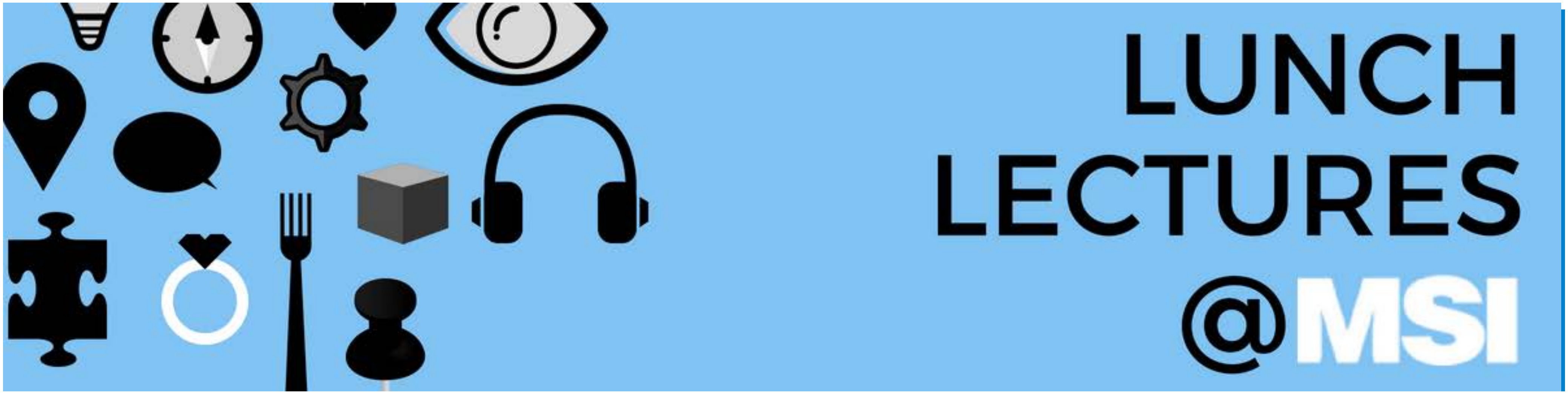
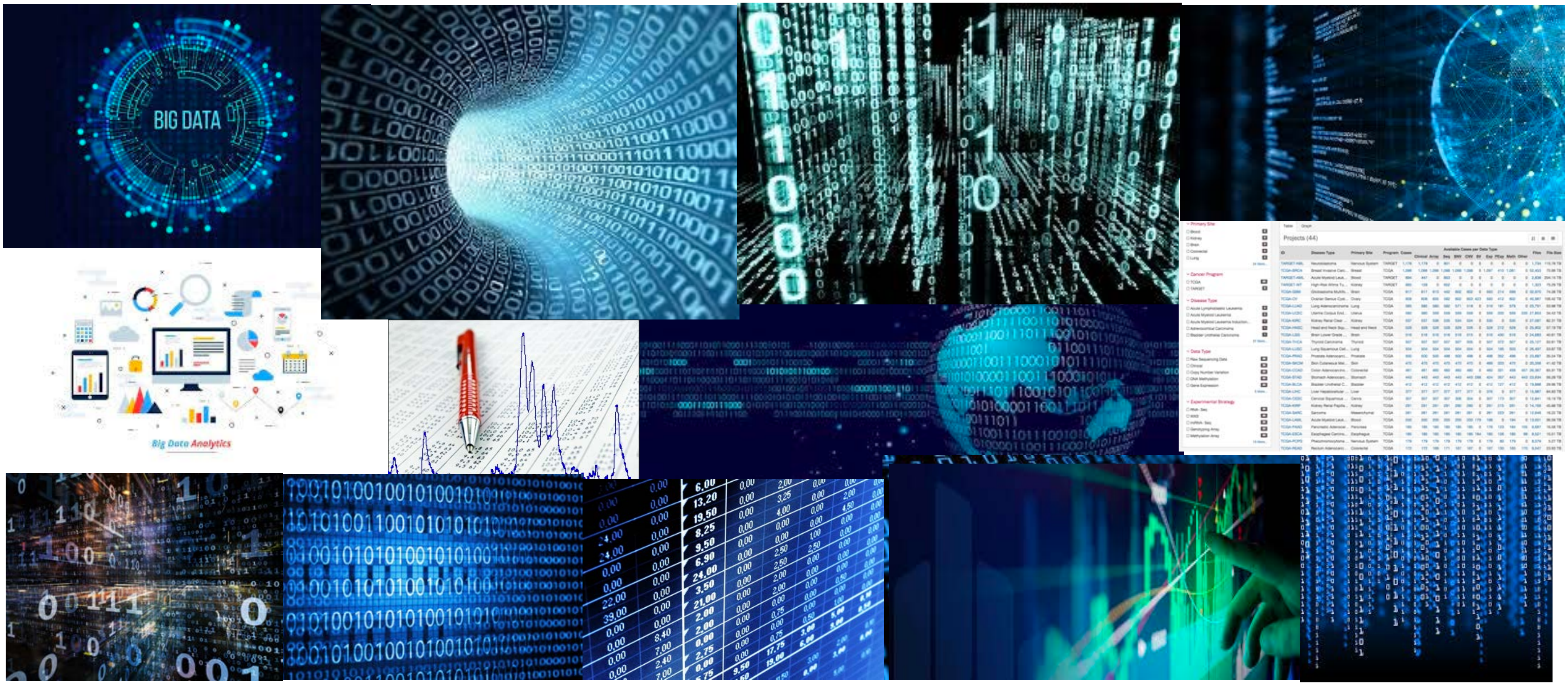
Columbia Business School
AT THE VERY CENTER OF BUSINESS™

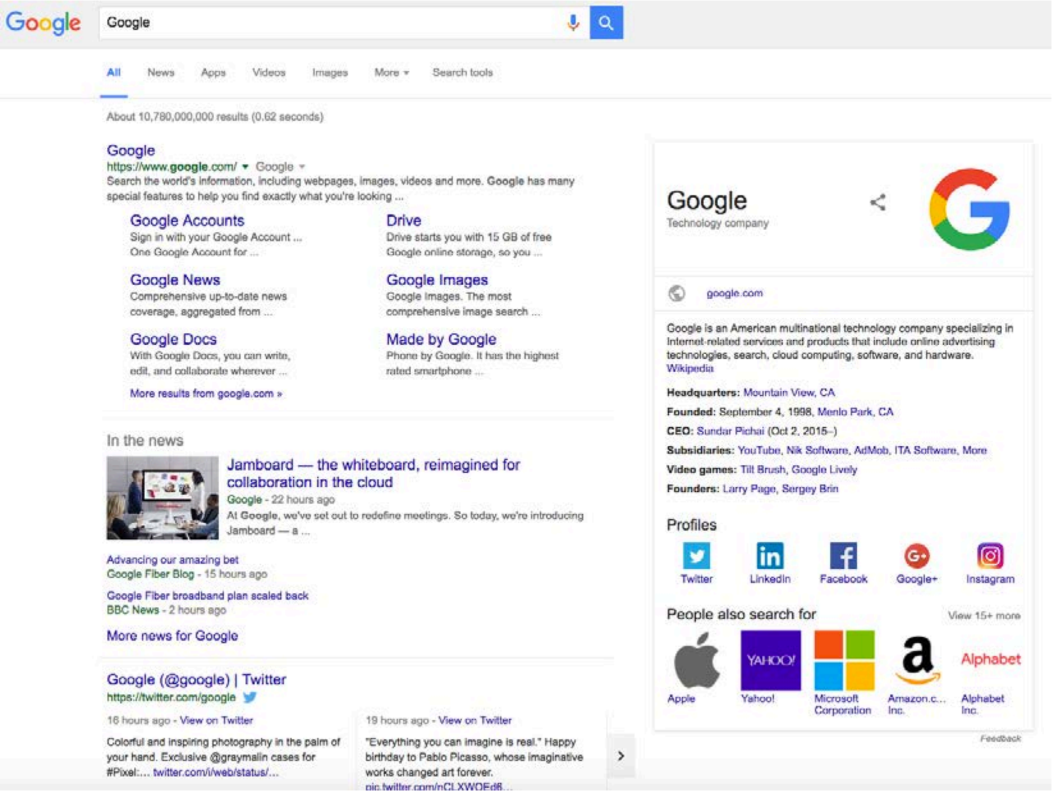
# Using Text for Business Insight

Oded Netzer
Columbia Business School



# The Image of Data





About 10,780,000,000 results (0.62 seconds)

Google
https://www.google.com/ ▾ Google ▾
Search the world's information, including webpages, images, videos and more. Google has many special features to help you find exactly what you're looking ...

Google Accounts
Sign in with your Google Account ... One Google Account for ...

Drive
Drive starts you with 15 GB of free Google online storage, so you ...

Google News
Comprehensive up-to-date news coverage, aggregated from ...

Google Images
Google Images. The most comprehensive image search ...

Google Docs
With Google Docs, you can write, edit, and collaborate wherever ...

Made by Google
Phone by Google. It has the highest rated smartphone ...

More results from google.com »

In the news

Jamboard — the whiteboard, reimagined for collaboration in the cloud
Google - 22 hours ago
At Google, we've set out to redefine meetings. So today, we're introducing Jamboard — a ...

Advancing our amazing bet
Google Fiber Blog - 15 hours ago

Google Fiber broadband plan scaled back
BBC News - 2 hours ago

More news for Google

Google (@google) | Twitter
https://twitter.com/google

16 hours ago - View on Twitter
Colorful and inspiring photography in the palm of your hand. Exclusive @graymalin cases for #Pixel:... twitter.com/i/web/status/...

19 hours ago - View on Twitter
"Everything you can imagine is real." Happy birthday to Pablo Picasso, whose imaginative works changed art forever. pic.twitter.com/nCLXWOEd8...

Google
Technology company

google.com

Google is an American multinational technology company specializing in Internet-related services and products that include online advertising technologies, search, cloud computing, software, and hardware. Wikipedia

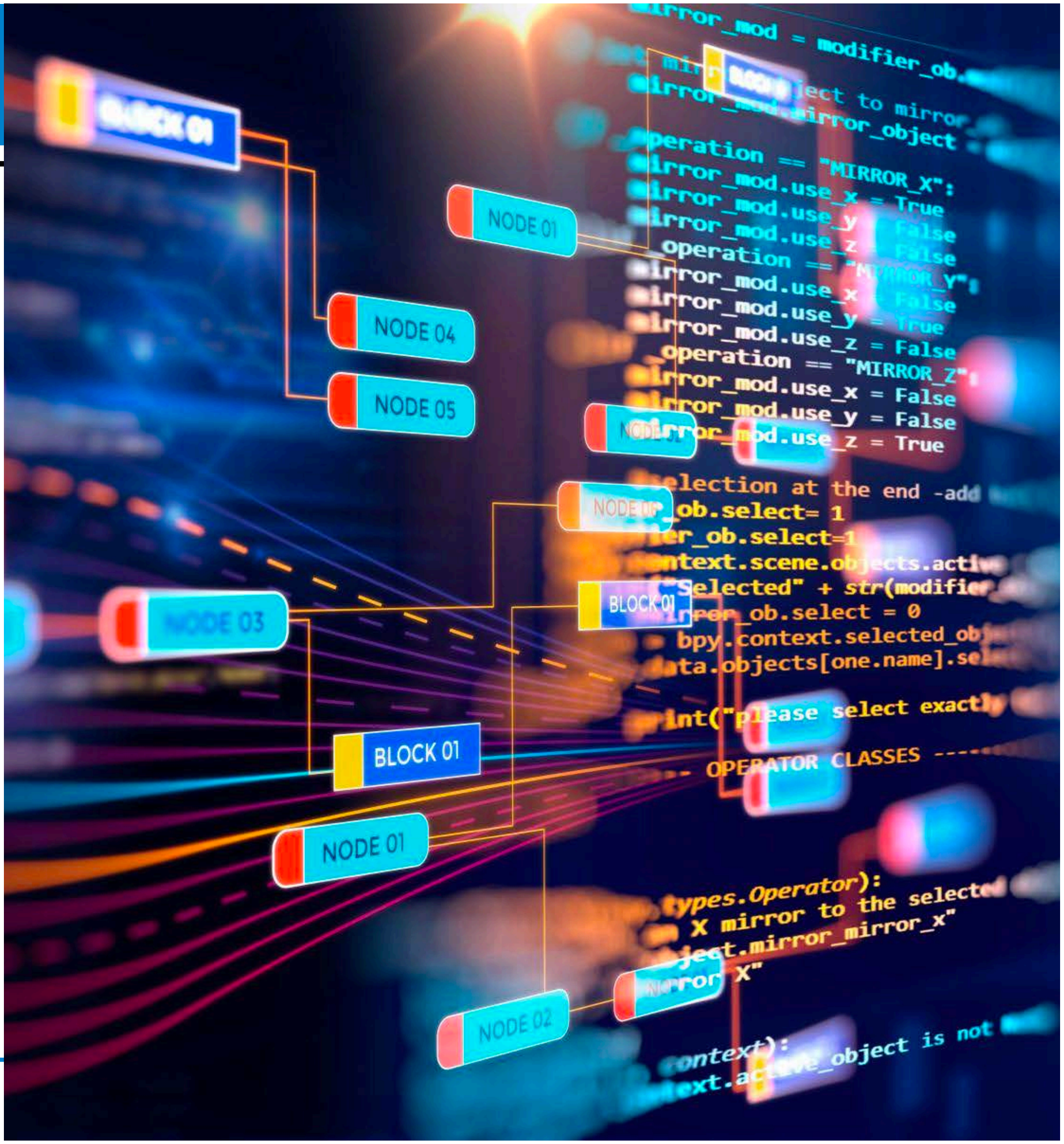
**Headquarters:** Mountain View, CA
**Founded:** September 4, 1998, Menlo Park, CA
**CEO:** Sundar Pichai (Oct 2, 2015–)
**Subsidiaries:** YouTube, Nik Software, AdMob, ITA Software, More
**Video games:** Tilt Brush, Google Lively
**Founders:** Larry Page, Sergey Brin

**80-95% of all usable information for businesses is in the form of unstructured data***

* Gandomi and Haider 2015

# Huge Amount of Text Data

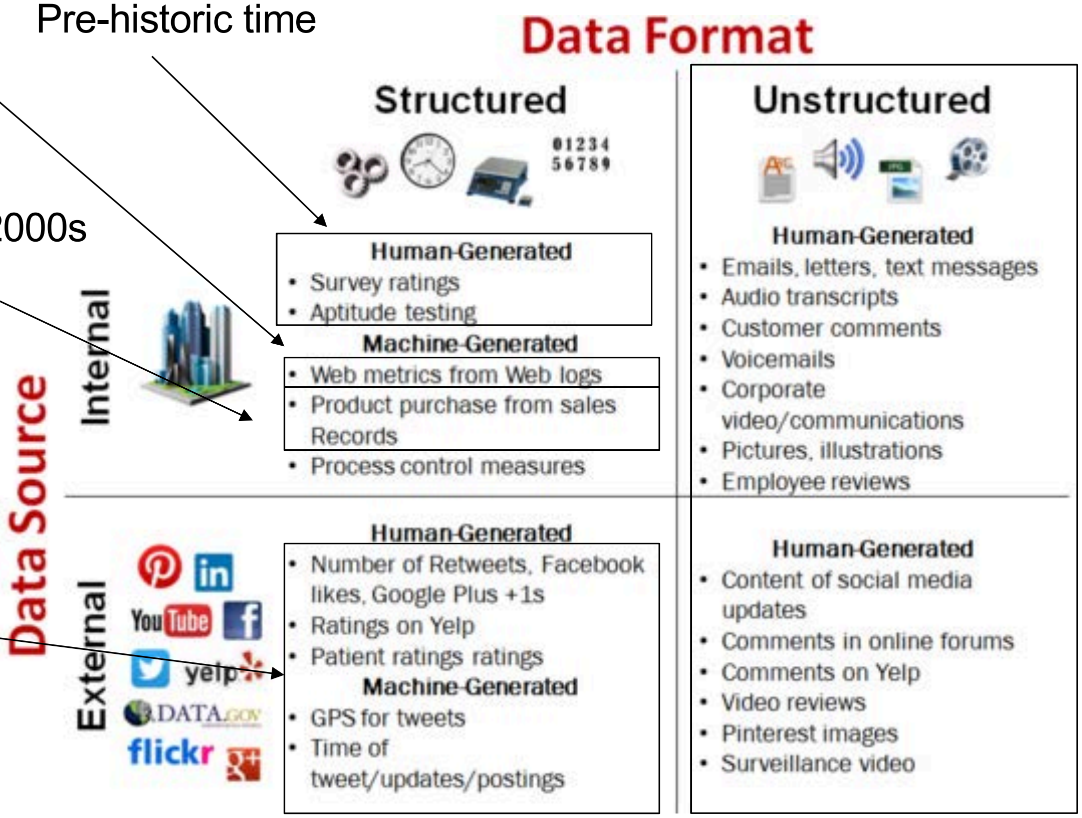
- Online reviews
- Social media posts
- Texts
- Customer service calls
- Open-ended survey questions
- Firm annual reports
- Advertisements
- Newspaper articles
- Movie scripts
- Song lyrics
- How can we extract insight?



Early 2000s

Pre-historic time

1983-early 2000s



| | | Data Format | |
|---|---|---|---|
| | | Structured | Unstructured |
| Data Source | Internal | **Human-Generated**<br>• Survey ratings<br>• Aptitude testing<br>**Machine-Generated**<br>• Web metrics from Web logs<br>• Product purchase from sales Records<br>• Process control measures | **Human-Generated**<br>• Emails, letters, text messages<br>• Audio transcripts<br>• Customer comments<br>• Voicemails<br>• Corporate video/communications<br>• Pictures, illustrations<br>• Employee reviews |
| | External | **Human-Generated**<br>• Number of Retweets, Facebook likes, Google Plus +1s<br>• Ratings on Yelp<br>• Patient ratings ratings<br>**Machine-Generated**<br>• GPS for tweets<br>• Time of tweet/updates/postings | **Human-Generated**<br>• Content of social media updates<br>• Comments in online forums<br>• Comments on Yelp<br>• Video reviews<br>• Pinterest images<br>• Surveillance video |

Mid 2000s…

Last few years

# Textual Data Goes Beyond Social Media…

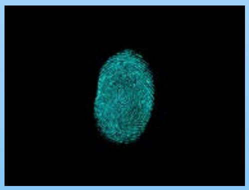
**Text Recipient**

**Text Generator**

| | **Consumers** | **Firms** | **Society** |
|---|---|---|---|
| **Consumers** | Reviews, social Media | Call centers, chats | Online comments |
| **Firms** | Product information, ads | Internal e-mails, financial reports | Financial reports, interviews |
| **Society** | News, songs, movies | Business news (e.g., HBR) | Congressional discussions |

# Two Main Ways we Can Use Text Data

Language reflects



Language affects

# Textual Data Unites the Marketing Tribes

## Consumer Behavior

Secondary data with a window into the why

## Quantitative Marketing

Thousands of new predictors and the methodological challenges

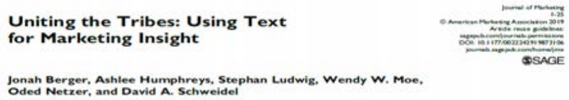
**Textual Data**

## Marketing Strategy

Rich data on both firms and consumers

## Consumer Culture

Finally can quantify CCT data



**Uniting the Tribes: Using Text for Marketing Insight**

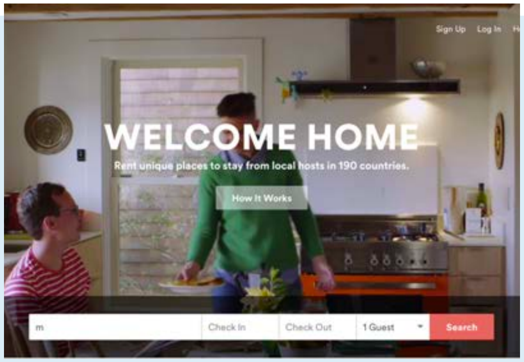
Jonah Berger, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer, and David A. Schweidel

Journal of Marketing
1-25
© American Marketing Association 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0022242919873106
journals.sagepub.com/home/jmx
SAGE

# Extracting Hosts' Motivations from Text - Bringing Motivations to Marketing Science

Chung, Li, Johar, Netzer and Peterson, 2020

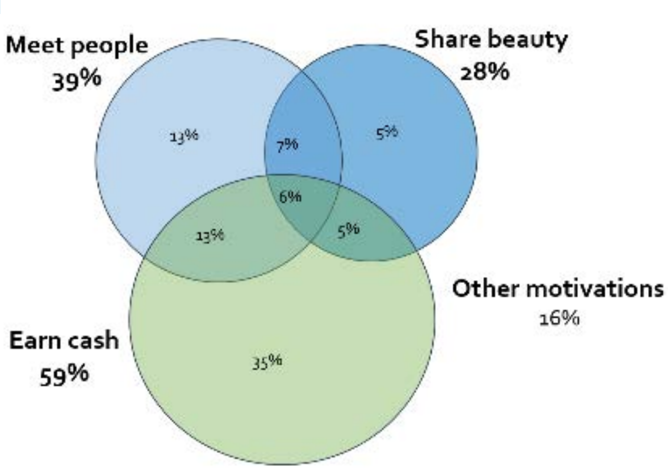
# What Motivates People to Share their Home?



Rent unique places to stay from local hosts in 190 countries.

Q: “Why did you start hosting?”

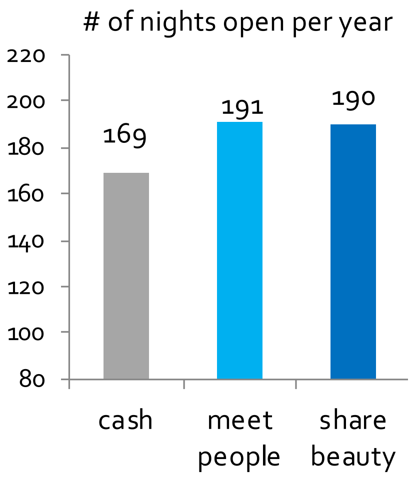
“To earn extra cash”
“I wanted to meet new people from all around the world!”
“To make a little money and share the beauty of this place with others for brief periods.”

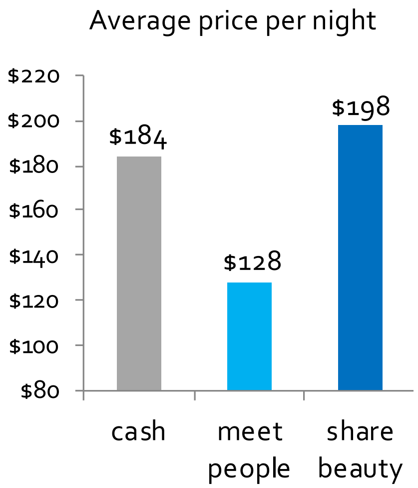


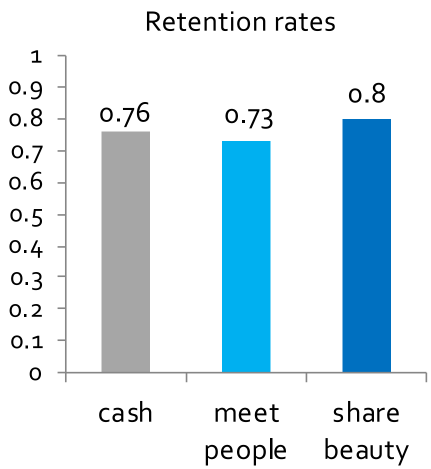
## Data

| | |
|---|---|
| Period | January, 2011 – October, 2015 |
| Hosts | 25,290 individuals |
| Observations | 561,628 reservations |
| Countries | 143 countries |

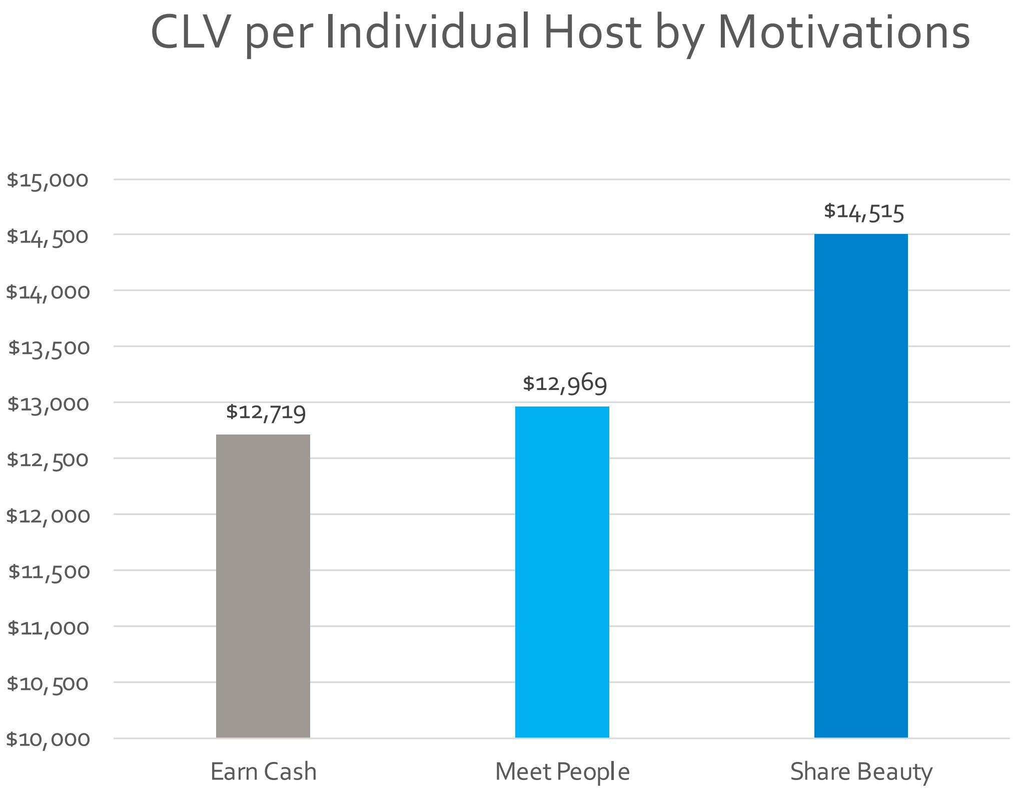
# Bringing Motivations to Marketing Science







# Customer Lifetime Value per motivation group

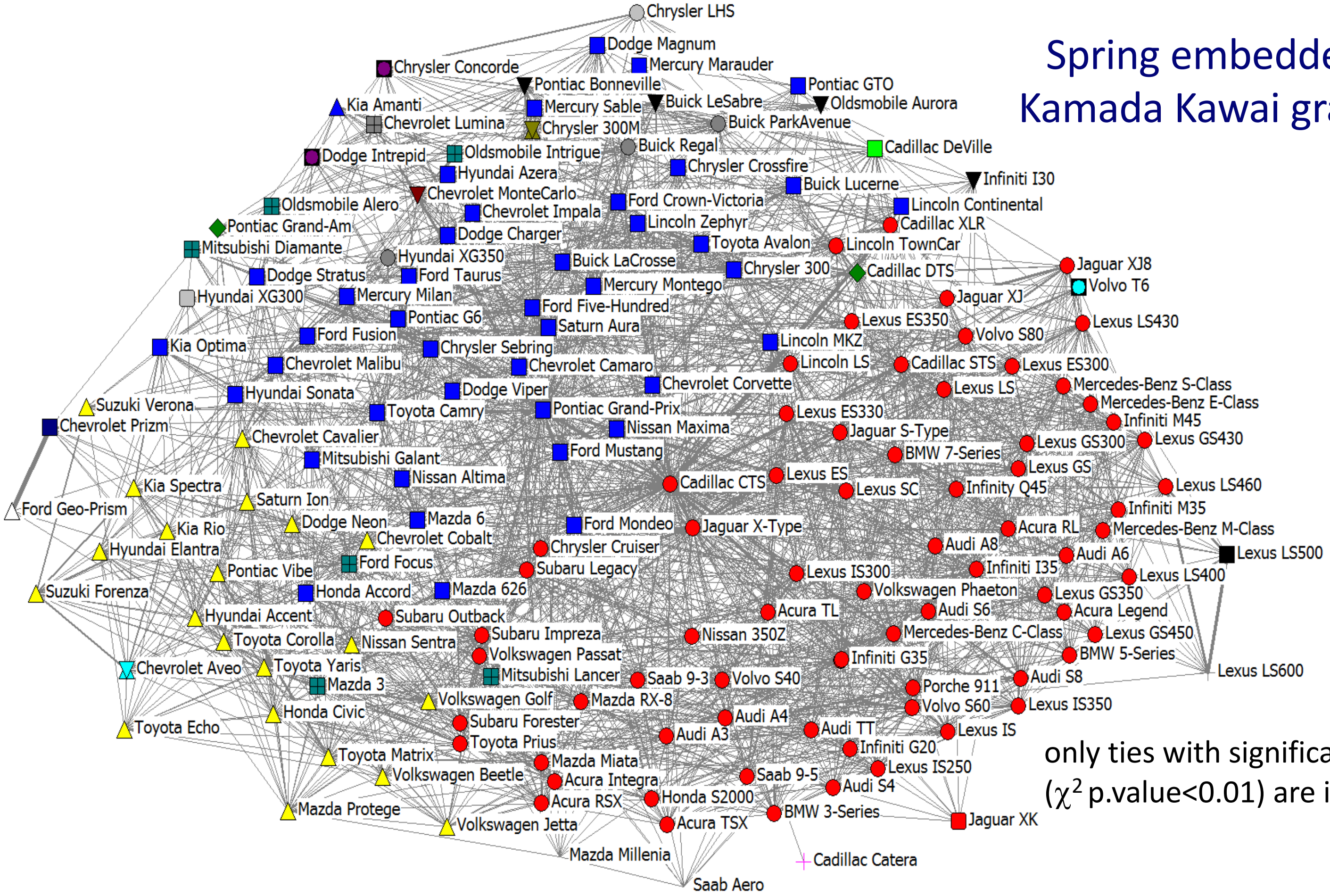


# Mine Your Own Business

Netzer, Feldman, Goldenberg, Fresko 2012

# The Car Market Competitive Landscape

Spring embedded
Kamada Kawai graph

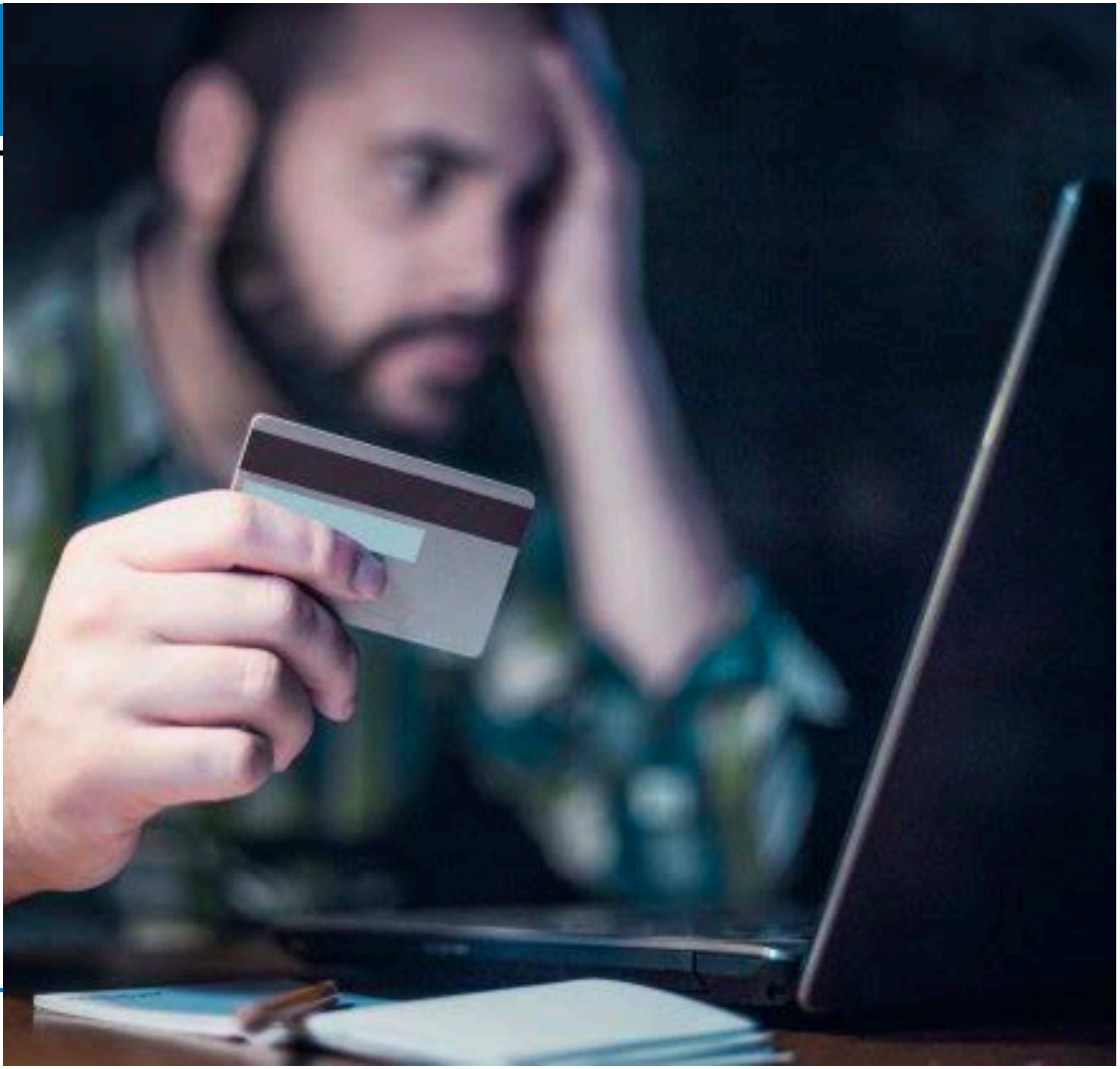


only ties with significant lifts ($\chi^2$ p.value<0.01) are included

Columbia Business School
AT THE VERY CENTER OF BUSINESS™

# Can Words Predict Default?

Netzer, Lemaire, Herzenstien 2019



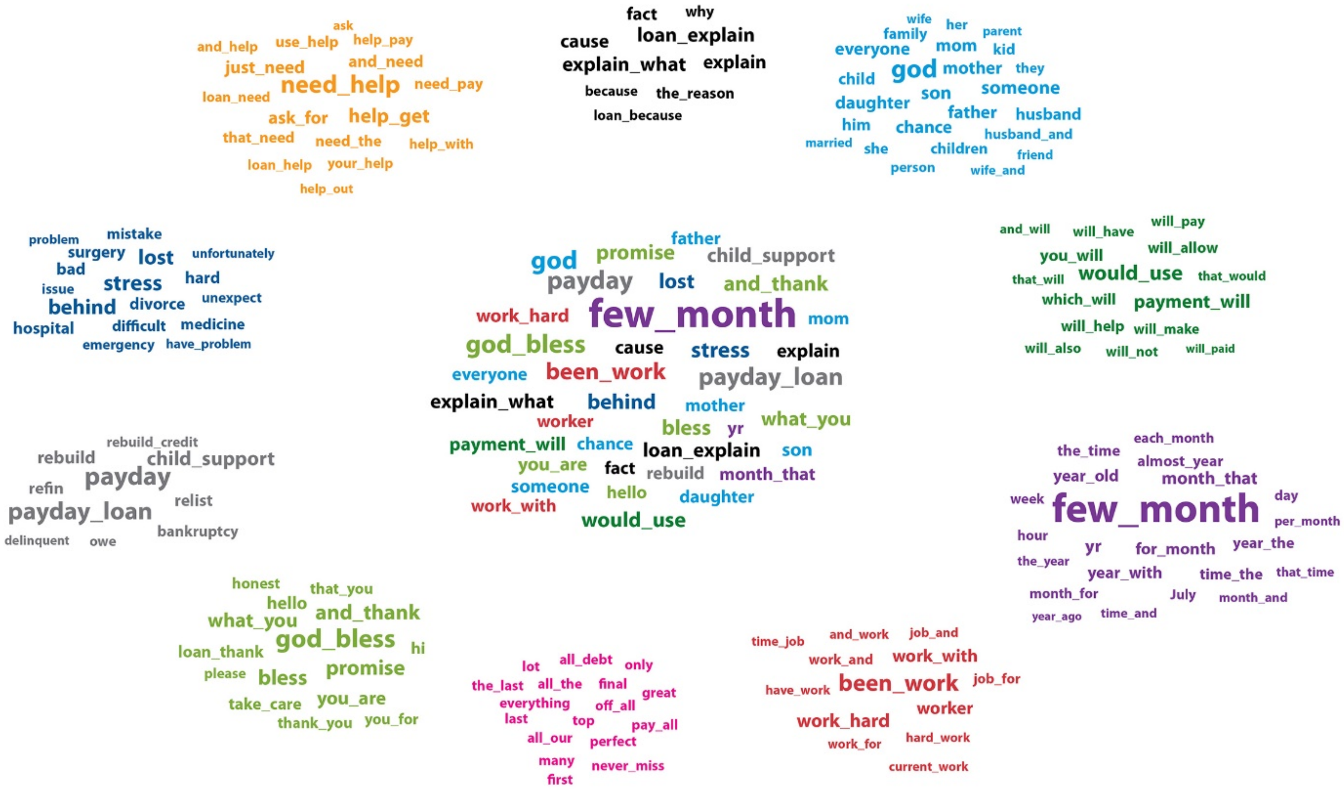
Columbia Business School
AT THE VERY CENTER OF BUSINESS™

# Can Words Predict Default?

***Can borrowers' loan application text—writing style and content—predict future repayment behavior?***

- What is the **predictive** power of writing styles?
- Which **words** are predictive of default?
- What is the **writing style** that is most associated with default?

# What do defaulters write about?


ask
and_help
use_help
help_pay
just_need
and_need
loan_need
need_help
need_pay
ask_for
help_get
that_need
need_the
help_with
loan_help
your_help
help_out
fact
why
cause
loan_explain
explain_what
explain
because
the_reason
loan_because
wife
her
family
parent
everyone
mom
kid
child
god
mother
they
daughter
son
someone
father
husband
him
chance
husband_and
married
she
children
friend
person
wife_and
problem
mistake
surgery
lost
unfortunately
bad
stress
hard
issue
behind
divorce
unexpect
hospital
difficult
medicine
emergency
have_problem
father
god
promise
child_support
payday
lost
and_thank
work_hard
few_month
mom
god_bless
cause
stress
explain
everyone
been_work
payday_loan
explain_what
behind
mother
worker
bless
yr
what_you
payment_will
chance
loan_explain
son
you_are
fact
rebuild
month_that
someone
hello
daughter
work_with
would_use
and_will
will_have
will_pay
you_will
will_allow
that_will
would_use
that_would
which_will
payment_will
will_help
will_make
will_also
will_not
will_paid
rebuild_credit
rebuild
child_support
payday
refin
relist
payday_loan
bankruptcy
delinquent
owe
the_time
each_month
almost_year
year_old
month_that
week
few_month
day
per_month
hour
yr
for_month
year_the
the_year
year_with
time_the
that_time
month_for
July
month_and
year_ago
time_and
honest
that_you
hello
what_you
and_thank
loan_thank
god_bless
hi
please
bless
promise
take_care
you_are
thank_you
you_for
lot
all_debt
only
the_last
all_the
final
great
everything
off_all
last
top
pay_all
all_our
perfect
many
never_miss
first
time_job
and_work
job_and
work_and
work_with
have_work
been_work
job_for
worker
work_hard
hard_work
work_for
current_work

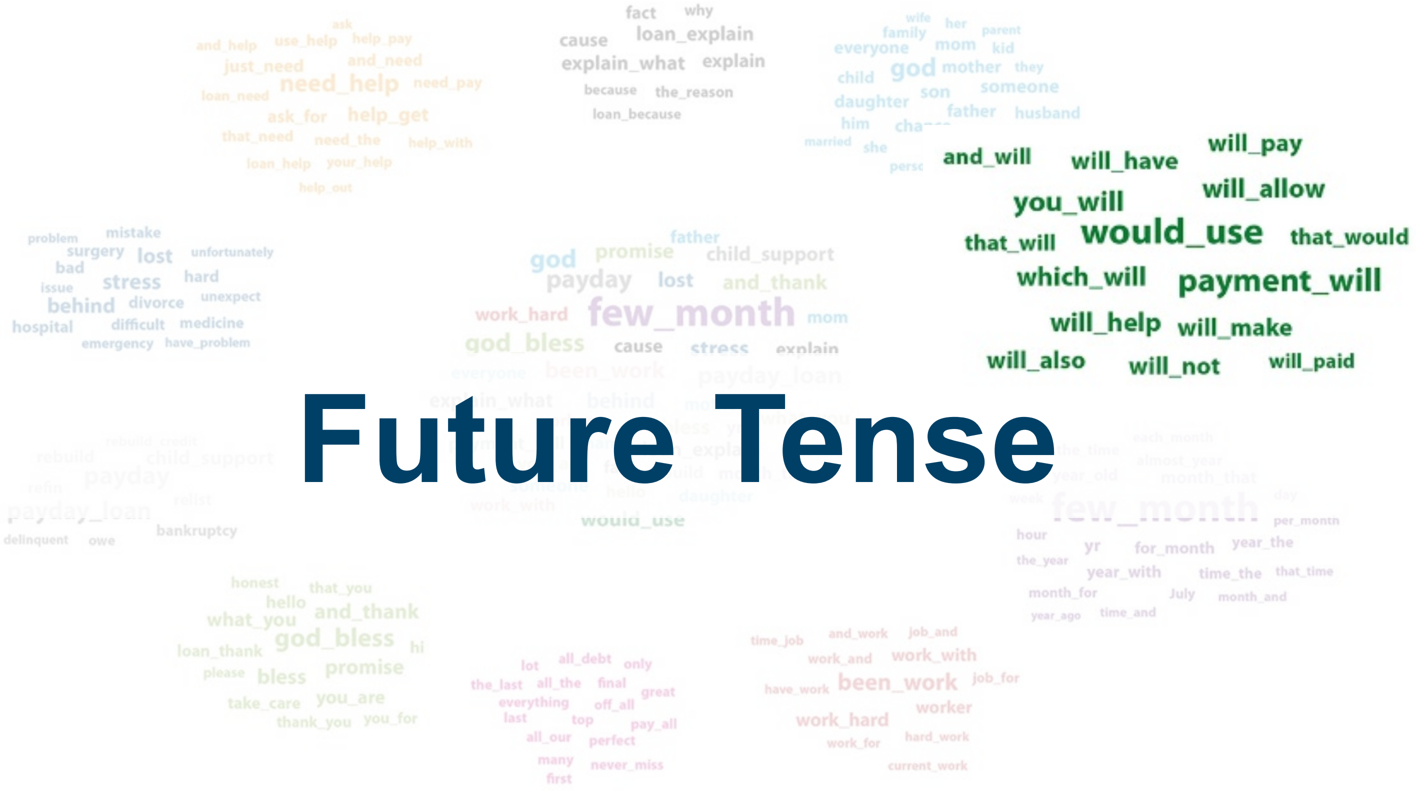
# Hardship & financial hardship

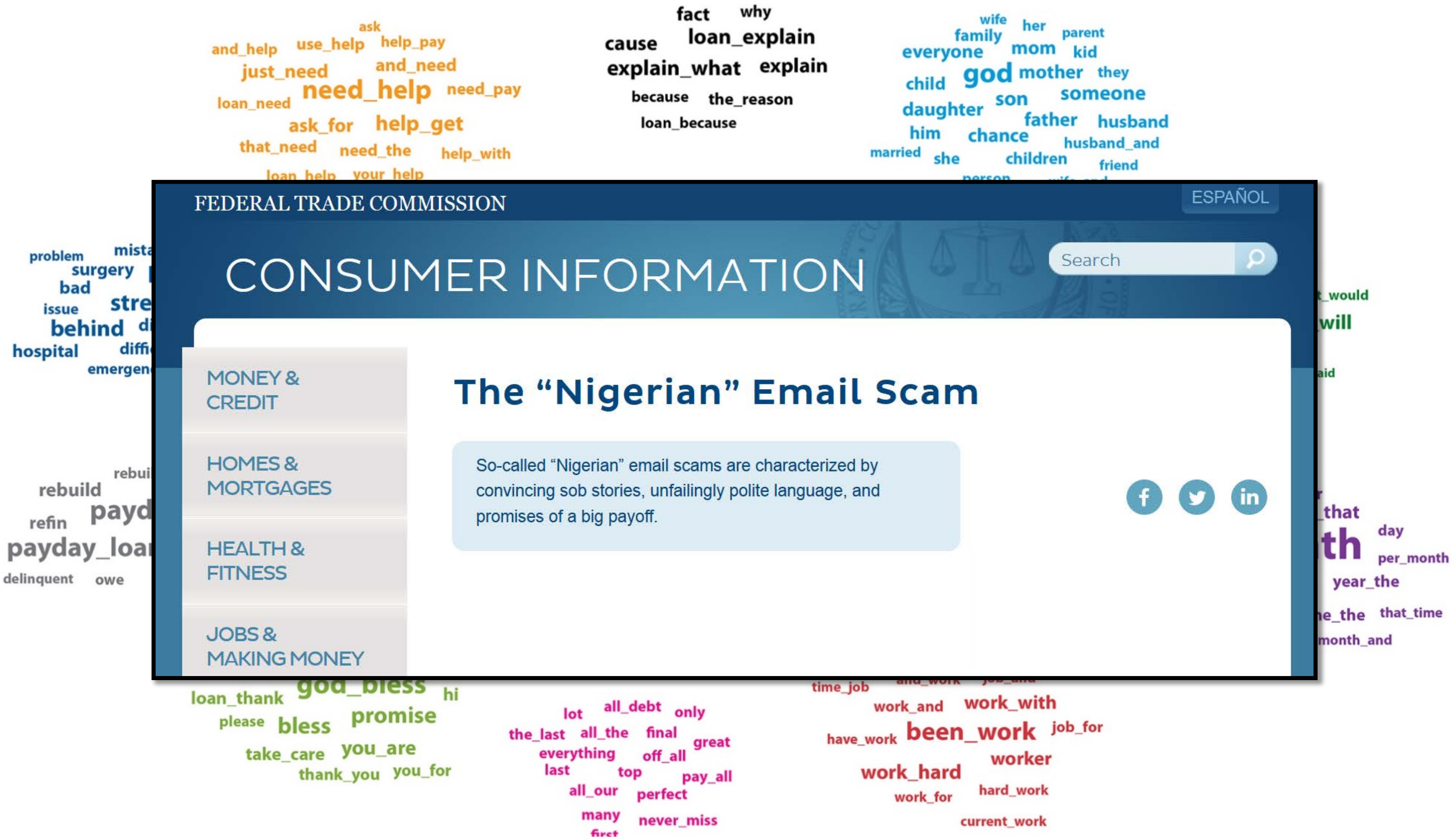
# Desperation & plea

# Explanation

# External influence & others

# Future Tense





FEDERAL TRADE COMMISSION

ESPAÑOL

CONSUMER INFORMATION

Search

MONEY & CREDIT

HOMES & MORTGAGES

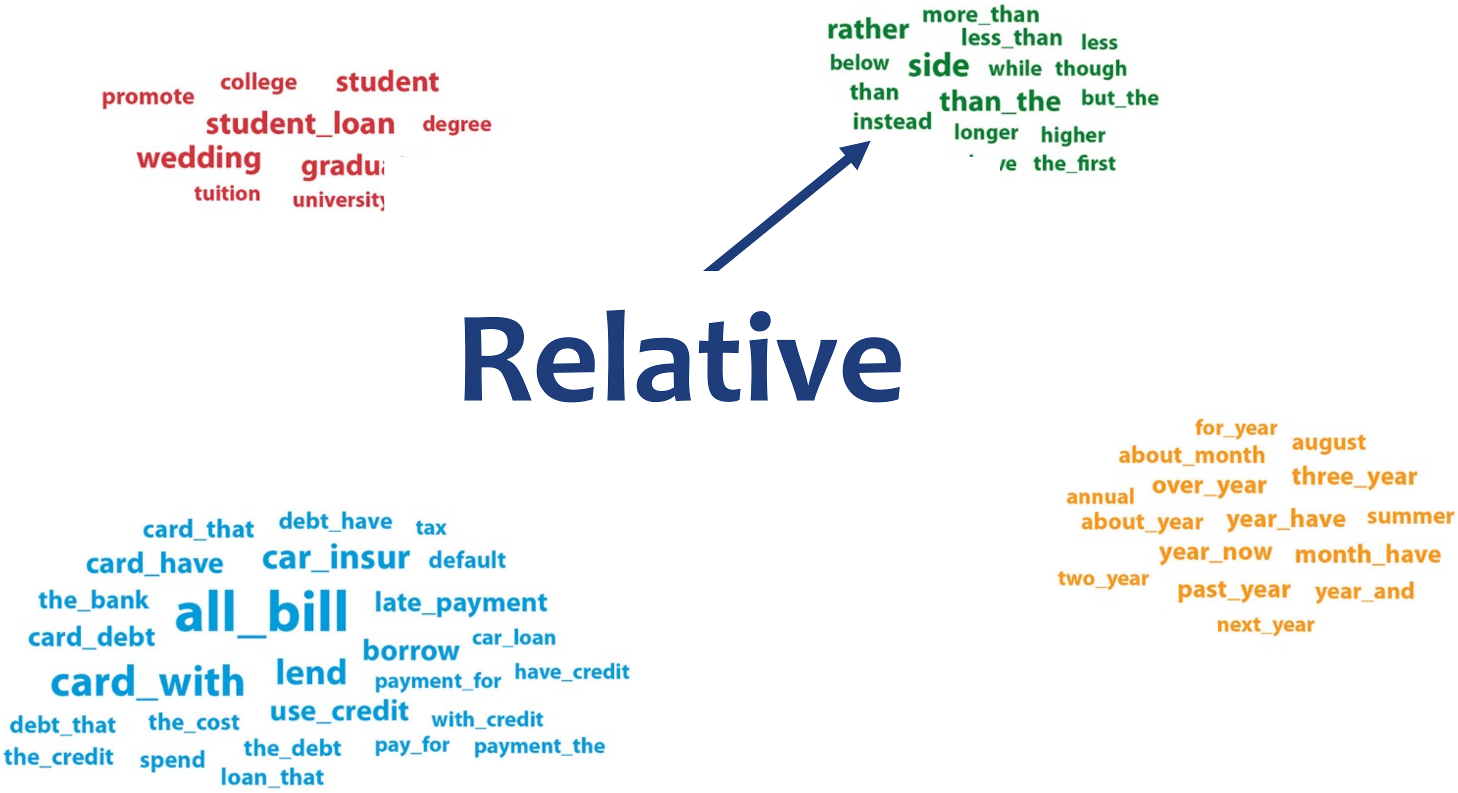
HEALTH & FITNESS

JOBS & MAKING MONEY

## The “Nigerian” Email Scam

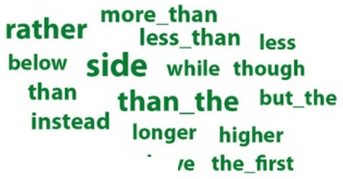
So-called “Nigerian” email scams are characterized by convincing sob stories, unfailingly polite language, and promises of a big payoff.

# What do repayers write


rather
more_than
less_than
less
below
side
while
though
than
than_the
but_the
instead
longer
higher
the_first
promote
college
student
student_loan
degree
wedding
gradu
tuition
university
for_year
about_month
august
over_year
three_year
annual
about_year
year_have
summer
year_now
month_have
two_year
past_year
year_and
next_year
card_that
debt_have
tax
card_have
car_insur
default
the_bank
all_bill
late_payment
card_debt
car_loan
borrow
card_with
lend
payment_for
have_credit
debt_that
the_cost
use_credit
with_credit
the_credit
spend
the_debt
pay_for
payment_the
loan_that

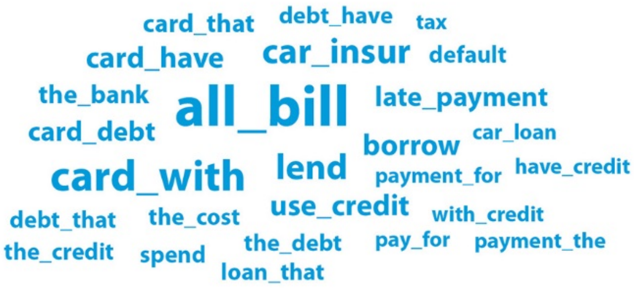
# Relative


promote
college
student
student_loan
degree
wedding
gradu
tuition
university

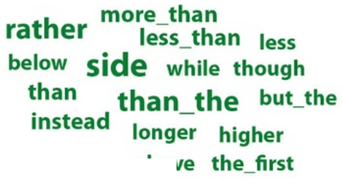

rather
more_than
less_than
less
below
side
while
though
than
than_the
but_the
instead
longer
higher
ve
the_first

# Time

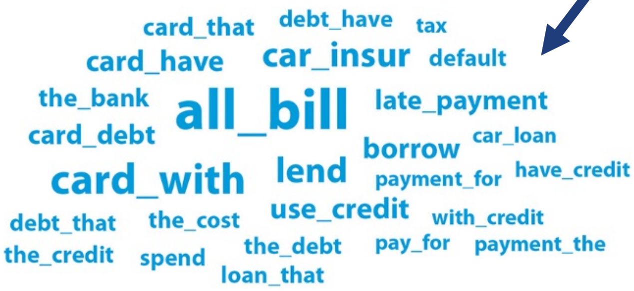

for_year
about_month
august
over_year
three_year
annual
about_year
year_have
summer
year_now
month_have
two_year
past_year
year_and
next_year


card_that
debt_have
tax
card_have
car_insur
default
the_bank
all_bill
late_payment
card_debt
borrow
car_loan
card_with
lend
payment_for
have_credit
use_credit
with_credit
debt_that
the_cost
the_debt
pay_for
payment_the
the_credit
spend
loan_that

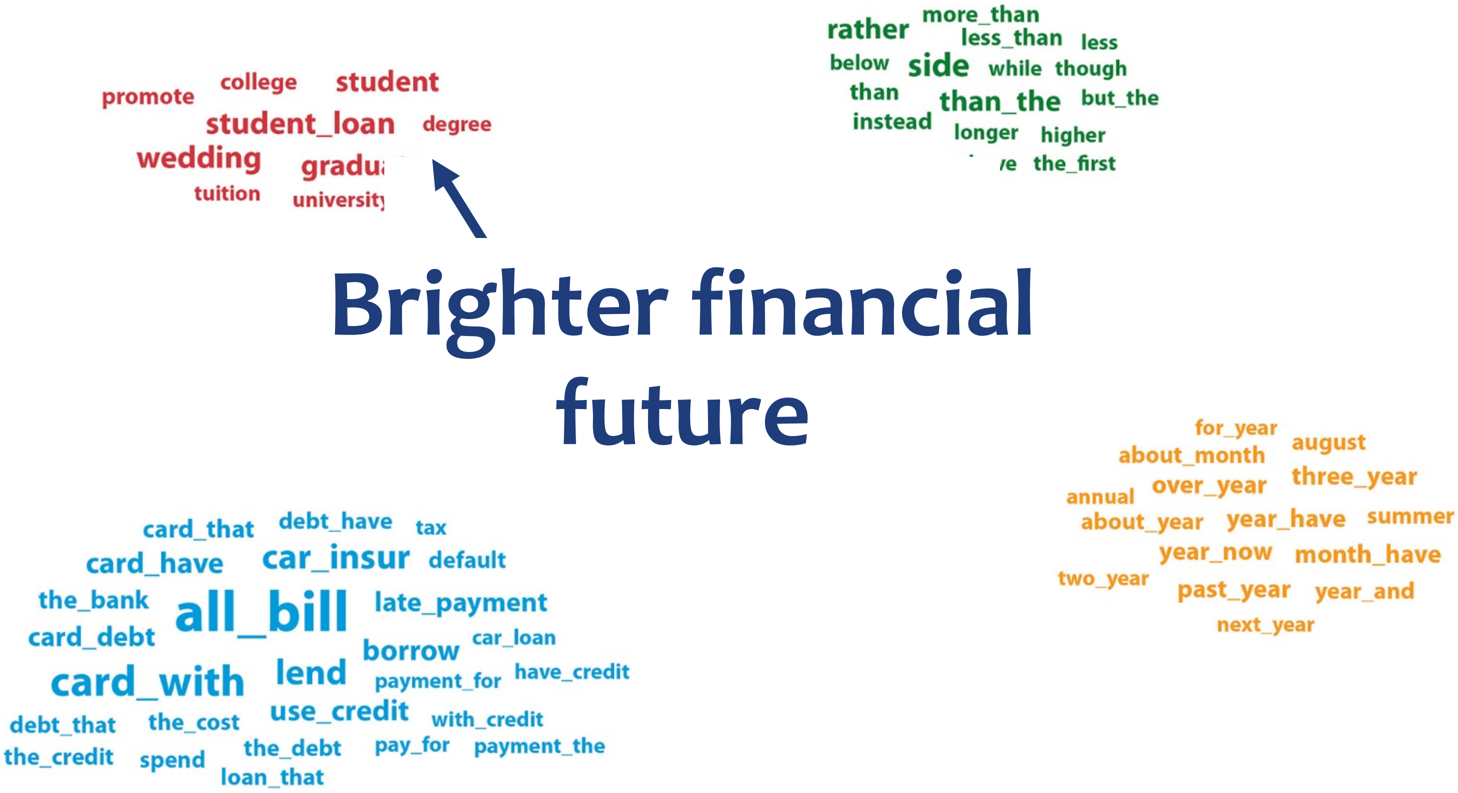

promote
college
student
student_loan
degree
wedding
gradu
tuition
university


rather
more_than
less_than
less
below
side
while
though
than
than_the
but_the
instead
longer
higher
ve
the_first

# Debt


card_that
debt_have
tax
card_have
car_insur
default
the_bank
all_bill
late_payment
card_debt
borrow
car_loan
card_with
lend
payment_for
have_credit
debt_that
the_cost
use_credit
with_credit
the_credit
spend
the_debt
pay_for
payment_the
loan_that


for_year
august
about_month
three_year
annual
over_year
about_year
year_have
summer
year_now
month_have
two_year
past_year
year_and
next_year



# Brighter financial future

# Extraversion

Defaulting loan requests have writing styles of extraverts

- More religious and body words (god, bless, cloth, eye)
- More social and human words (we, family)
- More motion words (arrive, car, increase)
- Less negation words (not, never)
- Less insight words (think, know, consider)

e.g., Pennebaker and King 1999; Hirsh and Peterson 2009; Yarkoni 2010; Schwartz, et al. 2013

# Deception

Defaulting loan requests have traces of deception

- More present and future tense
- More social words (we rather than I)
- More motion words (arrive, car, increase)
- Longer text

e.g., Newman, et al. 2003; Pasupathi 2007; Hancock, et al. 2008; Ott, et al. 2011

# Leveraging Text Mining for Idea Generation

Toubia, Netzer 2017

Columbia Business School
AT THE VERY CENTER OF BUSINESS™

# Leveraging Text Mining for Idea Generation

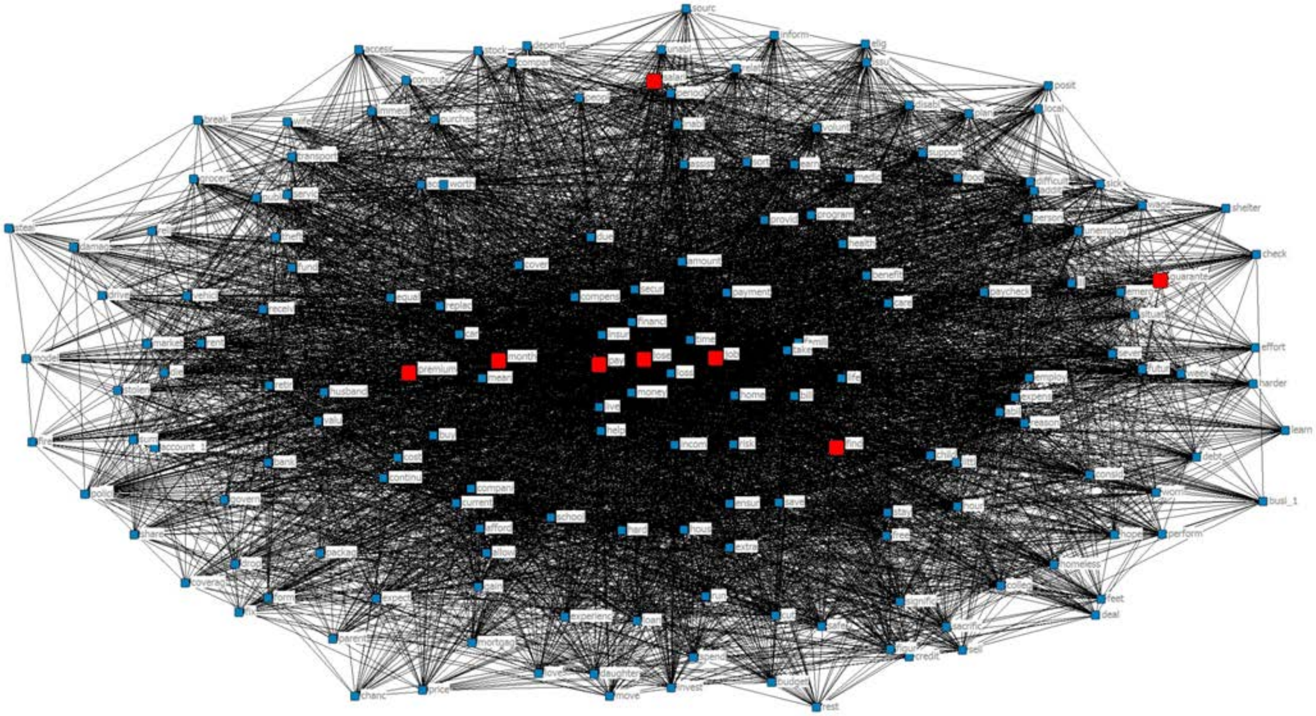
# Balancing Novelty with Familiarity

“Truly useful creativity may reflect a balance between novelty and a connection to previous ideas” (Ward 1995)

Semantic subnetwork between concepts in an idea

- Concepts close to each other in the semantic network → familiarity
- Concepts far from each other in the semantic network→ novelty

# Good Ideas Balance the Novel with the Familiar



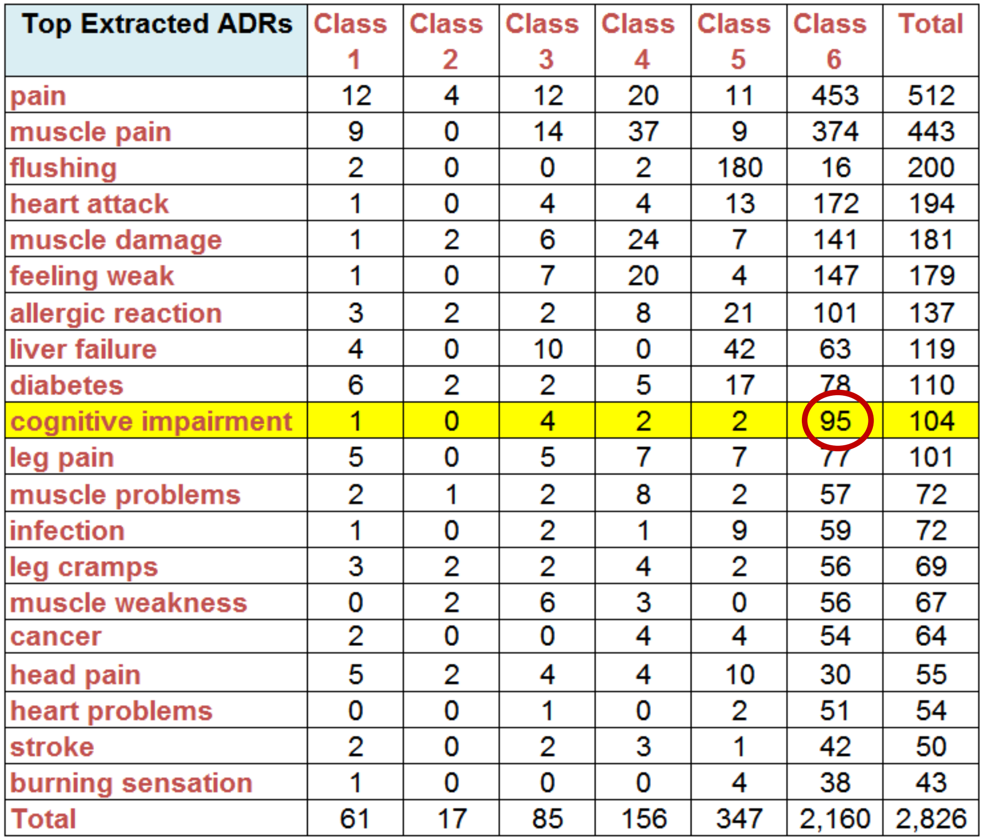
Idea #201: “-LOSE Job -GET A guarantee of 70% of their former salary for 5 years if they cannot find a job that paid as much as they were making. -GIVE A premium every month based on their salary.“

# Utilizing Text Mining to Predict Label Change due to Adverse Drug Reactions

Feldman, Netzer, Peretz, Rosenfeld, 2015

# 2011 Label Change Statins and Cognitive Impairment

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 |

# 2011 Label Change Statins and Cognitive Impairment

## 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 |

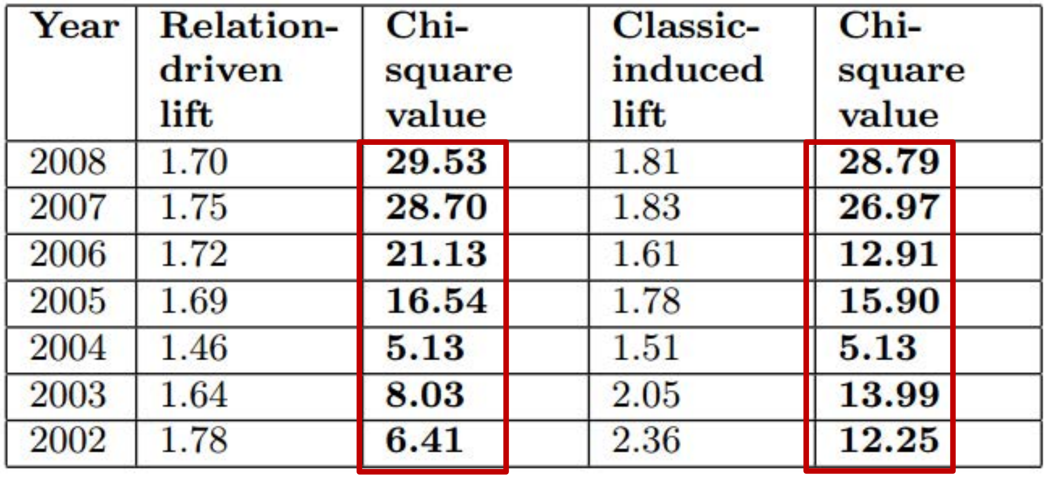
# How Early Could We Have Detected?

| 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 are significant at the 0.05 level

# Similar Results for Wellbutrin and Agitation

- How early could we have detected it?

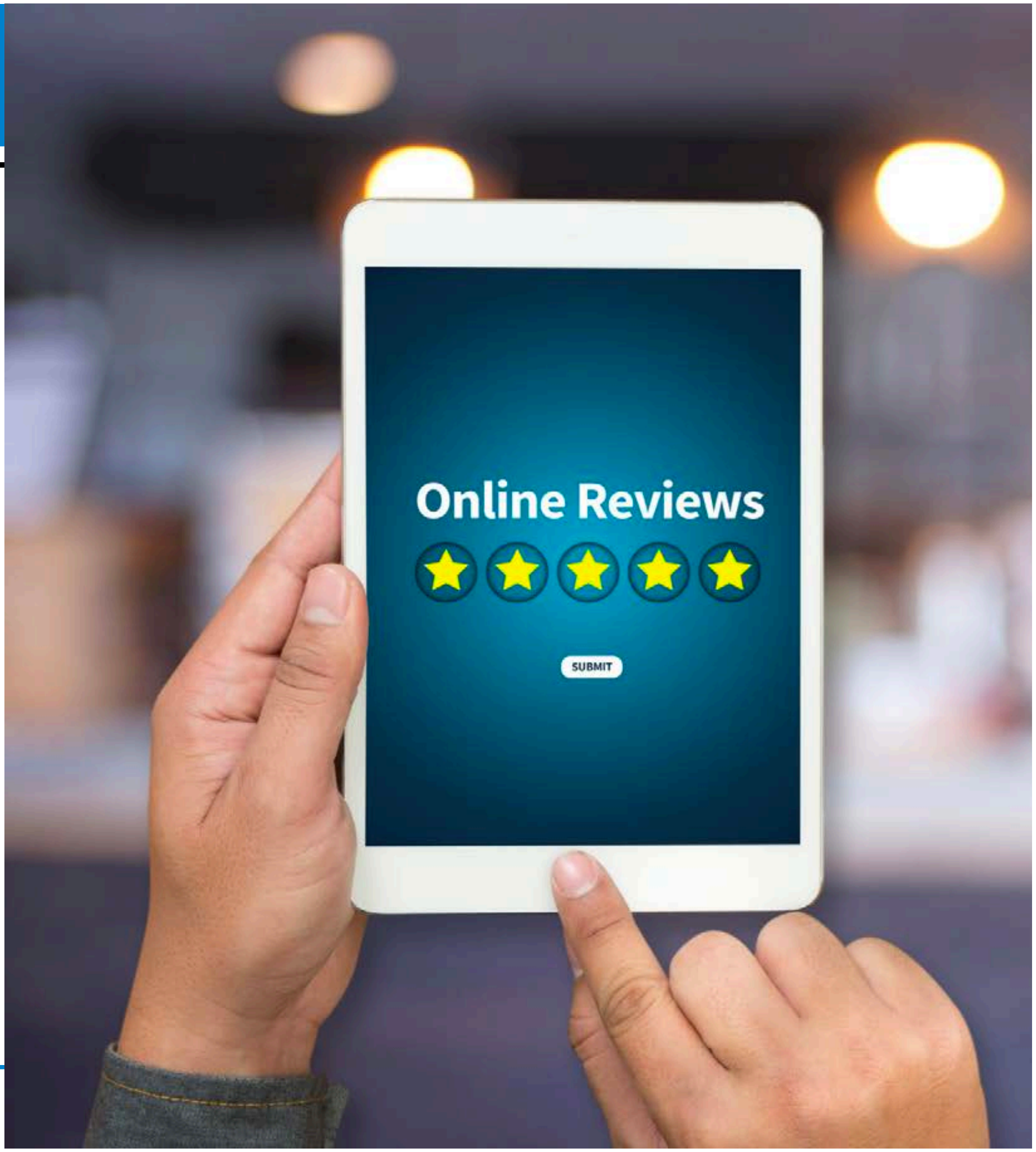


| Year | Relation-driven lift | Chi-square value | Classic-induced lift | Chi-square value |
|---|---|---|---|---|
| 2008 | 1.70 | **29.53** | 1.81 | **28.79** |
| 2007 | 1.75 | **28.70** | 1.83 | **26.97** |
| 2006 | 1.72 | **21.13** | 1.61 | **12.91** |
| 2005 | 1.69 | **16.54** | 1.78 | **15.90** |
| 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 are significant at the 0.05 level

# Matching Linguistics of Reviewers and Products

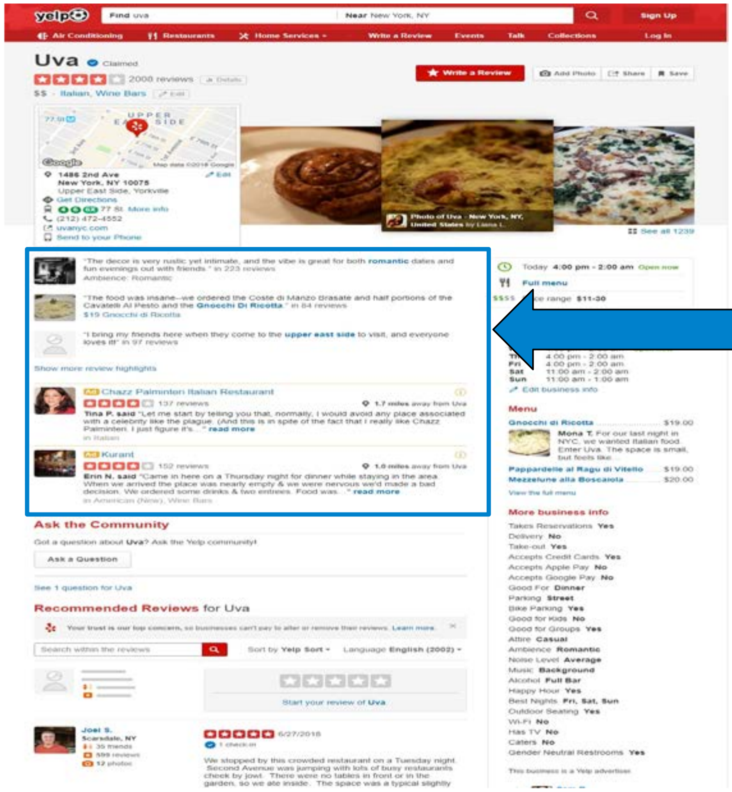
Lemaire and Netzer 2020



Columbia Business School
AT THE VERY CENTER OF BUSINESS™

# Matching Linguistics of Reviewers and Products

Reviews of the restaurants



"The decor is very rustic yet intimate, and the vibe is great for both romantic dates and fun evenings out with friends." in 223 reviews
Ambience: Romantic

"The food was insane--we ordered the Coste di Manzo Brasate and half portions of the Cavatelli Al Pesto and the Gnocchi Di Ricotta." in 84 reviews
$19 Gnocchi di Ricotta

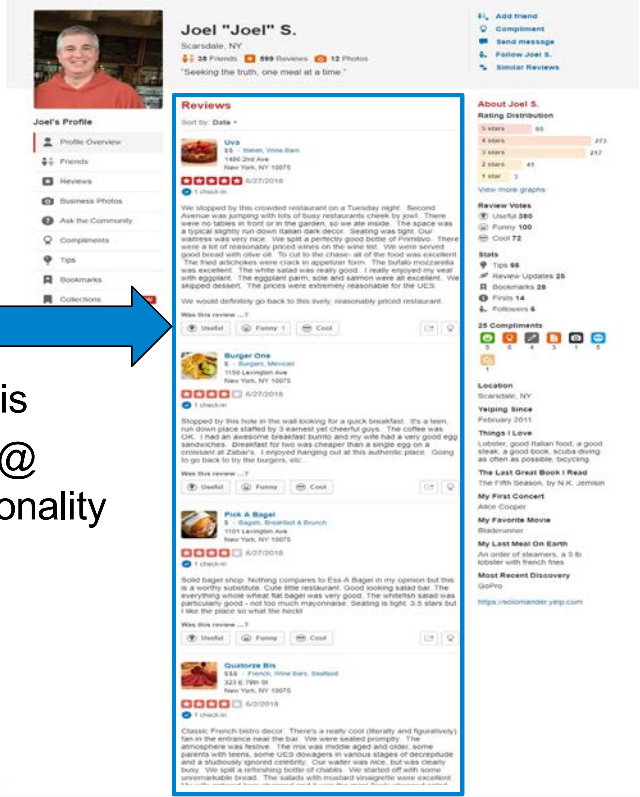
"I bring my friends here when they come to the upper east side to visit, and everyone loves it!" in 97 reviews

Show more review highlights

Chazz Palminteri Italian Restaurant
137 reviews · 1.7 miles away from Uva
**Tina P. said** "Let me start by telling you that, normally, I would avoid any place associated with a celebrity like the plague. (And this is in spite of the fact that I really like Chazz Palminteri. I just figure it's…" **read more**
in Italian

Kurant
152 reviews · 1.0 miles away from Uva
**Erin N. said** "Came in here on a Thursday night for dinner while staying in the area. When we arrived the place was nearly empty & we were nervous we'd made a bad decision. We ordered some drinks & two entrees. Food was…" **read more**
in American (New), Wine Bars

Reviews of the reviewer



**Joel "Joel" S.**
Scarsdale, NY
38 Friends · 599 Reviews · 12 Photos
"Seeking the truth, one meal at a time."

**Reviews**
Sort by: Date

Uva
$$ · Italian, Wine Bars
1486 2nd Ave
New York, NY 10075
6/27/2018
1 check-in

We stopped by this crowded restaurant on a Tuesday night. Second Avenue was jumping with lots of busy restaurants cheek by jowl. There were no tables in front or in the garden, so we ate inside. The space was a typical slightly run down Italian dark decor. Seating was tight. Our waitress was very nice. We split a perfectly good bottle of Primitivo. There were a lot of reasonably priced wines on the wine list. We were served good bread with olive oil. To cut to the chase- all of the food was excellent. The fried artichokes were crack in appetizer form. The bufalo mozzarella was excellent. The white salad was really good. I really enjoyed my veal with eggplant. The eggplant parm, sole and salmon were all excellent. We skipped dessert. The prices were extremely reasonable for the UES.

We would definitely go back to this lively, reasonably priced restaurant.

Burger One
$ · Burgers, Mexican
1150 Lexington Ave
New York, NY 10075
6/27/2018

Emojis

%$&@

Personality

# Summary

- Most data out there are unstructured (it goes beyond social media data)
- Text can help uniting the tribes
- Every text has a generator and an audience
- Text reflects versus text affects
- Other sources of unstructured data (image, audio, video) can provide additional value
- 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…

# What’s Next?

**Oded Netzer**

**Columbia Business School**

**onetzer@gsb.columbia.edu**

# References

- Berger, Jonah, Ashlee Humpherys, Stephan Ludwig, Wendy Moe, Oded Netzer and David Schweidel (2020) **“Uniting the Tribes, Using Text for Marketing Insights.”** *Journal of Marketing* 84(1), 1-25 http://www.columbia.edu/~on2110/Papers/Uniting%20The%20Tribes.pdf

- Oded Netzer, Alain Lemaire, Michal Herzenstein (2019) **"When Words Sweat: Written Words Can Predict Loan Default."** *Journal of Marketing Research* 56(6), 960-980 http://www.columbia.edu/~on2110/Papers/When%20Words%20Sweat1.pdf

- Toubia, Olivier and Oded Netzer (2017), **"Idea Generation, Creativity, and Prototypicality."** *Marketing Science*, 36 (1), 1-20 http://www.columbia.edu/~on2110/Papers/Toubia_Oded_2017.pdf

- Feldman Ronen, Oded Netzer, Aviv Peretz, Binyamin Rosenfeld (2015) **Utilizing Text Mining on Online Medical Forums to Predict Label Change Due to Adverse Drug Reactions,** *Proceedings of the 21st ACM SIGKDD International Conference of Knowledge Discovery and Data Mining,* 1779-1788. https://dl.acm.org/doi/10.1145/2783258.2788608

- Netzer, Oded, Ronen Feldman, Jacob Goldenberg and Moshe Fresko (2012), **"Mine Your Own Business: Market Structure Surveillance through Text Mining."** *Marketing Science*, 31 (3), 521-543 http://www.columbia.edu/~on2110/Papers/Netzer_Feldman_Goldenberg_Fresko_2012.pdf

- Jaeyeon Chung, Yanayn Li, Gita Johar Oded Netzer and Mathew Pearson (2020) **“Mining Consumer Minds: The Downstream Consequences of Airbnb Host Motivations.”**

- Lemaire, Alain and Oded Netzer (2020) **“Linguistic-Based Recommendation: The Role of Linguistic Match Between Users and Products.”**