

# **The analytics landscape: A personal view**

**Charles Elkan**

**December 20, 2011**



# What is analytics?

- Big data, business intelligence (BI), decision support (DSS), data warehousing, unstructured data, knowledge discovery in databases (KDD), information visualization, map-reduce.
- analytics = convert data into intelligence + capture value  
= statistics + optimization
- statistics = machine learning = data mining
- optimization = microeconomics + operations research

# Outline

1. Structured data (predictive, visual)
2. Unstructured data
3. The business of analytics
4. A research *and* business opportunity

# A basic distinction

## I. Structured data

Tables in databases  
Nodes and links in  
networks

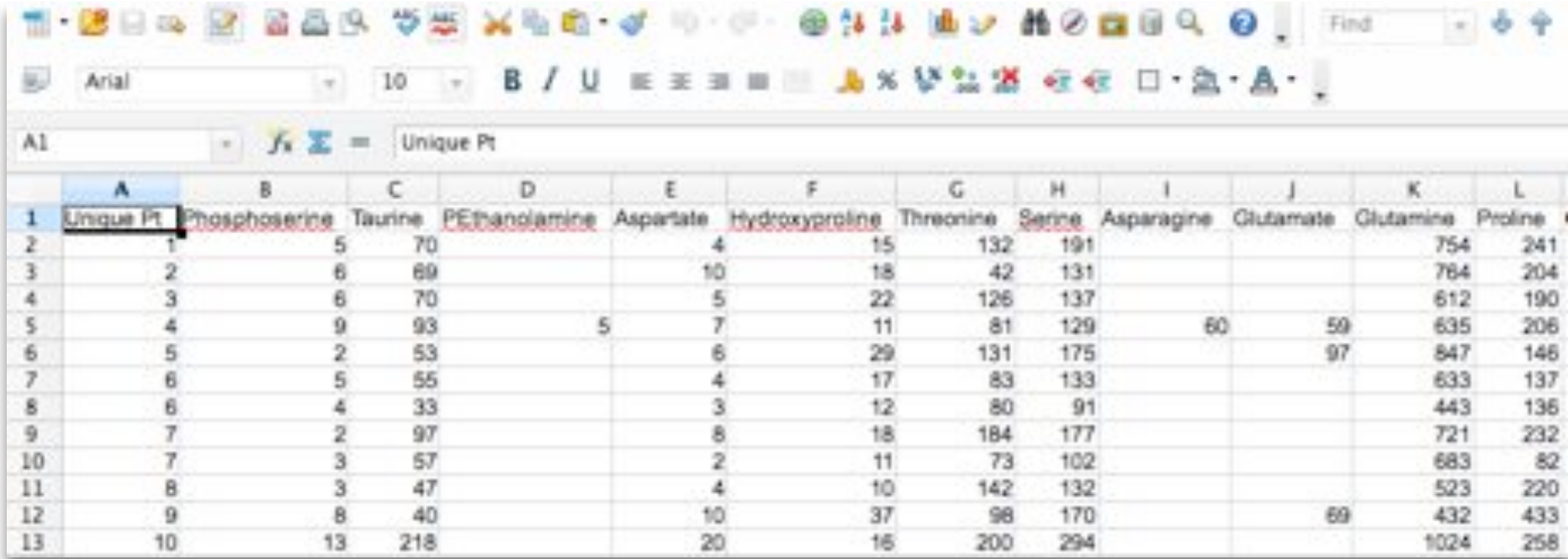


## II. Unstructured data

Text  
Videos  
Tables in web pages  
XML



# I. Structured data



The screenshot shows a spreadsheet application with a toolbar at the top and a data table below. The table has 13 columns labeled A through L and 13 rows labeled 1 through 13. The data is as follows:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Unique Pt	Phosphoserine	Taurine	PEthanolamine	Aspartate	Hydroxyproline	Threonine	Serine	Asparagine	Glutamate	Glutamine	Proline
2	1	5	70		4	15	132	191			754	241
3	2	6	69		10	18	42	131			764	204
4	3	6	70		5	22	126	137			612	190
5	4	9	93	5	7	11	81	129	60	59	635	206
6	5	2	53		6	29	131	175		97	847	146
7	6	5	55		4	17	83	133			633	137
8	6	4	33		3	12	80	91			443	136
9	7	2	97		8	18	184	177			721	232
10	7	3	57		2	11	73	102			683	82
11	8	3	47		4	10	142	132			523	220
12	9	8	40		10	37	98	170		69	432	433
13	10	13	218		20	16	200	294			1024	258

- A data warehouse is a cost center, not a profit center.
- How can structured data be a profit center?

1. Predictive analytics

2. Visual analytics

# 1. Predictive analytics

- So, what can we do with structured data?
- Answer: Make predictions, **then take actions.**
- Example:

IEEE TRANSACTIONS ON RELIABILITY, VOL. 51, NO. 3, SEPTEMBER 2002

## Improved Disk-Drive Failure Warnings

Gordon F. Hughes, *Fellow, IEEE*, Joseph F. Murray, Kenneth Kreutz-Delgado, *Senior Member, IEEE*, and Charles Elkan

- But, what are the costs and benefits of alternative actions?
- And, **who pays which costs?**



# Cost-sensitive learning

- Cross-domain theory of making optimal decisions given predictions:



The image shows a screenshot of a Google Scholar search result. At the top left is the Google Scholar logo. To its right is a search bar containing the text "cost-sensitive learning" and a "Search" button. Below the search bar is a navigation bar with the word "Scholar" on the left and several filters: "Articles and patents", "anytime", and "include citations", each with a dropdown arrow. To the right of these filters is an envelope icon and the text "Create email alert". Below the navigation bar is the search result for the paper "The foundations of cost-sensitive learning" by C. Elkan. The title is in blue and underlined. Below the title is the author's name and the conference information: "C Elkan - International Joint Conference on Artificial Intelligence, 2001 - Citeseer". A short abstract follows: "This paper revisits the problem of optimal learning and decision-making when different misclassification errors incur different penalties. We characterize precisely but intuitively when a cost matrix is reasonable, and we show how to avoid the mistake of defining a cost ...". At the bottom of the result are several links: "Cited by 510", "Related articles", "View as HTML", "BL Direct", and "All 21 versions".

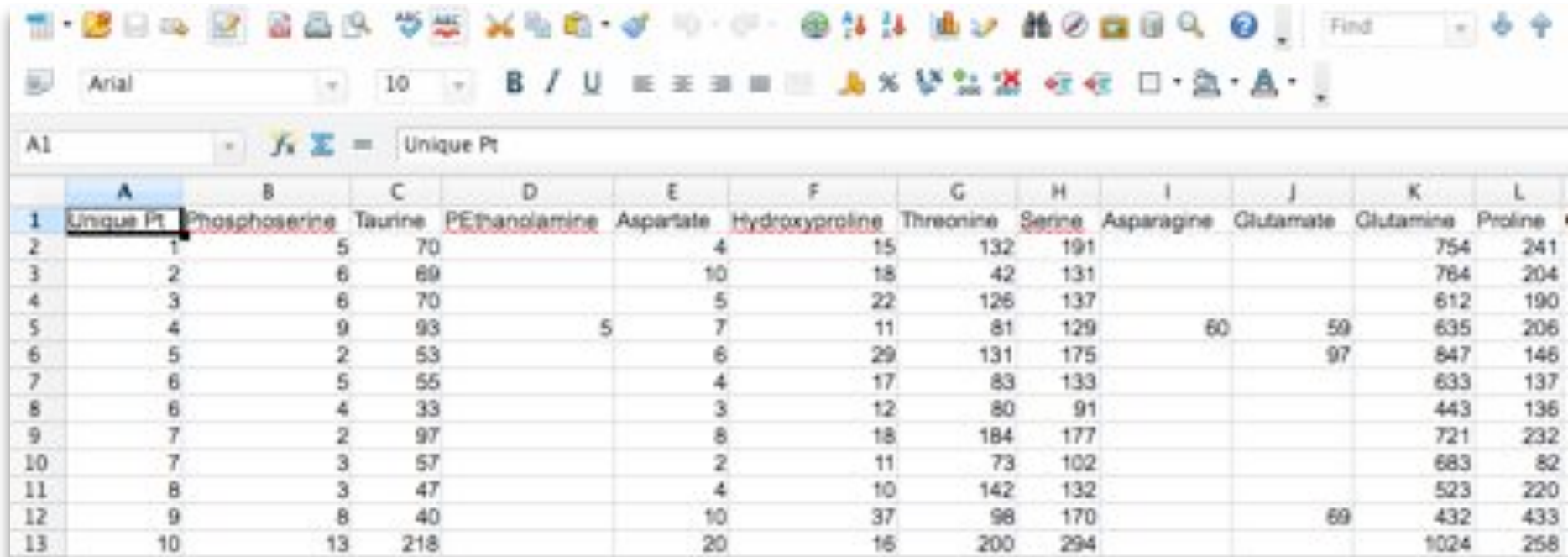
Google scholar

**Scholar**

[\[PDF\] The foundations of \*\*cost-sensitive learning\*\*](#)  
C Elkan - International Joint Conference on Artificial Intelligence, 2001 - Citeseer  
This paper revisits the problem of optimal learning and decision-making when different misclassification errors incur different penalties. We characterize precisely but intuitively when a cost matrix is reasonable, and we show how to avoid the mistake of defining a cost ...  
[Cited by 510](#) - [Related articles](#) - [View as HTML](#) - [BL Direct](#) - [All 21 versions](#)

## 2. Visual analytics

- So, what can we do with structured data?
- Answer: Find and display patterns; prompt human insight.



The image shows a screenshot of a spreadsheet application. The formula bar at the top displays the formula for cell A1: `= Unique Pt`. The spreadsheet contains the following data:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Unique Pt	Phosphoserine	Taurine	PEthanolamine	Aspartate	Hydroxyproline	Threonine	Serine	Asparagine	Glutamate	Glutamine	Proline
2	1	5	70		4	15	132	191			754	241
3	2	6	69		10	18	42	131			764	204
4	3	6	70		5	22	126	137			612	190
5	4	9	93	5	7	11	81	129	60	59	635	206
6	5	2	53		6	29	131	175		97	847	146
7	6	5	55		4	17	83	133			633	137
8	6	4	33		3	12	80	91			443	136
9	7	2	97		8	18	184	177			721	232
10	7	3	57		2	11	73	102			683	82
11	8	3	47		4	10	142	132			523	220
12	9	8	40		10	37	98	170		69	432	433
13	10	13	218		20	16	200	294			1024	258





# Information visualization

- “state of the art analytic tools to identify biomarkers”

## Clinical Metabolomics™

DIAGNOSTICS • THERAPEUTICS • DYNAMIC MEDICINE

HOME

ABOUT

CKD & DIABETES

METABOLOMICS

DRUG DISCOVERY

SOLUTIONS

TEAM

BOARD

### A biomarkers company leveraging Dynamic Systems Medicine

ClinMet™ is founded on the opportunity to harness the discoveries pioneered by Dr. Robert Naviaux and Dr. Kumar Sharma of University of California San Diego. Building on the long tradition of applying metabolomics to unravel inborn errors of metabolism at UCSD, ClinMet™ employs state of the art analytic tools to identify biomarkers that provide novel insight into disease states, and the mechanisms of action for therapeutics. By focusing on clinical samples and clinical trials, ClinMet™ enjoys a unique position in the field of Clinical Metabolomics.

## II. Unstructured data



Twitter

Home Find & Follow Public Timeline Settings Help Sign out

What are you doing? 140

Update

Archive Replies Retweet

**AGerada** wondering how big DFJ is - link to Draper Fisher Jurvetson - Premier Early Stage Venture Capital: they have capital commitments of 5.5 Billion \$ - not too bad 2 minutes ago from web

**zakieman** The cobbler's children. 4 minutes ago from web

**NewsGang** Apple: The angel vs. the reality (from Larry Dignan) - it's open season on Apple these days. The stock is - link to NewsGang 5:07:15 5 minutes ago from twitterfeed

**Moe** @Schubert I'm thinking about getting one. #sweep 8 minutes ago from Twitter in reply to Schubert

**wes** ah, warum kann denn Pölers immenoch nicht pagraden bei Fickert? 10 minutes ago from twitterfeed

**wes** @peteworldwide dockside or counter-dockside? 10 minutes ago from twitterfeed in reply to peteworldwide

**remarkk** kill me now 12 minutes ago from twitterfeed

**ipgentl** happy about the last release of - link to Communipedia ... 14 minutes ago from web

**silverg** @kriske webinale was VERY techie last year but some cool speakers. as they go web3.0 I'll be there 14 minutes ago from Twitter in reply to kriske

**silverg** there is still some room left at bercamp bodensee - link to byju.com 19 minutes ago from Twitter

**maskable** Gmail Chat Adds Invisible Mode - link to Gmail Chat Adds Invisible Mode 20 minutes ago from web

**fredwilson** @andyswan I told my reasons, please tell yours (blog post or twitter post) 23 minutes ago from web in reply to Andy Swan

**saachalobo** Ein Café in Pöchlauer Berg, drei Twitterer anwesend, ist

14 Following 53 Followers 1 Favorites 2 Direct Messages 82 Updates

People

Find friends search

A grid of small profile pictures of users.



# A case study

## **CHALLENGES:**

- Facing a highly publicized global recall, Toyota needed a way to understand its quality data – yet had an exponentially growing number of questions and a fraction of time to react
- Gave close to 1,000 users, from quality engineers to executive dashboard users the ability to analyze quality data from heterogeneous sources

## **RESULTS:**

- Allowed users to design reports and dashboards in minutes
- Delivered analytics on 6 years of structured and unstructured data from more than a dozen sources with 110 analytical dimensions, and 250 analytical components
- Will eliminate hundreds of thousands of hours of end-user wait time per year

# A general need: Task-oriented semantic search

LaVerne Council, CIO of Johnson & Johnson:

“... allow anyone to ask a question ... folks that have given us access to their email ... data mining for answers to that question

... help us solve a very hairy issue for one of our products ... one of the associates had completed his thesis in college on that very topic ... they weren't in the same company

... we were able to really come back with answers.”

# A grand vision

- “Open source intelligence (OSI)”

THE IDEA LOBBY

September 8, 2011

## Spy Agency Seeks Digital Mosaic to Divine Future

The U.S. intelligence community wants to mine lots and lots of the tidbits bopping around on the Internet to suss out trends before they make the news.

By Emily Badger

... Charles Elkan, a computer scientist at the University of California at San Diego, suspects the biggest value of all these tools will come from quantifying what social scientists already know.

“Social scientists have said for 100 years that revolutions happen at times of rising expectations,” he said. “If, for example, society is becoming more prosperous, people start have rising economic expectations, that spills over into rising political expectations, and that can make a revolution more likely. Social scientists have said this for a long time, and now it’s beginning to be possible to quantify it.”

[COMMENTS](#) | [PRINT](#) | [SHARE](#)



*U.S. intelligence agencies hope those tidbits bopping around on the Internet will help discover all kinds of trends before they make the news. (John Fox/Stockbyte)*



# A less grand vision

## THE WALL STREET JOURNAL.

Today's Paper - Video - Columns - Blogs - Graphics - Newsletters & Alerts - Journal Community

HOME U.S. WORLD BUSINESS MARKETS TECH PERSONAL FINANCE LIFE & STYLE OPINION CAREERS REAL ESTATE SMALL BUSINESS

BUSINESS TECHNOLOGY | NOVEMBER 23, 2010

989

## Using Software to Sift Digital Records

By NATHAN KOPPEL



"The biggest pain point in litigation is the amount of money spent on attorneys reviewing documents," said Jonathan Redgrave, a Washington, D.C., lawyer ...

... used so-called predictive coding software made by Recommind Inc. to respond to a government investigation ...

Attorneys at the firm started by reviewing a relatively small set of records to identify the important characteristics. ...

That knowledge was then coded into software, which was used to scan a larger universe of electronic records. Attorneys then reviewed the most relevant records to make the final determination about whether they should be disclosed.

... Recommind licenses its software for costs ranging from about \$650 per gigabyte of data analyzed to several million dollars annually for unlimited data.

# III. The business of analytics

- Analytics **applications** are valuable.

## Heritage Provider Network Announces the Official Launch of the \$3 million Dollar Heritage Health Prize

MARINA DEL REY, Calif., April 4, 2011 /PRNewswire/ -- Dr. Richard Merkin, President and CEO of Heritage Provider Network, announced today the official launch of the \$3 million dollar Heritage Health Prize, the world's largest predictive modeling contest. ...

Heritage Health Prize Advisory Board members include Arvind Narayanan, Stanford University; Charles Elkan, UC San Diego and Netflix prize judge and Claudia Perlich, MediaSixDegrees, winner and organizer of several data mining competitions

For more information on the HPN Health Prize go to: [www.heritagehealthprize.com](http://www.heritagehealthprize.com)

Analytics **companies** are valuable

## Oracle Endeca Deal Echoes HP's Autonomy Purchase

Oracle plans to use Endeca's technology to boost unstructured data access and analysis for Web commerce and business intelligence applications.

By [Doug Henschen](#) InformationWeek

October 18, 2011 03:50 PM

Oracle announced Tuesday that it plans to acquire Endeca Technologies in a deal that will help it bring unstructured data into e-commerce transactions and business intelligence analyses. It's no coincidence that the deal was announced just weeks after Hewlett-Packard finalized its \$10-billion-plus acquisition of Autonomy.



# Are valuations bubble-icious?

- HP compared to Autonomy:

Sales: \$128B versus \$963M

Income: \$12B versus \$343M

Value: \$50B versus \$11B

- Forrester: “The Autonomy IP is stagnant. There hasn’t been a major release in five years.”
- Zero recent patents for the core analytics.

# IV. A research *and* market opportunity

[Forrester Blogs](#) • [Information Technology](#) • [Content & Collaboration Professionals](#) • [Leslie Owens](#)

## What Is Autonomy, Without Its Marketing?

Posted by Leslie Owens on August 26, 2011

### AUTONOMY ARCHITECTURE OVERVIEW



Source: HP Investor Relations

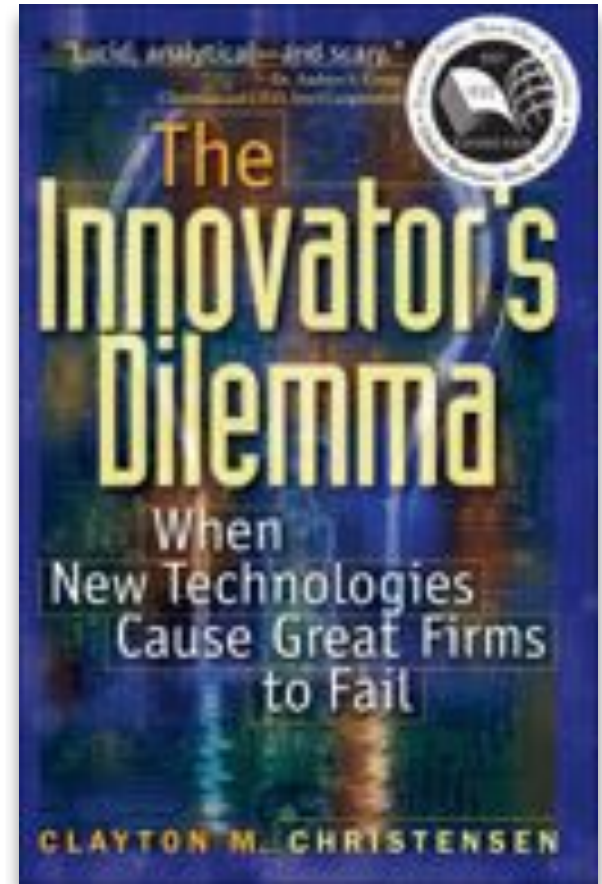
... At the heart of Autonomy's stack is the Intelligent Data Operating Layer (IDOL) – its brand name for search and content processing technology.

... The IDOL IP is stagnant. There hasn't been a major release of IDOL in over 5 years.



# Disruption from below

- New platform for diverse data
  - Cloud-based
  - Multiply the user base 10x:
    - Easy to use
    - Fun to use
- Opportunity: Add “secret sauce” to open-source software
  - Newer artificial intelligence
  - Patented artificial intelligence





# Simple, Secure Sharing

Share, manage and access all your business content online. [Learn More](#)



## Try It Free

Compare plans and pricing

## Talk To Sales

Call us at 1-877-729-4269

## Quick Tour

A short video about Box.net

### Simple. Open. Mobile.

Hear from Box co-founder Aaron Levie, our developers and designers on Box's simple, open and mobile approach to enterprise software.

• [Watch video](#)



### Security in the Cloud

With Box.net you can store and share your content with confidence. 99.9% up-time guarantee, SSL encryption, redundant storage, configurable permissions, and more.

- [View Security Whitepaper](#)
- [Learn more about Box.net's security](#)



### The Microsoft SharePoint Alternative

Forget expensive servers, training and excessive demands on your IT department and use Box to share, access and manage files in the cloud.

- [Box and SharePoint quick comparison](#)
- [Comparison whitepaper](#)

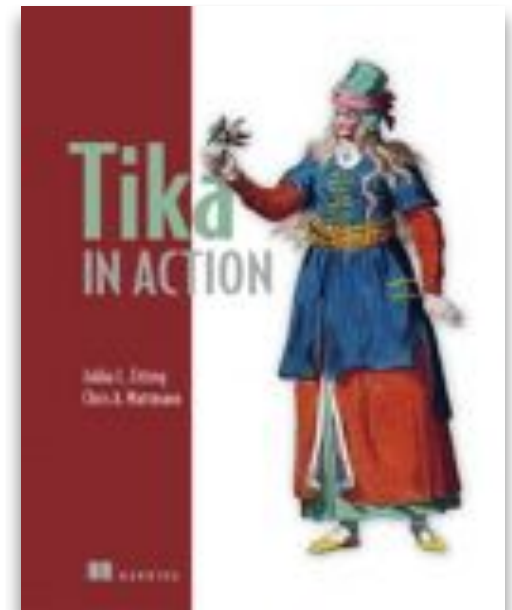


- A role model for cloud-based ease of use: Box.net
- \$650M valuation, but no intelligence.

# Disruption from below

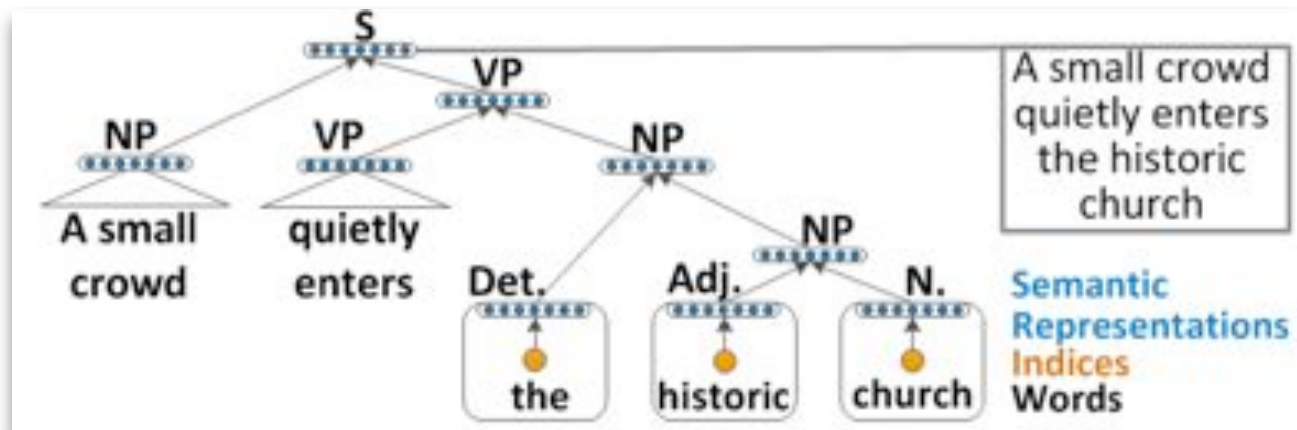
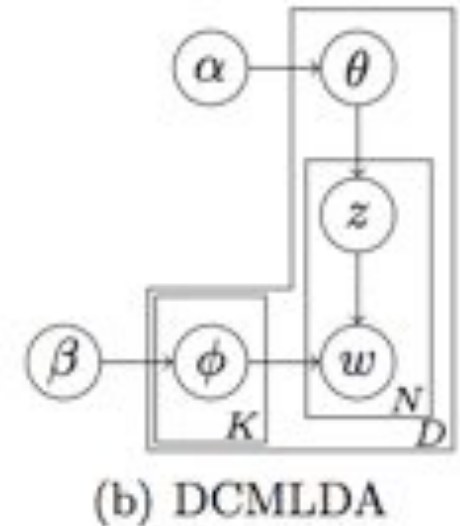
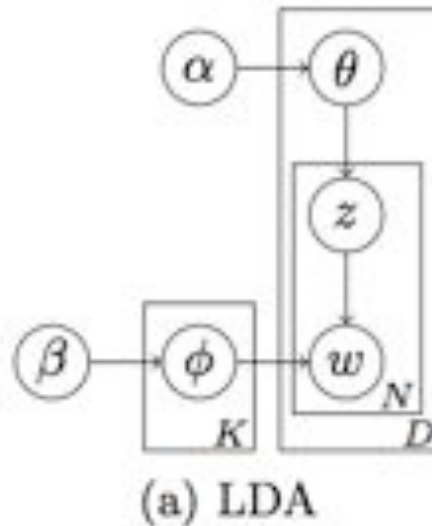


- Cloud-based software as a service (SaaS)
- Easy to use, fun to use
- Newer AI, patented AI
- **Open-source foundation:**
  - Lucene and Solr as backend
  - Tika for importing unstructured data



# Newer artificial intelligence

- Sentiment analysis
  - Topic models for organizing content
  - Recursive neural nets for deep understanding
- [www.socher.org/index.php/Main/ParsingNaturalScenesAndNaturalLanguageWithRecursiveNeuralNetworks](http://www.socher.org/index.php/Main/ParsingNaturalScenesAndNaturalLanguageWithRecursiveNeuralNetworks)



# Newer AI: Fewer topics, better fit

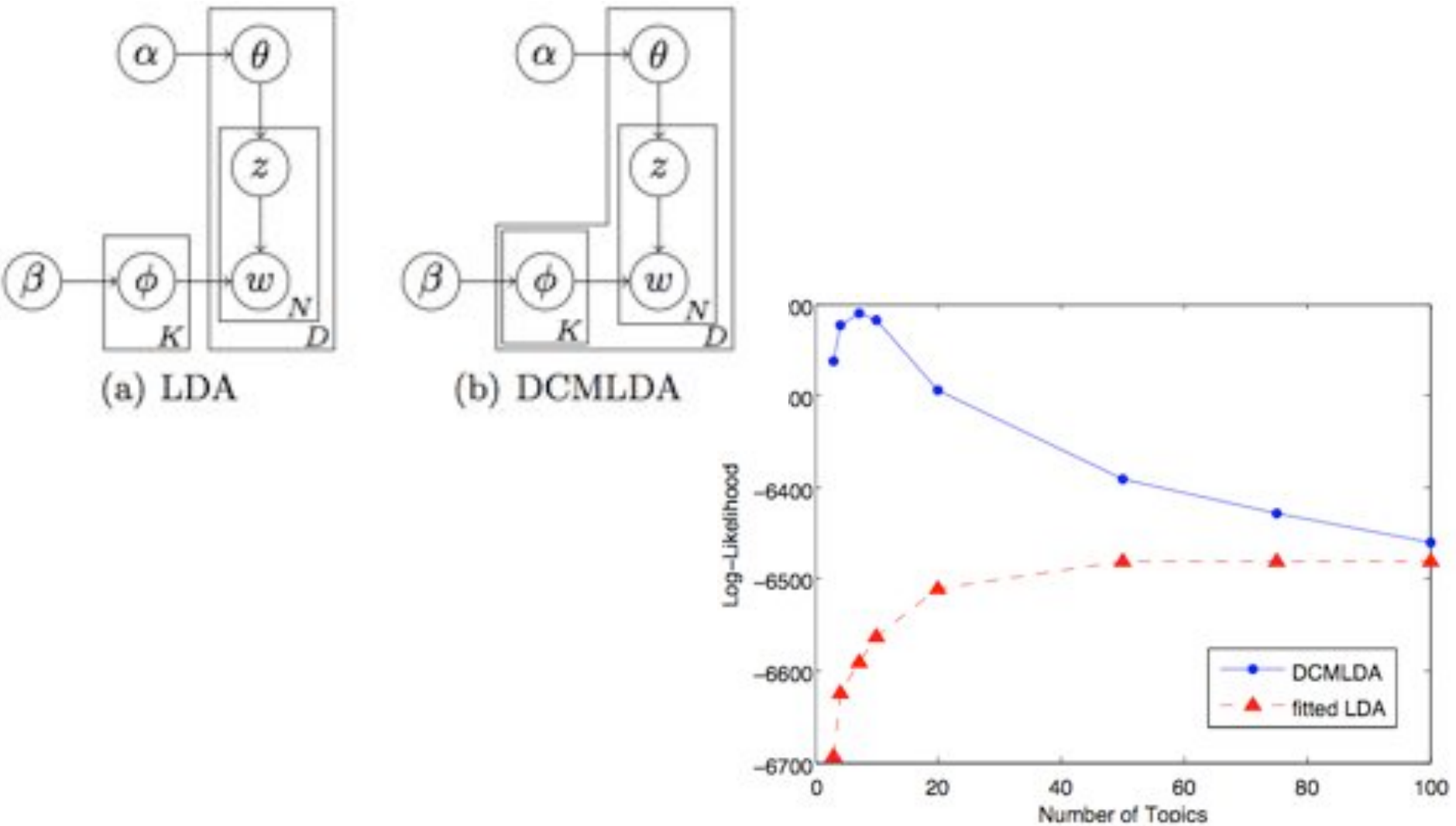


Figure 2. Mean per-document log-likelihood on the S&P500 dataset for DCMLDA and fitted LDA models.

# Patented AI: Sentiment analysis

- ... labels designate level of quality, such as interestingness, appropriateness, timeliness, humor, style of language, obscenity, **sentiment**
- ... a classifier means effective to **automatically associate a quality value to items of data**, wherein said quality value is indicative of the qualitative nature of said items of data

(12)	<b>United States Patent</b> <b>Elkan</b>
(54)	<b>METHOD AND SYSTEM FOR SELECTING DOCUMENTS BY MEASURING DOCUMENT QUALITY</b>
(75)	Inventor: <b>Charles Elkan</b> , San Diego, CA (US)
(73)	Assignee: <b>The Regents of the University of California</b> , Oakland, CA (US)
(* )	Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 455 days.
(21)	Appl. No.: <b>10/004,514</b>
(22)	Filed: <b>Nov. 2, 2001</b>



# Today in the *New York Times*



Ashley Pon/Bloomberg News

## **Apple Wins Partial Victory on Patent Claim**

By NICK WINGFIELD 12 minutes ago

The ruling by a U.S. agency, involving a set of important smartphone features, could force changes in Google's Android phones, including HTC's, above.



# SQUID

- Sentiment analysis
- Question answering
- Unstructured data organization
- Interactive insight
- Diverse entity extraction
  
- But what will be most beneficial and profitable?
  
- Historical answer: Specific vertical applications.

# Profit lies in verticals, I

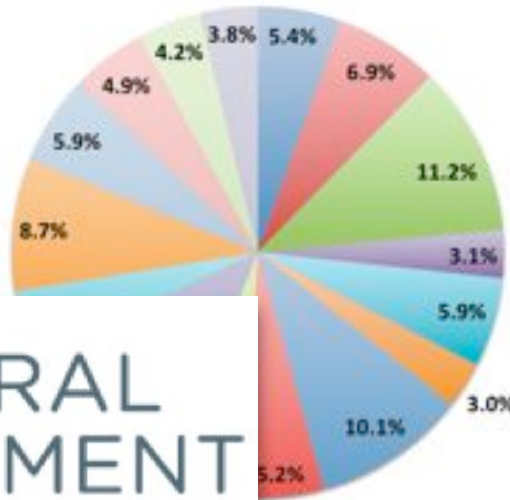


Many news and social media publishers create new content every day.

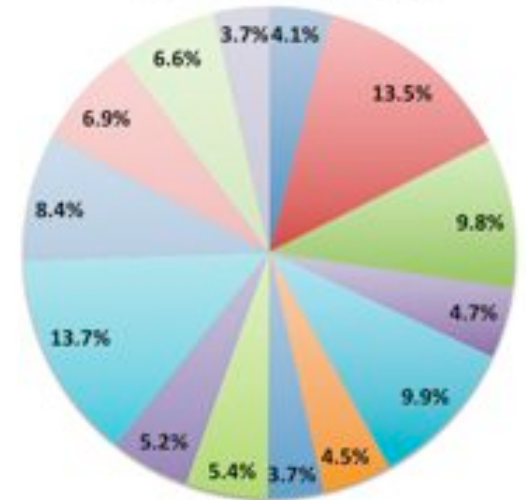
Sentence by sentence, General Sentiment detects the sentiment in each document related to a topic...

...giving you real-time, actionable information in our Executive Sentiment Dashboard and customized reports.

90210



American Dad!



GENERAL SENTIMENT

Top Match

**Mercedes-Benz**



Mercedes-Benz

Top Match

**Hyundai**



HYUNDAI

# Profit lies in verticals, II

December 8, 2011 1:32 pm

## Financial groups hit by flood of new rules

By Brooke Masters in London



Financial services firms worldwide are being hit with an average 60 regulatory changes every working day, a 16 per cent increase over last year, and no let up is in sight, a study has found.

Regulators around the world announced 14,215 changes in the twelve months to November, up from 12,179 for the same period a year earlier, according to new

research by the Thomson Reuters governance, risk and compliance unit.

### More

#### ON THIS STORY

[HSBC to widen mis-selling bond review](#)

[Regulators warn on banks' risk evaluation](#)

[Analysis European banks face a dual burden](#)

The study tracks everything from the passage of new laws and short-selling bans to the issuance of consultation papers and speeches that contain policy announcements; in short everything compliance officers are expected to keep abreast of. The rules range from global packages like the Basel III bank capital reforms to local rules in individual US states.

# Discussion

- Acknowledgement: Most images are due to other authors.