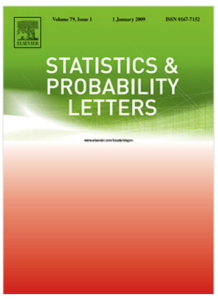
Accepted Manuscript



A practical guide to big data

Ekaterina Smirnova, Andrada Ivanescu, Jiawei Bai, Ciprian M. Crainiceanu

PII: S0167-7152(18)30059-2
DOI: https://doi.org/10.1016/j.spl.2018.02.014
Reference: STAPRO 8136

To appear in: *Statistics and Probability Letters*

Please cite this article as: Smirnova E., Ivanescu A., Bai J., Crainiceanu C.M., A practical guide to big data. *Statistics and Probability Letters* (2018), https://doi.org/10.1016/j.spl.2018.02.014

A practical guide to big data

Ekaterina Smirnova* Andrada Ivanescu† Jiawei Bai‡ Ciprian M. Crainiceanu§

November 9, 2017

Abstract

Big Data is increasingly prevalent in science and data analysis. We provide a short tutorial for adapting to these changes and making the necessary adjustments to the academic culture to keep Biostatistics truly impactful in scientific research.

1 Introduction

Big Data has been analyzed for a long time. Indeed, in a 1938 landmark paper, Raymond Pearl [9] used data on 6,813 men (2,094 non-smokers, 2,814 moderate smokers, and 1,905 heavy smokers) to show that tobacco smoking was "statistically associated with the impairment of life duration, and the amount of this impairment increased as the habitual amount of smoking increased". It took until January 11, 1964, for Luther L. Terry, M.D., Surgeon General of the U.S. Public to officially acknowledge that "cigarette smoking was cause of lung cancer and laryngeal cancer in men, a probable cause of lung cancer in women, and the most important cause of chronic bronchitis". In 1982 Allan Gittelsohn [5] published results on the distribution of underlying causes of death in the US using 21 million death records from 1968 to 1978. Currently, Biostatisticians routinely work with hundreds of Terabytes of data from genomics, brain imaging, and wearable sensors. Thus, one could think that the "Big Data" phenomenon is not new and is just a clever rebranding of the analysis of ever larger datasets generated by increasingly sophisticated new technologies. However, this would not explain the explosion in popularity of Big Data. What could explain it is the large amount of money it can generate when analyzing who will click a "like" button, what advertising to provide to a net surfer, or what smart phone to recommend to an online shopper. The sheer sexiness of money makes Big Data cool. We believe that this excitement should be captured, embraced, and directed to solving important societal problems. In this short paper we try to provide a practical guide to doing that, mention a few tautologies, and identify a few arbitrage opportunities. The recipe is simple, though the implementation is difficult because it requires actual work.

*Assistant Professor, Department of Mathematical Sciences, University of Montana, 32 Campus Dr, Missoula, MT 59812. E-mail: ekaterina.smirnova@mso.umt.edu

†Assistant Professor, Department of Mathematical Sciences, Montclair University, 1 Normal Avenue Montclair, NJ 07043; E-mail: ivanescua@montclair.edu

‡Assistant Scientist, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. Baltimore, MD 21205 USA. E-mail: jiawei.bai@jhu.edu

§Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. Baltimore, MD 21205 USA. E-mail: ccraini1@jhu.edu

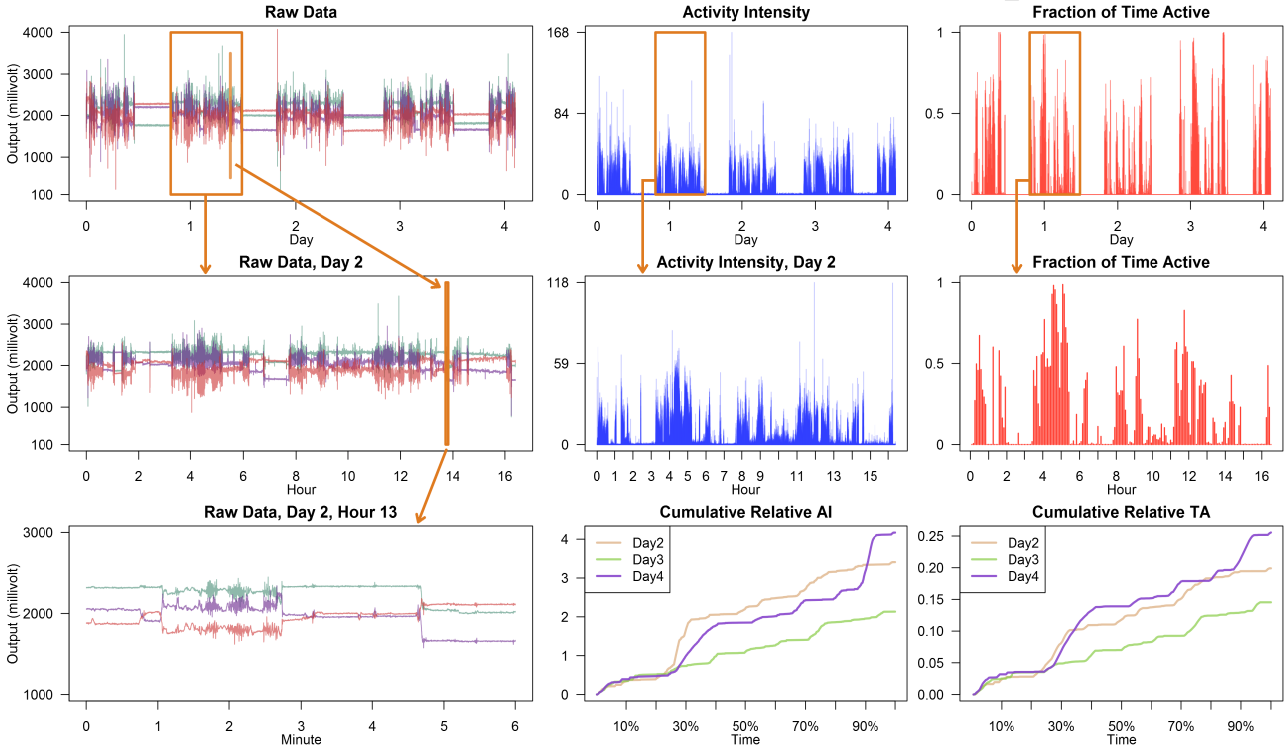


Figure 1: Accelerometry data for one subject followed over 5 days.

2 Place data front and center

To a junior investigator in search of a meaningful career there is nothing more powerful and interesting than seeing an important scientific problem and the data that could be interrogated, explored, and analyzed to solve that problem. Thus, being concrete is an important first step and we will do just that by showing one of the most exciting developments in Biostatistics: data generated by wearable and implantable technology (WIT). Figure 1 displays accelerometry data collected at a frequency of 10Hz for five days from a sensor placed on the hip of a person [3]. The top left panel shows data measured along three orthogonal axes (up-down, left-right, backward-forward in the device frame of reference). The five days correspond to long periods of higher amplitude signals that are clearly separated by four nights characterized by low amplitude signals. Thus, for this subject there are ≈ 13 million observations. To get a closer view, the orange box in the top-left panel in Figure 1 identifies day 2 of the data and a zoom-into these data is shown in the left-middle panel. An orange yellow line indicates a period of six minutes during day 2. This period is zoomed-into the left-bottom panel. As one looks at finer resolutions of the data, more patterns occur, could be identified and, possibly, used. These raw data are expressed in millivolts (mV), though most devices output raw data in Earth gravitational units ($g = 9.81\text{m/s}^2$). Working directly with raw data could be quite daunting and, in practice, data are often summarized as activity counts (or steps) per minute. The middle-top panel in Figure 1 provides such a summary measure at the minute level, while the middle-center panel displays the same measure for day 2. The minute-level data tends to be strongly skewed, very spiky, and highly non-stationary. While informative, overlaying such visualizations in the same panel will lead to over-plotting and loss of information when comparing different days or subjects or when displaying an

entire cohort. Instead, the middle-bottom plot shows the cumulative measure of activity up to a particular time of the day. This panel contains exactly the same information as the middle-center panel, but allows for joint plotting of multiple days and subjects. The right panels display similar information, though they focus on the proportion of time active per minute instead of activity intensity during that minute. The proportion of time active is obtained by calculating the activity intensity at the second level, applying a threshold on activity intensity that indicates active/inactive, and then computing the proportion of active seconds within that minute.

For large studies such as the UK Biobank, NHANES or BLSA, accelerometry data are collected for up to hundreds of thousands of participants, which creates very Big WIT Data. Moreover, the current trend is to conduct high frequency biological signal monitoring well beyond one or two weeks and use additional sensors, including heart rate monitors, GPS, video cameras, ecological momentary assessment (EMA), glucose monitors, or environmental polutant sensors. Continuous monitoring of WIT information produces a fundamental shift in volume, variety and velocity of Big Data. Indeed, we are witnessing one of the most exciting developments in measurement: large number of sensors that are silently producing enormous amounts of information, some of which may be relevant to health. The broader goals for collecting such data sets are to: 1) obtain detailed objective measurements of activity at the individual level in large samples; 2) quantify the patterns of activity and variability within- and between-individuals in the population; 3) extract scientifically meaningful features from the rich activity data; 4) assess potential associations between patterns of activity and health; and 5) conduct discovery research that could lead to new scientific hypotheses to be investigated in more targeted studies.

3 Obtain access to data when it matters

Fast access to data when the scientific community and the society care about the information contained in it provides a large advantage to individual research groups. Statisticians should agree that impact comes primarily from scientific findings and that analytic methods are the tools for extracting those findings. Indeed, there is nothing sadder than a paper analysing data collected during the last century and showing a slight improvement over the other 30 methods that analyzed the same data set. Statistical methods are important, but they should be viewed as the means for extracting information and not the end goal of research. Of course, fast access to data is not typically given away, but it can be earned by building long-term collaborations with high quality researchers.

The accelerometry data shown in Section 2 was obtained directly from the investigator who ran the experiment and whose initial goal was to explore whether specific activity types (e.g. walking, standing up from a chair, stair climbing) can be predicted from raw accelerometry data. In time, these goals have evolved and become more refined, but the focus remained on quantifying the subject- and population-specific characteristics of the accelerometry data and their potential associations with human health. To articulate and address these goals the team needed to work for many years to understand the measurement, conduct data pre-processing, and ensure data quality. This was done in direct and continuous collaboration with our scientific collaborators and provided us timely access to data. This effort provided our team a seat at the research table and allowed us to shape decisions about experimental design as well as data pre-processing and analysis.

Data pre-processing is a huge task that takes a lot of time and requires skillful and

computationally savvy Biostatisticians. Indeed, the effort required by this task could easily dwarf the modeling and analytic effort. Unfortunately, it has seldom been acknowledged or rewarded, most likely because the importance of this crucial task has not been historically recognized by Bio/Statistics Departments. The problem has been exponentially magnified by Big Data, which requires much larger efforts for pre-processing and organization. Therefore, not having early access to Big Data can substantially erode the relevance of downstream analyses, which have been tradionally conducted by Biostatisticians. We contend that pre-processing is part of the over-all Statistical analysis and the effects of incorrect or sub-optimal pre-processing could be devastating to downstream analyses; see, for example, pioneering work exposing pre-processing pitfalls [1, 2, 4, 7, 12]. Thus, we anticipate that teams that create and maintain well organized big data sets will have a competitive advantage over teams that wait for well organized data sets and well defined scientific questions. Indeed, why would a team that has the skills to organize and pre-process a vast and complex data set, such as a large cohort study involving accelerometry, not take the final step of analyzing and publishing the results?

4 Solutions are simple, but require actual work

Becoming and staying relevant in a fast moving data-centric world requires a few simple steps: 1) be involved with the best scientists and work on the most important scientific problems; 2) build a team dedicated to taking on multiple aspects of the problem simultaneously and in real time; 3) implement a multi-dimensional training system, where new members get involved quickly into projects and are mentored by colleagues and supervised by experts; 4) continuously search for and identify passionate, hard working researchers who are convinced of the importance of their work; 5) foster the diversity of ideas and honest-speak using a failure-tolerant atmosphere where ideas are considered, evaluated, discarded and recycled. This is simple to enumerate, but requires a lot of real work, daily interactions with collaborators, and exposure of students and faculty to important problems. The senior mentor/s should continue to get their hands dirty with data, get involved, and keep the pace with computational developments. This is necessary to keep the development of methods realistic and avoid falling into the over-complexification trap. At the end of the day, if the researcher cannot explain to their family what they are working on and why it is important then they should change the topic of their research.

The cornerstone of the solution is to build and maintain meaningful long term collaborations, which raises a legitimate question: how do you actually do that? The answer is to keep the office door and own mind open to new ideas, search for collaborators, and actively work with them on developing a common language. Identifying areas that are of mutual interest is crucial and requires special atention. Involving students early, bringing them to meetings, and slowly giving them more responsibilities tends to solidify the collaboration established during the original discussions. Working with the researchers to clarify the scientific questions, explore the data and decide what questions can be answered given the existent data, and proposing data collection solutions that could address remaining or emerging scientific questions could help maintain collaborations. We have discussed this process as the soft null hypothesis [11], which takes a softly defined scientific concept and transforms it into a simple, clean, addressable scientific question. To build trust and genuinely learn about the scientific area of research, we suggest to visit the lab of the collaborator together with students and junior investigators. The last author routinely takes his students to witness

surgeries or assist researchers conducting WIT research. Even if all these suggestions are implemented, one should remember that collaborations can fail. This is a normal part of the process, not unlike the social process of making friends: both aquiring and maintaining collaborators requires a lot of positive effort from both sides.

5 Methods follow problems, not the other way around

The scientific problem should be the focus of the investigation and existent methods should be refined and new methods developed to solve the scientific problem. This provides a level of clarity and sense of purpose that can be highly motivating for new investigators. Unfortunately, Statistical training is often method-, theory-, and algorithm-centric, which favors over-complexification, opaque mathematization, and may induce detachment from data reality. For example, during a visit at a famous Statistics department the last author was approached by an investigator who mentioned that he developed new prediction methods that are superior to existing approaches. The discussion was left at that, though it was obvious that the investigator made bold, context-free claims about the prediction performance of his methods. A few months later the last author received an email from the investigator who sent an R [10] package "to be applied to his EEG data". This would have been hilarious if it were not the manifestation of a much bigger problem: the development of methods in the search of an application. It is infinitely more satisfying to start with a problem, identify the hard scientific and statistical problems, define and attempt to solve them. Indeed, putting data and problems first and identifying methods that are useful later is the correct approach. This is especially true because Statistics is a mature field that already has an enormous array of data analytic methods. Assuming and embracing the maturity of our field and taking on challenging scientific problems is the simple, but hard, way forward. We conclude that the real difficulty is dealing with actual data, defining and clarifying the associated scientific problems, and providing clean Biostatistical solutions; not the mathematics.

6 There is no method without reproducible code

The importance of software should be acknowledged and rewarded. The definition of a method should include published, reproducible code that illustrates it, provides detailed examples, and is updated and maintained. Indeed, this is the only way to prove that ideas are not falling apart when new data analytic requirements meet theoretical concepts. This has become increasingly necessary with the rise of unnecessarily complicated models and formulae. The use of software can highlight serious problems lurking in otherwise unimpeacheable rationalizations of complexity and expose the fragility of pesty tuning parameters. If change is to be enacted then our own thinking about the importance of software development needs to change and become an important factor in recommendations of acceptance or rejection of papers and grant applications.

While software development should become a requirement for methods development, it remains largely unresolved how it should be assessed and rewarded by the community. Some of the solutions would be for Departments to provide explicit incentives for software development and for the community to develop impact factors for software, similar to the impact factors for papers. For example, it may be important to report how many papers use the software and how many scientific papers use the Statistical methodology and its

associated software. Methods that are developed just to impress fellow Statisticians, never make it into the mainstream, and are never used in scientific applications should simply be allowed to retire.

One could argue that writing code is no longer a serious problem in Statistics and that there has been an explosion in the number of packages as well as a marked improvement in their quality and impact. Indeed, R is probably the most important success of Statistics as a science. However, the system also suffers from having too many packages and a so-and-so ratio of decent to exceptional packages. To draw an analogy, the R world is like a semi-organized library with a huge number of books, but without clear catalogues for specific areas of interest. The solution is, of course, organization. This has been done successfully in Genomics by Bioconductor [6], which organizes the software for Genomics research. However, there are many emerging areas of research and reproducing Bioconductor is quite cumbersome and expensive. To address requirements specific to biomedical imaging we have developed Neuroconductor [8], which is based on GitHub and continuos integration software such as TRAVIS CI and AppVeyor. However, such efforts should be started in many different areas including WIT, electronical medical records (EMR), personalized medicine, and functional data analysis. Organizing the R packages dedicated to specific areas of research, imposing a minimum requirement of quality checks, and providing training materials is crucial to conducting high level Statistical research and creating impact. We consider that the community should invest in low-cost systems, such as Neuroconductor, to organize and provide up-to-date educational and computational materials developed for specific scientific areas. The R task views are a step in the right direction, but they do not provide all the necessary support to move entire fields computationally forward.

7 The incentive system

Alas, all is for naught if the incentive system is outdated. Indeed, junior investigators are smart and understand exactly what is expected for fast and painless promotion. Most of them are told that they need to publish many papers in the top 5 or 6 Bio/Statistical journals, which favor theory. Thus, the most creative 10 to 12 year period of the life of a Bio/Statistician is spent on things that have nothing to do with data, in general, or Big Data in particular. Indeed, some Departments go as far as placing the Annals of Statistics above other Statistical journals, most likely because it is a more theoretical journal. This message is unmistakable and can only lead to more junior investigators being disengaged from data, be it big, small or, even, moderate.

Instead, the incentive system needs to evolve, become more flexible, and more inclusive. In particular, Departments should encourage their faculty to publish in top tier journals, *irrespective of the area of research*. The idea that publishing in applied journals is easier should be abandoned and replaced by the requirement for the Bio/Statistician to be a leader in whatever area of scientific research they choose to work in. For example, if a Biostatistican works on nutrition research then they should participate in the most important conferences in nutrition research, be recognized in, contribute to, and drive the methodological research underlying nutrition research. In this context Statistical novelty may not be an absolutely new model or theory, but a carefuly crafted, targeted, highly impactful contribution to a new area of science. Novelty does not spring only from completely new methods, but also from carefully tuning and polishing existent methods for entire new areas of science and disseminating these approaches to our collaborators and colleagues. Simply waiting for

somebody to *discover our research and apply it into their area* is a very low probability shot in the increasingly faster game of scientific research. If we agree that these are the hard truths then solutions are relatively simple: 1) count papers in scientific journals at the same level with those in Bio/Statistical journals; 2) establish the relevance of the research by the position of the author (first for leading, last for senior); 3) actively and publicly encourage junior investigators to become leaders at whatever they choose to do; 4) build an atmosphere of tolerance for diversity of research foci; 5) reward faculty for building long term solid collaborations with world-leading scientific researchers.

8 Acknowledgement

We would like to thank one anonymous reviewer for their excellent comments that helped improve the first draft of the manuscript. All remaining typos or controversial statements are our own. We only apologize for the typos.

References

[1] J.M. Akey, S. Biswas, J.T. Leek, and J.D. Storey. On the design and analysis of gene expression studies in human populations. *Nature Genetics*, 39(7):807–808, 2007.

[2] K.A. Baggerly and K.R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, 3:1309–1334, 2009.

[3] J. Bai, J. Goldsmith, B. Caffo, T.A. Glass, and C.M. Crainiceanu. Movelets: A dictionary of movement. *Electronic Journal of Statistics*, 6:559578, 2012.

[4] J. Carp. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage*, 63(1):289–300, 2012.

[5] A.M. Gittelsohn. On the distribution of underlying causes of death. *American Journal of Public Health*, 72(2):133–140, 1982.

[6] Huber, W. and Carey, V.J. and Gentleman, R. and Anders, S. and Carlson, M. and Carvalho, B.S. and Bravo, H. C. and Davis, S. and Gatto, L. and Girke, T. and Gottardo, R. and Hahne, F. and Hansen, K.D. and Irizarry, R.A. and Lawrence, M. and Love, M.I. and MacDonald, J. and Obenchain, V. and Ole's., A.K. and Pag'es, H. and Reyes, A. and Shannon, P. and Smyth, G.K. and Tenenbaum, D. and Waldron, L. Morgan, M. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015.

[7] J.T. Leek, R.B. Schrapf, H. Corrada-Bravo, D. Simcha, B. Langmead, E. Johnson, D. Geman, K. Baggerly, and R.A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11:733739, 2010.

[8] J. Muschelli, J.-P. Fortin, A. Gherman, B. Avants, B. Whitcher, J. D. Clayden, B.. Caffo, and C.M. Crainiceanu. Neuroconductor: An R Platform for Medical Imaging Analysis. *Biostatistics*, page Submitted, 2017. preprint on webpage at `http://works.bepress.com/john_muschelli/6/`.

[9] R. Pearl. Tobacco smoking and longevity. *Science*, 87(2253):216–217, 1938.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[11] H. Shou, R.T. Shinohara, H. Liu, D.S. Reich, and C.M. Crainiceanu. Soft null hypotheses: A case study of image enhancement detection in brain lesions. *Journal of Computational and Graphical Statistics*, 25(2):570–588, 2016.

[12] S.C. Strother. Evaluating fMRI preprocessing pipelines. *IEEE Engineering in Medicine and Biology Magazine*, 25(2):27–41, 2006.