Exploring interviewer coding of reasons for unit nonresponse:

An application of text analysis

Wendy Martinez and Lucilla Tan
Bureau of Labor Statistics

IFD&TC Conference
May 18, 2015

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

Disclaimer: Opinions expressed in this paper are those of the authors and do not reflect official policy of the U.S. Bureau of Labor Statistics.

www.bls.gov

Background

- **Data source:**

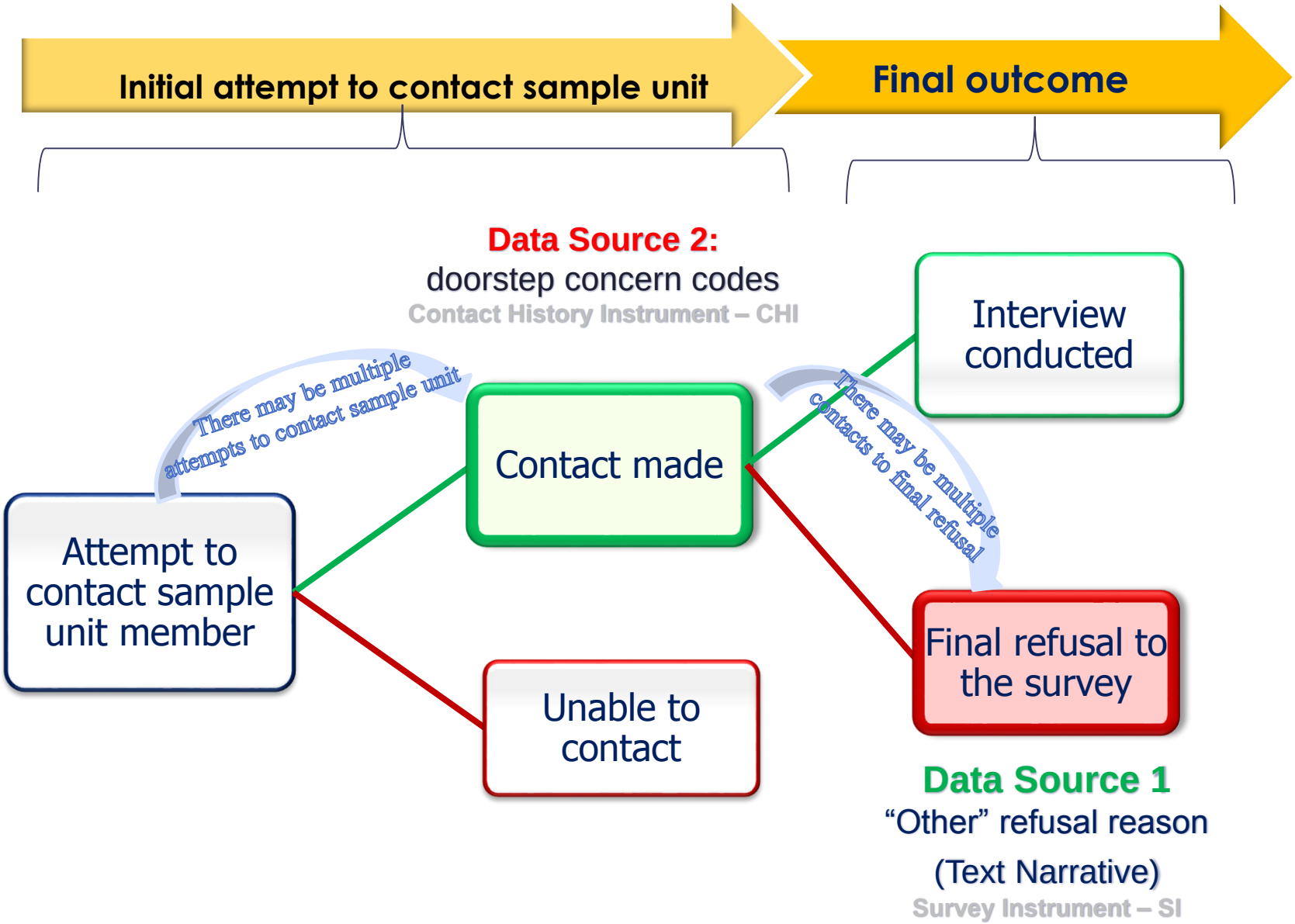
  The Consumer Expenditure Interview Survey (CE) – provides information on the buying habits of America's consumers, including data on expenditures, income, and demographics

  - For more details about the Consumer Expenditure program: http://www.bls.gov/cex

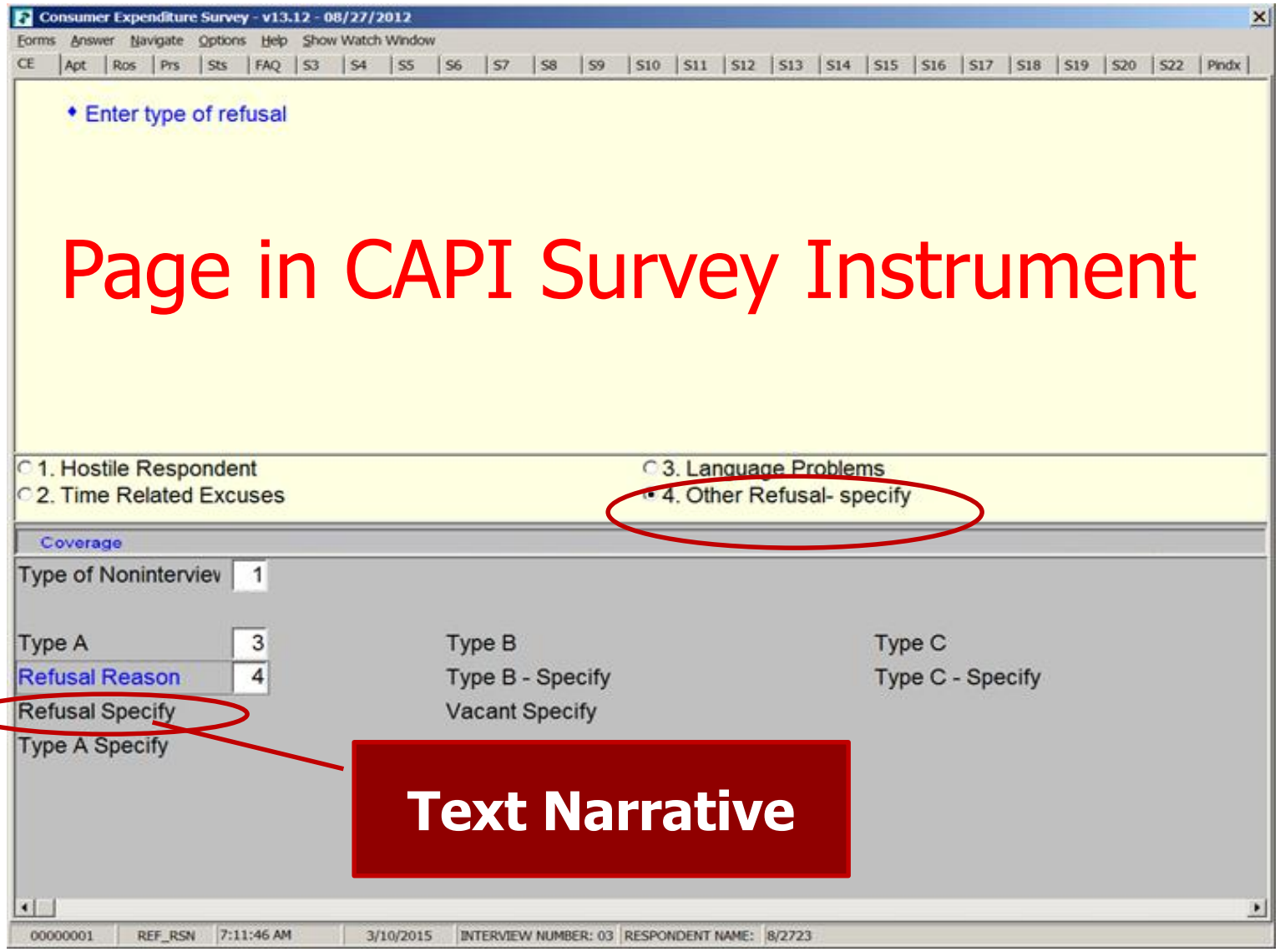
GOAL: Determine "Other" reasons for **final** non-response (Survey Instrument – SI text) and relate to reluctance at **initial** contact (Contact History Instrument – CHI)

BLS

Analysis Variables from 2 Instruments



Question: what are "Other" refusal reasons for survey participation among **contacted** sample units?



BLS

Study Sample

- Wave 1 sample units from CE collection April 2012 through March 2014
- 18,031 distinct sample units
- 25% were non-respondents
- 30% of non-respondents refused for 'other' reasons
- Only know reason for refusal through text analysis

BLS

Example of entries in the Refusal Reason text narrative

"DOESN'T DO SURVEYS". SOMEWHAT HOSTILE
"It's voluntary; I just don't want to do it."
"Special family situation"
"VOLUNTARY NO THANKS"
"just not interested"
"makes it a policy not to do such things"
100% Day
roomates and don't want bothered
old lady said dsnt wnt to participate
999999999999999999999999999999999999999
?? I'm closing this case for another FR
ALREADY DNE OTHER SURVEYS TOO INVASIVE
ANTI GOV
ATTORNEY TOLD THEM THEY DIDN'T HAVE TO DO IT
AVOIDANCE
AVOIDANCE, SILENT REFUSAL
Absolutely will not answer questions

BLS

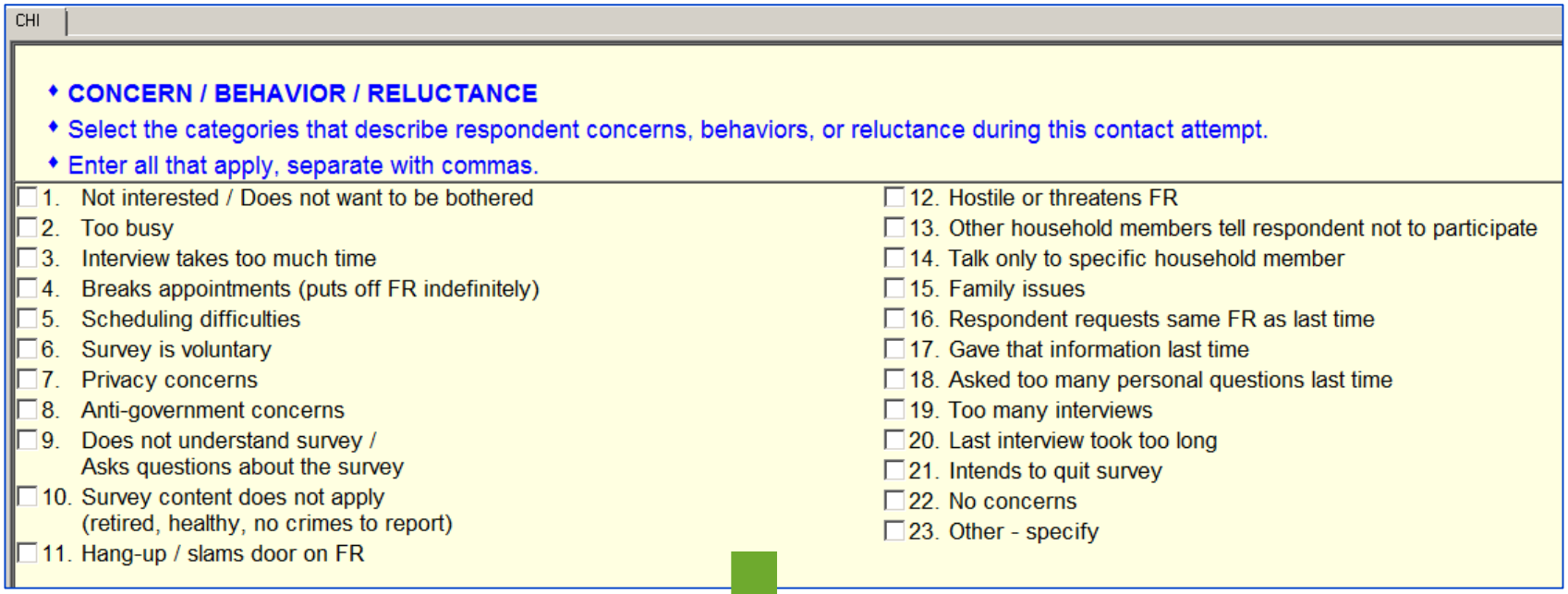
Highest Frequency Words

Most frequent words in the text narrative	
Highest Frequency (1 – 10)	**Highest Frequency (11 – 20)**
privacy	doesn
refusal	door
avoidance	government
silent	health
issues	voluntary
survey	concerns
participate	personal
refused	gov
not	govt
anti	family

BLS

"Doorstep concerns" from the Contact History Instrument (CHI)

Interviewers can report their observations of contacted sample unit member's reactions to the survey request in the CHI, as shown in the screenshot below:



ID # of CHI doorstep concern codes grouped to form theme	Doorstep concern theme (used in analysis)
1, 11, 12	**Not interested** / hostility
2, 3, 4, 5	Time
6, 7, 8, 9, 10	Survey **voluntary** / **privacy**
13, 14, 15	Gatekeeping
16, 17, 18, 19, 20, 21	Prior wave
23	Other

Any doorstep theme observed in the contact attempt history for a sample unit is flagged and rolled up into 1 record for the sample unit.

BLS

Pre-Process the Data

- Unstructured text from reason for refusal narrative is a "document."
- These are clustered – hopefully with similar reasons for refusal.
- Preprocessed text
  - Removed special characters
  - Converted to lower case
  - Removed stop words
- Size of corpus
  - 1,283 documents (descriptions of 'other' refusals)
  - Lexicon had 760 unique words

BLS

Exploratory Iterative Process

GOAL: Determine "Other" reasons for **final** non-response (SI text) and relate to reluctance at **initial** contact (CHI)

1. Initial clustering – **too noisy**
   - K-means
   - Agglomerative
2. Reducing dimensionality of the data
   - Nonlinear – ISOMAP
   - Singular value decomposition – SVD
   - Nonnegative matrix factorization
3. Cluster analysis
   - Same as above
   - Model-based clustering

- Software used – MATLAB

BLS

Analysis: outline of comparisons

CONTACT made with sample unit

Data source 2

Contact History Instrument: doorstep concern **CODES**

Sample unit REFUSES survey

Data source 1

Survey Instrument: Refusal Reason **text narrative**

S3. Connect CHI doorstep concern codes to refusal reasons in SI

S1. Cluster interviewer narratives and identify refusal reasons

S2. Identify major doorstep concern themes in clusters – using CHI

BLS

S1. Refusal Reasons from CAPI text narrative

Distribution of refusal reason by approach				
Cluster Topic – Refusal Reason (based on the **5 most frequently occurring words** in a cluster) **(n = 1,283)**	**Data dimensionality reduction & clustering method (K = 6)**			
	ISOMAP, Model-Based clustering	ISOMAP, K-Means clustering	SVD, K-Means clustering	NNMF, K-Means clustering
	Column percent distribution			
Firm refusal	4.5	4.4	10.1	10.1
Refused answer door	19.4			
Avoidance	9.5	10.9	5.7	6.4
Anti-government / voluntary	13.6	10.7	53.0	5.1
Privacy	29.9	56.3	12.2	5.6
Privacy / voluntary		7.2		6.8
Not interested	23.0		14.4	
Does not participate in surveys		10.5	4.5	65.9

BLS

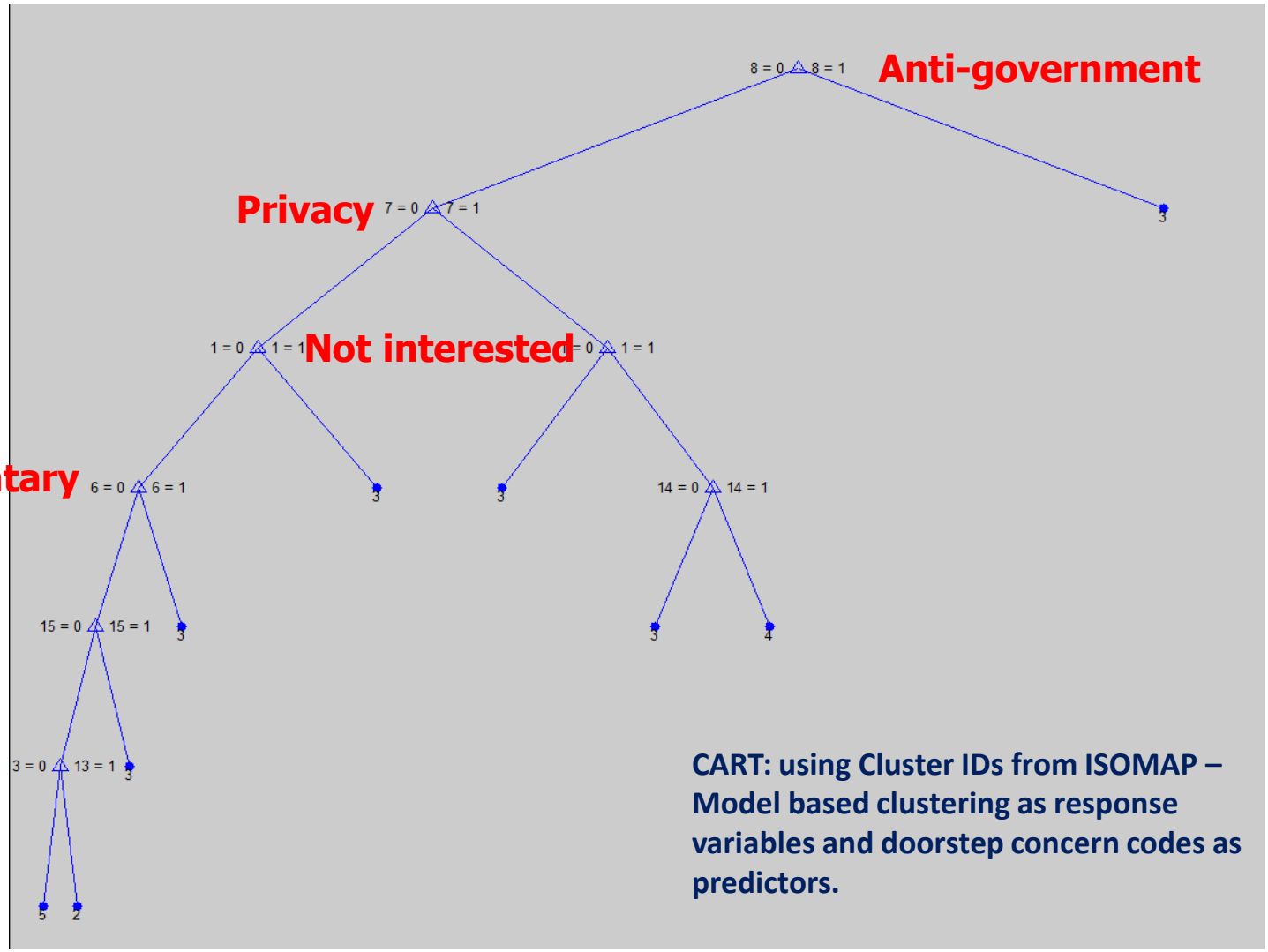
S2. Doorstep concern themes (CHI) observed for clusters

	Characterize clusters from ONE approach using CHI doorstep concern themes (clusters formed from SI text narrative)					
	1	2	3	4	5	6
No. of sample units in each cluster	58	175	384	249	122	295
Doorstep concern themes CHI	*Prevalence of themes (% observed among members of a cluster)*					
Not interested / hostility	**55.2**	72.0	62.8	**60.2**	35.2	**62.0**
Survey voluntary / privacy	51.7	**82.3**	**70.1**	55.4	35.2	61.0
Time	37.9	41.7	43.5	35.3	**44.3**	48.1
Gatekeeping	6.9	17.7	15.6	14.9	11.5	15.6
Prior wave	12.1	22.9	17.4	12.0	2.5	18.0
Other	25.9	26.3	27.1	32.5	20.5	29.8

- Among the 4 data dimensionality reduction – clustering methods, the ISOMAP-model based clustering resulted in relatively less unbalanced cluster sizes.
- More than 1 theme may be observed for a sample unit

BLS

S3. Doorstep concern codes (CHI) as predictors in CART



Top 3 levels of the tree show that the codes (analogous to demographics) most predictive of cluster membership were: ***8 anti-government, 7 privacy, 1 not interested, 6 voluntary.***

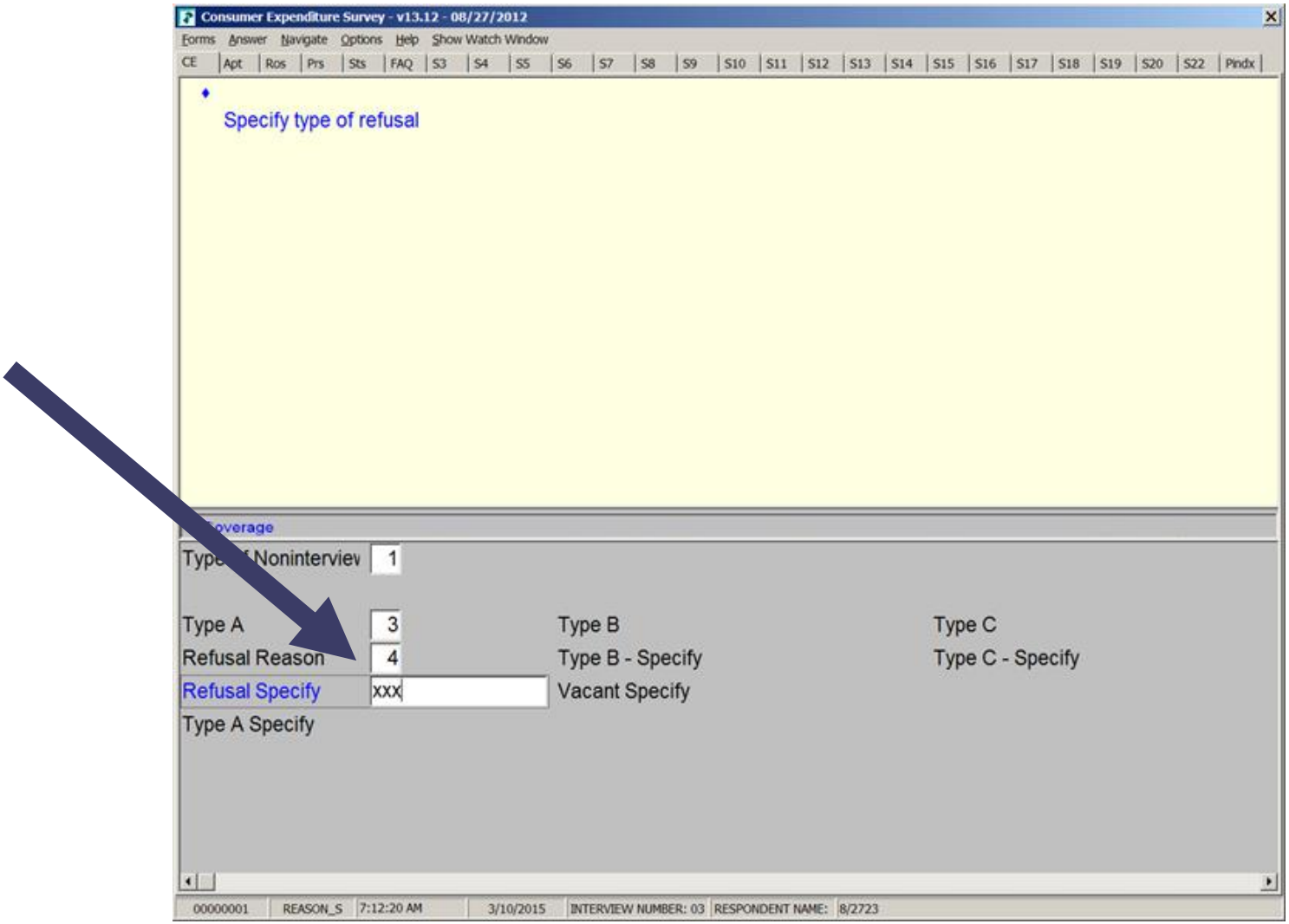
Limitations

1. Limited access to interviewer notes due to PII concerns
   a) No access to interviewer's case level notes
   b) No access to doorstep concern item "other-specify" description

2. Clustering method assigns a sample unit to membership in 1 unique cluster, but more than one doorstep concerns may be observed for a sample unit member

3. Text box for entering reason in SI is too small (usability perspective) resulting in short documents

Box for Text Narrative



Discussion/Application

"Other" refusal reasons consistently emerged from the SI text narrative: *not interested*, *privacy*, *anti-government*, and *voluntary* nature of the survey. This suggests:

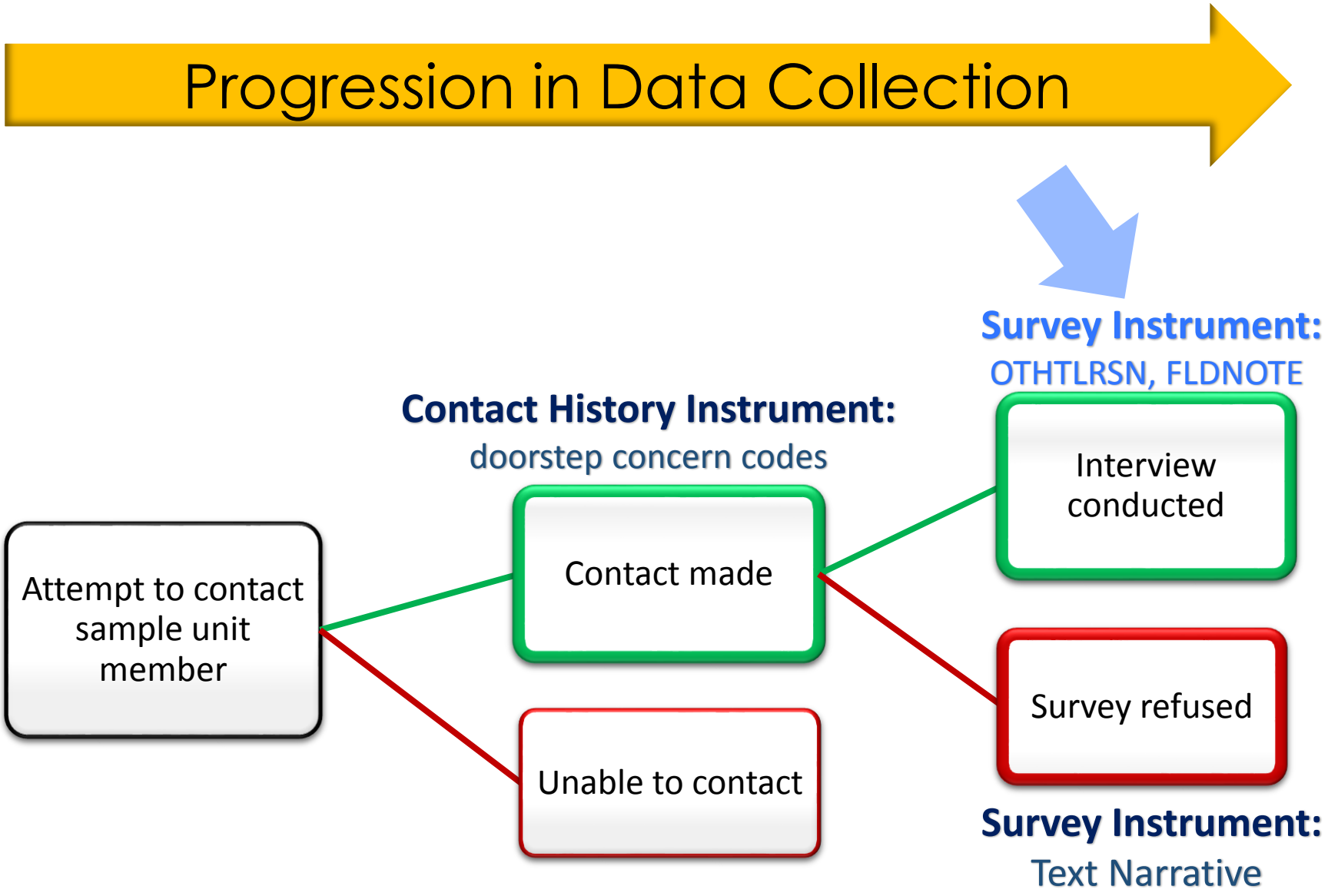
1. Interviewers are correctly filtering the types of refusal reasons to enter in the text narrative field.

2. There is an underlying structure in the text that can be used to enhance the SI instrument.
   - E.g. the response options for "reason for refusal" in the survey instrument can be expanded to include the additional pre-specified refusal categories ➔ saves interviewer data entry time; will facilitate analyses.

3. Mitigate non-response. Understand refusal reasons to better tailor information about the usefulness of government statistics and measures taken for privacy protection for sample units with these types of concerns.

BLS

Next Steps


Progression in Data Collection
Survey Instrument:
OTHTLRSN, FLDNOTE
Contact History Instrument:
doorstep concern codes
Attempt to contact sample unit member
Contact made
Unable to contact
Interview conducted
Survey refused
Survey Instrument:
Text Narrative

Contact Information

Wendy Martinez

Bureau of Labor Statistics
Office of Survey Methods Research

202-691-7400
martinez.wendy@bls.gov

BLS

Background

- Initial research question: *Exploring the use of interviewer notes from a later phase in the data collection process to corroborate pre-survey respondent attitudes observed at the survey request*
  - Change made to the current topic because of misunderstanding about the skip pattern of the first text field analyzed
- Current research question (next slide)

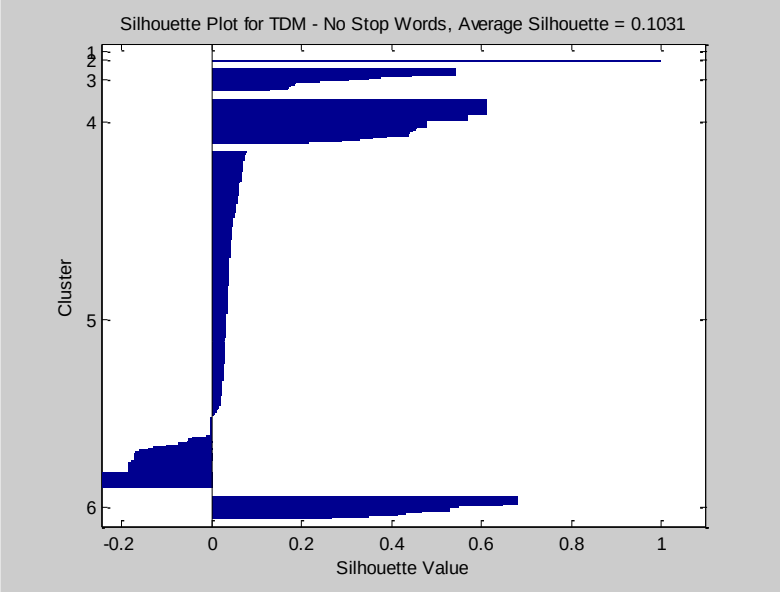
Encode the Text

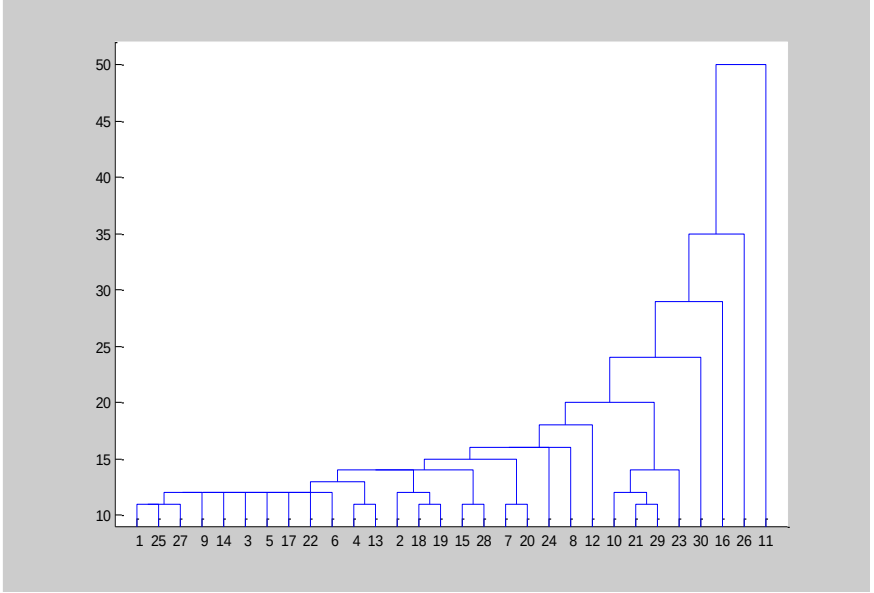
- The most common approach is the bag of words or term-document matrix.
- The rows correspond to words.
- The columns correspond to documents.
- The (i,j) -th entry in the matrix is the number of times the i -th word appears in the j -th document.
- These are the raw frequencies.

BLS

Initial Clustering

- Clustered without reducing the dimensionality
- Found the data to be too noisy





Reduce Dimensions

- Isomap: Nonlinear dimensionality reduction
  - Classical multidimensional scaling
  - Inputs are geodesic distances
- Nonnegative Matrix Factorization
  - Factors the term-document matrix
  - Factors are constrained to be nonnegative
  - Provides grouping (clusters)
  - Prespecify number of dimensions

BLS

Reduce Dimensions

- Singular value decomposition of term-document matrix

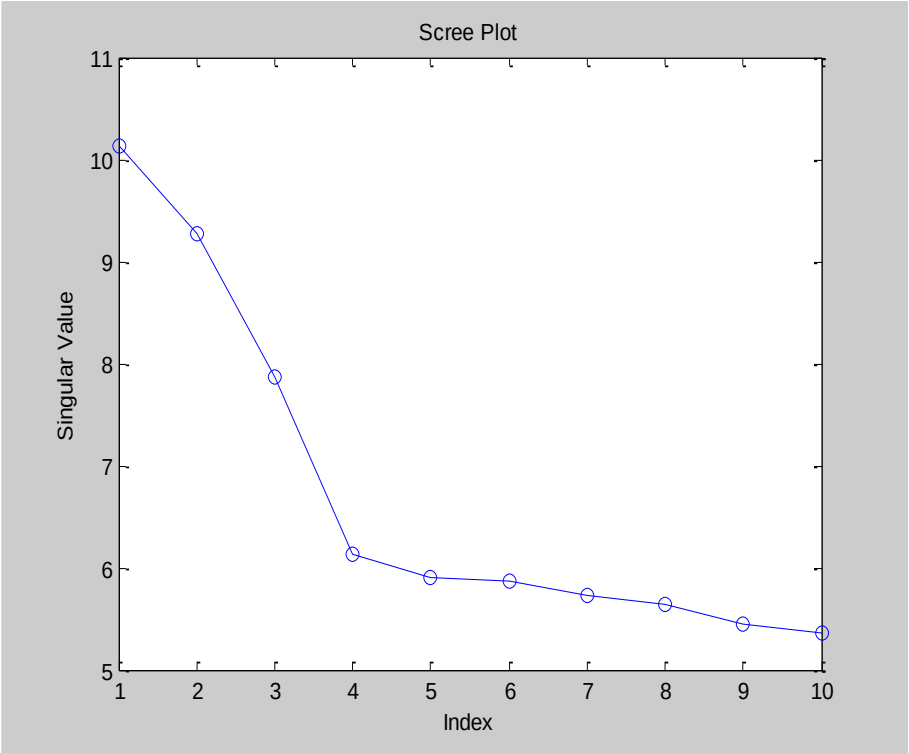
$$\mathbf{X} = \mathbf{TSD}^T$$

- Left singular vectors in **T** span the document space
- Right singular vectors in **D** span the word/term space
- Use matrix **D** to reduce dimensionality ~ Principal Component Analysis

BLS

Choosing the Number of Dimensions

- Use a scree plot
  - ISOMAP
  - SVD
- Look for 'elbow' in the curve
- Chose 4 dimensions
- NNMF – *a priori*



BLS

Cluster Documents

- *K*-Means
  - Specify the number of clusters *k*
  - Iteratively grouped with closest centroid
  - Tends to find spherical clusters
- Model-Based Clustering
  - Estimate a probability density function for cluster structure
  - Model is finite sum (mixture) of multivariate Gaussians
  - Each term is a cluster – very flexible structure
  - Provides estimate of number of groups

BLS