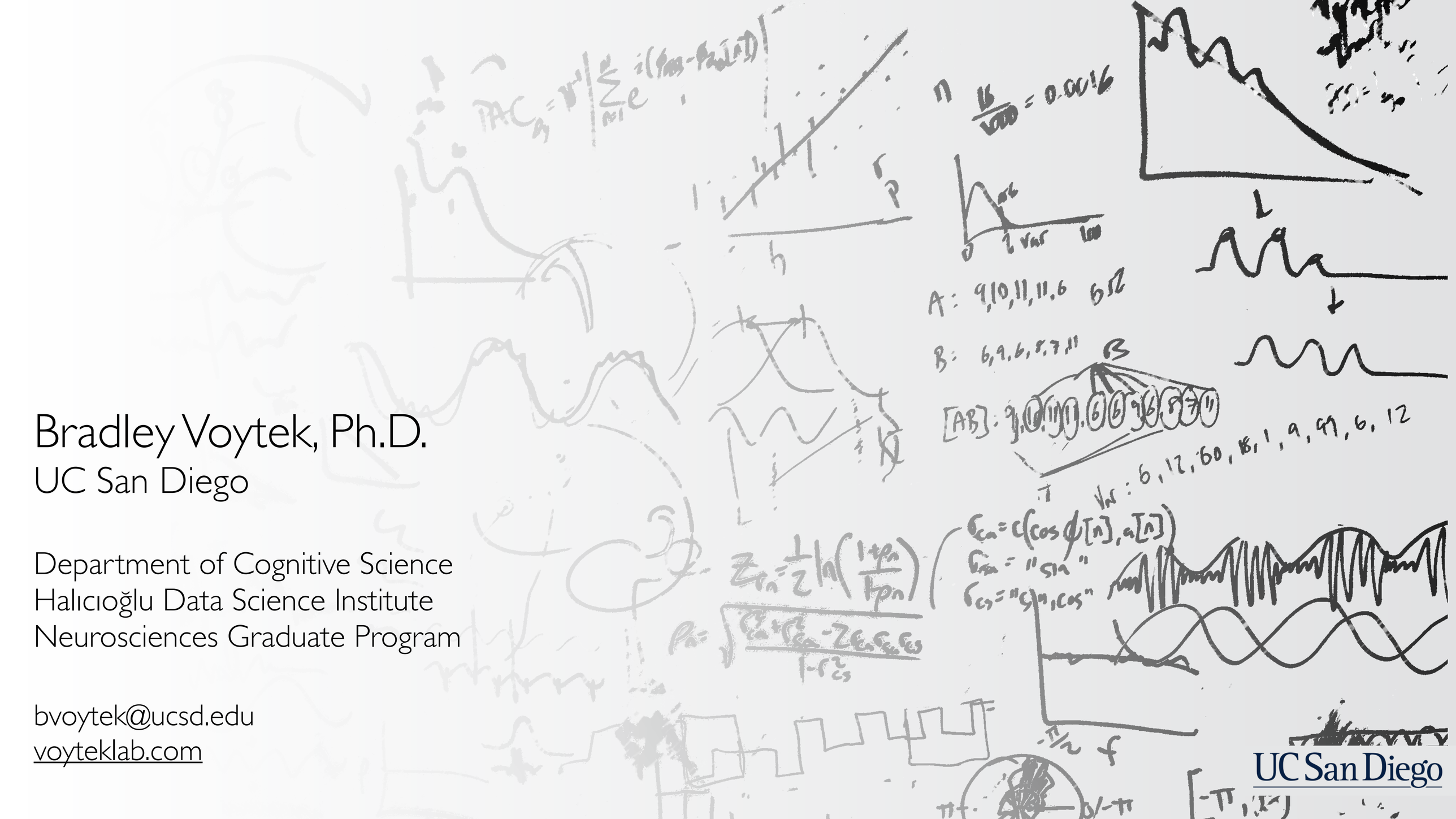
Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
voyteklab.com

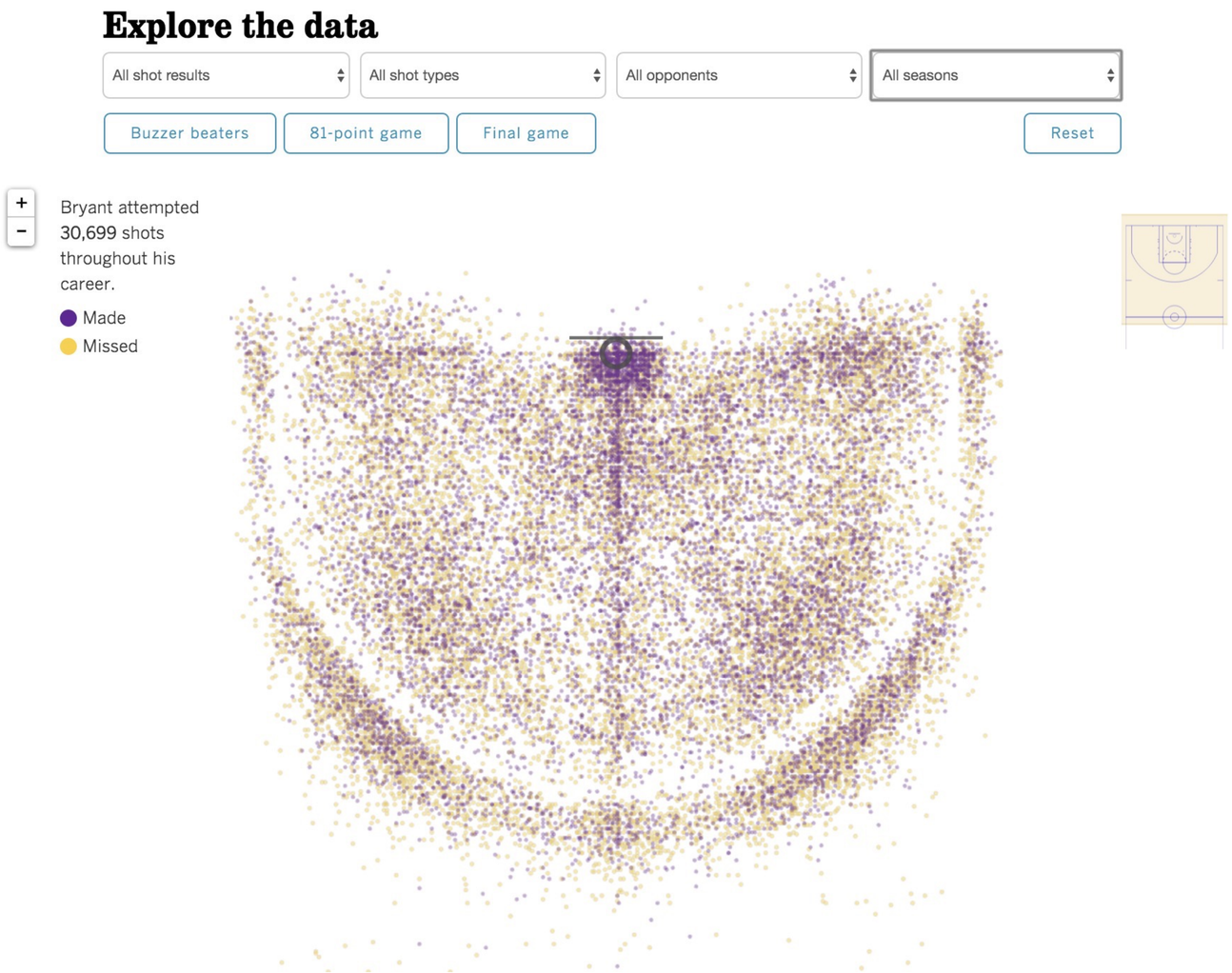


UC San Diego

Previous lecture

stray thoughts

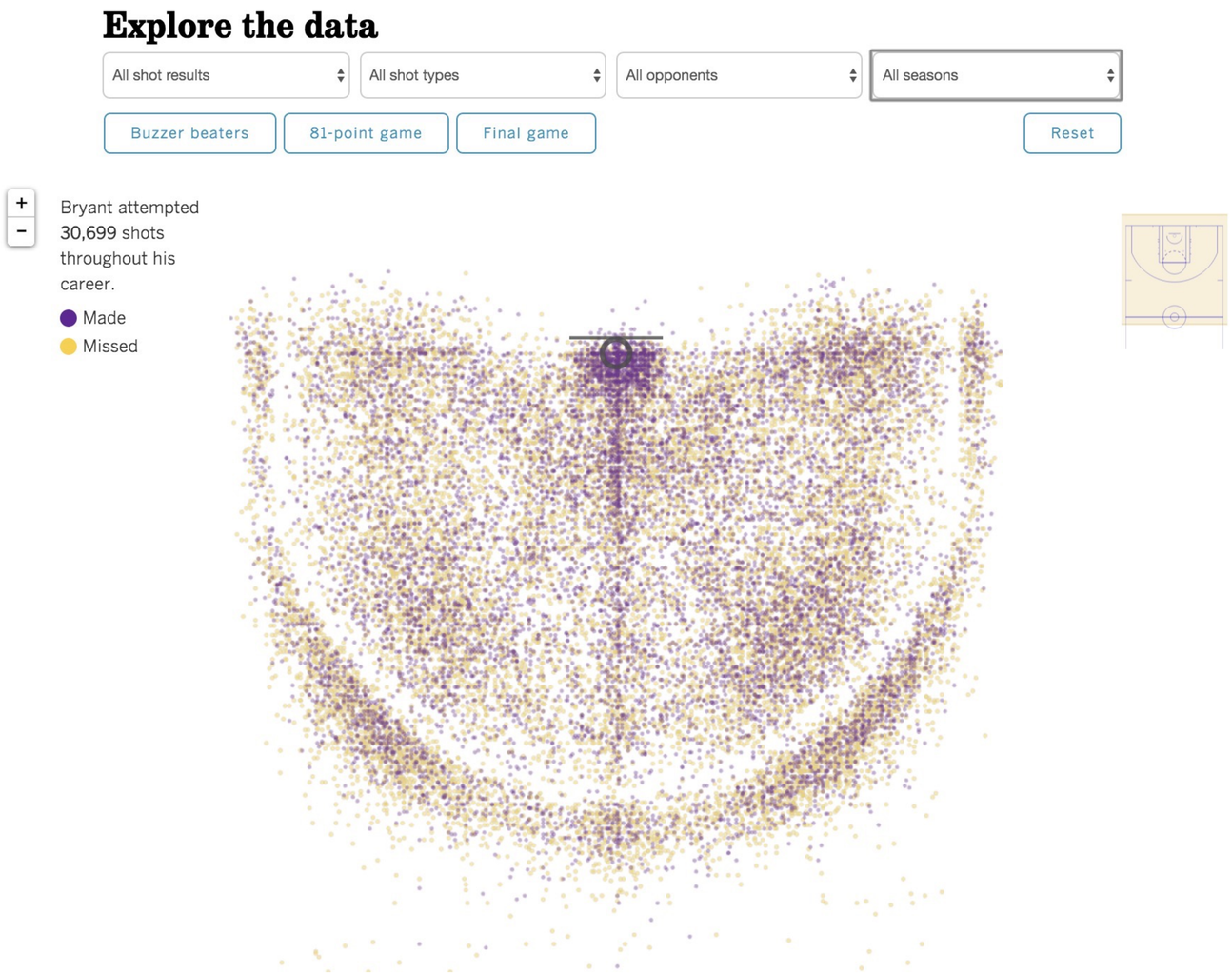
KOBE



source: http://graphics.latimes.com/kobe-every-shot-ever/

UC San Diego

KOBE



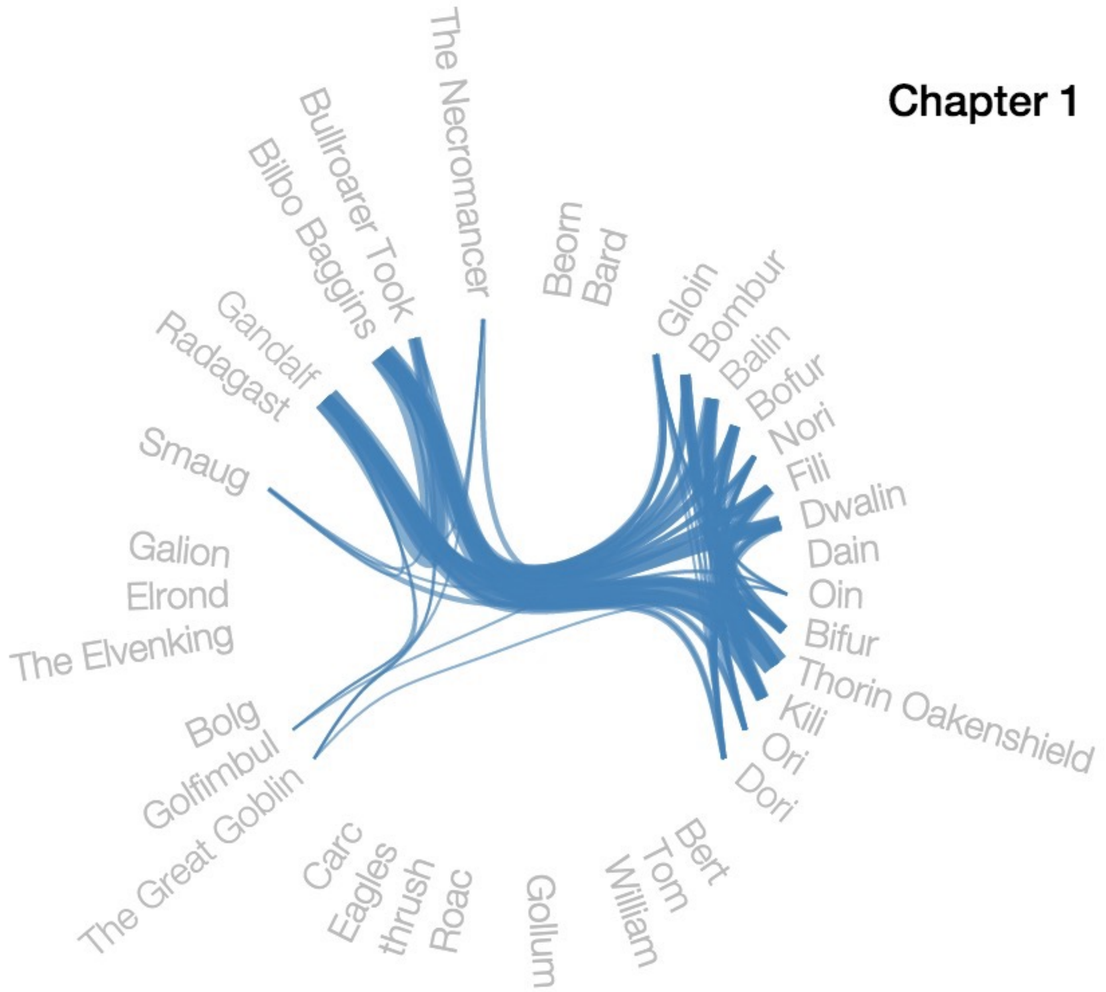
source: http://graphics.latimes.com/kobe-every-shot-ever/

UC San Diego

Visualizing *The Hobbit*

Hover over a character name in a graph.
Yellow links show connections in the selected chapter.
Yellow names show connections in the whole book.

The bars below show emotion intensity for each sentence.
Click on a character in the graph to see where they appear.
Hover over each bar to read the original sentence.



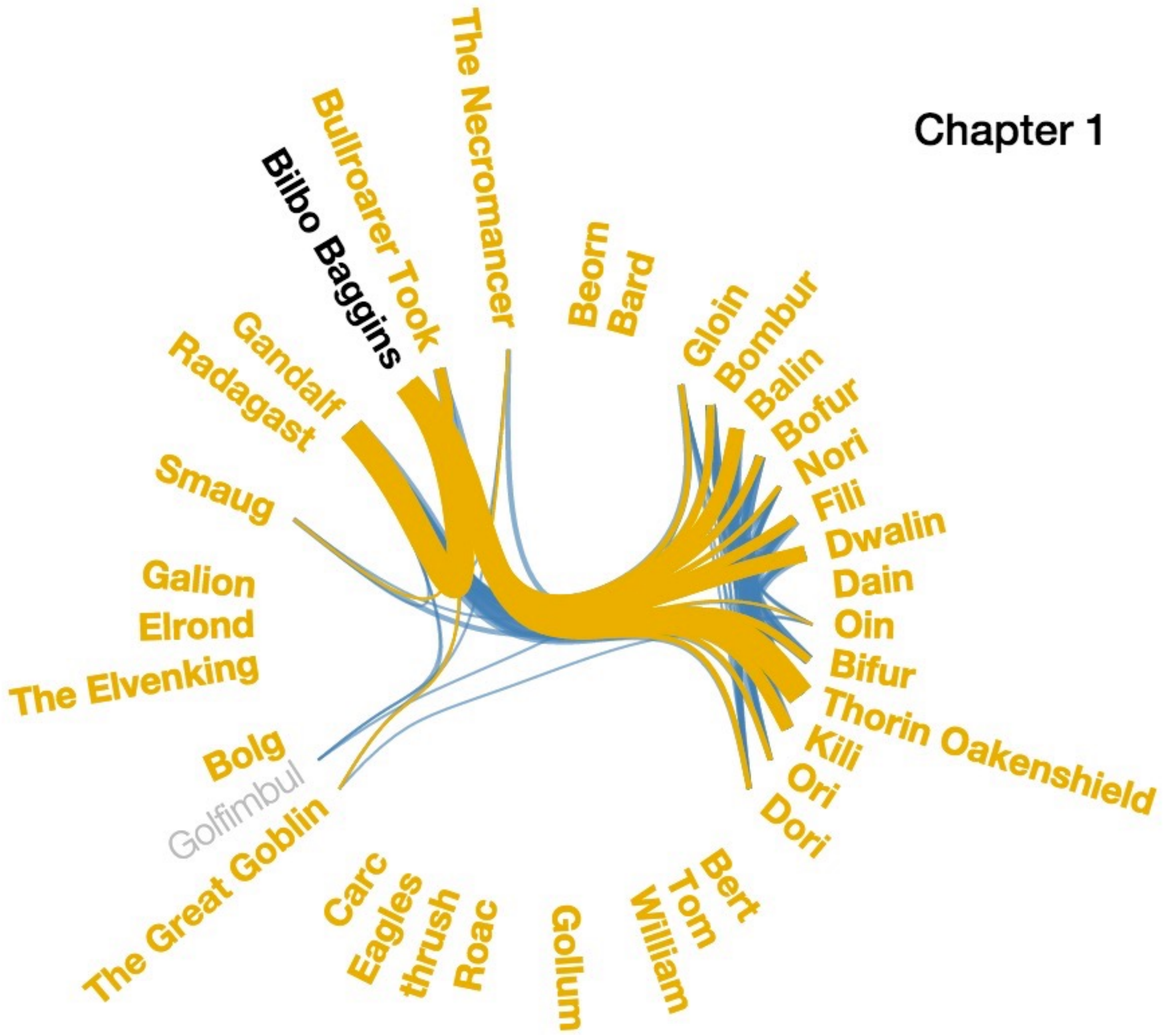
Source: Natalia Bilenko & Asako Miyakawa

UC San Diego

Visualizing *The Hobbit*

Hover over a character name in a graph.
Yellow links show connections in the selected chapter.
Yellow names show connections in the whole book.

The bars below show emotion intensity for each sentence.
Click on a character in the graph to see where they appear.
Hover over each bar to read the original sentence.



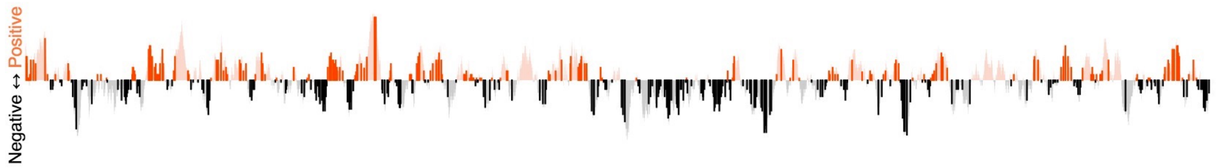
Source: Natalia Bilenko & Asako Miyakawa

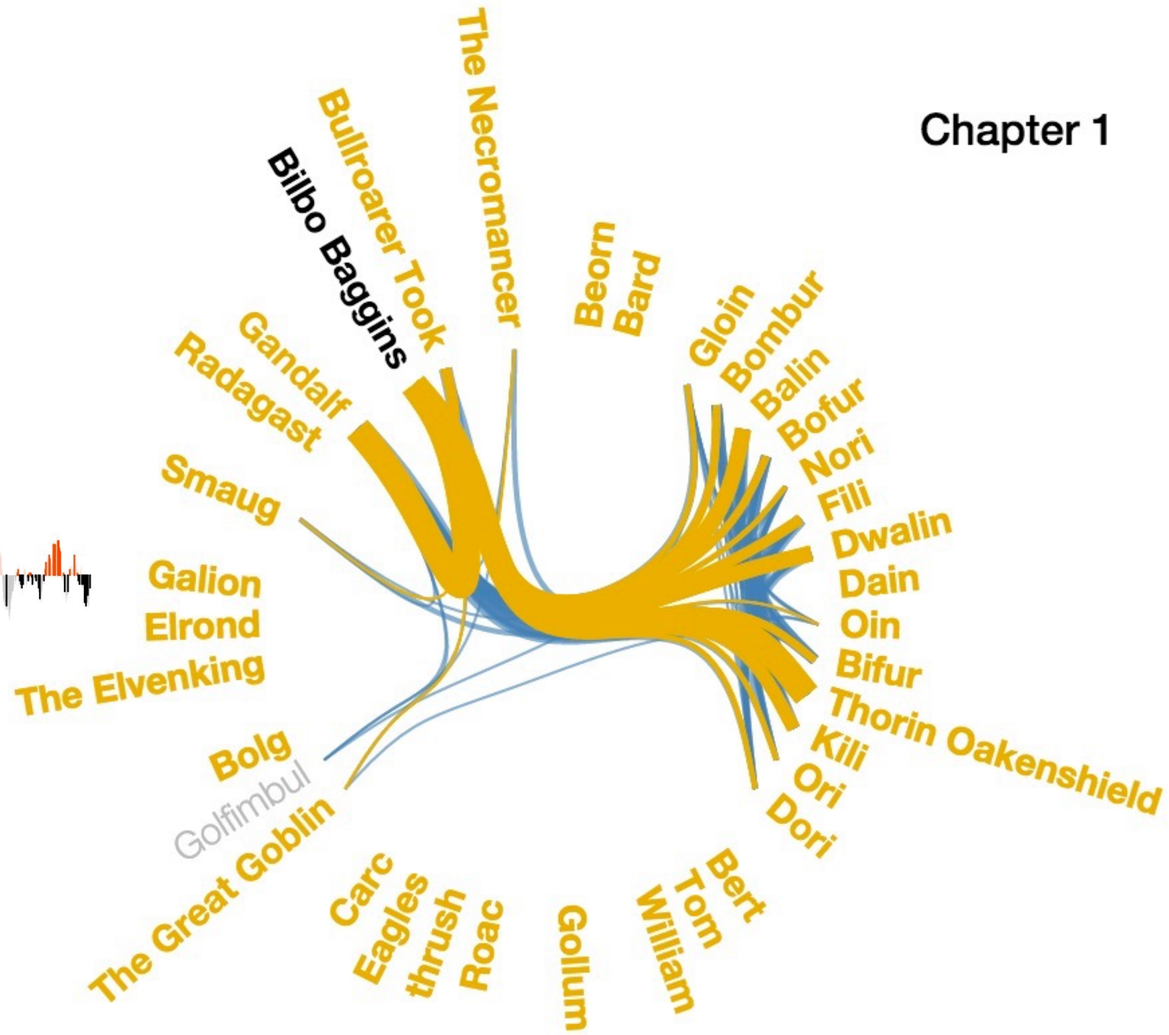
UC San Diego

Visualizing *The Hobbit*

Hover over a character name in a graph.
Yellow links show connections in the selected chapter.
Yellow names show connections in the whole book.

The bars below show emotion intensity for each sentence.
Click on a character in the graph to see where they appear.
Hover over each bar to read the original sentence.





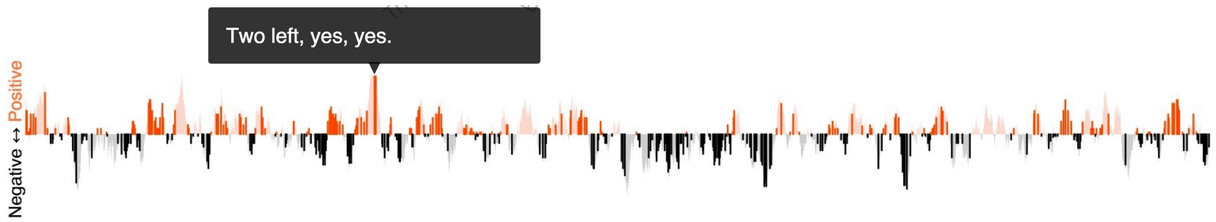
Source: Natalia Bilenko & Asako Miyakawa

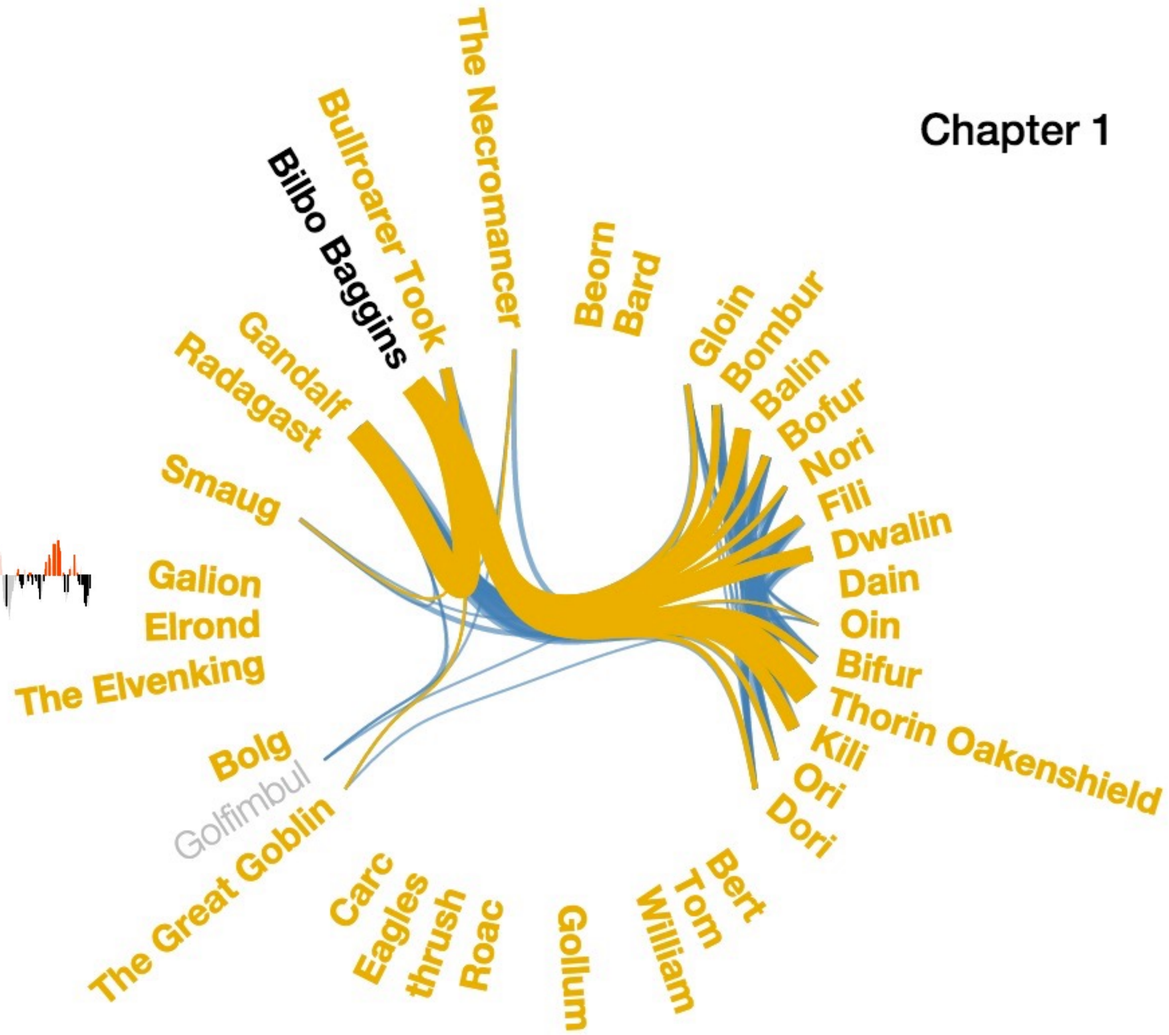
UC San Diego

Visualizing *The Hobbit*

Hover over a character name in a graph.
Yellow links show connections in the selected chapter.
Yellow names show connections in the whole book.

The bars below show emotion intensity for each sentence.
Click on a character in the graph to see where they appear.
Hover over each bar to read the original sentence.





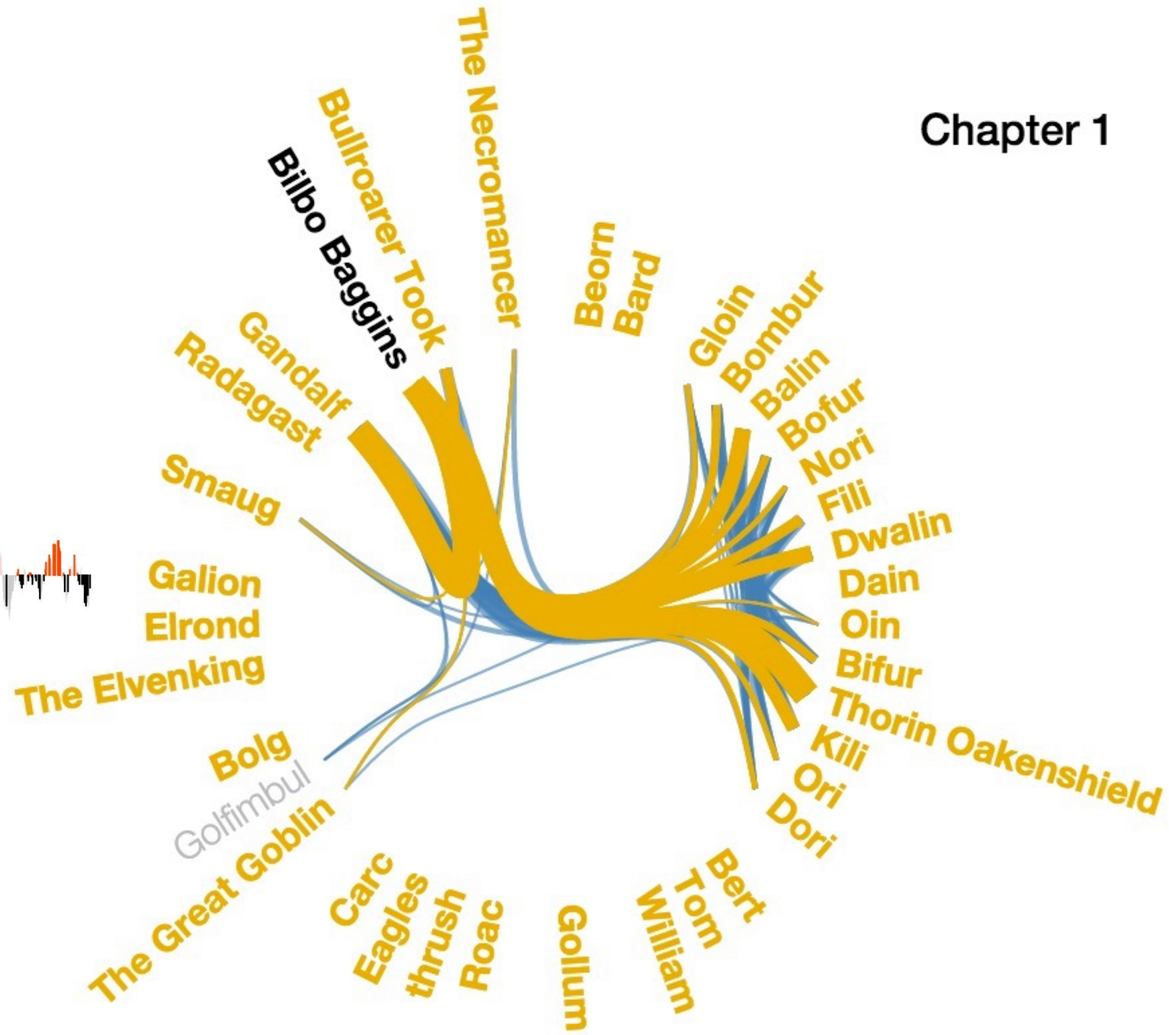
Source: Natalia Bilenko & Asako Miyakawa

UC San Diego

Visualizing *The Hobbit*

Hover over a character name in a graph.
Yellow links show connections in the selected chapter.
Yellow names show connections in the whole book.

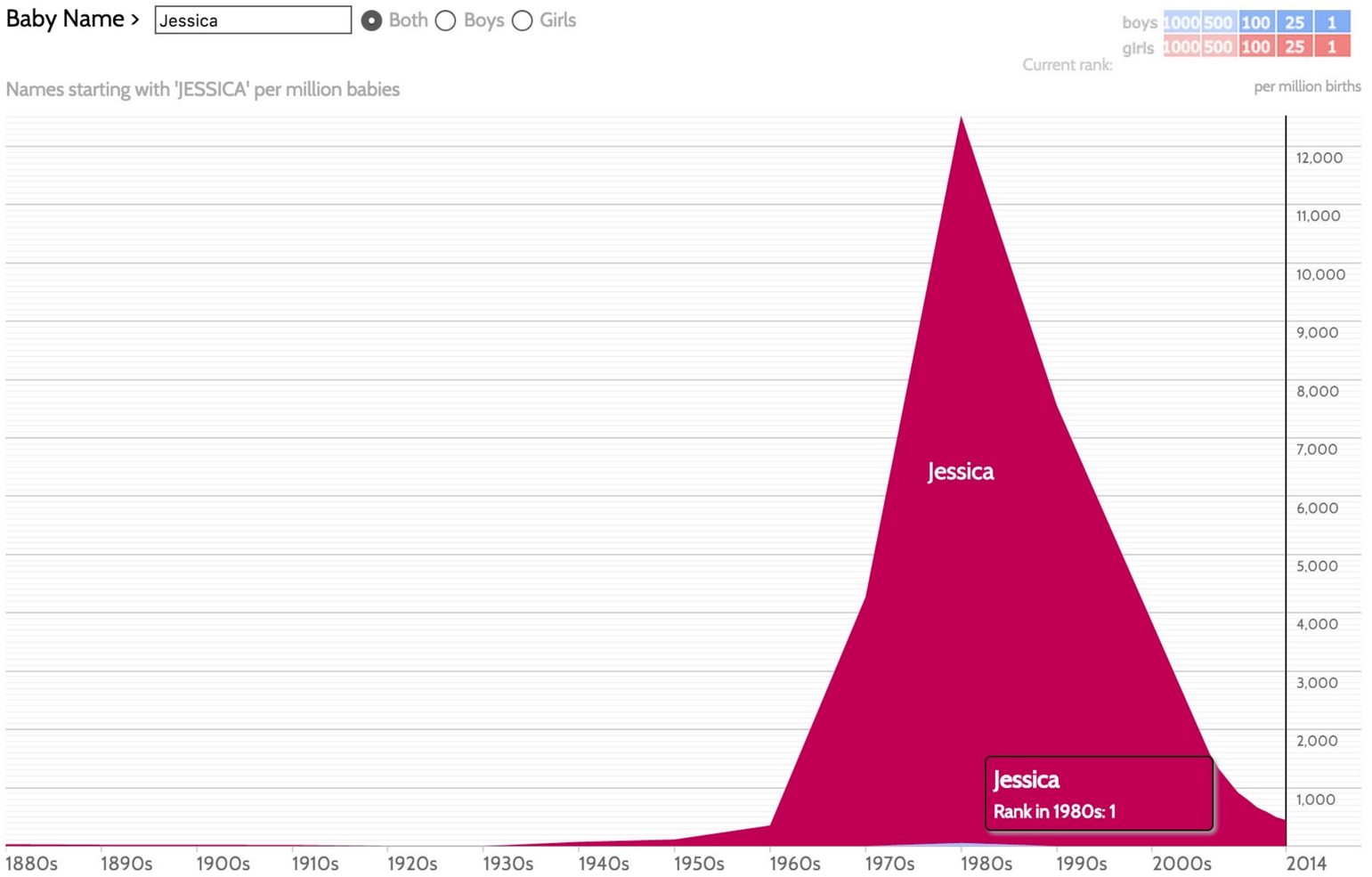
The bars below show emotion intensity for each sentence.
Click on a character in the graph to see where they appear.
Hover over each bar to read the original sentence.



Source: Natalia Bilenko & Asako Miyakawa

UC San Diego

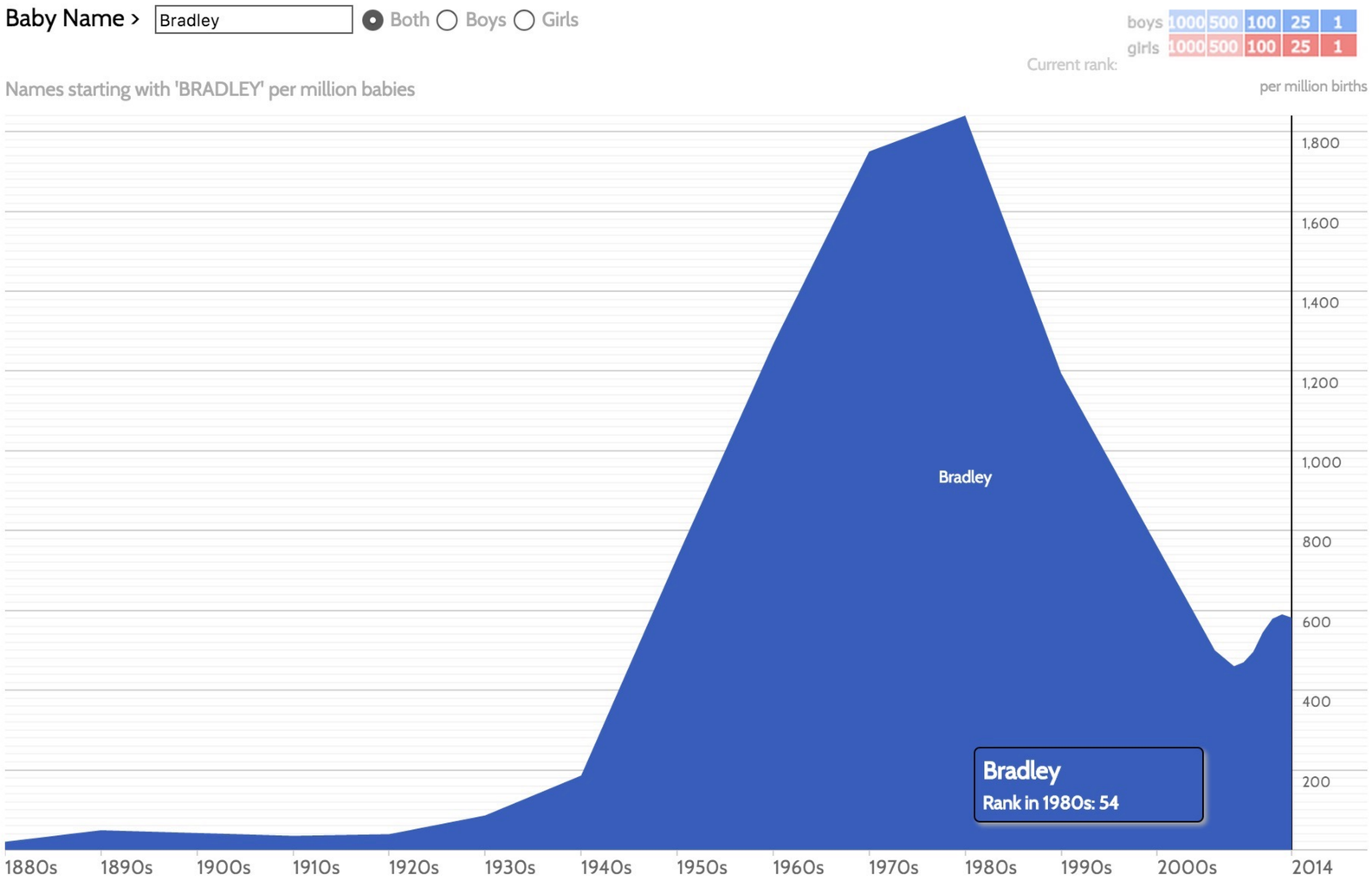
Names over time



Source: babynamewizard.com

UC San Diego

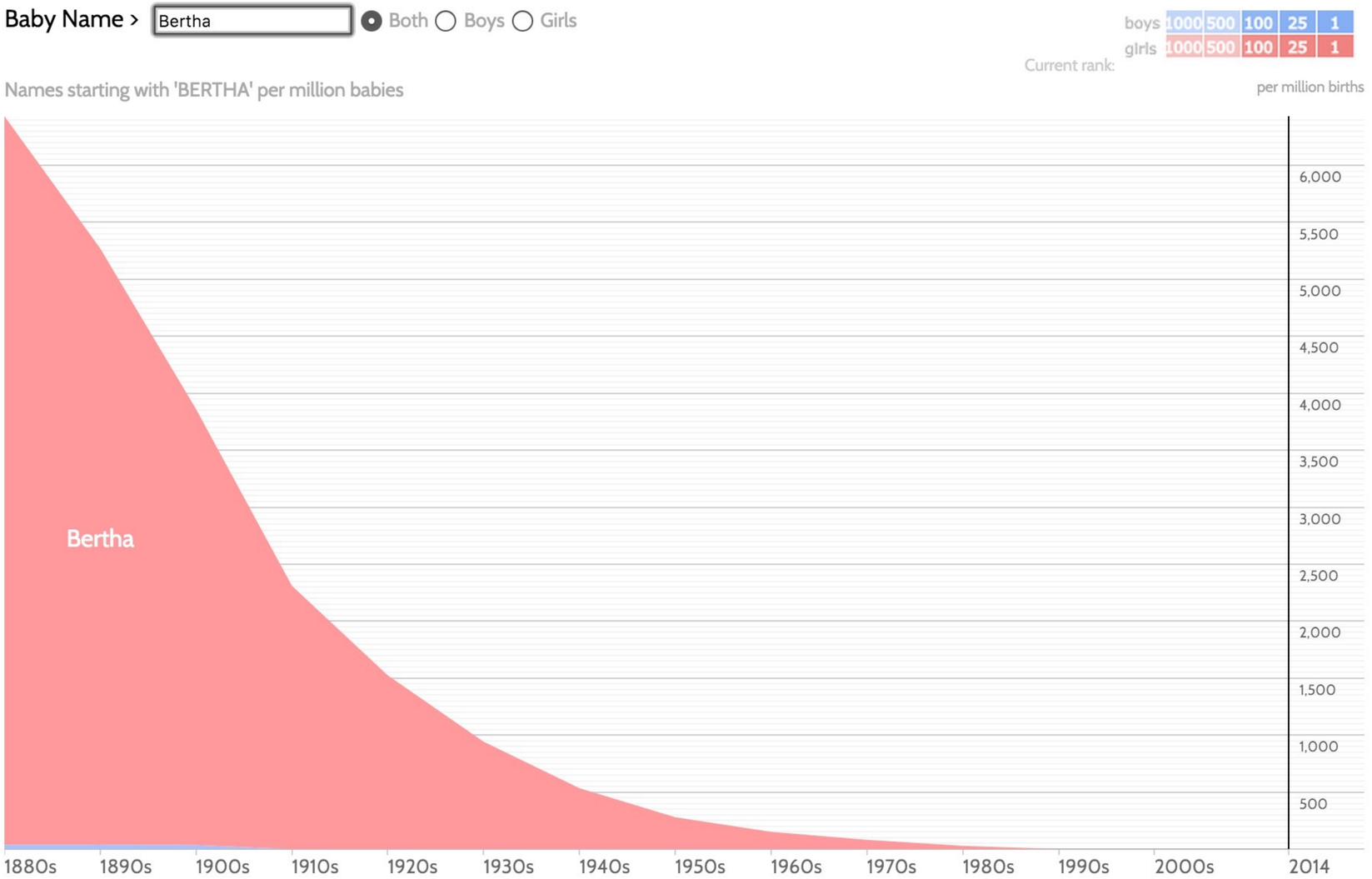
Names over time



Source: babynamewizard.com

UC San Diego

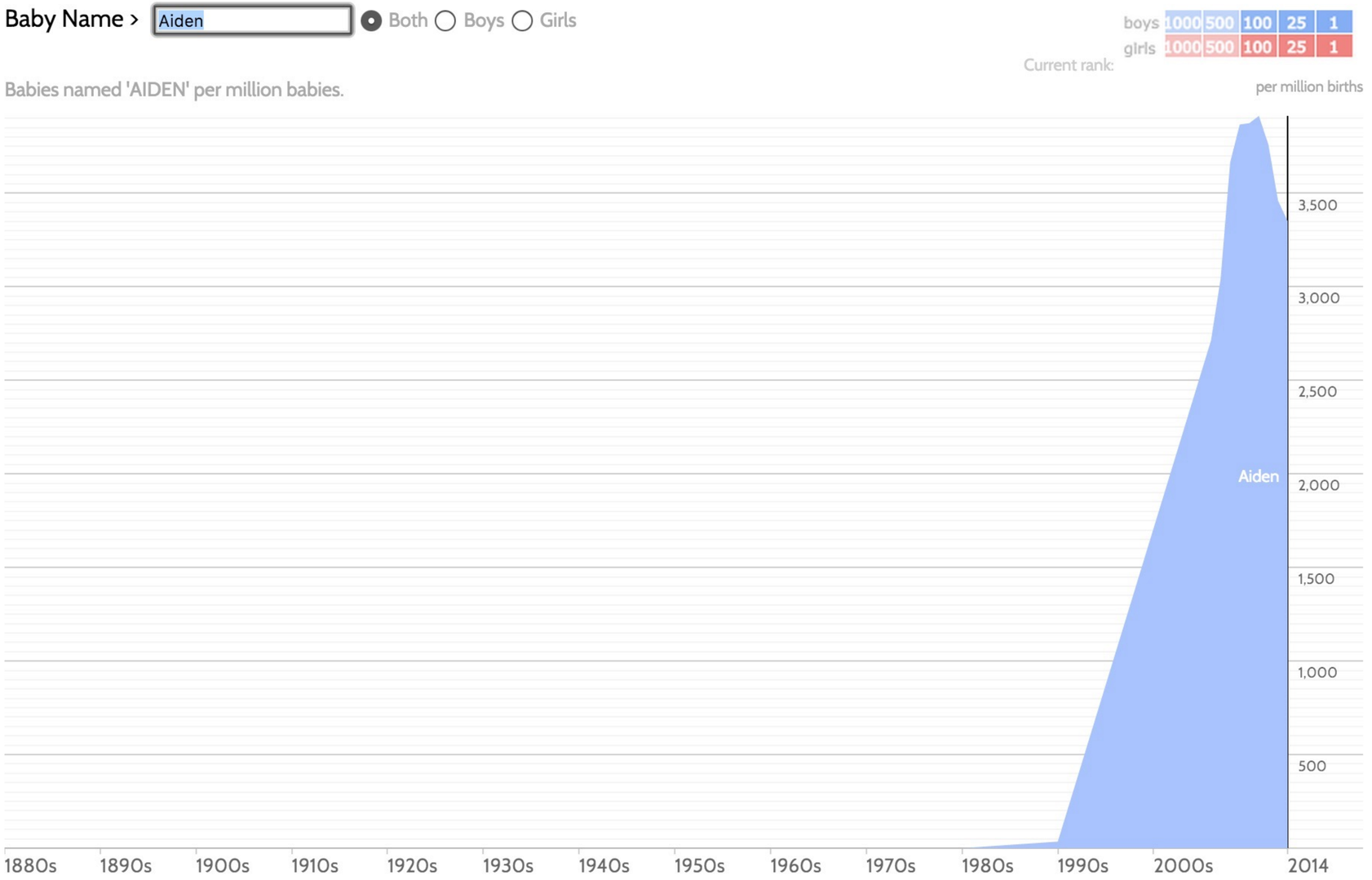
Names over time



Source: babynamewizard.com

UC San Diego

Names over time



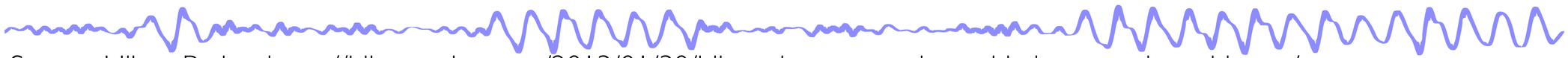
Source: babynamewizard.com

UC San Diego

Names

Looking at the absolute chance in percentages is interesting, but would not tell the full story. A change of, say 15% to 14% would be quite different and less drastic than a change from 2% to 1%, but the absolute change in percentage would measure those two things equally. Thus, I need a measure of the relative change in the percentages — that is, the percent change in percentages (confusing, I know). Fortunately the public health field has dealt with this problem for a long time, and has a measurement called the relative risk, where "risk" refers to the proportion of babies given a certain name. For example, let's say the percentage of babies named "Jane" is 1% of the population in 1990, and 1.2% of the population in 1991. The relative risk of being named "Jane" in 1991 versus 1990 is 1.2 (that is, it's (1.2/1)=1.2 times as probable, or (1.2-1)*100=20% more likely). In this case, however, I'm interested in instances where the percentage of children with a certain name decreases. The way to make the most sensible statistics in this case is to calculate the relative risk again, but in this case think of it as a decrease. That is, if "Jane" was at 1.5% in 1990 and 1.3% in 1991, then the relative risk of being named "Jane" in 1991 compared to 1990 is (1.3/1.5)=0.87. That is, it is (1-0.87)*100=13% less likely that a baby will be named "Jane" in 1991 compared to 1990.

(Note that I'm not doing any model fitting here because I'm not interested in any parameter estimates — I have my entire population! I'm just summarizing the data in a way that makes sense.)

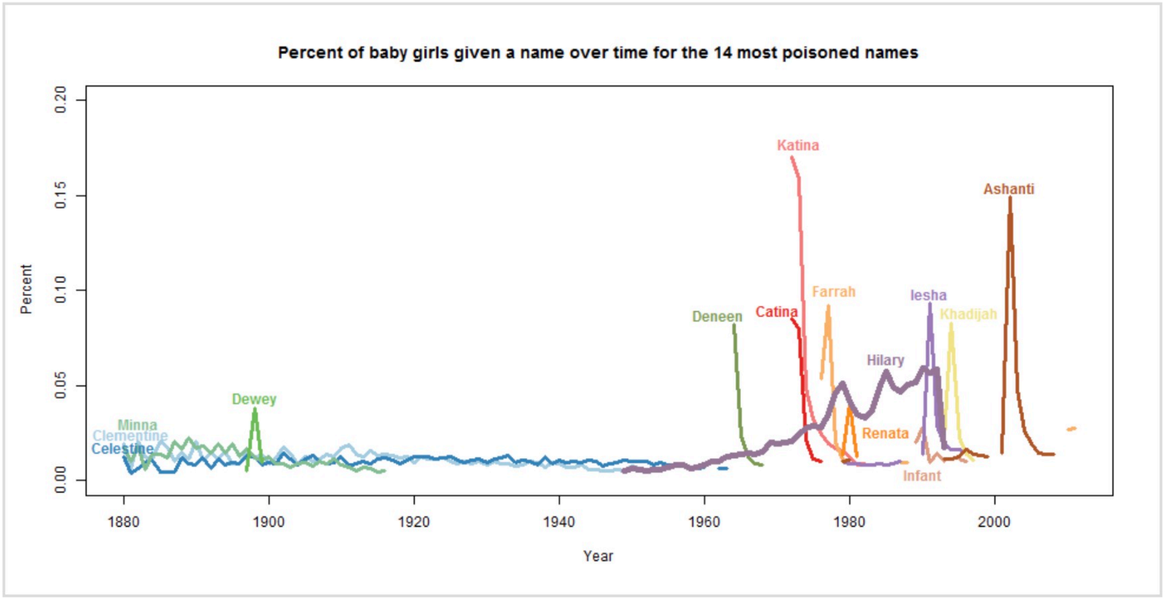


Source: Hilary Parker, https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/

Names

Looking at the absolute chance in percentages is interesting, but would not tell the full story. A change of, say 15% to 14% would be quite different and less drastic than a change from 2% to 1%, but the absolute change in percentage would measure those two things equally. Thus, I need a measure of the relative change in the percentages — that is, the percent change in percentages (confusing, I know). Fortunately the public health field has dealt with this problem for a long time, and has a measurement called the relative risk, where "risk" refers to the proportion of babies given a certain name. For example, let's say the percentage of babies named "Jane" is 1% of the population in 1990, and 1.2% of the population in 1991. The relative risk of being named "Jane" in 1991 versus 1990 is 1.2 (that is, it's (1.2/1)=1.2 times as probable, or (1.2-1)*100=20% more likely). In this case, however, I'm interested in instances where the percentage of children with a certain name decreases. The way to make the most sensible statistics in this case is to calculate the relative risk again, but in this case think of it as a decrease. That is, if "Jane" was at 1.5% in 1990 and 1.3% in 1991, then the relative risk of being named "Jane" in 1991 compared to 1990 is (1.3/1.5)=0.87. That is, it is (1-0.87)*100=13% less likely that a baby will be named "Jane" in 1991 compared to 1990.

(Note that I'm not doing any model fitting here because I'm not interested in any parameter estimates — I have my entire population! I'm just summarizing the data in a way that makes sense.)



These plots looked quite curious to me. While the names had very steep drop-offs, they also had very steep drop-ins as well.

This is where this project got deliriously fun. For each of the names that "dropped in" I did a little research on the name and the year. "Dewey" popped up in 1898 because of the Spanish-American War — people named their daughters after George Dewey. "Deneen" was one name of a duo with a one-hit wonder in 1968. "Katina" and "Catina" were wildly popular because in 1972 in the soap opera *Where the Heart Is* a character is born named

Source: Hilary Parker, https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/

UC San Diego

Names

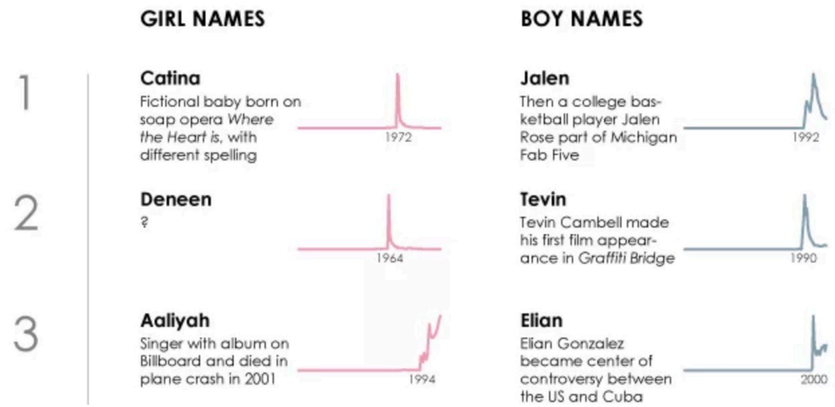
The Most Trendy Names in US History

POSTED TO DATA UNDERLOAD | TAGS: NAMES | NATHAN YAU

Names are incredibly personal things. It's how we identity ourselves. We associate others, places, and points in our past with names. Maybe you recall a family member, a celebrity, or a significant other.

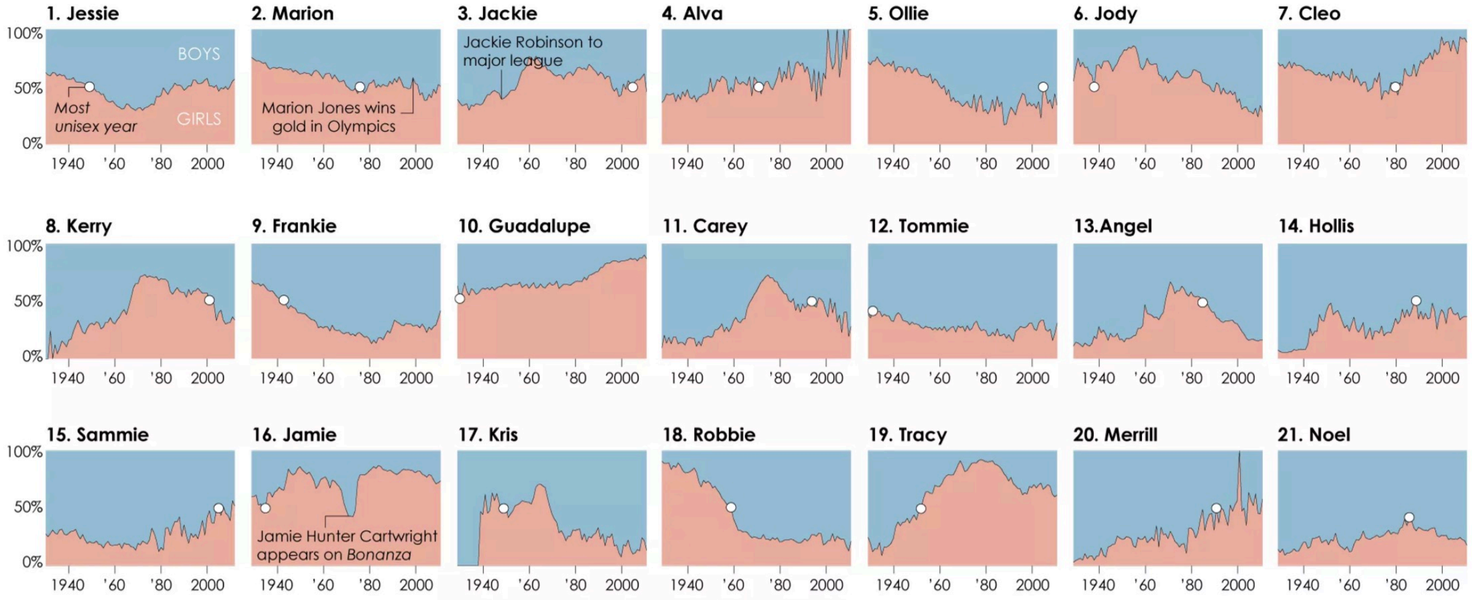
At the same time, it's not uncommon for two people with the same name to run into each other, and it's why gift shops can sell and profit from those mini license plates. Parents decide what they want to call their kid at some point. So as you walk through history, you end up with names that surge, some that die off, and some that come back again.

Hilary Parker already looked at the most poisoned name in US history (her own). Here we look at names from the other direction. The most trendy:



The Most Unisex Names in US History

BY NATHAN YAU / POSTED TO DATA UNDERLOAD / TAGS: NAMES

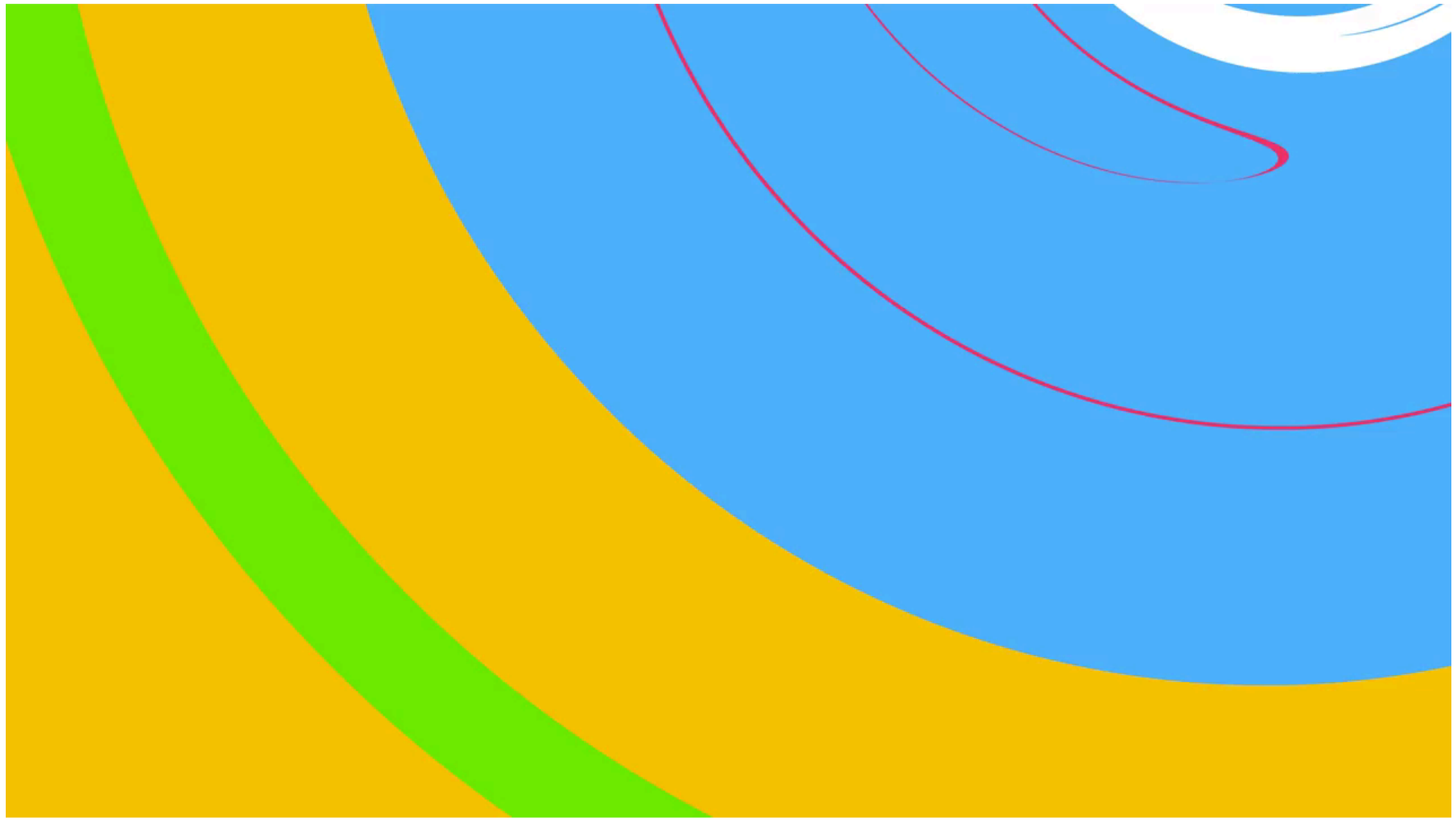


Source: Nathan Yau, http://flowingdata.com/2013/07/29/the-most-trendy-names-in-us-history/; /2013/09/25/the-most-unisex-names-in-us-history/

Names over time

Source: *Abacaba*

Names over time



Source: *Abacaba*

COGS 9
Introduction to Data Science

Data and information

UC San Diego

Today's Learning Objective

Be able to define "data", the two broad kinds of data, what data mining is, and how data mining and visualization might be used for decision-making.

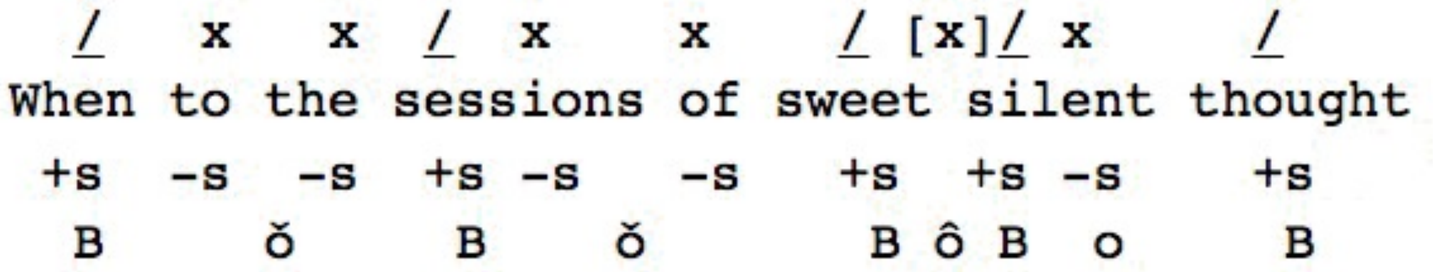
Today's Learning Objective

Be able to define "information", how it relates to uncertainty and "surprise", and how it can be transmitted

When to the sessions of sweet silent thought
I summon up remembrance of things past,
I sigh the lack of many a thing I sought,
And with old woes new wail my dear time's waste:
Then can I drown an eye, unused to flow,
For precious friends hid in death's dateless night,
And weep afresh love's long since cancelled woe,
And moan the expense of many a vanished sight:
Then can I grieve at grievances foregone,
And heavily from woe to woe tell o'er
The sad account of fore-bemoanèd moan,
Which I new pay as if not paid before.
But if the while I think on thee, dear friend,
All losses are restored and sorrows end.

Source: Shakespeare, Sonnet 30

Scansion



- "the method or practice of determining and (usually) graphically representing the metrical pattern of a line of verse.In classical poetry, these patterns are quantitative based on the different lengths of each syllable."

“I have a friend who's an artist and has sometimes taken a view which I don't agree with very well. He'll hold up a flower and say "look how beautiful it is," and I'll agree. Then he says "I as an artist can see how beautiful this is but you as a scientist take this all apart and it becomes a dull thing," and I think that he's kind of nutty. First of all, the beauty that he sees is available to other people and to me too, I believe.

Source: Feynman, *The Pleasure of Finding Things Out* 1999

When to the **sessions** of sweet silent thought
I summon up remembrance of things past,
I sigh the lack of many a thing I sought,

- "In Renaissance England, where the poem was written, the word "sessions" had a specific technical meaning: it refers to court sessions, the period of the year when magistrates and judges heard legal cases. The metaphor thus suggests that the speaker experiences "sweet silent thought" as a kind of tribunal: a place of trial and questioning. The metaphor might even suggest that the speaker starts the poem feeling guilty."

Source: Shakespeare, Sonnet 30; *LitCharts*

Although I may not be quite as refined aesthetically as he is ... I can appreciate the beauty of a flower. At the same time, I see much more about the flower than he sees. I could imagine the cells in there, the complicated actions inside, which also have a beauty. I mean it's not just beauty at this dimension, at one centimeter; there's also beauty at smaller dimensions, the inner structure, also the processes. The fact that the colors in the flower evolved in order to attract insects to pollinate it is interesting; it means that insects can see the color. It adds a question: does this aesthetic sense also exist in the lower forms? Why is it aesthetic? All kinds of interesting questions which the science knowledge only adds to the excitement, the mystery and the awe of a flower. It only adds. I don't understand how it subtracts."

Source: Feynman, *The Pleasure of Finding Things Out* 1999

The Pleasure of Finding Things Out

- This class is like data *poetry*.

The Pleasure of Finding Things Out

- This class is like data *poetry*.
- Its goal is to give you an *appreciation for* data.

The Pleasure of Finding Things Out

- This class is like data *poetry*.
- Its goal is to give you an *appreciation for* data.
- Your job, from here on out, is to learn the *hard stuff*.

For instance, I stand at the seashore, alone, and start to think. There are the rushing waves . . . mountains of molecules, each stupidly minding its own business . . . trillions apart . . . yet forming white surf in unison.

Ages on ages . . . before any eyes could see . . . year after year . . . thunderously pounding the shore as now. For whom, for what? . . . on a dead planet, with no life to entertain.

Never at rest . . . tortured by energy . . . wasted prodigiously by the sun . . . poured into space. A mite makes the sea roar.

Source: Feynman, *The Value of Science* 1955

Deep in the sea, all molecules repeat the patterns of one another till complex new ones are formed. They make others like themselves . . . and a new dance starts.

Growing in size and complexity . . . living things, masses of atoms, DNA, protein . . . dancing a pattern ever more intricate.

Out of the cradle onto the dry land . . . here it is standing . . . atoms with consciousness . . . matter with curiosity.

Stands at the sea . . . wonders at wondering . . . I . . . a universe of atoms . . . an atom in the universe.

Source: Feynman, *The Value of Science* 1955

Deeper appreciation

of data

Reasons for data visualization

- Getting a sense of your data
  - Is it just positive integers (like counting the number of stars)?
  - Or can it assume nearly any positive value (like the radius of a circle)?
  - Or is it constrained (like the number of children a person can have)?

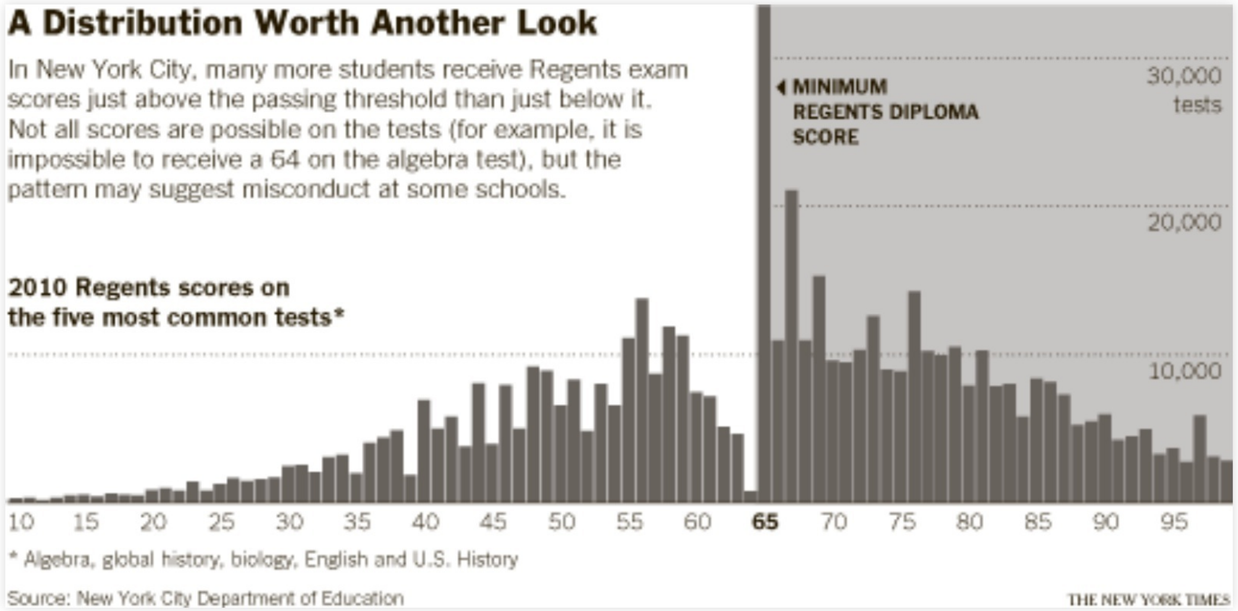
Reasons for data mining

- Data checking
  - Are there outliers?
  - How are the data distributed?

Data distributions

Grades from NYC Regents Exam

A **histogram** of these test scores *forces us to notice* something somewhat problematic



Benford's Law

- Probability of observing a first digit of d base B is $log_B(1+(1/d))$
- Base 10: about 30% are 1s (ones)

UC San Diego

Source: SJ Miller

Benford's Law

In 1881, Newcomb explained that his discovery of the significant-digit law was motivated by an observation that the pages of a book of logarithms were dirtiest in the beginning and progressively cleaner throughout. In

Source: Hill, *Chance* 1999

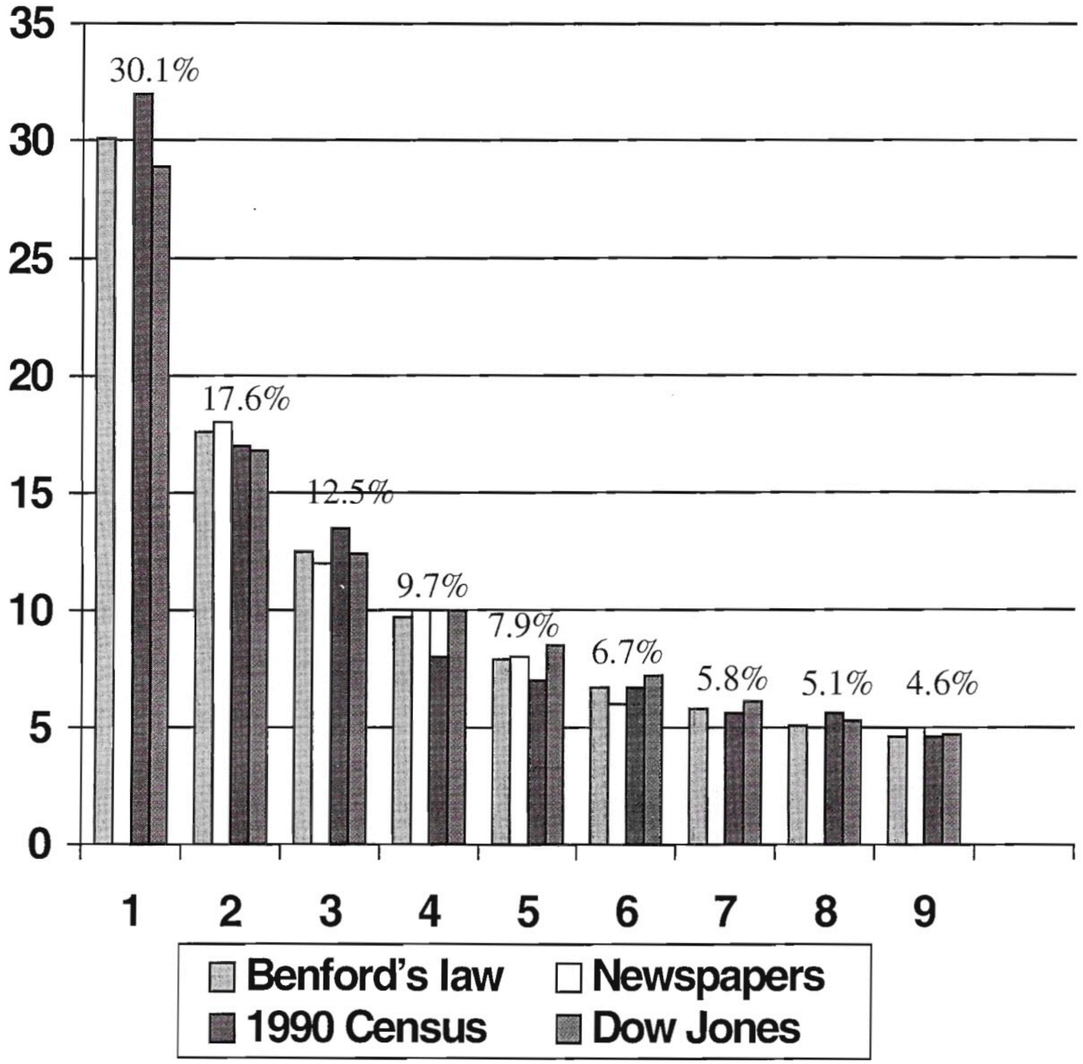


Figure 1. Benford's law predicts a decreasing frequency of first digits, from 1 through 9. The frequencies in datasets developed by Benford for numbers appearing on the front pages of newspapers, by Mark Nigrini of 3,141 county populations in the 1990 U.S. Census, and by Eduardo Ley of the Dow Jones Industrial Average from 1918–93 follows Benford's law (the numbers given atop each set of columns) within 2%.

Source: Hill, *Chance* 1999

Benford's Law applications

Exhibit 3: Check Fraud in Arizona

The table lists the checks that a manager in the office of the Arizona State Treasurer wrote to divert funds for his own use. The vendors to whom the checks were issued were fictitious.

Date of Check	Amount
October 9, 1992	$ 1,927.48
	27,902.31
October 14, 1992	86,241.90
	72,117.46
	81,321.75
	97,473.96
October 19, 1992	93,249.11
	89,658.17
	87,776.89
	92,105.83
	79,949.16
	87,602.93
	96,879.27
	91,806.47
	84,991.67
	90,831.83
	93,766.67
	88,338.72
	94,639.49
	83,709.28
	96,412.21
	88,432.86
	71,552.16
TOTAL	$ 1,878,687.58

Source: SJ Miller

Benford's Law applications

- Most amounts below $100,000 (critical threshold for data requiring additional scrutiny).

- Over 90% had first digit of 7, 8 or 9.

Source: SJ Miller

Flip some coins!

Benford's Law *mis*applications

EVERYTHINGNEWS NOVEMBER 10, 2020 / 1:25 PM / UPDATED 2 MONTHS AGO

Fact check: Deviation from Benford's Law does not prove election fraud

By Reuters Staff

9 MIN READ

UC San Diego

Benford's Law *mis*applications

Theodore P. Hill, Professor Emeritus of Mathematics at Georgia Tech, Atlanta, cautioned that regardless of the distribution uncovered, the application of Benford's Law would not provide definitive evidence that fraud took place.

"First, I'd like to stress that Benford's Law can NOT be used to "prove fraud"," he told Reuters by email. "It is only a Red Flag test, that can raise doubts. E.g., the IRS has been using it for decades to ferret out fraudsters, but only by identifying suspicious entries, at which time they put the auditors to work on the hard evidence. Whether or not a dataset follows BL proves nothing."