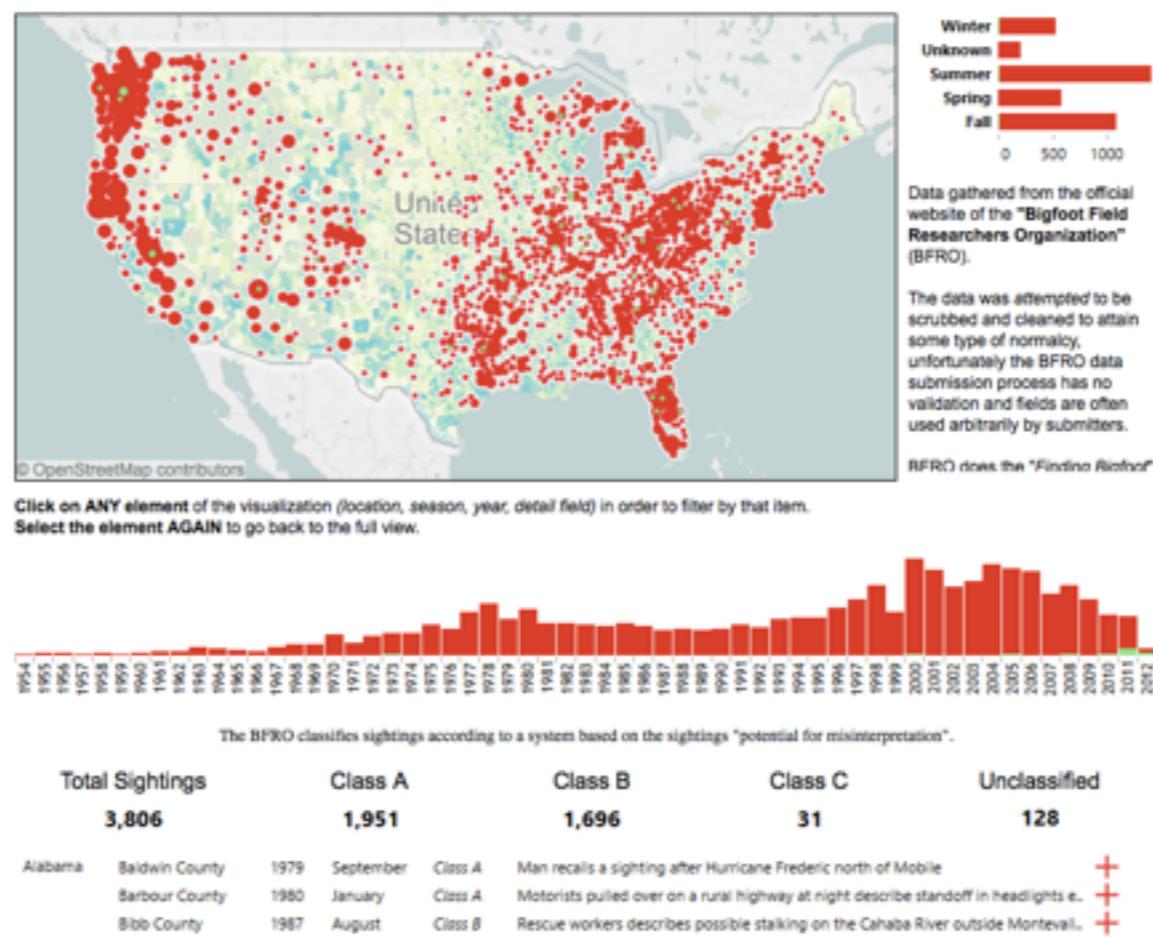


# CS I 09/Stat I 2I/AC209/E- I 09

# 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](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 NH<sub>2</sub>-terminal portion of the adenosinetriphosphatase (ATPase) subunit 6 gene has been identified as the animal equivalent of the recently discovered yeast H<sup>+</sup>-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.

Want to have a memorable easily understandable story/example.

# 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)

Avoid the generic

# 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)

Give a talk: Use VISUALS

Helps a lot, and can tell a story with pictures

# 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?**  
Final Project: Make it accessible to anyone  
Exciting for others: get them interested in your research  
Video: It is advertisement for the paper  
(which can be technical and complicated)
- **What questions are you answering?**
- **Why should the audience care?**
- **What are your major insights and surprises?**
- **What change to 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? ie, being a student -> teaching/learning
- What insights can you give them? What tools and “magical gifts”?

# Don't Make Them Think!

Don't require understanding of notation

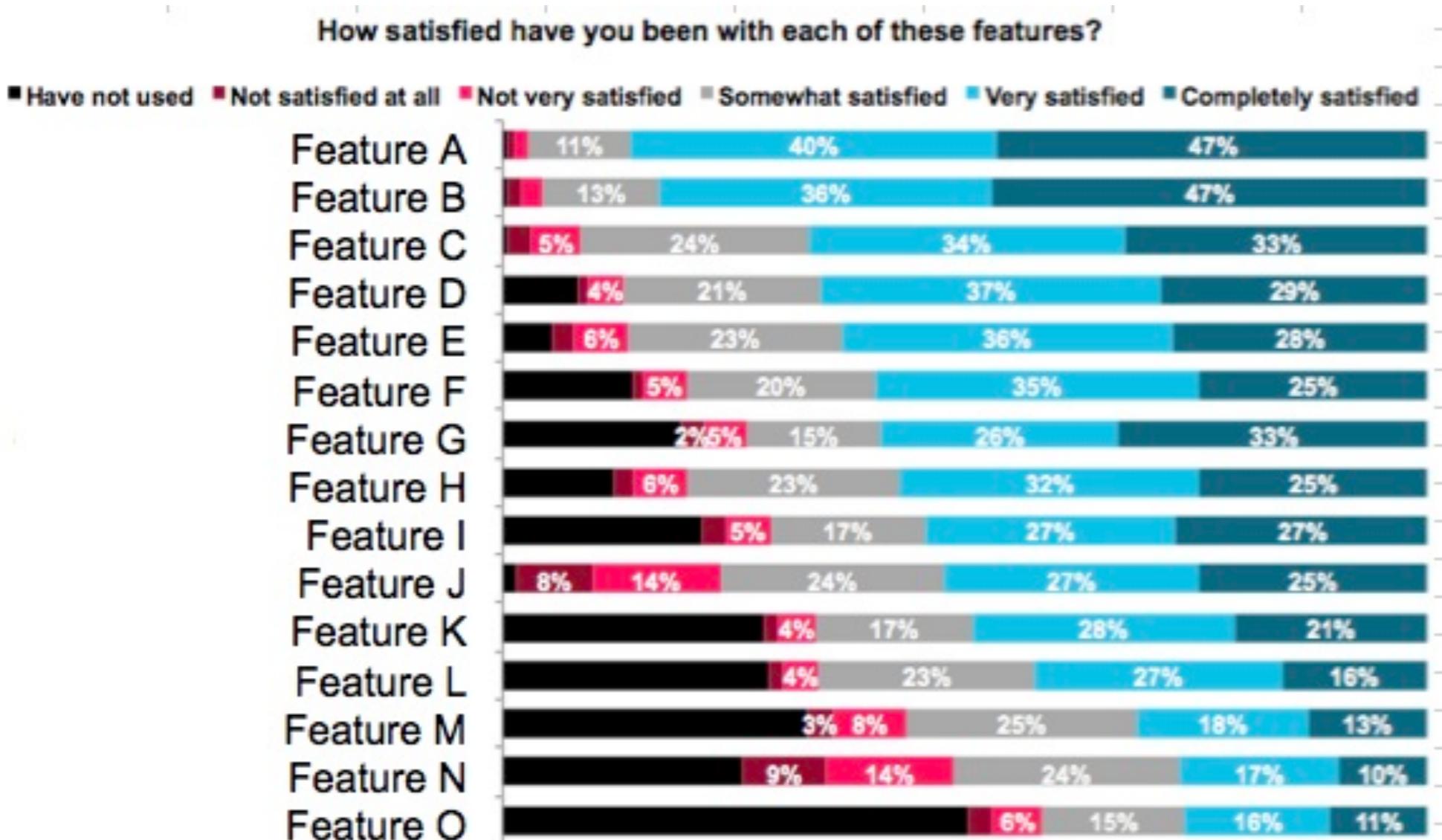
Keep everything at an easy to understand level

- 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

captions/annotations good in paper



# Don't Bury the Lead



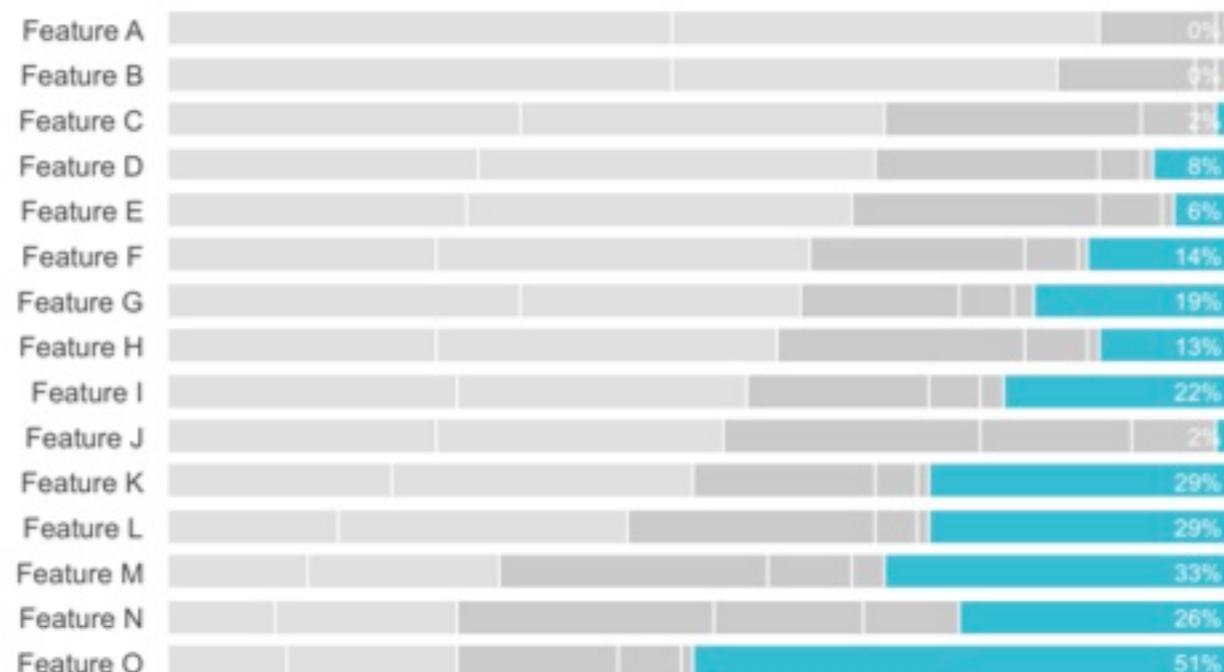
# Don't Bury the Lead

User satisfaction varies greatly by feature

Message is much clearer here

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?

Our Final project requires a 2 min video

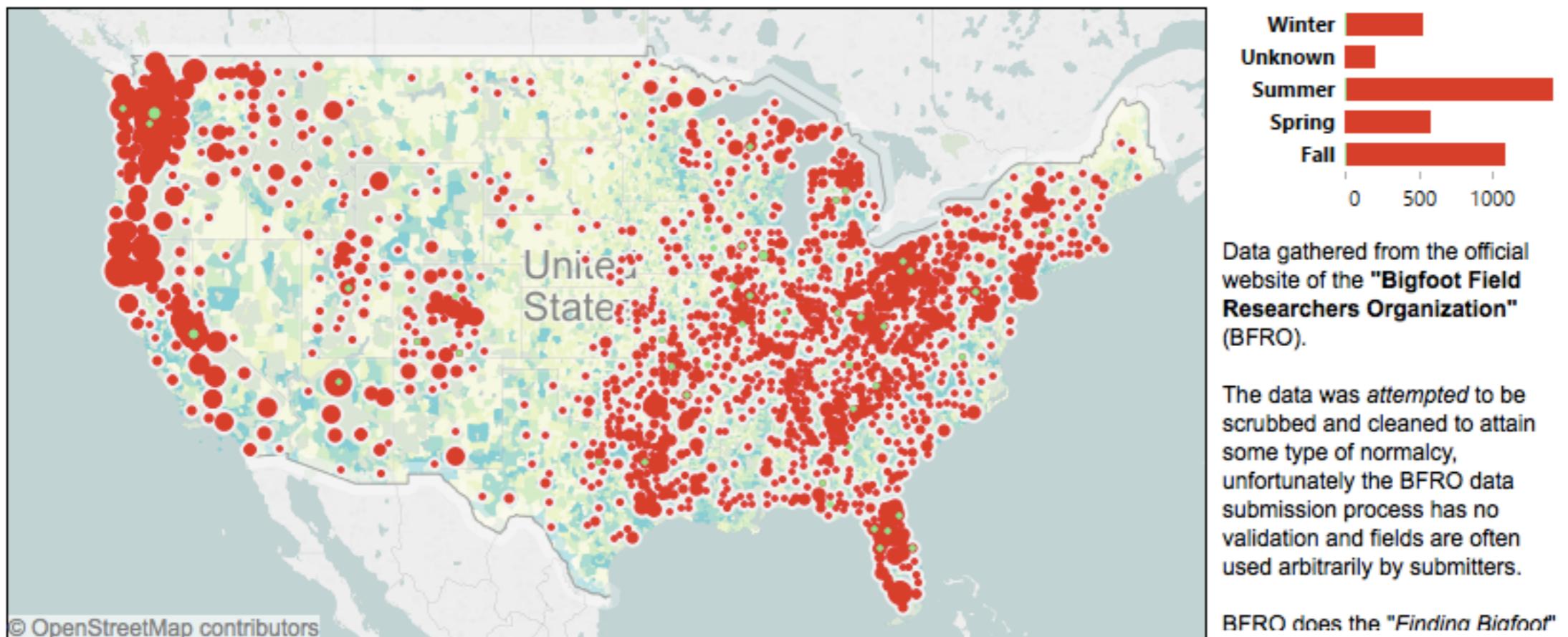
# David Jacopille



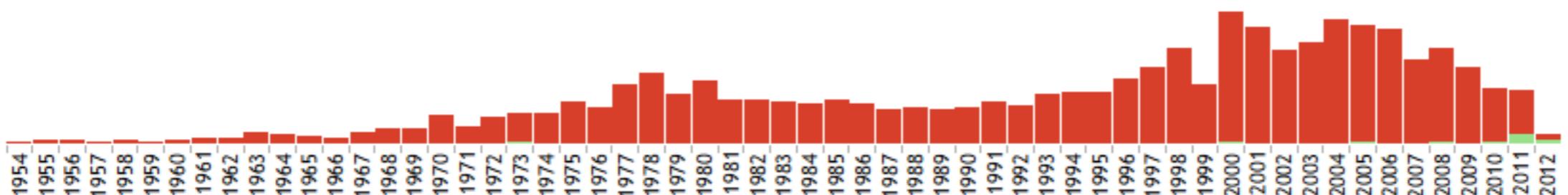
Ryan Robitaille

Bigfoot sighting correlated with population density

# 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.



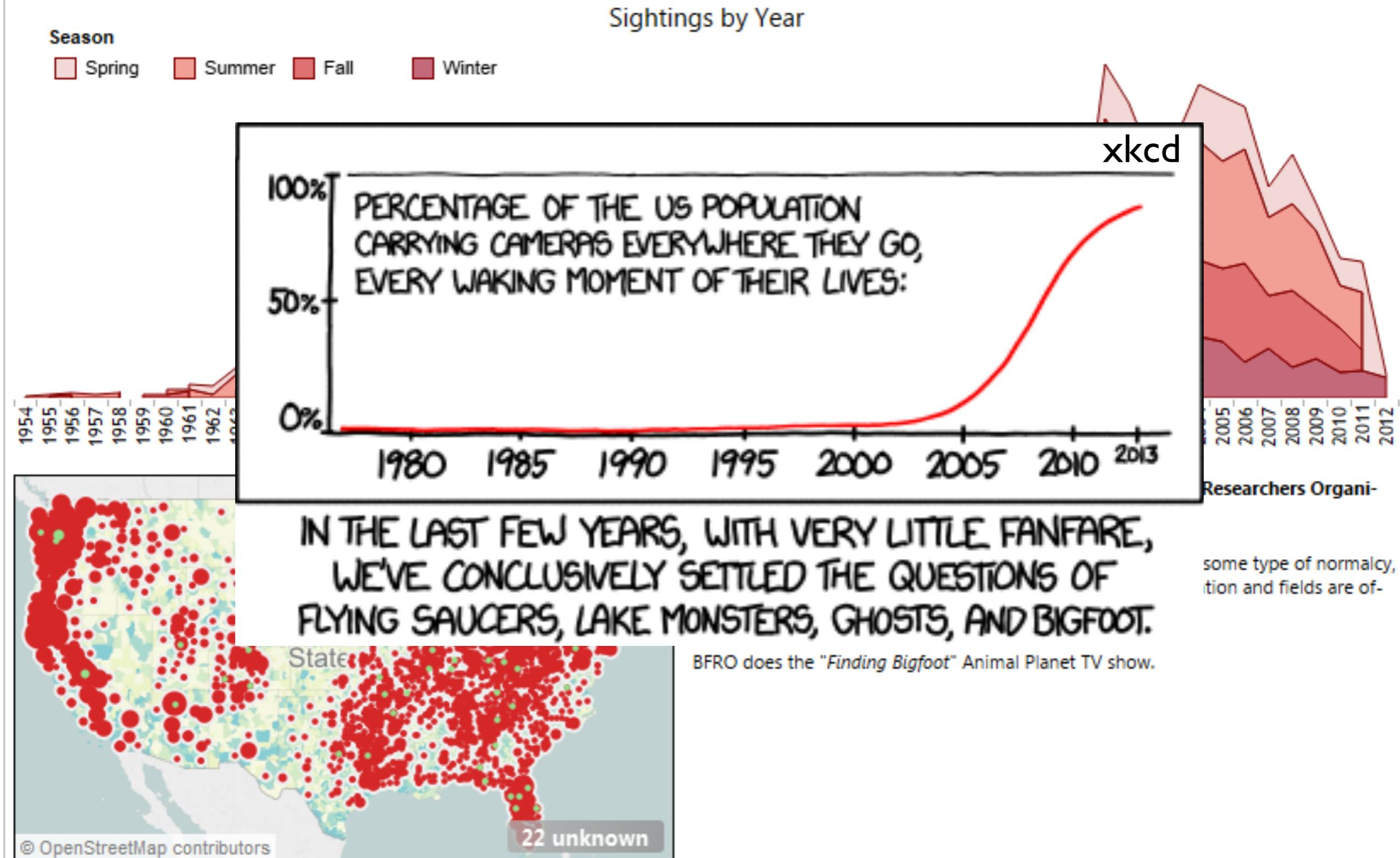
The BFRO classifies sightings according to a system based on the sightings "potential for misinterpretation".

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 Montevall..	+

Put explicit question in title and it should be interesting.

# Bigfoot sightings are in decline



Ryan Robitaille

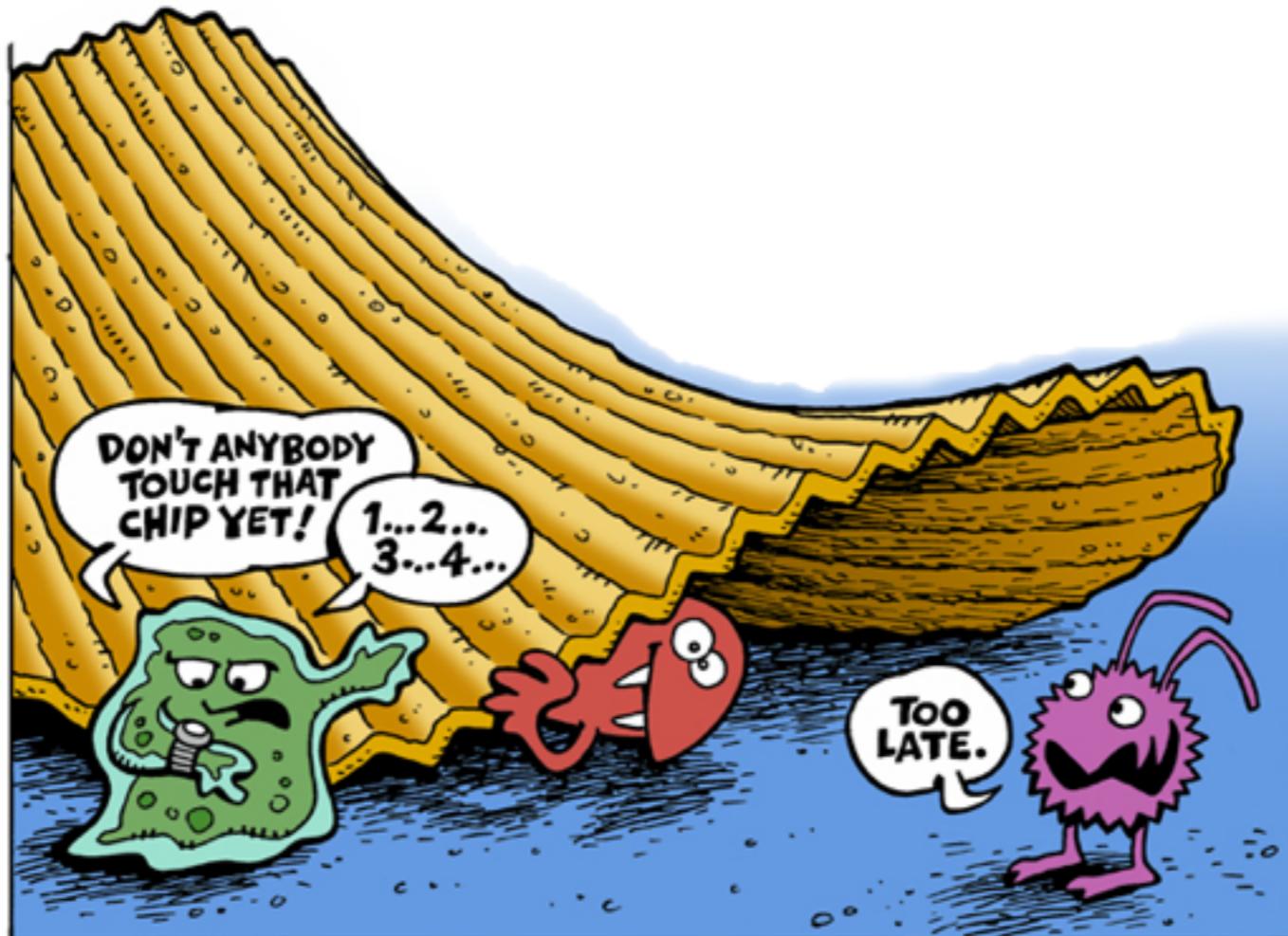
# Unexpectedness

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

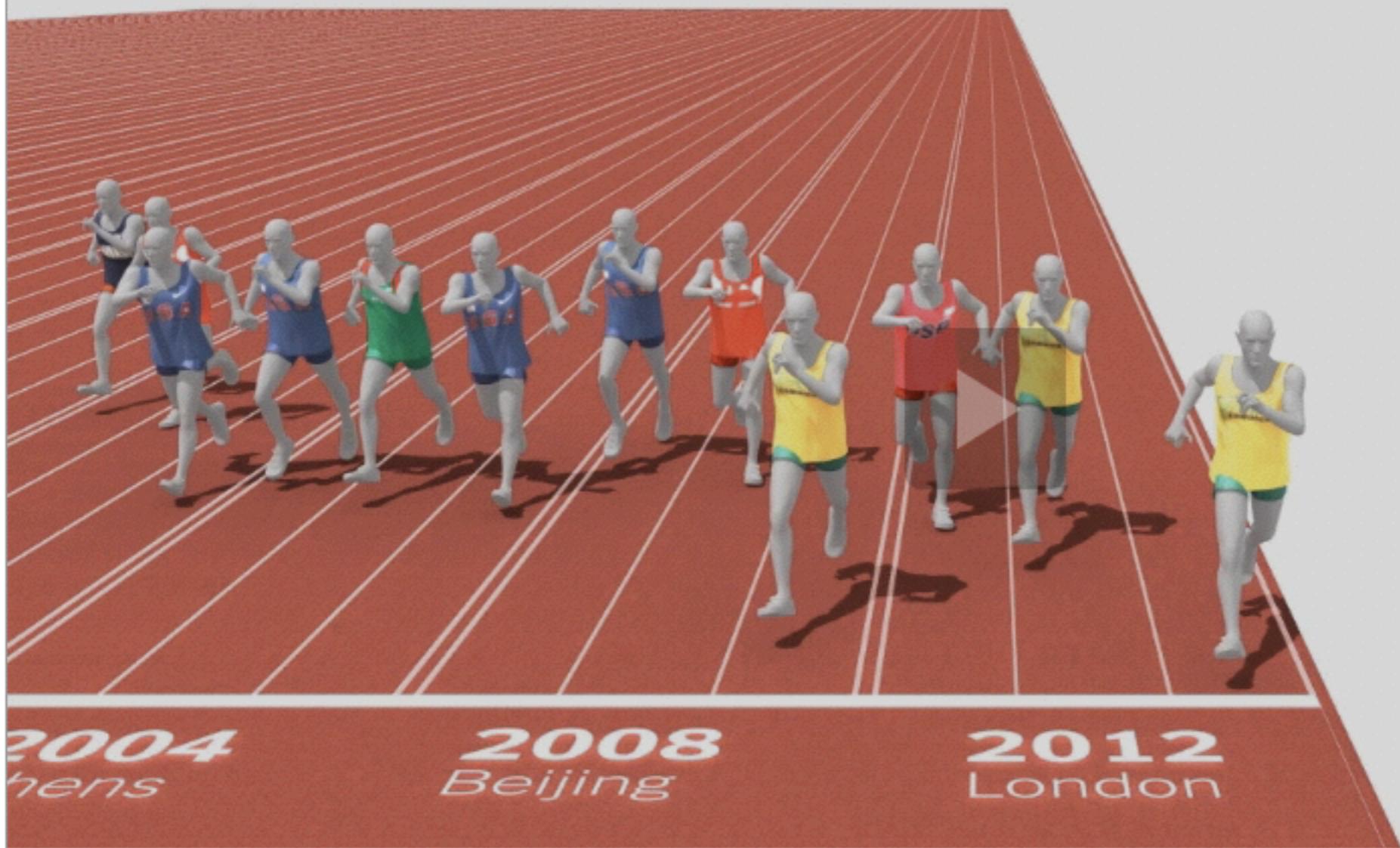
Use surprise to grab the audience's attention

“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](#)

Can enhance or detriment your presentation  
Must frame the story correctly.

# Messaging

very important to frame with type of story/presentation and audience can frame emotionally, logically, etc.

# Framing - Why should I care?

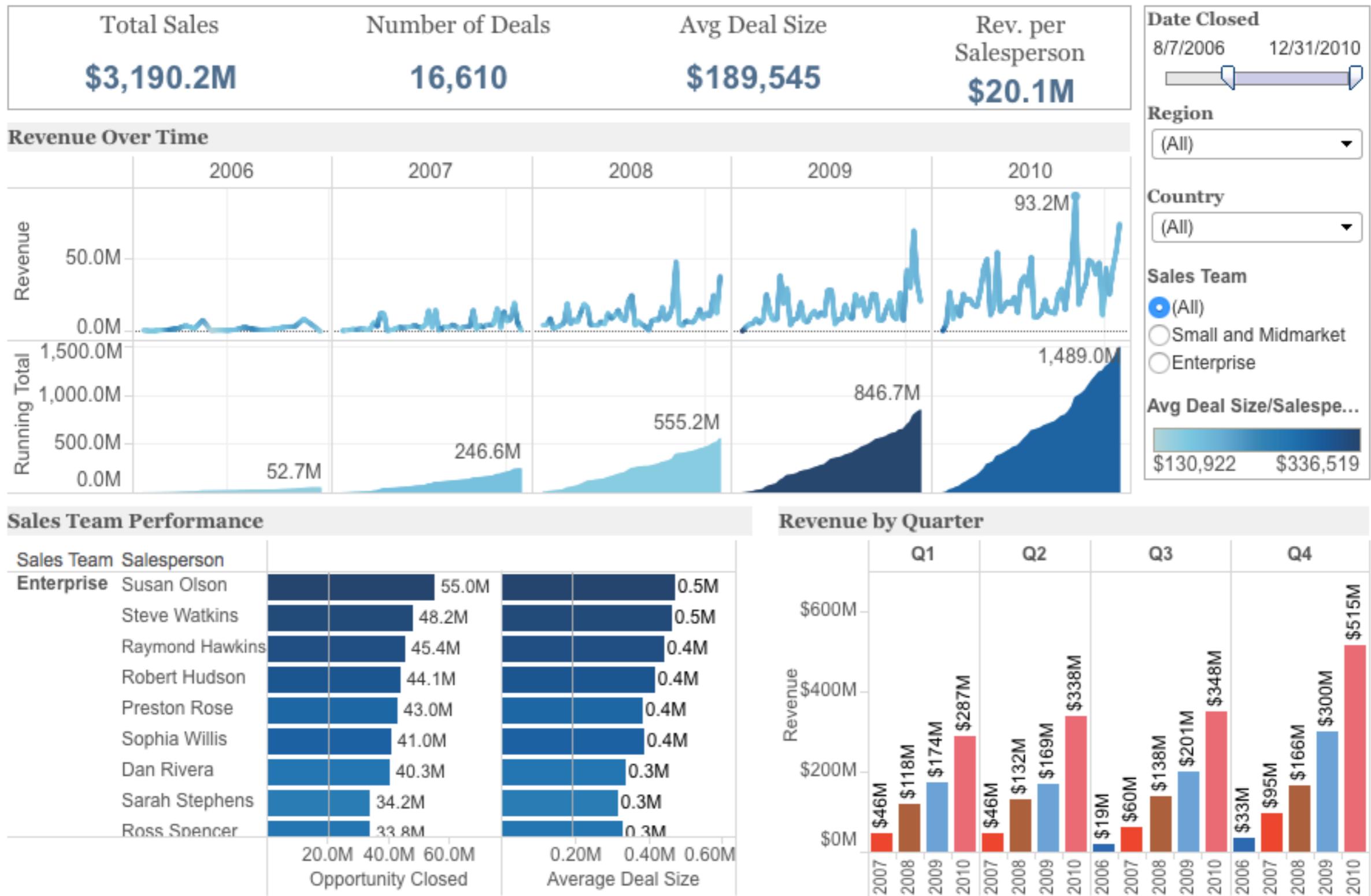
- Tell the audience: “Here is the right way to think about the problem I was trying to solve.”
- Catch the audience’s attention and frame the story using captions and annotations
- If done well, your insights will seem obvious given this framing. And that’s a good thing!



Opportunity Dashboard

Sales Dashboard

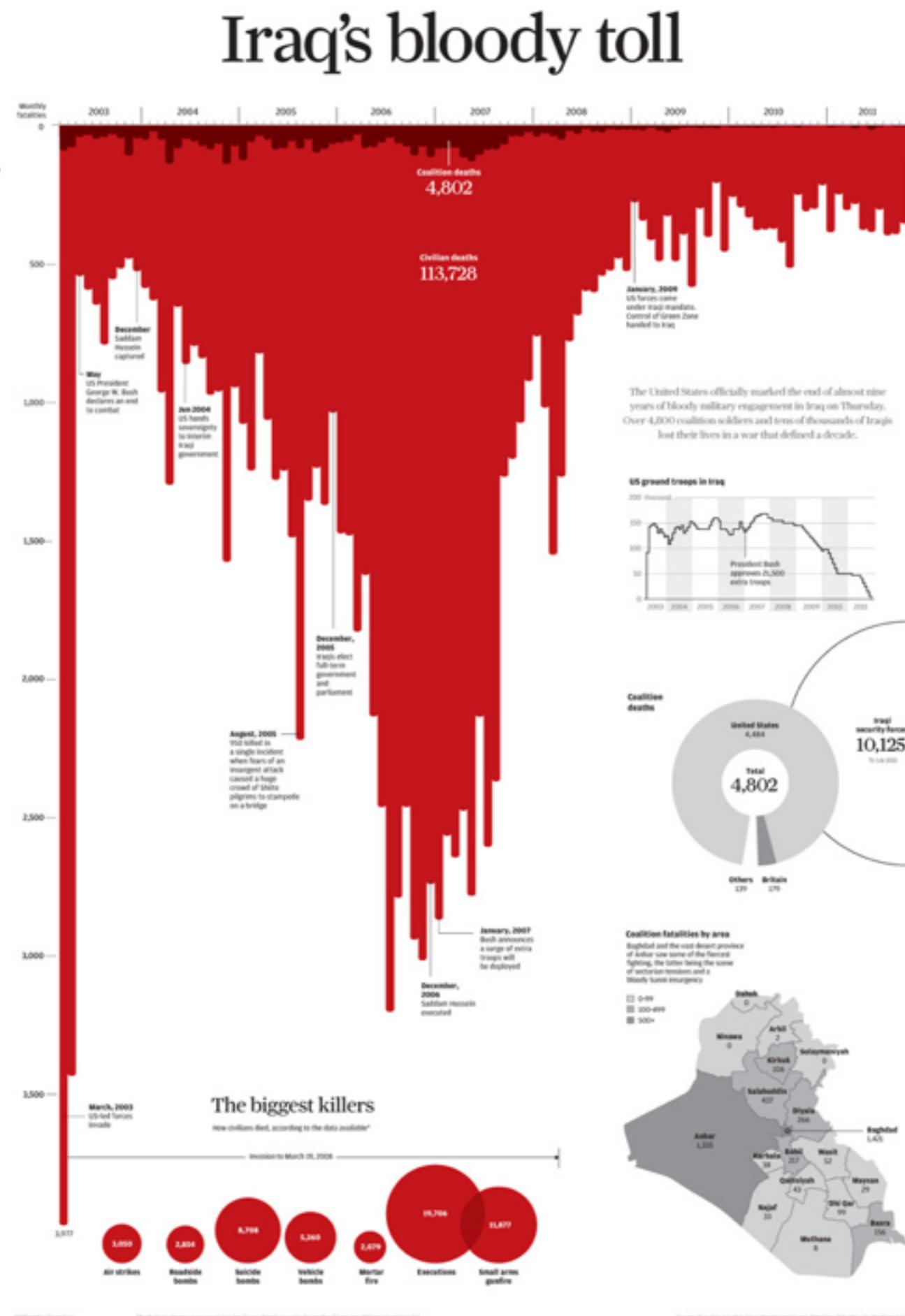
## Sales Dashboard



looks like blood dripping:  
frame: negative pov on  
life's lost  
-> risk to be opinionated,  
can turn off those who disagree

more emotionally engaging  
BUT, don't lose integrity in the  
visualization

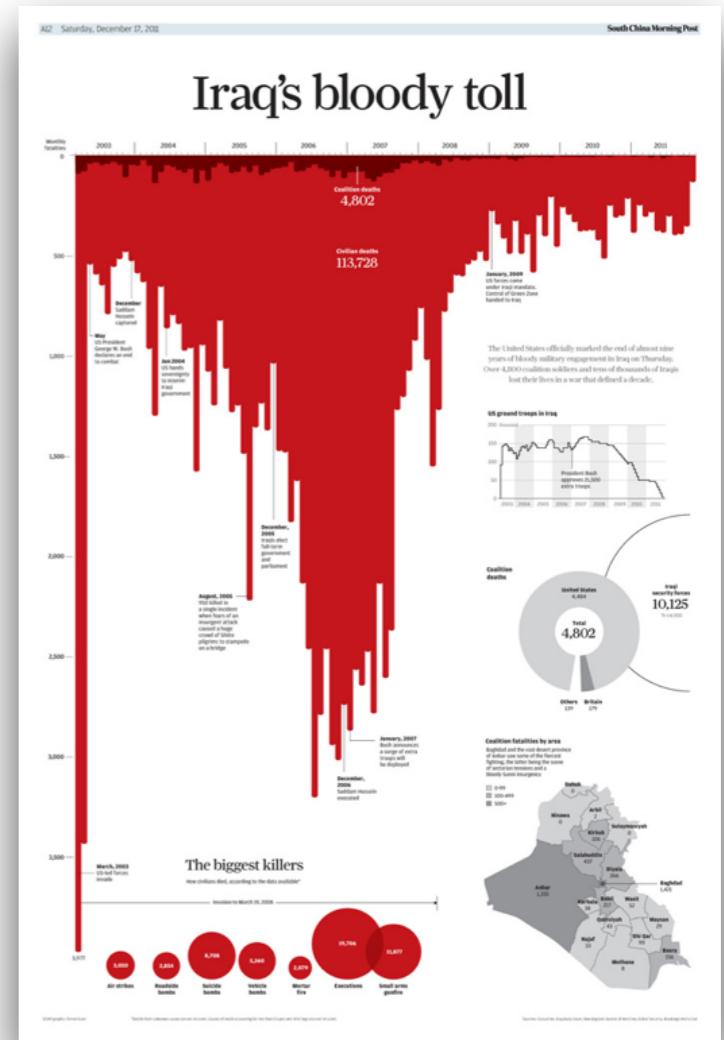
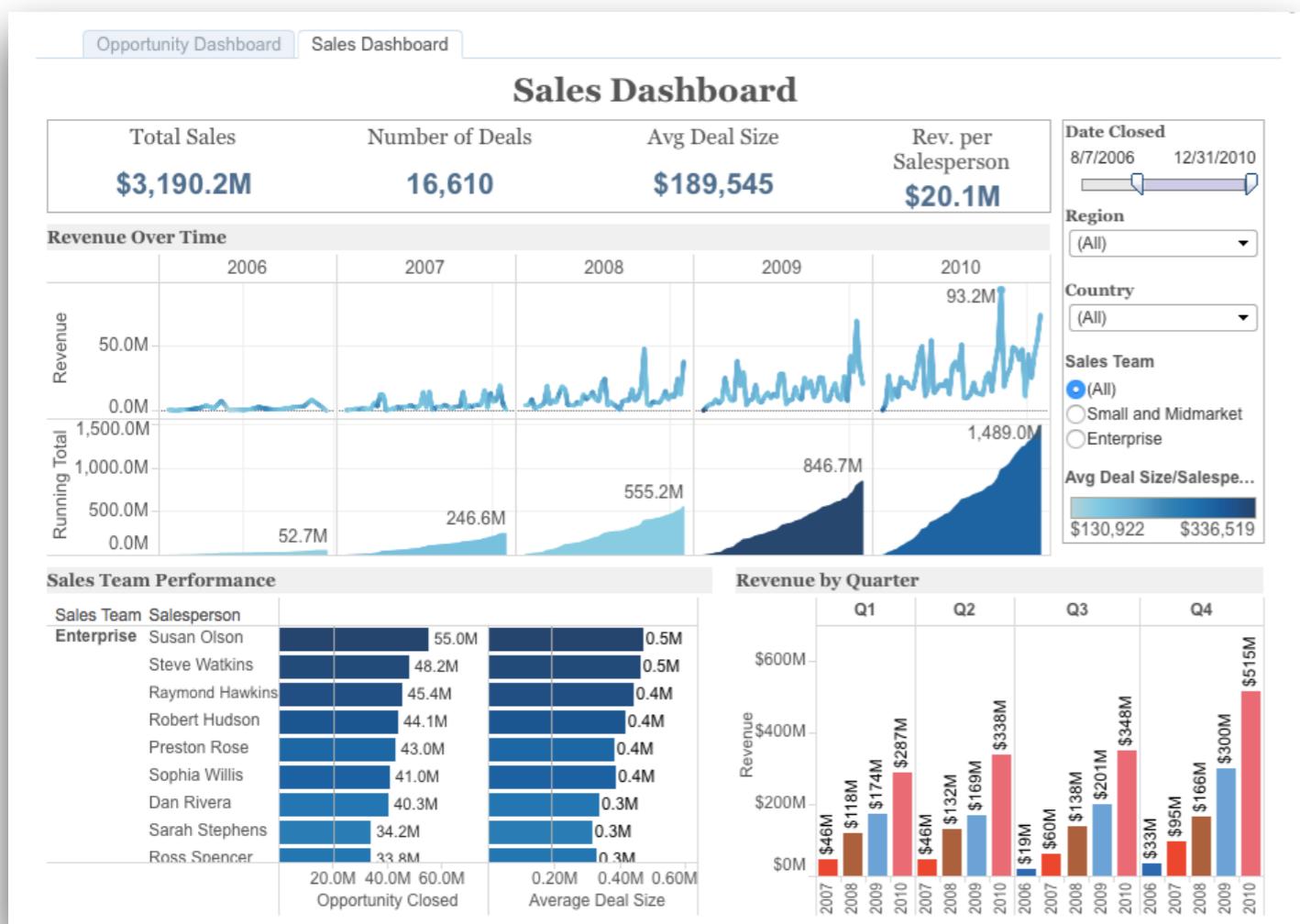
This example does not seem  
to distort data



# What is the message?

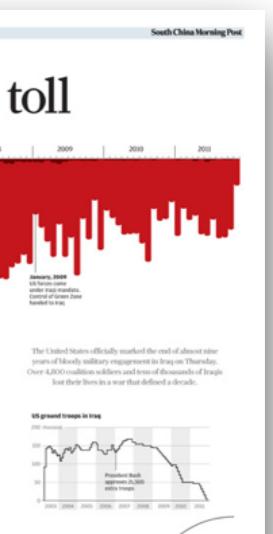
Exploratory  
Neutral

Explanatory  
Opinionated



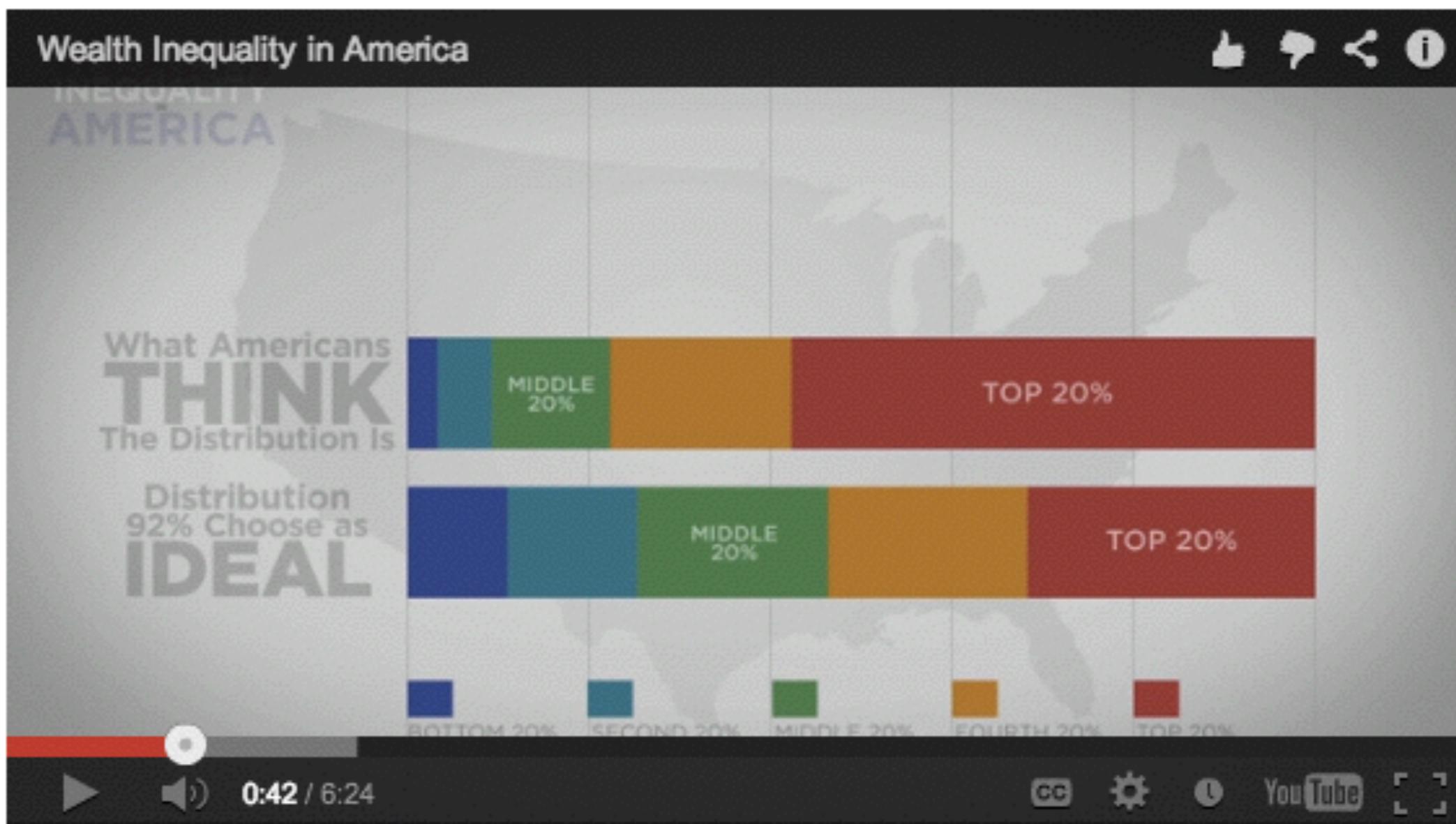
# Know Your Audience

This board is unrealistically diverse ....



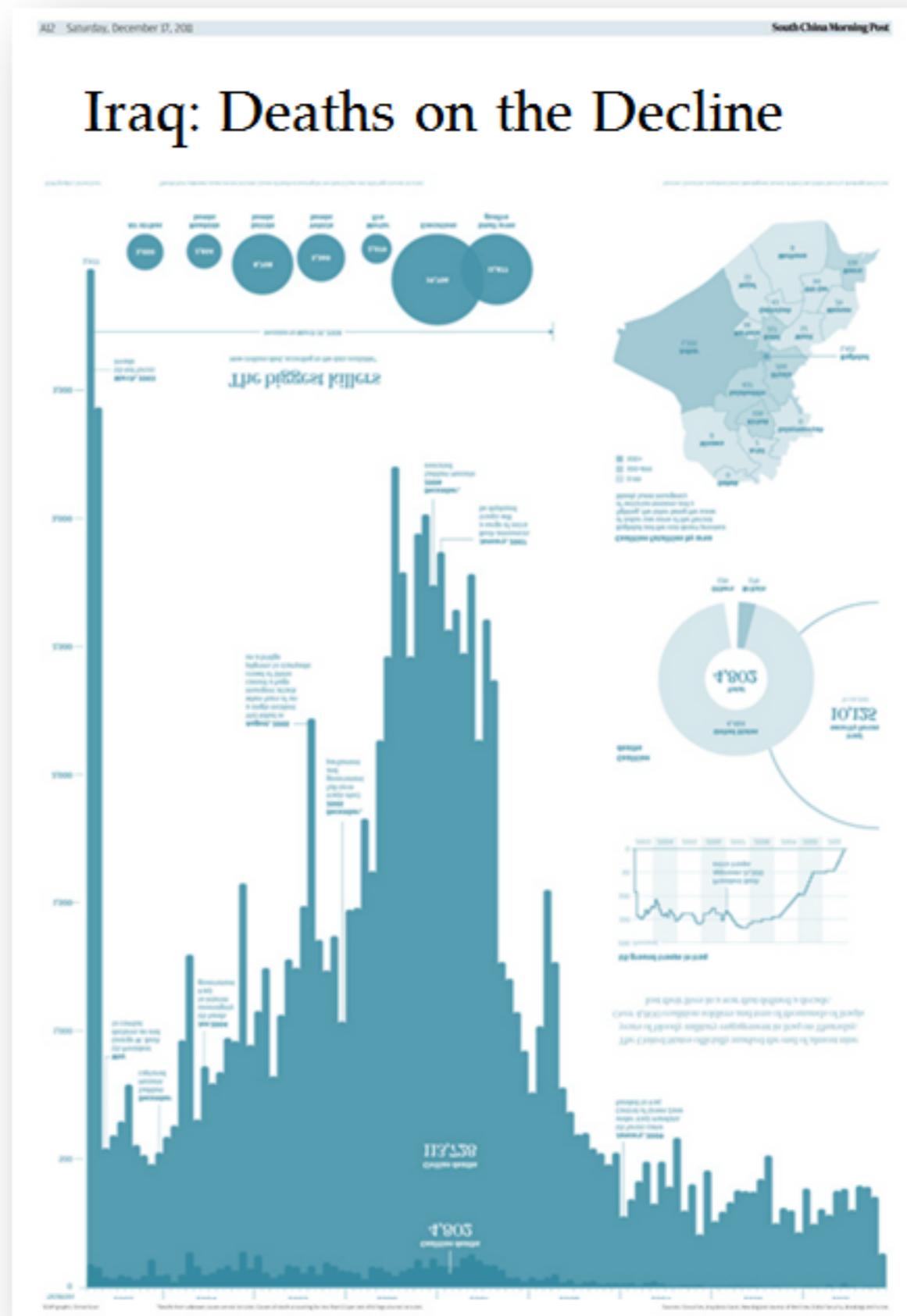
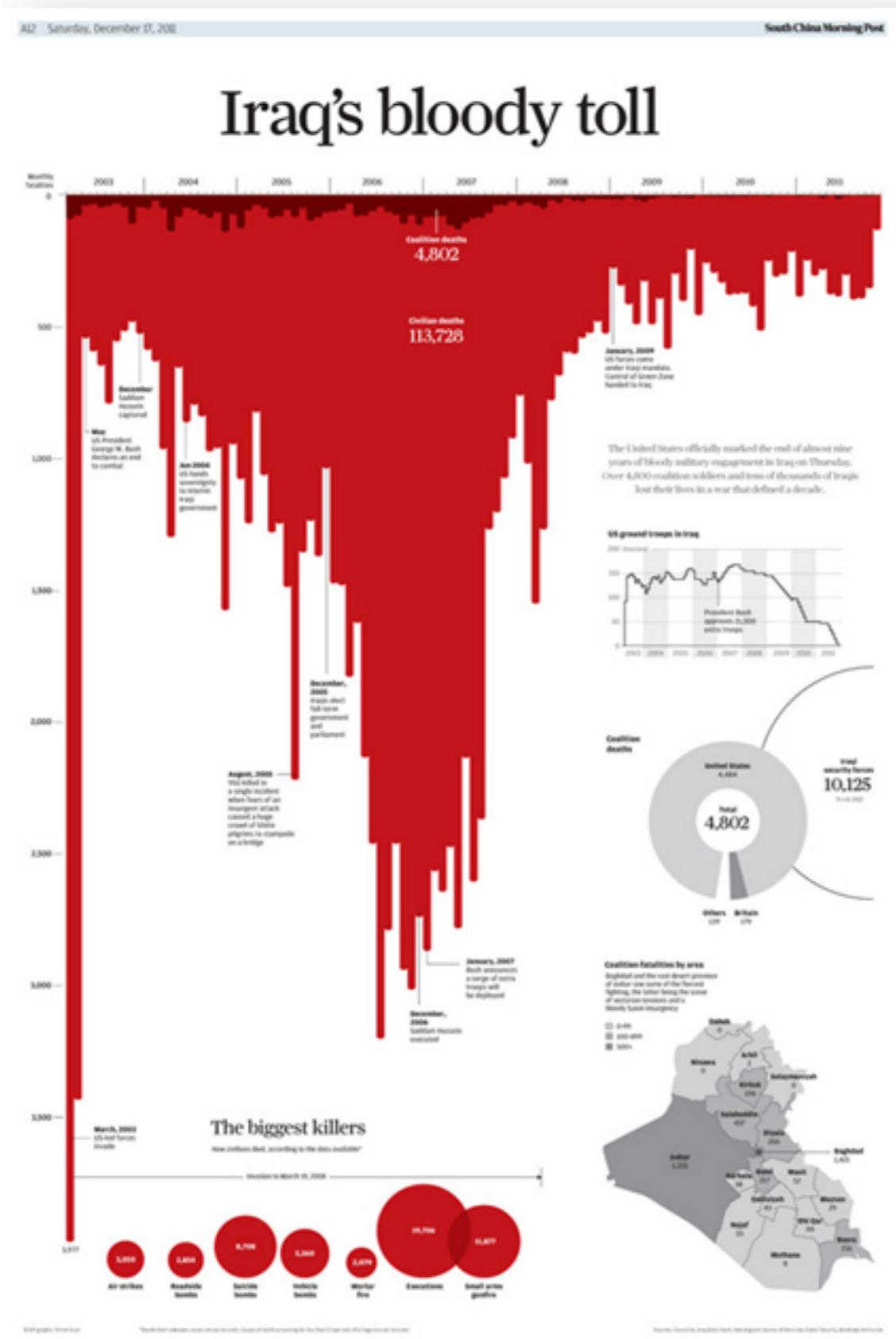
# Wealth Inequality in America

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



# Visual Story Design

choice of color, orientation, and title made all the difference (same underlying data)



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.

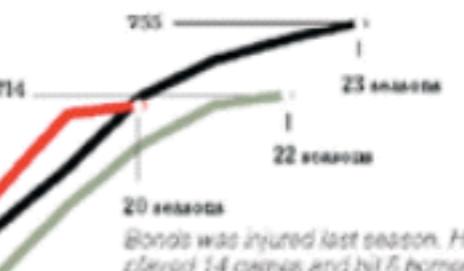


Bonds takes lead  
Home runs:  
after 16 seasons  
Bonds 587  
Aaron 554  
Ruth 516

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.



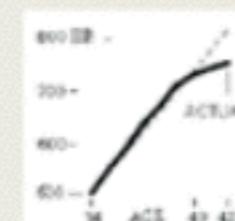
Text

### Homer Pace After Age 34

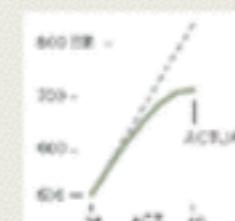
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

PROJECTION BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

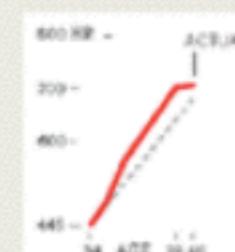
**Aaron**  
Actual homers slightly outpace projected homers for five seasons



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

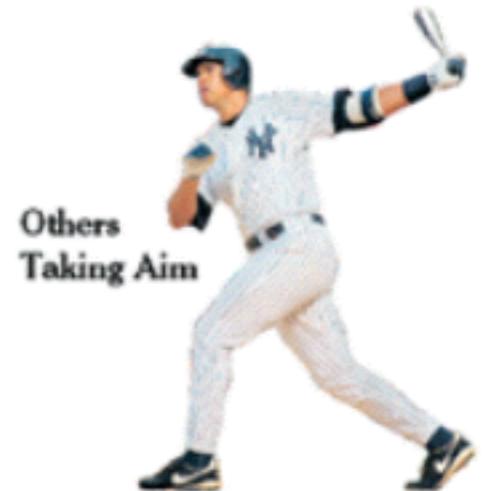


**Bonds**  
From age 35 to 39, he averaged 14 more homers a season than projected

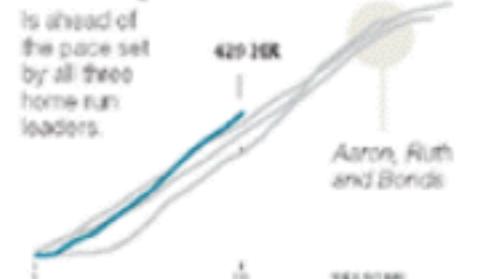


Note: Ages as of July 1 of each season

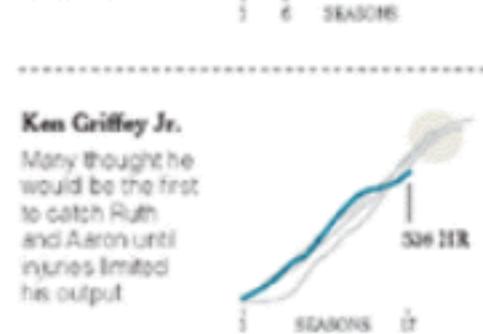
### Others Taking Aim



**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.

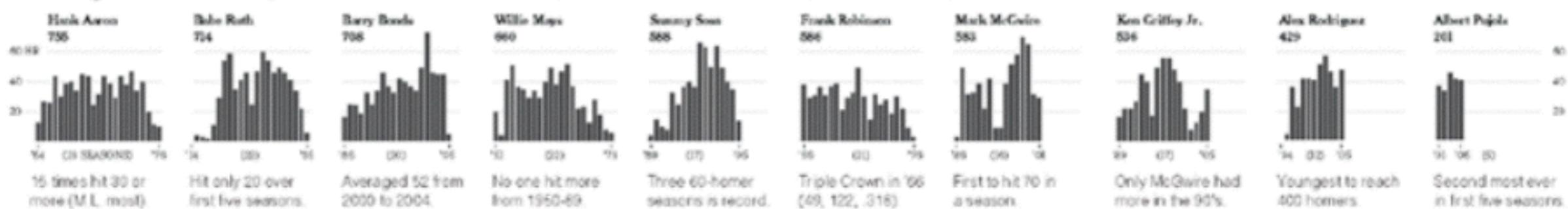


**Ken Griffey Jr.**  
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.

Note: Ages as of July 1 of each season

### 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).



Source: The Sporting News and New York Times

E. Segel

# 755

## Steroids or Not, the Pursuit Is On

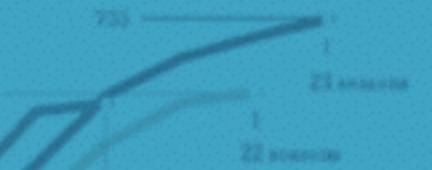
Every Bonds is being won at the career home run record. His record is only six more to tie Babe Ruth and 47 to equal Hank Aaron.

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

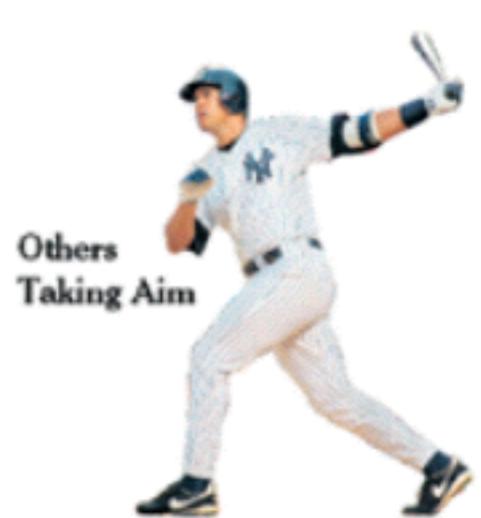
If the accusations are correct, Bonds was 24 in his first season on steroids. Here are projected home run paces for each player after age 34.

Accurately rates Bonds the fastest of previous five leaders.



Note: Ages as of July 1 of each season

### Others Taking Aim



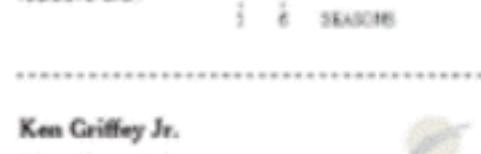
#### Alex Rodriguez

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



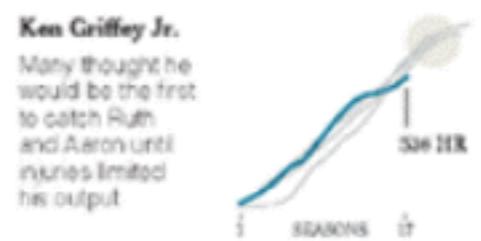
#### Albert Pujols

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



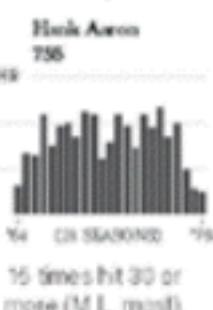
#### Ken Griffey Jr.

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

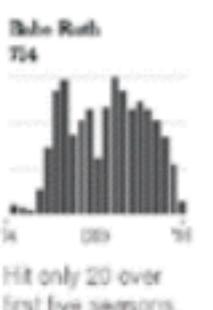


### Differing Paths to the Top of the Charts

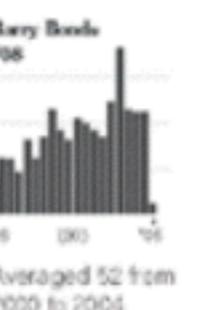
The top seven players on the career home run list, along with a look at Griffey (129), Rodriguez (37th) and Pujols (led 257th).



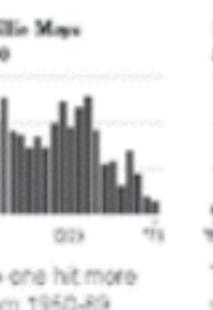
15 times hit 30 or more (M.L. most).



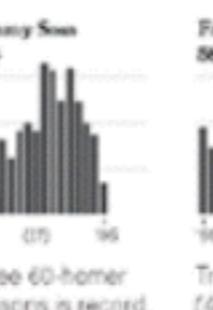
Hit only 20 over first five seasons.



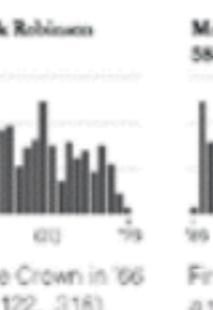
Averaged 52 from 2000 to 2004.



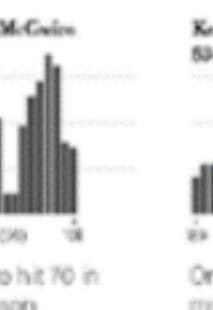
No one hit more from 1950-69.



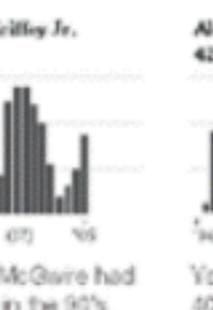
Three 60-homer seasons is record.



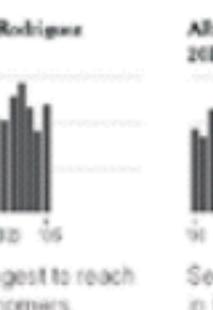
Triple Crown in '66 (49, 122, .316).



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

# 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



Bonds takes lead  
1990-91-92-93  
After 75 home runs  
Bonds: 547  
Aaron: 544  
Ruth: 538

600

650

700

750

800

850

900

950

1000

1050

1100

1150

1200

1250

1300

1350

1400

1450

1500

1550

1600

1650

1700

1750

1800

1850

1900

1950

2000

2050

2100

2150

2200

2250

2300

2350

2400

2450

2500

2550

2600

2650

2700

2750

2800

2850

2900

2950

3000

3050

3100

3150

3200

3250

3300

3350

3400

3450

3500

3550

3600

3650

3700

3750

3800

3850

3900

3950

4000

4050

4100

4150

4200

4250

4300

4350

4400

4450

4500

4550

4600

4650

4700

4750

4800

4850

4900

4950

5000

5050

5100

5150

5200

5250

5300

5350

5400

5450

5500

5550

5600

5650

5700

5750

5800

5850

5900

5950

6000

6050

6100

6150

6200

6250

6300

6350

6400

6450

6500

6550

6600

6650

6700

6750

6800

6850

6900

6950

7000

7050

7100

7150

7200

7250

7300

7350

7400

7450

7500

7550

7600

7650

7700

7750

7800

7850

7900

7950

8000

8050

8100

8150

8200

8250

8300

8350

8400

8450

8500

8550

8600

8650

8700

8750

8800

8850

8900

8950

9000

9050

9100

9150

9200

9250

9300

9350

9400

9450

9500

9550

9600

9650

9700

9750

9800

9850

9900

9950

10000

10050

10100

10150

10200

10250

10300

10350

10400

10450

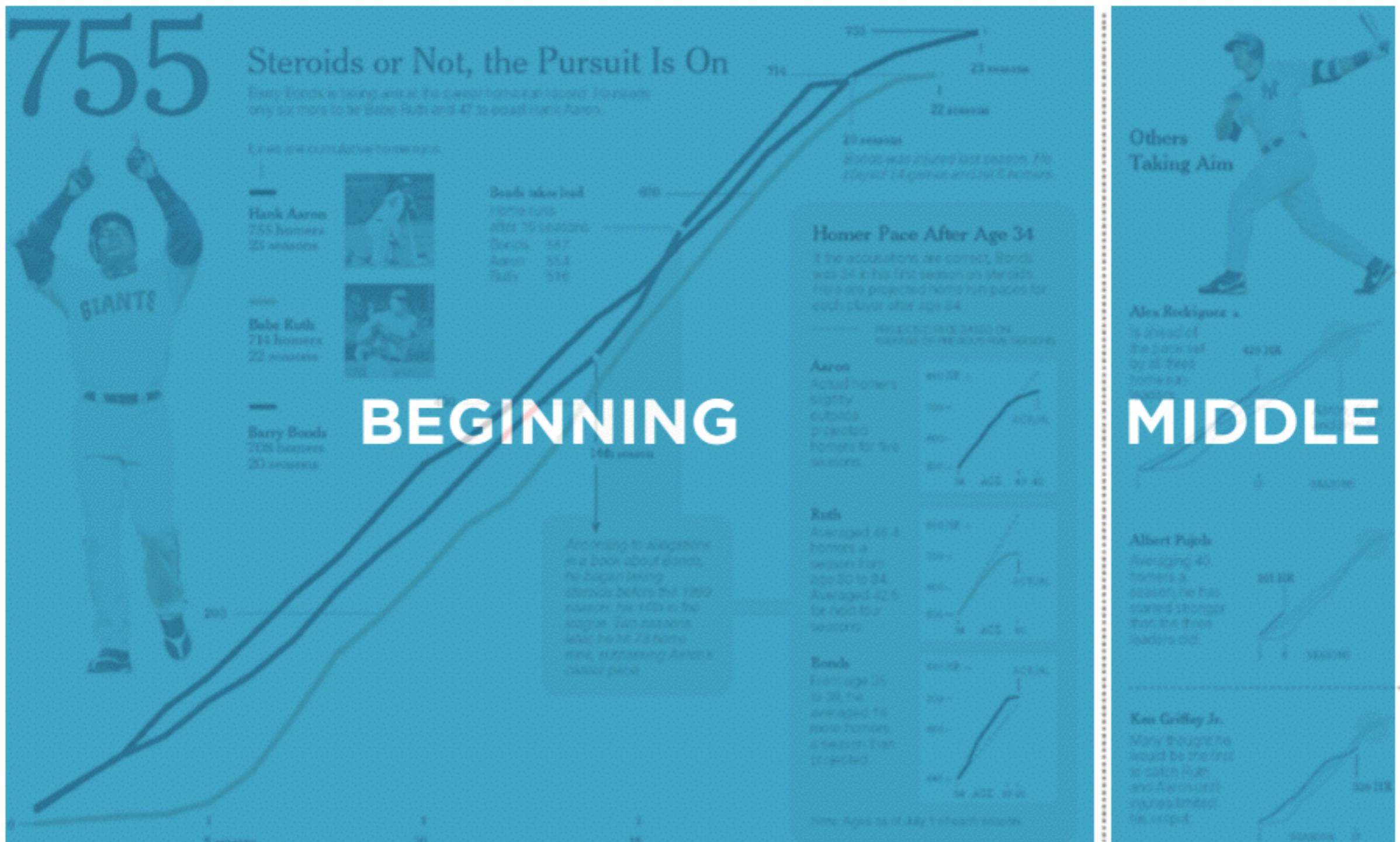
10500

10550

10600

10650

Put important stuff on the top: infinite scrolling, have all the space on the bottom.

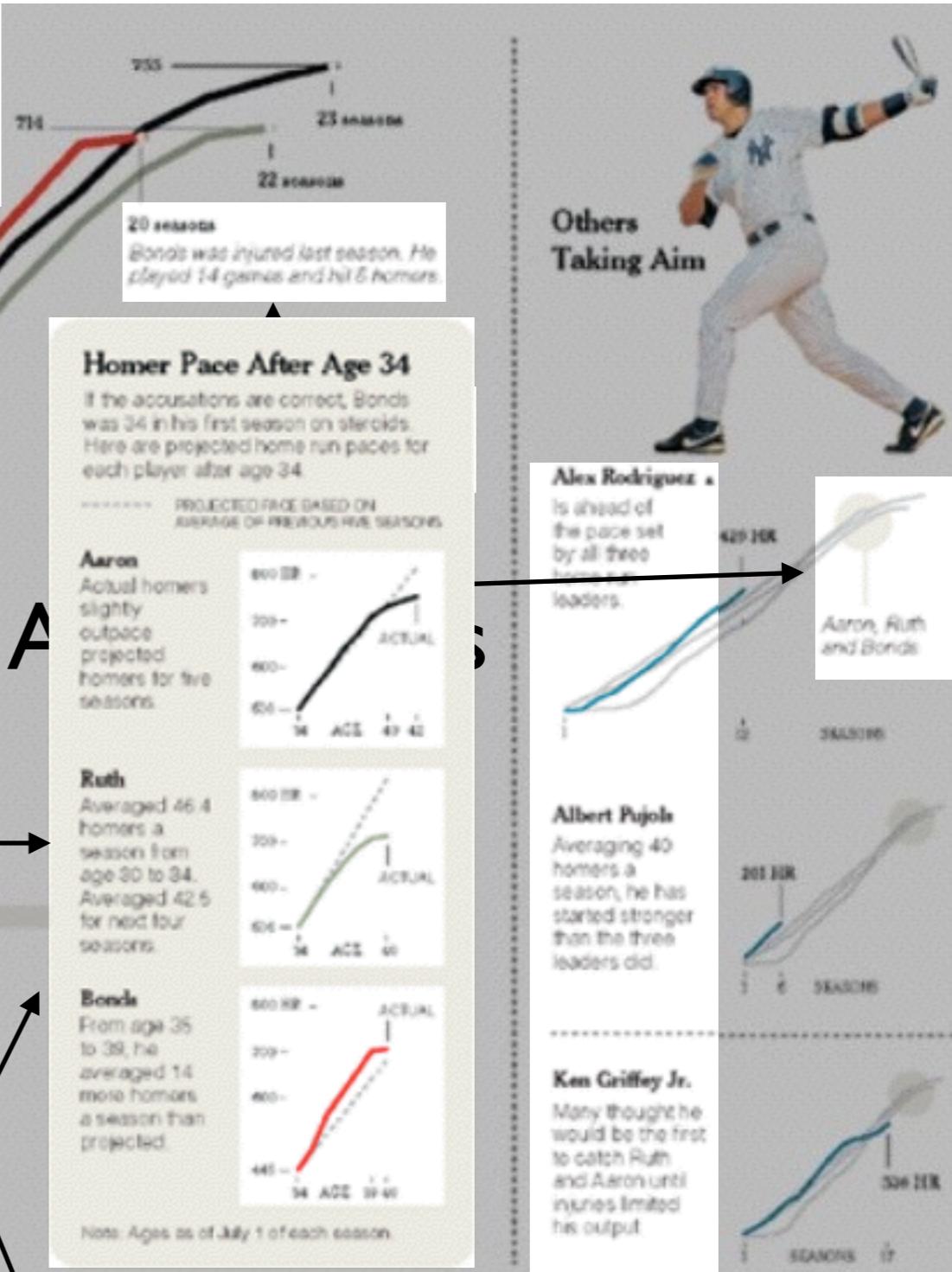
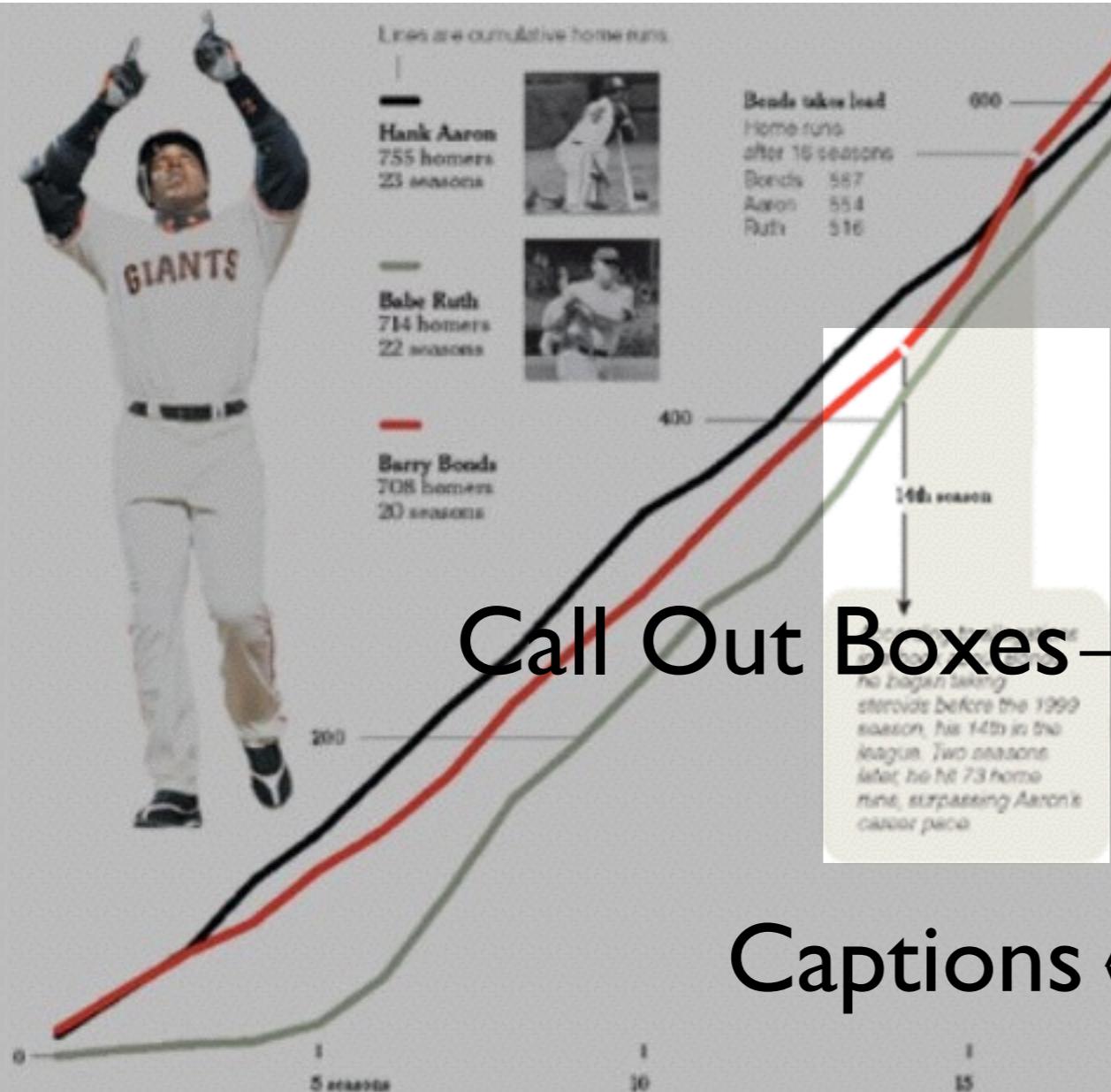


# 755

## Headline

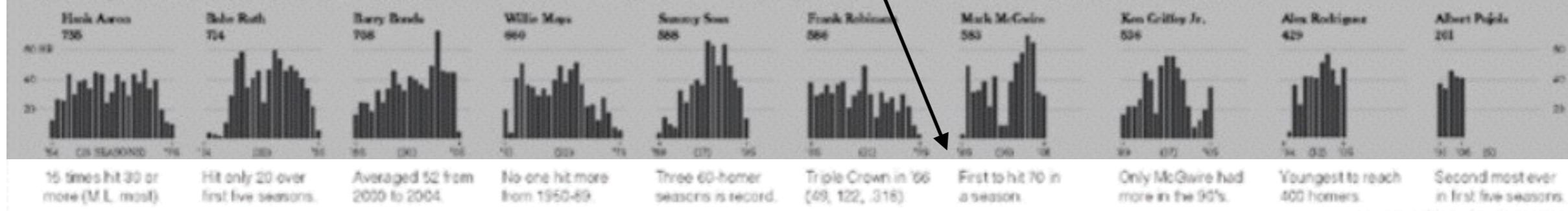
### Steroids or Not, the Pursuit Is On

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.



### 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).



# Where the Power Is Out and Returning Across the Northeast

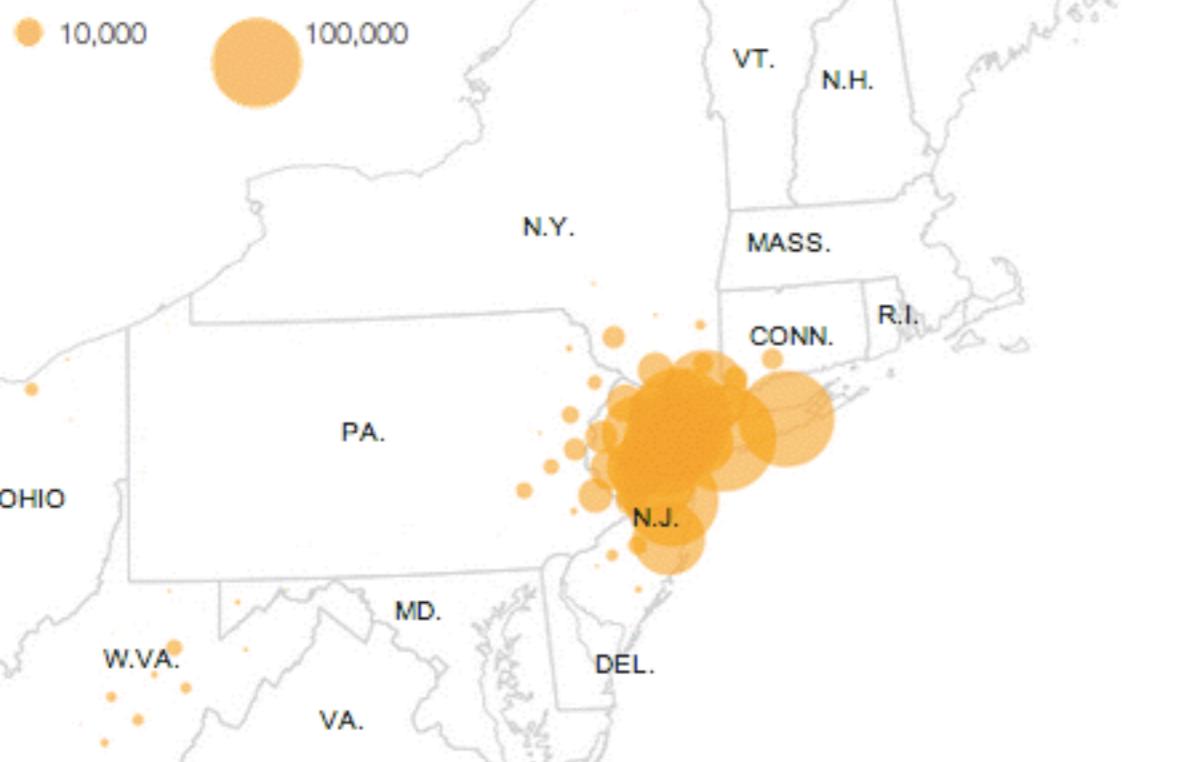
Updated Sunday, November 4 at 8:00 PM

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

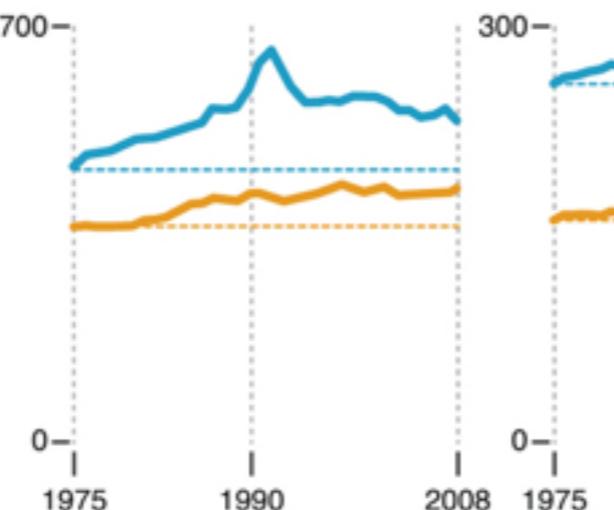
Cancer rates are up, but mortality is down. New diagnostics and treatments are responsible for part of this trend. But the greatest single contributing factor is the decline in smoking—rates are at their lowest level in 50 years.

Men      Women

## 1 Increased incidence

An aging population contributes to rising incidence of cancer.

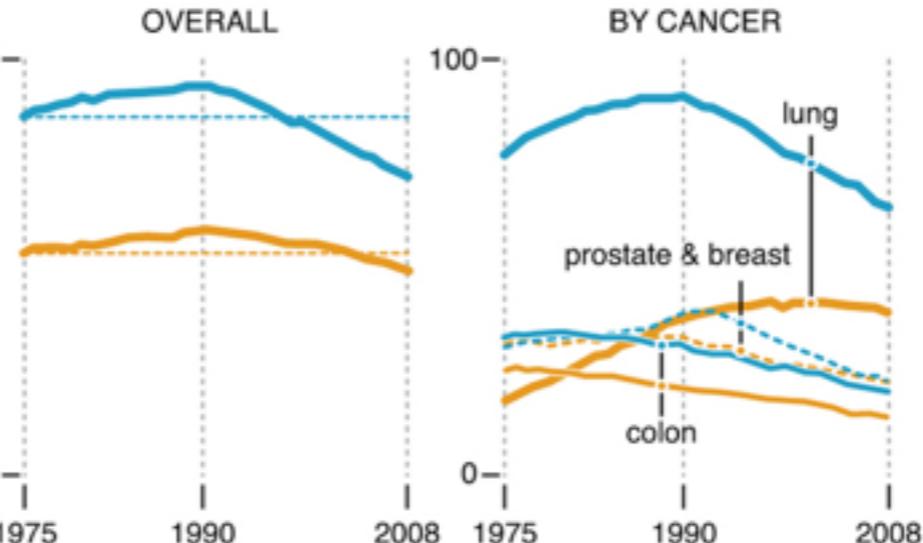
Cancer incidence rates (per 100,000)



## 2 Fewer deaths

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

Cancer death rates (per 100,000)



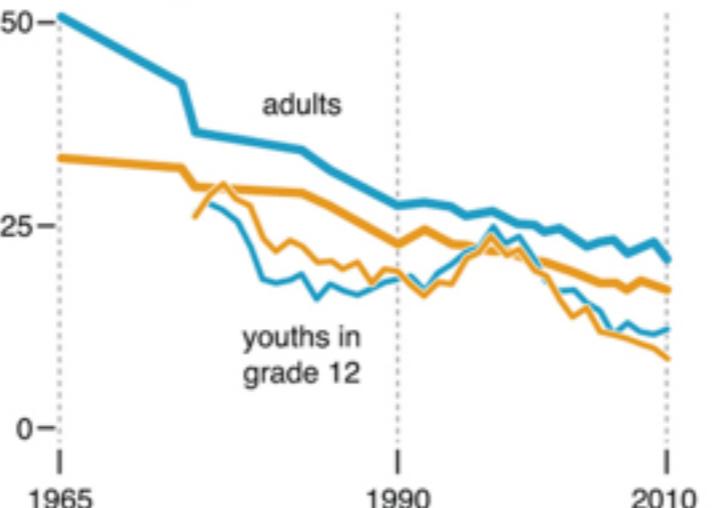
## 3 Decline of lung cancer

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.

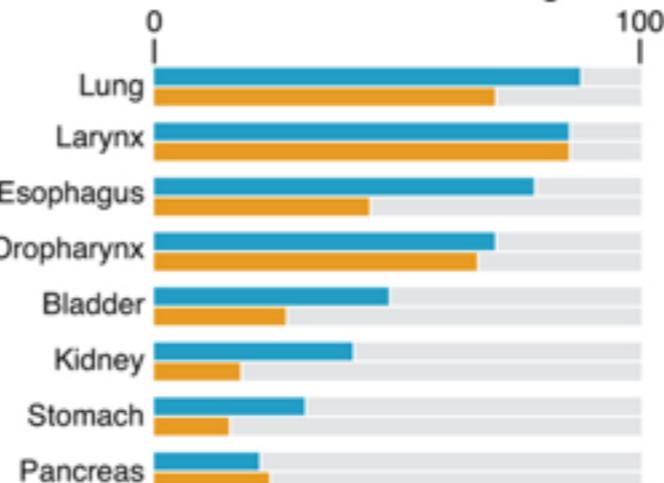
Smoking prevalence (%)



## 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

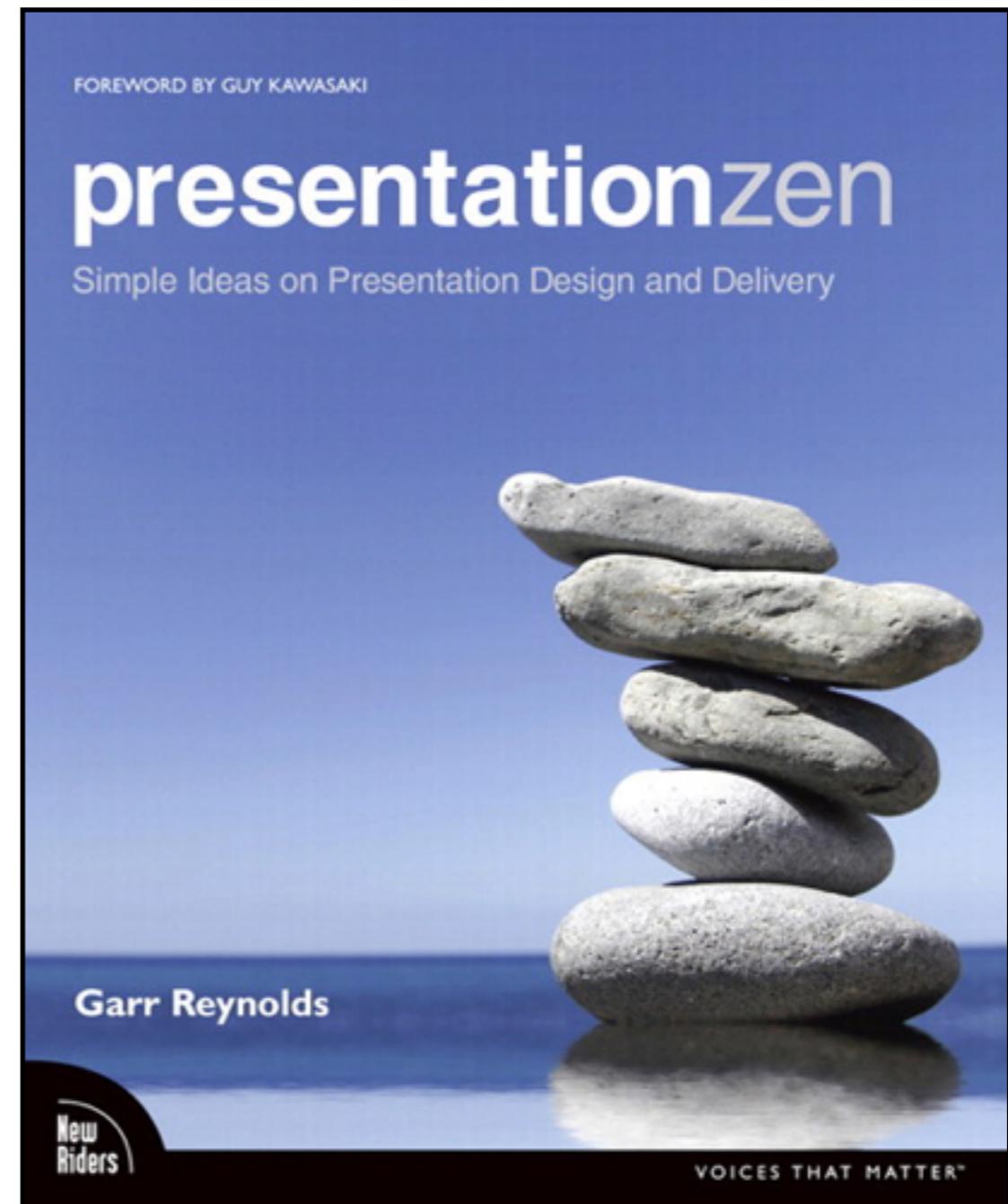
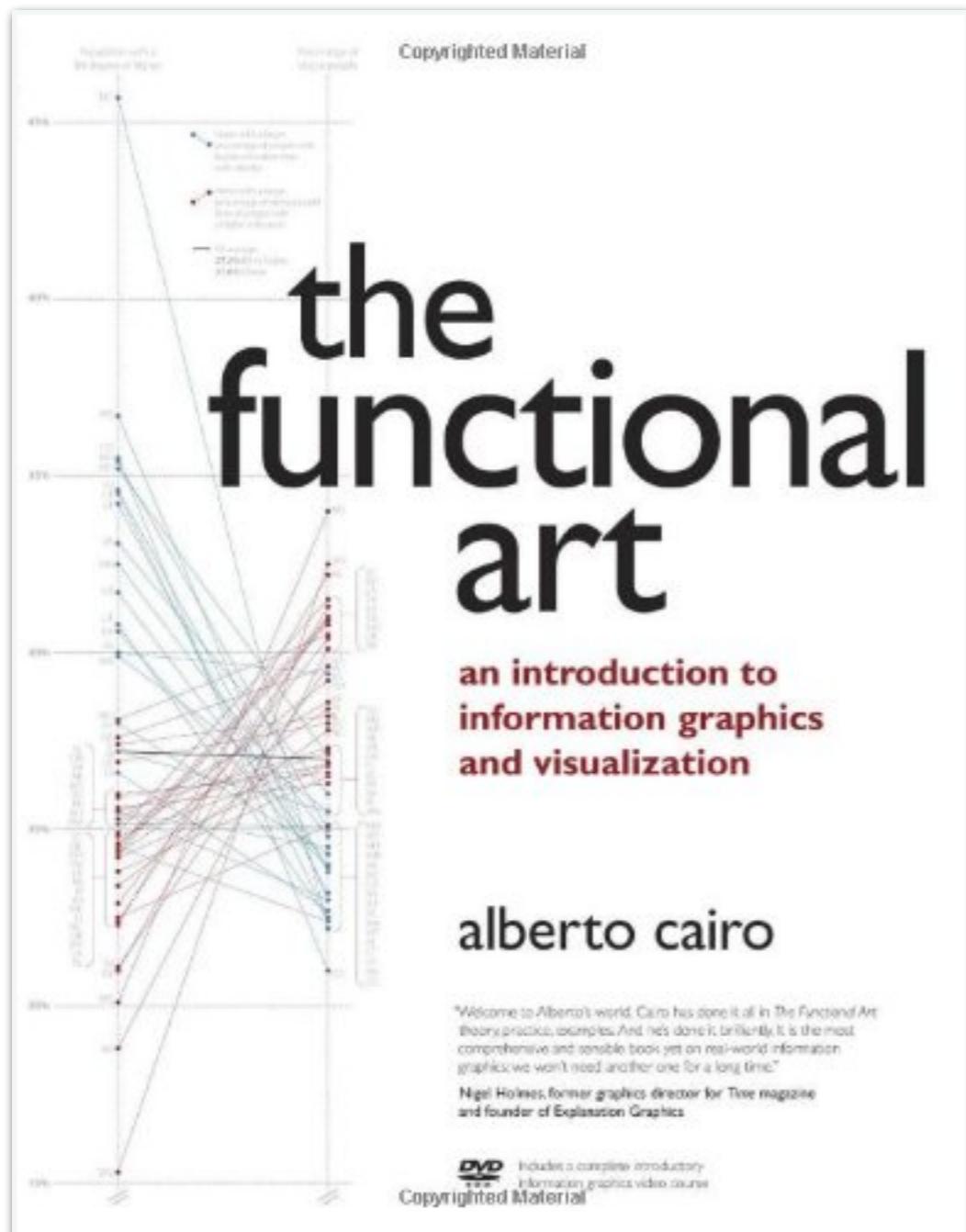


# Successful Data Stories...

- ....target the audience
- ...engage and are memorable      memorable stories
- ...answer concise questions
- ...are carefully designed
- ...move us to want to change the world  
add emotional impact to data visualization

Fewer words, more pictures

# Further Reading



# Further Reading

