

Sentiment Analysis: Example

- Veltri (2013) Microblogging and nanotweets: Nanotechnology on Twitter. *Public Understanding of Science* October 2013 vol. 22 no. 7 832-849

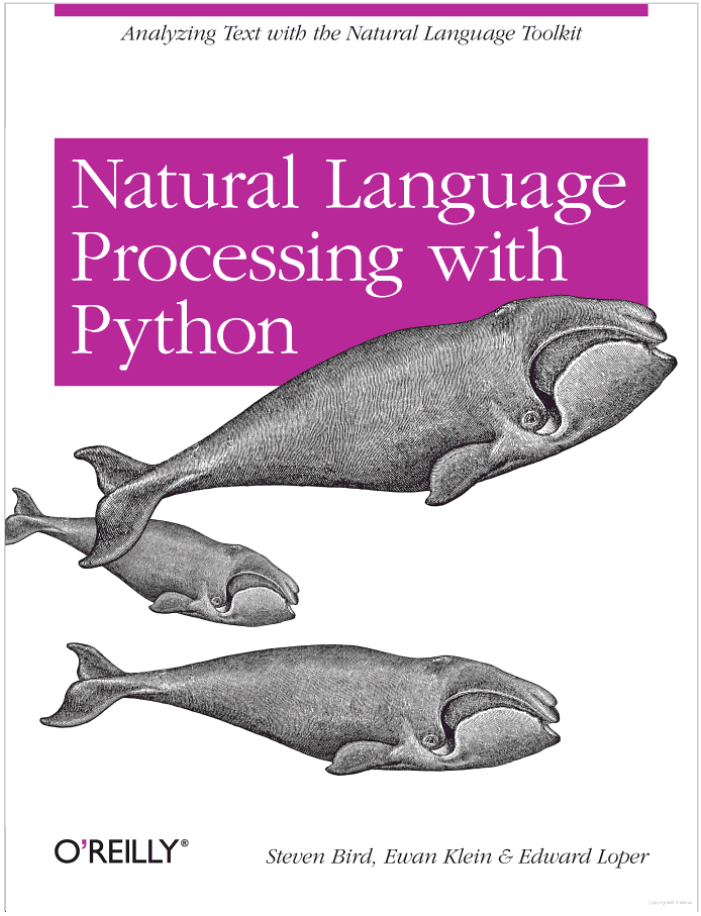
A Tool for Twitter Analysis

http://scraperwiki.com/dataset/flxnhta



Some Other Tools

- No coding is necessary
  - Import.io
  - iMacros
  - RapidMiner
  - SPSS Text Analytics
  - SAS Text Miner
  - Statistica Text Miner
  - VantagePoint
- Gold Standard (requires coding)
  - Phyton
  - R (tm package)



Web Mining

(Gök et al. 2014)

Webmining: Introduction

- Firms reveal a significant amount of information in their websites:
  - 63% of British firms, 73% British manufacturing firms
  - not only products and services but also capabilities, orientations, trends, strategies and relationships with other firms and organisations
- Historic websites are available at the Wayback Machine: http://archive.org/web/
- How useful is the website data in comparison with other data sources?
- How to retrieve, structure, clean, manipulate, conceptualise and interpret website data in understanding enterprise R&D activity

Webmining: Literature

- Three types of Web-mining:
  - Web structure analysis: (network) analysis of the hyper-linked structure of a set of webpages
    - role between the internet and innovation systems by utilising website-based indicators from webpage counts and links (Katz and Cothey, 2006)
    - website data in a future-oriented technology analysis to identify existing networks that are concerned with technological change (van de Lei and Cunningham, 2006)
    - landscape of online resources pertaining to emerging technologies: the top search terms and resulting top-ranked webpages from Google (Ladwig et al., 2010)
    - examining the relationships between, and prominence of, actors engaged in nanotechnology (Ackland et al., 2010)
  - Web usage analysis: data mining process involving the usage data of webpages
    - Webometrics movement (see Thelwall (2012) for an overview).

Webmining: Literature

- Web content analysis: analysis of unstructured text data within webpages to extract structured information
  - keyword occurrence in company websites from a cross-industry sample of SMEs, in order to identify commercialisation-focussed business models from highly-innovative firms (Libaers et al., 2010)
  - which organisations play a key role in the development of nanotechnology (Hyun Kim, 2012)
  - transition of nanotechnology from discovery to commercialisation (Youtie et al., 2012)
  - activities of SMEs in their pursuit of commercialising emerging technologies (Arora et al., 2013)
  - whether North American Industry Classification System code (NAICS) effectively shows the true industrial sectors of Fortune 500 firms by analysing their websites (Al-Hassan et al. 2013)
  - map geo-tagged geo-hazards, such as landslides, earthquakes and floods, by analysing online news (Battistini et al., 2013)
  - feasibility and desirability of the automated collection of official statistics, such as consumer price index, from websites (Hoekstra et al., 2012)
  - mining of political opinions from websites, forums and social media (Sobkowicz et al. 2012; Sobkowicz and Sobkowicz 2012)
  - content mining of website discussion forums to detect concern levels for HIV/AIDS (Sung et al. 2013)
  - mining social media to discover drug adverse effects (Yang et al. 2012)

WebMining: Questions

- How can the web content mining process be operationalised to study business R&D activities?
- How sensitive are the results to the web content mining procedure followed?
- How does the website-based R&D activity indicator compare with other R&D indicators?
- What are the relative advantages and disadvantages of website data over other data sources?

Data and Methodology

- 'Sustaining Growth for Innovative New Enterprises'
  - Focus: emerging green goods industries (firms whose outputs benefit the environment or conserve natural resources)
- Data Sources:
  - FAME time series database (derived from UK Companies House data) for financial information, including R&D expenditure
  - Technology Support Board (TSB) funded R&D projects database, which includes the extent to which identified firms receive support from the UK government for their R&D activities;
  - Publications and patents from the Scopus and Derwent databases;
  - Face-to-face interviews and;
  - Firm websites

Web Content Analysis Variables

- manufacturing strategy
  - Products
  - manufacturing intensity
  - customisation
  - greenness
- linkages
  - Universities
  - Partnerships
  - membership organisations
  - regional/extra-regional links
- investment strategy
  - venture capital
  - Investment
  - policy influence (regulation)

R&D Varaibles

- R&D expenditure data from the FAME database;
- R&D monetary support received from the UK government (TSB Database);
- Patents and publications from Scopus and Derwent databases;
- R&D activity data derived from firm websites

Coverage of Data

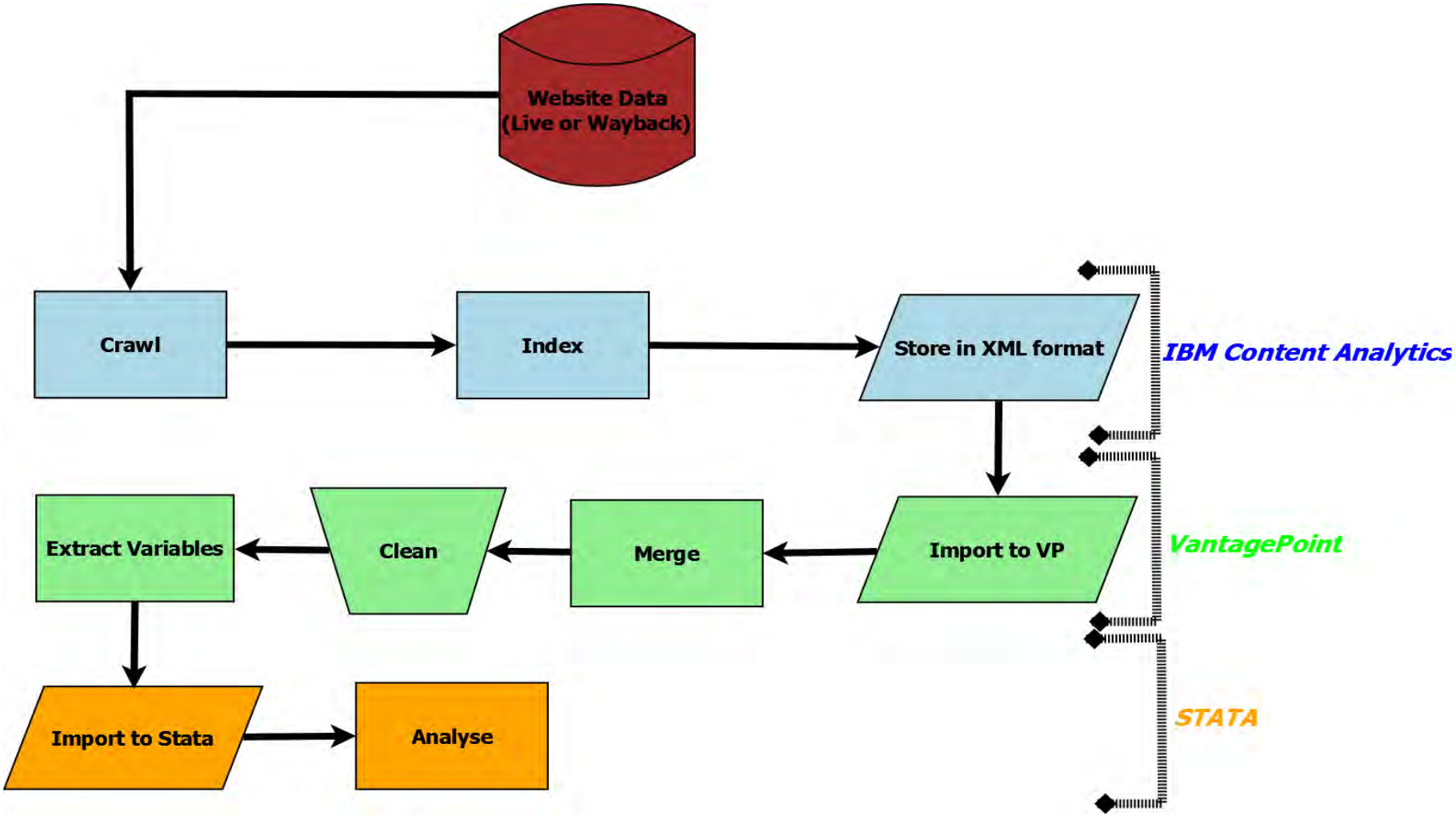
	Explanation	Number of Firms 1	Coverage of Firms	Number of Observations 2	Coverage of Observations
Patents	Total number of publications	43	14.53%	150	5.63%
Publications	Total number of patents	15	5.07%	33	1.24%
Grants	Amount of R&D Expenditure as reported in the FAME Database (GBP)	51	17.23%	91	3.42%
R&D Expenditure	Total number of grants from the TSB	66	22.30%	187	7.02%
rndweb6	Website-based variable, number of instances of keywords normalised by the number of noun phrases in websites	204	68.92%	909	34.12%

[1] Number of firms reporting a value for this variable at any year between 2004 and 2012.
[2] Number of non-missing observations

Data Included in the Web Content Analysis

Year	Number of Firms with websites	Number of Webpages	Number of Phrases
2004	125	14,919	1,880,173
2005	133	11,714	2,027,819
2006	131	15,790	1,965,263
2007	173	10,647	1,331,624
2008	173	12,786	1,232,770
2009	161	10,769	1,279,867
2010	163	12,991	1,617,074
2011	199	15,829	2,603,138
2012	237	51,683	10,320,032

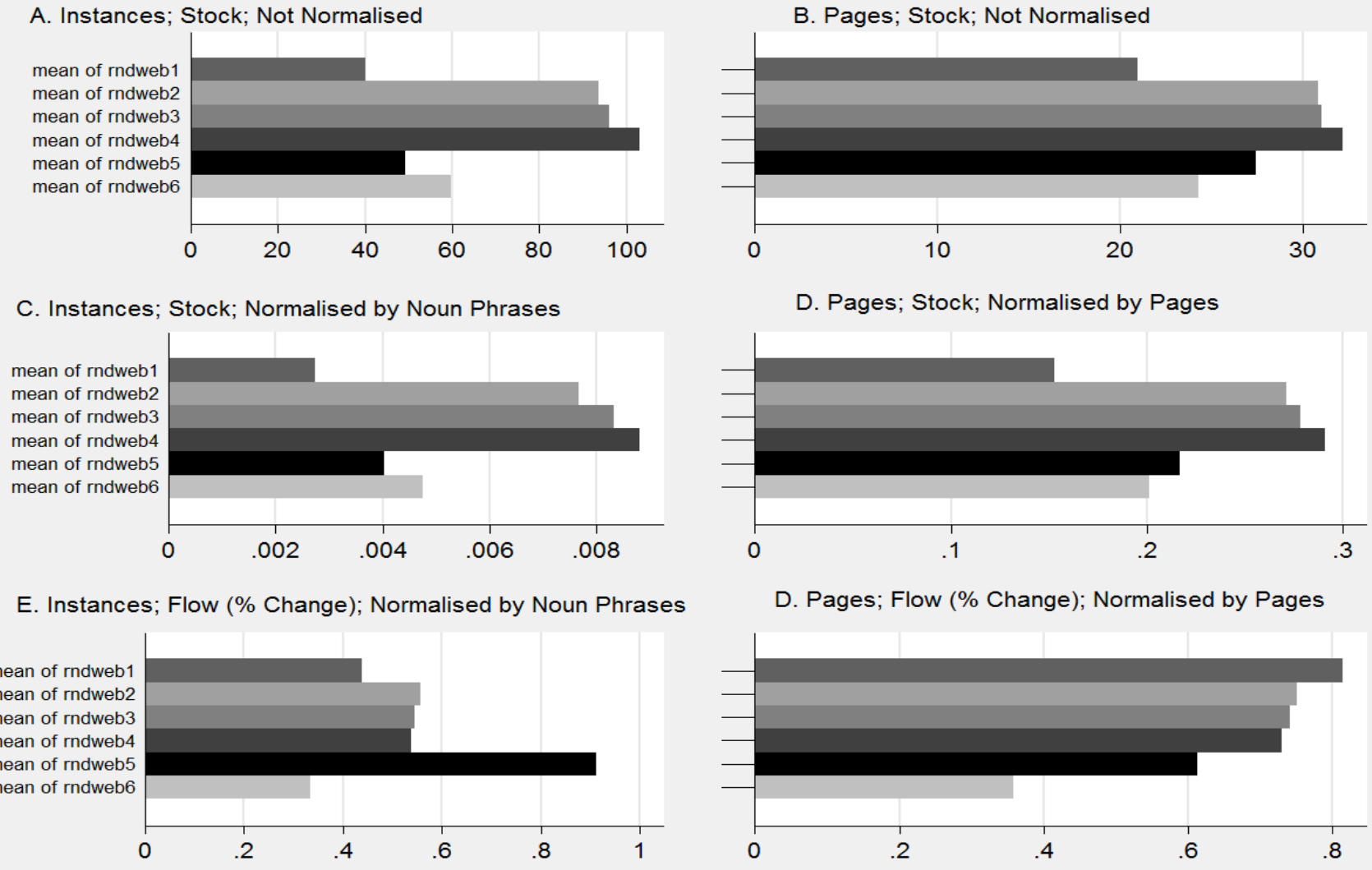
Web Content Analysis Process



Variables to Capture R&D Activity

Variable	Keywords	Difference from the previous set
rndweb1	research*	
rndweb2	research* AND development*	rndint1 + (development*)
rndweb3	research* AND development* AND R&D	rndint2 + (R&D)
rndweb4	research* AND development* AND R&D AND lab*, scientist*	rndint3 + (lab, laboratory, scientist)
rndweb5	research* AND (development NEARBY research) AND, R&D AND lab* AND scientist*	rndint4 – (development [NOT NEARBY] research)
rndweb6	(research and development) AND R&D AND lab* AND scientist* AND research AND researcher AND scientist* AND (product development*) AND (technology development*) AND (development phase) AND (technical development*) AND (development program*) AND (development process*) AND (development project*) AND (development cent*) AND (development facilit*) AND (technological development*) AND (development efforts) AND (development cycle) AND (development research) AND (research & development) AND (development activity)	rndint5 – (development [NEARBY] research) + (a set of development variants)

Different transformations and normalisations



Comparing Website Data with Other Data Sources

No	Variable	Explanation	1	2	3	4	5	6	7	8	9	10
1	Publications	Total number of publications	1									
2	Patents	Total number of patents		1								
3	R&D Expenditure	Amount of R&D Expenditure as reported in the FAME Database, in GBP		0.5431	1							
4	Number of Grants	Total number of grants from the TSB	0.9764	0.5367		1						
5	rndweb1	Website based variables, number of instances of keywords normalised by the number of noun phrases in websites					1					
6	rndweb2						0.8437	1				
7	rndweb3						0.8388	0.9974	1			
8	rndweb4						0.8265	0.9924	0.9957	1		
9	rndweb5						0.8317	0.8038	0.8230	0.8205	1	
10	rndweb6						0.8635	0.9108	0.9016	0.905	0.9551	1

Comparison of Different Data Sources

		R&D Expenditure	R&D Grants	Patents and Publications	Mention of R&D Activity in Websites
Source		Financial Database (FAME)	Government Database (TSB)	Web of Science	Current and historic websites
Indicator Type		Input	Input	Output	Process
Data Structure		Structured	Structured	Semi Structured	Unstructured
Data Quality Dimensions					
Completeness	Sufficient breadth and depth and scope				
Accuracy	Correct representation of the phenomenon				
Currency	How promptly data is updated				
Frequency	Data change frequency				
Consistency	Agreement among components				
Interpretability	Easiness of interpreting meaning				
Accessibility	Easiness of access and analysis				
Handling	Easiness of analysis				
Amount	Quantity of data				
Flexibility	Adaptable to different purposes				

[11] Adapted and extended from Batini and Scannapieco (2006)

Web Mining: Conclusions

- Website data is increasingly used and potentially very useful
- Potential advantages and disadvantages over other data sources
- The data retrieval, preparation, cleaning and analysis is cumbersome
- Once these steps are correctly performed, website data has significantly superior coverage
- Interpretation of the website data is important
- Different indicators show different facets of reality

Overall Conclusions

- Huge and unexploited potential for big (and unstructured data)
- Potential issues (Boyd and Crawford (2011)):
  - Big Data is not always more objective and accurate
    - Researchers still interpret data
    - Processing big data involves significant number of subjective choices
    - The bigger the data, the more possibility for error
  - Bigger Data are not always better data
    - Big data does not make most of the methodological issues disappear
    - Big data has additional methodological challenges
  - Big data is often proxy
  - There are still significant ethical challenges
  - Limited Access to Big Data Creates New Digital Divides

Overall Conclusions

- Big data changes the way we conduct research
- It also influences the questions we ask
- But also big data changes our understanding of the phenomenon we investigate
  - "Change the instruments, and you will change the entire social theory that goes with them" (Latour, 2009).
- Do numbers speak for themselves?
  - *"This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves."* (Anderson (2008) cited in Boyd and Crawford (2011))

Thank you!

Questions and comments:
abdullah.gok@manchester.ac.uk