Many of these lines were found to be of other carbocations, such as CH_3^+, $C_2H_3^+$, $C_2H_2^+$, and CH_2^+ [for a review, see (*8*)]. Each time I proudly reported these discoveries, Olah responded, "impressive, but what about CH_5^+?" After "weeding out" those thousands of understood spectral lines, the remaining messy spectrum was undecipherable, and the 900 lines of CH_5^+ were reported without assignment (*1*). Even purely empirical attempts at finding some regularity of the spectrum were not successful.

Asvany *et al.* have been able to determine the energy separation between several pairs of lowest levels using the action spectroscopy invented by Schlemmer and Gerlich (*9*). The proton affinity of CH_4 (5.72 eV) is slightly greater than that of CO_2 (5.68 eV) so the reaction $CH_5^+ + CO_2 \rightarrow CH_4 + CO_2H^+$ is endothermic. Addition of a resonant 3.3-μm (0.37 eV) laser photon makes this reaction exothermic. Thus, they can do spectroscopy by counting CO_2H^+ ions rather than photons. This paradigm shift from photon- to ion-counting spectroscopy has increased the sensitivity—instead of needing 10^{13} ions, 10^3 CH_5^+ suffice. Also, trapped ions can be cooled to cryogenic temperature, which leads to a 100 times increase in accuracy.

"As in Olah's chemistry on Earth, CH_5^+ is pivotal for producing hydrocarbons in space... I anticipate that this enfant terrible will be caught in interstellar space far ahead of its theoretical understanding..."

The 2897 lines observed by Asvany *et al.* at 10 K demonstrate the complexity of the CH_5^+ spectrum. In contrast, CH_4 at 10 K has only four rotational levels with quantum number $J < 3$ populated and only 10 transitions can be observed. The 300 times increase in spectral density from CH_4 to CH_5^+ is caused by proton scrambling and inversion motion. Rotational assignments could be made that differ from those they reported, but these are minor details—the CoDiff values are correct and the key to advancing our understanding.

In spite of the complexity, each quantum level can be specified by using the total proton spin quantum number and the parity (*10*). The scrambling of the five protons with spin 1/2 produces a total nuclear spin angular momentum I according to the formula

$$[D_{1/2}]^5 = D_{5/2} + 4D_{3/2} + 5D_{1/2}$$

This formula means that each of the levels of CH_5^+ have $I = 5/2$ (A_1), 3/2 (G_1), or 1/2 (H_1), and the number of levels are in the ratio of 1:4:5 and the CoDiffs 1:16:25. Each level also has a definite parity of + or – and their numbers are equal. The CoDiffs reported by Asvany *et al.* are for levels with $I = 3/2$ and 1/2 and the same parity. The next step will be to find CoDiffs with different parity, which the authors note could be tackled by applying their method for far-infrared spectroscopy.

The results by Asvany *et al.* put the experiment far ahead of the theory. To date, Wang and Carrington's computation (*11*), based on the potential energy surface (PES) of Jin *et al.* (*12*), seems to be the only frontal attack to this problem, but it does not include rotation. A brute-force variational calculation of the five protons with an accurate PES may be the way to solve this problem. Such treatment has been successful for H_3^+, but the formalism and computation will be much more demanding for a five-proton system.

As in Olah's chemistry on Earth, CH_5^+ is pivotal for producing hydrocarbons in space. The lines list by Asvany *et al.* suffices for detecting interstellar CH_5^+, but we badly need strongest lines for $I = 5/2$ (A_1). The classical CH_3^+ ion is yet to be detected, so detection of the nonclassical CH_5^+ will be difficult but worth a try. Once the far-infrared transitions are observed, including $I = 5/2$ levels, more sensitive observational techniques can be used. I anticipate that this enfant terrible will be caught in interstellar space far ahead of its theoretical understanding, which will take at least a few more decades. ■

Department of Chemistry and Department of Astronomy and Astrophysics, The Enrico Fermi Institute, University of Chicago, Chicago, IL 60637, USA. E-mail: t-oka@uchicago.edu

REFERENCES

1. E. T. White, J. Tang, T. Oka, *Science* **284**, 135 (1999).
2. P. R. Schreiner, S.-J. Kim, H. F. Schaefer, III, P. von Ragué Schleyer, *J. Chem. Phys.* **99**, 3716 (1993).
3. H. Müller, W. Kutzelnigg, J. Noga, W. Klopper, *J. Chem. Phys.* **106**, 1863 (1997).
4. O. Asvany, K. M. T. Yamada, S. Brünken, A. Potapov, S. Schlemmer, *Science* **347**, 1346 (2015).
5. T. Oka, *Phys. Rev. Lett.* **45**, 531 (1980).
6. G. A. Olah, *Angew. Chem. Int. Ed. Engl.* **34**, 1393 (1995).
7. C. S. Gudeman, M. H. Begemann, J. Pfaff, R. J. Saykally, *Phys. Rev. Lett.* **50**, 727 (1983).
8. T. Oka, *J. Phys. Chem. A* **117**, 9308 (2013).
9. S. Schlemmer, T. Kuhn, E. Lescop, D. Gerlich, *Int. J. Mass Spectrom.* **185–187**, 589 (1999).
10. P. R. Bunker, B. Ostojić, S. Yurchenko, *J. Mol. Struct.* **695–696**, 253 (2004).
11. X.-G. Wang, T. Carrington Jr., *J. Chem. Phys.* **129**, 234102 (2008).
12. Z. Jin, B. J. Braams, J. M. Bowman, *J. Phys. Chem. A* **110**, 1569 (2006).

10.1126/science.aaa6935

STATISTICS

What is the question?

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

By **Jeffery T. Leek** *and* **Roger D. Peng**

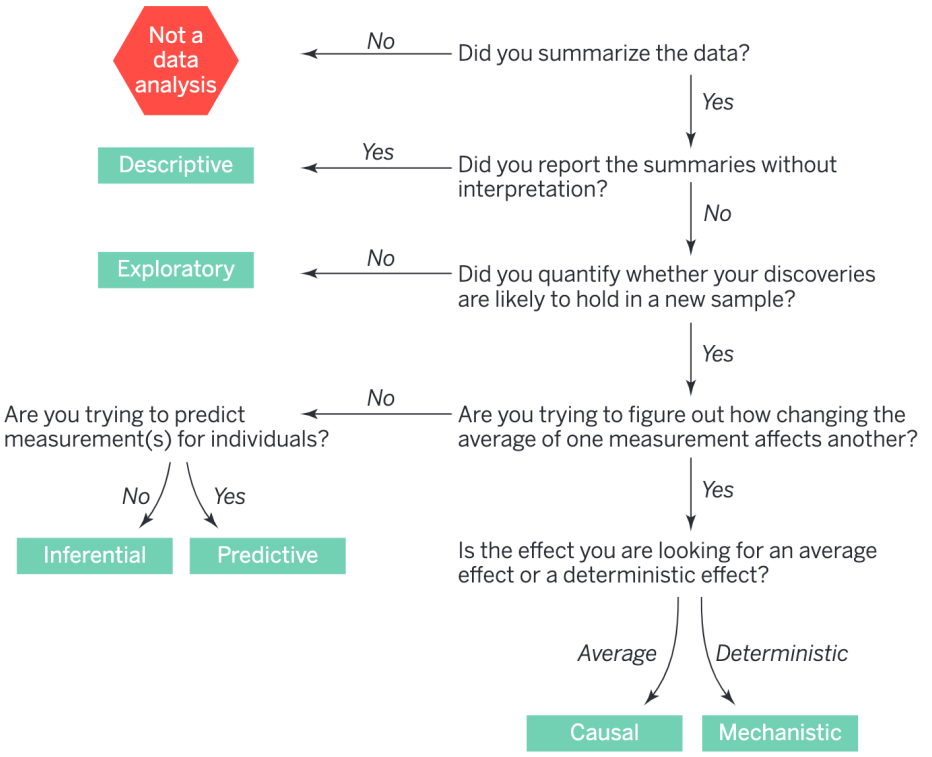
Over the past 2 years, increased focus on statistical analysis brought on by the era of big data has pushed the issue of reproducibility out of the pages of academic journals and into the popular consciousness (*1*). Just weeks ago, a paper about the relationship between tissue-specific cancer incidence 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 tools, 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

Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA. E-mail: jleek@jhsph.edu, jtleek@gmail.com

Data analysis flowchart



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.

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 interpreted 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).

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. ■

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"

REFERENCES

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).
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).
6. K. A. Baggerly, K. R. Coombes, *Ann. Appl. Stat.* **3**, 1309 (2009).
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).
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).
14. A. Gelman, K. O'Rourke, *Biostatistics* **15**, 18 (2014).

Published online 26 February 2015; 10.1126/science.aaa6146

Published by AAAS