Journal of Theoretical and Applied Electronic Commerce Research
ISSN 0718–1876 Electronic Version
VOL 12 / ISSUE 1 / JANUARY 2017 / I-V
© 2017 Universidad de Talca - Chile

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762017000100001

Editorial: What Can We Expect from Data Scientists?

***Kurt Englmeier*[1] *and Fionn Murtagh*[2]**
[1]*Schmalkalden University of Applied Science, Faculty of Computer Science, Schmalkalden, Germany, kurtenglmeier@acm.org*
[2]*University of Derby, Big Data Lab, Department of Electronics, Computing and Mathematics, UK, fmurtagh@acm.org*
Guest Editor
January 2017

Data Scientist - The New Profession

Data scientist is probably the most trendy job in Information Technology (IT) nowadays. This new profession emerged with the Big Data wave. Even though there is no such thing like an exact job profile, we expect that the data scientist can handle all the Big Data challenges that are novel to us. Without being a magician she or he shall help to deliver us all the magic Big Data promises. The data scientist capitalizes on unstructured data without taking the roles of a programmer, database expert, statistician, or content manager. All these professions are around for decades. So, why invent a new one?

“More than anything, what data scientists do is make discoveries while swimming in data […] they are able to bring structure to large quantities of formless data and make analysis possible.” [3] They sketch, orchestrate, and control the discovery process. The leading paradigm in this process it to find information that meets a certain need or provides an answer for a certain problem. “We need to avoid the temptation of following a data-driven approach instead of a problem-driven one.” [5] The data scientist has to develop an idea about the required information that meets our information need. We can expect that data scientists have a deep understanding of the foundations of both, the nature of information and the domain. They follow a mental model on the information demand that abstractly reflects the facts they expect to encounter in data and the way to detect and present them to us. We expect the data scientist to discover novel information that may provide us with new insights. And we want these insights to be true. It is thus part of the data scientist’s responsibility to make sure that the discovered information is not only novel but also trustworthy. The data scientists cannot prove data analysis models. That exceeds their capabilities. We cannot hold them liable for information that eventually turns out to be wrong. Nevertheless their skills should include a sound sensation of plausibility that helps them to raise doubts and to prompt a closer look when the results of data analysis seem too questionable to them. However, separating questionable results from plausible ones is a task that is far from being trivial.

Separating Data Science from Data Fiction

When driving a car we often encounter these nice roadside signs that sometimes make us realize that we’re driving too fast. Furthermore, on some occasions there is official personnel not far from these signs measuring our speed, explaining to us our traffic infraction face-to-face and documenting it on a speeding ticket. Doesn’t this sound a bit outdated? There are enough sensors in a car that measure location, speed, time, and more. Even if the car’s sensors don’t measure those, the driver’s cellphone can do it. Combining this sensor information with the cartographic data about roads and speed limits we can easily imagine that, by the end of day, the car or the phone exactly knows every infraction we committed and can automatically trigger the issuing of speeding tickets. Aren’t there just legal aspects that hamper this technically feasible scenario making its way into reality?

One trait of Big Data is the availability of sensor data and their combination with already available data generating new information, like the detection of traffic infraction. We can extrapolate this scenario in many directions, including more types of infraction, driving for an irresponsibly long time without a substantial break, or detecting patterns of aggressive driving. We can extend it also towards future possibilities if we think about sensor information from *smart* watches indicating a possibly problematic pulse rate or from the car’s air control sensor that the driver may drive under the influence of alcohol.

Much like in Data Mining, the strength of Big Data originates from the combination of facts producing new information. With sensor data we have much more new information sources available we can add to the knowledge our databases already harbor. Thus there are much more facts to analyze leading to new insights. Nevertheless the even more exciting source are texts, all types of comments, critiques, profiles, or descriptions on all sorts of platforms like Twitter, LinkedIn, Facebook, or TripAdvisor, just to name a few. These texts contain further valuable facts, they just need to be discovered. The sheer unlimited possibilities to combine our personal data stored digitally and our digital footprints, that we leave behind, using all sorts of electronic gadgets evoke some Big Data nightmares.

Imagine yourself dining in a restaurant: You register with your credit card and order one of your favorite dishes, let’s say roast pork with dumplings. For ordering you use your phone or the restaurant provides a tablet where it displays the menu. According to your recent medical diagnosis you should avoid too much of this type of heavy food. The restaurant's order management system automatically forwards the nutrition facts of your order to your personal health monitoring system you had to subscribe to in order to keep your monthly health insurance rate at bay. Sounds feasible, doesn't it? Taking the recent medial report and comparing your order with previous orders and the list of food you consumed at home, the monitoring system automatically rejects your order and selects from all available dishes the one that best suits your health condition. Instead of roasted pork with dumplings you get tomato soup with breadcrumbs. Now the story gets the trait of Data Fiction. First, the complexity of the model required to monitor your health condition together with your nutrition is quite massive. Even if we had all relevant data at hand, designing such a model that, fed into a reasoning system, leads to nutrition recommendations requires an extraordinary effort. Currently there is no silver lining on the horizon that IT will ever be in the position to develop such reasoning systems.

Take Google Flu Trends as an indicator that even less complex systems are too error prone to provide solid statements. It estimated the spread of a flu epidemic based on particular flu-related keywords in tweets combined with their authors' position data. Later the results of this analysis were proved to be somewhat far from reality [7]. Or take the Uber taxi service: it pretended to have detected service usage patterns indicating presumable one-night stands of its clients [11]. To presume that the detected correlation points to just this one cause is both embarrassing and scrupulous. The two examples shed light on a very Big Data problem besides the complexity problem: data analysis produces false positives. In Big Data analysis, it seems analysts do not investigate any further the broad variety of causes as long as correlations indicate new and spectacular insights. This is quite a real and actual threat of Big Data: we trust too fast and too much in the authority of analytic tools and neglect a thorough analysis of the causes that produced the phenomena appearing in the correlations. This may lead to severely false conclusions when we take the produced phenomena for granted too fast.

Correlations in data may stand for relationships between facts pointing to a phenomenon. If somebody is using keywords related to a certain disease in a tweet, we may reason that she or he is suffering from this disease and requires appropriate treatment. We see immediately, that the observed phenomenon is first of all an assumption about both, the person and the illness. And these assumptions can be wrong. This person may look up this information for somebody else, for instance. The combination of keywords may also stand for a completely different phenomenon.

Data scientists do not construct models for reasoning systems but should clearly be aware of the information they presume to produce. They should constantly check the results for false positives, in particular, when personal data are involved and/or expectations are high on the outcome of the data analysis. The data scientist is a scientist. She or he should take care that information is put to good use.

Data Science and Information Discovery

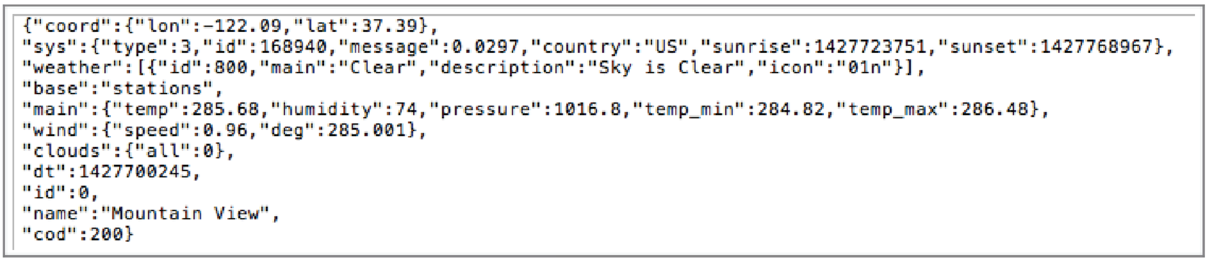
The complexity of data analysis models and the threat of false positives pose challenges for data sciences. Distilling factual information from unstructured data is a further one, in particular when facts are *hidden* in text. There is no general approach to information discovery from unstructured data. It is highly individual, theme-specific, and exposed to constant changes. This trait calls for an active involvement of the data scientist. She or he may delegate some subprocesses to machines, but discovery as a whole cannot be delegated to machines. Automatic information discovery is worthless if it bypasses substantial ingredients of the facts to be discovered. There are many discovery tasks that serve individual, ad hoc, and transient purposes. Automatic discovery can only focus on mainstream discovery that, in contrast, supports recurring discovery requests commonly shared by large user communities. Mainstream discovery addresses data like stock quotes from a stock market ticker service or weather data from a weather service, just to name a few. The corresponding process to identify, extract, and visualize these data can be reused for a broad community of users. The services addressed do not change their data representation too often. That raises the reuse of the discovery process, too.

Non-mainstream discovery, in contrast, satisfies only small user groups or even only individuals. Users may have to analyze from time to time dozens of failure descriptions or complaints, for instance. The corresponding information may be contained in bunches of text files or emails, where the required information is rendered in manifold ways. Dynamically changing small-scale requests would mean permanent system adaptation, which is too intricate and prohibitively expensive in the majority of cases. With a flexible self-service solution data scientists can reap the benefits of automatic information discovery and sharing and avoid the drawbacks of mainstream discovery. The smooth integration of domain and tool knowledge completes the picture of self-service discovery [10] that data scientists can easily handle [8].

The real challenge of the data scientist is the correct location of the right information ingredients in a sea of unstructured data. Insofar, information discovery's groundwork is mainly information extraction [2] or information filtering [1], because it needs to distill first the ingredients of the required facts from text. Subsequently, they need to be prepared for data analysis, that is, they need to be transformed into data formats suitable for data analysis. We may recall, that the concept of text includes also messages sent from sensors or provided by Application Programming Interfaces (API).

As we already mentioned, a data scientist is neither a programmer nor an expert in linguistics or Regular Expressions, albeit the two fields are extraordinary helpful in information extraction. Much like a content manager, the data scientist knows where the data come from and how they need to be combined and transformed. Moreover, the data scientist should be aware of concepts that reflect facts and their different forms of representation in texts. For instance, there are manifold representation forms of a date in texts. It can be expressed by *August 20, 2016, 2016-08-20, Saturday, 20, Friday last week*, just to name a few. Sensor information may show date and time as milliseconds since January 1, 1970 (see hour of sunrise and sunset in figure 1). Again, we do not expect a data scientist to code a function that turns the date representation in milliseconds into a more readable form for humans. Moreover, the data scientist should be aware of data concepts. She or he should know how granular concepts integrate into basic concepts and further into more complex ones. In the conceptualization of information lies the main focus of our expectations of a data scientist. This perspective on the data scientist's role inclines to the general understanding of problem conceptualization in Information Technology [6].

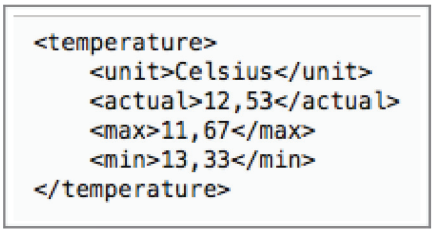
The granular form of a concept references a fact like date. Basic concepts, in turn, add more meaning by the inclusion of adjacent keywords or symbols forming concepts like birthday, sunset, due date, or the like. This means, the data scientist should be aware of the different representation forms of the fact date or time and should be in the position to annotate the facts accordingly, such as *sunrise* or *sunset*. The following examples help to understand the role of the data scientist in this context. A weather station sends its data in unstructured or semi-structured form, in text or JavaScript Option Notation (JSON) format for instance. The constituent parts of a message, namely the position of the station, air pressure, temperature, wind speed, etc., are, to some extent, human-readable as figure 1 shows. The same information from a different provider of weather information certainly comes also in a completely different form.



```
{"coord":{"lon":-122.09,"lat":37.39},
"sys":{"type":3,"id":168940,"message":0.0297,"country":"US","sunrise":1427723751,"sunset":1427768967},
"weather":[{"id":800,"main":"Clear","description":"Sky is Clear","icon":"01n"}],
"base":"stations",
"main":{"temp":285.68,"humidity":74,"pressure":1016.8,"temp_min":284.82,"temp_max":286.48},
"wind":{"speed":0.96,"deg":285.001},
"clouds":{"all":0},
"dt":1427700245,
"id":0,
"name":"Mountain View",
"cod":200}
```

Figure 1: Example of sensor information from a weather station

The data scientist has to find out how the required data are represented. This is essential when it comes to extract and transform data representations and to prepare them for the subsequent analysis. Discovery templates can help to design and communicate the extraction and transformation task. They reflect the syntactic pattern of data and include hints that help locating these data. To locate and extract the actual temperature in the example above the template may indicate the keywords *main* and *temp* to locate the required value *285.68*. Furthermore, the template indicates that the temperature in measured in Kelvin. A template contains also annotations that reflect the meaning of the data. Figure 2 shows an example transformation resulting from the application of the template that concentrates just on the temperature provided for the specific location.



```
<temperature>
    <unit>Celsius</unit>
    <actual>12,53</actual>
    <max>11,67</max>
    <min>13,33</min>
</temperature>
```

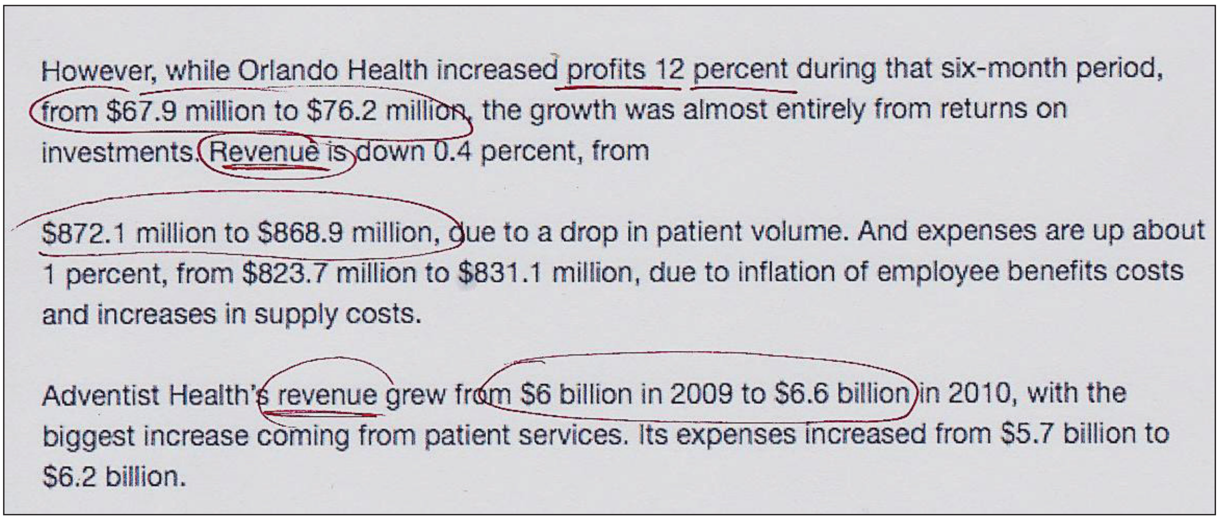
Figure 2: Example output from the application of a discovery template.

Discovery templates follow the paradigm of pattern recognition. Each template is thus linked to a descriptive pattern that can be rendered, for instance, as Regular Expression in combination with key terms. Regular Expressions are a powerful instrument to precisely detect all kinds of patterns in data. Their application is in particular useful for the detection of facts in unstructured information. There are several ways for programmers to derive machine-processable forms from those templates. There are other forms for output representation, too. They have to follow corporate standards that are usually defined by the people in charge of data analysis. The data scientist, in turn, is the person responsible for design, documentation, and communication of the templates.

When we consider articles, headlines, tweets, emails, etc. - in brief, texts that are authored by humans for human readers - things look different, but not too different. There is no standard way to express a fact in a text. Humans demonstrate an enormous creativity in illustrating facts in their own words. They tend to express the same thing in many different ways. However, to correctly analyze a complaint of a client about a product she recently bought the data scientist needs to include the discovery of facts like purchase date or product type, for instance. In each text, we can identify numerous facts of different complexities. Many of them can be identified in patterns that comprise facts located in close proximity. Information on a person may span over more basic facts like name, tax payer number, birthdate, address, etc. Some fact may comprise constituent facts that are widely scattered over a broader area of the data, even over more text pages or documents for instance. A data scientist knows about the concepts related to these facts, how their constituent parts integrate with each other, and how they can be represented in text.

To locate the right facts in a text, like increase or decrease in revenues, we need more than just keywords. Keyword search may detect that the piece of text (as shown in figure 3) deals with profits and revenues, but nothing more. However, it fails to locate essential information required to support the economic analysis of hospitals. Discovery templates focusing on information concepts like *light increase, moderate increase*, etc. help to identify the ingredients that the subsequent data analysis requires.

The examples show: there is no general approach to discovery. Even the most *standard* facts like a date are represented in manifold ways in different languages and cultures. Factual representations may also vary across domains. The concept *small lump* has certain attributes associated when appearing in the context of a lung diagnosis. It may cover instances like *lump of 2 mm in diameter* or *lump of up to 3 mm* and so on. In the context of cooking recipes *small lump* may be equivalent to *one or two tablespoons*.



However, while Orlando Health increased profits 12 percent during that six-month period, from $67.9 million to $76.2 million, the growth was almost entirely from returns on investments. Revenue is down 0.4 percent, from

$872.1 million to $868.9 million, due to a drop in patient volume. And expenses are up about 1 percent, from $823.7 million to $831.1 million, due to inflation of employee benefits costs and increases in supply costs.

Adventist Health's revenue grew from $6 billion in 2009 to $6.6 billion in 2010, with the biggest increase coming from patient services. Its expenses increased from $5.7 billion to $6.2 billion.

Figure 3: Typical representation of factual data in texts.

Information discovery has also a collective dimension the data scientist should manage. For, discovery starts with a hypothesis on the ingredients of information to be detected in data. This hypothesis leads the data scientists when sketching the required information that is finally manifested as a discovery blueprint. Data scientists initiate and control discovery by a certain belief-predisposition or bias reinforced by years of expertise - and gradually refine this belief. They gather data, try their blueprints in a sandbox first and check the results, and then, after sufficient iterations, they operationalize their blueprints in their individual world and then discuss them with their colleagues. After having thoroughly tested, they institutionalize them to their corporate world, that is, cultivate them in their information ecosystem.

The discovery blueprints thus serve three purposes: they reflect semantic qualities of the facts that need to be discovered, how representation of these can be located in data, and how the results are combined to support data integration and sharing for analysis purposes. While syndicating discovery blueprints along their domain competence, data scientists foster implicitly active compliance with organizational data governance policies.

Conclusion

There is no gold standard for the profile of a data scientist. We can roughly say that this person understands data and their value in the context of information discovery. Furthermore, this understanding is a common one, collectively shared over an organization. The collaborative trait of information discovery is not new. It has been discussed in related areas such as information extraction, filtering and gathering for decades (see [4] for instance). The collective management of discovery blueprints is comparable to the collective design of taxonomies that constitute the semantic layer of a corporate information ecosystem [9]. In the light of Big Data, this discussion gets new emphasis and importance. We expect the data scientist to design information discovery in a collective context.

Data scientists need a broad knowledge of the information concepts that constitute an organization's information landscape. In order to operationalize this knowledge an instrument that helps them to reap the benefits of information discovery and sharing and to avoid the drawbacks of mainstream discovery. Discovery templates serve to manifest and syndicate the information conceptualization in a corporate context. They also interface information concepts with machine-processable representations to support information location and extraction as well as transformation to suit its use in data analytic tools. We deliberately did not discuss technical details of the templates, because there are many forms they may take. For example, granular concept templates may use Regular Expressions for distilling facts. These granular concepts can be combined with keywords to address larger concepts. The keyword representation may include forms typically applied in information retrieval, namely stemming or synonym expansion to support standardization of template expressions. In any case, the concept gets proper annotations that, in turn, can serve the transformation of the extracted data with proper tags in Extensible Markup Language (XML).

The discussions on Big Data and Data Science have a strong focus on Data Mining, Statistics, even ontologies, etc. We do not deny that these fields provide useful instruments to benefit from Big Data. However, we may not neglect that many of these fields only partly address information discovery. We believe that there is no such thing like a discovery engine that automatically locates and distills all kinds of information. Conceptualization in the realm of information discovery is quite a human endeavor that qualitatively benefits if collectively undertaken. This leads to a more solid information governance that a data scientist should and can contribute to.

References

[1] N.J. Belkin and W.B. Croft, Information filtering and information retrieval: Two sides of the same coin?, Communications of the ACM, vol. 35, no. 12, pp. 29-38, 1992.
[2] J. Cowie and W. Lehnert, Information extraction, Communications of the ACM, vol. 39, no. 1, pp. 80-91, 1996.
[3] T.H. Davenport and D.J. Patil, Data scientist: The sexiest job of the 21st century, Harvard Business Review, vol. 90, no. 10, pp. 70-76, 2012.
[4] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry, Using collaborative filtering to weave an information tapestry, Communications of the ACM, vol. 35, no. 12, pp. 61-70, 1992.
[5] V.N. Gudivada, R. Baeza-Yates and V.V. Raghavan, Big data: Promises and problems. IEEE Computer, vol. 48, no. 3, pp. 20-23, 2015.
[6] J. Johnson and A. Henderson, Conceptual models: Begin by designing what to design, ACM Interactions, vol. 9, no. 1, pp. 25-32, 2002.
[7] S. Lohr, Google Flu Trends: The Limits of Big Data. New York: The New York Times, 2014
[8] T. Magaria and M. Hinchey, Simplicity in IT: The power of less, IEEE Computer, vol. 46, no. 11, pp. 23-25, 2013.
[9] F. Murtagh, M. Orlov and B. Mirkin. (2016, July) Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research. Cornell University Library. [Online]. Available: https://arxiv.org/abs/1607.03200
[10] J. Parenteau, R.L. Sallam, C. Howson, J. Tapadinhas, K. Schlegel, and T.W. Oestreich. (2016, February) Magic quadrant for business intelligence and analytics platforms. Gartner. [Online]. Available: www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204
[11] Z. Tufekci and B. King, We Can't Trust Uber. New York: The New York Times, 2014