Martin Hilbert on the world's information capacity

| Tweet | 0 | Like | 0 | g+1 | 3 |

April 22, 2014 | Luke Muehlhauser (https://intelligence.org/author/lukeprog/) | Conversations (https://intelligence.org/category/conversations/)

Martin Hilbert (http://www.martinhilbert.net) pursues a multidisciplinary approach to understanding the role of information, communication, and knowledge in the development of complex social systems. He holds doctorates in Economics and Social Sciences, and in Communication, a life-long appointment as Economic Affairs Officer of the United Nations Secretariat, and is part of the faculty of the University of California, Davis. Before joining UCD he created and coordinated the Information Society Programme of United Nations Regional Commission for Latin America and the Caribbean (http://www.cepal.org/SocInfo). He provided hands-on technical assistance to Presidents, government experts, legislators, diplomats, NGOs, and companies in over 20 countries. He has written several books about digital development and published in recognized academic journals such as Science, Psychological Bulletin, World Development, and Complexity. His research findings have been featured in popular outlets like Scientific American, WSJ, Washington Post, The Economist, NPR, and BBC, among others.

Luke Muehlhauser: You lead an ongoing research project (http://www.martinhilbert.net/WorldInfoCapacity.html) aimed at "estimating the global technological capacity to store, communicate and compute information." Your results have been published in *Science (http://www.sciencemag.org/content/332/6025/60)* and other journals, and we used your work heavily in The world's distribution of computation (https://intelligence.org/2014/02/28/the-worlds-distribution-of-computation-initial-findings/). What are you able to share in advance about the next few studies you plan to release through that project?
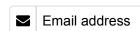
Martin Hilbert: When we first started out, we were rather surprised how little work had been done in the area of quantifying our information and communication capacity. We have statistics about everything and know how many cars and trees there are, and have estimates about the social and economic impact of shoe sales and carbon exhaust, but living in an information age, only very few pioneering studies have been done about how much information there is[1] . We felt the topic would deserve a more coherent treatment. So we set up three basic stages:

First, creating the basic database: how much is there? How much is stored, how much communicated, how much can we compute? This was by far the most tedious part and resulted in a 300 page methodological appendix, where we list the more than 1,100 sources and databases we combined to create these numbers   (http://www.martinhilbert.net /LopezHilbertSupportAppendix2012.pdf). We found some interesting things here, such as the fact that our computational capacity has grown between 2 to 3 times faster than our information and communication capacity since the 1980s. This is not only good news for the machine intelligence community, but also for human kind as a whole: while it currently seems like we are drowning in an information overload that stems from the sustained 25 – 30 % annual growth of information storage and communication capacities, we should be able to make use of the computational power (growing at 60 – 90 % per year) to make sense and eventually tame of all of this information.

Second, how can we describe it? We found several surprising things here. Some of the most basic assumptions of the digital revolution literature appeared in a totally new light. For example, usually, it is assumed that the digital revolution has increased global communication equality. The problem with this conclusion is that it based on the head-count of digital devices and subscriptions as the main indicator, so since there are more phones now than in the 1980s (with a current mobile phone penetration of 90% worldwide), the conclusion usually is that equality must have increased. However, not all phones are equal nowadays. So looking at the distribution of communication capacities, we found (http://www.martinhilbert.net/TechInfoInequality.pdf) that communication capacity in 1986 was actually more equally distributed than in the 1990s and the 2000s! In the 1980s there were only fixed line phones, but everybody had "equally little". Afterward, the myriad of communication technologies increased the inequality among countries and within countries. Only very recently have we re-establish the pre-1990 equality levels in terms of our bits-capacity. In other words: while we are all much better off in absolute terms ("we all have more"), relative information inequality in terms of information capacity continuously opens up with each new innovation ("we are not automatically more equal"). The digital divide turns out to be a moving target! We do not yet have any idea

yet about the social, economic and long-term political consequences of this ever-changing inequality in information and communication capacities among and within countries…

Another traditional assumption of the digital revolution literature is that we now live in a multimedia age, with an unprecedented share of moving videos and audio sounds. Looking at the evolution of the content of the world's information and communication capacity, we actually found (http://www.martinhilbert.net /WhatsTheContent_Hilbert.pdf) that the relative share of text and still images captures a larger portion of the total amount than before the digital age! Text merely represented 0.3% of the (optimally compressed) bits that flowed through global information channels in 1986 but grew to almost 30% in 2007. Back in the pre-digital age, text mainly appeared on paper, while telephone channels were filled with audio (voice) and many homes hoarded vast amounts of video material in VHS libraries, etc. The proliferation of alphanumeric text on the web and in vast databases in a phenomena of the digital age. The fact the digital age turns out to be a "text and image age" is good news for big-data analysts who extract intelligence from more easily analyzable text and image data.

And as a last example, we were able to parse out how much of the global information and communication explosion was driven by more, and how much by better technology (http://www.martinhilbert.net /HowMuchMoreORbetter_Hilbert.pdf). We found that technological progress has contributed between two to six times more than additional technological infrastructure to our global bits capacity. While infrastructure actually seems to reach a certain level of saturation (at roughly 20 storage devices per capita and 2 to 3 telecommunication subscriptions per capita), informational capacities are still expanding quickly. We also found that additionally to progress in better hardware, software for information compression turns out to be an important and often neglected driver of the global growth of technologically-mediated information and communication capacities. We estimate that better compression algorithms alone allowed us to triple our communication capacity: in the 2000s we could send 3 times as much information through the same channel than int he 1980s, thanks to compression. This underlines the importance to measure information and communication capacities directly in bits and bytes. Traditional statistics provided by the national telecom or science authorities (such as the FCC or NTIA) merely count devices and subscriptions. But this indicator does not tell as much anymore.

As a natural third step after this rather descriptive work, we are currently working on deepening our understanding of the social, economic and political impact of this information and communication flood. The first question here is: impact on what? Per definition, a general-purpose technology (like digital technology) affects all aspects of human conduct, which gives us the free choice

for the area of impact. The common theme is that this social change was produced by information, so we have to involve the [bit-metric]. With it, we can measure economic impact as [US$/kbps] or democratic participation by [participation/kbps]. These kind of measures show us that somebody makes more or less effective use of the same communication capacity than somebody else. The other way around, we can also ask about [kbps/US$] and try to understand why some have more communication capacity while starting from the same economic resources. We can then fine-tune the [bit-metric] and analyze how the communication capacity relates to additional attributes of interest of the capacity itself (e.g. mobile or fixed; individual or shared; private or public; always-on or sporadic, etc.), or to different content. It will enable us to take a more systematic approach to ideas like the information overflow: how much of which kind of content, from which kind of technology has which kind of impact on what? How does the supposed curve of "decreasing returns to information" look like empirically and in which task? More elaborate indexes and models can even integrate an arbitrary combination of these variables with communication capacity, just as economist have come up with a myriad of ways to evaluate the distribution of monetary currency with a society. In the statistical analysis of economics the unifying ingredient is naturally $, while in the statistical analysis of technologically mediated communication the unifying ingredient is naturally the bit. Obviously, bits only say "how much", not "how good". Once we understand the impact of "more" or "less" bits, we can then even go on and ask about "better" or "worse" bits (or more of less suitable kinds of bits). The "better" or "worse" will appear as an unexplained "residual" in our impact studies. In other words: instead of actively defining what is good and what is bad, we corner it by at least taking out the co-founder which stems from "more" or "less". Our main argument is that "quantity" is the lower hanging fruit, and that it must precede any question about quality. Otherwise we will helplessly confuse more- with better- information, and the other way around. So any impact must be normalized on the amount = [impact / bit], and for this we need to start measuring bits and bytes. Which brings us back to the reason why we started all of this…

Luke: In the course of your research so far, which trends related to information and information technology have you found to be roughly exponential during a certain period, and which trends have you found to be *not* exponential during some period?

Martin: Social systems are too complex to make identify pure distributions and dynamics, but within these limitation, they are all "roughly" exponential. Machines' application-specific capacity to compute information per capita has

roughly doubled every 14 months over the past decades in our sample, whereas the per capita capacity of the world's general-purpose computers has doubled every 18 months. The global telecommunication capacity per capita doubled every 34 months, and the world's storage capacity per capita required roughly 40 months. Per capita broadcast information has only doubled roughly every 12.3 years, but still grows exponentially. One the one hand, this stems from pure technological progress, such as Moore's laws. On the other hand, this stems from the diffusion of the technology through social networks. This diffusion mechanism follows a logistic S-shaped curve, starting of with exponential growth until an inflection point, after which saturation converts the process in a reverse exponential. So actually we have two exponential processes (technological progress and social diffusion).

In a recent study (http://martinhilbert.net /Powerlaw_ProgressDiffusion_Hilbert.pdf) I've shown that these two exponential processes combined can result in an authentic power-law distribution among the number of technological devices and their performance (so-called power-laws, scale-free-, Pareto- or Zipf distributions consist of two exponential distributions). I showed this with the distribution of supercomputers: there are exponentially few supercomputers with exponentially large computational power, and exponentially many supercomputers with exponentially lower computational capacity. Both line up in an almost spooky order, such as if some kind of super organizer subscribes the U.S. Department of Energy to order one supercomputer with performance x, and Los Alamos Laboratory, IBM and a couple of universities to order some with exactly lesser performance $x^{-\alpha}$, etc. Of course, there's no such super-organizer, but social complexity leads to this stable order, grown out of two complementary exponential processes. Recognizing such social patterns can be useful, since it provides predictive insights into the evolution of highly uncertain technology markets.

Luke: Which research investigations would you most like to see (in this line of work) over the next 5 years, whether conducted by yourself or others?

Martin: One very useful contribution would be the continuous reporting of the growth and nature of our information stock and our informational capacities. Together with my co-author Priscila Lopez, we were able to create 20 year long time series, covering over 60 technological families, but this was a two person effort, mainly driven out of curiosity. We have not counted with the resources to sustain this effort continuously. The level of detail and the scope of the inventory should also be extended.

For example, for storage and communication we normalized the amount of information on the available optimal level of compression. This allowed us to gain insights on the amount of information, not merely the available hardware infrastructure (the same bandwidth can store/communicate different amounts of information, depending on how compressed the content). For the case of computation, however, we simply had to use MIPS, which is a hardware metric. Of course, during recent decades, computational algorithms also became more efficient. The same hardware can certainly solve several of the same problems much faster now than 20 years back. We didn't have the resources to go into this distinction. The continuous effort of recording the growth and the nature of our information capacity will surely become an important corner stone of understanding reality in a digital world, and is therefore indispensable.

Besides this empirical effort, I think it will be important that we deepen our theoretical understanding of the way social organization and social dynamics are currently being "algorithmified". Social procedures, routines, habits, customs, and also laws have always been the central corner stones of civilization. These are currently being digitized, some in a more rigid, others in a less rigid fashion. Big Data is important here, as are agent-based computer simulations, and all kinds of decision support systems. This leads to profound changes what society is made of. We still lack a deeper understanding of the strengths, opportunities and threats of this ongoing process of social creative destruction.

Luke: Lack of information can be a major barrier for this kind of research. Sometimes the data you want to collect was simply never recorded by anyone, or perhaps it was recorded but never released publicly. If you could get your wish, what would change about how data about information and information technologies was recorded and disseminated? Could these changes be executed at a policy level, or an industry level, or some other level, if there was enough of a push for them?

Martin: I think this the conception of the lack of records and data is wrong.

It is true that data on ICT could be improved, and for the past 15 years I focused a large part of my effort at the United Nations Secretariat on adding ICT questions into household and business surveys worldwide. We got quite far and have achieved important improvements in this regard.[2]

However, after years of often fatiguing international policy dialogues (just imagine: you are not the only one lobbying for including "just one more question" into the national household survey!), and considering that the

collective policy dynamic often leads to the lowest common denominator (often obsolete indicators…), I came to the conclusion that it might be easier to simply start creating an alternative database from scratch and to lead by example. That was the starting point of our undertaking that eventually included over 1,100 different databases, business records, and statistics. When I first proposed the idea of taking a 20 years inventory of the "The World's Technological Capacity to Store, Communicate, and Compute Information (http://www.ris.org/uploadi /editor/13049382751297697294Science-2011-Hilbert-science.1200970.pdf)", we received cynic remarks even from befriended (and very recognized) colleagues in the field. The reaction included concepts like "utopian megalomania" (this also led to the Acknowledgments in the Science publication that states that we thank "colleagues who motivated us by doubting the feasibility of this undertaking", p.65). However, there is more information on ICT out there than we think, and the very same digital age often allows us to come up with proxies that enable us for very good estimates. The Big Data paradigm is not to be underestimated. The Big Data paradigm encourages us to embrace messiness of unstructured data, to look for highly correlating proxies, and to make up for it with redundancy from complementary sources. The wealth of "incomplete/ unstructured/messy" sources in a Big Data world often trumps the lack of one clean and centralized source. This also accounts for "datafying" the very own Big Data revolution!

This being said, if I "could get my wish", of course it would be nice if eventually both lines of work would concur, that is, if the global statistical machinery would start to also consider "information" as something worth of measuring continuously. Until now the UN and others have started to consider it (e.g. see Chapter 5 in *Measuring the Information Society* (http://www.itu.int/ITU-D /ict/publications/idi/material/2012/MIS2012_without_Annex_4.pdf)), but not yet embraced the idea fully. However, I think for this to happen it will still need much effort and we have to be proactive and show (a) that it's possible; (b) that it's worthwhile; and (c) which kind of stats are useful and which ones are not (which is still subject to an open trial and error process (http://ijoc.org /index.php/ijoc/article/viewFile/1318/746.pdf))

Luke: Thanks, Martin!

---

1. See this Special Section: Hilbert, M. 'How to Measure "How Much Information"? Theoretical, Methodological, and Statistical Challenges for the Social Sciences. Introduction. (http://ijoc.org/index.php/ijoc/article/viewFile/1318/746.pdf)' International Journal of Communication 6 (2012), 1042–1055. ↵
2. Partnership On Measuring ICT For Development (http://www.itu.int/ITU-D

/ict/partnership/) – The Global Information Society: a Statistical View (http://unctad.org/en/docs/LCW190_en.pdf) ↵

Did you like this post? You may enjoy our other Conversations (https://intelligence.org/category/conversations/) posts, including:

> Mike Frank on reversible computing (https://intelligence.org/2014/01/31/mike-frank-on-reversible-computing/)
> Randal Koene on whole brain emulation (https://intelligence.org/2014/03/20/randal-a-koene-on-whole-brain-emulation/)
> Suzana Herculano-Houzel on cognitive ability and brain size (https://intelligence.org/2014/04/22/suzana-herculano-houzel/)
> Michael Fisher on verifying autonomous systems (https://intelligence.org/2014/05/09/michael-fisher/)
> … and many more (https://intelligence.org/category/conversations/).

Tweet 0   Like 0   g+1 3

0 Comments　　Machine Intelligence Research Institute　　　💬 Login ▾

Sort by Best ▾　　　　　　　　　　　　　　　　Share 🔗　Favorite ⭐

Start the discussion…

Be the first to comment.

ALSO ON **MACHINE INTELLIGENCE RESEARCH INSTITUTE**　　WHAT'S THIS?

Superintelligence reading group
13 comments • 4 months ago

AvatarGuest — The idea of humans
devising methods for directly
controlling superintelligent AIs

New report: "Corrigibility"
15 comments • 3 months ago

AvatarCharles Tintera — If I could modify
myself, I'd get rid of my sleep
cycle, no matter how "natural" it

Randal Koene on whole brain emulation
1 comment • 10 months ago

AvatarPandorasBrain — "I can't parse his
argument to connect with any
demonstrated evidence." is a

Will MacAskill on normative uncertainty
4 comments • 9 months ago

AvatarEdwardTBabinski — Moral
theories, Borda rules? Morality is
not math. There are some