



Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
<http://economics.sas.upenn.edu/pier>

PIER Working Paper 13-003

“A Personal Perspective on the Origin(s) and Development of “Big Data”: The Phenomenon, the Term, and the Discipline”, Second Version

by

Francis X. Diebold

<http://ssrn.com/abstract=2202843>

A Personal Perspective on the Origin(s) and Development of “Big Data”: The Phenomenon, the Term, and the Discipline*

Francis X. Diebold
University of Pennsylvania
fdiebold@sas.upenn.edu

First Draft, August 2012
This Draft, November 26, 2012

Abstract: I investigate Big Data, the phenomenon, the term, and the discipline, with emphasis on origins of the term, in industry and academics, in computer science and statistics/econometrics. Big Data the phenomenon continues unabated, Big Data the term is now firmly entrenched, and Big Data the discipline is emerging.

Key words: Massive data, computing, statistics, econometrics

JEL codes: C81, C82

Contact Info: fdiebold@sas.upenn.edu

*For useful communications I thank – without implicating in any way – Larry Brown, Xu Cheng, Flavio Cunha, Susan Diebold, Dean Foster, Michael Halperin, Steve Lohr, John Mashey, Tom Nickolas, Lauris Olson, Mallesh Pai, Marco Pospiech, Frank Schorfheide, Minchul Shin, and Mike Steele. I also thank, again without implicating, Stephen Feinberg, Douglas Laney and Fred Shapiro, with whom I have not had the pleasure of communicating, but who are friends of friends, and whose insights were valuable. All referenced web addresses are clickable from pdf.

1 Introduction

Big Data is at the heart of modern science and business. Premier scientific groups are intensely focused on it, as evidenced for example by the August 2012 “Big Data Special Issue” of *Significance*, a joint publication of the American Statistical Association and the Royal Statistical Society.¹ Society at large is also intensely focused on it, as documented by major reports in the business and popular press, such as Steve Lohr’s “How Big Data Became so Big” (*New York Times*, August 12, 2012).²

2 Big Data the Phenomenon

Big Data the phenomenon marches onward.³ Indeed the necessity of grappling with Big Data, and the desirability of unlocking the information hidden within it, is now a key theme in all the sciences – arguably *the* key scientific theme of our times. Parts of my field of econometrics, to take a tiny example, are working furiously to develop methods for learning from the massive amount of tick-by-tick financial market data now available.⁴ In response to a question like “How big is your dataset?” in a financial econometric context, an answer like “90 observations on each of 10 variables” would have been common fifty years ago, but now it’s comically quaint. A modern answer is likely to be a file size rather than an observation count, and it’s more likely to be 200 GB than the 50 kB (say) of fifty years ago. And the explosion continues: the “Big Data” of fifteen years ago are most definitely small data by today’s standards, and moreover, someone reading this in twenty years will surely laugh at my implicit assertion that a 200 GB dataset is large.⁵

¹<http://www.significancemagazine.org/view/0/index.html>.

²<http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html>.

³By the “Big Data” phenomenon, I mean the recent phenomenon of explosive data growth and associated massive datasets. Diebold (2003) not only used the term extensively in precisely that way, but also defined it, noting that “Recently much good science, whether physical, biological, or social, has been forced to confront – and has often benefited from – the Big Data phenomenon. Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology.”

⁴For a recent overview, see Andersen et al. (2012).

⁵And of course the assertion that 200 GB is large by today’s standards is with reference to my field of econometrics. In other disciplines like physics, 200 GB is already small. The large hadron collider experiments that led to discovery of the Higgs boson, for example, produce a petabyte of data (10^{15} bytes) *per second*.

3 Big Data the Term

My interest was piqued when Marco Pospiech, a Ph.D. student studying the Big Data phenomenon at the Technical University of Freiberg, informed me in private correspondence that he had traced the use of the term (in the modern sense) to my paper, “‘Big Data’ Dynamic Factor Models for Macroeconomic Measurement and Forecasting,” presented at the Eighth World Congress of the Econometric Society in Seattle in August 2000, and subsequently published as Diebold (2003).⁶ Amused, I did a bit more digging. As regards my paper, what’s true with near certainty is that it is the first academic reference to Big Data in a title or abstract in the statistics, econometrics, or additional x -metrics (insert your favorite x) literatures.⁷ But deeper investigation reveals that the situation is more complicated – and more interesting – than it first appears: the origins of the term are intriguing and a bit murky, involving both industry and academics, computer science and statistics/econometrics. I play an early role, but I am not alone, and as it turns out, not first.

I stumbled on the term Big Data innocently enough, via discussion of two papers that took a new approach to macro-econometric dynamic factor models (DFMs), Reichlin (2003) and Watson (2003), presented back-to-back in an invited session of the 2000 World Congress of the Econometric Society.⁸ Older dynamic factor analyses included just a few variables, because parsimony was essential for tractability of numerical likelihood optimization. The new work by Reichlin and Watson, in contrast, showed how DFMs could be estimated using principal components, thereby dispensing with numerical optimization and opening the field to analysis of much larger datasets while nevertheless retaining a likelihood-based approach. My discussion had two overarching goals. First, I wanted to contrast the old and new macro-econometric DFM environments. Second, I wanted to emphasize that the driver of the new macro-econometric DFM developments matched the driver of many other recent scientific developments: *explosive growth in available data*. To that end, I wanted a concise term that conjured a stark image. I came up with “Big Data,” which seemed apt and resonant and intriguingly Orwellian (especially when capitalized), and which helped to promote both goals.

But there really *is* nothing new under the sun, and credit for the term Big Data must be shared. The appropriate allocation is open to debate, however, as there are issues of Big Data interpretation and context, and things get murkier if one includes unpublished and/or

⁶The November 2000 post-conference working paper, Diebold (2000), is available at <http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF>.

⁷Moreover, as progressively more searches find nothing, it’s becoming progressively more likely that it’s the first reference in those literatures, whether in the title, abstract or elsewhere.

⁸<http://www.econometricsociety.org/meetings/wc00/Invited.pdf>.

non-academic references. A few pre-2000 references to Big Data, both academic and non-academic, are intriguing but ultimately unconvincing, using the term but not thoroughly aware of the phenomenon.⁹ Conversely, academics were aware of the emerging phenomenon but not the term.¹⁰ There is, however, some pre-2000 (non-academic, unpublished) activity that is spot-on. In particular, Big Data the term, coupled with awareness of Big Data the phenomenon, was clearly percolating at Silicon Graphics (SGI) in the mid 1990s. John Mashey, retired former Chief Scientist at SGI, produced a 1998 SGI slide deck entitled “Big Data and the Next Wave of InfraStress,” which demonstrates clear awareness of Big Data the phenomenon.^{11,12} Related, SGI ran an ad that featured the term Big Data in *Black Enterprise* (March 1996, p. 60), several times in *Info World* (starting November 17, 1997, p. 30), and several times in *CIO* (starting February 15, 1998, p. 5). Clearly then, Mashey and the SGI community were on to Big Data early, using it both as a unifying theme for technical seminars and as an advertising hook.

There is also at least one more relevant pre-2000 Big Data reference in computer science. It is subsequent to Mashey *et al.*, but interestingly, it comes from the academic as opposed to industry part of the computer science community, and it not only uses the term but also demonstrates some awareness of the phenomenon. Weiss and Indurkha (1998) note that “... very large collections of data ... are now being compiled into centralized data warehouses, allowing analysts to make use of powerful methods to examine data more comprehensively. In theory, ‘Big Data’ can lead to much stronger conclusions for data-mining applications, but in practice many difficulties arise.”

⁹On the academic side, Tilly (1984) mentions Big Data, but his article is not about the Big Data phenomenon and demonstrates no awareness of it; rather, it is a discourse on whether statistical data analyses are of value to historians. On the non-academic side, the margin comments of a computer program posted to a newsgroup in 1987 mention a programming technique called “small code, big data.” (See https://groups.google.com/forum/?fromgroups#!msg/comp.sources.misc/d3EXP4D_VK8/x7WrVBMb5FgJ.) Fascinating, but off-mark. Next, Eric Larson provides an early popular-press mention in a 1989 *Washington Post* article about firms that assemble and sell lists to junk-mailers. He notes in passing that “The keepers of Big Data say they do it for the consumer’s benefit.” Again fascinating, but again off-mark. (See Eric Larson, “They’re Making a List: Data Companies and the Pigeonholing of America,” *Washington Post*, July 27, 1989.) Finally, a 1996 PR Newswire, Inc. release mentions network technology “for CPU clustering and Big Data applications...” Still off-mark, neither reporting on the Big Data phenomenon nor demonstrating awareness of it, instead reporting exclusively on a particular technology, the so-called high-performance parallel interface.

¹⁰See, for example, *Massive Data Sets: Proceedings of a Workshop*, Committee on Applied and Theoretical Statistics, National Research Council (National Academies Press, 1997), http://www.nap.edu/catalog.php?record_id=5505.

¹¹http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf.

¹²Mashey notes in private communication that the deck was for a “living talk” and hence updated regularly, so that the 1998 version is not the earliest. The earliest deck of which he is aware (and hence I am aware) is from 1997.

Finally, arriving on the scene later but also going beyond previous work in compelling ways, Laney (2001) highlighted the “Three V’s” of Big Data (Volume, Variety and Velocity) in an unpublished 2001 research note at META Group.¹³ Laney’s note is clearly relevant, and it goes beyond my exclusive focus on volume, producing a significantly enriched conceptualization of the Big Data phenomenon.¹⁴ In short, if Laney arrived slightly late, he nevertheless brought more to the table.

4 Big Data the Discipline

Big Data is now not only a phenomenon and term, but also a *discipline*. It leaves me with mixed, but ultimately positive, feelings. At first pass it sounds like frivolous fluff, as do other information technology sub-disciplines with catchy names like “artificial intelligence,” “data mining” and “machine learning.” Indeed it’s hard to resist smirking when told that Big Data has now arrived as a new discipline and business, and that major firms are rushing to create new executive titles like “Vice President for Big Data.”¹⁵ But as I have argued, the phenomenon behind the term is very real, so it may be natural and desirable for a corresponding new discipline to emerge, whatever its executive titles.

It’s not obvious, however, that a new discipline is required, or that Big Data *is* a new discipline. Skeptics will argue that traditional disciplines like computer science, statistics and x -metrics are perfectly capable of confronting the new phenomenon, so that Big Data as a discipline is redundant, merely drawing a box around some traditional disciplines. But it’s hard not to notice that the whole of the emerging Big Data discipline seems greater than the sum of its parts. That is, by drawing on perspectives from a variety of traditional disciplines, Big Data as a discipline is not merely taking us to bigger *traditional* places. Rather, it’s taking us to wildly *new* places, unimaginable only a short time ago, ranging from cloud computing and associated massively-parallel algorithms, to methods for controlling false-discovery rates when testing millions of hypotheses, with much in between. Indeed one could argue that, in a landscape littered with failed attempts at interdisciplinary collaboration, Big Data is emerging as a major interdisciplinary triumph.

¹³META is now part of Gartner.

¹⁴<http://goo.gl/Bo3GS>.

¹⁵Seriously. Lohr reports the title “Vice President for Big Data” in his earlier-mentioned *Times* piece, at <http://www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html>.

5 Conclusion

The term “Big Data,” which spans computer science and statistics/econometrics, probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid 1990s, in which John Mashey figured prominently. The first significant academic references are arguably Weiss and Indurkha (1998) in computer science and Diebold (2000) in statistics/econometrics. An unpublished 2001 research note by Douglas Laney at Gartner enriched the concept significantly. Hence the term “Big Data” appears reasonably attributed to Massey, Weiss and Indurkha, Diebold, and Laney. Big Data the phenomenon continues unabated, and Big Data the discipline is emerging.

References

- Andersen, T.G., T. Bollerslev, P.F. Christoffersen, and F.X. Diebold (2012), “Financial Risk Measurement for Financial Risk Management,” In M. Harris, G. Constantinides and R. Stulz (eds.), *Handbook of the Economics of Finance*, Elsevier, in press.
http://www.ssc.upenn.edu/~fdiebold/papers/paper107/ABCD_HOEF.pdf.
- Diebold, F.X. (2000), “Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting,” Discussion Read to the Eighth World Congress of the Econometric Society, Seattle, August.
<http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF>.
- Diebold, F.X. (2003), “Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting: A Discussion of the Papers by Reichlin and Watson,” In M. Dewatripont, L.P. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, Cambridge University Press, 115-122.
- Laney, D. (2001), “3-D Data Management: Controlling Data Volume, Velocity and Variety,” META Group Research Note, February 6.
<http://goo.gl/Bo3GS>.
- Reichlin, L. (2003), “Factor Models in Large Cross Sections of Time Series,” In M. Dewatripont, L.P. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, Cambridge University Press, 47-86.
- Sala-i-Martin, X. (1997), “I Just Ran Two Million Regressions,” *American Economic Review*, 87 (May), 187–183.
- Tilly, C. (1984), “The Old New Social History and the New Old Social History,” *Review (Fernand Braudel Center)*, 7, 363–406.
- Watson, M.W. (2003), “Macroeconomic Forecasting Using Many Predictors,” In M. Dewatripont, L.P. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, Cambridge University Press, 87-115.
- Weiss, S.M. and N. Indurkha (1998), *Predictive Data Mining: A Practical Guide*, Morgan Kaufmann Publishers, Inc.