Detecting Financial Misreporting in 2019

March 2019

Dr. Richard M. Crowley
rcrowley@smu.edu.sg
http://rmc.link/

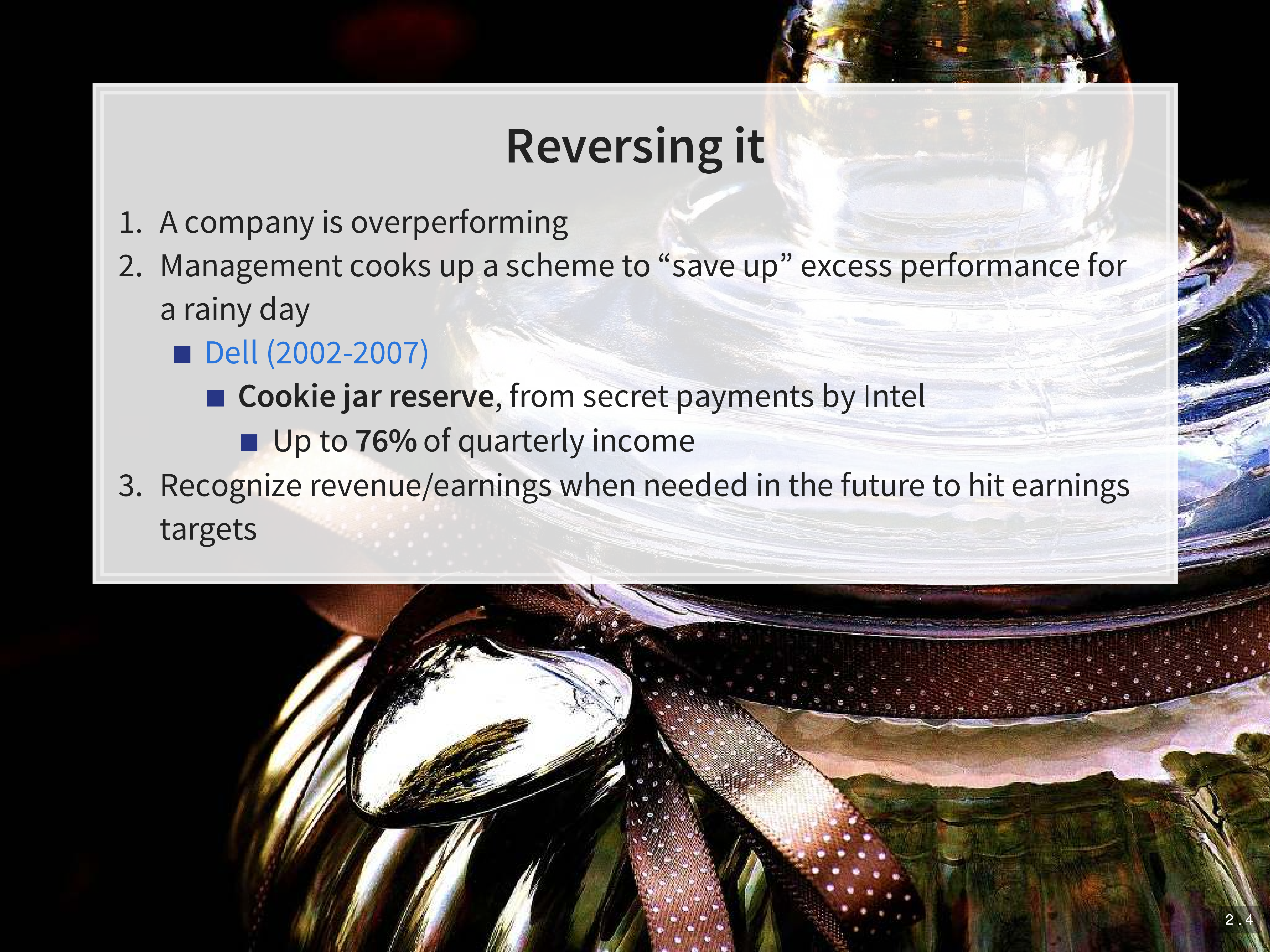
What is Misreporting?

Misreporting: Simple definition

> Misstatements that affect firms’ accounting statements and were done seemingly intentionally by management or other employees at the firm.

Traditional accounting fraud

1. A company is underperforming
2. Management cooks up some scheme to increase earnings
   - Wells Fargo (2011-2018?)
     - Fake/duplicate customers and transactions
3. Create accounting statements using the fake information



Reversing it

1. A company is overperforming
2. Management cooks up a scheme to “save up” excess performance for a rainy day
    - Dell (2002-2007)
        - **Cookie jar reserve**, from secret payments by Intel
            - Up to **76%** of quarterly income
3. Recognize revenue/earnings when needed in the future to hit earnings targets

Other accounting fraud types

- Apple (2001)
  - *Options backdating*
- China North East Petroleum Holdings Limited
  - *Related party transactions* (transferring funds to family members)
- Keppel O&M (2001-2014)
  - *Bribery* ($55M USD in bribes to Brazilian officials for contracts)
- CVS (2000)
  - *Improper accounting treatments* (Not using mark-to-market accounting to fair value stuffed animal inventories)
- Countryland Wellness Resorts, Inc. (1997-2000)
  - Gold reserves were actually… dirt.

The data

How do misstatements come to light?

1. The company/management admits to it publicly
2. A government entity forces the company to disclose
   - In more egregious cases, government agencies may disclose the fraud publicly as well
3. Investors sue the firm, forcing disclosure

> This is what we can leverage to detect fraud!

Where are these disclosed?

In the US:

1. SEC AAERs: Accounting and Auditing Enforcement Releases
   - Generally highlight larger or more important cases
   - Written by the SEC, not the company
   - To get a sense what these are, you can read the *Summary* section (starting on page 2) of this AAER against Sanofi
2. 10-K/A filings (/A means amendment)
   - Note: not all 10-K/A filings are caused by fraud!
     - Benign corrections or adjustments can also be filed as a 10-K/A
     - Audit Analytics' write-up on this for 2017
3. By the US government through a 13(b) action
4. In a note inside a 10-K filing
   - These are sometimes referred to as "little r" restatements
5. In a press release, which is later filed with the US SEC as an 8-K
   - 8-Ks are filed for many other reasons too though

Predicting Fraud

Main question

> How can we *detect* if a firm *is* involved in a major instance of missreporting?

- This is a pure forensic analytics question
- “Major instance of misreporting” will be implemented using AAERs

Approaches to detection

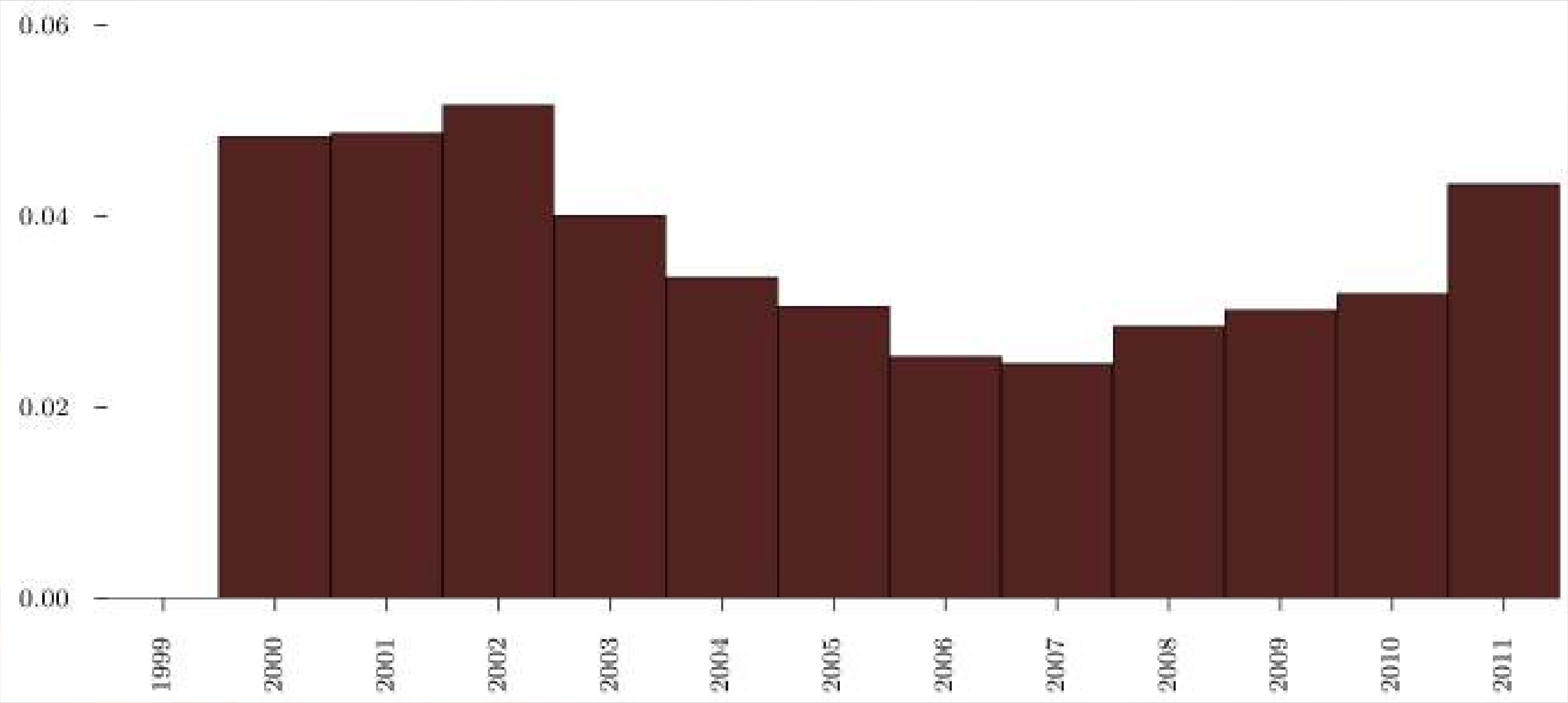
- 1990s: Financials and financial ratios
  - Misreporting firms' financials should be different than expected
- Late 2000s/early 2010s: Characteristics of firm's disclosures
  - How long, how positive, word choice, …
- Late 2010s: More holistic text-based machine learning measures of disclosures
  - Modeling exactly *what* the company talks about in their annual report

> All of these are discussed in Brown, Crowley and Elliott (2018) – I will refer to the paper as **BCE** for short

Changing methods

> Why did we shift away from accounting ratios?

- The old ways of doing fraud were too obvious
- Those committing fraud got smarter



Dealing with infrequent events

- Fraud is infrequent
- A few ways to handle this:
  1. Very careful model selection (keep it sufficiently simple)
  2. Sophisticated degenerate variable identification criterion + simulation to implement complex models that are just barely simple enough
     - The main method in BCE
  3. Automated methodologies for pairing down models (LASSO, XGBoost)
     - Also implemented in BCE

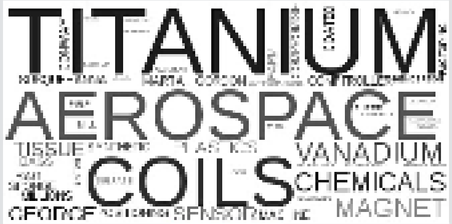
The models

The BCE model

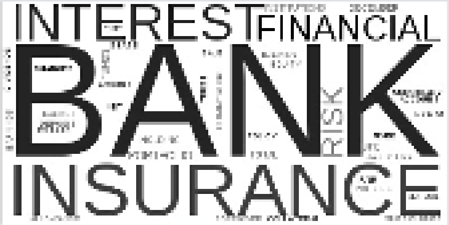
- Retain the variables from the previous models regressions
- Add in a machine-learning based measure quantifying how much documents talked about different topics common across all filings
  - Learned on filings from the 5 years prior
    - Optimal to have 31 topics per 5 years

> Topic

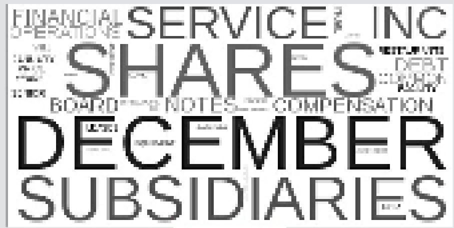
What the topics look like



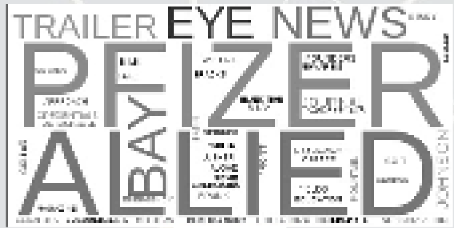
Topic 6



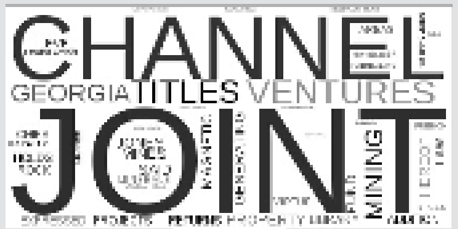
Topic 11



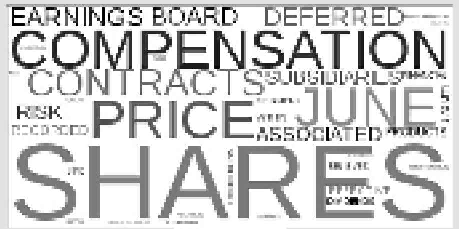
Topic 21



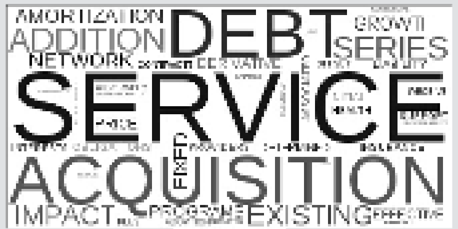
Topic 30



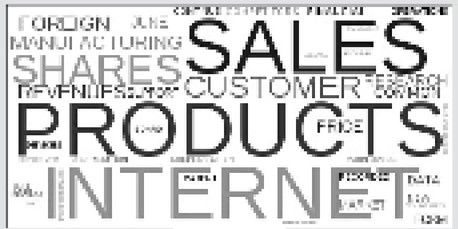
Topic 2



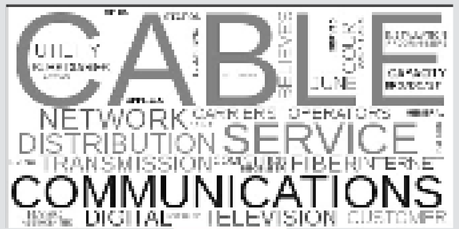
Topic 9



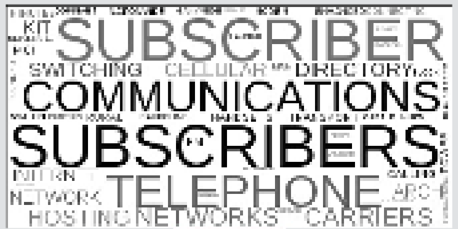
Topic 12



Topic 26



Topic 8



Topic 19

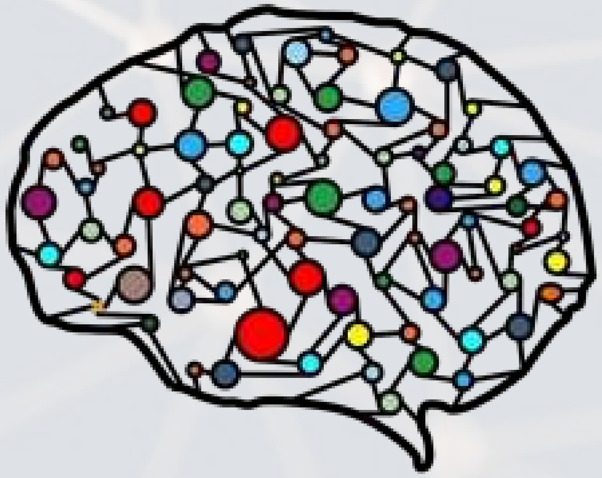
Theory behind the BCE model

- From communications and psychology:
  - When people are trying to deceive others, what they say is carefully picked
    - Topics chosen are intentional
- Putting this in a business context:
  - If you are manipulating inventory, you don't talk about it

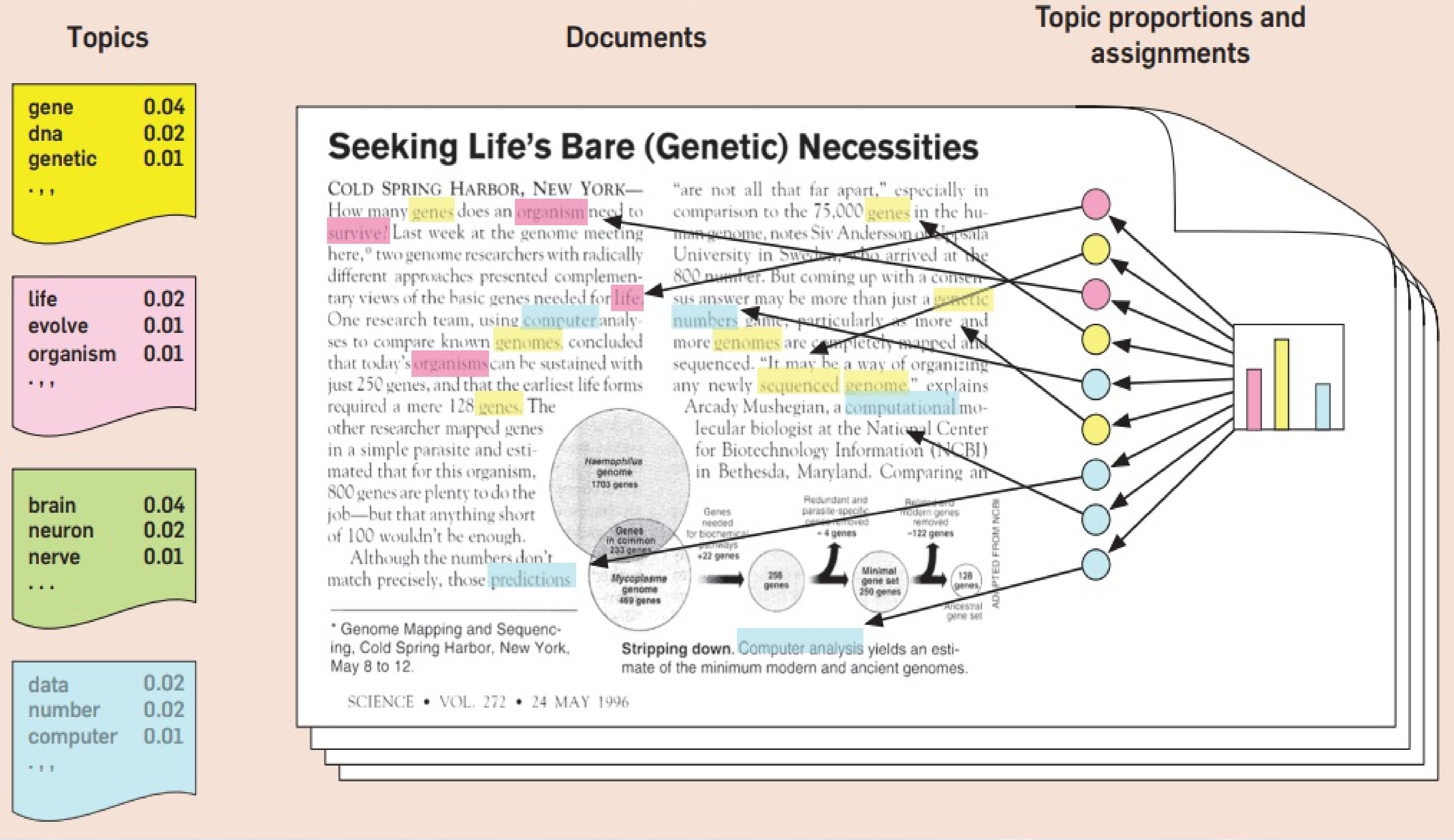
> Think like a fraudster!

How to do this: LDA

- LDA: Latent Dirichlet Allocation
  - Widely-used in linguistics and information retrieval
    - Available in C, C++, Python, Mathematica, Java, R, Hadoop, Spark, …
  - Used by Google and Bing to optimize internet searches
  - Used by Twitter and NYT for recommendations
- LDA reads documents all on its own! You just have to tell it how many topics to find



An example of LDA


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
. . .
life 0.02
evolve 0.01
organism 0.01
. . .
brain 0.04
neuron 0.02
nerve 0.01
. . .
data 0.02
number 0.02
computer 0.01
. . .
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

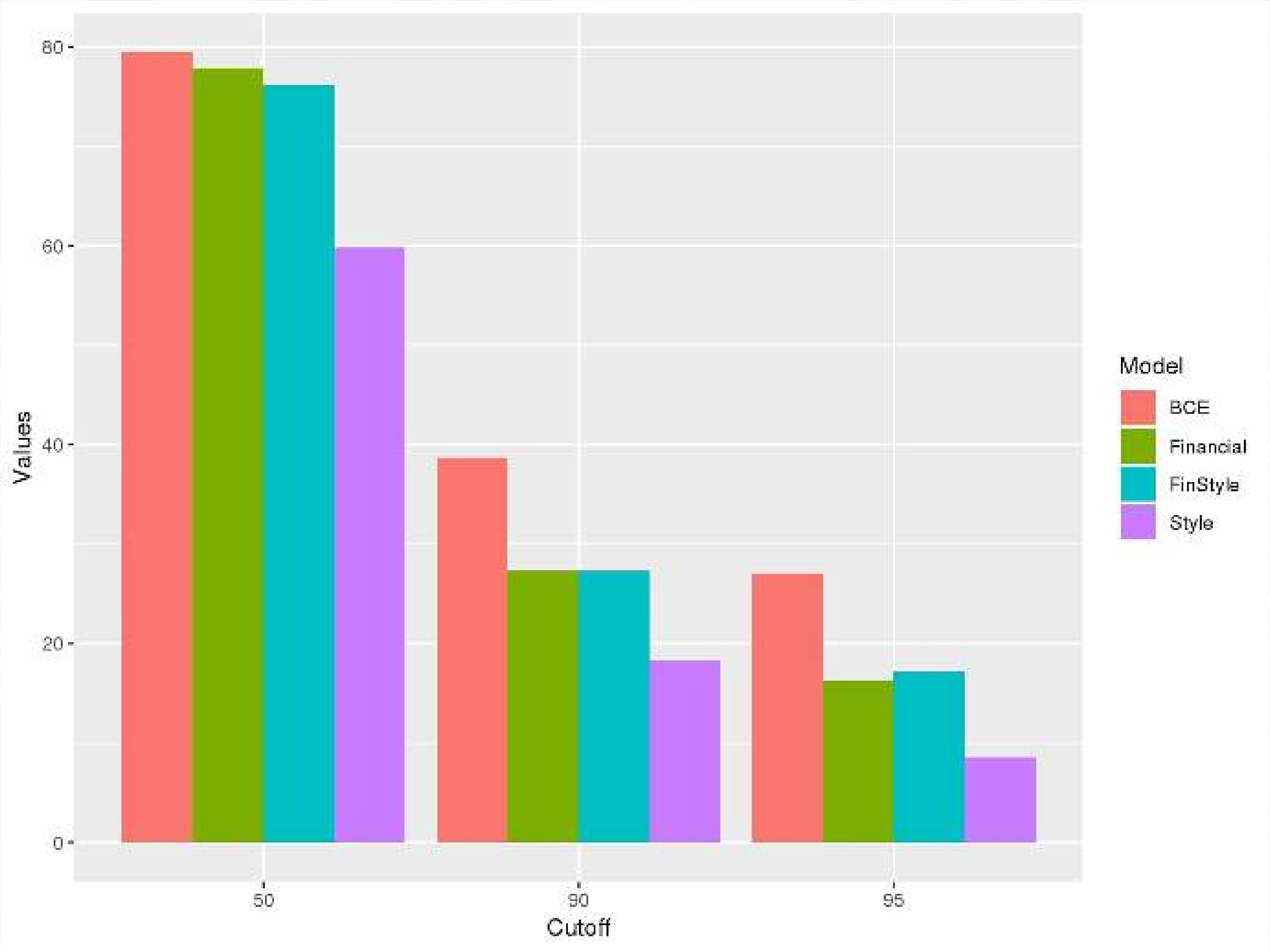
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

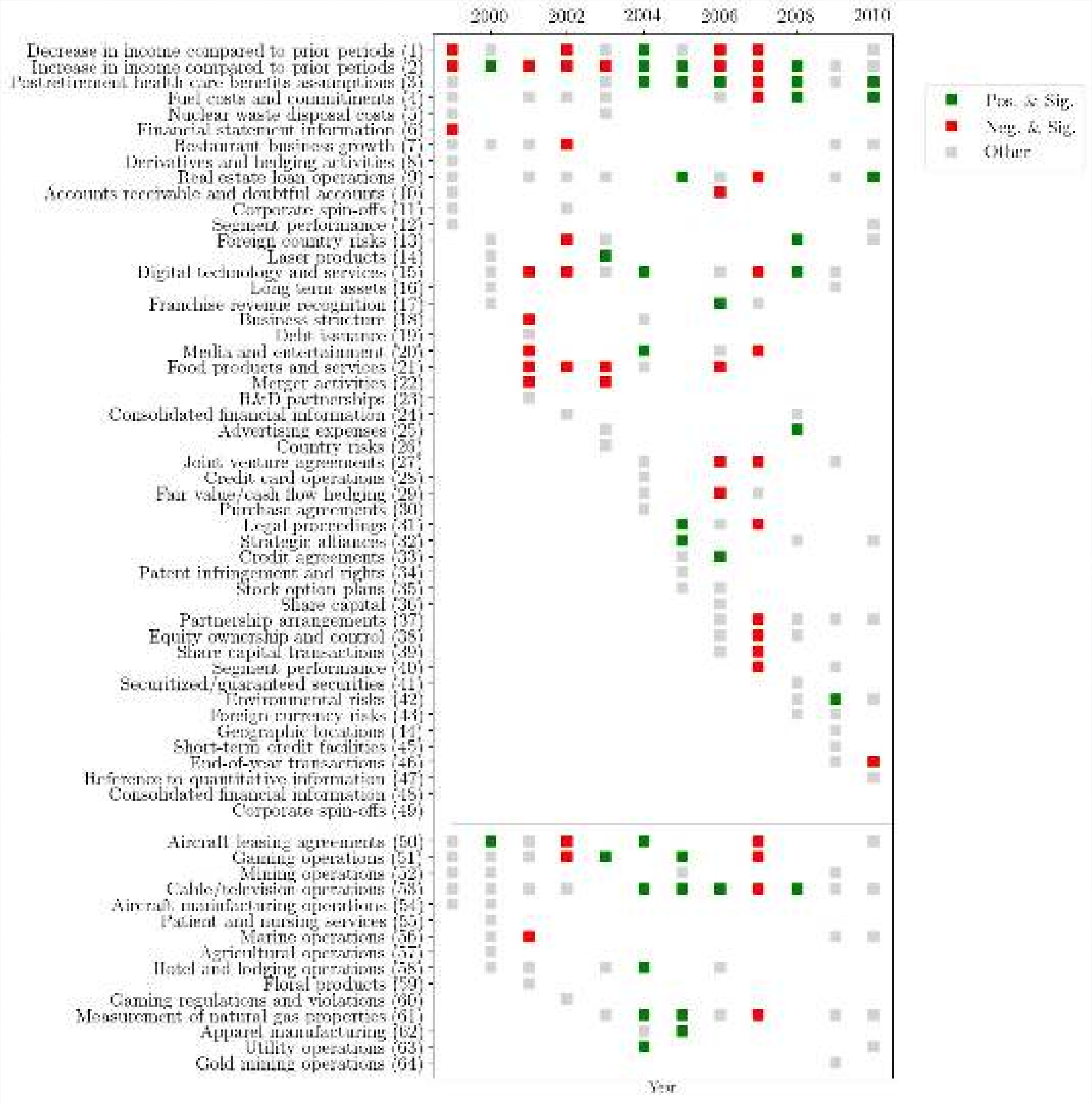
SCIENCE • VOL. 272 • 24 MAY 1996

From David Blei's website

How well does it work?



Topics driving our model



Case studies

ENRON®

- Prediction scores for 1998 and 1999 rank in the 93 and 98 percentiles
- *Increases in Income* topic and firm size are the biggest red flags

ZALES
THE DIAMOND STORE®

- Prediction scores for 2004 through 2009 rank 97 percentile or higher each year
- *Media* and *Digital Services* topics are the red flags
- Our algorithm detects this 4 years before misreporting ceased

End matter

To learn more

- Detail of how, exactly, to build this model will be presented later this month
  - Data Science Singapore (DSSG)
  - March 27, 7:00pm
  - Ngee Ann Kongsi Auditorium
  - Register on meetup.com
- Technical details publicly available at SSRN
- Some other details on rmc.link