Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Big data naturally rescaled

Ruedi Stoop[a,*], Karlis Kanders[a], Tom Lorimer[a], Jenny Held[a,b], Carlo Albert[b]

[a] *Institute of Neuroinformatics and Institute of Computational Science, University of Zürich and ETH Zürich, Winterthurerstr. 190, 8057 Zürich, Switzerland*
[b] *Eawag Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland*

ARTICLE INFO

Article history:
Received 20 December 2015
Revised 18 February 2016
Accepted 27 February 2016
Available online 20 April 2016

Keywords:
Complex systems
Complex networks
Dynamical systems
Physiological networks
Power laws
Biological modeling
Clustering algorithms

ABSTRACT

We propose that a handle could be put on big data by looking at the systems that actually generate the data, rather than the data itself, realizing that there may be only few generic processes involved in this, each one imprinting its very specific structures in the space of systems, the traces of which translate into feature space. From this, we propose a practical computational clustering approach, optimized for coping with such data, inspired by how the human cortex is known to approach the problem.

1. Introduction

In ‘big data’ we deal with the empirical observation of an explosion of the amount of information available to us, where this information stretches over multiple scales of characteristics. Much of this information was already around, but simply not sensed. In the past, many approaches have attempted to cope with similar problems; particularly, methods that put a focus on unstructured data, natural language, filtering, network and Bayesian modeling, and machine learning (such as neural networks). Other approaches have emphasized the importance of parallel or hierarchically scalable computational solutions. For the definition of big data [1], it is popular to refer to: a) a significant growth in the volume, b) velocity of arrival and c) variety and variability of data. For more than general recommendations and solutions, this seems, however, a too general problem formulation. The ‘big data’ we have in mind has several concrete characteristics, though not necessarily all of them at the same time: multi-source, multi-scale, high-dimensional, dynamic-state, and non-linear characteristics (cf. Fig. 1). These properties appear to be core aspects of the big data problem and require novel approaches and fresh views. Below, we present such an approach that is directly based on these aspects.

Current methodology in big data suggests the following sequence of processing steps: data acquisition; cleaning, extraction and annotation; aggregation, integration, and representation; modeling and analysis; interpretation. As tools, this methodology suggests to make heavy use of data-analyzing methods (e.g., data fusion, sub-sampling, filtering, dimension-reduction, sparse representation, parallel calculation, clustering; transfer-, online-, deep-learning methods, etc.). All of these methods have, however, unspoken assumptions regarding the direction the processing should push the data into; how this direction should be chosen, has, however, not been sufficiently addressed and is generally not as trivial and harmless as it may appear at first sight. Moreover, one key difficulty in big data is that many emerging data relate to compositions of signals from very different origins, recorded at the same time, but otherwise not necessarily closely related (cf. Fig. 1). The complexity of such a ‘signal space’ appears to be prohibitively difficult to get a grip on (in face of this difficulty, the mentioned data-analyzing tools are not sufficiently powerful). The wealth and breadth of data could, however, be handled if we are able to identify data structures that *intrinsically* bind data items into sets that we then can jointly process and find simple descriptors of. We push the viewpoint that, at least for a substantial subset of the big data cases related to ‘natural’ (i.e., physical) data, such generic and universal structures do exist, and can to a substantial extent, guide and substantially abbreviate the methodological sequence of data extraction, cleaning, aggregation and integration. Our hypothesis is that such big data could be much better understood, and more efficiently be dealt with, from the space composed by generic data-generating systems: the ‘system space’. The main goal of this paper is to exemplify how salient properties of the data, having simple origins in the space of systems, are traceable in the corresponding

[*] Corresponding author. Tel.: +41446353063.
E-mail address: ruedi@ini.phys.ethz.ch (R. Stoop).

http://dx.doi.org/10.1016/j.chaos.2016.02.035
0960-0779/

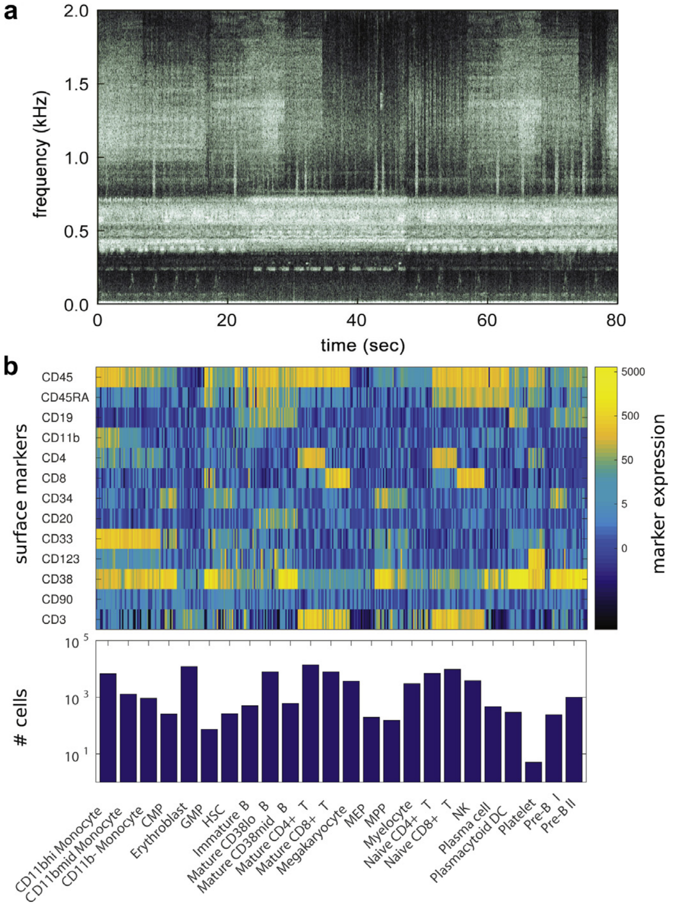

Fig. 1. Characteristic examples of big data in nature. (a) Forest ambient sound spectrogram. We are able to associate different animals to different co-temporally recorded sounds, due to sound cues. (b) Surface protein marker expression levels in healthy human bone marrow cells [2], cells (horizontal axis) already clustered into cell types (13 dimensions, high variation).

'feature space', and become, in this way, identifiable in the 'signal space'. For illustrating this perspective, we will merge earlier published [3–6] with newer, previously unpublished results, including a developed clustering tool that takes considerable advantage of these insights.

During evolution, biological systems already dealt with challenges similar to today's 'big data' problem. This leads one to expect that biology has consistently developed toward solutions (jointly in generation, reception and processing of vital signals) that are able to cope with such problems: solutions that are structurally robust in the mathematical sense, forgiving regarding detail omission and, as an essential consequence of robustness, promise to be simple. This not only strongly reduces the solution space of interest to us; the features that come along with these solutions will moreover express underlying universality. From the physics point of view, this implies that the features should follow simple scaling laws. From the modeling point of view, this implies that robust models based on a minimal amount of well-chosen measured biological data, could be used to deal with a substantial part of the signal space. In the following, we shall explore some instances where physics indeed provides such an insight and permits 'rescaling' the big data problem toward a few fundamental categories of data.

The ultimate way by which physics imprints structure into system space is through the data-generating process itself. Our foray will exhibit that while power law distributed features are natural and robust, they have only a few, but generic, origins in system space (whereas random compositions of signals from unrelated origins emerge as noise and can be naturally discarded). To explore this connection, we separate data-generating systems into: a) systems that only share a generic nonlinear building plan, but do not interact [4], b) systems that exchange nonlinear periodic interaction (our example will be Arnol'd tongues [3] or the biological cochlea [7]) and c) systems that interact in a more general, less time-dependent, manner, not requiring explicit description of node dynamics (examples would be air transport or 'www' networks) [5]. While we see this triage as fundamental, this includes no completeness claim. We will show that each of these cases imprints distinct structures in system space (with some properties even shared) and that this translates into corresponding properties of the feature space. We will then exhibit a few cases where biology implements these principles and demonstrate how our approach leads to a fresh view on neuronal coding and signaling. As a final demonstration of the approach's practical usefulness, we will discuss a derived clustering algorithm that is superior to the standard clustering tools used in big data.

2. Families of isolated nonlinear systems

Biological systems often exhibit strongly nonlinear building plans. A notion of connectivity among such systems is then implied by the similarity of their building plans. This situation may be interpreted as nodes having a continuity of variable strength of connections. In feature space, close proximity is then reflected by qualitative similarity, e.g., by equal periodicity or stability properties. For this case, our paradigm will be most thoroughly worked out.

On the parameter spaces underlying the definition of a nonlinear system, an infinity of system constructions are possible. Physics of the last century has however revealed that nonlinearity introduces order in the complexity that such constructions offer, by means of universality expressed in scaling behavior. We focus on the simplest generic nonlinear systems and ask the following question: where in the parameter space are the systems that share similar behavior and what universal scaling properties do they exhibit? In answering this question, we are naturally led to so-called shrimp-shaped parameter space domains [4] to which systems exhibiting identical periodicity are confined. The geometric form of these shrimp-shaped areas is surprisingly intricate, but all of them have a common building principle and follow scaling rules that manifest in feature space. For a single parameter, the set leading to periodic behavior is an interval. From this, a cartesian product of such intervals could be expected in higher dimensional parameter spaces (i.e., a square or a circle in dimension two). This conclusion underlies the main computational approaches of bioinformatics (e.g., *k*-means clustering), but is inappropriate. The shrimp-shaped periodicity domains underlying real-world systems have no affinity with the expected Gaussian cloud (cf. Fig. 2), but instead generalize Feigenbaum intervals to two parameters. Their existence and convex-concave form was already predicted by Shilnikov [8–11] and discussed in more details by Gaspard, Kapral and Nicolis [12]. Only recently, they were corroborated in real-world dynamical systems (electronic circuits [4,13,14], laser systems [15], biochemical systems [6,16–18] and in models of biological neurons [6]).

The multitude of scaled versions of the shrimp template reflect the simple building principle of the generating process. Shrimps express the interaction of two or more largely independent parameters in creating points that have a full set of zero partial derivatives, to ensure stable periodic behavior. From this observation, the shrimps phenomenon can be explained in a simple way for flows and maps [12]. For simplicity of argument we consider the discrete formulation [4,19]. The dissipative Hénon map [20] is the paradigm for two-dimensional dissipative nonlinear maps;