Discussion paper

# Data Science for Institutional and Organizational Economics

By Jens Prüfer and Patricia Prüfer

March, 2018

TILEC Discussion Paper No. 2018-011
CentER Discussion Paper No. 2018-016

ISSN 2213-9532
ISSN 2213-9419
http://ssrn.com/abstract=3137014

Understanding Society

8 March 2018

# Data Science for Institutional and Organizational Economics[1]

Jens Prüfer and Patricia Prüfer[2]

***Abstract***

*To which extent can data science methods – such as machine learning, text analysis, or sentiment analysis – push the research frontier in the social sciences? This essay briefly describes the most prominent data science techniques that lend themselves to analyses of institutional and organizational governance structures. We elaborate on several examples applying data science to analyze legal, political, and social institutions and sketch how specific data science techniques can be used to study important research questions that could not (to the same extent) be studied without these techniques. We conclude by comparing the main strengths and limitations of computational social science with traditional empirical research methods and its relation to theory.*

*Keywords: Data science, Machine learning, Institutions, Text analysis*

*JEL codes: C50, C53, C87, D02, K0*

## 1. Introduction

Is it possible to elicit the ideological positions of voters by having information about the structure of their social media networks? Can we make predictions about upcoming policy changes by analyzing the speeches of statesmen, although they could be considered cheap talk? Is there a way to identify individual biases of judges by analyzing all court decisions within a jurisdiction, and thereby suggest ways to make the legal system more impartial? Can we develop a reliable index of organized crime and subversion in industrial areas, typical hotbeds of such crimes, by using a wide range of Internet, social media and administrative data sources?

All these questions are within the domain of Institutional and Organizational Economics (IOE). And all of them could not be seriously studied, let alone answered, by traditional empirical methods. But today, our world is full of devices and applications that rapidly generate, store, and transmit huge amounts of information. Enormous quantities of data have become available for all kinds of analyses: sensor data in self-driving cars, "smart" homes, and office equipment, social media data, mobile data, data on browsing behavior, or digital camera images. These fast growing volumes and varieties of available data, paired with cheaper and more powerful computational processing and

---

[1] Forthcoming in Claude Menard and Mary Shirley (eds.), *A Research Agenda for New Institutional Economics*, Edward Elgar Publishing, forthcoming.

[2] J. Prüfer: Department of Economics, CentER, TILEC, Tilburg University; j.prufer@uvt.nl. P. Prüfer: CentERdata, Tilburg University; p.prufer@uvt.nl. Both: P.O. Box 90153, 5000 LE Tilburg, The Netherlands. We are grateful to Jan Boone, Marcel Das, Eleonora Freddi, Peter Fontein, Freek van Gils, Madina Kurmangaliyeva, Wieland Müller, Louis Raes, and the editors, who provided valuable feedback on an earlier draft of this essay. Research assistance of Pradeep Kumar is highly appreciated. All errors are our own.

affordable data storage, have constituted the "rise of big data" or "datafication" (Mayer-Schönberger and Cukier, 2013). With further developments in artificial intelligence, datafication will accelerate. It is already having disruptive effects on the social sciences, such as economics (Einav & Levin, 2014a) and management (George, Haas, & Pentland, 2014). It has created the new field of *computational social science* with its capacity to collect and analyze data at a scale that may reveal new patterns of individual and group behavior and allow scholars to model economic and social interactions more precisely (Lazer et al, 2009).

These developments have been noticed by many organizations, which are increasingly inclined to do more data driven and fact based work. In parallel, data and technology ubiquity is likely to have direct effects on individual and group behavior, the nature of social and organizational structures---and thereby on institutions. Thus, the very object of social science research may change, not only empirically but theoretically.

From the research perspective of IOE, datafication has led to strongly increased interest in two areas: (i) the consequences of big data for markets, polities, jurisdictions, and societies (e.g. Prüfer, 2013, 2017, Prüfer and Schottmüller, 2017; Sobbrio, 2018); (ii) the better understanding of data science techniques to improve our own research. While both topics deserve deep scholarly attention, this essay focuses on item (ii) and aims to outline to which extent data science techniques can push the research frontier in IOE further.

Section 2 briefly describes the most prominent data science methods that lend themselves to analyses of the governance structures of institutions and organizations. Section 3 elaborates on several examples using data science to analyze legal, political, and social institutions and sketches how specific data science techniques can be used to study important research questions that could not (to the same extent) be studied without these techniques. Section 4 concludes by comparing the main strengths and limitations of computational social science with traditional empirical research methods and its relation to theory. The Appendix provides links to literature and Internet resources and to the most relevant text mining tools and download sources, showing how to get started with data science methods independently.

## 2. Data Science Methods

### Background

Data science is a relatively new discipline that analyzes and interprets large amounts of complex and unstructured data. This discipline covers more than the conventional use of data and statistics. It relies heavily on computational power and computer science. In conventional statistical research, hypotheses are stated and tested with the help of data, assuming that the data are generated by a given stochastic data model. Data science, by contrast, "not only provides new tools, it solves a different problem" (Mullainathan and Spiess, 2017, p.88). It uses algorithmic models and treats the data mechanism as unknown thus discovering complex structures that were not specified in advance

(Breiman, 2001). In a nutshell, whereas conventional statistics is deductive, data science is inductive; the approaches are complementary.

The foundation of data science techniques is the use of pattern recognition in complex data. By making software autonomous or using iterative feedback to discover associations in data, generalizable patterns are found and anomalies are detected. Where the human brain can associate two or three dimensions of information with each other, data science methods allow us to relate hundreds of dimensions, thereby obtaining much more fine-grained associations. This leads to a system for finding clusters and classifications to extract meaningful information from the data, which can then be transformed into an understandable structure for further use.

Data science techniques allow us to leverage the exponentially growing data never seen before, so-called ‘big data’. The term big data relates to the following features: the volume (amount) of data that is being produced every second; the velocity (speed) with which new data are generated, collected, and analyzed; the variety (types) of data that are being used; the veracity of data, thus their quality and trustworthiness; and their complexity. These features cause big data to require special ways of storage, accessibility, and processing. Analyses are often done by using multiple computers and multiple calculation units, so-called 'high-performance computing', such as Hadoop clusters and Spark-Streaming, or parallel virtual environments.

Data science combines a multitude of different tools and techniques: we highlight the most interesting ones for research in IOE. Usually the basic steps for analysis include writing an algorithm, setting up an automated process (script), and linking it with open data protocols and Application Programming Interfaces (APIs). Collecting large amounts of unstructured information often provides a complex set of information. With the help of visualization techniques, such as chord charts and network graphs, we can observe clusters within that information and present results of data analyses.

Notably, data science techniques, such as data & text mining and machine learning, can be used not only for the aforementioned big data sources, but also for ‘traditional’ data sources such as surveys and large administrative data sets. We can also apply the techniques described below to ‘old data’ in order to find new patterns and insights, for example, by analyzing answers to surveys that were conducted to assess the level of democracy or the quality of institutions. The computational power of these techniques allows for a much broader and varied search on existing (administrative) data bases, which may lead to the revelation of new patterns even in traditional data sources. A notable example is the work of Raj Chetty and coauthors, which makes use of the large IRS data base with billions of observations on U.S. tax payers. In combination with other data sets, for example, mortality information from the Social Security Administration or test scores and teacher assignments from a large urban school district, these researchers disentangle the relationship between life expectancy and earnings, why some people cheat on their taxes, or the long-term effects of better teachers (Chetty et al., 2013, Chetty et al., 2014, Cutler et al., 2016).

## Key data science methods for IOE

About 80% of big data is available in unstructured text form, for example in blogs, websites, and social media (Cogburn and Hine, 2017). There are various useful tools and techniques to extract valuable information from these sources. Here, we describe how and what we can infer from the data, and discuss useful techniques for mining and analyzing text data. Our emphasis is on statistical and machine learning approaches that can be generally applied to arbitrary text data in any natural language with minimum human effort. In general, *Python* has very effective and useful open source libraries for these tasks.[3]

### Machine Learning

Machine learning algorithms are a key ingredient of data science techniques. *Artificial intelligence* (AI) and *machine learning* (ML) are sometimes confused. In fact, AI is a much broader concept in which machines mimic cognitive functions of learning and problem solving and are, therefore, able to carry out tasks in a way that we would consider "smart," that is, with human-like cognitive functions (OECD, 2017). ML, on the other hand, "is the field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959). The most important categories of ML are *supervised learning* and *unsupervised learning*.[4]

Technically, in a *supervised learning* task a computer learns a relation between some observed input (usually a vector of many predictors) and some desired output (one outcome variable) (Hastie et al., 2009). A supervised learning algorithm analyzes the labeled training data and produces an inferred function, which can be used for mapping of novel (test) data. *Classification* and *regression* are two categories of supervised learning. Regression techniques are used to predict continuous values, while classification techniques are used in discrete settings to identify to which set of categories (classes) a new observation belongs. Supervised learning can help to predict unseen patterns, but also to understand which input best predicts the outcome and/or to assess the quality of previously tested predictions/inferences. It can, therefore, also be used as a process of statistical dimensionality reduction. For example, supervised learning can predict legal court decisions. In a pioneering example, Martin et al. (2004) tested the performance of experts in legal prediction – law professors and attorneys – against a statistical model that had analyzed data about hundreds of past U.S. Supreme Court cases. In this early case, the ML algorithm significantly outperformed the experts in predictive ability: it correctly forecasted 75% of Supreme Court outcomes, while the experts only had a 59% success rate.

*Unsupervised learning* is a learning algorithm drawing inferences from unlabeled input data sets. The most common unsupervised learning method is *cluster analysis*. It is mostly used for

[3] A library is collection of pre-written programs, scripts, or functions which can be loaded on disk for immediate use. With the help of these libraries, one can implement (complex) algorithms by writing few lines of codes.

[4] Another category is *reinforcement learning* where training data given as feedback to the algorithm is in the form of rewards and punishments. *Deep learning* is a special class of supervised learning that is frequently used for feature extraction from complex, multidimensional data such as images or visuals. For instance, Google uses it to automatically suggest the next word(s) of a search term when one has started typing.

exploratory data analysis to detect hidden patterns or grouping in data. The clusters are modeled using a measure of similarity, which is defined by distance metrics. For instance, unsupervised learning can be used to group customers into segments, based on their characteristics, to improve and customize services.[5]

### Text analytics and web data scraping

A frequently used source of big data is scraping information from the Internet, for which Python is effective, for instance. But all other data sources that relate to natural language can be used for data science methods, too, such as open answers, text files, sections, reports, or e-mails. Internet log files and the metadata of search engines can provide interesting information about trends over time. Search engines register when and where a search query has been performed.[6] The numbers of searches on certain topics and the order of search results often show interesting patterns, which are used, for instance, by *Google Trends*. In a series of articles in *The New York Times*, Seth Stephens-Davidowitz used Google Trends to measure racism in different parts of the United States by using the search volume for the word "nigger(s)." Correlating this measure with Obama's vote share, he calculated that Obama lost about 4 percentage points due to racial animus in the 2008 presidential election.[7] In his recent book, he provides many more examples on how to use Google data, especially for sensitive issues such as sexual orientation, sexism, female teenage obesity, and other stereotypes (Stephenson-Davidowitz, 2017).

Mining, clustering, and analyzing these unstructured data sources requires the use of analytical techniques for natural language, for instance, *sentiment analysis*, which can extract subjective information from language. This so-called *Natural Language Processing* (NLP) can be performed in different languages, for example Python or R, for which well-established packages and toolboxes are available. *Box 1* in the Appendix lists the most common techniques and tools.

## 3. Data Science for Legal, Political, and Social Institutions

Here, we briefly highlight some recent papers in which data science methods have been applied to IOE research questions. For more policy relevant examples and further applications, see Einav and Levin (2014b) or Varian (2014).

The role of online and social media communication for the development of political opinions, their influence on elections, and even on the stability of democracy has received a lot of attention, especially since the 2016 U.S. presidential elections (Tucker et al., 2017). Questions about echo chambers, political ideologies, and online hate speech have been explored using Twitter data. For example, Barberá (2015) conducts an ideal point estimation of Twitter users' ideology and is able to estimate the ideological location of both elite and mass Twitter users in the U.S. and in five

[5] For an overview, see http://ieeexplore.ieee.org/document/7516129/.
[6] Economic consequences of this usage are studied in Prüfer and Schottmüller (2017).
[7] Stephens-Davidowitz, S., "How Racist Are We? Ask Google", The New York Times, June 9, 2012.

European countries. The estimated positions of legislators and political parties replicate conventional measures of ideology. The method is also able to successfully classify individuals who state their political preferences publicly and a sample of users matched with their party registration records. With this method, it is possible to estimate ideology for more actors than any existing alternative, at any point in time, and across many polities.

Ceron (2015) combines data from a Eurobarometer survey held in November 2012 across 27 European countries with a supervised sentiment analysis of online political information, based on data scraped from the Internet. This method allows the author to track differences between mediated and unmediated sources of information in the political debate (news media vs. social media), highlight the potential shortcoming of e-democracy, and analyze the relationship between Internet usage and political trust. It finds that consumption of news from information/news websites is positively associated with higher trust, while access to information available on social media is linked with lower trust.

Emotions captured by social media data can also be used for insights regarding the (temporal) happiness within a society, and also to measure social tensions and unrest.[8] Since 1973, Statistics Netherlands has compiled an annual *Safety Monitor*, which surveys a representative group of citizens about their feelings of security. Although terrorist attacks, a major driver of feelings of security, influence these perceptions, the social impact of such events cannot be directly derived from the results of the Safety Monitor. Therefore, the Center for Big Data Statistics at Statistics Netherlands is experimenting with a so-called *Social Tension Indicator* (STI), for which Twitter data have been streamed via an API since 2010.[9] The STI measures tensions or unrest within Dutch society, based on a validated glossary of words related to security. By using this type of fine-grained and real-time information on feelings of security, the STI is better suited to measure actually observed and recorded feelings than a survey conducted at most once a year.

Rios-Morales et al. (2009) use a combination of unsupervised and supervised learning to measure the relevance and usefulness of 'good governance' indicators to assess the political risk in a country. They integrate the *Good Governance Indicators* of the World Bank[10] with political instability classifications from the *Political Instability Task Force* (PITF) to create a rich data set from which countries can be clustered according to their political risk. This approach shows that indicators such as 'Rule of Law' and 'Control of Corruption' are the most important characteristics of stable countries, while the level of 'Political Stability' and 'Voice and Accountability' are the most important variables for countries at high levels of political risk.

[8] In 2008, an instrument that measures the happiness of the U.S. population in real-time based on Twitter data was launched: http://hedonometer.org/index.html. In the Netherlands, CentERdata started to perform sentiment analyses on tweets in 2017. A real-time measure of (temporal) average happiness in the Netherlands can be found at: https://www.centerdata.nl/en/projects-by-centerdata/temporal-happiness-score-of-dutch-tweets.

[9] https://www.cbs.nl/en-gb/our-services/innovation/project/social-tension-indicator-based-on-social-media.

[10] www.worldbank.org/wbi/governance/data.

By applying data science methods to a broad variety of data sources, we can analyze multifaceted topics on governance and (informal) institutions. In an ongoing project, Prüfer and Kumar (2018) attempt to predict indicators of organized crime and subversion. Based on a wide range of Internet, social media and administrative data sources and a broad combination of text analytics and machine learning algorithms, that paper disentangles patterns of organized crime and subversion in Dutch industrial areas. Based on these patterns and the best predictors, they develop a method with which policy makers can judge the level of criminal and subversive activities taking place in a certain area.

Finally, an exciting feature of data science techniques and big data is their potential for mass collaboration. Under the header, "What would happen if hundreds of social scientists and data scientists worked together on a scientific challenge to improve the lives of disadvantaged children in the United States?," Princeton University and Columbia University established the *Fragile Families Challenge*. This ongoing mass research collaboration uses big data (some 54 million data points) collected as part of Princeton University's Fragile Families and Child Wellbeing Study, studying a cohort of about 5,000 children born in large U.S. cities.[11]

## 4. Conclusion

Data science methods are able to support research on institutions and organizations by facilitating the automated collection of information, especially, but not only, on the Internet. Via text and sentiment analysis, computers can learn to "understand" the meaning of words, relate them to each other, and analyze them at scales that otherwise would require hordes of research assistants. The new techniques and technologies also allow us to use many more (unstructured) real-time data sources to conduct analyses that would not have been possible otherwise, for instance by using sensor data from mobile devices to measure economic activity (Blumenstock et al., 2015). Given that these methods are usually available for free and relatively easy to learn, data science techniques thereby contribute to a democratization of empirical research tools, where scholars or students with less resources have a higher chance to compete with established researchers from resource-rich countries.

However, given the current state of data science methods, they cannot completely substitute for human creativity and research design skills in the social sciences.[12] Regarding applied research with a policy angle, they appear to be especially well suited for inductive analyses to guide further research efforts, for instance by pointing researchers at relevant correlations and helping them to design better (field) experiments, to make better comparisons between more precise populations of interest, and to reveal behavior that was difficult to detect previously (Monroe et al., 2015).

Combining big data and machine learning with administrative and survey data also looks like a very fruitful avenue for further research. Data science techniques have been largely applied to Internet data (often by scraping and analyzing big social media data sets). This approach ignores both

[11] http://www.fragilefamilieschallenge.org/.

[12] This may change once an *Artificial General Intelligence* is developed - which is expected to take 10-100 years, according to OECD (2017).

potential selection effects due to differences between online (social media) users and the entire population, and measurement errors due to the unreliability of social media data as representative measures of social phenomena. Comparing the results of a (small) representative survey with results of (big) unrepresentative data, where the representativeness can even be assessed empirically, looks like an ideal way forward for empirical research.[13]

A general problem that grows with the power of data science methods and big data availability are privacy and confidentiality concerns (Acquisti et al., 2016; Dengler and Prüfer, 2017). As a direct policy response to this technological progress, the *General Data Protection Regulation* (GDPR) will become effective in the EU in May 2018, regulating legal use of privacy-sensitive data, especially those relating to Internet services.[14] However, since many data on the Internet are publicly available and voluntarily shared by data subjects, the risk of conflicting with the GDPR is reduced, yet not eliminated.

Notably, data science methods do not substitute for theoretical research or conventional statistics. The inductive, data-driven approach allows us to learn from actual observed behavior at unprecedented scale, pointing theorists at the key variables of interest for a specific question that deserve to be modeled. This may alleviate the need for expert interviews or the use of small, unrepresentative surveys to obtain a first understanding of the main factors influencing certain industries, jurisdictions, or the like. It may also reduce the risk that theorists fall victim to confirmation bias (Mahmoodi et al., 2017), which is especially a problem in closed research communities clustered around a narrow set of questions, where publication success depends on positive reviews by those peers who advanced their careers by studying the questions in a specific way. Dreaming ahead, the best applied theoretical researchers may regularly motivate their choice of key model variables by using the results of big data analyses.

## References

Acquisti, A., Taylor, C. and Wagman, L. (2016), The Economics of Privacy, *Journal of Economic Literature* 54(2): 442-92.

Barberá, P. (2015), Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data, *Political Analysis*, 23(1), 76-91.

Blumenstock, J., Cadamuro, G. and On, R. (2015), Predicting poverty and wealth from mobile phone metadata, *Science*, 350, 1073-1076.

Bishop, C. (2011), *Pattern Recognition and Machine Learning*, Springer, New York.

[13] Varian (2014, p.23) comments: "A good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard."

[14] Regulation (EU) 2016/679 of the European Parliament and of the European Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (http://ec.europa.eu/justice/data-protection/).

Breiman, L. (2001), Statistical Modeling: The Two Cultures, *Statistical Science*, 16(3), 199-231.

Ceron, A. (2015), Internet, News, and Political Trust: The Difference between Social Media and Online Media Outlets, *Journal of Computer-Mediated Communication* 20, 487-503.

Chetty, R., Friedman, J. and Saez E. (2013), Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings, *American Economic Review*, 103(7), 2683-2721.

Chetty, R., Friedman, J. and Rockoff, J. (2014), Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates, *American Economic Review*, 104(9), 2593-2632.

Cogburn, D. and Hine, M. (2017), Introduction to Text Mining in Big Data Analytics, *Proceedings of the 50th Hawaii International Conference on System Sciences*, HICSS 2017.

Cutler, D., Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N. and Bergeron, A. (2016), The Association Between Income and Life Expectancy in the United States, 2001-2014, *Journal of the American Medical Association*, 315(16), 1750-1766.

Dengler, S. and Prüfer, J. (2017), Consumers' Privacy Choices in the Era of Big Data, mimeo, Tilburg University.

Einav, L. and Levin, J. (2014a), Economics in the age of big data, *Science*, 346(6210), 1243089.

Einav, L. and Levin, J. (2014b), The Data Revolution and Economic Analysis, in: *Innovation Policy and the Economy*, Vol. 14, ed. by Lerner, J., Stern, S., University of Chicago Press, 1-24.

George, G., Haas, M. and A. Pentland (2014), Big data and management, *Academy of Management Journal*, 57(2), 321-32.

Hastie, T., Tibshirani, R. and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and M. Van Alstyne (2009), Computational Social Science, *Science*, 323(5915), 721-723.

Mahmoodi, J., Leckelt, M., Van Zalk, M., Geukes, K. and Black, M. (2017), Big Data approaches in social and behavioral science: four key trade-offs and a call for integration, *Current Opinion in Behavioral Sciences*, 18, 57–62.

Martin, A., Quinn, K., Ruger, T. and Kim, P. (2004), Competing Approaches to Predicting Supreme Court Decision Making, *Perspectives on Politics*, 2(4), 761-767.

Mayer-Schönberger, V. and Cukier, K. (2013), The Rise of Big Data, *Foreign Affairs*, May/June Issue.

Monroe, B.L., Pan, J., Roberts, M.E., Sen, M., and Sinclair, B. (2015), No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science, *PS-Political Science and Politics*, 48(1):71-74.

Mullainathan, S. and Spiess, J. (2017), Machine Learning: An Applied Econometric Approach, *Journal of Economic Perspectives*, 31(2), 87-106.

Murphy, K. (2012), *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge.

OECD (2017), *OECD Digital Economy Outlook 2017*, OECD Publishing, Paris, ch.7.

Prüfer, J. (2013), How to Govern the Cloud? *IEEE CloudCom 2013*, DOI 10.1109/CloudCom.2013.100, 33-38.

Prüfer, J. (2017), Trusting Privacy in the Cloud, mimeo, Tilburg University.

Prüfer, J. and Schottmüller, C. (2017) Competing with Big Data, *CentER Discussion Paper* No. 2017-007.

Prüfer, P. and Kumar, P. (2018), Predictors of Organized Crime and Subversion: A Machine Learning Approach, mimeo, CentERdata, Tilburg University.

Rios-Morales, R., Gamberger, D., Smuc, T. and Azuaje, F. (2009), Innovative methods in assessing political risk for business internationalization, *Research in International Business and Finance*, 23, 144-156.

Samuel, A. (1959), Some Studies in Machine Learning Using the Game of Checkers, *IBM Journal*, 3(3), 535-554.

Sobbrio, F. (2018), New media, new issues, in: Claude Menard and Mary Shirley, (eds.). *A Research Agenda for New Institutional Economics*, Edward Elgar Publishing, forthcoming.

Stephenson-Davidowitz, S. (2017), *Everybody Lies – Big Data, New Data, and What the Internet can tell us about who we really are*, Harper Collins, New York.

Tucker, J., Theocharis, Y., Roberts, M. and P. Barberá (2017), From Liberation to Turmoil: Social Media and Democracy, *Journal of Democracy*, 28(4), 46-59.

Varian, H. (2014), Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28(2), 3-27.

# Appendix

## Getting started yourself

To start working in the field of Data Science many online resources are available to learn from. However, to actually grasp these, one should have decent knowledge of basic linear algebra and probability theory. Useful resources to learn these methods are the video lectures of Andrew Ng at Stanford University, GitHub repositories, and Coursera, Udacity, Udemy, or edX courses. To gain practical experience with various kinds of challenging data, data science enthusiasts can try many open projects available at Kaggle (the biggest data science community in the world).

Bishop (2011), Hastie et al. (2009), and Murphy (2012) are excellent entry points in book form.

## Box 1: Techniques and tools for text mining

| | |
|---|---|
| **Basic Text Mining** | *Natural Language Toolkit (NLTK)* is a widely used open-source toolkit for text mining and NLP. It has several handy tools, gives access to many text corpora, and to the most suitable algorithms for such tasks. [http://www.nltk.org/] |
| **Web Scraping** | *BeautifulSoup* is a tool to work with web-based data. It facilitates the scraping, parsing, and reading of web data, as well as data access using web APIs in different formats of data, for example in HTML, XML, and JSON formats. [https://www.crummy.com/software/BeautifulSoup/bs4/doc/] |
| **Text Classification** | One of the typical tasks in *supervised* machine learning. Assigning categories to documents, which can be web pages, library books, media articles, etc. has many applications, for instance, spam filtering, e-mail routing, or sentiment analysis. Several toolkits are available for supervised text classification. *Scikit-learn*, an open-source machine learning library in Python, is a prominent one. [http://scikit-learn.org/stable/] |
| **Information Extraction (IE)** | The main goal of IE is to identify and extract fields of interest from free text. It is the first step in converting the unstructured text to more structured forms. The so-called *Stanford NLP* is a suite of very useful NLP tools for IE. [https://nlp.stanford.edu/software/] |
| **Semantic Similarity & Topic Modelling** | Algorithms to detect semantic similarity are used to group similar words into semantic concepts that have the same meaning, or appear to have the same meaning. For example, *currency – money – coin* are semantically similar. One useful resource for semantic similarity is WordNet, a semantic dictionary of words interlinked by semantic relationships.[15] Topic modeling is a widely used text-mining tool for discovering hidden patterns in a text body. A good topic model for example gives 'school', 'university', 'college', 'teacher', 'professor' for a topic "Education". |
| **Sentiment Analysis** | Opinion mining (sometimes known as sentiment analysis or emotion AI) refers to the use of NLP to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely used to analyze reviews, survey responses, and online and social media discussions (e.g. Twitter data). There are two ways to perform sentiment analysis: the lexicon-based approach and the machine-learning approach. For both approaches different tools and algorithms exist as well as data bases of positive and negative words.[16] |

[15] Primarily developed for English, WordNets are now available for many other languages. WordNet includes rich linguistic information e.g. part of speech, different meanings of the same word, synonyms, words with same meaning, hypernyms and hyponyms. WordNet is freely available in NLTK (http://www.nltk.org/howto/wordnet.html). It is extensively used in many natural language processing tasks and, more broadly, in text mining tasks.

[16] For example the Liu and Hu opinion lexicon contains around 6800 positive and negative opinion or sentiment words for English: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon. *SentiWordNet* is a good lexical resource for opinion mining that assigns three sentiment scores (positivity, negativity, and objectivity) to the words in WordNet. See http://sentiwordnet.isti.cnr.it/. *NLTK* and *TextBlob* are Python libraries, which are frequently used for sentiment analysis based on machine learning. TextBlob is built on the top of NLTK and is more convenient than NLTK for new users and has a lot of functionality in NLP tasks. Similar libraries are also available in R and RapidMiner.