CS 571 - Data Visualization & Exploration

Set and Text Visualizations

Instructor: Hamza Elhamdadi

UMassAmherst

Upcoming Dates

May 9:**

- **Homework 5 Due**
- **Project Screencast Submission Due**

May 12: Final Project Submission Due

****Only if you requested an extension by May 2**

5 Dataset Types

5 Dataset Types

5 Dataset Types

These are categorical data

A Metaphor..

A Metaphor..

Item: Lego

A Metaphor..

Item: Lego

Attributes: ???

A Metaphor..

Item: Lego

Attributes:

- color
- height
- width
- length
- shape

Dataset

Dataset

this is more realistic..

Dataset

Need to organize

How?

Dataset

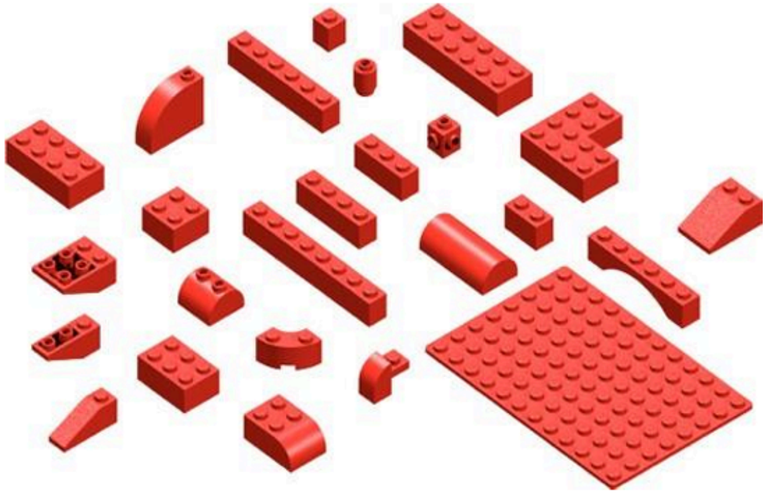
Need to organize

sort by **color**

Dataset

Need to organize

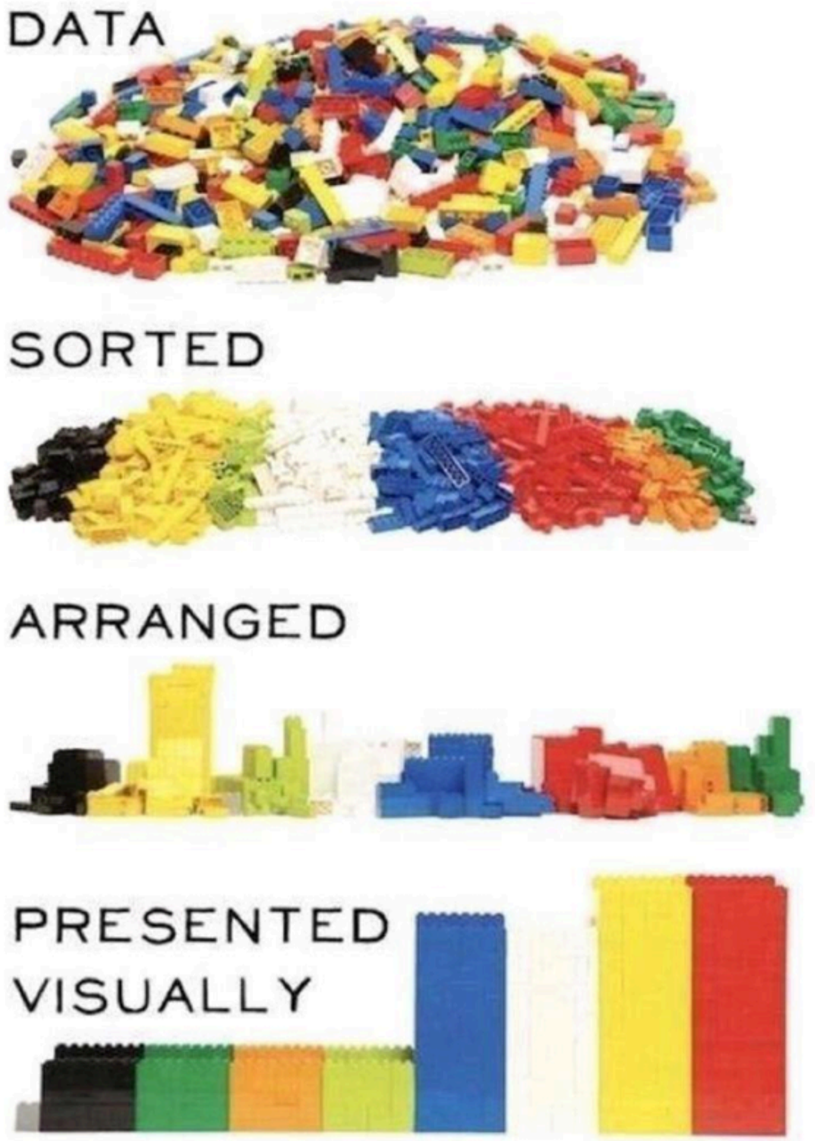
sort by **size/shape**



Sets

Organization leads us to sets

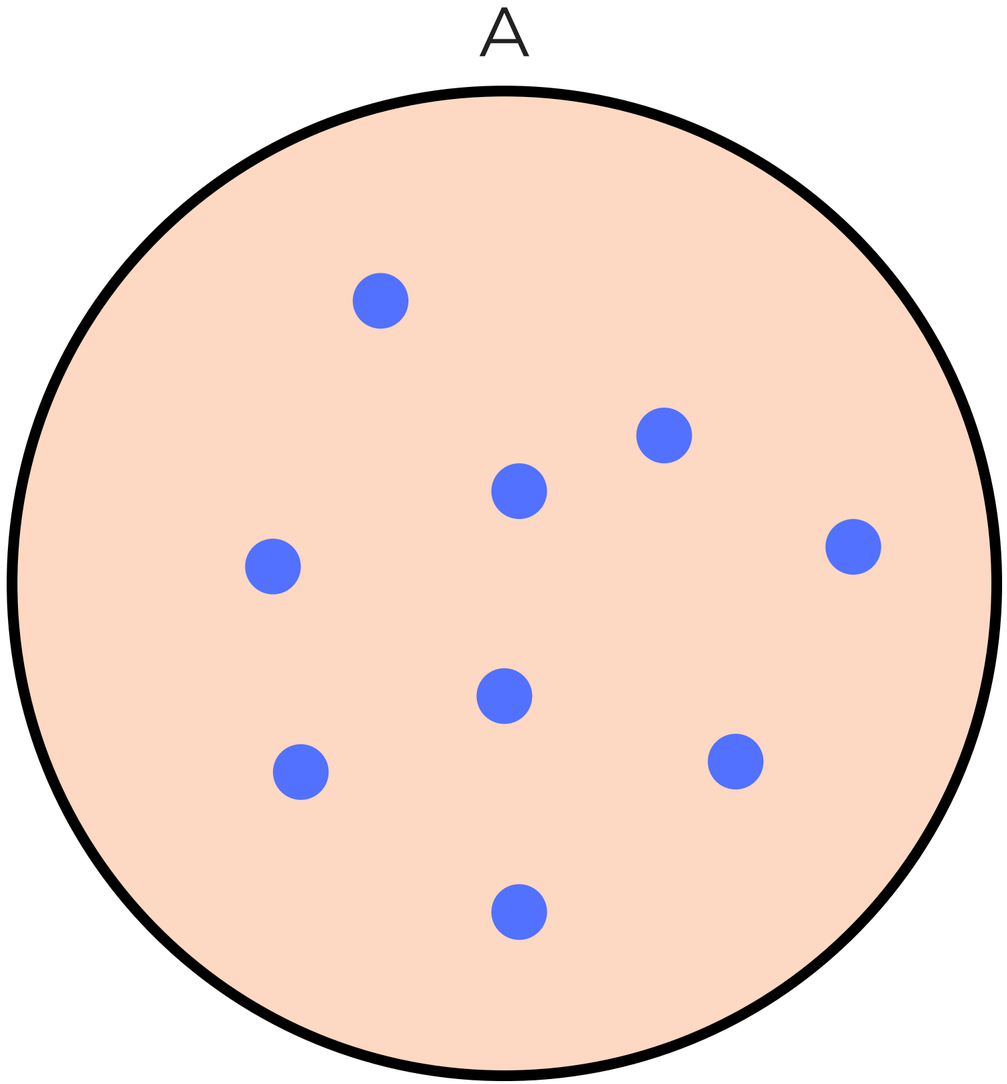
What are sets?



Set Theory

Set:

- A collection of objects
- e.g., set A



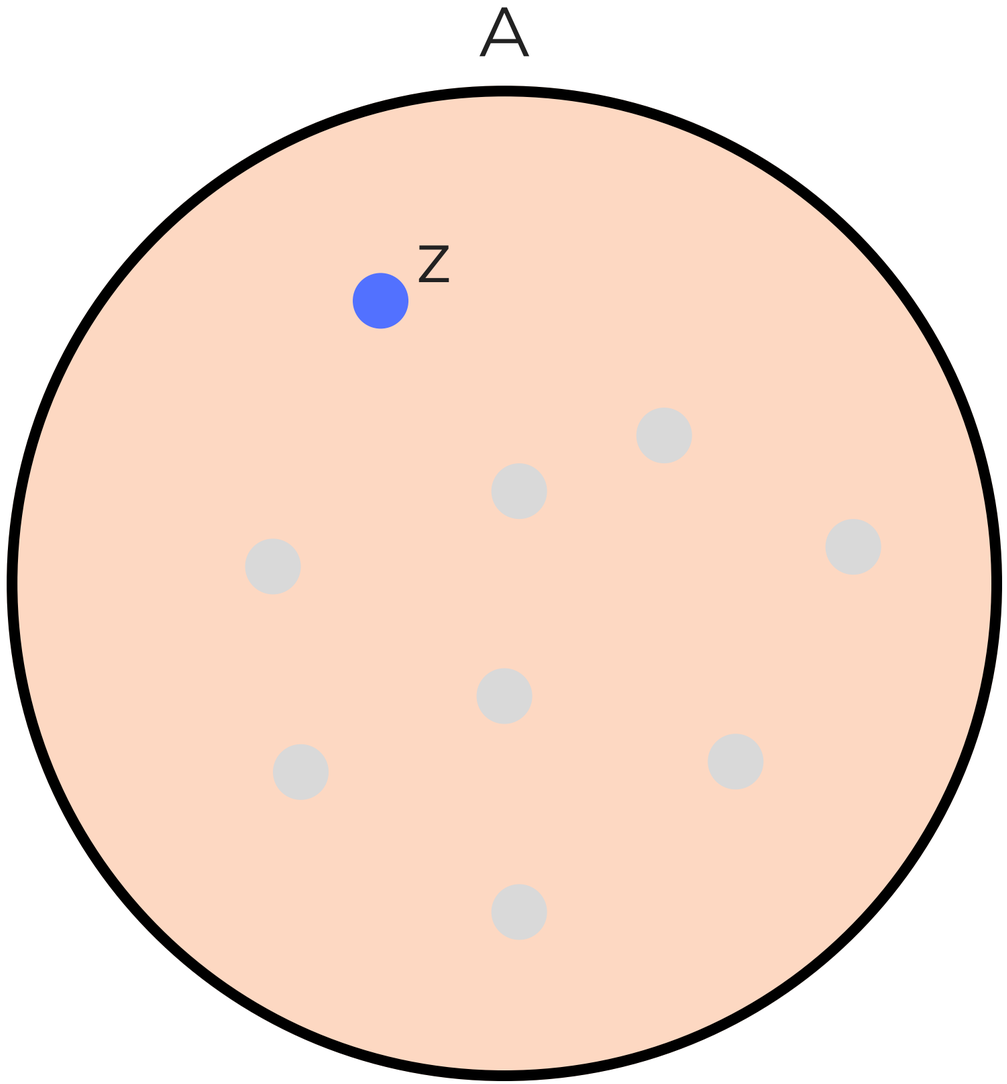
Set Theory

Set:

- A collection of objects
- e.g., set A

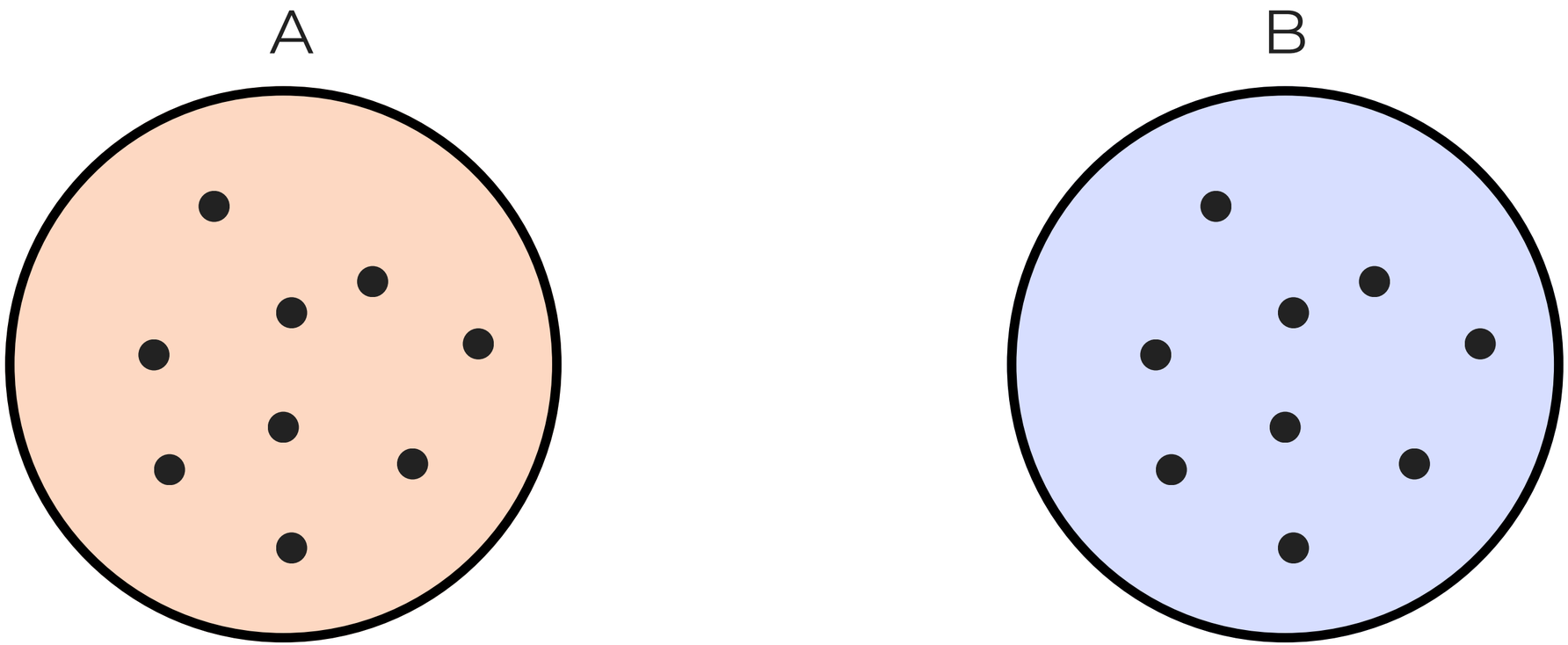
Object:

- element of a set
- e.g., $z \in A$



Set Theory

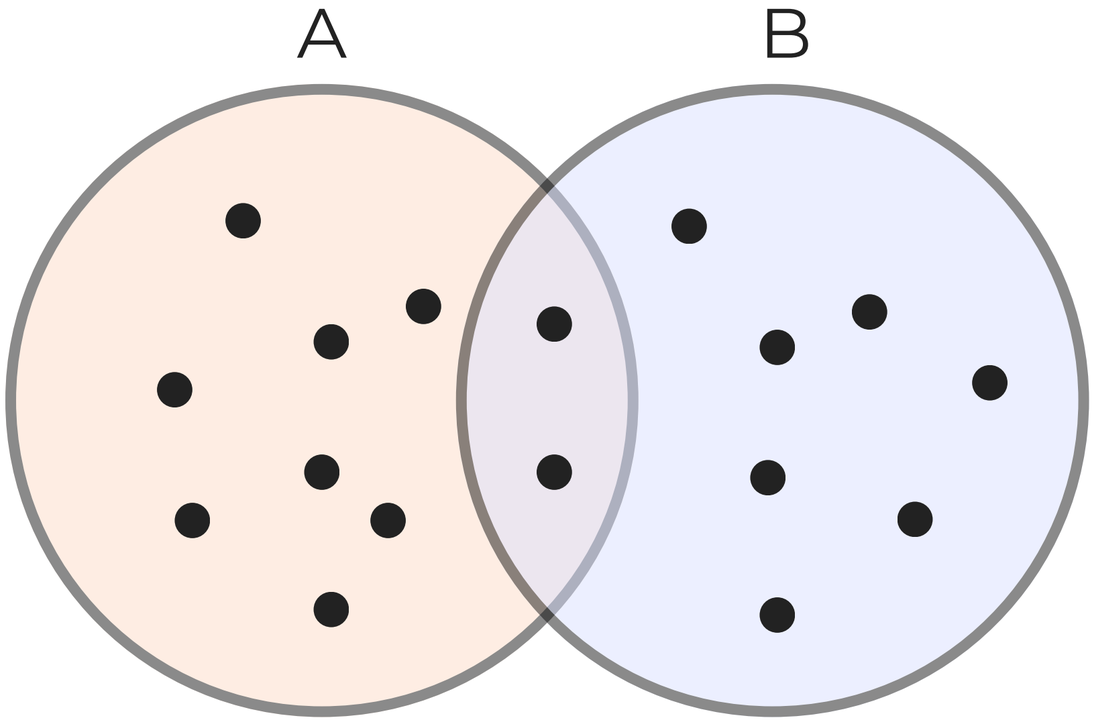
Given multiple sets (e.g., A and B)



Set Theory

Given multiple sets (e.g., A and B):

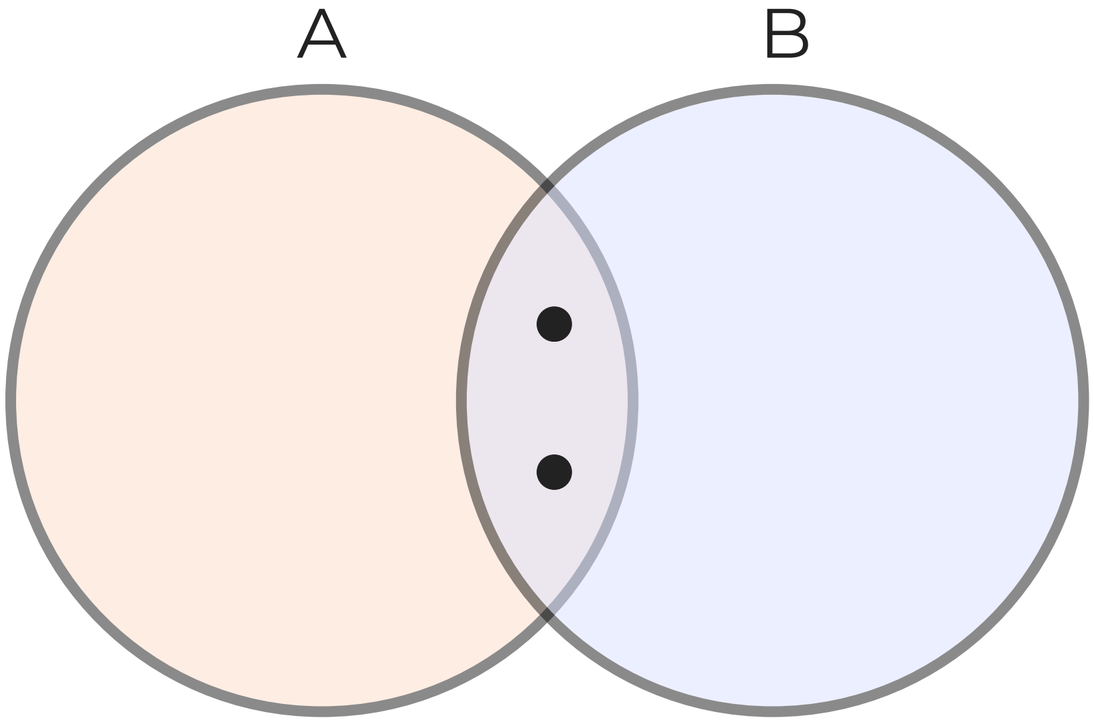
- union: A U B



Set Theory

Given multiple sets (e.g., A and B):

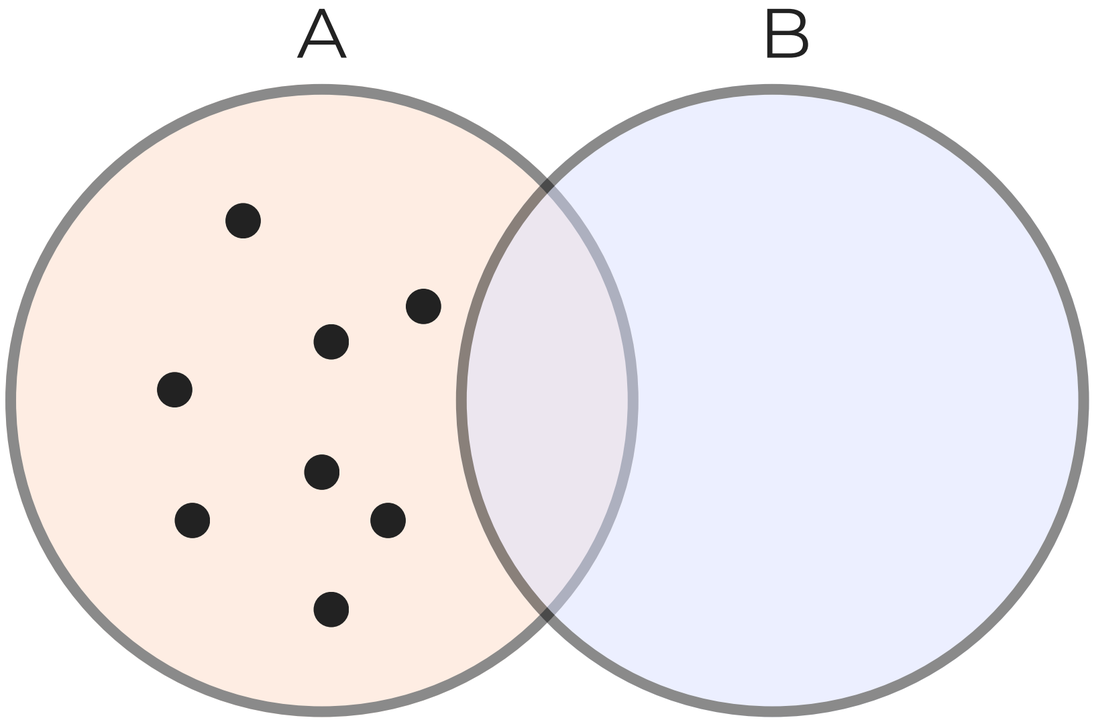
- intersection: $A \cap B$



Set Theory

Given multiple sets (e.g., A and B):

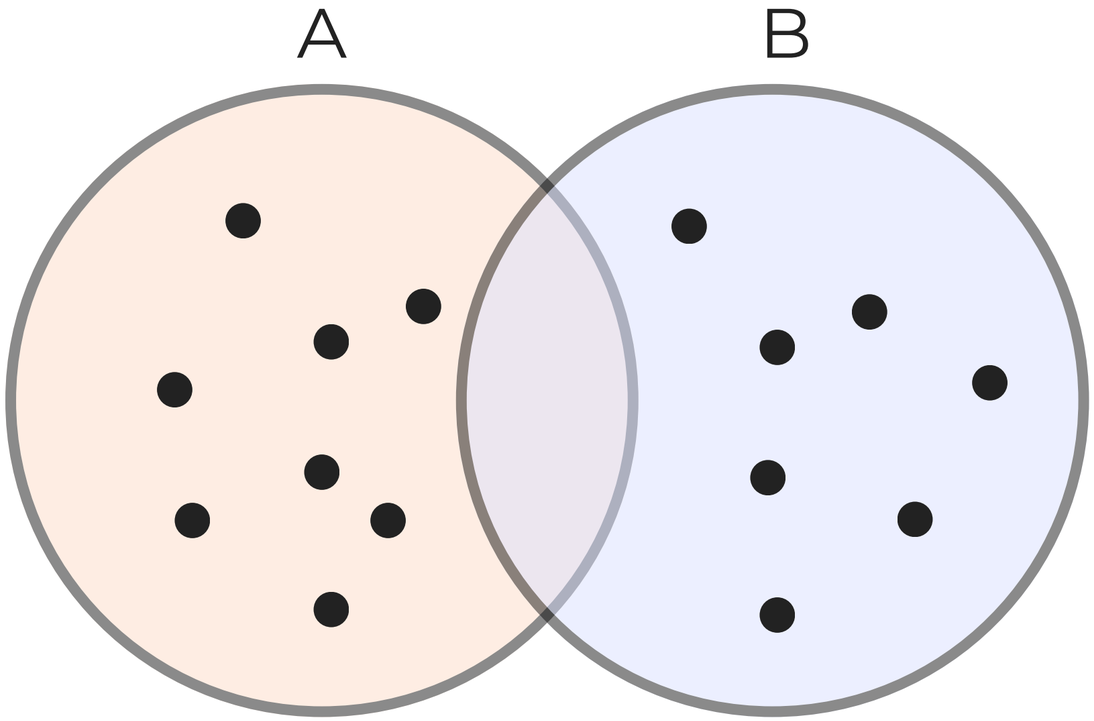
- set difference: A \ B



Set Theory

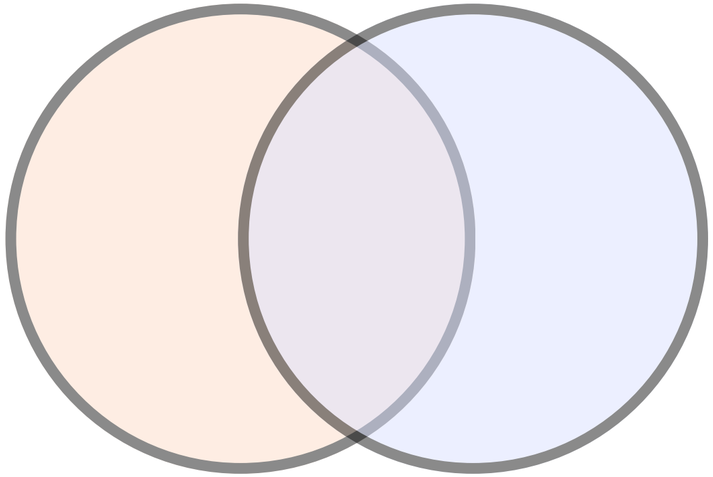
Given multiple sets (e.g., A and B):

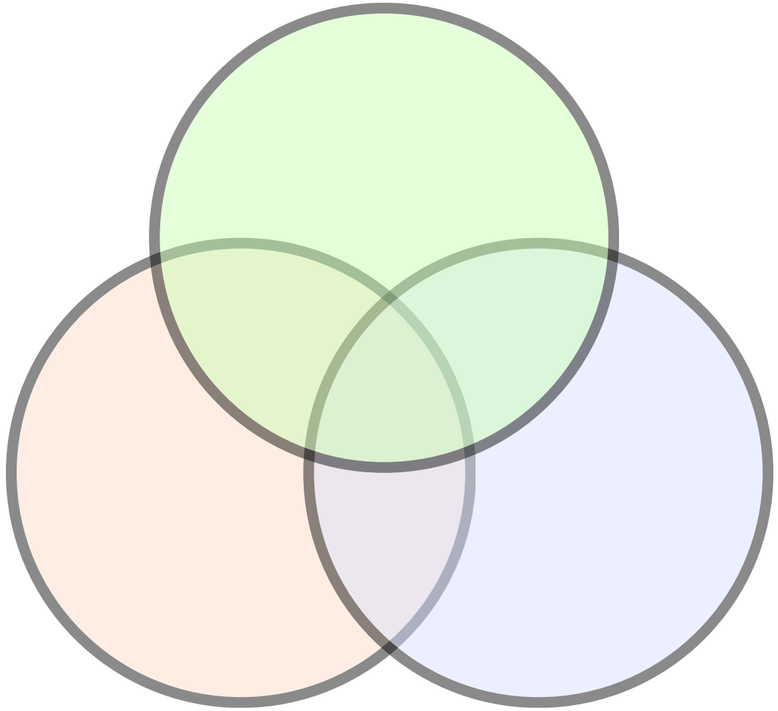
- symmetric difference: $A \ominus B$

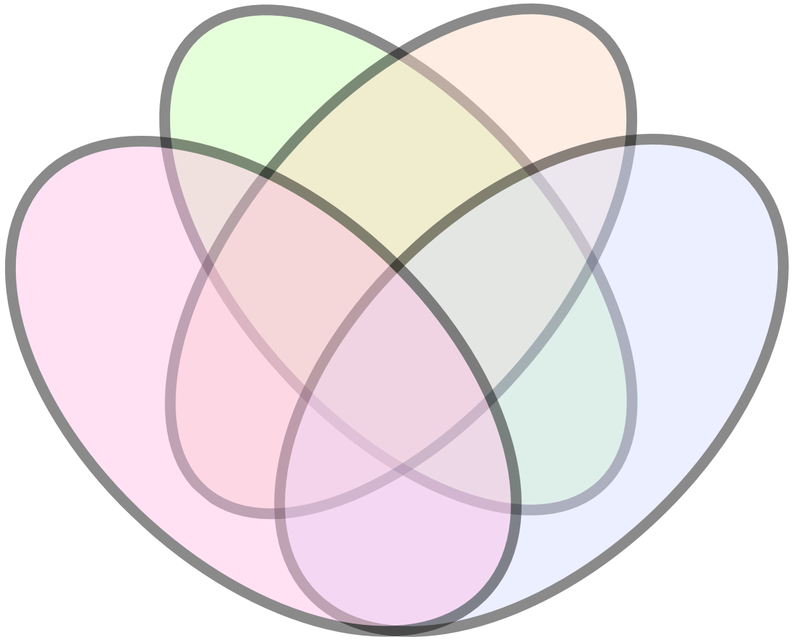


Venn Diagrams

Shows all possible relationships

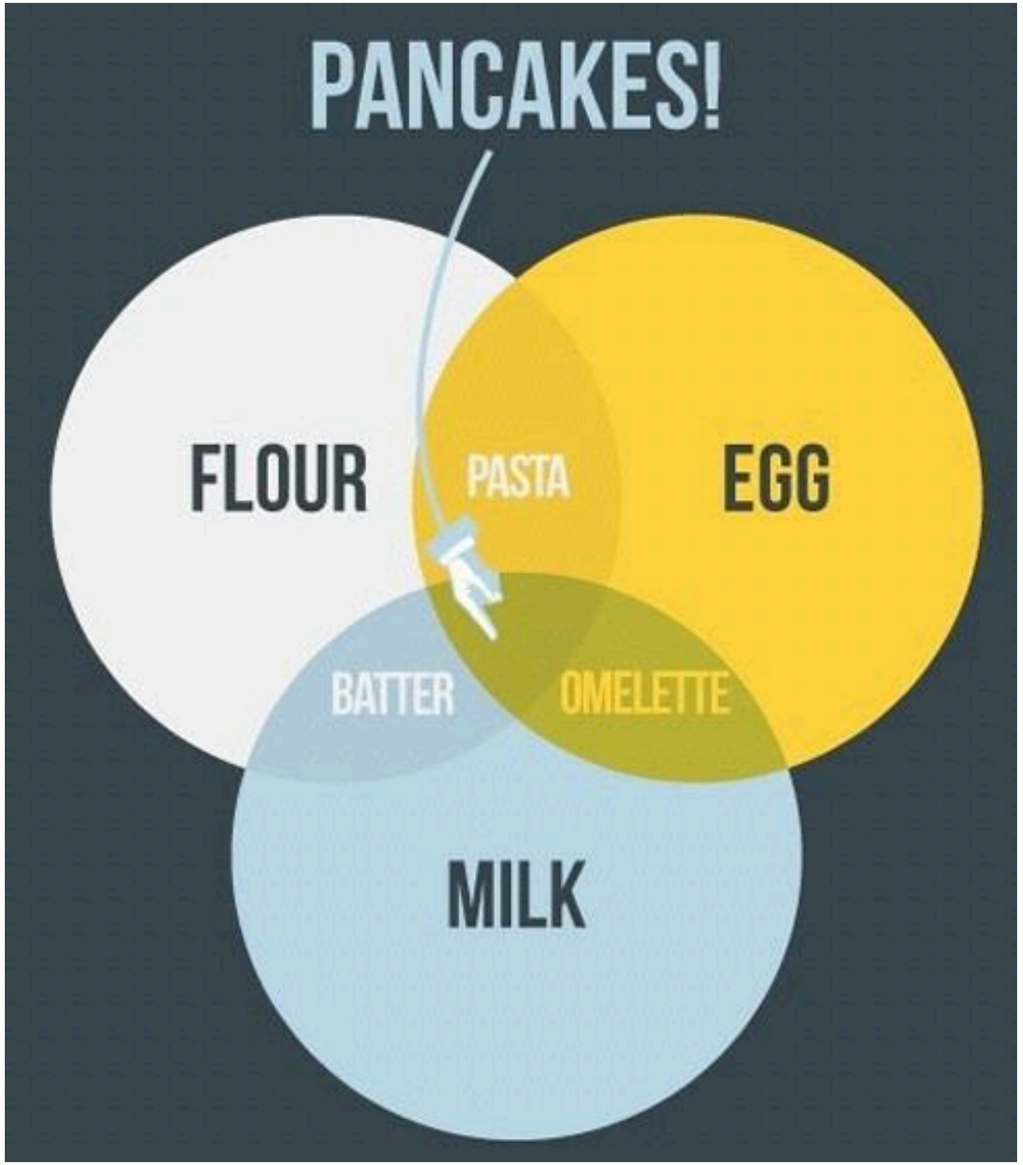






Venn Diagrams

“casual infovis”

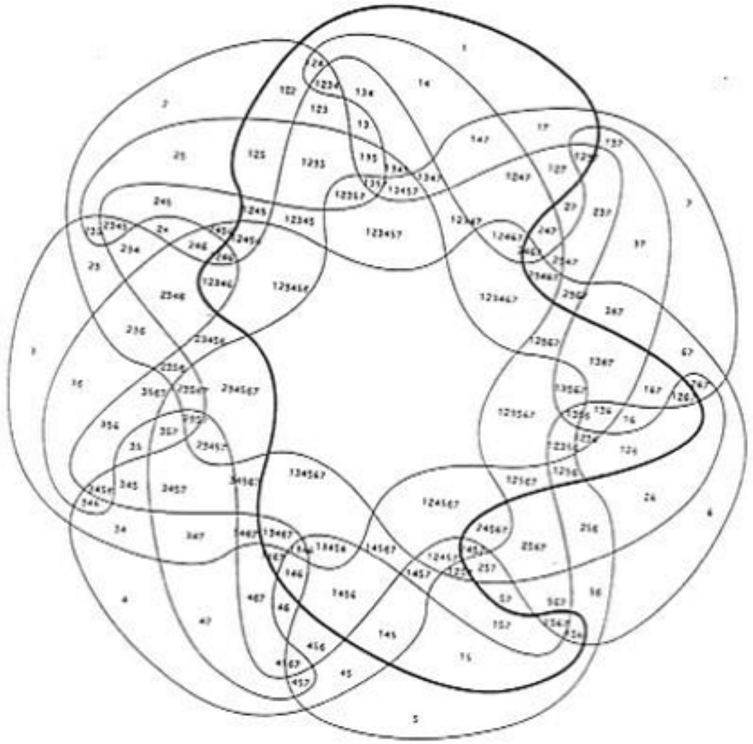


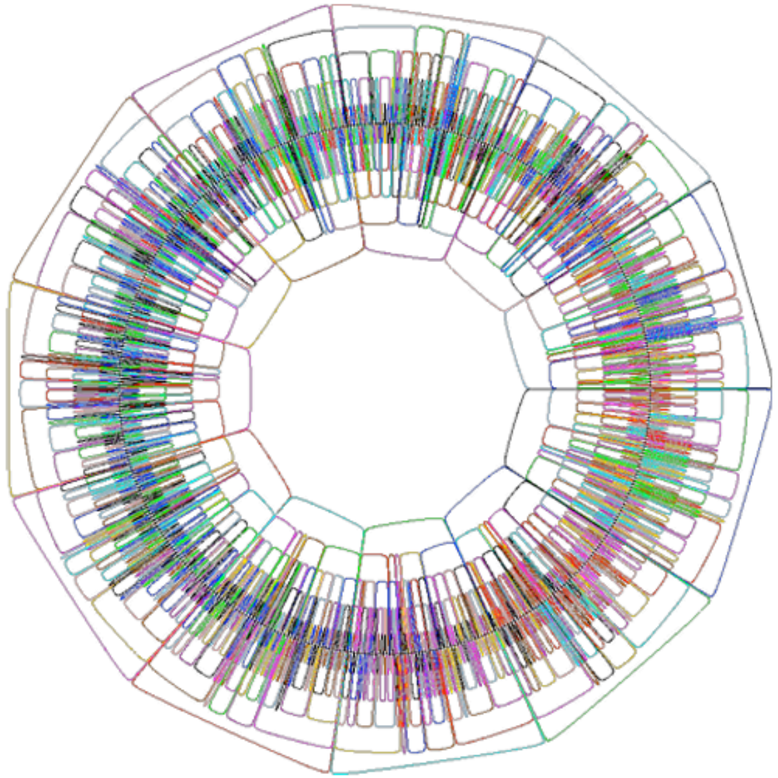
Venn Diagrams

get messy fast

Venn Diagrams

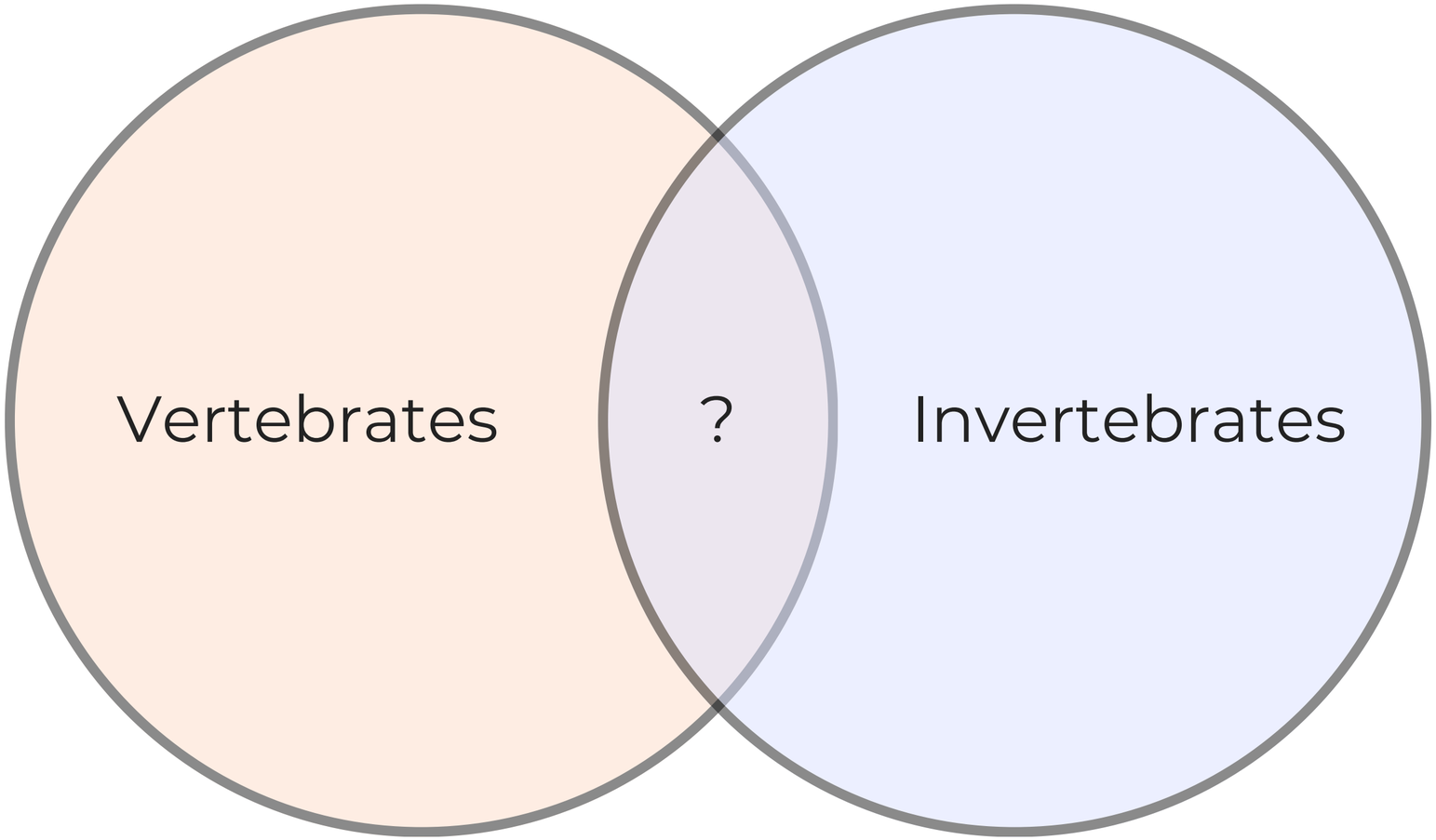
get messy fast





Venn Diagrams

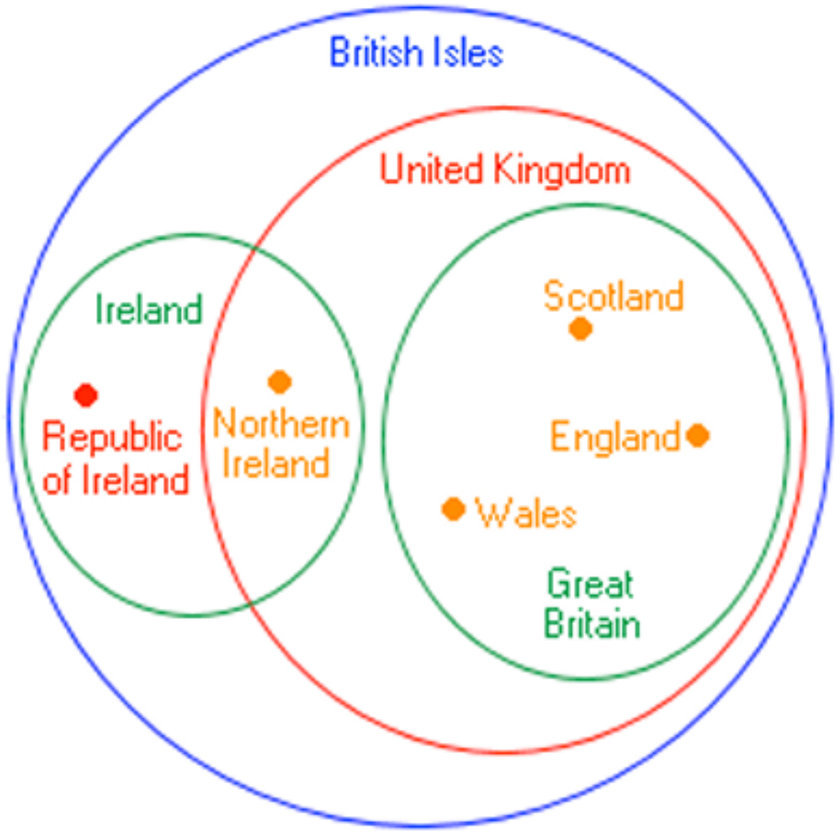
non-sensical?

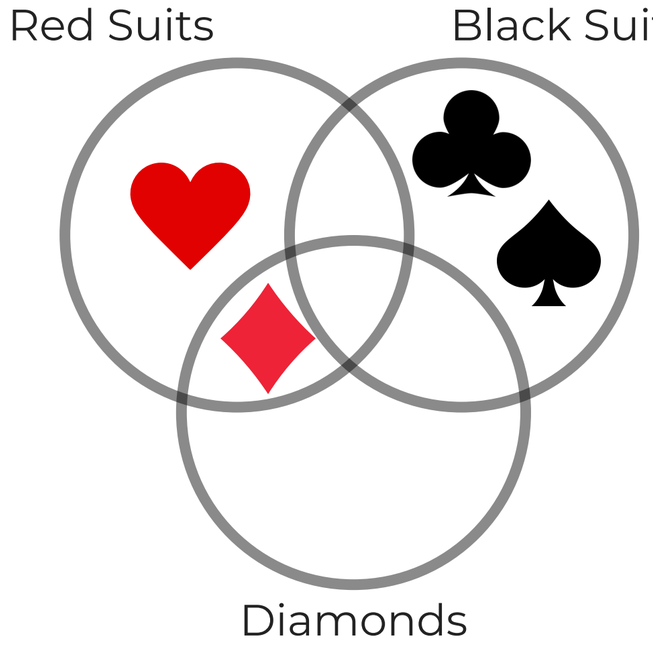


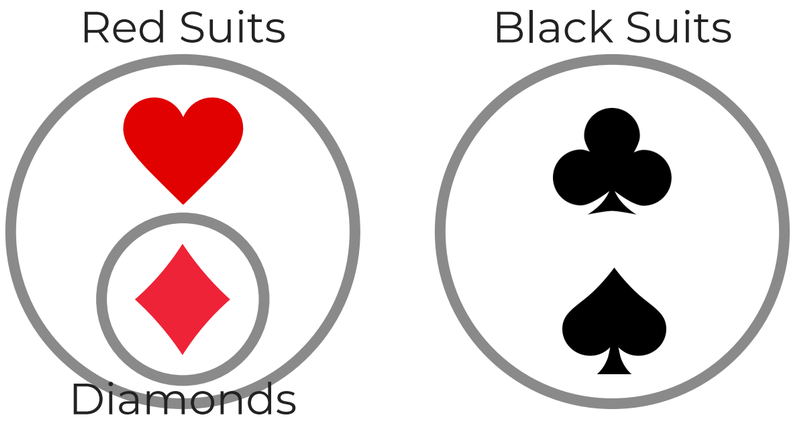
VENN

Euler Diagrams

only visualize existing relationships



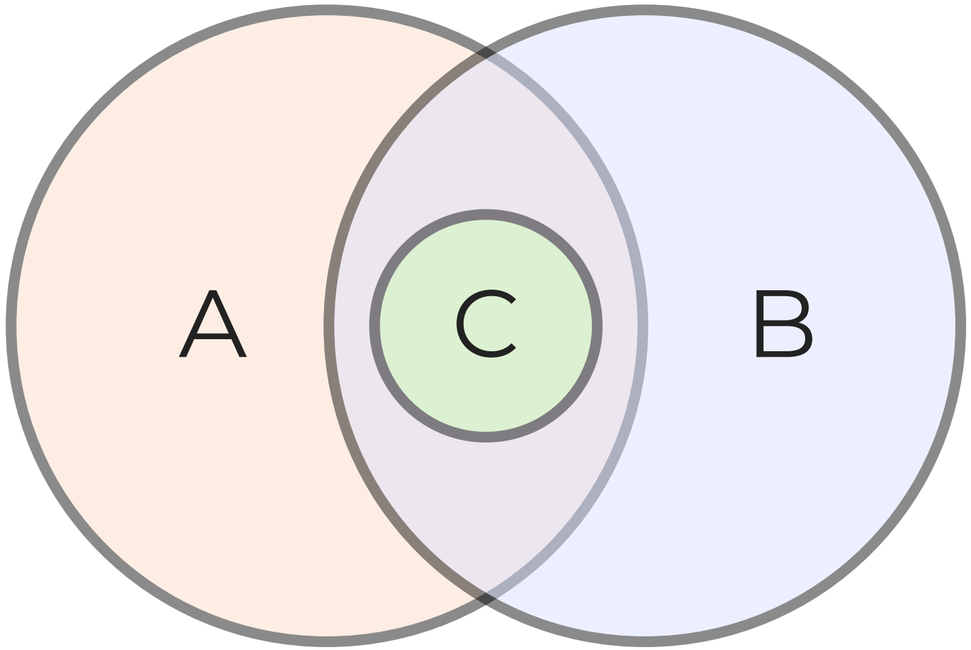




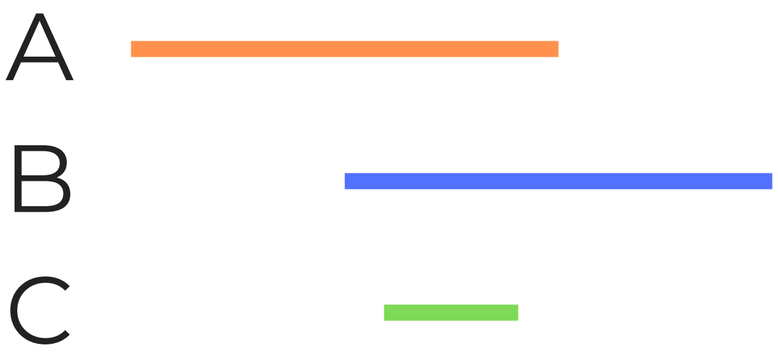
EULER

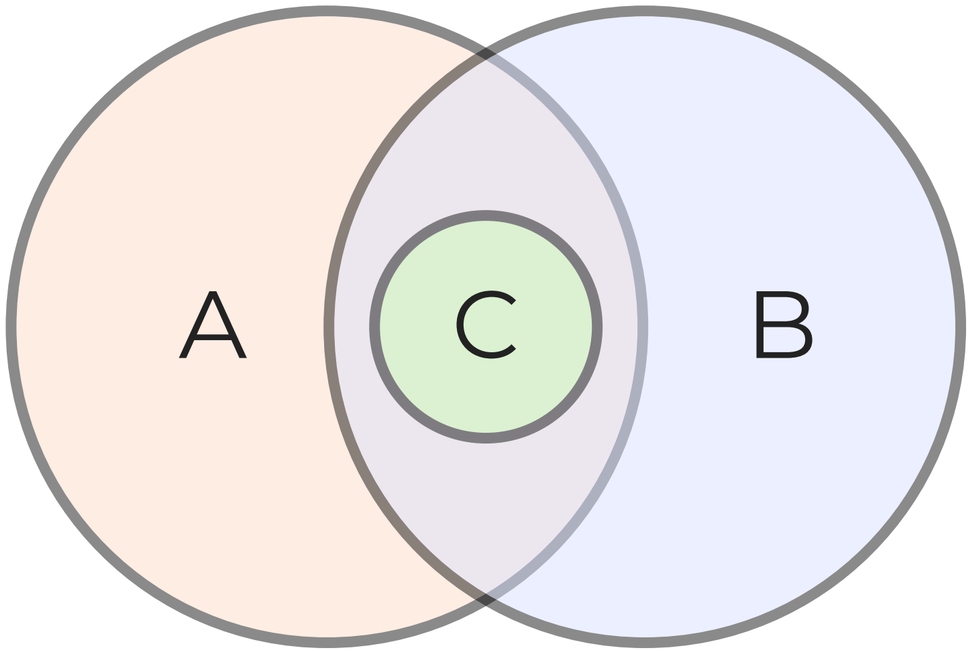
Euler Diagrams

A: People who **know what a Venn Diagram is**
B: People who **know what an Euler Diagram is**
C: People who **know the difference**

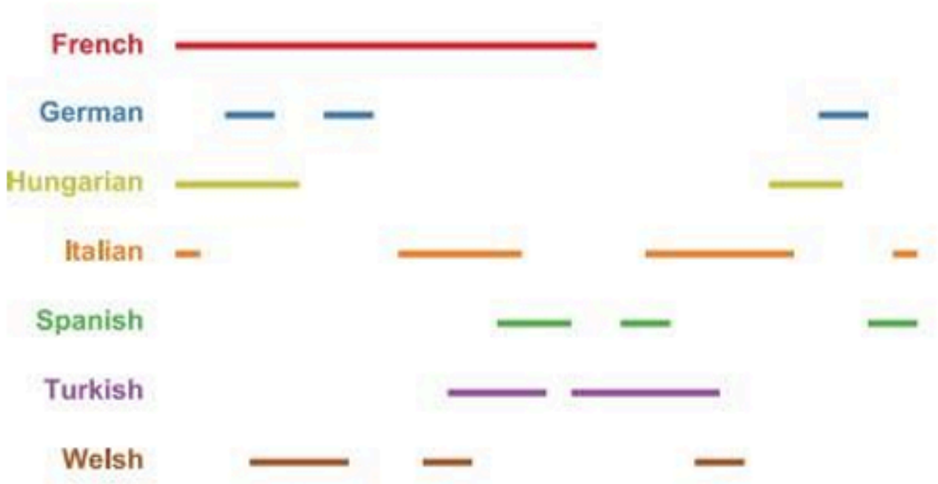


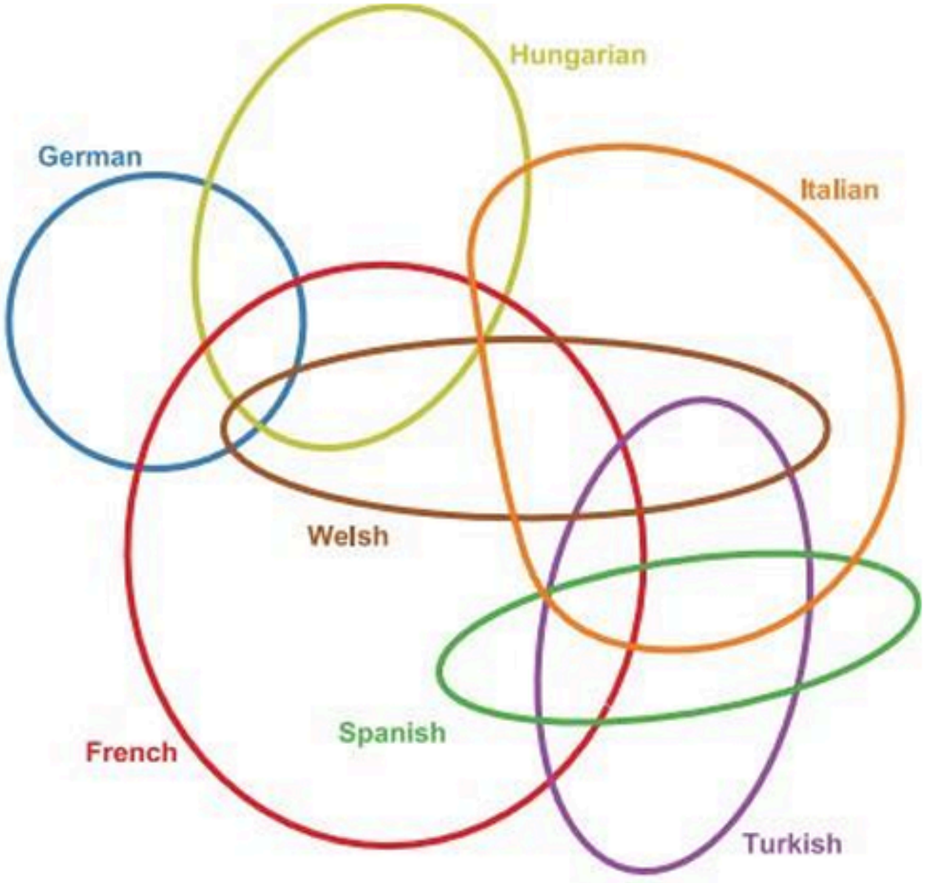
Linear Diagrams





Linear Diagrams

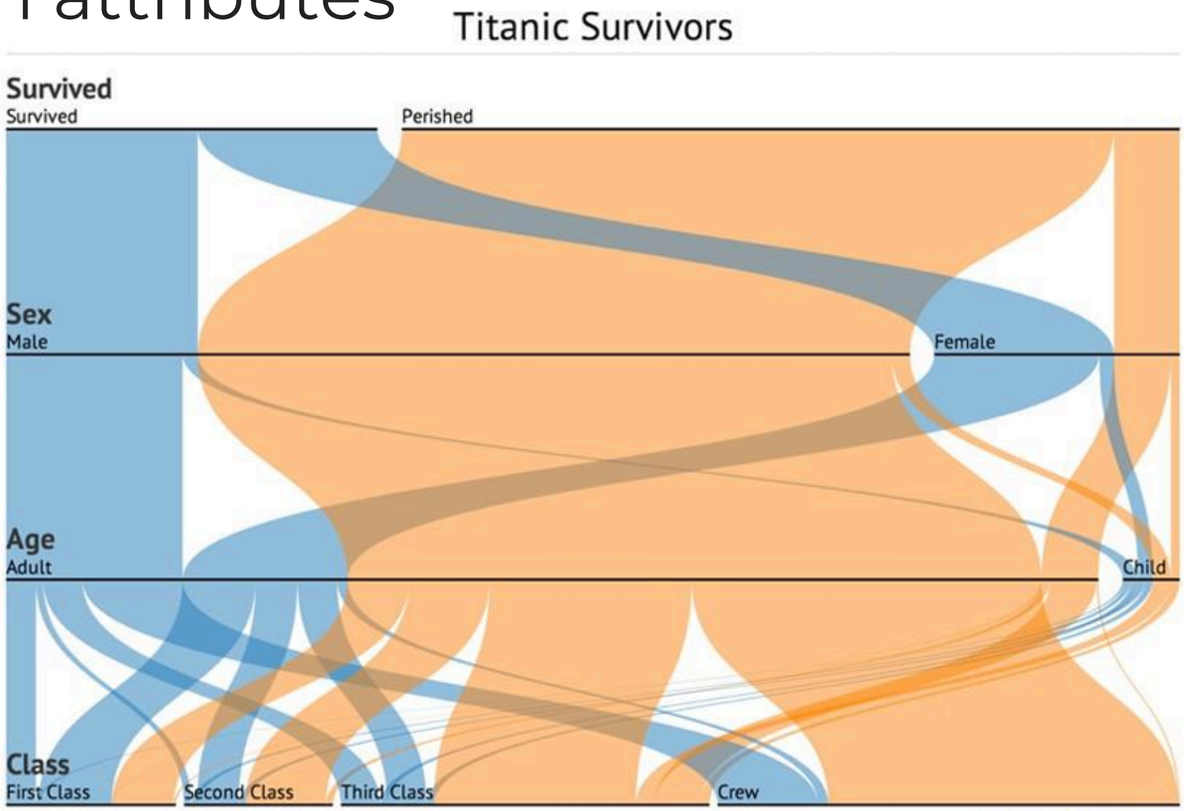




Parallel Sets

Like parallel coordinates, but for categorical data

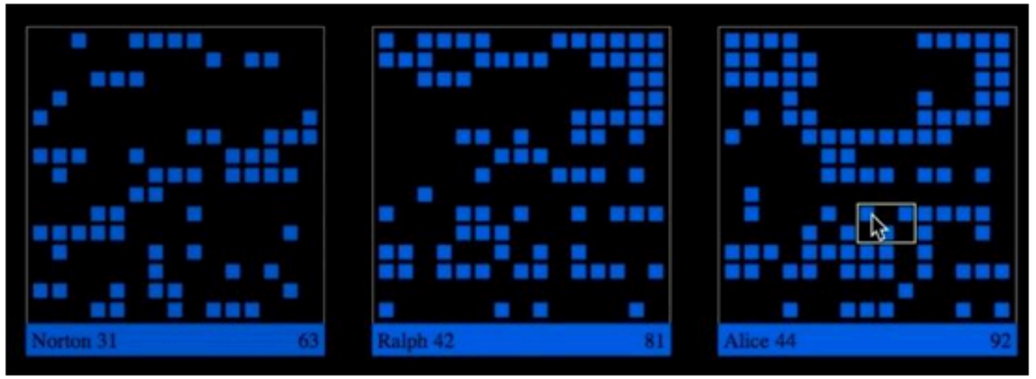
Task: Find relationship between attributes

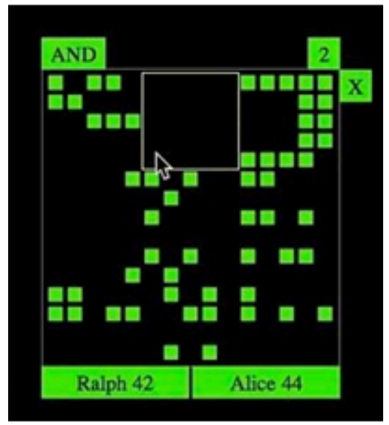


OnSet

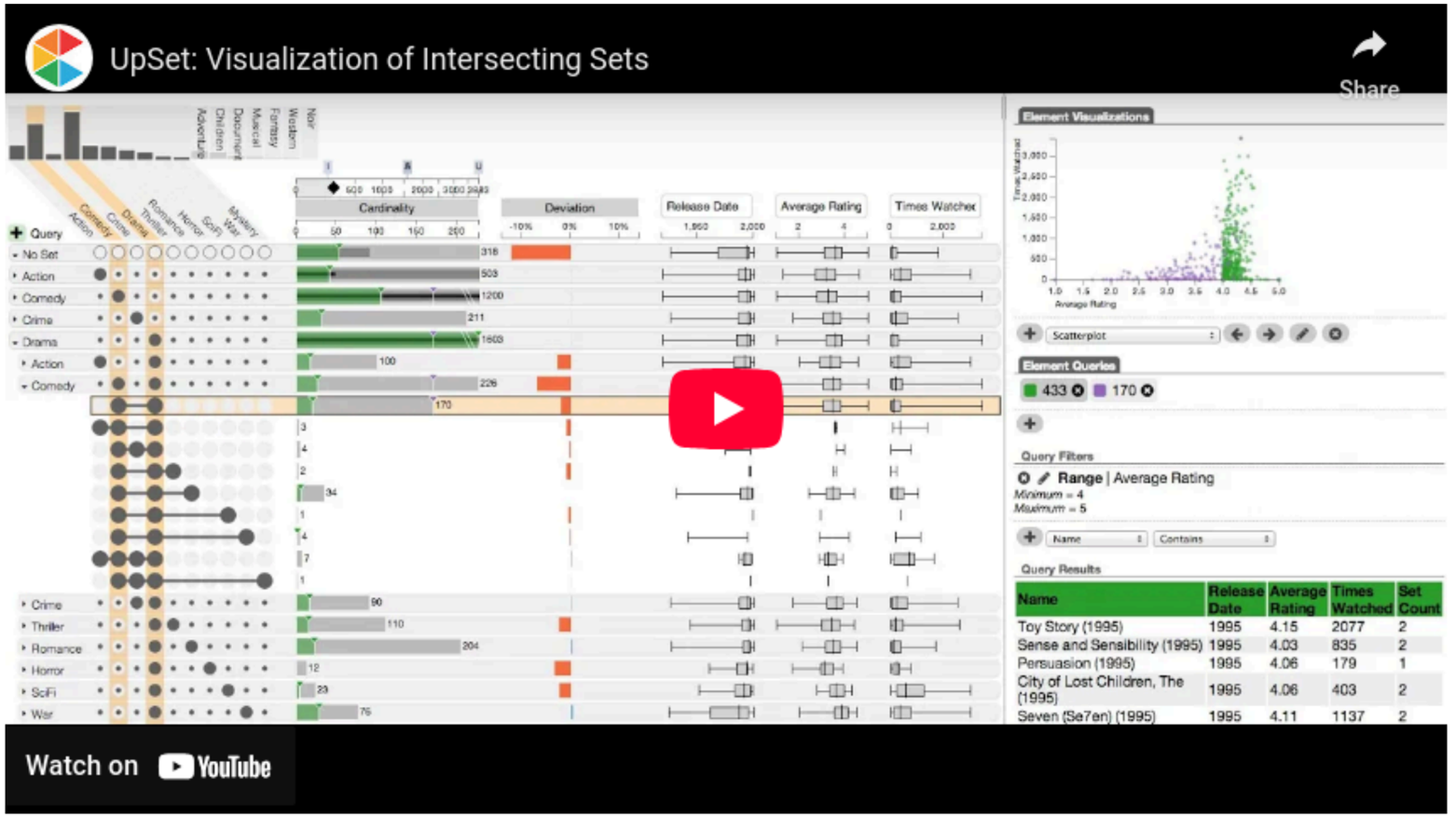
Set membership for each item (formatted as a matrix)

supports visual logic operations





UpSet



https://www.youtube.com/watch?v=-IfF2wGw7Qk

Set Visualization Takeaways:

Applies to many datasets:

- Many categorical datasets can be viewed as sets

Limited number of sets = more tractable

5 Dataset Types

5 Dataset Types (+1?)

Text is unique.

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists	Text
Items	Items (nodes)	Grids	Items	Items	
Attributes	Links	Positions	Positions		
	Attributes	Attributes			

Text is Everywhere

What kind of analysis questions might one ask about text and documents?

Which documents contain text about topic X?

Are there other documents that are similar to this one?

How are different words used in a document?

Which documents have an angry tone?

How does one set of documents differ from another set?

I want to understand the history of changes in a document.

We can think of text as data

Text is a different type of data than your typical quantities and categories

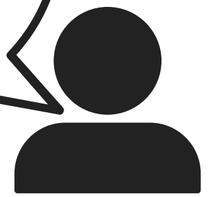
Text as Data

Visualizing text as data can include visualizing:

- individual documents
  - e.g., a news article, an email, a book, etc
- collections of documents
  - e.g., message exchanges, social media profiles, sets of academic publications, etc

Why should we visualize text?
Couldn’t we just read the raw text...

We visualize text to show **patterns, trends, and themes.**
Then, we can **focus our time reading only the parts**
of the text **that are relevant.**



Benefits of Visualizing Text

Fast-track understanding: Visualizing text allows us to get the **“gist” of a document** in a less linear fashion.

Group: Visualizing text allows us to **cluster** the documents to get an **overview**.

Compare: Visualizing text allows us to **compare collections of documents** or **inspect the same collection over time**.

Correlate: We might want to see **how patterns in text relate to other data** (e.g., frequency of a company mentioned in the media with that company’s stock prices).

Example: We can compare the transcripts of speeches from Clinton and Obama to see how their stances on healthcare differed

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of you about an issue that is central to that future – and that is the issue of health care.

I am not the first President to take up this cause, but I am determined to be the last. It has now been nearly a century since Theodore Roosevelt first called for health care reform. And ever since, nearly every President and Congress, whether Democrat or Republican, has attempted to meet this challenge in some way. A bill for comprehensive health reform was first introduced by John Dingell Sr. in 1943. Sixty-five years later, his son continues to introduce that same bill at the beginning of each session.

Clinton's Health Plan: 'Security, Simplicity, Savings, Choice, Quality and Responsibility'

Transcript of President's Address to Congress on Health Care

Following is a transcript of President Clinton's address to a joint session of Congress last night on his plan to overhaul the nation's health care system, as recorded by The New York Times.

Mr. Speaker, Mr. President, members of Congress, distinguished guests, my fellow Americans: Before I begin my words tonight, I would like to ask that we all bow in a moment of silent prayer for the memory of those who were killed and those who have been injured in the tragic train accident in Alabama today.

My fellow Americans, tonight we come together to write a new chapter in the American story.

Our forebears enshrined the American dream — life, liberty, the pursuit of happiness. Every generation of Americans has worked to strengthen that legacy, to make our country a place of freedom and opportunity, a place who — where people who work hard can rise to their full potential and a place where their children can have a better future.

From the settling of the frontier to the landing on the moon, ours has been a continuous story of challenges defined, obstacles overcome, new horizons secured. That is what makes America what it is and Americans what we are.

Now we are in a time of profound change and opportunity. The end of the cold war, the information age, the global economy have brought us both opportunity and hope and strife and uncertainty.

Our purpose in this dynamic age must be to change, to make change our friend and not our enemy.

To achieve that goal, we must face all our challenges with confidence, with faith and with discipline, whether we're reducing the deficit, creating tomorrow's jobs and training our people to fill them, converting from a high-tech defense to a high-tech domestic economy, expanding trade, reinventing government, making our streets safer or rewarding work over idleness. All these challenges require us to change.

If Americans are to have the courage to change in a difficult time, we must first be secure in our most basic needs.

Fixing a Broken System

Tonight I want to talk to you about the most critical thing we can do to build that security.

This health care system of ours is badly broken and it is time to fix it.

Despite the dedication of literally millions of talented health care professionals, our health care is too uncertain and too expensive, too bureaucratic and too wasteful. It has too much fraud and too much greed.

At long last, after decades of false starts, we must make this our most urgent priority: giving every American health security

Jose R. Lopez/The New York Times

President Clinton acknowledging applause last night. Behind him were Vice President Al Gore and Speaker Thomas S. Foley.

look up to. They are what is right with this health care system. But we also know that we can no longer afford to continue to ignore what is wrong.

Millions of Americans are just a pink slip away from losing their health insurance, and one serious illness away from losing all their savings. Millions more are locked into the

Under our plan every American would receive a health care security card that will guarantee a comprehensive package of benefits over the course of an entire lifetime roughly comparable to the benefit package offered by most Fortune 500 companies.

This health care security card will offer this package of benefits in a way that can

than they would if they had regular treatment in a way that only adequate medicine can provide.

I also believe that over time we should phase in long-term care for the disabled and the elderly on a comprehensive basis.

Toward a Simpler System

and we can change it.

And doctors, nurses and consumers shouldn't have to worry about the fine print. If we have this one simple form there won't be any fine print. People will know what it means.

Toward Savings in the System

The third principle is savings.

Reform must produce savings in this health care system — it has to. We are spending over 14 percent of our income on health care; Canada's at 10; nobody else is over 9. We're competing with all these people for the future and the other major countries, they cover everybody. And they cover them with services as generous as the best company policies here in this country.

Rampant medical inflation is eating away at our wages, our savings, our investment capital, our ability to create new jobs in the private sector and this public treasury.

You know the budget we just adopted had steep cuts in defense, a five-year freeze on the discretionary spending so critical to reeducating America and investing in jobs and helping us to convert from a defense to a domestic economy. But we passed the budget, which has Medicaid increases of between 16 and 11 percent a year over the next five years and Medicare increases of between 11 and 9 percent in an environment where we assume inflation will be at 4 percent or less.

We cannot continue to do this. Our competitiveness, our whole economy, the integrity of the way the Government works, and ultimately our living standards, depend upon our ability to achieve savings without harming the quality of health care. Unless we do this, our workers will lose $655 in income each year by the end of the decade. Small businesses will continue to face skyrocketing premiums, and a full third of small businesses now covering their employees say they will be forced to drop their insurance.

Large corporations will bear bigger disadvantages in global competition, and health care costs will devour more and more and more of our budget. Pretty soon all of you or the people who succeed you will be showing up here, and writing out checks for health care and interest on the debt and worrying about whether we've got enough defense, and that'll be it unless we have the courage to achieve the savings that are plainly there before us. Every state and local government will continue to cut back on everything from education to law enforcement to pay more and more for the same health care.

These rising costs are a special nightmare for our small businesses, the engine of our entrepreneurship and our job creation in American today. Health care premiums for small businesses are 35 percent higher than those of large corporations today, and they will keep rising at double-digit rates unless we act.

Word (Tag) Clouds

Visualize the frequency of words in a document

Obama’s Speech (2009)

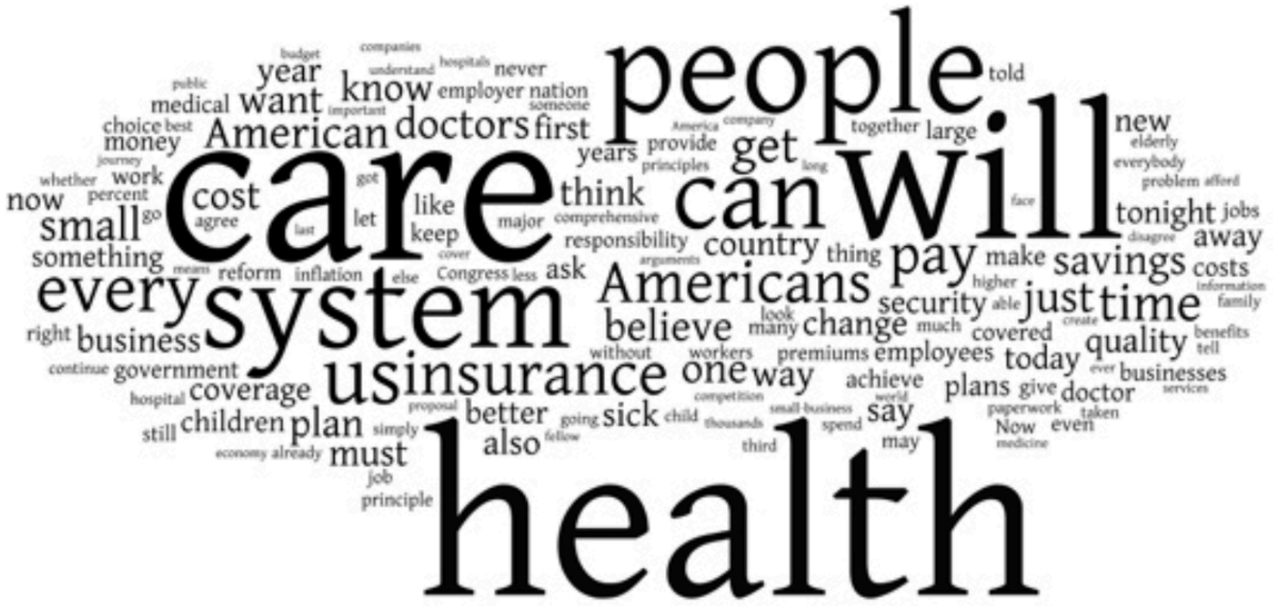
Word (Tag) Clouds

Visualize the frequency of words in a document

Clinton’s Speech (1993)

Obama’s Speech (2009)

Clinton’s Speech (1993)



Portrait of the candidate as a pile of words

What's the most frequent word on McCain's blog? "Obama."

BY JOHN SCHWENKLER

BOTH OF THIS year's major party presidential candidates have made official campaign blogs a central part of their virtual headquarters. John McCain campaign hired former Weekly Standard blogger Michael Goldfarb as its deputy communications director and made the "McCain Report" into his one-man show. The Obama team has taken a more collaborative approach, allowing supporters to create their own blogs and then moving some of that content, together with snippets of news coverage and messages from campaign officials, to the official "Obamablog" front page. The results are revealing. Online software at Wordle.net makes it easy to portray the blogs visually, with the most frequently used words displayed largest. A snapshot from late last week:

MCCAIN'S BLOG

obama Given the way the McCain campaign has been complaining about unbalanced media attention, it is puzzling to see that it has been as fixated on Obama as every one else. But while fawning media coverage is one thing, successfully turning the election into a national referendum on the relatively inexperienced junior senator from Illinois just might yield McCain an improbable victory. It could also backfire, though, if voters tired of excessively negative politicking and end up looking kindly on the fresher of the two faces.

visit The ongoing attempt to stir controversy over the cancellation of Obama's scheduled visit to the Landstuhl Army Hospital has also been a consistent part of the blogging strategy at the McCain Report, and in this case the growing backlash against these attacks (even Republican bloggers have called them unfair) has only managed to feed that tendency, with each round of explanations being picked through, point by point.

drilling Following the lead of the national GOP, the McCain Report has recently started hammering away at the need to drill for oil. This is a bit of an unnatural move for McCain—it hasn't been long since his most-publicized breaks from Republican orthodoxy on exactly this issue—but it may be one he has to make as both candidates look to establish a compelling message on the economy.

opposes Barack Obama's opposition to offshore oildrilling and the federal gas tax holiday has led the McCain Report to start calling him "Dr. No." Most analysts agree that such policies are unlikely to have any significant effect on fuel prices, but McCain's proposals nevertheless satisfy the widespread desire for the government to step up and do something, while Obama's calls for individual restraint are not what many voters want to hear.

OBAMA'S BLOG

future Barack Obama's youth and relative newness to national politics make him the perfect candidate to center his message on leaving the past behind and moving into a better future. This sort of progressive rhetoric has the effect of keeping people optimistic while still reminding them that the status quo just isn't good enough.

change Obama's trademark theme of change is still figuring prominently in his campaign rhetoric, and with good reason; the national mood is deeply pessimistic, and everyone is looking for signs of something different. While this sort of sloganeering is decidedly vague, it has worked well enough so far that there's no reason for the Obama campaign to abandon it now.

donate The Obama campaign's official blogging constantly emphasizes the importance of donating to supporting candidacy. While Obama's decision to waive his position on public financing drew a fair amount of criticism, the perception of a campaign funded by small donors rather than wealthy special interests has done a lot to make up for that. At this point, keeping the money flowing is a top priority.

supporters In addition to keeping a steady focus on the candidate himself, Obama's campaign blog often recounts the personal stories of his individual supporters. This is a way to keep that enthusiasm high, and make his grassroots backers feel they are doing something important.

John Schwenkler is a graduate student in philosophy at the University of California, Berkeley.

Apparently, both candidates mentioned Obama a lot

Word (Tag) Clouds

Visualize the frequency of words in a document

Word clouds are:

- social, friendly, approachable
- harmless when the goal is to have fun
- less harmless when used for communicating scientific or journalistic information

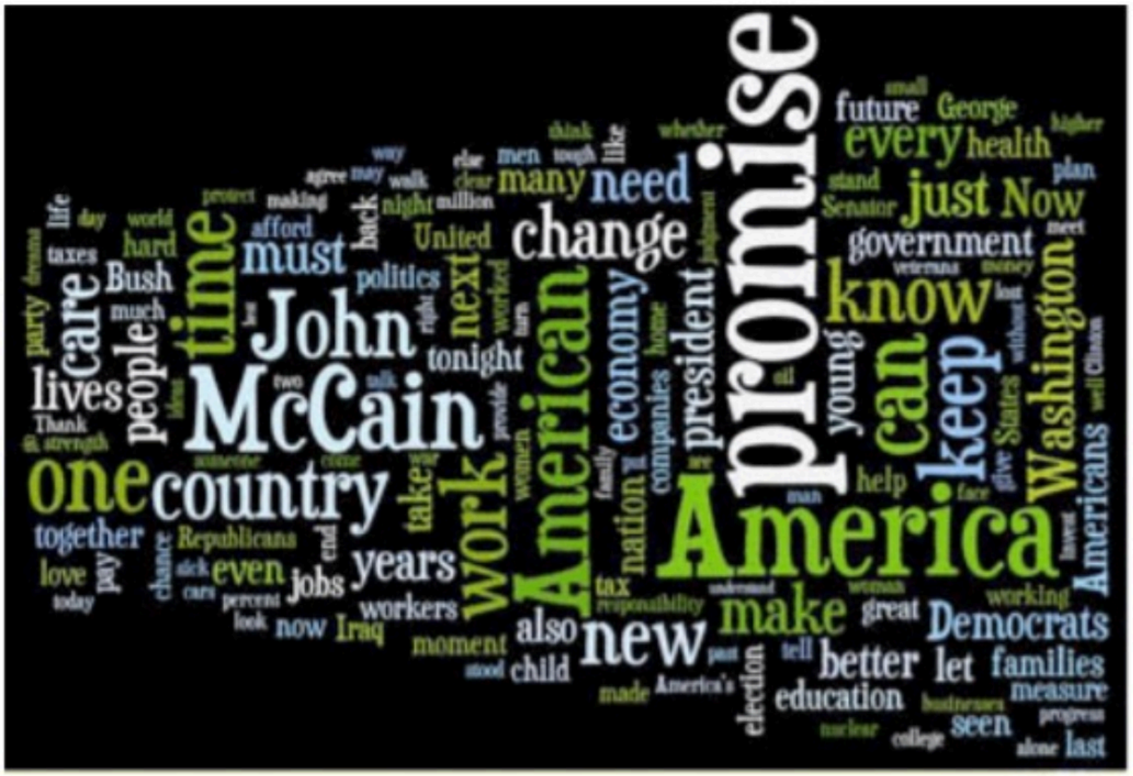
Word (Tag) Clouds: Issues

Word clouds:

- prioritize frequent words rather than important words
- prioritize words that have few alternatives
- ignores core words that are only explicitly mentioned a few times
- are perceptually difficult to parse
- confound the size channel with word length
- don’t actively support comparison

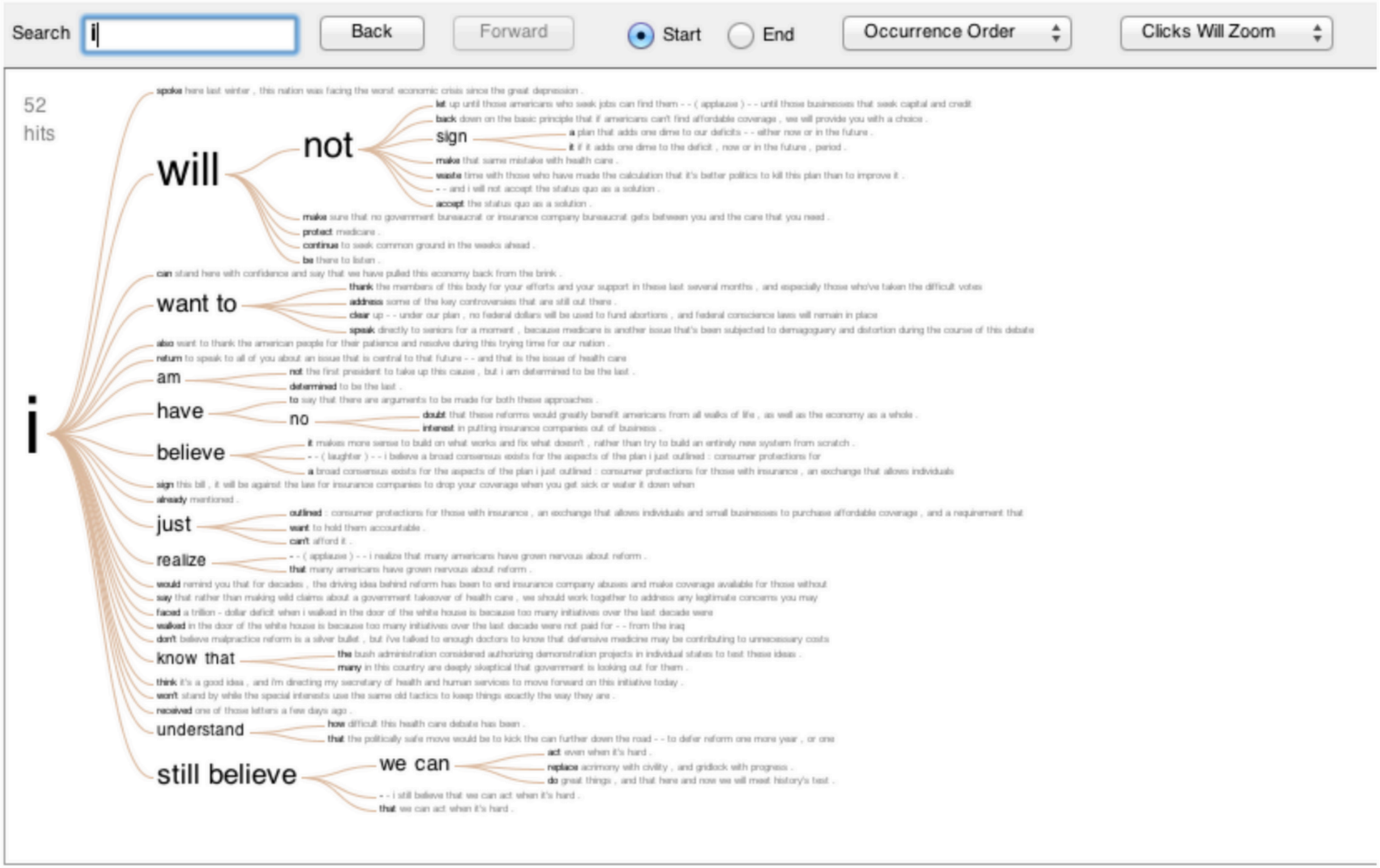
Some alternatives to word clouds

Sorted Word Clouds



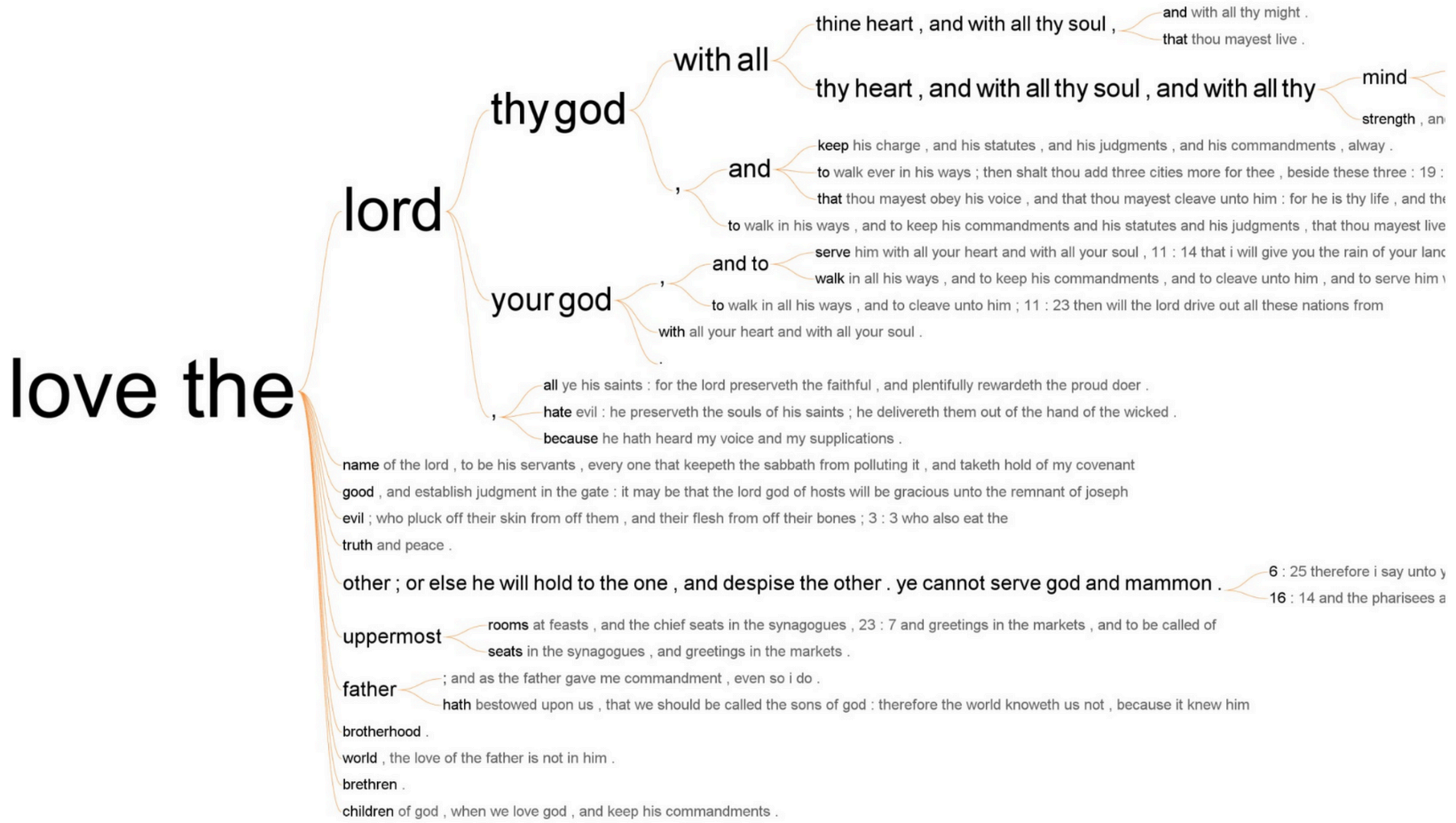
Sorted Word Clouds

WordTree



Obama’s Speech (2009)

WordTree



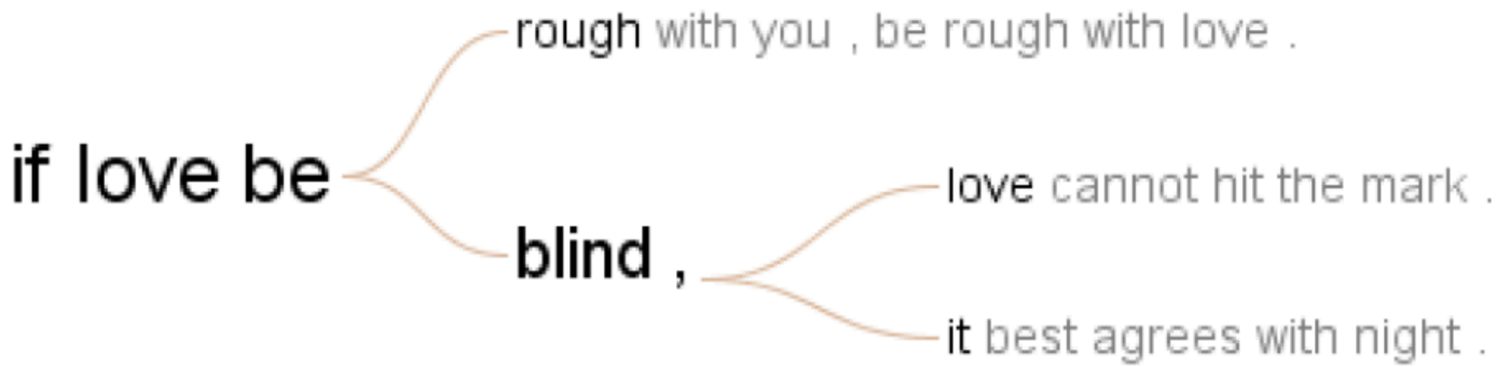
The Bible

WordTree

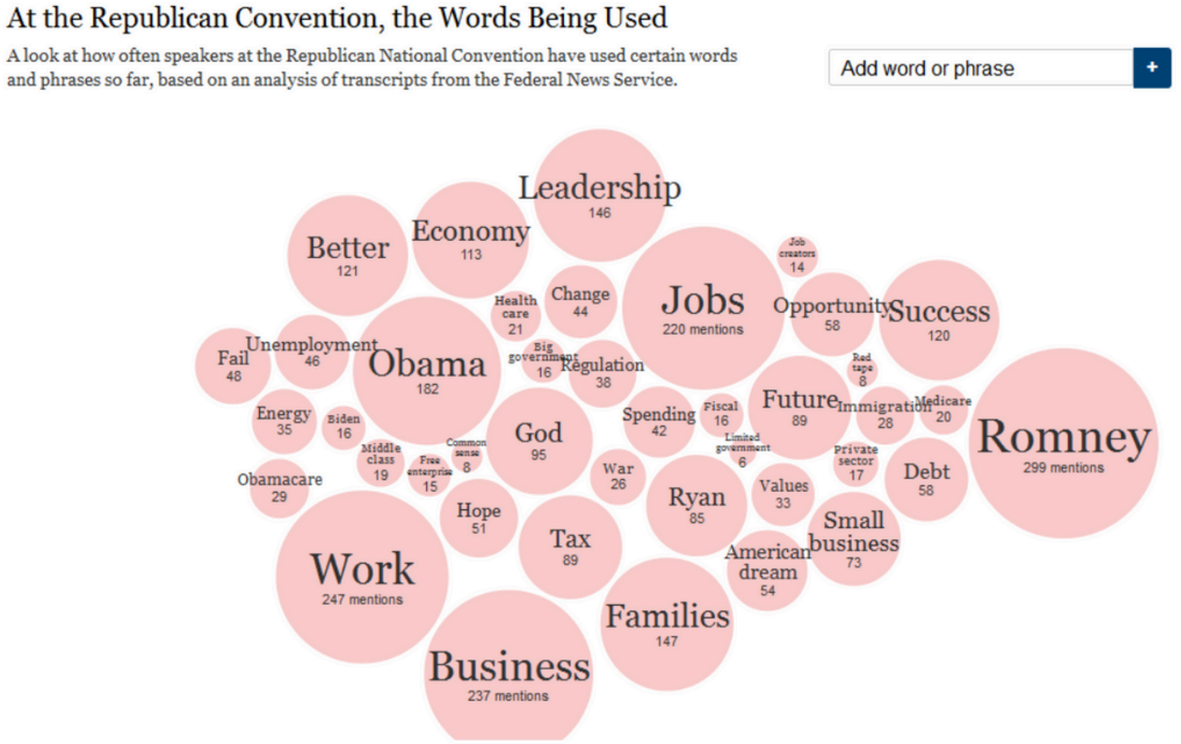
if love be rough with you , be rough with love .

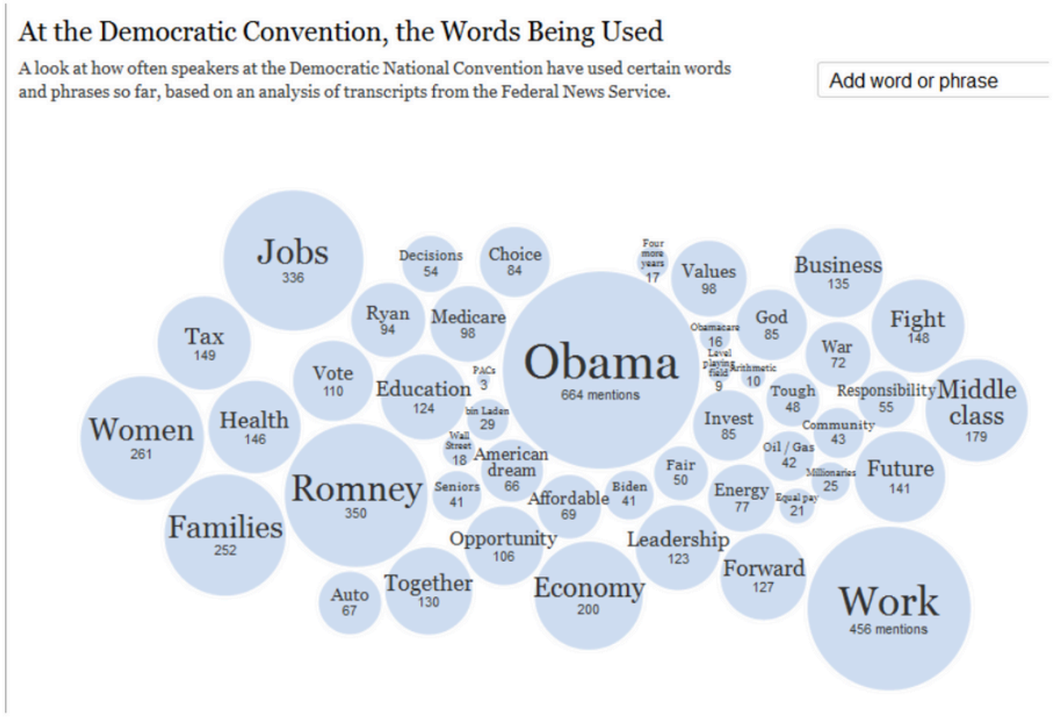
if love be blind , love cannot hit the mark .

if love be blind , it best agrees with night .

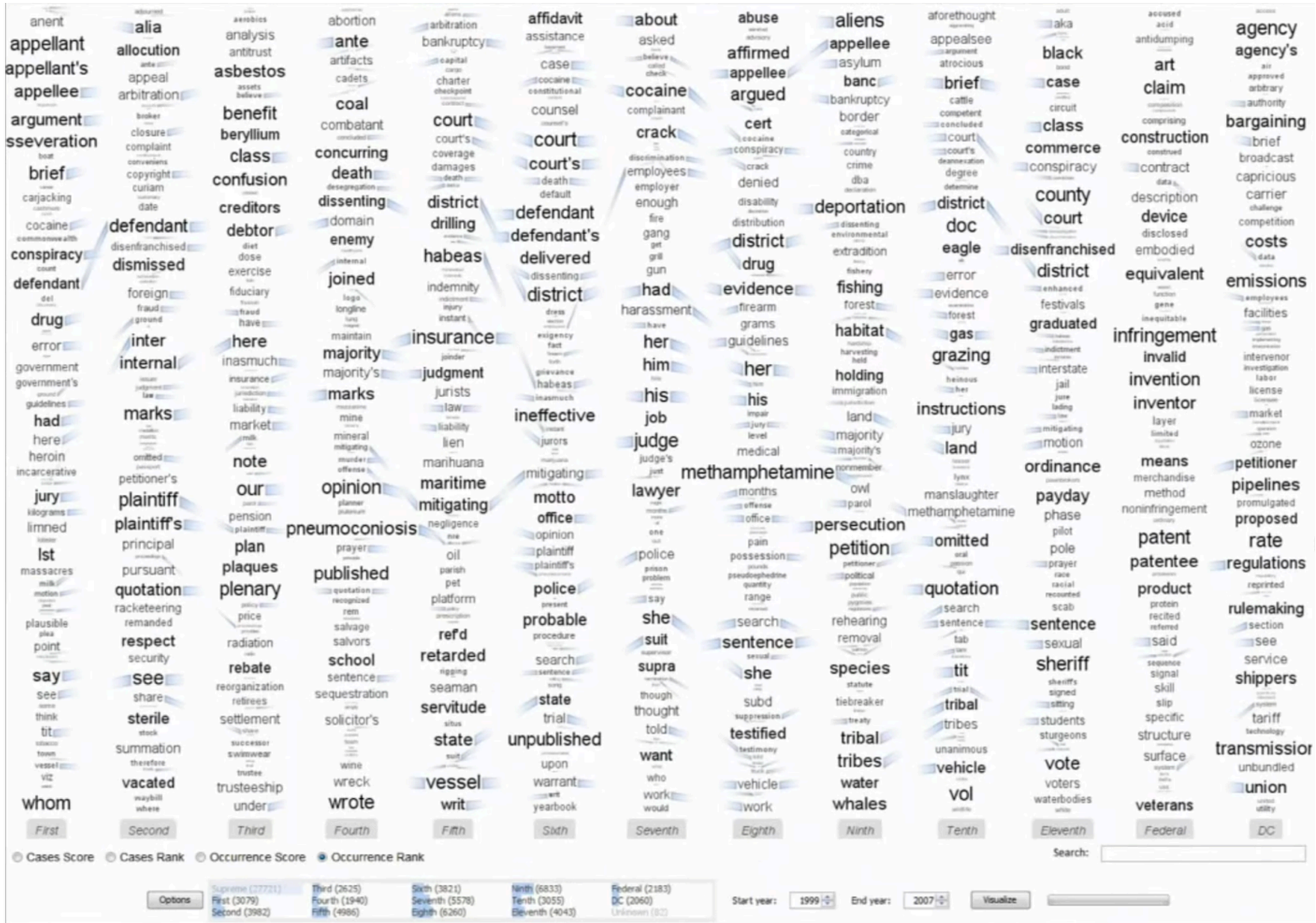


Bubble Charts

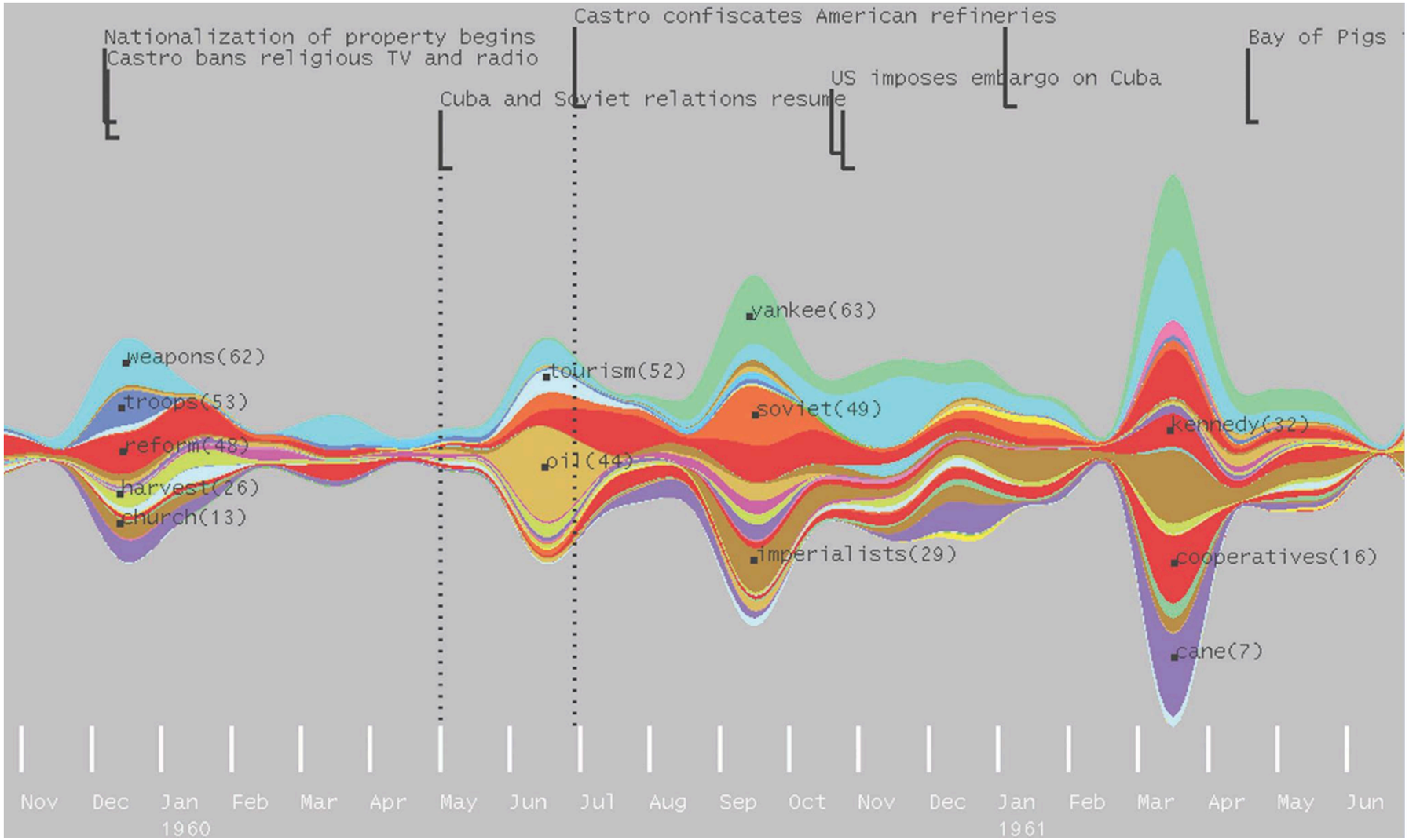




Parallel Tag Clouds



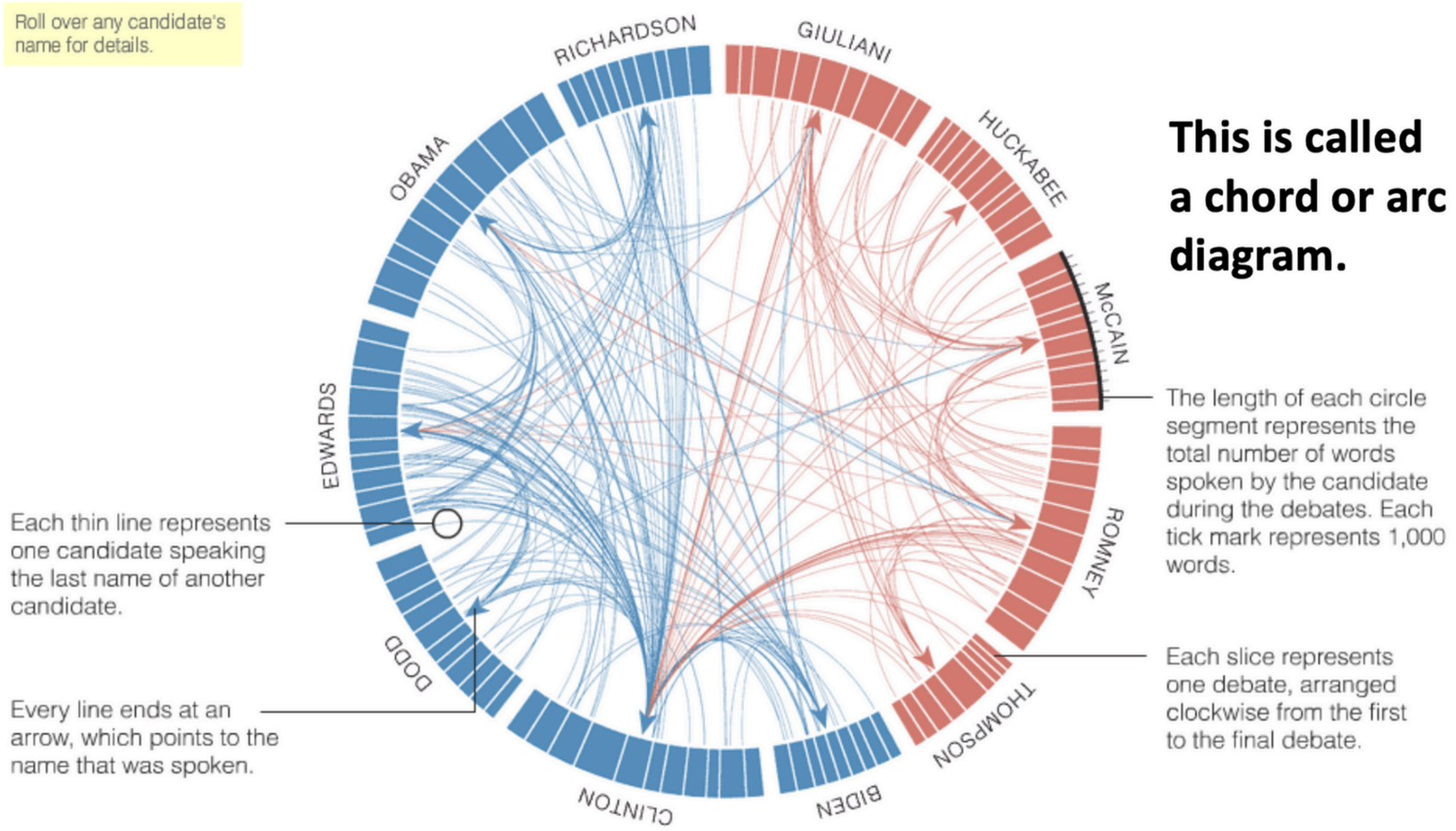
ThemeRiver


Nationalization of property begins
Castro bans religious TV and radio
Cuba and Soviet relations resume
Castro confiscates American refineries
US imposes embargo on Cuba
Bay of Pigs
weapons(62)
troops(53)
reform(48)
harvest(26)
church(13)
tourism(52)
oil(44)
yankee(63)
soviet(49)
imperialists(29)
kennedy(32)
cooperatives(16)
cane(7)
Nov
Dec
Jan
1960
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
1961
Feb
Mar
Apr
May
Jun

Chord/Arc Diagram

Naming Names

Names used by major presidential candidates in the series of Democratic and Republican debates leading up to the Iowa caucuses.



This is called a chord or arc diagram.

Source: Debate transcripts

Jonathan Corum and Farhana Hossain/The New York Times

Challenges in Text Visualization

Challenges in Text Visualization

1. Text is mostly unstructured (categorical, but atypical)

1. Text is Mostly Unstructured

Text is categorical, but not your typical categorical

Words can be categorically different, but:

- semantically similar (e.g., tennis, swimming, running)
- synonymous (e.g., easy, effortless, facile)
- ordered (e.g., January, February, etc.)
- and so on..

For example:

- Fall = Autumn (not the same word)
- Fall ?= Fall (not the same meaning)

Challenges in Text Visualization

1. Text is mostly unstructured (categorical, but atypical)
2. Textual Data can be very large and high-dimensional

2. Textual data is large and high-dimensional

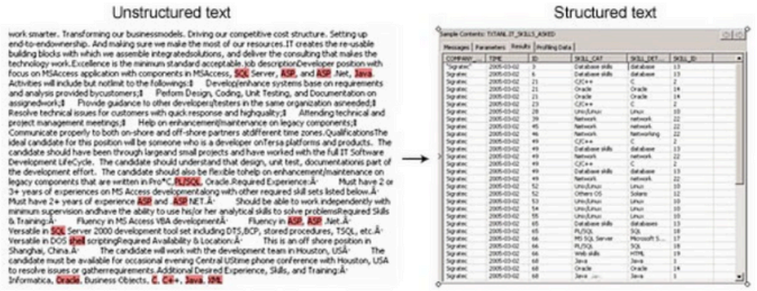
Unstructured Text

Extract Entities

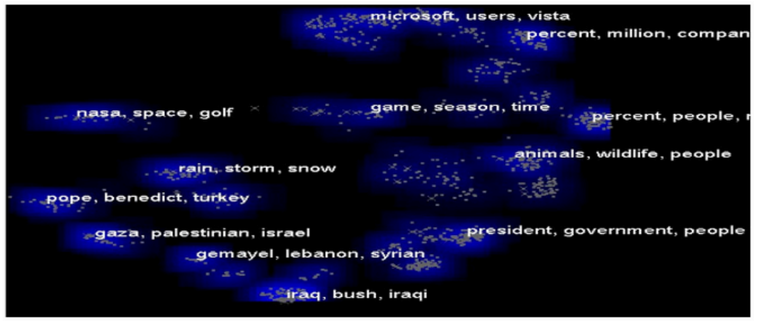
High Dimensional Data

Dimension Reduction Model

Choice of Visual Encodings/Metaphor



	BM1	BM2	BM3	BM4	BM5	BM6	BM7	SM1	SM2	SM3	SM4	SM5	SM6	BE	RM
BM1	—	78	62	76	76	68	71	65	79	62	54	65	74	47	55
BM2	0.75	—	54	70	64	63	67	60	61	68	60	60	62	40	37
BM3	0.55	0.51	—	72	81	78	62	83	76	68	57	68	67	58	57
BM4	0.66	0.65	0.63	—	83	78	73	77	75	73	71	84	77	58	57
BM5	0.67	0.57	0.70	0.68	—	94	85	84	87	78	74	83	79	68	63
BM6	0.58	0.57	0.67	0.65	0.76	—	74	85	81	77	68	79	83	67	64
BM7	0.63	0.63	0.55	0.63	0.71	0.63	—	68	63	60	68	74	67	46	53
SM1	0.57	0.56	0.73	0.66	0.70	0.72	0.59	—	92	78	72	75	73	67	55
SM2	0.68	0.55	0.65	0.63	0.70	0.67	0.53	0.78	—	78	73	79	74	67	63
SM3	0.57	0.66	0.62	0.65	0.67	0.68	0.54	0.70	0.68	—	83	80	72	52	52
SM4	0.50	0.59	0.52	0.63	0.64	0.60	0.62	0.65	0.64	0.78	—	74	81	51	53
SM5	0.58	0.56	0.59	0.71	0.68	0.66	0.64	0.65	0.66	0.71	0.66	—	88	61	63
SM6	0.65	0.58	0.59	0.66	0.65	0.70	0.58	0.63	0.62	0.65	0.73	0.76	—	64	61
BE	0.44	0.40	0.55	0.51	0.60	0.61	0.43	0.62	0.61	0.50	0.50	0.56	0.59	—	65
RM	0.51	0.37	0.53	0.54	0.55	0.57	0.49	0.50	0.56	0.50	0.51	0.57	0.56	0.64	—



The Text Processing Pipeline

Three Models:

- Bag of Words
- Keyword Weighting
- Term Commonness

Bag of Words

Text

Hogwarts School of Witchcraft and Wizardry, commonly shortened to Hogwarts, is a fictional British school of magic for students aged eleven to eighteen, and is the primary setting for the first six books in J. K. Rowling's Harry Potter series…

Albus Percival Wulfric Brian **Dumbledore** is a fictional character in J. K. Rowling's Harry Potter series. For most of the series, he is the headmaster of the wizarding school Hogwarts. As part of his backstory, it is revealed that he is the founder and leader of …

Collinwood Mansion is a fictional house featured in the Gothic horror soap opera Dark Shadows (1966–1971). Built in 1795 by Joshua Collins, Collinwood has been home to the Collins family—and other sometimes unwelcome supernatural visitors…

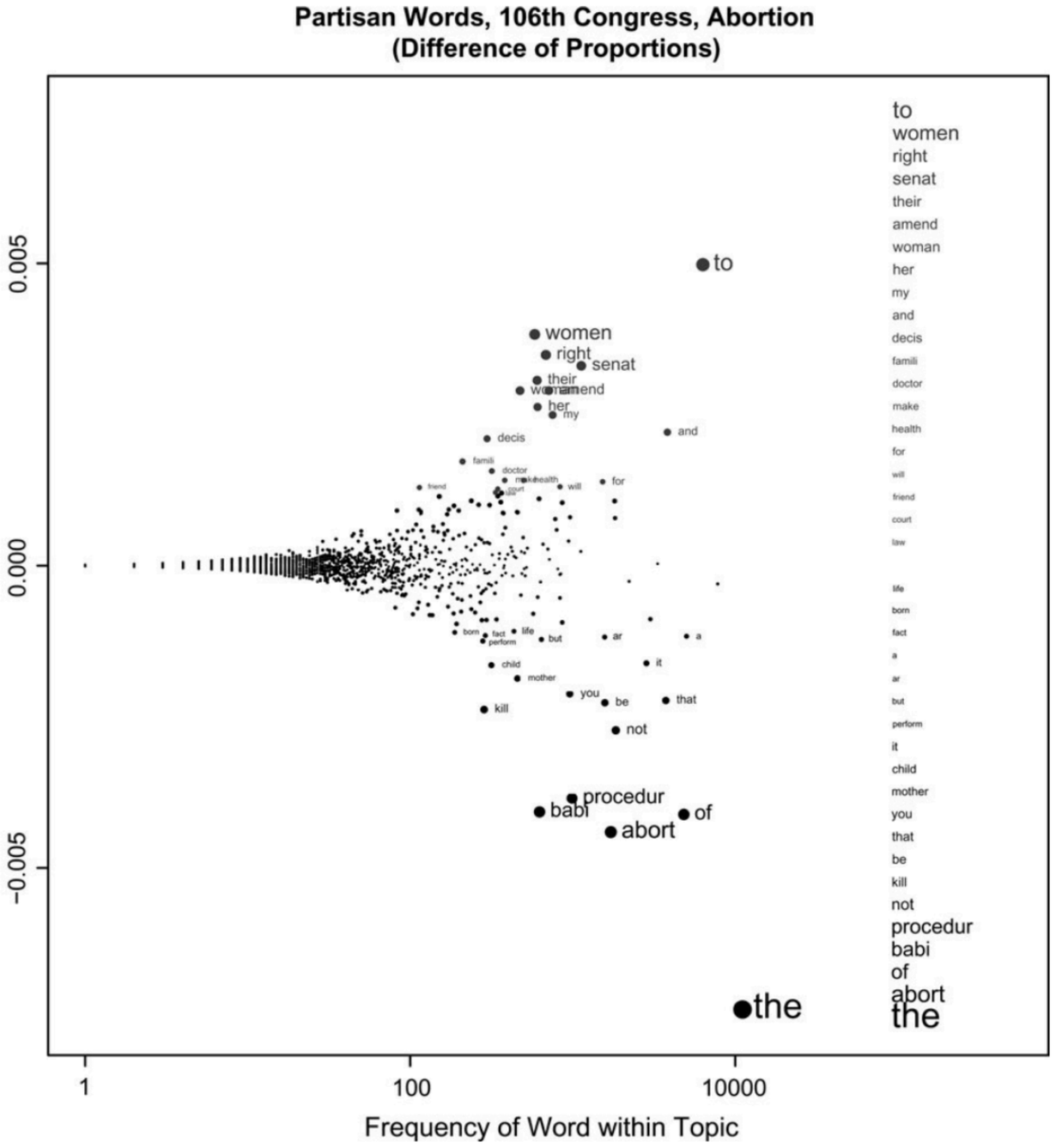
Data Table

	document		
term	Hogwarts (Top)	Dumbledore (Middle)	Collinwood (Bottom)
a	1	1	1
of	1	2	
in	1	1	2
is	2	4	1
fictional	1	1	1
school	1		
rowling's	1	1	
harry	1	1	
potter	1	1	
series	1	1	
house			1

Keyword Weighting

$$\text{Term frequency } (tf_{td}) = \text{number of times term } t \text{ occurs in document } d$$

$$\text{Normalized Term Frequency } (tf_{td_{norm}}) = \frac{tf_{td}}{\sum_t tf_{td}}$$



Democrats used more

Republicans used more

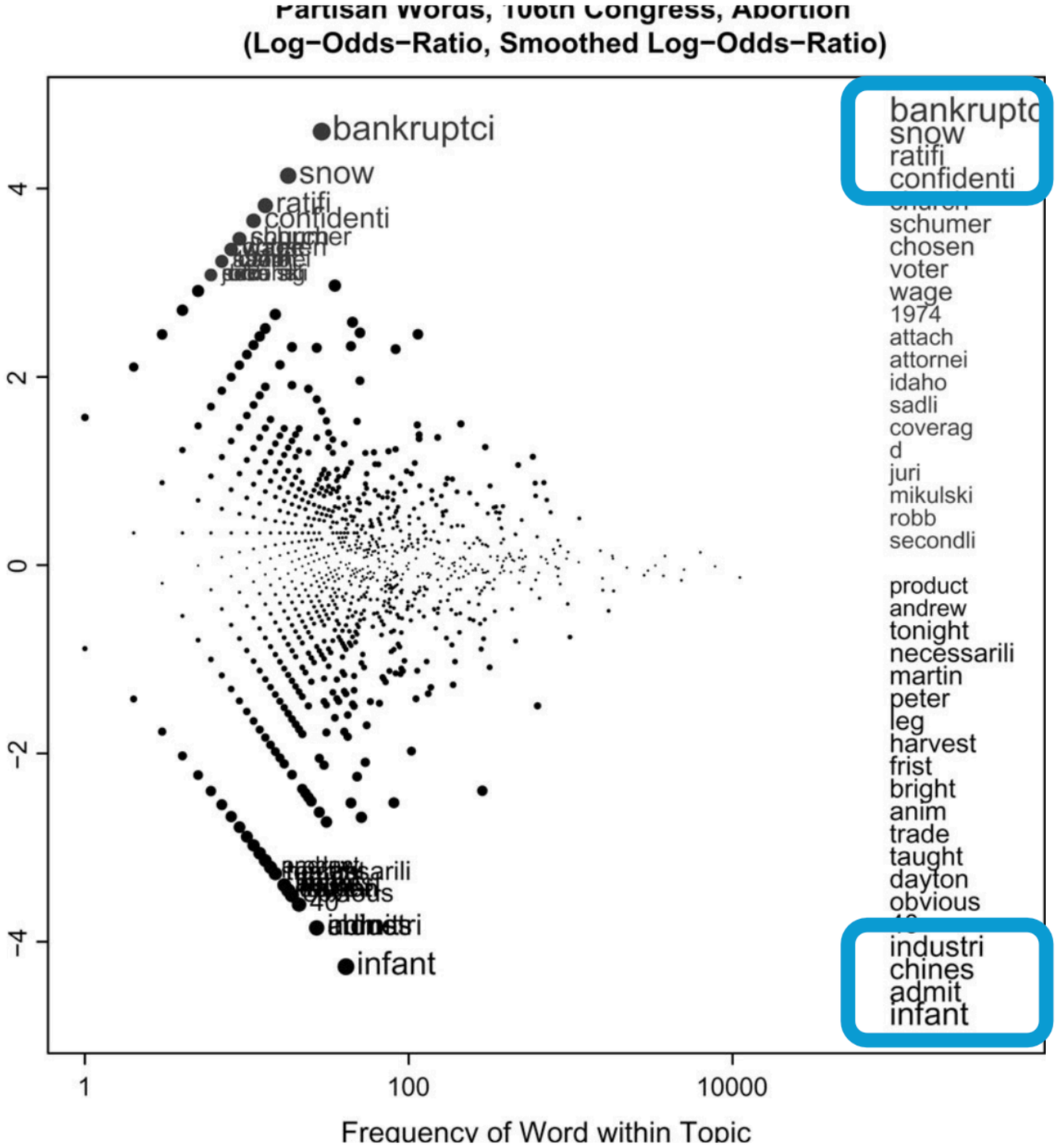
Keyword Weighting

Normalized Term Frequency considering the whole corpus ($tf_{td.wc}$)

$$= \log(1 + tf_{td}) \times \log(\frac{N}{d_t})$$

where d_t = number of documents containing term t

N = number of documents in the corpus



Democrats used more

Republicans used more

Term Commonness

Normalized Term Frequency considering the entire english language $(tf_{td.e})$

$$= \log(tf_{td}) \times \log(tf_{t.e})$$

The Text Processing Pipeline: Issues

All three models:

- rely only on term frequencies
- ignore contextual information like grammar, part of speech, negation, adjective-noun pairs, pronoun reference, etc.

Oblivion is a better game than Skyrim .
Is Oblivion a better game than Skyrim ?

Same frequencies, different meanings

Challenges in Text Visualization

1. Text is mostly unstructured (categorical, but atypical)
2. Textual Data can be very large and high-dimensional
3. Context is critical for understanding text
4. Summarizing, grouping, and parsing text is time-consuming

FIN

Upcoming Dates

May 9:**

- **Homework 5 Due**
- **Project Screencast Submission Due**

May 12: Final Project Submission Due

****Only if you requested an extension by May 2**