

Deconstruct and Reconstruct: Using Topic Modeling on an Analytics Corpus

LAK14

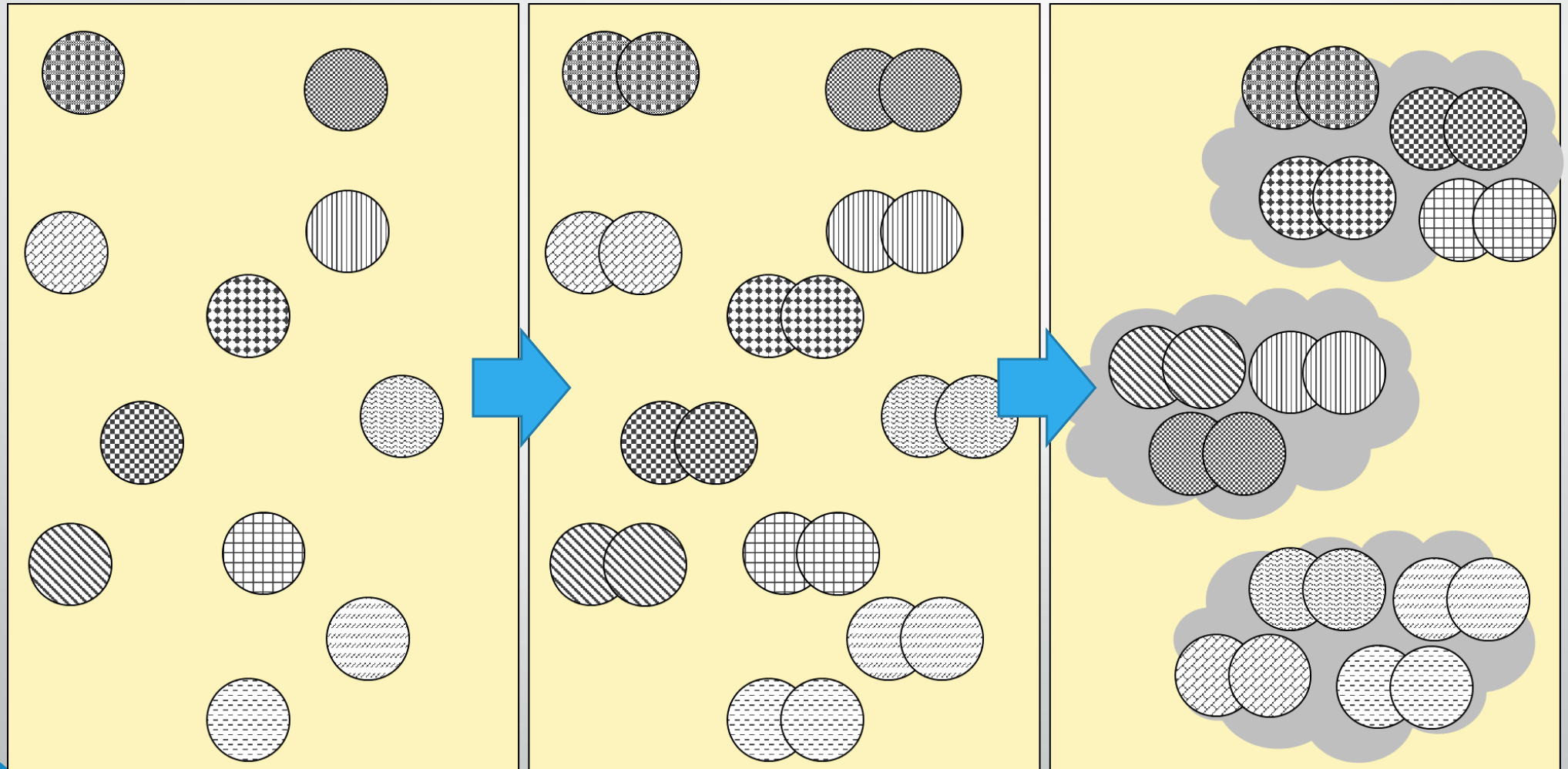
Mike Sharkey & Mohammed Ansari

March 25th, 2014

Major Themes

- Answer the challenge (research trends)
- Our approach (deconstruct/reconstruct)
- Systematize the process

Focus on Topics & Concepts



@mjshark
bluecanarydata.com

NLP Toolset

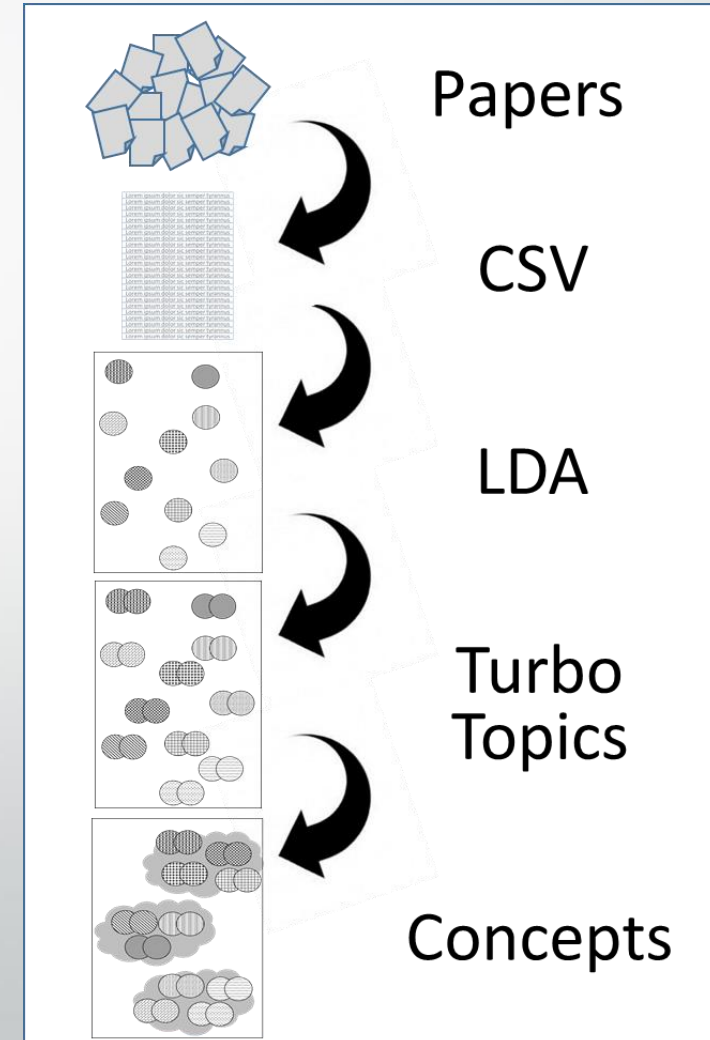
- LDA
- Turbo Topics
- Custom Java & Python



Acknowledgments to David Blei
<http://www.cs.princeton.edu/~blei/topicmodeling.html>

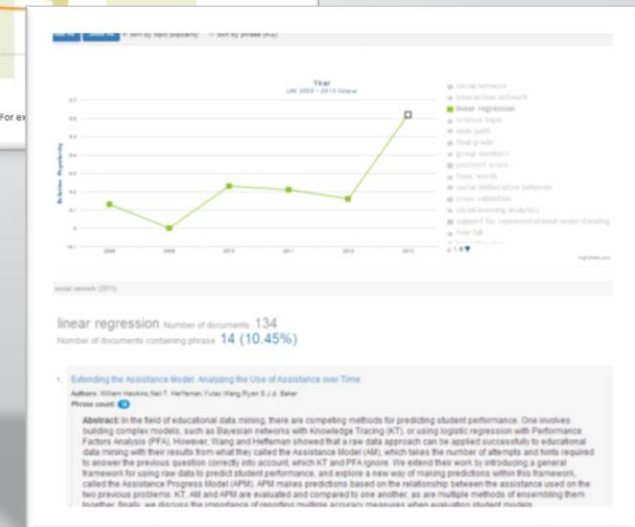
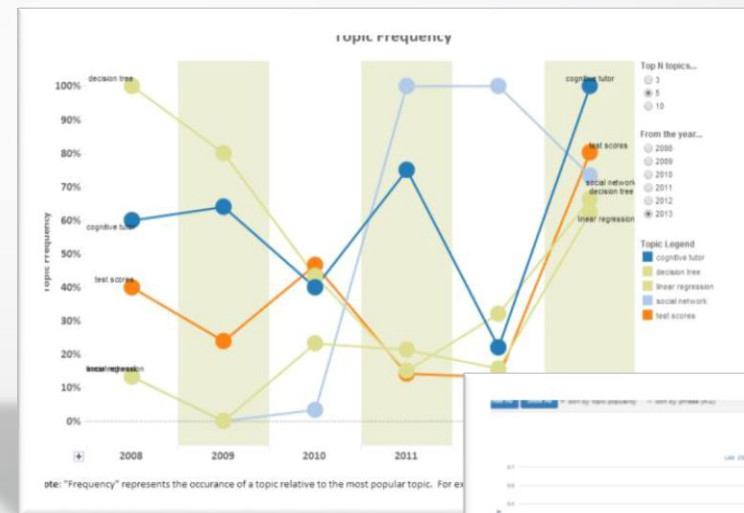
Overview of our Process

- 1) Use XML (courtesy of Taibi & Deitze, 2013)
- 2) CSV – one line per paper
- 3) Create stop word file ($50 < x < 200$)
- 4) Create vector of word counts
- 5) Run LDA using vector file
- 6) Use LDA output as input for Turbo Topics
- 7) Create n-gram result file
- 8) Visualize results



Data Presentation

- Context:
 - Who is viewing the data
 - What questions are they trying to answer
- Tableau reports
 - Interactive exploration of metrics
- Topic Browser & High Charts
 - Interactive drill down into corpus

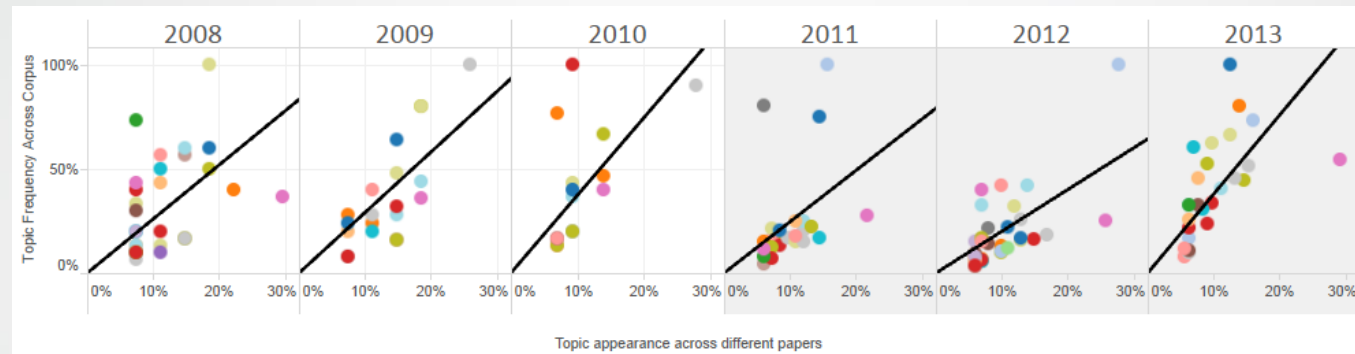


<http://lak14.bluecanarydata.com/>

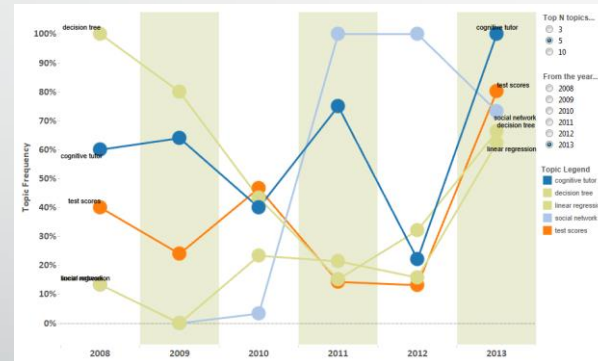
@mjshark
bluecanarydata.com

Findings

- Topic Convergence



- Concept Trends



Post Mortem

- 👉 Many areas for subjectivity (stop words, concept grouping). Open to biases
- 👉 Topic convergence chart is open to multiple interpretations

- 👍 Extremely strong team (coauthor: data janitor)
- 👍 Building on work from LAK13 Data Challenge
- 👍 Laid the foundation for automation/pipeline
- 👍 Applications in other areas
 - End of course survey analyses
 - White-label social networks

Thank You!

mike@bluecanarydata.com

www.bluecanarydata.com

@mjshark