KARK:
INTERNET

INTERNET

Christopher Kark
Stanford University

The Internet is everywhere, but nowhere in particular. The server farms dotting the North Dakota landscape, the tortuous web of fiber optic cables that straddle the ocean floor, the cellphone towers narrowly disguised as redwoods: these comprise a vast network that, for most of us, is amorphous. Space, in the geographic sense of the term, is passé. It's no longer a big deal that the Internet collapses geographic distances. Now, we spend our days swiping to the right. Now, it's all about parsing data, irrespective of where they originate. To grasp how such data are gathered and disseminated, to know how to manipulate them to one's advantage, is to wield a power fast disappearing. Yet we are, more starkly and nakedly than before, dependent on data to understand the world and ourselves. No longer can we count on our own wits to interpret all the information coming our way; there's far too much of it to take in. If we can speak of closeness on the Internet, it's about control over data—their collection, their distribution, their use. And while we have more on hand than ever, they always seem to escape our grasp (Zeynep Tufeski, «Engineering the Public: Big Data, Surveillance and Computational Politics,» *First Monday*, 7, 7 (2014), accessed February 5, 2015).

Information has, of course, abounded over the ages, appearing in cuneiform scratched on temple walls, in baskets of papyri piled in warehouses flanking the Alexandria docks, evened out by feverish outpourings one century and catastrophic losses the next. As early as the fifth century BCE, Plato wrote of Thamus, an Egyptian king, who upbraided the god Thoth (Hermes) for inventing writing because it would hamper our ability to remember, and consequently, destroy wisdom (Ryan Szpiech, «The Dagger of Faith in the Digital Age,» *Tablet*. October 7, 2014, accessed October 10, 2014). Yet precisely because it allows us to commit our thoughts to a medium that outlasts us, writing has proved to be one of the most potent technologies ever wrought. Indeed, it was the only way to record natural language until the invention of sound recording in mid-nineteenth century France. While the amount of writing (and now, audiovisual media) has mushroomed, our capacity to process it hasn't.

Such complaints about information overload resemble anxieties voiced as early as *Ecclesiastes* 12:12 and Seneca the Younger's writings in the first century CE (Ann Blair, «Information Overload, Then and Now,» *Chronicle of Higher Education*, November 28, 2010, accessed

December 30, 2014). Around the same time, Pliny the Elder took to compiling thousands of facts in his *Natural History* (77 CE). Centuries later, medieval scholars habitually bemoaned how much there was to read and how easy it is to forget. The invention of movable type the mid-fifteenth century substituted printed texts for the hefty tomes of previous eras. Printing was not only cheap; it was quick enough to meet the demands of increasingly literate populations. Formerly handwritten bestsellers flew off the printing presses just as new genres like the novel made private reading a pastime. Information overload as we know it came into being, as did the avalanche of content that now spills from our inboxes and timelines.

Velocity and volume are only two causes of information overload. No less crucial is how such information is organized. In medieval Europe, the push to preserve classical works for posterity raised the question of how best to sort them. Monks in monasteries scattered throughout Europe and scholars in Baghdad began to catalogue their collections over a thousand years before Google became a verb (ibid). By tinkering with ways to organize heaps of print, they cultivated the dull art of indexing that evolved into the card catalogues silently vanishing from today's libraries. In most places, they've been replaced by sophisticated query tools and dozens of subscription databases. The modern search engine is an extension of the card catalogue into the digital sphere, except that it also sorts every kind of data recognizable to query algorithms. What the average user understands as the Internet, then, is a nexus that weaves these data together into an ever-expanding whole. The result is an abundance of content that nobody can hope to consume in a lifetime.

Dwarfing even the biggest libraries of Antiquity, the amount of traffic on the Internet now totals more than zettabyte (1021 bytes), palpable in the burgeoning of content—or media readied for mass consumption—across websites, social networking platforms, and databases. Ours is a time of «content,» a time of self-published media featuring every kind of topic directed at every kind of person. There is far more content than time to consume it, let alone gauge its worth. A query about a celebrity meltdown can yield dozens of results that range from papers of record to budding blogs with a loyal, if small readership. A hashtag like «#JeSuisCharlie» on Twitter or Instagram might aggregate thousands of snippets about a developing news story. Photos accumulate, videos abound, and what begins with a single post snowballs into a sprawling web of media, threaded together by links and hashtags. Information overload is no longer just the scholar's nemesis, but a throbbing artery in our Zeitgeist.

Why is this the case? How, in the digital era, has technology altered our attitudes towards information—what comprises it, the parts worth our attention, the parts that aren't? What is the personal toll of information overload? With the ballooning amount of personal data available online, to what extent can we manage our reputations without incurring significant expense? Stepping back, how did we become accountable for all information published about us online in the first place? Assuming that we do bear these responsibilities, what tools exist to help us distill data into information, and information into knowledge? Too sweeping to address exhaustively, these questions have found solutions in Big Data platforms and reputation management services. But these solutions aren't answers: solutions need not account for causes. Often, it's not even in their best interest to do so. By contrast, causes are elusive and hard to quantify, but, when given due consideration, can illuminate shifts in social and cultural mores (Giorgio Agamben, «For a Theory of Destituent Power,» *Chronos Mag*, November 16, 2013, accessed January 3, 2015). Though information overload is nothing new, the burden

of responsibility for managing it in increasing amounts has never weighed heavier on us as it

does today. It began with the impact of digital self-publishing on traditional authorities.

The Internet has bypassed traditional gatekeepers who vouch for the quality of the material they publish. Ideas, both the illuminating and the incendiary, no longer face such barriers to publication. Anyone with a mobile phone and something to say can broadcast his or her thoughts to the world, joining millions of voices reverberating through cyberspace. The diminished importance of gatekeepers coincided with the rise of social media platforms, the boom of the sharing economy, and the vogue of making media in addition to consuming them. Hence the explosion of content ranging from the viral meme to award-winning citizen journalism. Authority has splintered, with the customer review and amateur demo on YouTube now carrying as much clout as expert opinion.

The era predating the Internet saw some fragmentation of authority with the explosion of cable television channels. But for anyone to broadcast one's ideas around the world in an instant was unthinkable—a pipe dream, even for the tech-savvy. For centuries, content distribution was limited to those with free access to a medium of communication. Only the literate could produce content in the first place; few would see theirs published on a large scale. As print publishing became widespread, presses and newspapers developed a set of practices to ensure quality by judging the author(s) and the content itself. The pen-wielding editor was born. This system of checks and balances remains in place: if a chief editor deems something inappropriate or inaccurate, it doesn't enter print until revisions are made. Likewise, to ensure that they meet audience expectations, movies and television shows undergo months of scrutiny before they debut.

These procedures guaranteed a measure of quality in most of the materials available for

mass distribution. By the same token, many books and films never saw the light of day because they voiced a controversial or subversive idea. The French philosopher René Descartes's works were censored throughout seventeenth-century Europe for the same reason most social media platforms are banned in China, North Korea, and Iran: ideas are powerful, as is the potential to reach millions with a few keystrokes. Outright prohibition is nonetheless the exception rather than the rule. Wherever freedom of expression is the norm, most people continue to look to traditional sources of authority in some cases. In others, they crowdsource. Teenagers, contrary to prevailing stereotypes, prefer their doctor's advice to recommendations from a reputable health website (Urs Gasser, Sandra Cortesi, Momin Malik, and Ashley Lee, «Youth and Digital Media: From Credibility to Information Quality,» *Berkman Center for Internet & Society*, 2012, accessed November 20, 2014). When it comes to news, however, we're prone to seek out publications that tell us what we want to hear. Meanwhile, we've thrown up a whole economy around the reviews (and reviews of reviews) of strangers when consulting Yelp to pick a *taquería* or Angie's List to find a carpenter who specializes in bamboo flooring (Andrew Flanagin and Miriam Metzler, «Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility,» in Digital Media, Youth, and Credibility, ed. Miriam Metzler and Andrew Flanagin, Cambridge, MA: MIT University Press, 2008, 5-28, 11).

We now rely on two modes of authority: one that filters information hierarchically, another that culls opinions from a thicket of unverified sources. In spite of claims that traditional gatekeepers have gone the way of the Dodo, they're just as revered as reviewers in the Yelp

Elite Squad. Hence the prosperity of legacy newspapers like the *New York Times,* as well as the vastly increased viewership of programs syndicated on subscription entertainment sites à la Hulu. What's changed is the premium placed on self-sufficiency. Public demands for more self-service have spurred a boom in automation, leading to the downsizing of support personnel across economic sectors. Rather than consult with a travel agent or call an airline directly, consumers flock to travel metasearch engines to book flights. Rather than confer with a stock broker, the casual investor might scroll through the blogs of a few finance gurus before rolling over her 401k. With greater self-service comes greater pressure on the public to seek out information for itself, a task made more cumbersome by the absence of physical cues online (David Lankes, «Trusting the Internet: New Approaches to Credibility Tools,» in Digital Media, Youth, and Credibility, ed. Miriam Metzler and Andrew Flanagin, Cambridge, MA: MIT University Press, 2008, 101-21, 104).

So we've banished the agents and experts who acted as intermediaries between our needs and the resources to satisfy them. In their place sit platforms that interface directly with those resources, but without the contextualized knowledge of a professional. Branded as covetous middlemen, such professionals are increasingly rare, whereas the customer review—carefully arranged with proprietary algorithms—is touted as a proxy of truth. Consequently, we place our trust in the opinions of strangers on an unprecedented scale. Sure, we might loyally cling to *NYT Now* for our daily news briefing, but we're liable to sift through hundreds of reviews on Amazon or a write-up in CNET before splurging on a Nikon Df.

A product rated at four-and-a-half stars is often enough to nudge a purchase, even if we haven't the faintest idea how credible that product's reviewers are. This is the burden of choice. Traditional authority abides—in this case, in ads—, but we can also leaf through scores of reviews before reaching a decision (ibid., 107). In the process, we're asked to make numerous judgments about the reliability of the opinions before us, accepting some, dismissing others, and ignoring those we don't want to hear. Sometimes, overwhelmed by the glut of information, we give up the search altogether. Yet the drive to call on multiple sources of information to make a decision or to learn about what's happening in the world isn't going away, and particularly as traditional authorities dodder.

More information self-sufficiency also carries the responsibility of control information about oneself. What you choose to watch or read is largely up to you; what others choose to watch or read about you online largely isn't. Naturally, reputation has always been a factor in human relations. At least half of plays of the plays staged during the Spanish Golden Age dealt with sanctimonious nobles on a warpath to avenge their honor. The still-popular joke about «pistols at dawn» evokes the absurdly genteel culture of the antebellum South. Today, however, nowhere is reputation more paramount than in concerns over online privacy, and with it, the opaque rating systems that undergird the on-demand economy. Career advice blogs admonish students and job applicants alike to rethink what they publish online, with each post leaving a footprint in a digital trail that may stretch over years. Such content may lie unnoticed for years until a college admissions committee or a hiring manager stumbles upon it. The consequences can be swift and their impact devastating.

Anxieties about privacy seem to clash with the on-demand economy's ethos of transparency, whereby everyone is a brand and star ratings are the rubric of quality. In question is not just how we should define privacy online, a debate

that, at any rate, will always outpace the legislative process. Privacy presupposes the right to hide a part, if not the whole of ourselves from the public gaze, unveiling only what we see fit.[1] Also at stake is distinguishing the person from the digital persona, a distinction hopelessly blurred as data become the measure of man (Farhad Manjoo, «Uber's Business Model Could Change the Way You Work,» *New York Times*, January 28, 2015, accessed February 16, 2015). In the years predating the Internet, there was scant information published about anyone. Of the little there was, most was factual (e.g., birth certificates, bank accounts, obituaries) and most could be checked against a reliable source. In short order, you could exhaust what the public could know about you.

That era has ended. As with celebrities in the pre-Internet years, your reputation now depends on what is said about you online (David Streitfeld, «Ratings Now Cut Both Ways, So Don't Sass Your Uber Driver,» *New York Times*, January 30, 2015, accessed January 31, 2015). Absent the barriers to self-publishing, hype and hearsay now compete freely with facts. Indeed, if the Tea Party movement proves anything, it's that facts themselves are now suspect. People have, of course, fallen victim to nasty rumors since time immemorial, but these were never available for the world to share and for search engines to cache. Rumors, once transmitted by word of mouth, warped by reinterpretation and cropped by memory lapses, would disappear into the ether as soon as everyone grew tired of hearing them. Few traces would remain to be trotted out later in a digital scarlet letter as they are today (Daniel Solove, «Speech, Privacy, and Reputation on the Internet,» in T*he Offensive Internet: Speech, Privacy, and Reputation, ed. Saul Levmore and Martha Nussbaum*. Cambridge, MA: Harvard University Press, 2010, 15-30, 16). Though new rumors would inevitably arise, these lacked the permanence of what show up when you Google your name.

As demands to guard one's online reputation grow, so, inexorably, has the burden of handling the data comprising it. We might spend hours—days, even—polishing our social media profiles into idealized avatars of ourselves, while at the same time untag ourselves from images of moments we regret posting (but perhaps not living), amp up the privacy settings on our accounts, or delete screeds posted to the blog we've neglected since Howard Dean screamed his way out of politics. Such digital self-fashioning is redolent of nobles preening themselves in the courts of Renaissance princes, where dissimulation, not certifiable fact, was the best defense against assaults on your reputation. Nowadays, you can hire reputation management firms to do some of the legwork for you. Like a discredited business seeking to make a comeback, you can (for a substantial fee) pay to scrub your digital footprints from search results, or at least relegate them to pages where nobody bothers looking (Evgeny Morozov, «Two Decades of the Web,» *Prospect Magazine*, June 22, 2011, accessed November 10, 2014).

Reputation has also become synonymous with quality in the on-demand economy, encoded in apps that provide consumer services—ride-sharing for a trip home, couchsurfing as a cheap alternative to hotels—and underwritten by an intricate rating system. The career of a ride-sharing driver hinges on this rating, as does the host of a room (or a couch) rented out to guests. To safeguard their jobs, drivers and hosts must strive to maintain a high rating at all costs, their careers weighed in the balance of aggregated customer reviews. Subjective and unverified though they may be, ratings like these are poised to expand beyond the on-demand economy to schools, hospitals, and, eventually, government, the *ne plus ultra* of datafication. It therefore makes little sense to speak of creating an online image as though it were somehow separate from us. Insofar as our persona is subject to ratings of one sort or another, the boundary separating

who we are online and who we are in real life is fast vanishing. The reputation of this convergent self rests on how much of the information about it we control, whether directly or indirectly. Amidst the tumult of data, maintaining a good reputation proves to be as cumbersome an endeavor as prying native ads from genuine journalism.

Beneath the surfeit of content, data have proved invaluable to those with the resources to analyze them. A century ago, gathering data was a manual undertaking. One of its earliest manifestations was public opinion polling. Polling grew in importance during the years following World War I, beginning when *Literary Digest* sent out postcards asking recipients who they believe would win the U.S. presidential election. With the notable exception of 1936, the returned postcards accurately predicted the outcome of each election between 1920 and 1932, signaling an interest in measuring public opinion objectively—an interest echoed in other countries with representative governments. It's no coincidence that Gallup, the consulting firm now synonymous with public opinion polling, was founded in 1935.

The information collected through Gallup polls was unrivaled for the time, but pales in comparison to databases pooled from websites and, increasingly, from the devices that nag us to go running or remind us that we spent way too much at Sur la Table. Aware that these data are far too abundant to parse manually, engineers have developed sophisticated procedures to surface patterns in them in the service of people and profit alike. These procedures, which read as litanies of if-then clauses, are commonly known as «algorithms.» Algorithms retrieve specific data from a database by following a list of well-defined computations. Although they've existed since the Persian mathematician Muḥammad ibn Mūsā al-Khwārizmī described them early in the ninth century CE, algorithms have become relevant in recent years with the advent of search engines and Big Data platforms. The same is true of colossal datasets gleaned from our click-through paths and health records, a momentous transformation in how we analyze human behavior.

Algorithms are invisible to the vast majority of people who use them. Paradoxically, this is one of the main reasons we trust them: we need ways to fish out information we want without dredging up tons of data we don't. The utility of data lies not merely in their collection but in the ability to retrieve meaningful information. Constantly updated by their proprietors, programmed to learn through user input, today's algorithms fuel research and organizational decision-making on an unprecedented scale. Consumers encounter them in browser advertisements, «recommended» movies to slot into their Netflix queue or items to toss into their Amazon cart. By sentencing unwanted email to their spam box, Gmail users can even teach algorithms how to discern spam from a relevant message (Evgeny Morozov, «The Rise of Data and Death of Politics,» *The Guardian*, July 19, 2014, accessed November 10, 2014).

At the same time, algorithms also fuel commerce and statecraft. Drawing on its vast store of consumer data, Walmart knew to flood its stores with junk food and flashlights in a hurricane-riddled region well before a destructive storm made landfall (Michiko Kakutani, «Watched by the Web: Surveillance Is Reborn 'Big Data,' by Viktor Mayer-Schönberger and Kenneth Cukier,» *New York Times*, June 10, 2013, accessed December 30, 2014). U.S. intelligence and special operations agencies have developed a target-selection program that assesses surveillance data to automate drone strikes, a decision that once fell to trained specialists (Cori Crider, «Killing in the Name of Algorithms: How Big Data Enables

the Obama Administration's Drone War,» *Al-Ja-zeera America*, March 4, 2014, accessed December 30, 2014). And those charged with designing these algorithms bring personal assumptions to bear on which categories they use to structure a database, which data go into these categories, and how such categories are interpreted (Tarleton Gillespie, «The Relevance of Algorithms,» *Culture Digitally*, November 26, 2012, accessed November 15, 2014). Each step in data collection and analysis lies exposed to personal bias, even if algorithms are seen as impervious to human error (Danah Boyd, *It's Complicated: The Social Lives of Networked Teens*, New Haven, CT: Yale University Press, 2014, 185).

More often than not, though, algorithms are the best hope we have for plucking something relevant from the welter of data. The Internet would be impossible to use without them, as would tracking down microfiche in a library flattened by a tornado. For all intents and purposes, the question isn't whether we need algorithms (we do) but what they can and can't reveal. This may come as unwelcome news to those already saddled by the heft of online content. But it

shouldn't come as a shock, as anyone who's used a search engine knows its limitations.

Every year, it's astonishing how much better search results become, so much so, in fact, that we scarcely have to venture beyond the results page to check the weather or compare airfares. Such tasks are what algorithms do well: in datasets with relatively straightforward properties—ones that vary slightly or not all over time—, algorithms are unbeatable (Gary Marcus, «Steamrolled by Data,» *The New Yorker*, March 29, 2013, accessed December 31, 2014). Still, in context-dependent systems like natural language processing or advanced translation, algorithms can only approximate exactness, and can often be wrong—or, at minimum, sound stifled. Algorithms can only parse data based on relevance and highlight correlations in datasets. They can't evaluate data for validity or soundness, nor can they assess the quality of those data or deduce causation (ibid). They certainly can't render moral judgement. Higher-level thinking, for the time being, still falls to humans.

A technical solution to the problem of excess data, algorithms reveal their shortcomings the moment pesky variables like context come into play. Context is rife with factors that defy predictable patterns. Hence the nuisance of ads that claim to anticipate our every desire by harvesting common words from our email and search queries. Since context is a human artifact, only the human mind is capable of grasping it. This distinction remains a pivotal one, a distinction characterized by much more than the threadbare man-versus-machine dichotomy. Technology only matters insofar as it is a means to human ends. Even with the help of all the automated tools that help us plow through information, we still must, at some point, be willing to think for ourselves.

Impossible though it may seem to buy a wearable health tracker or a self-driving car without pouring over every data point beforehand, we cope by understanding the nature of information itself. To assess content critically—the stuff of college humanities seminars, so misunderstood, still, by the way we think of the humanities in general—is far more potent than any algorithm. Anyone who understands the Internet as a dynamic ecosystem, and who understands, too, the ideological motives for creating and disseminating content, can weather any amount of it that comes her way.

So whether for laughs or for learning, it's important to know how to assess what we consume online and anywhere else. Scholars have taken to calling this skill set «media literacy,»

whose standard definition is «the ability to assess, analyze, evaluate, and create messages in a variety of forms» (Patricia Aufderheide, *Media Literacy: A Report of the National Leadership Conference on Media Literacy*, Washington, D.C.: Aspen Institute, Communications and Society Program, 1993, xx). Media literacy involves understanding content based on several criteria: who creates it, to whom it speaks, how it was created, what biases it contains, and how its creator wants people to think or act. The elements of classical rhetoric, these criteria help differentiate between a polished op-ed and a listicle spewed from a content mill, or distinguish the original version of a video from a half-dozen knockoffs. This not only enables people to take full advantage of content on the Internet, but to become its savviest consumers.

On social media, for instance, users often express themselves by sharing links that direct to content hosted anywhere on the Internet. But not all content is created equal: every time we read a blog post or watch a video posted on YouTube, we assess it. Frequently, it's for entertainment value, while at others we're looking for an expert opinion on something that matters to us. The sheer amount of content and the absence of traditional markers of validity can be bewildering and, as misunderstandings so often do, spark conflict.

More than making us sound intelligent, assessing media adeptly impacts how we think and behave. Without a sense of how or why they're

created, it's easy to mistake opinions for facts that shape our decisions (Flanagin and Metzler, 8; Gasser et al., 76-77; Lankes, 102). Yet the importance of searching for and evaluating information has never been greater. Public demands for self-service have spiked in recent years, leading to a spike in technological automation. And with direct access to information comes the challenge of finding what's relevant and evaluating it for ourselves.

Knowing how to appraise online content thoughtfully is not a given. Depending on which school they attend, students may receive a thorough grounding in media literacy or none at all (Boyd, 181-82; Harris, 156). This same learning gap branches into the adult world, with technology evolving at a pace so brisk that society struggles to keep up. For example, content on personal health can vary from a how-to guide sponsored by a major brand to a peer-reviewed journal article. Neither is necessarily better than the other, but it's crucial to be able to tell them apart. As with a research project for school, it may be hard to distinguish between various kinds of content and assess them accordingly. This can carry over into social media platforms, many of which thrive on sharing content from every corner of the Internet. Learning to distinguish different kinds of content may prevent misunderstandings and defuse conflicts. Moreover, it helps us learn how to evaluate information with a critical eye.

As solutions go, media literacy only goes so far. It teaches us to cope with information overload, not with data points that quietly determine our fate. We can speak in tongues of disruption; we can build all the Hadoop clusters we want; but without a robust debate about the ethics of data collection and use, information overload will be the least of our worries. For better or worse, the debate began over a de-

cade ago, but wouldn't reach fever pitch until Edward Snowden leaked evidence of the National Security Agency's widespread surveillance programs. The Snowden leaks exposed an ugly combination of data science and intelligence gathering. No sooner did Snowden flee to Russia than U.S. policymakers recoil by passing digital privacy legislation, with ten states ratifying no fewer than two dozen laws in short order

(Somini Sengupta, «No U.S. Action, So States Move on Privacy Law,» *New York Times*, October 31, 2013, accessed February 15, 2015). California passed three, two of which regulate website content removal, and another tracking signals on browsers. Oklahoma approved a bill protecting student data, while Delaware—riding the momentum of states before it—lowered restrictions on access to a dead person's data (Jacob Gershman, «Delaware Eases Access to Digital Data of Dead,» *Wall Street Journal*, August 20, 2014, accessed February 15, 2015).

At this time of writing, 15 states have enacted laws that require express consent or a search warrant to obtain user information. Others, like California, continue to introduce similar legislation in the face of veto threats (Philip Janquart, «Second Try at California Electronic Privacy Bill,» *Courthouse News Service*, February 10, 2015, accessed February 15, 2015). The fact that that it took a damning leak to galvanize lawmakers into action isn't surprising; the silence about the cultural shift it signals is. It's one thing to accept datafication as a natural order to which, once we download this app or polish that profile, we can adapt—maybe even thrive. It's another thing to consider where we've smeared our digital fingerprints and whether we have the right to know in the first place. After all, when we speak of data, we're always close by: for we speak of ourselves.

NOTES

1   Shaped by cultural currents over the centuries, privacy is an evolving concept. Legal scholars Saul Levmore and Martha Nussbaum ascribe four characteristics to privacy in its current iteration: seclusion, or freedom from public scrutiny; intimacy, or a space to share private information in a limited sense; secrecy, which restricts knowledge to a specific person or group; and autonomy, or choices made privately. See Saul Levmore and Martha Nussbaum, «Introduction,» in *The Offensive Internet: Speech, Privacy, and Reputation*, ed. Saul Levmore and Martha Nussbaum (Cambridge, MA: Harvard University Press, 2010), 1-11, 10.