# Natural Language Processing

**Lecture 11**

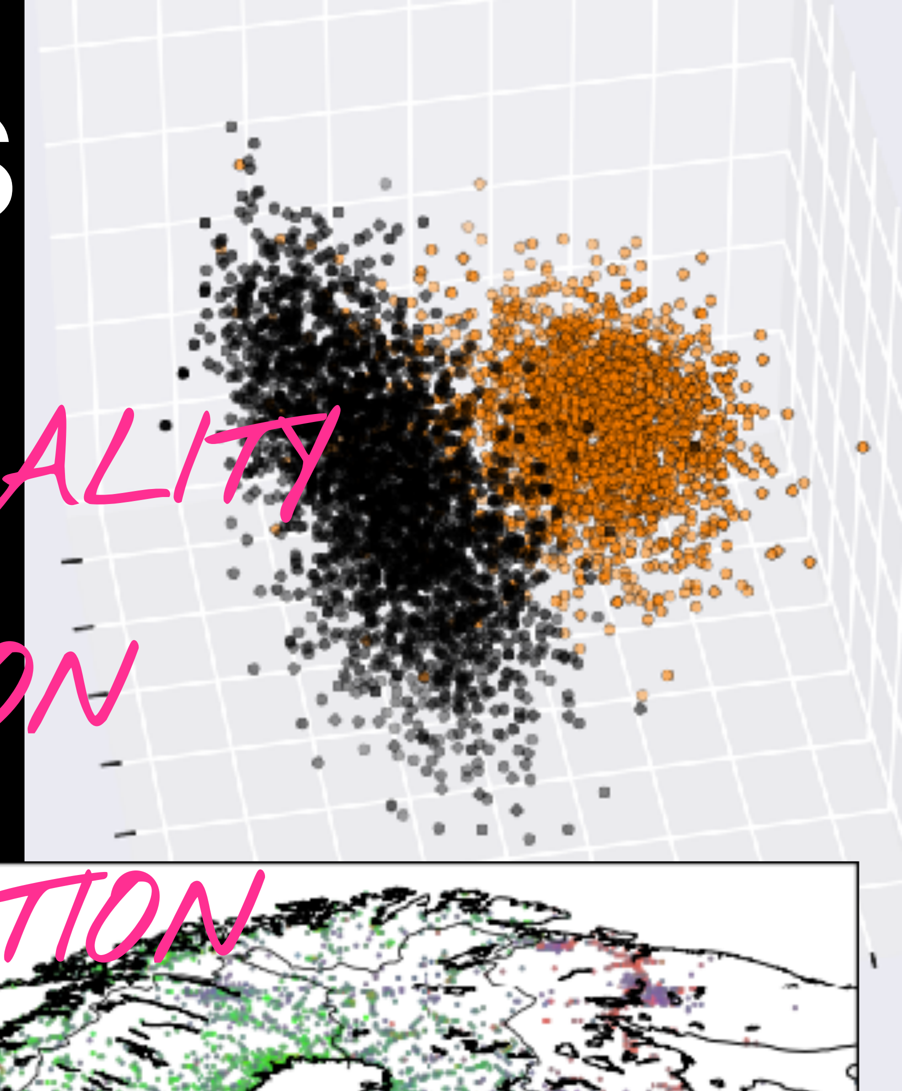
Dirk Hovy

dirk.hovy@unibocconi.it

@dirk_hovy

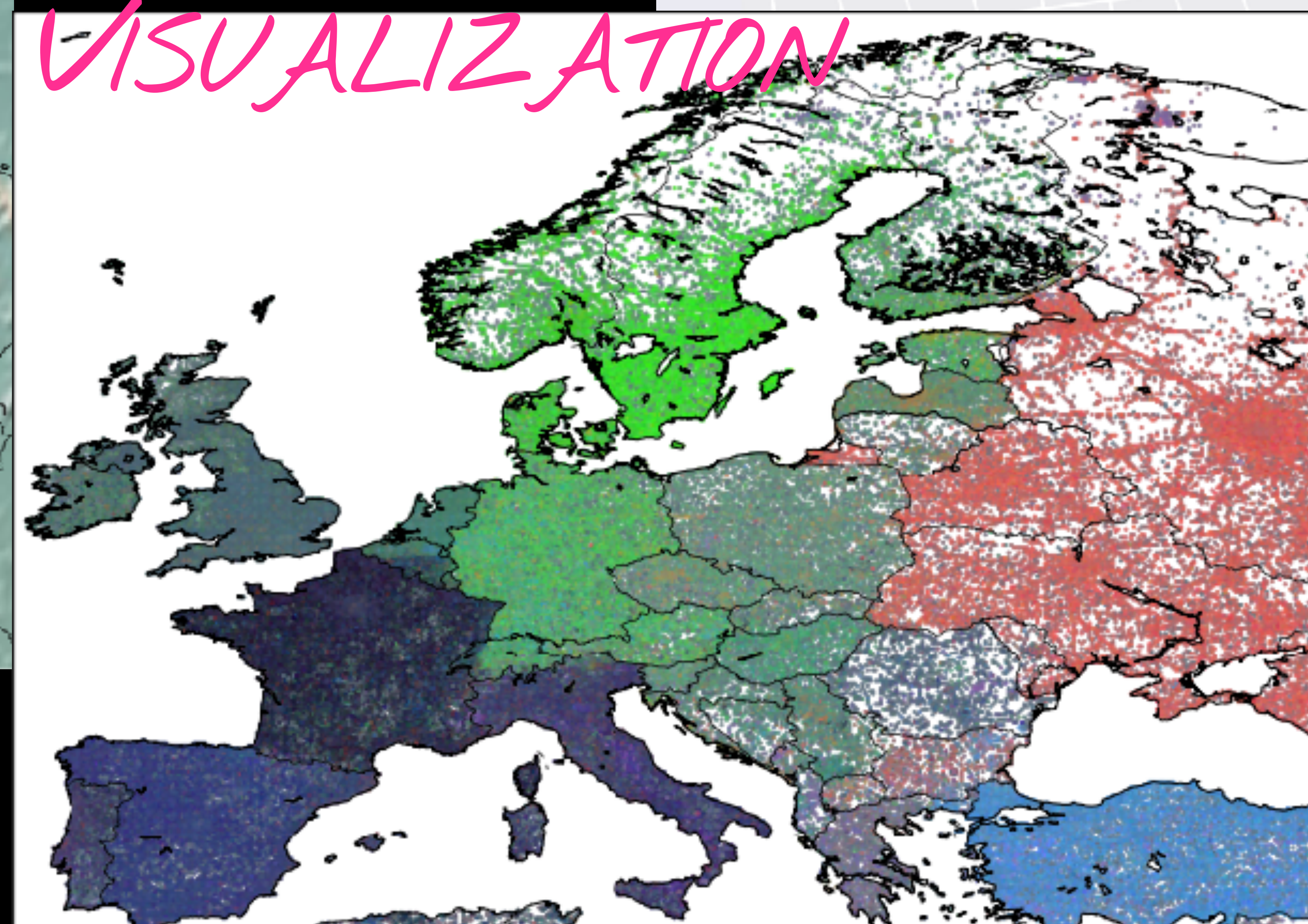
Bocconi

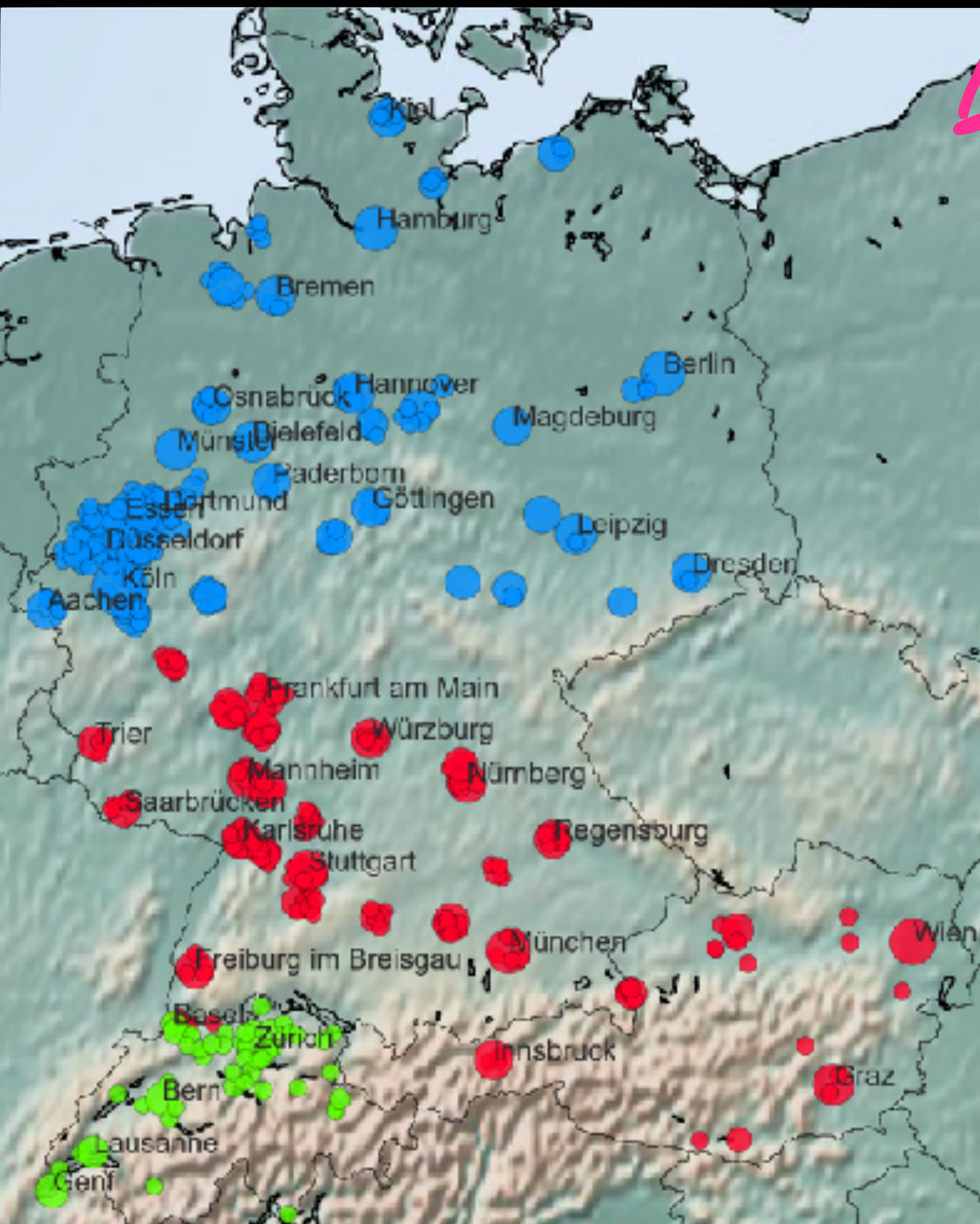
# Examples



DIMENSIONALITY REDUCTION

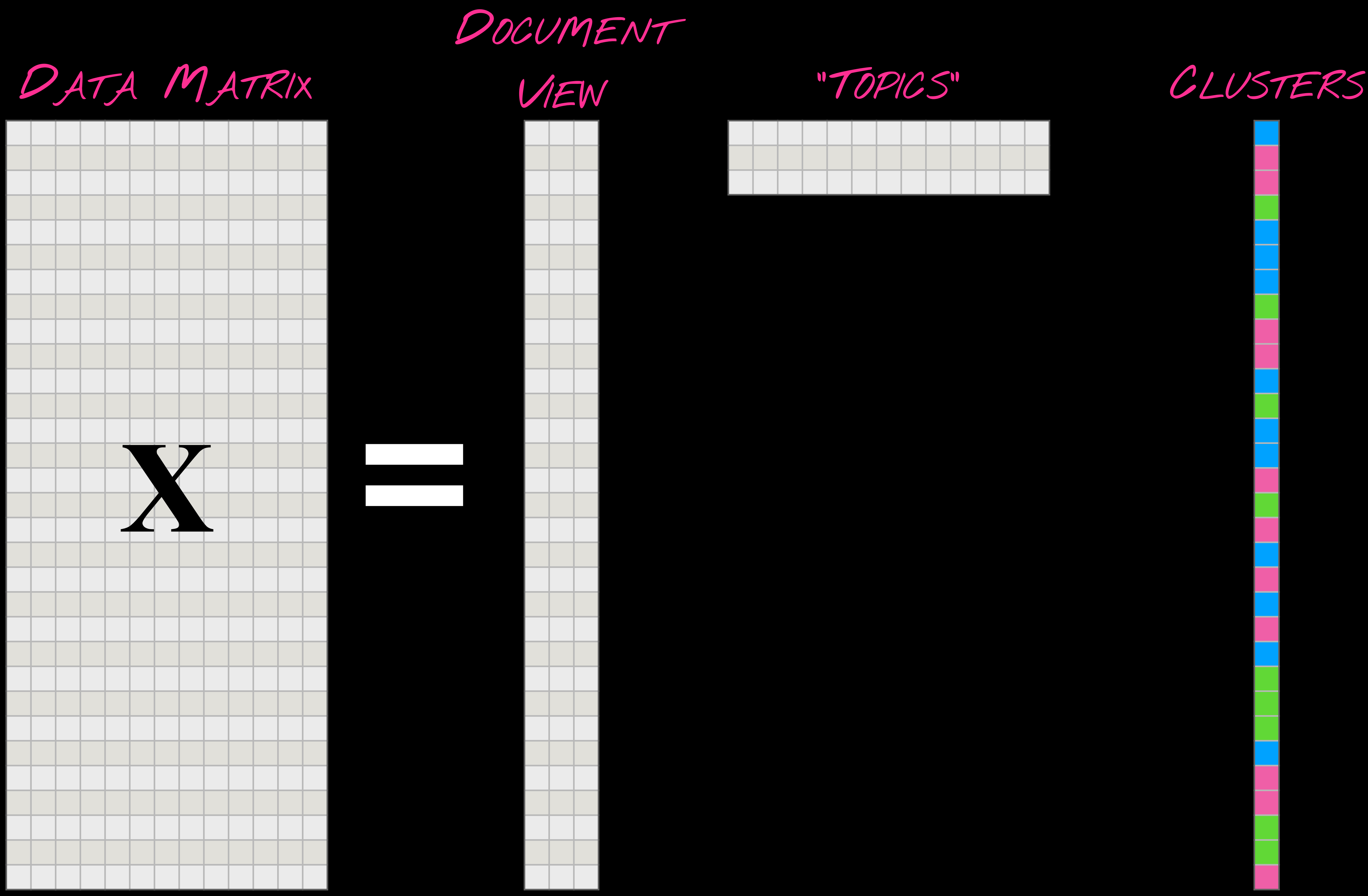
VISUALIZATION






CLUSTERING

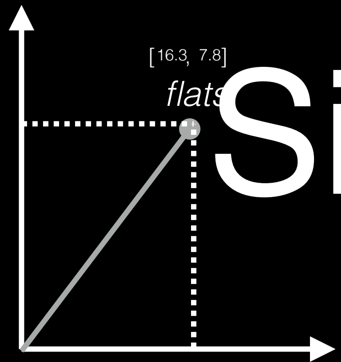
# Latent Dimensions



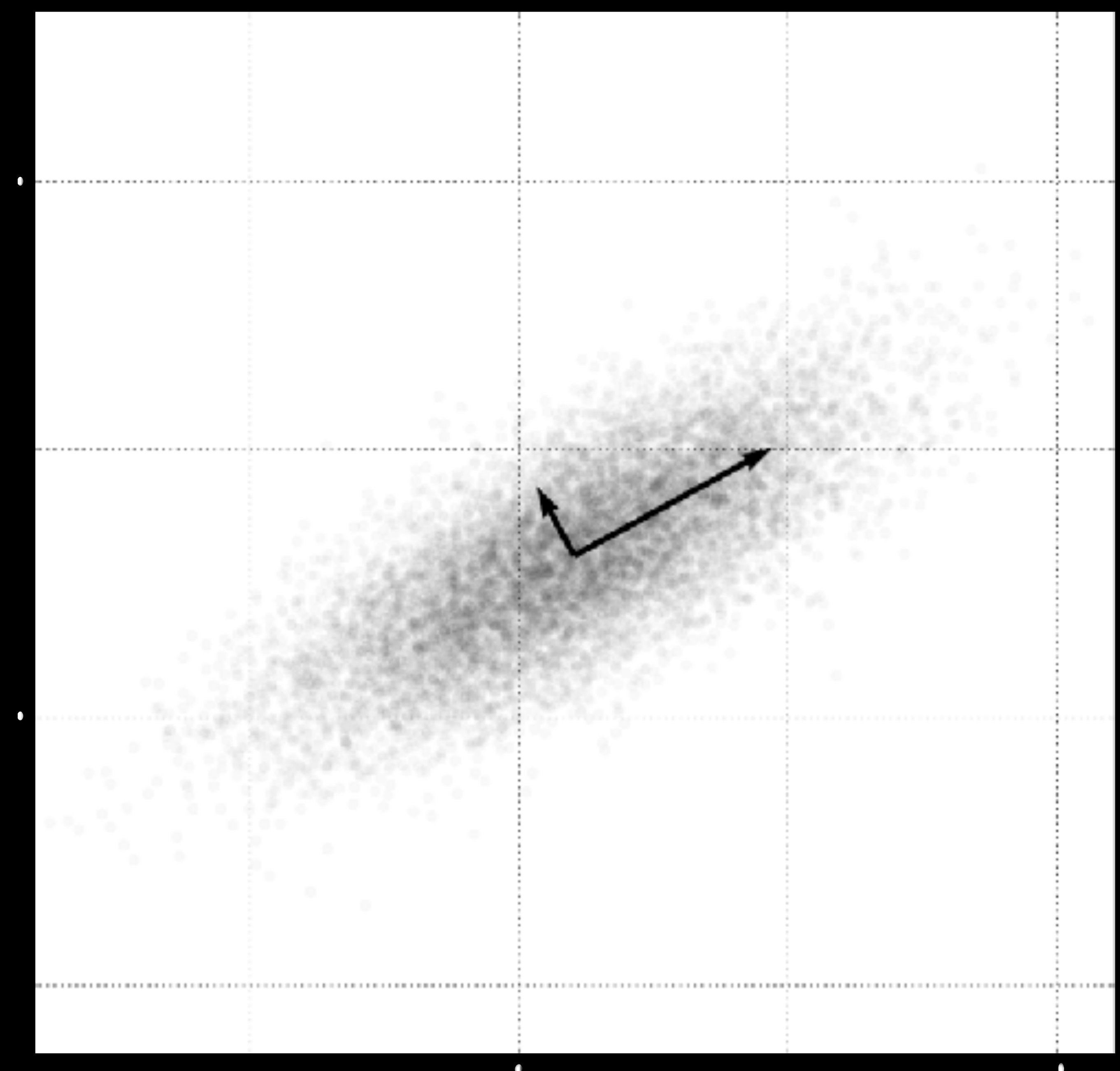
# Goals for Today

- Learn about **matrix factorization** and its use for **semantic similarity** and **visualization**
- Learn about ***k*-means** and **agglomerative clustering**
- Learn about **evaluation** criteria

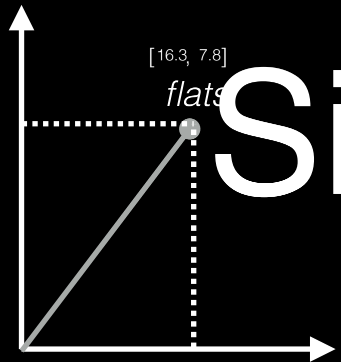
# Matrix Factorization

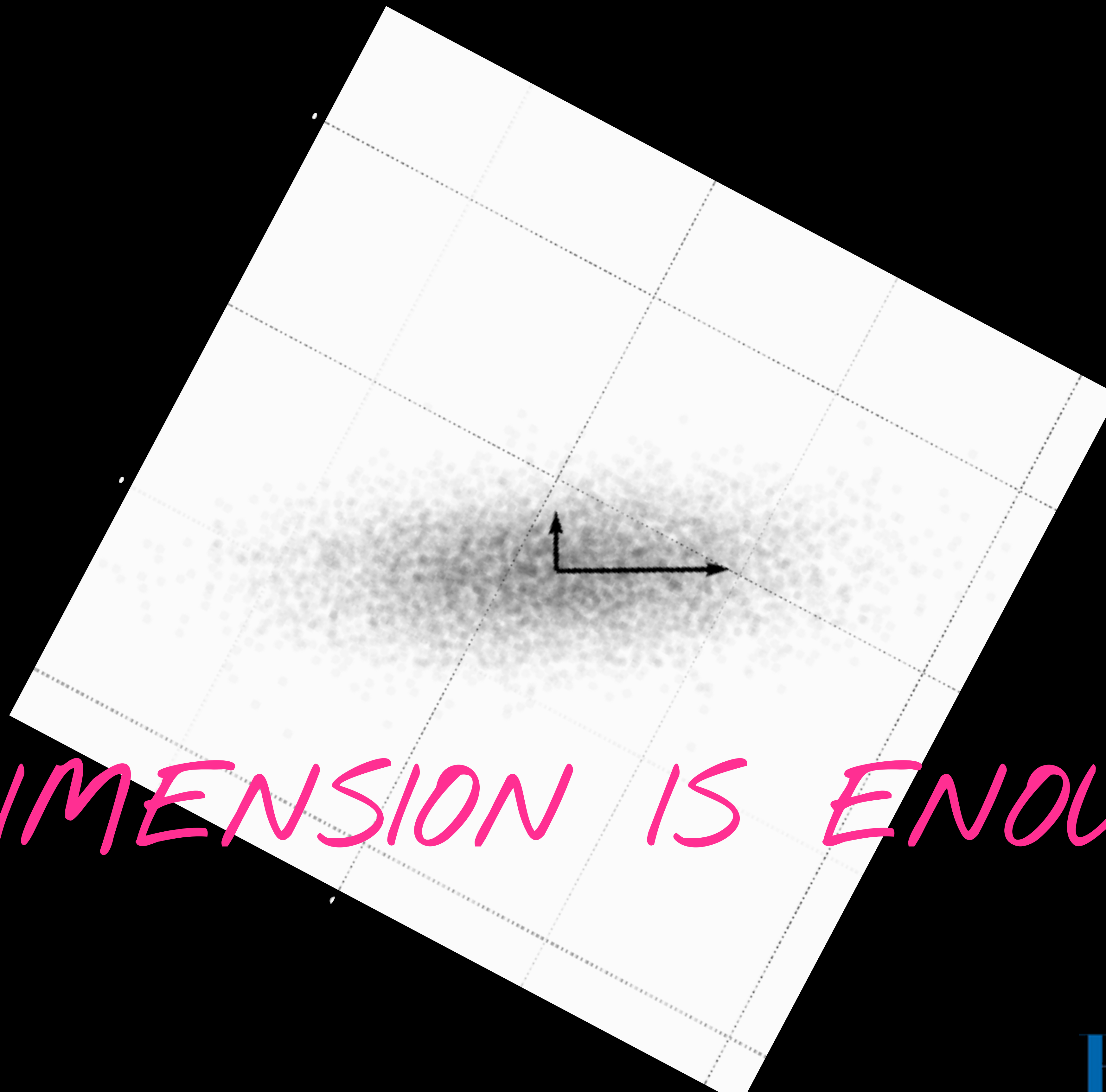



# Singular Value Decomposition

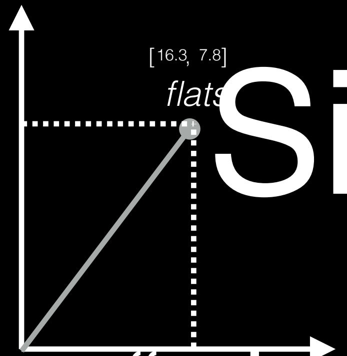


# Singular Value Decomposition



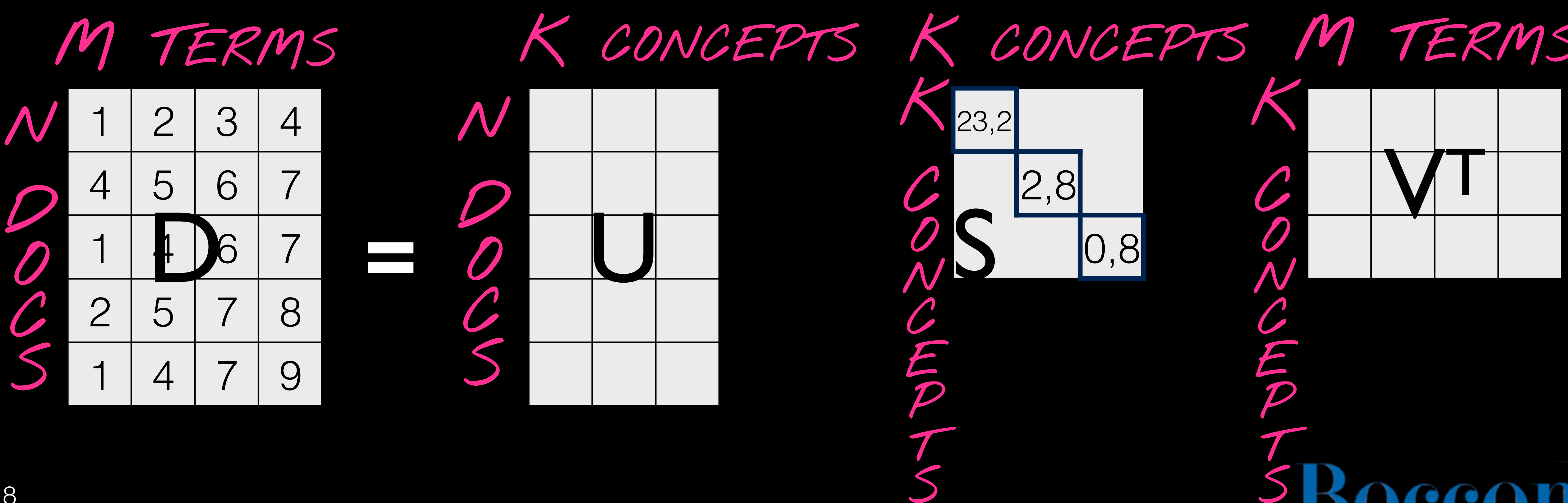



1 DIMENSION IS ENOUGH!

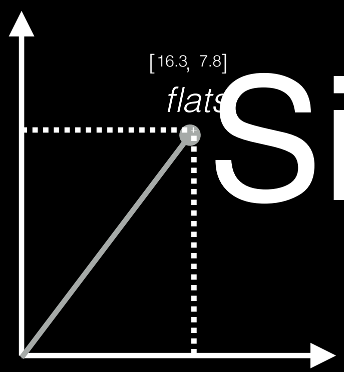



# Singular Value Decomposition

- "principal component analysis": discover the dimensions that matter
- idea: matrix is made up of few hidden dimensions
- Dimensions correspond to **documents**, **terms**, and latent **concepts**

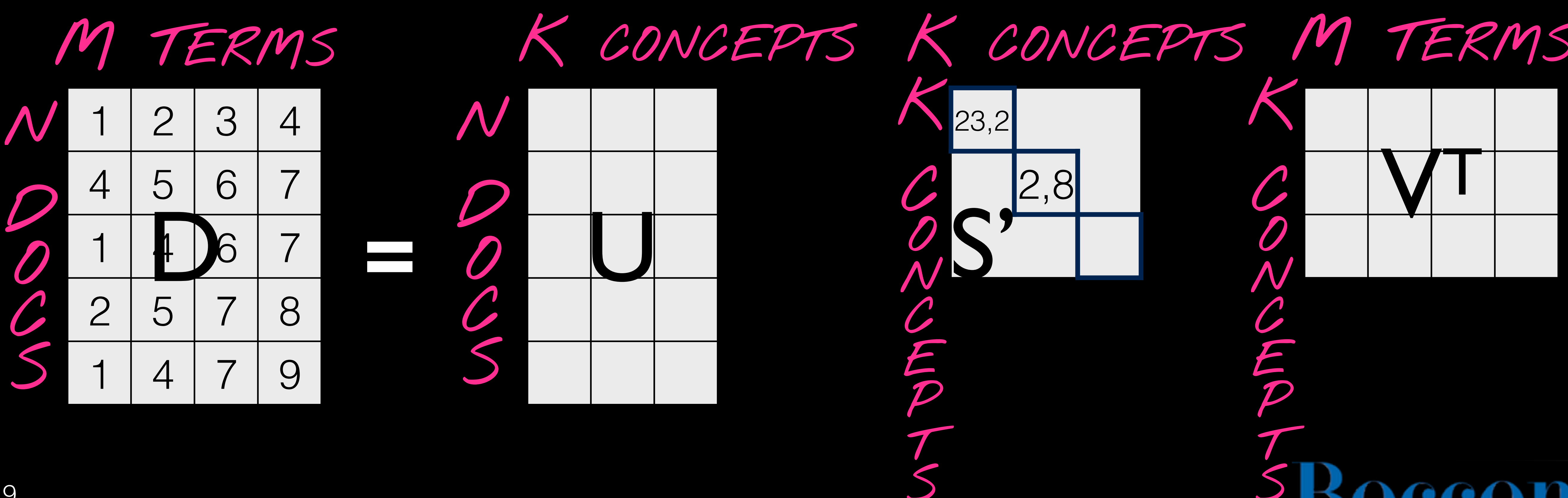


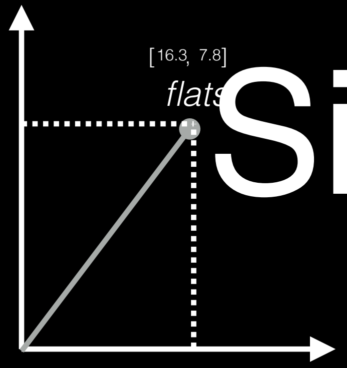

| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 4 | 5 | 6 | 7 |
| 1 | 4 | 6 | 7 |
| 2 | 5 | 7 | 8 |
| 1 | 4 | 7 | 9 |




# Singular Value Decomposition

- reduce principal components/concepts to smaller number



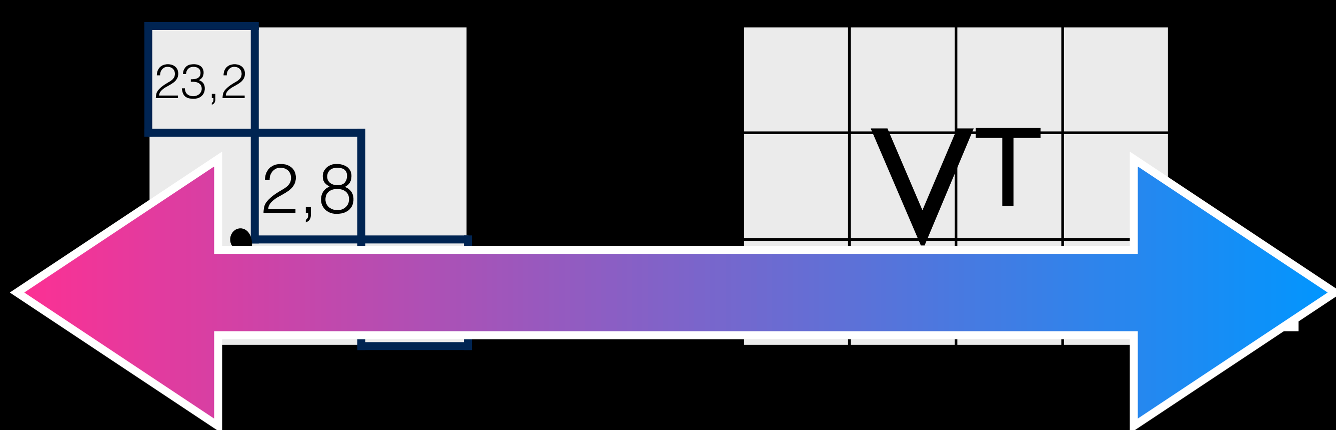



# Singular Value Decomposition

- reconstruct original matrix in new concept space:
  **Latent Semantic Analysis**

D

| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 4 | 5 | 6 | 7 |
| 1 | 4 | 6 | 7 |
| 2 | 5 | 7 | 8 |
| 1 | 4 | 7 | 9 |

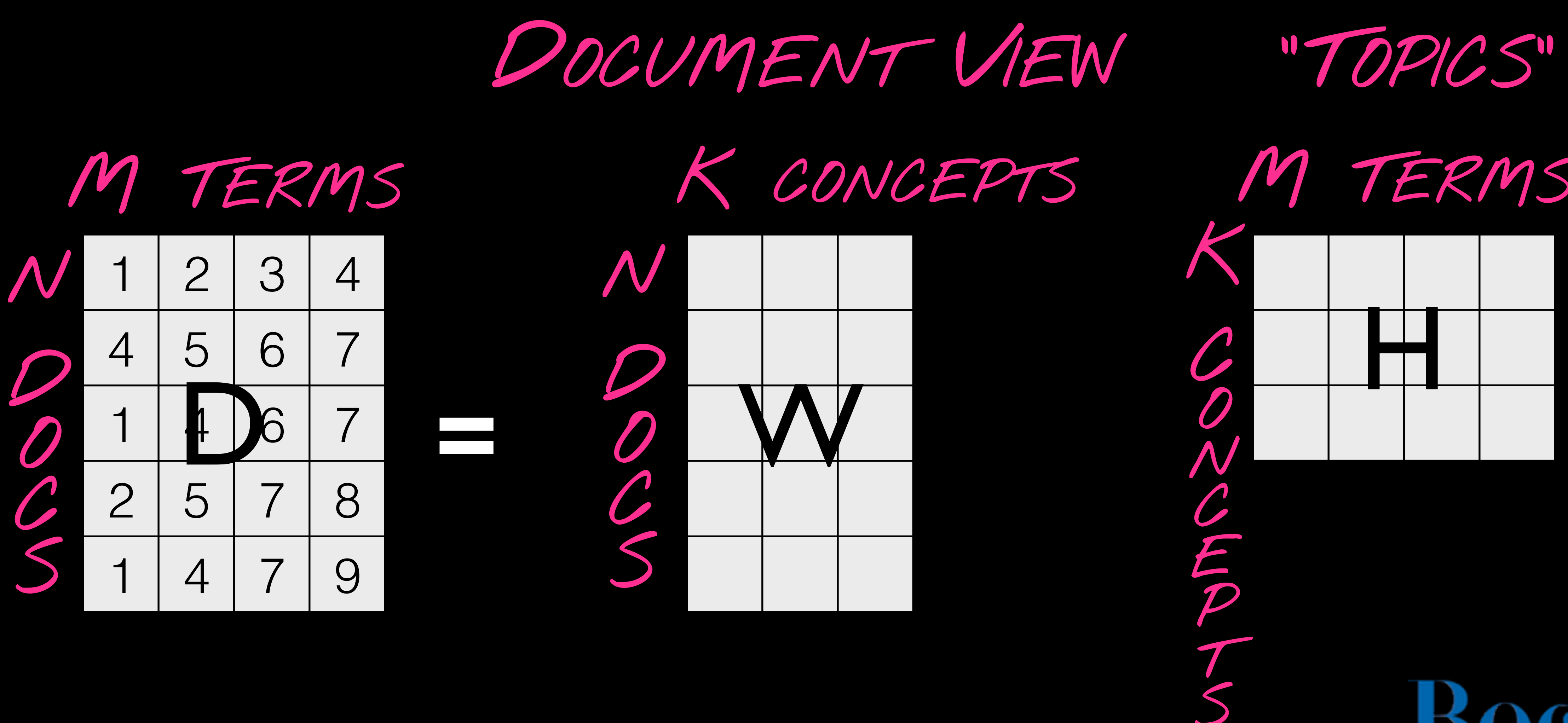



D’

| | | | |
|---|---|---|---|
| 0,9 | 2,1 | 3,2 | 3,8 |
| 3,9 | 5,1 | 6 | 6,9 |
| 1,1 | 3,8 | 5,9 | 7,2 |
| 2,2 | 4,7 | 6,9 | 8,2 |
| 0,8 | 4,3 | 7,1 | 8,8 |

# Non-negative Matrix Factorization

- Use only positive values
- Find approximation of two components






Bocconi

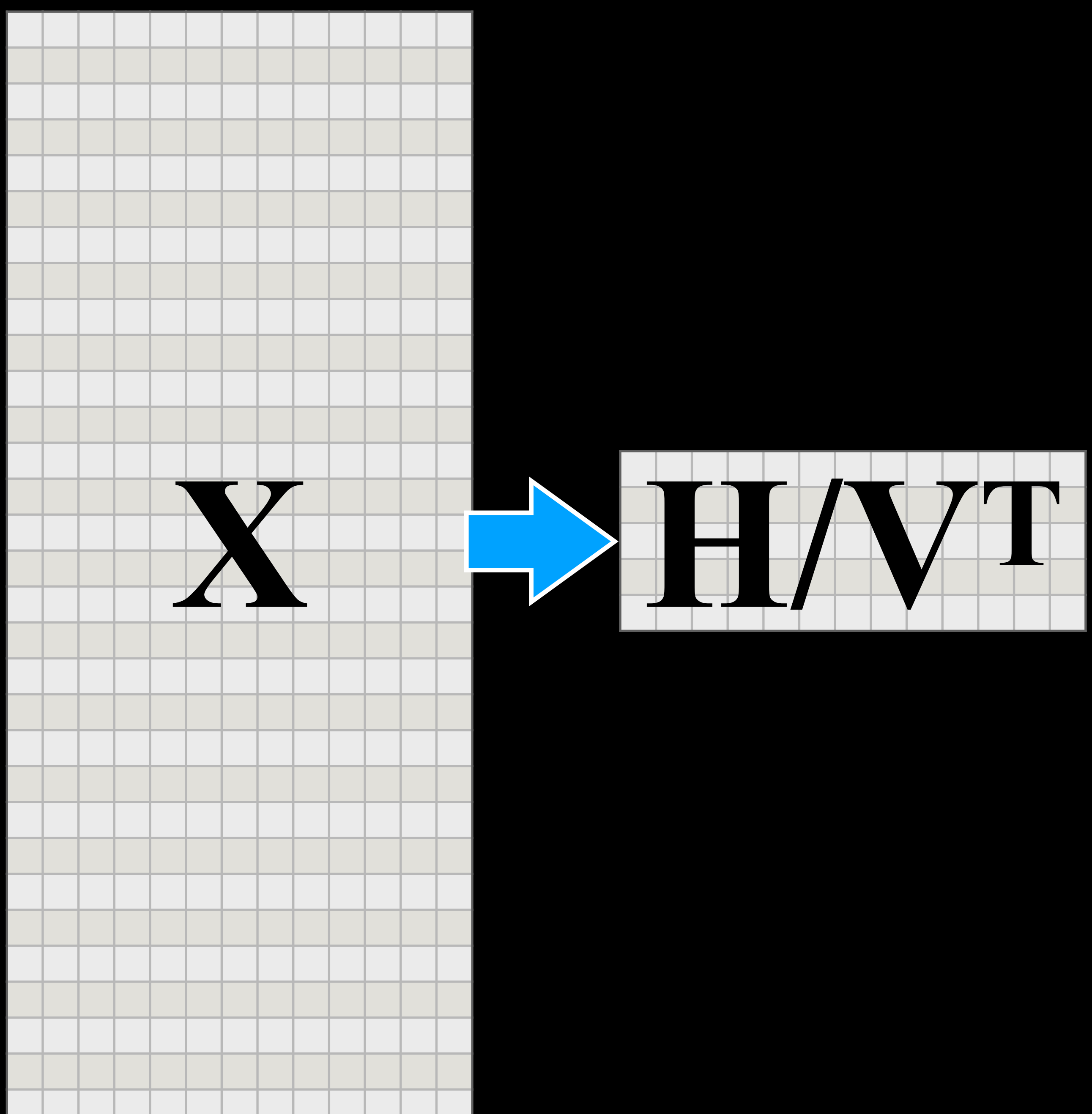
# Comparison

| | SVD | NMF |
|---|---|---|
| **Negative values (embeddings) as input?** | yes | no |
| **#components** | 3: $U, S, V$ | 2: $W, H$ |
| **document view?** | yes: $U$ | yes: W |
| **term view?** | yes: $V$ | yes: $H$ |
| **strength ranking?** | yes: $S$ | no |
| **exact?** | yes | no |
| **"topic" quality** | mixed | better |
| **sparsity** | low | medium |

# Yes, but: What is it Good for?

- Find top words for each latent dimension (≈ "topics")
- Find word similarity in latent space (alternative: Word2Vec)
- Find document similarity in latent space (alternative: Doc2Vec)
- Reduce dimensionality for visualization

# Latent Word Dimensions

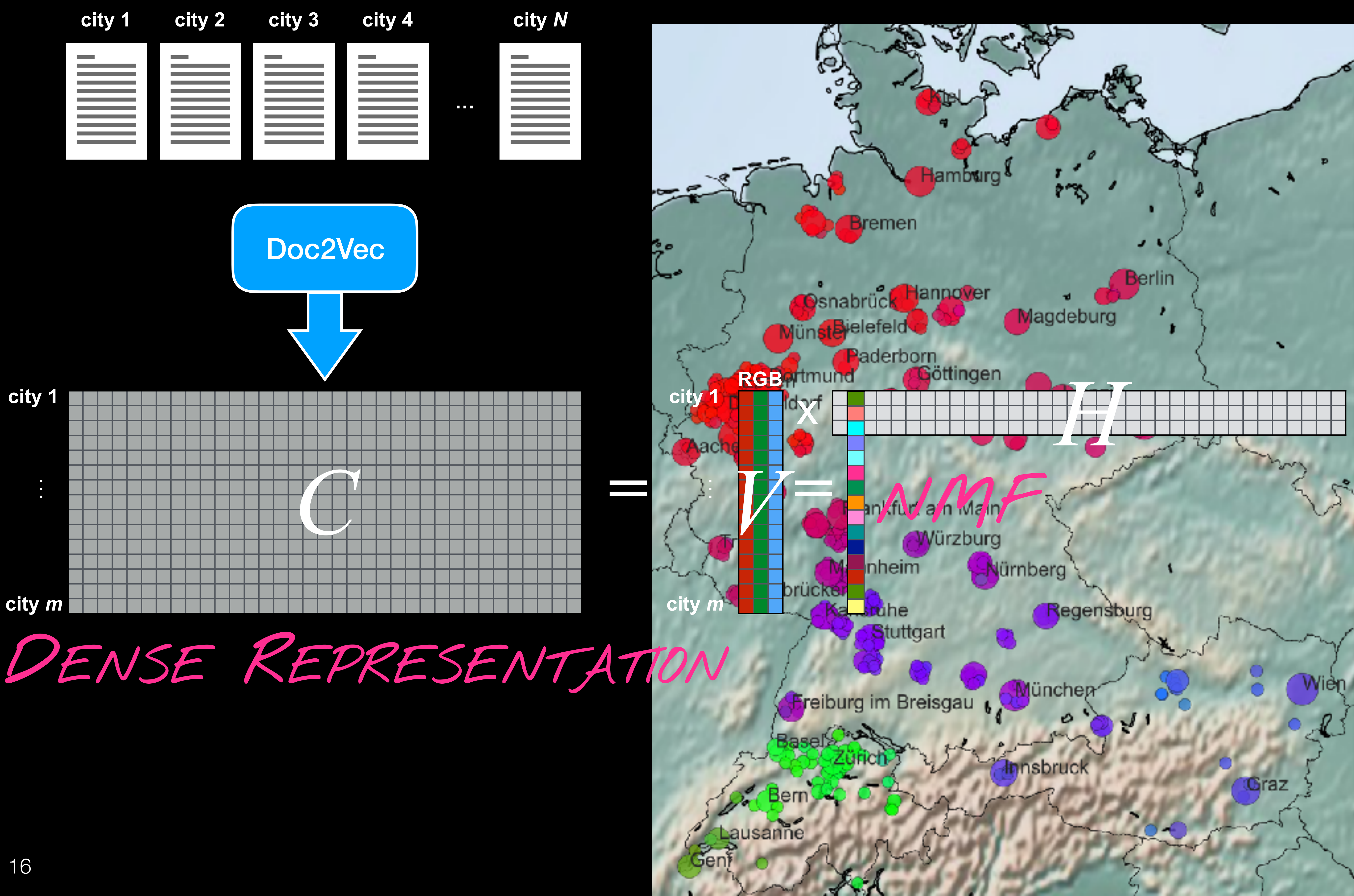



*NMF(10) for Moby Dick*

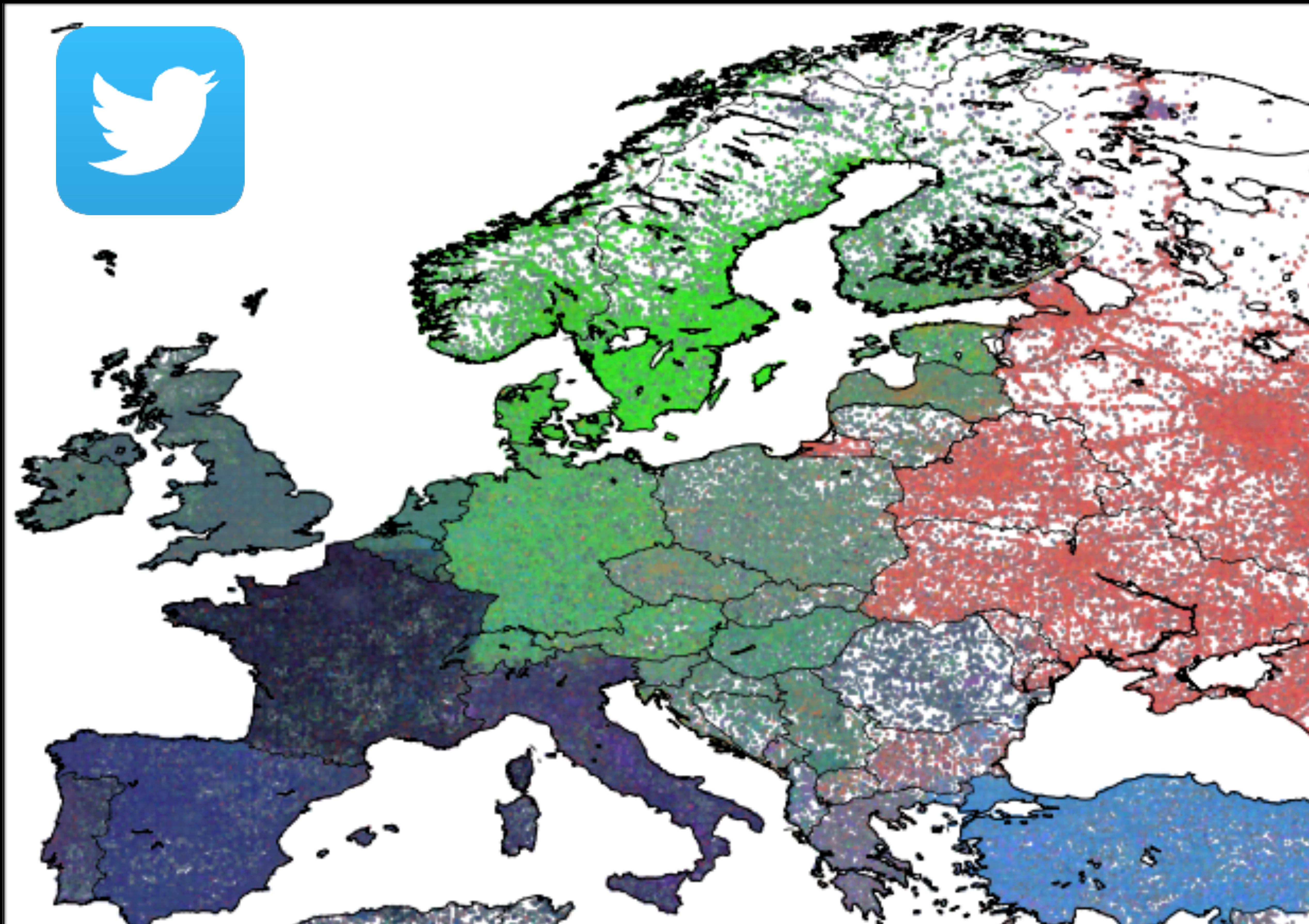
ahab, captain, cried, captain ahab, cried ahab
chapter, folio, octavo, ii, iii
like, ship, sea, time, way
man, old, old man, look, young man
oh, life, starbuck, sweet, god
said, stubb, queequeg, don, starbuck
sir, aye, let, shall, think
thou, thee, thy, st, god
whale, sperm, sperm whale, white, white whale
ye, look, say, ye ye, men

# Dimensionality Reduction for Visualizations

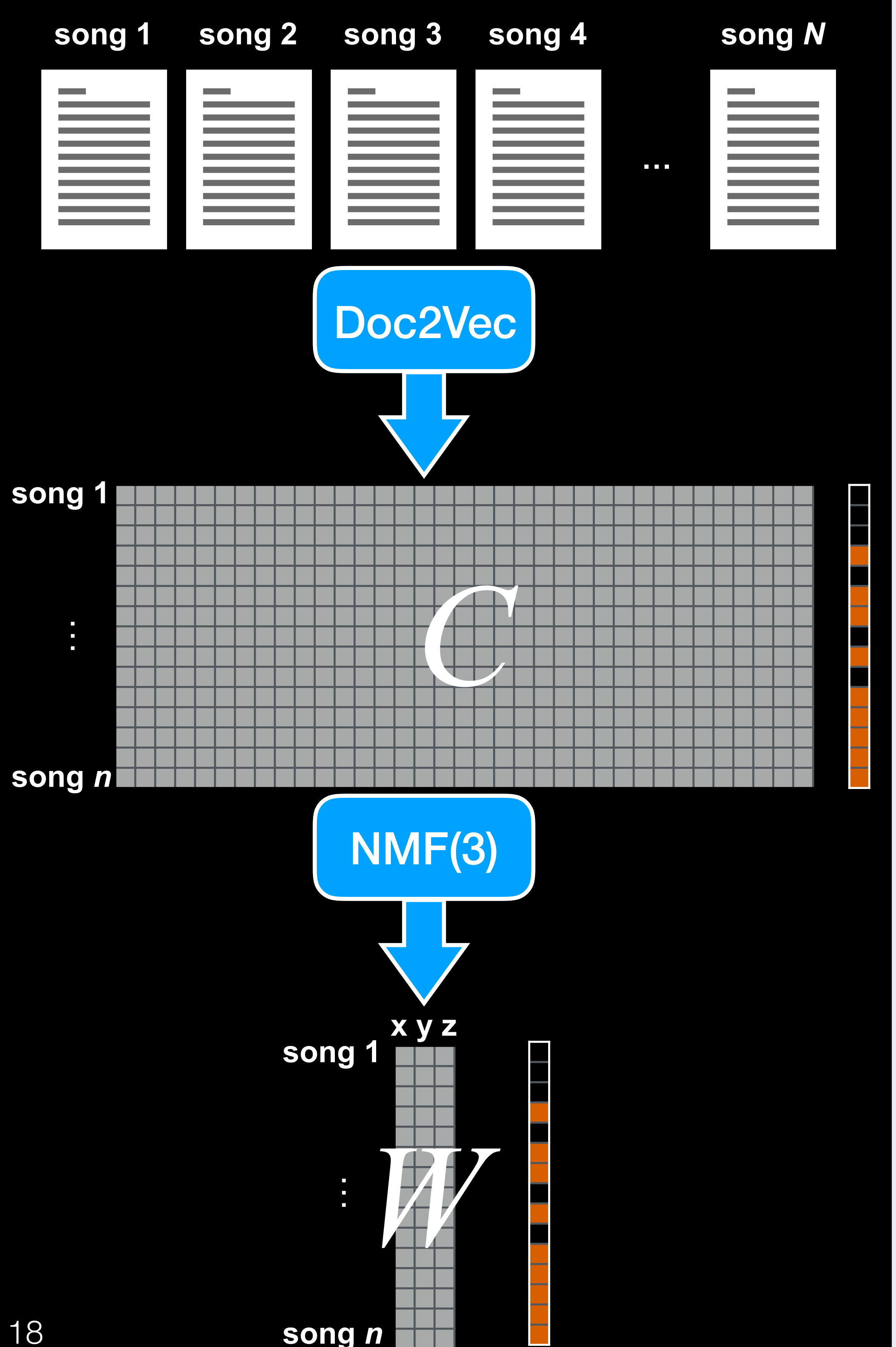
# Dimensions as RGB

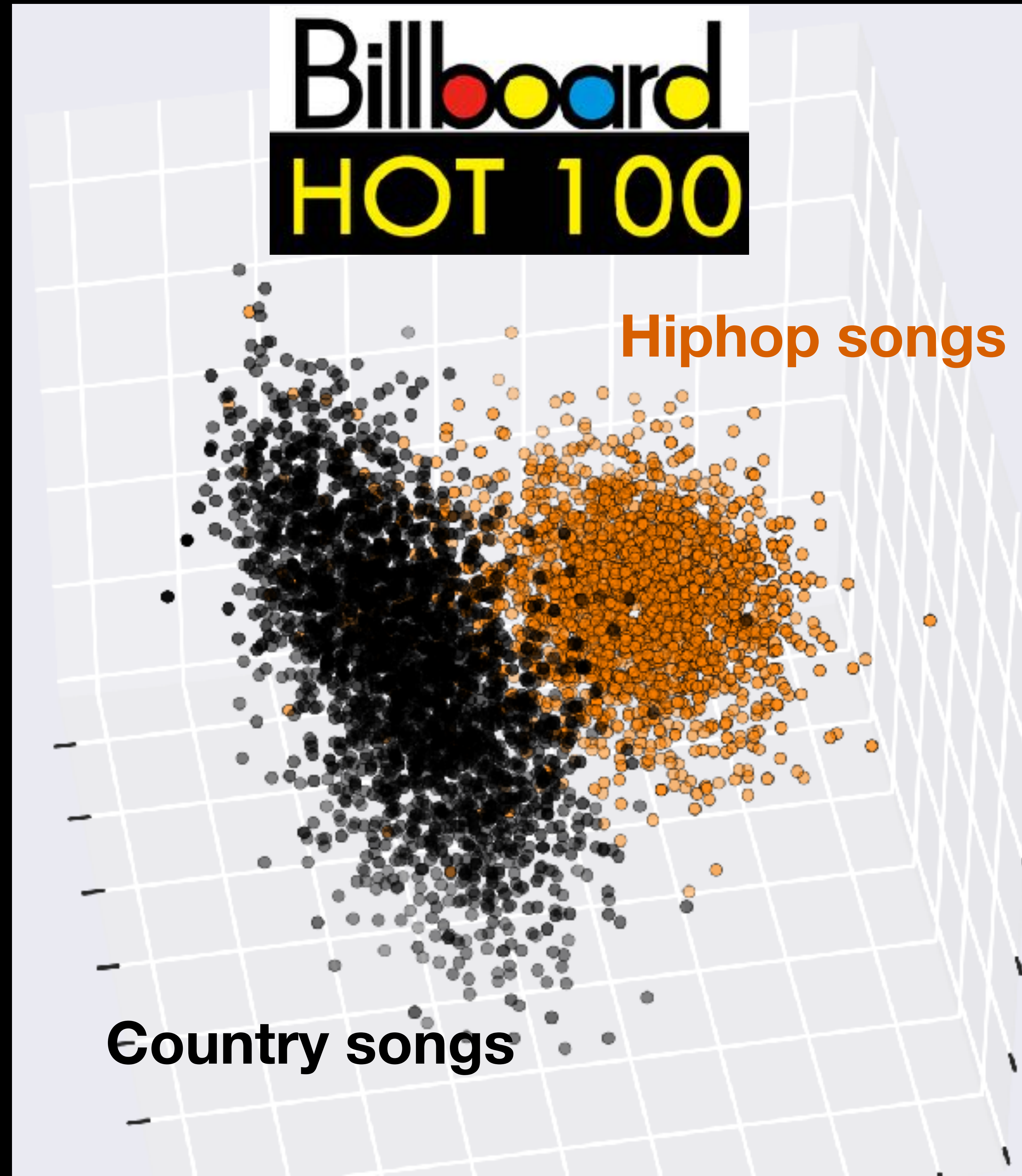


# Dimensions as RGB



# Dimensions as Position



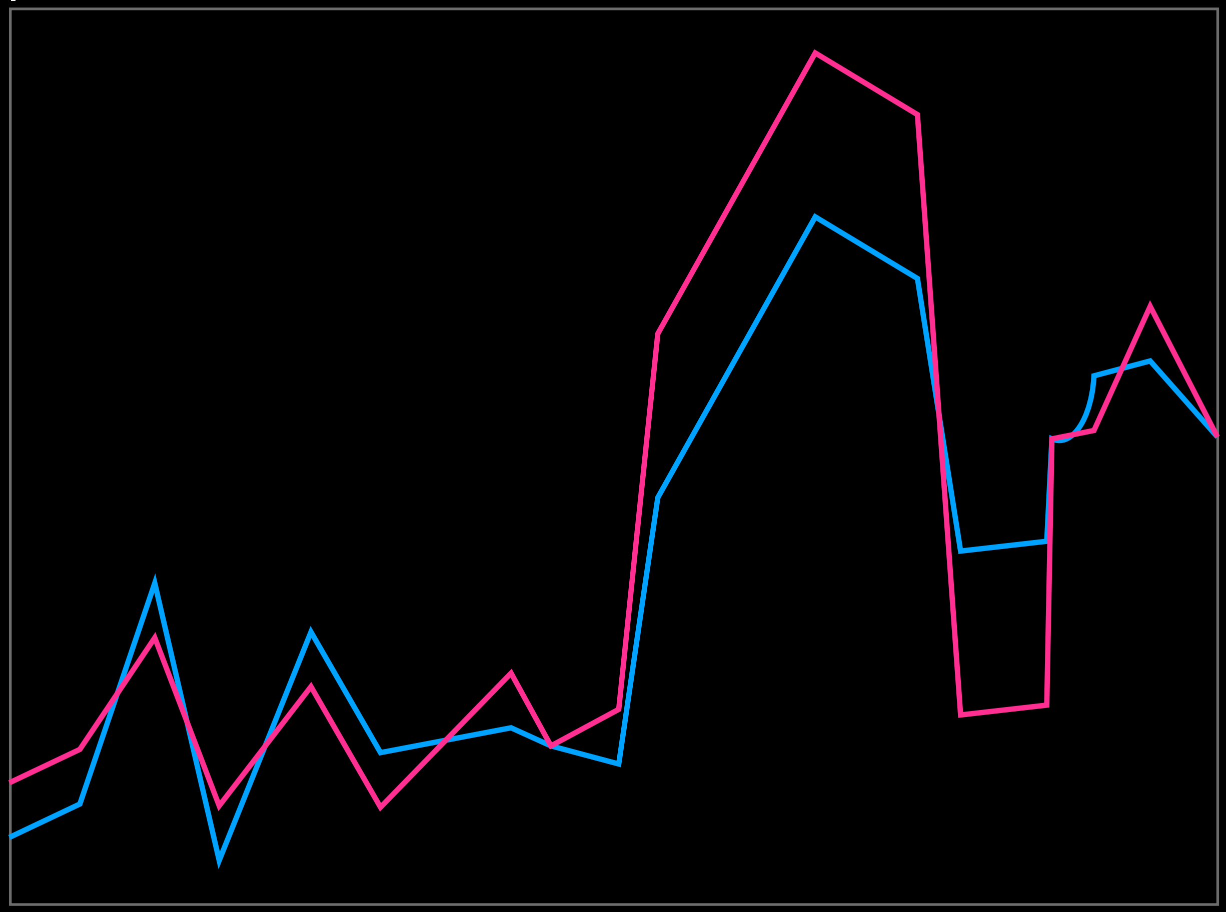



Van der Maaten & Hinton, 2008

# t-SNE

- Map/preserve neighborhood structure of high-dimensional space in lower-dimensional space
- Minimize difference between probability distributions over neighbors in both dimensions

$P(n_j|n_i)$



+ Nicer spatial images

– Stochastic

– Many parameters

# Clustering

# Making Sense of Clusters

NEUTRAL

NOUNS

perspicacious

lacerate

president

VERBS

kissed

fresh

ROMANCE ROOT

induce

rotten

GERMANIC ROOT

platypus

arsenic

mortify

mortgage

recognize

pie

blue

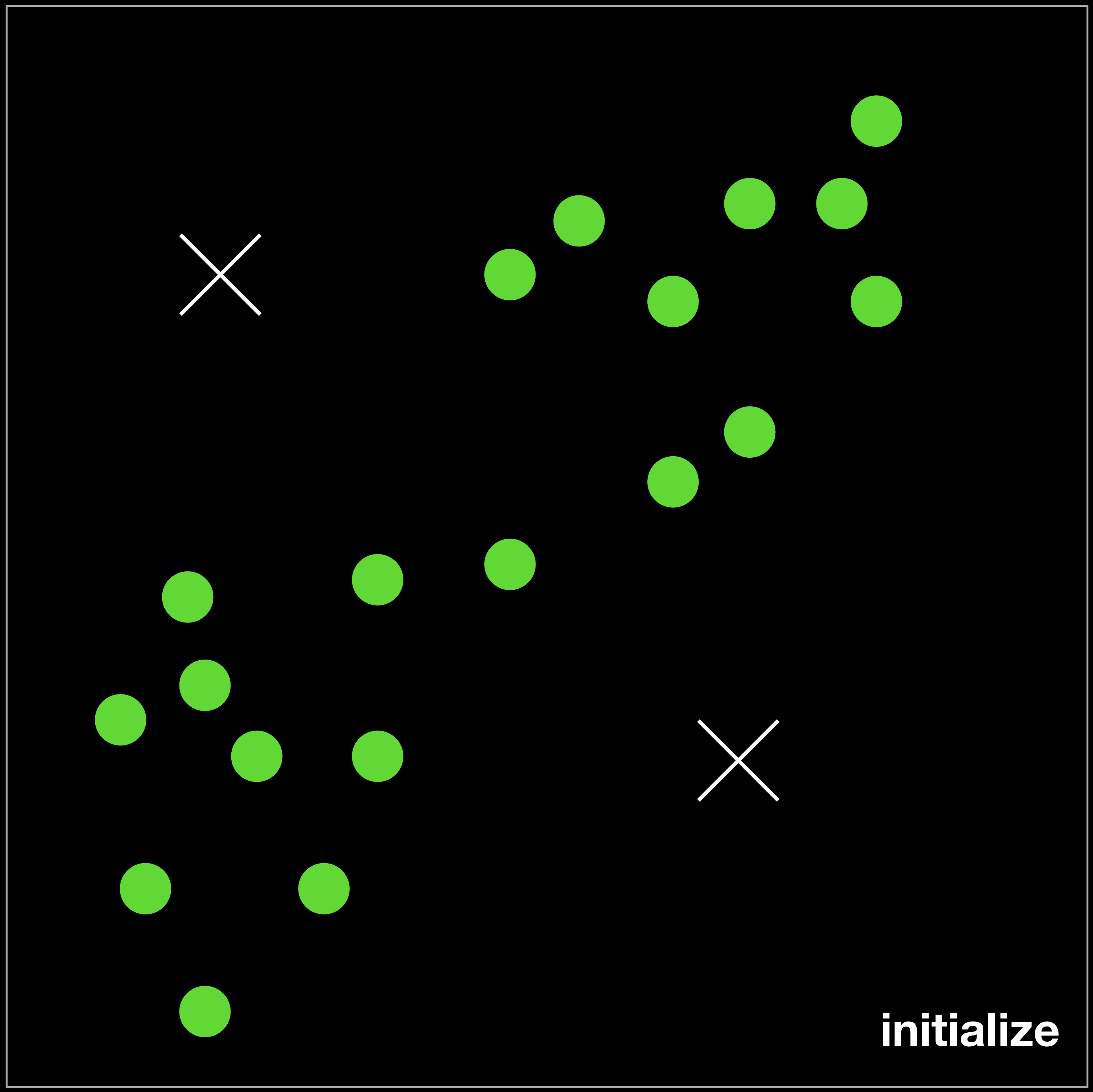
POSITIVE

ADJECTIVES

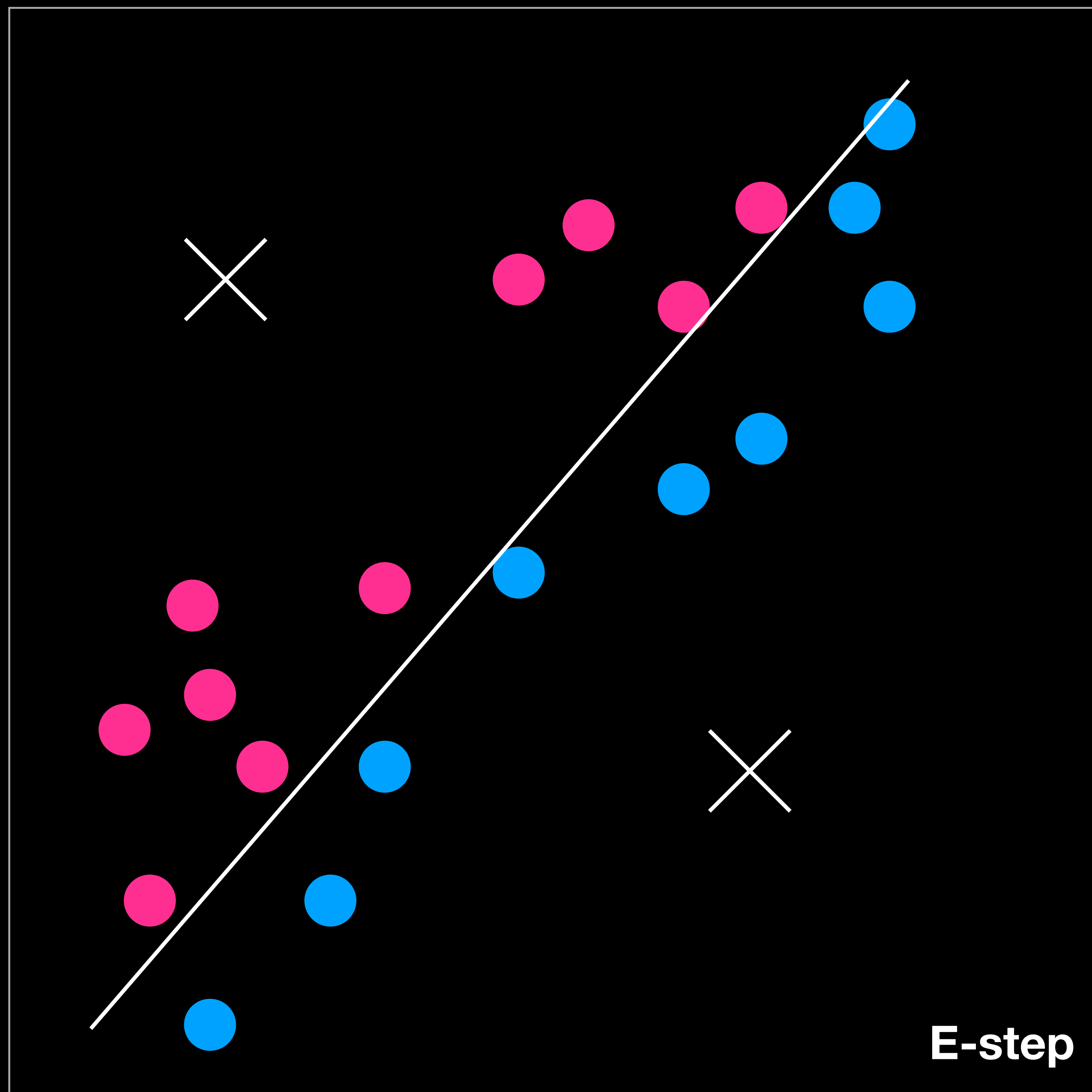
NEGATIVE

# *k*-Means Clustering

# *k*-Means



# *k*-Means

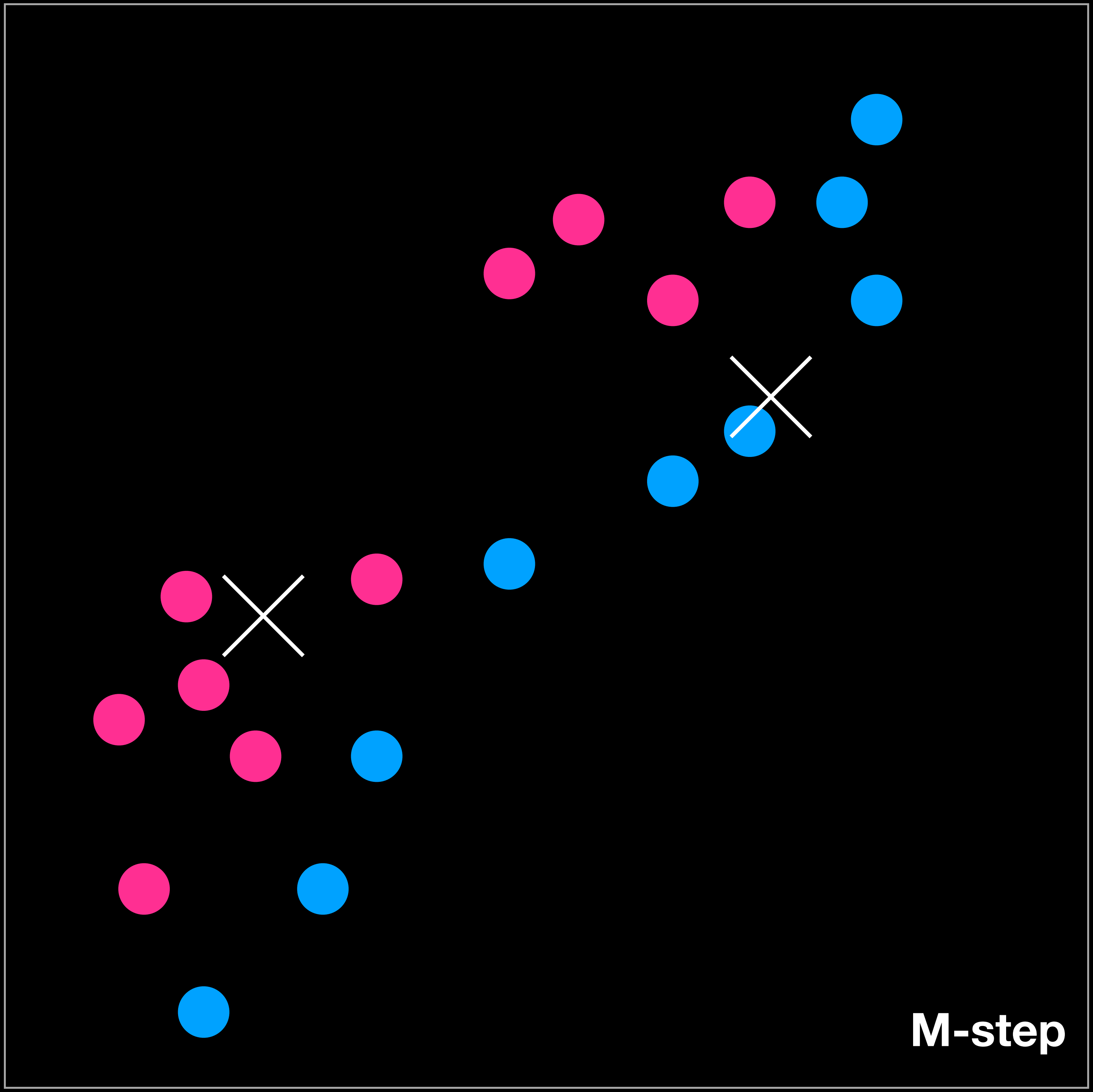



ASSIGN POINTS TO CENTROIDS



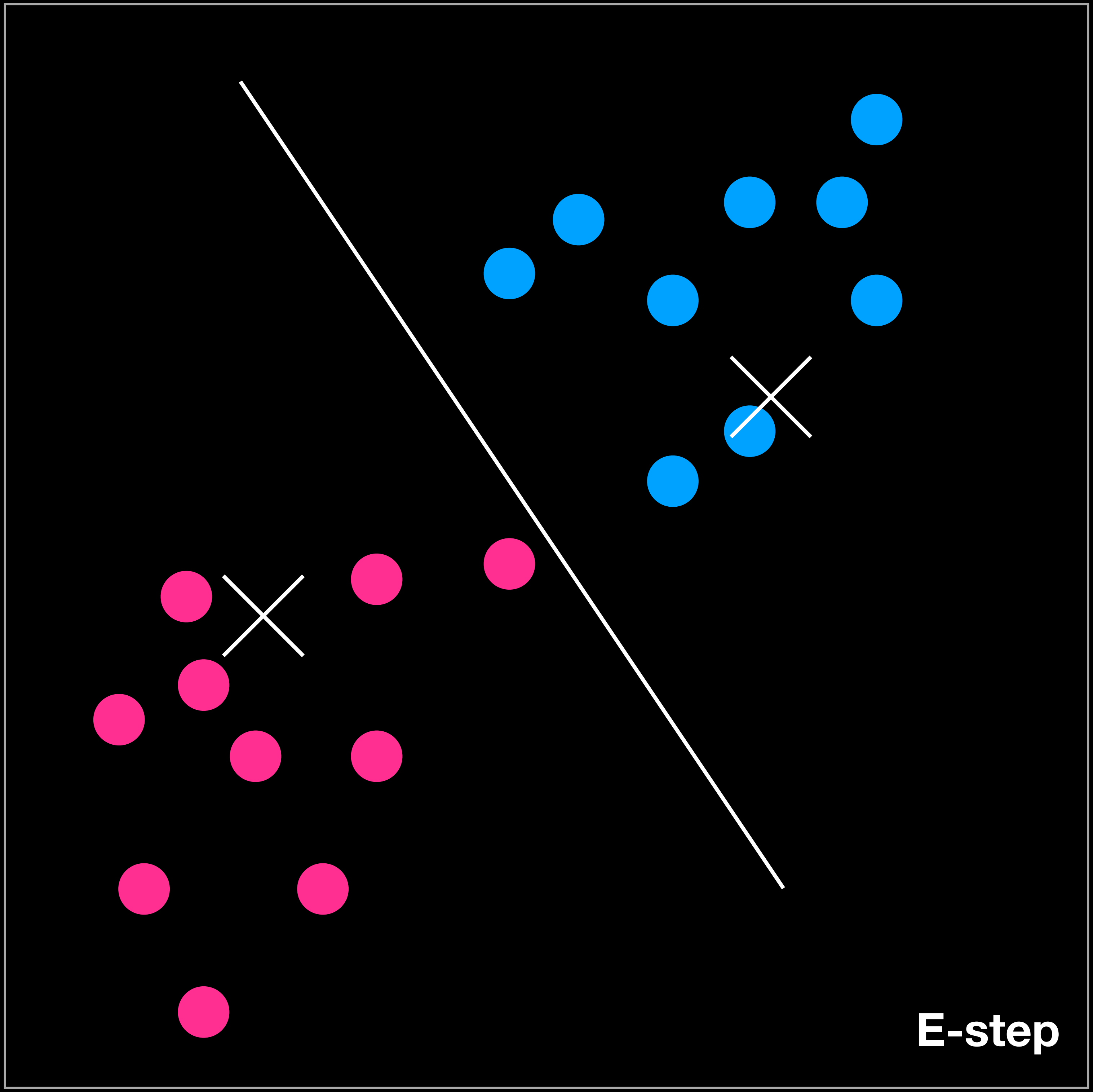
Bocconi

# *k*-Means

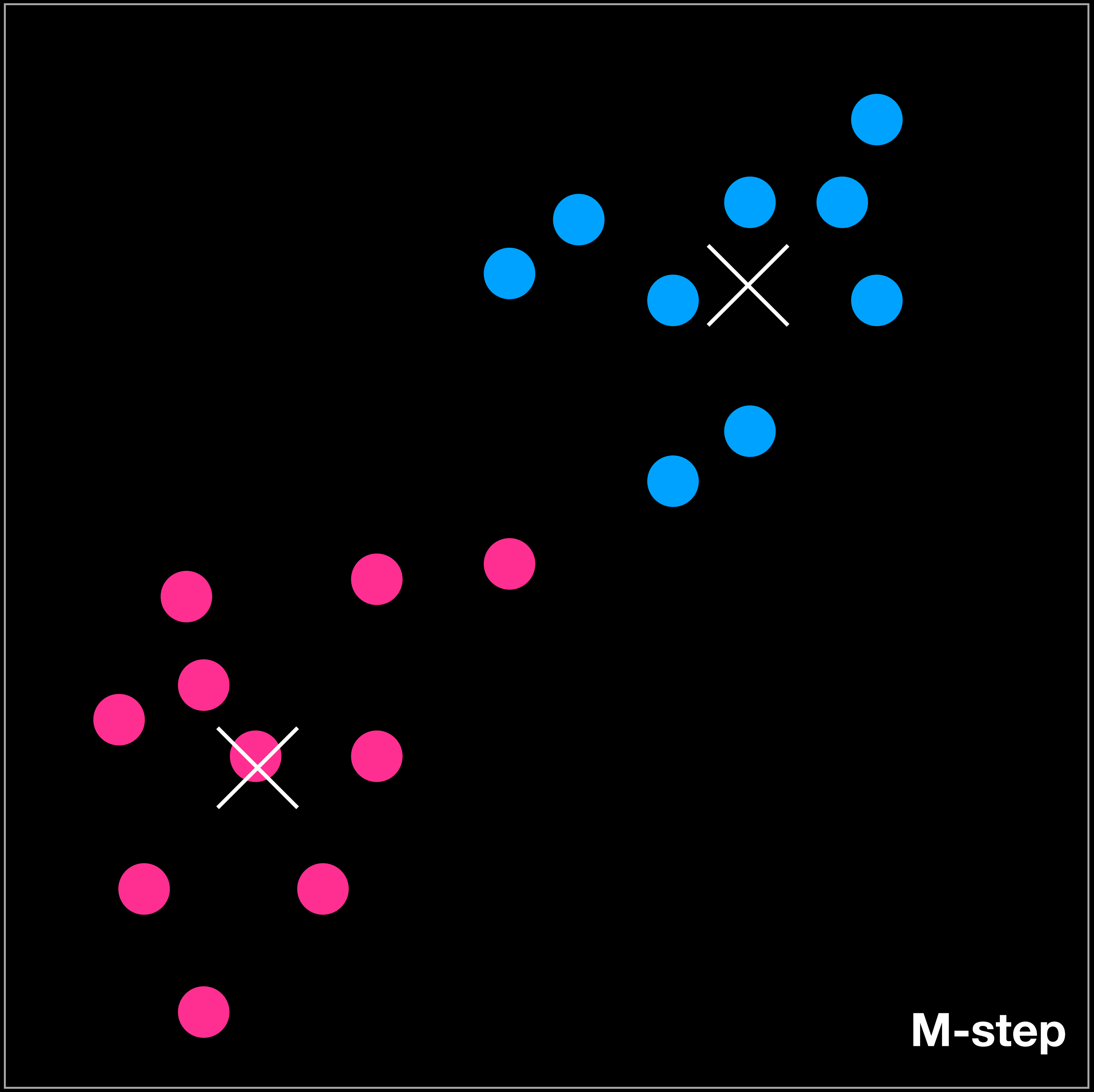



RECOMPUTE CENTROIDS

# *k*-Means

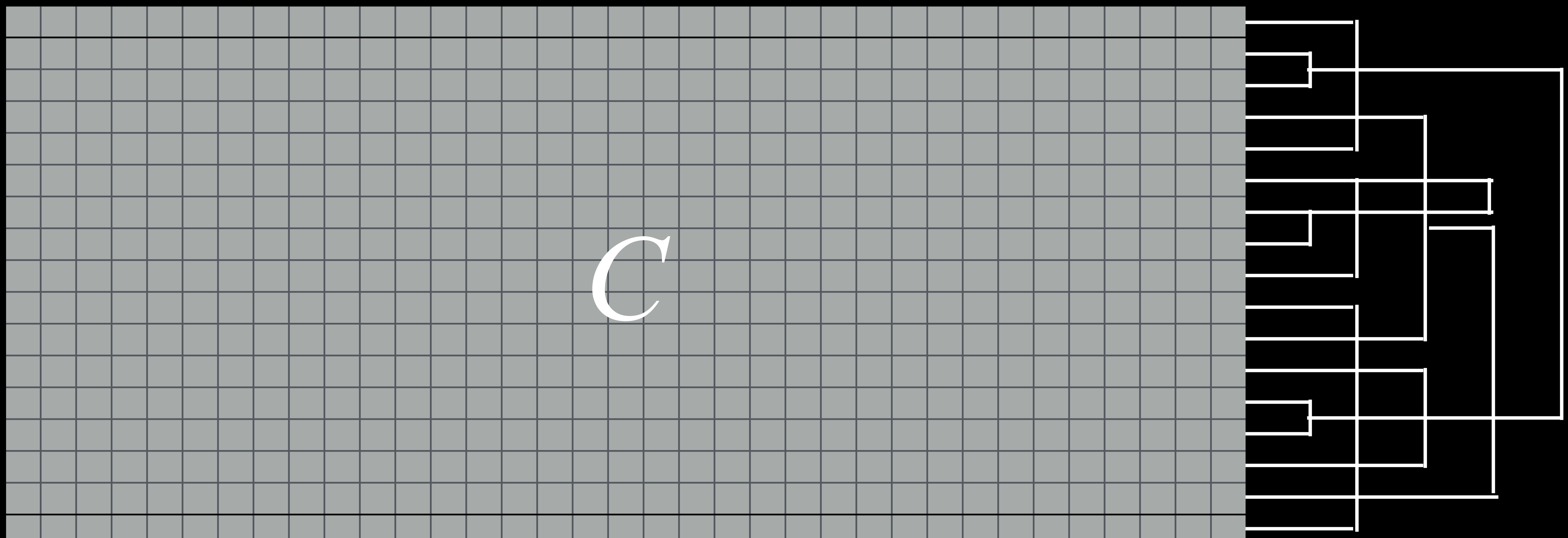


# *k*-Means

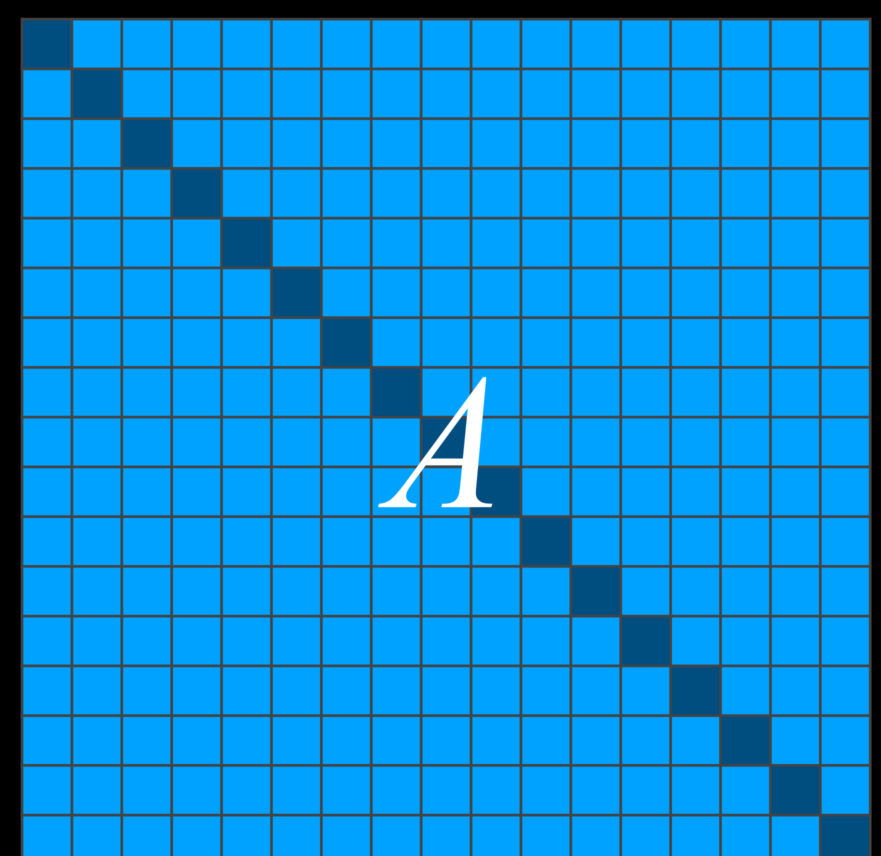


# Agglomerative Clustering

# Building Up

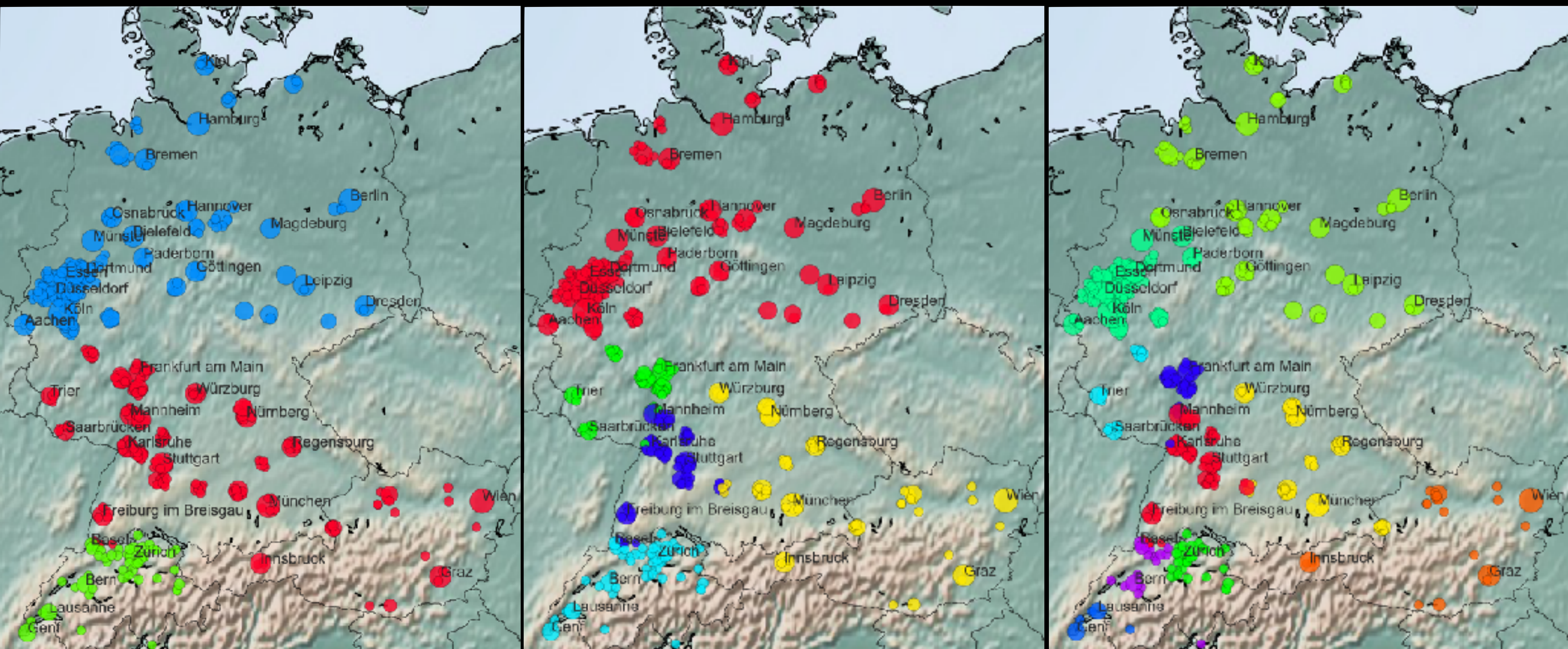



Adjacency Matrix



# Dialect Clusters

**3** **5** **10**





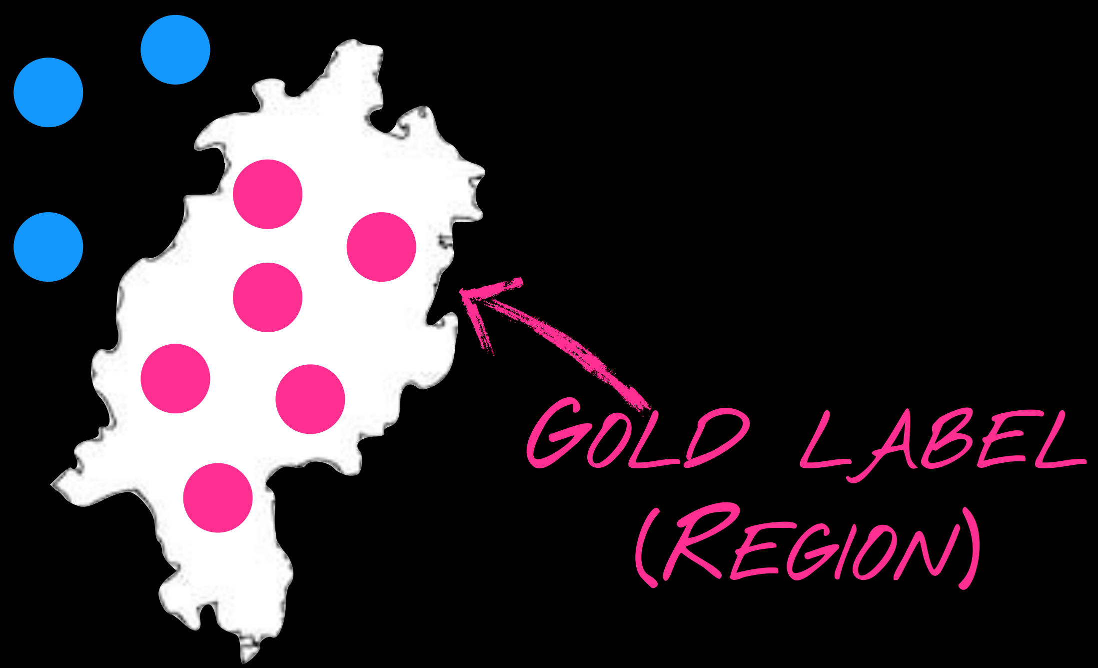

Bocconi

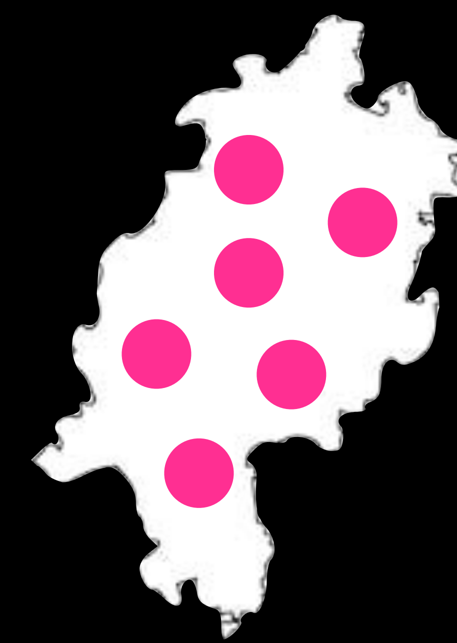
# Evaluation Metrics

**Homogeneity**
cluster has only 1 gold label

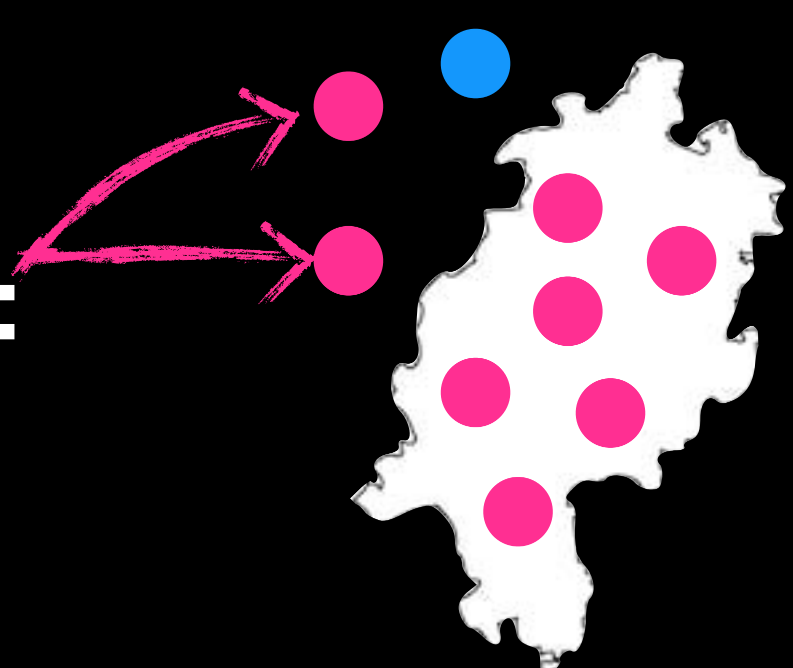
**Completeness**
gold label has only 1 cluster

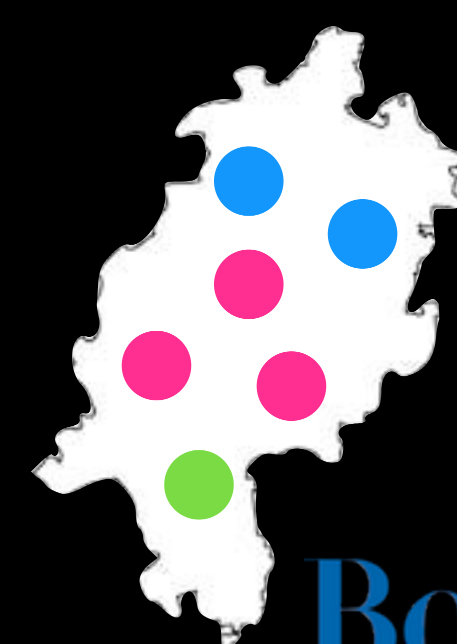
**Good:**



**Bad:**





# Comparison

| | *k*-means | Agg |
|---|---|---|
| **scalable** | yes | no (up to ~20k) |
| **repeatable result** | no | yes |
| **include external info** | no | yes |
| **Good on dense clusters?** | no | yes |

# Wrapping Up

# When to Use What

| | Discrete Features | Embeddings |
|---|---|---|
| **Latent topics** | NMF | *Not applicable* |
| **RGB translation** | NMF | SVD + scaling |
| **Plotting** | SVD | t-SNE |
| **Clustering** | *Reduce dimensions* | *Use as-is* |



Bocconi

# Take-Home Points

- **Matrix factorization** assumes latent concept dimensions
  - Can be used for semantic similarity (**LSA**)
  - Reduced components can be **visualized** in **graphs** or as **RGB** colors
- **Clusters** can group input in new ways
- Trade-off between speed and interpretability



Bocconi