
ADA project consortium
Data Science
Foundations and applications

ADA project consortium

DATA SCIENCE

Foundations and applications

Belgrade, 2020

ADA project consortium

Data Science

Foundations and applications

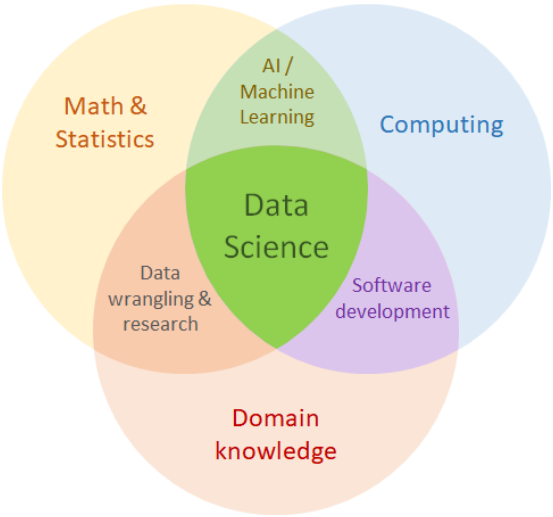
Belgrade, 2020

This booklet is published as a result of the project **Advanced Data Analytics in Business (ADA)** – EACEA 598829-EPP-1-2018-1-RS-EPPKA2-CBHE-JP, co-funded by the Erasmus+ programme of the European Union*.

What is data science – DS?

Although nowadays DS is one of the most frequently heard buzzwords in academia, business, and services, there is still no strict and widely accepted definition of it. In addition, whenever people talk about DS they also tend to use terms like data analysis, data mining, machine learning, artificial intelligence (AI), Big Data, etc. These are words that go together well, but are not exact synonyms with DS.

Still, many people knowledgeable in DS will agree that it is about creating, preparing, managing, curating, maintaining and using various datasets in order to analyze and make inferences about the phenomena and processes that the data is collected from and that the datasets represent. Doing DS requires knowledge and skills in math & statistics, in computing and in application domains.



Students might ask: Why would I want to study DS? Well, because of the great employment opportunities in the first place. There is already a high (and rapidly growing) demand for data scientists in companies and other institutions, and salaries are high. Also, because of very promising career building – from attractive starting positions, to quick

promotion to decision-making ones. Last but not least, because doing DS is so much fun! It is simply exciting to play with data and get some useful, previously unknown and often surprising insights.

Likewise, companies might wonder: Why do we need data scientists? The most concise answer has been summed by the *Wired* magazine that calls data "the oil of the digital economy". There is literally a flood of data everywhere, at all organisations, and a whole new workforce is needed to put in charge of controlling it and turning it into a business advantage. The potential of data is incredibly huge, at all levels, as companies need to use data to run and grow their everyday business. Time is money, they say, but so is data.

There is virtually no domain where DS is not needed: from business to energy industry, from education to demography, from healthcare to transport, from finance and insurance to retail, entertainment and more. For instance, oil giant Shell has managed to save millions of dollars a year in their inventory management thanks to a predictive modeling DS platform that analyzes numerous oil drilling machine parts that might fail. And Netflix recommends you the next movie to watch by using Big Data analytics over search and watch data coming from over 100 million subscribers.

So, who are data scientists and what do they actually do? Well, they know how to identify DS problems that organizations might tackle in order to increase their opportunities. They also use different tools to collect data from different sources, clean it, validate it and transform it for analysis. Then they use their skills in modeling and data mining in order to possibly discover patterns in the data and interpret them to discover solutions to the problems. They communicate their findings to stakeholders using different data visualization tools.

DS building blocks

Mathematical foundations

Mathematics is at the very foundation of science, including Data Science. Methods and techniques of modern DS deeply rely on mathematics, especially on linear algebra, discrete structures, calculus, probability, and statistics. In addition to providing models of real-world problems, mathematics also develops tools for solving the problems. Although we do not need to be experts in mathematics to be able to apply its tools in solving DS problems, understanding of what is going on behind the scene is always a great advantage.

Linear algebra is the field that could be named *Mathematics of data*. It essentially operates on data streams whatever their source and interpretation is. A transformation of a picture from one format to another and a recommendation of a movie by the aforementioned Netflix have something in common: matrix algebra. It includes different topics such as basic properties of matrices and vectors – scalar multiplication, linear transformation, inner and outer products, matrix multiplication and various algorithms, special matrices, vector space, eigenvalues, eigenvectors, and many others.

Linear algebra is not easy, it is necessary for Data Science, but - guess what! It also can be so much fun! Doing linear algebra on real data and with executable code is a lot of fun. And where linear algebra cannot solve a linear algebra problem, numerical methods come in handy. It is an amazing mathematical field on its own, perfectly tailored for computer programming, giving approximate and "good enough" solutions to many problems.

All modern DS is done with the help of computational systems, which are based on *discrete mathematics* and *discrete data structures*. Some key topics of discrete data structures are stacks, queues, graphs, arrays, hash tables, trees, and discrete mathematics in general - sets, graphs, recurrence relations and equations, asymptotic behaviour and O(n) notation concept.

When data start "thickening" up to continuity, such as child growth, car speed and the like, this is where *calculus* comes to play. We are in the calculus universe if surrounded with limits, continuity and differentiability, derivatives, acceleration, velocity, integrals, slopes, series… Calculus is used in many fields where data comes from. In economics, calculus is used to compute a rate of change of prices. In biology, it may be used for calculating birth and death rates. In physics, it is used when dealing with motion, electricity, etc., whereas in chemistry, calculus can be used to predict functions such as reaction rates. So, calculus concepts and applications are present in many different places in DS and machine learning. Data scientists use it for implementing various algorithms such as, for example, logistic regression.

Probability is a measure of the likelihood that an event will occur. Probability is sometimes counter-intuitive. That's why it is important for a data scientist to study it. Let's look at a couple of examples.

- If you roll a well-balanced dice 6 times, is the probability of getting the series of numbers 1 5 2 4 6 3 - higher / lower / the same as the probability of getting the series of numbers 6 6 6 6 6 6?
- If we pull out a number from the set of natural numbers N, is the probability of pulling out different numbers – the same? Should the sum of those probabilities still sum up to 1?
- If we guess the place of Maserati behind one of the three curtains (the other two hide donkeys), and, before being informed whether we are right or wrong, we change our mind when shown a curtain with a donkey, shell we increase our chances to win or not?

The most common topics of *probability theory* include: probability, conditional probability, random variables, Bayes's rule, probability distributions, normal distribution, etc.

DS often uses *statistical inferences* to predict or analyze trends from data. On the other hand, statistical inferences almost always use

probability distributions of data. So, probability and its applications are important for DS problems.

Statistics is an absolute must-know in DS. Many think of machine learning as statistical learning in general. Although statistics "exposes everything and does not show anything", it still helps decide whether to diet or exercise if one wishes to lose weight, will nicotine patches help quit smoking or not, does smoking during pregnancy influences the child's IQ, and the like.

Important concepts to master through studying statistics include basic probability, conditional probability, probability distribution functions, data summaries and descriptive statistics, correlation, hypothesis testing, confidence intervals, p-values, Analysis of Variance (ANOVA) models, t-test, linear regression.

Today's *DS algorithms* often rely on huge data sets. In order to deal with such a large data quantity, either the algorithms need to be adjusted to scale better or data itself needs to be transformed. Two standard approaches for decreasing the size of data are vertical and horizontal reduction. When data is arranged in a tabular form, the table columns correspond to data features, whereas rows represent separate records or vectors of features. Therefore, *vertical reduction* (or *dimensionality reduction*) reduction relates to feature set downsizing in such a way that the original data set informativeness is preserved as much as possible. The feature set can be modified in various ways: features can be removed or weighted, transformed via discretization or aggregation, combined in order to form new features, etc. A well-known method for dimensionality reduction is *Principal Component Analysis* (*PCA*). PCA transforms the original feature set into a new feature set of smaller size.

Numerosity reduction is *horizontal*, meaning that the original data set is reduced on a per-record base. One way to do it is to use sampling, i.e. selection of a subset of records which is to behave similarly as the whole set with regard to certain statistical aspects. Numerosity reduction can also be done by utilizing *cluster analysis*, i.e. forming

groups of similar records whose representatives (centroids) are later used to represent the original set of records.

Classification is one of the most frequently studied problems in DS. Classification problems occur often in real life. For example, when you go to a bank to apply for a loan, your case is most probably fed into some classification-based methodology, performed either by a human expert or an algorithm. If your previous history of using bank loans is clean, you are married, have children, healthy cash flows, etc., then there is a high chance that you will return the loan. This task can be modeled as a binary classification problem, meaning that the output of the classification algorithm is twofold (yes, approved /no, declined). Classification algorithms usually rely on previous, historical data, and some popular algorithms are *Support Vector Machines* (*SVM*), *k-Nearest-Neighbors* (*kNN*), *Artificial neural networks*, etc.

Regression methods are like classification methods, except that the predicted output does not necessarily belong to some fixed number of options. Let's say that instead of predicting whether the bank client is likely to return the loan (yes/no), the goal is to predict what is the chance that the client will return the loan, therefore, any value from the interval [0, 100%] is now possible. Regression methods can be used in the context of time series analysis as well, e.g. in the prediction of atmospheric conditions like air pressure and temperature, in stock markets to predict the stock prices, etc.

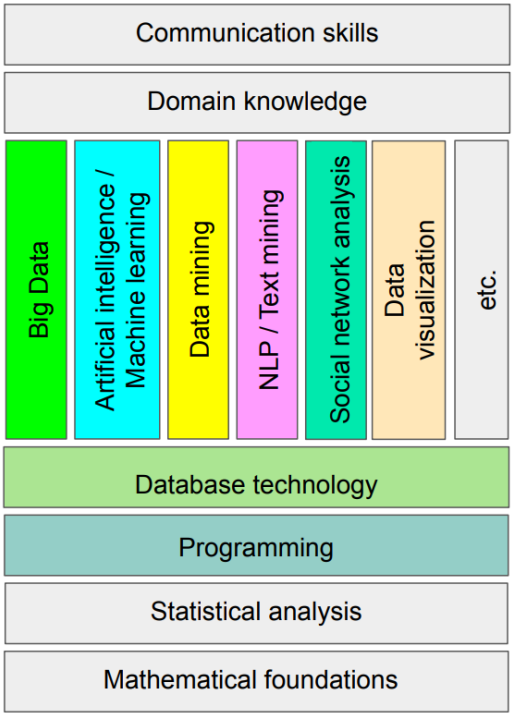
Cluster analysis relates to finding groups of data records that are in some way regarded as similar. Following the previous example, the task might be to perform cluster analysis on the set of all bank clients and to obtain separation such that clients inside the same group (cluster) are very similar among themselves, whereas for any two clients that belong to different groups there is a clear distinction. These insights can be later used to better understand the client needs in general and to address them appropriately by offering different kinds of loans across different groups.

Optimization of parameters of DS algorithms. Successful application of DS algorithms relies on setting appropriate parameters. For

example, the SVM algorithm usually depends on two parameters, the so-called regularization constant C and the kernel hyperparameter γ. In order to have high-quality classification results, SVM needs to be executed with an appropriate combination of these parameters. Since both parameters belong to a continuous domain, the solution space can be very difficult to search. The most basic approach, usually used as a bottom-line approach, is Grid search, an optimization algorithm in which a good combination of parameters is searched by systematic check of all possible combinations of parameters with a certain precision. When precision is high, the number of combinations gets very high as well. Another, more efficient approach would be to follow the gradient. However, this approach tends to "get stuck" in so-called local optima. Some more prominent approaches often avoid an exhaustive search and instead utilize random number generators to a certain degree. The extreme approach in this methodology would be to simply use the Random search algorithm. The best results, however, are usually obtained by using approaches that are somewhere in between exhaustive and random search, e.g. approximative algorithms such as metaheuristics (Genetic algorithm, Variable neighborhood search, etc.) or probabilistic algorithms such as Bayesian optimization.

Technological foundations

Data scientists often use the term *DS technology stack*. Intuitively, it includes a wide range of tools, technologies and related skills that help people run a DS project. Although some of these tools and technologies are alternatives to each other, it is necessary to have a big picture of the DS technology stack. It provides a wider technological context when it comes to tasks and processes like data collection, storage, cleaning, transformation, exploration, analysis and presentation, as well as to integrating these processes and their results into coherent DS applications.



All DS starts from *data*. Data itself comes from a variety of sources. It can be Websites, applications, mobile devices and apps, different services, sensor systems, social media, large volumes of text, large collections of audio and video recordings, etc. Some data is partially preprocessed and partially structured, other data can be largely unstructured. Some data has to be scraped from the Internet, and some other data can come from the Internet of Things (IoT) systems.

Wherever data comes from, it gets stored eventually in *databases* and/or *datasets*. Both databases and datasets contain data, but there is a difference between them. A database is an organized collection of data, typically controlled and accessed by a database management system, which is software that manages data storage, retrieval, and update in a structured way, as well as multi-user access to the database. Data in a database can be stored as a set of tables, typically

referencing each other, or they can be stored as documents (in different formats), graphs, key-value pairs, vectors and so on. There is a variety of techniques that enable fast access to data in large databases.

A dataset is just a set of data, typically organized as a table, i.e. arranged in rows and columns for processing by statistical software. The data in a dataset might have come from a database, but it might not. In DS, the data being analyzed often comes from datasets. Rows in datasets are often called observations or cases. Columns are typically called variables or features. We say that the data in a dataset represent different observations/cases of a phenomenon, each case being described by the same set of variables/features.

But what if these datasets grow too large or too complex to become difficult to process and present with traditional data-processing application software like spreadsheet applications, visualization packages, or widely used database management systems? Then we're talking about *Big Data*. In the world of Big Data, the volumes of data are huge (we're not talking terabytes, but petabytes or exabytes), the types of data is very diverse, the data velocity (the speed at which data changes and at which it has to be processed) much higher than with traditional data sources, and the data veracity (the amount of "noise" in the data) is critical.

Note, however, that Big Data must ultimately have a certain value for businesses and end users – unless we can turn huge volumes of data into value, it is useless. Understanding both the costs and benefits of Big Data cases is crucial.

Obviously, Big Data puts a range of challenges to users and analysts in terms of capturing, storing, sharing, querying, handling privacy, and the like. It is often not enough to have only specialized software to deal with Big Data, but also specialized hardware and/or hundreds or thousands of servers.

A necessary part of each data scientist's toolset is *data visualization* tools (and skills!). Big or not-so-big, data often needs to be visualized in order to make sense out of it. It is not only about a picture is worth

a thousand words, nor it is only about presenting data in a form more attractive than datasets/tables. It is primarily about enabling end users to grasp difficult concepts or identify new patterns, trends and previously unknown clusters in data. It is also about quick comprehension of large volumes of data, and effective communication of the meaning of data. Last but not least, it is about experimenting with different scenarios of future changes in data.

Data visualization technology is nowadays advanced and sophisticated. It enables drilling down into suitable visual representations of data for more detail, interaction with graphs and charts, changing the views and focuses, and watching animations of data dynamics. Data visualization tools enable creative and imaginative exploration of data in order to get new insights. 1D (linear), 2D (planar), 3D (volumetric), nD (multidimensional), temporal, tree/hierarchical and network visualizations are just broad categories of data visualization types; each category includes a number of specific visualization types. Selecting an appropriate visualization for the data at hand is a very intriguing thing, can largely speed up the process of spotting interesting patterns in the sea of data, and, yes – it is aesthetically appealing.

AI is a highly interdisciplinary area of study and practice focusing on the development of systems that can perform tasks normally requiring human intelligence – knowledge representation, reasoning about processes and phenomena, natural language processing, translation between languages, perception, vision, speech recognition, problem-solving, decision making, and so on.

Recently, the most popular field of AI is *machine learning* (*ML*). Wikipedia defines it as "the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead". Typical ML tasks include classification (building a model that can tell us with a higher or lower certainty what class from a pre-specified list data-item classes an observation – a data item – belongs to), clustering (identifying more or less coherent groups – clusters – of data from a dataset), regression (predicting the outcome

of an event or the value of the output variable based on the relationship between other variables in the dataset), and so on. Practical ML applications build a model based on a sample of data from a dataset, known as "training data", verify the model's accuracy based on another sample of data (the "test data") and then use it in order to make predictions or decisions on new or previously unknown data. ML is the most relevant field of AI for DS.

Neural networks (*NN*) are among the most popular ML systems. They are loosely modeled after human (animal) brain and nervous system. When a NN is shown a sufficient number of examples ("training data"), it can learn how to recognize patterns and then perform tasks like identification, classification, and clustering. For example, a NN can be trained to recognize fingerprints, handwritten characters, objects (even moving ones), human faces, and many more.

Technically, a NN is organized as a number of small processing units (neurons) interconnected and grouped in layers. Multiple layers in between the input and the output layers can enable *deep learning*, where each layer transforms its input (coming from the previous layer) into a slightly more abstract and composite representation. For example, a deep learning NN trained to recognize human hands can have a layer that abstracts edges from pixels, then the next one to recognize different arrangements of edges related to parts of the hand, followed by the layer that recognizes joints, then the fingers, palm, and finally the hand. Deep learning networks may require a large amount of training data and powerful computers, but enable building exciting applications, such as self-driving cars, automatic translation of text between different languages, automatic text and handwriting generation, breast cancer detection, earthquake prediction, and many, many more.

A concept very similar to that of ML is *data mining*. Data mining tools search for meaning in large amounts of data, finding patterns in seemingly unrelated data and putting them together to provide interpretation. It usually requires a lot of human interaction and driving, whereas ML systems, once trained, can perform autonomously. Data

mining can be thought of as "food" for ML, since ML often uses datasets formed from mined data.

Text mining (or *text data mining*, or *text analytics*) is the process of extracting knowledge from unstructured text. It usually includes structuring the input text first, discovering patterns in the structured text data, and evaluation and interpretation of the output. Intriguing applications and subfields of text mining include text categorization, text clustering, topic modeling, concept/entity extraction, taxonomy extraction, sentiment analysis, creating document summaries, and so on. Text mining overlaps with *natural language processing*, which focuses on analyzing large amounts of natural language data, speech recognition, natural language understanding, natural language translation, and natural language generation.

Social network analysis (*SNA*) uses networks/graph theory and data visualization to explore social structures, interactions, and patterns. The nodes in visualized social network structures (sociograms) are people (or any other kind of actors or interacting things) and the links represent interactions or relationships among them. SNA is getting increasing popularity in DS. It is used to analyze social media networks, friendship and acquaintance networks, business networks, disease transmission, sexual relationships and so on. Although the roots of SNA are in modern sociology, it is increasingly used in other social sciences (anthropology, demography, communication studies, economics, geography, history, organizational studies, political science, social psychology, development studies, sociolinguistics, and as well as in biology, information science, and computer science).

Programming and *software engineering* underlie all of these other technologies on the DS technology stack. Even if the objective of a DS project is not to develop an application, some programming is always involved in doing data analysis. Thus using a programming language like Python or R is always part of a DS project. Data scientists typically also use different programming tools and libraries that come with these languages, in order to do data cleaning, transformation, analysis, and visualization.

Now, there are some typical questions about programming in DS: Should a data scientist care more about analysis, interpretation, mathematics rather than programming? What's wrong in using a GUI-based tool? Well, there are not-so-obvious things that come with programming skills and are essential in DS. GUI-based tools are all fine up to a certain point. However, when it comes to production, all models require some fine-tuning, all analyses require some reproducibility, and only actual programming enables endless customization of our models. Likewise, programming is essential for developing computational thinking, which underlies almost all DS. Understanding how different algorithms work and what are numerous options in fitting a model is impossible without practical experience in programming.

Applications

There is virtually no field where DS cannot be used, as long as we have a suitable dataset to work with. What follows is but a small excerpt from exemplary DS applications.

Applications in business

If data is the new oil, then mastering the process of its refinement into business decisions is the key to unleashing its potential. DS allows examining large amounts of data to uncover hidden patterns, correlations and also to give insights so as to make proper business decisions.

Generally, most businesses have several goals when adopting DS projects. While the primary goal for most of them is to enhance customer satisfaction, other goals include cost reduction, better targeted marketing and making existing products and business processes more efficient.

Data Science Techniques as Part of a Product/Service

As previously mentioned, Netflix uses data analytics for targeted advertising, and for sending suggestions for the next movies a

subscriber should watch. Netflix does this by using users' past search and watch data, and thus can make suggestions for similar content. The company also uses its data insights to select what shows to buy, license, and what new content to create to cater the needs of their subscribers.

Spotify, one of the biggest digital music services, also uses data analytics to create personalized playlists for each subscriber. The suggestions that these playlists provide are very good, and here's why. Spotify data analysts analyze all artists, their styles and categorize them by loudness, danceability, energy, and more. Then, all artists are grouped into clusters based on these features, and ready for the recommendation engine. The recommendation engine takes the subscribers' favorite artists and gives additional recommendations in terms of artists found in the same cluster.

Rolls Royce, the world's second-largest manufacturer of aircraft and sea vessel propulsion systems, invests heavily into building the Internet of Things and data analytics technologies both on their ground facilities and all the fleet they are building in order to provide top-notch services. They have incorporated data analytics technologies into their main products, the aircraft engines ran by more than 500 civil and 150 military entities worldwide. Their engines are controlled by *Engine Health System*, system using up to 100 parameters from the data stored in the flight recorders and analyzed after the flight. The plan for the next generation of engines is to monitor more than 5,000 parameters and will be connected to the cloud monitoring solution, in order to prevent issues before they occur.

Buxtonco company is using data analytics to help others choose a good retail location. This is especially important for restaurants and stores, since some will never succeed because of a poorly selected one. Buxtonco does this by looking for where customers spend their time, and what they are doing in certain locations - something very similar to geofencing. This way, they can determine where it would be best to open the next business.

Gramener, a data visualization and predictive analytics company, implemented a Web-based AI program for the Nisqually River Foundation, a Washington-based nature conservation organization. This program measures and monitors the fish species present in the Nisqually river by automatically analyzing data from a video camera and infrared sensors in the water.

Cargill developed a mobile application that helps shrimp farmers reduce the mortality rate of their yields [4]. The application uses data analytics to predict biomass in shrimp ponds (based on temperature pH and nutrition) and works together with an automated shrimp feeding system to achieve optimal results. The data is uploaded from the application to the cloud, and then farmers can access the live dashboard that visualizes pond performance, providing key measures and predictive analytics that help them better manage shrimp health and increase yields.

Data Science Can Improve Customer Satisfaction

In 2015, Coca-Cola began using data analytics for its digital-led loyalty program. This enabled the company to create and serve different advertising content to different audiences: sport fans, music lovers, etc. Consequently, it had boosted Coca-Cola's customer acquisition and retention.

Optimizing HR with Data Science

Xerox is an international company that provides office appliances and accompanying services. For years the company was struggling with high attrition rates of its employees in support centers, as people were quitting the job regardless of the HR efforts, regardless of additional perks, benefits or bonuses the company has introduced. Experienced professionals, highly trained, working for a decent sum of money, were quitting the job. After applying the data analytics over the employee data, the company had discovered that the previous experience as a call center operator did not matter at all. Personality traits of the employees were discovered to be the key factor. Having team members with the spirit of partnership and mutual respect proved to be much more efficient than having superstars in the team, organizing

in-team competitions or offering bonuses for the extra cases. Such a shift in approach resulted in Xerox reorganizing their hiring approach and lowering their support personnel attrition rates by 20%, saving the company millions of dollars long-term.

Data Science in Other Domains

PepsiCo, being a big multinational company, relies on efficient supply chain management. They have taken the process of supply chain management to another level with data analytics by leveraging retailers' and POS sales and inventory reports to forecast the production and shipment needs. This way, the company ensures retailers have the right products, in the right volumes and at the right time.

UOB bank from Singapore uses big data to improve risk management, i.e. to reduce the risk calculation time. The whole process used to take up to 18 hours, but has been reduced to several minutes with the potential to carry out real-time risk analysis.

Transport for London (TfL) is the company operating millions of buses, cabs, subway trains, ferries, roads and traffic lights systems in London. They serve nearly 10 millions of Londoners and collect a vast amount of data daily. This data allows the company to understand how and when passengers commute, the long routes used, what roads and bridges are more loaded and what public transport routes face heavier traffic. Daily analysis of such data allows allocating more buses to more loaded routes in order to minimize the time spent commuting and improve the traffic flow; adjusting the traffic lights on the more loaded roads to allow them to increase throughput capacity and avoid traffic jams, and many more activities.

Applications in medicine and life sciences

Data analytics in life sciences has undergone revolutionary changes during the last 10 years, and these changes are most visible in medicine and healthcare. Introduction of novel and innovative information management systems has led to significant changes in the functioning of contemporary healthcare institutions such as clinical

centers and large hospitals. New strategies and approaches in data analysis have also enabled substantial advances in research in both biology and medicine.

Research in both biology and medicine is almost impossible without a competent data scientist. Often, after experiments and clinical trials, there is an abundance of raw data that need to be processed and structured. Medical doctors and biologists often lack professional knowledge in statistics, so any help from a data scientist is valuable. Statistical tests such as T-test, Chi-square, ANOVA, Pearson and Spearman correlations are all essential tools in fundamental and clinical medical research.

The most obvious example of the value of data analysis in medical research would be the introduction of novel diagnostic tests in the fields such as internal medicine, neurology, oncology or radiology. Before the test can be implemented in clinical practice, its sensitivity (ability to identify those with the disease), specificity (ability to identify those without the disease) and various other characteristics need to be determined. Each patient is unique, and while a test may produce outstanding results in a specific group of patients, in other circumstances its performance may be poor. When evaluating the future diagnostic test, researchers must deal with complex and heterogeneous data obtained under various conditions.

Big data in biomedicine, particularly those related to Genomics, Epigenomics and Proteomics, are today increasingly popular among medical doctors and researchers. With the development of novel and innovative information technologies, it has become possible to adequately organize, classify and transform these data, and extract information useful for future biomedical applications. Indeed, today there are numerous freely available databases with huge amounts of potentially valuable information waiting to be discovered and implemented in clinical practice. And the only reason for the delay is the lack of competent data scientists with adequate expertise in bioinformatics!

The *Human genome project* that was completed in 2003, was one of the world's largest scientific endeavors intended to map all the human genes and to assess both their structural and functional characteristics. The resulting sequences are now stored in many databases such as the one of the U.S. National Center for Biotechnology Information, Ensembl (European Bioinformatics Institute and the Wellcome Trust Sanger Institute database) and others. Interpreting genome data is very difficult without complex data analysis software and a skilled IT specialist. The amount of potentially useful information that one can obtain from these databases is truly immense and so is the potential application in future development of new medications and therapeutic strategies for many diseases. On the other hand, due to the lack of human, financial and other resources related to data analysis, the level of utilization of these databases is relatively low. Here lies the opportunity for a future data analyst to help integrate knowledge in the areas of informatics, biology and medicine in order to help future treatment and cure of various illnesses that are today considered incurable. And this is the major goal of today's medicine.

In large healthcare providers and institutions, doctors are often faced with enormous datasets containing information from thousands of past and current patients (4). Management of these datasets is crucial for the proper functioning of these institutions, and various strategies and information systems have been developed to serve this purpose. Data such as patient names, IDs, age, gender, diagnosis, appointments, assigned doctors, etc. need to be stored and easily accessible for future clinical and research work. Also, these systems may store large diagnostic data such as radiographic images (X-rays, Computerized Tomography, Ultrasound, Nuclear Magnetic Resonance) that can further be analyzed using various computer-based algorithms before making a definite diagnosis. Finally, issues such as patient privacy are of utmost importance in clinical practice and managing data security is a crucial skill each data analyst should possess.

In Serbia, a typical example of a large centralized database system for storage and editing of patient data is "Integrated health information

system" (Интегрисани здравствени информациони систем Републике Србије). It enables efficient management of specialist appointments, as well as patient electronic referrals to different doctors and medical institutions. Through this system a patient may ask for and choose an adequate time for appointment with his family physician / general practitioner. Also, general practitioners through this system interact with secondary and tertiary level medical institutions such as hospitals and clinical centers. All this significantly facilitates communication, saves time for both the patient and the doctor and may lead to timely diagnosis and treatment. And of course, the creation and maintenance of such as complex piece of information technology wouldn't be possible without a team of highly qualified and hardworking data analyst.

Apart from large, county-level integrated healthcare systems, many large hospitals and clinical centers have their own local databases where they store data related to their employees, patients, diagnostic procedures, projects etc. Although local, these databases tend to be very large and difficult to manage. There were many instances in the past where, due to the technical difficulties or security flaws related to database access, editing and storing, the hospital experienced severe disruptions in health services. An example would be the outage in 2015 in Hillingdon hospital in London (serves Heathrow Airport) where because of a network failure, medical staff were unable to access data required for patient treatment. System crashes occurred in many other hospitals such as the one in Boulder, Colorado in 2013 and some providers in Australia. In one survey covering 50 different healthcare institutions, in the last 3 years, 70% reported at least one IT system crash causing disruptions greater than 8 hours (CMAJ. 2014 Apr 15; 186(7): 493. doi: 10.1503/cmaj.109-4719). This is one of the reasons why a qualified data specialist is highly respected (and often paid as well) in many world hospitals.

Another important area of medicine where sophisticated data analyst skills are essential, is computer-based electronic generation of medication prescription, or simply e-prescribing. This technology enables a medical professional to electronically transmit drug

prescription or other drug-related authorization directly to pharmacy. In some countries such as Denmark and Sweden, this is a routine practice. Other countries such as the ones in Southeastern Europe are currently starting or have already started to implement it in their health system. Electronic prescribing, apart from saving time and reducing bureaucratic problems, significantly contributes to the overall physician effectiveness. Physician can now display the complete list of active and approved medications along with official recommendations and warnings (i.e. drug-drug interactions, allergy concerns) and wisely choose the best one. Also, the probability in making unintentional prescription errors is now substantially reduced as is the possibility of drug abuse. Managing e-prescribing as a new and innovative technology requires many experienced and skilled data analysts, and job opportunities in this field will definitely increase in the future years.

A study program in DS can be designed to include courses related to data analysis in medicine, biology and life sciences in general. In such cases, these courses help current and future information technology professionals build expertise in implementing IT knowledge in life sciences. The multidisciplinary nature of such a study program contributes to establishing adequate cooperation between IT engineers, medical doctors, and biologists, which greatly benefits both our current knowledge and practices in these areas.

Applications in social sciences and humanities

The field of social data science (SDS) is becoming increasingly popular. SDS revolutionizes the field of social and behavioural sciences. By using new sets of digitally generated data (“big social data”) and innovative analytical techniques, social data scientists are becoming more and more capable of addressing complex social problems. For SDS, maths, statistics and computing are essential but so are theoretical models, as well as methodological and ethical standards developed in social sciences. Someone might say, with a dose of humour, that SDS is social sciences “on steroids”, but we

would suggest that it represents social sciences fit to combat the global social problems of the 21st century.

What kind of data are used in SDS? Mostly the digital social data. Today, billions of people all around the world are using smartphones, different gadgets and the Internet to communicate, study, do business, find information, plan holidays, share content on social media, shop, pay bills and taxes. Every time we visit a website, shop online, search for a medical diagnosis, monitor our health using a medical app, listen to music on Youtube, binge-watch hit series on Netflix or comment on our friends' posts on Facebook, we, as users, generate potentially valuable data. All these daily activities leave behind our digital traces on the Internet. In this way, we are collectively creating a giant social selfie. However, our digital imprints represent an enormous quantity of messy and unstructured data, so to be able to see and understand this grand picture of society, we need social data scientists. Social data scientists use maths, various statistical modelling techniques and programming to gather, organize, clean, visualize, analyze and make sense of these data.

With some help from DS, social big data can be put in a good use. Big data analysis can assist policy makers in monitoring, understanding and solving social problems. For instance, SDS is successfully used to investigate hate speech and covert discrimination on Facebook, to explore trends in political communication during presidential election, to map political landscape on Twitter, to look into high-level corruption in public procurement in different countries. It was applied to discern fake news, but also to explore the profiles of deceased Facebook users (whose number will soon exceed the number of living) opening the ethical question of digital data preservation and curation. Being successful at understanding and finding solutions for complex social issues, SDS is particularly vigilant regarding the ethical aspects of the use of big (social) data.

Are people in a better mood in the morning or in the afternoon? On Mondays or on Wednesdays? What about weekends? Although earlier socio-psychological studies give some insights into the patterns of human affective rhythms, their findings tend to be inconclusive,

ungeneralizable (due to the small samples) and based on retrospective self-reports vulnerable to memory error. Given the shortcomings of the earlier research, two sociologists from Cornell University – Scott Golder and Michael Macy (2011) – decided to employ SDS techniques to explore changes in individuals' moods rhythms. Within the scope of their research study, Golder and Macy analyzed approximately half a billion public Tweets, from 2.4 million Twitter users living in 84 countries across the globe! Something completely unimaginable just a few years earlier when social scientists didn't have tools for real-time observation of individual behavior in large populations.The research showed that people tend to have better moods when they wake up and that their mood deteriorates as the day progresses. To answer one of the questions from the beginning - people are happier on weekends, but the morning peak in "good mood" is delayed by 2 hours, since people awake later on weekends.

Another good example of application of social big data analysis is development of social networks forensics for tracking cybercriminals and activities such as identity theft, cyberbullying, sexual harassment and hate speech. Scientists from the University of Nairobi used data collected from Twitter to predict hate speech. By using the key words of hate speech, such as ''kill'', ''rape'', ''murder'', ''assault'', ''kidnap'', ''shot'', ''gun'', ''crime'', etc., they made a selection of approximately 3 million Tweets that were then classified applying Naive Bayes Classifier in 3 categories: positive, neutral and negative hate speech/cyberbullying. The model proved to be successful in detecting and classifying hate speech on Twitter, which was mostly ethnic based. This study demonstrated how Twitter data can be collected and preserved within database for forensic analysis, ensuring the authenticity of Tweets that contain hate speech before the courts of law.

For social data scientists, achieving social good is of great importance. There are already many examples of successful use of digital social data in combating serious health, climate and social issues.

Tuberculosis (TB) is one of the leading global causes of death, taking up to 2 million lives every year. Since India accounts for almost a third of the total number of deaths from TB, the Indian Government decided to end up this disease by 2025 with the help of new technologies and data science. A joint initiative of public authorities, World Health Organization and the major mobile phone providers in the country, resulted in a successful project of mapping geographical locations at risk and detecting the patterns of disease spread by analyzing large sets of mobile network data. The scale, precision and immediacy of mobile data enabled the identification of areas with low TB incidence rates, but which are intensively connected to areas with high TB rates. Statical analysis of the collected data provided a valuable insight into the patterns of disease spread, showing that regular population movement is a more important factor of the TB spread than spatial proximity between high and low TB regions. This finding was exceptionally important in shaping and implementing targeted prevention and diagnosis measures in these areas and reducing the number of new TB infections in the country.

In a similar fashion, big social data were used in Colombia to reduce negative impacts of climate change, in Brazil to predict air pollution, and in Turkey and Japan in creating policies that aim at reducing negative impacts of natural disasters.

Social scientists are more and more interested in big datasets that are generated within public institutions. Data of the health system, public governance, police, education system, the census data, are increasingly being used to improve the performance of institutions. The goal is smart governance. For example, in order to adequately manage risks within a city, analysis of data on criminal activities (time, place, type of offense), offenders (sex, age, biography, etc.) and the city structure help to distribute police patrols adequately, as well as to develop preventive treatment in risky areas. Data of educational transitions, generated within the education system, can give us answers to what extent the inequality in access to education and consequently to the labor market are present.

Analysis of large quantities of digitally generated data, among other things, provide us with the opportunity to have more efficient and user friendly transport in big cities. An example of the successful application of new techniques of big data analysis is the city of London. Traditional methods of analysis could not accurately map the routes, and ways of using different types of transport (tube, bus, taxi, and others) in this city. In 2017, a four-week pilot project was conducted with the aim of getting a better understanding of how citizens navigate the Underground system in London. Based on the data collected through the iBus location system, Oyster card and Mobile applications made for this purpose, it was possible to accurately track the movement of passengers. According to TFL (Transport for London) officials benefits of using big data are numerous:

1. "allowing staff to better inform customers of the best way to avoid disruption or unnecessary crowding;
2. helping customers plan the route that best suits them - whether based on travel time, crowding or walking distance;
3. enabling greater sophistication in proving real-time information to customers as they travel across London;
4. helping further prioritise transport investment to improve services and address regular congestion points - ensuring the maximum benefits to customers
5. providing a better insight on customer flows which could increase commercial revenue from companies which advertise or rent retail units on the transport network."

Social data science integrates social science domain knowledge with the DS analytical techniques. It develops fast in a dynamic environment marked by a synergy of academic work, policy and business application. Social data science gives special attention to the ethical aspects of data management and processing: digital inequalities and exclusion; data openness, ownership and access; privacy; social causes and consequences of the algorithmic biases etc.

Erasmus+ project:
Advanced data analytics in business – ADA

If you take a look at the popular *datascience.community* Website (http://datascience.community/), you'll see lists of hundreds (if not thousands) of colleges and bootcamps that offer degrees and training in DS worldwide. Most of the colleges listed there offer MSc degrees in DS.

The DS hype is huge, and the buzz is gradually spreading over Serbia as well. The industry and services in Serbia are gradually becoming aware of their need for DS as their regular practice and their demand for DS specialists is growing. Simultaneously, many students have heard the buzz.

These two waves are brought in resonance in the European project *Advanced data analytics in business* (*ADA*), http://www.ada.ac.rs/, funded by the European Commission under the Erasmus+ program (project no. EACEA 598829-EPP-1-2018-1-RS-EPPKA2-CBHE-JP). The major objective of the ADA project is to develop several MSc programs in DS / data analytics, at different universities in Serbia, and in line with similar programs at European universities.

A breakdown of this major objective into a set of smaller, interconnected pieces, boils down to the following:

- The skills of experts and in DS / data analytics in Serbia must be raised to the level of EU experts. Some university teachers in Serbia already have this expertise, usually in specific subdomains of DS, but few of them have a good sense of the whole of DS and few have experience in teaching in a DS

study program. To this end, and with the help of project partners from the EU, the ADA project organizes periodic training for teachers, summer schools, visits to relevant labs at EU universities, discussions with international experts and students, and the like.

- Awareness of the importance of DS must be raised among potential students as well. Both the big picture of DS and wider social implications (new employment prospects, changes in the job market, economic and legal aspects of DS, etc.) must be explained to them thoroughly.
- Industry, services and public sector in Serbia must take an important role in defining the demand for new workforce and trained data scientists, as well as in specifying the business problems that the graduates of DS study programs should train the students for and in providing internship and practical placement for students and MSc candidates.
- Changes at different universities in Serbia must be introduced. It is not just about putting up another few study programs – it is also about a certain shift in the orientation of these universities in the education continuum from BSc/BA to PhD levels, about current trends in higher education worldwide, and about overall education approaches. Developing and running ADA MSc programs also means that several BSc programs and a number of courses should consider adapting their teaching approaches to creating a problem-solving attitude necessary for students to benefit from MSc programs in DS / data analytics.
- A wide range of current analytic tools must be introduced in teaching, from statistical analysis tools, complex data mining and predictive modelling applications, to data visualization tools, business intelligence reporting software, self-service analytics platforms and Big Data platforms.
- A great deal of flexibility and cooperation must be exercised, because the idea of ADA MSc programs is not to have "just another few study programs". To this end, intensive coordination of activities and cooperation of staff from

different faculties at all the relevant universities in Serbia and the relevant accreditation bodies is a must in order to achieve a great deal of interdisciplinarity, effectiveness, high quality and the ultimate acceptance of these new study programs.

- ADA programs are internationally oriented, accredited to be taught in both Serbian and English, and teaching staff don't come only from Serbian universities, but also from EU universities, which is currently a very rare case in Serbian higher education.
- Lifelong learning programs in DS / analytics for professionals in different fields should be introduced. All Serbian universities participating in the project also plan on running LLL programs and seminars for professionals already employed in different sectors (Serbian companies, public sector, healthcare, research institutions, and so on). Meetings of the ADA project consortium will help the project participants and different stakeholders select suitable topics to be covered during these LLL seminars.
- The ultimate goal of ADA MSc programs is to make a nation-wide and region-wide impact in higher education. Again, they are not "just another few study programs".
- Putting all these pieces of the puzzle together into a coherent whole requires coordinated approach that takes into account several key players and stakeholders (universities, industry, public sector, media, Ministry of Education, accreditation bodies, EU partners, etc.).

The ADA project partners include:

- University of Novi Sad (the project coordinator, Serbia)
- University of Belgrade (Serbia)
- University of Kragujevac (Serbia)
- University of Niš (Serbia)
- Vienna University of Economics and Business (Vienna, Austria)
- Graduate School of Computer Science and Mathematics Engineering (Cergy, France)

- University of Rome Tor Vergata (Rome, Italy)
- Centre for Research and Technology Hellas (Thessaloniki, Greece)
- Serbian Association of Managers (Belgrade, Serbia)

ADA study programs

Some common starting points in developing ADA study programs at universities in Serbia include an agreed upon desired skillset of graduates from these programs, an insight into a number of MSc programs in DS / Data analytics worldwide, as well as general guidelines for developing curricula of such study programs, published by authoritative institutions.

From the students' perspective ("Why should I apply for this study program?"), the most important point is that the skillset of graduates from ADA study programs is designed to match what is colloquially called *the third-wave data scientist*. This includes proficiency with current statistical toolbox and algorithms for data analysis, software engineering craftsmanship, the necessary soft skills (communication, problem solving, teamwork, explorative attitude, entrepreneurial spirit,...), and business mindset – taking DS job positions to create business value, not just to build models, prioritizing work and knowing when to stop, looking where the money flows in the organization, and experimenting and creating an innovative data culture in order to bring about real change.

With regard to the contents of the DS study programs at the MSc level, note that there are some varieties related to the emphasis in these programs. Many of them are business analytics programs where applications of DS in business dominates the program and course contents. There are also programs where quantitative and algorithmic aspects of data analysis are emphasized, as well as those where DS technology-oriented components are driving the program implementation. Expertise of teaching staff can bring preference for one application domain or another, or it can result in having a curriculum with multiple courses related to Big Data, or to statistical

analysis, or to data curation, security and transformation. There are also numerous examples of study programs that include electives from related fields, such as information theory, signal processing, sampling theory and the like.

The ADA project team has relied on DS curriculum guidelines published by associations like Association for Computing Machinery (ACM) and Park City Math Institute (PCMI). The team has also explored specific implementations of these guidelines in different MSc study programs, and has tried to stay as close as possible to "role model" implementations.

University of Novi Sad: Advanced Data Analytics in Business

The ADA MSc study program in Data Science developed at the University of Novi Sad is a full-time program taught entirely in English, with participation of lecturers from several European universities, aiming to prepare graduates interested in data management and analysis techniques. The program lasts 2 years and has 120 ECTS credits. All courses are one-semester courses. This study program combines an in-depth theoretical understanding of data acquisition and data systems architecture with more business-driven skills such as data analysis and visualizations. It is mainly oriented on business data analytics, therefore it covers all important aspects of mathematics and statistics, business application and IT. The study program meets the high standards of quality of the education system in Serbia and is also in accordance with the requirements of top knowledge in the field of data science and economics in European and world contexts. One of the most important goals of the study program is the education of experts, who in addition to considerable theoretical knowledge, has a sufficient experience in the application of data analytics on the real problems of modern business.

The purpose of this study program is to achieve the following general learning outcomes: mastering competencies in understanding big data, their preparation, modelling, evaluation and implementation of

solutions in business, through the application of programming, statistics, machine learning, data manipulation and visualization. The predicted learning outcomes are based on the experiences of several leading European universities and companies that apply data science in business. Outcomes are also in line with the needs of our economy, job market and the wider community.

The study program is designed to ensure the acquisition of competencies that are defined through communication with companies from Serbia and region, but also to be compatible with similar master studies across Europe, in order to attract international students. With the implementation of the study program defined in this way, business analysts acquire the ultimate combination of knowledge in economics, computer science and quantitative methods, that is acceptable in European and world contexts.

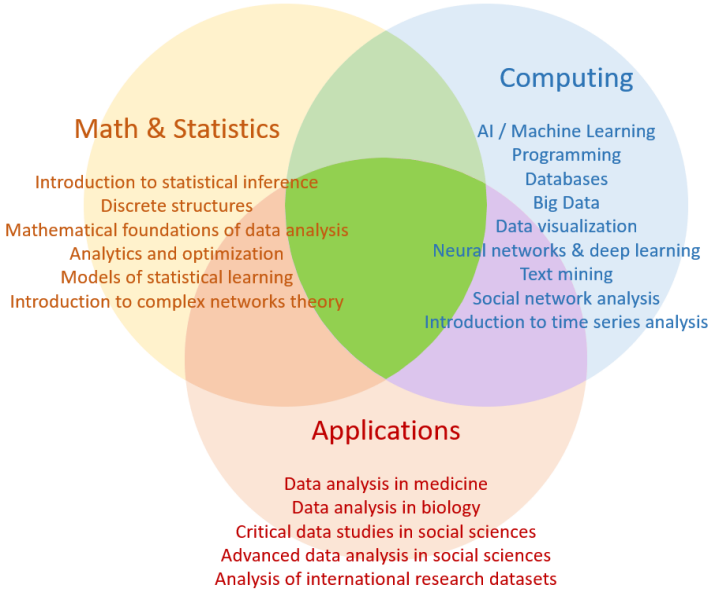
The program offers compulsory courses in the first two semesters, elective courses in third and work on the master thesis in the fourth semester. Before the beginning of the first semester one preparation course, Data Campus, is offered. Courses in the first semester are mainly related to computing foundations of modern data analytics (Big Data Fundamentals and Machine Learning) and specific topics and tools that help data analysts in working on practical problems (Social Media Analytics, Managing, Storage and Visualising Big Data, R for Data Science). Compulsory courses in the second semester are focused on quantitative modelling and statistical analysis (Time Series), together with practical application of data science in business (Business Cases). Third semester offers different elective courses that enable a variety of learning paths (from Data Analytics in Finance, Supply Chain, Marketing or Management to Deep Learning or Academic Writing). The fourth semester is reserved for practical work and for working on the master thesis.

University of Belgrade: Advanced data analytics

The ADA MSc study program in DS developed at the University of Belgrade, *Advanced data analytics*, starts from the general

understanding of what DS is, as explained in the introductory part of this booklet. Accordingly, it offers three distinct groups of courses – Math & Statistics, Computing, and Applications. The program has a strong technological component, implemented in the courses related to Computing. With regard to applications of DS, the program currently targets primarily the domains of life sciences, social sciences and scientific computing. Further releases of the program might also offer courses in business data analysis, Big Data in the domain of energy, and so on. Although the program primarily targets students who have already completed their BSc studies in a quantitative discipline, it is also open for students with backgrounds in other disciplines.

The courses offered in this study program are developed around three major pillars of modern data analytics: mathematical/statistical foundations, technological foundations, and applications. As part of several courses, special attention is given to ethical aspects of data management and processing (digital inequalities and exclusion; data openness, ownership and access; privacy; social causes and consequences of the algorithmic biases; etc.).



The program also includes mandatory internship (capstone project / practicum) for students to get practical experience in working on data analytics projects, as well as mandatory master thesis. Practical orientation also features many of the courses indicated in the figure above, since many of the core hours are actually hours of labs that focus on practical projects.

It is a 3-semester, 90-credits (ECTS) study program. All courses are one-semester courses. The program does not have modules, but the rich offer of elective courses allows students to select those courses that lead them towards the improvement of their knowledge in selected disciplines of quantitative sciences.

Prospective students can take a number learning paths in this study program. Carefully designed prerequisites for many courses enable this variety of learning paths. Essentially, many of the courses offered in the first semester are related to mastering probability, statistics and other mathematical foundations of DS (calculus, linear algebra, discrete structures and the like). Then in the second semester come courses related to computing foundations of modern data analytics (*Programming, Databases*, *Big data*) and different AI topics necessary for advanced data analytics (*Machine learning*, *Neural networks and deep learning*). Several courses in the second semester also cover specific topics and tools that help data analysts in working on practical problems (*Data visualization*, *Text mining* and *Social network analysis*). The third semester is reserved for application-oriented courses, with a lot of practical work, for internship, and for working on the master thesis. With the help of other partners in the project, a number of institutions provides internship for the students, and the internship is designed as the basis for the master thesis.

By defending his/her master thesis, the candidate receives the academic title of *Master of Advanced Data Analytics*. Students who have completed the master program Advanced Data Analytics at the University of Belgrade become competent in:

- independent work in analyzing datasets of different complexity in selected domains, with advanced use of current data analysis tools and technologies
- preparing, modifying, adapting and combining datasets for analysis, out of raw data produced from different applications and other sources

- involvement in various interdisciplinary working teams where data analysis skills in different disciplines and mastery of current data analysis tools and technologies are expected, not only in solving routine practical problems, but also in non-standard situations where creativity and research approach are required
- working with large sets of data

Subject-specific competencies of graduates include:

- ability to understand and analyze different datasets from the perspective of mathematical underpinnings of advanced data analysis (linear algebra, calculus, discrete mathematics, high-dimensional geometry, optimization, etc.)
- mastery of statistical underpinnings of advanced data analysis (data summaries, hypotheses testing, data variance, data correlation, probability and probability distribution functions, applying descriptive and inferential statistics to datasets, etc.)
- programming using state-of-the-art programming languages in data analytics
- skills in using advanced and appropriate data visualization techniques, as well as current software tools and technologies that enable creating rich data visualizations

ADA – employers' perspective

One of the ADA project partners is Serbian Association of Managers (SAM). The association includes more than 400 members. They are managers of successful companies that altogether employ more than 70.000 people in Serbia. In the Spring of 2019, SAM has conducted a survey among its managers and data scientists. The objective of the survey was to identify the needs of industry in Serbia in the context of DS, as well as the skills of data analysts needed in Serbian companies, given the existing organizational structure in these companies. The needs analysis resulting from the survey has shown interesting results.

What can companies and institutions in Serbia get from ADA? The obvious answer to this question is: educated data scientists, graduates from ADA study programs. But there is more to this end.

Interestingly, the managers who have participated in the survey have demonstrated only an average understanding of what exactly is the job of a data scientist, more so in medium-to-large companies and in those that do their business internationally or have foreign owners. Only every fifth company employs data scientists, and respondents generally do not know if they will be employed in the next year. Data scientists working in companies are quite scattered within organizations and work in different sectors: advisory, sales, program and sales management, management, operations, business intelligence, pricing, analytics, business data management, technology, controlling, finance, marketing, strategic management, segment management... There seems to be no universal sector/function where they work[1]. There is also a noticeable distinction between big and small companies: in smaller companies, the CEO is the one who is mostly in charge of managing data and using it to improve business decision-making, whereas in larger companies functions are much more diverse: most functions are related to finance & management. This clearly indicates the need for additional training of managers about how a data scientist can improve the job and what exactly he needs to do. And this justifies the ADA project idea of organizing lifelong learning programs in DS / analytics for professionals in different fields.

What do managers think about DS / advanced data analytics and to what extent it is used in their companies? Almost two-thirds of the managers who have participated in the survey (59%) argue that, according to their understanding and knowledge, their companies have not implemented any DS project so far. Those who have, they have used internal resources. Very few companies (only 2 in the overall sample) hired experts outside the company to assist in the implementation of DS projects. Logically, larger companies and

[1] In a similar study in the USA (Big Data and AI Executive Survey 2019, NewVantage Partners) to 65 of the Fortune 1000 companies, two functions were clearly distinguished: Chief Data Analytics Officer and Chief Information Officer.

foreign-owned companies are more often those that have already implemented DS projects.

What gaps have been identified? The managers who have participated in the survey estimate that the use of advanced analytics in their companies is at a fairly low level. Every fourth company has not yet begun implementing advanced analytics, which is particularly worrying since SAM brings together the most successful companies in Serbia (and therefore, it can be expected that, if the economy is taken as a whole, application is considerably lower). Again, larger companies, foreign-owned companies and companies that trade internationally are more advanced to this end.

What are the barriers and impediments in embracing DS? The basic barrier to applying advanced analytics is the lack of understanding of DS and how to apply it in business. Smaller businesses are also limited by the lack of financial resources, whereas larger companies struggle with organizational difficulties.

What are the priorities in introducing and applying advanced data analytics more intensively? The SAM survey has shown that the programming languages, technologies, cloud platforms or tools important for a data scientist position include: SQL, visualization tools (PowerBI, Tableau, GGPlot, Plotly, Qlik), Python and/or R and Cloud platforms (AWS, Google Cloud, private cloud). The tools and techniques that will be necessary for a data scientist to master perfectly in the future are: Hadoop, Scala, Hive, Spark, Tensor Flow, NoSQL, and natural language processing (NLP). The knowledge of SAS, SPSS and Excel is desirable because these tools are widely used in companies today, but data scientists do not need to be experts in using them. When asked what other skills (in addition to technical skills) data scientists should clearly demonstrate, the experts (data scientists who have participated in the survey) have singled out strategic management, value chain, business processes (because business processes are changing with science), fundamentals of finance (ROI, what is the balance and balance of success), as well as soft skills (presentation skills, communication skills, …).

The SAM survey has also resulted in a rather clear insight into the state-of-the-art of using DS in businesses in Serbia. These findings can be very useful for both the businesses themselves and the prospective students of ADA study programs.

Dominant industries and processes. Business processes where new data analytics approaches and tools are already used include: insight into the consumer profiles and more precise targeting (more pronounced among middle and large companies), financial planning and analysis (more pronounced for small businesses), pricing policy and profitability. Medium and small companies today have a focus on financial planning and analysis, whereas larger companies focus on creating consumer insights and more precise targeting. The sectors most often responsible for optimizing data management and improving business decision-making are finances (dominant), management and business intelligence.

Distribution of responsibilities. The experts who have participated in the survey disagree over whether and to what extent a data scientist should be a domain expert, as well as to be (or not to be) a good presenter. There is a wider and narrower interpretation of the scope of the work of a data scientist (two or one function), as well as whether it is necessary for them to deal with regular reporting or not. What the experts have agreed on is what is *not* the job of data scientist:

- to be an expert in programming or in a programming language that is not related to their work
- to be a database administrator
- to be a financial expert
- to do *everything* related to IT

Given that the data scientist's job description is not fully known to managers, experts have defined what a data scientist *should* be doing in a company:

- technical aspect
    - data preparation and cleansing
    - understanding databases

    - working with advanced tools
    - analyzing, modeling and manipulating data
    - transforming data into information
    - visualizing data
- business aspect
    - communication with people at other positions/functions in the company
    - understanding and participating in defining business processes and data that these processes generate

Most of the lessons learned from the responses to the SAM survey are related to the employment opportunities for graduates from the ADA study programs.

Benefits for students and graduates. Knowledge and skills offered by the ADA study programs and increasingly important in industry today include:

- data analysis, statistics and algebra, data visualization, descriptive analysis (transformation of data into information)
- openness to cooperation with domain experts,
- working with advanced data analysis tools
- understanding of data and its value for businesses
- capability to identify business problems and to model data accordingly
- knowledge of the basics of strategy and finance
- communication with other business sectors
- practice

Benefits for companies and institutions. The managers who have participated in the SAM survey believe that the fact that ADA study programs produce graduates with sound technical and communication skills, good knowledge of processes in industry and in different services, as well as good business knowledge is very important. Representatives of companies currently value more communication and visualization skills than the process of gradually generating a complete business solution (probably due to the lack of understanding of the spectrum of techniques that a data scientist can use).

ADA – Follow us!

ADA project Website: http://www.ada.ac.rs/

ADA on Facebook: https://www.facebook.com/ADA-Advance-Data-Analytics-in-Business-2319108561657462/

ADA on Instagram: https://www.instagram.com/advanceddataanalytics/

ADA on Twitter: https://twitter.com/AdaAdvanced

Resources

Study.EU: Your gateway to universities in Europe, *Why you should study a Masters in Data Science: 3 reasons*, https://www.study.eu/article/why-you-should-study-a-masters-in-data-science-3-reasons

Deepsense.ai, *Why do we need more data scientists and why should you become one?*, https://deepsense.ai/why-do-we-need-more-data-scientists-and-why-should-you-become-one/

Mentionlytics, *5 Real-World Examples of How Brands are Using Big Data Analytics*, https://www.mentionlytics.com/blog/5-real-world-examples-of-how-brands-are-using-big-data-analytics

CIO, *10 data analytics success stories: An inside look*, https://www.cio.com/article/3221621/6-data-analytics-success-stories-an-inside-look.html

University of Wisconsin Data Science Program, *What Do Data Scientists Do?*, https://datasciencedegree.wisconsin.edu/data-science/what-do-data-scientists-do/

Liip Data Science Stack, *Tools in the Data Science Stack*, http://datasciencestack.liip.ch/

Itsvit.com, *Real-Life Business Success Stories Based on Big Data*, https://itsvit.com/big-data/8-real-life-business-success-stories-based-big-data-part-1/

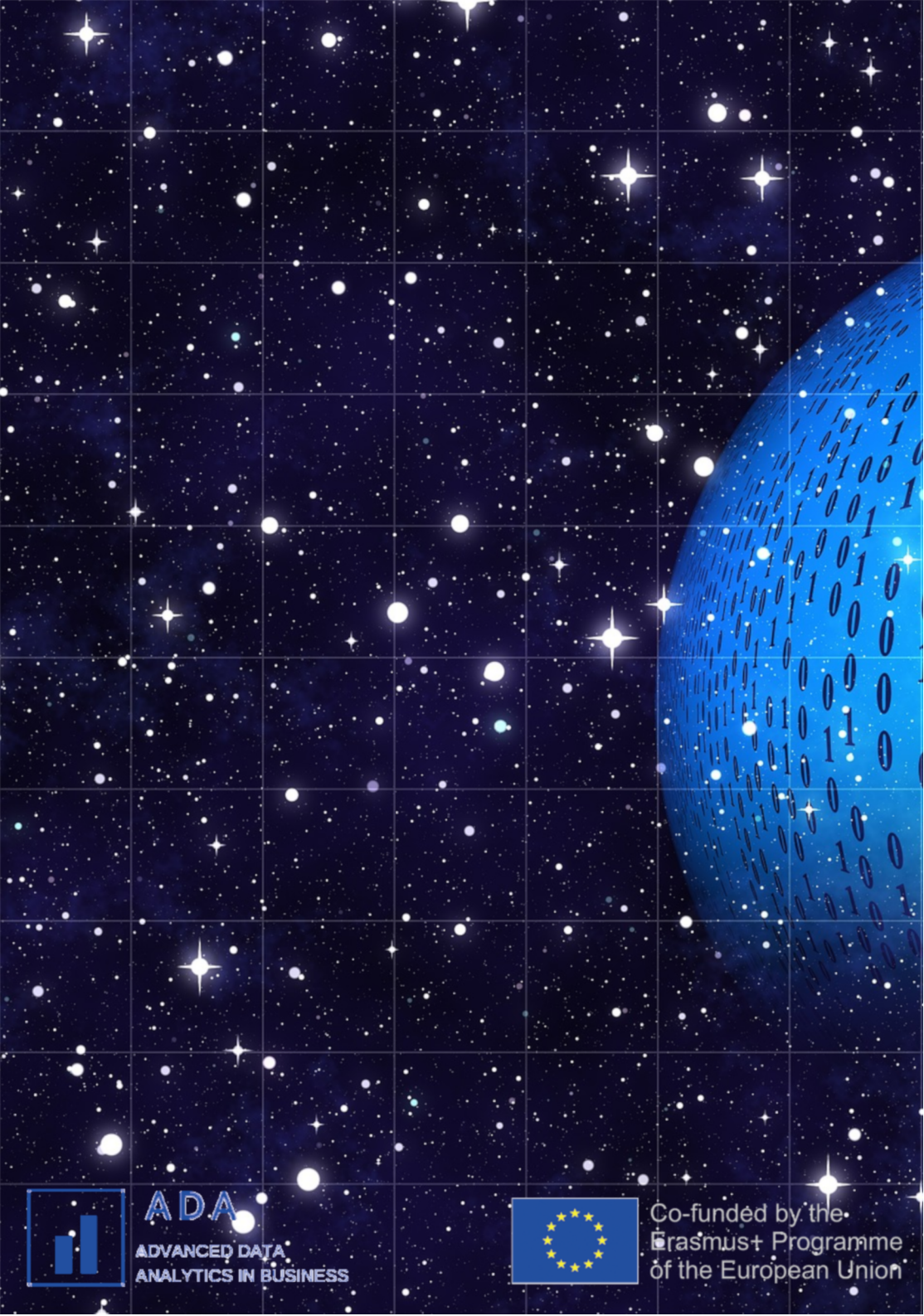
Data Science Community Website: http://datascience.community/

Data Science Graduate Programs, *Data Science Careers*, https://www.datasciencegraduateprograms.com/careers/

KDNuggets.com, *Education in Analytics, Big Data, Data Mining, Data Science, Machine Learning*, https://www.kdnuggets.com/education/index.html

Towards Data Science, *The Third Wave Data Scientist*, https://towardsdatascience.com/the-third-wave-data-scientist-1421df7433c9

Park City Math Institute (PCMI), *Curriculum Guidelines for Undergraduate Programs in Data Science*, https://arxiv.org/pdf/1801.06814.pdf


ADA
ADVANCED DATA
ANALYTICS IN BUSINESS
Co-funded by the
Erasmus+ Programme
of the European Union