Using Machine Learning to Detect Financial Fraud

6 May 2021

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

About me

- Assistant Professor of Accounting at SMU
  - **Research**
    - Accounting disclosure: What companies say, and why it matters
      - Fraud detection based on annual report content
      - Corporate and executive social media posting
      - Fine-grained measurement of context within annual reports
    - Approaching accounting disclosure problems using AI/ML
  - **Teaching**
    - Forecasting and Forensic Analytics
    - Accounting Theory
    - Financial Accounting
    - Machine Learning for Social Science
- Adviser to Fraud Factors, a local corporate governance data vendor

Corporate financial fraud

What our dicussion will focus on

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

- In other words, when a company is misrepresenting its finances to its investors
  - More precisely called *misreporting*

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
  - Lead 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

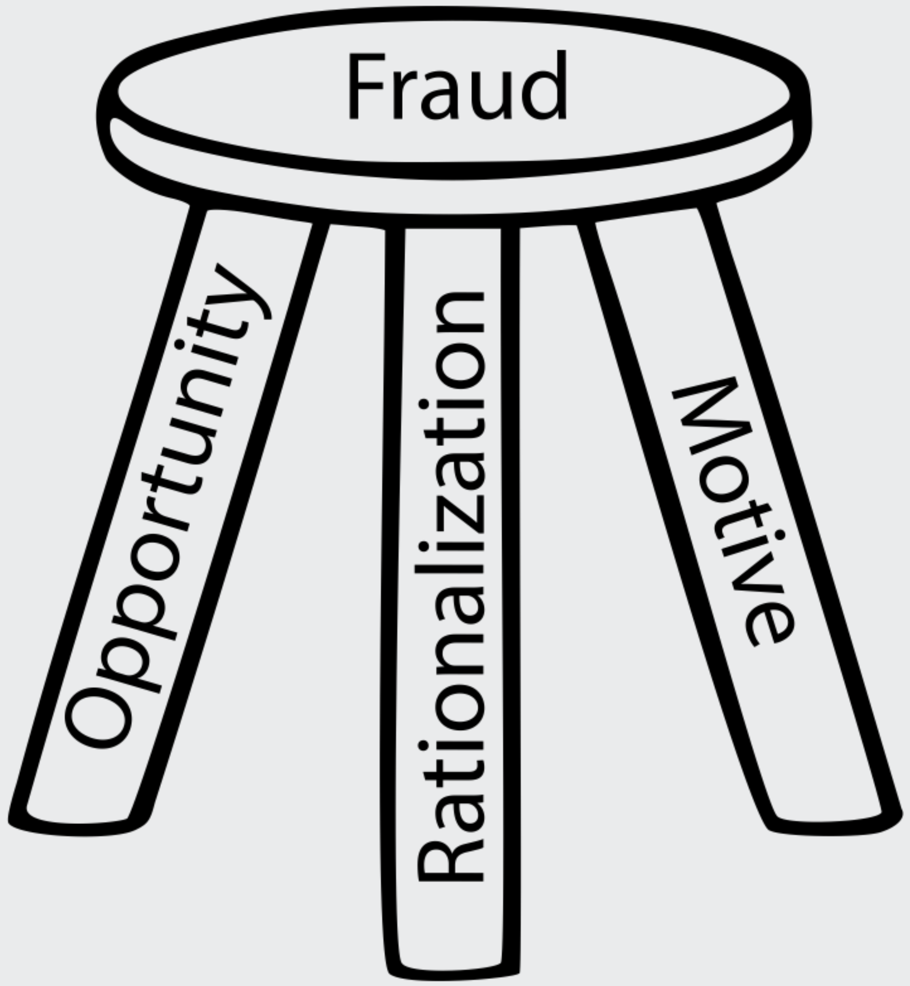
Singapore is not immune

- Coastal Oil
  - Forging contracts to secure loans from 8 banks
    - $320M USD worth of loans
- Keppel O&M
  - $55M USD bribery in Brazil for contracts
  - Highly profitable, until fines rolled in
    - Profit of $351.8M USD
    - Fines of $422M USD (to US, Brazil, Singapore)
  - 6 employees implicated
  - 1 Keppel lawyer pleaded guilty in USA for drafting bribery contracts

Why does financial fraud happen?

> Per the Fraud Triangle, fraud stems from having all of…

- Opportunity
  - Hole in the control system
  - Profitably exploitable
- Rationalization
  - Resentment of corporation
  - Poor culture
  - “Borrowing”
- Motivation
  - Family needs
  - Maintaining lifestyle
  - Maintaining performance



What can we do?

The problem

How can we *detect* if a firm is *misreporting*?

- *Detect*: There are usually companies misreporting any given year
  - E.g., 1.5-2% of US public companies misreport per year
- We will approach this with a mix of…

- Business insight
- Economic theory
- Psychology theory

- Statistics
- Machine learning

Careful consideration is needed throughout

Why is this a tough problem?

- Fraud happens in many ways, for many reasons
  - We saw 7 different types earlier
  - All of them are important to capture
  - All of them affect accounting numbers differently
  - None of the individual methods are frequent…

> Ideally we want a general method to capture all of these

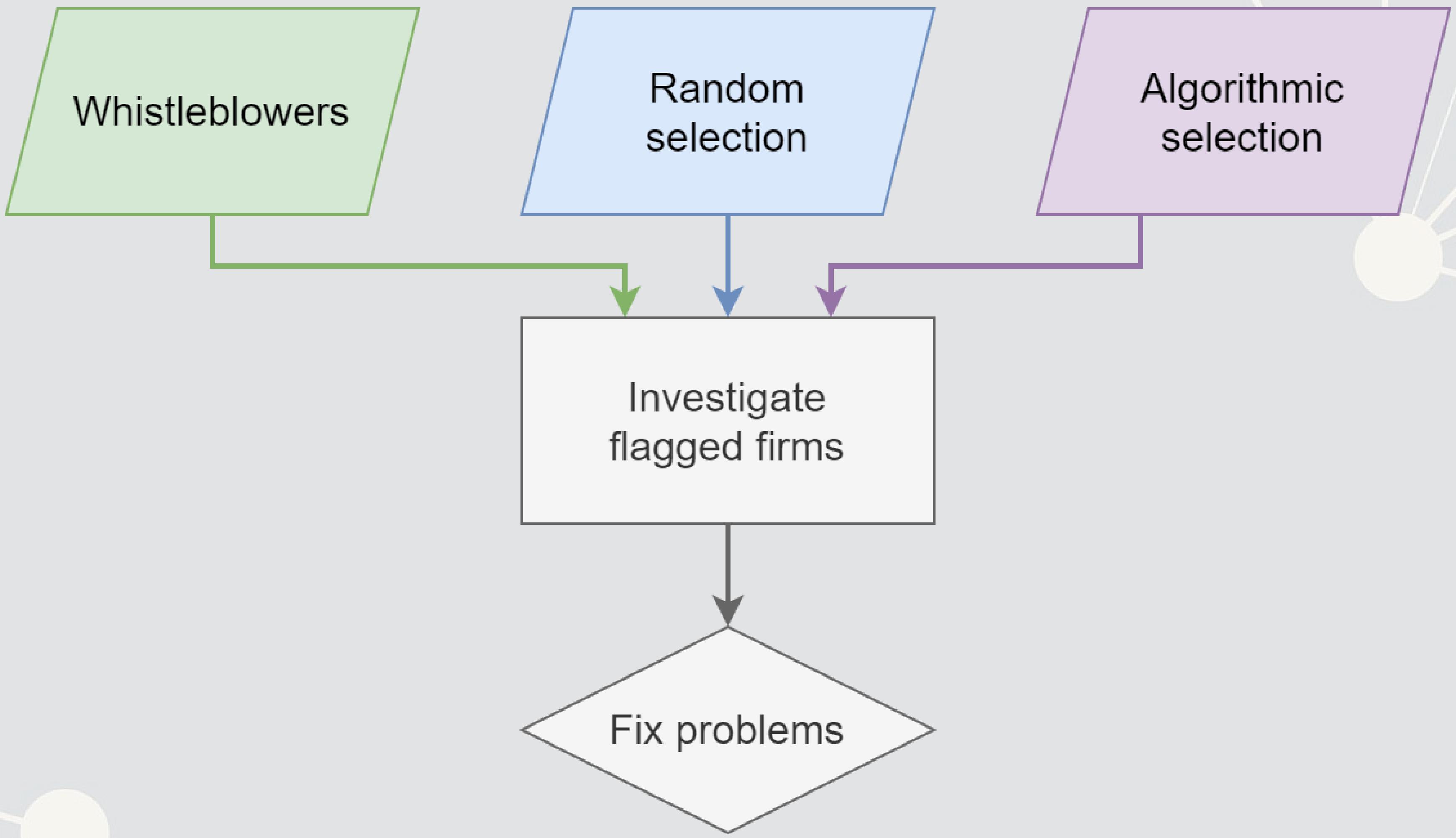
Ways to detect fraud

- Random checks
- 1990s: Focus on financial metrics
  - Are metrics it too good to be true?
  - Do metrics not make sense?
- 2000s: Look for certain peculiar behaviors of the company
- Modern approaches:
  - Purpose-built metrics to detect inconsistent corporate behavior
  - New statistical approaches to determine inconsistencies

> We will see how machine learning helps with both modern approaches

- Note that the two modern approaches aren't mutually exclusive: they can be combined!

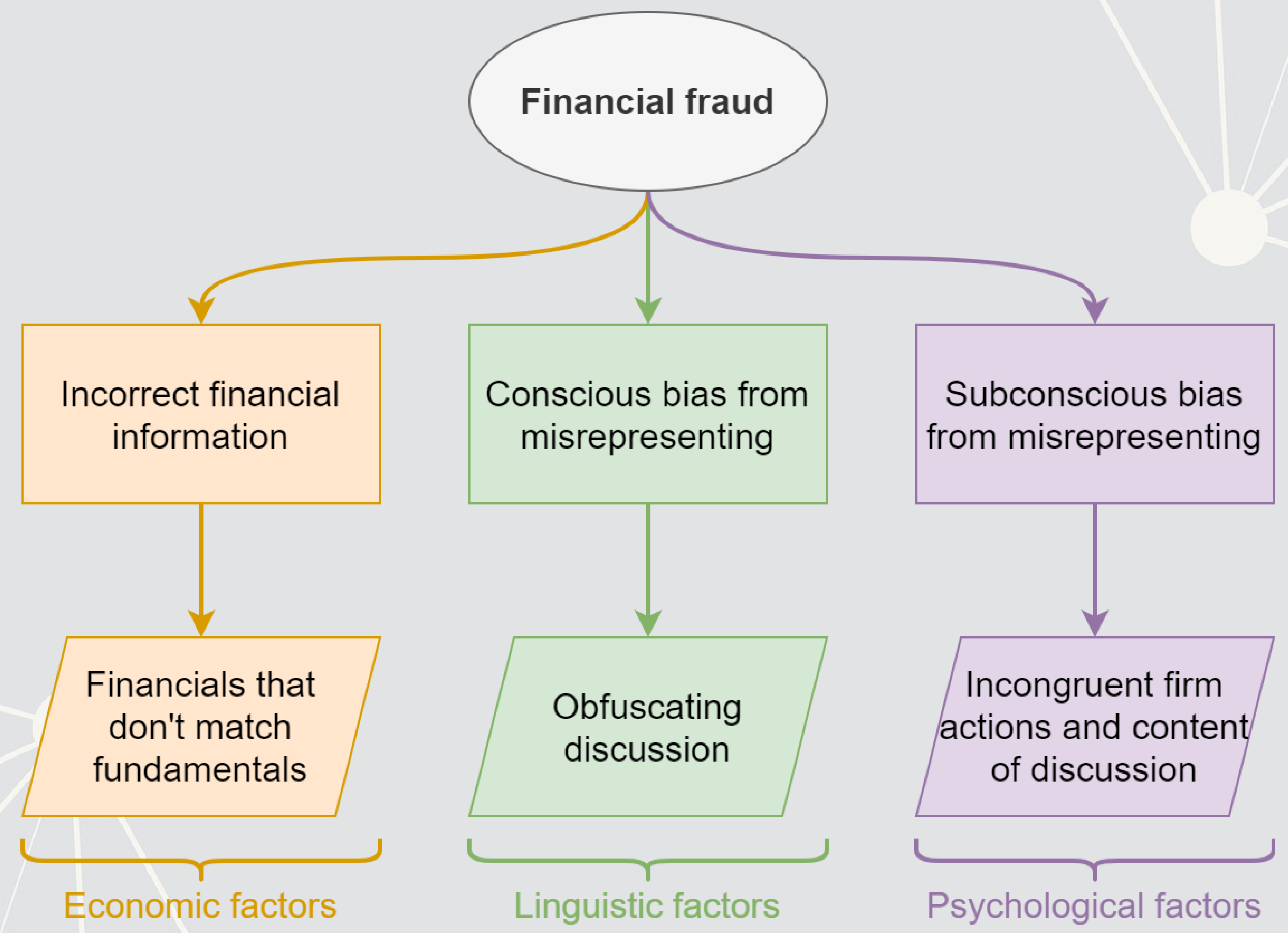
A practical modern approach



> Why a hybrid approach? Each approach has its own strengths.

Unlocking data through machine learning

Mental model of misreporting



The scientific method

- To effectively determine an approach to solving a problem as complex as defecting financial fraud, we leverage the scientific method:
  1. **Question**: What are we trying to determine?
     - "How can we *detect* if a firm is *misreporting*?"
  2. **Hypothesis**: What do we think will happen? Build a mental model
     - From our mental model:
       1. Some financial information will be incorrect
       2. Some aspects of obfuscation may be visible
       3. Certain discussion will be over- or under-discussed
  3. **Prediction**: What exactly will we test? Define goals; formalize model/statistical approach
  4. **Testing**: Test the model
  5. **Analysis**: Did it work? Why did it [not] work? How can we improve?

Putting our mental model into action

- We would like to gather data that best approximates the constructs from our mental model
- Constructs like "annual report content" are traditionally difficult to measure

> Machine learning can automate these processes

- For well defined constructs we can either create manual rules to flag it, or we can use *supervised* machine learning
  - E.g., "amount of discussion of loan loss provisions by banks"
- For broader constructs we can use *unsupervised machine learning*
  - E.g., "annual report content"

> We will focus on unsupervised machine learning first

How does machine learning help?

> Consider how to measure "annual report content"

- The traditional way:
  - Hire a team to manually examine annual reports
  - The team would assign scores to filings based on what was or was not covered in the filing
  - Time taken: *1,392 man-hours* per year of reports (on average)

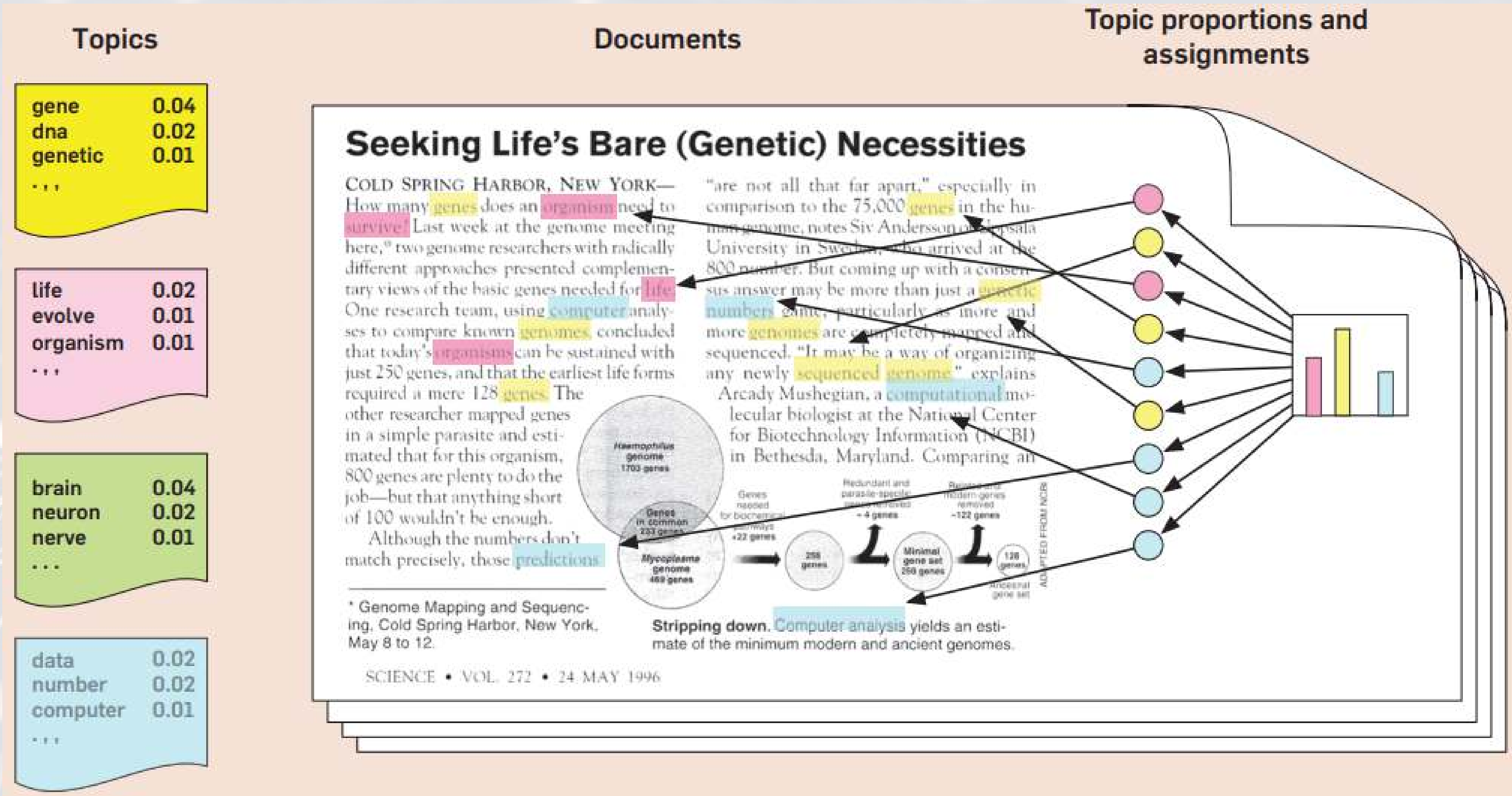
How does machine learning help?

> Consider how to measure "annual report content"

- The machine learning way (LDA):
  - Let the computer read every annual report
  - Based on the correlations between words within and across documents, the computer simultaneously determines:
    1. The types of discussion in the annual reports
    2. A weighted list of which words fit with which type of discussion
  - Apply this weighted list to each annual report to get each document's content weightings
  - Time taken: *a few hours* of coding and running the code

> Because of the ambiguity of our construct, human and computer performance is similar

Let’s take a look at the ML method



Source: Blei 2012

LDA output on annual reports

Topic 6

Topic 11

Topic 21

Topic 30

Topic 2

Topic 9

Topic 12

Topic 26

Topic 8

Topic 19

5.8

Executing our full mental model

- We model misreporting as a function of:
  - Financial metrics (as in the 1990s)
  - Linguistic characteristics (as in the 2000s)
  - The deviation of annual report discussion from industry norms
    - This is where LDA is used
- We use a logistic regression framework to test the model
  - Tested using data from 1994 through 2012

> This model is showcased in Brown, Crowley, and Elliott (2020)

How well does it work?

Adding in report content drastically increases performance

Lesson learned

1. Mental models are important in building predictive models
   - Ideally, we want the model we build to capture as much of our mental model as possible
2. Machine learning can make it easier to better approximate our mental model with data
   - We can capture broad constructs like *annual report content* with ease
3. A model that better captures our mental model should perform better
   - The modern model is much better at predicting fraud!

> Overall, machine learning can help improve the effectiveness of decision making for this problem by letting us more precisely utilize our mental model

Improving fraud detection algorithms

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!

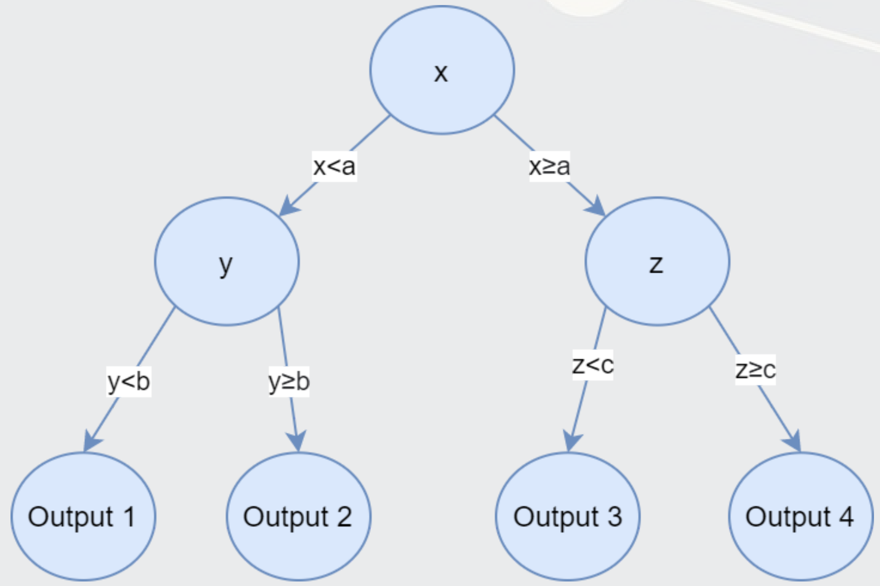
> If we want a better accuracy, we need to replace logistic regression

How ML helps with sparsity

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

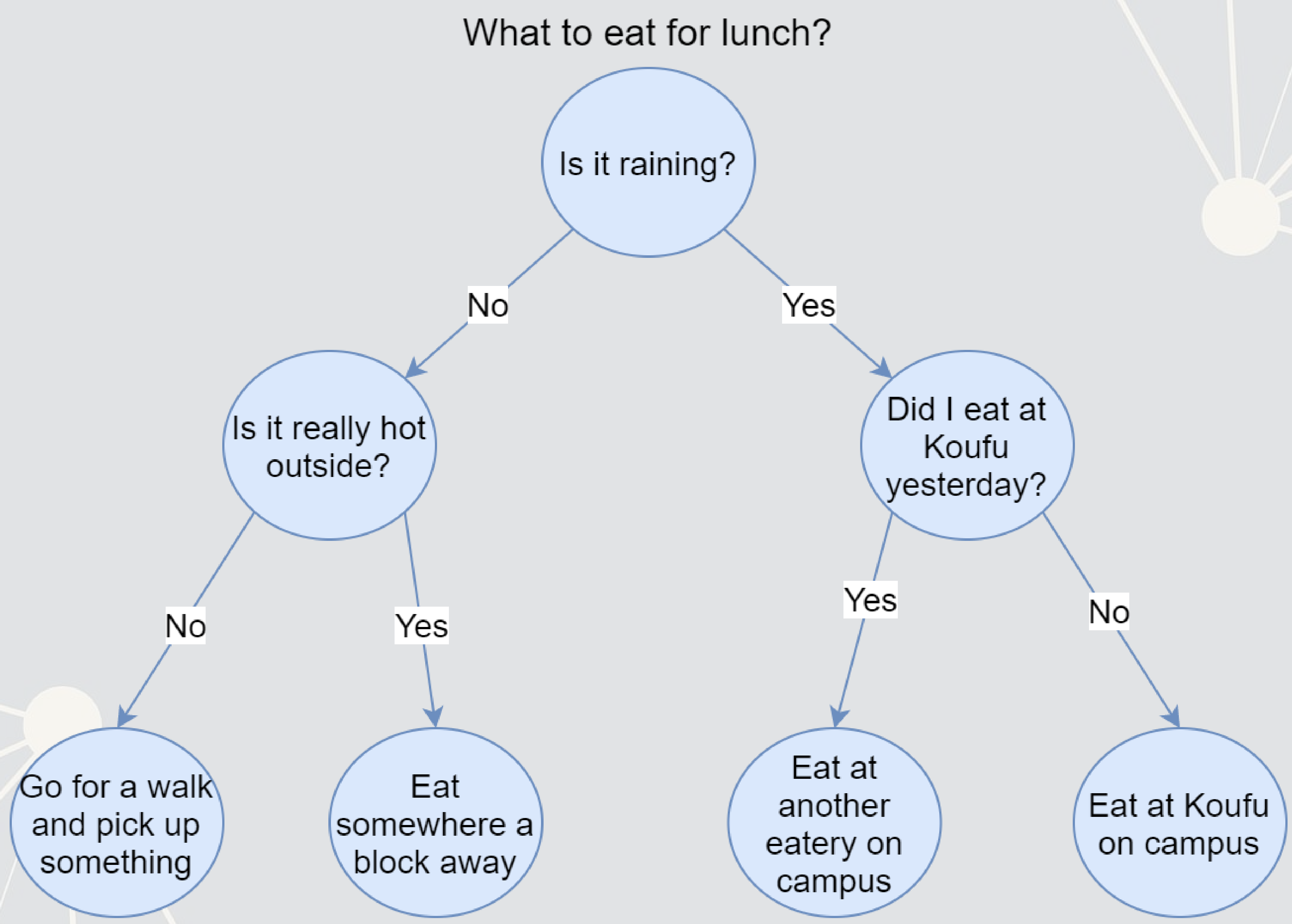
Decision trees

- 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?



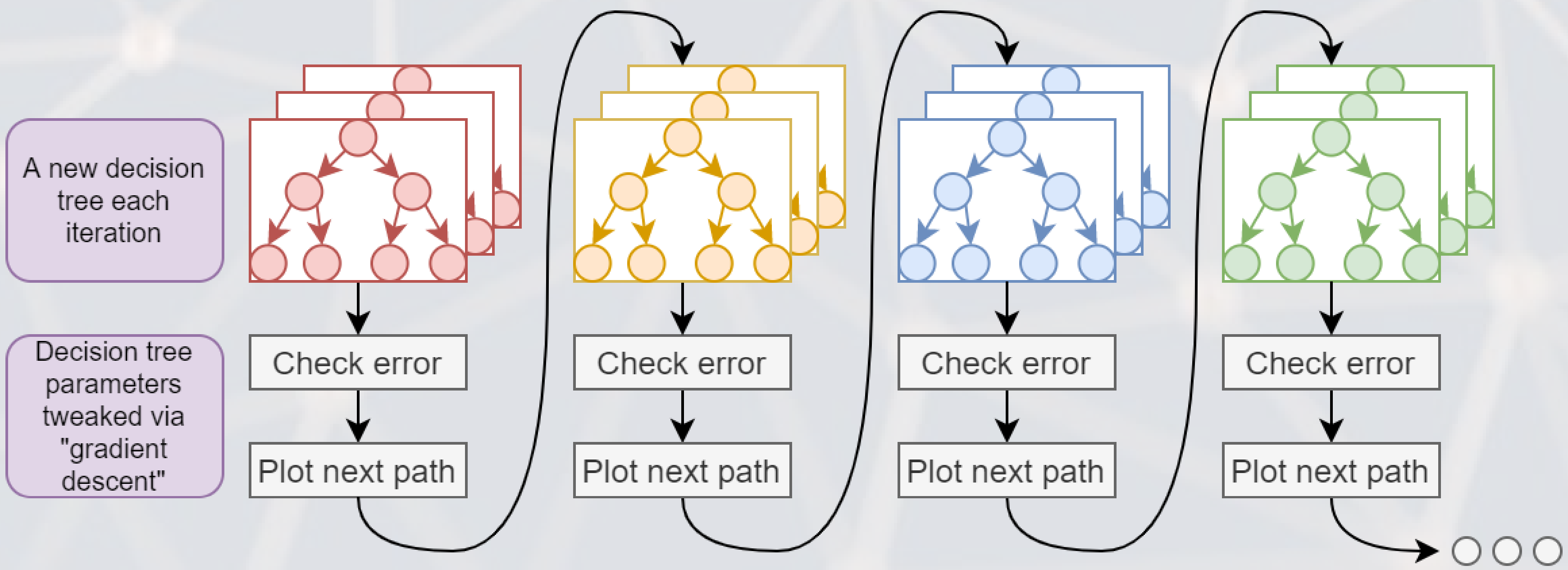
> The final approach will use a bunch of decision trees

A simple example



Applying trees to fraud detection

- Bao et al. 2020 take the following approach:
  1. Let financial data speak for itself, by using raw financial information
     - This is in contrast to the traditional approach of carefully selecting financial ratios to put in a model
  2. Toss the data to RUSBoost (AdaBoost variant), which is a tree-based machine learning classification method
     - Trees to allow for nonlinear/discontinuous effects
     - *R*andom *u*nder*s*ampling: to further help address sparsity



How well does this work?

> Improves statistical accuracy over logistic regression

Lesson learned

1. Traditional statistical approaches to binary classification aren't always appropriate
   - Logistic regression works best when at least 10% of all observations are in each group
2. Certain algorithms from machine learning can be appropriate drop-in replacements for traditional regression techniques
3. For sparse classification problems (events that occur < 10% of the time), algorithms based on ensembled decision trees work well
   - This is illustrated well by our second modern model

> Overall, machine learning can help improve the effectiveness of decision making for this problem by swapping out a standard regression approach for a machine learning approach in an automated process

Some final thoughts

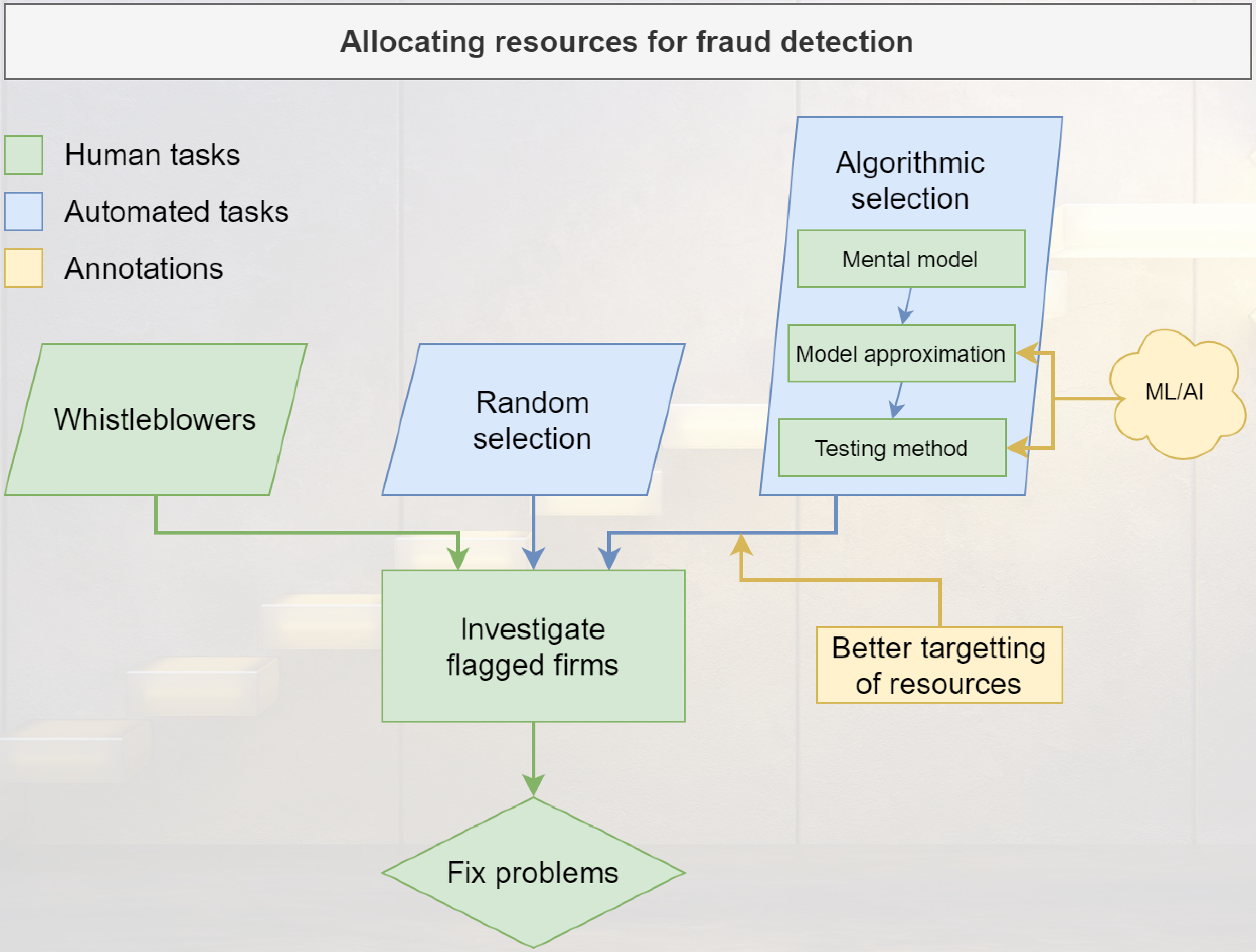
You can combine both methods!

> This material is covered in our *Forecasting and Forensic Analytics* course at SMU (html slides, pdf slides)

- On data from 1999-2003…
  - The best traditional model has an AUC of 73%
  - The first modern model has an AUC score of 76%
  - Replacing the logistic regression in the modern model with XGBoost yields an AUC of 81%!

> AUC: If I select to observations at random, what is the probability the algorithm correctly orders them?

Bringing everything together



Caveats

- Don't use machine learning tools just for the sake of using them
  - While the discussed tools are useful, it is always important to consider how appropriate the tool is for the job at hand
- Instead, carefully consider how exactly you expect the phenomenon you are trying to detect behaves
  - Do this in the absence of considerations about data or methodology!

> Once you have a firmed-up mental model, you can determine how to best measure the various factors from your model

Main takeaways

> #1: Machine learning can help unlock new fraud detection features

- Machine learning lets you build measures that more closely map to your mental model
  - Often times these features could be manually coded, but at the expense of hundreds to thousands of hours of work

> #2: Machine learning provides new ways to leverage existing data

- Even with the same data and measures, we can get better predictive ability, particularly when trying to detect sparse events (<10% frequency)

To learn more:

- The first modern approach is based on the following research paper:
  - Brown, Nerissa C., Richard M. Crowley, and W. Brooke Elliott. “What are you saying? Using topic to detect financial misreporting.” Journal of Accounting Research 58, no. 1 (2020): 237-291.
- The second modern approach is based on the following research paper:
  - Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. “Detecting accounting fraud in publicly traded US firms using a machine learning approach.” Journal of Accounting Research 58, no. 1 (2020): 199-235.
- To see an illustration combining the above, you can check out the following slide deck by Professor Crowley:
  - Html slides, PDF slides

Thanks!

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