


FROM DATA TO SOLUTIONS

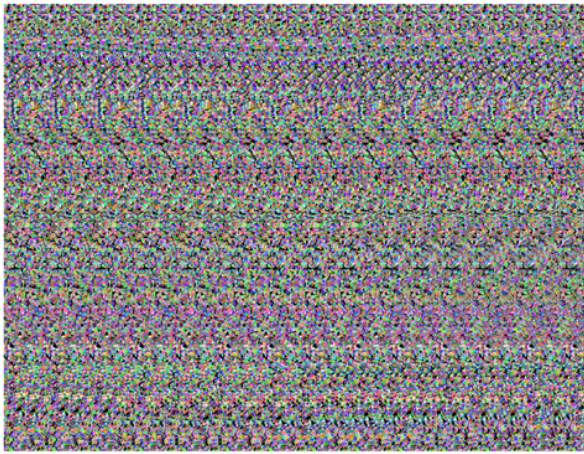
MINE YOUR OWN BUSINESS


Oded Netzer

Columbia Business School

 Columbia Business School

October 5, 2012



 Columbia Business School

FOLLOW THESE INSTRUCTIONS TO UNLOCK A HIDDEN MESSAGE!

① STARE AT YOUR COMPUTER FOR UNHEALTHY AMOUNTS OF TIME (IT'S CALLED "RESEARCH")

② ALLOW YOUR EYES TO GLAZE OVER AND YOUR MIND TO START QUESTIONING REALITY

③ START SEEING THINGS THAT ARE NOT REALLY THERE!

JORGE CHAM © 2012

WWW.PHDCOMICS.COM



Real Time
+10 posts this hour

Most Popular
The Case For The 'Unqualified'



TECH | 2/16/2012 @ 11:02AM | 1,549,799 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. [Target](#), for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the [New York Times](#) how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things



Target has got you in its aim



Journal of Advertising Research



Guest Editorial:
The Shape of Marketing Research in 2021

ANCA CRISTINA MICU, KIM DEDEKER, VAN LEWIS, ROBERT MORAN, GODEF NETZER, JOSEPH PLAMMER, and JOEL RUBINSON

INTRODUCTION

It is year 2011. A chief marketing officer (CMO) sits at his desk very early one morning. His consumer insights team's desk from a presentation the previous night still is in front of him. The CMO leans back in his chair, takes his glasses off, starts sipping, then thoughtfully, and dives deep into thought.

It's haunting when I think back 10 years, no broadband, no social media, no smartphones, no 50-inch LED TVs, no DVRs, no consoles, no iPads, and Google hadn't had its IPO. The term "conversion" hardly was taking off—now my company is into "conversioning" (Jensen, 2006; White, 2008). In the last decade, many industries went through what Andy Grove labeled "strategic-reflection points"—those moments when the balance of forces shifted from the old structure and the old ways of doing business and competing to new ones (Grove, 1996) the music business, the book business, the publishing business, even the original Internet leader, AOL. Will my business be next? What will be the "normal" 10 years from now? What will be the "next big things?"

I do know that "disruption of overabundance" will be the mantra. I am certain the rate of change will keep accelerating—the old, Facebook went from nothing to 500 million users in just a year. And so finally realized that we marketers are not in control anymore. I know that Internet across endpoints will be a given that geo-marketing will be pervasive; that retail environments will be transformed by digital technologies; that smartphone capabilities will be far more advanced; that RFID will have a big impact even now; and so on. And none of this even touches changes that won't be driven by technology. The global economic balance of power will shift substantially in the next decade, driven by the BRIC countries and led by China.

I also know that all of this is only the tip of the iceberg—I just can't see the right roads beneath the surface yet.

The basics of marketing don't change. I still need to identify, design, and market products and services that satisfy customer needs even as they keep ahead of the competition. I must do a better job in several ways. I need to be better at anticipating the future, at sensing consumer and customer needs, at being faster to market, at communicating and interacting with consumers and customers, at understanding and delivering against consumer needs around the world, and at recognizing potential reflection points that could either bring great potential or destroy my business.

I wonder what the "new normal" will be...

ECOLOGICAL OR STRATEGY?
Induced or Autonomous Adaptation

If anything else, the "new normal" will either constant change and adaptation. Focusing on strategic change in competition, some scholars have documented major epochs—periods of quantum change, and reorientations in strategy making—whereas others have documented the ongoing process of strategy making in organizations.

From an organizational strategy perspective, attributing changes to sweeping, environmental triggers or long-term strategic planning means taking either an ecological or strategic view-point. The ecological versus-strategic debate centers on the issue of environmental determinism versus strategic choice.

Whether forced by the environment or as the result of strategic planning, adaptation to change triggers sets of activities within a company grouped in an adaptation process. Adapting companies follow either induced or autonomous processes to adapt. The induced process concerns initiatives that are within the scope of the organization's current strategy and build on existing organizational learning whereas the autonomous process concerns initiatives that emerge outside of it and provide the potential for new organizational learning (Chatterman 1993).

The Shape of MR in 2021- The River



- ▣ Our analogy for the industry in 2021 is a RIVER
- ▣ The fundamental premise is that research in 2021 will represent a continuous and organic flow of knowledge
- ▣ There are 1,000's of tributaries that feed the river. These tributaries represent individual information sources
- ▣ In our new world, the knowledge exists before the business question is formed
- ▣ There will be a fundamental shift in how we approach business decision making and influence of strategy. We move away from a project orientation toward an ongoing process of knowledge access and utilization. Value creation is catalyzed from the organic knowledge found in the flow of the river

Where Do We Stand on Using Social Media for Business Decisions?



- ▣ Lots of counting, not enough evaluating
- ▣ Too much 'data', not enough 'solutions'
- ▣ Thinking Big Data is primarily an IT challenge
- ▣ Some success stories...



POTTERY BARN



Mine Your Own Business: Market Structure Surveillance (Netzer, Feldman, Goldenberg, and Fresko 2012)



- Can we use the Web as a marketing research playground?
- Can we quantify the rich, yet unstructured, information consumers post on the web?



Uncovering market structure from information consumers are posting on the Web

What Are We Going to Do?



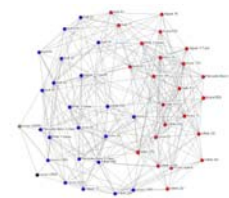
- Text mine consumer postings
- Use network analysis framework and other co-occurrence methods of analysis to reveal the underlying *market structure*



Text Mining

	compact	sport	old
Audi A6	67	345	56
Honda Civic	1384	539	245
Toyota Corolla	451	128	211

Co-occurrence and
Network Analysis Methods



Text Mining Background



- A rapid flow (**river**) of information available in digital format
- Managers have **less time** to absorb more information



Mining Consumer Forums



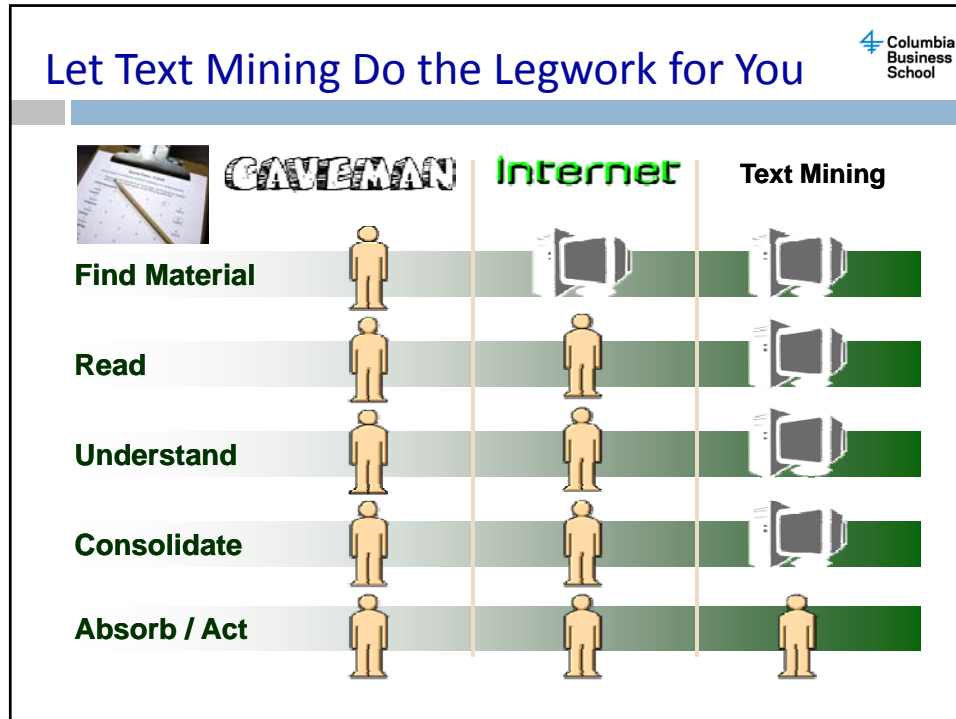
Opportunities

- A combination of observational and descriptive marketing research
- Permits both qualitative and quantitative information
- Non-invasive (no demand effect)
- Minimizes recall error
- Very rich data
- Sample size is not an issue
- Real time data

Difficulties

- Massive amount of data
- Data is all over the Web
- Data is unstructured
- Population may not be representative
- Topic of discussion may not be representative






The Text Mining Process


Columbia Business School

- ❑ **Downloading:** html-pages are downloaded from a given forum site
- ❑ **Cleaning:** html-like tags and non-textual information like images, commercials, etc. are cleaned from the downloaded pages
- ❑ **Chunking:** the textual parts are divided into informative units like threads, messages, and sentences
- ❑ **Information Extraction:** products and product attributes are extracted from the messages
- ❑ **Comparisons between products are extracted:** either by using co-occurrence analysis or by utilizing learned comparison patterns



The Text Mining Part

Edmunds.com sedan forum



where smart car buyers start

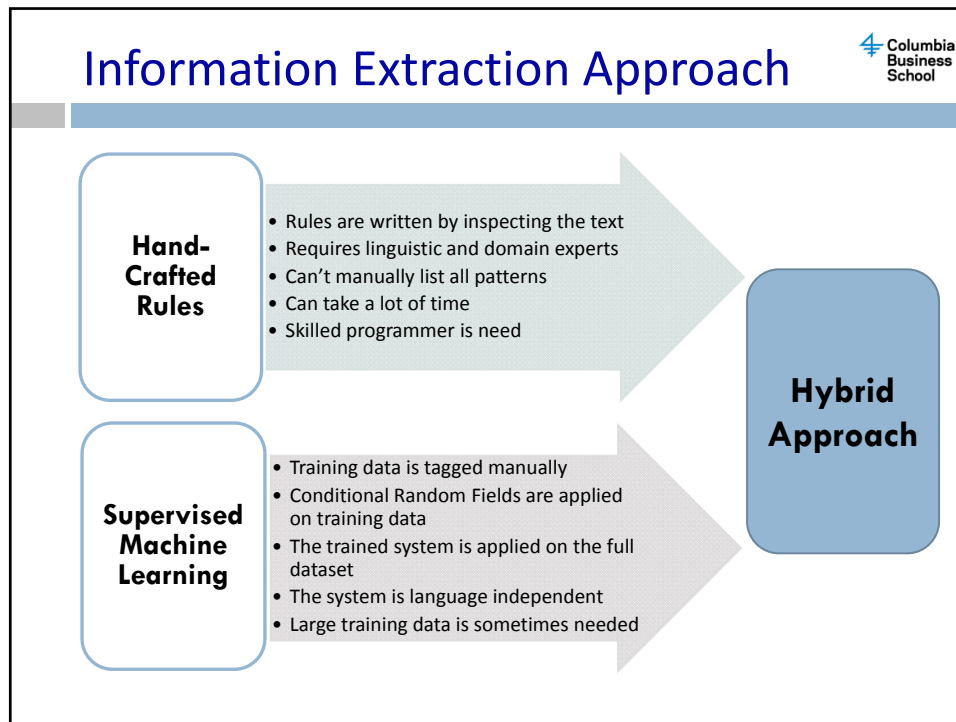
Honda Accords and Toyota Camrys are nice sedans, but hardly the best car on the road (for many people). It's just that they are very competent in their price range. So, a love fest of the best selling may not tell you what is "best".

- <Brand>Honda</Brand>
- <Model>Honda Accord</Model>
- <Brand>Toyota</Brand>
- <Model>Toyota Camry</Model>
- <Term>Sedans</Term>
- <Term>Best</Term>
- <Term>Competent</Term>
- <Term>Price</Term>
- <Term>Love</Term>
- <Term>Best selling</Term>
- <Term>Best</Term>

Major Issues

- Handling Negation
 - Prevent bone loss
- Deciding if a phrase is positive or negative, verbs alone are not enough, and nouns alone are not enough.
 - Reducing losses vs. Reducing forecasts
- Anaphora Resolution
 - The company
 - The 3rd biggest US oil producer (COP)
- Catching Meaningful Phrases





Some Text Mining Difficulties

 Columbia Business School

- ❑ We are interested in:
 - **Brand names** (e.g., car companies)
 - **Model names** (e.g., car models)
 - Some **common terms** (mostly noun-phrases and adjectives)
- ❑ **Brand names** - are relatively easy
 - Need to deal with abbreviations and spelling mistakes
- ❑ **Models** - are more complex
 - Variations in writing styles
 - Honda Civic could be written as "Honda Civic"; "Civic"; "Honda Civic LS"; "Honda Civic LE"; "LE"; "H. Civic"; "Hondah Sivik"
 - Model numbers can be written as: 5, V, Five
"The Audi A6 is great! the 6 is better than the 4"
 - Model can be referred to as numbers but numbers do not always refer to models (e.g., "1010 for New Balance 1010", but \$1010)

How Accurate Are We?

Information Type	Recall	Precision	Overall Accuracy
Car Brands	98%	98%	98%
Car Models	88%	95%	91%
Drugs	89%	100%	94%
Side Effects	74%	90%	82%
Drug--Side Effect	60%	96%	74%

Empirical Applications



Empirical Applications




Sedan Cars Application: Edmunds.com



“ Honda Accords and Toyota Camrys are nice sedans, but hardly the best car on the road (for many people). It's just that they are very compentant in their price range. So, a love fest of the best selling may not tell you what is "best". That depends very much on what is important to you. A car could have a quirk, that you would just love, but not be popular to many people. Thus, the best car for you might not sell many. If you are looking for resale value, then it might be a factor.”

Product Co-occurrence Data




Message #1199 Civic vs. Corolla by mcmanus Jul 21, 2007 (4:05 pm)



Yes DrFill, the Honda car model is sporty, reliable, and economical vs the **Corolla** that is just reliable and economical. Ironically its Toyota that is supplying 1.8L turbo ... Neon to his 16 year old brother. I drove it about 130 miles today. Boy does that put all this **Civic vs. Corolla** back in perspective! The Neon is very crudely designed and built, with no low ...

Audi A6	Honda Civic	252
Audi A6	Toyota Corolla	101
Honda Civic	Audi 6	252
Honda Civic	Toyota Corolla	2762
Toyota Corolla	Audi A6	101
Toyota Corolla	Honda Civic	2762

Associative Network




	Audi A6	Honda Civic	Toyota Corolla
Audi A6	---	252	101
Honda Civic	252	---	2762
Toyota Corolla	101	2762	---

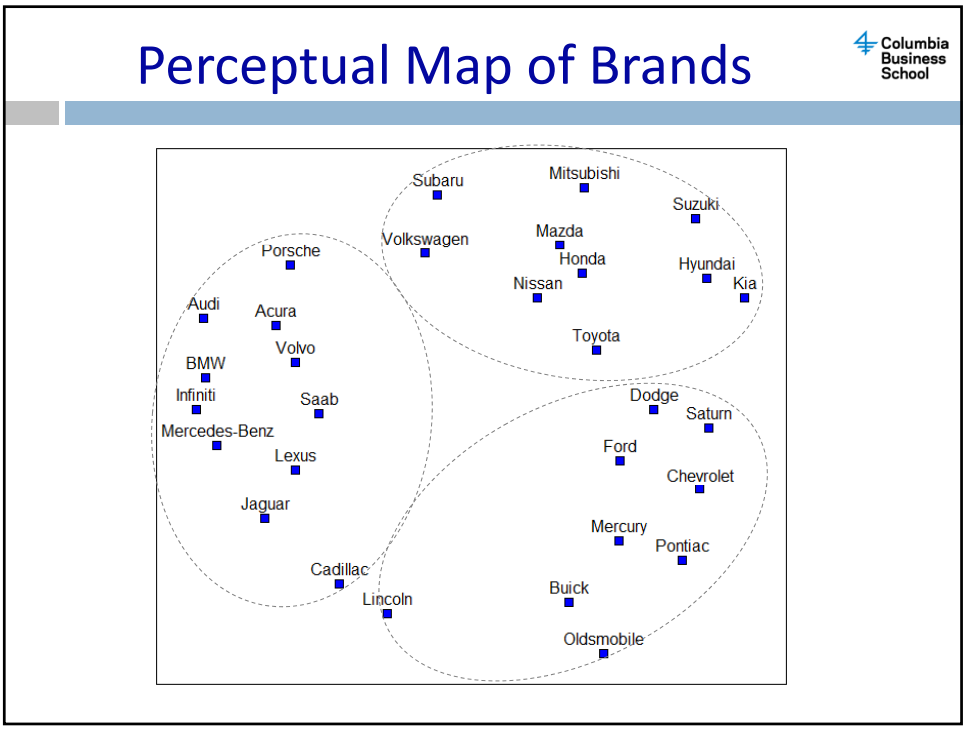
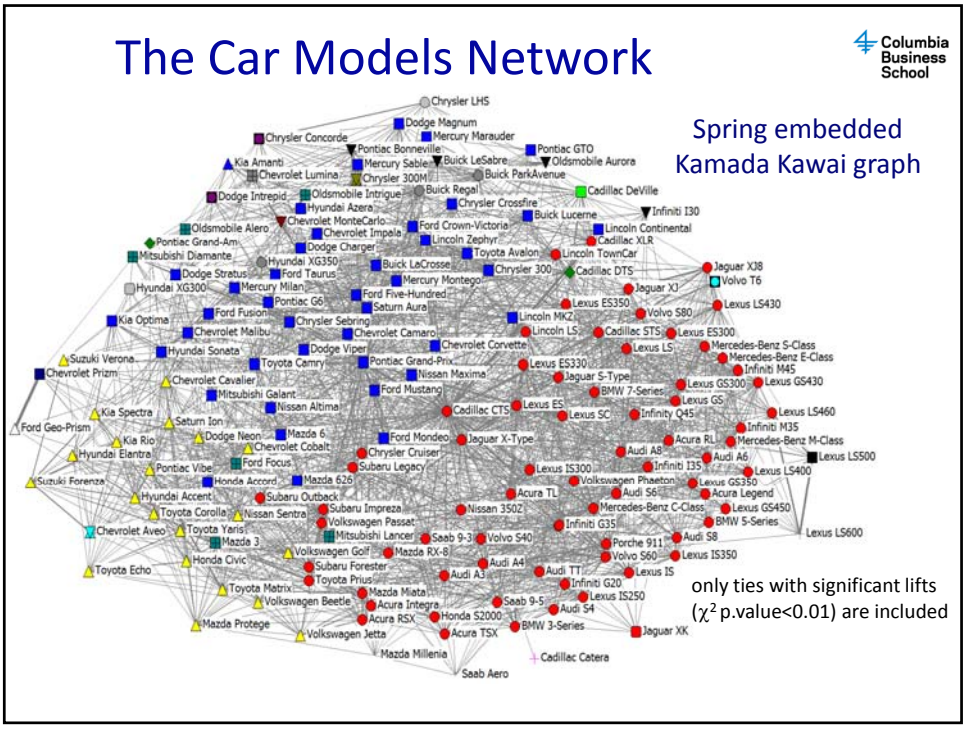



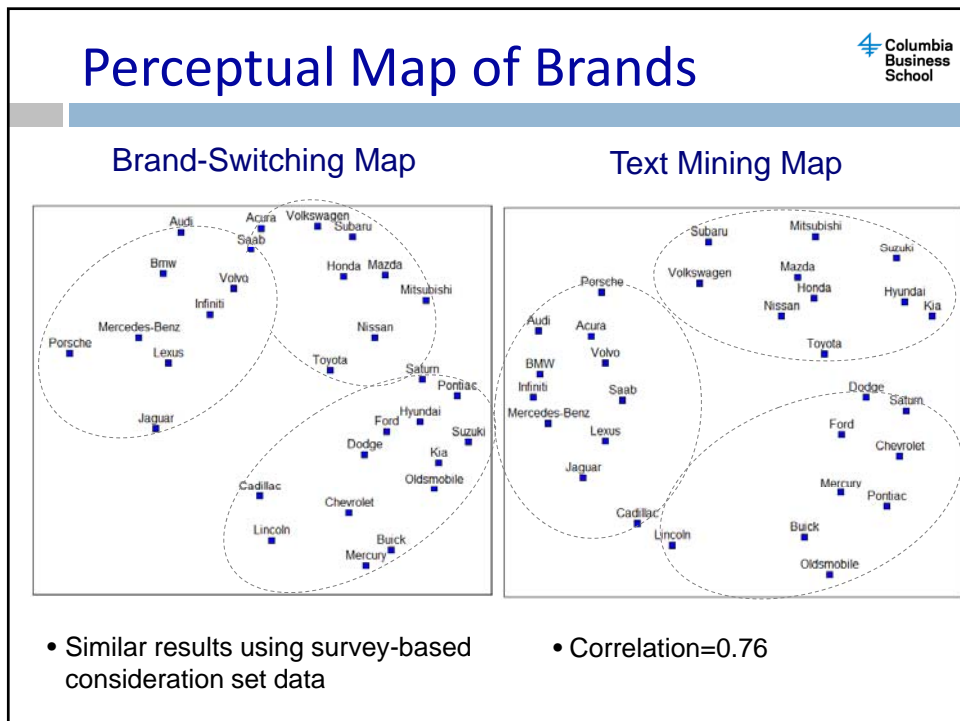
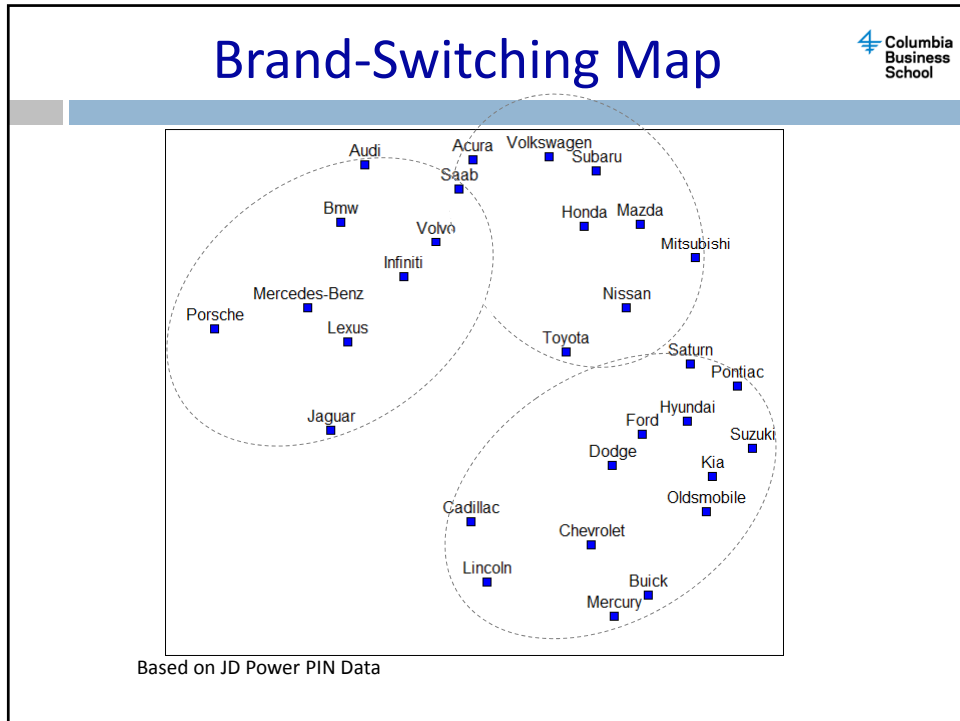
$$lift(A, B) = \frac{P(A, B)}{P(A) \times P(B)} = \frac{C(A, B)}{C(A) \times C(B)} \times N$$

Cars That Are Most Central to the Discussion



	Car model	Degree	Betweenness	Eigenvector	No. of occurrences
1	Honda Accord	100.00	0.954	13.56	60,546
2	Toyota Camry	99.41	0.891	13.54	36,038
3	Volkswagen Passat	99.41	0.861	13.53	17,151
4	BMW 3-Series	98.81	0.710	13.48	34,595
5	Lexus ES	98.81	0.794	13.45	10,630
6	Nissan Altima	98.81	0.441	13.49	13,191
7	Nissan Maxima	98.81	0.385	13.51	11,501
8	Toyota Avalon	98.81	0.412	13.48	13,153
9	Honda Civic	98.21	0.402	13.39	25,037
10	Infiniti G35	98.21	0.417	13.47	24,131
11	Mazda 6	97.02	0.372	13.29	10,664
12	Acura TL	96.43	0.363	13.36	27,262
13	BMW 5-Series	96.43	0.402	13.32	27,075
14	Chrysler 300	95.83	0.411	13.28	5,658
15	Toyota Corolla	95.83	0.342	13.21	7,256
16	Cadillac CTS	95.24	0.352	13.24	6,582
17	Lexus GS	95.24	0.284	13.12	8,843
18	Audi A4	94.64	0.298	13.18	11,974
19	Audi A6	94.64	0.278	13.13	13,848
20	Mercedes-Benz E-Class	94.64	0.325	13.12	9,220





BusinessWeek

The Second Coming Of Cadillac

Nov. 4, 2003
By David Welch

PUTTING A NEW SPIN ON CADDY

GM has taken significant strides toward making Cadillac a stronger rival to luxury import cars:

- 1 IMAGE** Ads featuring Led Zeppelin's rock music seized boomers' attention. Now Caddy will begin focusing more on its improved sporty ride and handling. It's also putting its cars front and center at glitzy events like the Oscars and Wimbledon.
- 2 QUALITY** GM's highly automated \$540 million Cadillac plant in Lansing, Mich., is one of the most efficient auto factories in the U.S. More important, the cars have earned consistently high marks in the J.D. Power & Associates quality survey.
- 3 PERFORMANCE** Beating BMW and Mercedes requires an upgrade under the hood. In January, Cadillac will start selling the CTS-V, a 400-horsepower version of the CTS that will hit the racing circuit.
- 4 PRESENTATION** A new incentive plan doubles bonuses for dealers who upgrade showrooms to a more cutting-edge look that includes black porcelain tile floors and black leather furniture.



CAR AND DRIVER

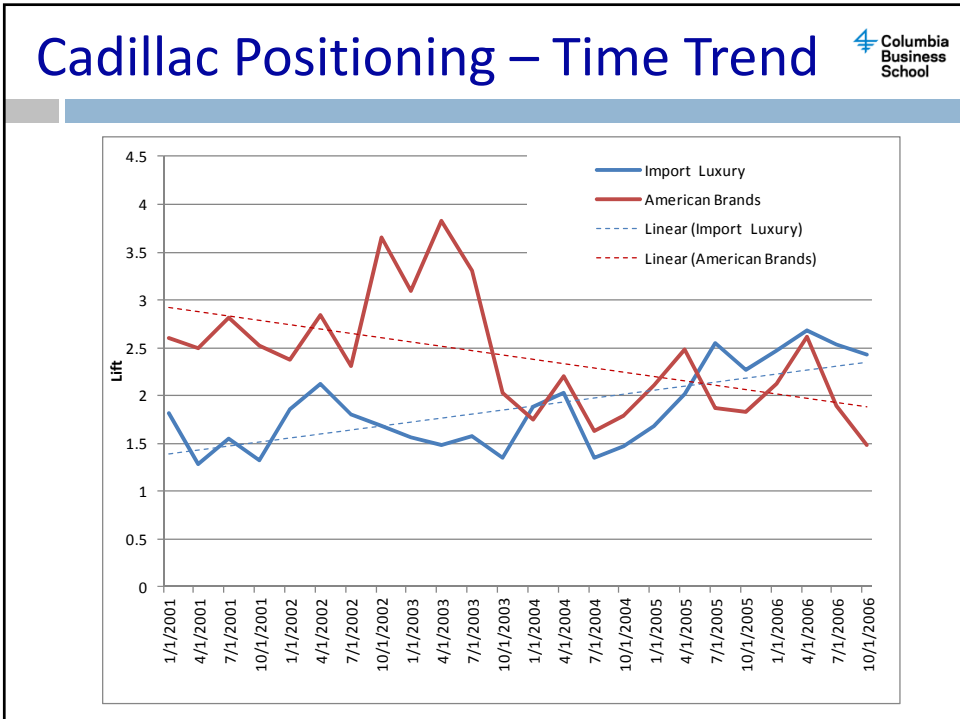
2004 CADILLAC XLR

Cadillac stakes a claim in the luxury-roadster arena.

June 2003
BY CSABA CSERE

"Looks like Cadillac intends to become a full-service luxury carmaker again."

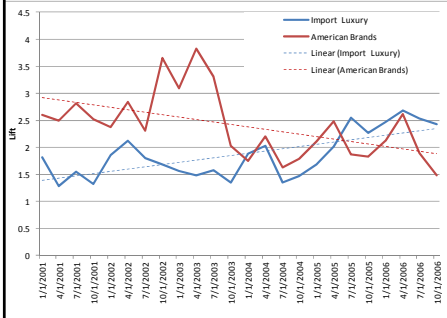




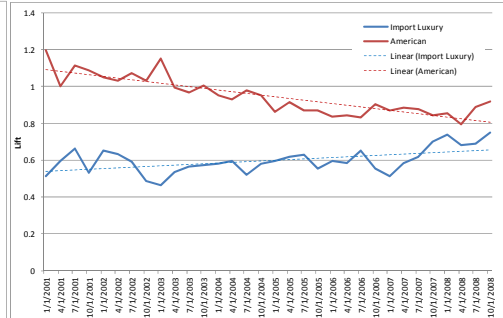
Cadillac Positioning – TM vs. Sales



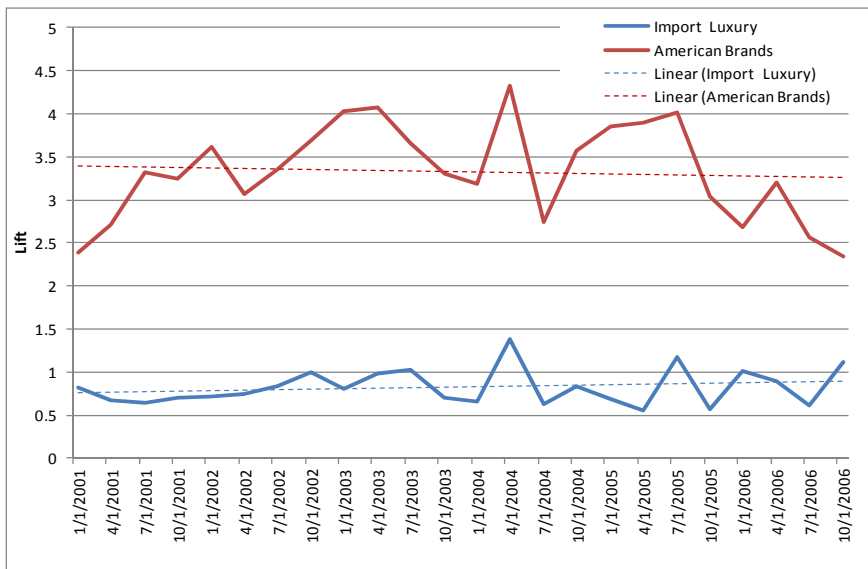
Text-mining-based trend



Sales-based trend



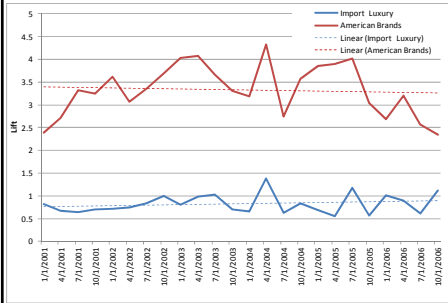
Buick Positioning – Time Trend



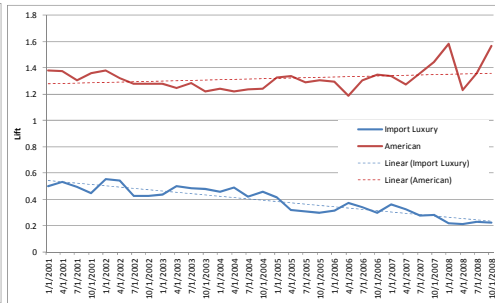
Buick Positioning – Sales vs. TM



Text-mining-based trend



Sales-based trend



Model-Term Data



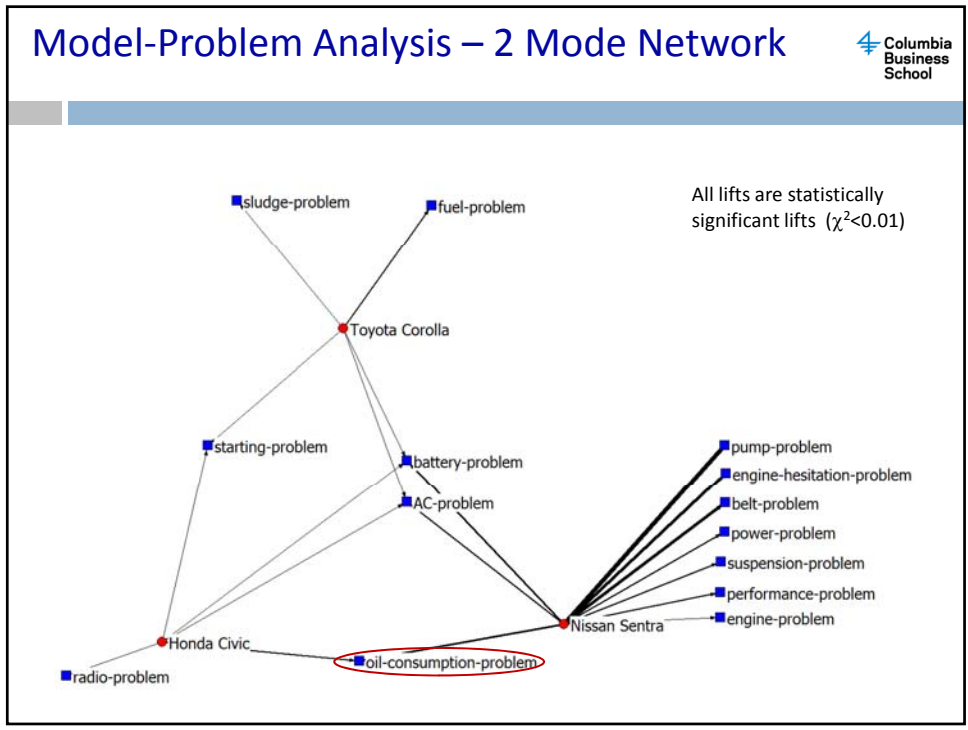
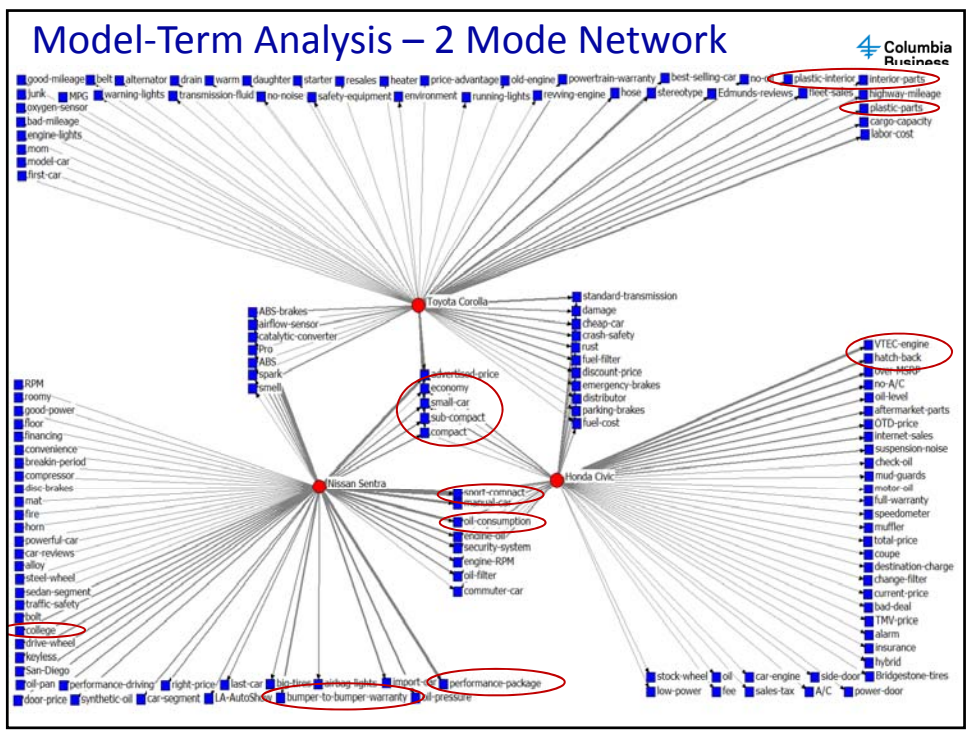
Message #1449 Bold and repulsive by eldaino Jul 20, 2007 (9:33 am)
 i agree with what robertsmx has said; a bold design does not guarantee a 'ooh thats hot!'. I will say this; when i got my **civic** (06) i went inside to a mcdonalds to eat and a gentleman asked me if that was a new **civic** and he commented on how sharp and **sparty** it looked. The **civic** sedan may not be as 'exciting' or '**sparty**' as the coupe, but ...



- Audi A6 compact 67
- Audi A6 sport 345
- Audi A6 old 56
- Honda Civic compact 1384
- Honda Civic sport 539
- Honda Civic old 245
- Toyota Corolla compact 451
- Toyota Corolla sport 128
- Toyota Corolla old 211



	compact	sport	old
Audi A6	67	345	56
Honda Civic	1384	539	245
Toyota Corolla	451	128	211



Explaining the Car Models Discussion



$$Lift(Y_i, Y_j) = f(\text{same_brand}, \text{same_manuf}, \text{country_of_origin}, \text{car_size}, \text{used})$$

- Observations are correlated – cannot use simple regression
- Converted $Lift(Y_i, Y_j)$ to $\log(1 + Lift(Y_i, Y_j))$
- Used Dekker Semi-Partialling QAP Regression

	Coefficient	Standardized Coefficient	Pseudo P. value
Intercept	0.2204	--	0.000
Same brand	0.3423	0.2031	0.000
Same manufacturer	0.1732	0.1708	0.000
Same country of origin	0.0903	0.1259	0.000
Same size class	0.1325	0.1826	0.100
Price Difference	-0.0054	-0.1629	0.000

Explaining the Car Models Discussion



$$\log(1 + Lift(Y_i, Y_j)) = f(\text{terms used to describe the cars})$$

- Use factor analysis to reduce the dimensionality of the terms
- Six main factors found

Upscale	Looks	Driving Experience	Customer Value	Forum Sentiment	Comfortable Ride
MSRP	Styling	Fun	Value	Fuel (neg.)	Ride
Premium	Interior	Handle/handling	Quality	Power (neg.)	Room
Bad (neg.)	Nice	Drive	Consumer	Engine (neg.)	Suspension
Luxury	Looks	Manual	Reliability	Love	Back-Seats
Navigation		Speed		Great	Leather
Sports		Automatic		Board	Seats
				Comment	

Explaining the Car Models Discussion



$$\log(1 + Lift(Y_i, Y_j)) = f(\text{terms used to describe the cars})$$

	Coefficient	Standardized Coefficient	Pseudo P. value
Intercept	0.5104	--	0.000
Occurrence	0.0003	0.0176	0.161
Factors			
Upscale	-0.0904	-0.2374	0.000
Looks	-0.0302	-0.0829	0.000
Driving experience	-0.0216	-0.0598	0.004
Consumer value	-0.0530	-0.1490	0.000
Forum sentiment	-0.0312	-0.0899	0.000
Comfortable ride	0.0022	0.0057	0.358

Explaining the Car Models Discussion



	Coefficient	Standardized Coefficient	Significance
Intercept	0.3952	--	0.000
Same brand	0.3197	0.1897	0.000
Same manufacturer	0.1778	0.1754	0.000
Same country of origin	0.0765	0.1067	0.000
Same size class	0.1332	0.1836	0.000
Price Difference	-0.0029	-0.0866	0.000
Occurrence	-0.0003	-0.0609	0.000
Factors			
Upscale	-0.0708	-0.1860	0.000
Looks	-0.0326	-0.0897	0.000
Driving experience	-0.0096	-0.0267	0.1259
Consumer value	-0.0396	-0.1114	0.000
Forum sentiment	-0.0287	-0.0827	0.000
Comfortable ride	0.0061	0.0163	0.2324

Robustness Check 1

How Much Data is Needed?



	# of messages	Correlation with the Full Dataset
1/16 of the messages	54,261	0.983
1/8 of the messages	108,52	0.989
1/4 of the messages	217,044	0.995
1/2 of the messages	434,088	0.997

Robustness Check 2

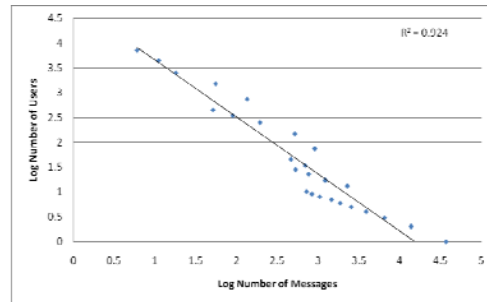
How Much Training Data is Needed?



	Recall	Precision	Overall Accuracy (F)
276 messages	91%	90%	90%
138 messages	86%	90%	87%
69 messages	84%	87%	86%
34 messages	81%	86%	83%
17 messages	73%	86%	79%

Robustness Check 3

Are all Forum Participants Equal?



- 10% of the users post over 80% of the content
- Participation in the forum follows the power law
- The correlation between “heavy” users and “light” users is $r=0.79$
- The correlation between short and long messages is $r=0.96$

Robustness Check 4

Alternative Measures of Association and Similarity



Other association and similarity measures

- *Jaccard Index*
$$Jaccard_{ij} = \frac{X_{ij}}{X_j + X_i - X_{ij}}$$
- *Cosine Similarity*
$$Cosine_{ij} = \frac{X_{ij}}{\sqrt{X_j X_i}}$$
- *TF-IDF*
$$CO(tf-idf)_{ij} = \sum_{m \in D} (tf_{jm} - idf_j \times tf_{im} - idf_i)$$

$$tf_{jm} = X_{jm} / N_m \quad idf_j = \log(D / M_j)$$
- *Pearson Correlation*
$$\rho_{ij} = \text{correl}(\mathbf{X}_i, \mathbf{X}_j)$$

Robustness Check 4

Alternative Measures of Association and Similarity

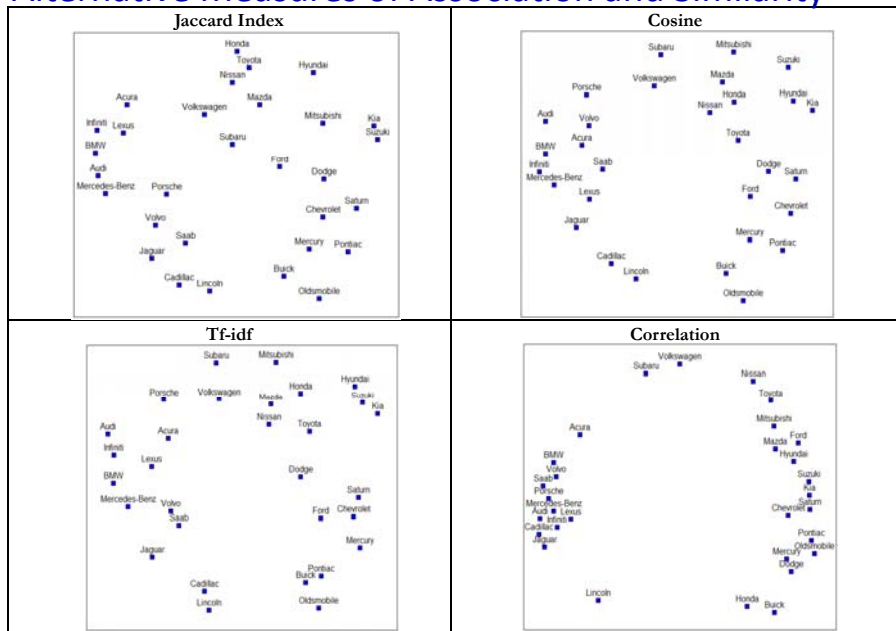


Correlations among similarity measures and with trade-ins

	Lift	Jaccard	Cosine	tf-idf	Correlation	Cars Trade-ins
Lift	--					0.753
Jaccard Index	0.970	--				0.708
Cosine	1.000	0.970	--			0.753
tf-idf	0.961	0.919	0.961	--		0.714
Correlation	0.623	0.575	0.623	0.473	--	0.578

Robustness Check 4





Alternative Measures of Association and Similarity



Pharmaceutical Application: Diabetes Drugs



Five diabetes forums

- diabetesforums.com 
- healthboards.com 
- forum.lowcarber.org 
- diabetes.blog.com
- diabetesdaily.com 



Over 670K messages

Side Effects of Diabetes Drugs



Text mining



Text mining



Drug	Effect	Lift*	WebMD	
			Frequency	Severity
Actos	Fluid retention	4.18	Infrequent	Severe
Actos	Swelling	3.55	Infrequent	Severe
Actos	Weight gain	2.98	Rare	Severe
Avandia	Edema	4.90	Rare	Severe
Avandia	Heart problems	4.45	Rare	Severe
Avandia	Swelling	4.45	Infrequent	Severe
Avandia	Fluid retention	4.17	Infrequent	Severe
Avandia	Weight gain	2.19	Rare	Severe
Byetta	Hair loss	4.20	Rare	Less severe
Byetta	Constipation	3.15	Rare	Less severe
Byetta	Nausea	3.01	Common	Less severe
Byetta	Cold symptoms	2.64	Doesn't exist	
Byetta	Fatigue	2.40	Infrequent	Less severe
Glucophage	Chest pain	13.16	Infrequent	Less severe
Glucophage	Leg pain	10.52	Infrequent	Less severe
Glucophage	Stomach cramps	4.30	Common	Less severe
Glucophage	Diarrhea	3.55	Common	Less severe
Glucophage	Lactic acid	3.40	Rare	Severe

Drug	Effect	Lift*	WebMD	
			Frequency	Severity
Humalog	Kidney problem	9.40	Doesn't exist	
Humalog	Allergic reaction	9.32	Common	Severe
Humalog	Lower blood count	7.99	Rare	severe
Humalog	Headaches	6.73	Doesn't exist	
Januvia	Irritability	11.93	Rare	Severe
Januvia	Nausea	3.62	Rare	Less severe
Lantus	Mood problem	9.72	Doesn't exist	
Lantus	Irritability	5.83	Rare	Severe
Lantus	Lower blood count	4.86	Rare	Severe
Lantus	Hypoglycemia	2.43	Common	Severe
Metformin	Muscle pain	4.12	Infrequent	Less severe
Metformin	Lactic acid	3.59	Rare	Severe
Metformin	Digestive disorders	3.09	Common	Less severe
Metformin	Diarrhea	2.68	Common	Less severe
Metformin	Leg pain	2.47	Infrequent	Less severe
Metformin	Stomach cramps	2.35	Common	Less severe
Novolog	Allergic reaction	12.40	Rare	severe
Novolog	Cold symptoms	11.05	Doesn't exist	

* Lift>1 (P.<0.05)

Drug Co-taking Analysis



Drug Co-taking with Byetta

Drug	Co-Occurrence	Lift
Januvia	67	2.96
Symlin	48	1.90
Glucovance	13	1.55
Metformin	448	1.51
Amaryl	42	1.11
Starlix	12	1.02
Glipizide	23	0.92
Actos	46	0.82
Glyburide	24	0.76
Glucophage	24	0.53
Avandia	20	0.47
Lantus	86	0.26
Humalog	28	0.15

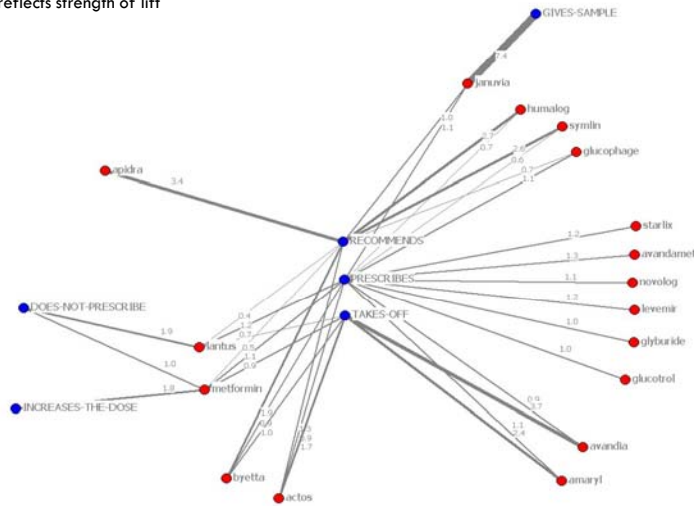
Byetta is mentioned as “taken together” more frequently than chance (lift>1): with

- Januvia
- Symlin
- Metformin (generic name)
- Amaryl
- Starlix

Exploring What is Happening at the Doctor's Office



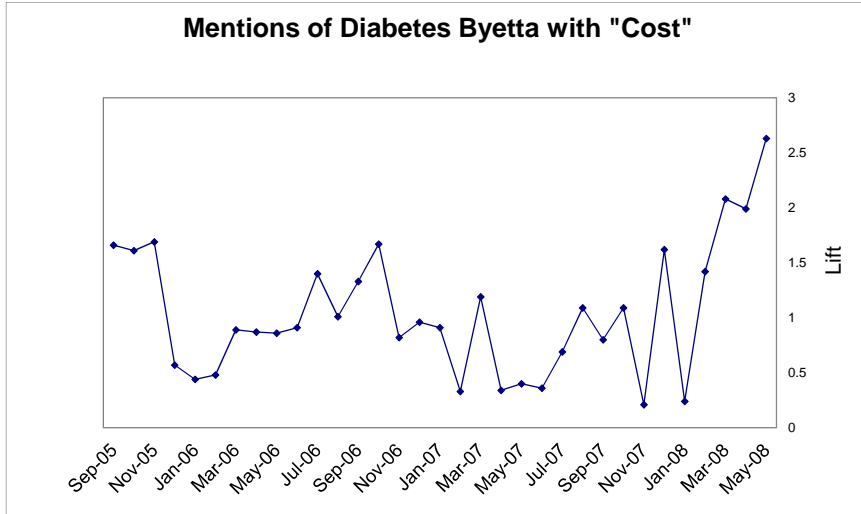
- ❖ Lifts larger than 0.5
- ❖ Width of edge reflects strength of lift



Time Trend Analysis



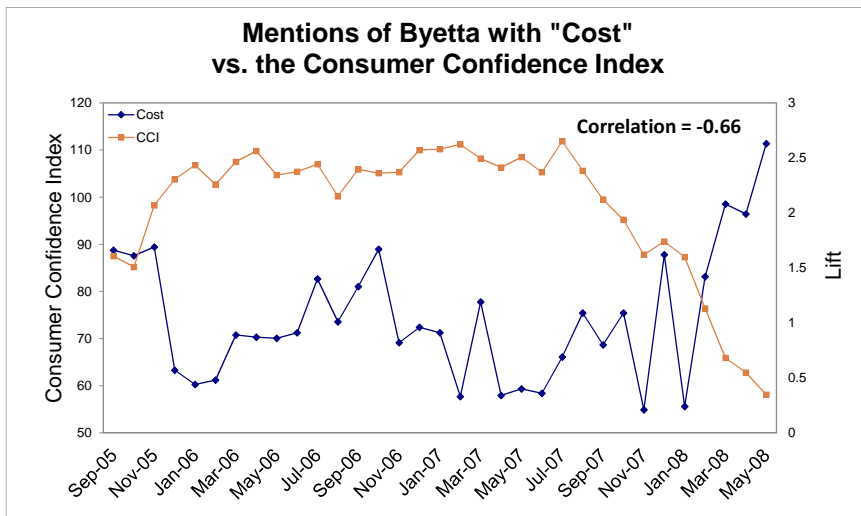
Mentions of Diabetes Byetta with "Cost"



Time Trend Analysis



Mentions of Byetta with "Cost" vs. the Consumer Confidence Index



Deriving Insights...



- ▣ Competitive landscape
- ▣ Building brand association maps
- ▣ Competitive intelligence
- ▣ Identifying customers (opinion leaders, potentially profitable, at risk)
- ▣ Brand monitoring
- ▣ “structured” exploratory research
- ▣ Tracking marketing campaign effectiveness
- ▣ Utilizing other textual information (e.g., call center)



What's Next?



Contact information: Oded Netzer
Columbia Business School
on2110@columbia.edu

