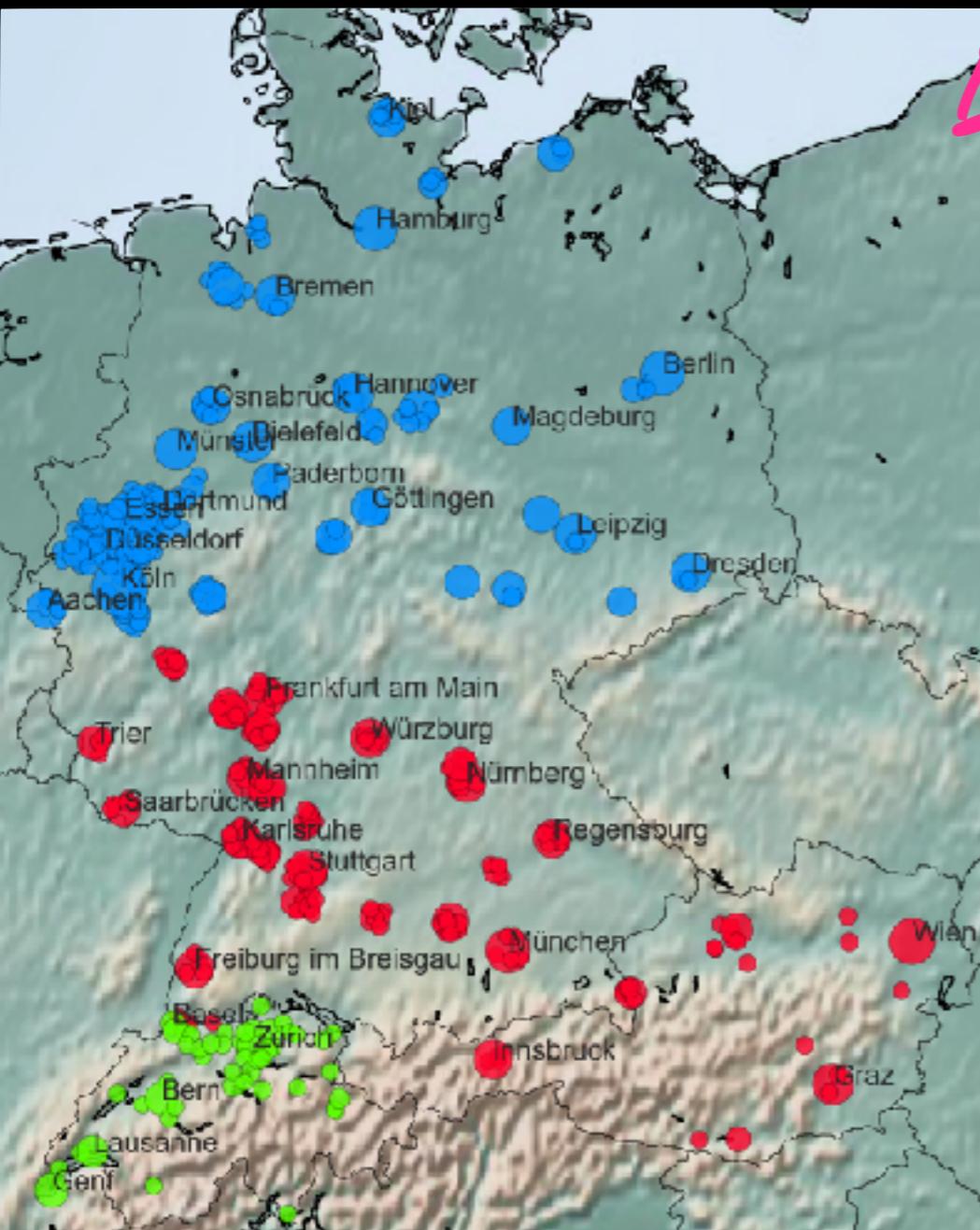


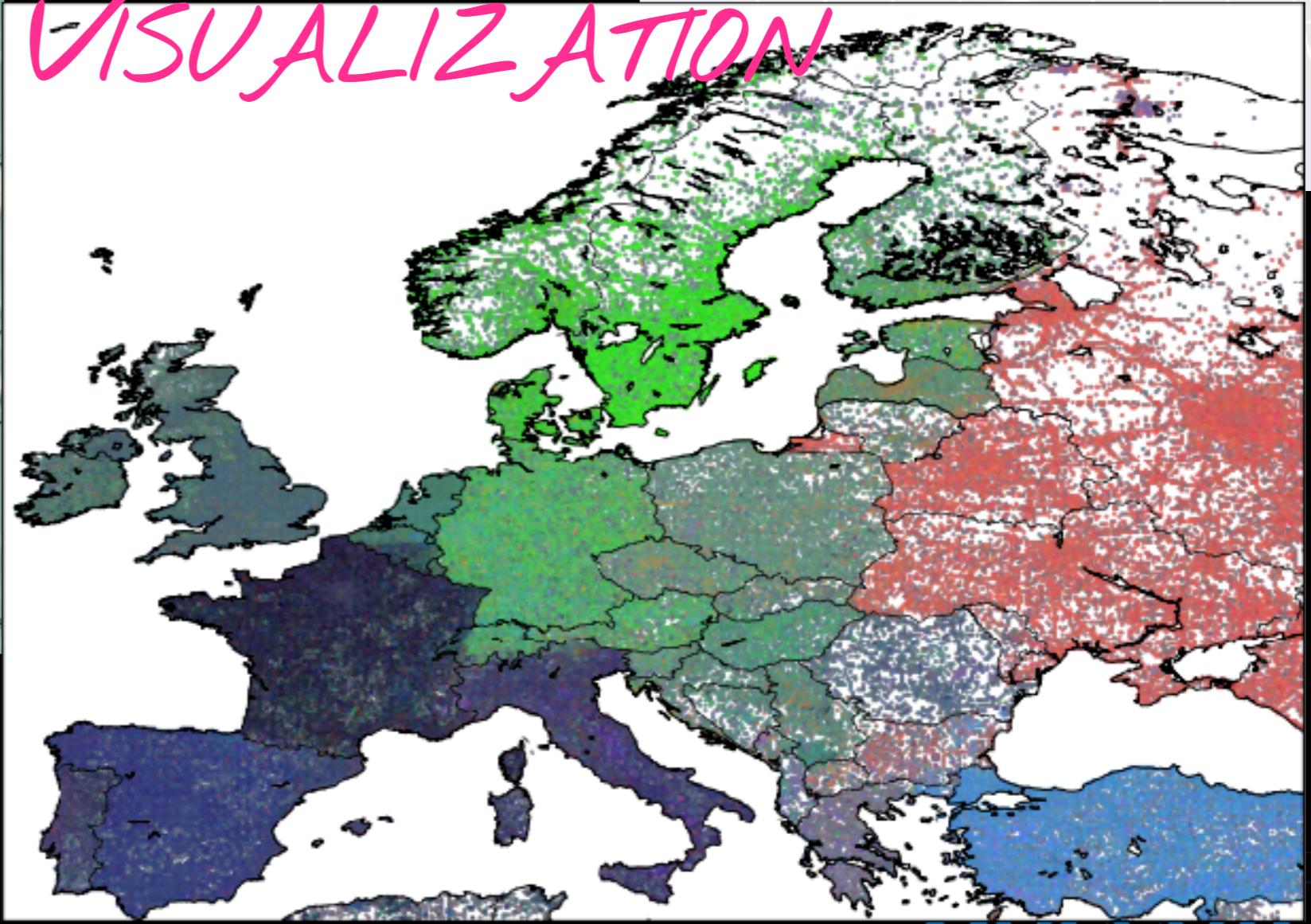
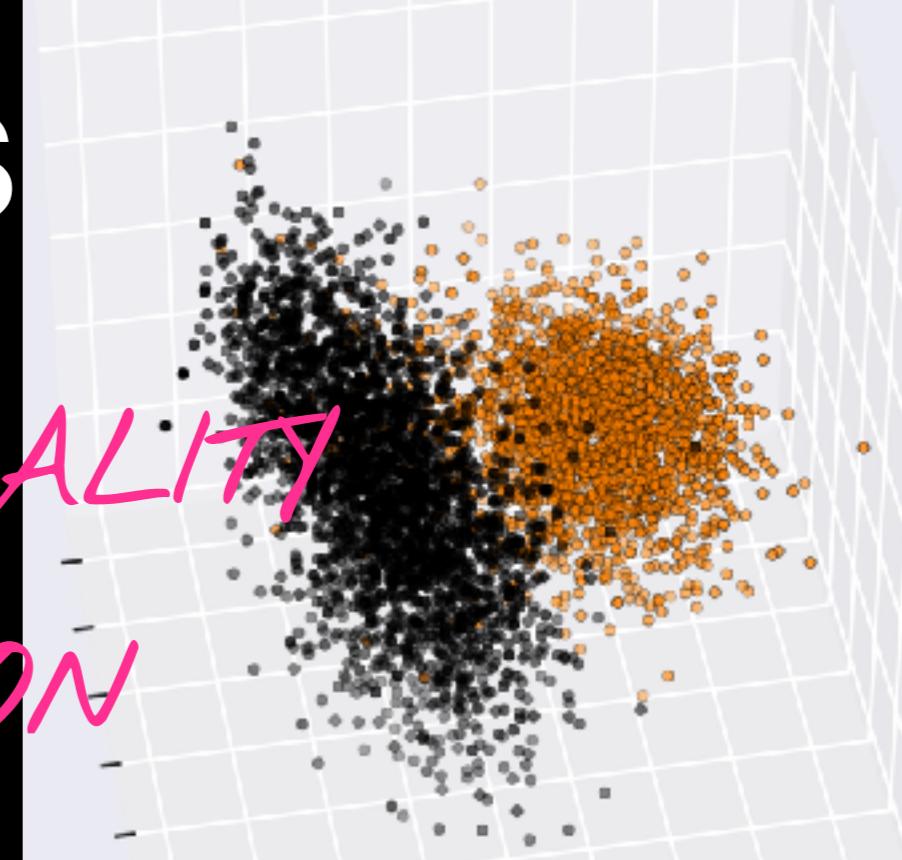
Examples



CLUSTERING

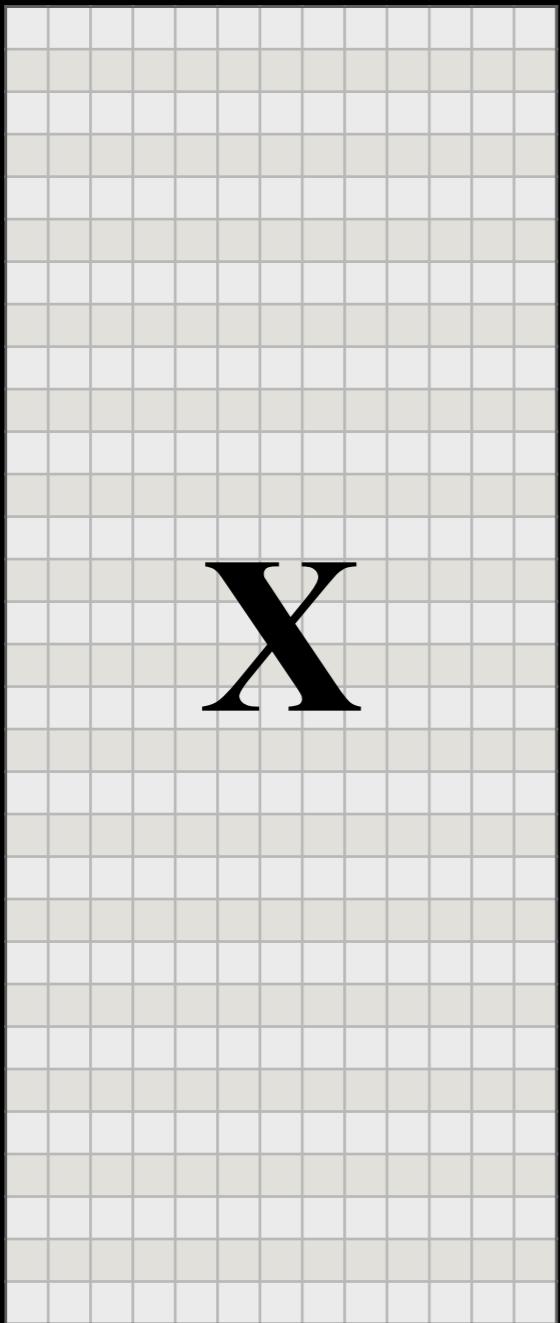
DIMENSIONALITY
REDUCTION

VISUALIZATION



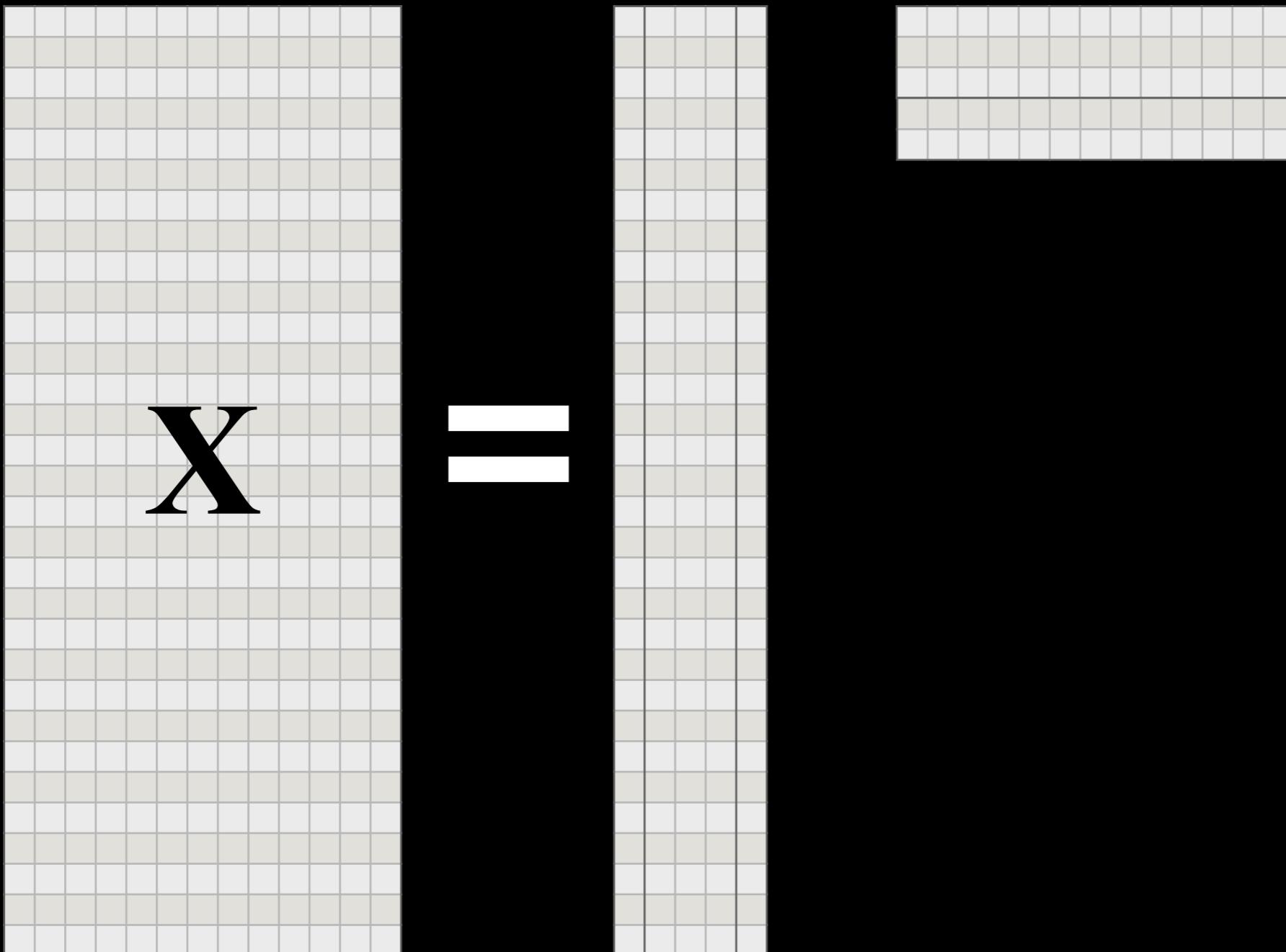
Latent Dimensions

DATA MATRIX

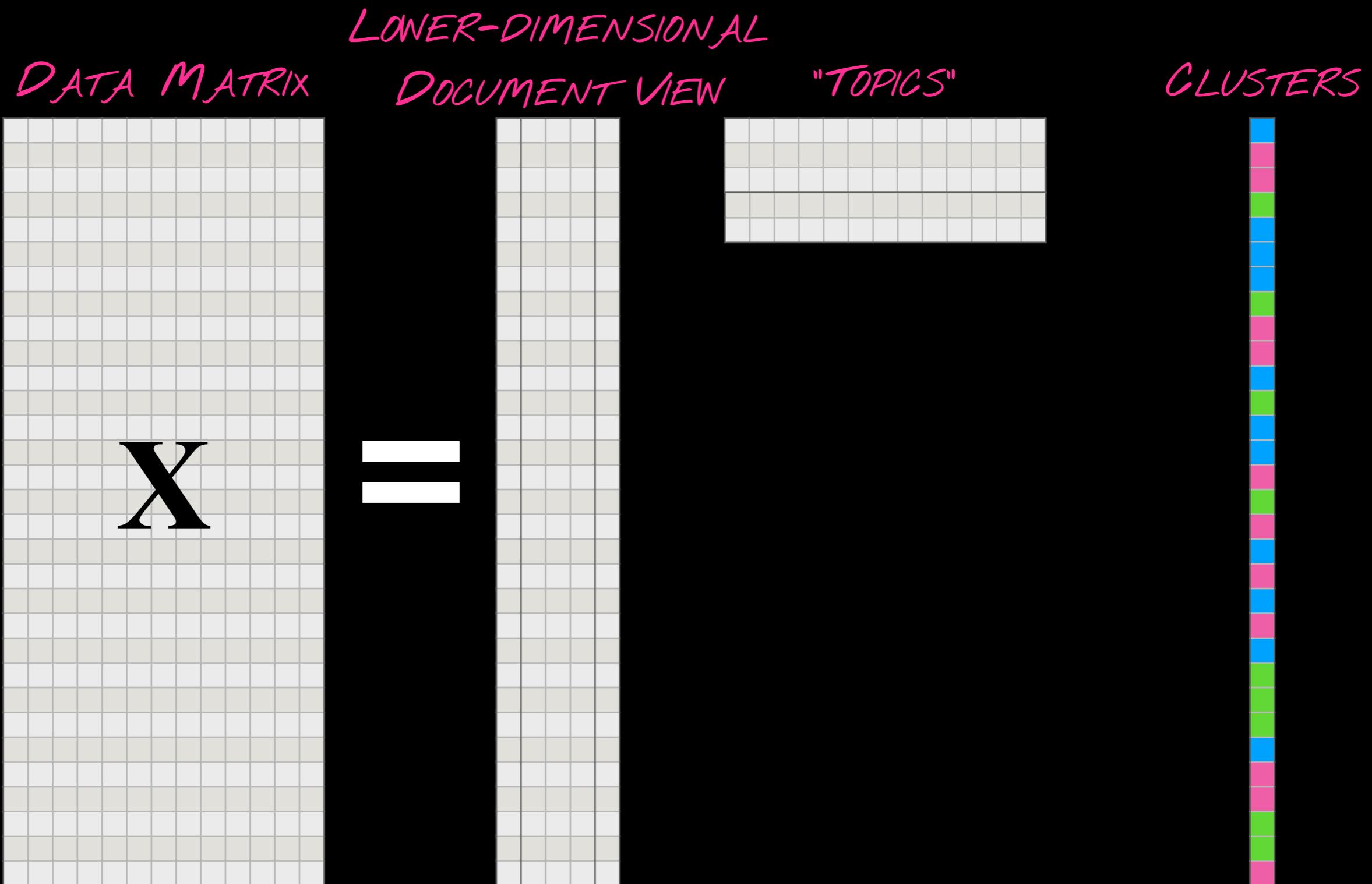


Latent Dimensions

LOWER-DIMENSIONAL
DATA MATRIX DOCUMENT VIEW "TOPICS"



Latent Dimensions

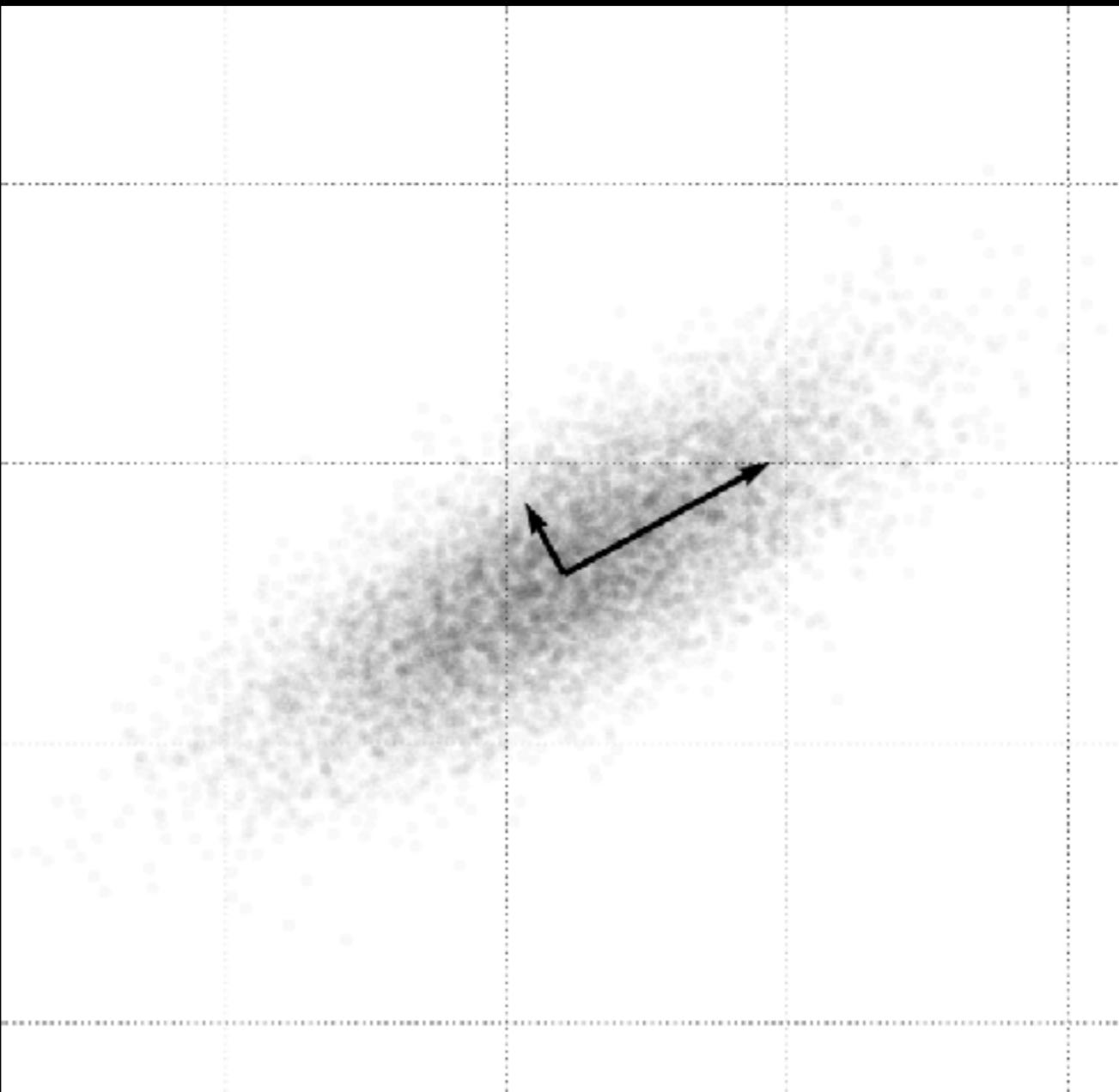


Goals for Today

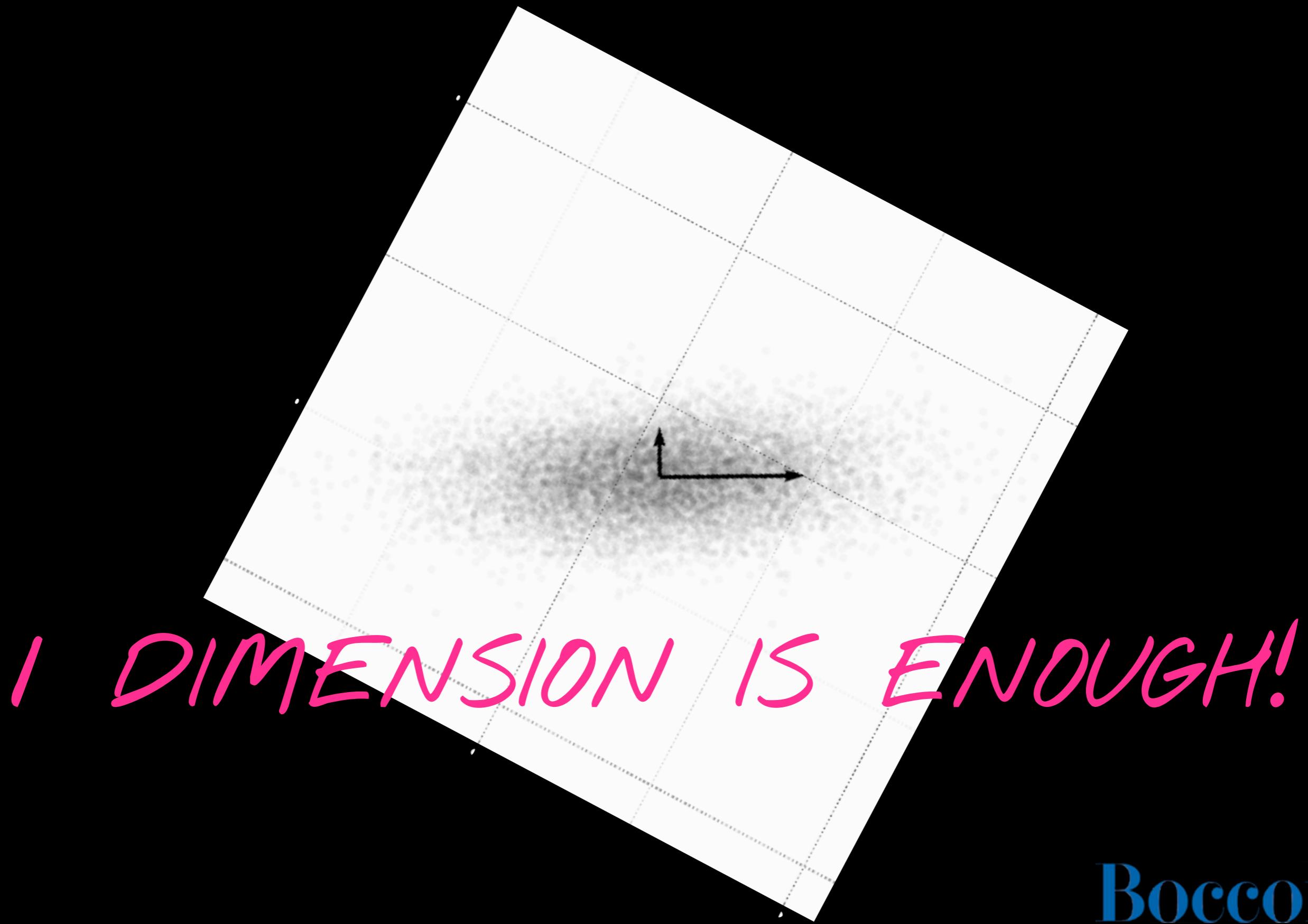
- Learn about **matrix factorization** and its use for **semantic similarity** and **visualization**
- Learn about **k-means** and **agglomerative clustering**
- Learn about **evaluation** criteria

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**

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**

M TERMS

$$\begin{matrix} N \\ D \\ O \\ C \\ S \end{matrix} \begin{matrix} 1 & 2 & 3 & 4 \\ 4 & 5 & 6 & 7 \\ 1 & 4 & 6 & 7 \\ 2 & 5 & 7 & 8 \\ 1 & 4 & 7 & 9 \end{matrix} = \begin{matrix} D \\ O \\ C \\ S \end{matrix}$$

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**

$$\begin{matrix} & \text{M TERMS} \\ \begin{matrix} N \\ D \\ O \\ C \\ S \end{matrix} & \end{matrix} \quad \begin{matrix} = \\ \text{D} \\ \text{O} \\ \text{C} \\ \text{S} \end{matrix} \quad \begin{matrix} K CONCEPTS \\ \begin{matrix} N \\ D \\ O \\ C \\ S \end{matrix} \end{matrix}$$

A diagram illustrating the Singular Value Decomposition (SVD) of a document-term matrix. On the left, a 5x4 matrix D is shown with columns labeled 'TERMS' (1, 2, 3, 4). The rows are labeled 'DOCS' (N, 1, 2, 3, 4) and have values 1, 2, 3, 4 respectively. A large black 'D' is placed over the first column. An equals sign follows, followed by two matrices: U (a 5x5 identity matrix) and S (a 5x4 diagonal matrix with values 1, 2, 3, 4, 5 along the diagonal).

Singular Value Decomposition

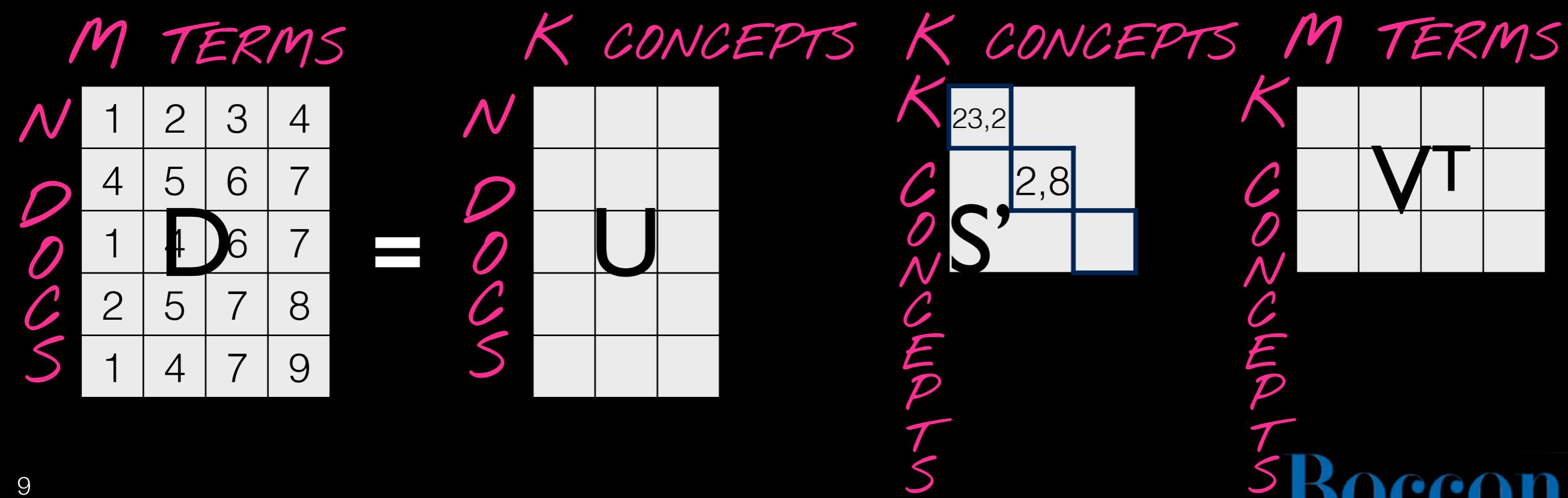
- “principal component analysis”: discover the dimensions that matter
- idea: matrix is made up of few hidden dimensions
- Dimensions correspond to **documents**, **terms**, and **latent concepts**

$$\begin{matrix} M \text{ TERMS} \\ N \\ D \\ O \\ C \\ S \end{matrix} \quad \begin{matrix} K \text{ CONCEPTS} \\ N \\ D \\ O \\ C \\ S \end{matrix} \quad = \quad \begin{matrix} K \text{ CONCEPTS} \\ K \\ C \\ O \\ N \\ C \\ E \\ P \\ T \\ S \end{matrix} \quad \begin{matrix} M \text{ TERMS} \\ K \\ C \\ O \\ N \\ C \\ E \\ P \\ T \\ S \end{matrix}$$

The diagram illustrates the Singular Value Decomposition (SVD) of a document-term matrix D . The matrix D has dimensions $M \times N$ (4 terms by 4 documents). It is decomposed into three components: U (orthogonal matrix of document concepts), S (diagonal matrix of singular values), and V^T (orthogonal matrix of term concepts). The matrix S is shown with its top-left element highlighted in blue, labeled $23,2$. Other elements in the matrix are labeled $2,8$ and $0,8$.

Singular Value Decomposition

- reduce principal components/concepts to smaller number



Singular Value Decomposition

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

U		

	23,2	
S'	2,8	

$$\mathbf{V}^T =$$

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

Singular Value Decomposition

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

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



0,9	2,1	3,2	3,8
3,9	5,1	6	6,9
1,1	3,8	4,9	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

M TERMS

N 1 2 3 4

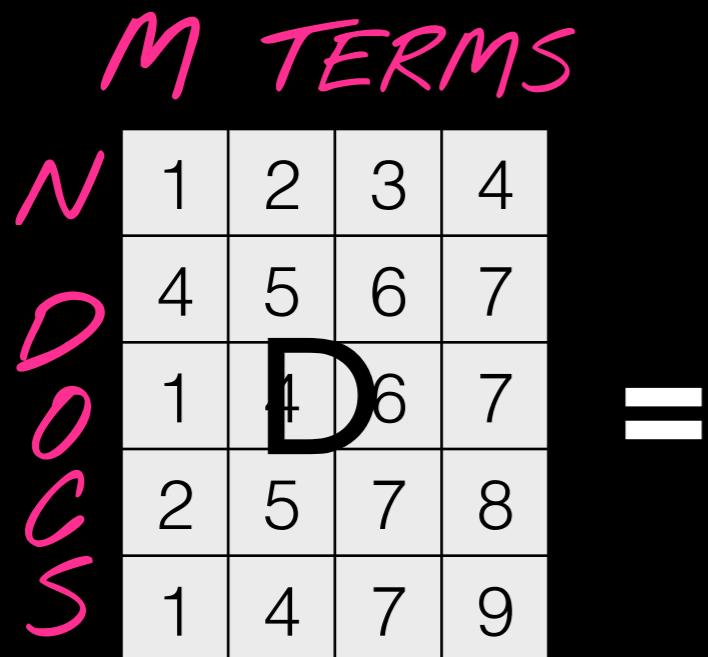
D 4 5 6 7

O 1 1 6 7

C 2 5 7 8

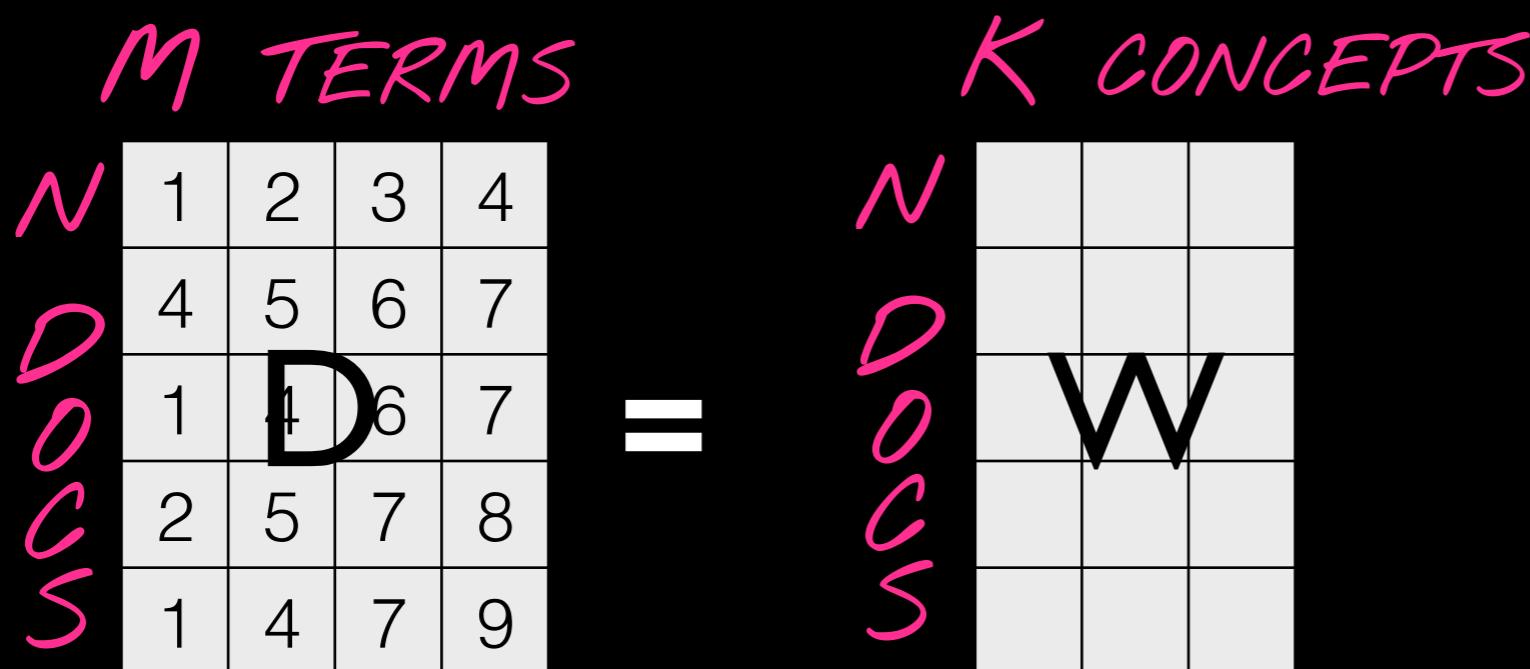
S 1 4 7 9

=



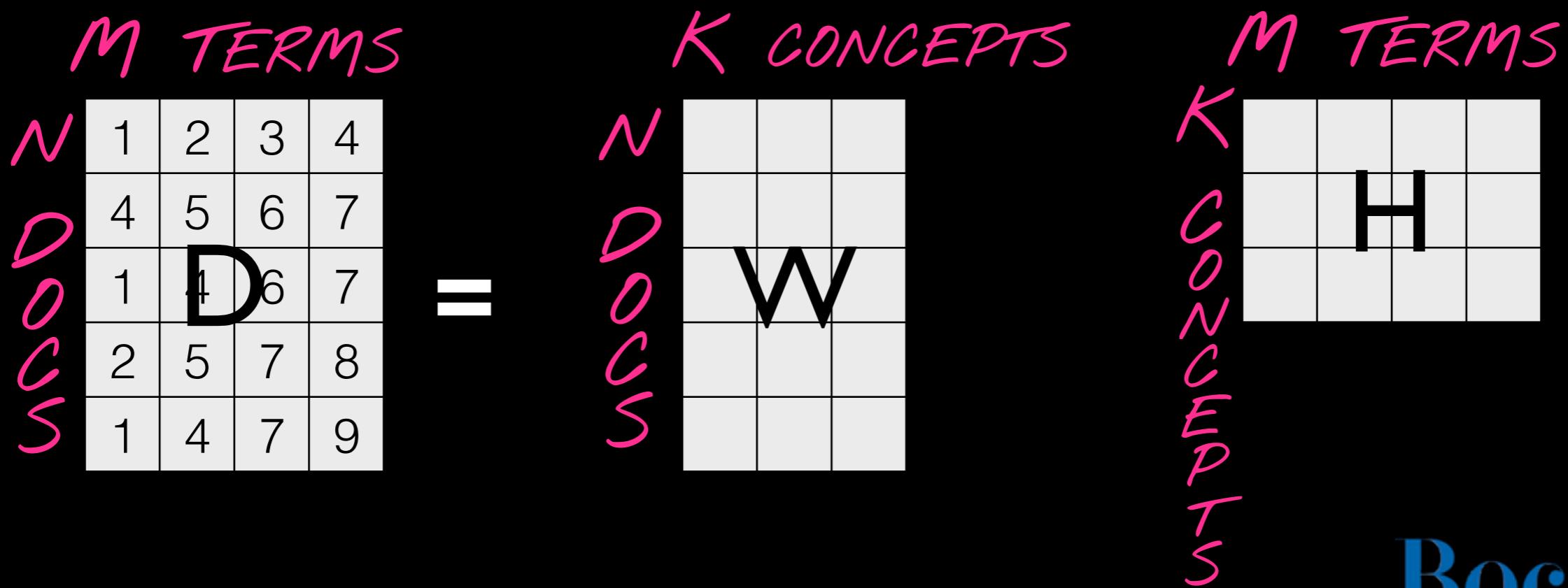
Non-negative Matrix Factorization

- Use only positive values
- Find approximation of two components



Non-negative Matrix Factorization

- Use only positive values
- Find approximation of two components



Non-negative Matrix Factorization

- Use only positive values
- Find approximation of two components

DOCUMENT VIEW

M TERMS	
N	1 2 3 4
D	4 5 6 7
O	1 4 6 7
C	2 5 7 8
S	1 4 7 9

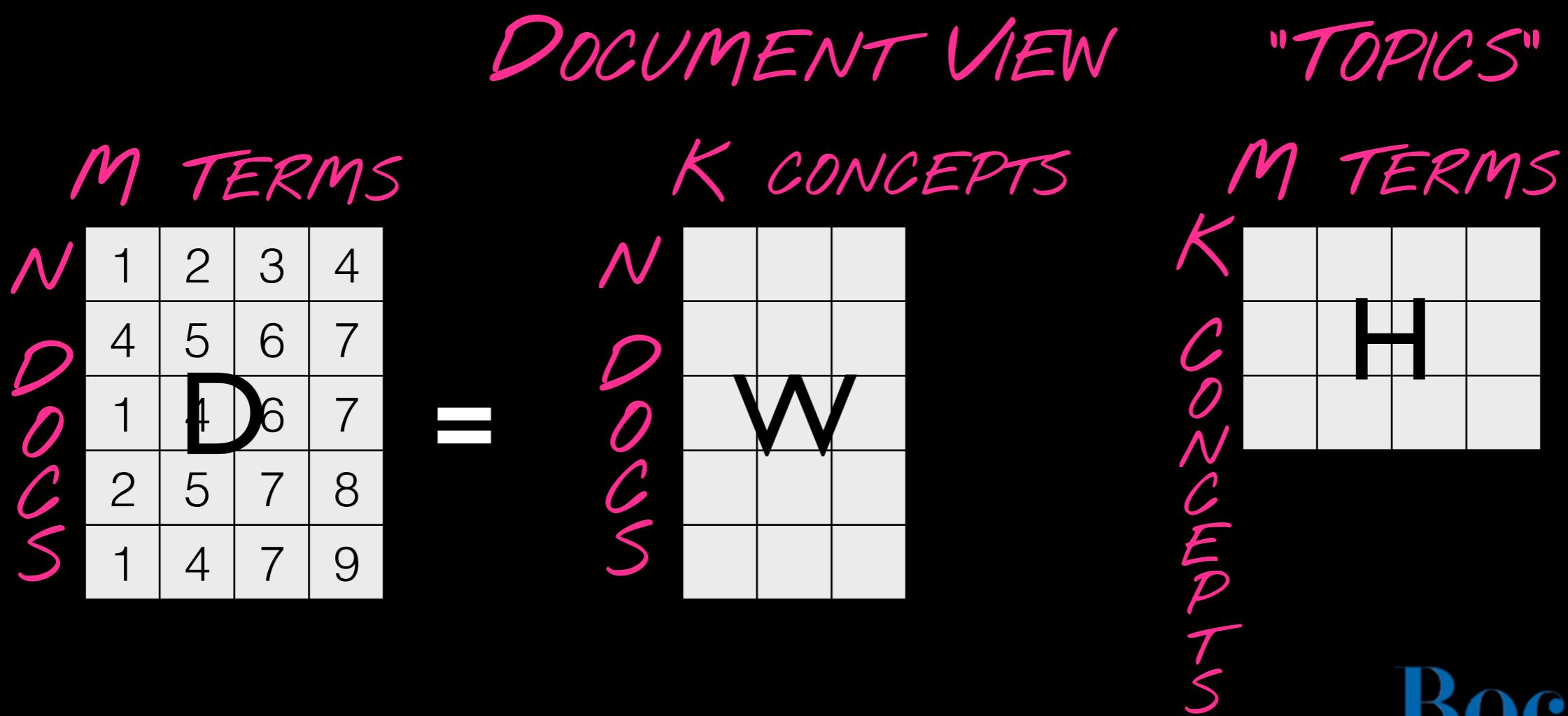
=

K CONCEPTS	
N	
D	
O	
C	
S	

M TERMS	
K	
C	H
N	
O	
C	
E	
P	
T	
S	

Non-negative Matrix Factorization

- Use only positive values
- Find approximation of two components



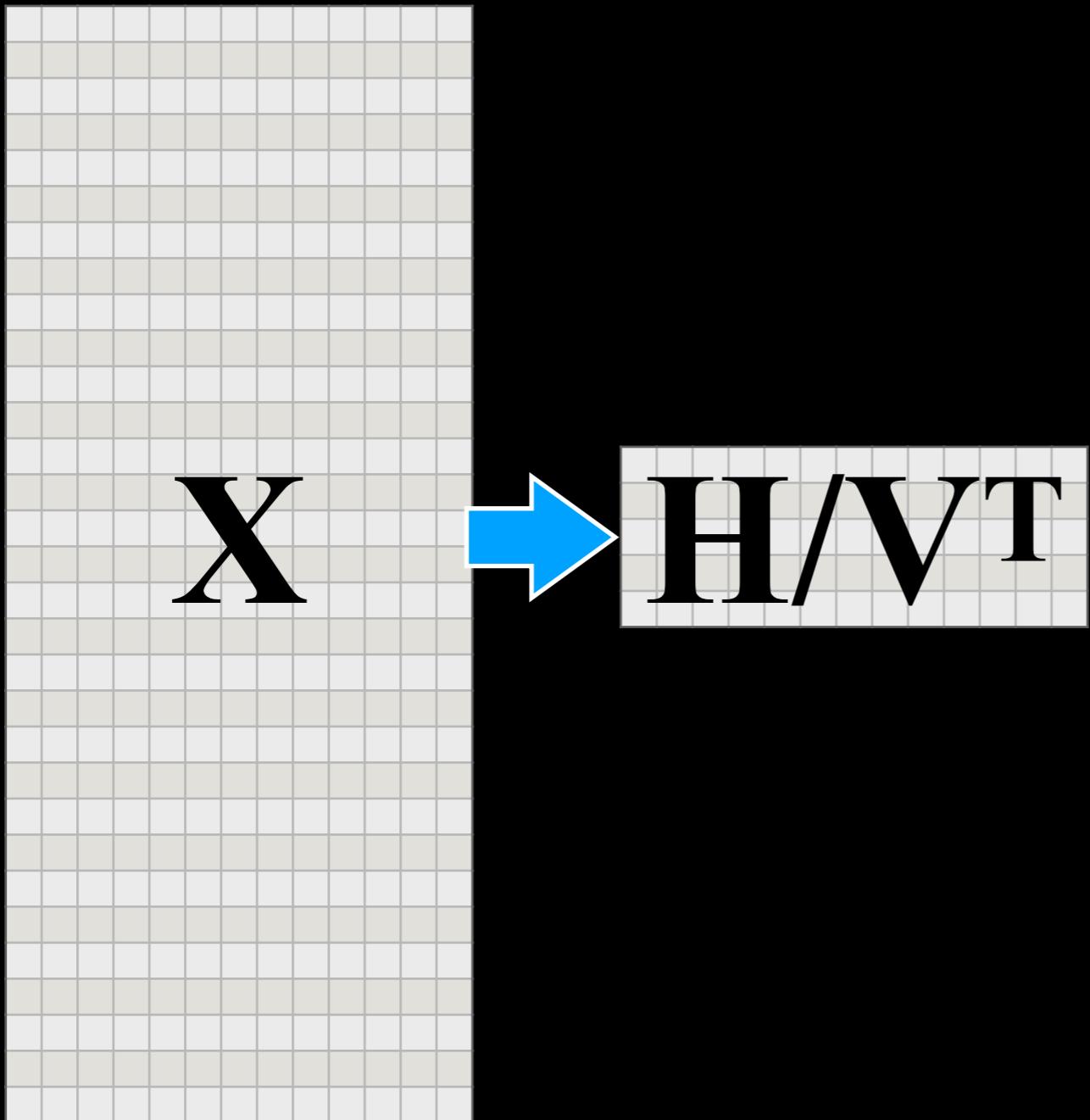
Comparison

	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

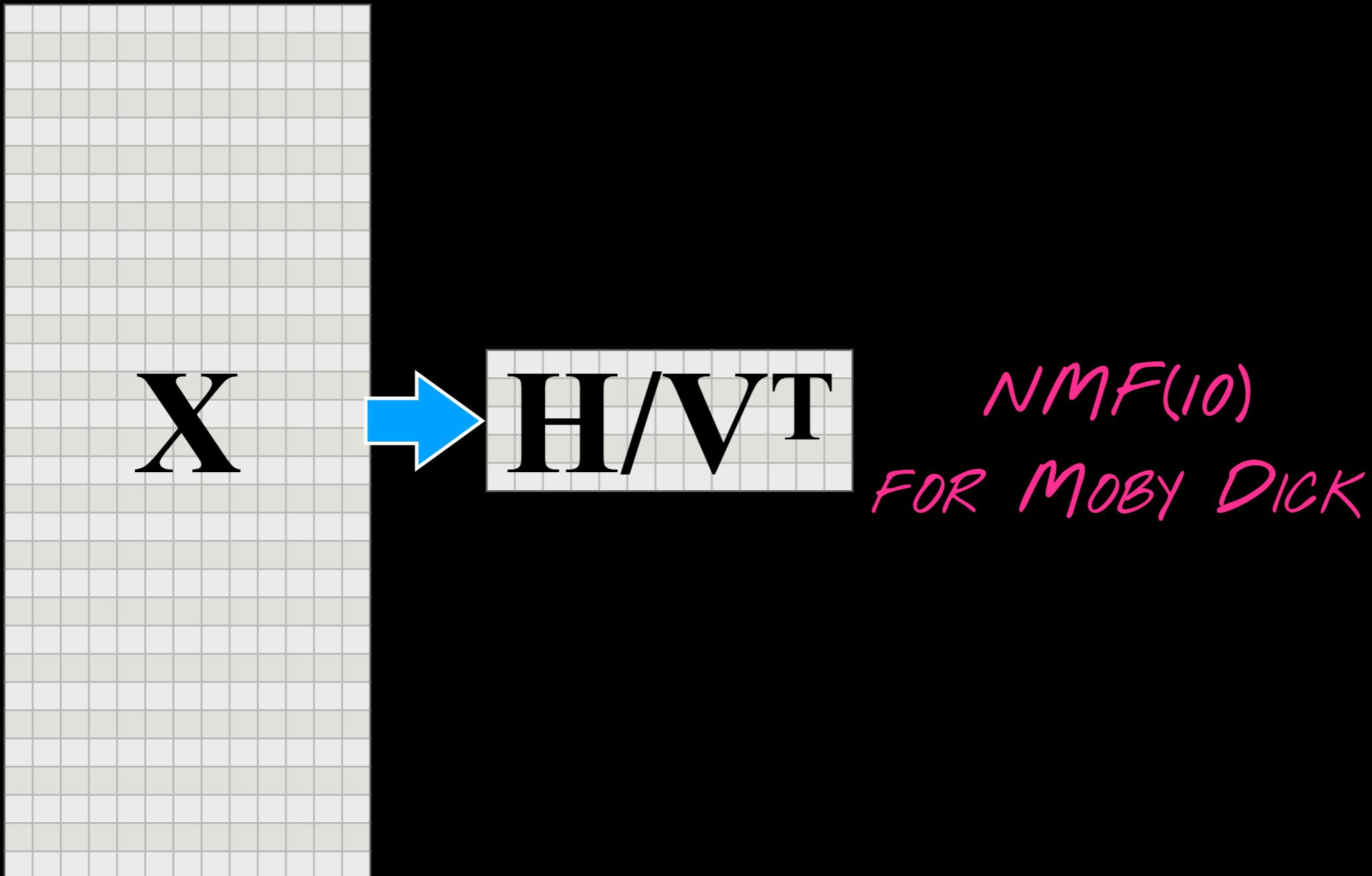
Yes, but: What is it Good for?

- Find latent **topic** dimensions (alternative: LDA)
- Find **word similarity** in latent space (alternative: Word2Vec)
- Find **document similarity** in latent space (alternative: Doc2Vec)
- Reduce dimensionality for **visualization**

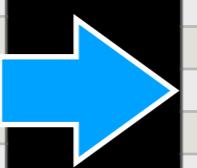
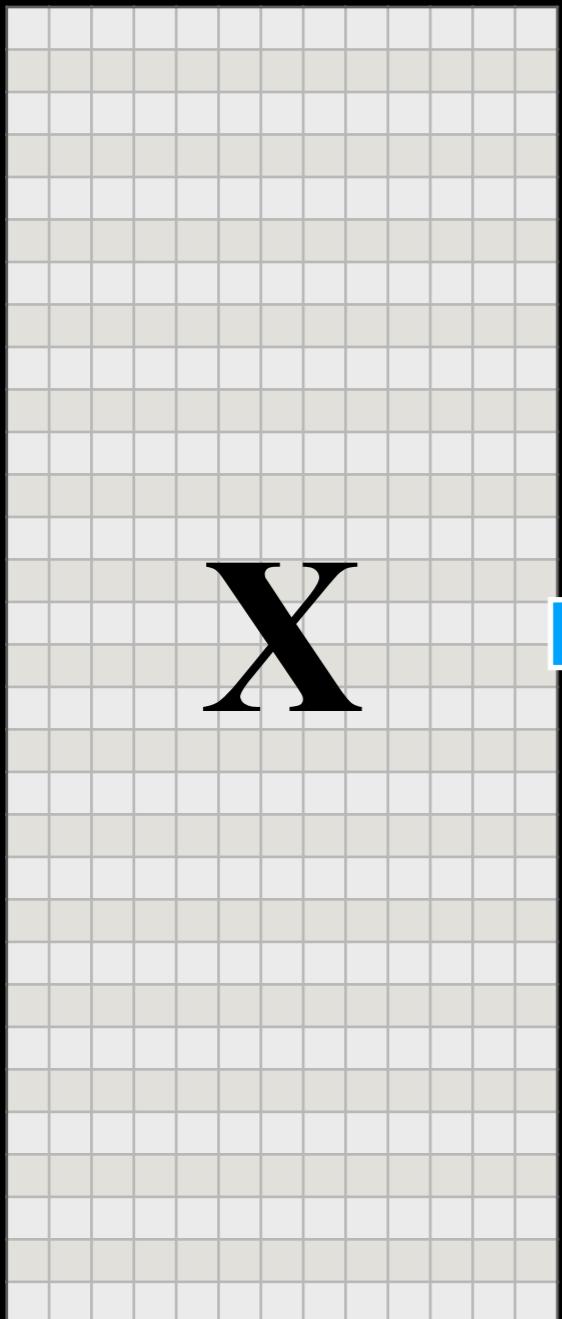
Latent Word Dimension Topics



Latent Word Dimension Topics



Latent Word Dimension Topics



EVT

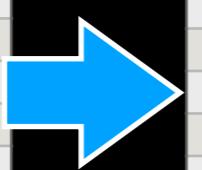
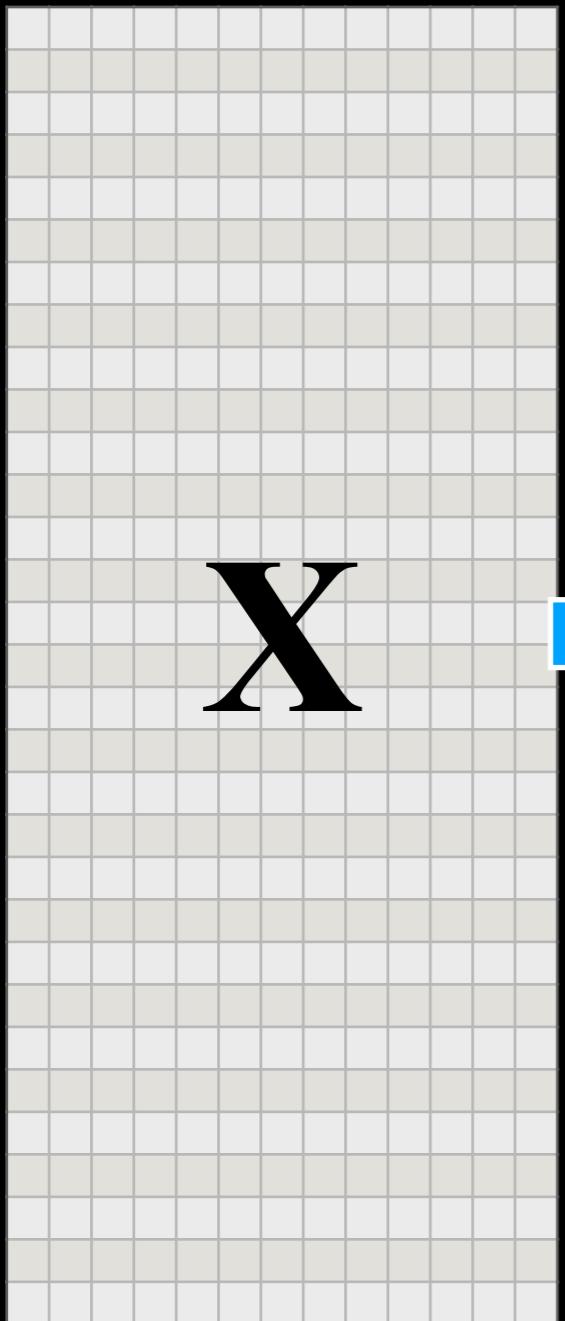
NMF(10)

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

Bocconi

Latent Word Dimension Topics



EVT

NMF(10)

FOR MOBY DICK

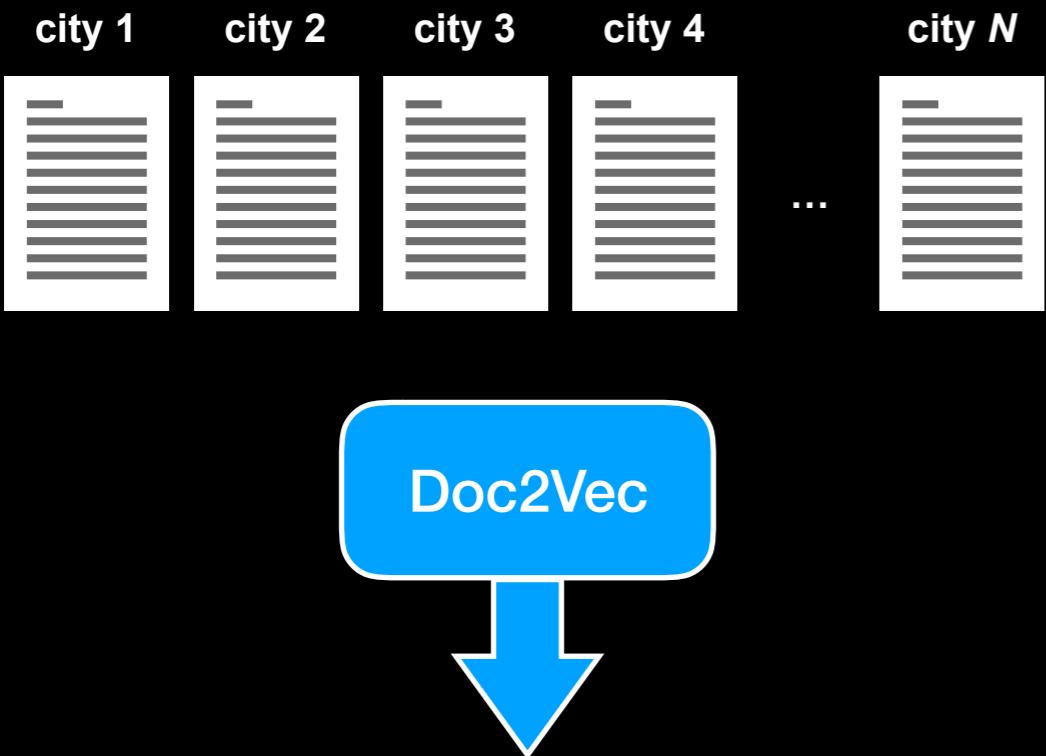
AHAB	ahab, captain, cried, captain ahab, cried ahab
STRUCTURE	chapter, folio, octavo, ii, iii
???	like, ship, sea, time, way
MEN	man, old, old man, look, young man
???	oh, life, starbuck, sweet, god
CHARACTERS	said, stubb, queequeg, don, starbuck
???	sir, aye, let, shall, think
OLD-TIMEY	thou, thee,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

Dimensionality Reduction for Visualizations

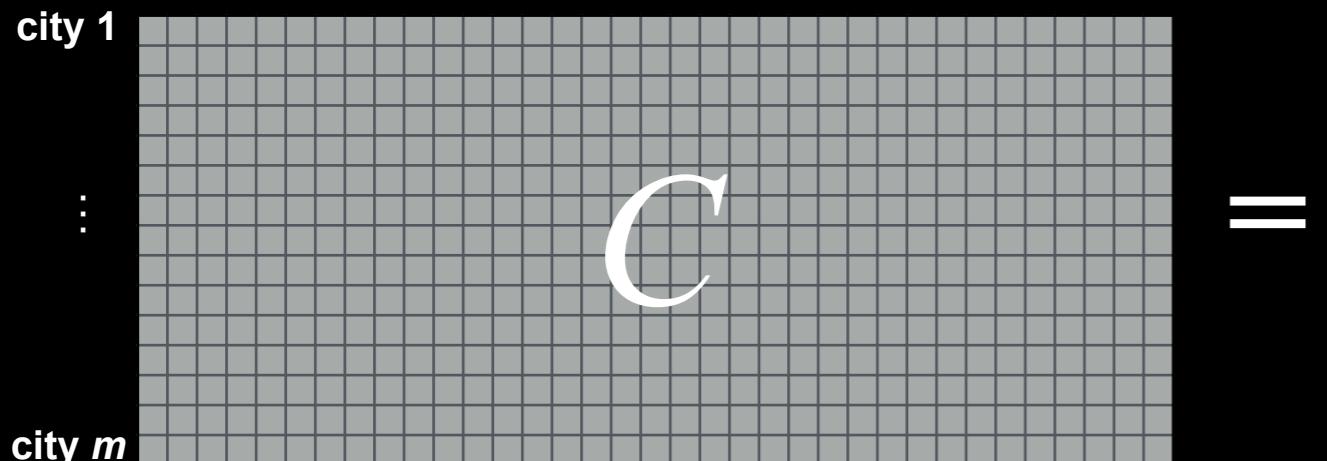
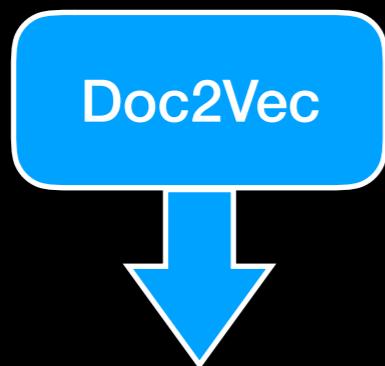
Dimensions as RGB



Dimensions as RGB

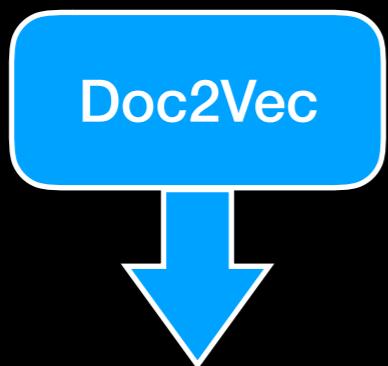
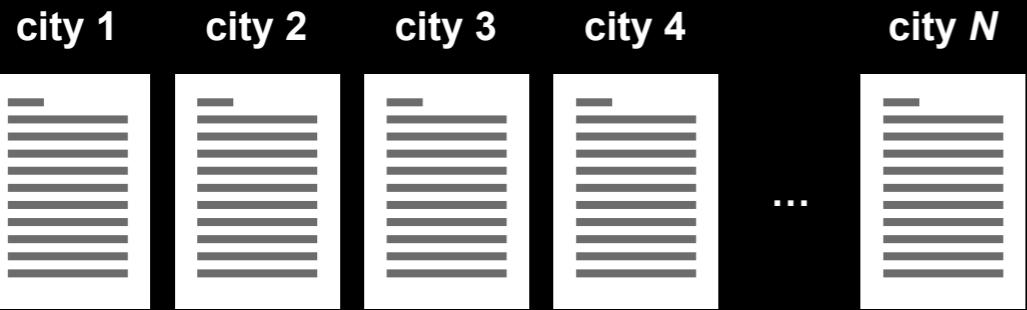


Dimensions as RGB



DENSE REPRESENTATION

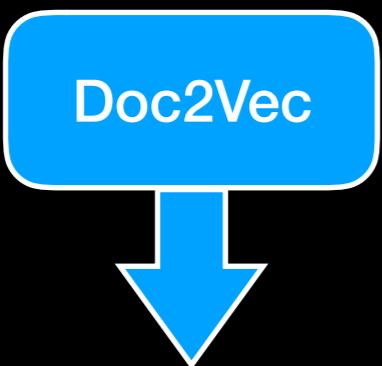
Dimensions as RGB



$$\begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \mathbf{C} \quad = \quad \begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \mathbf{V}$$

DENSE REPRESENTATION

Dimensions as RGB



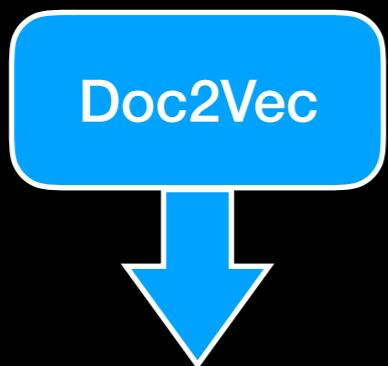
$$\begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \begin{matrix} C \\ \quad \end{matrix} = \quad \begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \begin{matrix} V \\ \quad \end{matrix} \times \quad \begin{matrix} H \\ \quad \end{matrix}$$

NMF

[Diagram showing the matrix factorization equation for Non-negative Matrix Factorization (NMF). On the left, a vertical column of cities from 1 to m is multiplied by a matrix C. This equals a vertical column of cities from 1 to m multiplied by a matrix V, which is then multiplied by a matrix H. The word 'NMF' is written in pink at the bottom right.]

DENSE REPRESENTATION

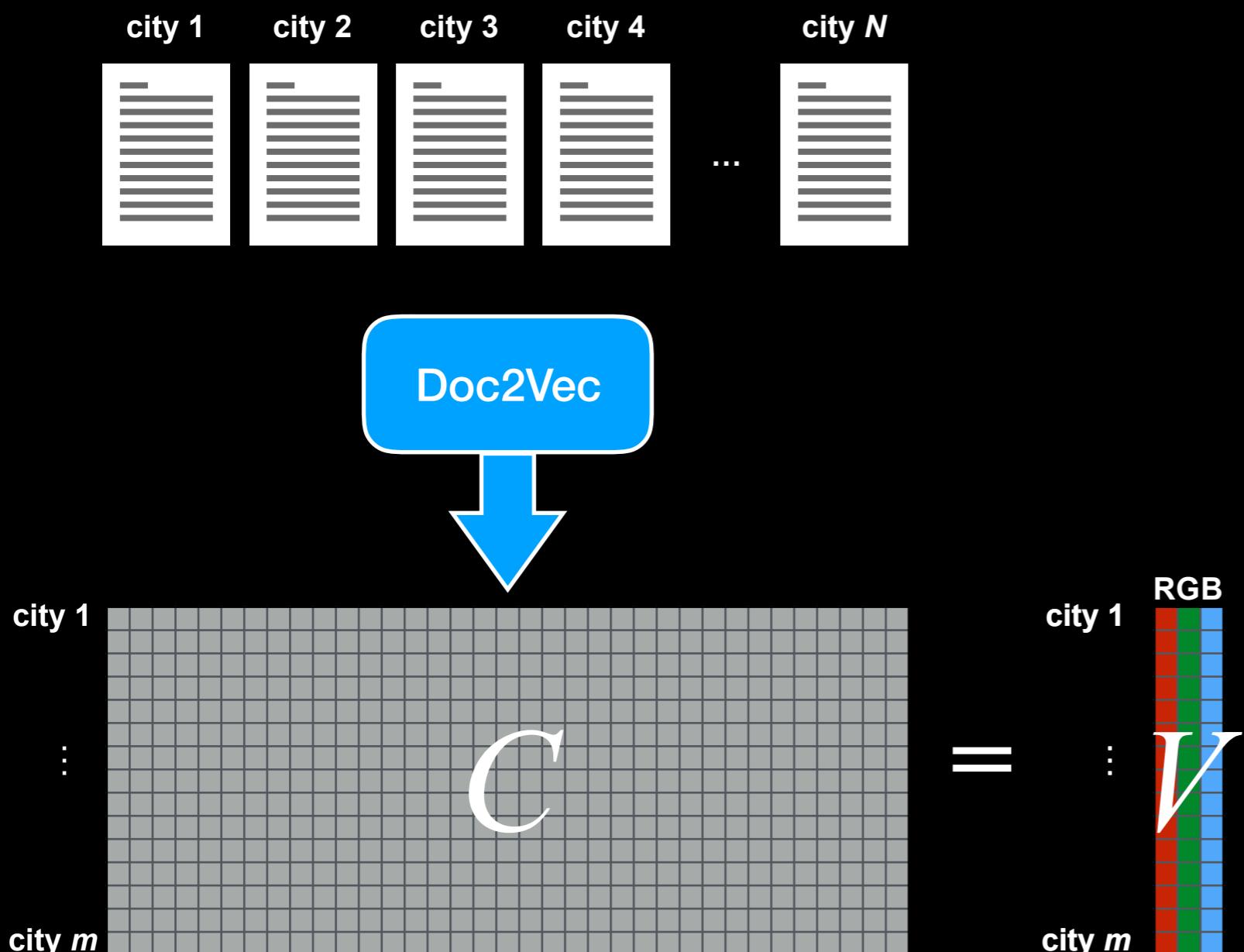
Dimensions as RGB



$$\begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \mathbf{C} \quad = \quad \begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \mathbf{V}$$

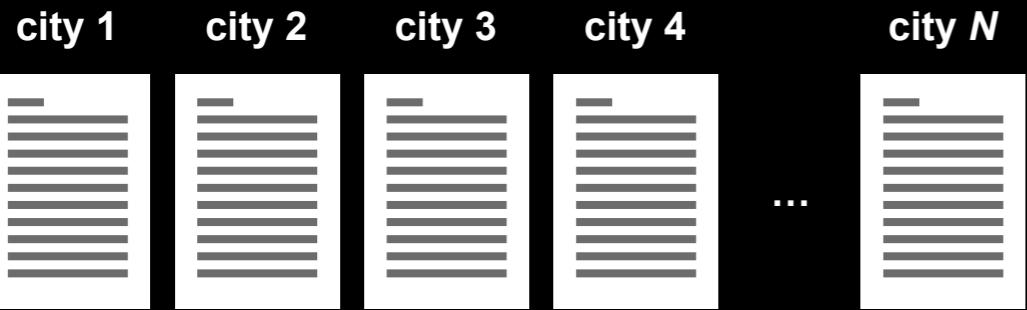
DENSE REPRESENTATION

Dimensions as RGB



DENSE REPRESENTATION

Dimensions as RGB

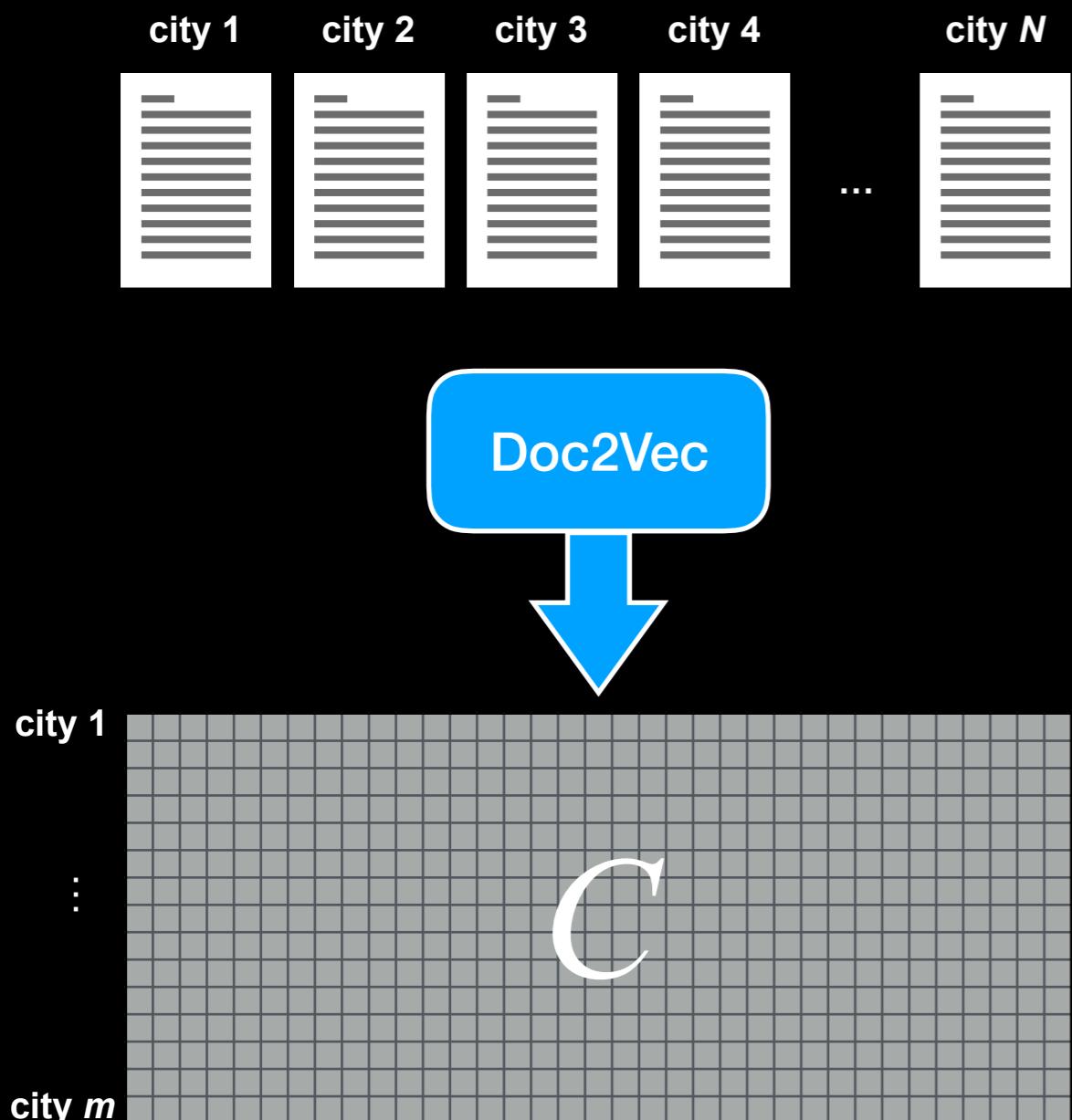


Doc2Vec

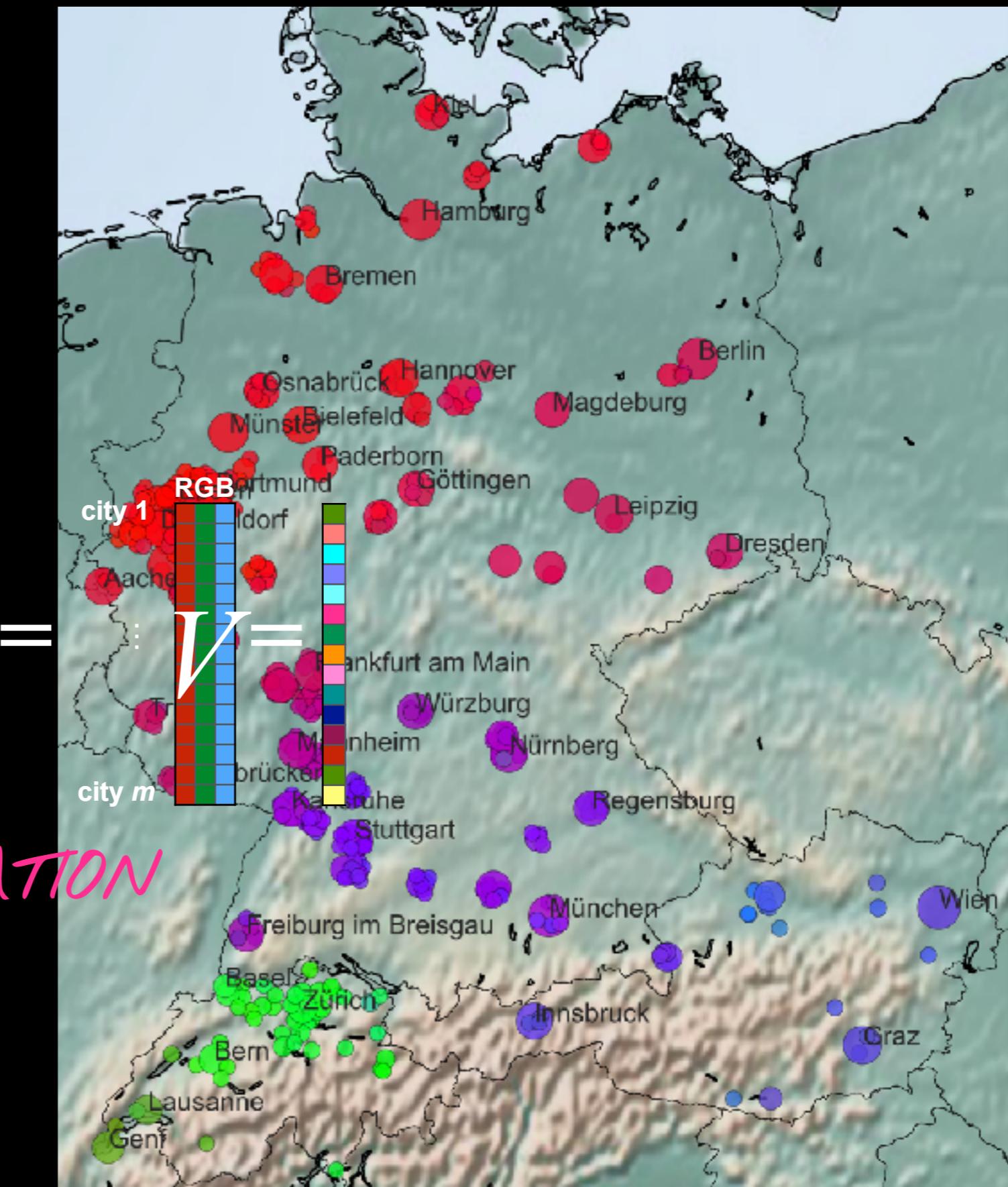
$$\begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \mathbf{C} = \begin{matrix} \text{city 1} \\ \vdots \\ \text{city } m \end{matrix} \quad \mathbf{V} = \begin{matrix} \text{RGB} \\ \vdots \\ \text{city } m \end{matrix}$$

DENSE REPRESENTATION

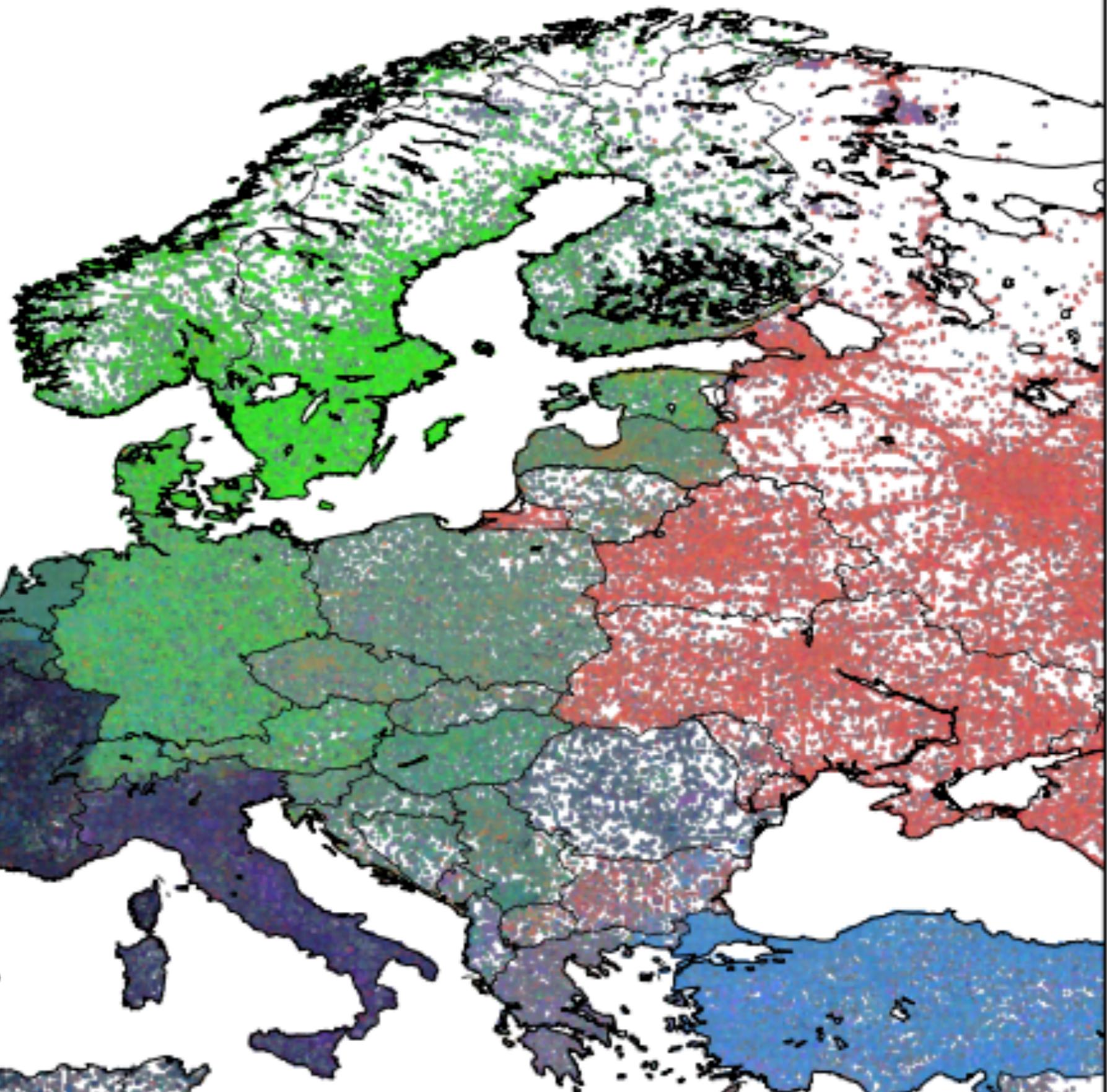
Dimensions as RGB



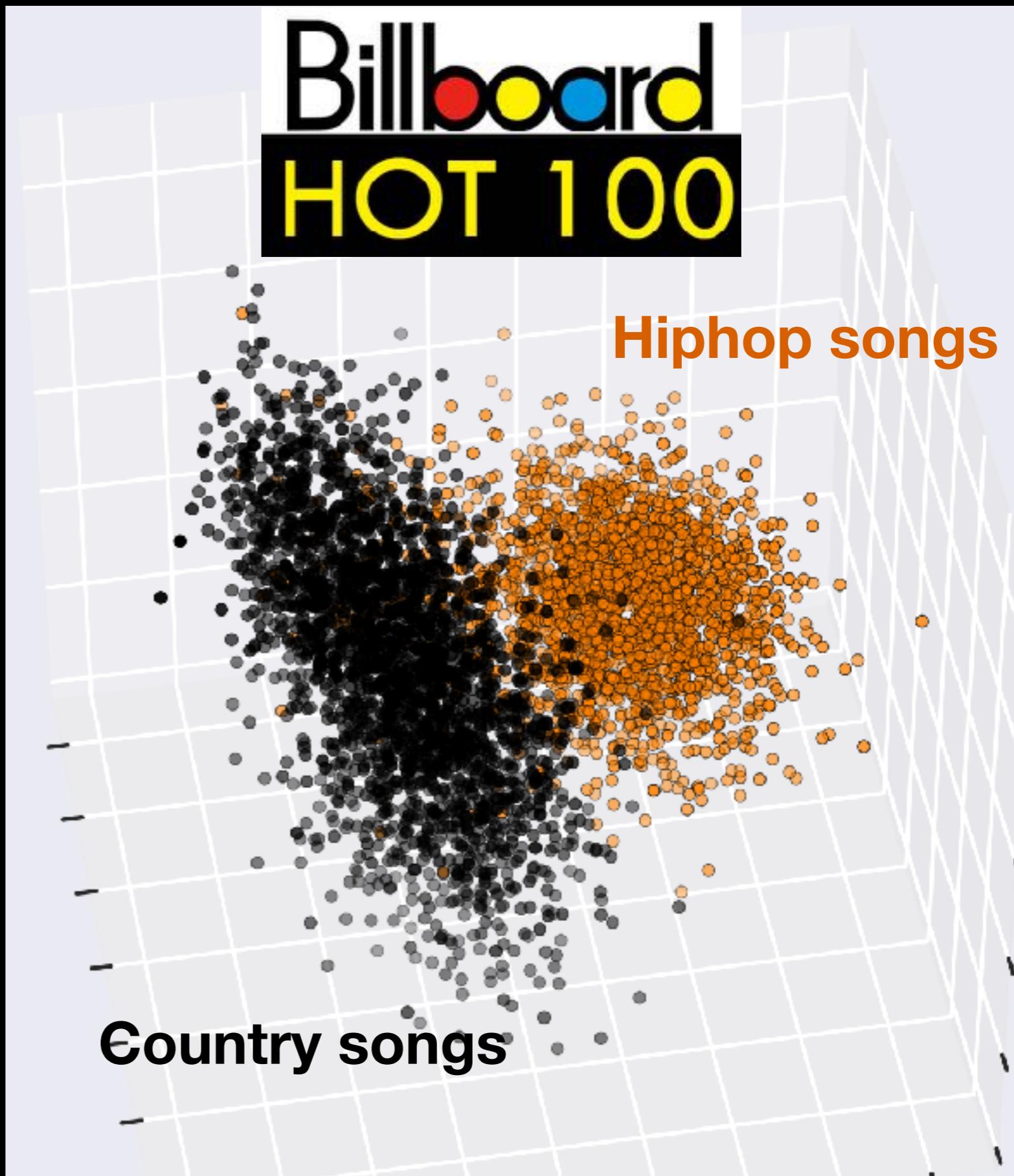
DENSE REPRESENTATION



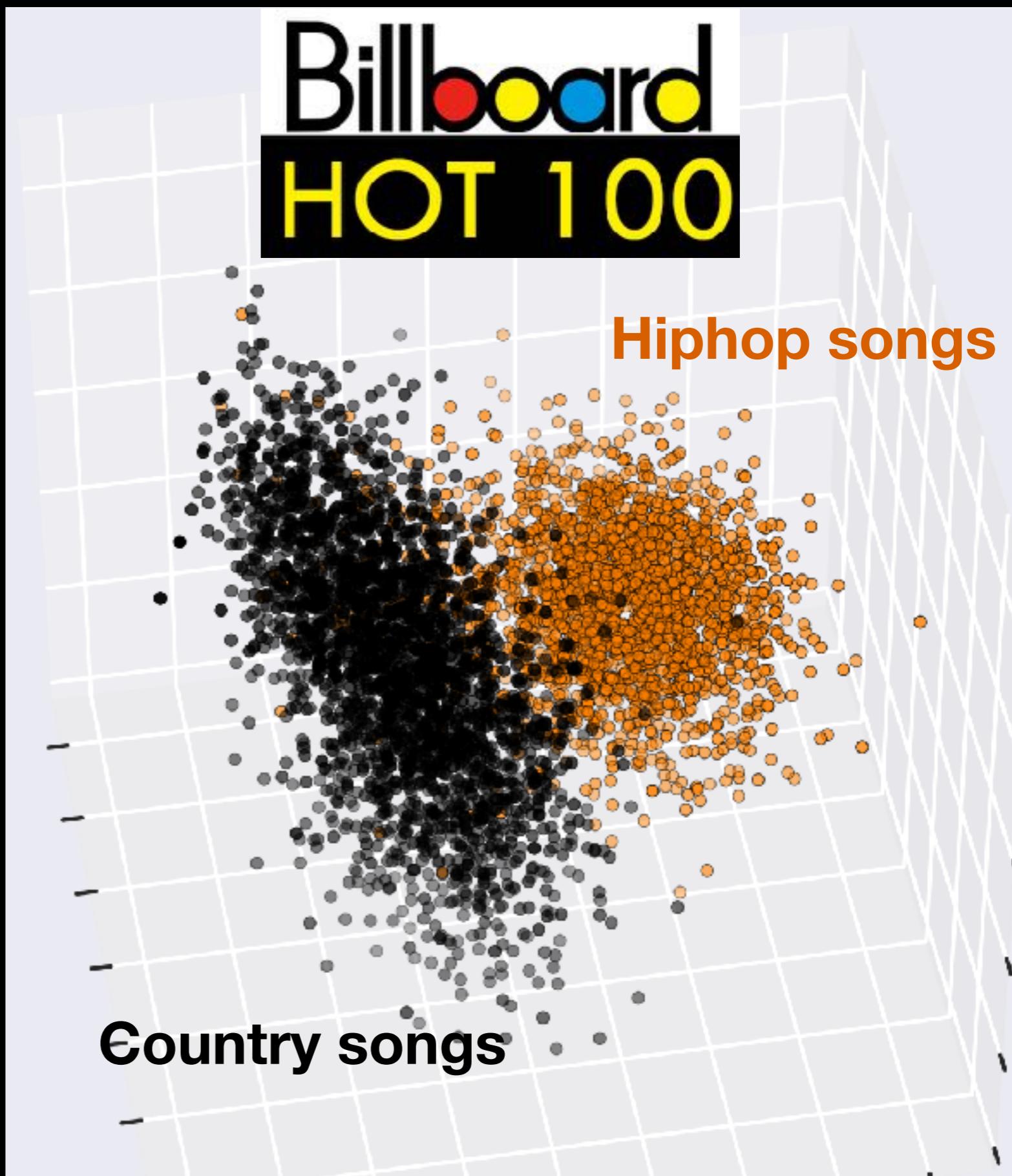
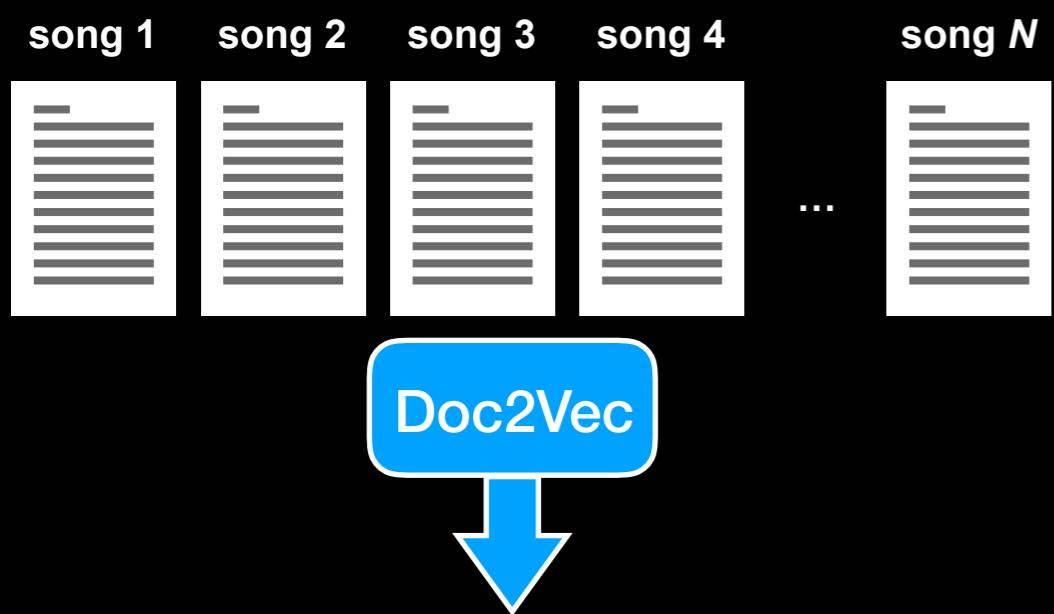
Dimensions as RGB



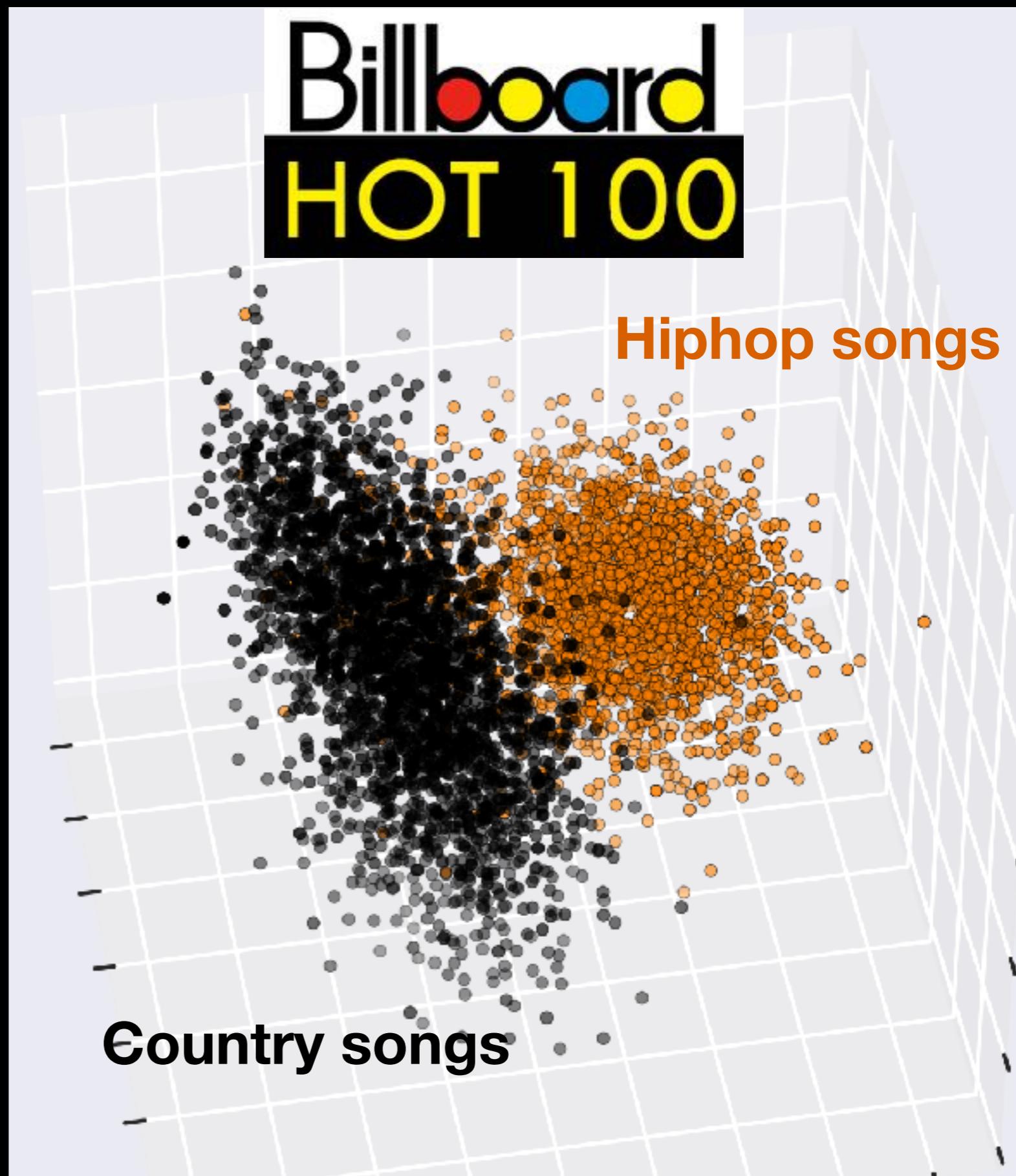
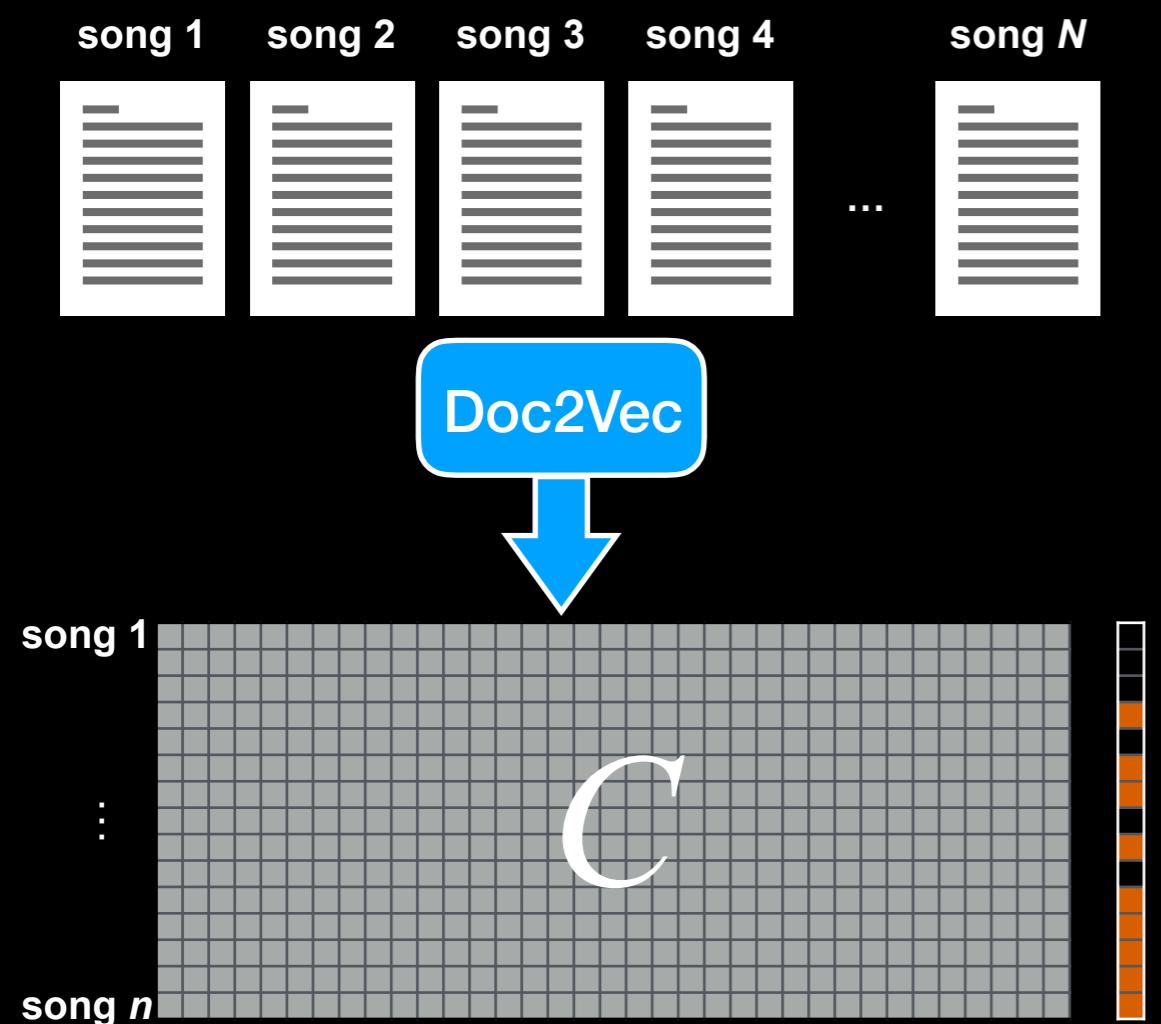
Dimensions as Coordinates



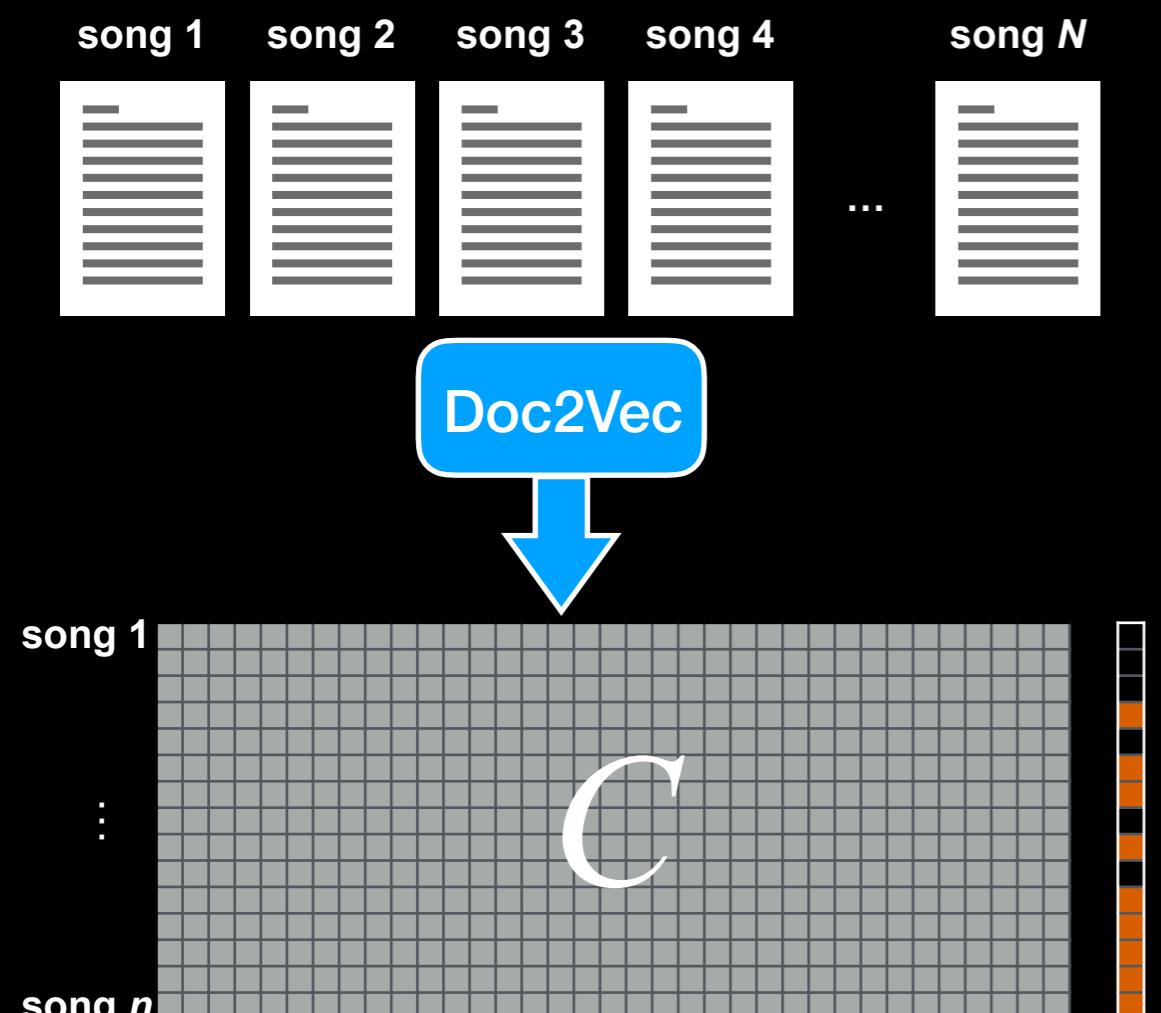
Dimensions as Coordinates



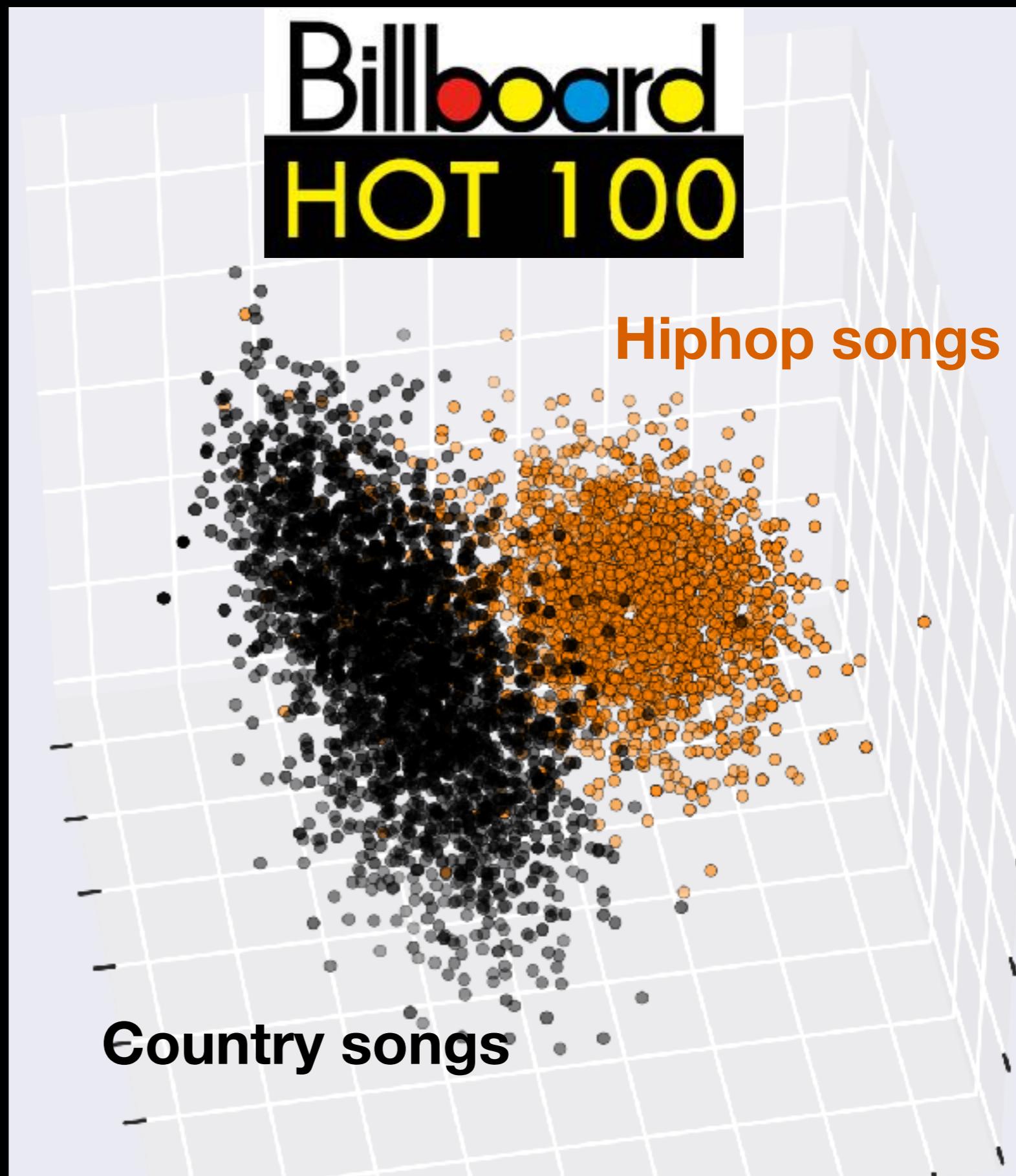
Dimensions as Coordinates



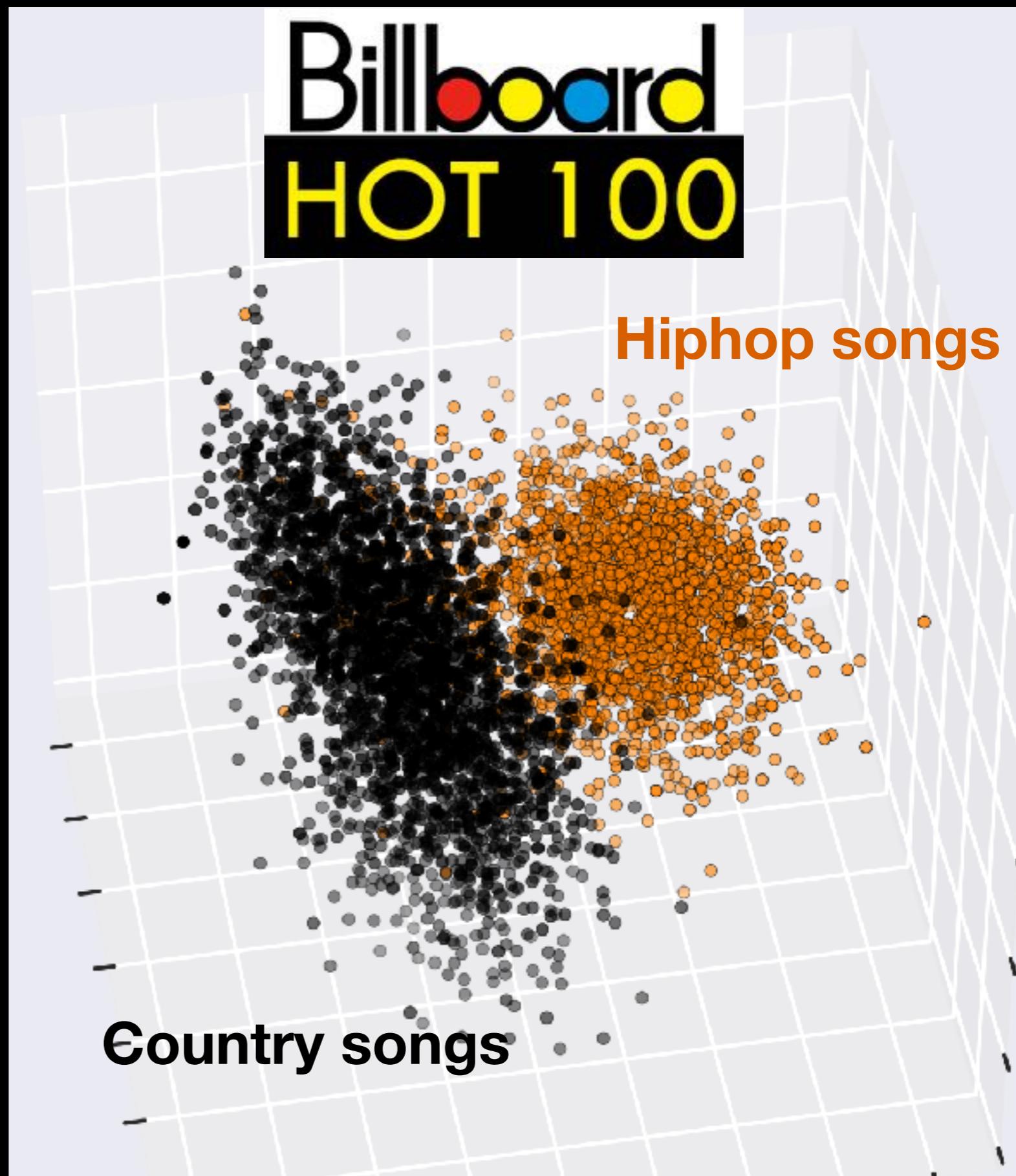
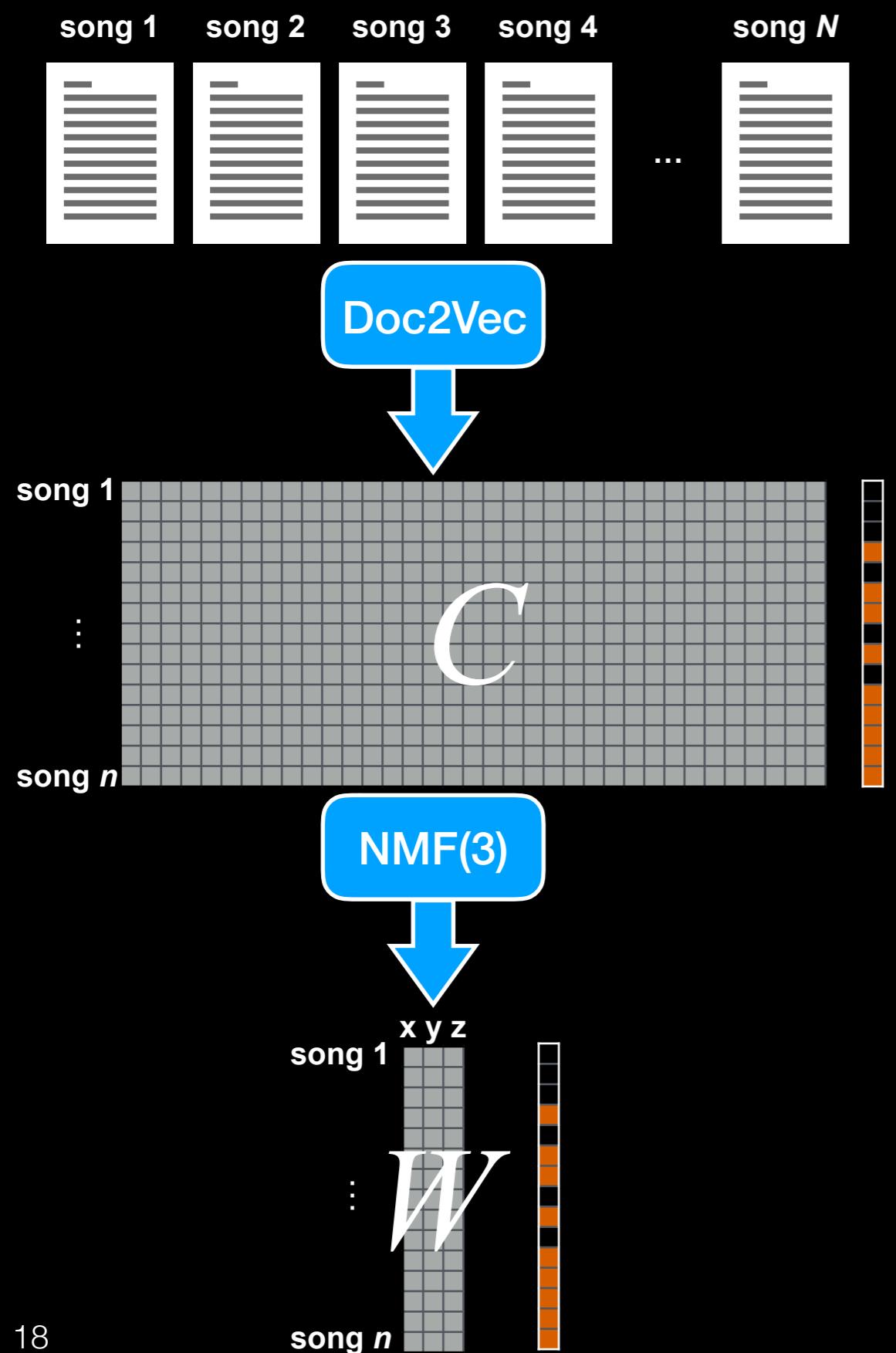
Dimensions as Coordinates



↓
NMF(3)



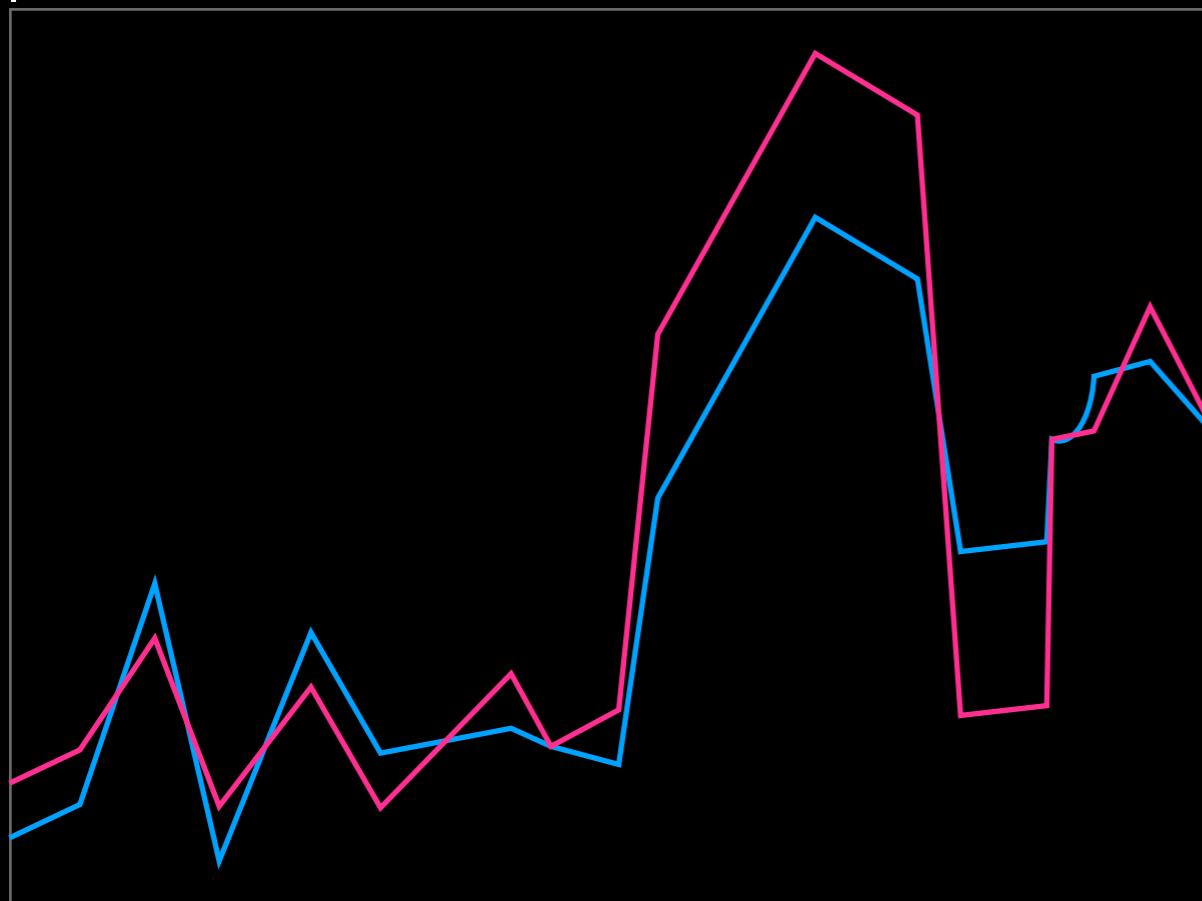
Dimensions as Coordinates



t-SNE

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

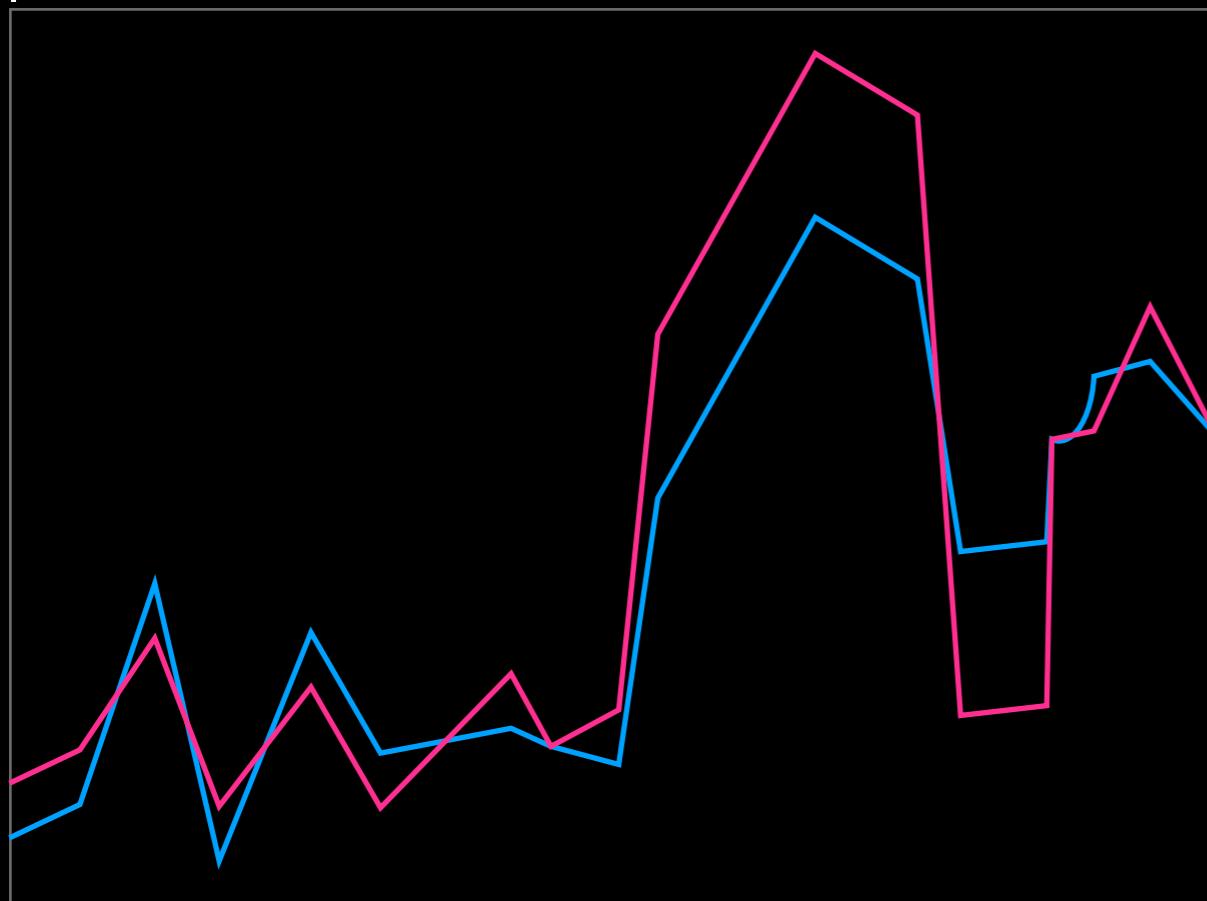
$P(n_j|n_i)$



t-SNE

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

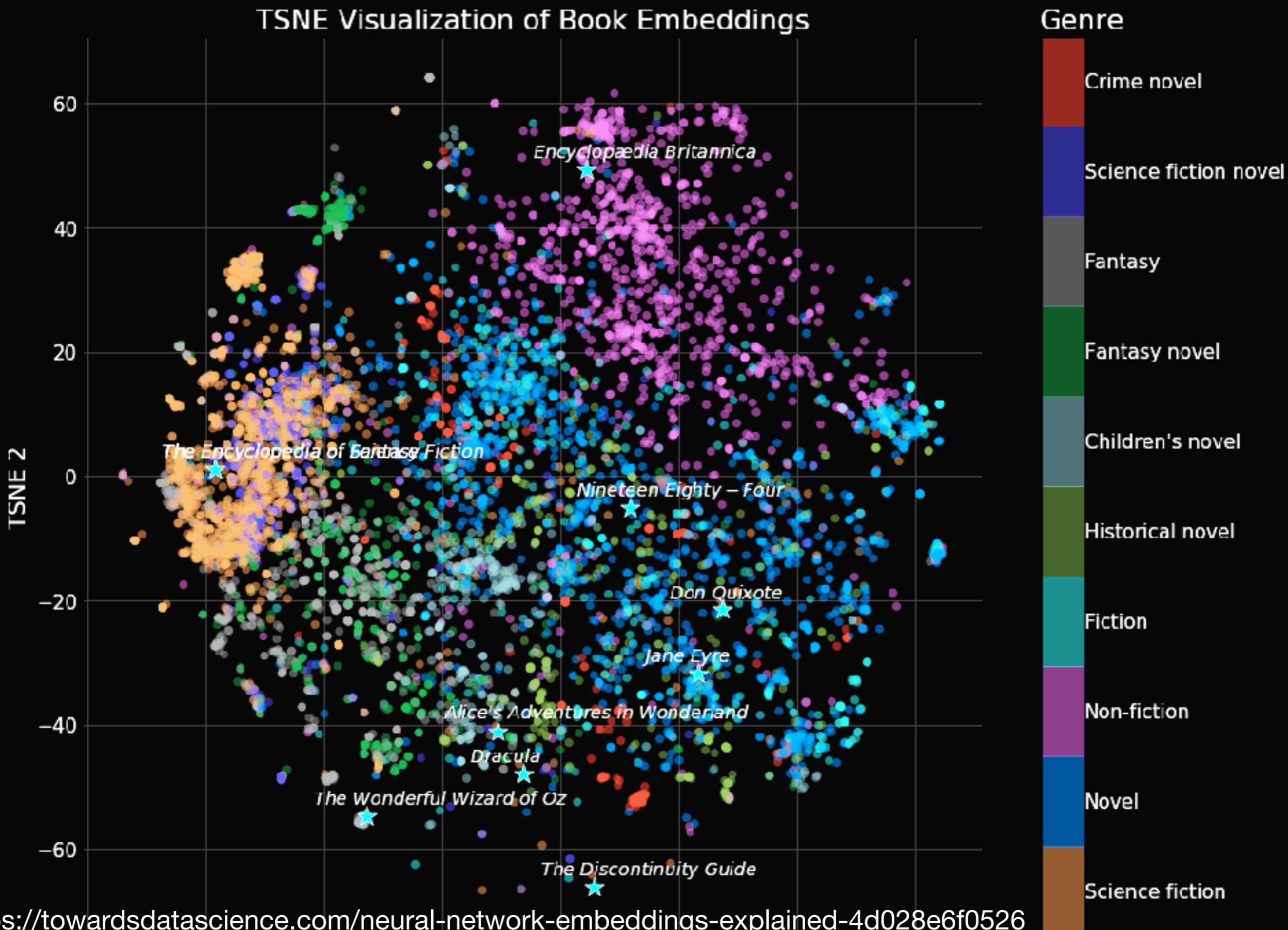
$P(n_j|n_i)$



- + Nicer spatial images
- Stochastic (many runs)
- Many parameters

t-SNE

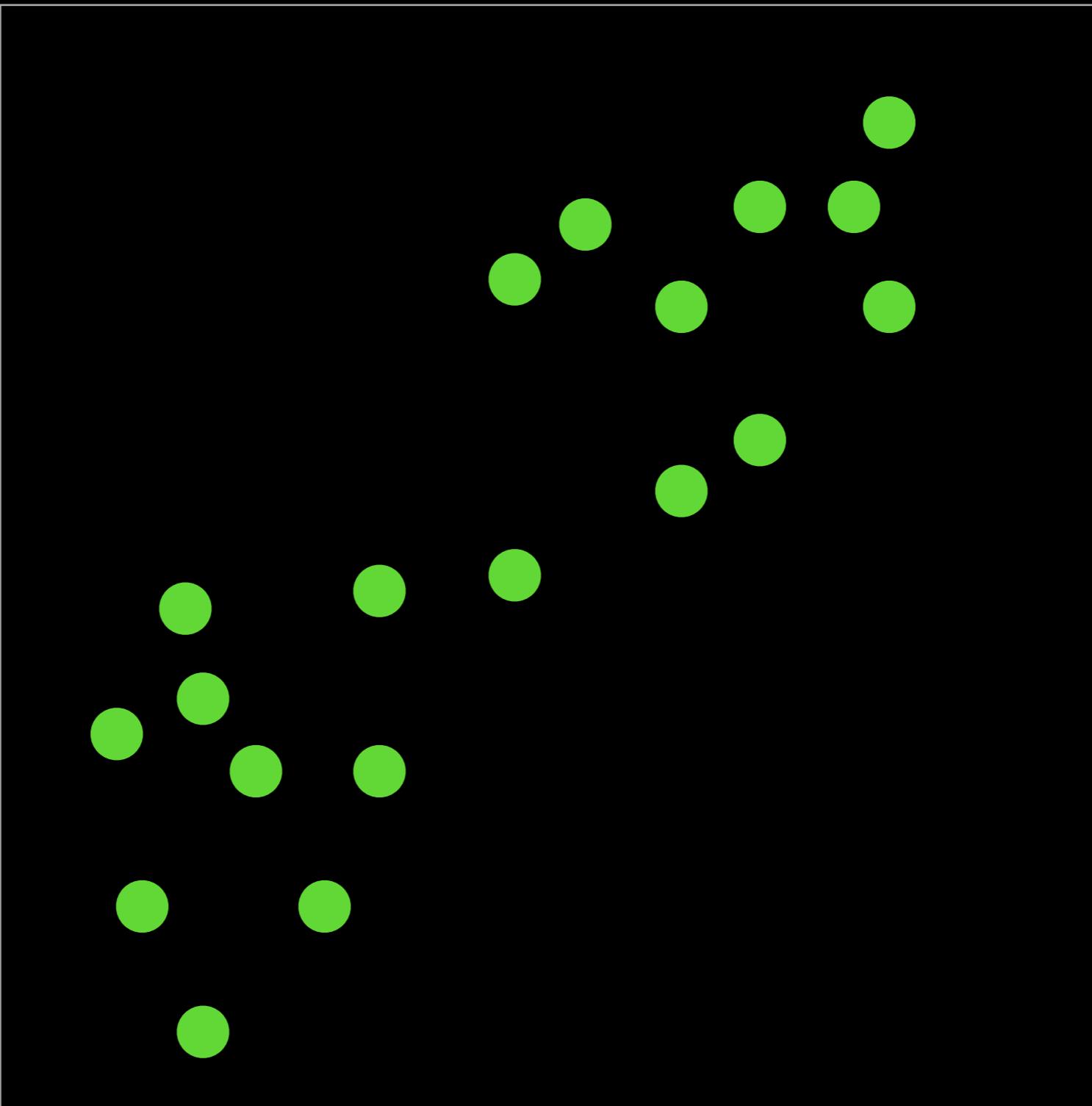
TSNE Visualization of Book Embeddings



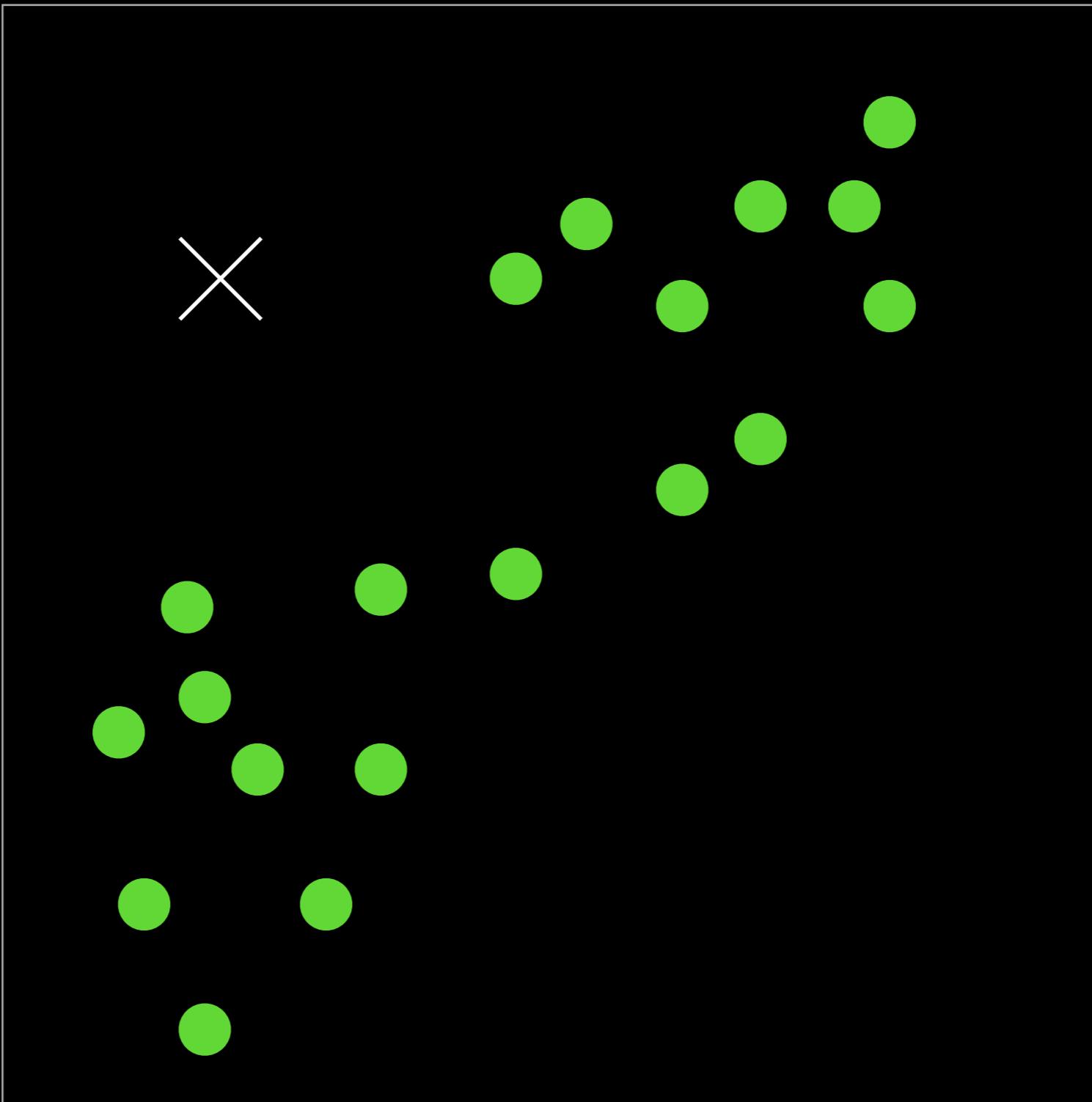
Clustering

k -Means Clustering

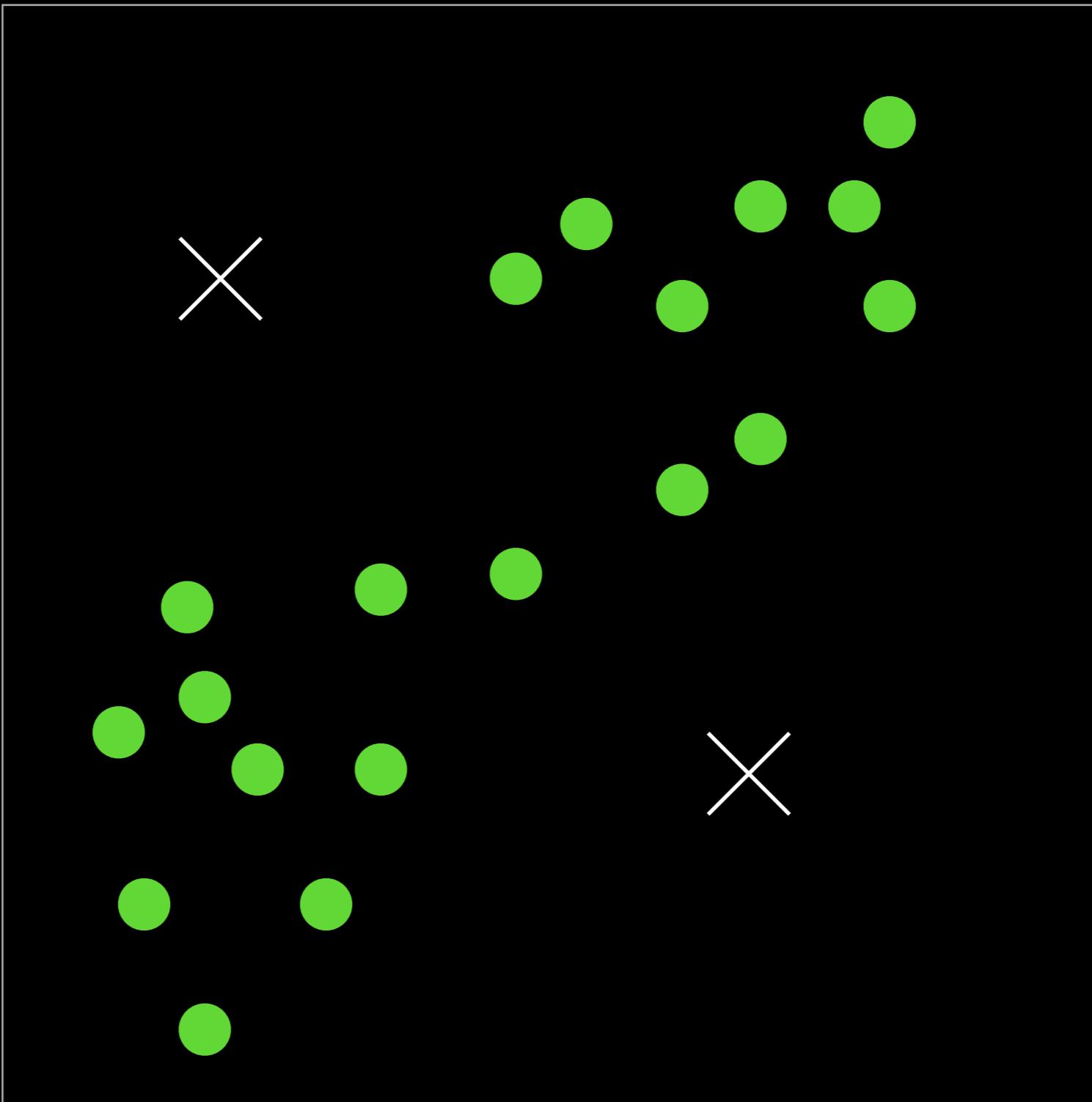
k -Means



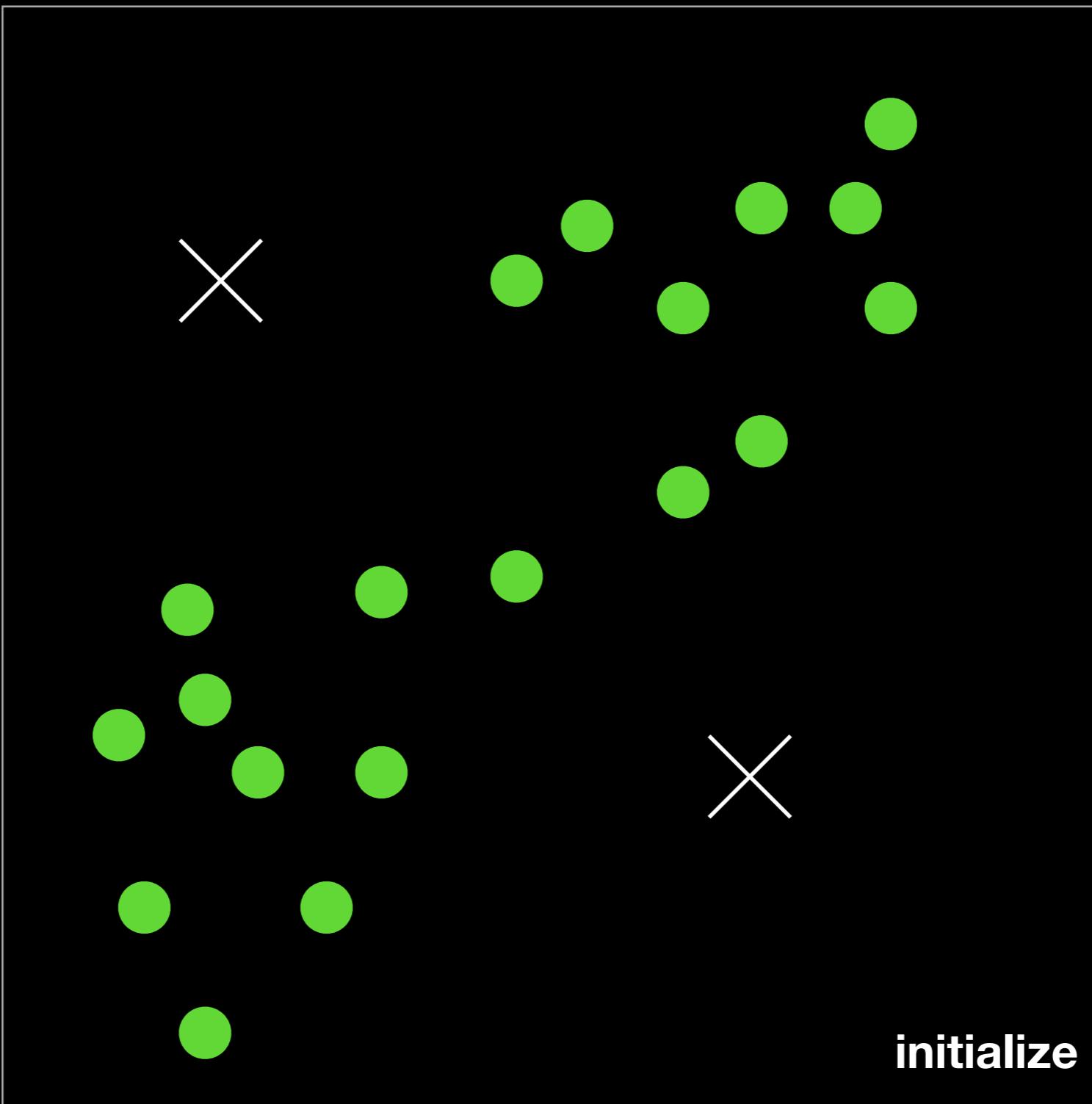
k -Means



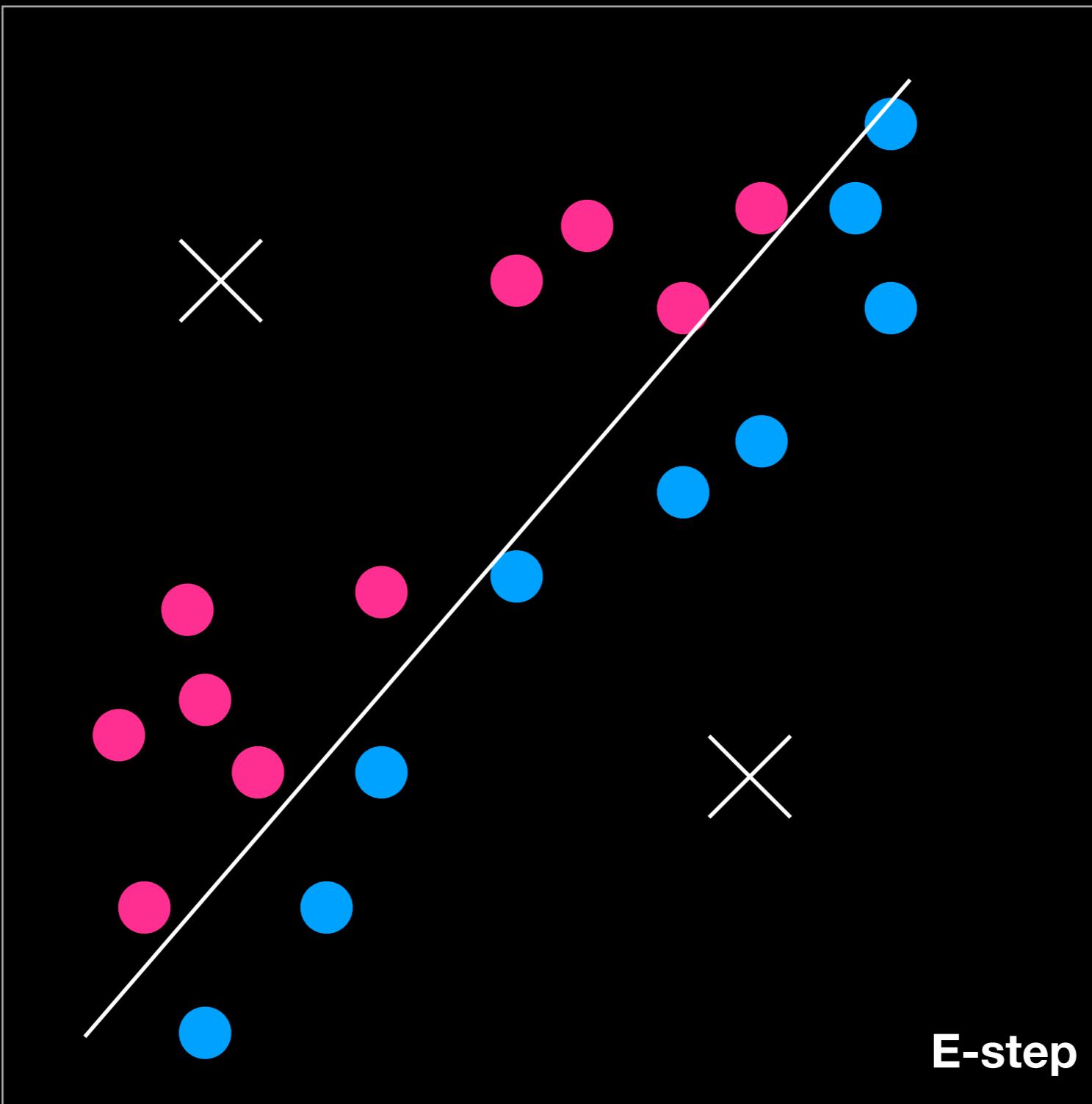
k -Means



k -Means

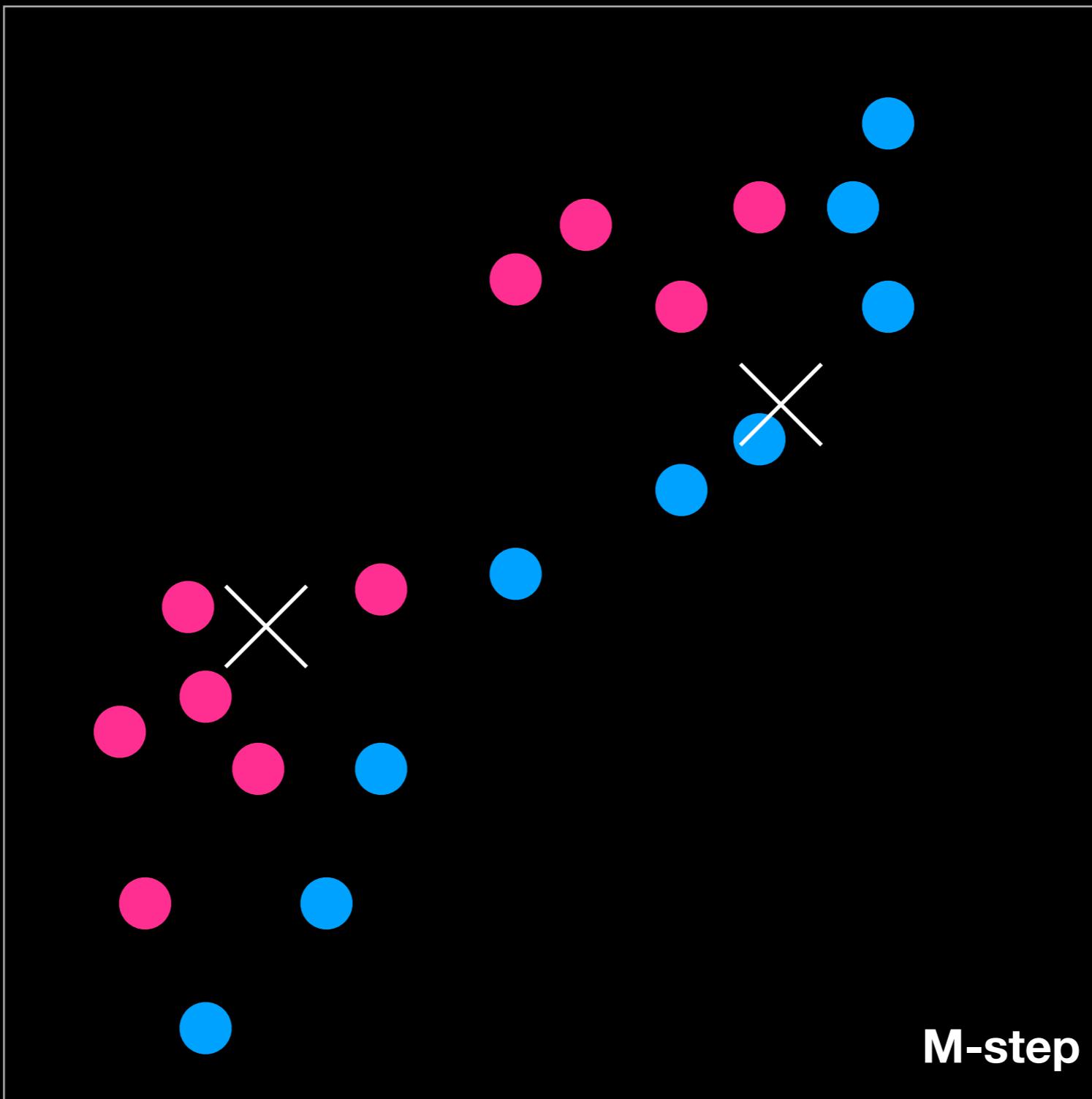


k -Means



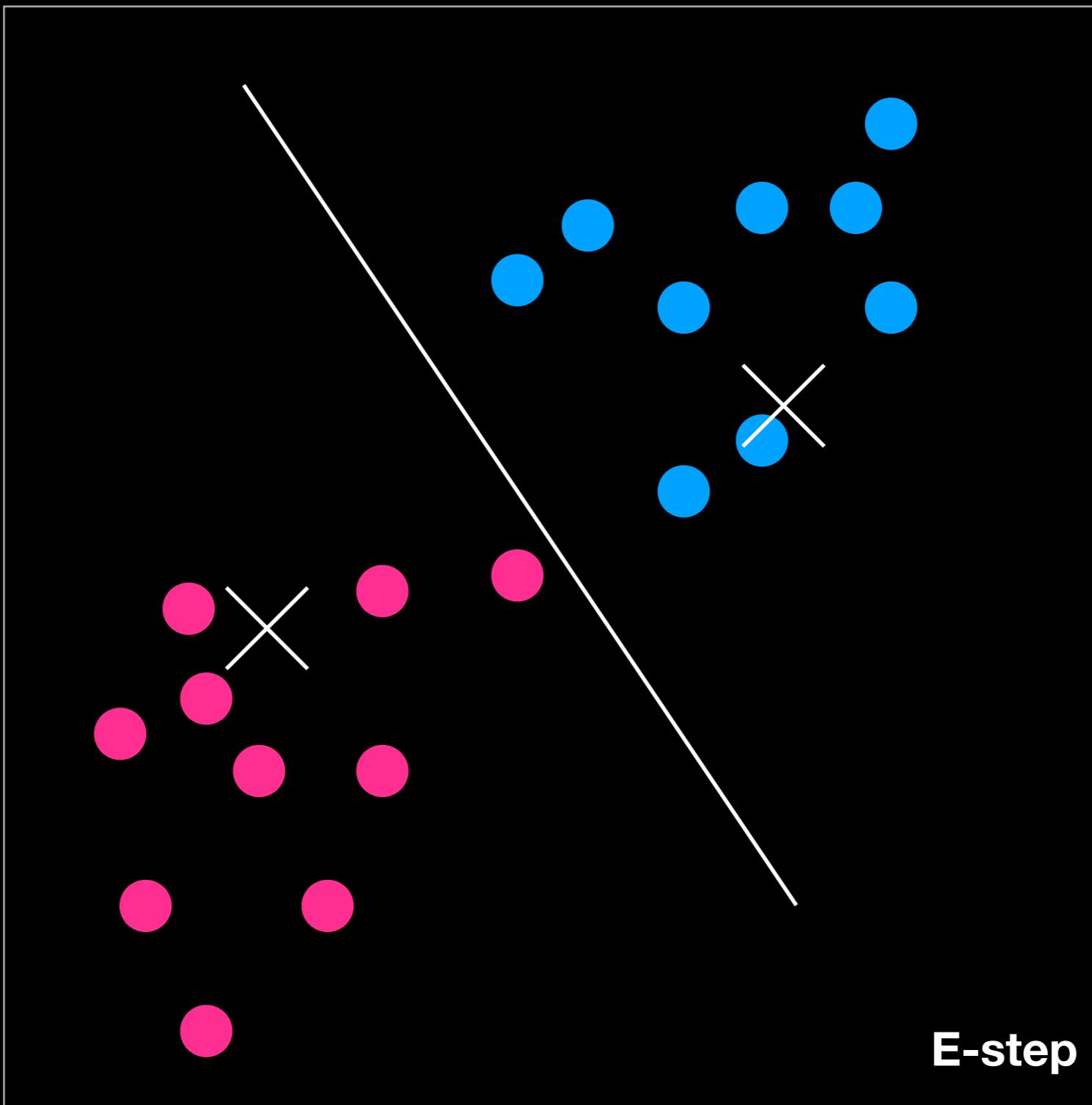
ASSIGN POINTS TO CENTROIDS

k -Means

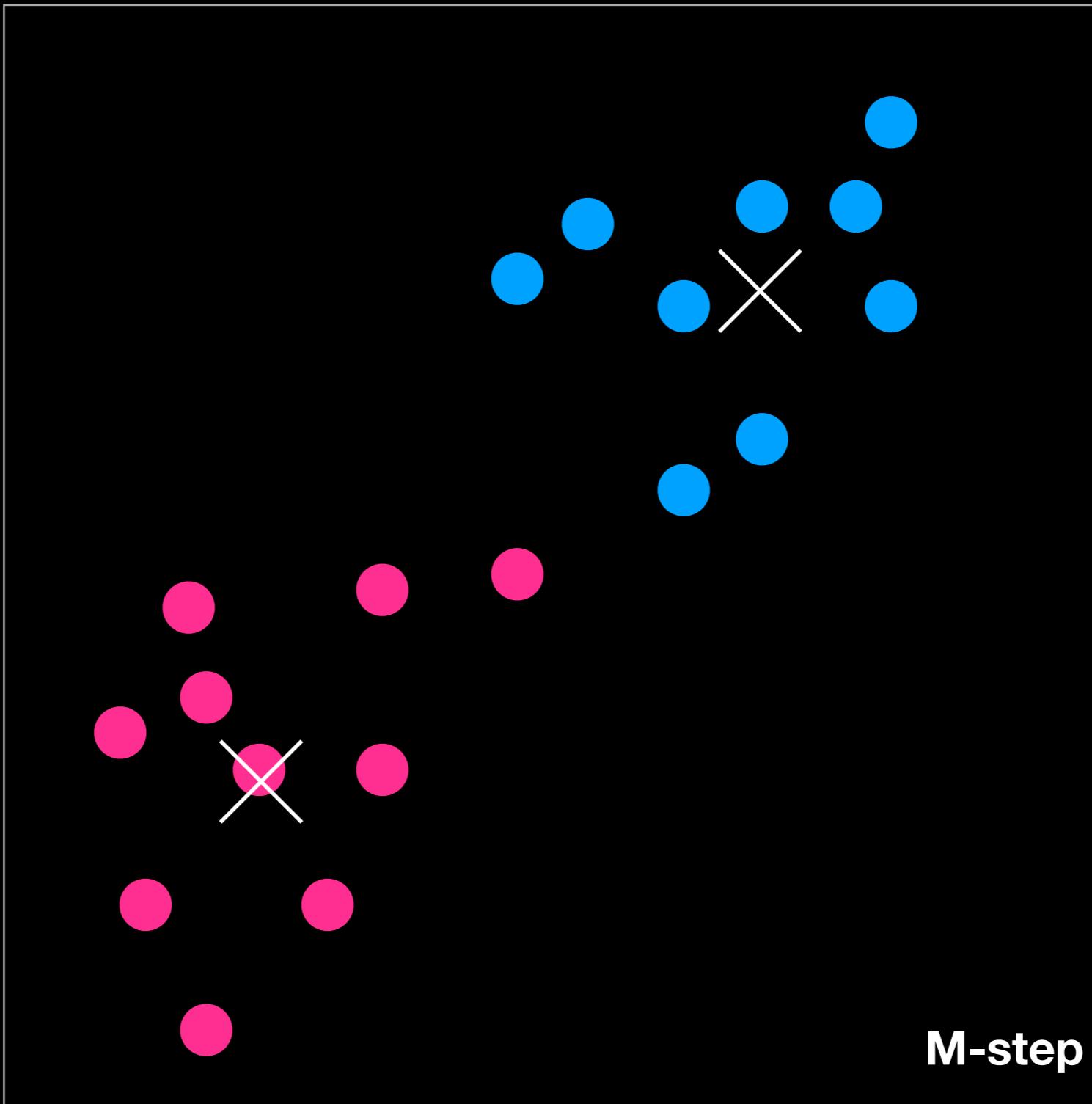


RECOMPUTE CENTROIDS

k -Means



k -Means

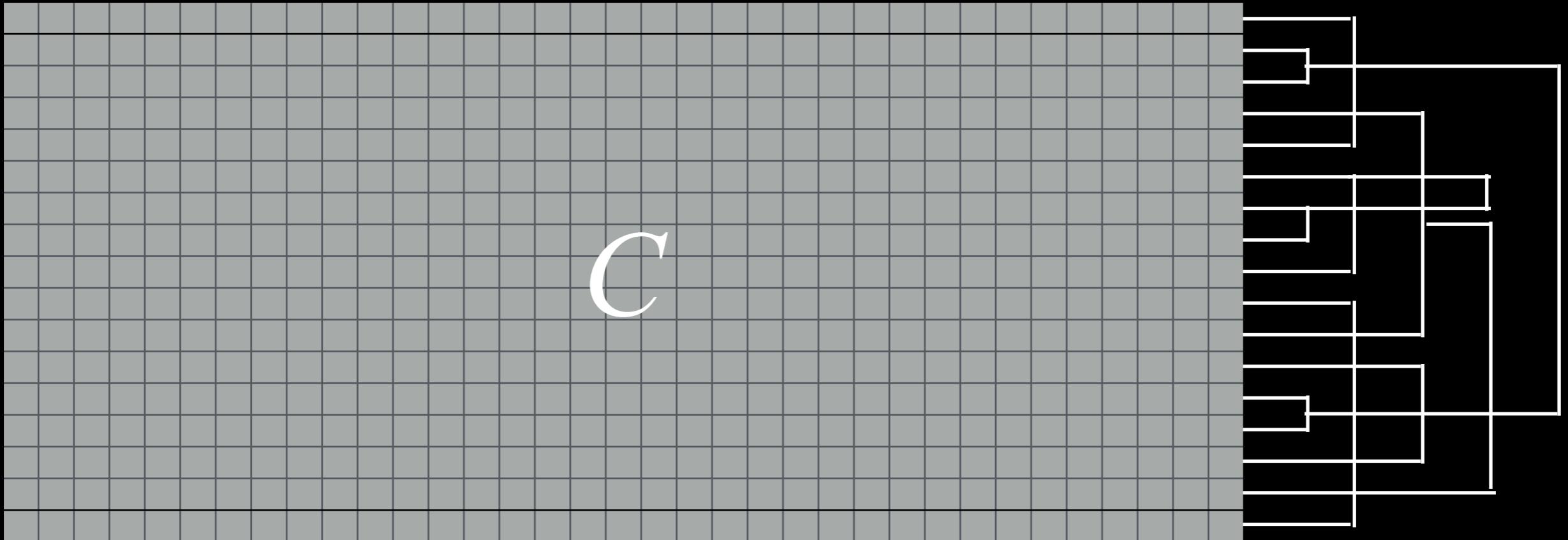


Agglomerative Clustering

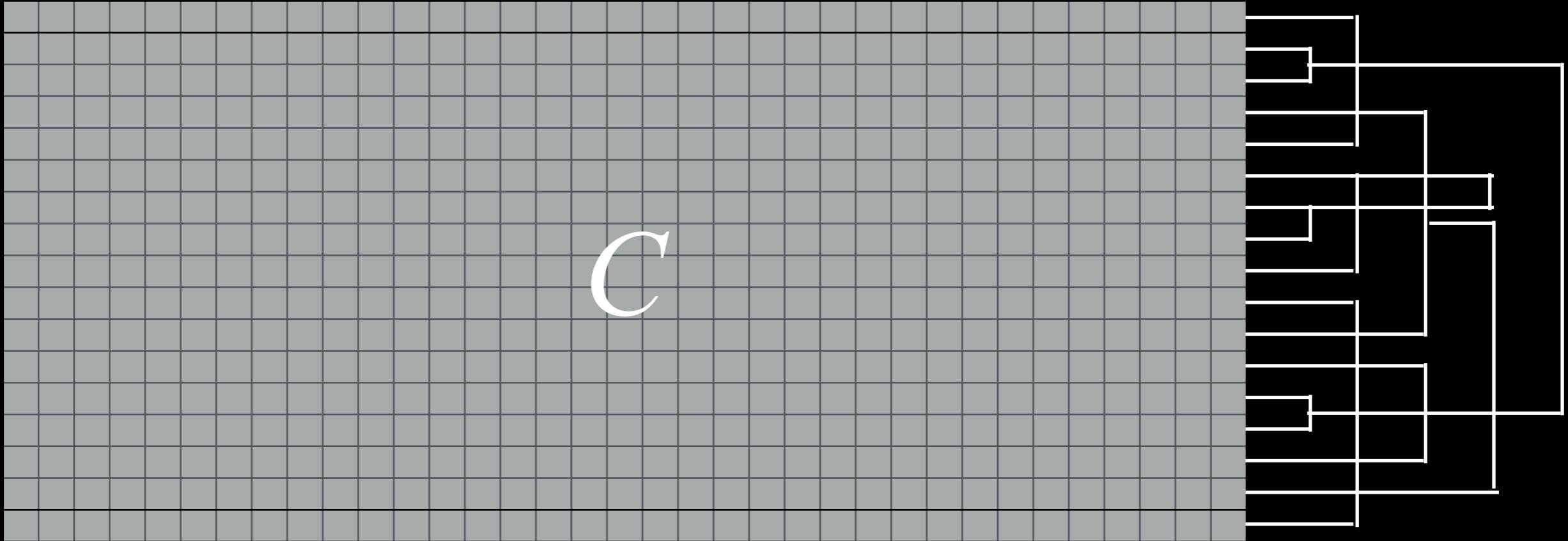
Building Up

C

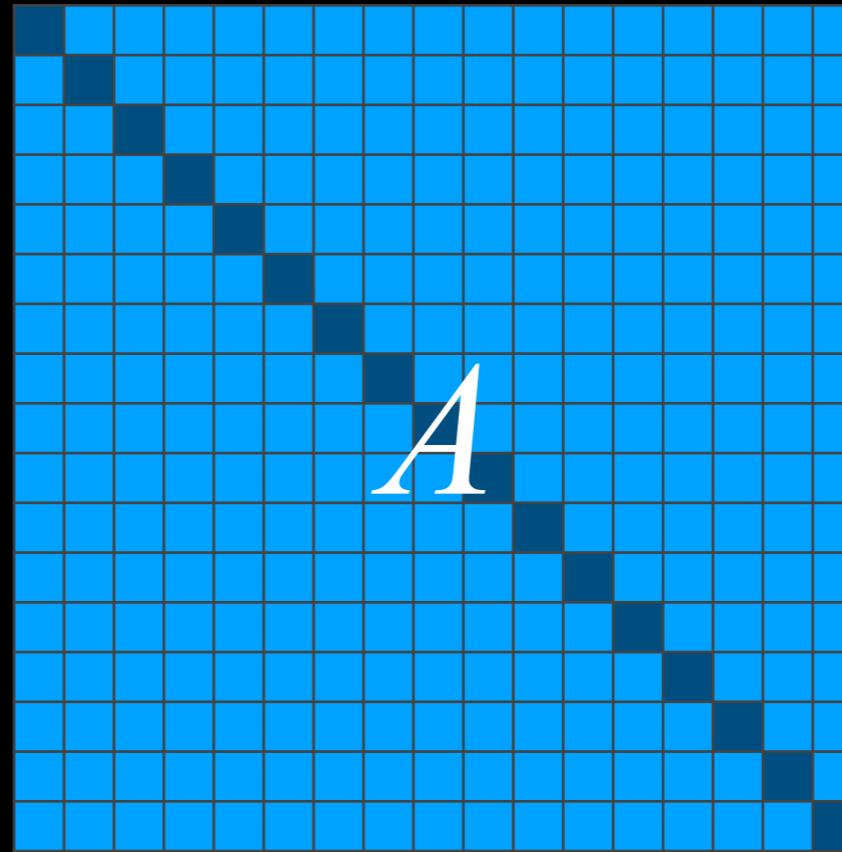
Building Up



Building Up

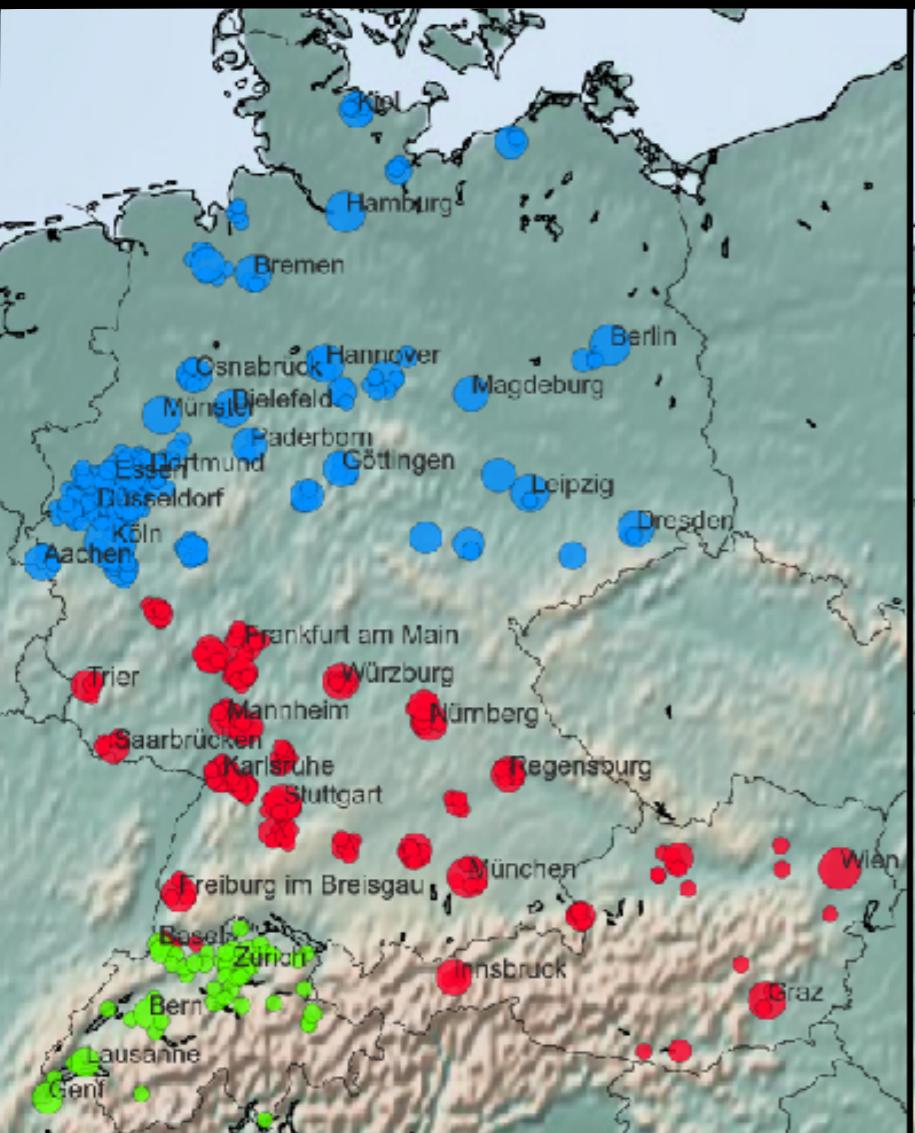


*ADJACENCY
MATRIX*

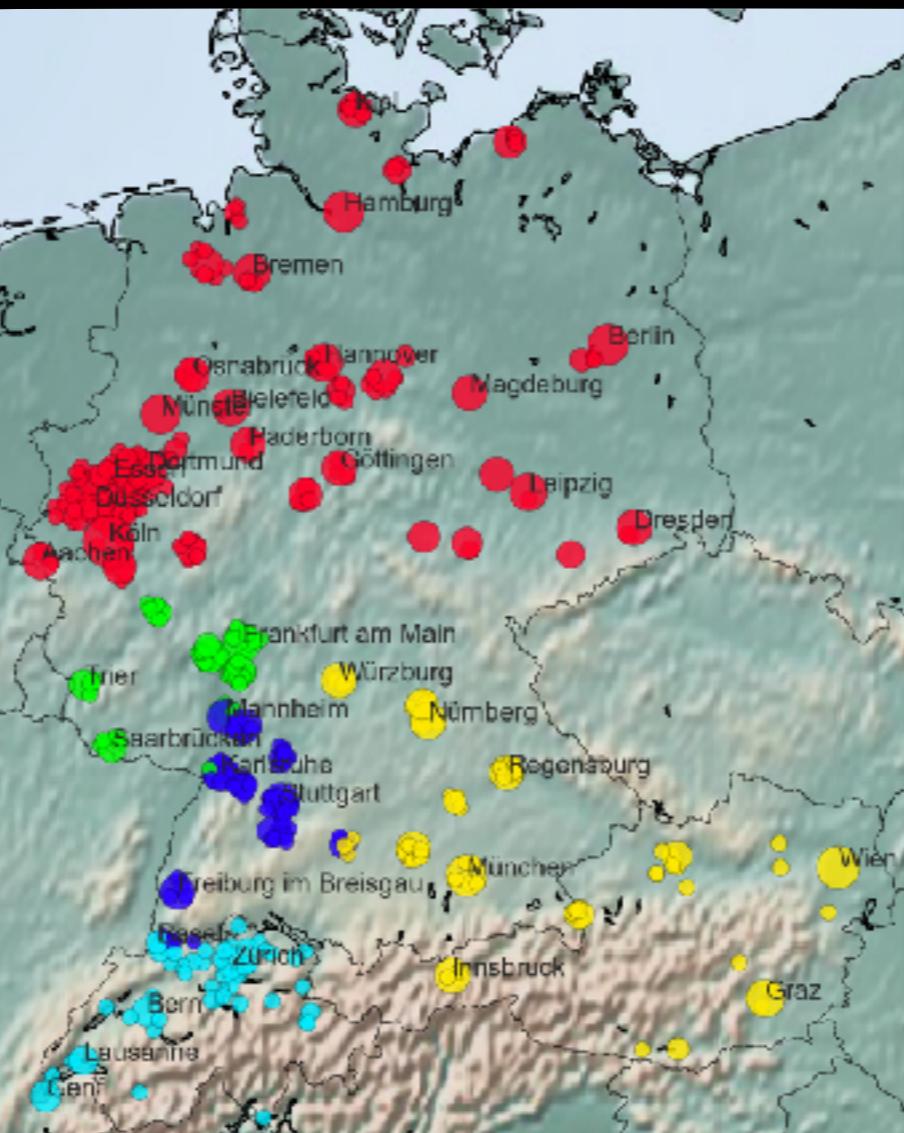


Dialect Clusters

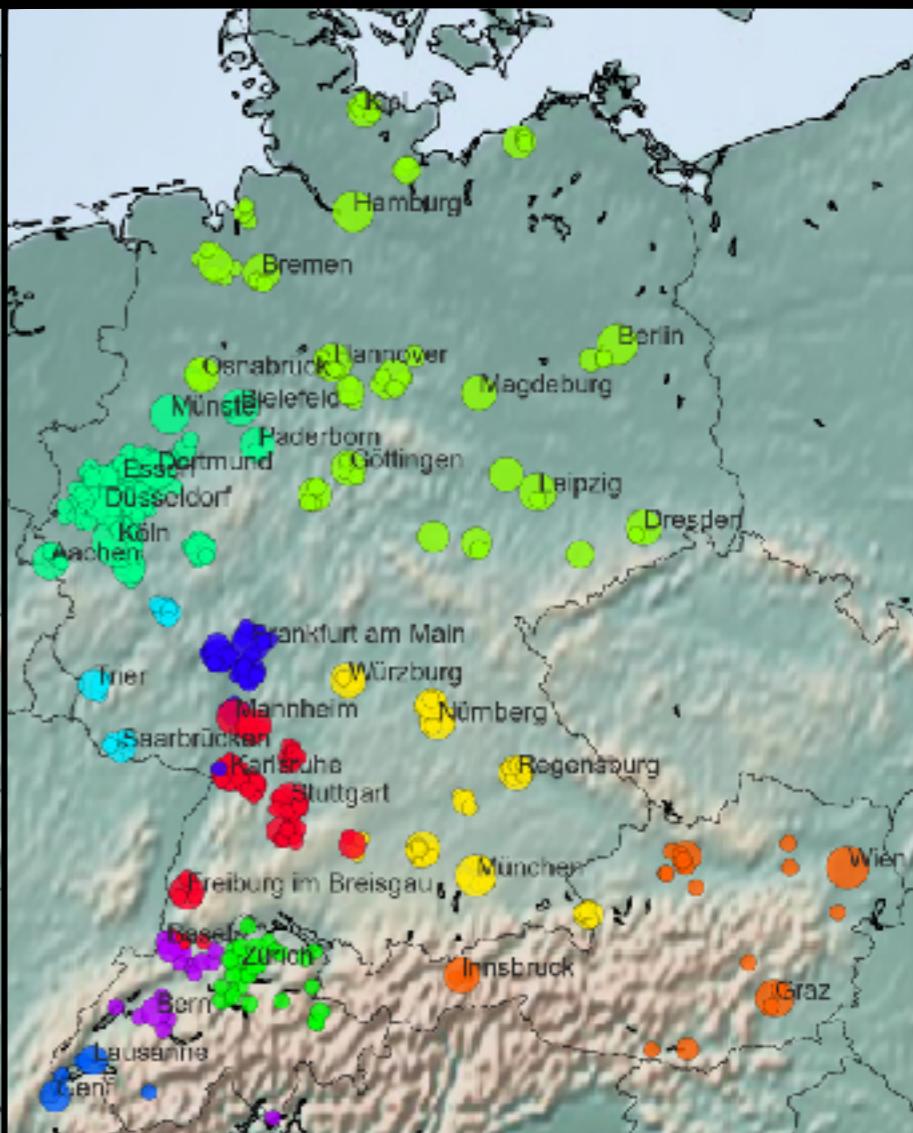
3



5



10



Evaluating Clusters

Making Sense of Clusters

fresh	arsenic	mortgage	blue
perspicacious	induce	platypus	recognize
lacerate	kissed	mortify	pie
president	rotten	pie	

Making Sense of Clusters

A 3D word cloud visualization showing clusters of words. The words are colored by cluster: red, green, blue, and orange. The clusters are roughly spherical and overlap.

president **platypus**
mortgage pie
fresh rotten perspicacious
blue arsenic induce
kissed mortify lacerate

Making Sense of Clusters

NOUNS

president platypus
 pie
mortgage

VERBS

recognize induce
 mortify
 kissed
 lacerate

rotten

perspicacious

fresh blue arsenic

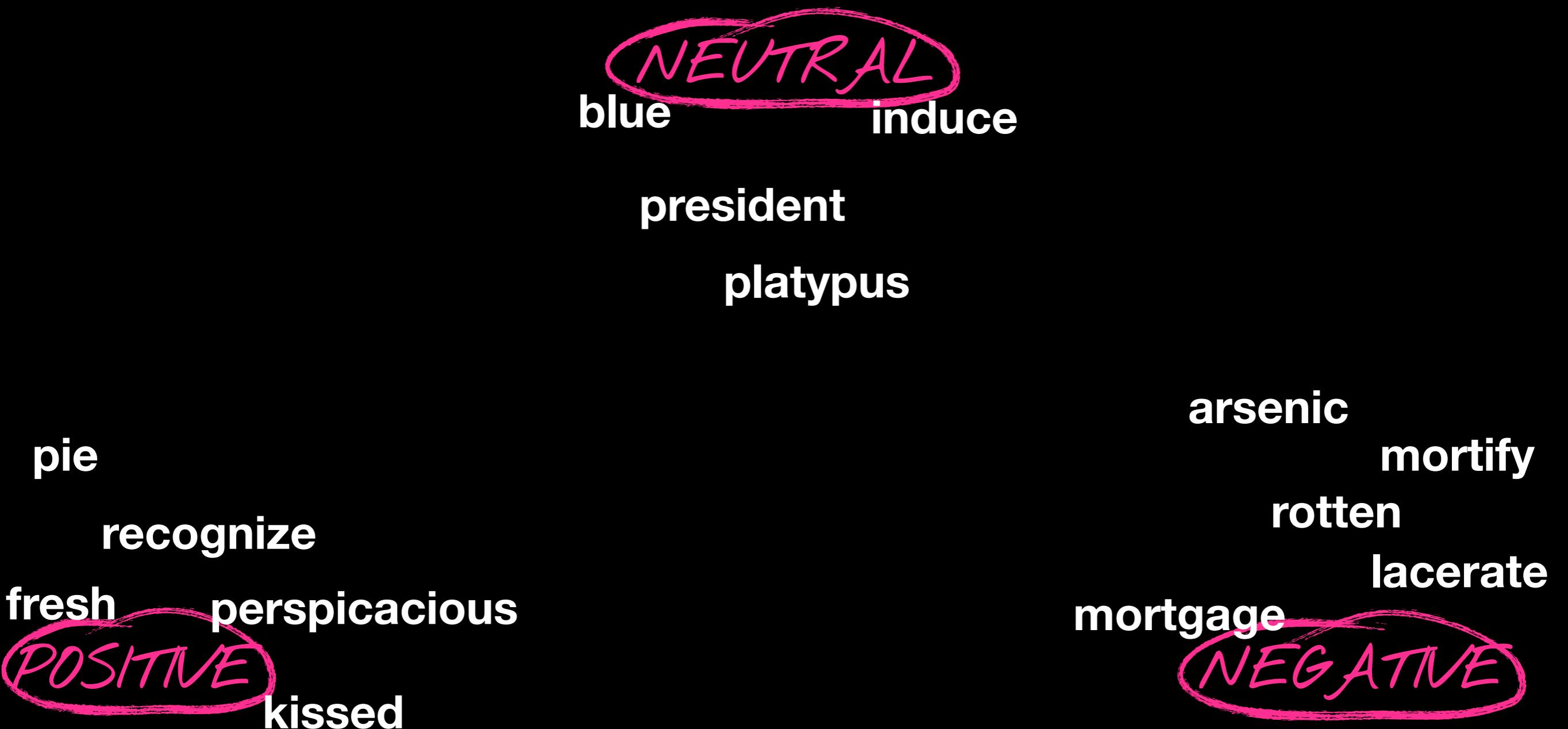
ADJECTIVES

Making Sense of Clusters

blue induce
president
platypus

pie arsenic
recognize mortify
fresh rotten
perspicacious lacerate
 mortgage
kissed

Making Sense of Clusters



Making Sense of Clusters

mortify

induce
president

arsenic

platypus

perspicacious recognize

lacerate

mortgage

blue

kissed

pie

fresh

rotten

Making Sense of Clusters

mortify

induce
president

ROMANCE
ROOT

arsenic
platypus

perspicacious recognize

lacerate
mortgage

blue

kissed

GERMANIC
ROOT

fresh

rotten

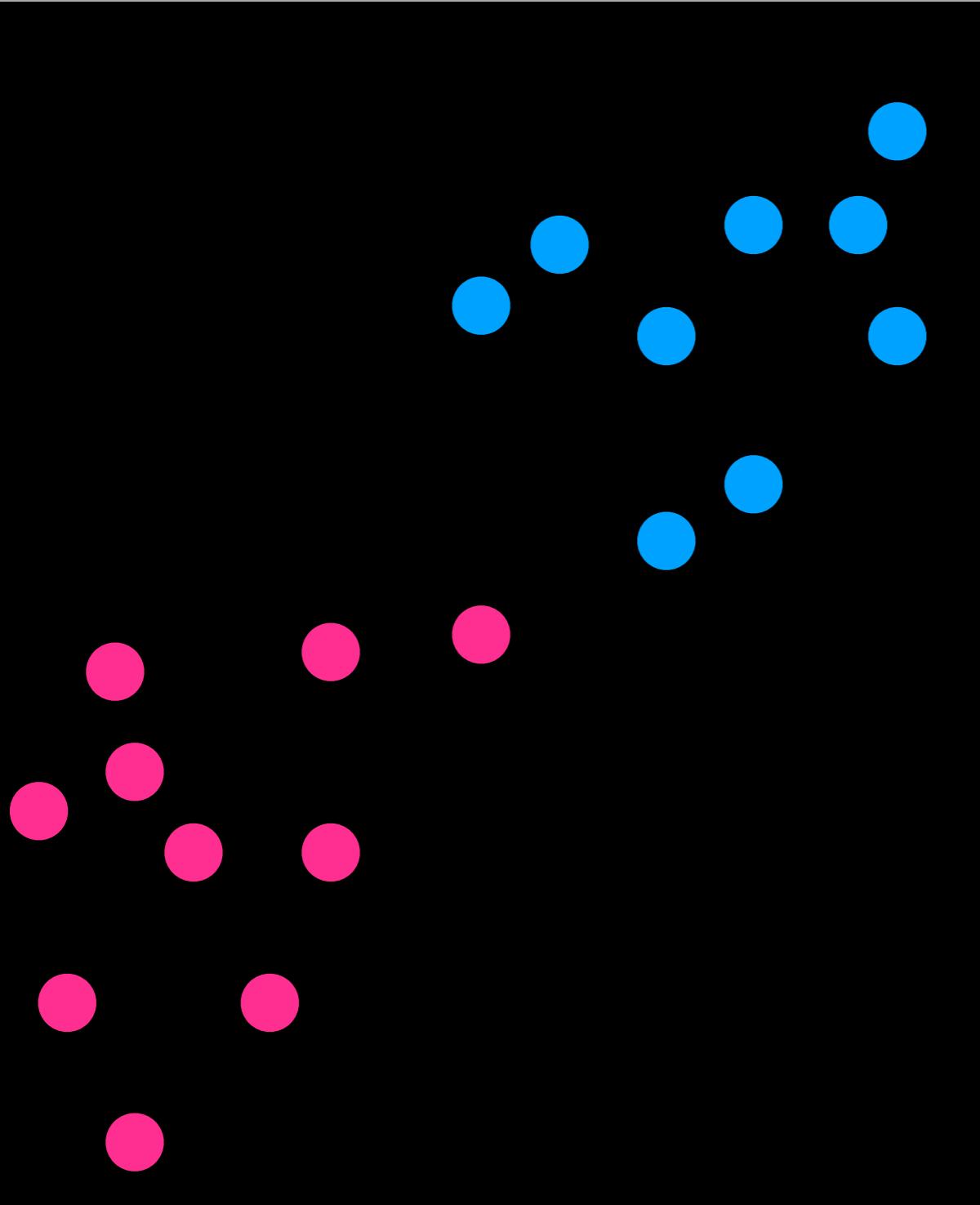
Making Sense of Clusters

perspicacious
lacerate
arsenic
president
mortgage
recognize
platypus
induce
kissed
mortify
rotten

fresh
blue
pie

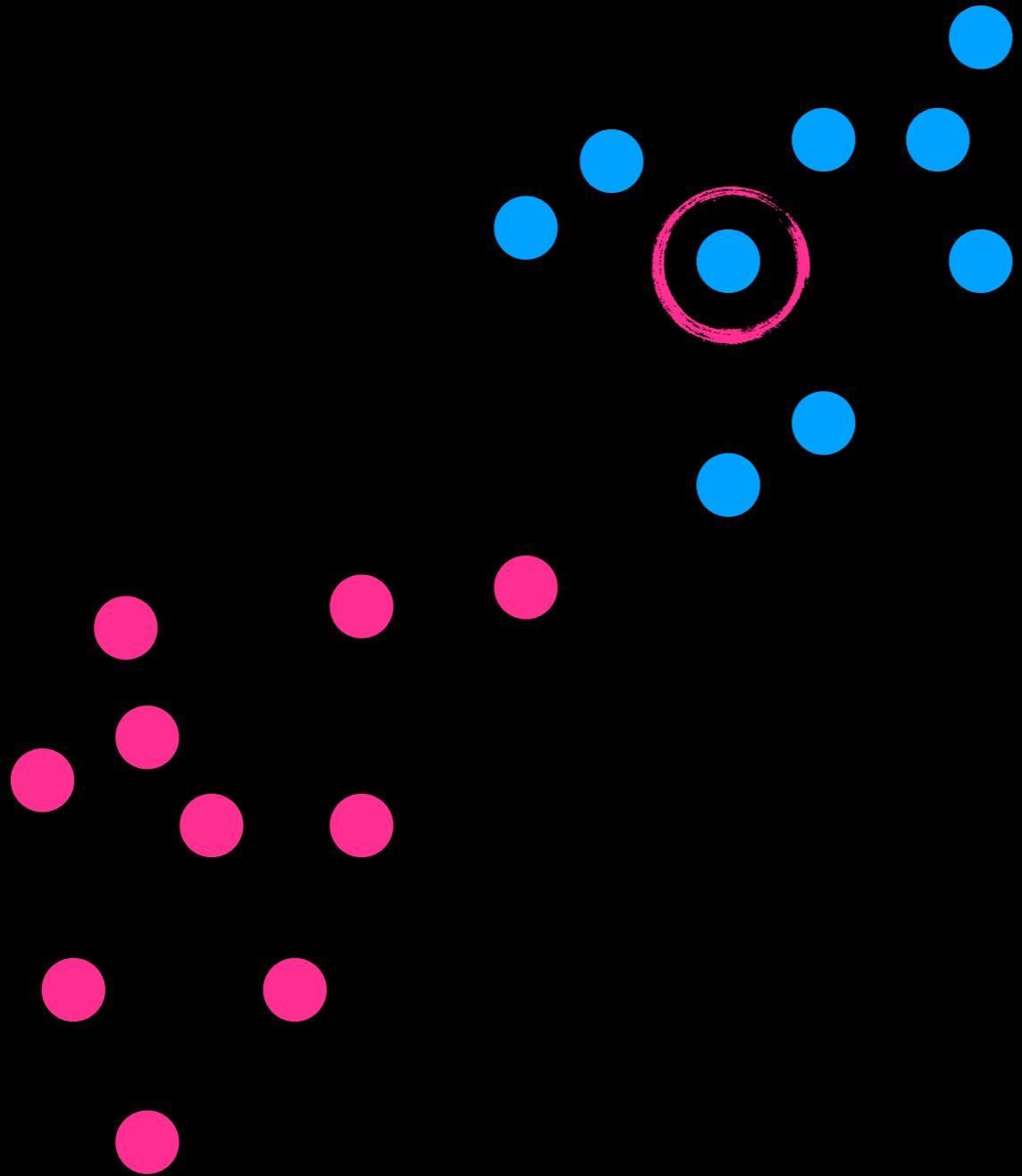
How Many Clusters?

Silhouette Score



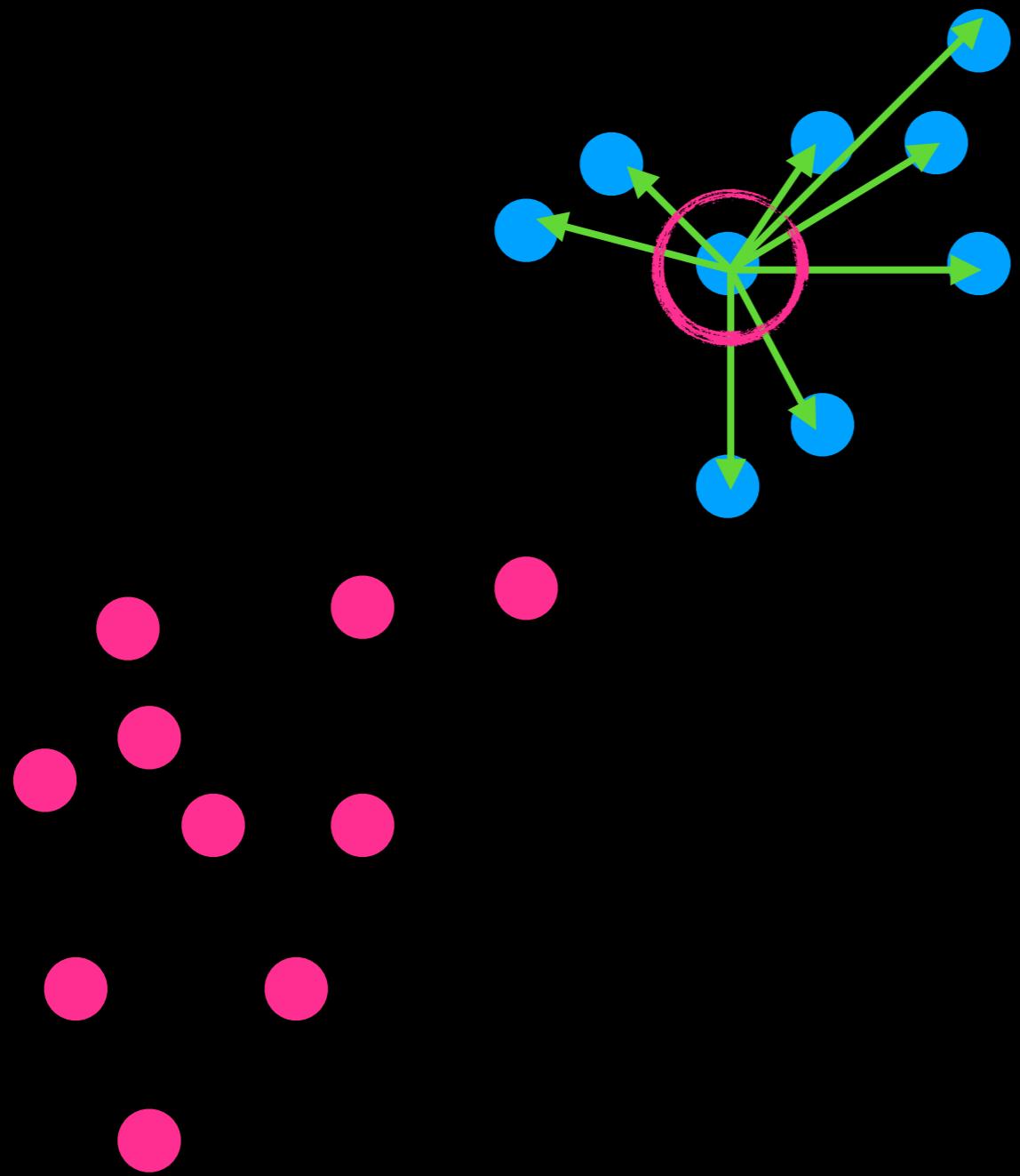
How Many Clusters?

Silhouette Score



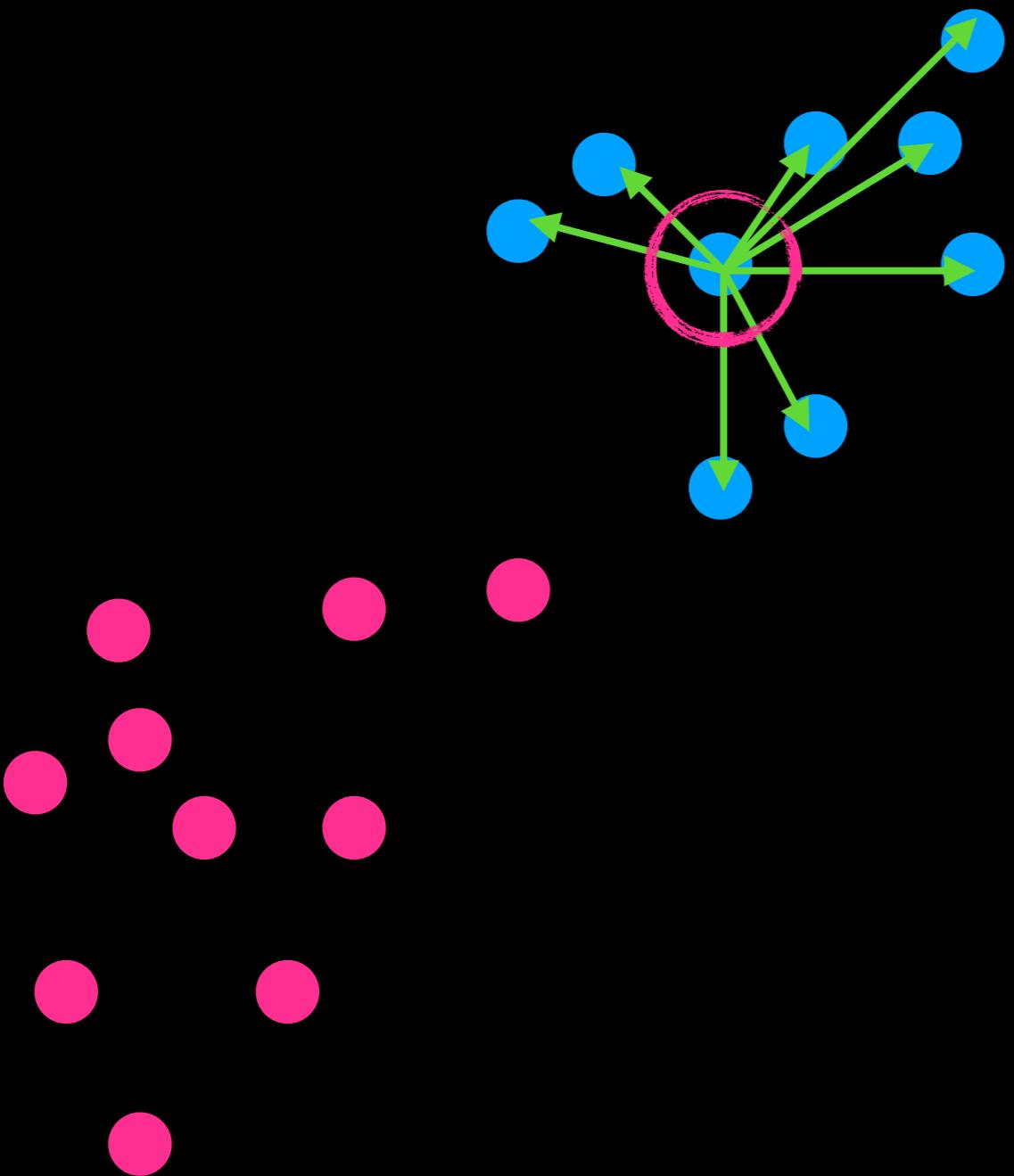
How Many Clusters?

Silhouette Score



How Many Clusters?

Silhouette Score

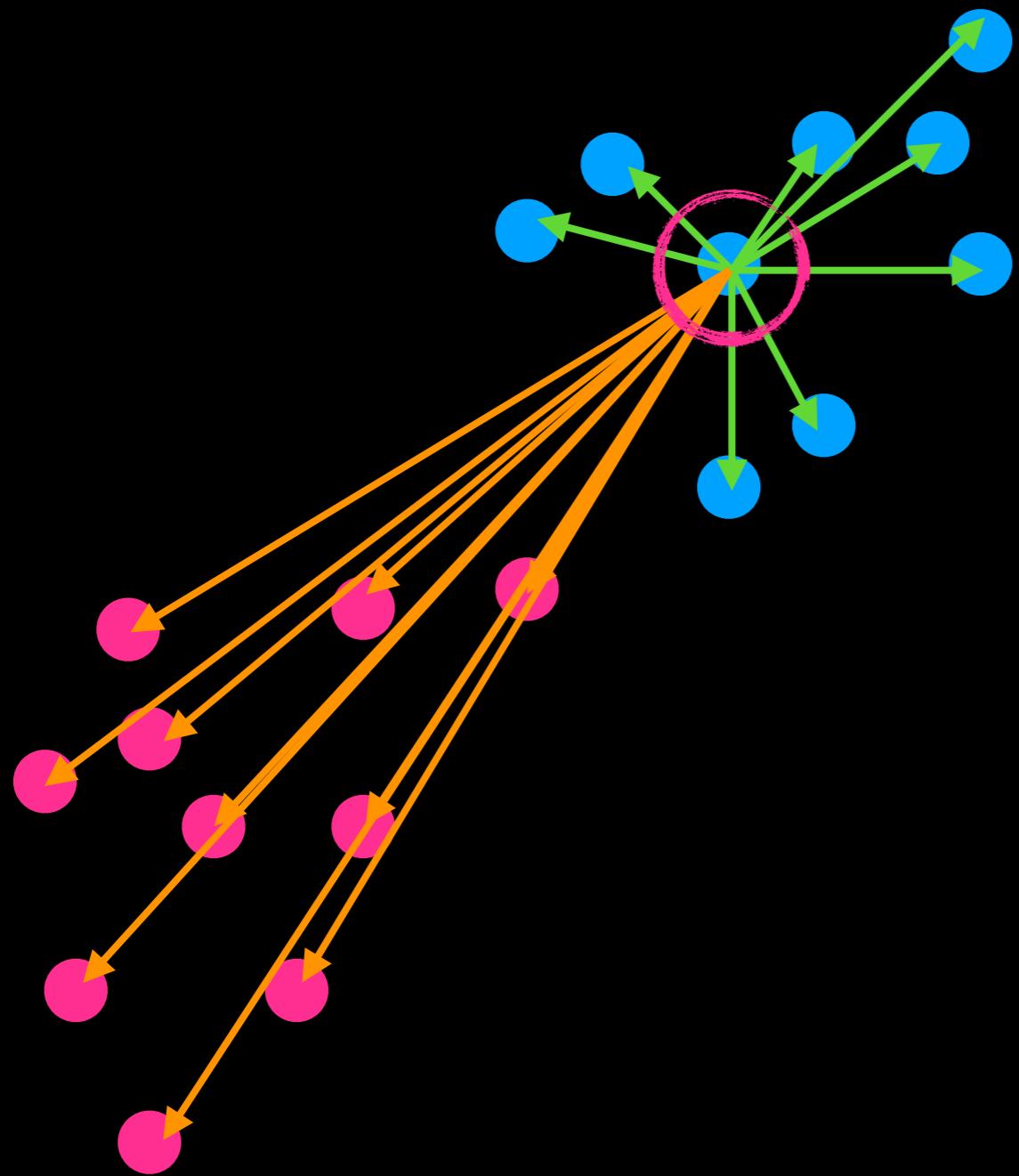


$a = \text{mean intra-cluster distance}$

How Many Clusters?

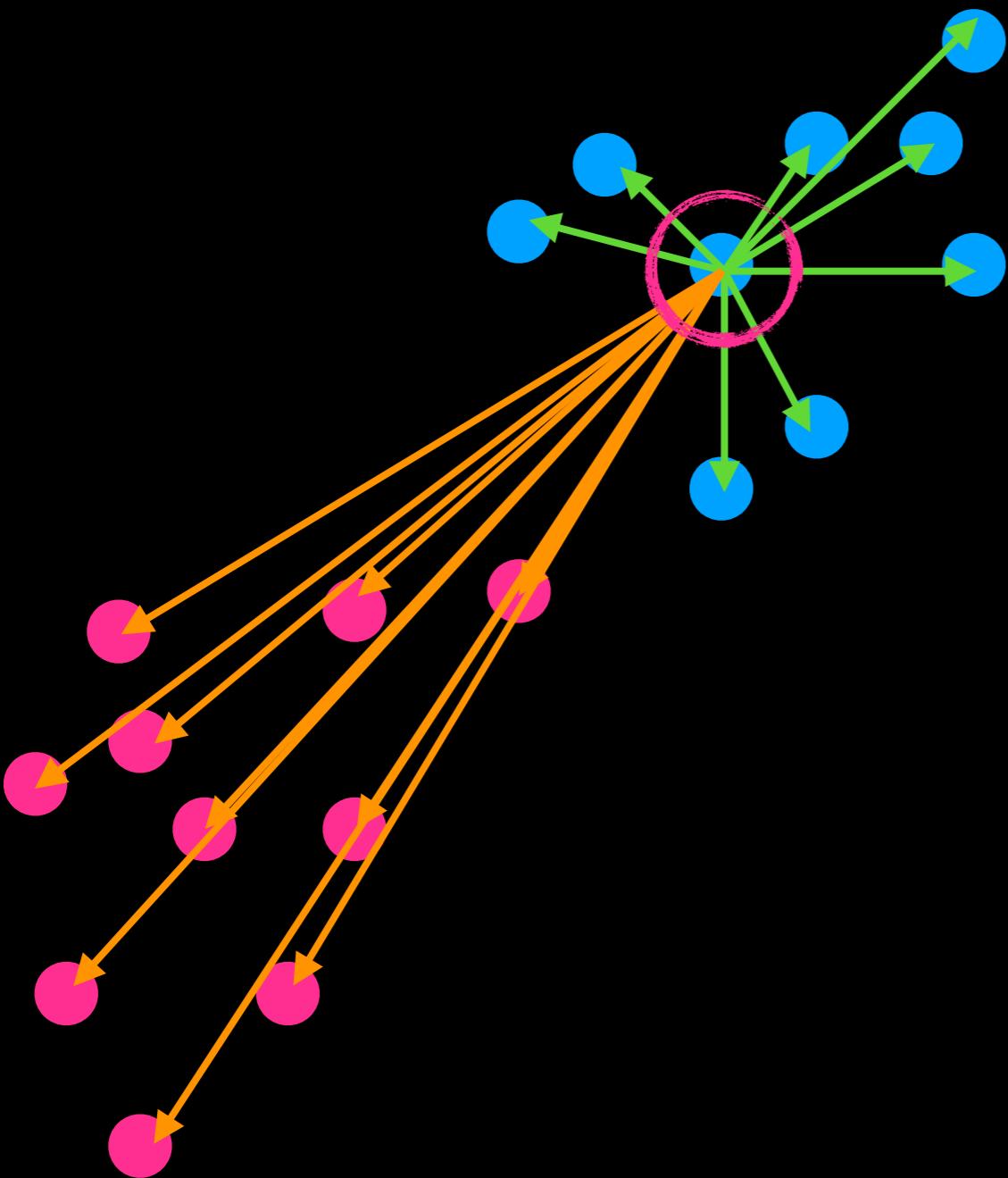
Silhouette Score

$a = \text{mean intra-cluster distance}$



How Many Clusters?

Silhouette Score

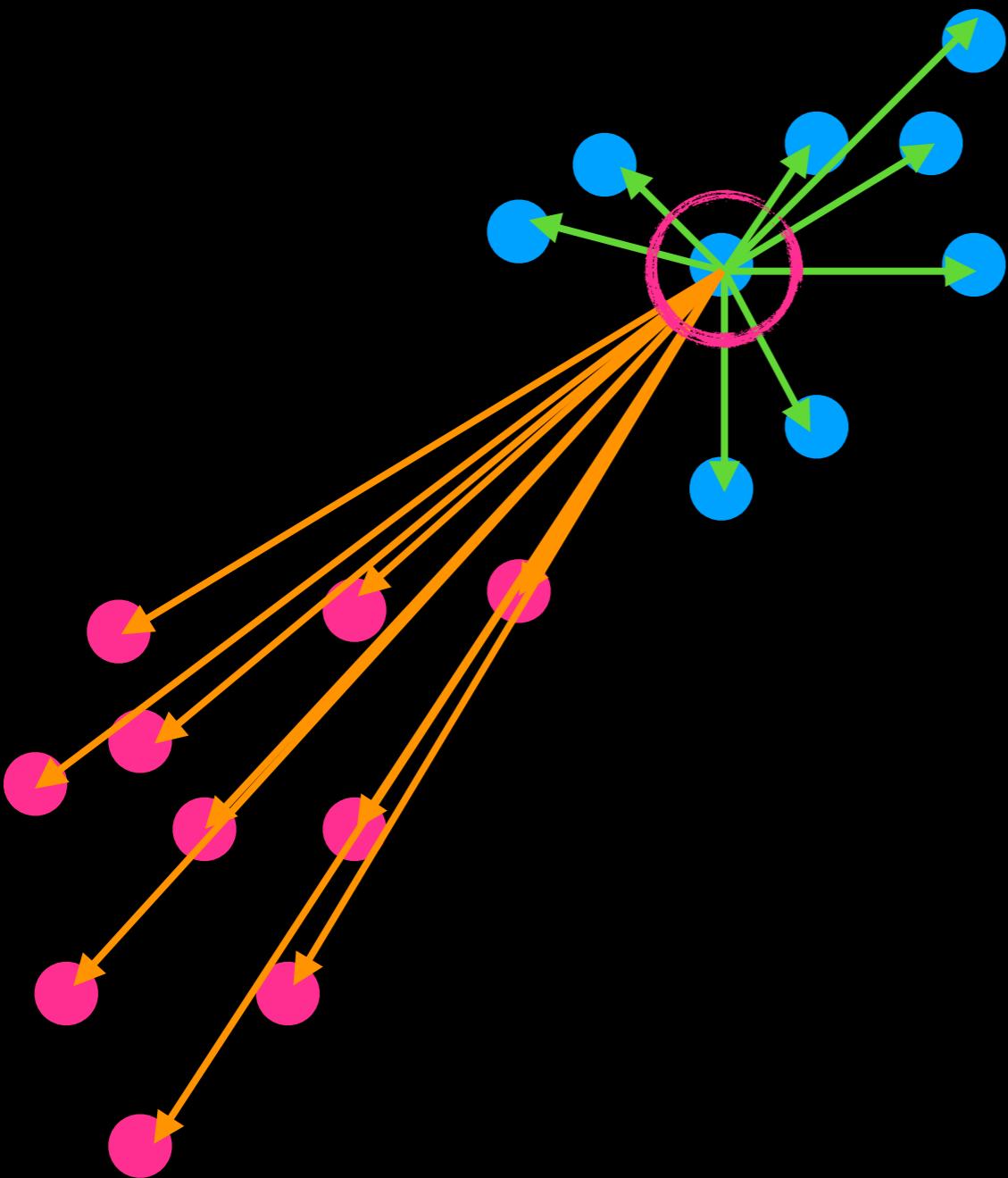


$a = \text{mean intra-cluster distance}$

$b = \text{mean dist. nearest cluster}$

How Many Clusters?

Silhouette Score



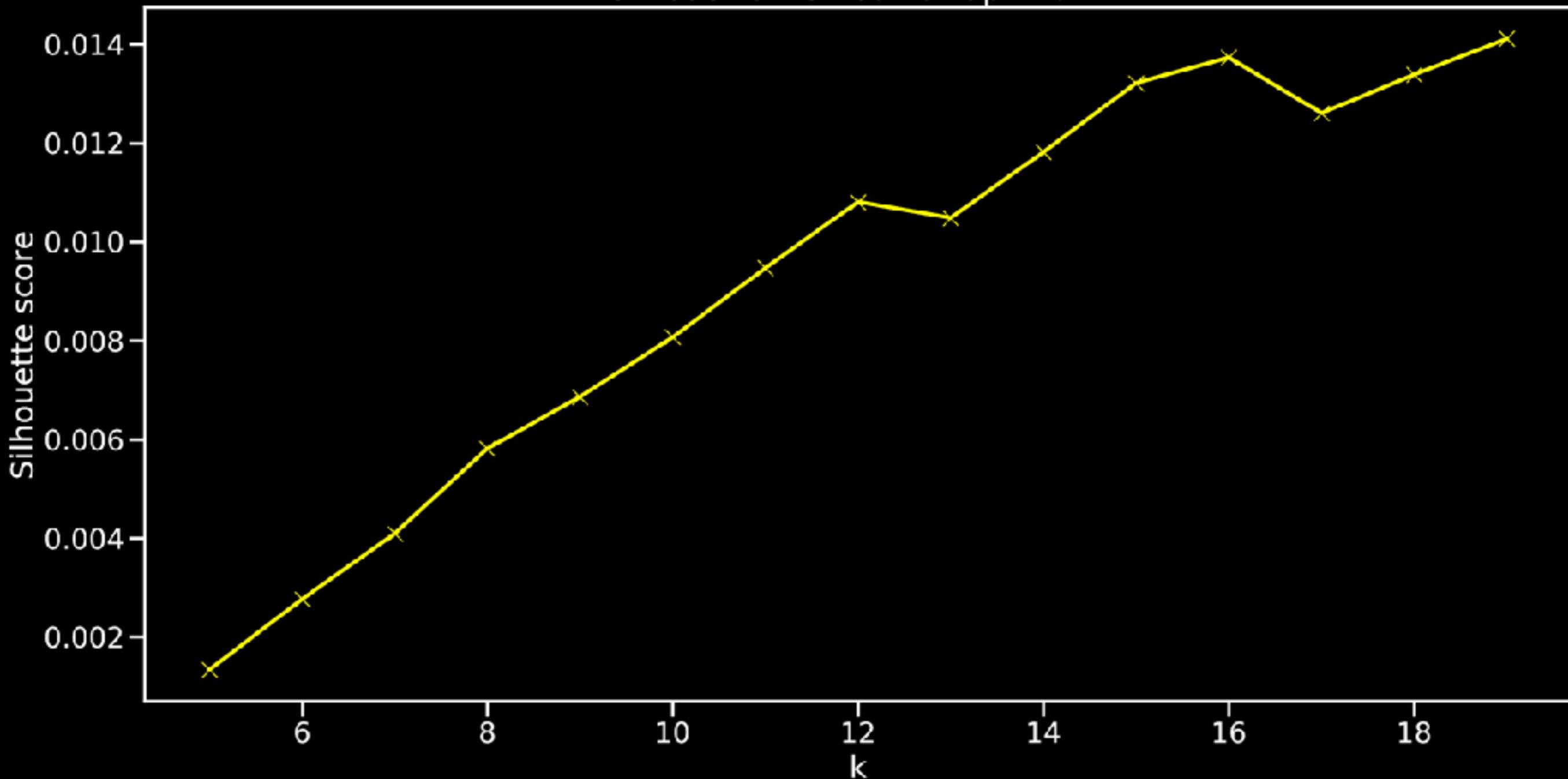
$a = \text{mean intra-cluster distance}$

$$S = \frac{(b - a)}{\max(a, b)}$$

$b = \text{mean dist. nearest cluster}$

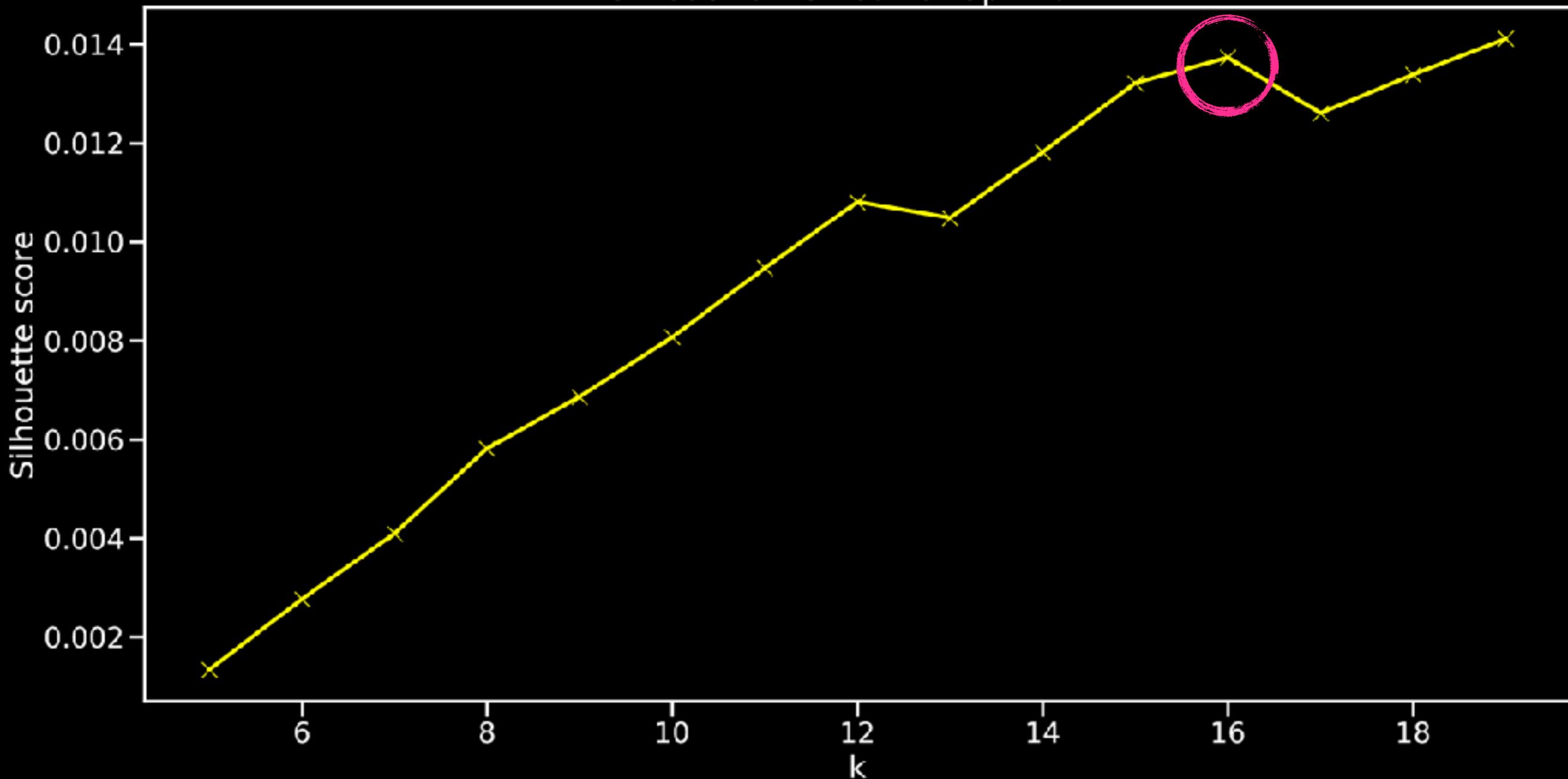
Silhouette Scores

Silhouette Method For Optimal k



Silhouette Scores

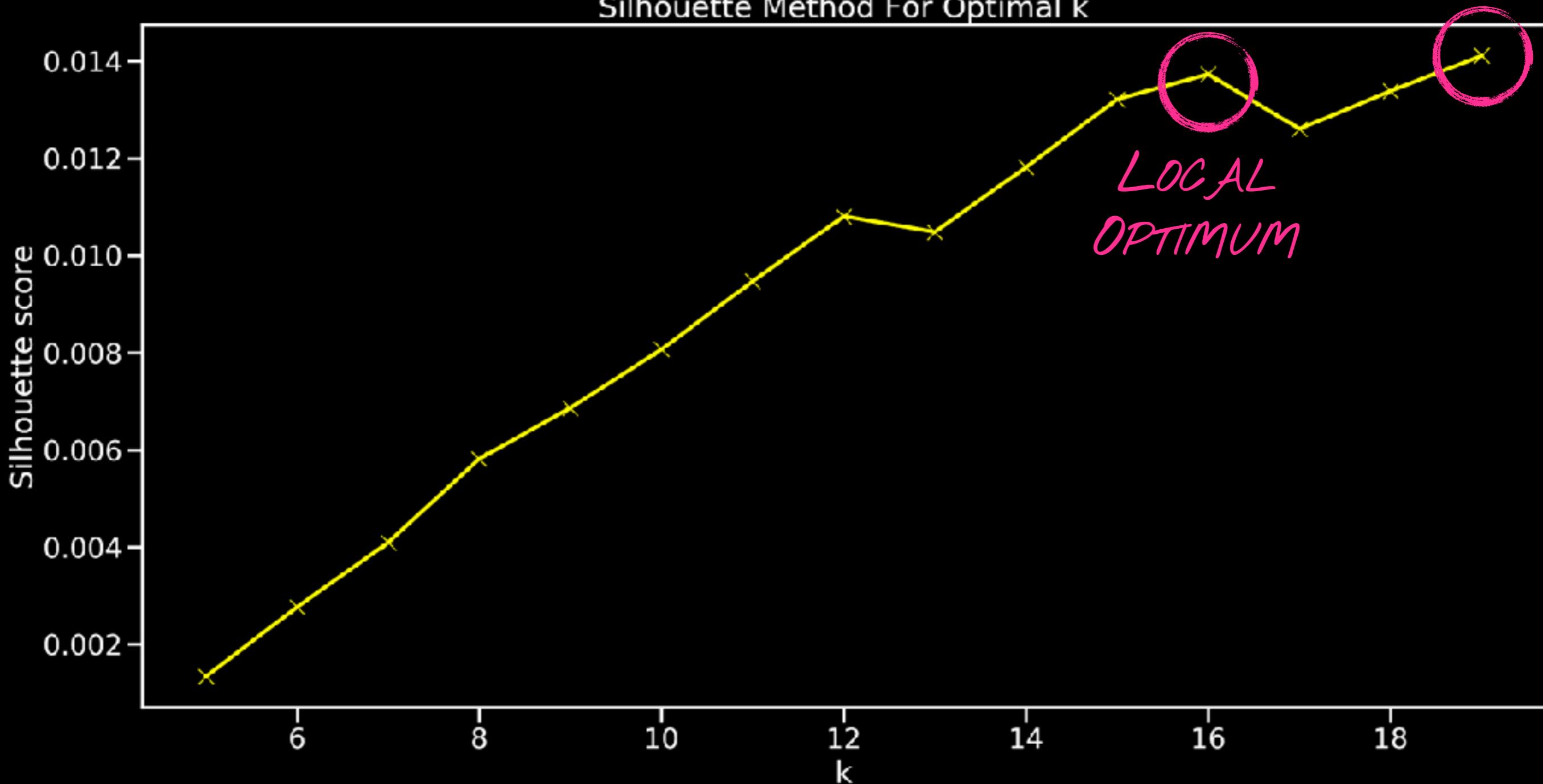
Silhouette Method For Optimal k



Silhouette Scores

DEPENDS ON PATIENCE/COMPUTE POWER

Silhouette Method For Optimal k



Supervised Evaluation Metrics

Homogeneity

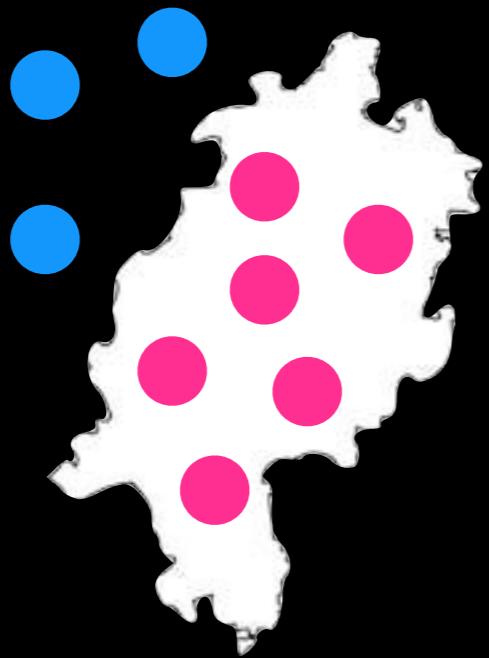
cluster has only 1 gold label

Supervised Evaluation Metrics

Homogeneity

cluster has only 1 gold label

Good:

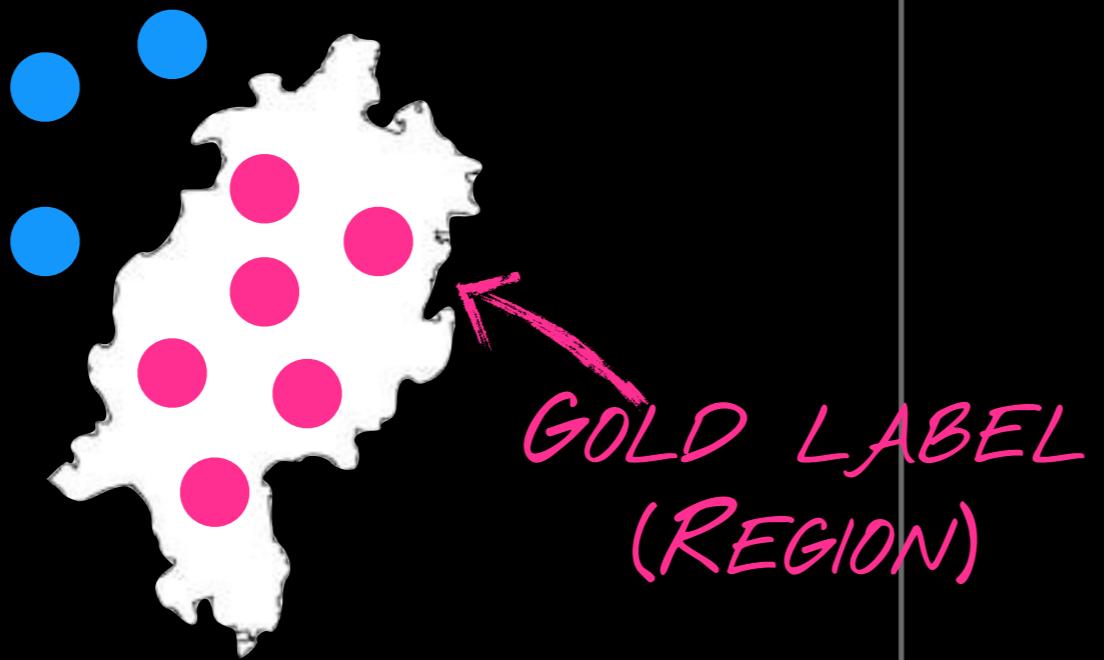


Supervised Evaluation Metrics

Homogeneity

cluster has only 1 gold label

Good:

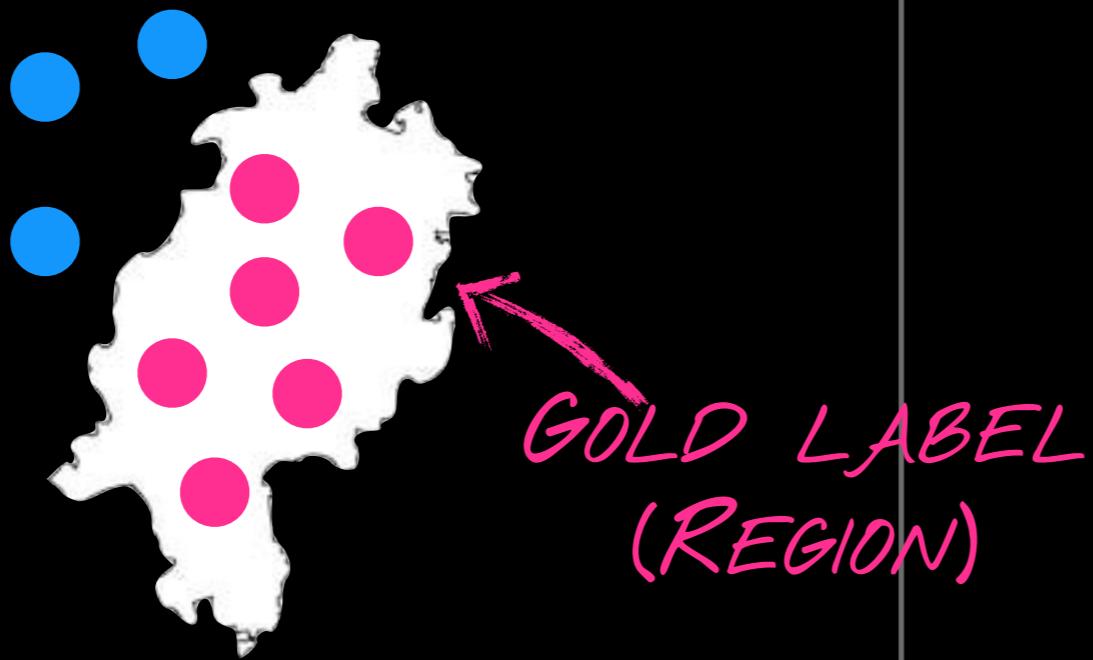


Supervised Evaluation Metrics

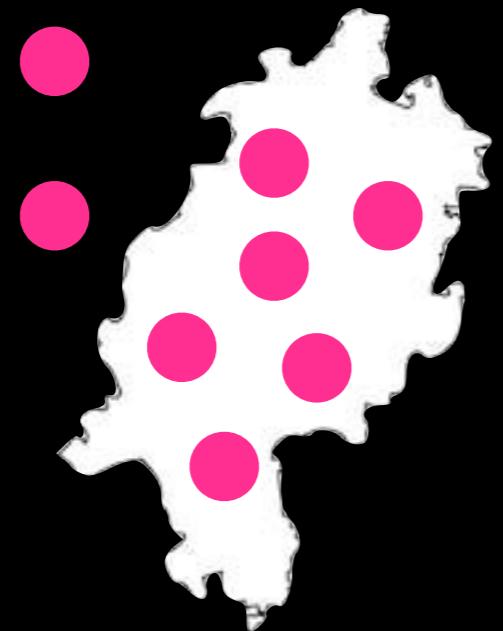
Homogeneity

cluster has only 1 gold label

Good:



Bad:

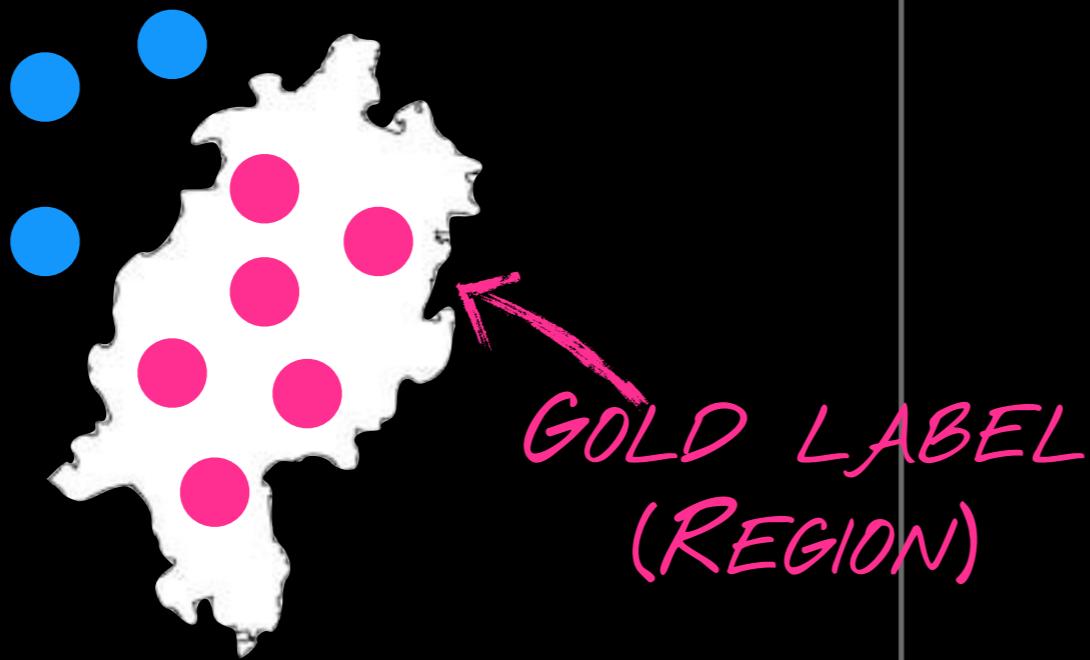


Supervised Evaluation Metrics

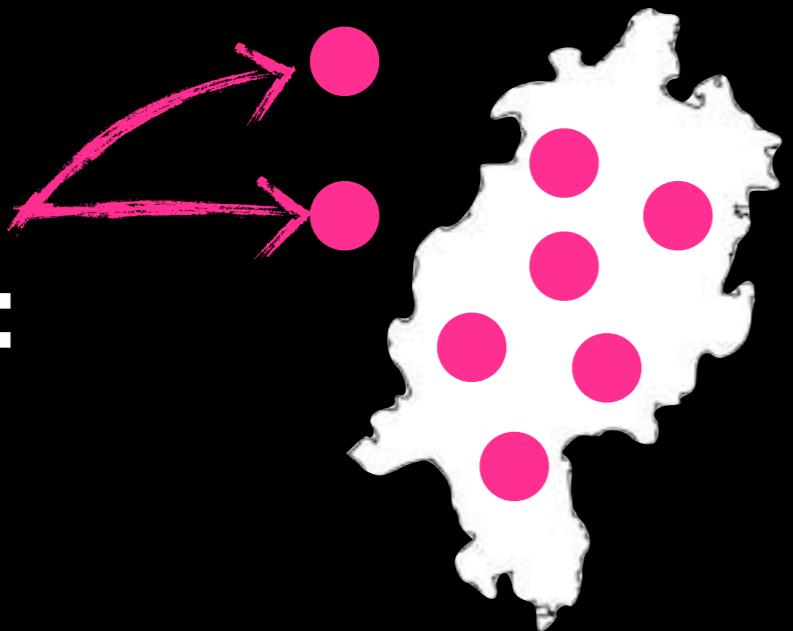
Homogeneity

cluster has only 1 gold label

Good:



Bad:



Supervised Evaluation Metrics

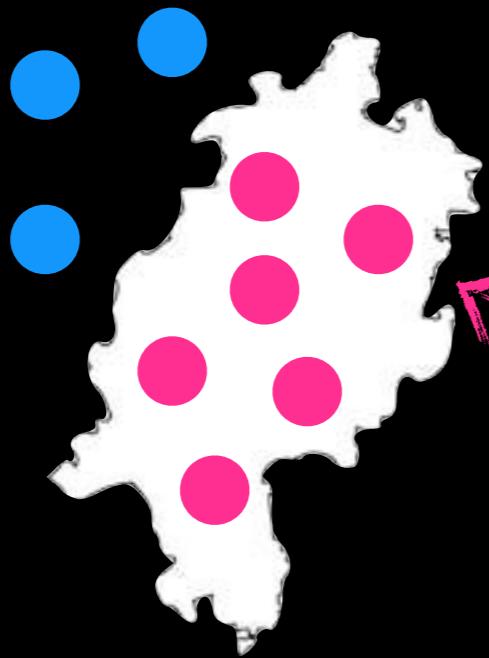
Homogeneity

cluster has only 1 gold label

Completeness

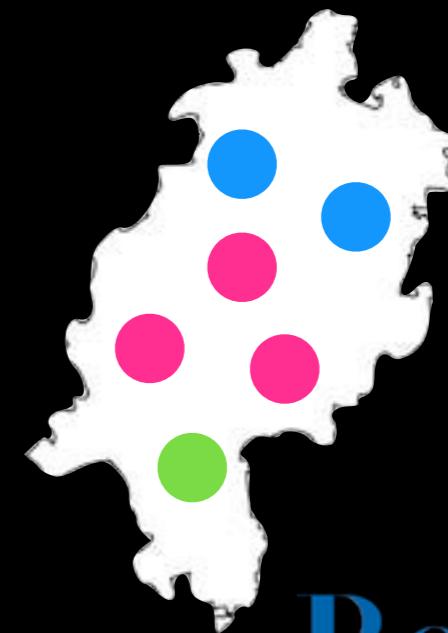
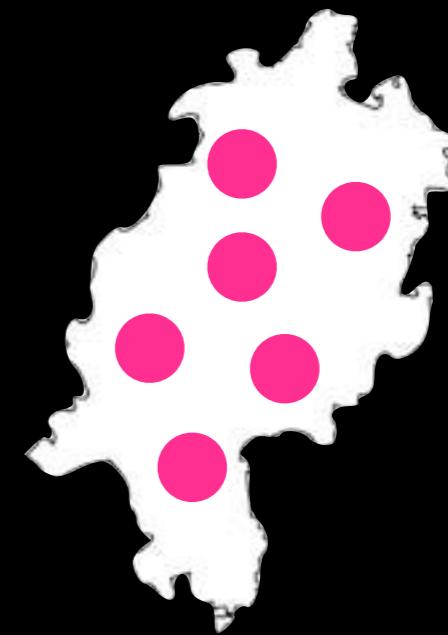
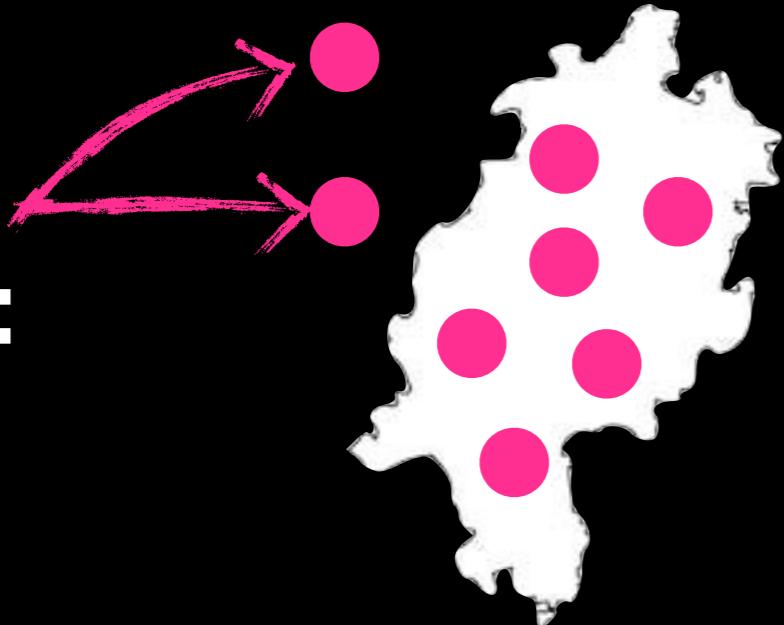
gold label has only 1 cluster

Good:



GOLD LABEL
(REGION)

Bad:



Comparison

	<i>k-means</i>	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	<i>Not applicable</i>
RGB translation	NMF	SVD + scaling
Plotting	SVD	t-SNE
Clustering	<i>Reduce dimensions</i>	<i>Use as-is</i>

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