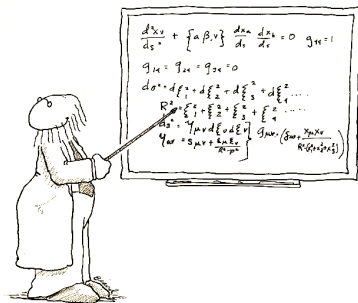


Doing Bayesian Text Analysis

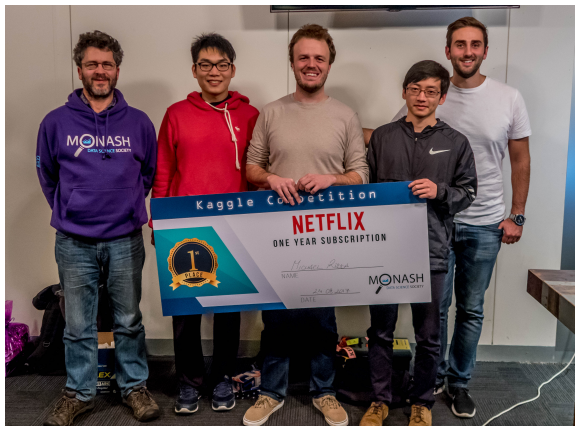
Wray Buntine
Monash University



<http://topicmodels.org>

August 26, 2017

Recent MDSS Kaggle



- ▶ [search "Facebook Monash MDSS"](#)
- ▶ [Youtube on Kaggle competition](#) (link from Facebook, finished 24/08)
- ▶ [Quantify - The University Datathon](#) (starting 26/09)

Outline



Career

Research

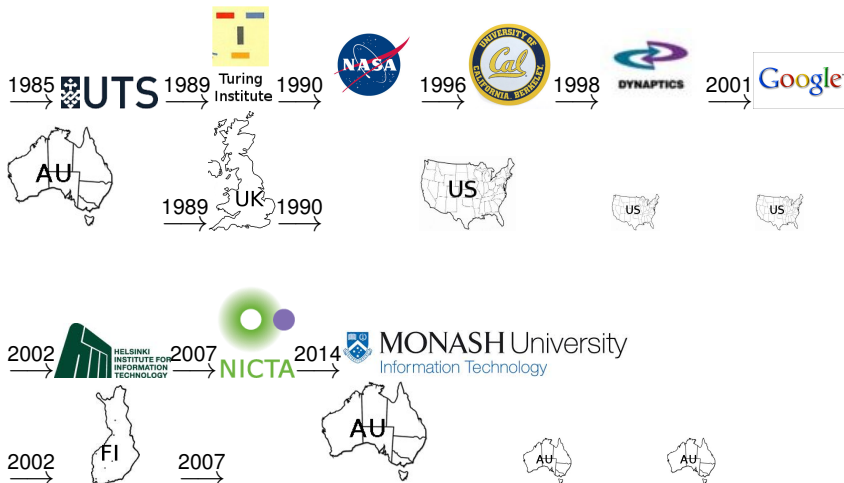
Motivating Text and Documents

In The News!

Working with Text

Recently

My Career



Creating a Web Profile

Blog on Wordpress: <https://topicmodels.org/>

Google Scholar: [Wray Buntine](#)

Quora: <https://www.quora.com/profile/Wray-Buntine>

GitHub: <https://github.com/wbuntine>

Know Your Strengths (and weaknesses)

- ▶ theory
- ▶ speaking
- ▶ management, networking
- ▶ business
- ▶ development
- ▶ teaching and supervision:
 - ▶ undergraduate
 - ▶ graduate
- ▶ entrepreneur

aim for the best, plan for the worst!

Long Term Planning

- ▶ R&D
- ▶ family and lifestyle
- ▶ health
- ▶ being multi-disciplinary
- ▶ develop application areas
- ▶ coding as a springboard
- ▶ dig in and build a team and develop connections
- ▶ be a hotshot in a mid-tier location or a struggler in a top-tier location?

Outline



Career

Research

Bayesian Model Averaging
Bayesian Non-parametrics
Other Issues

Motivating Text and Documents

In The News!

Working with Text

Recently

Background: Research Areas

- 1987-1990 inductive logic programming
- 1988-1993 Bayesian decision trees (inspired Breiman's "bagging")
- 1990-1995 (Bayesian) model averaging, BMA (major development in statistics)
- 1993-???? graphical models for machine learning
- 2001-???? machine learning in text, documents and search
- 2003-???? topic models
- 2010-???? nonparametric Bayesian methods
- 2014-???? machine learning for networks and semi-structured data

Outline



Career

Research

Bayesian Model Averaging

Bayesian Non-parametrics

Other Issues

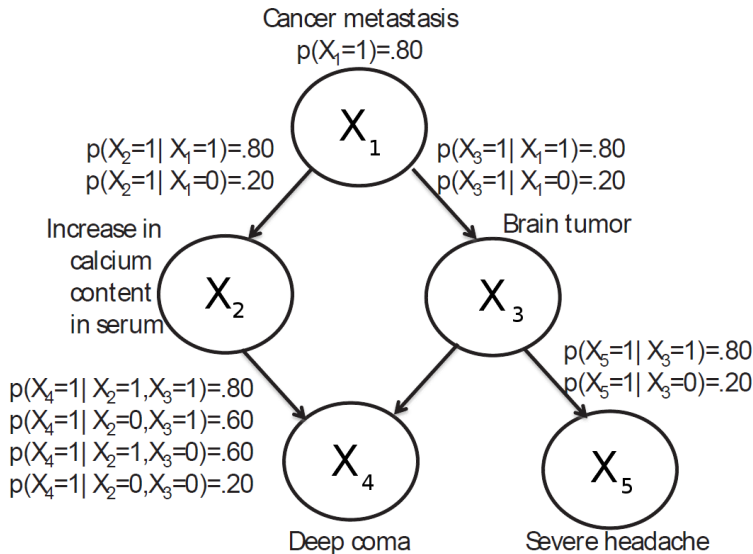
Motivating Text and Documents

In The News!

Working with Text

Recently




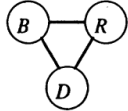
Bayesian Networks



Bayesian Model Averaging for Bayesian Networks

Summaries of the posterior distributions of N for the spina bifida data for all models with posterior probability greater than 0.01.

\hat{N} is a Bayes estimate, minimizing a relative squared error loss function

Model	Posterior	\hat{N}	2.5%, 97.5%
	Prob.		
	0.373	731	(701, 767)
	0.301	756	(714, 811)
	0.281	712	(681, 751)
	0.036	697	(628, 934)
Model Averaging	—	731	(682, 797)

From Madigan and York, 1995

Bayesian Model Averaging for Bayesian Networks, cont

Build a pool of “good” models (*i.e.*, the graphical structures) based on the training data: $\text{Good-Models}(\mathcal{X})$.

BMA estimate of probability for new data \vec{x} given training sample \mathcal{X} is

$$p(\vec{x}|\mathcal{X}) \approx \sum_{M \in \text{Good-Models}(\mathcal{X})} \frac{p(M|\mathcal{X})}{\sum_{M \in \text{Good-Models}(\mathcal{X})} p(M|\mathcal{X})} p(\vec{x}|M, \mathcal{X})$$

Question: how do we build “good” models given there are a combinatoric number?

Bayesian Model Averaging Storyline

- 1988: Sydney masters project students [Suk Wah Kwok](#) and [Chris Carter](#) help me do “averaging” of decision trees.
- 1990: [Graham Williams](#) PhD thesis, [ensembles](#) for decision trees.
- 1990: My PhD thesis, [BMA](#) for decision trees.
- 1990: York and [David Madigan](#) develop BMA for Bayesian networks.
- 1991: Ken Church (AT&T, 1990) tells me BMA cannot apply to n-grams!
- 1994: [Leo Breiman](#) developed [bagging](#) (or random forests, tree ensembles) for trees as a Frequentist response:
 - still one of the top performing classification algorithms
- 1995: Willems, Shtarkov, Tjalkens adapt BMA for n-grams, [context tree weighting \(CTW\)](#) for lossless compression.

Outline



Career

Research

Bayesian Model Averaging

Bayesian Non-parametrics

Other Issues

Motivating Text and Documents

In The News!

Working with Text

Recently

What are Parametric Methods?

By statistics: modelling with a fixed number of parameters,
e.g., linear regression

Common extension: add a dimension parameter and use
model selection ^{\mathcal{W}} to estimate,
e.g., linear regression with polynomials of order K
(unknown)

What are Non-parametric Methods?

By classical statistics: modelling **without** parameters,
e.g., nearest neighbour methods

What are Non-parametric Methods?

By classical statistics: modelling **without** parameters,
e.g., nearest neighbour methods

By Bayesian statistics: modelling **with an infinite number of**
parameters.

What are Non-parametric Methods?

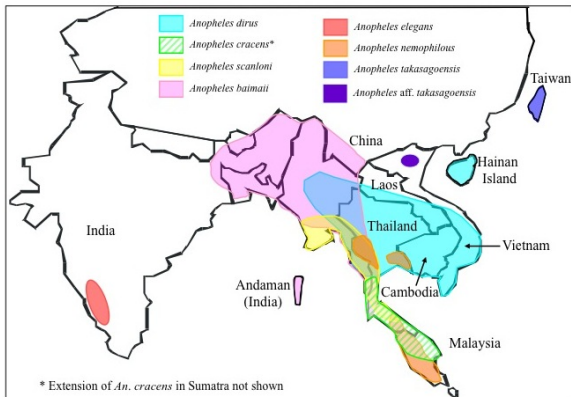
By classical statistics: modelling **without** parameters,
e.g., nearest neighbour methods

By Bayesian statistics: modelling **with an infinite number of**
parameters.

More accurately: modelling:

- ▶ **with a finite but variable number of**
parameters;
- ▶ more parameters are “unfurled” as needed.

How Many Species of Mosquitoes are There?



e.g. Given some measurement points about mosquitoes in Asia, how many species are there?

K=4? K=5? K=6 K=8?

How Many Words in the English Language are There?

... lastly, she pictured to herself how this same little sister of hers would, in the after-time, be herself a grown woman; and how she would keep, through all her riper years, the simple and loving heart of her childhood: and how she would gather about her other little children, and make their eyes bright and eager with many a strange tale, perhaps even with the dream of wonderland of long ago: ...

e.g. Given 10 gigabytes of English text, how many words are there in the English language?

$K=1,235,791?$ $K=1,719,765?$ $K=2,983,548?$

What is an Unknown Dimension?

- ▶ Some dimensions are fixed but unknown:
 - ▶ this uses parametric statistics
 - ▶ this is not Bayesian non-parametrics
- ▶ Some dimensions keep on growing as we get more data:
 - ▶ this is Bayesian non-parametrics

Modelling – What We Need

- ▶ matrices, tensors, graphs
- ▶ semi-structured data
- ▶ preferences, ratings, connections
- ▶ bioinformatics, social networks, bibliographic data
- ▶ ever expanding numbers of items/dimensions/nodes,
not small but unknown numbers

Bayesian Inference – General Methodology

Bayesian inference ^{\mathcal{W}} is particularly suited for intelligent systems in the context of the previous requirements:

- ▶ Bayesian model combination ^{\mathcal{W}} and Bayes factors ^{\mathcal{W}} for model selection ^{\mathcal{W}} can be used;
- ▶ marginal likelihood ^{\mathcal{W}} , a.k.a. the evidence for efficient estimation;
- ▶ collapsed Gibbs samplers ^{\mathcal{W}} , a.k.a. Rao-Blackwellised samplers, for Monte-Carlo Markov chain ^{\mathcal{W}} (MCMC) estimation;
- ▶ also blocked Gibbs samplers ^{\mathcal{W}} .

Bayesian Inference – General Methodology

Bayesian inference ^{\mathcal{W}} is particularly suited for intelligent systems in the context of the previous requirements:

- ▶ Bayesian model combination ^{\mathcal{W}} and Bayes factors ^{\mathcal{W}} for model selection ^{\mathcal{W}} can be used;
- ▶ marginal likelihood ^{\mathcal{W}} , a.k.a. the evidence for efficient estimation;
- ▶ collapsed Gibbs samplers ^{\mathcal{W}} , a.k.a. Rao-Blackwellised samplers, for Monte-Carlo Markov chain ^{\mathcal{W}} (MCMC) estimation;
- ▶ also blocked Gibbs samplers ^{\mathcal{W}} .

Wikipedia coverage of Bayesian non-parametrics is patchy.

Outline



Career

Research

Bayesian Model Averaging

Bayesian Non-parametrics

Other Issues

Motivating Text and Documents

In The News!

Working with Text

Recently

I'm not bad, I was just drawn that way



Jessica Rabbit in "Who Shot Roger Rabbit?"

I'm not complex, I was just written that way

For any topological space T , $\mathcal{B}(T)$ will denote the Borel σ -field of subsets of T . Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space and \mathbb{X} be complete, separable and endowed with a metric $d_{\mathbb{X}}$. Define on $(\Omega, \mathcal{F}, \mathbb{P})$ a Poisson random measure \tilde{N} on $\mathbb{S} = \mathbb{R}^+ \times \mathbb{X}$ with intensity measure ν . This means that

- (i) for any C in $\mathcal{B}(\mathbb{S})$ such that $\nu(C) = \mathbb{E}[\tilde{N}(C)] < \infty$, the probability distribution of the random variable $\tilde{N}(C)$ is Poisson($\nu(C)$);
- (ii) for any finite collection of pairwise disjoint sets, A_1, \dots, A_k , in $\mathcal{B}(\mathbb{S})$, the random variables $\tilde{N}(A_1), \dots, \tilde{N}(A_k)$ are mutually independent.

Moreover, the measure ν must satisfy the following conditions,

$$\int_{(0,1)} s \nu(ds, \mathbb{X}) < \infty, \quad \nu([1, \infty) \times \mathbb{X}) < \infty.$$

We refer to Daley & Vere-Jones (1988) for an exhaustive account on Poisson random measures.

From James, Lijoi & Prünster, 2009,

[“Posterior analysis for normalized random measures with independent](#)

Disclaimer: excellent paper, and the complexity is a necessary part of the precision required for mathematical analysis ... but its not needed in machine learning!

Outline



Career

Research

Motivating Text and Documents

In The News!

Working with Text

Recently

The Internet Society



Consumer information; targeted advertising; tracking blogs, tweets and search terms.

Document analysis and text mining has taken on a new life. Business, government and consumer ramifications still unfolding.

Social Media

- ▶ big data, rich text
- ▶ opinions, rants, reviews, gossip
- ▶ informal and formal language
- ▶ links and metadata
- ▶ linked data
- ▶ context and structure

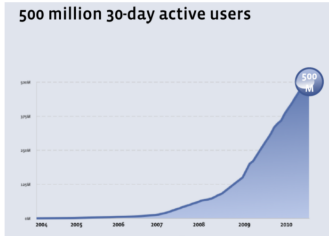


Linked Open Data (Semantic Web)



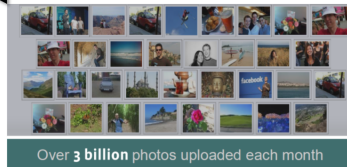
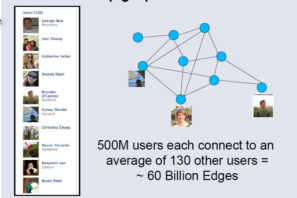
- ▶ Data stored in RDF (triples) and XML interlinked via URIs.
- ▶ Information Extraction is integrated (e.g., OpenCalais).
- ▶ Good structured content for semi-supervised training.

Facebook



facebook
Graphics from Lars Backstrom, ESWC 2012

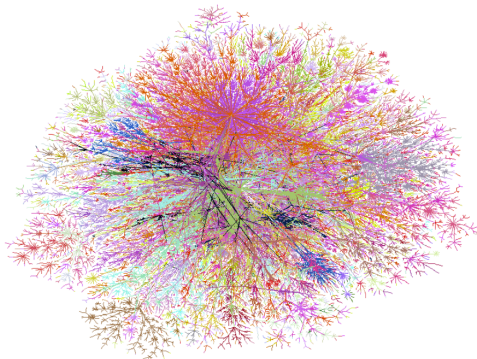
The Friendship graph



(from Smyth, "Probabilistic Models" invited talk at ECML-PKDD 2012)

Web Data

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, Y!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquared)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YIM, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



>10B useful webpages

(from Ahmed and Smola, “Graphical Models” tutorial at NIPS 2011)

Personal/Corporate Data

browse history

03/Aug/2009:21:51:12	[94.8.75.149]	/blog/2009-04-14-molecules-and-non-steam-engines/	http://www.enharden.com/blog/2009-04-14-molecules-and-non-steam-engines/
03/Aug/2009:21:46:00	[97.104.84.201]	/blog/category/aba/	http://www.enharden.com/
03/Aug/2009:21:45:55	[97.104.84.201]	/	-
03/Aug/2009:21:42:58	[97.104.84.201]	/blog/wp-blog.php?redirect_to=http://www.enharden.com/blog/2009-04-14-molecules-and-non-steam-engines/	-
03/Aug/2009:21:39:58	[66.249.68.12]	/html/gang2.asp	-
03/Aug/2009:21:39:40	[97.104.84.201]	/html/gang2.asp	-

tweets



Wray Buntine @wraylb

29 Nov 12

On #education, see OECD's Programme for Internl. Student Assessment, bit.ly/R4zr4T, AU and NZ do OK. Sourced huff.to/SbJlzX

[Hide summary](#)

Reply Retweets Favorites More

email

From: <it@nicta.com.au>
To: <Wray.Buntine@nicta.com.au>
Date: Mon, 6 Jan 2014 14:51:05 +1100
Subject: Optus Mobile Usage Report for Wray Buntine (BILLING PERIOD: December 2013)
Content-Type: text/html; charset="utf-8"
Content-Transfer-Encoding: base64
Message-ID: <da811487-ea6b-40e3-b2b1-c39ce3a36dda@ATP-EXCHCAS1.in.nicta.com.au>
Return-Path: it@nicta.com.au
X-MS-Exchange-Organization-AuthSource: atp-exchcas1.in.nicta.com.au
MIME-Version: 1.0

PGIldGEgaHR0cCl1cXVpdj0iQ29udG9udGVudC1UeXB1IiBjb250Zk50PSJ0ZDh0L2h0bWw7IGNoYXJzZXQ9dXRmLlTgipjxkaYgc3R5bGU9ImZvbnQtZmFtaXZlbnR1bWVsb3c3Y5bSaxN0IG9mIE5JQ1RBEI9wcmF5IEt1bRpbmUsIDxicj48YnI+UGxlYXNlIGZpbnQgYmVsb3c3Y5bSaxN0IG9mIE5JQ1RBEI9w

letters

Prof Zoubin Ghahramani
Computational and Biological Learning Lab
Department of Engineering, University of Cambridge



25th December, 2012

Dear Prof. Ghahramani

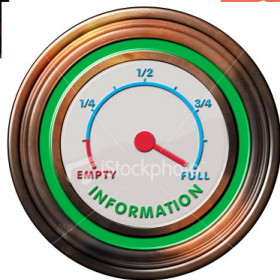
I have been Michael Jordan's supervisor for the last three years. In that time Michael has made great progress towards his PhD and obtained impressive results in Bayesian machine learning, jointly publishing dozens of journal papers each year.. While his masters as in psychology, his strong statistical and mathematical background meant he could quickly understand the material required. In fact, his first paper towards

The Data Society

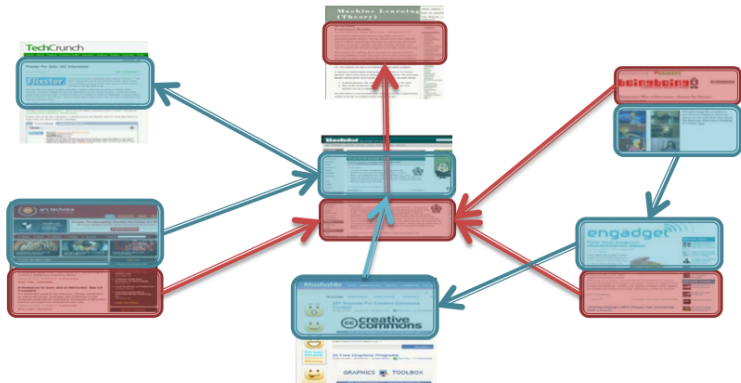
- ▶ **Your life is recorded**: phone calls, mobile, internet access, email, GPS, financial, health, and coming soon RFID tags, DNA and pharmaceuticals.
- ▶ The “web of things” is becoming a reality, following [Linked Open Data](#) and [the Internet](#).
- ▶ The British Academy wonders about *effects of the rapid rise of the data society* (Dec 2010).
- ▶ **Data Science** is here.

Data (and text) is now viewed as a key resource for business, its collection and analysis valued.

Information Overload



Information Flow



Information propagates.

- ▶ **Lifecycle of a meme** tracks its progress through tweets, blogs and articles.
- ▶ **Lifecycle of an article** tracks its change over time as other information integrated.

Information Warfare

Definition: "the use and management of information in pursuit of a competitive advantage over an opponent."

- ▶ Email spam, link spam, etc.
 - ▶ Whole websites are fabricated with fake content to trick search engines.
 - ▶ Spammers using social networks [to personalise attacks](#)
- ▶ BBC reports trust in information on the web is being damaged (Dec. 2011)
["by the huge numbers of people paid by companies to post comments"](#)
- ▶ People think "news organizations tend to favor one side," ... and "are often influenced by powerful people and organizations" says [Pew Research Center](#) (Sept. 2011)

It's an information war out there on the internet (between consumers, *i.e.*, you, companies, not-for-profits, voters, parties, news publishers, ...).

Language Technologies and ...

Language Technologies and Machine Learning are the front-line technologies for

- ▶ Information Access,
- ▶ analysis of Information Flow, and
- ▶ Information Warfare

on the Internet, Social Media and Linked Open Data.

We're the rocket scientists of the 2010's.

Outline



Career

Research

Motivating Text and Documents

In The News!

Working with Text

Recently

IBM/Watson on Jeopardy! in 2011

IBM's Watson supercomputer destroys all humans in Jeopardy! practice round (video!)

by Paul Miller | @futurepaul | January 13th 2011 at 4:43 pm



0



So, in February, IBM's Watson will be in an official Jeopardy! tournament this competition



FEATURED STORIES



(from Engadget.com 2011)

Time on Text Mining in 2012

TIME

Subscribe



Redeem

1 Oct - 15 Oct '14

Travel

1 Oct - 15 Dec '14

REDEEM YOUR FLIGHT NOW

Enrich

Terms and co

TECHNOLOGY & MEDIA

Why Text Mining May Be The Next Big Thing

By Gary Belsky @garybelsky | March 20, 2012 | 2 Comments

f Share

f Like 51

Tweet 128

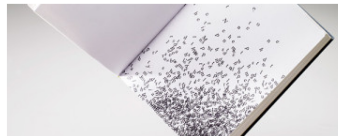
g+1 9

in Share 44

Pin It

Read Later

"Big Data" is a hot topic in the business world these days. But there's a subset of this broad field that has yet to take a turn in the spotlight. It's called "text mining," and you're probably going to be hearing a lot more about it over the coming months and years. Basically, text mining is the process of combing through



NYT and Deep Neural Networks in 2012

The New York Times

Business Day
Technology

WORLD

U.S.

N.Y. / REGION

BUSINESS

TECHNOLOGY

SCIENCE

HEALTH

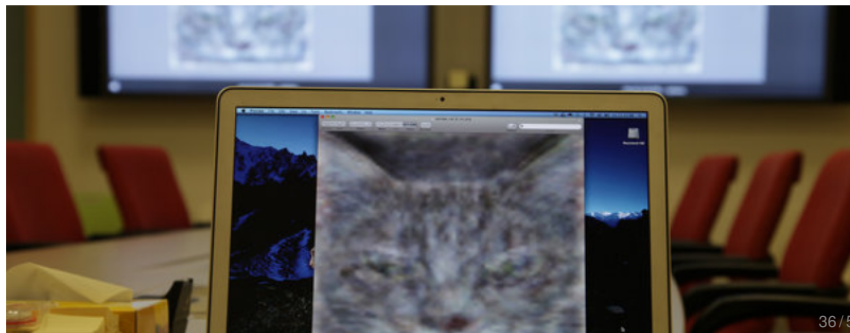
SPORTS

OPINION



Rolex-made in Switzerland

How Many Computers to Identify a Cat? 16,000



IBM and Watson in 2014

BUSINESSDAY

[HOME](#)[WORLD](#)[MARKETS](#)[COMPANIES](#)[ECONOMY](#)[INTELLIGENCE](#)[PERSONAL FINANCE](#)[Training](#)[Special Reports](#)[Contact Us](#)[About Us](#)[Stay Connected](#)

How Watson Changed IBM

📅 September 8, 2014 | 📁 Filed under: Monday Morning | 👤 Author: Editor



1



14



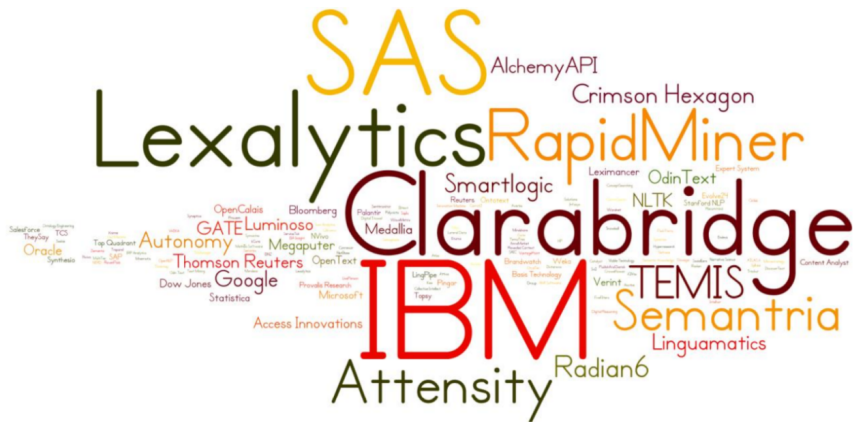
1

When IBM's "Watson" computer competed on "Jeopardy" and won, it demonstrated that machines could understand and interact in a natural-language, question-and-answer format and learn from their mistakes. That meant that machines could deal with the non-numeric information that is getting hard for humans to keep track of, including the knowledge coming out of human genome research and medical information in patient records.

So IBM asked how the fullest potential of this breakthrough could be realized and how the company could create and capture a significant portion of that value.

IBM decided to release Watson to the world as a platform, to run experiments with a variety of organizations to

Text Analytics Vendor Space in 2014



(from Grimes, "Text Analytics 2014" market study by Alta Plana)

Outline



Career

Research

Motivating Text and Documents

In The News!

Working with Text

Recently

What Can Companies Do with Text?

- ▶ customer interaction and analysis (web feedback, call centres)
- ▶ qualitative analysis of surveys (customers, staff, web)
- ▶ expertise finding and related employee knowledge management (email, intranet)
- ▶ website management (intranet)
- ▶ search enhancement
- ▶ legal search and discovery (client docs)
- ▶ financial (qualitative), policy and political analysis (external docs)
- ▶ business intelligence (external docs)
- ▶ record matching (names and addresses in databases)

What Can Companies Do with Text? *cont.*

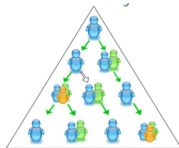
- Domains:
- ▶ healthcare and medicine
 - ▶ public administration, government
 - ▶ defence/security
 - ▶ consumer-facing business
 - ▶ research and intelligence
- Usage:
- ▶ own corporate data versus acquired data
 - ▶ in-house processing versus outsourced
- Tasks:
- ▶ generating taxonomies
 - ▶ classifying/tagging documents
 - ▶ information extraction
 - ▶ use of dictionaries and ontologies
 - ▶ sentiment extraction

Using Information



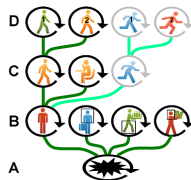
.....real-time citizen journalism

social media marketing

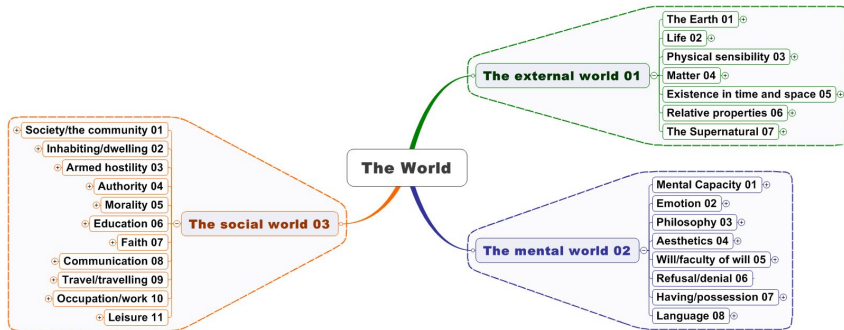


..... reputation management

behaviour analysis



Text Analysis Tasks: Taxonomies



Tasks:

- ▶ **creating** a suitable taxonomy for your collection
- ▶ **classifying** new documents into an existing taxonomy

Text Analysis Tasks: Named-Entities and Key Phrases

RT @user : the **iPhone**_{product} is so awesome!!! Emailing, texting, surfing the sametime! — Can do all tgat while talkin on the phone?...

It would appear that the **iPhone**_{product}, due to construction, is weak at holding signal. Combine that with a bullshit **3G network**_{phrase} in **Denver**_{city}.

Text Analysis Tasks: Sentiment

Positive	Negative
RT @user : the iPhone is so awesome !!! Emailing, texting, surfing the sametime! — Can do all tgat while talkin on the phone?...	@user awww thx! I can't send an email right now bc my iPhone is stupid with sending emails. Lol but I can tweet or dm u?
Ahhh! Tweeting on my gorgeous iPhone! I missed you! hehe am on my way home, put the kettle on will you pls :)	It would appear that the iPhone, due to construction, is weak at holding signal. Combine that with a bullshit 3G network in Denver.

Text Analysis: State of the Art

- ▶ deep neural networks has revolutionised text analysis and is now state of the art for most tasks.
- ▶ the book [*Deep Learning*](#) by Goodfellow, Bengio and Courville, 2017 covers methods (not text analysis)
- ▶ deep neural networks doesn't cover "smaller data" tasks and prior knowledge as well
 - ▶ traditional statistical machine learning still has a lot of work!

Outline



Career

Research

Motivating Text and Documents

In The News!

Working with Text

Recently

My Research: Discrete Bayesian Non-Parametrics

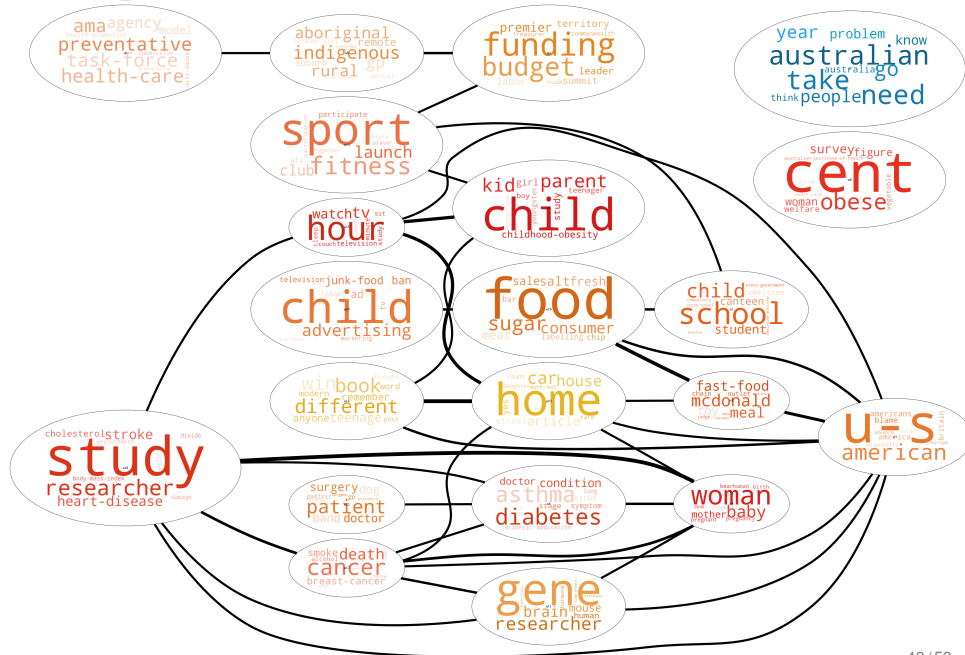
- ▶ using Discrete Bayesian Non-Parametrics
 - ▶ Pitman-Yor processes
 - ▶ Poisson process techniques
- ▶ on complex discrete structures and networks
- ▶ for data like
 - ▶ knowledge graphs
 - ▶ WordNet, parse trees
 - ▶ citation networks with abstracts
- ▶ yielding state-of-the-art software for several problems

Applied to Machine Learning by folks like Lancelot James (HKUST) and Yee Whye Teh (Oxford).

My Research: Analysing News Articles on Obesity

- ▶ Use 900 articles on the ABC website containing the keyword “obesity.”
- ▶ Analyse with non-parametric topic model to “summarise” main topics.
- ▶ Produce graphic.

Analysing News Articles on Obesity, cont.

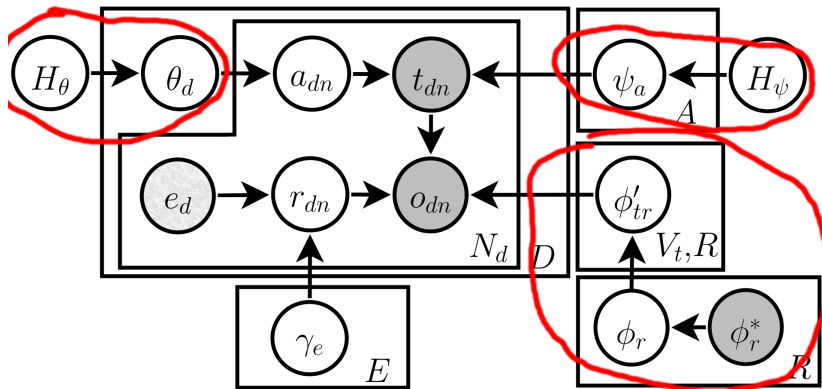


My Research: Clustering Tweets

- ▶ Use 1 million Tweet collection from ANU,
- ▶ incorporate hashtags and author networks into model,
- ▶ hierarchical information sharing from combination.

Clustering Tweets using Author Networks and Hashtags

hierarchical model integrating hashtags (y), word content (w), author network (x)



Clustering Tweets, Results

Table 7: Topical analysis on the T6 dataset with the TNTM, which displays the top three hashtags and the top n words on six topics. Instead of manually assigning a topic label to the topics, we find that the top hashtags can serve as the topic labels.

Topic	Top Hashtags	Top Words
Topic 1	finance, money, economy	finance, money, bank, marketwatch, stocks, china, group, shares, sales
Topic 2	politics, iranelection, tcot	politics, iran, iranelection, tcot, tlot, topprog, obama, musiceanewsfeed
Topic 3	music, folk, pop	music, folk, monster, head, pop, free, indie, album, gratuit, dernier
Topic 4	sports, women, asheville	sports, women, football, win, game, top, world, asheville, vols, team
Topic 5	tech, news, jobs	tech, news, jquery, jobs, hiring, gizmos, google, reuters
Topic 6	science, news, biology	science, news, source, study, scientists, cancer, researchers, brain, biology, health

Writing: Probabilistic Modelling for Data Science

- ▶ On sabbatical, but giving lectures currently at Clayton.
- ▶ Lecture slides can be found on Google drive, search [*“Wray’s Slides”*](#)
- ▶ Introduction to statistical reasoning for machine learning.
 - ▶ causality, independence, decision theory, information
 - ▶ exponential family distributions and inference
 - ▶ basic statistic processes with additive distributions
 - ▶ Monte Carlo Markov chain sampling
 - ▶ probability networks, inference, collapsing and augmenting distributions

Questions?

ngiyabonga
teşekkür ederim
dank je
danke
謝謝
thank you
gracias
moichakkeram
tapadh leat
go raibh maith agat
dakujem
merci
arigato
sukriya
kop khun krap
grazie
merci
terima kasih
감사합니다
obrigado
dziękuję
hvala
bedankt
спасибо



Questions?