

---

# INFO 2950: Intro to Data Science

(no whiteboards today!)

Lecture 24  
2023-11-20

---

# Agenda

1. Text data
  - a. SVD
  - b. TF-IDF
  - c. Word Embeddings
2. Experimentation!
  - a. Data streaming
  - b. A/B Tests
  - c. Bandits

# SVD: movie “concepts”

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

Matrix	Alien	Serenity	Casablanca	Amelie
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

$$=$$

0.13	0.02	-0.01
0.41	0.07	-0.03
0.55	0.09	-0.04
0.68	0.11	-0.05
0.15	-0.59	0.65
0.07	-0.73	-0.67
0.07	-0.29	0.32

$$\times$$

12.4	0	0
0	9.5	0
0	0	1.3

$$\times$$

0.56	0.59	0.56	0.09	0.09
0.12	-0.02	0.12	-0.69	-0.69
0.40	-0.80	0.40	0.09	0.09

# SVD: movie “concepts”

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

$U$  = User-to-concept similarity matrix

Matrix	Alien	Serenity	Casablanca	Amelie
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

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

User 1  
 User 2  
 User 3  
 User 4  
 User 5  
 User 6  
 User 7

# SVD: movie “concepts”

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

Matrix	Alien	Serenity	Casablanca	Amelie
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

$$=$$

0.13	0.02	-0.01
0.41	0.07	-0.03
0.55	0.09	-0.04
0.68	0.11	-0.05
0.15	-0.59	0.65
0.07	-0.73	-0.67
0.07	-0.29	0.32

$\times$

12.4	0	0
0	9.5	0
0	0	1.3

$\times$

Movie 1	Movie 2	Movie 3	Movie 4	Movie 5
0.56	0.59	0.56	0.09	0.09
0.12	-0.02	0.12	-0.69	-0.69
0.40	-0.80	0.40	0.09	0.09

$V$  = Movie-to-concept similarity matrix

# SVD: movie “concepts”

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

Matrix	Alien	Serenity	Casablanca	Amelie
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

$$= \begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix}$$

$\Sigma$  = Concept matrix

Concept 1

Concept 2

Concept 3

$$\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix}$$

$$\begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.40 & -0.80 & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

## Last time: Goodreads data

A =

	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			

Which matrix would this represent?

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

	Concept 0	Concept 1	Concept 2	...	Concept 49
The Fault in our Stars	0.22	213.34	0.67		2.33
Life of Pi	0.46	4.66	0.56		3.34
To the Lighthouse	3.44	0.33	0.48		1.03
Invisible Man	1.45	0.90	115.59		1.55
...					



What matrix would this represent?

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

Book-to-concept matrix

	Concept 0	Concept 1	Concept 2	...	Concept 49
The Fault in our Stars	0.22	213.34	0.67		2.33
Life of Pi	0.46	4.66	0.56		3.34
To the Lighthouse	3.44	0.33	0.48		1.03
Invisible Man	1.45	0.90	115.59		1.55
...					

Cosine similarity (X, Y)

Cosine similarity (Target book, every other book)

	Concept 0	Concept 1	Concept 2	...	Concept 49
The Fault in our Stars	0.22	213.34	0.67		2.33
Life of Pi	0.46	4.66	0.56		3.34
To the Lighthouse	3.44	0.33	0.48		1.03
Invisible Man	1.45	0.90	115.59		1.55
...					

X

Y

# 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

Cosine similarity (X, Y)

Cosine similarity (Target book, every other book)

	Concept 0	Concept 1	Concept 2	...	Concept 49
The Fault in our Stars	0.22	213.34	0.67		2.33
Life of Pi	0.46	4.66	0.56		3.34
To the Lighthouse	3.44	0.33	0.48		1.03
Invisible Man	1.45	0.90	115.59		1.55
...					

X

Y

Cosine similarity (X, Y)

Cosine similarity (Target book, every other book)

	Concept 0	Concept 1	Concept 2	...	Concept 49
The Fault in our Stars	0.22	213.34	0.67		2.33
Life of Pi	0.46	4.66	0.56		3.34
To the Lighthouse	3.44	0.33	0.48		1.03
Invisible Man	1.45	0.90	115.59		1.55
...					

X

Y

## Closest books by cosine similarity

1.00	The Alchemist
0.94	Life of Pi
0.91	The Kite Runner
0.89	Memoirs of a Geisha
0.87	The Little Prince
0.86	A Thousand Splendid Suns
0.86	The Da Vinci Code (Robert Langdon, #2)
0.86	Eat, Pray, Love
0.86	The Time Traveler's Wife
0.86	Les Misérables

**Closest  
books by  
cosine  
similarity**

1.00	A Portrait of the Artist as a Young Man
0.99	Dubliners
0.97	The Sound and the Fury
0.97	The Waste Land
0.97	To the Lighthouse
0.97	Waiting for Godot
0.97	Who's Afraid of Virginia Woolf?
0.96	Notes from Underground
0.96	The Death of Ivan Ilych
0.96	Absalom, Absalom!

# Closest books by cosine similarity

1.00	1984
0.99	Animal Farm
0.98	Brave New World
0.98	Lord of the Flies
0.97	Fahrenheit 451
0.96	The Catcher in the Rye
0.95	Of Mice and Men
0.95	Slaughterhouse-Five
0.94	Frankenstein
0.94	The Great Gatsby



## Closest books by cosine similarity

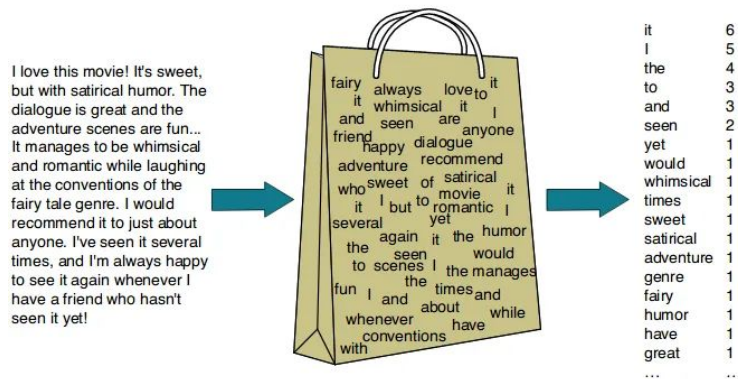
1.00	Parable of the Sower (Earthseed, #1)
0.94	Dawn (Xenogenesis, #1)
0.92	Kindred
0.88	The Sparrow (The Sparrow, #1)
0.85	Oryx and Crake (MaddAddam, #1)
0.84	The Left Hand of Darkness
0.84	The Dispossessed
0.83	Pattern Recognition (Blue Ant, #1)
0.82	The Year of the Flood (MaddAddam, #2)
0.81	Among Others

## Closest books by cosine similarity

1.00	Invisible Man
0.97	Native Son
0.96	As I Lay Dying
0.96	The Sound and the Fury
0.95	The Autobiography of Malcolm X
0.95	The Jungle
0.95	The Sun Also Rises
0.95	Survival in Auschwitz
0.95	Go Tell It on the Mountain
0.94	A Good Man is Hard to Find and Other Stories

# Working with text data

The “naive” approach: **bag-of-words**

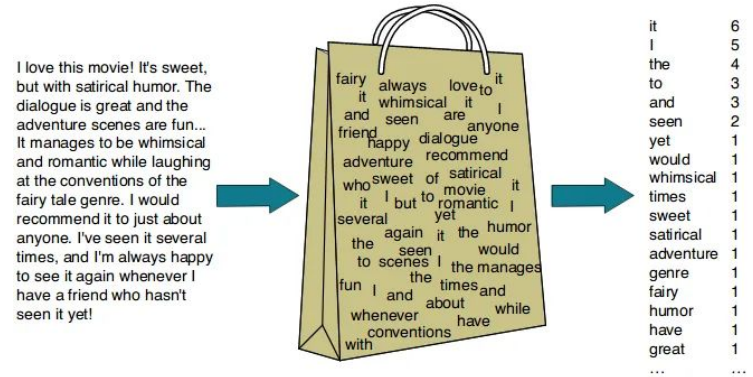


# Working with text data

The “naive” approach: **bag-of-words**

When does a naive approach fail?

1. Extremely common words may have more impact than we want
2. Slight differences in language (e.g. negation) make less of a difference
3. People use language creatively! (sarcasm, figurative language, idioms)



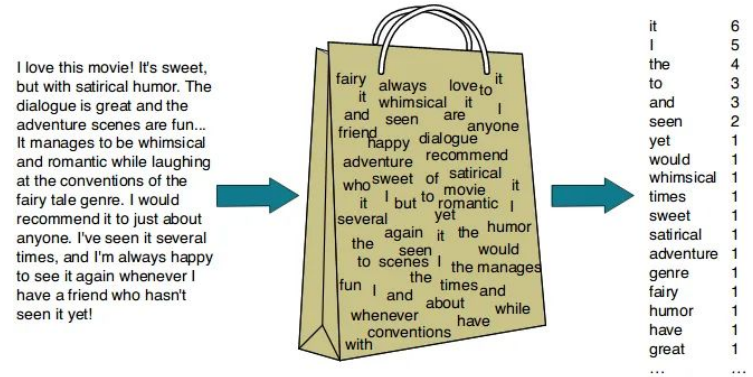
# Working with text data

The “naive” approach: **bag-of-words**

## When does a naive approach fail?

1. Extremely common words may have more impact than we want
2. Slight differences in language (e.g. negation) make less of a difference
3. People use language creatively! (sarcasm, figurative language, idioms)

*I do **not** love this movie. It wasn't sweet, with satirical humor. The dialogue could have been great or the adventure scenes fun...*



# Working with text data

The “naive” approach: **bag-of-words**

Are there alternatives to the bag-of-words approach?



---

## Are there alternatives to the bag-of-words approach?

### Alternative #1: Term Frequency-Inverse Document Frequency (TF-IDF)

- Problem: Some words are extremely frequent (e.g. “the”, “of”, “and”)
  - Learning that these words are common in a document isn’t all that interesting

### Alternative #2: Word Embeddings

---

## Are there alternatives to the bag-of-words approach?

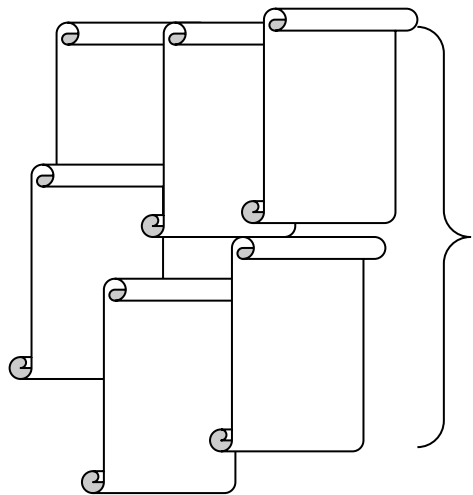
### Alternative #1: Term Frequency-Inverse Document Frequency (TF-IDF)

- Problem: Some words are extremely frequent (e.g. “the”, “of”, “and”)
  - Learning that these words are common in a document isn’t all that interesting
- TF-IDF is a way of measuring how important a word is to a **document** in a **collection** or “**corpus**”, adjusting for general word frequency
  - *What are the most distinctive words in this document?*



---

# Collections & documents

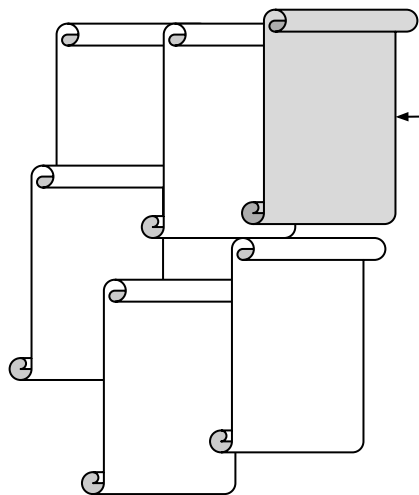


A **collection**  
(of books, tweets,  
Goodreads reviews,  
court opinions, etc.)

- Some words are *always* frequent (“it,” “and,” “of”)
- Other words are frequent in a collection (e.g. a collection of book reviews might often reference the words “book,” “review,” or “star”)

---

# Collections & documents



A **document** (one book, tweet, Goodreads review, court opinion, etc.)

Words that are uncommon in the **collection** but frequent in the **document** give us insight into what makes that document special

---


# Calculating TF-IDF

**TF-IDF = Term Frequency \* Inverse Document Frequency**

---

# Calculating TF-IDF

TF-IDF = **Term Frequency** \* Inverse Document Frequency



**Term Frequency** =  
number of times a given  
term appears in  
document / total words  
in document

---

# Calculating TF-IDF

TF-IDF = **Term Frequency** \* Inverse Document Frequency

**Term Frequency** =  
number of times a given  
term appears in  
document / total words  
in document

Rows: every  
**document** in  
collection


Columns: every word in **collection**

	the	cat	is	...	Total
Doc 0	3	0	2		20
Doc 1	3	3	3		30
Doc 2	2	0	1		10
...					

---

# Calculating TF-IDF

TF-IDF = **Term Frequency** \* Inverse Document Frequency

 **Term Frequency** =  
number of times a given  
term appears in  
document / total words  
in document

TF of “cat” in Doc 0 = 0 / 20

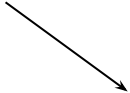
TF of “cat” in Doc 1 = 3/30

	the	cat	is	...	Total
Doc 0	3	0	2		20
Doc 1	3	3	3		30
Doc 2	2	0	1		10
...					

---

# Calculating TF-IDF

TF-IDF = Term Frequency \* **Inverse Document Frequency**

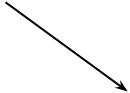


**Inverse Document Frequency** =  
log(Total Number of Documents /  
Number of Documents with Term)

---

# Calculating TF-IDF

**TF-IDF = Term Frequency \* Inverse Document Frequency**



**Inverse Document Frequency =**  
 $\log(\text{Total Number of Documents} / \text{Number of Documents with Term})$

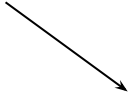
**Total Number of Documents =**  
number of documents (rows) in the  
entire collection



---

# Calculating TF-IDF

**TF-IDF = Term Frequency \* Inverse Document Frequency**



**Inverse Document Frequency =**  
log(Total Number of Documents /  
**Number of Documents with Term**)

**Total Number of Documents =**  
number of documents (rows) in the  
entire collection

**Number of Documents with Term =**  
for every word, the number of  
documents that have at least one  
instance of word

---

# Calculating TF-IDF

TF-IDF = Term Frequency \* **Inverse Document Frequency**

**Why take the inverse** (“flipped fraction”) **of document frequency?**

Inverse Document Frequency =  
 $\log(\text{Total Number of Documents} / \text{Number of Documents with Term})$

**Regular**

N docs with term

---

Total N docs

**Inverse**

Total N docs

---

N docs with term

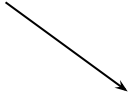
---

# Calculating TF-IDF

**TF-IDF = Term Frequency \* Inverse Document Frequency**

**Why take the inverse (“flipped fraction”) of document frequency?**

→ To boost the rarer words that occur in relatively few documents



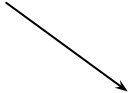
Inverse Document Frequency =  
 $\log(\text{Total Number of Documents} / \text{Number of Documents with Term})$

---

# Calculating TF-IDF

TF-IDF = Term Frequency \* Inverse Document Frequency

Taking the **log** smooshes our IDF so that small differences don't have an enormous effect



Inverse Document Frequency = **log**(Total Number of Documents / Number of Documents with Term)

---

# Calculating TF-IDF

	the	cat	is	...	Total
Doc 0	3	0	2		20
Doc 1	3	3	3		30
Doc 2	2	0	1		10

Inverse Document Frequency =  
 $\log(\text{Total Number of Documents}) /$   
 $(\text{Number of Documents with Term})$

## IDF of “cat”

Total number of documents = 3

Number of documents with “cat” = 1

$$\begin{aligned}\text{IDF} &= \log(3/1) \\ &= \log(3) \\ &= 1.099\end{aligned}$$

---

# Calculating TF-IDF

	the	cat	is	...	Total
Doc 0	3	0	2		20
Doc 1	3	3	3		30
Doc 2	2	0	1		10

Inverse Document Frequency =  
 $\log(\text{Total Number of Documents}) /$   
 $(\text{Number of Documents with Term})$

## IDF of “the”

Total number of documents = 3

Number of documents with “the” = 3

$$\begin{aligned}\text{IDF} &= \log(3/3) \\ &= \log(1) \\ &= 0\end{aligned}$$

# Calculating TF-IDF

	the	cat	is	...	Total
<b>Doc 0</b>	<b>3</b>	<b>0</b>	<b>2</b>		<b>20</b>
Doc 1	3	3	3		30
Doc 2	2	0	1		10

IDF of “cat” = 1.099  
IDF of “the” = 0

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

**Doc 0**

TF-IDF for “cat” =  $(0/20) * 1.099$   
= 0

TF-IDF for “the” =  $(3/20) * 0$   
= 0

# Calculating TF-IDF

	the	cat	is	...	Total
Doc 0	3	0	2		20
<b>Doc 1</b>	<b>3</b>	<b>3</b>	<b>3</b>		<b>30</b>
Doc 2	2	0	1		10

IDF of "cat" = 1.099  
IDF of "the" = 0

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

## **Doc 1**

$$\begin{aligned}\text{TF-IDF for "cat"} &= (3/30) * 1.099 \\ &= 0.1 * 1.099 \\ &= \mathbf{0.1099}\end{aligned}$$

$$\begin{aligned}\text{TF-IDF for "the"} &= (3/30) * 0 \\ &= 0.1 * 0 \\ &= \mathbf{0}\end{aligned}$$



---

# Calculating TF-IDF

	the	cat	is	...	Total
Doc 0	3	0	2		20
<b>Doc 1</b>	<b>3</b>	<b>3</b>	<b>3</b>		<b>30</b>
Doc 2	2	0	1		10

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

## **Doc 1**

TF-IDF for “cat” = 0.1099

TF-IDF for “the” = 0

IDF of “cat” = 1.099  
IDF of “the” = 0

**Same TF, different IDF! “Cat” is weighted as more important because it is less frequent across the collection**

---

# Calculating TF-IDF

	the	cat	is	...	Total
Doc 0	3	0	2		20
<b>Doc 1</b>	<b>3</b>	<b>3</b>	<b>3</b>		<b>30</b>
Doc 2	2	0	1		10

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

## **Doc 1**

TF-IDF for “cat” = 0.1099

TF-IDF for “the” = 0

**IDF of “cat” = 1.099**

**IDF of “the” = 0**

**Intuition: it is much more interesting  
that “cat” shows up in this document**

---

# scikit-learn's TF-IDF

TF-IDF = Term Frequency \* **Inverse Document Frequency**

 **Inverse Document Frequency** =  
 $\log((1 + \text{Total Number of Documents}) / (\text{Number of Documents with Term} + 1)) + 1$

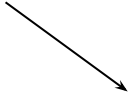
**Any guesses on why?**

**NOTE:** scikit-learn's IDF adds 1s to the numerator, denominator, and total IDF by default

---

# scikit-learn's TF-IDF

TF-IDF = Term Frequency \* **Inverse Document Frequency**



**Inverse Document Frequency** =  
 $\log((1 + \text{Total Number of Documents}) / (\text{Number of Documents with Term} + 1)) + 1$

**Any guesses on why?**

- Prevents zero division
- Helps avoid multiplying TF by zero

**NOTE:** scikit-learn's IDF adds 1s to the numerator, denominator, and total IDF by default

---

	the	cat	is	...	Total
Doc 0	3	0	2		20
Doc 1	3	3	3		30
Doc 2	2	0	1		10

**Inverse Document Frequency =**  
 $\log((1 + \text{Total Number of Documents}) / (\text{Number of Documents with Term} + 1)) + 1$

### IDF of “cat”

Total number of documents = 3

Number of documents with “cat” = 1

$$\begin{aligned}\text{IDF} &= \log((1 + 3) / (1 + 1)) + 1 \\ &= \log(2) + 1 \\ &= 1.693\end{aligned}$$

---

	the	cat	is	...	Total
Doc 0	3	0	2		20
Doc 1	3	3	3		30
Doc 2	2	0	1		10

**Inverse Document Frequency =**  
 $\log((1 + \text{Total Number of Documents}) / (\text{Number of Documents with Term} + 1)) + 1$

### IDF of “the”

Total number of documents = 3

Number of documents with “the” = 3

$$\begin{aligned}\text{IDF} &= \log((1 + 3) / (3 + 1)) + 1 \\ &= \log(1) + 1 \\ &= 1\end{aligned}$$

---

	the	cat	is	...	Total
<b>Doc 0</b>	<b>3</b>	<b>0</b>	<b>2</b>		<b>20</b>
Doc 1	3	3	3		30
Doc 2	2	0	1		10

**IDF of “cat” = 1.693**  
**IDF of “the” = 1**

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

**Doc 0**

TF-IDF for “cat” =  $(0/20) * 1.693$   
**= 0**

TF-IDF for “the” =  $(3/20) * 1$   
**= 0.15**

---

	the	cat	is	...	Total
Doc 0	3	0	2		20
<b>Doc 1</b>	<b>3</b>	<b>3</b>	<b>3</b>		<b>30</b>
Doc 2	2	0	1		10

IDF of “cat” = 1.693  
IDF of “the” = 1

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

**Doc 1**

$$\begin{aligned}\text{TF-IDF for “cat”} &= (3/30) * 1.693 \\ &= 0.1 * 1.693 \\ &= \mathbf{0.1693}\end{aligned}$$

$$\begin{aligned}\text{TF-IDF for “the”} &= (3/30) * 1 \\ &= 0.1 * 1 \\ &= \mathbf{0.1}\end{aligned}$$



---

	the	cat	is	...	Total
Doc 0	3	0	2		20
<b>Doc 1</b>	<b>3</b>	<b>3</b>	<b>3</b>		<b>30</b>
Doc 2	2	0	1		10

IDF of “cat” = 1.693  
IDF of “the” = 1

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

**Doc 1 (with 1s added)**  
TF-IDF for “cat” = **0.1693**  
TF-IDF for “the” = **0.1**

---

	the	cat	is	...	Total
Doc 0	3	0	2		20
<b>Doc 1</b>	<b>3</b>	<b>3</b>	<b>3</b>		<b>30</b>
Doc 2	2	0	1		10

IDF of “cat” = 1.693  
IDF of “the” = 1

**TF-IDF = Term Frequency \*  
Inverse Document Frequency**

**Doc 1 (original)**

TF-IDF for “cat” = 0.1099

TF-IDF for “the” = 0

**Doc 1 (with 1s added)**

TF-IDF for “cat” = 0.1693

TF-IDF for “the” = 0.1

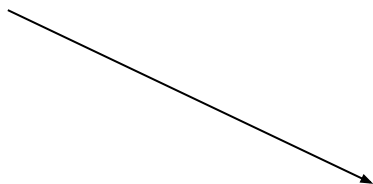
---

# TF-IDF in code

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
tfidf_vectorizer = TfidfVectorizer()
```

```
tfidf_vector = tfidf_vectorizer.fit_transform(text_files)
```



	Word 0	Word 1	Word 2
Doc 0	0.11	0.00	0.23
Doc 1	48.00	3.45	0.00
Doc 2	0.00	0.00	70.87
Doc 3	0.25	0.55	0.34

---

# Applying TF-IDF

Once you have TF-IDF scores for all words/documents in your collection, you can use them for many applications (e.g., clustering, classification)


	Word 0	Word 1	Word 2
Doc 0	0.11	0.00	0.23
Doc 1	48.00	3.45	0.00
Doc 2	0.00	0.00	70.87
Doc 3	0.25	0.55	0.34

---

# Applying TF-IDF

You can use TF-IDF  
weights as input (X) to a  
regression

X



	Word 0	Word 1	Word 2
Doc 0	0.11	0.00	0.23
Doc 1	48.00	3.45	0.00
Doc 2	0.00	0.00	70.87
Doc 3	0.25	0.55	0.34

---

# Applying TF-IDF

Input (X) = TF-IDF weights

Output (y) = binary 1 for spam, 0 for not spam

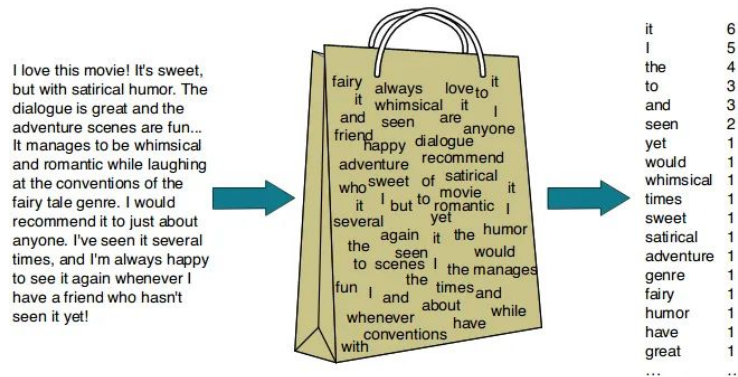
Example application:  
detect whether or not  
an email is spam  
based on TF-IDF  
scores

	X			y
	link	won	money	Spam
Email 0	0.11	0.00	0.23	0
Email 1	48.00	3.45	0.00	0
Email 2	0.00	0.00	70.87	1
Email 3	0.25	0.55	0.34	0

# Are there alternatives to the bag-of-words approach?

**Alternative #1: Term  
Frequency-Inverse Document  
Frequency (TF-IDF)**

**Alternative #2: Word Embeddings**



---

# Working with text data



“You shall know a word  
by the company it keeps!”

*J. R. Firth,  
A synopsis of linguistic theory (1957)*

<https://twitter.com/gianfrancocont9/status/1460130490015531008>

---



---

# Does context matter for word meaning?

- “You shall know a word by the company it keeps”
- A word’s meaning depends on its context (i.e. the words surrounding it)
- Many words have multiple meanings:
  - [financial] **bank**, [river] **bank**, [word] **bank**
  - [state] **fair**, [light] **fair**, [equitable] **fair**
  - [fruit] **date**, [time] **date**

---

# Does context matter for word meaning?

- “You shall know a word by the company it keeps”
- A word’s meaning depends on its context (i.e. the words surrounding it)
- Many words have multiple meanings:
  - [financial] **bank**, [river] **bank**, [word] **bank**
  - [state] **fair**, [light] **fair**, [equitable] **fair**
  - [fruit] **date**, [time] **date**

How can we capture a word’s meaning?

---

# Word embeddings

- A **word embedding model** learns associations between words, based on their usage (their context)
- Word embedding models are **trained** on enormous datasets (e.g. Wikipedia, books, news articles, webpages)
- Word embeddings help us differentiate between [financial] bank and [river] bank
- Word embeddings can also show which words are more/less similar in their usage

---

# Word embeddings

Every word is represented by a **vector**.

cat	0.1	-0.4	0.9	.01	0.2	0.3
-----	-----	------	-----	-----	-----	-----

---

# Word embeddings

Every word is represented by a **vector**.

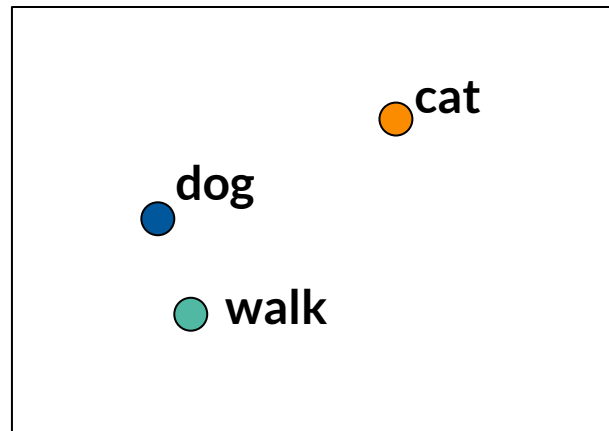
We can compare how different/similar words are by comparing their vectors.

cat	0.1	-0.4	0.9	.01	0.2	0.3
dog	0.5	0.9	0.3	.1	-0.1	0.2

---

# Word embeddings

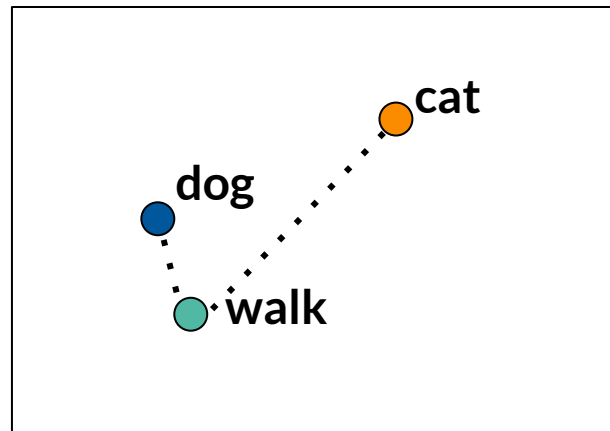
cat	0.1	-0.4	0.9	.01	0.2	0.3
dog	0.5	0.9	0.3	0.1	-0.1	0.2
walk	0.4	0.8	0.1	0.1	-0.1	0.2



---

# Word embeddings

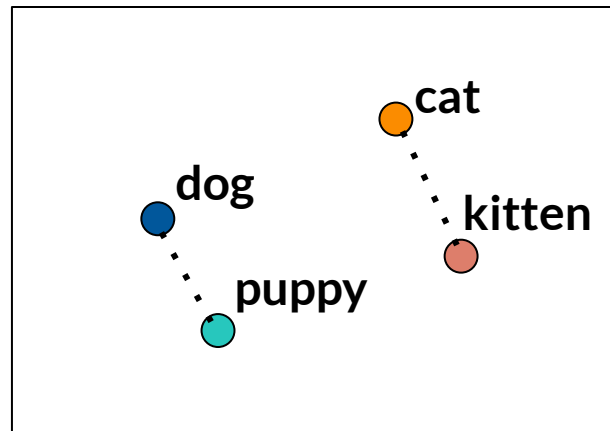
cat	0.1	-0.4	0.9	.01	0.2	0.3
dog	0.5	0.9	0.3	0.1	-0.1	0.2
walk	0.4	0.8	0.1	0.1	-0.1	0.2



We can compare which words are more similar

# Word embeddings

cat	0.1	-0.4	0.9	.01	0.2	0.3
dog	0.5	0.9	0.3	0.1	-0.1	0.2
kitten	0.2	-0.3	0.1	0.9	0.2	-.3
puppy	0.6	0.4	0.2	0.8	-0.1	-.4



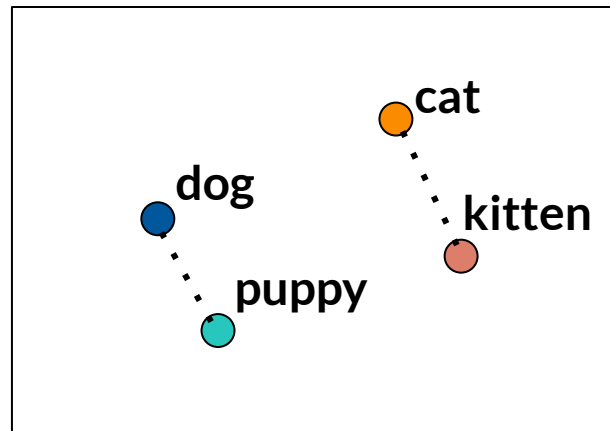
We can compare relationships  
between pairs of words



---

	<i>Barks</i>	<i>Goes on walks</i>	<i>Purrs</i>	<i>Small</i>	<i>Alive</i>	<i>Cute</i>
<b>cat</b>	0.1	-0.4	0.9	.2	0.5	0.9
<b>dog</b>	0.5	0.9	-.1	.01	0.6	0.9
<b>kitten</b>	0.2	-0.3	0.7	0.9	0.4	0.9
<b>puppy</b>	0.6	0.4	0.2	0.8	0.5	0.9

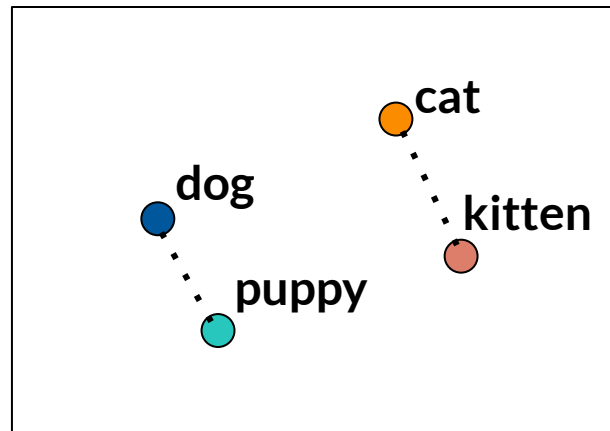
Each dimension of the vector represents some aspect of the word's meaning



---

	<i>Barks</i>	<i>Goes on walks</i>	<i>Purrs</i>	<i>Small</i>	<i>Alive</i>	<i>Cute</i>
cat	0.1	-0.4	0.9	.2	0.5	0.9
dog	0.5	0.9	-.1	.01	0.6	0.9
kitten	0.2	-0.3	0.7	0.9	0.4	0.9
puppy	0.6	0.4	0.2	0.8	0.5	0.9

Each dimension of the vector represents some aspect of the word's meaning



You can think of this as similar to SVD concepts!

---

# How does a word embedding model know what words are similar?

Generally, there are two options:

1. Train the word embeddings model on your own dataset
2. Load a **pre-trained** model
  - “**Pre-trained**” means that someone else already trained the model on an enormous dataset, like Wikipedia, news articles, court opinions, Google Books
  - The model learns associations between words in that dataset

Popular word embedding models: word2vec, GloVe

**Gensim** is a python library for accessing/using word embedding models

---

---

# Finding similar words

```
import gensim.downloader as api

wv = api.load('word2vec-google-news-300')

wv.most_similar("cat")

[('cats', 0.8099379539489746),
 ('dog', 0.760945737361908),
 ('kitten', 0.7464984655380249),
 ('feline', 0.7326233983039856),
 ('beagle', 0.7150582671165466),
 ('puppy', 0.7075453996658325),
 ('pup', 0.6934291124343872),
 ('pet', 0.6891531348228455),
 ('felines', 0.6755931377410889),
 ('chihuahua', 0.6709762215614319)]
```

---

# Finding similar words

```
import gensim.downloader as api  
wv = api.load('word2vec-google-news-300')  
wv.most_similar("cat")  
[('cats', 0.8099379539489746),  
 ('dog', 0.760945737361908),  
 ('kitten', 0.7464984655380249),  
 ('feline', 0.7326233983039856),  
 ('beagle', 0.7150582671165466),  
 ('puppy', 0.7075453996658325),  
 ('pup', 0.6934291124343872),  
 ('pet', 0.6891531348228455),  
 ('felines', 0.6755931377410889),  
 ('chihuahua', 0.6709762215614319)]
```

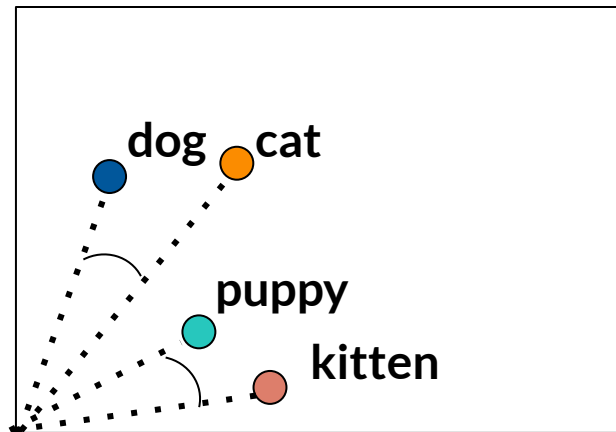
Word similarity based on  
Google News (~100  
billion words)

# Finding analogies

Dog: puppy :: cat: kitten

```
wv.most_similar_cosmul(positive = ["cat", "puppy"], negative = ["dog"])  
( 'kitten', 0.9333010911941528)
```

cat + puppy - dog = kitten



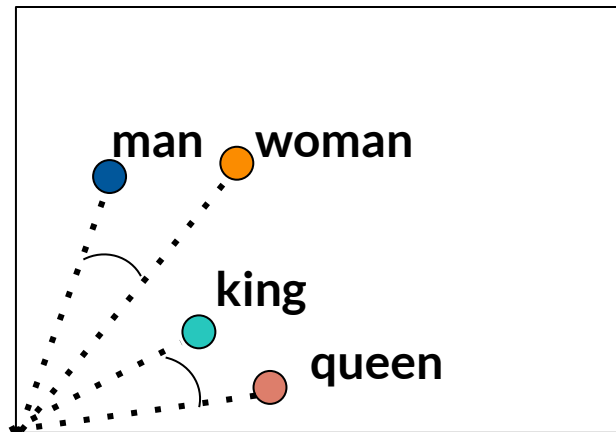
# Finding analogies

Man : king :: woman : queen

```
wv.most_similar_cosmul(positive = ["king", "woman"], negative = ["man"])
```

```
('queen', 0.9314123392105103),  
( 'monarch', 0.858533501625061),  
( 'princess', 0.8476566076278687),  
( 'Queen_Consort', 0.8150269985198975),  
( 'queens', 0.8099815249443054)
```

king - man + woman = queen



---

# Bias in Word Embeddings

Computer programmer - man + woman = ?

```
wv.most_similar_cosmul(  
    positive = ["computer_programmer", "woman"],  
    negative = ["man"])  
  
('homemaker', 0.8643969297409058),  
('paralegal', 0.8267658948898315),  
('registered_nurse', 0.8235622644424438),  
('housewife', 0.8165889978408813)
```

Computer programmer + man - woman = ?

```
wv.most_similar_cosmul(  
    positive = ["computer_programmer", "man"],  
    negative = ["woman"])  
  
(['mechanical_engineer', 0.8572883009910583),  
 ('programmer', 0.8219982981681824),  
 ('electrical_engineer', 0.8172740340232849),  
 ('engineer', 0.8136039972305298)
```



---

# Bias in Word Embeddings

- Word embedding models learn relationships between words from text
- Text like Wikipedia, books, news articles, etc. contain the biases of their creators
- Word embeddings learn **biased representations** of words/concepts in text

---

# 1 min break!



---

# Decision-making in the “real world”

- **Boss:** “Should we increase prices?”



---

# Decision-making in the “real world”

- Boss: “Should we increase prices?”
  - **Regression** (sales ~ prices)



---

# Decision-making in the “real world”



- **Boss:** “Should we increase prices?”
  - **Regression** (sales ~ prices)
- **Boss:** “We want to hire more employees at store location A if customers spend a longer time in location A than other locations. Should we?”

---

# Decision-making in the “real world”



- **Boss:** “Should we increase prices?”
  - **Regression** (sales ~ prices)
- **Boss:** “We want to hire more employees at store location A if customers spend a longer time in location A than other locations. Should we?”
  - **Hypothesis testing** (null  $H_0$ :  $\mu$  = [avg customer time in store, across locations])

---

# Decision-making in the “real world”



- Should we make “Tweet” button size bigger?
- Should we have videos autoplay in the feed?
- Should we change the ranking algorithm to prioritize controversial vs. recent content?

---

# Decision-making in the “real world”

- Should we switch from doing **A** to doing **B**?
  - Should we make “Tweet” button size bigger?
  - Should we have videos autoplay in the feed?
  - Should we change the ranking algorithm to prioritize controversial vs. recent content?



---

# Distill question into A and B

Question	A	B
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size

---

# Distill question into A and B

Question	A	B
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size
Should we have videos autoplay in the feed?		

---

# Distill question into A and B

Question	A	B
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size
Should we have videos autoplay in the feed?	Videos do not autoplay	Videos autoplay

---

# Distill question into A and B

Question	A	B
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size
Should we have videos autoplay in the feed?	Videos do not autoplay	Videos autoplay
Should we change the ranking algorithm to prioritize controversial vs. recent content?		

---

## Distill question into A and B

Question	A	B
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size
Should we have videos autoplay in the feed?	Videos do not autoplay	Videos autoplay
Should we change the ranking algorithm to prioritize controversial vs. recent content?	Prioritize controversial content on top of feed	Prioritize recent content on top of feed

---

---

# Life advice

- When you're trying to make a decision, explicitly write out the two detailed options you're deciding between

---

# Life advice

- When you're trying to make a decision, explicitly write out the two detailed options you're deciding between:
  - Procrastinate by doomscrolling for an hour
  - Study for an hour for the INFO2950 final on Dec. 10th at 2pm

---

# Decision-making in the “real world”

- Should we give a clinical patient Drug D?
  - What are A and B?



---

# Decision-making in the “real world”

- **Should we give a clinical patient Drug D?**
  - **A:** maybe no drugs, maybe existing set of medication, maybe a very low dose of Drug D?
  - **B:** Drug D

---

# Decision-making in the “real world”

- **Should we give a clinical patient Drug D?**
  - **A:** maybe no drugs, maybe existing set of medication, maybe a very low dose of Drug D?
  - **B:** Drug D
- **Lots of details to consider:** drug interactions, clinical history, efficacy for patient demographic

---

# Should we do A or B?

- Distill your decision into two “experiment arms”, A and B
  - This requires domain expertise!
- Now what?

---

# Should we do A or B?

- Distill your decision into two “experiment arms”, A and B
  - This requires domain expertise!
- Now what? **Do an A/B test!**

---

# A/B test (a.k.a. A/B experimentation)

- A/B test will inform you of whether arm A or arm B (or neither) performs better
  - This requires hypothesis testing! **What's the null?**

---

# A/B test (a.k.a. A/B experimentation)

- A/B test will inform you of whether arm A or arm B (or neither) performs better
  - $H_0$ : arm A is no different from arm B

---

# A/B test (a.k.a. A/B experimentation)

- A/B test will inform you of whether arm A or arm B (or neither) performs better
  - $H_0$ : arm A is no different from arm B
- Similar to **Randomized Controlled Trials (RCT)**, though generally done in different applications
  - A/B requires more assumptions, tends to be with online data that updates rapidly

---

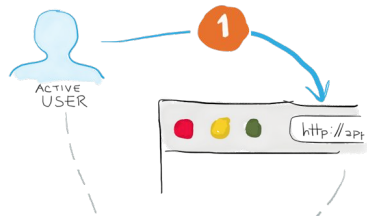
# Multiple meanings of “online”

- A/B experimentation is built for when **we don't know all the data in advance** (whereas we did in previous lectures on hypothesis testing)



---

# Multiple meanings of “online”

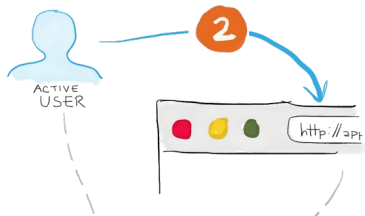
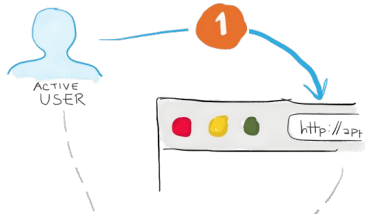


- A/B experimentation is built for when **we don't know all the data in advance** (whereas we did in previous lectures on hypothesis testing)
- To know whether users prefer a smaller or bigger button, we have to **wait for users to come to our website**, show them a button, and then make decisions

---

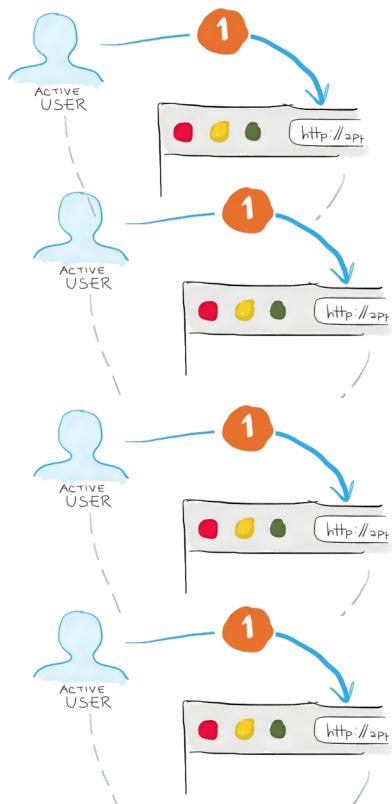
# Multiple meanings of “online”

- A/B experimentation is built for when **we don't know all the data in advance** (whereas we did in previous lectures on hypothesis testing)
- To know whether users prefer a smaller or bigger button, we have to **wait for users to come to our website**, show them a button, and then make decisions



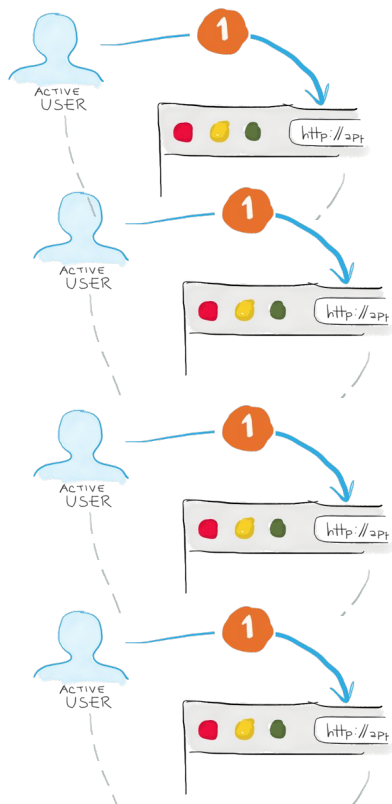
---

# Multiple meanings of “online”



- User data comes to us in “**data streams**” that we can’t ever store *all* of (dealing with this is a whole subfield of data science)

# Multiple meanings of “online”



- User data comes to us in “**data streams**” that we can’t ever store *all* of (dealing with this is a whole subfield of data science)
- In ML, this necessitates doing “Online Learning”
  - Train model on existing data
  - Do slow updates on your model (e.g. using SGD) as data comes in
- How does A/B testing work?

---

# How do we compare A and B?

Question	A	B
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size
Should we have videos autoplay in the feed?	Videos do not autoplay	Videos autoplay
Should we change the ranking algorithm to prioritize controversial vs. recent content?	Prioritize controversial content on top of feed	Prioritize recent content on top of feed

---

# How do we compare A and B?

- Can we have all users, as they come in, do **both** situations A and B **simultaneously**? (E.g., click both small and big buttons)?

---

# How do we compare A and B?

- Can we have all users, as they come in, do **both** situations A and B **simultaneously**? (E.g., click both small and big buttons)?
  - **No! A and B are mutually exclusive**

---

# How do we compare A and B?

- Can we have all users, as they come in, do both situations A and B **sequentially**? (E.g., click the small button and then the big button)?



---

# How do we compare A and B?

- Can we have all users, as they come in, do both situations A and B **sequentially**? (E.g., click the small button and then the big button)?
  - **No!** What if they only go to the website once?
  - **Plus**, you can't know if good performance is because of a delayed reaction to the first situation A, or only the second situation B

---

# How do we compare A and B?

- Can we have all users in arm A for a period of time, and then switch all users to arm B for the next period of time?

---

# How do we compare A and B?

- Can we have all users in arm A for a period of time, and then switch all users to arm B for the next period of time?
  - Technically, yes... but **this seems dangerous!**
  - What if B is a really bad and all your users quit?
  - Plus, time effect concerns: what if performance isn't about the experiment, but about the time?

---

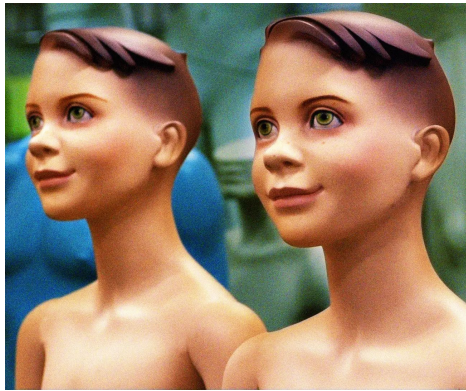
# Would be nice if we had 2 copies of every person!



- Users Koenecke\_1, Thalken\_1 → A
- Users Koenecke\_2, Thalken\_2 → B

---

# Would be nice if we had 2 copies of every person!



Mimno\_1

Mimno\_2

A

B

- Users Koenecke\_1, Thalken\_1 → A
- Users Koenecke\_2, Thalken\_2 → B
- Then, since the users are “the same” we can determine whether **A** or **B** is better

---

# Would be nice if we had 2 copies of every person!



- Users Koenecke\_1, Thalken\_1 → **A**
- Users Koenecke\_2, Thalken\_2 → **B**
- Then, since the users are “the same” we can determine whether **A** or **B** is better
- **Without clones, how do we simulate this?**

---

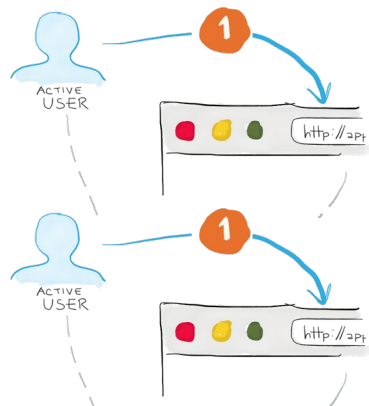
## Choosing who sees A, B

- What if we just **randomize** users so half of them see A, and half of them see B?
  - Without knowing more information about the users as they come in, this is pretty good!

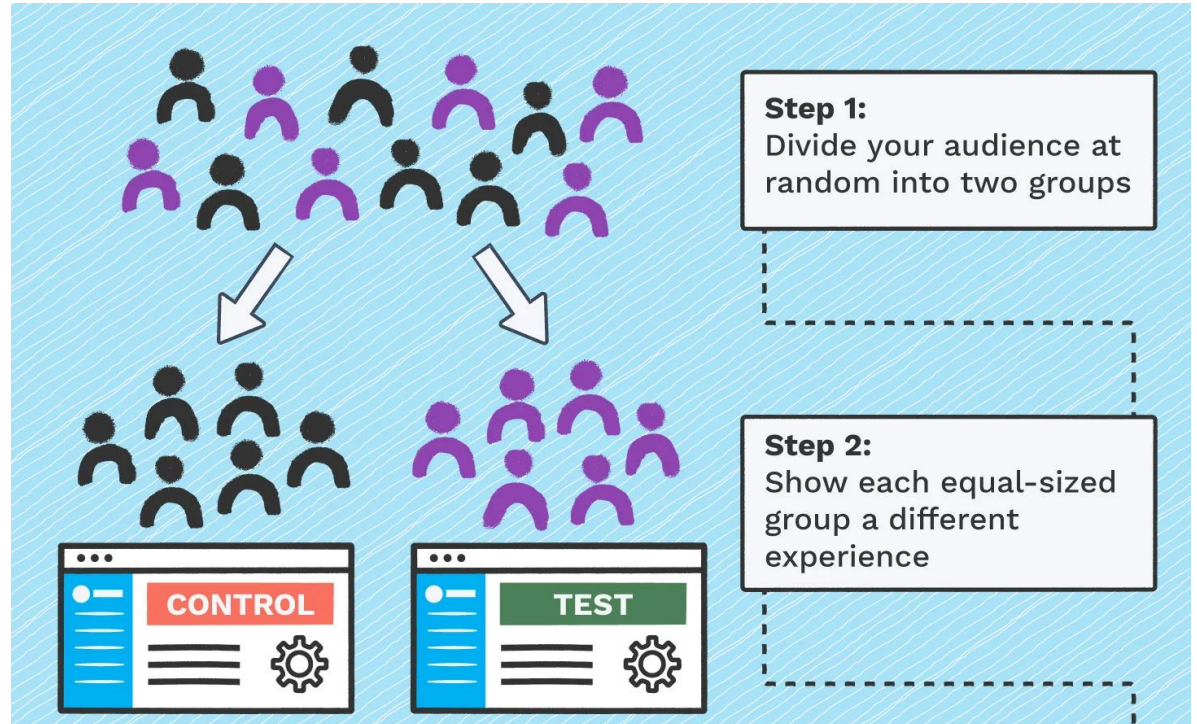
---

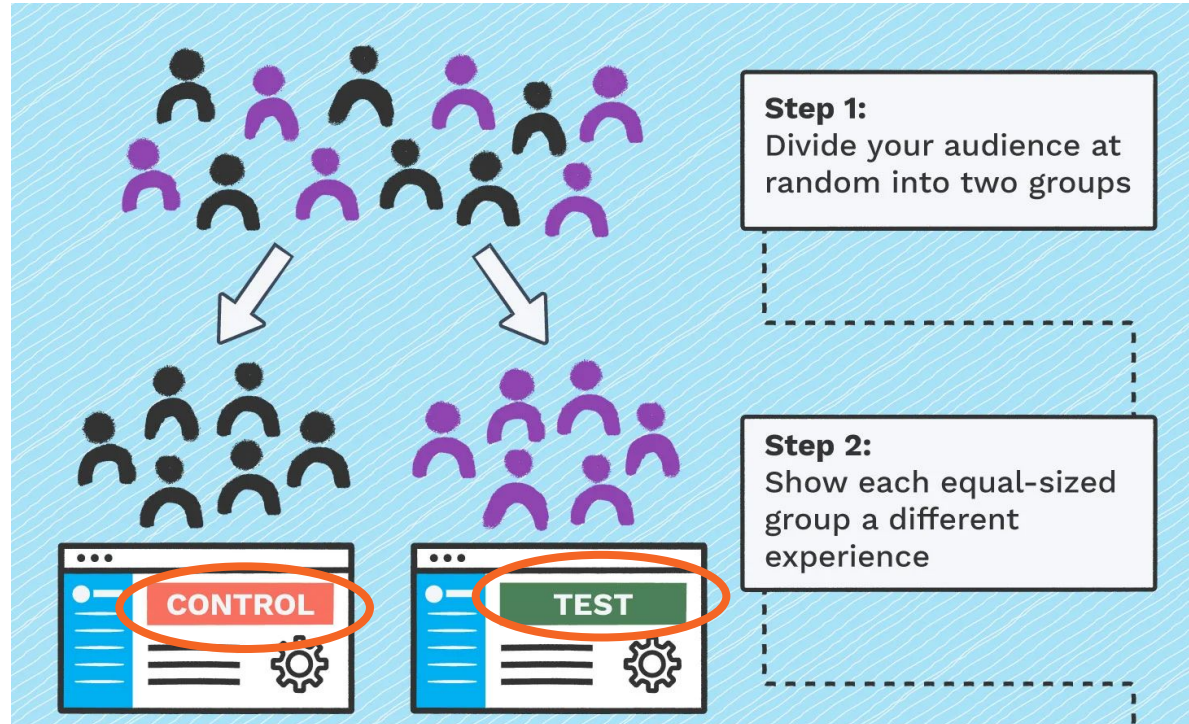
# Choosing who sees A, B

- What if we just **randomize** users so half of them see A, and half of them see B?
  - Without knowing more information about the users as they come in, this is pretty good!
- Data streaming: alternate users getting sent into A vs. B





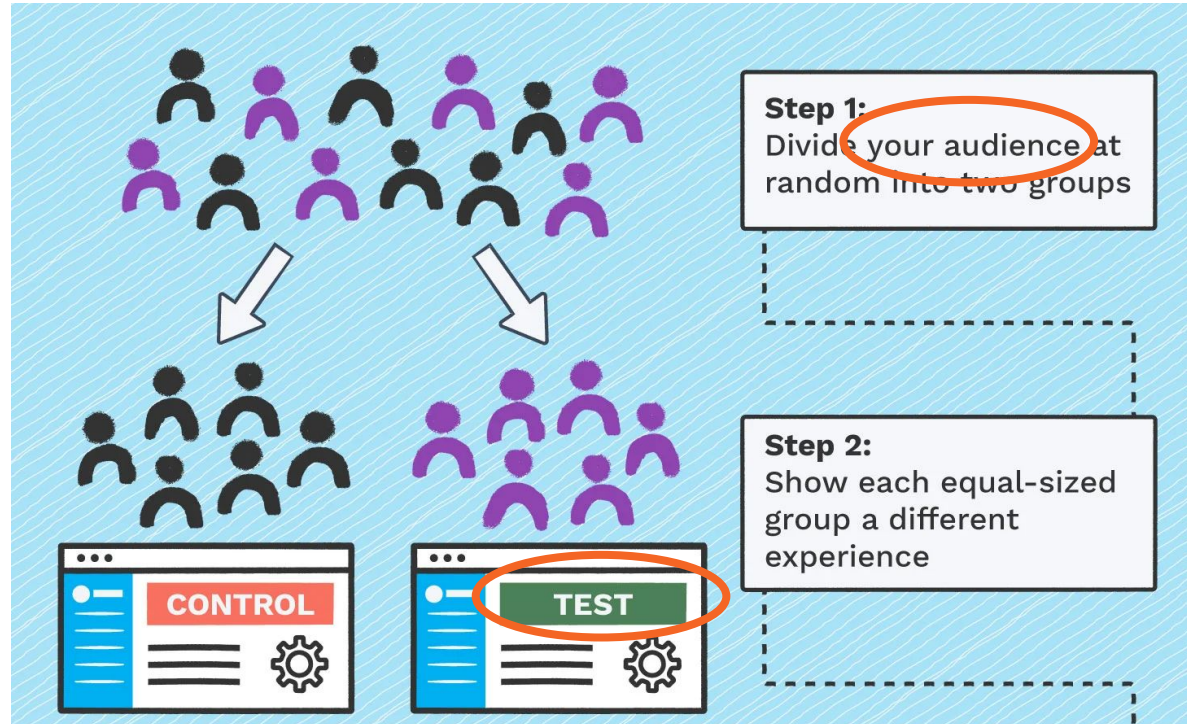




a.k.a.  
“A”

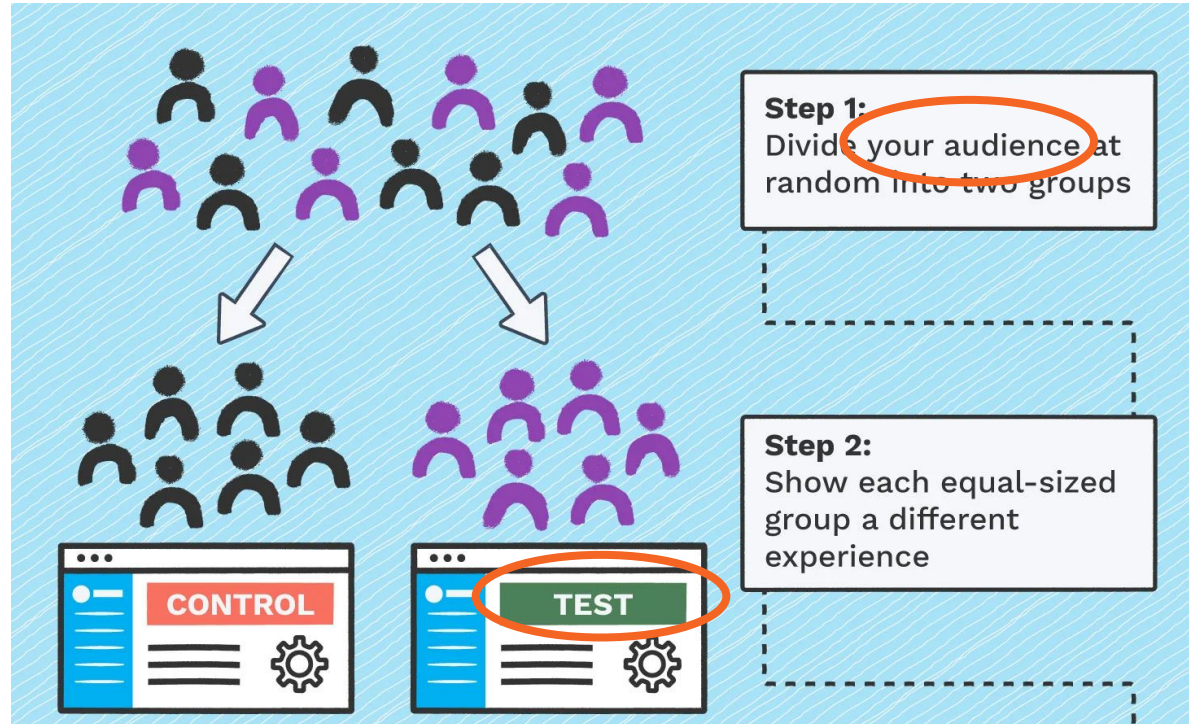
a.k.a. “B”, “variant”, “treatment”

## who's my audience in a user stream?

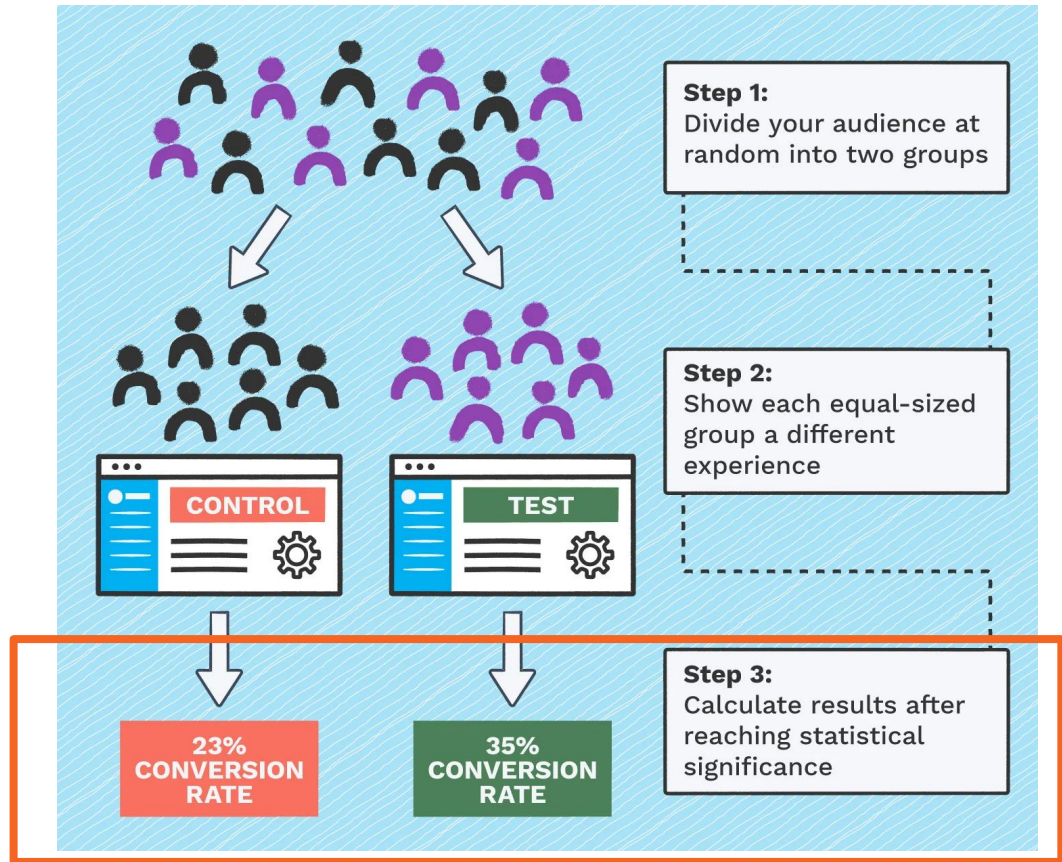




~ alternate users who show up

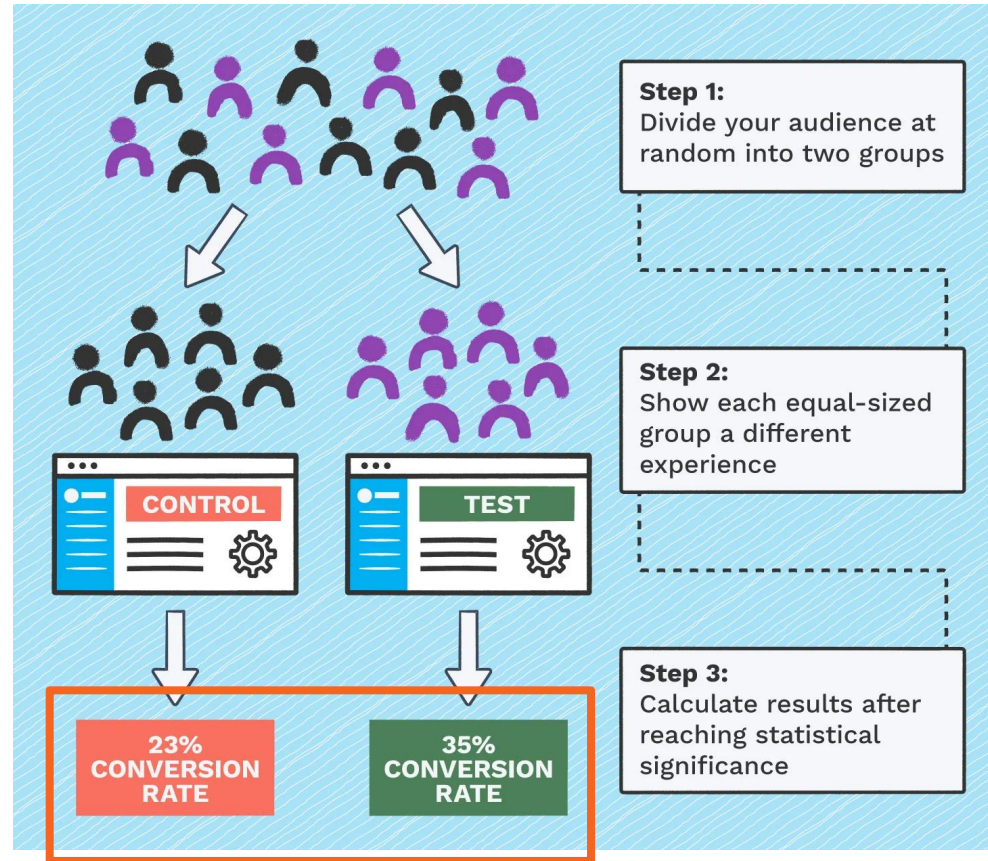


The “final step” of A/B testing,  
but what are we actually  
testing for?



The “final step” of A/B testing,  
but what are we actually  
testing for?

Choose your metric(s)!



---

# Common marketing metrics

- Website traffic
- Bounce rates
- Impressions
  - [Content, Ad] Consumption
- Click-through rates
- Conversion rates
- Cart abandonment
- Interactions

---

# Distill question into A and B

Question	A	B	Metric
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size	?
Should we have videos autoplay in the feed?	Videos do not autoplay	Videos autoplay	?
Should we change the ranking algorithm to prioritize controversial vs. recent content?	Prioritize controversial content on top of feed	Prioritize recent content on top of feed	?



---

# Distill question into A and B

Question	A	B	Metric
Should we make “Tweet” button size bigger?	Current (small) button size	New (bigger) button size	<ul style="list-style-type: none"><li>• # button clicks</li></ul>
Should we have videos autoplay in the feed?	Videos do not autoplay	Videos autoplay	<ul style="list-style-type: none"><li>• # seconds spent watching video?</li><li>• # interactions with video?</li></ul>
Should we change the ranking algorithm to prioritize controversial vs. recent content?	Prioritize controversial content on top of feed	Prioritize recent content on top of feed	<ul style="list-style-type: none"><li>• # seconds on platform?</li><li>• # interactions?</li><li>• # posts?</li><li>• \$ ad revenue?</li></ul>



We randomly hide toxic content on social media & find a 23% reduction in content consumption on Facebook and a 9% reduction in ad consumption on Twitter (beyond the intervention's mechanical effect), as well as a decrease in toxicity of content production:

[mstalinski.net](https://mstalinski.net)

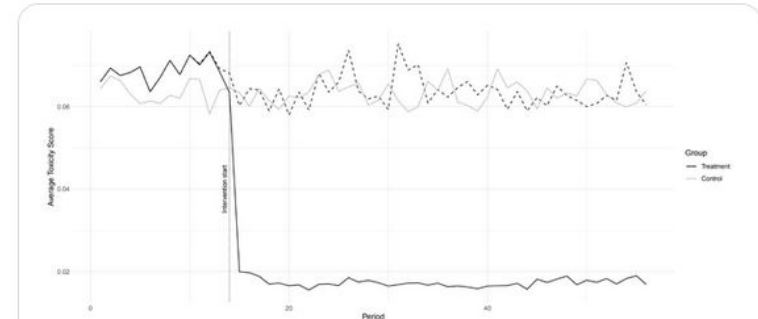


FIGURE 2: AVERAGE TOXICITY OF CONTENT SHOWN TO USERS DURING THE STUDY

Note: The figure depicts the average toxicity of posts, comments, and replies shown to users on each day of the study (relative to when a given participant started), separately for the control group and the treatment group. The dashed line for the treatment group demonstrates the average toxicity of elements that the platforms intended to show to the user before any hiding was applied by the intervention. The data presented here encompasses the three supported platforms (Twitter, Facebook, and YouTube). The dashed vertical line ("Intervention start") indicates day 15 – the first day of the intervention period.

3:17 PM · Nov 18, 2022 · Twitter Web App



Participants install a browser extension that, for some subset of participants, automatically filters out toxic content

Researchers quantify whether not seeing toxic content has an effect on participant behavior on FB, Twitter

We randomly hide toxic content on social media & find a 23% reduction in content consumption on Facebook and a 9% reduction in ad consumption on Twitter (beyond the intervention's mechanical effect), as well as a decrease in toxicity of content production:

[mstalinski.net](https://mstalinski.net)

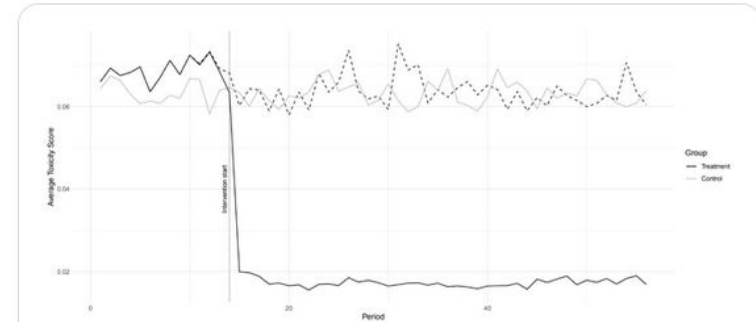


FIGURE 2: AVERAGE TOXICITY OF CONTENT SHOWN TO USERS DURING THE STUDY

Note: The figure depicts the average toxicity of posts, comments, and replies shown to users on each day of the study (relative to when a given participant started), separately for the control group and the treatment group. The dashed line for the treatment group demonstrates the average toxicity of elements that the platforms intended to show to the user before any hiding was applied by the intervention. The data presented here encompasses the three supported platforms (Twitter, Facebook, and YouTube). The dashed vertical line ("Intervention start") indicates day 15 – the first day of the intervention period.

3:17 PM · Nov 18, 2022 · Twitter Web App



Participants install a browser extension that, for some subset of participants, automatically filters out toxic content

Researchers quantify whether not seeing toxic content has an effect on participant behavior on FB, Twitter

What are A and B here?  
What are the relevant metrics?

We randomly hide toxic content on social media & find a 23% reduction in content consumption on Facebook and a 9% reduction in ad consumption on Twitter (beyond the intervention's mechanical effect), as well as a decrease in toxicity of content production:

[mstalinski.net](https://mstalinski.net)

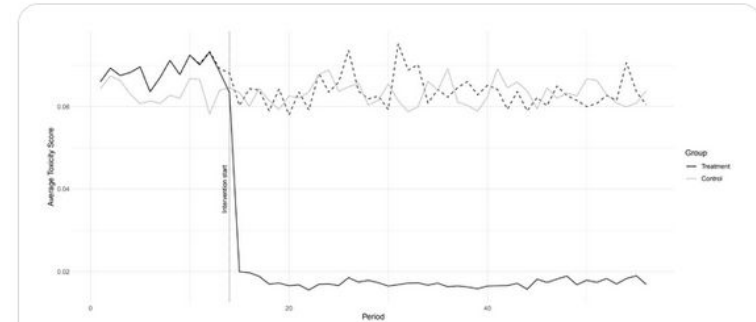


FIGURE 2: AVERAGE TOXICITY OF CONTENT SHOWN TO USERS DURING THE STUDY

Note: The figure depicts the average toxicity of posts, comments, and replies shown to users on each day of the study (relative to when a given participant started), separately for the control group and the treatment group. The dashed line for the treatment group demonstrates the average toxicity of elements that the platforms intended to show to the user before any hiding was applied by the intervention. The data presented here encompasses the three supported platforms (Twitter, Facebook, and YouTube). The dashed vertical line ("Intervention start") indicates day 15 – the first day of the intervention period.

3:17 PM · Nov 18, 2022 · Twitter Web App



## Facebook:

- A: show regular content
- B: hide toxic content
- Metric: content consumption

We randomly hide toxic content on social media & find a 23% reduction in content consumption on Facebook and a 9% reduction in ad consumption on Twitter (beyond the intervention's mechanical effect), as well as a decrease in toxicity of content production:

[mstalinski.net](https://mstalinski.net)

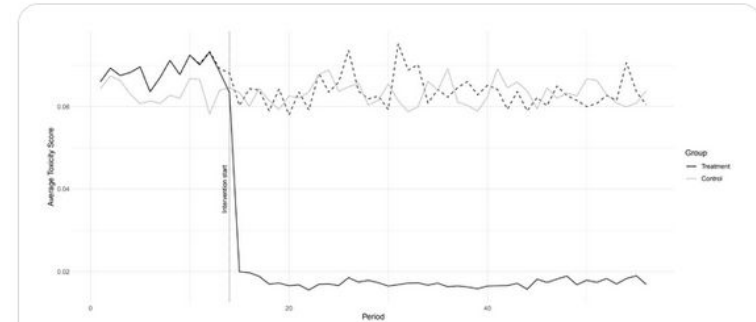


FIGURE 2: AVERAGE TOXICITY OF CONTENT SHOWN TO USERS DURING THE STUDY

Note: The figure depicts the average toxicity of posts, comments, and replies shown to users on each day of the study (relative to when a given participant started), separately for the control group and the treatment group. The dashed line for the treatment group demonstrates the average toxicity of elements that the platforms intended to show to the user before any hiding was applied by the intervention. The data presented here encompasses the three supported platforms (Twitter, Facebook, and YouTube). The dashed vertical line ("Intervention start") indicates day 15 – the first day of the intervention period.

3:17 PM · Nov 18, 2022 · Twitter Web App

## Twitter:

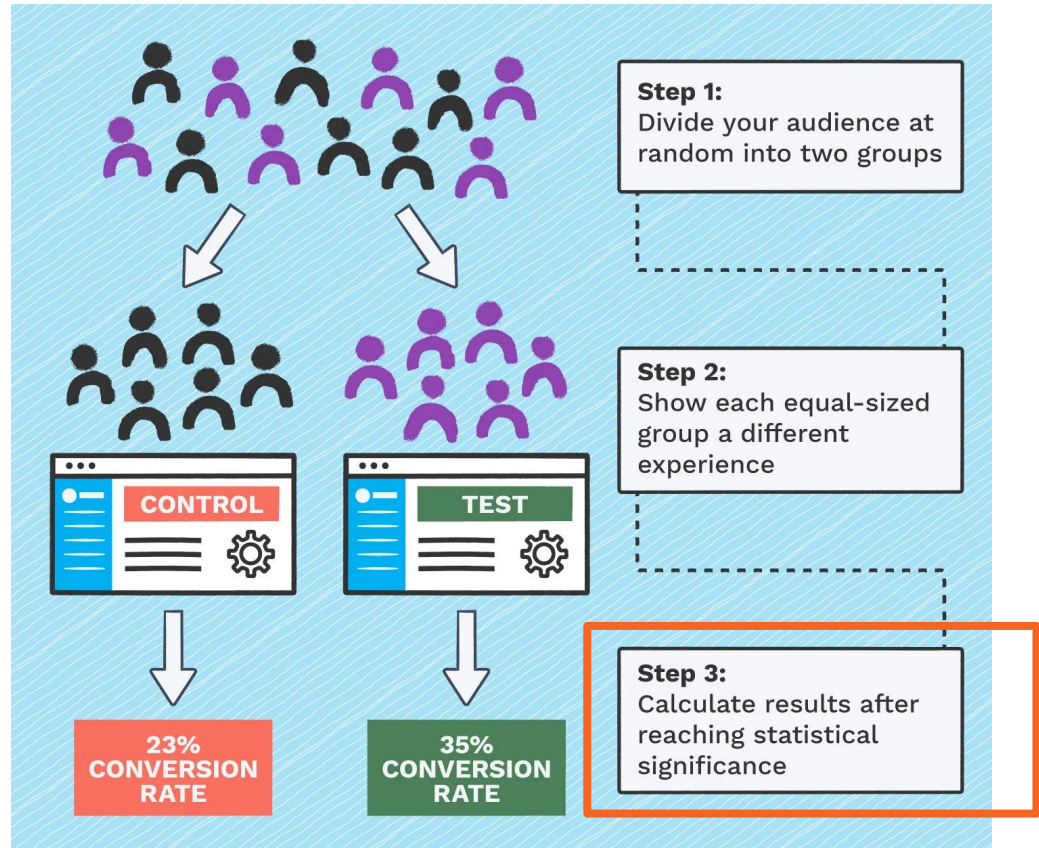
- A: show regular content
- B: hide toxic content
- Metric: ad consumption

---

# Common ~~marketing~~ metrics

- Metrics are dependent on your application!
- E.g. for clinical trials, metrics might include:
  - Did you have a heart attack? (binary)
  - Blood pressure (numeric)
  - # hospital visits (numeric)

How do we do the hypothesis test?



---

# Hypothesis tests for A/B

- $H_0$ : arm A is no different from arm B
- But, what distributions are we assuming?



---

# Distributions for A/B

Assumed Distribution	Example	Standard Test
<a href="#">Gaussian</a>	Average Revenue Per Paying User	<a href="#">Welch's t-test</a> (Unpaired t-test)
<a href="#">Binomial</a>	Click Through Rate	<a href="#">Fisher's exact test</a>
<a href="#">Poisson</a>	Transactions Per Paying User	E-test
<a href="#">Multinomial</a>	Number of each product purchased	<a href="#">Chi-squared test</a>

---

# Hypothesis tests for A/B

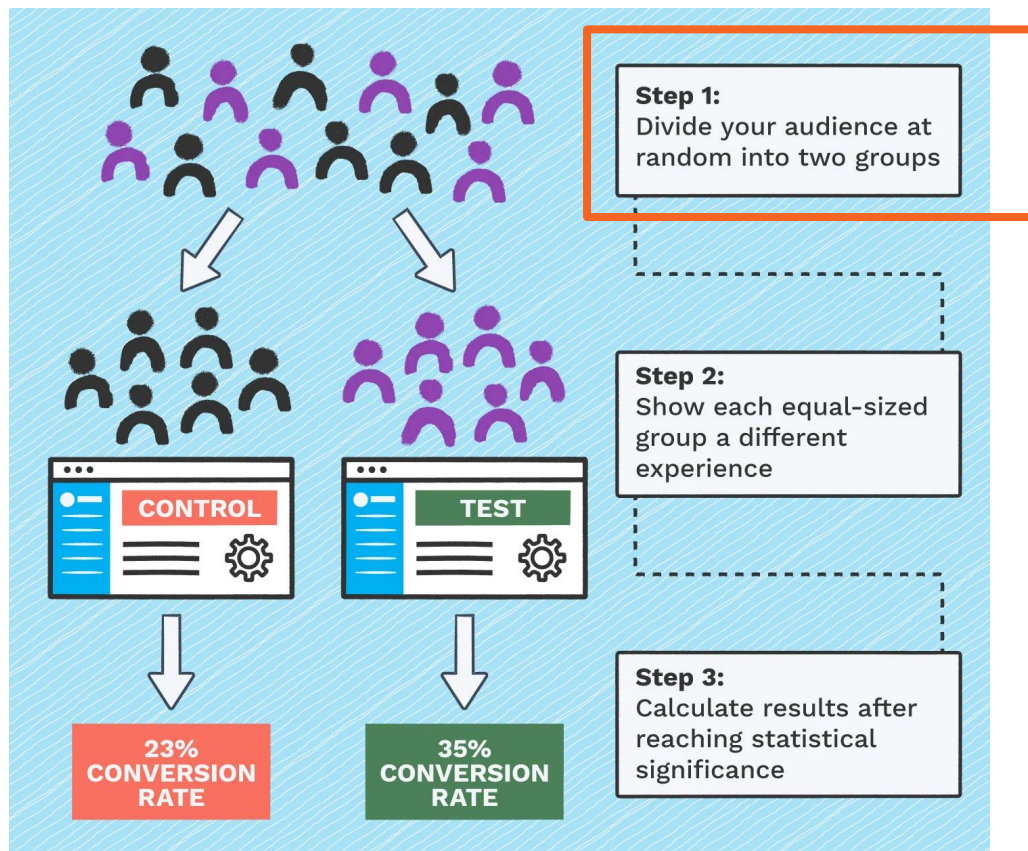
- $H_0$ : arm A is no different from arm B
- Choose a distribution and corresponding test **based on your metric**

---

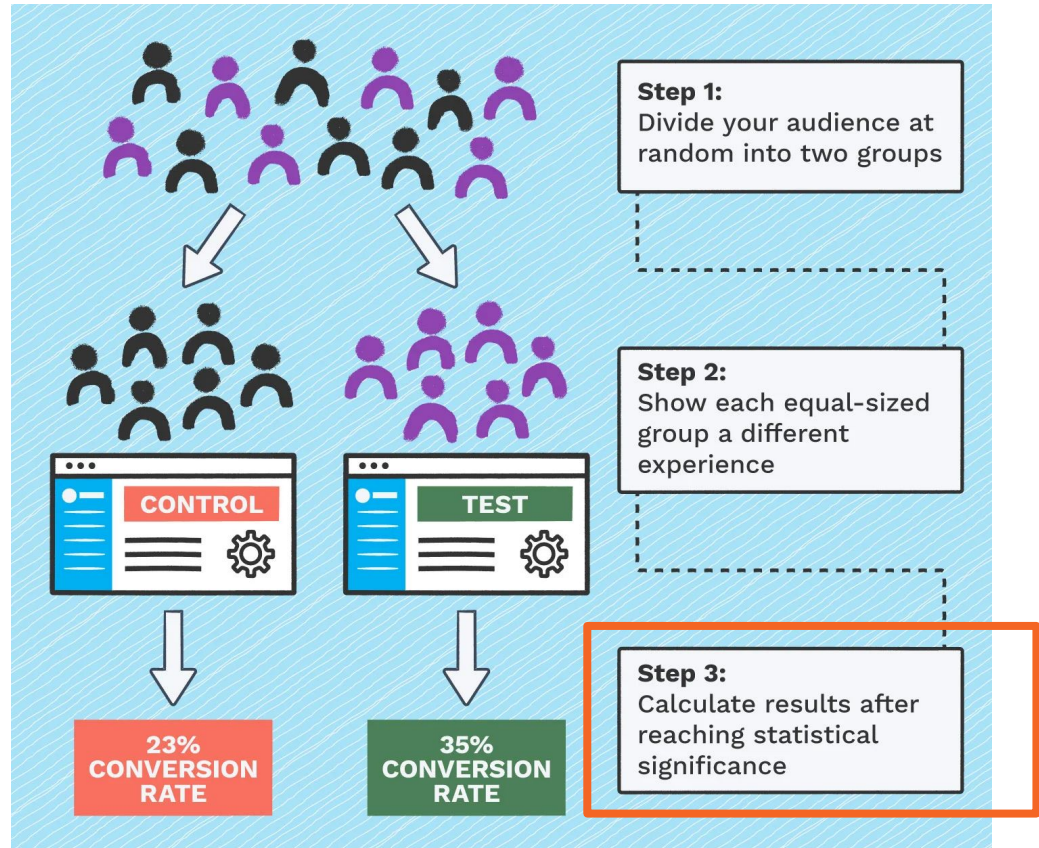
# Hypothesis tests for A/B

- $H_0$ : arm A is no different from arm B
- Choose a distribution and corresponding test **based on your metric**
- Note: instead of hypothesis testing, lots of companies these days use *Bayesian methods* (think: log-likelihood ratios) for sequential testing; not covered in this class

I have some users streaming in...



“after reaching statistical significance”??



---

## **But how do we know how many people to put in the A/B groups?**

- Want to determine a sample size that is large enough for statistical significance

---

## But how do we know how many people to put in the A/B groups?

- Want to determine a sample size that is large enough for statistical significance
- But sample size isn't necessarily only defined by the number of users!
  - **What else might affect significance?**

---

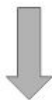
## But how do we know how many people to put in the A/B groups?

- Want to determine a sample size that is large enough for statistical significance
- But sample size isn't necessarily only defined by the number of users!
  - **How long we run our experiment (time!)**
  - **Longer experiment → more users**



# Example: blue vs. green buttons

A:



Project name Home About Contact Dropdown - Default Static top Fixed top

## Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Learn more

B:



Project name Home About Contact Dropdown - Default Static top Fixed top

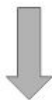
## Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

→ Learn more

# Example: blue vs. green buttons

A:



Project name Home About Contact Dropdown - Default Static top Fixed top

## Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Learn more

Conversion  
rate = 5%

B:



Project name Home About Contact Dropdown - Default Static top Fixed top

## Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

→ Learn more

Conversion  
rate = 4%

---

# Example A/B test

- **Goal:** establish with some level of confidence that experiment arm A is better than B
- **Question:** How many observations do we need?

---

# Example A/B test

- **Goal:** establish with some level of confidence that experiment arm A is better than B
- **Question:** How many observations do we need?

---

# Confidence vs. Power

**Confidence level.** The probability of failing to reject (i.e., retaining) the null hypothesis when it is true. Commonly set to 95%, this level implies that 5% of the time we will incorrectly conclude that there is a difference when there is none (Type I error). All else being equal, increasing this level reduces our power.

**Power.** The probability of correctly rejecting the null hypothesis,  $H_0$ , when it is false. Power measures our ability to detect a difference when it indeed exists. Commonly desired to be around 80–95%, although not directly controlled. If the Null Hypothesis is false, i.e., there is a difference in the arms, the power is the probability of determining that the difference is statistically significant.

---

**Sample Size Needed to Compare Two Binomial Proportions Using a Two-Sided Test with Significance Level  $\alpha$  and Power  $1 - \beta$ , Where One Sample ( $n_2$ ) Is  $k$  Times as Large as the Other Sample ( $n_1$ ) (Independent-Sample Case)**

To test the hypothesis  $H_0: p_1 = p_2$  vs.  $H_1: p_1 \neq p_2$  for the specific alternative  $|p_1 - p_2| = \Delta$ , with a significance level  $\alpha$  and power  $1 - \beta$ , the following sample size is required

$$n_1 = \left[ \sqrt{\bar{p}\bar{q}\left(1 + \frac{1}{k}\right)} z_{1-\alpha/2} + \sqrt{p_1 q_1 + \frac{p_2 q_2}{k}} z_{1-\beta} \right]^2 / \Delta^2$$

$$n_2 = k n_1$$

where  $p_1, p_2$  = projected true probabilities of success in the two groups

$$q_1, q_2 = 1 - p_1, 1 - p_2$$

$$\Delta = |p_2 - p_1|$$

$$\bar{p} = \frac{p_1 + k p_2}{1 + k}$$

$$\bar{q} = 1 - \bar{p}$$

**Sample Size Needed for Comparing the Means of Two Normally Distributed Samples of Equal Size Using a Two-Sided Test with Significance Level  $\alpha$  and Power  $1 - \beta$**

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2} = \text{sample size for each group}$$

where  $\Delta = |\mu_2 - \mu_1|$ . The means and variances of the two respective groups are  $(\mu_1, \sigma_1^2)$  and  $(\mu_2, \sigma_2^2)$ .

Do everything with  
software in this class :)

## Inference for Proportions: Comparing Two Independent Samples

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values (as decimal fractions) for  $p_1$  (proportion in population 1) and  $p_2$  (proportion in population 2) and, if calculating power, a sample size (assumed the same for each sample). You may also modify  $\alpha$  (type I error rate) and the power, if relevant. After making your entries, hit the **calculate** button at the bottom.

- ☒ Calculate Sample Size (for specified Power)
- ☐ Calculate Power (for specified Sample Size)

Enter a value for  $p_1$ :

Enter a value for  $p_2$ :

- ☐ 1 Sided Test
- ☒ 2 Sided Test

Enter a value for  $\alpha$  (default is .05):

Enter a value for desired power (default is .80):

The sample size (for each sample separately) is:

Calculate

Reference: The calculations are the customary ones based on the normal approximation to the binomial distribution. See for example *Hypothesis Testing: Categorical Data - Estimation of Sample Size and Power for Comparing Two Binomial Proportions* in Bernard Rosner's **Fundamentals of Biostatistics**.

---

## Inference for Proportions: Comparing Two Independent Samples

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values (as decimal fractions) for  $p_1$  (proportion in population 1) and  $p_2$  (proportion in population 2) and, if calculating power, a sample size (assumed the same for each sample). You may also modify  $\alpha$  (type I error rate) and the power, if relevant. After making your entries, hit the **calculate** button at the bottom.

- ☒ Calculate Sample Size (for specified Power)
- ☐ Calculate Power (for specified Sample Size)

Enter a value for  $p_1$ :

Enter a value for  $p_2$ :

- ☐ 1 Sided Test
- ☒ 2 Sided Test

Enter a value for  $\alpha$  (default is .05):

Enter a value for desired power (default is .80):

The sample size (for each sample separately) is:

Reference: The calculations are the customary ones based on the normal approximation to the binomial distribution. See for example *Hypothesis Testing: Categorical Data - Estimation of Sample Size and Power for Comparing Two Binomial Proportions* in Bernard Rosner's **Fundamentals of Biostatistics**.



---

# Green vs. Blue Button

- To have “sufficient” (95% power, 95% confidence) belief in our results, we need 22,332 observations
  - 11,166 in arm A
  - 11,166 in arm B
- How long will this take?
  - If we get 100 users per day on this website...
  - \_\_\_\_ **days!**

---

# Green vs. Blue Button

- To have “sufficient” (95% power, 95% confidence) belief in our results, we need 22,332 observations
  - 11,166 in arm A
  - 11,166 in arm B
- How long will this take?
  - If we get 100 users per day on this website...
  - 223 days!

---

# How to figure out significance?

If we did actually get 11,166 users in each of A and B, and found 5% conversion rate in A and 4% in B:

```
conversions_control = 11,166*0.05
```

```
total_users_control = 11,166
```

```
conversions_treatment = 11,166*0.04
```

```
total_users_treatment = 11,166
```

---

# In Python (for reference)

```
prob_pooled = (conversions_control + conversions_treatment) / (total_users_control + total_users_treatment)

#Calculate pooled standard error and margin of error
se_pooled = math.sqrt(prob_pooled * (1 - prob_pooled) * (1 / total_users_control + 1 / total_users_treatment))
z_score = st.norm.ppf(1 - confidence_level / 2)
margin_of_error = se_pooled * z_score

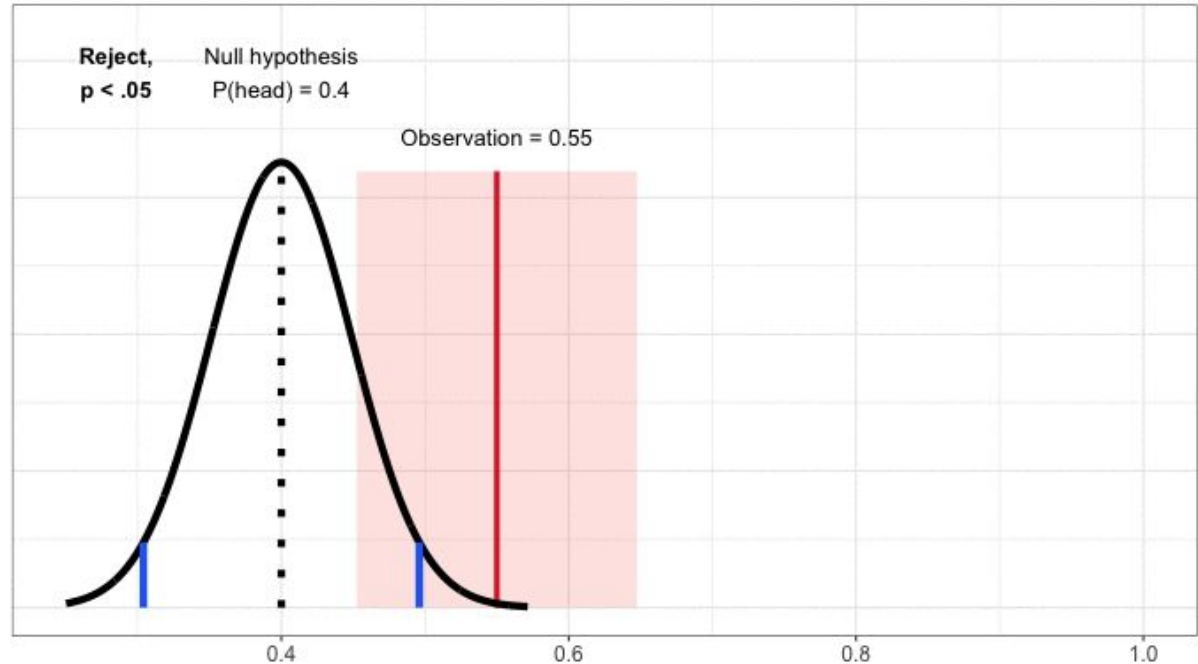
#Calculate dhat, the estimated difference between probability of conversions in the experiment and control groups
d_hat = (conversions_treatment / total_users_treatment) - (conversions_control / total_users_control)

#Test if we can reject the null hypothesis
lower_bound = d_hat - margin_of_error
upper_bound = d_hat + margin_of_error

if practical_significance < lower_bound:
    print("Reject null hypothesis")
else:
    print("Do not reject the null hypothesis")

print("The lower bound of the confidence interval is ", round(lower_bound * 100, 2), "%")
print("The upper bound of the confidence interval is ", round(upper_bound * 100, 2), "%")
```

# Visualizing the rejection region



# Example: blue vs. green buttons

A:



Project name Home About Contact Dropdown - Default Static top Fixed top

## Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Learn more

B:



Project name Home About Contact Dropdown - Default Static top Fixed top

## Welcome to our website

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

→ Learn more

---

# Example A/B test output

Metric  
(click through rate)

Control: 15% (+/- 2.1%)

Variation: 18% (+/- 2.3%)

---

# Example A/B test output

Control: 15% (+/- 2.1%)

Variation: 18% (+/- 2.3%)

Margin of error at  
5% significance



# The bootstrap confidence interval

```
poll = np.random.binomial(100, 0.28, 1000)
np.percentile(poll, [2.5, 97.5])
array([19., 37.])
```

95% of the time, **+/- 9ish**

This is called the “margin of error”

---

# Example A/B test output

**Control: 15% (+/- 2.1%)**

**Variation: 18% (+/- 2.3%)**

If you ran your A/B test 100 times, 95 of the ranges would capture the true CTR (e.g. here, the range that contains our CTR would be 15.7%-20.3%). CTR should fall outside the margin of error only 5 times.

---

# Example A/B test output

**Control: 15% (+/- 2.1%)**

**Variation: 18% (+/- 2.3%)**

If you ran your A/B test 100 times, 95 of the ranges would capture the true CTR (e.g. here, the range that contains our CTR would be 15.7%-20.3%). CTR should fall outside the margin of error only 5 times.

**Would you implement the button from arm B?**

---

# Example A/B test output

**Control: 15% (+/- 2.1%)**

**Variation: 18% (+/- 2.3%)**

If you ran your A/B test 100 times, 95 of the ranges would capture the true CTR (e.g. here, the range that contains our CTR would be 15.7%-20.3%). CTR should fall outside the margin of error only 5 times.

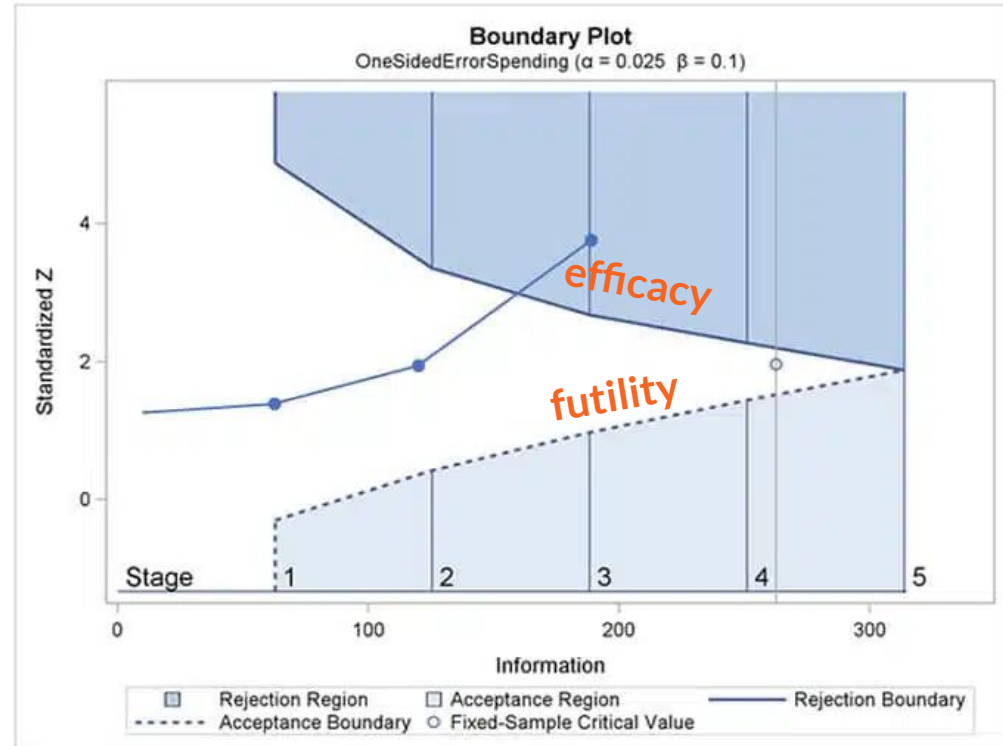
**If the cost to implement it is low, I'd consider implementing it!  
"3% lift" even though there's overlap between A and B CTRs  
(i.e. not statistically significantly different)**

---

# A/B experimentation...faster?

- There are 3 reasons you can potentially stop your A/B test early:
- **Benefit:**
  - Higher system usage, lower mortality
- **Harm:**
  - Lower system usage, serious drug side effects
- **Futility:**
  - No significant insight likely to come, experiment too \$\$\$

# Visualizing the rejection region



---

# Ethics: when is it ok to stop an experiment early?

## FDA wants to yank pregnancy drug. Firm argues Black women will suffer.

With no alternative treatments, some worry the move could deepen maternal and infant health inequities



By Ariana Eunjung Cha

Updated October 17, 2022 at 2:57 p.m. EDT | Published October 17, 2022 at 7:00 a.m. EDT

---

# Ethics: when is it ok to stop an experiment early?

- **Example: Preterm birth prevention drug (17 $\alpha$ -hydroxyprogesterone caproate)**
- First trial (Meis et al 2003) was stopped early for efficacy.
  - Trial population was **59% Black**



---

# Ethics: when is it ok to stop an experiment early?

- **Example: Preterm birth prevention drug (17 $\alpha$ -hydroxyprogesterone caproate)**
- First trial (Meis et al 2003) was stopped early for efficacy.
  - Trial population was **59% Black**
- Second trial (PROLONG, 2019) showed no effect.
  - Trial population was **7% Black**

---

# Ethics: when is it ok to stop an experiment early?

- **Example: Preterm birth prevention drug (17 $\alpha$ -hydroxyprogesterone caproate)**
- First trial (Meis et al 2003) was stopped early for efficacy.
  - Trial population was **59% Black**
- Second trial (PROLONG, 2019) showed no effect.
  - Trial population was **7% Black**
- Many varied criticisms of both studies, ranging from design to recruitment
- Could a longer duration have helped detect the null effect for non-Black patients?

---

# Stopping early

- This is an entire field with lots of statistical “sequential stopping” rules outside the scope of this class
  - Pocock bounds, O’Brien-Fleming, SPRT
- But you **must** preregister an early stopping method, otherwise you might just stop early because you *randomly* got really positive/negative results in arm B

---

# Stopping early... adaptively?

- What if we don't want to prespecify, and instead want to learn whether arm A or B is better on the fly, as users come in?
- Recall that our green button vs. blue button example will probably take us about 223 days to resolve
  - Can we do it faster?

---

# Multi-armed bandits



---

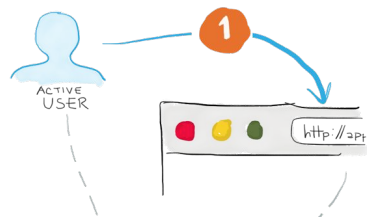
# Multi-armed bandits



- Want to maximize your payout \$\$\$
- On each slot machine (independently): win with probability  $p$ , lose with probability  $1-p$ 
  - But, you don't know  $p$
  - You can only estimate  $p$  by pulling the arm!

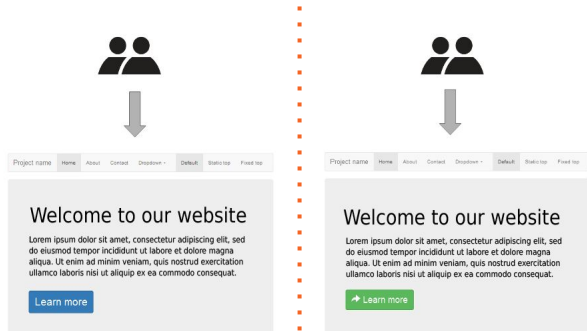
---

# Multi-armed bandits with users



- You can change how you distribute users into arms A and B!
- Instead of alternating 50/50, you can change the flow of traffic by looking at your metric:
  - If A is performing well, send more users to A and fewer to B
  - “Thompson sampling”

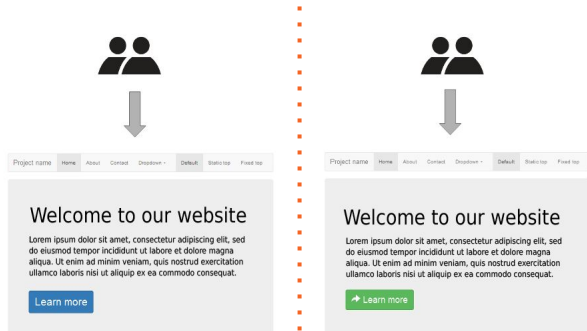
# Multi-armed bandits



- **Day 1:** 50 users  $\rightarrow$  A, 50 users  $\rightarrow$  B.
  - Arm A gets really lucky and has a 70% chance of being superior
- **Day 2:** assign 70% of users  $\rightarrow$  A, 30% of users  $\rightarrow$  B.
  - Now recompute probability of A being superior
- ...etc. for many days...



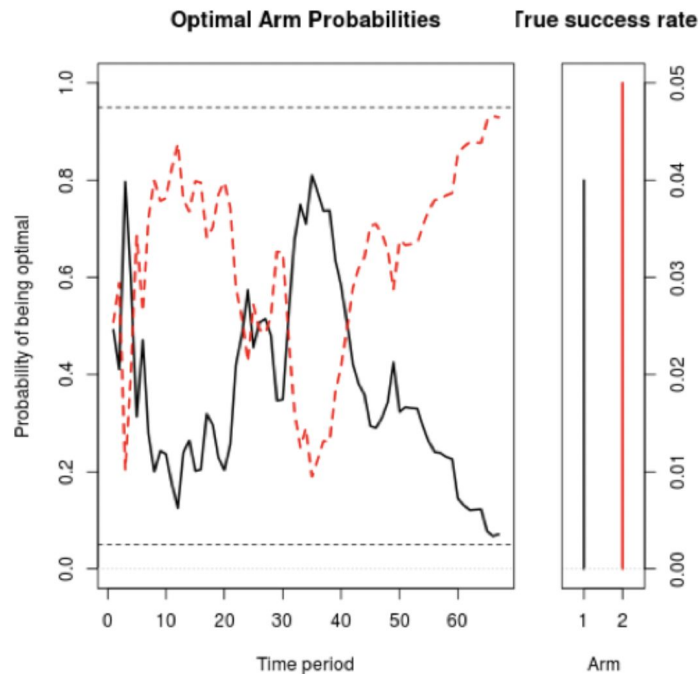
# Multi-armed bandits



- **Day 1:** 50 users  $\rightarrow$  A, 50 users  $\rightarrow$  B.
  - Arm A gets really lucky and has a 70% chance of being superior
- **Day 2:** assign 70% of users  $\rightarrow$  A, 30% of users  $\rightarrow$  B.
  - Now recompute probability of A being superior
- ...etc. for many days...
- My claim: we can simulate that we can finish this experiment in **66 days instead of 223!**

# Multi-armed bandits

A vs. B



---

# Experimentation Pitfall #1

- Stopping early when you shouldn't
  - Stopping as soon as you hit “significance” (e.g. after the first few users)
  - Experimenting on too few users
- Tip: preregister your early stopping methods; be careful even when using MAB

---

## Experimentation Pitfall #2

- Being unable to scale up to large numbers of users
  - Issues building software to do A/B testing “at scale”
- Tip: outsource to other companies that run “experimentation platforms” e.g. Eppo, Optimizely, etc.

---

## Experimentation Pitfall #3

- There are bugs in your code
  - Random assignment of people into A/B
  - Feature issues in A/B
  - Data collection
  - Data analysis pipeline
- Tip: audit your pipeline by doing an A/A test and make sure you don't get significant results!

---

# Experimentation Pitfall #4

- Failing to consider that your users are not a monolith
  - Mobile users might act differently than desktop users
  - Patients with different medical histories may react differently to a drug
- Tip: Segment your audiences (e.g. by device type, traffic source, etc.). Use “blocking”: when randomly assigning users to A and B, do so at the segment level.

---

# Experimentation Pitfall #5

- Failing to account for the “novelty effect”
  - Any shiny new feature will get a lot of clicks out of curiosity, even if it’s not a good feature
- Tip: Run your experiment for longer. Segment your audience by new vs. returning users (since all features are new for new users) and use “blocking”

---

## Experimentation Pitfall #6

- Running too many experiments at once and not accounting for interaction effects across new features
- Tip: you can do multivariate experiments, e.g. A/B/C testing, or A/B/C/D testing. You can also perform multivariate testing.



---

# Experimentation Pitfall #7

- Network effects: violating SUTVA
  - All experimental designs assume SUTVA (Stable Unit Treatment Value Assumption) = users in A and B are independent of each other, selected randomly, and only affected by the treatment and not external factors

---

# Experimentation Pitfall #7

- Network effects: violating SUTVA
  - All experimental designs assume SUTVA (Stable Unit Treatment Value Assumption) = users in A and B are independent of each other, selected randomly, and only affected by the treatment and not external factors
  - But if rolling out a payment feature in arm B, then only users who are both in arm B can try it (so you can only pay your friends who are also in arm B – maybe no one!)
- Tip: group similar users together by clustering, or...

# The New Zealand approach



---

# Admin

- No homework over break if you turn in HW6 on time!
- Happy Thanksgiving :)

---

# Attendance!



[tinyurl.com/29bjyd5v](https://tinyurl.com/29bjyd5v)