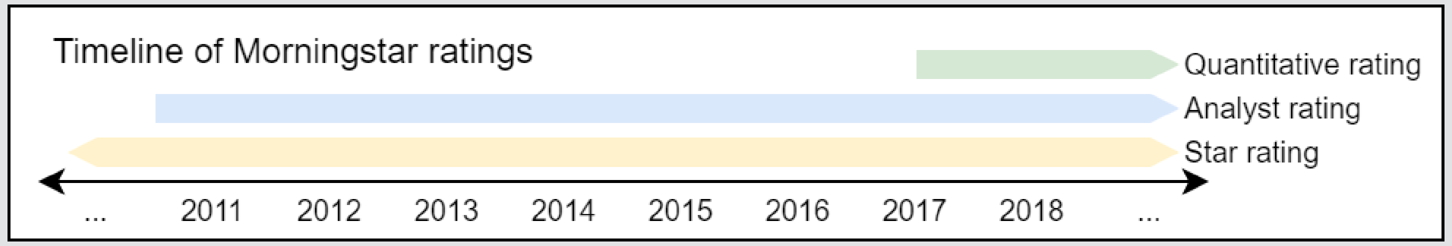
Discussion of "What Should Investors Care About?…"

ABFER, May 2021

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

Overview of the paper

My interpretation of the story



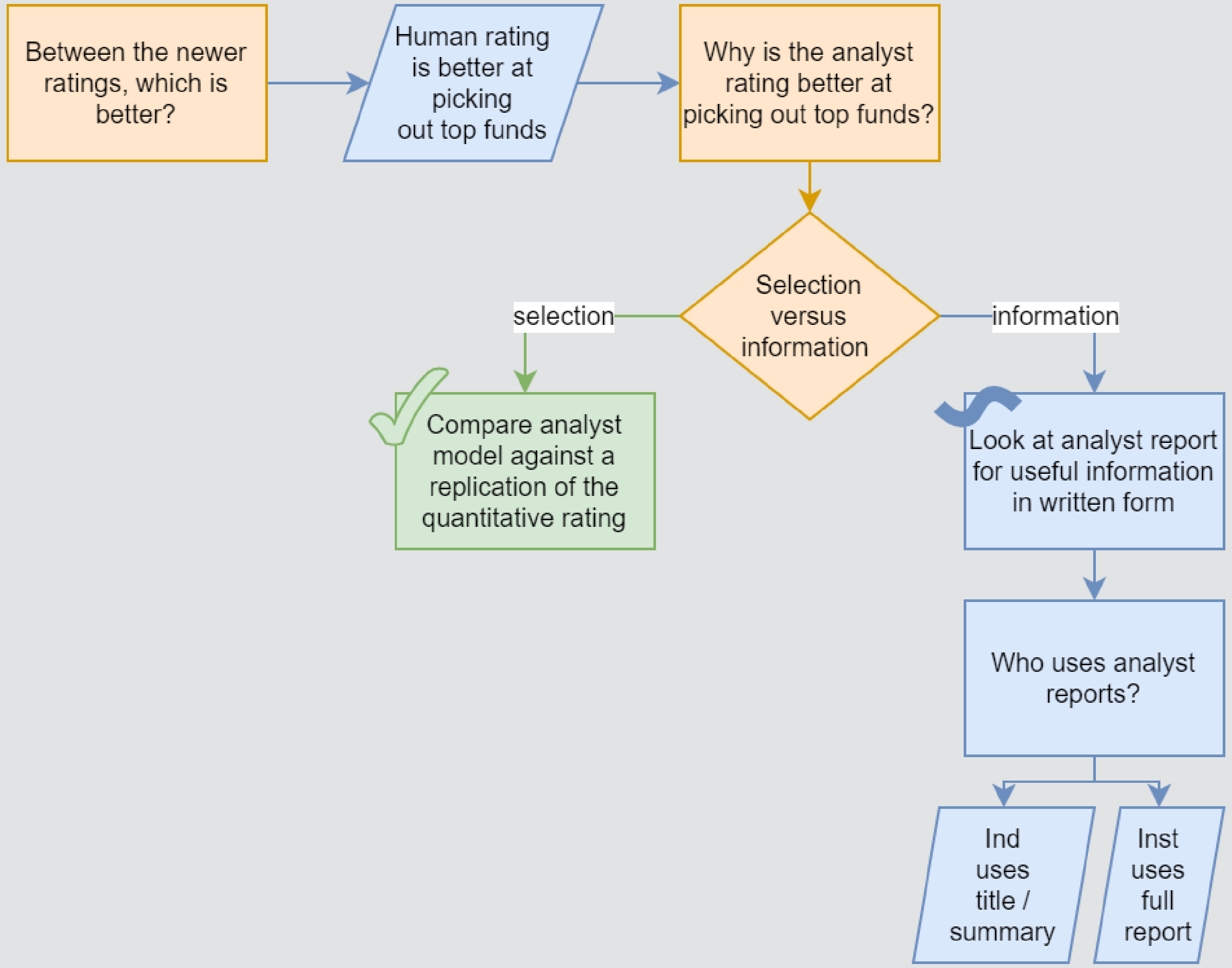
Timeline of Morningstar ratings

Quantitative rating

Analyst rating

Star rating

... 2011 2012 2013 2014 2015 2016 2017 2018 ...



Between the newer ratings, which is better?

Human rating is better at picking out top funds

Why is the analyst rating better at picking out top funds?

Selection versus information

selection

information

✓ Compare analyst model against a replication of the quantitative rating

Look at analyst report for useful information in written form

Who uses analyst reports?

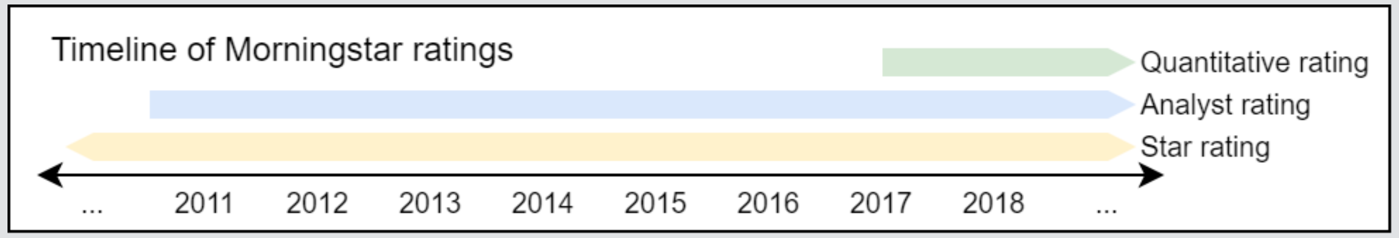
Ind uses title / summary

Inst uses full report

Highlights

1. If true, human raters being superior due to a selection mechanism rather than skill is an interesting result
2. Comparing human ratings against an ML-based rating that is *in use*, rather than a hypothetical benchmark
   - This is needed to convincingly point out shortcomings of ML methods
   - This is not needed to show potential benefits of ML methods though
3. The test showing the differential impact of the full document (useful for returns) and summary (useful for flows) is interesting
   - Perhaps an example of the effectiveness of push vs pull mechanisms
   - Also possibly related to frictions (cost of information)

My interpretation of the results



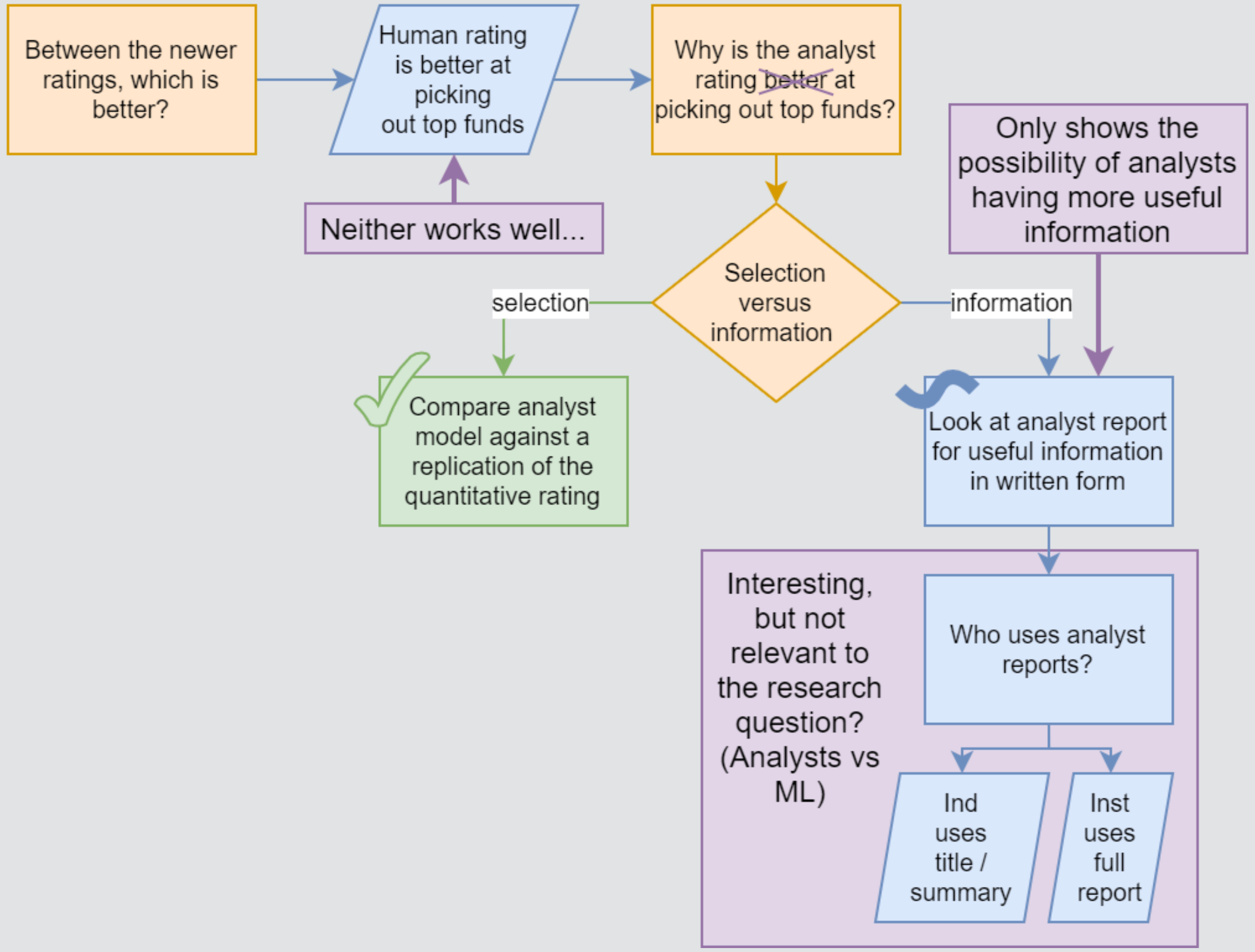
Timeline of Morningstar ratings

Quantitative rating

Analyst rating

Star rating

... 2011 2012 2013 2014 2015 2016 2017 2018 ...



Between the newer ratings, which is better?

Human rating is better at picking out top funds

Neither works well...

Why is the analyst rating ~~better~~ at picking out top funds?

Selection versus information

selection

Compare analyst model against a replication of the quantitative rating

information

Only shows the possibility of analysts having more useful information

Look at analyst report for useful information in written form

Interesting, but not relevant to the research question? (Analysts vs ML)

Who uses analyst reports?

Ind uses title / summary

Inst uses full report

Main comment #1

What is optimal for investors in mutual funds?

The result as is (univariate)

Rating type	1 month Style-adj	1 year style adj
Analyst rating, 2017+	0.122**	0.104**
Quantitative rating, 2017+	0.034	0.032
Star rating, 2017+	0.094**	0.073*

Note:

Data from Table 2 and Internet Appendix 2

> If your goal is to get the highest average return above similar funds, analyst rating and star rating could work if you have no other data

What about the spread of returns?

Spec	1 month style-adj	1 year style adj
Analyst rating, 2017+	0.350	0.508*
Quantitative rating, 2017+	0.128**	0.135**
Star rating, 2017+	0.499***	0.468***

Note:

Data from Table 2 and Internet Appendix 1

> If your goal is to avoid the lowest return compared to similar funds, analyst rating is perhaps the worst choice, statistically?

Separately, why is this result so different from Morningstar's own event study?

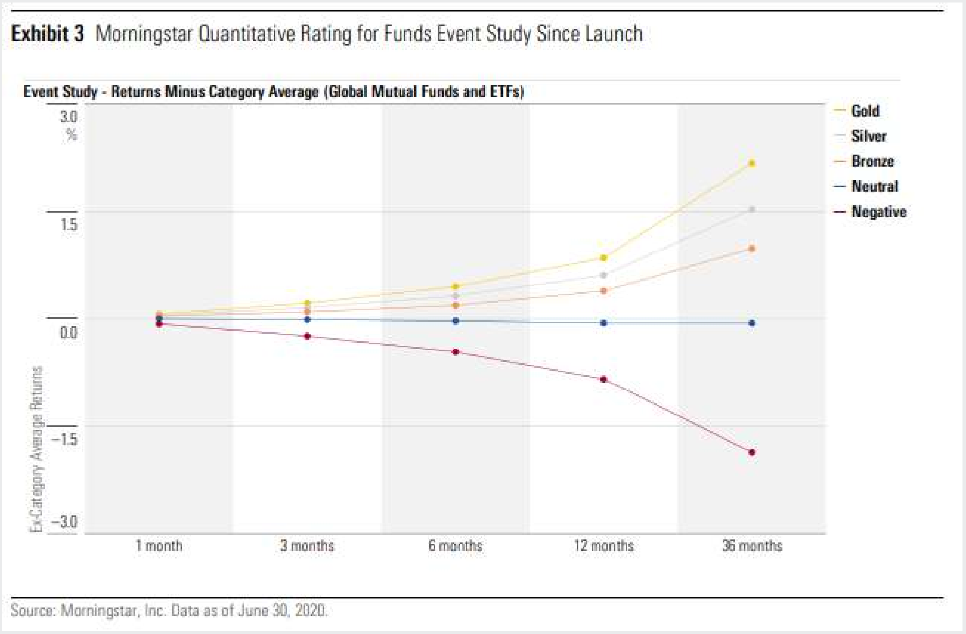


Exhibit 3 Morningstar Quantitative Rating for Funds Event Study Since Launch

Source: Morningstar, Inc. Data as of June 30, 2020.

But if you have other data…

	Analyst rating, 2017+	Analyst rating, 2017+	Quantitative rating, 2017+	Quantitative rating, 2017+
Variables	1 mo style-adj	1 mo style-adj	1 mo style-adj	1 mo style-adj
Rating	0.014		0.003	
Negative		-0.149		0.030
Bronze		0.020		0.008
Silver		0.018		0.024
Gold		0.054		0.031
Star rating	0.039	0.038	0.050**	0.053**
Additional controls	Yes	Yes	Yes	Yes

Note:

Data from Table 3

> Then why use any of these ratings?

- Need to show *incremental* contribution of analyst rating and/or quantitative ratings over other existing data
  - Though, for any measure than star rating, it seems unclear if it will be useful on a 1 month horizon

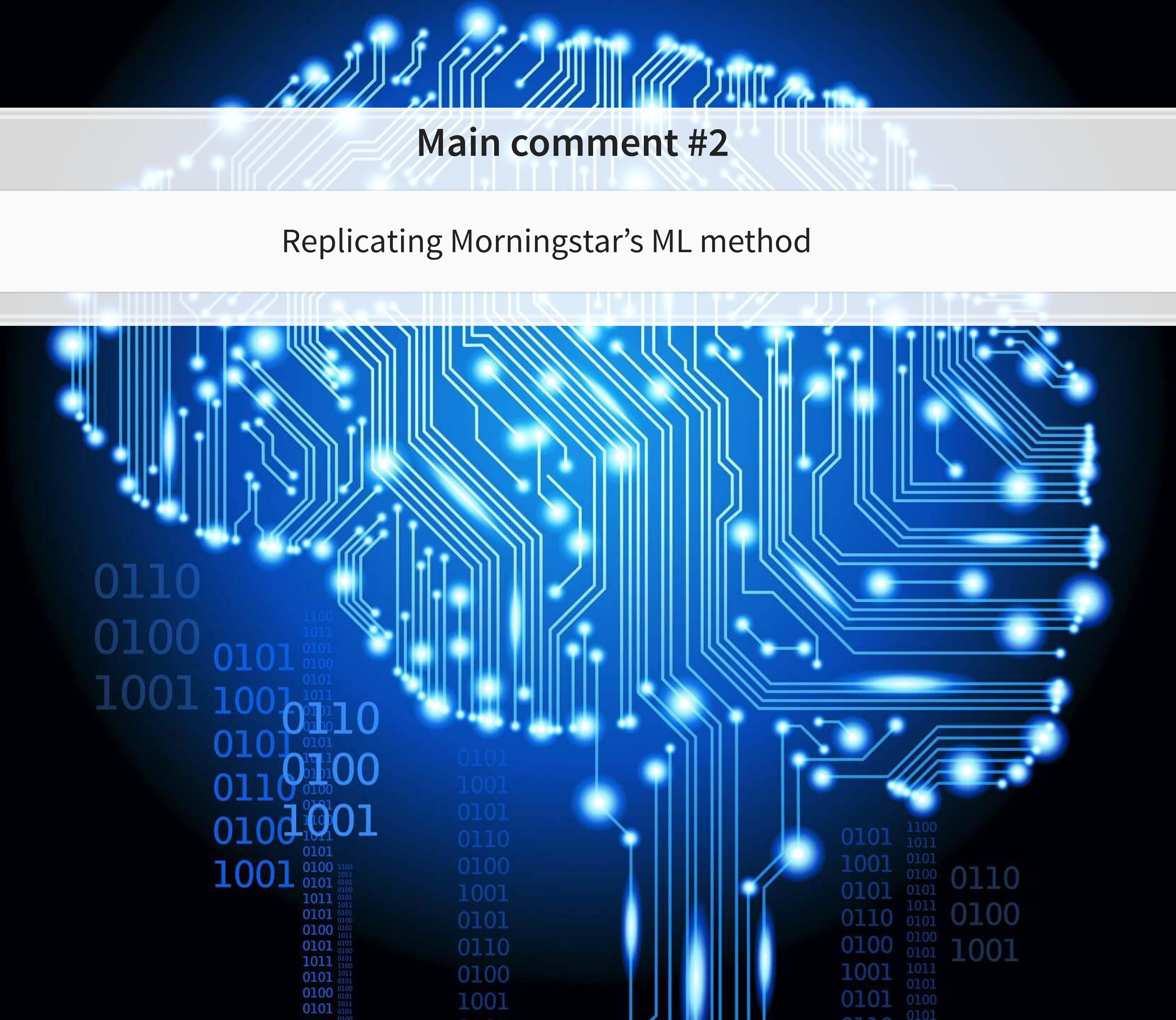
Why does this matter?

> The prior results cast doubt on the assertion that the analyst rating is “still highly valuable to individual investors in guiding their investment decisions.”

- The paper is motivated by a perceived superiority of Morningstar’s analysts’ ratings over Morningstar’s algorithmic ratings
  - However, there doesn’t appear to be much evidence of this, since the analyst ratings themselves don’t seem very useful

> How to re-frame the paper to be more consistent with the results?

1. Implement a story that doesn’t require analyst ratings to be more useful
   - Difficult to come up with one
2. Take a more agnostic and exploratory view
   - Seeing the relative performance analyst and quantitative ratings
   - Harder to motivate, but given that Morningstar produces both ratings, it is still an interesting setting



Main comment #2

Replicating Morningstar’s ML method

Replicating the quantitative rating

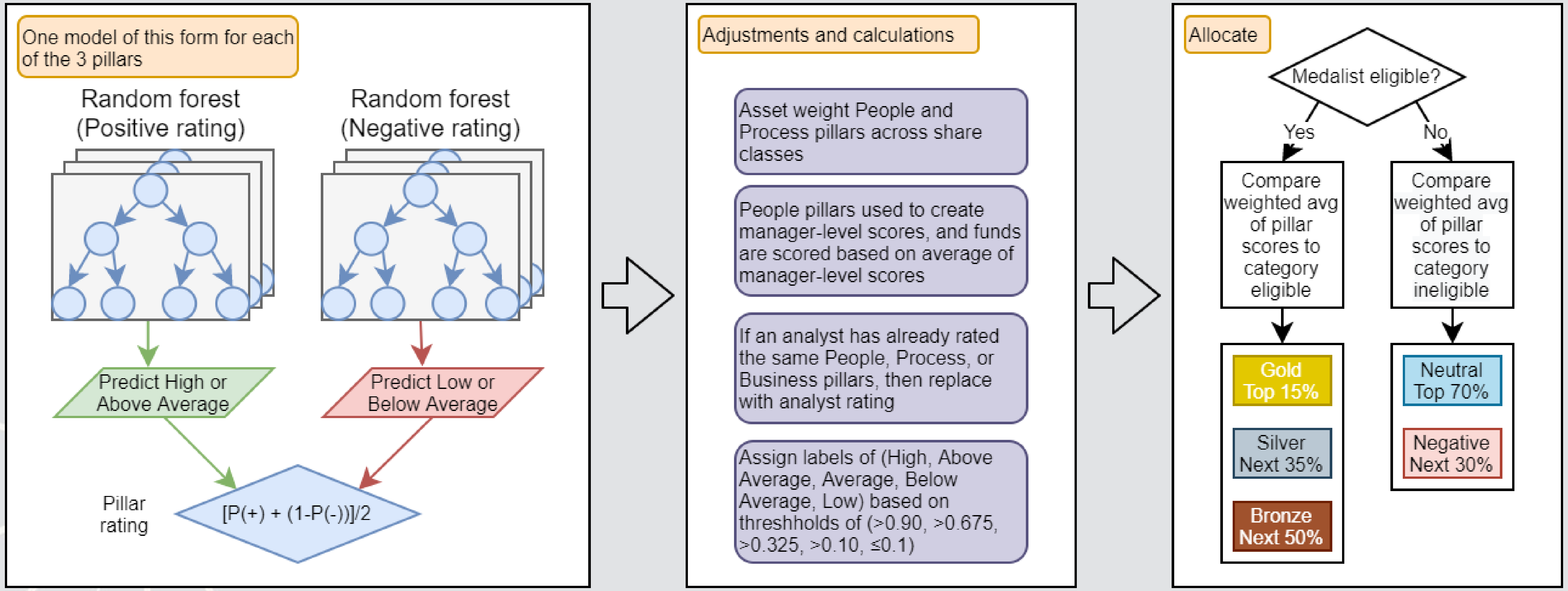
“Specifically, we train random forest models to predict analyst rating and apply to all funds. Our self-constructed quantitative rating starts from 2014, as we require at least 3 years for the training sample, and the forward rolling is made on a monthly basis.”

> The above is the sole description of the replication in the paper. What *exactly* is implemented? Is it following Morningstar’s method 100%?

- Furthermore, note that the method is a moving target!
  - The methodology in 2018 is relatively simple
  - The methodology in 2020 is a bit more refined, with attention paid to certain incongruent behaviors between a model and analyst behavior

Morningstar’s 2020 methodology

> Morningstar is *very thorough* in their documentation! They list the model inputs in the appendix too.



One model of this form for each of the 3 pillars

Random forest (Positive rating)

Random forest (Negative rating)

Predict High or Above Average

Predict Low or Below Average

Pillar rating

[P(+) + (1-P(-))]/2

Adjustments and calculations

Asset weight People and Process pillars across share classes

People pillars used to create manager-level scores, and funds are scored based on average of manager-level scores

If an analyst has already rated the same People, Process, or Business pillars, then replace with analyst rating

Assign labels of (High, Above Average, Average, Below Average, Low) based on threshholds of (>0.90, >0.675, >0.325, >0.10, ≤0.1)

Allocate

Medalist eligible?

Yes

No

Compare weighted avg of pillar scores to category eligible

Compare weighted avg of pillar scores to category ineligible

Gold Top 15%

Silver Next 35%

Bronze Next 50%

Neutral Top 70%

Negative Next 30%

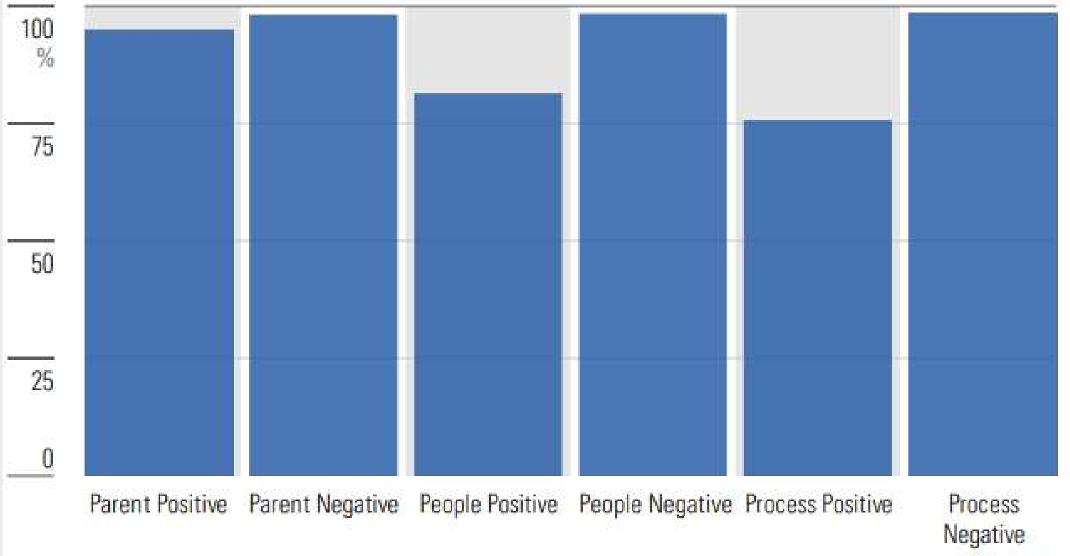
How well does the replication work?

- Easy to validate: Show the performance of your replication in predicting Morningstar’s quantitative ratings
  - What percent are correctly classified? Is the replication better for Gold ratings or other ratings? Can show a confusion matrix.
  - It would also be good to compare against analyst ratings, e.g., where does it agree and disagree?

> If you can replicate the method well, then you can reasonably extend the sample to earlier years or an overlapping sample

Morningstar has done some validation of their own which you can also benchmark against.

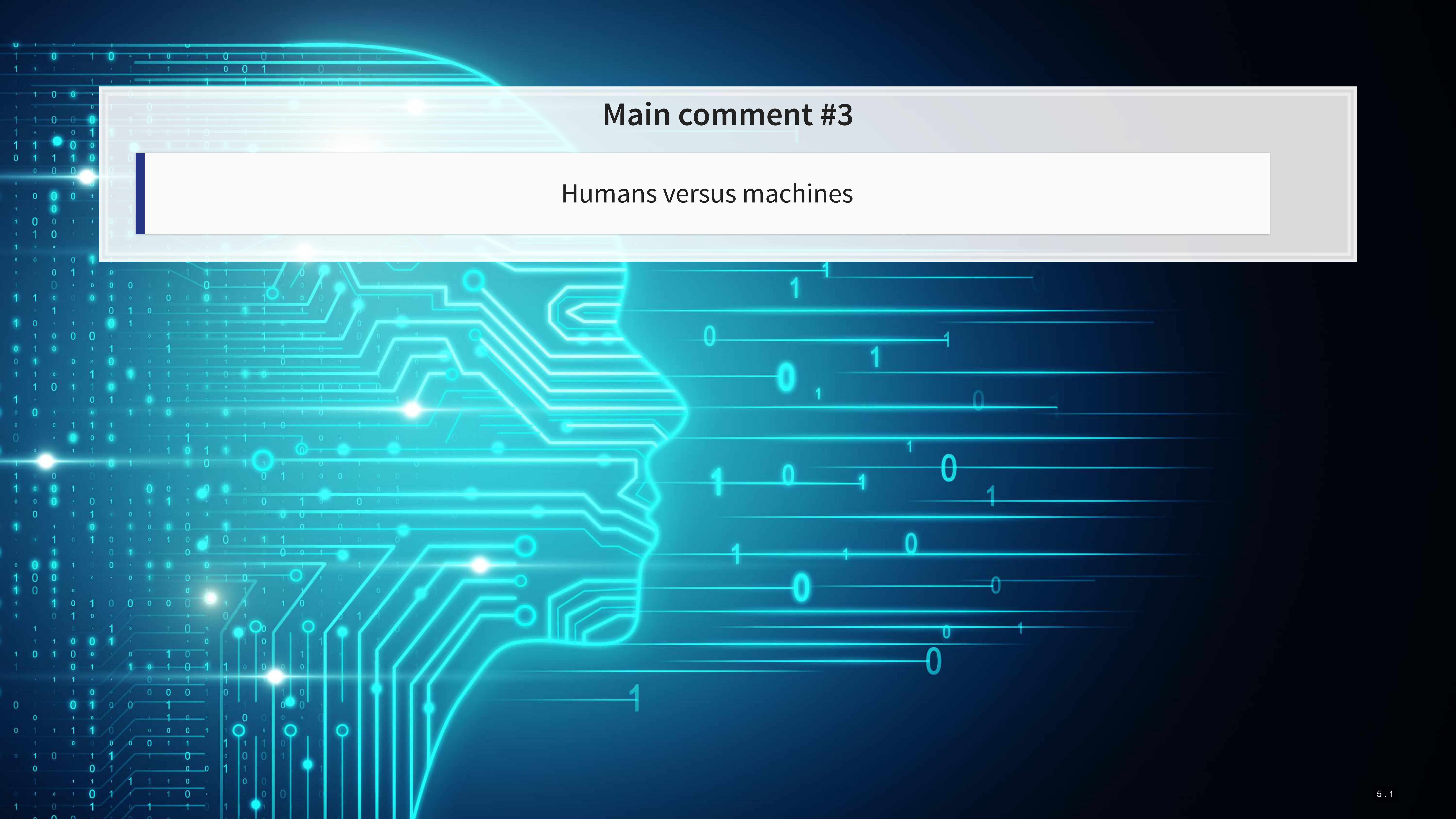
Exhibit 2 Percentage Accuracy Between Quantitative Pillar Ratings and the Analyst Pillar Ratings



Source: Morningstar, Inc. Data as of March 31, 2019.

Main comment #3

Humans versus machines



Humans vs machine in the context of the results

> Since minimal results show human (or machine) superiority, why might we care about their difference?

1. There are some sample differences that make the prior tests perhaps not wholly convincing
2. The analyst rating is not the *only* output by the analysts – maybe all hope is not lost on the human side
    - Tone is used to try to tease this out

> One other caution: Remember that the ML method is trying to *replicate analyst ratings*. It is not trying to *beat* analyst ratings. E.g., it is trained to replicate analyst's rating decisions, not optimal rating choices.

What exactly is tone measuring

	Independent Variable	Dependent Variable
Theory	Analyst soft information collection	Usefulness of information
Empirics	Tone of analyst report	Style adjusted 1 month return

Note:

Libby box of tone test (Table 5)

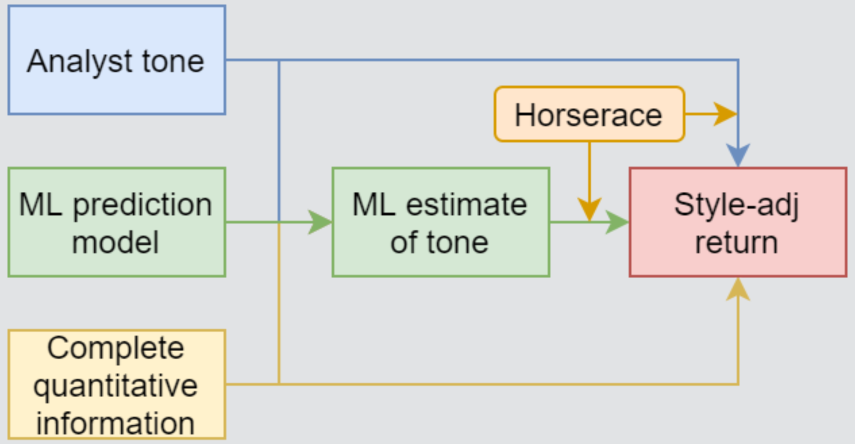
> Tone is at best a very tangential measure of soft information collection

- Information generation is a latent process, unless you can observe analyst actions
  - E.g., like Cheng, Du, Wang, and Wang (2016 RAS) in the context of China
- We could potentially tease out latent information by controlling for all public information
- The generating process of tone itself is latent
- At best, we can say tone is a rough proxy for soft information *generation*
  - There are better proxies, such as the LDA-based proxy from Shen, Jiang, and Kong (working)

Testing soft information collection via tone

> Option 1: Build an ML model to predict the tone of analyst reports, add that prediction in, and then see if the original tone measure is *incrementally* useful over the ML one

- Pro: With a well designed ML model, can convincingly show analyst report tone is [not] quantitatively driven
- Pro: Puts the analyst and machine on even footing – both get to generate the same metric
- Pro: Better relates to “human vs. machine” motivation
- Con: Difficult to properly model the machine learning side; it would need to be a custom design that is sufficiently precise and fine tuned to convince readers that this is a fair contest

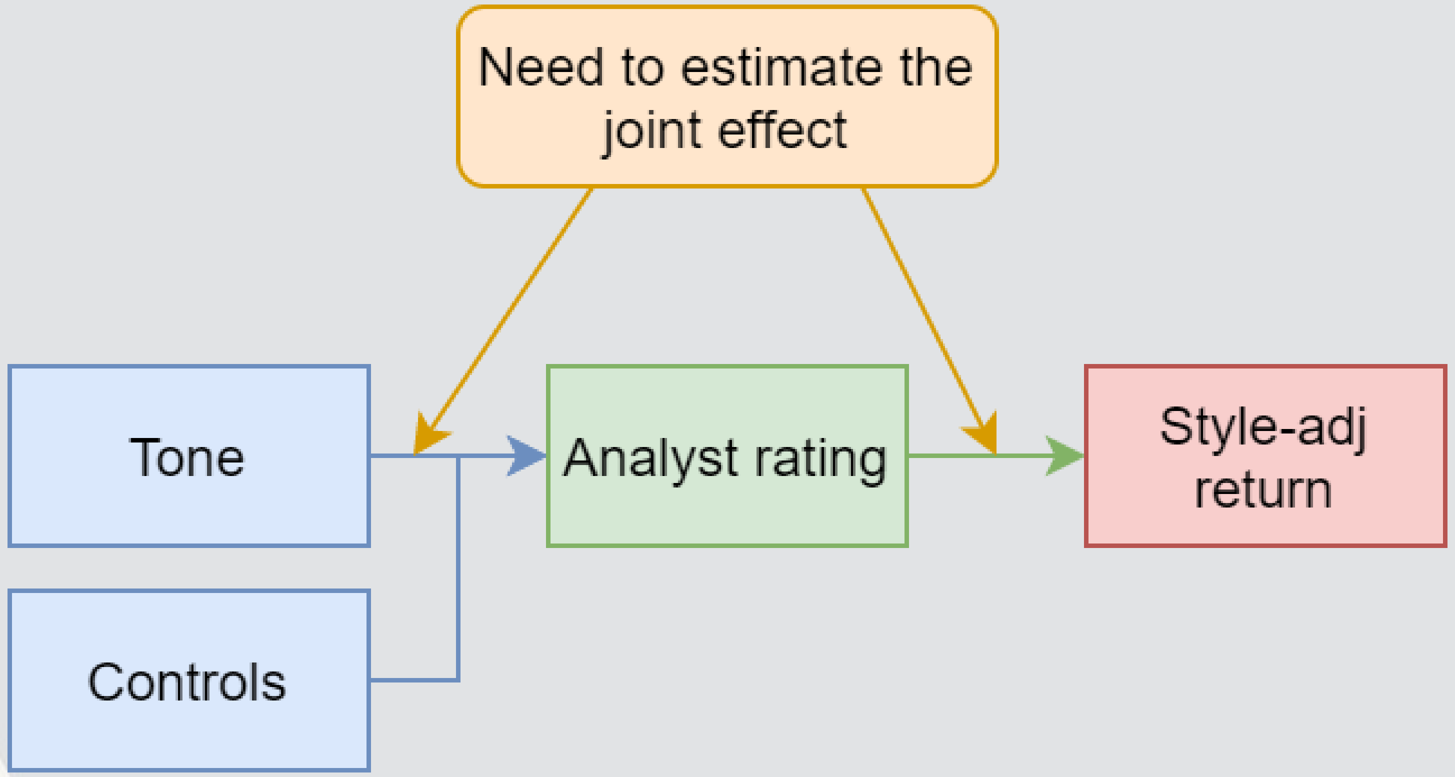


> E.g., instead of stating “the soft information acquired by analysts […] cannot be easily captured by the more sophisticated machine learning algorithm,” directly test it.

Testing soft information collection via tone

> Option 2: Isolate the impact of tone on analysts’ categorization and then show that said isolated component is incrementally valuable over the ML model’s prediction.

- Pro: Easier to test
- Con: Less direct – you would need the supposed soft information collection via tone to impact at least some analyst ratings
- Con: Would need to use something like SEM (Structural Equation Modeling) to test convincingly



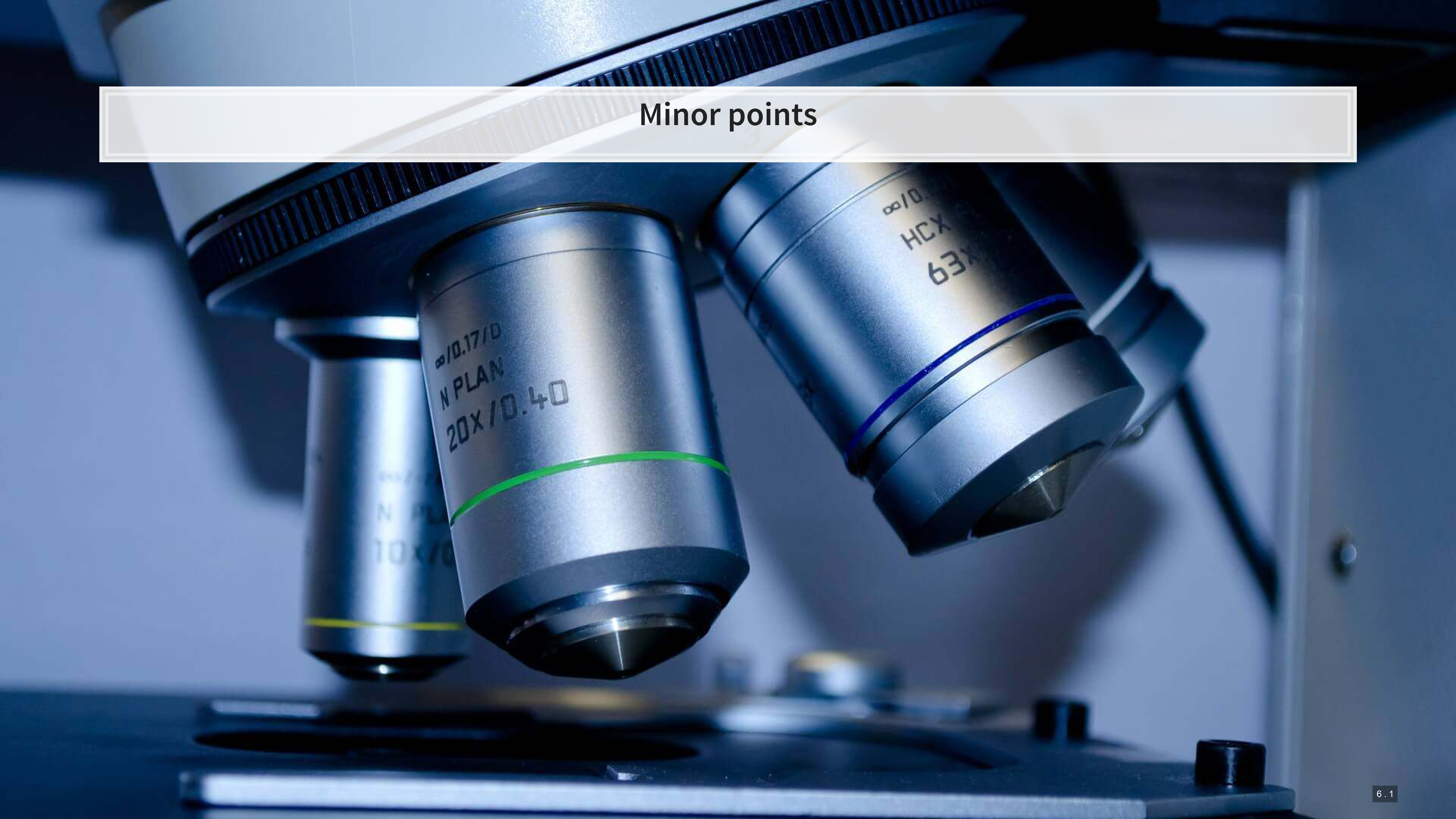
Some interesting but detached results

1. Full-report tone appears to largely lead to fund flows based on institutional investors
2. Tone of the summary and title appear to drive fund flows by individual investors
3. Looking at the tone by section provides a bit more context to the results
   - Proc: Adds some *context* to the measure!
   - Con, do we expect the tone measure to be equally well-specified for each section?

> Intuitive result, and an interesting example of processing costs or *frictions*

- May want to back away from calling investor behavior irrational though
  - Is it rationally worthwhile for your average investor to pay $199 to subscribe to Morningstar?
    - Plus, this may be why the star rating is more reacted to: it's free!
  - Perhaps this analysis substantiates why Morningstar can charge for Analyst Ratings though

Minor points



Tables

- A lot of tables aren’t very valuable/useful, e.g.:
  - Tables 4 and 6 aren’t very informative or don’t directly add to the paper’s story
  - Tables 5 and 7 could be cut down and combined into 1 table
  - Table 9, as is, is not very useful, though perhaps using it as a proper DID would be more convincing. That being said, it is very endogenous – is it about analyst choice, or is any effect because market participants noticing the analyst’s choice?
- “Unreported results confirm that the differences in those fund characteristics are statistically significant.” (p11) – this seems more important than the star rating comparisons in Table 1 panels A1, A2, B1, and B2. Why not focus more on this?
- For univariate comparisons, statistical tests (t-tests or the like) should be used to substantiate differences being significant

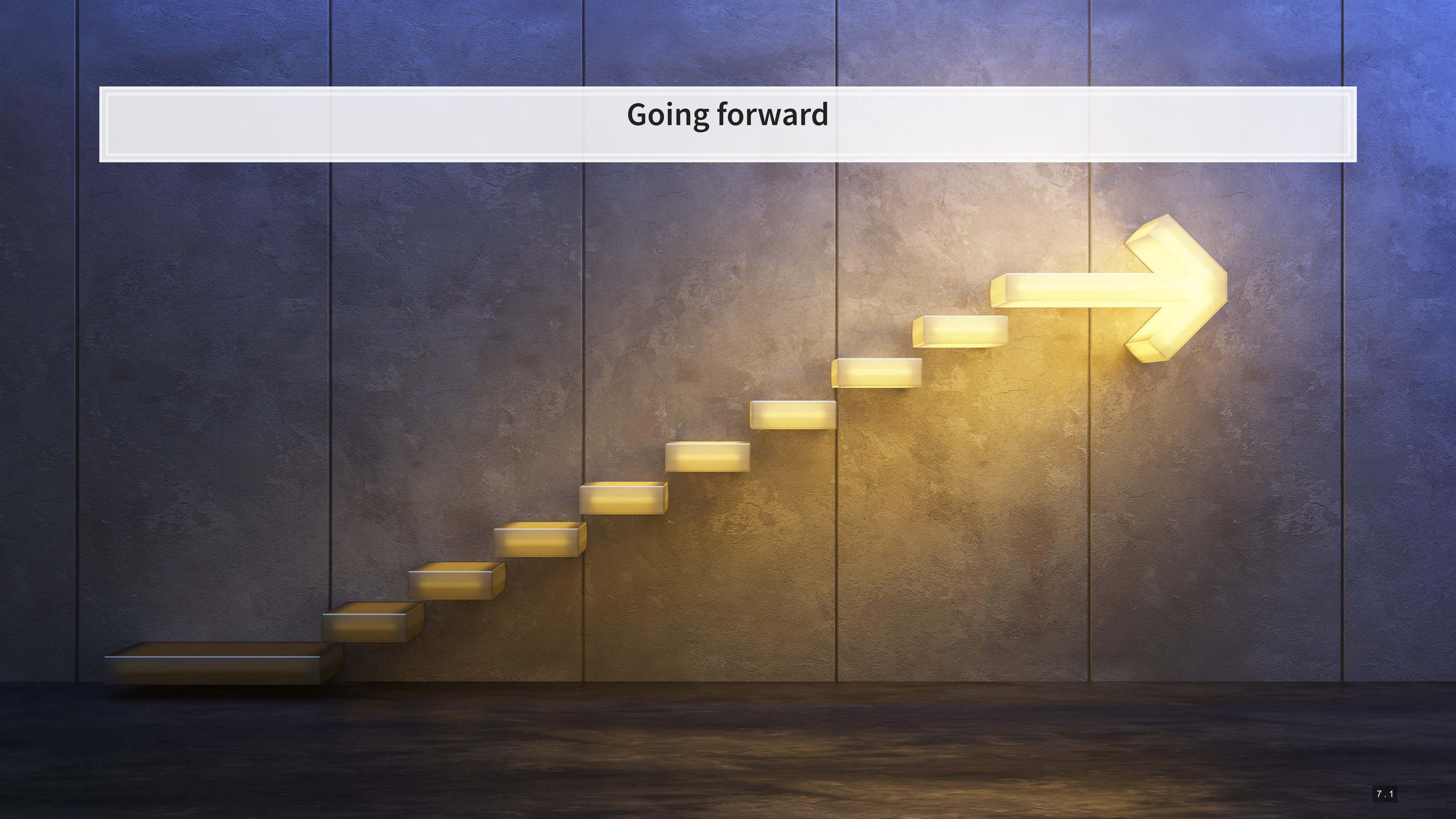
Sample representation

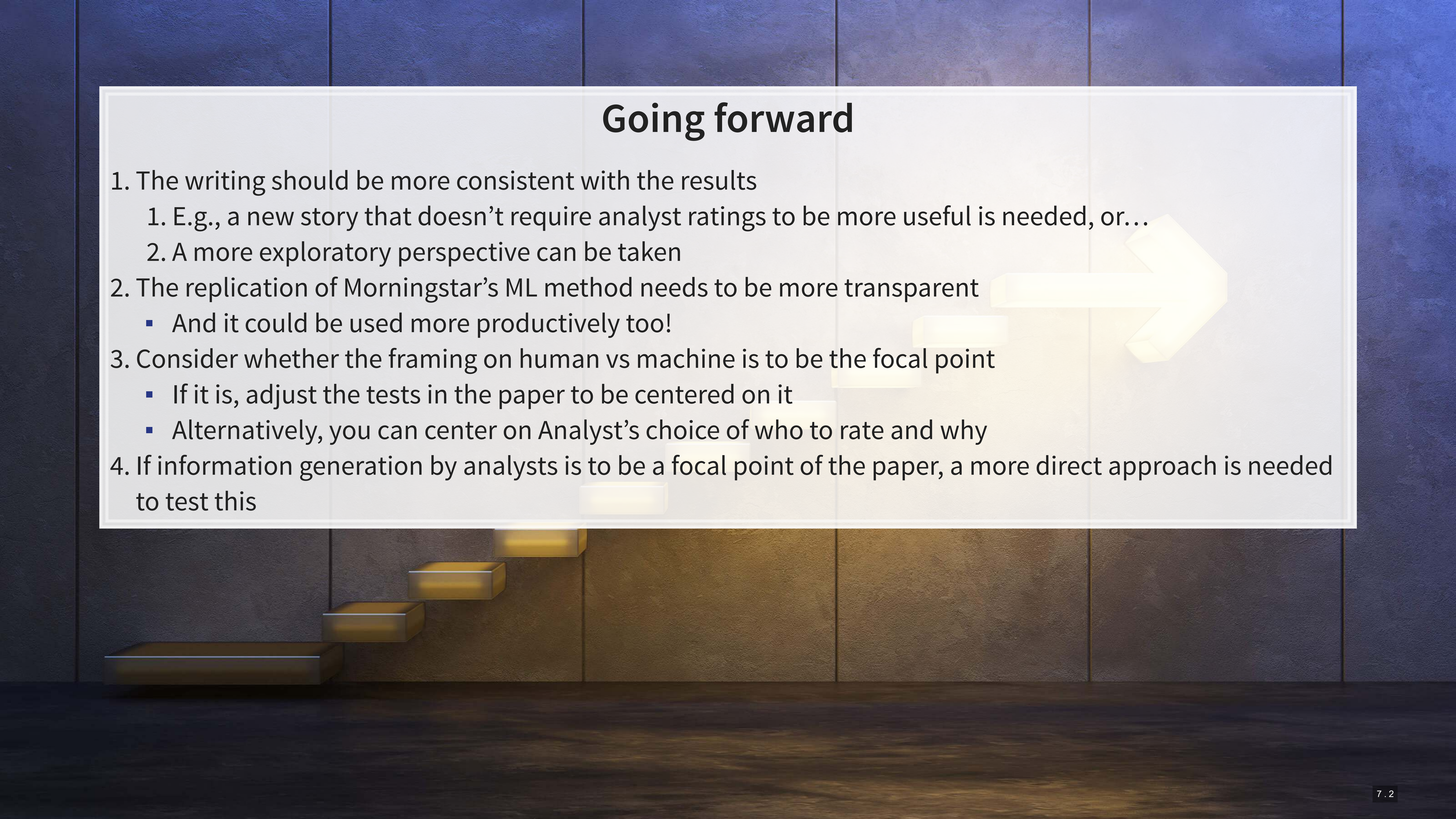
- If the algorithm was made in 2017, why include a sample of analyst info from 2011 to 2018 in most tests? It is irrelevant except when using the replicated quantitative rating (where 2011-2013 are still irrelevant)
- Table 1, Panel A2: % of Negative vs Neutral (1.5% vs 31.5%, i.e., 4.5/95.5) doesn't match Morningstar's allocation of 30/70 (21.4/88.6) for these categories in 2020 (2018). Medal classes are closer – should be 15/35/50 (16.7/33.3/50) for Gold/Silver/Bronze in 2020 (2018), but is instead 19.1/33.0/47.9 (slightly overweight on Gold)
  - On the other hand, Table 1, Panel B2's % of Negative vs Neutral should be 30/70, and is pretty much there. Similarly, the percent of Gold/Silver/Bronze is 12.5/34.9/52.7, which is quite close to the 15/35/50 target.
- Panel C1 – how are missing values handled? It appears that they are set to zero, since the median score (0) for Analyst rating couldn't be attained from the actual distribution from Table A1 otherwise.
- For the quantitative rating replication, did you stick to the time horizon used by Morningstar? They disclose that they based the algorithm on a sample of 10,000+ ratings – is your sample that large?

Writing

- The numbers in the discussion of Table 2, Panel A and Panel B, don’t match the table at all?
- Footnote 19: Morningstar has done a (presumably) out-of-sample test in their 2021 methodology document: https://www.morningstar.com/content/dam/marketing/shared/research/methodology/813568-QuantRatingForFundsMethodolgy.pdf
- First paragraph of p16 – generally unwarranted, since your results don’t show general superiority of human ratings in this context. The sentence starting with “For instance, it could be helpful” is however quite correct.
- “Compared to the widely adopted star rating, the analyst rating serves as a better tool to facilitate capital allocation for mutual fund investors.” – Internet Appendix 1 disagrees, since a five star rating appears to be statistically significant in both a 1 month and 1 year context.

Going forward



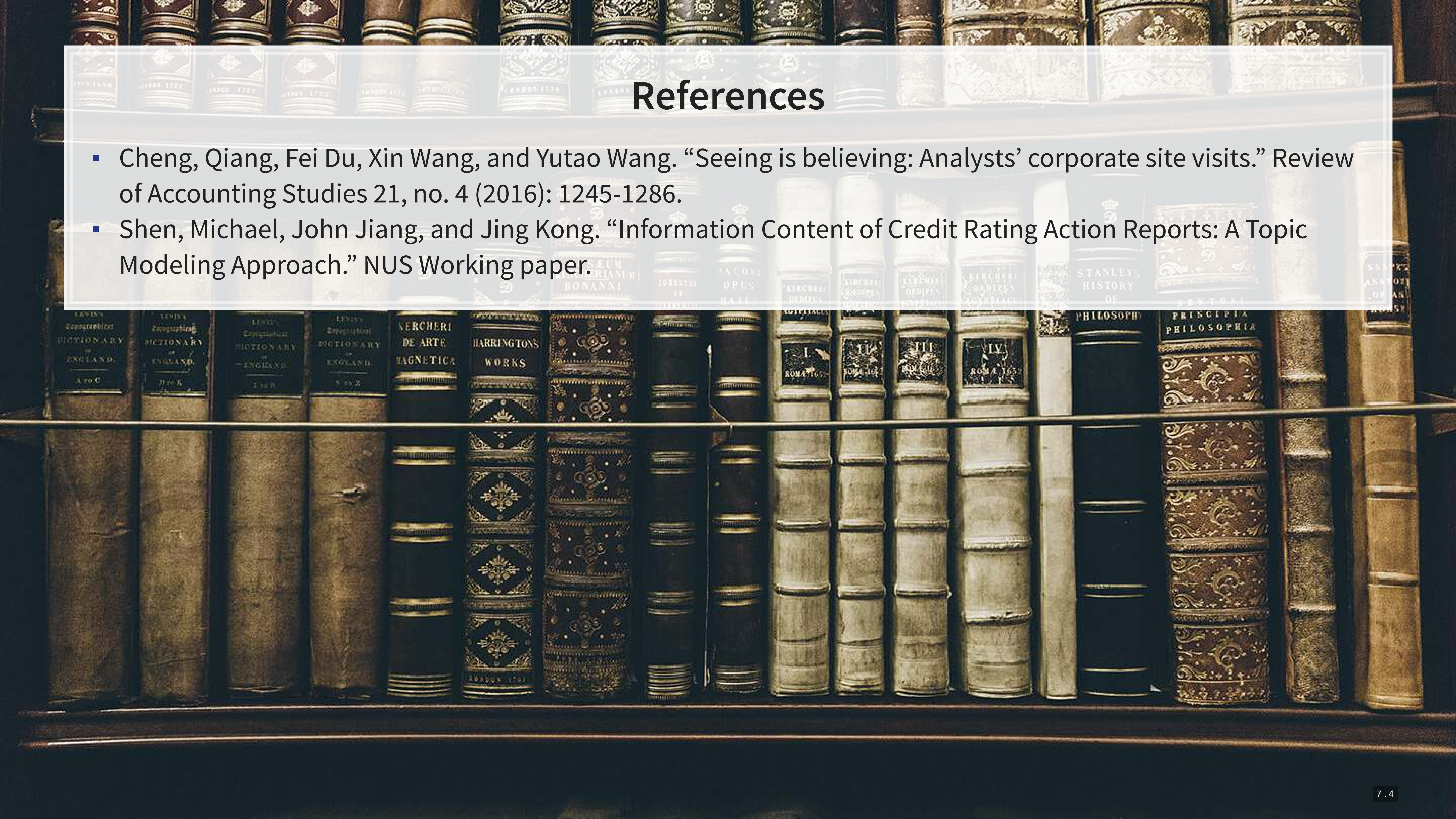


Going forward

1. The writing should be more consistent with the results
    1. E.g., a new story that doesn’t require analyst ratings to be more useful is needed, or…
    2. A more exploratory perspective can be taken
2. The replication of Morningstar’s ML method needs to be more transparent
    - And it could be used more productively too!
3. Consider whether the framing on human vs machine is to be the focal point
    - If it is, adjust the tests in the paper to be centered on it
    - Alternatively, you can center on Analyst’s choice of who to rate and why
4. If information generation by analysts is to be a focal point of the paper, a more direct approach is needed to test this

Thanks!

Richard M. Crowley
Singapore Management University
https://rmc.link/
@prof_rmc



References

- Cheng, Qiang, Fei Du, Xin Wang, and Yutao Wang. “Seeing is believing: Analysts’ corporate site visits.” Review of Accounting Studies 21, no. 4 (2016): 1245-1286.
- Shen, Michael, John Jiang, and Jing Kong. “Information Content of Credit Rating Action Reports: A Topic Modeling Approach.” NUS Working paper.

Packages used for these slides

- kableExtra
- knitr
- revealjs