Beyond Bag of Words

Taking Statistical Text Mining to the Next Level

Andrew Fast, Ph.D.
Chief Scientist

Elder Research Inc.
300 W Main St., Suite 301
Charlottesville, Virginia 22903
(434) 973-7673
fast@dataminingslab.com

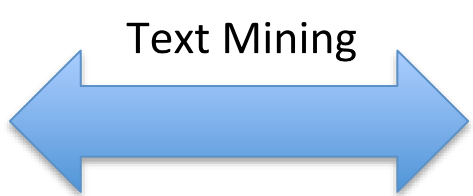
Text Analytics World, San Francisco, April 2013

Challenge of Text

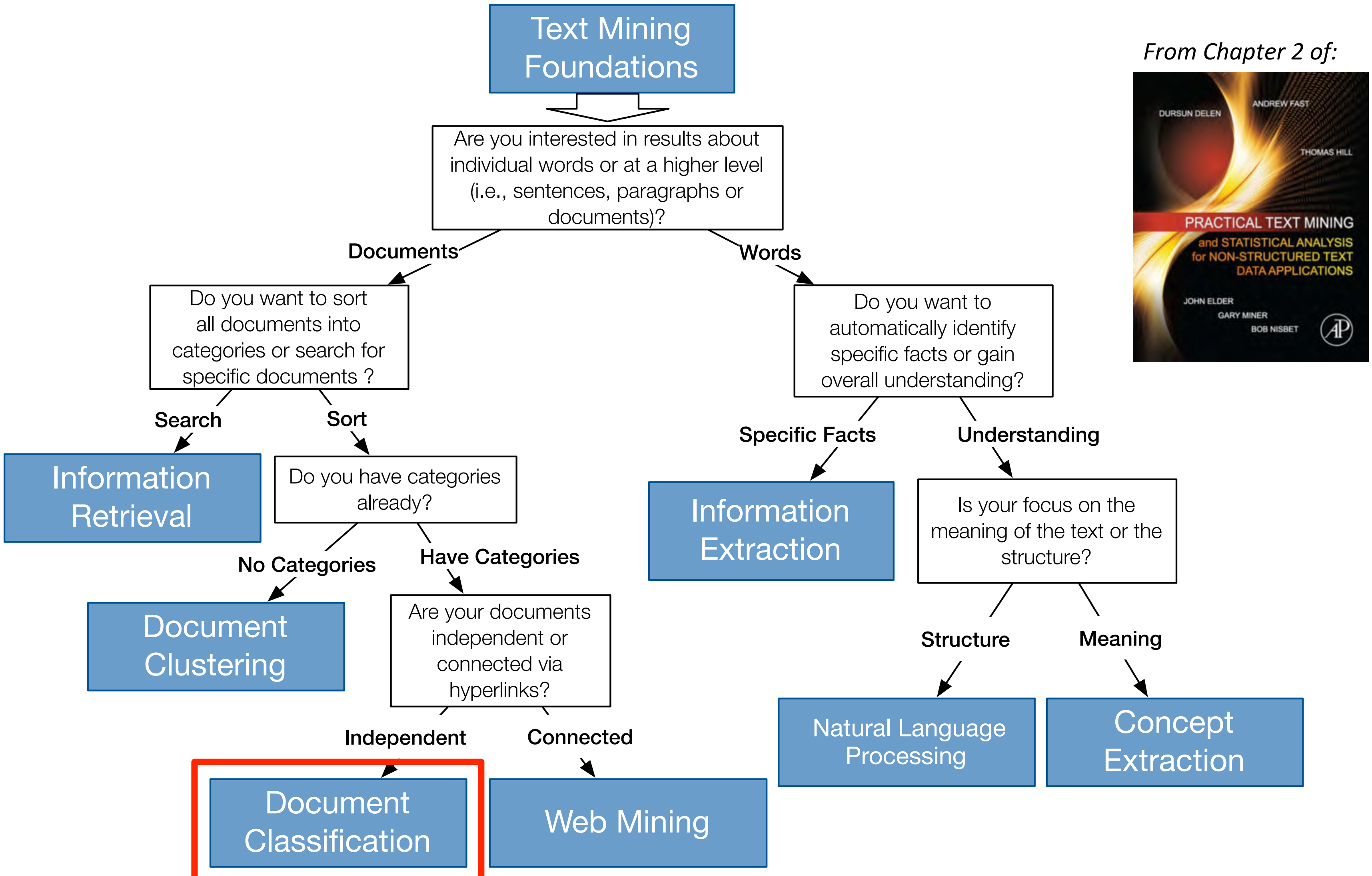
- Large amounts of unstructured textual data but still need understanding
- Do not know all useful features in advance
  - Useful features are unknown
  - Or, too labor intensive to enumerate all features
- Combination of structured (numerical) and unstructured data

Complementary Strengths

- Humans
  - Thoughtful
  - Nuance
  - Slow
- Machines
  - Repeatable
  - Brute Force
  - Fast



Focus on Document Classification



ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Text Mining vs. Text Analytics

	Text Mining	Both	Text Analytics
Approach	Statistical Machinery		Linguistic Rules
Inputs		Features of words and documents	
Performance		Equivalent (eventually)	
Effort	Creating training data		Tuning rule-sets
Strength	Flexibility		Human Understanding
Rule Generator	Algorithmic		Human

Naïve Bayes Classification

- A statistical text mining algorithm for document classification and categorization

Strengths

- Allows for identification of unknown rules
- Extends to new datasets with minimal work
- Incorporates multiple kinds of evidence
- Fast!

Weaknesses

- "Bag of Words" - Assumes all features are independent given the class labels
- No human insight or understanding into the text

“Bag of Words” Limits

“She can refuse to overlook our row,” he moped,
“unless I entrance her **with the** right present: **a** hit!”

Her moped is *presently* right **at the** entrance **to the**
overlook*;* she had hit **a** row **of** refuse *cans*!

Success Stories

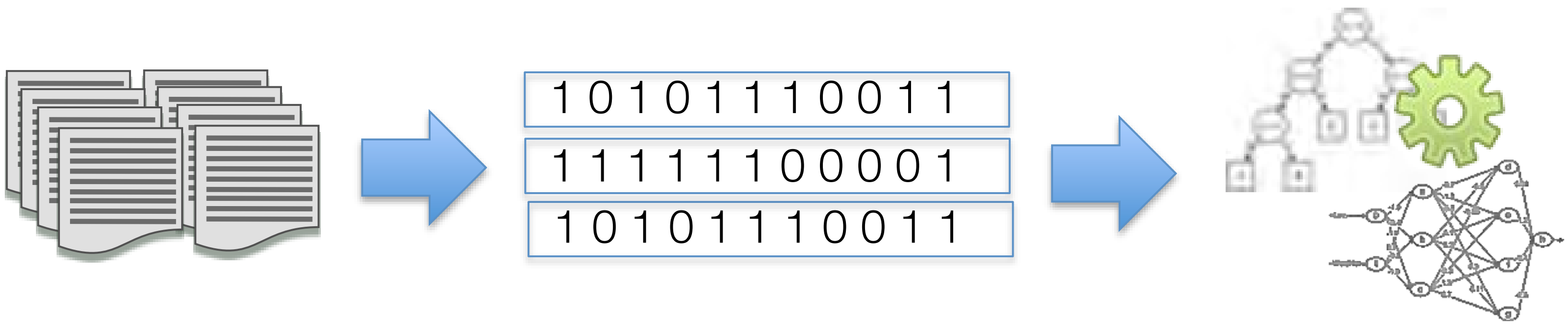
- Customer Satisfaction (sentiment analysis) for a major insurance company
- Predicting churn of mobile phone customers for nTelos
- Disability Approval for the Social Security Administration

Shared Characteristics

- Big Value
- Rich meaning
- Large collection of historical data
- Multiple meanings
- Mixed messages
- Messy text
- Short text
- Both structured data and unstructured text

"Bag of Words"

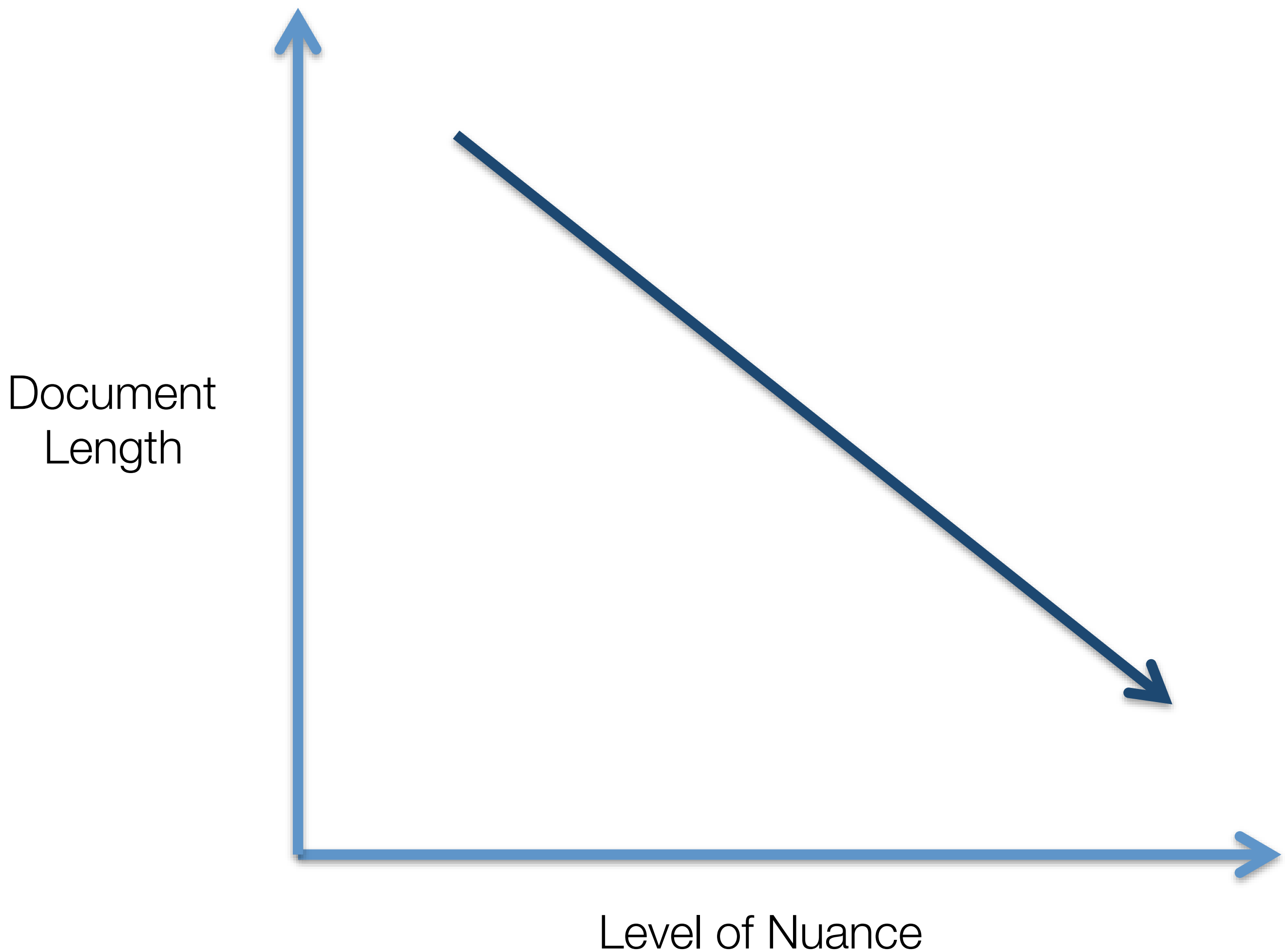
- Assumption that each word occurrence is independent as if drawn from a bag
  - Context and word order do not matter!



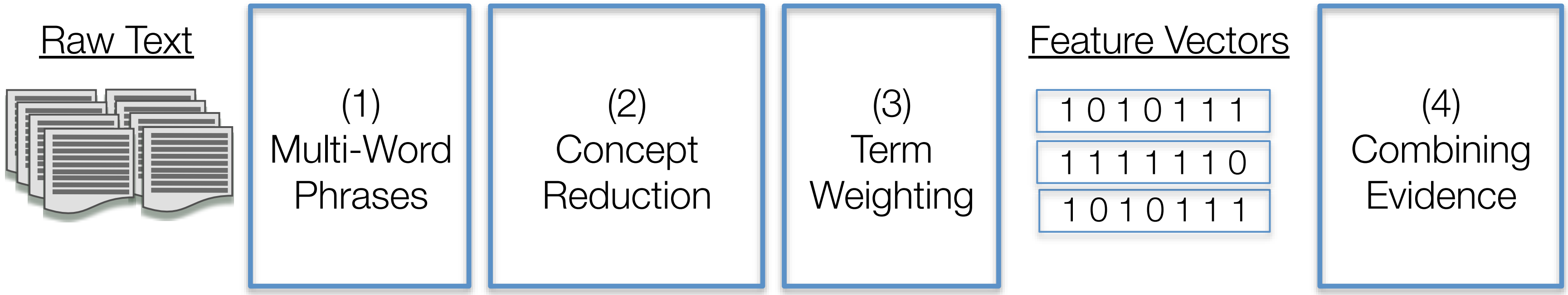
- Transform each document into a feature vector for input into statistical modeling algorithms

- Extremely high-dimensional space
  - Typically one dimension per word

Effectiveness of Bag of Words

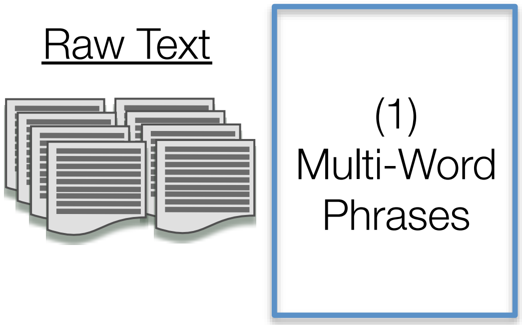


Beyond Bag of Words



- Focus on methods for improving the accuracy of statistical document classification by transforming the feature vector creation process

Multi-Word Phrases

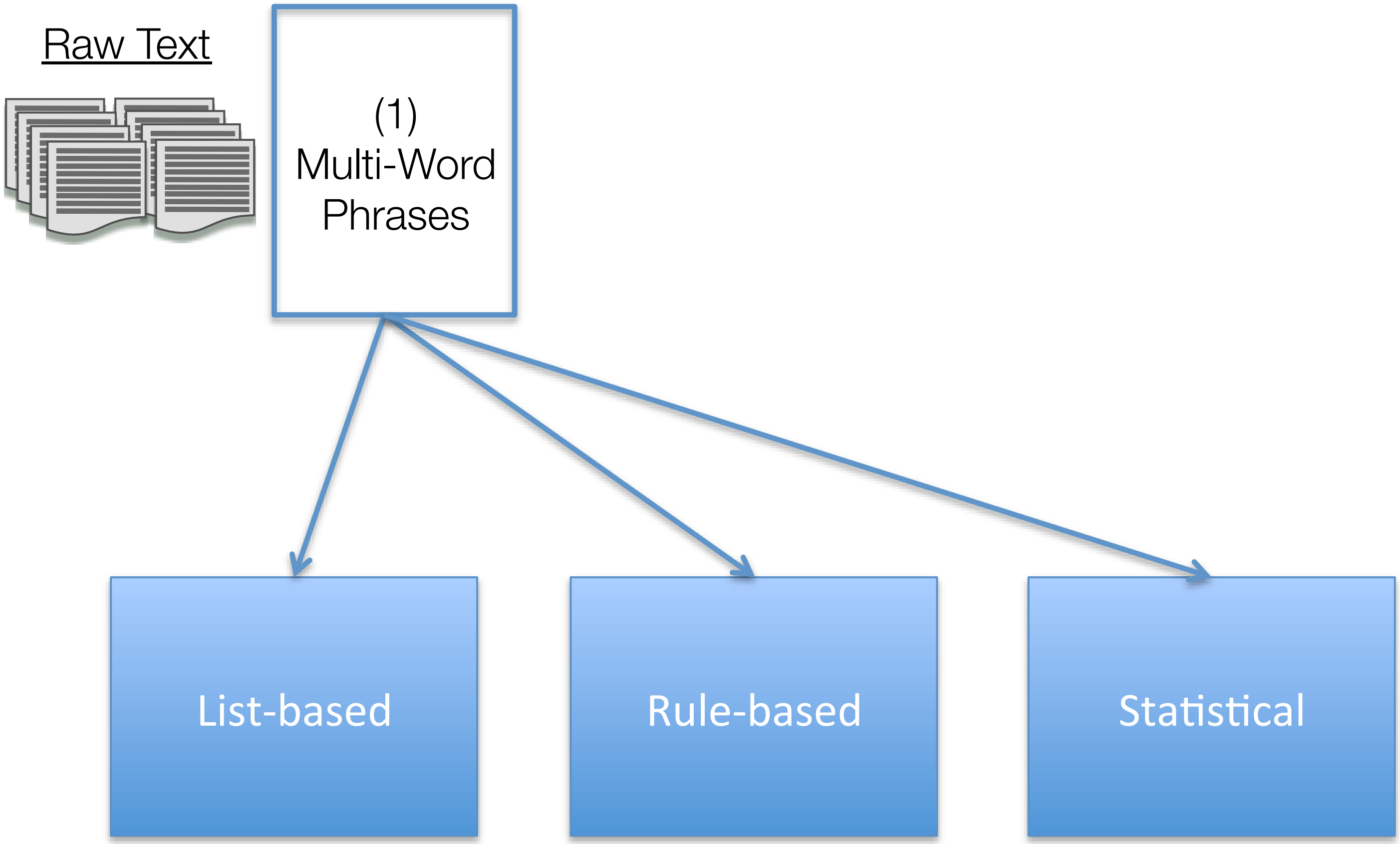


- Biggest exception for the bag of words assumption
- Also known as collocations

Examples of multi-word phrases

- “President of the United States”
- “moving violation”
- “homeowners insurance”
- “Lou Gehrig’s Disease”
- “Learning Disability”
- “Blackberry Pearl Flip”
- “HTC Desire”

Multi-Word Phrase Detection



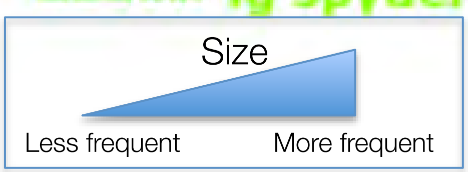
Case Study: nTelos Wireless

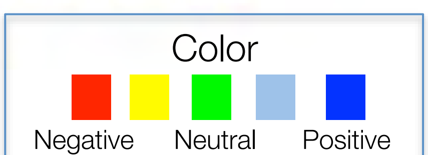
- *Task:* Identify customers likely to leave the network (“churn”)
- *Data:* Combination of structured and unstructured data
  - Text consists of customer service rep notes from customer interaction including summaries of customer responses.
- *Key Insight:* The model of the customers phone is major factor in customer satisfaction.

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

List-based Collocation

- Used a list of supported phones to group full phone models together





ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Overall Results: nTelos

- Adding textual features to structured data resulted in a 3.1% increase in prediction accuracy on hold-out data.

- Brand alone is churn neutral.

- Certain older phone models tied strongly to churn

- Also, customer provided equipment (CPE) is tied strongly to churn

Challenge: Unknown Phrases

- What happens when the helpful phrases are not known in advance?
  - Or, taxonomy too extensive to use efficiently?

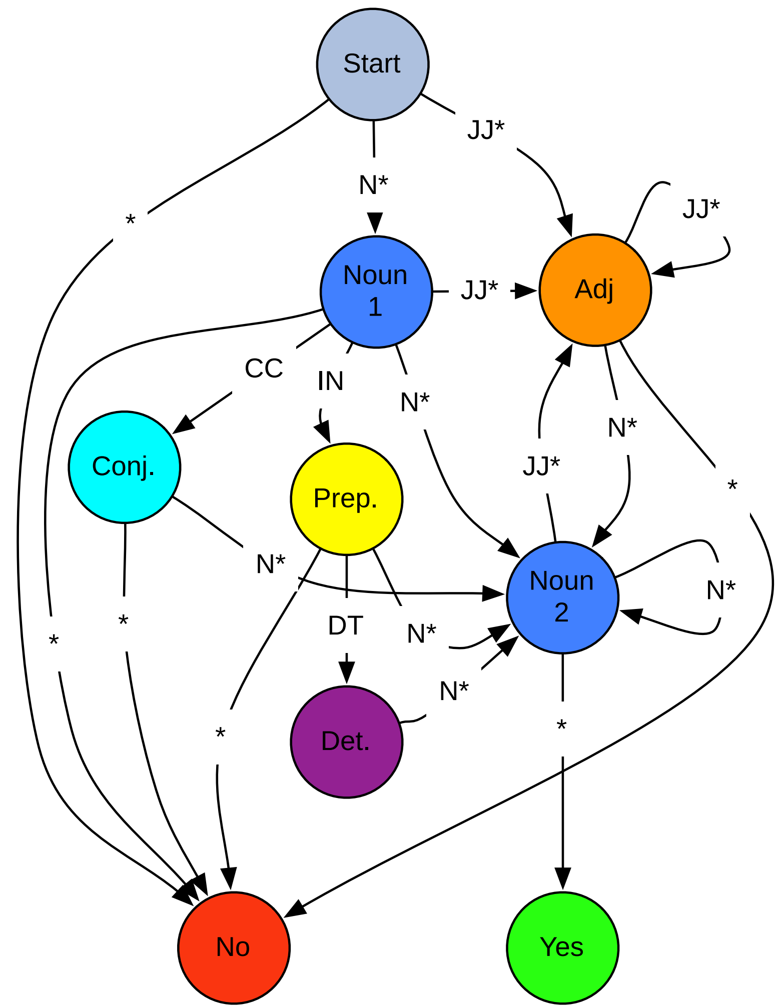
- SSA: medical concerns and diseases

- Insurance: positive and negative sentiment

Collocations using Rules

- Use regular expressions with Part-of-Speech tags
- Focused on Noun-based patterns

N* -> Nouns
JJ* -> Adjectives
CC -> Coordinating Conjunction
DT -> Determiner
IN -> Preposition



[Justeson and Katz 1995]

Collocations Using Statistics

- Use inductive modeling approach
  1. Use analysts to label training documents
  2. Build a statistical model from labeled data
  3. Apply model to new data

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Method Comparison

- you could lower my rates. I'm an excellent driver with no accidents or moving violations.

- $COMPANY$'s homeowner's coverage is good and priced reasonably. But $COMPANY$'s auto coverage is, at best, average yet overpriced. The same applies to our motorcycle policy - average yet overpriced.

Statistical Only	Both	Rule-based Only

Entity Extraction

- Entity Extraction is a specialized form of multi-word phrase detection
  - Limited to proper names, organizations, and places.
- Often more accurate than general purpose detectors for names
- Both rule-based and statistical approaches.

Concept Reduction

Raw Text

(1) Multi-Word Phrases

(2) Concept Reduction

- Combine multiple expressions of the same concept into a single term

Concept Reduction

- Replace tokens representing the same concept with a standard token
  - Synonyms
  - Dates
  - Specific proper nouns
  - Abbreviation Expansion

- How do you do this?
  - Lists, Regexes, Rules
  - Automatic synonym detection

Word Clustering

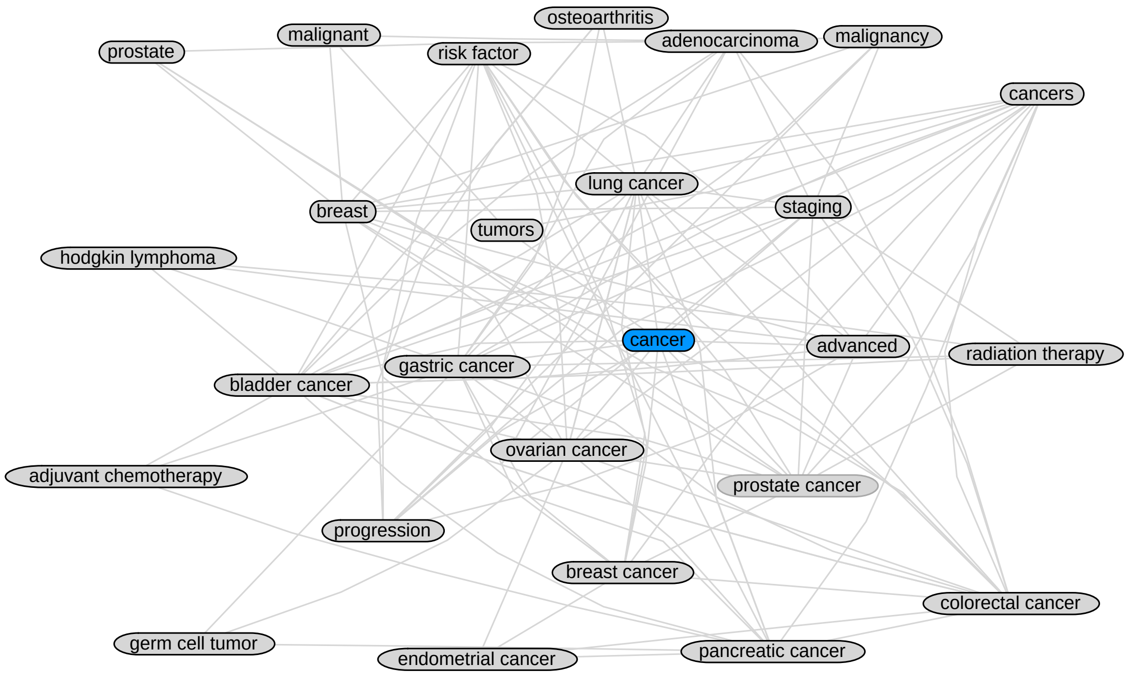
- Uses similar techniques to document clustering
- …but uses a "term-context" vector instead.

Document 1: My dog ate my homework.
Document 2: My cat ate the sandwich.
Document 3: A dolphin ate the homework.

	a	ate	cat	dog	dolphin	homework	my	sandwich	the
Document 1	0	1	0	1	0	1	2	0	0
Document 2	0	1	1	0	0	0	1	1	1
Document 3	1	1	0	0	1	1	0	0	1

	a	ate	cat	dog	dolphin	homework	my	sandwich	the
a	0	1	0	0	1	0	0	0	0
ate	0	0	0	0	0	1	1	2	2
cat	0	1	0	0	0	0	0	0	1
dog	0	1	0	0	0	0	1	0	0
dolphin	0	1	0	0	0	0	0	0	1
homework	0	0	0	0	0	0	0	0	0
my	0	2	1	1	0	1	0	0	0
sandwich	0	0	0	0	0	0	0	0	0
the	0	0	0	0	0	1	0	1	0

Concept Extraction



- Automatically detect conceptually related words using statistical clustering

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Case Study: Satisfaction Survey

- *Task*: Identify sentiment and level of satisfaction from combination of numerical and textual survey data for an insurance company
- *Data:* Combination of structured and unstructured data
  - Text consists of open-ended comments supplied by the customer
- *Key Insight*: Not all negative sentiment about the company, the survey itself was a major source of negative sentiment!

Understanding Customer Behavior

Statistical Categorization Summary

Area	Positive Experience	Positive	No Experience	Neutral	Mixed	Suggestion	Negative	Negative Experience	Off-Topic
Agency	1.4%	61.1%	17.8%	7.7%	0.3%	0.1%	9.2%	1.5%	0.9%
Billing	1.0%	47.5%	16.4%	18.5%	0.3%	2.6%	10.1%	1.2%	2.3%
Claims	2.0%	33.3%	40.7%	16.0%	0.6%	0.0%	3.4%	2.5%	1.5%
Coverage	1.0%	43.3%	14.2%	20.7%	1.0%	0.6%	14.4%	0.3%	4.5%
Onboarding	0.7%	29.2%	46.5%	16.0%	0.1%	0.0%	3.7%	0.4%	3.4%
Phone	0.7%	28.3%	20.2%	45.6%	0.1%	0.0%	2.2%	0.4%	2.5%
Policy	0.5%	42.3%	27.5%	22.1%	0.0%	0.3%	3.7%	0.4%	3.2%
Premium	0.9%	39.4%	26.2%	20.5%	0.0%	0.2%	9.5%	0.2%	3.1%
Renewal	0.7%	49.3%	18.6%	18.2%	0.1%	0.5%	8.4%	0.8%	3.2%
Value/Price	0.6%	38.1%	14.6%	16.6%	0.1%	0.4%	27.4%	0.3%	1.9%
Website	1.4%	30.8%	19.3%	40.5%	0.0%	0.5%	4.9%	0.1%	2.3%

Term Weighting

Raw Text

(1) Multi-Word Phrases

(2) Concept Reduction

(3) Term Weighting

- Focus on methods for improving the accuracy of statistical document classification by transforming the feature vector creation process

Challenge: Weighting Scores

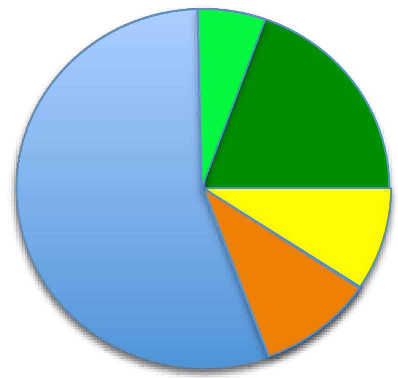
- How do you weight rare concepts?

- Naïve Bayes:
  - Count number of times a word is associated with a particular outcome
  - Leads to skewed weights on rare concepts

Case Study: SSA Disability Approval

- **Pain:** Approval process is long, bureaucratic

Up to 2 Years !



- **Goal:** Fast-track "easy" cases
- **Challenge:** Free-text on disability application
- **Result:** 20% of Approvals possible immediately and with greater consistency

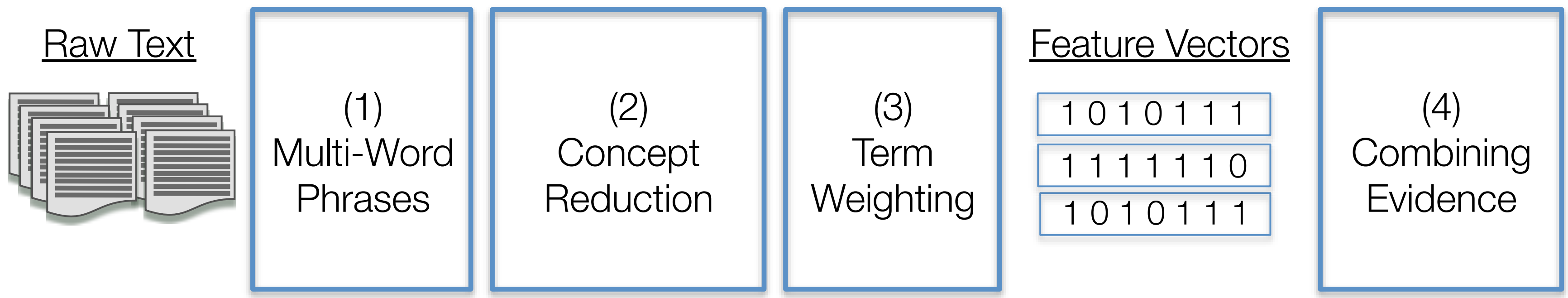
Weighting with a Prior

- Draw from Bayesian statistics and smooth the raw count with an empirical prior
  - Use baseline probability of the most probable classification
    - For SSA, roughly 33% of applications approved
  - Counts for each word are initialized with the baseline probability
    - Known as Shrinkage, James-Stein Estimator, Ridge Regression, etc.

- Hypothetical Example: Multiple Myeloma
  - Appears 5 times, 4 times was approved = 80% predicted weight
  - With prior of 33%, now weight is 5/8 = 62.5% predicted weight

- More data outweighs the prior

Combining Evidence



- How to score documents so that “strong” words are emphasized and “weak” words are ignored

Example

- “Multiple Myeloma I have been diagnosed with Multiple Myeloma (cancer of the bone marrow) and am currently undergoing treatment to prepare me for an autologous stem cell transplant. There has been a brain tumor associated with this, for which I have had....”

Combining Weights

- Common aggregations don't match medical domain requirements
  - SUM: many symptoms increases probability of predicting approval
  - MAX: ignores multiple serious symptoms
  - AVG: minor symptoms water down major symptoms

- Naïve Bayes uses *maximum a priori* (MAP) approach
  - All evidence combined equally

Our approach for SSA

If (no data), then use prior

Else If (max(probability) < 0.5) then use that max.

Else:

i. Ignore concepts with probability < 0.5

ii. Combine the remaining ones with a log-likelihood formula and use the resulting joint probability.

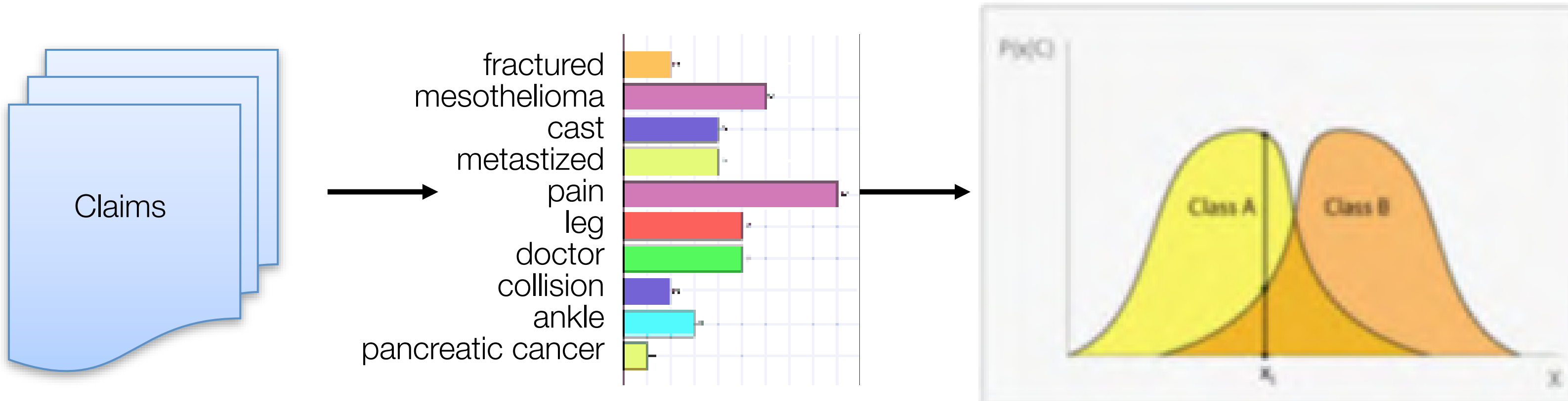
Example Weights

Table 2-2: Example problem case for QD target definition

TOKEN	PROBABILITY APPROVED QD	ln(a / ~a)
stem-cell transplant	93.8%	2.71
bone cancer	85.0%	1.74
multiple myeloma	76.6%	1.18
marrow	78.6%	1.30
brain tumor	63.0%	0.53
Score		7.46
Final Percentage	**99.94%**	

QD = 0

SSA Solution

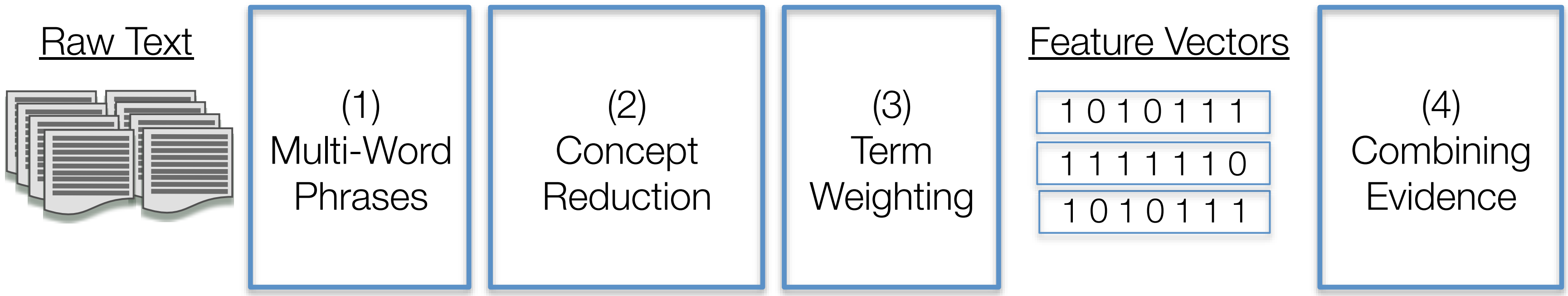


- Convert individual words and phrases to features
- Exploit custom method for combining evidence from multiple features
- Text classifier accuracy equivalent to committee of experts

  30% baseline -> 90% model accuracy

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Summary



- Described methods for improving the accuracy of statistical document classification by transforming the feature vector creation process

Contact Information

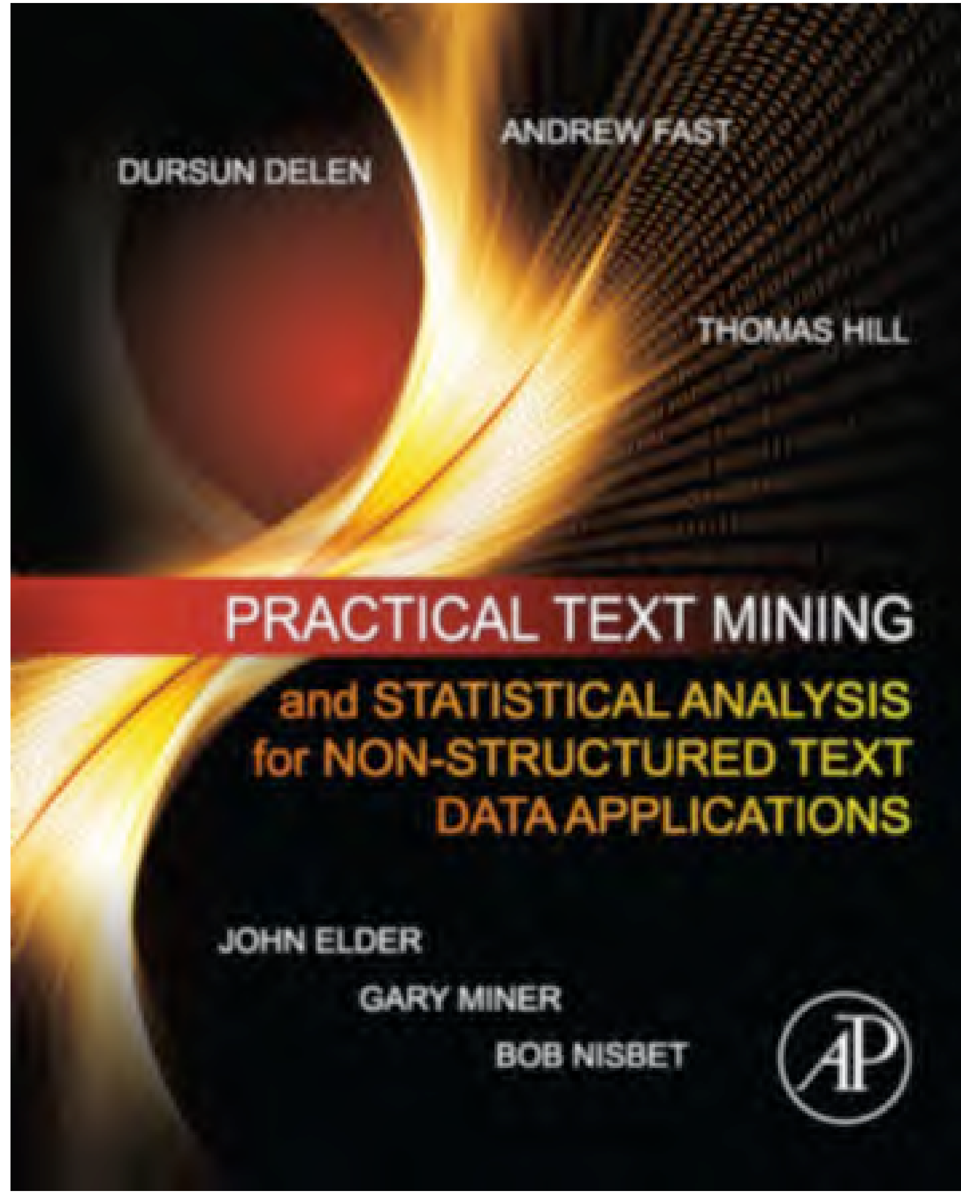
Andrew Fast, Ph.D.
Chief Scientist

fast@datamininglab.com
(434) 973-7673
www.datamininglab.com

ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Practical Text Mining

- Winner of the 2012 PROSE award for Computing and Information Science
- Written for a technical audience seeking more text experience
- Includes trial versions of major software tools



ELDER RESEARCH INC.
DATA MINING & PREDICTIVE ANALYTICS

Andrew Fast
Chief Scientist, Elder Research, Inc.

Dr. Andrew Fast leads research in Text Mining and Social Network Analysis at Elder Research, the nation's leading data mining consultancy. ERI was founded in 1995 and has offices in Charlottesville VA and Washington DC, (www.dataminingLab.com). ERI focuses on Federal, commercial, investment, and security applications of advanced analytics, including stock selection, image recognition, biometrics, process optimization, cross-selling, drug efficacy, credit scoring, risk management, and fraud detection.

Dr. Fast graduated Magna Cum Laude from Bethel University and earned Master's and Ph.D. degrees in Computer Science from the University of Massachusetts Amherst. There, his research focused on causal data mining and mining complex relational data such as social networks. At ERI, Andrew leads the development of new tools and algorithms for data and text mining for applications of capabilities assessment, fraud detection, and national security.

Dr. Fast has published on an array of applications including detecting securities fraud using the social network among brokers, and understanding the structure of criminal and violent groups. Other publications cover modeling peer-to-peer music file sharing networks, understanding how collective classification works, and predicting playoff success of NFL head coaches (work featured on ESPN.com). With John Elder and other co-authors, Andrew has written a book on Practical Text Mining, that was awarded the prose Award for Computing and Information Science in 2012.