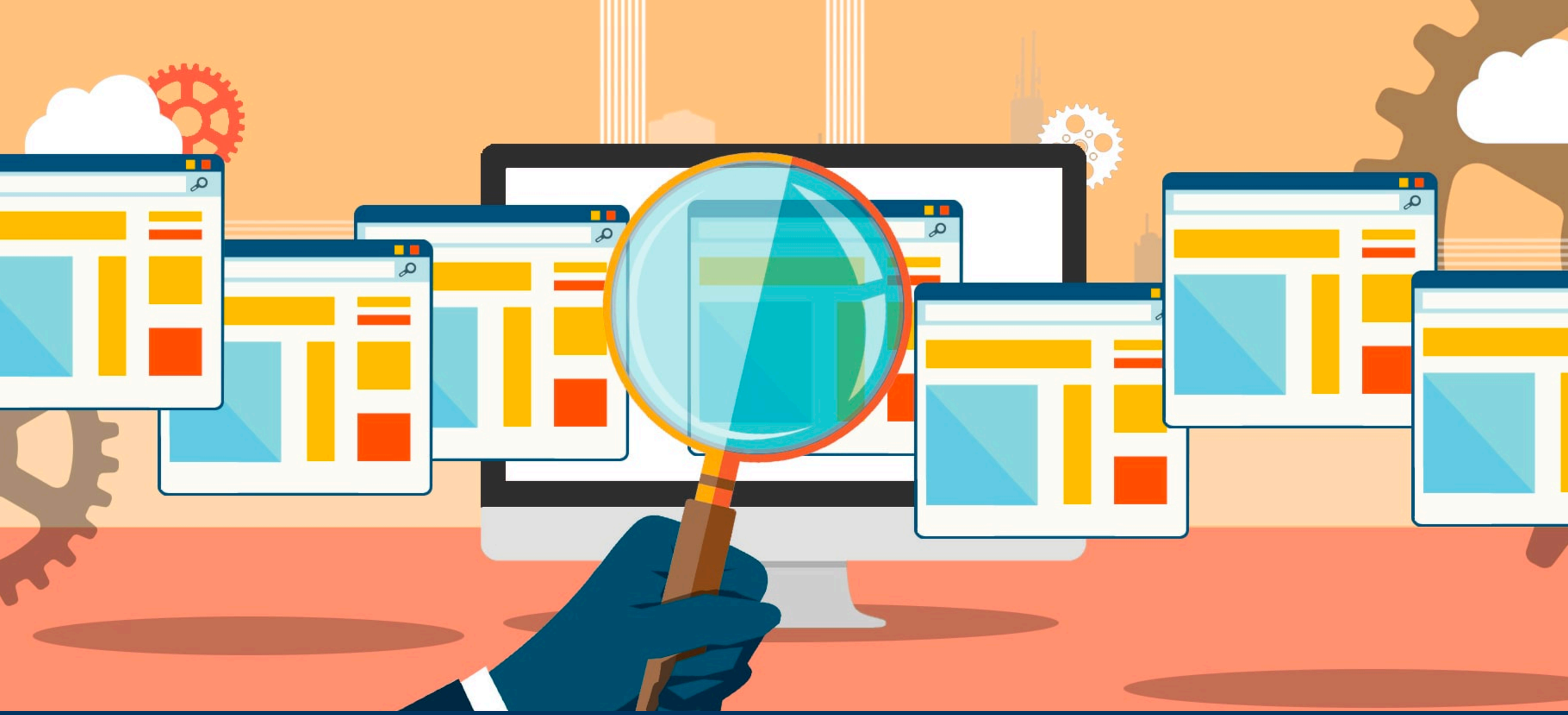


Distance Metrics

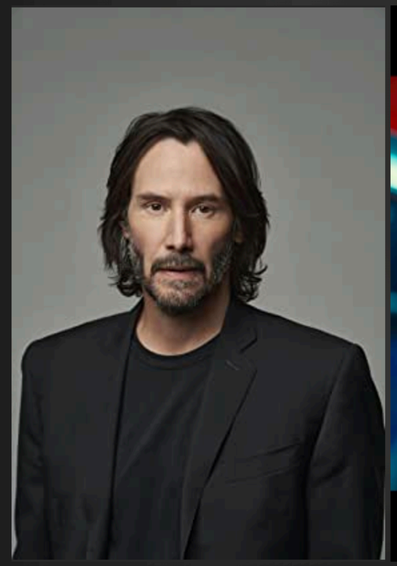
INST414 - Data Science Techniques

Motivating Question: What does it mean for two actors to be similar?

Keanu Reeves

Actor · Producer · Additional Crew

IMDbPro STARMETER Top 500 ▲ 6



99+ VIDEOS

99+ PHOTOS

Expand below

Actor

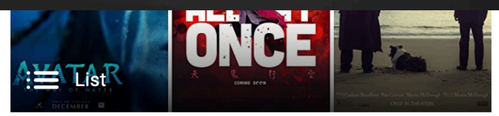
Upcoming · 7

Previous · 104

DC League of Super-Pets
Batman (voice)
2022

The Matrix Resurrections
Neo · Thomas Anderson
2021

The Matrix Awakens: An Unreal Engine 5 Experience
Neo · Keanu Reeves
Video Game
2021



Oscars 2023: Where to Watch the Best Picture Nominees

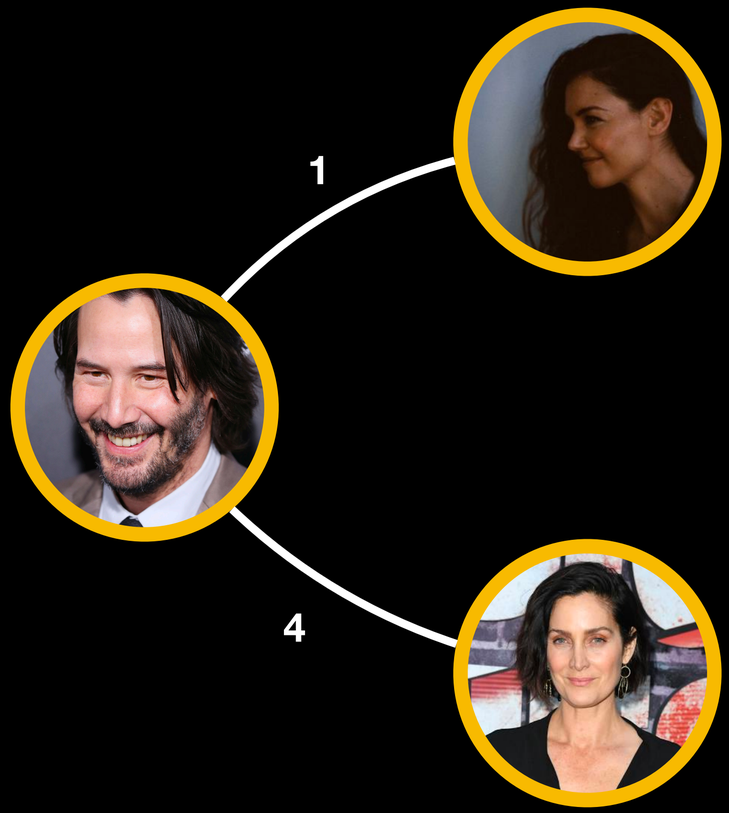
See the list

User lists >

Related lists from IMDb users

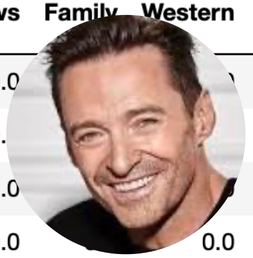
+ Create a list

Celebrity


1
4

```
actor_genre_df.head(20)
```

	Comedy	Fantasy	Romance	Action	Crime	Adventure	Mystery	Thriller	Drama	Biography	...	Sport		News	Family	Western	Short
nm0000212	16.0	3.0	16.0	5.0	4.0	2.0	5.0	3.0	16.0	2.0	...	0.0	0.0	0.0			0.0
nm0413168	8.0	3.0	6.0	14.0	6.0	11.0	5.0	2.0	13.0	5.0	...	0.0	0.0	0.			0.0
nm0000630	10.0	2.0	6.0	4.0	1.0	2.0	2.0	4.0	17.0	6.0	...	4.0	1.0	1.0			0.0
nm0005227	12.0	1.0	3.0	2.0	0.0	3.0	0.0	1.0	5.0	1.0	...	1.0	0.0	0.0		0.0	0.0
nm0697338	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm1300519	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0940707	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0625977	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0792032	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0496571	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm2868805	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm2866192	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0001379	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	...	1.0	0.0	0.0	1.0	0.0	0.0
nm0462648	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0000953	6.0	0.0	0.0	1.0	3.0	0.0	0.0	2.0	9.0	7.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0001782	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0005077	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0
nm0550626	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0177016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0907480	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0



```
In [2]: pd.read_csv("Actor data.csv")
```

Out[2]:

	Actor	Film	Year	Salary	Total income	Total income_cleaned
0	Keanu Reeves	The Matrix ReloadedThe Matrix Revolutions	2003	$30,000,000	$156,000,000	15,60,00,000
1	Bruce Willis	The Sixth Sense	1999	$14,000,000	$100,000,000	10,00,00,000
2	Tom Cruise	Mission: Impossible 2	2000	NaN	$100,000,000	10,00,00,000
3	Tom Cruise	War of the Worlds	2005	NaN	$100,000,000	10,00,00,000
4	Will Smith	Men in Black 3	2012	NaN	$100,000,000	10,00,00,000
5	Robert Downey Jr.	Avengers: Infinity War	2018	NaN	$75,000,000+	7,50,00,000
6	Robert Downey Jr.	Avengers: Endgame	2019	$20,000,000	$75,000,000	7,50,00,000
7	Robert Downey Jr.	Iron Man 3	2013	NaN	$75,000,000	7,50,00,000
8	Sandra Bullock	Gravity	2013	$20,000,000	$70,000,000+	7,00,00,000
9	Tom Hanks	Forrest Gump	1994	NaN	$70,000,000	7,00,00,000
10	Tom Cruise	Mission: Impossible	1996	NaN	$70,000,000	7,00,00,000
11	Harrison Ford	Indiana Jones and the Kingdom of the Crystal S...	2008	NaN	$65,000,000	6,50,00,000
12	Jack Nicholson	Batman	1989	$6,000,000	$60,000,000	6,00,00,000
13	Leonardo DiCaprio	Inception	2010	NaN	$59,000,000	5,90,00,000
14	Johnny Depp	Pirates of the Caribbean: On Stranger Tides	2011	$35,000,000	$55,000,000	5,50,00,000
15	Robert Downey Jr.	The Avengers	2012	NaN	$50,000,000	5,00,00,000
16	Cameron Diaz	Bad Teacher	2011	NaN	$42,000,000	4,20,00,000
17	Robert Downey Jr.	Captain America: Civil War	2016	$40,000,000	$40,000,000+	4,00,00,000
18	Robert Downey Jr.	Avengers: Age of Ultron	2015	NaN	$40,000,000	4,00,00,000
19	Leonardo DiCaprio	Titanic	1997	NaN	$40,000,000	4,00,00,000
20	Tom Hanks	Saving Private Ryan	1998	NaN	$40,000,000	4,00,00,000

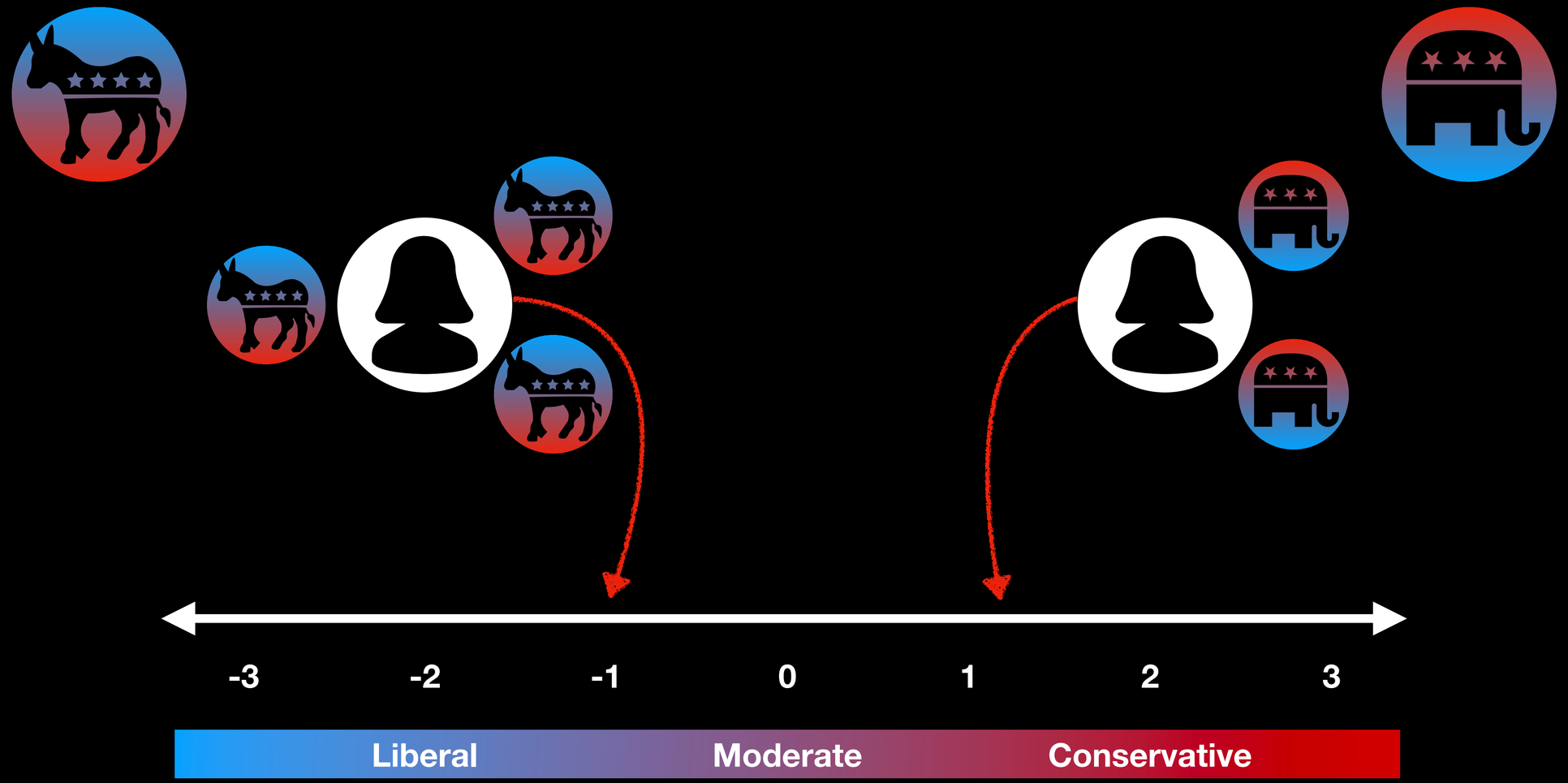
What does it mean for two politicians to be similar?

Elizabeth Warren

Lisa Murkowski

Bernie Sanders

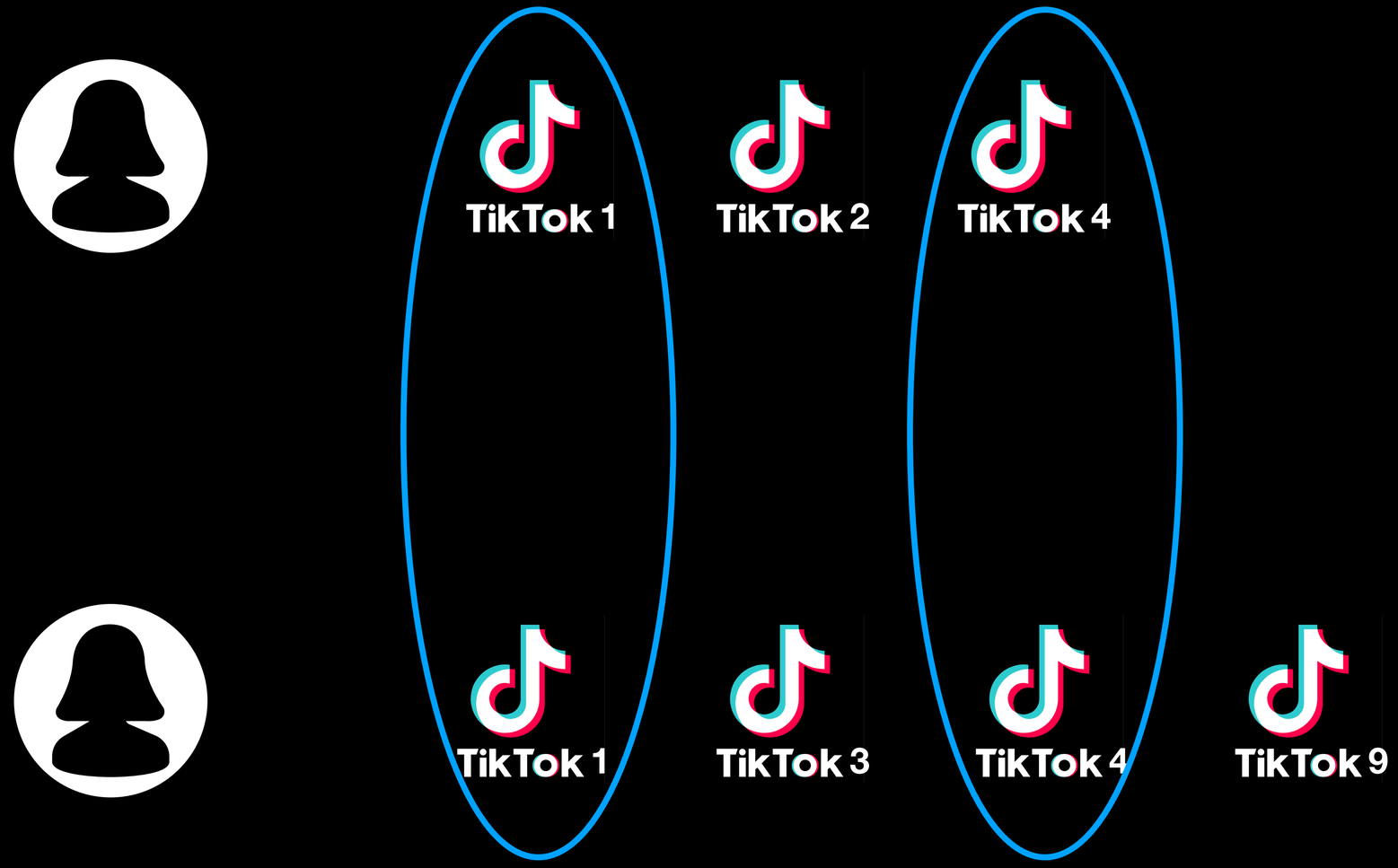
nytimes.com	reuters.com	bbc.co.uk	vox.com	HuffPost.com	breitbart.com	...	foxnews.com	npr.org
23	7	9	37	22	0	...	0	20
3	11	10	0	1	5	...	21	3
31	23	17	40	39	0	...	0	17

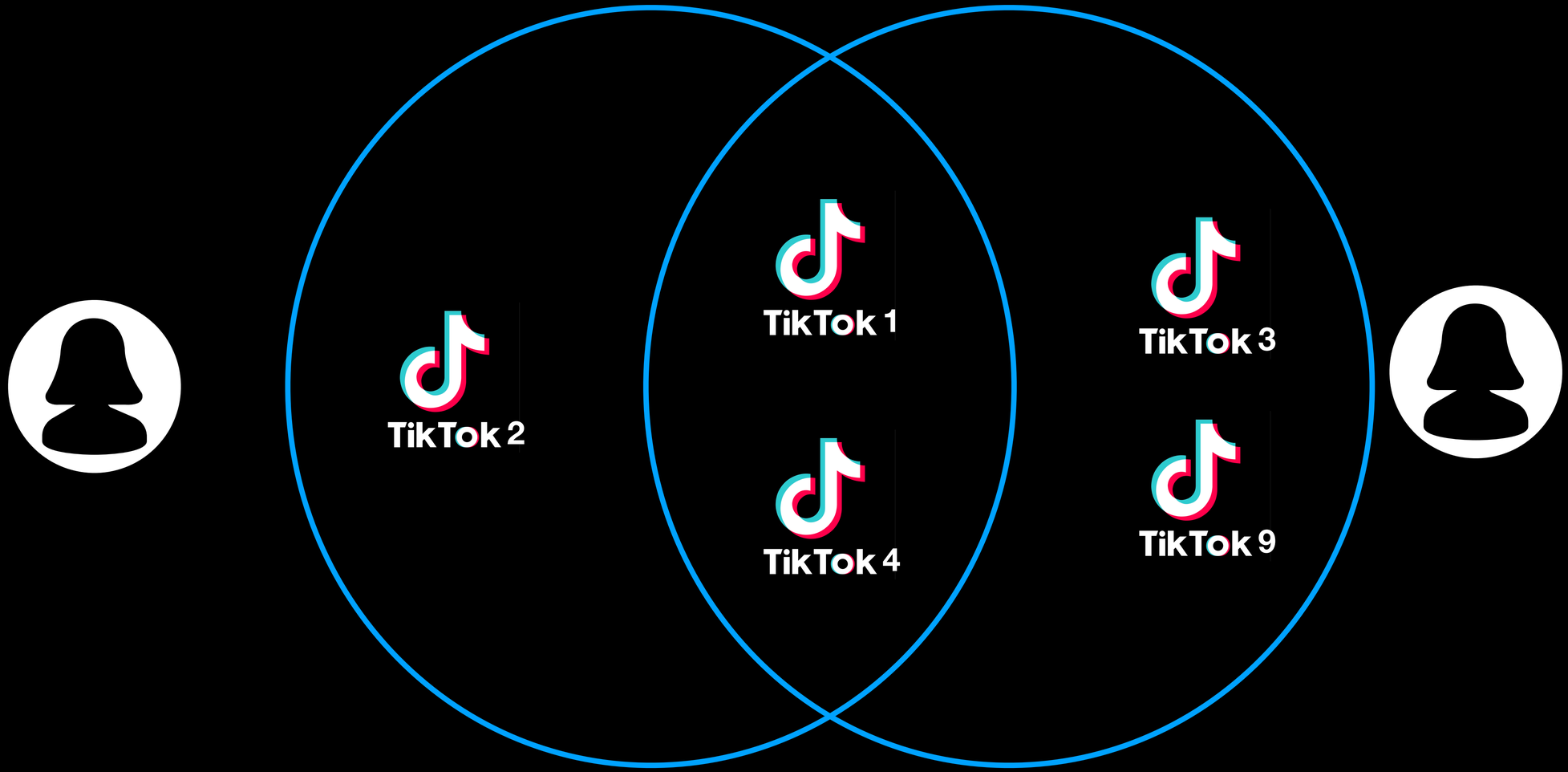

-3
-2
-1
0
1
2
3
Liberal
Moderate
Conservative

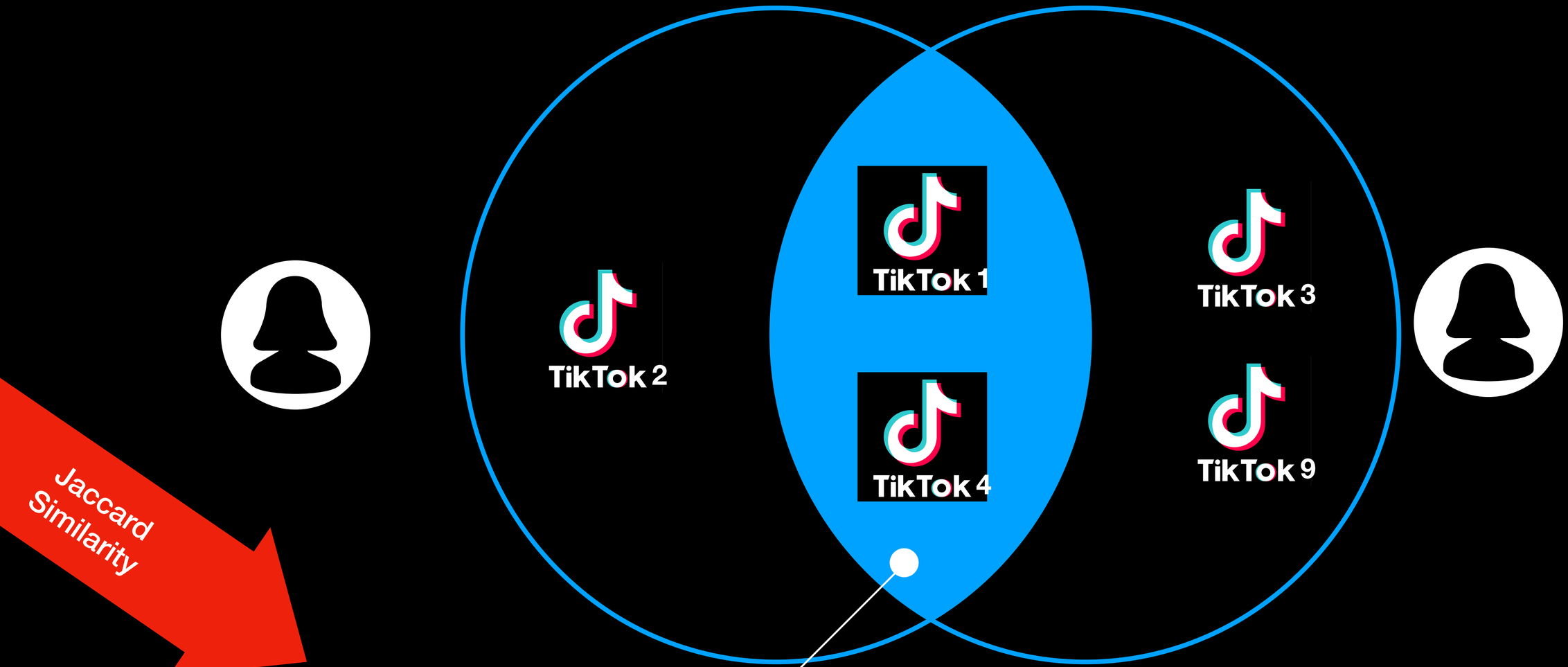
What does it mean for two people to be similar in Twitter/Instagram/TikTok?

	TikTok 1	TikTok 2	TikTok 3	TikTok 4	...	TikTok 9
	✓	✓		✓		
	✓		✓	✓		✓

How might you measure such similarity?

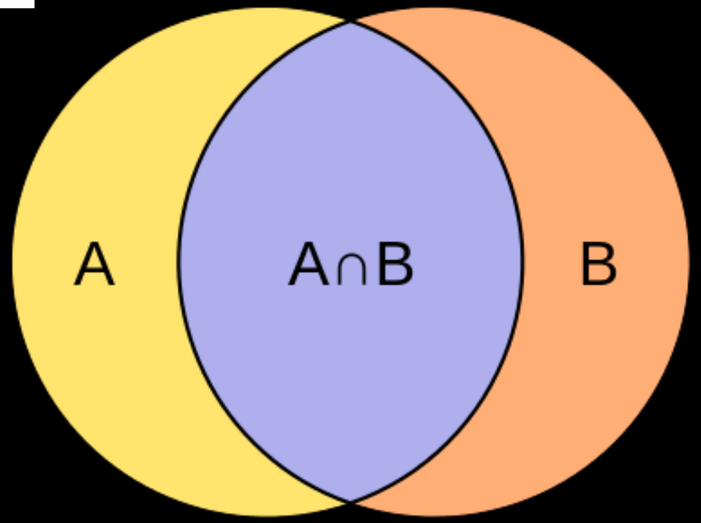

TikTok 1
TikTok 2
TikTok 4
TikTok 1
TikTok 3
TikTok 4
TikTok 9


TikTok 1
TikTok 2
TikTok 3
TikTok 4
TikTok 9


TikTok 1
TikTok 2
TikTok 3
TikTok 4
TikTok 9
Jaccard Similarity
Overlap = 2
All Watched Videos = 5
Similarity = 2/5 = 0.4

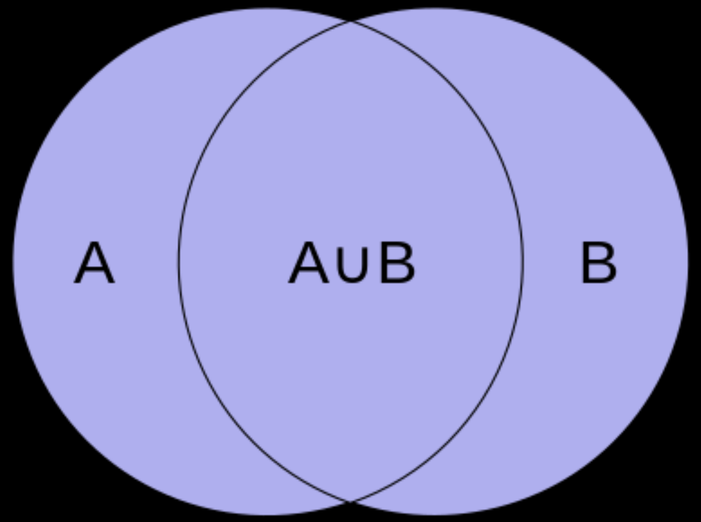
Defining Jaccard Similarity

Intersection of A and B



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Union of A and B



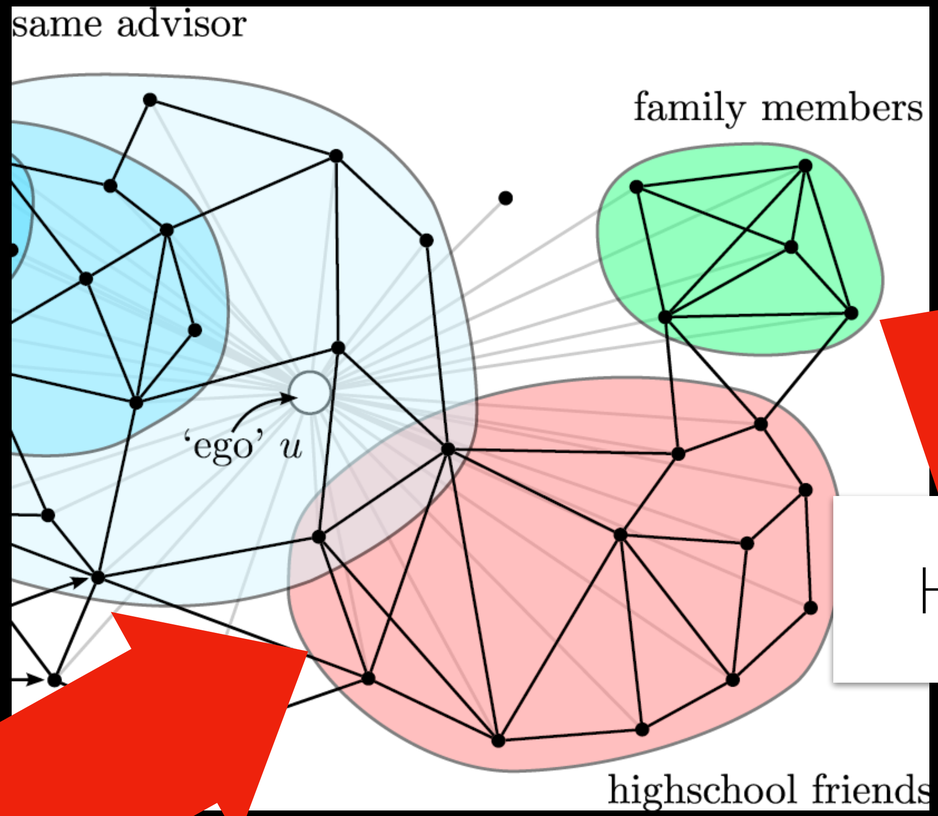
Six Core Learning Objectives

This Time

1. Collect and clean large-scale datasets
2. Articulate the math behind supervised and unsupervised techniques
3. Execute supervised and unsupervised machine learning techniques
4. Select and evaluate various types of machine learning techniques
5. Explain the results coming out of the models
6. Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach

Where are we?

2. Articulate the math behind supervised and unsupervised techniques



How similar are elements in these groups?

This Week's Learning Objectives

Define "homophily"

Give examples of at least three distance metrics

Define shingling and its utility in evaluating similarity

This Week's Learning Objectives

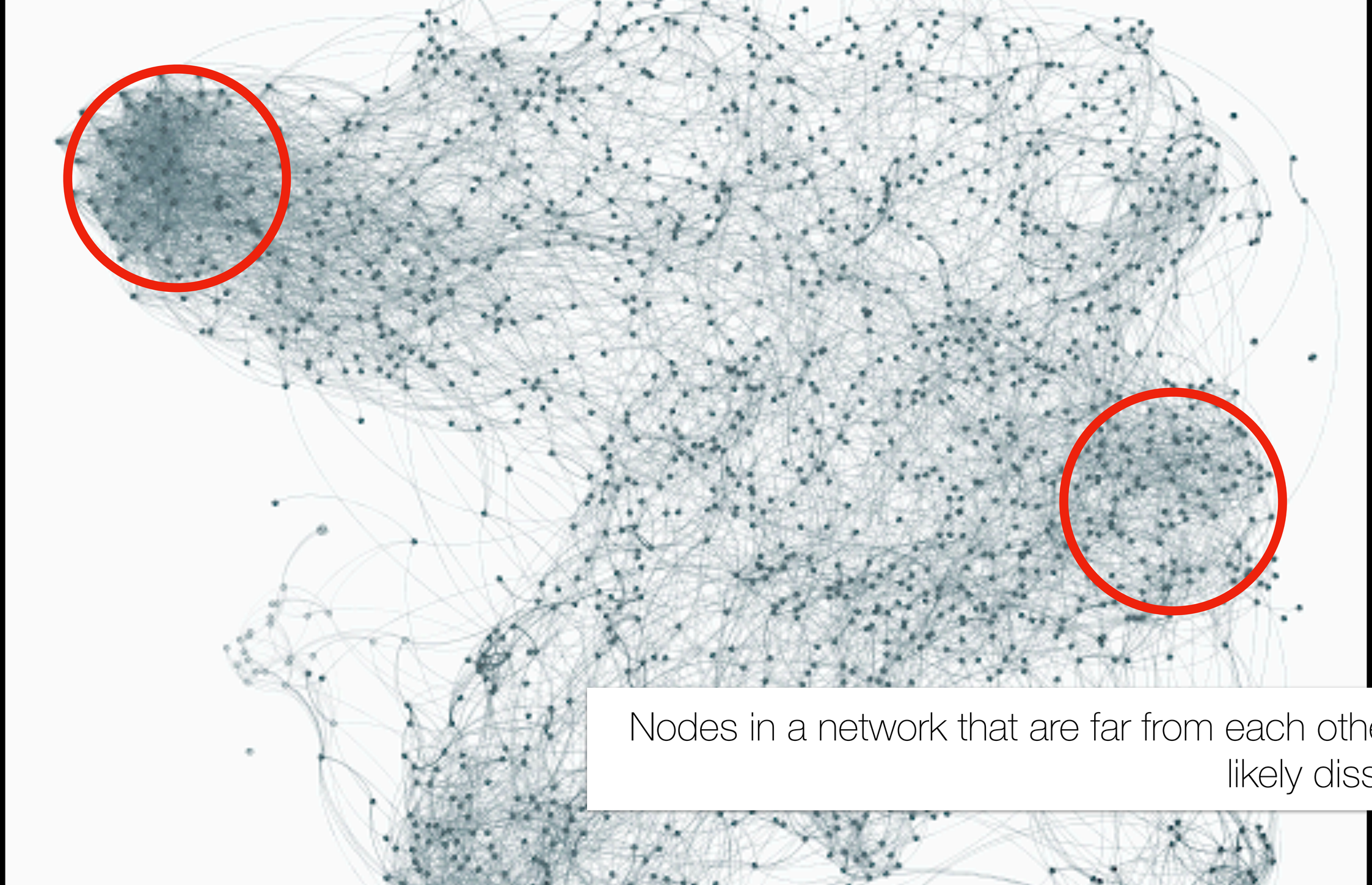
Define "homophily"

Give examples of at least three distance metrics

Define shingling and its utility in evaluating similarity

Motivation: How do can we find similar items in a dataset?

A network of products on Amazon, with edges denoting two products were purchased together



Nodes in a network that are far from each other are likely dissimilar

Homophily

18 languages

Article Talk Read Edit View history

From Wikipedia, the free encyclopedia

Contents [hide]

Not to be confused with Homophile.

Homophily (from Ancient Greek ὁμός *(homós)* 'same, common', and φιλία *(philía)* 'friendship, love') is a concept in sociology describing the tendency of individuals to associate and bond with similar others, as in the proverb "*birds of a feather flock together*".[1] The presence of homophily has been discovered in a vast array of network studies: over 100 studies have observed homophily in some form or another, and they establish that similarity is associated with connection.[2] The categories on which homophily occurs include age, gender, class, and organizational role.[3]

The opposite of homophily is heterophily or intermingling.[4] Individuals in homophilic relationships share common characteristics (beliefs, values, education, etc.) that make communication and relationship formation easier. Homophily between mated pairs in animals has been extensively studied in the field of evolutionary biology, where it is known as *assortative mating*. Homophily between mated pairs is common within natural animal mating populations.[5]

Homophily has a variety of consequences for social and economic outcomes.[6]

Types and dimensions [edit]

Baseline vs. inbreeding [edit]

To test the relevance of homophily, researchers have distinguished between two types:[2]

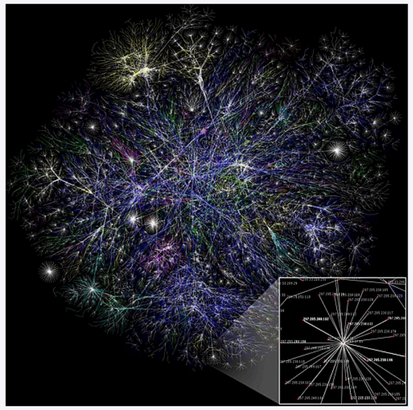
- **Baseline homophily**: simply the amount of homophily that would be expected by chance given an existing uneven distribution of people with varying characteristics; and
- **Inbreeding homophily**: the amount of homophily over and above this expected value, typically due to personal preferences and choices.[7]

Status vs. value [edit]

In their original formulation of homophily, Paul Lazarsfeld and Robert K. Merton (1954) distinguished between *status homophily* and *value homophily*; individuals with similar social status characteristics were more likely to associate with each other than by chance:[8][2]

Part of a series on

Network science



Theory

Graph · Complex network · Contagion · Small-world · Scale-free · Community structure · Percolation · Evolution · Controllability · Graph drawing · Social capital · Link analysis · Optimization · Reciprocity · Closure · **Homophily** · Transitivity · Preferential attachment · Balance theory · Network effect · Social influence

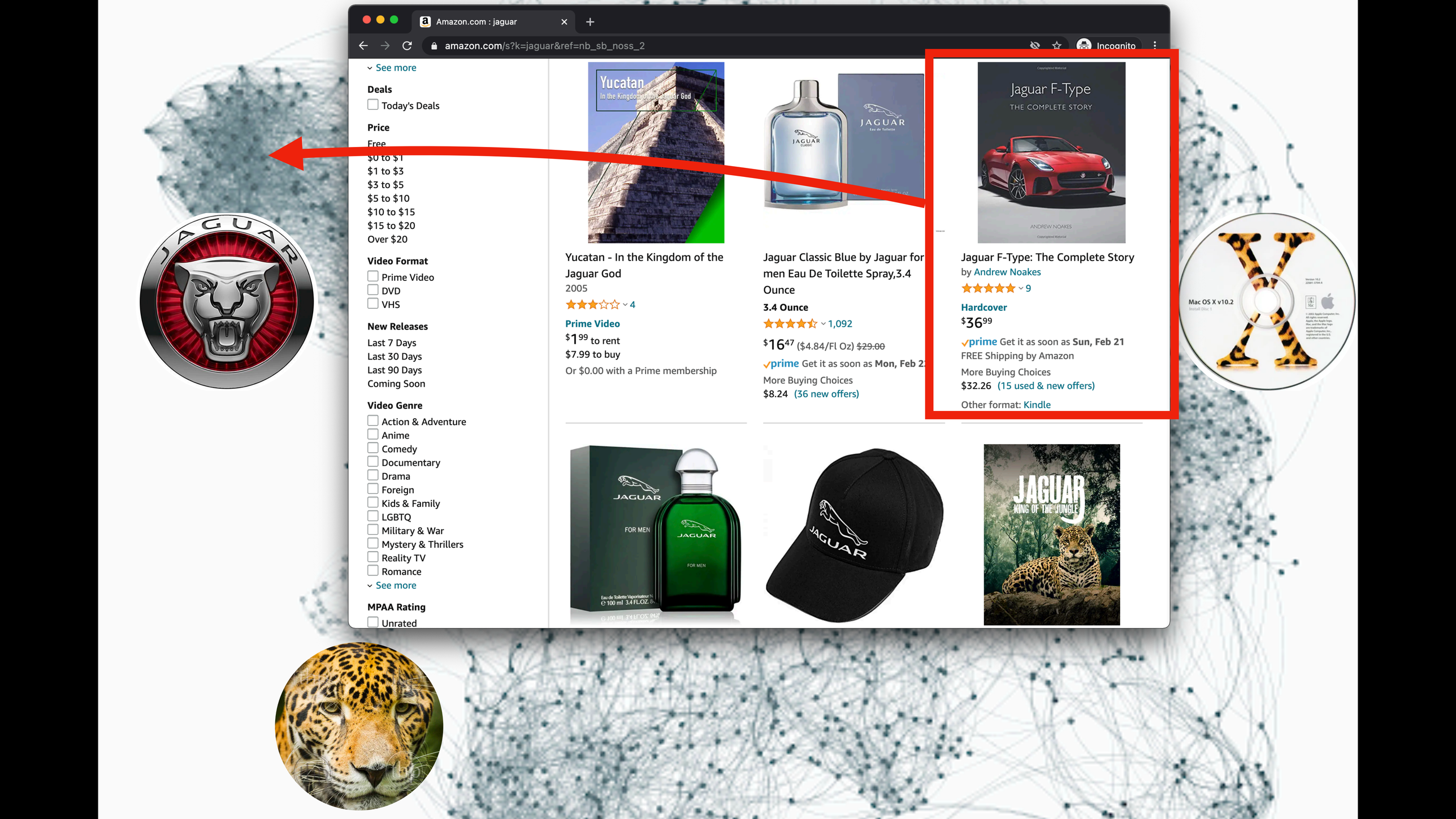
Network types

Informational (computing) · Telecommunication · Transport · Social · Scientific collaboration · Biological · Artificial neural · Interdependent · Semantic · Spatial · Dependency · Flow · on-Chip

Graphs

Features


JAGUAR
Mac OS X v10.2
Install Disc 1


Amazon.com : jaguar
amazon.com/s?k=jaguar&ref=nb_sb_noss_2
Incognito
See more
Deals
Today's Deals
Price
Free
$0 to $1
$1 to $3
$3 to $5
$5 to $10
$10 to $15
$15 to $20
Over $20
Video Format
Prime Video
DVD
VHS
New Releases
Last 7 Days
Last 30 Days
Last 90 Days
Coming Soon
Video Genre
Action & Adventure
Anime
Comedy
Documentary
Drama
Foreign
Kids & Family
LGBTQ
Military & War
Mystery & Thrillers
Reality TV
Romance
See more
MPAA Rating
Unrated
Yucatan - In the Kingdom of the Jaguar God
2005
4
Prime Video
$1.99 to rent
$7.99 to buy
Or $0.00 with a Prime membership
Jaguar Classic Blue by Jaguar for men Eu De Toilette Spray,3.4 Ounce
3.4 Ounce
1,092
$16.47 ($4.84/Fl Oz) $29.00
prime Get it as soon as Mon, Feb 2
More Buying Choices
$8.24 (36 new offers)
Jaguar F-Type: The Complete Story
by Andrew Noakes
9
Hardcover
$36.99
prime Get it as soon as Sun, Feb 21
FREE Shipping by Amazon
More Buying Choices
$32.26 (15 used & new offers)
Other format: Kindle
JAGUAR
FOR MEN
JAGUAR
KING OF THE JUNGLE
Mac OS X v10.2

Jaguar F-Type: The Complete

amazon.com/Jaguar-F-Type-Complete-Andrew-Noakes/dp/1785007319/ref=sr_1_6?dchild=1&keywords=jaguar&qid=1613595...

Incognito

inspire your little engineer! From turbochargers to electric cars - this book teaches car specifics while having fun!

94

Paperback

$6.99 prime

Paperback

$6.99 prime

Paperback

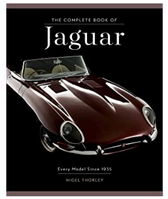
$7.99 prime

Kindle Edition

$14.95

Customers who viewed this item also viewed

Page 1 of 8



The Complete Book of Jaguar: Every Model Since 1935 (Complete Book Series)

Nigel Thorley

108

Hardcover

$27.99

Button Start Center Control Switch Knob Head Decorative Sequin Cap Cover Decal Trim fit for Jaguar XFL XE XEL XJ XJL F-PACE F-Type

286

$12.96

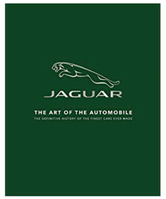
Jaguar Cars (First Gear)

› James Mann

40

Paperback

$11.28



Jaguar: The Art of the Automobile

› Zef Enault

64

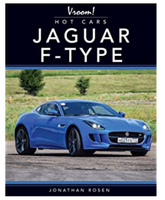
Hardcover

$45.65

DEFTEN Car Seat Gap Filler Premium PU Full Leather Seat Console Organizer, Car Seat Storage Box for Jaguar xj xf F-PACE (Black) (1…

60

$36.99



Jaguar F-TYPE (Vroom! Hot Cars)

› Jonathan Rosen

6

Library Binding

$32.79

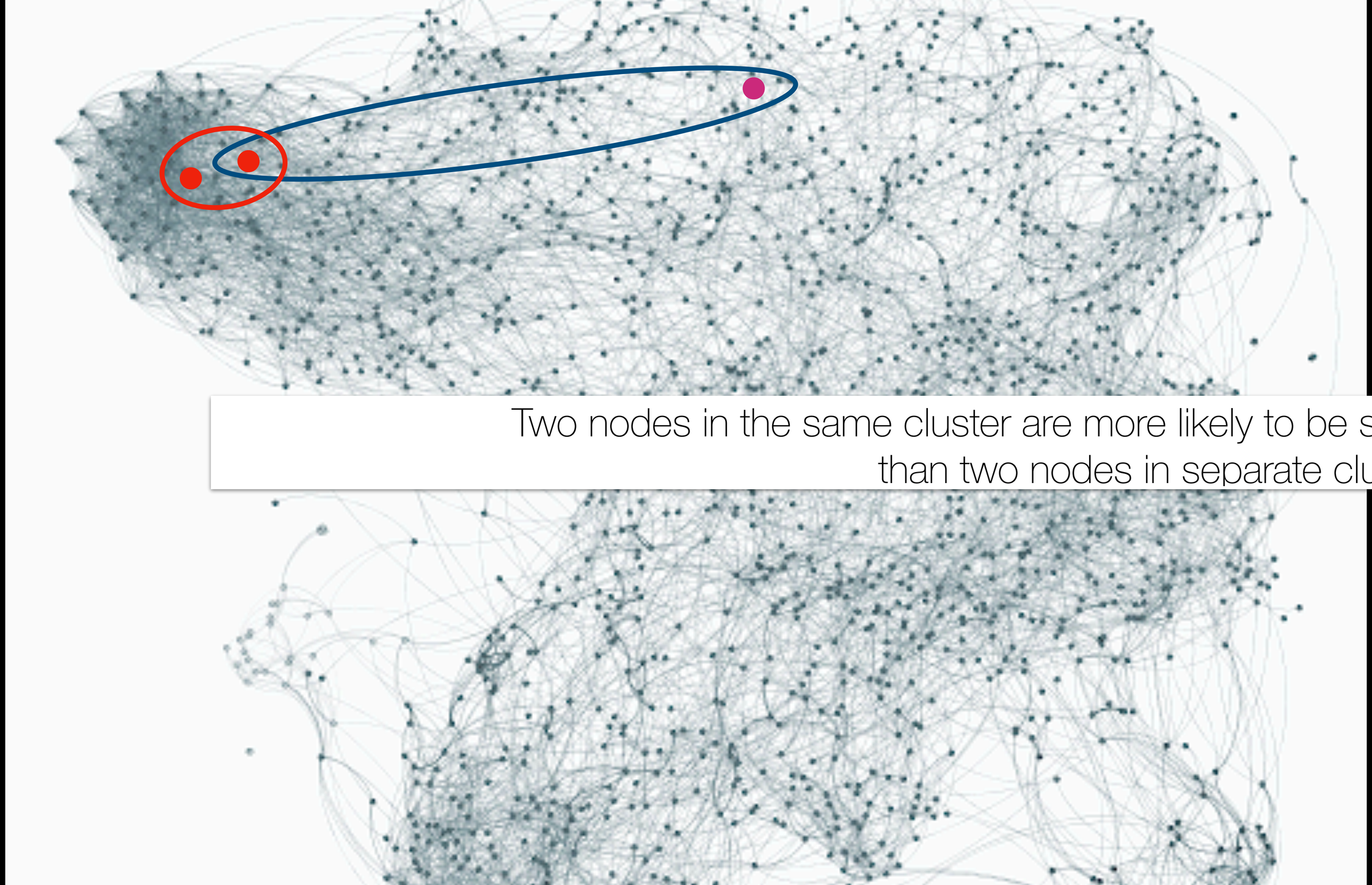
Editorial Reviews

About the Author

Andrew Noakes has been writing about cars for more than 20 years. He was trained as an automotive engineer but decided to go into motoring journalism, working for *Fast Car* magazine before launching the classic car title *Classics*. He has been freelance since 2002, writing on cars for a wide range of magazines and websites, and since 2006 has also taught motoring journalism at Coventry University. He is the chairman of the Guild of Motoring Writers in 2016-18.

Andrew is the author of more than a dozen motoring books including The Ultimate History of Aston Martin, Aston Martin DB7: The Complete Story and Aston Martin – Model By Model. He lives in Warwickshire.

Two nodes in the same group are likely similar in multiple ways



Two nodes in the same cluster are more likely to be similar
than two nodes in separate clusters

Homophily *(n)* – the tendency for people to seek out or be attracted to those who are similar to themselves

Implication: Things that are similar in one way are often similar in many ways

Motivation: How do can we find similar items in a dataset?

Groups (i.e., clusters) of entities should be similar

Corollary: Can build clusters from similar entities

How do we quantify similarity?

How do we quantify similarity?

It depends

This Week's Learning Objectives

Define "homophily"

Give examples of at least three distance metrics

Define shingling and its utility in evaluating similarity

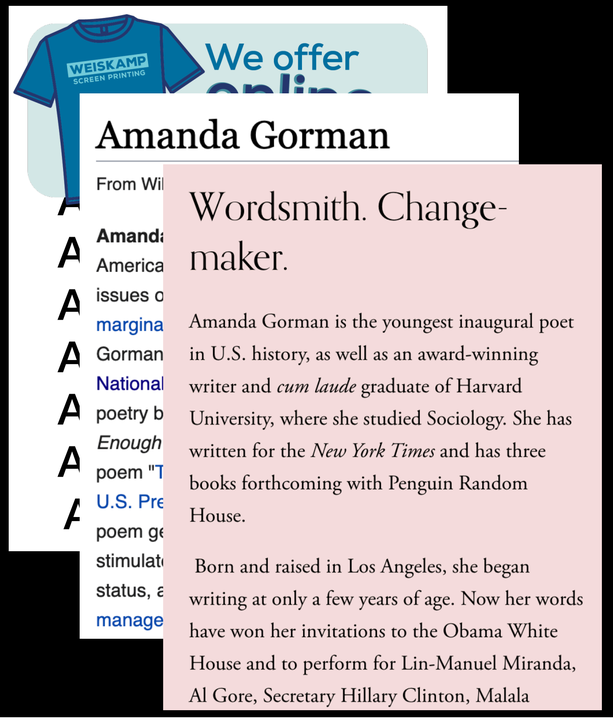
Alternate Definitions of Similarity

Adjacency Matrix *A*

	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0

Feature Matrix
|V| x D

$$\begin{bmatrix} r_A, r_B, \ldots r_E \end{bmatrix}$$



Nodes with similar sets of neighbors

Elements with “similar features"

Pages about the same topic

Adjacency Matrix *A*

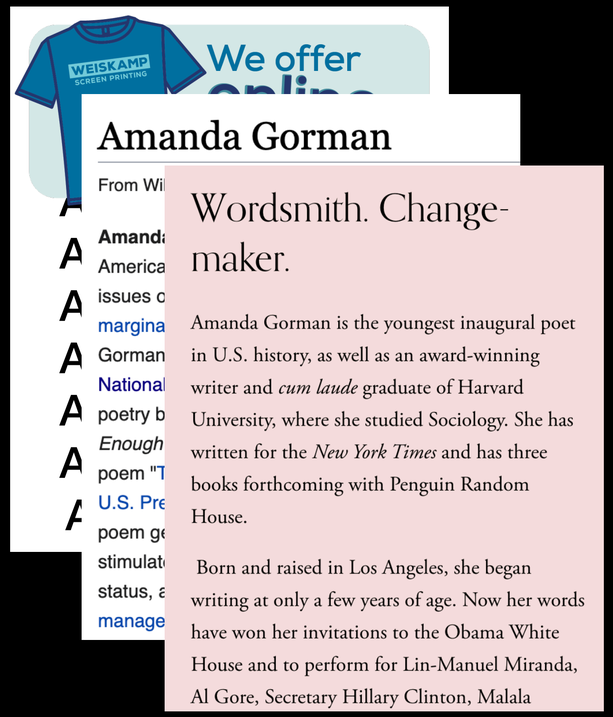
	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0

Similarity between two strings of words (i.e., documents)

Feature Matrix
|V| x D

$$\begin{bmatrix} r_A, r_B, \ldots r_E \end{bmatrix}$$

Or sets of neighbors



Similarity between two vectors of real numbers

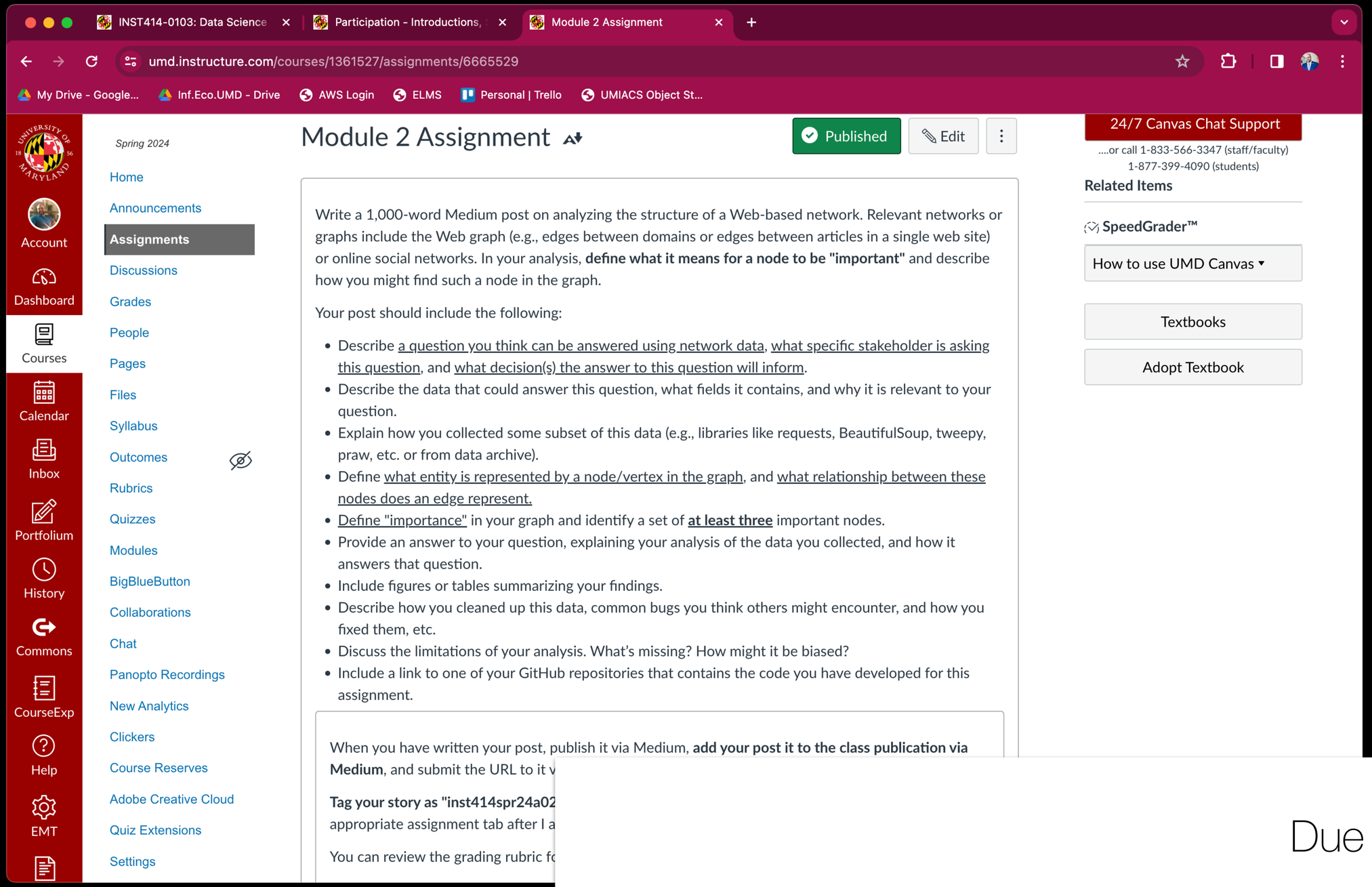
This Week's Learning Objectives

Define "homophily"

Give examples of at least three distance metrics

Define shingling and its utility in evaluating similarity

What questions do you have from Monday?



Due on Friday

Questions about the Module 2 assignment?

What is a “feature matrix”?

	d1	d2	d3	d4	d5	…	dD
x1	1	1	1	0	0		0
x2	2	1	1	1	1		0
x3	2	1	0	0	1		1
…							
xN	7	1	9	1	1		0

Standard terminology for a matrix of data

Rows tend to mean elements

	d1	d2	d3	d4	d5	…	dD
x1	1	1	1	0	0		0
x2	2	1	1	1	1		0
x3	2	1	0	0	1		1
…							
xN	7	1	9	1	1		0

Columns correspond to “features”

	d1	d2	d3	d4	d5	…	dD
x1	1	1	1	0	0		0
x2	2	1	1	1	1		0
x3	2	1	0	0	1		1
…							
xN	7	1	9	1	1		0

What is a “feature”?

Some characteristic of your data useful for analysis

AKA “feature”/“factor”/“independent variable”

“Feature engineering” is core to data science

To select Reels/TikToks to show to a user, what features are important?

This process is called “feature engineering”

“Feature engineering” is core to data science

Why do you think “feature engineering” is important?

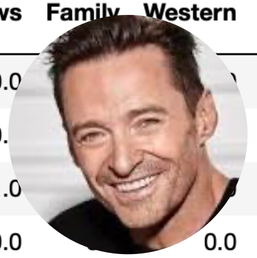
Requires your domain expertise

Want a parsimonious set of features

Higher dimensionality is bad

```
actor_genre_df.head(20)
```

	Comedy	Fantasy	Romance	Action	Crime	Adventure	Mystery	Thriller	Drama	Biography	...	Sport		News	Family	Western	Short
nm0000212	16.0	3.0	16.0	5.0	4.0	2.0	5.0	3.0	16.0	2.0	...	0.0	0.0	0.0			0.0
nm0413168	8.0	3.0	6.0	14.0	6.0	11.0	5.0	2.0	13.0	5.0	...	0.0	0.0	0.			0.0
nm0000630	10.0	2.0	6.0	4.0	1.0	2.0	2.0	4.0	17.0	6.0	...	4.0	1.0	1.0			0.0
nm0005227	12.0	1.0	3.0	2.0	0.0	3.0	0.0	1.0	5.0	1.0	...	1.0	0.0	0.0		0.0	0.0
nm0697338	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm1300519	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0940707	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0625977	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0792032	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0496571	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm2868805	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm2866192	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0001379	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0	...	1.0	0.0	0.0	1.0	0.0	0.0
nm0462648	1.0	0.0	1.0	0.0	0.0												
nm0000953	6.0	0.0	0.0	1.0	3.0												
nm0001782	0.0	0.0	0.0	1.0	0.0												
nm0005077	1.0	0.0	0.0	1.0	0.0												
nm0550626	0.0	0.0	0.0	0.0	0.0												
nm0177016	0.0	0.0	0.0	0.0	0.0												
nm0907480	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0



What are the elements here?

What are the features here?

Elizabeth Warren

Lisa Murkowski

Bernie Sanders

nytimes.com	reuters.com	bbc.co.uk	vox.com	HuffPost.com	breitbart.com	...	foxnews.com	npr.org
23	7	9	37	22	0	...	0	20
3	11	10	0	1	5	...	21	3

What are the elements here?

What are the features here?

$$y = mX + b$$

What is this equation?

X == feature matrix/independent var, y = dep var

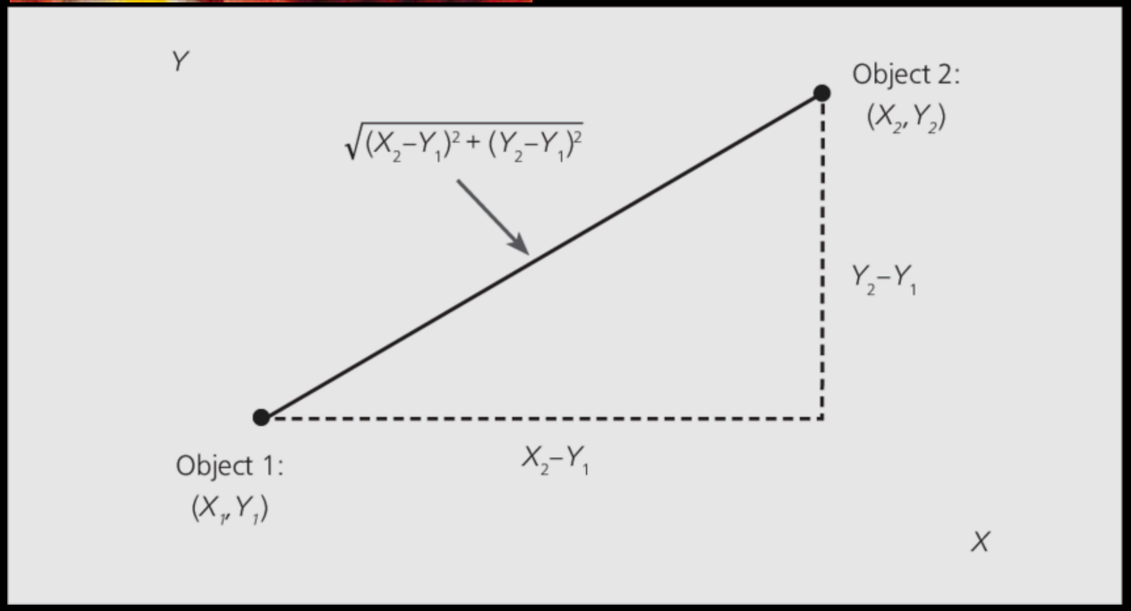
Euclidean Distance d(x,y):

Similarity: $1/d(r_A, r_B)$

Similarity —> ∞ as $d(r_A, r_B)$ —> 0

Feature Matrix
|V| x D

$$\begin{bmatrix} r_A, r_B, \ldots r_E \end{bmatrix}$$



Similarity between two vectors of real numbers

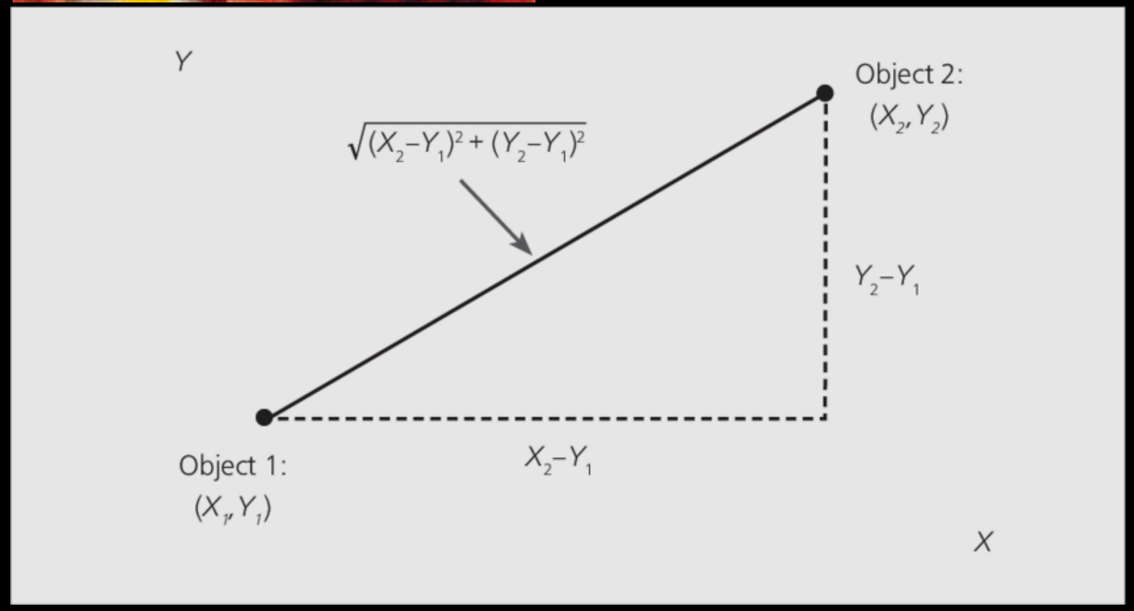
Euclidean Distance d(x,y):

Similarity: $1/d(r_A, r_B)$

Similarity $\rightarrow \infty$ as $d(r_A, r_B) \rightarrow 0$

Feature Matrix
For Models of Car

$$\begin{bmatrix} r_A, r_B, \ldots r_E \end{bmatrix}$$




```
In [1]: import pandas as pd

In [2]: pd.read_csv("auto-mpg.csv")
```

Out[2]:

	car name	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	130	3504	12.0	70	1
1	buick skylark 320	15.0	8	350.0	165	3693	11.5	70	1
2	plymouth satellite	18.0	8	318.0	150	3436	11.0	70	1
3	amc rebel sst	16.0	8	304.0	150	3433	12.0	70	1
4	ford torino	17.0	8	302.0	140	3449	10.5	70	1
...	...	...	...	...	...	...	...	...	...
393	ford mustang gl	27.0	4	140.0	86	2790	15.6	82	1
394	vw pickup	44.0	4	97.0	52	2130	24.6	82	2
395	dodge rampage	32.0	4	135.0	84	2295	11.6	82	1
396	ford ranger	28.0	4	120.0	79	2625	18.6	82	1
397	chevy s-10	31.0	4	119.0	82	2720	19.4	82	1

398 rows × 9 columns

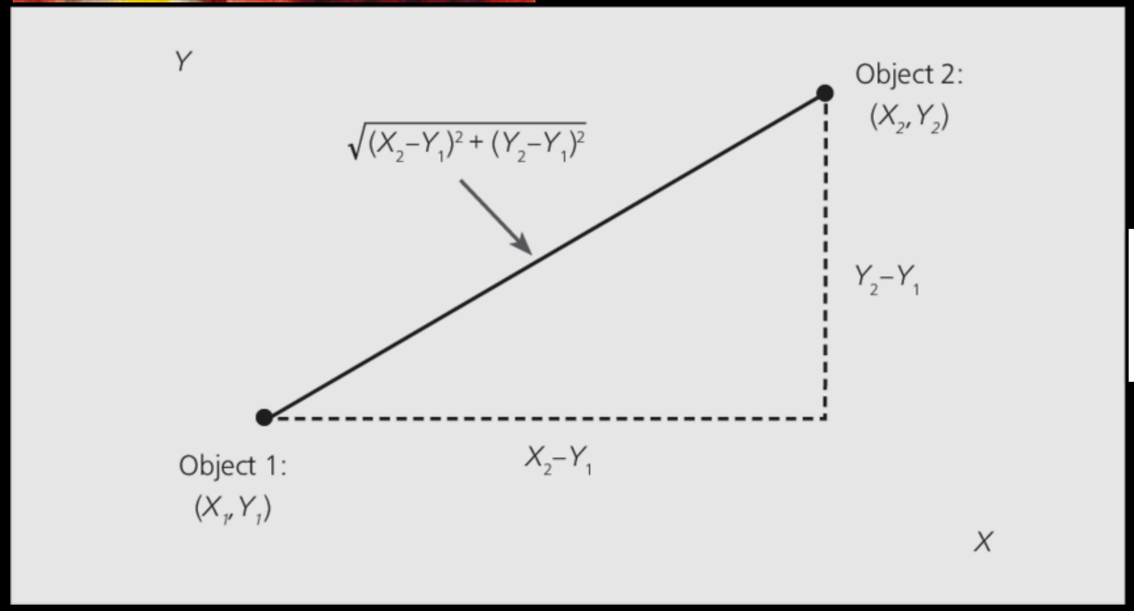
Euclidean Distance d(x,y):

Similarity: $1/d(r_A, r_B)$

Similarity $\rightarrow \infty$ as $d(r_A, r_B) \rightarrow 0$

Feature Matrix
For Models of Car

$$\begin{bmatrix} r_A, r_B, \ldots r_E \end{bmatrix}$$




```
In [1]: import pandas as pd

In [2]: pd.read_csv("auto-mpg.csv")

Out[2]:
```

	car name	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
0	chevrolet chevelle malibu	18.0	8	307.0	130	3504	12.0	70	1
1	buick skylark 320	15.0	8	350.0	165	3693	11.5	70	1
2	plymouth satellite	18.0	8	318.0	150	3436	11.0	70	1

Euclid is great. Why more distance metrics?

394	vw pickup	44.0	4	97.0	52	2130	24.6	82	2
395	dodge rampage	32.0	4	135.0	84	2295	11.6	82	1
396	ford ranger	28.0	4	120.0	79	2625	18.6	82	1
397	chevy s-10	31.0	4	119.0	82	2720	19.4	82	1

398 rows × 9 columns

Two main motivations for needing other distance metrics:

"Curse of dimensionality" issues

Sensitive to a data point's "scale"

DOI:10.1145/2347736.2347755

Tapping into the "folk knowledge" needed to advance machine learning applications.

BY PEDRO DOMINGOS

A Few Useful Things to Know About Machine Learning

MACHINE LEARNING SYSTEMS automatically learn programs from data. This is often a very attractive alternative to manually constructing them, and in the last decade the use of machine learning has spread rapidly throughout computer science and beyond.

Euclidean distance is problematic in higher dimensions

data mining or predictive analytics) will be the driver of the next big wave of innovation.[15] Several fine textbooks are available to interested practitioners and researchers (for example, Mitchell[16] and Witten et al.[24]). However, much of the "folk knowledge" that is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind up producing less-than-ideal results. Yet much of this folk knowledge is fairly easy to communicate. This is the purpose of this article.

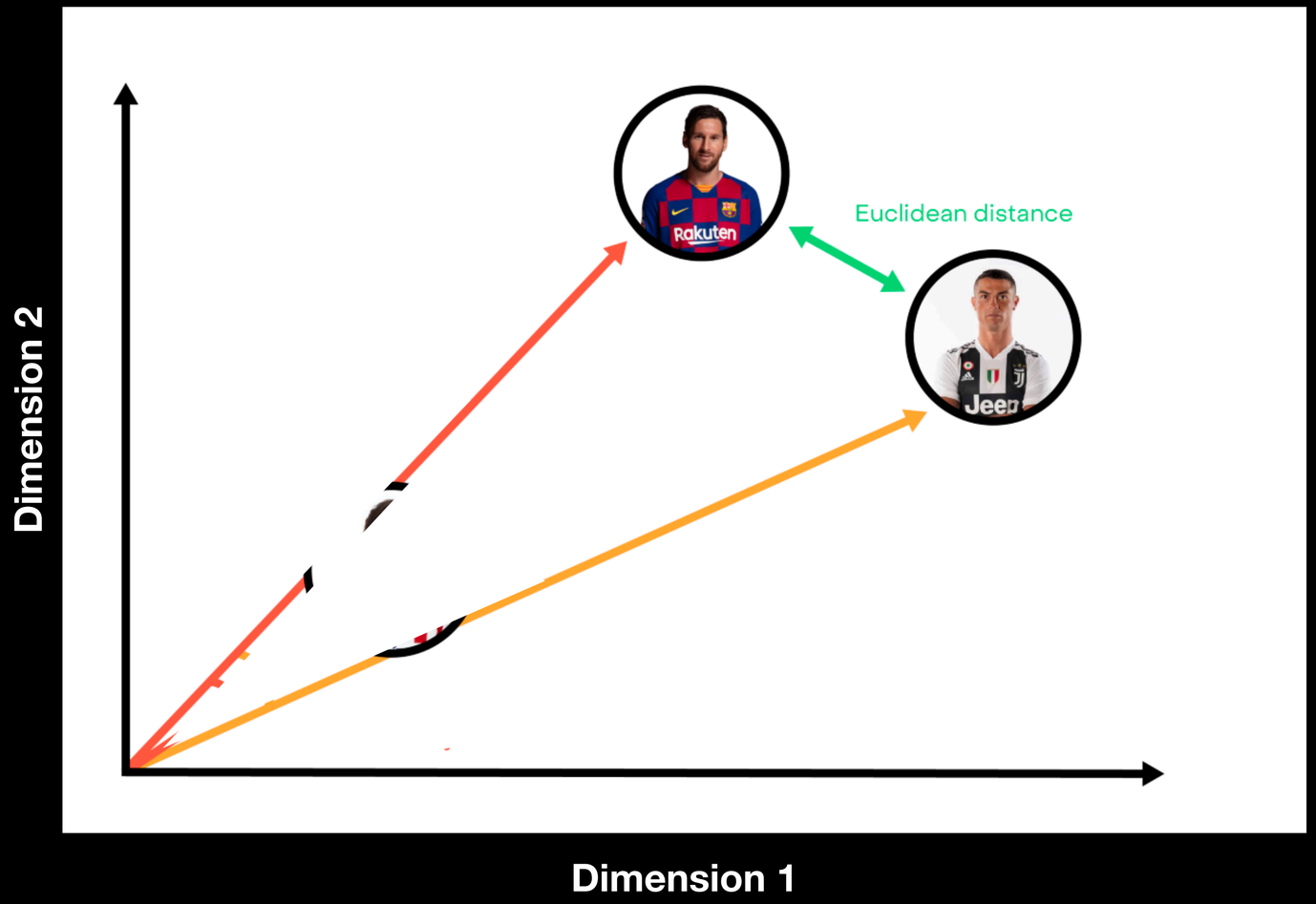
» key insights

- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples. This is
- machine learning applications requires a substantial amount of "black art" that is difficult to find in textbooks.
- This article summarizes 12 key lessons that machine learning researchers and practitioners have learned. These include pitfalls to avoid, important issues to focus on, and answers to common questions.

Adjacency Matrix *A*

	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0

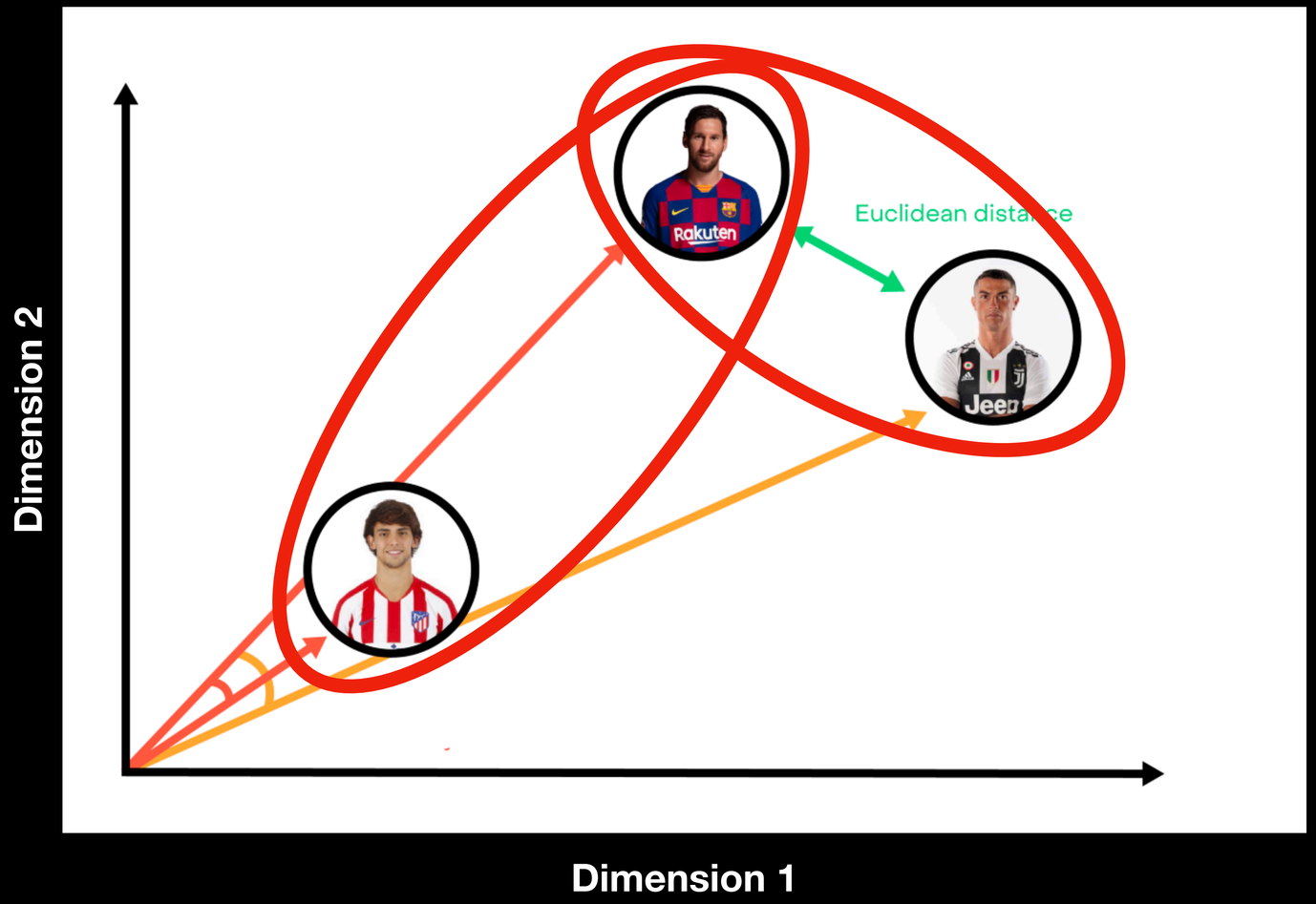
Similarity between two vectors of edge weights



Adjacency Matrix *A*

	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0

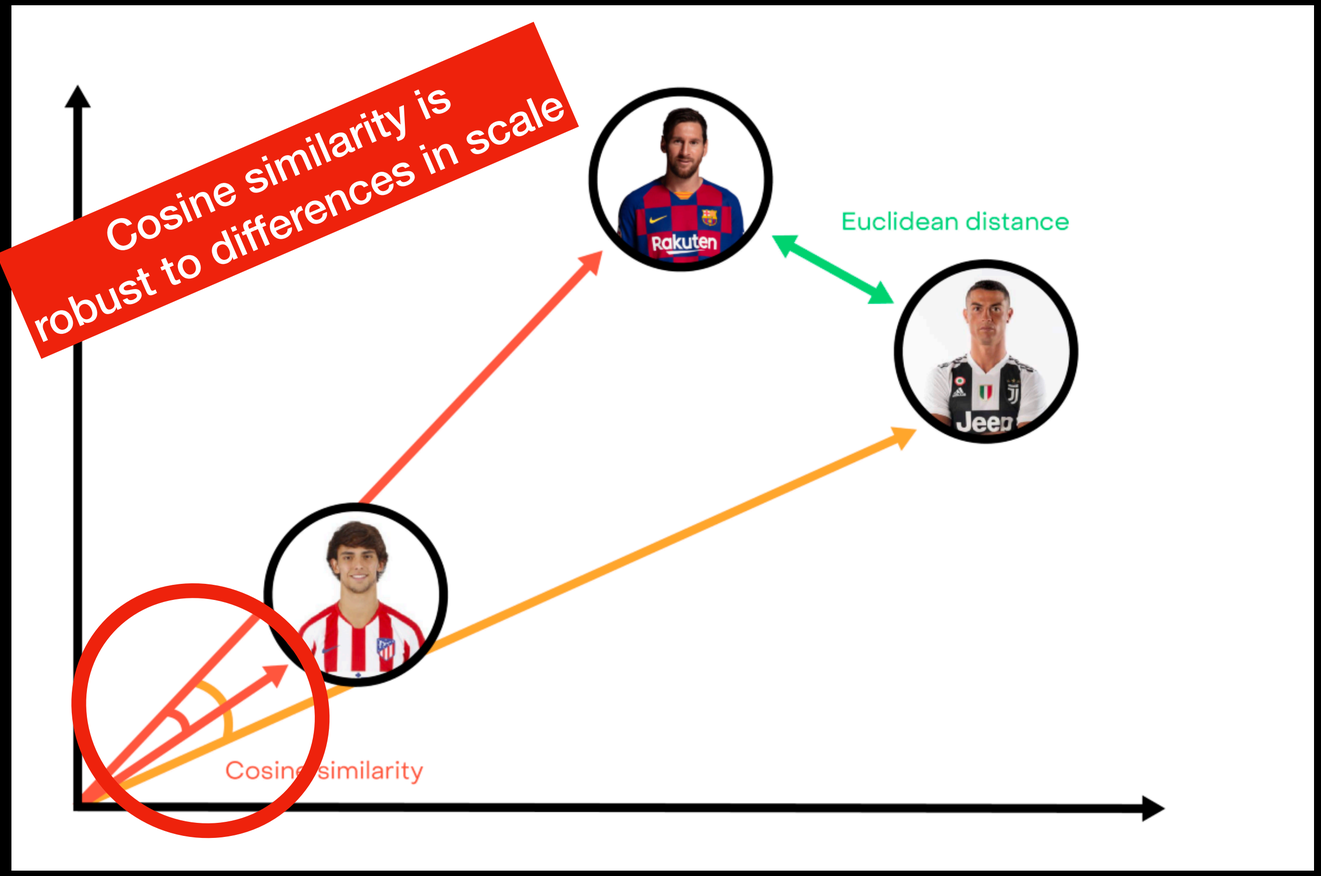
Similarity between two vectors of edge weights



Adjacency Matrix *A*

	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0

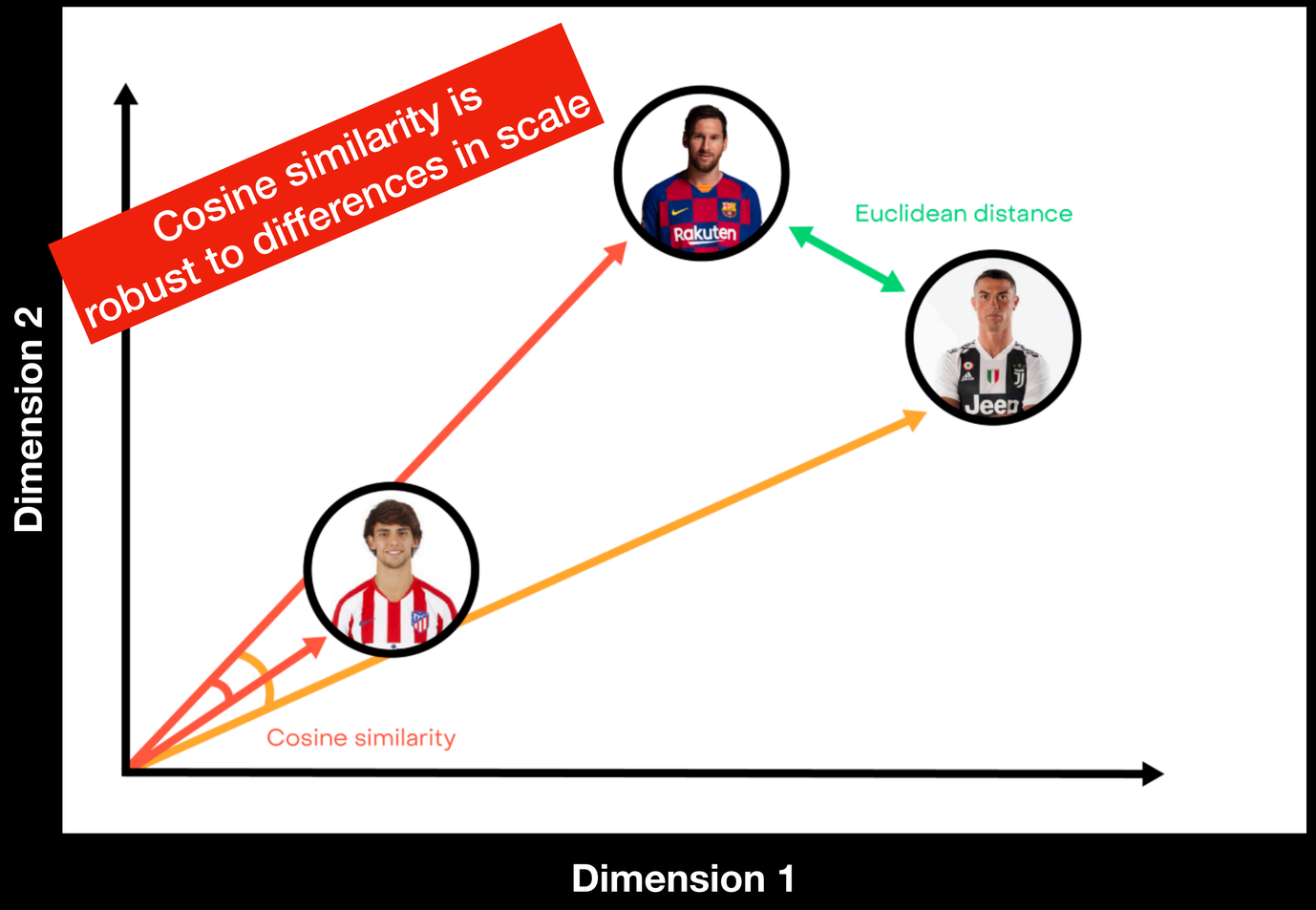
Similarity between two vectors of edge weights



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Adjacency Matrix *A*

	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0



cos(θ) is bounded in [-1, 1]

Cosine Distance = 1 - cos(θ)

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

This Week's Learning Objectives

Define "homophily"

Give examples of at least three distance metrics

Define shingling and its utility in evaluating similarity

Many Possible Distance Matrices

If i,j are in a graph…

$$D_{i,j} = \text{shortest\_path}(i, j)$$

Distance Matrix *D*

	A	B	C	D	E
A					
B					
C					
D					
E					

Many Possible Distance Matrices

If i,j are in a graph…

$$D_{i,j} = \text{shortest\_path}(i, j)$$

$$D_{i,j} = \text{euclidean\_distance}(i, j)$$

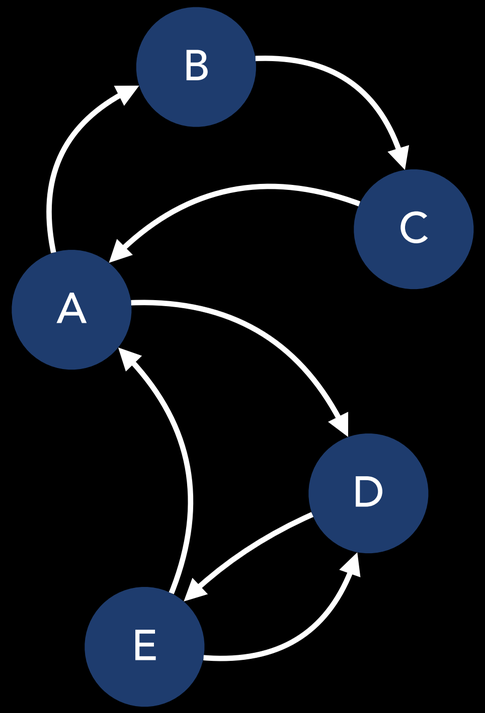
$$D_{i,j} = \text{hamming\_distance}(i, j)$$

$$D_{i,j} = 1 - \text{cosine\_similarity}(i, j)$$

...

Distance Matrix *D*

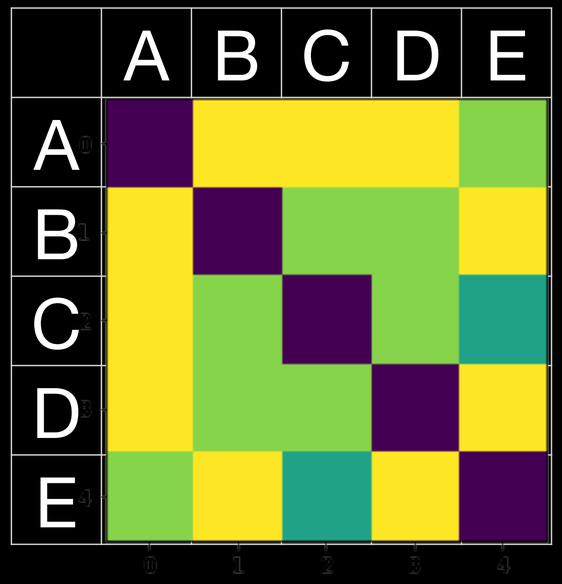
	A	B	C	D	E
A					
B					
C					
D					
E					



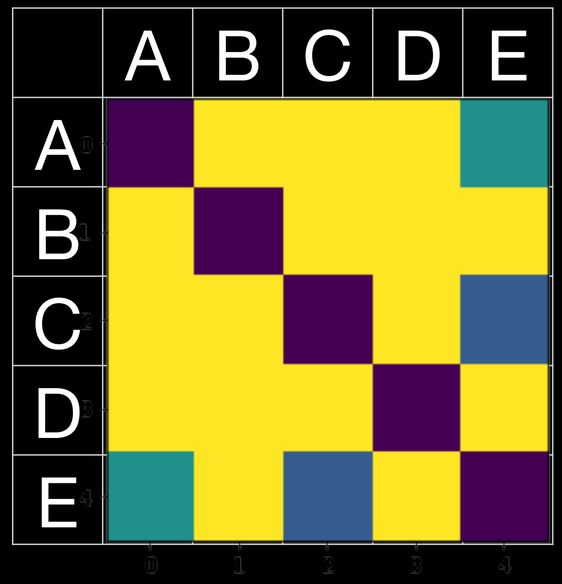
Adjacency Matrix *A*

	A	B	C	D	E
A	0	0	1	0	1
B	1	0	0	0	0
C	0	1	0	0	0
D	1	0	0	0	1
E	0	0	0	1	0

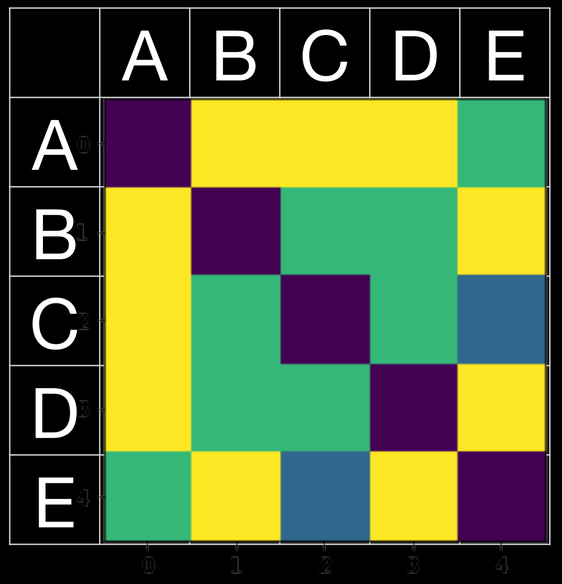
Euclidean Distance



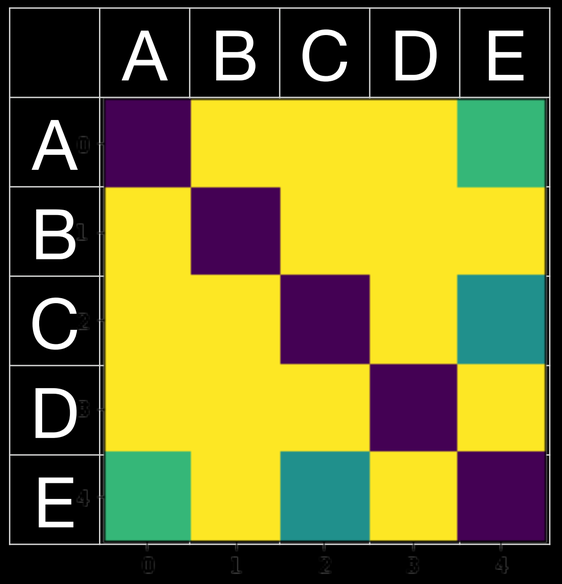
Cosine Distance



Hamming Distance



Jaccard Distance



jupyter 000-FbNetworkExtraction Last Checkpoint: 14 hours ago (autosaved)

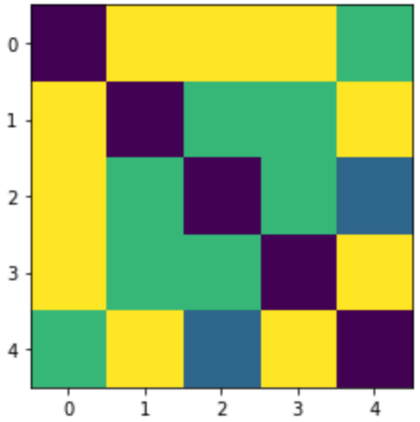
Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

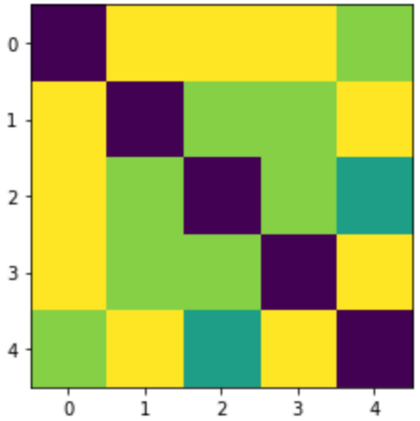
In [94]:
```
plt.imshow(sklearn.metrics.pairwise_distances(a, metric="hamming"))
```

Out[94]: <matplotlib.image.AxesImage at 0x7fd2a87294a8>



In [98]:
```
plt.imshow(sklearn.metrics.pairwise_distances(a, metric="euclidean"))
```

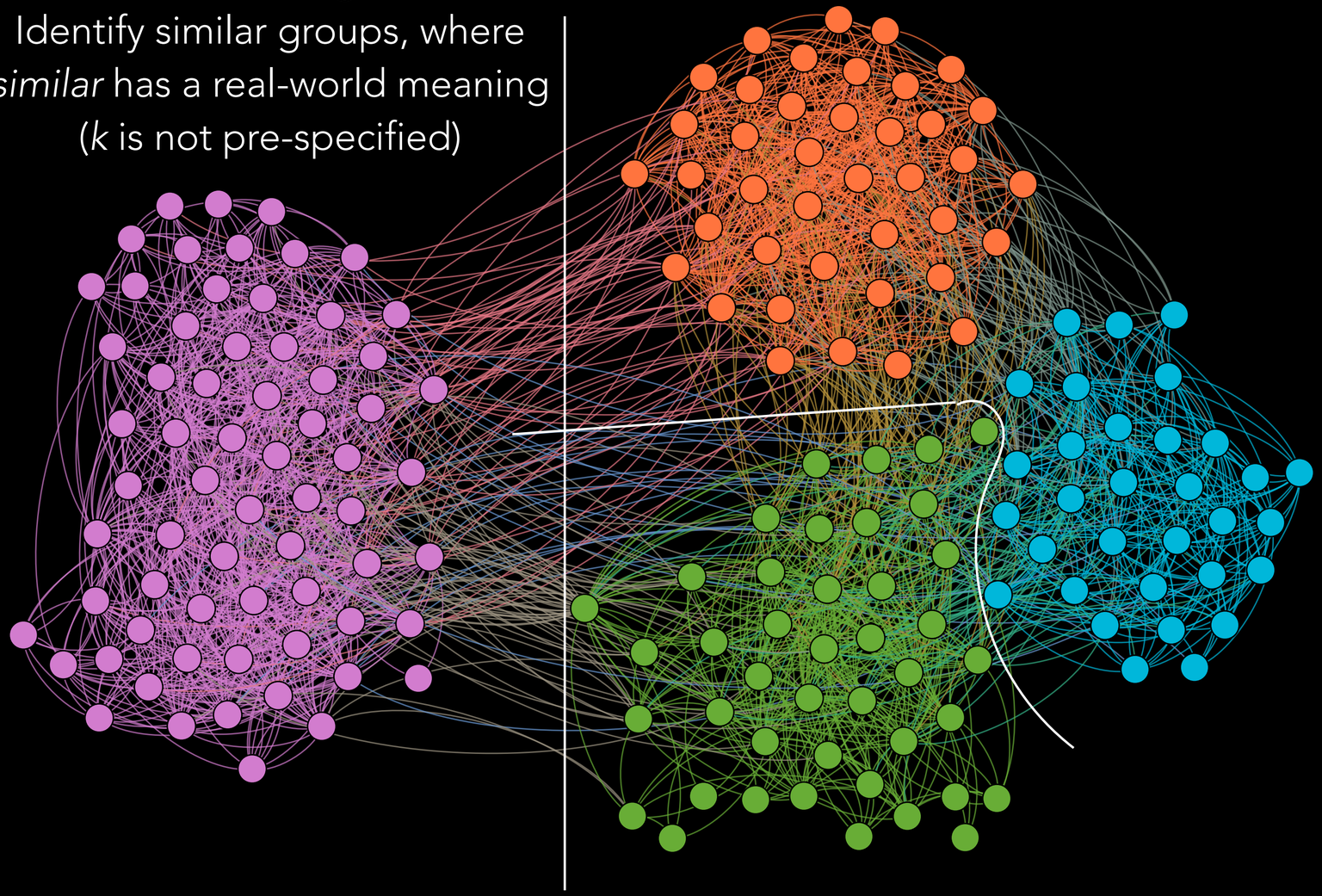
Out[98]: <matplotlib.image.AxesImage at 0x7fd2a4d93400>

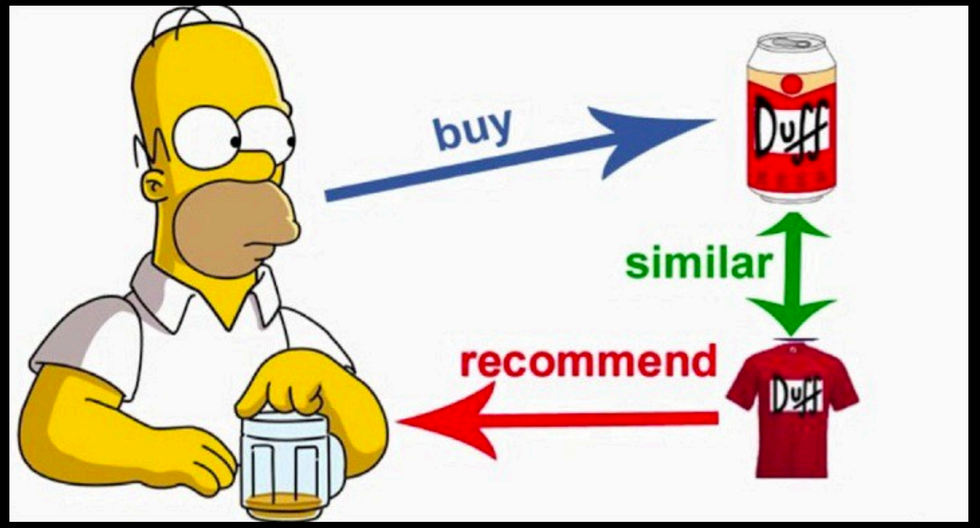


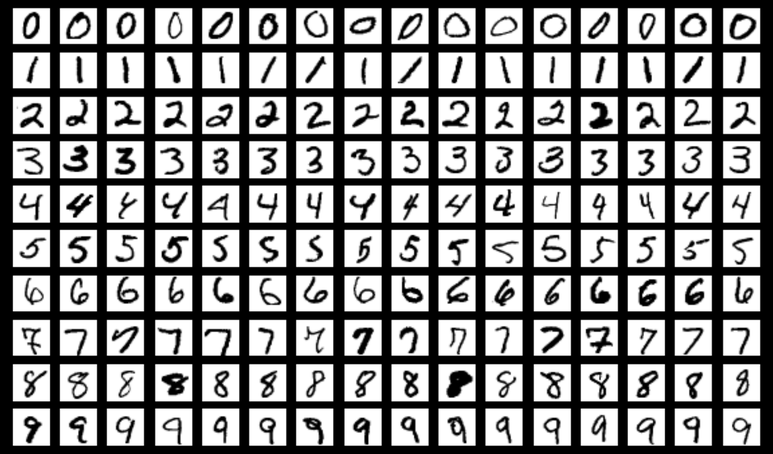
Why do these distance matrices matter?

Clustering:

Identify similar groups, where *similar* has a real-world meaning (*k* is not pre-specified)







Why do these distance matrices matter?

This Week's Learning Objectives

Define "homophily"

Give examples of at least three distance metrics

Define shingling and its utility in evaluating similarity

Questions?

Prof. Cody Buntain | @codybuntain | cbuntain@umd.edu
Director, Information Ecosystems Lab