Proceedings of the Informing Science + Information Technology Education Conference

An Official Publication of the Informing Science Institute InformingScience.org

InformingScience.org/Publications

June 30 – July 4, 2019, Jerusalem, Israel

MaDaScA: Instruction of Data Science to Managers

Sahahar Golan* Jerusalem College of Technology, Jerusalem, Israel sgolan@jct.ac.il

Dan Bouhnik Jerusalem College of Technology, Jerusalem, Israel bouhnik@jct.ac.il

* Corresponding author

Abstract

Aim/Purpose	Build a program that teaches prospect managers the skills that are relevant for leading data science activity.
Background	Data science becomes ubiquitous in organizations. It is imperative to train students in management departments in the skills that are relevant to this field. Most courses in data science focus on technical knowledge like model building methods, and neglect organizational knowledge such as team roles, ethical considerations and project stages. This work suggests a complementary program that supplies the students with the required knowledge. The authors believe that this program is most suitable for management-students, and that it can also be adapted to software engineering students, in order to provide them with a wider scope.
Contribution	We present the MaDaScA (Managing Data Science Activity) program. The program defines a list of topics that are required for managers' education in order to lead data science activity. This work suggests the content and take-away messages of each topic. The paper surveys several existing courses that teach data-science to managers.
Findings	All existing courses supply a part of the suggested topics, either focusing on technical aspects of data-science or on organizational aspects. In particular, only a small minority of the courses discuss ethical aspects of data science.
Recommendations for Practitioners	We recommend adopting MaDaScA in management departments in order to prepare managers for the challenges in data-science.

Accepting Editor: Eli Cohen │ Received: December 12, 2018 │ Revised: January 30, 2019 │ Accepted: February 2, 2019.

Cite as: Golan, S., & Bouhnik, D. (2019). MaDaScA: Instruction of data science to managers. *Proceedings of the Informing Science and Information Technology Education Conference, Jerusalem, Israel,* pp. 125-140. Santa Rosa, CA: Informing Science Institute. https://doi.org/10.28945/4271

Recommendations for Researchers We recommend adapting the MaDaScA model to the curriculum of the faculty of engineering, especially for the department of industrial engineering.

Impact on Society Educating prospect managers on the capabilities of data science and responsibilities that come with it is key for making sure organizations become much more data driven, efficient and ethical.

Future Research It is possible to make this program more effective by adding practical experience

Keywords data-science, data-science instruction, management

INTRODUCTION

Management of data science teams requires a combination of technical and soft skills. In many cases the team manager role is filled by an experienced data scientist who was promoted. Since the data science team may (and should) have high impact on the organization roadmap, it is important that the team leader have an academic background also in management. One approach would be to include management material in computer science programs. This work suggests a course plan for teaching data science principles as an organic part of management program.

The course aims at giving the manager tools that will enable him to understand the flow of a data science project, the responsibilities of each team member, the challenges that may arise during each stage and ways to overcome them. One important presumption is that the participants have no previous experience in programming. This requires finding creative solutions for presenting the main building blocks of data science activity.

The course is composed of four main topics that supply a holistic view of data science activity. The topics are elaborated in the following sections.

In the first topic, based on Kross, Peng, Caffo, Gooding, and Leek (2017), the course describes the data science team: its structure and its relationships with other parts of the organization. This topic gives the framework in which the other topics take place. The second topic describes common use cases of data science in organizations. These use cases include recommender systems, fraud detection and applications of data science methods to health care, human resources and logistics and maintenance. In the third topic the course delves into the different stages of data-science projects. The fourth topic touches on ethical issues that are involved when practicing data science: privacy concerns, requirements of experiments involving humans, data-driven discrimination and potential misuse of statistics.

This work proposes a method for assessing students based on Project/Problem Based Learning (PBL) and elaborates the different criteria the project must meet. In particular, it presents the value students get from hands-on experience and guidance from the course mentors.

The concluding section surveys a list of existing courses that aim at teaching data-science to managers and show how each focuses on a different subset of the topics presented above. Moreover, the survey shows that some courses emphasize the concrete, data-science oriented topics, while others emphasize the management-oriented ones. The survey shows how the proposed course spans both viewpoints and supplies a wide scope of the relevant areas.

THE DATA SCIENCE TEAM

Data science is a team sport, and it requires multidisciplinary skills. Moreover, data-science team management involves constructing the team, leading the internal dynamics in the team and positioning the team as a key contributor in the organization. The discussion of the data-science team is divided into two sub-topics:

- Building the team: Defining the different roles in the team and their responsibilities.
- Structuring the team's interactions within the organization.

Building the Data Science Team

The course describes the responsibilities and skill-set of the team and map them into three categories: engineering, science and management. Following is the list of responsibilities:

- Infrastructure establishment:
  - Hardware: Servers, Network, Storage.
  - Software: Parallelization,
  - Databases: Access control, Data architecture and query optimization
  - Cloud services
- Software development
  - Implementing data analysis algorithms
  - Model learning
  - Scripting languages (e.g., Python), Statistics analysis languages (e.g., R)
  - Parallelization: Map/Reduce, Hadoop, Pig
- Machine learning techniques
  - Supervised learning
  - Unsupervised learning
- Research skills
  - Data validation
  - Hypothesis generation
  - Statistical analysis
  - Experiment design
- Management skills
  - Team building
  - Empowering employees
  - Communication
  - Presenting results: visualization
- Leadership
  - Advocating Data Science
  - Leading the organization's culture to a data-driven decision-making approach

It is important to note that the mapping of the roles into job descriptions is highly dependent on the size and type of the organization. In particular, in very large and technology-oriented organizations, each responsibility might have a dedicated position. However, it is useful to group these responsibilities into roles that have common characteristics.

The infrastructure and software skills are mapped to the ***data engineer*** role. The data engineer is responsible for supplying the team with the hardware and software resources. He should be able to determine the number of servers, the storage size and hardware connectivity required for parallel processing of big data volumes. This role also requires database-administration skills, since the data should be available for large scale analysis and manipulation, and also should be secure against unauthorized access. In terms of software development, the data engineer should be able to develop data pipelines, learning algorithms, and data modeling solutions. The data engineer role usually comes from a quantitative background with a strong software orientation.

The machine-learning and research skills are mapped to the ***data-scientist*** role. The data scientist is responsible for experiment planning and analysis. He should define the research hypothesis that translates business goals into a scientific task and lead the model definition and optimization. The role requires deep understanding of statistics and machine learning techniques, including what types of models might underlie the sampled data, how to evaluate the quality of a model and what are the meta-parameters that may be employed for improving the learning process. In general, the data scientist has a more mathematical/statistical background that enables him to perform data analysis, but also software capabilities.

The management and leaderships skills are mapped to the ***data-science manager*** role. The data-science manager is responsible for the team's organic functioning and to its ability to collaborate with external teams and add value to the organization as a whole. He should define the priorities of the team, guide and support team members in the (many) challenging stages of the project and make sure that they are not blocked by technical or communication difficulties. An important part of this role is communicating the team's capabilities to the management and other team leaders, and the project's progress. On the highest level, the role includes the responsibility of leading the organization towards a data-driven decision-making mentality and positioning data-science as a valuable tool in strategy definition. The role requires excellent communication skills, the ability to tell a story when presenting a project, the passion to make a difference in the organizations' culture and to harness data science for making well based decisions. Stories are a valuable tool for modern organizations in general and for the information management area specifically. Stories allow organizations to efficiently pass on knowledge within the organization. They improve the organization's abilities to cope with complex situations, they encourage creativity and help attain the organizational vision. Stories are the most ancient tool available to mankind for conveyance of knowledge in a clear manner while emphasizing interpersonal connections, which are at the base of knowledge impartation and preservation. It is natural for the employees and helps recruit them to work toward achieving the organization's goals and fulfilling the processes which will bring about success. The data-science manager should be closely familiar with data science methodology and preferably have hands-on experience in industrial data-science projects.

Building a data science team requires locating candidates that match the required skill set. Interviews of candidates should also be adapted to reflect their skills in data science. A suggested format is to supply the candidate with a toy data-set and ask about what can be deduced from it.

Structuring the Team's Interactions

After establishing the team's structure, it is important to define what the team's role in the organization is and how to structure the team's interaction, both internally and externally, in order to make it effective and meaningful. In each stage of the process there are considerations that are data-science specific and should be taken into account.

The communication within the team should reflect the nature of data science as a field that involves many experiments with challenging situations and a constant need to overcome obstacles. The team manager should hold periodical individual meetings where he should communicate personal goals, expectations, and get updates on progress of the projects and potential obstacles. In addition, there are team meetings, in which the team can be updated on the big picture and discuss priorities. The team meeting can be leveraged to raising infrastructure issues that impact the team as a whole. In addition, this meeting is an opportunity for sharing ideas and milestones that have been achieved. Team meetings should be a place for secure discussion where one can raise new ideas and suggest new directions.

The goal of internal communication is to supply the team members with the supportive environment they need in order to make progress. On the one hand, there should be an open-door policy where the team member can raise quick questions he needs to resolve in order to remove obstacles. On the other hand, there shouldn't be too many meetings that hamper the work continuum. Since in many situations, data science activity may involve frustrating moments, where experiments fail or lead to unexpected results, the manager should use meetings for empowering the team, celebrating success and communicating reasonable expectations.

There are two approaches for structuring the communication between the data science team and other teams. According to the integrated approach, the team members are dispersed among other teams, each acting as a part of these teams. The advantage of this approach is that the team members are perfectly aligned with the organization's business-units' goals and can better aim their efforts to achieve the external team's goals. When incorporating the alternative Organic approach, the team members sit in the same area and collaborate closely in order to make progress. The advantage here is that each team member enjoys the support of his team members, who are familiar with the challenges he is tackling and can contribute to his efforts through brainstorming and advice.

The suggested approach is a hybrid of the two approaches, where on the one hand the team has a shared working space, where the team spends most of its time, closely communicating and collaborating. On the other hand, each team member is assigned a task in an external team, where he can contribute his capabilities in the many tasks that require data analysis, modeling and experimenting. Adopting this approach builds a strong data science team that empowers its members to conduct meaningful research and base their activity on the highest standards. At the same time it positions the team as an important stakeholder in the team's business strategy.

There are several tools that can enable the data science team to help the organization become more data driven. The team can hold lectures educating the employees on data-science capabilities and statistical-thinking. There can be Show-and-Tell events in which the team presents its recent results or case studies describing its activity. New employees orientation processes should include an introduction to the support that data-science can supply to the different teams. The data science team can extract and create new data-sources and make them available and accessible to teams that may benefit from them. The team can also develop tools and technologies that make data much more present in the working process: Dashboards, analysis tools and machine-learning platforms can help promote data-science to a central role in any team's progress.

Adopting the above-mentioned guidelines can help build a data science team that is capable of making a real change in the organization and making data science a leading factor in its core processes.

Data Science Use Cases

Following the description of a generic data-science project, the program delves into a description of several use-cases that are most common and have most potential for future development. For each use case the course describes the following properties:

- The motivation behind the use case
- Main characteristics
- The data that is used to develop a model
- Main methods that are used
- Challenges
- A Case-study

Table 1 summarizes the discussion content for each use case.

Table 1. Example of a table

Use case	Motivation	Characteristics	Data	Methods	Challenges	Case study
Recommender system	√	√	√	√	√	√
Fraud detection	√	√	√	√	√	
Human resources	√	√	√	√		
Health care	√	√	√	√	√	
Logistics	√	√	√	√		

Recommender Systems

The motivation for recommender systems stems from the plethora of options that surround consumers in any step they make. People are flooded by choices of commercial products (online shopping, restaurants), of content (books, movies, papers and even jokes) and social opportunities (collaboration, dating).

The input data for the recommender system learning is composed of ratings of users to items and of user and item profiles. Recommender systems vary by the number of users and number of products, the historical information they have on past transactions and the profile richness they maintain on both products and users. The rating information may be binary (for example purchase history), numerical (for example star-rating) or relative (comparing items) (Schafer, Konstan, & Riedl, 1999). The input might be explicit (solicited from the user) or implicit (click data, dwell time, sequence analysis) (Oard & Kim, 1998).

The trivial recommendation method relies on popular trends and rates each item by its current popularity. This approach does not take into account personalization considerations and can be used as a good baseline for other approaches. Personalized recommender systems are divided into content based and collaborative filtering. For the content-based approach, the course discusses how to build user/item profiles, and how to use user-item similarity and item-item similarity for the recommendation. For the collaborative filtering approach, the nearest-neighbors method and the matrix factorization method are defined. The course discusses the advantages and disadvantages for each approach in terms of the cost of profile creation, cold-start recommendation and computational requirements (Adomavicius & Tuzhilin, 2005; Koren, Bell, & Volinsky, 2009).

The challenges of recommender systems include creating diverse recommendations, identifying individual users from shared devices, user/item cold start and understanding user intent from implicit feedback (Ricci, Rokach, & Shapira, 2015).

This program includes a case study of a recommender system for mobile apps that compares alternative recommendation approaches and analyses the performance of each approach (Jannach, & Hegelich, 2009).

Fraud Detection

The motivation behind fraud-detection systems is presented in several reports that indicate that more than 30% of organizations fell victim to fraud and that the cost of fraud is about 5% of organizations income. Common fraud types involve credit cards, telecommunication, insurance, taxes and employee fraud in the workplace (Levi & Burrows, 2008; PWC, 2018).

Fraud detection stands out from other data-science use-cases, in the fact that it operates in an adversarial setting where the goal of the con-man is to avoid detection. Other characteristics of fraud detection problems are that they require constant adaptation to new behavior and that they involve very un-balanced data-sets, where most of the activity is legitimate (Laleh, & Azgomi, 2009; Phua, Lee, Smith, & Gayler, 2010).

The goals of fraud detection are diverse, and they include identifying as many frauds as possible, while avoiding misidentification of legitimate activity. Possible success metrics for fraud detection systems may include reducing the cost of fraud and misidentification (Stolfo, Fan, Lee, Prodromidis, & Chan, 2000), or minimizing the need for manual auditing of suspicious transactions.

Fraud detection methods can be divided into those learning from past fraudulent activity and trying to identify similar behavior (using supervised learning) and those who identify anomalies as suspicious (using unsupervised learning). The first approach is exemplified by Link Analysis (Bolton & Hand, 2002), the use connections between suspicious entities to discover fraud. For the second approach the course presents Break Point Analysis and Peer Group Analysis (Bolton & Hand, 2001).

HUMAN RESOURCES

Data science can assist Human Resources activity in many aspects. The motivation is to make HR processes more effective, more efficient and fairer. The relevant processes are preliminary candidate sifting, candidate evaluation, and employee assessment.

For potential candidates, the relevant data includes experience (Organization, tenure, role, technical skills), education (Major subject, grades), recommendations, publications (papers, open source content) and areas of interests. It is possible to employ NLP and modeling methods in order to filter multitudes of CV's to find the appropriate candidates. Automatic chatbots can make the communication with the candidates much more transparent and interactive,

The interviewing process is also rich in data. It is possible to weigh scores in many areas and from many interviewers. Moreover, once an interviewer has enough track record, his evaluation can be calibrated and become more standardized. The calibration process can remove any biases, conscious and unconscious and increase diversity. Once the evaluation process is over, it is possible to determine a salary range that matches the candidate's parameters.

For employees, periodical assessment and calibration involves many types of data: self-assessment, peer-assessment and manager assessment, textual, numerical or relative to other employees. In some scenarios, it is possible to analyze the content created by the employee. Here, too, it is important to employ data-driven processes to get an efficient and unbiased assessment (Angrave, Charlwood, Kirkpatrick, Lawrence, & Stuart, 2016).

HEALTH CARE

A recent research states that mis-diagnosis is responsible for 10% of patient mortality and 6-17% of medical complications (Makary, & Daniel, 2016). Data science assists in reducing these numbers and achieving more reliable decision making.

As a scientific field, health care is based on vast amounts of data of multiple types: Hundreds of years of accumulated knowledge, research, medical protocols and clinical experiment results. For a specific patient it is possible to use data of medical history, medical tests (imaging, biochemical, genetic and functional), family relations and demographic origin.

One important use case is early diagnosis of medical conditions that increases the treatment effectivity and chances of recovery. Image analysis applications – such as Ultrasound, MRI, Mammography, Tomography, Colonography, Angiography and X- rays are ubiquitous for diagnosis tasks. An interesting use case diagnoses diabetes using Ophthalmoscopy (analyzing images the retina). Signal processing is used in ECG and analyzing the input from digital stethoscopes (Arnoldi et al., 2010; Doi, 2006; Esteva et al., 2017; Kononenko, 2001).

A second important use case employs data science for developing new medical treatments, and in particular personalized treatments. It is possible to run thousands of experiments simultaneously and efficiently analyze the results (Raghupathi & Raghupathi, 2014).

Another use case is clinical decision-support-systems (Musen, Middleton, & Greenes, 2014). A review of the research in this field has shown a significant improvement in medical treatment when using a decision support system (Garg et al., 2005).The course presents the differences between standalone and integrated systems and discuss how standards are important for sharing clinical decision support content (Wright & Sittig, 2008).

A crucial challenge in data-science based health care lies in the strict regulations that governments enact in order to ensure patient privacy and safety. Another challenge is assimilating new technologies into medical facilities and the plethora of new tools physicians need to learn.

Logistics and Maintenance

Large systems that are composed of multiple parts pose a challenge for maintenance and management. Examples include aircrafts, trucks and civil infrastructure such as water and electricity. In addition, organizations like supermarkets also require efficient systems for managing logistic operation.

Many large systems collect data from sensors that continuously report the state of its parts. Modern technologies advance IoT as a fundamental source of information and GPS and GIS data allow elaborate analysis of location. Such systems use data science to predict malfunctions and for defining preventive care policy. An aircraft, for example, can continually monitor 5000 different parameters during its flight, which is equivalent to 2PB of data. The data can also be used for better evaluation of usage patterns and improve cost estimation. Rolls Royce harnessed data science in order to define a new business model of charging its clients based on the number of hours they operated their engines (Smith, 2013).

Retail stores also manage their supply chain using many sources of data: customer profiles, supplier data, store arrangement data, and stock data all play a critical role in optimizing the costs and profit of the store (Waller, & Fawcett, 2013).

The Data Science Project

The data science project flow is comprised of the following stages:

- Research question
- Data collection and cleaning
- Learning a model
- Model evaluation
- Results presentation
- Decision

The first stage in a data-science project is ***defining the research question*** (Peng & Matsui, 2015). The course describes the different types of questions that are asked in different stages of the flow and highlight the importance of differentiating between the types in order to set expectations and choose the correct methodology. There are six types of questions:

- Descriptive: What are the high-level characteristics of the dataset? This question requires understanding the distribution of the input features and is asked in the initial stages of the project.
- Exploratory: After answering the descriptive questions, the exploratory question is about deeper relationships and patterns that lie within the data. It concerns correlations between input fields and can be the base for the next types of questions.

- Inferential: Asking whether a pattern that was found in a given dataset is also valid for external situations. Answering this question requires information outside the scope of the dataset.
- Predictive: Can one field be predicted given the other field?
- Causal: Is there a causal effect between the values of the fields? This question is different from the predictive question in that it is sometimes possible to predict one field from other fields even if there is no causal connection between them.
- Mechanistic: If a causal connection exists, it might be relevant to ask why the input fields cause the output. Answering this question may require researching the real-world mechanism behind the causal connection (Psychology/Physics/Biology) and in many cases requires domain expertise rather than data-science methodology.

After the questions are defined, the next step is ***collecting the data and cleaning it***. Data acquisition requires either locating the data or generating it. The former may involve data that exists within the organization or from external resources. The latter may involve adding logging devices to existing operation flows or defining new experiments that result in new data. Often, data acquisition requires a data pipeline that combines information from several sources and coverts the data into a standard format. Once the data is available it requires validation and cleaning: taking care of missing values, data duplication and anomalies, converting the measurements to standard units of measurements, normalizing value ranges and grouping values into categorical fields (Rahm & Do, 2000).

Following the data acquisition, it is possible to start ***learning a model***. Since this program does not focus on programming skills, the machine learning techniques are presented using a cloud platform called BigML.com. The platform includes supervised-learning based models such as Decision-trees (Quinlan, 1986), Tree-Ensemble (Sollich & Krogh, 1996), Logistic Regression (Hosmer, Lemeshow, & Sturdivant, 2013) and Neural-Networks (Hagan, Demuth, Beale, & De Jesús, 1996), as well as unsupervised learning based models such as K-Means clustering (Forgy, 1965), Anomaly-detection (Chandola, Banerjee, & Kumar, 2009) and field-association. For each model type, the concept is explained, as well as the major advantages and disadvantages.

Model evaluation methodology is presented, starting with the usage of Train/Validation/Test sets for creating a model, configuring meta-parameters and evaluating the quality of the model. The course describes several evaluation metrics of models: Error rate, Precision/Recall and F1, AUC, RMSE and log-loss. The next step is to describe stratified sampling and cross validation methods. After discussing the methodology for evaluating a single model, the course discusses live experiments and the usage of A/B testing.

Once a model has been trained, configured and evaluated, the project's technical part is over. The data-science manager now presents the project results and main insights that were learned during its execution. Visualization takes a central role of the results presentation, and the course goes over the visual variables that can be used to clarify the presentation: location, size, shape, value (light/dark), orientation, color and texture. The course presents a variety of visualization classes that can be employed when presenting data: For distributions there are pie-charts, histograms, stacked bars, and sunburst charts. 2D plots include colored clusters and bubble charts. For process and flows, Sankey charts and funnels are introduced. Finally some non-orthodox visualizations are presented (for example the WorldMapper project), where the main motivation is to inspire the audience.

INTRODUCTION TO SQL

As part of the Project topic, the course includes a section dedicated to learning basic SQL. SQL is a tool that is used in many organizations to retrieve data and to analyze patterns. It presents a combination of a simple syntax and a wide functionality range. This introduction includes data-types, table-schema, simple/nested queries, filters, aggregation, ordering, joins and grouping. This topic also discusses how to define a table, manipulate its data and metadata and optimize its operation using indexes and views.

ETHICAL DATA SCIENCE

Data science is a powerful tool, and with great power comes great responsibility. It is crucial to include a discussion of ethical issues arising from data science in the program. Moreover, there is an increasing trend in many states that add legal regulations regarding the usage of data.

HOW (NOT) TO LIE WITH STATISTICS

The book *How to lie with statistics* (Huff, 1993) is used to demonstrate some of the most common ways of misusing statistical thinking.

Biased sample: There may be different reasons for getting a biased sample. Some populations may not be available for sampling. In other cases, the subjects do not supply truthful responses or do not respond at all. The person conducting the survey may also introduce bias by hinting at the "correct" opinion.

How to describe a distribution: Definition of average, mean, mode and how each one is not informative enough by itself. What are the statistics required for describing a distribution.

Data Dredging, Causation vs. Correlation: The common misunderstanding or mixing causation and correlation is discussed. Explaining how repeating the same stochastic experiment may result in improbable patterns that do not represent the real data.

Misleading visualizations: This section presents several ways that graphs can give a misleading impression. One example is that changing the graph baseline can cause trends to look more dramatic than they really are. Another example is the usage of 2D and 3D pictograms in order to exaggerate the effect of a comparison.

Percentage misconceptions: Misconceptions on percentages may lead to false notions. When working with percentages, it is most important to keep track of the whole that the percentage refers to. When one deals with a sequence of changes (increases or decreases) in percentages, the whole changes. When the reported change is very dramatic it may indicate that the relevant whole was small to begin with. In addition, when summing parts of the same whole, it is important to make sure they do not overlap.

ETHICAL CONSIDERATIONS

Experiments with humans: In recent decades there are increasingly more regulations limiting experiments in humans. It is imperative that the experiment subjects give their voluntary and informed consent for taking part in an experiment. The experiment should be planned in a way that it reduces the risks the subjects are exposed to (Kramer, Guillory, & Hancock, 2014).

Privacy and Anonymization: The course defines the difference between sensitive and non-sensitive information and shows that many organizations hold sensitive information about their users/customers (Barbaro, Zeller, & Hansell, 2006). Omitting some identifying details may still contain enough information to identify some of the persons. The course presents k-Anonymity (Sweeney, 2002), l-diversity (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006) and t-closeness criteria (Li, Li, & Venkatasubramanian, 2007) that measure possibility to identify and discuss their limitations.

Data based discrimination: Data science can be misused when historic patterns that result from discriminating behavior reflect biased information that supports discrimination. Data driven discrimination can be much harder to fight than "classical" discrimination based on unconscious bias. Two definitions of fairness are given and several ways of identifying and preventing discrimination are presented (Calmon, Wei, Vinzamuri, Ramamurthy, & Varshney, 2017; Hardt, Price, & Srebro, 2016; Fish, Kun, & Lelkes, 2016).

ASSESSMENT

The assessment of the students in the proposed course will be by the implementation of a project. The assessment itself is part of the learning process and can be categorized as Project (or Problem) Based Learning (PBL). This type of assessment allows the students to acquire knowledge via a continuous structured process surrounding an authentic question and concluding with the design of a product that mirrors the learning process. During this process the students deal with creative challenges and in order to attain achievements they must delve deeply into the subjects related to the project. All this, on their own. They must seek relevant material, research it and find answers to their questions.

In this paper, the term 'team' is vital and prominent. Therefore, also in the evaluation stages, it is important that the teamwork be entwined. Furthermore, many of the problems in the use cases are similar and the solution in one case may help solve the other. Towards this purpose, we chose the PBL evaluation method, which allows evaluation of teamwork and helps achieve the following goals: ability to learn new subjects: acquirement of problem solving capabilities; use of knowledge to solve problems; breaking down knowledge into parts; creative and critical thinking; development of an holistic approach to problems and situations; ability to work independently; ability to work in groups - collaborating and improving communication skills. Figure 1 depicts the reciprocal relations among the PBL integrated action circles.

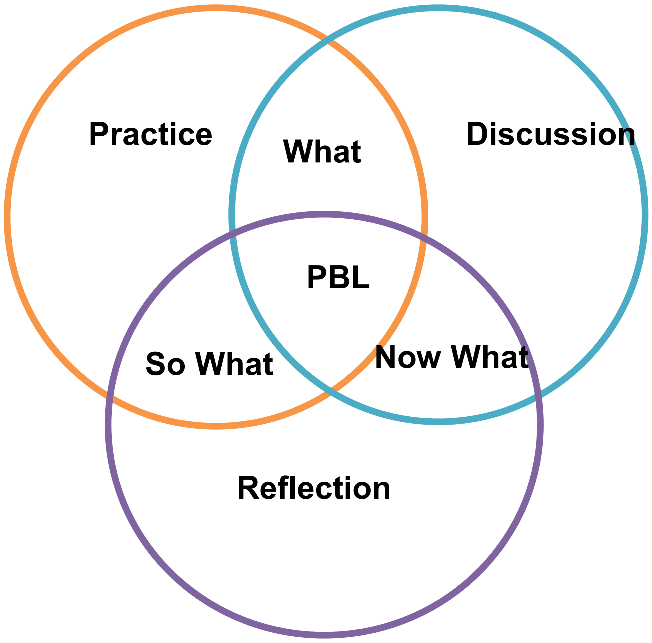

Figure 1: Reciprocal relations among the action circles in the PBL method

Self-feedback, which also exists in the PBL approach, may reveal changes in self-perception or apparent changes in behavior. At the basis of self-guided learning processes, lie reflection processes such as self-reflection, self-judgment and self-reaction (Zimmerman & Schunk, 2001)

It is important to note, that this process is not the end of the learning process. Just the opposite. It is the basis. While working on the project the material is studied in depth. The goal is to allocate a reasonable number of questions to the project and to include all necessary elements upon which the knowledge may be built and established

Evaluation of the project: The final evaluation does not relate only to the final product, but to the whole learning process. The process is based on chronological steps, including presentation of drafts, reflection processes and intermediate feedback.

The project should meet the following criteria:

- AL - Applied Learning – Learning that can be applied to future projects. Team work, interpersonal communication, presentations
- AE – Active Exploration – Learning that demands search and active movement, activity outside the classroom, such as reaching out to the community and specializing in relevant subjects.
- AC – Adult Connections – While working on the project the students will meet an advisor who specializes in the subject that they are researching.
- AR - Academic Rigor – Search for academic material and connection to knowledge acquired outside of the academic setting.
- AP – Assessment Practices – While working on the project the students will be evaluated at each step using various appropriate assessment tools.

This method creates an authentic experience from which they can learn and draw conclusions regarding students who have not experienced complicated situations.

EXISTING COURSES

The following section surveys several courses and programs that aim at teaching data science to managers/executives. Table 1 summarizes our findings. The surveyed courses are:

- Data Analytics for Managers (DAfM): Given by edx.org.
- Data Science for Executives (DSfE): Given by edx.org.
- Data Science for Managers (DSfM1): Given by Naya College.
- Executive Data Science Specialization (EDSS): Given by Coursera.
- Data Science for Managers (DSfM2): Given by Monash University
- Managing Data Science Activity (MaDaScA): The course presented here

The survey checks for the existence of the following components in the course:

- SQL introduction (SQL)
- Data science use cases (UC)
- Data science project: initial steps (PI)
- Data science project: Model creation and eval (PM)
- Data science project: presenting results (PP)
- Building the data science team (BT)
- Structuring the data science team in the organization (ST)
- Ethical data science (EDS)

For example, the course ‘Data Science for Managers’ (DSfM1) includes a discussion on the main use cases of data science and the steps required for creating a model (including the initial steps such as acquiring the data and pre-processing it). In addition to the content that MaDaScA includes, this course also touches upon programming skills and the students learn how to implement algorithms.

While the survey shows that the core components of all programs are about creating, evaluating models (Only DSfM1 does not discuss result-presentation), part from MaDaScA, only DAfM discusses SQL, only EDSS discusses organizational considerations and only DSfM2 discuss ethical considerations in data science.

The columns are arranged so that the more concrete, technical-oriented topics (Use-cases, SQL, project flow) are on the left and more high level, management oriented (team structure, ethical considerations) are on the right. From the table it is evident that the first three courses are focused on the data-science elements of the content, while the remaining are focused on the management components. This is also reflected by the additional subjects that some of the courses include, and the suggested course omits.

MaDaScA balances the two perspectives and gives a wide scope for managers who want to impact the organization's roadmap using data-driven decision-making.

Table 1. Courses contents

Course name	SQL	UC	PI	PM	PP	BT	ST	EDS	Additional subjects
DAfM	√	√		√	√				
DSfE		√		√	√				IoT
DSfM1		√	√	√					Programming, Algorithms implementation
EDSS			√	√	√	√	√		
DSfM2			√	√	√	√		√	History, business strategies
MaDaScA	√	√	√	√	√	√	√	√	

Conclusion

The paper presents MaDaScA program for teaching Data science to potential managers. The course aims at balancing professional knowledge of the process of data science project and organizational knowledge regarding the structure of the team, its role in the organization, and its duties and responsibilities. It is crucial that more professional managers join the data science field, so they may complement the technical capabilities that data science teams have. It is even more important that data-science managers have a solid background in data-driven decision making and ethical usage of the power that lies in big data and ever developing technologies. The combination of strong data-scientists and strong data-science manager may lead to the next level of data science capabilities.

Discussion

There are several ideas that might contribute additional value to this program. It would be interesting to experiment with an interactive session where the student simulates the learning process and tries deriving an intuitive (manual) model from a given dataset, given point after point. Promoting joined projects with experienced and active data-science managers can add value to the students and help them in their first steps in the field. It might be beneficial to add more case studies comparing different approaches. Moreover, it is important to learn also from "negative" case-studies, where methodological caused projects to fail. In general, it would be interesting to add more ways to make the program tangible and keep it in close correspondence with the industry.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, *6*, 734-749. https://doi.org/10.1109/tkde.2005.99

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal*, *26*(1), 1-11. https://doi.org/10.1111/1748-8583.12090

Arnoldi, E., Gebregziabher, M., Schoepf, U. J., Goldenberg, R., Ramos-Duran, L., Zwerner, P. L., ... & Thilo, C. (2010). Automated computer-aided stenosis detection at coronary CT angiography: initial experience. *European radiology*, *20*(5), 1160-1167. https://doi.org/10.1007/s00330-009-1644-7

Barbaro, M., Zeller, T., & Hansell, S. (2006, August 9). A face is exposed for AOL searcher no. 4417749. *New York Times,* p. 8. Retrieved from http://shawndra.pbworks.com/f/A+Face+Is+Exposed+for+AOL+Searcher+No.+4417749+-+New+York+T.pdf

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. *Credit Scoring and Credit Control VII*, 235-255. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.5743&rep=rep1&type=pdf

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 235-249. Retrieved from https://projecteuclid.org/download/pdf_1/euclid.ss/1042727940

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems* (pp. 3992-4001). Retrieved from http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 15. Retrieved from http://www.cs.umn.edu/sites/cs.umn.edu/files/tech_reports/07-017.pdf

Doi, K. (2006). Diagnostic imaging over the last 50 years: Research and development in medical imaging science and technology. *Physics in Medicine & Biology*, *51*(13), R5. Retrieved from https://www.uio.no/studier/emner/matnat/fys/nedlagte-emner/FYS4760/h08/undervisningsmateriale/Diagnostics%2050%20year%20pmb6_13_r02.pdf

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115. Retrieved from http://on-demand.gputechconf.com/gtc/2017/presentation/s7822-andre-esteva-dermatologiest-level-classification-of-skin-cancer.pdf

Fish, B., Kun, J., & Lelkes, Á. D. (2016, June). A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 144-152). Society for Industrial and Applied Mathematics. Retrieved from https://arxiv.org/pdf/1601.05764.pdf

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, *21*, 768-769.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., ... & Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. *Jama*, *293*(10), 1223-1238. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.3830&rep=rep1&type=pdf

Hagan, M. T., Demuth, H. B., Beale, M. H., & De Jesús, O. (1996). *Neural network design* (Vol. 20). Boston: Pws Pub.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323). Retrieved from http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Huff, D. (1993). *How to lie with statistics*. WW Norton & Company.

Jannach, D., & Hegelich, K. (2009, October). A case study on the effectiveness of recommendations in the mobile internet. In *Proceedings of the third ACM conference on Recommender systems* (pp. 205-208). ACM. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.452.3453&rep=rep1&type=pdf

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, *23*(1), 89-109. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.184&rep=rep1&type=pdf

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *8*, 30-37. Retrieved from https://datajobs.com/data-science-repo/Recommender-Systems-[Netflix].pdf

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 201320040. Retrieved from https://www.pnas.org/content/pnas/early/2014/05/29/1320040111.full.pdf

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., & Leek, J. T. (2017). *The democratization of data science education* (No. e3195v1). PeerJ Preprints. Retrieved from https://peerj.com/preprints/3195.pdf

Laleh, N., & Azgomi, M. A. (2009, March). A taxonomy of frauds and fraud detection techniques. In *International Conference on Information Systems, Technology and Management* (pp. 256-267). Springer, Berlin, Heidelberg. Retrieved from https://s3.amazonaws.com/academia.edu.documents/46467288/Laleh-2015-Profile1-Paper2.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1544185945&Signature=4E7c4WRxByvsMDQx9vUoXKjMv%2FA%3D&response-content-disposition=inline%3B%20filename%3DA_Taxonomy_of_Frauds_and_Fraud_Detection.pdf

Levi, M., & Burrows, J. (2008). Measuring the impact of fraud in the UK: A conceptual and empirical journey. *The British Journal of Criminology*, *48*(3), 293-318. doi:10.1093/bjc/azn001

Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on* (pp. 106-115). IEEE. Retrieved from http://www.utdallas.edu/~mxk055100/courses/privacy08f_files/tcloseness.pdf

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, April). l-Diversity: Privacy Beyond k-Anonymity. In *22nd International Conference on Data Engineering (ICDE'06)* (p. 24). IEEE. Retrieved from https://ptolemy.berkeley.edu/projects/truststc/pubs/465/L%20Diversity%20Privacy.pdf

Makary, M. A., & Daniel, M. (2016). Medical error—The third leading cause of death in the US. *Bmj*, *353*, i2139. Retrieved from http://healthofamericans.org/files/Medical_error.pdf

Musen, M. A., Middleton, B., & Greenes, R. A. (2014). Clinical decision-support systems. In *Biomedical Informatics* (pp. 643-674). Springer, London. Retrieved from https://www.researchgate.net/profile/Mark_Musen/publication/226706299_Clinical_Decision-Support_Systems/links/0fcfd5082f6e1def38000000.pdf

Oard, D. W., & Kim, J. (1998, July). Implicit feedback for recommender systems. In *Proceedings of the AAAI workshop on recommender systems* (Vol. 83). WoUongong. Retrieved from http://www.aaai.org/Papers/Workshops/1998/WS-98-08/WS98-08-021.pdf

PWC. (2018), Pulling fraud out of the shadows: Global economic crime and fraud survey 2018. Retrieved 18 December 2018 from http://www.pwc.com/gx/en/forensics/global-economic-crime-and-fraud-survey-2018.pdf

Peng, R. D., & Matsui, E. (2015). The art of data science. A guide for anyone who works with data. *Skybrude Consulting*, *200*, 162.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119. Retrieved* from https://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3. Retrieved from https://hal.archives-ouvertes.fr/hal-01663474/document

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3-13.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In *Recommender systems handbook* (pp. 1-34). Springer, Boston, MA. Retrieved from http://fumblog.um.ac.ir/gallery/1057/Recommender%20Systems_%20Introduction%20and%20Challenges.pdf

Schafer, J. B., Konstan, J., & Riedl, J. (1999, November). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce* (pp. 158-166). ACM. Retrieved from https://s3.amazonaws.com/academia.edu.documents/31095343/recommender-systems-e-com-merce.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1544183998&Signature=ufIz%2BHQMwiGBqApxbGF%2FPKCBGWg%3D&response-content-disposition=inline%3B%20filename%3DRecommender_systems_in_e-commerce.pdf

Smith, D. J. (2013). Power-by-the-hour: The role of technology in reshaping business strategy at Rolls-Royce. *Technology Analysis & Strategic Management*, *25*(8), 987-1007. Retrieved from http://irep.ntu.ac.uk/id/eprint/926/1/214516_Re-shaping%2520Business%2520Strategy_v1.6c.pdf

Sollich, P., & Krogh, A. (1996). Learning with ensembles: How overfitting can be useful. In *Advances in neural information processing systems* (pp. 190-196).

Stolfo, S. J., Fan, W., Lee, W., Prodromidis, A., & Chan, P. K. (2000). *Cost-based modeling for fraud and intrusion detection: Results from the JAM project*. Columbia University New York Department of Computer Science. Retrieved from https://pdfs.semanticscholar.org/7334/806e28edef38aadcc0a52e1b016dfae5fff6.pdf

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(05), 557-570. Retrieved from http://www.cs.pomona.edu/~sara/classes/cs190-fall12/k-anonymity.pdf

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77-84, from https://pdfs.semanticscholar.org/9c1b/9598f82f9ed7d75ef1a9e627496759aa2387.pdf

Wright, A., & Sittig, D. F. (2008). A four-phase model of the evolution of clinical decision support architectures. *International Journal of Medical Informatics*, *77*(10), 641-649. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2627782/pdf/nihms68821.pdf

Zimmerman, B. J., & Schunk, D. H. (2001). *Self-regulated learning and academic achievement: Theoretical perspective*. Lawrence Erlbaum Associates, New Jersey.

BIOGRAPHIES

Shahar Golan, PhD is a lecturer in Lev Academic Center (JCT), in the department of software engineering. Before joining JCT, he worked as a researcher and developer in Google, Yahoo Labs and HP Labs. His main research topics include Machine Learning (Recommender Systems in particular) and Constraint Satisfaction Problems.

Professor Dan Bouhnik is the head of the Computer Science department in the Jerusalem College of Technology. He is the author of a number of books used for teaching Advanced Computer Sciences. In his research he touches upon information security issues from a number of angles: anonymity, privacy, usability, personalization and the awareness level of the user to these issues.