Discovering Word Associations in News Media via Feature Selection and Sparse Classification

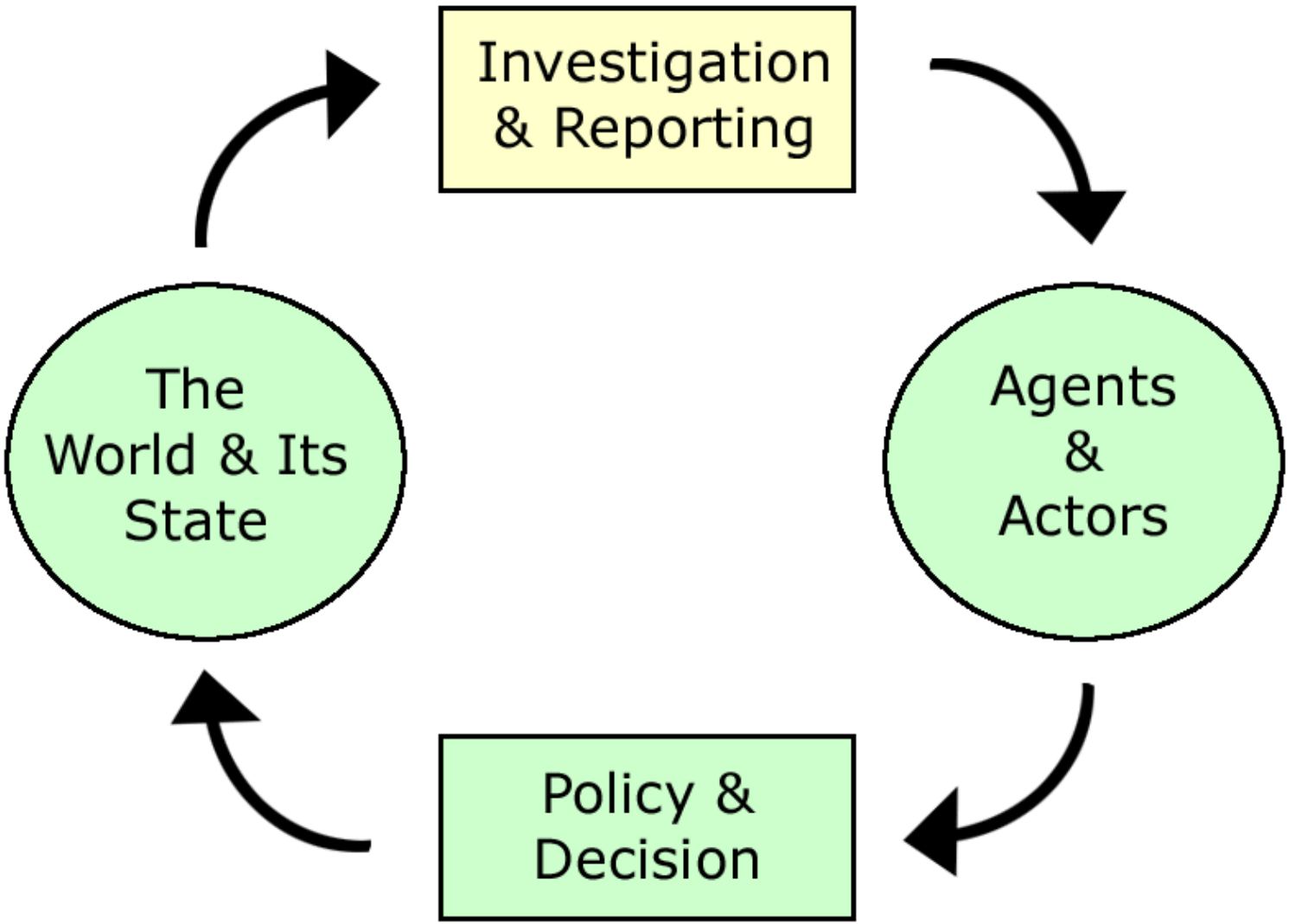
Brian Gawalt

UC Berkeley
Departments of
EECS & Statistics

Based on joint work with:

Jinzhu Jia (Stat)
Luke Miratrix (Stat)
Laurent El Ghaoui (EECS)
Bin Yu (Stat/EECS)
Sophie Clavier (SFSU, Intl. Relations)

A Tidy Model of How the World Works



Standards and Practices

... versus...

- To be persuasive we must be believable; to be believable we must be credible; to be credible we must be truthful.
  - *Edward R. Murrow*

- You supply the photographs, and I'll supply the war.
  - *William Randoph Hearst*

Improve News Media, Improve How the World Works

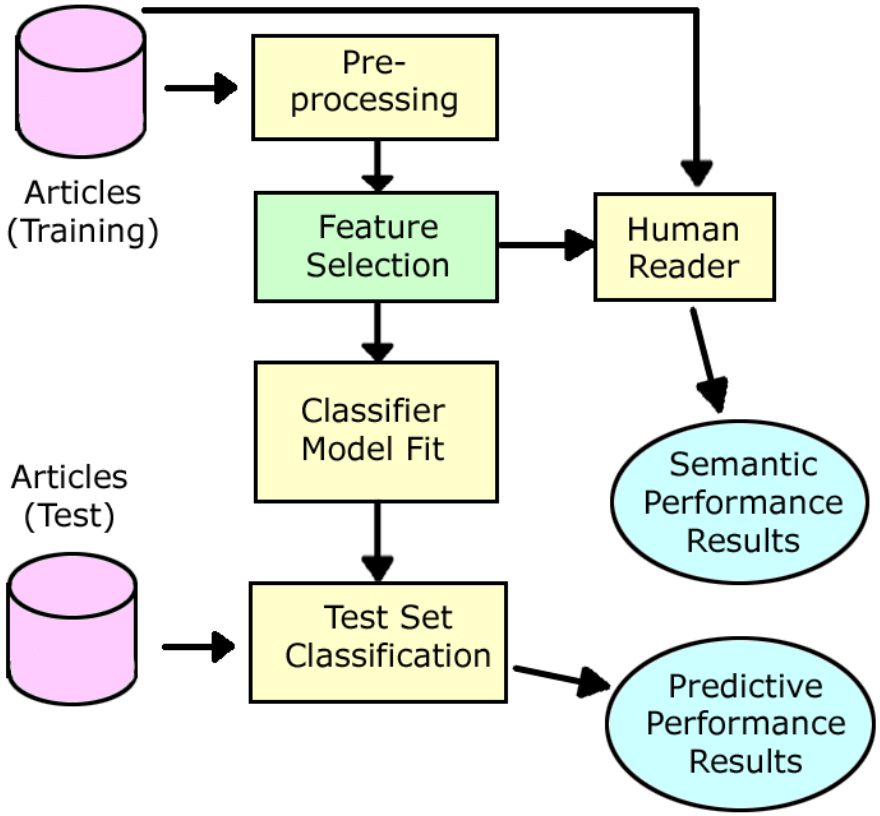
Improve News Media Analysis, Improve News Media, Improve How the World Works

- Holes in current approach
  - Time and labor constraints
  - Case study approach too prone to bias
- Statistical machine learning techniques
  - Fast, scale well
  - Reproducible results
  - Designed around predictive tasks
- Harness machine learning to power media studies
  - New predictive framework needed for media study
  - New design guidelines and metrics needed for machine learning

Our application: word image in the New York Times

- Word Image: a small set of words describing/distinguishing a topic
- As a predictive problem:
  - Predict appearance of a query word q in a document from the document's use of other words
- Predictive model must be interpretable
  - Predictor weights must directly and simply drive label
  - No. of predictors used must be few: sparse model **approximation**
  - The faster predictors can be computed, the better
- Chosen predictor words form a set known as the Word Image for q
- Word image must be evaluated two ways:
  - Can labels (appearance indicator for q) be effectively predicted?
  - Are the chosen words meaningful w.r.t. q?

Our approach: feature selection techniques from text classification



- Independent variable: Feature selection process
- Dependent variables: Semantic/predictive performance
- Experiment is conducted repeatedly across 47 queries in order to broadly test the effects of the choice of feature selection process.

Feature Selection Methods

Positive doc. set $I^+ = \{i \mid y_i = 1\}$; negative doc. set $I^- = \{i \mid y_i = -1\}$

- Co-occurrence (COOC):

$$c_j^+ = \sum_{i \in I^+} x_{ij}$$

  - Take 15 words appearing most often in positive documents (highest c_j^+ scores)

- Delta TF-IDF (DTF): [Martineau09]

$$d_j^\pm = \sum_{i \in I^\pm} \mathbb{I}(x_{ij} > 0)$$

$$\delta_j = c_j^+ \log \left(\frac{m^+}{d_j^+} \frac{d_j^-}{m^-} \right)$$

  - Appearances of rarer words now count more when finding top scorers

Feature Selection Methods

- Bi-normal Separation (BNS): [Forman03]

$$b_j = \Phi^{-1}\left(\frac{d_j^+}{m^+}\right) - \Phi^{-1}\left(\frac{d_j^-}{m^-}\right)$$

  - $\Phi(.)$ the inverse standard normal CDF
  - Selects words with strong divergence of between-class appearance-rate

- χ^2 log-likelihood (CHI):

$$\begin{aligned} f_j = \quad & d_j^+ \log\left(\frac{d_j^+}{m^+}\right) + [m^+ - d_j^+] \log\left(1 - \frac{d_j^+}{m^+}\right) + \\ & d_j^- \log\left(\frac{d_j^-}{m^-}\right) + [m^- - d_j^-] \log\left(1 - \frac{d_j^-}{m^-}\right) - \\ & [d_j^+ + d_j^-] \log\left(\frac{d_j^+ + d_j^-}{m}\right) - [m - d_j^+ - d_j^-] \log\left(1 - \frac{d_j^+ + d_j^-}{m}\right) \end{aligned}$$

  - Select words by ranked p-value for hypothesis "Word j appears at a different rate between the two classes"

Feature Selection: l_1 Regularized Logistic Regression (L1LR)

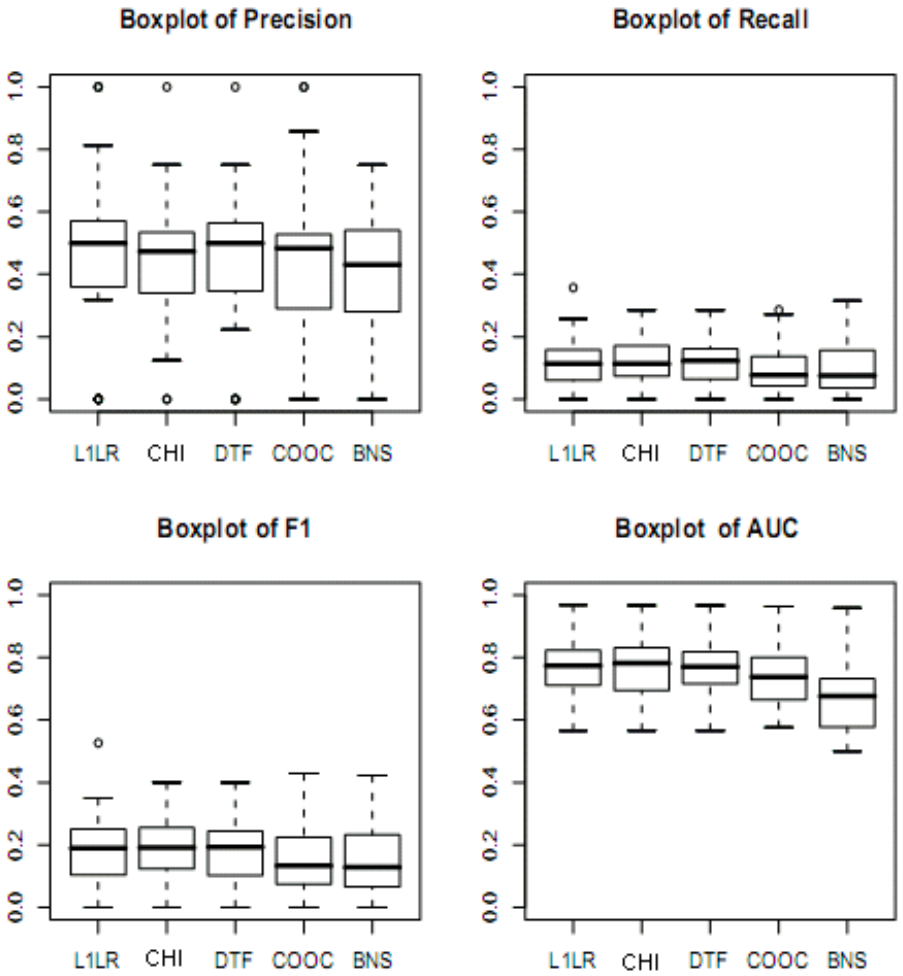
$$\mathcal{L}_{L1LR}(\beta) = \underbrace{-\sum_{i=1}^{m} \log\left(1 + \exp(-y_i(\beta_0 + x_i^T\beta))\right)}_{\text{classifier loss function}} + \underbrace{\lambda \sum_{j=1}^{n} |\beta_j|}_{\text{weight penalty}} \quad (1)$$

- L1LR loss function encourages fitting to the data, discourages non-zero values of β
- As $\lambda \to \infty$, $\beta_j \to 0 \;\; \forall j = 1, \ldots, n$
- By binary search, isolate value of λ which leaves ~ 15 nonzero predictors
- Greater computational complexity than previous four methods, but still solved efficiently

Selected features: $q =$ "CHINA"

COOC	DTF	BNS	CHI	L1LR
year	killing	[not] recur	[not] recurring	korea
chinas	institutions	[not] recurring	[not] purified	united
north	view	[not] stalins	[not] nazis	north
beijing	larger	[not] kenya	[not] marches	global
government	history	[not] marches	[not] holocaust	countries
states	outside	[not] eradicate	[not] perpetrators	russia
mr	place	[not] victims	[not] eradicate	states
united	death	[not] goldhagen	[not] kenya	chinas
chinese	russia	[not] holocaust	[not] stalins	beijing
said	world	[not] killing	[not] goldhagen	chinese

Predictive Performance Results



- L1LR, CHI, and DTF do not have significant differences from each other
- L1LR, CHI, and DTF all perform significantly better than both COOC and BNS

Human Reader Survey

Please read the following paragraphs:

Paragraph 1:

After nearly two decades of independence, Moldova's citizens are still at odds over the basic question of who they are. That division boiled over last week, when a huge anti-Communist demonstration turned violent. Its participants, in their teens and 20s, say they are desperate to escape a Soviet time warp and enter Europe. But many of their elders feel more affinity with Russia, and see the protests as a plot by their western neighbor Romania to snatch away Moldova's sovereignty.

Paragraph 2:

Mr. Lukyanov pointed out that the United States and Russia approach Iran from sharply different perspectives. Russia and Iran are neighbors, and the Kremlin has for many years had positive dealings with Iran on regional issues, including unrest in Chechnya and in Central Asia.

Paragraph 3:

Last week the government's point man on the economic crisis, the deputy prime minister Igor I. Shuvalov, seemed to underline that policy. He told an economic forum in Moscow that the government would withhold support from industry and cut the budget, allowing Russia to husband reserves to support the ruble.

Q8) Which of the following word lists is the most useful summary of the above paragraphs as described in the instruction sheet? (You may select two, one, or none as desired.)

List A	List B	List C	List D
NOT pakistans	baghdad	NOT enriched	georgia
NOT boldest	iraqi	NOT officials	moscow
NOT unfolded	war	NOT slated	ukraine
NOT consult	afghanistan	NOT stockpile	russian
NOT islamabad	american	NOT lightly	putin
NOT offensive	troops	NOT vienna	russias
NOT oversees	bush	NOT accord	europe
NOT arrived	oil	NOT reactor	china
NOT capital	military	NOT geneva	united
NOT head	invasion	NOT research	gas

Continue to Q8B)...

The paragraphs on the previous page are best described as focusing on the topic(s) of _____

If at least two of the three paragraphs focus on a topic, then consider them to be focusing on the topic overall.

(a) russia
(b) iraq
(c) both of the above
(d) neither of the above

Continue to Q9A)...

(Few questions were misidentified in part B)

Processing Survey Results

Example counts of survey respondent selections...

	Paragraph Queries	Decoy Queries
Scheme [One]	10	1
Scheme [Two]	4	0

	Paragraph Queries	Decoy Queries
Scheme [Three]	4	3
Scheme [Four]	5	3

- Toss out any misidentified paragraphs
- Two ways that, e.g., L1LR can demonstrate superior quality over COOC:
  - When head-to-head, L1LR is picked more frequently
  - L1LR is picked ahead of DTF, BNS, or CHI at a greater rate than COOC is picked ahead of DTF, BNS, or CHI
- Combine p-values from both these hypotheses across all 10 matchups

Human Survey Results

Scheme a	Scheme b	p-value
L1LR	COOC	0.151
L1LR	DTF	**0.002**
L1LR	CHI	**0.000**
L1LR	BNS	**0.000**
COOC	DTF	0.327
COOC	CHI	**0.000**
COOC	BNS	**0.000**
DTF	CHI	**0.003**
DTF	BNS	**0.001**
CHI	BNS	0.297

- L1LR significantly bests all but COOC
- COOC not significantly preferred over cousin DTF
- CHI and BNS roundly rejected, except between each other

Conclusions

- L1LR success indicates effectiveness of sophistication in ML approaches
- Traditional ML practices wouldn't yield these images – new design criteria were applied
- Scale and complexity can be easily accomodated
- Posing news media analysis problems in a predictive framework in a way that takes advantage of these and future tools should be encouraged