What is the question?

By **Jeff Leek*** *and* **Roger D. Peng**

Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA.

***Corresponding author. E-mail: jleek@jhsph.edu, jtleek@gmail.com**

Mistaking the type of question being considered is the most common error in data analysis.

Over the past 2 years, increased focus on statistical analysis brought on by the era of big data has pushed the issue of reproducibility out of the pages of academic journals and into the popular consciousness (*1*). Just weeks ago, a paper about the relationship between tissue-specific cancer incidence and stem cell divisions (*2*) was widely misreported because of misunderstandings about the primary statistical argument in the paper (*3*). Public pressure has contributed to the massive recent adoption of reproducible research, with corresponding improvements in reproducibility. But an analysis can be fully reproducible and still be wrong. Even the most spectacularly irreproducible analyses—like those underlying the ongoing lawsuits (*4*) over failed genomic signatures for chemotherapy assignment (*5*)—are ultimately reproducible (*6*). Once an analysis is reproducible, the key question we want to answer is, "Is this data analysis correct?" We have found that the most frequent failure in data analysis is mistaking the type of question being considered.

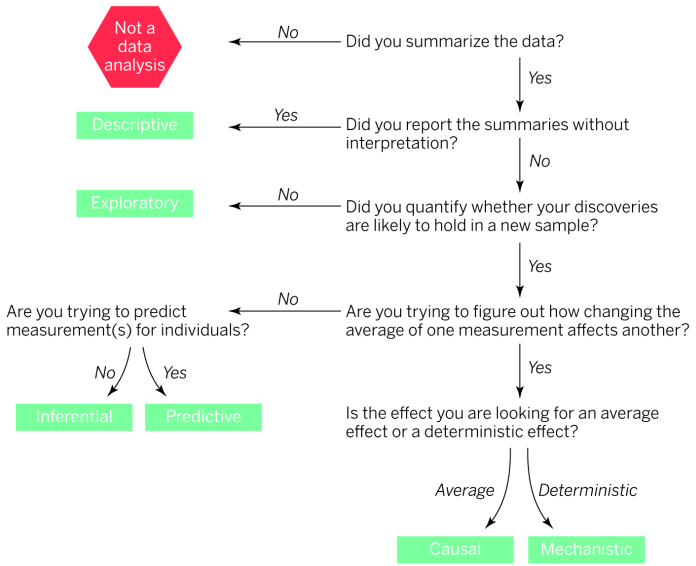
Any specific data analysis can be broadly classified into one of six types (see the figure). The least challenging of these is a descriptive data analysis, which seeks to summarize the measurements in a single data set without further interpretation. An example is the United States Census, which aims to describe how many people live in different parts of the United States, leaving the interpretation and use of these counts to Congress and the public.

An exploratory data analysis builds on a descriptive analysis by searching for discoveries, trends, correlations, or relationships between the measurements to generate ideas or hypotheses. The four-star planetary system Tatooine was discovered when amateur astronomers explored public astronomical data from the Kepler telescope (*7*). An exploratory analysis like this seeks to make discoveries, but can rarely confirm those discoveries. Follow-up studies and additional data were needed to confirm the existence of Tatooine (*8*).

An inferential data analysis quantifies whether an observed pattern will likely hold beyond the data set in hand. This is the most common statistical analysis in the formal scientific literature. An example is a study of whether air pollution correlates with life expectancy at the state level in the United States (*9*). In nonrandomized experiments, it is usually only possible to determine the existence of a relationship between two measurements, but not the underlying mechanism or the reason for it.

Going beyond an inferential data analysis, which quantifies the relationships at population scale, a predictive data analysis uses a subset of measurements (the features) to predict another measurement (the outcome) on a single person or unit. Web sites like FiveThirtyEight.com use polling data to predict how people will vote in an election. Predictive data analyses only show that you can predict one measurement from another; they do not necessarily explain why that choice of prediction works.

Data analysis flowchart



A causal data analysis seeks to find out what happens to one measurement on average if you make another measurement change. Such an analysis identifies both the magnitude and direction of relationships between variables on average. For example, decades of data show a clear causal relationship between smoking and cancer (*10*). If you smoke, it is certain that your risk of cancer will increase. The causal effect is real, but it affects your average risk.

Finally, a mechanistic data analysis seeks to show that changing one measurement always and exclusively leads to a specific, deterministic behavior in another. For example, data analysis has shown how wing design changes air flow over a wing, leading to decreased drag. Outside of engineering, mechanistic data analysis is extremely challenging and rarely achievable.

Mistakes in the type of data analysis and therefore the conclusions that can be drawn from data are made regularly. In the last 6 months, we have seen inferential analyses of the relationship between cellphones and brain cancer inter-

preted as causal (*11*) or the exploratory analysis of Google search terms related to flu outbreaks interpreted as a predictive analysis (*12*). The mistake is so common that it has been codified in standard phrases (see the table).

Common mistakes

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	"correlation does not imply causation"
Exploratory	Inferential	"data dredging"
Exploratory	Predictive	"overfitting"
Descriptive	Inferential	"n of 1 analysis"

Determining which question is being asked can be even more complicated when multiple analyses are performed in the same study or on the same data set. A key danger is causal creep—for example, when a randomized trial is used to infer causation for a primary analysis and data from secondary analyses are given the same weight. To accurately represent a data analysis, each step in the analysis should be labeled according to its original intent.

Confusion between data analytic question types is central to the ongoing replication crisis, misconstrued press releases describing scientific results, and the controversial claim that most published research findings are false (*13*, *14*). The solution is to ensure that data analytic education is a key component of research training. The most important step in that direction is to know the question.

REFERENCES AND NOTES

1. "How science goes wrong." *The Economist*, 19 October 2013; see www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong.
2. C. Tomasetti, B. Vogelstein, *Science* **347**, 78 (2015). Medline doi:10.1126/science.1260825
3. See www.bbc.com/news/magazine-30786970.
4. Duke's Legal Stance: We Did No Harm, *The Cancer Letter Publications* (2015); see www.cancerletter.com/articles/20150123_2.
5. A. Potti *et al.*, *Nat. Med.* **12**, 1294 (2006). Medline doi:10.1038/nm1491
6. K. A. Baggerly, K. R. Coombes, *Ann. Appl. Stat.* **3**, 1309 (2009). doi:10.1214/09-AOAS291
7. "Planet with four stars discovered by citizen astronomers," *Wired UK* (2012); see www.wired.co.uk/news/archive/2012-10/15/four-starred-planet.
8. M. E. Schwamb *et al.*, *http://arxiv.org/abs/1210.3612* (2013).
9. A. W. Correia *et al.*, *Epidemiology* **24**, 23 (2013). Medline doi:10.1097/EDE.0b013e3182770237
10. O. A. Panagiotou *et al.*, *Cancer Res.* **74**, 2157 (2014).
11. E. Oster, Cellphones Do Not Give You Brain Cancer, *FiveThirtyEight* (2015); see http://fivethirtyeight.com/features/cellphones-do-not-give-you-brain-cancer/.
12. D. M. Lazer, R. Kennedy, G. King, A. Vespignani, The Parable of Google Flu: Traps in Big Data Analysis (2014); see http://dash.harvard.edu/handle/1/12016836.
13. L. R. Jager, J. T. Leek, *Biostatistics* **15**, 1 (2014). Medline doi:10.1093/biostatistics/kxt007
14. A. Gelman, K. O'Rourke, *Biostatistics* **15**, 18, discussion 39 (2014). Medline doi:10.1093/biostatistics/kxt034

Published online 26 February 2015
10.1126/science.aaa6146