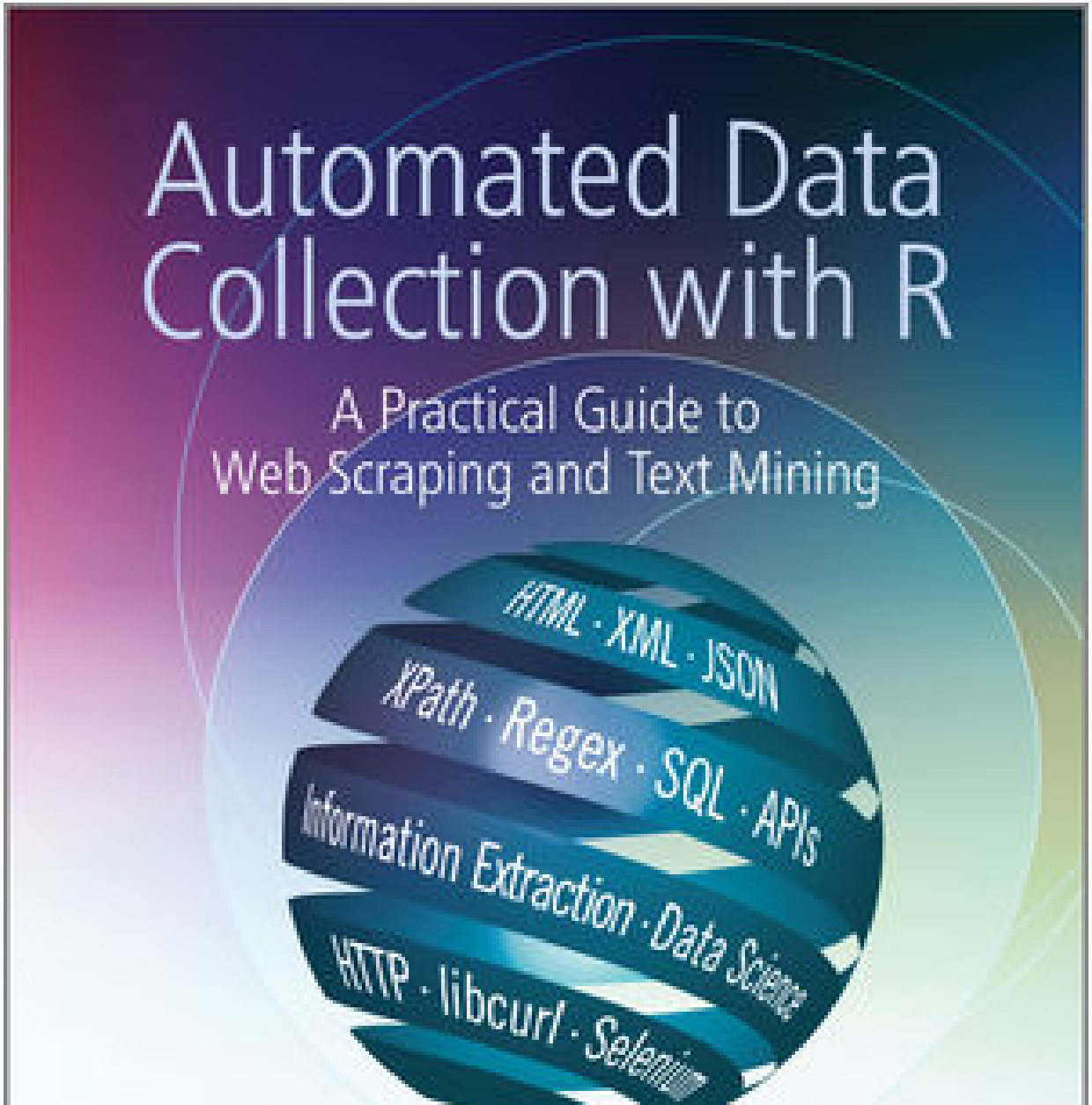
Computer-assisted content analysis

Will Lowe Princeton University
James Lo University of Southern California

May 17, 2016

IQMR 2016 Syracuse

Automated Data Collection: Web Scraping with R



Menu

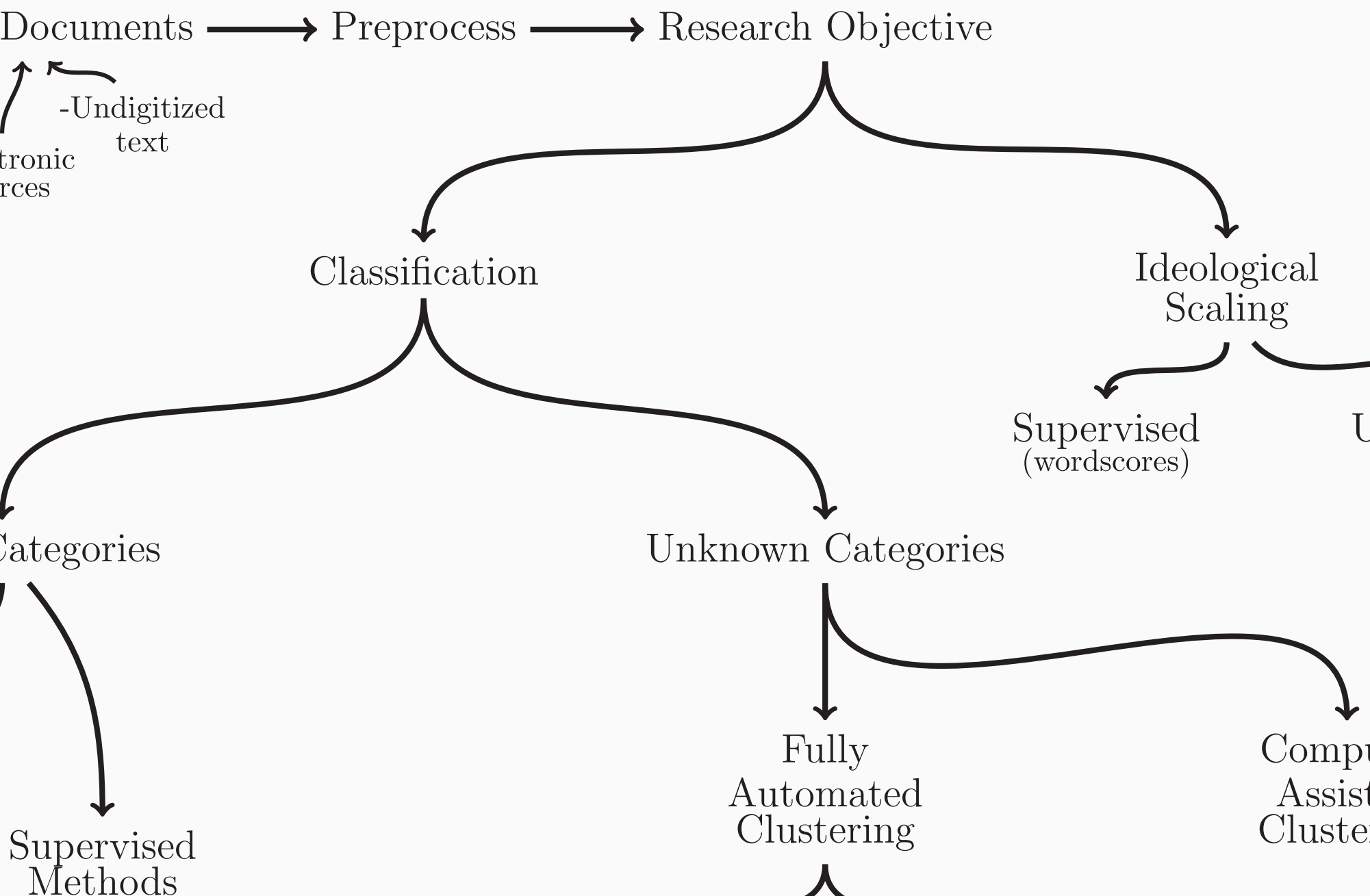
Session 1: Classical Content Analysis

Session 2:

- Classification for Known Categories: Supervised Methods
- Evaluation
- Content Analysis Research Design

Session 3: Topic and Scaling Models

Text as Data


Documents
Preprocess
Research Objective
-Undigitized text
tronic rces
Classification
Ideological Scaling
Supervised (wordscores)
ategories
Unknown Categories
Supervised Methods
Fully Automated Clustering
Compu Assist Cluste

Classification Approaches when Categories are Known

Examples:

Are campaign advertisements positive or negative?

What policy areas do newspaper editorials cover?

Are international statements belligerent or peaceful?

Do court letters represent liberal or conservative positions?

What language is this article written in?

Is this email spam?

Classification Approaches when Categories are Known

Earlier today, we talked how to do this using a dictionary approach.

An alternative are supervised machine learning methods.

The idea:

1. coders categorize a set of documents by hand
2. the algorithm "learns" how to sort the documents in categories
3. characteristics of training set are used to assign new documents to categories.

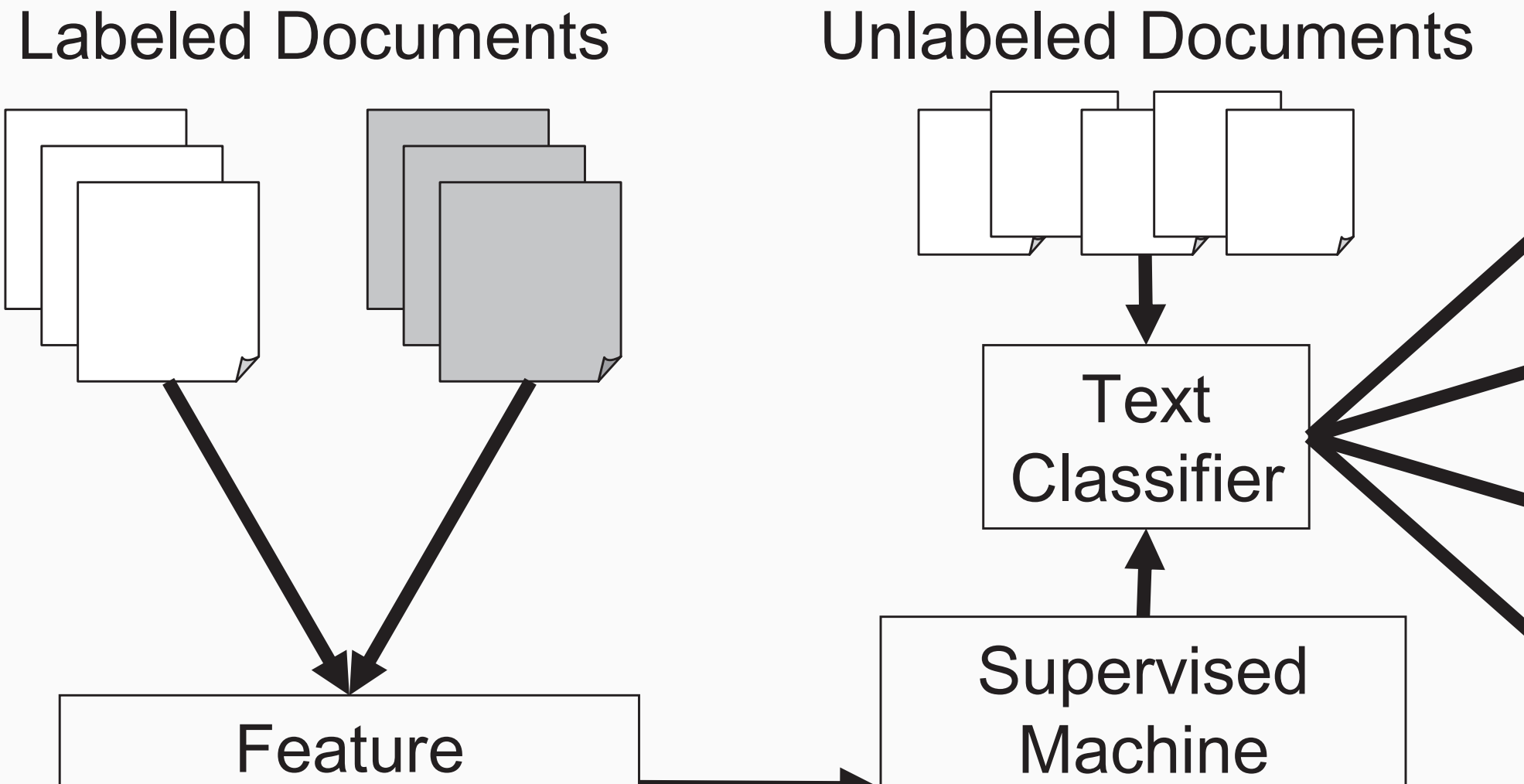
Classification Approaches when Categories are Known

Assume that each document has a *single* topic Z

Let θ_k be the *probability* that Z=k for each document

Assume that (some) topic labels are observed

The machine learning approach to text cla



Output

Desirable classification outcome:

	$P(Z = \text{'Domestic Policy'} \mid \text{D})$	$P(Z = \text{'Foreign Policy'} \mid \text{D})$
D_1	0.75	0.25
D_2	0.82	0.18
...	...	...
D_{N-1}	0.02	0.98
D_N	0.45	0.55

The Basic Steps

1. Construct a training set
   (a) create a coding scheme
   (b) select documents (ideally randomly sampled)
2. Apply the supervised learning method to learn features of a training set and infer labels
3. Validate and classify remaining documents

Caveats

Much of machine learning, computational linguistics, and AI deals with classification problems

We only touch on the issues here…

Ensembles: combine different individual classifiers to produce one that is superior to any of the individual ones.

Example of Naive Bayes Classifier: Amicus Curiae Briefs

Background:

- Amicus Curiae (friend of the court) briefs are submitted to an appellate court
- They usually present a legal argument for or against one of the parties to a case
- Amicus Curiae can be submitted by any group that feels that it has a stake in the case

Affirmative action

Evans et al. use cases about the constitutionality of ‘affirmative action’ programs at university level.

The arguments are as much *political* (state vs federal rights, social welfare, constitutional interpretation) as they are *legal*

Affirmative Action

This work uses document classification to answer two questions

- To what extent can the conservative/liberal *direction* of an AC brief be predicted on the basis of its words?
- What can we learn about *language* of each side of the case?

Naive Bayes

Naive Bayes is a relatively old ($\sim$1975) classification method

Suppose one had to guess whether document j is liberal (Z=L) or conservative (Z=C) based on its word profile W_j.

Probability can be derived by applying Bayes theorem:

$$P(Z = L \mid W_j) = \frac{P(W_j \mid Z = L)\, P(Z = L)}{P(W_j)}$$

Naive Bayes

$$P(Z = L \mid W_j) \propto P(W_j \mid Z = L)\, P(Z = L)$$

We can drop $P(W_j)$ since it is constant across categories.

Given a representative training set, estimating $P(Z = L)$ is easy:

$$\hat{P}(Z = L) = \frac{\text{\# training docs that are liberal}}{\text{\# of training docs}}$$

Naive Bayes

Estimating the probability that a word profile W_j occurs given that the document is liberal $P(W_j \mid Z = L)$ is more challenging, because any one word profile is likely to occur only once.

Solution: words are assumed to be generated *independently* given the category Z (the 'naive' and wrong assumption).

$$
\begin{aligned}
P(W_j \mid Z = L) &= \prod_i P(W_i \mid Z = L) \\
&= \prod_i P(W_i \mid \theta_L) \qquad \text{realised as Multinomial or Binomia}
\end{aligned}
$$

Naive Bayes

With this assumption, we can estimate the probability of observing a word *i* given that the document is liberal: proportion of word *i* in liberal training set.

The classifier then chooses the class Z (Liberal or Conservative) with the highest aggregate probability.

Note that every new word adds a bit of information that re-adjusts the conditional probabilities.

Naive Bayes

Note that with two classes (here: liberal and conservative) this has a rather neat interpretation:

$$\frac{P(Z = \text{'Liberal'} \mid W_j)}{P(Z = \text{'Conservative'} \mid W_j)} =$$

$$\prod_i \frac{P(W_i \mid Z = \textit{'Liberal'})}{P(W_i \mid Z = \textit{'Conservative'})} \times \frac{P(Z = \textit{'Liberal'})}{P(Z = \textit{'Conservative'})}$$

Logging this probability ratio, every new word *adds* a bit of information that pushes the ratio above or below 0

Evaluating Classifiers: Accuracy, Precision and Recall

How can we evaluate how good a supervised classifier works?

Use the confusion matrix (training set versus machine)

Evaluating Classifiers: Accuracy, Precision and Recall

		Machine	
		Liberal	Conservative
Training Data	Liberal	40	10
	Conservative	40	60

Accuracy = (40+60)/150 = .66

Precision ($Z_{machine}$=Liberal) = 40/(40+40) = 0.5

Precision ($Z_{machine}$=Cons) = 60/(10+60) = 0.86

Recall ($Z_{training}$=Liberal) = 40/(40+10) = 0.80

Recall ($Z_{training}$=Cons) = 60/(40+60) = 0.60

Evaluation

Use evaluation to improve classifier. Things you can do:

- refine coding scheme for human coding
- ensure representative training set
- use ensemble of classifiers

Distinctive Words

Detecting different rhetorical styles of liberal and conservative groups:

"Liberal groups use language emphasizing the impact of affirmative action polices, while conservative words indicate concern over legal-constitutional limits on administrative procedure" (Evans et al. 2007, p. 1029)

	Avg. Freq. per Lib. Brief	*Avg. Freq per Cons. Brief*	*Chi*^{2}	*Interpretive Code Example*
tive Words				
*	2.83	41.79	39.18	Proceduralist; Race/Gender N Justice
	0.07	1.17	36.14	Intent vs. Consequences; Cons
l*	14.86	25.04	24.13	Proceduralist; Race/Gender N

Compare and Contrast

Bara et al. (2007) abortion debate with a thematic dictionary

Vocabulary*	Liberals-restrictionists (1966)
Advocacy	-2.70
Legal	2.22
Medical	-0.34
Moral	-1.02
Rhetoric of debate	0.42
Social	1.48

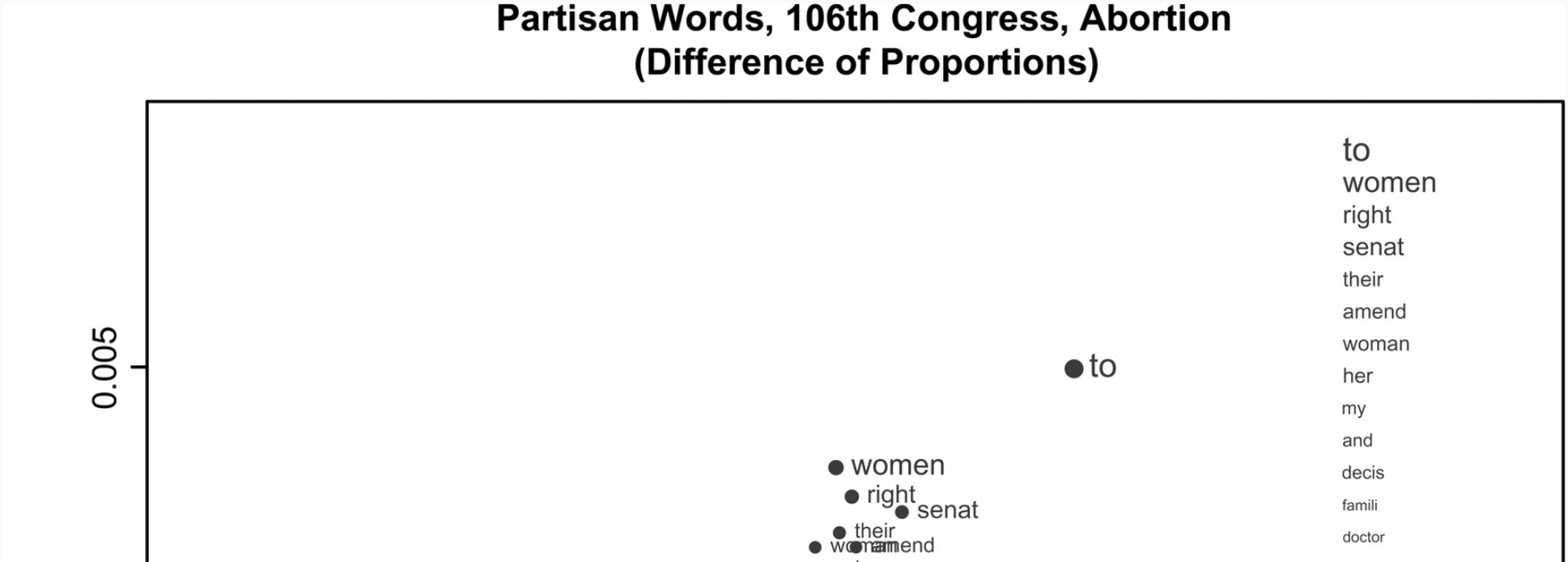
Vocabulary Usage

We can use known Z to characterize vocabulary usage directly, if we're careful…

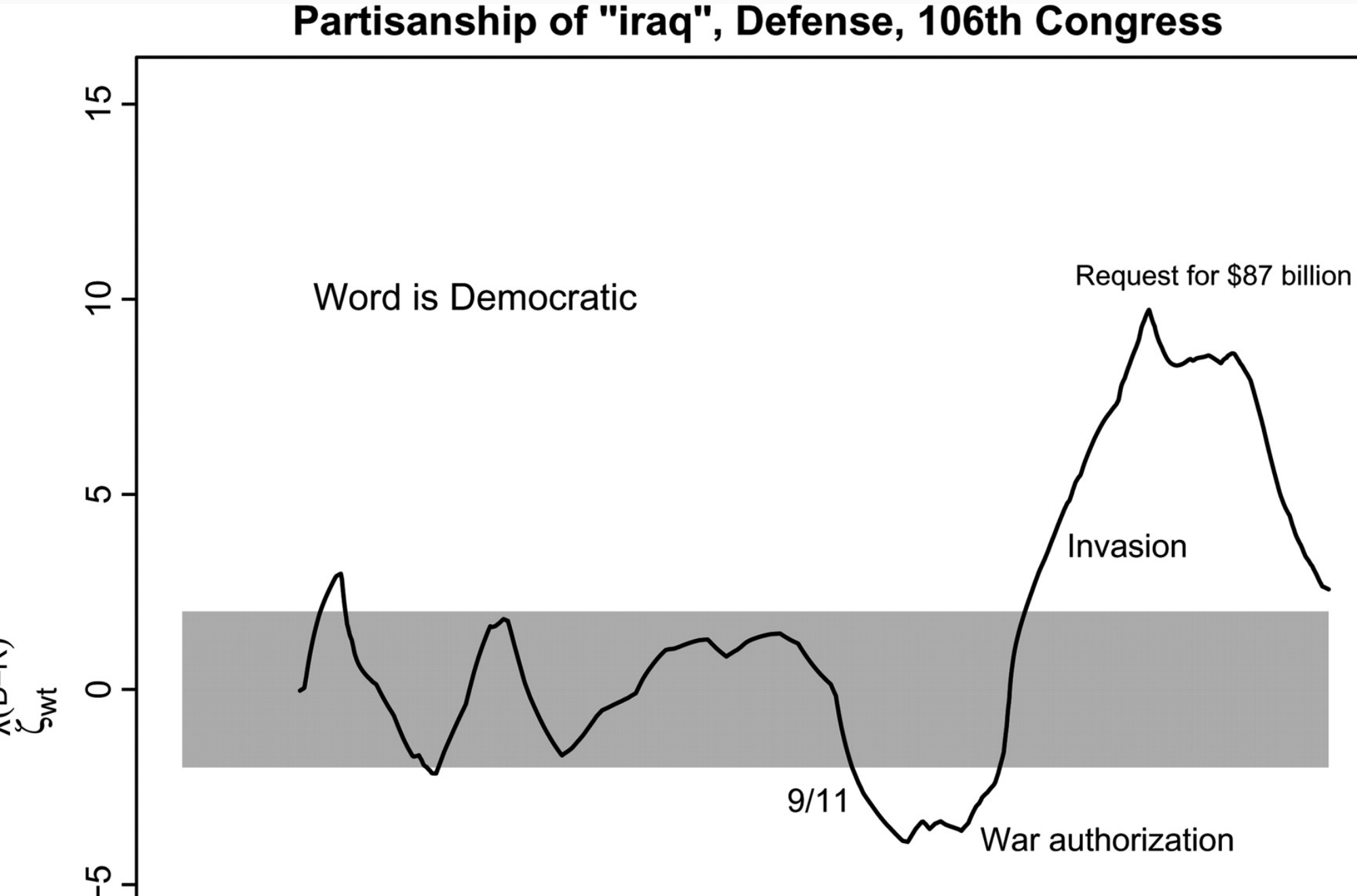
Monroe et al. (2008) compare different measures of 'partisan vocabulary' in abortion debates

Simple frequencies will be misleading

They settle for Laplace regularized odds-ratios



Lexical Instability



Designing a Content Analysis: Issue to Consider

Validity: does measurement reflect the truth?

Replicability: can you repeat the procedure?

Uncertainty: how uncertain are estimates?

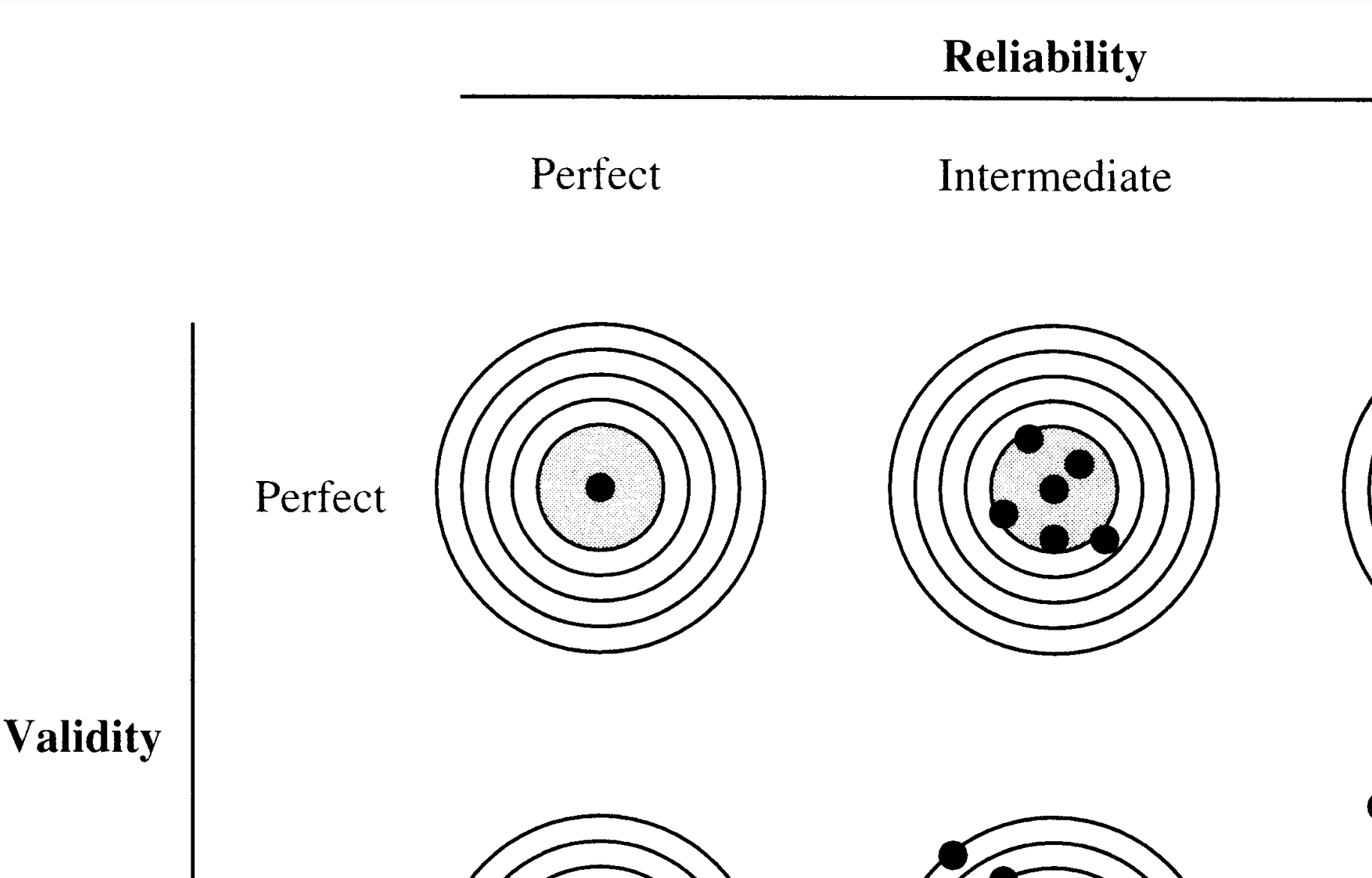
Accuracy: proportion of correctly classified documents?

Precision: proportion of documents assigned a category *i* that are actually about this category?

Recall: proportion of documents in category *i* classified correctly?

Dictionary: substantive, theory-driven, vocabulary from training set

Validity and Reliability



Validity and Reliability

Constructing Dictionaries

Always a trade-off between *precision* and *recall*.

> **Precision**: proportion of words used the way your dictionary assumes
> **Recall**: proportion of words used that way that are in your dictionary

Depends on *a priori* knowledge on categories

> Unknown categories: topic model $\rightarrow$ look at words strongly associated with discovered topics
> Known categories: human coding of entire document, discover discriminatory words from classification analysis
> Keyword in context analyses allow you to scan all contexts of a word
>
> How many of them are the sense or usage you want?