



# Is Big Data a Transient Problem?

Jimmy Lin

**W**hat's growing faster, Big Data or Moore's Law?

It's undeniable that the amount of data that organizations must store, process, organize, and analyze is growing rapidly. This requires increasingly larger clusters and data-centers, as well as increasingly complex software infrastructure to orchestrate the necessary computations. But is Big Data growing faster than Moore's Law is lowering the costs of computing capabilities to accomplish these tasks? For rhetorical convenience, I'm using Big Data to refer to all the things we want to *do* on massive collections of data, and Moore's Law to refer to exponential increases in computing capabilities for *doing it*. It's worth emphasizing that I don't literally mean the periodic doubling of transistors on a chip; I use Moore's Law as a convenient shorthand to refer to continued exponential advances in computing.

Logically, there are only three possibilities:

1. Big Data is growing faster than Moore's Law.
2. Big Data is growing at the same rate as Moore's Law.
3. Big Data is growing slower than Moore's Law.

The first two scenarios aren't particularly interesting: In the first case, what we can store will be bounded by Moore's Law and the rest of the data will need to be processed in real time (and then thrown away). The second case is essentially the status quo (hence, uninteresting). The third scenario, however, is intriguing: it suggests that computing capabilities are going to "catch up" to Big Data at some point. In other words, Big Data is a *transient problem*.

## Defining the Question

What do I mean by a transient problem? Here's an analogy that might resonate with many: I remember when digital music first burst upon the scene about two decades ago. At the time, storing

all those MP3s on my (gasp, desktop!) computer was a big deal. I distinctly remember my music collection consuming most of my hard drive, and having to sacrifice (delete) some files to make room for others. Over time, however, keeping MP3s around became less and less of a problem: storage technology improved many-fold, whereas the amount of music I could consume had a clear upper bound (24 hours in a day), and beyond a certain point, increased encoding quality didn't make a difference (at least to my ears). Today, all the music I could possibly want to listen to easily fits in my pocket (on my phone). In this sense, digital music storage was a transient problem that technology solved. Is Big Data the same way?

Of course, I'm assuming that Moore's Law will continue for some time, or more generally, exponential increases in computing technology will continue unabated. Obviously, there are physical limits,<sup>1</sup> but we're still pretty far from those. What do I mean by "for some time?" I have left this deliberately vague, because it depends on the particular prediction: when examining extrapolations of computing capabilities ("supply") with the demands of Big Data, it only matters if the pace of technological improvement will "hold up" until the anticipated crossover point. As to the broader question about continued technological progress (dire predictions about the end of Dennard Scaling notwithstanding), it depends on whether you're Cornucopian or Malthusian. This philosophical argument is beyond the scope of my piece, although I would note that Malthusians have essentially been wrong every time, because human civilization is still around and we don't (yet) live in a post-apocalyptic wasteland.

In this article, I only focus on *human-generated data* and leave aside data from scientific instruments (such as the Large Hadron Collider and the Square Kilometer Array), remote sensing (for

## Big Data Bites

“Big Data Bites” is a regular department in *IEEE Internet Computing* that aims to deliver thought-provoking and potentially controversial ideas about all aspects of Big Data. Interested in contributing? Drop me a line!

—Jimmy Lin

example, satellite imagery), surveillance (including traffic cameras), and related applications because the economics are quite different. Human-generated data benefit from what Jeff Bezos calls *the flywheel*: the virtuous cycle where insights from user-generated data are exploited to improve products and services, which lead to broader usage and even more user-generated data, thus closing the loop. Amazon, Google, Facebook, Uber, and countless other companies are all built on this powerful driver. In contrast, the economics of data not generated by humans are very different (and in some ways, less interesting).

Note that my definition of human-generated data is fairly expansive: it includes all forms of data generated by humans, including those in databases (for example, Amazon’s vast product catalog is human generated in the sense that the products for sale are produced and consumed by humans), behavior logs, personal medical records, and even some aspects of the Internet of Things (the data generated by connected appliances are ultimately derived from human activity).

### Data Bounds

The upper bound on human-generated data is the product of two terms: total human activity and the amount of data generated per unit time, or the *data density*. Let’s examine the first term: by most accounts, the human population will stabilize sometime relatively soon. The “medium” scenario of Samir KC and Wolfgang Lutz<sup>2</sup> shows a continued world population increase, resulting in 9.17 billion in 2050, peaking around 9.4 billion in the 2070s, and declining somewhat to 9 billion by 2100. A competing analysis by Patrick Gerland and his colleagues<sup>3</sup> is somewhat more pessimistic, arguing that world population stabilization is unlikely this century. According to their models, there’s an 80 percent probability that the world population will increase to between 9.6 and 12.3 billion in 2100.

Regardless, there appears to be a consensus that overall fertility rates are decreasing – the debate is mostly over how quickly – so the point remains that the human population on this planet won’t grow indefinitely. This means, in turn, that there’s a finite upper bound on human activity; after all, there are only 24 hours in a day. My analysis depends on the assumption that the human population won’t grow without bound, which means that when we start colonizing the galaxy, all bets are off!

Let’s look at the second term, the data density. As a specific case, consider the amount of human-generated textual data on the Web (for example, HTML pages): evidence suggests that it’s growing slower than Moore’s Law. Andrew Trotman and Jinglan Zhang present quite reasonable projections suggesting that by the middle of the next decade, “the storage capacity of a single hard drive will exceed the size of the index of the Web at that time,” and that “within another decade it will be possible to store the entire searchable text on the same hard drive.”<sup>4</sup> They explore the implications of this for the design of search engines, which is interesting but beyond the scope of the current discussion. You might quibble with the details of their projections, but the underlying point remains: when we talk about text, it’s not growing as fast as we have room to store and index it.

The challenge in extending this argument to all human-generated media is that there’s no upper bound to data density except for special cases like text, since we can arbitrarily improve sensor resolution (we’ll even-

tually run into quantum limits, I suppose, but we’re far from those). The argument with textual data “works” because text has a low and constant data density – which isn’t the case with images and video, for example. What if we include all human-generated images and video on the Web? Imagine a dystopian future where all humanity does is create YouTube videos all day long: although the content’s length in hours would be bounded, the data’s size wouldn’t, since the resolution could be made arbitrarily better. You might counter with the observation that, beyond a certain resolution, the human visual system can’t tell, so the bandwidth of the human perceptual system might provide a natural upper bound. But what if I wanted to zoom in on a previously captured image or video? Then I’d want as high a resolution as physically possible. (Perhaps it wouldn’t matter if nobody was watching!)

Why stop at video? What about a personal magnetic resonance image (MRI) scanner that continuously monitors and captures our physiological state? Or a swarm of nanobots living inside us that gathers detailed measurements of our molecular functionings? The limits imposed by physics aside, it’s difficult to see an end to increased data density. However, it’s important to remember that many of the technological trends that give rise to higher-resolution sensors are either directly or indirectly related to Moore’s Law (for example, the increase in megapixels in digital cameras). Could it be that what Moore’s Law giveth, data density taketh? In which case, the real question is: What’s the growth in our ability to capture data at finer resolutions

compared to increases in computing power? The human population is simply the “constant” in the equation (albeit a fairly large constant), but if we’re talking about exponential growth, the constant is basically irrelevant.

At this point, we might shift the argument to focus on useful data as opposed to all data. Suppose I went around capturing 4K-quality video of my every waking moment (technically possible today) – who cares and why would I possibly want that? Perhaps not now, but this is a failure of imagination: much of data science and Big Data analytics today is built on data we thought was useless two decades ago (in fact, some people call it *data exhaust*). One day, questioning the usefulness of certain data-collection activities could sound as quaint as asking: Why would we ever want to keep around click logs? What possible use could we have for them?

However, a more nuanced way to think about this issue is to compare the growth of Big Data with the extent to which we can exploit the data practically. Numerous studies have found roughly a log-linear relationship between the amount of data analyzed and its effectiveness in an application.<sup>5,6</sup> That is, achieving the next increment in effectiveness (for example, accuracy in a classification task) requires a multiple-fold increase in the amount of data. The relationship between Moore’s Law and the slope of this effectiveness line is important. For example, if making an algorithm incrementally better requires four times more data, then one Moore’s Law cycle (doubling capabilities) is insufficient to improve our algorithm. However, in rough terms, it does make the current problem half as difficult. In this case, we might say that practically exploitable Big Data is growing slower than Moore’s Law. Yet, there’s a hole in this argument, since it assumes that there won’t be significant algorithmic improvements in the future. Perhaps

some brilliant researcher will devise entirely new classes of algorithms that exploit Big Data much more efficiently?

### Implications

So, is Big Data growing slower than Moore’s Law? Hopefully, I’ve shown that it’s plausible, at least in a suitably qualified or more restrictive form. Thus, it’s worthwhile to consider some of the implications on future computing systems for Big Data.

The most important implication is what I call “the revenge of scale up.” A nearly unquestioned assumption in the design of data processing systems today is the superiority of scaling “out” on a cluster of commodity machines as opposed to scaling “up” on a single “beefy” machine (more memory, more cores). Previously, scaling up simply wasn’t an option because no single machine, no matter how powerful, was sufficient to handle the data-processing task at hand. Scaling out, however, incurs large costs in terms of synchronization, communication, and fault tolerance. If Big Data is indeed growing slower than Moore’s Law, then we need to revisit the scale out versus scale up debate, because at some point, a single machine might become powerful enough to handle Big Data.

In fact, this debate is already under way. According to the analysis of Antony Rowstron and his colleagues,<sup>7</sup> at least two analytics production clusters (at Microsoft and Yahoo) have median job input sizes under 14 gigabytes and 90 percent of jobs on a Facebook cluster have input sizes under 100 gigabytes (in 2012). A study of enterprise Hadoop clusters at around the same time shows that the workloads are dominated by relatively small jobs.<sup>8</sup> So why are we still using distributed processing frameworks such as MapReduce or Spark when the data easily can be held in memory on a single machine? As my colleague Jens Dittrich puts it, why are we all obsessed with building a 1,000-horsepower supercar

just to make a two-mile trip to the supermarket? Indeed, we’re seeing a resurgence of interest in scale-up approaches, particularly from the academic community.<sup>9-12</sup>

So then, what’s with all the petabytes that we’re accumulating in our vast data warehouses? As it turns out, the process of extracting features (or “signals”) from raw data is quite distinct from data mining and machine-learning algorithms for deriving insights from those features. In the first, we typically distill raw data into sparse feature vectors; during this process there’s typically many orders of magnitude reduction in data size. The feature vectors then serve as input to machine-learning or data-mining algorithms. We still need large clusters for feature extraction, since the raw data are often immense and we need the aggregate throughput of disk spindles across many machines. However, the distilled feature vectors are quite manageable. For example, state-of-the-art large-scale machine learning today talks about billions of training examples with millions of parameters,<sup>13</sup> on the order of a trillion nonzero features in total (since the feature vectors are sparse). A trillion floating-point values occupy four terabytes of main memory: any day now, we’ll purchase commodity machines with that much memory.

Similarly, consider a graph with a trillion edges: stored in the most naive manner as (source, destination) pairs, it would take eight terabytes. We’ll purchase a commodity machine with that much memory soon enough (one Moore’s Law cycle later, in fact). In general, graphs of human social relations are bounded by population size, which suggests that graph problems are progressively becoming easier with each generation of hardware. As a concrete example, Twitter’s production graph recommendation service began with a scale-up approach, holding the entire follower graph in memory on a single machine (and exploiting replication for increased throughput).<sup>14</sup> Examples

of impressively fast machine learning on individual machines include Vowpal Wabbit (see [https://github.com/JohnLangford/vowpal\\_wabbit](https://github.com/JohnLangford/vowpal_wabbit)), the lock-free “Hogwild” method for parallelizing parameter updates,<sup>15</sup> and recent work in matrix factorization.<sup>16</sup>

Decoupling feature extraction and machine learning suggests a heterogeneous architecture where we exploit clusters to munge the raw data, and then bring extracted features over to a single machine to perform the actual machine learning – in other words, scale out for data cleaning, feature extraction, and so on, and scale up for machine learning.

This architecture, however, raises two interesting questions: First, data scientists loathe multiple processing frameworks, which introduce impedance mismatches into their daily activities. Having one framework for feature extraction and another framework for machine learning introduces friction. Thus, it would be ideal to have a single framework that both scales up and out. Second, datacenter operations engineers prefer consolidated clusters with a homogeneous hardware configuration, from both the perspective of economics and management overhead. Modern cluster-management software such as Mesos<sup>17</sup> (and Google’s equivalents) work best with homogeneous fleets of servers. This doesn’t mean they’re unable to handle workloads where certain jobs (that require lots of memory) can run only on certain machines (that have enough memory) – but it does add an element of complexity in scheduling and coordination.

Finally, if Big Data indeed is growing slower than Moore’s Law, this means that the Big Data of today will fit in my pocket tomorrow – in the same way that my music collection, which occupied most of the disk on my desktop machine about 15 years ago, fits in my pocket easily today. How would information seeking change if we could store a cache of the Web in a mobile device we carry around all

the time? We’re already proceeding down this path: Ask yourself, when was the last time you searched Google just to go to Wikipedia? Or when you used a search engine as a bookmark to return to a page you’ve visited before (what the information retrieval community calls “refinding”<sup>18</sup>)? In both cases, perhaps a local cache of the Web might do the job just as well, and has the additional advantages of freeing us from flaky connectivity and network latencies.

Already today, so-called low-power “wimpy” devices (such as mobile phones and tablets) are far more prevalent than traditional servers and PCs. The technology research firm Gartner forecasts that worldwide shipments of PCs in 2015 will total around 320 million units, compared to 2.3 billion mobile phones and tablets ([www.gartner.com/newsroom/id/2791017](http://www.gartner.com/newsroom/id/2791017)). Thus, it’s worthwhile to explore how infrastructure designed for “brawny” servers in a traditional datacenter might run in wimpy environments, and the implications of many thousands of wimpy devices within a relatively small area (say, in a Manhattan city block or sporting venue). Interesting work along these lines include deploying full-text search engines<sup>19</sup> and transactional databases<sup>20</sup> on mobile phones, and Web archiving infrastructure on Raspberry Pis.<sup>21</sup> In addition to scaling out and up, it’s worthwhile to think about scaling “down” Big Data technology.

**W**hat does the future hold for Big Data? It could be the same qualitatively, just bigger and better, or there might be fundamentally disruptive forces that completely reshape the computing landscape. Trying to predict the future, of course, is a perilous exercise. At best, this discussion provides some deep insight on future developments in Big Data. At worst, it makes for an interesting cocktail conversation. Either way, it’s worth the rumination. □

## Acknowledgments


I’d like to thank Andrew Trotman for various engaging discussions; and Charlie Clarke, Craig Murray, and Arnab Nandi for comments on previous drafts of this piece.

## References

1. S. Lloyd, “Ultimate Physical Limits to Computation,” *Nature*, vol. 406, 2000, pp. 1047–1054.
2. S. KC and W. Lutz, “The Human Core of the Shared Socioeconomic Pathways: Population Scenarios by Age, Sex, and Level of Education for All Countries to 2100,” *Global Environmental Change*, 2014; doi:10.1016/j.gloenvcha.2014.06.004.
3. P. Gerland et al., “World Population Stabilization Unlikely This Century,” *Science*, vol. 346, no. 6206, 2014, pp. 234–237.
4. A. Trotman and J. Zhang, “Future Web Growth and Its Consequences for Web Search Architectures,” 2013; arXiv:1307.1179v1.
5. M. Banko and E. Brill, “Scaling to Very Very Large Corpora for Natural Language Disambiguation,” *Proc. 39th Ann. Meeting of the Assoc. for Computational Linguistics*, 2001, pp. 26–33.
6. T. Brants et al., “Large Language Models in Machine Translation,” *Proc. 2007 Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 858–867.
7. A. Rowstron et al., “Nobody Ever Got Fired for Using Hadoop on a Cluster,” *Proc. 1st International Workshop on Hot Topics in Cloud Data Processing*, 2012, article no. 2.
8. Y. Chen, S. Alspaugh, and R. Katz, “Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads,” *Proc. 38th Int’l Conf. Very Large Data Bases*, 2012, pp. 1802–1813.
9. R. Appuswamy et al., “Scale-Up vs Scale-Out for Hadoop: Time to Rethink?” *Proc. 4th ACM Symp. Cloud Computing*, 2013.
10. J. Shun and G.E. Blelloch, “Ligra: A Lightweight Graph Processing Framework for Shared Memory,” *Proc. 18th ACM SIGPLAN Symp. Principles and Practice of Parallel Programming*, 2013, pp. 135–146.
11. K.A. Kumar et al., “Optimization Techniques for “Scaling Down” Hadoop on Multi-Core, Shared-Memory Systems,” *Proc. 17th Int’l Conf. Extending Database Technology*, 2014, pp. 13–24.

12. F. Chen et al., "Palette: Enabling Scalable Analytics for Big-Memory, Multicore Machines," *Proc. 2014 ACM SIGMOD Int'l Conf. Management of Data*, 2014, pp. 705–708.
13. A. Agarwal et al., "A Reliable Effective Terascale Linear Learning System," *J. Machine Learning Research*, vol. 15, Mar. 2014, pp. 1111–1133.
14. P. Gupta et al., "WTF: The Who to Follow Service at Twitter," *Proc. 22nd Int'l World Wide Web Conf.*, 2013, pp. 505–514.
15. F. Niu et al., "Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent," *Advances in Neural Information Processing Systems 24*, 2011, pp. 693–701.
16. Z. Liu, Y.-X. Wang, and A.J. Smola, "Fast Differentially Private Matrix Factorization," 2015; arXiv:1505.01419v2.
17. B. Hindman et al., "Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center," *Proc. 8th USENIX Symp. Networked Systems Design and Implementation*, 2011.
18. S.K. Tyler and J. Teevan, "Large-Scale Query Log Analysis of Re-Finding," *Proc. 3rd ACM Int'l Conf. Web Search and Data Mining*, 2010, pp. 191–200.
19. A. Balasubramanian et al., "FindAll: A Local Search Engine for Mobile Phones," *Proc. 8th Int'l Conf. Emerging Networking Experiments and Technologies*, 2012, pp. 277–288.
20. T. Mühlbauer et al., "One DBMS for All: The Brawny Few and the Wimpy Crowd," *Proc. 2014 ACM SIGMOD Int'l Conf. Management of Data*, 2014, pp. 697–700.
21. J. Lin, "Scaling Down Distributed Infrastructure on Wimpy Machines for Personal Web Archiving," *Proc. 24th Int'l World Wide Web Conf. Companion*, 2015, pp. 1351–1355.

**Jimmy Lin** holds the David R. Cheriton Chair in the David R. Cheriton School of Computer Science at the University of Waterloo. His research lies at the intersection of information retrieval and natural language processing, with a particular focus on Big Data and large-scale distributed infrastructure for text processing. Lin has a PhD in electrical engineering and computer science from MIT. Contact him at [jimmylin@uwaterloo.ca](mailto:jimmylin@uwaterloo.ca).

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

## ADVERTISER INFORMATION

### Advertising Personnel

Marian Anderson: Sr. Advertising Coordinator  
Email: [manderson@computer.org](mailto:manderson@computer.org)  
Phone: +1 714 816 2139 | Fax: +1 714 821 4010

Sandy Brown: Sr. Business Development Mgr.  
Email: [sbrown@computer.org](mailto:sbrown@computer.org)  
Phone: +1 714 816 2144 | Fax: +1 714 821 4010

### Advertising Sales Representatives (display)

Central, Northwest, Far East:  
Eric Kincaid  
Email: [e.kincaid@computer.org](mailto:e.kincaid@computer.org)  
Phone: +1 214 673 3742  
Fax: +1 888 886 8599

Northeast, Midwest, Europe, Middle East:  
Ann & David Schissler  
Email: [a.schissler@computer.org](mailto:a.schissler@computer.org), [d.schissler@computer.org](mailto:d.schissler@computer.org)  
Phone: +1 508 394 4026  
Fax: +1 508 394 1707

Southwest, California:  
Mike Hughes  
Email: [mikehughes@computer.org](mailto:mikehughes@computer.org)  
Phone: +1 805 529 6790

Southeast:  
Heather Buonadies  
Email: [h.buonadies@computer.org](mailto:h.buonadies@computer.org)  
Phone: +1 973 304 4123  
Fax: +1 973 585 7071

### Advertising Sales Representatives (Classified Line)

Heather Buonadies  
Email: [h.buonadies@computer.org](mailto:h.buonadies@computer.org)  
Phone: +1 973 304 4123  
Fax: +1 973 585 7071

### Advertising Sales Representatives (Jobs Board)

Heather Buonadies  
Email: [h.buonadies@computer.org](mailto:h.buonadies@computer.org)  
Phone: +1 973 304 4123  
Fax: +1 973 585 7071