City Data: Big, Open and Linked

Mark S. Fox

Professor of Industrial Engineering and Computer Science
Senior Fellow, Global Cities Institute

Department of Mechanical and Industrial Engineering
University of Toronto
5 King's College Road, Toronto ON, M5S 2H2, Canada
tel: +1-416-978-6823, email: msf@eil.utoronto.ca

26 September 2013

Technical Report, Enterprise Integration Lab, University of Toronto, Sept. 2013.
http://www.eil.utoronto.ca/smartcities/papers/City-Data-v5.pdf
A version appeared in: *Municipal Interfaces*, November 2013, pp. 19-25.
http://www.nxtbook.com/naylor/MISF/MISF0513/index.php?startid=19#/26

Abstract

Cities are moving towards policy-making based on data. They are publishing data using Open Data standards, linking data from disparate sources, allowing the crowd to update their data with Smart Phone Apps that use Open APIs, and applying "Big Data" analytics to discover relationships that lead to greater efficiencies. This paper provides an introduction to City Data that is "Big, Open and Linked". It review sthe basics while drawing on examples from cities. It then explores the gap between the data's availability and its usability. Data from different sources cannot be used, i.e., linked nor analyzed, nor can longitudinal and transversal analyses be performed, unless they share common vocabulary and semantics. We explore the role of the Semantic Web and Ontologies in bridging this gap with examples drawn from Global City Indicators.

Keywords

City Data, Open Data, Big Data, Linked Data, Semantic Web, Ontology.

Introduction

Cities are moving towards policy-making based on data. They are publishing data using Open Data standards, linking data from disparate sources, allowing the crowd to update their data with Smart Phone Apps that use Open APIs, and applying "Big Data" techniques to discover relationships that lead to greater efficiencies.

One Big City Data example is from New York City (Schönberger & Cukier, 2013). Building owners were illegally converting their buildings into rooming houses that contained 10 times the number people they were designed for. These buildings posed a number of problems, including fire hazards, drugs, crime, disease and pest infestations. There are over 900,000 properties in New York City and only 200 inspectors who received over 25,000 illegal conversion complaints per year. The challenge was to distinguish nuisance complaints from those worth investigating where current methods were resulting in only 13% of the inspections resulting in vacate orders.

New York's Analytics team created a dataset that combined data from 19 agencies including buildings, preservation, police, fire, tax, and building permits. By combining data analysis with expertise gleaned from inspectors (e.g., buildings that recently received a building permit were less likely to be a problem as they were being well maintained), the team was able to develop a rating system for complaints. Based on their analysis of this data, they were able to rate complaints such that in 70% of their visits, inspectors issued vacate orders; a fivefold increase in efficiency.

Another example from New York City is their "grease disposal" compliance program[1]. 61% of all sewer backups are caused by improperly disposed of restaurant grease. The city combined sewer data, grease hauling licenses data and grease producer data to generate a prioritize list of places to inspect. This list resulted in a 95% success rate in detecting sources of grease entering the sewer system.

Big Data is the buzzword *du jour*. It is rare to read a newspaper or magazine without seeing an article on it. But these articles focus primarily on benefits with little being said about the actual technology. This paper provides an introduction to the concepts that underlie Big City Data. It explains the concepts of Open, Unified, Linked and Grounded data that lie at the heart of the Semantic Web. It then builds on this by discussing Data Analytics, which includes Statistics, Pattern Recognition and Machine Learning. Finally we discuss Big Data as the extension of Data Analytics to the Cloud where massive amounts of computing power and storage are available for processing large data sets. We use city data to illustrate each.

[1] http://www.nyc.gov/html/bic/downloads/pdf/pr/nyc_bic_dep_mayoroff_policy_10_18_12.pdf

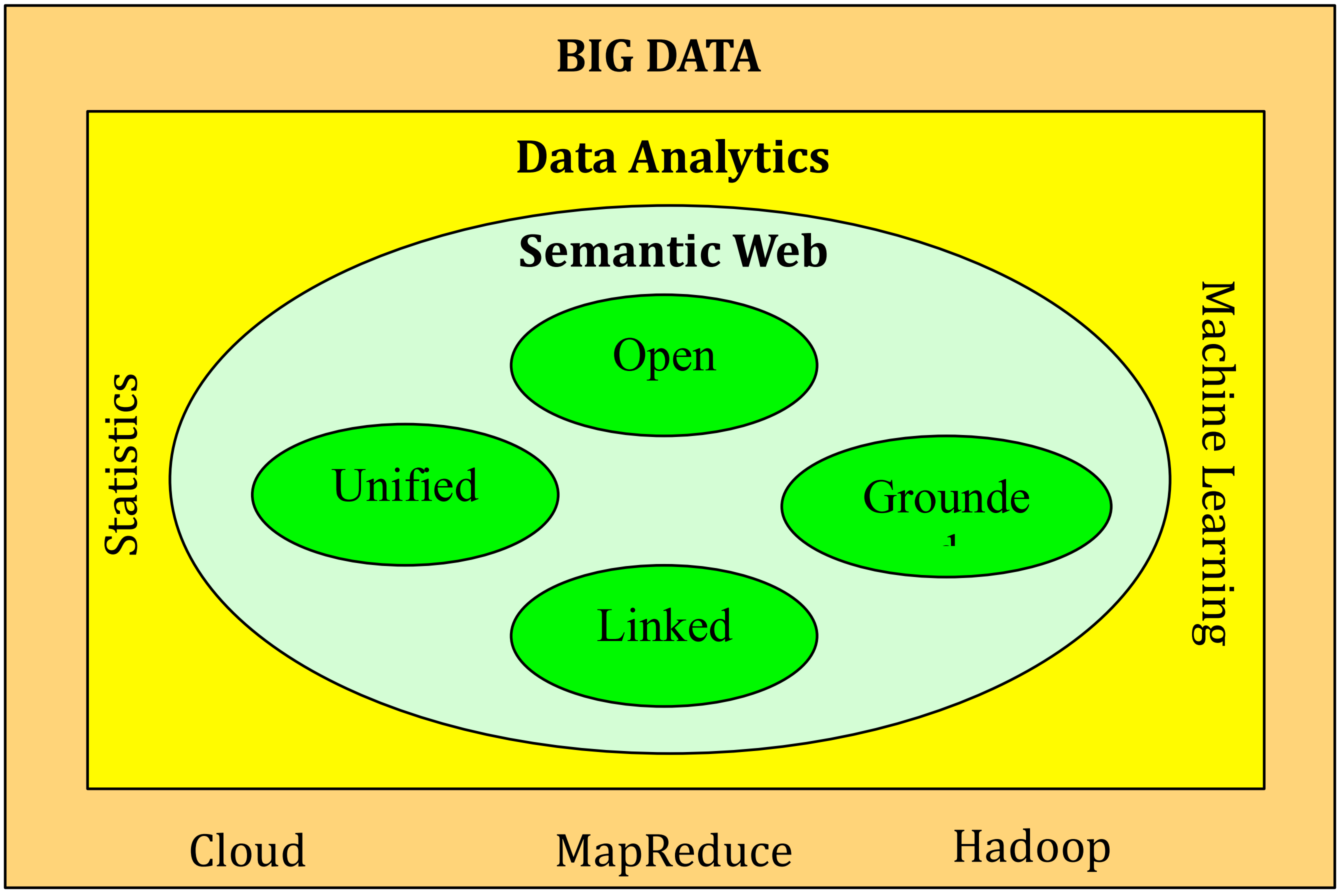


Figure 1

Open

The Open Data movement is part of the broader Open Government movement where the belief is that making data publicly available will lead to more effective public oversight. Yet there are other benefits than simply oversight. With more “eyes on the data”, waste and inefficiencies can be detected, crowd-based solutions suggested, and the crowd harnessed to implement some of the solutions.

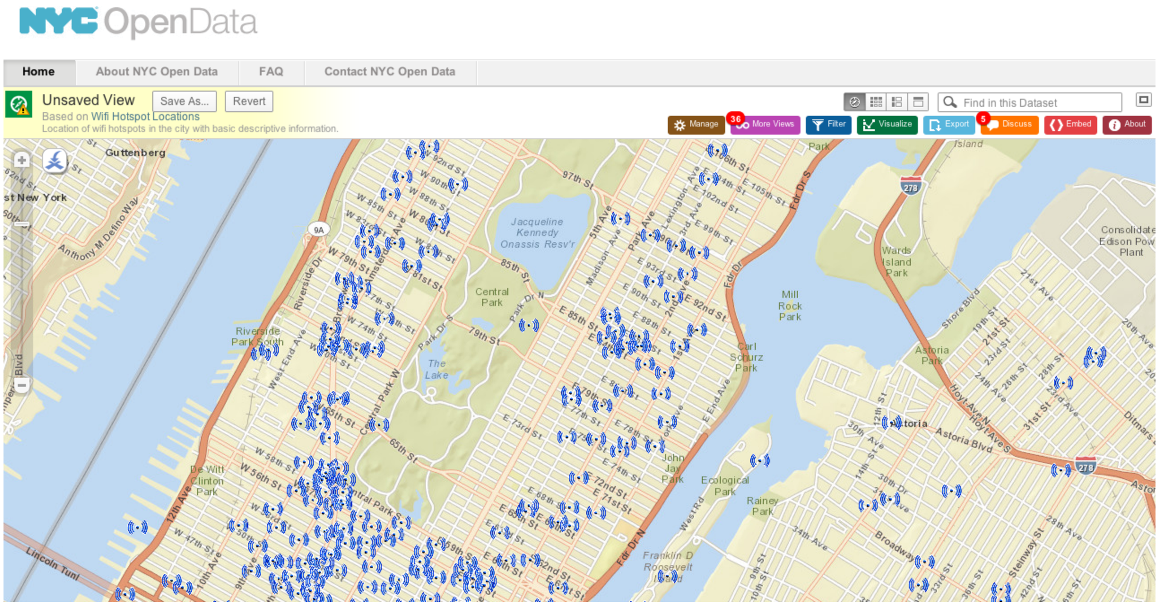
Within this context, the open data movement is composed of three efforts. The first and most pervasive effort is the publishing of datasets on city web sites. Canadian cities such as Ottawa, Toronto, Vancouver and Winnipeg, and American cities such as Chicago, New York, San Francisco and Seattle, all have major efforts underway to make city data publicly available[2]. Datasets available include 311 calls, 911 calls, police incident reports, building permits, traffic flow counts, bicycle flow counts, drop in centres, and much more.

The datasets are published in many forms, with spreadsheets (e.g., xls, csv) and XML being the most pervasive. An example is 311 Toronto, the non-emergency information and reporting service for the city of Toronto. 311 Toronto has published datasets covering Service Request Codes, a log of requests that have come into their call centre, and 311 Contact

[2] http://data.gov maintains a list of open government data efforts around the world.

Centre performance metrics, all in spreadsheet format. These files can be downloaded from their web site.

New York City has taken open data a few steps further. Not only is their data publicly available, it is *dynamic,* as it is being updated continuously. Secondly, they make their data *usable* by making it available in a variety of formats, including RDF, which is the dominant format of the Semantic Web. Thirdly, the are making the data visual. They are providing mashups of data with maps. For example, depicting the location of wireless hot spots on a map of New York (Figure 2).



Unsaved View

Based on Wifi Hotspot Locations

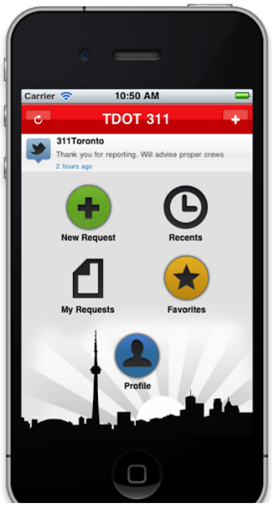
Location of wifi hotspots in the city with basic descriptive information.

Figure 2

The second effort is the publishing of public APIs through which applications such as smart phone apps or web sites can both retrieve and post data. 311 Toronto participates in Open 311[3], a global project that creates standard APIs for accessing information and services, and reporting. For example, GeoReport[4] is a standard API for reporting non-emergency issues such as potholes, graffiti, and street cleaning. 311 Toronto has implemented some of the APIs for reporting:

- Minor pothole damage, and
- Graffiti on private property, City roads, City sidewalks, City bridges, and City litter bins.

Mobile app providers have implemented apps conforming to these standards. One example (Figure 3) is TDOT 311[5], which reports pothole and graffiti to the city via its open 311 APIs. Hack-a-thons are taking place on a regular basis across the world where smart app



311Toronto

Thank you for reporting. Will advise proper crews

2 hours ago

[3] http://open311.org/

[4] http://wiki.open311.org/GeoReport_v2

[5] http://www.publicleaf.com/tdot311

developers over a period of days build new apps based on open city APIs.

The third, less developed effort, is the definition of standardized data models for city data, i.e., standard attributes and values. API specifications require a standard data format, and APIs such as Open311's GeoReport API use XML as their primary format, but have yet to define an XML schema for their format. We will address this issue in the next section.

Figure 3

Unify

Almost every example of Big City Data combines data from multiple datasets. With the availability of Open Data, how do we unify (or merge) data from multiple open datasets. In order to unify two or more datasets we have to be able to match up their respective attributes and values. Consider three datasets, two from Toronto and the other from New York City, containing 311 call centre requests. For each dataset we show a subset of attributes plus an example of a value for each attribute:

Dataset 1 (Toronto):

ID	Creation Date	Service Request Name
TO-Request1	01-02-2010 10:35:03	Missing/Damaged Signs

Dataset 2 (Toronto):

ID	Division	Section	Service Request Name	Problem Code
SAM-01	Transportation Services	TMC – Signs & Markings	Missing/Damaged Signs	SAM-01

Dataset 3 (New York):

ID	Creation Date	Agency	Complaint Type	Descriptor
NYC-Request1	4/30/2013 12:00:00AM	DOT	Street Sign – Damaged	Stop

Datasets 1 & 2 are from Toronto. The first dataset provides a list of call centre requests for the year 2010. Each request has its request time and its service request name. To determine which division/section of the city is responsible for responding to the request, you have to look into Dataset 2 and use the Service Request Name as the link (key). Dataset 3 from NYC substitutes a single dataset for the information found in Toronto's two datasets, where the request date, type and Division/Section (Agency) are associated with each request. One could imagine that the designer of the Toronto dataset decided to use two datasets so that the Division and Section information is not repeated in each request, thereby saving storage.

If the goal is to compare and contrast calls between Toronto and New York, there are a couple of issues that have to be addressed in unifying the three datasets. First is the mapping between attributes. Is NYC's "**agency**" attribute equivalent to Toronto's "**Division**", "**Section**" or both? Is Toronto's "**Service Request Name**" equivalent to NYC's "**Complaint Type**"? If we are to merge or compare the datasets, we have to determine the mapping between attribute names in one dataset to another, assuming a mapping is possible. The second issue is determining the mapping between values of equivalent attributes. Assuming that Toronto's "**Service Request Name**" attribute is equivalent to NYC's "**Complaint Type**" attribute, does there exist a mapping between values? Is Toronto's "**Missing/Damaged Signs**" equivalent to NYC's "**Street Sign – Damaged**". Without being able to map equivalent attributes and values across two or more datasets, it is not possible to merge, compare nor analyse the data.

The most common method for addressing the unification problem is to introduce standard vocabularies. A vocabulary is a list of terms providing standard Attributes (e.g., Complaint Type), Values (e.g., Missing/Damage Signs) and Objects (e.g., Citizen). Vocabularies can be local to the organization that created the dataset, or can be standardized within an industry. For example, Open 311 could have (but has not) defined a standard vocabulary for request types; this vocabulary would be used by cities across the world for specifying 311 request types. Or a vocabulary could be "universal", meaning that it is sufficiently abstract that it applies across all industries. The Dublin Core[6] is an example of a universal vocabulary that provides standard attributes such as the creator and creation date for describing web pages. DCAT[7], a product of the Government Linked Data Working Group[8], provides a vocabulary for describing government datasets.

The Government Strategic Reference Model (GSRM), a Treasury Board of Canada initiative (Buchinski et al., 2007), defines a standard vocabulary of government services terms. These terms include: policy, program, service and resource pool. Figure 4[9] depicts the core concepts and properties.

The Global City Indicator Facility[10] at the University of Toronto has defined a set of about 100 indicators for measuring city performance. These indicators entail a city metrics vocabulary that is rapidly gaining acceptance, with over 350 cities across the world having adopted them.

[6] http://dublincore.org/documents/dces/

[7] http://www.w3.org/TR/vocab-dcat/

[8] http://www.w3.org/2011/gld/wiki/Main_Page.

[9] Reprinted from Buchinski et al. (2007).

[10] http://www.cityindicators.org/.

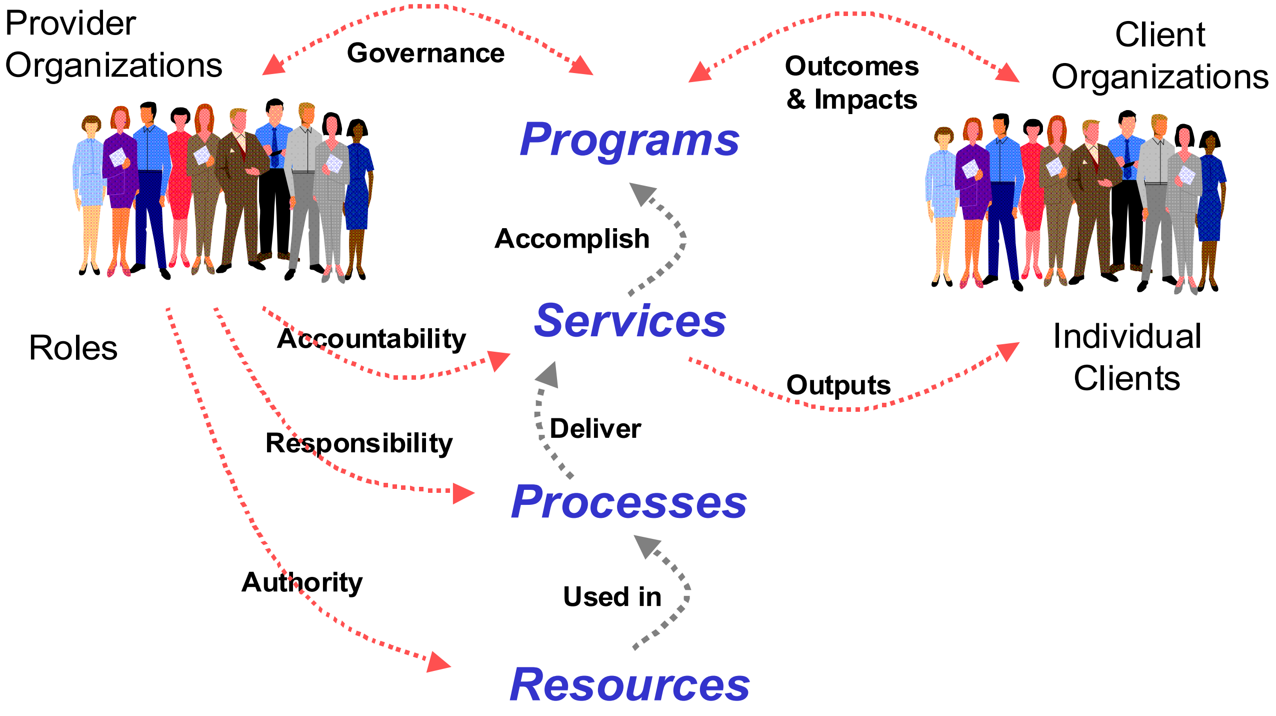


Figure 4

If cities adopt the same vocabularies for their datasets, then it simplifies the unification process. Otherwise, the unification will have to be done manually, with people often guessing as to what objects, attributes and values map onto each other.

Link

In the previous section, we made the case for adopting standard vocabularies for objects, attributes and values. If we want to unify data that is openly available across the internet,

- How do we share vocabularies across the internet? Where would a 311 vocabulary reside? How do we advertise their availability?
- How do we uniquely identify objects, attributes and values in these vocabularies? How do we refer to “Service Request Name” in Toronto and NYC?
- How do we access data on the internet whose structure (i.e., data model) is unknown to us?

Linked data (Heath & Bizer, 2011) provides a web-based solution to this problem. Vocabularies are stored as files accessible via the web. These vocabulary files are referred to using an IRI: International Resource Identifier. For example, lets assume that we create a file named “311.owl” that contains a vocabulary for 311 data using the OWL format (more on this later), and that the file is stored on an Internet server named “ontology.eil.utoronto.ca”. The IRI for this vocabulary file would be: **http://ontology.eil.utoronto/311.owl**.

There exist languages for defining vocabularies. RDF is the dominant language but OWL (Hitzler, P., et al., 2012) is quickly superseding it. Once a vocabulary is constructed form some domain, it is necessary to make it open. But how can this be done? How can anyone find your vocabulary file? Of course, one could you Google to search for a vocabulary file and that may be the best way. But there are web sites that are repositories for vocabularies. One is “Linked Open Vocabularies” (http://lov.okfn.org/dataset/lov/). Another is “Swoogle” (http://swoogle.umbc.edu/). Each can be searched for vocabularies that contain one or more keywords.

Lets assume that in the 311.owl vocabulary there is an attributed named “Request”. If we want to refer to the “Request” attribute defined in the 311.owl vocabulary file, we use an IRI with the term appended to the end with a #, e.g, **http://ontology.eil.utoronto/311.owl#Request**. This IRI uniquely identifies the **Request** attribute contained in the vocabulary file **http://ontology.eil.utoronto/311.owl**. If this IRI is used in multiple datasets, then each use refers to the exact same “thing”, namely a **Request**.

Another problem we face is how to access open data. How do we properly construct queries using languages such as SQL? In order to construct a query we have to know the structure of the data (i.e., its data model). If each dataset has its own unique structure, which they often do, we have a problem; we would have to construct a custom query for each dataset. If you are unifying 20 datasets, then you would have to create 20 custom queries. Linked data addresses this problem by simplifying the structure of the dataset. Linked data transforms datasets into a basic triple format: Subject, Property, and Value. For example, the three datasets in our example would be transformed into the following triples:

Subject	**Property**	**Value**
TO-Request1	Date	01-02-2010 10:35:03
TO-Request1	Service Request Name	Missing/Damaged Signs
SAM-01	Division	Transportation Services
SAM-01	Section	TMC – Signs & Markings
SAM-01	Service Request Name	Missing/Damaged Signs
NYC-Request1	Creation Date	4/30/2013 12:00:00AM
NYC-Request1	Agency	DOT
NYC-Request1	Complaint Type	Street Sign - Damaged
NYC-Request1	Descriptor	Stop

If all datasets adopt this triple representation, then constructing queries is greatly simplified.

The next step in linking the data is transforming the triples so that they adopt the same vocabularies. We do this by replacing Subject, Property and Values with IRIs where appropriate:

Subject	**Property**	**Value**
TO-Request1	http://purl.org/dc/elements/1.1/date	01-02-2010 10:35:03
TO-Request1	http://ontology.eil.utoronto.ca/311.owl#Request	http://ontology.eil.utoronto.ca/311.owl#DamagedSign
SAM-01	http://ontology.eil.utoronto.ca/organization#Division	Transportation Services
SAM-01	http://ontology.eil.utoronto.ca/organization#Section	TMC – Signs & Markings
SAM-01	Service Request Name	Missing/Damaged Signs
NYC-Request1	http://purl.org/dc/elements/1.1/date	30-04-2013 00:00:00
NYC-Request1	http://ontology.eil.utoronto.ca/organization#Division	DOT
NYC-Request1	http://ontology.eil.utoronto.ca/311.owl#Request	http://ontology.eil.utoronto.ca/311.owl#DamagedSign
NYC-Request1	Descriptor	Stop

In the above, the **Date** and **Creation Date** properties have been replaced by a single IRI, http://purl.org/dc/elements/1.1/date, which is a broadly accepted IRI for the date property. Where there exist IRIs that have been accepted by cities as appropriate for their data, the corresponding properties and values are replaced by those IRIs. Note that the **Agency** and **Division** are mapped onto the same IRI and **"Street Sign – Damaged"** and "**Missing/Damaged Signs"** have also been mapped onto the same IRI.

The process of transforming a dataset into linked data can be arduous, but tools are emerging that simplify the process. One tool is the open source software Open Refine[11] (originally Google Refine). Originally designed to clean up messy datasets, it has an extension that allows you to transform a data into IRIs and publish as triples. Best practices for publishing open+linked data have begun to appear. A W3C note on this topic can be found at: http://www.w3.org/TR/gld-bp/.

Berners-Lee[12] has introduced a 5 star rating scheme for Linked Open Data to encourage the development of good linked data:

1. "Available on the web (whatever format) but with an open license, to be Open Data,
2. Available as machine-readable structured data (e.g. excel instead of image scan of a table),
3. As (2) plus non-proprietary format (e.g. CSV instead of excel),
4. All the above, plus use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff,
5. All the above, plus: Link your data to other people's data to provide context."

Much of the city data being published today is at a 2 star level. There are copious amounts but limited to mostly spreadsheet formats. Few cities have reached the four star level. New York City is at the 4 star level as they provide their data in multiple formats including RDF. In addition, many of their datasets are being updated dynamically, such as 311 data.

Ground

Consider the problem of selecting an Education vocabulary. You want to use the vocabulary's term "Teacher" but are not sure what they mean by it. Does the term "Teacher" include both full and part time teachers? Does it include special education teachers along with regular teachers? Does it include nursery and kindergarten teachers? To answer this question, you most likely will have to search for a document that contains definitions of the terms in the vocabulary. If the document exists and is accessible, it is most likely written in a natural language (e.g., English) and is not machine readable. Consider the situation of trying to merge two datasets which both use the term "Teacher" but from drawn from different vocabularies. How can our software determine if these two versions of the term "Teacher" are equivalent? How can we provide definitions of the terms in the vocabulary so that they are machine readable, understandable and hence comparable? The answer is Ontologies.

[11] https://github.com/OpenRefine

[12] http://www.w3.org/DesignIssues/LinkedData.html

An Ontology is an “explicit representation of shared understanding” (Gruber, 1993). It “consists of a representational vocabulary with precise definitions of the meanings of the terms of this vocabulary plus a set of formal axioms that constrain interpretation and well-formed use of these terms” (Campbell & Shapiro, 1995). What distinguishes simple vocabularies from ontologies is the latter adds definitions of the terms and constraints on their interpretation using a computational language. Key to the creation of an Ontology is *grounding* the definitions of terms in lower level, more concrete terms (Jakulin & Mladenic, 2005). Consider defining the term “Student/Teacher Ratio”. The process of grounding the definition follows a thought process as follows:

- A student/teacher ratio is composed of the division of the number of students by the number of teachers.
- The number of students is defined by a count of the students in a particular city.
- A student is defined as a full time student in grades 1 thru 12.
- Full time is defined as spending at least 1500 hours a year in school.

As you can see, the definition of the original term, Student/Teacher Ratio, is composed of other terms in the vocabulary that, in turn, need to be defined.

The next question is: how do we create ontologies? This is the focus of Ontology Engineering. The primary goal of ontology engineering is to develop a shareable representation of knowledge. The belief is that by engineering the terms and axioms properly, they will be reusable across a broad spectrum of applications. With reuse, we can achieve interoperability, namely the ability to access, analyse and merge data from diverse sources across the web because they use the same ontology or specify a mapping between their ontology and other more broadly used ontologies.

Ontology engineering begins by determining the competency of the target ontology, which is defined by a set of questions that the ontology must be able to answer (Grüninger & Fox, 1995). Based on these competency questions, the terms (i.e., objects, attributes and values) and axioms are developed. Development takes a layered approach where application specific ontologies (e.g., manufacturing ontologies) are defined in terms (i.e., grounded in) of more foundational ontologies such as time, activity, resource, location, etc. For example, a manufacturing operation would be defined in terms of more general classes such as activities and resources. Secondly, if an ontology already exists that satisfies the some or all of the competency requirements, then it will be reused.

Once the competency of the ontology is defined, the terms are identified that are to be used to answer the competency questions, and are organized into a taxonomy of classes and properties. For example, for a 311 ontology, both `Division` and `Section` are classes (Figure 5) and are subclasses of `Organization_Unit`. Classes have properties whose values can be numbers or strings or other classes. An `Organization_Unit` can have a property `has_manager` with a value that is a string. If `Section` is a subclass of `Organization_Unit`, then `Section` “inherits” the property `has_manager` from `Organization_Unit`. Properties may also have values that are restricted to instances of other classes. For example, `Section` may have a property `processes_request` where the value is an instance of the class `Request` (e.g., the actual request `request1145` is an instance of class `Missing/Damaged Signs` which is a subclass of `Request`).

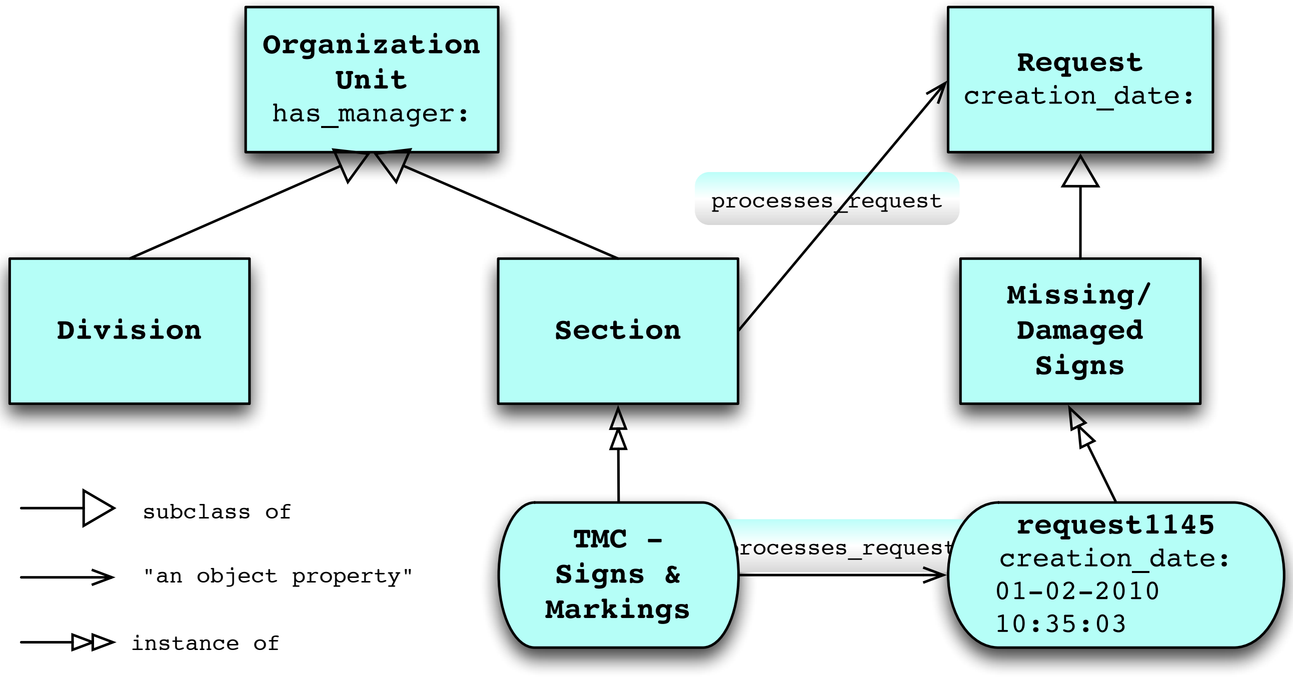


Figure 5

Axioms both define and constrain the interpretation of the terms. For example, a `Section` is defined as having to process at least one type of request. The `processes_request` property that links `Section` to `Request` is an axiom; it defines `Section` as having a property whose value is constrained to being a `Request`. While its use in linking the instance `TMC – Signs & Markings` to `request1145` is an assertion of a fact that `TMC – Signs & Markings` processed `request1145`.

A variety of methods exist for specifying ontologies. The most common method is a version of Description Logics (Nardi & Brachman, 2002) implemented in the Semantic Web language OWL 2 (Hitzler et al., 2012). For each method a variety of logic theorem provers exist to evaluate the consistency of what is being represented based on the axioms. The specification of terms and axioms, and their consistency testing is facilitated using a graphical, interactive tool such as Protegé (Noy et al., 2001). Most ontologies found on the Semantic Web are little more than taxonomies of classes and properties without axioms.

A notable exception is the SCRIBE ontology for modeling cities (Uceda-Sosa et al., 2012). It has classes and properties representing city organization and services, flow of events and messages, and key performance indicators. OWL definitions of the classes and properties are provided. Axiomatization is limited and so its use of foundational ontologies.

Global City Indicator Ontology

An example of a city focused ontology, is the Global City Indicator Ontology (Fox, 2013). "Today there are thousands of different sets of city (or urban) indicators and hundreds of agencies compiling and reviewing them. Most cities already have some degree of performance measurement in place. However, these indicators are usually not standardized, consistent or comparable (over time or across cities), nor do they have sufficient endorsement to be used as ongoing benchmarks" (Hoornweg et al., 2007). In response to this challenge, the Global City Indicator Facility (GCIF) was created at the University of Toronto, to define a set of city indicators that can be consistently applied globally. Over 350 cities worldwide are participating in this effort.

Hoornweg et al. (2007) identified that a good indicator must be Objective, Relevant, Measurable, Replicable, Auditable, Statistically representative, Comparable, Flexible, Potentially predictive, Effective, Economical, Interrelated, Consistent and Sustainable over time. Never the less, today's indicators are just numbers stored in a spreadsheet or database. Any information as to:

- the geographic area they were drawn from,
- what measurement scales were used and whether they were consistent,
- when the data was gathered,
- how the data was gathered, what population was sampled,
- who gathered the data, to what degree can they be trusted, and
- how valid is the data,

is often not documented, or if documented online, is in a form that cannot be processed by machines (e.g., MS Word document).

The Global City Indicator Ontology is designed to provide a precise representation with an unambiguous semantic interpretation that enables machine processing of City Indicator data over the Web. It integrates seven ontologies from across the semantic web. Figure 6 depicts the ontology dependencies.

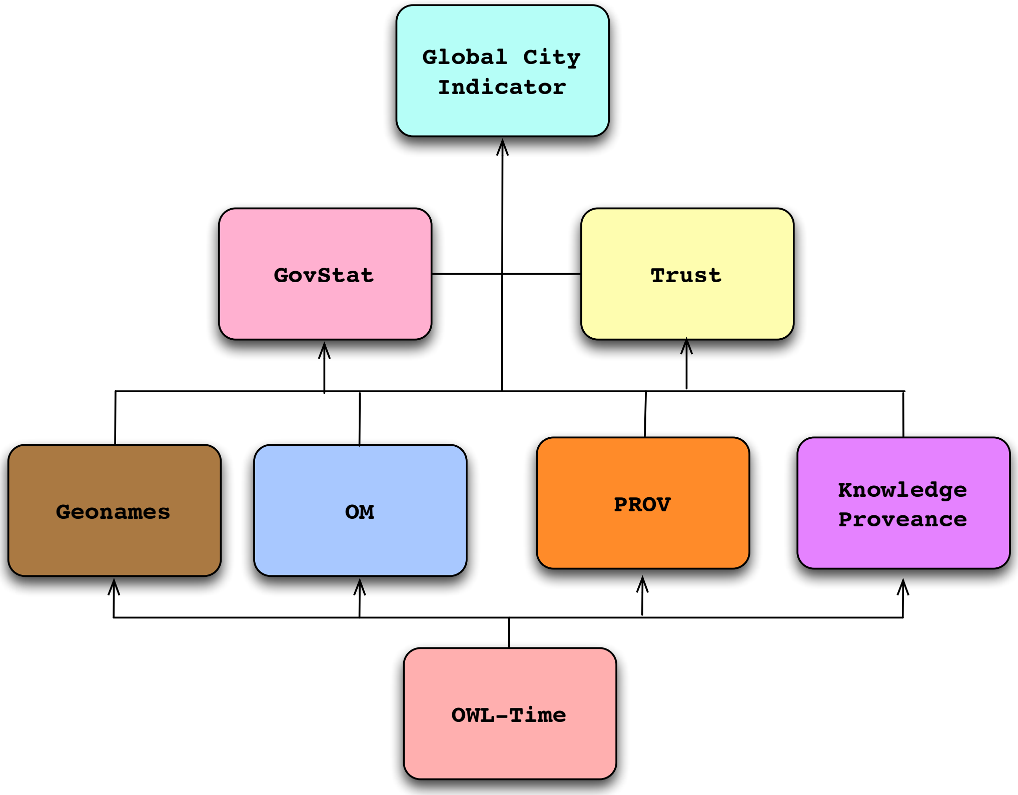


Figure 6

- **OWL-Time:** Basic representation of time points and time intervals, along with basic date and time representations (Hobbs & Pan, 2006).
  - Over what period of time was this GCI constructed?
  - How long did it take?
  - Was the teacher population sizing done during the same time that the student population sizing was done?
- **Geonames**[13]: An ontology for naming places along with a database of over 10 million named places.
  - What is the city being measured?
  - What version of the city? Time period?
  - What area does it cover?
  - What places does it contain?
- **OM**[14]: An ontology for measurement theory (Rijgersberg et al., 2011).
  - How is the indicator constructed?
  - What is it scale? Can one indicator be 2x another?
  - What are its units? Mega, kilo?
- **PROV**[15]: An ontology for representing the provenance of datasets (Belhajjame et al., 2012).
  - Who created the actual value of the GCI?
  - When was it created?
  - What process was used to create it?
  - Has this GCI been revised?
- **Knowledge Provenance**[16]: An ontology for representing the validity of data (Huang & Fox, 2004a; 2004b).
  - Is the GCI believed to be an accurate measure by its creator?
  - Over what time is it believed to be accurate?
- **GovStat**[17]: A basic ontology for representing statistics (Pattuelli, 2009).
  - What defines the members of the population?
  - What is its unit of measure?
  - Where is the population being counted?
- **Trust**[18]: An ontology for representing the degree of trust one has in data (Huang & Fox, 2006).
  - Trust the creator of the GCI?
  - Trust the process that created the GCI?
- **Global City Indicator**[19]: An ontology for representing indicators (Fox, 2013). It extends the other ontologies, where appropriate, to satisfy the Global City Indicator competency requirements.

[13] The Geonames Ontology is available at: http://www.geonames.org/ontology/ontology_v3.1.rdf.

[14] The OM ontology can be found at: http://www.wurvoc.org/vocabularies/om-1.8/.

[15] The PROV Ontology can be found at: http://www.w3.org/ns/prov.

[16] The Knowledge Provenance Ontology can be found at: http://ontology.eil.utoronto.ca/kp.owl.

[17] The GovStat Ontology is not available online, but a version with the GCI extensions can be found at: http://ontology.eil.utoronto.ca/govstat.owl.

[18] The Trust Ontology can be found at: http://ontology.eil.utoronto.ca/trust.owl.

[19] The Global City Indicator Ontology can be found at: http://ontology.eil.utoronto.ca/GCI-v1.owl.

Figure 7 shows a small portion of the representation of the educational indicator "Student/Teacher Ratio." It incorporates the definition of the ratio has having a numerator and denominator being a measure of student population size and teacher population size respectively. It also includes information on its provenance and degree of trust.

The GCI ontology is defined in OWL, and implemented in a prolog RDF server. In addition, a set of consistency axioms are defined and implemented to perform tests not possible using the OWL axiomatization.

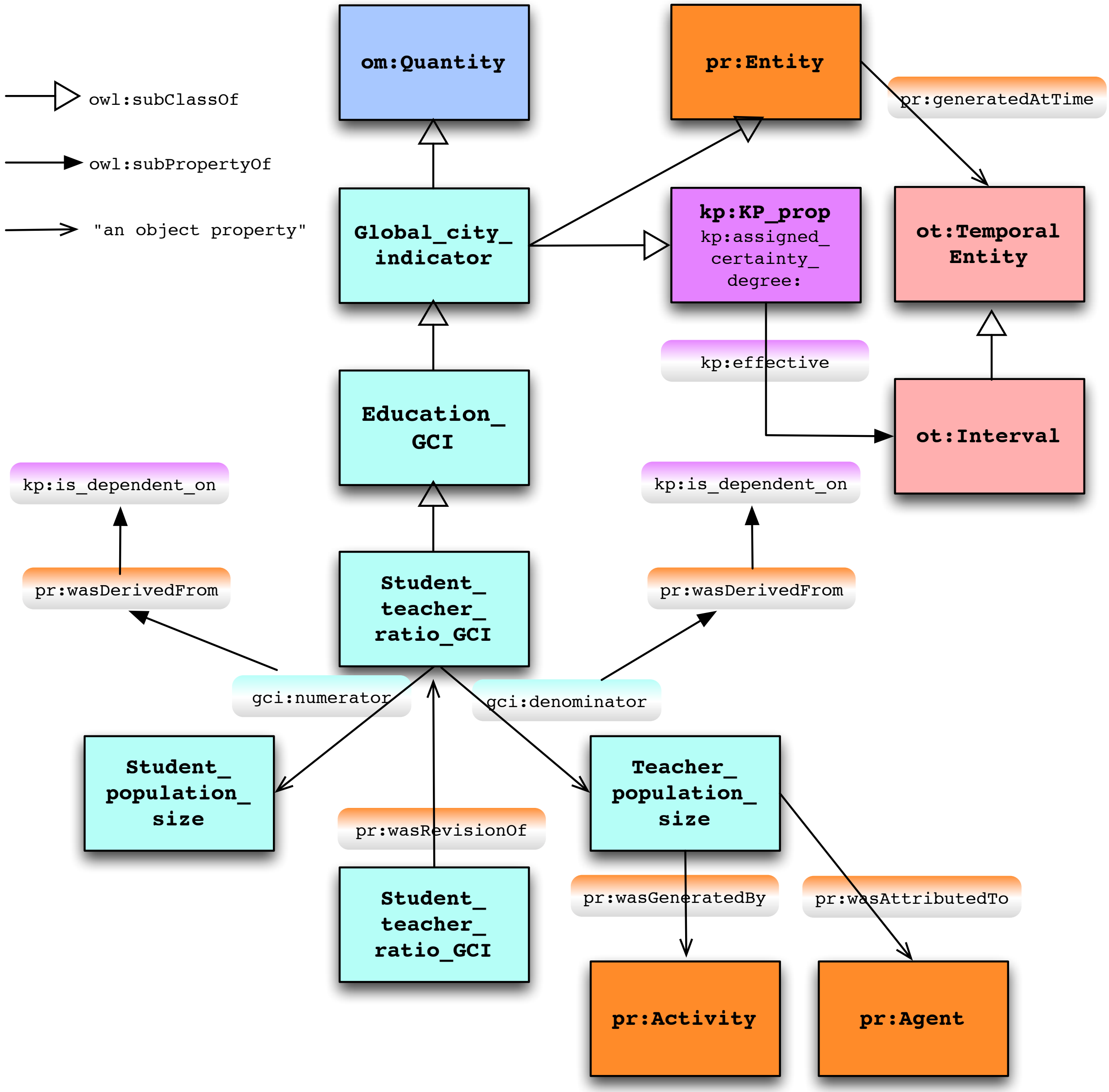
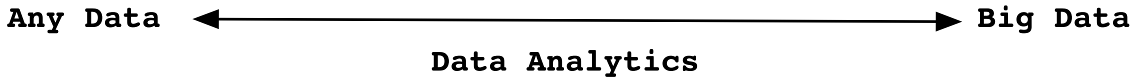


Figure 7

Analyze

Big Data has rapidly become a catchall phrase for any type of data analysis. At one end of the spectrum are organizations that for the first time are gathering data to analyze – I call this "Any Data" as they are looking for *any* data to work with. At the other end of the spectrum is the analysis of datasets measured in terabytes to petabytes.



Any Data ◄——► **Big Data**

Data Analytics

In between we have conventional Data Analytics aka Business Intelligence.

Data Analytics relies upon a broad set of techniques that have evolved over the decades from Statistics, Computer Science, Operations Research and other disciplines. For example, Data Mining (Han & Kamber, 2001) provides methods for discovering concise and succinct summarizations of concepts within a dataset. For example, what are the key characteristics of people who use rapid transit. Data mining also provides methods for learning association rules. For example, what is the probability that if someone reports a water main break, there will also be reports of potholes. Other techniques derived from Statistics (e.g., regression analysis) and Machine Learning (e.g., decision tree learning, neural nets) are also used in data analytics. Key to the use of these techniques is a strong understanding of what each techniques is good for and how to apply it. This has led to the creation of the field of study called "Data Science" which combines aspects of Computer Science, Statistics and Artificial Intelligence to create experts in analyzing data.

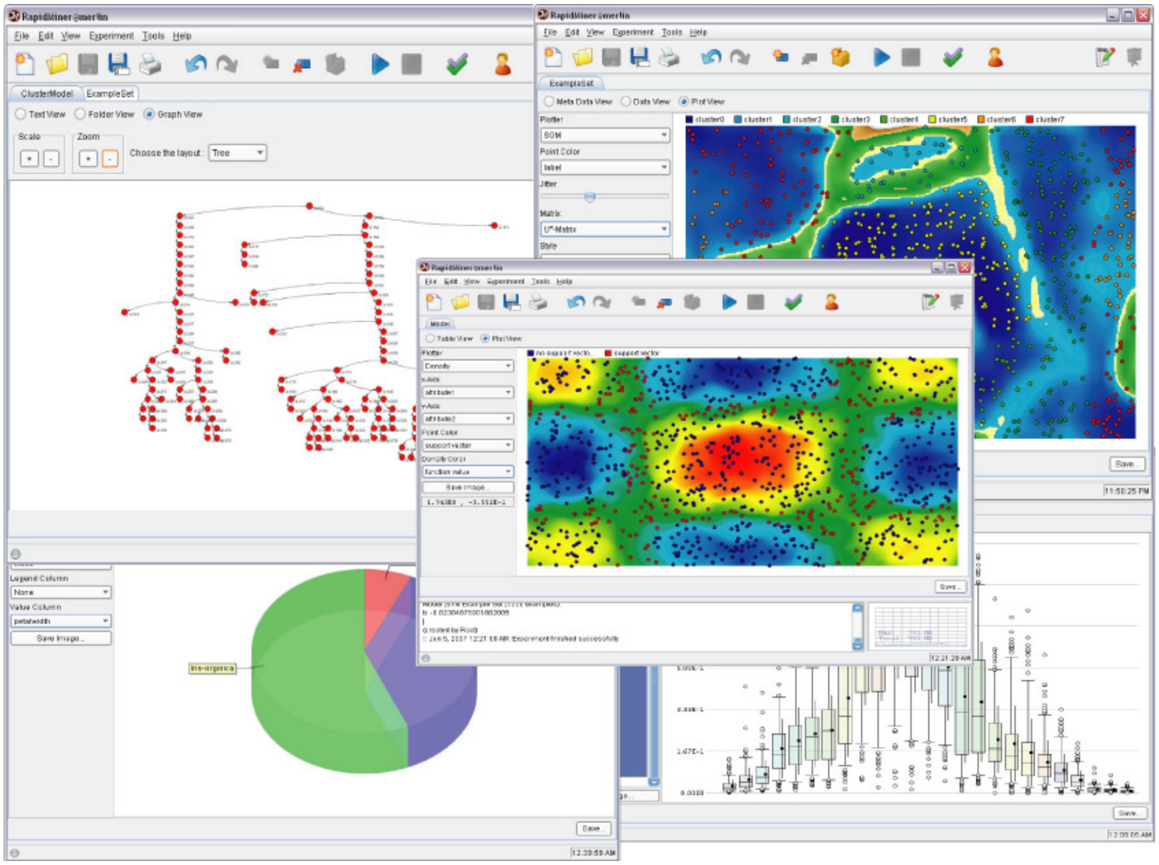
The two cases from New York City described in the introduction are good examples of data analytics. A third case is work done by IBM with the City of Boston (Srivastava, et al., 2013). In this case, traffic data from many sources, including manual counts, embedded road sensors and cameras, were cleaned, unified using an integrated data model and analyzed both visually and using data mining techniques. In the latter, time-series clustering algorithms were used to classify road patterns into six traffic patterns: commuting, going-home, anomaly, early-bird, night-owl and busy. Based on these classes, the planning of road maintenance, truck deliveries and other city operations can be optimized.

There exist a variety of commercial and open source tools available for analyzing data. One open source tool is Rapidminer[20] pictured in Figure 8. In order to apply these systems successfully, one must have a good understanding of the various methods they employ. Consequently, Data Science degree programs have begun to appear to teach basic courses in Statistics, Machine Learning, etc., and the tools for applying them.

One of the interesting outcomes of the growth of Data Analytics is the substitution of correlation for causation (Cukier & Mayer-Schoenberger, 2013). In other words, these techniques make it possible to discover interesting patterns, correlations, etc., but do not bring us any closer to understanding the underlying causal model that explains why. An example of this is the analysis of crime and abortion data in Levitt & Dubner (2005). The

[20] Visit http://rapid-i.com/ to download the tool.

correlation between the two leads them to believe that abortion has prenatally reduced the population of potential criminals. Whether this is truly the cause continues to be argued.



b: -1.02304675901083005

(created by Root)

Jan 5, 2007 12:21:08 AM: Experiment finished successfully

Figure 8

Big Data

Big Data refers to the analysis of large amounts of data to discover useful relationships and/or patterns. The recent ascendency of Big Data is due to the “Datafication” of society (Cukier & Mayer-Schoenberger, 2013) combined with the availability of huge amounts of on-demand processing power (i.e., the Cloud). The amount of digital data created daily is enormous: tweets generate terabytes; sensors, whether they be security cameras, traffic sensors, or Mars rover cameras, generate terabytes; web logs generate enormous amounts.

Big data analysis techniques build upon existing data analytics techniques, such as data mining. Because of the large size of the datasets, big data has modified these techniques using large scale parallel processing technologies (often cloud based) such as Google’s MapReduce and the Hadoop programming environment. MapReduce is a framework for distributing processing tasks across large numbers of processors. Hadoop, as depicted in

Figure 9[21], is a Java environment for distributed programming that uses MapReduce to parallelize the processing across multiple servers.

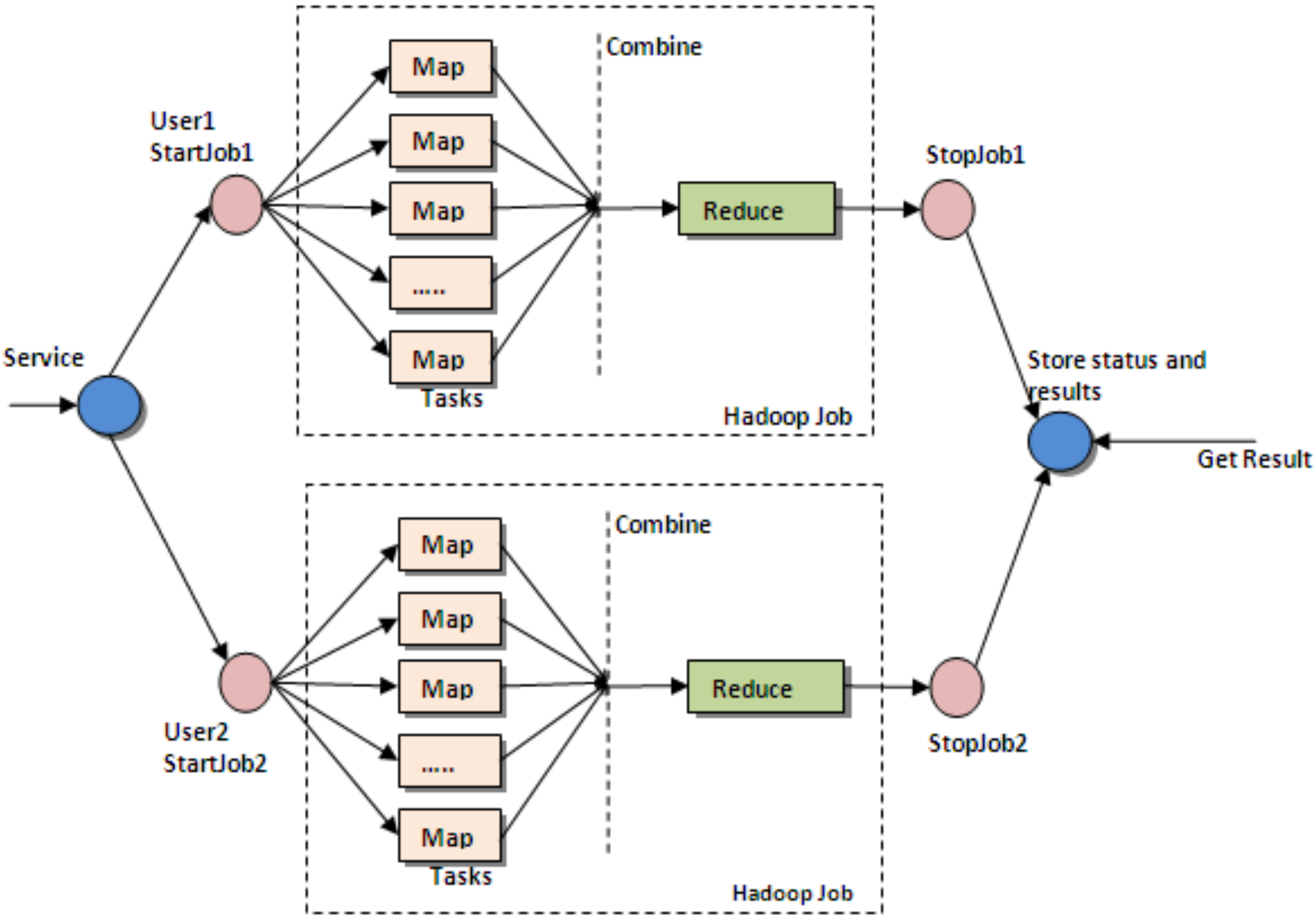


Figure 9

As noted in the previous section, the vast majority of examples found in the popular press are not examples of Big Data, but conventional data analytics where the datasets are measured in mega or gigabytes. Nothing beyond the processing power of a few processors is needed. What is new is the growing focus on making decisions based on data as opposed to gut feel.

Conclusion

The "Democratization of Data", as envisioned in the 1990s, was to be facilitated by three requirements: 1) "the locus of computing power and data access is broadening", 2) "less demanding level of skill that is now required to turn raw data into information", and 3) "the locus of applications must move closer to the citizenry" (Sawicki & Craig, 1996). In the last decade we have seen the democratization process accelerate at a blistering pace, facilitated by global accessibility enabled by the Internet/Web, massive amounts of cheap computer

[21] Reprinted from Stat221 lecture notes by Sergiy Nesterko. http://nesterko.com/lectures/stat221-2012/lecture17/#/

power provided by Cloud Computing, ubiquitous access to computing in the form of tablets, pads and smart phones, and the opening of vast amounts of data.

Never the less, the second condition of reducing the skill required to turn data into information is still nascent and requires a transformation on how data is defined and represented. The basis for this transformation is the Semantic Web (Berners-Lee et al., 2001). The World Wide Web was originally conceived as a global information space comprised of linked documents residing in servers accessible via the Internet. The Semantic Web envisions a "web of data that can be processed directly and indirectly by machines." By adopting the principles of the Semantic Web:

- Data being *openly* available over the internet,
- Data being *unifiable* using common vocabularies,
- Data being *linkable* using International Resource Identifiers,
- Data being *accessible* using a common data structure, namely triples,
- Data being semantically *grounded* using Ontologies,

we can develop the tools that will enable anyone to analyze data, both big and small.

Acknowledgements

I would like to thank Biplav Srivastava for his extensive comments on this paper.

References

Belhajjame, K., Deus, H., Garijo, D., Klyne, G., Missier, P., Soiland-Reyes, S., and Zednik,S., (2012), "PROV Model Primer", http://www.w3.org/TR/prov-primer.

Berners-Lee, T., Hendler, J., and Lassila, O., (2001), "The Semantic Web", Scientific American, May.

Bloomberg, M., and Merchant, R.N., (2012), "Open Data Policy and Technical Standards Manual", New York City Information Technology & Telecommunications, September 2012, http://www.nyc.gov/html/doitt/downloads/pdf/nyc_open_data_tsm.pdf.

Buchinski, E., Miller, G., Desmarais, C., Desmarais, J-M., and Jones, D., (2007), "Government Strategic Reference Model (GSRM)", Treasury Board of Canada Secretariat.

Campbell, A.E., and Schapiro, S.C., (1995), "Ontologic Mediation: An Overview", *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knoweldge Sharing*, Menlo Park CA, USA: AAAI Press.

Cukier, K., and Mayer-Schoenberger, V., (2013), "The Rise of Big Data", *Foreign Affairs*, May/June, pp. 28-40.

Fox, M.S., (2013), "An Ontology for Global City Indicators", Global Cities Institute Working Paper, Volume 1, Number 4: 1-45. Global Cities Institute, University of Toronto, to appear.

Gruber, T. R., (1993), "Towards Principles for the Design of Ontologies used for Knowledge Sharing." *Proceedings of the International Workshop on Formal Ontology*, Padova, Italy.

Grüninger, M., and Fox, M. S., (1995), "Methodology for the Design and Evaluation of Ontologies." *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, IJCAI-95, Montreal, Canada.

Han, J., and Kamber, M., (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Pub. Co.

Heath, T., and Bizer, C., (2011), *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool Pub.

Hitzler, P., et al., (2012), "OWL 2 Web Ontology Language Primer (2nd Edition)", http://www.w3.org/TR/owl-primer.

Hobbs, J.R., and Pan, F., (2006), "Time Ontology in OWL", http:www.w3.org/TR/owl-time/.

Huang, J., and Fox, M.S., (2004a). "Dynamic Knowledge Provenance", *Proceedings of Business Agents and Semantic Web Workshop*, pp. 372-387, National Research Council of Canada.

Huang, J., and Fox, M.S., (2004b), "Uncertainty in Knowledge Provenance", *Proceedings of the European Semantic Web Symposium*, Springer Lecture Notes in Computer Science.

Huang, J., and Fox, M.S, (2006), "An Ontology of Trust – Formal Semantics and Transitivity," *Proceedings of the International Conference on Electronic Commerce,* pp. 259-270. http://www.eil.utoronto.ca/km/papers/huang-ec06.pdf

Hoornweg, D., Nunez, F., Freire, M., Palugyai, N., Herrera, E.W., and Villaveces, M., (2007), "City Indicators: Now to Nanjing", World Bank Policy Research Working Paper 4114.

Jakulin, A., and Mladenic, D., (2005), "Ontology Grounding"*, Proceedings of the Conference on Data Mining and Data Warehouses*, Ljubljana, Slovenia.

Levitt, S.D., and Dubner, S.J., (2005), *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, New York: William Morrow.

Nardi, D., and Brachman, R.J., (2002), "Introduction to Description Logics", In *Description Logic Handbook*, edited by F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider, Cambridge University Press, pp. 5-44.

Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protege-2000. *Intelligent Systems*, IEEE, Vol. 16, No. 2,, pp. 60-71.

Pattuelli, M.C., (2003), “The GovStat Ontology: Technical Report”. The GovStat Project, Integration Design Laboratory, School of Information and Library Science, University of North Carolina at Chapel Hill, http://ils.unc.edu/govstat/papers/govstatontology.doc.

Rijgersberg, H., Wigham, M., and Top, J.L., (2011), “How Semantics can Improve Engineering Processes: A Case of Units of Measure and Quantities”, *Advanced Engineering Informatics*, Vol. 25, pp. 276-287.

Sawicki, D., and Craig, W., (1996), “The Democratization of Data: Bridging the Gap for Community Groups”, *Journal of the American Planning Association*, Vol. 62, No. 4, pp. 512-523.

Schönberger, V., and Cukier, K., (2013), “Big Data in the Big Apple*”, Slate.com*, 6 March 2013, http://www.slate.com/articles/technology/future_tense/2013/03/big_data_excerpt_how_mike_flowers_revolutionized_new_york_s_building_inspections.html

Srivastava, B., Rudy, R., Xu, J., Miller, B., Giacomel, A., Wysmuller, S., Gupta, V., Jacob, N., Osgood, C., Parker, K., Gately, C., and Hutyra, L., (2013), “A General Apporach to Exploit Available Traffic Data for a Smarter City”, ITS World Congress 2013, Tokyo.

Uceda-Sosa, R., Srivastava, B., and Schloss, B., (2012), “Building a Highly Consumable Semantic Model for Smarter Cities”, In *Proceedings of the workshop on AI for an Intelligent Planet*, ACM.

Yu, J., Benatallah, B., Casati, F., and Daniel, F., (2008), “Understanding Mashup Development”, *IEEE Internet Computing*, pp. 44-52.