Data Science Questions

Data Analysis

Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.

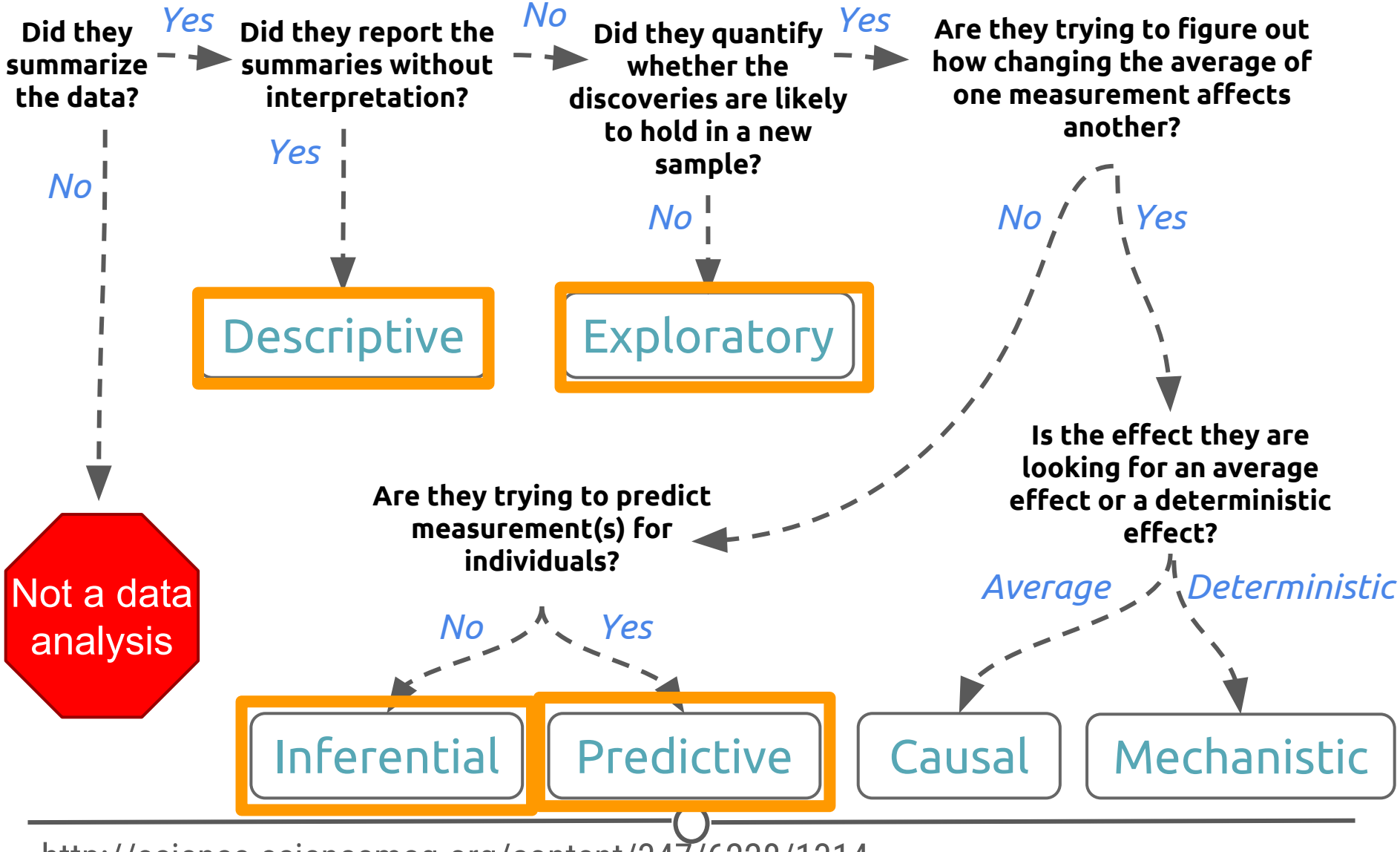
https://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-that-is-a-big-problem-for-data-scientists/

1. Define the question you want to ask the data
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

https://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-that-is-a-big-problem-for-data-scientists/

1. Define the question you want to ask the data
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

https://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-that-is-a-big-problem-for-data-scientists/



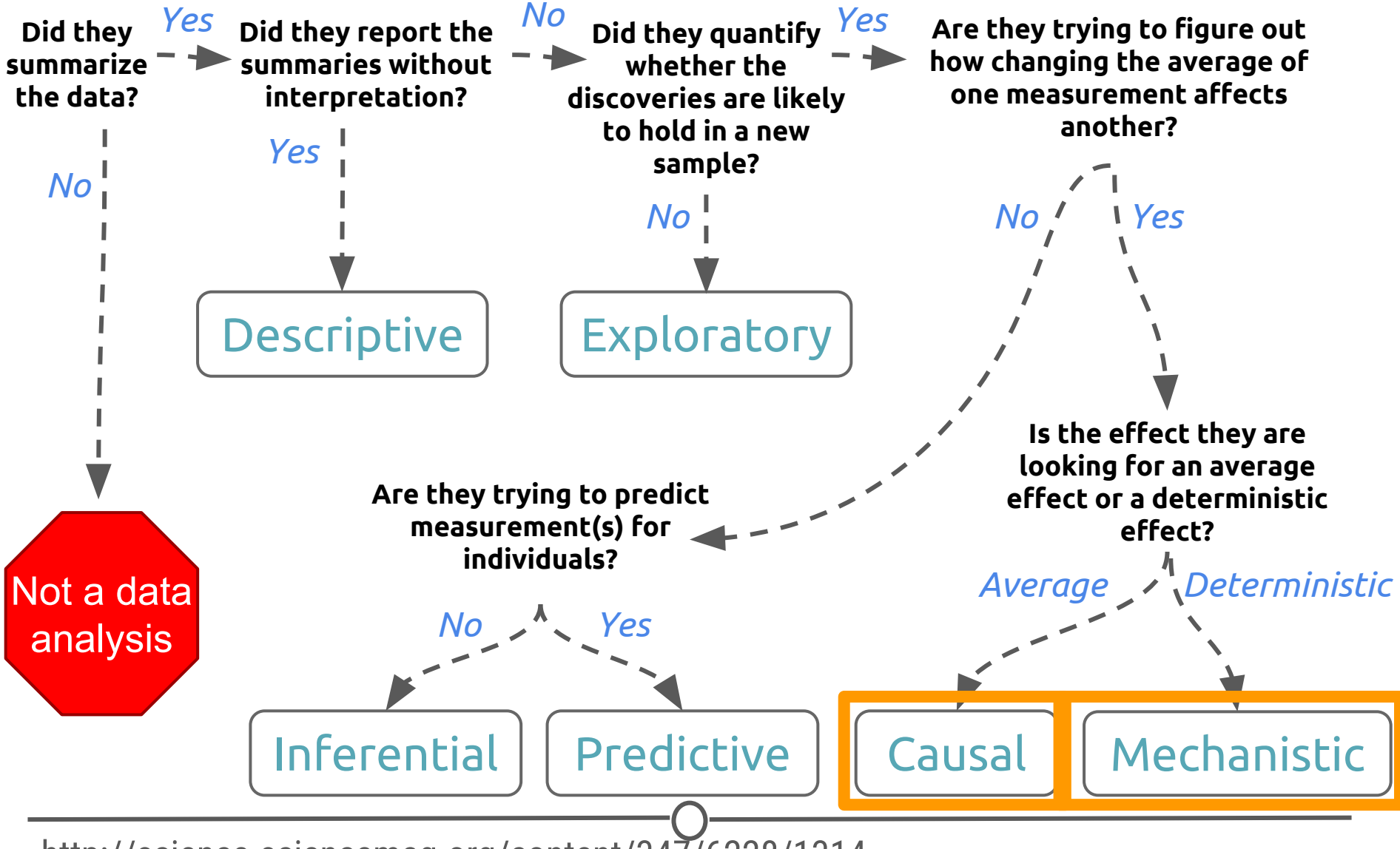
http://science.sciencemag.org/content/347/6228/1314

Descriptive: The goal of descriptive data science questions is to understand the components of a data set, describe what they are, and explain that description to others who might want to understand the data. This is the simplest type of data analysis.

Exploratory: The goal of exploratory data science questions is to find unknown relationships between the different variables you have measured in your data set. Exploratory analysis is open ended and designed to find expected or unexpected relationships between different measurements.

Inferential: The goal of inferential data science questions is to is to use a small sample of data to say something about what would happen if we collected more data. Inferential questions come up because we want to understand the relationships between different variables but it is too expensive or difficult to collect data on every person or object.

Predictive: The goal of predictive data science question is to use data from a large collection to predict values for new individuals. This might be predicting what will happen in the future or predicting characteristics that are difficult to measure. Predictive data science is sometimes called machine learning.

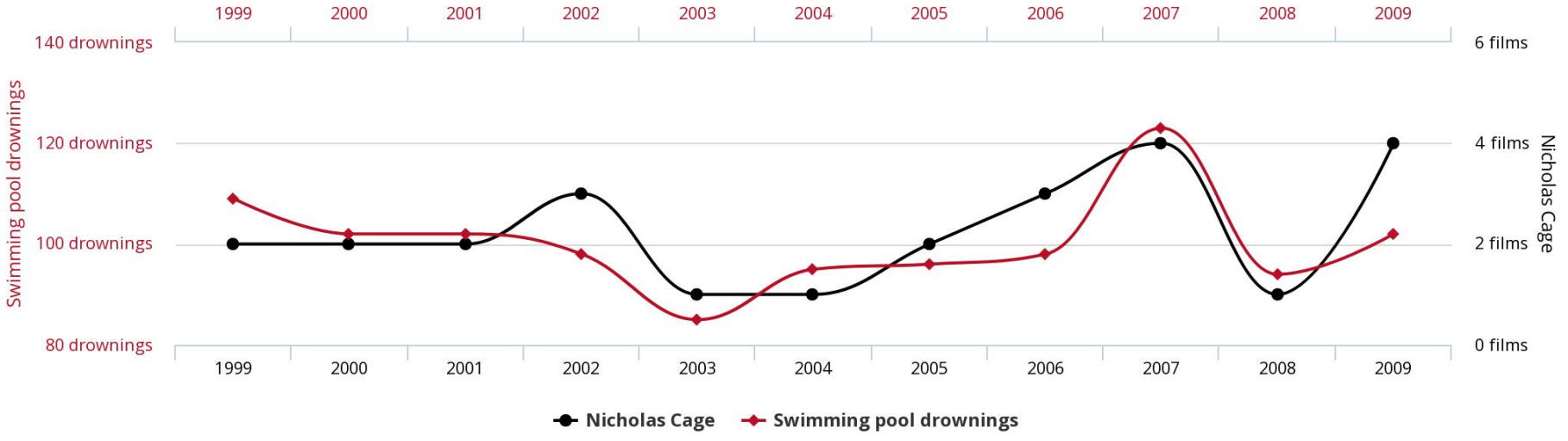


http://science.sciencemag.org/content/347/6228/1314

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



tylervigen.com

- **Problem:** Detecting whether credit card charges are fraudulent.
- **Data science question:** Can we use the time of the charge, the location of the charge, and the price of the charge to predict whether that charge is fraudulent or not?
- **Type of analysis:** Predictive analysis

Image: Freepik.com

- **Problem:** Understanding whether users are nice or mean on Youtube
- **Data science question:** Are the words that people use in their comments more frequently positive words (great, awesome, nice, useful) or negative words (bad, stupid, lame, awful)?
- **Type of analysis:** Descriptive analysis

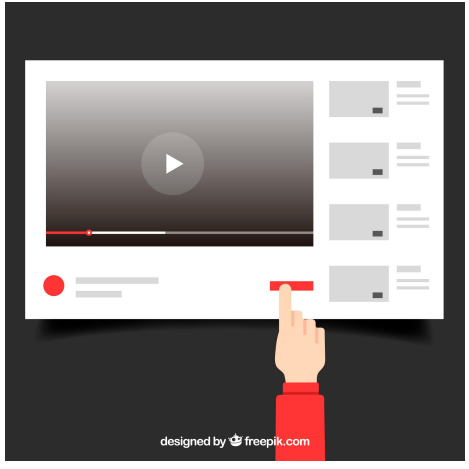


Image: Freepik.com

- **Problem:** Does Sesame Street and kids brain development
- **Data science question:** Is there a relationship between watching Sesame Street and test scores among children?
- **Type of analysis:** Inferential analysis

Image: Wikimedia Commons, Walter Lim

Built with blogdown and Hugo. Theme Blackburn.

What statistics should do about big data: problem forward not solution backward

Jeff Leek 2013/05/29

There has been a lot of discussion among statisticians about big data and what statistics should do to get involved. Recently Steve M. and Larry W. took up the same issue on their blog. I have been thinking about this for a while, since I work in genomics, which almost always comes with "big data". It is also one area of big data where statistics and statisticians have played a huge role.

A question that naturally arises is, "why have statisticians been so successful in genomics?" I think a major reason is the phrase I borrowed from Brian C. (who may have borrowed it from Ron B.)

> problem first, not solution backward

One of the reasons that "big data" is even a term is that there is that data are less expensive than they were a few years ago. One example is the dramatic drop in the price of DNA-sequencing. But there are many many more examples. The quantified self movement and Fitbits, Google Books, social network data from Twitter, etc. are all areas where data that cost us a huge amount to collect 10 years ago can now be collected and stored very cheaply.

As statisticians we look for generalizable principles; I would say that you have to zoom pretty far out to generalize from social networks to genomics but here are two:

https://simplystatistics.org/2013/05/29/what-statistics-should-do-about-big-data-problem-forward-not-solution-backward/