

# **TED Talks** **for Data Scientists**

A 97-Day Seminar

Hosted by Brian Turner  
Spring 2017

A middle-aged man with short, light-colored hair and glasses is standing on a stage. He is wearing a dark suit jacket over a light blue button-down shirt. He has his hands raised in a gesturing motion, palms facing up. The background is a bright red wall with some blurred blue and white shapes. The text "Why are you here?" is overlaid in white, bold, sans-serif font across the center of the image.

**Why are you here?**

# ...To Be Original

Top Professions
Writers
Designers
Journalists
Artists
Architects

Not enough Data Scientists giving **TED** Talks

Day 1: Topic Overload  
i.e. what **NOT** to talk about  
if you want to **be original**

# Analyzing overly used topics

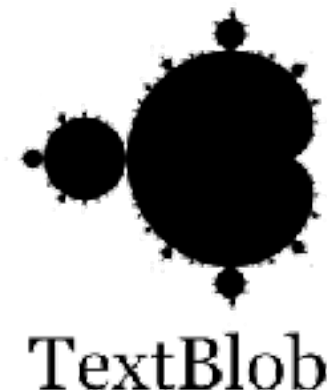
2300 **TED** transcripts scraped using BeautifulSoup

Raw corpus of 5 million words

Topic modeling with sci-kit learn

Visualizations with D3

Sentiment analysis with TextBlob



# Topic Modeling Process

Preprocess and tokenize words with NLTK



TF-IDF Vectorizer - 5500 token features

Uni-grams and max word document-frequency: 65%



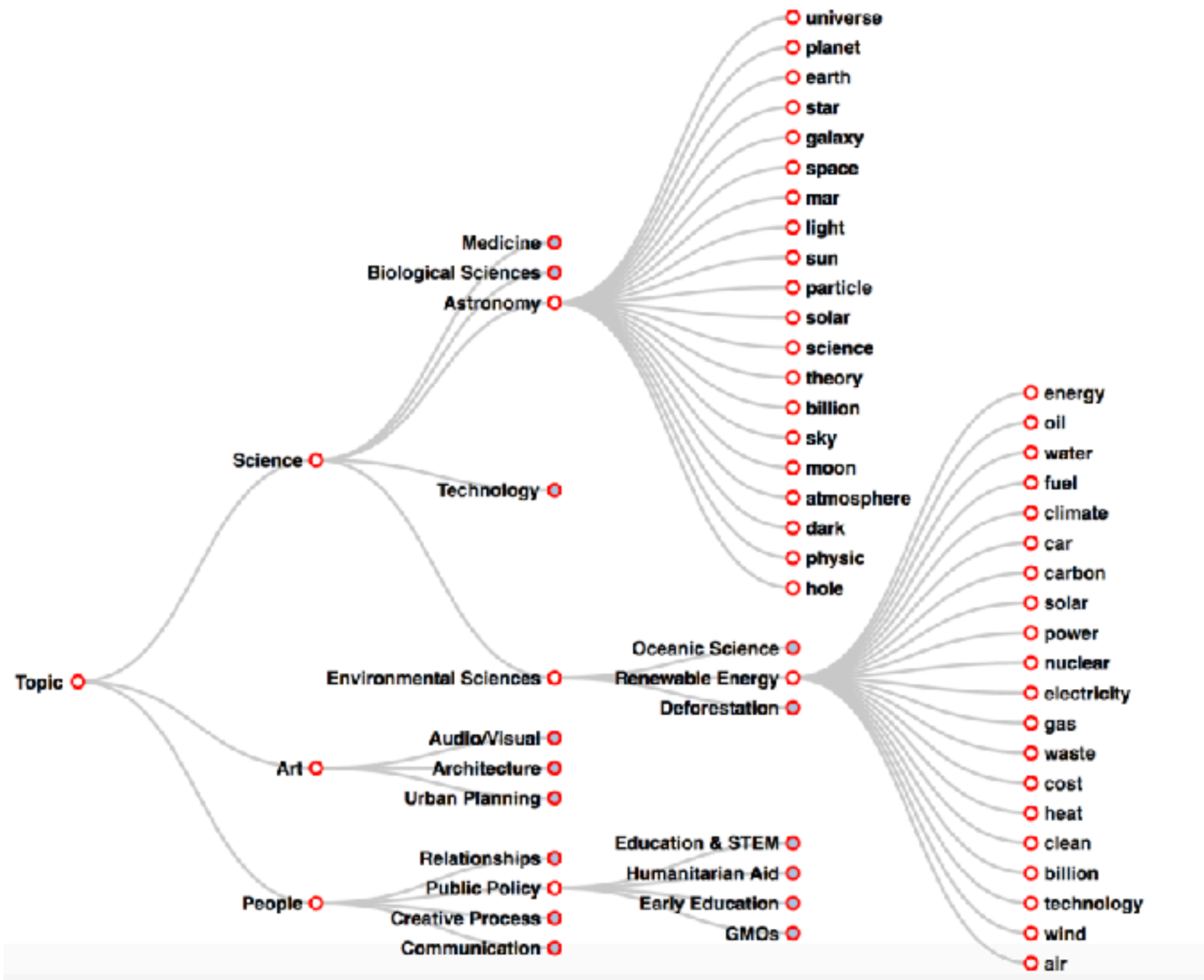
LDA vs. Non-Negative Matrix Factorization (NMF)



NMF Output: 25 topics to avoid

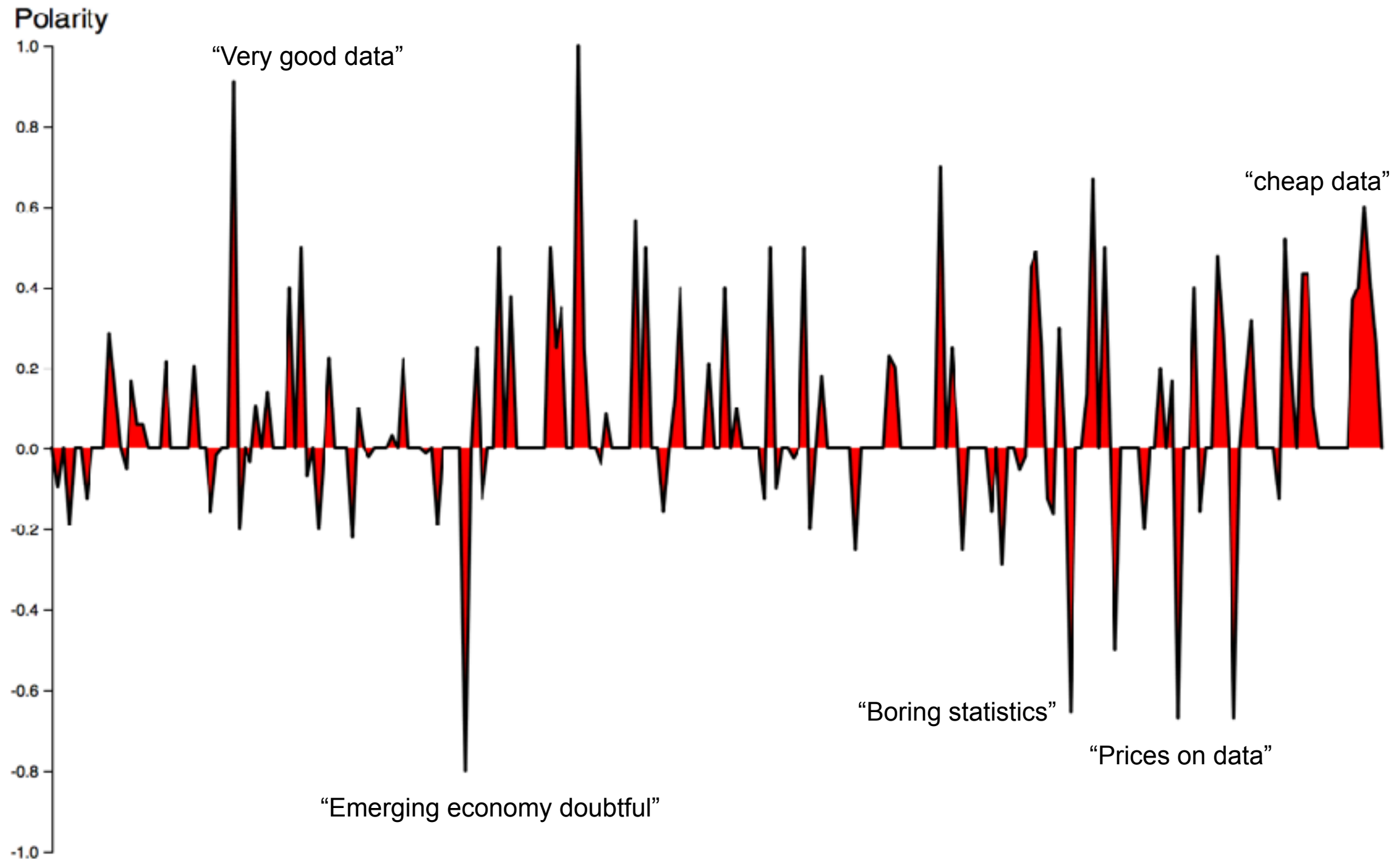
# 25 topics worth stopping

<http://bl.ocks.org/anonymous/21033ce472e9f6699acd7597c47325ec>



# Hans Rosling's TED Talk

Sentiments on the need for better data





# Tomorrow...Day 2:

# Words to avoid in your **TED** Talk



**Thanks and Have Fun!**