Visualization for Communication

cs109a

(CNN)

the previous day…

MTI ASSESSMENT OF TEMPERATURE CONCERN ON SRM-25 (51L) LAUNCH

- 0 CALCULATIONS SHOW THAT SRM-25 O-RINGS WILL BE 20° COLDER THAN SRM-15 O-RINGS
- 0 TEMPERATURE DATA NOT CONCLUSIVE ON PREDICTING PRIMARY O-RING BLOW-BY
- 0 ENGINEERING ASSESSMENT IS THAT:
  - 0 COLDER O-RINGS WILL HAVE INCREASED EFFECTIVE DUROMETER ("HARDER")
  - 0 "HARDER" O-RINGS WILL TAKE LONGER TO "SEAT"
    - 0 MORE GAS MAY PASS PRIMARY O-RING BEFORE THE PRIMARY SEAL SEATS (RELATIVE TO SRM-15)
      - 0 DEMONSTRATED SEALING THRESHOLD IS 3 TIMES GREATER THAN 0.038" EROSION EXPERIENCED ON SRM-15
  - 0 IF THE PRIMARY SEAL DOES NOT SEAT, THE SECONDARY SEAL WILL SEAT
    - 0 PRESSURE WILL GET TO SECONDARY SEAL BEFORE THE METAL PARTS ROTATE
      - 0 O-RING PRESSURE LEAK CHECK PLACES SECONDARY SEAL IN OUTBOARD POSITION WHICH MINIMIZES SEALING TIME
- 0 MTI RECOMMENDS STS-51L LAUNCH PROCEED ON 28 JANUARY 1986
  - 0 SRM-25 WILL NOT BE SIGNIFICANTLY DIFFERENT FROM SRM-15

JOE C. KILMINSTER, VICE PRESIDENT
SPACE BOOSTER PROGRAMS

MORTON THIOKOL INC.
Wasatch Division

INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

(PCSSCA)

Less than 1 second after ignition, a puff of smoke appeared at the aft joint of the right booster, indicating that the O-rings burned through and failed to seal. At this point, all was lost.

On the launch pad, the leak lasted only about 2 seconds and then apparently was plugged by putty and insulation as the shuttle rose, flying through rather strong cross-winds. Then 58.788 seconds after ignition, when the Challenger was 6 miles up, a flicker of flame emerged from the leaky joint. Within seconds, the flame grew and engulfed the fuel tank (containing liquid hydrogen and liquid oxygen). That tank ruptured and exploded, destroying the shuttle.

As the shuttle exploded and broke up at approximately 73 seconds after launch, the two booster rockets crisscrossed and continued flying wildly. The right booster, identifiable by its failure plume, is now to the left of its non-defective counterpart.

The flight crew of Challenger 51-L. Front row, left to right: Michael J. Smith, pilot; Francis R. (Dick) Scobee, commander; Ronald E. McNair. Back row: Ellison S. Onizuka, S. Christa McAuliffe, Gregory B. Jarvis, Judith A. Resnik.

(VST, Tufte)

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36°--66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

Engineer deck, the previous day…

BLOW BY HISTORY

SRM-15 WORST BLOW-BY
- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY
- 2 CASE JOINTS (30-40°)

SRM-13A, 15, 16A, 18, 23A 24A
- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES-F)

MOTOR	MBT	AMB	O-RING	WIND
DM-4	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

(PCSSCA)

CONCLUSIONS:

- TEMPERATURE OF O-RING IS NOT ONLY PARAMETER CONTROLLING BLOW-BY

  SRM 15 WITH BLOW-BY HAD AN O-RING TEMP AT 53°F
  SRM 22 WITH BLOW-BY HAD AN O-RING TEMP AT 75°F
  FOUR DEVELOPMENT MOTORS WITH NO BLOW-BY WERE TESTED AT O-RING TEMP OF 47° TO 52°F

  DEVELOPMENT MOTORS HAD PUTTY PACKING WHICH RESULTED IN BETTER PERFORMANCE

- AT ABOUT 50°F BLOW-BY COULD BE EXPERIENCED IN CASE JOINTS

- TEMP FOR SRM 25 ON 1-28-86 LAUNCH WILL BE 29°F 9 AM
  38°F 2 PM

- HAVE NO DATA THAT WOULD INDICATE SRM 25 IS DIFFERENT THAN SRM 15 OTHER THAN TEMP

RECOMMENDATIONS:

- O-RING TEMP MUST BE ≥ 53°F AT LAUNCH

  DEVELOPMENT MOTORS AT 47° TO 52°F WITH PUTTY PACKING HAD NO BLOW-BY
  SRM 15 (THE BEST SIMULATION) WORKED AT 53°F

- PROJECT AMBIENT CONDITIONS (TEMP & WIND) TO DETERMINE LAUNCH TIME

(PCSSCA)

[redacted]

Flight	Date	Temperature °F	Erosion incidents	Blow-by incidents	Damage index	Comments
51-C	01.24.85	53°	3	2	11	Most erosion any flight; blow-by; back-up rings heated.
41-B	02.03.84	57°	1		4	Deep, extensive erosion.
61-C	01.12.86	58°	1		4	O-ring erosion on launch two weeks before Challenger.
41-C	04.06.84	63°	1		2	O-rings showed signs of heating, but no damage.
1	04.12.81	66°			0	Coolest (66°) launch without O-ring problems.
6	04.04.83	67°			0	
51-A	11.08.84	67°			0	
51-D	04.12.85	67°			0	
5	11.11.82	68°			0	
3	03.22.82	69°			0	
2	11.12.81	70°	1		4	Extent of erosion not fully known.
9	11.28.83	70°			0	
41-D	08.30.84	70°	1		4	
51-G	06.17.85	70°			0	
7	06.18.83	72°			0	
8	08.30.83	73°			0	
51-B	04.29.85	75°			0	
61-A	10.30.85	75°		2	4	No erosion. Soot found behind two primary O-rings.
51-I	08.27.85	76°			0	
61-B	11.26.85	76°			0	
41-G	10.05.84	78°			0	
51-J	10.03.85	79°			0	
	06.27.82	80°			?	O-ring condition unknown; rocket casing lost at sea.
51-F	07.29.85	8				

RISK ASSESSMENT?

(VST, Tufte)

Chartjunk at hearings

(PCSSCA)

Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Visualization Goals

Communicate (Explanatory)

Present data and ideas

Explain and inform

Provide evidence and support

Influence and persuade

Analyze (Exploratory)

Explore the data

Assess a situation

Determine how to proceed

Decide what to do

Communicate

Napoleon's March to Russia

Next, the temperature experienced by his troops when winter settled in on the return trip.

https://robots.thoughtbot.com/analyzing-minards-visualization-of-napoleons-1812-march

Minard's Graphic on Napoleon's Russia Campaign

(from wikipedia)

Minard's Graphic on Napoleon's Russia Campaign

(from wikipedia)

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** to you want to affect?

Effective Visualizations

1. Have graphical integrity
2. Keep it simple
3. Use the right display
4. Use color strategically
5. Know your audience

WRONG

CITIZENS FOR
TRUMP
THE INSIDE STORY
OF THE PEOPLE'S MOVEMENT
TO TAKE BACK AMERICA
★ ★ ★ ★ ★
JACK POSOBIEC

RIGHT

COUNTIES FOR
TRUMP
THE INSIDE STORY
OF 46% OF VOTERS' MOVEMENT
TO TAKE BACK AMERICA
★ ★ ★ ★ ★
JACK POSOBIEC

The Persuasive Power of Data Visualization

Anshul Vikram Pandey
New York University

Anjali Manivannan
New York University

Oded Nov
New York University

Margaret L. Satterthwaite
NYU School of Law, satterth@exchange.law.nyu.edu

Enrico Bertini
New York University

After looking into common effects in attitude formation and change we searched for specific mentions to the graphical appearance of charts as a driver for persuasion. Some of the comments we collected seem to back up the findings we found in our results. Some participants explicitly mention the charts as being the main reason for their change: "*I already knew that increased incarceration didn't lower crime, but I wasn't sure of the statistics. To see it on the graphs is really eye opening.*"; "*I was influenced by the bar graph showing the reasons why the survey respondents played video games.*"; "*I would not know exact numbers on this issue - the graphs gave a visual and helped identify the numbers*"; "*Seeing the graphs conflicted with my previous opinion, so I feel like I need to reevaluate my stance in a way.*"

It is also important to mention that the graphical appearance of charts is not the only factor that has a strong impact on people's attitude. In our collected feedback, we found numerous references to statistics and numbers, suggesting that mere exposure to data does have a persuasive effect – maybe at least partially due to the increased sense of objectivity evidence supported by numbers carries. We found comments like: "*It was concrete data that seemed compelling.*; "*Seeing numbers is a good indicator of change rather than just reading what someone has to say*"; "*It showed a large amount of different sources, which made it more credible*". More research is needed to disentangle what kind of specific effects each of these components have on persuasion.

http://lsr.nellco.org/cgi/viewcontent.cgi?article=1476&context=nyu_plltwp

Keep it Simple

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

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

Cole Nussbaumer

Use the right display

C. Mulbrandon
VisualizingEconomics.com

Most Effective

VisualizingEconomics.com

Less Effective

VisualizingEconomics.com

Possible solution to cases when you have data that diverge a lot

Use color strategically

Colors for Categories

Do not use more than 5-8 colors at once

Ware, "Information Visualization"

Colors for Ordinal Data

Vary luminance and saturation

Zeilis et al, 2009, "Escaping RGBland: Selecting Colors for Statistical Graphics"

Colors for Quantitative Data

Hue
(Rainbow)

Luminance

Luminance
& Hue

Rogowitz and Treinish, Why should engineers and scientists be worried about color?

Rainbow Colormap

Rainbow Colormap

R. Simmon

Gray

Colormap evaluation: gray
The colormap in its glory
Black-and-white printed
Sample images
Moderate deuter.
Perceptual deltas (total: 100.01)
Perceptual lightness deltas (total: 100.00)
Moderate deuteranomaly
Moderate protanomaly
Complete deuteranopia
Complete protanopia
J'/K (white -> black)
a' (green -> red)
b' (blue -> yellow)
Toggle gamut

Color Blindness

Protanope Deuteranope Tritanope

Red / green deficiencies

Blue / Yellow deficiency

Based on slide from Stone

Color Blindness

Normal Protanope Deuteranope Lightness

Based on slide from Stone

Viridis

Colormap evaluation: option_d.py
The colormap in its glory
Perceptual deltas (total: 141.79)
Black-and-white printed
Perceptual lightness deltas (total: 73.63)
Sample images
Moderate deuter.
Moderate deuteranomaly
Moderate protanomaly
Complete deuteranopia
Complete protanopia
J'/K (white -> black)
b' (blue -> yellow)
a' (green -> red)
Toggle gamut

Color Brewer

Nominal

Ordinal

Cynthia Brewer, Color Use Guidelines for Data Representation

number of data classes on your map

3 learn more >

the nature of your data

sequential learn more >

pick a color scheme: GnBu

multihue

single hue

(optional) only show schemes that are:

- [x] colorblind safe
- [ ] print friendly
- [ ] photocopy-able

learn more >

pick a color system

224, 243, 219

168, 221, 181

67, 162, 202

(•) RGB ( ) CMYK ( ) HEX

adjust map context

- [ ] roads
- [ ] cities
- [x] borders

select a background

(•) solid color

( ) terrain

color transparency

learn more >

how to use | updates | credits

COLORBREWER 2.0

color advice for cartography

EXPORT YOUR COLORS >>

SCORE CARD

© Cynthia Brewer, Mark Harrower and The Pennsylvania State University

Support

Back to ColorBrewer 1.0

axm

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”?

What is the message?

Exploratory
Neutral

Explanatory
Opinionated

Andy Cotgreave, Tableau

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!

Gun Deaths in 2010

Tools for interactive graphics

- R/shiny
- plotly/dash
- Tableau
- d3.js
- vega-lite/vega

Is there a story?

Surface it….even if it is incomplete

2014 Gun Deaths

PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

(XKCD)

Figure 1: Percent of murders
Percent of Murders
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
1%
3%
5%
10%
12%
68%
50%
Next 10%
Next 10%
Next 10%
Next 10%
Next 5%
Last 5%
Safest 70% of the Counties
Next 25% of Counties
Most Dangero Countie
Percent of Counties
Crime Prevention Research Center
crimeresearch.org

Los Angeles
Pasadena
Glendale
Burbank
Beverly Hills
Santa Monica
Alhambra
Rosemead
El Monte
West Covina
Baldwin Park
Covina
Azusa
Glendora
Pomona
Chino
Montebello
Pico Rivera
Whittier
Huntington Park
South Gate
Bell Gardens
Downey
Norwalk
La Mirada
La Habra
Brea
Fullerton
Placentia
Yorba Linda
Anaheim
Orange
Santa Ana
Tustin
Buena Park
Cerritos
Lakewood
Long Beach
Torrance
Carson
Gardena
Hawthorne
Inglewood
Los Angeles International Airport
Redondo Beach
Manhattan Beach
El Segundo
Cypress
Stanton
Fountain Valley
Seal Beach
Signal Hill
Lomita
Rolling Hills
Rancho Palos Verdes
Hacienda Heights
Rowland Heights
Walnut
Claremont
Montclair
San Dimas
La Verde
Sierra Madre
Monrovia
Altadena
La Cañada Flintridge
La Crescenta-Montrose
Hidden Springs
Angeles National Forest
San Gabriel Mountains
Falling Springs
Kagel Canyon
San Fernando
Deer Lake Highlands
Topanga State Park
Chino Hills State Park
Limestone Canyon Regional Park
Santiago Canyon
North Tustin
Villa Park
Orange Park Acres
El Modena
Map data ©2017 Google
Terms of Use

Per capita sales by county

Annual gun sales per 100,000 residents are generally higher in California's rural and northern counties.

Local counties

Placer	3,693
Sacramento	3,196
Sutter	2,571
Yuba	2,430
El Dorado	1,741
Yolo	600

Source: California Department of Justice

The Sacramento Bee

Deaths by county, 2014

(crimeresearch.org)

Whether Crime Is Up or Down Depends on Data Being Used

By TIMOTHY WILLIAMS SEPT. 27, 2016

RELATED COVERAGE

U.S. Murders Surged in 2015, F.B.I. Finds
SEPT. 26, 2016

The news from the F.B.I. crime data was alarming: The murder rate rose sharply last year, driven by jumps in several major cities.

Four urban areas — Baltimore, Chicago, Milwaukee and Washington — accounted for about a fifth of the increase in homicides in 2015. Those cities, however, make up only about 1 percent of the nation's population.

But whether crime is up or down depends on what data is being looked at — and who is doing the looking.

The F.B.I. data showed that violent crime rose about 4 percent last year from 2014, and homicides increased 10.8 percent. Yet crime over all fell in 2015 for the 14th consecutive year.

And the total number of homicides last year was fewer than 20 years ago even as the country's population increased, criminologists said. There were 19,645 homicides in 1996 in a nation of 265 million; in 2015, there were 15,696 in a population of 321 million.

What that data means, criminologists and police officials said, is that the decline in homicides has been so significant in the last quarter century that sudden increases in the number of killings in just a few cities can skew the entire national picture, even as the country has one of its safest periods on record.

"It isn't a national trend, it's a city trend, and it's not even a city trend, but a problem in certain neighborhoods," said Richard A. Berk, a professor of statistics and criminology at the University of Pennsylvania. "Certainly, people around the country should not be worried. People in Chicago shouldn't be worried. But people in certain neighborhoods might be."

Criminologists and police officials point out that homicides do not usually disrupt entire cities. Instead, they occur in particular neighborhoods — and on the same blocks — leaving much of the rest of the city relatively untouched.

Explanations for the increase in homicides in certain American cities are largely guesswork. Criminologists acknowledge that the required analysis has not been done in the neighborhoods where killings are occurring — or even an agreement of what such a study should include — to arrive at any but the broadest conclusions.

Careful with amalgamation paradoxes and with outliers

http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00513/full

Ask Ask Ask

- Is the exact distribution of guns really the important concern?
- did we check the uncertainties?
- Should we be looking at this from a “risk” perspective?
- we tend to believe what we believe and look for confirmation.
- we need to be disciplined about interrogating ourselves
- it is ok (and not against simplicity) to surface our process

a woman's age vs. the age of the men who look best to her

(from Dataclysm)

a man's age vs. the age of the women who look best to him

20	20
21	20
22	21
23	21
24	21
25	21
26	22
27	21
28	20
29	20
30	20
31	20
32	20
33	20
34	20
35	20
36	20
37	22
38	20
39	20
40	21
41	21
42	20
43	23
44	21
45	24
46	20
47	20
48	23
49	20
50	22

Sample of 100 men of 40
vs. the age of the women
who look best to them

Sample of 100 men of 40
vs. the age of the women
who look best to them

▬ =1 of men

Structure of communication graphics

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 tie 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 was injured last season. He played 14 games and hit 6 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

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

BEGINNING

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

Bonds was injured last season. He played 14 games and hit 6 homers.

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).

Player	Homers	Note
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 from 1960-69.
Sammy Sosa	588	Three 60-homer seasons is record.
Frank Robinson	586	Triple Crown in '66 (49, 122, .316).
Mark McGwire	583	First to hit 70 in a season.
Ken Griffey Jr.	534	Only McGwire had more in the 90's.
Alex Rodriguez	429	Youngest to reach 400 homers.
Albert Pujols	261	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 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 19 seasons
Bonds 567
Aaron 554
Ruth 516
600
200
0
5 seasons
10
15
20
714
755
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 6 homers.
BEGINNING
14th season
According to allegations in a book about Bonds, he began taking steroids before the 1999 season. He 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.
ACTUAL
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.
429 HR
MIDDLE
Albert Pujols
Averaging 49 homers a season, he has started stronger than the three leaders did.
201 HR
Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.
536 HR
SEASONS
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).
Hank Aaron
755
Babe Ruth
714
Barry Bonds
708
Willie Mays
660
Sammy Sosa
588
Frank Robinson
586
Mark McGwire
583
Ken Griffey Jr.
536
Alex Rodriguez
429
Albert Pujols
201
END
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

Headline
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 tie 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
5 seasons
10
15
14th season
According to allegations in a book about Bonds, he began taking steroids in the 1998 season. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.
755
714
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 6 homers.
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 [illegible] 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.
ACTUAL
AGE
Note: Ages as of July 1 of each season.
Others Taking Aim
Alex Rodriguez
Is ahead of the pace set by all three [illegible] 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.
420 HR
201 HR
526 HR
Aaron, Ruth and Bonds
SEASONS
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).
Hank Aaron 755
Babe Ruth 714
Barry Bonds 708
Willie Mays 660
Sammy Sosa 588
Frank Robinson 586
Mark McGwire 583
Ken Griffey Jr. 526
Alex Rodriguez 429
Albert Pujols 201
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 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.
Annotations
Call Out Boxes
Captions

E. Segel

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

Application to modeling

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 understand how it is working? Do we need to iterate and improve the model? What are the limitations and future directions?

(from Foster and Fawcett)

Breast Cancer on a Mammogram

- False positives OK
- False Negatives are disaster
- More people dont have it

Predicted

Observed		0	1	
	0	TN True Negative	FP False Positive	ON Observed Negative
	1	FN False Negative	TP True Positive	OP Observed Positive
		PN Predicted Negative	PP Predicted Positive	

Which Model is Better?

Communicating a model

Telecom Churn Problem

Survey 1000 customers , with an offer with an administrative cost of $3 and an offer cost of $100, an incentive for the customer to stay with us.

Want to predict for our 100000 customer base.

If a customer leaves us, we lose the customer lifetime value, which is some kind of measure of the lost profit from that customer.

Lets assume this is the average number of months a customer stays with the telecom times the net revenue from the customer per month. We'll assume 3 years and $30/month margin per user lost, for roughly a $1000 loss.

```
admin_cost=3
offer_cost=100
clv=1000 # customer lifetime value
```

- TN=people we predicted not to churn who wont churn. We associate no cost with this as they continue being our customers
- FP=people we predict to churn. Who wont. Lets associate a `admin_cost+offer_cost` cost per customer with this as we will spend some money on getting them not to churn, but we will lose this money.
- FN=people we predict wont churn. And we send them nothing. But they will. This is the big loss, the `clv`
- TP= people who we predict will churn. And they will. These are the people we can do something with. So we make them an offer. Say a fraction f accept it. Our cost is `admin_cost + f*offer_cost + (1-f)*clv`.

```
f = 0.5
tnc = 0.
fpc = admin_cost+offer_cost
fnc = clv
tpc = admin_cost + f*offer_cost + (1. - f)*clv
```

Predicted

Observed	0	1	
0	TN True Negative	FP False Positive	ON Observed Negative
1	FN False Negative	TP True Positive	OP Observed Positive
	PN Predicted Negative	PP Predicted Positive	

Predicted

Observed	0	1
0	TNC True Negative Cost	FPC False Positive Cost
1	FNC False Negative Cost	TPC True Positive Cost

Average Cost = TN x TNC
+ TP x TPC +
FN x FNC + TP x TPC

Predicted

Observed	0	1
0	TNC 0 True Negative Cost	FPC 103 False Positive Cost
1	FNC 1000 False Negative Cost	TPC 553 True Positive Cost

170
160
150
140
130
120
110
1.0
0.91
0.59
0.27
0.06
cost curve for gnb
0
20
40
60
80
100

Annotated Diagram

Loss

made with Preview

Reduce churn and our cost by sending customers an offer

Loss

Making offers within Budget

This study was made on a pilot survey of 1000 customers from our 100000 customer base.

Make an offer with an **administrative cost of $3** and an **offer cost of $100**, an incentive for the customer to stay with us.

If a customer leaves us, we lose the customer lifetime value (CLV), a **roughly $1000 loss**.

We assume that 50% of those customers targeted will stay with us.

If we do nothing we lose $150 per customer including CLV

We ***choose which customers to target*** according to 2 different models, **dt** and **gnb:**

- Making an offer to 13% of our most likely to leave customers will cut this cost to a lowest value of $103 per customer according to the **dt** model, for a total cost of $1.34 million.
- If we only target 10% of the customers (Budget 1) using the **dt** model, we get by in 1.03 million but incur a loss of $110 per customer including CLV.
- If we target 40% of our customers, we need a budget (Budget 2) of $4.2 million. Here the **gnb** model performs better and we will choose customers according to it. We incur a loss of $116 per customer including CLV.

made with Pages

StoryTelling

Edward Tufte

Stephen Few

Show Me the Numbers
Designing Tables and Graphs to Enlighten
Stephen Few

INFORMATION
DASHBOARD
DESIGN
The Effective Visual Communication of Data
Stephen Few
O'REILLY

Now you see it
Stephen Few

2012

2016

I've always believed in the power of data visualization (the representation of information by means of charts, diagrams, maps, etc.) to enable understanding