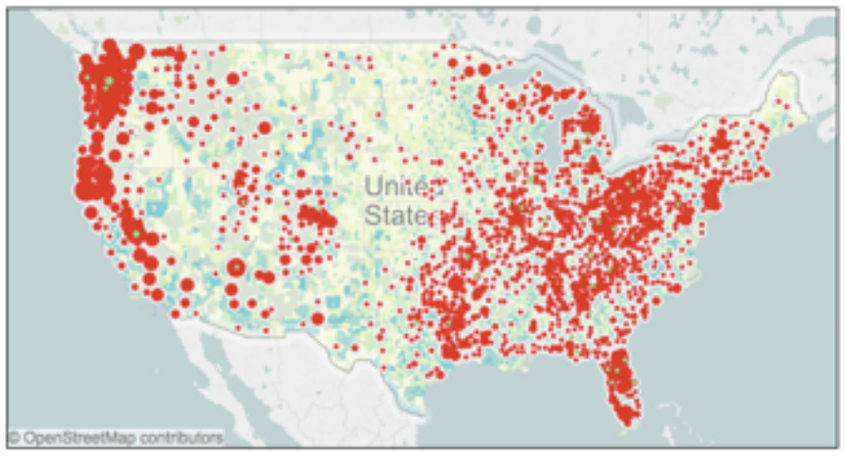
CS109/Stat121/AC209/E-109
Data Science
Communication and Storytelling

Hanspeter Pfister, Joe Blitzstein, and Verena Kaynig

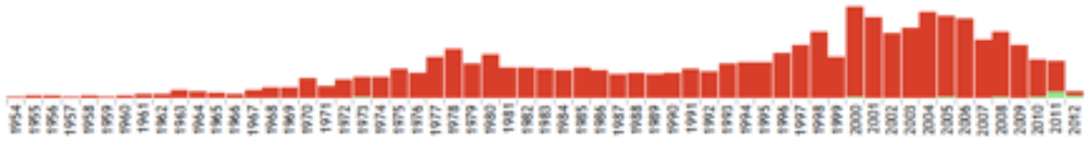


Data gathered from the official website of the **"Bigfoot Field Researchers Organization"** (BFRO).

The data was *attempted* to be scrubbed and cleaned to attain some type of normalcy, unfortunately the BFRO data submission process has no validation and fields are often used arbitrarily by submitters.

BFRO does the "Finding Bigfoot"

Click on ANY element of the visualization *(location, season, year, detail field)* in order to filter by that item.
Select the element AGAIN to go back to the full view.



The BFRO classifies sightings according to a system based on the sightings "potential for misinterpretation".

Total Sightings	Class A	Class B	Class C	Unclassified
3,806	**1,951**	**1,696**	**31**	**128**

Alabama	Baldwin County	1979	September	*Class A*	Man recalls a sighting after Hurricane Frederic north of Mobile
	Barbour County	1980	January	*Class A*	Motorists pulled over on a rural highway at night describe standoff in headlights e..
	Bibb County	1987	August	*Class B*	Rescue workers describes possible stalking on the Cahaba River outside Montevall..

This Week

- HW1 is due this Thursday (September 24) at 11:59 pm (Eastern Time)
- Remember to attend section. If you need to change sections, try to swap with someone (there is a Piazza thread for this).
- Make sure you have read the homework policies on the syllabus, and follow the homework submission procedure carefully. See https://piazza.com/class/icf0cypdc3243c?cid=451
- Always check your submission.
- Late days are calculated based on the time stamp of the *last* push to your respository.

Two Fundamental Questions

1. What is the goal?

- predict future data?
- explain and understand a phenomenon?
- test a hypothesis?
- compare two groups?
- dimension reduction?
- build a good recommendation system?
- decide on a course of action or a policy?

Two Fundamental Questions

2. Who cares?

IMAC

I: inferential goal (scientific question of interest)
M: model (all models are wrong, some are useful)
A: algorithms
C: conclusions and checking

The C is crucial: what did we learn? Was the model useful, and how well does it fit? How do we know whether the method is working? Do we need to iterate and improve the model? What are the limitations and future directions?

Some Key Principles

- remember The Golden Rule
- know your audience
- tell a story
- choose and use notation carefully
- read great writers
- create good sense of direction (with the help of *signposts*), with clear flow of logic

Notation, notation, notation

It was said of Jordan's writings that if he had four things on the same footing (as a, b, c, d) they would appear as $a, M_3', \epsilon_2, \Pi_{1,2}''$.

- J.E. Littlewood

Halmos' nightmare: $n_\varepsilon \to 0$ as $\varepsilon \to \infty$.

Four useful references on scientific writing

Marie Davidian: http://www4.stat.ncsu.edu/~davidian/st810a/written_handout.pdf

Rod Little: http://sitemaker.umich.edu/rlittle/files/styletips.pdf

Paul Halmos: http://www.matem.unam.mx/ernesto/LIBROS/Halmos-How-To-Write%20Mathematics.pdf

George Gopen and Judith Swan: http://engineering.missouri.edu/civil/files/science-of-writing.pdf

The Science of Scientific Writing (Gopen-Swan)

http://engineering.missouri.edu/civil/files/science-of-writing.pdf

The smallest of the URF's (URFA6L), a 207-nucleotide (nt) reading frame overlapping out of phase the NH2-terminal portion of the adenosinetriphosphatase (ATPase) subunit 6 gene has been identified as the animal equivalent of the recently discovered yeast H+-ATPase subunit 8 gene. The functional significance of the other URF's has been, on the contrary, elusive. Recently, however, immunoprecipitation experiments with antibodies to purified, rotenone-sensitive NADH-ubiquinone oxido-reductase [hereafter referred to as respiratory chain NADH dehydrogenase or complex I] from bovine heart, as well as enzyme fractionation studies, have indicated that six human URF's (that is, URF1, URF2, URF3, URF4, URF4L, and URF5, hereafter referred to as ND1, ND2, ND3, ND4, ND4L, and ND5) encode subunits of complex I. This is a large complex that also contains many subunits synthesized in the cytoplasm.

But what about *structure*, not just jargon?

The smallest of the URF's, an [A] has been identified as a [B] subunit 8 gene. The functional significance of the other URF's has been, on the contrary, elusive. Recently, however, [C] experiments, as well as [D] studies, have indicated that six human URF's [1-6] encode subunits of Complex I. This is a large complex that also contains many subunits synthesized in the cytoplasm.

How are these sentences connected?
What is the emphasis?

The Science of Scientific Writing (Gopen-Swan)

http://engineering.missouri.edu/civil/files/science-of-writing.pdf

Recently, however, immunoprecipitation experiments with antibodies to purified, rotenone-sensitive NADH-ubiquinone oxido-reductase [hereafter referred to as respiratory chain NADH dehydrogenase or complex I] from bovine heart, as well as enzyme fractionation studies, have indicated that six human URF's (that is, URF1, URF2, URF3, URF4, URF4L, and URF5, hereafter referred to as ND1, ND2, ND3, ND4, ND4L, and ND5) encode subunits of complex I.

The Science of Scientific Writing (Gopen-Swan)

http://engineering.missouri.edu/civil/files/science-of-writing.pdf

Recently, however, immunoprecipitation experiments with antibodies to purified, rotenone-sensitive NADH-ubiquinone oxido-reductase [hereafter referred to as respiratory chain NADH dehydrogenase or complex I] from bovine heart, as well as enzyme fractionation studies, have indicated that six human URF's (that is, URF1, URF2, URF3, URF4, URF4L, and URF5, hereafter referred to as ND1, ND2, ND3, ND4, ND4L, and ND5) encode subunits of complex I.

The Science of Scientific Writing (Gopen-Swan)

http://engineering.missouri.edu/civil/files/science-of-writing.pdf

Recently, however, immunoprecipitation experiments with antibodies to purified, rotenone-sensitive NADH-ubiquinone oxido-reductase [hereafter referred to as respiratory chain NADH dehydrogenase or complex I] from bovine heart, as well as enzyme fractionation studies, have indicated that six human URF's (that is, URF1, URF2, URF3, URF4, URF4L, and URF5, hereafter referred to as ND1, ND2, ND3, ND4, ND4L, and ND5) encode subunits of complex I.

Linda the Bank Teller (Kahneman-Tversky)

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

The option with the least conditions is of course the most probable

85% of Stanford Business School students participating in the study said option 2 is more probable.

Statistica Sinica **7** (1997), 211-220

THE EVOLUTION OF A PROBLEM

Henry S. Baird and Colin L. Mallows

Bell Laboratories, AT & T Laboratories

Dedicated to Herbert Robbins on the occasion of his 80th birthday

Abstract: This paper describes several problems, all arising from one real-world problem. Some of these problems have been solved, others offer interesting challenges.

Abstract MadLibs!!

This paper presents a _____ method for _____
(synonym for *new*) (sciencey verb)
the _____. Using _____, the
(noun few people have heard of) (something you didn't invent)
_____ was measured to be _____ +/- _____
(property) (number) (number)
_____. Results show _____ agreement with
(units) (sexy adjective)
theoretical predictions and significant improvement over
previous efforts by _____, et al. The work presented
(Loser)
here has profound implications for future studies of
_____ and may one day help solve the problem of
(buzzword)
_____.
(supreme sociological concern)

Keywords: _____, _____, _____
(buzzword) (buzzword) (buzzword)

title: "Abstract Mad Libs" - originally published 1/14/2009

Tell a Story!

Any story has a beginning, a middle, and an end.

- introduce interesting characters
- put them in a predicament
- resolve the predicament
- but leave room for sequels! (Limitations and future work)

Tell a Story with Data

Stories

Stories are the most powerful delivery tool for information, more powerful and enduring than any other art form

New York Times

Key Considerations

- Who is your **audience**?
- What **questions** are you answering?
- Why should the audience **care**? get them excited about the topic
- What are your major **insights** and surprises?
- What **change** to you want to affect?

Know Your Audience

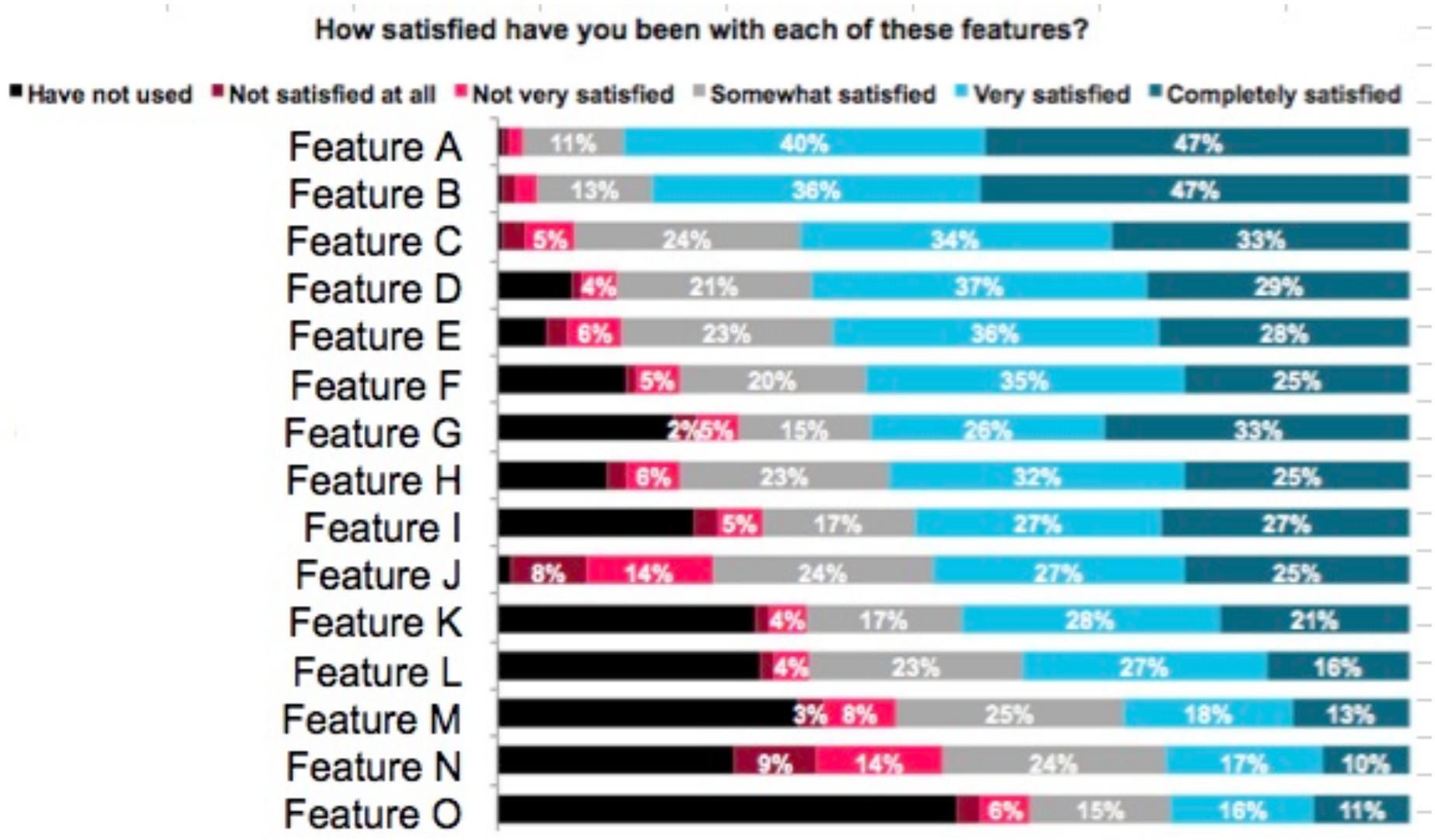
Know Your Audience

- What do they know?
- What motivates them? What do they desire?
- What experiences do you share? What are common goals?
- What insights can you give them? What tools and “magical gifts”?

Don’t Make Them Think!

- Your audience does not want to spend cognitive effort on things you know and can just show them
- Lead them through the major steps of your story
- Point out interesting key facts and insights using captions and annotations

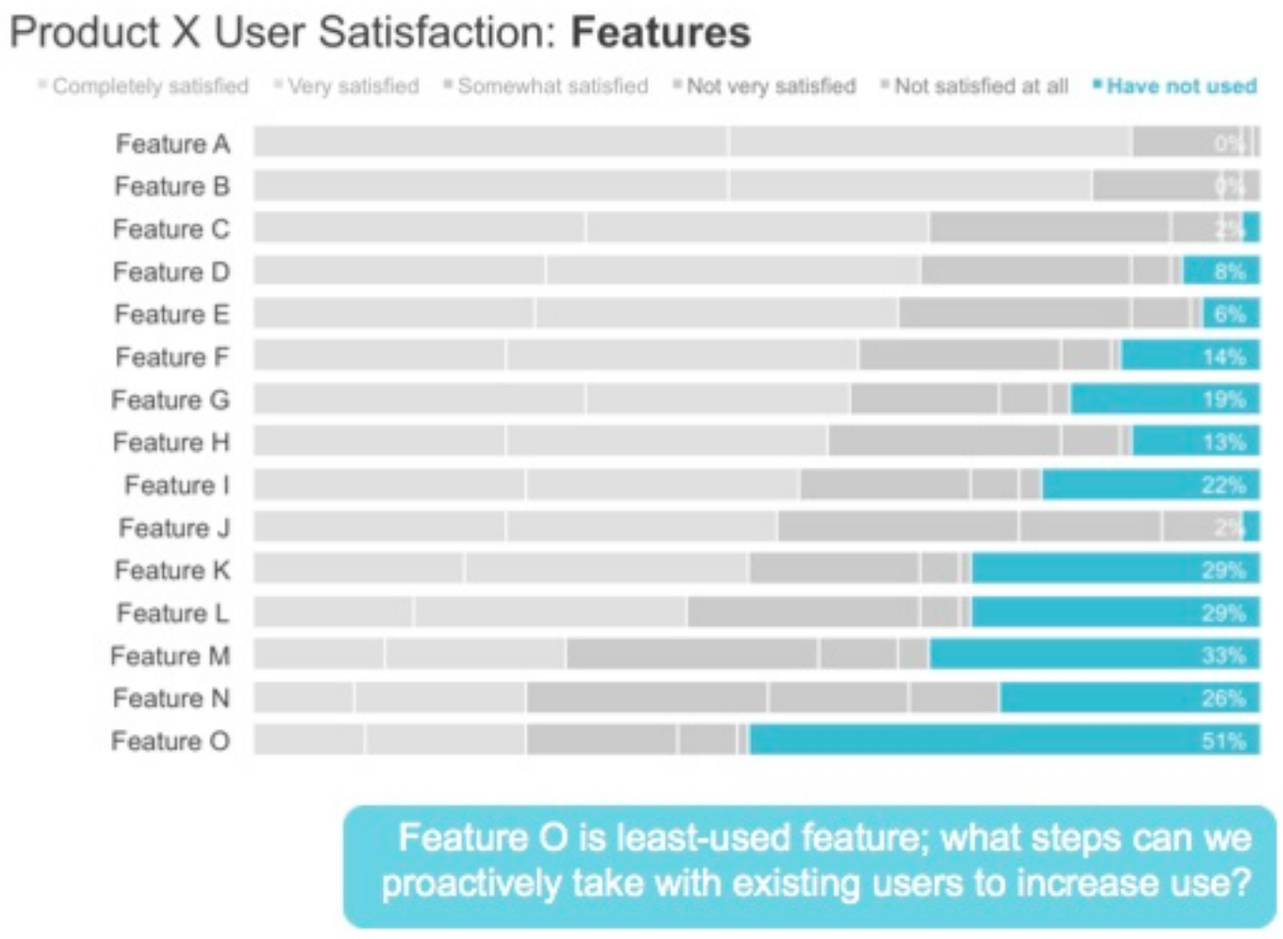
Don’t Bury the Lead



Cole Nussbaumer

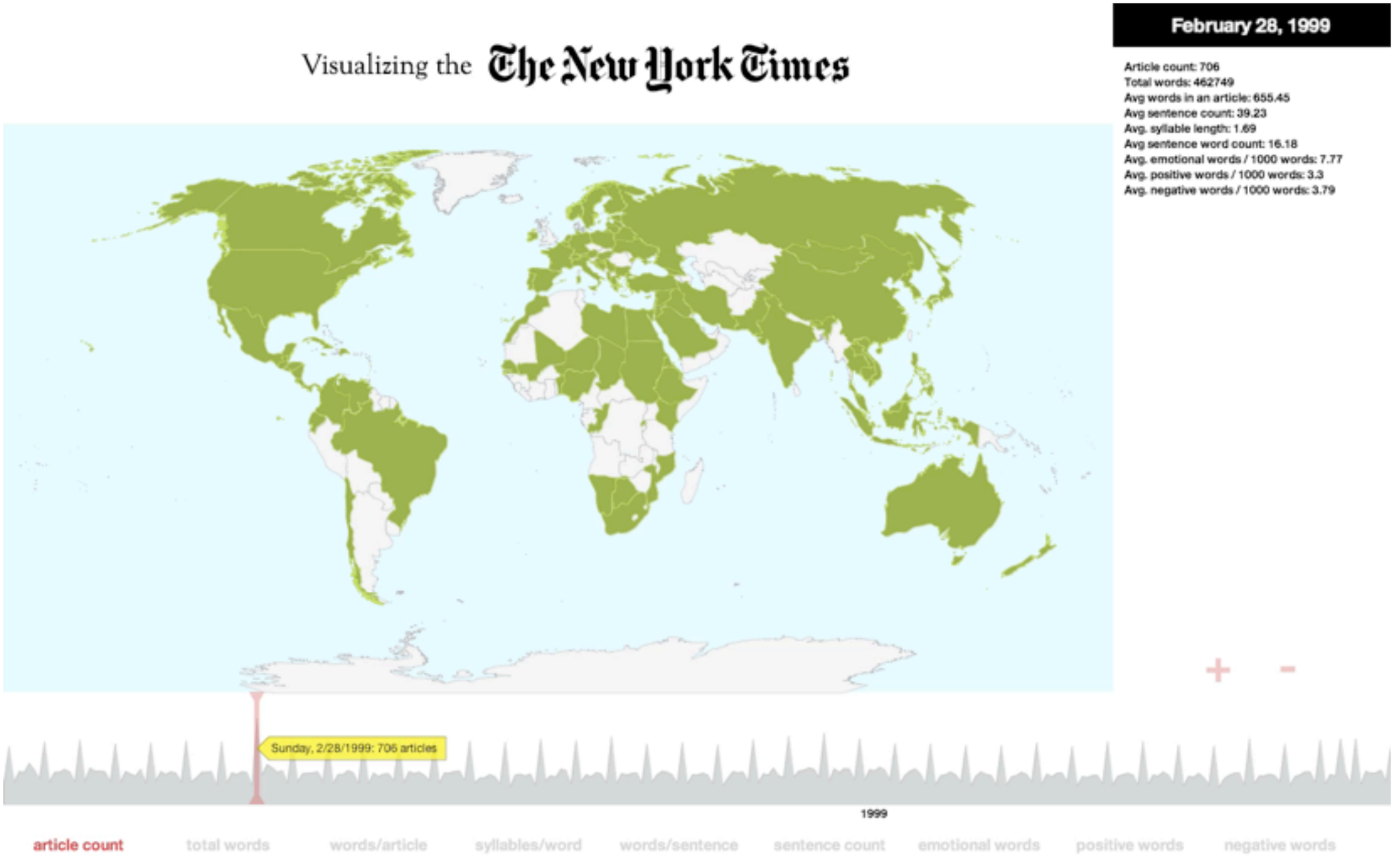
Don’t Bury the Lead

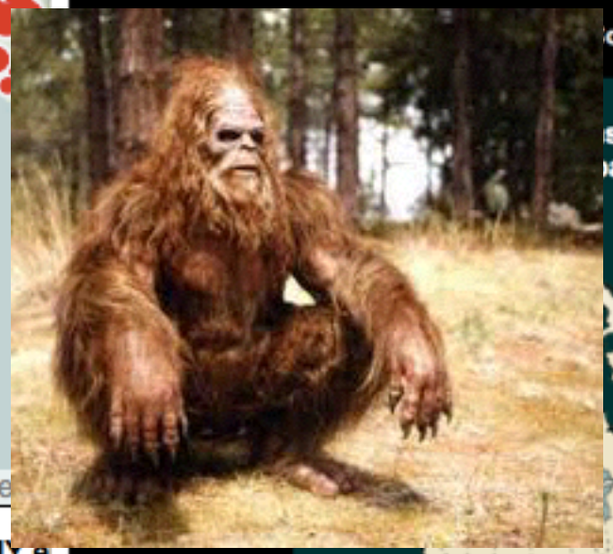
User satisfaction varies greatly by feature

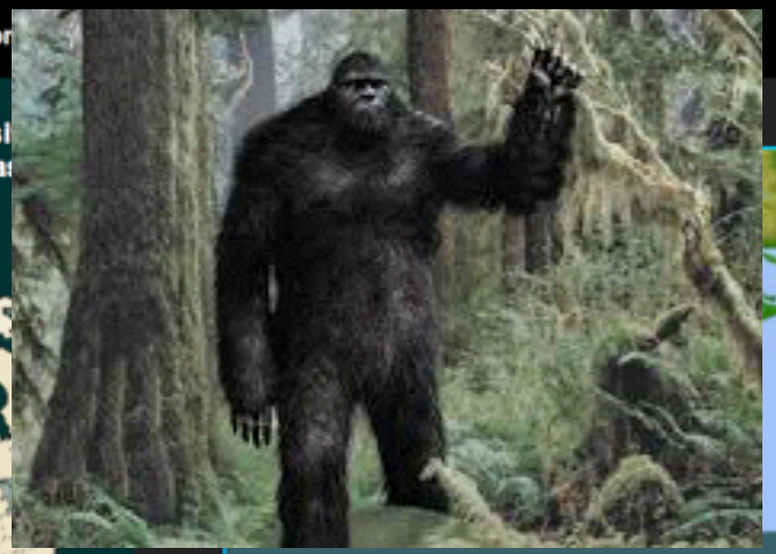


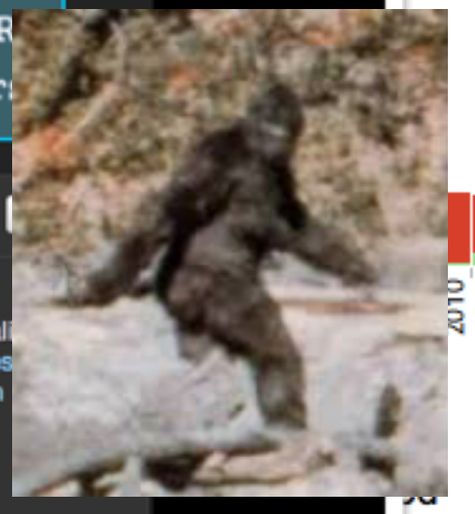
Cole Nussbaumer

David Jacopille



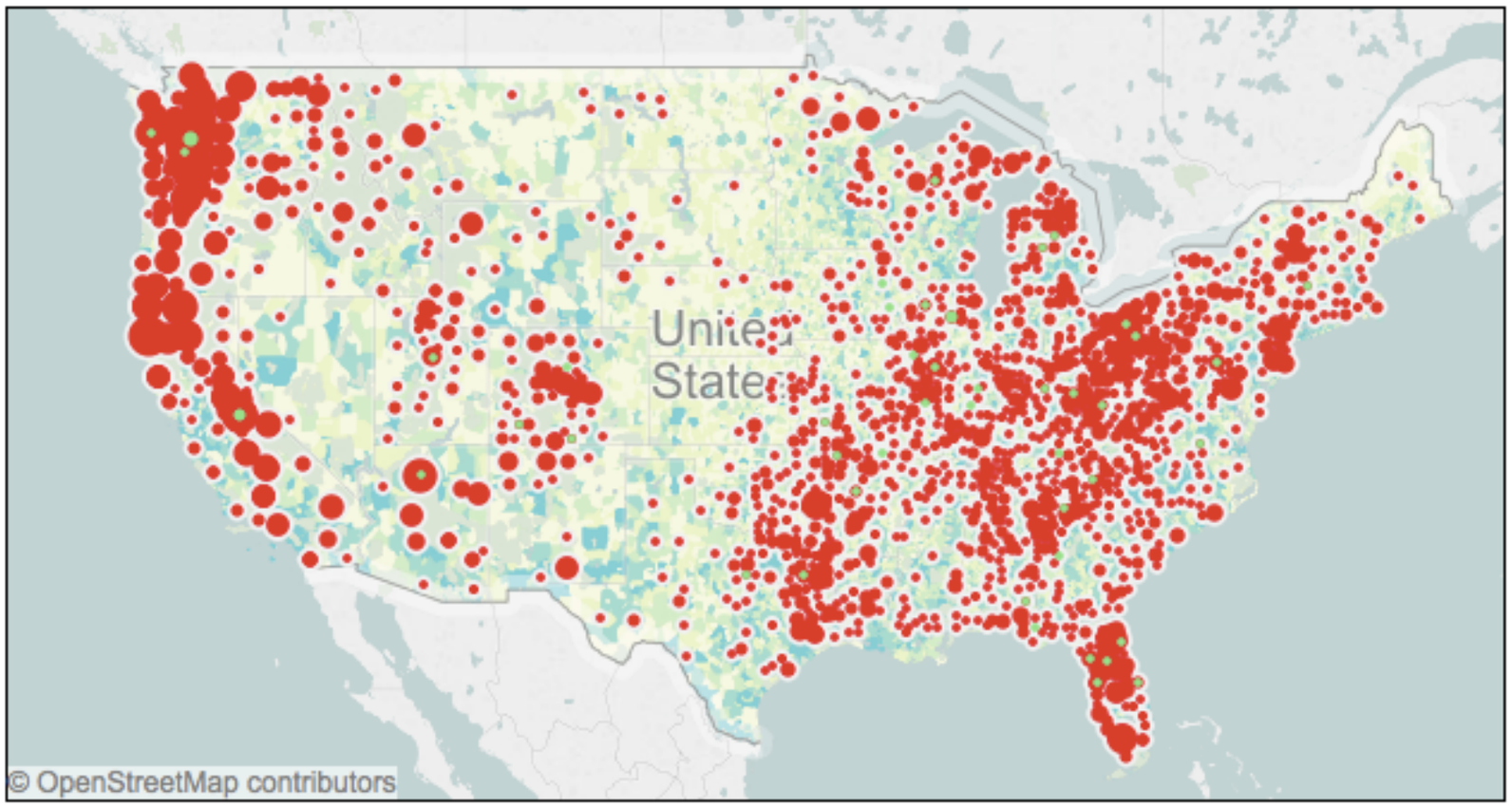


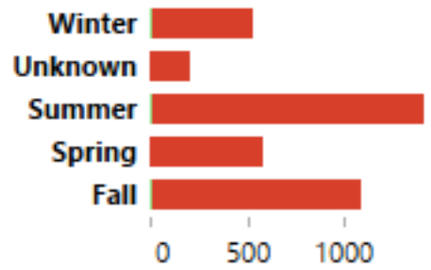




Ryan Robitalle

Where is Bigfoot seen in the USA?



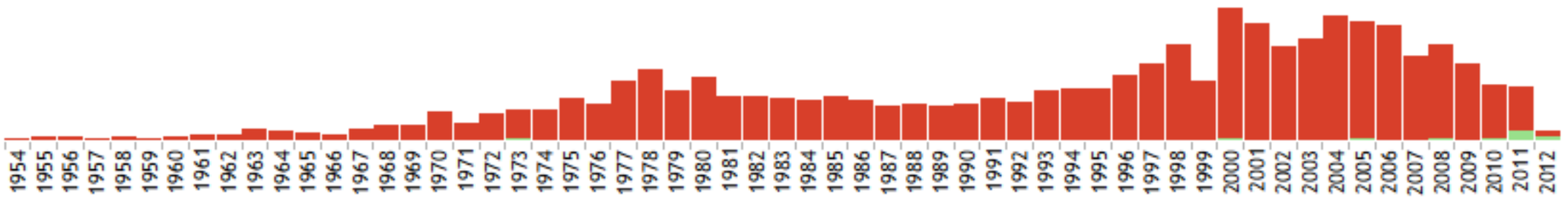


Data gathered from the official website of the **"Bigfoot Field Researchers Organization"** (BFRO).

The data was *attempted* to be scrubbed and cleaned to attain some type of normalcy, unfortunately the BFRO data submission process has no validation and fields are often used arbitrarily by submitters.

BFRO does the "*Finding Bigfoot*"

Click on ANY element of the visualization *(location, season, year, detail field)* in order to filter by that item.
Select the element AGAIN to go back to the full view.



The BFRO classifies sightings according to a system based on the sightings "potential for misinterpretation".

Total Sightings	Class A	Class B	Class C	Unclassified
3,806	**1,951**	**1,696**	**31**	**128**

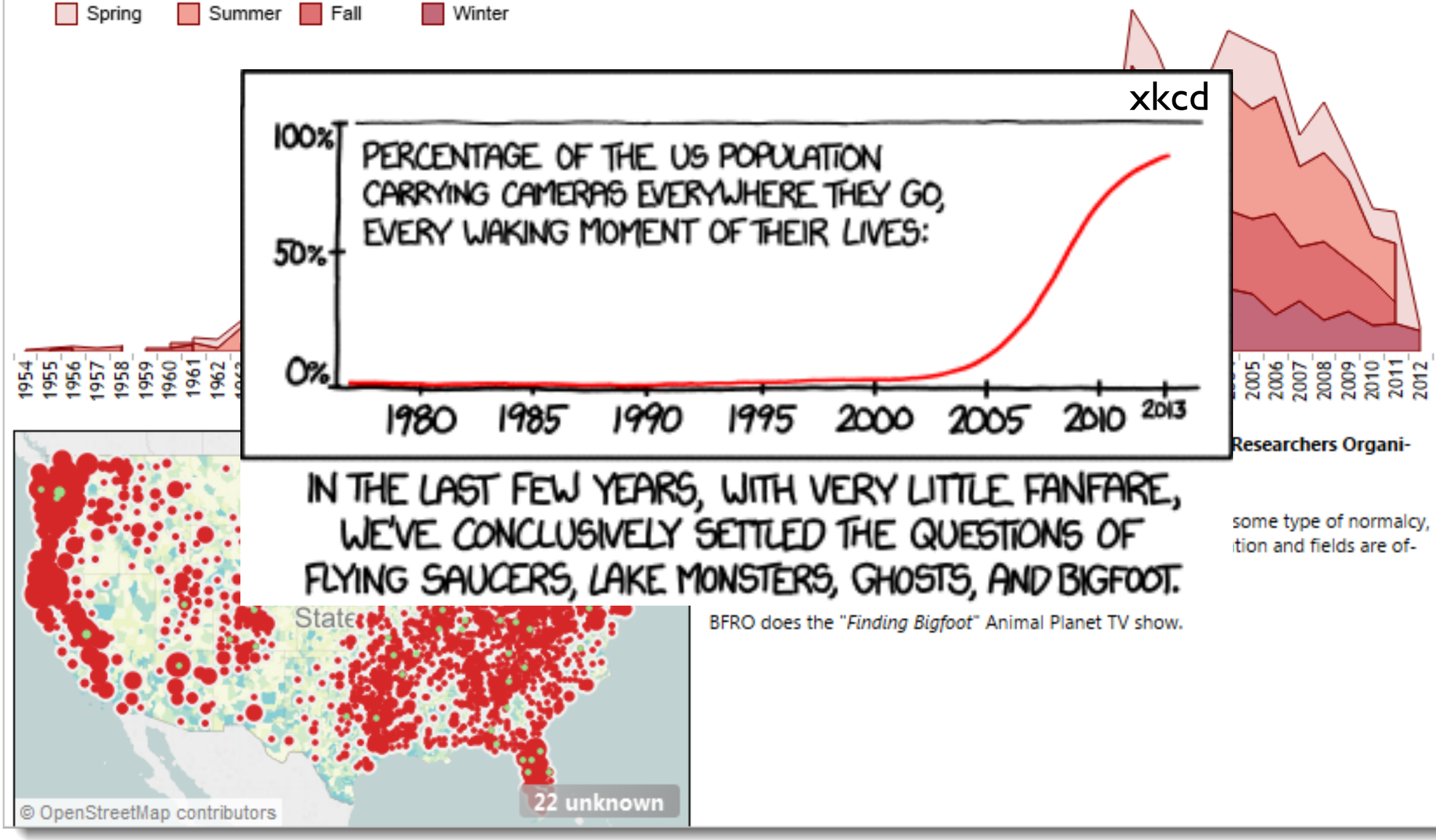
Alabama	Baldwin County	1979	September	*Class A*	Man recalls a sighting after Hurricane Frederic north of Mobile
	Barbour County	1980	January	*Class A*	Motorists pulled over on a rural highway at night describe standoff in headlights e..
	Bibb County	1987	August	*Class B*	Rescue workers describes possible stalking on the Cahaba River outside Montevall..

Bigfoot sightings are in decline

Sightings by Year

Season

Spring Summer Fall Winter

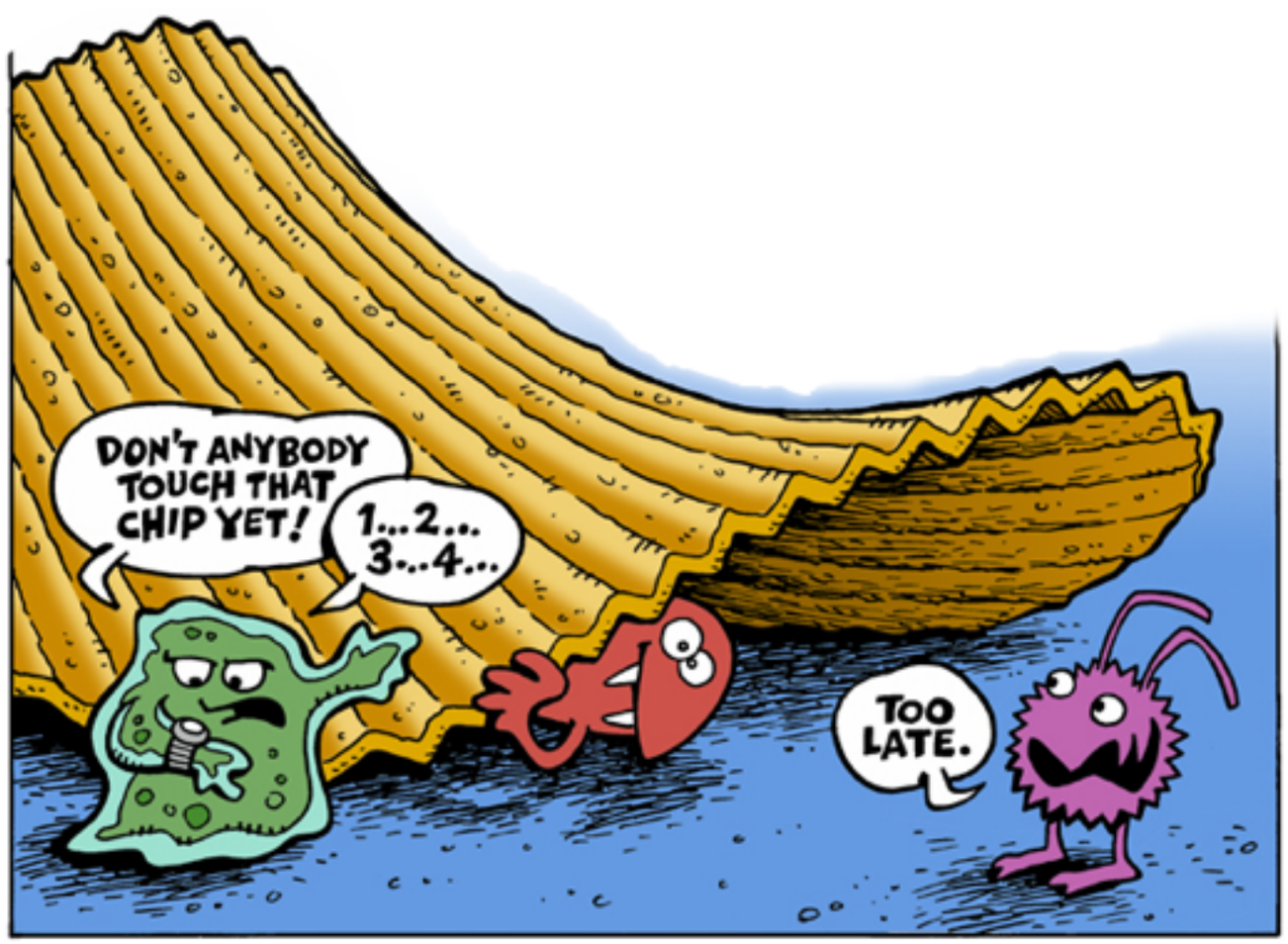


Ryan Robitalle

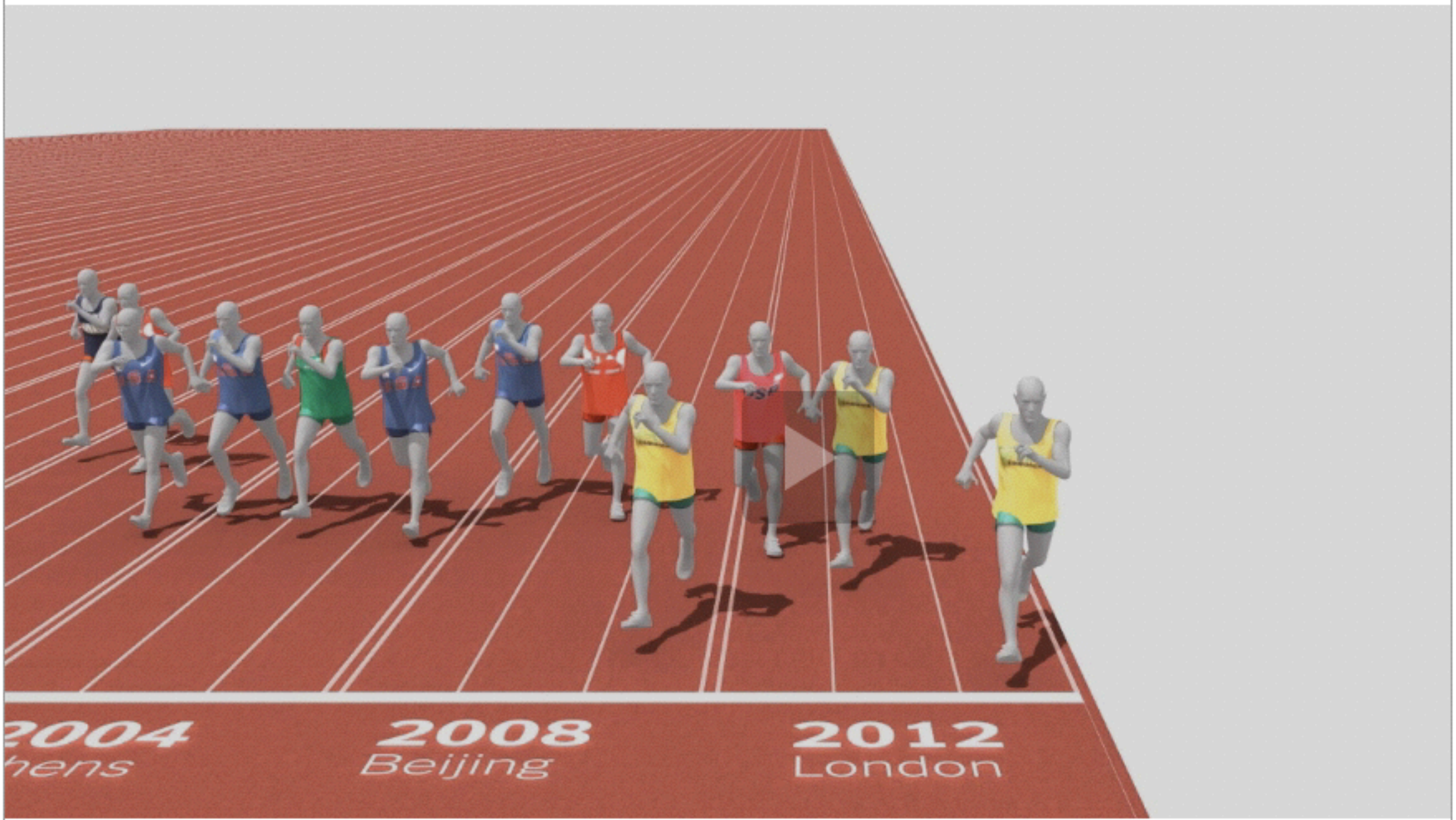
Unexpectedness

Make the audience aware that there is something they didn't know they didn't know

- Use surprise to grab the audience's attention
- "You might think you know this, but here's a new angle on it"
- Curiosity happens when we feel a gap in our knowledge



All the Medalists: **Men's 100-Meter Sprint**



Sources: "The Complete Book of the Olympics" by David Wallechinsky and Jaime Loucky, International Olympic Committee; Amateur Athletic Assocation; Photographs: Chang W. Lee/The New York Times, Getty Images, International Olympic Committee

Messaging

Framing - Why should I care?

- Tell the audience: “Here is the right way to think about the problem I was trying to solve.”
- Catch the audience’s attention and frame the story using captions and annotations
- If done well, your insights will seem obvious given this framing. And that’s a good thing!

Opportunity Dashboard | Sales Dashboard

Sales Dashboard

Total Sales	Number of Deals	Avg Deal Size	Rev. per Salesperson
$3,190.2M	16,610	$189,545	$20.1M

Date Closed
8/7/2006 — 12/31/2010

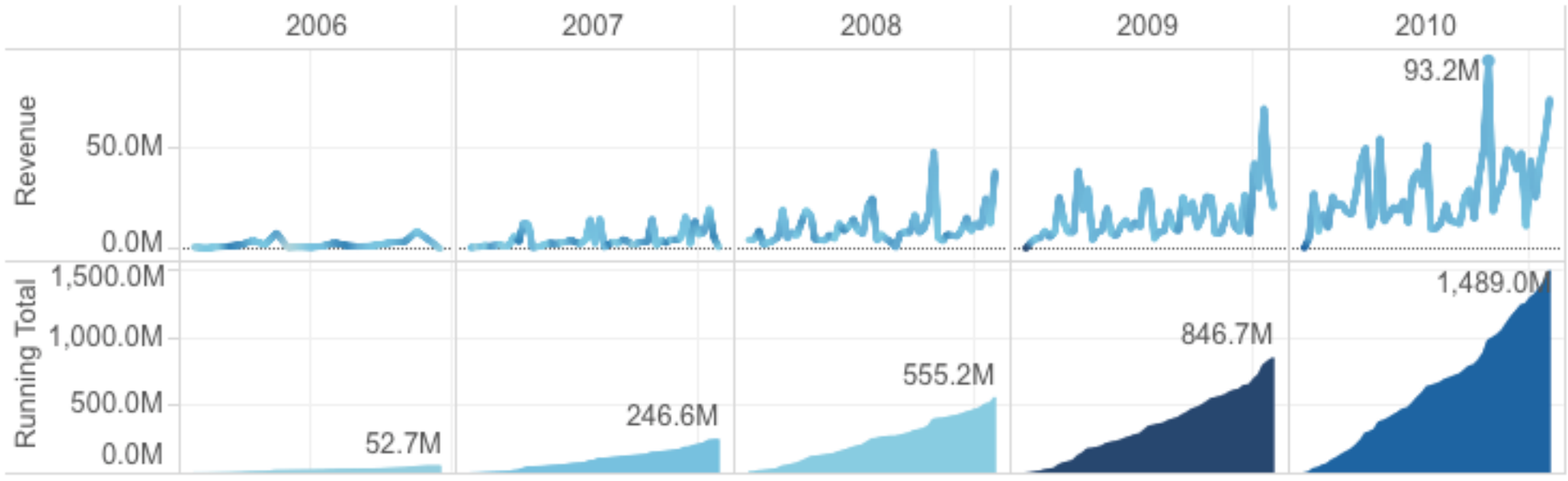
Region
(All)

Country
(All)

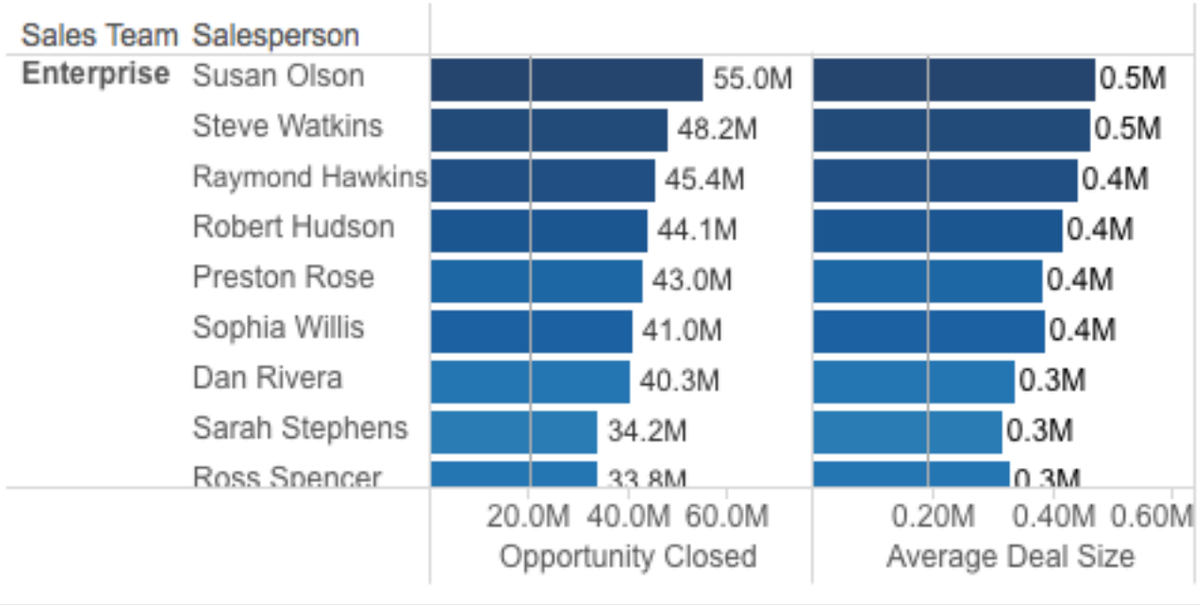
Sales Team
- (All)
- Small and Midmarket
- Enterprise

Avg Deal Size/Salespe...
$130,922 — $336,519

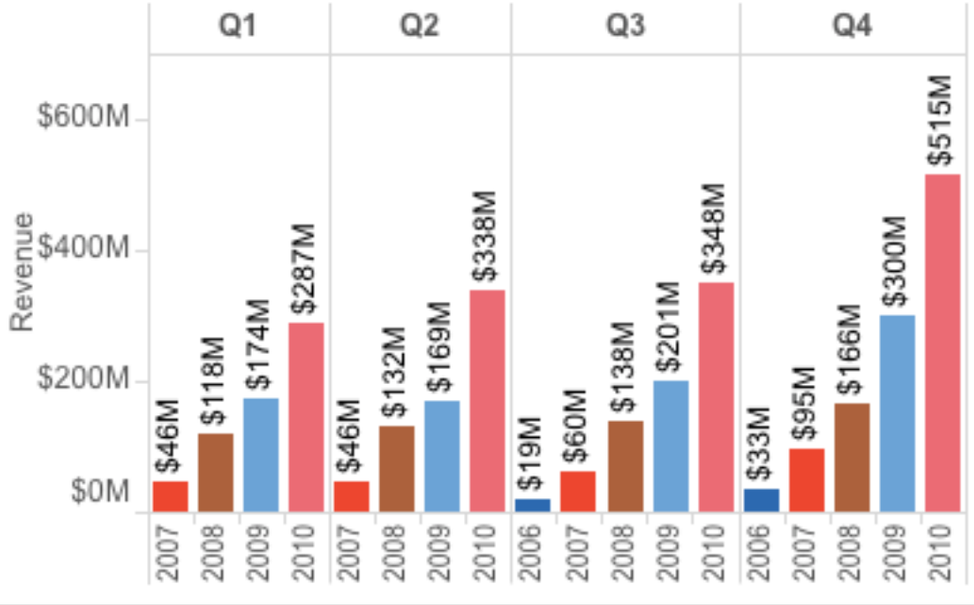
Revenue Over Time



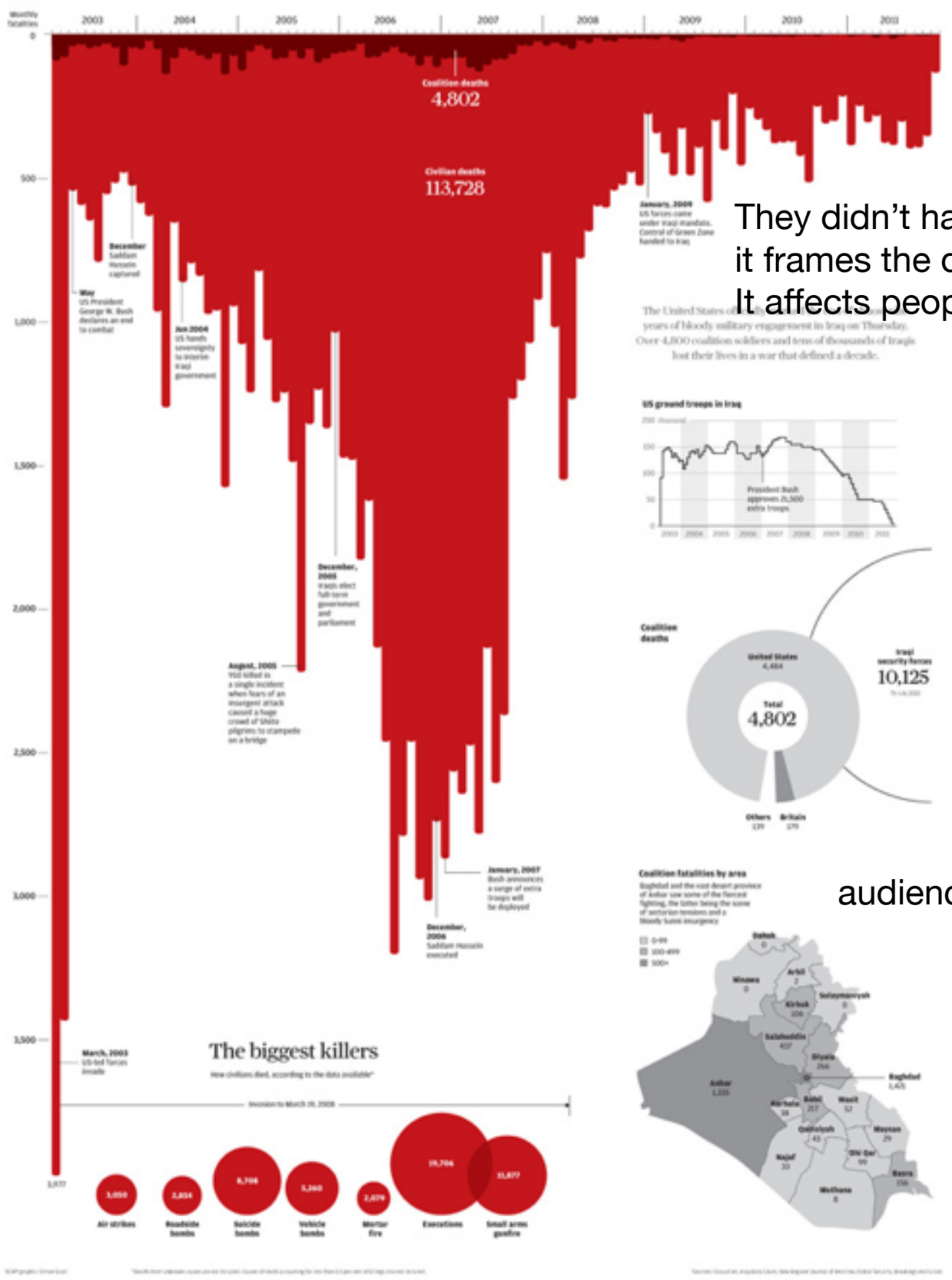
Sales Team Performance



Revenue by Quarter



Iraq's bloody toll

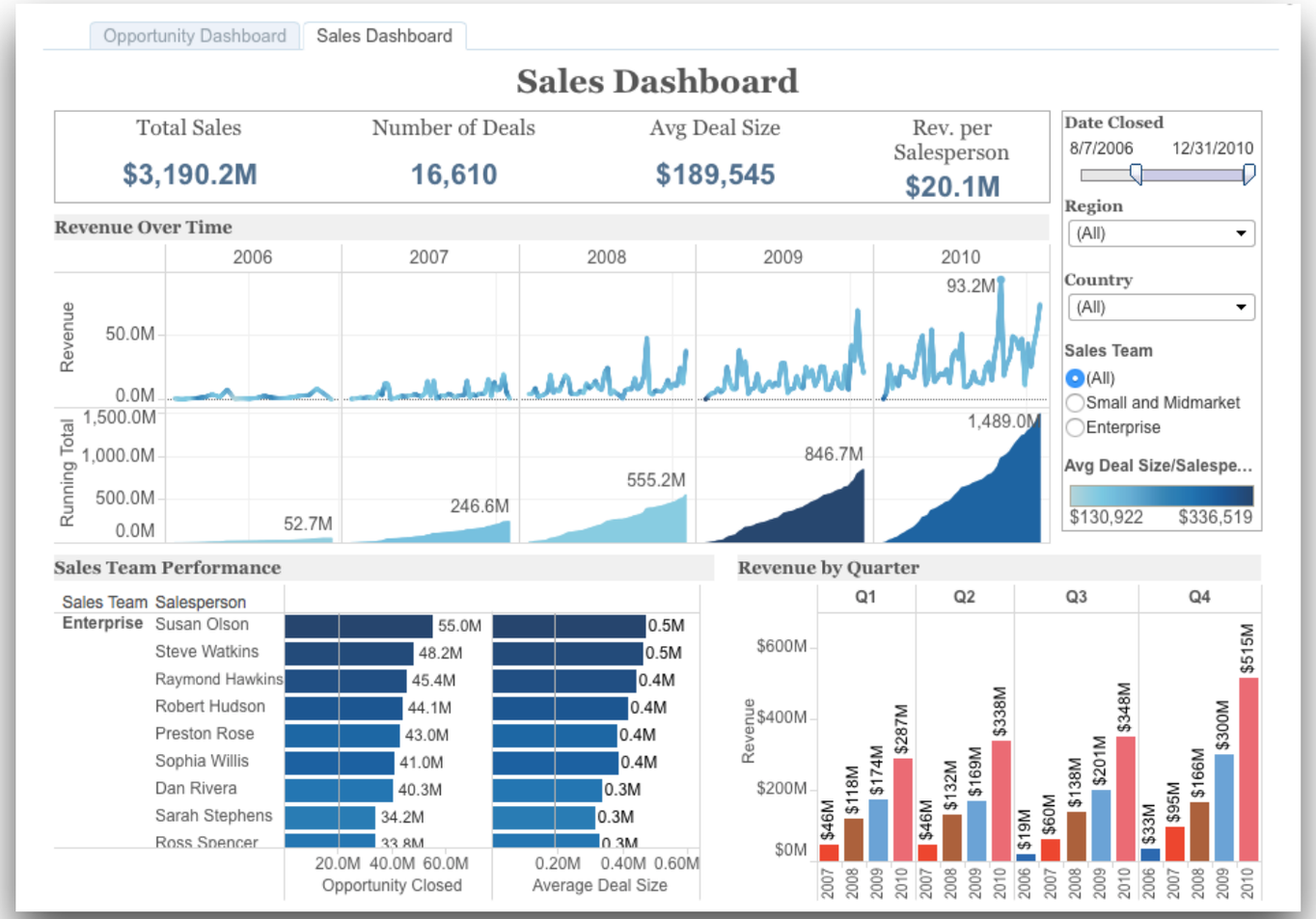


They didn't have to put the bars downwards
it frames the data
It affects people's emotions, it involves them

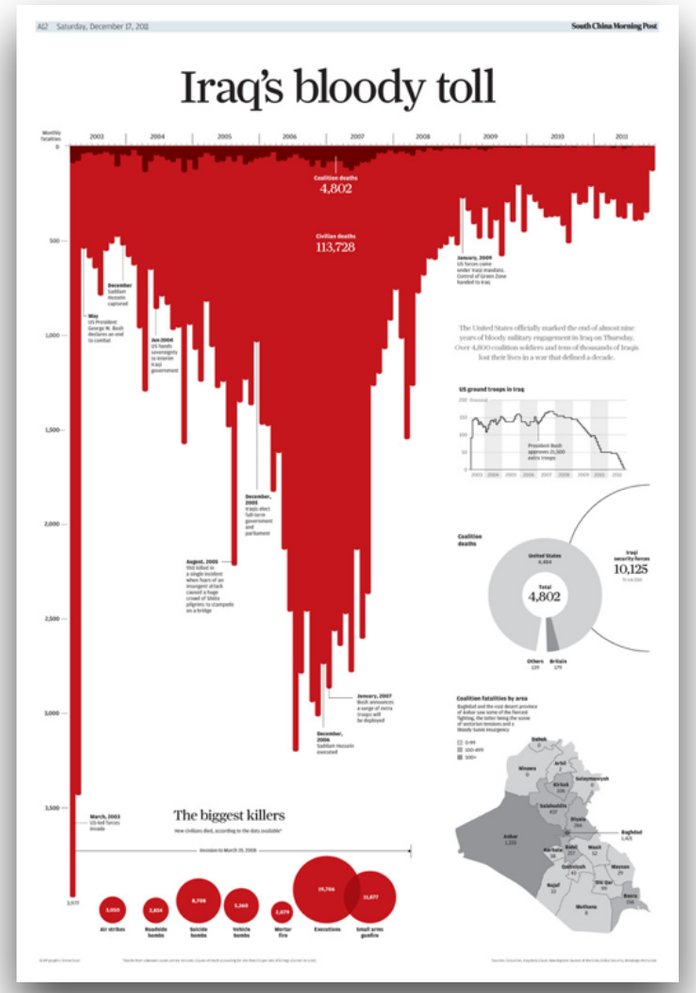
audience is not your phd defense

What is the message?

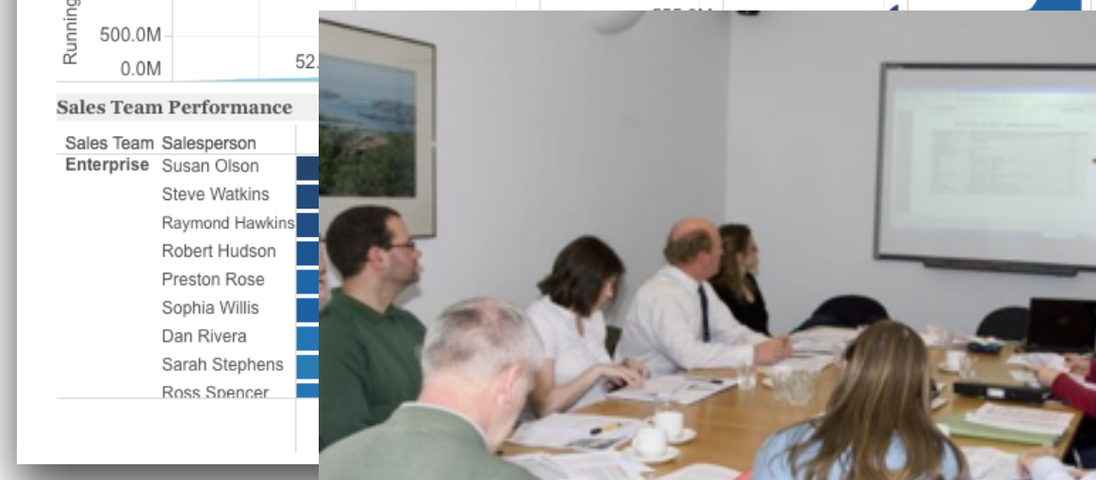
Exploratory
Neutral



Explanatory
Opinionated

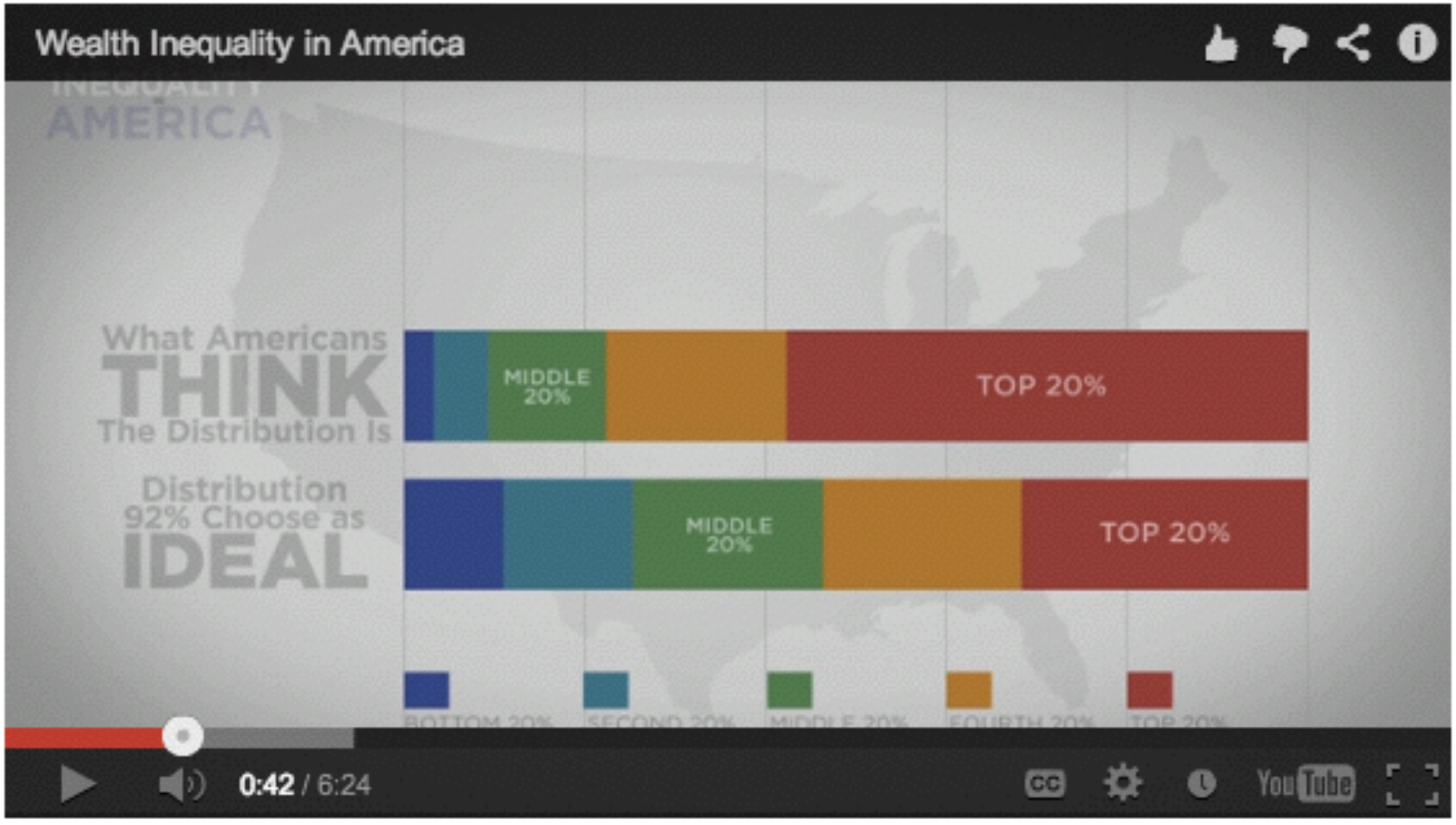


Know Your Audience



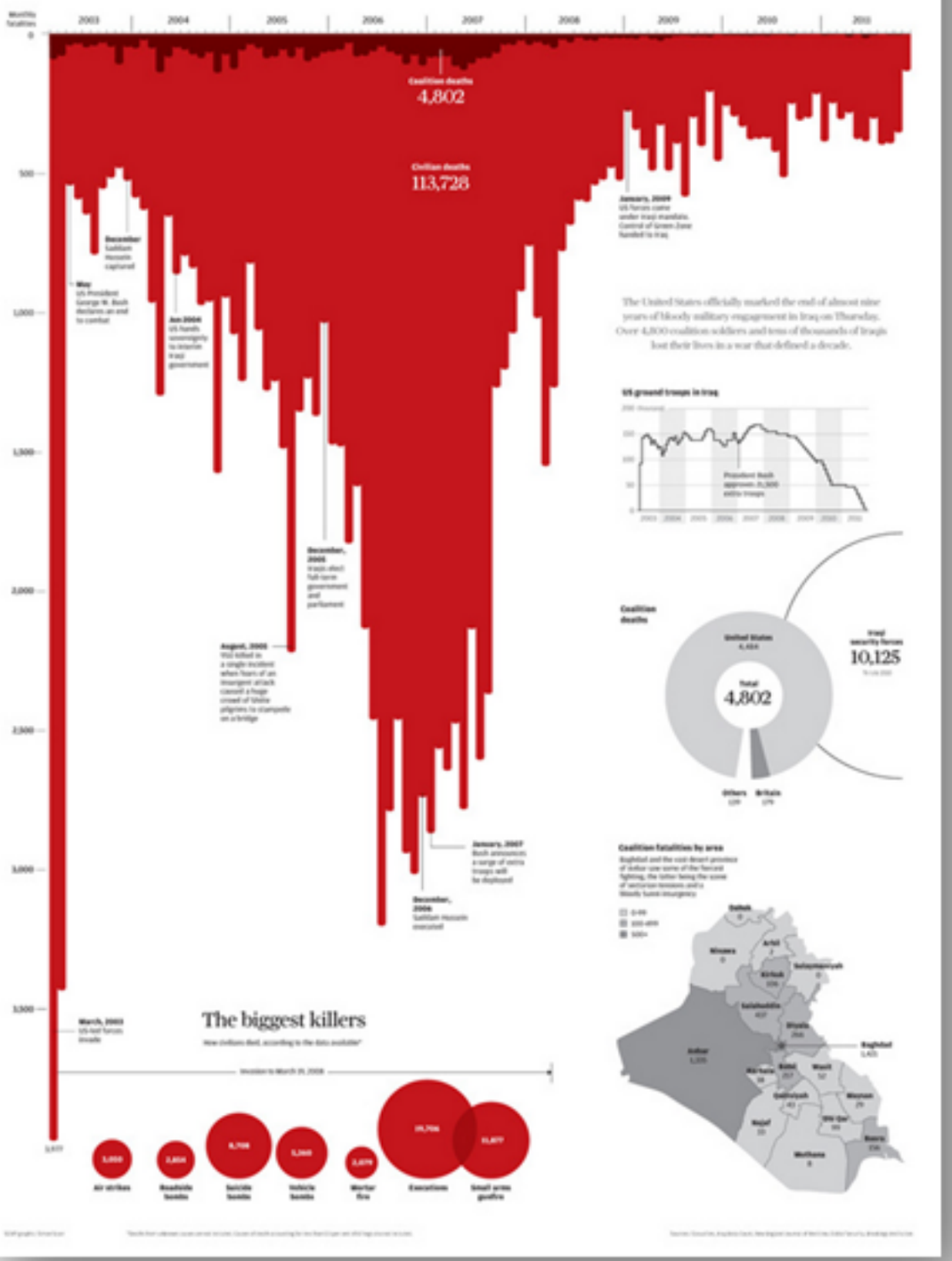
Wealth Inequality in America

RANDY | MONDAY, MARCH 11, 2013 AT 8:08AM PERMALINK

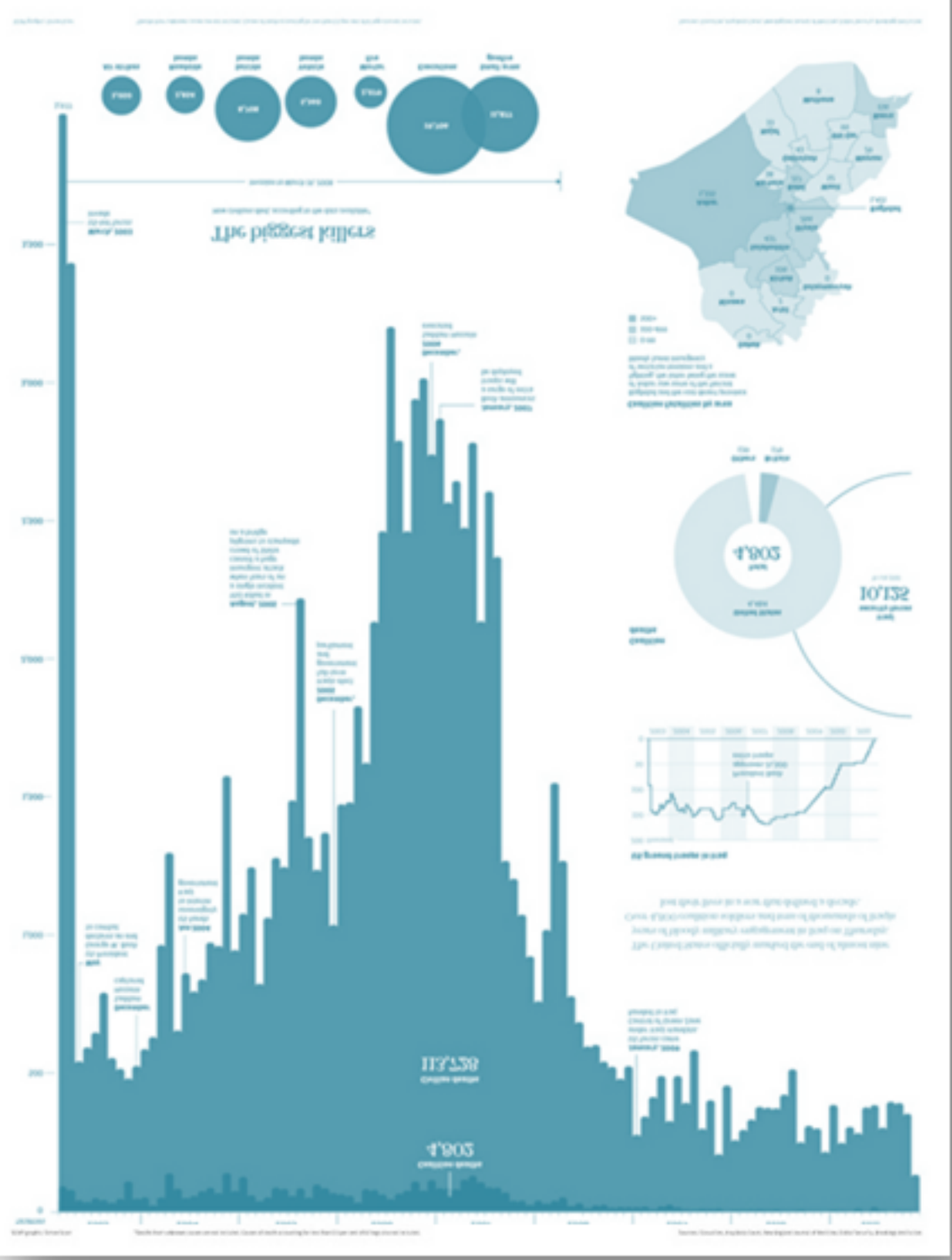


Visual Story Design

Iraq's bloody toll



Iraq: Deaths on the Decline



Andy Cotgreave, Tableau

755

Steroids or Not, the Pursuit Is On

Barry Bonds is taking aim at the career home run record. He needs only six more to be Babe Ruth and 47 to equal Hank Aaron.

Lines are cumulative home runs.

Hank Aaron
755 homers
23 seasons

Babe Ruth
714 homers
22 seasons

Barry Bonds
708 homers
20 seasons

Bonds takes lead
Home runs after 16 seasons

Bonds	567
Aaron	554
Ruth	516

Bonds was injured last season. He played 14 games and hit 5 homers.

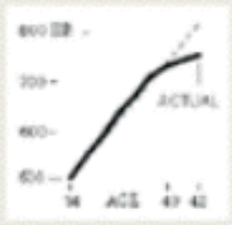
According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.

Homer Pace After Age 34

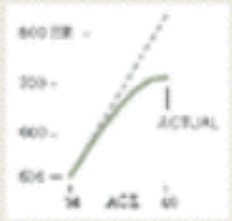
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

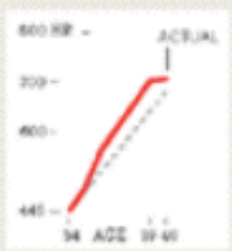
Aaron
Actual homers slightly outpace projected homers for five seasons.



Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.

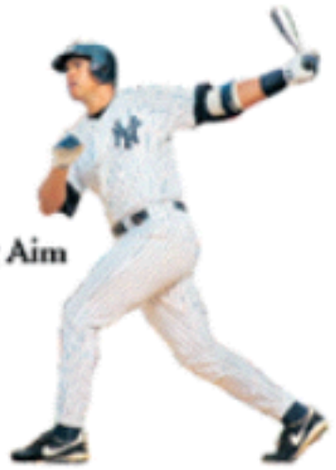


Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.

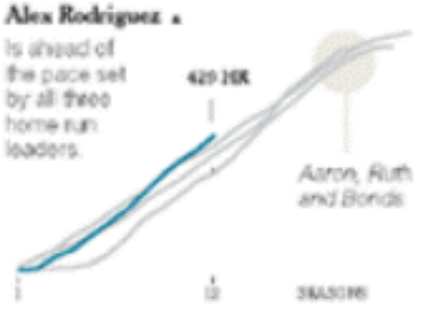


Note: Ages as of July 1 of each season.

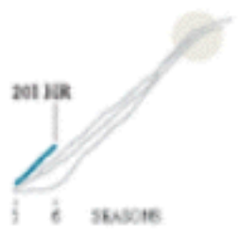
Others Taking Aim



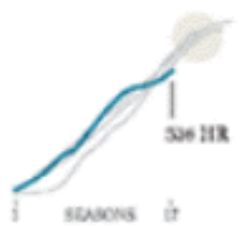
Alex Rodriguez ▲
Is ahead of the pace set by all three home run leaders.



Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.

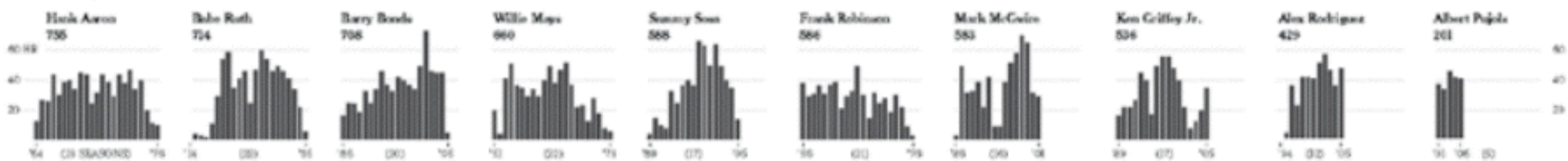


Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.

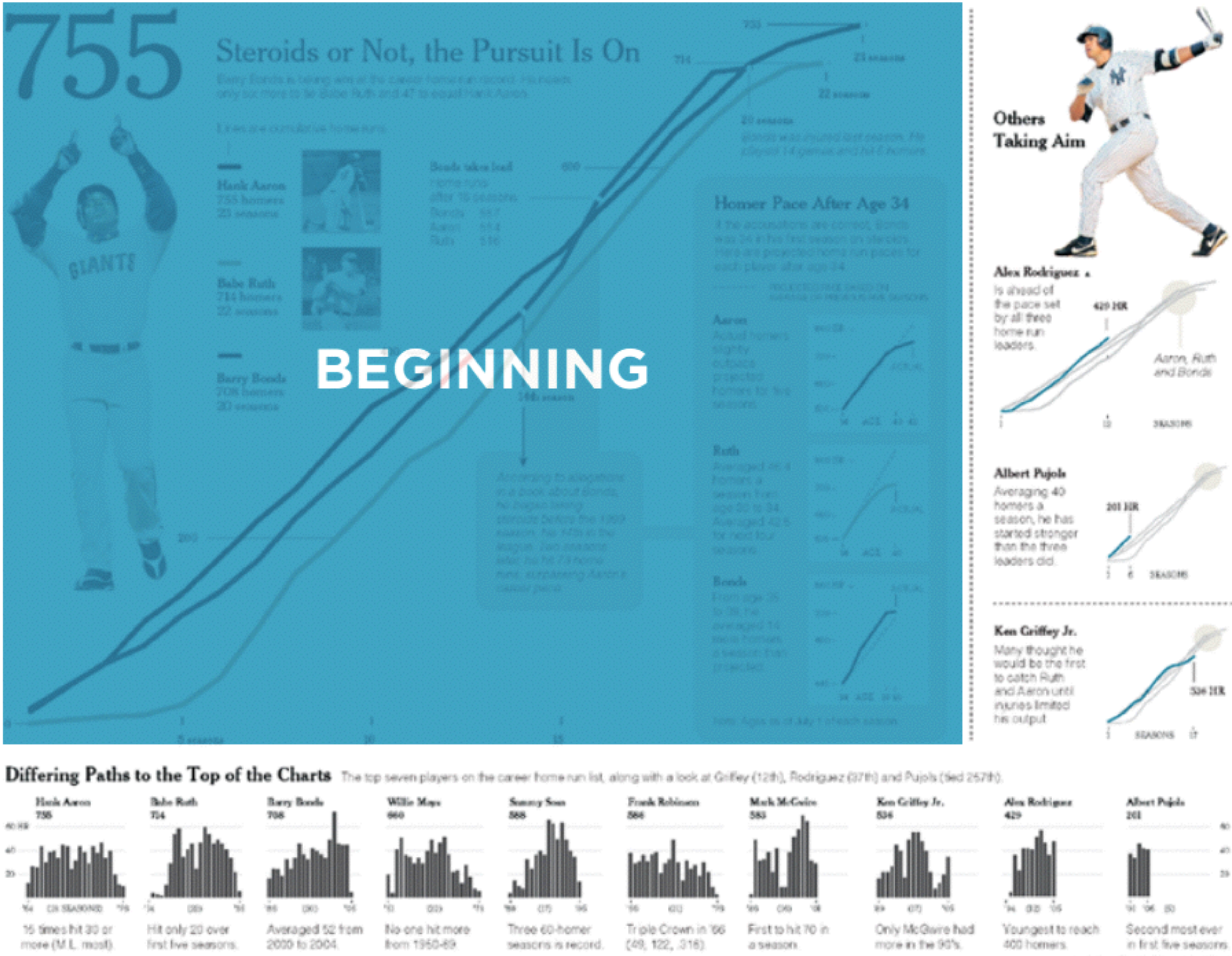


Differing Paths to the Top of the Charts

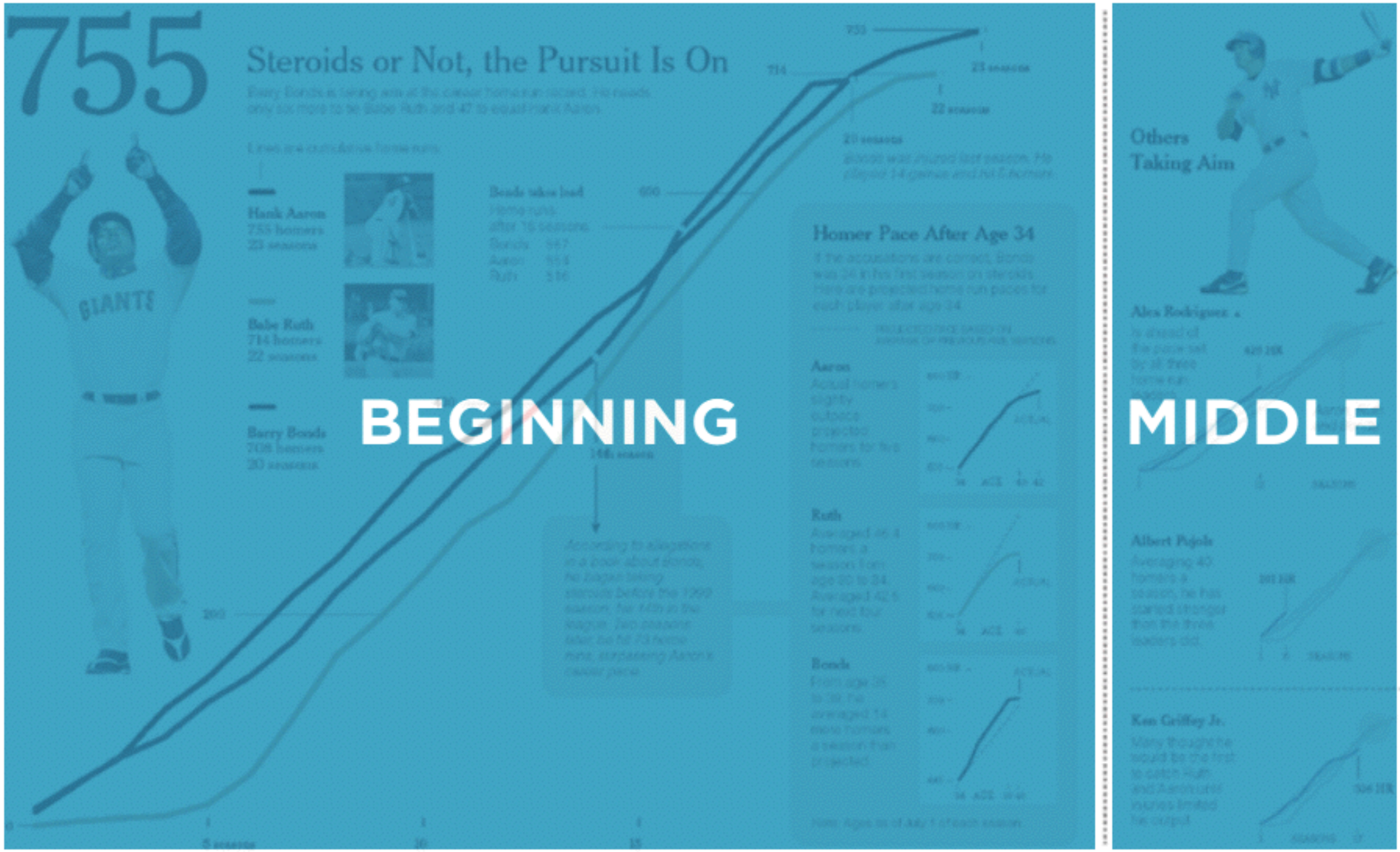
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).



E. Segel

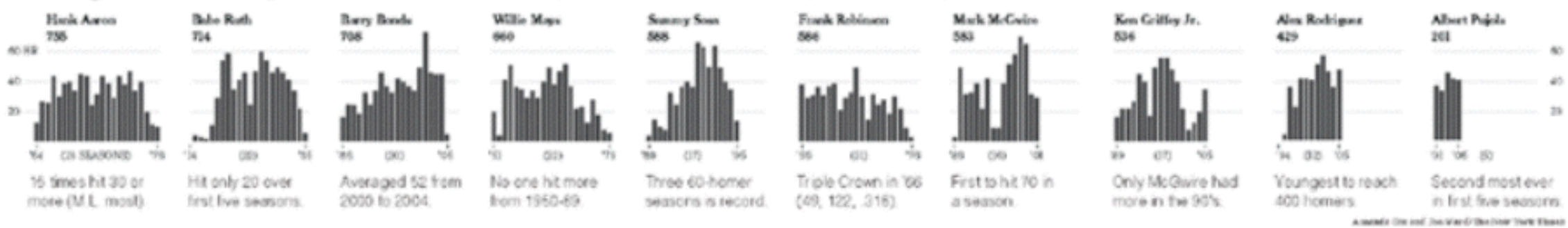

BEGINNING
755
Steroids or Not, the Pursuit Is On
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
Homer Pace After Age 34
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three home run leaders.
Aaron, Ruth and Bonds
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output
Differing Paths to the Top of the Charts
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).
15 times hit 30 or more (M.L. most).
Hit only 20 over first five seasons.
Averaged 52 from 2000 to 2004.
No one hit more from 1950-69.
Three 60-homer seasons is record.
Triple Crown in '66 (49, 122, .316)
First to hit 70 in a season.
Only McGwire had more in the 90's.
Youngest to reach 400 homers.
Second most ever in first five seasons.

E. Segel

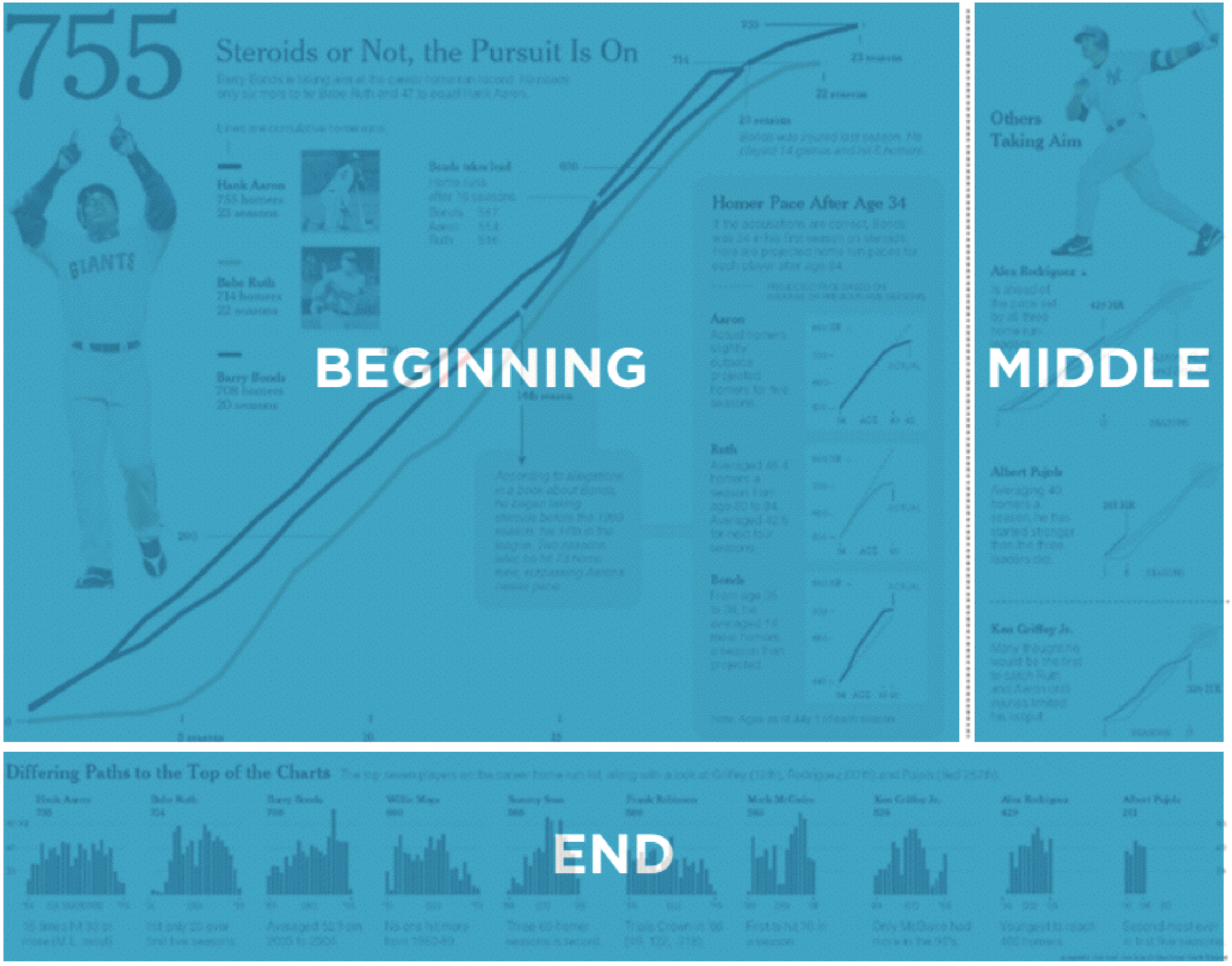


Differing Paths to the Top of the Charts

The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).



E. Segel


755
Steroids or Not, the Pursuit Is On
Barry Bonds is racing ever at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.
Hank Aaron
755 homers
23 seasons
Babe Ruth
714 homers
22 seasons
Barry Bonds
708 homers
20 seasons
BEGINNING
Homer Pace After Age 34
MIDDLE
Others Taking Aim
Alex Rodriguez
Albert Pujols
Ken Griffey Jr.
Differing Paths to the Top of the Charts
END

E. Segel

Headline = Answer to the most important question you could find

755

Steroids or Not, the Pursuit Is On

Barry Bonds is taking aim at the career home run record. He needs only six more to be Babe Ruth and 47 to equal Hank Aaron.

Lines are cumulative home runs

Hank Aaron
755 homers
23 seasons

Babe Ruth
714 homers
22 seasons

Barry Bonds
708 homers
20 seasons

Bonds takes lead
Home runs after 16 seasons
Bonds 587
Aaron 554
Ruth 516

600
400
200
0
1 season
5 seasons
10
15

Call Out Boxes →

Captions →

20 seasons
Bonds was injured last season. He played 14 games and hit 5 homers.

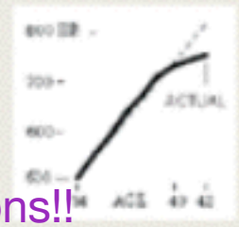
He began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.

Homer Pace After Age 34

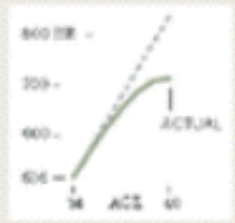
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

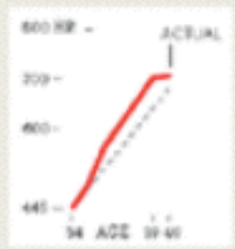
Aaron
Actual homers slightly outpace projected homers for five seasons.



Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.

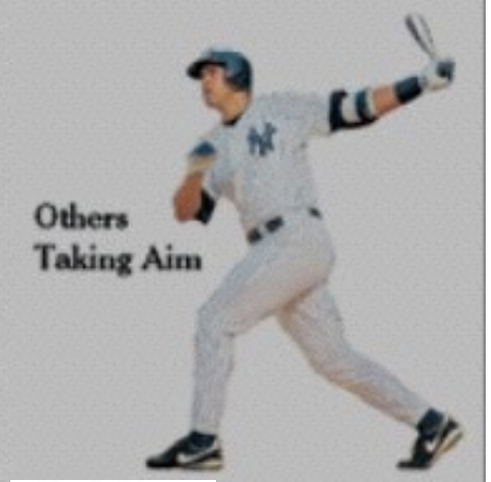


Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.



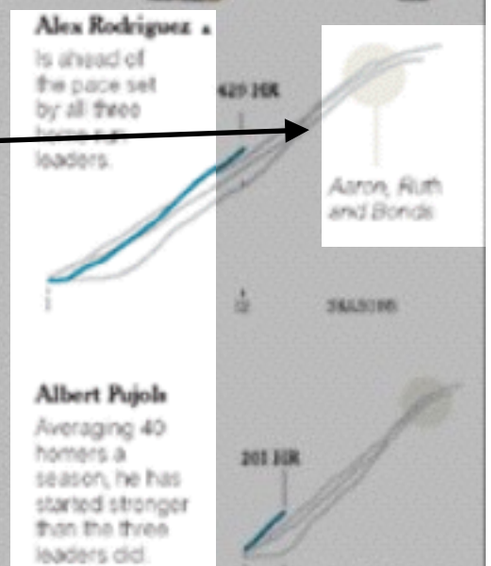
Note: Ages as of July 1 of each season.

Captions!!



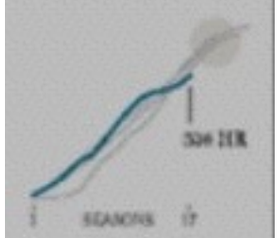
Others Taking Aim

Alex Rodriguez
Is ahead of the pace set by all three leaders.



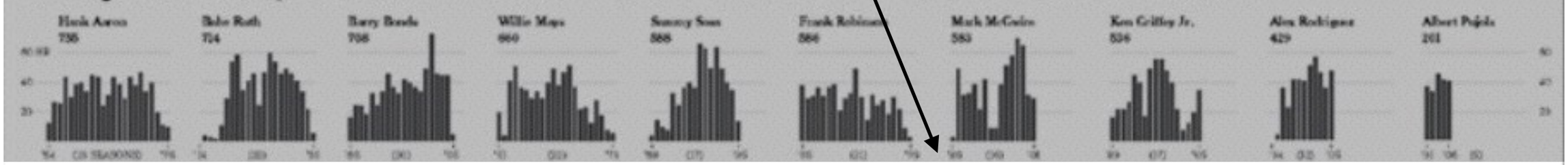
Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.

Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.



Differing Paths to the Top of the Charts

The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (3rd 257th).



16 times hit 30 or more (M.L. record).

Hit only 20 over first five seasons.

Averaged 52 from 2000 to 2004.

No one hit more from 1960-69.

Three 60-homer seasons is record.

Triple Crown in '66 (49, 122, .316).

First to hit 70 in a season.

Only McGwire had more in the 90's.

Youngest to reach 400 homers.

Second most ever in first five seasons.

E. Segel

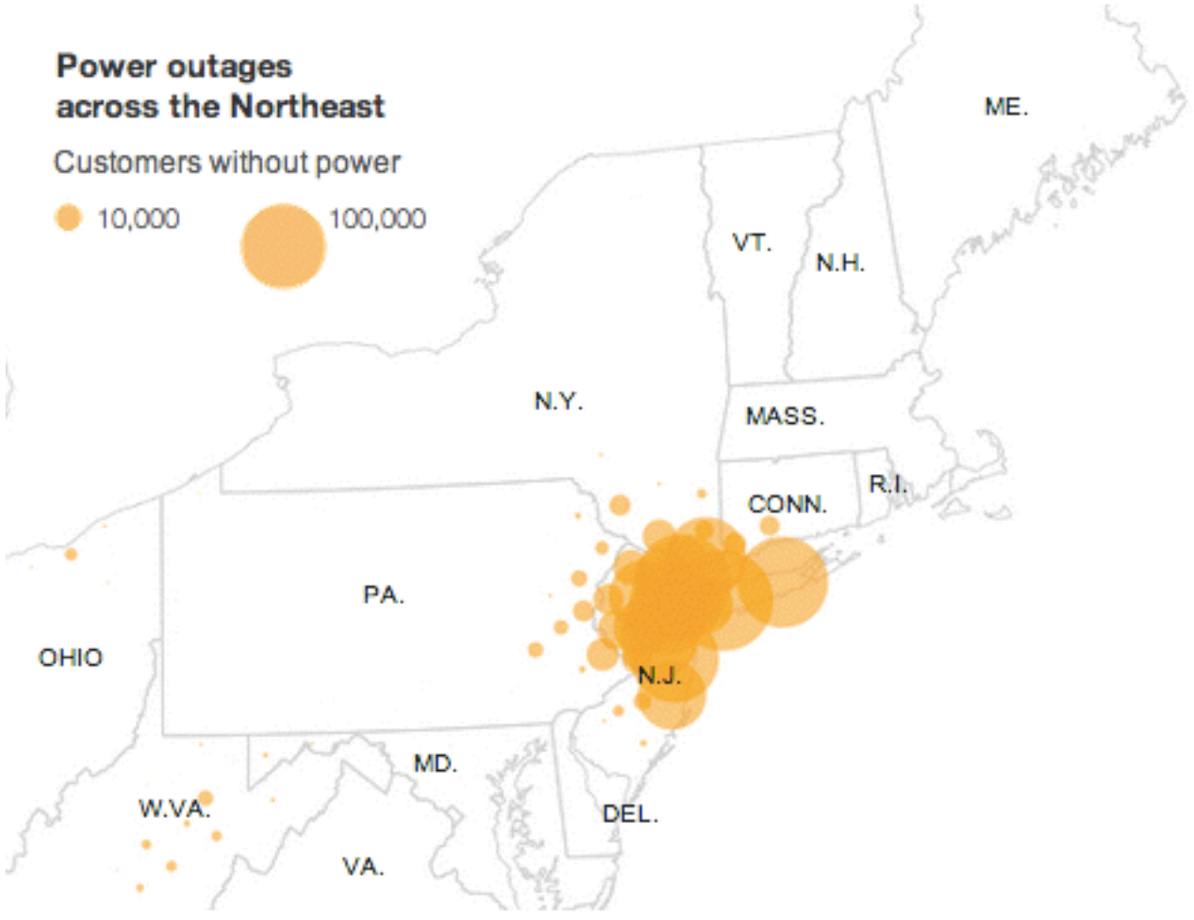
UPDATED November 1, 2012

FACEBOOK TWITTER GOOGLE+ E-MAIL SHARE

Where the Power Is Out and Returning Across the Northeast

Updated Sunday, November 4 at 8:00 PM

Hurricane Sandy felled trees, downed power lines and flooded substations. The storm led to power failures in at least 17 states. Here's the restoration status in areas with significant power failures.



TRI-STATE AREA

PSE&G

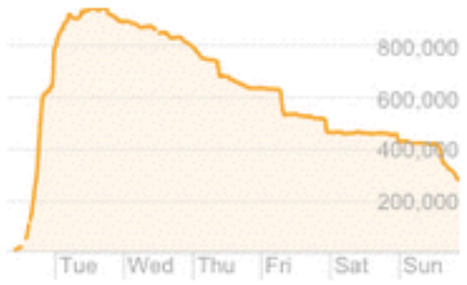
501,074 customers affected

Jersey Central Power & Light

405,816 customers affected

Long Island Power Authority

281,900 customers affected



NYT

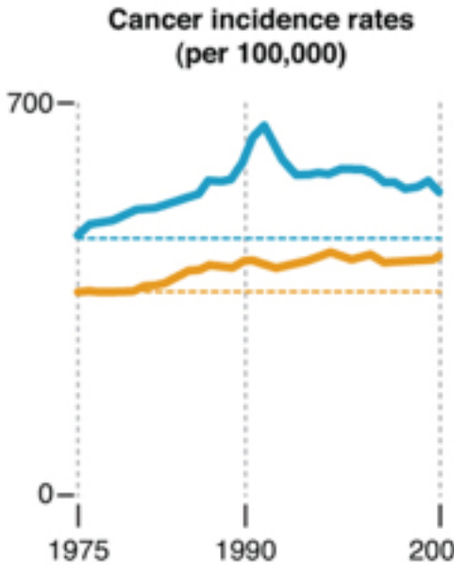
WHERE THERE'S SMOKE—THERE'S CANCER

Cancer rates are up, but mortality is down. New diagnostics and treatments are responsible for part of this trend. But the greatest single contributing factor is the decline in smoking—rates are at their lowest level in 50 years.

Men Women

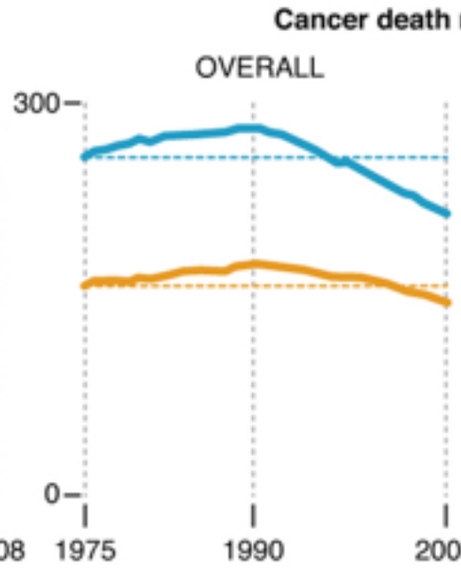
1 Increased incidence

An aging population contributes to rising incidence of cancer.



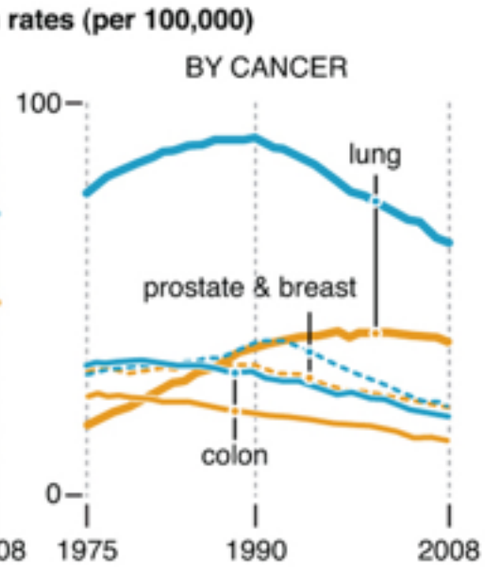
2 Fewer deaths

Cancer deaths have been dropping since 1991, especially in males.



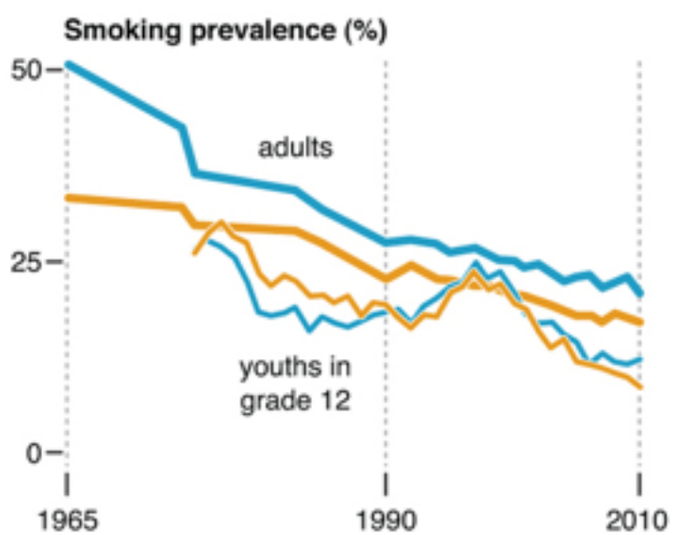
3 Decline of lung cancer

Drop in lung cancer deaths in males is the primary reason why death rates are down.



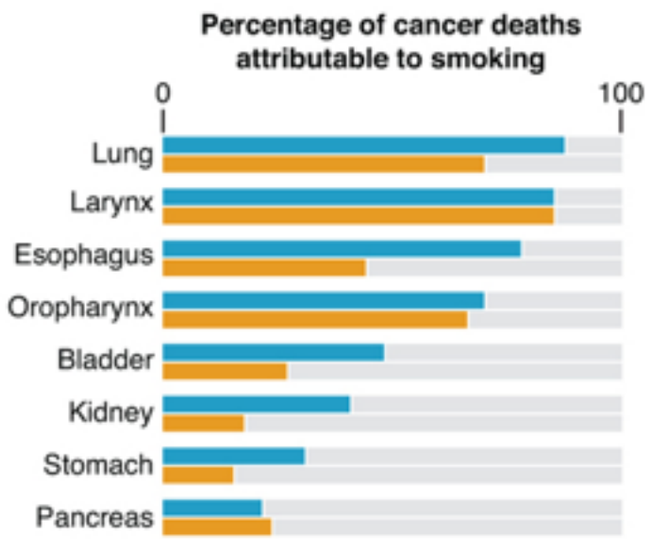
4 Decline in smoking

Since the 1964 first Surgeon General's report, smoking rates have been dropping. By 2010, the rate among males was down to 20%, from 50% at its peak. Among youths, rates have been on an even steeper decline since 1997.



5 Impact of smoking on cancer deaths

Smoking is a major risk factor for many types of cancer and significant contributor to cancer-related deaths. It remains the single largest preventable cause of disease and premature death in the US.



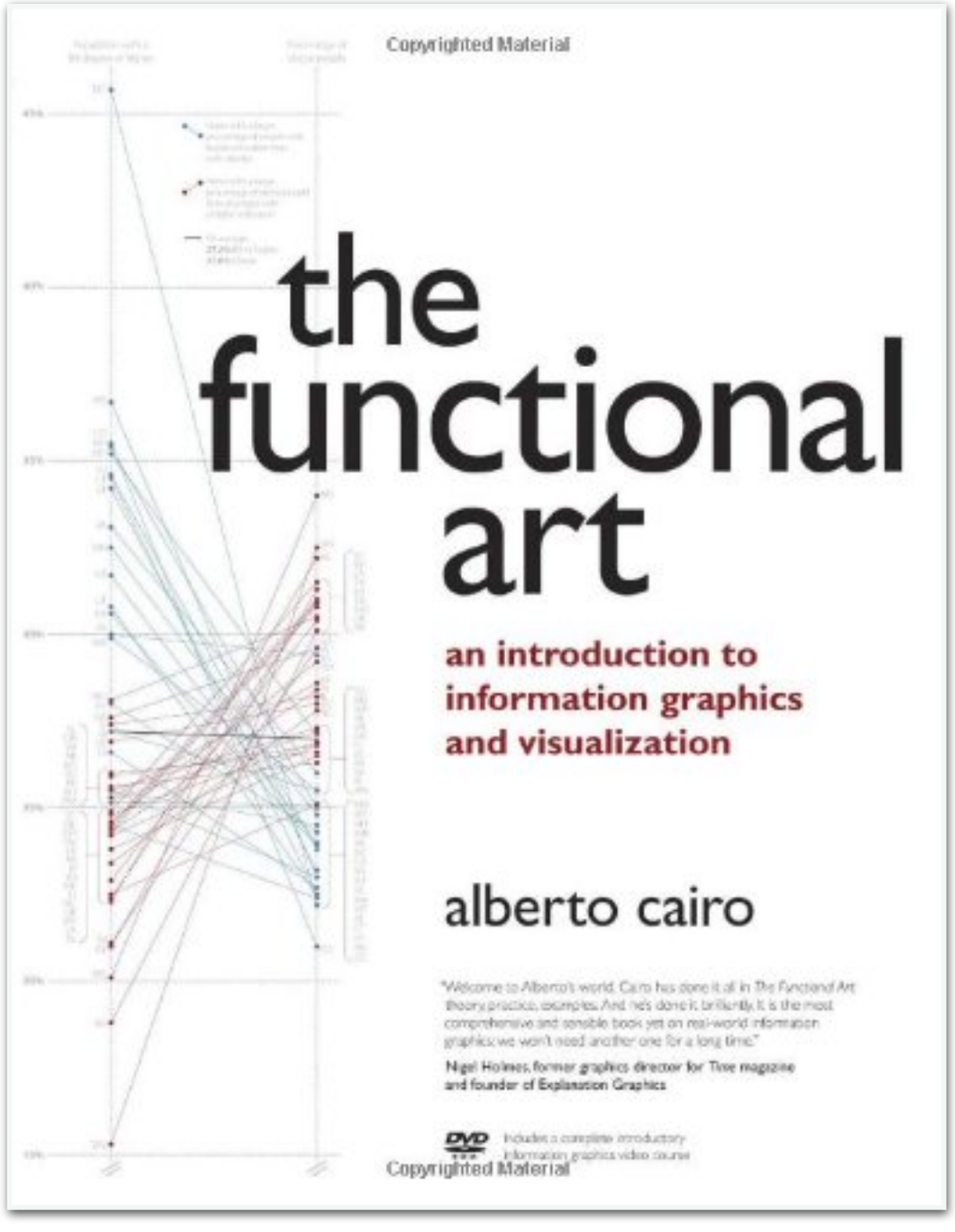
source: American Cancer Society Cancer Statistics 2012; Monitoring the Future (University of Michigan).

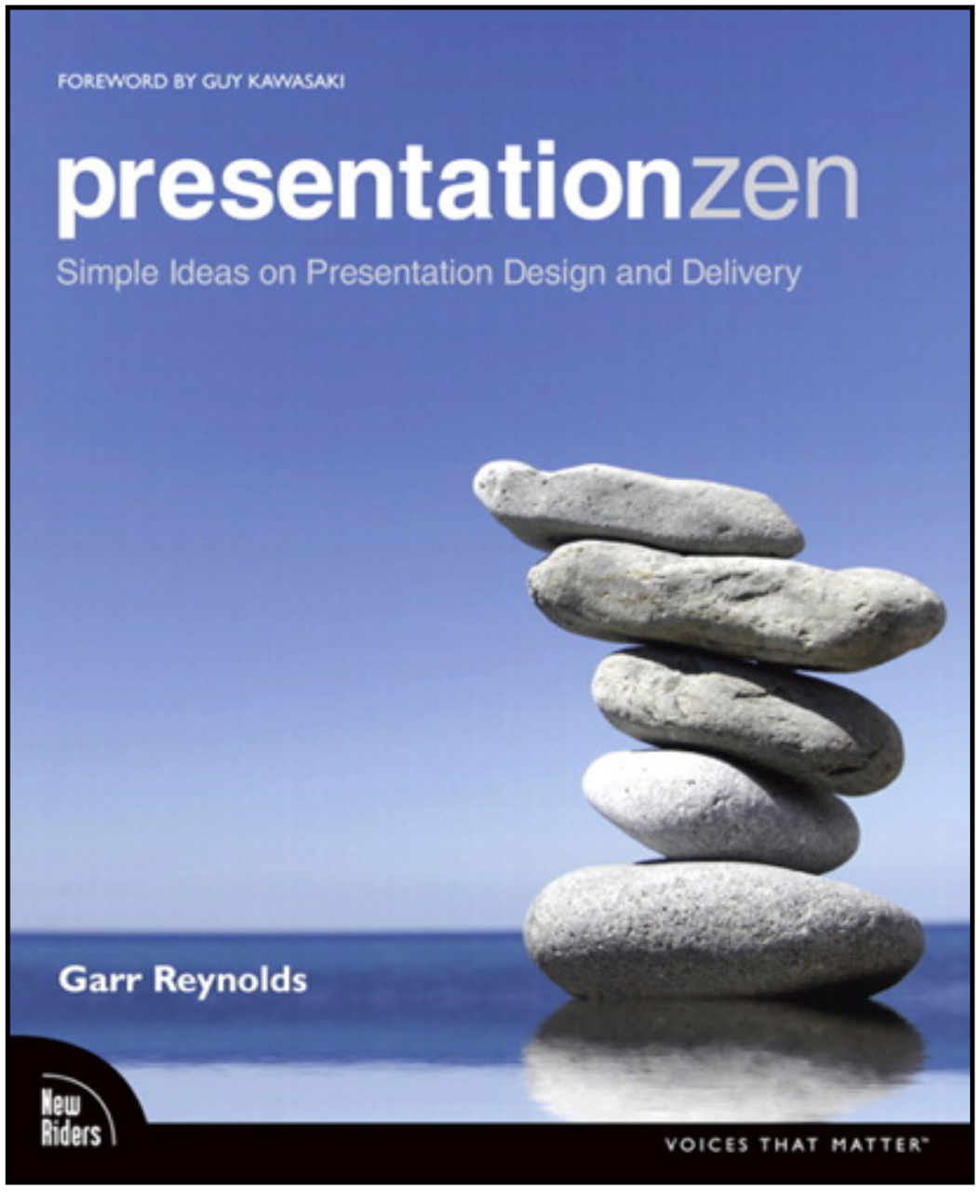
M. Krzywinski & A. Cairo

Successful Data Stories…

- …target the audience
- …engage and are memorable
- …answer concise questions
- …are carefully designed
- …move us to want to change the world

Further Reading





Further Reading