Modeling Topics in Survey Interviewer Notes

Wendy Martinez
Terrance Savitsky
US Bureau of Labor Statistics

SDSS 2018

The views expressed are those of the authors and do not necessarily reflect policies of the U.S. Bureau of Labor Statistics.

Origin

- Graduate courses in
  - Computational statistics
  - Exploratory data analysis
- Other Ed Wegman students
  - Jeffrey Solka – finite mixture models
  - Angel Martinez – text analysis
- Builds on prior work with Lucilla Tan

Major Points of Analysis

- Use two data sources
  - Sample unit behavior
  - Text describing reason for refusal
- Use two types of text encodings
  - Term-document matrix
  - Bigram proximity matrices
- Cluster text using
  - Model-based clustering
  - Bayes clustering
- Find important concerns in clusters using classification trees
  - Cluster IDs are ‘class labels’
  - Coded behaviors are ‘features’

Background – CE

- **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.

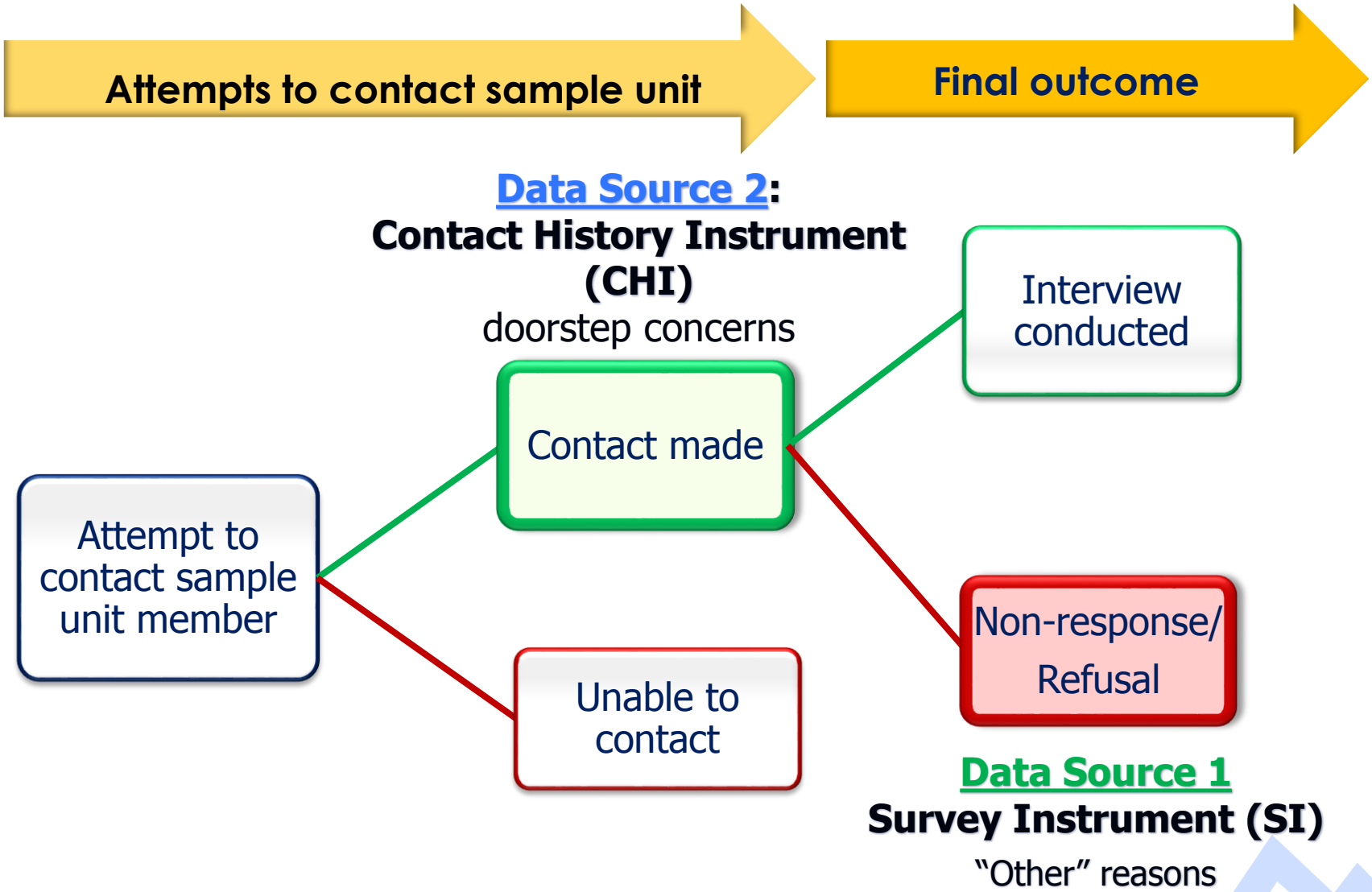
  - For more details about the Consumer Expenditure program: http://www.bls.gov/cex

GOAL: Associate a sample unit's sentiment (doorstep concerns from Contact History Instrument) regarding the survey with the reasons for non-response (Survey Instrument – SI)

Study Sample

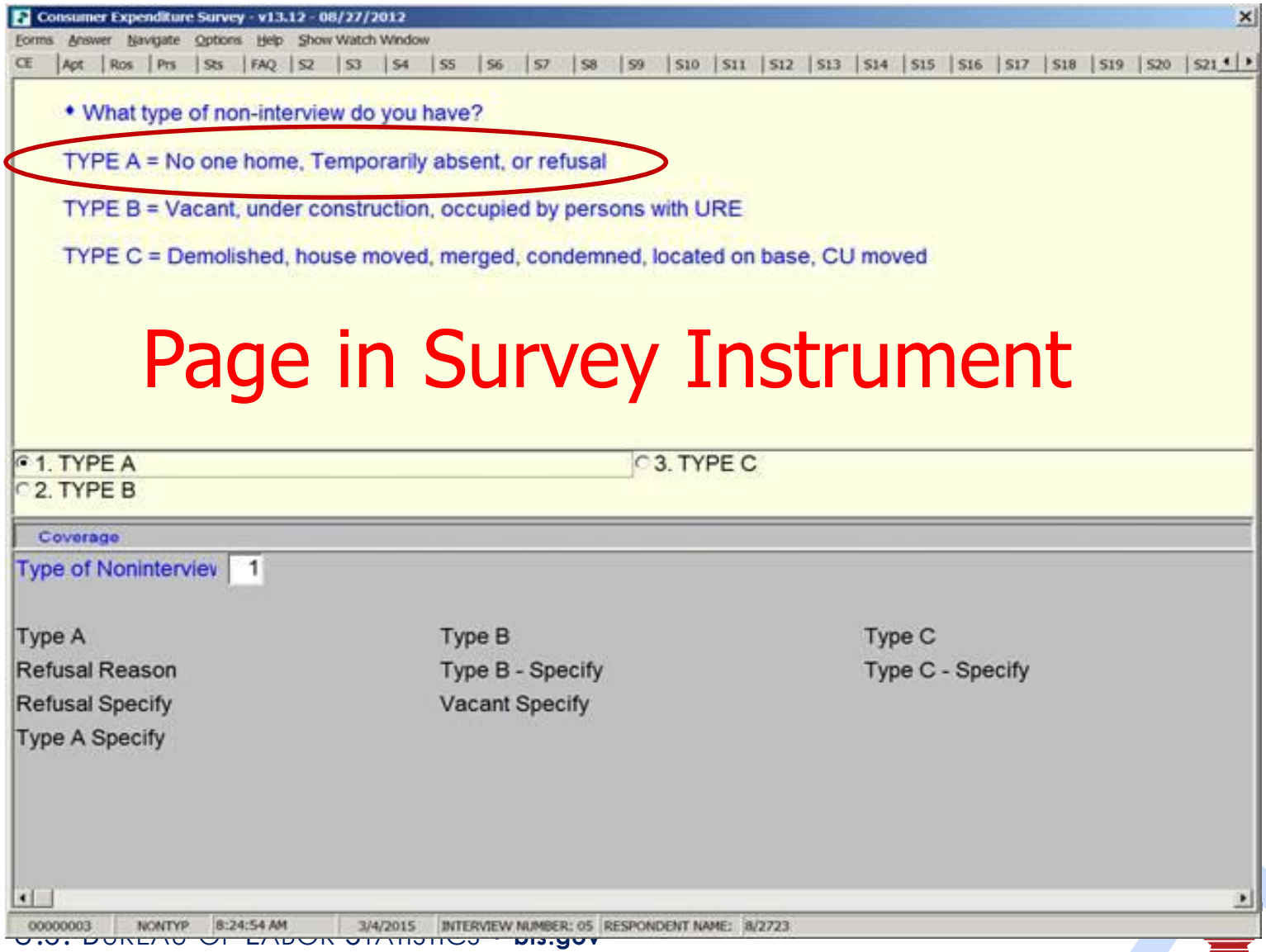
- Wave 1 sample units from CE collection April 2012 through March 2014
- 18,031 distinct sample units
- 25% were non-respondents
  - 30% of these refused for Other reasons
- Reasons not captured by codes in SI
- Only know reason through text analysis

Data Sources – 2 Instruments



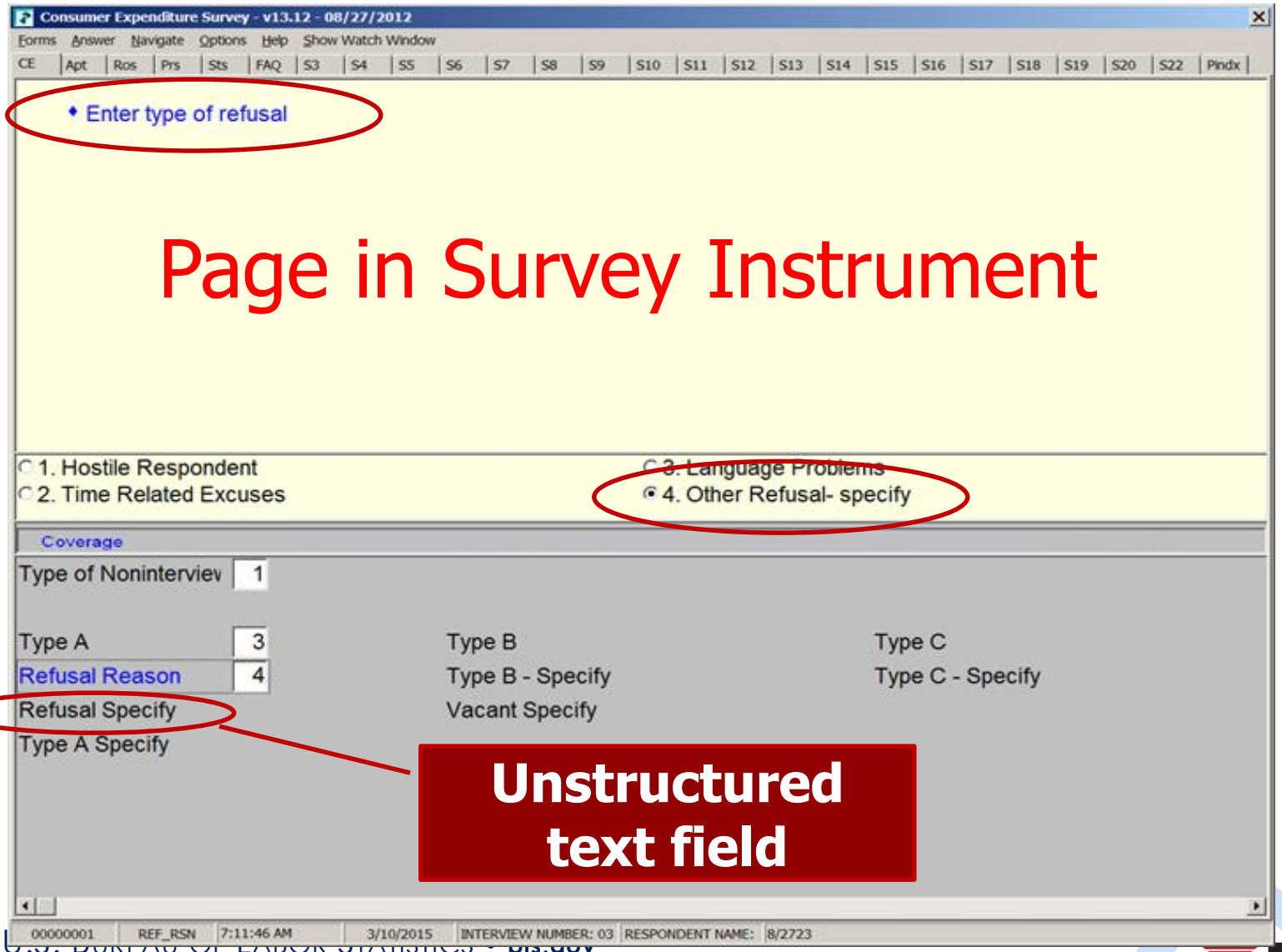
Survey Nonresponse Inputs

Data Source 1 (SI)



"Other Refusal" Reason – A Document

Data Source 1 (SI)



Examples of text documents from Data Source 1

"DOESNT DO SURVEYS". SOMEWHAT HOSTILE
"Its voluntary; I just dont 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 dont want bothered
old lady said dsnt wnt to participate
99999999999999999999999999999999999999999
?? Im closing this case for another FR
ALREADY DNE OTHER SURVEYS TOO INVASIVE
ANTI GOV
ATTORNEY TOLD THEM THEY DIDNT HAVE TO DO IT
AVOIDANCE
AVOIDANCE, SILENT REFUSAL
Absolutely will not answer questions

Highest Frequency Words
Refusal Corpus (SI)

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

"Doorstep concern" indicators from Data Source 2 (CHI)

- Interviewers report observations of sample unit reactions to the survey request.
- Associate concern codes with refusal reasons
- CHI revised after 2013 data collections – fewer items.

CHI

♦ **CONCERN / BEHAVIOR / RELUCTANCE**

♦ Select the categories that describe respondent concerns, behaviors, or reluctance during this contact attempt.

♦ Enter all that apply, separate with commas.

☐ 1. Not interested / Does not want to be bothered	☐ 12. Hostile or threatens FR
☐ 2. Too busy	☐ 13. Other household members tell respondent not to participate
☐ 3. Interview takes too much time	☐ 14. Talk only to specific household member
☐ 4. Breaks appointments (puts off FR indefinitely)	☐ 15. Family issues
☐ 5. Scheduling difficulties	☐ 16. Respondent requests same FR as last time
☐ 6. Survey is voluntary	☐ 17. Gave that information last time
☐ 7. Privacy concerns	☐ 18. Asked too many personal questions last time
☐ 8. Anti-government concerns	☐ 19. Too many interviews
☐ 9. Does not understand survey / Asks questions about the survey	☐ 20. Last interview took too long
☐ 10. Survey content does not apply (retired, healthy, no crimes to report)	☐ 21. Intends to quit survey
☐ 11. Hang-up / slams door on FR	☐ 22. No concerns
	☐ 23. Other - specify

Process the Text

- Used MATLAB and R
- Text narrative from a non-responding sample unit is a “document.”
- Preprocessed text
  - Removed special characters and stop words
  - Converted to lower case
- Size of corpus
  - 1,283 documents (*n*)
  - 760 unique words (*p*)

Exploratory Process

1. Encode documents using raw frequencies
    1. TDM – Term-document Matrix
    2. BPM – Bigram Proximity Matrix
2. Reduce dimensionality–ISOMAP – nonlinear approach
    - Chose $d = 4$
    - Used cosine distance
3. Conduct cluster analysis
    1. Model-based Clustering
    2. Bayes Clustering
4. Associate clusters of interviewer notes (refusal reasons) with doorstep concerns

Encode the Text – TDM

- The most common approach is the bag of words or term-document matrix (TDM).
- 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.

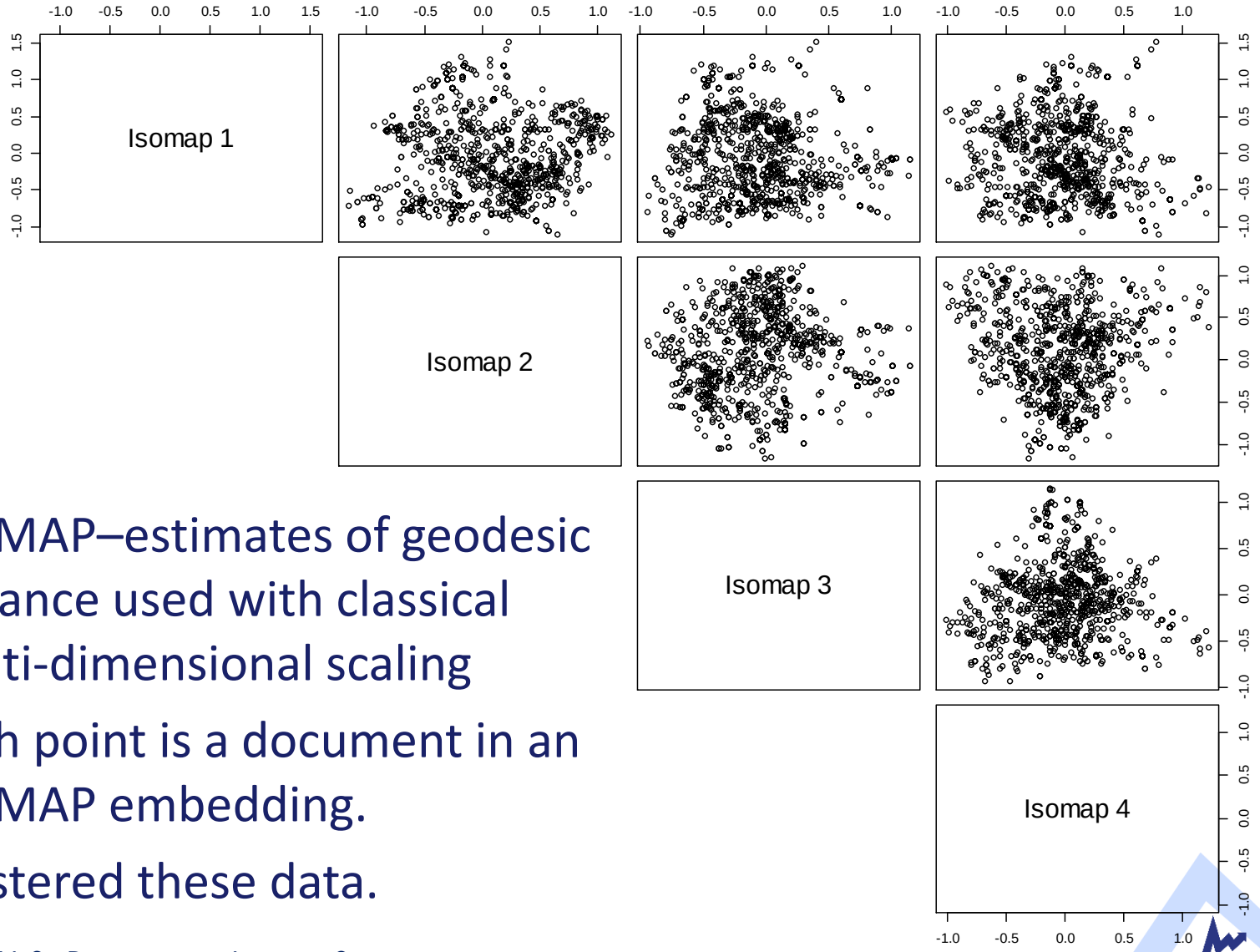
Encode the Text – BPM

- TDM – each document coded as a vector
- Bigram Proximity Matrix (BPM) – each document coded as a matrix
- The rows and columns in BPM_k (k-th document) correspond to words.
- The (i,j) -th entry in the matrix BPM_k is the number of times the i -th word appears before the j -th word.
- BPM_k is reshaped as a row in the data matrix.

The Data

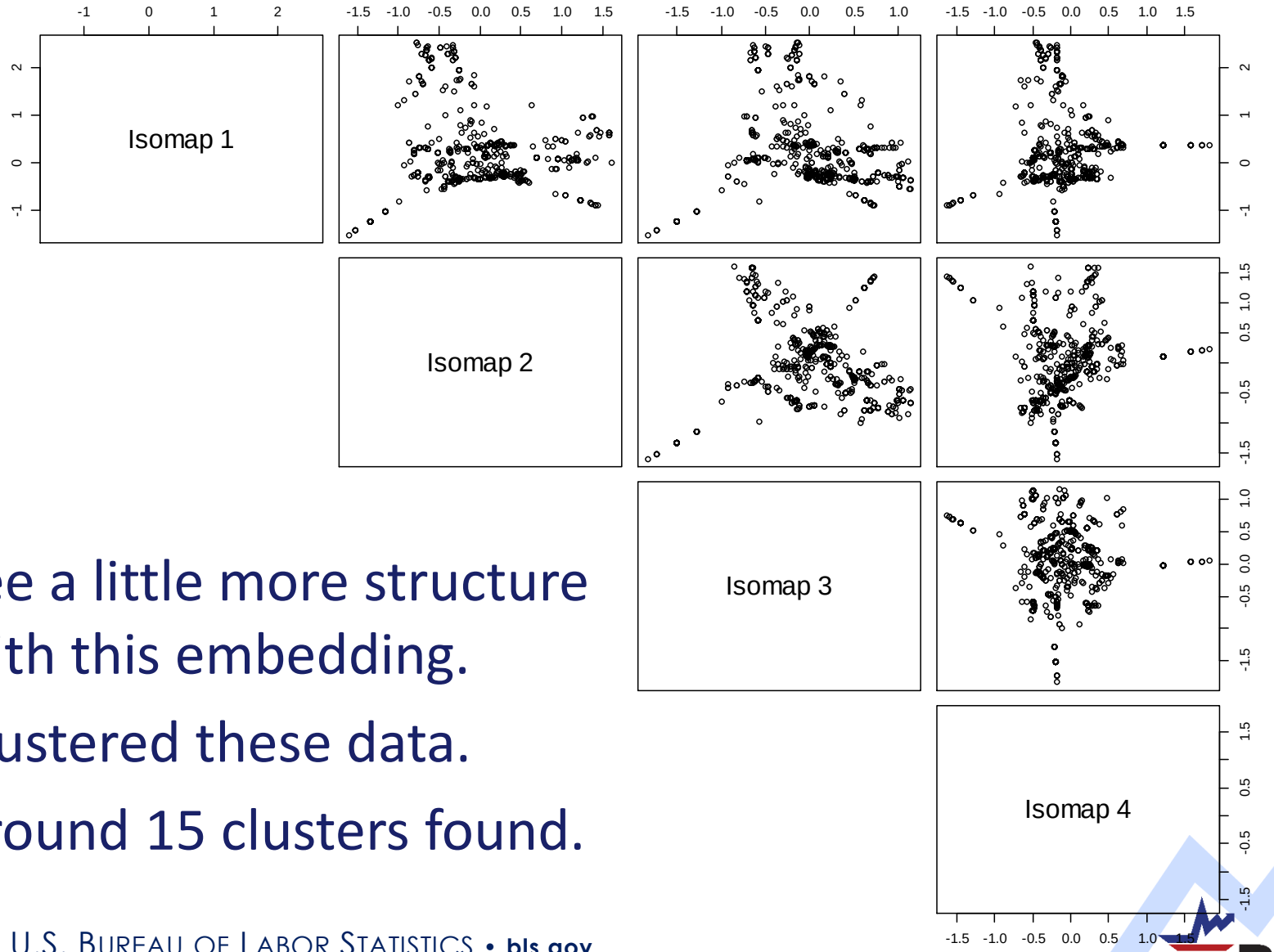
- Interviewer note in survey instrument is a document.
- Recall the size of corpus:
  - 1,283 documents (*n*)
  - 760 unique words (*p*)
- Size of data matrix using TDM is 1,283 x 760
- Size of data matrix using BPM encoding is 1,283 x 579,121
  - The BPM uses the period for all end of sentence punctuation.
  - The period is counted as a word.

ISOMAP Dimensions for TDM Encoding



- ISOMAP–estimates of geodesic distance used with classical multi-dimensional scaling
- Each point is a document in an ISOMAP embedding.
- Clustered these data.

ISOMAP Dimensions for BPM Encoding



- See a little more structure with this embedding.
- Clustered these data.
- Around 15 clusters found.

Cluster Analysis

- 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
- Bayes Clustering:
  - Limit of a Dirichlet process (DP) model as the noise variance contracts on zero
  - Converts posterior distribution to penalized optimization
  - Use Carlinski-Harabaz statistic to select penalty parameter (in turn determines number of clusters)
  - Connects DP to *k*-means

Connect Clusters with Concerns

- Cluster ID for each narrative of non-response
- Construct classification trees
  - Use cluster IDs (SI) as class labels
  - Use doorstep concerns (CHI) as features
  - Variable chosen to 'best' split into subsets
  - Indication of 'importance'

Feature/Predictor

Contact History:

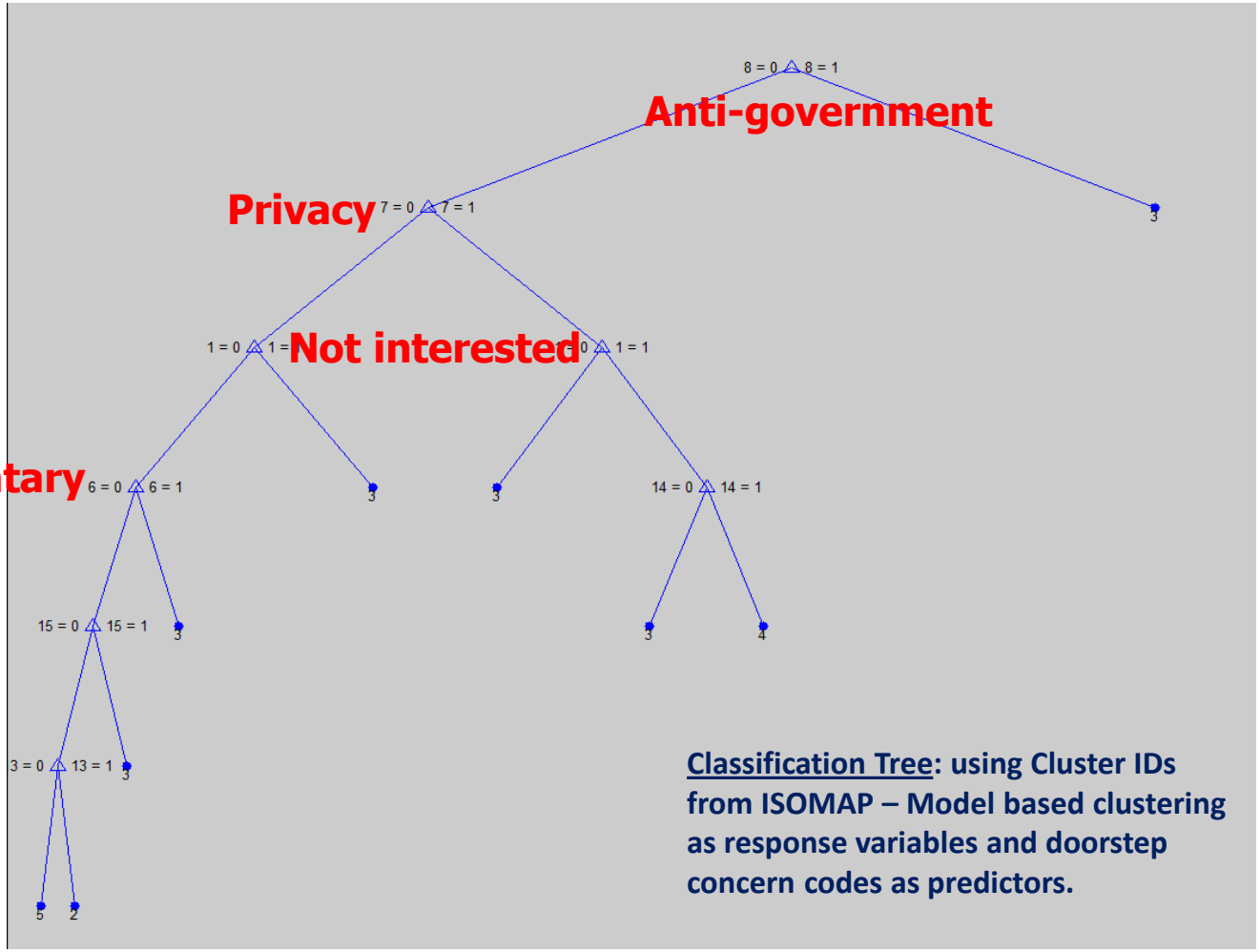
Doorstep concern codes

Class/Response

Survey Instrument:

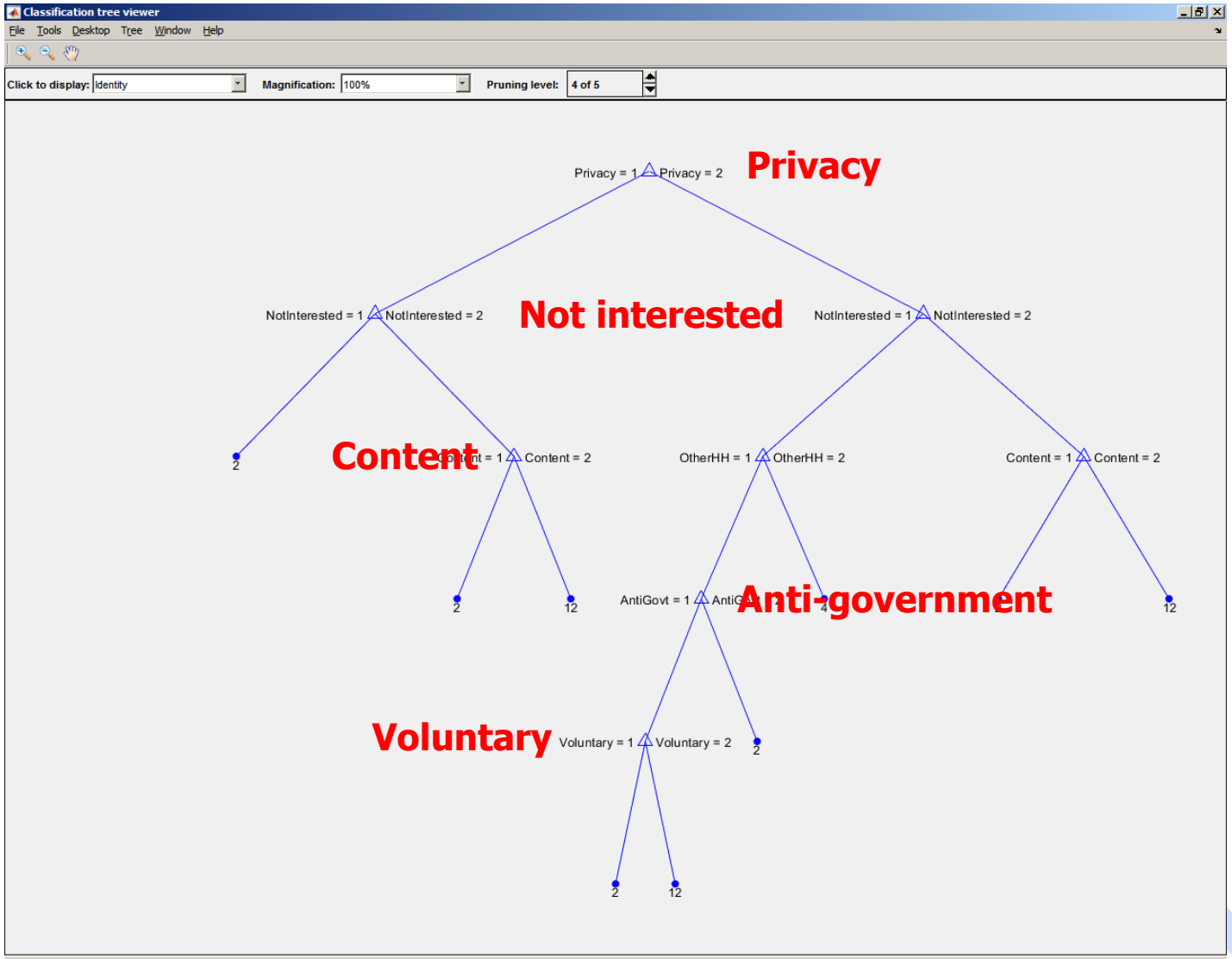
Cluster ID for text narrative

Model-Based Clustering using TDM Data Matrix

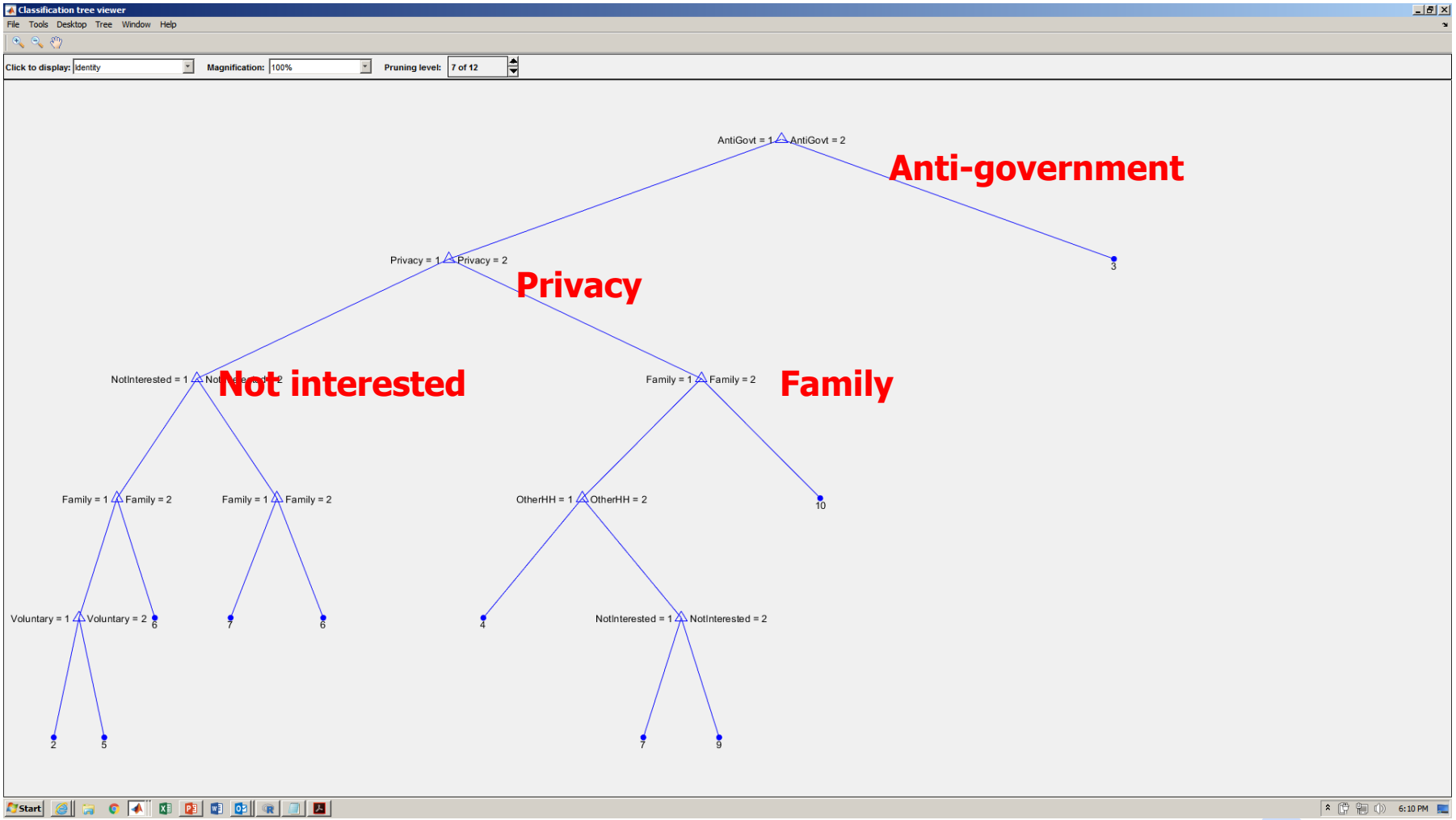


Classification Tree: using Cluster IDs from ISOMAP – Model based clustering as response variables and doorstep concern codes as predictors.

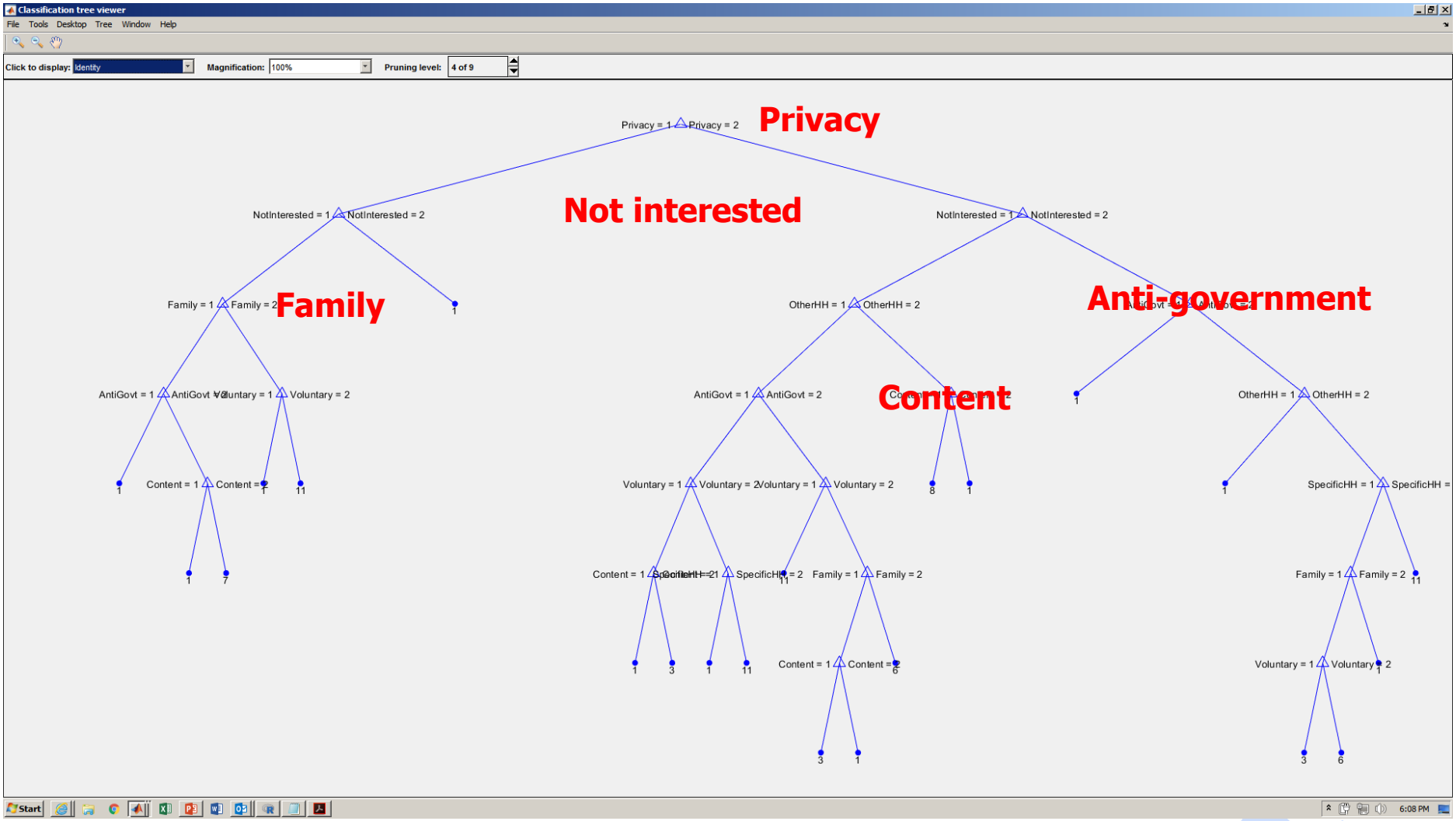
Model-Based Clustering using BPM Data Matrix



Bayes Clustering using TDM Data Matrix



Bayes Clustering using BPM Data Matrix



Recap

- Used two types of encodings
  - Term-document matrix
  - Bigram proximity matrices – captures some word order
- Explored two types of cluster approaches – both estimate number of clusters
- Model-based clustering
  - Flexible clusters
  - Not appropriate for high-dimensional data
- Bayes clustering
  - Similar to *k*-means clustering – looks for spherical clusters
  - Can be used with high-dimensional data
- Associated clusters with sample unit behavior

Discussion

- Compare cluster approaches – MBC and Bayes
  - Similar estimates on the number of clusters ~ 15
  - Same major concerns – Privacy, anti-government, not interested, voluntary
  - Bayes – different concern – Family
- Compare encodings – TDM and BPM
  - Some similar concerns
  - BPM uncovered different concern – Survey content

Application

- Important reasons for nonresponse are not captured by the existing codes – enhance the survey instrument.
- Missing these reasons could adversely affect non-response bias analyses.
- Understand refusal reasons and sentiment to better tailor information about the usefulness of government statistics and measures taken for privacy protection.
- Use information from text, doorstep concerns, and other variables to estimate propensity to respond.

References

- Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Boca Raton, FL: CRC Press.
- Solka, J. 1995. *Matching Model Information Content to Data Information*, PhD Dissertation, George Mason University.
- Tenenbaum, de Silva, & Langford. 2000. "A global geometric framework for nonlinear dimensionality reduction," *Science*, **290**:2318-2323.
- Martinez, A., 2002. *A Framework for the Representation of Semantics*, PhD Dissertation, George Mason University.
- Fraley & Raftery, 2002. "Model-based clustering, discriminant analysis, and density estimation: MCLUST," *Journal of the American Statistical Association*, **97**:611-631.
- MBC: https://www.stat.Washington,edu/rafter/Research/mbc.html
- Martinez W. and L. Tan, "Categorizing sentiment using unstructured text," Joint Statistical Meetings, 2015.
- Savitsky, T.D. 2016. "Scalable Approximate Bayesian Inference for Outlier Detection under Informative Sampling," *Journal of Machine Learning Research*, **17:**1-49.

Contact Information

Wendy Martinez

Bureau of Labor Statistics
Office of Survey Methods Research

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

Limitations

1. Limited access to interviewer notes due to PII concerns
   a) No access to interviewers 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