DOI:10.1145/2160718.2160737

Applied to almost 3,500 articles it reveals computing's (and *Communications*') culture, identity, and evolution.

BY DANIEL S. SOPER AND OFIR TUREL

An *n*-Gram Analysis of *Communications* 2000–2010

GAINING A DEEP, tacit understanding of an institution's identity is a challenge, especially when the institution has a long history, large geographic footprint, or regular turnover among its employees or members. Though the cultural themes and historical concepts that have shaped an institution may be embedded in its archived documents, the volume of this material may be too much for institutional decision makers to grasp. But without a solid, detailed understanding of the institution's identity, how can they expect to make fully informed decisions on behalf of the institution?

Many scientific disciplines suffer from this phenomenon, with the problem especially pronounced in computing.[1] A constant influx of new technologies, buzzwords, and trends produces an environment marked by rapid change, with the resulting instability making it difficult to establish a stable identity.[2,3] As archived institutional artifacts, articles in journals and other media reflect and chronicle the field's evolving identity. Unfortunately, humans are simply unable to digest it all. However, by leveraging a computational method known as *n*-gram analysis, it may be possible for computer scientists and scholars alike to unlock the secrets within these vast collections and gain insight that might otherwise be lost. If the history and identity of computing are encoded on the pages of journals, systematically analyzing them is likely to yield a better understanding of where the field has been and where it might

» key insights

- ***N*-gram analysis is a simple but extremely useful method of extracting knowledge about an institution's culture and identity from its archived historical documents.**
- ***N*-gram analysis can reveal surprising and long-hidden trends that show how an institution has evolved.**
- **Knowledge gained from *n*-gram analyses can substantially improve managerial decision making.**

be headed. After all, "We are what we write, we are what we read, and we are what we make of what we read."[4]

Here, we address the identity problem by prescribing the same medicine for ourselves—technology and algorithms—we often prescribe for others. We present a culturomic analysis[5] of *Communications* showing how natural language processing can be used to quantitatively explore the identity and culture of an institution over time, inspired by the *n*-gram project released in 2010 by Google labs (http://ngrams.googlelabs.com). In natural language processing, an *n*-gram can be viewed as a sequence of words of length *n* extracted from a larger sequence of words.[12] Google's project allows for quantitative study of cultural trends based on combinations of words and phrases (*n*-grams) appearing in a corpus of more than five million books published as early as the 15th century. The central theory behind the project is that when taken together, the words appearing in a collection of books reveal something about human culture at the time the books were written. Analyzing these words computationally makes it possible to study cultural evolution over time.

Applying a similar analytical approach to the articles in *Communications* would allow us to better understand the culture, identity, and evolution of computing. With a view toward portraying its value for institutional-identity data mining, we present several findings that emerged from our *n*-gram analysis of *Communications*. Though our effort here focuses on a single scientific journal, we hope it engenders future studies that evaluate and compare the cultures and identities of a basket of institutions, providing a deeper understanding of their history and evolution over time.

Figure 1. Structural changes in *Communications* from 2000 to 2010.

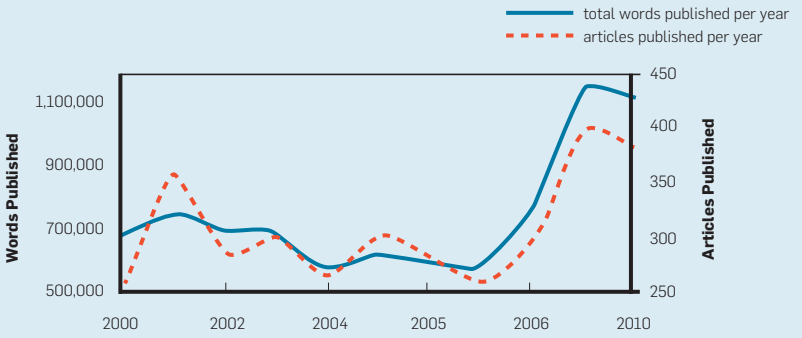


Table 1. Major trends (growth and decline) in terms published in *Communications* from 2000 to 2010.

Growing in Popularity		Declining in Popularity	
Term	**% Change**	**Term**	**% Change**
Google	+18,151%	perceptual	-9,459%
queue	+11,160%	wrapper	-9,459%
cloud	+10,833%	biometrics	-9,295%
VM	+6,439%	CORBA	-9,295%
IT professionals	+5,703%	telemedicine	-8,969%
parity	+5,151%	disintermediation	-7,991%
workload	+5,151%	multimedia	-7,665%
venue	+4,844%	transcription	-7,502%
polynomial time	+4,599%	personalization	-6,425%
DRAM	+4,292%	user profile	-6,197%
test cases	+4,231%	e-commerce	-5,382%
theorem	+4,016%	e-business	-4,281%
OOP	+3,741%	satellites	-4,240%
science and engineering	+3,557%	AOL	-4,158%
emulator	+3,434%	OCR	-4,077%

Method

To appreciate how the identity of *Communications* has evolved, we first constructed a corpus of the complete text of every article it published from 2000 to 2010.[a] We also collected metadata for all these articles, including title, author(s), year published, volume, and issue. In total, our corpus contained 3,367 articles comprising more than 8.1 million words. To put this in perspective, consider that if you were to spend 40 hours per week reading *Communications*, you would need more than four months to read every article published from 2000 to 2010.

With our corpus complete, we next constructed a software system to tokenize, or split the text of each article into a series of *n*-grams. For example, René Descartes' famous phrase "cogito ergo sum"[10] can be subdivided into three *1*-grams (cogito, ergo, and sum), two *2*-grams (cogito ergo, and ergo sum), and one *3*-gram (cogito ergo sum). As this illustrates, the number of *n*-grams that could potentially be extracted from a large corpus of text greatly exceeds the number of words in the corpus itself. This situation has serious scaling and performance implications for a corpus with millions of words, so to avoid them we limited our analysis to include *n*-grams with a maximum length of $n = 4$.

To address the challenges of punctuation, we adopted the same method used by the developers of Google's *n*-gram project for digitized books,[13] treating most punctuation marks as separate words during the *n*-gram construction process. The phrase "Elementary, my dear Watson" would be tokenized as, say, five words:

Elementary , my dear Watson

a Only the text of each article was included in the database; excluded was trailing matter (such as acknowledgements and references).

PHOTOGRAPH BY ALICIA KUBISTA

Notable exceptions to this rule include currency symbols, decimal components of numbers, and apostrophes indicating possessive case. A term like "$5.95" would be treated as a *1*-gram, while "Euler's constant" would be treated as a *2*-gram. For a more general rule for tokenization, developers might consider splitting tokens that contain a special character only if the character is adjacent to whitespace or a linefeed.

We ignored case in the construction of our *n*-gram corpus. Had we retained case sensitivity, a term (such as "computer science") would have been treated as distinct from the term "Computer Science." While ignoring case vastly reduced the potential number of *n*-grams the system might encounter, it also involved a few negative implications for search specificity. Without case sensitivity, the term "IT" (for information technology) would be considered identical to, say, the word "it." Despite this drawback, we concluded that the overall benefit of ignoring case outweighed its cost.

Broadly speaking, our analysis of how *Communications* evolved from 2000 to 2010 was predicated on the idea that the level of importance or relevance of a particular concept is reflected in how often the concept is mentioned over time. We therefore had to compute the frequency with which every *n*-gram in the corpus appeared in *Communications* during each year of the analysis. For example, if the *n*-gram "e-commerce" was mentioned 273 times in 2000 but only 23 times in 2010,[b] we might infer the concept of e-commerce had become less important in *Communications* over time. However, direct frequency comparisons can be deceiving because they do not account for potential growth or decline in the number of words *Communications* published over time. It was therefore necessary for us to calculate relative frequencies for each *n*-gram. We thus divided *n*-gram frequencies for each year by the total number of words appearing in the corpus during that year in order to produce a standardized measure of frequency that would allow valid comparisons between *n*-grams from year to year.[13] The standardized frequency values resulting from this process indicated how often a particular *n*-gram appeared in *Communications* during a particular year relative to the total quantity of text published in it that year. Standardized frequencies are not, of course, the only means *n*-grams can be compared over time. Indeed, other, more sophisticated information-theoretic measures (such as entropy and cross-entropy) can also be used for this purpose.

The result was a vast database containing more than 160 million *n*-grams and their associated years and standardized frequencies. From it we then selected the one million unique *n*-grams exhibiting the most absolute change over time, reasoning that the frequencies of less-interesting

b These were the actual *n*-gram frequencies for "e-commerce" during 2000 and 2010.

n-grams (such as "how" and "the") would remain relatively stable from year to year. Alternatively, a linguistically informed approach to reducing the size of the search space could be done through part-of-speech (POS) tagging, such that only those *n*-grams identified as noun phrases would be added to the dataset. However, state-of-the-art POS taggers are only about 95% accurate, implying that many interesting *n*-grams could have been overlooked had we taken this approach. Nevertheless, POS-based *n*-gram identification remains an option, especially when the corpus to be analyzed is extremely large.

Finally, we constructed a Web-based system to enable us to query, graph, and explore our *Communications* *n*-gram database, plot and analyze multiple *n*-grams simultaneously, and combine related search terms into a single result. For example, the search phrase "cellphone+cellphones, smartphone+smartphones" would produce a graph containing two lines, one representing the combined frequencies of the terms "cellphone" and "cellphones" over time, the other representing the combined frequencies of the terms "smartphone" and "smartphones" over time. To try out our *Communications* *n*-gram tool, see http://www.invivo.co/ngrams/cacm.aspx.

If, in the aggregate, *Communications* reflects what is happening in computing, then perhaps existing industry standards should be refined to more closely approximate real-world practice.

Findings

Though we cannot expect to identify all ways the computing field has evolved in a single article, we do aim to provide a point of embarkation for future research. Beginning with big-picture considerations, we are confident saying the structure and content of *Communications* evolved significantly from 2000 to 2010. An analysis of our metadata revealed several striking, large-scale structural changes from 2000 to 2010. Over that time, *Communications* published an average of 306 articles per year, each containing an average of about 2,400 words. However, these averages obscured underlying trends showing that both the number of articles published per year and the average length of each article grew significantly, especially in more recent years. These trends (see Figure 1) imply *Communications* was providing more value to its readers than in it had previously, since more recent issues contained more articles and words than earlier issues.

Changing Focus

Continuing our investigation, we next extracted the 15 terms that experienced the most growth or decline in popularity in *Communications* from 2000 to 2010 (see Table 1). We hope you find at least a few trends in the table that are unexpected or interesting; indeed finding them is a primary goal of large-scale data mining. For us, we noticed that several of the terms showing the most growth were related to science and technology, while several of the declining terms were related to business and management. But is this observation anecdotal or a broader pattern in *Communications*? To answer, and to show how *n*-gram analyses can be integrated with more traditional analytic techniques, we conducted an interaction analysis comparing the *n*-gram frequencies for terms related to business and management with those related to science and technology. We identified related terms using Thinkmap's Visual Thesaurus software (http://www.visualthesaurus.com), which is specifically designed for this purpose. We then extracted *n*-gram frequencies for the resulting lists of related terms, using these values to conduct our interaction analysis (see Figure 2). As shown in the figure, the average frequency of business- and management-related terms declined steadily from 2000 to 2010, while science- and technology-related terms became more common. Our interaction analysis indicated that the observed disparity was highly significant ($t_{[5052]} = 2.834, p < 0.01$), providing statistical evidence of *Communications*' evolving identity.

Changes in Style

The style of the writing in *Communications* also evolved from 2000 to 2010. Authors seemed to be transitioning from the traditional academic style of writing, adopting instead a less-formal, more personal voice. Evidence of this change can be seen in the increasing use of words that refer directly to an article's author(s) (such as "I" +143% and "we" +137%) and in the increased frequency authors spoke

directly to their readers through such words as "you" (+222%) and "your" (+205%). Interesting to note is while the content of *Communications* became more scientific and technical, it was presented in a way that was less scientific and technical. A possible effect of this change is that the content of *Communications* became more accessible to a wider, more diverse audience. We, too, found ourselves adopting this more personal style when writing this article.

Our *n*-gram analysis also revealed changes in *Communications*' use of gender-related terms from 2000 to 2010. On average, masculine pronouns (such as "he," "his," and "him") appeared 277% more often than feminine pronouns (such as "she," "hers," and "her"). Moreover, the gap widened from 190% in 2000 to more than 290% in 2010. One possible explanation is the gender gap between male and female computing professionals also grew and was wider in 2010 than it was at any time in the previous 25 years.[14]

World Events

When major events occur somewhere in the world, how and to what extent does *Communications* respond? Do such events influence the computing profession? To answer, we conducted an *n*-gram analysis that included three types of world events: a natural disaster (Hurricane Katrina), a terrorist attack (9/11), and a health crisis (2003 SARS outbreak); Figure 3 shows a common pattern with respect to how *Communications* reacted to such events. Specifically, a major world event would first appear in *Communications* shortly after it occured, with discussion of the event—measured by how often it was mentioned—growing quickly for a short time thereafter. This finding indicates *Communications* is not insulated from major world events but rather embraces them and actively contributes to their discussion in the global forum. After a few years, however, *Communications*' interest in a major event would decline sharply. Nevertheless, even after experiencing this precipitous drop, major world events still tend to be mentioned occasionally in *Communications* over the following years.

Systems Development Life Cycle

The systems development life cycle (SDLC) is one of the most ubiquitous and enduring components of the computing profession. With respect to the four principal phases of SDLC—planning, analysis, design, and implementation—industry standards generally recommend that approximately 15% of time and resources budgeted for a systems development project go to planning, 20% to analysis, 35% to design, and 30% to implementation.[9] To what extent, then, does discussion of them in *Communications* mirror the level of interest recommended by industry standards? To answer, we again turned to *n*-gram analysis to compute the average frequency with which each SDLC phase was mentioned in *Communications* from 2000 to 2010. Dividing the value for each phase by the overall sum of the average frequencies yielded the relative average frequencies for *Communications* in Table 2.

If we accept industry standards as canonical, then the values in the table suggest *Communications* underemphasized the SDLC planning and implementation phases while overemphasizing analysis and design. However, *Communications* would seem to agree

Figure 2. *Communications*' changing focus.

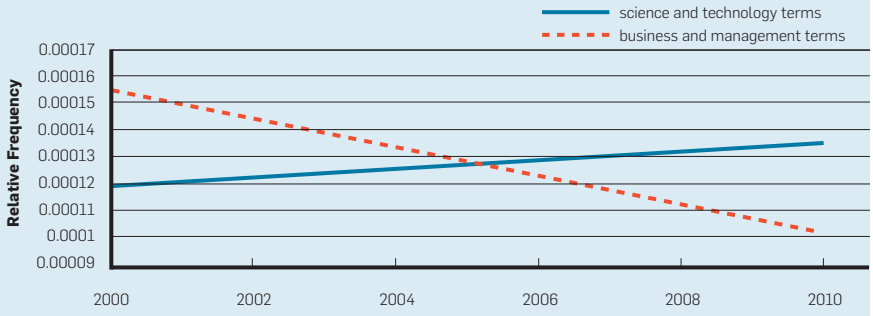


Figure 3. *Communications*' response to major world events.

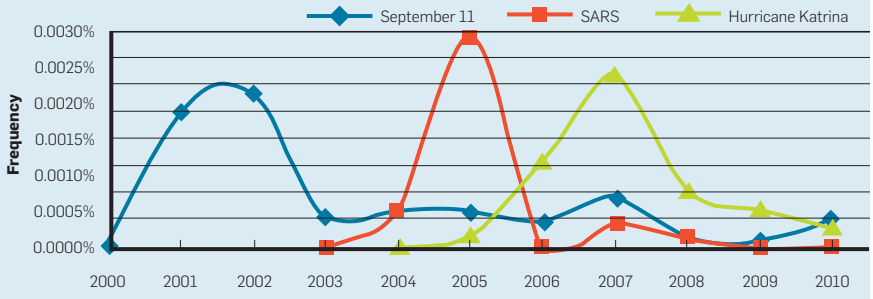


Table 2. Interest in the phases of SDLC in *Communications* compared to industry standards.

	Level of Interest	
SDLC Phase	**Industry Standard**	**Communications**
Planning	15%	8%
Analysis	20%	29%
Design	35%	46%
Implementation	30%	17%

in principle with industry standards that design deserves the most attention and planning the least. The overall discrepancies between these two sources also raise another interesting point: If, in the aggregate, *Communications* reflects what is happening in computing, then perhaps existing industry standards should be refined to more closely approximate real-world practice.

Special Sections

From 2000 to 2010, *Communications* featured special sections that included a number of articles on a specific topic. But did these sections engender long-term interest in the topic being addressed or was the effect more fleeting? To answer, we selected three topics that were the focus of special sections: spyware,[6] digital rights management,[7] and democracy.[8] Our only criterion in selecting them was that they were published closer to 2000 than to 2010, making it easier to identify long-term effects (see Figure 4).

The figure shows special sections generated a spike of interest in the topic during the calendar year the section was published. This interest was sustained for a short time before declining rapidly, eventually returning to a near-zero steady state. Although they can be expected to increase the visibility of a topic in the short-term, special sections did not seem to engender lasting interest in the topics they addressed, at least in *Communications*. Whether this observation holds for other journals is an interesting empirical question but cannot be answered within the scope of this article.

Technology Preferences

If the articles published in *Communications* truly reflect the state of the art in computing, then a *Communications* *n*-gram analysis focused on specific technologies should help reveal differences between the technological preferences of computing professionals and those of the general public. To this end, we compared *Communications* *n*-gram frequencies for different Web browsers, operating systems, and search engines in 2010 against the market shares of the same products among the general public during the same year.[15] The results, which speak to the comparative popularity of different technologies among computing professionals and the general public, are outlined in Figure 5.

The figure indicates that the preferences of computing professionals with respect to Web browsers and operating systems differed markedly from those of the general public. Specifically, there appeared to be more diversity among the preferences of computing professionals regarding the technologies, while usage patterns among the general public were much more homogeneous. One explanation might be that computing professionals simply have a more nuanced understanding of the advantages and disadvantages of the various technology options available to them and are more likely to orient their preferences toward the technologies that serve their needs best. If this is indeed the case, then the search engine results in the figure represent a powerful testament to the perceived superiority of Google's search technology.

Technology Life Cycles

Finally, we wondered whether *Communications*' interest in a particular technology proxies the location of the technology along its product life cycle. To answer, we conducted an *n*-gram analysis of three mass-market commercial Apple products—iPod, iPhone, and iPad—that were arguably at different points in their respective life cycles (see Figure 6).

As shown in the figure, the frequency a particular product appeared in *Communications* spoke to the location of the technology along its own unique trajectory. For example, interest in the iPod in *Communications* grew sharply from 2004 to 2005, corresponding to a 500% increase in sales during that pe-

Figure 4. Effect of special sections on long-term topic interest.

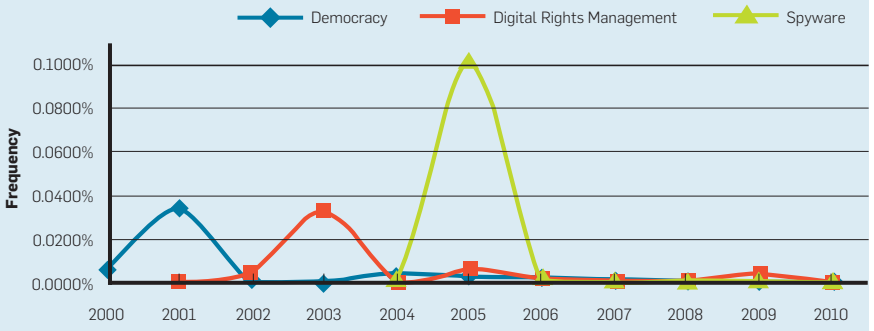


Figure 5. Technology preferences compared.

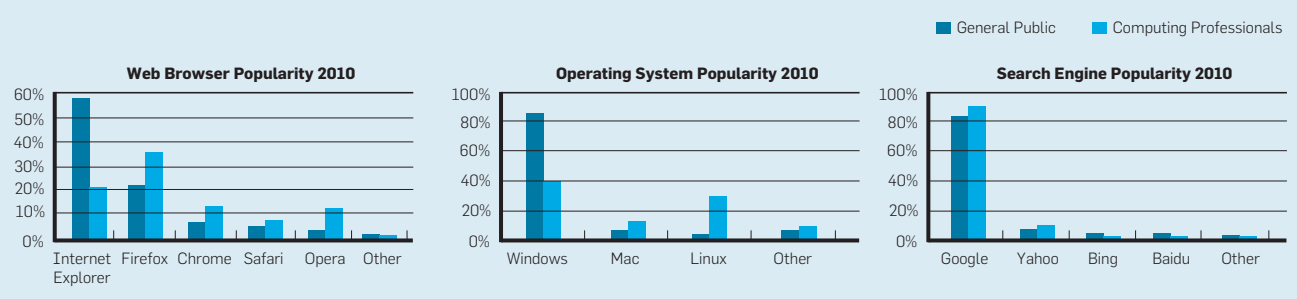
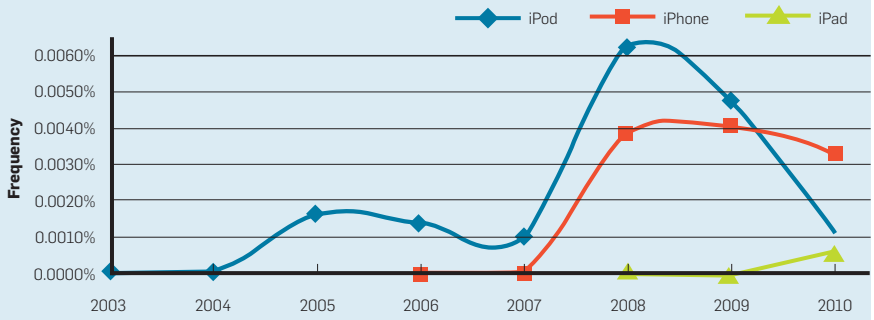


Figure 6. Interest in mass-market commercial Apple products.



riod.[11] However, after the sudden enormous popularity of the iPhone in 2007 and 2008, interest in the iPod began to wane, as the ongoing incorporation of digital music capabilities into smartphones (such as the iPhone) made the iPod an increasingly redundant technology. Beginning around 2009 interest in the iPhone also began to decline among computing professionals, while at the same time interest in Apple's most current innovation—iPad—continued to increase. Together, these results demonstrate the viability of using *n*-gram analyses to help study technological evolution over time.

Conclusion

Findings of a 2010 U.S. Department of Labor longitudinal study indicate college-educated workers hold an average of 11 different jobs by the time they are 44.[16] This surprising statistic has important ramifications for all organizations, since it implies skilled workers (such as managers) will, on average, hold their positions of leadership and productivity for only a few years before moving on. Reflecting such short tenure, managers inevitably struggle to develop the sort of deep, tacit understanding of their organizations that is critical to strategic decision making. If such knowledge can be found within the archived document artifacts produced by an institution over time, then *n*-gram analyses like those we have presented here may prove invaluable for sifting through the content of these vast collections of archival documents, as well as contribute to improvements in managerial decision making.

With respect to our conclusions, we found more recent issues of *Communications* contained substantially more content than earlier issues. The nature of this content is also changing, as articles published in *Communications* are trending away from managerial and business subjects to focus instead on more technical and computational subjects. Despite this trend, the writing style in *Communications* became less formal and more personal over time, helping it reach a wider audience.

In addition to these structural changes, we also pursued an assortment of analyses that, together, demonstrate the potential of the *n*-gram method for institutional data mining. We invite you to conduct your own *Communications* analyses using our *n*-gram tool at http://www.invivo.co/ngrams/cacm.aspx. Broadly speaking, *n*-grams are a powerful tool for gaining insight into textual data that might otherwise go unnoticed. Though the analyses we discussed here target a single publication, the *n*-gram method itself is suitable for a range of analytic situations, including virtually any large corpus of text data. For example, historical documents could be analyzed to identify long-hidden trends, fads, and modes of thought. Blogs or tweets could be analyzed to produce a near-real-time snapshot of the global consciousness. And software source code could be analyzed for style and/or efficiency. Combined with optical character recognition, *n*-grams could be leveraged to sift through mountains of archived paper documents. Indeed, the possibilities for *n*-gram analyses are almost limitless.

Finally, we hope this article serves as evidence for why scholarly journals should not be viewed as immutable objects but as living entities evolving and responding to their environments. To this end, analyses like ours should be conducted periodically to codify and chronicle a publication's history, helping refine its identity over time. This is especially important with long-lived journals (such as *Communications*) that serve as standard-bearers for their respective disciplines. When applied to a broad portfolio of publications, such analyses would help readers, authors, advertisers, and editors alike better understand journals more objectively and perhaps even glimpse what the future holds for disciplines like computer science. C

References

1. Agarwal, R. and Lucas, J.H.C. The information systems identity crisis: Focusing on high-visibility and high-impact research. *MIS Quarterly 29*, 3 (Sept. 2005) 381–398.
2. Benbasat, I. and Zmud, R.W. The identity crisis within the IS discipline: Defining and communicating the discipline's core properties. *MIS Quarterly 27*, 2 (June 2003) 183–194.
3. Bhattacherjee, A. and Gill, G. Whom are we informing? Issues and recommendations for MIS research from an informing sciences perspective. *MIS Quarterly 33*, 2 (June 2009), 217–235.
4. Bloomer, M., Hodkinson, P., and Billett, S. The significance of ontogeny and habitus in constructing theories of learning. *Studies in Continuing Education 26*, 1 (Mar. 2004), 19–43.
5. Bohannon, J. Google Books, Wikipedia, and the future of culturomics. *Science 331*, 6014 (Jan. 14, 2011), 135.
6. Crawford, D. Editorial pointers. *Commun. ACM 48*, 8 (Aug. 2005), 5.
7. Crawford, D. Editorial pointers. *Commun. ACM 46*, 4 (Apr. 2003), 5.
8. Crawford, D. Editorial pointers. *Commun. ACM 44*, 1 (Jan. 2001), 5.
9. Dennis, A., Wixom, B.H., and Roth, R.M. *Systems Analysis and Design, Fourth Edition*. John Wiley & Sons, Inc., Hoboken, NJ, 2009.
10. Descartes, R. *Principia Philosophiae*. Kluwer Academic, Norwell, MA, 1984, originally published 1644.
11. Dowling, S. Apple Reports First Quarter Results: Quarterly Revenue & Net Income Highest in Apple's History. Apple Computer, Cupertino, CA, 2005; http://blog.ttnet.net/resources/2005/01/13/apple-reports-first-quarter-results-quarterly-revenue-net-income-highest-in-apples-history-20050113
12. Manning, C.D. and Schütze, H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999.
13. Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Team, T.G.B., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Aiden, E.L. Quantitative analysis of culture using millions of digitized books. *Science 331*, 6014 (Jan. 14, 2011), 176.
14. Misa, T.J. *Gender Codes: Why Women Are Leaving Computing*. John Wiley & Sons, Inc., Hoboken, NJ, 2010.
15. NetMarketShare. Usage Share Statistics for Internet Technologies. NetApplications.com, Aliso Viejo, CA, 2010; http://netmarketshare.com/
16. U.S. Bureau of Labor Statistics. *Number of Jobs Held, Labor Market Activity, and Earnings Growth Among the Youngest Baby Boomers: Results From a Longitudinal Survey*. U.S. Department of Labor, Washington, D.C., 2010; http://www.bls.gov/news.release/nlsoy.nr0.htm

Daniel S. Soper (dsoper@fullerton.edu) is an assistant professor in the Information Systems and Decision Sciences Department of the Mihaylo College of Business and Economics at California State University, Fullerton.

Ofir Turel (oturel@fullerton.edu) is a professor in the Information Systems and Decision Sciences Department of the Mihaylo College of Business and Economics at California State University, Fullerton.