
# FROM DATA TO SOLUTIONS

# MINE YOUR OWN BUSINESS

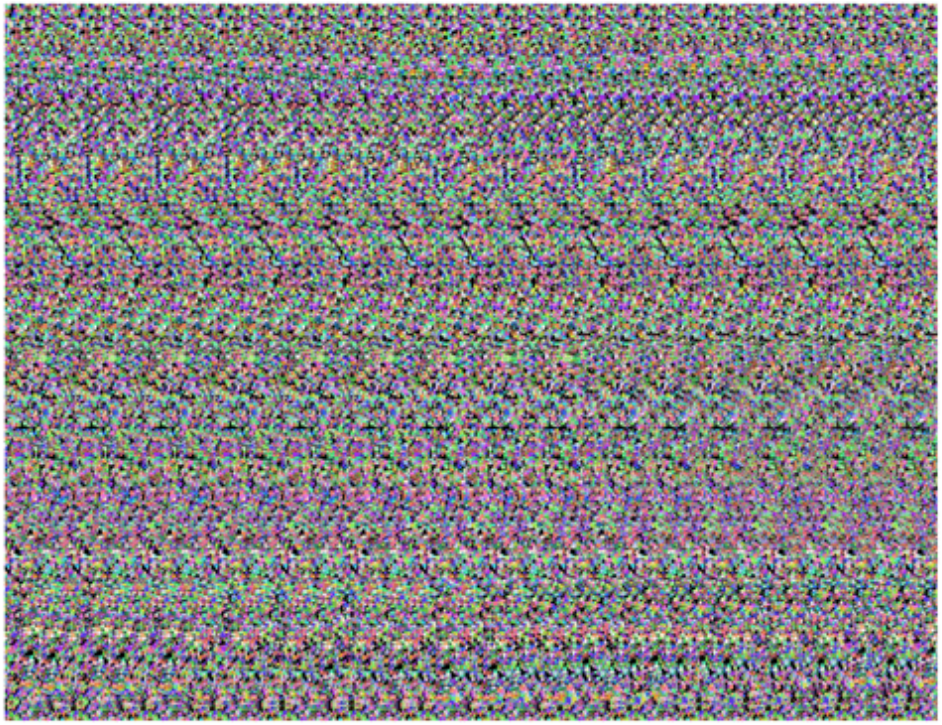
Oded Netzer

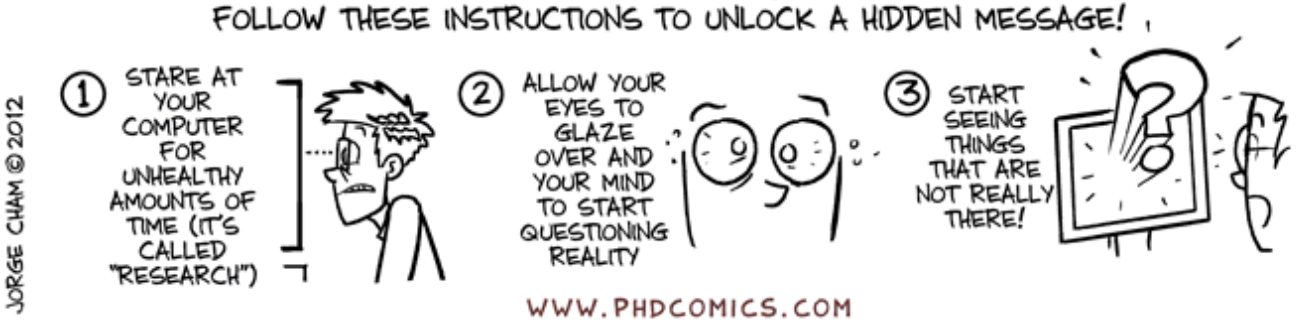
Columbia Business School

Columbia Business School

February, 2016

Columbia Business School





Columbia Business School

"Leadership is to see what everybody else has seen, and to think what nobody else has thought."

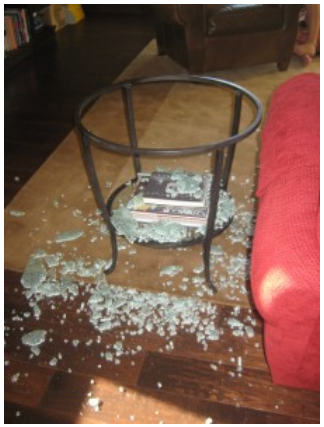
Source: Drinking from the fire hose

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

Columbia Business School

- Lots of counting, not enough evaluating
  - "If you can't measure it you can't manage it" ≠"if you can measure it you should manage it"
- Too much 'data', not enough 'insights'
- Thinking Big Data is primarily an IT challenge
- Some success stories…




# MINE YOUR OWN BUSINESS

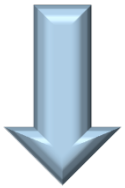
## Netzer, Feldman, Goldenberg, Fresko 2012

Columbia Business School

# The Business Problem

Columbia Business School

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



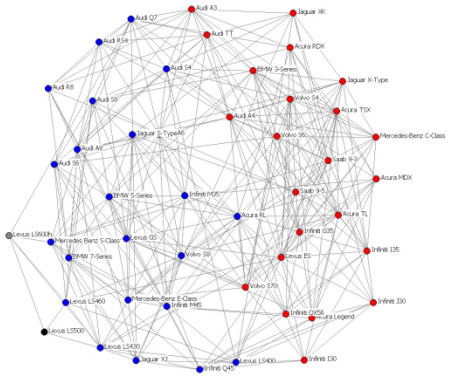
Uncovering market structure from information consumers are posting on the Web

# What Are We Going to Do?

Columbia Business School

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


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



Text Mining

Co-occurrence and Network Analysis Methods

# Mining Consumer Forums

Columbia Business School

## Opportunities

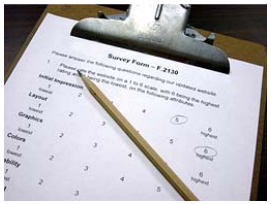
- A combination of observational and descriptive marketing research
- Permits both qualitative and quantitative information
- Non-invasive (no demand effect)
- Minimizes recall error
- Very rich data
- Sample size is not an issue
- Real time data

## 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

# Let Text Mining Do the Legwork for You

Columbia Business School



CAVEMAN | Internet | Text Mining

Find Material

Read

Understand

Consolidate

Absorb / Act

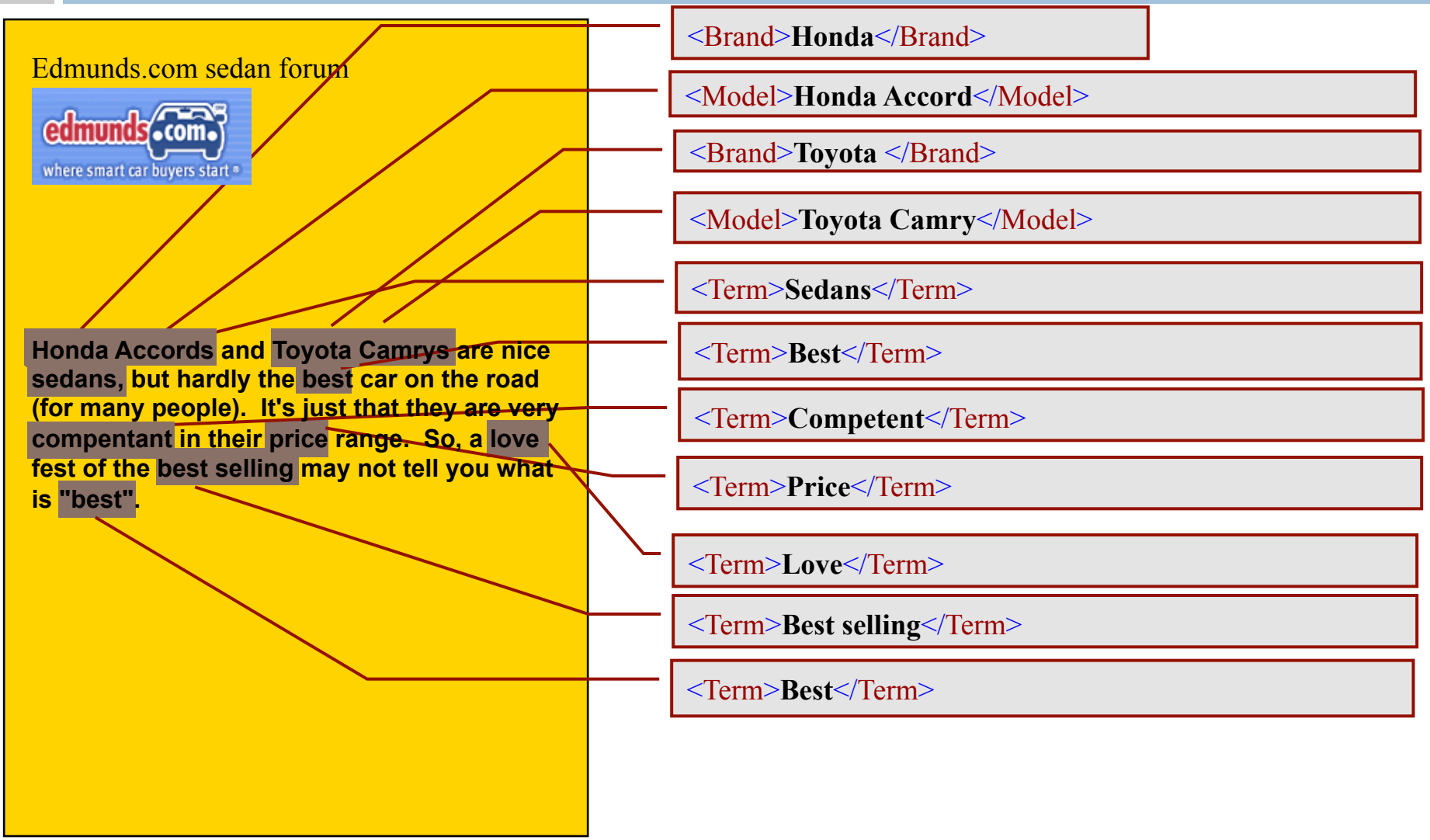
# The Text Mining Process

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

Columbia Business School



# Major Issues

Columbia Business School

- 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?

Columbia Business School

| 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

Columbia Business School

RUNNER'S WORLD

GM

edmunds.com where smart car buyers start

Consumer Driven® edmunds CARSPACE


D&B

Ford

REUTERS

STANDARD &POOR'S

# Sedan Cars Application: Edmunds.com

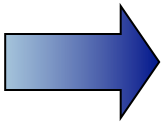
Columbia Business School


***" 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

Columbia Business School

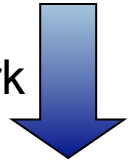
Message #1199 **Civic vs. Corolla**
by mcmanus *Jul 21, 2007 (4:05 pm)*
Yes DrFill, the Honda car model is sporty, reliable, and economical vs the **Corolla** that is just reliable and economical. Ironically its Toyota that is supplying 1.8L turbo ... Neon to his 16 year old brother. I drove it about 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 |
| Audi A6 | Toyota Corolla |
| Honda Civic | Audi 6 |
| Honda Civic | Toyota Corolla |
| Toyota Corolla | Audi A6 |
| Toyota Corolla | Honda Civic |

| | |
|---|---|
| 252 | |
| 101 | |
| 252 | |
| 2762 | |
| 101 | |
| 2762 | |

Associative Network

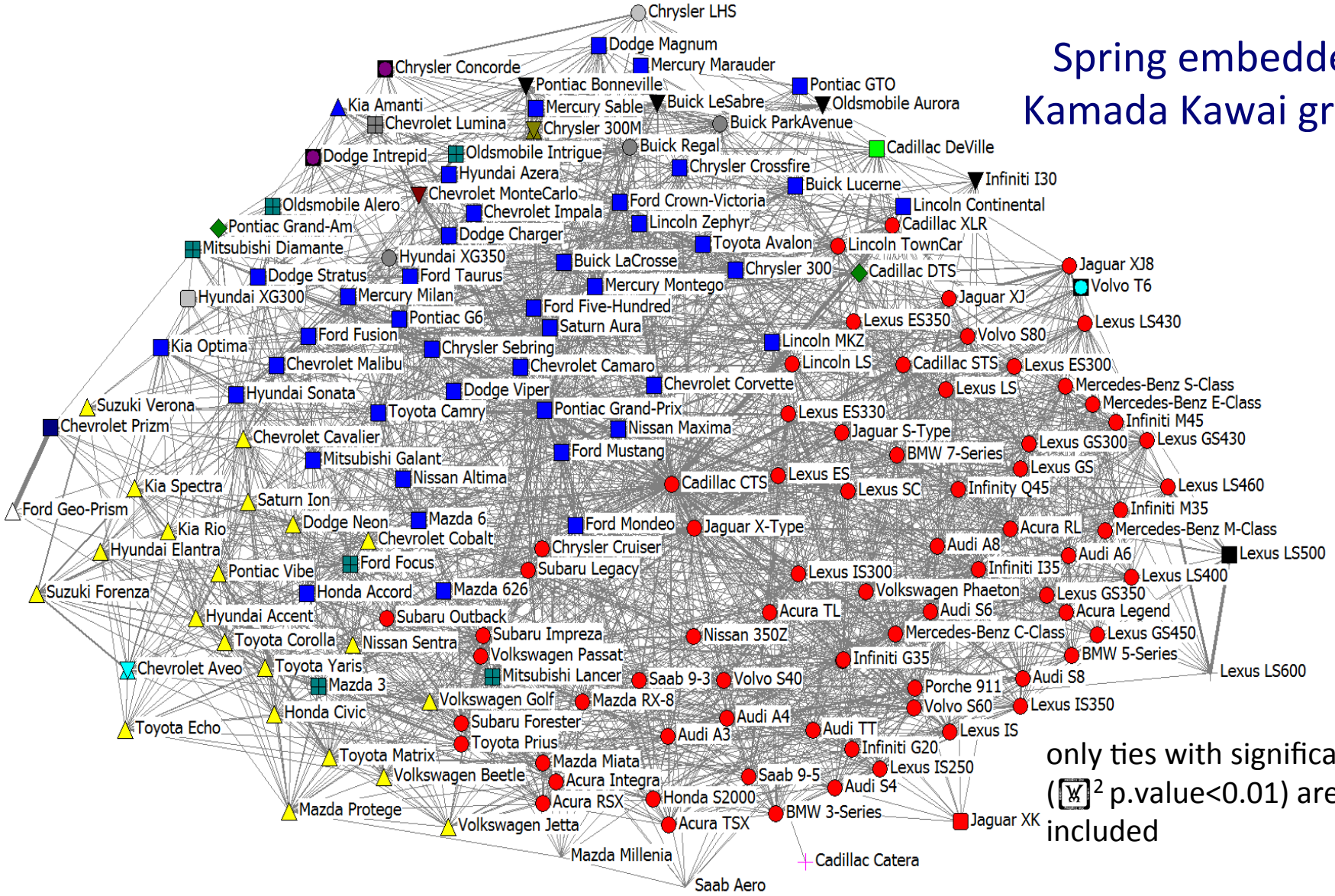


| | Audi A6 | Honda Civic | Toyota Corolla |
|---|---|---|---|
| **Audi A6** | --- | 252 | 101 |
| **Honda Civic** | 252 | --- | 2762 |
| **Toyota Corolla** | 101 | 2762 | --- |

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

# The Car Models Network

Columbia Business School

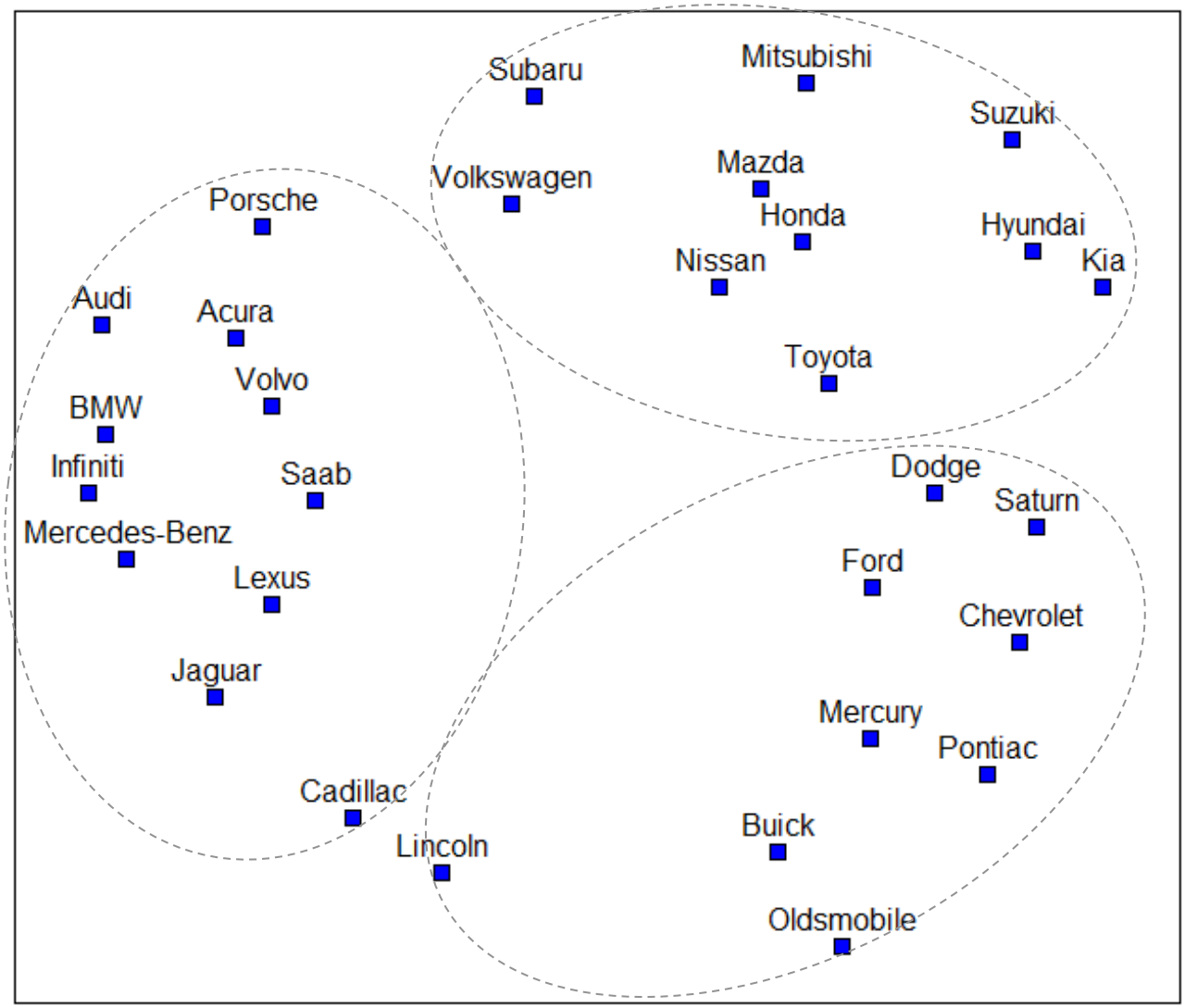



Spring embedded Kamada Kawai graph

only ties with significant lifts ($\chi^2$ p.value<0.01) are included

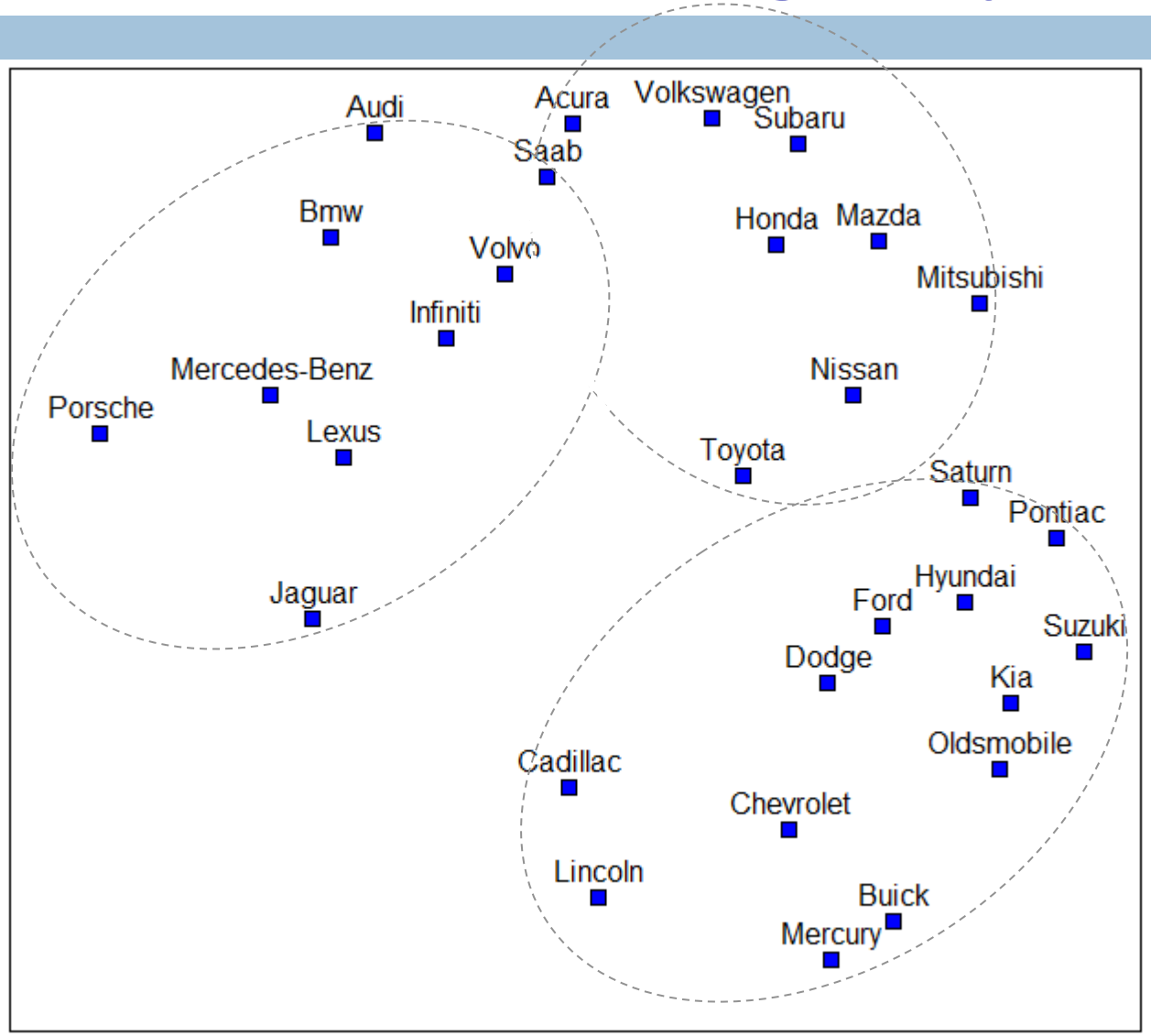
# Perceptual Map of Brands

Columbia Business School



# Brand-Switching Map

Columbia Business School

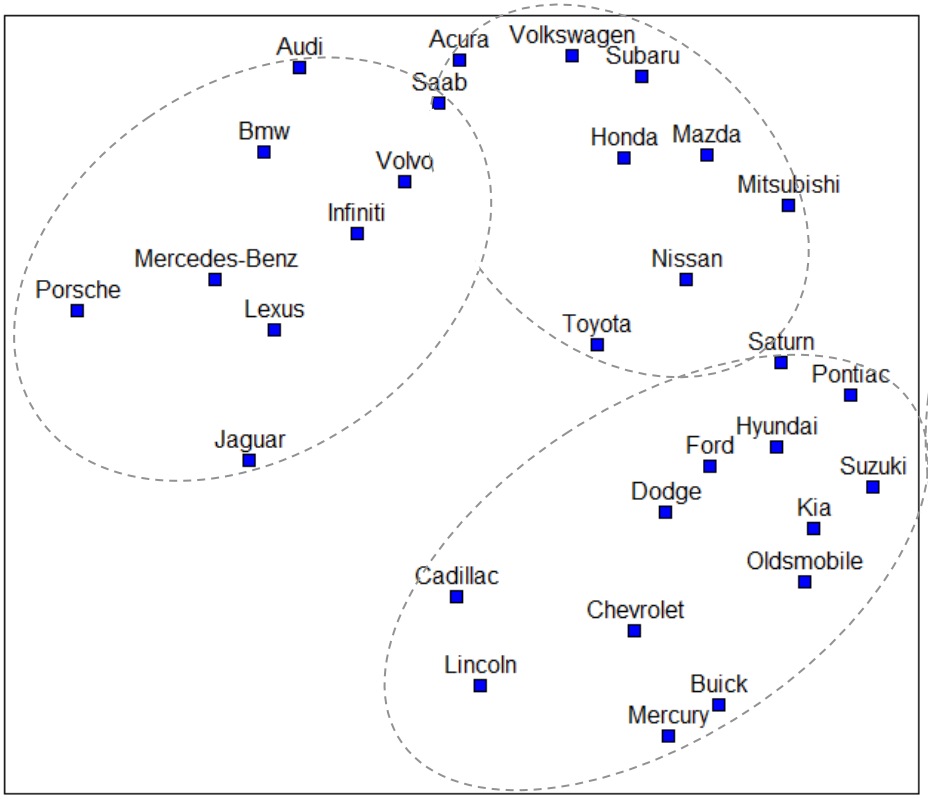



Based on JD Power PIN Data

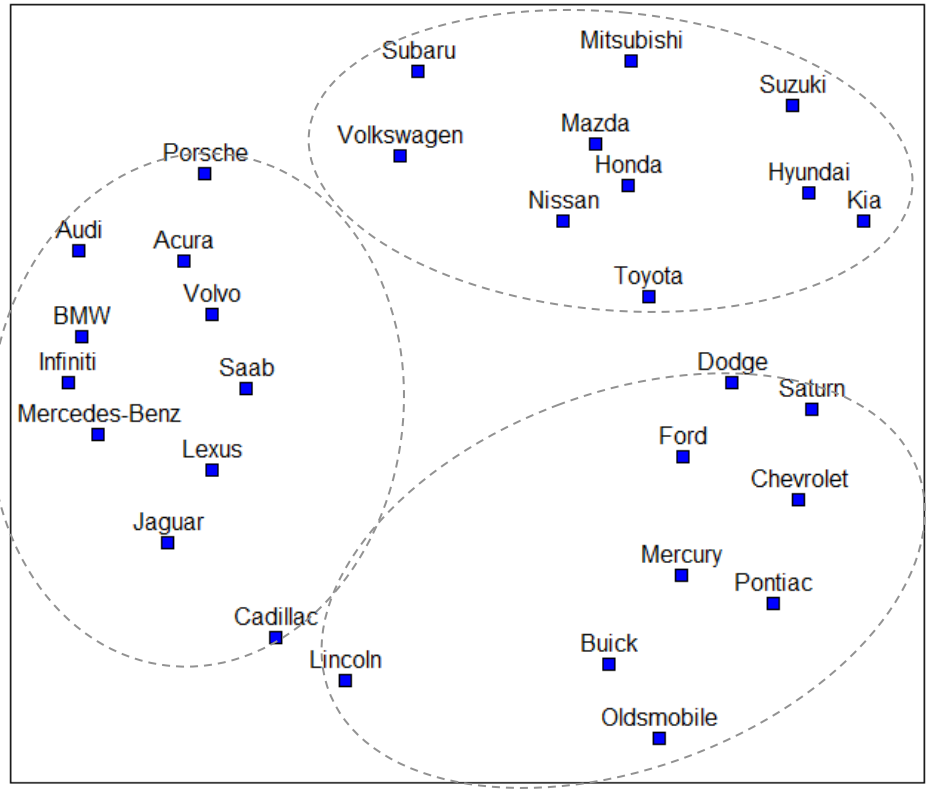
# Perceptual Map of Brands

Columbia Business School

## Brand-Switching Map




## Text Mining Map




- Similar results using survey-based consideration set data

- Correlation=0.76

# Robustness Check 1
# Alternative Measures of Association and Similarity

Columbia Business School

## Other association and similarity measures

- *Jaccard Index* $$Jaccard_{ij} = \frac{X_{ij}}{X_j + X_i - X_{ij}}$$

- *Cosine Similarity* $$\text{Cosine}_{ij} = \frac{X_{ij}}{\sqrt{X_j X_i}}$$

- *TF-IDF* $$CO(tf-idf)_{ij} = \sum_{m \in D} \left( tf_{jm} - idf_j \times tf_{im} - idf_i \right)$$

$$tf_{jm} = X_{jm} / N_m \qquad idf_j = \log(D / M_j)$$

- *Pearson Correlation* $$\rho_{ij} = correl(\mathbf{X_i}, \mathbf{X_j})$$

# Robustness Check 1
# Alternative Measures of Association and Similarity

Columbia Business School

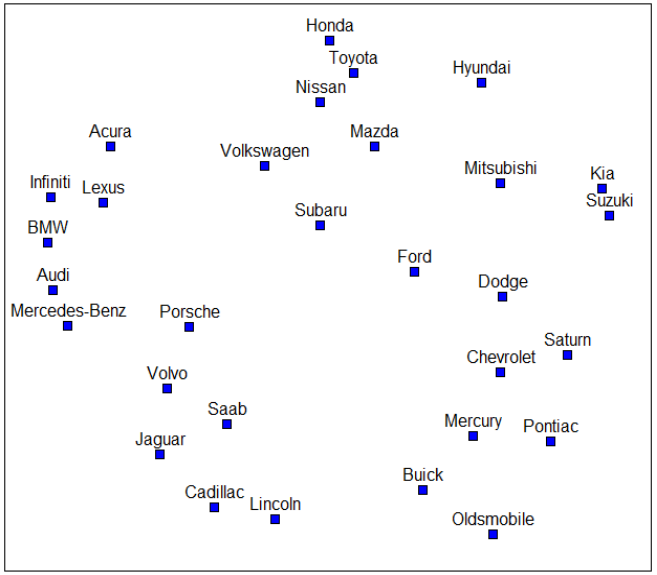
Correlations among similarity measures and with trade-ins

| | 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 |

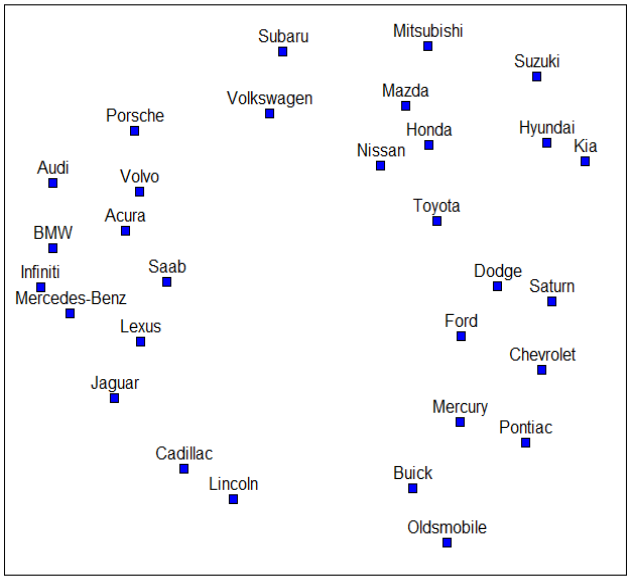
# Robustness Check 1
# Alternative Measures of Association and Similarity

Columbia Business School

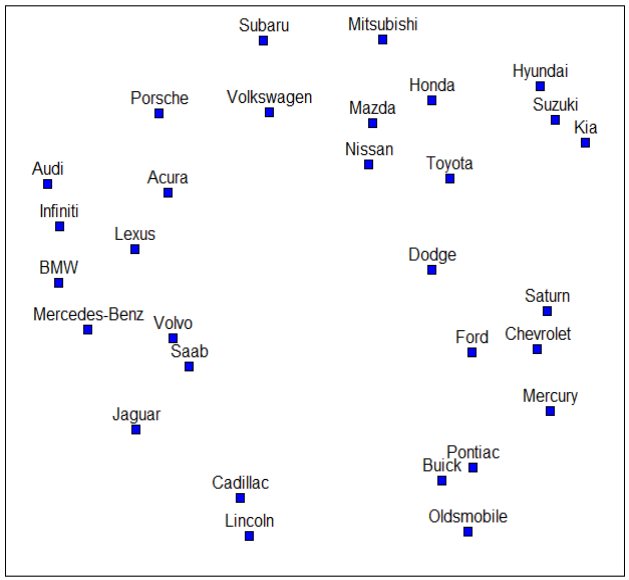
**Jaccard Index**



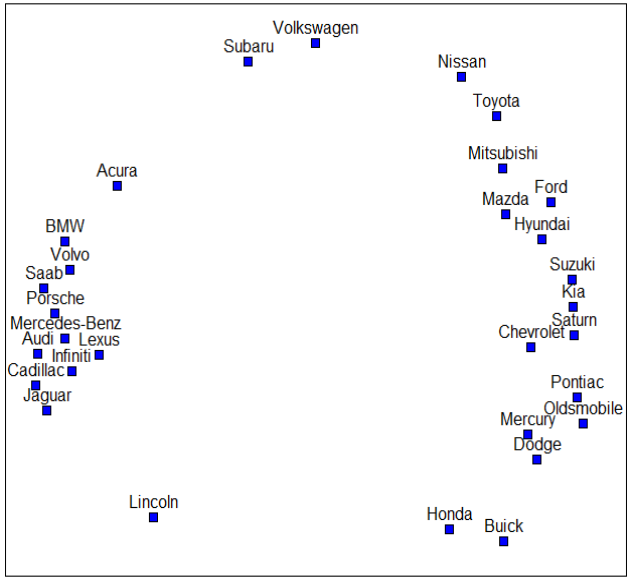

**Cosine**




**Tf-idf**




**Correlation**



# Robustness Check 2

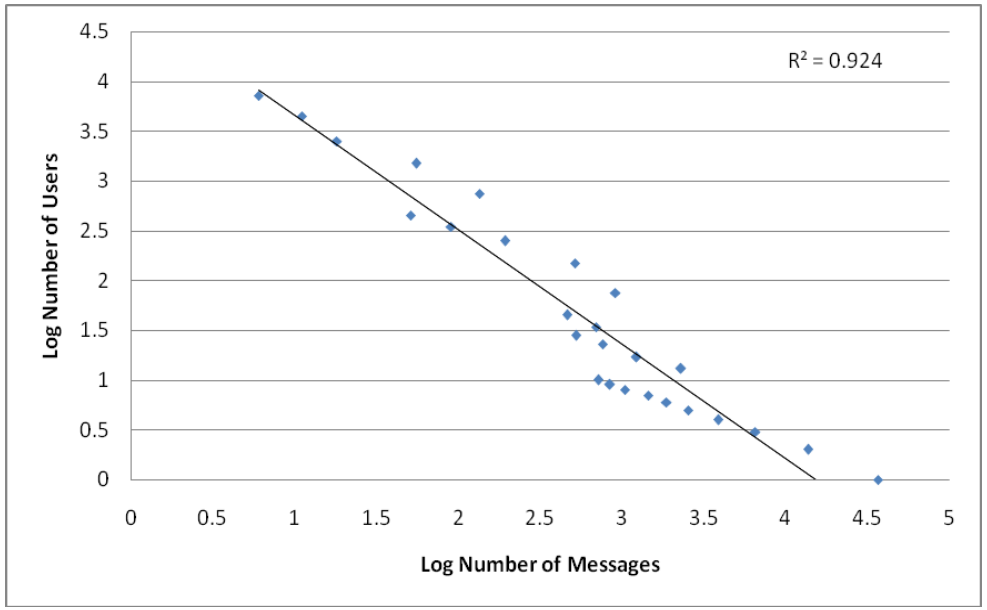
## How Much Training Data is Needed?

Columbia Business School

| | 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?

Columbia Business School




- 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

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.com

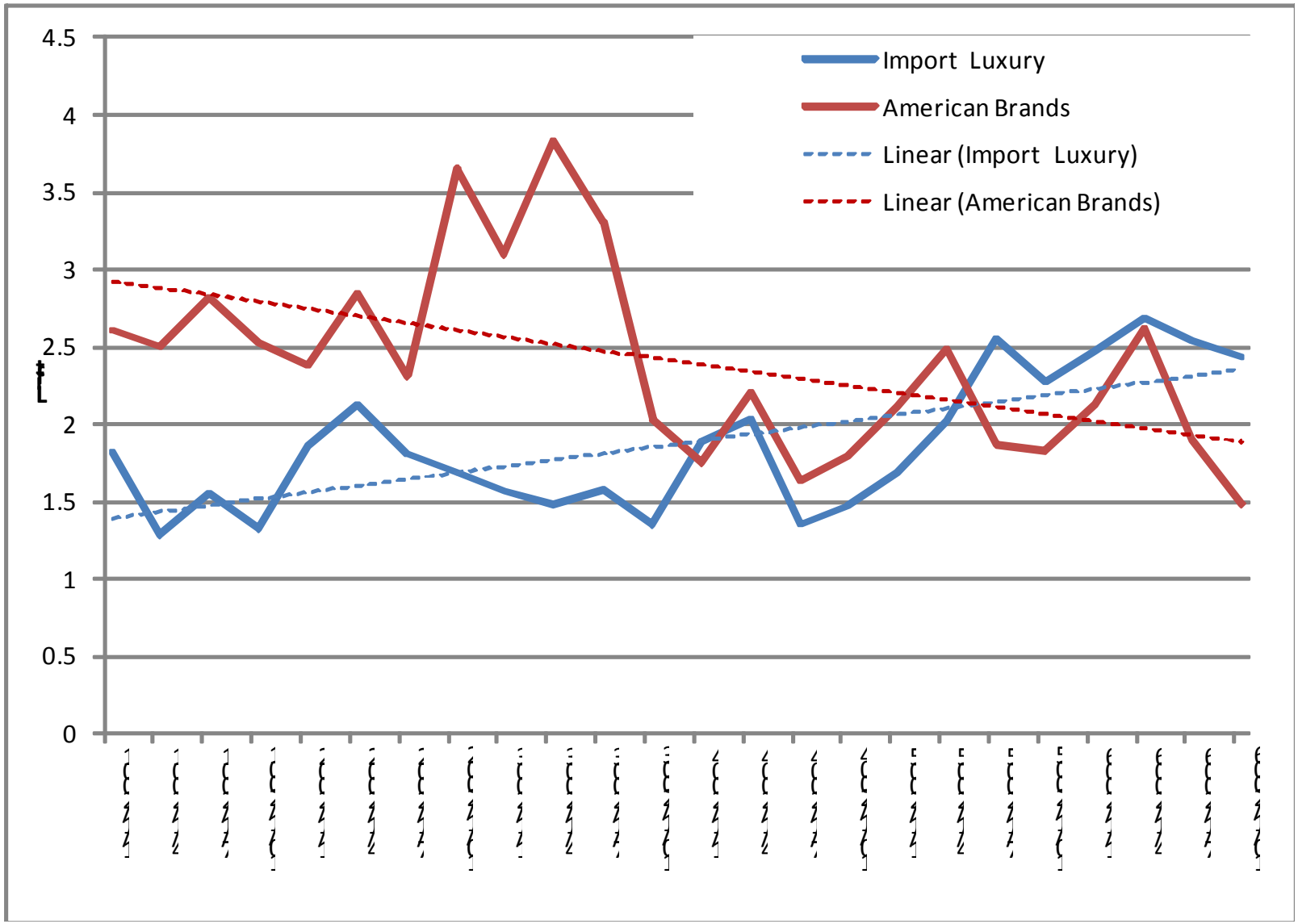
# 2004 CADILLAC XLR

## 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."

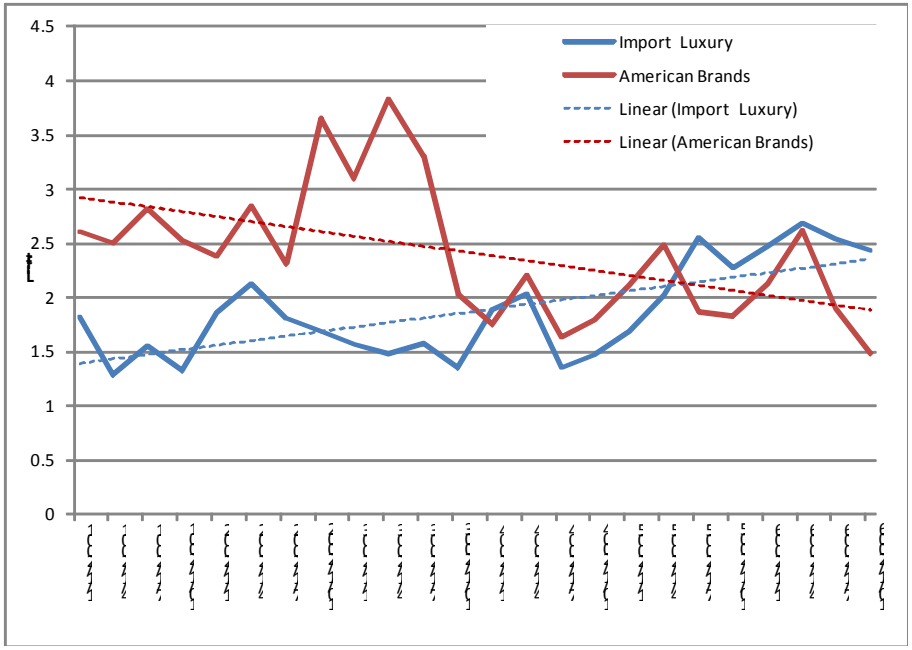
# Cadillac Positioning – Time Trend



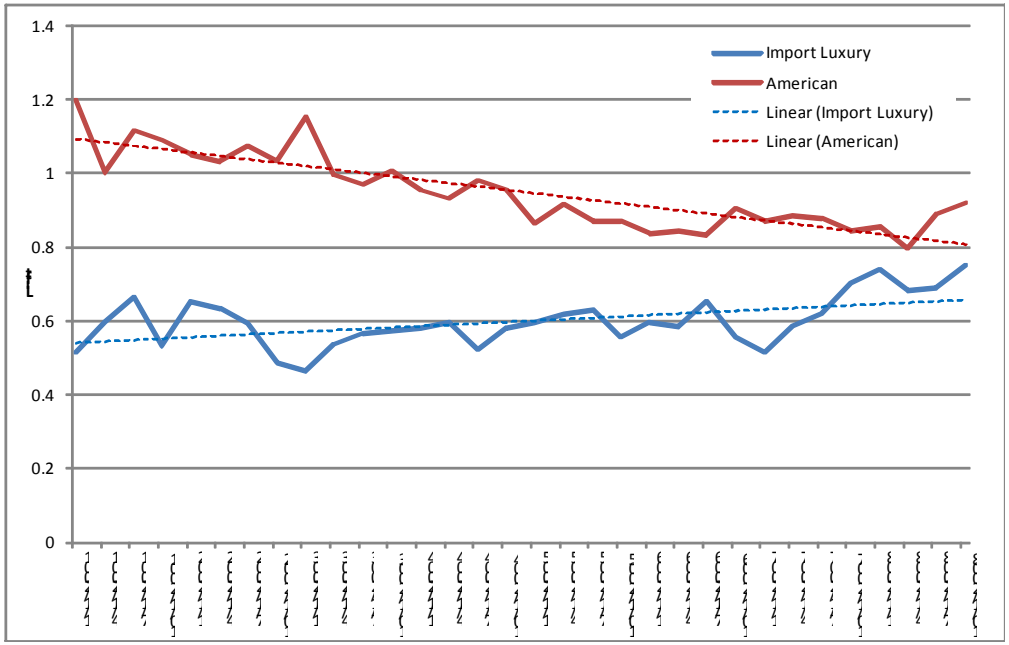
# Cadillac Positioning – TM vs. Sales

Columbia Business School

## Text-mining-based trend

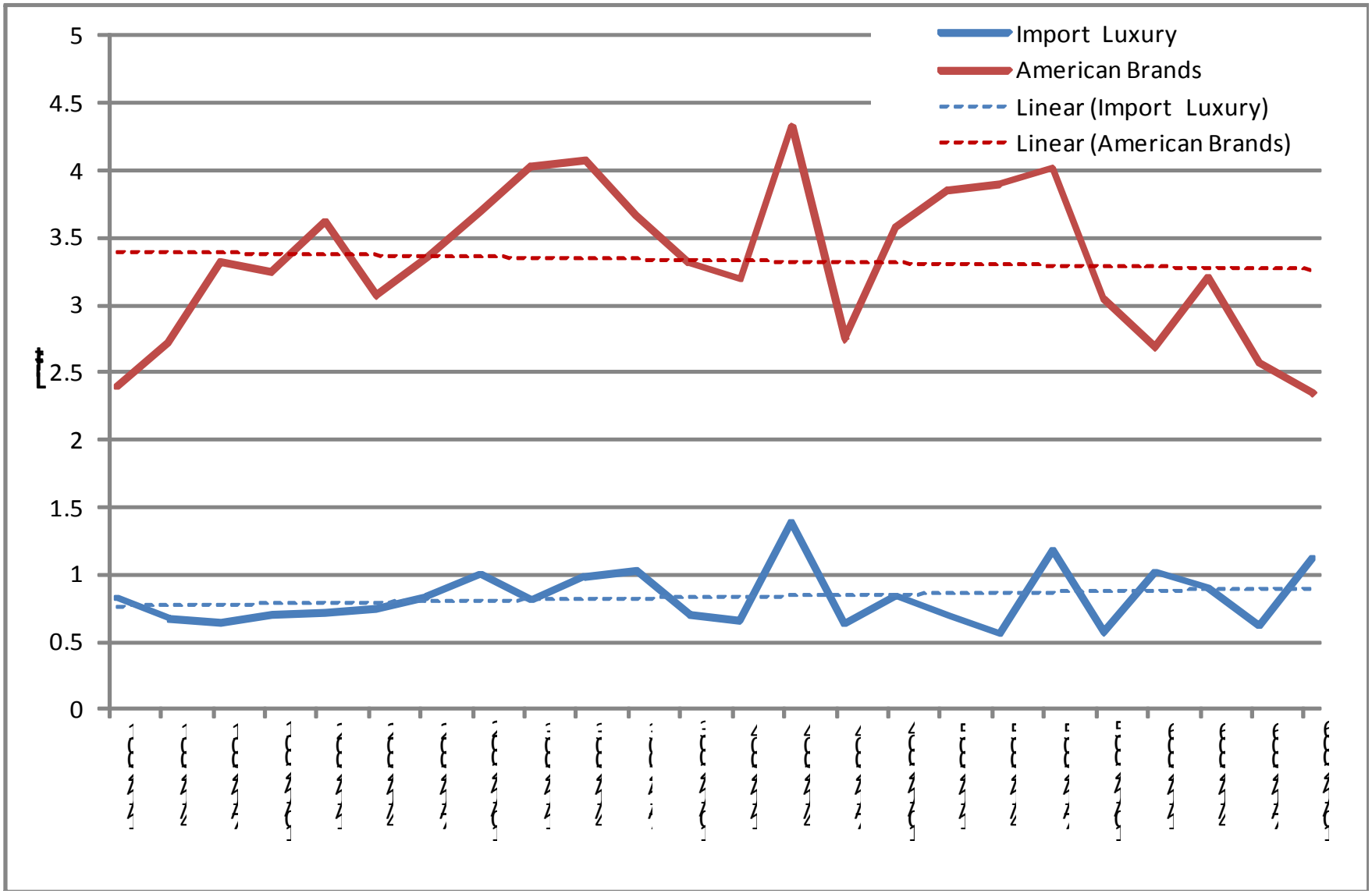



## Sales-based trend



# Buick Positioning – Time Trend

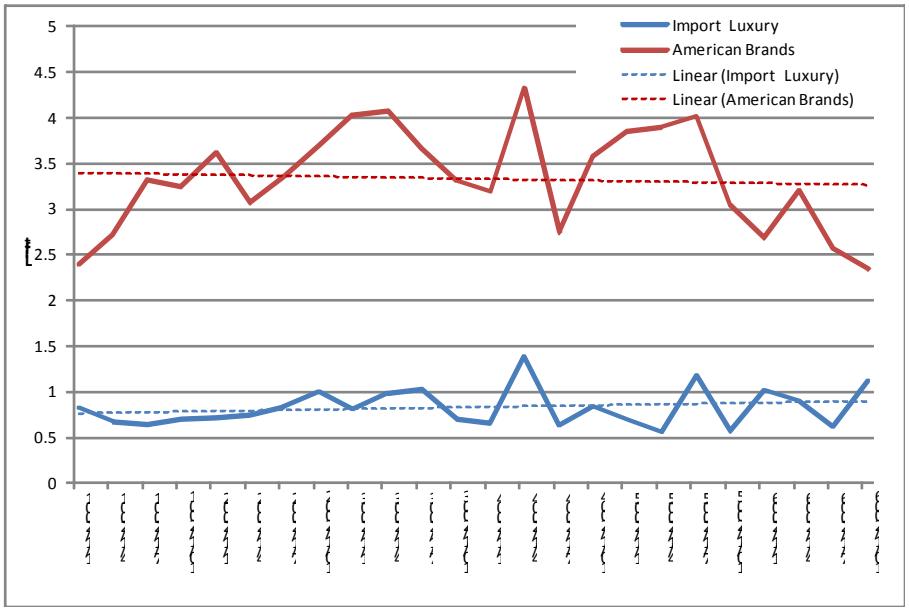
Columbia Business School



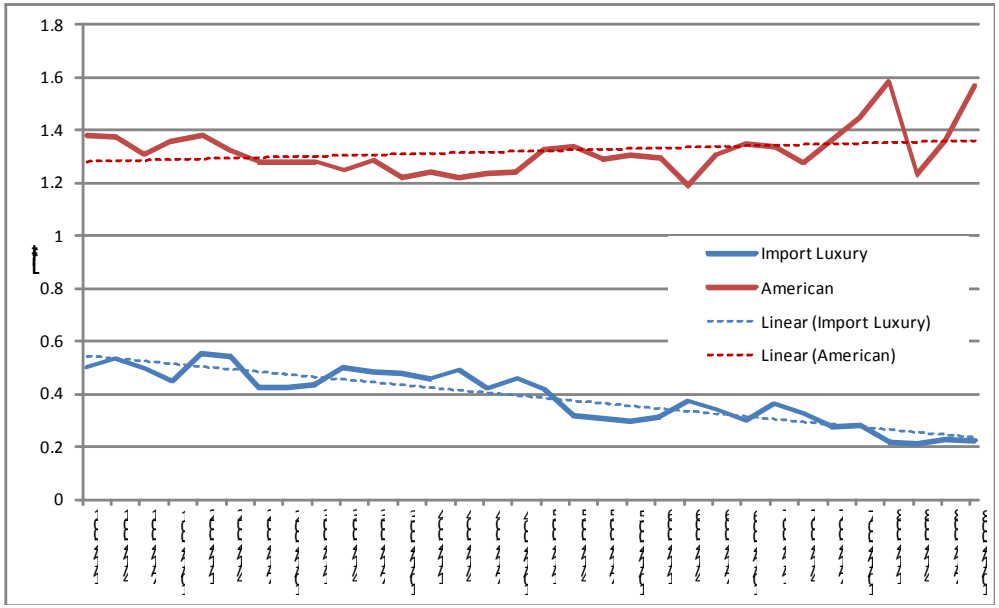
# Buick Positioning – Sales vs. TM

Columbia Business School

## Text-mining-based trend



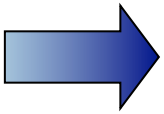

## Sales-based trend



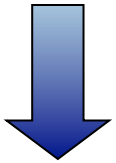
# Model-Term Data

Columbia Business School

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 **sporty** it looked. The **civic** sedan may not be as 'exciting' or '**sporty**' as the coupe, but ...

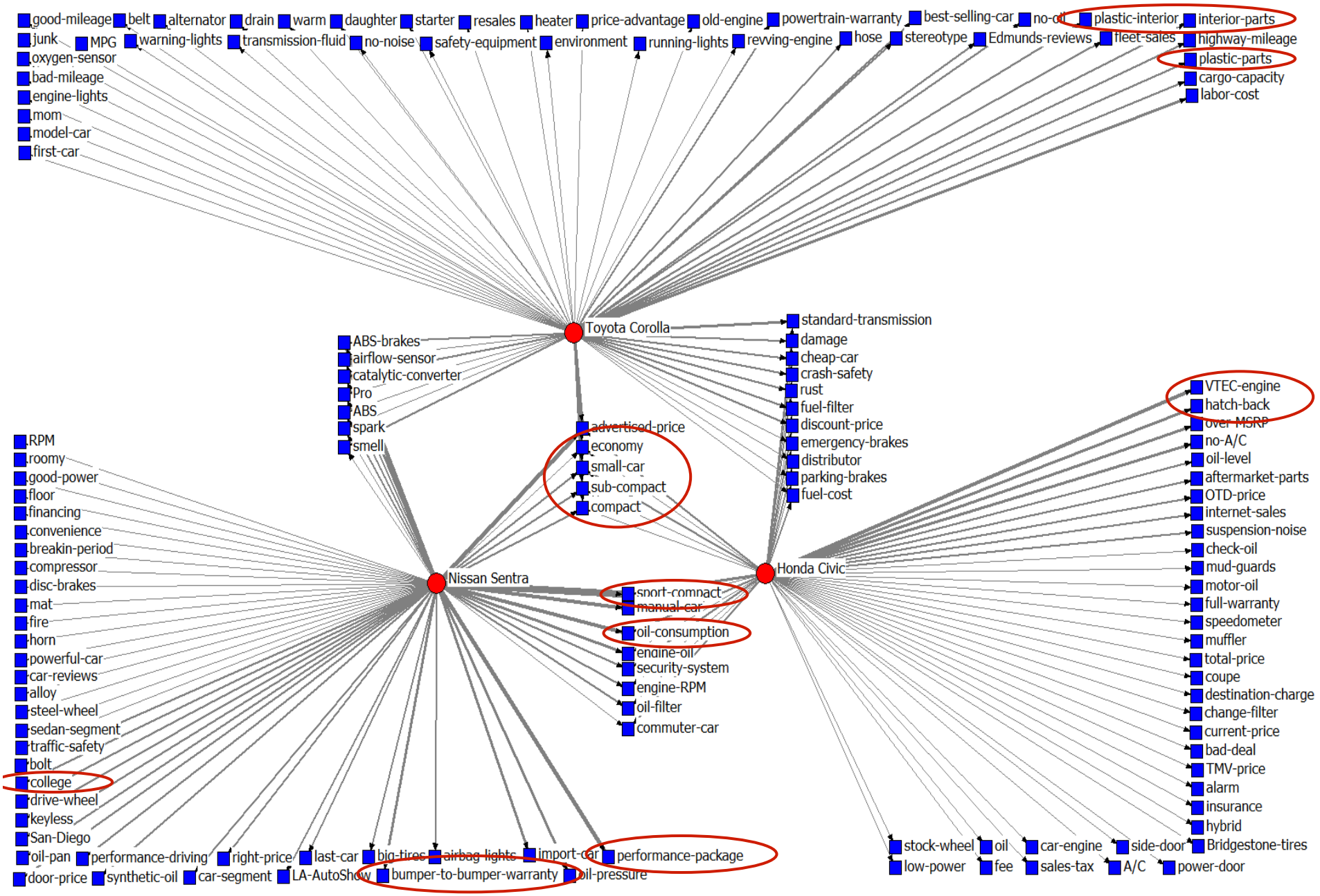


| | | |
|---|---|---|
| Audi A6 | compact | 67 |
| Audi A6 | sport | 345 |
| Audi A6 | old | 56 |
| Honda Civic | compact | 1384 |
| Honda Civic | sport | 539 |
| Honda Civic | old | 245 |
| Toyota Corolla | compact | 451 |
| Toyota Corolla | sport | 128 |
| Toyota Corolla | old | 211 |



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

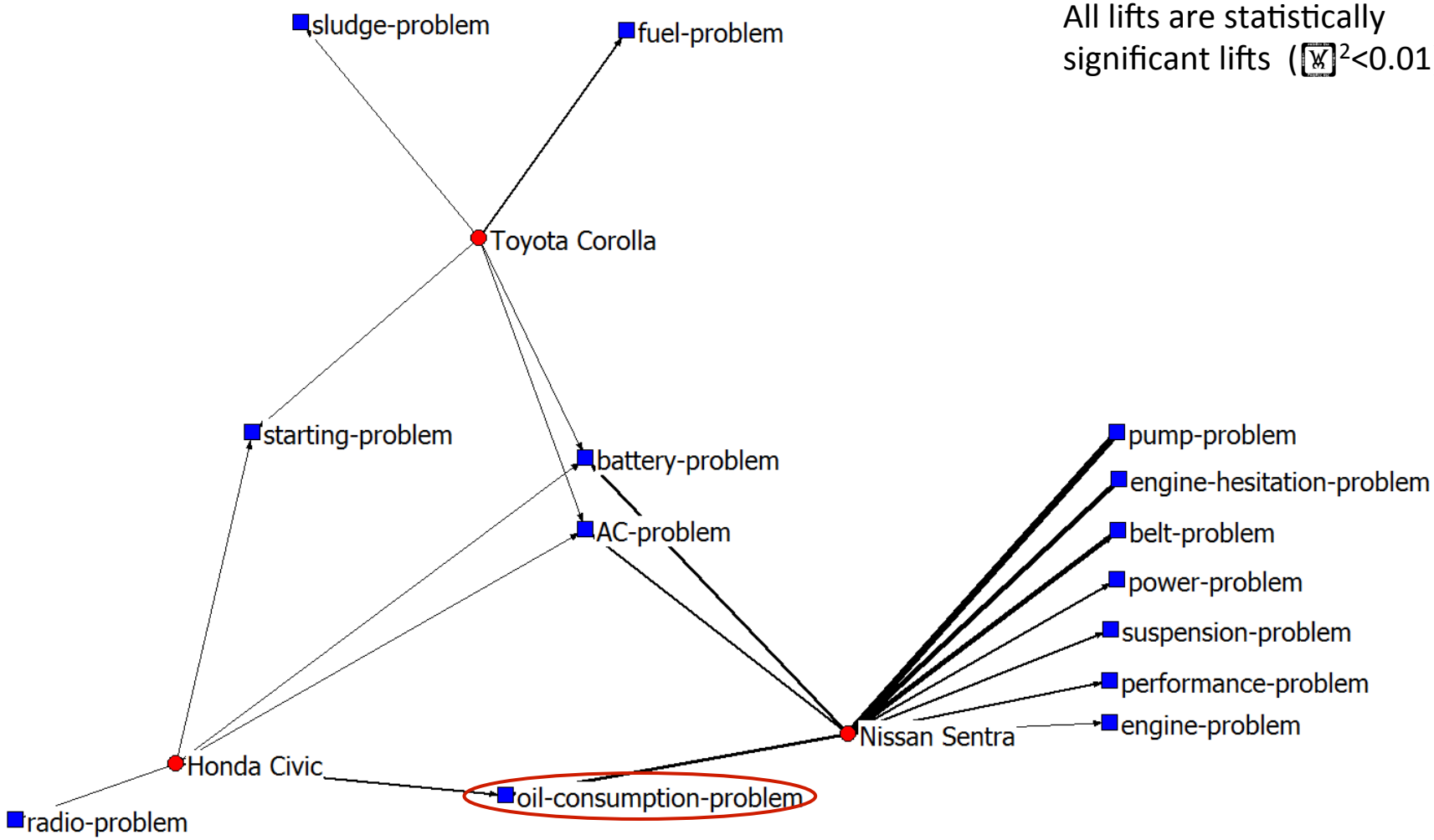
# Model-Term Analysis – 2 Mode Network



# Model-Problem Analysis – 2 Mode Network

Columbia Business School

All lifts are statistically significant lifts ($\chi^2$<0.01)




# UTILIZING TEXT MINING ON ONLINE MEDICAL FORUMS TO PREDICT LABEL CHANGE DUE TO ADVERSE DRUG REACTIONS

Feldman, Netzer, Peretz, Rosenfled 2015

Columbia Business School

# Adverse Drug Reactions

Columbia Business School

- Adverse drug events comprise the 4th leading cause of death in the US.
- Clinical trials are often limited in terms of the number of participants and scope. Accordingly, these trials sometimes fail to indicate adverse drug reactions (ADRs) associated with a particular drug.
- ADRs that were not found in the clinical trials are often reported (years) later on as FDA label changes

# Objective

Columbia Business School

Present a text mining methodology that will allow rather unlaborious, yet reliable, detection of unreported ADRs that are likely to be identified in the future.

Put our approach to a predictive test: Empirically demonstrate the ability of our proposed methodology to credibly predict ADRs prior to their reporting by the FDA.

# Methodology

Columbia Business School

- Processing
- Relation Pattern Acquisition
- Extraction
- Post-processing

# Methodology

Columbia Business School

- Processing
- Relation Pattern Acquisition
- Extraction
- Post-processing

# Relation Pattern Acquisition

Columbia Business School

- Run Unsupervised Relationship Extraction (URE) in order to learn Drug-ADR relation patterns between the following entity types: Person, Drug, Symptom and, when relevant, Disease

- Manually remove irrelevant relations

# Methodology

Columbia Business School

- Processing
- Relation Pattern Acquisition
- Extraction
- Post-processing

# Extraction

Columbia Business School

- The output of this process includes all extracted relevant entities and relations in the form of HPSG semantic structures

# Methodology

Columbia Business School

- Processing
- Relation Pattern Acquisition
- Extraction
- Post-processing

# Post Processing

Columbia Business School

- Many valuable relations that are mentioned in more indirect, elusive ways are missed.
- Try and catch those undetected relations
- For example, merging the two partial relations Person_take_Drug and Person_suffer_Symtpom:

"I took Lipitor and suffered muscle weakness and memory loss"

# Case Study 1 - Statins and Cognitive Impairment

Columbia Business School

- February 2012: FDA approves safety label changes for statins. Among others, addition of label information with regard to the potential for non-serious and reversible cognitive side effects.
  - Memory loss, forgetfulness, amnesia, memory impairment, confusion, etc.

# Case Study 1 - Statins and Cognitive Impairment

Columbia Business School

## Co-mentions up to 2011

| Top Extracted ADRs | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 | Total |
|---|---|---|---|---|---|---|---|
| pain | 12 | 4 | 12 | 20 | 11 | 453 | 512 |
| muscle pain | 9 | 0 | 14 | 37 | 9 | 374 | 443 |
| flushing | 2 | 0 | 0 | 2 | 180 | 16 | 200 |
| heart attack | 1 | 0 | 4 | 4 | 13 | 172 | 194 |
| muscle damage | 1 | 2 | 6 | 24 | 7 | 141 | 181 |
| feeling weak | 1 | 0 | 7 | 20 | 4 | 147 | 179 |
| allergic reaction | 3 | 2 | 2 | 8 | 21 | 101 | 137 |
| liver failure | 4 | 0 | 10 | 0 | 42 | 63 | 119 |
| diabetes | 6 | 2 | 2 | 5 | 17 | 78 | 110 |
| cognitive impairment | 1 | 0 | 4 | 2 | 2 | 95 | 104 |
| leg pain | 5 | 0 | 5 | 7 | 7 | 77 | 101 |
| muscle problems | 2 | 1 | 2 | 8 | 2 | 57 | 72 |
| infection | 1 | 0 | 2 | 1 | 9 | 59 | 72 |
| leg cramps | 3 | 2 | 2 | 4 | 2 | 56 | 69 |
| muscle weakness | 0 | 2 | 6 | 3 | 0 | 56 | 67 |
| cancer | 2 | 0 | 0 | 4 | 4 | 54 | 64 |
| head pain | 5 | 2 | 4 | 4 | 10 | 30 | 55 |
| heart problems | 0 | 0 | 1 | 0 | 2 | 51 | 54 |
| stroke | 2 | 0 | 2 | 3 | 1 | 42 | 50 |
| burning sensation | 1 | 0 | 0 | 0 | 4 | 38 | 43 |
| Total | 61 | 17 | 85 | 156 | 347 | 2,160 | 2,826 |

# Case Study 1 - Statins and Cognitive Impairment

Columbia Business School

## Lift up to 2011

| Relation-Driven Lift | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pain | 1.1 | 1.3 | 0.8 | 0.7 | 0.2 | 1.2 | 0.1 | 0.3 | 0.0 | 0.0 | 0.0 | 50.3 |
| muscle pain | 0.9 | 0.0 | 1.1 | 1.5 | 0.2 | 1.1 | 0.0 | 0.0 | 0.0 | 8.1 | 0.0 | 18.6 |
| flushing | 0.5 | 0.0 | 0.0 | 0.2 | 7.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 1207.0 | 0.0 |
| heart attack | 0.2 | 0.0 | 0.7 | 0.4 | 0.5 | 1.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 17.3 |
| muscle damage | 0.3 | 1.8 | 1.1 | 2.4 | 0.3 | 1.0 | 0.0 | 0.8 | 0.1 | 22.2 | 0.0 | 0.2 |
| feeling weak | 0.3 | 0.0 | 1.3 | 2.0 | 0.2 | 1.1 | 0.0 | 0.0 | 0.5 | 11.7 | 0.0 | 3.4 |
| allergic reaction | 1.0 | 2.4 | 0.5 | 1.1 | 1.2 | 1.0 | 0.0 | 1.8 | 0.0 | 0.0 | 1.2 | 0.0 |
| liver failure | 1.6 | 0.0 | 2.8 | 0.0 | 2.9 | 0.7 | 0.9 | 0.0 | 12.4 | 0.0 | 61.1 | 0.0 |
| diabetes | 2.5 | 3.0 | 0.6 | 0.8 | 1.3 | 0.9 | 5.9 | 2.8 | 0.0 | 0.0 | 1.1 | 0.0 |
| cognitive impairment | 0.4 | 0.0 | 1.3 | 0.3 | 0.2 | 1.2 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | 13.3 |
| leg pain | 2.3 | 0.0 | 1.6 | 1.3 | 0.6 | 1.0 | 3.9 | 0.0 | 1.4 | 0.4 | 0.0 | 0.0 |
| muscle problems | 1.3 | 2.3 | 0.9 | 2.0 | 0.2 | 1.0 | 0.1 | 0.8 | 0.0 | 4.4 | 0.0 | 0.3 |
| infection | 0.6 | 0.0 | 0.9 | 0.3 | 1.0 | 1.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.2 |
| leg cramps | 2.0 | 4.8 | 1.0 | 1.1 | 0.2 | 1.1 | 1.6 | 6.2 | 0.0 | 0.0 | 0.0 | 0.9 |
| muscle weakness | 0.0 | 5.0 | 3.0 | 0.8 | 0.0 | 1.1 | 0.0 | 6.5 | 8.3 | 0.0 | 0.0 | 1.9 |

| Critical Values | | |
|---|---|---|
| lift | chi-square value | p-value |
| 1.00 | 3.84 | 0.05 |
| | 6.64 | 0.01 |

# Case Study 1 - Statins and Cognitive Impairment

Columbia Business School

- How early could have detected?

| Year | Relation-driven lift | Chi-square value | Classic-induced lift | Chi-square value |
|---|---|---|---|---|
| 2011 | 1.20 | **13.33** | 1.99 | **49.28** |
| 2010 | 1.21 | **13.24** | 1.94 | **42.21** |
| 2009 | 1.22 | **13.35** | 1.97 | **40.03** |
| 2008 | 1.21 | **10.70** | 1.89 | **31.42** |
| 2007 | 1.20 | **9.95** | 2.00 | **36.46** |
| 2006 | 1.21 | **10.30** | 1.89 | **28.20** |
| 2005 | 1.20 | **6.63** | 2.04 | **25.12** |
| 2004 | 1.25 | 3.46 | 2.18 | **12.93** |
| 2003 | 1.27 | 1.55 | 2.16 | **5.79** |

- All values in bold are chi-square values ≥ 3.85 or ≥ 6 .64, corresponding p-value ≤ 0 .05 or ≤ 0 .01.

# Case Study 2 - Wellbutrin and Agitation

Columbia Business School

- July 2009: FDA alert informing that manufacturers of Wellbutrin were required to add new boxed warnings highlighting the risk of serious neuropsychiatric symptoms in patients using the drug.
- Among those symptoms were agitation and hostility.

# Case Study 2 - Wellbutrin and Agitation

Columbia Business School

## Co-mentions 1999-2008

| Top Extracted ADRs | celexa | effexor | pristiq | wellbutrin | xanax | zoloft | Total |
|---|---|---|---|---|---|---|---|
| anxiety | 217 | 359 | 3 | 343 | 423 | 492 | 1,837 |
| weight gain | 71 | 144 | 1 | 130 | 11 | 169 | 526 |
| head pain | 56 | 127 | 1 | 100 | 21 | 102 | 407 |
| panic state | 44 | 70 | 0 | 19 | 133 | 96 | 362 |
| sleep disorder | 43 | 78 | 1 | 86 | 49 | 104 | 361 |
| allergic reaction | 46 | 69 | 0 | 67 | 23 | 98 | 303 |
| feeling weak | 40 | 74 | 3 | 48 | 15 | 60 | 240 |
| pain | 21 | 74 | 0 | 38 | 30 | 60 | 223 |
| tremors | 29 | 59 | 0 | 30 | 43 | 57 | 218 |
| agitation | 24 | 46 | 0 | 77 | 11 | 54 | 212 |
| nausea | 22 | 89 | 2 | 33 | 7 | 51 | 204 |
| sweating | 22 | 103 | 1 | 21 | 5 | 42 | 194 |
| seizure | 5 | 16 | 0 | 113 | 40 | 16 | 190 |
| dizziness | 18 | 66 | 1 | 25 | 9 | 50 | 169 |
| suicidality | 15 | 67 | 0 | 18 | 9 | 36 | 145 |
| sexual dysfunction | 21 | 37 | 0 | 46 | 2 | 28 | 134 |
| cognitive impairment | 10 | 41 | 1 | 17 | 23 | 21 | 113 |
| weight loss | 5 | 23 | 0 | 61 | 2 | 18 | 109 |
| mood swings | 18 | 36 | 0 | 21 | 3 | 27 | 105 |
| sleepiness | 26 | 32 | 0 | 17 | 5 | 17 | 97 |
| Total | 753 | 1,610 | 14 | 1,310 | 864 | 1,598 | 6,149 |

# Case Study 2 - Wellbutrin and Agitation

Columbia Business School

## Lift up to 2011

| Relation-Driven Lift | C | E | P | W | X | Z |
|---|---|---|---|---|---|---|
| anxiety | 1.0 | 0.7 | 0.7 | 0.9 | 1.6 | 1.0 |
| weight gain | 1.1 | 1.0 | 0.8 | 1.2 | 0.1 | 1.2 |
| head pain | 1.1 | 1.2 | 1.1 | 1.2 | 0.4 | 1.0 |
| panic state | 1.0 | 0.7 | 0.0 | 0.2 | 2.6 | 1.0 |
| sleep disorder | 1.0 | 0.8 | 1.2 | 1.1 | 1.0 | 1.1 |
| allergic reaction | 1.2 | 0.9 | 0.0 | 1.0 | 0.5 | 1.2 |
| feeling weak | 1.4 | 1.2 | 5.5 | 0.9 | 0.4 | 1.0 |
| pain | 0.8 | 1.3 | 0.0 | 0.8 | 1.0 | 1.0 |
| tremors | 1.1 | 1.0 | 0.0 | 0.6 | 1.4 | 1.0 |
| agitation | 0.9 | 0.8 | 0.0 | 1.7 | 0.4 | 1.0 |
| nausea | 0.9 | 1.7 | 4.3 | 0.8 | 0.2 | 1.0 |
| sweating | 0.9 | 2.0 | 2.3 | 0.5 | 0.2 | 0.8 |
| seizure | 0.2 | 0.3 | 0.0 | 2.8 | 1.5 | 0.3 |
| dizziness | 0.9 | 1.5 | 2.6 | 0.7 | 0.4 | 1.1 |
| suicidality | 0.8 | 1.8 | 0.0 | 0.6 | 0.4 | 1.0 |

| C | E | P | W | X | Z |
|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.0 | 0.0 | 174.7 | 0.9 |
| 0.8 | 0.4 | 0.0 | 4.0 | 0.0 | 11.3 |
| 0.9 | 5.7 | 0.0 | 2.8 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 164.0 | 0.1 |
| 0.0 | 0.0 | 0.0 | 1.5 | 0.0 | 1.6 |
| 2.6 | 0.0 | 0.0 | 0.1 | 0.0 | 6.7 |
| 4.5 | 2.8 | 11.5 | 0.0 | 0.0 | 0.0 |
| 0.0 | 5.9 | 0.0 | 0.0 | 0.0 | 0.1 |
| 0.2 | 0.1 | 0.0 | 0.0 | 6.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 29.5 | 0.0 | 0.0 |
| 0.0 | 33.2 | 5.3 | 0.0 | 0.0 | 0.0 |
| 0.0 | 75.1 | 0.7 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 170.4 | 8.0 | 0.0 |
| 0.0 | 14.9 | 1.0 | 0.0 | 0.0 | 1.2 |
| 0.0 | 30.8 | 0.0 | 0.0 | 0.0 | 0.0 |

| Critical Values | | |
|---|---|---|
| lift | chi-square value | p-value |
| 1.00 | 3.84 | 0.05 |
| | 6.64 | 0.01 |

# Case Study 2 - Wellbutrin and Agitation

Columbia Business School

- How early could have detected?

| Year | Relation-driven lift | Chi-square value | Classic-induced lift | Chi-square value |
|---|---|---|---|---|
| 2008 | 1.70 | **29.53** | 1.81 | **28.79** |
| 2007 | 1.75 | **28.70** | 1.83 | **26.97** |
| 2006 | 1.72 | **21.13** | 1.61 | **12.91** |
| 2005 | 1.69 | **16.54** | 1.78 | **15.90** |
| 2004 | 1.46 | **5.13** | 1.51 | **5.13** |
| 2003 | 1.64 | **8.03** | 2.05 | **13.99** |
| 2002 | 1.78 | **6.41** | 2.36 | **12.25** |

- All values in bold are chi-square values ≥ 3.85 or ≥ 6 .64, corresponding p-value ≤ 0 .05 or ≤ 0 .01.

# Summary

Columbia Business School

- We are facing a river flow of information
- We need to develop fishing rods to fish the river for insights
- Text mining tools are constantly advancing
- Complex textual relationships require more advanced text mining skills/tools
- Social media and other sources of textual data:
  - are VERY large and messy
  - keep coming in real time
  - can be extremely useful if we learn how to listen…

# Deriving Insights…

Columbia Business School

- 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?

Columbia Business School

Columbia Business School

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