Data Analysis

ID 413: Information Graphics and Data Visualization

Spring 2016

Venkatesh Rajamanickam (@venkatrajam)
venkatra@iitb.ac.in
http://info-design-lab.github.io/ID413-DataViz/

Objectives

- This is by no means a technical lecture in either statistics or data analysis
- I will however talk about the essentials, fallacies, paradoxes and curiosities in data collection & analysis
- And why good data and sound analysis matters a great deal to good visualization

Abraham Wald and the missing bullet holes

- A classified WW2 program called *The Statistical Research Group (SRG)* assembled might of American statisticians to the war effort—something like the Manhattan Project, except the weapons being developed were equations, not explosives

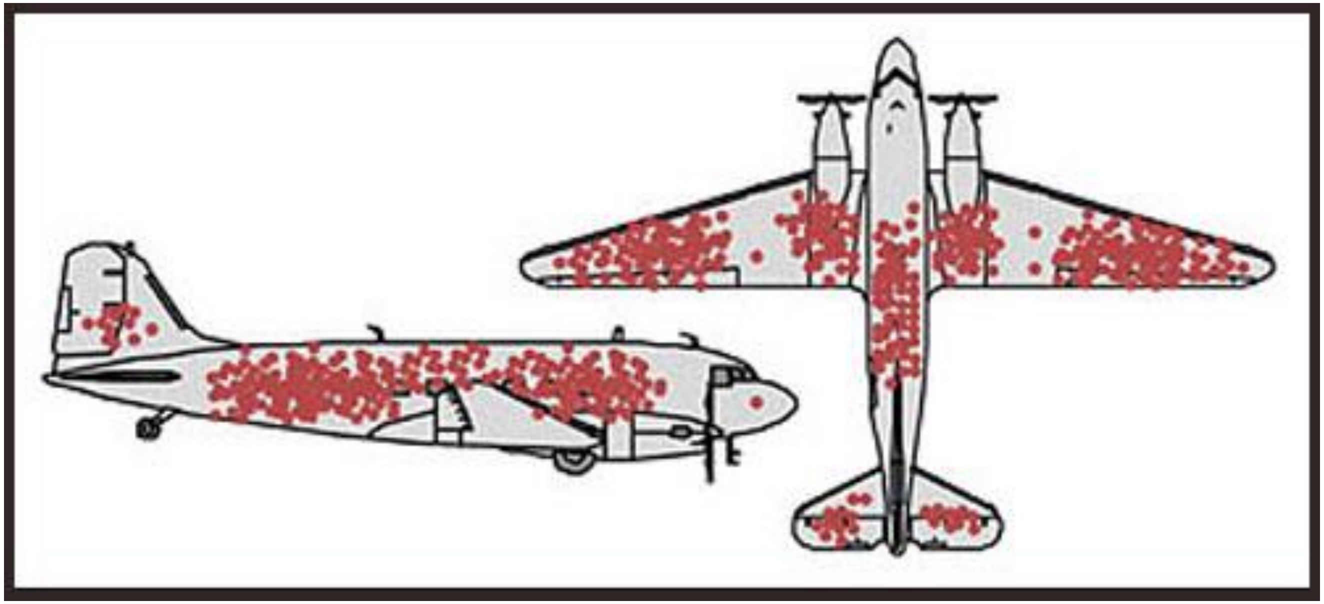
- Frederick Mosteller (founder Harvard's Statistics department), Leonard Jimmie Savage (father of Bayesian statistics), Norbert Wiener (MIT mathematician and the creator of cybernetics), Milton Friedman, (future Nobelist in economics)

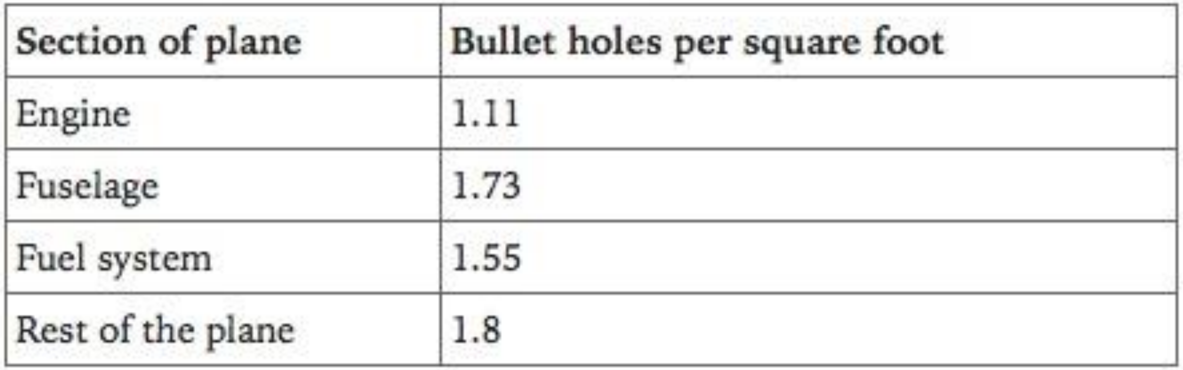
- Optimization: The winners in any war are usually the side which get 5% fewer of their planes shot down, or use 5% less fuel, or get 5% more nutrition into their infantry at 95% of the cost

Abraham Wald and the missing bullet holes

Abraham Wald and the missing bullet holes

Abraham Wald and the missing bullet holes





Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

Abraham Wald and the missing bullet holes

- The officers saw an opportunity for efficiency.

- They wanted to put the armour on parts of plane with the greatest need, where the planes are getting hit the most.

- But exactly how much more armour to put on those parts of the plane? That was the answer they came to Wald for.

- The armour, said Wald, doesn’t go where the bullet holes are. It goes where the bullet holes aren’t.

- What is the insight Wald was hoping to get by asking: where are the missing holes?

- What is the error the officers were making?

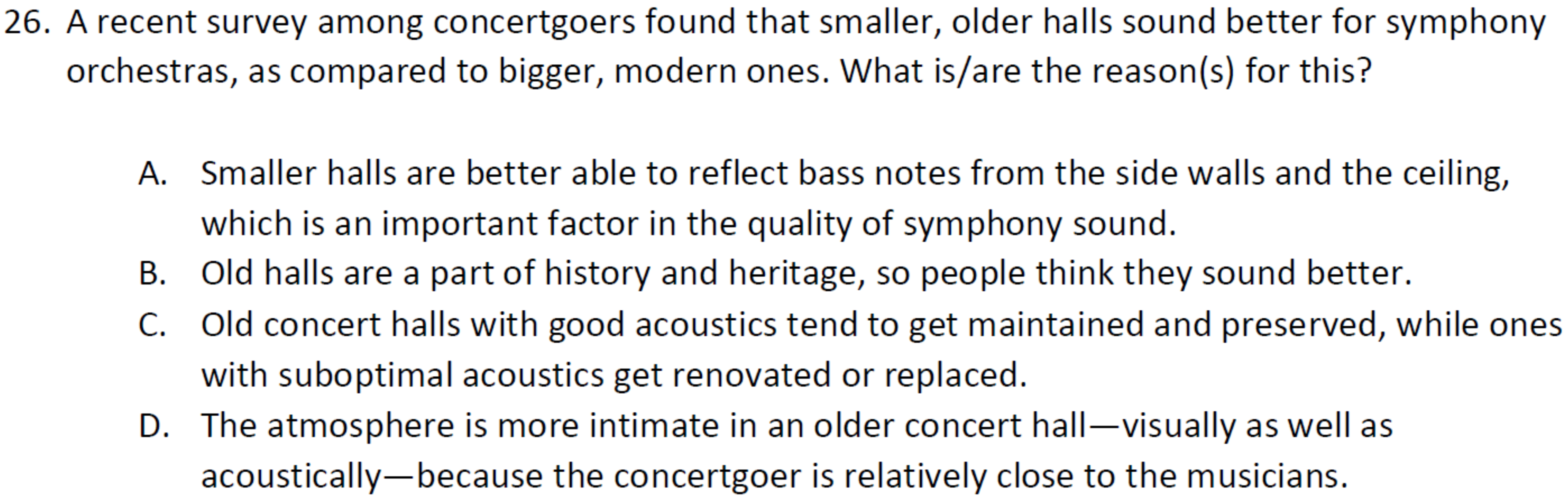
Abraham Wald and the missing bullet holes

- The officers saw an opportunity for efficiency
- They wanted to put the armour on parts of plane with the greatest need, where the planes are getting hit the most
- But exactly how much more armour yo put on those parts of the plane? That was the answer they came to Wald for.
- The armour, said Wald, doesn’t go where the bullet holes are. It goes where the bullet holes aren’t
- What is the insight Wald was hoping to get by asking: where are the missing holes?
- What is the error the officers were making?

The planes that came back were a random sample of all the planes

A question from UCEED 2016

Multiple Select Question – one or more than one correct answer(s)



26. A recent survey among concertgoers found that smaller, older halls sound better for symphony orchestras, as compared to bigger, modern ones. What is/are the reason(s) for this?

    A. Smaller halls are better able to reflect bass notes from the side walls and the ceiling, which is an important factor in the quality of symphony sound.
    B. Old halls are a part of history and heritage, so people think they sound better.
    C. Old concert halls with good acoustics tend to get maintained and preserved, while ones with suboptimal acoustics get renovated or replaced.
    D. The atmosphere is more intimate in an older concert hall—visually as well as acoustically—because the concertgoer is relatively close to the musicians.

More examples of erroneous analysis of data

1. A certain company discovered that 40% of all sick days were taken on a Friday or a Monday. They immediately clamped down on sick leave before they realized their mistake. What was the mistake?

2. Before the introduction of tin helmets during the First World War soldiers only had cloth hats to wear. After the introduction of tin helmets, the number of injuries to the head increased dramatically leading authorities to mistakenly conclude that the design of helmet may be flawed. What was the mistake?

3. A study in the U.S found 70% of black babies were born out of wedlock (compared to 40% national average) and concluded that all the poverty and educational problems of African-Americans are caused by the fact that too many of their children are born and raised out of wedlock, and presumably by single parents, and that it’s better to promote “traditional marriage” instead of affirmative action, welfare etc. What data did the conclusion omit?

More examples of erroneous analysis of data

4. A study was conducted on four-year-olds, comparing those who went to pre-school and socialised with other children, with those that stayed at home with their mothers. It measured aggressive behaviour such as stealing toys, pushing other children and starting fights.

   It showed that children who went to pre-school were three times more likely to be aggressive than those who stayed at home with their mothers. The report was used to persuade parents to keep their children at home until they start school, aged five.

5. A study found a strong correlation between accidental deaths due to drownings and the amount of ice cream sold by street vendors! (Which is the cause and which is the effect?)

More examples of erroneous analysis of data

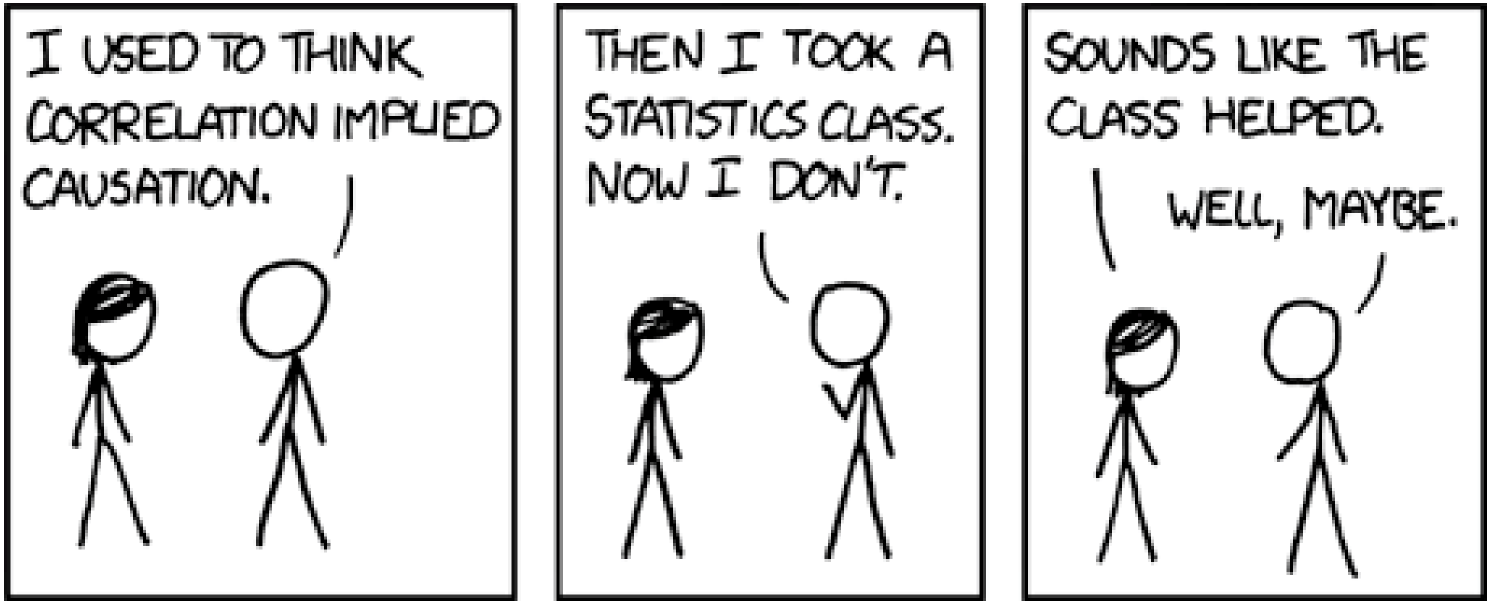
4. A study was conducted on four-year-olds, comparing those who went to pre-school and socialised with other children, with those that stayed at home with their mothers. It measured aggressive behaviour such as stealing toys, pushing other children and starting fights.

   It showed that children who went to pre-school were three times more likely to be aggressive than those who stayed at home with their mothers. The report was used to persuade parents to keep their children at home until they start school, aged five.

5. A study found a strong correlation between accidental deaths due to drownings and the amount of ice cream sold by street vendors! (Which is the cause and which is the effect?)

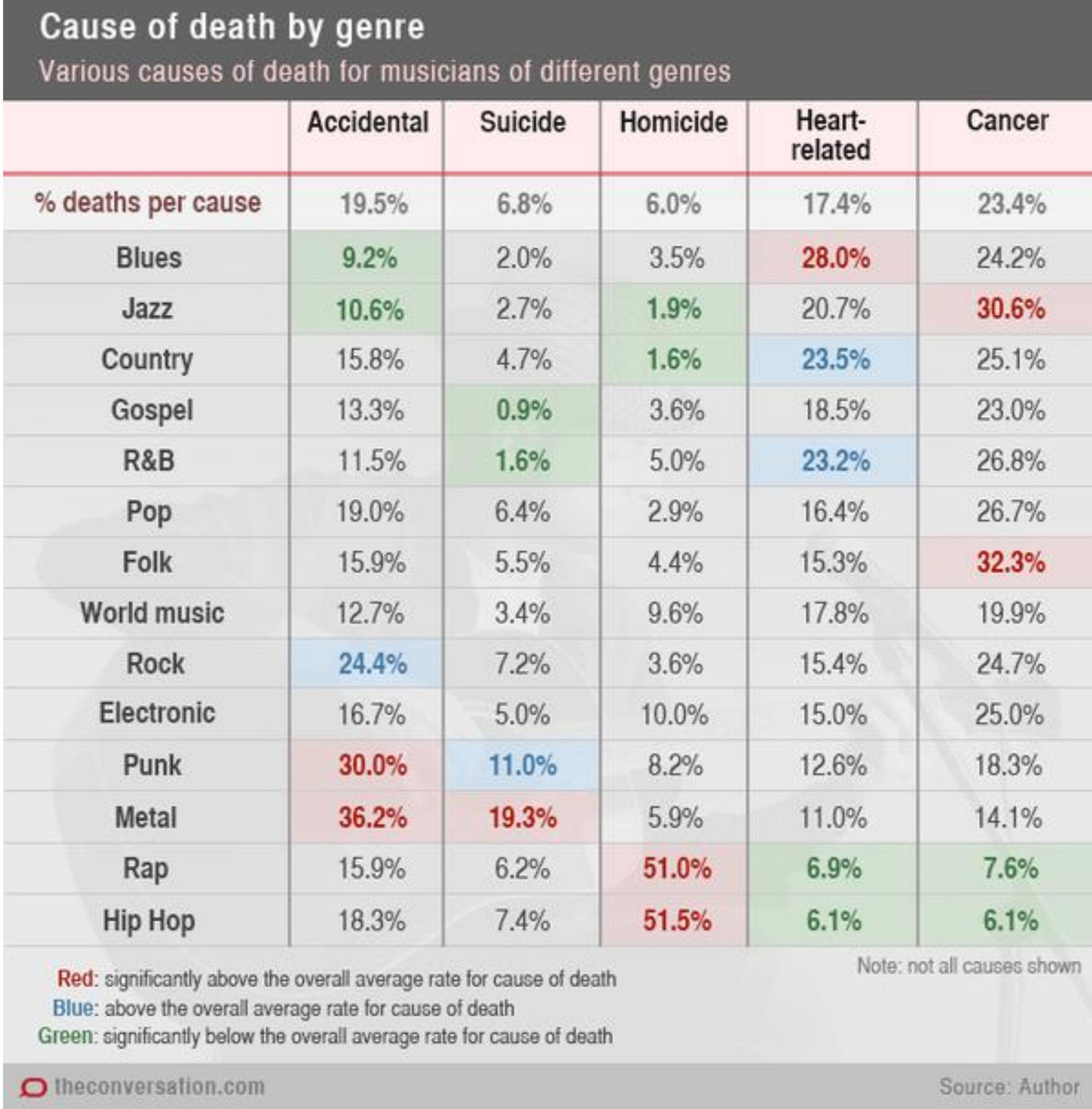
   Obviously, there was an unobserved variable causing both. Summers and good weather days are when people go for a swim, and when the most ice cream is sold.

More examples of erroneous analysis of data


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Analyse this!

Why are Rap and Hip Hop artists overwhelmingly die due to homicide?



Cause of death by genre

Various causes of death for musicians of different genres

	Accidental	Suicide	Homicide	Heart-related	Cancer
% deaths per cause	19.5%	6.8%	6.0%	17.4%	23.4%
Blues	9.2%	2.0%	3.5%	28.0%	24.2%
Jazz	10.6%	2.7%	1.9%	20.7%	30.6%
Country	15.8%	4.7%	1.6%	23.5%	25.1%
Gospel	13.3%	0.9%	3.6%	18.5%	23.0%
R&B	11.5%	1.6%	5.0%	23.2%	26.8%
Pop	19.0%	6.4%	2.9%	16.4%	26.7%
Folk	15.9%	5.5%	4.4%	15.3%	32.3%
World music	12.7%	3.4%	9.6%	17.8%	19.9%
Rock	24.4%	7.2%	3.6%	15.4%	24.7%
Electronic	16.7%	5.0%	10.0%	15.0%	25.0%
Punk	30.0%	11.0%	8.2%	12.6%	18.3%
Metal	36.2%	19.3%	5.9%	11.0%	14.1%
Rap	15.9%	6.2%	51.0%	6.9%	7.6%
Hip Hop	18.3%	7.4%	51.5%	6.1%	6.1%

Red: significantly above the overall average rate for cause of death
Blue: above the overall average rate for cause of death
Green: significantly below the overall average rate for cause of death

Note: not all causes shown

theconversation.com

Source: Author

How a genre of music affects life expectancy of famous musicians in that genre

- For male musicians across all genres, accidental death (including all vehicular incidents and accidental overdose) accounted for almost 20% of all deaths. But accidental death for rock musicians was higher than this (24.4%) and for metal musicians higher still (36.2%).

- Suicide accounted for almost 7% of all deaths in the total sample. However, for punk musicians, suicide accounted for 11% of deaths; for metal musicians, a staggering 19.3%. At just 0.9%, gospel musicians had the lowest suicide rate of all the genres studied.

- Murder accounted for 6.0% of deaths across the sample, but was the cause of 51% of deaths in rap musicians and 51.5% of deaths for hip hop musicians, to date.

- So bluesmen and country singers are most likely to die of broken hearts?

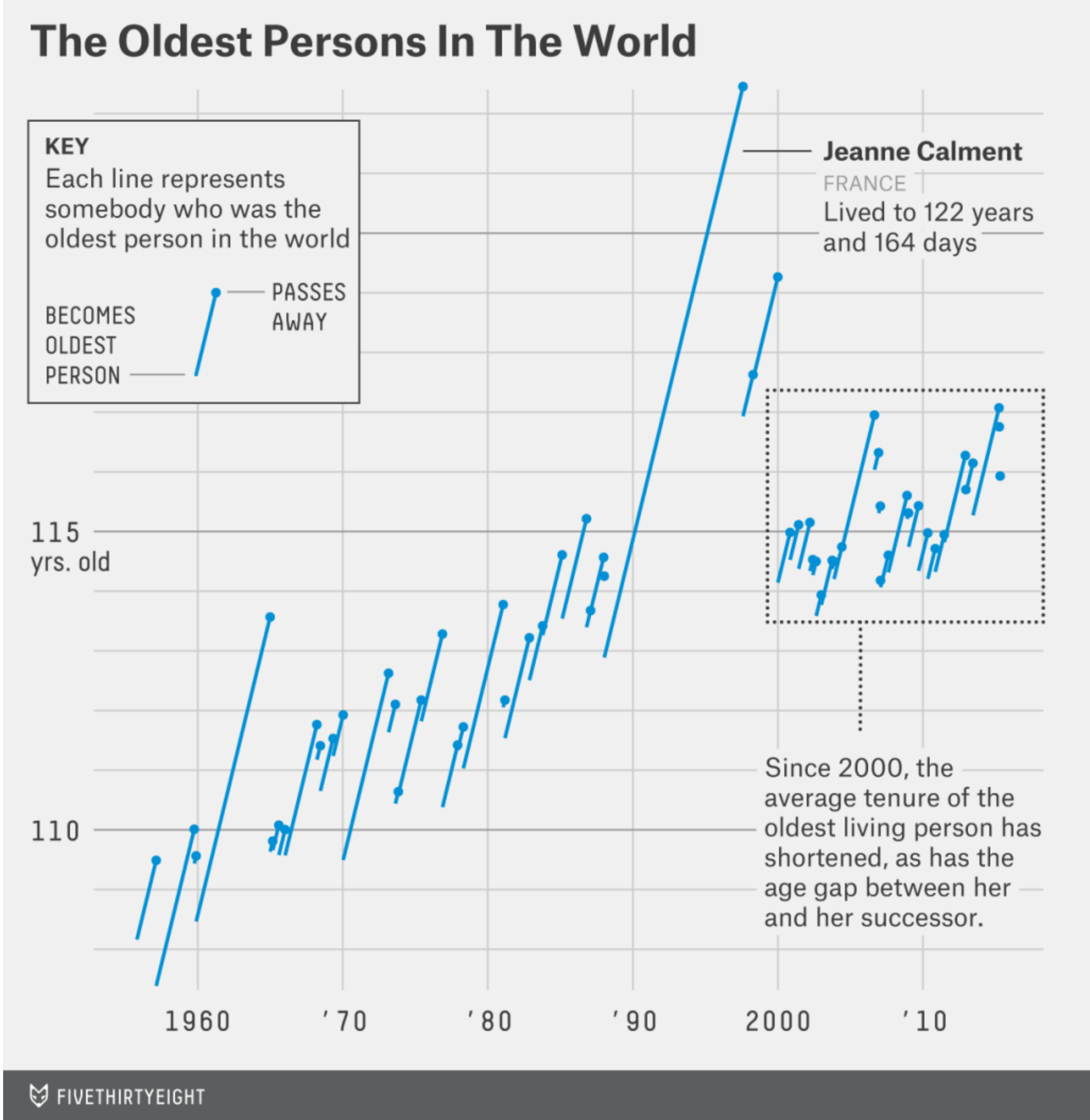
- Beware selection, because of course most rap musicians aren’t dead yet. This problem will be more extreme, the younger is the genre

How a genre of music affects life expectancy of famous musicians in that genre

- Beware selection, because of course most rap musicians aren’t dead yet.
- This problem will be more extreme, the younger is the genre.
- Another selection effect may be that getting killed, or dying in an unusual way, contributes to your fame.
- Remember at least half of rappers and hip hoppers who ever lived are still a live. They are too young to die of old age.
- Accident rate appears to track group size; genres with more solo artists (R&B, gospel) having lower rates and genres with a tendency for groups of 4-5 having higher rates. i.e., more “musicians” die when a rock group’s plane crashes than when an R&B singer’s does.

Analyse this!

Why the oldest person in the world keeps on dying?



The Oldest Persons In The World

KEY
Each line represents somebody who was the oldest person in the world

Jeanne Calment
FRANCE
Lived to 122 years and 164 days

Since 2000, the average tenure of the oldest living person has shortened, as has the age gap between her and her successor.

FIVETHIRTYEIGHT

4 critical questions

1. Where did the data come from?
   (Who ran the survey? Who payed for it? Do they have an ulterior motive for having the result go one way?)

2. Have the data been peer-reviewed?
   (Is it reliable? Was it interesting to warrant a peer review?)

3. How was the data collected?
   (What questions were asked? Who asked them? How did they ask them? Who was asked?)

4. Is data taken out of context?
   (Is the data cherry-picked? Are numbers taken out of context to support a foregone conclusion?)

We ask of data 3 things

1. That they be a representative sample of some group or population.
   (truly random)

2. That they provide some sense of comparison.
   (isolating the impact of one specific attribute)

3. That they can be described statistically.
   (mean, median, mode, standard deviation, distribution, index)

Descriptive Statistics

- Descriptive statistics help to frame the issue and to simplify, which always implies some loss of nuance or detail.
- The point of statistics is not do rigorous mathematical calculations, the point is to gain insight into meaningful social phenomena.
- Statistical inference is really just the marriage of two concepts—data and probability.
- The power of statistical inference derives from observing the most likely explanation for that outcome.

Descriptive Statistics

- Mean is the mathematical average of a data set and is the most commonly used form of central tendency.

- It is also the most commonly used descriptive statistic.

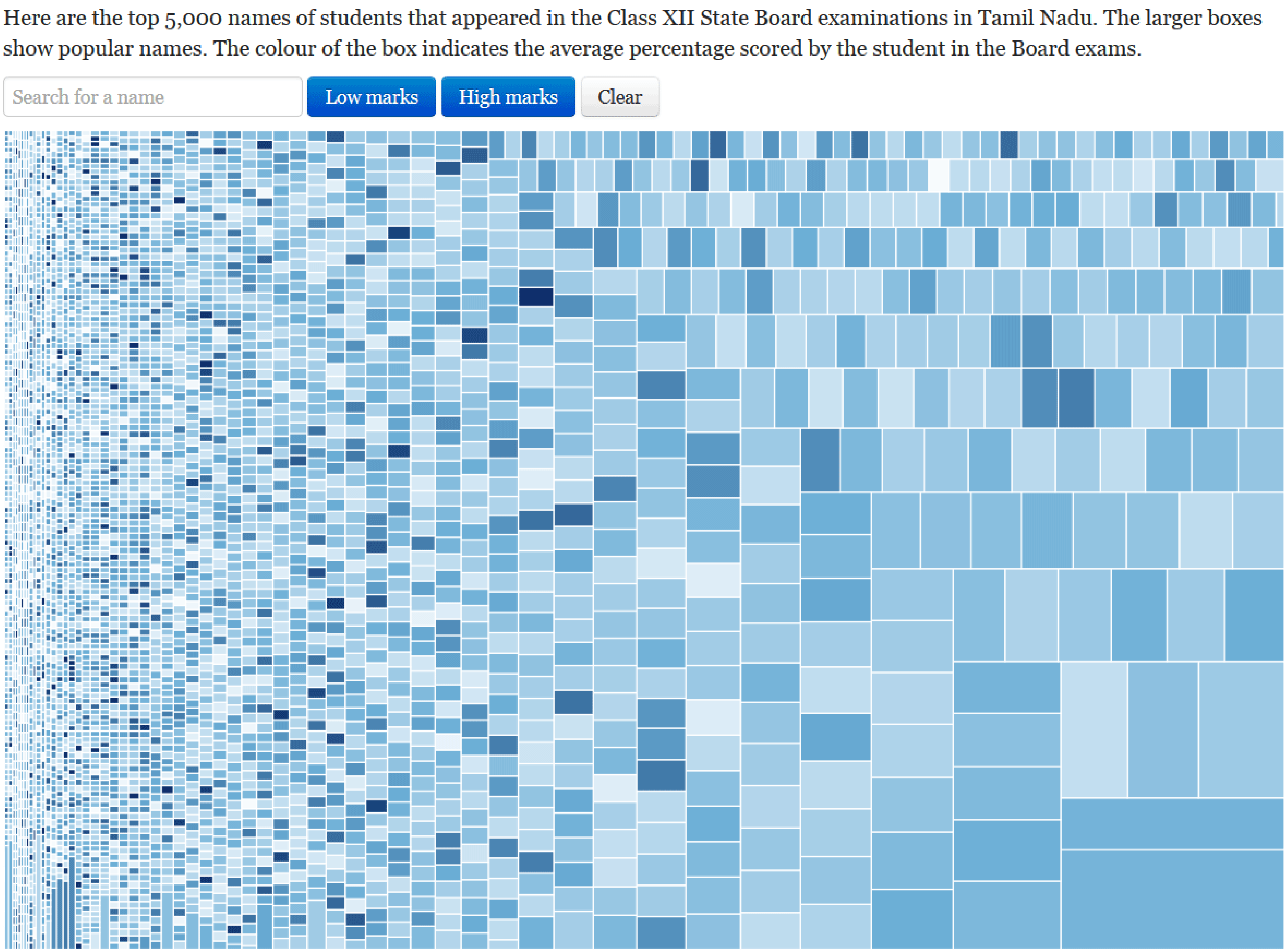
- Standard deviation is how dispersed the data are from their mean.

- Median is the point that divides a distribution in half, meaning half the observations lie above the median and half lie below.

- The mode is the value that appears most often in a data set

- This is the only measure of central tendency that can be applied to non-numeric values.

- While mean is sensitive to outliers, mode and median are not.



Here are the top 5,000 names of students that appeared in the Class XII State Board examinations in Tamil Nadu. The larger boxes show popular names. The colour of the box indicates the average percentage scored by the student in the Board exams.

Descriptive Statistics

- An Index is a descriptive statistic made up of other descriptive statistics.
- The advantage of any index is that it consolidates lots of complex information into a single number.
- We can then rank things that otherwise defy simple comparison.
- The disadvantage of any index is that it consolidates lots of complex information into a single number.

Descriptive Statistics

- Precision reflects the exactitude with which we can express something.

- Accuracy is a measure of whether a figure is broadly consistent with the truth - hence the danger of confusion precision with accuracy.

- If an answer is accurate, then more precision is usually better. But no amount of precision can make up for inaccuracy.

- From a standpoint of accuracy, the median versus mean question revolves around whether the outliers in a distribution distort what is being described or are instead an important part of the message.

Biases in data & analysis

- Selection bias
  (self-selection, not representative, anecdotal etc.)
- Publication bias
  (positive findings are more likely to be published than negative findings, video games-colon cancer, something does not cause cancer is not interesting)
- Recall bias
  (memory is fallible, 1993 Harvard longitudinal cancer study)
- Survivorship bias
  (school drop outs, mutual finds, older concert halls)
- Healthy user bias
  (people who take vitamins regularly are more healthy than those who do not, parable of purple pyjamas, Van Halen-no brown M&Ms)

Displaying Scientific Evidence for Making Valid Decisions

Northwestern University CLIMB

(Collaborative Learning and Integrated Mentoring in the Biosciences)

Displaying Scientific Evidence for Making Valid Decisions: Lessons from Two Case Studies