Matrix Decompositions: A Powerful Tool for Data-Driven Topic Modeling in Federal Surveys

Irina Belyaeva, Ph.D

Chief Architect, AI and Statistical Applications
Research and Methodology
U.S Census Bureau

October 22, 2024

2024 Federal Committee On Statistical Methodology Conference
College Park, MD

United States Census Bureau | U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Outline

Federal Surveys as Essential Statistical Products

Federal surveys, like the U.S. Census American Community Survey (ACS) are *crucial statistical products*

Importance of Federal Surveys as Statistical Products

- Comprehensive Data Collection

✓ Gather large-scale data on demographics, economics, housing, and more

- Inform Public Policy

✓ Provide crucial insights that shape national and local policy decisions

- Academic and Social Research

✓ Key resources for researchers studying societal trends and behaviors

Purpose

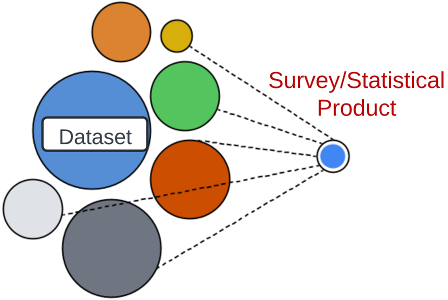
- To aid in *informed decision-making* and policy development across multiple sectors

United States Census Bureau | U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Why is Topic Modeling Important for Federal Surveys?

Federal surveys generate complex, *high-dimensional* statistical products, making *pattern (theme)* discovery difficult

Challenges in Statistical Products Complexity



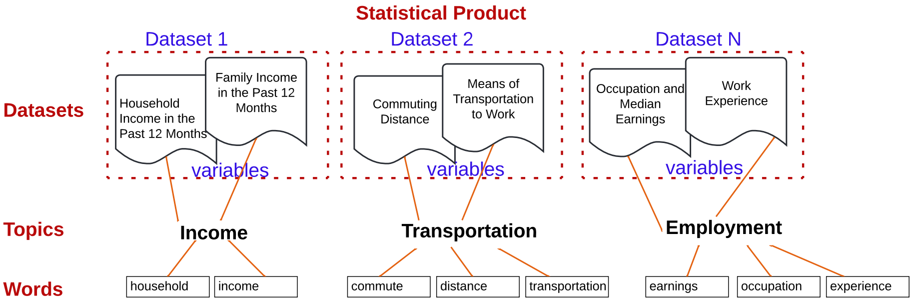
- Variety of statistical variables
- Volume of statistical variables
- Complexity of *manual* annotation and ontology creation

Outline

Role of Data-Driven Topic Modeling

Topic modeling uncovers latent patterns within statistical data products by grouping related variables

Concept of Topic Modeling



Role of Data-Driven Topic Modeling

Topic modeling uncovers latent patterns within statistical data products by grouping related variables

Concept of Topic Modeling

Efficient Data Analysis

- Data-driven annotation of large-scale surveys

Benefits

- Improves surveys interpretability
- Identifies key patterns that can inform policy-making
- Improves statistical products organization and retrieval

Role of Data-Driven Topic Modeling

Topic modeling uncovers latent patterns within statistical data products by grouping related variables

Concept of Topic Modeling

Efficient Data Analysis

- Data-driven annotation of large-scale surveys

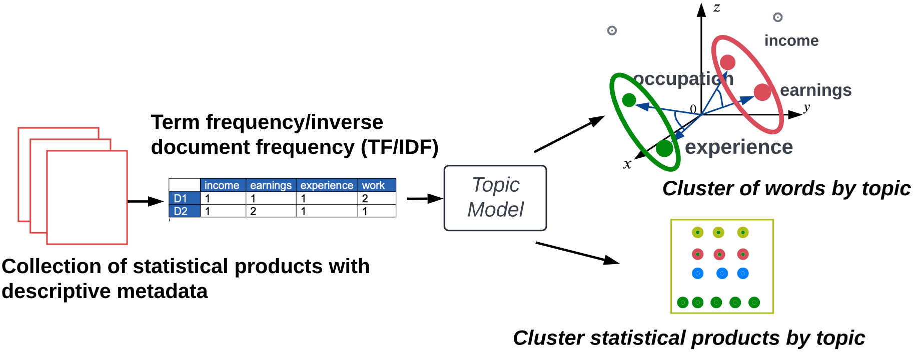
Benefits

- Improves surveys interpretability
- Identifies key patterns that can inform policy-making
- Improves statistical products organization and retrieval

Concept of Topic Modeling

Topic modeling uncovers latent patterns within statistical data products by grouping related variables

Concept of Topic Modeling



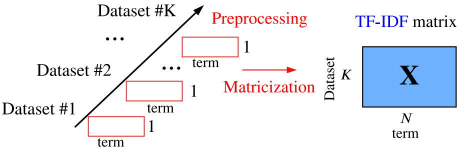
	income	earnings	experience	work
D1	1	1	1	2
D2	1	2	1	1

Outline

Topic Modeling via Non-Negative Matrix Factorization

$\mathbf{X} \in \mathbb{R}^{K \times N}$

Goal: Estimate common topics across group of datasets to learn typical themes and patterns



Bag of Words Model

- Metadata Tokenization
- Count word frequencies
- Numerical Encoding
- Creation of Term/Frequency Inverse Document Frequency Matrix

Term Frequency (TF)/Inverse Document Frequency (IDF)

TF-IDF

$$\text{TF-IDF}(w, d, D) = \text{TF}(w, d) \times \text{IDF}(d, D)$$

$$\text{TF-IDF}(w, d, D) = \frac{f_{w,d}}{\sum_{w' \in d} f_{w',d}}$$

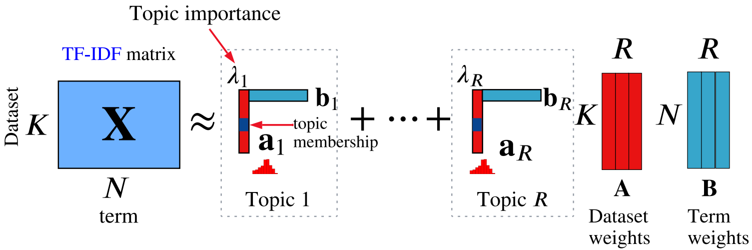
$$\text{IDF}(d, D) = \log \frac{|D|}{|d : d \in D \text{ and } w \in d|}$$

w word/term or token, d a metadata for a single dataset, D the entire set of datasets

Topic Modeling via Non-Negative Matrix Factorization

$\mathbf{X} \in \mathbb{R}^{K \times N}$

Goal: Estimate common topics across group of datasets to learn typical themes and patterns



Distinct Topic Pattern

$$\mathbf{X}_r = \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \in \mathbb{R}^{K \times N}$$

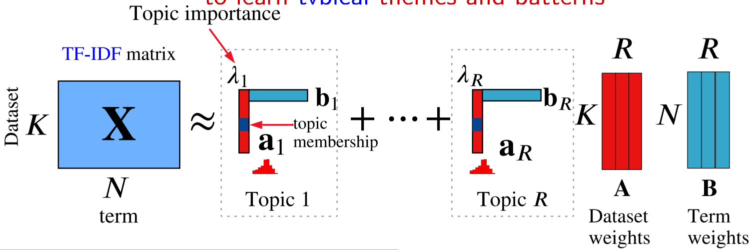
K # of datasets, N # of terms/words, R # of latent topics components
$\mathbf{a}_r \in \mathbb{R}^K$ dataset topic weights, $\mathbf{b}_r \in \mathbb{R}^N$ term topic weights, λ_r topic scale factor

United States Census Bureau
U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Topic Modeling via Non-Negative Matrix Factorization

$\mathbf{X} \in \mathbb{R}^{K \times N}$

Goal: Estimate common topics across group of datasets to learn typical themes and patterns



Non-Negative Matrix Factorization (NMF)

$$\mathbf{X} \approx \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r$$

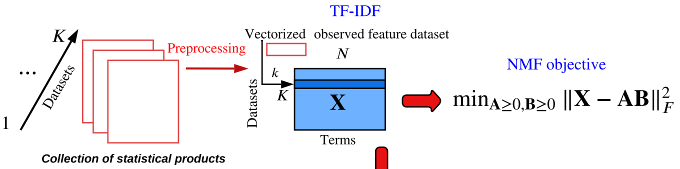
s.t $\|\mathbf{a}_r\|_2 = \|\mathbf{b}_r\|_2 = 1,\ \mathbf{a}_r \geq 0, \mathbf{b}_r \geq 0.$

- Feature (TF-IDF) matrix is simultaneously factorized into Dataset and Term topic components by fitting the NMF model

A Dataset Factors, **B** Term Factors

Topics Identification Outline

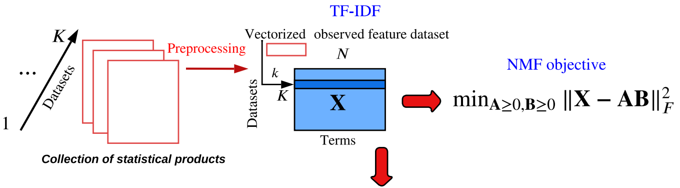
Goal: Identify *key* topic patterns across multiple datasets to uncover high-level themes and patterns within a statistical product



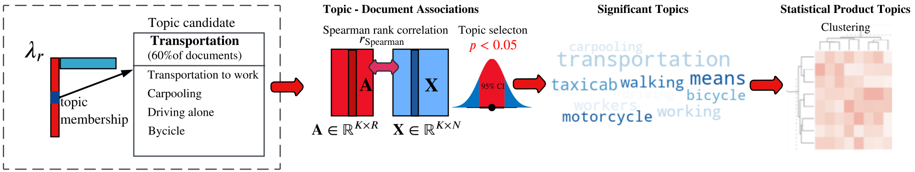
Topics Selection

Topics Identification Outline

Goal: Identify *key* topic patterns across multiple datasets to uncover high-level themes and patterns within a statistical product



Topics Selection



Outline

1. Background/Motivation
2. Introduction to Topic Modeling
3. Non-Negative Matrix Factorization for Topic Modeling
4. Demonstration of Topic Modeling: American Community Survey
5. Discussion

American Community Survey: Key Socioeconomic Patterns

Extracted topics *reveal* broad socioeconomic patterns in American Community Survey (ACS)

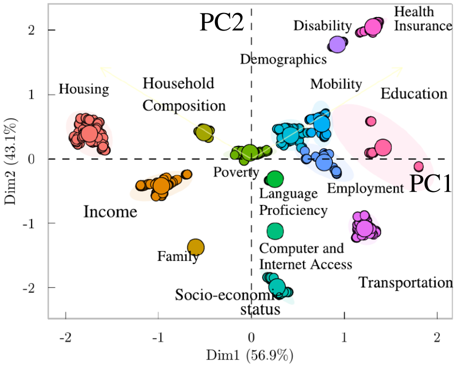
Topic embeddings

- 1600 ACS's datasets were used for topics extraction
- 100 topics were extracted via NMF

American Community Survey: Key Socioeconomic Patterns

Extracted topics *reveal* broad socioeconomic patterns in American Community Survey (ACS)

Topic embeddings



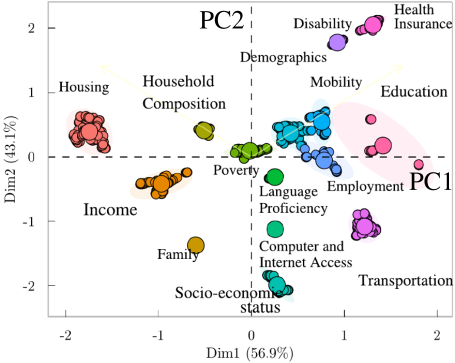
What are the main themes of the ACS product?

- Demographic and socioeconomic aspects

American Community Survey: Key Socioeconomic Patterns

Extracted topics *reveal* broad socioeconomic patterns in American Community Survey (ACS)

Topic embeddings



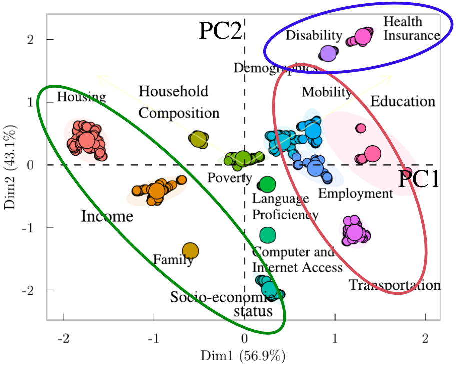
What are the main themes of the ACS product?

- Demographic and socioeconomic aspects

American Community Survey: Key Socioeconomic Patterns

Extracted topics *reveal* broad socioeconomic patterns in American Community Survey (ACS)

Topic embeddings



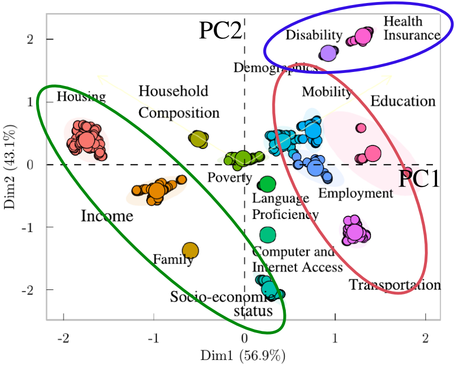
What are the primary topics?

- Socio-Economic and Household Dynamics

American Community Survey: Key Socioeconomic Patterns

Extracted topics *reveal* broad socioeconomic patterns in American Community Survey (ACS)

Topic embeddings

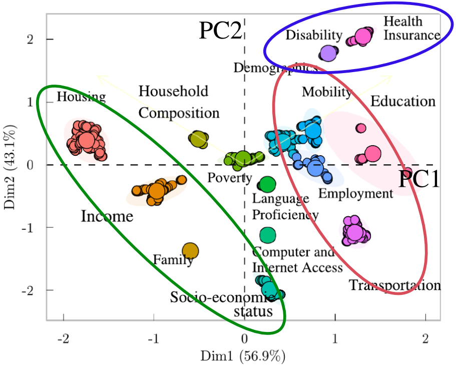


- Socio-Economic and Household Dynamics
  - Income and Housing
  - Family
  - Socio-economic status
- Employment, Education, and Resource Accessibility

American Community Survey: Key Socioeconomic Patterns

Extracted topics *reveal* broad socioeconomic patterns in American Community Survey (ACS)

Topic embeddings

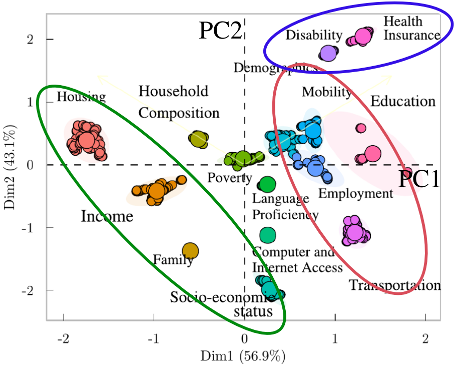


- Socio-Economic and Household Dynamics
  - Income and Housing
  - Family
  - Socio-economic status
- Employment, Education, and Resource Accessibility
  - Employment, Education
  - Mobility
  - Transportation
  - Computer and Internet Access
- Health and Insurance Coverage

American Community Survey: Key Socioeconomic Patterns

Extracted topics *reveal* broad socioeconomic patterns in American Community Survey (ACS)

Topic embeddings

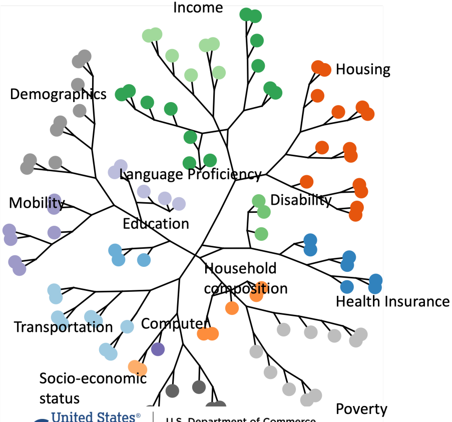


- **Socio-Economic and Household Dynamics**
  - Income and Housing
  - Family
  - Socio-economic status
- **Employment, Education, and Resource Accessibility**
  - Employment, Education
  - Mobility
  - Transportation
  - Computer and Internet Access
- **Health and Insurance Coverage**
  - Disability
  - Health Insurance

Interconnected Social and Demographic Factors Influencing Well-being

Topic modeling *reveals* important aspects of a *community profile*

Linked Socio-Demographic Factors



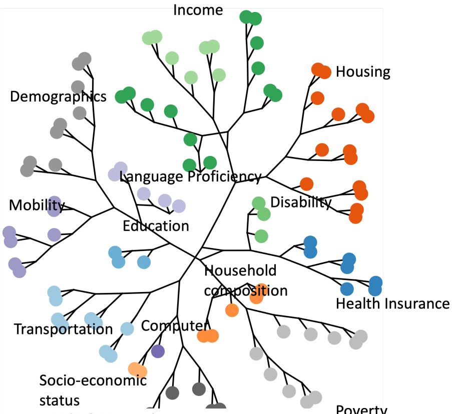
What are the major connections between socio-demographic factors?

- Language proficiency is linked to *education* and *socio-economic opportunities*

Interconnected Social and Demographic Factors Influencing Well-being

Topic modeling *reveals* important aspects of a *community profile*

Linked Socio-Demographic Factors



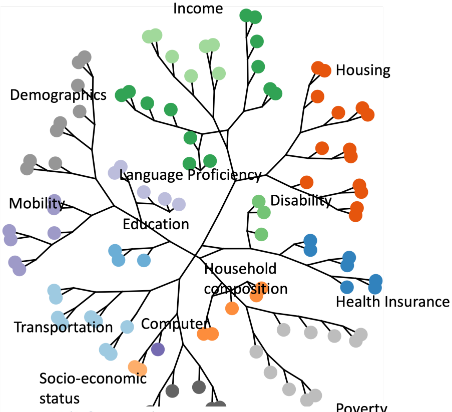
What are the major connections between socio-demographic factors?

- Language proficiency is linked to *education* and *socio-economic opportunities*
- Education directly tied to *income* and *language proficiency*

Interconnected Social and Demographic Factors Influencing Well-being

Topic modeling *reveals* important aspects of a *community profile*

Linked Socio-Demographic Factors



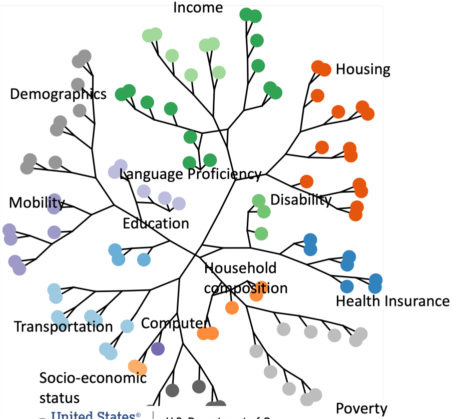
What are the major connections between socio-demographic factors?

- **Language proficiency** is linked to *education* and *socio-economic opportunities*
- **Education** directly tied to *income* and *language proficiency*
- **Housing** is crucial *social determinant* of well-being, influencing *household stability* and *health*

Interconnected Social and Demographic Factors Influencing Well-being

Topic modeling *reveals* important aspects of a *community profile*

Linked Socio-Demographic Factors



What are the major connections between socio-demographic factors?

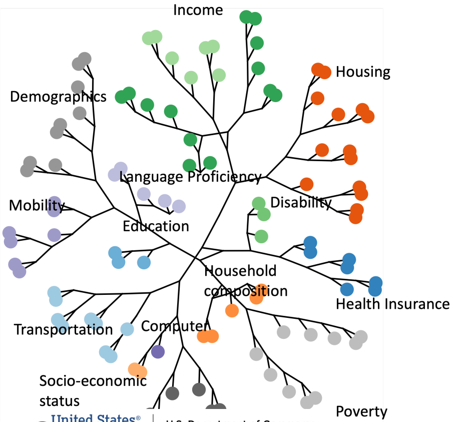
- **Language proficiency** is linked to *education* and *socio-economic opportunities*
- **Education** directly tied to *income* and *language proficiency*
- **Housing** is crucial *social determinant* of well-being, influencing *household stability* and *health*
- **Digital access** impacts *education* and *socio-economic* standing

United States® Census Bureau
U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Interconnected Social and Demographic Factors Influencing Well-being

Topic modeling *reveals* important aspects of a *community profile*

Linked Socio-Demographic Factors



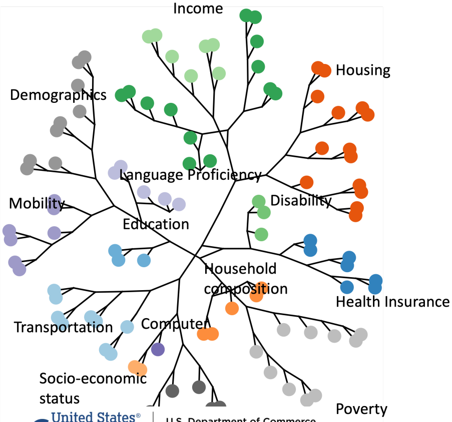
What are the major connections between socio-demographic factors?

- Language proficiency is linked to *education* and *socio-economic opportunities*
- Education directly tied to *income* and *language proficiency*
- Housing is crucial *social determinant* of well-being, influencing *household stability* and *health*
- Digital access impacts *education* and *socio-economic* standing
- Poverty is strongly tied to *income* and *socio-economic* status

Interconnected Social and Demographic Factors Influencing Well-being

Topic modeling *reveals* important aspects of a *community profile*

Linked Socio-Demographic Factors



What are the major connections between socio-demographic factors?

- Language proficiency is linked to *education* and *socio-economic opportunities*
- Education directly tied to *income* and *language proficiency*
- Housing is crucial *social determinant* of well-being, influencing *household stability* and *health*
- Digital access impacts *education* and *socio-economic* standing
- Poverty is strongly tied to *income* and *socio-economic* status

United States® Census Bureau | U.S. Department of Commerce U.S. CENSUS BUREAU census.gov

Outline

1. Background/Motivation
2. Introduction to Topic Modeling
3. Non-Negative Matrix Factorization for Topic Modeling
4. Demonstration of Topic Modeling: American Community Survey
5. Discussion

Leveraging Unsupervised Topic Modeling to Enhance Policy Decisions

- Uncovers hidden patterns
  - ➔ Automatically identifies key topics in large datasets without predefined categories
- Informs data-driven decisions
  - ➔ Helps policymakers prioritize areas requiring intervention based on data-driven insights
- Reveals emerging trends
  - ➔ Detects shifts in public sentiment or issues that may not be immediately apparent through traditional analysis
- Supports proactive policy development
  - ➔ Enables anticipation of future challenges and the formulation of timely, effective policies
- Enhances transparency and equity
  - ➔ Contributes to more informed, transparent, and equitable decision-making

Leveraging Unsupervised Topic Modeling to Enhance Policy Decisions

- Uncovers hidden patterns
  - ➔ Automatically identifies key topics in large datasets without predefined categories
- Informs data-driven decisions
  - ➔ Helps policymakers prioritize areas requiring intervention based on data-driven insights
- Reveals emerging trends
  - ➔ Detects shifts in public sentiment or issues that may not be immediately apparent through traditional analysis
- Supports proactive policy development
  - ➔ Enables anticipation of future challenges and the formulation of timely, effective policies
- Enhances transparency and equity
  - ➔ Contributes to more informed, transparent, and equitable decision-making

Leveraging Unsupervised Topic Modeling to Enhance Policy Decisions

- Uncovers hidden patterns
  - ➔ Automatically identifies key topics in large datasets without predefined categories
- Informs data-driven decisions
  - ➔ Helps policymakers prioritize areas requiring intervention based on data-driven insights
- Reveals emerging trends
  - ➔ Detects shifts in public sentiment or issues that may not be immediately apparent through traditional analysis
- Supports proactive policy development
  - ➔ Enables anticipation of future challenges and the formulation of timely, effective policies
- Enhances transparency and equity
  - ➔ Contributes to more informed, transparent, and equitable decision-making

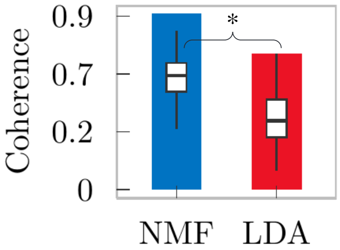
Leveraging Unsupervised Topic Modeling to Enhance Policy Decisions

- Uncovers hidden patterns
  - ➔ Automatically identifies key topics in large datasets without predefined categories
- Informs data-driven decisions
  - ➔ Helps policymakers prioritize areas requiring intervention based on data-driven insights
- Reveals emerging trends
  - ➔ Detects shifts in public sentiment or issues that may not be immediately apparent through traditional analysis
- Supports proactive policy development
  - ➔ Enables anticipation of future challenges and the formulation of timely, effective policies
- Enhances transparency and equity
  - ➔ Contributes to more informed, transparent, and equitable decision-making

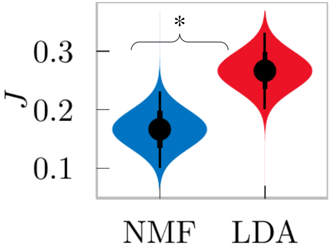
Data-Driven Methods Provide Significant Gains in Topic Modeling

Data-driven (NMF) vs. Probabilistic Topic Modeling (Latent Dirichlet Allocation)

Topic Coherence



Topic Diversity



Topic Coherence ↑

$$C(T) = \sum \log \frac{DF(w_i, w_j)}{DF(w_i)}$$

Topic Diversity (Jaccard Similarity) ↓

$$J(T_i, T_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|}$$

$DF(w_i)$ document frequency of word w_k, $DF(w_i, w_j)$ co-document frequency
T_i, T_j words for topic i and j

Acknowledgements

This work was supported:

Richie Wang, U.S Census Bureau, Assistant Center Chief

Christopher Carino, Ph.D, U.S Census Bureau, Senior Computer Scientist

Thank you

Irina Belyaeva, Ph.D,
Chief Architect, AI and Statistical Applications
Research and Methodology
U.S. Census Bureau
irina.belyaeva@census.gov