A reprint from

American Scientist
the magazine of Sigma Xi, The Scientific Research Society

To Throw Away Data: Plagiarism as a Statistical Crime

Andrew Gelman and Thomas Basbøll

"The distortion of a text," says Freud in Moses and Monotheism, *"is not unlike a murder. The difficulty lies not in the execution of the deed but in doing away with the traces."*
—James Wood

Much has been written on the ethics of plagiarism. One aspect that has received less notice is plagiarism's role in corrupting our ability to learn from data: We propose that plagiarism is a *statistical* crime. It involves the hiding of important information regarding the source and context of the copied work in its original form. Such information can dramatically alter the statistical inferences made about the work.

In statistics, throwing away data is a no-no. From a classical perspective, inferences are determined by the sampling process: point estimates, confidence intervals and hypothesis tests all require knowledge of (or assumptions about) the probability distribution of the observed data. In a Bayesian analysis, it is necessary to include in the model all variables that are relevant to the data-collection process. In either case, we are generally led to faulty inferences if we are given data from urn A and told they came from urn B.

Andrew Gelman is a professor in the departments of statistics and political science at Columbia University, New York, and the author of Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do *(Princeton University Press, 2008). Thomas Basbøll is an independent writing coach and an external lecturer in the department of management, politics and philosophy at the Copenhagen Business School. Address for Gelman: Columbia University, 1016 Social Work Building, New York, NY 10027. E-mail: gelman@stat.columbia.edu*

Whether data are numerical or narrative, removing them from their context represents an act of plagiarism

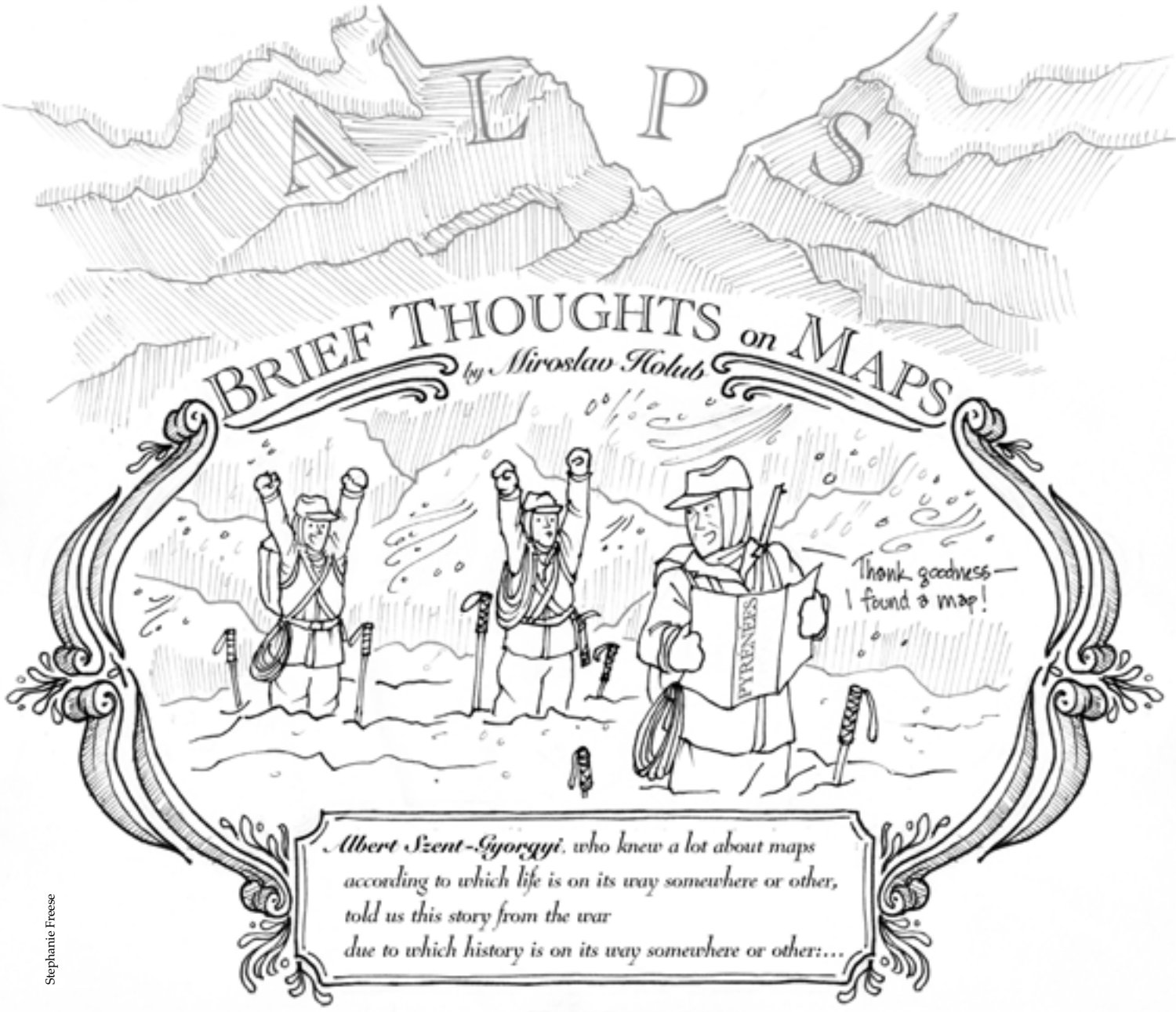
A statistical perspective on plagiarism might seem relevant only to cases in which raw data are unceremoniously and secretively transferred from one urn to another. But statistical consequences also result from plagiarism of a very different kind of material: stories. To underestimate the importance of contextual information, even when it does not concern numbers, is dangerous.

Perhaps the most prominent statistician to have repeatedly published material written by others without attribution is Edward Wegman, formerly of the Office of Naval Research and currently a professor at George Mason University. The case is especially interesting because Wegman has a distinguished record of public service and scholarship (he received the Founders Award in 2002 from the American Statistical Association) and because one of the plagiarized documents was part of a report on climate change delivered to the U.S. Congress. The ethical dimensions of this copying seem clear enough: By taking others' work without giving credit—even copying from Wikipedia at one point (see the appendix to this essay at *American Scientist*'s website)—Wegman and his research team were implicitly claiming expertise on subjects in which they were not experts. Wegman continues to deny having plagiarized, even in the face of direct evidence that several of his publications (on topics ranging from network analysis to color vision) include unattributed material previously published by others.

We shall avoid speculating about the motives for plagiarism here. Generally, however, the ethical dilemma seems to be analogous to the person who robs a store to feed his or her family, or the politician who lies to achieve a larger political goal. In all of these cases, the behavior in question is generally recognized to be unethical, so if the broader context in which the action takes place is deemed ethical, it can only be thus because the unethical action serves some larger, more important goal. In Wegman's case, no such argument about a larger context has been made (perhaps because that would require admitting the ethical violation in the first place).

The Wegman case came to public notice after the Canadian blog *Deep Climate* found the first few pages of material in the report to be plagiarized from a book by Ray Bradley, one of the authors whose work was attacked in that document. The blog post stirred others to study this and other documents written by Wegman and his students, at which point additional incidents of copying without attribution turned up. In 2011, a related article by Wegman and a collaborator in the journal *Computational Science and Data Analysis* was formally retracted by the publisher on grounds of plagiarism.

Despite the human and political drama of the Wegman case, it may not appear immediately interesting from the standpoint of statistics. Perhaps counterintuitively, a purely qualitative example reveals why this appearance is wrong.



Stephanie Freese

A poem by Miroslav Holub that appeared in the *Times Literary Supplement* in 1977 tells the story of a Hungarian reconaissance unit caught in a snowstorm in the Alps. Holub recounts how the lieutenant who sent the unit out feared for their lives—but that the unit returned after three days, saying that one of their number had a map. The map, however, turns out to be of the Pyrénées, not the Alps. The story has been widely retold in the field of organization studies. Whether and how its source is cited, the authors argue, is a matter of statistical concern.

Snowstorm, Map, Conundrum

An anecdote that has been widely circulated in the organization studies literature goes something like this: A group of soldiers are sent out by their leader and get lost in a snowstorm in the Alps. After discovering that one of their number has a map, they regain their confidence, wait out the storm and return to camp. Only afterward do they realize that the map was not of the Alps but of the Pyrénées.

This story has made the rounds in management circles, often accompanied by the slogan, "When you are lost, any old map will do." It was even retold by noted psychologist Daniel Kahneman at the 2009 Digital Live Design conference as part of an account of the importance of confidence. Kahneman attributed the story to the "famous organizational psychologist Karl Weick." Weick, like Wegman, is an award-winning and highly regarded scholar in his field, and he is the commonly cited source for the anecdote in the organization studies literature. But, as Kahneman noted in his talk, some irregularities in Weick's referencing (or lack thereof) have emerged.

In 2006, one of us (Basbøll), and a Ph.D. student in his department, Henrik Graham, published a paper showing that Weick had simply transcribed the story from a poem by Miroslav Holub that had been published in 1977 in the *Times Literary Supplement*. The text has minor changes but is nearly identical to Holub's—without the line breaks, of course. (See the online appendix to

this essay.) In his earliest uses of the anecdote, Weick provided no reference to Holub whatsoever, despite the fact that his account was a nearly verbatim reproduction of the poem. In later versions, he mentioned Holub's poem but continued to represent the story as his own prose, without enclosing it in quotation marks.

Importantly, Weick also began to alter Holub's framing of the story. Like Holub, he invoked Albert Szent-Gyorgyi, the Nobel Prize–winning physiologist, as the original source of the story (though he did not clearly cite Holub as the source for this source). Holub described the anecdote as a "story from the war," whereas Weick repeatedly called it "an incident that happened during military maneuvers in Switzerland." With this phrasing, not only did he conceal the nature of his evidence from his readers (it is a poem with a unique author, not a story recounted aloud or included in some unspecified report), he also exaggerated the veracity of the account (and gave the war story an implausible Swiss setting, perhaps by associating the mention of the Alps with Switzerland).

The article set off a back-and-forth of publications. The journal that published Basbøll and Graham's 2006 article, *ephemera,* printed a response from Weick in the same issue. In it, he dismissed the charge. In 2010, Basbøll published a response to Weick, along with further examples of plagiarism, which Weick again dismissed. In 2012 Basbøll published a rhetorical analysis of the exchanges so far.

Weick claimed that by the time he realized the anecdote had relevance to his work, he had forgotten where he first encountered it, and that he "reconstructed the story as best [he] could." It seems unlikely that a scholar would add to his own writing a nearly word-for-word copy of a text whose citation he did not have—and this in the era before computer copy-and-paste. Beyond this, Weick's reaction when the news came out also gives us reason to doubt his account. Instead of being embarrassed and bending over backward to add a clear, apologetic citation in subsequent appearances of the material, he seemed all too eager to explain the event away.

Gelman had never heard of any of the people involved in these incidents before Basbøll drew his attention to the case of plagiarism. What brought us together was a shared frustration with an especially slippery aspect of the case, and others like it: the denial or avoidance of the topic by colleagues of the offenders. Weick is influential in his field, known for his counterintuitive management advice. Often, when people who attain such stature misbehave, others find it hard to believe or don't want to hear about it. The assumption, perhaps, is that any misbehavior was for the greater good.

In the wake of the paper's publication in *ephemera,* Basbøll and Graham were mocked by organizational strategy professors Teppo Felin and Omar Lizardo (the latter referred to them as "what's his name and watchumacallit") on the *orgtheory* blog. When Basbøll tried to mention Weick's plagiarism on the online correspondence site of the *Journal of Management Studies,* he was rejected on the grounds that Weick might sue the journal. And the American Statistical Society, which presented its Founders Award to Wegman in 2002, has not to our knowledge commented publicly on the issue.

Learning that part of a corpus of work is plagiarized can degrade one's trust in the rest of the work. This is not just a moral or psychological argument of the sort that one might legitimately use against a scientist known to have fabricated or misrepresented data, such as Diederik Stapel or Marc Hauser—if the guy cheated with data in one place, you can't trust his other statements either. Indeed, Basbøll found that the first four pages of one of Weick's most widely cited books, *Sensemaking in Organizations* (1995), reproduce the work of several other scholars without adequate attribution. The book also includes an instance of the Holub plagiarism.

But we are saying something more: If Weick represented a story recounted in a poem as if it were a historical event, that casts doubt on his rules of evidence. It's not that an unsourced anecdote has more authority than a published poem. Rather, obscuring the source makes the story free-floating, immune from any detail-based examination. Meanwhile, Weick's reputation as an original thinker is threatened if it turns out that he was appropriating others' ideas while concealing his debt to them. In a 2004 article in the journal *Organization Studies,* Weick explains his reputation in terms of the "hidden connections" that exist between his own work and that of his precursors. Acknowledging one's precursors is good, but it's better when their names are given and their work recognized as their own.

Similarly, if Wegman, a nonexpert in network analysis, plagiarizes a description of the field (and, as the blogger known as Deep Climate noted, in the process introduces a typo that wrecks one of the mathematical expressions), that casts doubt on any empirical studies he performs using network analysis. Ultimately, such analyses must be evaluated on their own terms—but without the nudge toward acceptance that might come from the knowledge that they were performed by an eminent statistician. In the Weick case, the copier was getting credit for an interesting story, as well as credit for Holub's writing style—indeed, for certain very specific turns of phrase. In addition, by obscuring the source, he became more free to alter its meaning in different tellings.

Some organization theorists, such as Barbara Czarniawska, have argued that the truth or falsity of the original story has no bearing on the reception of Weick's theory. But we disagree. We believe, for example, that Weick's argument would not have been so well received if he had presented the material as the poem it was rather than calling it "an incident that happened during military maneuvers in Switzerland." In a sense, the vaguer attribution, by placing the story in the category of folklore, gives it an implication of broader significance—in the same way it can be disappointing to learn that a purported folk ballad was in fact the product of a forgotten songwriter.

Decoupling Story and Source

To see more clearly how plagiarism is a crime against statistics, we need to examine how it helps to decouple the story from the source. In Weick's case, this distancing allowed him to convey a message that was virtually the opposite of the story's original meaning. Weick first told the story in 1982 when, five years after the appearance of Holub's poem, Robert Swieringa and he published an article in the *Journal of Accounting Research* including a nearly word-for-word transcription of the poem text, but not using quotation marks or acknowledging Holub at all. In a 1987 essay, Weick added a "twist" to the story that had resulted from a conversation with Robert Engel, a Wall Street executive. Engel, he relates, suggested the possibility that the leader who was out with the troops

might have known that the map was false and still used it effectively. Weick concurs with Engel and expounds on the implications as follows:

> What is interesting about Engel's twist to the story is that he has described the basic situation that most leaders face. Followers are often lost and even the leader is not sure where to go. All the leader knows is that the plan or the map he has in front of him is not sufficient by itself to get them out. What he has to do, when faced with this situation, is instill some confidence in people, get them moving in some general direction, and be sure they look closely at what actually happens, so that they learn where they were and get some better idea of where they are and where they want to be.

He goes on to suggest that the key in this kind of situation is to "get people moving." But in Holub's poem—Weick's primary source material—the soldiers' recounting stands in direct opposition to this interpretation: They say that the map "calmed us down" and that they "pitched camp, lasted out the snowstorm."

Making speculations about what might have happened differently in a situation is not an invalid strategy in all settings; it's just a nonempirical one. In this case the line between fact and supposition was blurred so badly that no such distinction could be made. But facts exist that can be adduced to determine whether Engel's supposition was correct. Assuming that any such event actually occurred, then his notion about what happened is either right or wrong. As it turns out, versions of the story that predate Holub's poem appear in reports given by medical researchers Oscar Hechter and Bernard Pullman at scientific symposia in the early 1970s. These versions suggest that the anecdote as told by Szent-Gyorgyi had the troops' immediate leader thinking it was a map of the Alps, too. Those versions rule out Engel's interpretation.

That interpretation may, of course, be more appealing to Wall Street executives. Given the evolution of modern finance since the mid-1980s, the fact that they appear to have thought that "any old map will do" is somewhat disturbing. But Engel's idea was generated in a problematic context, one in which Weick had, in effect, taken ownership of the story and arrogated for himself the right to alter it at will. The act of plagiarism was the first step in a process that unmoored the story from its sources and removed its evidential value.

A Statistical Crime

Returning to the statistical language of probability and likelihood, to falsify the provenance of a story is to imply an incorrect likelihood function and thus to lose inferential validity. (Statistically speaking, systematically excluding data without revealing the exclusion is a misspecification of the model.) As one of us (Basbøll) eventually showed, any telling of the story is a selection from several possible versions of it. By not sourcing it properly, Weick hides the opportunism of his sampling and sets Engel up to propose a convenient (for top management) "truth" about corporate strategy. This is not to say that, had Weick cited Holub appropriately, he would not have ultimately used it to draw lessons about leadership, even ones that executives would find useful. But if he had done so, he would have had to justify his argument, rather than merely retell the story in his own way to suit his purposes.

Scholars in fields ranging from psychology to history to computer science have recognized that stories are part of how people understand the world. As statisticians, we can consider reasoning from stories as a form of approximate inference. From this perspective, statistical principles should provide some approximate guidance about the potential biases and precision of such inferences. One key principle is not to throw away information and, if discarding data is for some reason necessary, to describe as clearly as possible the mechanism by which the relevant information was excluded. Plagiarism violates both these rules and, as such, is a violation of statistical ethics, beyond any other considerations of moral behavior.

Acknowledgment

Parts of this essay are adapted from Gelman's blog, *Statistical Modeling, Causal Inference, and Social Science*, at http://andrewgelman.com.

Bibliography

Basbøll, T. 2010. JMS suppresses scholarly debate. *Research as a Second Language* blog. May 25. http://secondlanguage.blogspot.dk/2010/05/jms-suppresses-scholarly-debate.html

Basbøll, T. 2010. Softly constrained imagination: Plagiarism and misprision in the theory of organizational sensemaking. *Culture and Organization* 16:163–178.

Basbøll, T. 2012. Any old map won't do: Improving the credibility of storytelling in sensemaking scholarship. WMO Working Paper Series, Copenhagen Business School.

Basbøll, T. 2012. Legitimate peripheral irritations. *Journal of Organizational Change Management* 25:220–235.

Basbøll, T., and H. Graham. 2006. Substitutes for strategy research: Notes on the source of Karl Weick's anecdote of the young lieutenant and the map. *ephemera* 6(2):194–204.

Czarniawska, B. 2005. Karl Weick: Concepts, style and reflection. *Sociological Review* 53: 267–278.

Deep Climate. 2011. Wegman and Said 2011: Yet more dubious scholarship in full colour, part 1. *Deep Climate* blog. March 26. http://deepclimate.org/2011/03/26/wegman-and-said-2011-dubious-scholarship-in-full-colour/

Deep Climate. 2011. Said and Wegman 2009: Suboptimal scholarship. *Deep Climate* blog. Oct. 4. http://deepclimate.org/2011/10/04/said-and-wegman-2009-suboptimal-scholarship/

Felin, T. 2006. Charges of plagiarism in org theory. *Orgtheory* blog. July 22. http://orgtheory.wordpress.com/2006/07/22/charges-of-plagiarism-in-org-theory/

Hechter, O. 1972. Reflections on General Membrane Structure: The Conference in Review. *Annals of the New York Academy of Sciences* 195:506–519.

Holub, M. 1977. Brief thoughts on maps. Translated by J. and I. Milner. *Times Literary Supplement*. Issue 3908:118. February 4.

Mallon, T. 1989. *Stolen Words: Forays into the Origins and Ravages of Plagiarism*. New York: Ticknor & Fields.

Pullman, B. 1974. Summary of the chemical aspects of carcinogenesis. In *Chemical Carcinogenesis*. P. O. P. Ts'o and J. A. DiPaolo, eds. New York: Marcel Dekker.

Swieringa, R., and K. E. Weick. 1982. An assessment of laboratory experiments in accounting. *Journal of Accounting Research* 20 (supplement):56–101.

Vergano, D. 2011. Experts claim 2006 climate report plagiarized. *USA Today*. November 22.

Weick, K. E. 1987. Substitutes for strategy. In *The Competitive Challenge: Strategies for Industrial Innovation and Renewal*, ed. D. J. Teece. Cambridge, MA: Ballinger. pp. 222–233.

Weick, K. E. 1995. *Sensemaking in Organizations*. Thousand Oaks, CA: Sage Publications.

Weick, K. E. 2001. *Making Sense of the Organization*. Oxford: Blackwell Publishing.

Weick, K. E. 2004. Mundane poetics: Searching for wisdom in organization studies. *Organization Studies* 25:653–668.

Weick, K. E. 2006. Dear editor: A reply to Basbøll and Graham. *ephemera* 6(2):193.

Weick, K. E. 2010. Comment on "softly constrained imagination." *Culture and Organization* 16:179.

Wood, James. 2009. James Wood writes about the manipulations of Ian McEwan. *London Review of Books* 31(8):14–16.