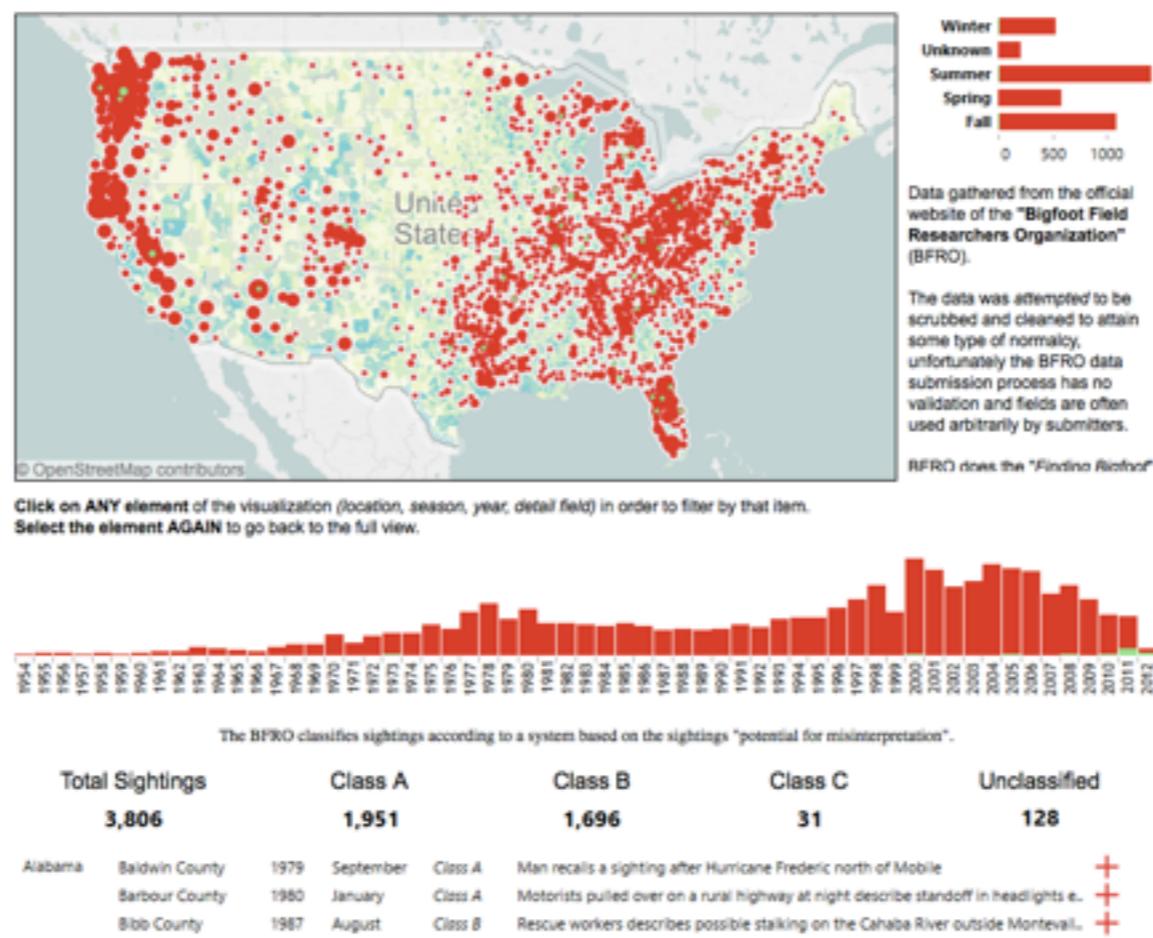


CS I 09/Stat I 2I/AC209/E- I 09

Data Science Communication and Storytelling

Hanspeter Pfister, Joe Blitzstein, and Verena Kaynig



This Week

- HWI 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 repository.

Two Fundamental Questions

I.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: $n_\varepsilon \rightarrow 0$ as $\varepsilon \rightarrow \infty$.

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 (URF4L), 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.

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

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)

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?
- What are your major insights and surprises?
- What change do you want to affect?

Know Your Audience



People you don't know are difficult to influence

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”?

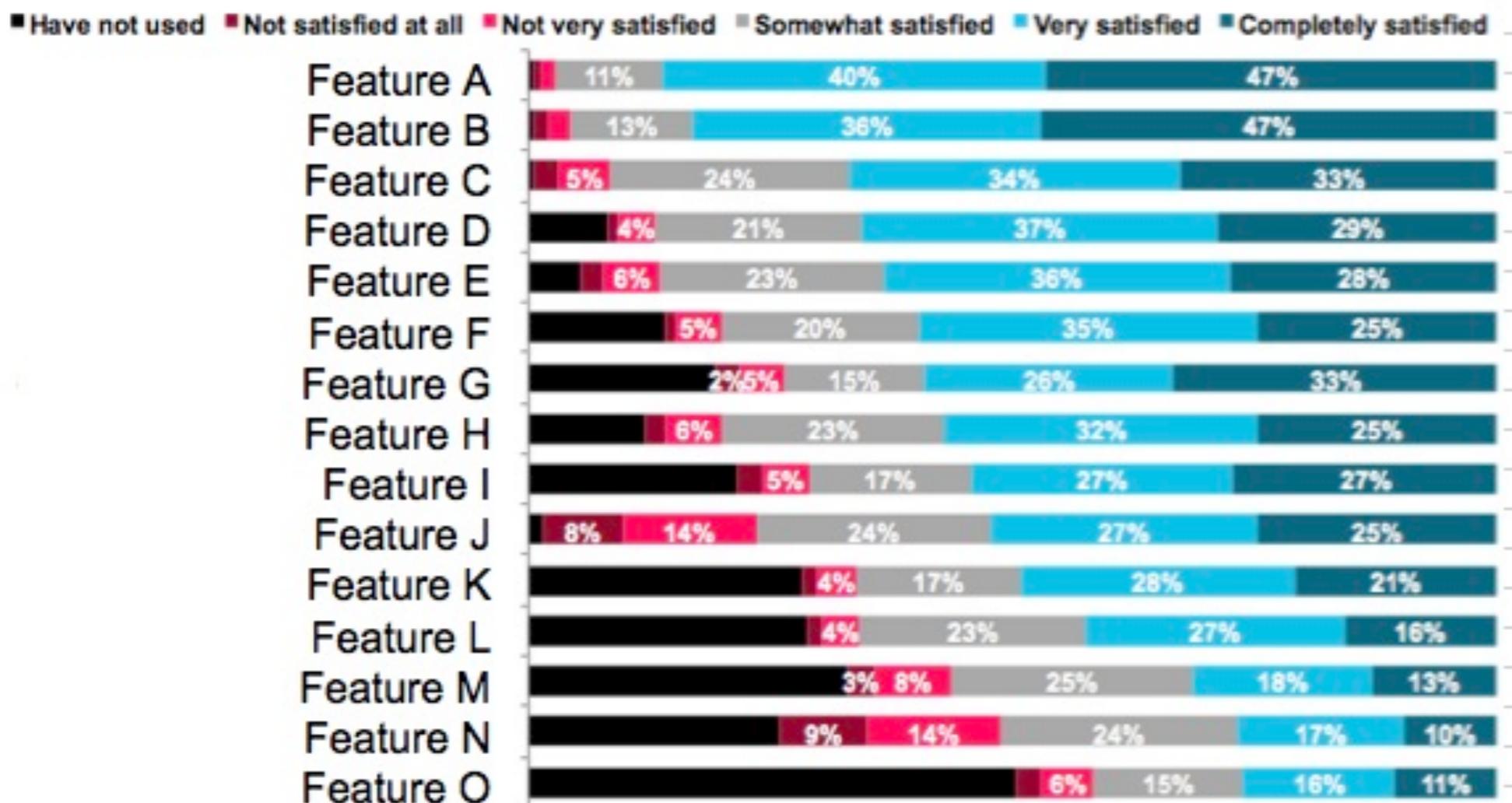
Don't Make Them Think!

- 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

How satisfied have you been with each of these features?

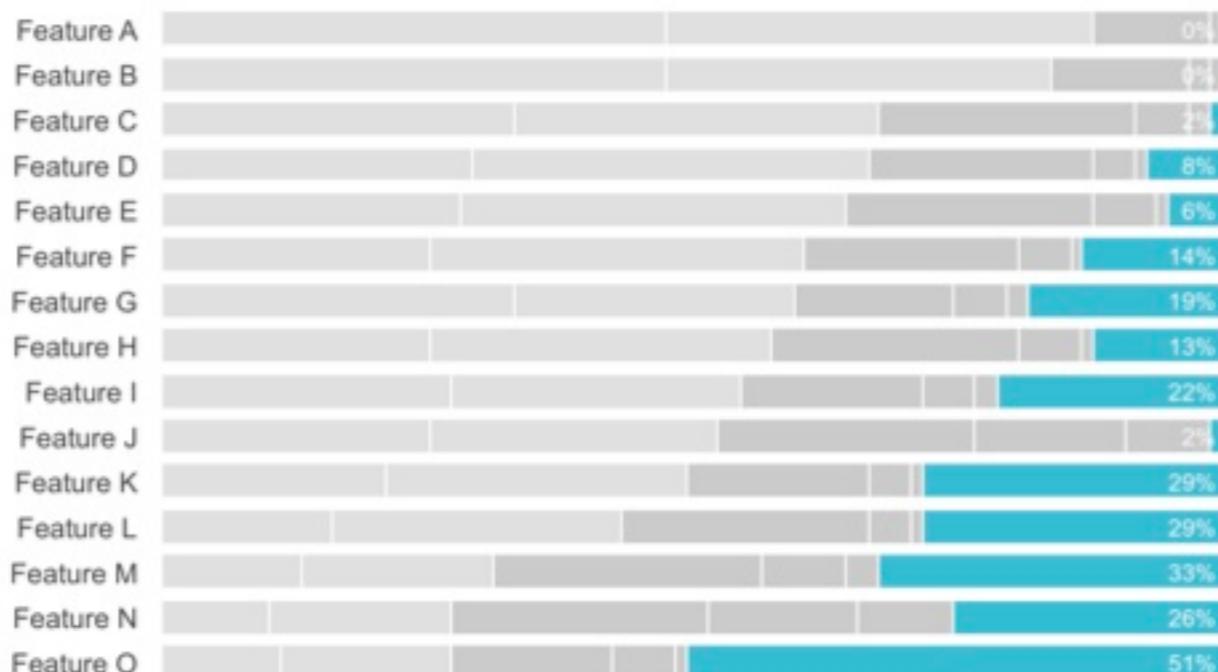


Don't Bury the Lead

User satisfaction varies greatly by feature

Product X User Satisfaction: Features

* Completely satisfied * Very satisfied * Somewhat satisfied * Not very satisfied * Not satisfied at all * Have not used



Feature O is least-used feature; what steps can we proactively take with existing users to increase use?

David Jacopille



V

THE BIGFOOT FIELD RESEARCHERS ORGANIZATION

Founded in 1995 -- The only scientific research organization exploring the bigfoot/sasquatch mystery.

Click on ANY image to Select the entire image

1954
1955
1956
1957
1958

Total sightings: 3,8

Alabama
Ba
Ba
Bit

Chilton County 2000 May Class A Daylight sighting while riding ATV on his private property outside Verbena

TO REPORT A SIGHTING

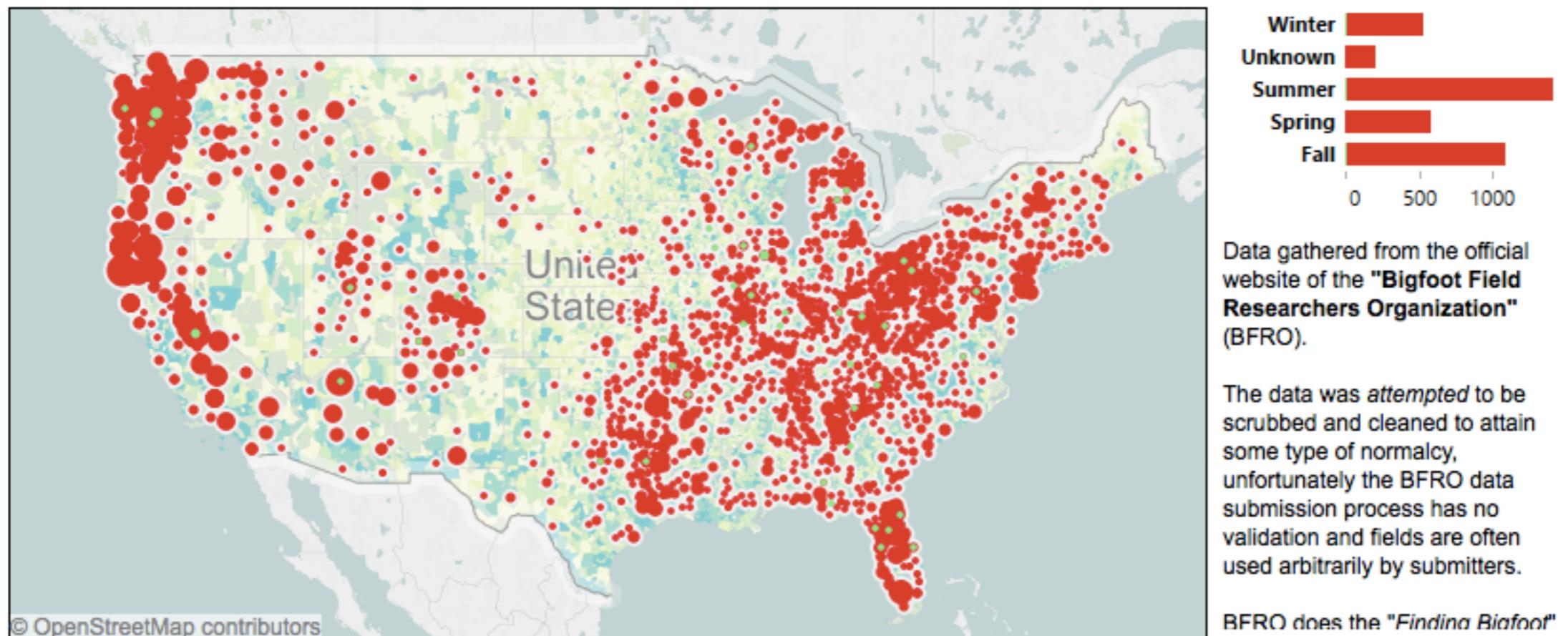
Florida BFRO @FL_BFRO 11 Sep
MA class A! Seen on Saturday! @MIBFRO
@NewJersey_BFRO @BFRO_Updates
bfro.net/GDB/show_repor...
Retweeted by BFRO UPDATES
Expand

2010
2011
2012

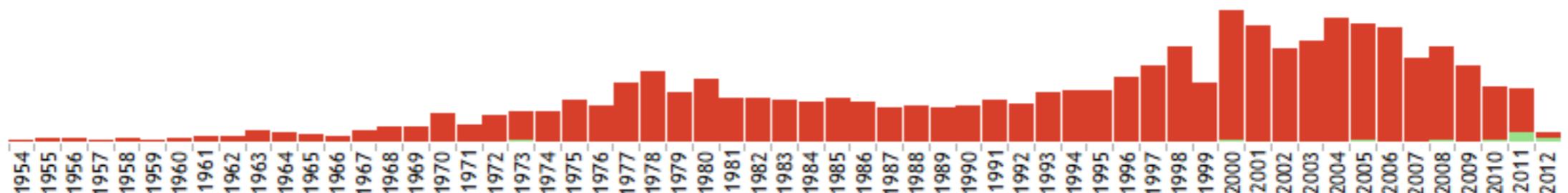
?

Ryan Robitaille

Where is Bigfoot seen in the USA?



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.

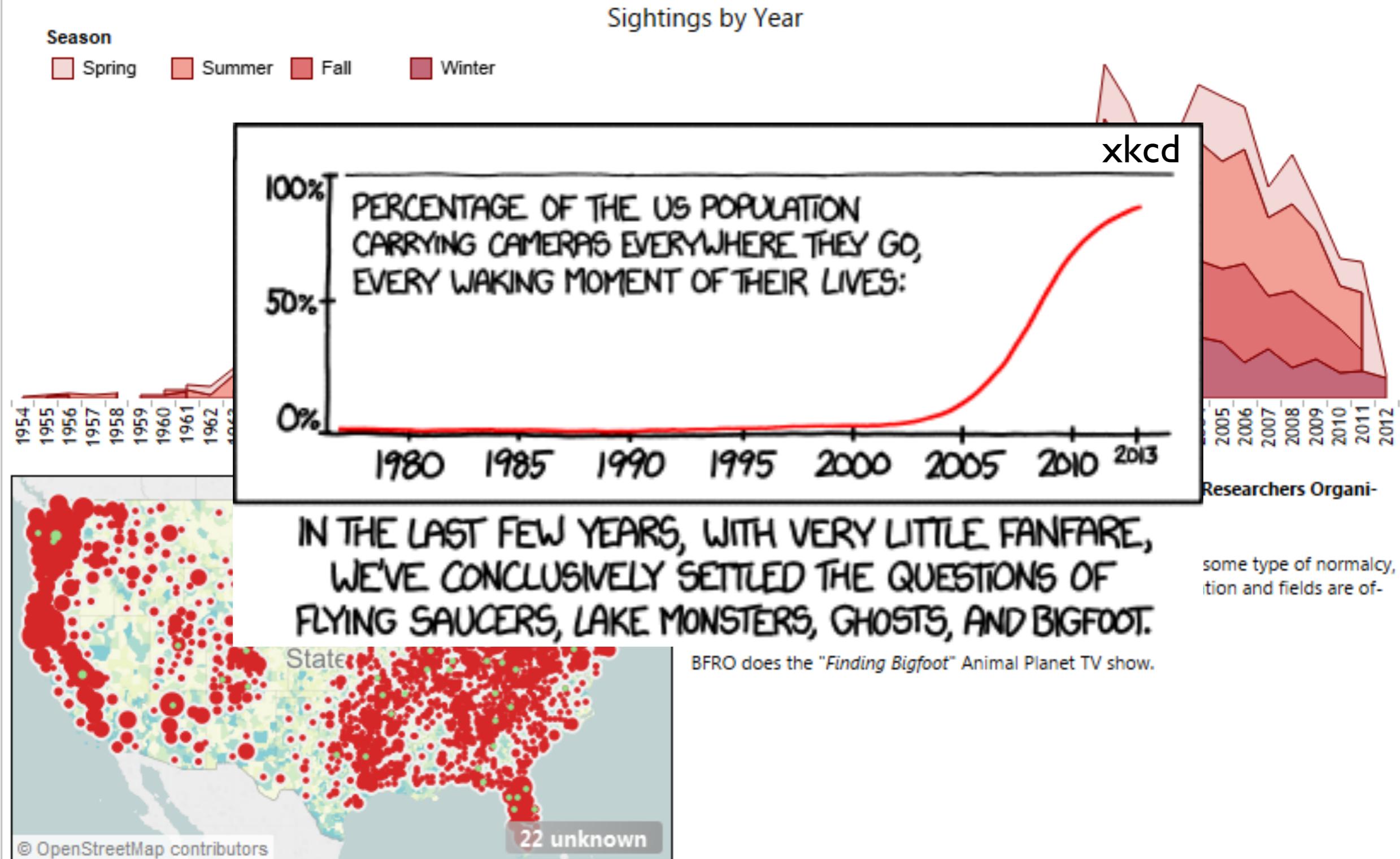


The BFRO classifies sightings according to a system based on the sightings "potential for misinterpretation".

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 Monteval..	+

Bigfoot sightings are in decline



Ryan Robitaille

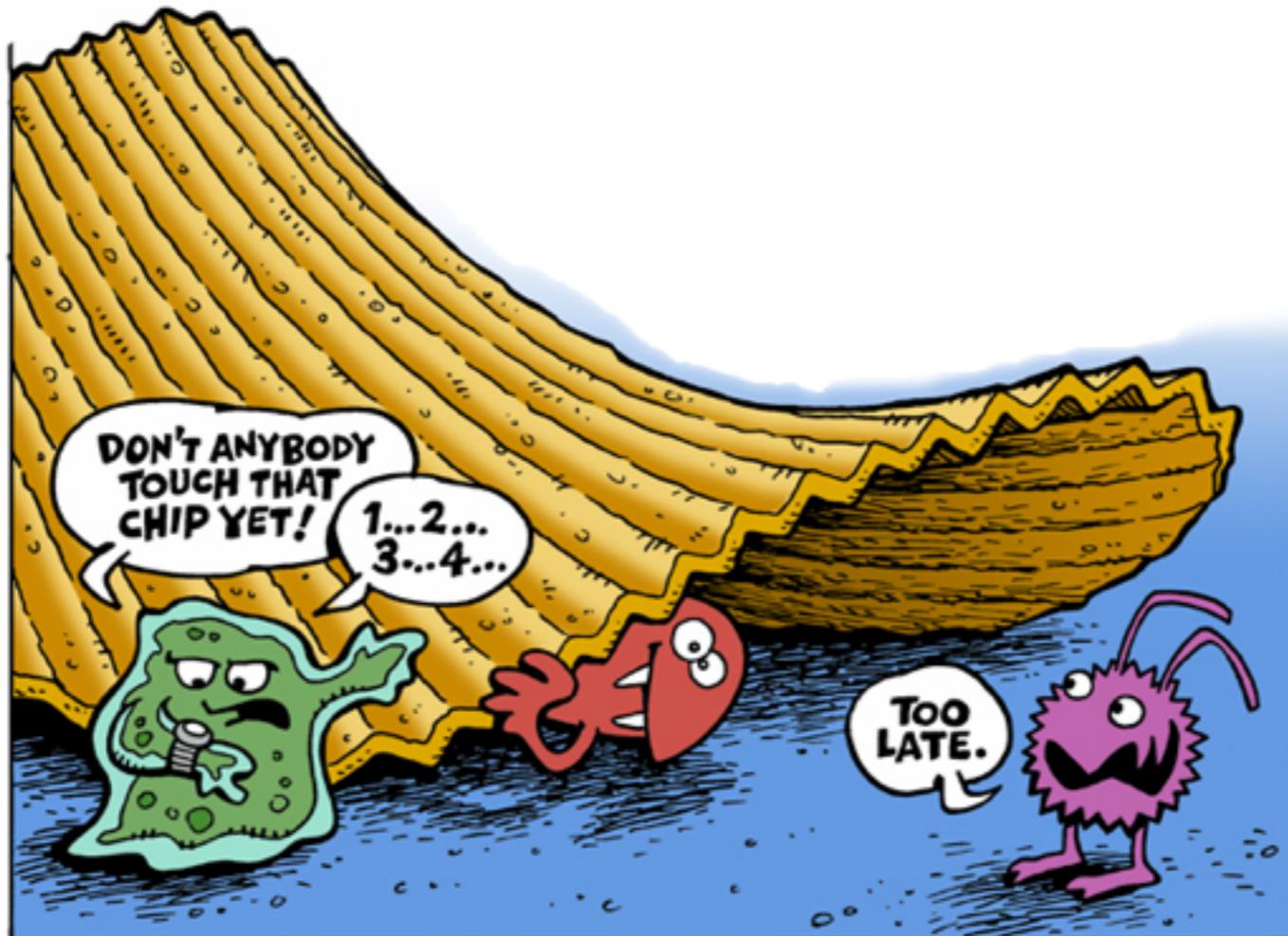
Unexpectedness

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

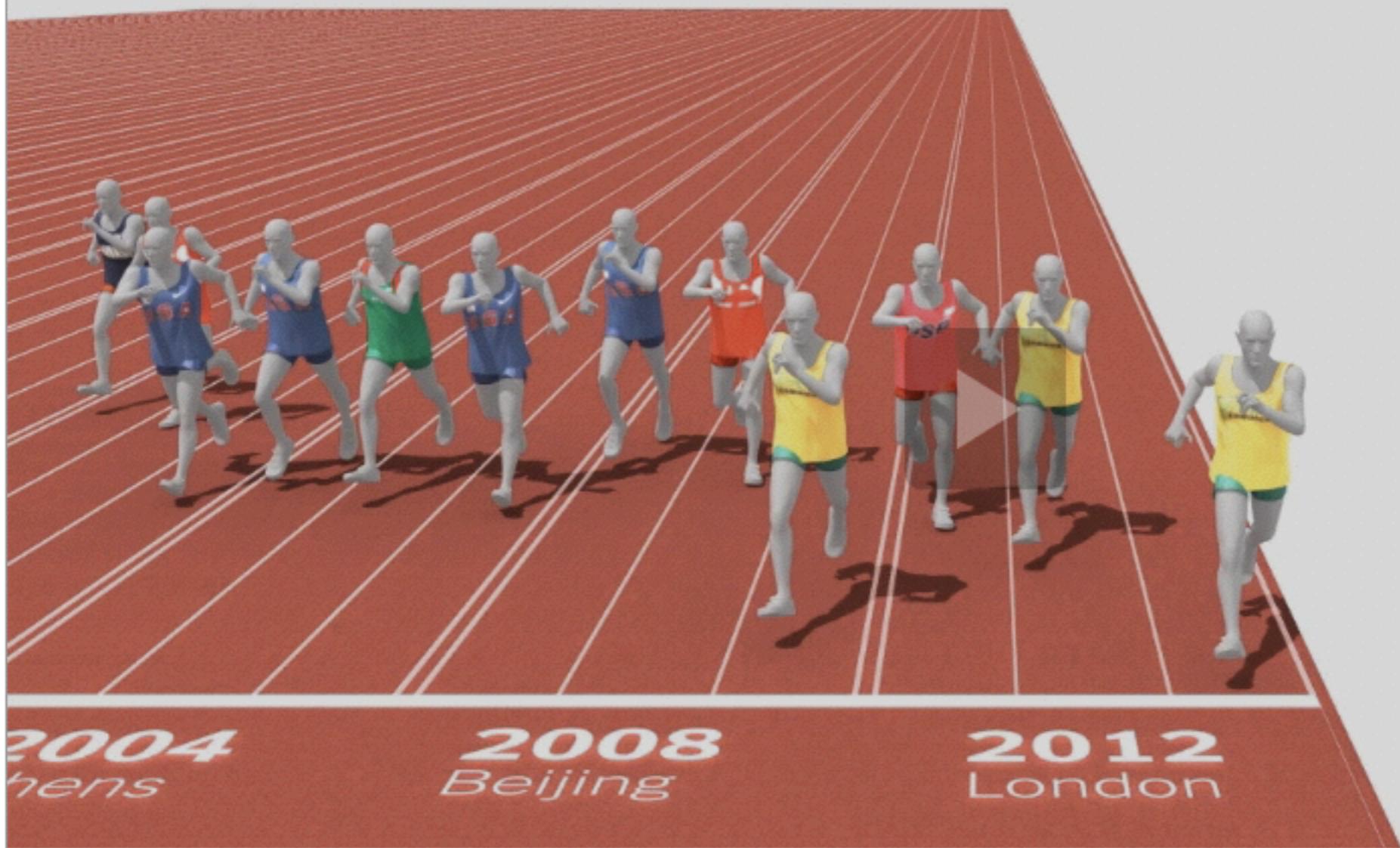
Use surprise to grab the audience's attention

“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 Association; Photographs: Chang W. Lee/The New York Times, Getty Images, International Olympic Committee

[FACEBOOK](#) [TWITTER](#) [GOOGLE+](#) [E-MAIL](#) [SHARE](#)

Messaging

Framing - Why should I care?

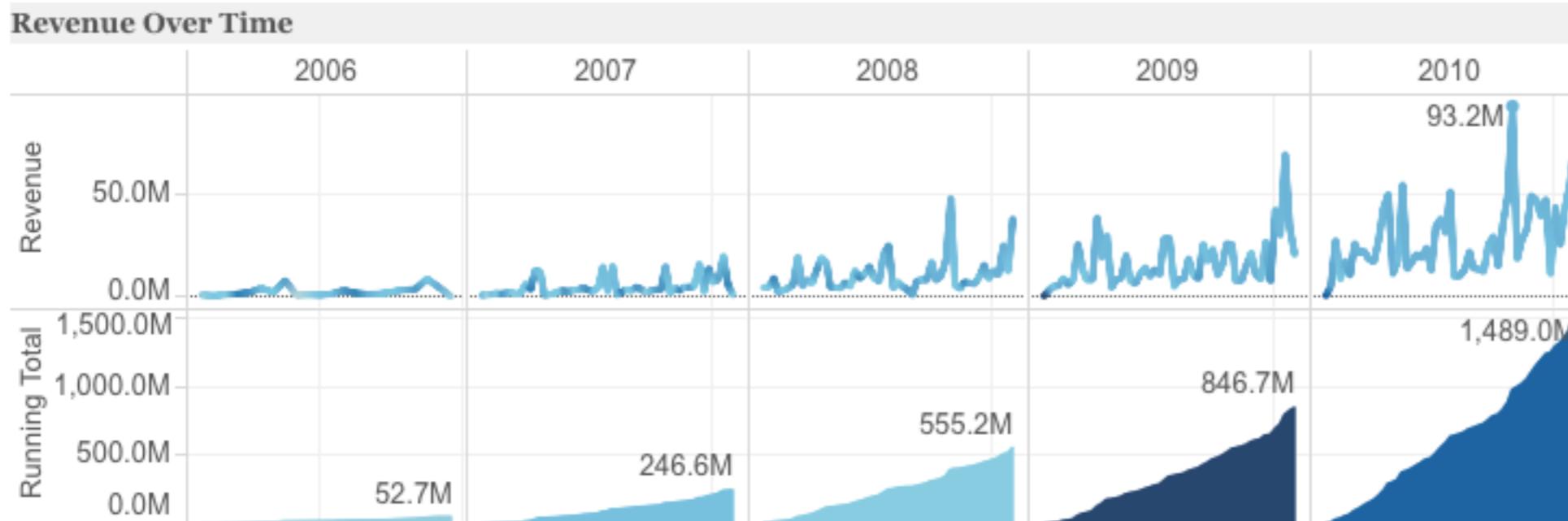
- Tell the audience: “Here is the right way to think about the problem I was trying to solve.”
- Catch the audience’s attention and frame the story using captions and annotations
- If done well, your insights will seem obvious given this framing. And that’s a good thing!



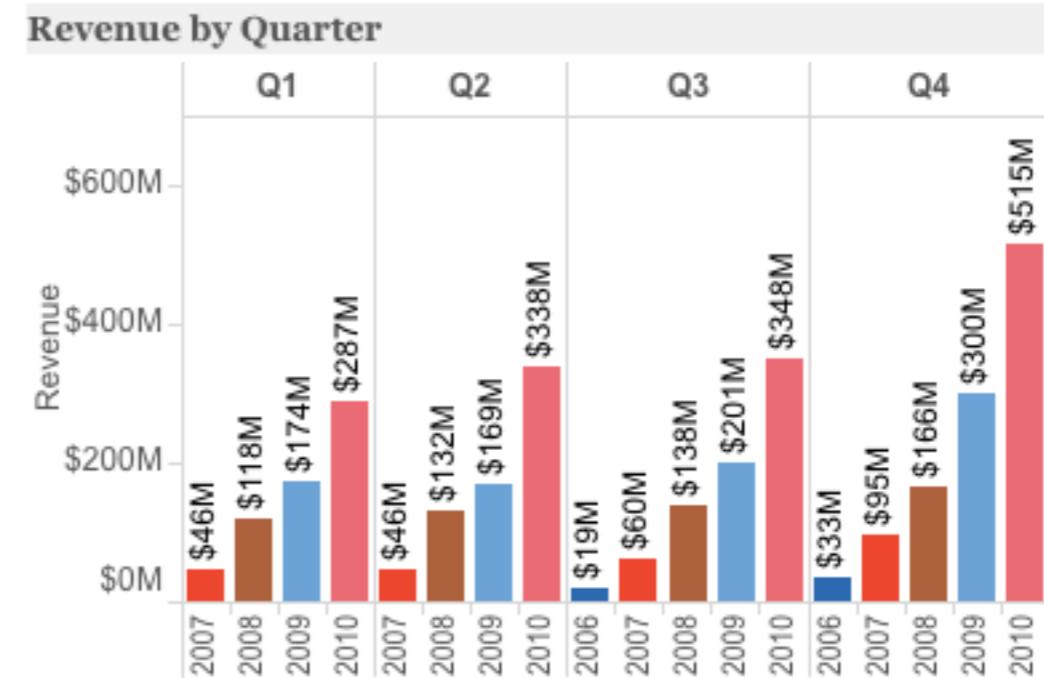
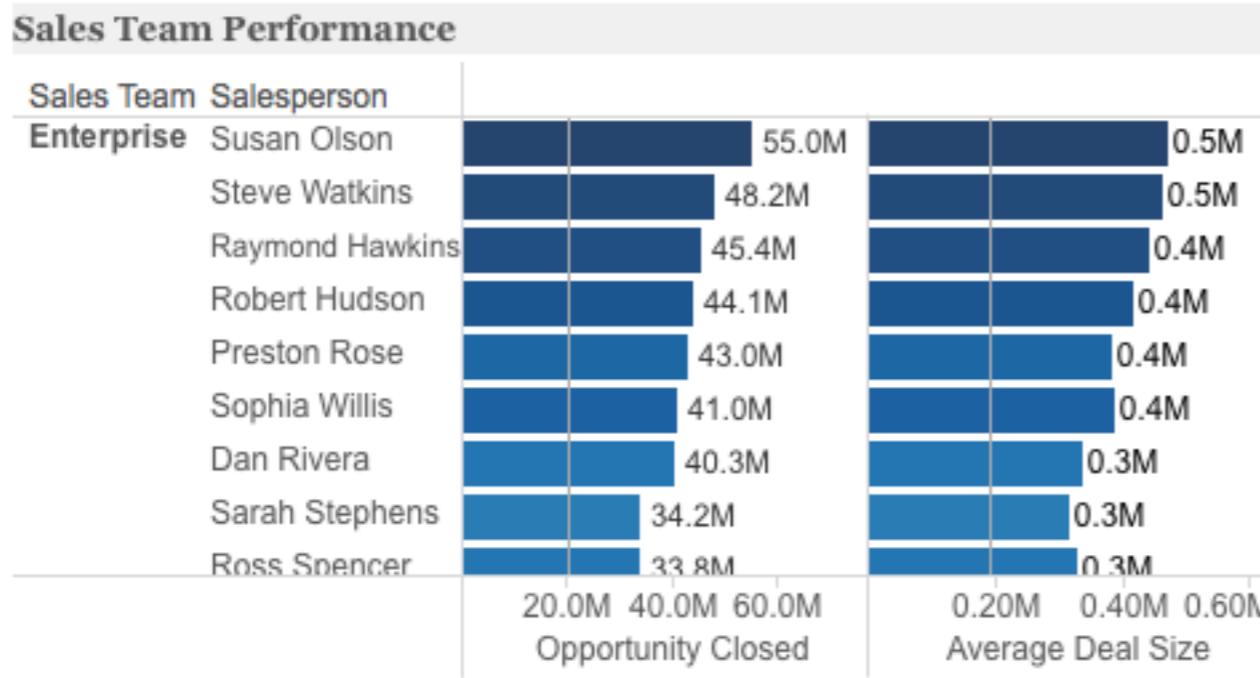
[Opportunity Dashboard](#)[Sales Dashboard](#)

Sales Dashboard

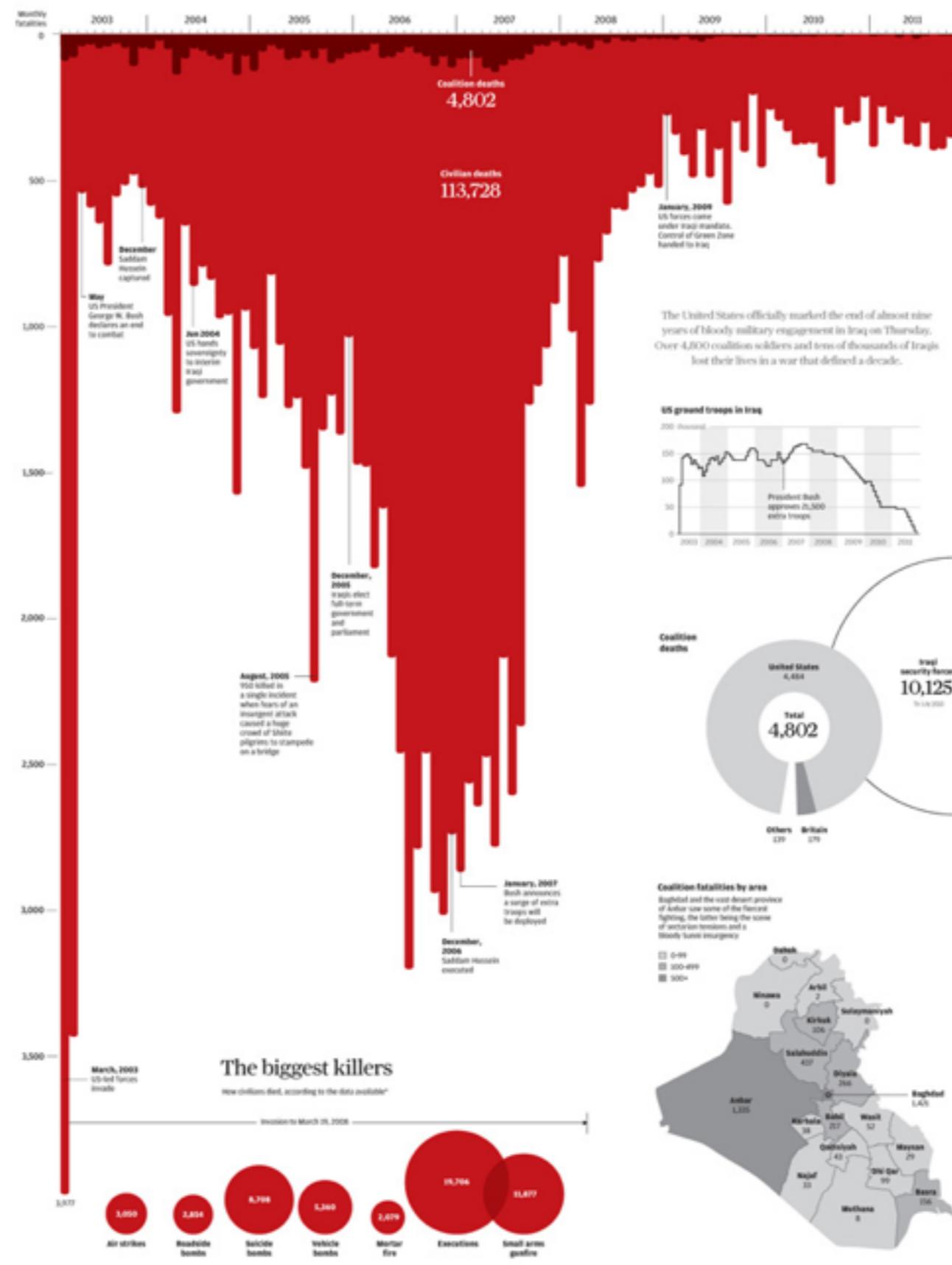
Total Sales	Number of Deals	Avg Deal Size	Rev. per Salesperson
\$3,190.2M	16,610	\$189,545	\$20.1M



Date Closed	8/7/2006	12/31/2010
<input type="text"/>	<input type="text"/>	<input type="text"/>
Region	(All)	<input type="button" value="▼"/>
Country	(All)	<input type="button" value="▼"/>
Sales Team	(All)	<input type="radio"/>
Small and Midmarket	<input type="radio"/>	<input type="radio"/>
Enterprise	<input type="radio"/>	<input type="radio"/>
Avg Deal Size/Salespe...	\$130,922	\$336,519



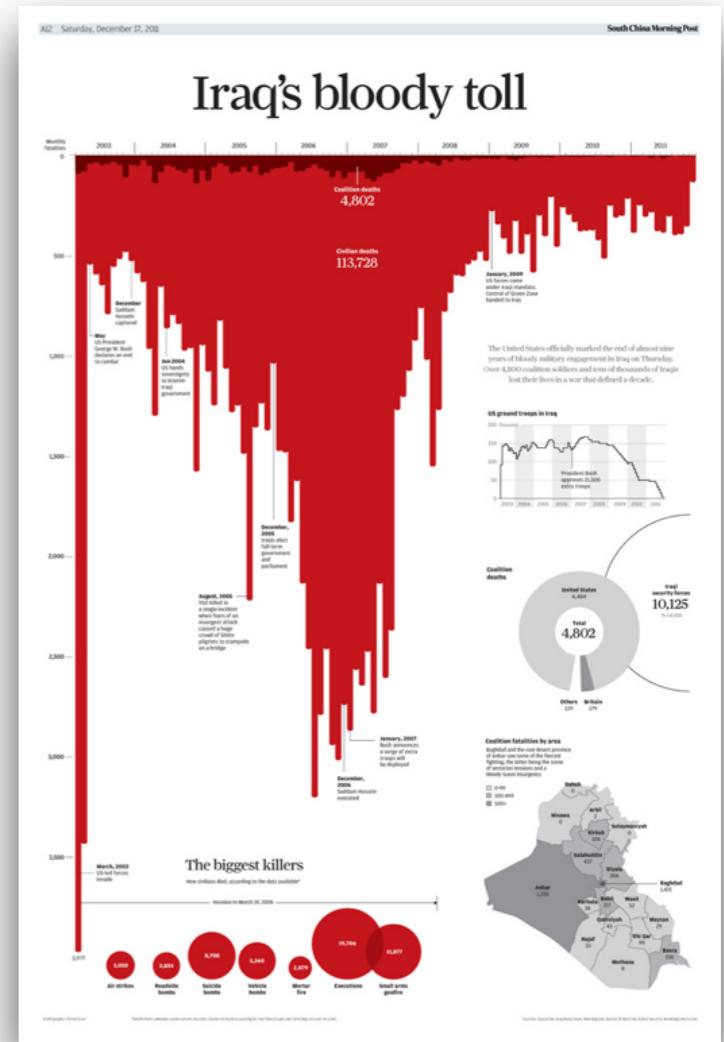
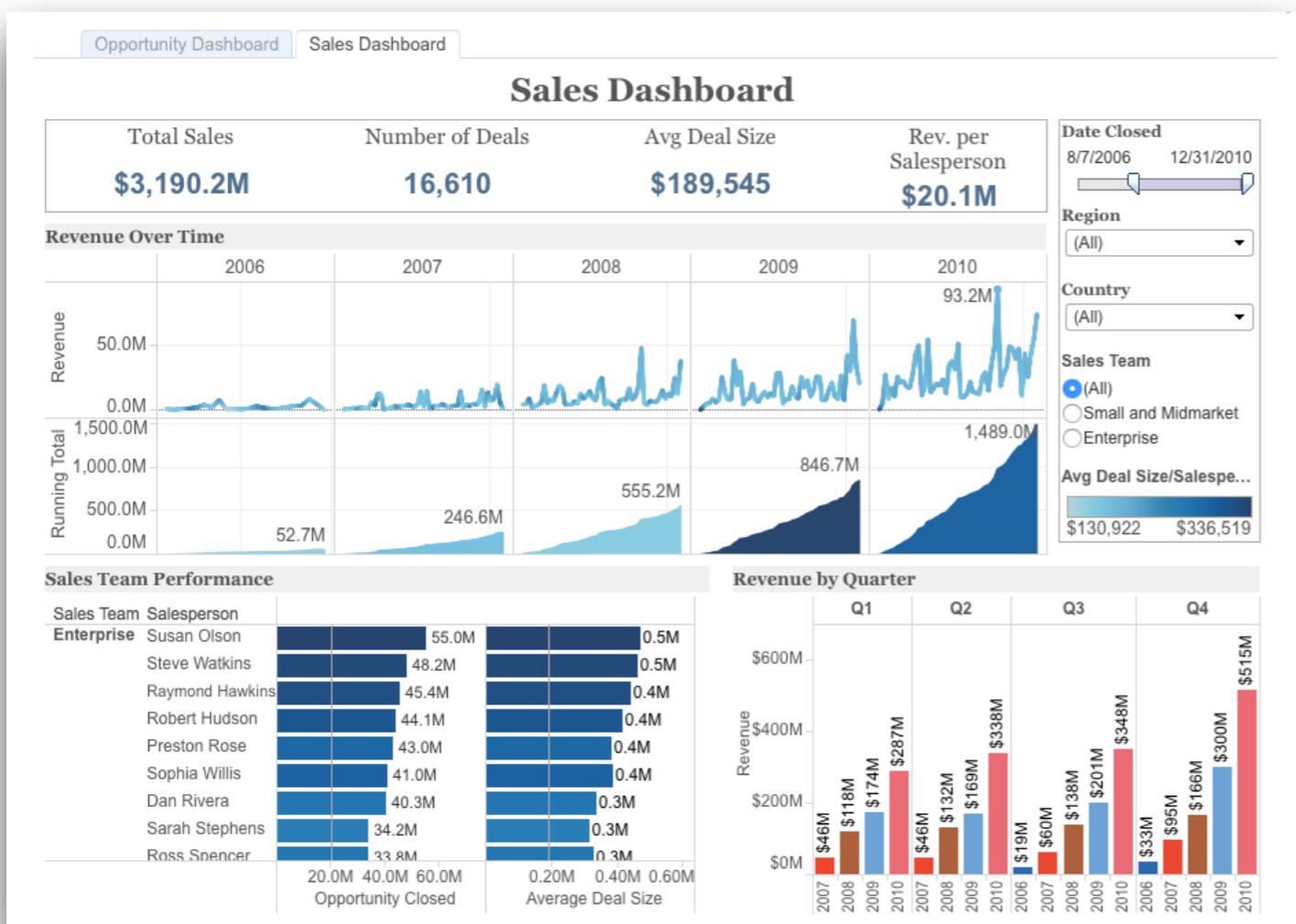
Iraq's bloody toll



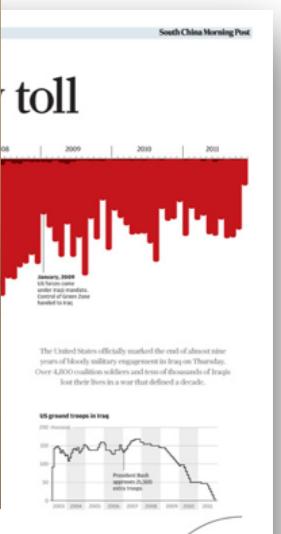
What is the message?

Exploratory
Neutral

Explanatory
Opinionated

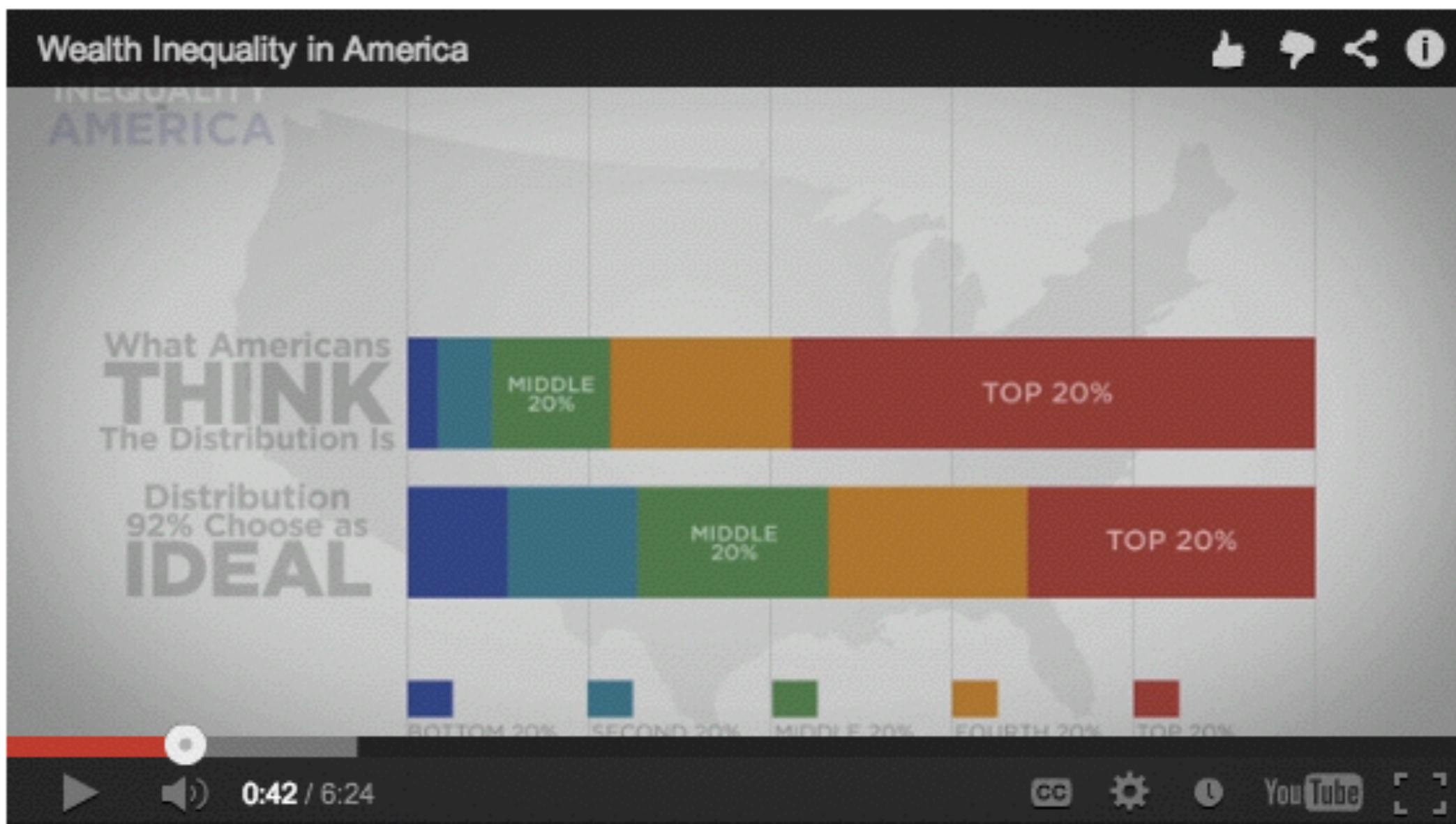


Know Your Audience



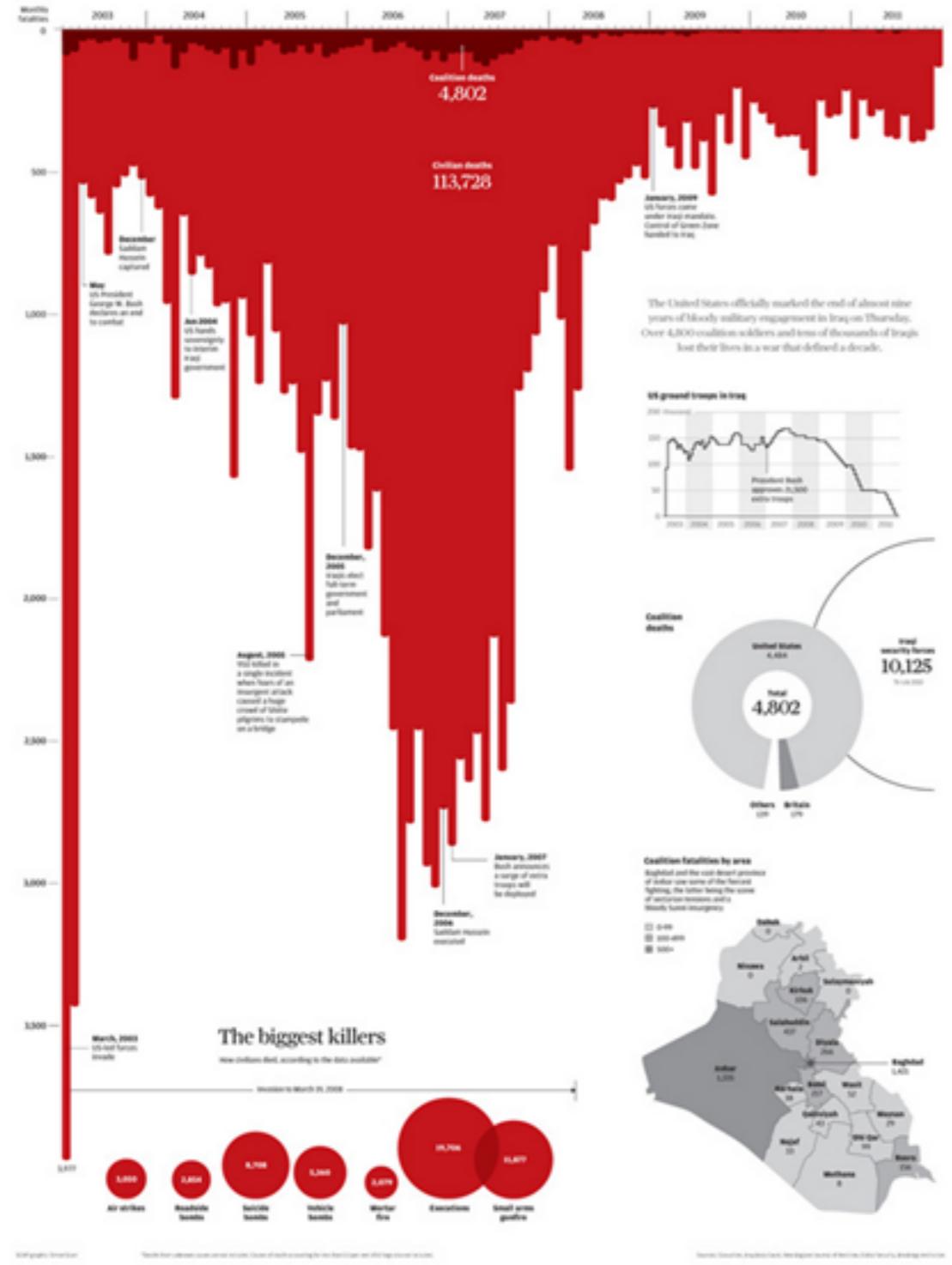
Wealth Inequality in America

RANDY | MONDAY, MARCH 11, 2013 AT 8:08AM [PERMALINK](#)

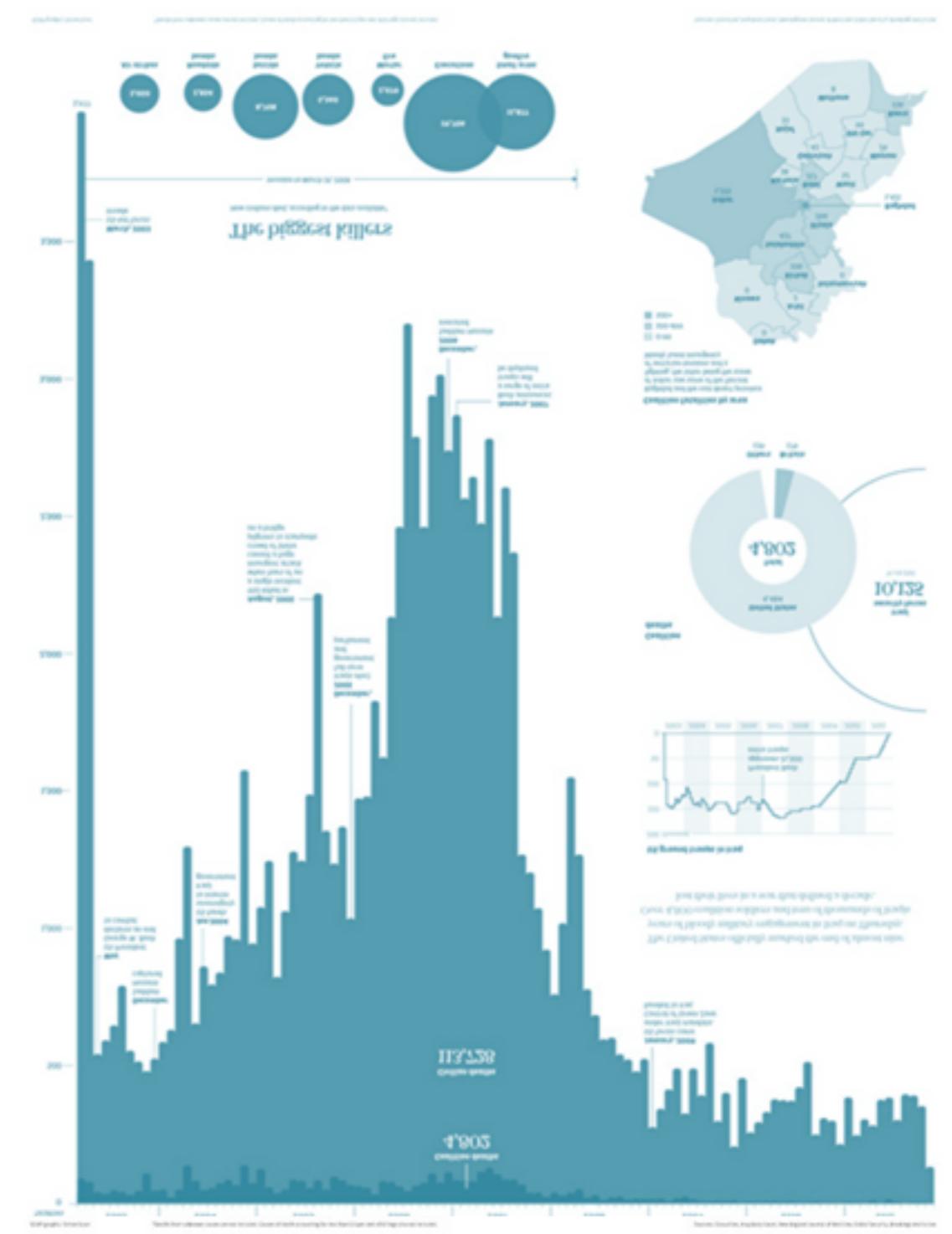


Visual Story Design

Iraq's bloody toll



Iraq: Deaths on the Decline



Andy Cotgreave, Tableau

755



Steroids or Not, the Pursuit Is On

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

Lines are cumulative home runs.

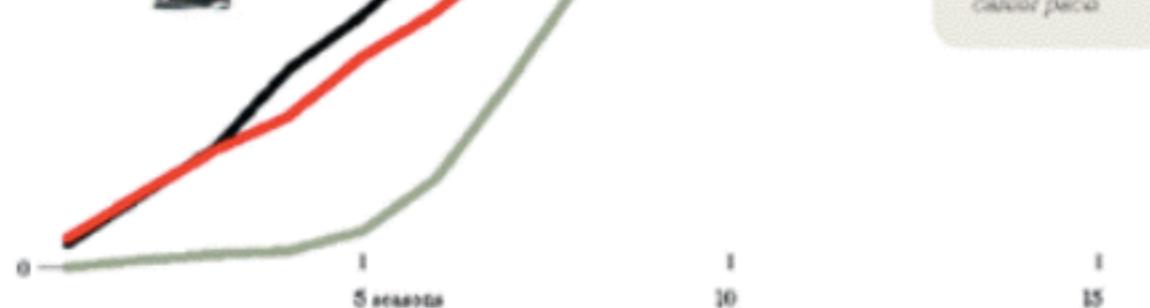
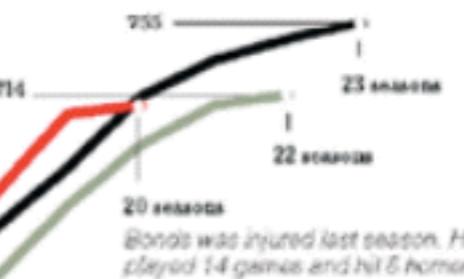


Bonds takes lead
Home runs:
after 16 seasons
Bonds 587
Aaron 554
Ruth 516

600

14th season

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.

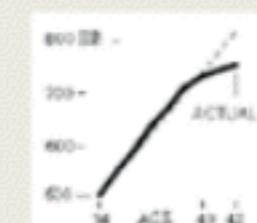


Homer Pace After Age 34

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.

PROJECTION BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

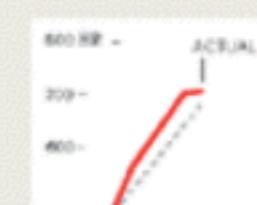
Aaron
Actual homers slightly outpace projected homers for five seasons



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

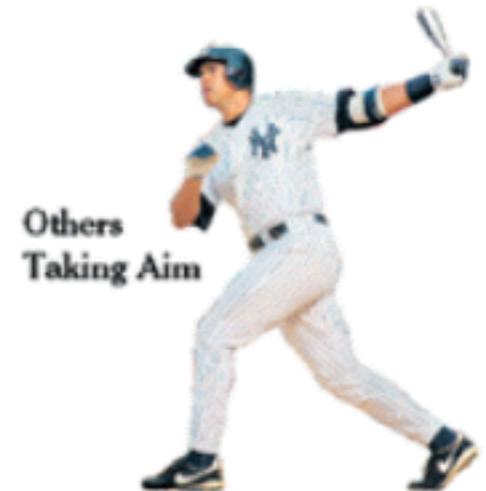


Bonds
From age 35 to 39, he averaged 14 more homers a season than projected



Note: Ages as of July 1 of each season

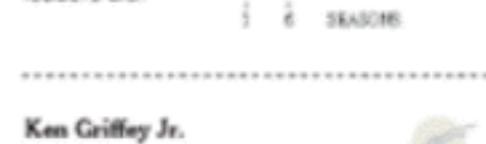
Others Taking Aim



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.

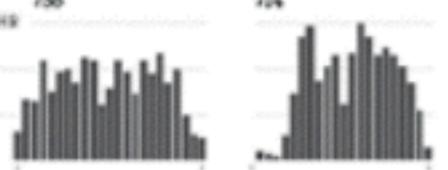


Differing Paths to the Top of the Charts

The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (56th).

Hank Aaron

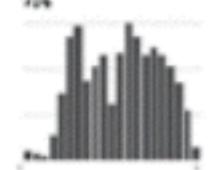
755



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

Babe Ruth

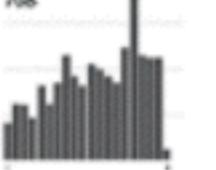
714



Hit only 20 over first five seasons.

Barry Bonds

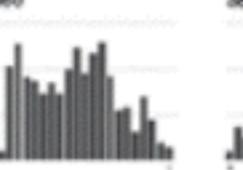
708



Averaged 52 from 2000 to 2004.

Willie Mays

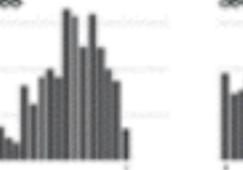
660



No one hit more than 1950-69.

Sunny Sosa

588



Three 60-homer seasons is record.

Frank Robinson

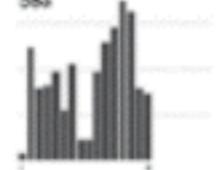
586



Triple Crown in '66 (49, 122, .316). First to hit 70 in a season.

Mark McGwire

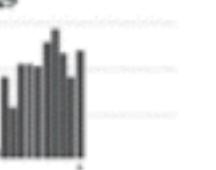
583



Only McGwire had more in the 90's. Youngest to reach 400 homers.

Ken Griffey Jr.

536



Youngest to reach 400 homers. Second most ever in first five seasons.

Source: The New York Yankees Yearbook

E. Segel

755

Steroids or Not, the Pursuit Is On

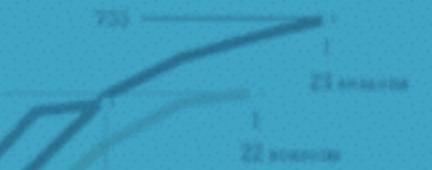
Every Bonds is being won at the career home run record. His record is only six more to tie Babe Ruth and 47 to equal Hank Aaron.

It's a remarkable home run



BEGINNING

According to accusations in a book about steroids, he began using steroids before the 1993 season. He hit 439 in the league two seasons after his 61 in 1994, then surpassing Aaron's career pace.



20 seasons
Bonds was around last season. He already had 14 games and had 6 home runs.

Homer Pace After Age 34

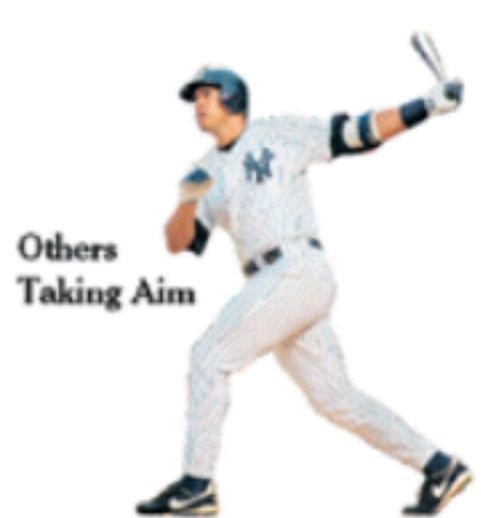
If the accusations are correct, Bonds was 24 in his first season on steroids. Here are projected home run paces for each player after age 34.

Accurately rates Bonds the fastest of previous five leaders.



Note: Ages as of July 1 of each season

Others Taking Aim



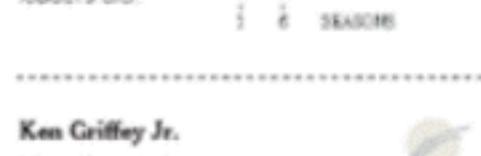
Alex Rodriguez

is ahead of the pace set by all three home run leaders.



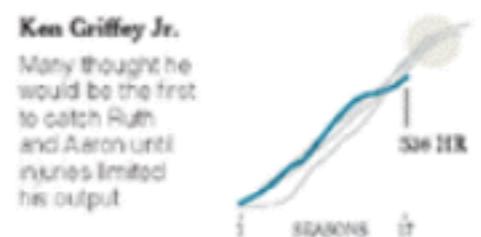
Albert Pujols

Averaging 40 homers a season from age 20 to 34, he has started stronger than the three leaders did.



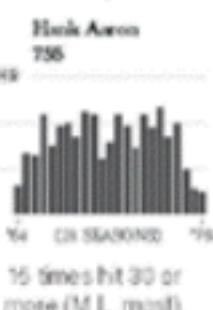
Ken Griffey Jr.

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

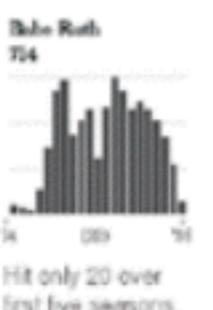


Differing Paths to the Top of the Charts

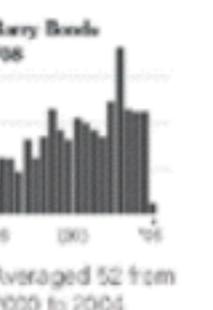
The top seven players on the career home run list, along with a look at Griffey (128), Rodriguez (37th) and Pujols (led 257th).



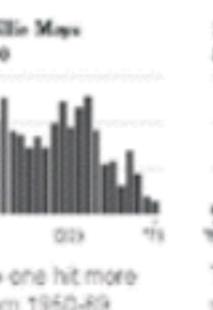
15 times hit 30 or more (M.L. most).



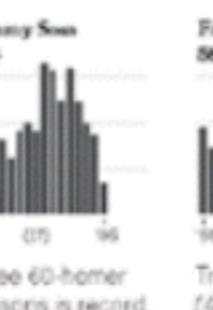
Hit only 20 over first five seasons.



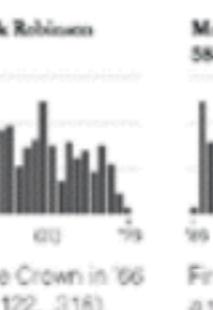
Averaged 52 from 2000 to 2004.



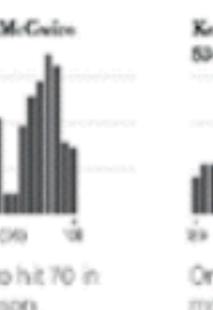
No one hit more from 1950-69.



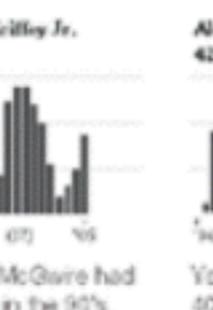
Three 60-homer seasons is record.



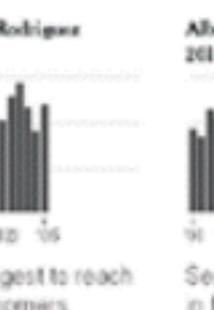
Triple Crown in '66 (49, 122, .316).



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.

Source: The 2005 World Series New York Times

E. Segel

755

Steroids or Not, the Pursuit Is On

Every Bonds hit seems like the closer home run is scored. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.

Line shows cumulative home runs.

Hank Aaron
755 home runs
23 seasons



Babe Ruth
714 home runs
22 seasons



Barry Bonds
709 home runs
20 seasons



BEGINNING

According to accusations in a grand jury about steroids, he began hitting records before the 1990 season. The 400 in the major leagues was passed when he hit 70 home runs, surpassing Aaron's career pace.

600

1400

15

10

5

0

5 seasons

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

00

05

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

755

Steroids or Not, the Pursuit Is On

Barry Bonds is racing toward the career home run record. Perhaps only six more to be above Ruth and 47 to beat Hank Aaron.

It's been an extraordinary career.

Hank Aaron
755 home runs
23 seasons



Babe Ruth
714 home runs
22 seasons



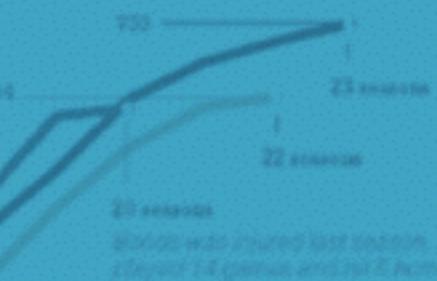
Barry Bonds
708 home runs
20 seasons



BEGINNING



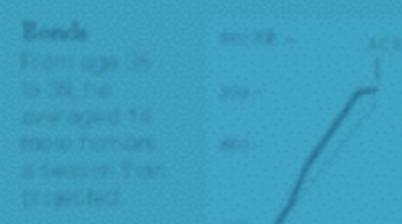
According to Allegations in a book about steroids, no player having entered before the 1990 season has averaged 50 or more home runs per season, let alone 567. (Aaron, however, surpassed 50 home runs in his first year.)



Homer Pace After Age 34

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

PROJECTED BASED ON HOME RUNS PER SEASON



NOTE: Ages as of July 5 of each season.

Others Taking Aim

Alex Rodriguez: 31 instead of 30 home runs. 429 total.

MIDDLE

Albert Pujols: 361 total.

Albert Pujols: Averaging 49 home runs 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 Bonds came along.

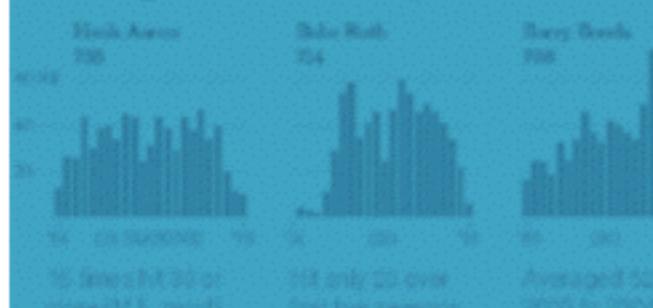
END

Diving deep into the career home run lists, along with a look at Griffey (123), Rodriguez (1078) and Pujols (632578).

Player	Total Home Runs	Seasons
Hank Aaron	755	23
Babe Ruth	714	22
Barry Bonds	708	20
Willie Mays	660	20
Swing Shift	585	20
Frank Robinson	580	20
Mark McGwire	566	20
Ken Griffey Jr.	556	20
Alex Rodriguez	429	16
Albert Pujols	361	16

Youngest to reach 400 home runs: Griffey (19); Second-most ever to hit 500: Bonds (20).

Differing Paths to the Top of the Charts



END

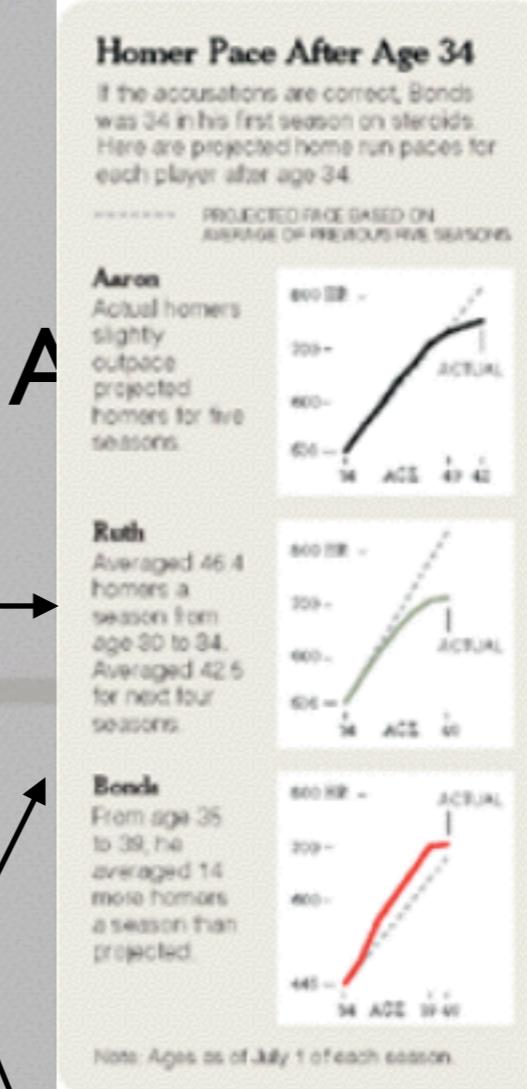
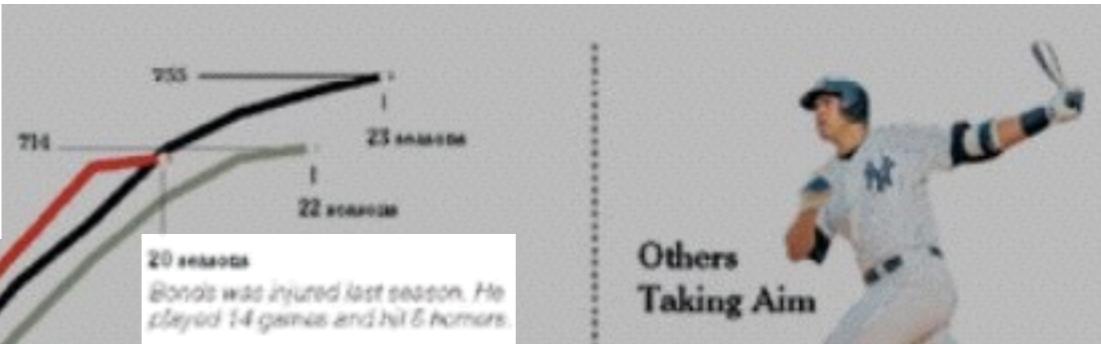
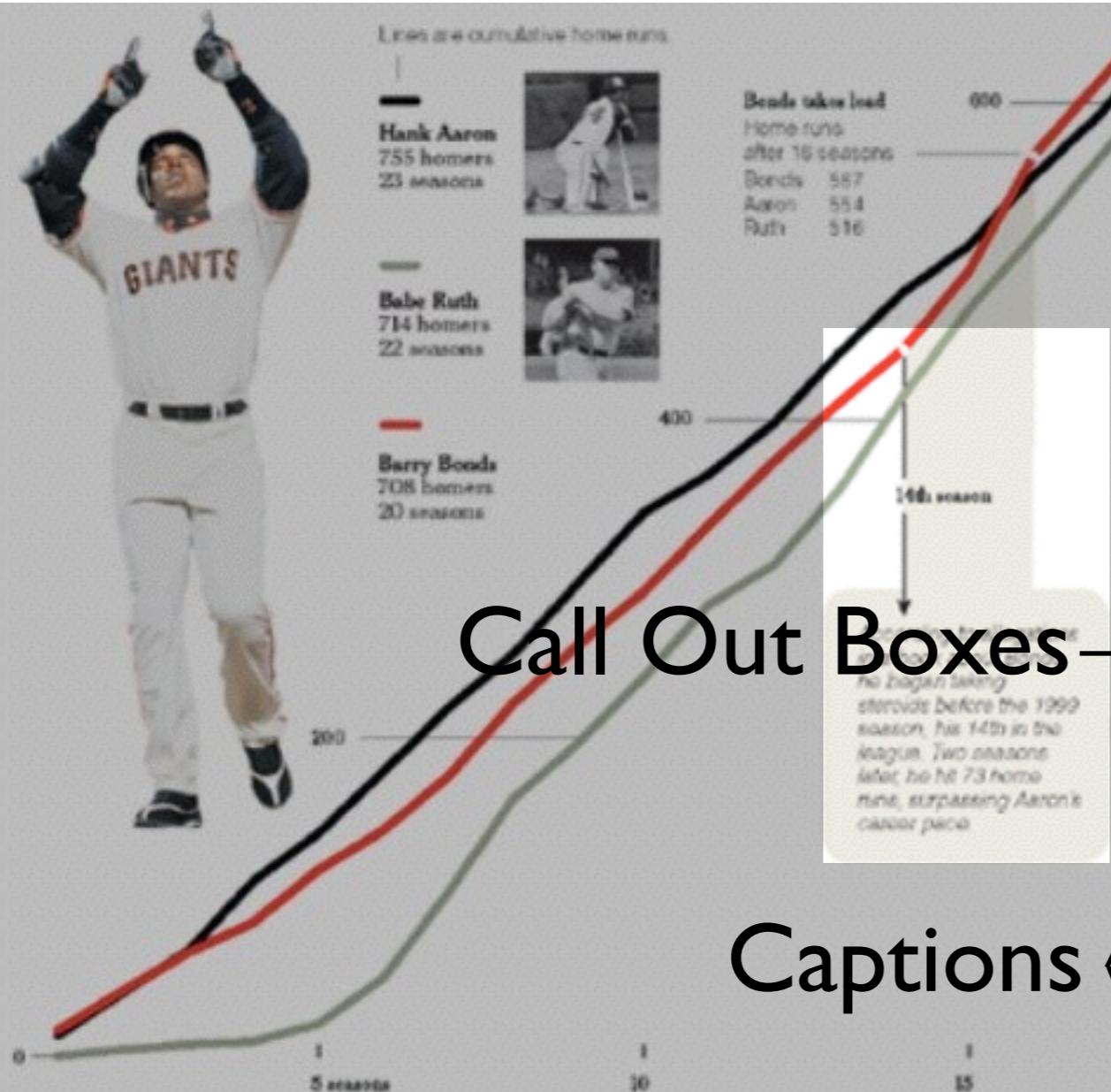
E. Segel

755

Headline

Steroids or Not, the Pursuit Is On

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



Others Taking Aim

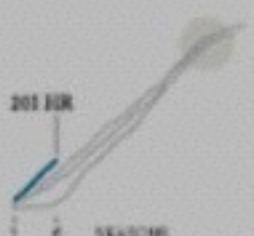
Alex Rodriguez

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



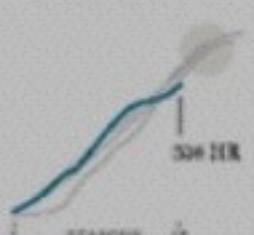
Albert Pujols

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



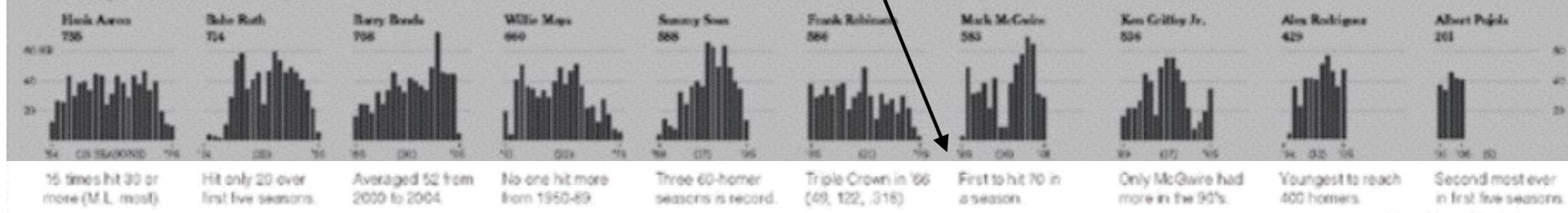
Ken Griffey Jr.

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 (129), Rodriguez (37th) and Pujols (5ed 257th).



© 2005 The New York Times Company

Where the Power Is Out and Returning Across the Northeast

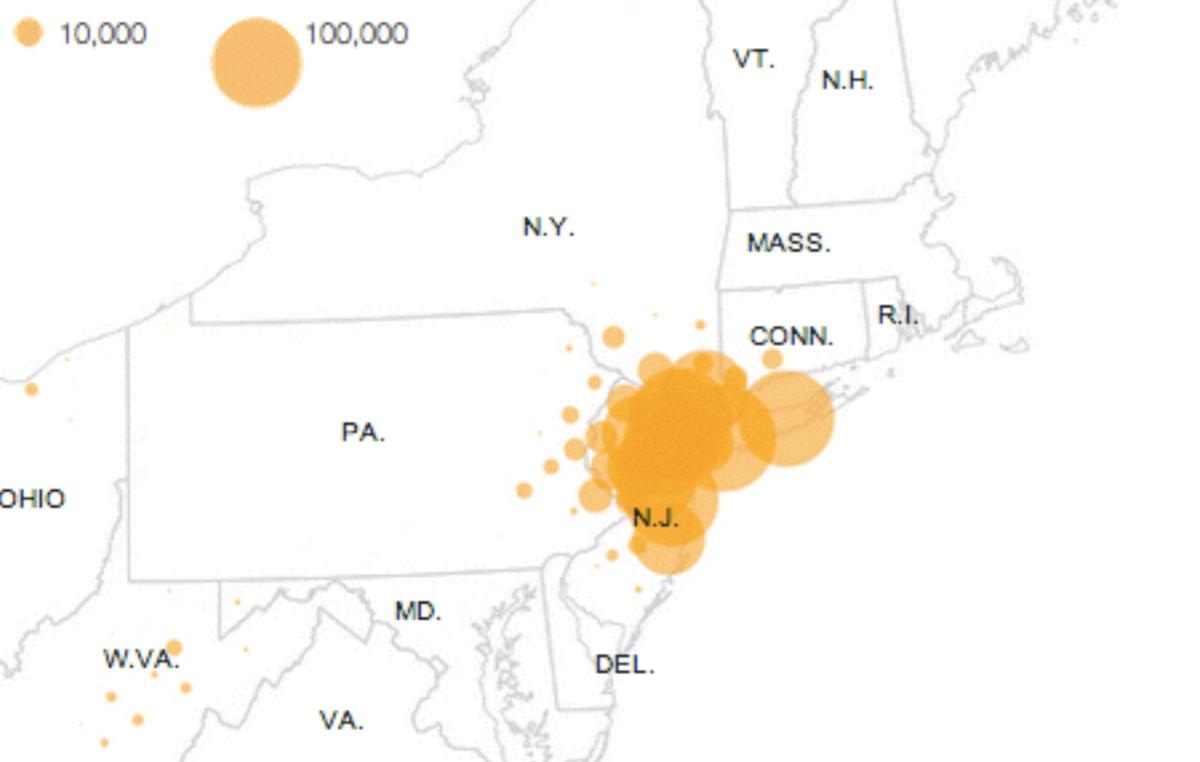
Updated Sunday, November 4 at 8:00 PM

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.

Power outages across the Northeast

Customers without power

● 10,000



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



WHERE THERE'S SMOKE—THERE'S CANCER

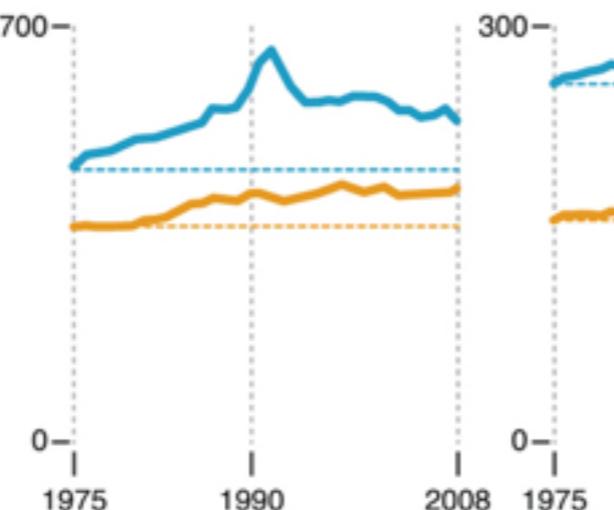
Cancer rates are up, but mortality is down. New diagnostics and treatments are responsible for part of this trend. But the greatest single contributing factor is the decline in smoking—rates are at their lowest level in 50 years.

Men Women

1 Increased incidence

An aging population contributes to rising incidence of cancer.

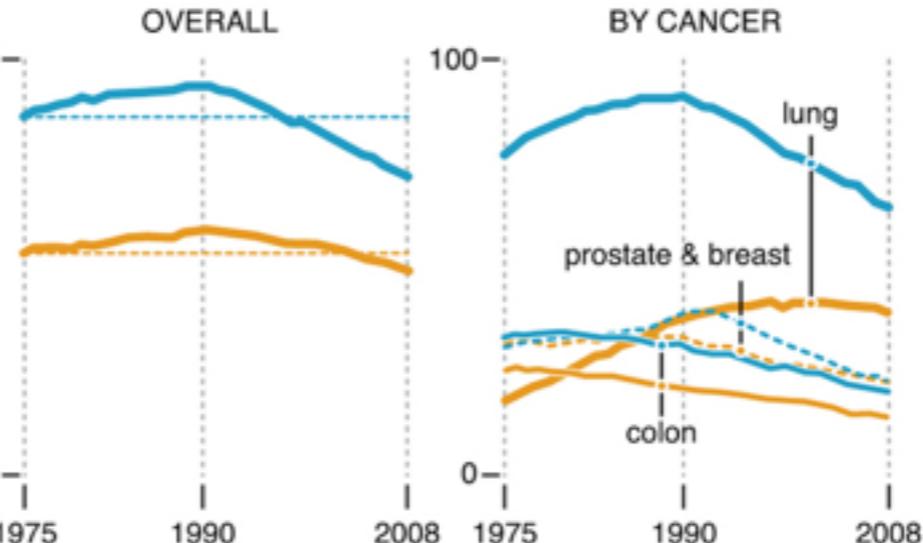
Cancer incidence rates (per 100,000)



2 Fewer deaths

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

Cancer death rates (per 100,000)



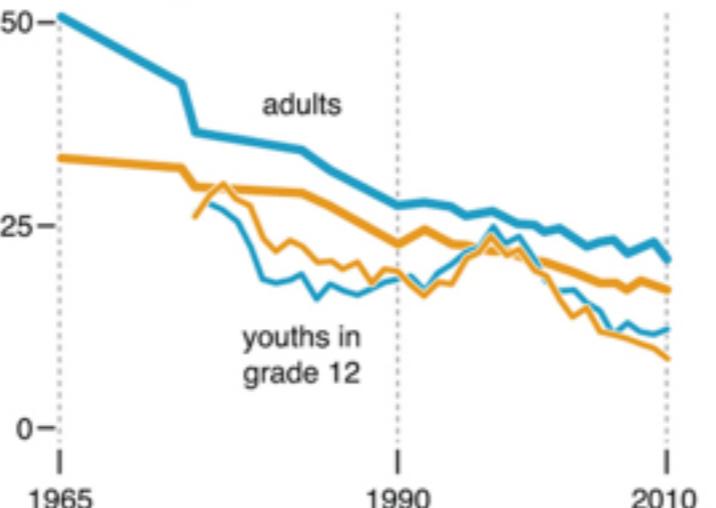
3 Decline of lung cancer

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.

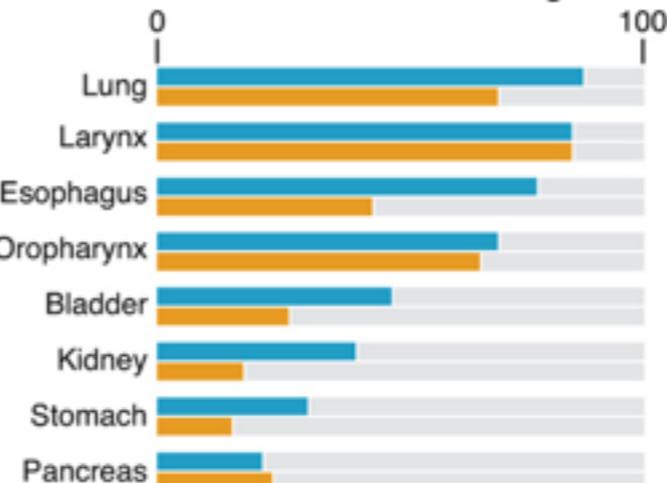
Smoking prevalence (%)



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.

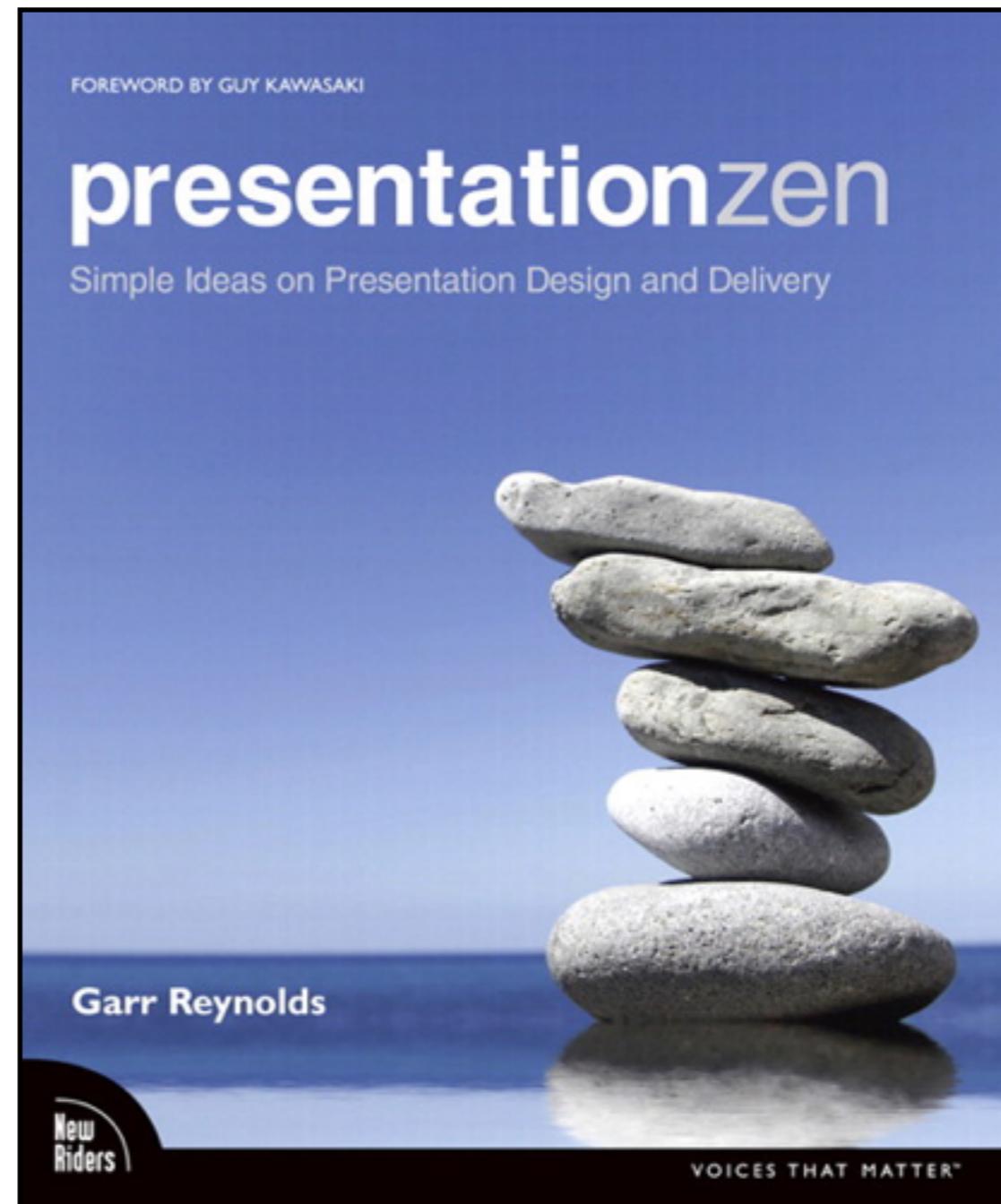
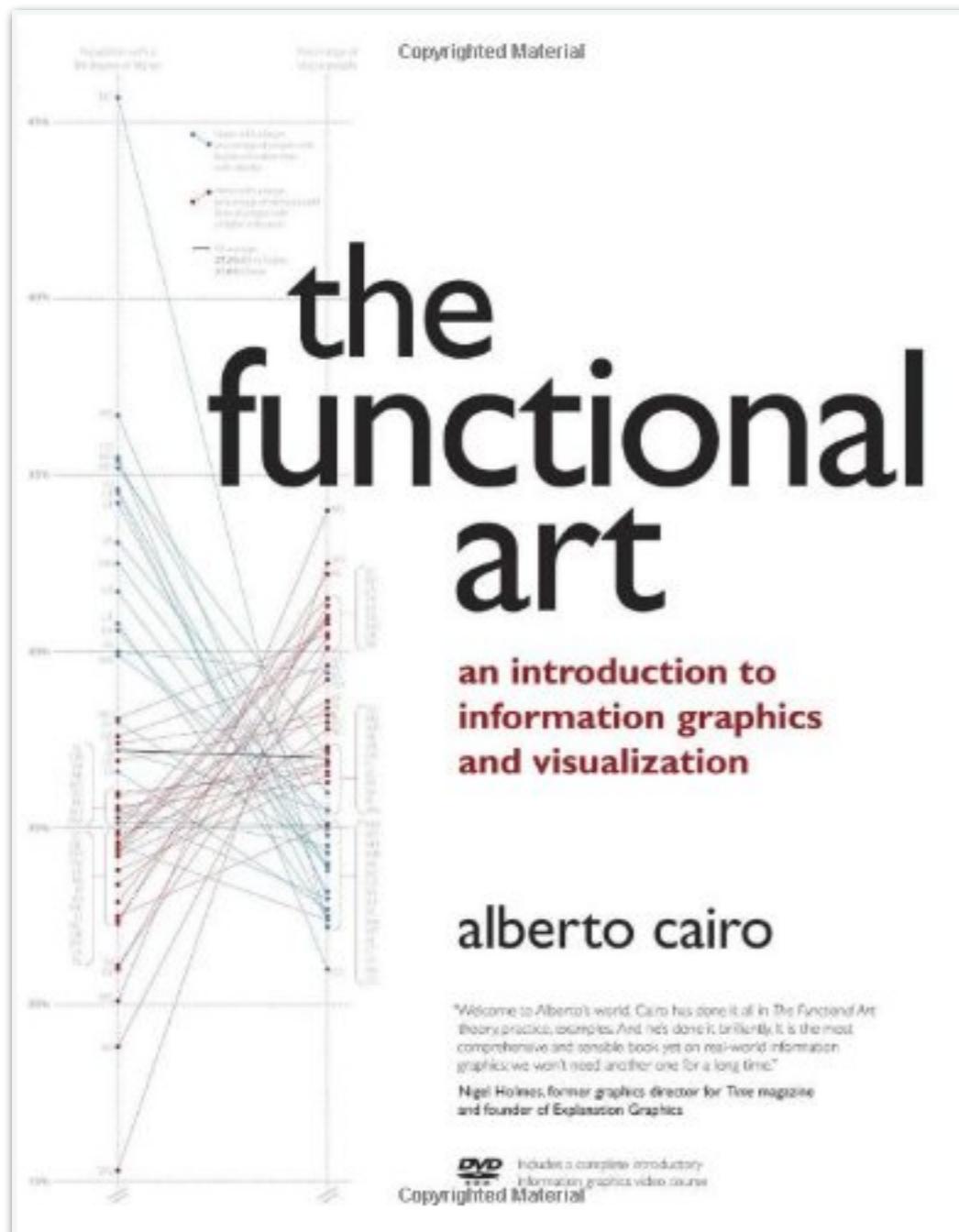
Percentage of cancer deaths attributable to smoking



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

