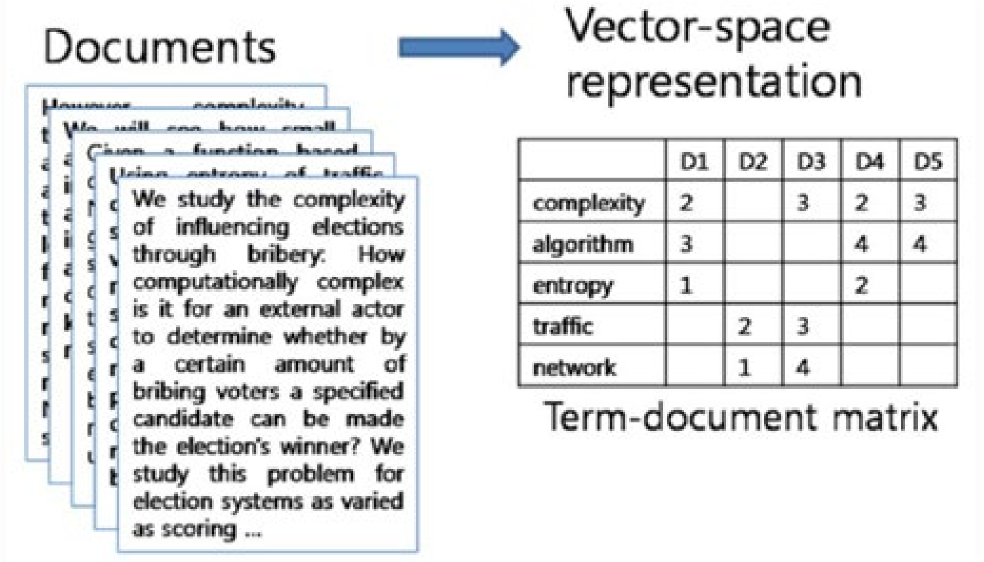
# Dimensional reduction

SVD

# Dimensional Reduction

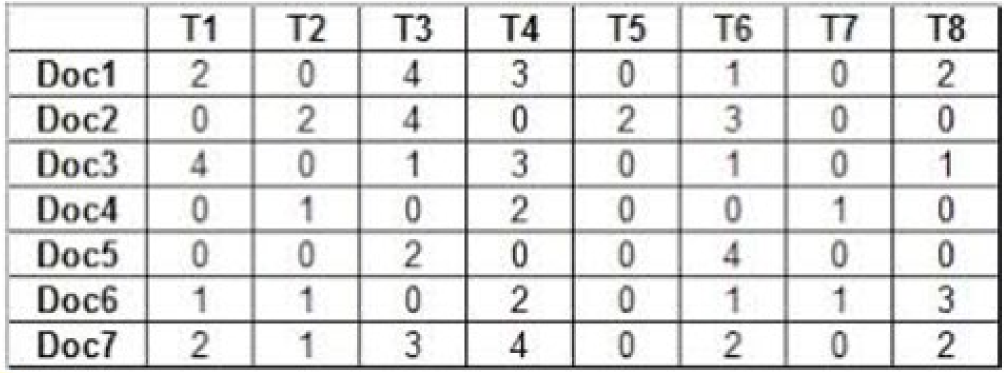


We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

| | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| complexity | 2 | | 3 | 2 | 3 |
| algorithm | 3 | | | 4 | 4 |
| entropy | 1 | | | 2 | |
| traffic | | 2 | 3 | | |
| network | | 1 | 4 | | |

Term-document matrix

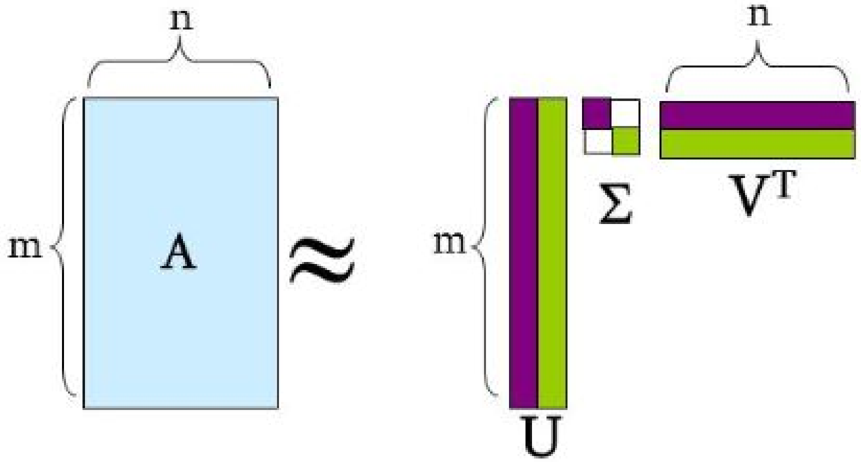
- **Sparse**
- **High dimension**
- **When lots of documents**



| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 |
|---|---|---|---|---|---|---|---|---|
| Doc1 | 2 | 0 | 4 | 3 | 0 | 1 | 0 | 2 |
| Doc2 | 0 | 2 | 4 | 0 | 2 | 3 | 0 | 0 |
| Doc3 | 4 | 0 | 1 | 3 | 0 | 1 | 0 | 1 |
| Doc4 | 0 | 1 | 0 | 2 | 0 | 0 | 1 | 0 |
| Doc5 | 0 | 0 | 2 | 0 | 0 | 4 | 0 | 0 |
| Doc6 | 1 | 1 | 0 | 2 | 0 | 1 | 1 | 3 |
| Doc7 | 2 | 1 | 3 | 4 | 0 | 2 | 0 | 2 |

# Singular Value Decomposition

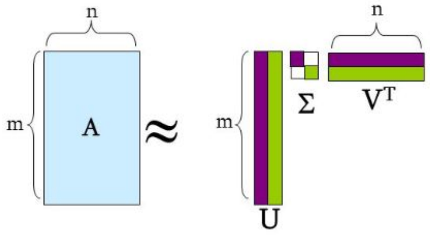
$$\mathbf{A} \approx \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^{\mathsf{T}}$$

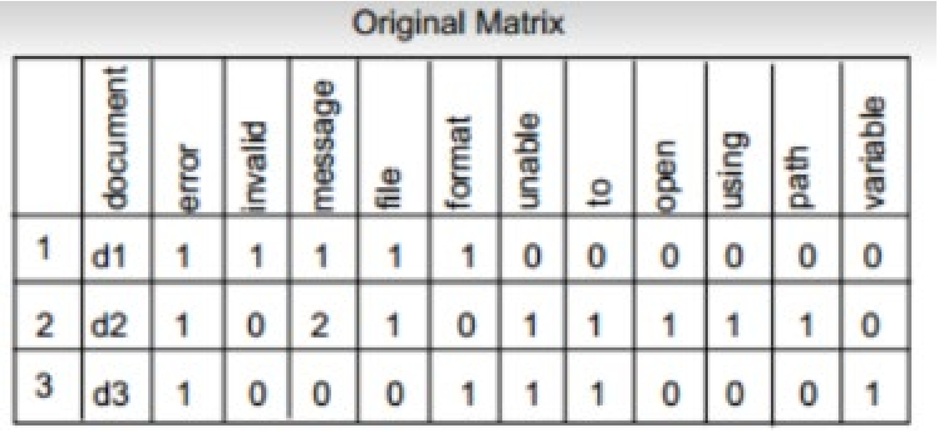


- **U,V**
  - Columns are orthogonal and unit vectors
- **Σ**
  - Entries (singular values) are positive and sorted in decreasing order of importance

# Singular Value Decomposition

$$\mathbf{A} \approx \mathbf{U}\Sigma\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \circ \mathbf{v}_i^T$$



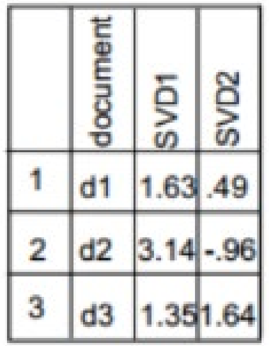


Original Matrix

| | | document | error | invalid | message | file | format | unable | to | open | using | path | variable |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | d1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | d2 | 1 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| 3 | d3 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |

3,11

A ≈



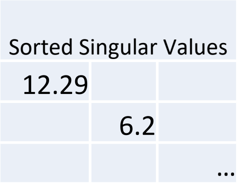
| | document | SVD1 | SVD2 |
|---|---|---|---|
| 1 | d1 | 1.63 | .49 |
| 2 | d2 | 3.14 | -.96 |
| 3 | d3 | 1.35 | 1.64 |

3,N

U

×

When N=2



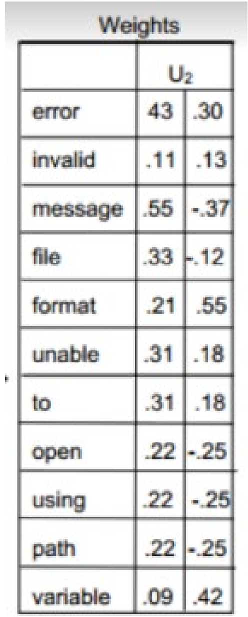
| Sorted Singular Values | | |
|---|---|---|
| 12.29 | | |
| | 6.2 | |
| | | ... |

N,N

Σ

×

T



Weights

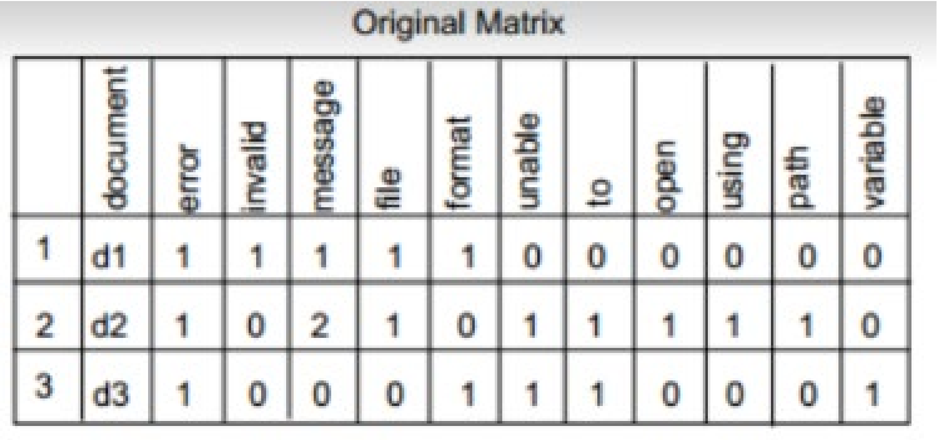
| | U₂ | |
|---|---|---|
| error | 43 | .30 |
| invalid | .11 | .13 |
| message | .55 | -.37 |
| file | .33 | -.12 |
| format | .21 | .55 |
| unable | .31 | .18 |
| to | .31 | .18 |
| open | .22 | -.25 |
| using | .22 | -.25 |
| path | .22 | -.25 |
| variable | .09 | .42 |

$V^T$

T

11,N

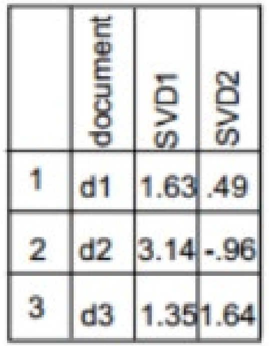
# Singular Value Decomposition



Original Matrix

| | document | error | invalid | message | file | format | unable | to | open | using | path | variable |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | d1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | d2 | 1 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| 3 | d3 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |

- Dimensions reduced from 11 to N=2



| | document | SVD1 | SVD2 |
|---|---|---|---|
| 1 | d1 | 1.63 | .49 |
| 2 | d2 | 3.14 | -.96 |
| 3 | d3 | 1.35 | 1.64 |

( 3,N )

X

Sorted Singular Values

| | | |
|---|---|---|
| 12.29 | | |
| | 6.2 | |
| | | ... |

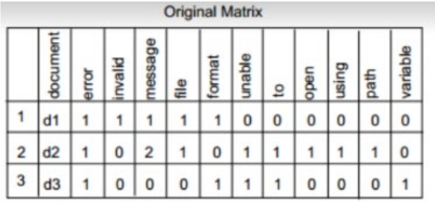
( N,N )

( 3,N )

New Matrix

Dense & low

# Dimension Reduction

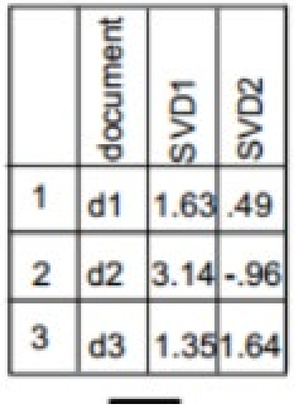


Original Matrix

| | document | error | invalid | message | file | format | unable | to | open | using | path | variable |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | d1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | d2 | 1 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |
| 3 | d3 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |

Step1. Apply SVD/PCA

You get SVDs/Concepts.

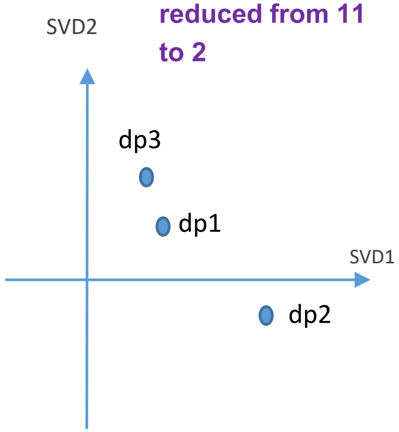


| | document | SVD1 | SVD2 |
|---|---|---|---|
| 1 | d1 | 1.63 | .49 |
| 2 | d2 | 3.14 | -.96 |
| 3 | d3 | 1.35 | 1.64 |

×

| Sorted Singular Values | |
|---|---|
| 12.29 | |
| | 6.2 |

**Dimensions reduced from 11 to 2**



**DataPoint1 = [1,1,1,1,1,0,0,0,0,0,0]**

**DataPoint2 = [1,1,2,1,0,1,1,1,1,1,0]**

**DataPoint3 = [1,0,0,0,1,1,1,0,0,0,1]**

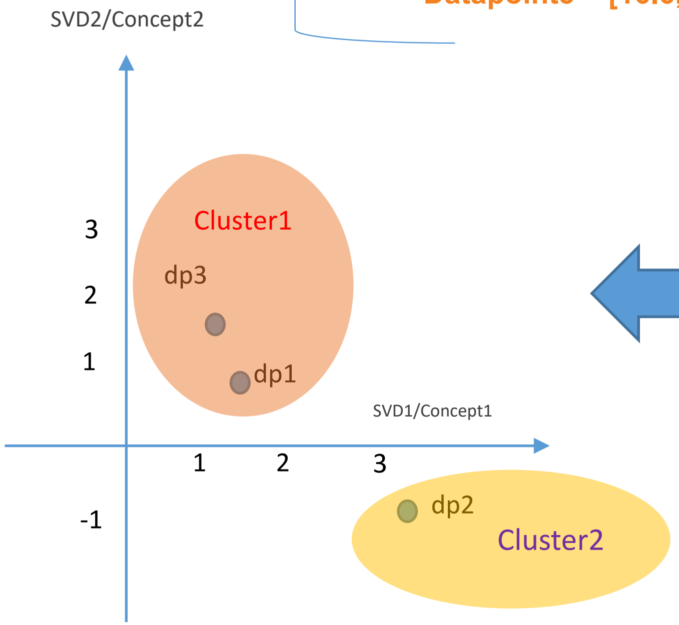
≈

**Datapoint1 = [20.1, 3.0]**

**Datapoint2 = [38.6,-5.95]**

**Datapoint3 = [16.6, 10.2]**

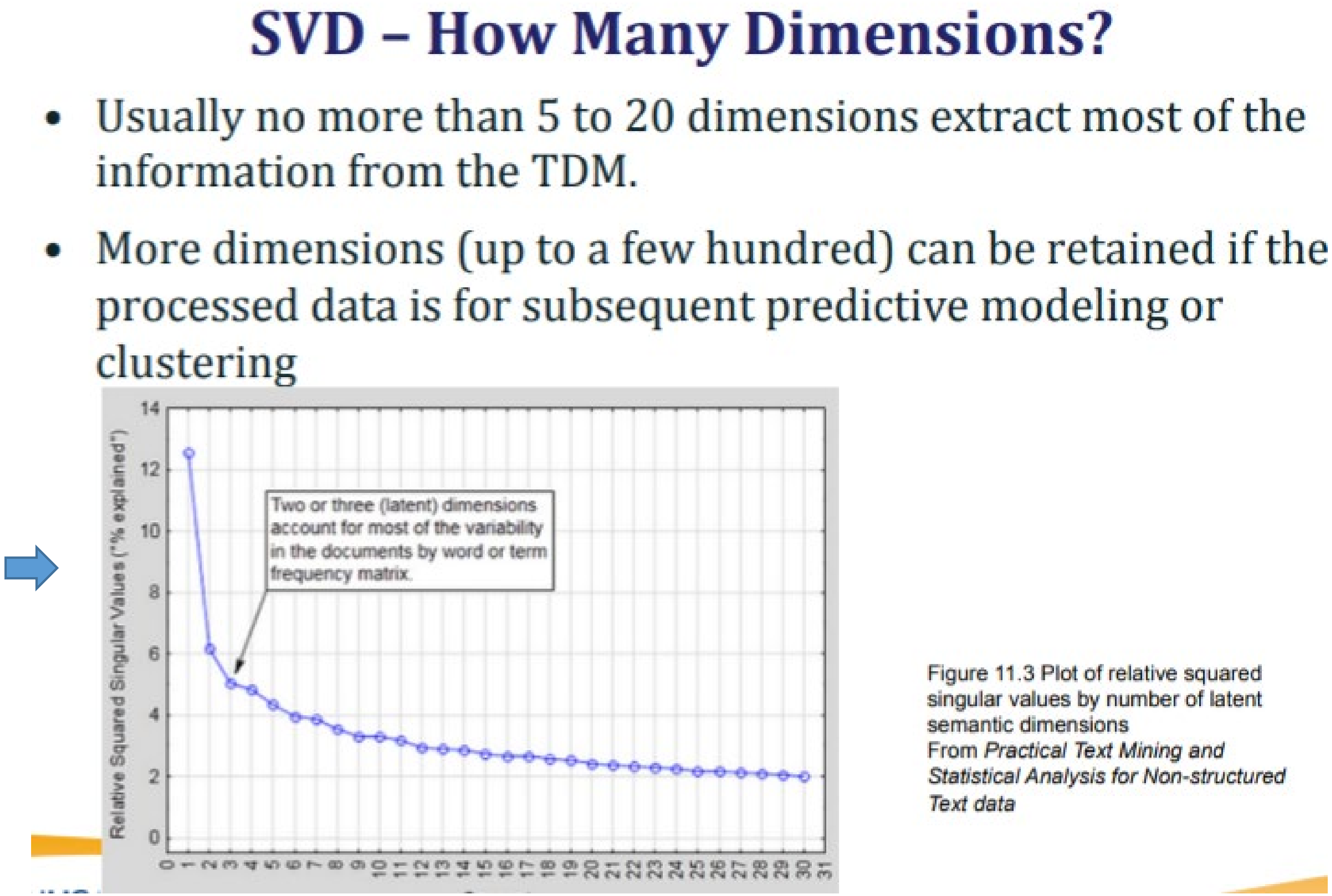
Step 2. Apply KM
Or other classifiers



**Concept# /SVD#** ≠ **Cluster#**

# Singular Value Decomposition

| Sorted Singular Values | | |
|---|---|---|
| 12.29 | | |
| | 6.2 | |
| | | … |



## SVD – How Many Dimensions?

- Usually no more than 5 to 20 dimensions extract most of the information from the TDM.
- More dimensions (up to a few hundred) can be retained if the processed data is for subsequent predictive modeling or clustering

Two or three (latent) dimensions account for most of the variability in the documents by word or term frequency matrix.

Figure 11.3 Plot of relative squared singular values by number of latent semantic dimensions
From *Practical Text Mining and Statistical Analysis for Non-structured Text data*