Administrivia

Visualization Critique

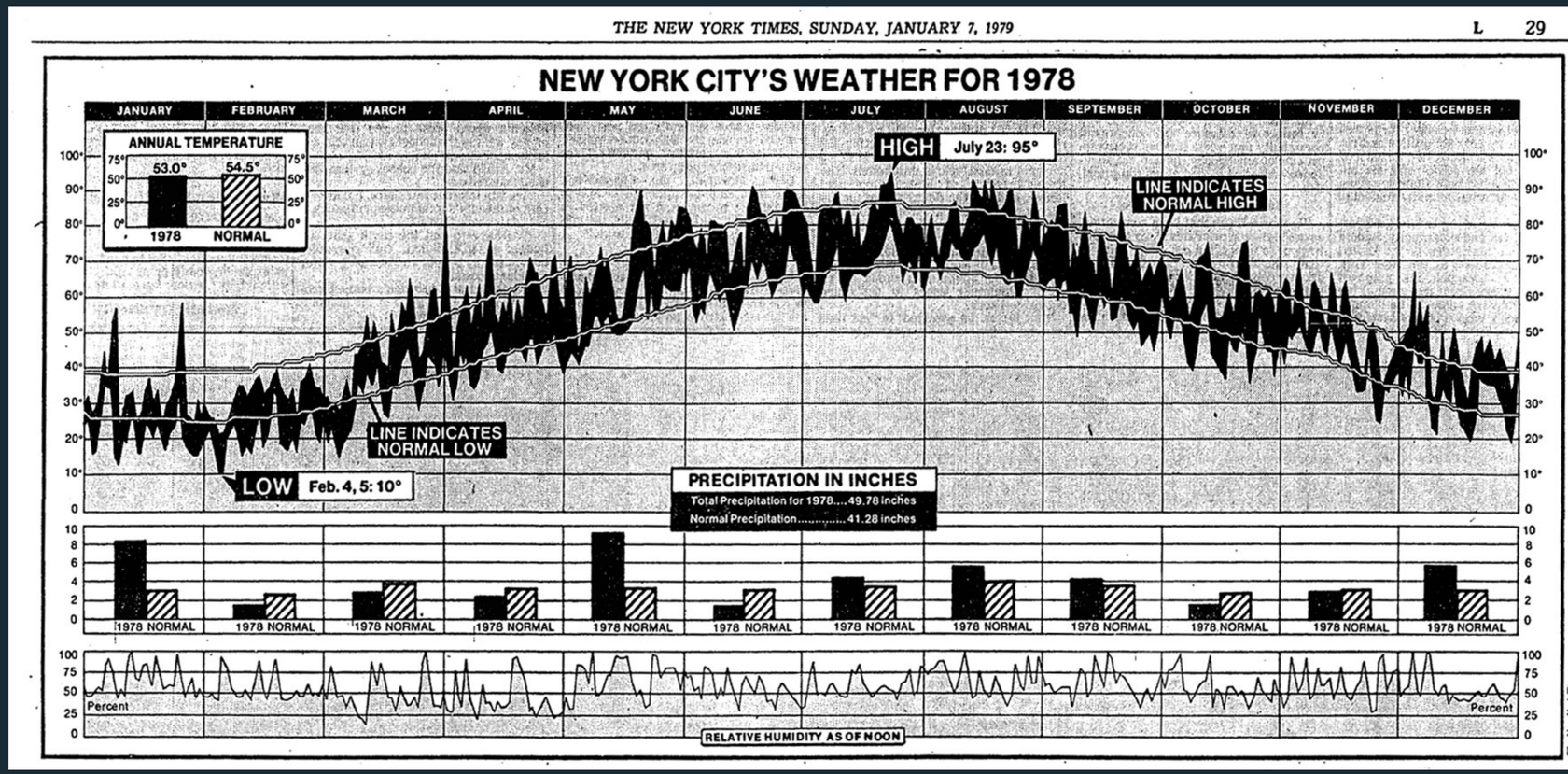
What is this a visualization of?

What visual encodings (mappings from data -> channel) are used?

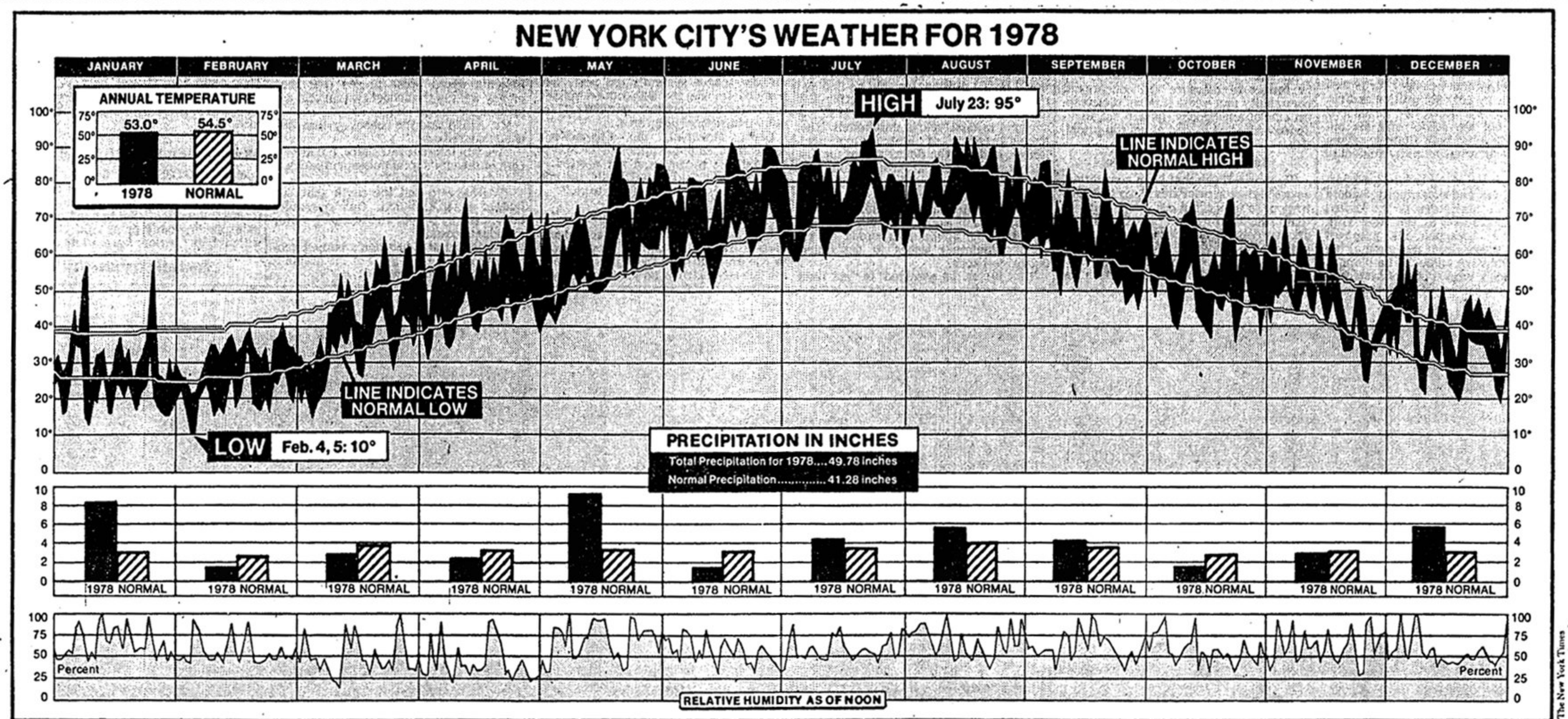
To help structure your critique:

- *"I like..."*
- *"I wish..."*
- *"What if...?"*

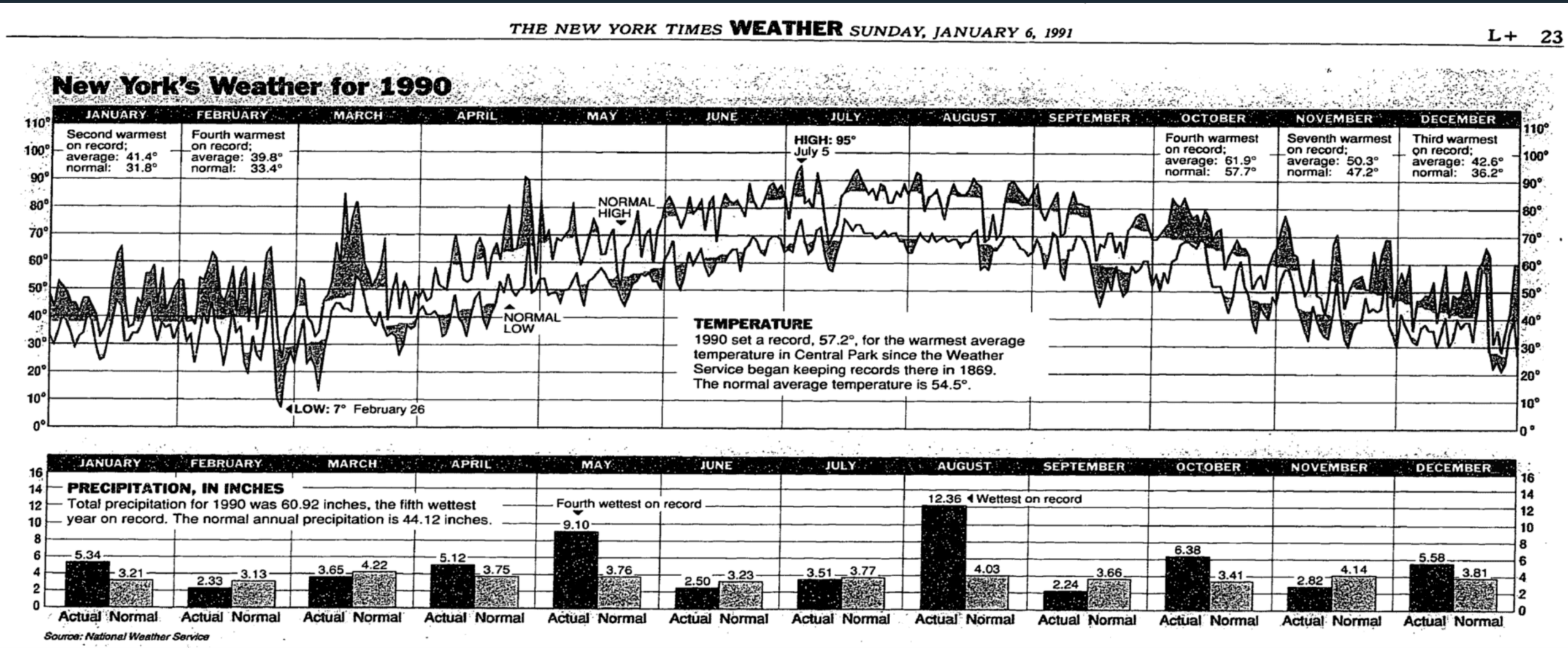
Think (~ 3 mins).
Pair (~7 mins).
Share.



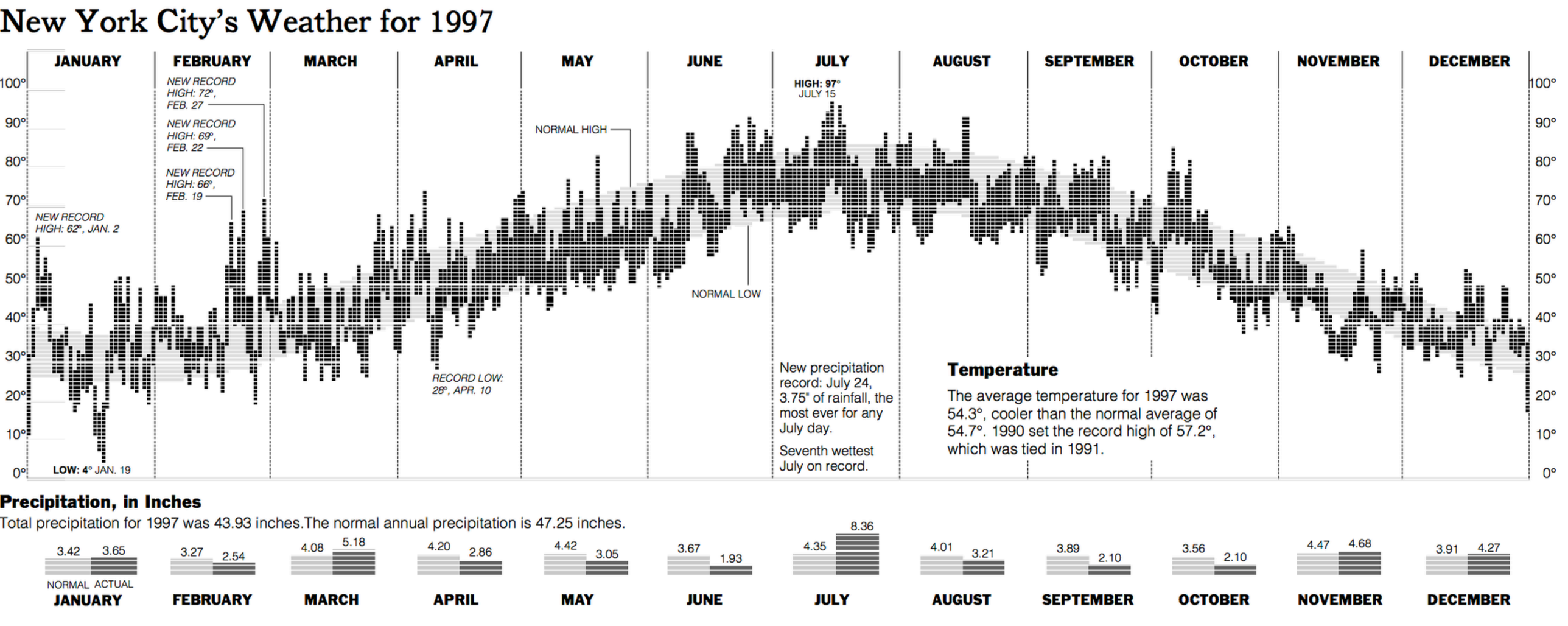
Visualization Critique



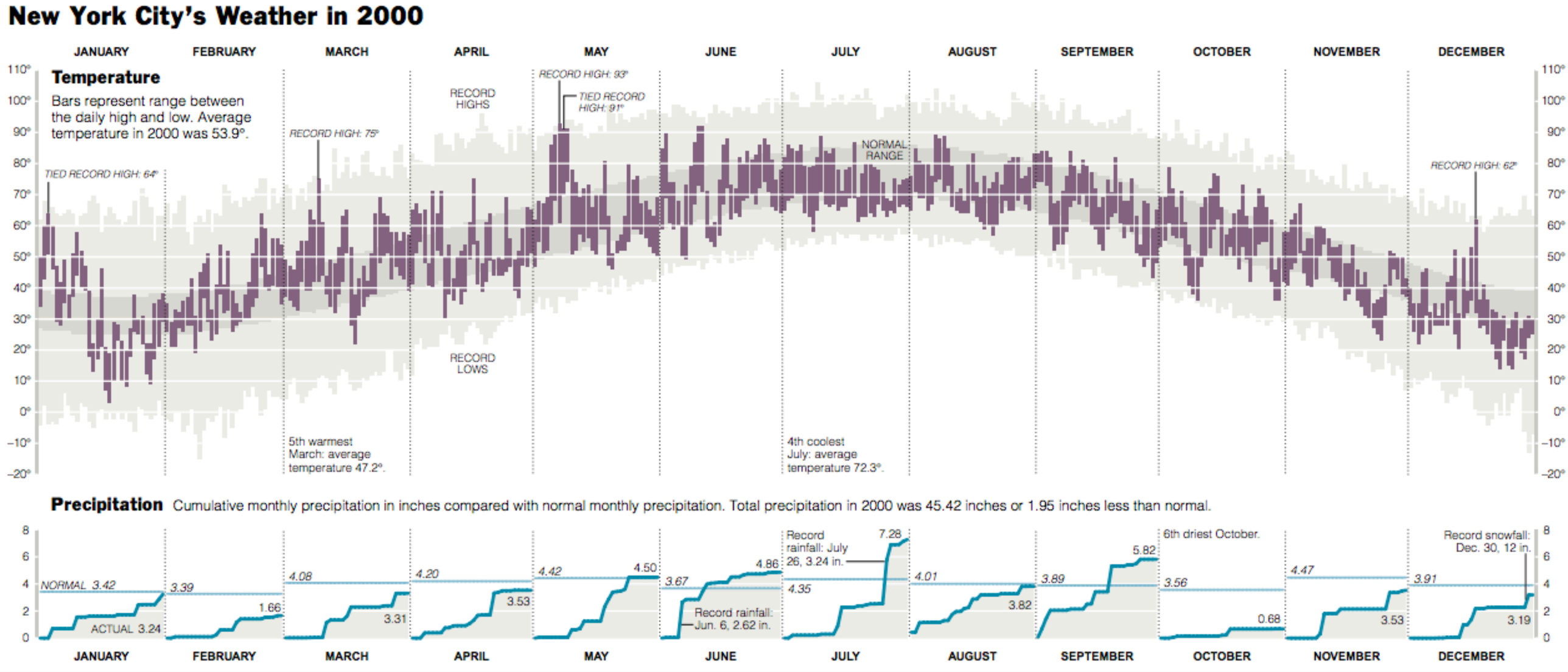
Visualization Critique



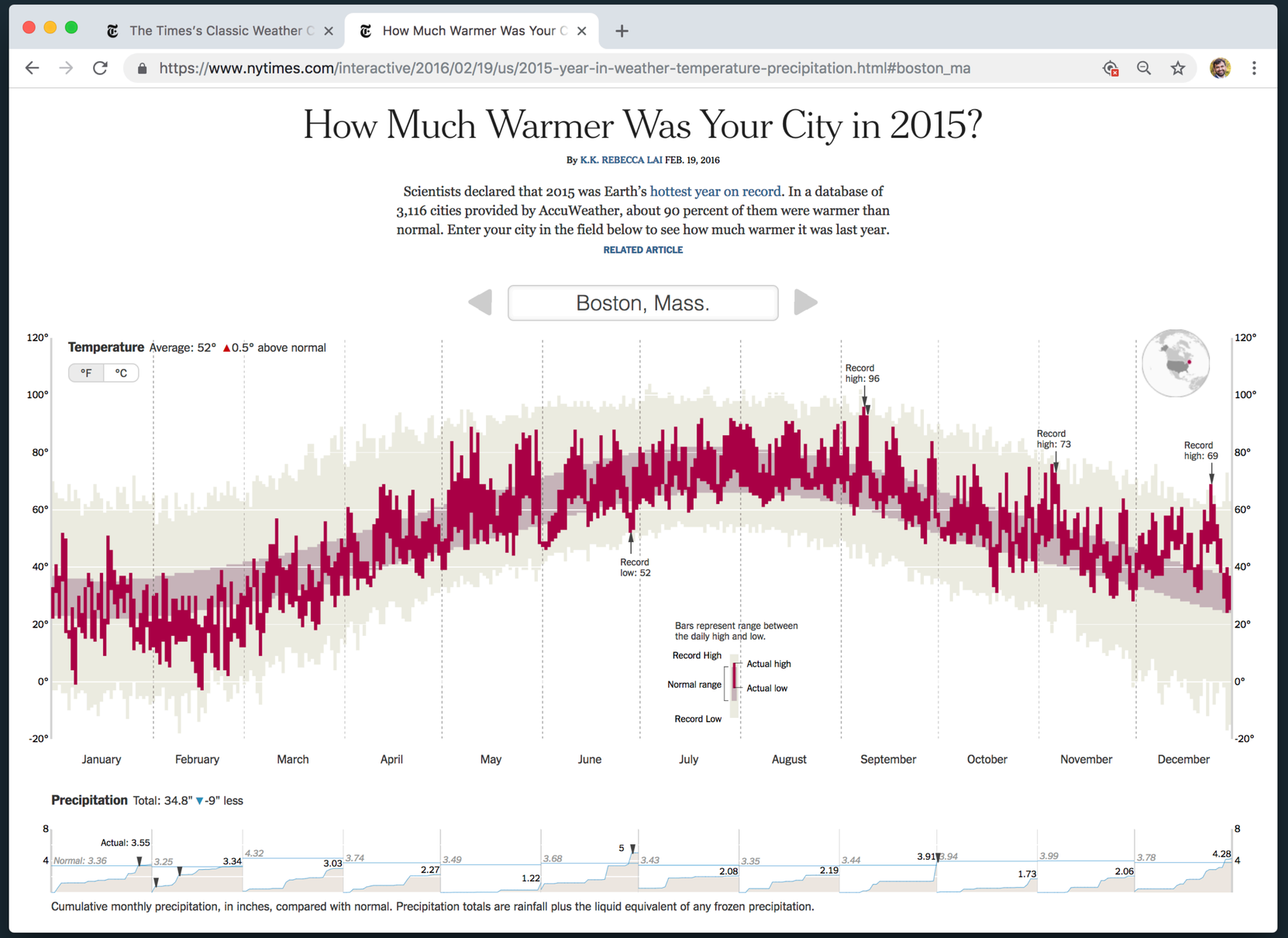
Visualization Critique



Visualization Critique



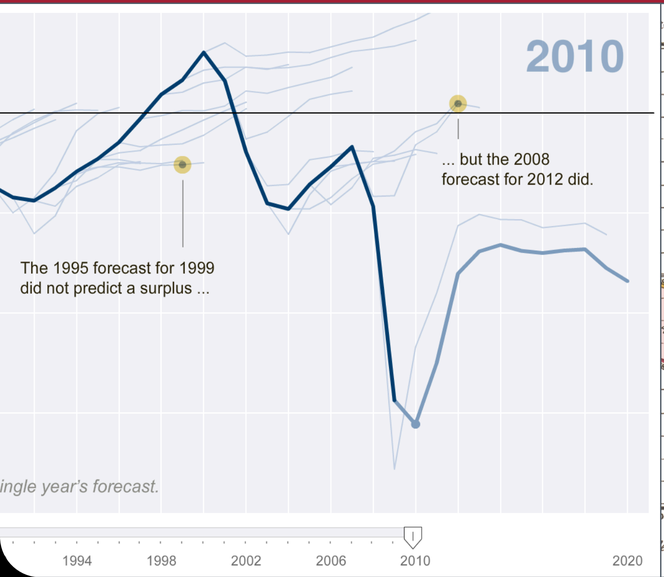
Visualization Critique

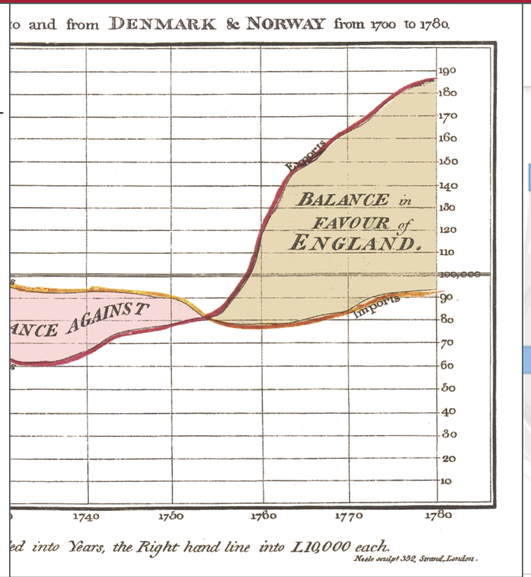


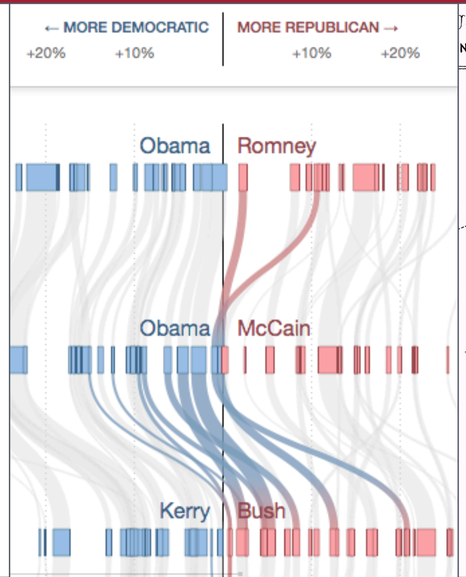
6.894: Interactive Data Visualization

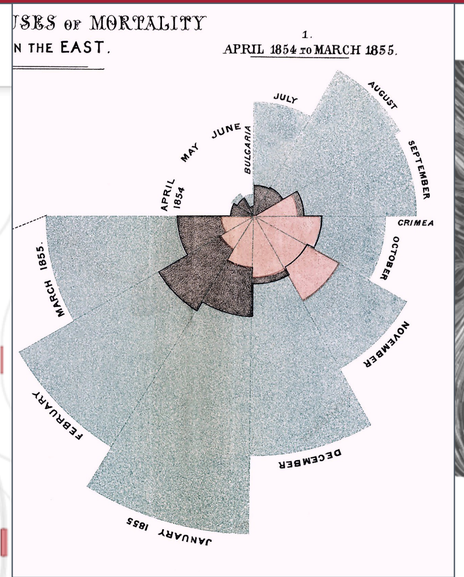
Exploratory Data Analysis

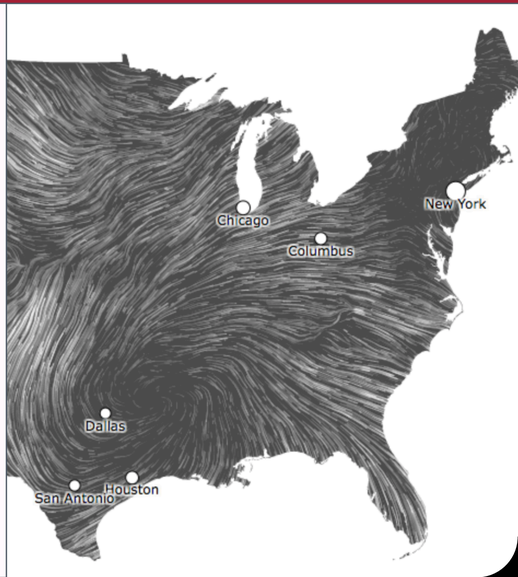
Arvind Satyanarayan











Mapping or Visual Encoding

Data

Visual

Physical Data Types
`int, float, string`

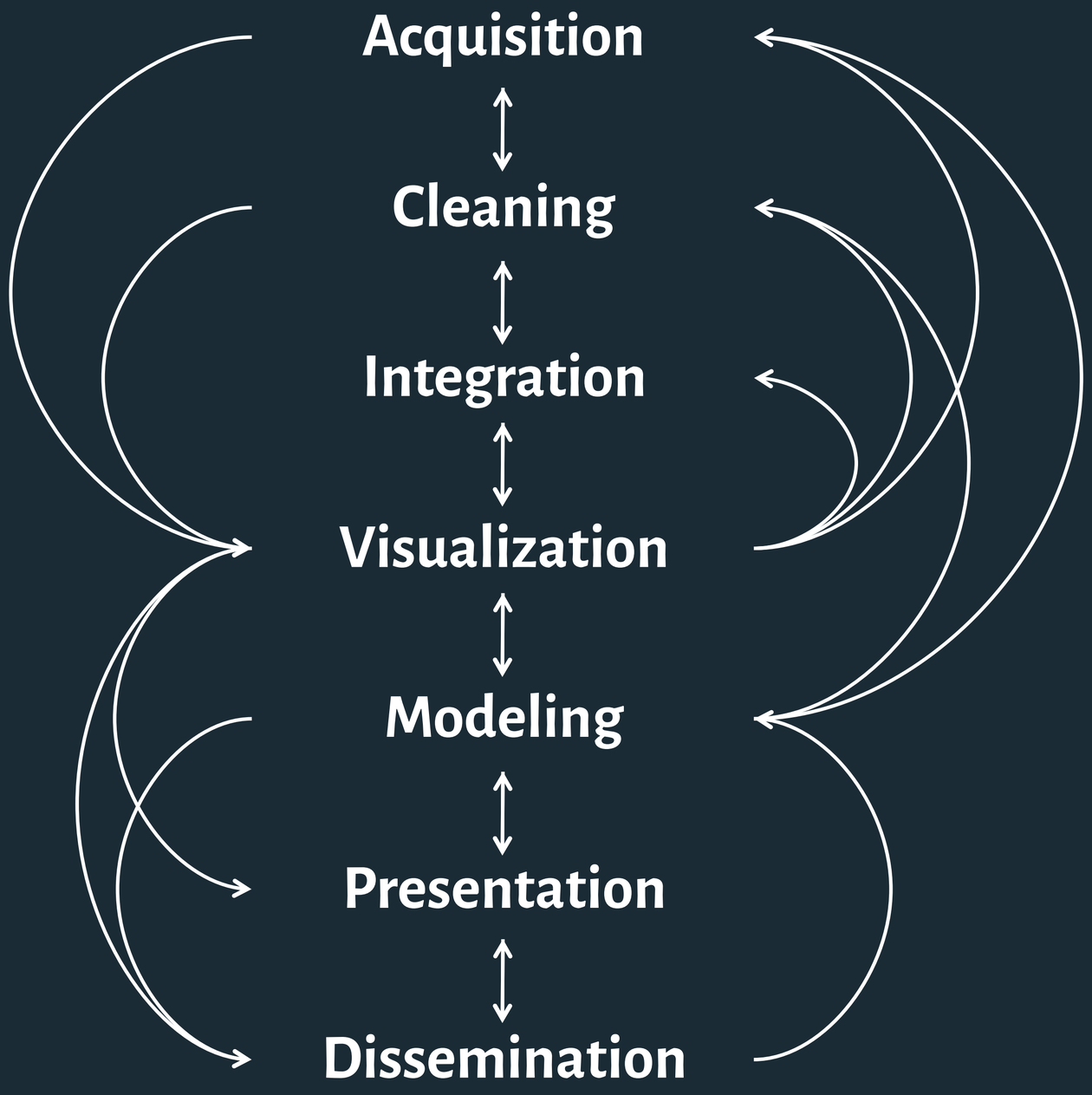
Conceptual Data Types
`temperature, location`

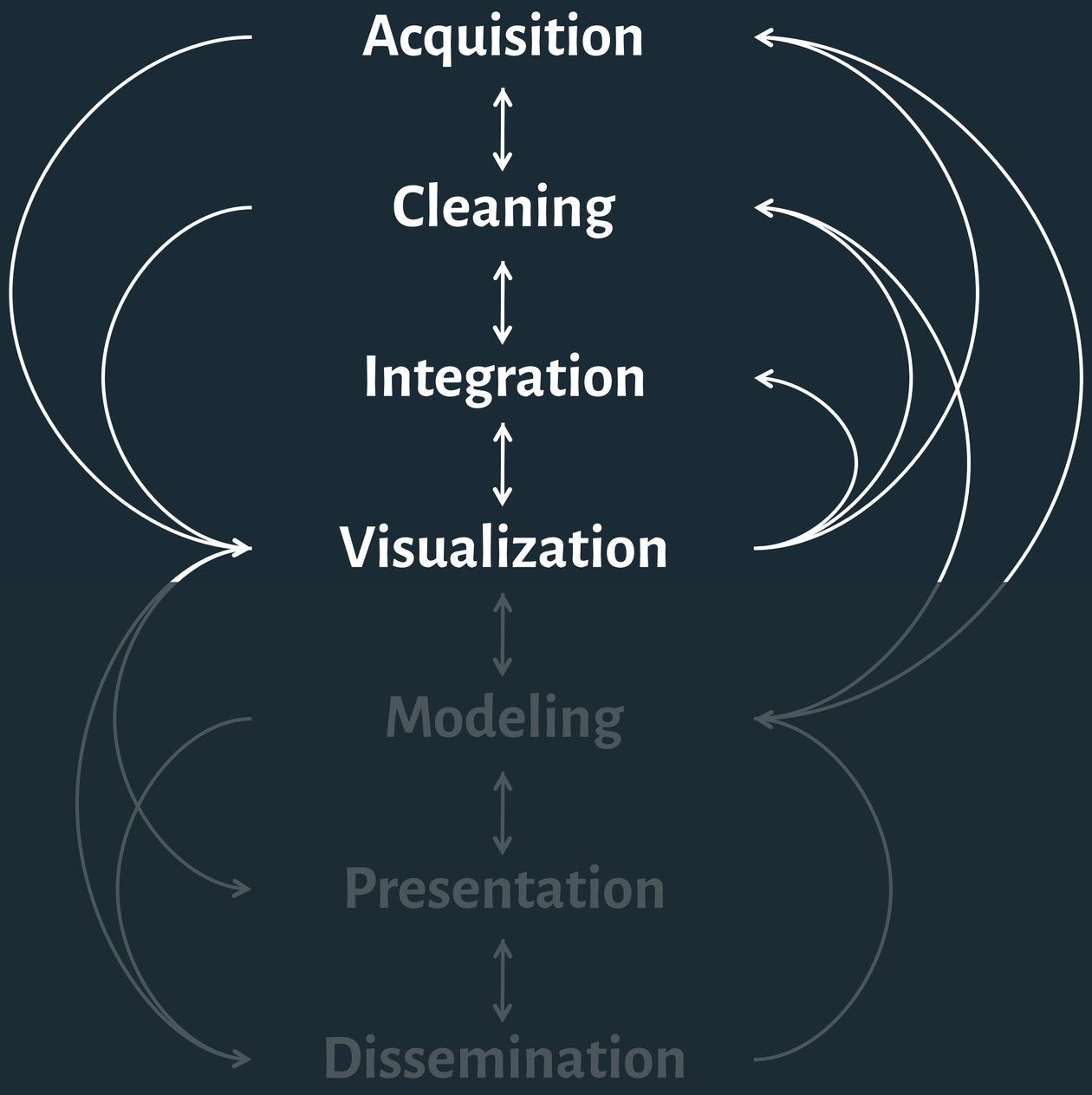
Attribute Types
`nominal, ordinal, quantitative (ratio or interval)`

Graphical **Marks**
`rect, line, point, area`

Visual **Channels**
`x, y, color, opacity`

Data Visualization


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination

Activity: What data quality issues may arise?

Think (~ 2 mins).
Pair (~5 mins).
Share.

Data Skepticism

"The wave of bullshit data is rising, and now it's our turn to figure out how not to get swept away."

—**Jacob Harris**

Co-founder of NYT Interactive Newsroom Tech Team

Nieman Lab, Dec 2014

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?)

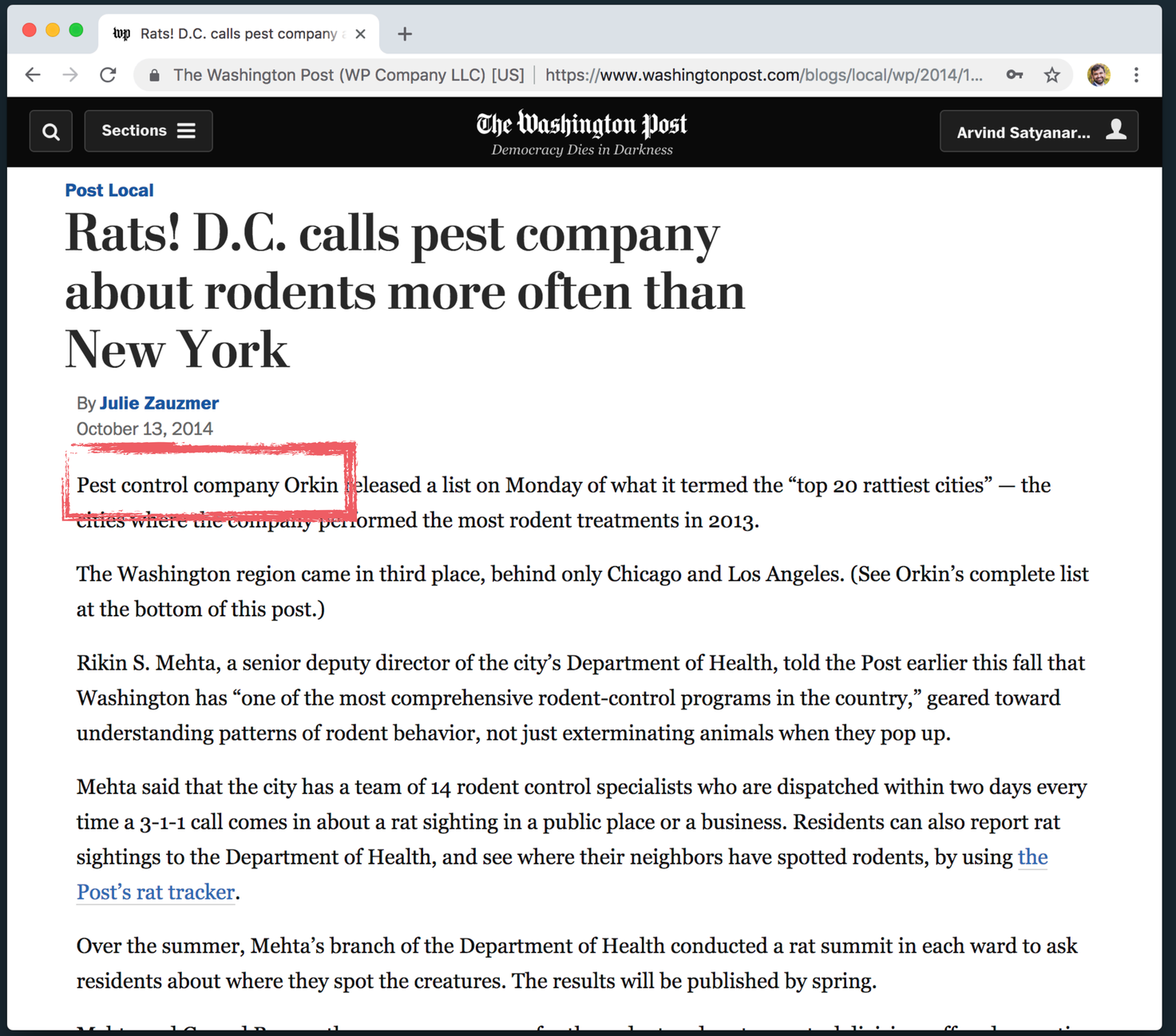
Data Provenance

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?)



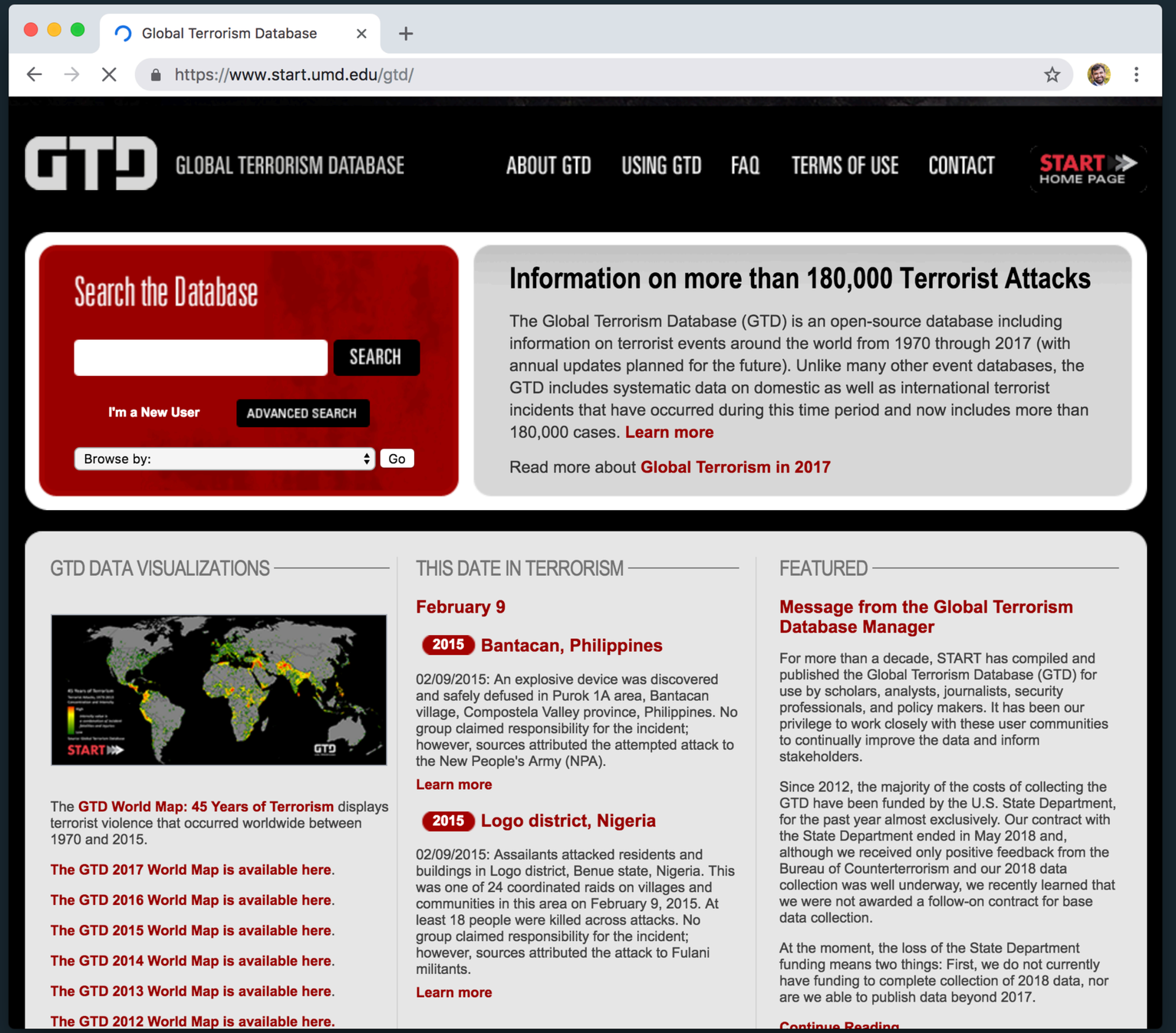
Data Provenance

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.

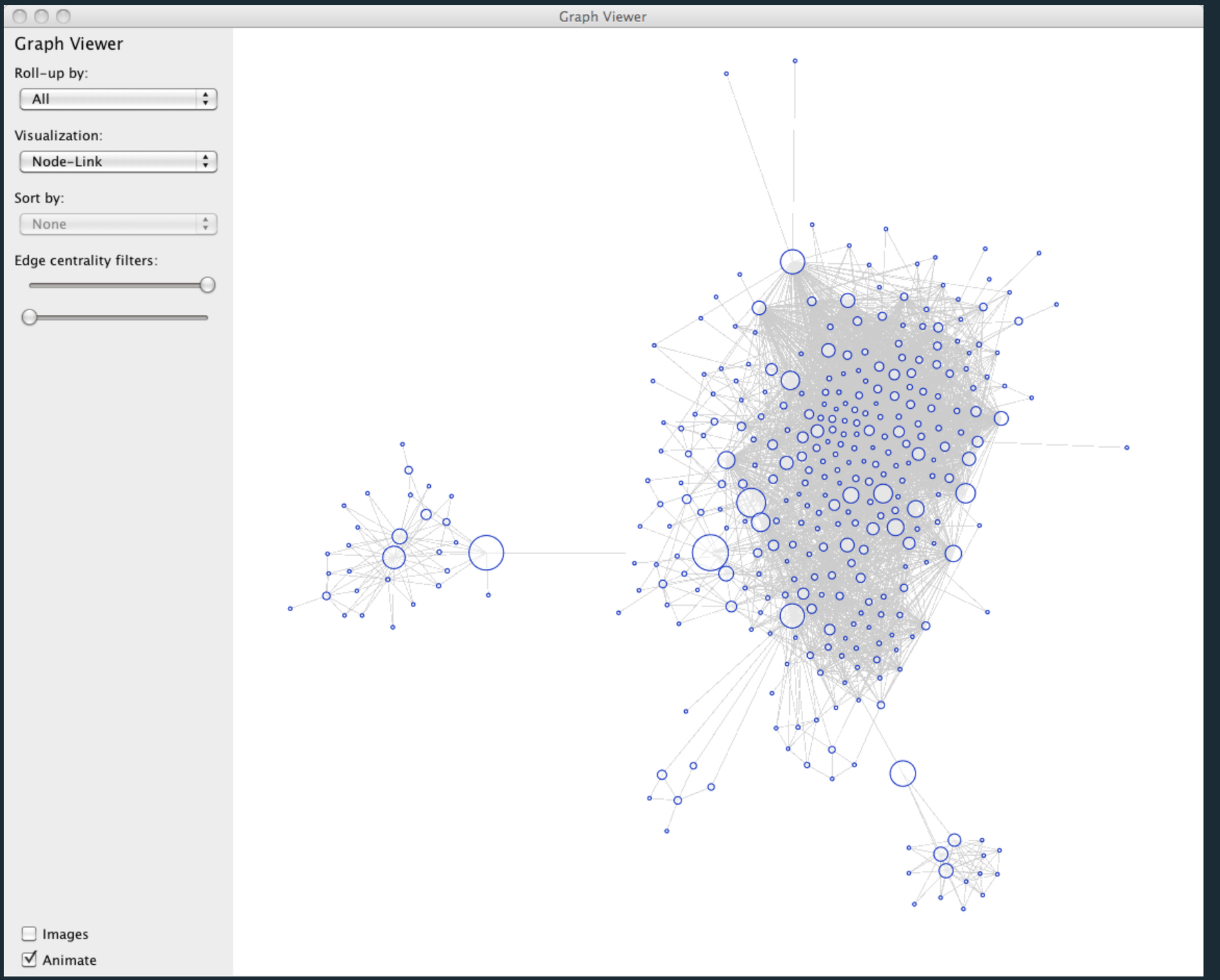
Definitions/interpretations in quantification can change over time.

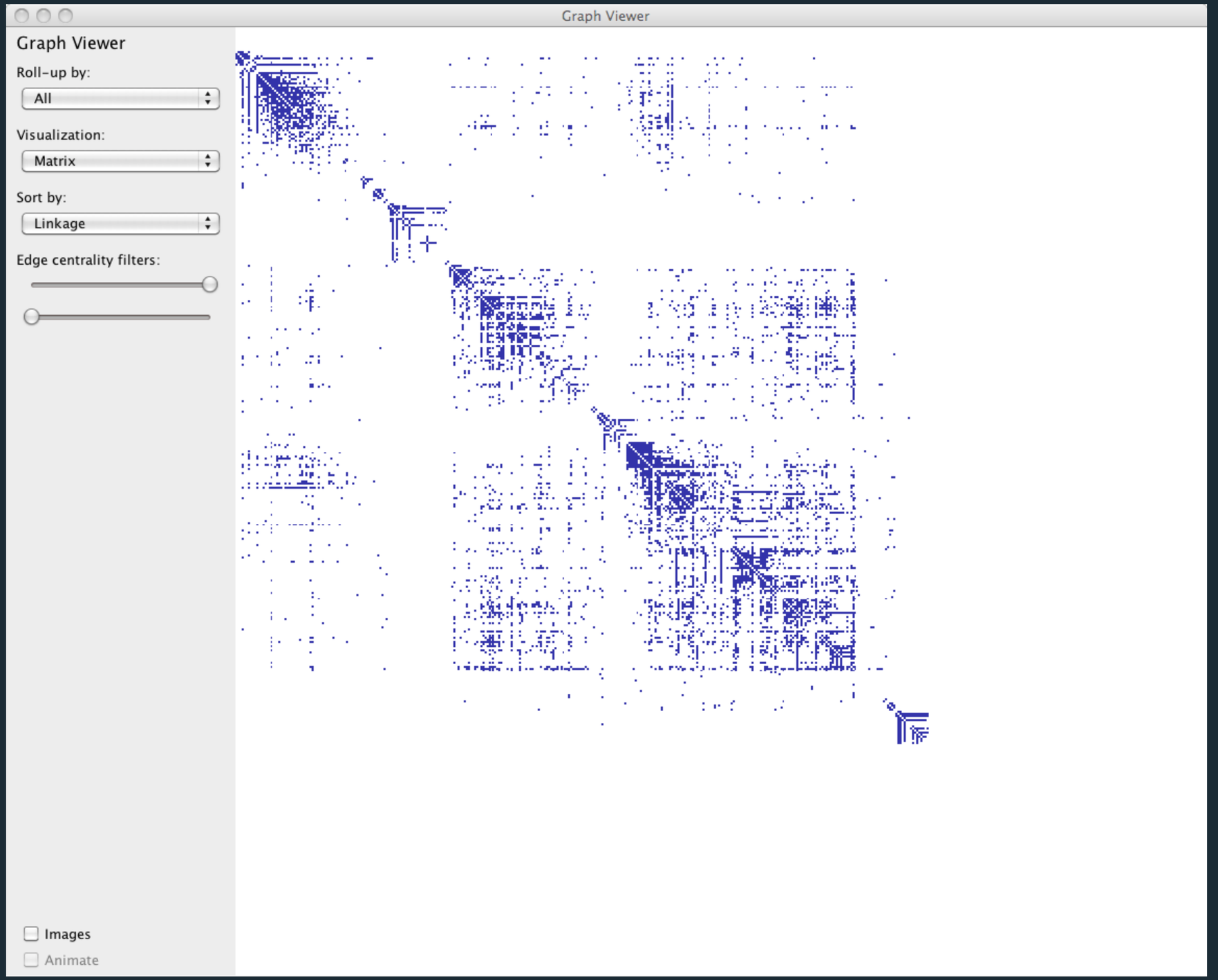
Is the data recent, and how much does that matter to the insight you wish to convey?

The Three-Year Plunge

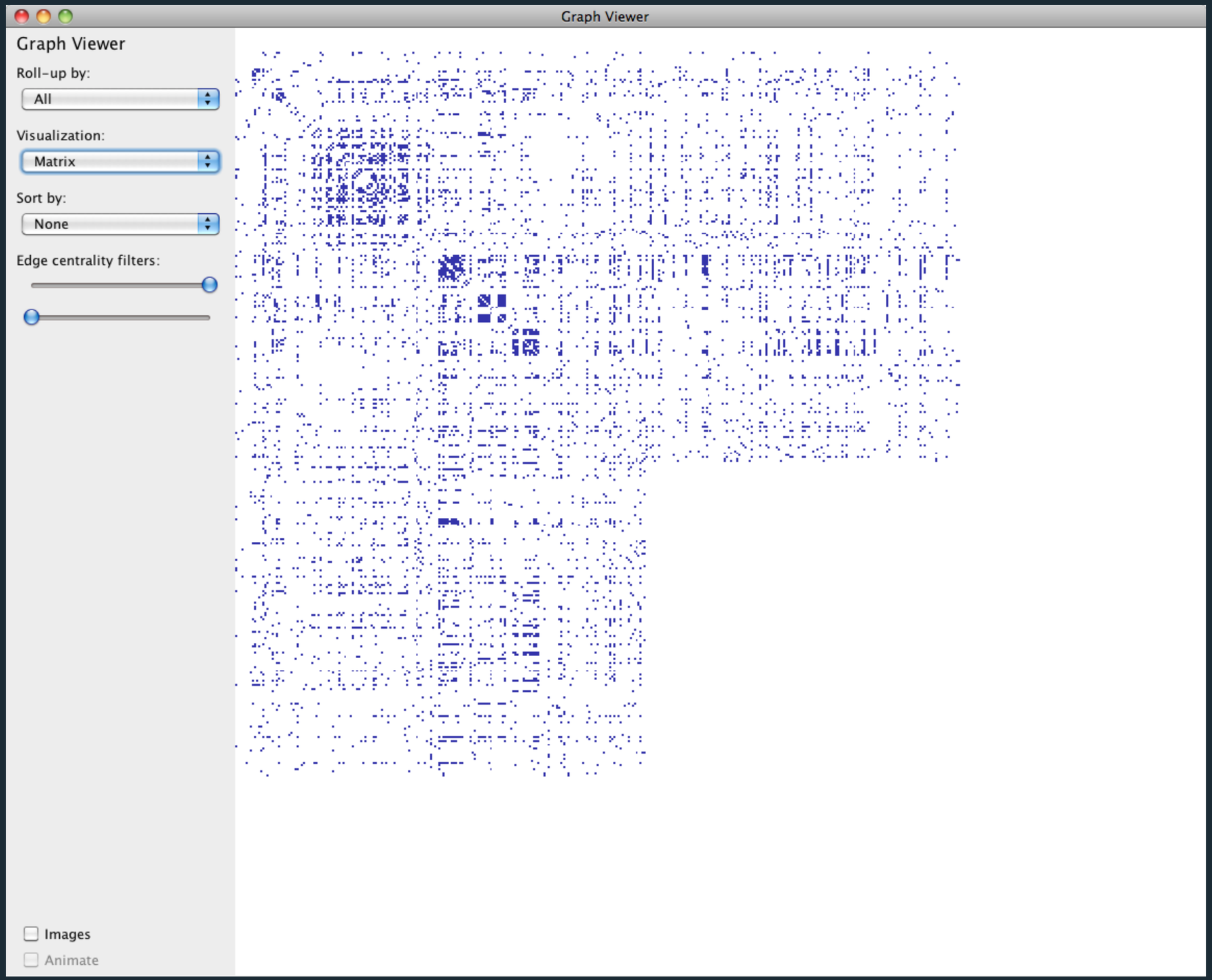
To help gauge each city's overall crime level, the FBI tracks eight "index crimes." From 1993 to 2010, Chicago's annual total dropped by 47 percent. But from 2010 to 2013, it dropped a stunning 56 percent, or nearly 19 percent per year, according to data from the Chicago Police Department.

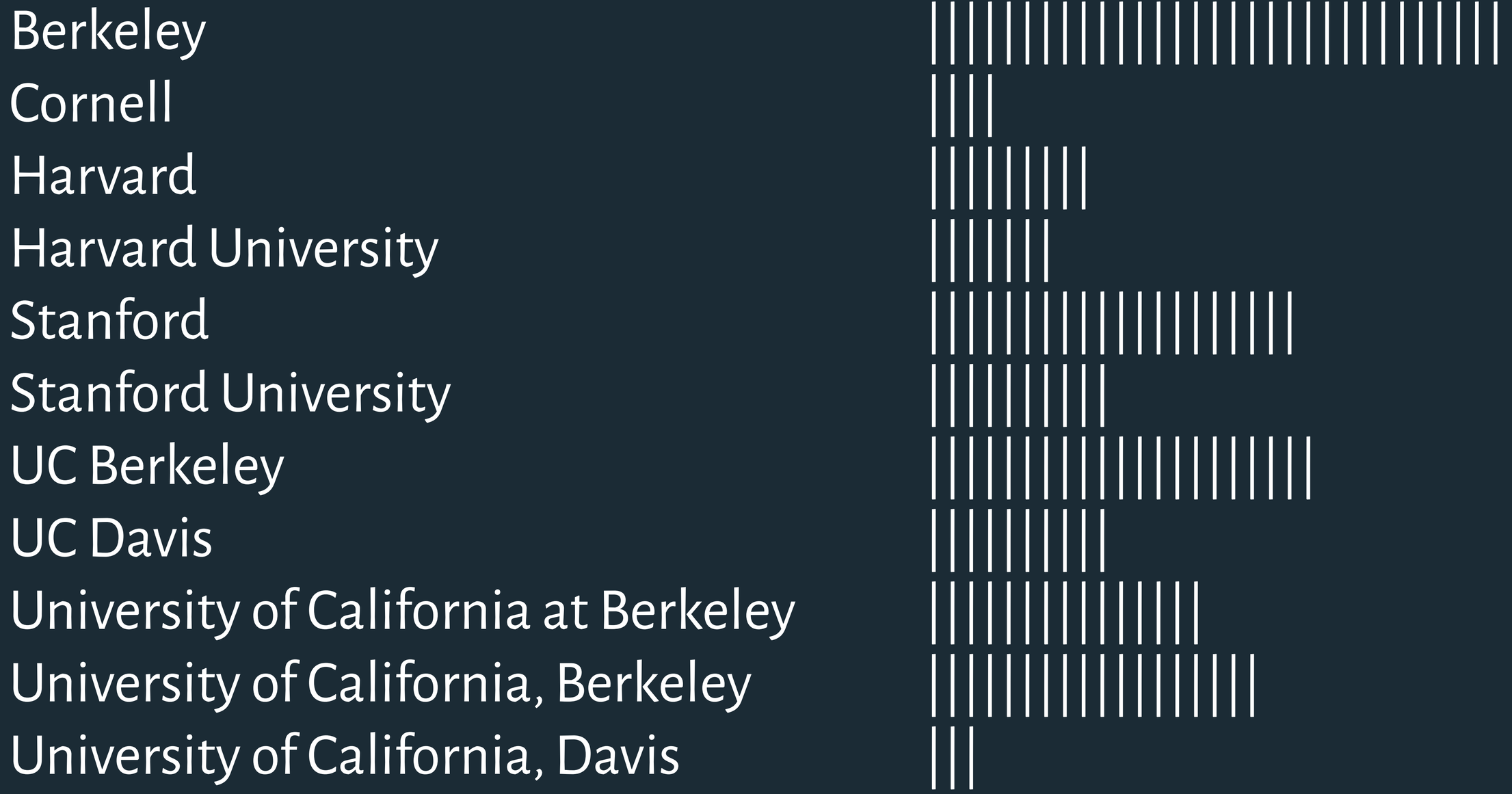
David Berstein & Noah Isackson. *Chicago Magazine*, April 2014.


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



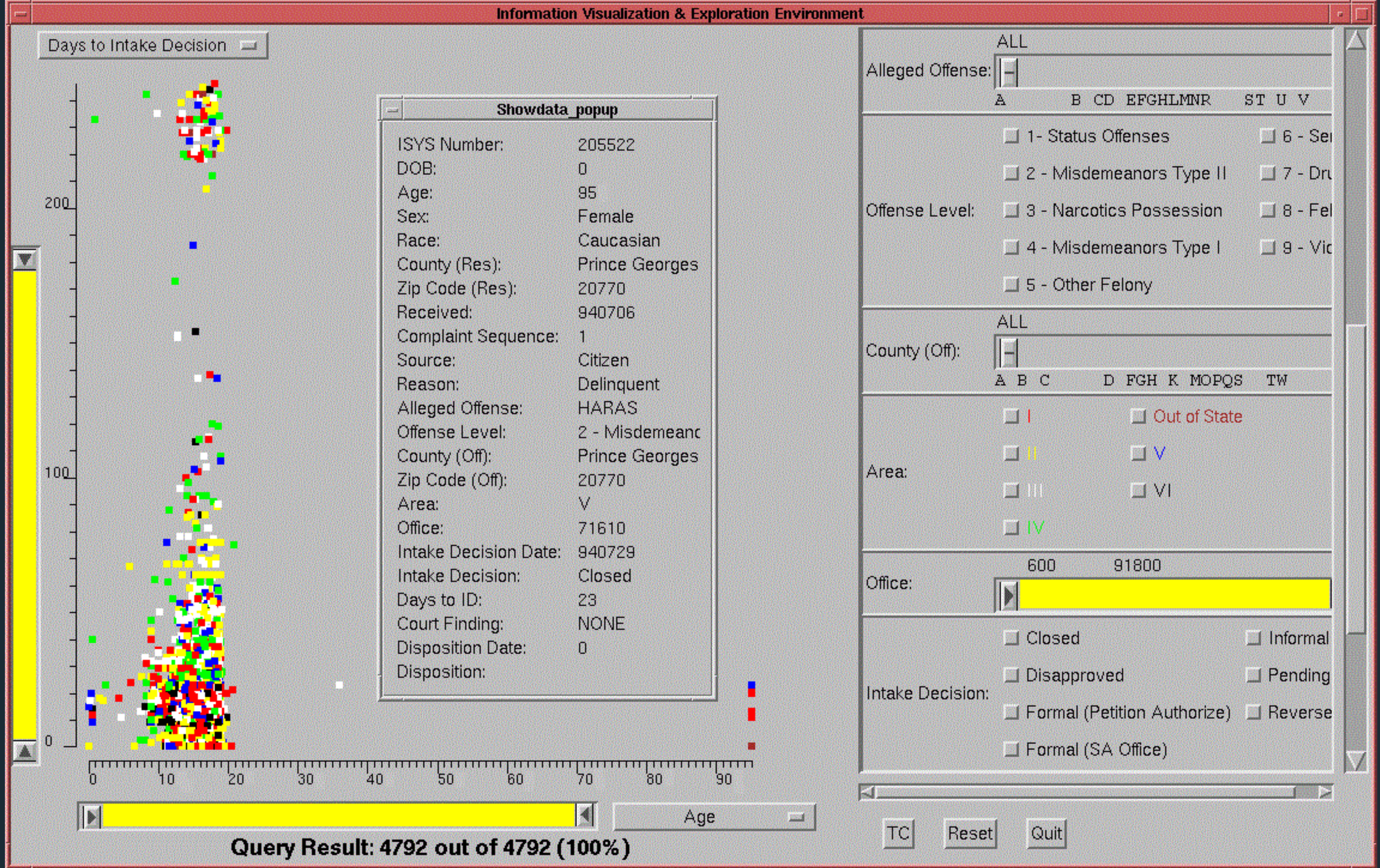

Berkeley
Cornell
Harvard
Harvard University
Stanford
Stanford University
UC Berkeley
UC Davis
University of California at Berkeley
University of California, Berkeley
University of California, Davis

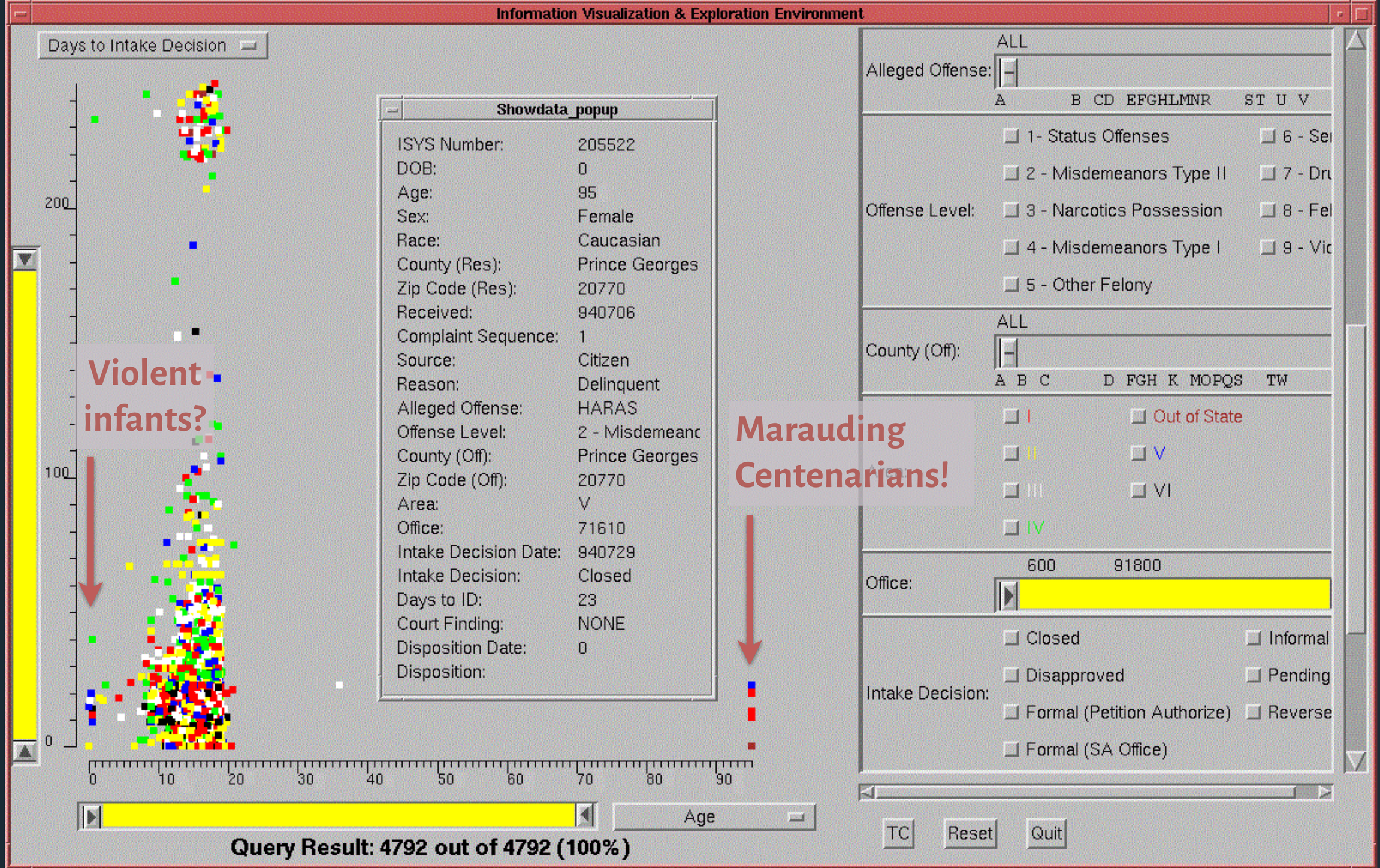
“The first sign that a visualization is good is that **it shows you a problem in your data**. Every successful visualization that I've been involved with has had this stage where you realize, ***"Oh my God, this data is not what I thought it would be!"*** So already, you've discovered something.”

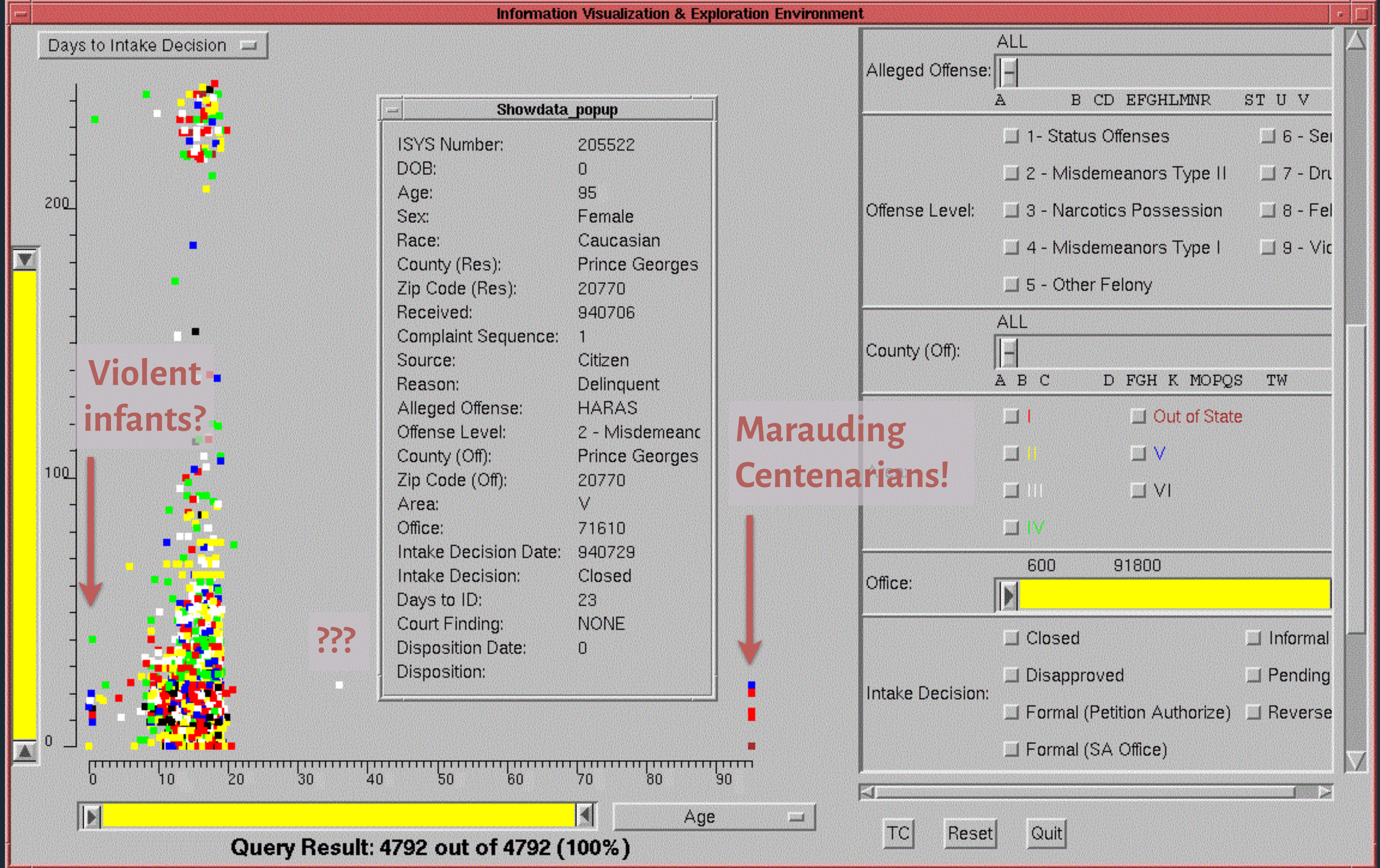
— **Martin Wattenberg**

Co-lead of Google's People + AI Initiative

ACM Queue, Mar 2010


Information Visualization & Exploration Environment
Days to Intake Decision
Showdata_popup
ISYS Number: 205522
DOB: 0
Age: 95
Sex: Female
Race: Caucasian
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanc
County (Off): Prince Georges
Zip Code (Off): 20770
Area: V
Office: 71610
Intake Decision Date: 940729
Intake Decision: Closed
Days to ID: 23
Court Finding: NONE
Disposition Date: 0
Disposition:
0
100
200
0 10 20 30 40 50 60 70 80 90
Age
Query Result: 4792 out of 4792 (100%)
ALL
Alleged Offense:
A B CD EFGHLMNR ST U V
Offense Level:
1- Status Offenses
2 - Misdemeanors Type II
3 - Narcotics Possession
4 - Misdemeanors Type I
5 - Other Felony
6 - Se
7 - Dru
8 - Fel
9 - Vic
ALL
County (Off):
A B C D FGH K MOPQS TW
Area:
I
II
III
IV
Out of State
V
VI
600 91800
Office:
Intake Decision:
Closed
Disapproved
Formal (Petition Authorize)
Formal (SA Office)
Informal
Pending
Reverse
TC
Reset
Quit


Information Visualization & Exploration Environment
Days to Intake Decision
Showdata_popup
ISYS Number: 205522
DOB: 0
Age: 95
Sex: Female
Race: Caucasian
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanc
County (Off): Prince Georges
Zip Code (Off): 20770
Area: V
Office: 71610
Intake Decision Date: 940729
Intake Decision: Closed
Days to ID: 23
Court Finding: NONE
Disposition Date: 0
Disposition:
Violent infants?
Marauding Centenarians!
Alleged Offense:
ALL
A B CD EFGHLMNR ST U V
Offense Level:
1- Status Offenses
2 - Misdemeanors Type II
3 - Narcotics Possession
4 - Misdemeanors Type I
5 - Other Felony
6 - Se
7 - Dru
8 - Fel
9 - Vic
County (Off):
ALL
A B C D FGH K MOPQS TW
I
II
III
IV
Out of State
V
VI
Office:
600 91800
Intake Decision:
Closed
Disapproved
Formal (Petition Authorize)
Formal (SA Office)
Informal
Pending
Reverse
0 10 20 30 40 50 60 70 80 90
Age
TC
Reset
Quit
Query Result: 4792 out of 4792 (100%)


Information Visualization & Exploration Environment
Days to Intake Decision
200
100
0
Violent infants?
???
Marauding Centenarians!
Showdata_popup
ISYS Number: 205522
DOB: 0
Age: 95
Sex: Female
Race: Caucasian
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanc
County (Off): Prince Georges
Zip Code (Off): 20770
Area: V
Office: 71610
Intake Decision Date: 940729
Intake Decision: Closed
Days to ID: 23
Court Finding: NONE
Disposition Date: 0
Disposition:
0 10 20 30 40 50 60 70 80 90
Age
Query Result: 4792 out of 4792 (100%)
Alleged Offense: ALL
A B CD EFGHLMNR ST U V
Offense Level:
1- Status Offenses
2 - Misdemeanors Type II
3 - Narcotics Possession
4 - Misdemeanors Type I
5 - Other Felony
6 - Se
7 - Dru
8 - Fel
9 - Vic
County (Off): ALL
A B C D FGH K MOPQS TW
I
II
III
IV
Out of State
V
VI
Office: 600 91800
Intake Decision:
Closed
Disapproved
Formal (Petition Authorize)
Formal (SA Office)
Informal
Pending
Reverse
TC
Reset
Quit

“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]

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

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

Suggestions

Delete **rows 8,10**

Delete **empty rows**

Delete **rows where Property_crime_rate is null**

Delete **rows where Year is null**

Script Export

Split **data repeatedly** on **newline** into **rows**

Split **data repeatedly** on **','**

rows: 408 prev next

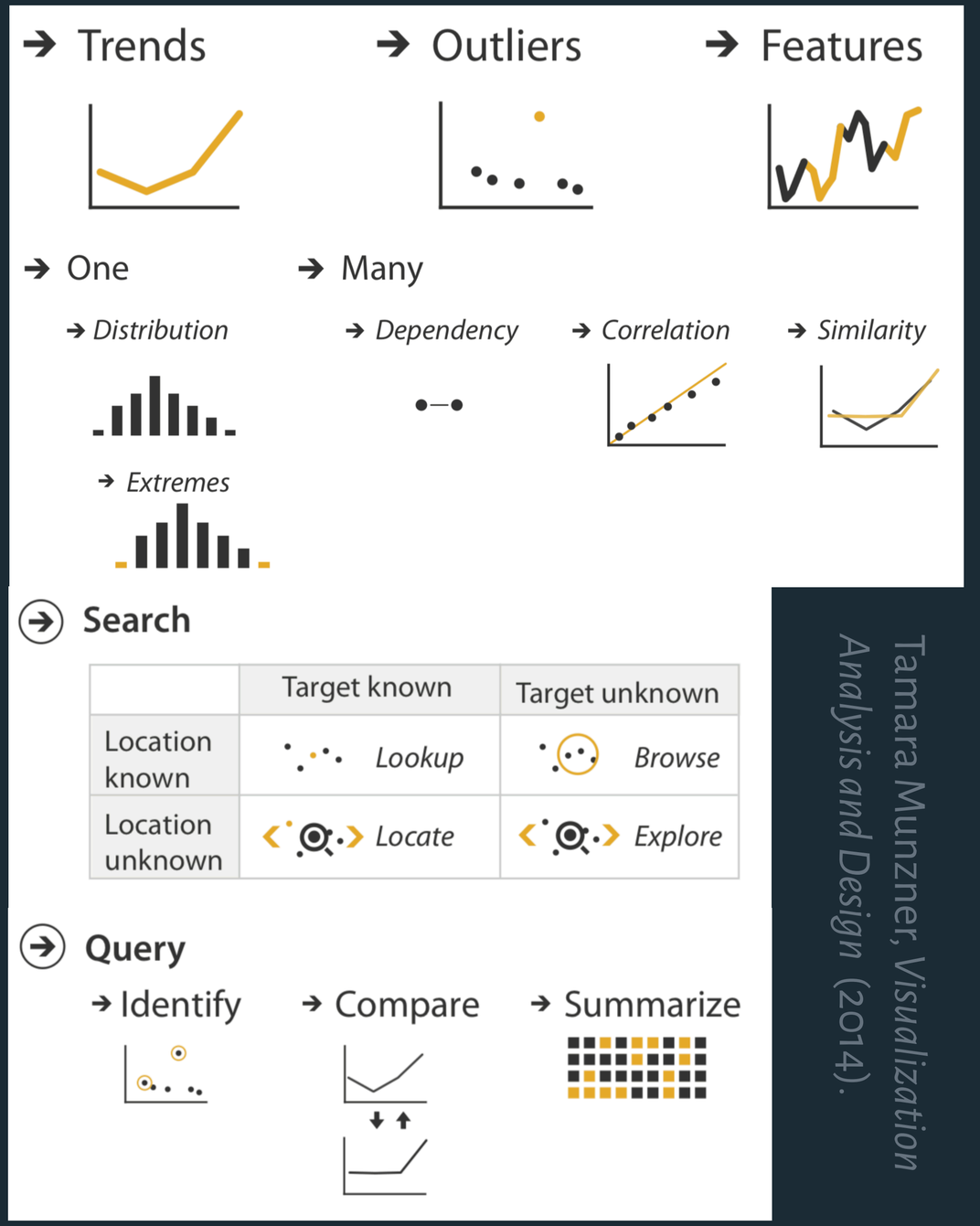
	Year	Property_crime_rate
1	Reported crime in Alabama	
2		
3	2004	4029.3
4	2005	3900
5	2006	3937
6	2007	3974.9
7	2008	4081.9
8		
9	Reported crime in Alaska	
10		
11	2004	3370.9
12	2005	3615
13	2006	3582
14	2007	3373.9

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).

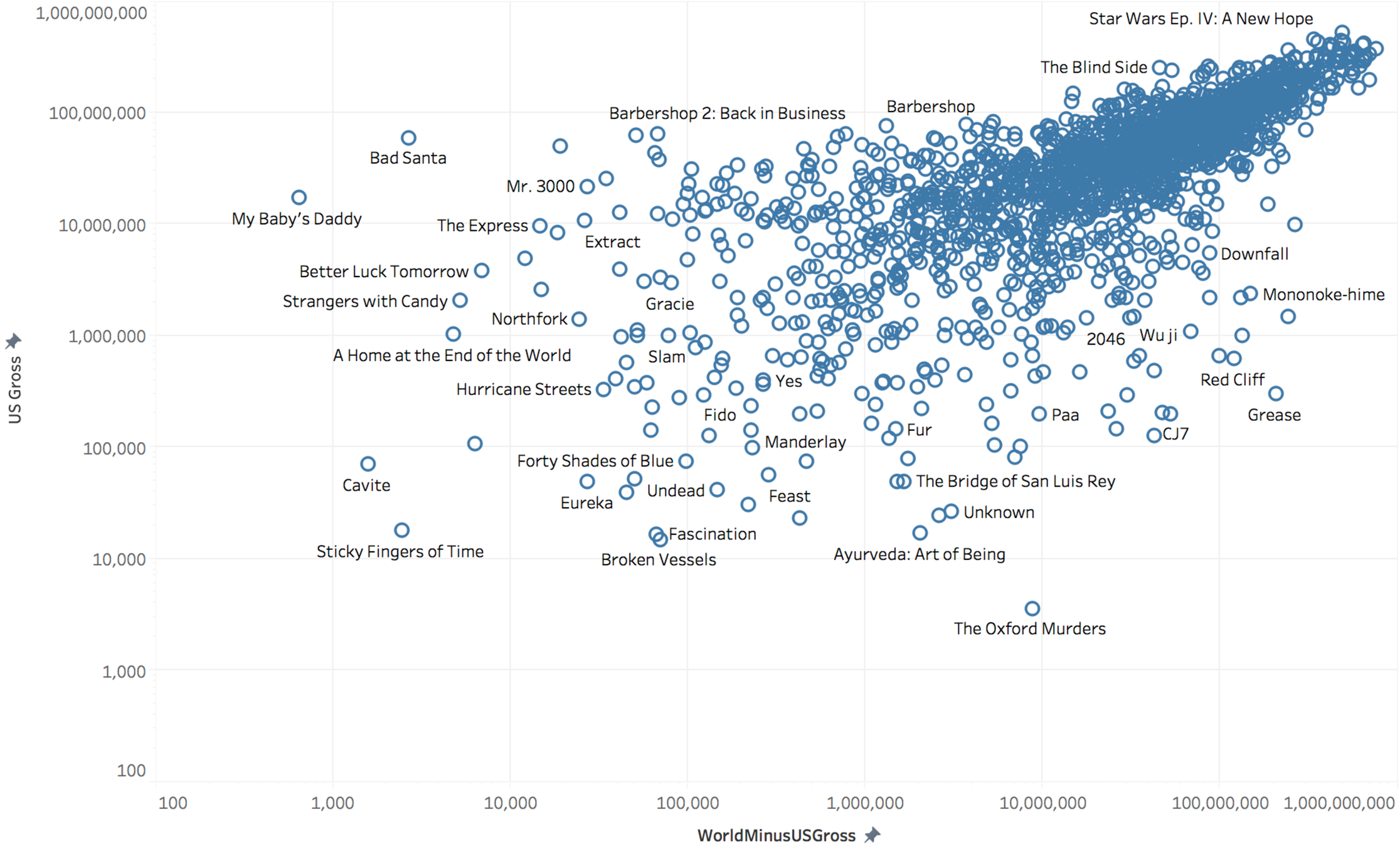
Analysis Example:
Motion Pictures Data

Analysis Example: Motion Pictures Data

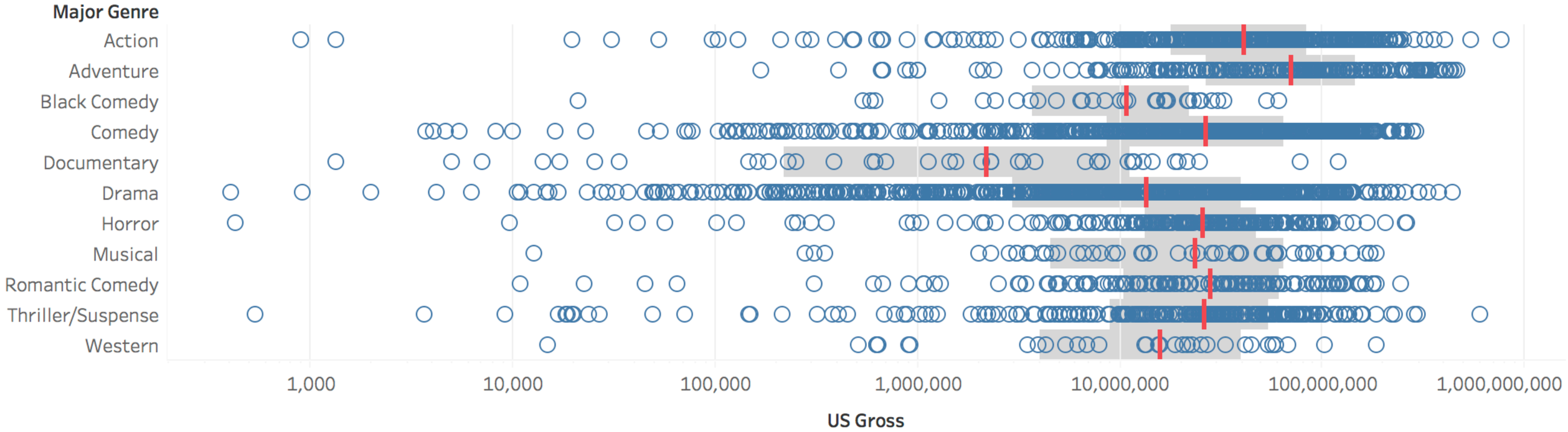
A sample of 3,201 movies collected in 2010.

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)

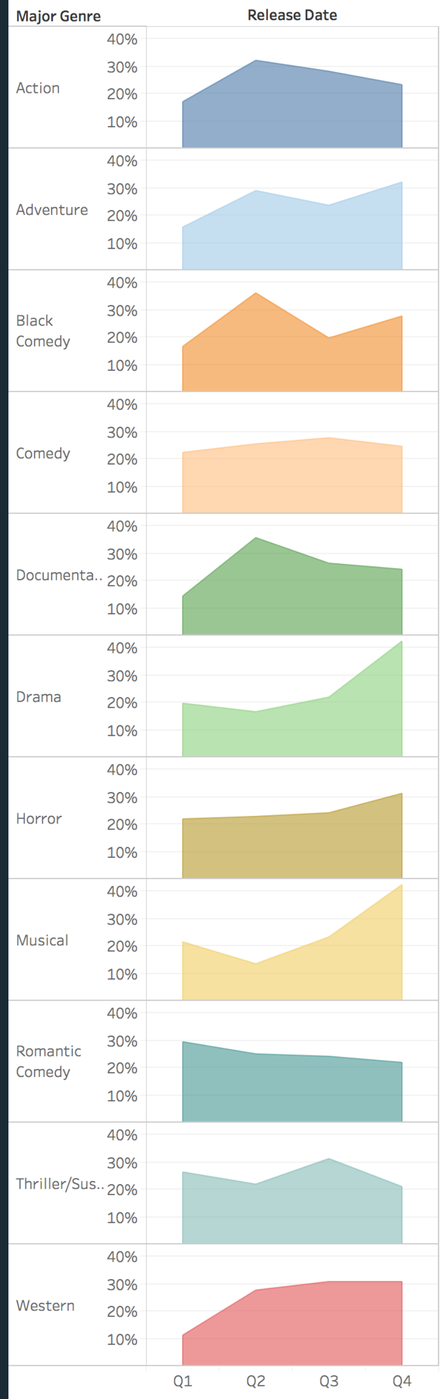
US vs WW

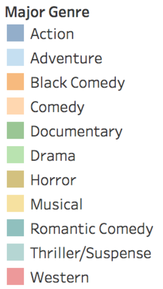

1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
US Gross
100
1,000
10,000
100,000
1,000,000
10,000,000
100,000,000
1,000,000,000
WorldMinusUSGross
Star Wars Ep. IV: A New Hope
The Blind Side
Barbershop
Barbershop 2: Back in Business
Bad Santa
Mr. 3000
My Baby's Daddy
The Express
Extract
Better Luck Tomorrow
Strangers with Candy
Gracie
Northfork
A Home at the End of the World
Slam
Hurricane Streets
Yes
Fido
Fur
Manderlay
Forty Shades of Blue
Cavite
Eureka
Undead
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
Paa
CJ7

Distribution of US Gross by Genre

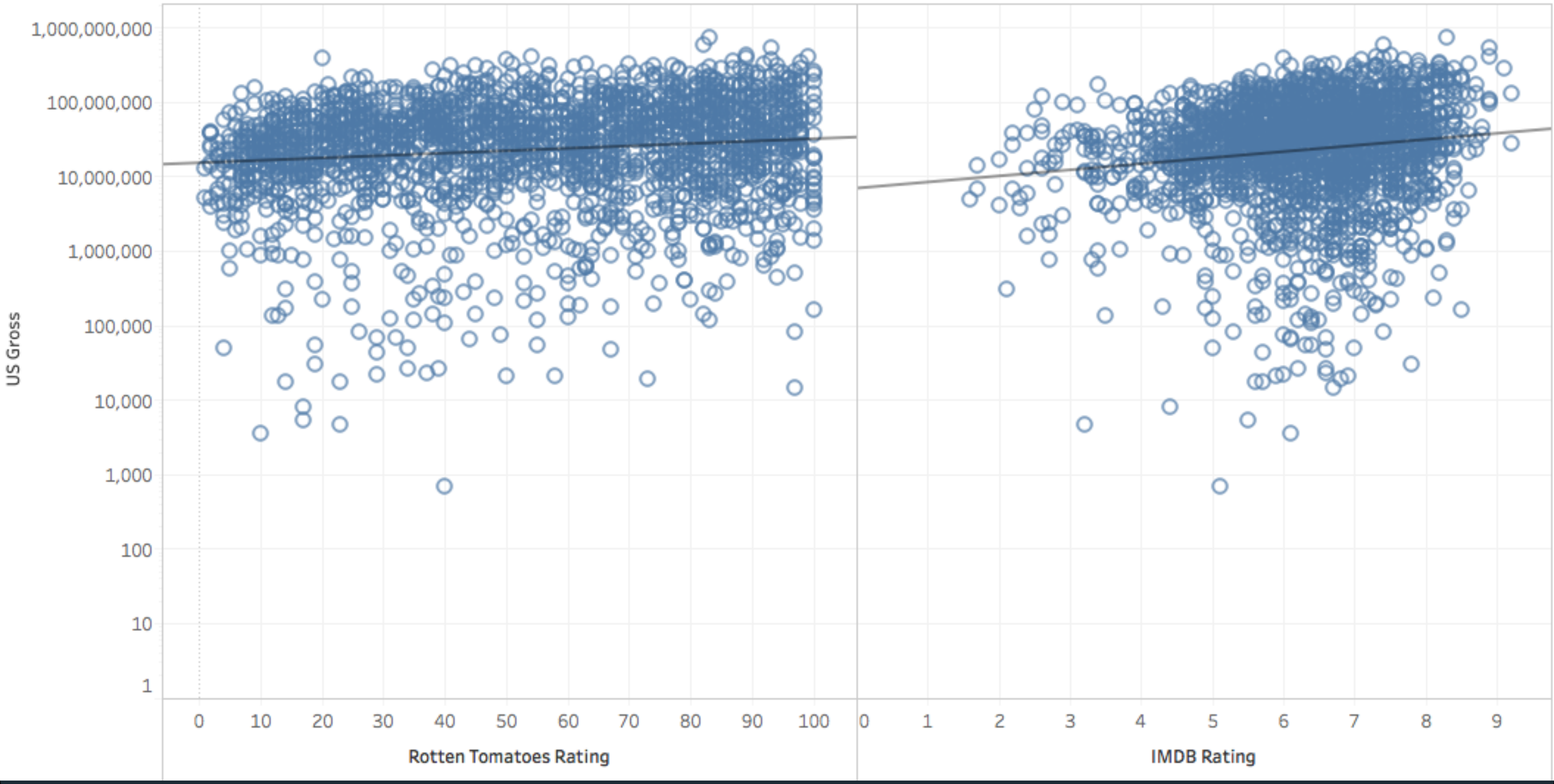


% Releases per Quarter per Genre

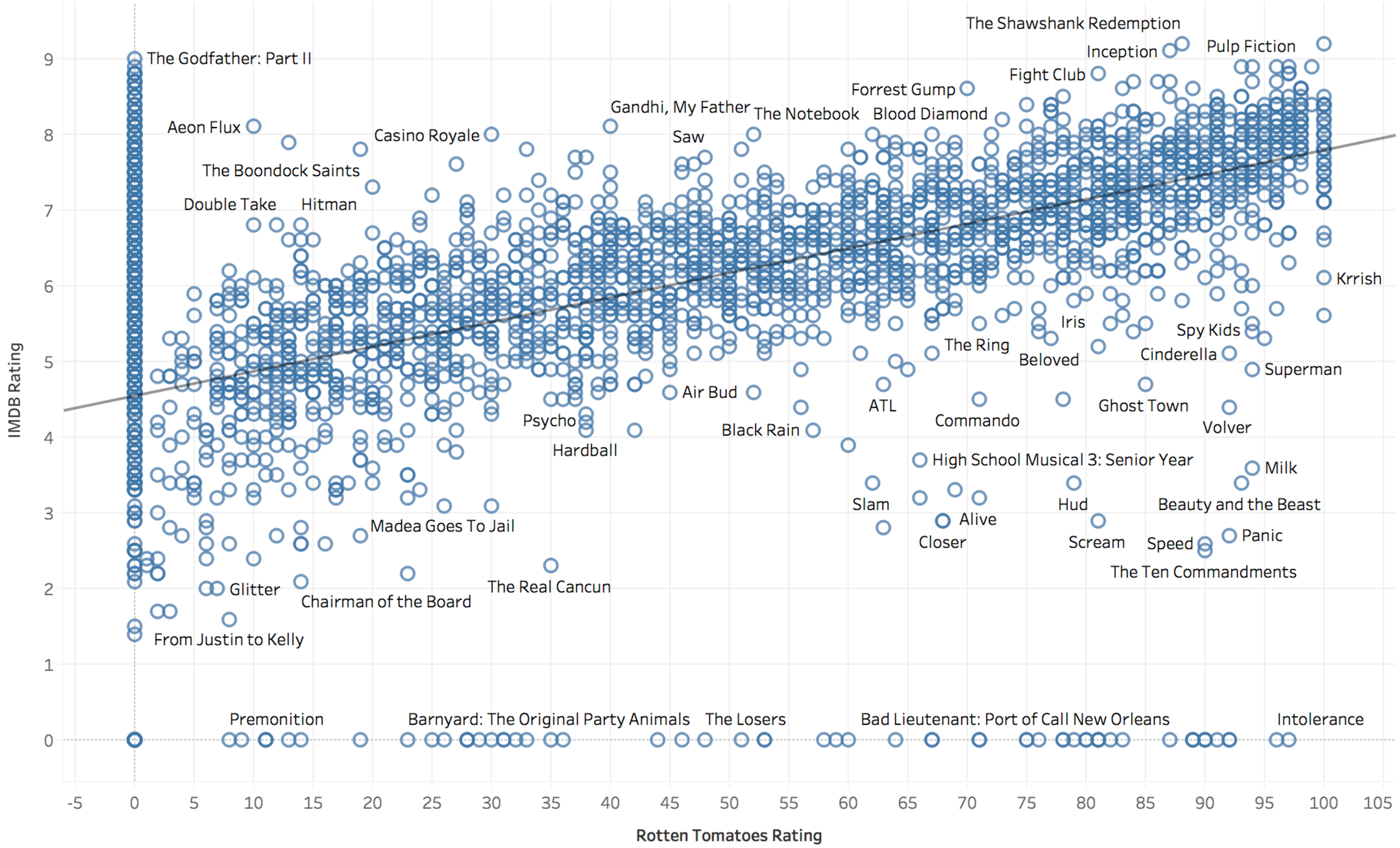


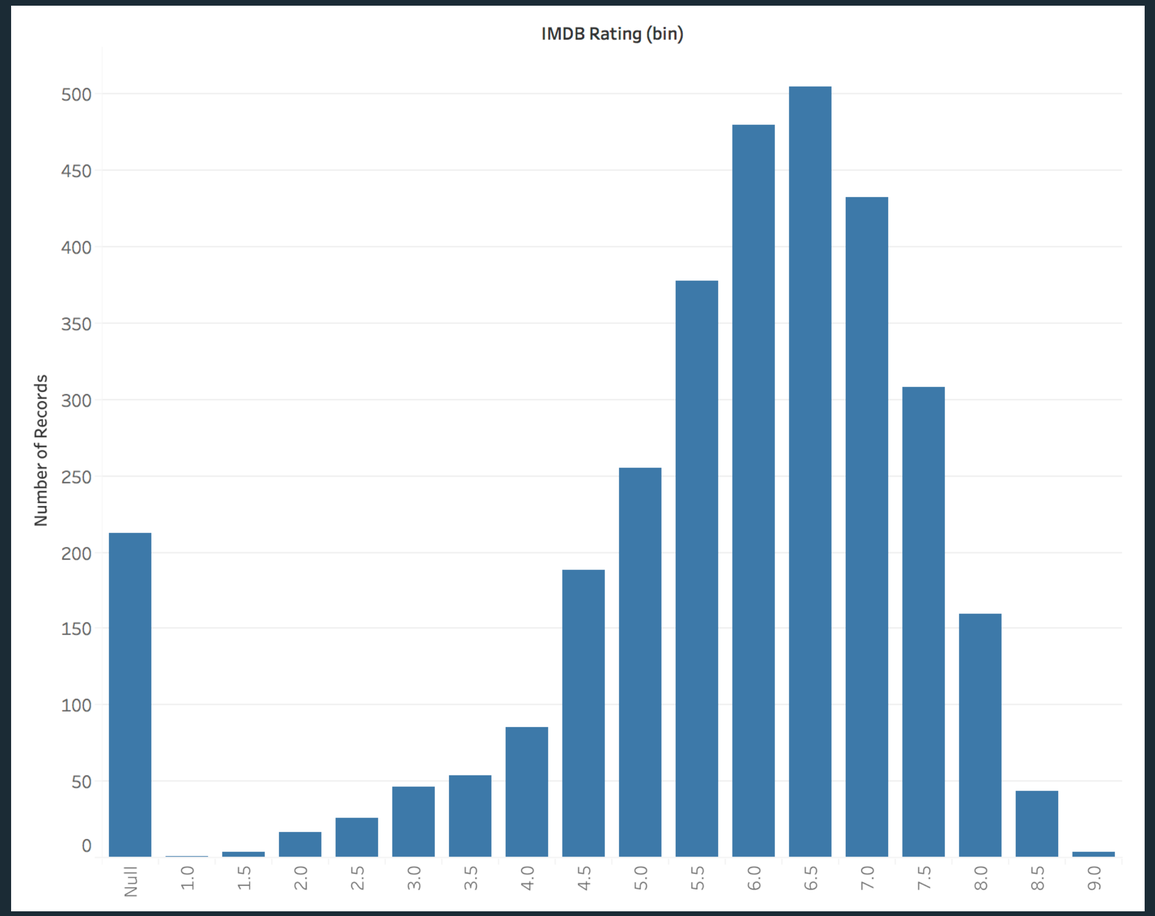


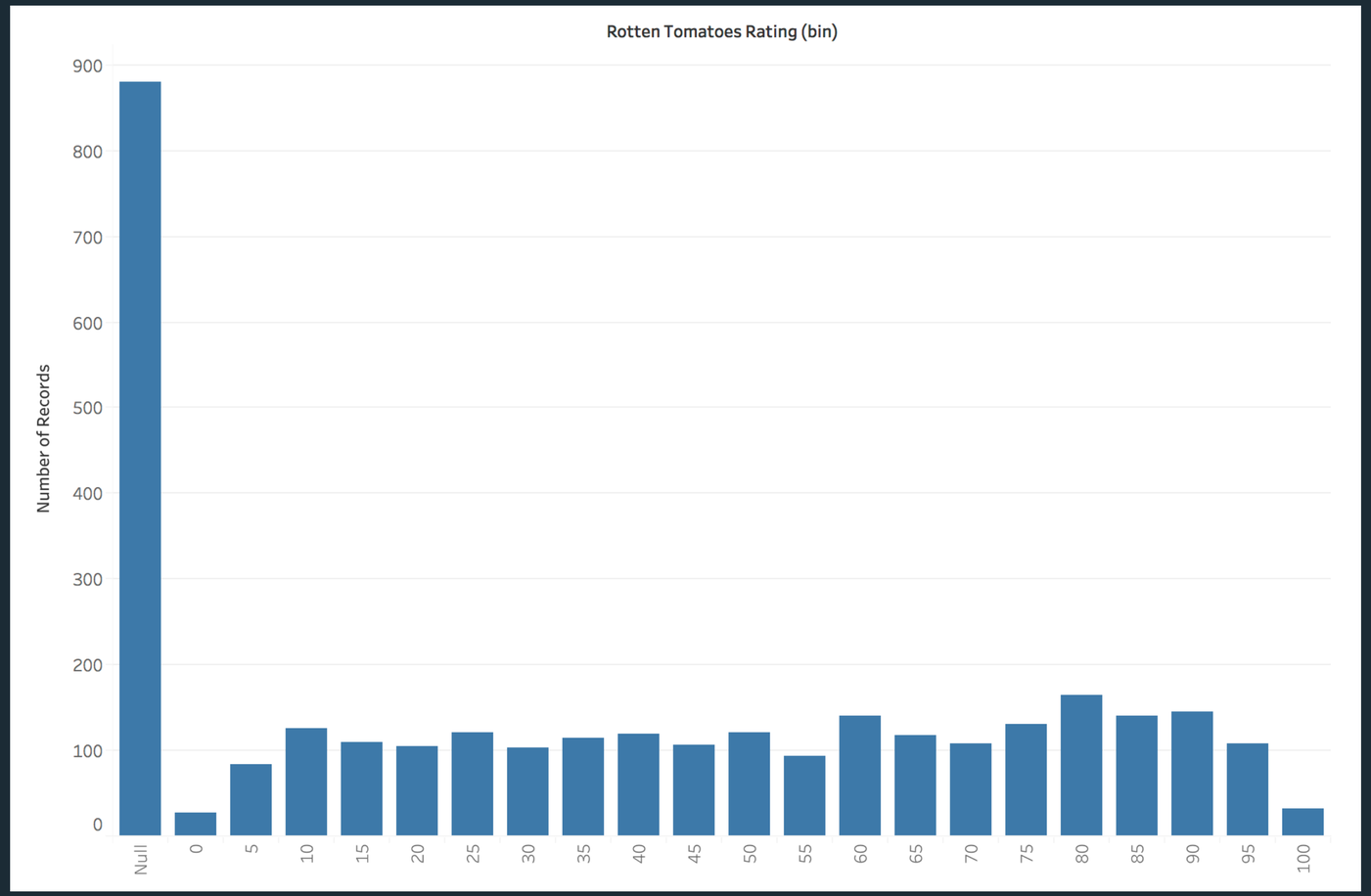
US Gross by Ratings

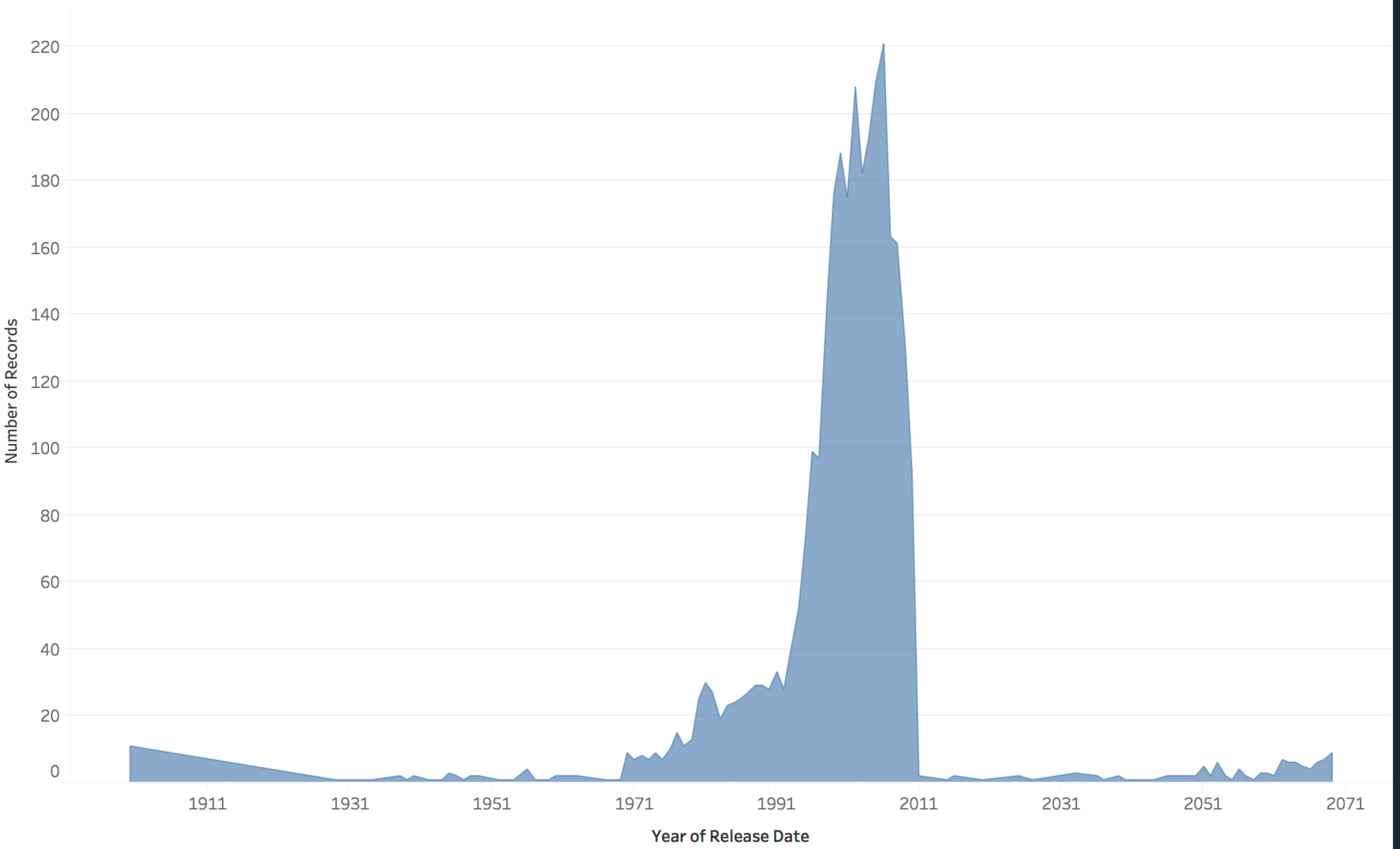


Audience vs. Critic Ratings




IMDB Rating (bin)
Number of Records
0
50
100
150
200
250
300
350
400
450
500
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
0
100
200
300
400
500
600
700
800
900
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

Exploratory Visual Analysis

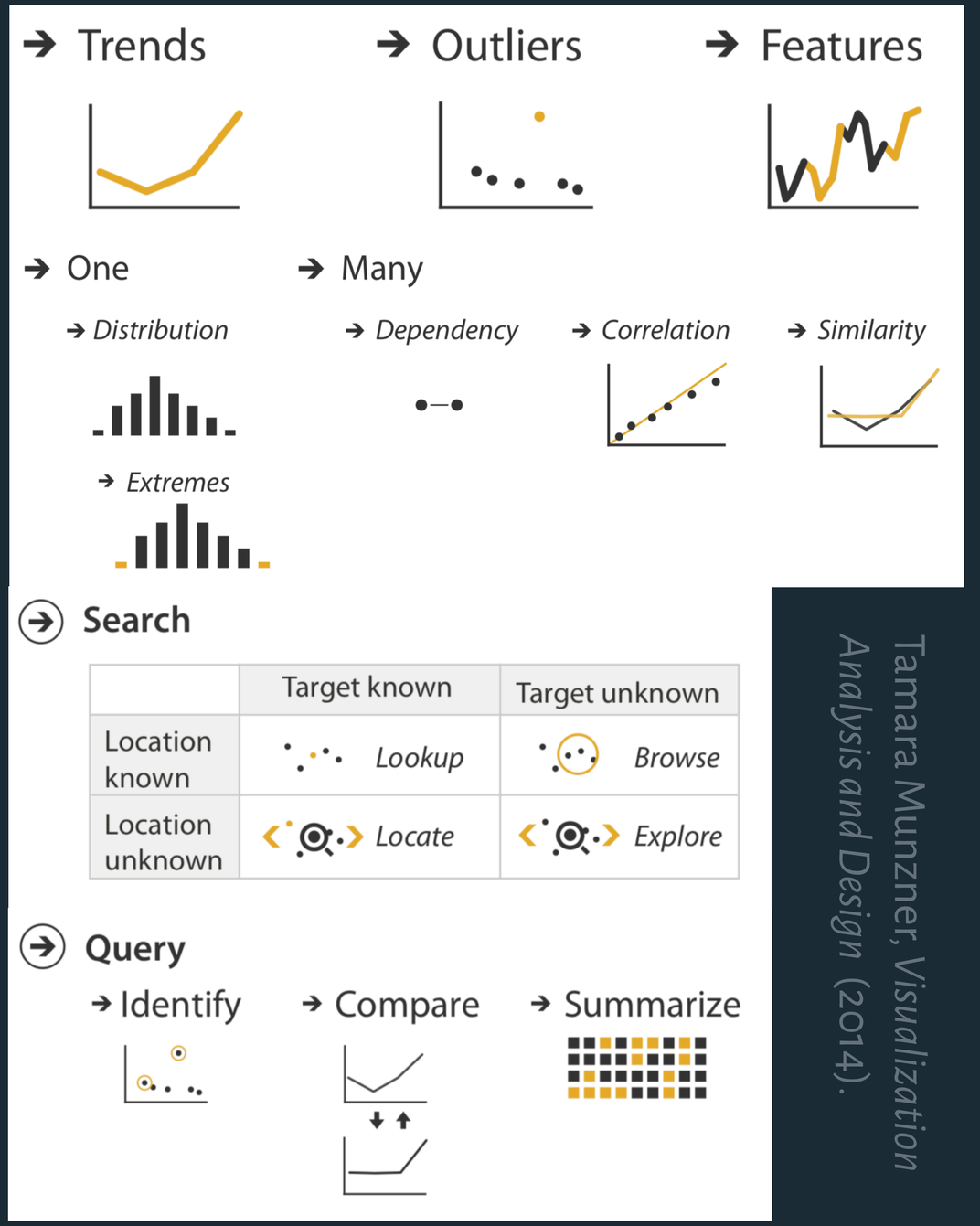
Process

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

What questions might we ask?

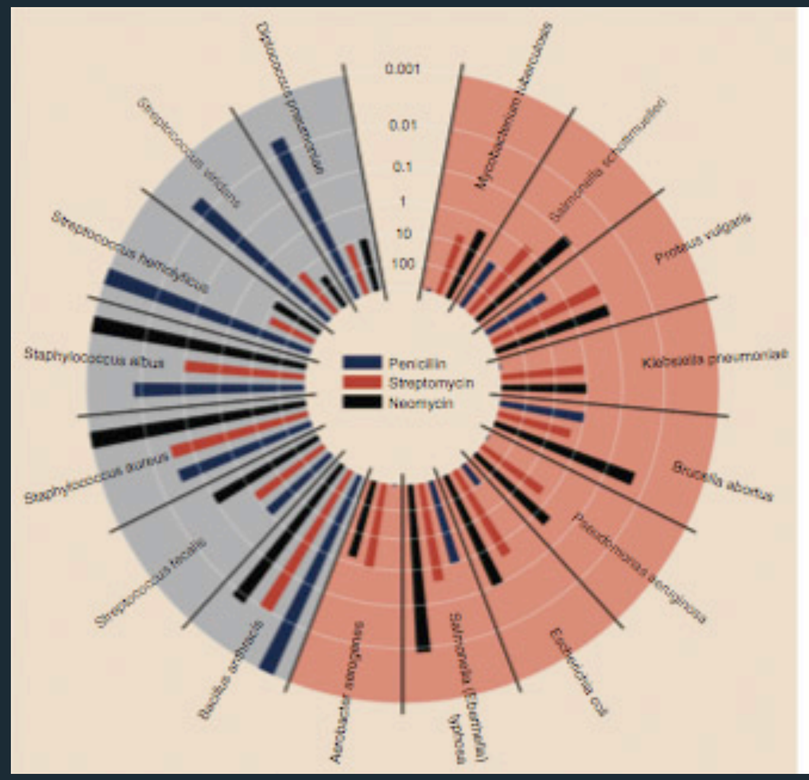
Collected prior to 1951

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	+

Encodings

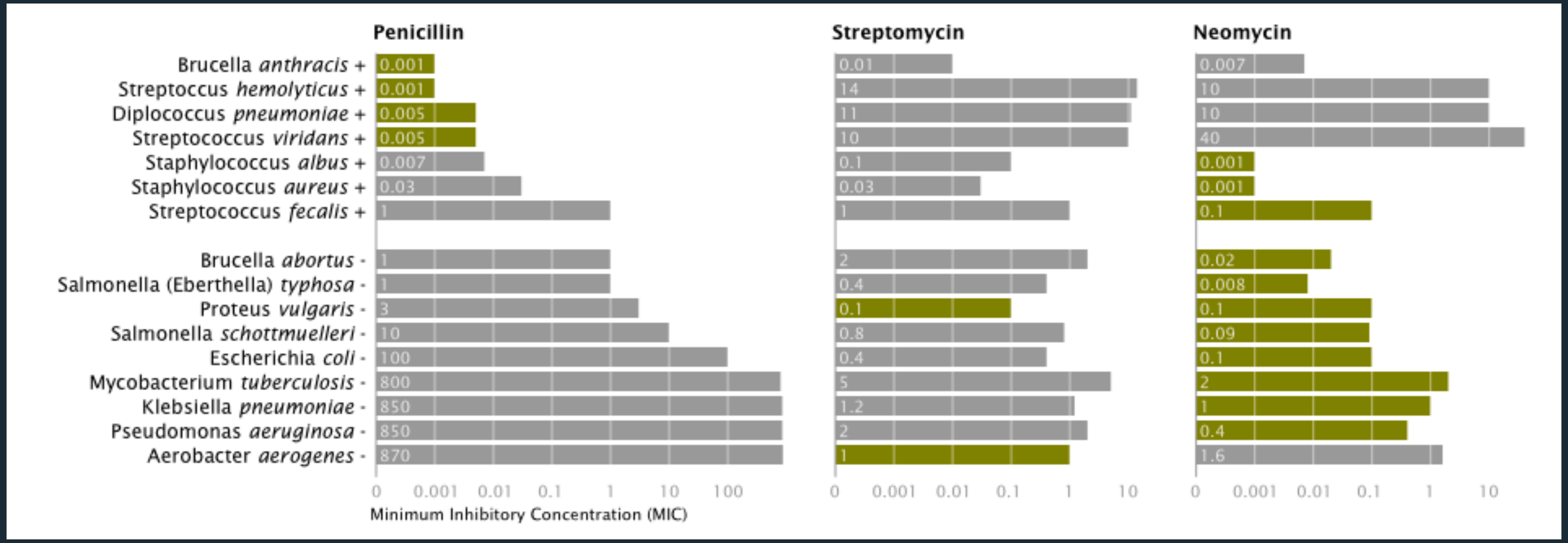
Radius: 1 / log(MIC)

Bar Color: Antibiotic

Background Color: Gram Staining

Original graphic by Will Burtin, 1951.

How do the drugs compare?



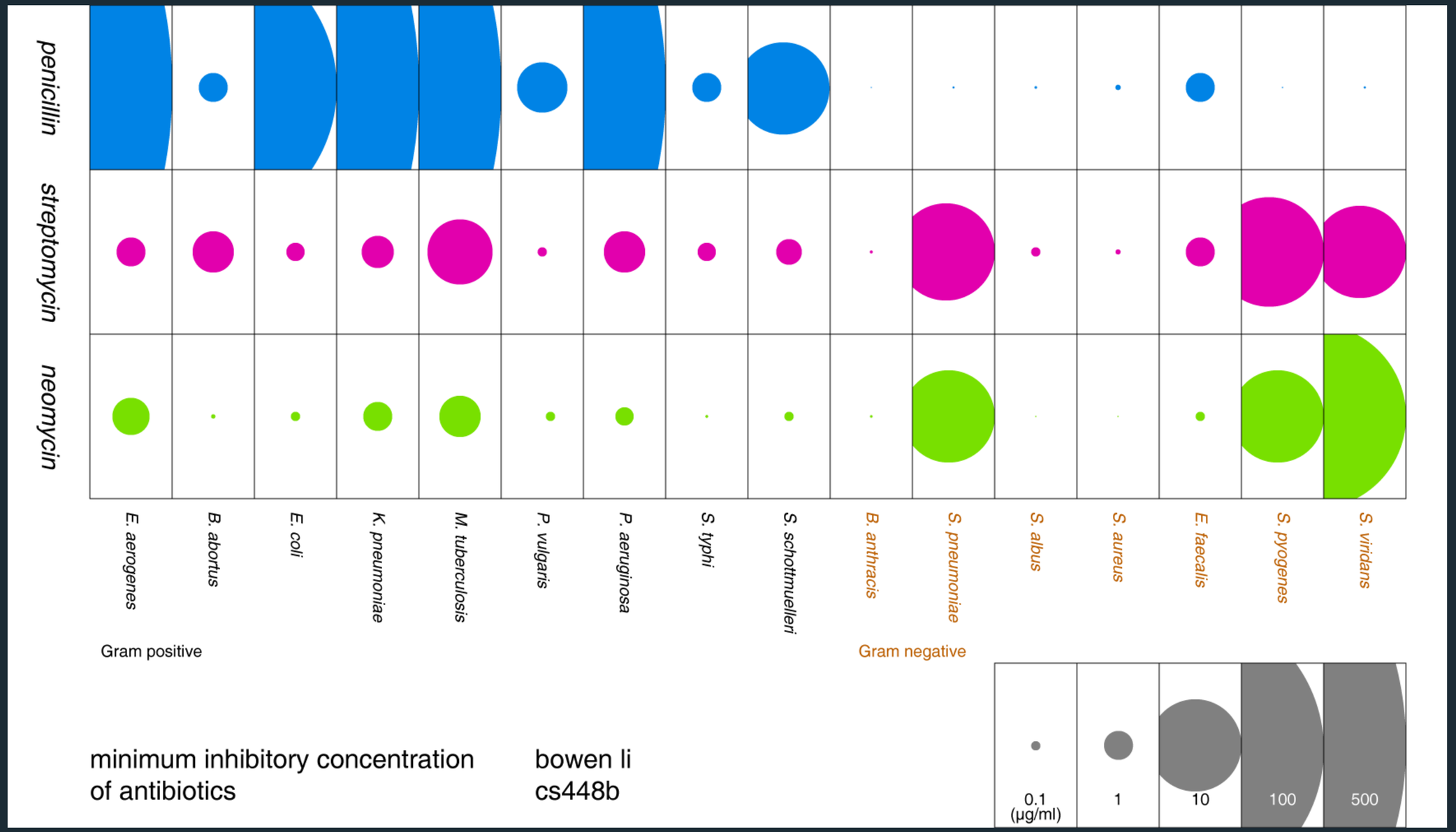
Mike Bostock, *Stanford CS448b* (Winter 2009).

X-Axis: Antibiotic | log(MIC)

Y-Axis: Gram-Staining | Species

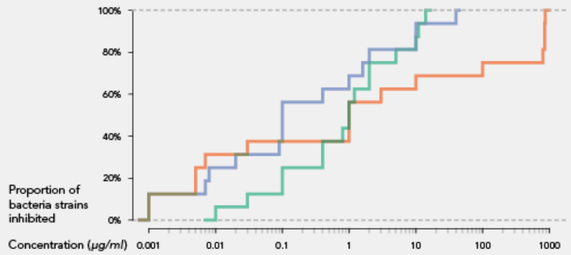
Color: Most Effective?

How do the drugs compare?



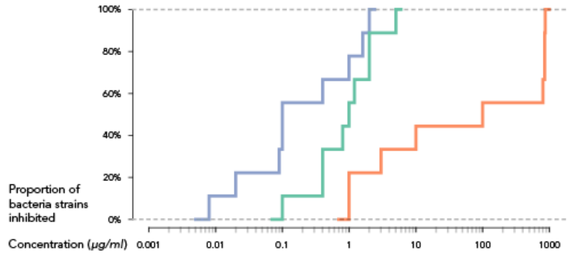
Bowen Li, *Stanford CS448b* (Fall 2009).

All bacteria



Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.

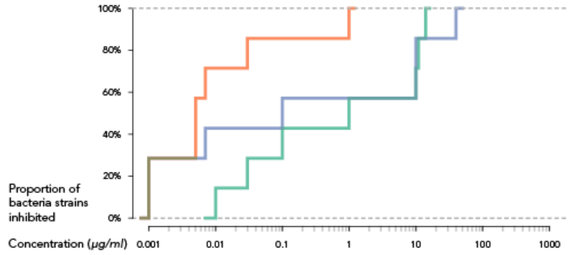
Gram-negative bacteria only



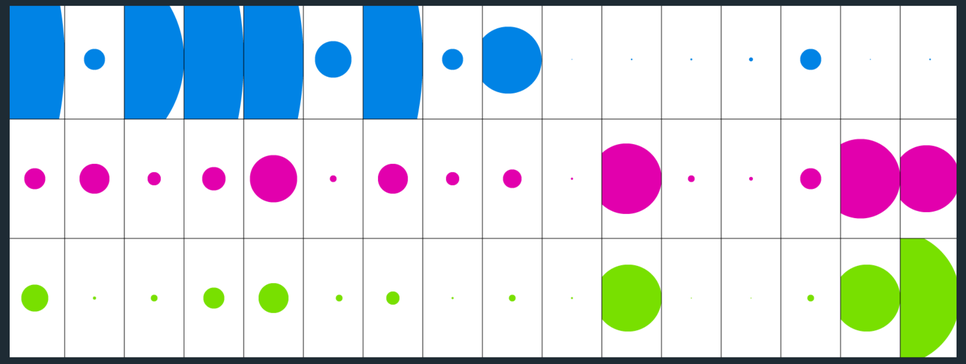
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.

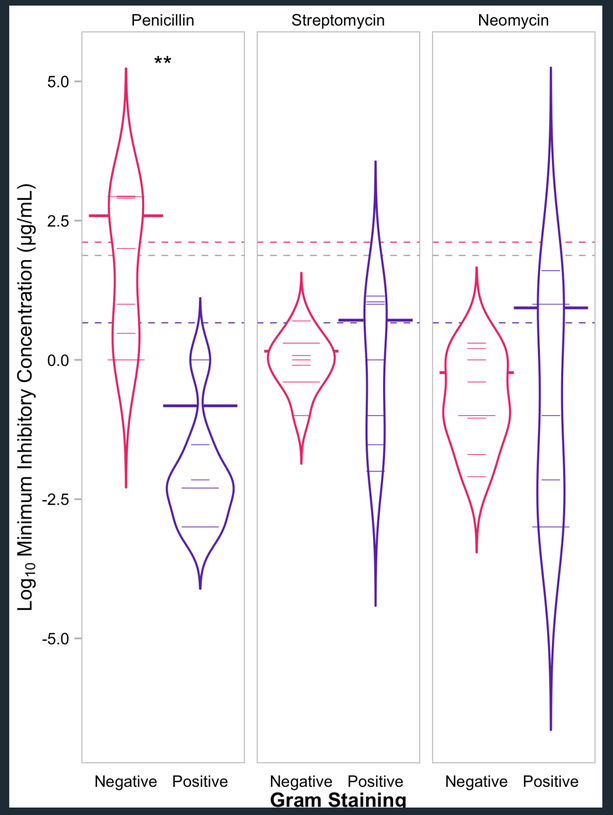
Gram staining quickly identifies bacteria as **Gram-negative** or **Gram-positive**, which can be used to find a more efficient antibiotic and dosage.

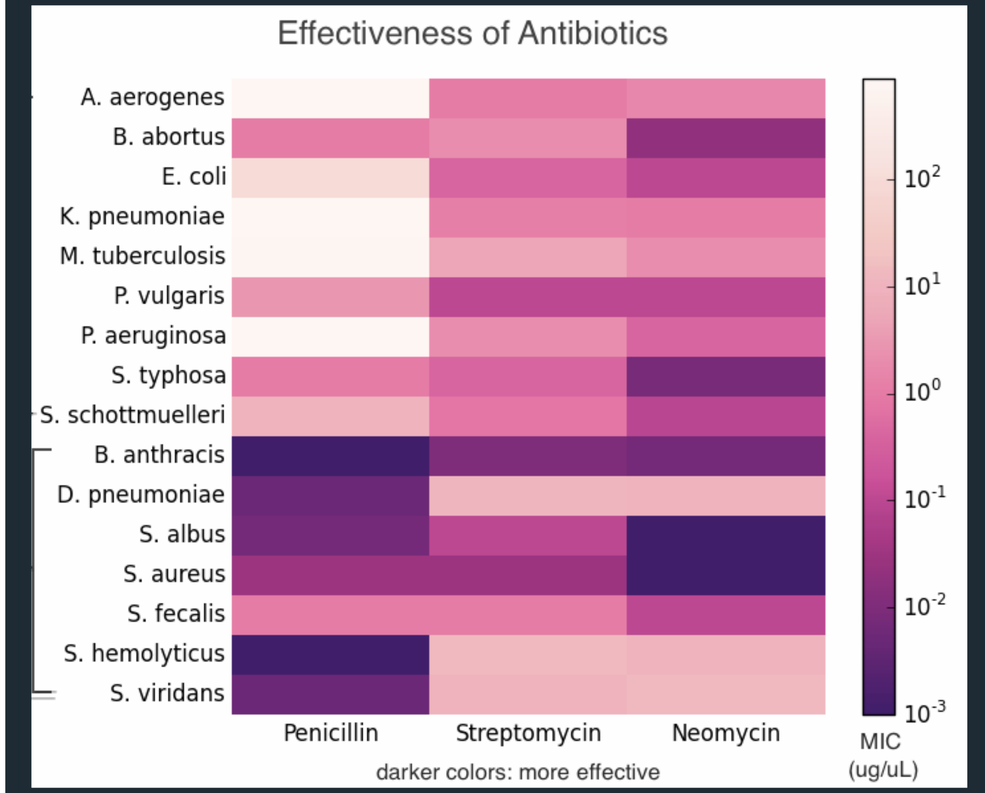
Gram-positive bacteria only

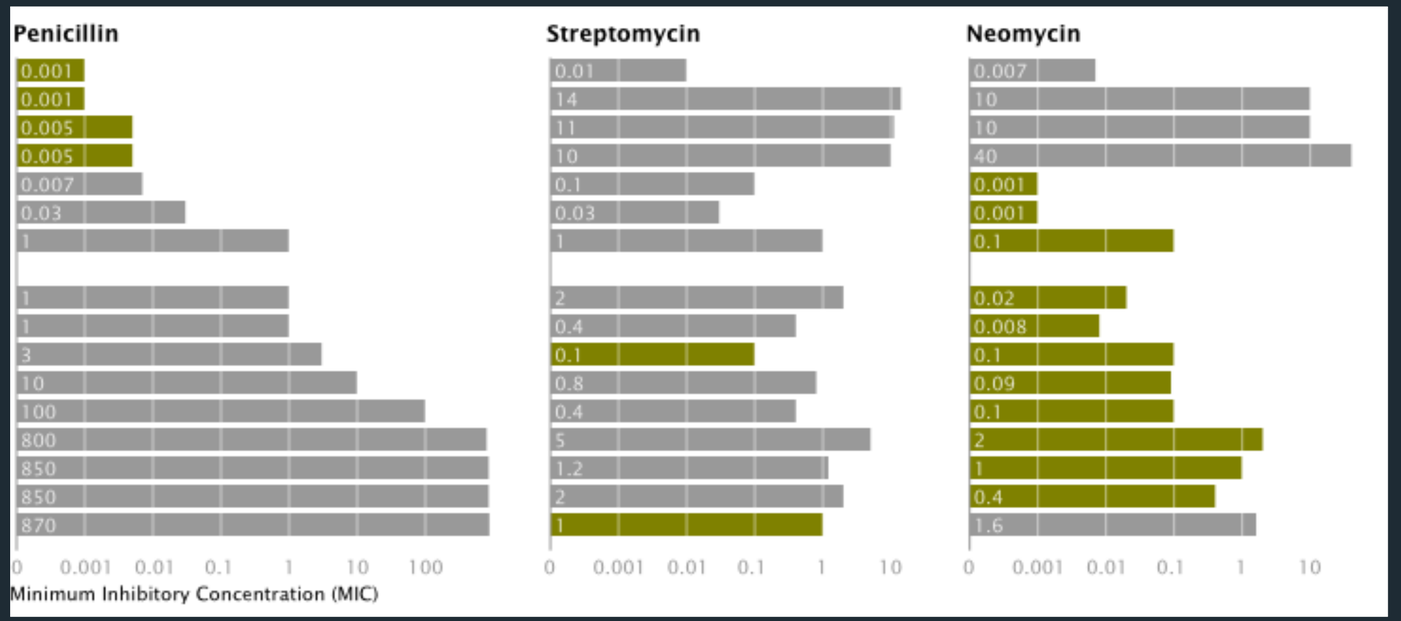


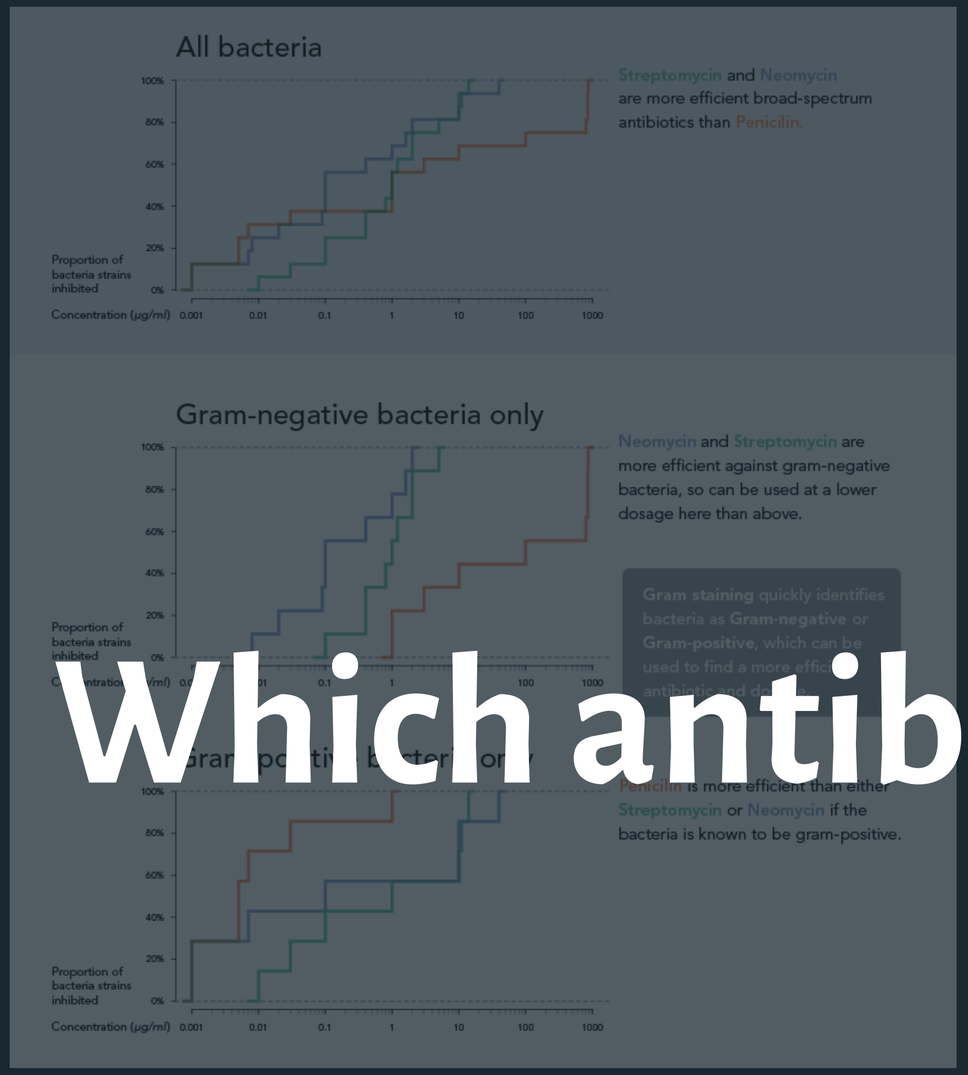
Penicilin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.

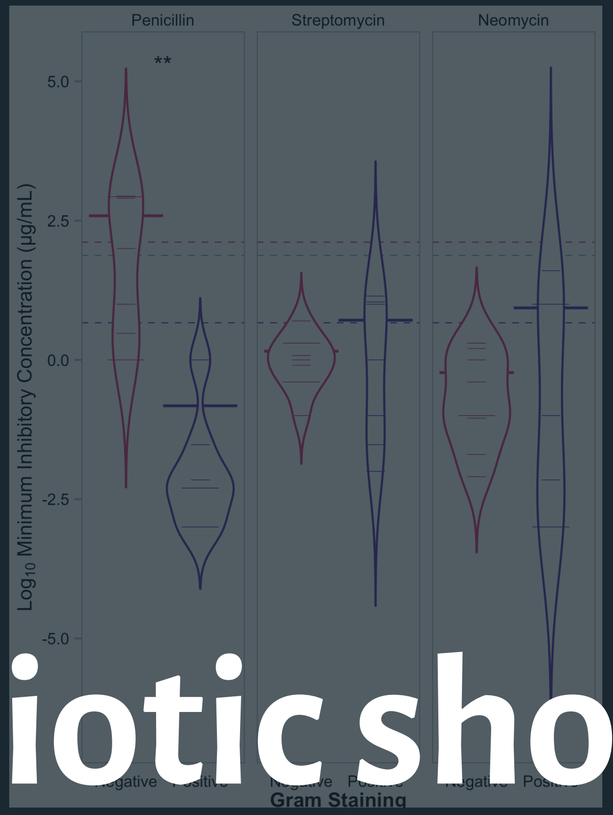


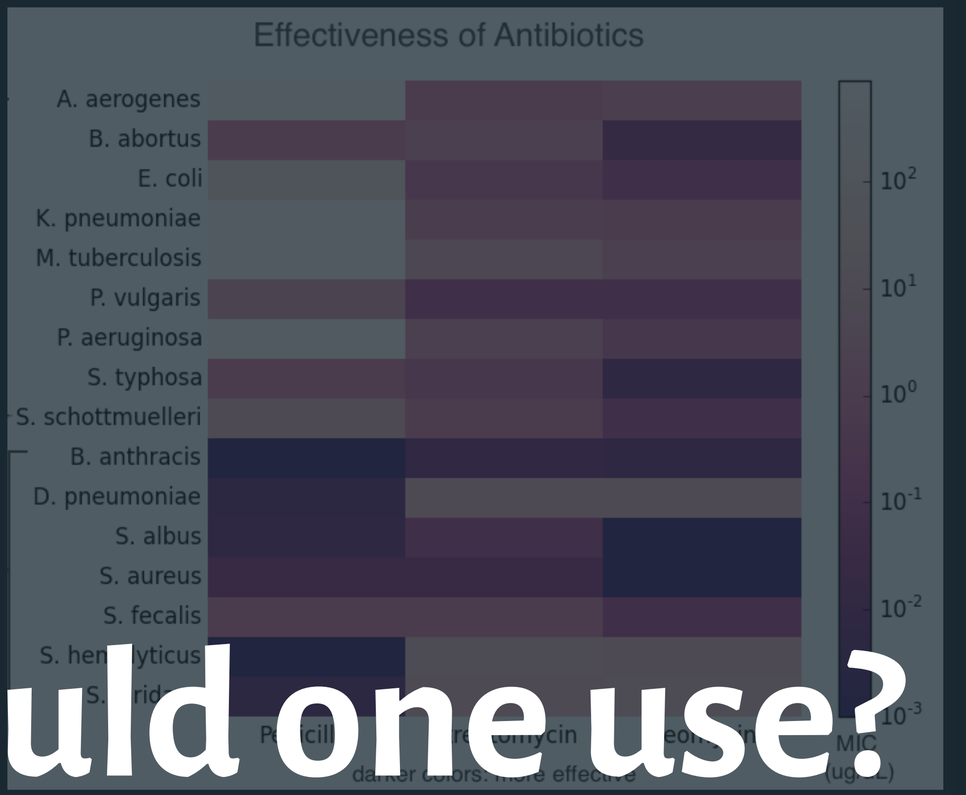




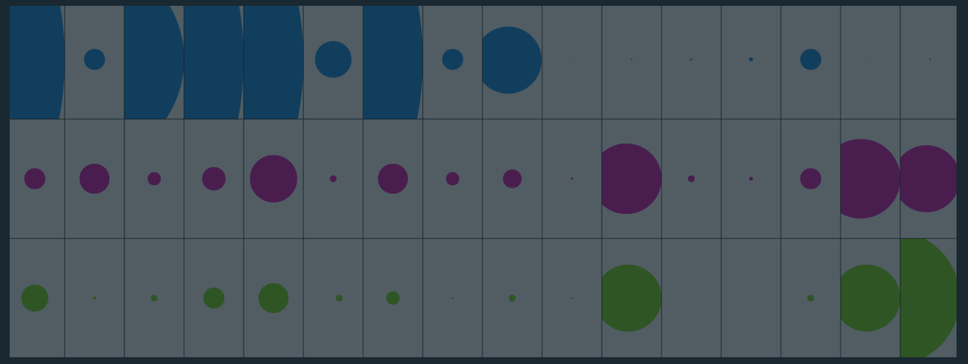


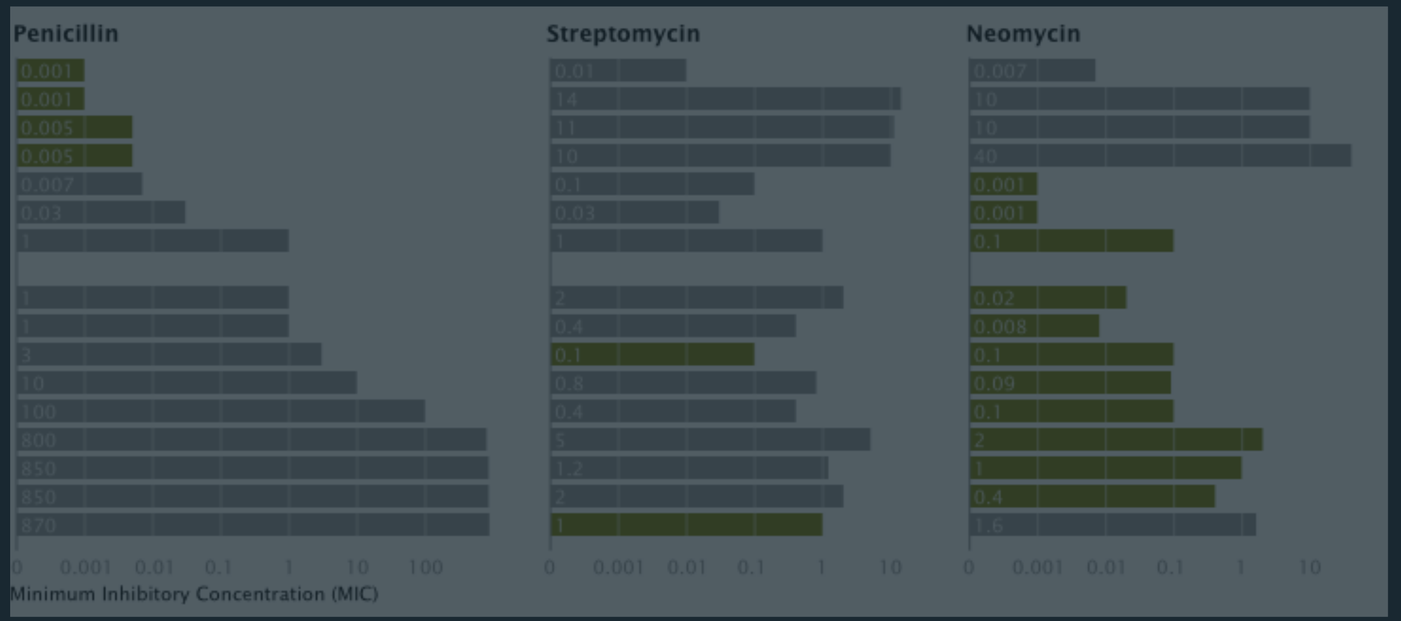




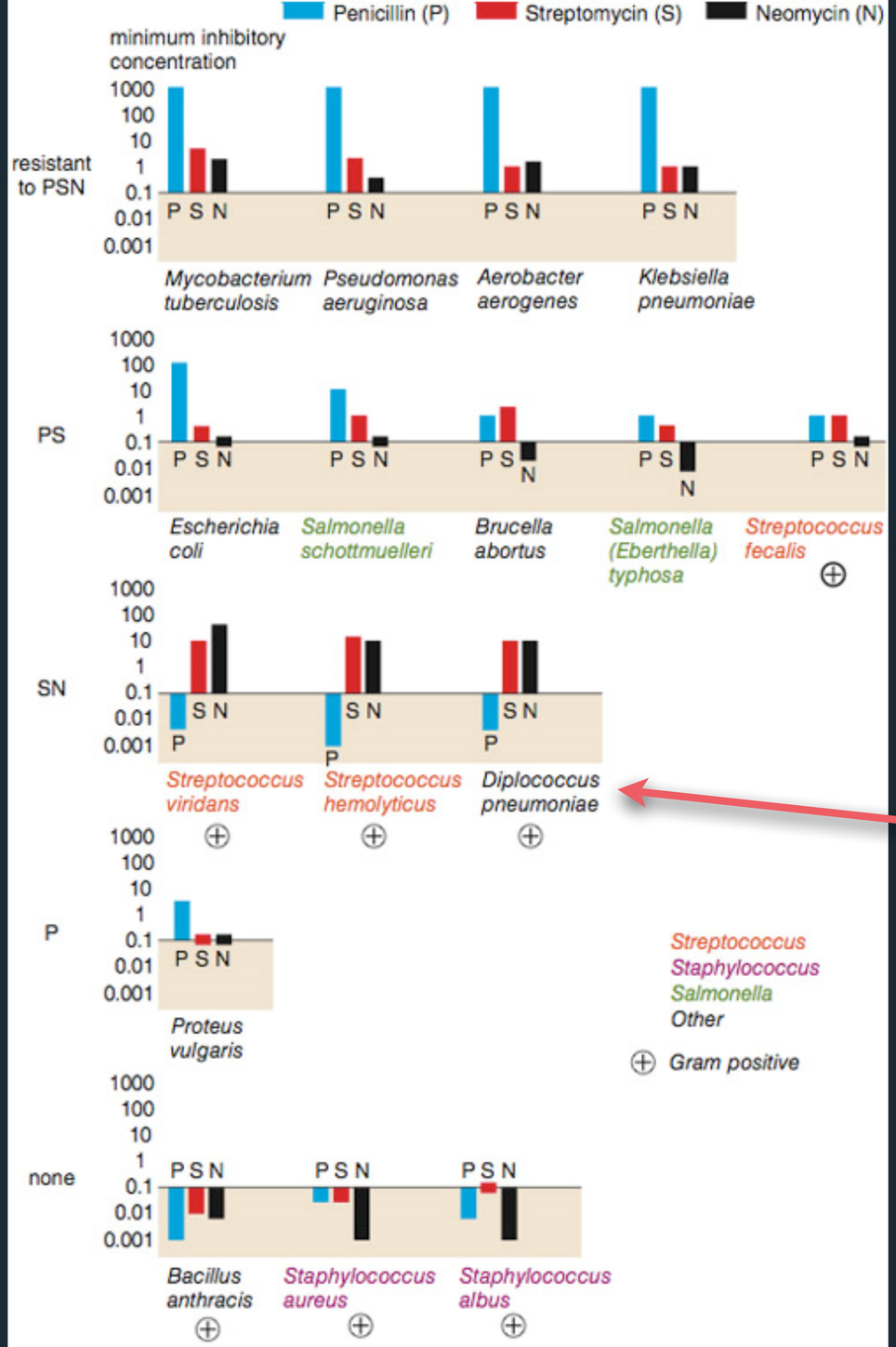


Which antibiotic should one use?





Wainer & Lysen. *American Scientist*, 2009



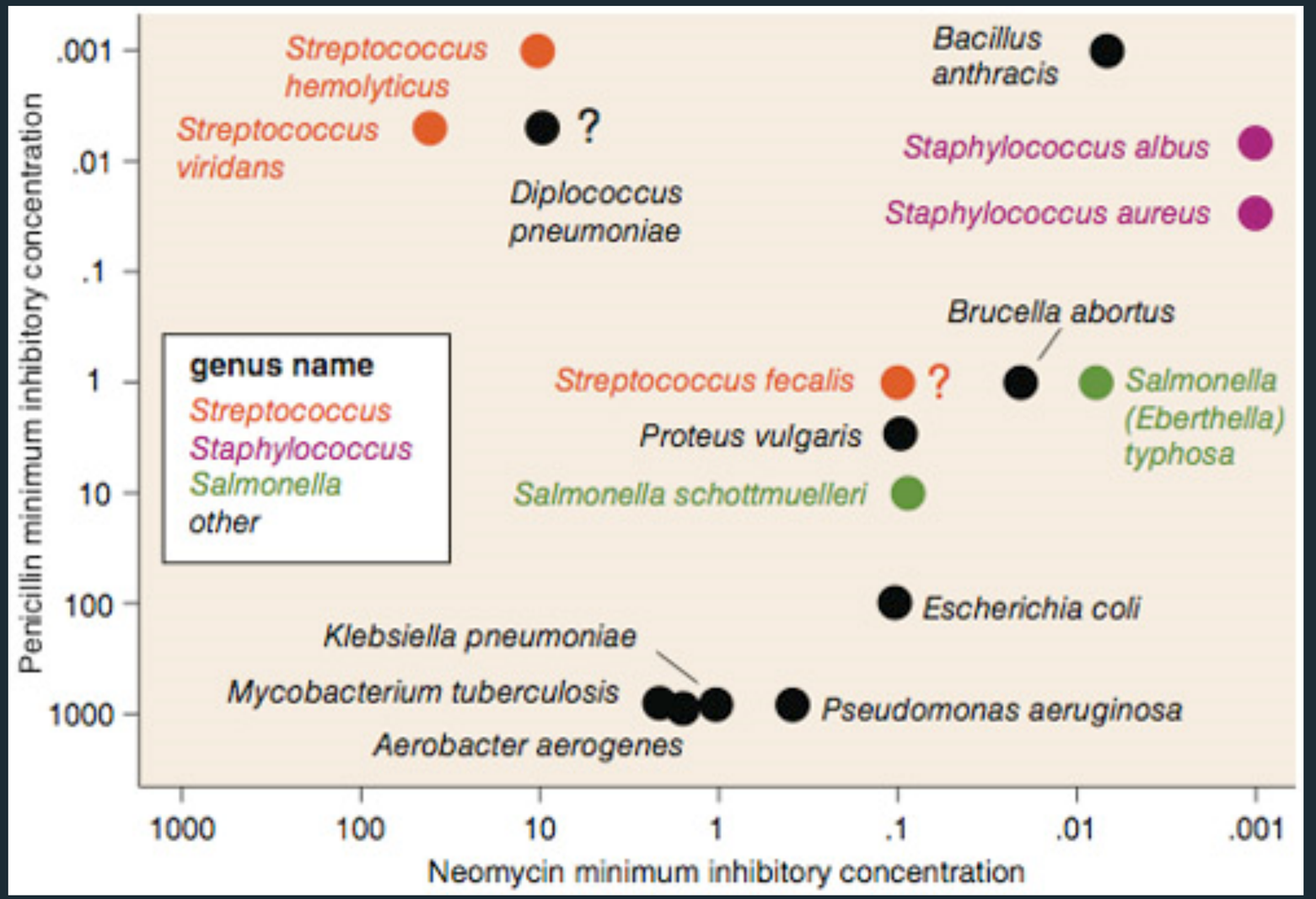
Do the bacteria group by antibiotic resistance?

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

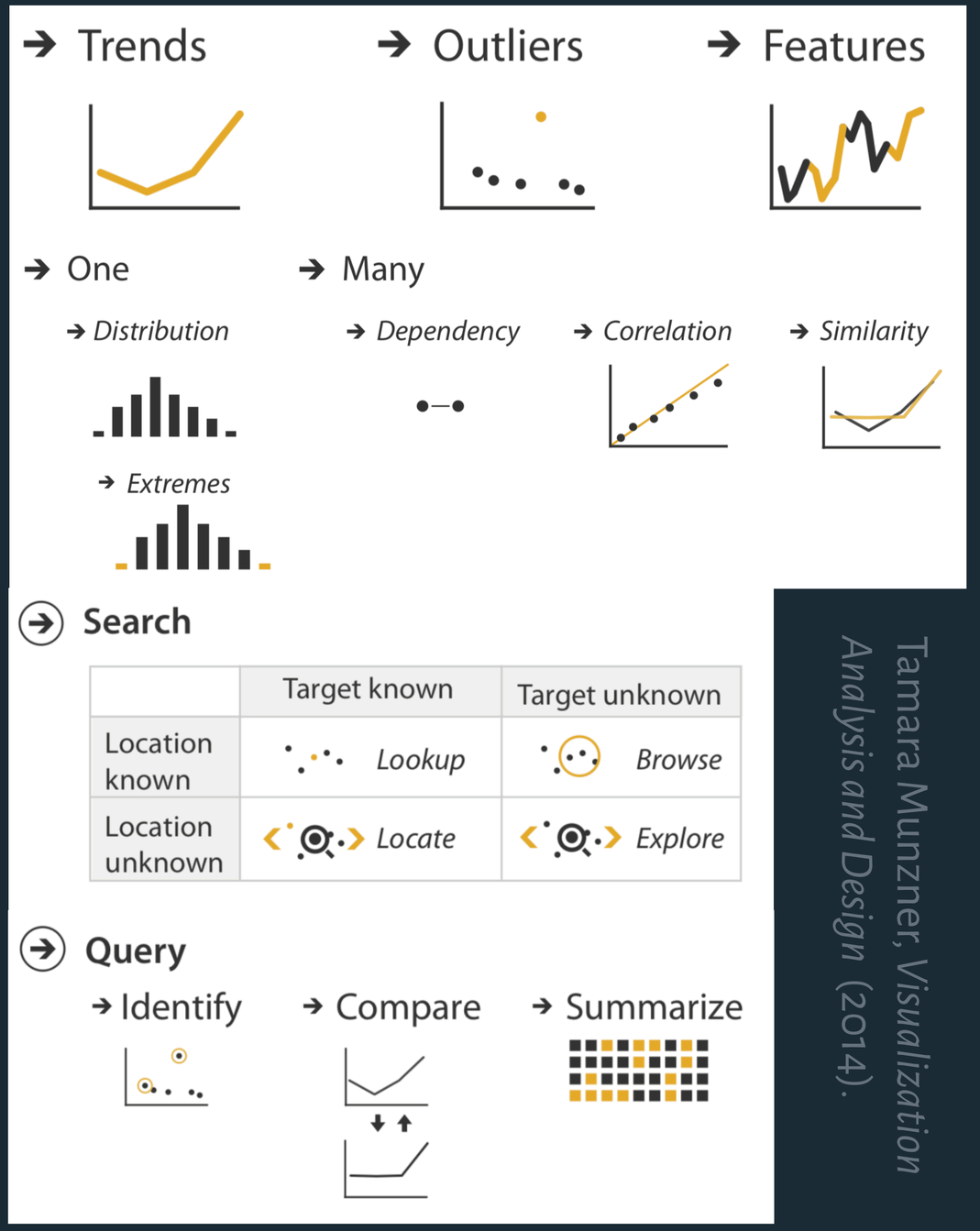
1. Construct graphics to address questions.
2. Inspect "answer" and ask new questions.
3. Iterate…

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).