# INFO 2950: Intro to Data Science

Lecture 23
2023-11-15

# Agenda

1. Singular value decomposition
   a. recommendations
   b. image compression
   c. penguin compression
   d. text compression
2. More text data!

# Singular Value Decomposition

| | |
|---|---|
| 1 | 0 |
| 0 | 1 |

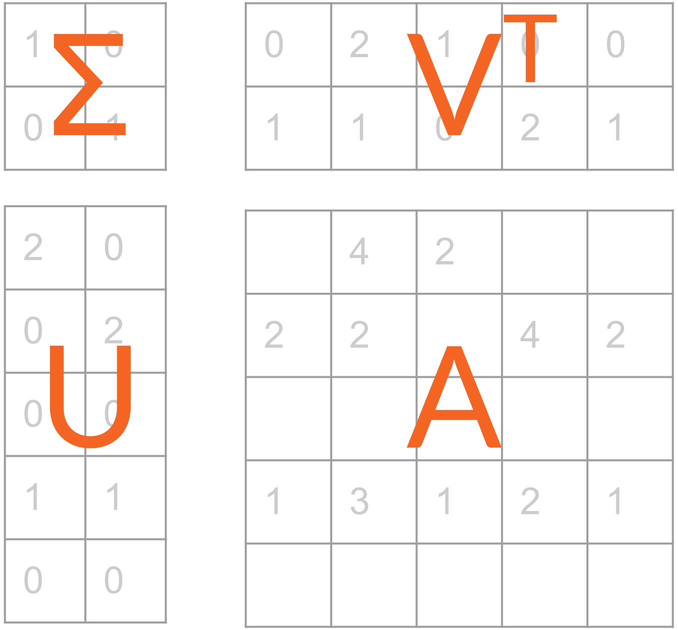
| | | | | |
|---|---|---|---|---|
| 0 | 2 | 1 | 0 | 0 |
| 1 | 1 | 0 | 2 | 1 |

| | |
|---|---|
| 2 | 0 |
| 0 | 2 |
| 0 | 0 |
| 1 | 1 |
| 0 | 0 |

| | | | | |
|---|---|---|---|---|
| | 4 | 2 | | |
| 2 | 2 | | 4 | 2 |
| | | | | |
| 1 | 3 | 1 | 2 | 1 |
| | | | | |

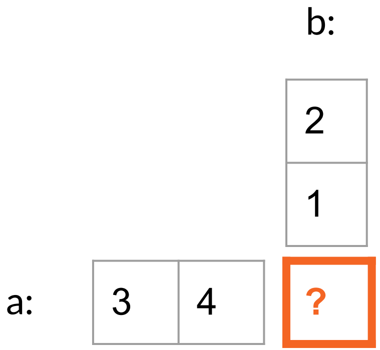
SVD gives us components and weights *from* a matrix

# Singular Value Decomposition




**SVD gives us components and weights *from* a matrix**

# Review: inner product <a,b>



# Review: inner product <a,b>

b:

| |
|---|
| 2 |
| 1 |

a:

| | |
|---|---|
| 3 | 4 |

**3*2 + 4*1 = 10**

# Review: inner product <a,b>

b:

| |
|---|
| 2 |
| 0 |
| 0 |
| 1 |
| 3 |

a:

| | | | | |
|---|---|---|---|---|
| 0 | 2 | 1 | 0 | 1 |

?

# Review: inner product <a,b>

b:

| |
|---|
| 2 |
| 0 |
| 0 |
| 1 |
| **3** |

a:

| | | | | |
|---|---|---|---|---|
| 0 | 2 | 1 | 0 | **1** |

**0+0+0+0+3 = 3**

# Review: inner product <a,b>

b:

| |
|---|
| 7 |
| 4 |

a:

| | |
|---|---|
| -4 | 7 |

$\boxed{?}$

**Are a and b orthogonal?**

# Review: inner product <a,b>

b:

| |
|---|
| 7 |
| 4 |

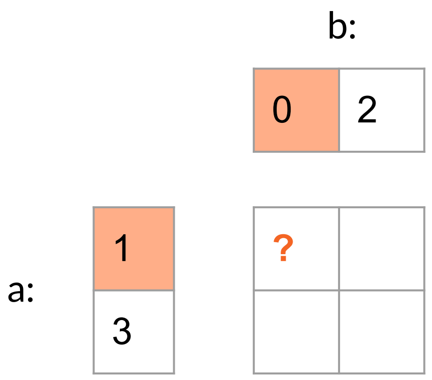
a:

| | |
|---|---|
| -4 | 7 |

**-4*7+7*4 = 0**

**Inner product is 0,**
**a and b are orthogonal**

# Review: outer product a⊗b




Hint: ? is **inner product** of 0, 1

# Review: outer product a⊗b

b:

$$\begin{bmatrix} 0 & 2 \end{bmatrix}$$

a:

$$\begin{bmatrix} 1 \\ 3 \end{bmatrix} \qquad \begin{bmatrix} \mathbf{0} & ? \\ ? & ? \end{bmatrix}$$

Top left: 1*0 = 0

# Review: outer product a⊗b

b:

| | |
|---|---|
| 0 | 2 |

a:

| |
|---|
| 1 |
| 3 |

| | |
|---|---|
| **0** | **2** |
| **0** | **6** |

# Review: outer product a⊗b

b:

| | | |
|---|---|---|
| 0 | 2 | 1 |

a:

| |
|---|
| 2 |
| 0 |
| 3 |

| | | |
|---|---|---|
| | | |
| | | |
| | | |

**Fill in the matrix**

**What is nnz?**

# Review: outer product a⊗b

b:

| | | |
|---|---|---|
| 0 | 2 | 1 |

a:

| |
|---|
| 2 |
| 0 |
| 3 |

| | | |
|---|---|---|
| 0 | 4 | 2 |
| 0 | 0 | 0 |
| 0 | 6 | 3 |

Number of non-zeros (nnz) = 4

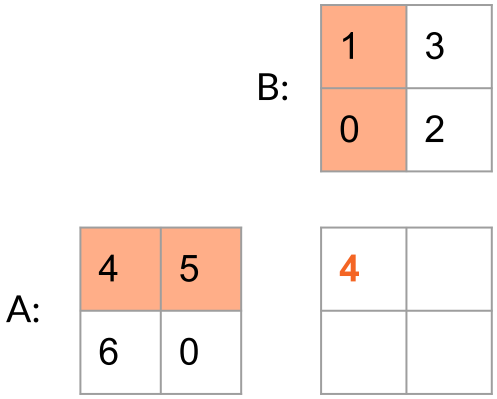
# Review: product AB

B: $\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}$

A: $\begin{bmatrix} 4 & 5 \\ 6 & 0 \end{bmatrix}$ $\begin{bmatrix} ? & \; \\ \; & \; \end{bmatrix}$

Hint: ? is **inner product** of orange inputs

# Review: product AB




**Inner product** of orange inputs is $4*1 + 5*0 = 4$

# Review: product AB

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

A:

| | |
|---|---|
| 4 | 5 |
| 6 | 0 |

| | |
|---|---|
| **4** | |
| ? | |

Fill in the ?

# Review: product AB

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

$? = 6*1+0*0 = 6$

A:

| | |
|---|---|
| 4 | 5 |
| 6 | 0 |

| | |
|---|---|
| **4** | |
| **6** | |

# Review: product AB

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

A:

| | |
|---|---|
| 4 | 5 |
| 6 | 0 |

| | |
|---|---|
| **4** | ? |
| **6** | ? |

# Review: product AB

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

A:

| | |
|---|---|
| 4 | 5 |
| 6 | 0 |

| | |
|---|---|
| **4** | **22** |
| **6** | **18** |

4*3+5*2 = 22

6*3+0*2 = 18

# Does AB = BA?

B: $\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}$

A: $\begin{bmatrix} 4 & 5 \\ 6 & 0 \end{bmatrix}$ $\begin{bmatrix} \mathbf{4} & \mathbf{22} \\ \mathbf{6} & \mathbf{18} \end{bmatrix}$

A: $\begin{bmatrix} 4 & 5 \\ 6 & 0 \end{bmatrix}$

B: $\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}$ $\begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix}$

# Does AB = BA?

B: $\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}$

A: $\begin{bmatrix} 4 & 5 \\ 6 & 0 \end{bmatrix}$ $\begin{bmatrix} \mathbf{4} & \mathbf{22} \\ \mathbf{6} & \mathbf{18} \end{bmatrix}$

A: $\begin{bmatrix} 4 & 5 \\ 6 & 0 \end{bmatrix}$

B: $\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}$ $\begin{bmatrix} \mathbf{22} & \mathbf{5} \\ \mathbf{12} & \mathbf{0} \end{bmatrix}$

**No, order matters!**

# Can we do this product AB?

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

A:

| | | |
|---|---|---|
| 4 | 5 | 1 |
| 6 | 0 | 2 |

| | |
|---|---|
| **?** | **?** |
| **?** | **?** |

# Can we do this product AB?

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

A:

| | | |
|---|---|---|
| 4 | 5 | 1 |
| 6 | 0 | 2 |

| | |
|---|---|
| ? | ? |
| ? | ? |

**No, you can’t do an inner product of a 3-length vector with a 2-length vector!**

# Can we do this product AB?

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

A:

| | |
|---|---|
| 4 | 5 |
| 6 | 0 |
| 1 | 2 |

| | |
|---|---|
| | |
| | |
| | |

# Can we do this product AB?

B:

| | |
|---|---|
| 1 | 3 |
| 0 | 2 |

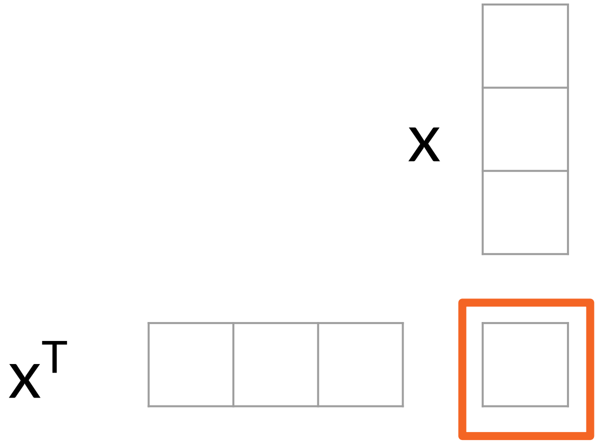
A:

| | |
|---|---|
| 4 | 5 |
| 6 | 0 |
| 1 | 2 |

| | |
|---|---|
| **4** | **22** |
| **6** | **18** |
| **1** | **7** |

**Yes! We can multiply (3x2) with (2x2) since the 2-dimension is shared**

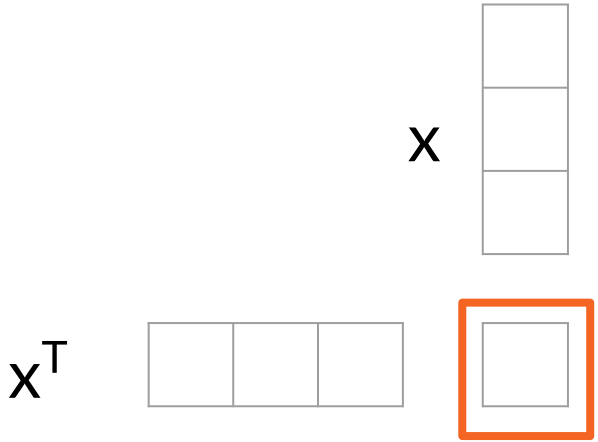
# Review: What is this value?




Assume the vector x has mean=0

What does $x^T x/3$ calculate?

# Review: What is this value?

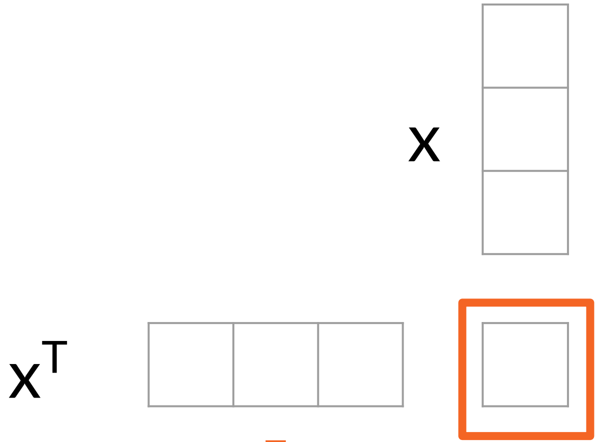



Assume the vector x has mean=0

What does $x^Tx/3$ calculate?

Variance! $$\frac{\Sigma(x_i-\mu)^2}{n}$$

# Review: What is this value?



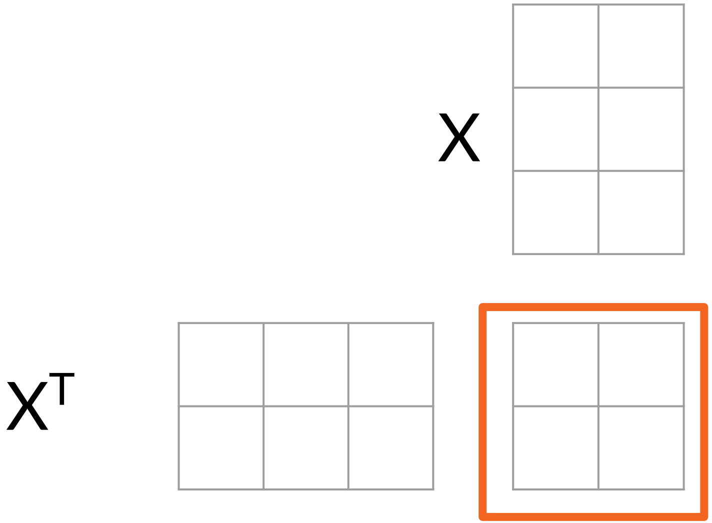

Assume the vector x has mean=0

What does $x^T x/3$ calculate?

Variance! $\frac{\Sigma(x_i-\mu)^2}{n}$

Remember, $x^T x$ gives you the sum of squared entries. We're given that μ=0, and len(x) = 3.

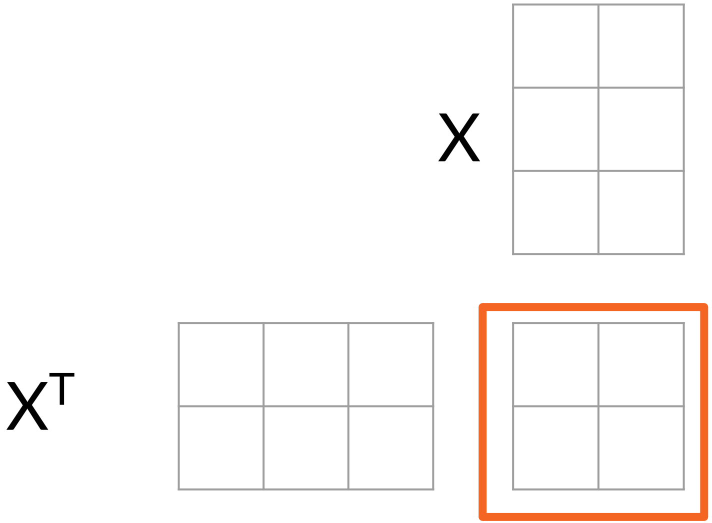
# Review: What is this matrix?




Assume the columns of X represent variables with mean=0

What does $X^TX/3$ calculate?

# Review: What is this matrix?




Assume the columns of X represent variables with mean=0

What does $X^TX/3$ calculate?
Covariance matrix!

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

# Covariance matrix

$X$

| Score | Age |
|---|---|
| 68 | 29 |
| 60 | 26 |
| 58 | 30 |
| 40 | 35 |

$X^T X / 4 =$

| | |
|---|---|
| Var(Score) = 104.75 | cov(Score, Age) = -27 |
| cov(Age, Score) = -27 | Var(Age) = 10.5 |

$X^T$

| | | | |
|---|---|---|---|
| 68 | 60 | 58 | 40 |
| 29 | 26 | 30 | 35 |

https://www.cuemath.com/algebra/covariance-matrix/

# Covariance matrix

$$\frac{\Sigma(x_i-\mu)^2}{n}$$

$\mu_{Score}$ = 56.5, n = 4
$[(68-56.5)^2 + (60-56.5)^2 + (58-56.5)^2 + (40-56.5)^2] / 4 =$
104.75

X

| Score | Age |
|---|---|
| 68 | 29 |
| 60 | 26 |
| 58 | 30 |
| 40 | 35 |

$X^T$

| | | | |
|---|---|---|---|
| 68 | 60 | 58 | 40 |
| 29 | 26 | 30 | 35 |

| | |
|---|---|
| Var(Score) = 104.75 | cov(Score, Age) = -27 |
| cov(Age, Score) = -27 | Var(Age) = 10.5 |

https://www.cuemath.com/algebra/covariance-matrix/

# Covariance matrix

$$\frac{\Sigma(x_i-\mu)^2}{n}$$

$\mu_{Age}$ =30, n = 4
$[(29-30)^2 + (26-30)^2 + (30-30)^2 + (35-30)^2]/4 = 10.5$

X

| Score | Age |
|---|---|
| 68 | 29 |
| 60 | 26 |
| 58 | 30 |
| 40 | 35 |

$X^T$

| | | | |
|---|---|---|---|
| 68 | 60 | 58 | 40 |
| 29 | 26 | 30 | 35 |

| | |
|---|---|
| Var(Score) = 104.75 | cov(Score, Age) = -27 |
| cov(Age, Score) = -27 | Var(Age) = 10.5 |

https://www.cuemath.com/algebra/covariance-matrix/

# Covariance matrix

$$\text{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

X

| Score | Age |
|---|---|
| 68 | 29 |
| 60 | 26 |
| 58 | 30 |
| 40 | 35 |

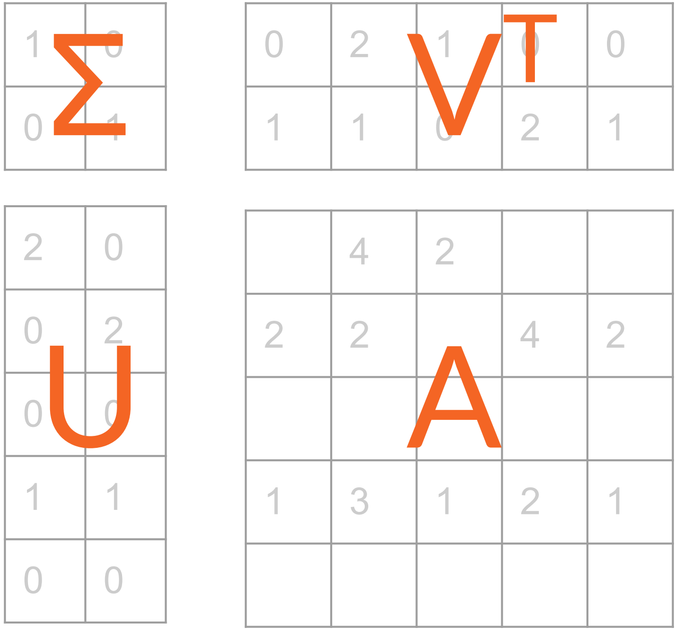
$X^T$

| | | | |
|---|---|---|---|
| 68 | 60 | 58 | 40 |
| 29 | 26 | 30 | 35 |

| | |
|---|---|
| Var(Score) = 104.75 | cov(Score, Age) = -27 |
| cov(Age, Score) = -27 | Var(Age) = 10.5 |

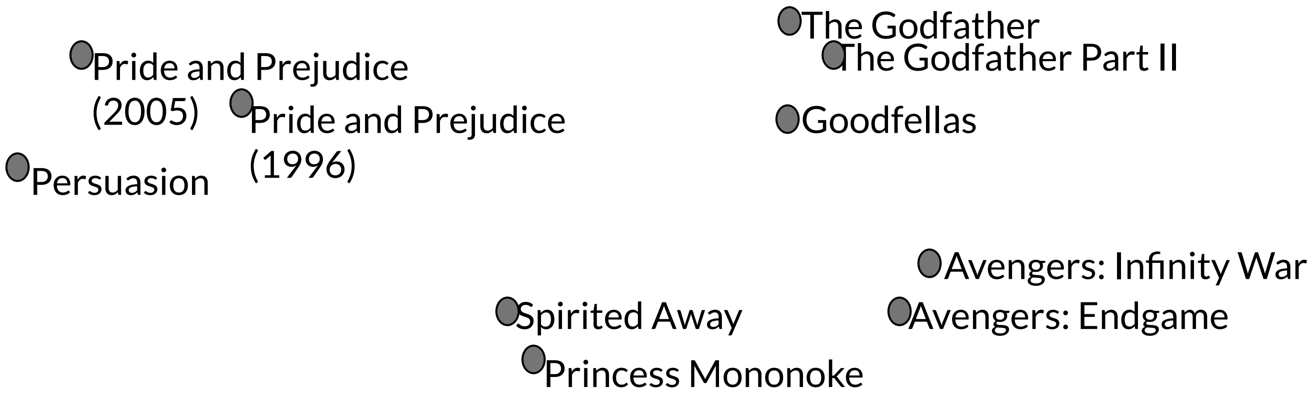
https://www.cuemath.com/algebra/covariance-matrix/

# Singular Value Decomposition




SVD gives us components and weights *from* a matrix

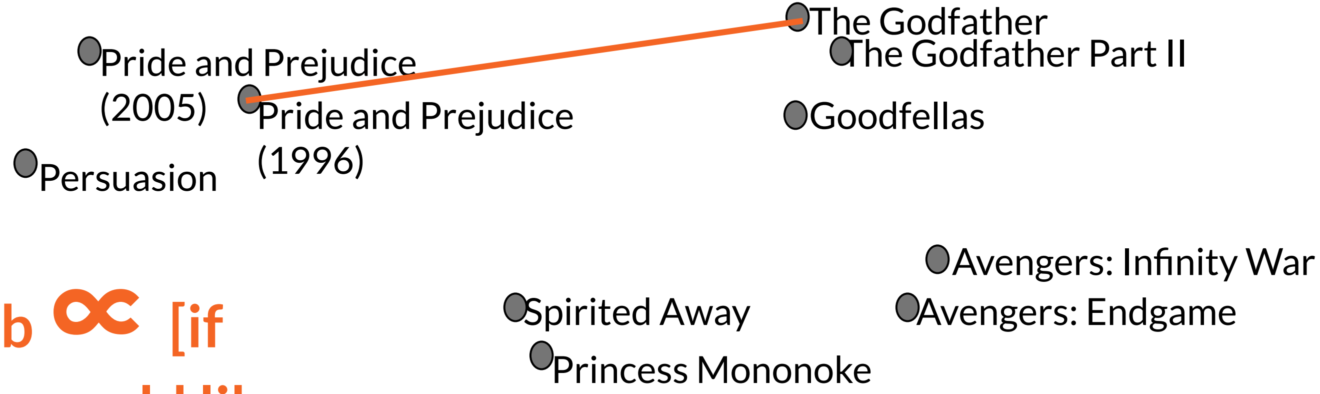
# Goal: find a smaller representation that preserves similarity



# Goal: find a smaller representation that preserves similarity

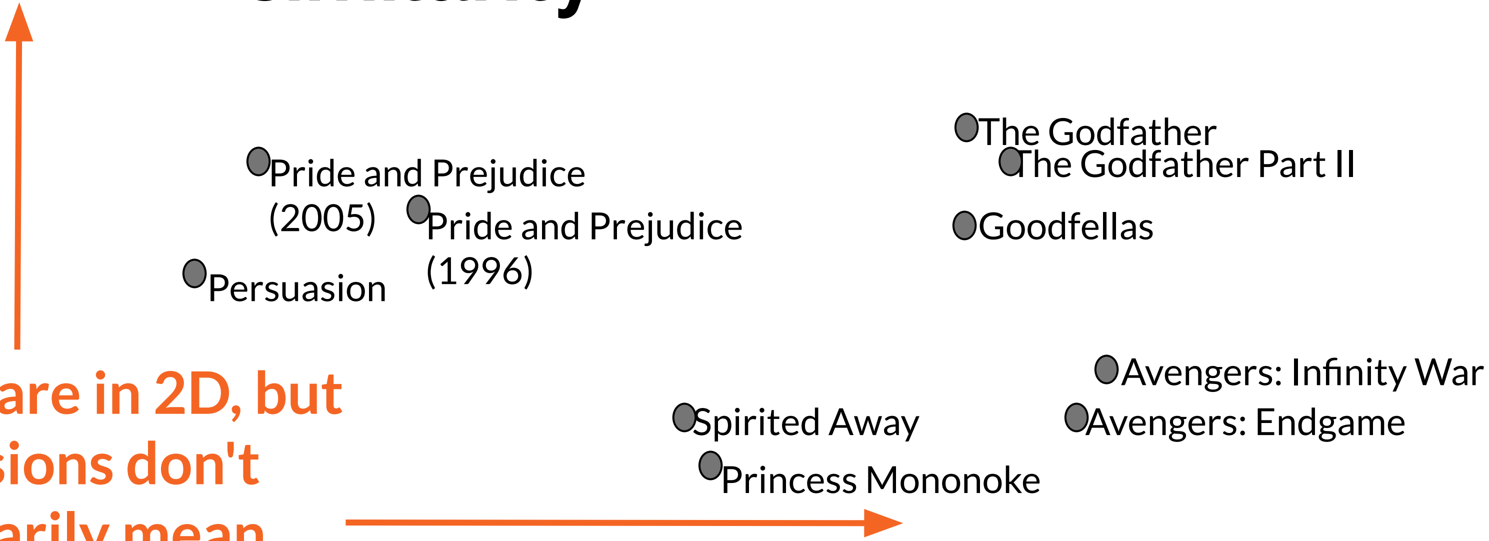

Distance a to b $\propto$ [if you like a, you would like b]

# Goal: find a smaller representation that preserves similarity




Distance a to b $\propto$ [if you like a, you would like b]

# Goal: find a smaller representation that preserves similarity




Points are in 2D, but dimensions don't necessarily mean anything

# Summarize user patterns

| | User 1 | User 2 | User 3 | User 4 | ... | User 13435 |
|---|---|---|---|---|---|---|
| Airplane! | 9 | 6 | | 7 | | |
| Akira | | 4 | 7 | 8 | | 8 |
| Aladdin | 6 | | | 7 | | |
| Alexander Nevsky | | | | 6 | | |
| ... | | | | | | |
| Zoolander | | | 9 | 5 | | 7 |

# Summarize user patterns

| | Axis 1 | Axis 2 | … |
|---|---|---|---|
| Airplane! | 1.3 | 3.1 | |
| Akira | -2.6 | 4.2 | |
| Aladdin | -2.3 | 3.3 | |
| Alexander Nevsky | 1.8 | -1.6 | |
| … | | | |
| Zoolander | -0.02 | -1.8 | |

# Make recommendations

| | Axis 1 | Axis 2 | ... |
|---|---|---|---|
| Airplane! | 1.3 | 3.1 | |
| Akira | -2.6 | 4.2 | |
| Aladdin | -2.3 | 3.3 | |
| Alexander Nevsky | 1.8 | -1.6 | |
| ... | | | |
| Zoolander | -0.02 | -1.8 | |

# How do we do this?

- By using the SVD!
  - SVD = singular value decomposition
- *"The SVD is like a matrix X-ray"*
  - Daniela Witten

https://www.imaginghealthcare.com/diagnostic-imaging/digital-x-ray/

https://twitter.com/WomenInStat/status/1285610321747611653

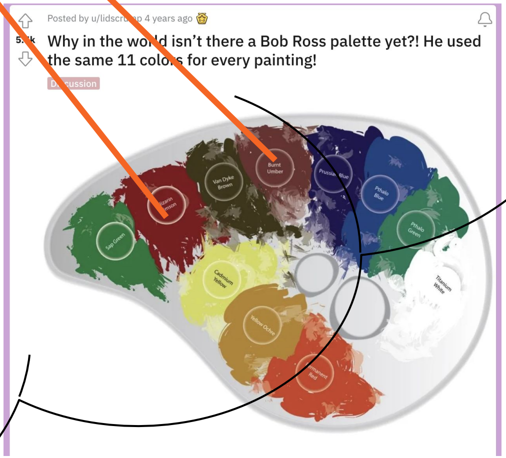
# Parts of a matrix factorization

| | |
|---|---|
| 2 | 0 |
| 0 | 1 |

| | | | | |
|---|---|---|---|---|
| 0 | 2 | 1 | 0 | 0 |
| 1 | 1 | 0 | 2 | 1 |

| | |
|---|---|
| 2 | 0 |
| 0 | 2 |
| 0 | 0 |
| 1 | 1 |
| 0 | 0 |

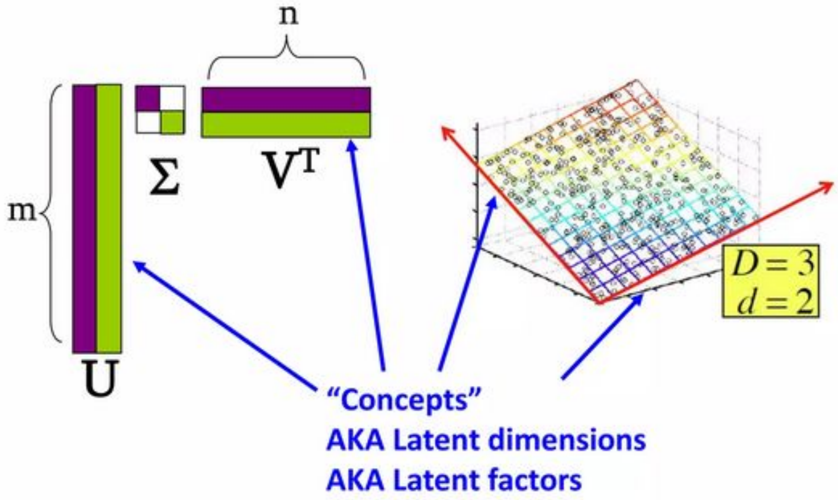
Components are like colors, component weights are how much of each color you are mixing



# SVD: movie “concepts”

- $A = U\,\Sigma\,V^T$ - example: Users to Movies

$$
\begin{array}{c}
\begin{matrix} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{matrix} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
=
$$



# SVD: movie "concepts"

$$A = U \, \Sigma \, V^T$$

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

# SVD: movie “concepts”

$$A = U \Sigma V^T$$

**U = User-to-concept similarity matrix**

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
=
\begin{array}{cc}
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
&
\begin{array}{l}
\text{User 1} \\ \text{User 2} \\ \text{User 3} \\ \text{User 4} \\ \text{User 5} \\ \text{User 6} \\ \text{User 7}
\end{array}
\end{array}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

# SVD: movie "concepts"

$A = U \, \Sigma \, V^T$

U = User-to-concept similarity matrix

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
\end{array}
=
\begin{array}{c}
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\end{array}
\begin{array}{l}
\text{User 1} \\ \text{User 2} \\ \text{User 3} \\ \text{User 4} \\ \text{User 5} \\ \text{User 6} \\ \text{User 7}
\end{array}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

# SVD: movie "concepts"

$$A = U \Sigma V^T$$

V = Movie-to-concept similarity matrix

$$\begin{array}{c} \begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \end{array} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \times \begin{array}{c} \begin{array}{ccccc} \text{Movie 1} & \text{Movie 2} & \text{Movie 3} & \text{Movie 4} & \text{Movie 5} \end{array} \\ \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix} \end{array}$$

# SVD: movie “concepts”

$$A = U \Sigma V^T$$

**Σ = Concept matrix**

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\
\end{array}
$$

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

Concept 1

Concept 2

Concept 3

# SVD: movie "concepts"

$$A = U \Sigma V^T$$

What rank is matrix A?

$$\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\ \end{array}$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

Concept 1

Concept 2

Concept 3

# SVD: movie “concepts”

$A = U \Sigma V^T$

What rank is matrix A?

Rank 3 (there are 3 concepts)

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

Concept 1

Concept 2

Concept 3

# SVD: movie “concepts”

$A = U \Sigma V^T$

SciFi-concept

(smaller magnitude → less effect)

Users that tend to watch sci-fi movies

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
\boxed{\begin{matrix} 1 & 1 & 1 \\ 3 & 3 & 3 \\ 4 & 4 & 4 \\ 5 & 5 & 5 \end{matrix}} & \begin{matrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{matrix} \\
\begin{matrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{matrix} & \begin{matrix} 4 & 4 \\ 5 & 5 \\ 2 & 2 \end{matrix}
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\boxed{\begin{matrix} 0.13 \\ 0.41 \\ 0.55 \\ 0.68 \end{matrix}} & \begin{matrix} 0.02 & -0.01 \\ 0.07 & -0.03 \\ 0.09 & -0.04 \\ 0.11 & -0.05 \end{matrix} \\
\begin{matrix} 0.15 \\ 0.07 \\ 0.07 \end{matrix} & \begin{matrix} \mathbf{-0.59} & \mathbf{0.65} \\ \mathbf{-0.73} & \mathbf{-0.67} \\ \mathbf{-0.29} & \mathbf{0.32} \end{matrix}
\end{bmatrix}
\times
\begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 & 1.3
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

# SVD: movie "concepts"

$$A = U \Sigma V^T$$

Users that tend to watch romance movies

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
=
\begin{array}{c}
\begin{array}{cc} \text{SciFi-concept} & \text{Romance-concept} \end{array} \\
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\end{array}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

# SVD: movie “concepts”

$$A = U \Sigma V^T$$

SciFi-concept

Romance-concept

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

Some 3rd concept, unclear what – applies positively to Users 5 and 7 but negatively to User 6

# SVD: movie "concepts"

$A = U \Sigma V^T$

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\left[\begin{array}{ccccc}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{array}\right]
\end{array}
=
\left[\begin{array}{ccc}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{array}\right]
\times
\left[\begin{array}{ccc}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{array}\right]
\times
\left[\begin{array}{ccccc}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{array}\right]
$$

SciFi-concept for movies

Movies that seem to be SciFi

# SVD: movie "concepts"

$$A = U \Sigma V^T$$

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\
\end{array}
$$

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \boxed{\mathbf{-0.69}} & \boxed{\mathbf{-0.69}} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

Romance-concept for movies

Movies that seem to be romance

Adapted from Jure Leskovec, Stanford CS246: Mining Massive Datasets

# SVD: movie “concepts”

$A = U \Sigma V^T$

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\
\end{array}
$$

$$
\left[\begin{array}{ccccc}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{array}\right]
=
\left[\begin{array}{ccc}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{array}\right]
\times
\left[\begin{array}{ccc}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{array}\right]
\times
\left[\begin{array}{ccccc}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{array}\right]
$$

Movies that seem to be romance

Some 3rd concept for movies that Matrix and Serenity are similar on, but other movies are quite different

# SVD: movie “concepts”

$$A = U \Sigma V^T$$

SciFi-concept

“Strength” of scifi-concept

$$\begin{array}{c} \begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \end{array} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

# SVD: movie “concepts”

$$A = U \Sigma V^T$$

Romance-concept

“Strength” of romance-concept

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie}
\end{array}
$$

$$
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

Adapted from Jure Leskovec, Stanford CS246: Mining Massive Datasets

# SVD: movie "concepts"

$$A = U \Sigma V^T$$

"Strength" of unknown-3rd-concept

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie}
\end{array}
$$

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

# SVD: movie "concepts"

$$A = U \Sigma V^T$$

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

$\Sigma$ = 3x3 concept matrix
indicating strength of concepts

# SVD: movie “concepts”

$A = \boxed{U \Sigma V^T}$

Which matrix represents “user-to-concept”?

Which matrix represents “movie-to-concept”?

$$\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array}$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

# SVD: movie “concepts”

$A = U \Sigma V^T$

Which matrix represents “user-to-concept”? U

Which matrix represents “movie-to-concept”? V

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\
\end{array}
$$

$$
\begin{bmatrix}
\mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\
\mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\
\mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\
\mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\
0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\
0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\
0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2}
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 & -0.01 \\
\mathbf{0.41} & 0.07 & -0.03 \\
\mathbf{0.55} & 0.09 & -0.04 \\
\mathbf{0.68} & 0.11 & -0.05 \\
0.15 & \mathbf{-0.59} & \mathbf{0.65} \\
0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\
0.07 & \mathbf{-0.29} & \mathbf{0.32}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{12.4} & 0 & 0 \\
0 & \mathbf{9.5} & 0 \\
0 & 0 & \mathbf{1.3}
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\
0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09
\end{bmatrix}
$$

(This is $V^T$, which is concept-to-movie)

# SVD: movie "concepts"

$$A = U \Sigma V^T$$

Dimensions:
(7users x 3concepts)
x (3concepts x 3concepts)
x (3concepts x 5movies)

$$\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\ \end{array}$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

Σ = 3x3 concept matrix
indicating strength of concepts

# SVD: movie “concepts”

$A = U \Sigma V^T$

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie}
\end{array}
$$

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0.13} & 0.02 \\
\mathbf{0.41} & 0.07 \\
\mathbf{0.55} & 0.09 \\
\mathbf{0.68} & 0.11 \\
0.15 & \mathbf{-0.59} \\
0.07 & \mathbf{-0.73} \\
0.07 & \mathbf{-0.29}
\end{bmatrix}
\times
\begin{bmatrix}
12.4 & 0 & 0 \\
0 & 9.5 & 0 \\
0 & 0 &
\end{bmatrix}
\times
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69}
\end{bmatrix}
$$

What if we just get rid of the unknown concept with low “strength”?


**Me:** using SVD decomposition to reduce one dimesion of my data
**Deleted dimension:**
SABC 1
Am I a joke to you?

# SVD: reducing "concepts" dimension

New U

New Σ

New $V^T$

$$\begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 \\ 0 & \mathbf{9.5} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}$$

Now we only have 2 concepts: scifi and romance

# SVD: reducing "concepts" dimension

New U — New Σ — New $V^T$

$$\begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 \\ 0 & \mathbf{9.5} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}$$

New Dimensions:
(7users x 2concepts)
x (2concepts x 2concepts)
x (2concepts x 5movies)

Now we only have 2 concepts: scifi and romance

# SVD: reducing "concepts" dimension

New U   New Σ   New $V^T$

$$\begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix} \times \begin{bmatrix} \mathbf{12.4} & 0 \\ 0 & \mathbf{9.5} \end{bmatrix} \times \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}$$

**What rank is this new $U\Sigma V^T$?**

# SVD: reducing "concepts" dimension

$$\underbrace{\begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix}}_{\text{New U}} \times \underbrace{\begin{bmatrix} \mathbf{12.4} & 0 \\ 0 & \mathbf{9.5} \end{bmatrix}}_{\text{New } \Sigma} \times \underbrace{\begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}}_{\text{New } V^T}$$

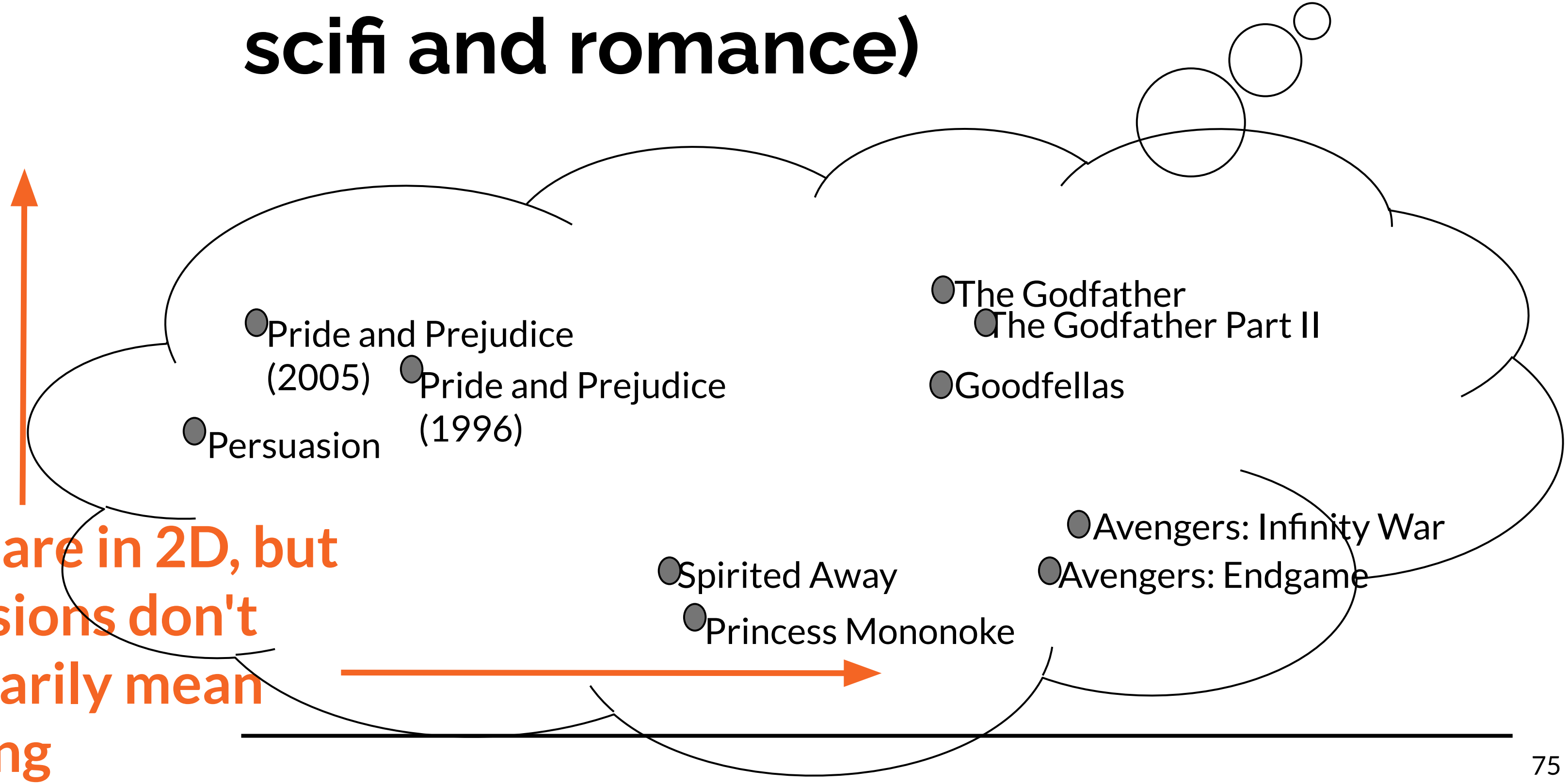
Now we only have matrix *rank* 2 (only scifi and romance "concepts")

# SVD: reducing "concepts" dimension

$$\underbrace{\begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix}}_{\text{New U}} \times \underbrace{\begin{bmatrix} \mathbf{12.4} & 0 \\ 0 & \mathbf{9.5} \end{bmatrix}}_{\text{New } \Sigma} \times \underbrace{\begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}}_{\text{New } V^T}$$

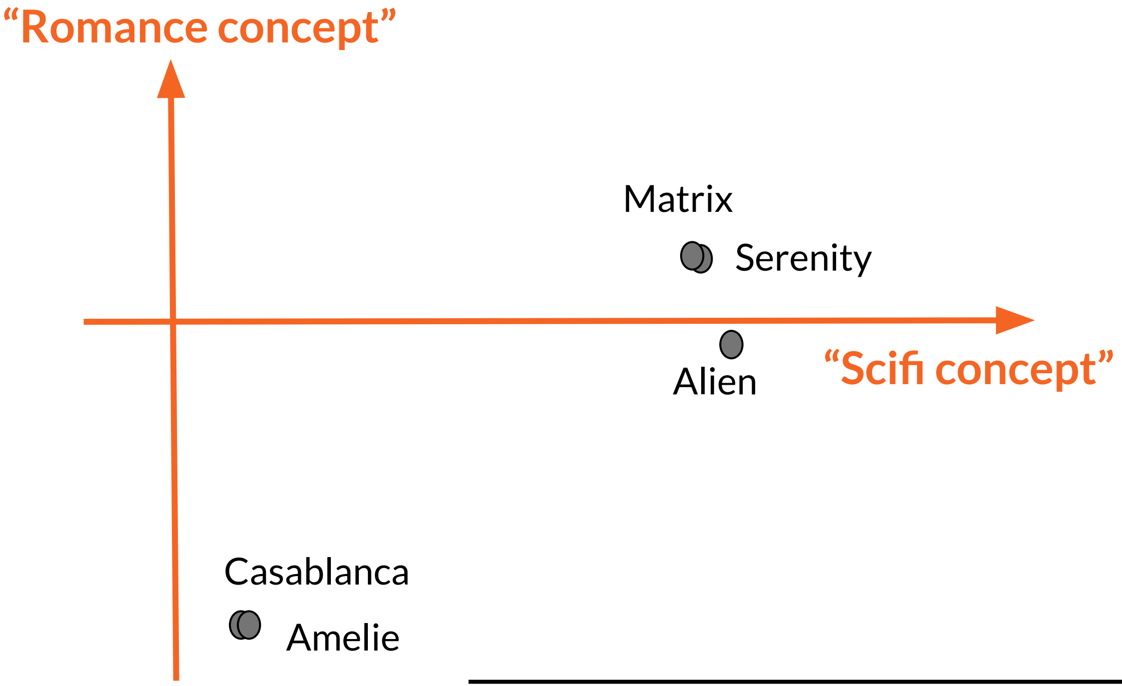
What does a "concept-to-movie" similarity *mean*?

# Map to 2 “concepts” (e.g. maybe scifi and romance)




Points are in 2D, but dimensions don't necessarily mean anything

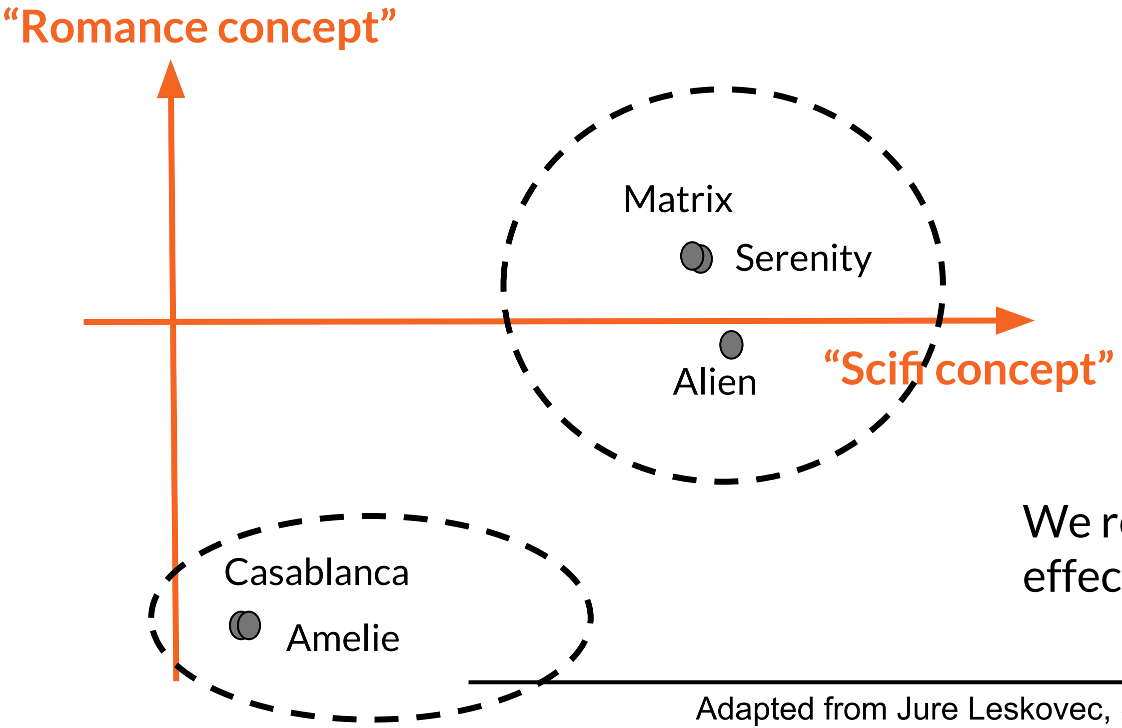
# SVD: reducing “concepts” dimension




## New $V^T$

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\
\end{array}
$$
$$
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69}
\end{bmatrix}
$$

# SVD: reducing “concepts” dimension




**New $V^T$**

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\
\end{array}
$$
$$
\begin{bmatrix}
\mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\
0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69}
\end{bmatrix}
$$

We reduced dimensions so we could effectively “cluster”!

# SVD as a matrix X-ray

- SVD gives you the best way to **approximate** any matrix (by decomposing it)
- **Principal components analysis** (PCA) is simply SVD after you normalize columns to mean 0
- If your df is missing values at random, **fill in missing elements** using e.g. column means, compute SVD, replace missing elements with SVD approximation, and iterate until convergence

https://twitter.com/WomenInStat/status/1285610321747611653

# SVD as a matrix X-ray

Efficient!

Interpretable!

Allows you to impute missing data!

- SVD gives you the best way to **approximate** any matrix (by decomposing it)
- **Principal components analysis** (PCA) is simply SVD after you normalize columns to mean 0
- If your df is missing values at random, **fill in missing elements** using e.g. column means, compute SVD, replace missing elements with SVD approximation, and iterate until convergence

https://twitter.com/WomenInStat/status/1285610321747611653

# But… why would we want to reduce “concepts”?

- Sometimes we only really need a good-enough approximation of our data

- Efficient storage matters a lot in massive datasets!

# SVD: reducing "concepts" dimension

$$\underbrace{\begin{bmatrix} \mathbf{0.13} & 0.02 \\ \mathbf{0.41} & 0.07 \\ \mathbf{0.55} & 0.09 \\ \mathbf{0.68} & 0.11 \\ 0.15 & \mathbf{-0.59} \\ 0.07 & \mathbf{-0.73} \\ 0.07 & \mathbf{-0.29} \end{bmatrix}}_{\text{New U}} \times \underbrace{\begin{bmatrix} \mathbf{12.4} & 0 \\ 0 & \mathbf{9.5} \end{bmatrix}}_{\text{New } \Sigma} \times \underbrace{\begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \end{bmatrix}}_{\text{New } V^T}$$

What happens if we multiply these new decompositions together?

# SVD: dimension reduction

Same matrix shape as original A

$$
\begin{array}{ccccc}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie}
\end{array}
\qquad
\begin{array}{cc}
\text{sci-fi} & \text{romance}
\end{array}
$$

$$
\underbrace{\left[\begin{array}{ccccc}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 0 & 0 & 2 & 2
\end{array}\right]}_{\text{Rank-2 Approximated } A}
\begin{array}{c}
\text{user 1} \\ \text{user 2} \\ \text{user 3} \\ \text{user 4} \\ \text{user 5} \\ \text{user 6} \\ \text{user 7}
\end{array}
=
\underbrace{\left[\begin{array}{cc}
.14 & 0 \\
.42 & 0 \\
.56 & 0 \\
.70 & 0 \\
0 & .60 \\
0 & .75 \\
0 & .30
\end{array}\right]}_{\substack{U \\ \text{(reduced)}}}
\overset{\text{strength}}{\underbrace{\left[\begin{array}{cc}
12.4 & 0 \\
0 & 9.5
\end{array}\right]}_{\substack{\Sigma \\ \text{(reduced)}}}}
\underbrace{\left[\begin{array}{ccccc}
.58 & .58 & .58 & 0 & 0 \\
0 & 0 & 0 & .71 & .71
\end{array}\right]}_{\substack{V^T \\ \text{(reduced)}}}
$$

# SVD: original vs. approximation

$$\begin{array}{c} \begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\ \left[\begin{array}{ccccc} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{array}\right] \end{array}$$

Original $A$

# SVD: original vs. approximation

$$\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array}$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}$$

Original $A$

Use SVD to get U, Σ, and $V^T$

# SVD: original vs. approximation

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 2 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 1 & 0 & 2 & 2
\end{bmatrix}
\end{array}
$$

Original $A$

Use SVD to get U, Σ, and $V^T$

Reduce rank to get new U', Σ', and $V^{T}$'

# SVD: original vs. approximation

$$\begin{array}{c} \begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix} \end{array}$$

Original $A$

Use SVD to get U, Σ, and $V^T$

Reduce rank to get new U', Σ', and $V^{T}$'

Multiply so new A' = U' Σ'$V^{T}$'

# SVD: original vs. approximation

$$\begin{array}{c}\begin{matrix} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}\end{array}$$

Original $A$

Use SVD to get U, Σ, and $V^T$

Reduce rank to get new U', Σ', and $V^{T}$'

Multiply so new A' = U' Σ'$V^{T}$'

$$\begin{array}{c}\begin{matrix} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}\end{array}$$

Rank-2 Approximated $A$

# SVD: original vs. approximation

$$\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array}$$
$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & \boxed{2} & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & \boxed{1} & 0 & 2 & 2 \end{bmatrix}$$

Original $A$

Dimension reduction loses some information but keeps the most important features intact (scifi, romance)!

$$\begin{array}{ccccc} \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \end{array}$$
$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & \boxed{0} & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & \boxed{0} & 0 & 2 & 2 \end{bmatrix}$$

Rank-2 Approximated $A$

Adapted from Jure Leskovec, Stanford CS246: Mining Massive Datasets

# SVD: dimension reduction

Now we get 0's in our prediction of A (same matrix shape though!)

$$
\begin{array}{ccccc|c}
\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} & \\
\end{array}
$$

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 0 & 0 & 2 & 2
\end{bmatrix}
\begin{array}{l}
\text{user 1} \\ \text{user 2} \\ \text{user 3} \\ \text{user 4} \\ \text{user 5} \\ \text{user 6} \\ \text{user 7}
\end{array}
\begin{array}{c}
\begin{array}{cc} \text{sci-fi} & \text{romance} \end{array} \\
\begin{bmatrix}
.14 & 0 \\
.42 & 0 \\
.56 & 0 \\
.70 & 0 \\
0 & .60 \\
0 & .75 \\
0 & .30
\end{bmatrix}
\end{array}
\begin{array}{c}
\text{strength} \\
\begin{bmatrix}
12.4 & 0 \\
0 & 9.5
\end{bmatrix}
\end{array}
\begin{bmatrix}
.58 & .58 & .58 & 0 & 0 \\
0 & 0 & 0 & .71 & .71
\end{bmatrix}
$$

Approximated $A$ $\qquad$ $U$ (reduced) $\qquad$ $\Sigma$ (reduced) $\qquad$ $V^T$ (reduced)

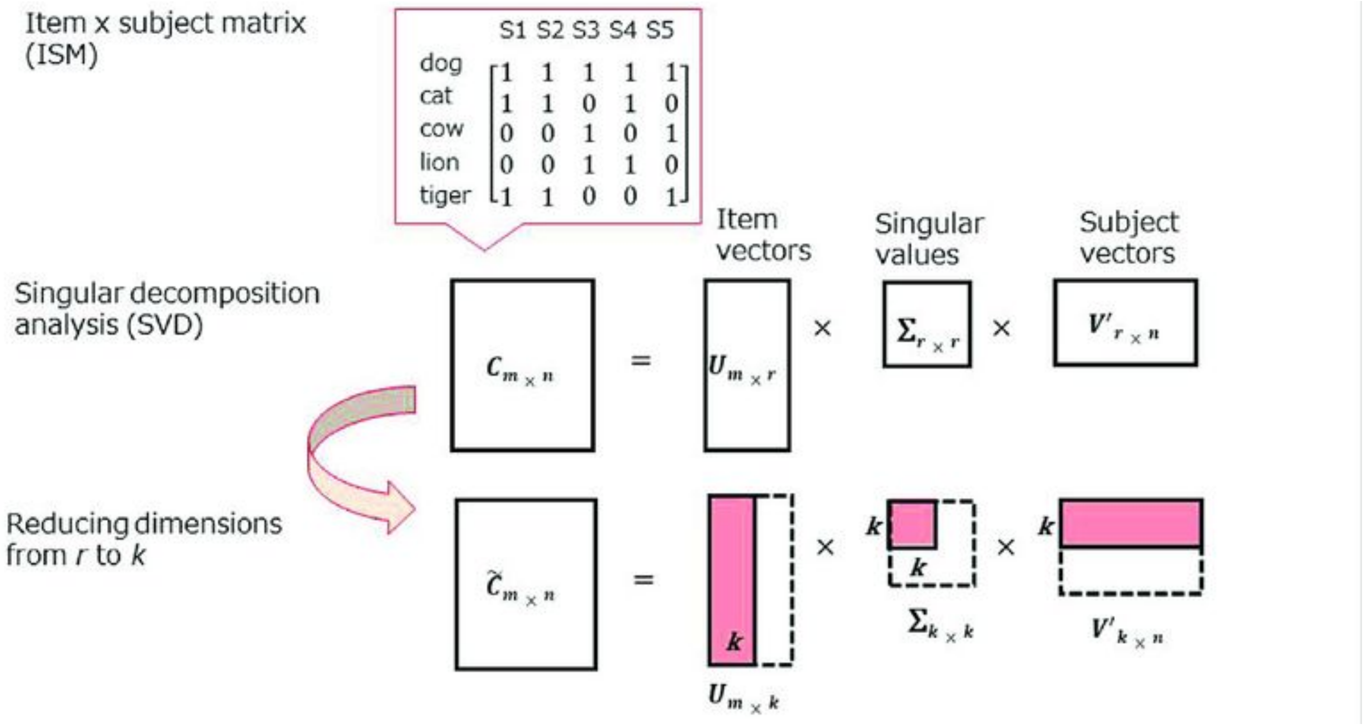
# SVD: dimension reduction

$$
\begin{array}{c}
\begin{array}{ccccc}\text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie}\end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 0 & 0 & 2 & 2
\end{bmatrix}
\end{array}
\begin{array}{l}
\text{user 1} \\ \text{user 2} \\ \text{user 3} \\ \text{user 4} \\ \text{user 5} \\ \text{user 6} \\ \text{user 7}
\end{array}
\begin{array}{c}
\begin{array}{cc}\text{sci-fi} & \text{romance}\end{array} \\
\begin{bmatrix}
.14 & 0 \\
.42 & 0 \\
.56 & 0 \\
.70 & 0 \\
0 & .60 \\
0 & .75 \\
0 & .30
\end{bmatrix}
\end{array}
\begin{array}{c}
\text{strength} \\
\begin{bmatrix}
12.4 & 0 \\
0 & 9.5
\end{bmatrix}
\end{array}
\begin{bmatrix}
\boxed{\begin{array}{ccccc}
.58 & .58 & .58 & 0 & 0 \\
0 & 0 & 0 & .71 & .71
\end{array}}
\end{bmatrix}
$$

Less nnz storage is needed!

$$
\begin{array}{cccc}
\text{Approximated } A & \underset{\text{(reduced)}}{U} & \underset{\text{(reduced)}}{\Sigma} & \underset{\text{(reduced)}}{V^T}
\end{array}
$$

# SVD: dimension reduction recap




https://www.i2tutorials.com/what-do-you-mean-by-singular-value-decomposition/

# Example: image compression

## cameraman

Unknown creator

Download

cameraman.tif (63.71Kb)

Alternate file

Cameraman Non-CC TOU (2.443Kb)

URI

https://hdl.handle.net/1721.3/195767

Date

1978

Abstract

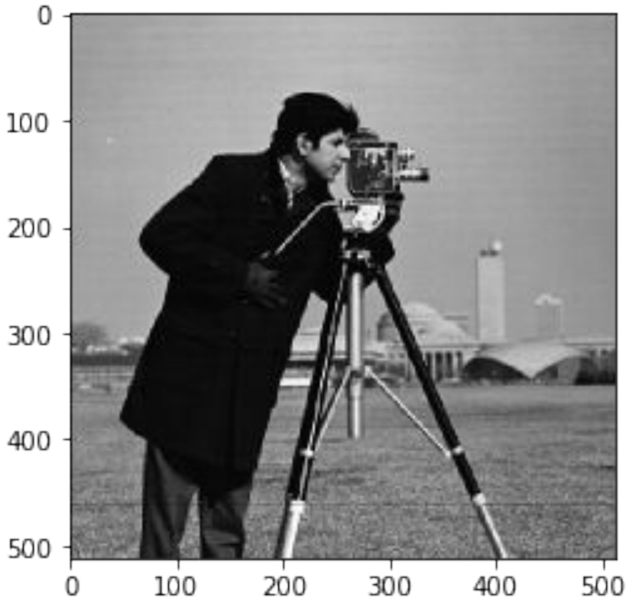
Image frequently used as a test image for image processing and compression algorithms. First known appearance in William F Schreiber's "Image Processing for Quality Improvement" in the Proceedings of IEEE, Vol. 6, No. 12, December 1978.

Rights

Creative Commons Attribution Non-Commerical. For use in journal publications or trade and educational book publishers that might fall outside the CC license terms, the terms provided in the additional download apply (see Cameraman Non-CC TOU). https://creativecommons.org/licenses/by-nc/4.0/

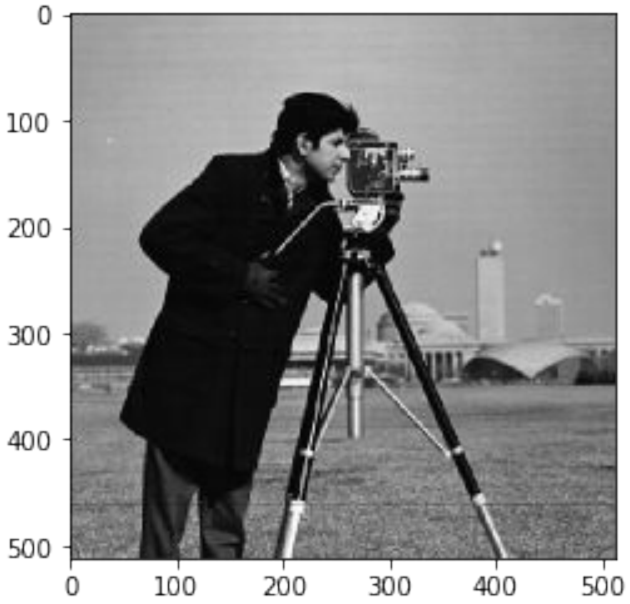
https://dome.mit.edu/handle/1721.3/195767

# Example: image compression



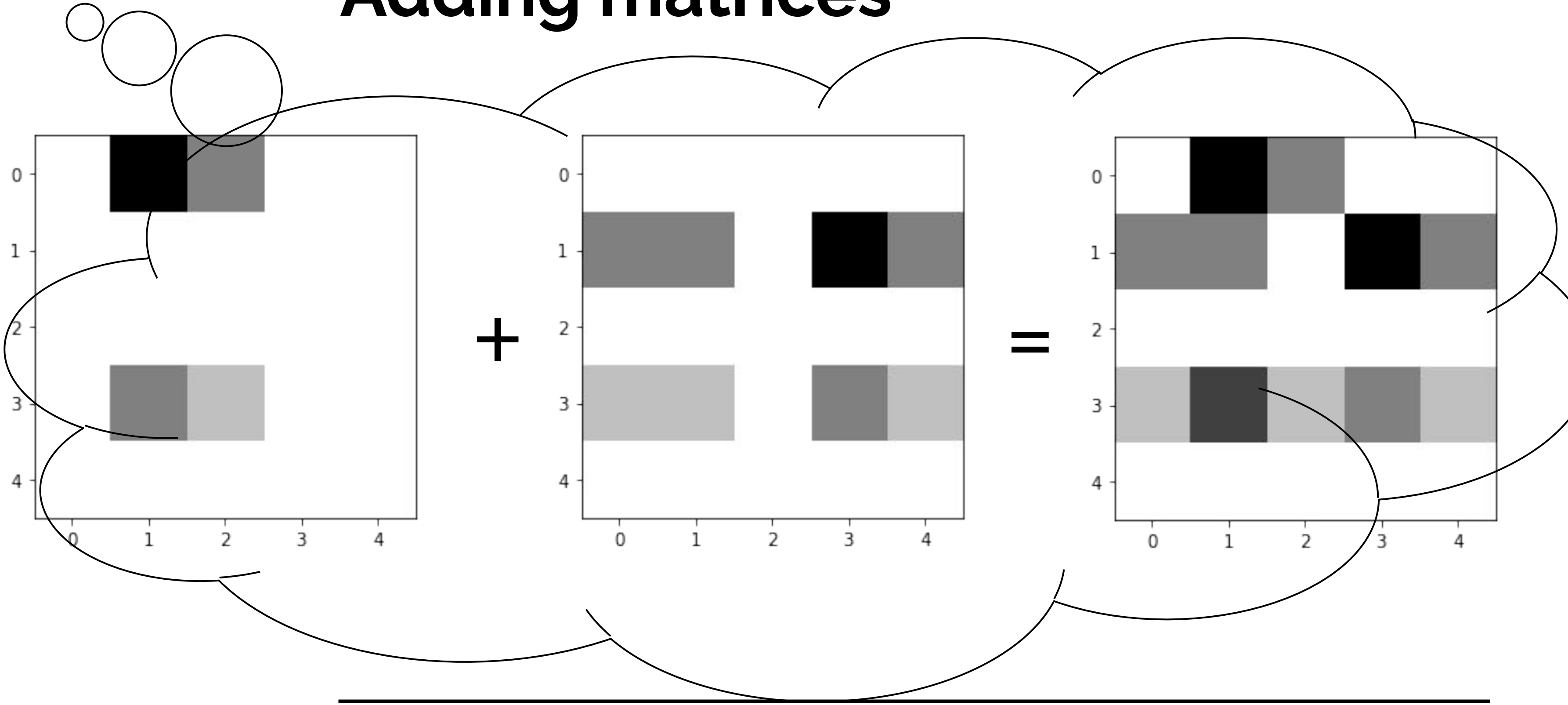

https://dome.mit.edu/handle/1721.3/195767

# Example: image compression

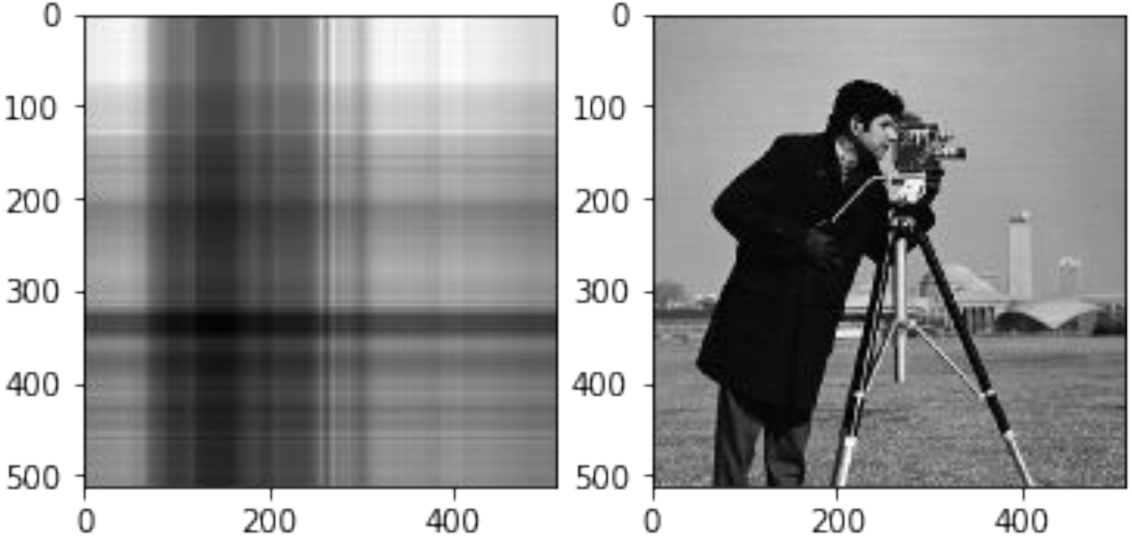



512 x 512 pixels =
262,144 numbers

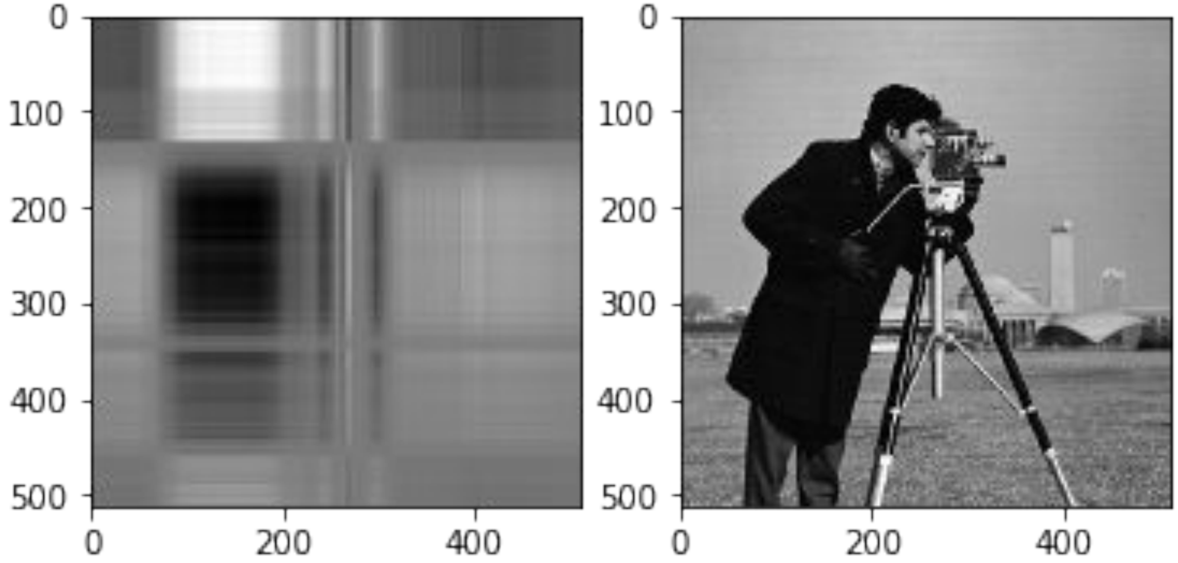
# Adding matrices



# Component #1



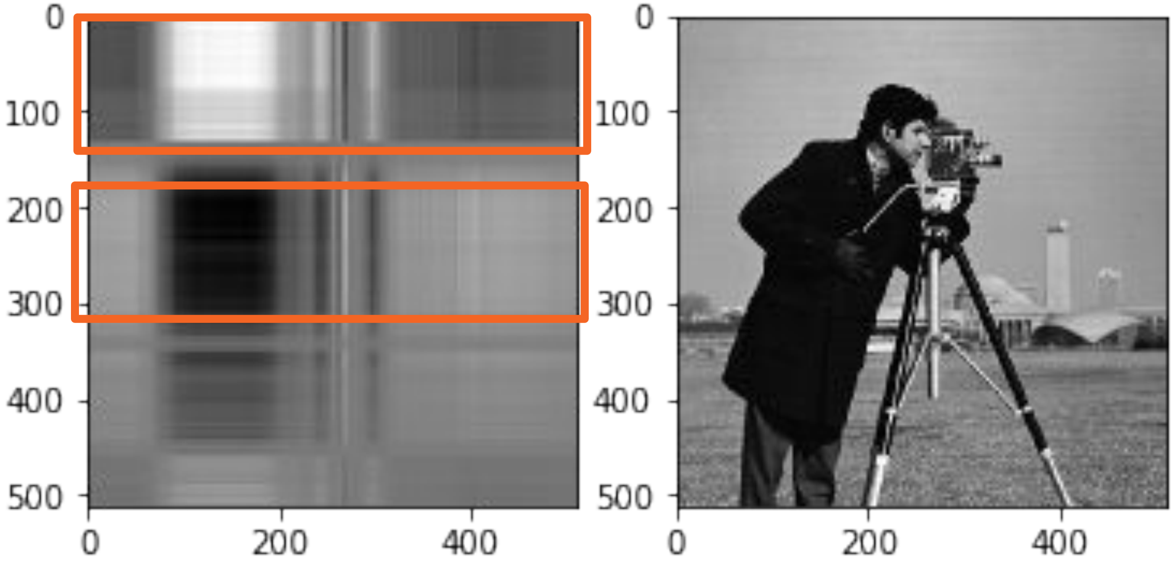
# Component #2



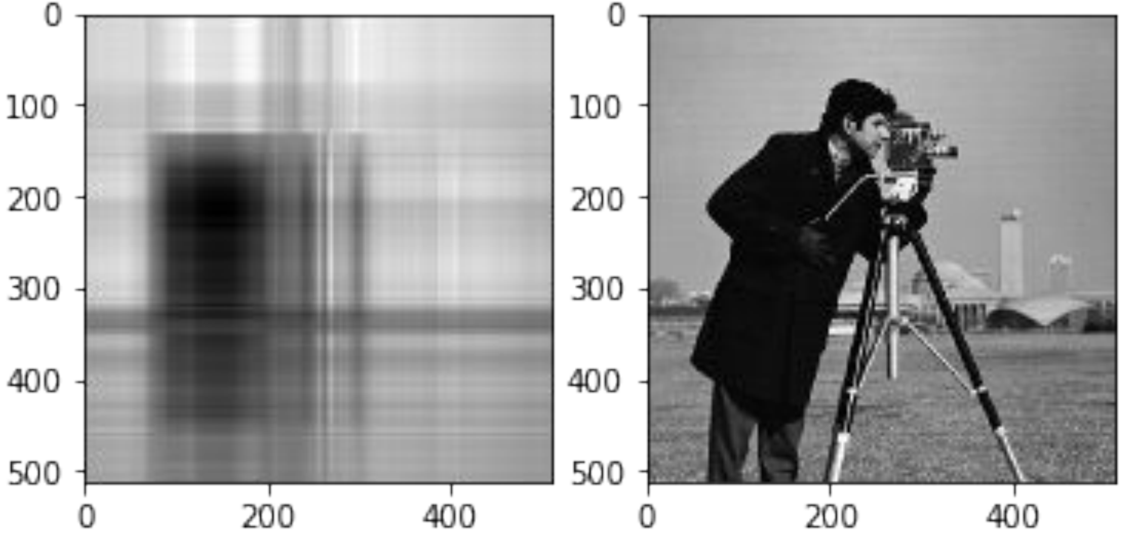
# Component #2

Values can be negative

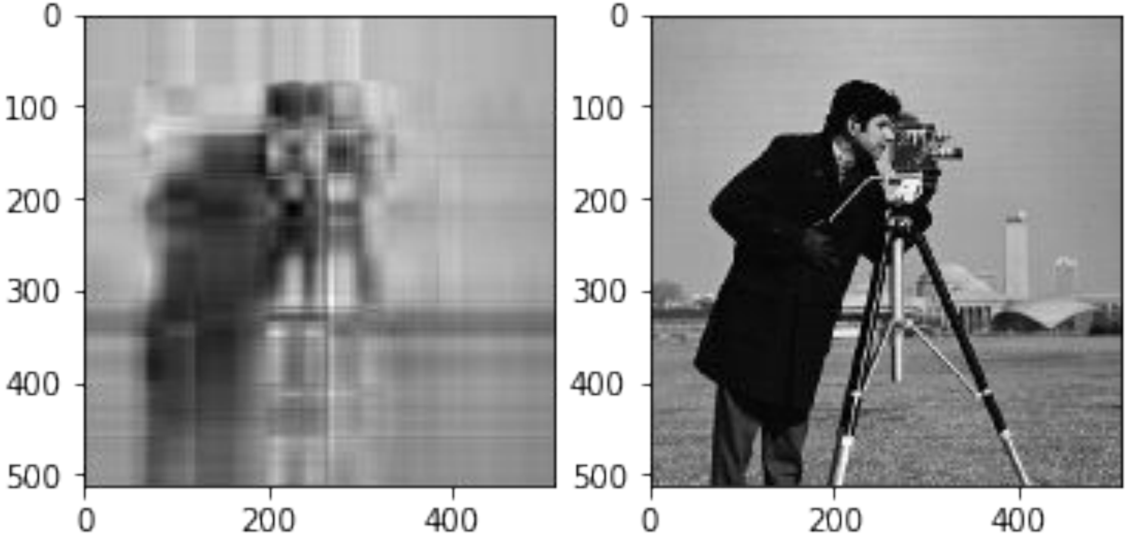
or positive



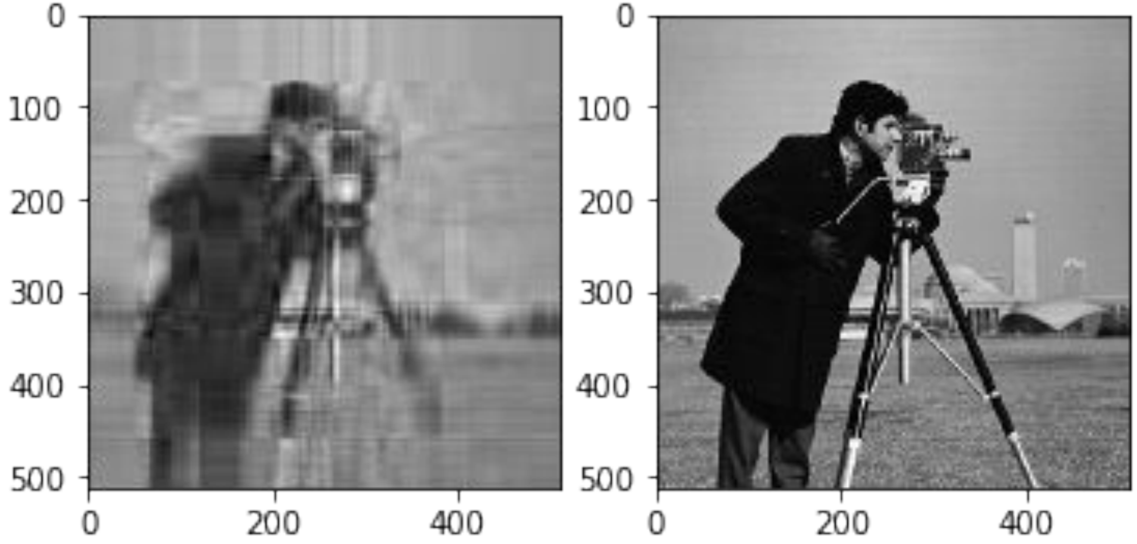
# #1 + #2, rank 2 approximation



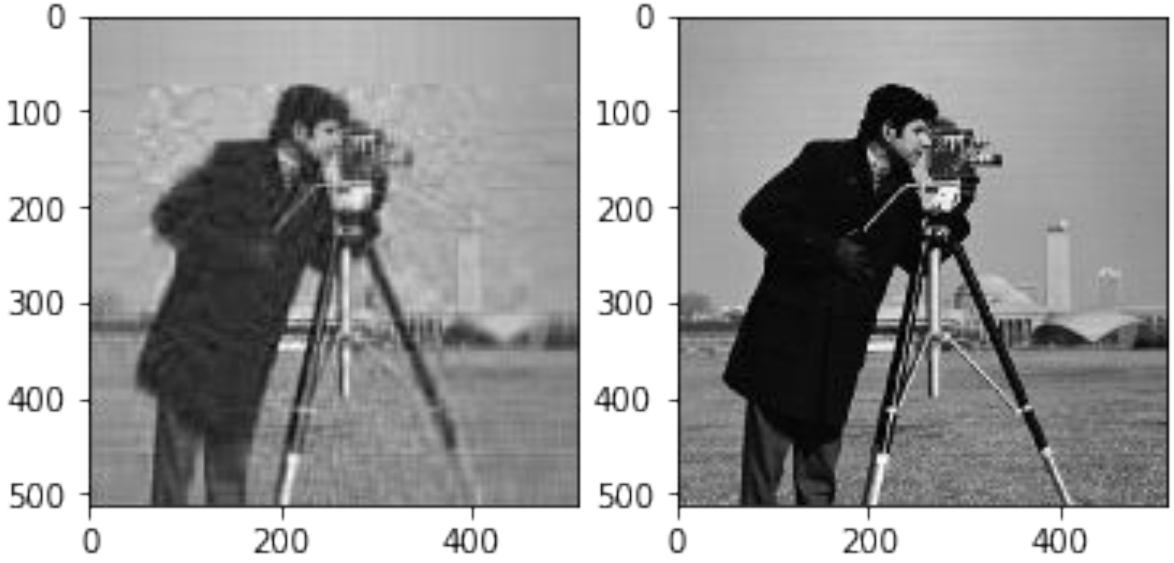
# Rank 5 approximation



# Rank 10 approximation



# Rank 20

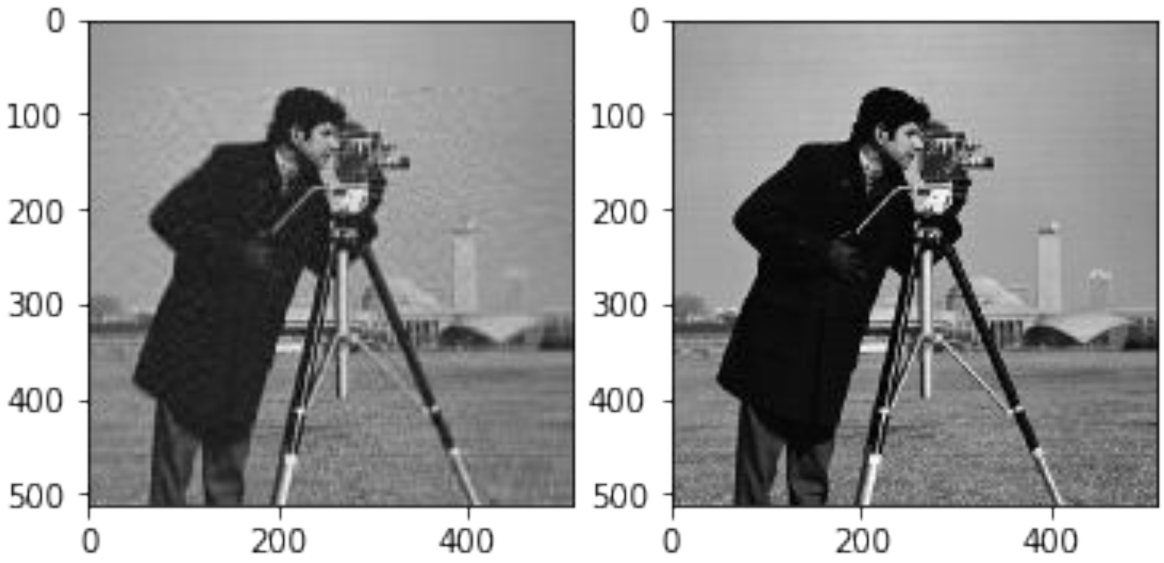


# Rank 40

512 x 512 pixels =
262,144 numbers

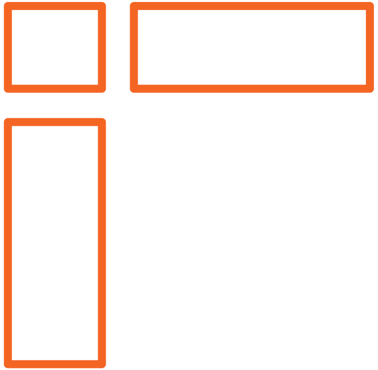
vs.

512 x 40 x 2 + 40
= 41,000 numbers



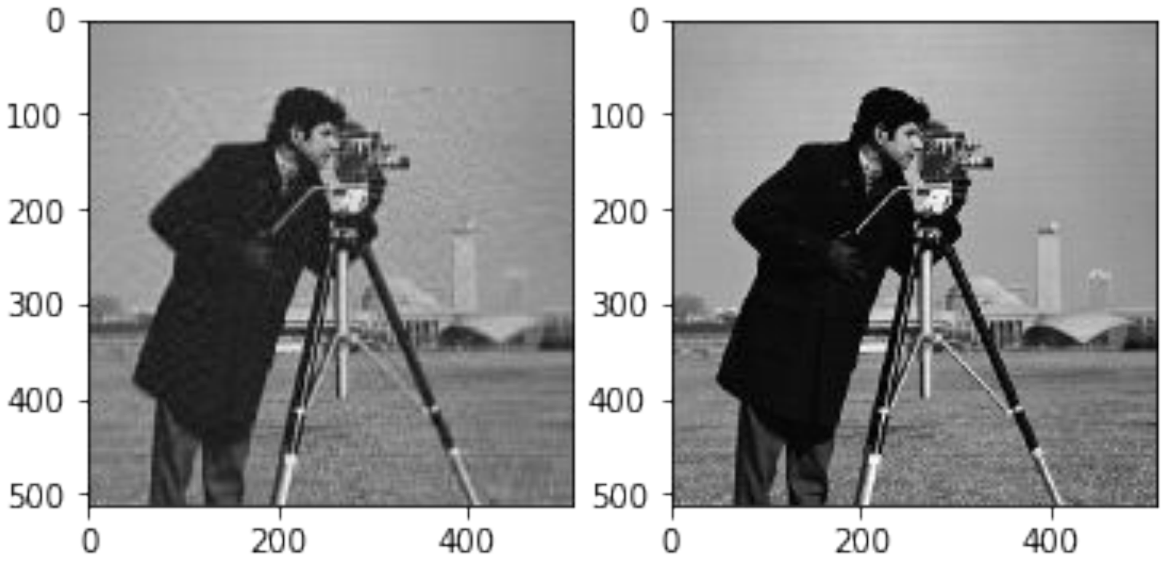
# Rank 40

(40x 40) (40 x 512)



(512 x 40)

512 x 40 x 2 + 40
= 41,000 numbers



# Rank 40

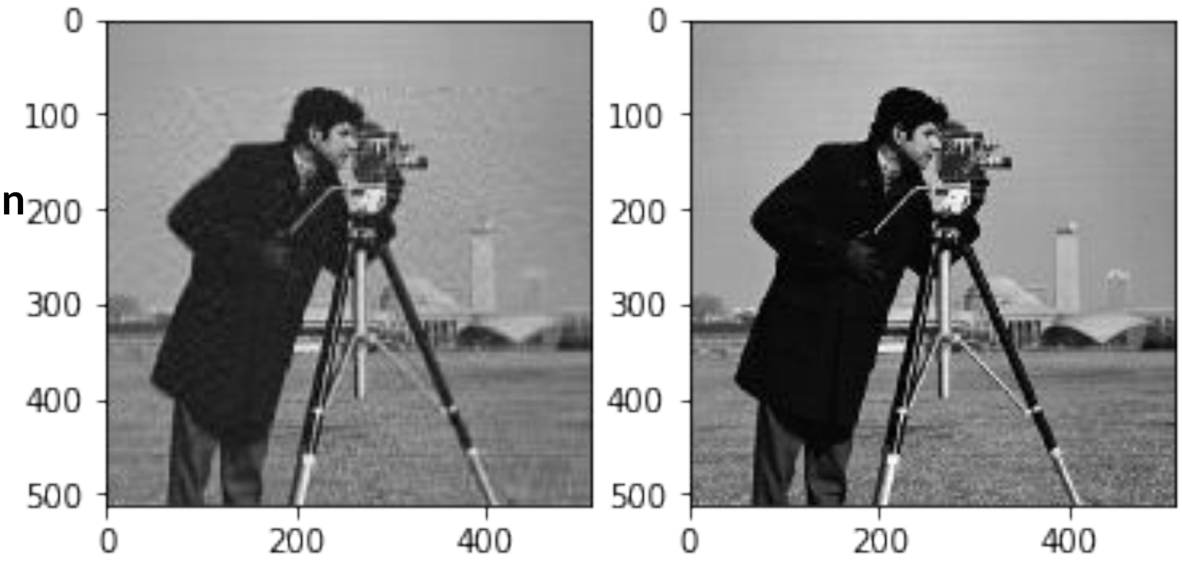
(40x 40) (40 x 512)

Only 40 nnz (on the diagonal)!

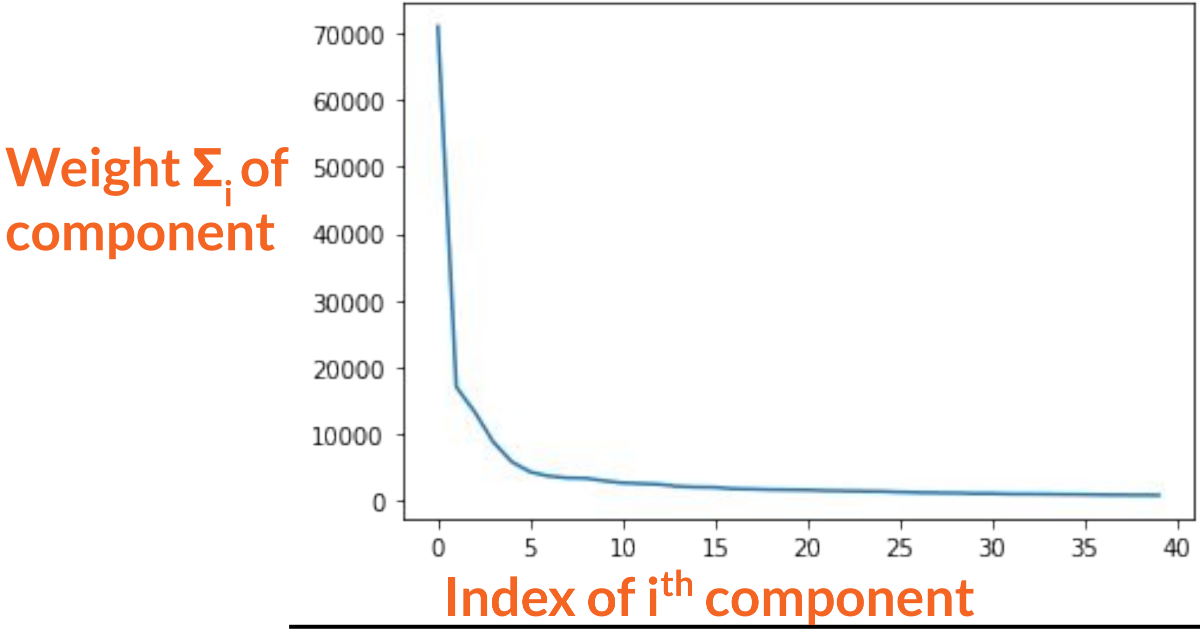
(512 x 40)

512 x 40 x 2 + 40

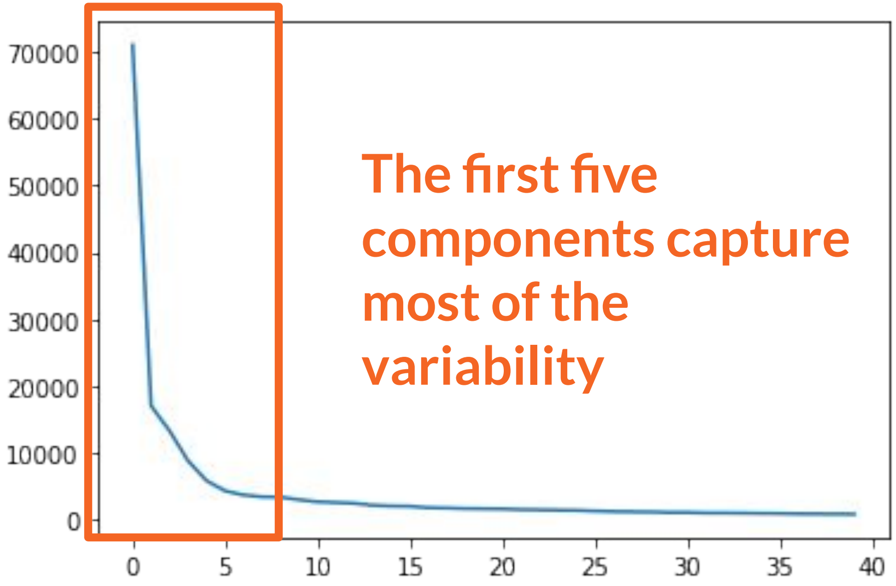
= 41,000 numbers



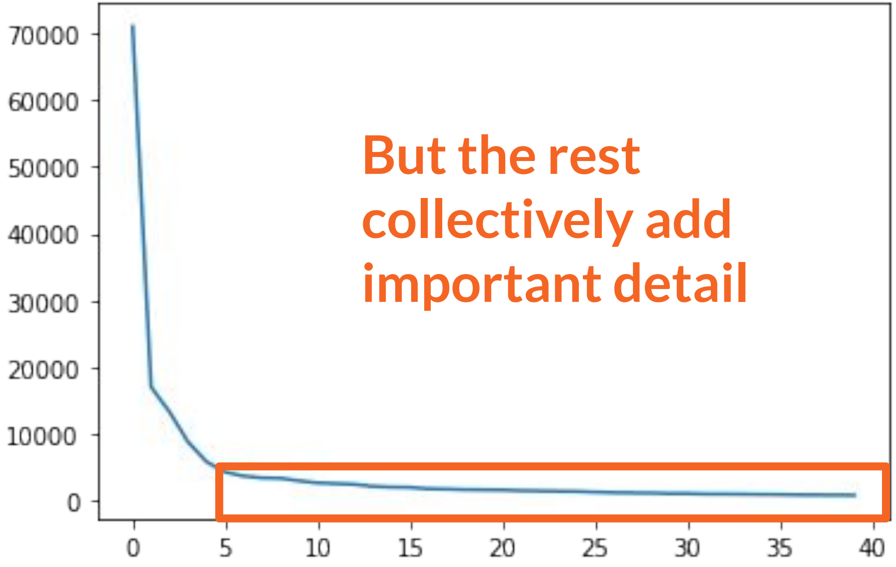
# How much does each component contribute?



# How much does each component contribute?



# How much does each component contribute?



# On image ethics: *lenna.jpg*

https://en.wikipedia.org/wiki/Lenna

# On image ethics: *lenna.jpg*

“This is one of the most widely used images in computer science. If you’ve ever taken a computer science class that worked with images, there’s a good chance you’ve used it. It also has a lesser-known, controversial history. The image comes from a 1973 *Playboy* centerfold. **It was originally used in a computer science paper because a bunch of USC scientists were writing a paper in a hurry and just needed an image to add as an example, and someone happened to walk in with a *Playboy*.** The image has been widely used ever since then, and there have been complaints for decades that it’s sexist to use it as a standard test image.”

- *Prof. Emma Pierson*

https://epierson9.github.io/playboy-kittens-and-deep-learning/

# On image ethics: *lenna.jpg*

- 2017: *Journal of Modern Optics* suggests the Cameraman image as an alternative to Lenna

# On image ethics: *lenna.jpg*

- 2017: *Journal of Modern Optics* suggests the Cameraman image as an alternative to Lenna
- 2018: *Nature Nanotechnology* announces they “no longer consider articles using the Lenna image.”

https://en.wikipedia.org/wiki/Lenna

# On image ethics: *lenna.jpg*

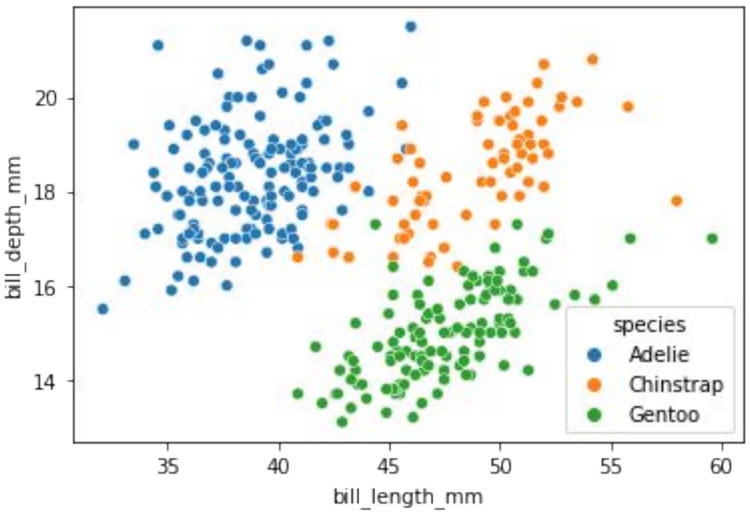
- 2017: *Journal of Modern Optics* suggests the Cameraman image as an alternative to Lenna
- 2018: *Nature Nanotechnology* announces they "no longer consider articles using the Lenna image."
- 2019: Lena Forsén, in film documentary *Losing Lena*, states "*I retired from modeling a long time ago. It's time I retired from tech, too... Let's commit to losing me.*"

https://en.wikipedia.org/wiki/Lenna

# 1 min break & attendance

tinyurl.com/un2n4xuh

# Example: penguins 🐧

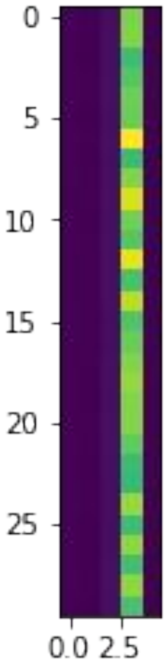


# Can we show more information?

| | species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---|---|---|---|---|---|---|
| **0** | Adelie | Torgersen | 39.1 | 18.7 | 181.0 | 3750.0 | Male |
| **1** | Adelie | Torgersen | 39.5 | 17.4 | 186.0 | 3800.0 | Female |
| **2** | Adelie | Torgersen | 40.3 | 18.0 | 195.0 | 3250.0 | Female |
| **4** | Adelie | Torgersen | 36.7 | 19.3 | 193.0 | 3450.0 | Female |
| **5** | Adelie | Torgersen | 39.3 | 20.6 | 190.0 | 3650.0 | Male |

# Data table as image

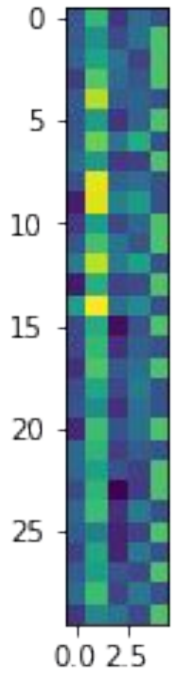
| | species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---|---|---|---|---|---|---|
| **0** | Adelie | Torgersen | 39.1 | 18.7 | 181.0 | 3750.0 | Male |
| **1** | Adelie | Torgersen | 39.5 | 17.4 | 186.0 | 3800.0 | Female |
| **2** | Adelie | Torgersen | 40.3 | 18.0 | 195.0 | 3250.0 | Female |
| **4** | Adelie | Torgersen | 36.7 | 19.3 | 193.0 | 3450.0 | Female |
| **5** | Adelie | Torgersen | 39.3 | 20.6 | 190.0 | 3650.0 | Male |



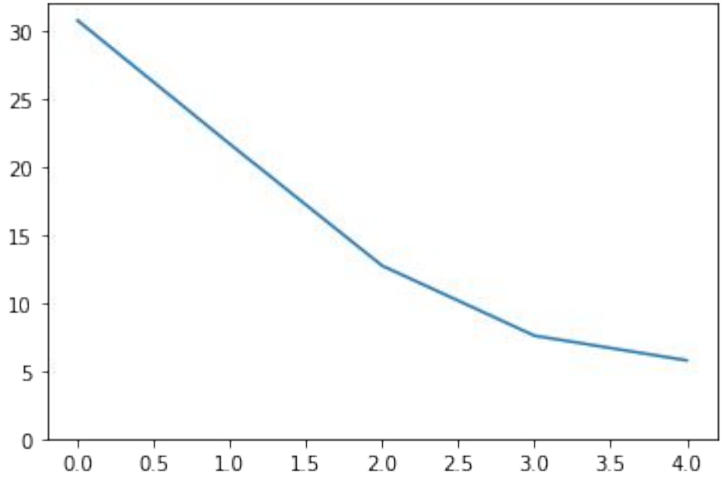
# Data table as image

| | species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---|---|---|---|---|---|---|
| **0** | Adelie | Torgersen | 39.1 | 18.7 | 181.0 | 3750.0 | Male |
| **1** | Adelie | Torgersen | 39.5 | 17.4 | 186.0 | 3800.0 | Female |
| **2** | Adelie | Torgersen | 40.3 | 18.0 | 195.0 | 3250.0 | Female |
| **4** | Adelie | Torgersen | 36.7 | 19.3 | 193.0 | 3450.0 | Female |
| **5** | Adelie | Torgersen | 39.3 | 20.6 | 190.0 | 3650.0 | Male |

Replace raw values with z-scores



# Component weights

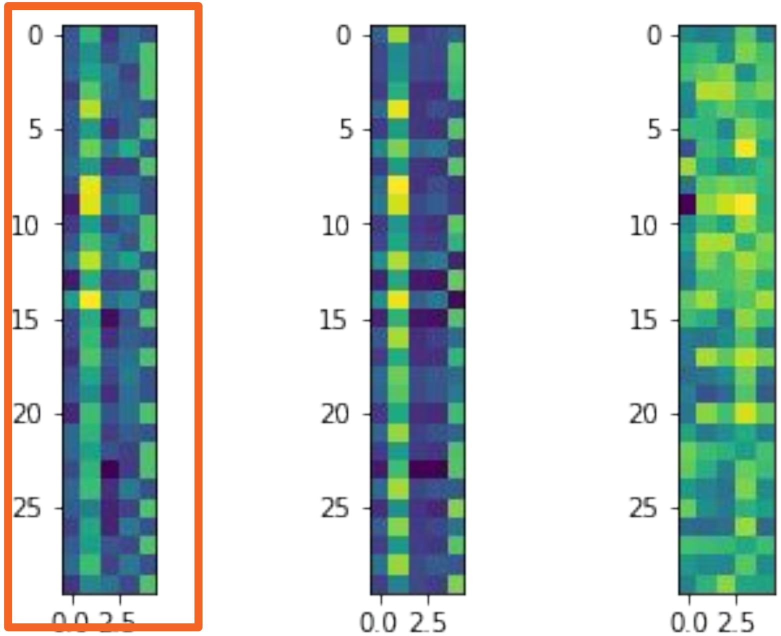



Original data is rank 5, so we can only get five components

Much less steep drop off than image!

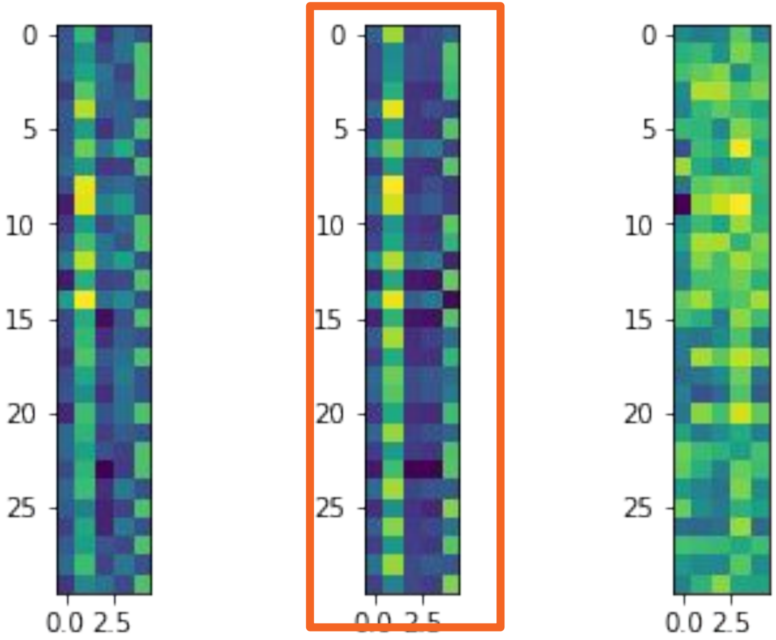
# Visualizing approximate penguins

Real penguin data

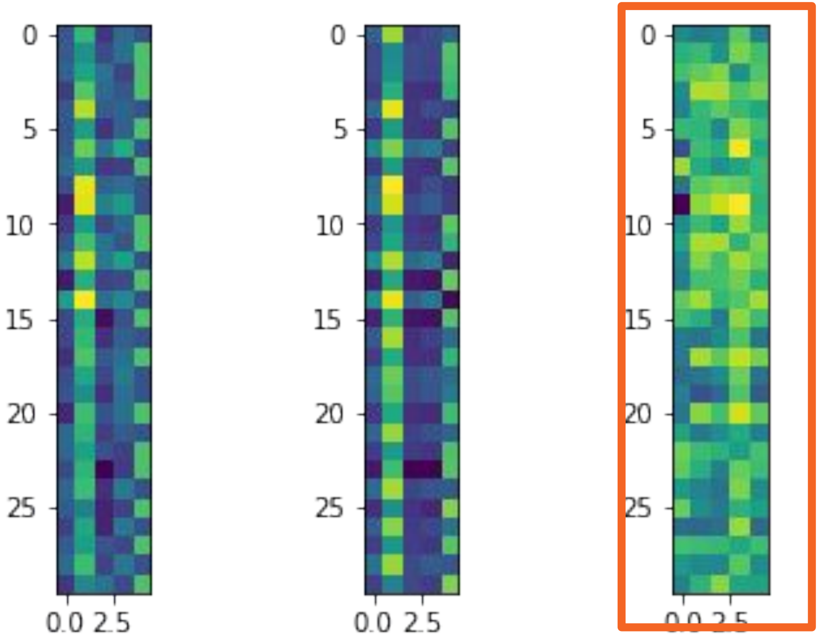


# Visualizing approximate penguins

Rank 2 approximation with SVD



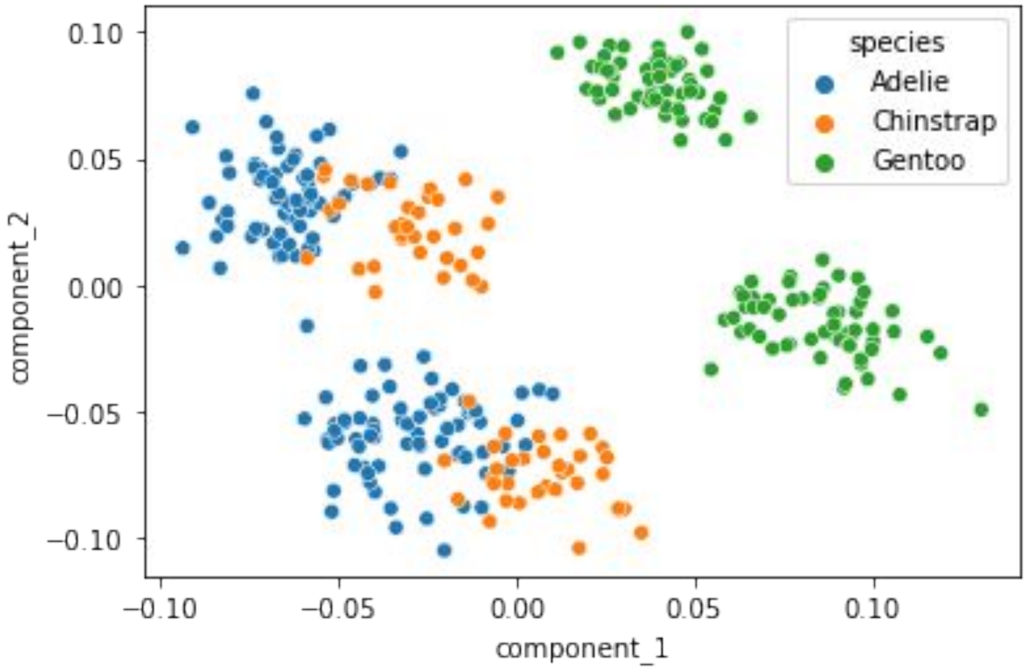
# Visualizing approximate penguins



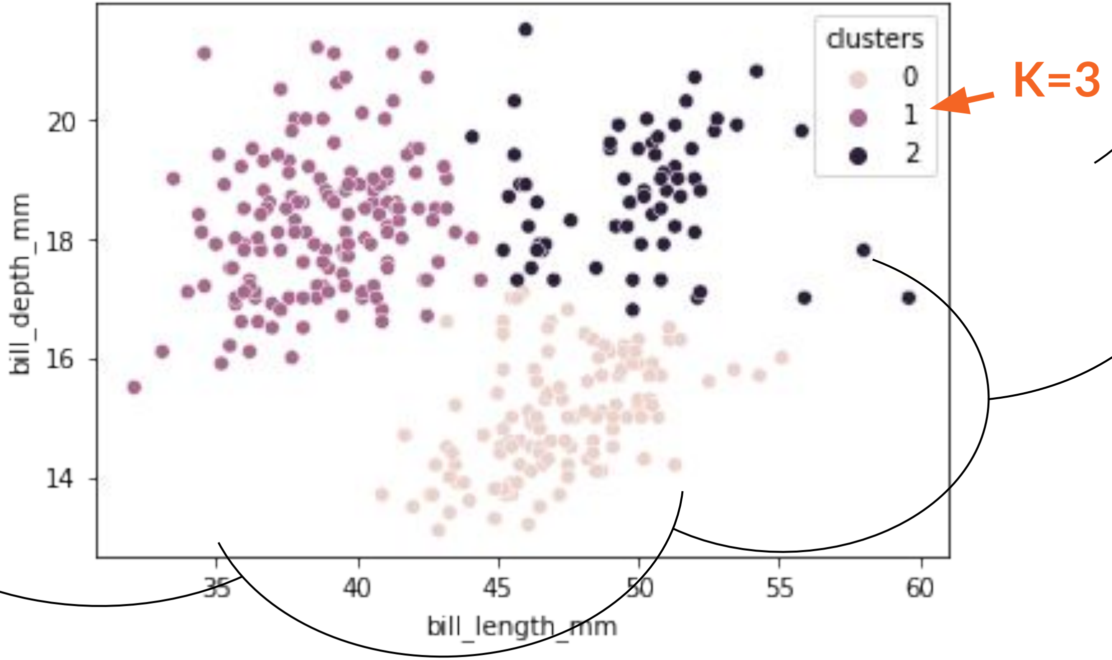

Error of rank 2 approximation

# Penguin similarity at rank 2

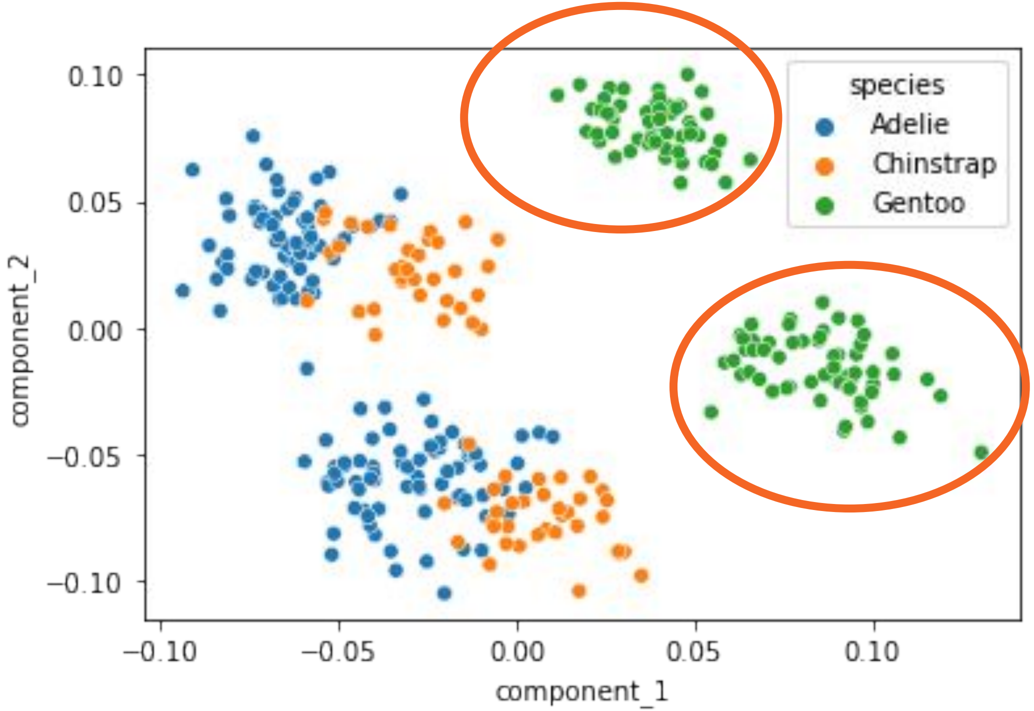
X, Y positions now represent five input variables



# Output: k-means clustering



# Penguin similarity at rank 2




SVD captures much more detail than just k=3 clustering

# Case study: Goodreads reviews

**Can we use SVD to recommend books based on Goodreads user ratings?**

# Case study: Goodreads reviews

## Goodreads Book Graph Datasets

### Overview

These datasets were collected in late 2017 from goodreads.com, where we only scraped users' public shelves, i.e. everyone can see it on web without login. User IDs and review IDs are anonymized. We collected these datasets for academic use only. Please do not redistribute them or use for commercial purposes.

We collected three groups of datasets: (1) meta-data of the books, (2) user-book interactions (users' public shelves) and (3) users' detailed book reviews. These datasets can be merged together by joining on book/user/review ids.

Basic Statistics of the Complete Book Graph:

- 2,360,655 books (1,521,962 works, 400,390 book series, 829,529 authors)
- 876,145 users; 228,648,342 user-book interactions in users' shelves (include 112,131,203 reads and 104,551,549 ratings)

**Download links** to these datasets can be found in the Datasets section below.

Note the complete interaction dataset is very large! We extracted several medium-size subsets by genre and recommend using these subsets for experimentation first (see "**By Genre**" in the Datasets section for details).

https://mengtingwan.github.io/data/goodreads.html

**How do we get the data?**

- Scraping is difficult but possible
- Is it ethical to use user/social media data?
- Already scraped data exists (but is outdated): UCSD book graph

# Users, books, and ratings

Decisions to make:

- What matrix do I want to start with?
- What constitutes an interaction?
    - On a shelf? Has read? Has rated? Has reviewed?
- Do I care about rating, or just binary interaction?

# Goodreads User Ratings

| | User 0 | User 1 | User 2 | User 3 | User 4 | …. | User N |
|---|---|---|---|---|---|---|---|
| Book 0 | 5 | | | 5 | 1 | | |
| Book 1 | | | | | | | |
| Book 2 | 2 | 1 | | 3 | | | 3 |
| Book 3 | | 1 | | | | | |
| Book 4 | | | 4 | 3 | 2 | | |
| … | | | | | | | |
| Book N | 1 | | | 2 | | | |

# Goodreads User Ratings

| | User 0 | User 1 | User 2 | User 3 | User 4 | …. | User N |
|---|---|---|---|---|---|---|---|
| Book 0 | 5 | | | 5 | 1 | | |
| Book 1 | | | | | | | |
| Book 2 | 2 | 1 | | 3 | | | 3 |
| Book 3 | | 1 | | | | | |
| Book 4 | | | 4 | 3 | 2 | | |
| … | | | | | | | |
| Book N | 1 | | | 2 | | | |

Columns are reviewers

# Goodreads User Ratings

Rows are books

| | User 0 | User 1 | User 2 | User 3 | User 4 | …. | User N |
|---|---|---|---|---|---|---|---|
| Book 0 | 5 | | | 5 | 1 | | |
| Book 1 | | | | | | | |
| Book 2 | 2 | 1 | | 3 | | | 3 |
| Book 3 | | 1 | | | | | |
| Book 4 | | | 4 | 3 | 2 | | |
| … | | | | | | | |
| Book N | 1 | | | 2 | | | |

# Goodreads User Ratings

| | User 0 | User 1 | User 2 | User 3 | User 4 | …. | User N |
|---|---|---|---|---|---|---|---|
| Book 0 | 5 | | | 5 | 1 | | |
| Book 1 | | | | | | | |
| Book 2 | 2 | 1 | | 3 | | | 3 |
| Book 3 | | 1 | | | | | |
| Book 4 | | | 4 | 3 | 2 | | |
| … | | | | | | | |
| Book N | 1 | | | 2 | | | |

Values are ratings (out of 5)

Missing value == no rating

# Issues with Goodreads data

**Scale** (in this dataset)

836,434 users

2,339,816 books

228,648,343 total interactions (4.3GB)

**Sparsity** (in this dataset)

Top book has 285k interactions

Median book has 5 user interactions

500000 books have one (!) interaction

# Strategy for making recommendations

1. **Count** user and book interactions
2. **Prioritize** most common books and most prolific reviewers
3. **Filter** to 5000 books x 10000 users
4. Create **sparse** matrix
5. Use **approximate** truncated SVD

# Strategy for making recommendations

How many MB @ 8 bytes if dense matrix?

1. **Count** user and book interactions
2. **Prioritize** most common books and most prolific reviewers
3. **Filter** to 5000 books x 10000 users
4. Create **sparse** matrix
5. Use **approximate** truncated SVD

# Strategy for making recommendations

How many MB @ 8 bytes if dense matrix?

1. **Count** user and book interactions
2. **Prioritize** most common books and most prolific reviewers
3. **Filter** to 5000 books x 10000 users
4. Create **sparse** matrix
5. Use **approximate** truncated SVD

**400 MB**

(5,000 x 10,000 x 8) / 1,000,000

1. **Count** user and book interactions
2. **Prioritize** most common books and most prolific reviewers
3. **Filter** to 5000 books x 10000 users

| | User 0 | User 1 | User 2 | User 3 | User 4 | … | User N |
|---|---|---|---|---|---|---|---|
| Book 0 | 5 | | | 5 | 1 | | |
| Book 1 | | | | | | | |
| Book 2 | 2 | 1 | | 3 | | | 3 |
| Book 3 | | 1 | | | | | |
| Book 4 | | | 4 | 3 | 2 | | 1 |
| … | | | | | | | |
| Book N | 1 | | | 2 | | | |

836,434 users x 2,339,816 books

| | User 0 | User 3 | User 4 | … | User N |
|---|---|---|---|---|---|
| Book 0 | 5 | 5 | 1 | | |
| Book 2 | 2 | 3 | | | 3 |
| Book 4 | | 3 | 2 | | 1 |
| … | | | | | |
| Book N | 1 | 2 | | | |

10000 users x 5000 books

# 4. Create **sparse** matrix

| | User 0 | User 3 | User 4 | User N |
|---|---|---|---|---|
| Book 0 | 5 | 5 | 1 | |
| Book 2 | 2 | 3 | | 3 |
| Book 4 | | 3 | 2 | 1 |
| Book N | 1 | 2 | | |

Value = [ 5, 5, 1, 2, 3, 3, 3, 2, 1, 1, 2 ]
Column_Index = [ 0, 1, 2, 0, 1, 3, 1, 2, 3, 0, 1]
Row_Index = [ 0, 3, 6, 9,  11]

Is this sparse matrix in:

**A)** Coordinate List (COO) format
**B)** Compressed Sparse Row (CSR) format

# 4. Create **sparse** matrix

| | User 0 | User 3 | User 4 | User N |
|---|---|---|---|---|
| Book 0 | 5 | 5 | 1 | |
| Book 2 | 2 | 3 | | 3 |
| Book 4 | | 3 | 2 | 1 |
| Book N | 1 | 2 | | |

**Value = [ 5, 5, 1, 2, 3, 3, 3, 2, 1, 1, 2 ]**
**Column_Index = [ 0, 1, 2, 0, 1, 3, 1, 2, 3, 0, 1]**
**Row_Index = [ 0, 3, 6, 9, 11]**

**Row_Index is shorter than Column_Index!**

Is this sparse matrix in:

**A)** Coordinate List (COO) format
**B)** Compressed Sparse Row (CSR) format

# Compressed Sparse Row format

Length 8 → Value = [ 2, 2, 4, 2, 1, 1, 2, 1 ]
Length 8 → Column_Index = [ 0, 1, 3, 4, 0, 1, 3, 4]
Length 6 → Row_Index = [ 0, 0, 4, 4, 8, 8]

Now we only have to store 8+8+6=22 numbers instead of 25!

| | | | | |
|---|---|---|---|---|
| | | | | |
| 2 | 2 | | 4 | 2 |
| | | | | |
| 1 | 1 | | 2 | 1 |
| | | | | |

A = np.array([[0,0,0,0,0],
[2,2,0,4,2],
[0,0,0,0,0],
[1,1,0,2,1],
[0,0,0,0,0]])

# 4. Create **sparse** matrix

```
from scipy.sparse import csr_matrix

shape = (len(books), len(users))
matrix = csr_matrix((data, (book_idx, user_idx)), shape=shape)
```

(5000, 10000)

## scipy.sparse.csr_matrix

*class* scipy.sparse.**csr_matrix**(*arg1*, *shape=None*, *dtype=None*, *copy=False*) [source]

Compressed Sparse Row matrix

**csr_array((data, (row_ind, col_ind)), [shape=(M, N)])**

where `data`, `row_ind` and `col_ind` satisfy the relationship `a[row_ind[k], col_ind[k]] = data[k]`.

## 5. Use **approximate** truncated SVD

# `sklearn.decomposition`.TruncatedSVD

*class* `sklearn.decomposition.`**`TruncatedSVD`**(*n_components=2*, ***, *algorithm='randomized'*, *n_iter=5*, *n_oversamples=10*, *power_iteration_normalizer='auto'*, *random_state=None*, *tol=0.0*) [source]

Dimensionality reduction using truncated SVD (aka LSA).

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

**Rank of transformed matrix: 50**

## 5. Use **approximate** truncated SVD

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

`fit_transform` is a two step method:

1. Fit the `TruncatedSVD` to our matrix
2. Transform the matrix to 50 components

**`fit_transform`**(*X, y=None, **fit_params*) [source]

Fit to data, then transform it.

Fits transformer to `X` and `y` with optional parameters `fit_params` and returns a transformed version of `X`.

| | |
|---|---|
| **Parameters:** | **X :** ***array-like of shape (n_samples, n_features)***<br>Input samples.<br><br>**y :** ***array-like of shape (n_samples,) or (n_samples, n_outputs), default=None***<br>Target values (None for unsupervised transformations).<br><br>****fit_params :** ***dict***<br>Additional fit parameters. |
| **Returns:** | **X_new :** ***ndarray array of shape (n_samples, n_features_new)***<br>Transformed array. |

5. Use **approximate** truncated SVD

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

You can get the **weights of the concepts** with:

```
concept_weights = svd.singular_values_
```

## 5. Use **approximate** truncated SVD

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

You can get the **weights of the concepts** with:

```
concept_weights = svd.singular_values_
print(len(concept_weights))
 50

print(concept_weights)
[2829.1, 1624.4, 1141.0, …, 288.5]
```

## 5. Use **approximate** truncated SVD

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

You can get the **weights of the concepts** with:

```
concept_weights = svd.singular_values_
print(len(concept_weights))
 50

print(concept_weights)
[2829.1, 1624.4, 1141.0, …, 288.5]
```

Weight of concept 1

5. Use **approximate** truncated SVD

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

You can get the **weights of the concepts** with:

```
concept_weights = svd.singular_values_
print(len(concept_weights))
 50

print(concept_weights)
[2829.1, 1624.4, 1141.0, …, 288.5]
```

Weight of concept 2

5. Use **approximate** truncated SVD

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

You can get the **weights of the concepts** with:

```
concept_weights = svd.singular_values_
print(len(concept_weights))
 50

print(concept_weights)
[2829.1, 1624.4, 1141.0, …, 288.5]
```

Weight of concept 50

## 5. Use **approximate** truncated SVD

```
from sklearn.decomposition import TruncatedSVD

svd = TruncatedSVD(n_components=50, n_iter=20)
truncated_matrix = svd.fit_transform(matrix)
```

You can get the **weights of the concepts** with:

```
concept_weights = svd.singular_values_
print(len(concept_weights))
 50

print(concept_weights)
[2829.1, 1624.4, 1141.0, …, 288.5]
```

**As concept number increases, concept weight decreases**

# Analyzing concepts

```
truncated_df = pd.DataFrame(truncated_matrix)
truncated_df["book_title"] = titles

truncated_df[[0, "book_title"]].sort_values(by = 0)[:10]
truncated_df[[0, "book_title"]].sort_values(by = 0)[-10:]
```

Link the book metadata (title) to the new matrix

Be careful to not mix up IDs!

# Analyzing concepts

```
truncated_df = pd.DataFrame(truncated_matrix)
truncated_df["book_title"] = titles

truncated_df[[0, "book_title"]].sort_values(by = 0)[:10]
truncated_df[[0, "book_title"]].sort_values(by = 0)[-10:]
```

For the first concept, sort values in the first column (index 0) and get first and last sorted rows

# Analyzing concepts

```
truncated_df = pd.DataFrame(truncated_matrix)
truncated_df["book_title"] = titles

truncated_df[[1, "book_title"]].sort_values(by = 1)[:10]
truncated_df[[1, "book_title"]].sort_values(by = 1)[-10:]
```

For the second concept, sort values in the second column (index 1) and get first and last sorted rows

## Concept 1

$\boxed{2829.1}$

Weight of concept 1

| | |
|---|---|
| 279.13 | Harry Potter and the Sorcerer's Stone (Harry Potter, #1) |
| 271.89 | The Hunger Games (The Hunger Games, #1) |
| 253.5 | Harry Potter and the Deathly Hallows (Harry Potter, #7) |
| 249.47 | Harry Potter and the Prisoner of Azkaban (Harry Potter, #3) |
| 248.07 | Harry Potter and the Goblet of Fire (Harry Potter, #4) |
| 247.45 | Harry Potter and the Chamber of Secrets (Harry Potter, #2) |
| 244.24 | Harry Potter and the Half-Blood Prince (Harry Potter, #6) |
| 240.69 | Harry Potter and the Order of the Phoenix (Harry Potter, #5) |
| 236.33 | Catching Fire (The Hunger Games, #2) |
| 226.66 | To Kill a Mockingbird |
| ... | |
| 0.32 | فلتغفري |
| 0.29 | قواعد العشق الأربعون: رواية عن جلال الدين الرومي |
| 0.28 | Perahu Kertas |
| 0.26 | في قلبي أنثى عبرية |
| 0.26 | أحببتك أكثر مما ينبغي |
| 0.25 | 1919 |
| 0.24 | 28 حرف |
| 0.22 | في ديسمبر تنتهي كل الأحلام |
| 0.15 | Kürk Mantolu Madonna |
| 0.14 | الخيميائي |

**Concept 1**
2829.1

Concept 1 values by book

| | |
|---|---|
| 279.13 | Harry Potter and the Sorcerer's Stone (Harry Potter, #1) |
| 271.89 | The Hunger Games (The Hunger Games, #1) |
| 253.5 | Harry Potter and the Deathly Hallows (Harry Potter, #7) |
| 249.47 | Harry Potter and the Prisoner of Azkaban (Harry Potter, #3) |
| 248.07 | Harry Potter and the Goblet of Fire (Harry Potter, #4) |
| 247.45 | Harry Potter and the Chamber of Secrets (Harry Potter, #2) |
| 244.24 | Harry Potter and the Half-Blood Prince (Harry Potter, #6) |
| 240.69 | Harry Potter and the Order of the Phoenix (Harry Potter, #5) |
| 236.33 | Catching Fire (The Hunger Games, #2) |
| 226.66 | To Kill a Mockingbird |
| ... | |
| 0.32 | فلتغفري |
| 0.29 | قواعد العشق الأربعون: رواية عن جلال الدين الرومي |
| 0.28 | Perahu Kertas |
| 0.26 | في قلبي أنثى عبرية |
| 0.26 | أحببتك أكثر مما ينبغي |
| 0.25 | 1919 |
| 0.24 | 28 حرف |
| 0.22 | في ديسمبر تنتهي كل الأحلام |
| 0.15 | Kürk Mantolu Madonna |
| 0.14 | الخيميائي |

# Concept 1

2829.1

Any guesses on what concept 1 represents?

| | |
|---|---|
| 279.13 | Harry Potter and the Sorcerer's Stone (Harry Potter, #1) |
| 271.89 | The Hunger Games (The Hunger Games, #1) |
| 253.5 | Harry Potter and the Deathly Hallows (Harry Potter, #7) |
| 249.47 | Harry Potter and the Prisoner of Azkaban (Harry Potter, #3) |
| 248.07 | Harry Potter and the Goblet of Fire (Harry Potter, #4) |
| 247.45 | Harry Potter and the Chamber of Secrets (Harry Potter, #2) |
| 244.24 | Harry Potter and the Half-Blood Prince (Harry Potter, #6) |
| 240.69 | Harry Potter and the Order of the Phoenix (Harry Potter, #5) |
| 236.33 | Catching Fire (The Hunger Games, #2) |
| 226.66 | To Kill a Mockingbird |
| ... | |
| 0.32 | فلتغفري |
| 0.29 | قواعد العشق الأربعون: رواية عن جلال الدين الرومي |
| 0.28 | Perahu Kertas |
| 0.26 | في قلبي أنثى عبرية |
| 0.26 | أحببتك أكثر مما ينبغي |
| 0.25 | 1919 |
| 0.24 | حرف 28 |
| 0.22 | في ديسمبر تنتهي كل الأحلام |
| 0.15 | Kürk Mantolu Madonna |
| 0.14 | الخيميائي |

# Concept 1
2829.1

Any guesses on what concept 1 represents?

Probably a mixture of popularity + language

279.13 Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
271.89 The Hunger Games (The Hunger Games, #1)
253.5 Harry Potter and the Deathly Hallows (Harry Potter, #7)
249.47 Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)
248.07 Harry Potter and the Goblet of Fire (Harry Potter, #4)
247.45 Harry Potter and the Chamber of Secrets (Harry Potter, #2)
244.24 Harry Potter and the Half-Blood Prince (Harry Potter, #6)
240.69 Harry Potter and the Order of the Phoenix (Harry Potter, #5)
236.33 Catching Fire (The Hunger Games, #2)
226.66 To Kill a Mockingbird
...
0.32 فلتغفري
0.29 قواعد العشق الأربعون: رواية عن جلال الدين الرومي
0.28 Perahu Kertas
0.26 في قلبي أنثى عبرية
0.26 أحببتك أكثر مما ينبغي
0.25 1919
0.24 حرف 28
0.22 في ديسمبر تنتهي كل الأحلام
0.15 Kürk Mantolu Madonna
0.14 الخيميائي

# Concept 2
# 1624.4

What is captured by concept 2?

| | |
|---|---|
| -73.03 | 1984 |
| -65.91 | Animal Farm |
| -65.36 | To Kill a Mockingbird |
| -63.09 | The Handmaid's Tale |
| -61.96 | The Adventures of Huckleberry Finn |
| -61.47 | The Great Gatsby |
| -61.46 | Brave New World |
| -60.13 | Where the Wild Things Are |
| -60.02 | The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide #1) |
| -59.79 | Charlotte's Web |
| ... | |
| 97.17 | One Foot in the Grave (Night Huntress, #2) |
| 101.02 | Fifty Shades Darker (Fifty Shades, #2) |
| 101.99 | Fifty Shades of Grey (Fifty Shades, #1) |
| 103.33 | Lover Revealed (Black Dagger Brotherhood, #4) |
| 103.96 | Halfway to the Grave (Night Huntress, #1) |
| 105.18 | Lover Unbound (Black Dagger Brotherhood, #5) |
| 106.55 | Beautiful Disaster (Beautiful, #1) |
| 117.95 | Lover Eternal (Black Dagger Brotherhood, #2) |
| 119.25 | Lover Awakened (Black Dagger Brotherhood, #3) |
| 124.89 | Dark Lover (Black Dagger Brotherhood, #1) |

## Concept 2
1624.4

**What is captured by concept 2?**

**Maybe classics/assigned in school and dark/fantasy romance novels**

| | |
|---|---|
| -73.03 | 1984 |
| -65.91 | Animal Farm |
| -65.36 | To Kill a Mockingbird |
| -63.09 | The Handmaid's Tale |
| -61.96 | The Adventures of Huckleberry Finn |
| -61.47 | The Great Gatsby |
| -61.46 | Brave New World |
| -60.13 | Where the Wild Things Are |
| -60.02 | The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide #1) |
| -59.79 | Charlotte's Web |
| ... | |
| 97.17 | One Foot in the Grave (Night Huntress, #2) |
| 101.02 | Fifty Shades Darker (Fifty Shades, #2) |
| 101.99 | Fifty Shades of Grey (Fifty Shades, #1) |
| 103.33 | Lover Revealed (Black Dagger Brotherhood, #4) |
| 103.96 | Halfway to the Grave (Night Huntress, #1) |
| 105.18 | Lover Unbound (Black Dagger Brotherhood, #5) |
| 106.55 | Beautiful Disaster (Beautiful, #1) |
| 117.95 | Lover Eternal (Black Dagger Brotherhood, #2) |
| 119.25 | Lover Awakened (Black Dagger Brotherhood, #3) |
| 124.89 | Dark Lover (Black Dagger Brotherhood, #1) |

# Concept 3
1141.9

**Many series grouped together**

**(Remember, this is all based on user data!)**

| | |
|---|---|
| -80.98 | Beautiful Disaster (Beautiful, #1) |
| -75.40 | Hopeless (Hopeless, #1) |
| -74.79 | Fallen Too Far (Rosemary Beach, #1; Too Far, #1) |
| -69.88 | Never Too Far (Rosemary Beach, #2; Too Far, #2) |
| -69.25 | The Fault in Our Stars |
| -68.37 | Slammed (Slammed, #1) |
| -67.58 | Real (Real, #1) |
| -65.18 | Walking Disaster (Beautiful, #2) |
| -64.50 | Rule (Marked Men, #1) |
| -64.32 | Wait for You (Wait for You, #1) |
| ... | |
| 67.02 | Magic Bleeds (Kate Daniels, #4) |
| 67.66 | Magic Strikes (Kate Daniels, #3) |
| 67.75 | Bone Crossed (Mercy Thompson, #4) |
| 68.40 | Magic Bites (Kate Daniels, #1) |
| 70.51 | Cry Wolf (Alpha & Omega, #1) |
| 74.58 | Dead Witch Walking (The Hollows, #1) |
| 76.88 | Silver Borne (Mercy Thompson, #5) |
| 82.34 | Blood Bound (Mercy Thompson, #2) |
| 83.55 | Moon Called (Mercy Thompson, #1) |
| 83.56 | Iron Kissed (Mercy Thompson, #3) |

**Concept 4**
1073.0

| | |
|---|---|
| -47.25 | Motorcycle Man (Dream Man, #4) |
| -46.24 | Mystery Man (Dream Man, #1) |
| -45.87 | Own the Wind (Chaos, #1) |
| -45.86 | Fifty Shades of Grey (Fifty Shades, #1) |
| -45.12 | Knight (Unfinished Hero, #1) |
| -43.31 | Fifty Shades Darker (Fifty Shades, #2) |
| -43.15 | Reflected in You (Crossfire, #2) |
| -43.14 | Real (Real, #1) |
| -42.70 | Sweet Dreams (Colorado Mountain, #2) |
| -42.27 | Law Man (Dream Man, #3) |
| ... | |
| 62.58 | The Fault in Our Stars |
| 63.04 | Legend (Legend, #1) |
| 63.17 | Insurgent (Divergent, #2) |
| 64.48 | Cress (The Lunar Chronicles, #3) |
| 65.71 | Daughter of Smoke & Bone (Daughter of Smoke & Bone, #1) |
| 66.34 | Divergent (Divergent, #1) |
| 69.35 | Scarlet (The Lunar Chronicles, #2) |
| 70.44 | Matched (Matched, #1) |
| 74.76 | Graceling (Graceling Realm, #1) |
| 81.43 | Cinder (The Lunar Chronicles, #1) |

# Concept 5
893.7

| | |
|---|---|
| -59.18 | Ender's Game (Ender's Saga, #1) |
| -58.54 | Good Omens |
| -56.01 | Watchmen |
| -54.58 | The Hitchhiker's Guide to the Galaxy (Hitchhiker's Guide, #1) |
| -53.42 | Dune (Dune Chronicles #1) |
| -52.96 | American Gods (American Gods, #1) |
| -51.27 | The Name of the Wind (The Kingkiller Chronicle, #1) |
| -49.18 | A Game of Thrones (A Song of Ice and Fire, #1) |
| -47.70 | Neverwhere |
| -47.51 | The Eye of the World (Wheel of Time, #1) |
| ... | |
| 42.63 | Seven Up (Stephanie Plum, #7) |
| 43.90 | Hot Six (Stephanie Plum, #6) |
| 44.33 | Gone with the Wind |
| 44.33 | My Sister's Keeper |
| 44.35 | Four to Score (Stephanie Plum, #4) |
| 45.10 | Three to Get Deadly (Stephanie Plum, #3) |
| 46.22 | Water for Elephants |
| 48.66 | The Secret Life of Bees |
| 51.41 | One for the Money (Stephanie Plum, #1) |
| 63.21 | The Help |

# Concept 50
288.5

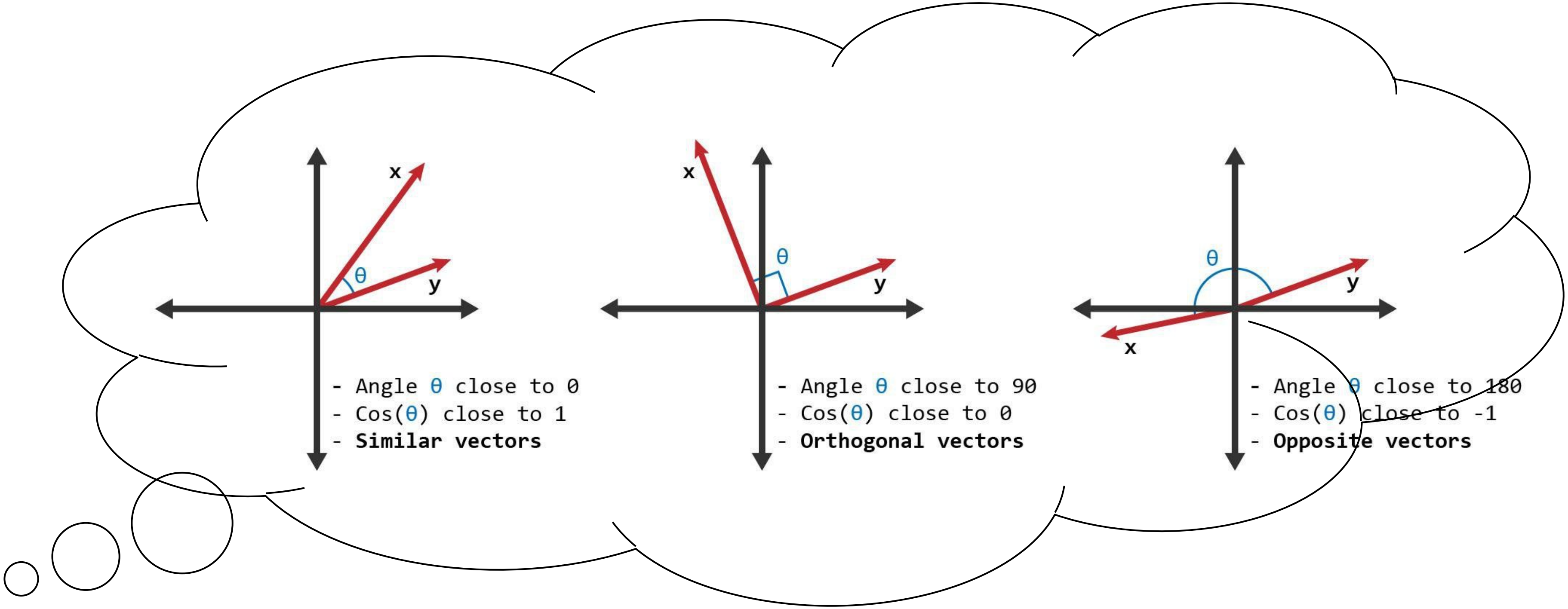
| | |
|---|---|
| -31.17 | Catching Fire (The Hunger Games, #2) |
| -29.40 | Mockingjay (The Hunger Games, #3) |
| -29.11 | The Hunger Games (The Hunger Games, #1) |
| -26.08 | Shadow Kiss (Vampire Academy, #3) |
| -25.99 | Frostbite (Vampire Academy, #2) |
| -25.04 | Blood Promise (Vampire Academy, #4) |
| -24.89 | Caressed by Ice (Psy-Changeling #3) |
| -24.39 | Vampire Academy (Vampire Academy, #1) |
| -23.93 | Visions of Heat (Psy-Changeling #2) |
| -23.73 | Slave to Sensation (Psy-Changeling #1) |
| ... | |
| 29.72 | Ever After (The Hollows, #11) |
| 30.30 | A Perfect Blood (The Hollows, #10) |
| 33.67 | Black Magic Sanction (The Hollows, #8) |
| 33.81 | The Good, the Bad, and the Undead (The Hollows, #2) |
| 33.93 | A Fistful of Charms (The Hollows, #4) |
| 34.04 | Pale Demon (The Hollows, #9) |
| 34.95 | Every Which Way But Dead (The Hollows, #3) |
| 35.00 | White Witch, Black Curse (The Hollows, #7) |
| 36.19 | For a Few Demons More (The Hollows, #5) |
| 36.26 | The Outlaw Demon Wails (The Hollows, #6) |

**What if we know that a user likes *The Fault in Our Stars* and we want to recommend them similar books?**

## What if we know that a user likes *The Fault in Our Stars* and we want to recommend them similar books?

**We can use our approximated matrix (5000 books x 50 concepts) with cosine similarity**

# Cosine similarity




https://www.learndatasci.com/glossary/cosine-similarity/

# Cosine similarity

$$\text{similarity}(x, y) = \frac{x^T y}{\|x\| \, \|y\|}$$

# Cosine similarity

$$\text{similarity}(x, y) = \frac{\boxed{x^T y}}{\|x\| \, \|y\|}$$

**Get similarity of a query vector to all vectors, divide by lengths, and sort**

query

| |
|---|
| 1 |
| 2 |

| | |
|---|---|
| 4 | 5 |
| 6 | 0 |
| 1 | 2 |

| |
|---|
| ? |
| ? |
| ? |

# Cosine similarity

$$\text{similarity}(x, y) = \frac{x^T y}{\|x\| \boxed{\|y\|}}$$

"length" of y = sqrt($y^T y$)

# Cosine similarity

$$\text{similarity}(x, y) = \frac{x^T y}{\|x\| \, \|y\|}$$

What is the cosine similarity of a vector with itself?

# Cosine similarity

$$\text{similarity}(x, y) = \frac{x^T y}{\|x\| \, \|y\|}$$

**What is the cosine similarity of a vector with itself?**

**$\cos(0) = 1.0$**

# Searching with cosine similarity

*The Fault in Our Stars* book ID = 45

Calculate cosine similarity between TFIOS and every other book (row)

```
book_id = 45

inner_products = truncated_matrix.dot(truncated_matrix[ book_id,: ])
lengths = np.linalg.norm(truncated_matrix, axis=1)
cosine_sims = inner_products / (lengths * lengths[book_id])

title_scores = sorted(zip(cosine_sims, titles))
```

# Searching with cosine similarity

*The Fault in Our Stars* book ID = 45

Calculate cosine similarity between TFIOS and every other book (row)

```
book_id = 45

inner_products = truncated_matrix.dot(truncated_matrix[ book_id,: ])
lengths = np.linalg.norm(truncated_matrix, axis=1)
cosine_sims = inner_products / (lengths * lengths[book_id])

title_scores = sorted(zip(cosine_sims, titles))
```

**We assess cosine similarity across all 50 concepts!**

# Closest books by cosine similarity

| | |
|---|---|
| 1.00 | The Fault in Our Stars |
| 0.95 | Looking for Alaska |
| 0.95 | Eleanor & Park |
| 0.92 | Paper Towns |
| 0.92 | We Were Liars |
| 0.92 | The Perks of Being a Wallflower |
| 0.91 | If I Stay (If I Stay, #1) |
| 0.91 | Fangirl |
| 0.90 | Thirteen Reasons Why |
| 0.89 | The Book Thief |

# Closest books by cosine similarity

| | |
|---|---|
| $\boxed{1.00}$ | The Fault in Our Stars |
| 0.95 | Looking for Alaska |
| 0.95 | Eleanor & Park |
| 0.92 | Paper Towns |
| 0.92 | We Were Liars |
| 0.92 | The Perks of Being a Wallflower |
| 0.91 | If I Stay (If I Stay, #1) |
| 0.91 | Fangirl |
| 0.90 | Thirteen Reasons Why |
| 0.89 | The Book Thief |

**What is the cosine of a vector with itself?**

**cos(0) = 1.0**