Using Machine Learning to Analyze Disclosure Narratives

1st Annual Rotman CPA Ontario Centre for Accounting Innovation Research Conference

Jared Jennings

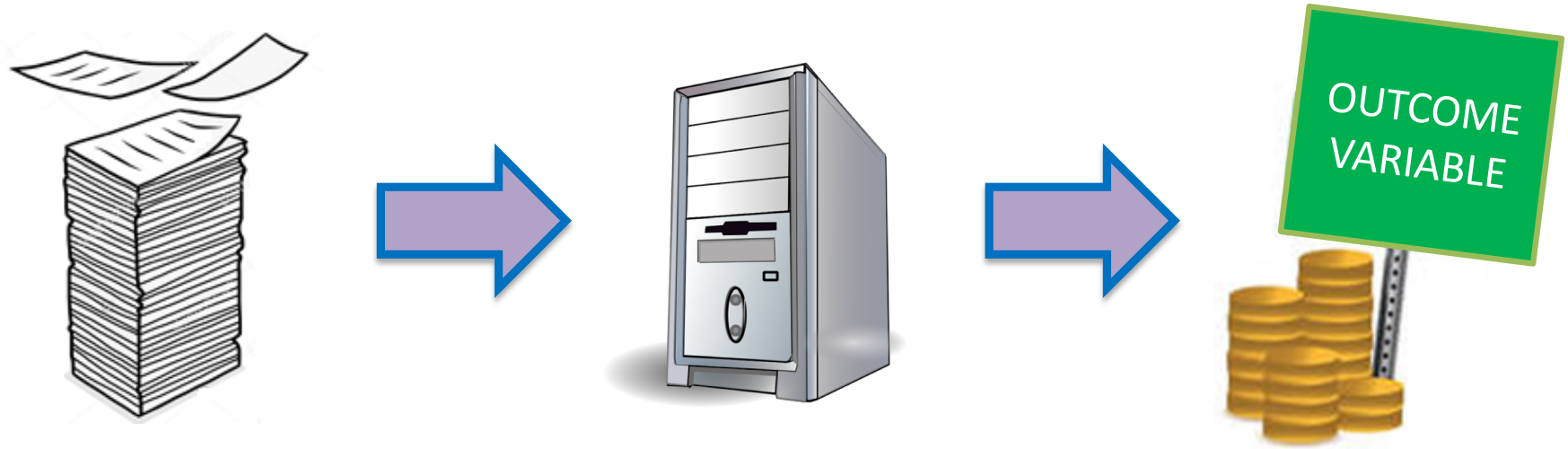
Washington University in St. Louis

Importance of Machine Learning (ML)

- "Machine learning combined with natural language processing can tell portfolio managers how bullish a CEO sounds in an earnings call *by mining transcripts for specific language it was trained to identify.*" (Barron's 4/7/18)

- "What we're not doing is automating investing decisions. We're exploring and trying to enhance our existing models…." (Tim Cohen – Fidelity)

What is supervised ML?



Language from Firm Disclosures

1) Future cash flows
2) Future earnings surprises
3) Credit Default Swaps (CDS)

Advantages of ML?

- Allows the researcher to examine relations that might not be possible with standard statistical techniques
- Automates the identification of narrative patterns using minimal researcher intervention
- May help to identify relevant words that are unknown or difficult to identify
- Researchers can apply ML to a variety of
  - Outcome variables
  - Languages
  - Disclosures

Disadvantages of ML?

- Relies on statistical methods to build a model vs. researcher intuition
  - Many words identified may not intuitively relate to the outcome variable
  - May introduce “noise” into our measure

- Several things can be done to reduce the likelihood that one is simply identifying a statistical relation
  - Examine word lists
  - Hold-out sample (i.e., out-of-sample tests)

Machine-learning Methods

- Support vector regressions (SVR)
- Random Forest Regression Trees (RF)
- Supervised Latent Dirichlet Allocation (sLDA)
- Unsupervised Latent Dirichlet Allocation (LDA)

Machine Learning Techniques (SVR)

$Outcome\ Variable_{i,q+1} = w_0 + \boldsymbol{w} \cdot \boldsymbol{x_{i,t}} + e_{i,t}$

– Where **x** is a matrix of one and two-word phrase counts and

w is a vector of regression coefficients

w cannot be estimated with OLS

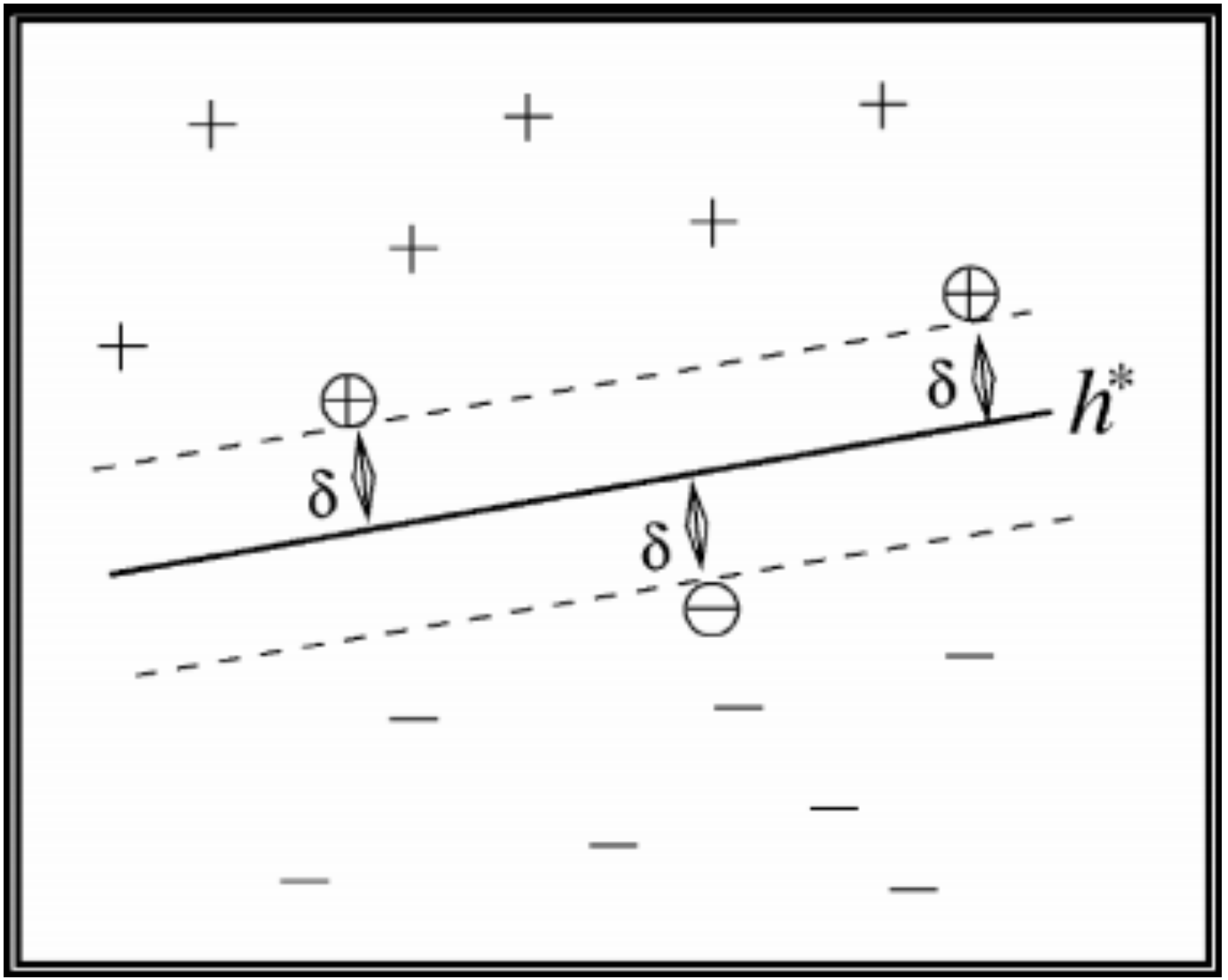
SVR can estimate ***w***

Coefficient Vector Magnitude

Prediction Error

$$\underset{\omega}{minimize}\ \frac{1}{2}\|\omega\|^2 + C \sum_{t \in Train}^{n} g_\epsilon(e_t)$$

Machine Learning Techniques (SVR)



Machine Learning Techniques (sLDA)

- sLDA jointly models the language in documents and a response variable
  - Finds latent topics that best predict responses for out-of-sample documents (Blei and McAuliffe, 2007)

- sLDA identifies ***predictive*** topics by:
  - Assessing the co-occurrence of words within documents
  - Allowing the response to be a function of the topic frequencies in the documents

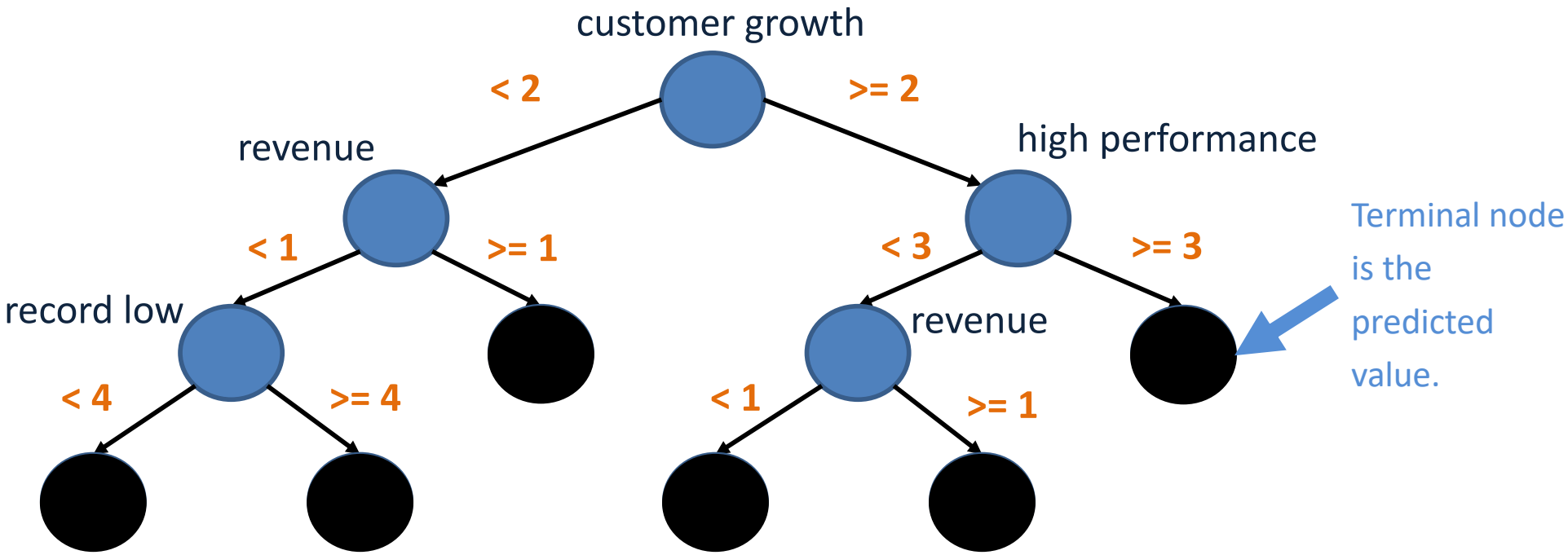
sLDA vs. LDA

- LDA – text categorization
  - Unsupervised LDA topics better at identifying genres (e.g., drama, action, horror)

- sLDA – prediction
  - Supervised LDA topics better at identifying sentiment (e.g., excellent, terrible, average)

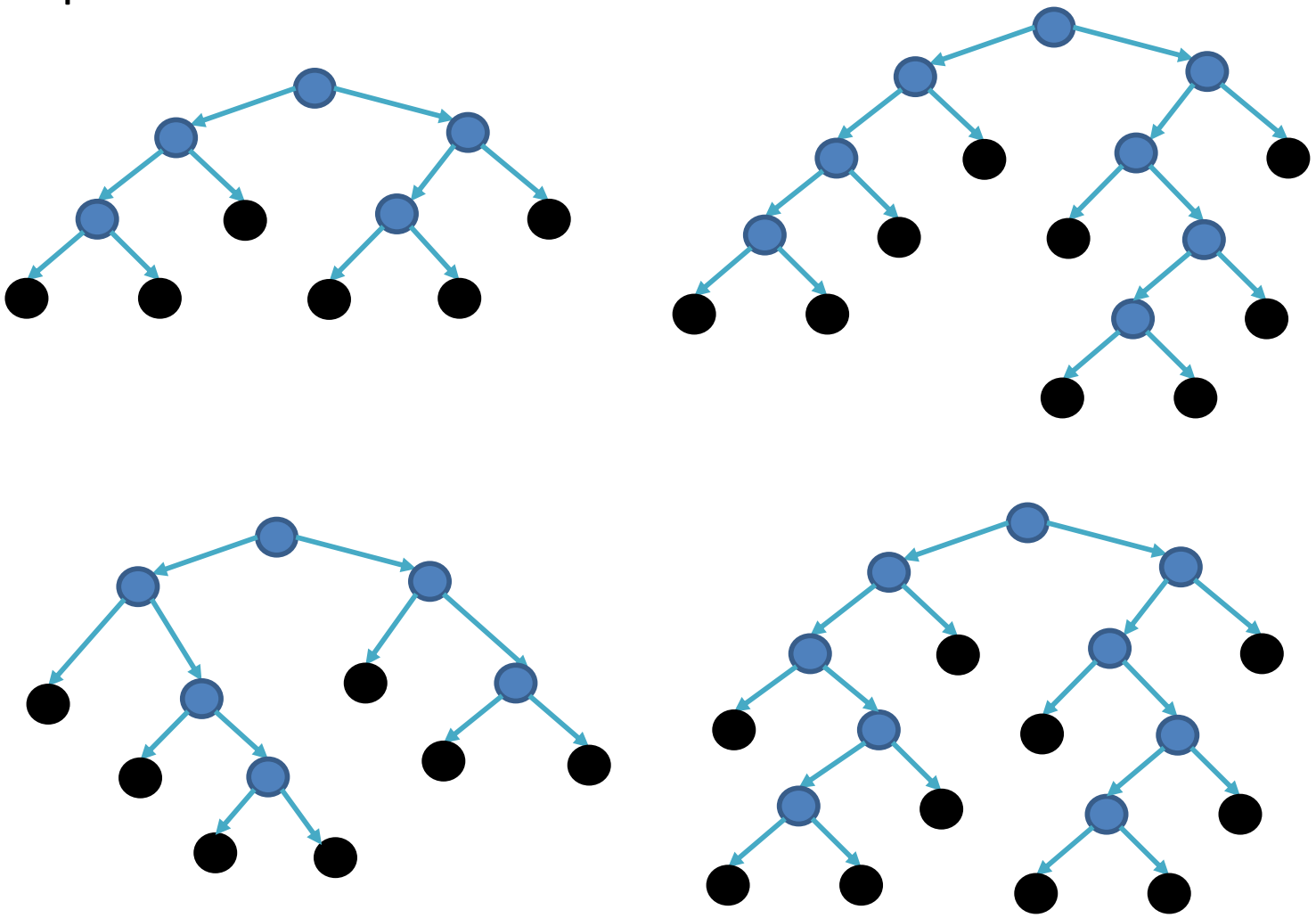
Machine Learning Techniques (RF)

- Randomly select a subset of observations and a subset of all available features (i.e., one- and two-word phrases)
- Builds a tree picking the partition at each node that minimizes the dependent variable's in-sample sum of squared error within the resulting subsets



Machine Learning Techniques (RF)

Repeat



The average predicted value for all trees is the final predicted value.

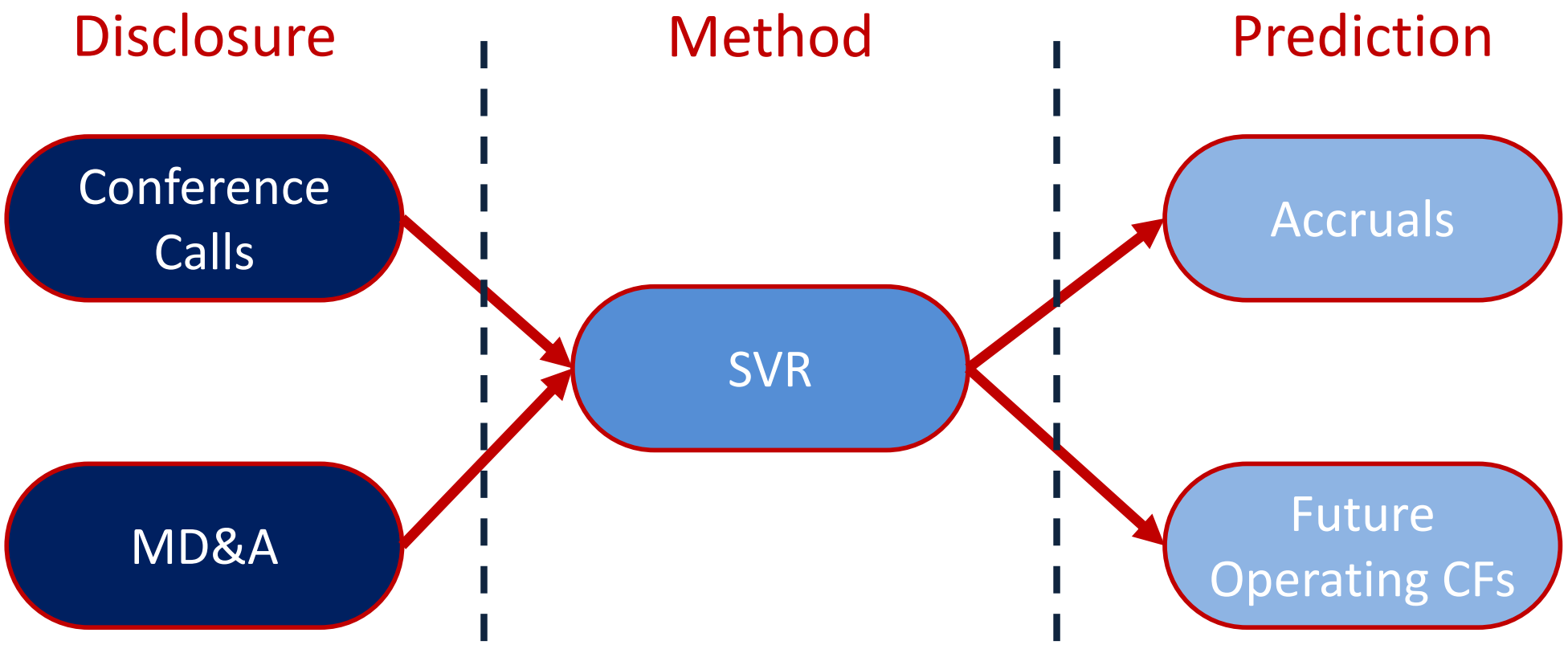
Research Papers

- Frankel, R., Jennings, J., Lee, J. 2016. Using unstructured and qualitative disclosures to explain accruals. Journal of Accounting Economics, 62(2), pg. 209-227.

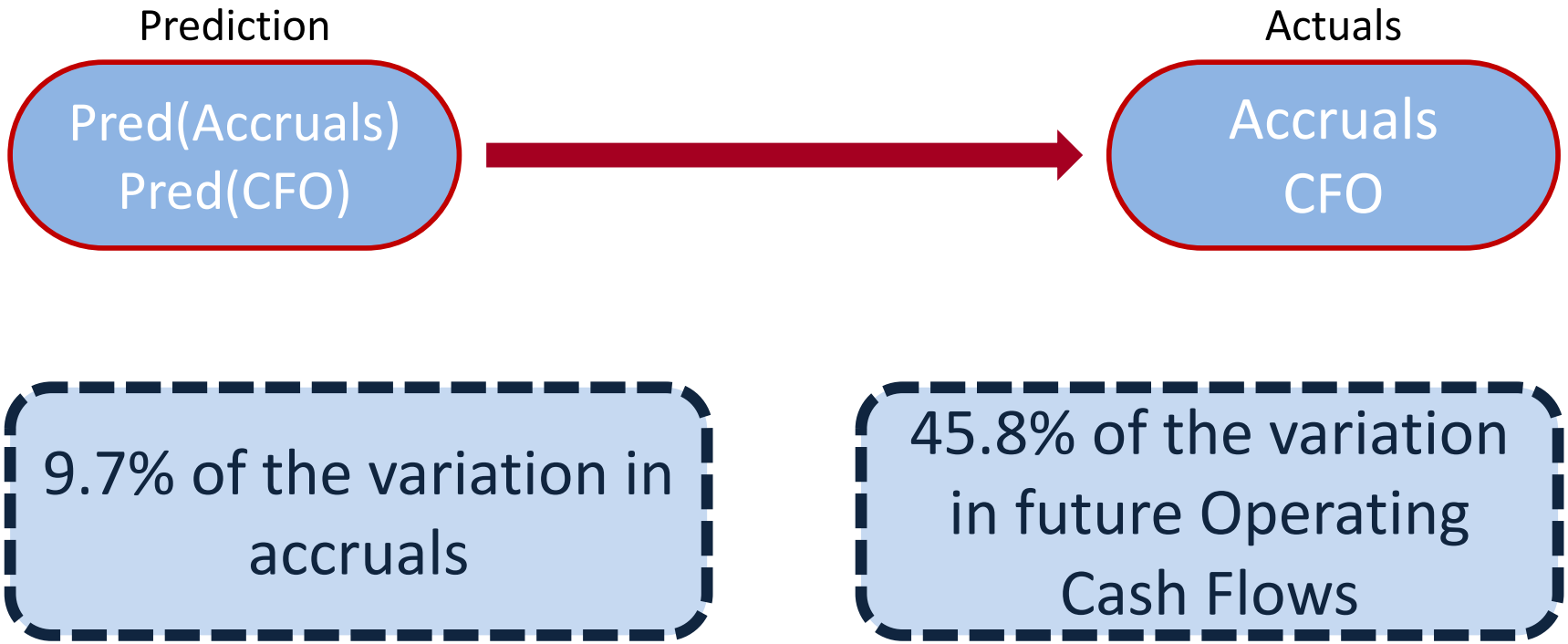
- Donovan, J., Jennings, J., Koharki, K., Lee, J. 2018. Determining credit risk using qualitative disclosure. Working Paper.

- Frankel, R., Jennings, J., Lee, J. 2018. Assessing the relative explanatory power of narrative content measures using conference calls, earnings predictions and analyst revisions. Working Paper.

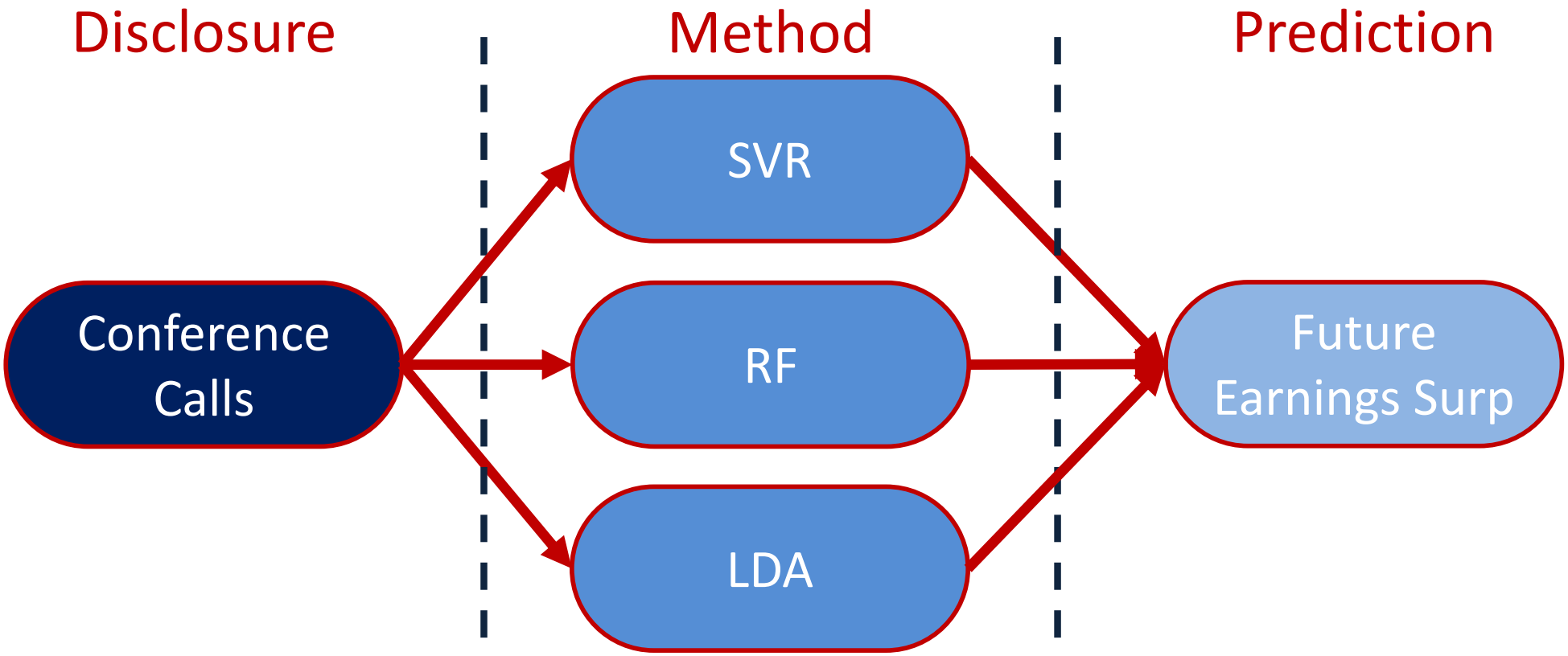
Frankel, Jennings, Lee (2016)



RESULTS: Frankel, Jennings, Lee (2016)



Frankel, Jennings, Lee (2018)

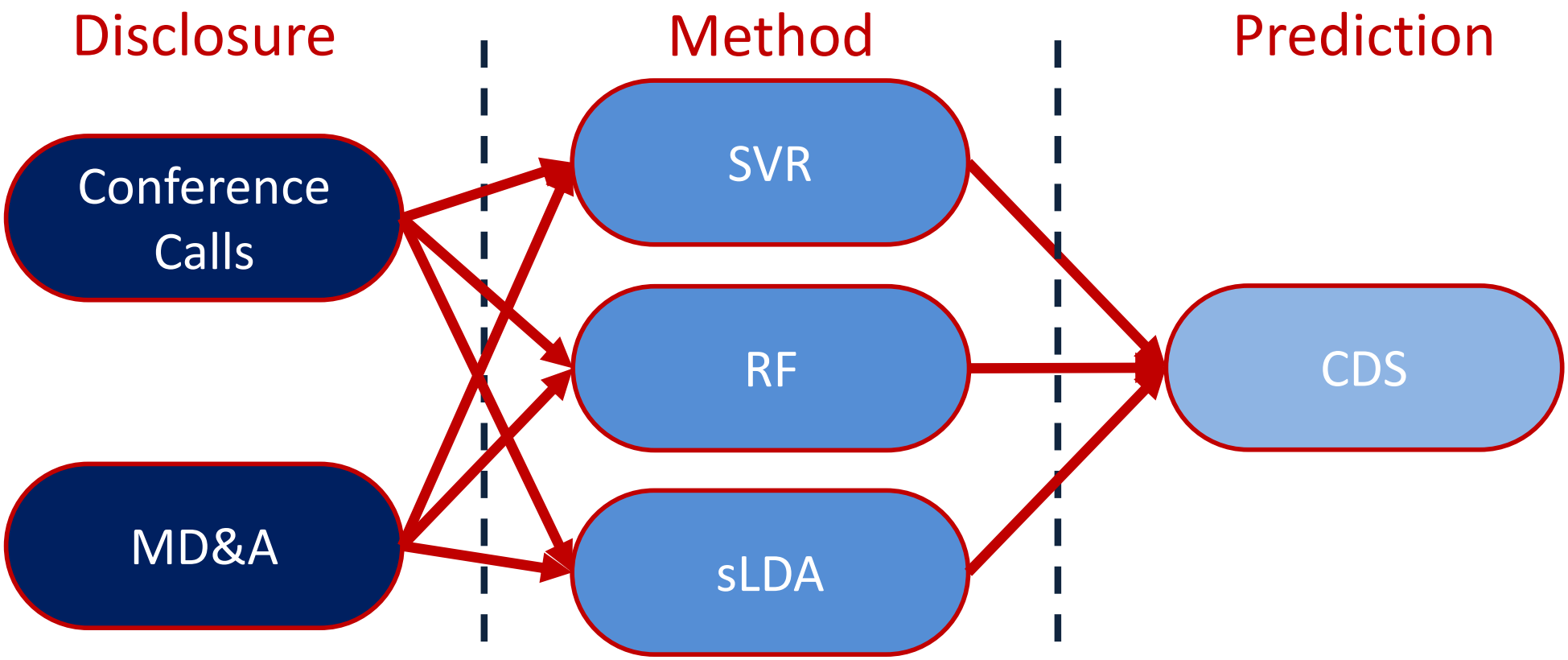


RESULTS: Frankel, Jennings, Lee (2018)

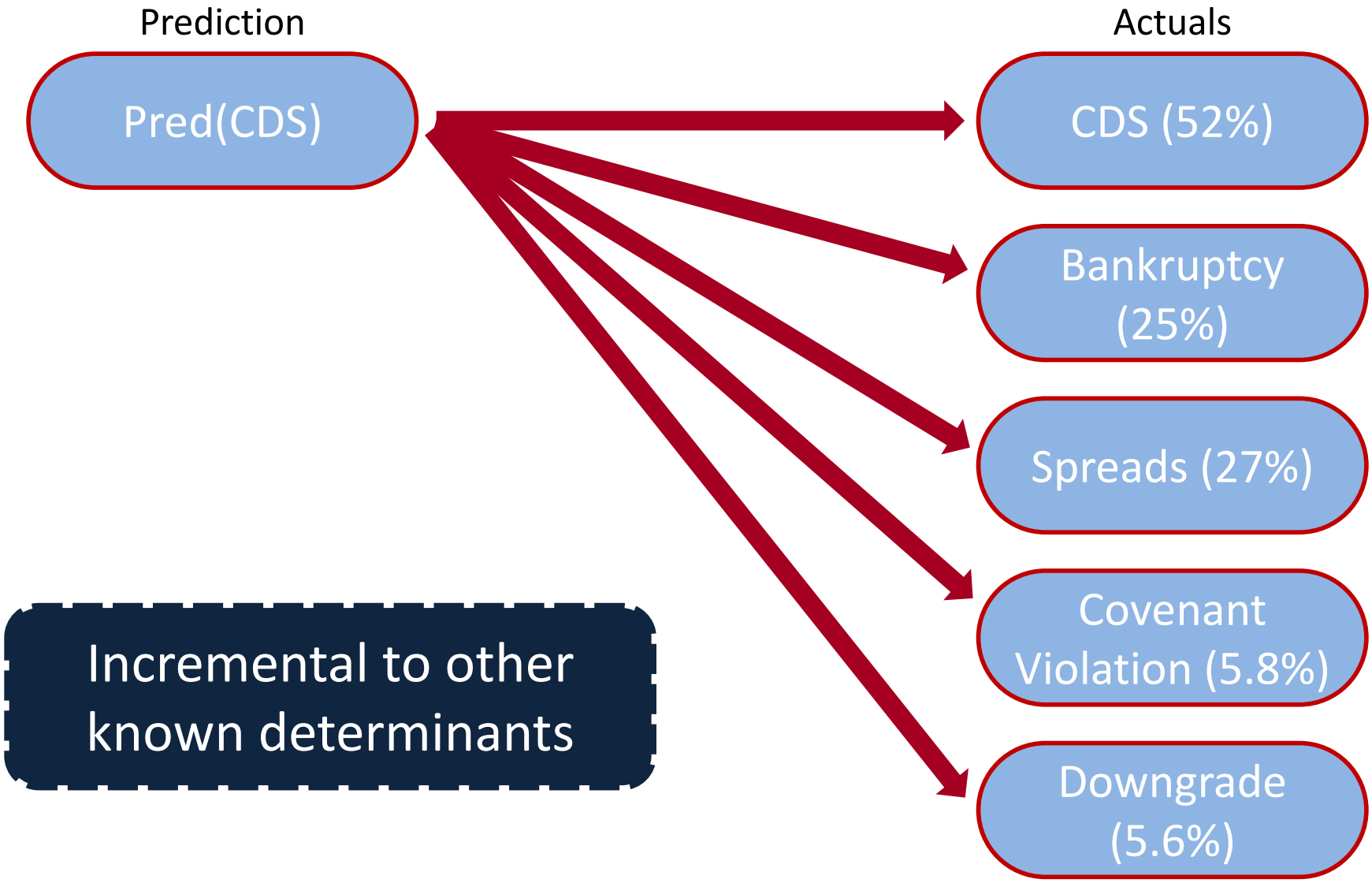
7.6% of the variation in the future earnings surprise

Incremental to other known determinants

Donovan, Jennings, Koharki, Lee (2018)



RESULTS: Frankel, Jennings, Lee (2018)



Takeaways

- ML can be useful when enhancing existing models
- ML can explain an economically significant portion of the variation in an outcome variable
- ML can identify information that is not captured by other accounting or economic signals

THANK YOU!