Distributed Representation of Documents with Explicit Explanatory Features

September 14th, 2015

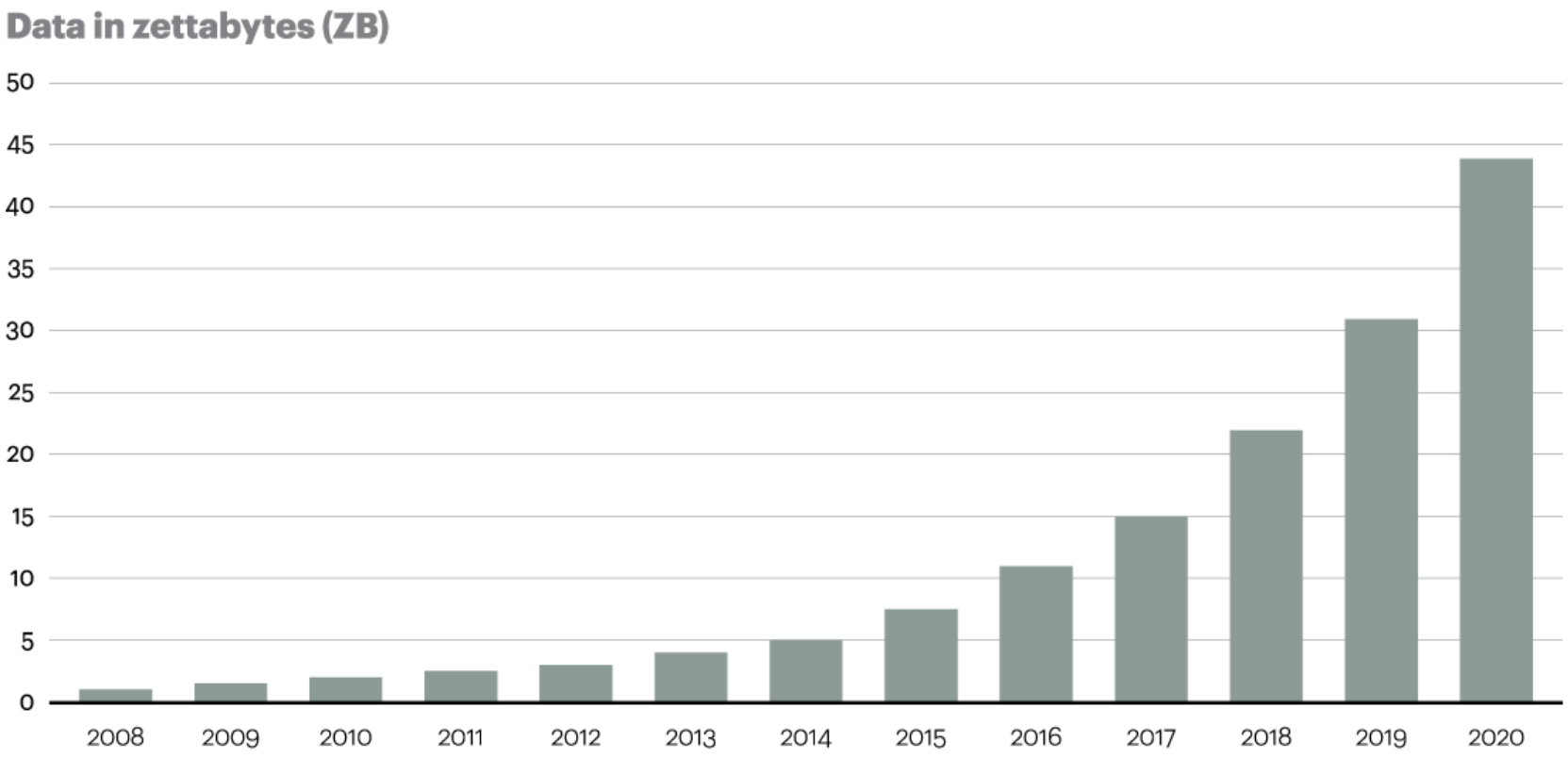
SNU Data Mining Center

Han Kyul Kim

Ascendance of Data

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020

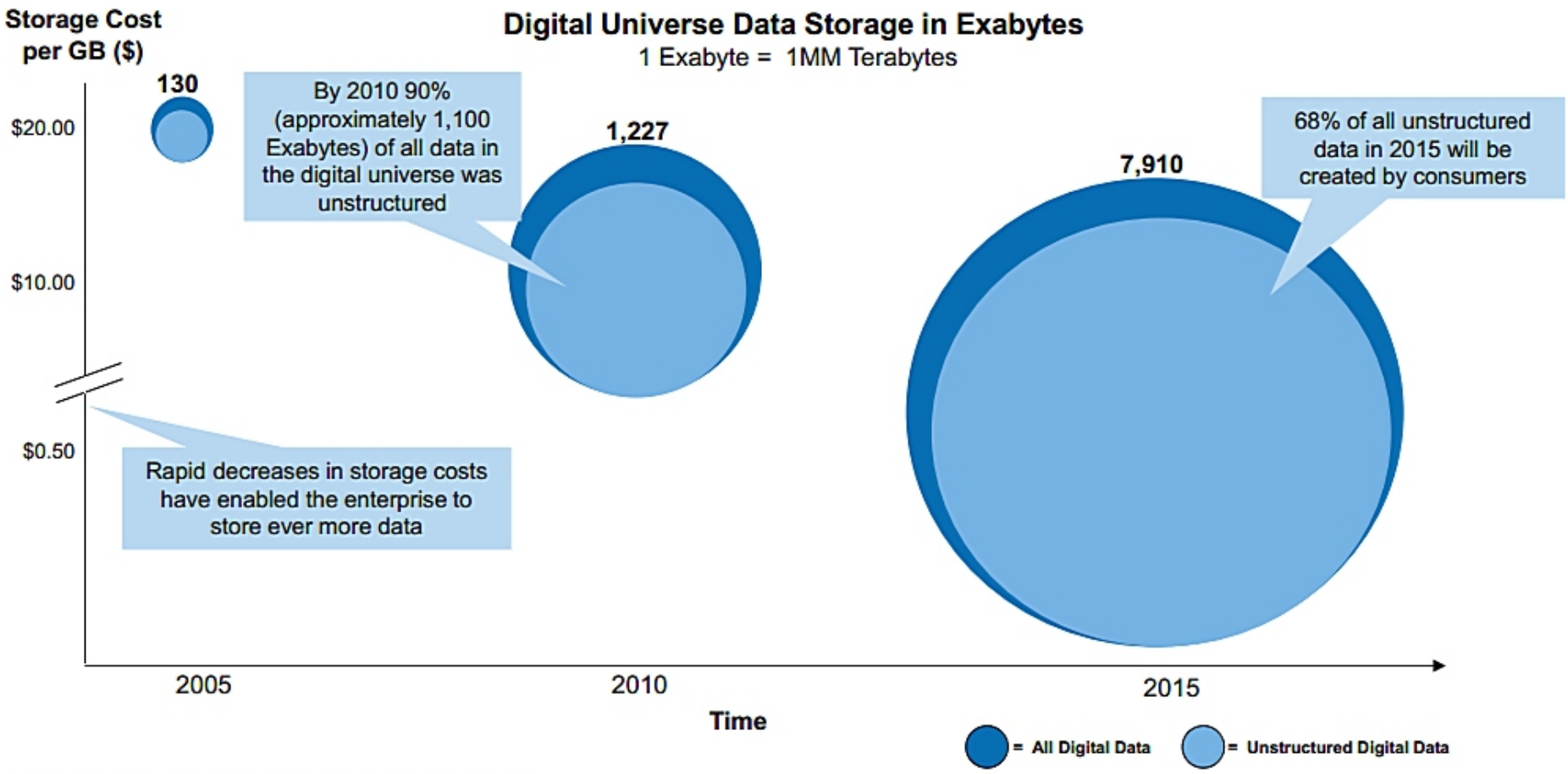


Source: Oracle, 2012

*** 1 ZB = 1 billion TB**

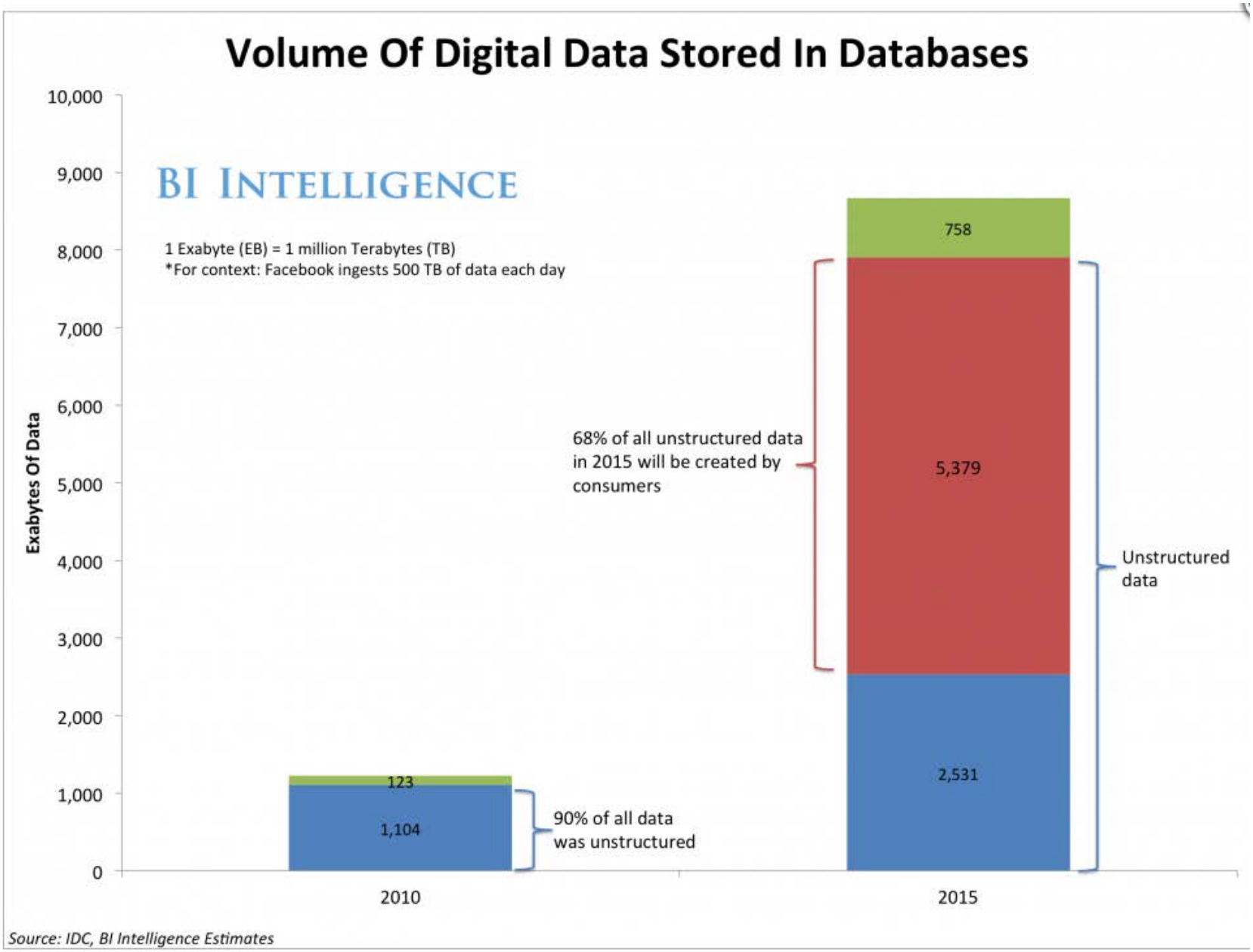
Ascendance of "Unstructured Data"

Unstructured consumer data, called Big Data, represents majority of growth in data volume, up 56% CAGR since 2005



Source: IDC's Digital Universe Study, sponsored by EMC, June 2011

Ascendance of "Unstructured Data"



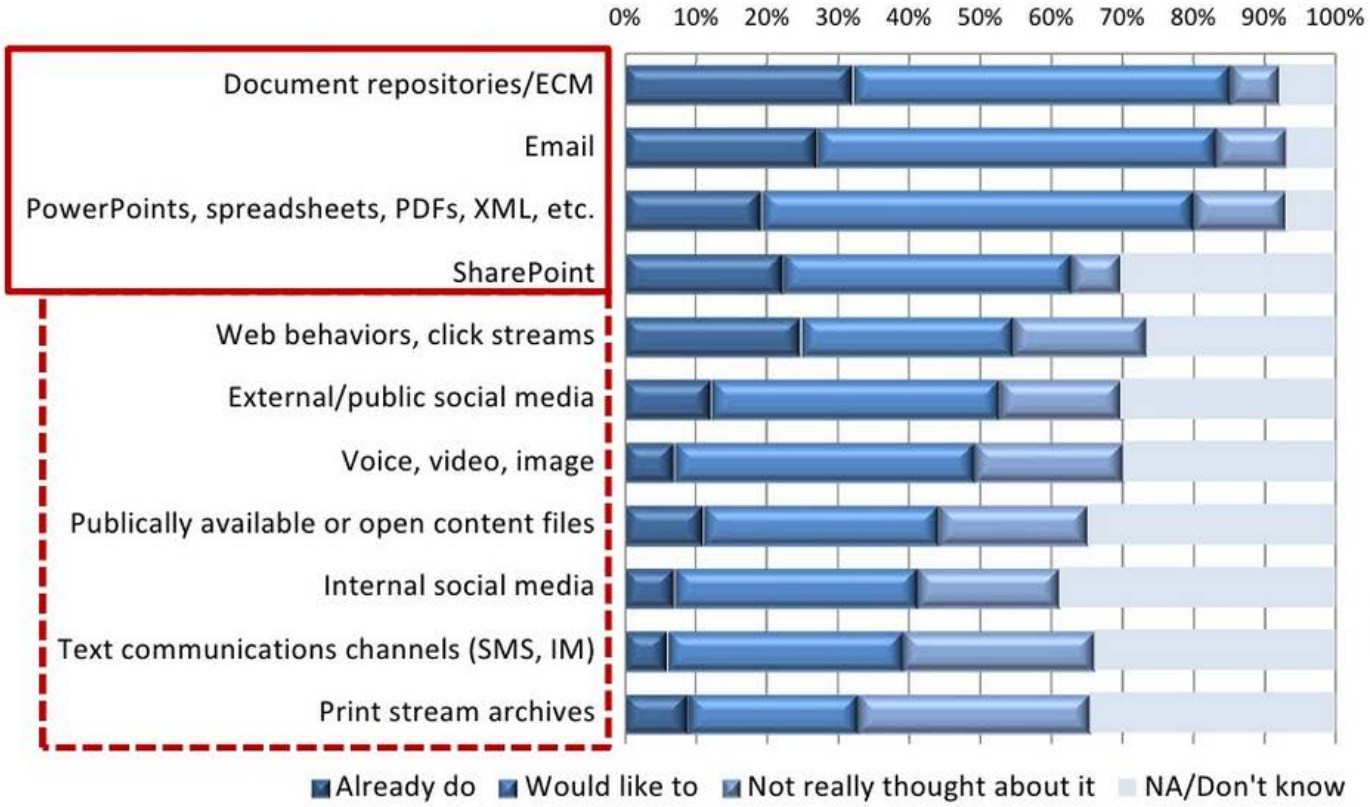
Are companies ready?

Considered overall, to what degree does your organization exploit its information assets for analysis and decision making purposes?

Source: Online survey of Register readers, 122 respondents, first half of November 2011, Freeform Dynamics

What kind of unstructured data are they interested in?

Are there large unstructured or semi-structured data repositories (ie, text, rich media, etc.) in your business that you would like to analyze, monitor or query - as opposed to search/retrieve?



Most popular are the basic ones – maybe some crossover "search/analyze."

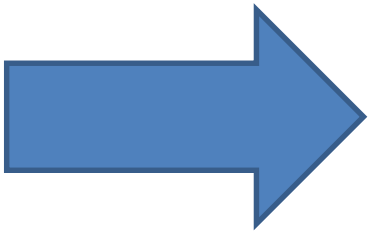
But 30-50% interested in a wide range of other applications.

N=300

But how should we represent a document?

- Number of Raw Words
- Number of Paragraphs
- Semantic Frequency
- Word Order
- Part of Speech
- Topic
- Sentiment

……



Doc 1: [1, 4, 3, 0, …]
Doc 2: [5, 2, 8 0, …]
……

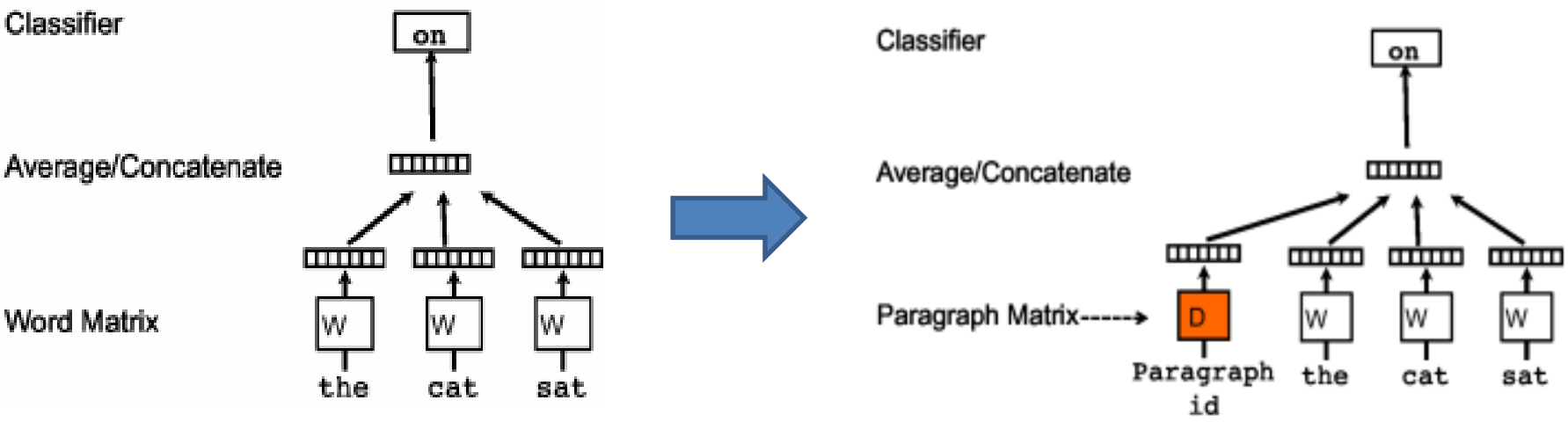
Term-Document Matrix Representation

- Bag of Word Hypothesis:
  - Frequencies of words in a document tend to indicate the relevance of documents

- Examples:
  1. Bag-of-words
     - Word order & semantic information are lost
     - Different sentences have exactly same representation as long as the same words are used
  2. Bag-of-n-grams
     - Restricted to short interval of texts
     - High dimensionality
     - Semantic information lost
  3. Weighted Average of word vectors
     - Loses the word order
     - Not that much different from BOW
  4. Word vectors + Parse Tree
     - Preserves the word order
     - Restricted to representing at the sentence level

- Yet, it still is one of the most popular method for representing document
  - Lucene: Term-Document Matrix method

Distributed Representation - Doc2Vec



<Word2Vec> <Paragraph2Vec>

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents

- Extension of Word2Vec
- Based on Distributional Hypothesis (Harris, 1954)
  - Words that occur in similar contexts tend to have similar meanings
- Able to represent input sequence of **variable** length
- Captures **semantic information** of words within paragraphs
- Doesn't rely on weight functions or parse trees
- Trained from **unlabeled data**

Next Generation of Document Representation?

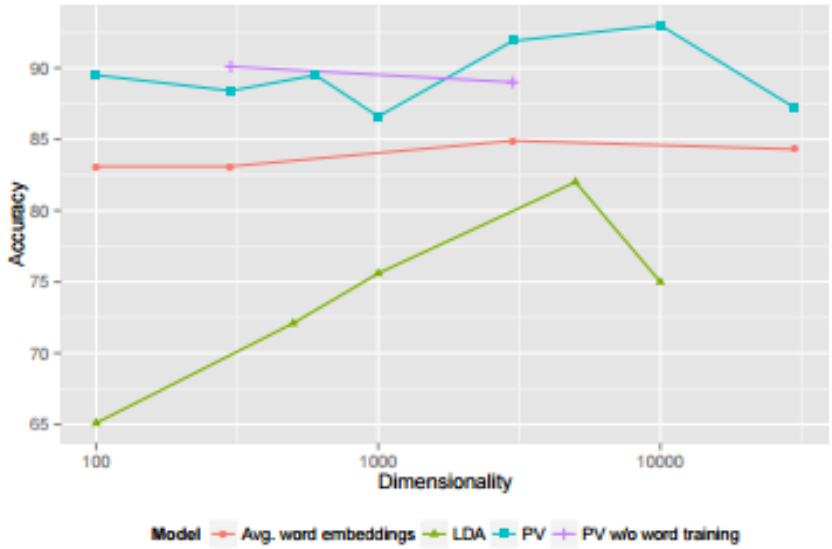


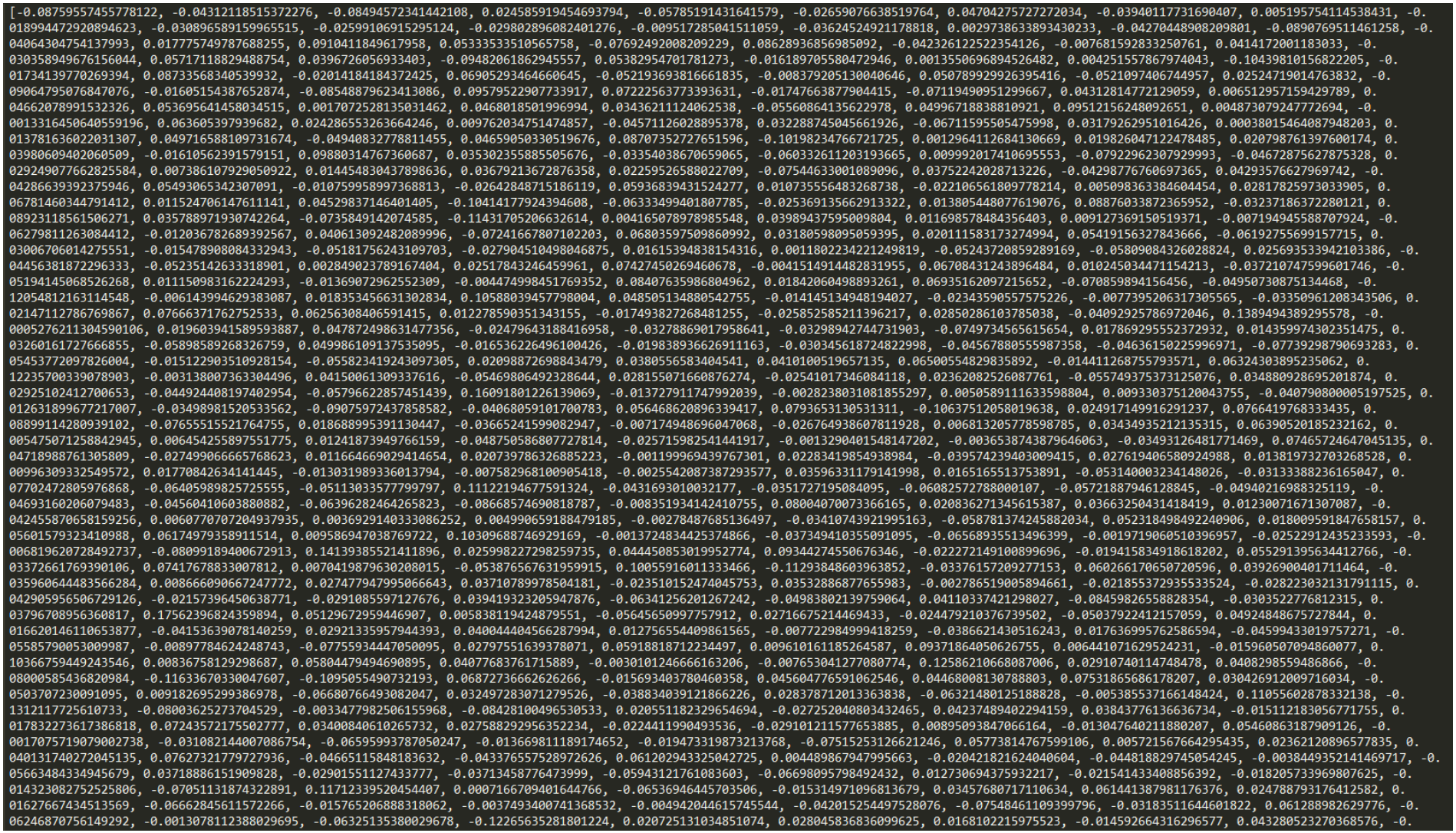
Figure 4: Results of experiments on the hand-built Wikipedia triplet dataset.

Table 3: Performances of different methods on hand-built triplets of Wikipedia articles on the best performing dimensionality.

Model	Embedding dimensions/topics	Accuracy
Paragraph vectors	10000	93.0%
LDA	5000	82%
Averaged word embeddings	3000	84.9%
Bag of words		86.0%

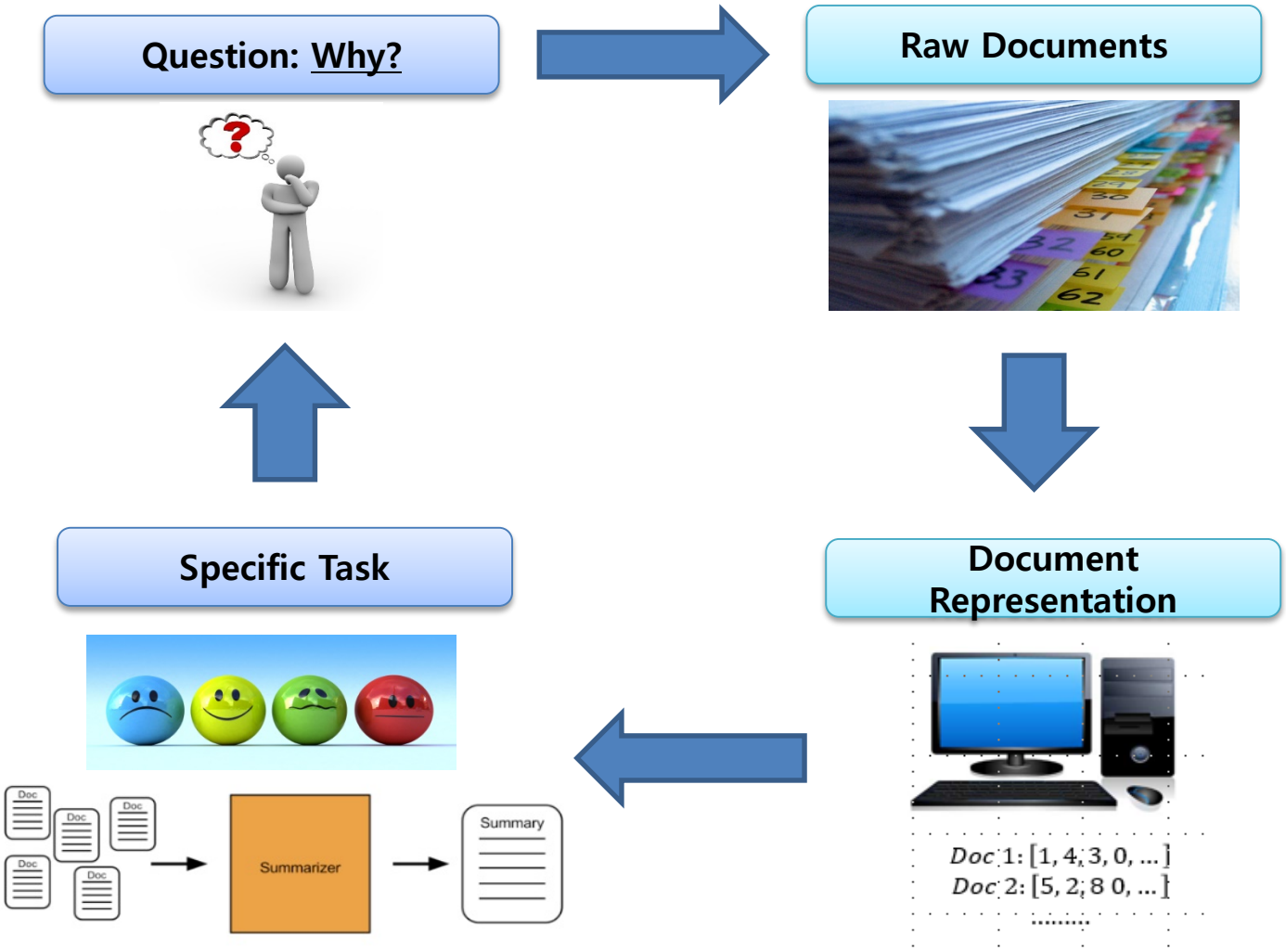
Dai, Andrew M., et al. "Document Embedding with Paragraph Vectors." NIPS Deep Learning Workshop. 2014.

Drawbacks of Doc2Vec



Having a good representation form itself is not the ultimate goal of document representation!

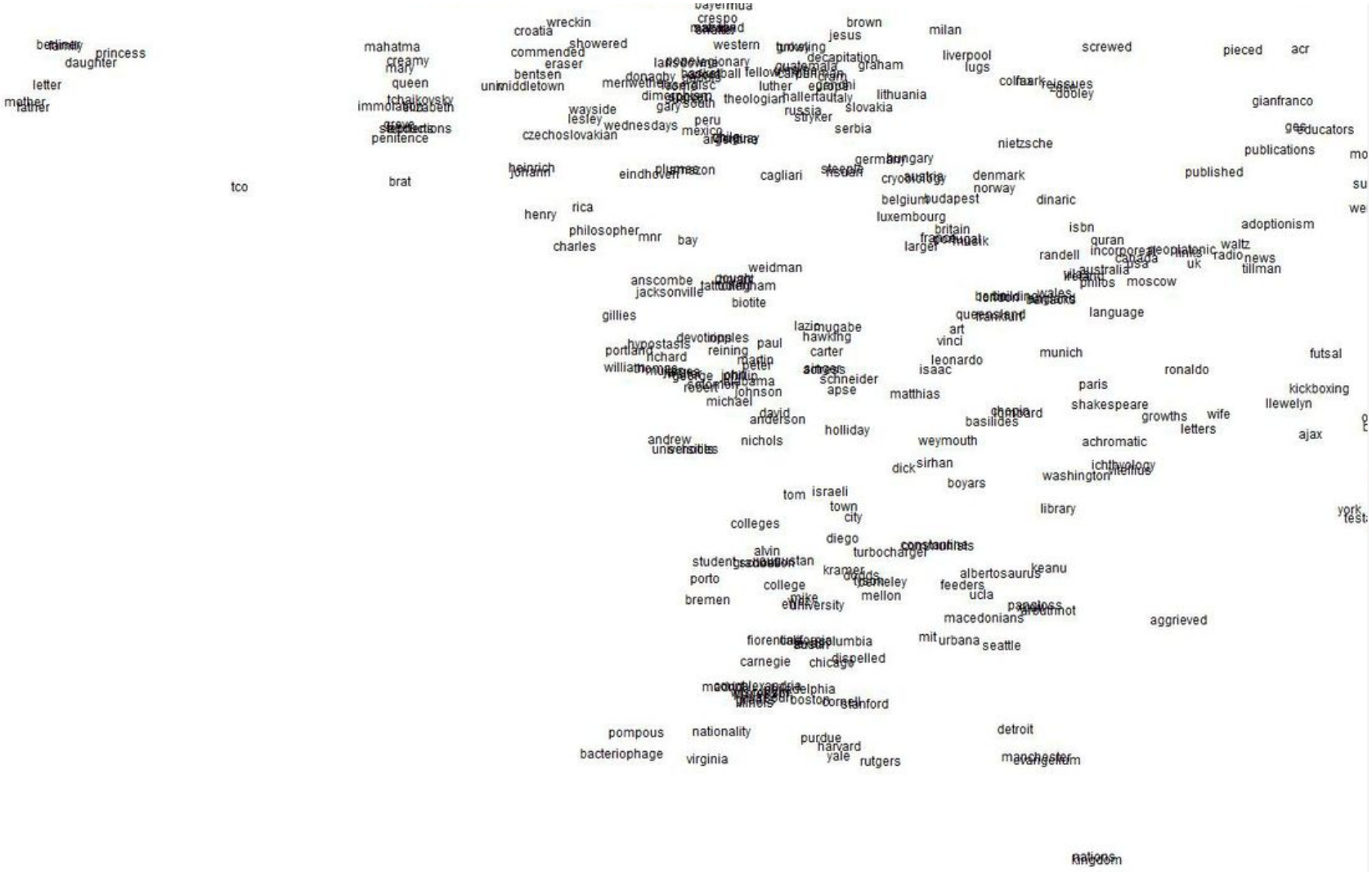
Drawbacks of Doc2Vec



Doc2Vec Representation provides with an insight to what's happening amongst the documents, but not how and why it's happening

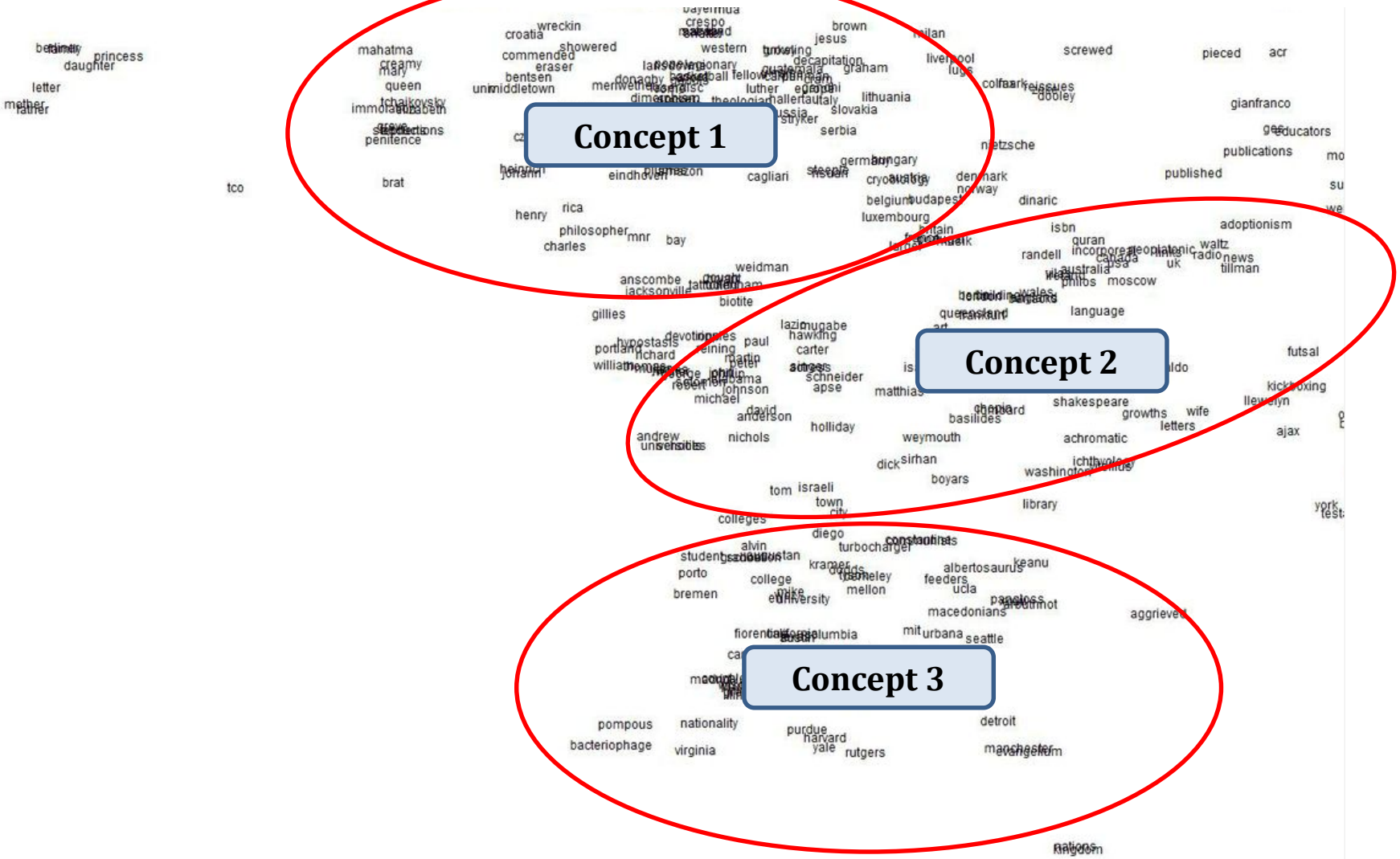
Proposed Framework

1. **Train Word2Vec with the collection of documents**



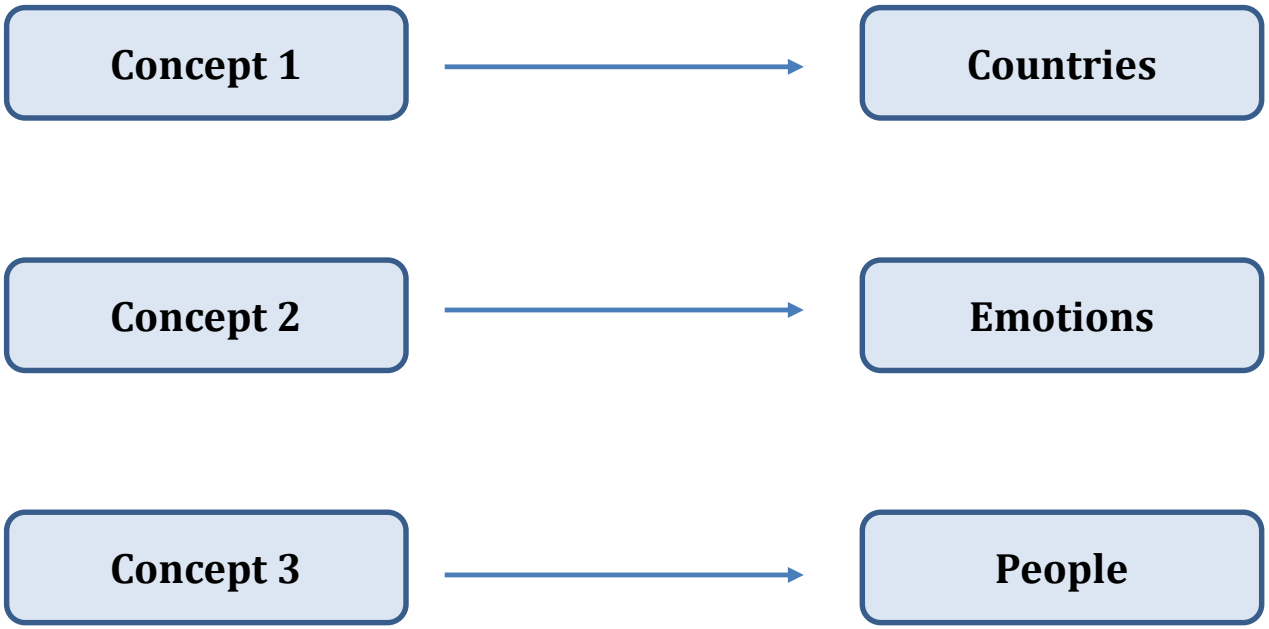
Proposed Framework

2. Cluster word2vec generated vectors to create clusters of concepts



Proposed Framework

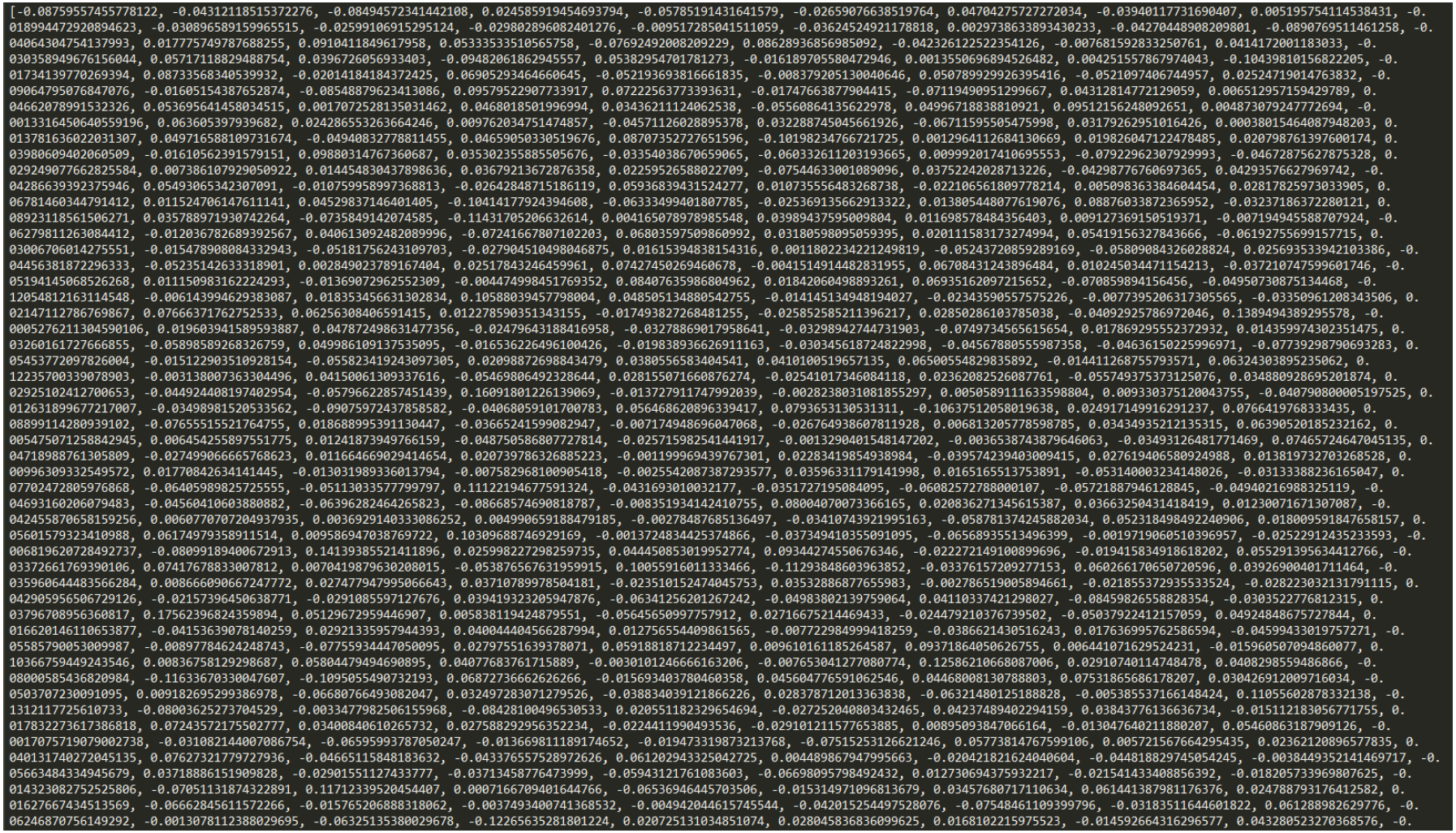
3. **Label the concepts using the words associated with each cluster**



Proposed Framework

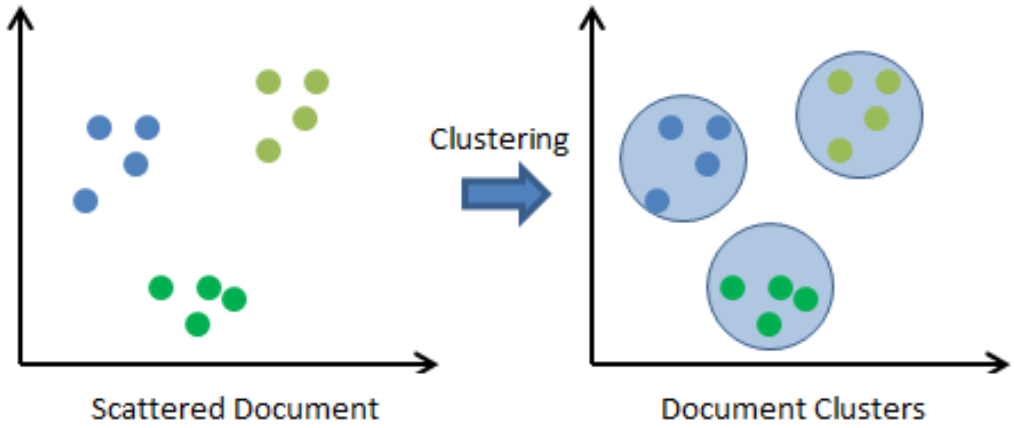
4. Represent the documents using the counts of these concepts

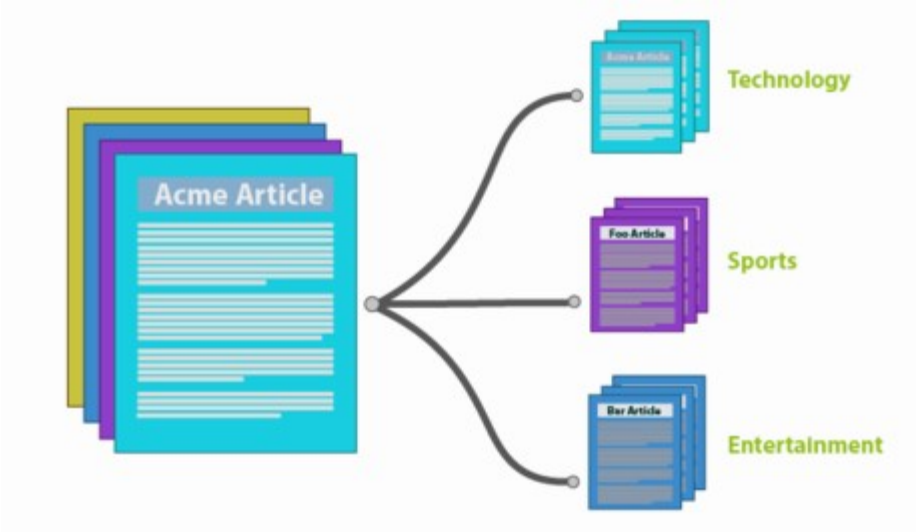
Doc 1 = [Countries, Emotions, People …]



Proposed Framework

5. **Test the effectiveness of the document representation through document clustering and classification**





Contribution

This proposed framework…

- Maintains the representational power of the distributed representation, while providing explicit and intuitive features
- Provides reasons and logic behind the representation
- Includes more holistic semantic information about documents
- Suggests language independent framework for representing documents

Reference

Dai, Andrew M., et al. "Document Embedding with Paragraph Vectors." NIPS Deep Learning Workshop. 2014.

Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." arXiv preprint arXiv:1405.4053 (2014).

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems. 2013.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. "Exploiting similarities among languages for machine translation." arXiv preprint arXiv:1309.4168(2013).

Rong, Xin. "word2vec Parameter Learning Explained." arXiv preprint arXiv:1411.2738 (2014).

Turney, Peter D., and Patrick Pantel. "From frequency to meaning: Vector space models of semantics." Journal of artificial intelligence research 37.1 (2010): 141-188.