Special issue on Data Literacy: Articles

Data Literacy defined pro populo: To read this article, please provide a little information

David Crusoe

Independent Researcher, United States
dave.crusoe@post.harvard.edu

Data literacy is of fundamental importance in societies that emphasize extensive use of data for information and decision-making. Yet, prior definitions for data literacy fall short of addressing the myriad ways individuals are shepherds of, and subjects to, data. This article proposes a definition to accurately reflect the individual in society, including knowledge of what data are, how they are collected, analyzed, visualized and shared, and the understanding of how data are applied for benefit or detriment, within the cultural context of security and privacy. The article concludes by proposing opportunities, strengths, limitations and directions for future research.

Introduction

The thermostat knows when its owner is at home; the lights adjust accordingly, as do the air vents. Baby socks talk to smartphones, and the smartwatch to an insurance company which sells its data onward to a marketing firm incorporated abroad. The proliferation of data-based and data-driven systems is undeniable, largely unregulated and, significantly, applied by a myriad of stakeholders to make significant and impactful determinations (Pasquale, 2015).

Indeed, increasingly, individuals are continuously stewards of, and servants to, their data. In the personal context, individuals use data, or the outputs of analyzed data, broadly, daily. For example, one might measure health and wellbeing, understand workplace trends and track expenses. Commercially, individuals are targets for advertising displays and campaigns, but also subjects of algorithms, "*a process or set of rules to be followed in calculations or other problem-solving operations* ..." (Oxford Dictionaries, n.d.), that use data to determine diverse opportunities, such as credit, career, health and insurance. Governments, in their turn, examine data about the individual and group, including social connections, to identify past or future potential undesirable behavior (Lyon, 2014). The individual's connection to data is

Crusoe, D. (2016). Data literacy defined pro populo: to read this article, please provide a little information. *The Journal of Community Informatics*, 12(3), 27-46.

Date submitted: 2015-12-17. Date accepted: 2016-06-13.

clearly sophisticated; and it is important, even critical, for the appropriate population to have at least a basic, broad understanding of their involvement in the "great data revolution" (U.S. Chamber of Commerce Foundation, 2014).

Data literacy is the instructional domain that addresses skills and competencies one might apply to thrive in a complex data world. Yet, as this study will illuminate, the definitions for data literacy are myriad and narrow. Over time, the definition may be expanding, and it may be time to unfurl it fully. To accomplish this feat, this study presents a new definition for data literacy that accurately represents how an individual, within society, comprehensively and continuously interacts with, and is engaged by, data.

This study begins by defining the relevant population (*populus*) and describing how the member of the *populus* are simultaneously stewards and subjects of data in their individual, commercial and governmental interactions. Next, this study presents the methodology and analysis for a review of current data literacy definitions that identifies shortcomings with current approaches. Such limitations are not unrecognized. Perrotta (2013: 118) proposed that data literacy "*... should be encouraged across schools*[1] *and local communities ... [to] involve elements of digital literacy, citizenship and varying degrees of methodological knowledge ...*" and Twidale et al. (2013) recognize the importance for "lay people" to increase their understanding.

Thereafter, this study addresses the population-level need by presenting a revised definition for data literacy. This definition is theoretical, and based upon observations, trends and applications of information and communication technology (ICT). It encompasses those skills, understandings and comprehensions a population at large may require and indicates what, specifically, might be addressed through education. Finally, this study concludes by presenting strengths and weaknesses related to this approach, and indicates where further research will be important.

The Webs of our Data World: Stewards and Subjects

Geertz (1973: 5) described culture as "... *webs of significance that he himself has spun …*" The *populus* for whom data literacy is relevant are those individuals who experience dataculture, or "webs of significance spun" from an individual's data by the individual him or herself, or by external social, commercial or governmental systems that act upon his or her data. Population membership is therefore dictated as much by external forces as by internal motivation, but isn't binary.

The total size of the *populus* is significant. Figure 1 indicates the penetration of communication technologies worldwide, including mobile, Internet, fixed-phone and broadband access utilization. While mobile phone penetration is high (80% ownership in the developing countries vs. 98% ownership in developed countries), internet access is mixed (31% in developing countries vs. "80% in high-income countries") across populations (World Bank, 2016: 6).

[1] Note that data literacy, as characterized by this work, differs significantly from the current economic, political, media and educational focus on pushing US and international primary and secondary schools to provide core computer science education.

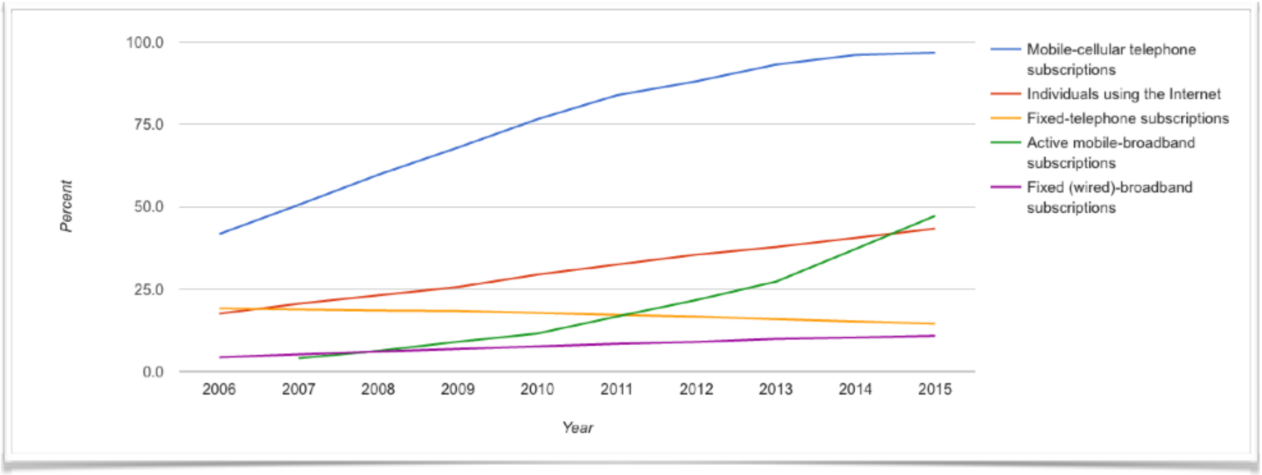

Figure 1: Global ICT developments, 2006-2015

Those individuals with higher levels of connectivity are most likely to comprise the *populus* though the extent to which an individual experiences dataculture may vary greatly. Consider the following:

- An adult who is not internet-connected, but who utilizes a spreadsheet to quantify agricultural data;
- An adult who is internet-connected, and shares agricultural data with other adults;
- A youth who accesses the Internet from school alone, does not use social networks, but who lives within a governmental and commercial system in which his health, citizenship and a myriad of other personal information is stored in digital and networked formats;
- An adult who does not use or access the web, but whose health information is stored by her doctors in a networked, digital format; and whose information is tapped by insurance companies to make decisions about medical care;
- An individual who pays for items with a credit or debit card, or who has, and utilizes, a personal library borrower's card;
- An urban adult who is digitally-engaged and constantly communicating through video, text and audio through her mobile device; who posts to social networks and has installed an internet-connected heating/cooling device that she manipulates from her mobile device.

These personas illustrate possible variance for the members of the *populus* engaged in dataculture, and for whom data literacy may be important. Individual participants can be characterized by some modicum of personal choice; but the vast majority of their participation is through their experience in the interconnected personal, social, commercial or governmental systems that collect, value and apply data or the abstraction of data for an end.

The volume of data produced by, about and for the *populus* is astonishing. In 2014, it was estimated that daily, humans captured 2.5 *quintillion*[2] bytes of data about "*... our words, system states, physical locations and personal interactions*" (Hayashi, 2013). In the resultant dataculture, humans are both stewards–responsible parties–and subjects–topics of, and controlled by, data systems.

As Stewards

Personal stewardship of data is a commonplace artifact of dataculture, and includes routine creation, management and mundane consumption. Workplace tasks, for example, include creating, managing or consuming data and data-based information through reports, spreadsheets or data-centric records. In home life, one might purchase clothing to monitor an infant's vital signs (Chernova, 2014). Such a device creates data about one's child, and provides a reporting interface and the means through which a user might recognize patterns or problems. The use of a pedometer is similar: movements result in the creation of data, which can be managed and consumed through a predetermined interface. Especially expert users can, at option, export and further manipulate their datasets[3].

In these and similar cases–of which there are too many to enumerate–human skill, knowledge and understanding is applied to create, manage and analyze data, and to consume resultant information. Most applications are relatively non-expert. That is, the responsible person must know how to power-up a device; utilize the device as needed to track what's desired; access the location for the stored data; make sense of the data, as provided by an interface; act responsibly to ensure the security and privacy of the data as desired; and understand, if possible, how else those data may be utilized, shared or applied.

As Subjects

As subjects, we are enmeshed firmly, perhaps beyond accurate perception, within webs of dataculture. Commercial entities utilize data as a means to increase efficiency, productivity and sales; governments utilize data for the same, and additionally, to govern. Let us turn first to commercial enterprises, within which are sectors as diverse as health care, children's toys and farming.

In writing about data and analytics, Spencer (2015) outlines four commercial applications for data and analytics as they relate to the external consumer: targeted advertising, price discrimination (see also: Danna et al., 2002), customer segmentation and eligibility determination.

Targeted advertising is advertising that matches specific audience variables with content so as to maximize reach and effectiveness. While delivered heavily through the web, the web is by no means the only way that advertising is generated to target an individual. Increasingly, for example, marketers utilize consumers' active wifi, Bluetooth and other technologies to

[2] In relative terms, this equates to 2.5 billion gigabytes: a very large number of stored instances about, for example, who sent whom which cat meme, when or, perhaps, how many times one's pedometer measured noticeable site-stand motion during a normal workday.

[3] There are interesting individual cases of "expert" data literacy; for example, of an individual who exported GPS data to successfully contest a traffic ticket: See https://skatter.com/2011/02/how-my-smart-phone-got-me-out-of-a-speeding-ticket-in-traffic-court/

identify and track movements, including to measure a consumer's location within a store, the time a consumer might spend in front of a particular display, whether a consumer removes an item from a display and whether or not the removal results in a closed sale (Turow et al., 2015; Rouhollahzadeh, 2001; Decker et al., 2003).

Price discrimination is the practice of "t*he sale of two or more similar goods at prices which are in different ratios to marginal cost*" (Stigler, 1966 as cited in Danna, 2002: 380). Amazon's inverse pricing of products based upon customer loyalty, through which loyal customers received higher prices (Chen, 2005), is an example of price discrimination. More recently, Orbitz was found to use customers' operating system data to generate hotel options and prices; the algorithmic assumption was that Apple users might spend more per room (Mattioli, 2012). And, Amazon currently faces a class-action lawsuit that alleges the retailer raised the price of Amazon Prime members' goods to cover members' "free expedited shipping." (Duryee, 2014).

Price revision is a variation of price discrimination, in which the initial price for a good or service sold to a consumer is revised based upon the consumer's later reported activity. Insurance giant John Hancock provides some members with tools to track and report their fitness levels as a means to gain "vitality points." In turn, the insurer utilizes "vitality points" to determine potential insurance discounts (John Hancock Insurance, 2015); thus John Hancock offers data-driven price revision. Allstate Insurance, a car insurance provider, offers its customers a "*telematics program" that that records "... information about your driving habits* …" (Allstate, 2014) to offer discounts.

Customer segmentation is the practice of creating specific groups of customers based upon demographic profiles. Like targeted advertising, it stems from the construction of demographic-based stereotypes, including, but not limited to, gender, ethnicity, social media use or weight. These constructs are segments for which products or product lines can be customized and marketed.

Finally, eligibility determination is the practice of utilizing data to understand whether an individual or customer segment may quality for an opportunity. These opportunities may range from the benign, such as a coupon, to the impactful, e.g. a job opening. In the construct of commercial use of data to determine consumer outcomes, as I will later describe, eligibility determination is potentially one of the more nefarious tools in play.

These four constructs illustrate methods by which the *populus*, as non-experts, are subjects of commercial dataculture. These webs are ubiquitous, myriad, visible and invisible, and indicate the very significant nature of what one might need to know so as to navigate a digitally-mediated life.

As a consumer, for instance, one must be conscious of what is shared, with whom, under what ownership agreement, how, and of what might result. By sharing demographic information, one might unconsciously place oneself at disadvantage when commercial algorithms dictate that certain opportunities are not valuable or appropriate to share. Or, one might take best advantage of price revision through personal tracking to access lower health insurance rates. There is little commonality across commercial constructs; a demographic that might be of benefit through one application could, potentially, become a liability elsewhere.

Within the corporate context, there is one additional element to add as it relates to the external consumer: that of data sharing. Data are not utilized by the originating agencies alone; they are packaged and sold, or can be stolen and sold. Data brokers, or entities that purchase, package and resell data, constitute a "multi-billion dollar industry" (U. S. Senate Committee on Commerce, Science and Transportation, 2013: 3) that is essentially unregulated[4]. Brokers craft segments as telling as "American Royalty," "Rural and Barely Making It" and "Mid-Life Strugglers: Families" (U. S. Senate Committee on Commerce, Science and Transportation, 2013: 24). These data are sold to firms that utilize the sets consistent with Spencer's (2015) construct and, as this study will later describe, to potentially detrimental ends.

Commercial data can also be stolen and shared. In 2015, 781 reported U.S. data breaches compromised an estimated 164.4 million social security records and 800,000 debit or credit records (Identity Theft Resource Center, 2016). These corporate records are packaged and sold through online black markets; purchasers include "... individuals, criminal organizations, [and] commercial vendors" (Ablon et al., 2014: 5). Therefore, the corporate data sharing can be enacted purposefully, consistent with terms and conditions a member of the *populus* may have agreed to so as to access a service or good, or enacted accidentally, through a data breach and subsequent black market distribution.

There are, of course, also uses for data within a corporation by a person; that is, that within a corporate context, a member of the *populus* might be required to read or write data as part of workplace expectations. Certainly, this is commonplace. Survey Monkey reported over "... 1*4 million free users, 360,000 paying customers and 65 million monthly visitors to its website* …" in 2013 (Savitz, 2013) to include both survey creators and survey takers, and Microsoft reports 1.2 billion Office users (Microsoft, 2014); this provides only a snapshot for how pervasive the non-expert utilization of data creation, management, understanding and display might be.

Members of the *populus*, however, are not only subjects to corporate data use; they are subject, also, to the use of data by governments. Government priority is to govern, and to do so, elements of government capture data as diverse as income, travel, driving, or criminal activity, not to mention demographic or health information.

These datasets are not going untapped. Governmental requests for data from providers including Twitter, Facebook and Tumblr are rising year-over-year[5], and many providers do not yet provide any window into government use and/or access of their members' data. One of the largest data brokers, Acxiom, contributed to the more than 200 million records sought by U.S. government agencies. Acxion, however, does not provide details analogous to those provided by Twitter and others. Freedom of Information requests indicate the close tie between

[4] Consumer data may be subject to agreements, including terms and conditions and privacy policies, agreed to when accessing a good or service. Whether or how such policies prevent, limit or shape information sharing, and an evaluation of whether such policies are followed by corporate actors, is beyond the scope of this study. The U. S. Senate Committee on Commerce, Science and Transportation (2013) and Pasquale (2015) are indicative of the probable state of affairs.

[5] Twitter reports data requests at https://transparency.twitter.com/; Facebook requests are available at https://govtrequests.facebook.com/; Tumblr's transparency report is available at https://www.tumblr.com/transparency .

commercial aggregators and government[6]. The bonds between commerce and government is tight; as Pasquale (2015: 46) writes, commerce enables "*...the government, in the name of 'information sharing,' to supplement its constitutionally constrained data-gathering activities with the unregulated collections of private industry*."

Furthermore, government utilizes predictive algorithms to identify those who may commit crimes (Citron et al., 2014: 4) and, within education, personalized academic predictions based upon past performance are growing in scope and scale. As Perrotta (2013: 117) writes, data "*... appear to offer the promise of accurate predictions, personalized recommendations and dramatic increases in the efficiency and effectiveness of [an educational] provision.*"

Thus, a population needs to understand not only that their data are captured by governments, but that their data, including their personal and commercial data, might be utilized for purposes of governing.

The Webs of our Data World: Imperfections

As stewards and as subjects, one's datacultural experience is further complicated by biases and imperfections. As Citron et al. (2014: 3) write, "*... private and public entities rely on predictive algorithmic assessments to make important decisions about individuals*." Data-based decisions are only as perfect as their underlying elements.

Centrally, data systems are only able to make decisions they have been algorithmically programmed to make, based upon instructions that humans have created by developing theories about what will be, based on what was (Citron et al., 2014). Saurwein, Just and Latzer (2015: 37) identify nine risks related to the application of data to make decisions, including "*manipulation, diminishing variety, limitations on freedom, surveillance and threats to data protection, discrimination, validation of Intellectual Property (IP), abuse of market power, effects on cognition and loss of human sovereignty*" . Van Wel et al. (2004) hypothesize similar deleterious impacts of profiling and data mining, including de-individualization, deleterious impacts to data privacy, locus of control over data, the creation of harmful data about a person through the synthesis of multiple datasets, and questions about data fidelity and accuracy. In the following section, I will touch upon these classifications as groups and briefly illustrate why each is not merely a perceived risk, but in fact, a concern of import to the population today.

Research and popular media indicate these risks as well-founded and applicable to lived experience. Algorithms, at their outset, are logical bits or mathematical models prone to human error in construction or, worse, direct manipulation. Even well-conceived predictive models applied to directing marketing or other resources are accurate only in a small percentage of all cases (Boire, 2013) and most are entirely inaccessible to independent audit (Sandvig, 2014). Techno-optimism is founded not on the certainty of what algorithms indicate, but instead upon incrementally more accurate models than earlier methods provided.

[6] See, for example, Freedom of Information Request to DARPA as modified by February 5, 2003 letter to Directorate [Letter written January 23, 2004 to Mr. David L. Sobel]. (n.d.). Retrieved March 23, 2016, from https://epic.org/privacy/profiling/tia/darpaacxiom.pdf as cited in The Surveillance-Industrial Complex: How the American Government Is Conscripting Businesses and Individuals in the Construction of a Surveillance Society (Rep.). (2004, August). Retrieved March 23, 2016, from The American Civil Liberties Union website: https://www.aclu.org/files/FilesPDFs/surveillance_report.pdf.

This understanding is critical for a population to know, if merely factually, as non-experts, and essential to the application of data in making decisions for, and about, people, as described earlier.

Specifically, fears about de-individualization and digital discrimination are not unfounded. "*... people are not being treated as individuals capable of making a rational choice in their own interest*" (Danna et al., 2002). In fact, predictive analytics treat people based upon patterns of their group; this practice violates the premise of individual agency within a society (van Wel et al., 2004). Citron et al. (2014: 15; well-described by Knutson, 2002) cite a case in which Allstate Insurance was found to be relying on a credit scoring algorithm that automatically placed non-whites in higher-premium categories. The plaintiff succeeded, and Allstate must now provide some transparency into how it develops its calculation. Nonetheless, this is an exception.

Furthermore, algorithmic approaches dictate far more than access to credit or the price of goods. In their work on data and civil rights, Rosenblat et al. (2014) describe how police departments utilize algorithmically-assessed data to target neighborhoods for aggressive policing, explore judges' use of data reporting to develop sentences for criminal behavior, and that tools used by police to identify offenders is prone to error. As U. S. Attorney General Eric Holder described,

> ... basing a sentence on something other than the conduct of the person involved and the person's record ... factors like the person's education level, what neighborhood the person comes from … [judges are] using this as a predictor of how likely this person as an individual is going to be a recidivist. I'm not at all certain that I'm comfortable with that … I think the result is fundamental unfairness. (Holder, August 1 2014).

Nonetheless, systems are built upon opaque decisions; these decisions impact people across society. Yet having too much data is not the only problem, having too little can hurt access too.

Being "thin-filed," or not having significant past data required by a vendor that utilizes data to make a decision, can also lead to detriments (Wessels, 2015). For example, access to credit requires a credit history, which is established by obtaining and managing credit. In California in 2007, amendments to public utility legislation allowed ⅔ of ineligible "thin-file" consumers to access credit by allowing the use of their utility payment history for credit assessment (California, 2007). The downsides to consumers are not limited, however, to what's done with data; detriments also originate from where data are stored, and how they are shared.

Data protection and privacy fears are well-founded. Data protection refers to the security and encryption applied to stored data; privacy relates to the availability of data to other parties, including individuals, corporations or others. A brief glance at the headlines makes clear how profoundly insecure data systems are, both in terms of protection (Open Security Foundation, n.d.) and privacy (U.S. Senate Committee on Commerce, Science and Transportation, 2013). Both are relevant, for the deleterious impacts of data "in the wrong hands" is evident. For example, medical data can be used to create faulty medical bills to defraud insurance or other companies (Krebs, 2014), and tax information has been used extensively to defraud the U.S. Internal Revenue Service through false returns (Internal Revenue Service, 2014). In sum, the security and privacy of one's data is paramount and one must live with the expectation that, at some point, data will be compromised. Thus, knowing which additional steps to take is simply a part of current life experience.

As this study has illustrated, data use by individuals, commercial interests, researchers and governments is pervasive and can bring about significant deleterious impacts that require knowledge to understand, if not overcome. Others (e.g. Pasquale, 2015) have proposed legal or regulatory approaches to alleviate related challenges. Next, we explore the existing definitions for data literacy that will make very clear the need for a revised, population-level educational approach to the same.

Definitions for Data Literacy

Methodology

This study is based upon the analysis of publications gathered through an extensive survey of related literature. Specifically, materials were accessed through Academic Search Complete, Google Scholar, JSTOR and other indexes June through December 2015. Resources were identified by using "data literacy," "data base [sic] literacy," "statistical literacy," "data instruction," "big data," "data privacy," "open data," "big data literacy" and "data security" as keywords. Irrelevant but related terminology, including "information literacy" and "digital literacy" were checked, but did not lead to immediately relevant resources. Searches uncovered a range of subject foci (education, ethics, business, medicine) and topical foci (data practices, security, primary and secondary education, university education). News items and other supplemental materials were surfaced through searches using Google's search engine.

The initial search yielded a significant breadth of articles, though those exploring "data literacy" specifically, rather than related practices (like security or privacy) were more limited in number. Of them, sixteen sources were selected for their detailed analysis of data literacy, and for their status as a reference to other sources. Many articles' authors adopted a similar definition approach (see "Current Definitions" below) and only one article, Mandinach et al. (2012) presented an exhaustive analysis and review of related discourse. To limit scope, those articles deemed most relevant to the discourse were ultimately incorporated into this discussion.

Limitations

This methodology precluded a full, qualitative literature review for all existing definitions and their derivation. Thorough, generalized reviews are limited in number. The latest review, the result of a sponsored research group, was conducted in 2012 and focused primarily on defining "*... what it means to be data literate in education*" (Mandinach et al., 2012: 1). Specifically, its intent was to define the term and related skills for school administrators and educators. This group did not report consensus for the definition of data literacy, nor did it conduct a domain-general review of the use of the term. Such a review may be needed.

Furthermore, the contents of this article are limited by time. The origins for digital terminology, including "data literacy," can be obscured by the slow move to digitize and index articles and other periodicals originally published in print alone. Historical indices may be expanding, but do not yet appear comprehensive[7].

Finally, research is published frequently. Therefore, some limitation in this review is due to the nature of when searches were conducted, and the possibility for technical advance or publication since.

Analysis

Definitions are important, even if tedious, because they are used to build and evaluate underlying theories and assumptions; they may define approach or practice; or they may lead to intended, or result in unintentional, outcomes (Pellegrini, 1992). Educators and others turn to definitions to form the basis of approach for teaching and learning (Mandinach et al., 2012).

As a term, data literacy originated in literature that discussed the application of "data bases" [sic] to inform educational decision-making (Burstein, 1983) and "... *how a company acquires data, and the contents and meanings of the data, as well as its translation into information …* " (Hartnett Jr., 1989: 21). As technologies advanced, it was applied to describe the educational need related to government use of graphical information systems (National Research Council, 1997).

Modern literature reflects the historical underpinnings for the application-centric focus, and there is strong commonality amongst definitions. The strong commonality may be due to a small number of original working definitions that have evolved in citation; for example, Erwin (2015) utilized Gunter (2007), who in turn utilized Schield (2004).

Table 1 illustrates "central" definitions that serve as source material for others' work. These central definitions share characteristics of "finding," "manipulating," "managing," "interpreting" and "applying" data so as to take action. While relatively consistent in definition, target audience does differ. For example, several authors have a primary and secondary student body in mind; others write for those seeking specialized degrees.

Table 1: Comparison of definitions for Data Literacy

Source	Definition	Audience
Schield, M. (2004: 8)	"... accessing, converting and manipulating data…"	College students

[7] For example, the resource Karten, Naomi. "Upload, Download," *Information Strategy: The Executive's Journal*, 4: 36-32 (Spring 1988) may contain an early reference to data literacy, but the resource was unavailable digitally or through interlibrary loan.

Vahey, P., Yarnall, L., Patton, C., Zalles, D., Swan, K. (2006: 1)	"... formulate and answer questions using data as part of evidence-based thinking; use appropriate data, tools, and representations to support this thinking; interpret information from data; develop and evaluate data-based inferences and explanations; and use data to solve real problems and communicate their solutions… ."	Primary and secondary students
Carlson, J., Fosmire, M., Miller, C. C., Nelson, M. S. (2011: 5)	"... what data mean, including how to read graphs and charts appropriately, draw correct conclusions from data, and recognize when data are being used in misleading or inappropriate ways… ."	College students
Harris, J. (2012: 1)	"... competence in finding, manipulating, managing, and interpreting data, including not just numbers, but also text and images."	Business
McAuley, D., Rahemtulla, H., Goulding, J., Souch, C. (2012: 53)	"... ability to identify, retrieve, evaluate and use information to both ask and answer meaningful questions… ."	Higher education
van't Hooft, M., Vahey, P., Swan, K., Kratcoski, A., Cook, D., Rafanan, K., Stanford, T., Yarnall, L. (2012: 20)	"... the ability to formulate and answer data-based questions; use appropriate data, tools, and representations; interpret information from data; develop and evaluate data-based inferences and expectations; and use data to solve real problems and communicate their solutions."	Primary and secondary students
Perrotta, C. (2013: 3)	"… digital literacy, citizenship and varying degrees of methodological knowledge, which together arguably represent a crucial '21st century skill' for a more active and informed participation not only in education, but in many other domains increasingly characterised by pervasive data collection and manipulation… ."	General
Deahl, E. S. (2014: 41)	"... the ability to understand, find, collect, interpret, visualize, and support arguments using quantitative and qualitative data."	Primary and secondary students
D'Ignazio, C., Bhargava, R. (2015: 2-3)	"reading data," "working with data," "analyzing data," "arguing with data" and incorporate three "big data" dimensions, including "identifying data collection," "understanding algorithmic manipulation" and "weighing ethical impacts."	Nonprofits in the social sector
Koltay, T. (2015: 403)	"... access, interpret, critically assess, manage, handle and ethically use data… .“	Researchers and specialists

Over time, it appears that breadth has increased to encompass aspects of dataculture. For example, Carlson et al. (2011: 5) include a dimension of truth recognition; "*... recognize when data are being used in misleading or inappropriate ways...*". Perrotta (2013) and D'Ignazio et al. (2015) begin to capture the experience of dataculture. Specifically, their elaborations for data literacy make specific mention of the personal, commercial and governmental application of big data collection, algorithmic sophistication and related ethical quandaries.

Of all approaches, these three begin to capture the elements needed to suit the population need. Specifically, a foundation for data literacy should be comprised of the essential human understanding made necessary by the growth of data systems that translate the messiness of existence and experience into exacting, if voluminous, data that reside within a digital system, and in turn, that various human-programmed digital systems reinterpret and reapply to inform or impact the lived human experience.

These needs include the interaction of an individual with data as a steward, that is, a creator, manager and consumer, and as a subject; an individual whose experiences within society are, in part, dictated by the data and by data cultural systems with the strengths and limitations extensively described earlier in this work. Therefore, I will redefine data literacy *pro populo*.

A Needed Definition for Data Literacy

Data literacy is the knowledge of what data are, how they are collected, analyzed, visualized and shared, and is the understanding of how data are applied for benefit or detriment, within the cultural context of security and privacy.

As one may observe, this model promotes a *less expert* outcome than preceding definitions strive to attain. The term "understanding" is used intentionally to reflect the broad, daily and constant "lay" (Twidale et al., 2013) or "*populus*" application of skills, knowledge and competence. This is to say that while population individuals will most certainly be stewards of, and subjects about, data, the common call, as I have outlined earlier, is for a relatively unsophisticated level of application. Thus, the proposed definition weaves a wider net around core components of expert data literacy, including the skill to find, manipulate, manage, interpret and act upon data. These six points will not be described in detail.

Firstly, the definition begins by highlighting the importance for knowledge of *what data are*. Knowledge of *what data are* is intended to highlight the exceptionally broad data-sourcing as it relates to all of the things one might know and capture. Key concepts include:

- Data can be captured about many aspects of life, and are not always explicitly numeric in origin, but may be converted into numeric representations for purposes of later analysis. Location tracking mechanisms convert geographic position into numeric representations, to be sure; but location itself is the data. Similarly, a series of images are data in the form of facial features; these facial features may later be analyzed algorithmically to determine if they represent the same person.
- Data are small, but interrelated in that when single-point data (such as a baby's current temperature, when captured with a digital sock; or a patient's temperature in a hospital, when captured with a plastic-covered thermometer; or the political climate by way of an online survey) is stored, it is stored with other information, such as name, birth date or an identification number, to ease later use.

The second component of this definition emphasizes a knowledge of *how data are collected*. How data are collected presents the opportunity to explore more traditional data collection forms, including through surveys or, say, a mailing list. Yet it also opens the opportunity to explore the pervasive data-collecting technologies that, knowingly or not, individuals come into contact with daily. These include mobile phones ("smart" or not), traffic light cameras, credit or ATM card systems used to make purchases, smartwatches, fitness trackers, web technologies, including advertising tracking mechanisms, the internet provider or even government tracking systems. Core elements of knowledge about *how data are collected* include:

- Individuals can act as stewards, to explicitly identify a data source to use for collection purposes.
- Individuals are also subjects; systems that we can and can't perceive collect a vast quantity of data.
- Broadly, members of the *populus* may interact with data through data intermediaries, or actors that unlock value to "effective use" (Gurstein, 2011; van Schalkwyk et al., 2015) by a less-resourced population.
- Ownership over data has not been fully established in case law (Pomerantz et al., 2015), but there are a variety of legal approaches that differ by culture.

Thirdly, this definition incorporates a need for knowledge about how data are analyzed and visualized. These components invite the individual to understand that data are acted upon statistically, to glean trends, patterns or other results of import. Further, it implies that results of statistical analysis may be represented visually, so as *to tell a story* for the purposes of demonstrating what statistical analysis has revealed. In turn, core elements of knowledge about the analysis of data include:

- One might possess some knowledge for how simple datasets can be acted upon, as through a spreadsheet, to glean basic insights about mathematical operations learned in primary and secondary school. This may include integration with basic statistical and computer literacy competency.
- It is also essential to know, conceptually, that sophisticated descriptive and predictive analyses can be brought to bear on data.
- That infomediaries may be available to assist in learning about, or analyzing, data, including through government initiatives, libraries or independent "hackathons, to facilitate understanding (Magalhaes et al., 2013).

And, for visualizing data, one might know:

- The process for converting data into simple visual representations, including charts and graphs.
- The means to evaluate reliability and veracity of information presented through data so as to understand if, and how, the display might be story-telling.
- The conceptual understanding that data displays are not merely limited to charts and graphs, but include a wide variety of visualization mechanisms.

Fourthly, the definition indicates the importance for understanding how data are shared. The implication is that some data are, indeed, stagnant; created and used by a single source. However, the necessary further understanding is that data are, indeed, bought, sold, shared and interrelated so as to derive additional value from them. Key conceptual elements include:

- An understanding that data have value to a variety of audiences, not only the originating data agent.
- An understanding that privacy policies govern data collection, and that subjects of data do have some–though varied–authority over how data are used
- An understanding that additional value can be derived from data through its relationship to other matched data.

Fifthly, the definition explores the benefits and detriments associated with data in society. It is clear that dataculture prides itself on the ability to find efficiencies through data use; to better, more rapidly understand a more complete picture than was possible without that data. Nonetheless, the reality is nuanced:

- There are a variety of challenges related to utilizing large-scale datasets to produce information (Busch, L. 2014).
- Algorithms, including descriptive and predictive, are utilized by the individual, and for commercial, research and governmental purposes.
- Algorithms are opaque to independent verification, and predictive algorithms only correctly describe a small percentage of cases overall.
- Algorithms utilize stereotypical user models (Konstan et al., 2012) or user matching models (Lops et al., 2011) to identify an individual by group characteristics; thus the individual is de-individualized from her or his own volition.
- Algorithms act upon primary and secondary data. Secondary data are those expected to correlate one behavior with another. For example, "... people who buy small felt pads that adhere to the bottom of chair legs (to protect the floor) are more likely than others to be good credit risks …" (Siegel, 2013 as cited in Hayashi, 2014, p. 36).
- The results of algorithms are utilized to direct opportunities and information to individuals, and the individual may gain or lose based upon what the algorithm determines; the individual should be aware of, and watch for, data-originating discrimination.
- Our data tell a story about us that we may not agree with, or which may paint an inaccurate picture of our self.

This proposed definition also implicates "the cultural context of security and privacy." Security, for our purposes, describes the nature of how the data are stored to repel unwanted access; privacy describes the control over who can see what about the data. And, as Perrotta (2013) rightly implies, data literacy must, by its population-impact nature, involve components necessary to evaluating how one's data are shared, stored and utilized. Context is important. The European Union and its citizens have pursued a far more stringent regulation regime for data use and privacy than their seemingly-lackadaisical counterparts, the United States Government and its citizens (Boehm et al., 2015). Therefore, while specific understanding may need to be adjusted to context, this definition embeds several core assumptions about privacy and security:

- Data stewardship requires responsible security and privacy measures, adjusted to suit the data.
- Data subjects are afforded opportunities to make decisions about both security and privacy in relation to their data; for example, whether to share a specific bit of information with a service, or whether to use the service at all.
- Subjects of data have little control over what happens with their data once relinquished to a third-party. In fact, a reasonable expectation is that data, once shared, will be re-shared.

Finally, there is an ethical component implied by the synthesis of cultural contexts of security and privacy, with an exploration of how data can be applied to benefit or detriment. This leaves room for discussion about the combination of the five elements, and the ethical components at the center, without adding ethics as a specific sixth element.

Conclusions

As Wolff et al. (2007: 186) write, "*Without [an] ... action plan ... societies are destined to continue to reinforce patterns of entrenched privilege and disadvantage, widening gaps between rich and poor, and the perpetuation of disadvantage*.". This, at a time when analytics predict that 90% of people will have some form of health tracking (not to mention even more powerful smartphones) by 2023 (Miner et al., 2014). The knowledge foundation for how to navigate in this future world must be established through data literacy education, the 21st century equivalent of learning how to balance a checkbook in a modern home economics course. Yet education will only be realized if we begin with the correct basis for a definition for data literacy, which this essay has proposed.

The basis for a definition should be drawn not from domain-specific needs or literature, but from the lived experience of our dataculture; it should draw its inspiration from the data-enabled webs of all that we engage, and all that engages us. This essay has set forth the rationale for determining a definition for data literacy as the knowledge and understanding that forms the basis for how people experience dataculture as stewards and subjects.

Nonetheless, this is the beginning, rather than the end, of exploration. Fundamentally, a thorough literature review of definitions should characterize audience, content, application and applicability. More specifically, research should be conducted to explore whether this particular definition–itself a hypothesis–bears the weight of data. Just which elements of data literacy are members of the public at large—the *populus*—called upon to perform routinely, daily? What does the citizenry understand of these requirements, and how shallow or deep is their understanding? Does it differ by nation? Few studies exist, and those that do report discouraging results (see: Turow et al., 2015).

Studies may also more formally characterize the size, scope and scale of the *populus* in terms of its current breadth and depth. For example, which constituents of a nation comprise those for whom data literacy is important? How do these constituencies vary by more- and less-developed nations? How is data literacy prioritized in the educational context? Finally, to engage educators, a mapping could be made between the knowledge stated herein and the classroom. How do components of data literacy map into the rigid topic-focused curricular regimen of schools? Given current constraints in educational markets, data literacy education may need to take place not in the classroom, but in formal out-of-school or home contexts.

These and many other questions remain to be explored. Nonetheless, the interaction of data and human experience will continue to evolve. These evolutions raise the need for widespread data literacy to the fore for a growing population, the *populo*, and this essay outlines a direction for data literacy to adopt so as to address current and potential future educational needs.

References

Ablon, L., Libicki, M. C., Golay, A. A. (2014). Markets for Cybercrime Tools and Stolen Data: Hackers bazaar. RAND National Security and Research Division. Santa Monica, CA. Retrieved from http://www.rand.org/pubs/research_reports/RR610.html.

Allstate Insurance. (2014, January). What is a Telematics Device? Retrieved from https://www.allstate.com/tools-and-resources/car-insurance/telematics-device.aspx

Boehm, F. (2015). A comparison between US and EU data protection legislation for law enforcement purposes. Retrieved from Directorate General for Internal Policies, Policy Department C: Citizens' Rights and Constitutional Affairs website: http://www.europarl.europa.eu/RegData/etudes/STUD/2015/536459/IPOL_STU(2015)536459_EN.pdf

Boire, R. (2013). Is predictive analytics for marketers really that accurate? *Journal of Marketing Analytics*, 1(2), 118-123.

Busch, L. (2014). Big Data, Big Questions - A Dozen Ways to Get Lost in Translation: Inherent Challenges in Large Scale Data Sets. *International Journal of Communication*, 8, 18.

California (State). Legislation. Assembly. Credit history: public utilities. AB-588. 2007. Reg. Sess. (July 7, 2007). California State Assembly. Retrieved from http://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=200720080AB588 .

Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. *Libraries and the Academy*, 11(2), 629-657.

Chen, Y. (2005). Oligopoly price discrimination by purchase history. In Swedish Competition Authority (Ed.), *The Pros and Cons of Price Discrimination*, 101-130. Sweden: Elanders Gotab AB

Chernova, Y. (2014, May 12). Oh, Baby: Wearables Track Infants' Vital Signs. *The Wall Street Journal*. Retrieved from http://www.wsj.com

Citron, D. K., & Pasquale, F. A. (2014). The scored society: due process for automated predictions. *Washington Law Review*, 89.

Danna, A., & Gandy Jr., O. H. (2002). All that glitters is not gold: Digging beneath the surface of data mining. *Journal of Business Ethics*, 40(4), 373-386.

Deahl, E. S. (2014). Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments. Available at SSRN 2445621.

Decker, C., Kubach, U., & Beigl, M. (2003). Revealing the retail black box by interaction sensing. Proceedings of the ICDCS 2003, Providence, Rhode Island.

D'Ignazio, C., Bhargava, R. (2015). Approaches to Building Big Data Literacy. Proceedings from Bloomberg Data for Good Exchange Conference. New York City, New York.

Duryee, T. (2014, February 24). Lawsuit alleges Amazon Prime third-party prices are inflated to cover shipping. Retrieved from http://www.geekwire.com/2014/lawsuit-alleges-amazon-prime-third-party-prices-inflated-cover-shipping/

Erwin, R. W. (2015). Data Literacy: Real-World Learning Through Problem Solving with Data Sets. *American Secondary Education*. 43:2, Spring.

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays*. New York, NY: Basic Books.

Gunter, G. A. (2007). Building student data literacy: An essential critical-thinking skill for the 21st century. *Multimedia & Internet @ Schools*. 14:3, 24.

Gurstein M (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday* 16:2.

Harris, J. (2012, September 13). Data Is Useless Without the Skills to Analyze It. *Harvard Business Review*. Retrieved from https://hbr.org/2012/09/data-is-useless-without-the-skills

Hartnett Jr, R. J. (1989). Project Management Software: Proper Selection for Use Within Air Force Systems Command (No. AFIT/GSM/LSQ/89S-17). Air Force Institute of Technology Wright - Patterson Air Force Base Ohio School of Systems and Logistics.

Hayashi, A. M. (2013, December 9). Thriving in a Big Data World. *MIT Sloan Management Review*. Retrieved from http://sloanreview.mit.edu/article/thriving-in-a-big-data-world/

Holder, E. M. (2014, August 1). Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference. Retrieved from http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

Identity Theft Resource Center. (2016). 2015 Data Breaches. Retrieved from http://www.idtheftcenter.org/ITRC-Surveys-Studies/2015databreaches.html .

Internal Revenue Service. (2014, February 19). IRS Releases the "Dirty Dozen" Tax Scams for 2014; Identity Theft, Phone Scams Lead List. Retrieved from https://www.irs.gov/uac/Newsroom/IRS-Releases-the-%E2%80%9CDirty-Dozen%E2%80%9D-Tax-Scams-for-2014%3B-Identity-Theft,-Phone-Scams-Lead-List

International Telecommunications Union. (2016). Graph illustration of global ICT developments, 2001-2016. Retrieved from https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

John Hancock Insurance. (2015, August). John Hancock Vitality Program: A Step-by-Step Resource. Retrieved from http://jh1.jhlifeinsurance.com/jhl-ext-templating/filedetail?vgnextoid=471e12d80815f410VgnVCM1000003e86fa0aRCRD&siteName=JHSalesNet

Koltay, T. 2015. Data literacy: in search of a name and identity. *Journal of Documentation*, 71:2, 401-415

Konstan, J. A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2), 101-123.

Krebs, B. (2014, September 18). Medical Records For Sale in Underground Stolen From Texas Life Insurance Firm. Retrieved from http://krebsonsecurity.com/2014/09/medical-records-for-sale-in-underground-stolen-from-texas-life-insurance-firm/

Knutson, J. H. (2002). Credit Scoring in the Insurance Industry: Discrimination or Good Business. Loyola *Consumer Law Review*, 15, 315.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.). Recommender Systems Handbook, 73-105. US: Springer.

Lyon, D. (2014). Surveillance, snowden, and big data: capacities, consequences, critique. *Big Data & Society*, 1(2).

Magalhaes, G., Roseira, C., Strover, S. (2013). Open government data intermediaries: a terminology framework. In Janowski, T., Holm, J., Estevez, E. (Eds.). *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance* (ICEGOV '13), 330-333. New York, NY: ACM.

Mandinach, E. B., & Gummer, E. S. (2012). Navigating the landscape of data literacy: it is complex. Washington, DC and Portland, OR: WestEd and Education Northwest.

Mattioli, D. (2012, August 23). On Orbitz, Mac Users Steered to Pricier Hotels. *The Wall Street Journal*. Retrieved from http://www.wsj.com

McAuley, D., Rahemtulla, H., Goulding, J., & Souch, C. (2012). How Open Data, data literacy and Linked Data will revolutionise higher education. In Coiait, L., & Hill, J. (2012). *Blue Skies: New thinking about the future of higher education in the Asia Pacific region*. (pp. 52-57). Hong Kong: Pearson.

Microsoft. (2014, November 13). Microsoft by the Numbers. Retrieved from https://news.microsoft.com/bythenumbers/ms_numbers.pdf

Miner, L., Bolding, P., Hilbe, J., Goldstein, M., Hill, T., Nisbet, R., & Miner, G. (2014). *Practical Predictive Analytics and Decisioning Systems for Medicine: Informatics Accuracy and Cost-effectiveness for Healthcare Administration and Delivery Including Medical Research*. Academic Press.

National Research Council. (1997). The Future of Spatial Data and Society. *National Research Council*, Washington D.C. DOI: 10.17226/5581

Open Security Foundation. (n.d.). DataLossDB. Retrieved December 16, 2015, from http://datalossdb.org/

Oxford Dictionaries. (n.d.). Algorithm: definition. Retrieved from http://www.oxforddictionaries.com/us/definition/american_english/algorithm

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, Massachusetts: Harvard University Press.

Pellegrini, A.D. & Dressden, J. (1992). Play in school? Yes, we're serious! In V.J. Dimidjian (Ed.). *Play's Place in Public Education for Young Children* (pp. 19-25). Washington, DC : National Education Association of the United States.

Perrotta, C. (2013). Assessment, technology and democratic education in the age of data. *Learning, media and technology*, 38(1), 116-122.

Pomerantz, F. J., & Aisen, A. J. (n.d.). Auto Insurance Telematics Data Privacy And Ownership. *Mealey's Data Privacy Law Report*, 1(1), 1-13.

Rosenblat, A., Wikelius, K., boyd, d., Gangadharan, S. P., & Yu, C. (2014). Data & Civil Rights: Criminal Justice Primer. Retrieved from http://www.datacivilrights.org/pubs/2014-1030/CriminalJustice.pdf

Rouhollahzadeh, B., & Bhatia, R. (2001). U.S. Patent No. 6,208,866. Washington, DC: U.S. Patent and Trademark Office.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Paper presented at preconference to Data and Discrimination: Converting Critical Concerns into Productive Inquiry, 5 May. Seattle, Washington.

Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: options and limitations. *Info*, 17(6), 35-49.

Savitz, E. (2013, January 17). SurveyMonkey To Raise $794M In Recap; Valuation $1.35 Billion (Updated). *Forbes*. Retrieved from http://www.forbes.com

Schield, M. (2004). Information Literacy, Statistical Literacy, Data Literacy. *IASSIST Quarterly*, Summer / Fall 2004, 7-11.

Spencer, S. B. (2015). Privacy and Predictive Analytics in E-Commerce. *New England Law Review*, 49, 629-723.

Twidale, M.B., Blake, C. & Gant, J. (2013). Towards a data literate citizenry. *iConference 2013 Proceedings* (pp. 247-257).

Turow, J., Hennessy, M., & Draper, N. (2015, June).The Tradeoff Fallacy: How Marketers Are Misrepresenting American Consumers and Opening Them Up to Exploitation. Retrieved from Annenberg School for Communication, University of Pennsylvania website: https://www.asc.upenn.edu/sites/default/files/TradeoffFallacy_1.pdf

U. S. Chamber of Commerce Foundation. (2014). The Future of Data-Driven Innovation. Retrieved from https://www.uschamberfoundation.org/sites/default/files/The%20Future%20of%20Data-Driven%20Innovation.pdf

U. S. Senate Committee on Commerce, Science and Transportation. (2013). A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes. Retrieved from Office of Oversight and Investigations website: http://www.commerce.senate.gov/public/index.cfm/reports?ID=57C428EC-8F20-44EE-BFB8-A570E9BE0CCC

Vahey, P., Yarnall, L., Patton, C., Zalles, D., & Swan, K. (2006, April). Mathematizing middle school: Results from a cross-disciplinary study of data literacy. Presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

van Schalkwyk, F., Canares, M., Chattapadhyay, S., Andrason, A. (2015). Open Data Intermediaries in Developing Countries. Retrieved from https://dx.doi.org/10.6084/m9.figshare.1449222.v1

van’t Hooft, M., Vahey, P., Swan, K., Kratcoski, A., Cook, D., Rafanan, K., Stanford, T., Yarnall, L. 2012. A Cross-Curricular Approach to the Development of Data Literacy in the Middle Grades: The thinking with data project. *Middle Grades Research Journal*, 7(3), pp. 19-33.

van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, 6(2), 129-140.

Wessels, B. (2015). Authentication, Status, and Power in a Digitally Organized Society. *International Journal of Communication* (19328036), 9.

Wolff, J., de-Shalit, A. (2007). *Disadvantage*. New York, NY: Oxford University Press.

World Bank. (2016). World Development Report 2016: Digital Dividends. Washington, DC: World Bank. doi:10.1596/978-1-4648-0671-1.