# Topic Modeling

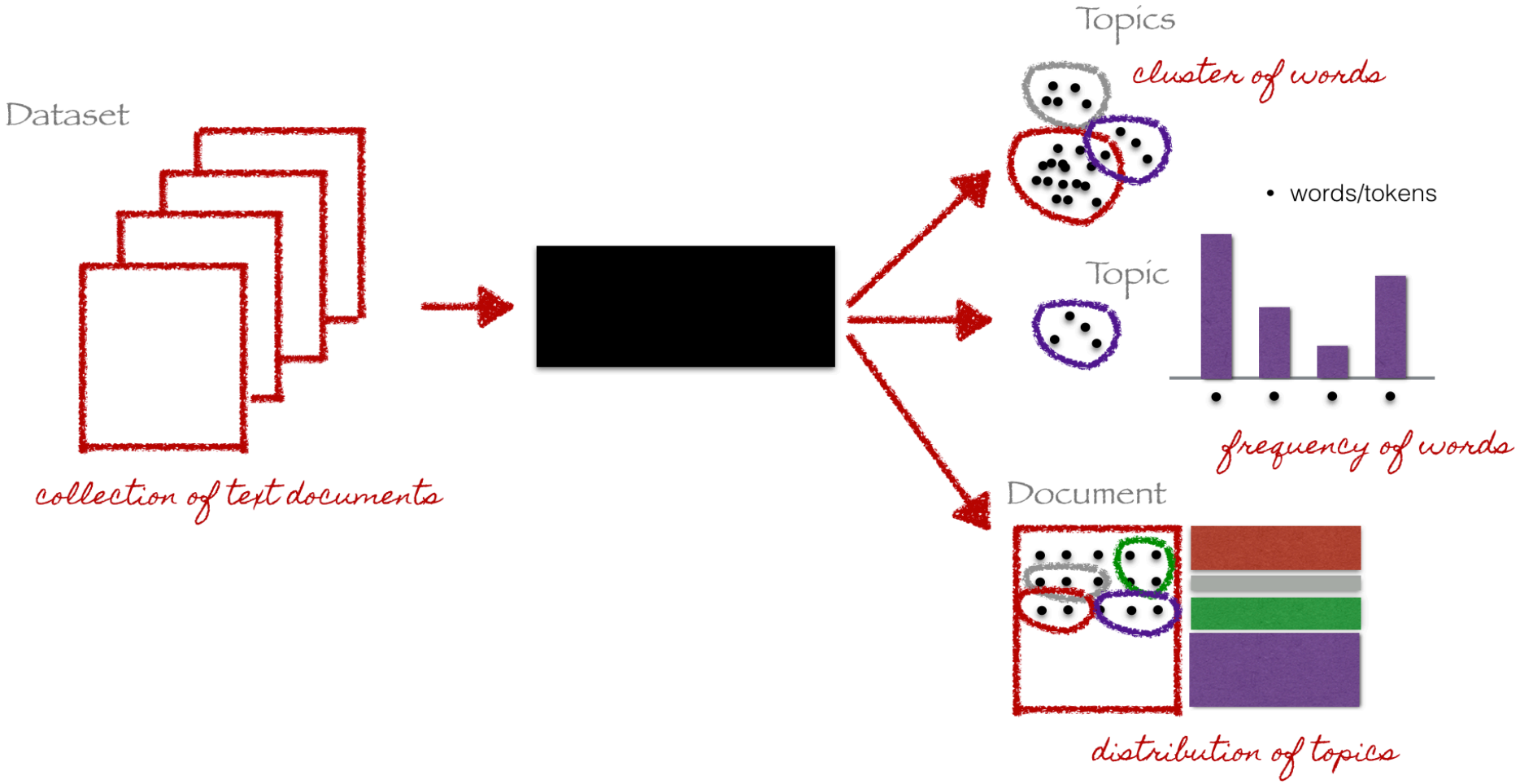
Gopika Jayadev [GGJ236]
Deepak Mahapatra [DKM2227]
Mridula Maddukuri [MM82959]

# Outline

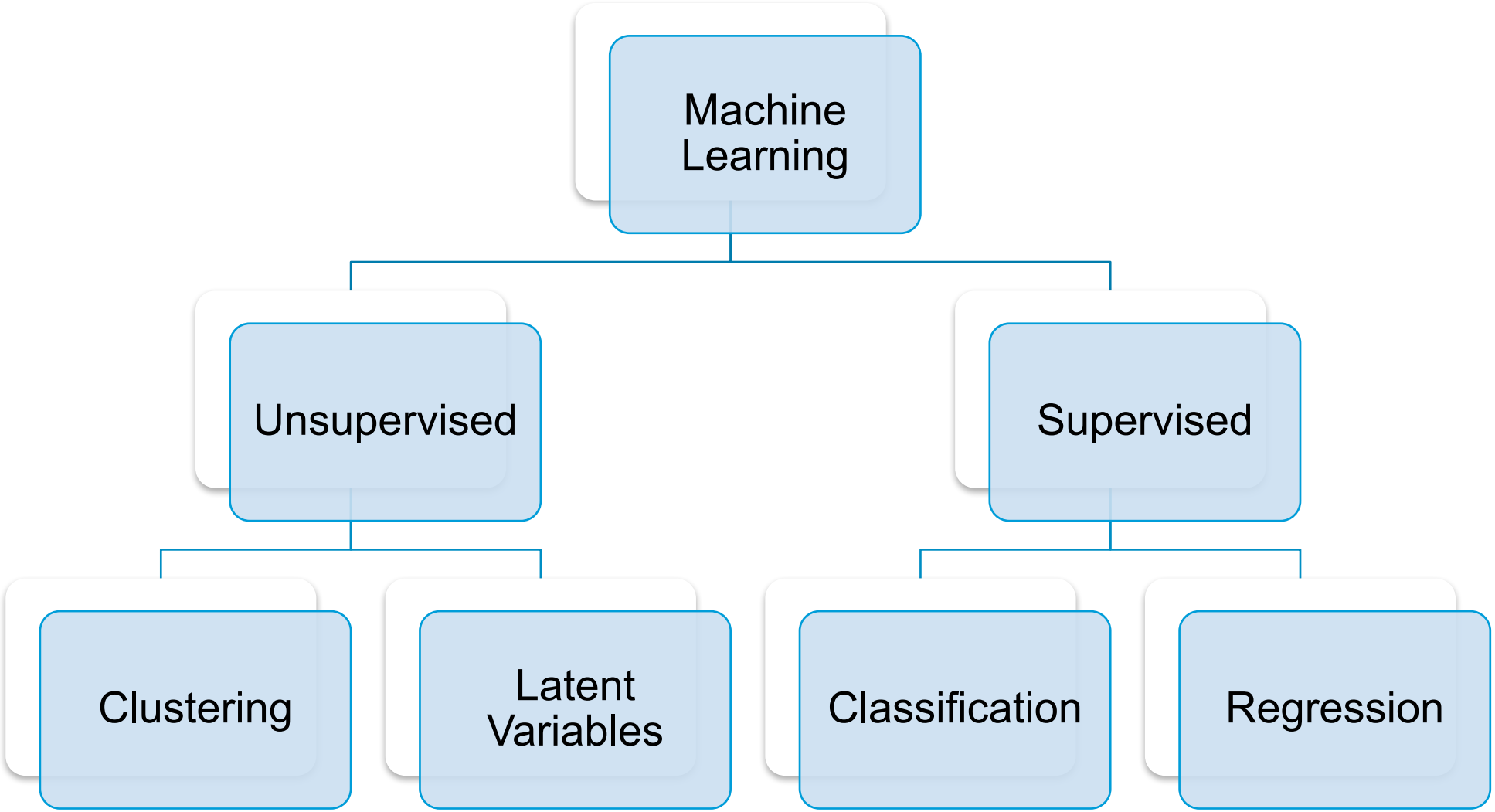
- Introduction
- Preprocessing
- Algorithms
  - NMF
  - LDA
- Results

# Topic Modeling

- Identify semantic structures in a corpus




Courtesy:
Introduction to
Topic Modeling in
Python
by Christine Doig




LDA: Latent Dirichlet Allocation
NMF: Nonnegative Matrix Factorization

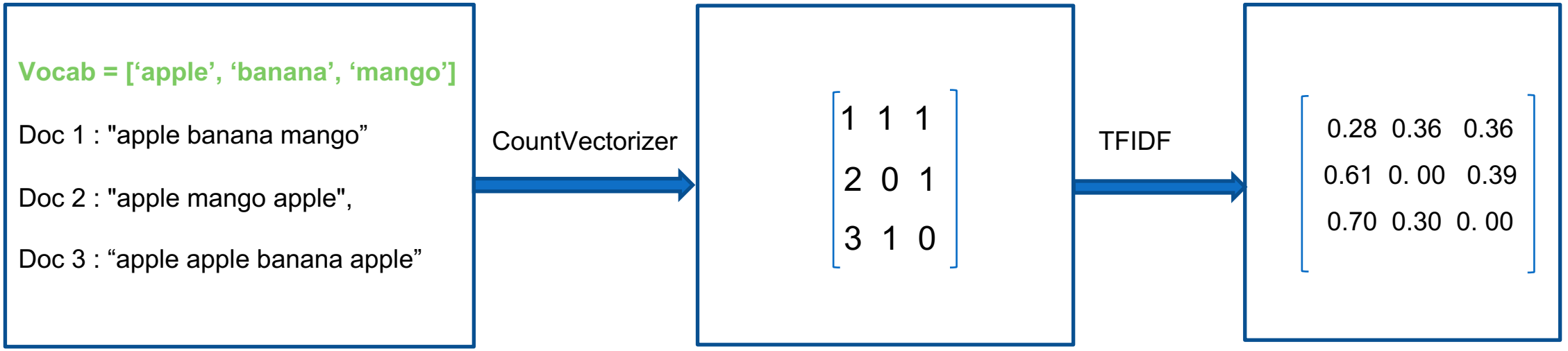
# Preprocessing

Cleaning your documents

- Tokenizing the document.
- Stop words removal.
- Stemming: merging words that are equivalent in meaning

```
['Donald Trump is an American businessman, author, politician and current President-elect of the United
States. Since 1971 he has chaired The Trump Organization, the principal holding company for his real
estate ventures and other business interests. During his business career, Trump has built office towers,
hotels, casinos, golf courses, and other branded facilities worldwide. He was elected as the 45th U.S.
president in the 2016 election on the Republican ticket, defeating Democratic nominee Hillary Clinton,
and is scheduled to take office on January 20, 2017. At 70 years old, Trump will be the oldest person to
assume the presidency.']
Vocabulary:
['person', 'interests', 'republican', 'elect', 'trump', 'estate', 'chaired', '2016', 'years', 'states',
'facilities', 'donald', 'election', 'golf', 'oldest', 'principal', 'united', 'built', 'author', 'since',
'current', 'take', 'holding', 'courses', '2017', 'hotels', 'real', '45th', 'business', 'towers',
'defeating', 'company', 'hillary', 'clinton', 'nominee', 'businessman', 'president', 'ticket', 'office',
'branded', 'politician', 'presidency', 'elected', 'ventures', 'january', '1971', 'career', 'american',
'casinos', 'assume', 'scheduled', 'organization', 'worldwide', 'democratic']
```

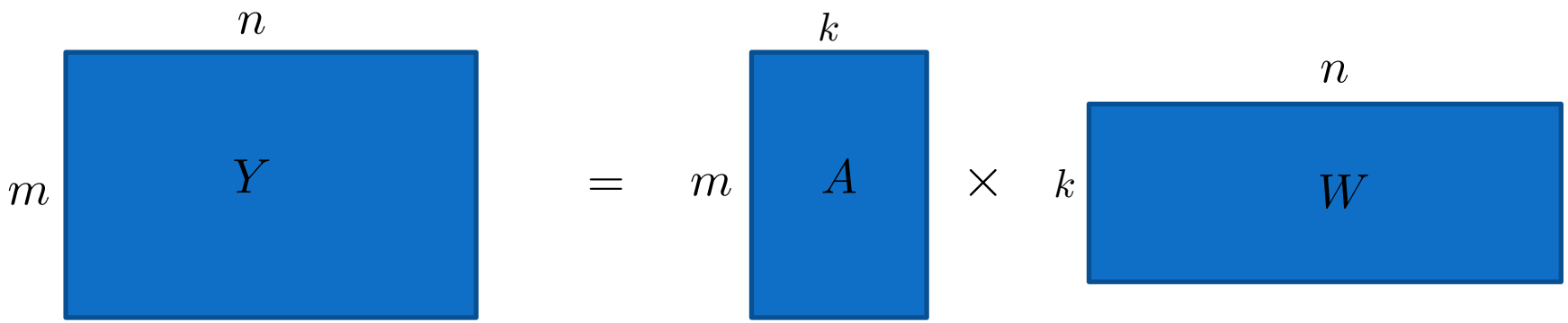
# Representation of corpus of documents




- Just counting word gives too much importance on words that are common and not document specific ( eg. he, she, said, man)
- TFIDF reweights the counts by number of documents a word appears in.
- TF weight of term t in documents d : $\propto f_{t,d}$
- IDF of term t : $\propto \frac{1}{n_t}$
- TFIDF = TF*IDF

$N$ : Total documents
$n_t$ : Number of documents with term $t$
$f_{t,d}$ : frequency of $t$ in $d$

# NMF in Topic Modeling




$$Y_{i,:} = A_{i,1}W_{1,:} + A_{i,2}W_{2,:} + \cdots + A_{i,k}W_{k,:}$$

$\sum_{j=1}^{k} A_{i,j} = 1, \forall i$ (All rows of Y are convex combinations of rows of W)

**Goal**: Given Y find the non-negative factors A and W

*Assumptions* under which NMF is efficiently solvable:

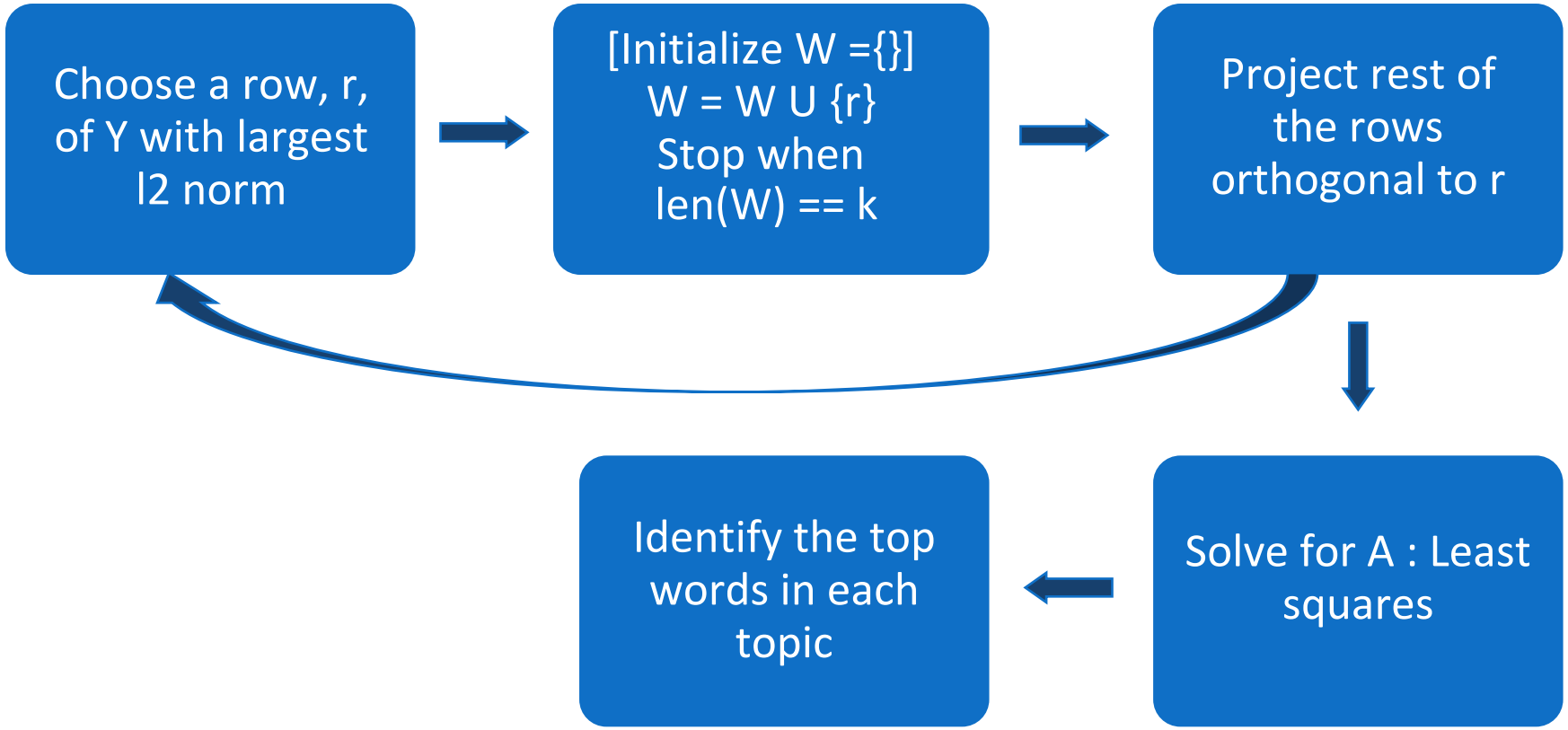
- W consists of some subset of columns of Y : Pure documents

  ( This means some k rows of A form an Identity matrix)

- Rows of W are not too close.

  (Each row is far from the convex hull of other rows = Topics are distinct)

# Robust Recursive NMF[1]

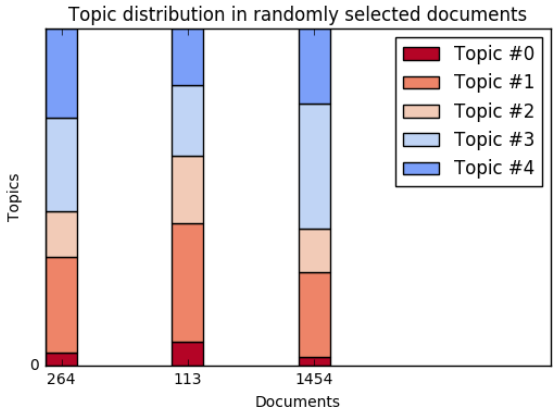



[1] Gillis, Nicolas, and Stephen A. Vavasis. "Fast and robust recursive algorithms for separable nonnegative matrix factorization." *IEEE transactions on pattern analysis and machine intelligence* 36.4 (2014): 698-714.

# Results on NIPS Dataset

Number of documents: 1500 Length of vocabulary: 12419

## Recursive NMF




**Topic 0** : ['neurosci', 'vocal', 'neuron', 'vocalization', 'template', 'auditory', 'memorized', 'nuclei', 'sparrow', 'bird', 'song'] **Speech Recognition**

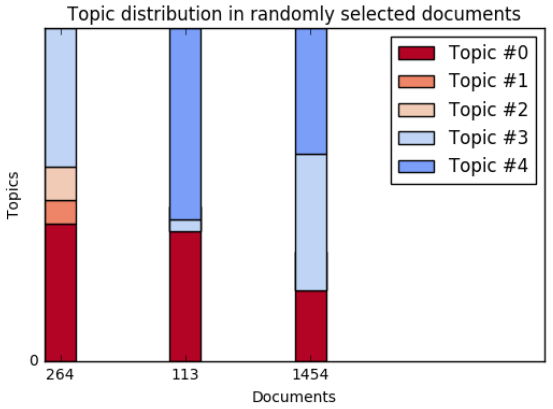
**Topic 1** : [‘adaptive', 'vector', 'basis', 'neural', 'handwriting', 'atypical', 'output', 'word', 'network', 'recognizer', 'character', 'adaptation’] **Text Recognition**

**Topic 2 :** ['contextual', 'visual', 'axis', 'responses', 'effect', 'ground', 'ripple', 'ill', 'region', 'cell', 'lamme, 'texture', 'iii', 'border’] **Image Processing**

**Topic 3 :** ['false', 'data', 'german', 'assert, 'dollar', 'target', 'virtual', 'trading’, 'performance', 'financial', 'symmetry', 'learning', 'market’] **Finance**

**Topic 4 :** [ 'learning', 'ortho', 'model', 'transformation', 'subspace', 'algorithm', 'centroid', 'discriminant', 'vector', 'tangent'] **Theoretical ML**

## SKlearn NMF




**Topic 0** : ['word', 'recognition', 'recurrent', 'speech', 'pattern’,'layer', 'neural', 'output', 'hidden', 'weight', 'training', 'input', 'network'] **RNN**

**Topic 1** : ['response', 'excitatory', 'activity', 'signal', 'network', 'voltage', 'synapses', 'analog', 'chip', 'synaptic', 'circuit', 'cell''] **System design**

**Topic 2 :** ['pomdp', 'dynamic', 'optimal', 'robot', 'mdp', 'reward', 'algorithm', 'control', 'reinforcement', 'action', 'policy', 'learning'] **MDP**

**Topic 3 :** ['bayesian', 'density', 'probability', 'classifier', 'learning', 'training', 'likelihood', 'vector', 'parameter', 'gaussian', 'model'] **Probabilistic models**

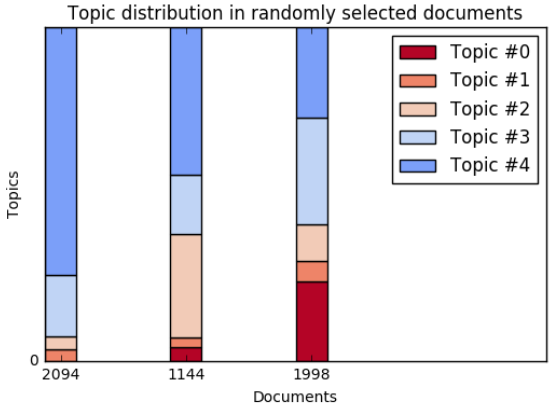
**Topic 4 :** ['receptive', 'pixel', 'recognition', 'spatial', 'map', 'view', 'orientation', 'direction', 'field', 'eye', 'images', 'model', 'cell', 'visual', 'motion', 'image'] **Image Processing**

# Results on AP news Dataset

Number of documents: 2243

Length of vocabulary: 37172

## Recursive NMF




**Topic 0** : ['firings', 'emaciated', 'infamous', 'freedom-loving', 'organized', 'perpetrators' 'workforce', 'unmask', 'insane'] **Crime News**

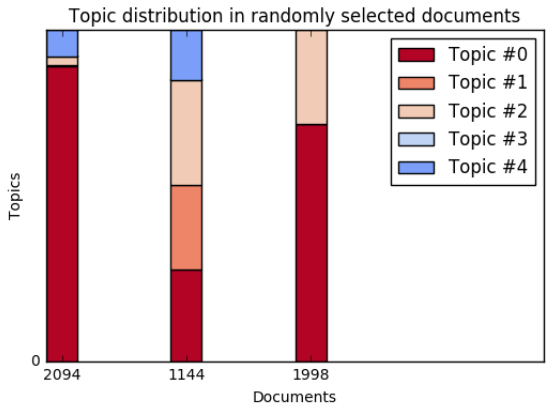
**Topic 1** :['protester', 'commitments', 'desegregation', 'life-saving', 'costume', 'feminism', 'smiling' ] **Social Issues**

**Topic 2 :** ['democratized', 'white-supremacist', 'outfitted', 'most-favored-nation', 'aids', 'preserving''] **Politics**

**Topic 3 :** ['restaurants', 'commutes', 'subway', 'retorts', 'wraps', 'natives', 'government-financed'] **Local news**

**Topic 4 :** ['acetominophen', 'capsules', 'athletes', 'cojuangco', 'gujarat', 'totalitarianism', 'galesburg'] **International news**

## SKlearn NMF




**Topic 0** : [armand', 'algirdas', 'algerie', 'saca', 'rica', 'toledo', 'bomb',' decades'] **International news**

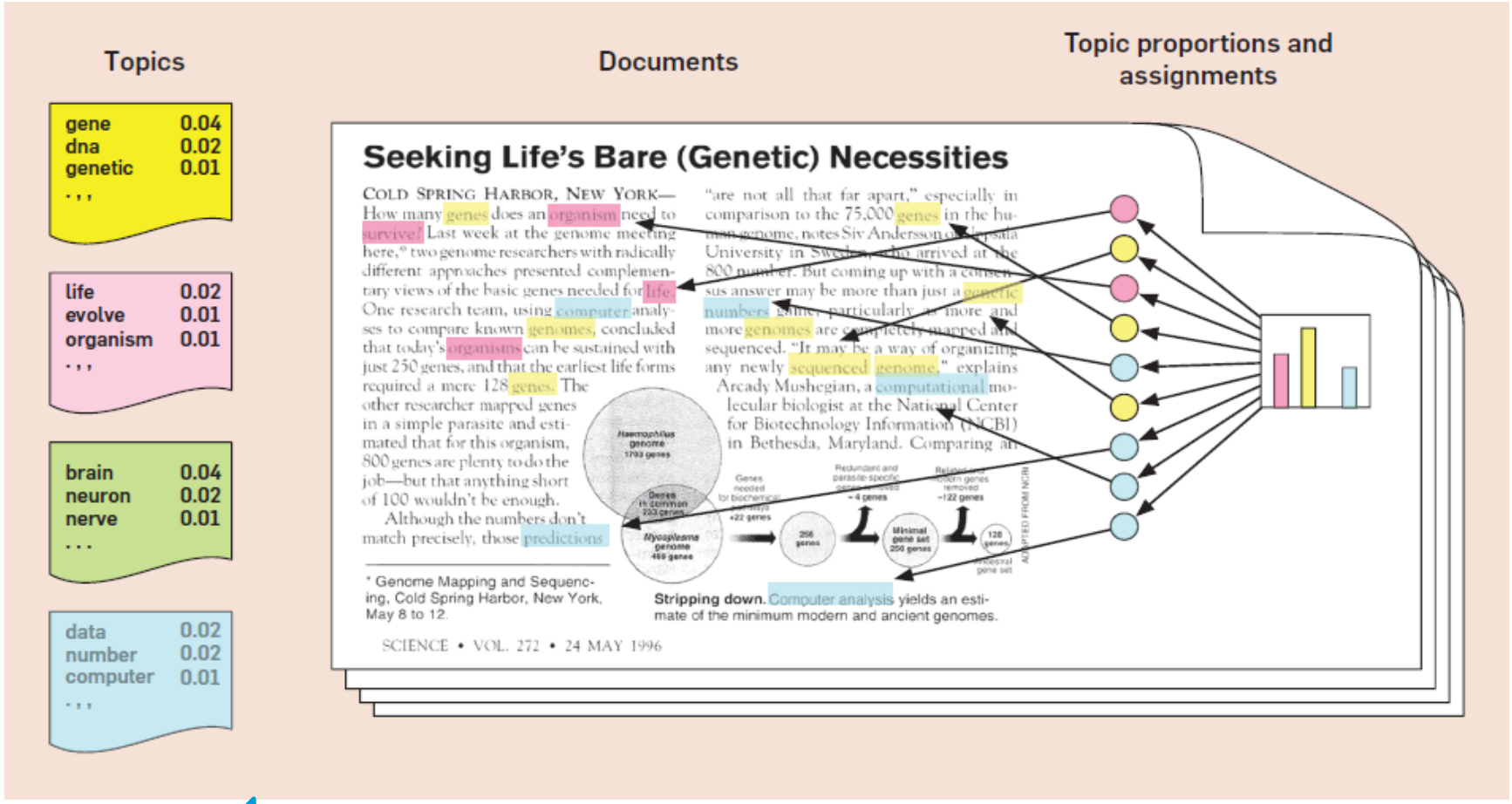
**Topic 1** : ['born', 'bypassed', 'declaration', 'lefty', 'success', 'traditionalist', 'coffeville', 'minnesota''] **National policies**

**Topic 2 :** ['newspaper', 'bypassed', 'undergraduate', 'marketplace', 'most-favored-nation', 'contained'] **Economics/Education**

**Topic 3 :** ['macnicol', 'feminism', "d'alene", 'chile', 'life-saving', 'smiling', 'costume'] **Social issues**

**Topic 4 :** ['declaration', 'commentators', 'last-stage','speculators', 'relations', 'non-profit', 'shout', 'possibility'] **Financial news**

# LDA (Latent Dirichlet allocation)



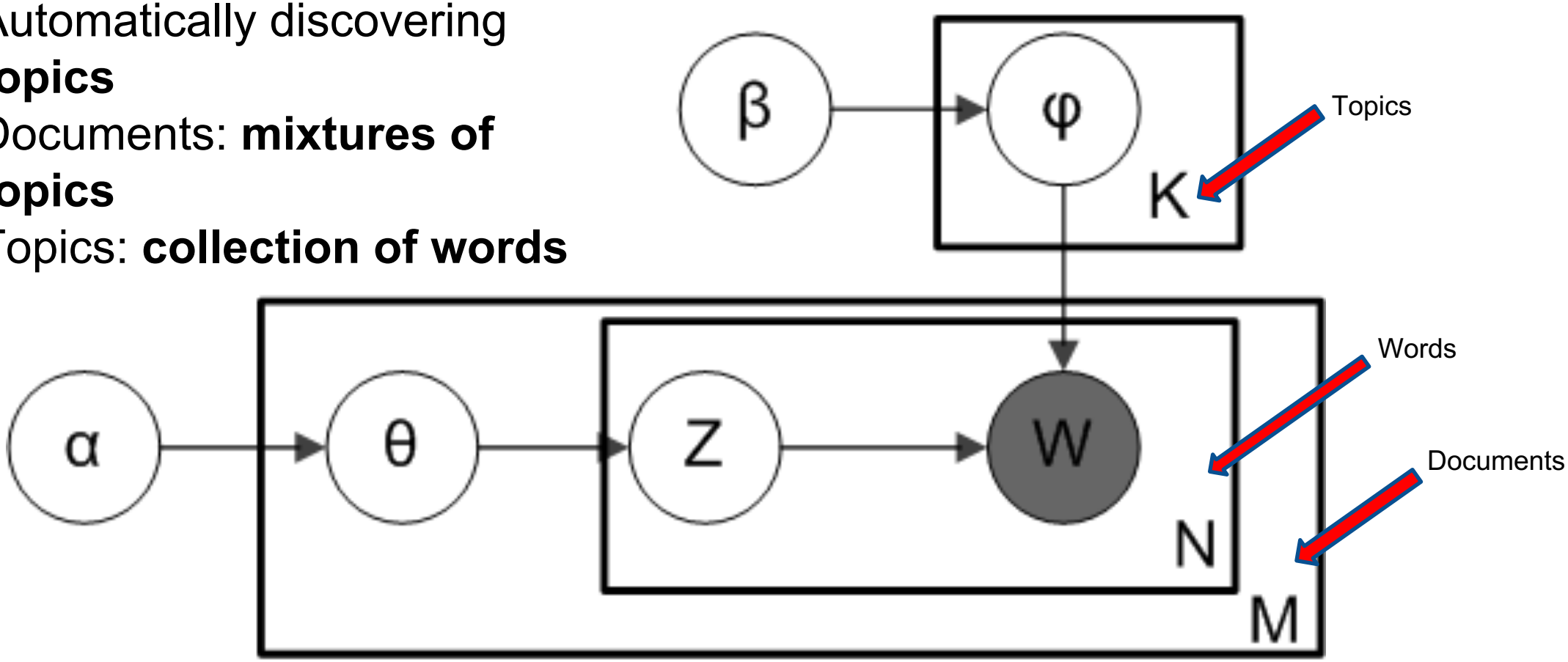

LDA Generates Topics

Takes a collection of documents and learns a model that describes it best ..

Courtesy: Probabilistic Topic Models by David M. Blei

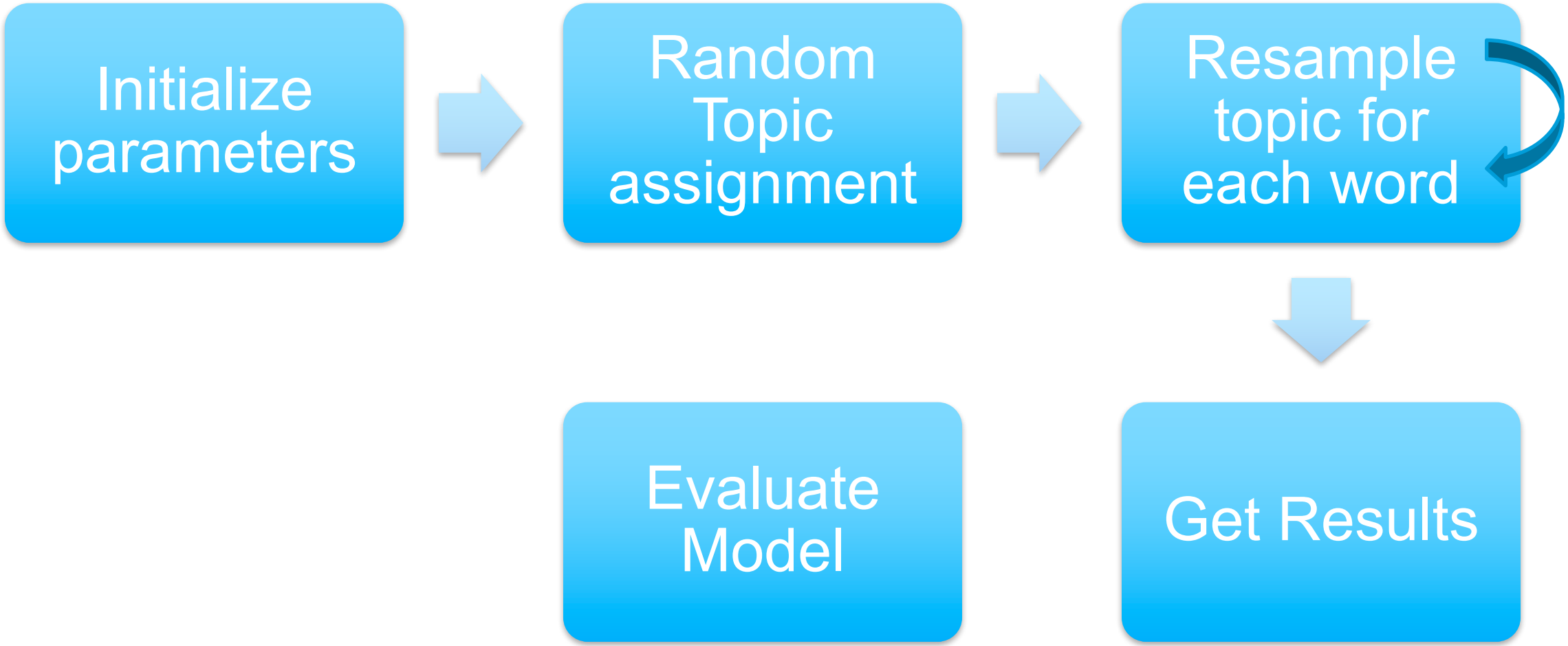
# LDA in Topic Modeling

- Automatically discovering **topics**
- Documents: **mixtures of topics**
- Topics: **collection of words**

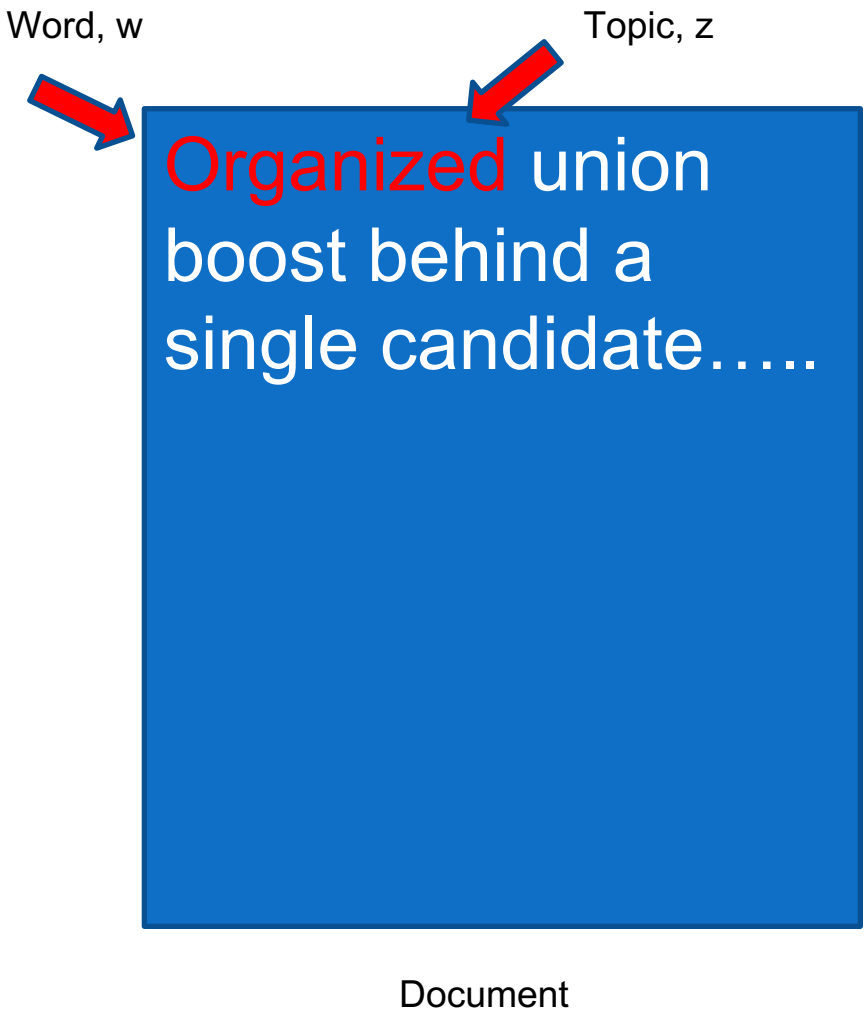



Courtesy:
Wikipedia.org

# LDA Algorithm



# LDA: Generation of Topics




For each word in each document:

Assign topic randomly

Resample based on:

- How prevalent is that word across topics?
- How prevalent are topics in the document?

# Topic words for 6 cluster:

## Topic 1

- Percent
- **Million**
- Year
- **Billion**
- **Market**
- New
- **Company**
- **Stock**
- **Prices**
- last..

## Topic 2

- **Government**
- **President**
- Soviet
- United
- **States**
- **Party**
- Bush
- Union
- new
- Also..

## Topic 3

- One
- **Year**
- Years
- **People**
- Time
- Old
- New
- Two
- **School**
- **family**..

## Topic 4

- **Court**
- **Federal**
- **Case**
- **Attorney**
- **Judge**
- State
- Department
- Trial
- Drug
- office..

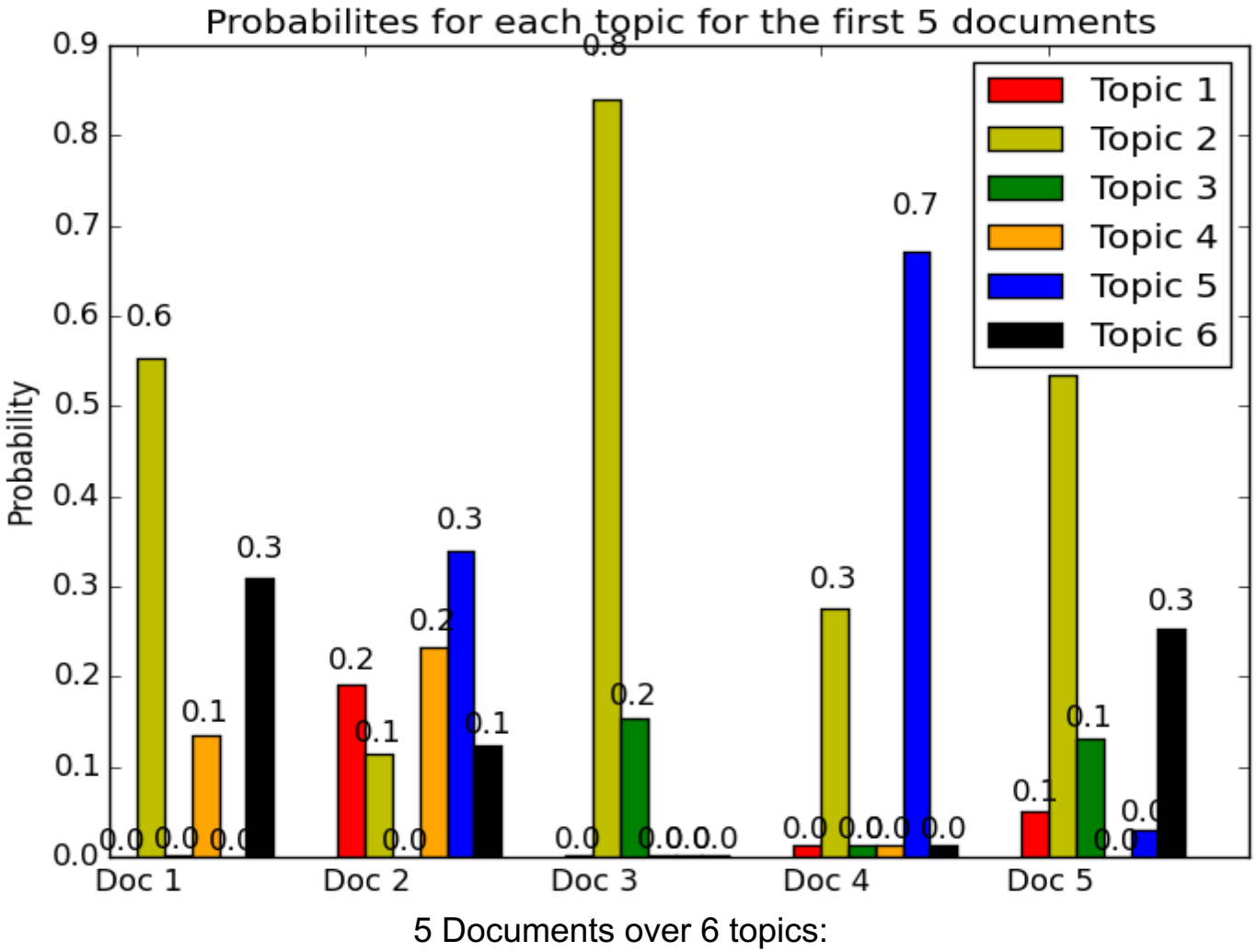
## Topic 5

- People
- **Water**
- **One**
- **Health**
- New
- **Study**
- Could
- Officials
- State
- Area..

## Topic 6

- **Police**
- **People**
- Government
- Two
- One
- **Military**
- **Officials**
- **Killed**
- Army
- since..

# Results: Topic-Document Distribution




Doc 1: Seizure of tax payments made by U.S. businesses operating in Panama

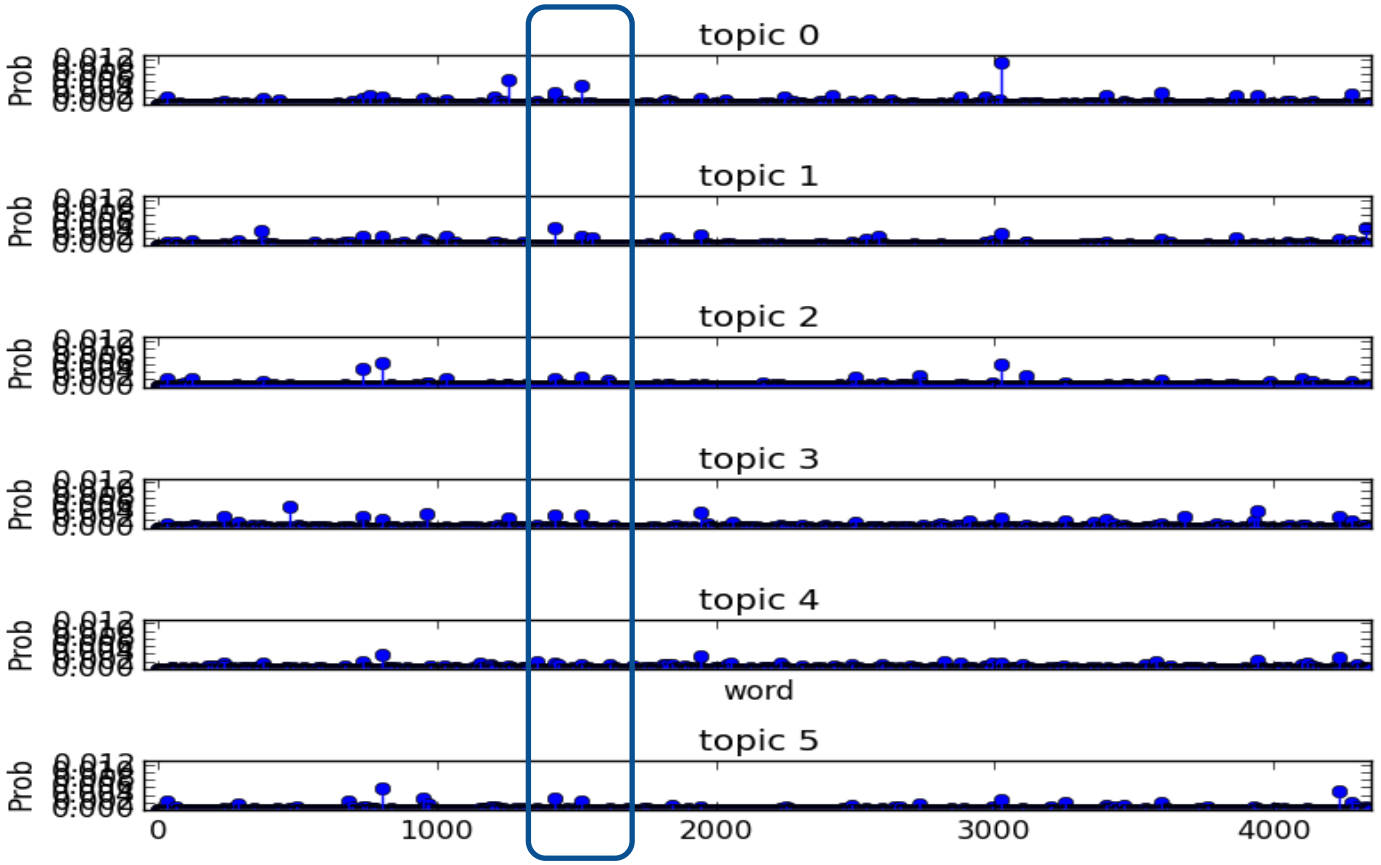
Doc 2: Protected species of alligator were smuggled

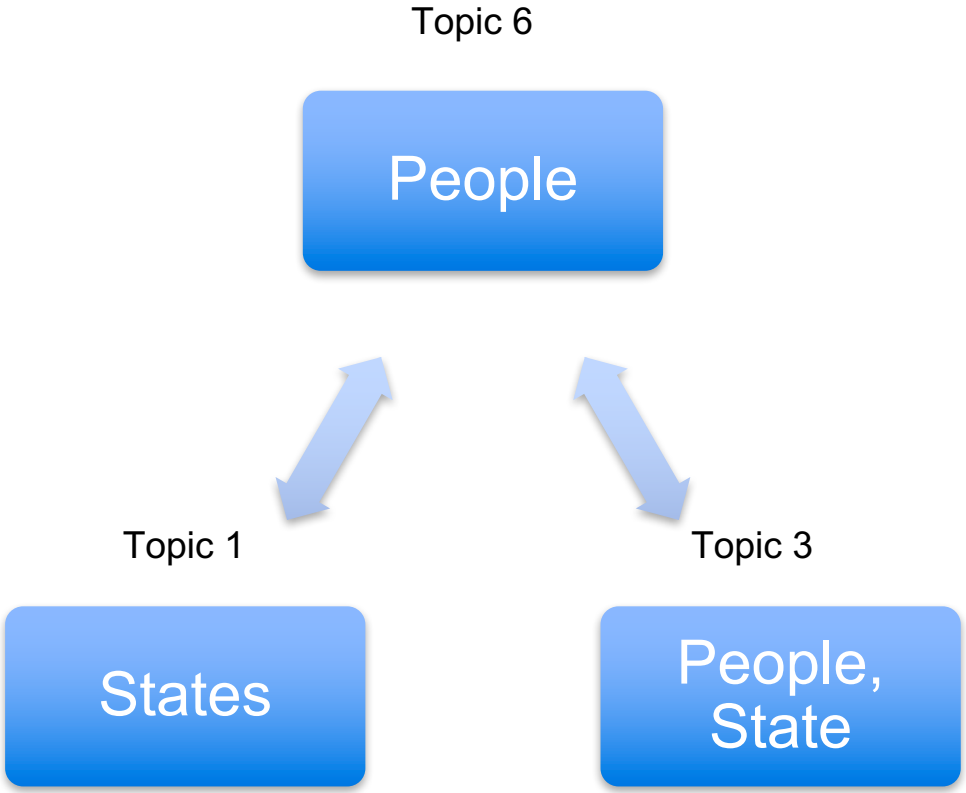
Doc 3: Democratic caucuses in Michigan

Doc 4:Death of National Front after party leader Jean-Marie Le Pen

Doc 5: Flood in some states in India

# Topic-Word Distribution



# A Sample Document:


- There will be no organized union boost behind a single candidate in Saturday's ***Democratic caucuses in Michigan***, a state where union members can wield more clout than almost anywhere else. While national labor leaders are assuming Michael Dukakis will be the ***eventual nominee***, they are prevented from endorsing him by what appears to be growing rank-and-file support for Jesse Jackson, who has gotten more union votes than any of the other ***candidates in primaries*** so far. Richard Gephardt also has considerable union support…….

# Evaluate model

- Hard: Unsupervised learning. No labels.
- Human-in-the-loop
- **Word intrusion:** Can human find the added word

# Reference


- Introduction to Topic Modeling in Python by Christine Doig
- Probabilistic Topic Models by David M. Blei
- Wikipedia.org