



knowledge without boundaries

Text and Data Mining: what librarians need to know

EIFL-Licensing/EIFL-IP webinar, 6 February 2014

Ben White
Ben O'Steen
British Library

The image features a silhouette of an oil pumpjack against a purple and blue gradient sky. The pumpjack is positioned on the right side of the frame, with its long arm extending towards the center. The overall scene is dark and atmospheric, suggesting a sunset or sunrise.

DATA

is the new oil

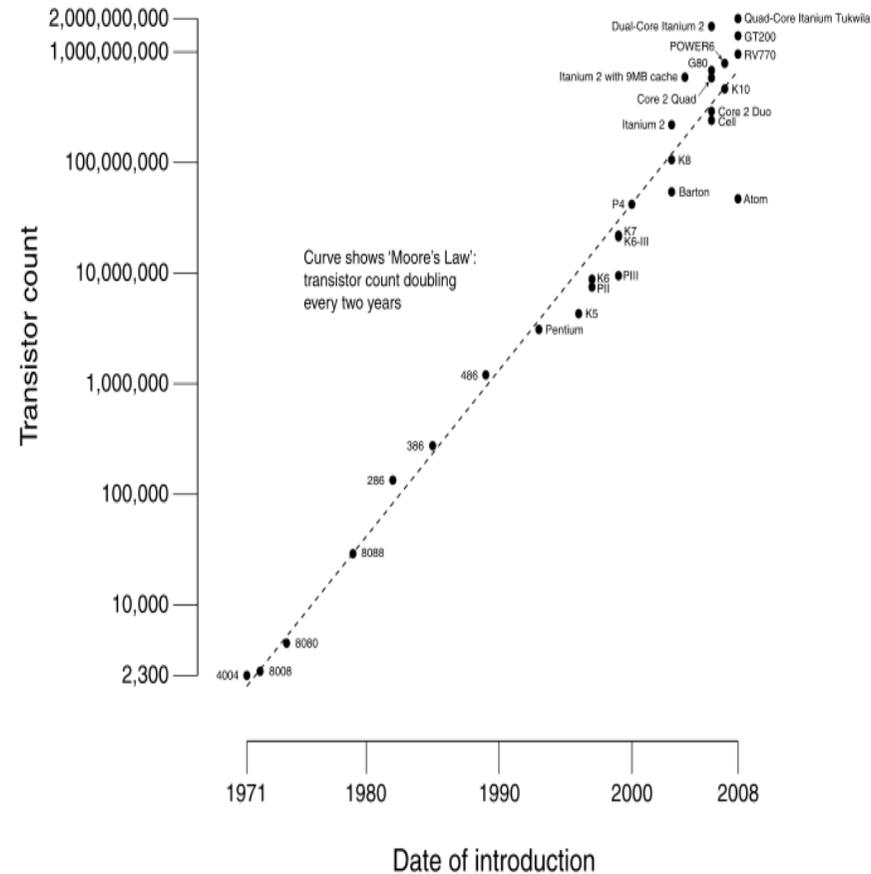
How Much Data is there?

2013

1.8 zetabytes?

And 80% is unstructured.

CPU Transistor Counts 1971-2008 & Moore's Law





Learning and Research

- For millennia learning has been based on people reading;
- Taking notes;
- Extracting facts and data; and
- Organising information.

Pre mid 1990s = pen, pencil and eyes

▪

Computers can now read



© Woodguy

www.bl.uk

And a lot faster than humans



How to Do Research in 2013?

Post mid 1990s = pen, pencil, eyes **AND** computers.

Are off the shelf text and data mining tools from software providers, but researchers write their own programmes too.

What is Text and Data Mining?

(**NOT** search by a search engine)

Algorithms are “intelligently” analysing and reading the text / data (using statistics, probabilities, computational linguistics etc) to do **amongst** other things:

- i) Make assumptions what text strings are about - (e.g. Is the “tree” a piece of wood, a family tree, the tree of life (biology)?);
- ii) Analyse what the entire text is about;
- iii) See if there is a +ve or –ve relationship between two pre-selected variables.

Text Mining Shakespeare

Data: Summary of association rules (Scene 1)*

Summary of association rules (Scene 1.sta)
 Min. support = 5.0%, Min. confidence = 5.0%, Min. correlation = 5.0%
 Max. size of body = 10, Max. size of head = 10

	Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
154	and, that	==>	like	6.94444	83.3333	91.28709
126	like	==>	and, that	6.94444	100.0000	91.28709
163	and, PAROLLES	==>	will	5.55556	80.0000	73.02967
148	will	==>	and, PAROLLES	5.55556	66.6667	73.02967
155	and, you	==>	your	5.55556	80.0000	67.61234
122	your	==>	and, virginity	5.55556	57.1429	67.61234
164	and, virginity	==>	your	5.55556	80.0000	67.61234
121	your	==>	and, you	5.55556	57.1429	67.61234
73	that	==>	like	6.94444	41.6667	64.54972
75	that	==>	and, like	6.94444	41.6667	64.54972
161	and, like	==>	that	6.94444	100.0000	64.54972

What is Text and Data Mining?

This allows for example people to:

- i) See if there is some kind of relationship between a chemical / enzyme etc and a medical disease;
- ii) Discover some previously undiscovered use for a drug or a chemical compound;
- iii) Allow organisations to organise electronic data by subject category etc.

TDM & Libraries

Libraries important as they provide access to scholarly information.

A lot of text and data on the web but also very valuable content in books and journals.

People want to hold the data locally and work on it using their own tools.

Text and Data Mining – Big Business

Video Time!

(hopefully)

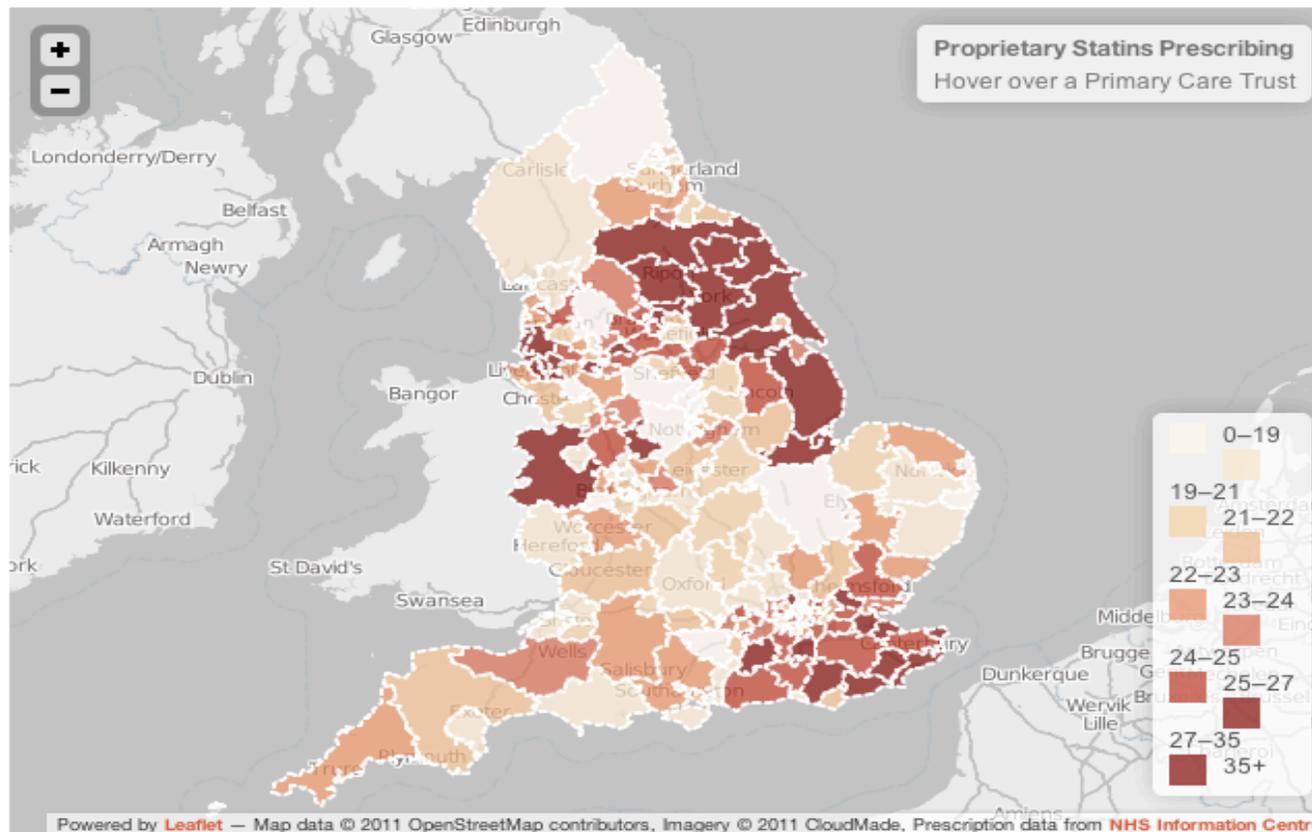
http://www.youtube.com/watch?v=2YQNNQ_GLe9Q

Savings in the Health Sector

NHS efficiency savings: the role of prescribing analytics

The NHS has been challenged to make £20 billion in “efficiency savings” by 2015 (1). £10 billion a year are spent by the NHS on essential drugs. Often, there’s a choice between a cheap “generic” medication, or an expensive “branded” one. Branded drugs can cost over ten times as much, for the same therapeutic benefit. “Prescribing Advisors” in the NHS, with the support of NICE, encourage doctors to use the most cost effective treatments. We have analysed exactly how much is spent on expensive “branded” medicines, for one class of drugs, namely statins, in England.

Percentage of proprietary statin prescribing by CCG Sep 2011 - May 2012



[Home](#)

[Aims & Objectives](#)

[NaCTeM Services](#)

[Text Mining Tools](#)

[Resources](#)

[Terms & Conditions](#)

[FAQ](#)

[News & Events](#)

[People](#)

[Projects](#)

[Publications](#)

[Community](#)

[External Collaboration](#)

[Vacancies](#)

[Teaching & Tutorials](#)

[Feedback](#)

[How to Find Us](#)

[Site Map](#)

Welcome to NaCTeM

The National Centre for Text Mining (NaCTeM) is the first publicly-funded text mining centre in the world. We provide text mining services in response to the requirements of the UK academic community. NaCTeM is operated by the University of Manchester.

On our website, you can find pointers to sources of information about text mining such as links to

- text mining services provided by NaCTeM
- software tools, both those developed by the NaCTeM team and by other text mining groups
- seminars, general events, conferences and workshops
- tutorials and demonstrations
- text mining publications

Let us know if you would like to include any of the above in our website.

What text mining can do for you

Text mining offers a solution to the challenge of 'data deluge', information overload and information overlook. For more information, please see:

- [NaCTeM Brochure](#),
- [Text Mining Briefing Paper](#),
- [National Centre for Text Mining: an introduction to tools for researchers](#),
- [Vision for the Future](#),
- [Mining Biomedical Literature](#).
- [Event extraction for systems biology by text mining the literature](#)
- [Supporting the education evidence portal via text mining](#)

NaCTeM has developed text mining services and service exemplars for the UK academic community. Our services are underpinned by a

Featured News

- [New paper and resources to support anatomical entity recognition at literature scale](#)
- [Keynote speech Pharma Documentation Ring special meeting in Bruges](#)
- [COLING 2014](#)
- [NaCTeM success at BioCreative IV](#)
- [Participation in Workshop on Text and Data Mining for Data Driven Innovation - Highlights available](#)
- [NaCTeM student selected to participate in Global Young Scientists Summit](#)

New Medical Discoveries

Text mining suggests new uses for thalidomide

Marc Weeber and colleagues used automated text mining tools to infer that the drug thalidomide could treat several diseases it had not been associated with before. Thalidomide was taken off the market 40 years ago, but is still the subject of research because it seems to benefit leprosy patients via their immune systems. Weeber and Grietje Molema, an immunologist, used text mining tools to search the literature for papers on thalidomide and then pick out those containing concepts related to immunology. One concept, concerning thalidomide's ability to inhibit Interleukin-12 (IL-12), a chemical involved in the launch of an immune response, struck Molema as particularly interesting. A second automated search for diseases that improve when the action of IL-12 is blocked revealed several not previously linked with thalidomide, including chronic hepatitis, myasthenia gravis and a type of gastritis.

'Type in thalidomide and you get 2–3000 hits. Type in disease and you get 40,000 hits. With automated text mining tools we only had to read 100–200 abstracts and 20 or 30 full papers. We've created hypotheses for others to follow up,' says Weeber.

Weeber et al. J Am Med Inform Assoc. 2003 10 252–259

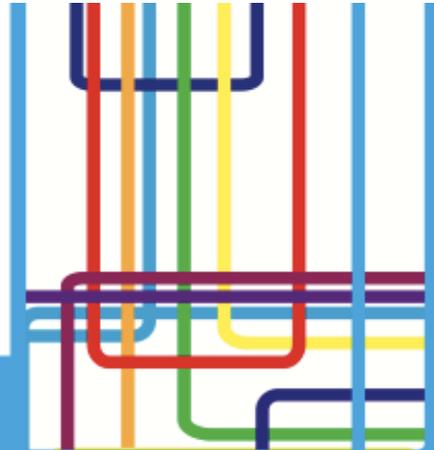
Reduces Reading Times Exponentially

JISC

Theme: Digital Infrastructure

The Value and Benefits of Text Mining

This is a **Digital Infrastructure Directions** report into the Value and Benefits of Text Mining to UK Further and Higher Education.



Not Just Computer Scientists Either



© South Wiltshire Girls School

The Right to Read is the Right to Mine?

- Facts and data not subject to copyright and database rights
- But computers have to copy in order to mine the data – so is it a licensable activity? (EU has an “internet browser” exception as browsers cache ...)
- European Union Commission stakeholder dialogue on TDM / “Licences for Europe” – Research / Library, Technology Sector and Open Access Publishers boycotted.



The Right to Read is the Right to Mine?

- How would you license the internet?
- UKPMC – 75 publishers had articles with the word “malaria” in the title. BL’s estimate that from experience of negotiating a new licence it takes 16 months on average.
- TDM goes across thousands / tens of thousands of articles which you **ALREADY** have legal access to. How can you **renegotiate** this with all publishers concerned?
- UK universities experiencing server access being suspended automatically when abnormal access is being detected.



Thank you



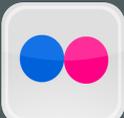
(unless indicated otherwise)

***Now it's
question time!***

Further information

- Find out more about the EIFL-Licensing programme
 - www.eifl.net/licensing
- Find out more about the EIFL-IP programme
 - www.eifl.net/copyright

Stay connected

- Visit our website - www.EIFL.net
- Subscribe to our newsletter - www.EIFL.net/subscribe
- Join email lists for EIFL programmes
-  facebook.com/EIFLnet
-  twitter.com/EIFLnet
-  www.flickr.com/photos/EIFL