A deep dive into Brown, Crowley, and Elliott (2020)

7 May 2021

Dr. Richard M. Crowley
rcrowley@smu.edu.sg
@prof_rmc
rmc.link/iim

Frontmatter

About me

- Assistant Professor of Accounting at SMU since 2016
- **Research**: Approaching accounting disclosure problems using AI/ML
  - Fraud detection based on annual report content
  - Corporate and executive social media posting
  - Fine-grained measurement of context within annual reports
  - WIP: COVID-19 social media discussion
  - WIP: Impact of fake news legistlation
- **Grants**: Singapore, Hong Kong, Canada
- **Research talks**: 20 across 7 countries/regions + 13 discussions
- **Visits**: Toronto and CMU (Accounting); Humboldt (Statistics)
- **Teaching**
  - PhD: Machine Learning for Social Science; Accounting Theory
  - UG: Forecasting and Forensic Analytics; Financial Accounting

Agenda

1. A bit about misreporting to set the stage
2. Idea generation
3. Sketch of paper's results
4. The paper's path to publication
5. Methodology: Machine learning
6. Methodology: Econometrics
7. Extension: Better econometrics through ML
8. Some final Thoughts

Misreporting

Misreporting: A simple definition

> Errors that affect firms' accounting statements or disclosures which were done seemingly *intentionally* by management or other employees at the firm.

Traditional accounting fraud

1. A company is underperforming
2. Someone at the company cooks up some scheme to increase earnings
3. Create accounting statements using the fake information

- Wells Fargo's opening of accounts without customer's consent from 2002-2016 is a standard, though extreme, example
    - Led to a $3B USD settlement with the US government

Other accounting fraud types

- Dell (2002-2007)
  - *Cookie jar reserve* (secret payments by Intel of up to **76%** of quarterly income)
    1. The company is overperforming
    2. “Save up” excess performance for a rainy day
    3. Recognize revenue/earnings when needed to hit future targets
- Apple (2001)
  - *Options backdating*
- China North East Petroleum Holdings Limited
  - *Related party transactions* (transferring 59M USD from the firm to family members over 176 transactions)
- Countryland Wellness Resorts, Inc. (1997-2000)
  - Gold reserves were actually… dirt

Why do we care?

> The 10 most expensive US corporate frauds cost *shareholders* **12.85B USD**

- The above figure is missing:
  - *GDP impacts*: Enron's collapse cost **~35B USD**
  - *Societal costs*: Lost jobs, lost confidence in the economy and government
  - Any *negative externalities*, e.g. new compliance costs borne by others
  - *Inflation*: In current dollars it is even higher

> Catching even 1 major fraud **as they happen** could save billions of dollars

What misreporting measures are there? (US)

1. US SEC AAERs: Accounting and Auditing Enforcement Releases
   - Highlight larger/more important cases, written by the SEC
   - Example: The *Summary* section of this AAER against Sanofi
2. 10-K/A filings ("10-K" ⇒ annual report, "/A" ⇒ 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
     - Note: 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

> Original disclosure motivated by management admission, government investigation, or shareholder lawsuit

Idea generation

Through what lenses can we view misreporting?

- The traditional approach to detecting misreporting is to use financial ratios
  - As such, the traditional lens is an *economic* or *accounting* lens
    - Misreporting $\Rightarrow$ financials are off $\Rightarrow$ Look for suspicious financial ratios

> There are other lenses we can use though!

> Let’s brainstorm a bit!

What lenses do we use?

- Economics
  - Accounting/Finance perspective on the relationship between fraud and accounting figures
- Linguistic
  - Conscious bias from misrepresenting financials leads to potential linguistic artifacts
    - Obfuscating language
    - Sentiment?
- Psychology theory
  - Subconscious bias from misrepresenting financials leads to intentional choices of topics to discuss

What was the original inspiration for the paper?

> Original inspiration was Bayesian spam filtering

- In particular, the idea of using the text of a document to identify documents exhibiting unwanted characteristics
  - I.e., equating spam and misreporting

Wait a minute…

> The inspiration was Naive Bayes, but the paper doesn't use it?

- The inspiration was just on the use of text for fraud detection
- Then we dug into various literatures:
  - Fraud detection
  - Linguistics
  - Psychology

> Based on the above, our original plan was to to apply Naive Bayes to n-grams

Where did LDA come from?

> Discussion with a CS PhD student

- After reading the Blei (2003) paper, it was clear that this method was a better way to to capture what we hoped to capture using Naive Bayes

> LDA is an unsupervised ML approach to quanitfying the content of a document

A sketch of the paper

Main question and approaches

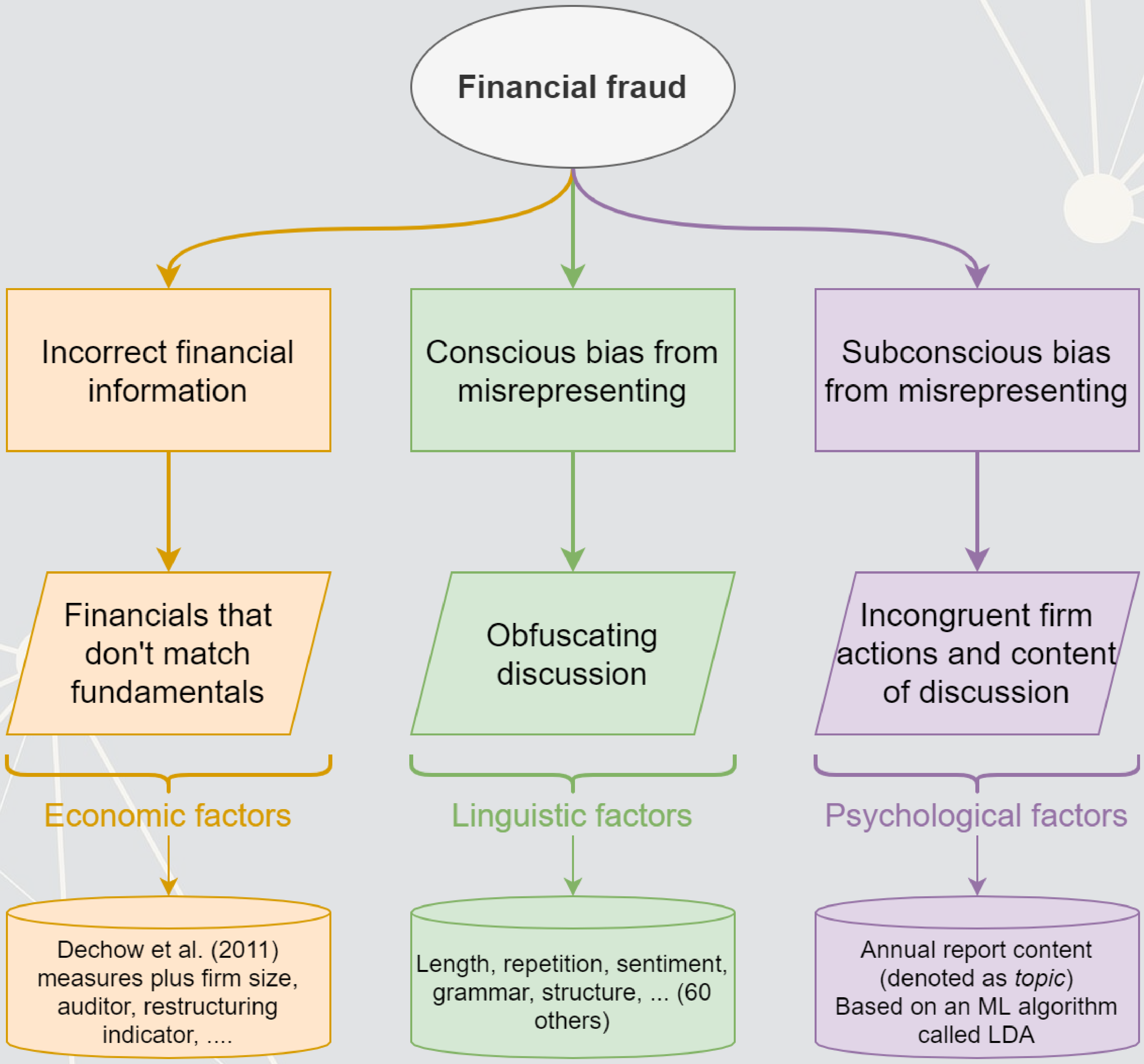
> How can we *detect* if a firm is *currently* involved in a major instance of *misreporting*?

- 1990s: Financials and financial ratios
  - Misreporting firms' financials should be different than expected
- Late 2000s/early 2010s: Characteristics of firm disclosures
  - **Annual report** length, sentiment, word choice, …
- Late 2010s: More holistic text-based ML measures of disclosures
  - Modeling *what* the company discusses in their **annual report**

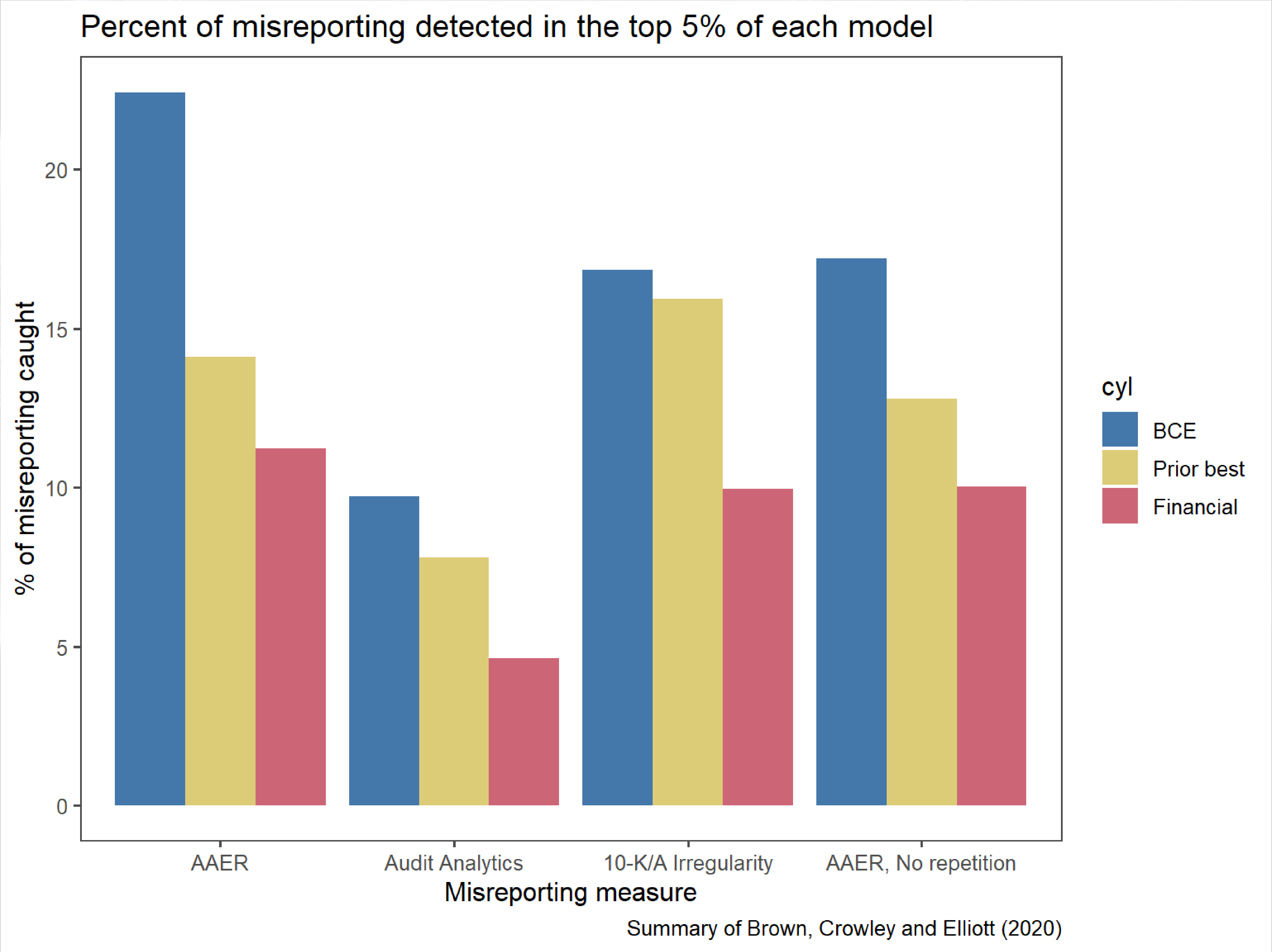
What we need to address:

1. Detecting varied events
   - “Careful” feature selection (via econometrics)
   - Intelligent feature design (partially via ML)
2. For business users… Interpretability matters
   - We develop a psychology-style experiment
     - And a quasi-experiment
3. Predictive model
   - Clean out of sample designs + backtesting
   - Windowed design – data from 1998 won’t help today, but it would in 1999
4. Infrequent events
   - Good for society, bad for modeling
   - Requires careful econometrics

Approach


Financial fraud
Incorrect financial information
Conscious bias from misrepresenting
Subconscious bias from misrepresenting
Financials that don't match fundamentals
Obfuscating discussion
Incongruent firm actions and content of discussion
Economic factors
Linguistic factors
Psychological factors
Dechow et al. (2011) measures plus firm size, auditor, restructuring indicator, ...
Length, repetition, sentiment, grammar, structure, ... (60 others)
Annual report content (denoted as *topic*) Based on an ML algorithm called LDA

Main results



The paper's path

How the paper improved

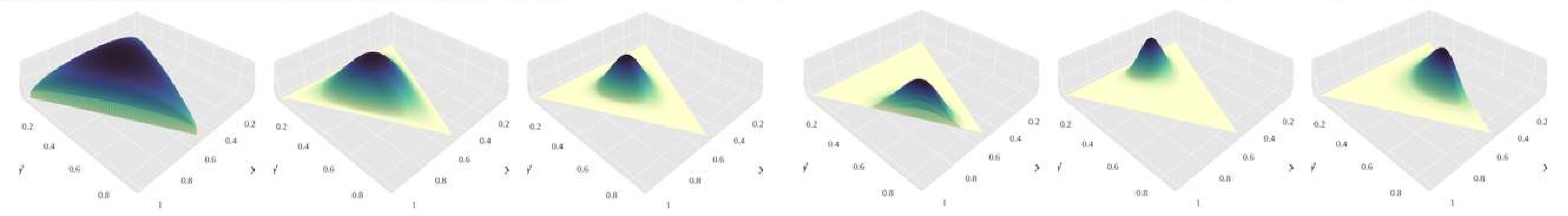
- The first draft only looked at 10-K/A irregularities
  - AAERs were added shortly after
  - Other DVs were added throughout the review process
- The original test statistics only included a Variance-gamma distribution test
  - This test is not present in the final draft
  - Replaced with a bootstrapped ROC AUC comparison
- The validation of our topic measure was relatively light initially
  - Significantly increased in response to workshops and reviewer comments

Methodology: Machine learning

> Let’s see where we all are at

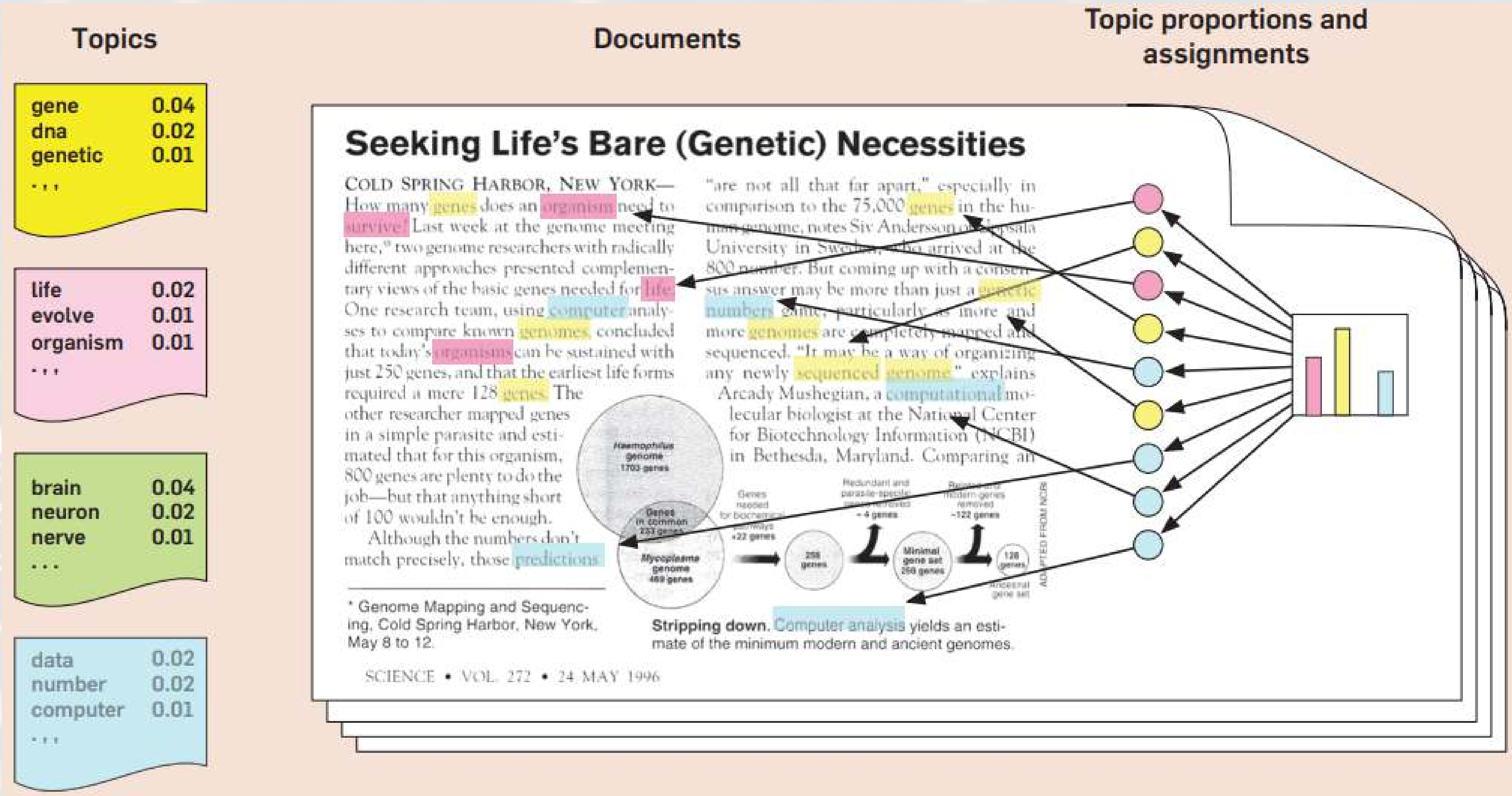
What is LDA?

- **L**atent **D**irichlet **A**llocation
- One of the most popular methods under the field of *topic modeling*
- LDA is a Bayesian method of assessing the content of a document
- LDA assumes there are a set of topics in each document, and that this set follows a *Dirichlet* prior for each document
  - Words within topics also have a *Dirichlet* prior



More details from the creator

A simple LDA example:



Source: Blei 2012

How does it work?

1. Reads all the documents
   - Calculates counts of each word within the document, tied to a specific ID used across all documents
2. Uses variation in words within and across documents to infer topics
   - By using a Gibbs sampler to simulate the underlying distributions
     - An MCMC method

- It boils down to a system where generating a document follows a couple rules:
  1. Topics in a document follow a multinomial/categorical distribution
  2. Words in a topic follow a multinomial/categorical distribution
- Use words' covariance within and across documents to back out topics in a Bayesian manner

> Caveat: Need to specify the number of topics *ex ante*

What the topics look like

Topic 6

Topic 11

Topic 21

Topic 30

Topic 2

Topic 9

Topic 12

Topic 26

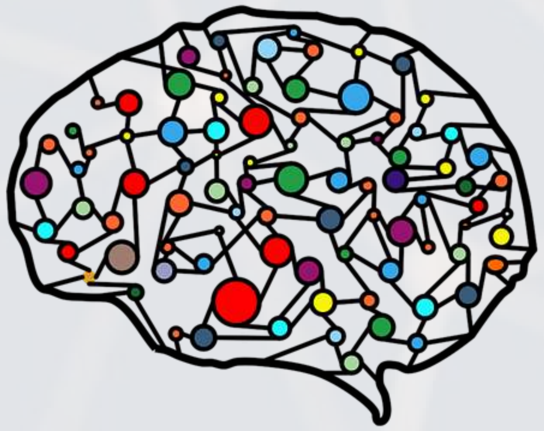
Topic 8

Topic 19

> An interactive illustration of a 10 topic model

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, …
    - We used `onlineldavb`
    - `Gensim` is great for python; `STM` is great for R
  - 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



Implementation details

> The usual addage that data cleaning takes the longest still holds true

1. Annual reports are a mess
   - Fixed width text files; proper html; html exported from MS Word…
   - Embedded hex images
   - Solution: Regexes, regexes, regexes
     - Detailed in the paper’s web appendix
2. Stemming, tokenizing, stopwords
3. Feed to LDA
4. Tune hyperparameters (# of topics is most crucial)
   - Tune this by maximizing in-sample prediction ability
5. Finally implement the model

Other considerations

1. LDA provides the *weight* on each topic, but documents vary a lot by length
   - Solution: Normalize to a percentage between 0 and 1
2. There is a mechanical component to topics due to firms' industries
   - Solution: Orthogonalize topics to industry
     - Run a linear regression and retain $\varepsilon_{i,firm}$:

$$topic_{i,firm} = \alpha + \sum_j \beta_{i,j} Industry_{j,firm} + \varepsilon_{i,firm}$$

LDA Validation

- LDA is well validated on general text, no question
- One key is to present some details of the topics to ensure comfort
- Another key is having prior evidence to fall back on
  - Whether LDA works on business-specific documents is not so well studied
    - Most studies ask people whether they agree with the hand-coded topic categorizations
    - Need evidence that the topics are separable coherently

> We decided to fill this gap (after some nudging)

Experimental design

Instrument: A word intrusion task

- Which word doesn't belong?

1. Commodity, Bank, Gold, Mining
2. Aircraft, Pharmaceutical, Drug, Manufacturing
3. Collateral, Iowa, Residential, Adjustable

Participants

- 100 individuals on Amazon Turk (20 questions each)
  - Human but not specialized

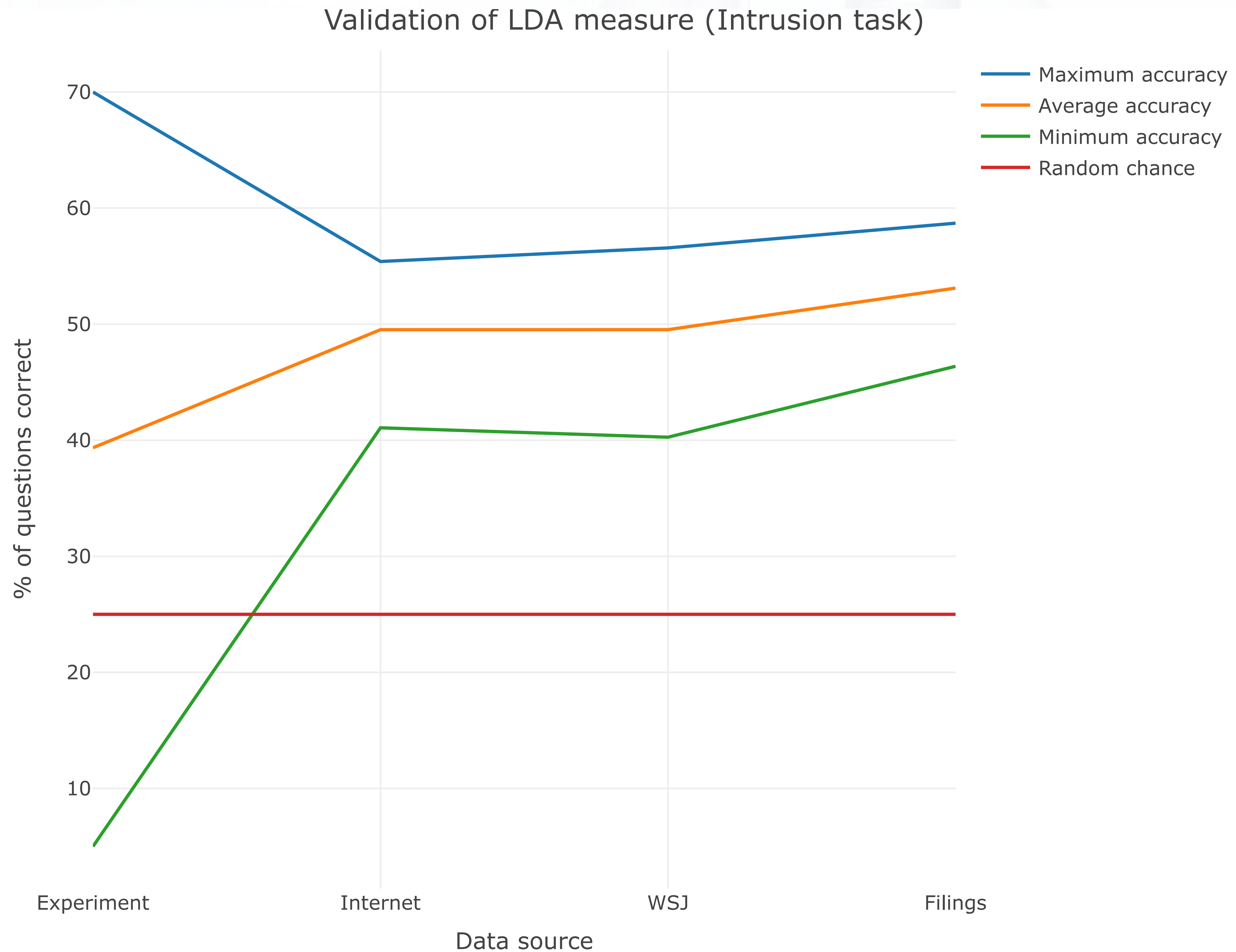
Quasi-experimental design

- 3 Computer algorithms (>10M questions each)
  - Not human but specialized
  1. GloVe on general website content
     - Less specific but more broad
  2. Word2vec trained on Wall Street Journal articles
     - More specific, business oriented
  3. Word2vec directly on annual reports
     - Most specific

> These learn the “meaning” of words in a given context

> Run the *exact same* experiment as on humans

Experimental results



Related constructs

	Scale	Construct	Precision
Topic	Document level (can go as fine as paragraph)	Distribution of content	Noisy but captures all information
Word counts	Document level (can go as fine as sentence)	Index-dependent	Only captures what the researcher is aware of
Specific phrase mentions	Document level (can do any subset)	Scale	Precise if construct is well defined AND terminology is unique
Embedding methods	Word, sentence, and document available	Scale	Noisy unless used to estimate 1 outcome in a supervised manner
Context (by itself)	Clause level	Content at clause level	Less precise than LDA for large documents, better for small snippets
Context (paired with word counts)	Clause level	Word counts' meaning	Precise at capturing content dependency, some noise in measurement

Methodology: Econometrics

Past models

Financial model based on [Dechow, et al. (2011)]

- 17 measures including:
  - Log of assets
  - % change in cash sales
  - Indicator for mergers
- Theory: Purely economic
  - Misreporting firms' financials should be different than expected

Textual style model based on various papers

- 20 measures including:
  - Length and repetition
  - Sentiment
  - Grammar and structure
- Theory: Linguistics
  - Style reflects complexity and unintentional biases
  - Some measures ad hoc

> We tested an additional 26 financial & 60 style variables

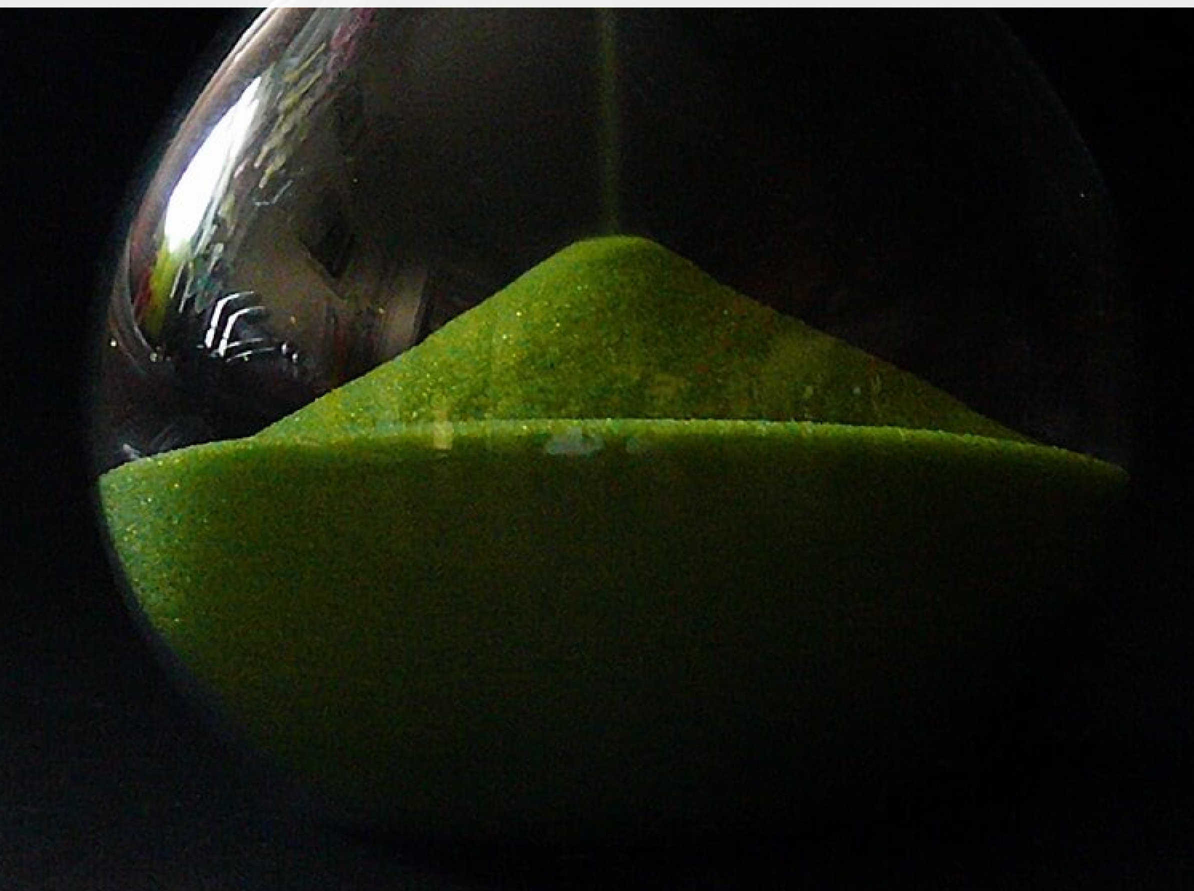
The BCE model

1. Retain the variables from the previous models' regressions
   - Forms a useful baseline
2. Add in our *topic* measure to quantify how much each **annual report** (~20-300 pages) talks about different *topics*
   - We train this on 5-year windows
     - Balance data staleness, data availability, and quantity of text
     - Optimal to have 31 topics per 5 years
       - Based on in-sample logistic regression optimization

Backtesting

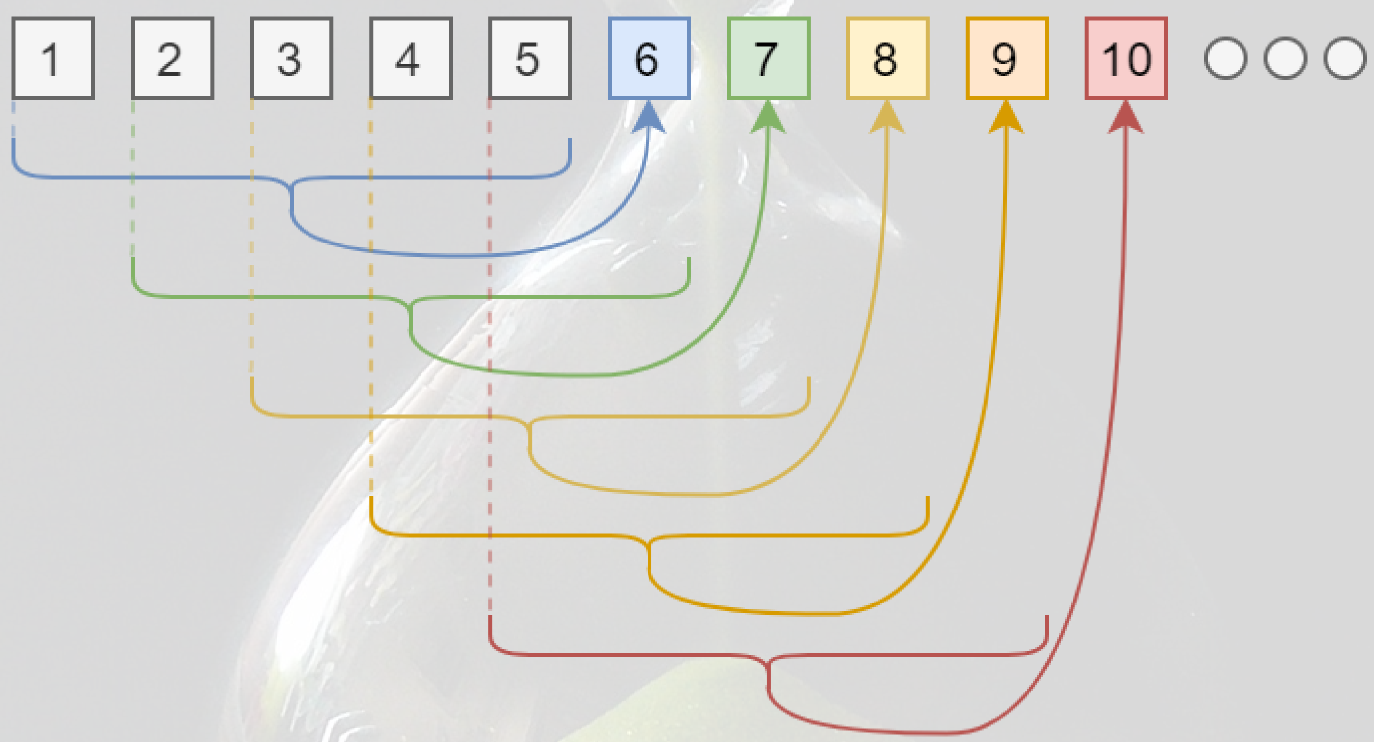
> We don't know who is misreporting today

- So, we will backtest
  - Use historical contexts to validate our model
- Problems:
  1. Misreporting changes over time
  2. Misreporting is unobservable (until it's observable)



Moving target

- Implement a moving window approach
  - 5 years for training + 1 year for testing
  - We use data from 1994 through 2012: 14 possible windows
- Ex.: to predict misreporting in 2010, train on data from 2005 to 2009



> Problem: Now we have 14 models…

Observability

- The other issue is that, as of a given year, say 2009, we do not know every firm that was misreporting
  - We could build an algorithm with perfect information, but it may fall flat on current, noisy data!
  - It could also give us a false impression of an algorithm's effectiveness when backtesting
  - Misreporting can take a long time to discover: Zale's started in 2004, finished in 2009, and was disclosed in 2011!

> Solution: Censor our data to what was known at that time

- Use data on when a misreporting case was first disclosed
  - If the fraud wasn't known by the end of the window, train as if that was 0 (as it was unobservable back then)
  - Mimics our current situation

Dealing with infrequent events

- Fraud is infrequent
  - E.g.: Out of 37,806 firm-years of data, there are 505 firm-years subject to AAERs
- Key issue: We may have more variables than events in a window…
  - Even if we don't, convergence is iffy using a logistic model
- 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)
     - Quite promising

Degenerate variable identification

1. Toss every input into a model
2. Check independentness using a QR decomposition
   - This will let us determine an order for dropping inputs
   - $A = Q \times R$, where A is our feature matrix, Q is an orthogonal matrix, and R is the transformation
     - More weight on the diagonal element in R means more independent (effectively)
     - Same underlying method as a Gram-Schmidt process
3. Remove excess inputs if too few 1s
   - Why? Because logit can't converge if there are more inputs than events (or non-events) in the data

> Independentness is a useful criterion for removing features with lower likelihood of being useful

Logistic iteration

1. Run a logit using a Newton-Raphson solver for 50 iterations
2. Check convergence for signs of quasi-completeness
   - Standard errors will be in the millions if quasi-complete
   - If quasi-complete, drop the next least independent variable and restart
3. Run a 500 iteration logit using a Newton-Raphson solver
4. Recheck convergence
   - If failed, drop the next least independent variable and restart

> We will essentially get the most complex feasible model with the most independent set of features

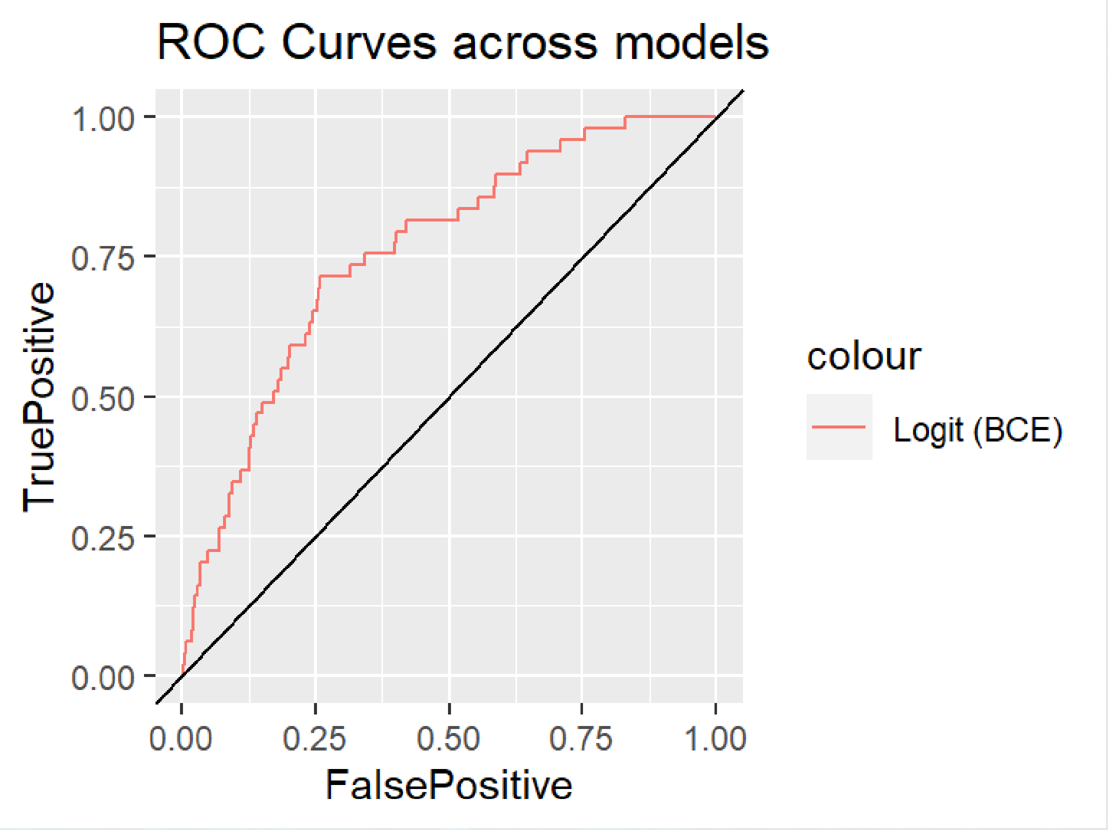
Comparing multiple models

- Performance measures:
  1. **ROC AUC**
  2. Fisher statistics
  3. Performance at a reasonable cutoff (5%)
  4. `NDCG@k` (usually used in ranking problems)

> ROC AUC and Fisher statistics also allow us to statistically compare across models

ROC AUC for windowed approaches

- **R**eceiver **O**perator **C**urve
  - ROC curve compares sensitivity and specificity of a model
    - Sensitivity: True positive rate
    - Specificity: True negative rate



- **A**rea **U**nder the **C**urve
  - What is the probability that a randomly selected `misreport=1` is ranked higher than a randomly selected `misreport=0`
  - A good score is above 0.70

Comparing with ROC AUC

- Can aggregate ROC AUCs via pooling predictions together
  - With clustering by year
- Higher aggregate AUC is better, but direct comparison is tricky
- Bootstrapping allows for generating test statistics for ROC AUCs, which can be compared with a Wald test
  - Available in Stata as part of `rocreg`

Comparing with Fisher Statistics

- Fisher (1934) provides a solution to aggregating p-values into a X^2 test statistic

$$-2\sum_{i=1}^{k}\log\left(p\text{-}value_i\right) \sim X^2_{2k}$$

- The difference of X^2 distributed variables follows a Variance Gamma distribution
- For 2 Fisher statistics X_1 and X_2 each with k observations:

$$\mathbb{P}\left(X_1 < X_2\right) = \int_{-\infty}^{X_1-X_2} \frac{1}{2^k\sqrt{\pi}\Gamma(k)}|z|^{k-1/2}K_{k-1/2}(|z|)dz$$

- where Γ is the gamma function and $K_{k-1/2} is the modified Bessel funciton of the second kind

Other methods of measuring performance

> NDCG @k: **N**ormalized **D**iscounted **C**umulative **G**ain @k

- Measures *ranking* quality, used for search engine optimization
- k is a specified percentile or # of observations
- “DCG” measures the # of true positives in k of the prediction score
- “N” is to divide the DCG by the theoretically optimal DCG to normalize to a [0,1] interval

> Counts at different thresholds

- E.g., at a 95% cutoff, the BCE model captures 96 AAERS, whereas traditional models only capture 70
- Easy to interpret economically
- Maps well to what regulators do in practice

Extension: Better econometrics

Augmenting our statistical analysis

- Traditionally, binary classification problems in statistics are solved using logistic regression
  - This is what we saw in the previous example

Pros of logistic regression

- Regression approaches are familiar
- Easy to run
  - You could even do it in Excel
- Easy to interpret

Cons of logistic regression

- Logistic regression handles *sparse* data poorly
- Ideally you want at least 10% of your data in each group
- Fraud is sparse!

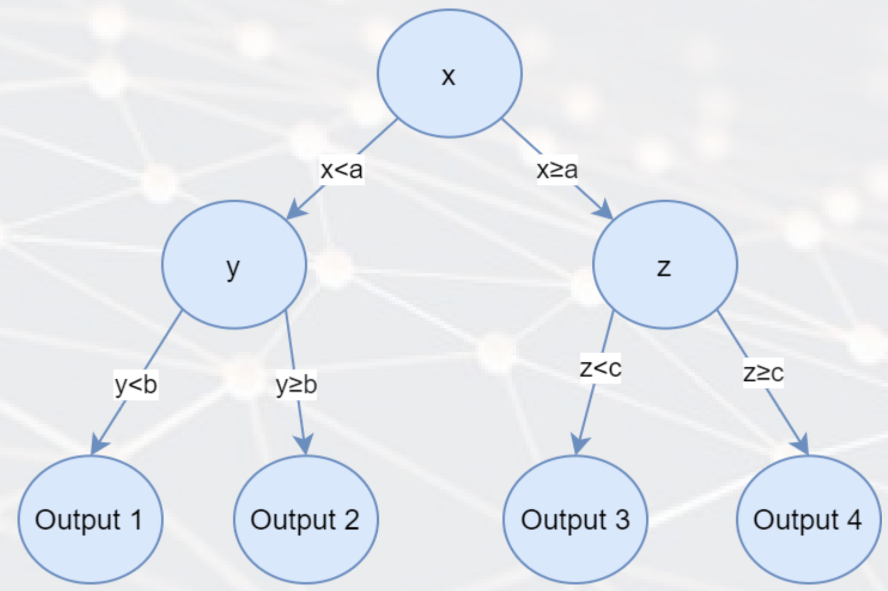
> If we want a better accuracy, we need to replace logistic regression

How ML helps with sparsity

- Certain machine learning methods are less sensitive to sparsity
  - Ensembled decision trees are one example

Decision trees

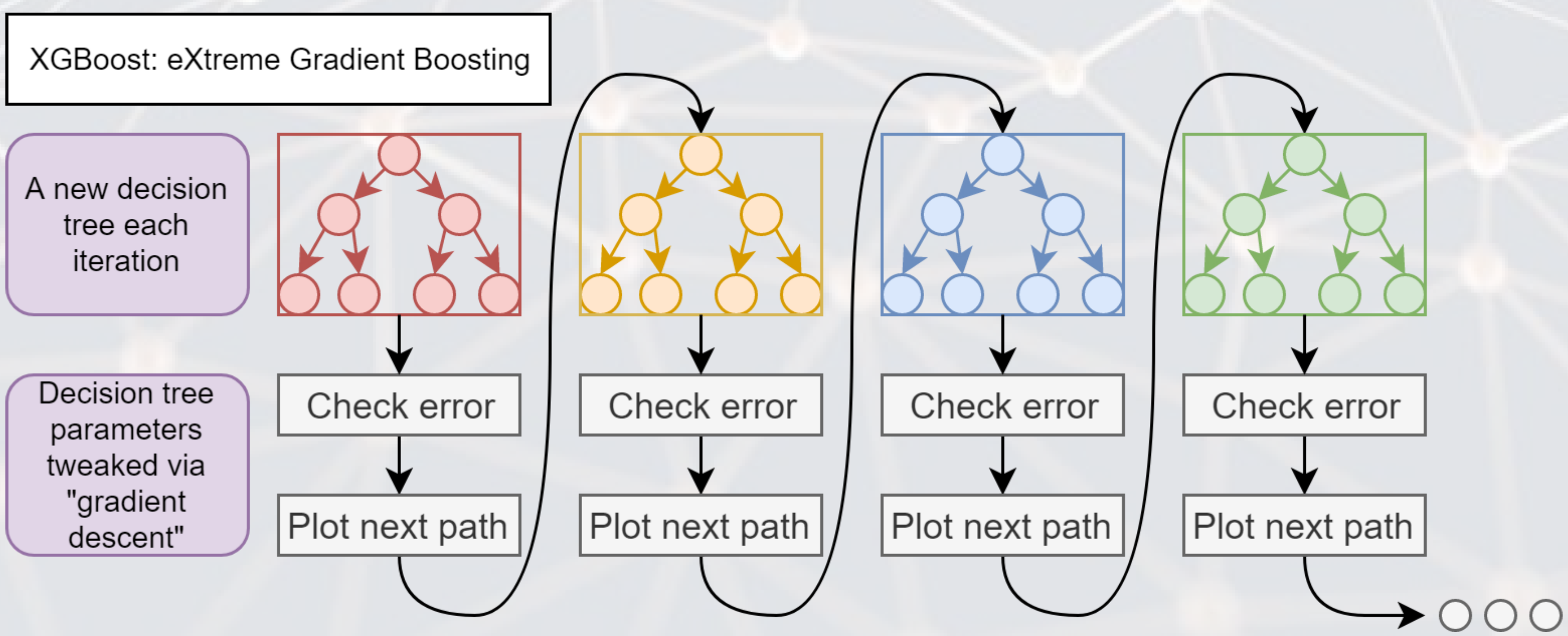
- Traverse from top to bottom
- Consider the impact of individual inputs…
  - If input is higher than X, what should we do?
  - If input is lower than X, what should we do?



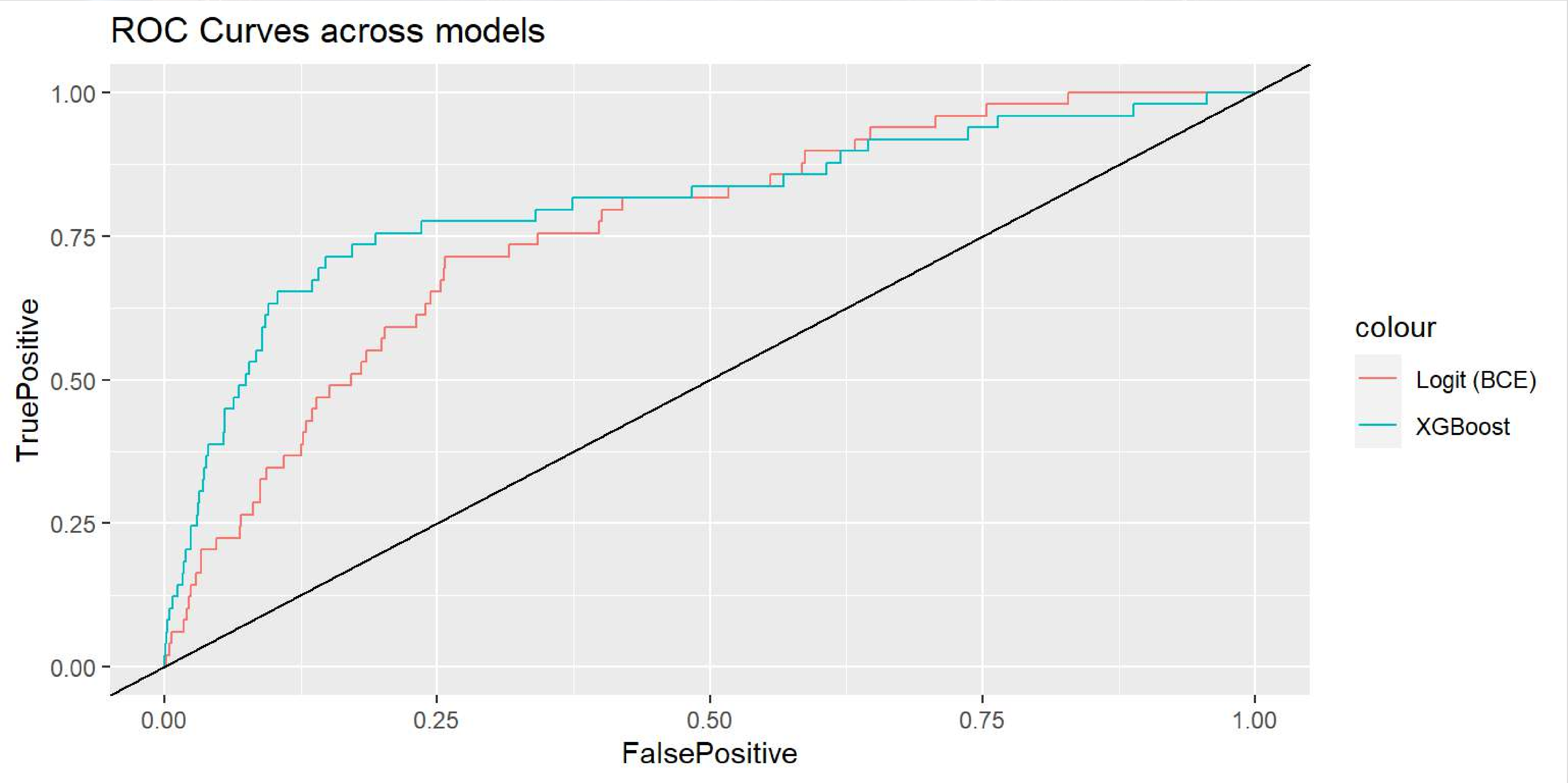
> We can combine a bunch of decision trees

A specific implementation: XGBoost

- e**X**treme **G**radient **B**oosting
- A simple explanation:
  1. Start with 1 or more decision trees & check error
  2. Make more decision trees & check error
  3. Use the difference in error to guess a another model
  4. Repeat #2 and #3 until the model’s error is stable



Prediction comparison: 2004



- AUC for standard BCE model: 0.76
- AUC for XGBoost BCE model: 0.81

Conclusion

Some ways to improve our model

1. Use a better tokenizer such as spaCy
   - Our tokenizer didn’t detect noun phrases
2. Use econometric (ML) methods that are better suited for sparsity
   - E.g.: XGBoost as shown earlier
3. Consider other lenses that we didn’t include
4. Consider examining text at a more precise scale than document-level?
5. Consider examining other sources of information than the annual report

> Final note: The motivation behind our work was not to build a better mousetrap, but to illustrate the usefulness documents’ content to better understand company/manager behavior

Thanks!

Dr. Richard M. Crowley
rcrowley@smu.edu.sg
@prof_rmc
rmc.link/iim

Packages used for these slides

- kableExtra
- knitr
- revealjs
- ROCR
- tidyverse