International Journal of Population Data Science

Journal Website: www.ijpds.org

A Position Statement on Population Data Science: The Science of Data about People

Kimberlyn M McGrail[1*], Kerina Jones[2], Ashley Akbari[2], Tellen D Bennett[3], Andy Boyd[4], Fabrizio Carinci[5], Xinjie Cui[6], Spiros Denaxas[7], Nadine Dougall[8], David Ford[2], Russell Kirby[10], Hye-Chung Kum[11], Rachael Moorin[12], Ros Moran[13], Christine M O'Keefe[14], David Preen[15], Hude Quan[9], Claudia Sanmartin[16], Michael Schull[17], Mark Smith[18], Christine Williams[19], Tyler Williamson[9], Grant MA Wyper[20], and Milton Kotelchuck[21]

Submission History

Submitted:	01/09/2017
Accepted:	30/11/2017
Published:	22/02/2018

[1]The University of British Columbia
[2]Swansea University
[3]University of Colorado School of Medicine
[4]University of Bristol
[5]University of Bologna
[6]PolicyWise for Children & Families
[7]University College London
[8]Edinburgh Napier University
[9]University of Calgary
[10] University of South Florida
[11]Texas A&M University
[12]Curtin University
[13]Health Research Board, Ireland
[14] CSIRO, Australia
[15]University of Western Australia
[16]Statistics Canada
[17]Institute for Clinical Evaluative Sciences (ICES)
[18]University of Manitoba, Manitoba Centre for Health Policy
[19]Australian Bureau of Statistics
[20] Public Health and Intelligence, NHS National Services Scotland
[21]Harvard Medical School

Abstract

Information is increasingly digital, creating opportunities to respond to pressing issues about human populations using linked datasets that are large, complex, and diverse. The potential social and individual benefits that can come from data-intensive science are large, but raise challenges of balancing individual privacy and the public good, building appropriate socio-technical systems to support data-intensive science, and determining whether defining a new field of inquiry might help move those collective interests and activities forward. A combination of expert engagement, literature review, and iterative conversations led to our conclusion that defining the field of Population Data Science (challenge 3) will help address the other two challenges as well. We define Population Data Science succinctly as ***the science of data about people*** and note that it is related to but distinct from the fields of data science and informatics. A broader definition names four characteristics of: data use for positive impact on citizens and society; bringing together and analyzing data from multiple sources; finding population-level insights; and developing safe, privacy-sensitive and ethical infrastructure to support research. One implication of these characteristics is that few people possess all of the requisite knowledge and skills of Population Data Science, so this is by nature a multi-disciplinary field. Other implications include the need to advance various aspects of science, such as data linkage technology, various forms of analytics, and methods of public engagement. These implications are the beginnings of a research agenda for Population Data Science, which if approached as a collective field, can catalyze significant advances in our understanding of trends in society, health, and human behavior.

Introduction

Developments in information and communications technologies have altered the research capabilities of almost every academic field. While advances are not new, the pace of change has increased rapidly over the last few decades. The real differences for research come from the exponential increases in computer storage and the digitization of information: <1 % of the world's information was estimated to be in digital form in 1986, compared to 94% in 2007, as shown on Figure 1 (1). Readily available digital information creates new opportunities to answer questions, on an ever-increasing population scale, about human health and well-being, the delivery of public services, and the functioning of societies.

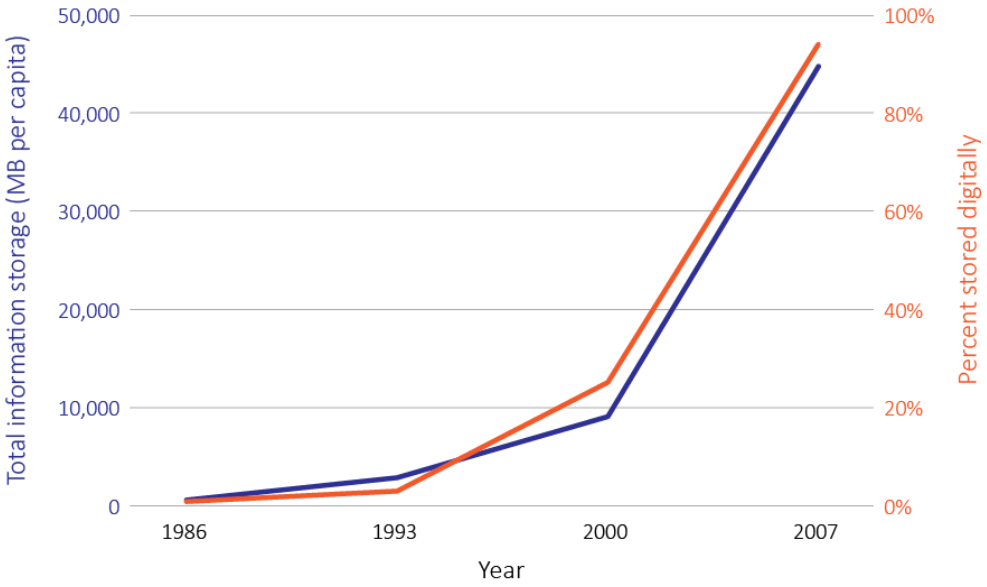
Digital storage and collection technologies also translate to amassing more information, with one repeated assertion being that 90% of the world's information has been collected in the last two years alone (2). Every interaction, service contact, device use, social media post, and clinical encounter is construed as a data resource from which we can extract information or meaning. More traditional administrative data, in a range of sectors, are also becoming increasingly digitized and available, with the capacity to link these data at the unit-record level now commonplace in many countries (3). Linked datasets that are large, complex, diverse, and increasingly available in near real-time open new challenges and new possibilities for the pursuit of scientific knowledge.

Data fuels the knowledge economy, and there is an increasing tendency of both private industry and the public sector to view data as an asset as well as a "frontier for innovation" (4). The size and complexity of datasets, particularly when they derive from multiple sources, makes assembling data for

*Corresponding Author:
Email Address: kim.mcgrail@ubc.ca (K McGrail)

http://dx.doi.org/10.23889/ijpds.v3i1.415

Figure 1: The increase in total information and percent digital, 1986-2007



Source: Hilbert M, López P. The World's Technological Capacity to Store, Communicate, and Compute Information. S [Internet]. 2012;332(60):60-5

individual research projects daunting. Changes in the nature and source of data lead to some suggesting there should be more onus on the uses and users of data (5), but these requirements are likely beyond the ability of individuals researchers to meet. There is thus an emerging role for data *systems* to be developed and embedded as fundamental components of science infrastructure, including cyberinfrastructure (6).

Data-intensive science requires multi-disciplinary collaboration and ongoing research and development both for building science infrastructure and for using that infrastructure to derive new knowledge. There have in fact been significant investments in centres that are charged with holding, linking and/or integrating, and making data available for research. The focus of these investments are data that are mainly individual-focused, and often population-based; and scientific studies that use those resources. Significant recent examples of investments can be found in Australia (the Population Health Research Network) (7,8), New Zealand (Integrated Data Infrastructure) (9), Canada (the Strategy for Patient Oriented Research and the Statistics Canada Social Data Linkage Environment) (10,11), United States (The national patient-centered clinical research network, PCORnet) (12) and the UK (the Farr Institute, the Administrative Data Research Network, and most recently Health Data Research UK) (13,14).

Challenges we face in the brave new world of data

The potential social and individual benefits that can come from data-intensive science are large (15) but realizing them is not straightforward. We identify three broad challenges, of equal importance.

Challenge one is achieving a balance between the need for individual privacy (protecting access to and use of data) and the public good (16-21). This is a consideration for many data-intensive domains of science, and across virtually all countries, but will have a different focus or importance depending on the users of data, the types and source of data, and the research questions being addressed. This challenge is complicated by differences in legal and regulatory structures across jurisdictions, varying levels of societal understanding relating to confidentiality safeguards, and inconsistent views as to when or how those safeguards are put into place (22).

Challenge two is nurturing appropriate socio-technical systems to support data-intensive science. The use of complex, linked data for research is a comprehensive technical enterprise, depending on powerful computational systems and advanced statistical methods. The first challenge of balancing privacy and the public good is just one example of why we need to think beyond these technical issues to the ethical and procedural challenges as well. Data-intensive research is one way to advance our understanding of individuals and societies, and given the leaps in data and analytics over the past decade it is difficult to imagine a future where data disappears. If anything, we are likely to become more reliant on data and analytics - both in general and in academic pursuits. Consequently, the broad systems in place that support data-intensive science must increasingly reflect attention to legal, ethical, financial, data quality, data curation, provenance, and other organizational issues and processes (23).

Challenge three is a conceptual one, encompassing questions concerning the evolution of scientific inquiry, and whether collective interests in population-focused data-intensive science are fully captured by, or contained within, any single existing field or discipline. The specific challenge is to determine whether defining a new field of inquiry might help

move those collective interests and activities forward. This challenge is meaningful in that it can help identify the resources needed to develop and nurture scientific ambitions, accommodating both current needs and future priorities.

In working through these challenges, we conclude that the increasingly popularized term "data science" is both too vague in describing our collective interests and activities, and too encompassing of broader and more general interests than are inherently relevant to secondary uses of multi-source, linked and person-focused data. We propose "Population Data Science" as a more meaningful term for this emerging field. This paper is the consequence of working through each of these challenges, and is focused on four specific aims:

1. To create a concise definition of what we see as an evolving field of Population Data Science.
2. To highlight the specific characteristics and challenges of Population Data Science.
3. To differentiate Population Data Science from existing fields of data science and of informatics in its various forms.
4. To discuss the implications and future opportunities for Population Data Science to address the three challenges outlined above.

Approach

We used a combination of expert engagement, literature review, and iterative conversations to work towards the information presented in this paper. These efforts and activities were organized through the International Population Data Linkage Network (see http://www.ipdln.org/) and so focus on its members, but draw on conversations and experience outside the Network as well.

More specifically, Aims 1 and 2 were derived largely through member engagement that was informed by both literature and experience, Aim 3 largely depended on a literature review, and Aim 4 was achieved through iterative responses to the evolving draft of this manuscript. The following provides a brief summary of information and interactions that result in this paper.

An initial workshop with 35 participants was held in August 2015 at the Farr Institute international conference in St. Andrews, Scotland (24). The discussion focused on identifying commonality of interests, and whether these constituted being a distinct field. The concept of "Population Data Science" was proposed as a possible 'umbrella' identity, with commitment to further exploration of the benefits and drawbacks of proposing a distinct field.

Following this initial meeting, one author (KM) conducted a rapid literature review to identify definitions of existing fields and to ascertain whether they sufficiently encompassed our self-described work area. This was a structured scoping review around the two disciplines that had the most likelihood of encompassing our work: informatics and data science. A broad search of academic and grey literature was conducted using a combination of "informatics" or "data science" with any or all of the following: "population", "health", "medical", "privacy" and "personal".

This review was summarized in a working paper outlining existing terminology and the fit of definitions with views expressed at the first workshop. At this point we established a core team (KHJ and KM) to lead the further development of the work.

A second workshop was held in February 2016 in the Data Science Centre at Swansea University in Wales, with staff from across several data infrastructure initiatives, including the SAIL databank [https://saildatabank.com/], ADRC-Wales [https://adrn.ac.uk/about/network/wales/] and Farr@CIPHER [http://www.farrinstitute.org/]. Participants were asked to prepare a concise definition to encapsulate the work area in which we are engaged. This was followed by a second international workshop of 40+ individuals during the third International Population Data Linkage Network conference in Swansea, Wales, in 2016 (25). There was iterative engagement with participants over a period of several months leading up to the workshop, including review and comment on a draft description of 'Population Data Science' and a request for concise descriptions to feed into the evolving definition. Summary findings of the workshop and a consensus definition of Population Data Science were circulated to the full IPDLN membership (n>600) with an invitation to contribute to the development of this position paper.

Results

This section summarizes collective thinking around Aims 1, 2 and 3, while Aim 4 is addressed in the Discussion section, drawing all the pieces together.

Aim 1: A concise definition

The definition we agreed upon is:
Population Data Science is a multi-disciplinary field aimed at obtaining population-level insights with public value by organizing, linking or otherwise integrating and analyzing data that pertain to individuals and their social, economic, biological and environmental characteristics and contexts.

It can be concisely described as ***the science of data about people***.

Defining the field in this way implies additional characteristics. Population Data Science is oriented to data use that will have a positive impact on citizens and society, and which places a high value on the inclusion of the public voice in the governance of data access and use. Given the focus on people and public value, Population Data Science also includes a focus on the development of safe, privacy-sensitive and ethical infrastructure to support science.

We do not envision that individual Population Data Scientists have expertise in every aspect of this definition. Some Population Data Scientists develop and implement infrastruc-

[1]Linking refers to the technical process of identifying the same individual in multiple data sets. Integrating, in contrast, refers to bringing together data sets and using them without necessarily linking at the individual level. We might, for example, link individual experience across health and education and then integrate those data with ecological information about neighbourhoods. Both terms are important to Population Data Science, but have distinct meanings which we adhere to throughout the paper.

ture and analytical tools to manage, curate, link or integrate[1] and provide secure access to data in accordance with legal and ethical obligations, and consistent with expectations of the public those data represent (26). Others use that infrastructure and the data available through them to develop insights into the development, maintenance and improvement of human and societal well-being (27).

Aim 2: Specific characteristics and ambitions of Population Data Science

The ultimate aim of Population Data Science is to have a positive impact on citizens and society through population-level insights obtained from the linkage and analysis of varied and often complex data sets pertaining to the social and biological circumstances and natural environment of individuals. Population Data Science utilizes data across different scales: from whole population datasets (e.g. the SAIL databank of Welsh citizen records (28)); through cohort studies (e.g, the Avon Longitudinal Study of Parents and Children (29)); targeted sub-populations that may be drawn from either of the above (e.g. clinical trial participants); and, qualitative in-depth interviews on purposively sampled populations (e.g. Wellcome Trust 'The One-Way Mirror' report on public attitudes to commercial access to health records (30)). Frequently these differing scales can interact, such as in-depth interviews or cohort studies being nested in wider samples. These characteristics introduce their own scientific challenges without which, of course, there would be no need for collective action.

The four characteristics of Population Data Science flow directly from these observations: 1) a focus on people and population-level insights; 2) linking multiple datasets and types, and building and using analytical tools to use these data; 3) the use of data for positive impact on citizens and society; and 4) development of technical and policy infrastructure built around legal, ethical and privacy norms as well as public expectations.

People and population-level insights

In this paper we use the term "population" broadly, recognizing that populations can be defined in various ways and through many different scales. What is significant here is that Population Data Science has a focus on collections of individuals, and the biological, economic, social, and environmental experiences that shape their health and well-being. This is in contrast, for example, to data scientists who might be interested in astrophysics, or informaticians who might focus on users and their adoption of technology. Insights are similarly broad, as they might come from a range of approaches, such as epidemiology, social science, predictive analytics, or machine learning, to name just a few. Population Data Scientists are also involved with developing novel methods for generating insights, for example distributed analytics, visual analytics, and the use of virtual reality, which are areas with connections to both data science and informatics.

The implication is that few people possess all of the requisite knowledge and skills needed to conduct research that will move our understanding of the human condition forward. Population Data Science is by its nature a 'team sport', and will only be possible with multi-disciplinary, multi-site and multi-jurisdictional research collaborations. Our field includes professionals from varied disciplines including, but not limited to: computer and information sciences, statistics, epidemiology, public health, clinical medicine, pharmacy, social welfare, ethics, law, information governance, geography, economics, social science, and knowledge translation. As noted, the field is also oriented towards inclusion of public and stakeholder voices through various means such as patient or consumer groups, public advisory panels and in-depth qualitative surveys.

Linking multiple data sources and types

The objectives of Population Data Science are frequently focused around the secondary use of data. The traditional backbone of many long-standing linked data systems is administrative data, and more specifically health care data, that reflect the use of services or payments for those services. The ambitions of Population Data Science, however, stretch well beyond health care or even health, and in terms of data, beyond administrative data to include data that are individually reported or objectively observed; that derive from traditional data capture mechanisms such as questionnaires, interviews, imaging, biosamples; and/or that derive from emerging data gathering technologies such as digital imaging, natural language processing, geospatial capturing, 'omics, sensors, wearable devices, consumer records and social media. In other words, the data of the future can be expected to be more varied in origin, structure, content and size. Population Data Science might include "big data", as defined by size, but more often is focused on the integration and interpretation of a number of different data sources and types to enable new and innovative research and solutions. As such, we consider the field to be defined as utilizing "complex" and "varied" rather than necessarily "big" data.

The implication is a need for both linking and/or integrating data and for the analytic approaches that complex data will require. Addressing the linking challenges will come from infrastructure development (see below) but also by advancing the science of linking technology itself, for example through tools such as privacy-preserving record linkage, which is a way to mask identities even while data are being integrated (31,32). On the analytics side, Population Data Scientists will develop new techniques for using and analyzing complex data, including methods that will enable analysis without moving data, and in some cases without seeing them, and will help train others in these approaches. This will include ensuring the ability to take advantage of the powerful longitudinal nature of the data, which means adding the element of time to the already complex data sources and types.

The use of data for positive impact on citizens and society

Our definition of Population Data Science emphasizes the importance of research that has public value, with the specific

intent of developing insights into the health and well-being of individuals and communities. We view the public value aspect of Population Data Science as fundamental because the data in question provide often quite detailed information about people and their surroundings. The ability to gather, link and use these data must, we assert, be built on a relationship of trust with the public, and the emerging scientific literature clearly indicates the public is far more willing to trust the conduct of data-intensive science that has public value (16,30). This does not mean that uses must have *exclusively* public value, as there is acceptance that the same piece of research may provide both private and public benefit, but the public benefit is paramount.

The implication is that the Population Data Science research agenda will include advancing the science of public and patient engagement, continuing to develop robust methods to ascertain public values and expectations, building those into the way we organize and operate data systems, and ensuring that we change as capabilities and expectations evolve. Population Data Scientists will do this through a commitment to advancing the science related to all of the above, and to involving the public and other key stakeholders in all aspects of our work, from informing infrastructure developments to advising the research which uses that infrastructure (22,23,28,33-38).

Technical and policy infrastructure built around legal, ethical and privacy norms

The preceding characteristics of Population Data Science all underscore the importance of infrastructure for amassing, storing, linking, providing and using data. In this sense "infrastructure" relates to far more than the bricks and mortar or technical computational systems involved: the data must be stored on secure systems (35); those systems need robust, proportionate governance plans and authority; there is a need for data access decision-making frameworks (39) and models with processes to implement those (40,41); people using data must be provided with information on data origins and quality and other metadata (42); analysts must have appropriate skills; and the public must be engaged to ensure these practices align with their values and expectations.

The implication here is again that the required knowledge and skills are not likely to exist in a single individual. This underlines that Population Data Science is a multi-disciplinary endeavour, which means nurturing collaborative teams but also placing emphasis on cross-disciplinary training and communication. In addition, there is a need to work closely with policy-makers and others who are responsible for the original sources of data. As data sources become more varied and complex, the number of data providers and interested policy-makers will also grow, implying that stakeholder engagement will continue to be a necessary and important feature of Population Data Science. Finally, there is a need for Population Data Scientists both to recognize and respect the sensitive nature of the data they use, and to have the skills and analytic tools available that make the best use of those data.

Aim 3: Differentiating Population Data Science

Asserting that we are embarking on the development of a new field of inquiry requires a review of existing fields and specifying what is missing or imperfectly emphasized in comparison to aspirations. We start with definitions of existing fields and then compare and contrast these with Population Data Science. This is not to suggest an absence of overlap among different fields, particularly at the margins, but enough distinctiveness to warrant a separate label and definition.

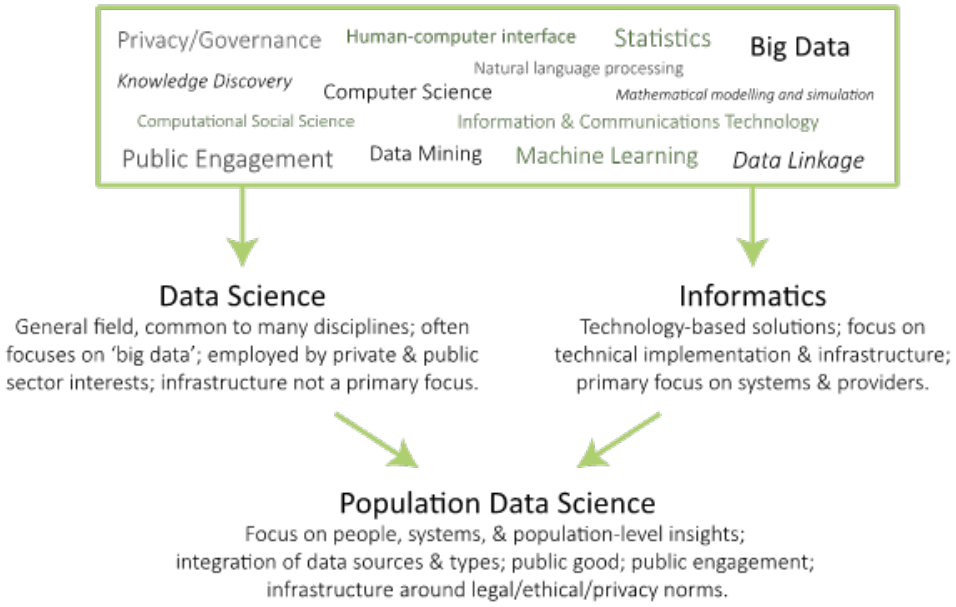
Data science. Data science is described as a combination of many different disciplines, with greatest emphasis on computer science and statistics (43). Another way of saying this is that data science can be viewed as an approach that combines the skills and knowledge of statistics, computer science, and some kind of domain expertise such as health or education or sociology. Others identify data scientists as people with computer science and statistical / mathematical skills who also have the ability to identify problems or questions of interest and to communicate their findings to relevant audiences (44).

Data science is the process of timely extraction of knowledge from data, and has emerged because of the advent of computing power sufficient to exploit these data; the data themselves may have existed for decades, but the difference now is the technical ability to link and analyze them. The term science underscores that, in attempts to extract meaning from data, robust methodological approaches and analytical designs are required (i.e. problems and questions are approached systematically). In addition, there must be caution against over-fitting or over-generalizing from data, as well as against uncritically attributing causation to correlation (44).

Informatics: The broad discipline of informatics has been defined as "...the study of the structure, the behaviour, and the interactions of natural and engineered computational systems" (45). It draws from disciplines such as cognitive science, computer science, and artificial intelligence. Similar to data science, informatics is often a combined field, with the connecting feature being the use of information and communication technology to support a discipline, such as development informatics (46), construction informatics (47), and perhaps most prominently, health informatics. The International Medical Informatics Association defines health informatics as "...the discipline that deals with health related data...and with how computers, software and telecommunication technologies are used to support the delivery of health care services" (48). This field has a much longer history than data science, and continues to grow, fuelled by digitization in the health care sector.

Kum et al. (49) have defined the terms population informatics and social genome as being counterparts to bioinformatics and human genome. They define Population Informatics as "the burgeoning field at the intersection of social sciences, health sciences, computer science, and statistics that applies quantitative methods and computational tools to answer questions about human populations." and social genome as "the digital footprints of our society". They are clearly engaged in areas of interest and relevance to the Population Data Science community, such as data integration (including

Figure 2: Relationship of Population Data Science to Data Science and Informatics



privacy preserving interactive data linkage), privacy and confidentiality more generally, data analysis, and orientation to the idea of using linked person-level data for population research.

In summary, there are many terms that refer to overlapping concepts all of which have some relevance to using primary or secondary data for population insights. Some of the distinctions, particularly between certain areas of data science and informatics, do not appear to be obvious even to people who align with those groups. The interests of Population Data Science fit at least partially into all of these areas, but always with a special focus on the problem domain of *populations* and use of *technical* and policy infrastructure that supports analysis of *multi-source, linked and person-oriented data* for research that has *public value*. Significantly, our literature review revealed no documented usage of the term Population Data Science, further supporting the use of this term to distinguish our field since it is not currently being used in any contradictory way.

Figure 2 shows the derivation of focus and themes that inform Population Data Science and their relationship to data science and informatics. All of these fields have some overlaps and interrelationships as well as clear distinctions. The name Population Data Science implies perhaps some greater alignment with the broader area of data science, but with some additional aims and focus that draw from informatics and elsewhere.

The four key characteristics of Population Data Science help identify our field, both in offering some focus and in differentiating our specific interests from the broader fields of data science and informatics. These similarities and differences are summarized in Table 1. These boundaries are of course porous and there are overlapping interests. At the same time, the ambitions of Population Data Science are distinct, and defining them will help our collective activities in identifying and advancing the field.

Discussion

Our world is increasingly digital, with more and more information about every aspect of our lives being captured and stored. This creates enormous opportunity for data-intensive science and great responsibility for that science to respect the people and communities those data represent. We identify three challenges for data-intensive science about people and populations: balancing individual privacy and public good; nurturing socio-technical systems that can support data-intensive science; and determining whether the scope and nature of our interests implies a need to define a new field of science.

Responding to the last challenge first, we propose that there is a need to define common interests as a way to promote collective action that can move science forward in the many areas that relate to data-intensive research using complex, multi-source, and person-focused linked data. The differentiation of Population Data Science from the related fields of data science and informatics is intended to help define and shape our evolving capacities to improve the human condition. Defining Population Data Science concisely as the "science of data about people" is part of our commitment to ensure the collection, storage, and use of data is rigorous and scientific, including in its reflection of public input and expectations.

More fully, we define Population Data Science as a multidisciplinary field aimed at obtaining population-level insights by organizing, linking and analyzing data that pertain to the lives of individuals and their social, economic, biological and environmental characteristics and contexts. Embedded in this definition are the four characteristics of Population Data Science of a focus on people and population-level insights, linking and analyzing multiple datasets and types, the use data for positive impact on citizens and society, and development of technical and policy infrastructure to support science. These characteristics clearly respond to the other two challenges of

Table 1: Existing fields of inquiry assessed against characteristics of Population Data Science

	Is the focus on people?	**Do data come from multiple sources?**	**What is the primary aim of research?**	**Is technical and policy infrastructure a focus?**
Population Data Science	Focus on people, systems and population-level insights	Yes, a primary objective is linking and/or integrating data from multiple data sources and data of different types	Research must be seen to have public value, with potential for positive impact on citizens and society	Infrastructure is a key focus, in particular with respect to legal, ethical, and privacy norms and public expectations. This covers all aspects of data from collection to storage and use.
Data science	The focus is on the data themselves. The field is general to all disciplines, though often focuses on "big data"	Linking can be, but is not necessarily a focus. Data from a single source (e.g. a private company) are often the focus.	Focus is use of data for actionable information. Data science techniques are often (though not exclusively) used by private, proprietary interests.	Not generally a focus outside of legal commitments to protect privacy
Informatics	General (though not exclusive) focus on providers or systems as much as people represented in the data.	Sometimes there is linking, though often from an operational perspective	Public good is present but often as a secondary objective behind (for example) implementation of technology-based solutions to improve health care delivery	Infrastructure focus is on database / technical development and implementation

Table 2: Population Data Science characteristics and implications for collective action

Focus:	Implication(s)	Response to implication(s)
People and population-level insights	Required knowledge and skills are not likely to exist in a single discipline	Population Data Science is a multi-disciplinary field that encourages collective action, including the public's voice.
Linking and interpreting multiple data sources / types	Technical approach to linkage - bringing together disparate data without common identifiers, preserving privacy	Population Data Science will include advancing the science of linkage technology.
	Analysis of complex data	Population Data Science will develop new tools for data analysis and will promote the training of practitioners.
		Population data science will develop methods for data analysis that do not require movement or (in some cases) direct viewing of sensitive data
	Interpreting data in a secondary context	Population Data Science will develop the assessment and reporting frameworks needed to document data with sufficient detail to inform accurate assessments
The use of data for public good, for positive impact on citizens and society	Understanding the values and expectations of the public and other key stakeholders, and then building systems to meet those.	Population Data Science is committed to public and stakeholder involvement and engagement in its many forms.
		Population Data Scientists will advance the science of public engagement to promote public understanding of data usage.
Technical and policy infrastructure built around legal, ethical, and privacy norms	Required knowledge and skills are not likely to exist in a single discipline	Population Data Science is a multi-disciplinary field that has a strong focus on stakeholder engagement and commitment to capacity building.
	Data are complex and sensitive.	Population Data Scientists commit to ethical and rigorous science within the legal framework of their jurisdictions.

data-intensive science. Balancing privacy and the public good is a theme that cuts across all characteristics, and as such is an intrinsic part of our identity and commitment to fellow citizens. Nurturing socio-technical systems that can support data-intensive science is a central aim of Population Data Science, really defining *how* the aim of science for public good can be achieved.

With that definition in mind, Table 2 summarizes the implications of the four characteristics and some initial thoughts on the ways that Population Data Science can and should respond. This can be interpreted as the beginnings of a research agenda, and a charge to Population Data Scientists on areas that will benefit from our collective action.

Providing definition to an emerging field is inherently complicated. It is for this reason that we engaged in an iterative process to debate this concept that included consultations, a literature review of definitions and boundaries of existing fields of inquiry, and an iterative approach to the development of this paper. Each of these steps had significant influence on how we thought about and ultimately presented the material here, and resulted in a strong endorsement of the content of this paper. While this work was done mainly within the membership of the International Population Data Linkage Network (IPDLN), we do not envision that all Population Data Scientists will become part of the Network (though we would welcome them) or that IPDLN "owns" Population Data Science. The IPDLN is, however, poised to take concrete steps to use the definition of Population Data Science as a way to create a framework for this discipline, help identify and increase its membership, and nurture our growing international community and its capacity to contribute to scientific, economic and social progress.

Our goal is for Population Data Science insights to contribute to the greater understanding of the root causes of social and public health problems, help predict the downstream effects of different policy options, identify upstream opportunities for interventions, and assist in allocating our collective resources for the greatest impact to benefit our global society. The emergence of new forms of data, new technical capabilities, and a group of Population Data Scientists committed to develop this field opens many new possibilities. Just as bioinformatics revolutionized biological research, Population Data Science can catalyze significant advances in our understanding of trends in society, health, and human behavior.

References

1. Hilbert M, López P. The World's Technological Capacity to Store, Communicate, and Compute Information. S [Internet]. 2012;332(60):60-5. Available from: http://www.martinhilbert.net/LopezHilbertSupportAppendix2012.pdf

2. SINTEF. Big Data, for better or worse: 90% of world's data generated over last two years – ScienceDaily [Internet]. Science Daily. 2013 [cited 2017 Aug 8]. Available from: https://www.sciencedaily.com/releases/2013/05/130522085217.htm

3. IPDLN. Data linkage centres | www.ipdln.org [Internet]. [cited 2017 Aug 8]. Available from: http://www.ipdln.org/data-linkage-centres

4. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: The next frontier for innovation, competition, and productivity. 2011;

5. Mayer-Schönberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.

6. Hey T, Trefethen AE. Cyberinfrastructure for e-Science. Science (80-). 2005;308(May):817-22.

7. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB, et al. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res [Internet]. 2012 Dec 29 [cited 2016 Oct 3];12(1):480. Available from: http://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-12-480

8. Population Health Research Network. Our Funders. 2011.

9. Stats NZ. Integrated Data Infrastructure [Internet]. 2017 [cited 2017 Aug 23]. Available from: http://m.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx

10. Canadian Institutes of Health Research. Canada's Strategy for Patient-Oriented Research: Improving health outcomes through evidence-informed care. 2011;(August):40. Available from: http://www.cihr-irsc.gc.ca/e/documents/P-O_Research_Strategy-eng.pdf

11. Statistics Canada. Social Data Linkage Environment (SDLE) [Internet]. 2017 [cited 2017 Aug 23]. Available from: http://www.statcan.gc.ca/eng/sdle/index

12. Daugherty SE, Wahba S, Fleurence R, Avillach P, Buelow J, Colletti R, et al. Patient-powered research networks: building capacity for conducting patient-centered clinical outcomes research. J Am Med Informatics Assoc [Internet]. 2014 Jul 1 [cited 2017 Aug 23];21(4):583-6. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2014-002758

13. Economic and Social Research Council. Big Data Investment: Capital funding. 2013.

14. Medical Research Council. £20 million for new health informatics research institute. 2013.

15. Stanley F. Data for health. In: Sykes H, editor. A Love of Ideas. Sydney (NSW): Future Leaders; 2014. p. 13-25.

16. Aitken M, Jorre JDS, Pagliari C, Jepson R, Cunningham-burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics [Internet]. 2016; Available from: http://dx.doi.org/10.1186/s12910-016-0153-x

17. Health Data Governance [Internet]. OECD Publishing; 2015 [cited 2017 Aug 24]. (OECD Health Policy Studies). Available from: http://www.oecd-ilibrary.org/social-issues-migration-health/health-data-governance_9789264244566-en

18. Nuffield Council on Bioethics. The collection, linking and use of data in biomedical research and health care: ethical issues. London, United Kingdom; 2015.

19. O'Keefe CM, Rubin DB. Individual privacy versus public good: Protecting confidentiality in health research. Stat Med. 2015;34(23):3081-103.

20. Oderkirk J, Ronchi E, Klazinga N. International comparisons of health system performance among OECD countries: Opportunities and data privacy protection challenges. Health Policy (New York) [Internet]. 2013 Sep [cited 2017 Aug 24];112(1-2):9-18. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168851013001681

21. OECD. Recommendation of the OECD Council on Health Data Governance [Internet]. 2017 [cited 2017 Aug 24]. Available from: http://www.oecd.org/health/health-systems/Recommendation-of-OECD-Council-on-Health-Data-Governance-Booklet.pdf

22. Audrey S, Brown L, Campbell R, Boyd A, Macleod J, Uk S. Young people's views about the purpose and composition of research ethics committees: findings from the PEARL qualitative study. BMC Med Ethics [Internet]. 2016 [cited 2017 Aug 23];17. Available from: https://bmcmedethics.biomedcentral.com/track/pdf/10.1186/s12910-016-0133-1?site=bmcmedethics.biomedcentral.com

23. Burton PR, Murtagh MJ, Boyd A, Williams JB, Dove ES, Wallace SE, et al. Data Safe Havens in health research and healthcare. Bioinformatics [Internet]. 2015 Oct 15 [cited 2017 Aug 23];31(20):3241-8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv279

24. The Farr Institute. Farr Institute | The Farr Institute International Conference 2015 [Internet]. [cited 2017 Aug 8]. Available from: http://www.farrinstitute.org/events-courses/event/the-farr-institute-international-conference-2015

25. IPDLN. Internationl Population Data Linkage Network Conference, 2016. 2016.

26. Smith M, Roos LL, Burchill C, Turner K, Towns DG, Hong SP, et al. Health services data: Managing the data warehouse: 25 years of experience at the Manitoba Center for Health Policy. In: Sobolev B, Levy A, Goring S, editors. Data and Measures in Health Services Research. SpringerLi. 2016. p. 1-26.

27. Roos LL, Jarmasz JS, Martens PJ, Katz A, Fransoo R, Soodeen R-A, et al. Health services information: From data to policy impact (25 years of health services and population health research at the Manitoba Center for Health Policy). In: Sobolev B, Levy A, Goring S, editors. Data and Measures in Health Services Research. SpringerLink; 2016. p. 1-20.

28. Ford D V, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC Health Serv Res. 2009 Jan;9(1):157.

29. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: The "Children of the 90s"-the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol [Internet]. 2013 Feb 1 [cited 2017 Aug 23];42(1):111-27. Available from: https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dys064

30. The Wellcome Trust. The One-Way Mirror: Public attitudes to commercial access to health data. 2016.

31. Randall SM, Brown AP, Ferrante AM, Boyd JH, Semmens JB. Privacy preserving record linkage using homomorphic encryption. In: First International Workshop on Population Informatics for Big Data (PopInfo'15). Sydney; 2015.

32. Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In: Handbook of Big Data Technologies [Internet]. Cham: Springer International Publishing; 2017 [cited 2017 Aug 9]. p. 851-95. Available from: http://link.springer.com/10.1007/978-3-319-49340-4_25

33. Jones KH, Ford D V., Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform [Internet]. 2014 Aug [cited 2017 Aug 24];50:196-204. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1532046414000045

34. Jones KH, McNerney CL, Ford D V. Involving consumers in the work of a data linkage research unit. Int J Consum Stud [Internet]. 2014 Jan 1 [cited 2017 Aug 24];38(1):45-51. Available from: http://doi.wiley.com/10.1111/ijcs.12062

35. Kum H-C, Ahalt S. Privacy-by-Design: Understanding Data Access Models for Secondary Data. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci [Internet]. 2013 [cited 2017 Aug 10];2013:126-30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24303251

36. Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future. Health Serv Res [Internet]. 2010 Oct [cited 2015 Mar 26];45(5 Pt 2):1456-67.

Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2965886&tool=pmcentrez&rendertype=abstract

37. Pencarrick Hertzman C, Meagher N, McGrail KM. Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. J Am Med Inform Assoc [Internet]. 2013 Jan 1 [cited 2016 Oct 3];20(1):25-8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22935136

38. Wellcome Trust Ltd. Understanding Patient Data [Internet]. 2017 [cited 2017 Aug 23]. Available from: https://understandingpatientdata.org.uk/

39. Elliot M, Mackey E, O'Hara K, Tudor C. The Anonymisation Decision-Making Framework [Internet]. Manchester; 2015 [cited 2017 Aug 23]. Available from: http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

40. McGrail KM, Gutteridge K, Meagher NL. Building on Principles: The Case for Comprehensive, Proportionate Governance of Data Access. In: A. G-D, G L, editors. Medical Data Prviacy Handbook. Springer; 2015.

41. Sethi N, Laurie GT. Delivering proportionate governance in the era of eHealth. Med Law Int [Internet]. 2013 Jun 27 [cited 2017 Aug 23];13(2-3):168-204. Available from: http://journals.sagepub.com/doi/10.1177/0968533213508974

42. Smith M, Lix LM, Azimaee M, E Enns J, Orr J, Hong S, et al. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. J Am Med Informatics Assoc. 2017;In press.

43. NIST Big Data Public Working Group. NIST Special Publication 1500-1 - NIST Big Data Interoperability Framework: Volume 1, Definitions. NIST Spec Publ [Internet]. 2015;1:32. Available from: http://dx.doi.org/10.6028/NIST.SP.1500-1

44. Provost F, Fawcett T. Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big Data [Internet]. 2013;1(1):51-9. Available from: http://online.liebertpub.com/doi/abs/10.1089/big.2013.1508

45. The University of Edinburgh. What is Informatics? | The University of Edinburgh [Internet]. [cited 2017 Aug 8]. Available from: http://www.ed.ac.uk/informatics/about/what-is-informatics

46. Walsham G. Development Informatics in a Changing World: Reflections from ICTD 2010 / 2012. Inf Technol Int Dev. 2013;9(1):49-54.

47. Turk Ž. Construction informatics: Definition and ontology. Adv Eng Informatics [Internet]. 2006 Apr [cited 2017 Aug 8];20(2):187-99. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1474034605000911

48. Kluge E-H. The IMIA Code of Ethics for Health Information Professionals. IMIA Code Ethics [Internet]. 2016; Available from: http://imia-medinfo.org/wp/wp-content/uploads/2015/07/IMIA-Code-of-Ethics-2016.pdf

49. Kum HC, Krishnamurthy A, Machanavajjhala A, Ahalt SC. Social genome: Putting big data to work for population informatics. Computer (Long Beach Calif). 2014;47(1):56-63.