AN INTRODUCTORY LOOK AT STATISTICAL TEXT MINING FOR HEALTH SERVICES RESEARCHERS

James McCart, PhD and Stephen Luther, PhD

Question 1

- What is your area of expertise?
  1. Clinician
  2. Informatics Researcher
  3. Other HSR&D Researcher
  4. Administrator
  5. Other

Question 2

- Which of the following methods have you used? (Select all that apply)
  1. Natural language processing (NLP)
  2. Data mining
  3. Statistical text mining
  4. None of the above

Goals of Presentation

- Describe how studies of statistical text mining (STM) relate to traditional HSR research
- Provide an overview of the STM process
- Discussion and demo of software

Acknowledgments

- HSR&D HIR 09-002: Consortium for Healthcare Informatics Research (CHIR)
- HSR&D/RR&D Center of Excellence: Maximizing Rehabilitation Outcomes (Tampa)
- HSR&D IIR 05-120: Using Knowledge Discovery Strategies to Identify Fall-Related Injuries

- This presentation is based on an HSR&D Workshop given at 2011 Annual Meeting
  - James McCart, Jay Jarman, Dezon Finch, & Steve Luther

CHIR

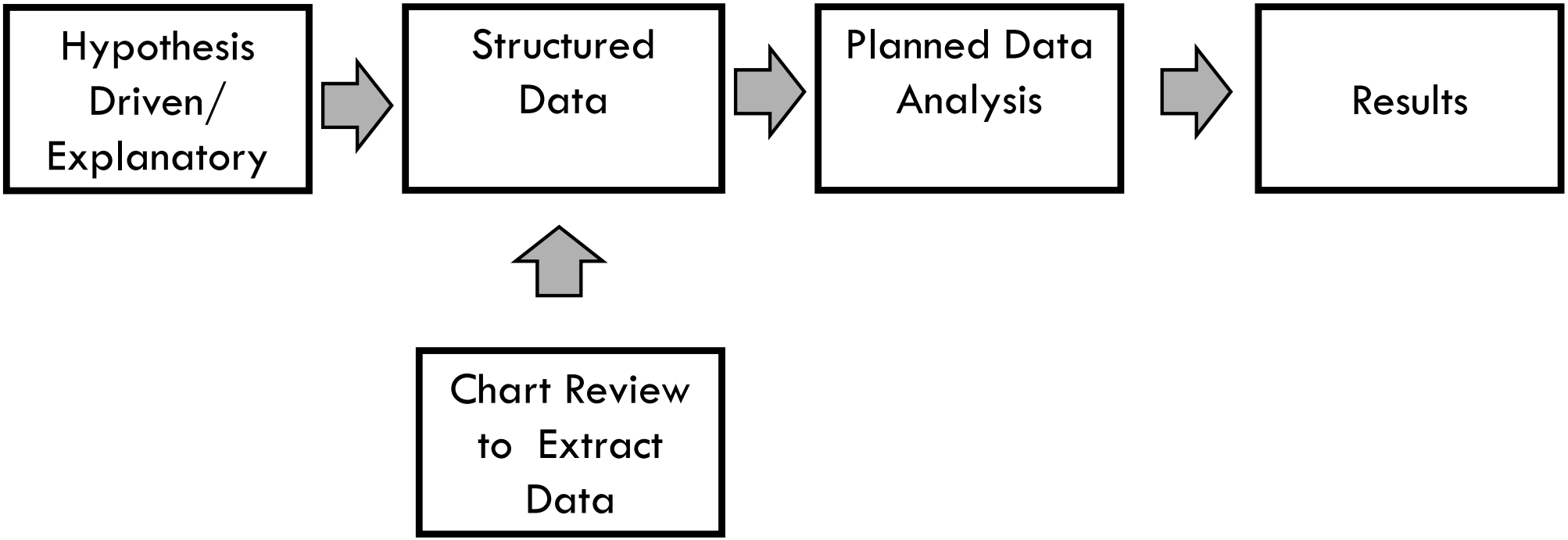
HSR&D/RR&D Research Center of Excellence
Maximizing Rehabilitation Outcomes
Tampa, FL

James Haley Veterans' Hospital and Clinics Tampa

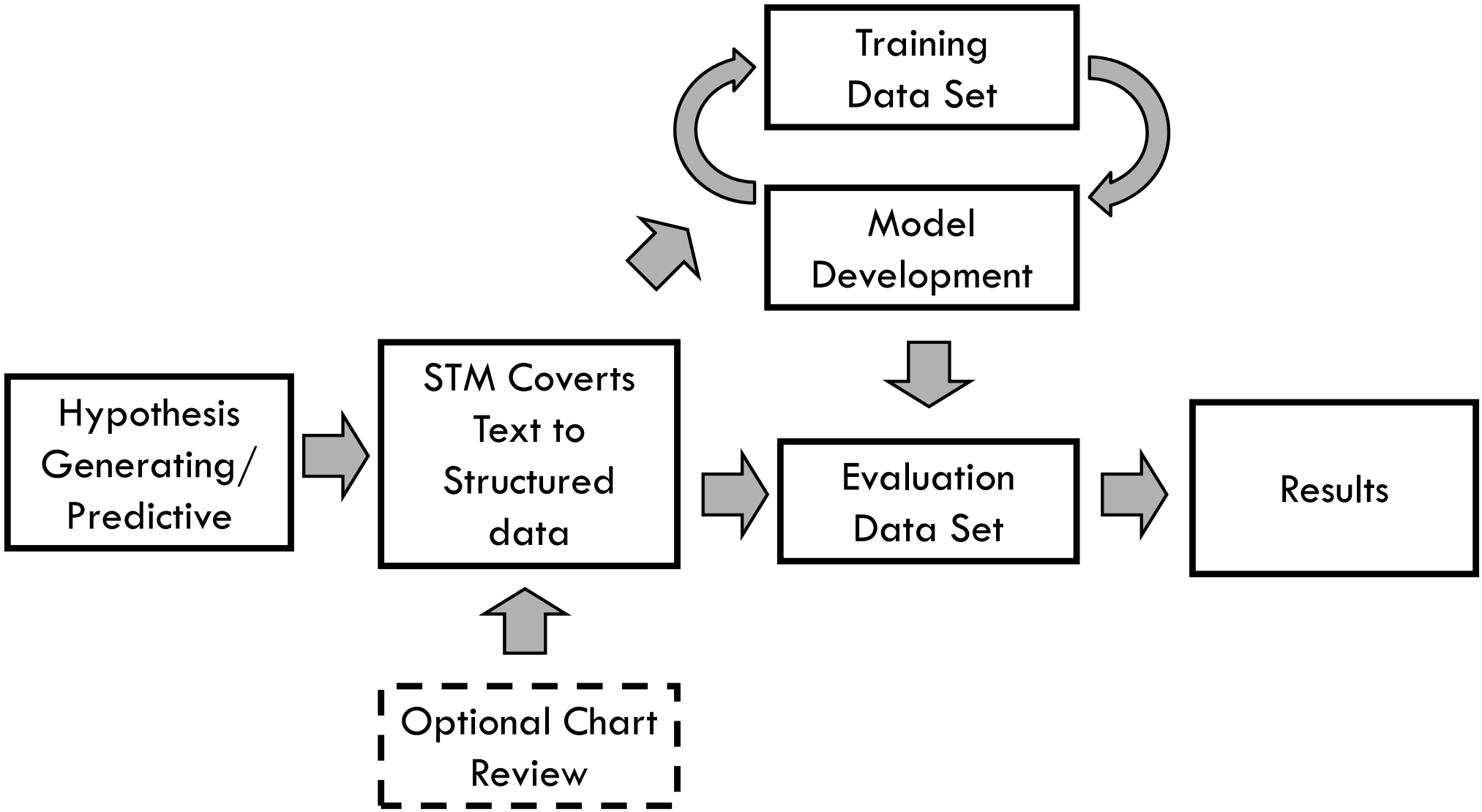
Natural Language Processing & Statistical Text Mining

- Natural Language Processing (NLP)
  - Analyze and understand natural language
  - E.g., information extraction (chart review)
- Statistical Text Mining (STM)
  - Extract patterns from documents
  - E.g., prediction (classification)
  - Similar to data mining

Text in Traditional HSR



Statistical Text Mining in HSR

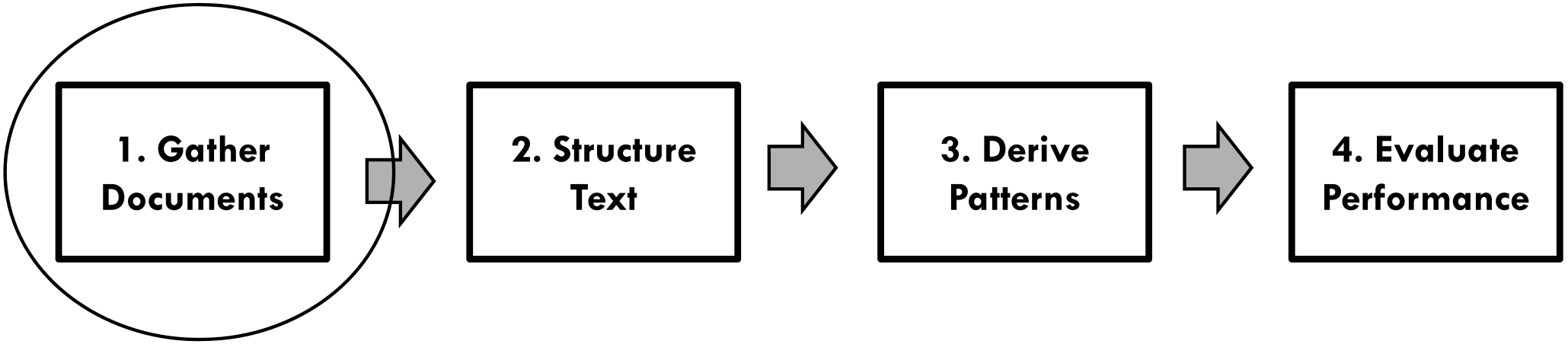


Applications in Research

- Classification
  - Genomic studies
  - Disease surveillance
  - Risk assessment
  - Cohort identification
- Knowledge Discovery
  - Hypothesis generation

STATISTICAL TEXT MINING PROCESS

Statistical Text Mining Process

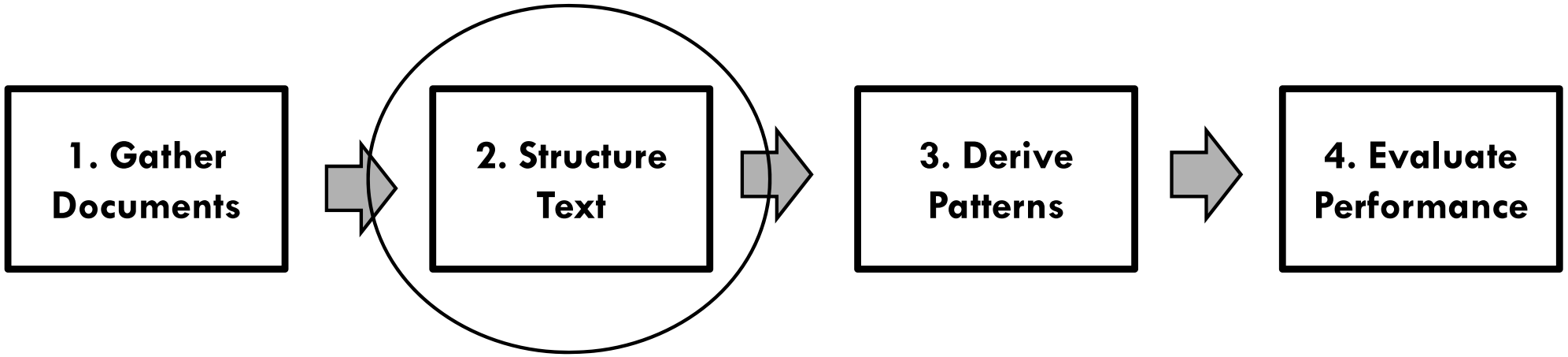


Step 1. Gather Documents

- Collection of documents

- Classification task – training and evaluation
  - Reference standard – label assigned to each document in the dataset
    - Label not available?
      - Annotate using Subject Matter Experts (SMEs)
      - Can be time consuming and expensive

Statistical Text Mining Process



Step 2. Structure Text

- Convert unstructured text into structured data
- Four sub-steps
  a) Create a term-by-document matrix
  b) Split data
  c) Weight the matrix
  d) Perform dimension reduction

Example Document Collection

Doc 1: smoking two packs per day

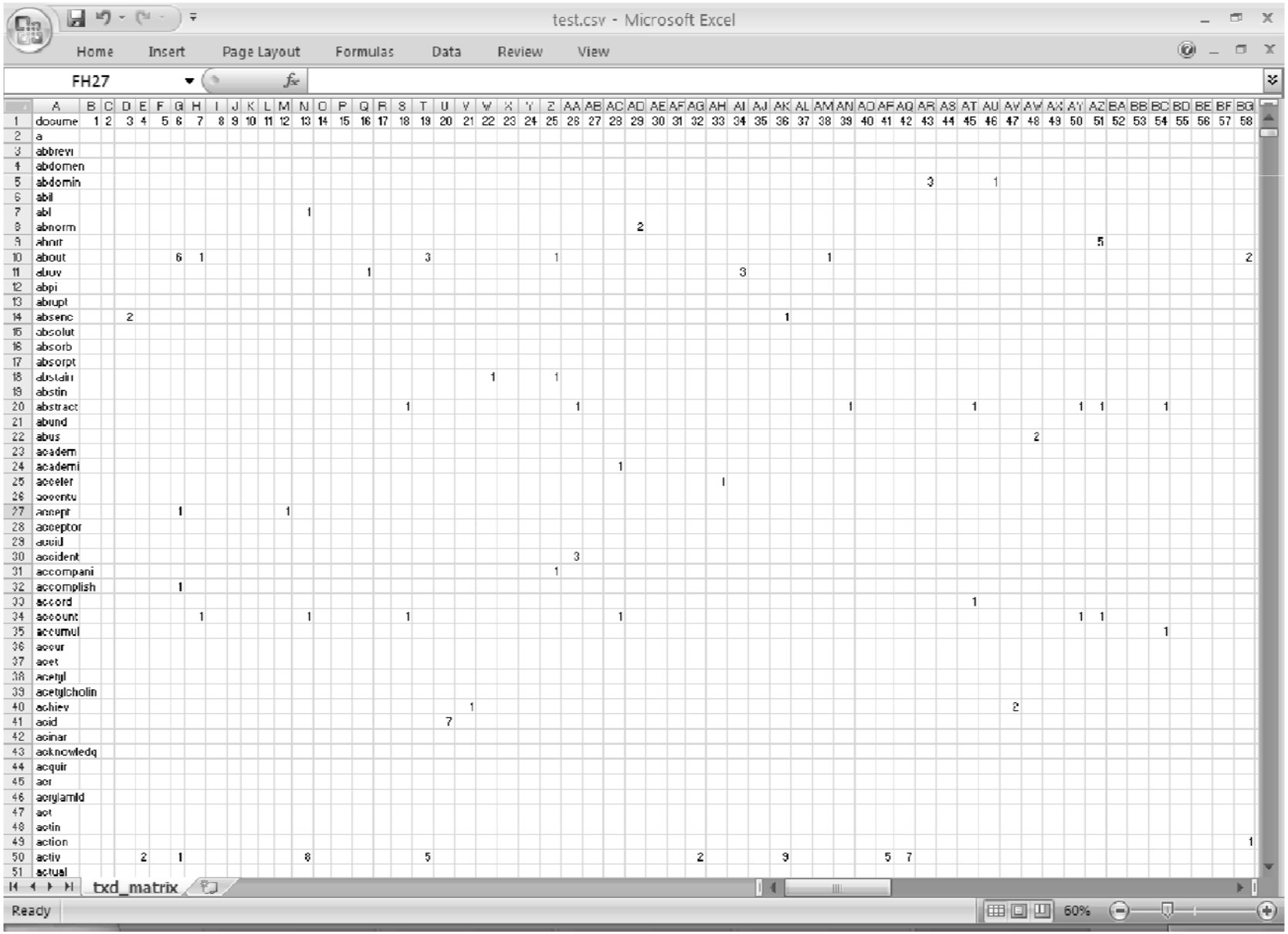
Doc 2: cough persisted for two weeks

Doc 3: motivated to quit smoking

Step 2a. Create a Term-by-Document Matrix

		Documents		
		Doc 1	Doc 2	Doc 3
Terms	cough	0	1	0
	day	1	0	0
	for	0	1	0
	motivated	0	0	1
	packs	1	0	0
	per	1	0	0
	persisted	0	1	0
	quit	0	0	1
	smoking	1	0	1
	to	0	0	1
	two	1	1	0
	weeks	0	1	0

Example Term-by-Document Matrix



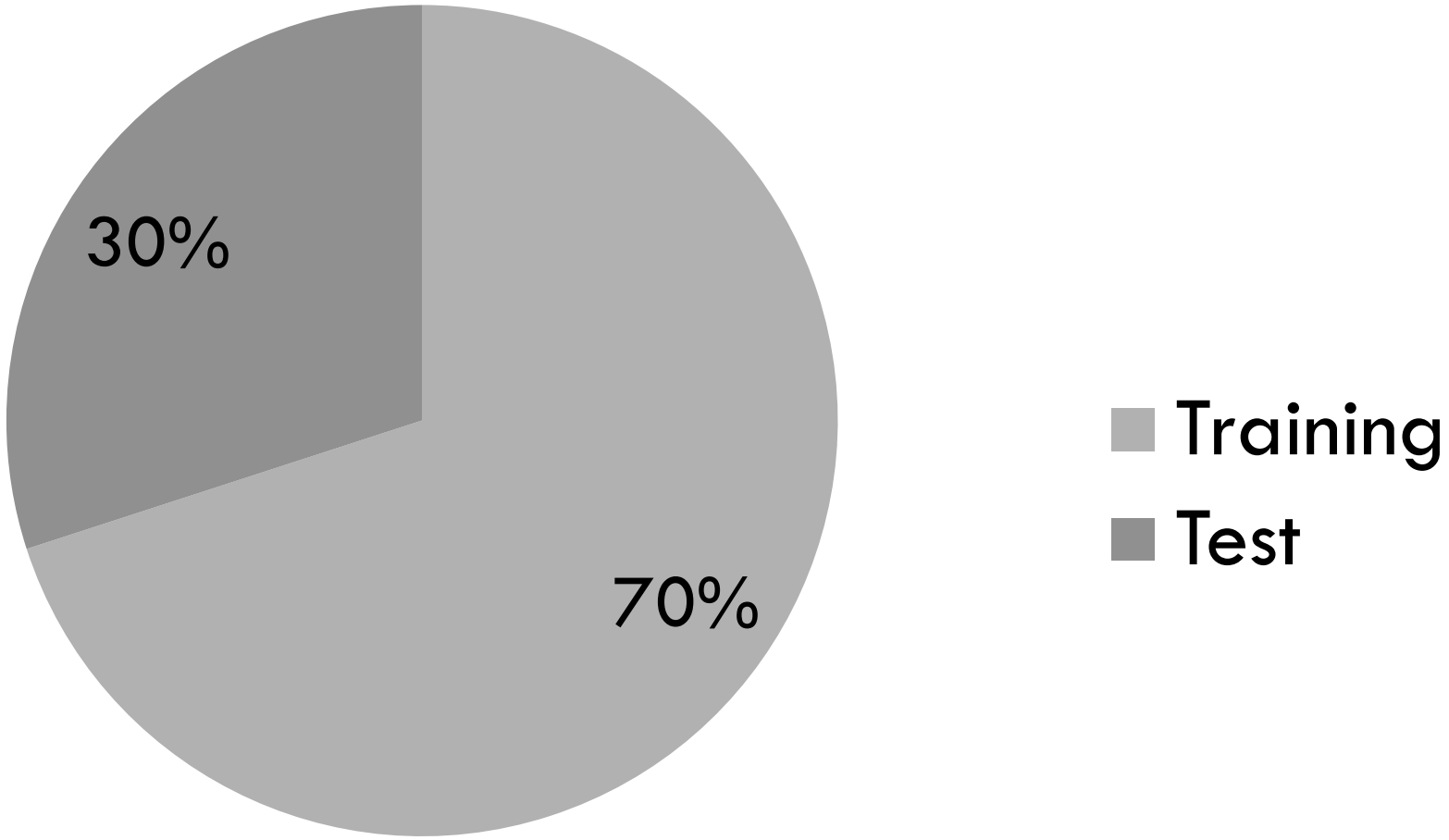
Creating a Term-by-Document Matrix

- Stop list – remove common words
  - "and", "the", "a"
- Filter terms by length
- Stemming – reduce words to their base form
  - "administer", "administers", "administered" → "administ"
- *n*-grams – include phrases of *n* sequential words
  - "…regional medical center" → "regional medical", "medical center"

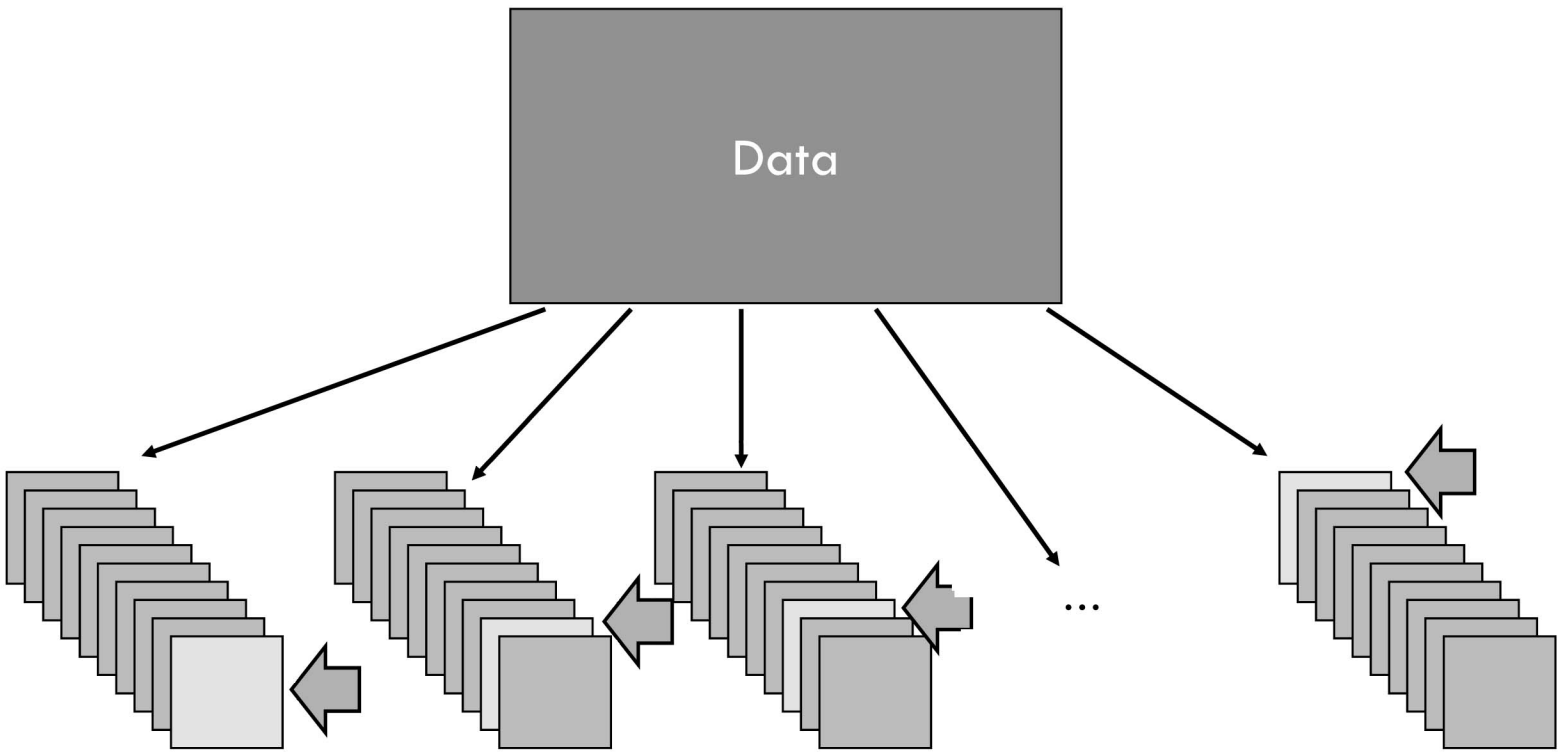
Step 2b. Split the Data

- Keep part of data separate to evaluate model
  - Overfitting
- Techniques
  - Training / Test split
  - X-fold Cross Validation

Training / Test Split



X-Fold Cross Validation



Step 2c. Weight the Matrix

- Importance
- Three components
  - Local
    - How informative a term is in a document
  - Global
    - How informative a term is across all documents
  - Normalization
    - Reduces impact of document length

weight = local * global * normalization

Example Local Weighting

Term Frequency

	Doc 1	Doc 2	Doc 3
Term 1	1	2	0
Term 2	0	4	3
Term 3	1	0	0

Binary

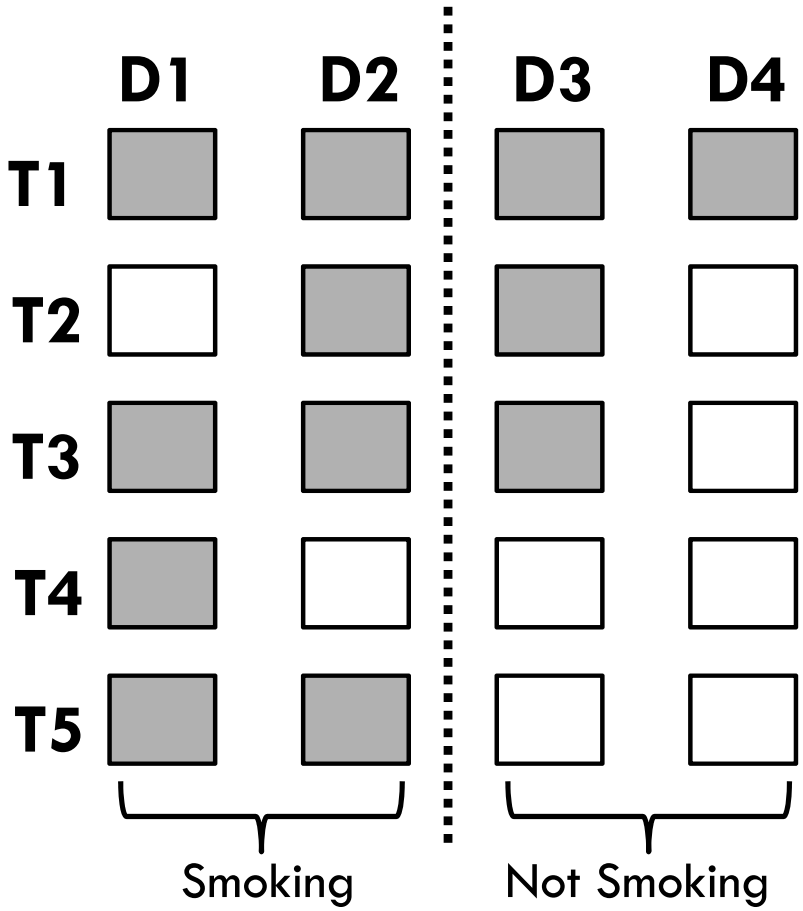
	Doc 1	Doc 2	Doc 3
Term 1	1	1	0
Term 2	0	1	1
Term 3	1	0	0

Log Term Frequency

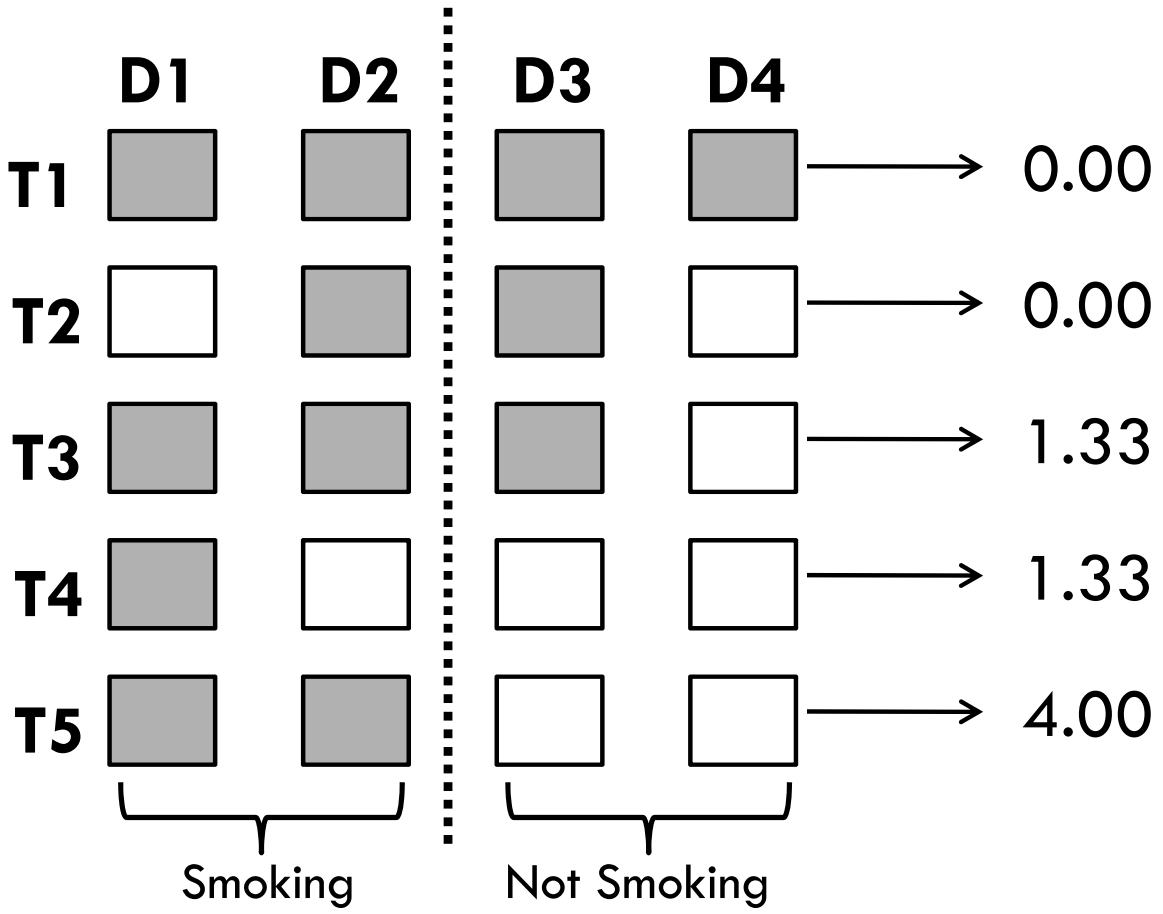
	Doc 1	Doc 2	Doc 3
Term 1	1.00	1.30	0.00
Term 2	0.00	1.60	1.48
Term 3	1.00	0.00	0.00

$$if\ (tf > 0)$$
$$1 + \log(tf)$$

Example Global Weighting



Example Weighting – X^2



Step 2d. Perform Dimension Reduction

- Characteristics of term-by-document matrix
  - Sparse
  - High dimensionality – "curse of dimensionality"

- Reduce number of dimensions
  - Remove terms occurring in only 1 document
  - Retain top N terms
  - Latent Semantic Analysis (LSA)

Warning!

- **Geeky statistical information ahead**
- **Proceed with caution**

Image from http://openclipart.org/detail/122119/warning-sign-by-dear_theophilus

Latent Semantic Analysis

- Singular Value Decomposition (SVD)
  - Similar to Principal Component Analysis or Factor Analysis
  - Creates dimensions (vectors) that summarize term / document information
    - Select k dimensions for analysis

Example – SVD

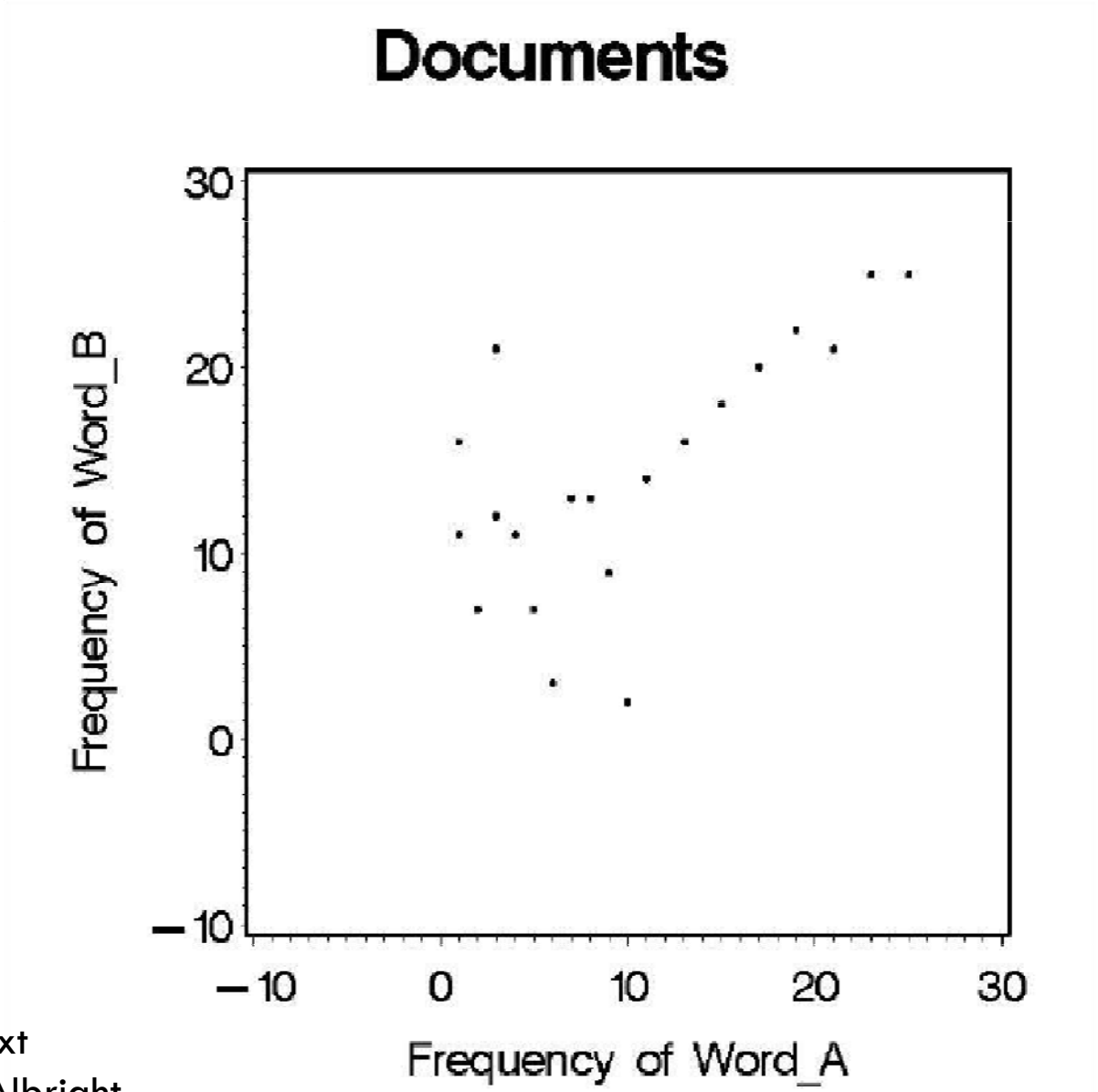


Figure from "Taming Text with the SVD" by Russ Albright

Example – SVD

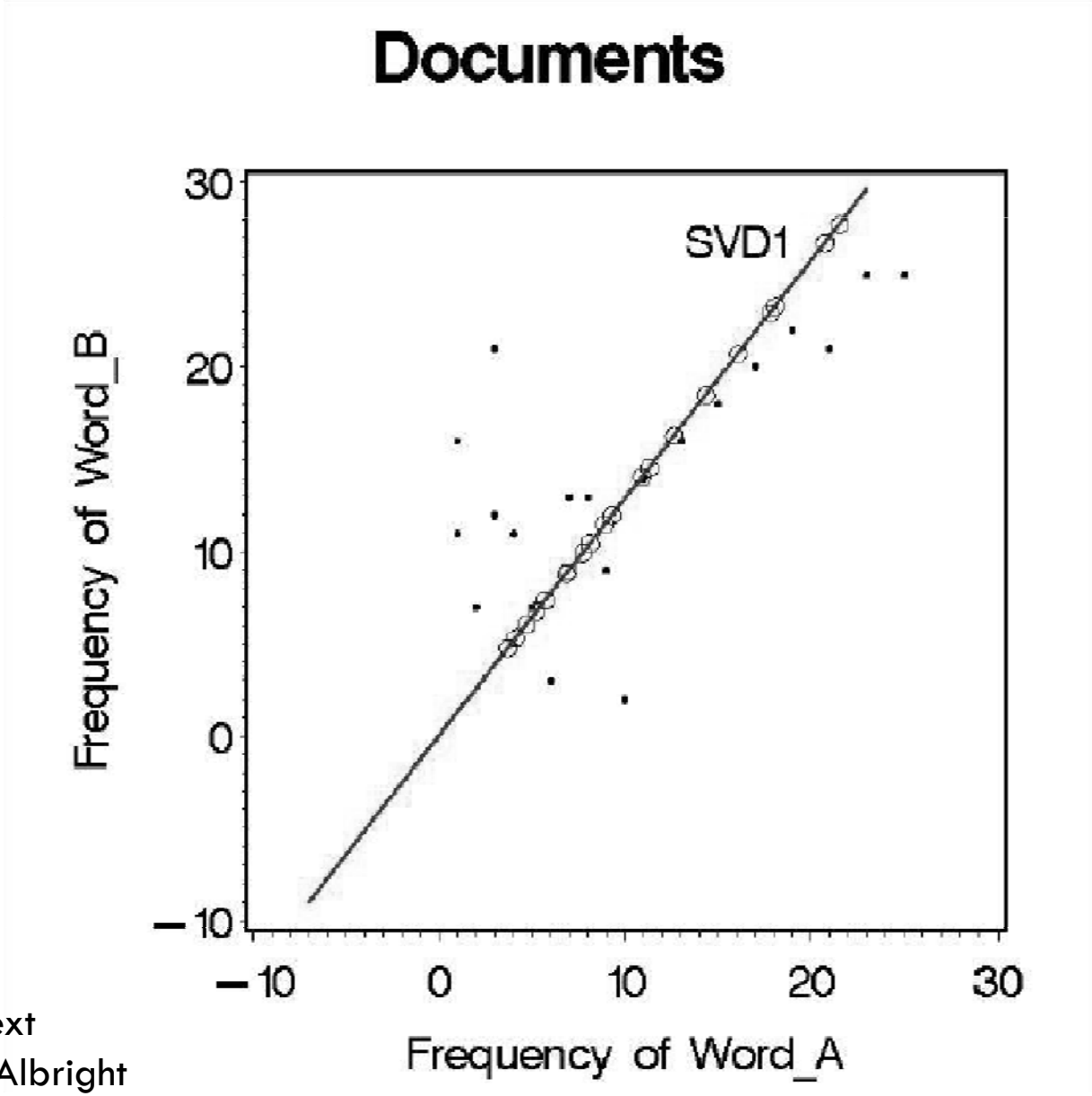
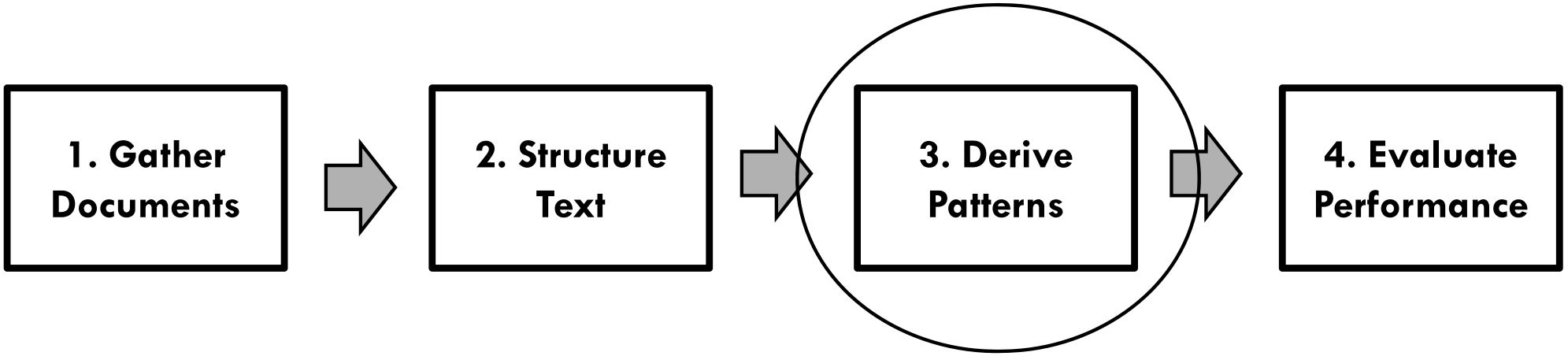


Figure from "Taming Text with the SVD" by Russ Albright

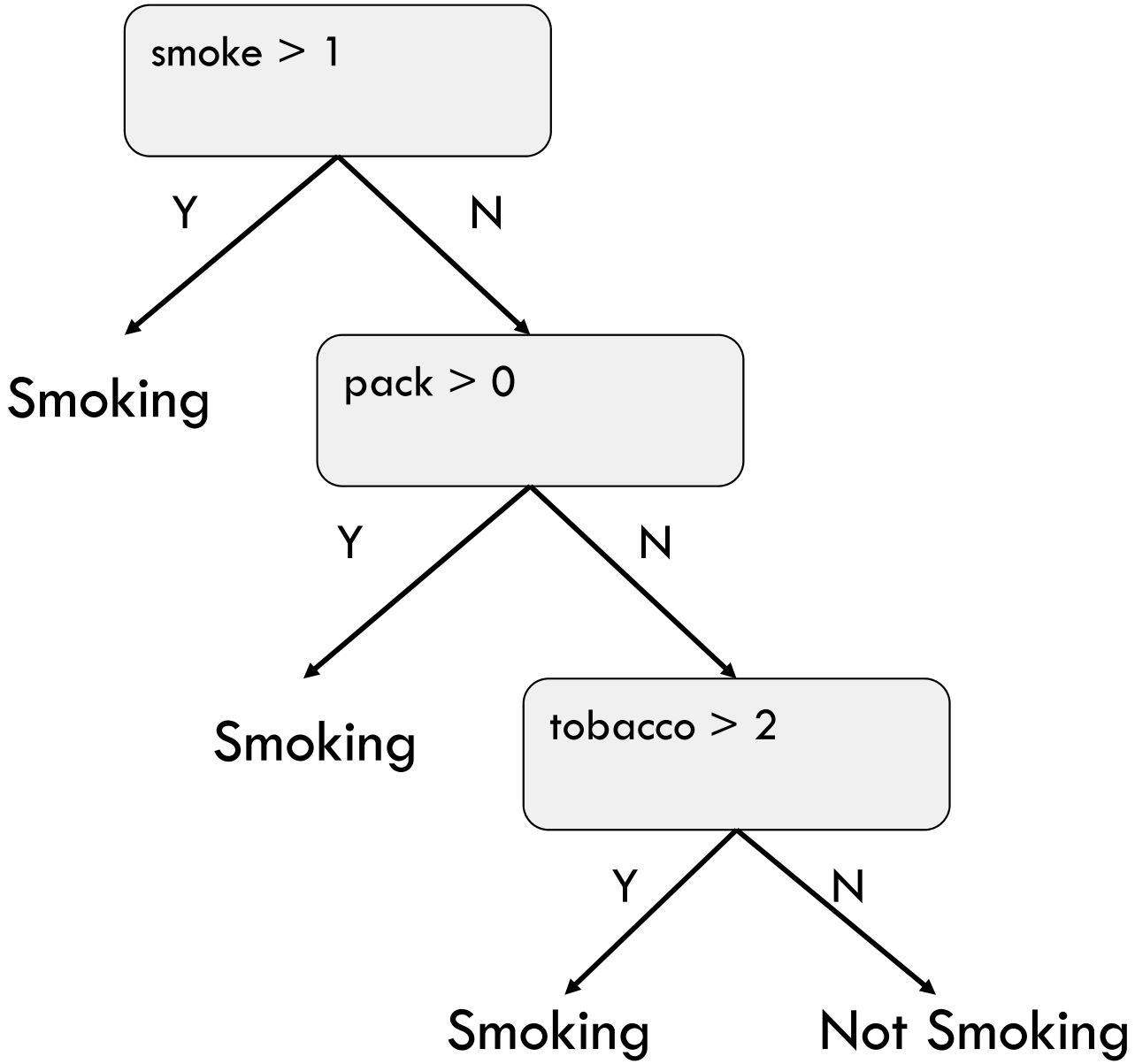
Statistical Text Mining Process



Step 3. Derive Patterns

- Terms and/or SVD dimensions used as inputs
- Classification algorithm
  - Common options
    - Naïve Bayes
    - Support Vector Machines (SVM)
    - Decision Trees
    - Logistic Regression
  - Empirically chosen based on prediction

Decision Tree



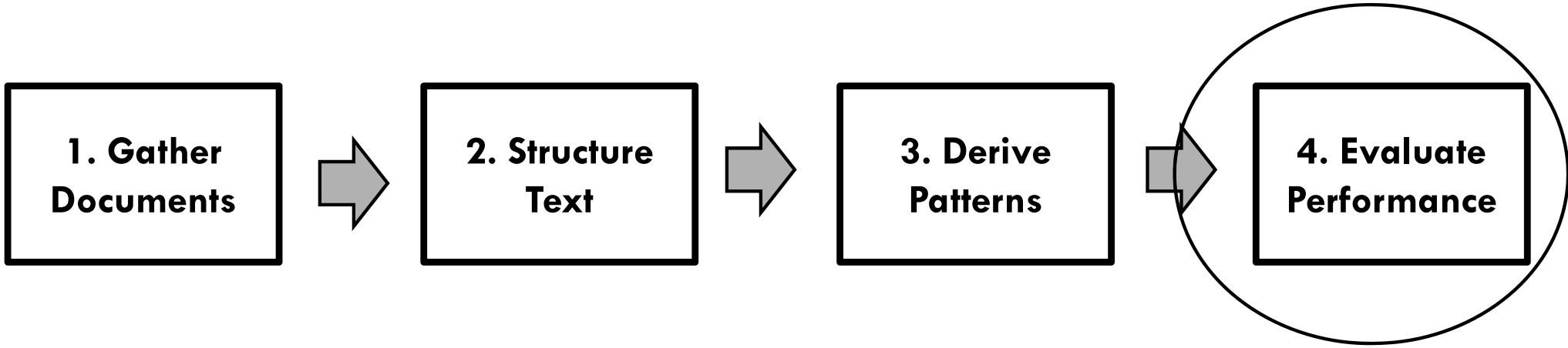
Logistic Regression

- Terms and/or SVD dimensions become variables in regression

$$\text{Smoking (Y or N)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$\text{Smoking (Y or N)} = 2.34 + 1.54(\text{smoke}) + 0.35(\text{pack}) + 0.09(\text{tobacco})$$

Statistical Text Mining Process



Step 4. Evaluate Performance

- 2x2 Contingency Table

		Reference Standard		Stats
		True	**False**	
Model Classification	**True**	TP	FP (Type I)	PPV (Precision) *TP/(TP+FP)*
	False	FN (Type II)	TN	NPV *TN/(TN+FN)*
Stats		Sensitivity (Recall) *TP/(TP+FN)*	Specificity *TN/(TN+FP)*	Accuracy *(TP+TN) / (TP+TN+FP+FN)*

Error Analysis

- Examine documents with incorrect classifications
  - False positives
  - False negatives
- Look for patterns / categories of error
  - Understand limitations of trained model
  - Make better informed decisions in STM process

STATISTICAL TEXT MINING SOFTWARE

STM Software

- As part of our HSR&D funded research
  - Identify open source STM product
  - Develop modules to facilitate use by HSR&D investigators

- Selected software
  - RapidMiner (rapid-i.com) – tool for data mining, STM, time series analysis, etc.

- Other software options
  - SAS Text Miner (proprietary)
  - IBM SPSS Modeler (proprietary)
  - Knime (open source)

VA RapidMiner Plugin

- Created operators to enhance STM capabilities of RapidMiner
  - Examples
    - Term-by-document matrix weighting
    - Term selection
    - Latent Semantic Analysis

RAPIDMINER DEMO

RapidMiner Demo

- Classification of 200 MEDLINE abstracts
  - 100 smoking / 100 not related to smoking

Questions

- Questions?

- Questions later?
  - James McCart – James.McCart@va.gov
  - Steve Luther – Steve.Luther@va.gov

1/12/2012	2:00pm	An Introductory Look at Statistical Text Mining for Health Services Researchers	Consortium for Healthcare Informatics Research	Luther, Stephen McCart, James

Q1. How do you split your documents into model building and testing sets?

Generally, the statistical text mining tool handles the splitting of documents into training and testing sets. The user typically specifies what percentage of documents to keep in the training set (for training/testing split) or how many folds to create (for X-fold cross validation). In addition, stratified sampling is generally used to keep the same proportion of positive and negative documents between the splits. In RapidMiner, the Split Validation operator is used for simple training and testing splits and the X-validation operator is used for X-fold cross validation.

Q2. What are the costs?

In general, the highest cost in statistical text mining projects come from building the reference standard – i.e., having a subject matter expert(s) read through each document and specify a label for the document. Other costs may include the personnel to perform statistical text mining and the cost of the statistical text mining program (or any custom modifications required to the program).

Q3. I am thinking that a perfect possible use for this would be to apply this when determining whether a fall is related to toileting, and searching for several key words, e.g. Toilet, Bathroom, Commode, etc. Am I correct in my thinking?

Yes, you are correct that statistical text mining can be used to determine the mechanism of a fall. (We are just finishing up a project that uses statistical text mining to look at various aspects of fall-related injuries.) Using your example, a reference standard would need to be created first, where documents would be labeled as a fall related to toileting or not (more than two labels are possible if you are interested in more than just toileting). Then a model would be built that would automatically determine which terms are most predictive (you would not need to pre-specify terms).

There might also be an advantage to combining statistical text mining with natural language processing (NLP). Since you would have a relatively straight forward set of terms you would look for, NLP might be used to find which falls were related to which mechanism.

Q4. Do you know of any shortcuts for getting CPRS Notes text data easily transferred into the STM software packages?

The best way at this point is probably to use the VINCI. We have also heard that folks at Ann Arbor VA have developed an algorithm to extract progress notes, but we don't have any direct experience with that method.

Q5. Can you add Data from Data Warehouse or SQL?

Structured data (data from a data warehouse or other similar types of databases) may be included in models with data derived from text. However, the structured and textual data may be at different levels

of analysis that make incorporating the data less than straightforward. For instance, statistical text mining is typically done at the document level of analysis. If the structured data is associated with a clinical event, then all documents for that event may need to be rolled up. We are currently examining different ways of incorporating structured and textual data.

Q6. Is there a spellchecker involved in any sort of cleaning, such as when you fix capitalization? so that FOOBAR is the same as FOOBARE?

Spellchecking can be done as a pre-processing step. However, a spellchecker is not currently available in RapidMiner. (I'm not sure whether the other statistical text mining programs have spellchecking capabilities or not.) Our research group does not generally spellcheck documents prior to statistical text mining for two reasons. (1) It can be difficult to automatically determine the correct word given a misspelled word. (2) Statistical text mining tends to perform well even in the presence of some misspelled words. However, a couple options for spellcheckers are listed below in case you're interested.

- HunSpell (http://hunspell.sourceforge.net/)
- gSpell (http://lexsrv3.nlm.nih.gov/LexSysGroup/Summary/gSpell.html).

Q7. In your projects, have you have dealt with sensitive information that needs to be de-identified? Was the approval process to use such data particularly difficult?

We have taken steps to de-identify documents in our studies using a home-grown program that searches for names, telephone numbers, etc. and also scrambles dates. The CHIR also has a group that has been working on a "best of breed" software package for this kind of work. I am not sure if their final product has been made available yet. I would check with the CHIR De-Identification project to see about the status of the software.

Q8. What's that in word count?

I'm not exactly sure what this question is asking about. If you would like to clarify the question please contact the presenters at James.McCart@va.gov and/or Steve.Luther@va.gov.

The word count (or term frequency) is one of the possible local weighting schemes that may be used in weighting the term-by-document matrix. Other common options include a log transformation of term frequency or binary.

Q9. I used Rapid miner for my research analyzing medical records. I used Clustering Method (K-means). But how do I prove that my approach gave an accuracy results??

A reference standard, where a subject matter expert(s) specifies the label for each document, will be needed to judge how well clustering performed. Once a reference standard exists a number of

evaluation measures may be used. For example, purity is a simple evaluation metric that measures accuracy of cluster assignment. Each cluster is assigned a class given the most frequent class found within the cluster. The total number of correctly assigned documents (within each cluster) is summed and then divided by the total number of all documents.

A description of purity and other evaluation measures for clustering can be found in chapter 16 section 3 of the book "Introduction to Information Retrieval" by Manning, Raghavan, and Schutze. The book is available online (and in bookstores) at http://nlp.stanford.edu/IR-book/.

Q10. Has the Consortium for Health Informatics had any collaboration with NSF?

Not that I am aware of. You might check with the CHIR staff in Salt Lake City and see f there are collaborations of which I am not aware.