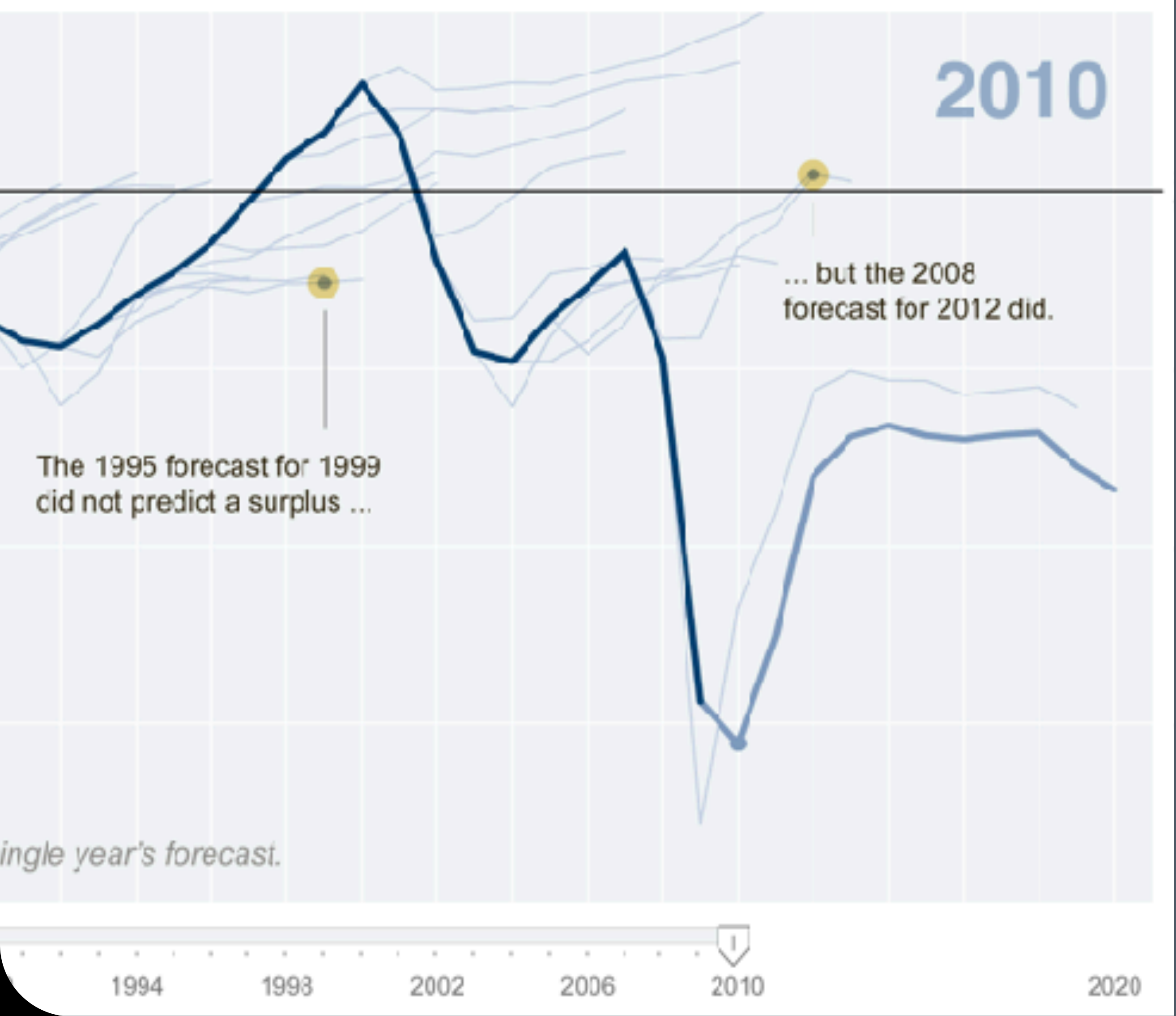
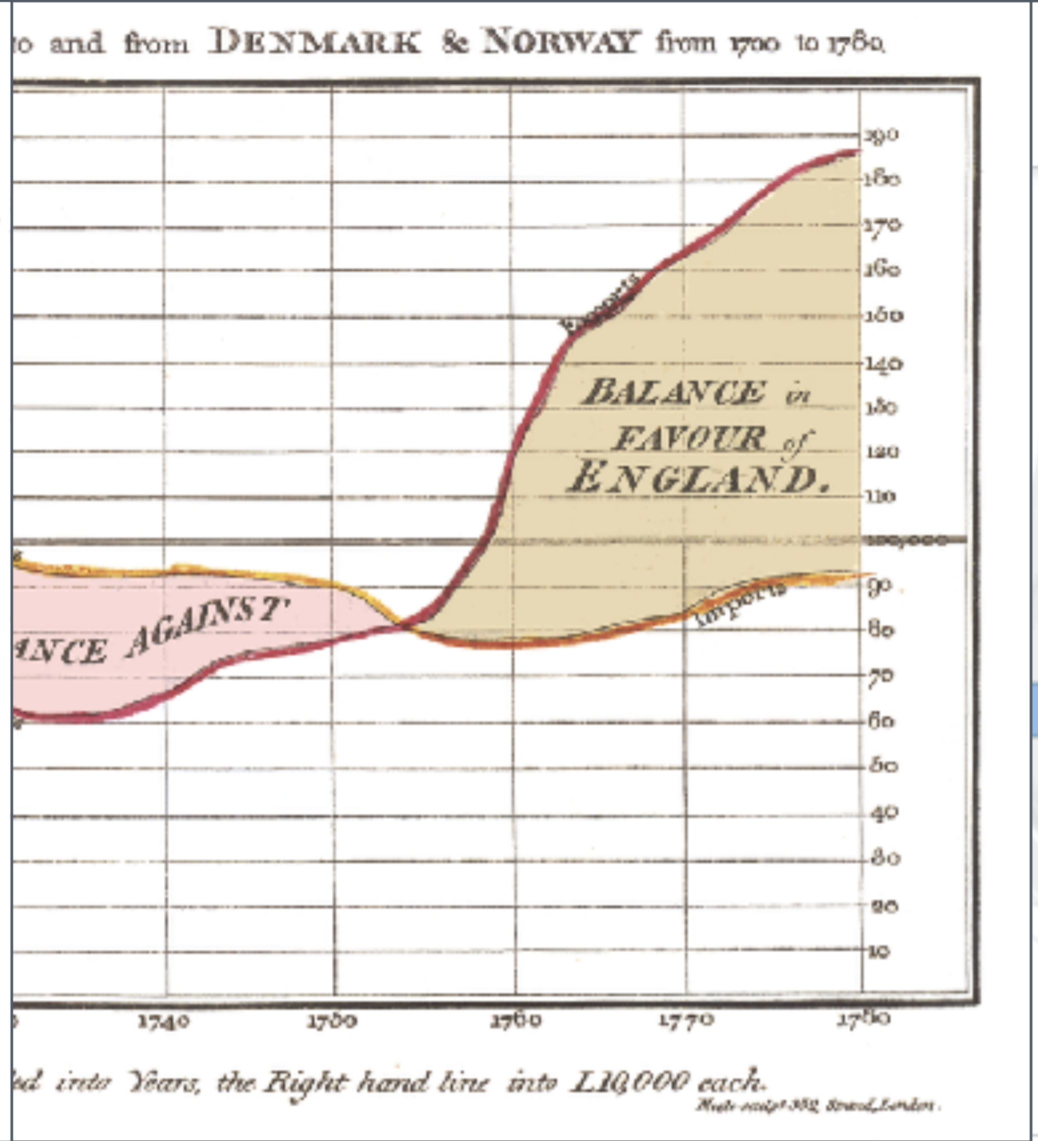
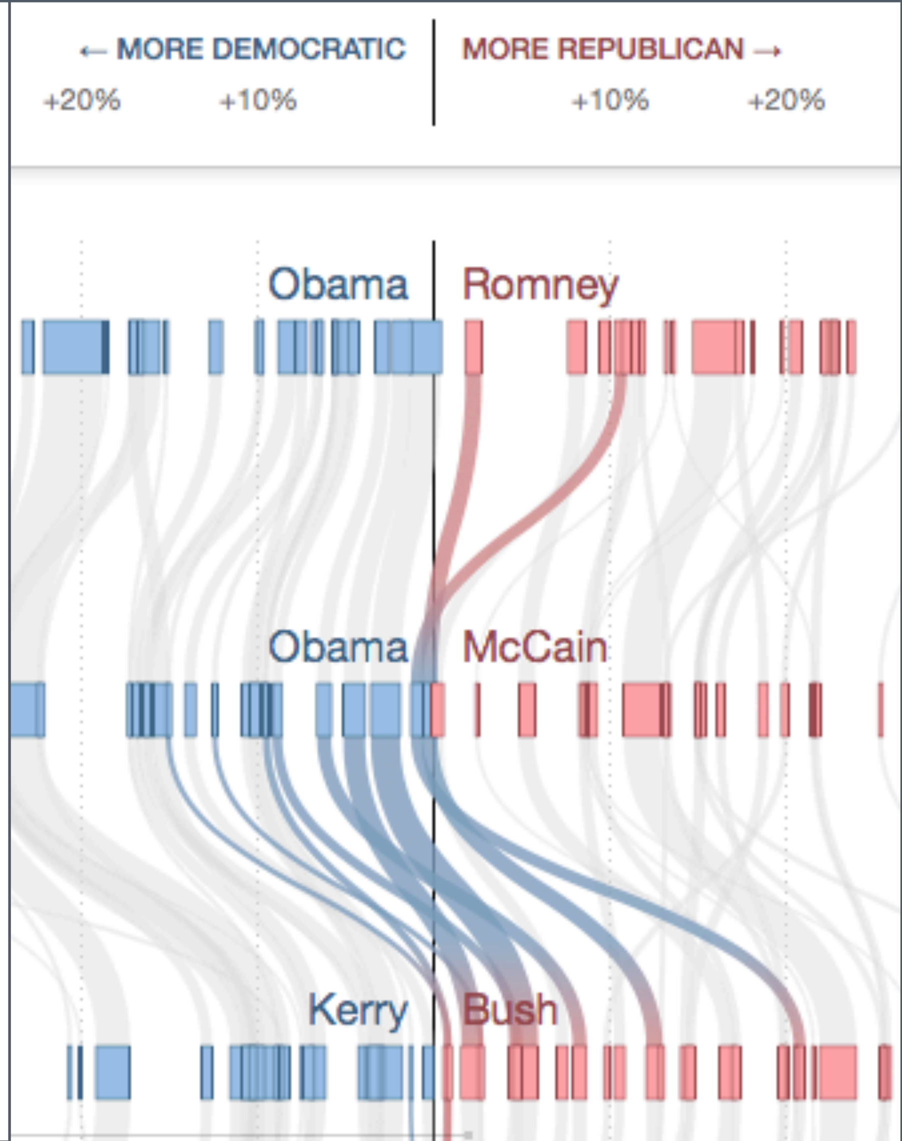
6.859: Interactive Data Visualization

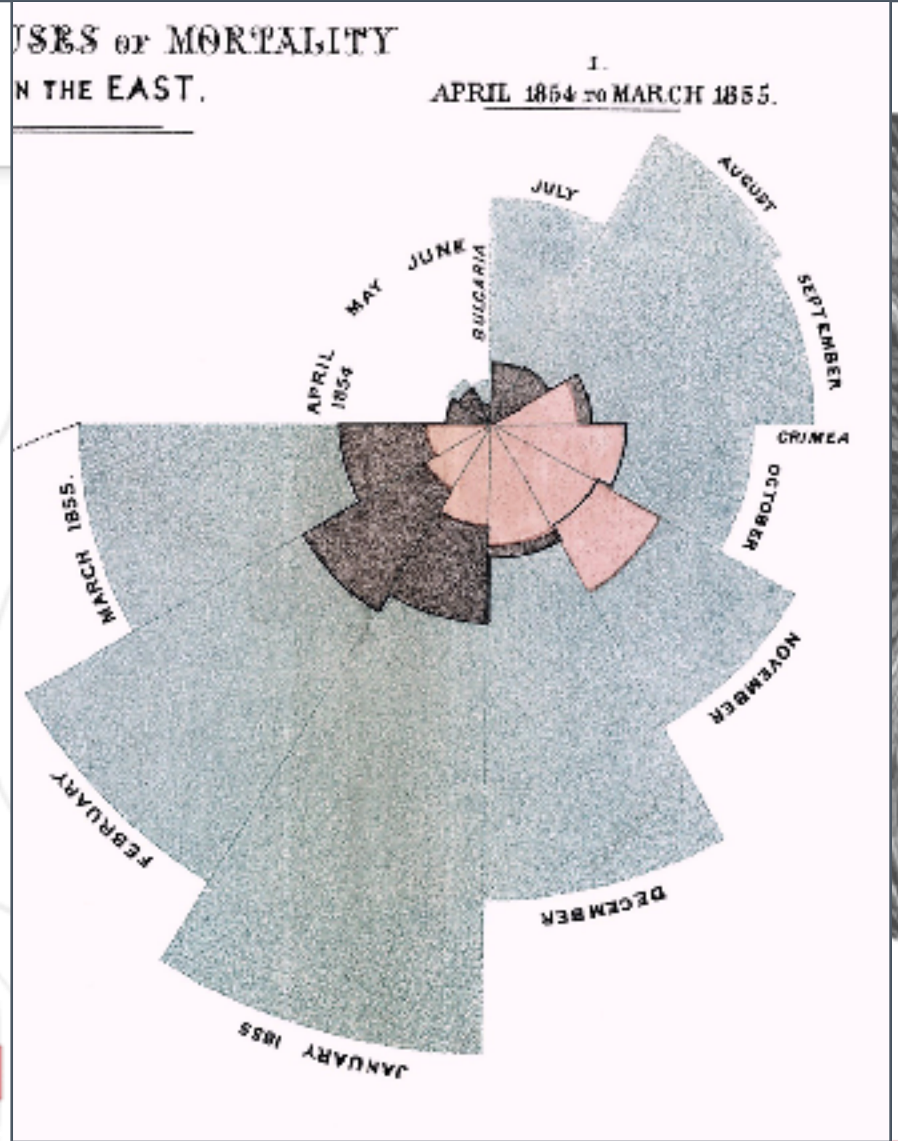
Exploratory Data Analysis

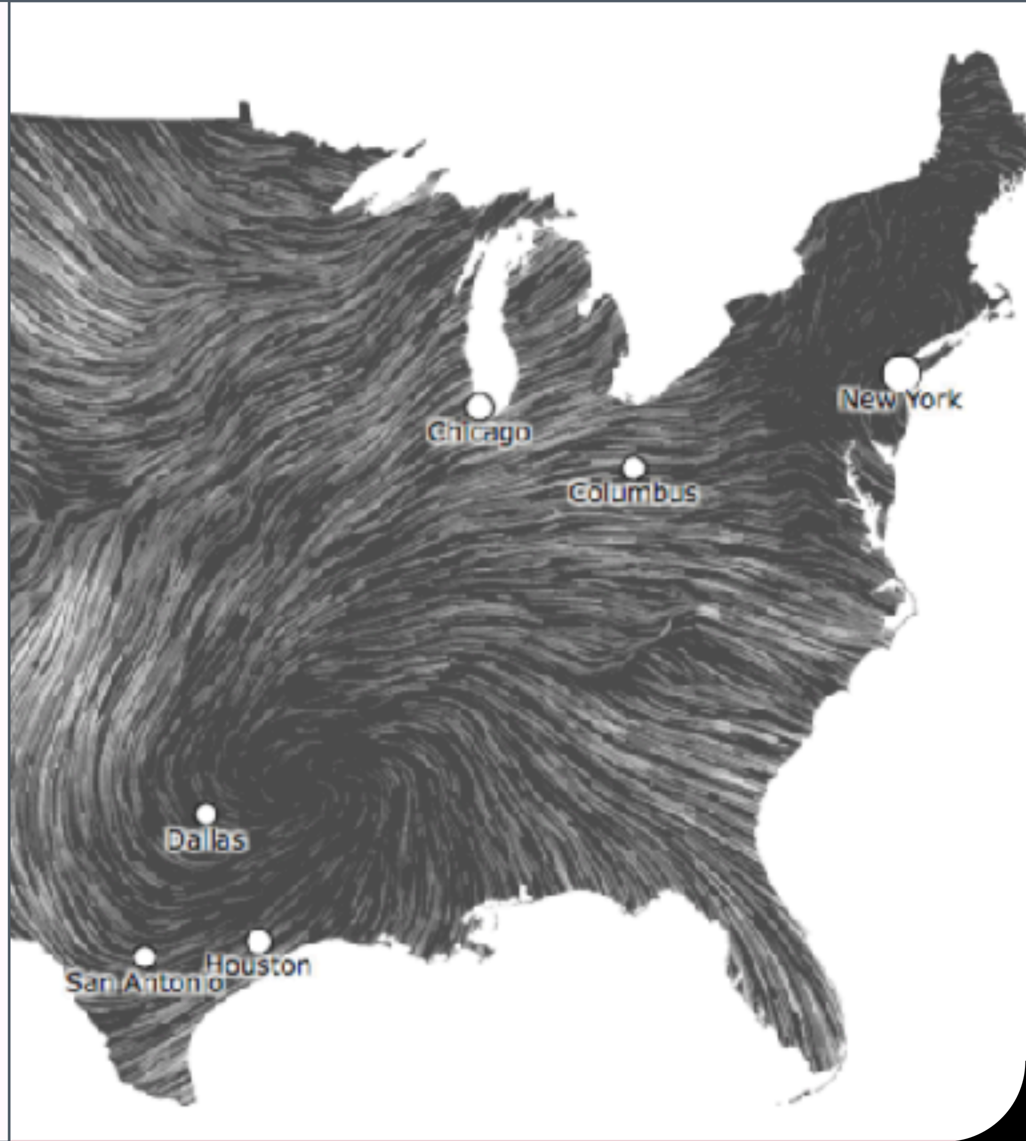
Arvind Satyanarayan











Course Grading

Class Participation	5%
Reading Commentaries	5%
A0: Sketching Visualizations	2%
A1: Visualization Design	3%
A2: Exploratory Data Analysis	10%
A3: White/Black Hat Visualization	15%
A4: Interactive Narratives	20%
Final Project	40%
Proposal	
MVP + Presentations	
Poster Session + Final Deliverables	

5 slack days which can be used as you wish for assignments.

Slack days should cover minor illnesses, special occasions (including religious holidays).

Additional extensions only granted for serious issues with a written note of support from S3 or GradSupport @ OGE.

Share your work on Slack to inspire your classmates + receive design feedback!

Mapping or Visual Encoding

Data → **Visual**

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express *all the facts in the set of data, and only the facts in the data*.

Data models give us a way of talking about this.

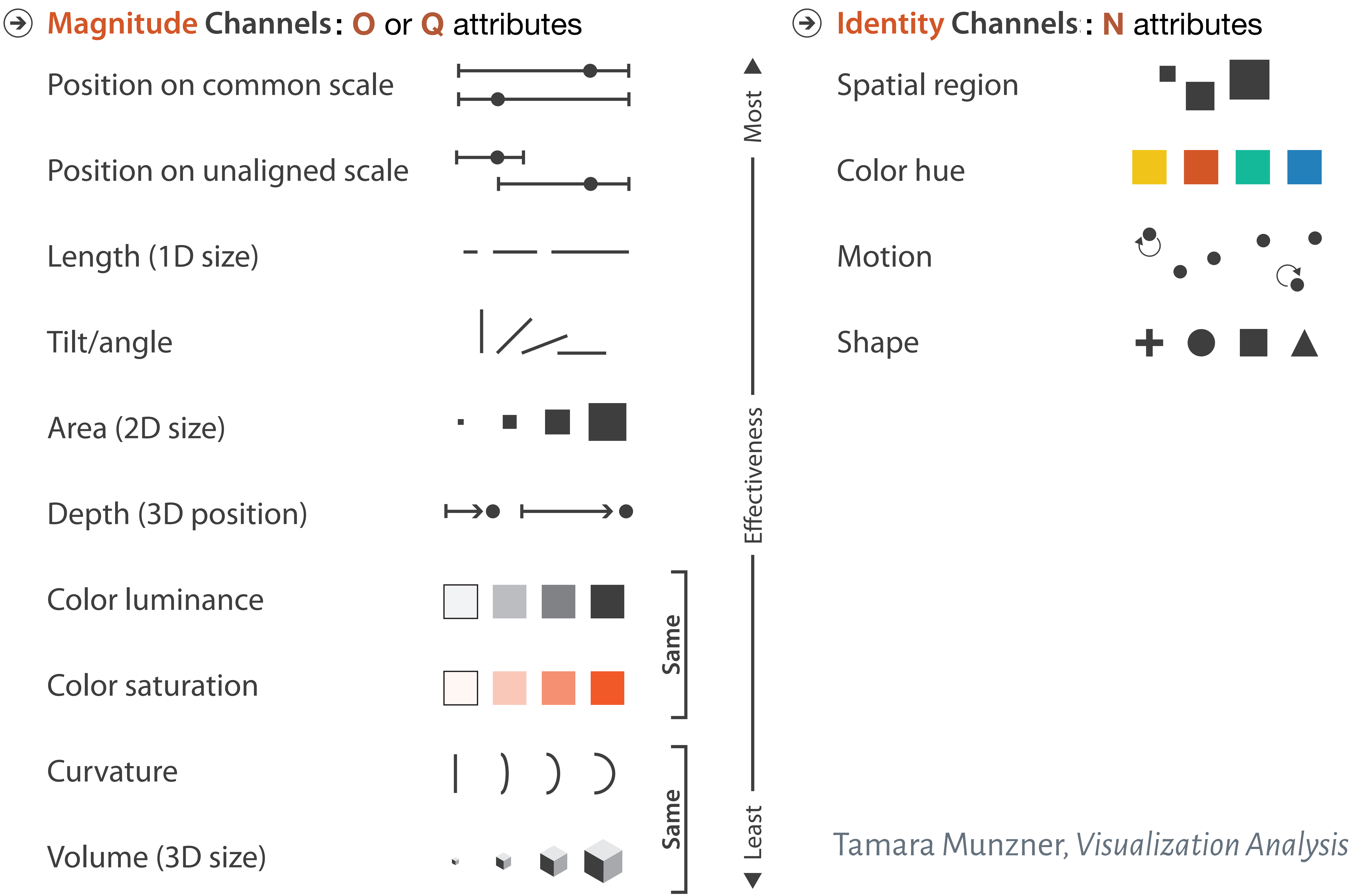
Effectiveness

A visualization is more *effective* than another if the information it conveys *is more readily perceived* than the information in the other visualization

Image models give us a way of talking about this.

[Mackinlay 1986]

Channels: Expressiveness Types and Effectiveness Ranks



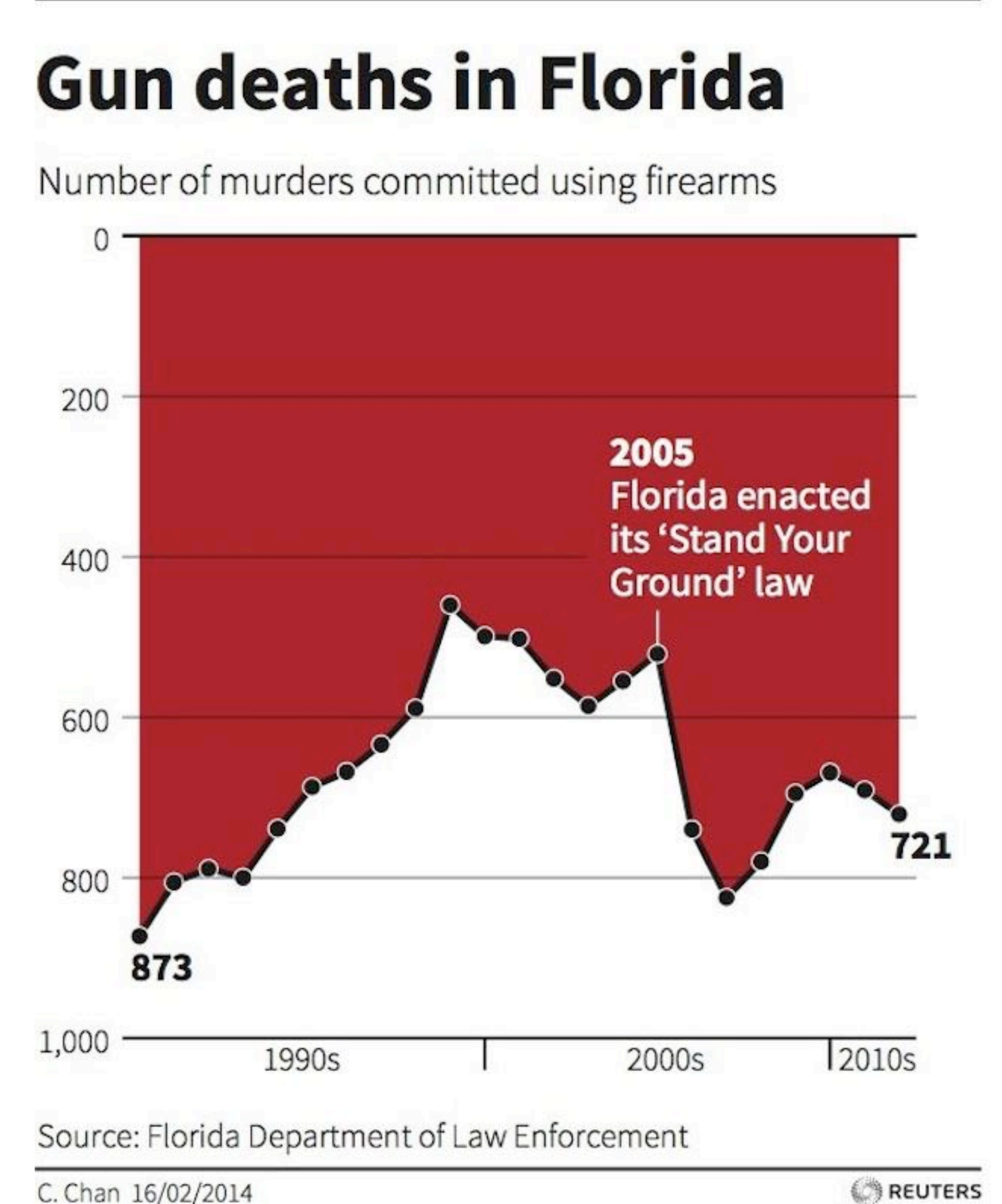
Tamara Munzner, *Visualization Analysis and Design* (2014).

Visualization Critique

What is this a visualization of?

Which of the visual encoding techniques we've discussed this week are being used? How effective or ineffective are they?

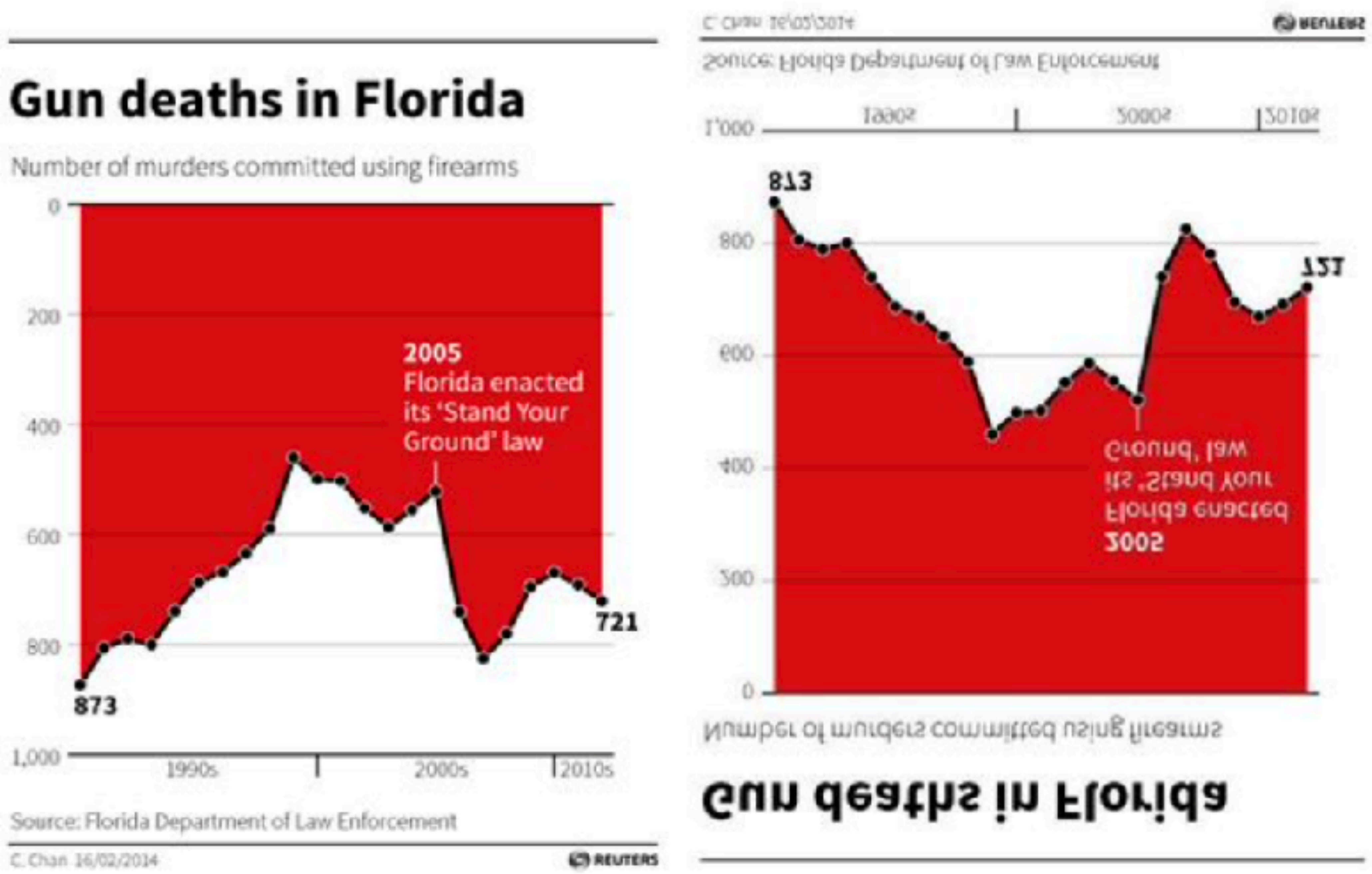
To help structure your critique:
- *"I like..."*
- *"I wish..."*
- *"What if...?"*

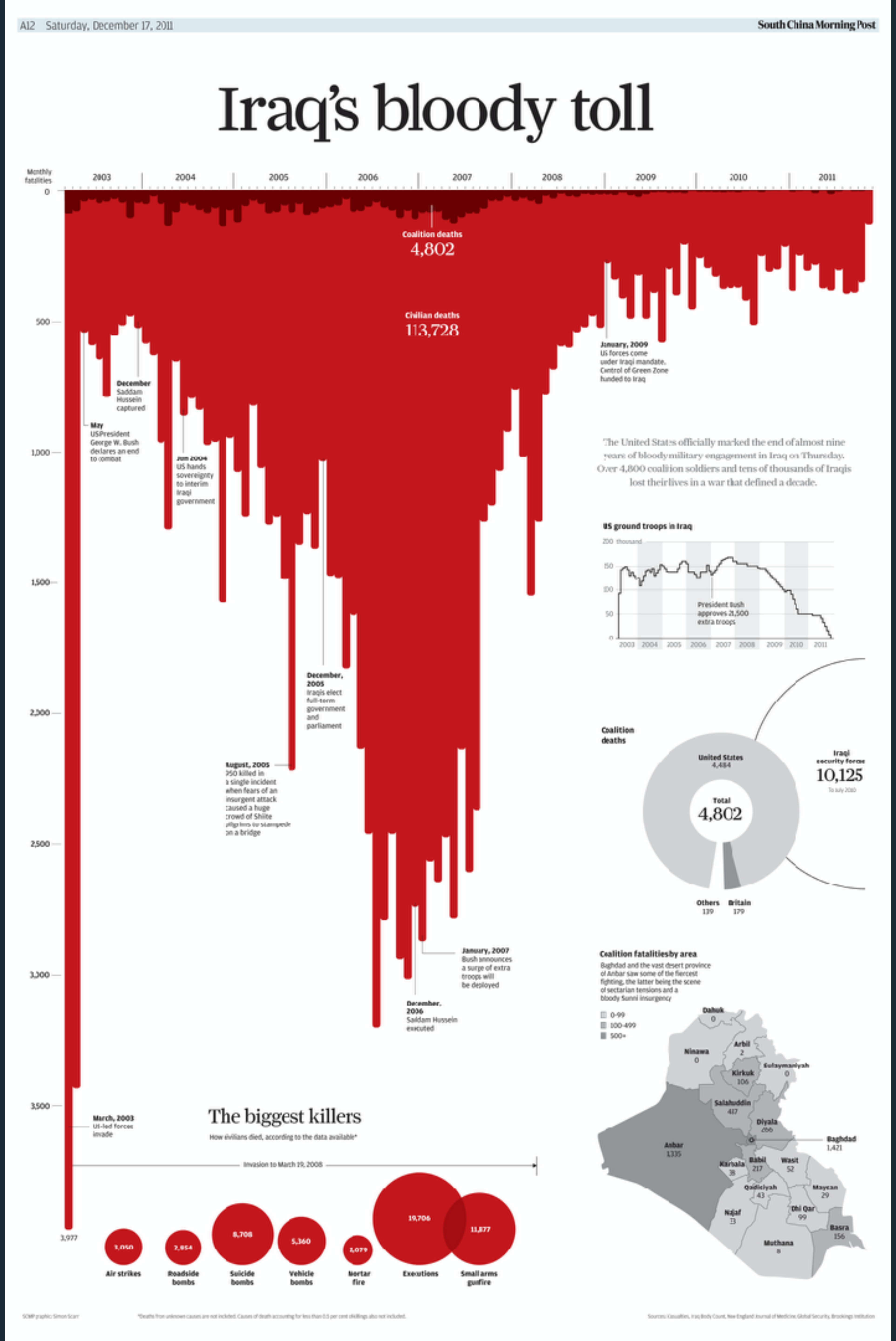


How to Lie with Statistics: Stand Your Ground and Gun Deaths

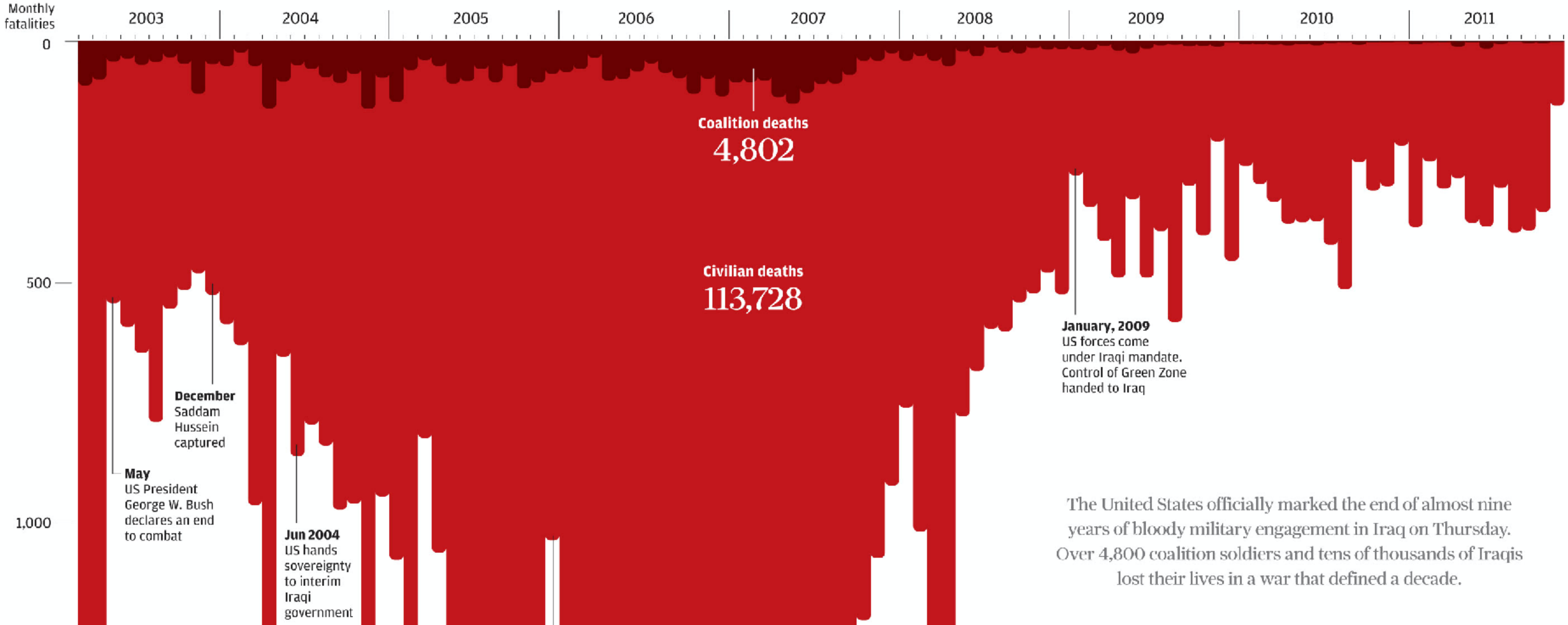
Lisa Wade, PhD on December 28, 2014

At Junk Charts, Kaiser Fung drew my attention to a graph released by Reuters. It is so deeply misleading that I loathe to expose your eyeballs to it. So, I offer you this:

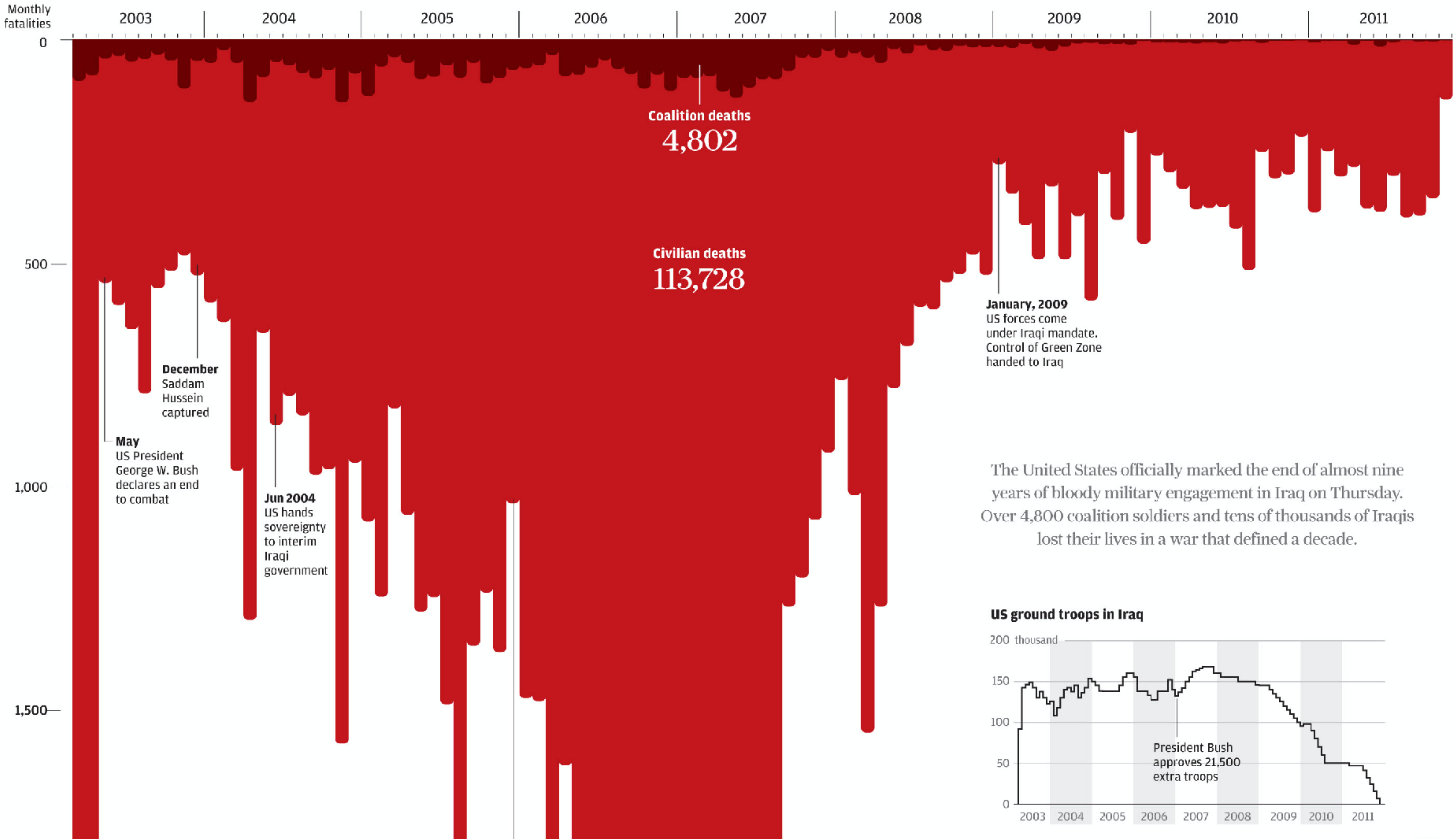



A12 Saturday, December 17, 2011
South China Morning Post
Iraq's bloody toll
Monthly fatalities
2003
2004
2005
2006
2007
2008
2009
2010
2011
0
500
1,000
1,500
2,000
2,500
3,500
3,977
Coalition deaths
4,802
Civilian deaths
113,728
March, 2003
US-led forces invade
May
US President George W. Bush declares an end to combat
December
Saddam Hussein captured
US hands sovereignty to interim Iraqi government
December, 2005
Iraqis elect full-term government and parliament
January, 2007
Bush announces a surge of extra troops will be deployed
December, 2006
Saddam Hussein executed
January, 2009
US forces come under Iraqi mandate. Control of Green Zone handed to Iraq
The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.
US ground troops in Iraq
200 thousand
150
100
50
0
President Bush approves 21,500 extra troops
Coalition deaths
United States
4,484
Total
4,802
Others
139
Britain
179
Iraqi security forces
10,125
To July 2010
Coalition fatalities by area
Baghdad and the vast desert province of Anbar saw some of the fiercest fighting, the latter being the scene of sectarian tensions and a bloody Sunni insurgency
0-99
100-499
500+
Dahuk
0
Ninawa
0
Arbil
2
Kirkuk
106
Sulaymaniyah
0
Salahuddin
417
Anbar
1335
Baghdad
1,421
Babil
217
Wasit
52
Karbala
38
Qadisiyah
43
Maysan
29
Najaf
13
Dhi Qar
99
Basra
156
Muthana
8
The biggest killers
How civilians died, according to the data available*
Invasion to March 19, 2008
8,708
5,360
19,706
11,877
Air strikes
Roadside bombs
Suicide bombs
Vehicle bombs
Mortar fire
Executions
Small arms gunfire
SCMP graphic: Simon Scarr
*Deaths from unknown causes are not included. Causes of death accounting for less than 0.5 per cent of killings also not included.
Sources: Casualties, Iraq Body Count, New England Journal of Medicine, Global Security, Brookings Institution

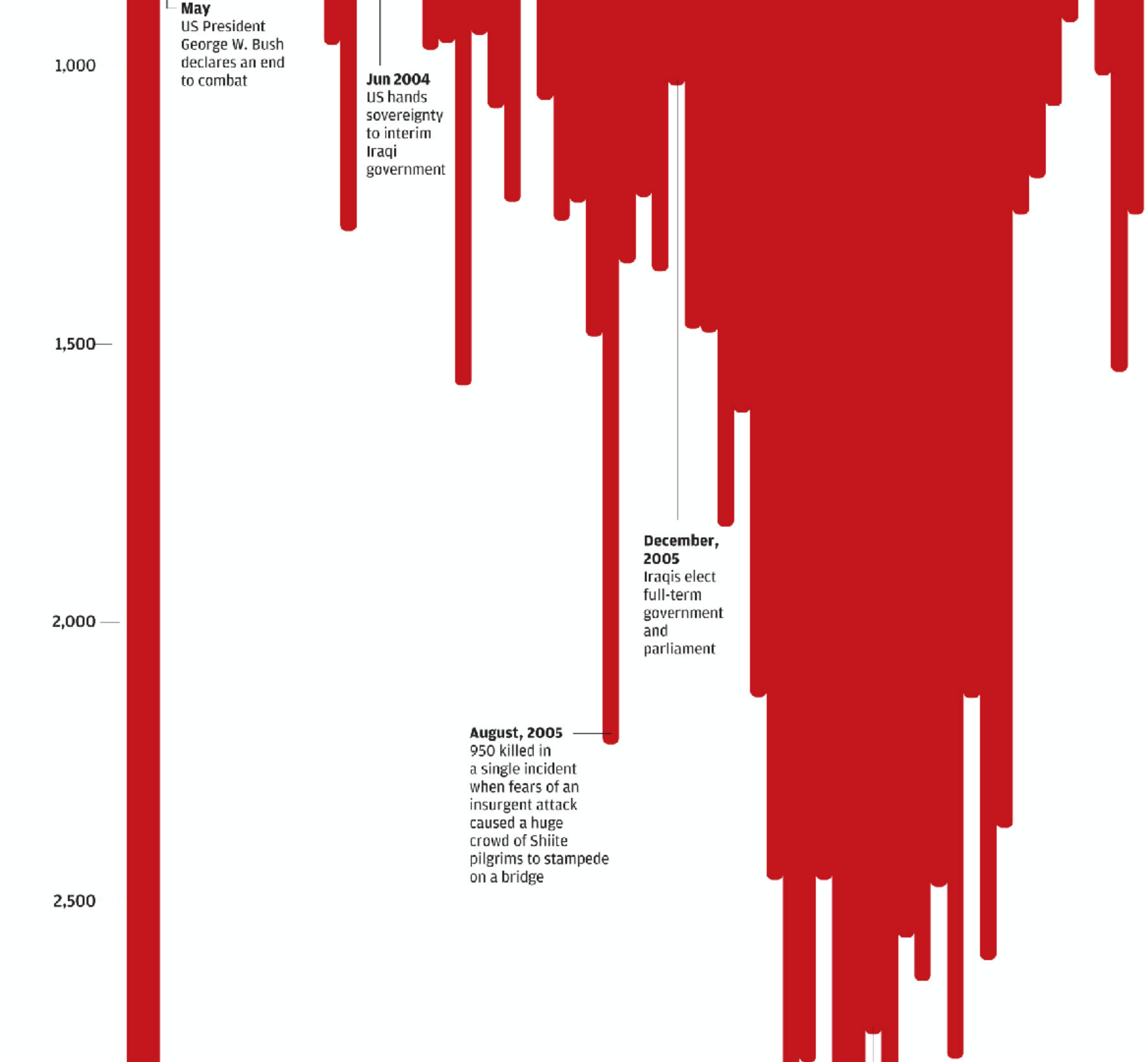
Iraq's bloody toll



The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.

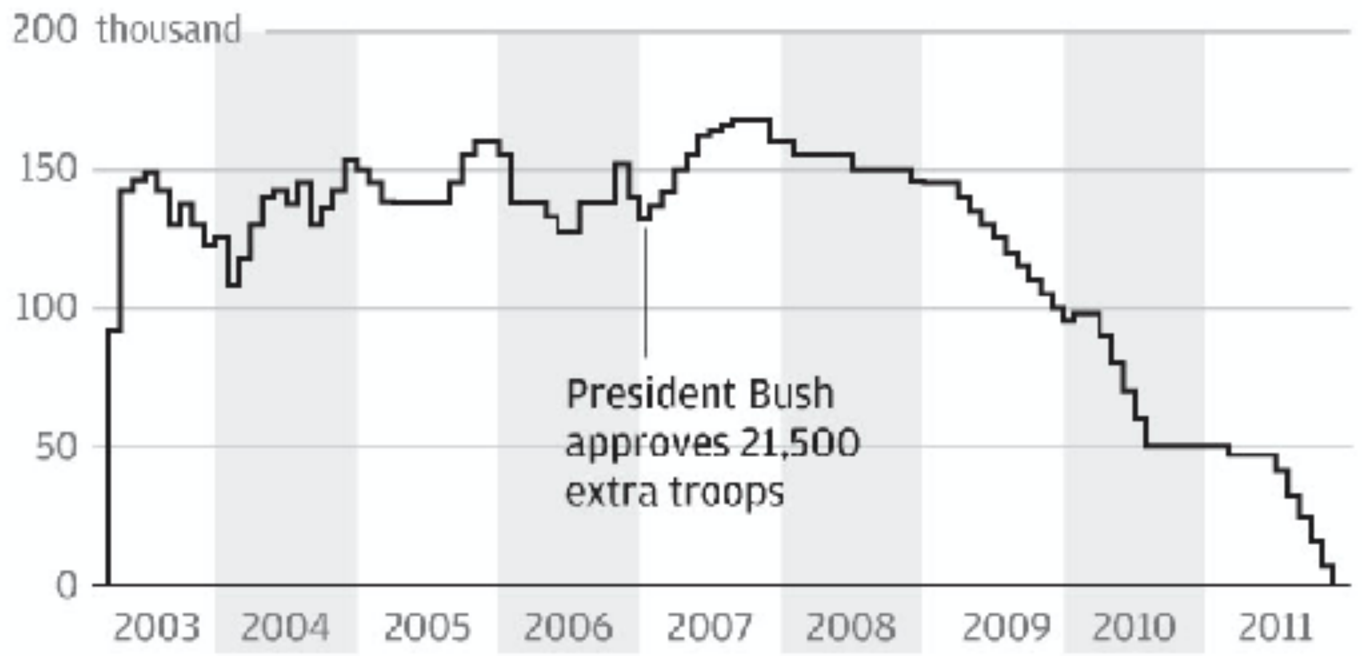


The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.

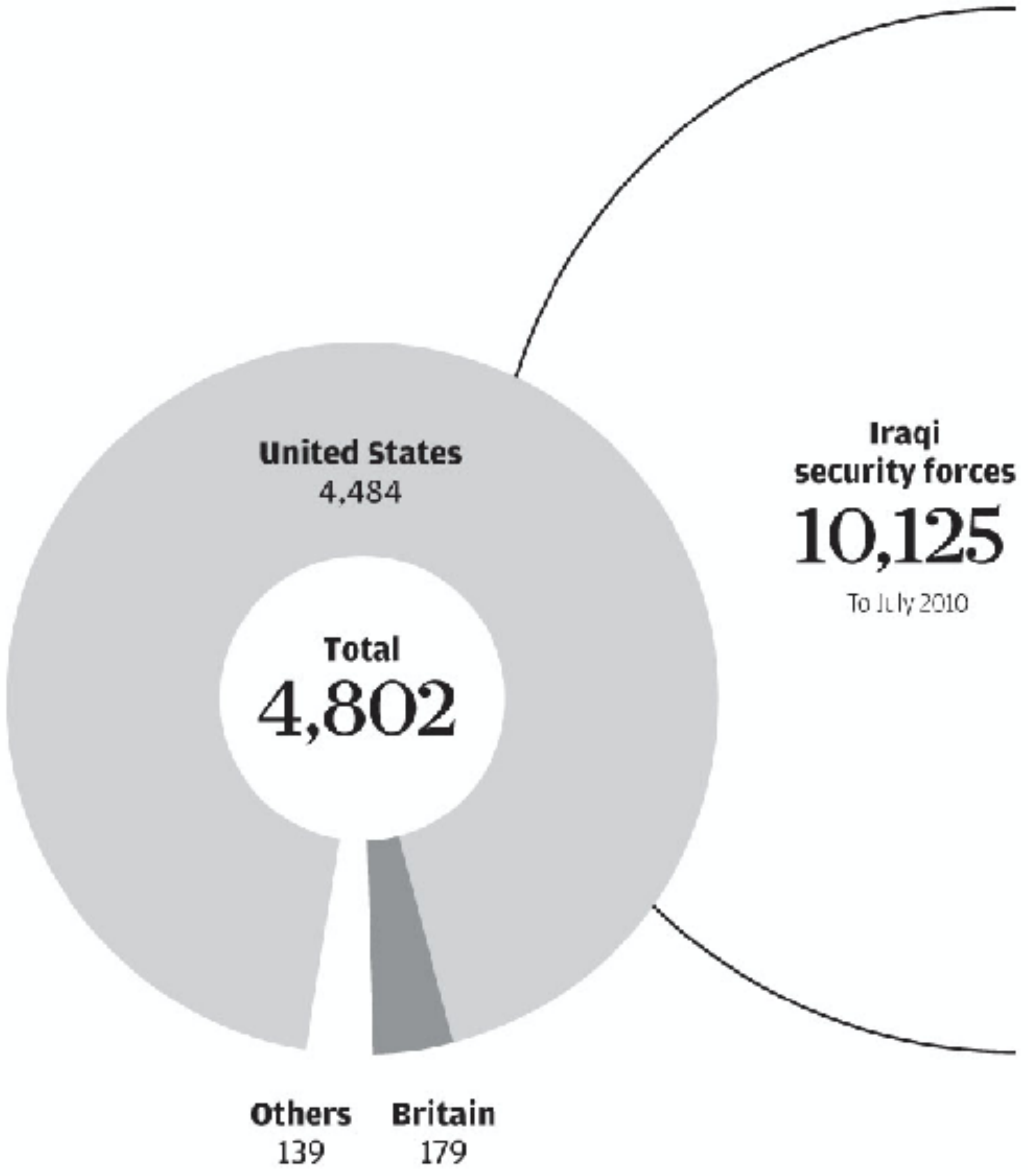


The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.

US ground troops in Iraq

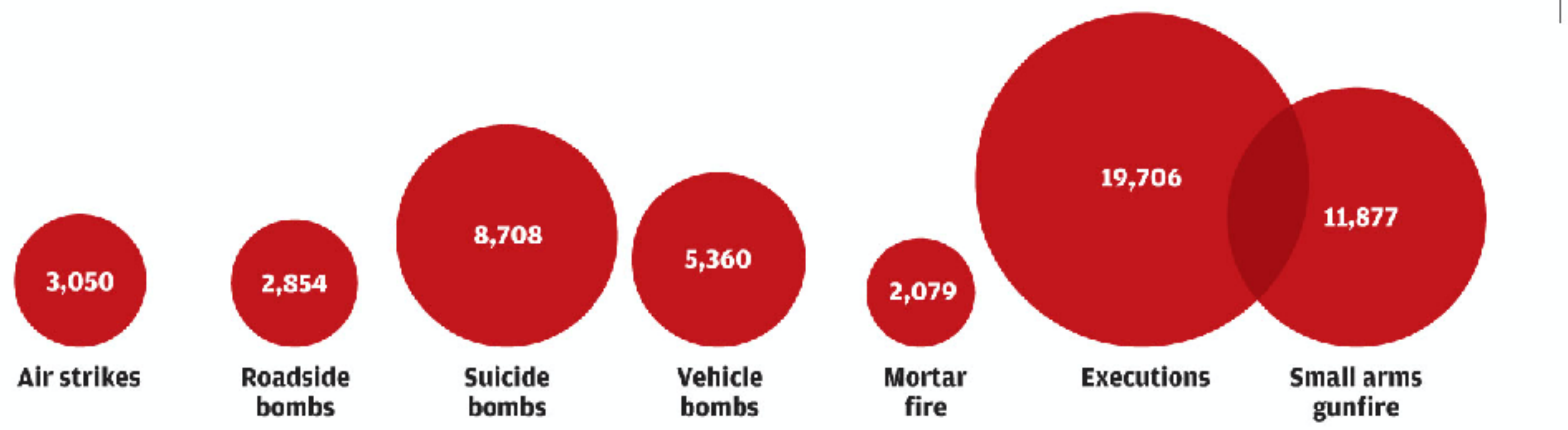


Coalition deaths



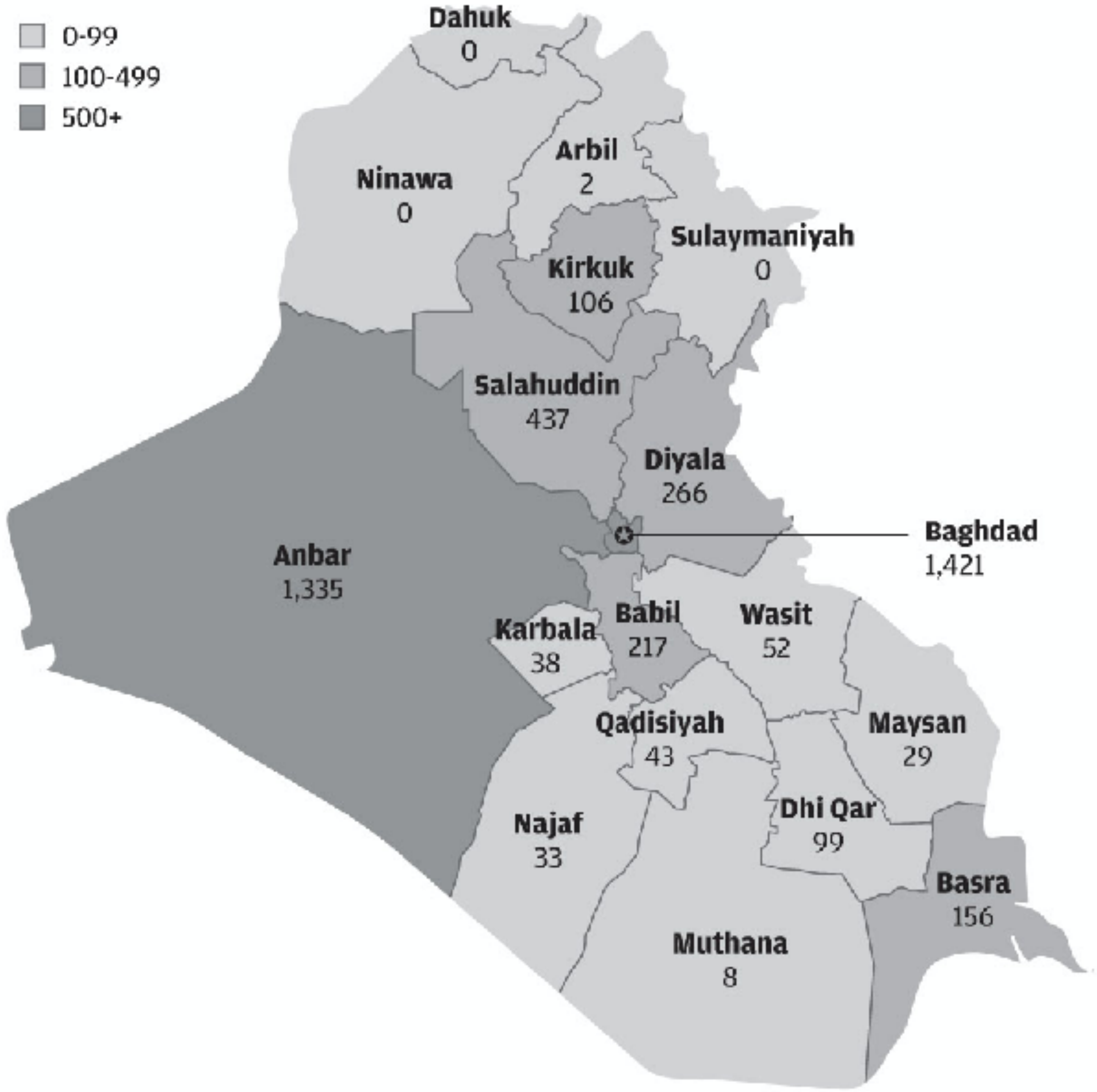
The biggest killers

How civilians died, according to the data available*



Coalition fatalities by area

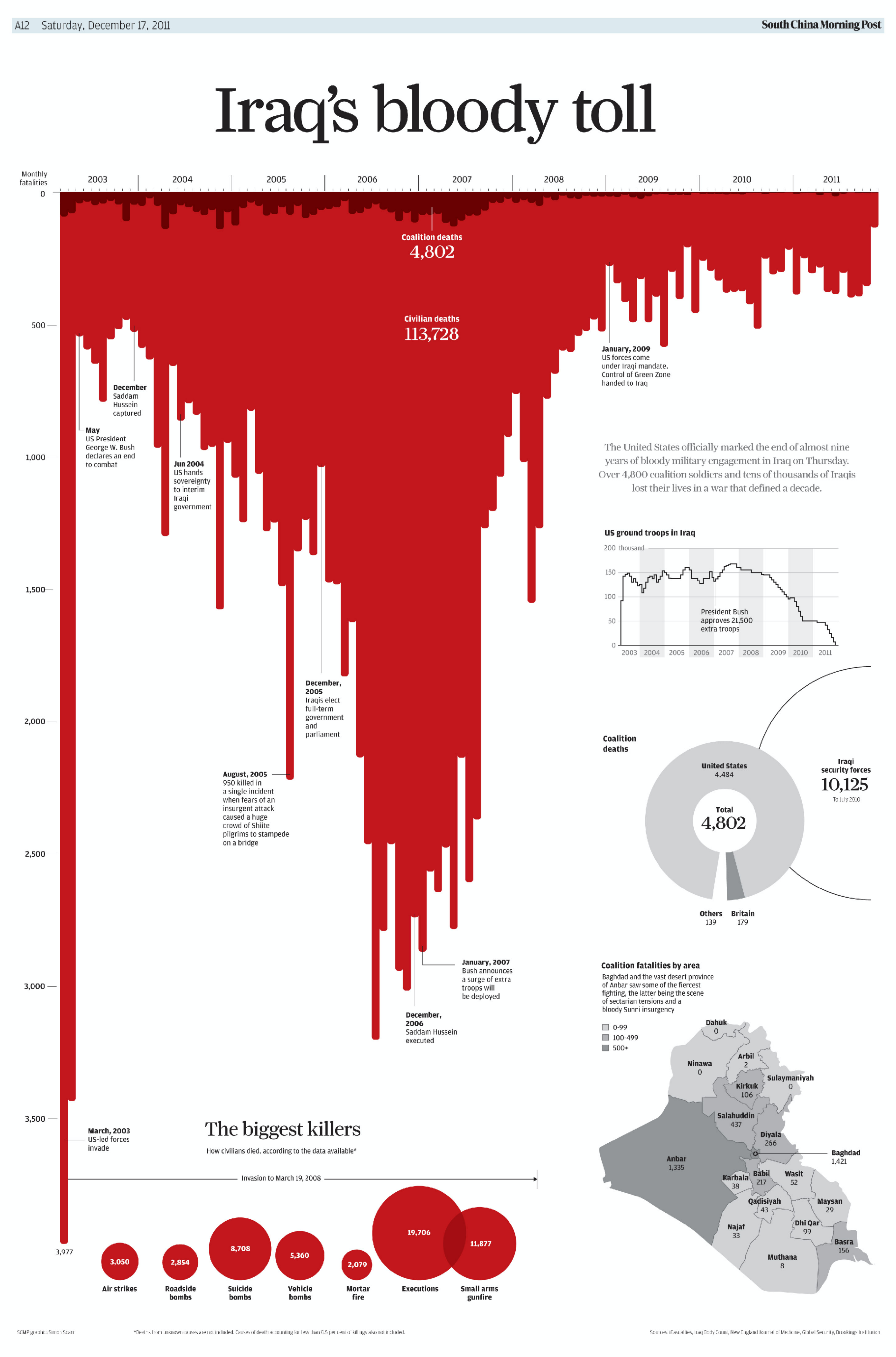
Baghdad and the vast desert province of Anbar saw some of the fiercest fighting, the latter being the scene of sectarian tensions and a bloody Sunni insurgency

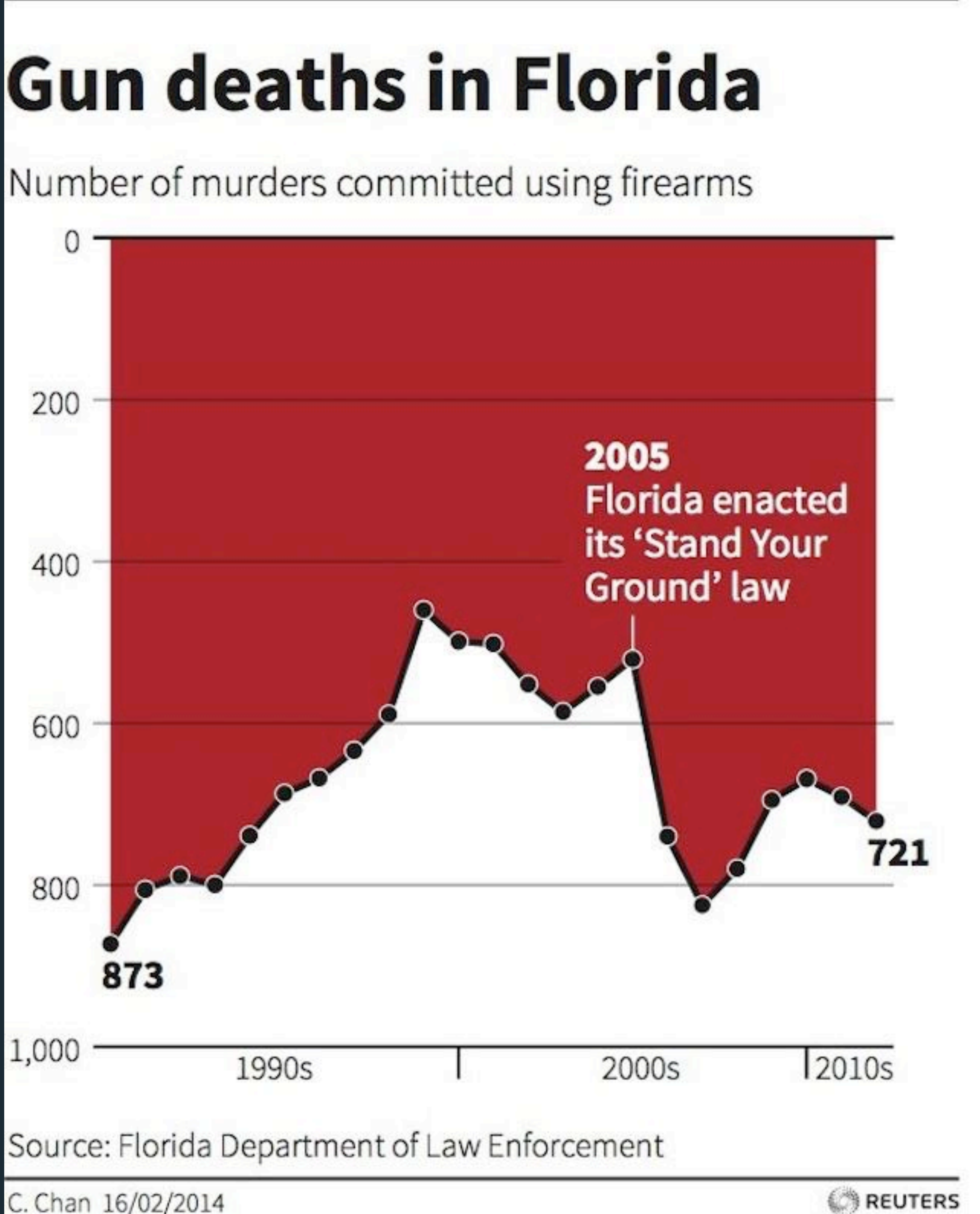


SCMP graphic: Simon Scarr

*Deaths from unknown causes are not included. Causes of death accounting for less than 0.5 per cent of killings also not included.

Sources: iCasualties, Iraq Body Count, New England Journal of Medicine, Global Security, Brookings Institution

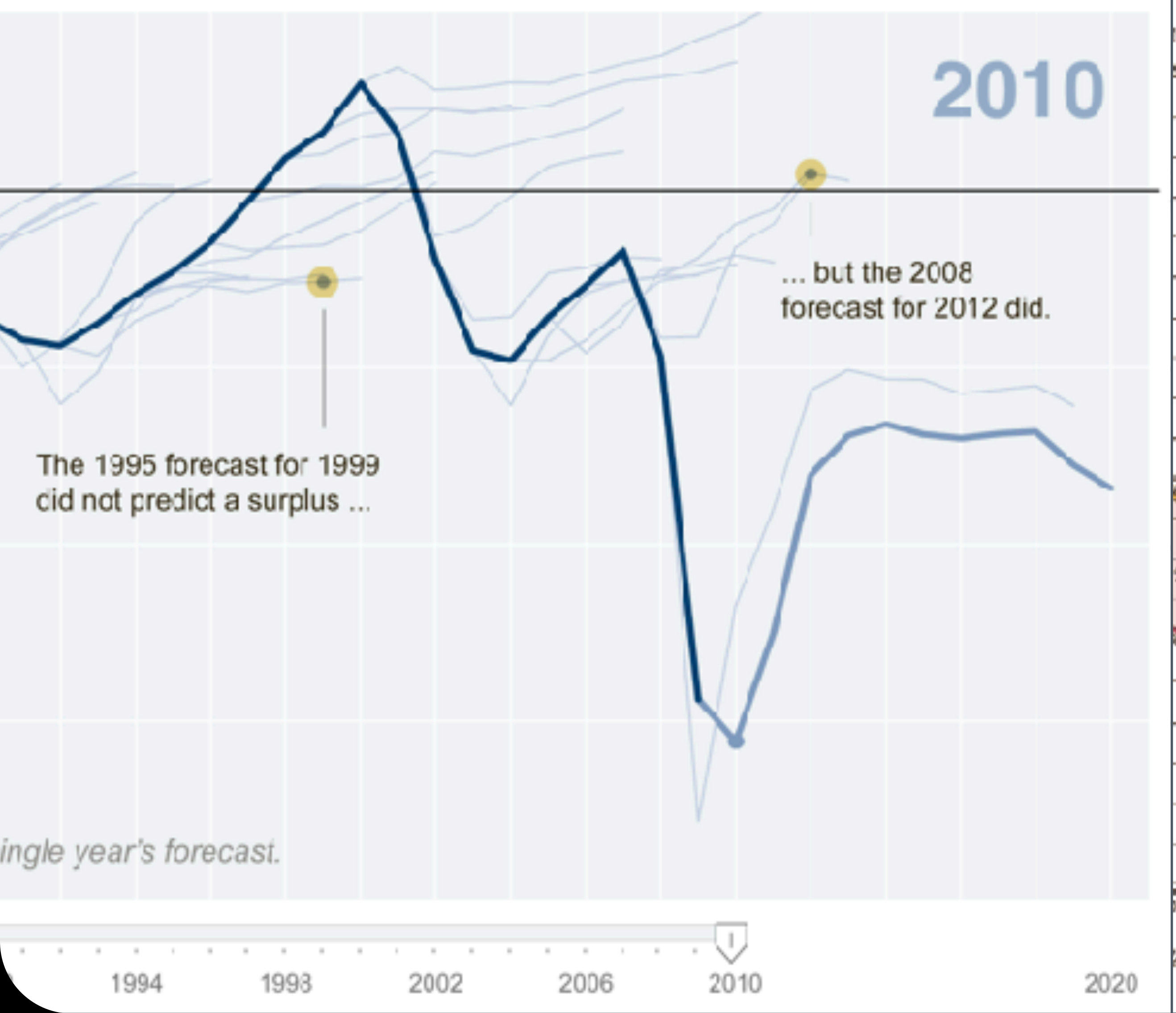

A12 Saturday, December 17, 2011
South China Morning Post
Iraq's bloody toll
Monthly fatalities
2003
2004
2005
2006
2007
2008
2009
2010
2011
Coalition deaths
4,802
Civilian deaths
113,728
December Saddam Hussein captured
May US President George W. Bush declares an end to combat
Jun 2004 US hands sovereignty to interim Iraqi government
January, 2009 US forces come under Iraqi mandate. Control of Green Zone handed to Iraq
The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.
US ground troops in Iraq
President Bush approves 21,500 extra troops
December, 2005 Iraqis elect full-term government and parliament
August, 2005 950 killed in a single incident when fears of an insurgent attack caused a huge crowd of Shiite pilgrims to stampede on a bridge
Coalition deaths
United States 4,484
Total 4,802
Iraqi security forces 10,125
Others 139
Britain 179
January, 2007 Bush announces a surge of extra troops will be deployed
December, 2006 Saddam Hussein executed
Coalition fatalities by area
Baghdad and the vast desert province of Anbar saw some of the fiercest fighting, the latter being the scene of sectarian violence and a bloody Sunni insurgency
0-99
100-499
500+
Dahuk 0
Ninawa 0
Arbil 2
Sulaymaniyah 0
Kirkuk 106
Salahuddin 437
Diyala 266
Baghdad 1,421
Anbar 1,335
Karbala 38
Babil 217
Wasit 52
Qadisiyah 43
Maysan 29
Najaf 33
Dhi Qar 99
Basra 156
Muthana 8
March, 2003 US-led forces invade
The biggest killers
How civilians died, according to the data available*
Invasion to March 19, 2008
3,977
3,050 Air strikes
2,854 Roadside bombs
8,708 Suicide bombs
5,360 Vehicle bombs
2,079 Mortar fire
19,706 Executions
11,877 Small arms gunfire
500
1,000
1,500
2,000
2,500
3,000
3,500

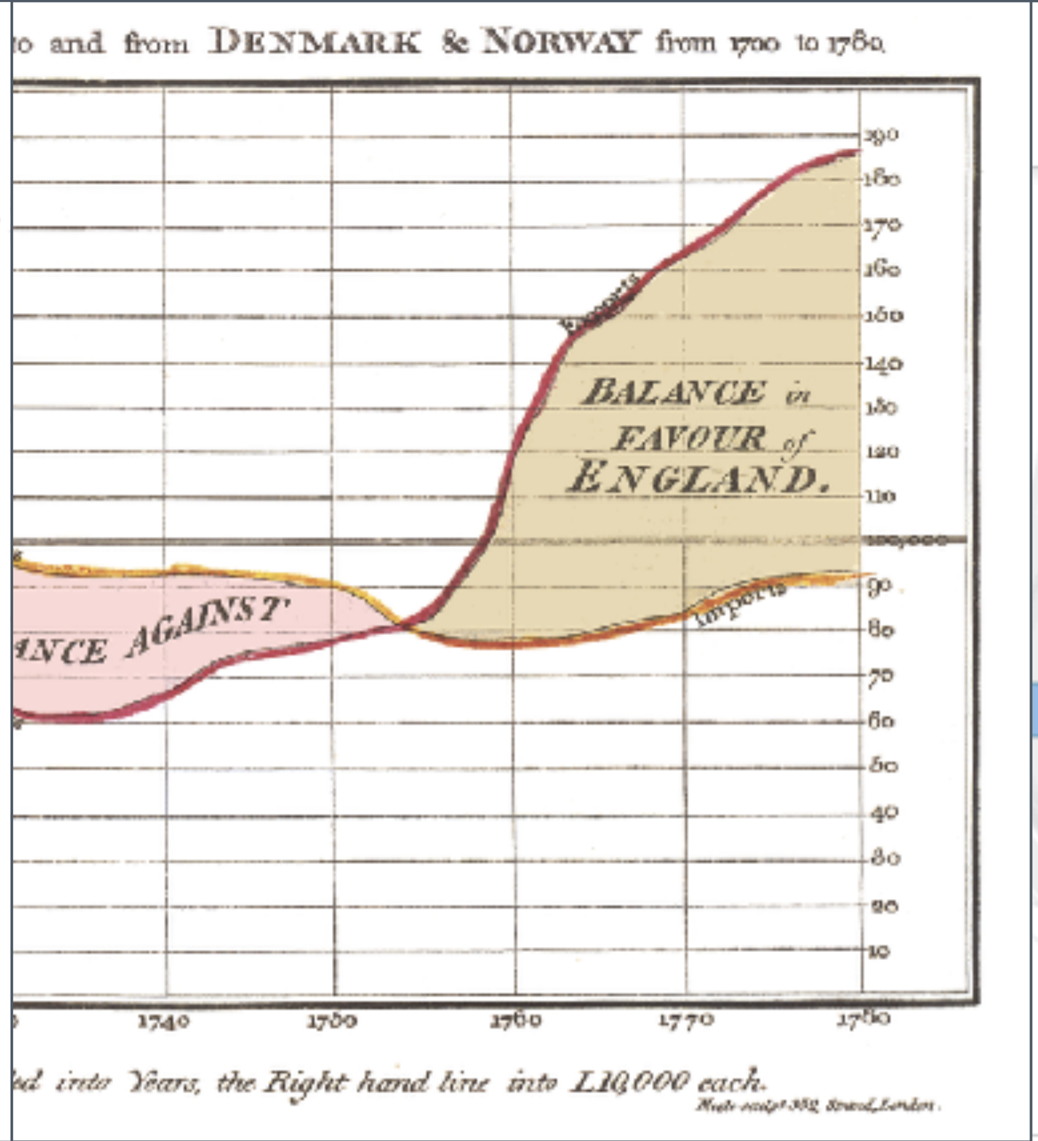

Gun deaths in Florida
Number of murders committed using firearms
0
200
400
600
800
1,000
2005
Florida enacted its 'Stand Your Ground' law
873
721
1990s
2000s
2010s
Source: Florida Department of Law Enforcement
C. Chan 16/02/2014
REUTERS

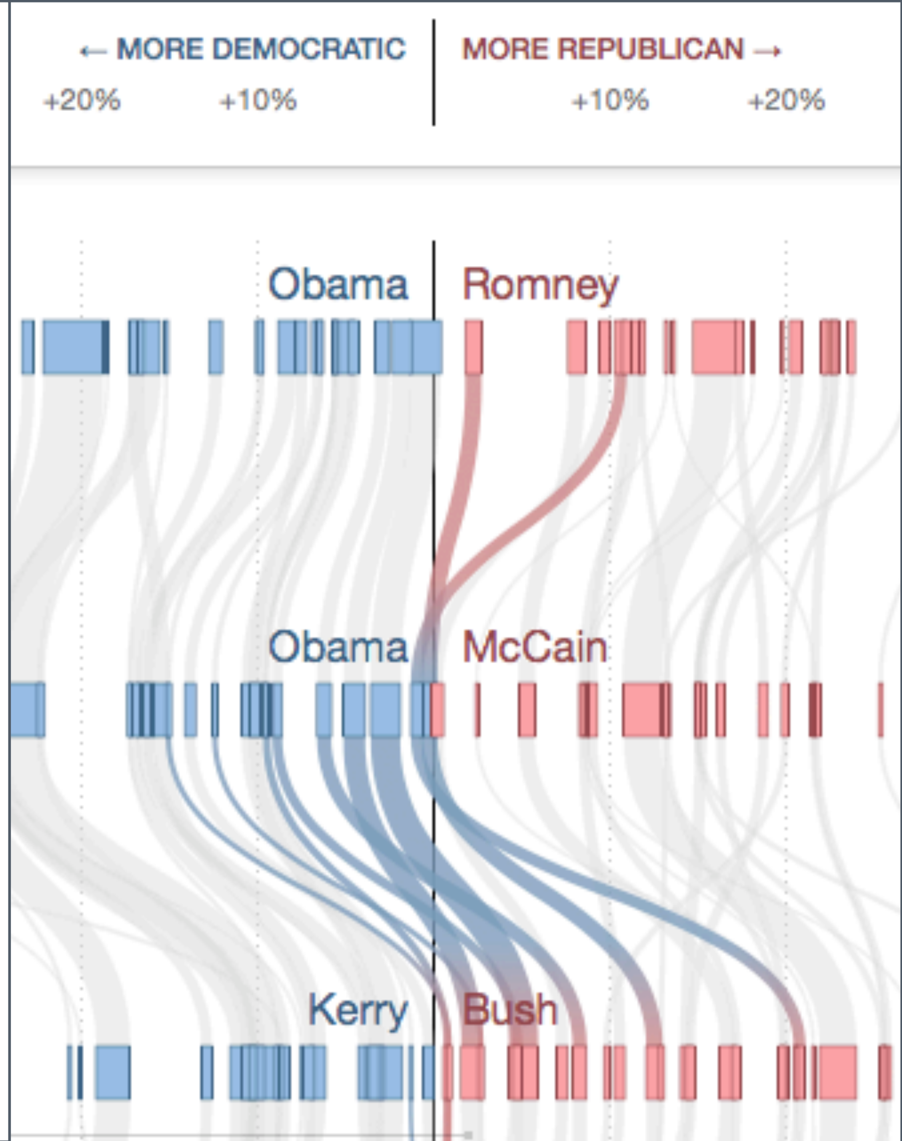
6.859: Interactive Data Visualization

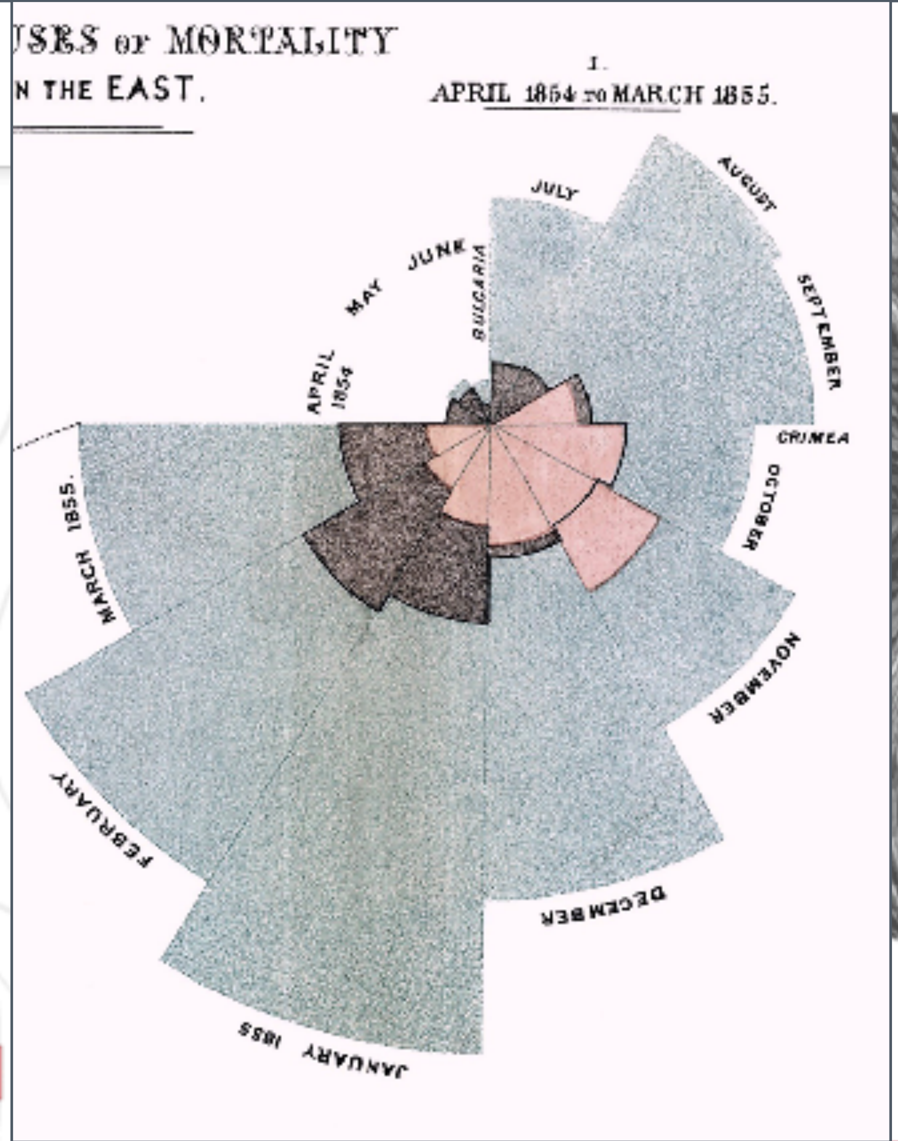
Exploratory Data Analysis

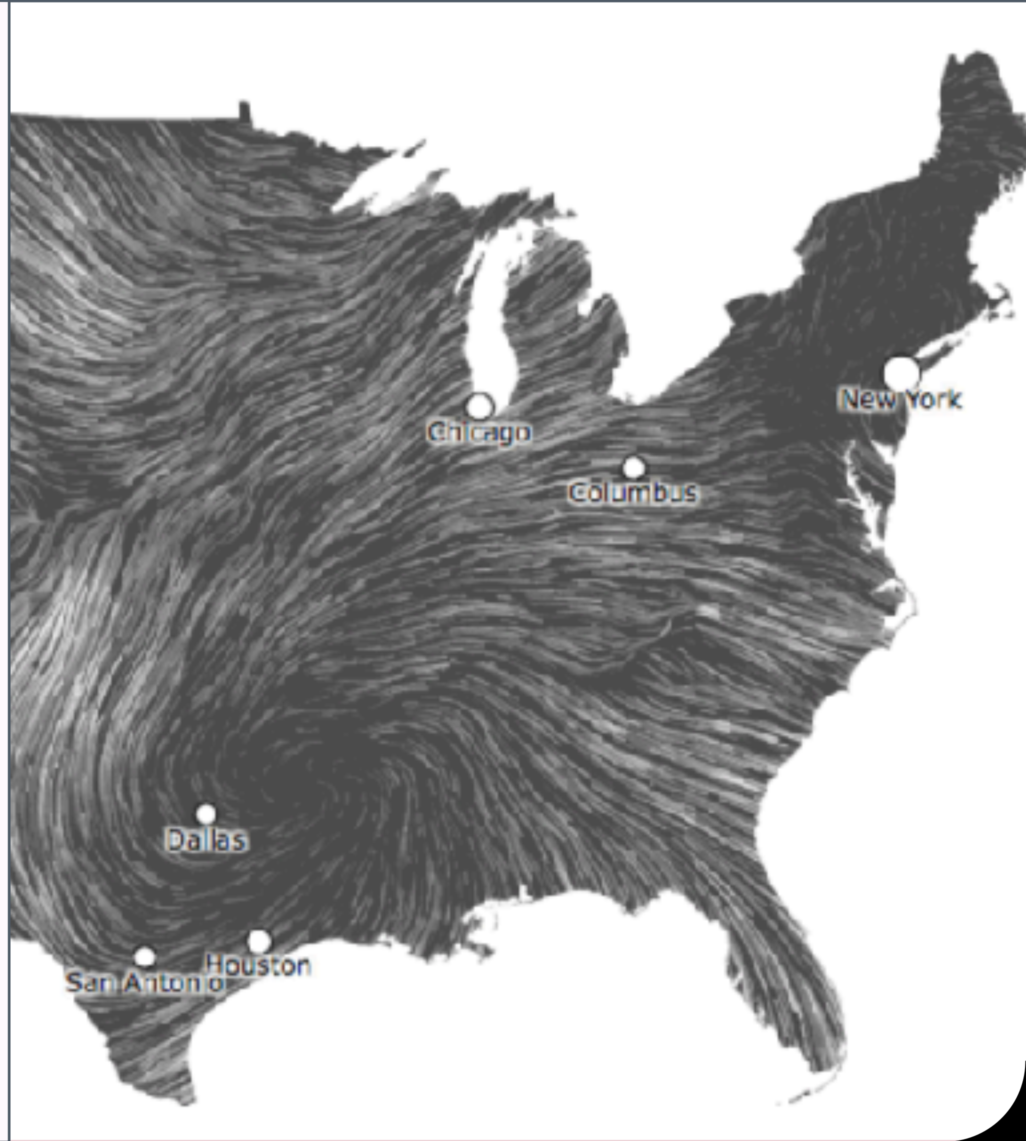
Arvind Satyanarayan



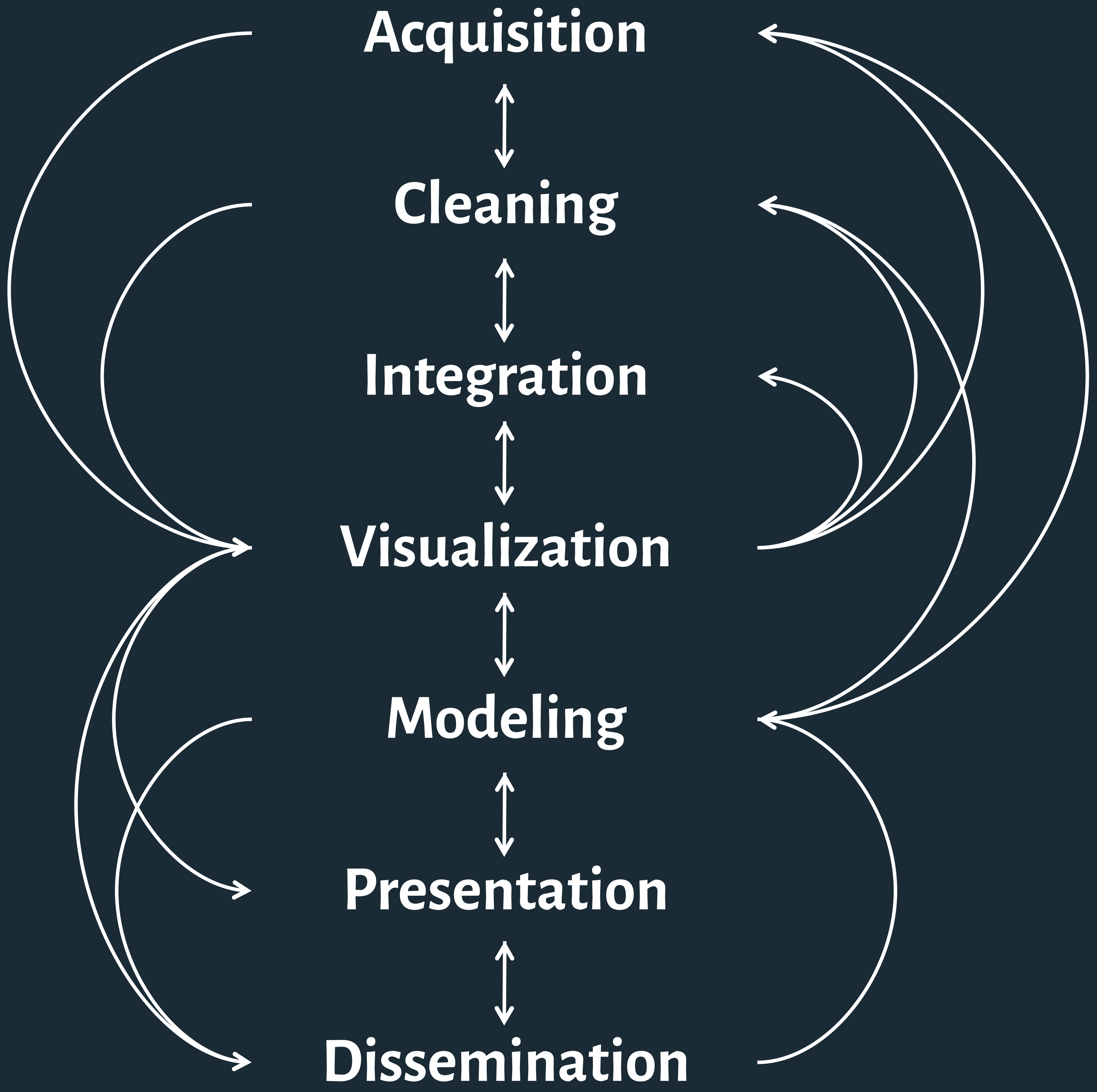


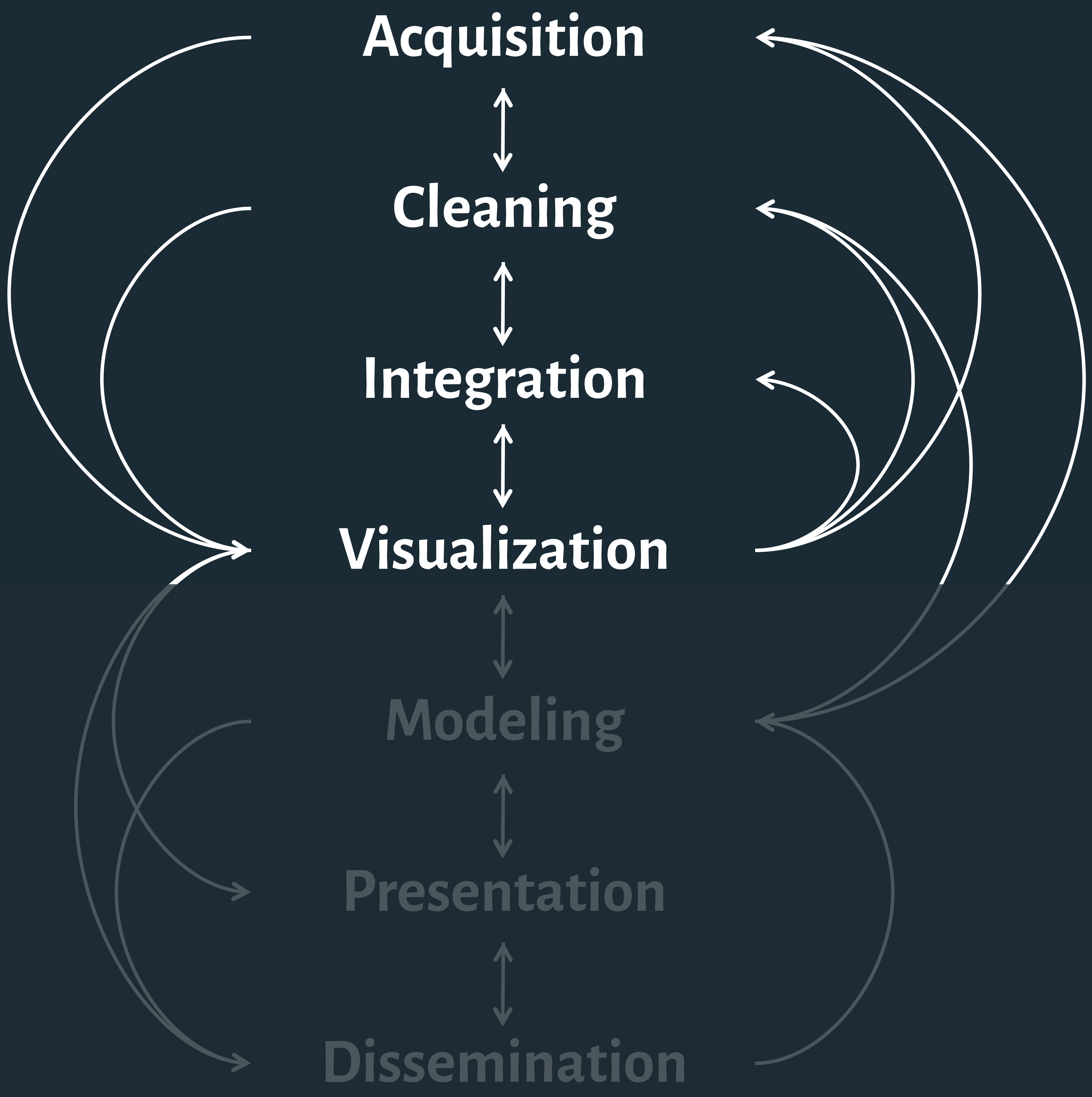






Data Visualization


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

Data Provenance

Who collected or produced it?

What was their intent?

Is it a reputable source?

What are the motives of the data producer (are they can advocate or lobbyist?)

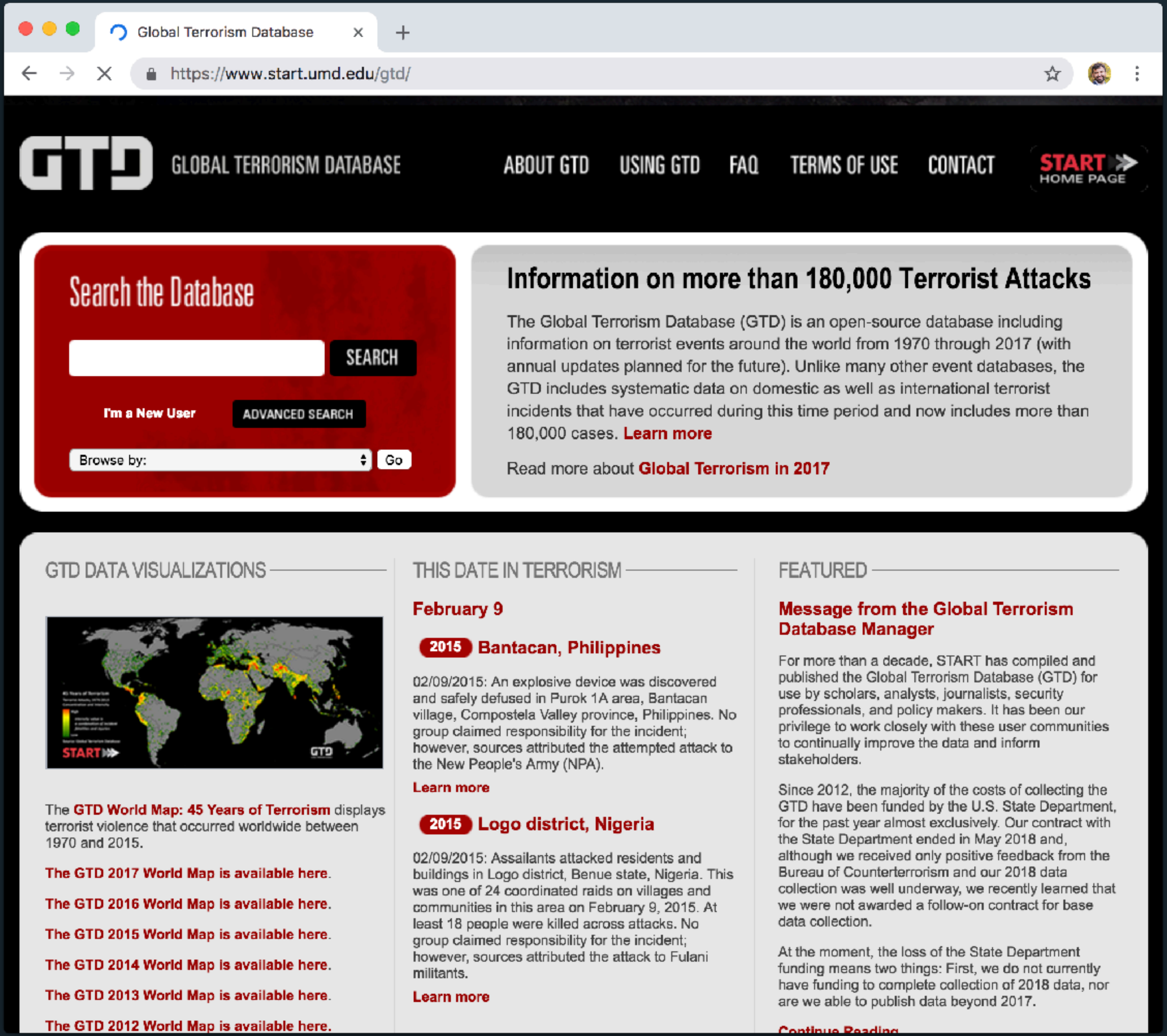
Data Provenance

Who collected or produced it?

What was their intent?

Is it a reputable source?

What are the motives of the data producer (are they can advocate or lobbyist?)



Data Provenance

When was the data collected?

Measurements can change over time.

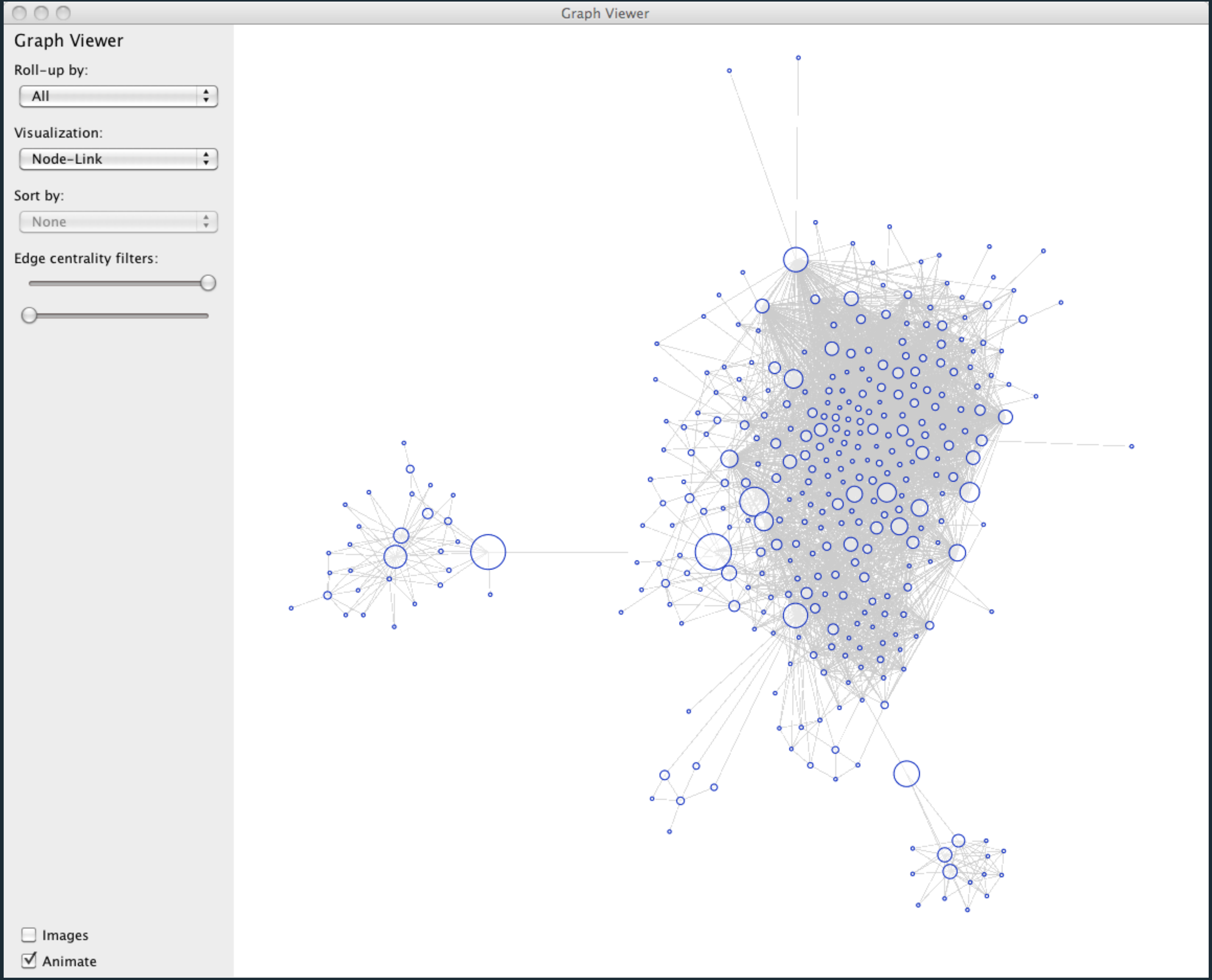
Definitions/interpretations in quantification can change over time.

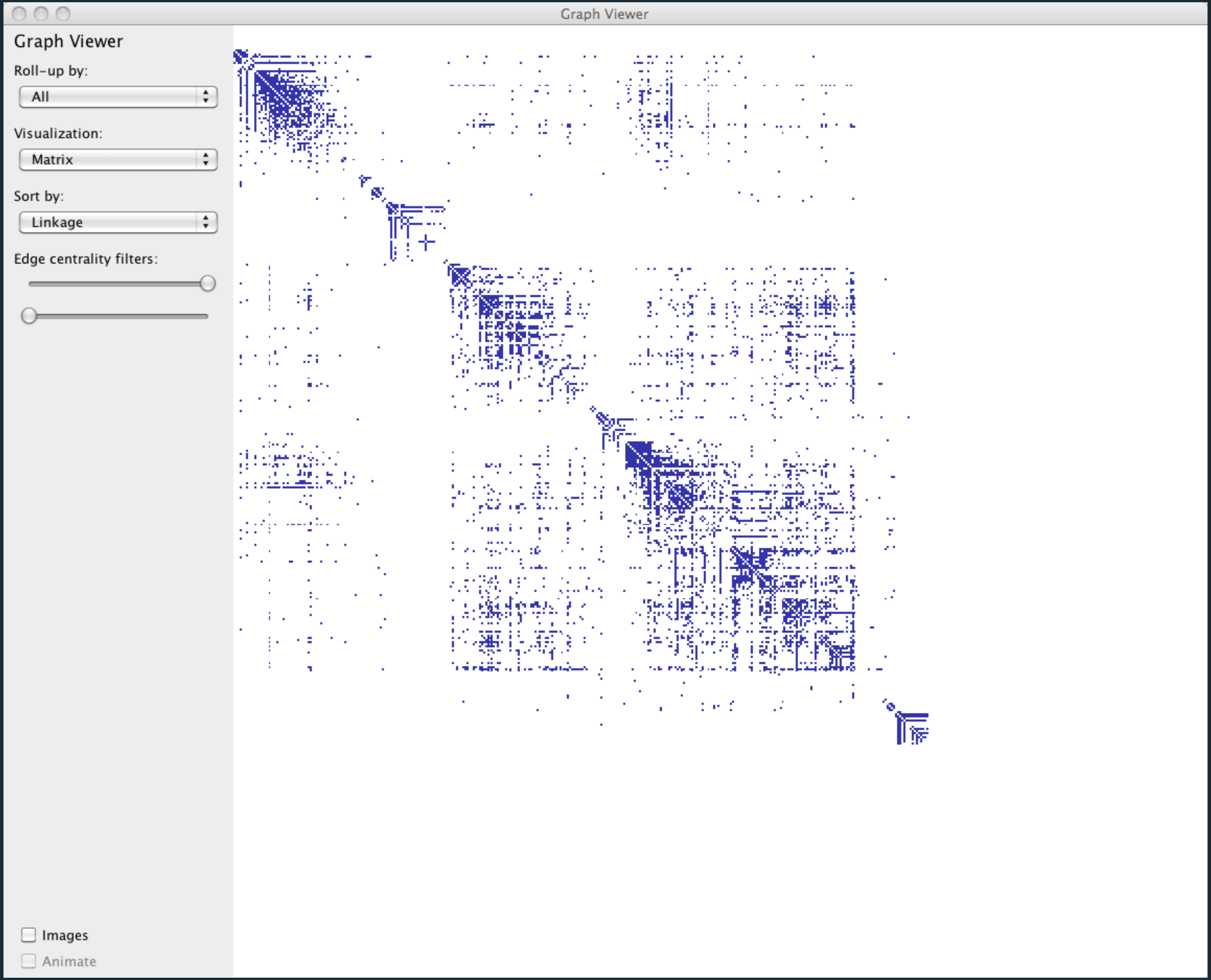
Is the data recent, and how much does that matter to the insight you wish to convey?

The Three-Year Plunge

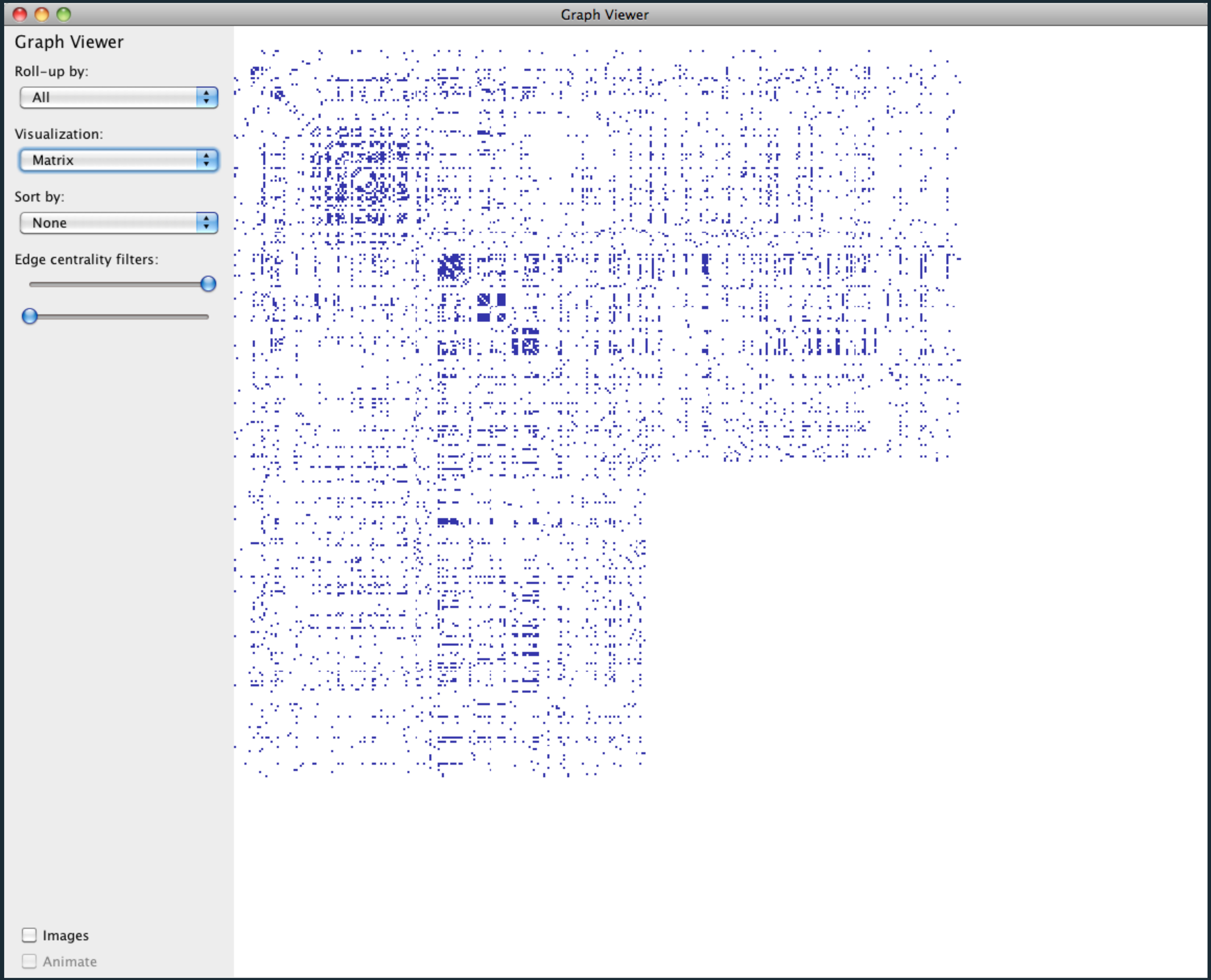
To help gauge each city's overall crime level, the FBI tracks eight "index crimes." From 1993 to 2010, Chicago's annual total dropped by 47 percent. But from 2010 to 2013, it dropped a stunning 56 percent, or nearly 19 percent per year, according to data from the Chicago Police Department.

David Berstein & Noah Isackson. *Chicago Magazine*, April 2014.


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
Linkage
Edge centrality filters:
Images
Animate

Missing Values



Berkeley	|||||||||||||||||||||||||||||||
Cornell	||||
Harvard	|||||||||
Harvard University	|||||||
Stanford	||||||||||||||||||||
Stanford University	||||||||||
UC Berkeley	|||||||||||||||||||||
UC Davis	||||||||||
University of California at Berkeley	|||||||||||||||
University of California, Berkeley	||||||||||||||||||
University of California, Davis	|||

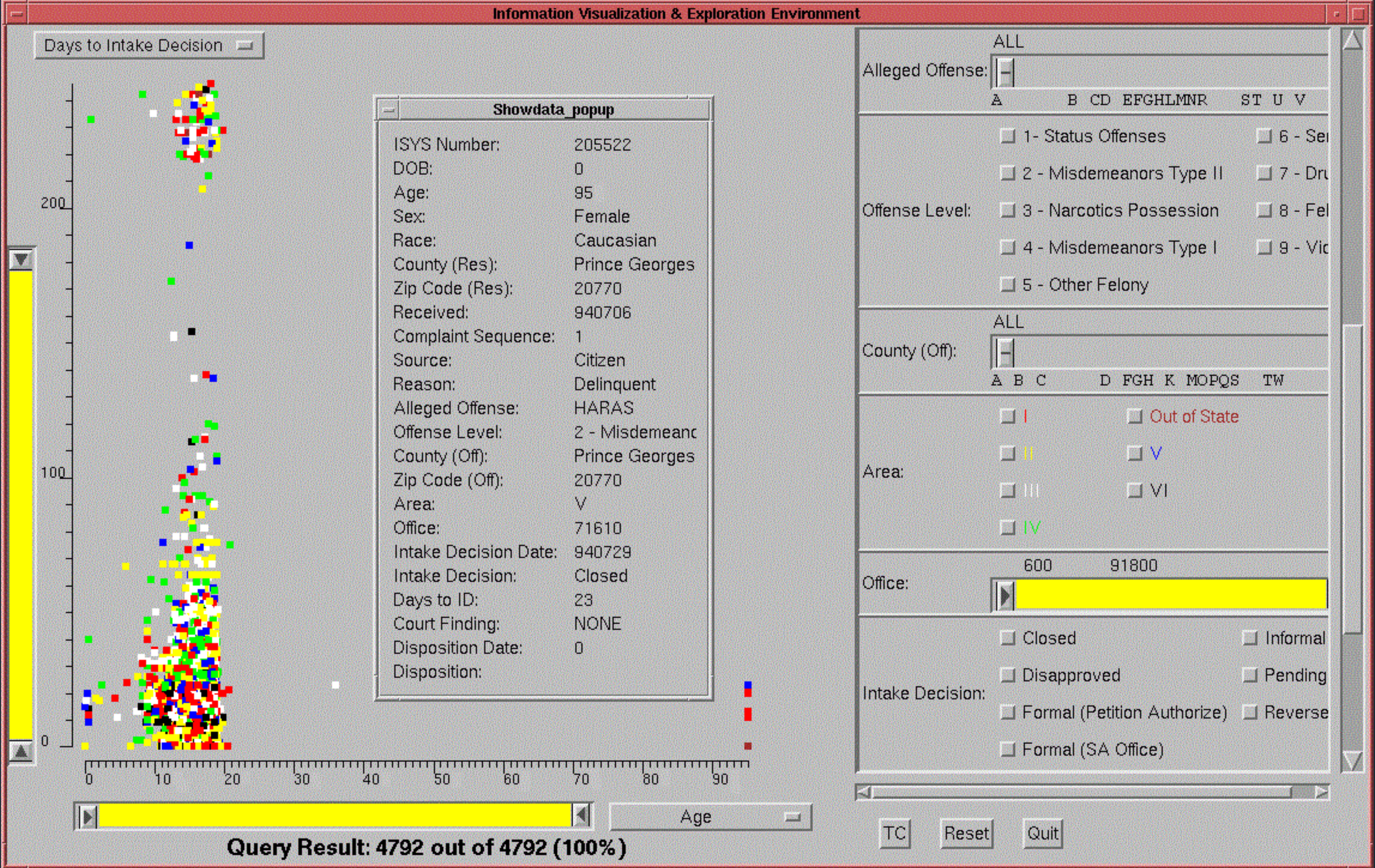
“The first sign that a visualization is good is that **it shows you a problem in your data**. Every successful visualization that I've been involved with has had this stage where you realize, ***"Oh my God, this data is not what I thought it would be!"*** So already, you've discovered something.”

– **Martin Wattenberg**

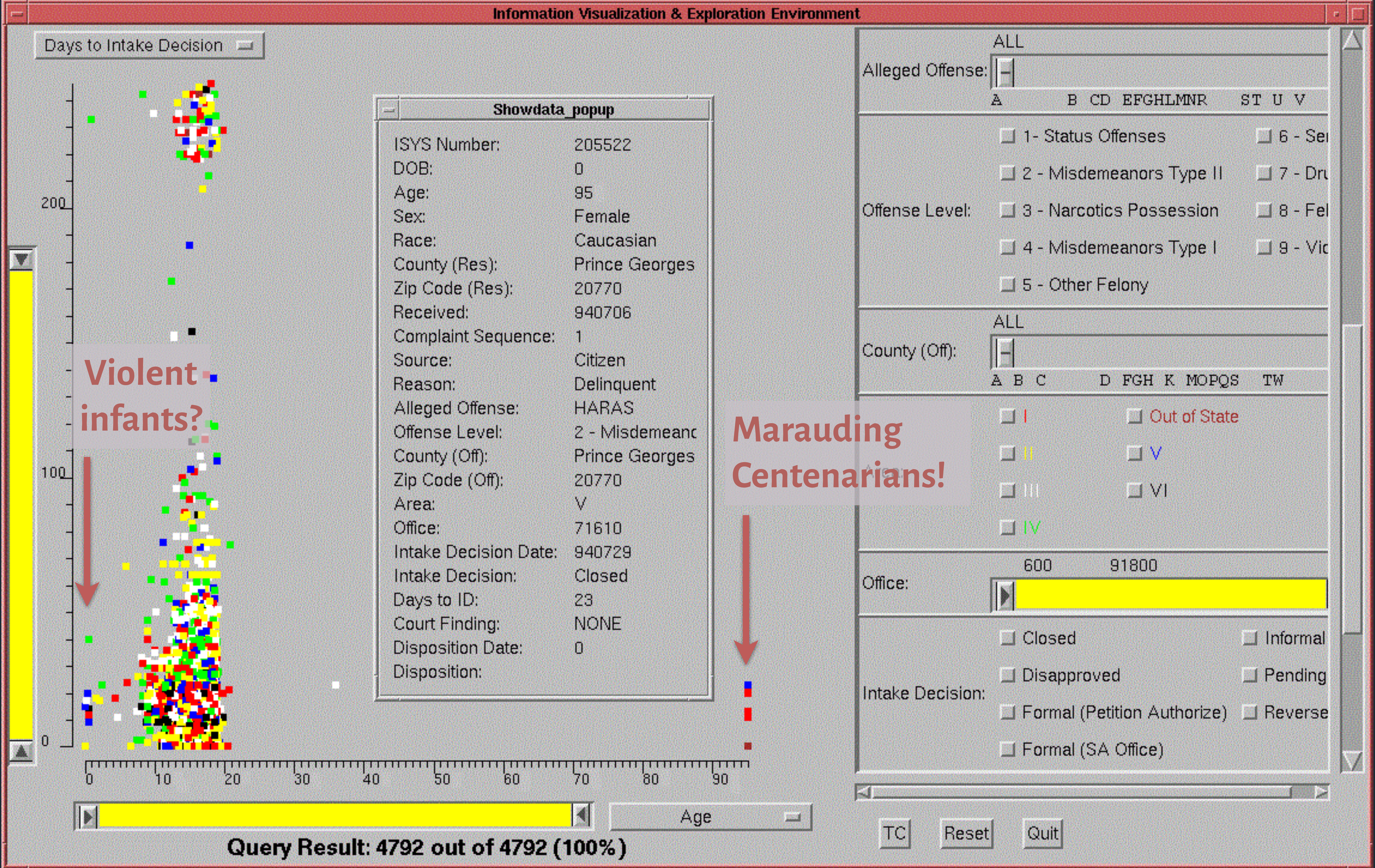
Co-lead of Google's People + AI Initiative

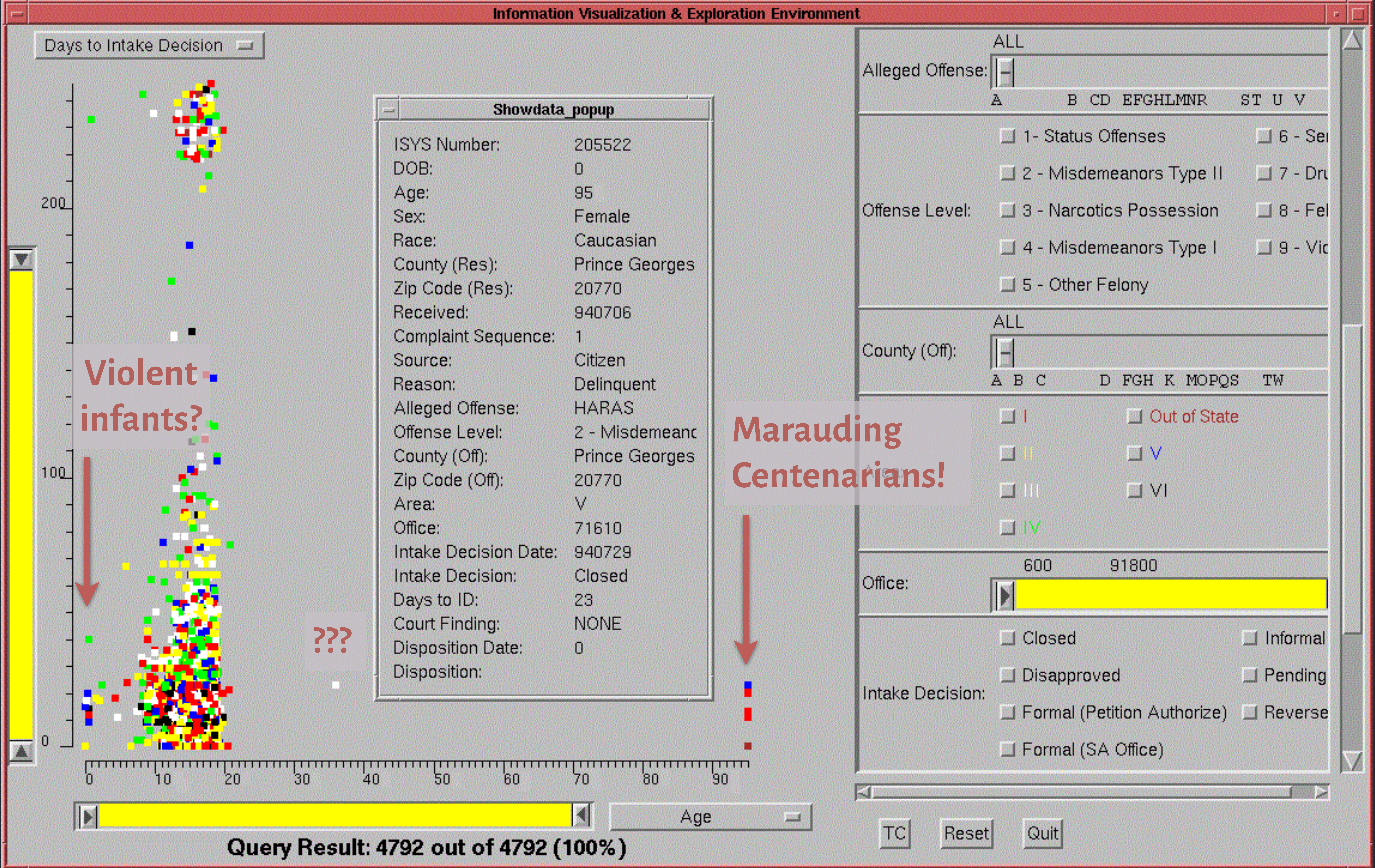
ACM Queue, Mar 2010

Ch 15: Visualization Data Mining Models. Thearling et al., *Information Visualization in Data Mining & Knowledge Discovery* 2002.



Ch 15: Visualization Data Mining Models. Thearling et al., *Information Visualization in Data Mining & Knowledge Discovery* 2002.





Ch 15: Visualization Data Mining Models. Thearling et al., *Information Visualization in Data Mining & Knowledge Discovery* 2002.

Big Data Borat
@BigDataBorat

Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

6:47 PM - 26 Feb 2013

540 Retweets **343** Likes

12 540 343

"I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any "analysis" at all."

– **Anonymous Data Scientist**

[Kandel et al. VAST 2012]

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

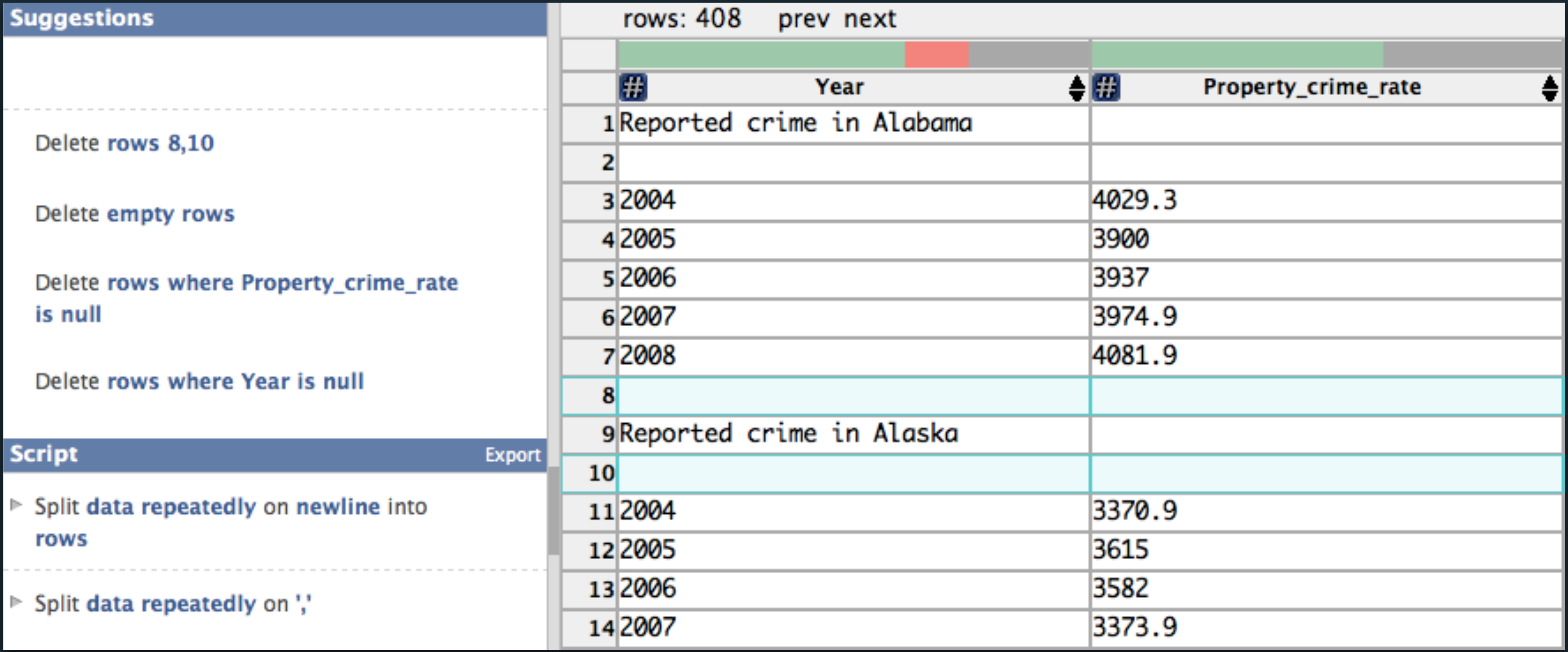
Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

Reported crime in Colorado

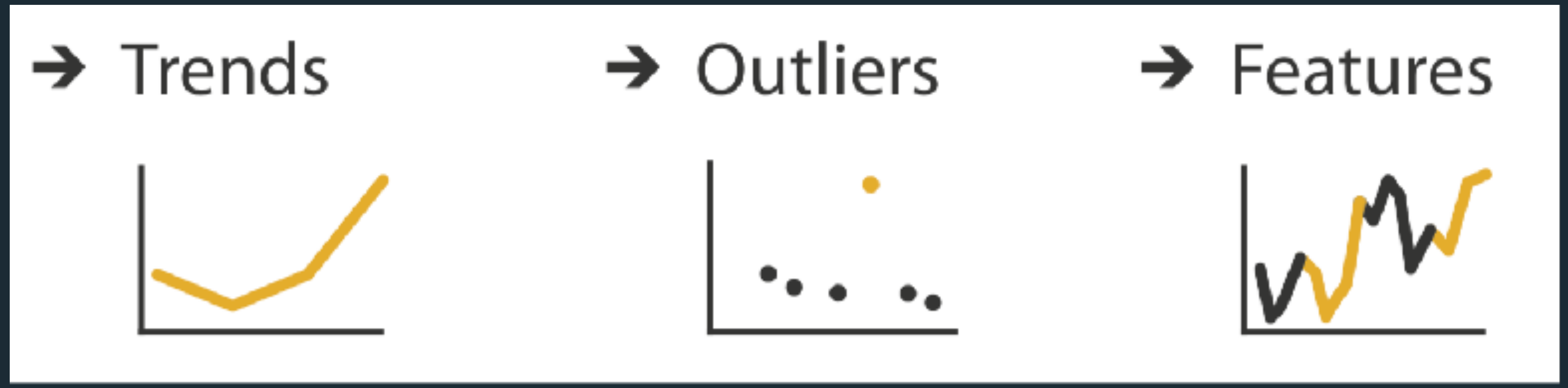
Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

DataWrangler



Wrangler: Interactive Visual Specification of Data Transformation Scripts. Sean Kandel et al., *ACM CHI* 2011.

Exploratory Visual Analysis



Process

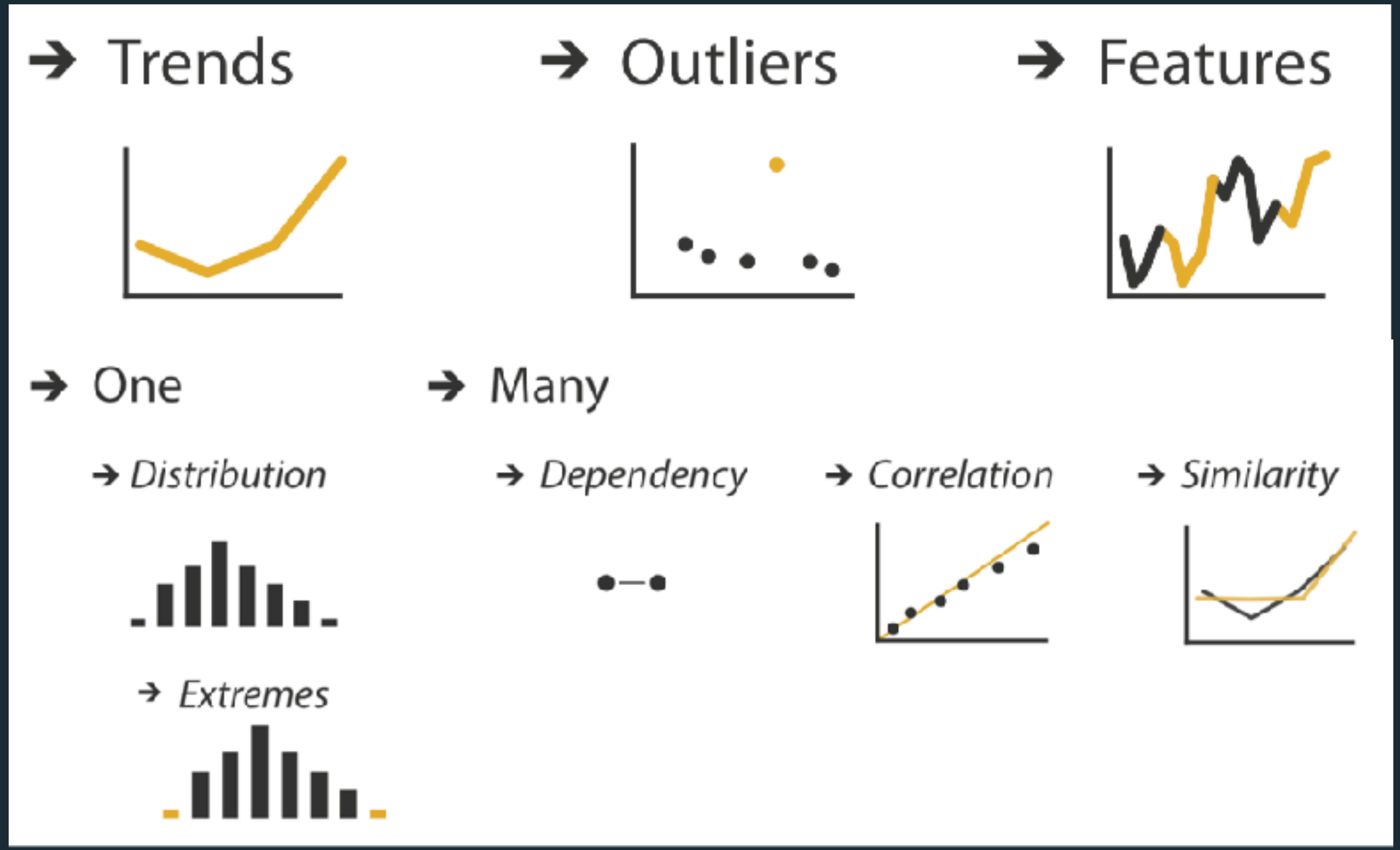
1. Construct graphics to address questions.
2. Inspect "answer" and ask new questions.
3. Iterate…

Tamara Munzner, *Visualization Analysis and Design* (2014).

Exploratory Visual Analysis

Process

1. Construct graphics to address questions.
2. Inspect "answer" and ask new questions.
3. Iterate...

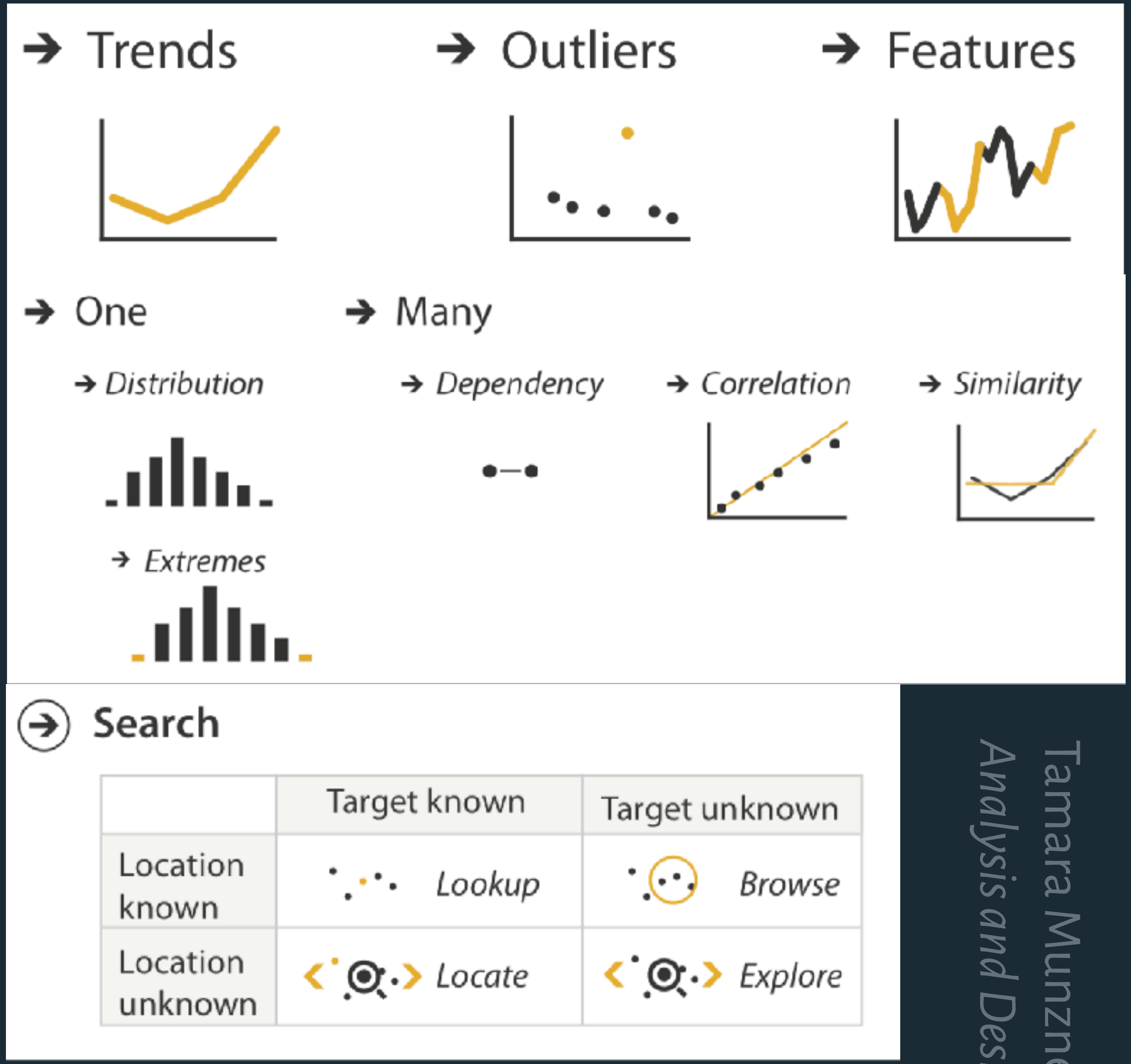


Tamara Munzner, *Visualization Analysis and Design* (2014).

Exploratory Visual Analysis

Process

1. Construct graphics to address questions.
2. Inspect "answer" and ask new questions.
3. Iterate…

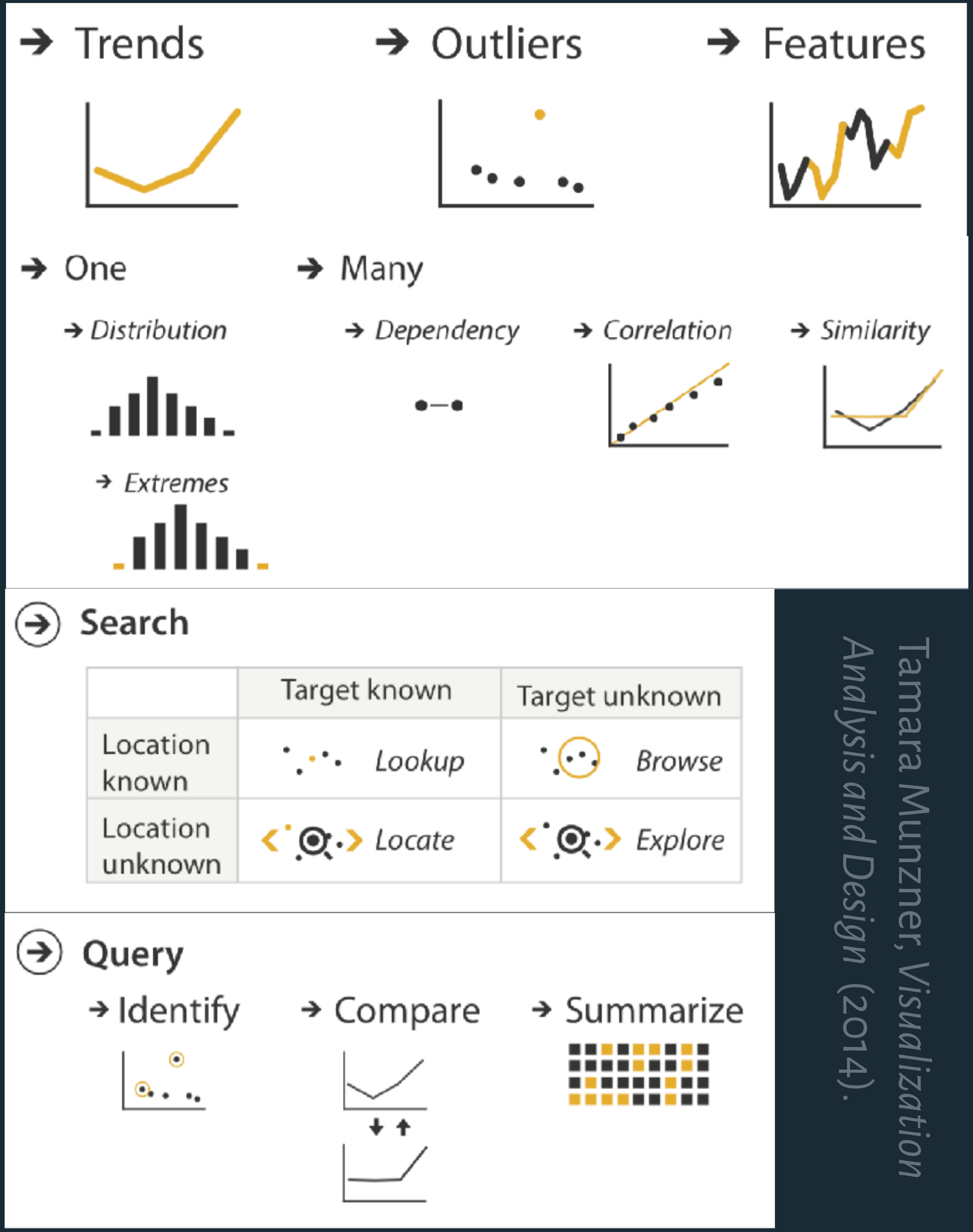


Tamara Munzner, *Visualization Analysis and Design* (2014).

Exploratory Visual Analysis

Process

1. Construct graphics to address questions.
2. Inspect "answer" and ask new questions.
3. Iterate…



Tamara Munzner, *Visualization Analysis and Design* (2014).

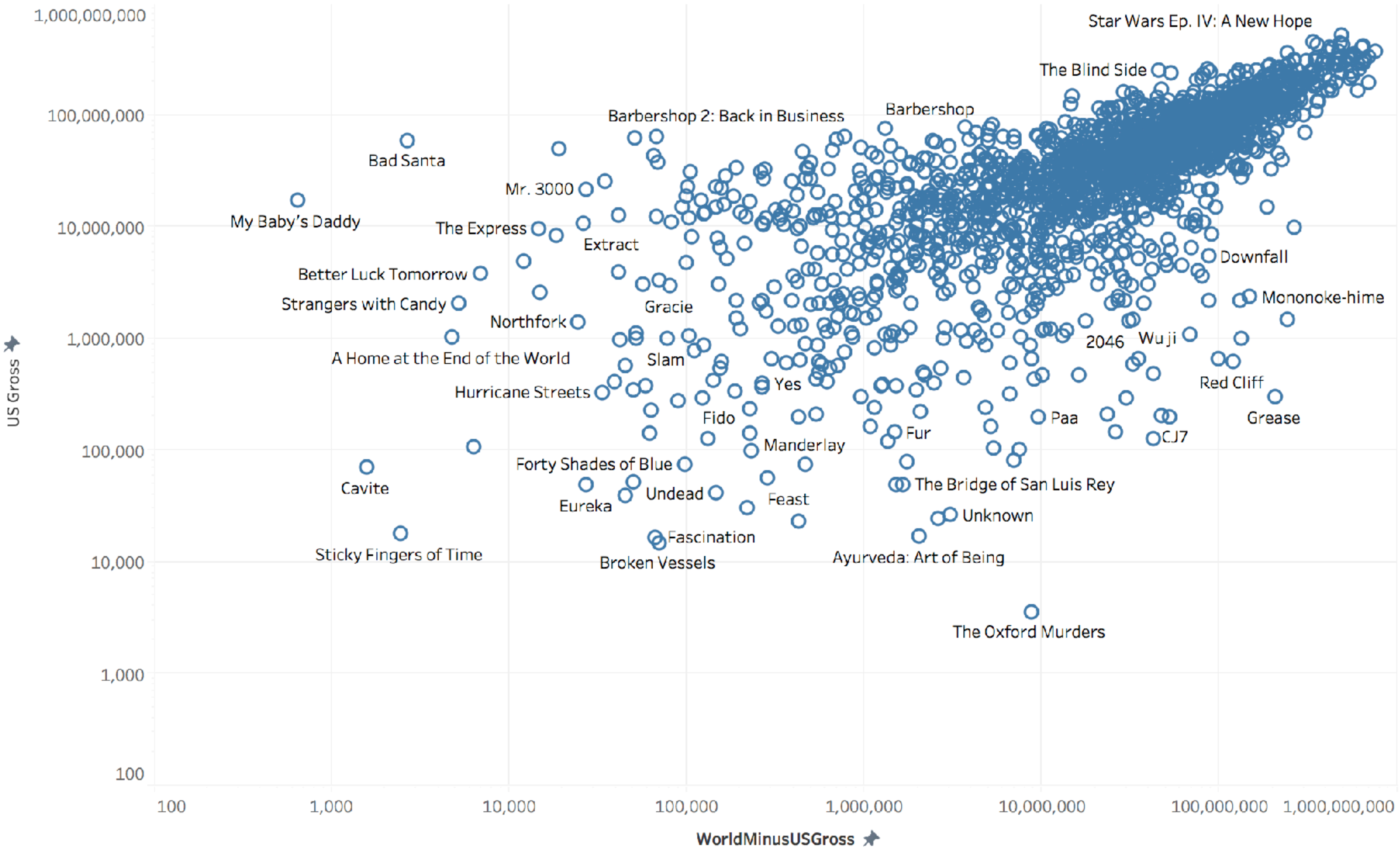
Analysis Example: Motion Pictures Data

Analysis Example: Motion Pictures Data

A sample of 3,201 movies collected in 2010.

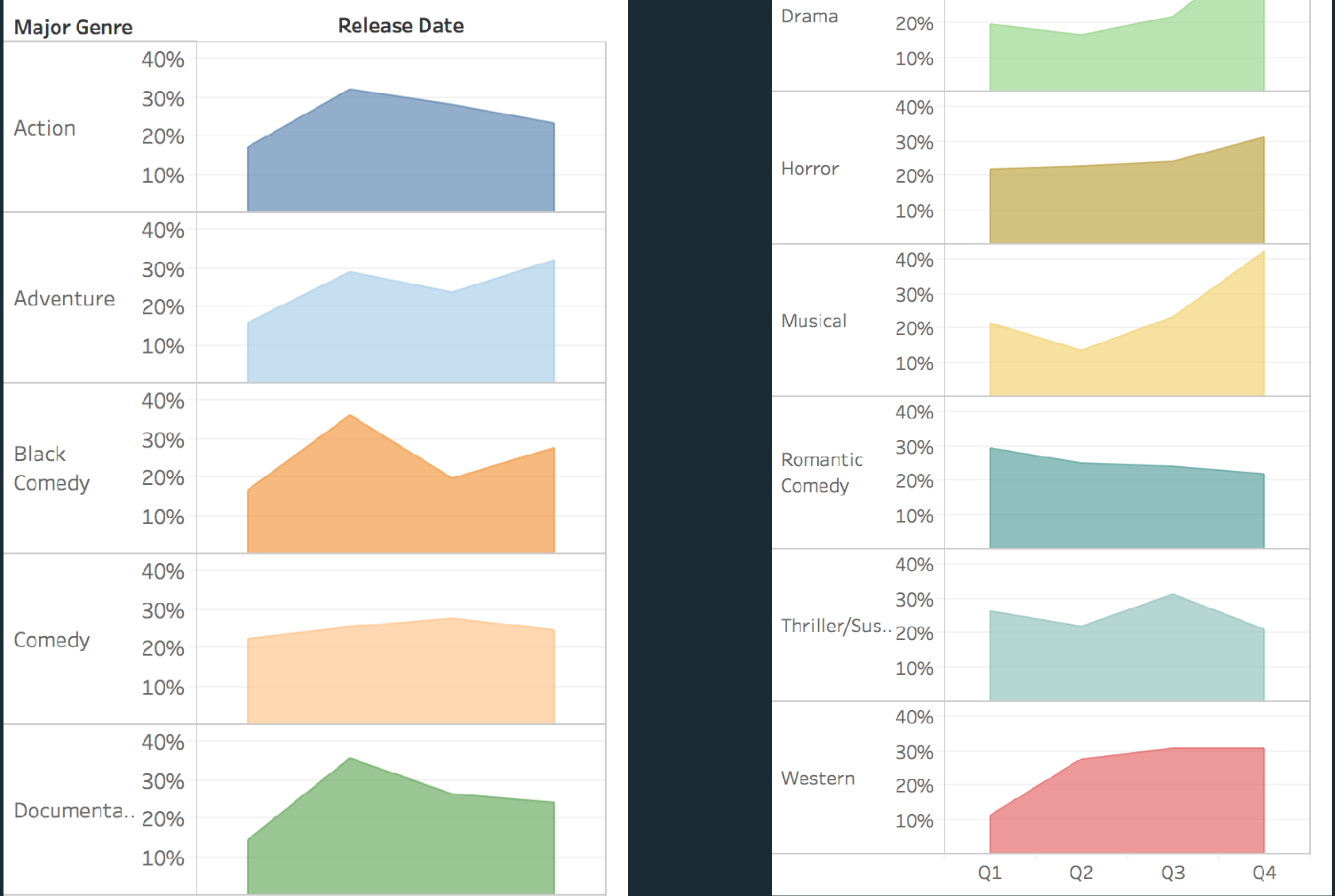
Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
Genre	String (N)
Release Date	Date (T)
US Gross	Number (Q)
Worldwide Gross	Number (Q)

US vs WW


US Gross
WorldMinusUSGross
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
100
1,000
10,000
100,000
1,000,000
10,000,000
100,000,000
1,000,000,000
Star Wars Ep. IV: A New Hope
The Blind Side
Barbershop
Barbershop 2: Back in Business
Bad Santa
Mr. 3000
My Baby's Daddy
The Express
Extract
Better Luck Tomorrow
Strangers with Candy
Gracie
Northfork
A Home at the End of the World
Slam
Hurricane Streets
Yes
Fido
Manderlay
Fur
Paa
Forty Shades of Blue
Cavite
Undead
Eureka
Feast
The Bridge of San Luis Rey
Unknown
Fascination
Sticky Fingers of Time
Broken Vessels
Ayurveda: Art of Being
The Oxford Murders
Downfall
Mononoke-hime
2046
Wu ji
Red Cliff
Grease
CJ7

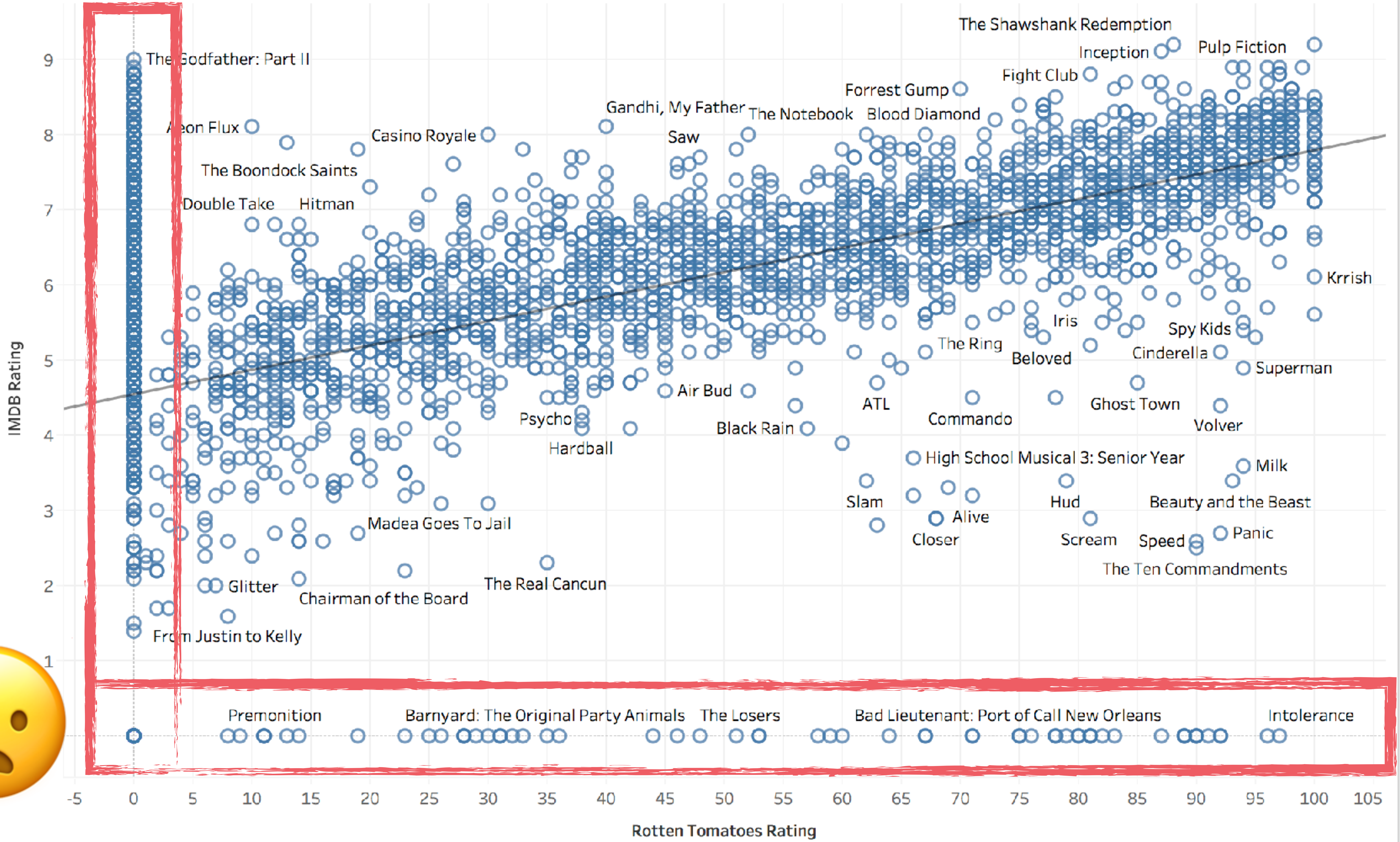
Distribution of US Gross by Genre

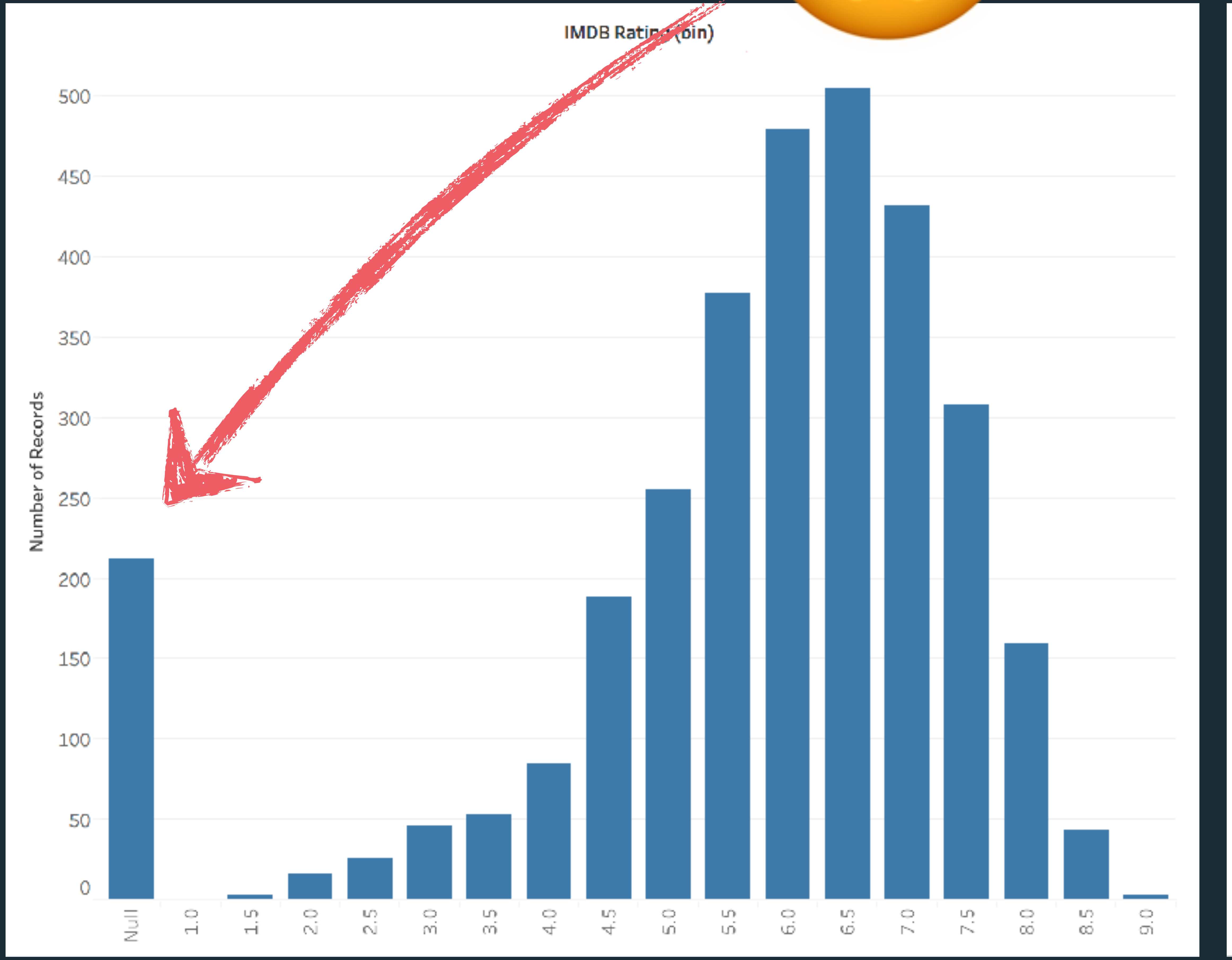
% Releases per Quarter per Genre

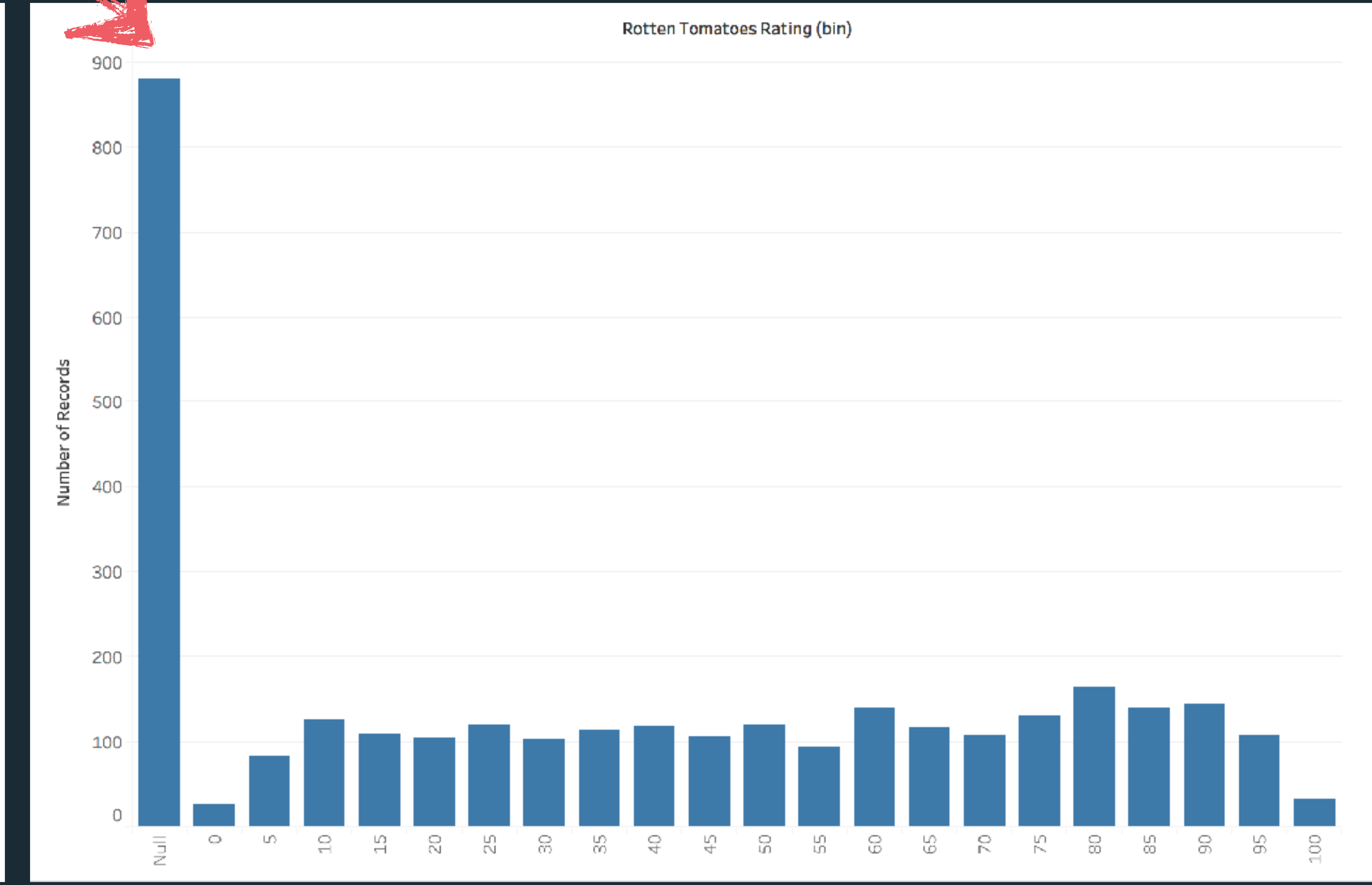


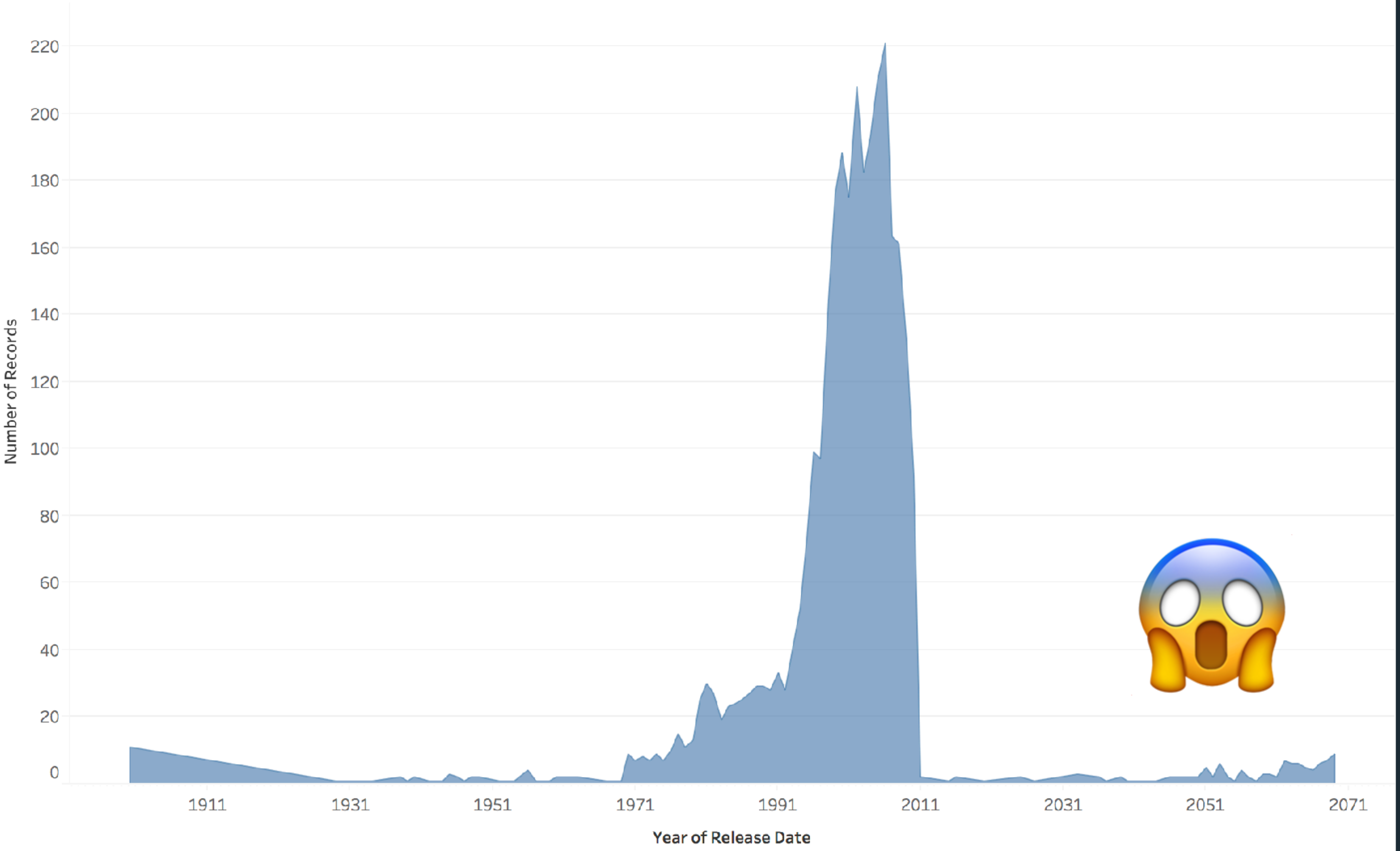
US Gross by Ratings

Audience vs. Critic Ratings




IMDB Rating (bin)
Number of Records
500
450
400
350
300
250
200
150
100
50
0
Null
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0


Rotten Tomatoes Rating (bin)
Number of Records
900
800
700
600
500
400
300
200
100
0
Null
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100



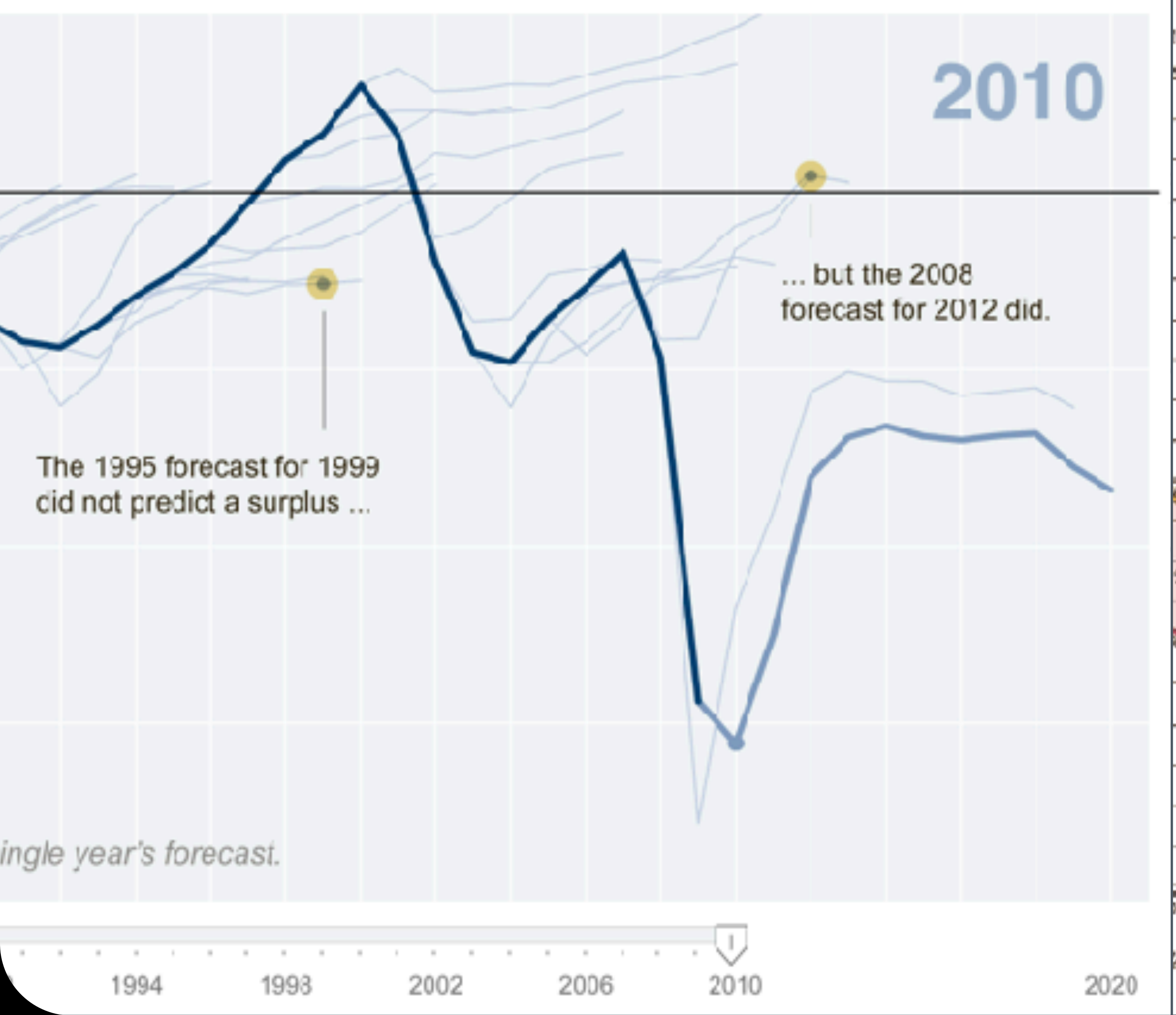
220
200
180
160
140
120
100
80
60
40
20
0
Number of Records
1911
1931
1951
1971
1991
2011
2031
2051
2071
Year of Release Date

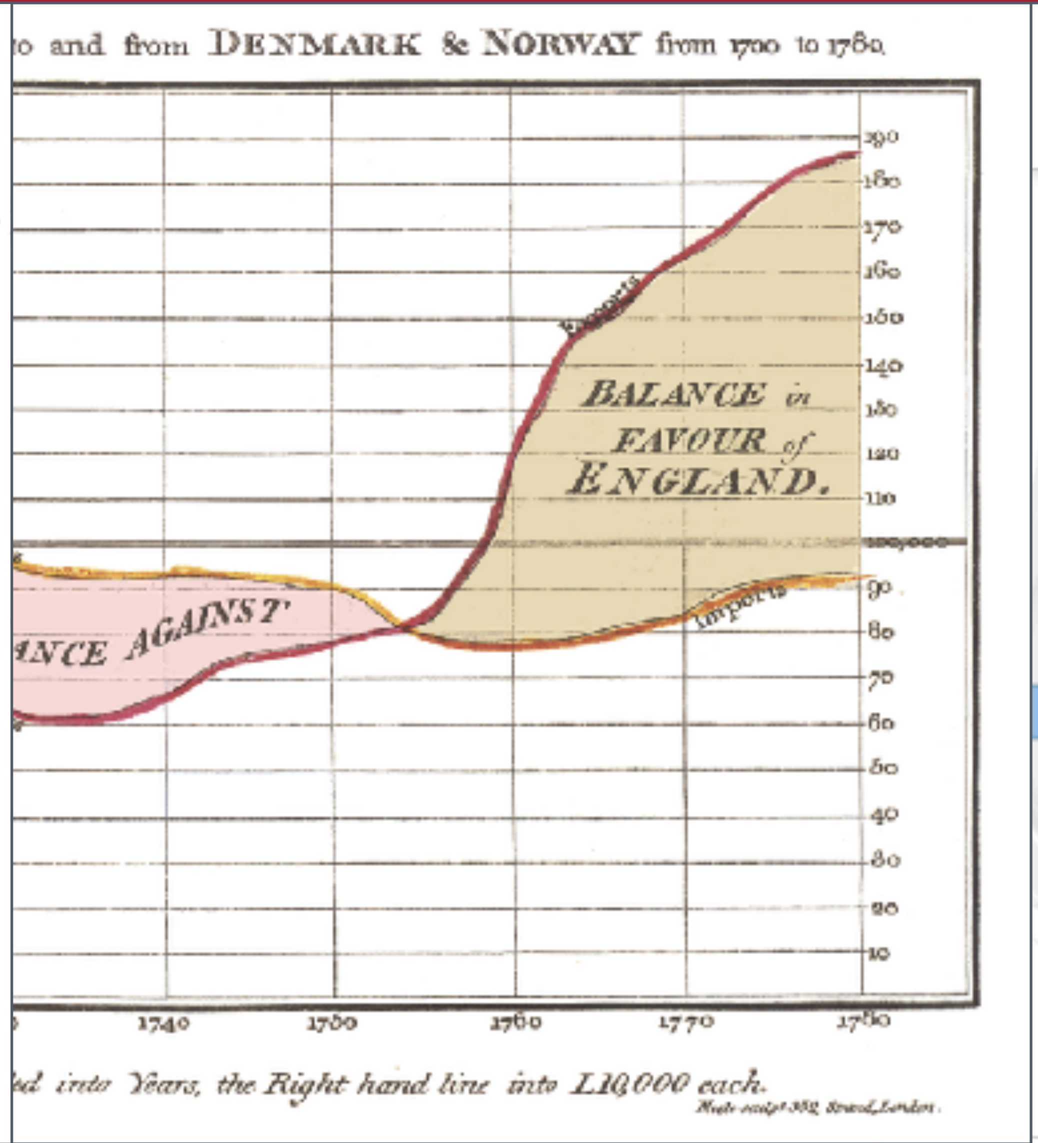
6.859: Interactive Data Visualization

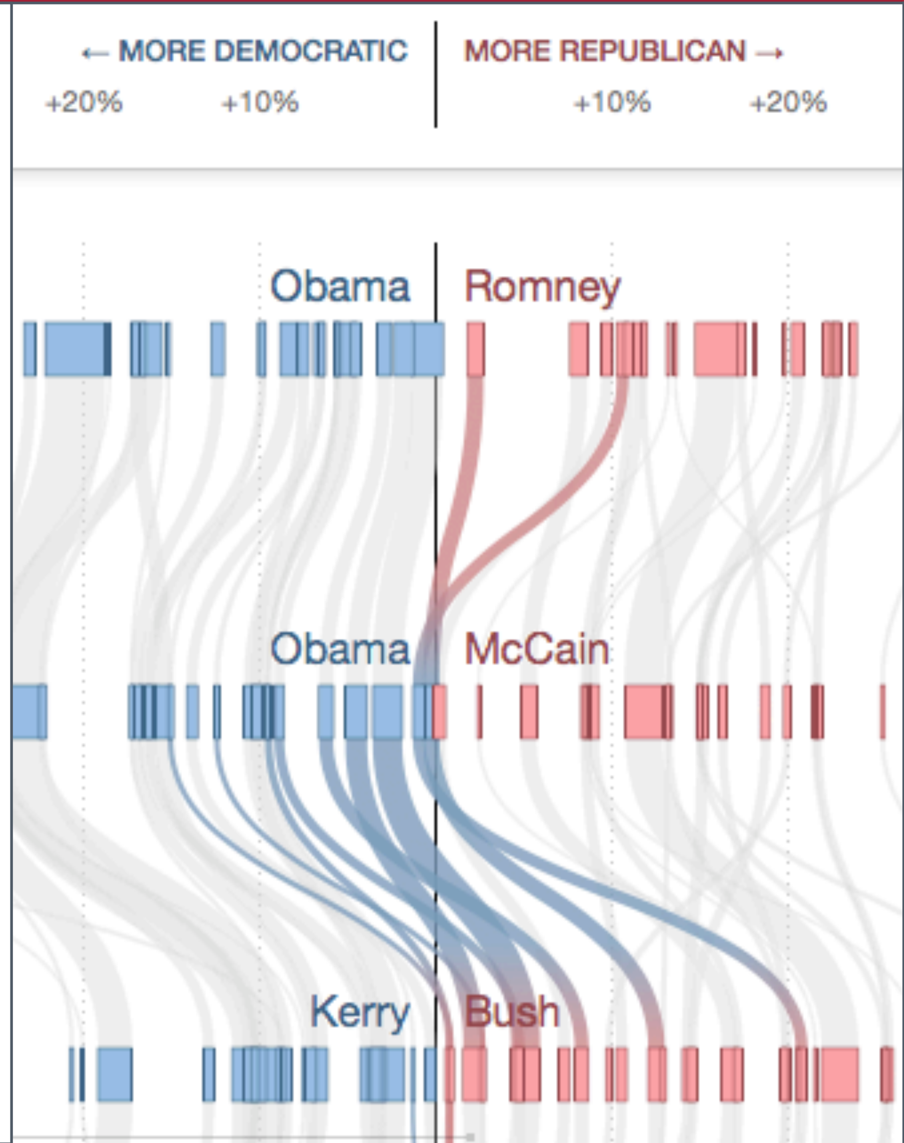
Exploratory Data Analysis

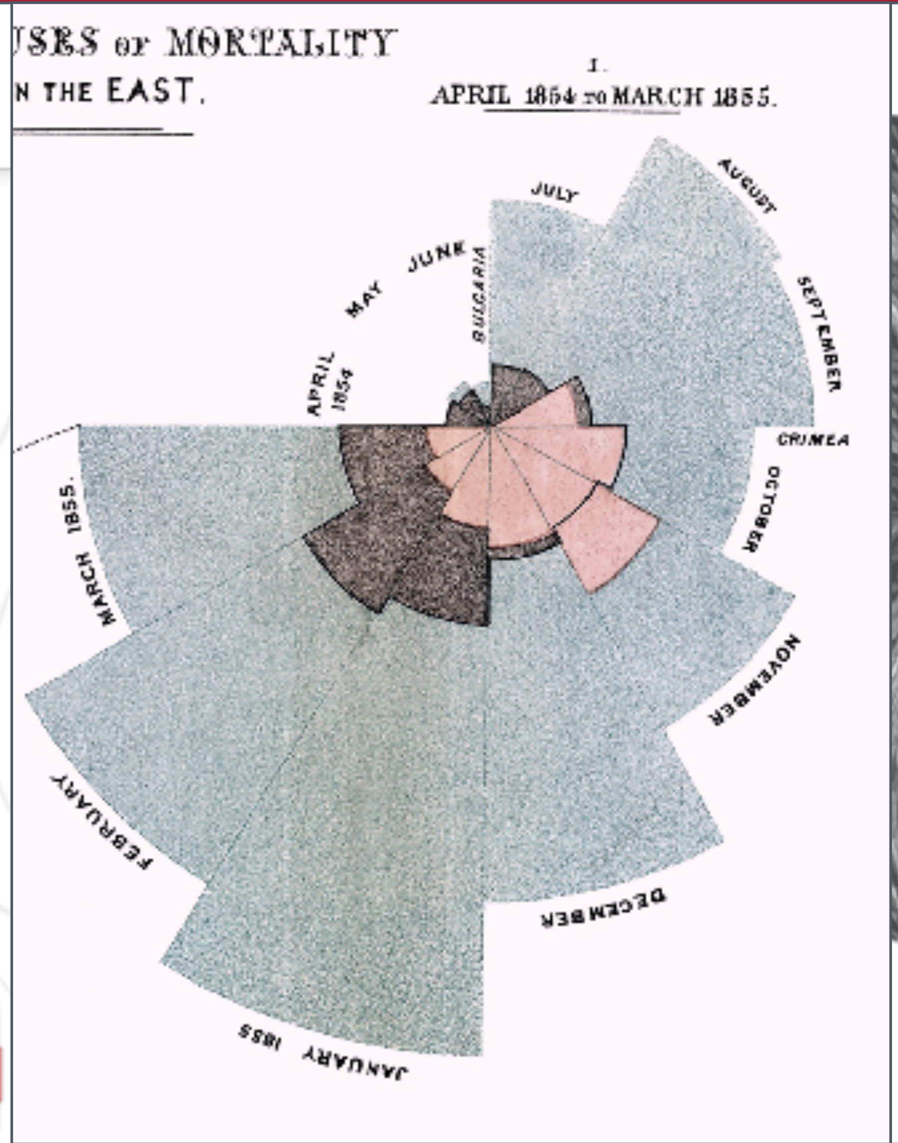
Arvind Satyanarayan

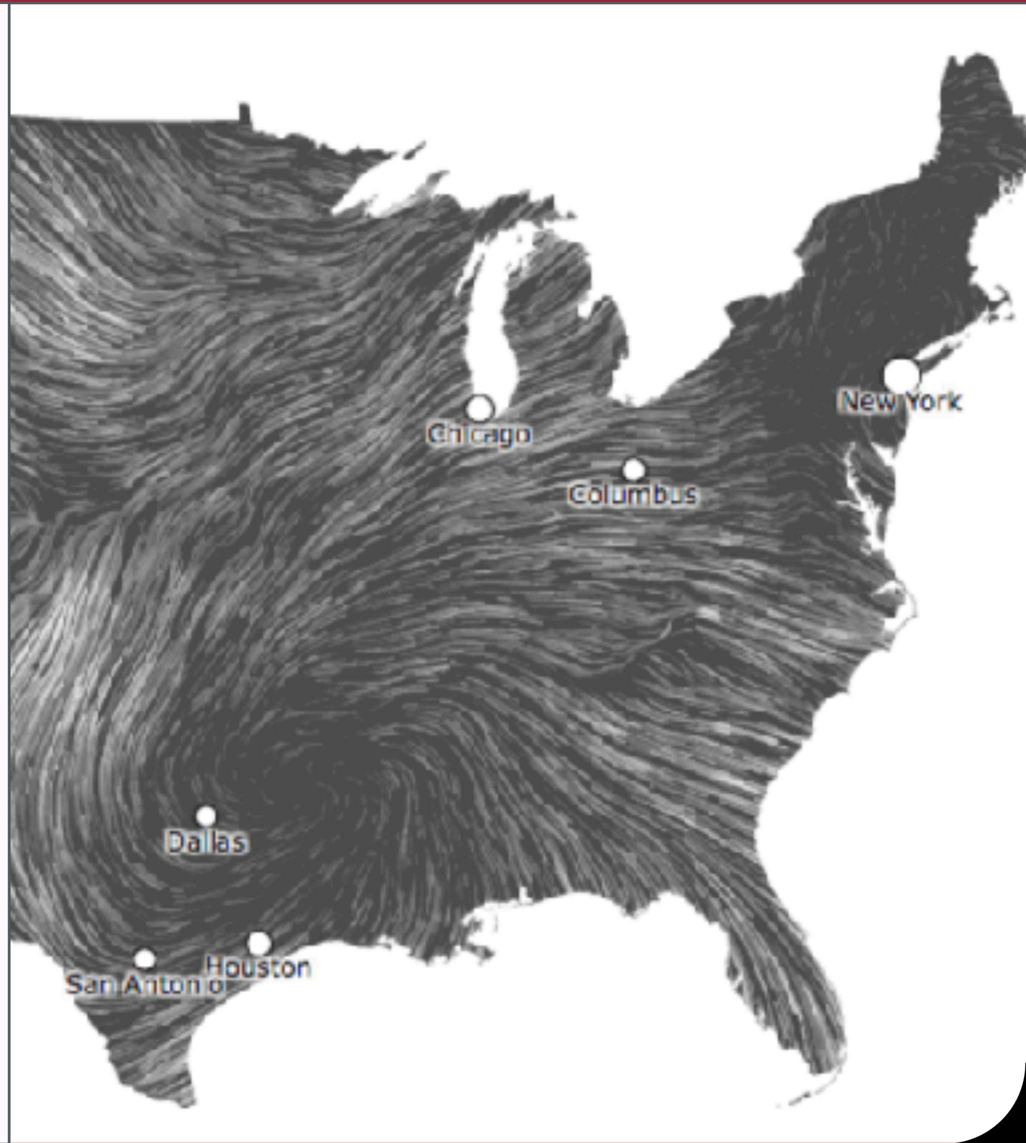
Download data for today's activity:
www.yellkey.com/free











Exploratory Visual Analysis

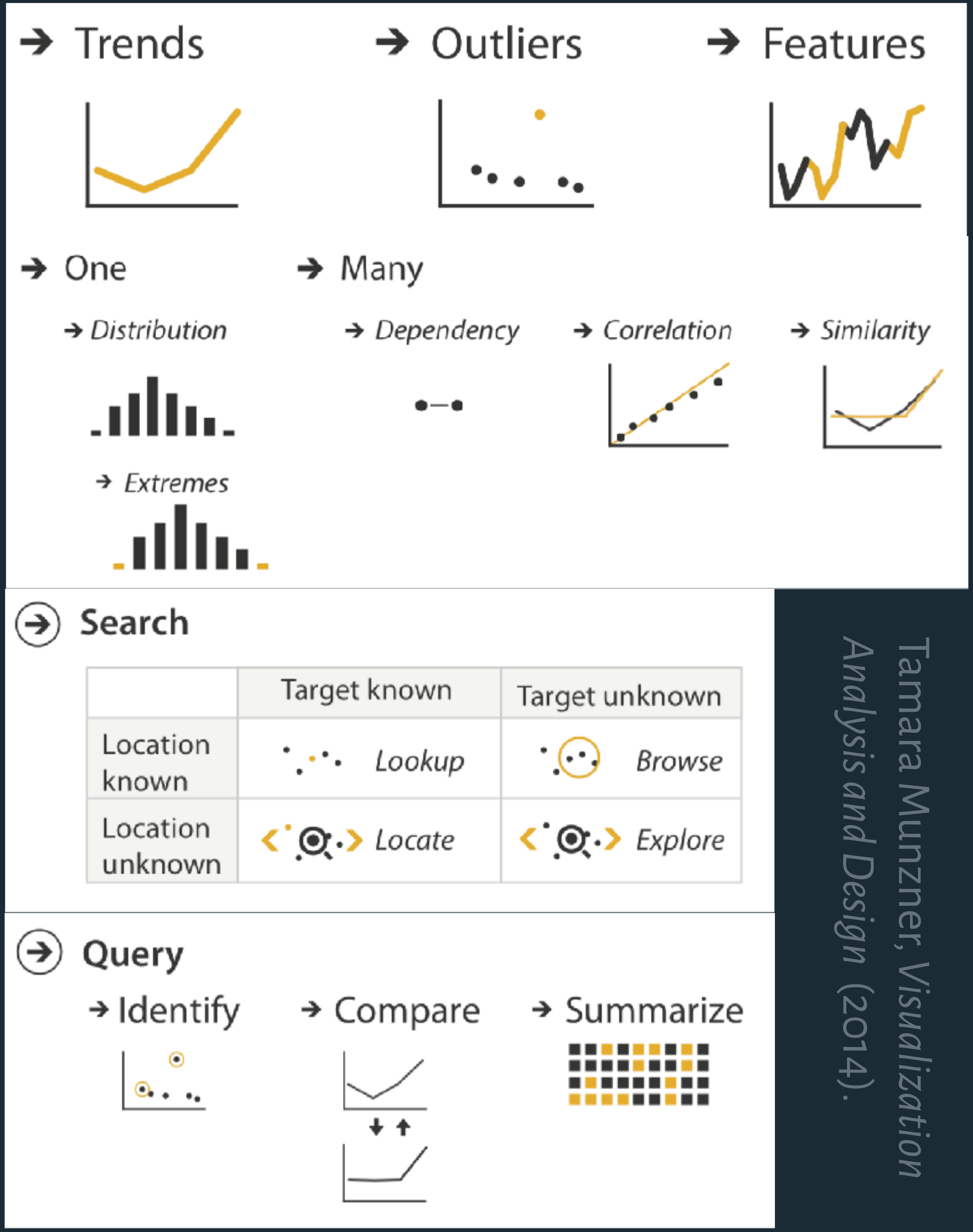
Process

1. Construct graphics to address questions.
2. Inspect "answer" and ask new questions.
3. Iterate…

Lessons

✓ Check **data quality** and your **assumptions**.

✓ Start with **univariate summaries**, then consider **relationships between variables**.



Tamara Munzner, Visualization Analysis and Design (2014).

Analysis Example:
Antibiotic Effectiveness

Analysis Example: Antibiotic Effectiveness

Collected prior to 1951

Genus of Bacteria	String (N)
Species of Bacteria	String (N)
Antibiotic Applied	String (N)
Gram-Staining?	Pos / Neg (N)
Min. Inhibitory Con. (g)	Number (Q)

Analysis Example: Antibiotic Effectiveness

Collected prior to 1951

Genus of Bacteria	String (N)
Species of Bacteria	String (N)
Antibiotic Applied	String (N)
Gram-Staining?	Pos / Neg (N)
Min. Inhibitory Con. (g)	Number (Q)

Table 1—Burtin's Data

	Antibiotic			
Bacteria	Penicillin	Streptomycin	Neomycin	Gram Staining
Aerobacter aerogenes	870	1	1.6	negative
Brucella abortus	1	2	0.02	negative
Brucella anthracis	0.001	0.01	0.007	positive
Diplococcus pneumoniae	0.005	11	10	positive
Escherichia coli	100	0.4	0.1	negative
Klebsiella pneumoniae	850	1.2	1	negative
Mycobacterium tuberculosis	800	5	2	negative
Proteus vulgaris	3	0.1	0.1	negative
Pseudomonas aeruginosa	850	2	0.4	negative
Salmonella (Eberthella) typhosa	1	0.4	0.008	negative
Salmonella schottmuelleri	10	0.8	0.09	negative
Staphylococcus albus	0.007	0.1	0.001	positive
Staphylococcus aureus	0.03	0.03	0.001	positive
Streptococcus fecalis	1	1	0.1	positive
Streptococcus hemolyticus	0.001	14	10	positive
Streptococcus viridans	0.005	10	40	positive

What questions might we ask?

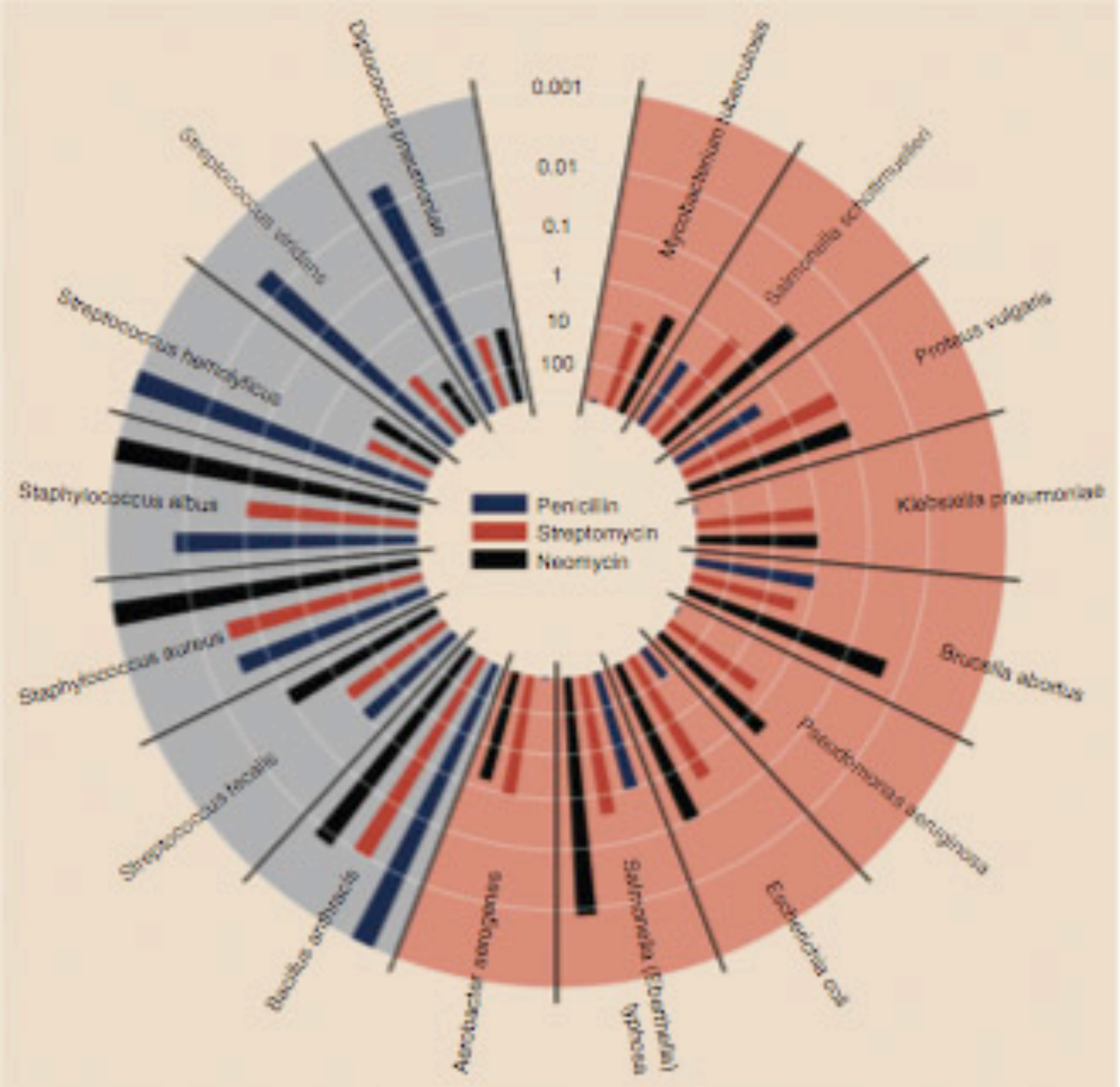
Collected prior to 1951

Genus of Bacteria	String (N)
Species of Bacteria	String (N)
Antibiotic Applied	String (N)
Gram-Staining?	Pos / Neg (N)
Min. Inhibitory Con. (g)	Number (Q)

Table 1—Burtin's Data

	Antibiotic			
Bacteria	**Penicillin**	**Streptomycin**	**Neomycin**	**Gram Staining**
Aerobacter aerogenes	870	1	1.6	negative
Brucella abortus	1	2	0.02	negative
Brucella anthracis	0.001	0.01	0.007	positive
Diplococcus pneumoniae	0.005	11	10	positive
Escherichia coli	100	0.4	0.1	negative
Klebsiella pneumoniae	850	1.2	1	negative
Mycobacterium tuberculosis	800	5	2	negative
Proteus vulgaris	3	0.1	0.1	negative
Pseudomonas aeruginosa	850	2	0.4	negative
Salmonella (Eberthella) typhosa	1	0.4	0.008	negative
Salmonella schottmuelleri	10	0.8	0.09	negative
Staphylococcus albus	0.007	0.1	0.001	positive
Staphylococcus aureus	0.03	0.03	0.001	positive
Streptococcus fecalis	1	1	0.1	positive
Streptococcus hemolyticus	0.001	14	10	positive
Streptococcus viridans	0.005	10	40	positive

How do the drugs compare?



Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

Original graphic by Will Burtin, 1951.

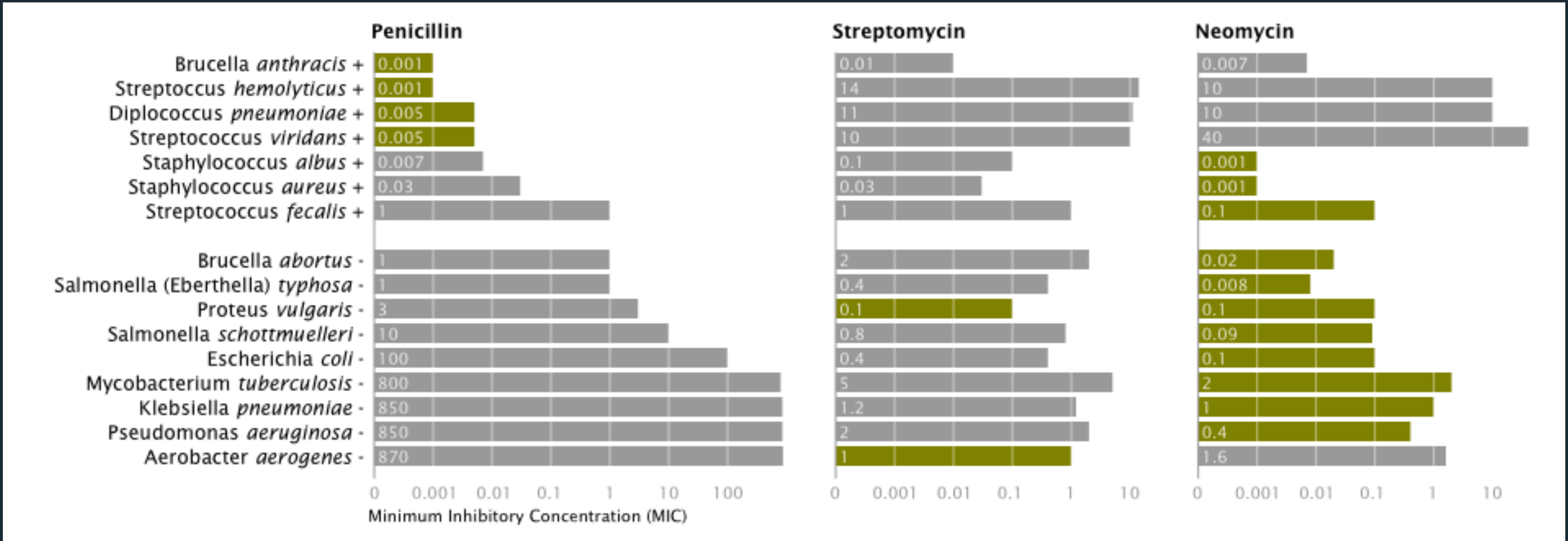
Encodings

Radius: 1 / log(MIC)

Bar Color: Antibiotic

Background Color: Gram Staining

How do the drugs compare?



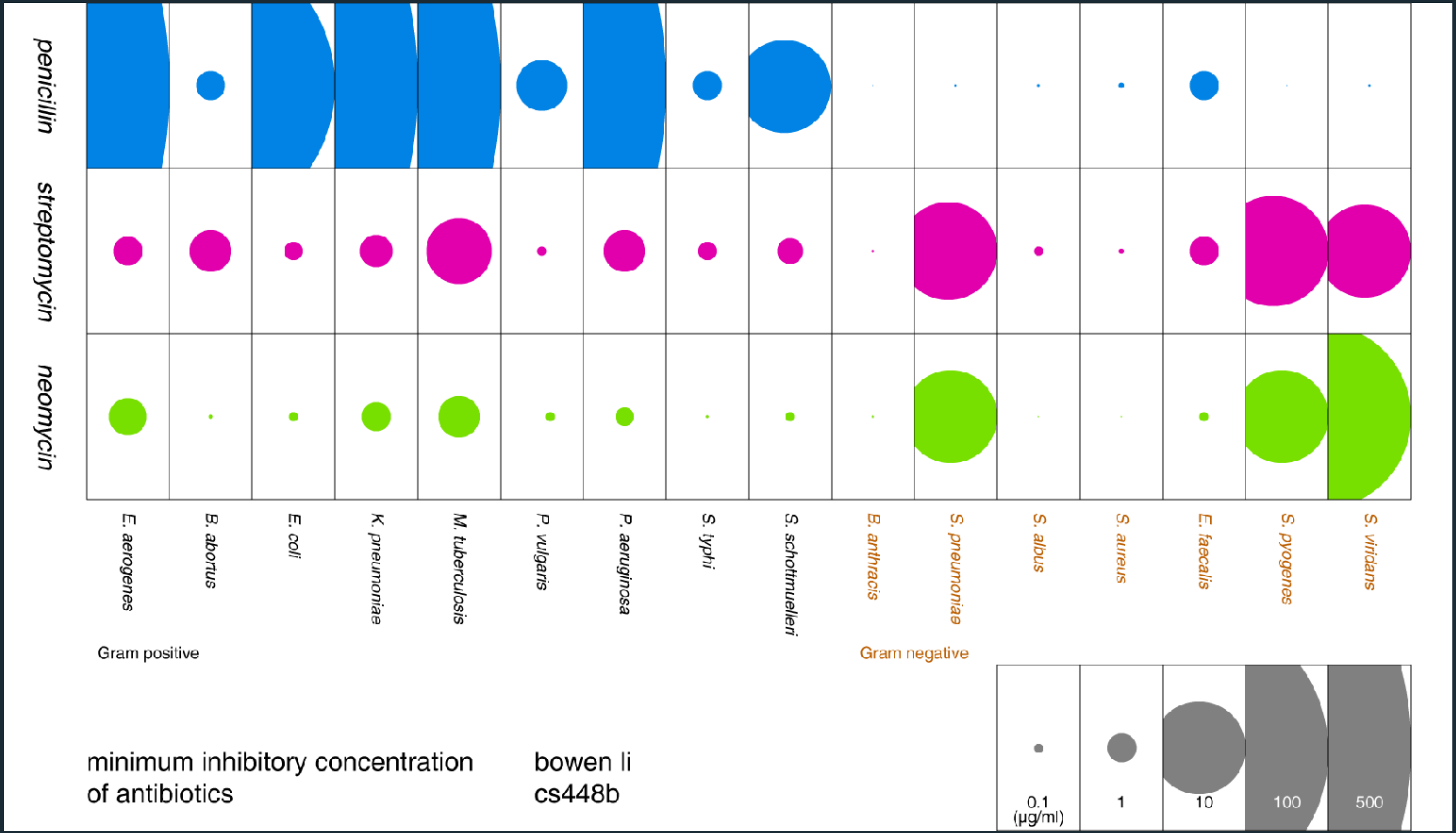
Mike Bostock, *Stanford CS448b* (Winter 2009).

X-Axis: Antibiotic | log(MIC)

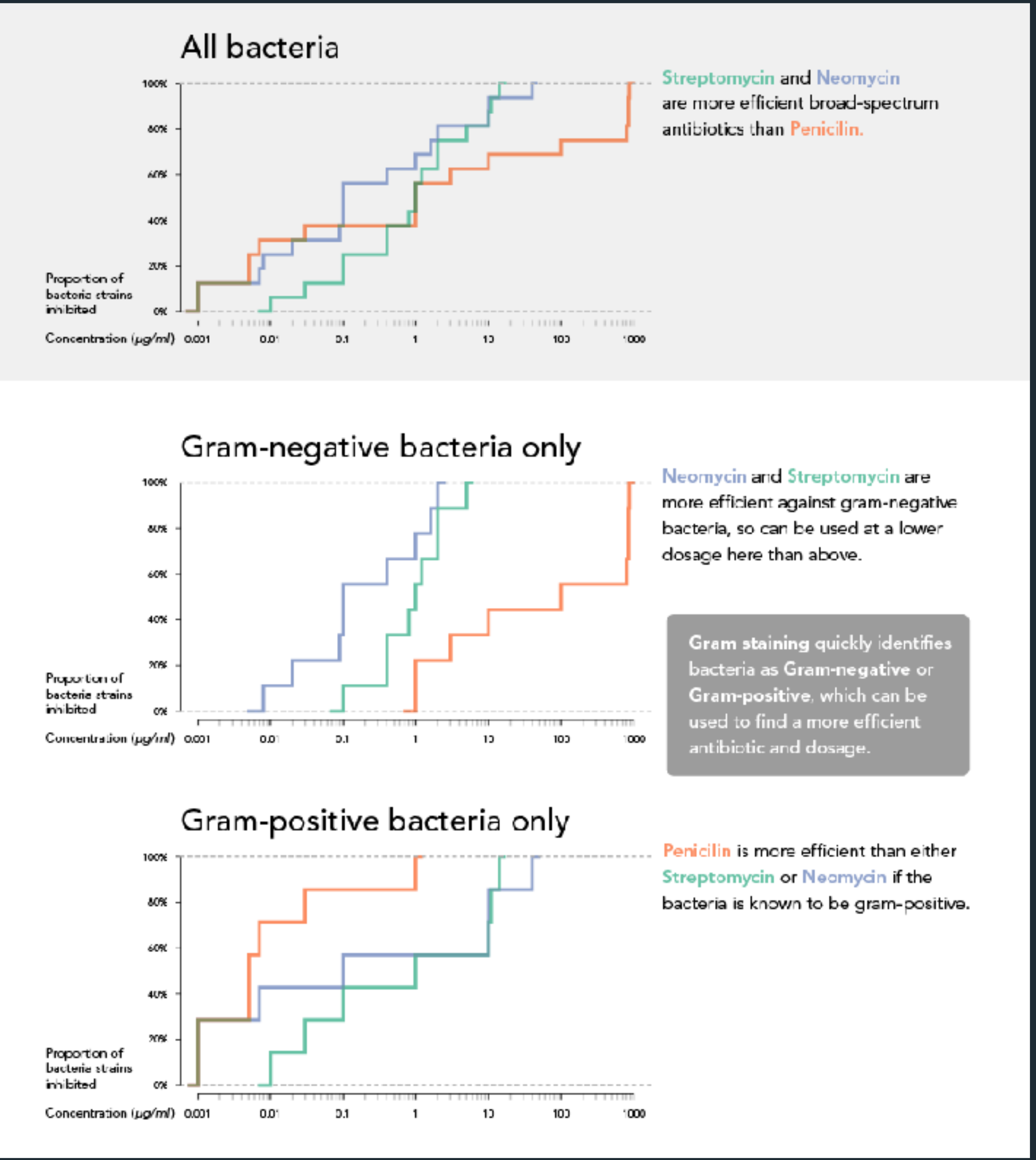
Y-Axis: Gram-Staining | Species

Color: Most Effective?

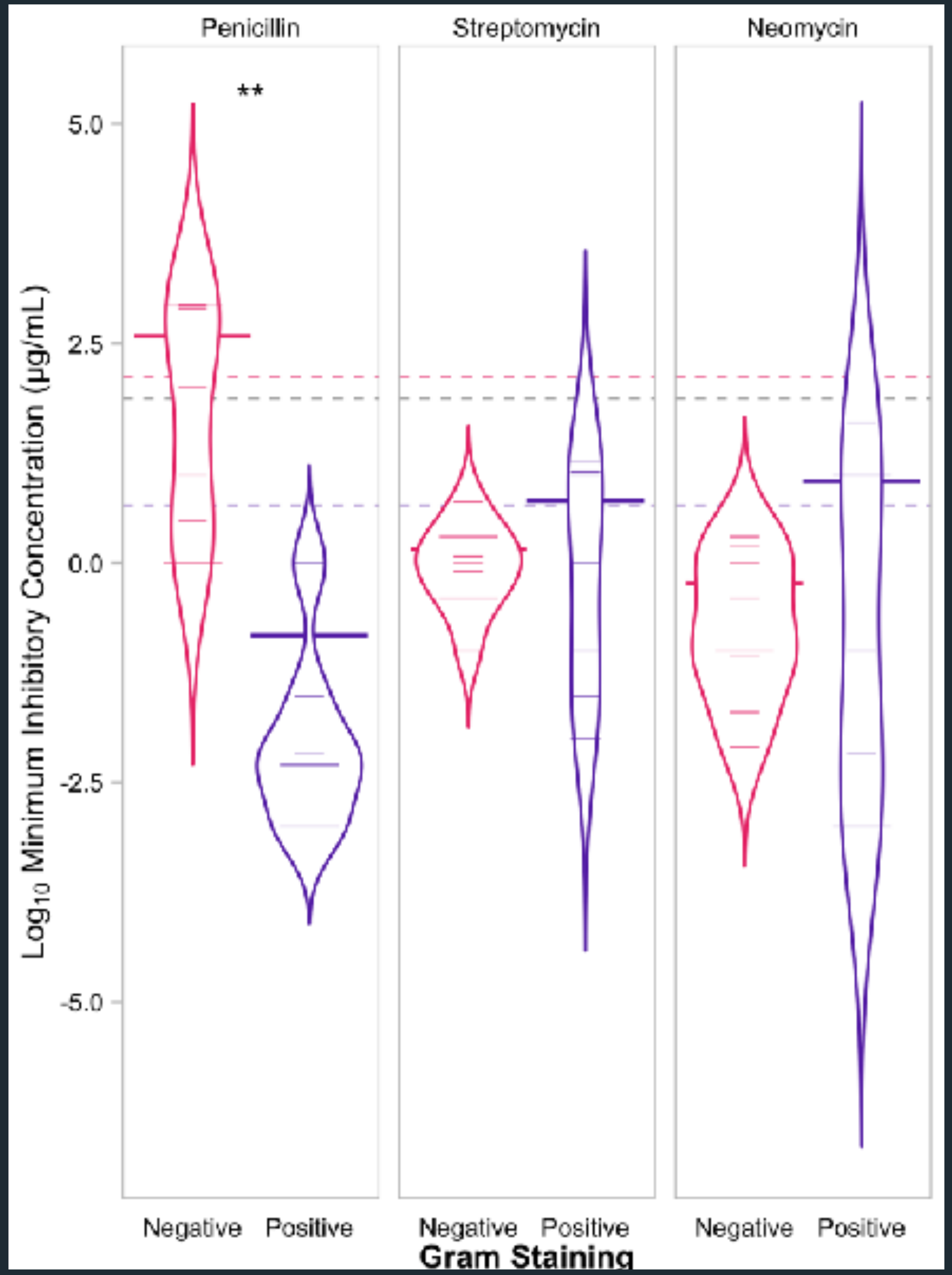
How do the drugs compare?



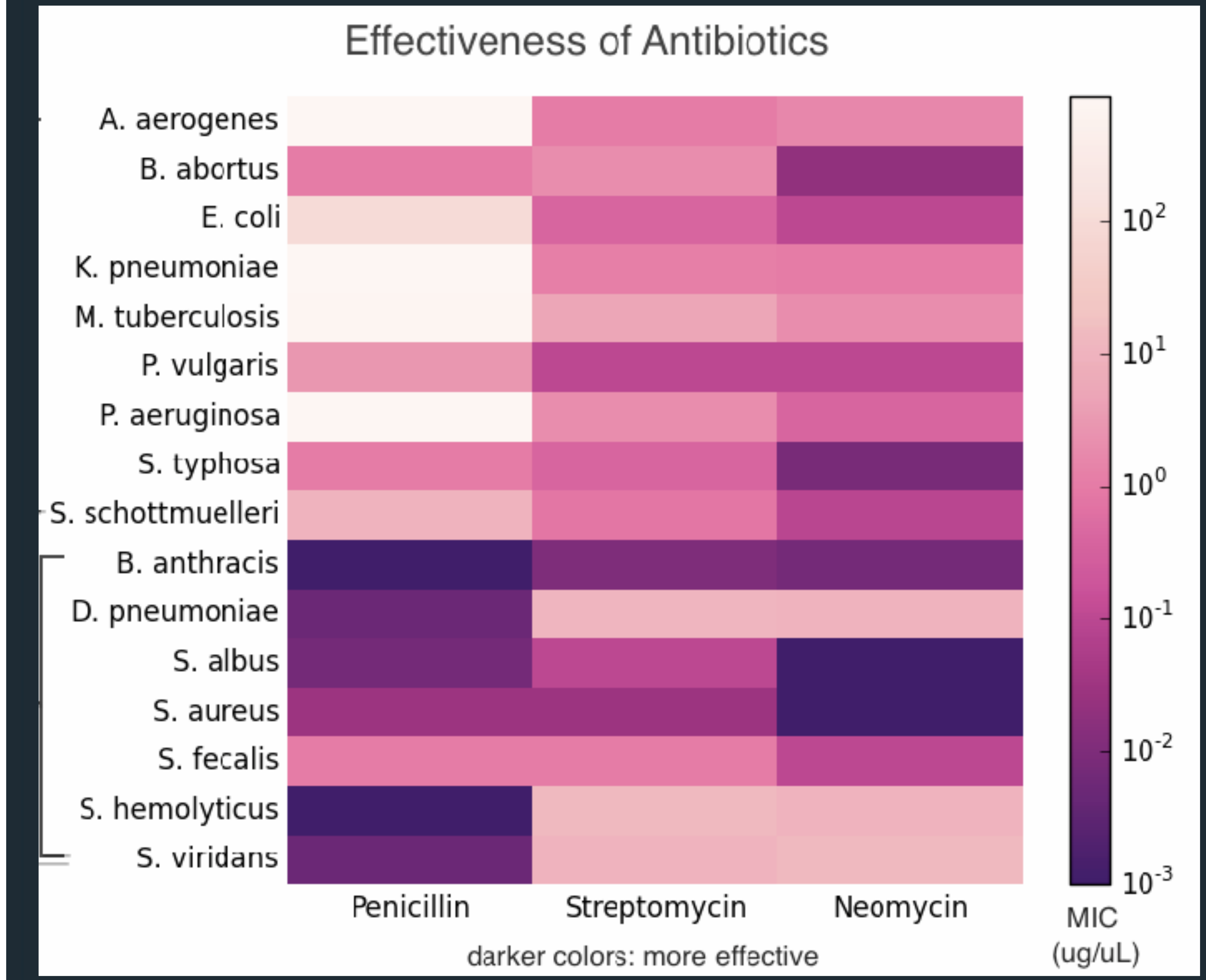
Bowen Li, *Stanford* CS448b (Fall 2009).



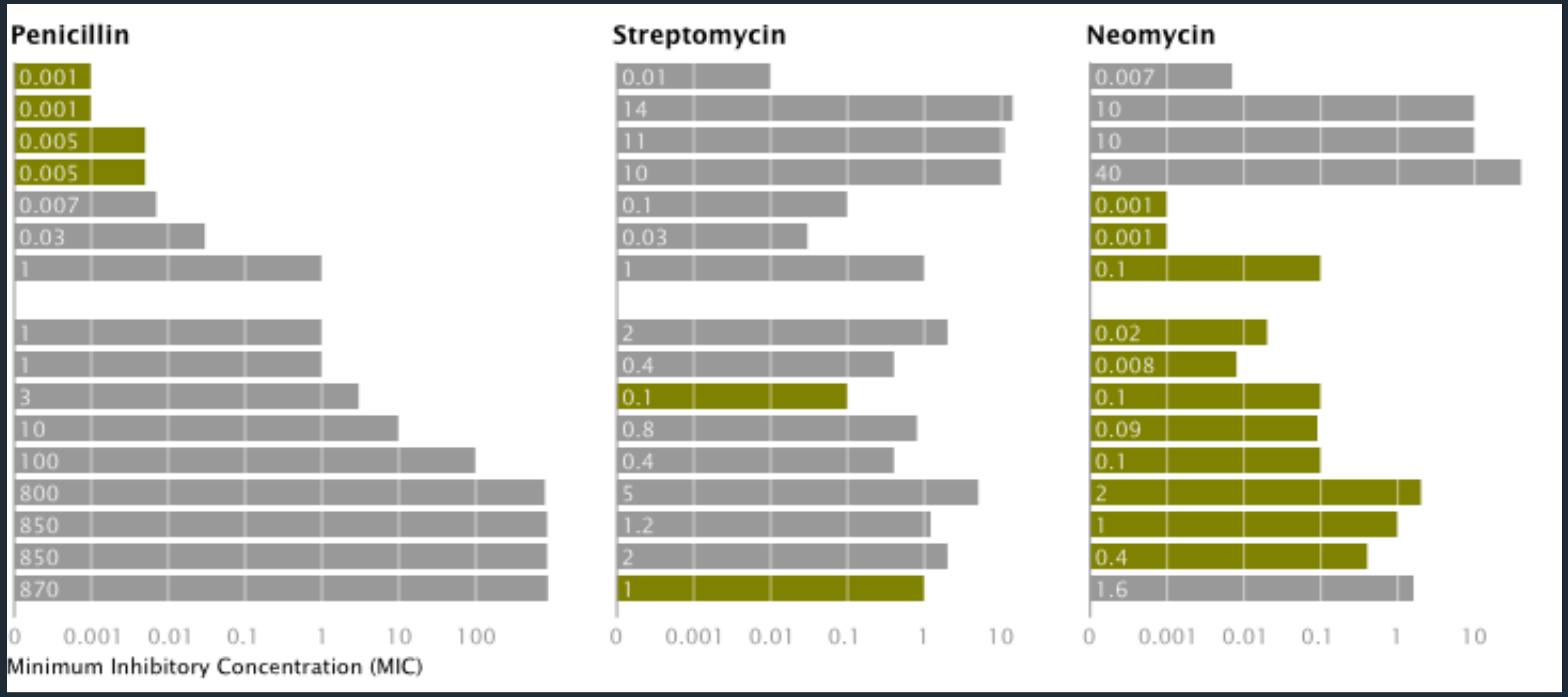
All bacteria
Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.
Proportion of bacteria strains inhibited
Concentration (µg/ml)
Gram-negative bacteria only
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.
Gram staining quickly identifies bacteria as Gram-negative or Gram-positive, which can be used to find a more efficient antibiotic and dosage.
Gram-positive bacteria only
Penicilin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.



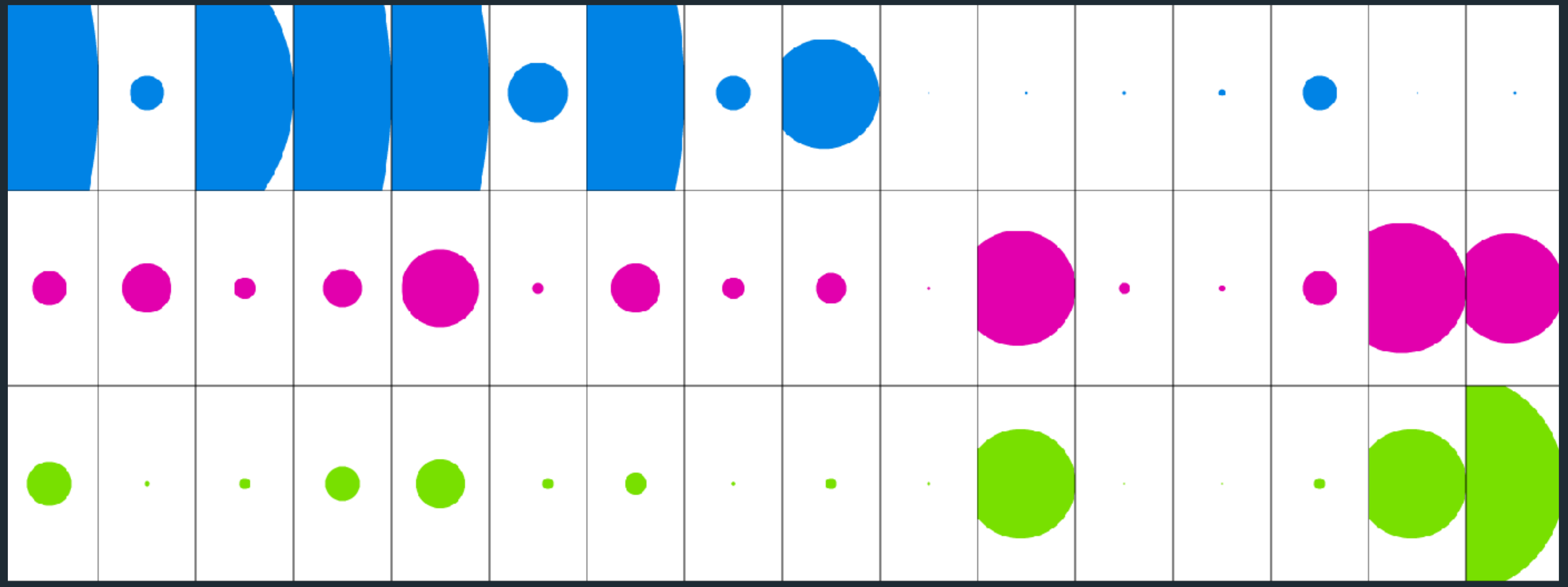
Penicillin
Streptomycin
Neomycin
**
5.0
2.5
0.0
-2.5
-5.0
Log10 Minimum Inhibitory Concentration (µg/mL)
Negative
Positive
Gram Staining



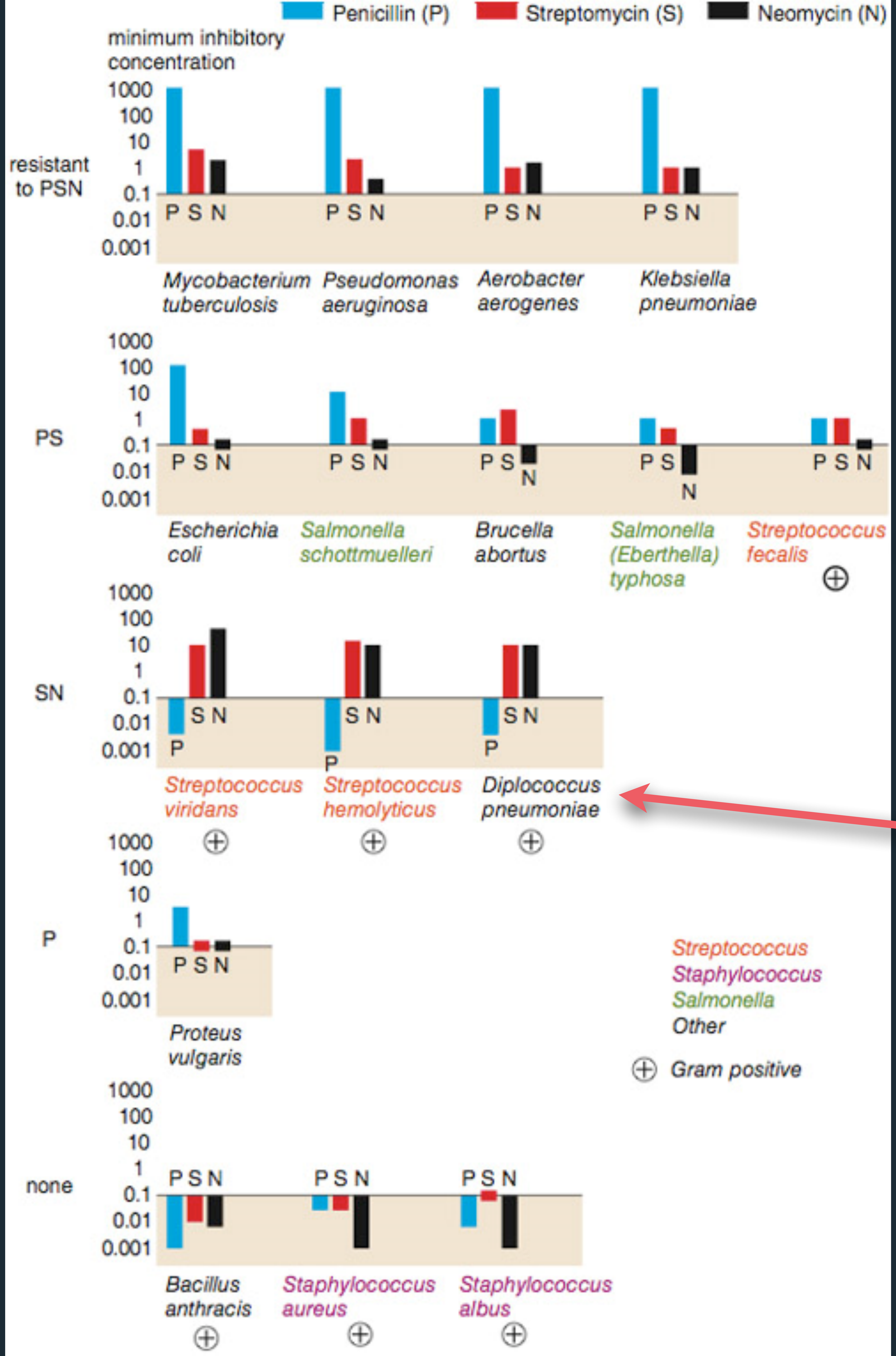
Effectiveness of Antibiotics
A. aerogenes
B. abortus
E. coli
K. pneumoniae
M. tuberculosis
P. vulgaris
P. aeruginosa
S. typhosa
S. schottmuelleri
B. anthracis
D. pneumoniae
S. albus
S. aureus
S. fecalis
S. hemolyticus
S. viridans
Penicillin
Streptomycin
Neomycin
darker colors: more effective
MIC (ug/uL)



Penicillin
Streptomycin
Neomycin
Minimum Inhibitory Concentration (MIC)



Which antibiotic should one use?



Wainer & Lysen. *American Scientist*, 2009

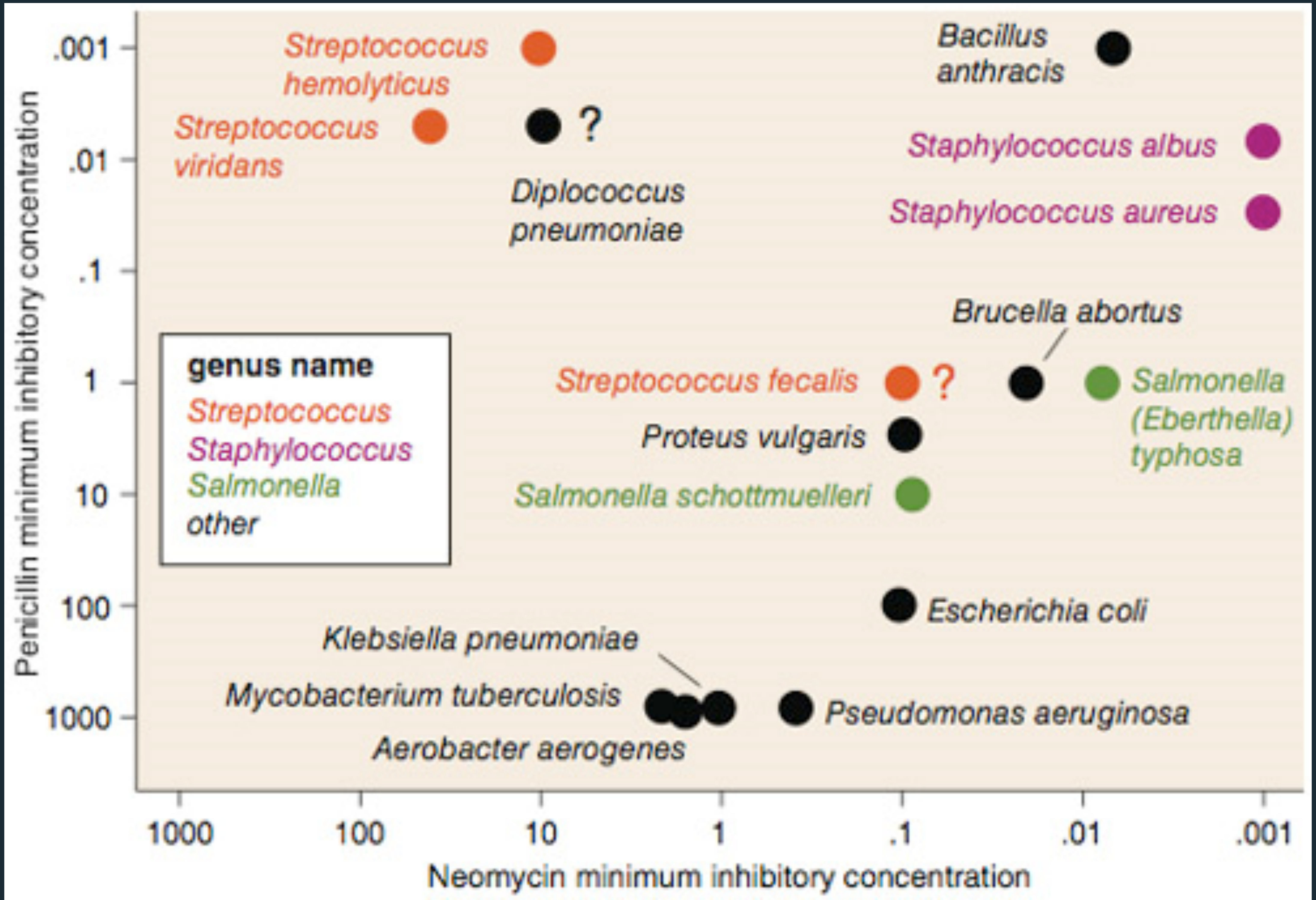
Do the bacteria group by antibiotic resistance?

Not a streptococcus!
(realized ~30 yrs later)

Really a streptococcus!
(realized ~20 yrs later)

Do the bacteria group by resistance?
Do different drugs correlate?

Wainer & Lysen. *American Scientist*, 2009



Exploratory Visual Analysis

Process

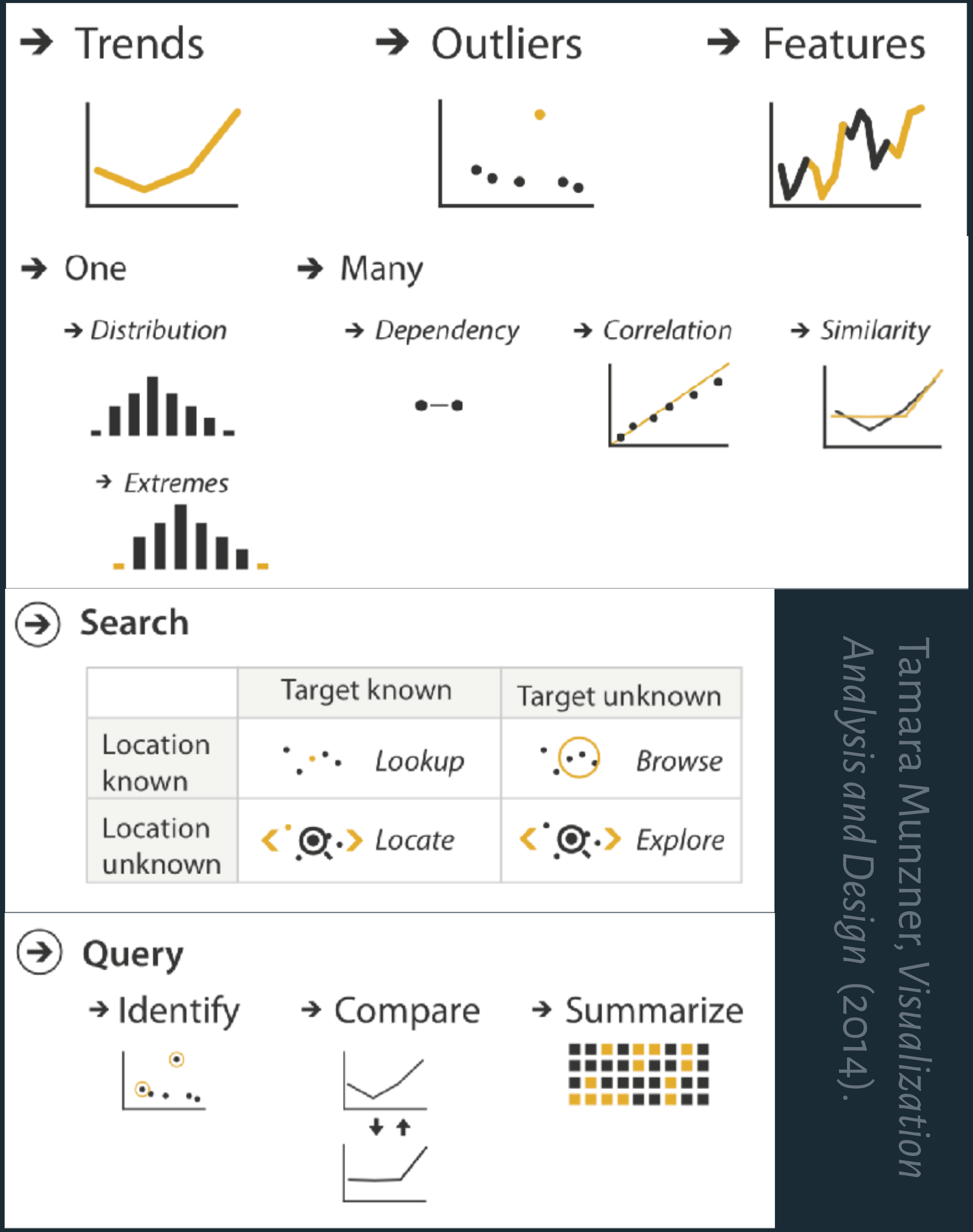
1. Construct graphics to address questions.
2. Inspect "answer" and ask new questions.
3. Iterate…

Lessons

✓ Check **data quality** and your **assumptions**.

✓ Start with **univariate summaries**, then consider **relationships between variables**.

✓ Avoid **premature fixation:** balance **data variation** and **design variation**.



Tamara Munzner, *Visualization Analysis and Design* (2014).

Is EDA/EVA fishing?

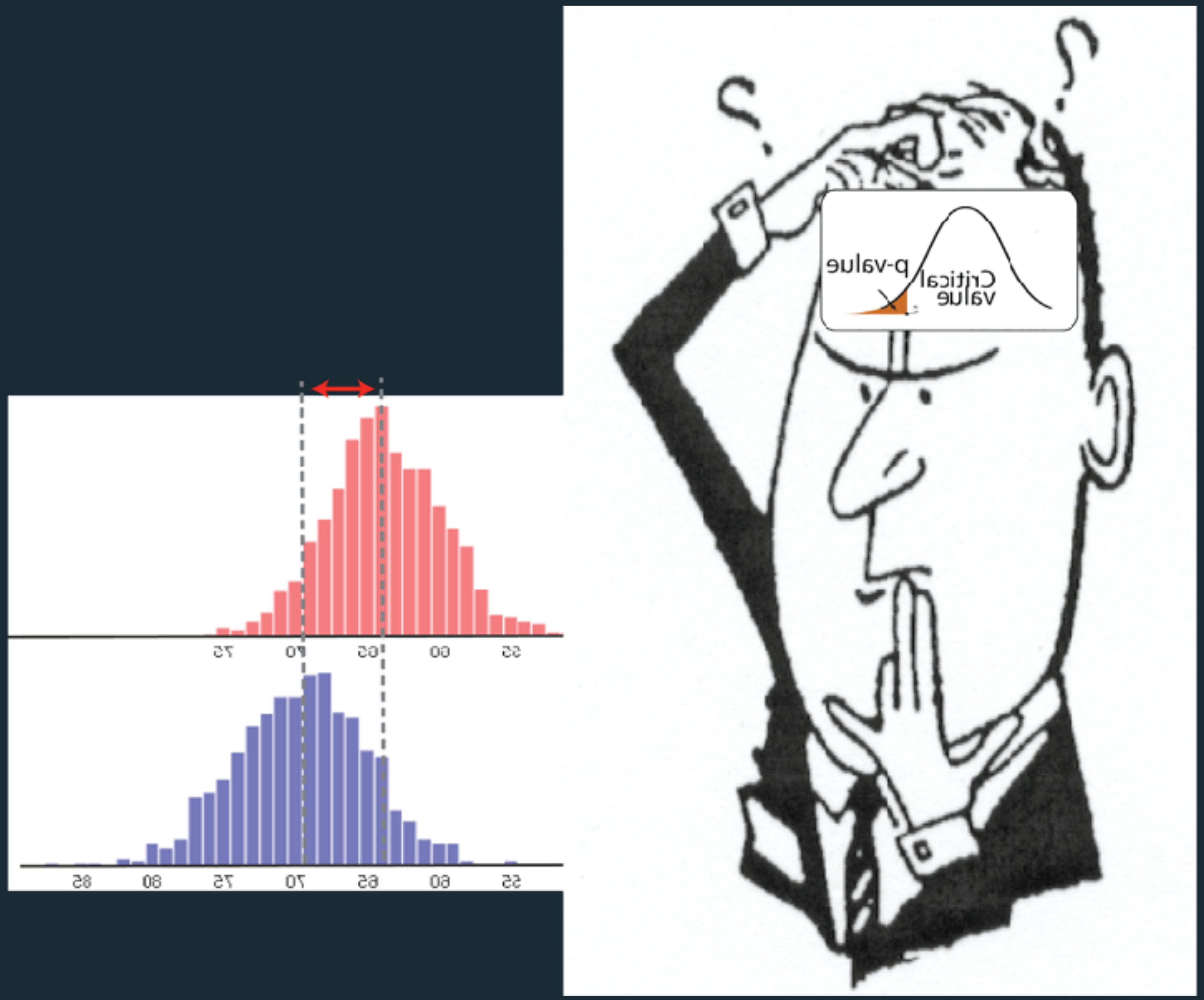
🎣

Is EDA/EVA Fishing? 🎣

Some statisticians have proposed that when we look for patterns in visualizations, we're doing a series of *visual hypothesis tests*.

Multiple comparisons problem: the more hypothesis tests, the greater the chance of a cpurious finding (since the Null Hypothesis Significant Test admits 5% false positives).

No, because there's not a clear separation between *exploratory* and *confirmatory* analysis.

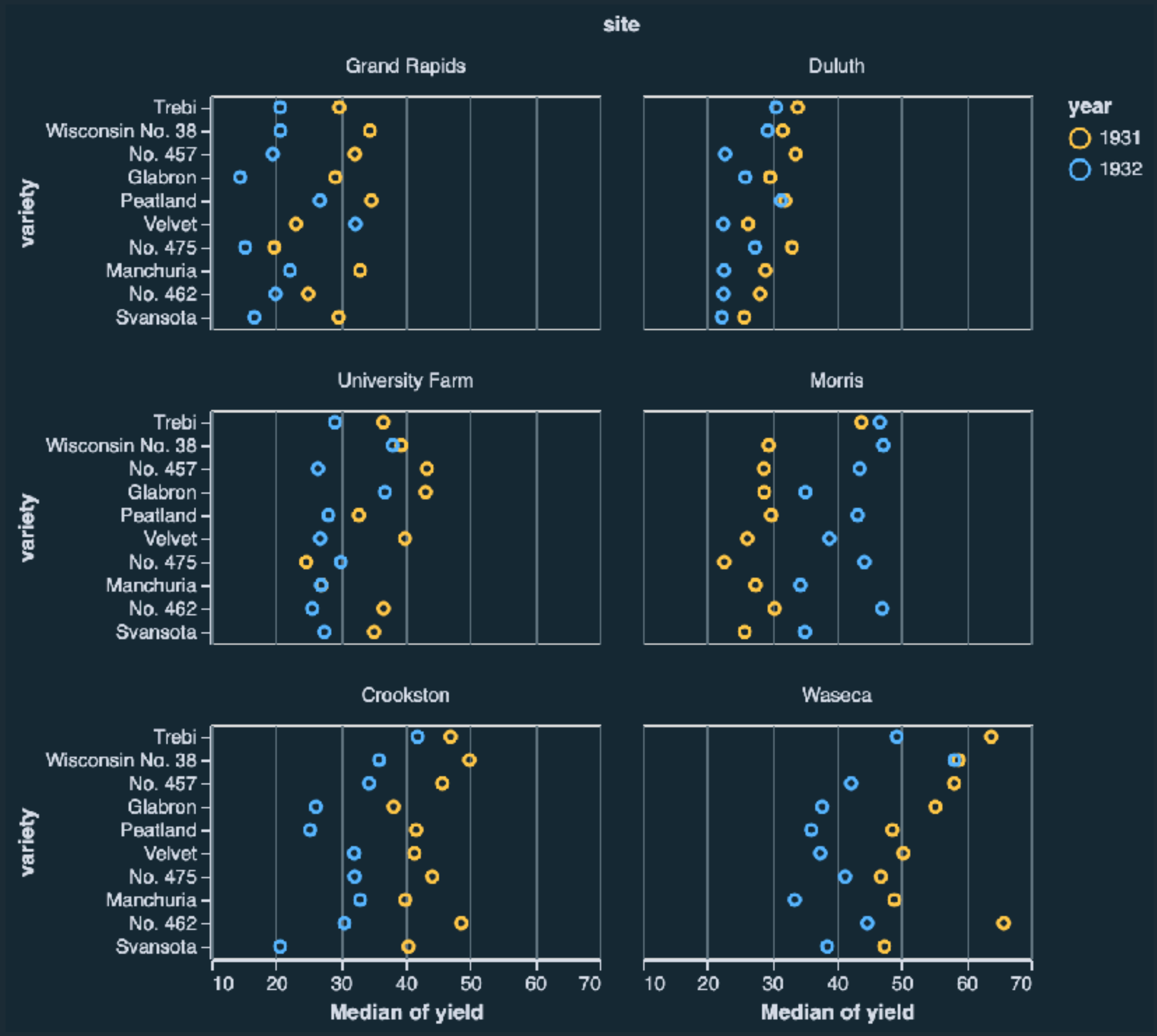


Based on slides by Jessica Hullman

Is EDA/EVA Fishing? 🎣

Some statisticians have proposed that when we look for patterns in visualizations, we're doing a series of *visual hypothesis tests*.

Multiple comparisons problem: the more hypothesis tests, the greater the chance of a cpurious finding (since the Null Hypothesis Significant Test admits 5% false positives).



No, because there's not a clear separation between *exploratory* and *confirmatory* analysis.

Is EDA/EVA Fishing? 🎣🐟

Some statisticians have proposed that when we look for patterns in visualizations, we're doing a series of *visual hypothesis tests*.

Multiple comparisons problem: the more hypothesis tests, the greater the chance of a cpurious finding (since the Null Hypothesis Significant Test admits 5% false positives).

No, because there's not a clear separation between *exploratory* and *confirmatory* analysis.

Sort of if you use the same dataset to make hunches *and* then test them (i.e., you cannot collect more data). Try a hold out set (e.g., separate training vs. test data).

Based on slides by Jessica Hullman