“Pattern Recognition and Anomaly Detection in Bookkeeping Data”

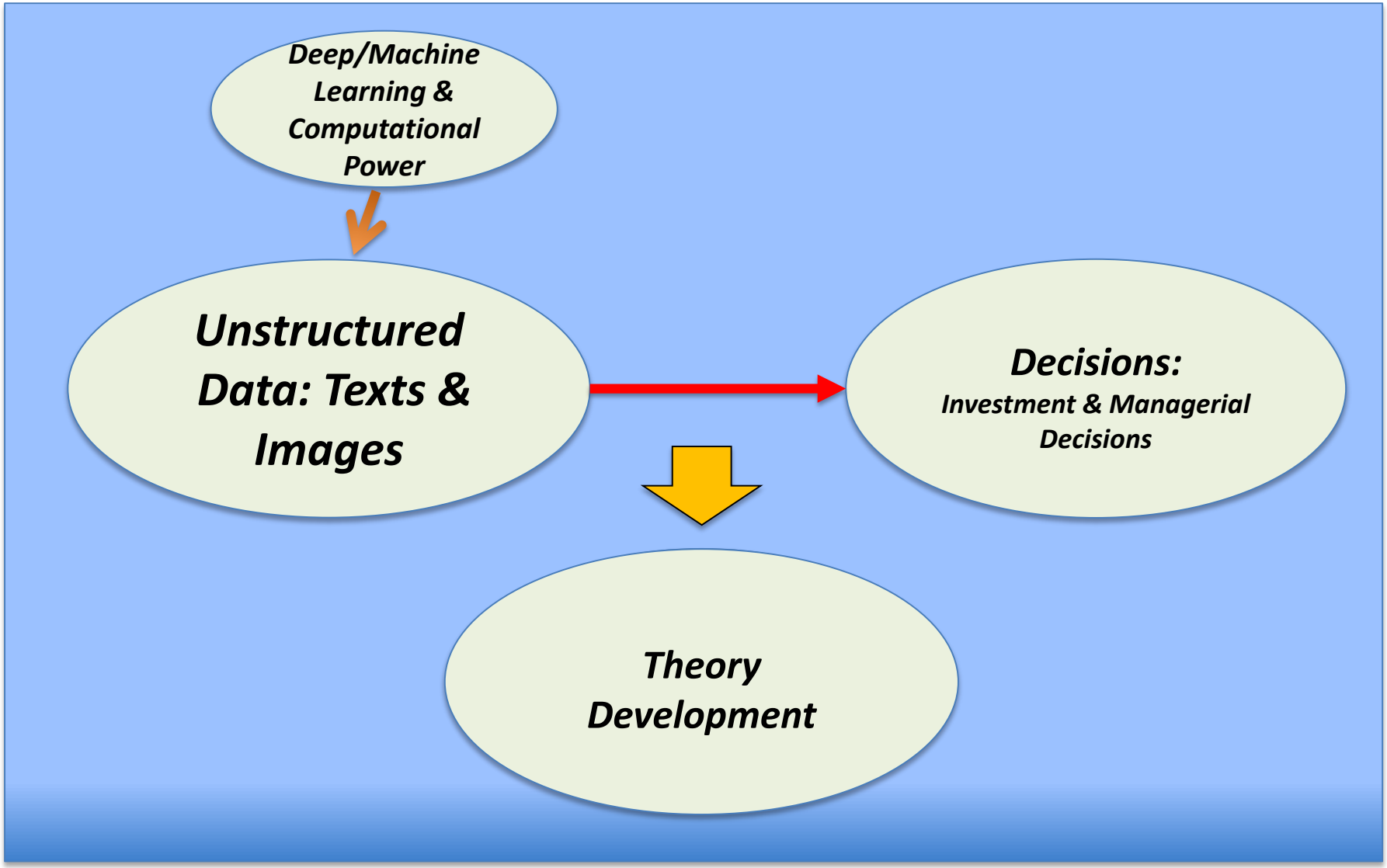
by Liang, Wang, Akoglu and Faloutsos

Discussed by Sean Cao

Georgia State university

May 26, 2021

Information Retrieval



Common Challenges in ML studies

- What is the contribution to economic theories in addition to methodology advancement?
  - The paper is important for both industry professionals and academics
- Why only this technology can do the job? Why not others?
- Can we make the technology understandable to a broad audience beyond the CS audience?

See more at 2020 GSU-RFS Conference summarizing 300 ML papers

2020 GSU-RFS FinTech Conference

*with **dual submission option to RFS***

February 28-29, 2020 | Atlanta, Georgia

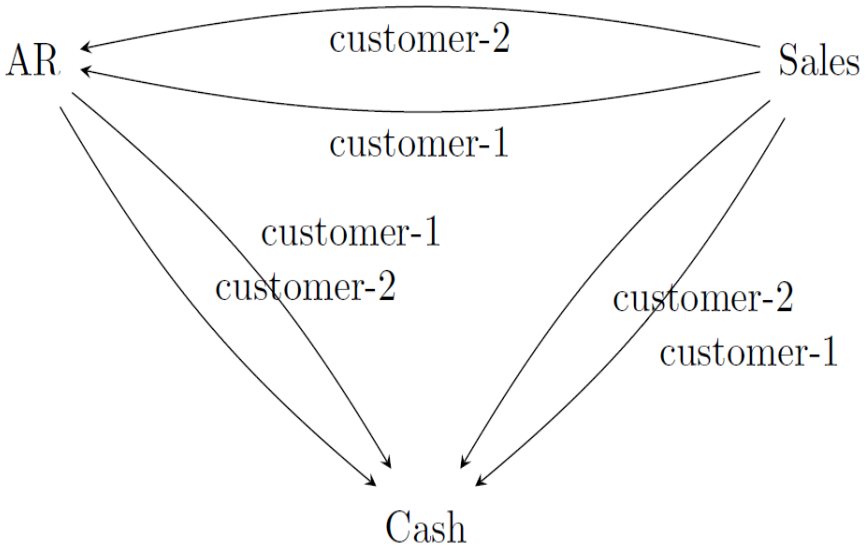
***Sponsoring Editors of the RFS: Itay Goldstein** (University of Pennsylvania) and **Manju Puri** (Duke University)*

Georgia State University

The Review of Financial Studies

Interpretation of the technology

- **Transaction level data**
  a. Represent transaction level data using graphs
  b. MDL (unsupervised) to detect anomalies



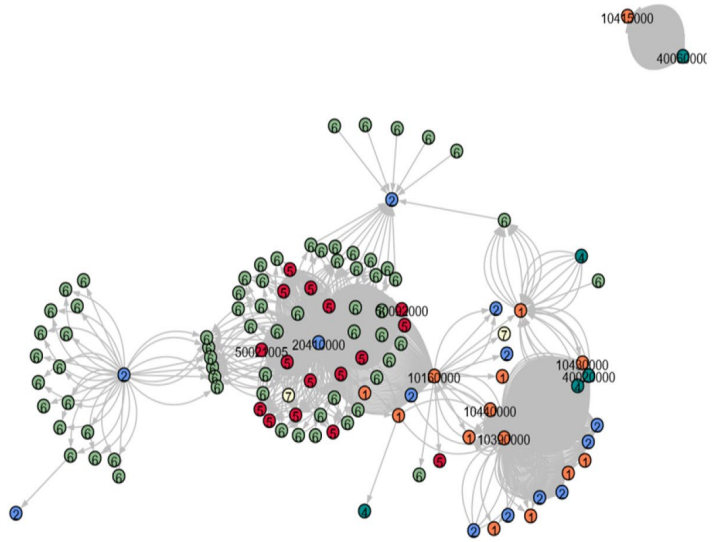


Figure 2: Bookkeeping Graph for a 10-day window of Data-set 1

Interpretation of the technology

- **Supervised vs unsupervised machine learning**

-- A simpler example:

- **Supervised LDA:**

a) human classify words into topics for a small sample

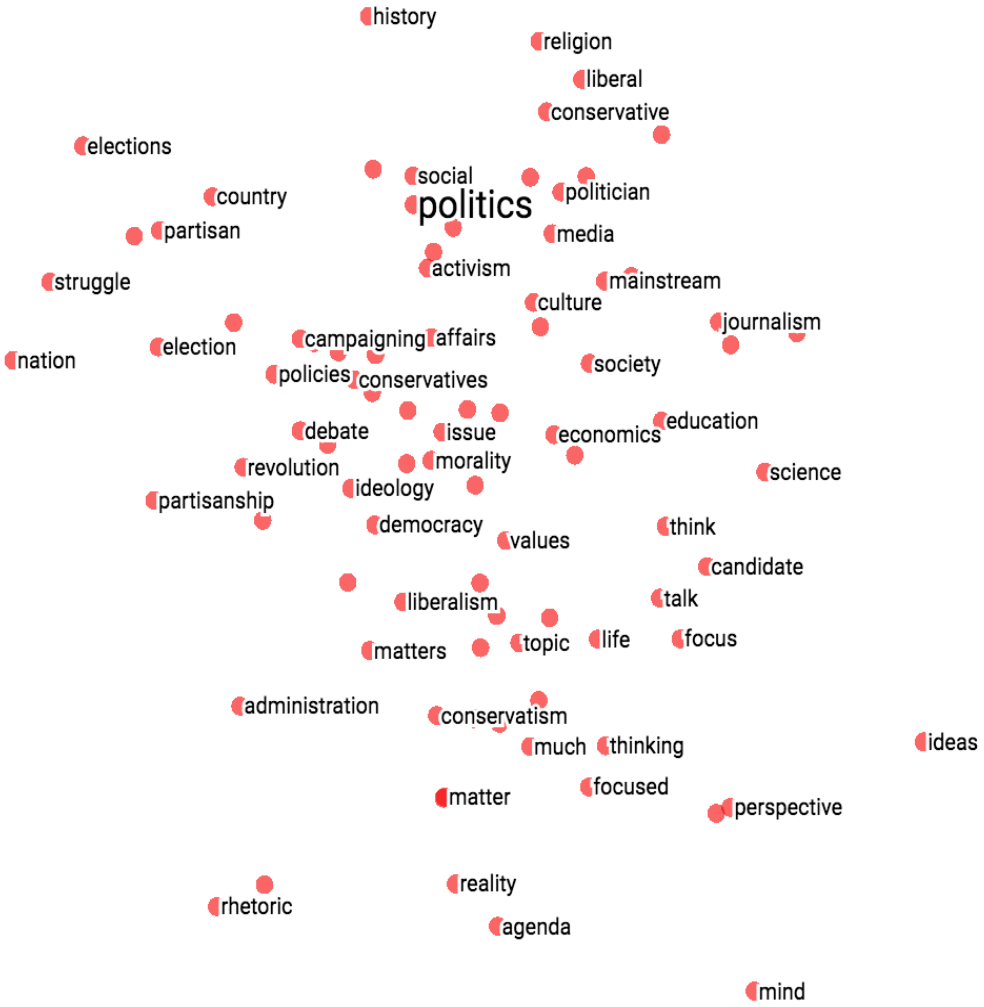
b) machine mimics human intelligence of classifying words into defined topics

c) given topics are defined, we can verify the accuracy of machine classification

- **Unsupervised LDA:**

a) machine classifies words into groups by minimizing distance

b) <u>given topics are not defined, we need to make sure machine-classified topics make sense</u>



Interpretation of the technology

- **Supervised vs unsupervised machine learning**

The paper did a very good job validating the classification

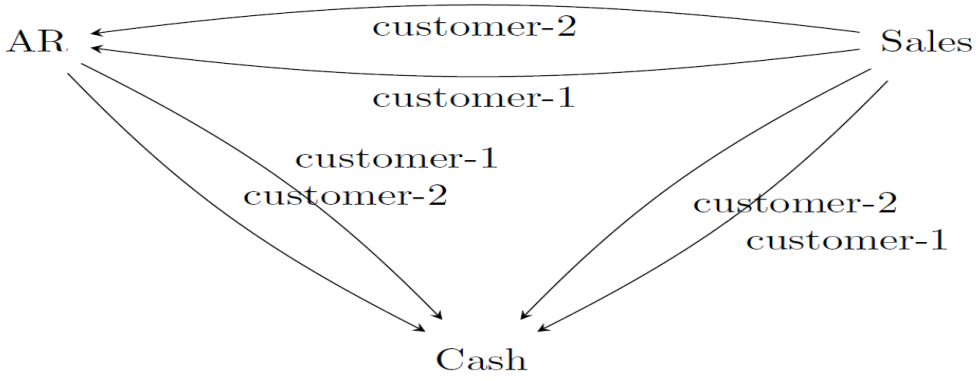
a. Quantitative: injected anomalies designed by experimenter
b. Quantitative: injected anomalies designed by practitioners
c. Qualitative: detected anomaly case studies

Discussion or comparison with supervised approach

a. Using the above injected anomalies as training sample and build a supervised ML model
b. Using cross-validation to report the accuracy of detecting out-of-sample anomalies – not available for unsupervised
c. Interpretation of feature importance, nonlinear interaction, model selections – advantage of supervised approach
d. At a higher level, why machine can detect some anomaly patterns that human cannot – more interpretation on what anomaly rules machine learned from the data that human are not aware of?

Interpretation of the technology

- Highlight why graph representation
  - Represent words with sematic vectors – Word2Vec/word embedding
    - NLP studies show the advantage of such vector representation
  - Consider showing the advantage of graph representation
    - Compare to a model directly built on transaction data without graph representation
    - Can machine (unlike human) do it without graphs?



Interpretation of the technology

- Highlight more on the customization of MDL for accounting application
  - medical vs finance

    Huge difference in model selection configuration, feature selection, performance
  - domain knowledge is critical
  - The paper did a great job in this regard: more highlights