A guideline for a public-private partnership on urban big data sharing

Daniel Sarasa Funes
Polytechnic University of Madrid (UPM)
daniel@openyourcity.com

ABSTRACT

This paper explores the construction of public-private partnerships on the subject of sharing big data in cities. It considers data as an strategic asset whose exploitation benefits are not sufficiently permeating our cities neither in the form of better local jobs, new scientific knowledge or well-informed urban operations. It analyzes the barriers and inhibitors of such a sharing agreement between key urban players, especially privacy concerns and cooperation dynamics, and goes over the potential advantages that mixing big data sources in the urban context could have.

The work presents and develops a set of implementation principles for the system, including agents and roles, building blocks, governance and the new kind of professional profiles needed, building a case for a stronger public action in the field of big (urban) data. We argue in favor of the City Hall as the best positioned institution to take the leadership flag in this endeavor, which can be pursued by combining a soft regulation strategy with the activation of other facilitation tools. We also highlight the contour of data-driven government alongside possible success metrics.

Finally, we list the conclusions of our work as a set of guidelines for those cities interested in progressing in this idea, signaling as future lines of research both the study of optimum locations for its practical implementation and the detailed study of the business model and design of a viable prototype.

1. INTRODUCTION

We live in the age of cities and big data. Privatization, out-contracting and the booming Internet have resulted in distributed 'de facto' city operations and management schemes. Today, digital businesses like Google, Amazon or telcos, the so-called 'shared economy' companies or traditional businesses like banks or utilities (through their new smart metering units), sense and know partially how the city operates, although its opacity prevent city halls to use that knowledge for a better urban operation.

1.1 Knowledge cities vs. 'data-driven' cities

Urban operation, as a result of the multiplicity of agents involved and overlapping layers, and due to urban growth, is an increasingly complex task. But, at the same time, with city activity operating twenty four hours a day its importance can not be sufficiently highlighted. The conclusion that, as more and more people move to cities, there is no better way to improve people's life than improving life in cities, is pretty straightforward. As urban concentration grows, the multiple risks that our planet and societies face at the environmental, social or economic spheres can be better addressed by adopting more sustainable, more innovative and better-informed urban policies.

Nowadays, those policies can not ignore the potential offered by the set of processes and technologies grouped under the generic 'big data' buzz-word. Big data, alongside other emerging technologies like, smartphones, Internet of Things (IoT) devices and machine (and deep) learning may have a substantial impact in how cities are understood, planned and managed. IoT, for example, is expanding exponentially the amount and diversity of the data collected. Although its acronym relates to *things*, IoT devices and networks increasingly harvest information about us, people, through a wide variety of sensors that track (with or without our consent) our daily actions. Smartphones and apps complement IoT devices by further extending the personal data that we release, including opinions, habits, relationships and health. Big data glues all that unstructured information together extracting the relevant traits of individuals, either considered as customers, voters, or fully recognized citizens. Finally, machine and deep learning automate the consumption of that information and many of the subsequent actions or decisions that happen in the organizations that hold and/or analyze the data.

The novelty is that these emerging technologies may join to provide the deepest level of comprehension so far about the physical and human systems and subsystems that form the city, that this can be achieved in real time and that it may be possible to better forecast its short and long term evolution. The nature of the difficulties that stand in the way of such a decisive jump in urban practice are more organizational than technical and require soft-skilled, multi-threaded and highly creative professionals rather than pure technological capacities. Professionals capable of grasping, for example, the subtle implications that arise when we blindly add machine and deep learning to big data and we apply the resulting combination to automated decision making processes in the urban context. By doing so, we can advance significantly towards an old futuristic idea that has been lately renewed: the data-driven government, although we might as well question if we can afford the price of losing

equity and human understanding of human needs along the way.

As stated before, cities are probably our most precious tool to tackle and eventually fix the main problems of mankind: environment, democracy, economy or decent life conditions. But, paradoxically, and at the same time, they showcase the crudest representations of those very same problems: pollution, corruption, unemployment, inequalities, poverty, isolation, etc. Hence, the question that quickly arises when deepening in the relationships between big data and cities is up to what extent big data contributes to solve those crucial challenges without creating new problems. Or, formulated differently, if a city would improve at all if we were able to scientifically test *all* decisions, and by how much.

1.2 The inflationary phase of big data

If we take an historical angle, after the 'big bang' explosion of data, the new 'data-verse' is experiencing a quick inflationary phase. This inflationary expansion, as shown later, is everything but homogeneous through organizations, sectors or processes.

In terms of organizations and sectors, the Internet businesses are clearly leading the way, fueled, first, by the inherent use of big data related technologies, some of which are even powering the development of the general concept of big data itself, and, second, by the highly competitive and innovative markets in which those companies operate. Obviously, a business logic works here, and the use of big data that those Internet giants make is driven by their needs to gain a competitive advantage. To illustrate this, let us take the case of a company like Google, which bases its whole business operation on big data, investing a huge quantity of resources in the development of the technology. Just as an example, at the base of big data storage and retrieving techniques, we find MapReduce, which is a Google patent [1]. In addition, known Google's businesses are based on data that users release, either openly (as in Google's search engine or Google Maps) or privately (as in Gmail).

There is another interesting set of businesses, grouped under the label of the 'shared economy' which is worth analyzing, since some of these new Internet companies are disrupting local economies in areas such as transportation or accommodation while somehow managing to surf or bypass local regulations. Über is a visible example of a new paradox. The serious blow that it afflicts to the community of local cab drivers not only affects self-employment in the city but also leads to a reduction in the overall local tax collection. To rub a little salt into the wound, their systems may very well use, for instance, the data about road outages that the city hall releases in open data formats for routing optimization purposes, while the company locks the vast amount of valuable information gathered through their daily trips around the city.

1.3 The 'data sharing value gap'

Über, like many other companies, are free to use government data available from the numerous open data initiatives that many public administrations, at local, regional, national or continental levels are implementing, being such reuse one of the goals of those open data policies, under the assumption that open data has the potential to create new business opportunities. The European Union [2] estimates that opening government data could add up to 40 billion euros to the European economy. Given that the Europe's GDP is 24% of the world's GDP, we could estimate the impact worldwide of opening government data at ~200 billion euros. Although this is just a rough estimation, it is perfectly valid for our purpose of pointing out the distance between the value that can be unlocked by considering opening just public data (let's call this *public data value*) and what can be obtained by unlocking the potential of collaborative data policies between the private and the public sector.

A report from McKinsey Global Institute [3] estimates that a joint policy between private and public sectors to cooperate on opening their respective data silos could unlock between $3 Trillion and $5 Trillion a year in additional value across just seven business domains. Let's call this figure *public-private data value*. Now, even if those figures are rough estimations, we see that the difference between *public data value* and *public-private data value* is in the range of several trillion dollars.

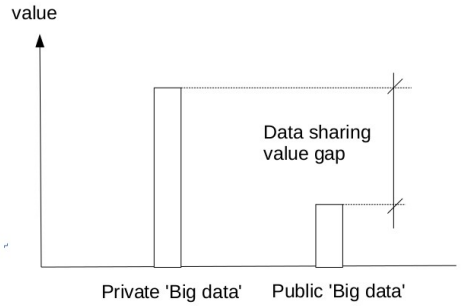

Figure 1. Data sharing value gap.

Narrowing what is shown in figure 1 as the *data sharing value gap* is therefore the main purpose of our work, since, as shown earlier, there is a clear potential to increase the economic and social dividends of big data by including the private sector at the core of the open data policies. This will imply not just considering companies as data consumers or providers of data analytics capabilities. Private companies need to be recognized as the main data producers, with (if not equal) comparable rights and responsibilities regarding open data than those of the public sector. In the absence of city, state, or nation-wide regulations about data sharing

schemes, this paper explores an organizational and conversational approach to a mutual relationship between key city players (public and private) that allows sharing 'at some extent' the knowledge that big data brings in ways that favor public interests while protecting individual privacy and legitimate business assets.

1.4 Data as a public good

We will deal in subsequent sections in more detail with individual privacy; let us just make now the general consideration that, in a digital world, privacy does not exist any more for the vast majority of us. Even in the absence of malicious or accidental information disclosure, at least a dozen of big corporations, and therefore hundreds of people working on them, have access to the records that precisely form our daily lives. Aggregated over millions of other users and thoroughly analyzed, that information constitutes a secret goldmine.

The term *data mining* was originally coined some decades ago to illustrate the nature of the processes that deal with retrieving relevant information out of large data bases. In the age of big data, and taking into account the analysis and projections about the value that unlocking the full potential of data can bring to the world economy, the term *data mining* acquires a new relevance, conveying a clue about the strategic place that data holds for our economies. Considering data as a strategic asset leads naturally to develop a normative framework that would recognize data as an essential public good whose exploitation rights and mechanisms have to be revised.

2. URBAN BIG DATA

2.1 Big data as a relative concept

There are several definitions of big data [4]. The most common big data definition encompasses three main features: volume, velocity and variety (the 3V rule). Lately, it has been added a 'fourth V' (value), highlighting the importance that industry concedes to big data. It would be pointless to give absolute magnitudes for those characteristics, since innovation speed these days would out-date any value by the time this work is published.

A logistic operator with terabytes of supply chain records may claim to be dealing with big data. But then, how about the trillions of stock exchange operations performed in real time by an army of autonomous algorithms operating globally? From an organizational perspective, which is the one this paper addresses, absolute values mean very little. Therefore, a complementary definition depicts big data as 'datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.' For the purpose of our research, centered in the potential, barriers and challenges that big data implies for the main urban players, we will extend that technical view to the organizational dimension, and provide an alternative definition of big data as

'the data whose volume, velocity and variety establishes new challenges for an organization or business, by the opening of new prospects but also by requiring new efforts and skills for its treatment, setting it out of its comfort zone at many levels.'

This definition deals with the original three V's (volume, velocity and variety) of big data as relative magnitudes linked to the level of maturity that organizations present in the field of data analytics, incorporates the fourth V (value) through the opening of new business opportunities, and expands the requirement to innovate that big data brings to the whole organization. As a result of this definition, understanding the level of progress, expectations and needs of every stakeholder pertaining to our data-sharing system with regards to the use of big data will necessarily be a key aspect of study when building a practical implementation of such a system.

2.2 Key processes of urban big data

The technical literature about big data is broad and deep. Let us just picture a simplified building block diagram of the processes involved in big data treatment, as depicted in [5]:

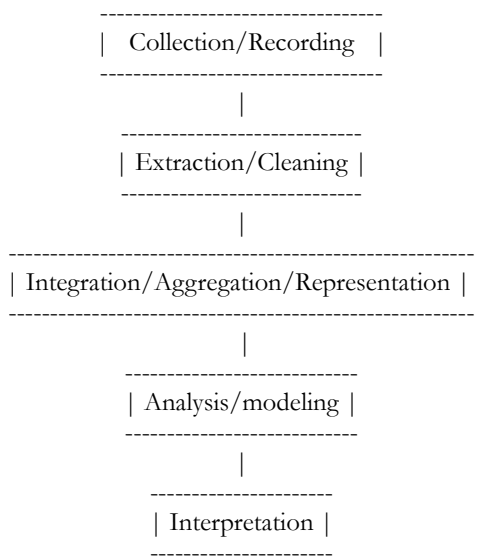

Figure 2. Canonical processes of Big Data

We will not discuss the convenience of this standard model for intra-organization purposes, where big data is produced, collected, stored and interpreted within the boundaries of an organization. However, we will argue in this research that if falls short to grasp the subtle and specific demands of the urban millieu. Hyper-urbanization, ultra-technification and skyrocketing innovation have joint forces to keep

permanently pushing the three V's of big data (volume, velocity and variety) to new limits. As a consequence, the urban 'data-verse', already immense, keeps expanding at increasing speed. Unfortunately, the resources available for cities to deal with those larger V's do not follow the same pattern. Thus leaner, smarter strategies are needed.

In the following chapters, we will examine closely the relationship between big data and cities, adding some key processes to the previously depicted standard big data architecture:

- the process of questioning
- the process of learning (and the related process of feedback)
- the set of processes related to governance
- the (slow) process of expanding the sense of citizenship

2.3 What big data could add to cities

Cities are one of the most complex ecosystems in nature, one of the closest to us, humans, and one of the least understood. The different urban disciplines (architecture, urban planning, social sciences, traffic engineering, telecommunications, urban economics...) have been traditionally devoted to study the city as a collection of either physical objects or human livings.

It was Jane Jacobs [6] who first pointed out the misalignments derived from the 'narrowness' of these approach, providing a broader understanding of the relationships that govern the mutual feedback between humans and objects in cities. A public space entomologist like Jan Gehl [7] followed and performed sound observations about interactions between people and 'physicalities in cities'. Gehl's empirical discovery 'first life, then spaces, then buildings' anticipates the thought that places are a result of interactions, and not the opposite. Saskia Sassen has brilliantly re-situated the importance of cities and places in the global economy [8] and as links of a 'global value chain'. In parallel and very closely related, Manuel Castells' [9] introduced the concept of flows as a governing phenomena to study thoroughly for a better understanding of cities. His influential socio-economic perspective of the city-verse as a 'space of flows' is at the basis of the new science of cities that a geographer like Michael Batty [10] is trying to build.

We sustain that this new comprehensive science of cities is just beginning to be constructed, in parallel to the growing perception that the solution to most acute problems of our age are to be sought in the urban context. As is shown by Anthony Townsend in [11], new urban studies are launched globally, applying interdisciplinary research to the question of cities.

Scientific progress in the question of cities is coming from unexpected directions. Mathematics, biology, and astrophysics promise to bring new theoretical tools to advance in understanding how cities work. Social networks, Internet of Things, and big data are sending much of the information about flows between humans and between humans and objects at local and distant scales. But, in screening the universe of cities, our observation artifacts are maybe too narrow and rudimentary. By looking at individual data sources separately, we are clearly trying to unveil urban mysteries through narrow lenses. We are still on the pre-Hubble phase.

A deep understanding of urban flows, and thus, a more significant progress in the construction of the new science of cities needs systems that can cross-examine the vast amount of information produced in cities through the broad lenses of interdisciplinary and collaborative work. This system needs to work in a close relationship with the city itself. Fog, clouds and atmospheric pollution do not allow an optimal observation of our universe. Stars are better observed from space. Cities are better observed from cities.

2.4 What cities add to big data

The Center for Advanced Spatial Analysis (CASA) at the London's University College (UCL) does not deal with astrophysics but with city sciences. Their work with urban data leans heavily on spatial representation of flows and interactions. Projects like 'Pulse of the City' maps the uses of the Oyster Card against the layout of the tube network and depicts the load of lines and stations over time. Geography helps humans draw a mind map of the territory we inhabit. If the territory changes rapidly, as cities do, then the time dimension needs to be added to the bi or tri-dimensional geographical representations of the urban space. Combined with powerful visualization, four dimensional geography is essential to understand the meaning of big data.

The second contribution which is worth highlighting from a more close connection between cities and big data is 'feedback'. BBVA Data Analytics is a company formed by a staff of data scientists that works both for the BBVA bank and for external customers. Amongst the work they do for cities, the 'event analysis' tool is especially inspiring. By analyzing card transactions during a certain period of time and by comparing the analysis with another period of normality, the company can infer the impact of a given event or abnormal situation in the retail sector of a given area. The tool opens a vast window for experimentation. A transportation authority could test the impact that doubling the size of light-rail wagons have on the sales of downtown shops, in real time, with real data, at eventually no cost.

The ability to conduct quick experiments at very low cost

would allow to validate hypothesis about changes in urban policies in areas like transportation, telecommunications, social sciences, infrastructures or tourism, in the way lean start-up thinking proposes. Lean start-up thinking, combined with agile development, are already key factors for innovation in the business sector. A closer relationship between big data and cities would incorporate them to the urban practice. Learning processes for the companies and institutions operating at city level would be more sound, and quicker, pushing the idea of the city as an innovation platform into which third parties can connect and run their experiments.

Another contribution by cities to the general big data phenomena has to do with the dissemination of knowledge. As explained in the introduction, currently big data provides precious insight to internal business processes to a broad range of organizations, but only to a very limited set of individuals inside those organizations. What big data tells about us is hidden to the general public, even its most coarse traits. Big data is still mostly dark data. City halls, through their cultural and civic institutions, can help in illuminating the big data-verse, connecting the new knowledge with people.

However, we should not expect that the general public would consume large, complicated reports. The process of extracting the fundamental traits from big data and to show them in a meaningful and simple way will require complementary skills to those of the data scientist. Skills that belong to the arts domain. As we will explain later, artists may play a central role in our proposal, not just by providing a deeper or more emotional insight for the general public about what big data reveals, but also by finding new, maybe lateral, paths to overcome some of the problems that may arise. From Claude Cézanne's intuitions about how the visual cortex works to Marcel Proust's and Virginia Woolf's accurate description of neuroscience and mental illnesses, artists have a proven record in anticipating some of the fundamental questions of science [12]. In our days, Collide@CERN and Arts@CERN are two joint initiatives between Ars Electronica Center in Linz and CERN aimed at provoking and exploring 'creative connections between the worlds of science, the arts and technology'. During its four years it has proven successful in reaching trans-disciplinary artistic excellence and fruitful exchanges between artists and scientists.

3. BIG DATA CHALLENGES IN CITIES

A collaborative white paper [5] published in 2012 described some of the challenges that were impeding progress on the big data field at the time, stating that heterogeneity, scale, timeliness, complexity and privacy were inhibiting the realization of much of the big data potential.

3.1 Privacy

All of the aforementioned inhibitors, except privacy, are somehow linked to the limits and costs of technology. However, privacy, due to its civic nature, is a more complex issue in which technology does not play a solving role, at least in a straightforward manner. On the contrary, technology, driven primarily by the (still) undefeated Moore's Law [13], poses a direct threat on privacy, as the capacity to store our personal information and the power to compute our present and future behavior keep doubling every 18-24 months.

Using the current processing and storage power, traffic cameras and sensors could potentially track every car in the city, banks could analyze all our financial footprints, and Google would be able to read all our emails (which effectively does).

Privacy concerns become more serious in the scenario in which several data flows (produced from formally independent data sources) about a user are intermingled. Although a worthless resource for researchers, mixing a patient's medical records with his or her supermarket transactions could affect the individual rights to health care or insurances.

The previous scenario shows clearly the power of cross-source analysis of data, both on the business and research sides. Luckily or not, regulations impose severe constraints on data sharing. In Europe [14], treatment of personal data is mainly affected by two principles: 1) no data that can potentially identify an individual should be disclosed to third parties, and 2) the collected data should be used for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. This imposes a severe restriction on the type of analysis that can be performed over our data, fire-walling, in theory, company users' databases from external algorithmic power.

3.1.1 Anonymation

Releasing anonymous individual data for understanding urban patterns has revealed as a useful technique to bypass the privacy barriers. However, several researches have drawn the limits of this technique.

In the field of card transactions, a recent work by M.I.T. Medialab researchers [15] shows that '4 spatio-temporal points are enough to identify 90% of individuals' or, alternatively, that given just the time stamp and location of 4 payment records from the same user, the probability of identifying her or him with certainty is 0.9. The same research shows that the probability of uniquely identifying individuals increases by 22% if the price information is added to the record, which is readily available in most cases.

In the area of urban mobility, another research [16] shows that urban mobility patterns of individuals are highly repetitive and that, 'despite the diversity of their travel history, humans follow simple reproducible patterns.' Although this predictability may bring new insights into solving urban issues such as emergency response or epidemic propagation, it also means that, once the trajectory of an individual is known on a certain weekday, the probability of predicting accurately the position of that individual at a given time in the future is known and high.

Other 'natural' techniques to preserve privacy of the users whose information is queried from statistical databases are the removal of identifiable attributes (to avoid identification of individuals by confronting the records with information known from public sources) and sub-sampling. Nevertheless, as shown before, with the increasing availability of public personal information, especially through social networks, it is feasible for advanced programmers with enough skills and time to complete the missing information in user records and compromise one's identity [17].

3.1.2 Aggregation and differential privacy

An alternative approach to tackle the privacy issue is to make only available group information under the obvious assumption that, the larger the group, the better privacy protection. The fact that such a system only responds to aggregate queries does not eliminate privacy breaches as shown in [18]. In the same article, Dwork brings mathematical rigor to the problem of privacy-preserving analysis of data and proposes the concept of differential privacy as the increase in his or her overall risk that a user suffers when his or her personal information is included into a database. Even if the database only responds to aggregate queries, a malicious attacker can infer personal information. To control the risk, noise is added to the database responses at the cost of losing accuracy, in a way that the responses of the database when a certain user record *is* into the database and when it *is not*, are indistinguishable in practice. By adequately selecting the noise function, the system designer can get the appropriate trade-off between privacy breaches and the utility of information.

The use of aggregated data and the subsequent loss in the utility of information is acceptable in many urban contexts. Consider, for example, the information about energy efficiency in dwellings provided by the City of Chicago, which can be found at http://energymap.cityofchicago.org. It provides the energy use per sq feet for both natural gas and electricity at the district and block levels, providing to external parties valuable information to help in Chicago's retrofitting strategy.

3.2 User consent for data sharing

Since the ultimate goal of this paper is to design a feasible big data sharing architecture amongst multiple organizations operating at the city level, it is worth noting at this point that, as shown above, whether that system chooses anonymation or aggregation techniques to preserve privacy of user records, the risk of privacy breach can not be zeroed, even in the absence of information leakages or malicious attacks. So it seems that this is as far as we can get following a purely technical path.

Fortunately, cities are about people, and social and political sciences play a predominant role in them. What if the user would voluntarily give away a portion of his or her privacy? What type of incentives would move him or her to do so? What are the 'consented' risks? And, what can public institutions do to mitigate them?

A smart workaround to the technical obstacle of privacy breaches would consist in facing the technological and mathematical constraints that publication of data pose to privacy up-front and obtain the user's complicity to share his or her personal data. This would imply an explicit acknowledgement prior to the data collection, and would certainly reduce the quantity of records published but, in turn, would shield our system against painful legal problems.

Internet of Things (IoT) company Nest (a top manufacturer of domestic smart metering units) has recently reported [19] that it would share user data with Google. Users would be informed and opt-in to keep control of privacy, and would be allowed easily to opt-out at any moment by unlinking their Nest devices from their Google accounts. The incentive given to the user in this case is that Google Now, Google's personal assistant (a machine learning based service), would provide the 'extra' service of allowing an easier and more integrated control of Nest thermostat. Similar examples can be found in all sorts of digital businesses, from on-line travel planners to social networks. In short, and despite regulations about personal data protection, on-line and off-line privacy is being privatized with our consent.

This volunteer data disclosure, although an exercise of free will and undoubtedly beneficial for business, has received a well-deserved criticism from those who alert against the dystopian effect that such a transfer of privacy to distant dominant market players may have in our civic health [20].

Other experiences in the area of the so-called 'smart cities' suggest that users could have also incentives related with cooperation to share user related data. The project Smartcitizen.me (https://smartcitizen.me) was developed in Barcelona and allows the user, via a simple Arduino-based electronic board, to monitor certain environmental conditions such as pollution, luminosity, humidity, etc, along with location data. More than 1.000 devices have been

shipped around the world, reporting more than 5.000 sensing signals and populating a database of more than 52 Million of records. People adhered to the project make their information public for the sake of cooperation.

Environmental data from smart sensors convey little personal information, so it can be argued that its publication matters little in terms of privacy. But humans can give away even its most sensitive information such as health records for the sake of contributing to a collaborative endeavor like the progress of medical science. A health research institution as the Sanger Institute [21] feeds with medical data freely provided by patients. As other members of the "European Data in Health Research Alliance", they pledge for a less restrictive EU regulation in terms of data sharing, something beyond the current obligation to re-consent on a case by case basis which, driven mainly by the desire of the European parliament to protect individuals against privacy abuses from Internet companies, would prevent the digital records of death patients to be used for science. After all, if we donate organs beyond life, why not donate data?

A data-sharing system, like the one sketched in this paper, would seek to contribute decisively to an area that is key for our future: urban sciences. Appropriate user-consent alongside with a combination of data anonymation and aggregation techniques would constitute solid foundations to build a crowd-source generator of knowledge about our cities.

3.3 Urban data silos

We have shown so far that anonymation, aggregation and user consent constitute viable tools from a technical, legal and social point of view to be used in the construction of an urban data-sharing system. Either for being better serviced or for common interest purposes, individuals may feel compelled to suffer a certain privacy loss and contribute to such a system with their personal data. But, are organizations ready to share?

The widespread out-contracting process that urban services have experimented in the last decades has created, over time, silos of urban data. Transport, lightning, cleaning, waste, water, energy, telecommunications... are former public services now mostly privatized.

3.3.1 Intra-organization cooperation

The 'silo effect' is a well documented impediment for transversal and integrated actions within an organization. Organizations tend to develop vertically around the various areas of the business, which reflects directly in the way information is stored and treated. Clearly, non-cooperative behaviors thicken the walls of the information silos, while cooperation can bring some of the walls down. Cooperation between individuals within groups is highly influenced by a strong phenomena from game theory called 'the prisoner's dilemma'. A simple formulation of this principle states that individuals find sound reasons to act selfishly in situations where cooperation would be mutually beneficial but some kind of extortion or defection can also be allowed. To put it into mathematical terms, given two isolated individuals facing this kind of choice, let:

P be the payoff for each upon mutual defection,
R be the payoff for each upon mutual cooperation

In case of one cooperating and the other defecting:
S be the payoff for the cooperator,
T be the payoff for the defector

It holds that:

$$T > R > P > S$$

which explains the natural tendency for individualistic behavior in the absence of external inputs.

Fortunately, the real world differs from this simplistic scenario. Individuals do rarely live in isolated pairs and there exists multiple external inputs that affect us. We live in a networked society and have dependency links with one another. The impact of dependency links in the outcome of the prisoner's dilemma has been studied by [22], resulting in an interesting proposition. From the four possible network topologies studied: 1) ring, 2) random, 3) scale-free, and 4) square lattice, it is only the introduction of dependency relationships in square lattice networks that ultimately promotes cooperation.

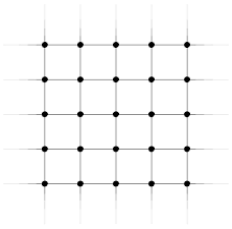

Figure 2. Square lattice network

As opposed to other topologies, square lattice networks are quite horizontal, equally balanced and do not present a hierarchy in its dependency relationship. Let us retain these important characteristics for the later formulation of the design of our data-sharing system, where the nodes in the lattice will be represented by each of its stakeholders.

In terms of external influences, recent research suggests that there is a thin line between cooperation and individualism in populations, as shown in the general framework developed in [23] by Stewart and Plotkin. Although there are powerful social, organizational, cultural and political influences that strengthen cooperation, the process can be more easily reversed than previously thought.

In evolving populations, the framework shows that very small variations in the variables of the system may cause cooperation to collapse. Without entering into the intricate mathematical depths of this new insight, let's keep in mind that latest developments in game theory support the idea that a very soft, delicate governance of the system will be required if we intend to keep it on the track of mutual cooperation dynamics both at the individual and intra-organizational (departmental) level.

3.3.2 Inter-organization sharing

Although the characteristics of horizontality and 'softness' in its governance principles introduced so far also applies at the stakeholder level, the general case of cooperation between organizations requires to address the 'business side' of things as well. To start with, we can consider two types of transactions between stakeholders pertaining to our system: information and value.

Information, either raw, anonymized, aggregated, or processed data, is what stakeholders contribute to the system. In turn, they receive some sort of value. Having the previous chapters dealt with how this information may look like, it is time now to take a look at 'the flip side of the coin': value.

Value is the expected outcome of all the stakeholders participating in the urban data sharing system which is the subject of this research. However, value can take different forms and have different attributes depending on the nature and motivations of each stakeholder which, in our system, might be any relevant urban player contributing to the city operations activity. For simplicity, we will specifically consider a limited set of contributing partners or stakeholders: the City Hall, the scientific research ecosystem (represented by research institutions linked to the local universities), the main telecommunication operators, those banks owning a significant share of local financial transactions, the (quasi monopolistic) utility companies and the less structured communities of entrepreneurs, data journalists and digital artists. By thinking over how value can be perceived, symbolized and exchanged between this set of actors we will advance on the task of depicting our urban data sharing architecture.

The most obvious representation of value is money. Money symbolizes the abstract concept of value and is commonly accepted as a ways of acknowledging the extent to which a service or goods provided by a certain supplier is valuable to a given customer. In the field of urban data, some incipient customer-provider relationships have already been established between organizations of the nature mentioned above.

The M.I.T.'s Senseable City Lab has carried out significant work in helping cities to comprehend urban patterns through the analysis of data. The worldwide recognized scientific and visualization capabilities of the Lab service cities and big corporations with knowledge and branding. These clients correspond by funding many of the projects in the Lab which, in turn, fuels future research.

This transaction scheme, knowledge in exchange of money and data, works equally in the case of other big data generators like banks, telecommunication companies or utilities. Lacking in-house advanced data analytics skills, they often lean on external data analytics companies to get insight over their own business operations. The conclusion is that the knowledge gained through big data is perceived as valuable, and thus organizations are willing to pay for it. Let k_n denote the knowledge gained by a stakeholder n through the analysis of a certain dataset d_n whose origin is the operational activities of that given stakeholder. Let $f(d)$ be the function that represent the analysis and interpretation of those data. Then:

$$k_n = f(d_n)$$

In a real scenario, k_n is the knowledge supplied by the research lab or data analytics company on a given project, and d_n is typically supplied (and extracted) by the customer.

In our proposed system (formed by several stakeholders providing separate datasets) we have an aggregate set of datasets D, such that

$$D = \Sigma_n \, d_n'$$

In such a system, we need to find a new set of functions f'_n, along with possible arrangements between N stakeholders that provide such k_n' that:

$$k_n' = f'_n(D)$$

where it holds that $k_n' > k_n$ for every stakeholder n. This implies several upgrades from the original customer-provider binomial system. In some cases, new datasets or information d_n' will have to be at the systems disposal. But what is central is to find new functions f'_n applied over the new set of datasets D such that the increase of knowledge received to every stakeholder allows the system to work without any monetary exchange. As money is zeroed, there is no such roles of customers and providers between stakeholders. All of them are partners bound by a mutual cooperation interest.

3.4 The quest for relevant questions

We have shown that in a shared environment of mutual cooperation between urban stakeholders, new datasets d_n' have to be extracted, summed up into an aggregated set of datasets D, and that new 'knowledge functions' f'_n need to be unveiled. Let us focus at the quest for new 'knowledge

functions' and try to imagine how such a process can take place.

It is widely accepted that the new knowledge from data is likely to originate from the highly creative communities of digital entrepreneurs, and so one could think that the celebration of myriads of hackathons around the world would cast light on the potential of big data. However, the outcomes of this wide movement in terms of new discoveries has been, when compared to what the digital economy brings in terms of services and wealth creation, modest. Conceived both as a branding mechanism and as a way to promote the politically healthy open data movement, hackathons witness how cities and institutions throw over and over again the same type of data to the communities of geeks. Not surprisingly, these produce local flavors of, essentially, the same type of apps everywhere.

Nigel Jacob leads a guerrilla group inside Boston's Mayor's Office called New Urban Mechanics that is using third party data from the routing app Waze.com, which claims to possess one of the largest community of users. The data-sharing agreement between the City of Boston and Waze.com includes joining city and app data to assess on the planning and execution of civil works, or to discourage double parking. The on-line community of data scientists Innocentive.com also cooperates in the scheme, accepting challenges from the city and providing advance and distributed knowledge skills.

This represents a new approach in the quest for relevant questions. Data-driven local hackathons could be combined with challenge-driven processes mixing local and distributed communities. The innovative example of Boston could be further enhanced through a wider participation. In challenge-driven hackathons challenges do not necessarily need to come from the city officials. Local civic communities have much to say when detecting the most acute local challenges.

In this reverse process, the new 'knowledge functions' trigger the emergence of hidden datasets. Formulated in an intuitive manner, the process of generating knowledge has to be triggered by the act of questioning.

3.5 Stabilizing cooperation

We start to have a glimpse of how a urban data sharing system may start operations: around a table, the big urban players acting mainly as data providers and knowledge consumers, and civic entrepreneurs and researchers transforming data into social and economic value. (The notion of stakeholders sitting around a table represents the flat, non-hierarchical governance scheme.) Periodical challenge-driven hackathons combined with the crowd-sourced power of an on-line community of data scientists are fueling the system with questions and with analytical skills, constructing an initial civic base.

Now that an initial version of our system is running, we might ask ourselves how to ensure that cooperation dynamics continues over time, making the system evolve and enlarge its stakeholder base. At this stage, we will pursue the path of game's theory and examine the concept of the 'Nash equilibrium'.

The 'Nash equilibrium' can be defined as the situation in which none of the players in a game has anything to gain by changing their strategy. To simplify, the set of strategies in our data-sharing system is:

{ Cooperation (sharing), no cooperation (not sharing)}

Given that the 'Nash equilibrium' is a stable situation, our problem of stabilizing cooperation can be formulated in terms of game's theory as finding the appropriate set of incentives and/or regulations so that our data sharing system finds its situation of 'Nash equilibrium' when the strategies of all stakeholders (players) are set to 'Cooperation (sharing)'.

As explained above, this means that any change to a 'No cooperation' strategy would imply a loss for any player pertaining to our system, hence the stability of the cooperation dynamics.

Now, the reader may note that the concept of *regulation* has appeared for the first time in this work. It will be treated in more detail in section 5.

4. IMPLEMENTATION PRINCIPLES OF AN URBAN BIG DATA SHARING SYSTEM

4.1 The 'platform approach'

Amongst the projects in which analysis and visualization of urban data is allowing a deeper understanding of how cities function, the 'Live Singapore' project is one of a kind (http://senseable.mit.edu/livesingapore). Within a few years time-frame, it has evolved from a (rather advanced) tool to visualize the city's changing geography (both in space and time) through the tracking of taxis, into a platform fed by multi-stream data sources that allows third party connections through an API (Application Programmer Interface). The project, framed within the broad Singapore - M.I.T. alliance, is on track to effectively build an 'urban real time data platform' fed by various data sources, combined and analyzed by M.I.T. data scientists and accessible by external agents.

4.1.1 Physical, digital, human

The concepts of platforms and APIs are important, but in terms of platforms and APIs we are less concerned by the

choice of a particular set of software or hardware in the traditional, rather narrow IT sense, than by the broader approach of putting our assets (physical facilities, systems, data and knowledge) at the service of third parties to innovate upon. In this context, the API is the way in which third parties interact and connect with our system and is not constrained to a piece of software. A 'conversational API', i.e. a way in which the technical staff of our system are readily available to help civic entrepreneurs in the design of new services, would be perfectly valid, and be even more valuable than a perfectly documented piece of code.

Before depicting some characteristics of the type of platform we intend for, a short foreword about city platforms imposes. Some of today's 'smart city' platforms have evolved from SOA (Service Oriented Architectures) systems, originally designed to manage digital businesses, while others come from the industrial sector where they were used to deal accurately with complex fabrication processes. If we look at cities, we will see that they are composed of physical objects, like roads, irrigating valves, or traffic lights, so 'physical' platforms (such as Scada) may be indeed useful. However, in the last decades, a digital layer has appeared. Digital objects like data or apps belong to this new layer, and not surprisingly, platforms that foster this digital side of cities are gaining momentum. But, we should not forget that, most of all, cities are inhabited by humans and that those humans demand nowadays the highest degree of participation in history. Therefore, human, face-to-face, off-line platforms and APIs are required.

Zaragoza's Open Urban Lab [24] held an extremely inspiring workshop on "Improving the mobility of schoolchildren in the city", School representatives, public servants from the Department of Mobility, the municipal technicians that were working on the city-wide program of "Safe routes to schools" and people from the Smart City Department discussed and worked in groups on innovative solutions for promoting a healthier and greener mobility around schools. An improvement for the classical traffic lights in the city was proposed, a traffic light that could extend its duration when a line of schoolchildren was about to cross, provided that it was led by an authorized adult. The leading adult had a special permission on its citizen card that granted her/him the possibility to modify the traffic light times on certain tranches of the day, usually around the hours of the beginning or end of classes.

The presence of the city technicians was key. The technicians of the Smart City Department opened up the possibility of using the citizen card as an authorization and authentication mechanism. The technicians of the Mobility Department explained the consequences of changing a traffic light duration in traffic flows over the city and how this effects propagates, imposing conditions to identifying the best suitable crossings where this system could be implemented at.

The Live Singapore platform has an astoundingly powerful technology and data scientific skills while a relatively low degree of civic engagement. On the opposite side, Zaragoza has a relatively thin open data API providing simple functions for developers, but has a long tradition of citizen participation and one of the thickest networks of civic and community centers. Fostering participation in a city, and shifting people's and institutional behavior might be an overwhelming task that needs time and deep political and cultural changes. Acquiring technology in a city where civic participation is one of the most recognizable traits of the place is far more easy. It just needs funding and resources.

4.1.2 A platform was originally a service

As we explained in the previous section, a platform must allow third parties to connect and interact. We add that a successful platform is a platform that is effectively used by the communities of possible contributors. But the path to a successful platform is not an easy task. The smoothness, level of service and functionality that platforms need are sometimes only achieved if the system is first used as a service. That is how a connectivity service like Internet became one of our most successful innovation platforms, and there are many examples that follow a similar pattern.

This tells us something about phasing the implementation of our system. It has to function primarily to service the founding stakeholders, and service them well. Only then other parties will be keen to connect.

4.1.3 'Open' platforms vs. 'Open source' platforms

Open platforms are an oxymoron, since the contrary, *closed* platforms, can not perform as such. Therefore, the *openness* of a platform adds little to its description. Open source platforms, on the contrary, are far more interesting, especially in the context of civic innovation. They are the natural fit to the crowd-sourced knowledge flow described in chapter 3.

As Anthony Townsend puts it [25] "when you create urban software, make it simple, modular, and open source […]". Some of the most sophisticated urban management platforms in the market promise to behave as the 'brain' of the smart city. Undeniably, emerging technologies such as machine learning and deep learning will keep growing and advancing towards the decision centers of cities, and although we acknowledge the fascination that data-driven government casts on many urban thinkers and technologists, we must be very careful at pushing automatic decision making in the urban ground. Let's not forget, after all, that cities are about people, and that it is far easier to write code for performance optimization tasks than for preserving abstract and complex values such as equity, inclusion or cultural background.

We still believe that human beings must still do the thinking at the brain of the smart city, surely helped by machines. Cities can not be built without the IT industry, but it would not be wise to let the IT industry run cities. Using open source code for our data sharing system has the additional advantage that anyone can audit how it performs and whether it respects ethics and/or civic values. There is still margin to adopt technologies and processes that lead to better-informed decision making in cities, without falling into a futuristic illusion.

Although open source refers normally to software or hardware, its concept can be extended to other assets such as data (open data), places, buildings and processes. Whatever the object is, an open source nature implies accessibility, understandability, reconfiguration and, finally, participation, in the sense that it is the community of users (with different degrees of involvement) who ultimately makes it work. In the following sections that deal with actual implementation issues we will treat the importance of processes and of buildings. Both concepts, related to governance and places, are as important in our system's design as the software or the data.

4.2 Leadership, stakeholders and communities

Urban thinkers like Jaime Lerner [26], Edward Glaeser, [27] Benjamin Barber [28], or the aforementioned Manuel Castells and Saskia Sassen, have stressed the role of cities as 'solution providers' and central nodes in the network of economic, social and political flows. We walk steadily towards a revisited times of the ancient Greece's *polis*. At the institutional level, the main activities of these twenty-first century polis are managed by a complex set of private and public agents (big and small private companies, some of them public contractors, agencies, foundations, public enterprises, research and cultural institutions, universities, etc.) At the center, the City Hall is the agent with the highest degree of responsibilities regarding city operations, and is accordingly recognized by citizens as the closest administration.

Although not traditionally perceived as innovation agents, in the last years we have witnessed an important shift in the mindset of city halls towards innovation: e-administration, open government, start-up incubators, innovation hubs, open data policies, and the general smart city industry trend, are some of the visible elements of these change. A change of mindset that is driven by factors such as global competitiveness, growing citizen aspirations, branding, financial constraints, climate change and technification, among others. It is no surprise anymore to find the words 'City Hall' and 'innovation' in the same sentence. And, in some cases, they are starting to galvanize the innovation ecosystem at the metropolitan scale. They are thus prepared to take the leadership role to spark and engine our data-sharing system, and to contribute with tons of data as any other 'big' urban player.

But, besides City Hall, who might be the other 'founding' stakeholders? To answer this question, let's come back to the subject of flows: social interactions, energy transfers, personal mobility, economic transactions and information exchanges are the main flows that can help to represent how people relate to the city and with each other. An energy company with a deployed network of domestic smart metering units, a bank with an extensive network of ATMs and Point Of Sale terminals, and a (wireless) telecommunication operator can join forces with the City Hall to form the main stakeholder base.

The third type of agent are the *communities*. As all the rest, they service the system and benefit from their participation. The following key communities have been identified so far: researchers, entrepreneurs, children and youngsters, artists and data journalists.

Researchers have the role of consuming data and producing knowledge. Entrepreneurs consume data and knowledge and transform them into new business models and, hopefully, new and better jobs. Children and youngsters are net knowledge consumers, they play and learn and their enthusiasm can act as a magnet for their families. Artists, as explained in section 2.4, work with researchers in the provision of new insights and help to extract and present meaningful information. Data journalists is a small community, but key to explain how the discoveries affect us.

4.3 Building blocks of an urban big data sharing platform

From section 2.2 we recall the set of processes that are key in the overall task of extracting value from urban big data, in the sense this work advocates:

- the process of questioning
- the process of learning (and the related process of feedback)
- the set of processes related to governance
- the (slow) process of expanding the sense of citizenship

4.3.1 Questioning. The *human API*

It is a key process. As shown in section 3.4, it triggers the whole knowledge cycle. Questioning is allowed through a *human API* that listens to civic demands and that helps to organize challenge-driven hackathons, that ensures that those events gather together civic communities, city officials, stakeholders and entrepreneurs, that mediate in the process of fulfilling the requests of those communities (sometimes through soft negotiation skills) and that, finally, make sure that deals are respected in reasonable time and form.

A special kind of technical staff is needed to fulfill this role of Human API. They will interface between the most dynamic communities and the city officials (mostly public servants) and must be prepared to soften the frictions that naturally may arise. They must cast a collective vision on things and be able to create a climate that favors understanding and empathy. With their attitude and communication skills, they foster a proactive approach to problems and demands. To the outside world, they transmit the values and constraints of the public institution openly. To the inside, they make sure that every demand from the communities is treated with an open mindset. They are conscious that their role is not to accept or deny petitions from the communities, but to transmit them and make sure that the flow of ideas from the civic communities permeate the institution and are used to effectively push the city forward.

The professionals forming this Human API can be seen as mediators, but also as integrators with a holistic vision of the urban field. Discreetly, they fuel the cooperation spirit between citizens, institutions and companies that a collaborative smart city needs to attain its full potential.

4.3.2 Learning. An *observatory* and a *laboratory*

Learning is a distributed process. It can be achieved through observation and through experimentation, which naturally leads to two important building blocks, which we will call *observatory* and *laboratory*. In our system, the observatory will perform the functions of the first three phases of big data treatment as depicted in section 2.2: a) collection and recording, b) extraction and cleaning, and c) integration, aggregation and representation, to which we will add the necessary anonymation processes explained in 3.1.1 to minimize privacy concerns.

And the lab is the city itself. A place to test hypothesis and prototypes. Experiments in the city have to be granted by a municipal authority (after all, the city hall is responsible of maintaining the public space) that, in collaboration with the stakeholders and the civic communities, establishes the time windows, places and detailed procedures to conduct the trials and pilots.

We propose the framework of lean start-up thinking to conceptualize this learning process. In essence, lean start-up thinking is a way to apply the scientific method to the launching and operation of new activities, encouraging confronting hypothesis with experimentation in the most simple and quickest way. Designing adequate experiments is therefore crucial. Given that lean start-up thinking has sped up the innovation cycle in many businesses, and is already part of the culture of the start-up ecosystems, there is no reason to think that it can not be applied in its fundamental principles to the urban millieu. It is a fact that there is a growing difference in speed at which innovation develops between city halls and the new digital big players. When city halls lose the innovation race, public action suffers, it gets weaker, and markets dominate over citizens' aspirations. So city halls need to build new organizational culture and tools that shorten the innovation gap, not only as a way to re-balance power between the public and the private sector, but also between the local and the global spheres.

We recognize that, up to now, only limited initiatives have been conducted in applying lean start-up thinking to urban planning. We forecast that future urban practice will deepen this line of work.

4.3.3 Governance. A 'soft' institutional architecture. The *agora* and the *board*

At this point, it is necessary to make a common definition of what the generic concept of governance encompasses. As found in [29], governance refers to a set of institutions and actors that are drawn from but also beyond government, identifies the blurring of boundaries and responsibilities to tackle social and economic problems, can be exerted by means of autonomous self-governing networks of actors and recognizes the capacity of getting things done which does not rest on the power of government to command or use its authority. Governance has more to do with steering and guiding that with ruling. For this purpose, governments, in this case city authorities, must incorporate new tools.

As shown in section 3.3.1, governance of cooperation agreements, which is our case, must be flat and soft. Flat refers to the topology of the network between stakeholders, while soft means that those relationships between actors lean more on flexible, often non-written rules and on interpersonal skills rather than on fixed norms. The system, however, accepts a minimum level of regulation, as explained in section 3.5, but only as a positive incentive aimed at stabilizing the system into a cooperative track. This aspect of regulation is further treated in chapter 5.

There are two key elements responsible for system's governance. One is the *agora*, a distributed on-line and off-line permanent assembly where stakeholders representatives, researchers, members of the civic, artistic and entrepreneurial communities meet and deliberate about past achievements, current problems and future milestones. The other is the *board*, a lean, flat table where stakeholders representatives sit in equal terms. The *board* addresses mainly the issues related to funding and ensures that the system meets the conditions that make cooperation between stakeholders happen, while the *agora* catches community needs.

As explained before, the activities of the the *agora* take place on-line and off-line, therefore needing a physical settlement,

a place where the members of the different communities, i.e. researchers, civic entrepreneurs, city officials, artists, etc can have a collaborative workplace.

As William J. Mitchell beautifully puts it [30], 'the twenty-first century will steed need agoras -maybe more than ever.' Agoras operating both at the local and the global scales, and where the configuration and character of the public space will determine their success at performing their functions, regardless if those places are 'virtual, physical, or some new and complex combination of the two'. What is essential, Mitchell states, is that they allow for both 'freedom of access and freedom of expression'.

In the last decade, a new set of public facilities have appeared in our cities: they are public civic-innovation hubs promoted and managed by city halls, which confers them the character of public space. Places where the communities of innovators, geeks, social and tech entrepreneurs meet and discuss physically and on-line. Accessible, understandable and, more often than not, reconfigurable. Open source facilities that can be a perfect fit for our *agora*.

4.3.4 Enhancing citizenship: children, workshops and a *datadome*

The final process outlined in section 2.2 deals with expanding the sense of citizenship. If one of our system's goals is to unveil the mysteries about how cities function, then those findings must be disseminated to the people living in the city. The understanding that better cities mean a better world is driving the interest of the scientific community towards them, but very little of that interest is still permeating downwards the education chain, specially in primary and secondary schools. It is important that children too get to know how cities work. What urban big data tells about cities, it tells about us.

In China, Shanghai's Urban Planning Exhibition Center (www.supec.org) is an example of good practice in visualizing urban planning as a key tool to explain the story of the city and its future. The several stores of the museum take the visitor on a journey along the pass, present and future of one of the world's most striving metropolis. By understanding Shanghai through its urbanism, it accrues the sense of belonging and citizenship.

In the UK, the project of Bristol's data-dome is visionary. To be launched in October 2015 and funded by Bristol's City Council, it is part of the broader 'Bristol is Open' initiative and aims at upgrading the Planetarium to create a city data visualization facility. The new *datadome* is conceived, not only as a data visualization facility for citizens, but also as a platform for businesses and for the digital creators to develop professionally. Through it the city openly recognizes data as a city asset upon which to project Bristol locally and internationally. The data visualization projects at Bristol's data-dome are triggered through *workshops* in which geeks and digital creators meet to organize and work the challenges of representing data in a collaborative way. By creating an outstanding facility Bristol has managed to extend the reach of the communities potentially interested to a national scope. A community of data scientists that, as a mechanism to underpin sustainability, is ready to open its data representation skills to private corporations.

In Linz, Austria, Ars Electronica Center holds the Futurelab. On the program, two initiatives may illustrate how urban planning, data and new visualization techniques can join forces to make citizens aware of the uniqueness of the flows and experiences that cities hold. On one side, the interactive exhibition 'Experience Big Data', a collaborative project with SAP, one of Europe's largest software provider, has succeeded in developing interfaces which work as translators between the digital realm, the architectural space of the SAP Pavilion as well as the visitors. On the other, 'GeoPulse Linz' is a state-of-the-art simulation and visualization tool, through urban data, of the culture, history, migration flows and future of the city of Linz.

4.3.5 Wrapping up the data-sharing system's building blocks

The following picture depicts the system's building blocks and its correspondence with the processes involved in dealing with urban big data, as explained in section 2.2. It is to be noted that the canonical processes common to big data and which are independent of the urban context are embedded under the generic *learning* process and performed by the *observatory*. It is therefore this building block that holds the IT gear, software and databases needed for performing such tasks. The location of those systems (local, remote) or whether they use cloud services for some of its functions is not significant for our architecture purposes.

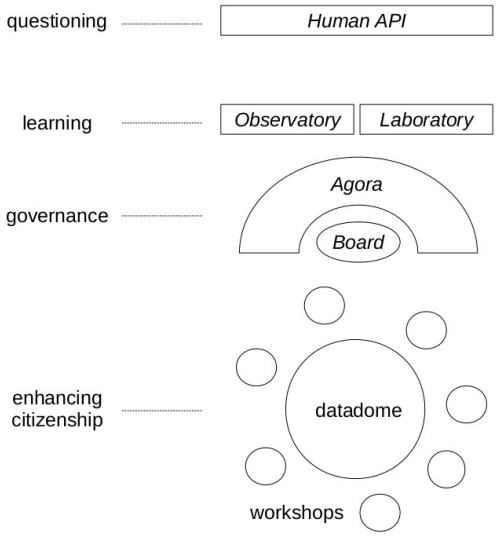

Figure 3. Urban big data processes and system's building blocks

4.4 Profiles

An institutional data sharing system like the one proposed in this work requires a wide range of different profiles. Undeniably, technology (alongside globalization) is driving many of the processes happening nowadays in cities, big data being an example of this. Thus advanced IT skills are required to do the technical job. However, it must be highlighted that cities are extremely complex ecosystems, with multiple dimensions or layers that intertwine: social, economic, technical (infrastructures), cultural, physical, but, most importantly, human, since they are created for and by people. It is the complexity of cities which make of city making an art [31], and it is the human dimension of cities that bring the emotional and political requirements to the table.

It is at the intersection of art, technology and social sciences where new discoveries can be applied to strengthen civic life and where new emotional narratives about urban life can be written. Data journalists, digital artists, political scientists, technological activists, cultural mediators, civic business mentors and holistic city planners are some of the new profiles needed. Little of them correspond with formal university degrees, but are rather shaped through a mix of field work and thirst of transversal vision. They belong to the set of new *freestyle* professions that will push our cities forward.

5. FINAL CONSIDERATIONS ABOUT PUBLIC ACTION AND BIG DATA

In the last decades, we have witnessed how the public sector has retreated from the direct delivery of urban services, in a steady process of privatization and out-contracting. We will not discuss here the appropriateness of such process. Instead, we intend to highlight that, pushed by globalization and technification, new areas of public action appear where markets fail to provide adequate service. There are several situations that might justify public intervention. One is the case of 'externalities', i.e. the costs (negative externalities) or benefits (positive externalities) produced by the economic activity of an agent are supported by third parties without a fair compensation.

5.1 Regulation

In the case of the data produced by human activity, we might face a positive externality without adequate compensation when that data is used solely for private purposes other than the operation of the business for which the data was collected, which is often the case. This is one of the particular reasons that would underpin a public intervention in the field of big data. In the idea of a data-sharing agreement between parties backed by a public authority such as a city hall, the research and entrepreneurial ecosystems would be the mechanism to compensate citizens for those externalities, in the form of social and economic value.

Amongst the instruments to public intervention on a given market: 1) actions on the supply side, 2) actions on the demand side, and 3) mediation actions between supply and demand (regulation), we must carefully choose the most appropriate. In order to do so, it is useful to re-examine the conclusions of section 3.5, which introduced the concept of regulation as a necessary strategy to stabilize the data-sharing system into the dynamics of cooperation. This regulatory action is furthermore supported by an evident lack of an adequate compensation for positive externalities in the business of big data. We find some hints of possible regulatory paths in the modifications that a city like Zaragoza (Spain) introduced in public procurement, establishing that all private companies holding a public concession should comply to the high quality open data formats that the city enforces.

In order to compensate to the positive externalities mentioned above, similar rules could apply to companies doing private businesses in town event if they have not been granted with a public concession or contract. AirBnB's 'Get the data' portal (http://insideairbnb.com/get-the-data.html) offers a glimpse of a possible practical implementation of this idea: an on-line tool that allows anonymized data from the company's operations to be downloaded, geographically filtered and visualized for certain cities. A regulation which imposed higher requirements on the data, e.g. richer datasets, higher quality, dynamic data, etc at least for the research community would therefore mean a significant progress.

There is another aspect in which a stronger public action is needed: the question of privacy. As explained in sections 1.4 and 3.2, privacy, in our hyper-connected, on-line world, has shifted hands. It does not belong to individuals any more, but to the companies that provide the services we use. This phenomena can be described as a *privatization of privacy*. It implies that our private data are stored in distant databases, that, although the service providers should comply with local regulations in terms of personal data protection, the user would not know how to exert his o her rights in case of conflict or under which jurisdiction a possible lawsuit would take place, or even how to address to the customer care of faraway companies like Über, AirBnB, or even if such customer care exists and speaks his or her language. This privatization of privacy, in practice, leaves individuals defenseless. A data sharing regulation should also take this into consideration and enforce stronger guarantees over the personal data that the main private urban players hold about us. In this sense, city halls can have delegated responsibilities for personal data protection. They are the closest administration to the citizens, who are accustomed to their off-line and on-line procedures. By empowering city halls to

act as guardians of their citizens privacy, the rights of the urban dwellers can be better protected, and the process of privatization of privacy can start to be reversed.

But, besides regulation, city authorities can do probably more.

5.2 Can city hall help our local cab drivers to defeat Über?

Über has disrupted the rather stable taxi business in many cities. Using, essentially, the same technology than the traditional taxi (cars on four wheels), the company's success is based on a mix of borderline regulation tactics and efficient operations. Now, since every percentage of market share that goes into Über's bottom line implies decreasing local jobs and local taxes in cities, these may be interested in finding ways to help their local cab drivers stay competitive. Unfortunately, just releasing mobile apps will not likely impact Über's growth in a significant way.

There are two projects that may signal creative paths for city councils to help local taxi drivers stay competitive. In the U.S., a team from M.I.T. Senseable City Lab worked with 160 million records of taxi trips in Manhattan to asses on more efficient operations of the Big Apple's taxi fleet [32]. Their goal was to investigate to what extent taxi-sharing would represent an opportunity for the taxi sector. Working on time and GPS data of pick-up and drop-off locations and computing billions of alternatives, their findings show that there is an opportunity to implement a sharing mode 40% more efficient and affordable than the current single passenger mode. Their study points out that a psychological factor such as privacy could be the main barrier for such a system, and leaves for future work the search of the adequate incentives to overcome it.

In Spain, Zaragoza has developed an innovative tool to implement incentives and cross-policies through city services. Zaragoza's citizen card integrates around 20 city services: light-rail, bus, public bikes, parking, city Wi-Fi, swimming pools, public libraries, theaters, etc... and the taxi service. A good example of the power of such a tool is the 'taxi for disabled people' project, where people with disabilities can use specially adapted taxis for their daily routines paying each time only the price of a bus trip, the rest being transferred automatically from a budget created for that purpose to the taxi driver. Smaller, 'disabled ready' taxi cabs are used instead of the big and expensive yellow buses, thus giving a more efficient service. Taxi drivers, city hall and, specially, disabled people, win with a more demanded, cheaper, and more agile service. The citizen card acts as a platform over which similar innovative policies can be applied to specific user groups across a wide range of city services. In addition, it represents a gold mine of urban big data.

5.3 Funding

It is not the purpose of this work to establish a business model for this system, since that will be the goal of a complementary and future work. Instead, we will just mention, as a general guidelines, that capital expenditures can be faced through public funds. Programs such as the Horizon 2020 that the European Commission launched in 2014 (http://ec.europa.eu/programmes/horizon2020/) are seeking for the implementation of innovative programs and infrastructures around smart cities and innovation, but there are also a multiplicity of programs at other levels that can complement and fit in this funding scope.

It is the cost sustainability model which is more delicate, due to the highly skilled personnel needed for the system to operate at full performance. The fact that, as explained in section 3.5, there is no monetary exchanges between stakeholders, does not mean that the overall system can not look for revenue sources to fuel its own growth. Business services appear as a potentially interesting revenue sources. As an example, Bristol's datadome intends to offer to corporations the service of visualizing annual reports in spectacular 3D representations. But also, some research projects can be converted into off-the-shelf products. For instance, the work in [16] could be transformed into a product that could eventually study transport demands in urban or interurban mobility plans. Mobility plans are mandatory in many countries, and its periodic renovation implies high costs that could be decreased with such automated analysis, making for a business case which is worth exploring.

5.4 The costs of risk aversion

The aim of this work is to settle the foundations to implement a data-sharing system between the main city operational players. We have seen so far that this system already exists partially in different projects around the world. We have also gone through possible impediments along the way: data privacy, the silo effect, the sustainability of cooperation dynamics, or funding. But, at the same time, we have been able to identify a vast array of opportunities: the maturity of technologies around big data, the knowledge potential of mixing data flows, the crowd-sourced talented communities, the new horizontal and soft profiles on the rise, etc. To sum up, the operation presents some risks while it grasps a great opportunity.

To further limit the risks, we would propose an iterative launching plan, from a minimum viable product or system with two or three stakeholders (typically the city council plus one of the more proactive urban stakeholders) and small projects in which the benefits of data flow mixing could be easily identified. This initial progress would eventually validate the general concept, test technology and deal with legal issues (privacy, for instance) and would settle the

foundations for expanding the system's reach.

However, organizations (and, especially, public institutions such as city councils) are risk averse, which is not surprsing given that risk aversion is deeply rooted in our brains. As noted by [33] there are many examples in which risk aversion manifests. Take a gamble in which we are given 200$ if we win or are taken 100$ if we lose. Under these circumstances, a rational choice would be to reject the gamble, since the prospect of losing 100$ is stronger than the prospect of winning 200$. Many of us would even reject the gamble even if probabilities, instead of being 50/50 are, say, 40/60. But the scenario changes if we are given twenty of these gambles. A rational person knows that, statistically, we would end up winning. However, neither people nor organizations tend to consider risk policies as a series of decisions along time; on the contrary, risk assessment focus on individual choices or operations. We have seen that if potential earnings are bigger than potential losses the rational strategy is to take the risks even in 50% gambles (which does not exclude adopting risk mitigation policies). In cases where the probabilities of winning are greater than 50% and the reward in case of earning is greater than the loss in case of failing, not adopting innovative policies may cost public institutions large amounts of tax dollars.

The government of Canada seems to have grasped some of this thinking in its Blueprint 2020 strategy [34], when it states that 'through the Blueprint 2020 process, departments, agencies and communities have already committed to a broad range of actions that will directly benefit Canada now and into the future. Moving forward, as we foster a culture of innovation in the workplace, public servants will continue to identify and implement new ways to improve services, partnerships and communication.' Other cities, regions and states are issuing similar strategies to promote risk taking and innovation internally.

5.5 Data-driven government and the 'observer effect'

We have so far proposed the basis of a big data sharing system that can bring benefits to the economic, scientific, social and civic tissue of the city. On the institutional side, it may mean a significant progress towards the idea of data-driven government, which is, undoubtedly, a significant progress in government.

However, it may be convenient to reflect on some blurring limits that data-driven government needs to consider. The *observer effect* is a well-known phenomena in physics that states, basically, that measurement inherently alters the experiment. The observer effect works also when observing or measuring human behavior. Consider, for instance, how placing a camera on a street impacts security, or how openly monitoring a sales department's performance may drive employees to increase their sales record. The observer effect in human behavior multiplies when combined with incentives. Policies in data-driven government must be designed very carefully to avoid perverse effects such as the cheating scandal in Atlanta's school system [35], where an aggressive performance-driven strategy of incentives for schools ended up in a systemic cheating scandal. School principals and educators routinely boosted students marks competing for budget allocation from the city's administration.

Governments should use data to adopt better informed decisions, but experiments and subsequent policies need to be carefully designed to preserve equity.

5.6 Measuring failure and success: metrics

In the previous section we have dealt with both the convenience and limits of metrics in government action. To sum up, we have shown that measuring the impact of policies with the appropiate data is a significant progress, but that we have to carefully consider pushing data-driven decisions too far, specially when combined with incentives. Efficiency should not govern over equity.

It is not our purpose to design with full precision the metrics that should apply to our system. Having established upper and bottom limits for metrics, let's advance towards a series of general considerations that can help to draw the contours of such design.

Latest developments in processes and quality assurance in the software world counsel to define metrics up-front even before start writing code. By thinking of metrics beforehand the system's design benefits from a clearer view, since the inherent problems of project implementation inevitably biases the choosing of the appropriate indicators. Lean start-up adds to this view the notion of *vanity metrics*, an easy temptation when setting up the measurements of success. An example of vanity metrics is to measure the number of downloads when launching a new app. It suffices to drive a powerful social media campaign to get many people to download our app, but that figure says little about the future profitability of our business. A more interesting metric would be to know how many users recommend our app to a friend, how they use it, or how many actually pay for its premium features, as those would be factors that impact directly over the sustainability of our business model.

Metrics should be therefore able to measure the engagement of users rather than its number, and should respond to the main goals: creating social and economic value. On the social side, we need to answer to key questions such as: how are the stakeholders contributing to the system and what is their satisfaction level? How available datasets increase and improve? How are the research activities improving? What is the level of engagement of the different communities? How are urban services increasing its quality, equity and efficiency?

On the economic side we need to measure the impact in job creation (both quality and quantity). An example of this is the 'Wealth Generation Report' for the public CIEM Zaragoza Startup Incubator, located in the Digital Mile innovation district [36]. Located in a zero-emissions building and managed with a cooperative vision, it is obtained with a mix of qualitative interviews and raw data, and includes the following group of indicators:

- economic sustainability,
- social sustainability
- partnership and business co-operation,
- environmental sustainability

The example above shows that the subtleties inherent to complex and multi-layer urban policies can be incorporated to a measurement system. Of course, whatever metrics are chosen, they must be public. Open data is a must.

6. CONCLUSIONS

This work depicts some of the most important characteristics of an urban big data sharing system, leaving possible implementations for future work. The following list of conclusions can be taken as a material to help cities reflect on the convenience of building such a system, as well as a (non-comprehensive) general guidelines for its design, implementation and governance:

1. A 'value gap' to be filled. The modest dividends that big data is paying in the development of cities compared to what it brings to private companies, as well as the lack of powerful observatories for a new science of cities capable of scrutinizing what big data has to tell about urban flows make a strong case for building a data sharing platform between the main urban stakeholders capable of generating both scientific knowledge and new business and economic and social dividends for locals. A sort of 'Hubble' of cities, connected also with the civic and entrepreneurial ecosystem, to whom it brings value and opportunities.

2. Data as a public good and strategic asset. And therefore 'data mining' should be an activity that could be better regulated, to ensure that its benefits permeate the local communities where data is generated but also to reinforce privacy.

3. Four-dimensional geography and representation of flows. Cities change, and they change quick. The time dimension must be added to the geographical representation of urban phenomena. As in thermodynamics, it is the heat resulting from friction between people and between people and objects that produce the necessary energy to create and re-shape places, therefore the study of flows and its representation is an essential task in our system.

4. The feedback loop. Change, test, learn. The system needs to allow running cheap, quick experiments, in a simple trial-and-error mode. This turns the system into a agile, lean start-up tool, two concepts that cities can use as a way to avoid losing ground in front of the speed at which new digital businesses are deployed in our cities.

5. The city as a learning and innovation platform. City innovation services and facilities (networks, digital services, innovation centers...) can be conceived as 'final' services, but under certain conditions they can evolve into platforms open to third parties. The ultimate goal is that the city itself turns into an innovation platform intended to nurture and foster local talent, skills and knowledge.

6. Knowledge dissemination. As urbanization increases, there is a growing academic and research interest in cities. New knowledge about how cities function is being created, but that knowledge has to be disseminated to the citizens. By doing so, the process of slow permeation of this new science of cities into the cultural layer of our urban society will have started.

7. Live with privacy breaches. Privacy breaches can not be zeroed, but anonymation, aggregation, transparency, user consent and a strong public vigilance can help decrease the risks.

8. Erode data silos through 'soft' governance. In complex environments with a variety of actors and normative arrangements such as the world of urban services, government can no longer impose its rules. Governance implies softer techniques that 'erode' more than 'break' the inter and intra organizational barriers that inhibit sharing.

9. Flat organizational structure. Accordingly, the position of the stakeholders in our urban big data sharing system is that of partners in equal terms that voluntarily participate through incentives and win-win situations.

9. Connect with digital entrepreneurs through challenge-driven hackathons. A common flaw in open data hackathons is that they work on similar datasets. We propose to reverse the process. First the questions or challenges to solve, then the data.

10. a) Regulation: the cooperation stabilizer. The prisoner dilemma (and real experience) shows that cooperation at all scales might be difficult to sustain in time. Regulation plays a role here, in the way of underpinning cooperation dynamics. The goal is to make data sharing the longterm winning strategy of all stakeholders. Regulation is further justified from an economics point of view due to the 'data sharing value gap', or the market failure to provide social and economic returns with big data.

10. b) Regulation. Where is my data? Regulation can also help to reverse the process of 'privatization of privacy'. City halls can play a stronger role due to its closeness and accessibility to citizens, acting as proxies between users and, at least, the main companies that have the urban soil as its business operations field (telcos, mobility companies, utilities, etc). That role can reinforce the rights that relate to personal data.

11. Human APIs. Many cities are building, buying or coding 'smart' digital platforms, but digital platforms need to be complemented with human APIs that help making digital infrastructures accessible for the different user communities. Human APIs are sometimes neglected when designing digital infrastructures, but are probably the most effective path towards the idea of the city as an innovation platform.

12. Open source. Is a fundamental characteristic of the proposed data sharing system. Although it refers primarily to the ICT part, the open source concept can be extended to other layers such as physical buildings or processes. It brings accessibility, understandability, reconfigurability and transparency (accountability). In addition, the inherent work dynamics around open source are collaborative.

13 City hall's innovation pro-activity versus protectionism. Nowadays, innovation and city hall appear often in the same sentence. Despite not always being well understood by national governments, city halls are a source of innovation in fields such as politics, energy, transport, economy, education and technology, just to mention a few. Big data, combined with research can help to shape innovative urban services capable of facing external threats. Protecting local economies through innovation instead of shielding them through protectionism is at an arm's reach.

15. Can we afford the costs of risk aversion? If it is true that opening public and private data has the potential to unlock trillions of dollars, then launching a minimum viable product that implements this data sharing scheme should not represent a huge financial obstacle, given the multiple sources of funding that exists (smart city public funding programs, revenues from services, gains in business efficiency, etc) and the relatively low incremental cost of the proposed architecture.

16. Subtle and non-vanity metrics. Metrics should be setup beforehand, and have to be open. Due to the unavoidable 'observer effect' which is specially intense in cities, designers must be careful to choose adequate metrics so that the goals of the system are fostered and not corrupted. They must reflect the subtleties of the urban millieu and be linked to the core purposes of our data sharing arrangement.

7. FUTURE WORK

After setting in this paper the convenience and principles that may guide an urban big data sharing system, future work should deepen into practical implementation issues and into the business model.

A second line of work would conduct a thorough analysis of worldwide innovation hubs in order to identify those cities that better meet the conditions for implementing this system: strong public leadership, a participative and collaborative business ecosystem, a thriving civic community, as well as an open and advanced vision with regards to data, suitable facilities and access to funding opportunities.

REFERENCES:

[1] MapReduce: Simplified Data Processing on Large Clusters. 2004. Google, Inc

[2] Open data for economic growth. The World Bank. 2014

[3] Open Data. Unlocking innovation and performance with liquid information. McKinsey Global Institute. 2013

[4] Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. Han Hu, Yonggan Wen, Tat-Seng Chua. IEEE. 2014

[5] Challenges and opportunities with Big Data. Several authors. 2012

[6] Death and life of the great American cities. Jane Jacobs. 1961.

[7] Life between buildings. Jan Gehl. 1971

[8] Global networks, linked cities. Saskia Sassen. 2002

[9] The Information Age: Economy, Society and Culture. Vol. I: The Rise of the Network Society. Manuel Castells. 2002

[10] The new science of cities. The M.I.T. Press. Michael Batty. 2013

[11] Making sense of the new urban science. Anthony Townsend. 2015

[12] Proust Was a Neuroscientist. Jonah Lehrer. 2007

[13] Moore's Law Keeps Going, Defying Expectations. Annie Sneed. Scientific American 2015

[14] EU Directive 95/46. Protection on personal data. 1995

[15] Unique in the shopping mall: On the reidentifiability of credit card metadata. Yves-Alexandre de Montjoye, Laura

Radaelli, Vivek Kumar Singh and Alex “Sandy” Pentland. Science. 2015

[16] Understanding individual human mobility patterns. Marta C. González, César A. Hidalgo, Albert-László Barabási

[17] Why you can't really anonymize your data. Peter Warden. O'Reilly Radar. 2011

[18] Differential privacy, Cynthia Dwork. Microsoft research. 2006.

[19] Nest to share information with Google for the first time. Wall Street Journal. 2014

[20] To save everything, click here. Evgeny Mozorov. Penguin. 2014

[21] Can sharing your personal data protect your freedom? Sarlon Bowers. Sanger Institute. 2015

[22] Dependency Links Can Hinder the Evolution of Cooperation in the Prisoner’s Dilemma Game on Lattices and Networks. Xuwen Wang, Sen Nie and Binghong Wan. 2015.

[23] Collapse of cooperation in evolving games. Alexander J. Stewart and Joshua B. Plotkin. Proceedings of the National Academy of Sciences U.S.A. 2014

[24] Zaragoza's Open Urban Lab. The city as a platform for innovation . Daniel Sarasa Funes. 2015

[25] Smart cities. Big data, civic hackers and the wuest for a new utopia. Anthony Townsend. 2013

[26] Urban Acupuncture. Jaime Lerner. 2003

[27] The Triumph of the Cities. Edward Glaeser. 2011

[28] If Mayors Ruled the World. Benjamin Barber. 2013

[29] Governance as theory: five propositions about governance. Gerry Stoker. 1998

[30] E-topia. Urban Life, Jim – but not as we know it. William J. Mitchell. 1999.

[31] The art of city making. Charles Landry. 2006

[32] Quantifying the benefits of taxi pooling with shareability networks. Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevski, Steven Strogatz, Carlo Ratti. 2014

[33] Thinking, fast and slow. Daniel Kahneman. 2011

[34] Destination 2020. Government of Canada. 2014.

[35] Wrong Answer. Rachel Aviv. The New Yorker. 2014

[36] Wealth Generation Report. Digital Mile Business Incubation Centre. Init and Zaragoza City Hall. 2014

CONVERSATIONS:

The following persons had the patience and time to help in this research by means of fruitful conversations:

Elena Alfaro, BBVA Data Analytics.
Carlos Alocén Alcalde, Zaragoza City Hall
José Carlos Arnal Losilla, Zaragoza City of Knowledge Foundation
Claudio Feijóo, Polytechnic University of Madrid (UPM)
Dennis Frenchmann, Leventhal Professor of Urban Design and Planning - M.I.T.
Nigel Jacob, Mayor's Office of New Urban Mechanics, City of Boston
Michael Joroff, M.I.T.
Ana Jiménez Train, Zaragoza City Hall
Marta de Miguel Esponera, University of Zaragoza
Enrique Morgades, CIRCE (University of Zaragoza)
Kevin O'Malley. Bristol City Council.
José Manuel Páez, Polytechnic University of Madrid (UPM)

ACKNOWLEDGEMENTS:

This research was possible thanks of the support of Real Colegio Complutense, a joint institution between several Spanish universities and Harvard University at Cambridge, MA (U.S.A.).

Special thanks to the Sergio Ramos (Polytechnic University of Madrid – UPM) as academic supervisor and José Carlos Arnal (Zaragoza City of Knowledge Foundation) for their thorough revisions.