

# The U.S. Presidential Race

## Comparing Coverage Across Three Major Sources

Thomas Brawner

Galvanize, Inc.

September 2015

# Two objectives

- *Is there a difference in coverage?*

→ Classification Problem

- *What is the difference?*

→ Topic Modeling

→ Sentiment Analysis

# Data collection

- # Scrape *Guardian*, *New York Times*, *Wall Street Journal*
- # January to August 2015
- # Raw data → MongoDB
- # Filter on candidates
- # TF-IDF, 2000 terms

# Multiclass classification accuracy

	Global	Guardian	New York Times	Wall Street Journal
Naive Bayes	0.703	0.414	0.965	0.340
Random Forest	0.905	0.783	0.998	0.809
Gradient Boosting	0.921	0.856	0.983	0.833
<i>N</i>	3757	1018	2087	652

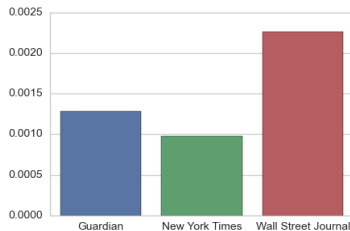
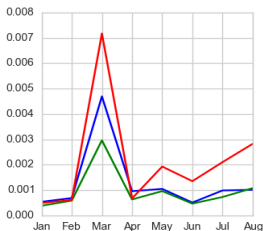
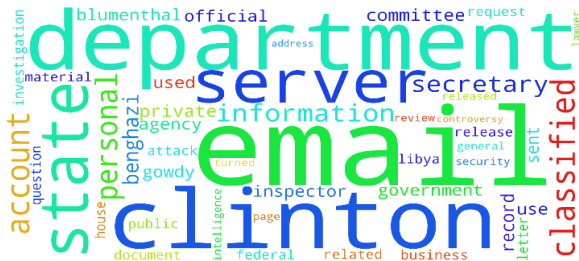
# Topic modeling strategy

$$\begin{bmatrix} v_{11} & \cdots & v_{1t} & \cdots & v_{1T} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{i1} & \cdots & v_{it} & \cdots & v_{iT} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{Nt} & \cdots & v_{NT} \end{bmatrix} =$$
$$\begin{bmatrix} w_{11} & \cdots & w_{1K} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{iK} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{NK} \end{bmatrix} \times \begin{bmatrix} h_{11} & \cdots & h_{1t} & \cdots & h_{1T} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ h_{K1} & \cdots & h_{Kt} & \cdots & h_{KT} \end{bmatrix}$$

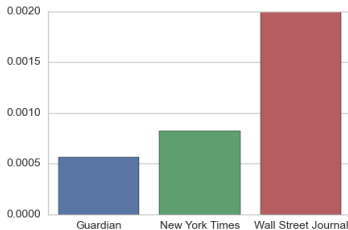
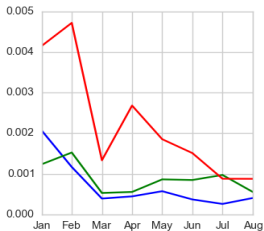
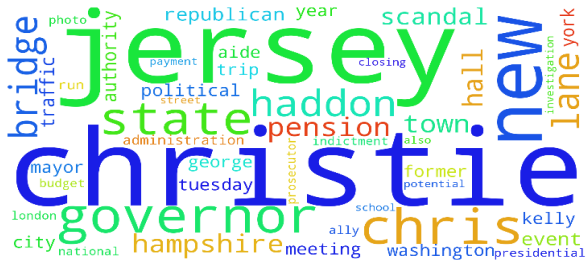
# Topic modeling strategy (cont.)

- With  $\mathbf{H}_{K \times T}$ 
  - sort descending for each  $k$
  - top 50 terms  $\rightarrow$  Word Cloud
- With  $\mathbf{W}_{N \times K}$ 
  - group by source, month  $\rightarrow$  time series
  - group by source  $\rightarrow$  bar plots

# Hillary's Email

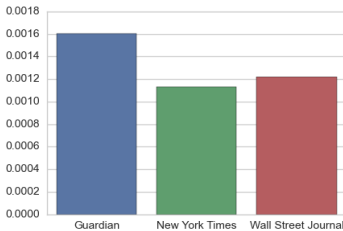
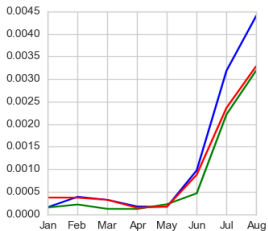
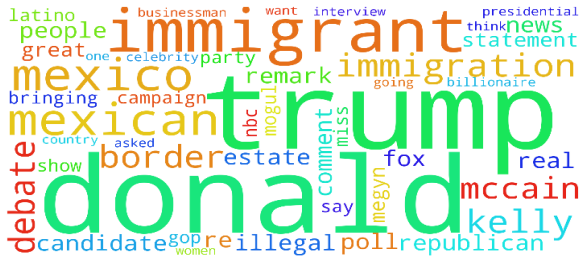


# Chris Christie

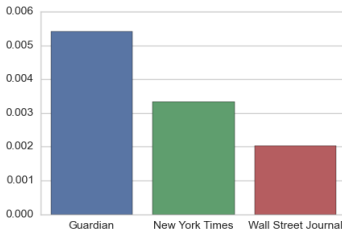
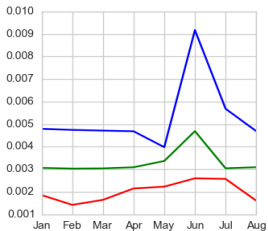
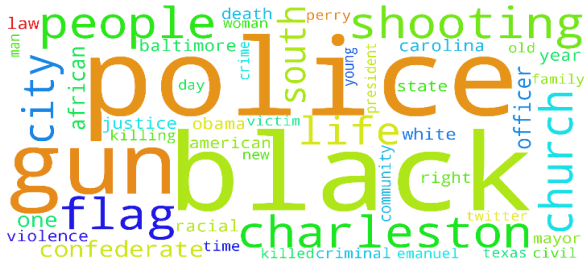




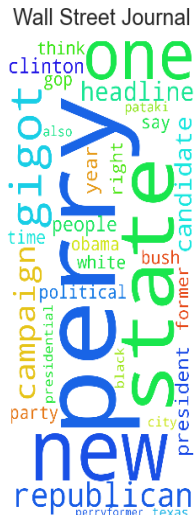
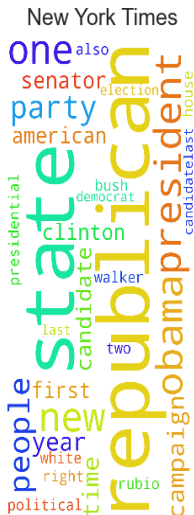
# The Rise of Trump



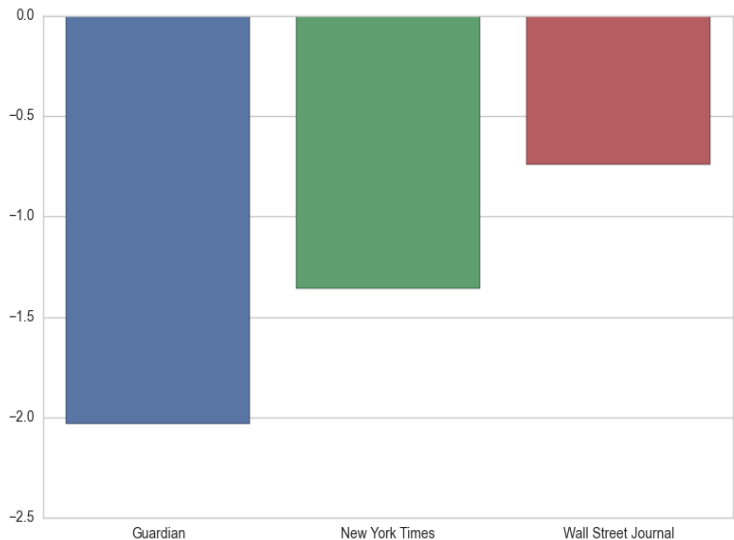
# Gun Violence



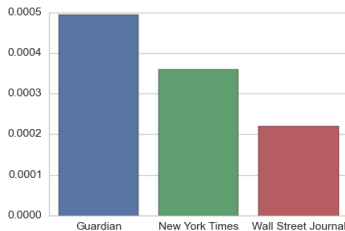
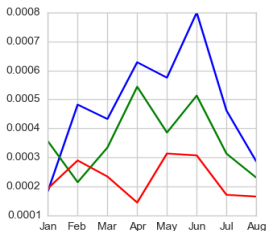
# Gun Violence by Outlet



# Gun Violence by Outlet



# Gay Marriage



# Gay Marriage by Outlet

Guardian

sex supreme gay republican candidate also state presidential cruz religious decision clerk issue former president couple one governor sex former law ruling people legal obama texas huckabee right related

marriage court

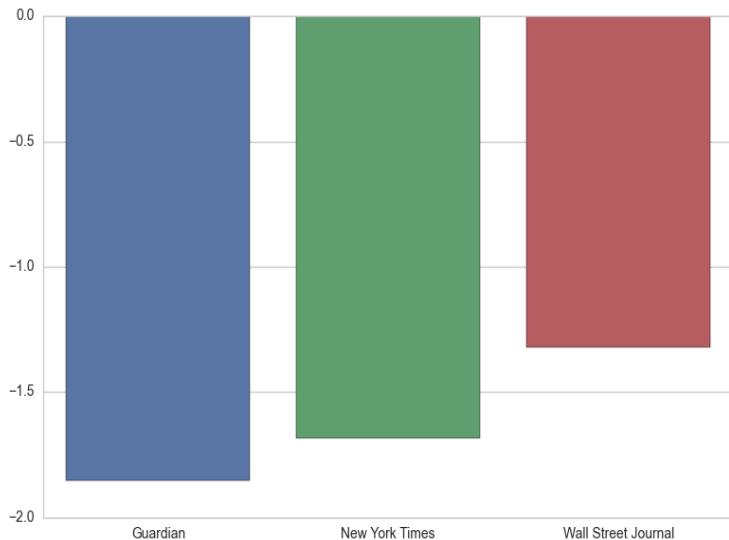
New York Times

A word cloud of terms related to the 2004 US Presidential election. The most prominent words are 'marriage' and 'republican' in large blue font. Other significant words include 'gay', 'court', 'supreme', 'decision', 'cruz', 'senator', 'religious', 'issue', 'conservative', 'candidate', 'support', 'amendment', 'justice', 'state', 'right', 'iowa', 'people', 'bush', 'party', 'law', 'reiser', 'one', and 'courtship'. The words are arranged in a dense, overlapping manner with various colors and orientations.

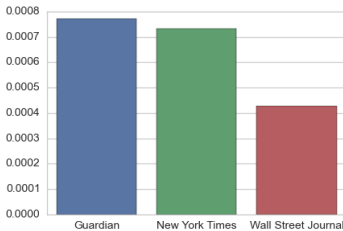
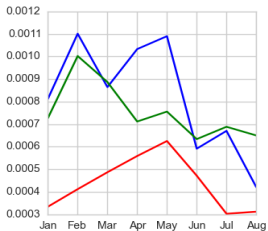
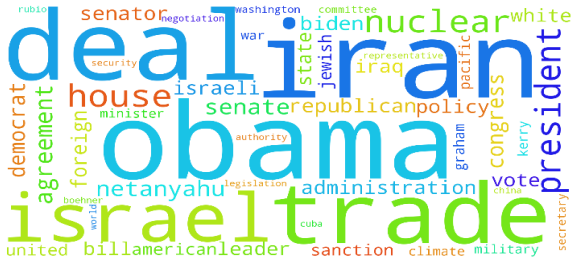
Wall Street Journal

A word cloud of terms related to the Supreme Court case on gay marriage. The words are arranged in a circular pattern around a central 'Court'. The words include: ruling, conservative, born, cruz, care, seeking, supreme, issue, former, florida, de, bush, moines, repeal, marriage, support, public, primary, iowa, friday, decision, presidential, gay, and republican health candidate.

# Gay Marriage Sentiment by Outlet

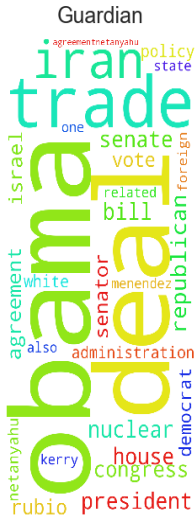


# Iran Nuclear Deal

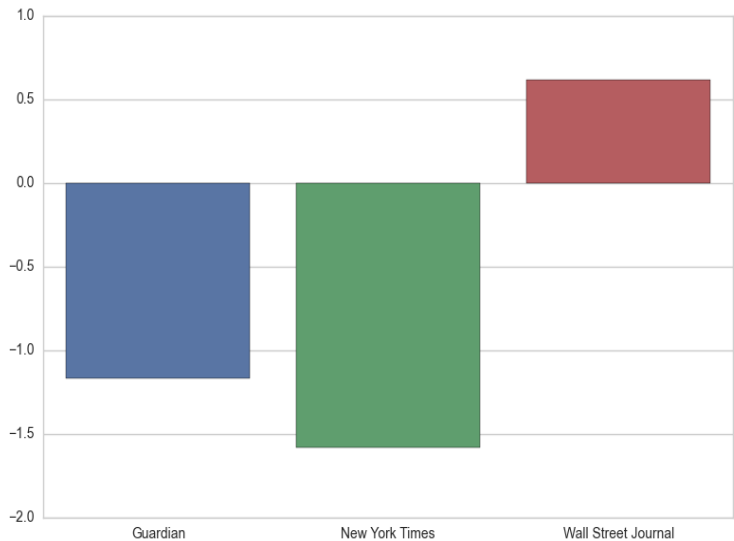




# Iran Nuclear Deal by Outlet



# Iran Nuclear Deal by Outlet



# Next steps

- # More data: collect through duration of election
- # More data: *Washington Post*, *Financial Times*, etc.
- # Finer classification model tuning
- # What are we misclassifying?