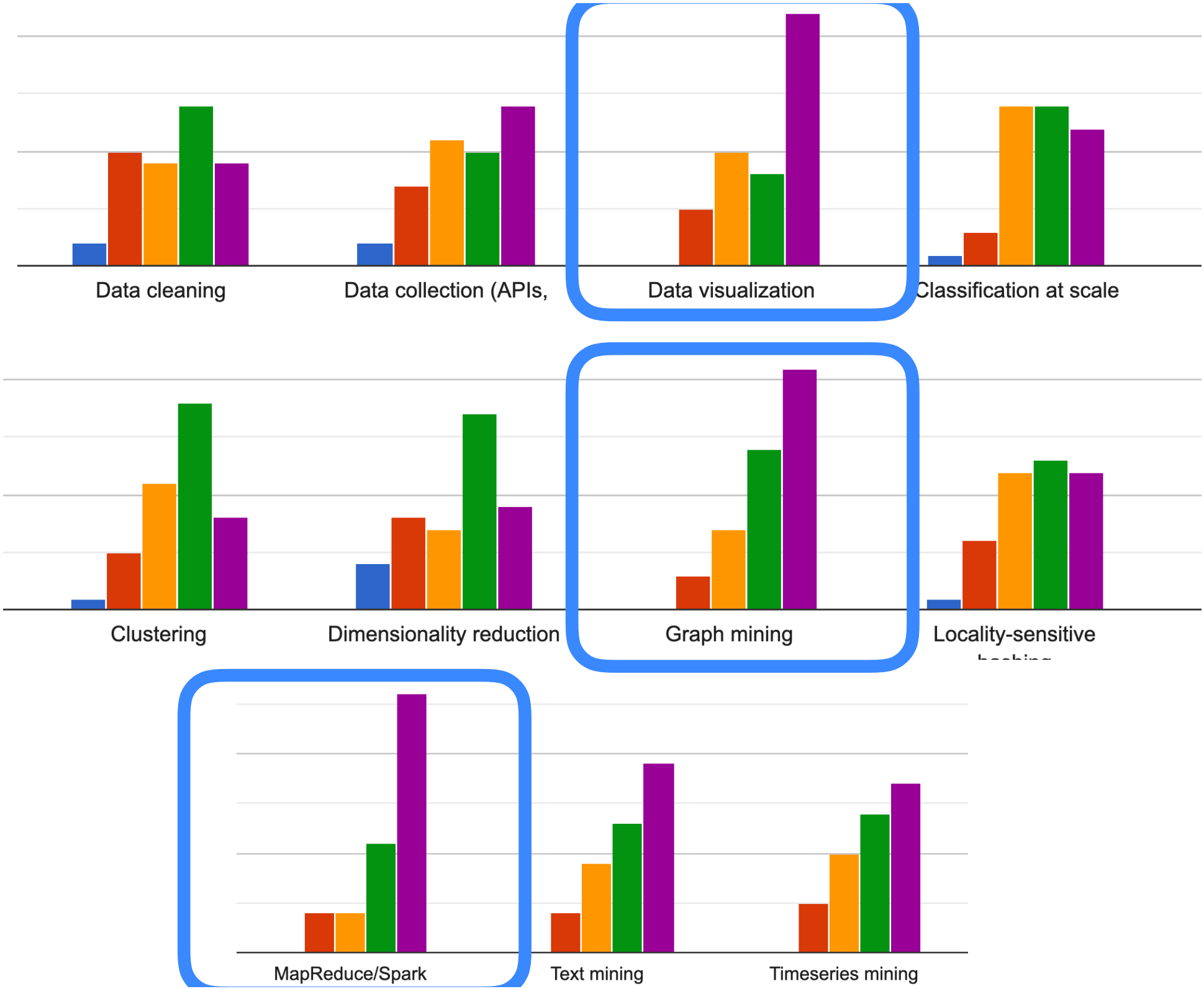
Data Mining and Analysis

Data Collection + Bias

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee


Data cleaning
Data collection (APIs,
Data visualization
Classification at scale
Clustering
Dimensionality reduction
Graph mining
Locality-sensitive
MapReduce/Spark
Text mining
Timeseries mining

Data Collection + Bias

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

Raw Data is Collected / Recorded

email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages

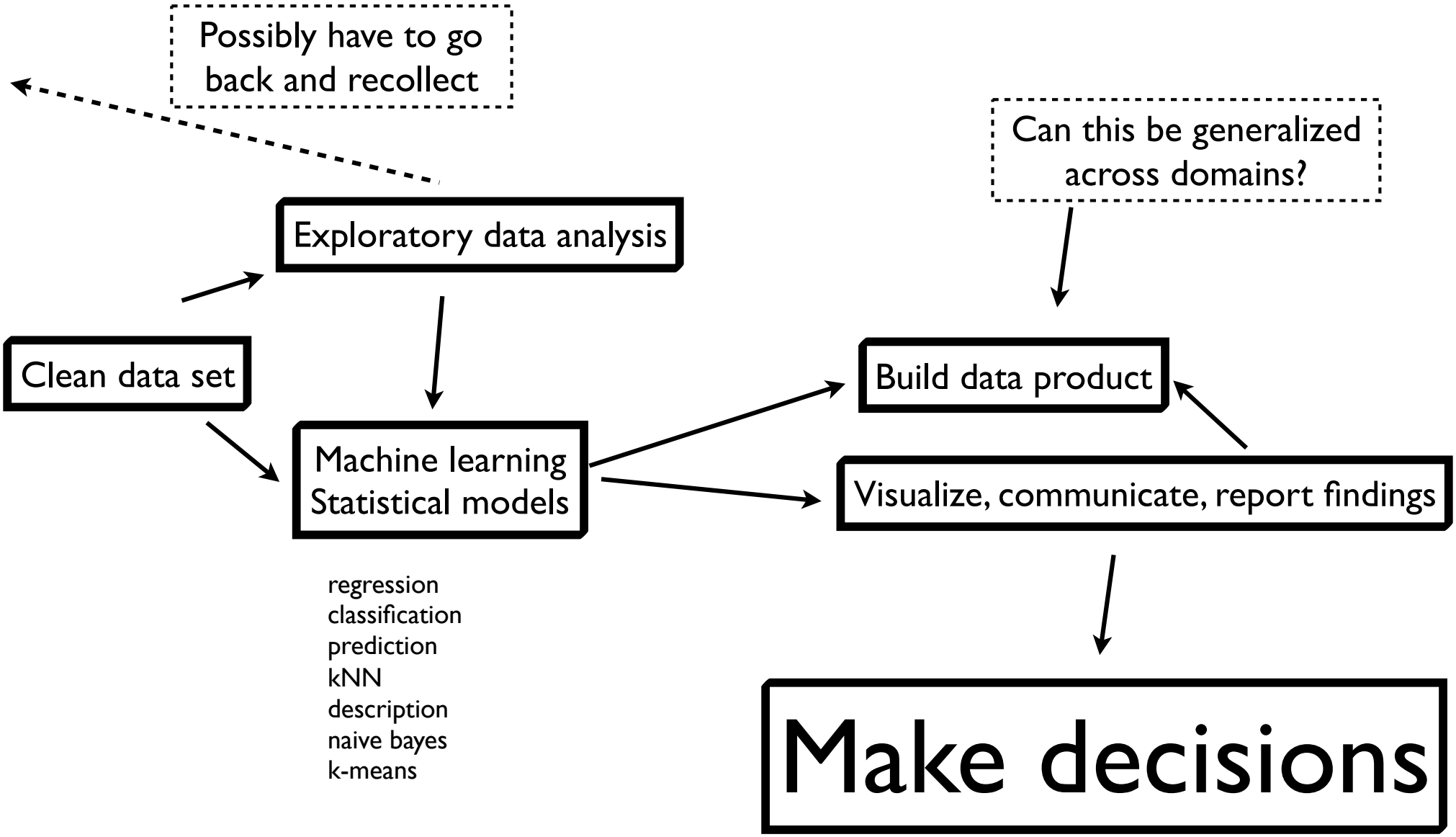
Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

Data set

"clean"
table

http://columbiadatascience.com/2012/09/24/reflections-after-jakes-lecture/


Possibly have to go back and recollect
Can this be generalized across domains?
Exploratory data analysis
Clean data set
Build data product
Machine learning
Statistical models
Visualize, communicate, report findings
regression
classification
prediction
kNN
description
naive bayes
k-means
Make decisions

Real World

Raw Data is
Collected / Recorded

How to collect data?

Re-use existing dataset (just download!)

Use an API

Build a crawler

Collect yourself!

- Instrument the world (e.g., sensors)
- Commission a survey

Augment an existing dataset:

- Use AMT to annotate an existing dataset

Data you can download

https://www.data.gov/

https://archive.ics.uci.edu/ml/datasets.php

https://www.kaggle.com/datasets

https://github.com/awesomedata/awesome-public-datasets

https://snap.stanford.edu/data/

https://toolbox.google.com/datasetsearch

…

DATA.GOV

DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 236,389 DATASETS

Monthly House Price Indexes

BROWSE TOPICS

Agriculture Climate Consumer Ecosystems Education Energy Finance

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Browse Through: 481 Data Sets

Default Task

Classification (356)
Regression (103)
Clustering (88)
Other (55)

Attribute Type

Categorical (38)
Numerical (318)
Mixed (55)

Data Type

Multivariate (367)
Univariate (24)
Sequential (49)
Time-Series (94)
Text (55)
Domain-Theory (23)
Other (21)

Area

Life Sciences (109)
Physical Sciences (49)
CS / Engineering (177)
Social Sciences (26)
Business (31)
Game (10)
Other (74)

 # Attributes

Less than 10 (117)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987

kaggle Search Competitions Datasets Notebooks Discussion Courses ••• Sign in

Datasets

Documentation New Dataset

Search 20,446 datasets Feedback Filter

PUBLIC Sort by: Hottest

PGA Tour Golf Data 182

Brad Klassen

6 days 92 MB 10.0 2 Files (CSV)

Crimes in Boston 284

Analyze Boston

a year 10 MB 8.2 2 Files (CSV)

Gas Prices in Brazil 183

Matheus Eduardo Freitag

a month 3 MB 10.0 1 File (other)

Goodreads-books 414

Awesome Public Datasets

awesome

NOTICE: This repo is automatically generated by apd-core. Please **DO NOT** modify this file directly. We have provided a new way to contribute to Awesome Public Datasets. The original PR entrance directly on repo is closed forever.

- I am well.
- Please fix me.

This list of a topic-centric public data sources in high quality. They are collected and tidied from blogs, answers, and user responses. Most of the data sets listed below are free, however, some are not. Other amazingly awesome lists can be found in sindresorhus's awesome list.

Table of Contents

- Agriculture
- Biology
- Climate+Weather
- ComplexNetworks
- ComputerNetworks
- DataChallenges
- EarthScience
- Economics
- Education
- Energy

snap.stanford.edu/data/

By *Jure Leskovec*

STANFORD UNIVERSITY

SNAP

- SNAP for C++
- SNAP for Python
- SNAP Datasets
- BIOSNAP Datasets
- What's new
- People
- Papers
- Projects
- Citing SNAP
- Links
- About
- Contact us

Open positions

Open research positions

Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks, articles, and metadata : Talk, editing, voting, and article data from Wikipedia
- Temporal networks : networks where edges have timestamps
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon
- Face-to-face communication networks : nodes are people and edges are face-to-face (non-online) interactions

SNAP networks are also available from SuiteSparse Matrix Collection by Tim Davis.

Google Dataset Search Beta

Search for Datasets

Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.

Data via an API

Nice collection of APIs at:

https://github.com/public-apis/public-apis

Public APIs build passing

A collective list of free APIs for use in software and web development.

Sponsor:

A public API for this project can be found here - thanks to DigitalOcean for helping us provide this service!

For information on contributing to this project, please see the contributing guide.

Example API: NYTimes

News

API	Description	Auth	HTTPS	CORS
Chronicling America	Provides access to millions of pages of historic US newspapers from the Library of Congress	No	No	Unknown
Currents	Latest news published in various news sources, blogs and forums	`apiKey`	Yes	Yes
Feedbin	RSS reader	`OAuth`	Yes	Unknown
Feedster	Searchable and categorized collections of RSS feeds	`apiKey`	Yes	Unknown
New York Times	Provides news	`apiKey`	Yes	Unknown
News	Headlines currently published on a range of news sources and blogs	`apiKey`	Yes	Unknown
NPR One	Personalized news listening experience from NPR	`OAuth`	Yes	Unknown
The Guardian	Access all the content the Guardian creates, categorised by tags and section	`apiKey`	Yes	Unknown
The Old Reader	RSS reader	`apiKey`	Yes	Unknown

developer.nytimes.com

Developers

Home APIs Get Started Sign In

The New York Times Developer Network

All the APIs Fit to POST

Get Started

Learn how to sign up for an API key.

GET STARTED

APIs

Learn about and try out NYT's APIs.

VIEW APIS

FAQ

Get answers to frequently asked questions.

FAQ

Branding

Read about NYT's branding & attribution

API REFERENCE

/articlesearch.json GET

RESOURCES

Overview

Article

Byline

Headline

Keyword

Multimedia

Person

Article Search

Use the Article Search API to look up articles by keyword. You can refine your search using filters and facets.

```
/articlesearch.json?q={query}&fq={filter}
```

Example Call

```
https://api.nytimes.com/svc/search/v2/articlesearch.json?q=election&api-key=yourkey
```

FILTERING YOUR SEARCH

Use filters to narrow the scope of your search. You can specify the fields and the values that your query will be filtered on. The Article Search API uses Elasticsearch, so the filter query uses standard Lucene syntax. Separate the filter field name and value with a colon, and surround multiple values with parentheses.

```
field-name:("value1" "value2" ... "value n")
```

The default connector for values in parentheses is OR. If you declare an explicit boolean value, it must be capitalized. You can filter on multiple values and fields.

```
field-name-1:("value1") AND field-name-2:("value2" "value3")
```

For a list of all fields you can filter on, see the Filter Query Fields table below.

Data you need to scrape

Goodreads

Spotify

ESPN

Google Play

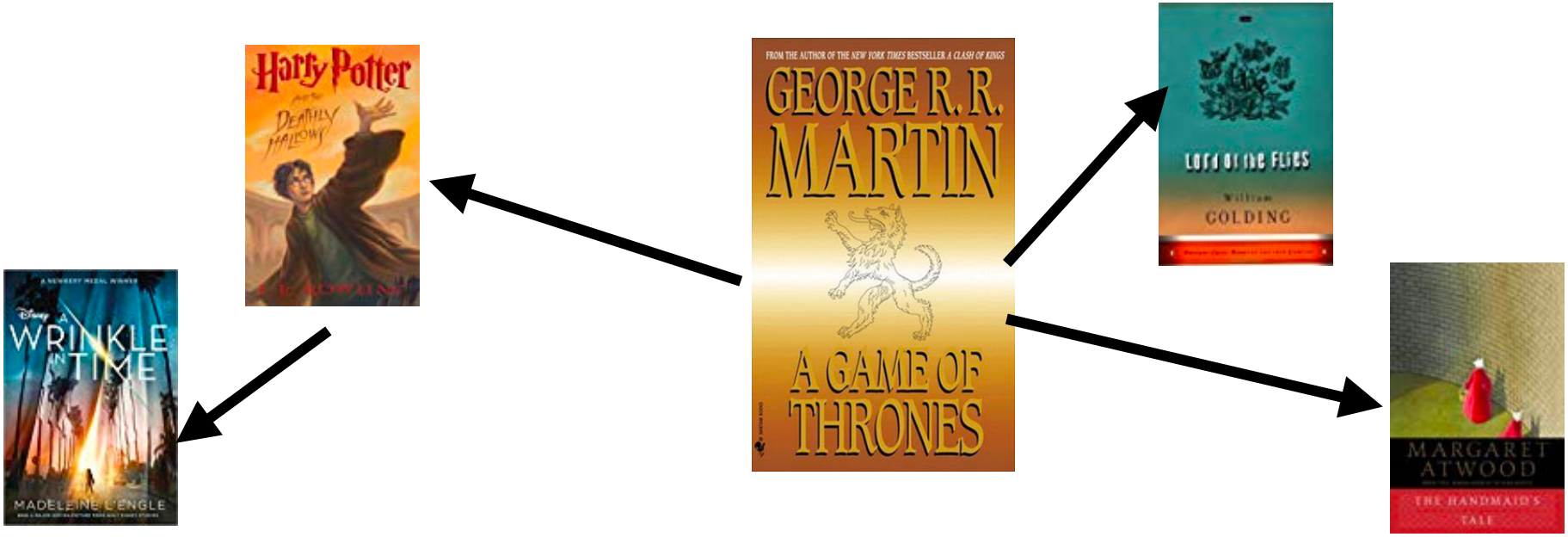
...

Example: Goodreads

Goal: extract a large collection of books that are related to each other

Each node is a book

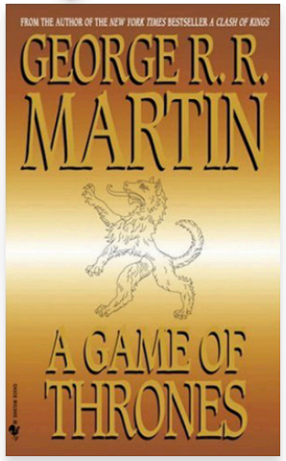
An edge connects “similar” books



Example: Goodreads

https://www.goodreads.com/book/show/13496.A_Game_of_Thrones

goodreads Home My Books Browse Community Search books Sign In



Want to Read

Rate this book

Open Preview

A Game of Thrones

(A Song of Ice and Fire #1)

by George R.R. Martin

4.45 · Rating details · 1,793,707 ratings · 51,795 reviews

Here is the first volume in George R. R. Martin's magnificent cycle of novels that includes A Clash of Kings and A Storm of Swords. As a whole, this series comprises a genuine masterpiece of modern fantasy, bringing together the best the genre has to offer. Magic, mystery, intrigue, romance, and adventure fill these pages and transport us to a world unlike any we have ever ...more

GET A COPY

Kindle Store $8.99 | Amazon | Stores | Libraries

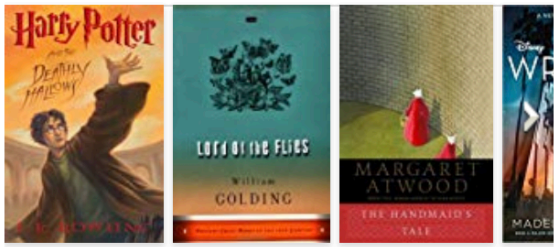
Mass Market Paperback, 848 pages

Published August 2005 by Bantam (first published August 6th 1996)

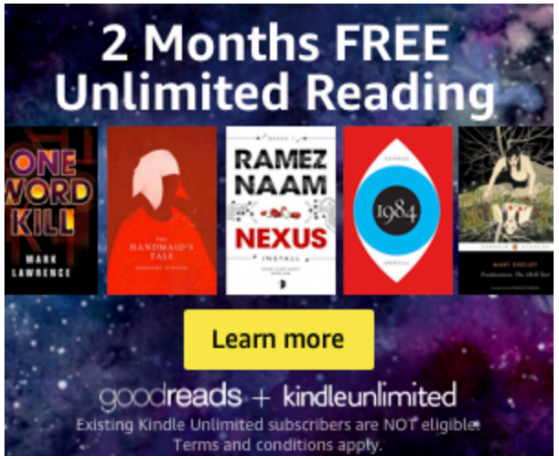
More Details... Edit Details

Share

READERS ALSO ENJOYED



See similar books...



sponsored books

A Noble Woman

Digital edition is available worldwide for

"Sci-fi thriller." Kirkus Reviews

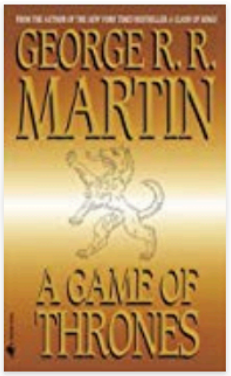
Warrior. Hero. Slave. Fighting to restore

https://www.goodreads.com/book/similar/1466917-a-game-of-thrones

goodreads Home My Books Browse Community Search books

George R.R. Martin > A Game of Thrones > Similar books

Books similar to A Game of Thrones (A Song of Ice and Fire, #1)



A Game of Thrones (A Song of Ice and Fire, #1)

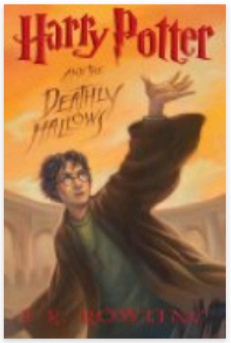
by George R.R. Martin

4.45 avg. rating • 1,630,964 Ratings

Here is the first volume in George R. R. Martin's magnificent cycle of novels that includes A Clash of Kings and A Storm of Swords. As a whole, this series comprises a genuine masterpiece of modern f... More

Want to Read Rate it:

Goodreads members who liked this book also liked:



Harry Potter and the Deathly Hallows (Harry Potter, #7)

by J.K. Rowling

4.61 avg. rating • 2,264,423 Ratings

Harry Potter is leaving Privet Drive for the last time. But as he climbs into the sidecar of Hagrid's motorbike and they take to the skies, he knows Lord Voldemort and the Death Eaters will not be fa... More

Want to Read Rate it:

class="wtrStarsContainer wtrStarsContainer--alwaysDisplay" data-reactid=".w61pjng83k.$394add486c44b811aba8aa75c1df827c50a1783f.1.0.3.0.1"> Rate it: <button class="userRating__star" title="Rate 1 star" data-reactid=".w61pjng83k.$394add486c44b811aba8aa75c1df827c50a1783f.1.0.3.0.1.0.1.$1"></button><button class="userRating__star" title="Rate 2 stars" data-reactid=".w61pjng83k.$394add486c44b811aba8aa75c1df827c50a1783f.1.0.3.0.1.0.1.$2"></button><button class="userRating__star" title="Rate 3 stars" data-reactid=".w61pjng83k.$394add486c44b811aba8aa75c1df827c50a1783f.1.0.3.0.1.0.1.$3"></button><button class="userRating__star" title="Rate 4 stars" data-reactid=".w61pjng83k.$394add486c44b811aba8aa75c1df827c50a1783f.1.0.3.0.1.0.1.$4"></button><button class="userRating__star" title="Rate 5 stars" data-reactid=".w61pjng83k.$394add486c44b811aba8aa75c1df827c50a1783f.1.0.3.0.1.0.1.$5"></button>

If you decide to scrape …

Lots of helpful libraries:

e.g., Selenium, Beautiful Soup, Scrapy, JSoup, …

Respect robots.txt

Don’t overwhelm any single site (e.g., use a delay between requests) … or you may get BLOCKED!

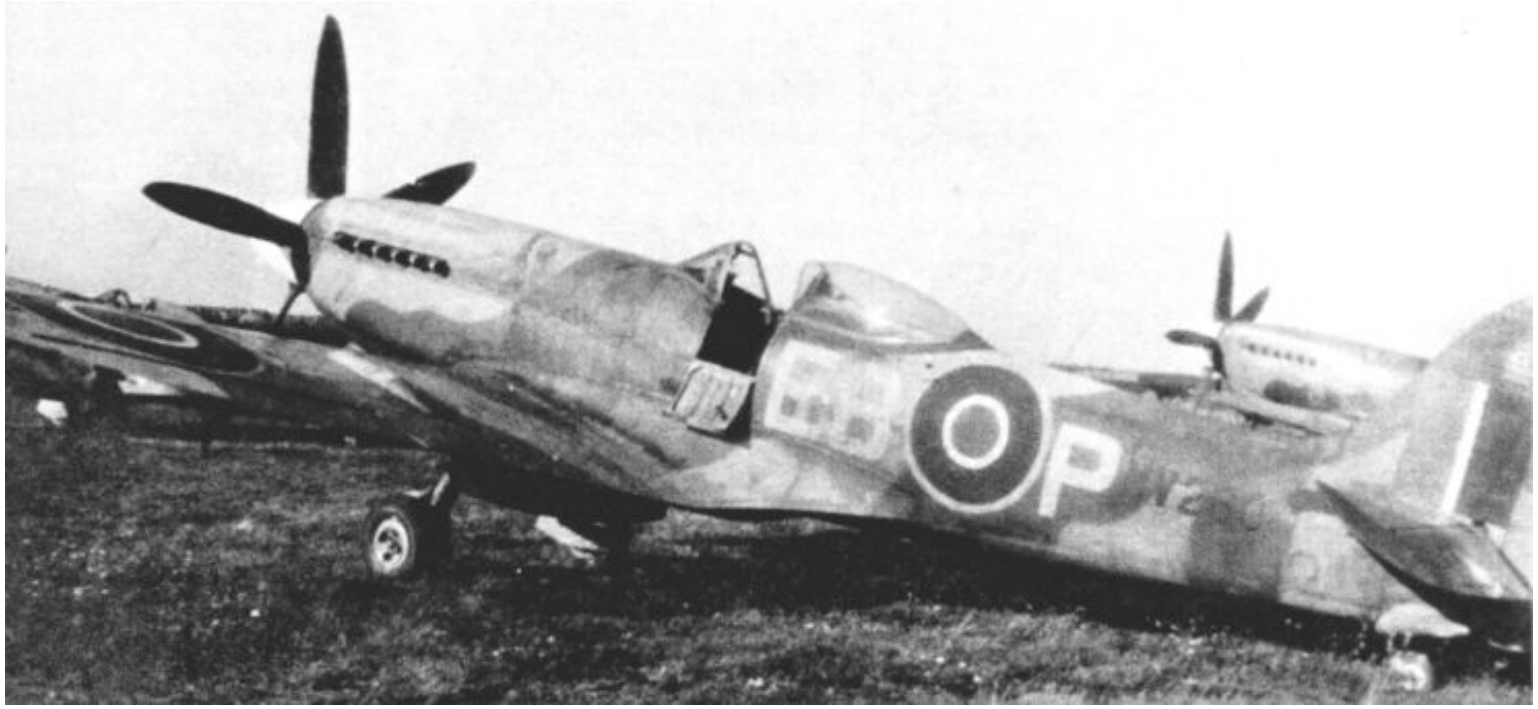
Data Collection Issues

Abraham Wald and the Missing Bullet Holes

An excerpt from How Not To Be Wrong by Jordan Ellenberg

https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d#.t1rhend8d

WWII:

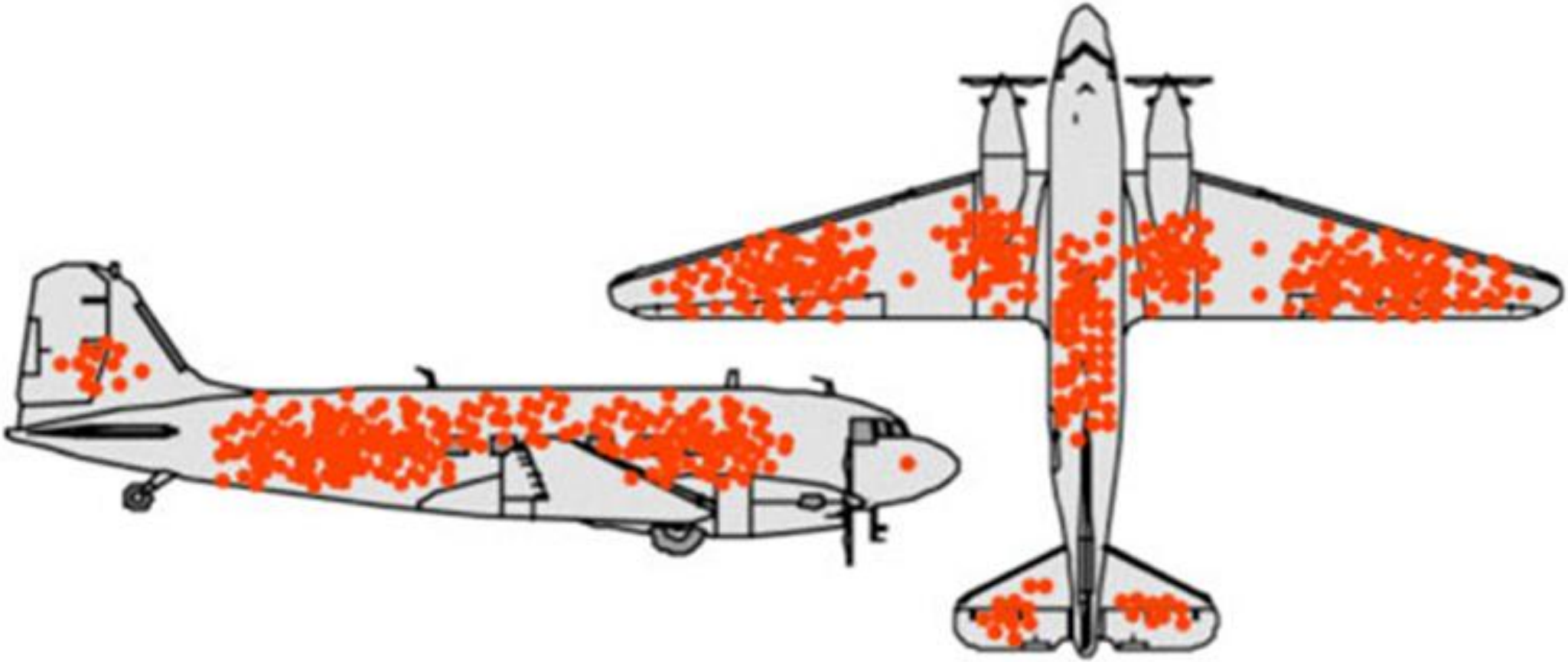


The Challenge

You don’t want your planes to get shot down by enemy fighters, so you armor them. But armor makes the plane heavier, and heavier planes are less maneuverable and use more fuel. Armoring the planes too much is a problem; armoring the planes too little is a problem. Somewhere in between there’s an optimum.

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

Where does the armor go?



Where does the armor go?

Wald's insight:

> The armor doesn't go where the bullet holes are. It goes where the bullet holes aren't: on the engines.

An illustration of Survivorship bias

Survivorship Bias

Other examples?

http://news.stanford.edu/2005/06/14/jobs-061505/

“music was better in the past”

Stanford | News

Search Stanford news…

Home Find Stories For Journalists Contact

JUNE 14, 2005

'You've got to find what you love,' Jobs says

This is a prepared text of the Commencement address delivered by Steve Jobs, CEO of Apple Computer and of Pixar Animation Studios, on June 12, 2005.

Goal: estimate the prevalence of UFOs

National UFO Reporting Center
Report Index For Post Date 160816
Click on links for details

NUFORC Home

Date / Time	City	State	Shape	Duration	Summary	Posted
8/16/16 05:30	Providence	RI	Circle	5 minutes	Red orb moving quickly, then stops multiple times over Providence, RI.	8/16/16
8/16/16 04:10	Edison	NJ	Triangle	Minutes (2 sightings)	Lighted Triangular Crafts sighted within minutes of each other HIGH OVERHEAD.	8/16/16
8/16/16 02:47	New York City (Manhattan)	NY	Unknown	15-20 minutes	I was on the roof of my apartment building eating a bag a potato chips around 2:45 am this morning. I was just absentmindedly looking u	8/16/16
8/15/16 23:15	Cornelius	NC	Circle	3 seconds	One round bright light, brighter than the moon, appearing from nowhere and descending rapidly straight down followed by two smaller bri	8/16/16
8/15/16 23:00	Sumter	SC		45 seconds	As I was laying in my bed around eleven O clock pm I strangely got The urge to go outside.I ignored the felling and layed back down I s	8/16/16
8/15/16 21:36	San Diego	CA	Fireball	1 minute	Flashing fireball slowly moving over Mission Valley around 9:36 pm on Monday night 8/15/2016. Looked like the flash of fire wold repeat	8/16/16
8/15/16 21:15	Boulder	CO	Unknown	30 seconds	Light traveling beneath the clouds without marker lights of regular aircraft. Light looked like like it was traveling in our dorection	8/16/16
8/15/16 20:05	Gilbert	AZ	Circle	5 minutes	Round shaped object with no blinking lights, just white, first high in air, then lower instantly, moving across the sky, disappears.	8/16/16
8/15/16 03:10	Nashville	TN	Light	~2 minutes	Bright orange light in the sky, appeared to be two separate lights when zoomed in. Both shaped in straight lines.	8/16/16
8/14/16 22:45	Marina Del Rey	CA	Light	30 seconds	Walking west on Washington blvd at Palawan way 6 orange lights were seen moving in out of a straight line. Then they formed a straight	8/16/16
8/14/16 22:40	Wassaic	NY	Fireball	15 seconds	Fireball above Wassaic, NY.	8/16/16
8/14/16 21:30	Stoughton	MA	Light	>2 hours	((HOAX??)) I saw four lights in the sky in a straight angled line. All difrent colors. I have pictures. ((anonymous report))	8/16/16
8/14/16 21:30	Ogallala	NE	Triangle	>15 minutes	Spotted Triangle of lights above No Name Bay at Lake McConaughey in Ogallala NE	8/16/16
8/14/16 21:00	Wheatland	CA	Circle	Several minutes	Observed strange bluish light. North to south direction. Unknown altitude. Very Fast. Horizon to horizon in approx. two minutes.	8/16/16
8/14/16 21:00	Camden	NJ	Circle	5 minutes	Red and orange light near Philadelphia international and the phillies stadium	8/16/16
8/14/16 19:45	Sedro Woolley	WA	Other	6 seconds	Big black tranparent blib in sky.	8/16/16
8/14/16 19:00	Nashua	NH	Circle	20 minute	object is a commercial airliner, above a contrail. PD))	8/16/16
8/14/16 18:00	Vancouver	WA	Egg	5 minutes	ied to get video but I think I was to far away.	8/16/16
8/14/16 17:00	Allyn	WA	Disk	10-15 seco	qyickly, sped off.	8/16/16
8/14/16 16:30	Saratoga	WY	Sphere	20	t up into the woods a little ways. ((anonymous report))	8/16/16
8/14/16 01:57	Vancouver	WA	Light	2× 1/2 min	pears again in Vancouver sky. Total time 1min.	8/16/16
8/14/16 00:00	Chester	SC	Triangle	Hours	em.	8/16/16
8/13/16 00:00	Berea	OH	Other	?	us report))	8/16/16
8/13/16 23:33	Fargo	ND	Sphere	6 minutes	few sec., then went away 6-10 minutes.	8/16/16
8/13/16 23:00	Denver	CO	Other	3 minutes	I walked out to my backyard and saw a very bright light that looked like a star but i noticed it was moving slow. Before my eyes the cr	8/16/16
8/13/16 22:15	Jacksonville	FL	Light	>15 minutes	My girlfriend noticed a bright star over one of the trees and pointed it out to me.	8/16/16
8/13/16 21:45	Bourbonnais	IL	Formation	45 seconds	10-12 yellow/orange globes traveling N to S Kankakee Co.; disappeared quickly in S. ((anonymous report))	8/16/16
8/13/16 21:40	Hurricane	WV	Light	5 minutes	Bright single light that went from northwest to southeast. No aircraft flashing lights. ((NUFORC Note: Intl. Space Station. PD))	8/16/16
8/13/16 21:40	Madison	WI	Light	<10 minutes	Red/orange light seen east of Madison, WI.	8/16/16
8/13/16 20:45	Mukwonago	WI	Light	5 minutes	Bright light that turned deep red.	8/16/16
8/13/16 19:10	Miami (Bal Harbor)	FL	Other	5-10 seconds	Invisible flying saucer exhausts and top revealed by sunsent over miami	8/16/16
8/13/16 13:00	Leawood	KS	Disk	1 hour	Objects circling in sky, daytime, other adults saw crafts. Too high up and far away to attempt to photograph.	8/16/16
8/13/16 05:15	Raleigh	NC	Sphere	~30 seconds	Object 300 meters in the air, crossing the sky like a rocket opens the walls of earth and explodes into thin air leaving just a trial	8/16/16
8/13/16 05:15	Alexandria	MN	Light	5 minutes	Blinking satellite type objects.	8/16/16
8/13/16 04:00	Fort Collins	CO	Chevron	~5 seconds	Was watching the meteor shower from a field at the end of the road. ~ 4AM an object flew overhead which had 5 lights in a V formation.	8/16/16
8/13/16 03:25	Harrisburg	PA	Circle	3 seconds	Flash of white/green light, moving very fast over the Susquehanna River. Suddenly disappears. No debris. ((anonymous report))	8/16/16
8/13/16 02:15	Longmont	CO	Rectangle	2 minutes	Big, rectangular, flying object that changed direction.	8/16/16
8/13/16 00:47	Albany	OR	Rectangle	20 seconds	Large reddish orange craft moving SW at fast rate of speed. Heard roaring almost like engine. Noisy neighborhood though.	8/16/16
8/12/16 23:30	Derby (UK/England)		Unknown	3 seconds	4 lights not recorded on camera.	8/16/16
8/12/16 23:00	Fairfax	VA	Triangle	90 minutes	From about 23:00 to 00:30 I saw two triangle shaped figures in the sky. The appearance of the UFOs were that they had three white light	8/16/16
8/12/16 22:30	Portland	OR	Light	Ongoing	Satellite not what it appears to be.	8/16/16
8/12/16 22:30	Jackson	MI	Light	50 seconds	Very, very bright led type light moving slowly west to east. High in the atmosphere.	8/16/16
8/12/16 21:00	Lake Roesiger	WA	Light	<1 minute	Silent floating white orb over Lake Roesiger, WA.	8/16/16
8/12/16 21:00	Albuquerque	NM	Light	3 minutes	Bright red light in the sky moving ENE disappearing in mid flight.	8/16/16
8/12/16 21:00	Inkster	MI	Disk	7 minutes	Saw a black saucer in the sky that started with being a red light	8/16/16
8/12/16 20:30	New York City (Middle Village)	NY	Unknown	10 seconds	Four red lights with two blinking green lighted object in sky	8/16/16
8/12/16 19:00	St. Peters	MO	Circle	8 seconds	We thought it was a plane falling from the sky, but as it became closer to the ground we realized the bright lit red ball did not have	8/16/16

Concerns?

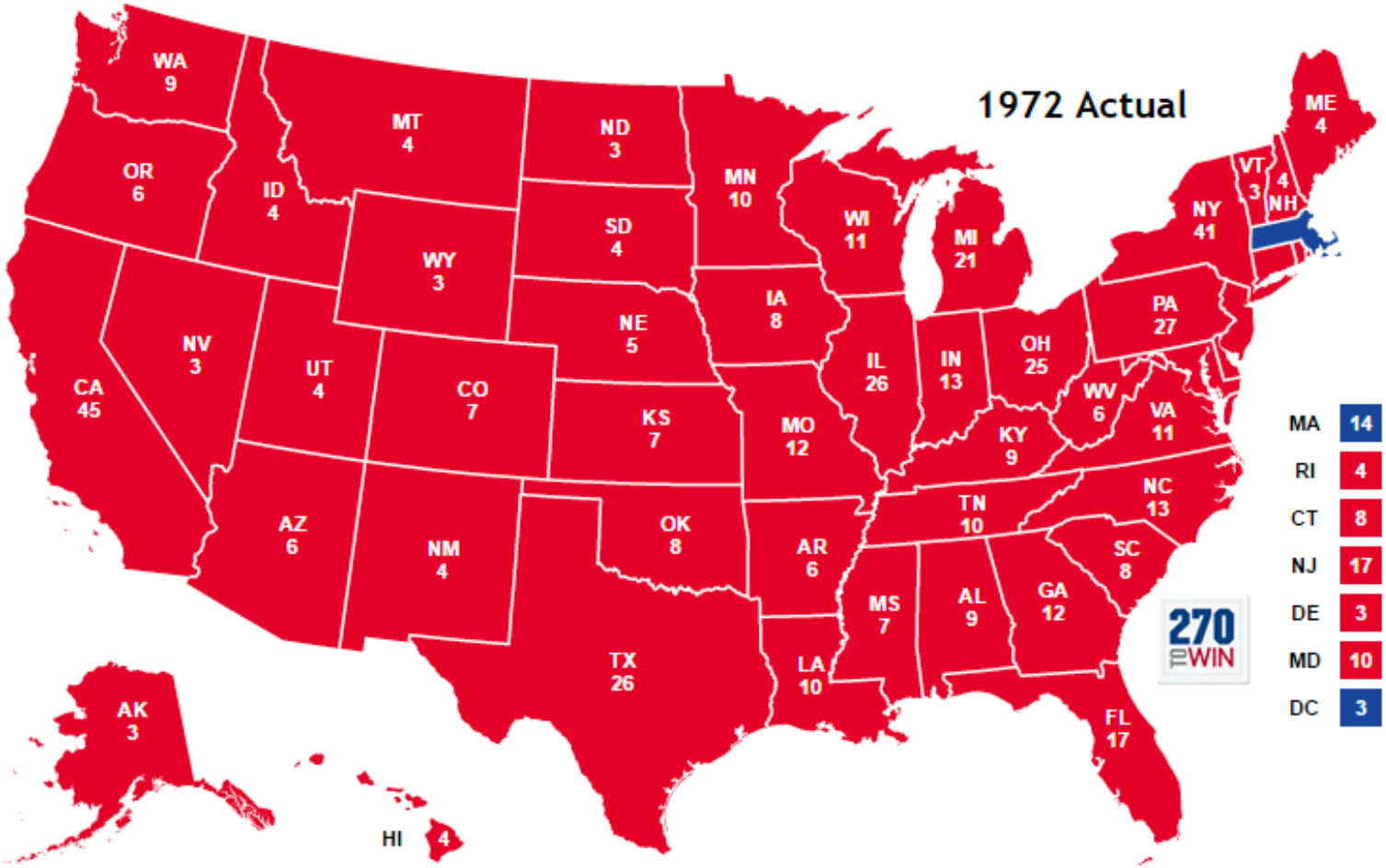
Selection Bias

Selection bias is the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed.

“I live in a rather special world. I only know one person who voted for Nixon. Where they are I don’t know. They’re outside my ken. But sometimes when I’m in a theater I can feel them.”

Pauline Kael, New Yorker film critic

1972 Electoral Map



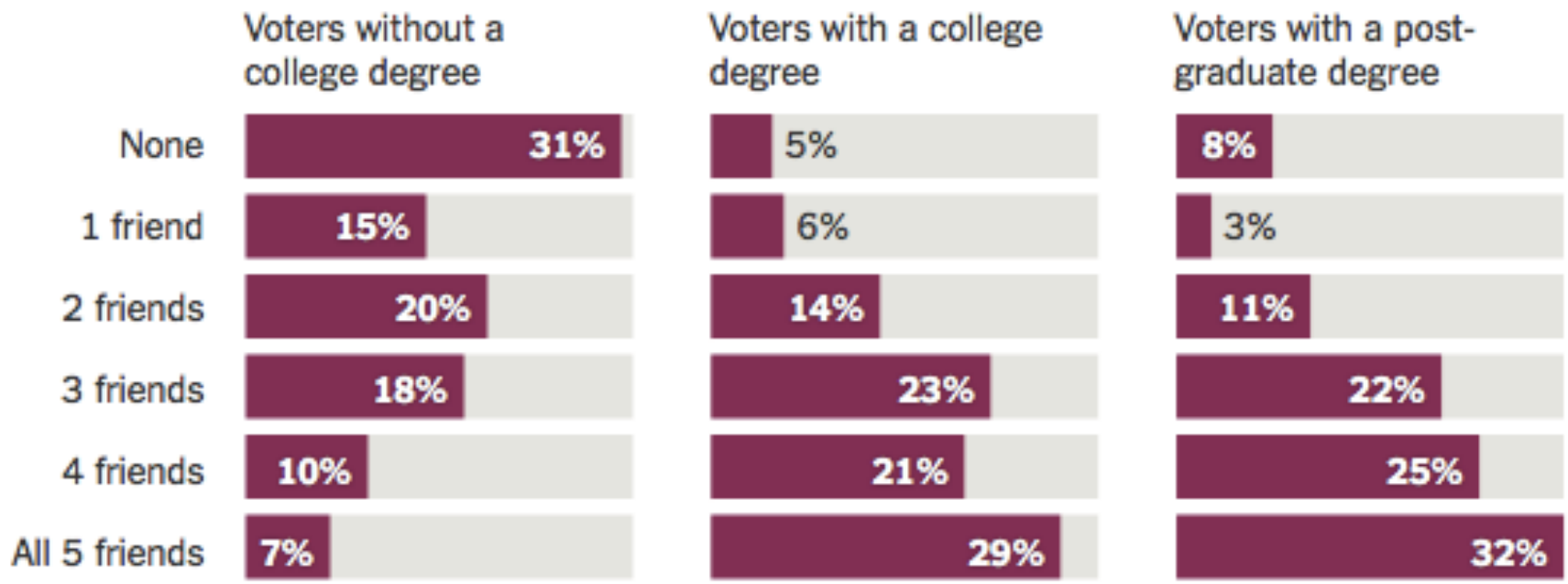
A Question About Friends Reveals a Lot About Class Divides

Damon Darlin @darlin SEPT. 1, 2016

5

Friends With College Degrees

Registered voters were asked how many of their five closest friends have four-year college degrees. The divide was stark.

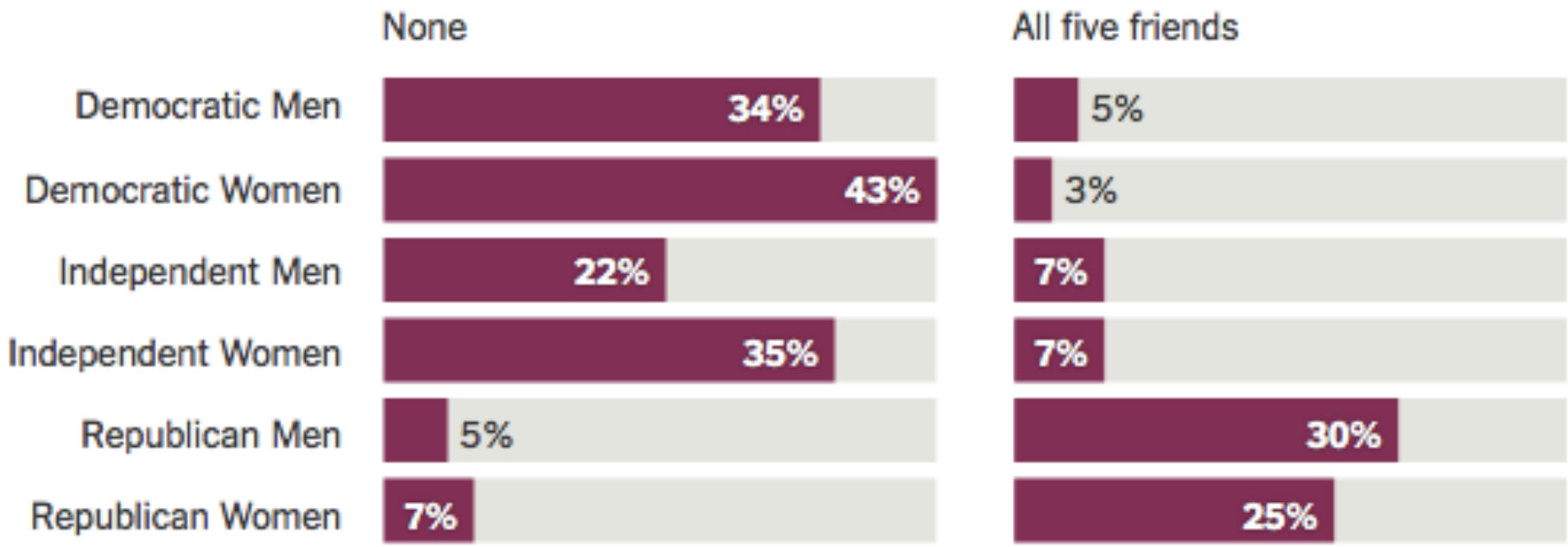


Source: Morning Consult

http://www.nytimes.com/2016/09/01/upshot/a-question-about-friends-reveals-a-lot-about-class-divides.html

Divided by Party

The divide was also sharply drawn when asked: How many of your five closest friends are Republicans?



Source: Morning Consult

United States presidential election, 1936

1932 ← November 3, 1936 → 1940

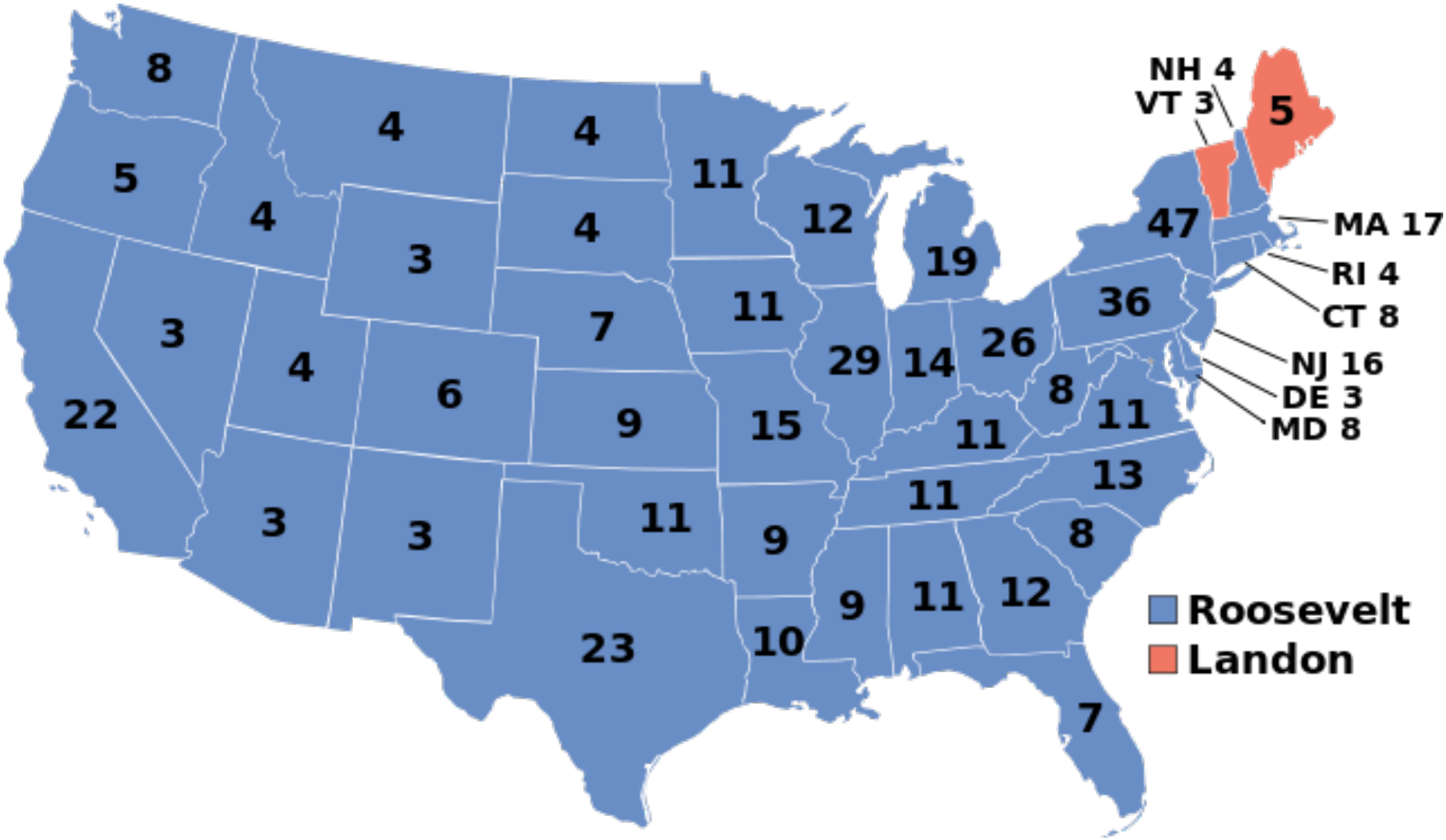
531 electoral votes of the Electoral College

266 electoral votes needed to win

Nominee	**Franklin D. Roosevelt**	Alf Landon
Party	Democratic	Republican
Home state	New York	Kansas
Running mate	**John Nance Garner**	Frank Knox

In the 1936 U.S. presidential election, The Literary Digest mailed out 10 million questionnaires, of which 2.3 million were returned. Based on this, they predicted that Republican Alf Landon would win with 370 of 531 electoral votes.

Actual Results … whoops



Poll of 10 million people —> 2.4 million responses

Who was polled?

- readers of Literary Digest
- registered automobile owners
- telephone users

What's the problem?

An illustration of Non-response bias

FOR IMMEDIATE RELEASE **July 30, 2016**

INTERVIEWS: Tom Jensen 919-744-6312

IF YOU HAVE BASIC METHODOLOGICAL QUESTIONS, PLEASE E-MAIL information@publicpolicypolling.com, OR CONSULT THE FINAL PARAGRAPH OF THE PRESS RELEASE

Clinton Image Improves Following Conventions; Leads Trump by 5

Raleigh, N.C. – PPP's new national poll, taken completely after both party's conventions, finds that Hillary Clinton emerged with a much more positive image than she had a month ago. Donald Trump meanwhile is just as unpopular as he was before the conventions.

Q9 Do you have a favorable or unfavorable opinion of Harambe?

Favorable	5%
Unfavorable	27%
Not sure	69%

Q10 If the candidates for President were Democrat Hillary Clinton, Republican Donald Trump, and independent Harambe, who would you vote for?

Hillary Clinton	48%
Donald Trump	43%
Harambe	5%
Not sure	4%

An illustration of Response bias

Other types of bias

Post hoc cherry picking

Selecting data with some special knowledge (violating randomness)

Publication bias (akin to survivorship bias)

Confirmation bias