Data Scientist - Manager of the Discovery Lifecycle

Kurt Englmeier[1] and Fionn Murtagh[2]
[1]*Faculty of Computer Science, Schmalkalden University of Applied Science, Blechhammer, Schmalkalden, Germany*
[2]*Department of Computing and Mathematics, University of Derby, Derby, U.K.*

Keywords: Data Science, Big Data, Information Discovery, Data Analysis, Data Mining, Text Mining, Information Extraction, Information Usability, Semantic Layer.

Abstract: Data Scientists are the masters of Big Data. Analyzing masses of versatile data leads to insights that, in turn, may connect to successful business strategies, crime prevention, or better health care just to name a few. Big Data is primarily approached as mathematical and technical challenge. This may lead to technology design that enables useful insights from Big Data. However, this technology-driven approach does not meet completely and consistently enough the variety of information consumer requirements. To catch up with the versatility of user needs, the technology aspect should probably be secondary. If we adopt a user-driven approach, we are more in the position to cope with the individual expectations and exigencies of information consumers. This article takes information discovery as the overarching paradigm in data science and explains how this perspective change may impact the view on the profession of the data scientist and, resulting from that, the curriculum for the education in data science. It reflects the result from discussions with companies participating in our student project cooperation program. These results are groundwork for the development of a curriculum framework for Applied Data Science.

1 INTRODUCTION

Data Scientists have the most attractive job, these days, haven't they? Not necessarily. In any case, they have one of the most demanded and most demanding jobs in IT. From health over finance and marketing to crime prevention, in almost all sectors we see an astonishingly rising need for data scientists. It is true, with Big Data we have a lot more data to delve into and data scientists can help us not to enjoy these data and not to get drowned. Our smartphones turned into a communication centre. We stay not only in touch with our beloveds, colleagues, and business partners, but also with society as a whole through Twitter and the like. Apart from that, our phones can also communicate with all kinds of devices and trigger actions, like turning on the light at home or the heating or telling us the nearest available parking lot in the area. A deluge of sensor data, biomedical data, including location data, retail transaction data, and the like merge with all kinds of text messages and data on social media activities and form a yet unseen sea of data. It hosts an equally unseen variety of information that has the capability to give us astonishing insights about the reality surrounding us.

We expect much from Big Data. To meet our expectations we count on approved disciplines, most prominently data mining, on new tools, most prominently Hadoop, and on new professions, most prominently the data scientist. Data science has strong roots in artificial intelligence. Education in data science nowadays focusses on Bachelor and Master programmes developed around topics like Data Mining, Neural Networks, Artificial Intelligence (AI), and Machine Learning. However, is it really necessary that the data scientist must be an expert in these fields? Big Data roused expectations that seemed dormant so far, despite Knowledge Mining, Predictive Analysis, Machine Learning, etc. being around for decades now. Since that time, we have analysts and statisticians working hard with all kinds of mining tools to provide us with insights emerging from the data around us. Do we fear they may fail in Big Data Analytics? Is Big Data too big for them? Why do we yearn for a new kind of profession? Data Science is certainly more than just the extension of data mining, machine learning and the like towards unstructured data and non-SQL databases. We propose to take information discovery as the overarching paradigm. This broadens the perspective on data science

Englmeier, K. and Murtagh, F.
Data Scientist - Manager of the Discovery Lifecycle.
DOI: 10.5220/0006393801330140
In *Proceedings of the 6th International Conference on Data Science, Technology and Applications (DATA 2017)*, pages 133-140
ISBN: 978-989-758-255-4

and brings the expectations into focus that information consumers conjoin with Big Data.

Through our student cooperation projects where students develop practical solutions in close cooperation with partners from the industry we learned that information discovery is a prominent theme for professionals and practitioners in industry when approaching the Big Data challenge. Viewing data science from the perspective of information discovery reveals the semantic and organisational dimension that reaches beyond the mere AI focus. From this perspective we develop a job profile for data scientists and give recommendations for a corresponding structure of a curriculum for the education in data science.

This article thus sketches the role of the data scientist as enabler of actionable insights emerging from Big Data. We see the data scientist's responsibility

- in the design of an overarching semantic layer addressing data and analysis tools,
- in identifying suitable data sources and data patterns that correspond to the appearance of structured and unstructured data, and
- in the management of the information discovery lifecycle and discovery teams.

It is important to note that we rather consider information discovery as a cooperative effort where discovery processes and tools are continuously evaluated and improved by discovery teams. These consist of experts with statistical and computational knowledge and skills, domain experts, and finally the data scientist. Discovery improves through experience, and the more intense the mutual exchange of discovery experiences then the better the teams perform.

The article also highlights the semantic dimension of data science with its ramifications into text mining, opinion mining, sentiment analysis and the like. It outlines the discovery lifecycle that is the process to manage information discovery as a commodity on workgroup or corporate level. Practical examples from the economic analysis give a hint on the variety of discovery and provide an impression why a user-driven approach may lead to data analysis solutions that meet the requirements of information consumers. Finally, the article presents our recommendation for a curriculum in the education of data science.

2 INFORMATION CONSUMER'S VIEW ON DATA SCIENCE

The data scientist must not necessarily take the role of a statistician, data analyst, database expert, content manager, or programmer. We have all these professions around for decades. If it is for these professional qualities, do we really require a new profession? Is it because Big Data means more unstructured data and more non-SQL access to data? Many university curricula in data science cover exactly these areas. Does it mean, data scientists have to be professionals in all these areas or does it suffice for them to know just a bit of everything? Shall they be just all-rounders in AI? There is no doubt, AI provides a lot of useful techniques and methods that help us to master the Big Data challenge. Nevertheless if we view data science from the perspective of AI's capabilities to contribute to Big Data solutions, we may bypass some expectations whereby information consumers connect with Big Data. Even though we cannot satisfy all their expectations it could be quite useful to view data science from the information consumers', perspective. In general, they expect to discover information in data that reveals insights to facts reflected in these data. Furthermore, they expect the data scientists to take care of their expectations (Englmeier and Murtagh, 2016). Decision makers must hold their own with their theories, their views, and their practices. Marketing managers, for instance, have to bolster their decisions on a particular advertisement campaign, that is, their theory on how many potential customers they reach and how this translates into increasing sales of a particular product. In the past, they filled some Excel spreadsheets with sales data from their company's database and plugged in their assumptions on customer behaviour in different segments and regions. If their forecasts proved wrong, it was their assumptions that were blamed first together with their model on how the campaign impacts customers and their purchases. This, in turn, negatively affected their reputation. No wonder decision makers want to keep an eye on their assumptions and their theory behind, that is, on well-reasoned insights that help to make their assumptions defensible. Therefore, they want data scientists who understand their needs and discover useful information.

Data scientists are first of all professionals in information discovery. "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." (Davenport and Patil, 2012) They sketch, orchestrate, and control the discovery process. The leading paradigm in this process is to find information that promises the insights information consumers require. "We need to avoid the temptation of following a data-driven approach instead of a problem-driven one." (Gudivada et al., 2015). Data scientists, thus,

should listen to people, understand their information need, and manage the Information Discovery Lifecycle that supports this need.

3 SEMANTIC DIMENSION OF DATA SCIENCE

Data scientists are data explorers, discoverers who detect and locate facts in–mostly unstructured–data (mainly texts) and prepare them for data analysis. On the other end, they have to ensure that the excitement about discovery in Big Data pays off in well-reasoned insights, that is, they have to ensure that decision makers, or information consumers in general, get solid information that leads to solid and reasonable decisions. The data scientist's challenge is to discover facts–to reveal whereabouts and facets of information in the sea of structured and (mainly) unstructured data. This makes the process of information discovery getting essential in data science. Each fact may take a variety of forms of appearance in data and that may vary even more over different data sources. They range from different spellings in texts, over different representations in database schemata, to special encodings on digital cards like the health insurance card or the passport. If we recall how many forms even a simple fact may take, then we get an impression that handling the variety of appearances of facts is far from being trivial. Without transforming all these different forms into one standardized representation we cannot even feed the corresponding fact into our data analytic systems. Information extraction and normalization is thus a broad and important field the data scientist has to manage.

The data scientist maps a certain variety of data patterns of a particular information concept to its standardized form reflecting the corresponding fact. The facts map to superior concepts with relationships among them. Data analysis takes facts and relationships and tries to derive new information from the results. It is again the data scientist who takes all this information together and presents it to the people who require it. The information must be useful, that is, it must enable to its consumers well-reasoned insights. Information usability is the key indicator for the quality of the discovery process and it depends on the needs of the information consumers. The data scientists design, manage, and control the discovery process, in short, they manage the Discovery Lifecycle. Therefore, this new profession has a closer link to enterprise information integration and human-computer interaction than to artificial intelligence. Data Science is much more than just an extended version of data mining and machine learning covering increased volume, variety, and velocity of data unleashed by Big Data. It has to manage the semantic dimension of this rich and volatile data ecosystem. This dimension is not only important because the vast majority of Big Data is text data. In fact, Big Data has a strong language background. We want to discover facts and the relationships among them that reveal the phenomenona we are interested in. That can include opinions, sentiments, lines of arguments, etc. The facts are usually terms or phrases in combination with numerical data. We need a strong semantic model that helps us to describe the fabric of impact among the facts. This model guides not only our exploring of the data environment but also the discovery processes we use to locate and extract the data. There are many recurring elements these process models may have, but there are no standard models. Usually, they need to be constantly adapted to the changing data environment.

4 DISCOVERY LIFECYCLE

Big Data is not just one homogenous source of information, it is masses of heterogeneous sources spawning avalanches of data. Discovering right sources and valuable data is not a trivial thing. It emerges from a combination of three knowledge areas, namely domain knowledge, technical knowledge and managerial capabilities. The domain knowledge serves to get an abstract conception of the information environment where discovery takes place. This abstract conception helps the data scientist to develop an overarching mental model that, in turn, guides the development of a semantic data layer on top of the data sources. This layer is a building block for a classification system that handles not only data but also analysis tools. It supports data scientists to determine what tool and what kinds of data are best suited for a specific purpose. The technical knowledge covers the knowledge on how facts appear domain-independently in data and how to standardize them. Checking positive or negative comments about a product, for instance, requires domain knowledge on the characteristics of the product and its competitive products, such as price, labels, etc., the full array of qualities producers and consumers attribute to this product. The knowledge on data patterns manifests the data scientist's technical knowledge. Product characteristics like price, size, power, energy consumption, for example, are expressed by specific data patterns. They are domain-independent by nature. The same holds for word patterns that turn a statement into a positive or negative comment. Finally, the data

scientist requires managerial knowledge to transform this knowledge into a cooperative discovery strategy. Managing the discovery lifecycle includes the management of analytics teams, too. Decentralized teams can be assigned to specific analysis tasks. At the same time, they can be moved closer to domain experts or business functions, for instance, to have them better meet their specific needs. These three knowledge areas constitute the proficiency we expect from data scientists.

The data scientist is more an information conceptualist who concentrates on definition, integration, and sharing of information concepts and less on analytic modelling. Even though information discovery requires a bit of everything, from coding and querying databases over statistical modelling to data mining, the role of the data scientist starts beyond that, it targets storytelling about discovery that may feature episodes from all these fields. These stories are about how we get insights from data, where to discover those data, and how to merge them into suitable visualizations.

The data scientists' mental models (Norman, 1987) on data patterns reflect information concepts. Usually there is a variety of data patterns that map onto a single concept. Discovery starts with a hypothesis on the concepts that meet the need of the information consumer. Data scientists sketch information concepts by blueprints that guide the discovery process. Discovery engines then execute these blueprints, that is, they locate the data that map to each of the required information concepts and extract and annotate these data. Discovery blueprints are far from being programming instructions. They are pattern descriptions, that is, schemas of how the required information may look like in data. For example, a blueprint may address the different facets of the mentioning of the most recent revenue of a certain company in tweets, newspaper articles, and the like. It may also describe the mapping of geospatial locations to possible descriptions in crime reports. After distilling—locating, extracting, and standardizing—the data they can be fed into statistical tools, data mining or text mining tools. A suitable visualization of the results from analysis finally provides the insights the decision maker—or the information consumer in general—expects.

Data scientists barely produce blueprints alone and in one step. It is an experimenting and iterative process. They initiate it by a certain belief—predisposition or bias reinforced by years of experience—and gradually refine this belief through cycles of gathering of instances of information concepts and checking them against their hypotheses. The blueprints describe the sources of information (news channel, location sensor, email, etc.) and the way, how data related to the different concepts must be extracted. We can call this process "operationalization". The data scientist designs one or more patterns that map data to a particular information concept and applies them to different data sources. Each of these tests checks the coverage of each pattern, that is, the potential of the pattern to detect all instances of the information concept. After sufficient iterations, data scientists operationalize their blueprints in their individual environment and discuss them with their colleagues. The cooperative approach to this process is necessary in order to incorporate reflections of other data scientists and stakeholders about this particular information discovery (Elbeshausen et al., 2014). After further reflections on workgroup or corporate

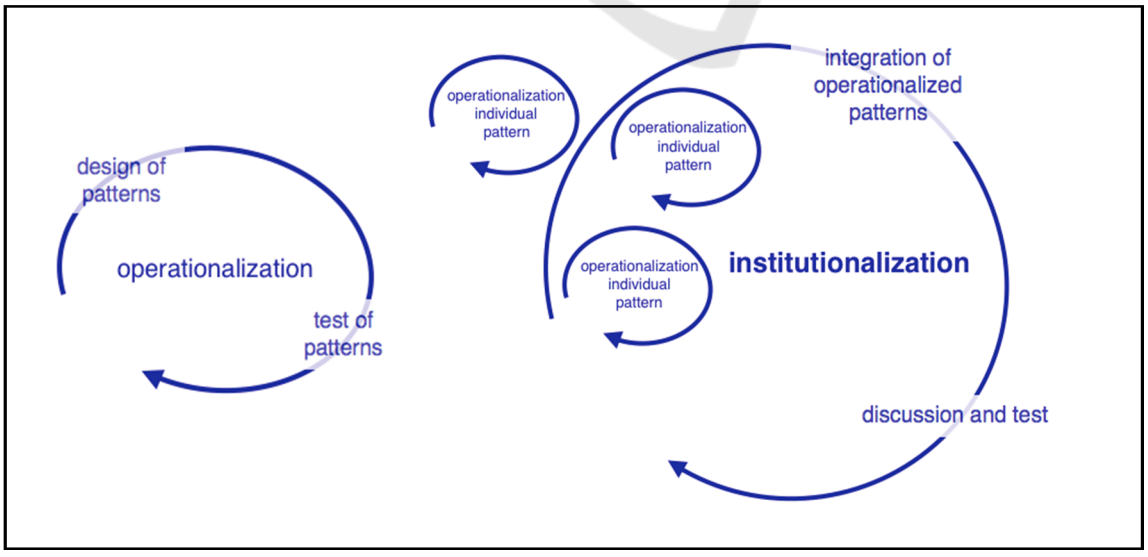

Figure 1: Schema of operationalization and institutionalization of discovery blueprints.

level they institutionalize their blueprints. It is obvious that each pattern may also be a composition of existing patterns.

From a different angle, we can say the data scientists are in charge with the corporate's discovery governance, which means, with the design, implementation, management, control, and improvement of information discovery. Discovery governance rests thus on the following phases that are managed by the data scientists (cf figure 1):

- Conceptualization: the data scientists design and manage the information concepts in a corporate thesaurus.
- Operationalization: they develop and maintain discovery blueprints that serve mapping of data onto information concepts.
- Institutionalization: they organize sharing and discussing of blueprints as a corporate commodity.
- Transformation: they organize the application of analytical tools/models to produce new information.
- Visualization: data scientists are also responsible to ensure that information consumers understand the findings, understand the "message" in the results. Proper visualization of key performance indicators, for instance, may faster impart the performance of a marketing campaign than mere numerical data. In general, proper visualizations rise the information usability (Rosenbaum and Ramey, 2014).

In general, the process of discovery passes iteratively through some or all of these phases. At the end of each phase, results are carefully checked for plausibility and usefulness. Each phase is thus completed by a quality gate where data scientists evaluate the quality of the results achieved and approve or reject the respective phase activities.

5 DISCOVERY LIFECYCLE IN TEXT MINING

There is no general approach to information discovery. It is highly individual, theme-specific, and volatile. The data scientist may delegate some subprocesses to machines, but there is no reason to think that the machine's algorithm could discover all the ingredients needed to develop a clear picture on the facts the information consumer expect. It is an illusion to believe that discovery as a whole or even to a large extent can 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 that appear in quite a standardized form like stock quotes from a ticker service or weather data from a weather service. It works fine when discovery tasks are well-defined and the data patterns to look after are highly recurring over thematic domains.

Just under half of respondents expressed solidarity

At 26.9 percent and around 190 billion euros of subscribed capital, Germany has by far the largest share of guarantees for the stabilization of crisis countries in the euro area.[5] Thus, the attitude of the German people to fiscal solidarity (see Box 1) can be seen as a special test case to show how much support there is for a European community of solidarity among the citizens of Europe—particularly in the current donor countries. Public opinion in Germany plays a key role here. Figure 1 shows that, in 2015, 48 percent of all respondents were in agreement with Germany providing financial aid to EU countries in crisis while 31 percent opposed the move and 21 percent had no firm opinion on the matter. Those in support of providing aid are clearly the largest group but they still do not constitute a majority among all respondents.

Figure 2: Two text snippets showing the stance of German people on fiscal solidarity (source: Langfeld, Kroh, 2016).

In practical data science, text mining has quite a prominent position (Bedathur et al., 2010; Fan et al., 2006). It starts with extracting statements from texts that follow a certain pattern of appearance. The statements are then normalized, that is, they are transformed into a standardized form of representation or are annotated by terms reflecting their content in a standardized form. The combination of factual (numerical) data with text data has its particular appeal. A statistical analysis may come to a certain findings. Text mining can help, in parallel, to find statements in articles, news, or Twitter messages that underpin or refute these findings. Numerical analysis, for example, may observe a certain stock by applying time series analysis to measure the probability that its value will rise or drop. Accompanying text analysis sifts through texts and looks for signals that indicate whether this stock is about to take off or drop in value. Identifying these signals and merging

them with the numerical analysis rest on quite an array of discovery tasks.

Discovering and comparing opinions on a particular issue is an interesting topic nowadays. Opinion Mining (or sentiment analysis) tries to identify people's stance in favour or against politics, products, persons, etc. (Feldman, 2013; Wright, 2009; Turney, 2002). It also tries to find facts (features) that underpin these opinions. Furthermore, fake news on social media seem to become a growing threat. Even based on false assumptions they can have an impact on public perception. In particular during the most recent election campaign in USA, much fake news appeared in social media (for example: Rosenberg, 2016).

Opinion mining (Feldman, 2013; Evangelopoulos, Visinescu, 2012) is a good example to highlight the data scientists' role in managing the semantic dimension in discovery. They have to organize the development (or reuse) of discovery blueprints that guide the discovery of the ingredients of a particular opinion. The result takes the form of a semantic skeleton that guides the discovery engine. Opinion mining requires discovery tasks but also integration tasks. Both are relevant for listing in the data scientists' job description. They detect, for instance, different facets of an opinion in a text, the relationship to a particular opinion holder and the overall theme of the opinion. We may take an opinion like "We expect that in about five years, we will see about 50 percent of the refugees in employment." and relate it to the theme "Refugee Crisis in Germany" and the opinion holder "Jürgen Schupp", a scientist at the German Institute for Economic Research (example taken from Wittenberg, 2016). If we see the attributes of this opinion in relationship with related topics, such as education level, language skills, etc. we can set up a heat map highlighting the essential concepts around the theme "employment of refugees". The map helps follow this opinion and all its thematic ramifications. The following design of blueprints relates to the example shown in figure 2. The opinion about solidarity with crisis countries is reflected in objective sentences containing factual information. To discover the fact whether a majority or minority is favouring solidarity we have to look after clusters of facts that are tightly related. We can discern the corresponding concepts for positive ("agreement"), negative ("opposition"), and neutral stance ("no opinion") over solidarity with crisis countries. We also identify the respective quantifications together with the variations of the key expression "fiscal solidarity". The second snippet uses in addition the concepts "increase" and "decrease" to express change in people's opinion over time.

Discovering the facts is the result of a number of text analysis processes:

- Identification of sentiment expressions (indicated in red in figure 2).
- Identification of factual (theme independent) information (indicated in blue)
- Identification of aspects or features (indicated in green)
- Extraction of the facts

The examples show that text analysis can get quite individual, hardly covered by mainstream discovery. Users may have to analyze, for instance, from time to time dozens of failure descriptions or complaints, for instance. These appear in many different facets over a variety of sources. Serving small-scale requests would mean permanent system adaptation, which is too intricate and prohibitively expensive in most of these cases. The same holds for data science as a service (DSaaS). They offer only standard discovery and analytical processes that cannot cover the broad variety of discovery required by information consumers.

6 PROPOSAL FOR A CURRICULUM IN DATA SCIENCE

Industry and Science is demanding a new profession that designs, orchestrates, and controls statistical and computational skills and tools within an organization to carry out information discovery on a large scale. Education in data science shall provide technical and managerial knowledge and skills for this profession. Inclined to the current layouts of curricula in data science and considering the discussion of this article we propose the following framework for a curriculum of data science. It is structured along six knowledge areas.

The broad field of methods and tools from AI supporting information discovery constitutes an integral part of data science. However, is it really necessary that the data scientists are expert in AI, down to all its mathematical and statistical foundations? Doesn't it suffice if they roughly know features and capabilities of AI tools and let the AI practitioners handle the particular problems that can be solved with these tools? Is it even necessary that data scientists are experts in Hadoop, R and the like, if there are correspondingly trained programmers available? We may not forget that building data mining or predictive models for AI tools is time-consuming and quite difficult. It usually involves testing and verifying different model hypotheses with a large array of variables, before the

analytic results approve the underlying overall hypothesis. The degree of difficulty ratchets up further the more complex the hypothesis becomes. Furthermore, the findings need to be validated carefully. In the end, the reputation of the whole discovery team is at stake if the discovery project goes awry. If results are approved too early it can damage the team's credibility. The failure of the Google Flu Trends project is a prominent example in this context (Lohr, 2014). It takes experienced analysts to develop and verify the core of complex analysis models. Knowledge and skills required for predictive analysis, deep learning and data mining are part of the formation in artificial intelligence or domain-specific fields like economics or statistics. It may be reasonable to employ data scientists if the analysis models are less complex. In general, however, we expect the data scientists to scrutinize model assumptions and to detect imperfections of predictive models. They can support the discovery team to focus on transparency. The data scientist need to understand how an analysis model came to its conclusions. The degree of uncertainty needs to be communicated, too. Even if the team supposes 90% certainty, there are chances the outcome will turn out to be false.

We think the main focus of the data scientists should be more on managerial capabilities concerning the information discovery lifecycle.

They should know

- where the required data come from,
- how they need to be transformed or standardized,
- how they must be semantically integrated,
- which kind of analytical model can be applied to them, and
- how they are best visualized to the information consumers.

We know that we cannot draw a clear line between the mathematical, statistical, and computational fundamentals a data scientist should know and the more overarching managerial capabilities to orchestrate discovery processes. There may be situations when the data scientist needs to assist in the design of special analytic tools. However, this will be more the exception than the rule.

Based on this profession profile of the data scientist, we propose that education in data science should address the following formation areas:

- Information Extraction and Standardization. Information-related data must be identified, in particular in unstructured data. The data scientist should know what forms of appearance the required information may take.
- Data Mining. This is the traditional and most widely adopted area for data science. It may be sufficient if the data scientist knows the nature of the analytic models, the required input, and the results they produce.
- Text Mining. It is currently underrepresented in education related to data science. This is quite strange, because the vast majority of Big Data are text data. Furthermore, it is key for text analysis like Sentiment Analysis, Opinion Mining, and Fake News Detection, just to name a few.
- Management of the Information Discovery Lifecycle. It starts with analyzing the needs of the information consumers. From their requirements the data scientist designs, guides, and controls the discovery process. This includes the composition of the discovery team to ensure that discovery is based on the necessary domain and technical knowledge.
- Quality and Ethics in Data Science. Privacy issues around big data will only become more prominent in the years ahead. More and more we adopt tools, like devices for connected home, driving assistants, and fitness or health trackers. They produce vast troves of interesting data about us that could theoretically be considered useful by criminal investigators, health care institutions, or national security agencies. Therefore, we expect the data scientists to take care of privacy concerns. We also expect them to check the assumptions of statistical methods and tools applied in analysis and to check the plausibility of their results. Through the semantics of all of the context of the work that is underway, to study and to fully respect the individual's role as well as the aggregate and collective roles.
- Information Visualization helps to impart analysis results to information consumers. In general, it shall ensure information usability. Consumers expect that results are presented in a way that helps them to understand them and to translate them into suitable actions and decisions.

7 CONCLUSION

There is no gold standard for the job profile of data scientists. In general, we expect them to manage the

discovery of information enabling us valuable insights from Big Data. AI claims to be in the position to provide information consumers with useful insights. There is no doubt, data mining, machine learning, deep learning, etc. are essential in data science. Their methods and tools produce a lot of useful data, which, in turn, lead to the expected insights. However, focusing exclusively on AI would restrict us to insights that we may gain from the results of AI tools. The capabilities of the available AI tools determine the range of insights.

By viewing data science as the discipline of information discovery performed along the requirements of information consumers and managed by data scientists we bring humans into the loop of technology design and use. The user-centred approach shifts the focus from a purely data-driven approach towards a problem-driven approach. We see practical data science rather as the result of a cooperative effort of discovery team representing domain and technical knowledge and managed by the data scientist. In this article we sketched the profession of the data scientist from this perspective that conceives the data scientists rather as a master of the information discovery lifecycle than data mining expert or the like.

Data scientists do not construct models for all kinds of analysis tools and reasoning systems but should clearly be aware of the information those systems 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. We can expect that she or he assumes ethical responsibility and makes sure that information provided is reliable. The "everything is possible" attitude that shines through in many Big Data discussions is dangerous and rises false expectations. The polls during the recent elections in USA showed that sometimes even the whole statistical and data mining machinery of an entire country can fail quite embarrassingly. Data Scientists need to develop a sound sensation of plausibility that helps them to rise doubts and to prompt a closer look when the results of data analysis seem too questionable to them.

Finally, we advocate for the education in data science a stronger focus on text mining, information visualization, ethics aspects rised by Big Data, and the management of information discovery.

REFERENCES

Bedathur, S., Berberich, K., Dittrich, J., Mamoulis, N., Weikum, G., 2010. Interesting-phrase mining for ad-hoc text analytics. In: *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, 1348-1357.

Cowie, J., Lehnert, W., 1996. Information Extraction. *Communications of the ACM*, vol. 39, no. 1, 80-91.

T.H. Davenport, T. H., Patil, D.J. 2010. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, vol. 90, no. 10, pp. 70-76.

Elbeshausen, S., Womser-Hacker, C., Mandl, T., 2014. Searcher heterogeneity in collaborative information seeking within the context of work tasks. In: *Proceedings of the 5th Information Interaction in Context Symposium (IIiX)*, 327-329.

Englmeier, K., Murtagh, F., 2016. Interaction for Information Discovery Empowering Information Consumers. In: S. Yamamoto (ed.): *Human Interface and the Management of Information: Information, Design and Interaction*. Volume 9734, Lecture Notes in Computer Science, 252-262.

Evangelopoulos, N, Visinescu, L., 2012. Text-Mining the Voice of the People. *Communications of the ACM*, vol. 55, no. 2, 62-29.

Fan, W., Wallace, L., Rich, S., Zhang, Z., 2006. Tapping the power of text mining. *Communications of the ACM*, vol. 49, no. 9, 76-82.

Feldman, R., 2013. Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, vol. 56 no. 4, 82-89.

Gudivada, V. N., Baeza-Yates, R., Raghavan, V.V. 2015. Big data: Promises and problems. *IEEE Computer*, vol. 48, no. 3, pp. 20-23.

Langfeld, H., Kroh, M., 2016. Solidarity with EU countries in crisis: results of a 2015 Socio-Economic Panel (SOEP) survey. *DIW Economic Bulletin*, no. 39, September 30, 2016, 473-479.

Lohr, S., 2014. Google Flu Trends: The Limits of Big Data. In: *The New York Times,* March 24, 2014.

McCallum, A., 2005. Information Extraction: Distilling Structured Data from Unstructured Text. *ACM Queue - Social Computing*. vol. 3, no. 9, 48-57.

Norman, D., 1987. Some observations on mental models. In: D. Gentner; A. Stevens, (Eds.) *Mental Models*, Lawrence Erlbaum, Hillsdale, NJ.

Rosenberg, E., 2016. Fake New York Times Article Claims Elizabeth Warren Endorsed Bernie Sanders. *The New York Times*, March 1, 2016.

Rosenbaum, S. and Ramey, J., 2014. Current Issues in Assessing and Improving Information Usability. In: *Proceedings of the CHI'14, Extended Abstracts*, 1119-1122.

Turney, P., 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the Association for Computational Linguistics*. pp. 417–424.

Wittenberg, E., 2016. Eight Questions for Jürgen Schupp, Refugees have a strong educational orientation, *DIW Economic Bulletin,* no 48/2016, December 6, 2016, 557-558.

Wright, A., 2009. Our Sentiments, exactly, *Communications of the ACM*, vol. 52 no. 4, 14-15 (2009)