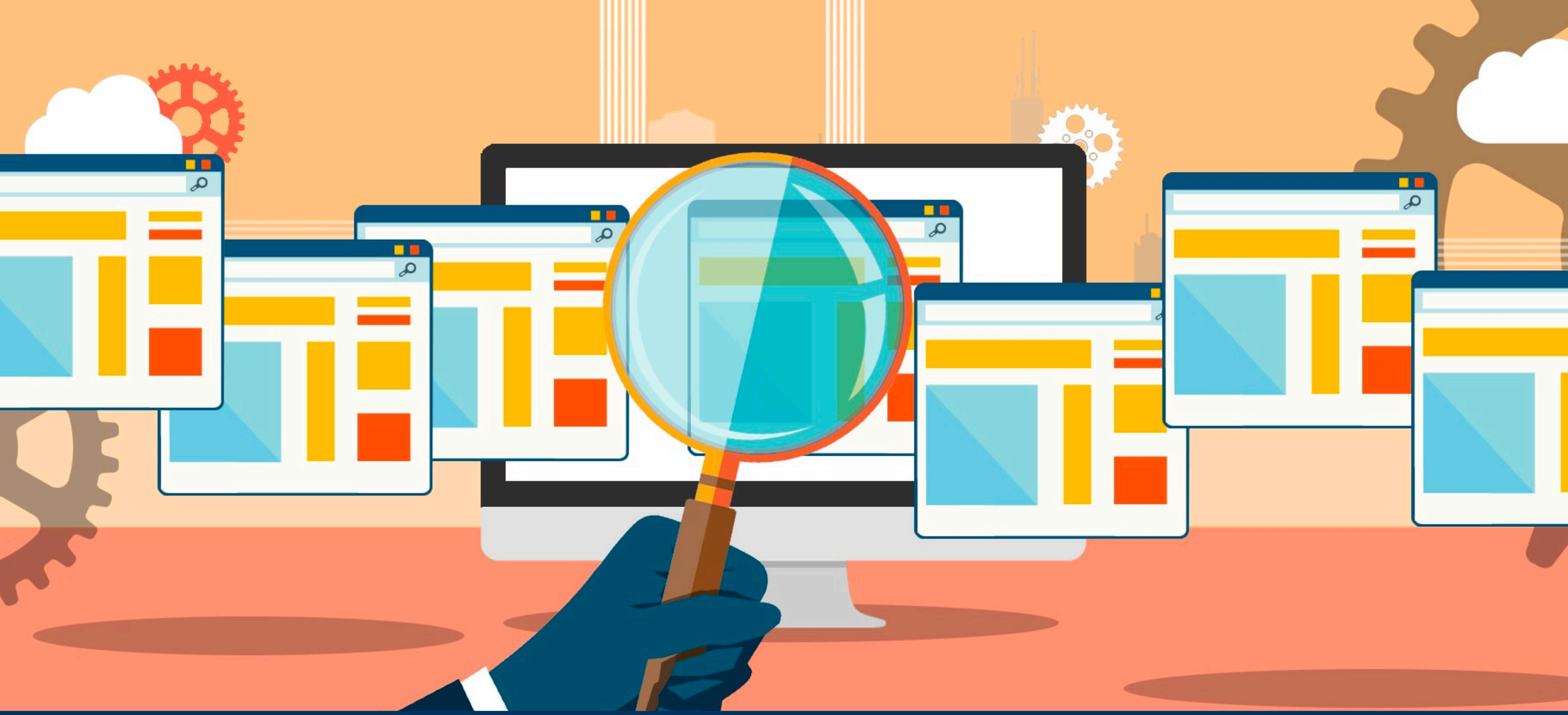


Probabilities and Clusters

INST414 - Data Science Techniques

This Module's Learning Objectives

Part 1

- Define discrete random variable's sample space and probability distribution
- Identify whether two events are probabilistically independent
- Calculate the conditional probability of an event given another event
- Calculate probability of a sample being in a particular cluster

This Module's Learning Objectives

Part 1

Define discrete random variable's sample space and probability distribution

Identify whether two events are probabilistically independent

Calculate the conditional probability of an event given another event

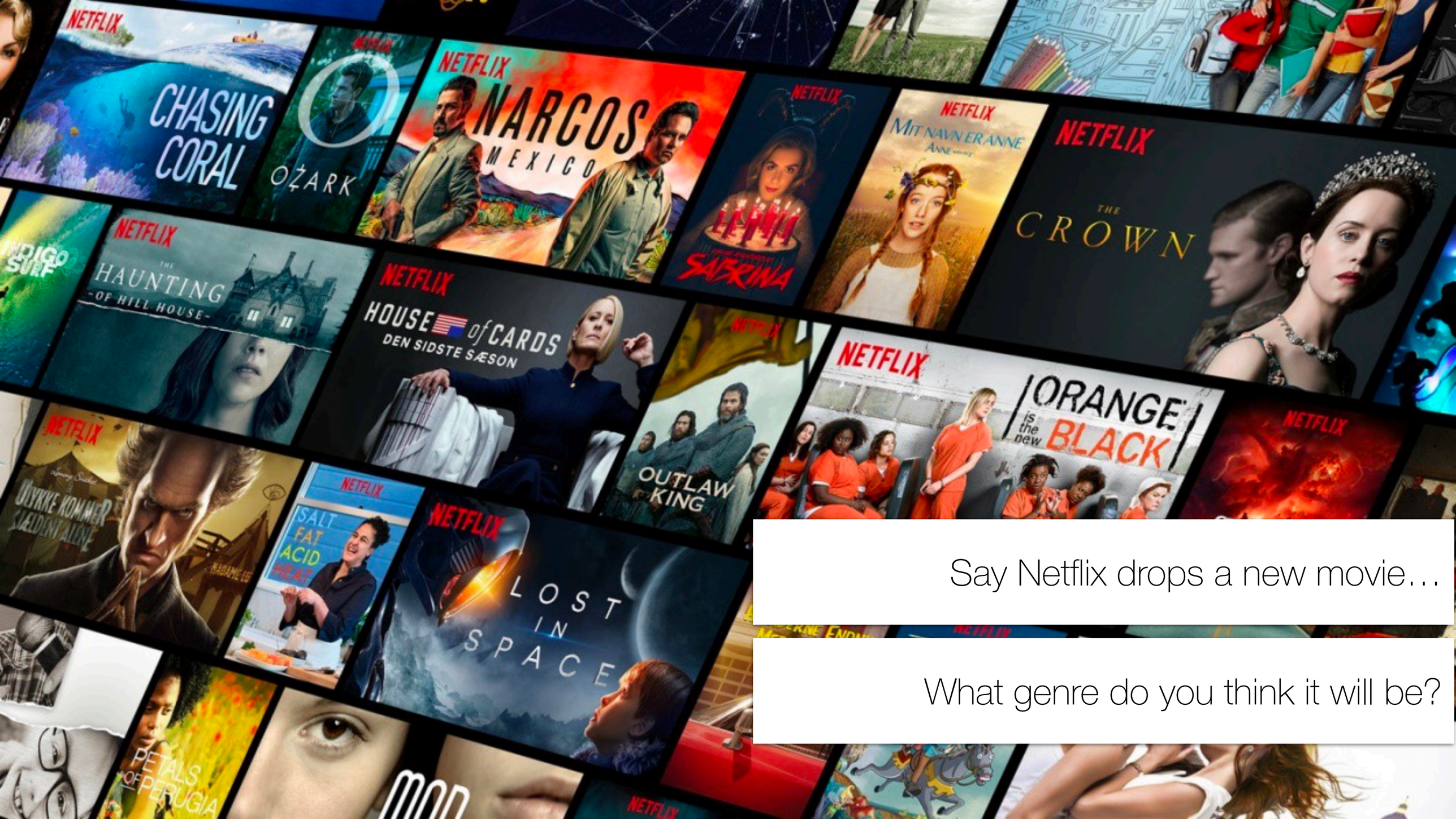
Calculate probability of a sample being in a particular cluster

What is a “variable”?

“A factor that is liable to vary or change”

“An abstract storage location …and symbolic name”

“Variable” in the probabilistic context is different



Say Netflix drops a new movie…

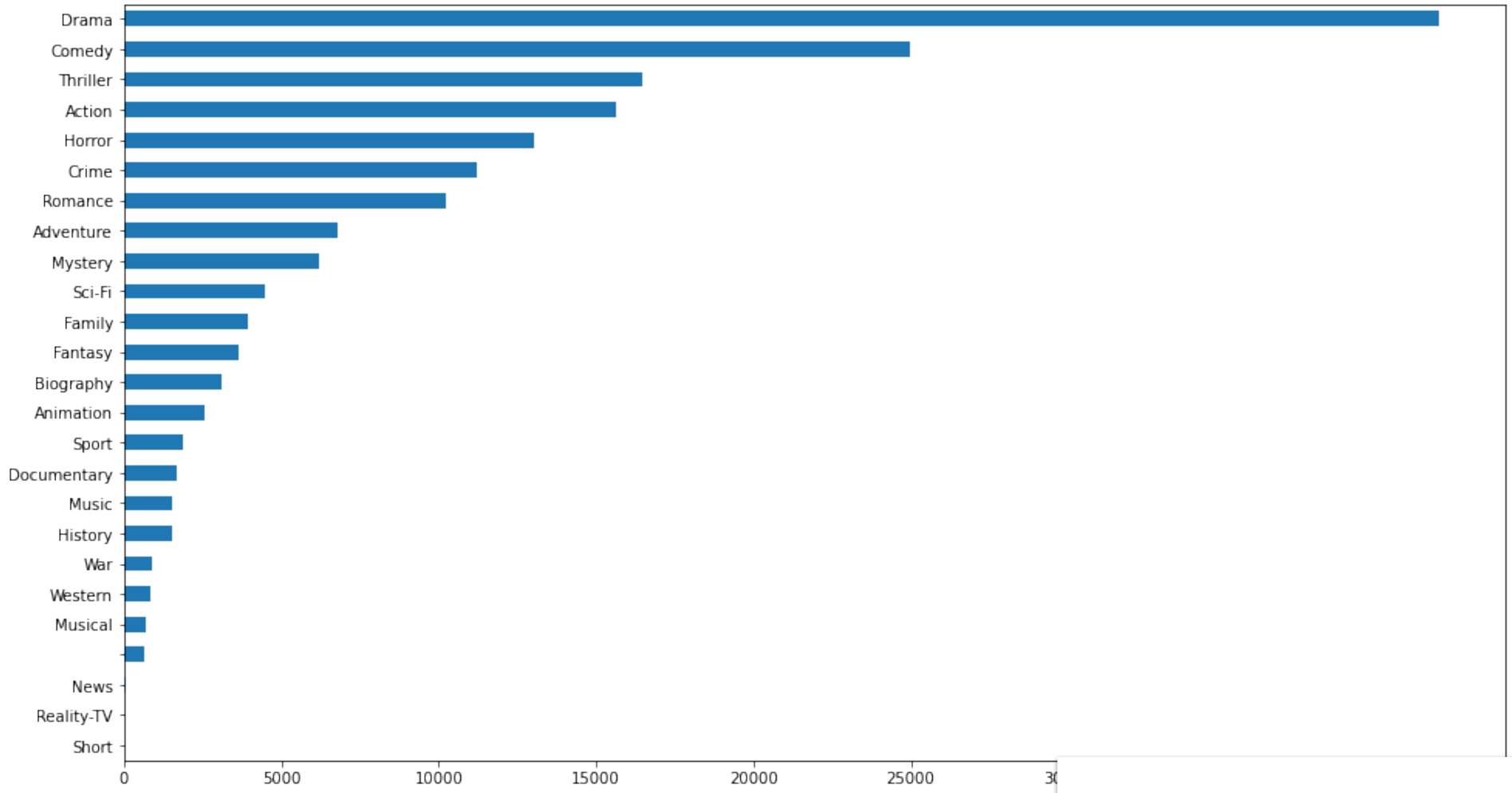
What genre do you think it will be?

```
In [19]: actor_genre_df.head(20)
```

Out[19]:

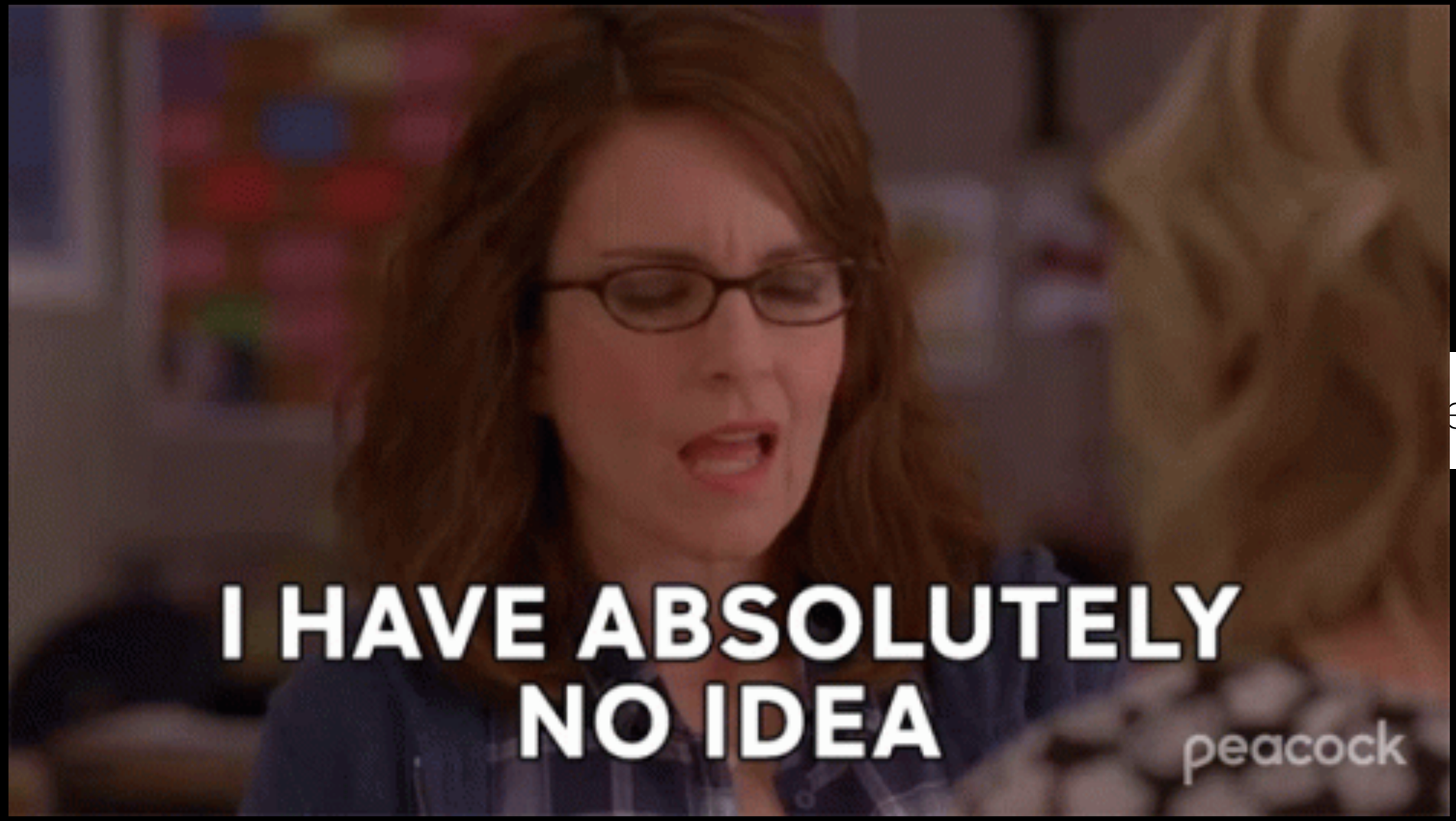
	Comedy	antasy	Romanc	Action	rime	Adventure	Mystery	Thriller	Drama	Biography	...	Sport		News	Family	Western	Short
nm0000212	16.0	3.0	16.	5.0	4.0	2.0	5.0	3.0	16.0	2.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0413168	8.0	3.0	6.	14.0	6.0	11.0	5.0	2.0	13.0	5.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0000630	10.0	2.0	6.	4.0	1.0	2.0	2.0	4.0	17.0	6.0	...	4.0	1.0	1.0	0.0	0.0	0.0
nm0005227	12.0	1.0	3.	2.0	0.0	3.0	0.0	1.0	5.0	1.0	...	1.0	0.0	0.0	3.0	0.0	0.0
nm0697338	1.0	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	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	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	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	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	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	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	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	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	2.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
nm0000953	6.0	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.	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.	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								
nm0177016	0.0	0.0	0.	0.0	0.0	0.0	0.0	0.0	0.0								
nm0907480	0.0	0.0	0.	1.0	0.0	1.0	0.0	0.0	0.0								

Let's sum up counts for these genres



Most likely to be a Drama

What is the Hollywood process that produces these genres?



ese genres?

What is the Hollywood process that produces these genres?

We'll call this process a random process

This process produces new movies with genres over time

What is a “random variable”?

A function mapping the outcome of a random experiment…

To a real number…

That has some uncertainty about its value

What is a “random variable”?

A function mapping the outcome of a random experiment…

…al number…

That has some uncertainty about its value

Confusing language, but we can’t observe this function

Can’t know the value beforehand

Defining a random variable as a function allows us to:

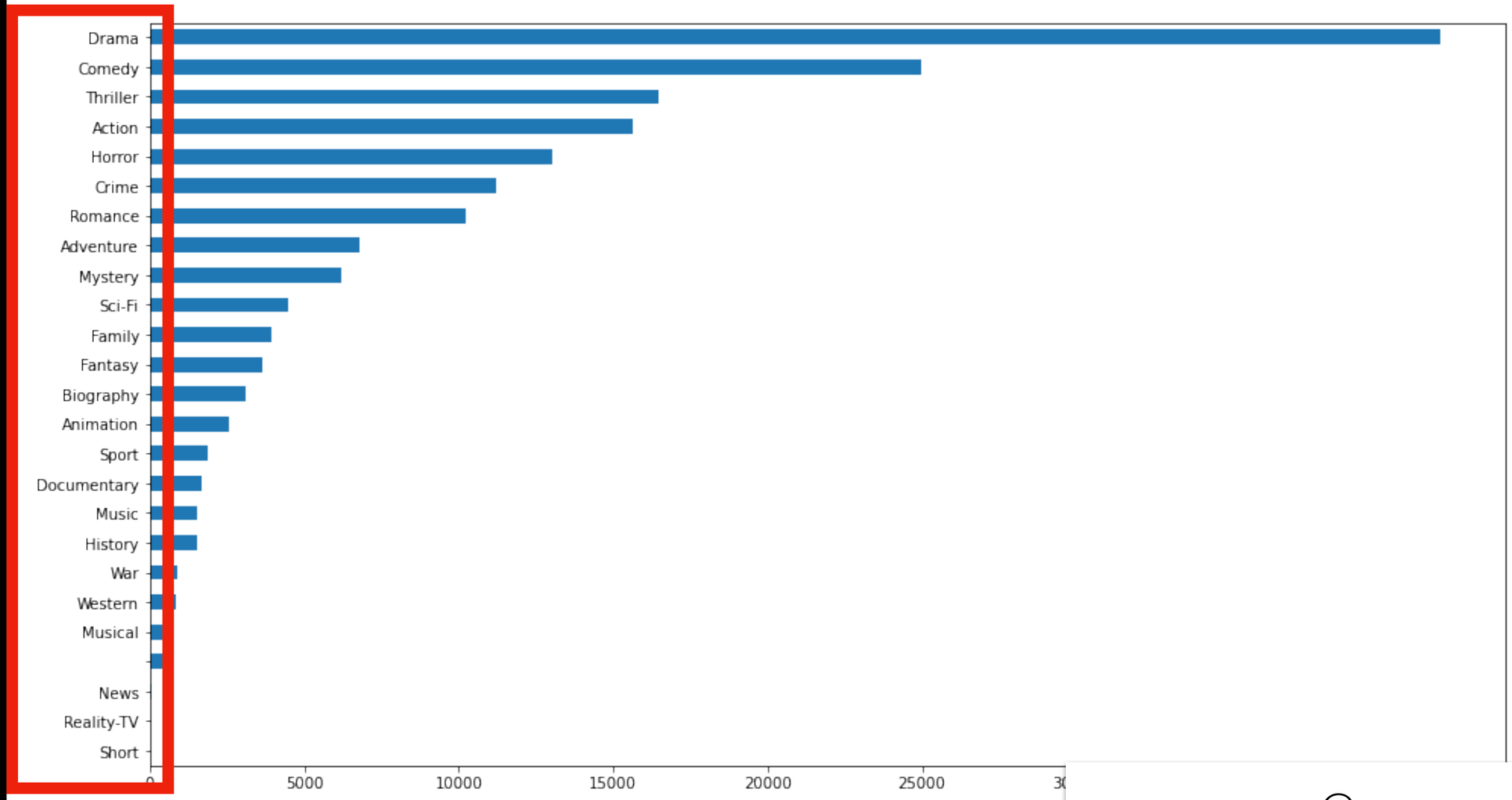
Define its domain and range

Random variable's "domain" ==> sample space

Random variable's "range" ==> possible outcomes

E.g., random variable *X* is a binary 0/1 for whether the next movie to be released is a sci-fi film

X = 1 <= shorthand => X = Sci-Fi

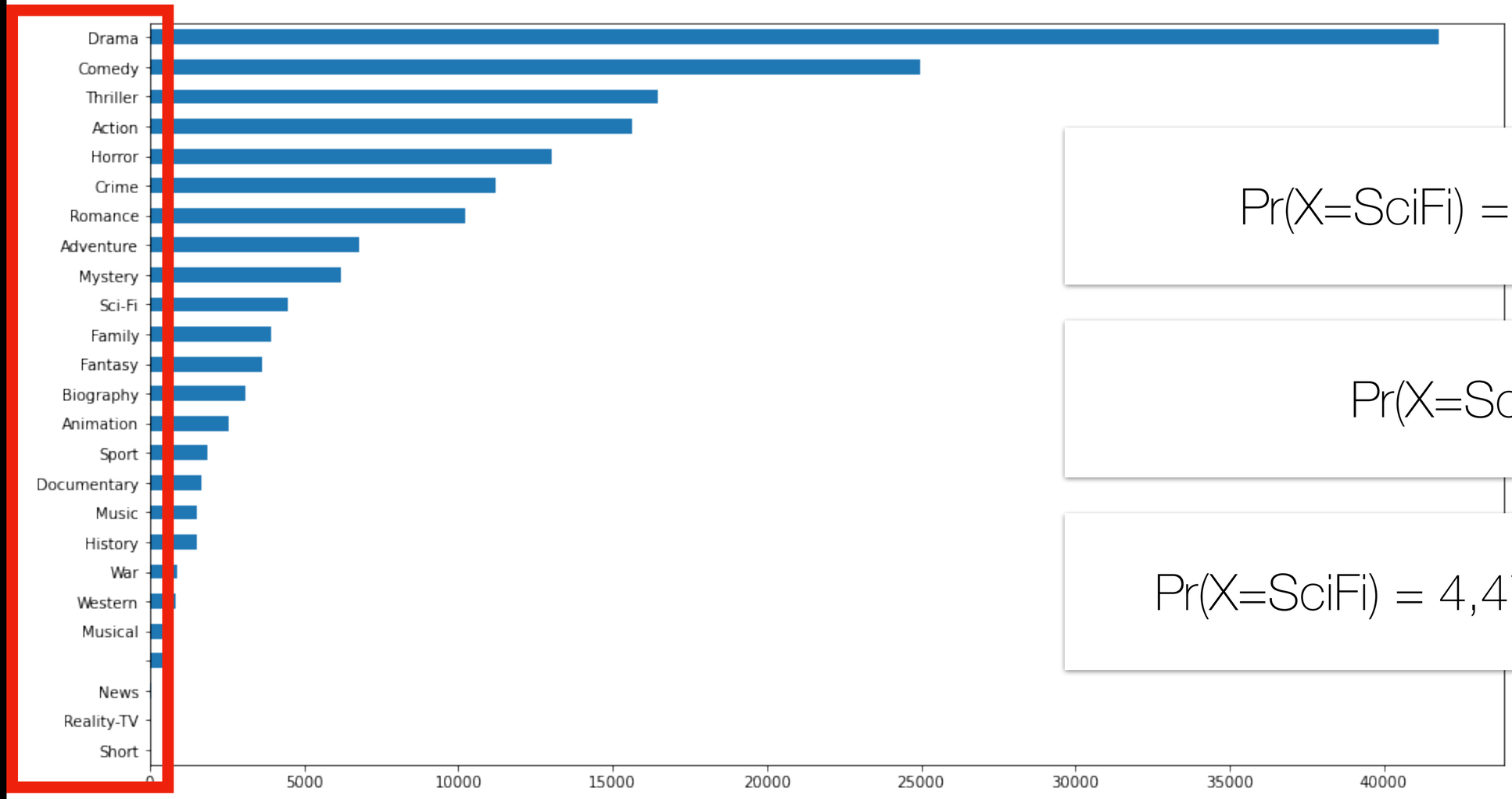


Genres represent the domain/sample space

A probability distribution for random variable X…

Is a function mapping random variable to its probability

Must sum to 1 across all values of X



Pr(X=SciFi) = Pr(X=1) = ?

Pr(X=SciFi) = 1/25?

Pr(X=SciFi) = 4,478/33,609?

This Module's Learning Objectives

Part 1

Define discrete random variable's sample space and probability distribution

Identify whether two events are probabilistically independent

Calculate the conditional probability of an event given another event

Calculate probability of a sample being in a particular cluster

Random variables X and Y are **independent** iff $P(X=x, Y=y) = P(X=x)P(Y=y)$

If I flip a coin twice, is the second outcome independent from the first outcome?

If I draw two socks from my laundry, is the color of the first sock independent from the color of the second sock?

Are “You use a Mac” and “the Green Line is on schedule” independent?

What about “Snowfall in the Himalayas / your favorite color is blue”?

What about “You follow @TomBrady / you watch football”?

"There is a traffic jam Baltimore / Orioles are playing a home game”?

Both of these are independent

“You use a Mac” / “the Green Line is on schedule”

“Snowfall in the Himalayas” / “your favorite color is blue”

These are **de**pendent

“You follow @TomBrady / you watch football”

"Traffic jam Baltimore / Orioles are playing a home game”

Other examples

The values of two dice (ignoring gravity!)

Whether it is raining and the number of taxi cabs

Whether it is raining and the amount of time it takes me to hail a cab

Two neighboring words in a sentence

A math example about social media…

Do emoji in tweets (E) change alter if they have hashtags (H)?

Are E and H independent of each other?

Does P(E=1, H=1) =?= P(E=1)P(H=1)?

Event E

Event H
#Swiftie

A math example about social media…

Event E

E	Pr(E)
E=1	0.75
E=0	0.25

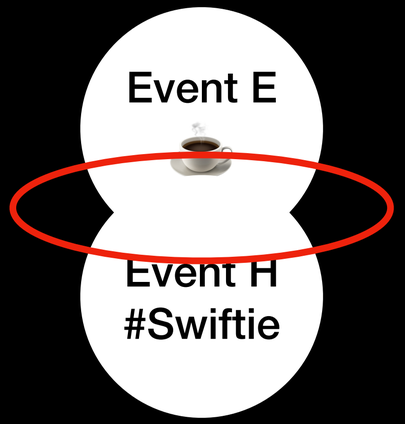
How often do tweets have emoji?

Event H
#Swiftie

H	Pr(H)
H=1	0.6
H=0	0.4

How often do tweets have hashtags?

A math example about social media…

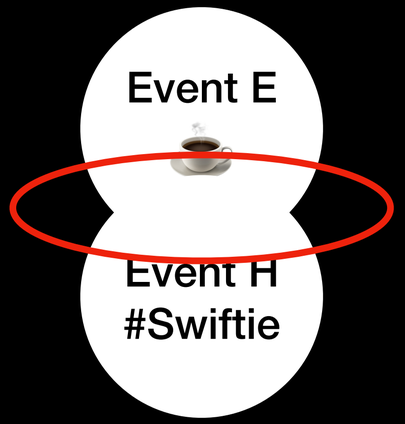


How often do tweets have emoji **AND** hashtags?

Tweet contains Emoji (E) or Hashtag (H)

	H=1	H=0	Marginal
E=1	0.45	0.3	0.75
E=0	0.15	0.1	0.25
Marginal	0.6	0.4	

A math example about social media…



E	Pr(E)
E=1	0.75
E=0	0.25

H	Pr(H)
H=1	0.6
H=0	0.4

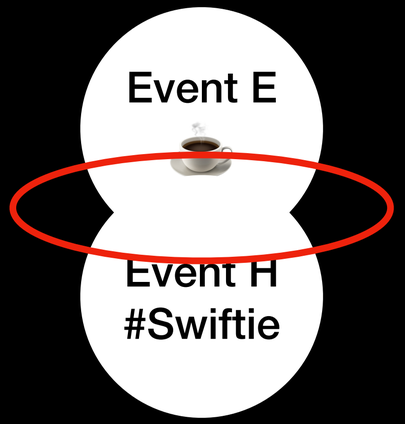
How often do tweets have emoji **AND** hashtags?

Tweet contains Emoji (E) or Hashtag (H)

	H=1	H=0	Marginal
E=1	0.45	0.3	0.75
E=0	0.15	0.1	0.25
Marginal	0.6	0.4	

A math example about social media…

Does P(E=1, H=1) =?= P(E=1)P(H=1)?



E	Pr(E)
E=1	0.75
E=0	0.25

H	Pr(H)
H=1	0.6
H=0	0.4

P(E=1, H=1) = 0.45

P(E=1)P(H=1) = 0.75 x 0.6 = 0.45

So yes, emoji and hashtag appear to be independent

Tweet contains Emoji (E) or Hashtag (H)

	H=1	H=0	Marginal
E=1	0.45	0.3	0.75
E=0	0.15	0.1	0.25
Marginal	0.6	0.4	

This Module's Learning Objectives

Part 1

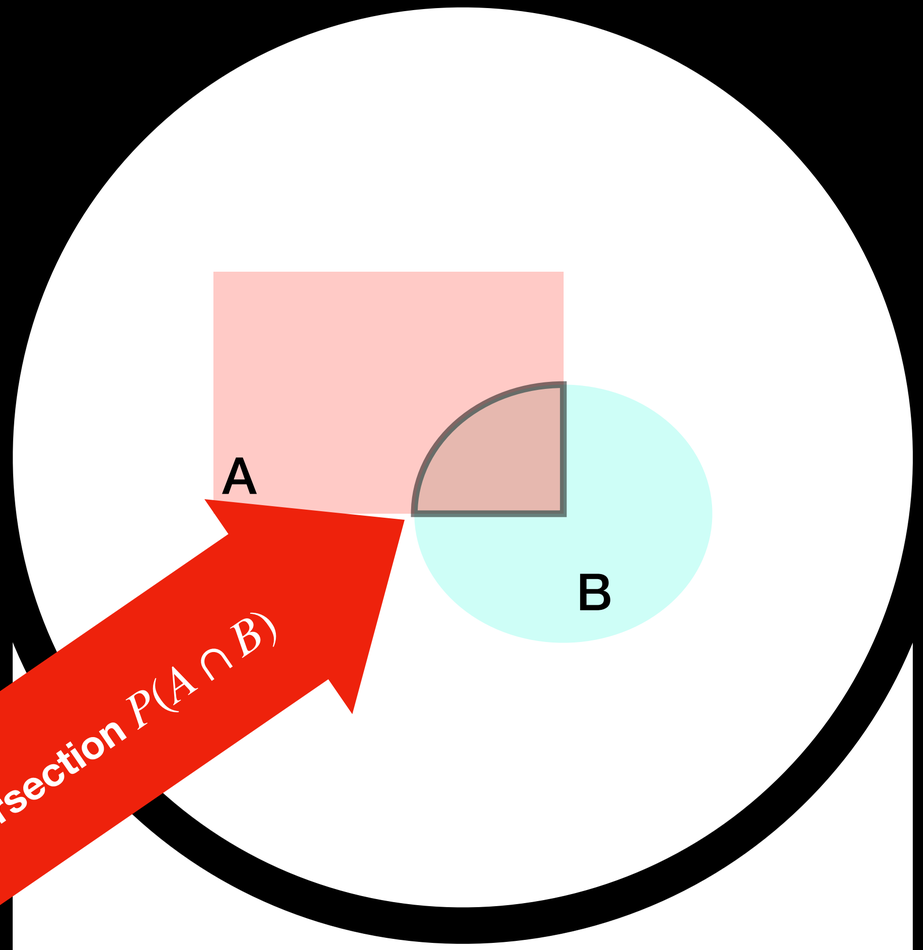
- Define discrete random variable's sample space and probability distribution
- Identify whether two events are probabilistically independent
- Calculate the conditional probability of an event given another event
- Calculate probability of a sample being in a particular cluster

An **event** is a set of outcomes to which a probability is assigned

Drawing a black card from a deck of cards

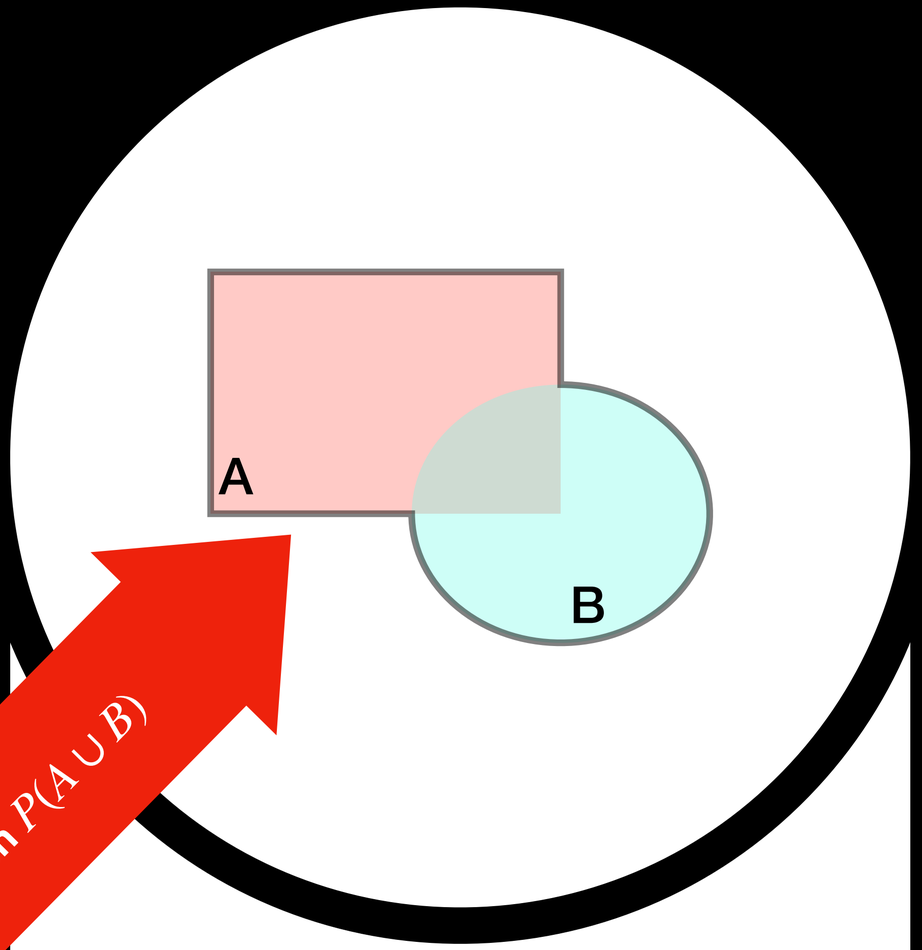
Drawing a King of Hearts

What else might be an event?



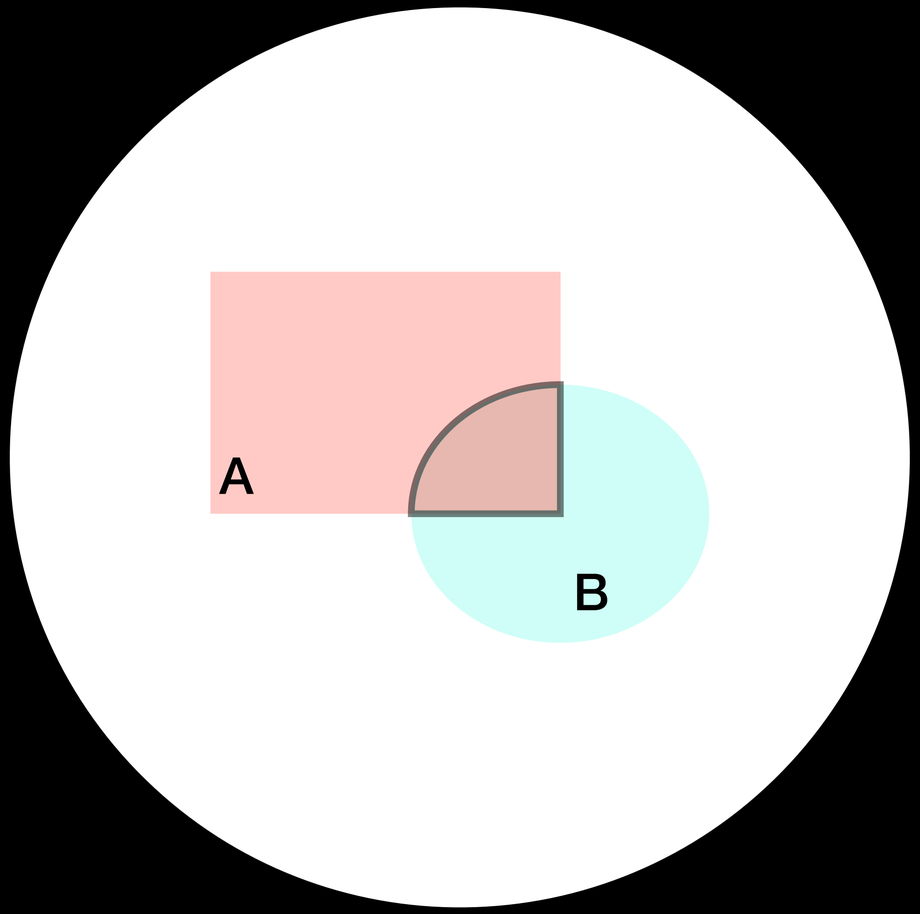
Intersection $P(A \cap B)$

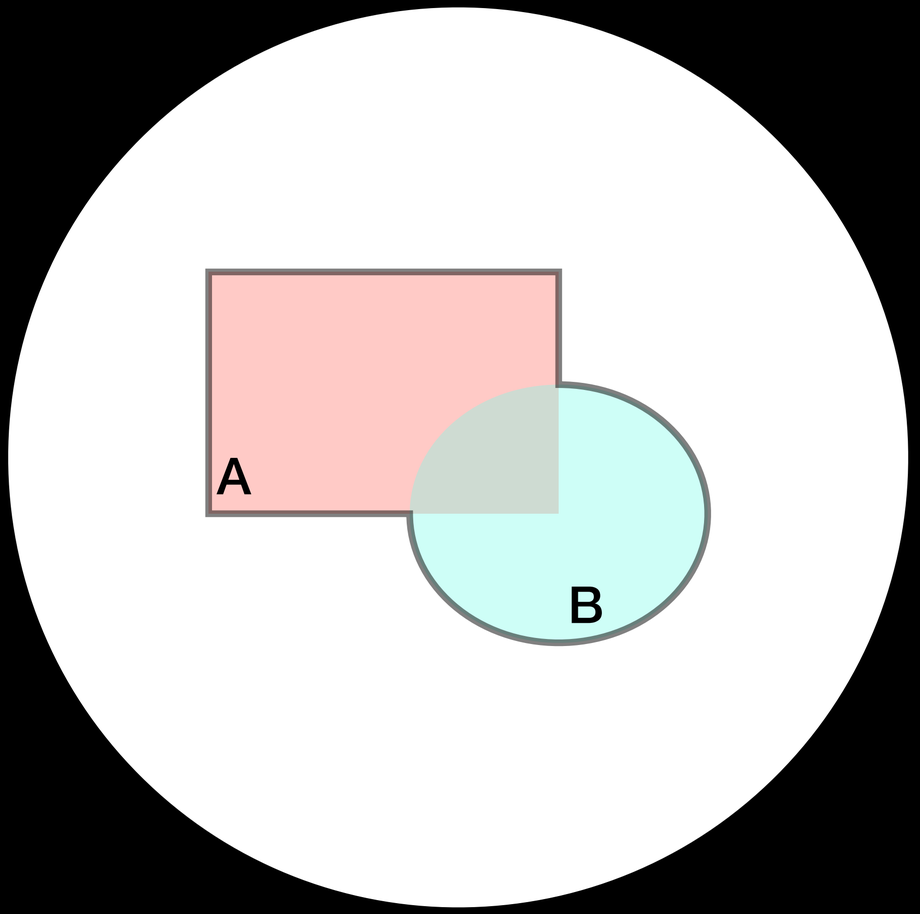
Draw a single card:
P(King of Hearts)



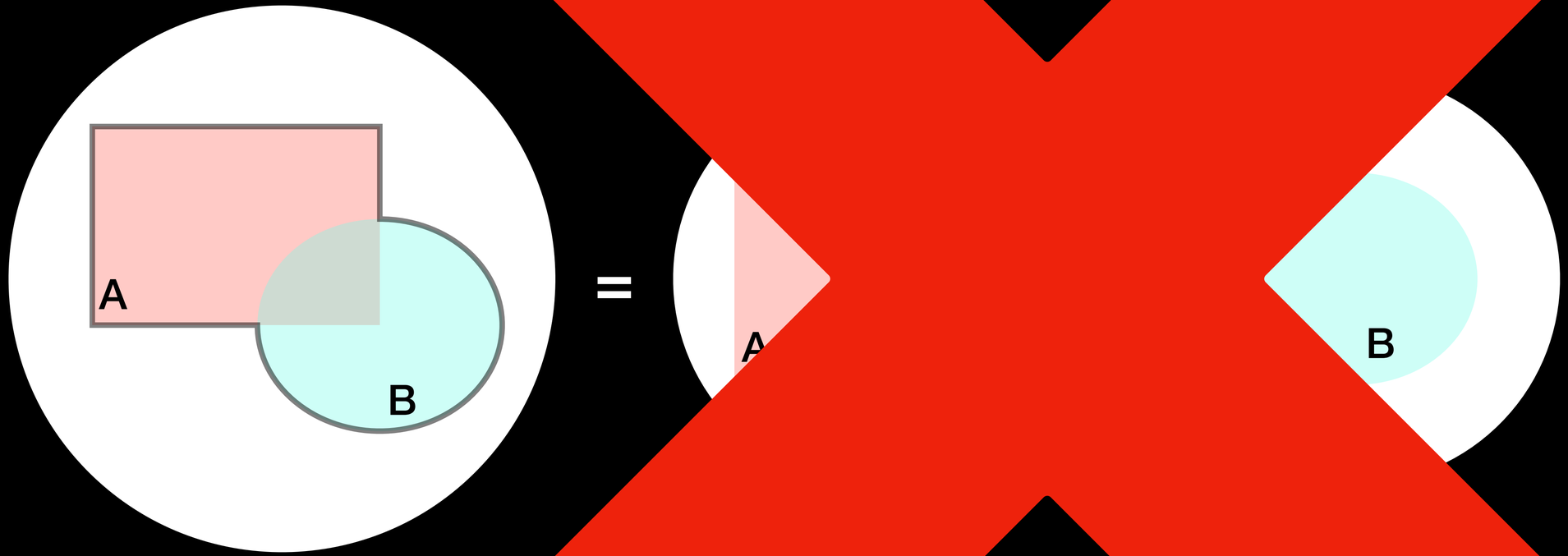
Union $P(A \cup B)$

Drawing a single card:
P(King or a Spade)

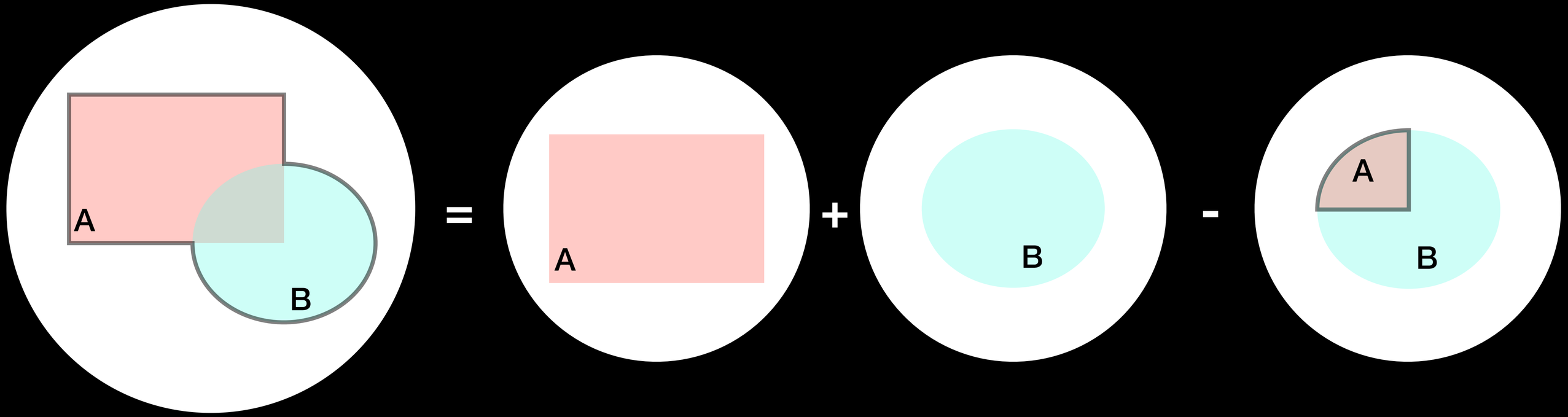




What is the relationship between these two quantities?



Is this correct?


A
B
=
A
+
B
-
A
B

Data science is often worried about “if-then” questions

If my e-mail looks like this, is it spam?

If I buy this stock, will my portfolio improve?

Language of **conditional probability** P(X | Y)

P(X = spam | Y = <email content>)

If my e-mail looks like this, is it spam?


P(X = <value increases> | Y = <portfolio>)
If I buy this stock, will my portfolio improve?

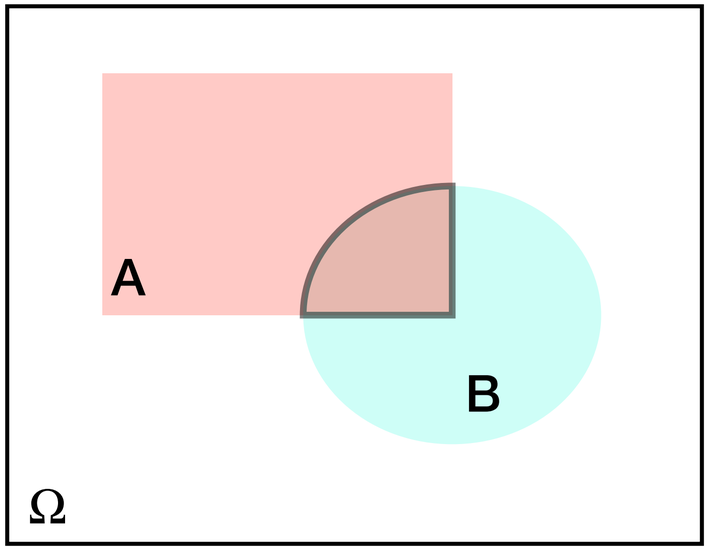
Conditional Probability

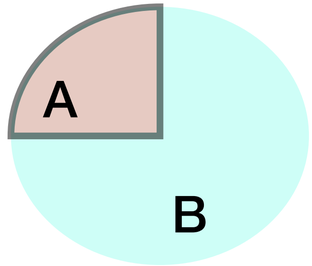
- The *conditional probability* of event A given event B is the probability of A when B is known to occur,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Joint Probability

Marginal Probability





New outcome space

An important connection exists between independence and conditional probability

Recall **independence** means $P(X=x, Y=y) = P(X=x)P(Y=y)$

$$P(X = x, Y = y) =$$

$$P(X = x, Y = y) = P(X = x \mid Y = y)P(Y = y)$$

If X and Y are independent…

$$P(X = x)P(Y = y) = P(X = x \mid Y = y)P(Y = y)$$

If X and Y are independent…

$$\frac{P(X = x)P(Y = y)}{P(Y = Y)} = \frac{P((X = x \mid Y = y)P(Y = y)}{P(Y = y)}$$

If X and Y are independent…

Knowing Y tells us nothing about X

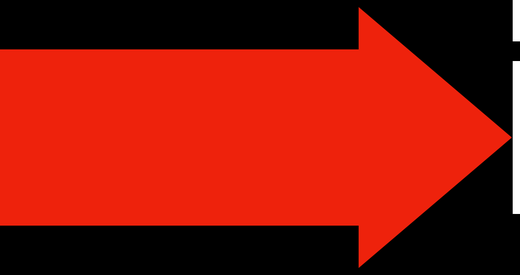
$$P(X = x) = P(X = x \mid Y = y)$$

If X and Y are independent…

This Module's Learning Objectives

Part 1

- Define discrete random variable's sample space and probability distribution
- Identify whether two events are probabilistically independent
- Calculate the conditional probability of an event given another event
- Calculate probability of a sample being in a particular cluster



Bridging Probabilities and Movie Clusters

jupyter 02-kMeans Last Checkpoint: 8 minutes ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted Python 3

Run Code

k-Means Example

Using the IMDB data, feature matrix and apply dimensionality reduction to this matrix via PCA and SVD.

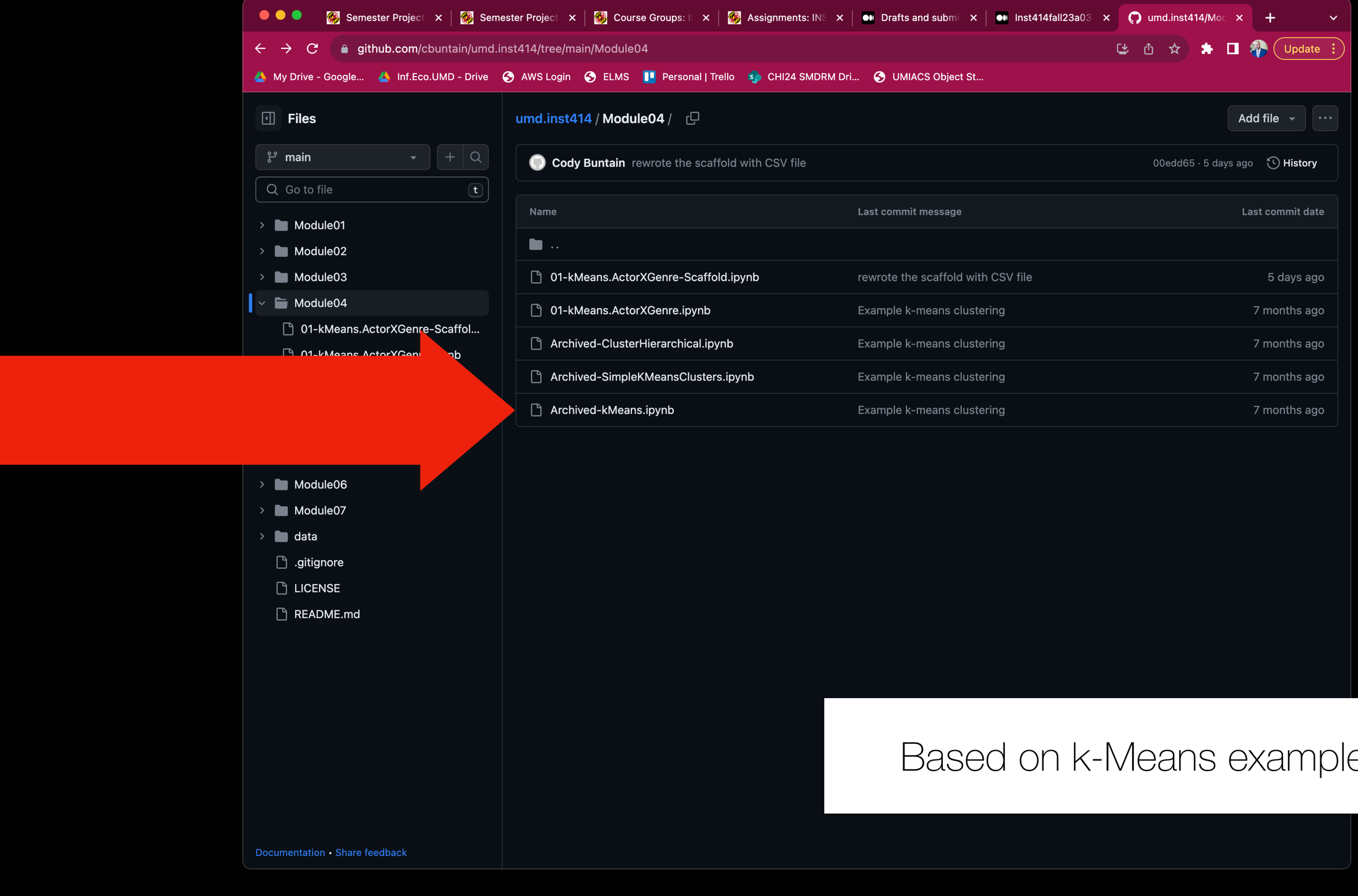
In [2]:
```
%matplotlib inline
```

In [3]:
```
import json
import random
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt

from collections import Counter
```

Based on k-Means example from GitHub



Based on k-Means example from GitHub

jupyter 02-kMeans Last Checkpoint: 10 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

Apply Clustering

We'll use AgglomerativeClustering from Sklearn to cluster this data.

In [60]:
```
from sklearn.cluster import KMeans

k = 32

model = KMeans(n_clusters=k)
```

In [61]:
```
model.fit(matrix_reduced)
```

Out[61]:
```
KMeans(n_clusters=32)
```

In [ ]:
```

```

In [62]:
```
reduced_df = df
```

In [63]:
```
reduced_df["cluster"] = model.labels_
```

In [64]:
```
reduced_df[["cluster"]]
```

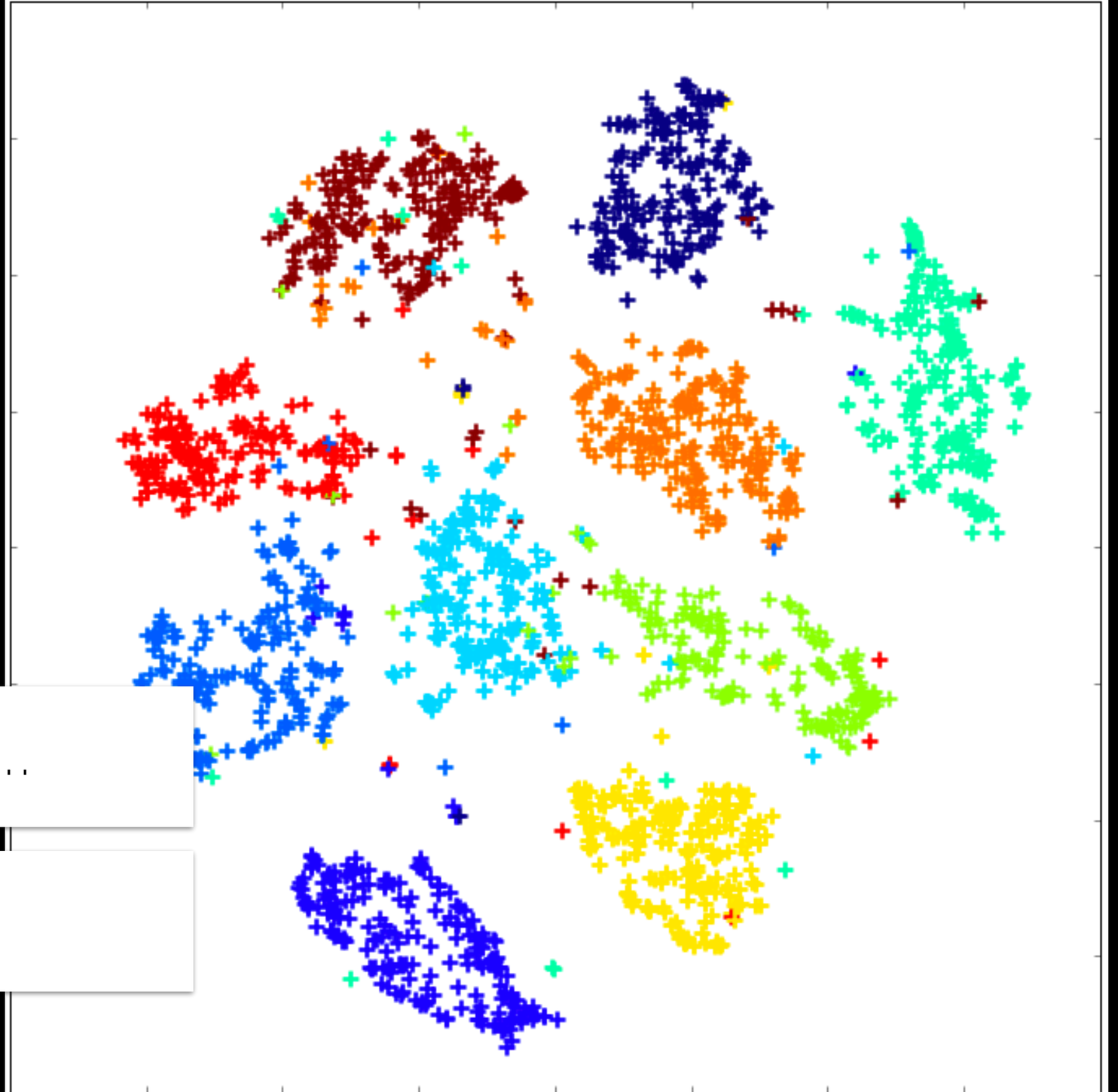
Out[64]:

	cluster
tt0069049	3
tt0088751	3
tt0093119	3
tt0094859	3
tt0096056	3
...	...
tt9914942	3
tt9916170	3
tt9916190	3
tt9916362	3
tt9916428	3

102445 rows × 1 columns

In [65]:
```
reduced_df["cluster"].value_counts()
```

With k-means, if k=32, what's the probability of a movie belonging to a cluster?



If we know nothing about k clusters…

“Prior” probability Pr(X)?

Random Variable

$$Pr(X) = 1/k$$

$$k = 32, Pr(X) = 1/32 = 0.03125$$

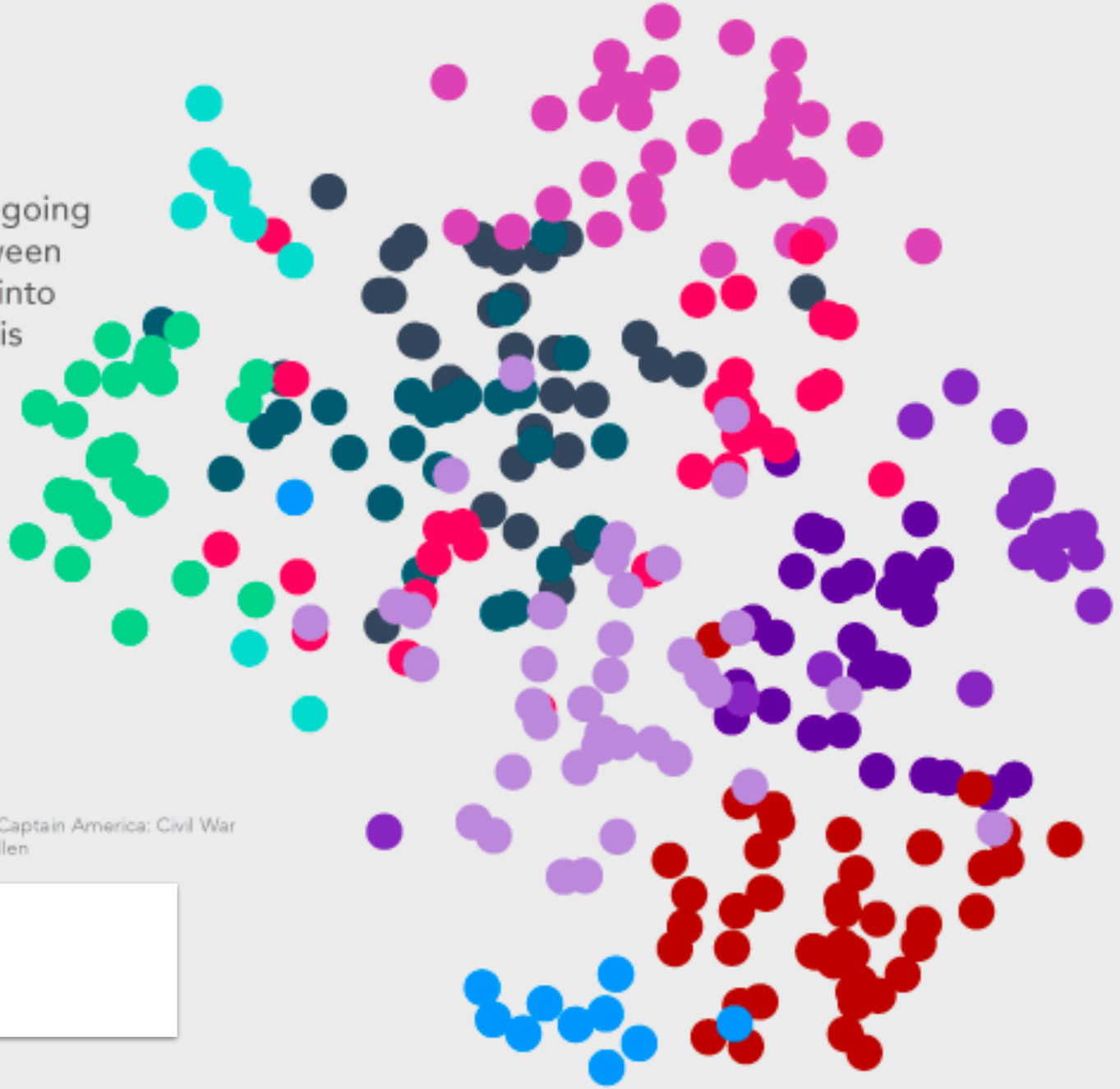
Pr(X) => “Probability of being in the Xth cluster”

Without additional info, equivalent to rolling 32-sided die

MOVIE CLUSTERS

In order to better understand the film landscape, we used our members' moviegoing history to divide 299 major releases between January 1, 2015 and September 1, 2016 into 11 different clusters*. The Horror cluster is made up of 26 movies, including titles such as The Conjuring 2, Crimson Peak, and The Shallows.

- Tentpole
- Adventure
- Drama
- Indie
- Horror
- Animation
- Comedy
- Christian
- Art-House
- African American
- Action



Tentpole: Star Wars: Episode VII - The Force Awakens, Jurassic World, Captain America: Civil War
Adventure: The Legend of Tarzan, Central Intelligence, London Has Fallen
Drama: Sully, The Intern, Bridge of Spies

Action: American Sniper, The Revenant, Creed

After running clustering…

jupyter 01-kMeans.MovieXGenre (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

Out[9]:
```
KMeans
KMeans(n_clusters=32)
```

In [ ]:

In [10]:
```
cluster_labels = cluster_model.predict(df)
movie_cluster_df = pd.DataFrame(cluster_labe
```

In [11]:
```
movie_cluster_df["cluster"].value_counts()
```

Out[11]:
```
cluster
1     2855
3     1771
10    1477
8     1211
2     1136
25     974
5      786
4      734
22     689
15     673
21     673
16     640
6      635
7      627
23     608
20     483
0      413
13     413
27     407
19     397
12     368
11     357
17     340
28     318
30     278
24     232
18     230
9      225
14     213
29     165
31     148
26     144
Name: count, dtype: int64
```

In k-Means example, we have 51,022 movies after some filtering

Fit k=32 to these 21k movies

Now what is the Pr(X=0)? Pr(X=1)?

jupyter 01-kMeans.MovieXGenre (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

Out[9]:
```
KMeans
KMeans(n_clusters=32)
```

In [ ]:

In [10]:
```
cluster_labels = cluster_model.predict(df)
movie_cluster_df = pd.DataFrame(cluster_labe
```

In [11]:
```
movie_cluster_df["cluster"].value_counts()
```

Out[11]:
```
cluster
1     2855
3     1771
10    1477
8     1211
2     1136
25     974
5      786
4      734
22     689
15     673
21     673
16     640
6      635
7      627
23     608
20     483
0      413
13     413
27     407
19     397
12     368
11     357
17     340
28     318
30     278
24     232
18     230
9      225
14     213
29     165
31     148
26     144
Name: count, dtype: int64
```

Pr(X=0) = 413/20,620 ~= 0.02

jupyter 01-kMeans.MovieXGenre (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

Out[9]:
```
KMeans
KMeans(n_clusters=32)
```

In [ ]:

In [10]:
```
cluster_labels = cluster_model.predict(df)
movie_cluster_df = pd.DataFrame(cluster_labe
```

In [11]:
```
movie_cluster_df["cluster"].value_counts()
```

Out[11]:

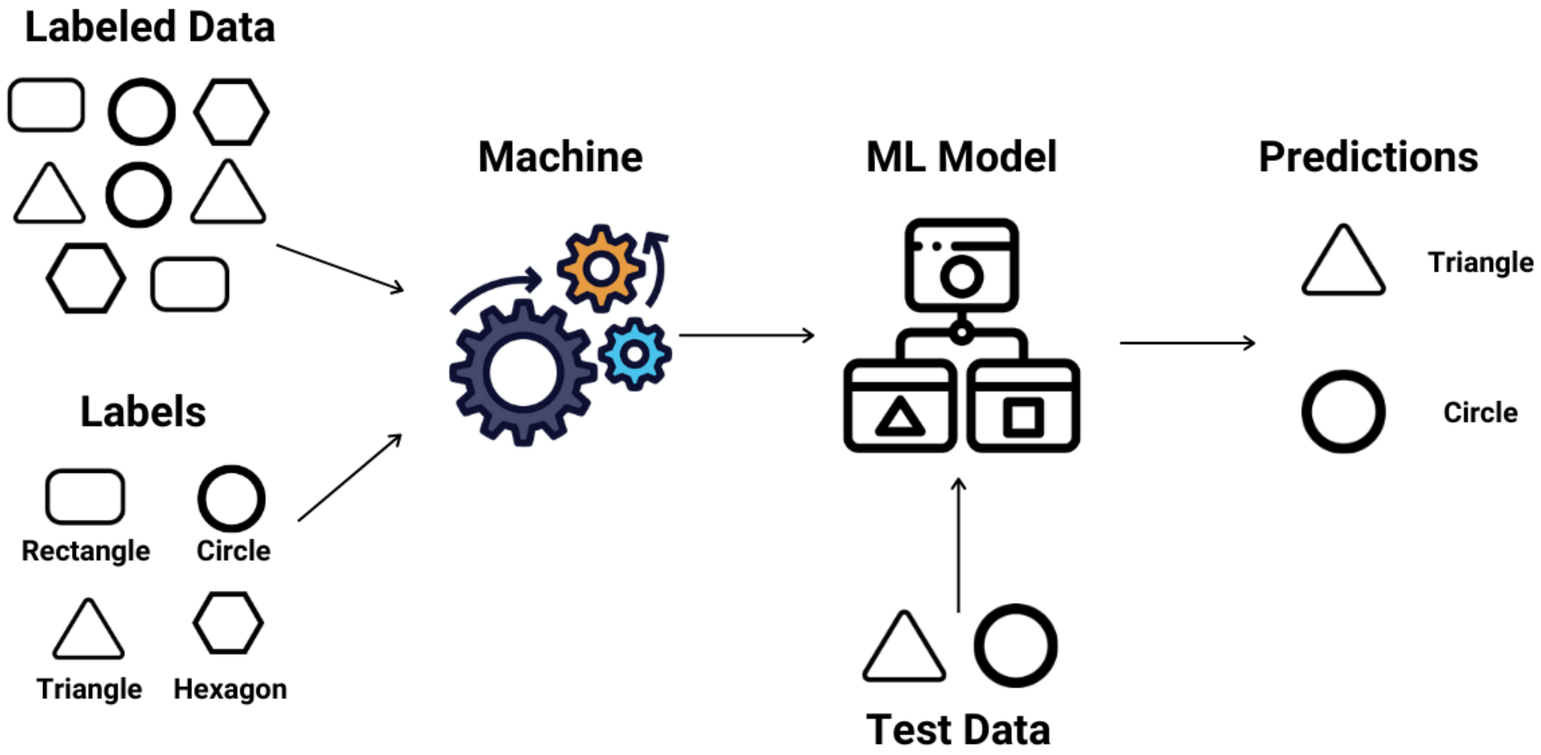
cluster	
1	2855
3	1771
10	1477
8	1211
2	1136
25	974
5	786
4	734
22	689
15	673
21	673
16	640
6	635
7	627
23	608
20	483
0	413
13	413
27	407
19	397
12	368
11	357
17	340
28	318
30	278
24	232
18	230
9	225
14	213
29	165
31	148
26	144

Name: count, dtype: int64

Pr(X=1) = 2,855/20,620 ~= 0.14

Supervised Learning

enjoy algorithms



a new movie?

ly cluster is C1

What is the most likely cluster given a new movie?

Pr(X=1) ~= 0.14, most likely cluster is C1

What if we know the genre of the film?

What if we know the genre of the film?

$$Pr(X|Genre = \ Sci\text{-}Fi \ )$$

Cluster	Probability
5	0.378679
19	0.284500
2	0.206017
21	0.139307
14	0.131458
1	0.119686
23	0.098103
22	0.092217
3	0.092217
8	0.084369
7	0.080445

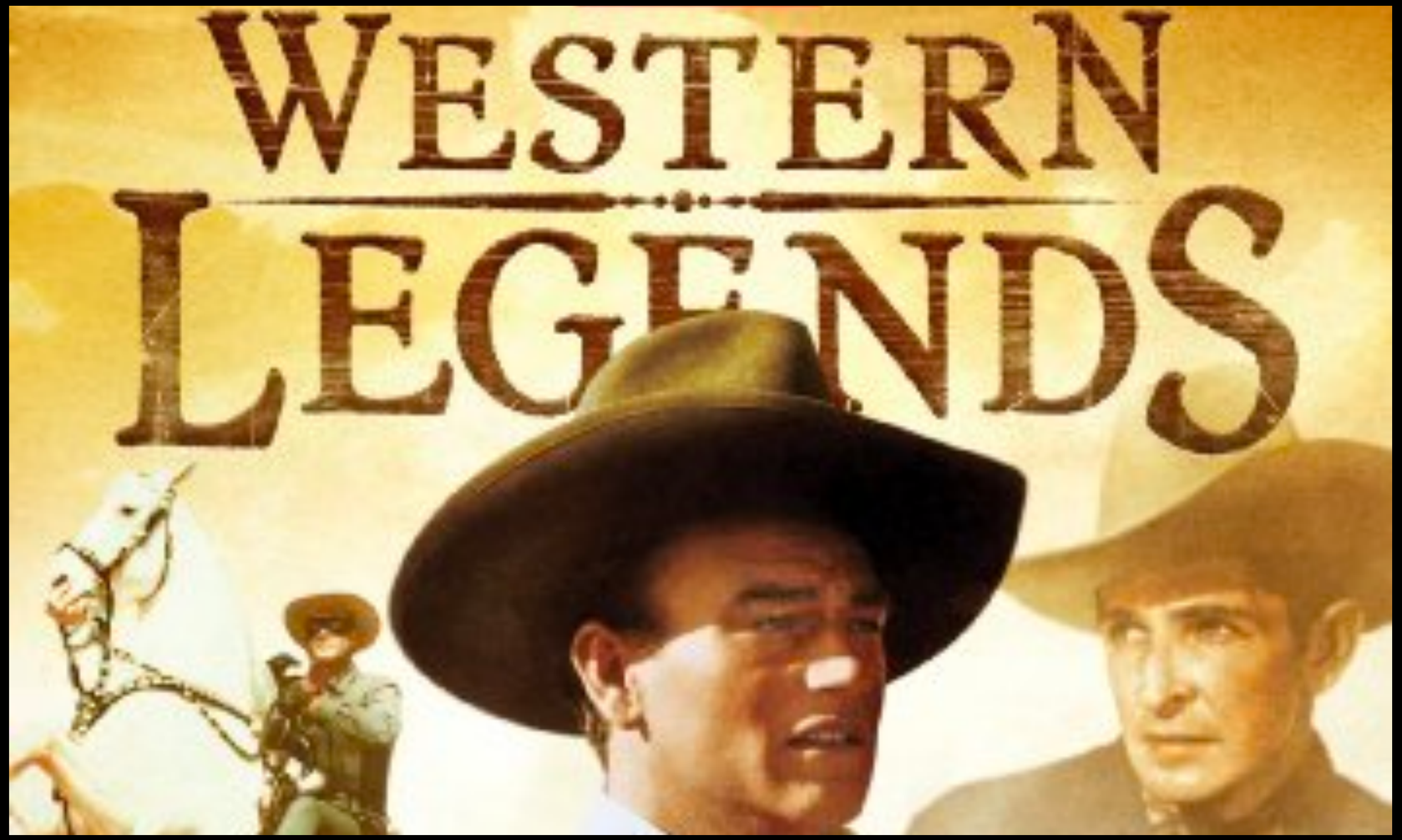
What if we know the genre of the film?

$$Pr(X|Genre = Sport)$$

Cluster	Pr
19	0.398976
1	0.398976
20	0.248773
3	0.215916
4	0.215916
15	0.150203
10	0.126734
17	0.093877
16	0.061020
25	0.061020

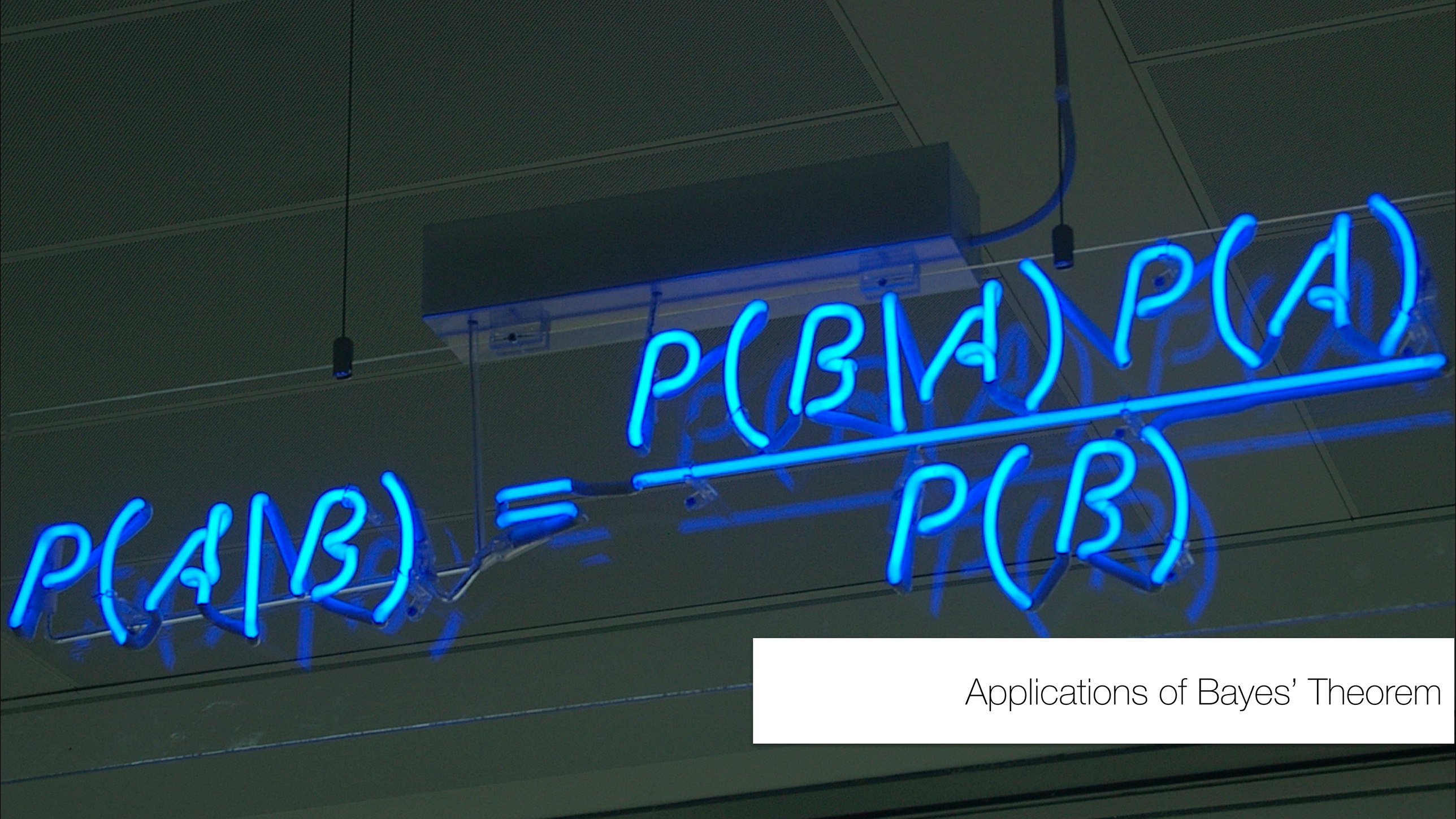
What if we know the genre of the film?

$$Pr(X|Genre = Western)$$



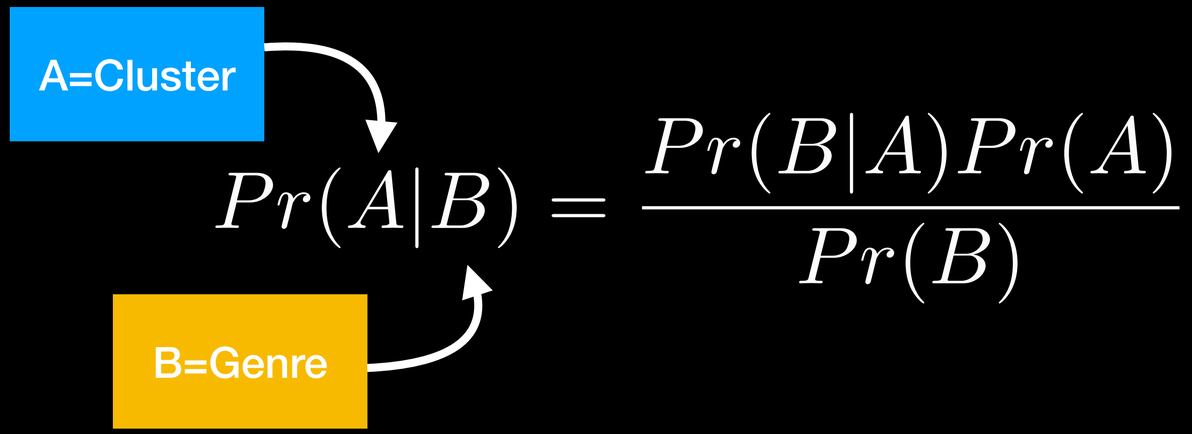
Cluster	Probability
19	0.843689
1	0.304120
5	0.166776
15	0.098103
2	0.088293
22	0.088293
25	0.068672
3	0.068672
17	0.049052
8	0.049052

Classification method based on **maximum likelihood estimation**



Applications of Bayes' Theorem

What is the most likely cluster given a movie in a genre or with an actor?



$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

$$Pr(Cluster = X|Genre = Y) = \frac{Pr(Genre = Y|Cluster = X)Pr(Cluster = X)}{Pr(Genre = Y)}$$

$$= \frac{Pr(Genre = Y, Cluster = X)}{Pr(Genre = Y)}$$

You can do this…

$$Pr(Cluster = X|Genre = Y) = \frac{Pr(Genre = Y|Cluster = X)Pr(Cluster = X)}{Pr(Genre = Y)}$$

$$Pr(Cluster = X|Genre = Y) \cdot Pr(Genre = Y) = Pr(Genre = Y|Cluster = X) \cdot Pr(Cluster = X)$$

Because this…

$$=$$

$$Pr(Cluster = X, Genre = Y)$$

Into which cluster will a Sports movie go?

$$Pr(Cluster = X | Genre = Sport)$$

$$= \frac{Pr(Genre = Sport | Cluster = X) Pr(Cluster = X)}{Pr(Genre = Sport)}$$

$$\text{argmax}_{X \in Clusters} Pr(Cluster = X | Genre = Sport)$$

Probability of a specific cluster given Sport genre...

Really want this:

$$\text{argmax}_{X \in Clusters} Pr(Cluster = X | Genre = Sport)$$

Really want this:

Maximize Over:

$$= \frac{Pr(Genre = Sport | Cluster = 1) Pr(Cluster = 1)}{Pr(Genre = Sport)}$$

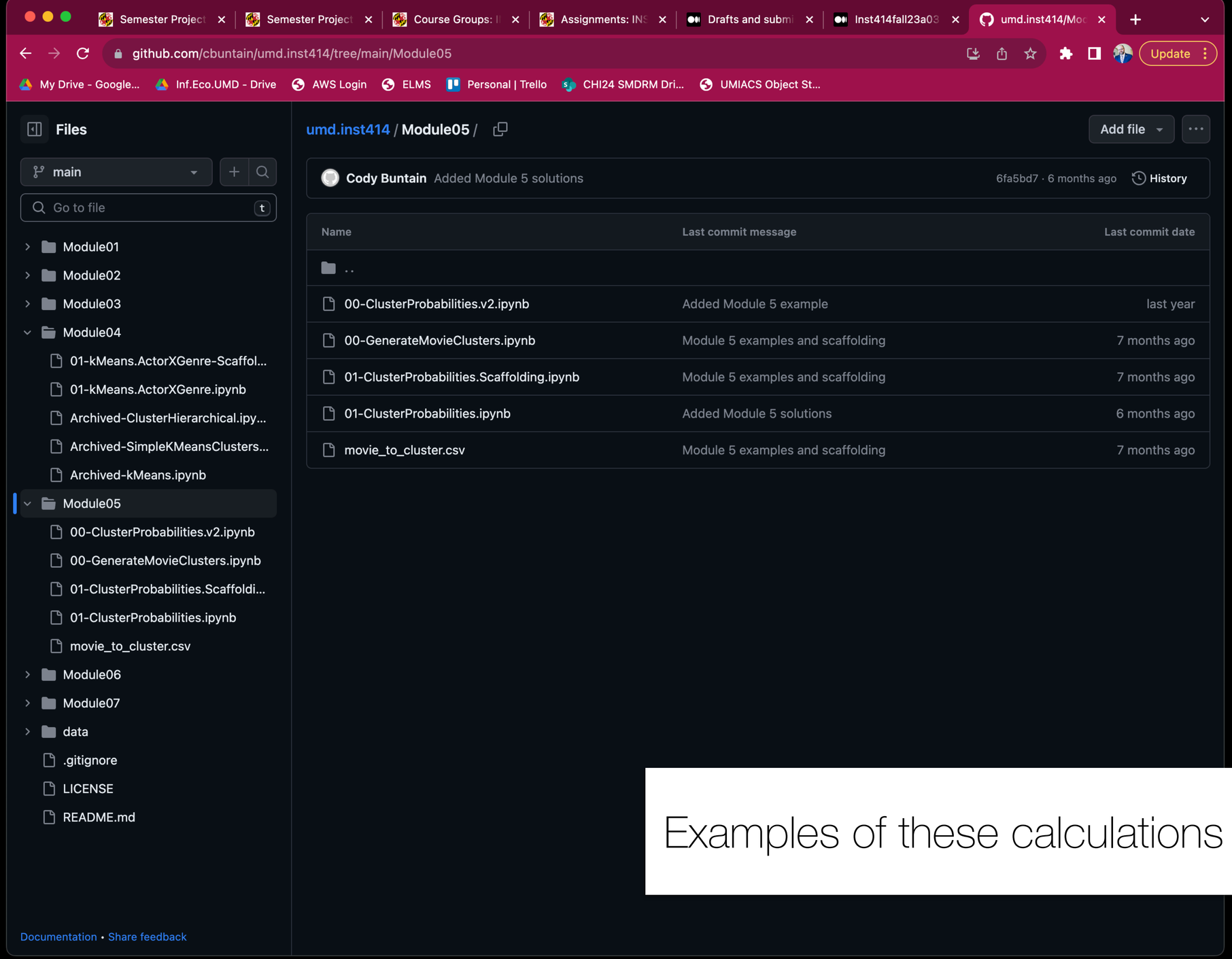
$$= \frac{Pr(Genre = Sport | Cluster = 2) Pr(Cluster = 2)}{Pr(Genre = Sport)}$$

...

$$= \frac{Pr(Genre = Sport | Cluster = C_n) Pr(Cluster = C_n)}{Pr(Genre = Sport)}$$

Each value has the same denominator

Given some genre, we can determine most probable cluster



Examples of these calculations are in GitHub

Into which cluster will a Sci-Fi movie go?

In [20]:
```
for cluster_id,cluster_genre_pr in enumerate(per_cluster_prs):

    pr_cluster_given_genre = cluster_genre_pr / genre_pr_map[target_genre]

    print("Pr[Cluster %02d | %s]:" % (cluster_id, target_genre), "\t", pr_cluster_given_genre)
```

```
Pr[Cluster 00 | Sci-Fi]:         0.4345047923322684
Pr[Cluster 01 | Sci-Fi]:         0.0031948881789137383
Pr[Cluster 02 | Sci-Fi]:         0.0
Pr[Cluster 03 | Sci-Fi]:         0.0
Pr[Cluster 04 | Sci-Fi]:         0.0
Pr[Cluster 05 | Sci-Fi]:         0.012779552715654953
Pr[Cluster 06 | Sci-Fi]:         0.028753993610223644
Pr[Cluster 07 | Sci-Fi]:         0.012779552715654953
Pr[Cluster 08 | Sci-Fi]:         0.015974440894568693
Pr[Cluster 09 | Sci-Fi]:         0.006389776357827477
Pr[Cluster 10 | Sci-Fi]:         0.47284345047923326
Pr[Cluster 11 | Sci-Fi]:         0.012779552715654953
Pr[Cluster 12 | Sci-Fi]:         0.0
Pr[Cluster 13 | Sci-Fi]:         0.0
Pr[Cluster 14 | Sci-Fi]:         0.0
Pr[Cluster 15 | Sci-Fi]:         0.0
```

Can think of this process as a supervised classification task

Movies' cluster assignments are labels

"Maximize pr of cluster label given a genre"
is our supervised model

What questions do you have?

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