Demystifying analytics, AI, and machine learning

Dr. Richard M. Crowley

rcrowley@smu.edu.sg

Slides available at https://rmc.link/GIC

Front matter

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
  - Fine-grained measurement of context within annual reports
  - Corporate and executive social media posting
  - Impact of fake news legislation on capital markets and social media
  - Policy implications from COVID-19 social media discussion
- **Teaching**
  - PhD: Machine Learning for Social Science; Accounting Theory
  - UG: Forecasting and Forensic Analytics; Financial Accounting

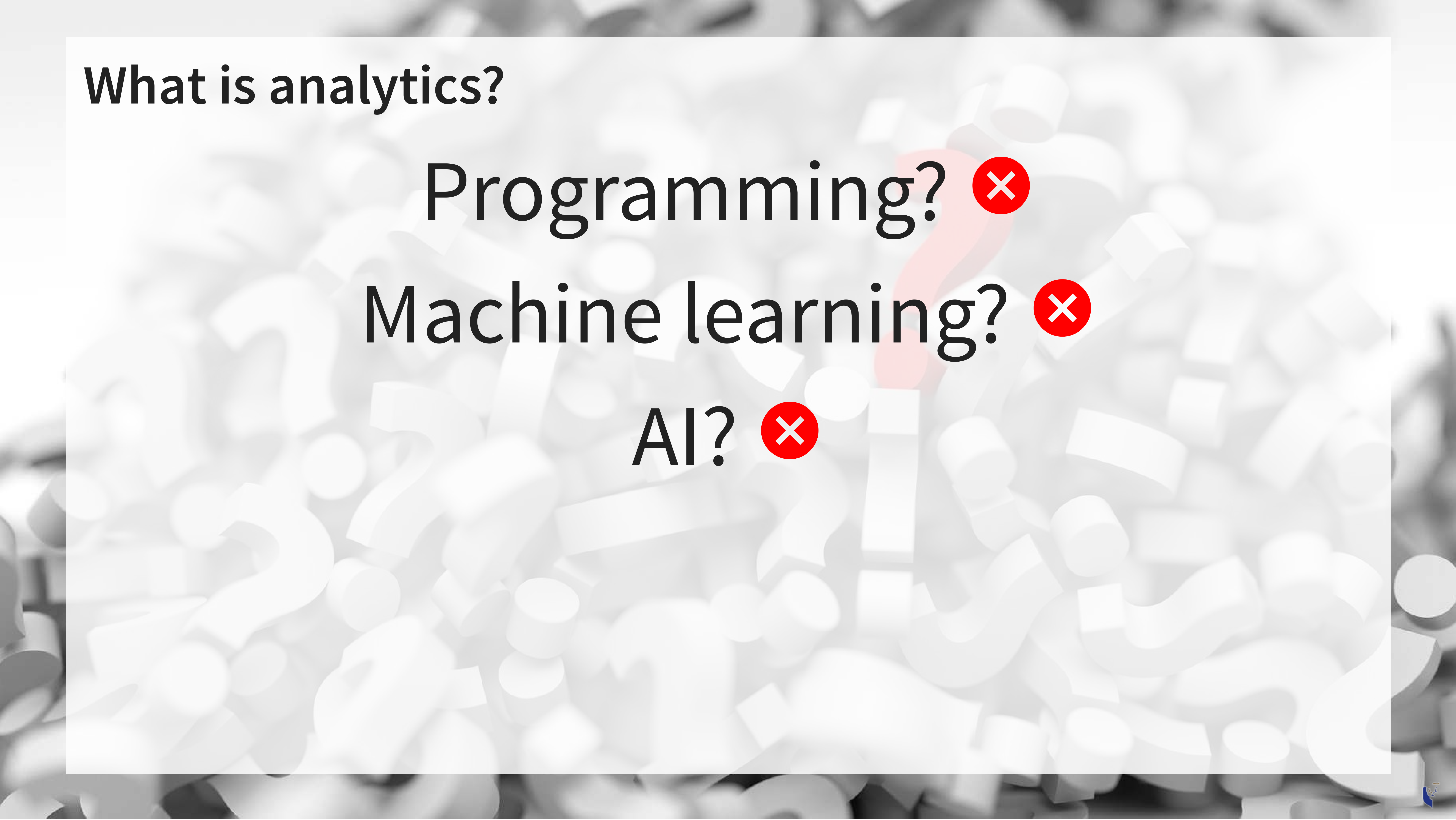
> General focus: How non-quantitative information flows in economies and society, with an eye to methodology (NLP; ML for econometrics)

Today’s objectives

1. Overview of data analytics
   - What is analytics?
   - Approaches to data analytics
   - AI and ML
2. Applications of analytics
   - Modern statistics and portfolio allocation
   - Fraud detection and machine learning
   - Working with text data: Transformer models

Data science

What is analytics?



Programming? ✖

Machine learning? ✖

AI? ✖

What is analytics?

> “The systematic computational analysis of data or statistics
>
> — **Oxford Dictionary**

> “The method of logical analysis
>
> — **Webster’s Dictionary**

> “Catch-all term for a variety of different business intelligence […] and application-related initiatives
>
> — **Gartner**

A simple definition

[Data] Analytics is simply **answering questions using data**

Refining our definition

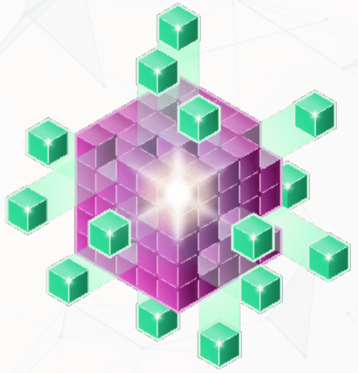
Analytics: Answering questions using data

Requirements

- A question
- Logic
- Reasoning
- Data

Tool box

- Univariate analysis
- Visualization
- Statistics and econometrics
- AI and machine learning

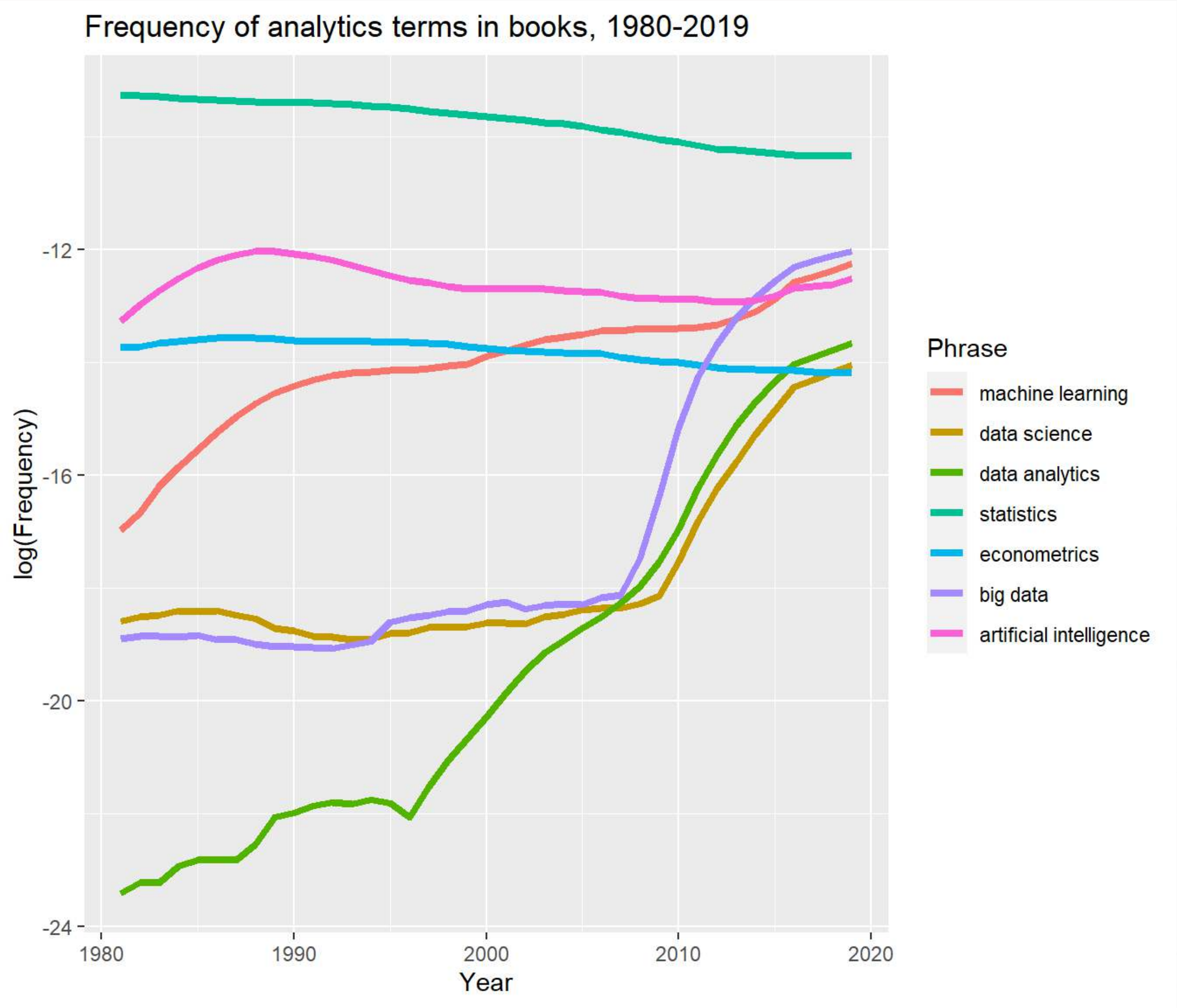


Embellishing our definition

We can add many different layers to this definition:

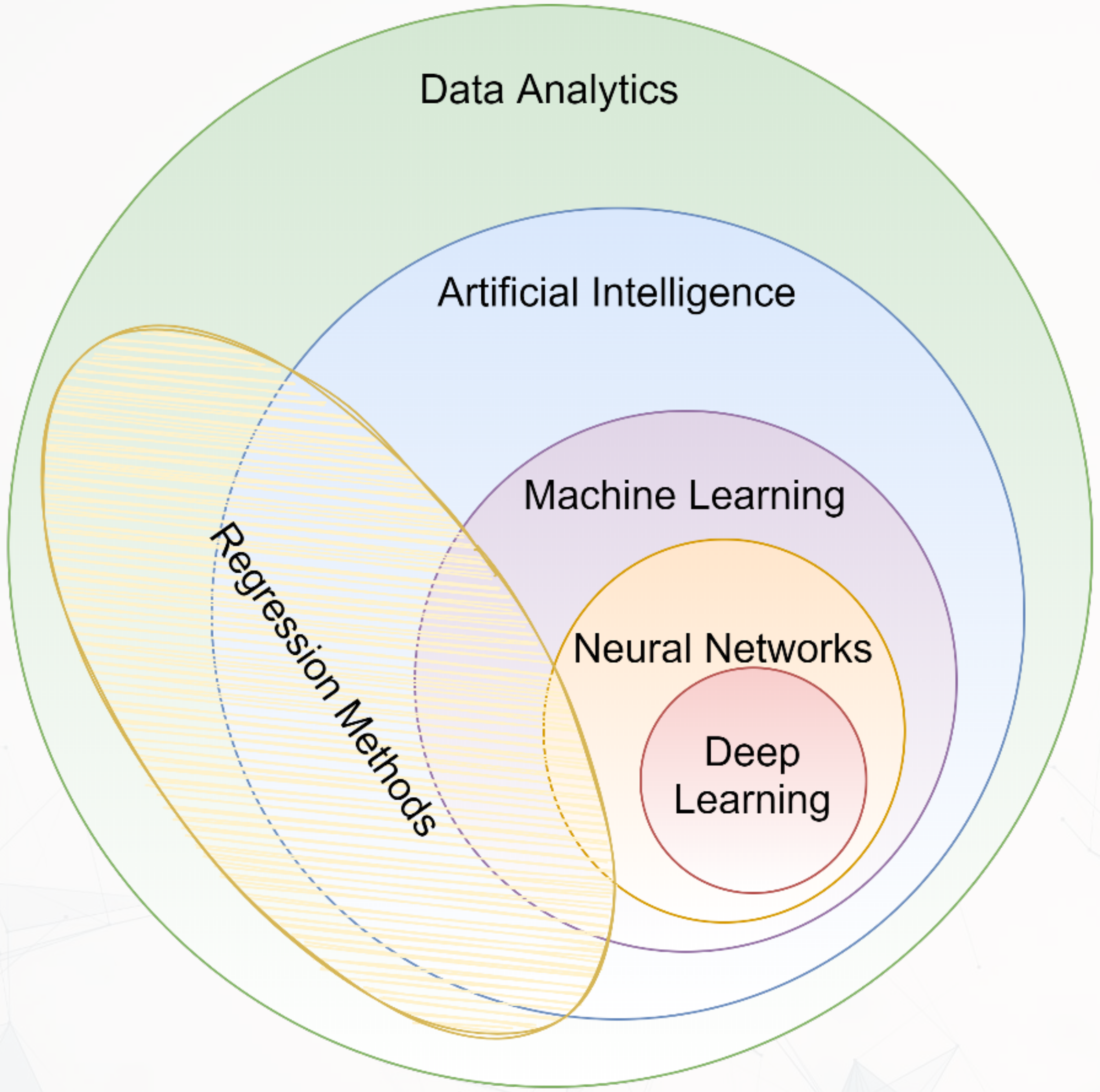
Answering questions using …

- data *and computers*
- *a lot of* data
- data *and statistics* or *econometrics*
- data *and ML* or *AI*



Made using `package:ngramr`

Analytics methods



- AI encompasses anything approaching *intelligence* by a computer
- ML Requires *learning* by a computer to be a part of the method
- Neural networks are a class of algorithms within ML
- Deep neural networks are simply more complex in terms of number of layers to the algorithm

Statistics vs machine learning

Benefits of statistics

- Well understood mathematical properties
  - Some methods are very reliable under set assumptions
- Has a focus on estimators and functional forms
- Bayesian approaches offers flexibility in estimation

Benefits of ML

- Can better handle sparse outcome
- Can better handle large numbers of [correlated] inputs
- Flexible measure creation for unstructured data
- Nonparametric methods are more approachable

⚠ **Takeaway**

Both statistics and machine learning are tools in the analytics tool box. Just because it is in fashion to use machine learning doesn’t mean you should. Instead, use the tool that best fits the job.

Application of statistics: Portfolio balancing

The problem

> How to allocate funds across a selection of stocks in an efficient manner?

- E.g. if you have $1M to allocate across 50 assets, how do you decide
  1. Which assets to invest in
  2. How much funds to allocate to each asset

Markowitz / mean-variance portfolios / modern portfolio theory / MPT

The go-to approach in finance is to allocate based on expected return, risk, and one’s preferences over return and risk.[1]

- But does it work in practice?

[1] Markowitz, H. “Portfolio Selection” J Finance (1952).

Practicalities of modern portfolio theory

- If you only need to allocate a portfolio once, yes!
  - Subject to critiques over differences in preferences over upside versus downside risk

> What if you want to allocate multiple times, such as daily?

- Traditionally, modern portfolio theory is purely a one-shot statistical exercise
  - It does not factor in your prior portfolio
  - Any overlap across allocations is due to shared statistical properties over time
  - If you iterate it, you often completely change portfolios each time

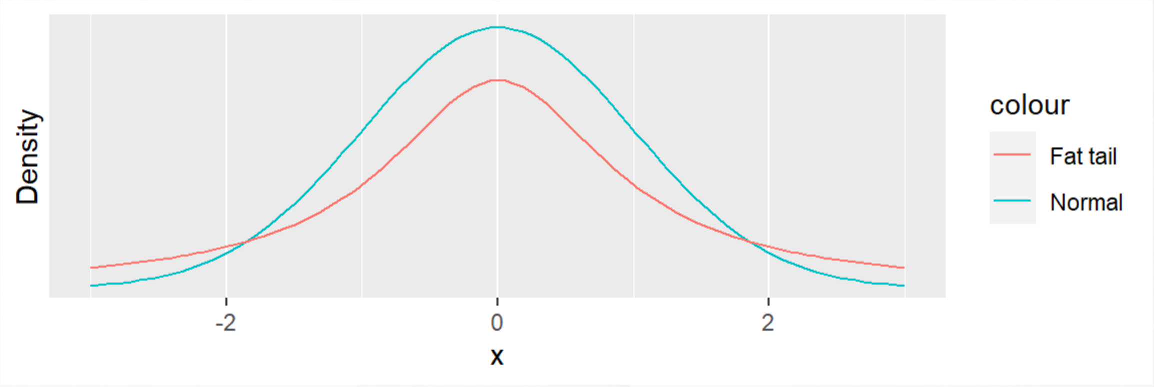
> ! **The big problem**
>
> This leads to excessive rebalancing and thus excessive transactions costs. It isn’t practical.

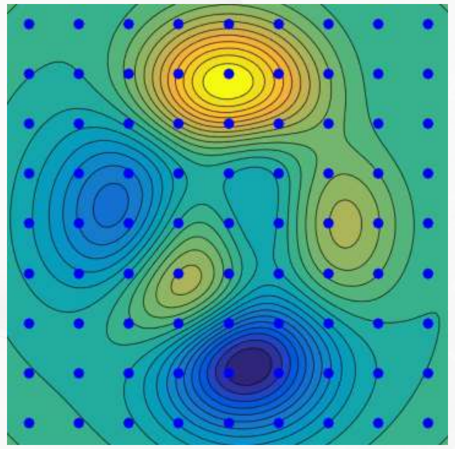
A solution with statistics and data

> "Robustifying Markowitz" (2023)[1]

Components of the solution

- Median-of-means
  - Helps with fat tails
- Robust estimators
  - Helps avoid rebalancing
- Gradient descent
  - Often used for optimizing ML algorithms





Gradient descent video is from: https://en.wikipedia.org/wiki/Gradient_descent

[1] Petukhina, Klochkov, Hardle, and Zhivotovskiy. "Robustifying Markowitz." Journal of Econometrics (2023).

Why is the solution so recent?

- The core methodology was developed by 1983[1]
  - But the mathematical complexity made it impractical to implement
- Some breakthroughs came in 2020[2]
  - Allowed for tractable computation of the measure on smaller data sets
  - Not particularly feasible on large sets of assets or long time horizons
- The “Robustifying Markowitz” paper has three main points:
  - Their approach is computationally reasonable on large data sets
  - Their approach is tractable when the number of assets $>>$ number of days of data
  - Empirically their approach leads to allocations with more desirable properties

[1] Nemirovsky and Yudin. “Problem Complexity and Method Efficiency in Optimization.” (1983).

[2] Hopkins. “Mean estimation with sub-Gaussian rates in polynomial time.” The Annals of Statistics (2020).

An empirical example

> Suppose you are investing in the Russel 3000. You are investing in 600 stocks at a time, rebalancing monthly[1]

	Turnover	Active TO	Net Sharpe Ratio
Robust version	40.51%	27.58%	4.44%
Traditional, linear	127.22%	144.97%	2.94%
Traditional, nonlinear	91.75%	102.25%	3.08%
Traditional, long only	19.73%	19.00%	4.05%
Equal weighting	26.81%	0%	2.42%

[1] Estimates are computed from 1 Jan 2002 through 31 Dec 2020

An empirical example

> How does the allocation look, per asset?

	Lowest	Highest	Std. Dev.
Robust version	-0.69%	1.04%	0.27%
Traditional, linear	-2.90%	4.30%	0.95%
Traditional, nonlinear	-1.85%	3.19%	0.67%
Traditional, long only	0.00%	13.83%	1.00%
Equal weighting	0.17%	0.17%	0.00%

Takeaway

The robust version derived from modern statistics theory significantly lowers turnover, lowering transaction costs and improving risk-adjusted return. It also leads to more diversified portfolios by not overweighting individual assets.

Application of machine learning with numeric data: Fraud detection

The problem

> How can we leverage firms' annual reports to detect misreporting[1]

- **Misreporting**: errors that affect firms' accounting statements or disclosures which were done seemingly *intentionally* by management or other employees at the firm.
  - We won't focus on small accounting errors or mistakes

Traditional misreporting

1. Company is under-performing
2. Someone initiates a scheme to increase earnings
3. Accounting statements are disseminated using fake information

Other types

- Cookie jar reserves
- Options backdating
- Related party transactions (that violate governance requirements)
- Bribery

[1] This discussion follows from Brown, Crowley and Elliott (2020) Journal of Accounting Research as well as my Hong Kong RGC IIDS lecture "Misreporting Detection: Past to Future".

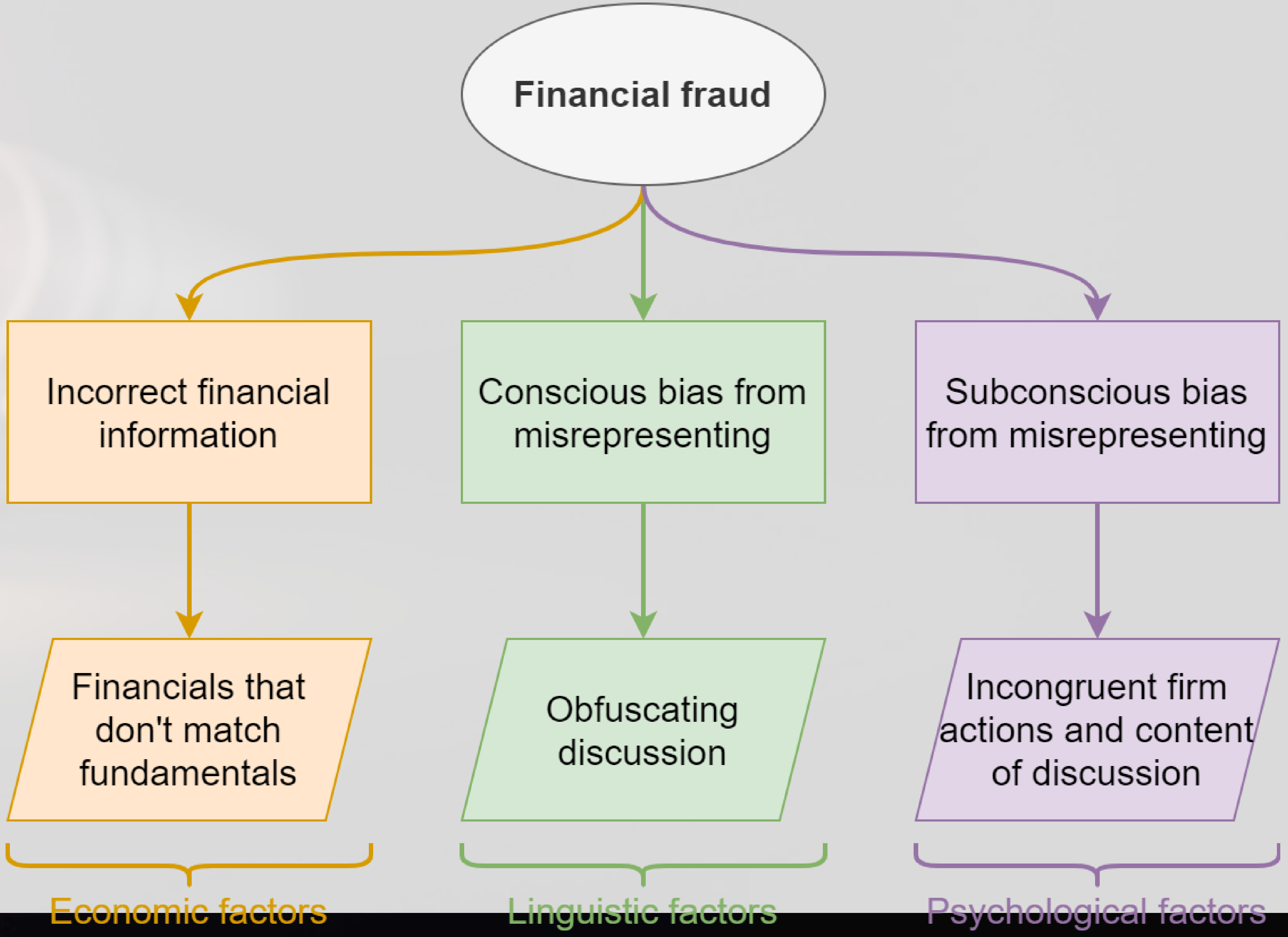
Some motivating details (based on US data)

- Misreporting is disclosed through many sources
  - By the U.S. SEC through *Accounting and Auditing Enforcement Releases* (AAERs)
  - Self-disclosed in *10-K/A* filings, *10-K* footnotes, and *8-K* releases
  - By the US government through *13(b)* actions
- Around 0.5-2% of public firms misreport per year
- It takes ~2 years, on average, for misreporting to come to light
- The economic impact of fraud is massive
  - Just the top 10 frauds cost shareholders **12.85B USD** (not inflation adjusted)
  - Enron alone cost **~35B USD** in GDP[1]
  - Follow-on costs: lost jobs, lost economic confidence, increased regulatory burden

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

[1] https://www.brookings.edu/research/cooking-the-books-the-cost-to-the-economy/

Thinking through the problem


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

Statistical issues: Sparsity

Problem

- 0.5% to 2% of public firms misreporting per year
- ⇒ difficult to detect
  - This a referred to as a *sparse* outcome
  - Sparse outcomes are difficult to model with traditional statistical methods

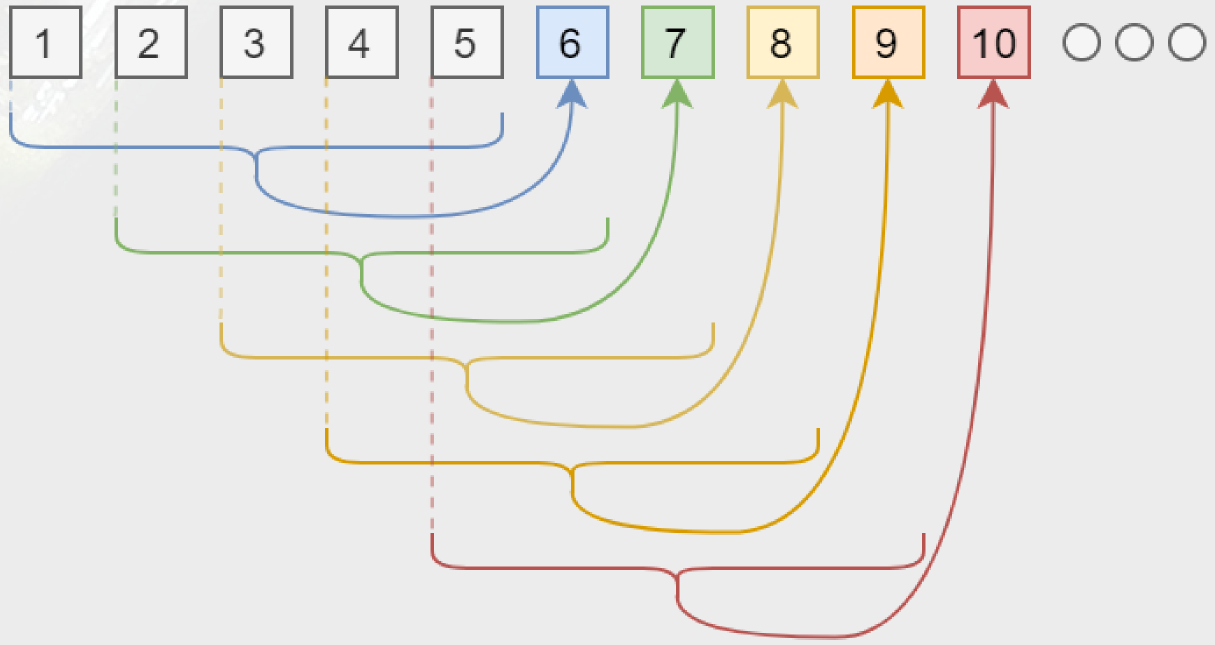
Solutions

1. Careful model selection
   - Used by traditional approaches
2. Sophisticated statistical simulation
   - Used in the paper
3. Machine learning (boosted decision trees)
   - Used in the RGC IIDS talk

> This is solveable through more advanced methods

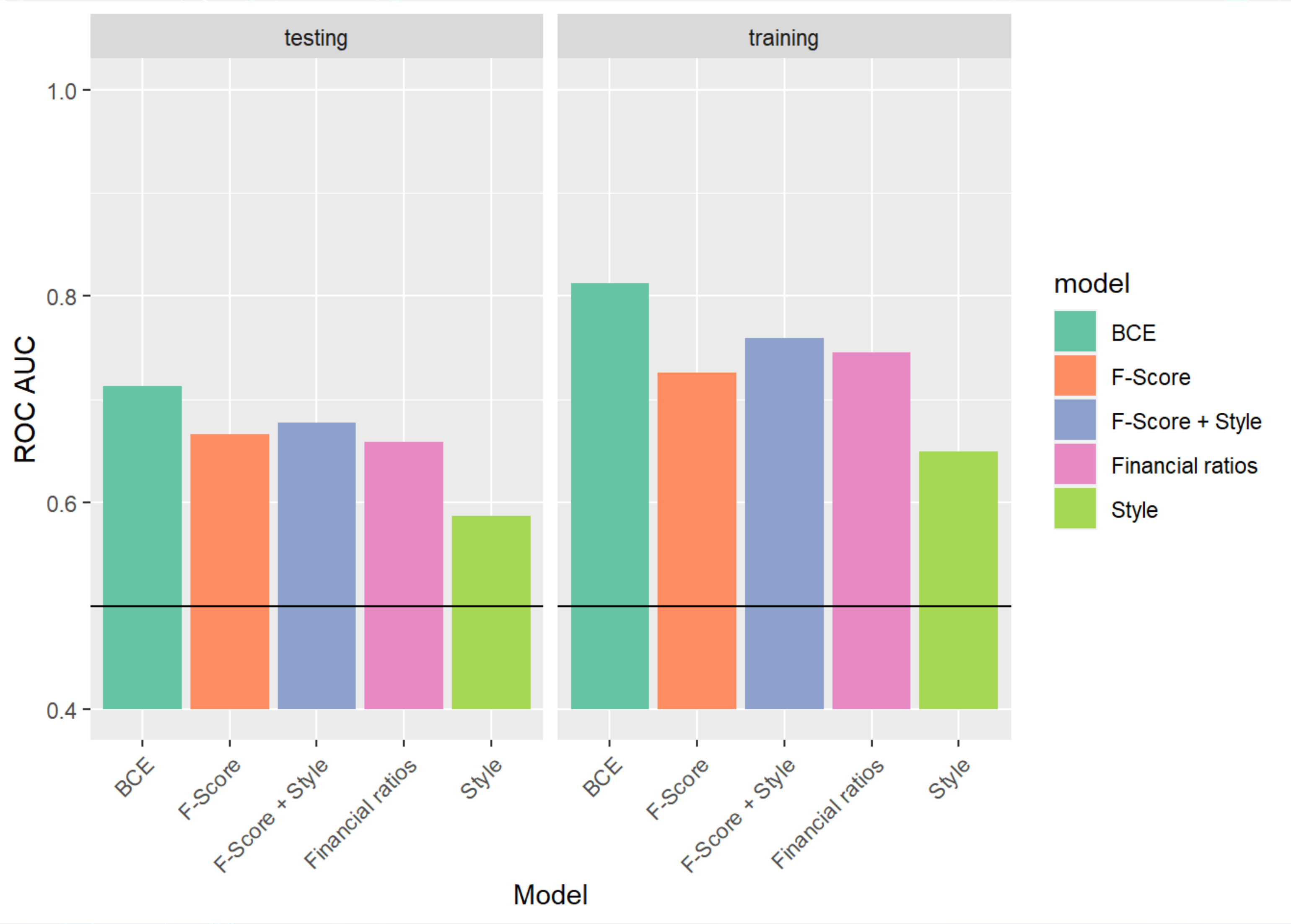
Statistical issues: Data quality

- 2 year gap between misreporting and discovery
- ⇒ need to create synthetic data to mimic that actual problem
  - We need to “go back in time” and use the data that was actually available to build a model
  - But test on the real outcomes!



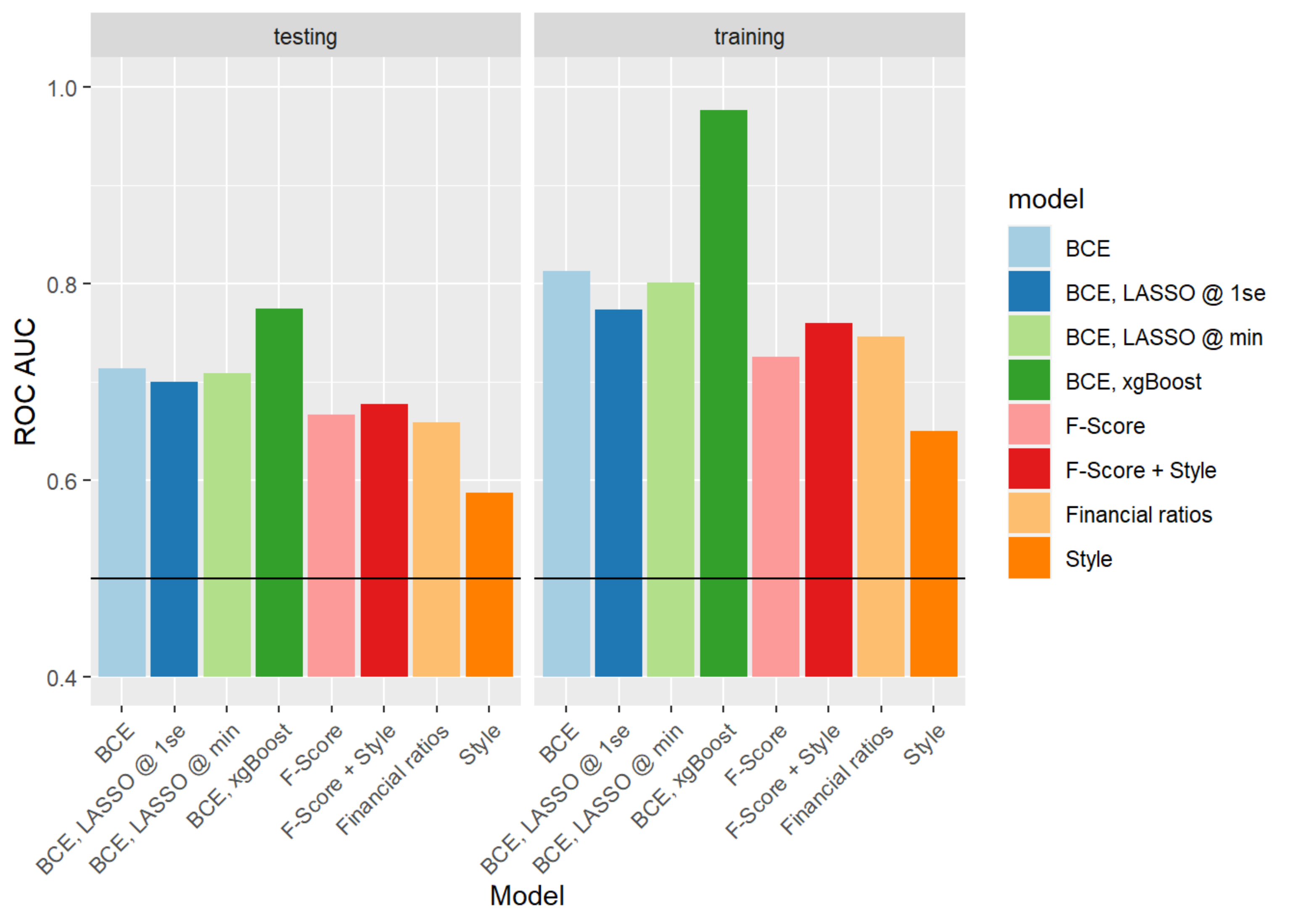
> This is solvable as a sequence of models

Outcomes using statistical approaches



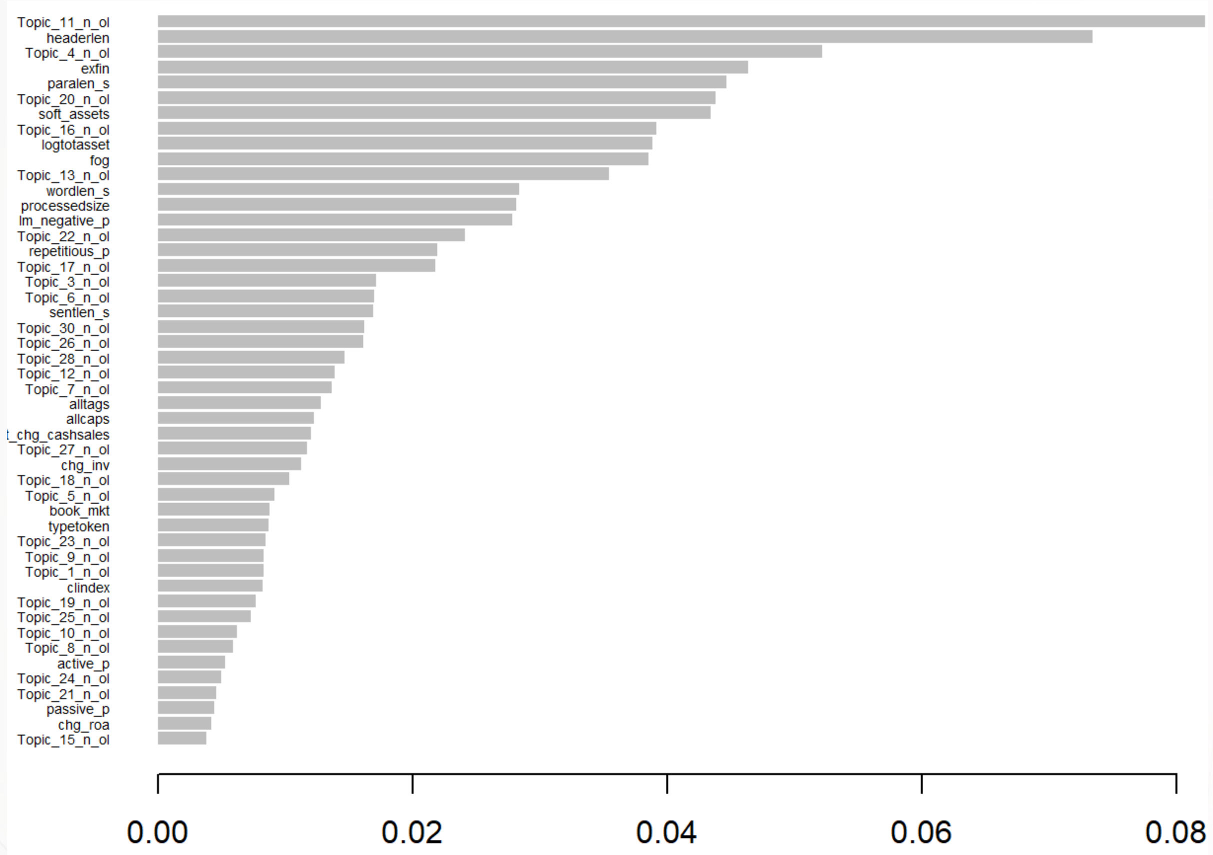
Financial ratios follows the traditional approach used since the 1990s of using intuition with a variety of ratios to build a fraud detection model. *F-Score* is a financial measure based fraud score. *Style* refers to the way in which a document was written. *BCE* is from Brown, Crowley and Elliott (2020) and uses F-Score, Style, as well as annual report content to detect misreporting.

Outcomes including machine learning estimation



xgBoost is a method based on simulation and aggregation of decision trees. LASSO is a machine learning variable selection approach to simplify models.

Explainability of ML outcomes: Importance



Explainability

Many machine learning approaches are not black boxes – computer science, statistics, and econometrics have worked hard to create easy-to-follow explanations of the methods. For instance, XGBoost has *importance* as a metric showing the overall usefulness of each type of data in the model. It can also explain exactly why a specific observation is given the score it has.

What can a regulator do with this?

> All of these methods provide a probability of misreporting

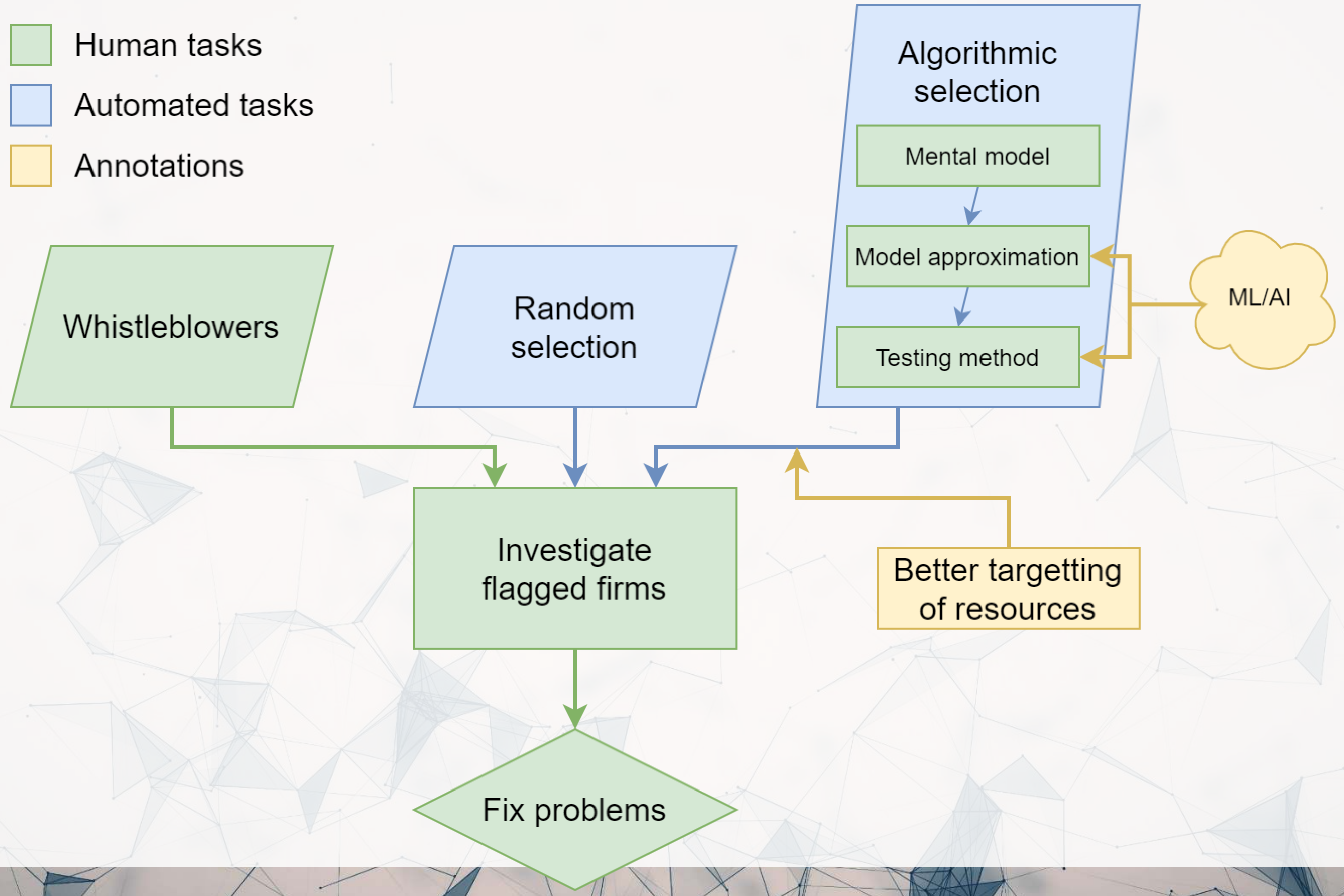
- This means a regulator must *set a threshold* calibrated to their needs
- A simple approach is to look at the top 10% of firms in a given year
  - These are much more likely to have misreporting than other firms.
  - Traditional approaches capture 18.8% of misreporting per year in the top 10%
  - The BCE approach captures 29.7% of misreporting per year
    - Including machine learning estimation pushes this even higher!

How it has been implemented

At least 1 regulator is using a hybrid approach to misreporting detection. They use both random assignment (as they previously did) for their base set of checks, and they include a model like the BCE model to identify the most likely misreporting firms to also check. This enables them to fine-tune their analytics, build comfort with them, and still reap some benefit in the mean time.

Implementation diagram for regulators

Allocating resources for fraud detection



🟩 Human tasks

🟦 Automated tasks

🟨 Annotations

Whistleblowers

Random selection

Algorithmic selection

Mental model

Model approximation

Testing method

ML/AI

Investigate flagged firms

Better targetting of resources

Fix problems

Application of machine learning for measure creation: Text analytics

Example use cases

1. Linking text data across sources
    - Article relevance for information search
    - Company comparables based on product descriptions[1]
2. Creating metrics
    - Simple document scores
        - How positive is it?
        - Do they talk about ESG?
    - Comprehensive document measures
        - What content is discussed?
        - How is that content discussed?
    - Complex metrics
        - How different is this year's annual report from last year's?

[1] Hoberg and Philips. "Text-based network industries and endogenous product differentiation." Journal of Political Economy (2016).

What about [GPT variant]?

> ChatGPT, GPT-4, GPT-3.5, GPT-3, etc.

- GPT models are a type of *large language model*
  - *Large*: the amount of parameters the model has
  - *Language*: the models are trained by seeing a large amount of written text
    - They infer everything from language
- There are use cases for these models in other fields (e.g., in robotics)
  - Chatbots like Chat-GPT
  - Personalized chatbots for customer service
- There are use cases in day-to-day work
  - Pattern matching, e.g., code completion or bug detection

Drawbacks of GPT models

- They are *language models*
  - It learns from reading and making associations across letters, words, phrases, etc.
- They lack causality (reasoning)
- They will always have an answer
  - It could be correct, incorrect due to incorrect data, or entirely hallucinated

> Recall the requirements for data analytics: A question, logic, reasoning, data

GPT has…

- A question: what you ask
- Data: what it was trained on
  - But is this the right data for your question?

GPT does not really have…

- Logic & reasoning: Both are based on language patterns
  - This not logical reasoning as we would expect

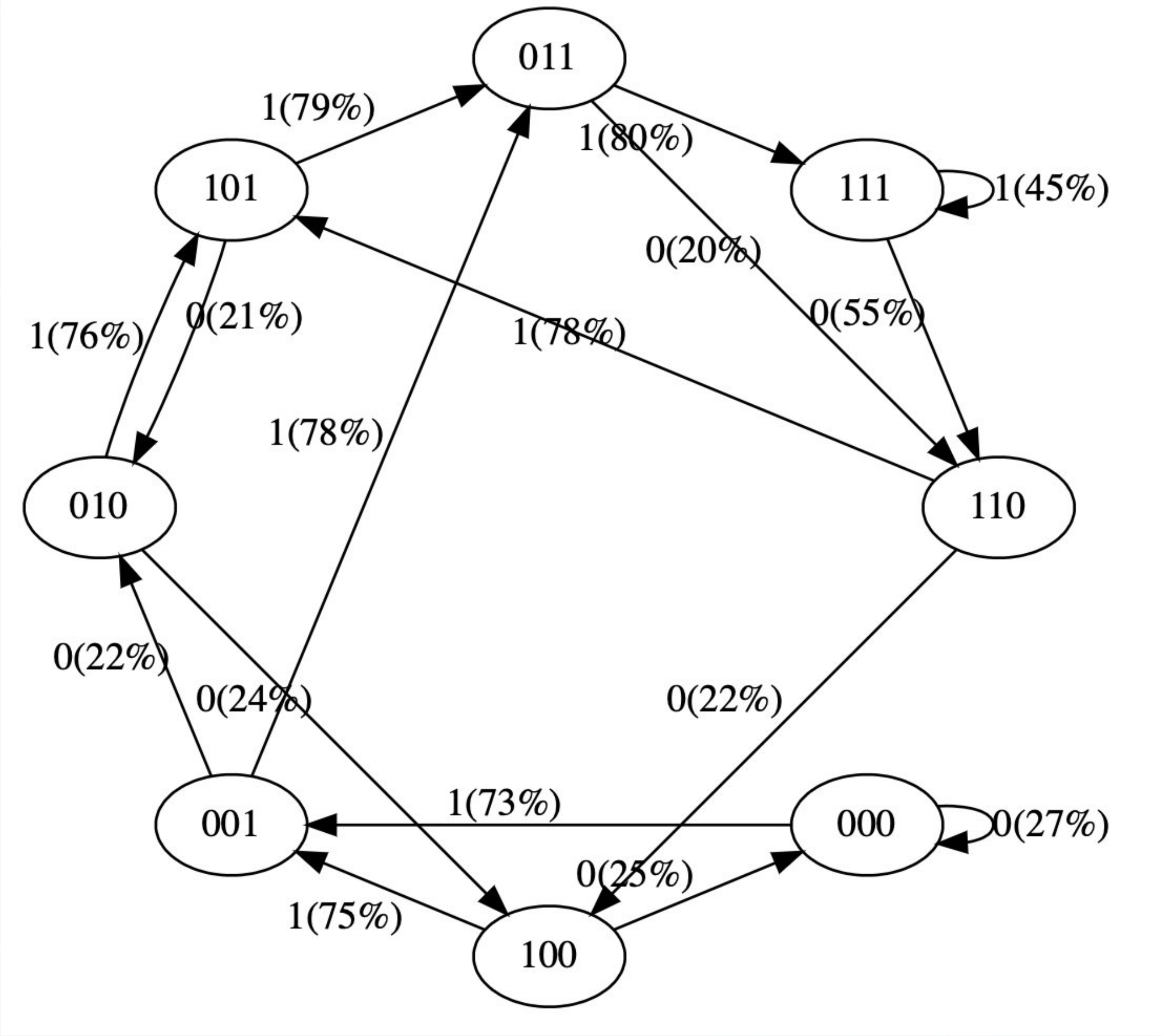
An illustration of a simple GPT model[1]

1. Trained on sequences of binary digits
   - 2 tokens: 1 or 0
2. Memory of length 3
3. Data: `111101111011110`

8 possible states for the model, but only 4 are in the data (no `00` or `000`)

What is the most likely next digit, given the past 3 digits?

The output model:



[1] Model and figure by Andrej Karpathy, code here

Practical uses of transformers in analytics

> GPT models belong to a large class of methods based on transformer neural networks

- Transformers are extremely useful in analyzing text
  - Some recent work has also found them to be useful in analyzing images (vision transformers)
- What are these models good at?
  - Definable problems
  - Problems you can “fine-tune” to
    - If you can replicate what you need by hand, then this is an option
  - Representations of text
    - If you need a numerical approximation of text, they can do it well

Transformer example: FinBERT

- Fin: **Fin**ancial
- BERT: **Bidirectional E**ncoder **R**epresentations from **T**ransformers
- FinBERT is two things:

1. A model trained on financial language (10-K filings, analyst reports, conference calls)
   - It can represent text within this set of information
2. A set of algorithms to calculate various text measures at sentence level
   - The sentiment
   - If there is ESG content
   - If the sentence is forward looking

- The paper behind it[1] shows that FinBERT outperforms simpler, more commonly used approaches for classifying financial text

[1] Huang, Wang and Yang. "FinBERT: A large language model for extracting information from financial text." Contemporary Accounting Research (2022).

Hands on with FinBERT

Go to this link: https://rmc.link/GICdemo

- Provide example sentences to the application, and it will analyze them with FinBERT
- Consider giving sentences about different financial content:
  - Positive or negative
  - Environment, social, governance, or unrelated to ESG
  - Forward looking or backward looking

> Does the model match your intuition?

Why do we care about a methods like FinBERT?

- Just doing 1 computation like this doesn’t matter – we can do it ourselves
  - But how many sentences are there in annual reports? (get figure for US)
    - Would you want to hire enough people to hand code this?
- These models allow for automating tasks that are practically infeasible
  - On a GPU that costs ~$2,000, you can run 1,000s of sentences per second
    - Cheaper and faster than labor
      - But would you actually do this if it needed human intervention?
    - And the accuracy is pretty good!
      - The benchmark is human accuracy, not perfection

> This opens new sources of data for analytics!

Takeaways

Main takeaways

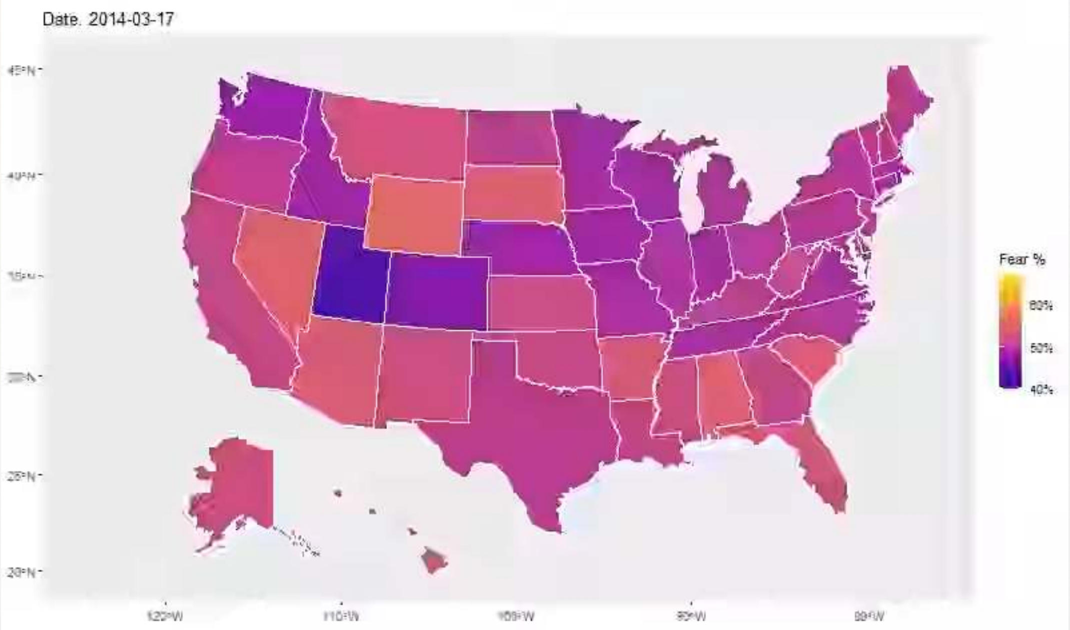
1. Data analytics is a broad field
2. Use what makes sense
   - Just because it is talked up in the news doesn’t mean it works well for your problem.
3. New tools can give new ways to feasibly approach analytics problems
   - But you still need the underlying logic

From our examples:

1. Analytics can improve operations
2. Hybrid methods using traditional approaches augmented by analytics are feasible to build comfort with analytics.
3. Text methods in machine learning are good enough for specific tasks like measurement

Overview of AI/ML research in Economics-based fields

1. Analyzing information content: annual reports, analyst reports, conference calls
2. Social media dissemination by firms or investors
   - Extract post content, analyze images, or calculate similarity between parties’ posts
3. ML for causality
4. ML for time series forecasting: Earnings or cost of capital up to 10 years ahead
5. Large language models in finance
6. Fake news measurement for financial news
7. Population characteristics and policy implications



Thanks!

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

Packages used for these slides

- `kableExtra`
- `knitr`
- `ggplot2`
- `ggthemes`
- `ngramr`
- `quarto`
  - `quarto-qrcode`
  - `material-icons`
- `revealjs`