Istat | Istituto Nazionale di Statistica

Valencia, 26-28 June 2024

CARMA 2024

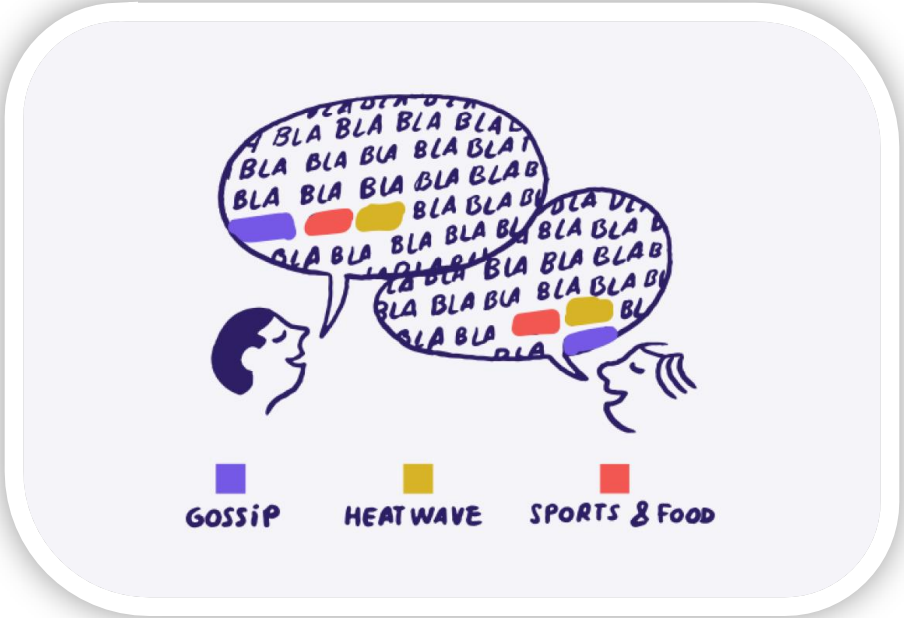
TOPIC MODELING LAB

A **HANDS-ON** REVIEW OF THE MOST POPULAR TOPIC MODELING TECHNIQUES

M. BRUNO, F. PUGLIESE, E. CATANESE, F. ORTAME
Istat | DCME

What is Topic Modeling?

- Topic modeling is a type of statistical modeling used to discover **abstract topics within a collection of documents**.
- This technique is widely used in natural language processing (NLP) to uncover **hidden patterns** and structure in large textual datasets.
- The primary goal of topic modeling is to automatically identify topics present in a corpus and to organize the documents according to these topics.



The code of the lab is available on github!!

A few key concepts…

- Documents and Words (Tokens)

The basic units of topic modeling. Documents are the individual pieces of text (e.g., Tweets, reviews...), and words are the tokens or terms within these documents.

- Corpus

A corpus is a collection of documents. Topic modeling algorithms analyze the corpus to identify the topics.

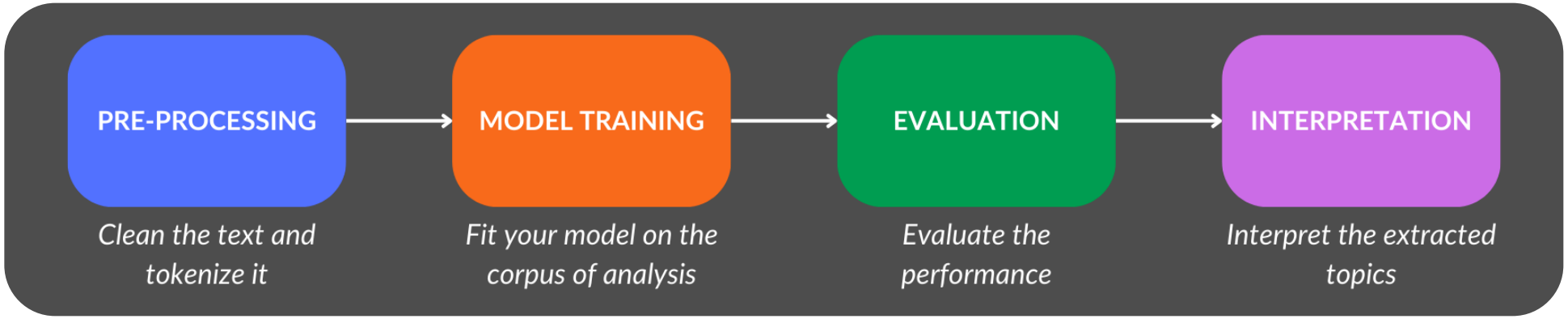
- Topics

A topic is a **distribution over a fixed vocabulary**. It is characterized by a set of words that frequently appear together. **Each topic can be seen as a pattern of co-occurrence of words**.

- Latent Variables

These are variables that are not directly observed (e.g., **the topics**) but are inferred from other variables that are observed (e.g., **the words**).

Data processing pipeline



1) Pre-processing: Clean the text data (e.g., remove stopwords and tokenize)

2) Model Training: Apply a topic modeling algorithm to the preprocessed data

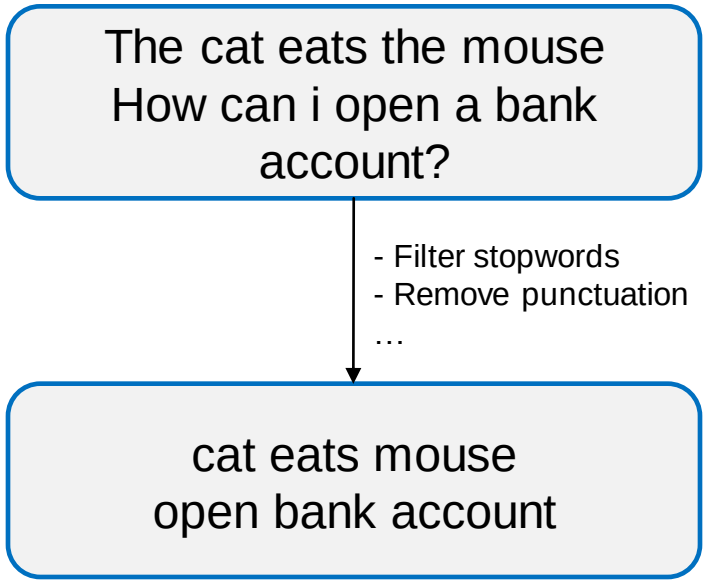
3) Evaluation: Assess the quality of the topics using **coherence scores**, **human judgement**, or other metrics

4) Interpretation: Analyze the topics and assign meaningful labels or descriptions

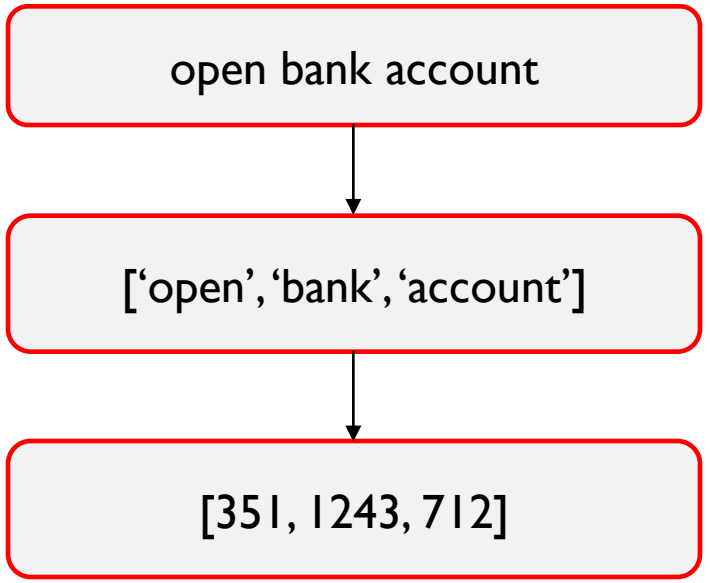
Data pre-processing for traditional methods

Before training our models, we need to ensure that the data is in the correct format.

TEXT CLEANING



TOKENIZATION



VECTOR-SPACE REPRESENTATION

Term-Document Matrix

	D1	D2
712 (*account*)	0	1
1243 (*bank*)	0	1
145 (*cat*)	1	0
1378 (*eats*)	1	0
449 (*mouse*)	1	0
351 (*open*)	0	1

Istat

Data pre-processing for traditional methods

Document Term Matrix (DTM):

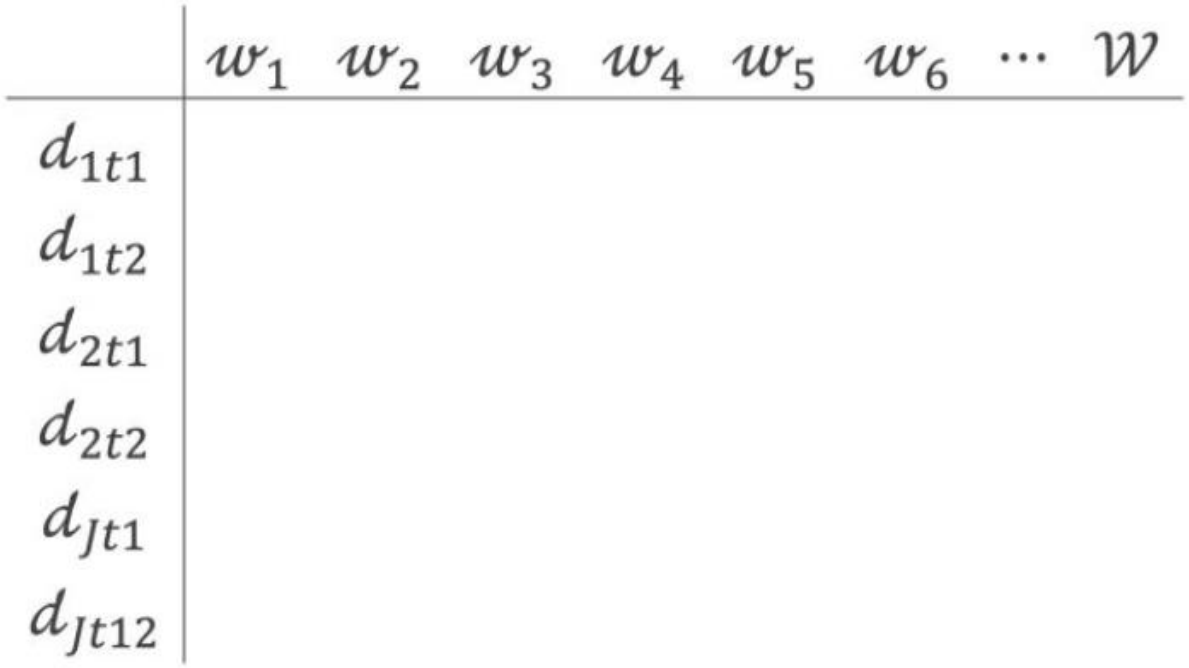
In a typical NLP task, DTM can be huge!!

More precisely:

Rows: documents in the corpus (**N**)

Columns: The columns represent **unique words**, which means, of course, each word onl shows up one time (**M**).

In our lab we will analyze a dataset of **100k** Tweets, the size of the vocabulary is more or less **50k**



DTM = (NxM) = 100K x 50K

Istat

Data pre-processing for traditional methods

Document Term Matrix (DTM):

In a typical NLP task, DTM can be huge!!

More precisely:

Rows: documents in the corpus (**N**)

Columns: The columns represent **unique words**, which means, of course, each word only shows up one time (**M**).

In our lab we will analyze a dataset of **100k** Tweets, the size of the vocabulary is more or less **50k**

	w_1	w_2	w_3	w_4	w_5	w_6	...	w
d_{1t1}	4	0	0	0	1	2		0
d_{1t2}	0	1	7	0	0	1		3
d_{2t1}	0	5	0	3	0	9		1
d_{2t2}	3	0	8	0	1	0		0
d_{Jt1}	0	2	1	0	2	12		4
d_{Jt12}	1	0	0	4	0	0		2

DTM = (NxM) = 100K x 50K

Istat

Data pre-processing for traditional methods

Document Term Matrix (DTM):

In a typical NLP task, DTM can be huge!!

More precisely:

Rows: documents in the corpus (**N**)

Columns: The columns represent **unique words**, which means, of course, each word only shows up one time (**M**).

In our lab we will analyze a dataset of **100k** Tweets, the size of the vocabulary is more or less **50k**

	w_1^*	w_2^*	w_3^*	w_4^*	w_5^*	w_6^*	...	W^*
d_{1t1}^*	2	0	0	0	1	0		0
d_{1t2}^*	0	1	0	0	0	1		3
d_{2t1}^*	0	0	0	3	0	2		1
d_{2t2}^*	1	0	1	0	1	0		0
d_{Jt1}^*	0	2	1	0	2	0		0
d_{Jt12}^*	1	0	0	1	0	0		1

DTM = (NxM) = 100K x 50K

Istat

Traditional Topic Modeling Techniques

Traditional topic modeling techniques rely on statistical methods to uncover hidden topics in a corpus.

We will explore:

- **Latent Dirichlet Allocation** (LDA)
- **Hierarchical Dirichlet Process** (HDP)
- **Non-negative Matrix Factorization** (NMF)

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that assumes documents to be mixtures of topics, and topics to be mixtures of words.

LDA involves the following steps:

- **Parameter Initialization**: LDA initializes the topics, the topic distribution for each document, and the word distribution for each topic.
- **Training**: Using an iterative process (usually Gibbs sampling), LDA refines these distributions to fit the observed data better.
- **Topic Inference**: After several iterations, LDA infers the topic distribution for each document and the word distribution for each topic.

In LDA, the number of latent topics to be extracted needs to be defined *a priori*.

Hierarchical Dirichlet Process (HDP)

Hierarchical Dirichlet Process (HDP) is a non-parametric Bayesian approach to topic modeling. Unlike LDA, HDP automatically determines the number of topics based on the data.

HDP involves the following steps:

- **Initialization**: HDP starts with an initial guess on the topic distribution.
- **Iterative Refinement**: Using a hierarchical process, HDP refines the topic distribution at both the document and corpus levels.
- **Dynamic Topic Adjustment**: HDP adjusts the number of topics dynamically as more data is processed.

In HDP, the number of latent topics to be extracted is *not* defined *a priori*.

Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) is a non-probabilistic technique, particularly suitable for large, sparse datasets. Unlike probabilistic models like LDA and HDP, **NMF is a linear algebra-based method** that decomposes the document-term matrix into two lower-dimensional matrices, one representing the topics and the other representing the topic distribution for each document.

NMF involves the following steps:

- **Matrix Decomposition**: NMF decomposes the document-term matrix into two non-negative matrices.
- **Iterative Optimization**: Using iterative optimization techniques, NMF refines these matrices to minimize the reconstruction error
- **Topic Extraction**: The resulting matrices are used to extract the topics and their distribution across documents.

Likewise LDA, NMF requires the number of topics to extract to be defined *a priori*.

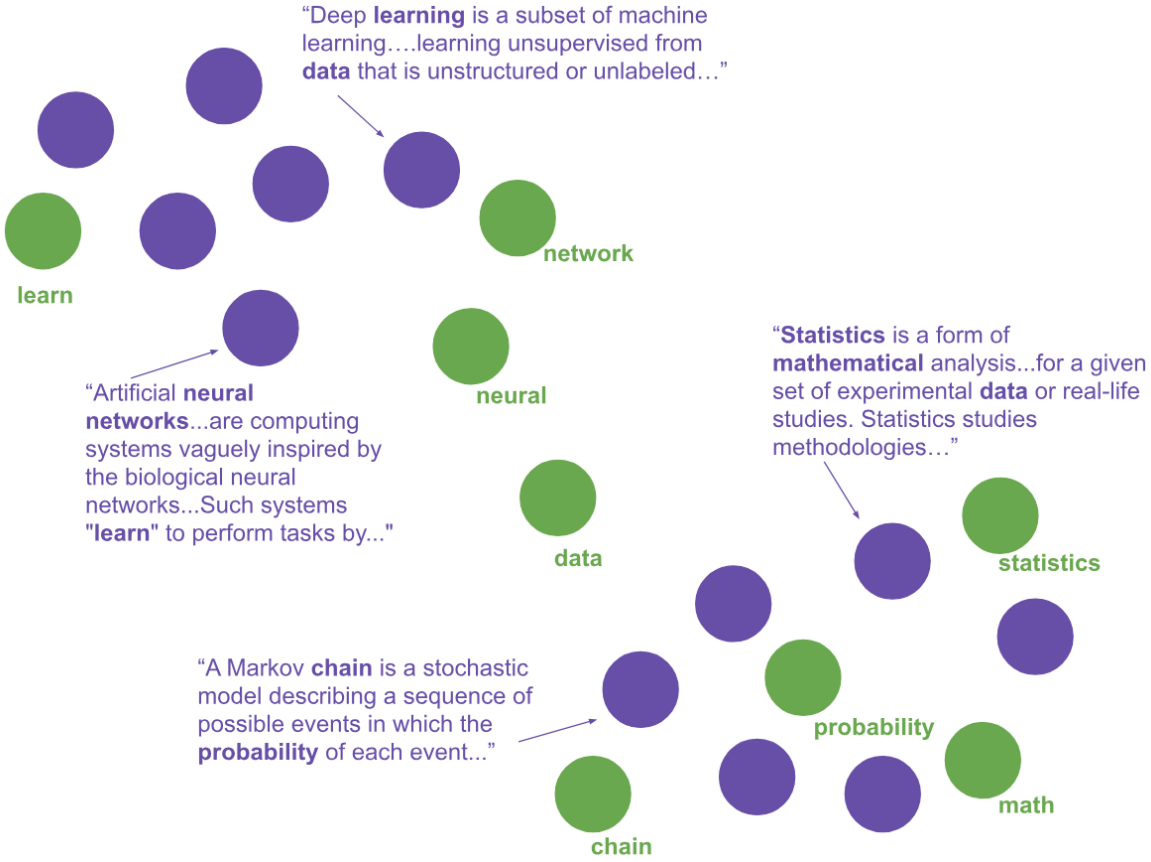
Clustering Algorithms on Embedding Spaces

Unlike traditional techniques, clustering algorithms on embedding spaces leverage **neural network-based word embeddings to discover latent topics in text data**.

These approaches rely on dense word representations to capture **semantic relationships** more effectively.

We will explore:

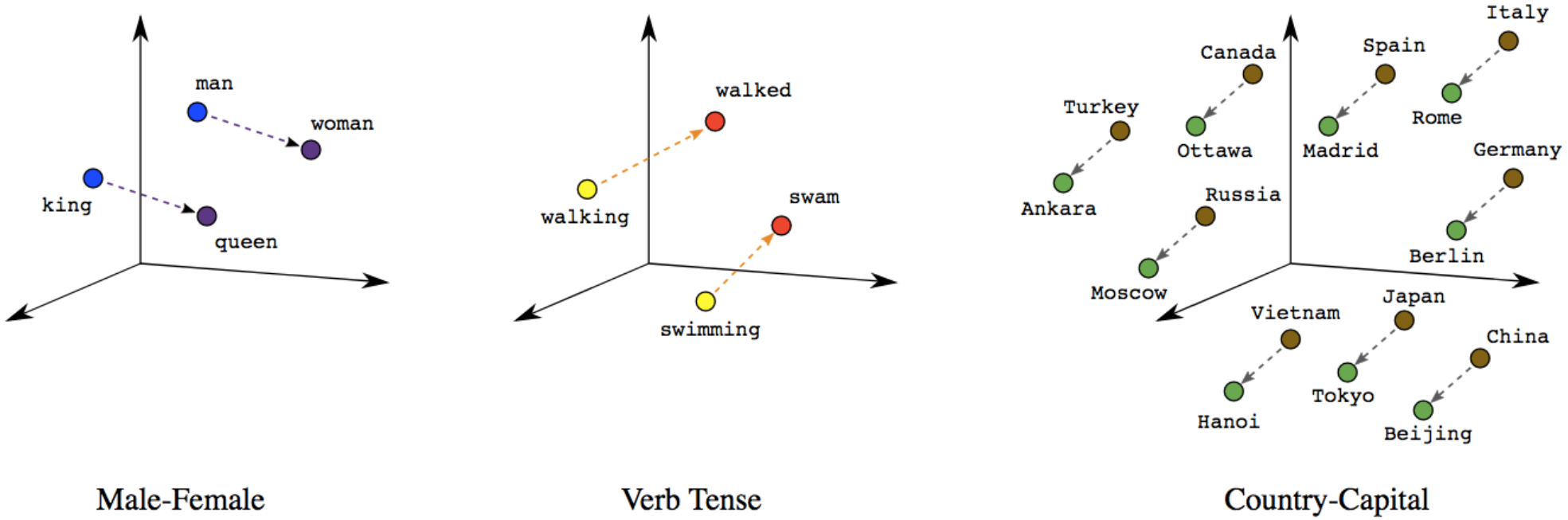
- **Top2Vec**
- **BERTopic**



Source: Top2Vec (GitHub)

Word Embeddings

Word embeddings are dense vector representations of words, **capturing semantic relationships by placing similar words closer in the vector space**.
They are trained on large text corpora to understand context and meaning.



Male-Female Verb Tense Country-Capital

Source: Towards Data Science

Top2Vec

Top2Vec is an algorithm that simultaneously learns the topic representations and the word embeddings. By mapping documents to a continuous vector space, **Top2Vec identifies clusters of documents that share similar themes without requiring a pre-defined number of topics**. This approach allows for the discovery of natural and meaningful topics directly from the data.

Top2Vec involves the following steps:

- **Embedding Creation**: Top2Vec uses word embeddings to create document vectors.
- **Dimensionality Reduction**: The document vectors are reduced to a lower-dimensional space using techniques like UMAP.
- **Clustering**: The reduced vectors are clustered to identify topics.
- **Topic Words Identification**: The algorithm finds words that are closest to the cluster centroids, representing the topics.

BERTopic

BERTopic leverages transformer-based embeddings to create document representations and applies clustering algorithms to discover topics. It combines BERT embeddings with clustering techniques like HDBSCAN and dimensionality reduction methods like UMAP to generate coherent topics from text data.

BERTopic involves the following steps:

- **Embedding Creation**: BERTopic uses transformer models to create document embeddings.
- **Dimensionality Reduction**: The embeddings are reduced in dimensionality using UMAP.
- **Clustering**: Density-based clustering algorithm such as HDBSCAN are applied to form clusters from the reduced embeddings.
- **Topic Representation**: The algorithm generates topics based on the clustered documents and their embeddings. In this step, LLMs can be used for automatic and meaningful topic labeling.

Thank You!

Mauro Bruno | mbruno@istat.it

Francesco Pugliese | frpuglie@istat.it

Francesco Ortame | francesco.ortame@istat.it

Elena Catanese | catanese@istat.it

Istat | Istituto Nazionale di Statistica