Accepted Manuscript

Big data in forensic science and medicine

Thomas Lefèvre

PII: S1752-928X(17)30115-4

DOI: 10.1016/j.jflm.2017.08.001

Reference: YJFLM 1536

To appear in: *Journal of Forensic and Legal Medicine*

Received Date: 1752-928X 1752-928X

Revised Date: 1752-928X 1752-928X

Accepted Date: 1752-928X 1752-928X

Please cite this article as: Lefèvre T, Big data in forensic science and medicine, *Journal of Forensic and Legal Medicine* (2017), doi: 10.1016/j.jflm.2017.08.001.

Big data in forensic science and medicine

Author: Thomas Lefèvre[1,2]

[1]Hôpital Jean-Verdier (AP-HP), Department of Forensic Science and Medicine, F-93140 Bondy, France

[2]IRIS - Institut de recherches interdisciplinaires sur les enjeux sociaux (UMR 8156-723), Bobigny, France

Email addresses: thomas.lefevre@univ-paris13.fr

Corresponding author: Thomas Lefèvre

Phone number: 33148026325

Fax number: 33148026557

Word count: 3792

Abstract

In less than a decade, big data in medicine has become quite a phenomenon and many biomedical disciplines got their own tribune on the topic. Perspectives and debates are flourishing while there is a lack for a consensual definition for big data. The 3Vs paradigm is frequently evoked to define the big data principles and stands for Volume, Variety and Velocity. Even according to this paradigm, genuine big data studies are still scarce in medicine and may not meet all expectations. On one hand, techniques usually presented as specific to the big data such as machine learning techniques are supposed to support the ambition of personalized, predictive and preventive medicines. These techniques are mostly far from been new and are more than 50 years old for the most ancient. On the other hand, several issues closely related to the properties of big data and inherited from other scientific fields such as artificial intelligence are often underestimated if not ignored. Besides, a few papers temper the almost unanimous big data enthusiasm and are worth attention since they delineate what is at stakes. In this context, forensic science is still awaiting for its position papers as well as for a comprehensive outline of what kind of contribution big data could bring to the field. The present situation calls for definitions and actions to rationally guide research and practice in big data. It is an opportunity for grounding a true interdisciplinary approach in forensic science and medicine that is mainly based on evidence.

Key words: forensic science; big data; personalized medicine; predictive medicine; machine learning; dimensionality

Big data in forensic science and medicine

Author: Thomas Lefèvre[1,2]

[1]Hôpital Jean-Verdier (AP-HP), Department of Forensic Science and Medicine, F-93140 Bondy, France

[2]IRIS - Institut de recherches interdisciplinaires sur les enjeux sociaux (UMR 8156-723), Bobigny, France

Email addresses: thomas.lefevre@univ-paris13.fr

Corresponding author: Thomas Lefèvre

Phone number: 33148026325

Fax number: 33148026557

Word count: 3792

Abstract

In less than a decade, big data in medicine has become quite a phenomenon and many biomedical disciplines got their own tribune on the topic. Perspectives and debates are flourishing while there is a lack for a consensual definition for big data. The 3Vs paradigm is frequently evoked to define the big data principles and stands for Volume, Variety and Velocity. Even according to this paradigm, genuine big data studies are still scarce in medicine and may not meet all expectations. On one hand, techniques usually presented as specific to the big data such as machine learning techniques are supposed to support the ambition of personalized, predictive and preventive medicines. These techniques are mostly far from been new and are more than 50 years old for the most ancient. On the other hand, several issues closely related to the properties of big data and inherited from other scientific fields such as artificial intelligence are often underestimated if not ignored. Besides, a few papers temper the almost unanimous big data enthusiasm and are worth attention since they delineate what is at stakes. In this context, forensic science is still awaiting for its position papers as well as for a comprehensive outline of what kind of contribution big data could bring to the field. The present situation calls for definitions and actions to rationally guide research and practice in big data. It is an opportunity for grounding a true interdisciplinary approach in forensic science and medicine that is mainly based on evidence.

Key words: forensic science; big data; personalized medicine; predictive medicine; machine learning; dimensionality

Introduction

In less than a decade, big data in medicine has become quite a phenomenon and many biomedical disciplines got their own tribune on the topic. Perspectives and debates are flourishing while there is a lack for a consensual definition for big data. Big data presents all the attributes of a buzz word, but it should be thoroughly investigated so that it can be decided whether big data is a mere trend or the premises of a true revolution for research and routine practice. Origins of the term “big data” are unclear and we propose here to go many years back in time to search for them. Techniques usually presented as specific to big data such as machine learning techniques are supposed to support the ambition of personalized, predictive and preventive medicines. These techniques are mostly far from being new and are more than 50 years old for the most ancient. On the other hand, several issues closely related to the properties of big data and inherited from other scientific fields such as artificial intelligence are often underestimated if not ignored. In this paper, we expose the most important of these issues. Finally, in a context of general enthusiasm about the big data phenomenon, forensic science is still awaiting for their position papers as well as for a comprehensive outline of what kind of contribution big data could bring to the field. The present situation calls for definitions and actions to rationally guide research and practice in big data. Here, we briefly present what is at stakes for forensic science if it wants to embrace the philosophy of the big data era while avoiding its main pitfalls.

1 – Is big data more than a buzz word and where does it come from?

No one can determine where or when the use of the term “big data” originated (1). It exponentially spread and contaminated all scientific and non-scientific fields within the past decade.

1.1 Origins of big data: definitions and practices of big data in the past decade.

Big data is a vague and generic term that can encompass several distinct and non-exclusive properties. The origin of the big data concept is often attributed to a short technical report from the META group which is an American consulting firm, since become Gartner (2). This report was written by Doug Laney in 2001 and presented the challenge of the “3D Data management” which evolved into the 3Vs concept: Volume – Variety – Velocity. Other Vs have been proposed since then, such as Veracity. All Vs refer to data: the challenges that faces the information society are bound to great Volumes of Various and heterogeneous data to process in real-time (Velocity). Practically, there is no consensual definition of big data. Even if many recognized the 3Vs terminology, not all understand the same meaning for each V. Volume may be the most cited property of big data, maybe because echoing to the “big” part. What is Volume? If data are figured by a 2-dimensional table (columns for variables and lines for observations or patients) is a big Volume of data a table with many observations (lines) or with many variables (columns)? Or both? Baro and colleagues also suggested defining Volume as a combination of both these dimensions (3) (figure1). The distinction seems theoretic but has strong implications, since it is the number of variables that defines the dimensionality of data, and not the number of observations (see part 2 – the curse of dimensionality). Variety mostly accounts for the heterogeneity of data. Heterogeneous data are data

made up of data collected from various methods and measures: clinical data, biological data, imaging data or genetic data and so on but also presented with various formats: from highly structured data, all numerically coded and standardized, to totally unstructured data such as handwritten texts in natural language. Velocity specifies how quick data can be accessed or processed: asynchronously, synchronously, in real time and so on.

If we cannot define uniquely big data, maybe we can track down the uses made by the companies or laboratories that claim practicing routinely big data. Big data is used by the so-called GAFA/GAFAMS: Google-Apple-Facebook-Amazon/Microsoft-Samsung. Google and Amazon have been prominent actors of the technological part of big data with the development of cloud computing and the promotion of dedicated software and frameworks such as Hadoop, mapReduce or NoSQL (4). The use of big data is then mainly oriented by the necessities of the best possible performances of search engines (Google) or the best matching strategy between advertising or services and a web user (Google, Amazon, Facebook). They use data provided when filling up a personal profile: age, gender or geolocation to adjust their services to the estimated needs of the users. In the past few years, Apple integrated a new health app to its iOS for its phones or tablets. This app can aggregate health data provided by the user or collected by tier apps. Google developed a similar app and framework for Androids (Google Fit). Firms like Google are also contracting public/private partnership to access health data. An example is given by the recent partnership between the NHS (National Health Service, United Kingdom) and Google for a better care of patients with kidney failure (5). Firms are developing platform for data gathering, storage and analysis: the artificial intelligence algorithm Watson from IBM is coupled to the Watson health cloud (6). Samsung has made available the SAMI/Artik platform as well (7). A step further is to make public the core algorithms that GAFAMS developed and use. Google (8), Facebook, Apple… have recently made their algorithm public and the code is now open source.

The emergence of big data is also associated with the rise of the social networks, e.g., Facebook or Twitter. Facebook already made a few attempts to exploit the individual data it collects for other purposes than the usual marketing uses. It got criticized for its unethical use of personal data in a published study about the impact of social network on users' mood (9). In addition to these areas, we can mention the example of finances and markets, with their automated processing of vast amounts of data in real time, signing the advent of the high frequency trading. Last, physicists working on fundamental particles are used to manipulating tons of data and adapting their framework to process them which led to the recent experimental discovery of the mediatized Higgs boson.

In comparison, big data in medicine is a relatively new topic, where big data is usually assimilated to precision, predictive or personalized medicine.

1.2 Big data in medicine: conceptions and misconceptions.

If the Pubmed database is searched for articles dealing with big data from inception to date (07/31/2017), about 3500 articles are retrieved, the first ones having been published in 2008 in Nature. Their number is growing on a steady pace since 2012 (2012: 41 articles, 2013: 201, 2014: 463, 2015: 723, 2016: 1186 and 2017: 825). The vast majority are position papers, personal opinions

and perspectives, with rare but valuable exceptions (10,11). Even if considered papers are based on a broad definition of big data, only a handful of publications are research articles and most of them are technical papers, e.g., describing an algorithm dedicated or adapted to big data. Most researchers seem ready to enter the big data game and subscribe to its promises while a few try to temper the many high expectations (12–14). Of course, physicians and researchers are tempted to include all genome-wide association studies (GWAS) into the big data paradigm (3). It is artificial since the only "big" thing in GWAS is the amount of data. The situation is nonetheless ambiguous because big data is seen as the means to develop precision, predictive or personalized medicines, which in turn are mostly if non-exclusively based on genomics or other "omics".

The new promise of genomics in the big data perspective is to deliver highly personalized risk factors, meaning that for one person, physicians should be able to "predict" the occurrence of such or such condition accurately. The question of whether it could apply to very specific and rare diseases or to from the least common to the most common of pathological conditions is not yet answered – and not really debated.

Another remarkable aspect of the use of big data in medicine would be found in epidemiological studies and epidemics monitoring, such as the avian flu pandemics. When Google claimed its search engine performed better than the Centers for Disease Control (CDC) (15), it was then opposed that it was not doing it accurately and only provided correct trends, not precise estimates (16). Additionally, epidemiologists do not seem unanimously enthusiastic about the big data perspective (17,18), although they may be the first medical scientists to have ever approached the principles of a big data dedicated to medicine. So far, big data and medicine have therefore many promises to keep.

2 – Is big data less than a revolution?

Before big data became a hot topic in medicine, several fields of research provided tools to deal with huge amounts of data, heterogeneity and complexity. Artificial intelligence, cognitive sciences or information theory were among the first to get interested in coping with such challenges. From these fields, we inherited several unanswered questions about different aspects of big data.

- Specific and intrinsic issues due to the essence of big data exist that we must be aware of and that should be characterized and addressed at best. We can cite among others heterogeneity, high dimensionality (many variables to deal with) and unstructured data.
- High dimensionality and data sparsity lead to unattended and unwanted mathematical behaviors. For example, the so-called curse of dimensionality states that the higher the dimensionality, the more difficult to discriminate between two individuals, whatever how different they may be from each other. Classical tools based on distances and metrics may not apply in such high dimensional spaces. Their reliability should not be taken for granted.
- The role and future of causality in science: some people argue that with big data, since we record "everything about everyone", there is no need for a science that explains the world we live in (19). From this perspective, a correlational approach is all we need to process data and derive high quality predictions about any topic of interest. Stated differently, it asks to choose between a predictive approach versus an explaining approach of science, without letting much choice of the technique. Interestingly, to date, there is no automated and

simple way to detect non-linear associations between variables (20): in many cases, all associations, all phenomena observed are then supposed to be proportional or linear and never more complex.

- What is the meaning of predicting events if there is no possible or identified means to prevent their occurrence? Indeed, if big data is only about searching for correlations and features that can predict particular events, it will not provide any clue about causal mechanisms. It seems unrealistic to ground education and rational strategies against unwanted outcomes without any form of understanding. In crime prevention, if purely predictive, how many "statistical" offenders should be watched over to avoid one crime? So far, predictive algorithms are better at predicting past events than future events, without talking about predicting novelty or adaptive behaviors (21).
- It is not obvious that « bigger is better » in every aspects and every contexts. When do we need many subjects, when do we need many variables to characterize a particular scientific problem? Does big data need to be exhaustive?
- How to extract and merge data from various and heterogeneous sources is a key issue for big data, and may not be fully satisfactory yet. Data quality and overall quality throughout the whole data processing flow is another matter of concern that still needs to be addressed.

If the advent of the big data as an outstanding service is not obvious so far, it is still true that the underlying prerequisites of such a revolution are almost fulfilled. Indeed, the automation and digitalization of many aspects both of our everyday lives and of medical data provide the elements needed for a big data approach of health. The integration of data from different sources is yet to be achieved and should among other issues integrate and overcome ethical and legal obstacles before becoming a reality. Forensic science has to face the same situation.

3 – Topics in forensic science and big data: what is at stake?

Biomedicine successively shifted from a pathology-centered paradigm to a syndrome-based representation and more recently to a risk factor approach of health. It can be read as a will to personalize medicine, to adapt prevention and treatment to each unique individual case. Forensic science also deals with personalization: from generic knowledge, one may want to apply its skills to each particular individual or case. The task is all the more difficult when it comes to diverse and entangled mechanisms that led to violence and sometimes to the death of an individual. Classical statistical approaches are not fit to address such an individual-tailored approach but a social and a judicial demand exist that have to be dealt with. Over the past decade, several authors questioned the forensic reasoning (22–25), while a Bayesian logic approach has been suggested and introduced in courts by many forensic experts (26–28). Individualization has also been discussed (29).

Beyond these considerations, all topics in forensic science can be impacted by the advent of big data. Some of them are even intimately intricate with it, such as ethics and laws, professional secrecy and privacy: the access and the publicity of personal data, with different degrees of anonymization, are matter of great concerns for health professionals and lawyers.

An unexpected asset of the application of big data to forensic science may lie in alternative methods for high levels of evidence. Indeed, because ethically or practically impossible to set up, forensic

science can rarely benefit from study designs providing the highest levels of evidence, such as randomized controlled trials. Paradoxically, there is a huge pressure on forensic science and high expectations from the society to provide the best possible quality evidence: courts and social equity call for robust results and knowledge to ground fair decisions. Since data are sparse and diverse, sometimes and somehow unique, a big data approach may help gathering and comparing data that would have never been available otherwise. Evidence will be sought, gathered and compared in texts, in clinical trials, in everyday life personal experience and assembled so that new knowledge can answer accurately and specifically many questions that arise from practice. For example, toxicology could benefit from sparse cases that have not been directly compared, published and unpublished as well as from pharmaceutical surveillance data. In the same way, autopsic findings often suffer from small series or biased study designs: sparsity and high heterogeneity could be addressed in a big data framework. Another example would be given in clinical legal medicine, where practitioners could access a more uniform framework to assess with more confidence and equity the objective and subjective aspects of the functional impairment of assault survivors, based on their own observations and skills as well as on individualized data obtained in context.

4 – A typology for big data studies and alike

For the physician and the epidemiologist, defining what is or could be a genuine big data study can be a difficult task. The most “genuine” types of big data studies are certainly the search for unexpected correlations across wide sets of characteristics and the search for weak signals in huge volumes of data. These two types apart, we can define four types of big data studies and alike.

- Enriched studies: data available from open data for example, can enrich data collected in a particular study, initially designed for a specific purpose, e.g., a clinical trial assessing an intervention for patients with COPD for which road traffic and air pollution data have been added.
- Combined studies: personal data, for example derived from mobile devices such as smartphones or medical devices, can be used in the context of a particular study, initially designed for a specific purpose, e.g. a clinical trial.
- Augmented studies: data collected with classical clinical, epidemiological studies or coming from databases (claims, healthcare institutes and hospitals) are reused or analyzed with non-classical, nonlinear methods such as machine learning methods, predictive and causal methods (30)
- Virtual interventions: data collected for other purpose than research are reused to simulate clinical trial and interventions (31). It could be achieved considering that in all existing practices, many differences can occur in prescribing drugs and then, “virtual” interventional groups and control groups can be identified that allows comparing efficacy and effectiveness.

The current international trends concerning the alliance between classical studies and big data studies seems to be the use of electronic health records (EHR) to enrich clinical trials, if not to be a complete substitute to the classical approach (32). Concerns are raised about ethical, legal issues but also about data quality and adequacy to such a use.

5 – A Special Issue dedicated to big data in medicine and forensic science: a JFLM initiative

What could bring a big data approach to forensic science? That was the question we wanted to answer – at least, we wanted to search for practical answers. Several contributors answered positively to our call for a special issue dedicated to big data in medicine and forensic science, others acknowledged the importance of addressing this topic while not being in position to date to contribute to this first effort. We wanted to invite researchers with true openness to other application domains than their usual ones, while being highly skilled in one or several aspects of big data. Naturally, we also invited forensic scientists and practitioners to contribute. We thank all of them for their contribution.

We felt important to first delineate and define the current landscape and what is at stakes for big data in medicine and forensic science: that is the main purpose of this text. Big data presents some fundamental and technical aspects that makes it different from classical studies: the mathematician PA Maugis (University College London, UK) explains what some of the main pitfalls are to avoid when practicing big data analysis (33).

Big data offers our community to switch from local and daily practice to data and evidence-based practice and multicentric, interdisciplinary research. Research and practice will be more and more intricate in the future. A forensic practitioners team, Dang and colleagues (Jean Verdier teaching hospital, France), explains how data can be captured in a daily practice in clinical legal medicine (34). Data scientists, Laugier and colleagues (Tekliko, Singapore), show how to take advantage of any kind of data collections, such as presented by Dang. They illustrate how clinical legal medicine can as soon as today go individualized or contextualized using the appropriate tools in daily practice (35). Their tool also integrates new techniques for high dimensional data vizualisation, previously introduced in (36). Bioinformaticians, Jaulent and colleagues (Inserm, France), go further and suggest how going global for data sharing and interoperability between practices and systems (37).

Examples provided by Dang, Laugier and Jaulent and their colleagues are so far mostly based on data collection at the level of forensic unit or department, or at the level of a gathering of such entities. Obviously, these sources of data are not the only ones available and usable. The sociologist and criminologist DeLisi (Iowa State University, USA) demonstrates that large, existing databases can be used to connect diverse aspects of a person's life, for instance in criminology (38). Smidt et al (Amsterdam Public Health Service, Netherlands) used a similar approach in a linkage study to investigate associations between the health of former police detainees in Amsterdam and their cause of death (39). To this, the big data researcher Liu and psychologist Young (University of California, Los Angeles, USA) add the new opportunity to connect devices to broaden the classical sources of personal data and illustrate their purpose by writing about social media data analysis (40). Also, data and databases linkage are a pivotal issue in big data. Another example is given by epidemiologists, Rey and colleagues (Inserm, France), regarding a better system for registering and elucidating causes of death (41). This example also echoes with Jaulent and colleagues' work about system and databases interoperability. Mujtaba and colleagues (Department of Information Systems, Kuala Lumpur, Malaysia) also explain how they analyzed autopsy reports with automatic text analysis to extract the causes of death (42). Finally, if big data is often said to be about personalization, we should not forget that individuals are not living alone and without any contact with each other: mathematician and social epidemiologists, Dimeglio and colleagues (Inserm, France), go further than

the individual horizon and expose why social and collective characteristics are important in big data (43).

Of course, many other aspects could have been presented in the present special issue. Among the most important ones, we could cite: genomics and metagenomics, toxicology, Bayesian frameworks, thanatology, mental health and forensic reasoning. We hope that part or all of these topics will be addressed in future issues and that the present articles collection will encourage researchers to share their point of view and work, present or future.

6 – A call for definitions and actions

Now that we got more familiar with several and somehow undetermined aspects of big data, it may be time to assess its relevance to medicine and forensics sciences in a scientific, rigorous way. We suggest several actions to ensure the development of a fair, useful and sustainable big data framework in forensic science (table 1). Nonetheless, this work cannot be the work of a few people but should be grounded on a broad concertation.

Big data is an opportunity for researchers to practice a genuine interdisciplinary approach of their work, based on both observations and evidence and techniques adapted to handle vast amounts of heterogeneous, unstructured and distributed data. Big data in medicine should be grounded on both personalized approaches, at several scales (genetic, phenotypic, epigenetic and psychological scales) and social approaches, all based on observations and aiming at predicting and explaining events.

More specifically, big data is for forensic science an unprecedented means for reuniting research, practice and education, both for health professionals and patients. It can provide an excellent framework that abolishes frontiers between narrower specialties, e.g. toxicology, thanatology or victimology and that allows every practitioners working with common, standardized tools on evidence data. It should encourage transparency in research and practice methodologies, in data and expertise sharing, and enhance the reproducibility capability that any science needs to remain sound and sane. Finally, it may favor international collaborations for the best of this field.

Acknowledgments

None.

Conflicts of interest

None.

References

1. Ward JS, Barker A. Undefined By Data: A Survey of Big Data Definitions. ArXiv13095821 Cs 2013 http://arxiv.org/abs/1309.5821 Accessed July 31st 2017

2. Laney D. 3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf Accessed July 31st 2017

3. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. BioMed Res Int. 2015;2015:639021.

4. Goli-Malekabadi Z, Sargolzaei-Javan M, Akbari MK. An effective model for store and retrieve big health data in cloud computing. Comput Methods Programs Biomed 2016;132:75‑82.

5. Hawkes N. NHS data sharing deal with Google prompts concern. BMJ 2016;353:i2573.

6. ODH, Inc. and IBM Watson Health Introduce Mentrics, a Population Health Management Platform to Transform Behavioral Healthcare. http://www-03.ibm.com/press/us/en/pressrelease/49564.wss Accessed July 31st 2017

7. https://www.artik.io/2016/03/artik-and-sami-form-an-integrated-smart-home/ Accessed July 31st 2017

8. https://www.tensorflow.org/ Accessed July 31st 2017

9. Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci U S A 2014;111(24):8788‑90.

10. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med 2015;372(26):2481‑98.

11. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med 2015;3(1):42‑52.

12. Joyner MJ, Paneth N. Seven questions for personalized medicine. JAMA 2015 dx.doi.org/10.1001/jama.2015.7725 Accessed July 31st 2017

13. Coote JH, Joyner MJ. Is precision medicine the route to a healthy world? The Lancet. 2015;385(9978):1617.

14. Kohane IS. Ten things we have to do to achieve precision medicine. Science. 2015;349(6243):37‑8.

15. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012‑4.

16. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science 2014;343(6176):1203‑5.

17. Chiolero A. Big data in epidemiology: too big to fail? Epidemiology 2013;24(6):938‑9.

18. Toh S, Platt R. Big data in epidemiology: too big to fail? Epidemiology 2013;24(6):939.

19. Sterling B. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. WIRED 2008. http://www.wired.com/2008/06/the-end-of-theo/ Accessed July 31st 2017

20. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting Novel Associations in Large Data Sets. Science 2011;334(6062):1518‑24.

21. Schwartz SE, Charlson RJ, Rodhe H. Quantifying climate change — too rosy a picture? Nat Rep Clim Change 2007;(0707):23‑4.

22. Jackson G, Jones S, Booth G, Champod C, Evett IW. The nature of forensic science opinion--a possible framework to guide thinking and practice in investigations and in court proceedings. Sci Justice J Forensic Sci Soc 2006;46(1):33‑44.

23. Houck MM. Intellectual infrastructure: a modest critique of forensic science. Sci Justice J Forensic Sci Soc 2013;53(1):1.

24. Dror I. The ambition to be scientific: human expert performance and objectivity. Sci Justice J Forensic Sci Soc 2013;53(2):81‑2.

25. Taroni F, Biedermann A, Garbolino P, Aitken CGG. A general approach to Bayesian networks for the interpretation of evidence. Forensic Sci Int 2004;139(1):5‑16.

26. Biedermann A, Bozza S, Garbolino P, Taroni F. Decision-theoretic analysis of forensic sampling criteria using bayesian decision networks. Forensic Sci Int 2012;223(1‑3):217‑27.

27. Biedermann A, Taroni F. Bayesian networks and probabilistic reasoning about scientific evidence when there is a lack of data. Forensic Sci Int 2006;157(2‑3):163‑7.

28. Taroni PF, Aitken C, Garbolino PP, Biedermann DA. Bayesian Networks and Probabilistic Inference in Forensic Science. Chichester, England ; Hoboken, NJ: Wiley-Blackwell; 2006. 372 p.

29. Biedermann A, Garbolino P, Taroni F. The subjectivist interpretation of probability and the problem of individualisation in forensic science. Sci Justice J Forensic Sci Soc 2013;53(2):192‑200.

30. Lefèvre T, Lepresle A, Chariot P. Detangling complex relationships in forensic data: principles and use of causal networks and their application to clinical forensic science. Int J Legal Med. 2015;129(5):1163-72. doi: 10.1007/s00414-015-1164-8

31. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016;183(8):758‑64.

32. Fiore LD, Lavori PW. Integrating Randomized Comparative Effectiveness Research with Patient Care. N Engl J Med 2016;374(22):2152‑8.

33. Maugis PA, Big data uncertainties. Doi : 10.1016/j.jflm.2016.09.005

34. Dang C, Phuong T, Beddag M, Vega A, Denis C, A data model for clinical legal medicine practice and the development of a dedicated software for both practitioners and researchers. doi: 10.1016/j.jflm.2016.11.002

35. Laugier V, Stindel E, Lichterowicz A, Ansart A, Lefèvre T. Making the best of data derived from a daily practice in clinical legal medicine for research and practice – the example of Spe3dLab. 2017. arXiv:1707.08454

36. Lefèvre T, Chariot P, Chauvin P. Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons. Forensic Sci Int. 2016;266:581.e1-9. doi: 10.1016/j.forsciint.2016.05.014

37. Jaulent MC, Leprovost D, Charlet J, Choquet R. Semantic Interoperability Challenges to process Large Amount of Data - Perspectives in Forensic and Legal Medicine. Doi: 10.1016/j.jflm.2016.10.002

38. DeLisi M, The Big Data Potential of Epidemiological Studies for Criminology and Forensics. Doi: 10.1016/j.jflm.2016.09.004

39. Smidt DB, Dorn T, Reijnders U. A record linkage study on former police detainees who died in Amsterdam between 2013 and 2015. JCFM17-88R1

40. Liu S, Young SD, A Survey of Social Media Data Analysis for Physical Activity Surveillance. Doi: 10.1016/j.jflm.2016.10.019

42. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of Cause of Death from Forensic Autopsy Reports using Text Classification Techniques: A Comparative Study. Doi: 10.1016/j.jflm.2017.07.001

41. Rey G, Bounebache K, Rondet C. Causes of deaths data, linkages and big data perspectives. Doi: 10.1016/j.jflm.2016.12.004

43. Dimeglio C, Kelly-Irving M, Lang T, Delpierre C. Expectations and boundaries for Big Data approaches in social medicine. Doi: 10.1016/j.jflm.2016.11.003

Figure 1 What is the big part in big data? Volume and variety as cardinal characteristics of big data.

Volume has not much interest if defined as a lot of lines or observations, but few columns or variables. Dimensionality of data is defined as the number of variables needed to describe a phenomenon. Raw data dimensionality is therefore the number of variables (number of columns). Additionally, genome-wide associations studies are based on the whole human genome, and there are as many variables as genes. There are not “genuine” big data studies either since they do not present variety of sources and formats of data (structured, unstructured, textual, numerical…).

Table 1 Actions to develop a fait and sustainable big data framework in forensic science

To ensure a fair and sustainable big data environment for forensic science, a minimal set of actions and concerted efforts must be decided. Three main areas of actions should be considered at least: i) infrastructure and information processing capabilities, ii) training, skills and literacy and iii) law and ethics.

Areas of action	Actions
Infrastructure and information processing capabilities	- Develop shared terminologies and ontologies for interoperability o e.g., for automated extraction of standardized information and variables from distributed, unstructured and heterogeneous data in different languages (medical reports, scientific publications…) - Develop shared procedures to ensure data access, security and processing - Ensure data quality and adequacy to reuse case - Organize an effective data sharing and resources reporting o At least, build a shared repository listing all available datasets and resources in forensic science o On top of that repository, build a gateway for searching adequate data and access data processing capabilities
Training, skills and literacy	- For researchers: o Develop adequate, shared methodologies and elaborate shared standards of good practice - Assess the digital literacy of all actors: o Physicians, victims, perpetrators, police officers, policy makers, magistrates and lawyers - Develop capability to adjust literacy of all actors for a fair use and understanding of big data studies and products
Law and ethics	- Respect of high standard of ethics for the access and reuse of data - Investigate the impact of digital tools and products and data-based evidence in court and along the whole judiciary process

2 columns = 2 variables
=
dimensionality of data

Height	Weight
172	62
161	80
194	90
175	75
163	54
186	84
...	...
...	...
...	...
201	110
153	49

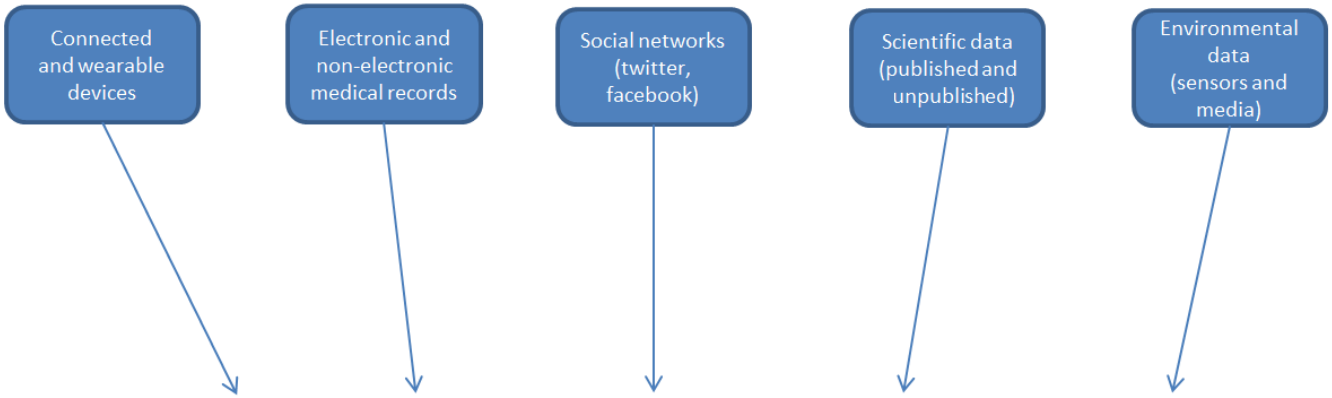
n lines
=
n observations

p columns = p variables
=
dimensionality of data

Height	Weight	Gender	Age	Eyes color	Hair color	Blood type	Rhesus	Blood pressure	...	...	...	X
172	62	M	62	G	Bl	A	-	115	...	...	...	E
161	80	M	18	B	Bck	B	+	121	...	...	...	R
194	90	F	26	Br	Bck	O	-	132	...	...	...	F

3 lines
=
3 observations

Variety of data sources and variety of formats contribute to define big data



Height	Weight	Gender	Age	Eyes color	Hair color	Blood type	Rhesus	Blood pressure	...	...	...	X
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...

Highlights

- Big data has yet no consensual definition
- Big data is full of promises that must be assessed
- Specific questions arise from data linkage and reuse with new techniques
- Big data is an opportunity to develop interdisciplinarity in forensic science
- We suggest several actions for a fair and useful big data framework