Data Science Workgroup Report
on

Opportunities in Data Science for Bilkent University

Ece Göztepe Çelebi
Savaş Dayanık
Çiğdem Gündüz Demir
Emin Karagözoğlu
S. Serdar Kozat

31 July 2019

Table of Contents

1) Introduction

With recent advances in technology, it is now possible to produce, store, and process huge amounts of data. This facilitates almost all domains to make improvements in their designs and systems, provided that they are able to effectively analyze and employ those data. As a result, data science is emerging as a very important field and finds many application areas in different domains. This report summarizes the efforts and needs in this newly emerged field, from both worldwide and Bilkent's perspectives.

Data science covers a broad range of topics from different disciplines. Thus, different subfields of data science require different levels of knowledge and skill sets. Section 2 focuses on the research and teaching of data science fundamentals. Section 3 is devoted to research and teaching in applied data science in engineering and social sciences. Both sections discuss these activities together with a list of suggestions. Ethical considerations that may arise as a result of engaging data are discussed in Section 4. The advances in data science may also help Bilkent University increase the efficiency of its operational and administrative activities and some opportunities are discussed in Section 5. We conclude with a summary of suggestions.

2) Core Data Science for Developers

In order for developers to propose and design new solutions for the data science problems and also to realize these solutions, they should be equipped with sufficient training. For all these activities, the developers should be equipped with a strong solid technical background, up-to-date knowledge, and advanced skills. The proposal and design parts may require the developers to get more focus on data science research whereas the realization part may require them to get more focus on the use of high-tech technologies. Regardless of what they focus on, there is a very high demand for these developers all over the world, of course, including Turkey.

This section will briefly summarize the research and teaching activities that have been carried out in the world and in Turkey in the perspective of preparing developers. Then, it summarizes the current situation at Bilkent and compiles a list of suggestions for possible improvements.

The design and development efforts in data science are carried out in three main directions. In the first direction, researchers concentrate on designing and developing algorithms, tools, methods, and architectures to collect, store, and process data. This direction involves both work on storage and data processing technologies as well as on computing architectures. Especially, companies such as Google, Amazon, and Facebook invest substantial amount of resources in order to advance technologies for data warehouses, highly accessible storage spaces (e.g., AWS), and highly efficient databases and data access methodologies (e.g., Cassandra, ElasticSearch). Great amount of work is also undertaken on advancing computing technologies in order to provide different architectures as well as hardware for efficient processing of data.

The second direction mainly deals with design and development of new algorithms to process data in order to extract relevant information and produce results for different applications both new and old. Here,

the main goal is to design and develop effective and efficient data processing algorithms, mainly machine learning algorithms, to provide solutions to real life applications. This direction of data science mainly focuses on developing novel techniques in order to generate value from raw data. This part of the data science is currently the most extensively publicized one due to the recent success of deep learning approaches and several different new technologies made possible by them including self driving cars, AI based game technologies, personal assistants such as Siri, Alexa.

The final direction deals with visualization technologies. Here, the main goal is to develop different approaches in order to facilitate access and interpretation of the knowledge generated by the data processing algorithms. Especially, interpreting the information from big data is the most challenging task still unattained in different applications. Hence, a substantial amount of work is carried out in order to develop visualization technologies to make this information readily accessible and interpretable by the naive or even for advance users.

To this end, we next detail both teaching and research efforts in data science in order to successfully meet both the current and emerging needs.

2.1) Teaching

All types of data scientists, regardless of whether they work as users or developers, should know the basic data science techniques and tools. In particular, the data scientists should know what the available techniques and tools are, what they do, how to select them for a given problem, and how to use them. However, in addition to these requirements, the developers should be equipped with a sufficient technical background such that they should also know ***how these techniques and tools work***, also understanding its underlying mathematical and computational foundations. So that, they can understand how to **implement** these techniques and tools and know how to **make further improvements** on them.

To this end, the necessary teaching components can be grouped into the following:

1. Key mathematical and statistical foundations. These correspond to the courses usually offered by the mathematics and statistics departments.

2. Key computational foundations including programming, data structures, algorithmic thinking and algorithm designs. These courses are usually offered by the computer science department. Here programming and algorithmic concepts are to be learned regardless of the selected programming language. As a developer with advanced skills, one is expected to quickly learn other programming languages when s/he has learned the selected language.

3. Data management concepts including how to store data and quick access it. It involves learning the skills related with modern databases (both relational databases and ones to keep unstructured data). This part is covered in the courses related with database management usually offered by the computer science department.

4. Software engineering concepts including how to design large systems, how to follow the necessary documentation and code standards, and how to use version control systems. This

teaching component is especially important for the developers who aim at working in large scale projects. It also needs to cover the principles followed when working in a team.

These first four groups of the teaching components are mainly covered by a typical computer science curriculum. Their main parts are also covered by the CS curriculum at Bilkent. Additionally, the CS department offers the technical elective courses (e.g., related with software engineering) to support the required curriculum.

5. Advanced algorithmic concepts related with more effectively and efficiently processing the data. These teaching components need to cover concepts such as cloud computing, distributed system designs, and parallel processing. They also need to cover advanced computer architecture courses to understand how to advance hardware designs for more efficiently processing the data.

6. Data modeling and assessment concepts. This teaching component includes equipping the developers with the state-of-the-art methods on the areas of machine learning and deep learning. The developers are expected not only to use these methods but also to be able to change them with respect to the given problems. To this end, they need to be taught the underlying mathematical and computational foundations of the machine learning.

7. Data analysis and visualization concepts. This teaching component covers to teach the techniques and tools to understand the data and to get it prepared for further analysis. As a developer with advanced skills, one is expected to visualize high-dimensional data as well as to prepare data using the state-of-the-art methods related with data cleansing, data transformation, and correcting missing and conflicting data. These issues have been becoming more important since many of today's applications require working on data with noise and large variance.

These three groups are mainly covered by technical electives offered in different departments. At Bilkent, such kinds of courses are mainly offered by the CS department at the undergraduate and graduate levels. Additionally, the EE and IE departments offer some courses especially related with machine learning.

8. Soft-skills. The teaching components need to improve written and oral communication skills as well as presentation skills. They also need to prepare the developers to be good listeners to understand the client needs and to effectively communicate the non-technical people.

9. Ethics. The developers need to learn their ethical responsibilities when they develop a technique and a tool and/or when they use data especially containing private and confidential data. They also need to acquire the ability to identify "junk" science and algorithmic biases.

These last two groups are covered by the support departments at Bilkent. Some parts of them are also covered by additional assignments given in almost any of the courses offered at Bilkent.

The following are the suggestions for possible improvements in teaching at Bilkent. As also indicated in many reports (prepared by different consortiums), one of the most important challenges here is to find necessary faculty to teach the existing and new courses and also to find the necessary TA support. The difficulty comes from the fact that there is high-demand for such personnel and the supply is not sufficient for this high-demand.

2.1.1) Short Term Suggestions

- Many of the aforementioned courses especially related with the basic concepts have already been offered at Bilkent. The contents of these courses may be updated such that in their last week(s), more recent concepts may be revisited.

- There are many machine learning related courses offered at Bilkent. They are technical electives and offered both at the undergraduate and graduate levels. The contents of these courses may be revised and updated if necessary (they may have already been updated for this purpose).

- The TA support is very important especially for designing courses that include hands-on projects. However, there is high-demand for data scientists (generally speaking, computer scientists) from the industry. As the salaries of the TAs are much much lower than the ones given in the industry, it may be challenging to find the necessary amount of TAs. Bilkent may think of giving bonuses to the TAs that support such kind of courses.

2.1.2) Medium Term Suggestions

- Incentives may be given to the faculty for offering new technical elective courses. The teaching load of the faculty may be decreased when s/he wants to design and offer a new course. This may be given especially when the faculty designs this course from scratch.

- There are different machine learning related courses that are currently being offered at Bilkent. The number of other courses related to the different aspects of data science is relatively low. These courses may include the topics related with data visualization, data cleansing, cloud computing, data security and privacy, parallel computing, and distributed computing. The number of such courses may be increased.

2.1.3) Long Term Suggestions

- As mentioned before, to increase the number and variety of the technical elective courses focusing on more advanced topics, the number of the faculty is the most important factor. Thus, it is important for Bilkent to hire new faculty working on these topics.

- There are many applications that require the use of data science principles. Although they use the same principles, they may necessitate specializing on a particular topic. Technical electives may also be offered for a particular application area (e.g., deep learning for natural language processing, deep learning for medical images, and generative deep learning models). Of course, here the number of the faculty also becomes an important factor.

2.2) Research

We next detail the research efforts in three main directions of data science. Here, for each research direction, we first provide a brief overview of the current situation as well as the trend worldwide, in Turkey and at Bilkent. We then provide our suggestions in order to meet the challenges in data science research and how we can handle them.

2.2.1) Research on data collection, processing and storage technologies:

Worldwide

With the recent advances in sensor technologies, internet and other tools, we now have access to huge amounts of data, which are leveraged for different applications. Those big data are defined by velocity, volume and versatility. As an example, Google crawls and indexes over 30 trillion web pages, where such a huge amount of information is readily accessed by millions of users in real time. The amount of data continues to increase in variety, volume, and velocity. Hence, there is a significant effort in order to develop technologies in order to store, effectively access and process data. Currently, the main research effort is fueled by the leading US tech companies including Amazon, Google, and Facebook with a different database and/or search engine invented yearly. Especially, distributed and parallel computing and storage software architectures are the main focus in order to process huge amounts of data. Research on storage and access technologies are mainly conducted in the mainstream and well established software domain, where the contribution of AI or AI related methods are limited.

Parallel to research on storage and processing technologies in the software domain, there exists substantial effort in the hardware domain in order to increase computing and processing capabilities tailored to big data. Especially, the GPU hardware is currently the main focus, where most of the deep learning architectures are running on. Most of the super computers with dedicated GPUs are located not in the US but in China showing the intense recent interest in producing such advanced hardware.

The cloud based approaches incorporate both hardware and software technologies in order to effectively and efficiently answer the challenges in big data processing. Here, both the hardware and software resources are dynamically dedicated to different users reducing downtime significantly, while also providing data security. Currently, 90 percent of all commercial AI based model training activities are performed on the AWS, where this number is expected to increase. The AWS is now extensively used for both data storage and processing as well as providing service to different customers. Companies such as IBM are following the footsteps of Amazon.

Research in Turkey

Research on big data technologies are mainly carried out in universities and government research labs in Turkey. There exist several well established research groups that work on parallel computing architectures. Parallel computing architectures such as Truba are run by the government and provide services to research activities. Several state universities including ODTÜ, İTÜ and Gazi University have a

wide range of computing facilities and provide services to other universities although less easily accessible compared to the government operated facilities. However, research on distributed architectures or databases are limited compared to the parallel computing. The effort in this field is more on system building rather than on novel research and mainly conducted in relatively big tech companies.

Cloud architectures are also focus of different research agencies both private and government supported. As an example, B3Lab conducts research on cloud technologies as well as provide support to build distributed big data architectures. The objective is to gather computing powers of different entities in order to build a private cloud to meet data processing needs similar to AWS, however, the data will be stored in Turkey.

Research on hardware technologies are limited in Turkey. There are a few companies such as Aselsan that produce ICs, however, this effort is not focused on production of parallel computing or GPU hardware.

Research at Bilkent

Parallel computing research is extensively carried out at Bilkent University in the Computer Science Department. Several faculties are involved in high performance computing. However, hardware design or research on cloud computing is limited.

Suggestions

- In order to extend our research on computing technologies, we need to extend our computing resources. The computing resources at Bilkent University is limited unlike other universities such as Koç University and/or Gazi University, which have significant computing resources.

- The resources to construct such an infrastructure can be supplied through different resources such as TÜBİTAK or T.C. Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı through university wide collaborative projects, which needs to be actively supported and endorsed by the upper management of Bilkent University.

2.2.2) Research on design and development of new algorithms to process data in order to extract relevant information

Research Worldwide

Recently, the data processing or the information extraction part of the data science has become the most publicized part due to recent success of the AI algorithms especially deep learning technologies. After the spectacular success of deep learning based vision algorithms that surplussed human level performances in different machine learning tasks, we now have machines that can diagnose specific diseases better than the doctors on average. As shown by the recently announced performance of the AlphGo, the autonomous machines can learn very complex games and beat expert human players. Due to these

significant advances, applications previously considered as science fiction such as self driving cars, humanoid robots and swarm bots are now available for service.

This part of the data science research involving AI and machine learning is the main focus of research in different countries and companies in the world. The research on AI is not only carried out by big tech companies but also extensively supported by governments and universities since applications not only involves commercial side but also security and defense. As an example, most of the cyber security research is recently shifted to developing AI based algorithms in order to detect and eliminate cyber attacks. Clearly, these algorithms will be used in different applications including surveillance, building robot armies (recently announced by the US government) and similar defense applications. MIT announced that a new faculty is coming solely focused on AI research and development. Chinese government invests enormous amounts of resources to advance their AI technologies and see this field as their main focus in technology strategy.

Research in Turkey

AI and machine learning are extensively studied in Turkey both in universities, private companies and government labs. We have several different research labs that solely work on AI and AI applications especially in vision and natural language processing. In these labs, state-of-the art image processing, speech processing, and text processing algorithms are developed. The main research trust is coming from the defense and security industry.

Research at Bilkent

Machine learning is one of the main focus of research at Bilkent University. The AI research is carried out in different domains with different real life applications. We have both very successful vision groups that work in different applications such as medical and defense domains. There exist also theoretical studies in this field.

Suggestions:

- Research on machine learning algorithms especially on deep learning requires specialized hardware, i.e., GPUs. This kind of hardware is essential for training and testing different algorithms and is usually very expensive. University wide dedicated GPU servers are essential in order to advance our research output in this field. The deep learning field is one of the main research fields where Bilkent University can lead the research and development efforts both Turkey and produce significant international recognition.

- Researchers in this field are paid very high salaries compared to the more classical fields in both Turkey and world wide. We see the effects of this situation currently in Turkey as the migration of deep learning researchers as well as our students to EU and USA. The current salary of our graduate students is significantly lower than the current salaries of the ML practitioners in Turkey. Such huge discrepancy greatly diminishes both the quantity and quality of our research outcome. Hence, strong and immediate measures are needed to rectify the current situation.

- The research in this field is usually application oriented, hence collaborations with tech companies are essential. The collaboration effort is carried out by joint research projects. The overhead policies for different projects including hardware, student and PI salaries are not

regularized and can change even after the budget of the project is determined. The overhead in certain components such as for the PI salaries greatly diminishes the motivation and incentive of faculty to perform such projects.

- Researchers and faculty in this field are highly sought after. Hence, in order to attract world quality researchers and faculty, we need to offer substantial salary packages.

2.2.1) Research on data collection, processing and storage technologies

Worldwide

Research on visualization technologies are mainly driven by the US and Asian companies, especially with ones dealing with big data and gaming technologies. The companies see the visualization techniques as the main component of creating value from the big data collected from different sources. Especially graph technologies are the main focus in this field. For this purpose, both dedicated hardware and software technologies are developed.

Research in Turkey

The research on visualization techniques are mainly carried out in the private companies in Turkey. There is currently a significant effort especially in IT vendors in order to analyze and visualize huge amounts of data generated by the telecom customers. Different technologies and algorithms are developed. In parallel to big data applications, there is a significant effort in gaming companies to generate better visualization approaches. Although, such approaches are tailored to generate graphics, the same approaches are currently used in map technologies.

Research at Bilkent

The research on visualization technologies are limited at Bilkent University.

Suggestions:

- The research in this field is usually application oriented, hence collaborations with tech companies are essential. The collaboration effort is carried out by joint research projects. The overhead policies for different projects including hardware, student and PI salaries are not regularized and can change even after the budget of the project is determined. The overhead in certain components such as for the PI salaries greatly diminishes the motivation and incentive of faculty to perform such projects.

- Researchers and faculty in this field are highly sought after. Hence, in order to attract world quality researchers and faculty, we need to offer substantial salary packages.

3) Data Science Applied to Engineering and Social Sciences

With the advances in data collections, for any real applications, one can easily obtain large amounts of data. Dealing with large data in applications brought new challenges. We summarized challenges in research and teaching in two separate sections after consulting to the findings in the reports [1] presented to the National Academies of Sciences in the United States in 2013 and [2] prepared by McKinsey Global Institute on the industry-wide impact of big data in 2012.

3.1) Research

Google, Yahoo, Microsoft produce data in exabytes (bytes). Facebook, YouTube, Twitter have hundreds of millions of users. Traditional methods do not scale up to deal effectively with such big data. Data mining of massive data changed the way we think about crisis response, marketing, entertainment, cyber-security, national intelligence, storage and retrieval of data.

Collections of documents, images, videos, and networks are potential sources of discovery and knowledge. Their processing requires techniques that go beyond indexing and counting keywords. Those data need to be mined for relations and semantic interpretation.

Mining big data is likely to help science (astronomy, biology) extend each reach. Technology will become more adaptive, personalized, robust. Individual genomics, cellular, environment data may be stored. Science can then make more effective personalized treatment recommendations.

Businesses can observe and mine the individual preferences, learn the details of the specific good, skills, and service in need, and create new markets.

High-impact applications can be built in the near future with the help of parallel, distributed, cloud computing and database management systems. New challenges for massive data go beyond storage, indexing, and querying (classical database, search engine, information retrieval systems) and hinge on the ambitious goal of inference (turning data into knowledge).

Research Challenges

One challenge in data management is tracking the origin of data, from generation to preparation, validating data, working with different data formats and structures, and enabling data discovery and integration. In data analysis, coping with sampling biases and heterogeneity is a big challenge. Data may not represent the population. If data were assembled from various sources, then fields may differ.

Fast amount of data necessitates developing algorithms for parallel and distributed architectures, new methods for visualizing massive data, scalable and incremental algorithms to cope with the need for

real-time analysis and decision-making (product recommendation, quick re-routing of jobs in the face of random unexpected disruptions in supply chains or sudden shifts in demand due to economic sanctions or government incentives). On the policy-site, ensuring data integrity and data security and enabling data sharing are the other big challenges.

As big data are turned into knowledge, statistical rigor is necessary to control the errors in inferences. However, statistical principles may not immediately apply to massive data. For example, sampling may have introduced bias. Data may have been collected according to some criterion (e.g., in a way that favors “larger” items over “smaller” items, or through online surveys may exclude population without access to the internet), but inferences may refer to a different sampling criterion. This is especially a severe problem for massive data because data consist of collections sampled under different criteria.

Data may not be original; for example, missing values in the original data may have been filled at some intermediate step. This can result in circularity and underestimated variance.

Controlling the errors will be difficult when many hypotheses are considered. As data size grows, more patterns are expected to be found. This translates to more hypotheses to be tested, but the law of large numbers may not then apply to error rate estimation.

Existing statistical theory may not apply if the sampling assumptions were violated during the assembly of mass data. The statistical methods may not scale up because those methods may need large computational resources themselves. For example, nonparametric smoothing algorithms need to access the entire data set every time response has to be estimated at new regressor values. Exploratory analysis often use nonparametric methods, and it is a challenge to explore big data with traditional methods.

Humans should still be kept in the loop because “knowledge” is often subjective and context-dependent. Human intelligence does not seem to be replaceable by machines in the near future.

On the one hand, data/computer scientists needs deeper awareness of inferential issues (bias modeling, error rate calculation). Assertions of knowledge require control over errors. “Knowledge” can turn into garbage if error rates are not under control. In massive data, there is limited information about each individual. Heterogeneity among individuals lead to an explosion of hypotheses. Errors become more difficult to control.

On the other hand, statisticians should worry about scalability, algorithmic issues, real-time decision making. Control over statistical error also demand for computational resources. Assumptions of classical statistical methods are likely to break down. For the stream-lined data, data cannot be stored for a retrospective analysis.

Mathematicians should work on scalable applied linear algebra, optimization algorithms that can in turn be used by the data scientists and statisticians to quantify and control the false discovery rates.

Social scientists should work on psychological, judicial consequences. Data privacy is an issue. Certain combinations of values may identify people. Anyone dealing with data should have technical skills in both computational and statistical thinking and awareness of privacy, ethics at risk.

On the technical side, we need benchmark data, repositories of data and software for comparisons. Computational infrastructure is needed to train the next generation of “data scientists,” to help

researchers meet real-world massive data problems. Significant new ideas emerge only if researchers play with real-world data.

Massive data analysis cannot be reduced of a turnkey procedures that consumers can use without thought. New generation of engineers need mass data management skills including modeling decisions, approximation skills, and attention to diagnostics and robustness.

3.2) Teaching

As McKinsey Global Institute [2] indicates, one of the major challenges we face is the shortage of talent. More precisely, data are abundant but there is a deep shortage of people with necessary skills in big data statistics, machine learning, AI and other relevant tools. To answer how this challenge and others regarding the effective utilization of data can/should be tackled, the same report lays down action plans at various sophistication levels. Some of those involve making people realize the benefits of using big data, placing data into standard forms, making data available, applying basic data analytics, and utilizing advanced analytics. We believe that contributing to the elimination of shortage of talent through university education requires a multi-layered approach, as well.

3.2.1) Current Situation in Turkey

Developments in Turkey are relatively recent. İstanbul Technical University, Middle East Technical University, Sabancı University, TOBB-ETU, Hacettepe University, Bahçeşehir University, Gazi University, Yeditepe University, TED University, Fırat University, İstanbul Şehir University, Akdeniz University, and MEF University are the universities in Turkey, which invested in Data Science or related MSc programs and/or data science labs/research centers. Most of these MSc programs are non-academic and aiming at working professionals rather than full-time graduate students. In addition, their curricula are more oriented towards business (e.g., finance, marketing, operations management) applications rather than hardcore engineering or computer science. Some universities offer MSc degrees with or without thesis. METU is the only university, which offers a PhD degree (on scientific computing with a specialization on data science).

As for research centers/labs, the list is, again, short:

- Bahçeşehir University started “BAU İstanbul Big Data Eğitim ve Araştırma Merkezi” with funding from İstanbul Kalkınma Ajansı and TC Kalkınma Bakanlığı,
- Sabancı University started “Center of Excellence in Data Analytics,”
- TOBB-ETU, in collaboration with IBM, started “IBM Büyük Veri Analiz Laboratuvarı,”
- Gazi University has the “Big Data and Information Security Center Research Lab,”
- Fırat University has “Büyük Veri ve Yapay Zeka Laboratuvarı”,

Undergraduate programs in data science:

- Hacettepe University Department of Artificial Intelligence

- TOBB University Department of Artificial Intelligence

Graduate programs in data science

- Hacettepe University, Veri ve Bilgi Mühendisliği Tezsiz Yüksek Lisans Programı
- İTÜ, Master of Science (One Year) in Big Data and Business Analytics
- METU, Master of Science in Scientific Computing (One Year without Thesis, Two Years with Thesis), where one of the specialization areas is Data Science
- Sabancı University, Data Analytics (One Year) MA Program
- Yeditepe University, Master of Science in Data Science (with or without thesis)

3.2.2) Suggestions

Our short, medium, and long-term suggestions below will reflect this approach.

3.2.2.1) Short Term Suggestions

Here, we list our teaching-related suggestions that can be implemented with a reasonable effort in the short run.

I. **University-wide Interdisciplinary Course:** One such suggestion is designing and offering a university-wide, interdisciplinary course on data science and related topics. This course should be designed and offered jointly by faculty members from various disciplines such as computer science, electrical and electronics engineering, industrial engineering, psychology, economics, law, management, and philosophy. Each faculty member will lecture 2-3 weeks on the importance and usage of data-driven tools, analyses, applications, and research questions. The main objective of this course will be raising awareness about the value from adapting a data-centered approach, and triggering interest about current developments in data science as well as its reflections (or repercussions) on other disciplines, rather than providing technical skills (e.g., programming, coding skills).

II. An introductory level data analysis course, preferably designed and offered by non-CS departments according to their needs. The targeted audience of this course will be non-CS students that will have taken an introductory level programming course (for example, CS 115). This course will aim to equip students with basic skills about how they apply their knowledge on programming to real-world problems which require the use of big data. It will also cover how to access and manipulate big data sets. Needless to say, offering such a course will require additional teaching staff and assistants.

III. **Data Analytics Course for Social Sciences.** An applied course on qualitative data analysis with MaxQA, Atlas will be helpful to undergraduate and graduate students in undertaking more effective and in-depth analyses of their subject matter.

IV. Non-CS students who are willing to work on the applied data science can be encouraged and better quided. For undergraduates, this guidance includes compiling a list of elective courses that they should take, informing them which prerequisites they should satisfy to be successful in these courses, and giving a lecture series on the applications of data science. This may also include organizing seminars given by people from the business world. For graduates, this may include offering them theses on interdisciplinary topics. Such a thesis may be joint-work between two (or more) engineering/physical sciences departments or joint-work between an engineering/physical sciences department and a social sciences department. To give incentives to such interdisciplinary theses, Bilkent may give initial funding to the related faculty to start such joint-work and/or may offer scholarships for attracting students to work on these interdisciplinary theses.

V. **Business Analytics Program.** Business now collect abundant scan data after every purchase of their customers, hold data on promotion dates, prices, and sales. They would like to develop effective promotion strategies. Promoted prices may increase the sales of targeted products, but reduce the total profits because of cannibalization of demand for substitutes. Analysing the market baskets, addressing the heterogeneity of customers with advanced statistical models are indispensable for finding the effective promotion strategies. Online stores face the challenge of designing individualized blend of products to be shown to their arriving potential customers. Most of those problems have their roots in understanding, modeling, measuring the customer choice among alternative products. Online auctions for search-based advertisement, revenue/yield management in airlines, hotels, sports events, gas and electricity pricing in deregulated markets, individualized healthcare management depend on vast data produced by the transactions in the relevant markets or records kept by the key institutions (exchanges, healthcare institutions) require good command of statistics, optimization, and data wrangling skills. A business analytics program to be jointly developed with Industrial Engineering, Management Science, and Economics Departments can raise students capable of producing high quality solutions for all of those new and exciting problems.

3.2.2.2) Medium Term Suggestions

I. **Curriculum Changes:** Our first medium-term suggestion can be considered as a follow up on our short-term suggestions. We suggest that –especially- in departments such as economics, psychology, political science, management, industrial engineering, law, and philosophy, where there is a lot of room for research on applications of machine learning, deep learning, artificial intelligence or essential/core questions about research on data science (e.g., ethics, privacy, philosophy), new faculty with an expertise/specialization on data science, big data, machine learning, and artificial intelligence should be hired, so that data-centered approaches using state-of-the-art toolboxes can find a spot in the curricula of these departments. We believe that there are important synergies to be exploited if the strategic plan is executed at the departmental level as well. Otherwise, unused knowledge will perish, and the fruits from aforementioned general-audience courses and programming courses will not be effectively realized. Needless to say, incorporating such courses to the curricula of various departments and hiring new faculty members with expertise in these fields will have valuable side-benefits on research collaborations involving faculty members from different disciplines; we think that generating that synergy will

create a significant value added.

II. **Improving the Education Experience:** Major improvements can be made in the education service Bilkent provides to its students. Standardizing, preparing, and using historical data from thousands of students taking the same set of classes over the years, the university can provide personalized recommendation systems to students as well as individual mentoring prepared by AI, which uses the aforementioned data. By computerizing most assignments and importing exam papers to an online platform, students' particular weaknesses can be specified during the semester and they can be assigned extra exercises in a timely fashion to improve in those dimensions. Another extension of efforts in this direction could be the utilization of AI-supported grading algorithms, which could minimize grading lags and reduce faculty member's as well as teaching assistants' burden. Needless to say, moving assignments and exams to an online platform will require non-negligible infrastructure investments, which is why we classify our suggestions here as not short but mostly medium (and to some extent long) term.

III. Every discipline now works more closely with data. It is important that both faculty and students can reach, store, and handle data with modern data analytic tools, most of which are brought home by the packages in Python, R, among other open-source tools. Bilkent can help everyone learn how to access and use those tools by reorganizing its computational infrastructure amenable to modern applied computational needs:

- Abandon teaching Windows, MSOffice, Matlab, Minitab, SPSS
- Move to open-source software
    - **Linux/Ubuntu desktop and server:** Linux was designed for handling high-end computation, which is what everyone is trying to do even on a laptop nowadays. With the user-friendly completely free, and very reliable desktop versions, the transition from Windows to Linux environment can be made very smoothly. MS Office users can start using Libreoffice, and within five years, all university offices are likely to be sharing Libreoffice documents without any problems.

    - **SQLite, PostgreSQL for database applications:** data are so abundant that even data received from Turkish companies for the senior student final projects do not fit in most up-to-date Excel spreadsheets. Excel spreadsheets do not allow indexing data rows with respect to various variables for fast sort, search, join operations. For effective data analysis, faculty and students should turn to modern database management systems. Even for visualization, R developers, for example, are building new tools that complete the computations on the database server side in order to speed up the renderings of the plots.

    - **Octave and Python for scientific computing**

    - **R for statistics**

Switching to open source is beneficial. On Matlab/SPSS developments, only core Matlab/SPSS software engineers work. However, on Octave/Python/R, more people, especially the end-users, the true owners of problems, work on the development. A lot of feature requests are implemented in parallel by people who need them. Because the source codes are open, with a slight twist or improvement, the existing software can be quickly adapted to work with a new algorithm. This helps applied research progress significantly faster. There is an abundance of experience shared

across sites like Stack Overflow and Scikit-learn.org. After graduation students can transfer what they learn at Bilkent to working environment easily and fast because open source is accessible from everywhere.

4) Ethical Concerns Invoked after Wide-Spread Use of AI

As a sub-branch of philosophy, ethics is concerned with evaluation of human behavior. Distinction between good and bad, human dignity, autonomy of individuals and free will are the main questions that ethics tries to find an answer. In this respect, it would not be wrong to say that ethics is not at first instance interested in the “rules” but in the effect and value of rules. AI algorithms, on the other hand, despite foreseeing a certain rule, do not include any value or ethical norm. For this reason, the relation and boundary between what is technologically possible and what is ethically permissible, emerges as a question that is likely to constantly accompany technological developments. For example, as a warning example in terms of the relation between ethics and AI, Emeritus Prof. of AI and Robotics Noel Sharkey, who is among the pioneers of AI research, cites the fact that one of the first products that emerged after the production of robots capable of emotional simulation is robots for pedophiles in the size of children. Considering that pornography sector is at the top with respect to AI, it is apparent that the data collected, their areas and purposes of use, should be subject to a critical discussion.

The principle that should be agreed upon, in the relation between AI and ethics, is that technology is not a purpose in itself and it is a tool that serves a certain purpose. So, the researchers should be mindful of the knowledge that smart machines are working tools for people. It is in the power of people to ensure that these machines do not hurt others, do not discriminate or do not destroy living things. The crux of the matter in AI research is that the relation between these algorithms and the creative power of science people do not get disconnected at some point.

There is slowly increasing a consciousness as to including “digital ethics” in the curricula of engineering faculties. Oxford University assumed a leading role by including „Digital Ethics Lab“[1] course into its curriculum.

4.1) AI, Transparency and Danger of Misuse

Transparency of the algorithms within the scope of AI, constitutes one of the most important dimensions of AI-ethics relation. With what purpose are the algorithms developed, after what kind of evaluation is data entered, which prejudices have had an impact on the development of the algorithm are very important questions. For this reason, many authors warn that, if it cannot be answered “why” an algorithm produces a certain outcome, then it should not be used especially in sensitive areas (security, privacy of people etc.). They especially make warnings as to the relation between AI and ethics, with respect to generation

[1] https://www.oii.ox.ac.uk/research/digital-ethics-lab.

of sexism, racism and xenophobia. For instance, who would be recognized as “human” or “beautiful human” in face recognition programs, is related to the data entered into the algorithm to be developed. The examples of classification of black people as “gorilla” instead of human being in a program developed in the USA; or classification of only Western Civilization artefacts and disregarding of other art works in the world while transferring world cultural heritage to digital medium, demonstrate how decisive ethical preferences are at the stage of development of algorithm.

Cathy O’Neil, a mathematician from the USA, classifies algorithms in four levels according to their degree of vice and states that each of the arising results are consequences of prejudices and subjective decisions that in fact, people carry inside them. At the top of the list, there are the algorithms that are qualified as vice because of the data inserted which created “unintended” stereotypes. For example, the systems that show more baby sitter positions to women seeking jobs and more manager positions to men. Secondly, there are algorithms, which create bad results due to negligence. For instance, in the case of software that fail to arrange the working hours of part-time workers at an optimum level and prevent these people from attending night school or spending more time with their families. Third is the algorithms that especially advertising industry uses and find out when the consumers are inclined to spend more money, when do they spend more money emotionally. Finally, O’Neil gives the example of illegal algorithms or algorithms that serve the purpose of surveillance and control over people[2]. Whereas the criteria of these four categories are open to discussion, they are meaningful in terms of demonstrating that each of them arise from a “human” choice at the beginning.

In order to prevent misuse of algorithms, a supervision platform (https://algorithmwatch.org/en/) has been established. The purpose of this organization, even the name of which reminds the rooted human rights organization Human Rights Watch, is to show the scope of human rights extending with AI. As a matter of fact, Human Rights Watch initiated an international campaign titled “Campaign to Stop Killer Robots” in 2013 for prohibition of autonomous arm systems and as of today more than sixty NGOs support this campaign[3].

Additionally, the fundamental rights and freedoms of individuals are tried to be adjusted to digital world in parallel to technological developments. With an initiative launched within the European Union (https://digitalcharta.eu/sprachen/), technological developments are addressed together with digital rights.

4.2) AI and its Effects on Political Orders

The relation between AI and ethics is actually the version of problems of history of science and science ethics adapted to new technologies. The effect of scientific developments on the needs of society, whether the contribution of technological developments to authoritarianism of democratic political orders is problematic, or to put it differently, the question of whether science can be “unbiased” in real sense, are permanent questions that accompany technological questions.

[2] Cathy O’Neil (2017): How can we stop algorithms telling lies?, The Guardian (16.07.2017).
[3] In this term, the “UN Convention on Certain Conventional Weapons” has been revised for the purpose of including also the autonomous weapons.

In this respect, it is very important to determine, within an ethics discussion, the limits of a technology that aims to abolish democratic political orders of developed modern societies or even if does not clearly have this aim, that supports authoritarian/totalitarian political aims. “Social credit system” that was put in operation in China (its first commercial application is “Sesame Credit” system), making elector focused manipulative news selection in the USA elections or prevention of access to certain information by governments are just a few examples.

4.3) Relation between AI, Science and Commercial Profit

In courses and projects concerning AI, there should be open discussions as to what kind of a power that the knowledge generated by science, may turn into in the hands of firms that work solely for the purpose of commercial profit. The potential dangers of knowledge power that is held by these firms which are not transparent and which hold vast amounts of data despite not being open to public inspection, is an important phenomenon that the researchers should be conscious about.

For the purpose of ensuring that the firms act responsible regarding AI and ethics, the platform “Partnership on AI” (https://www.partnershiponai.org/) has been established with the participation of research institutes, NGOs, politicians, lawyers and firms. The participants of the platform are in constant dialogue on issues such as the importance of AI for politics and society, ethical dimensions of research and technologies, protection of data and security. Also, the initiative of Microsoft „AI for Earth“ (https://www.microsoft.com/en-us/ai/ai-for-earth?activetab=pivot1%3aprimaryr6) serves the purpose of introduction of AI to extensive portions of society and democratization of AI.

23 Asilomar AI Principles (https://futureoflife.org/ai-principles/) emerged from the result of the 2017 Conference convened with the efforts of Future of Life Institute and brings together the principles of AI research. Security, legal transparency, respect for privacy, supportable technology, prohibition of discrimination, sharing of wealth, are the most important ones of these principles.

4.4) AI and Law

Legal discipline, which aims to find solutions to legal disputes which are increasing and getting complicated, by adapting the developments in technical field to its own needs, tries to limit the development of technology within the frame of the classical function of law, whereas it also benefits from the capacity of technology to handle standard works. Additionally, it is observed that, with the help of artificial intelligence, the decisions of judiciary which are biased due to social prejudices, assumptions and fixated values, are tried to be made more objective.

First thing that is aimed at law with machine learning, is to raise consciousness about the assumptions in law, with the help of information collected systematically. The patterns determined on a systematic basis, will help self-reflection of decision-making organs. Secondly, training of decision-makers in law on data

analysis, and ensuring knowledge on collection method and systematization of collected data, are important[4].

Although the main purpose of legal training is not to teach machine learning, programming etc. courses to law students with their technical aspect, it is possible to opt for an interdisciplinary cooperation by carrying out matching at the stage of evaluation and systematization of data with technical knowledge of engineering students. This way, instead of a blind belief of technicality, it is possible to ensure acquiring the knowledge of how technicality may be directed/manipulated. Because, it is possible and already practiced that human prejudices shape and are reflected at the stage of programming.

Clarifying the relation between big data and AI, is very significant for legal training. The purpose for systemization of big and yet non-systematized information and its evolution towards powerful or weak AI, as well as learning the methods for this, is very important for anticipating the problems that may arise in the legal field. Within this context, while it is possible to statistically put forward the decision-making patterns of AI software, it is not possible to determine the principles/rules that form their basis. So, there is no complete transparency concerning basic data of decision-making mechanism.

4.5) Suggestions

1. Organizing movie screenings followed by discussion with inter-disciplinary participation (especially with the cooperation of philosophy, law and cinema departments),
2. Planning a conference series for the academic terms 2019-2020 and 2020-2021, to which speakers from different disciplines should be invited,
3. Linking thesis subjects with AI in all departments,
4. Conference series for and by lawyers on AI and information technologies,
5. Inviting lawyers from big law firms to introduce AI technologies/programs used in offices. Also discussion on needs and problematic areas with the participation of engineers,
6. This way, constant updating of expectations about program, literature etc. competencies that lawyers find beneficial to teach to students,
7. Ensuring that court decisions are made open to common use by removing personal data, in cooperation especially with high judicial organs and Ministry of Justice,
8. Ensuring that law faculty students take basic courses related to AI in computer and engineering sciences,
9. Ensuring that academics that may work together to have a connection with each other through establishment of an “AI Working Platform” (interdisciplinary),
10. Listing the online courses on AI in the world and submitting this to the information of students,
11. Establishing inter-disciplinary reading groups within the university,
12. Organizing a meeting with external stakeholders to conduct a needs assessment.

[4] Chen, Machine Learning and the Rule of Law, s. 7.

5) Broader Impact of Data Science to Bilkent University

Bilkent University has to plan and deliver basic services to the students and faculty on the campus every day. The rich data collected on the operations in the university servers every day can be analyzed with data science methods. The results of the analysis can be turned into actionable policy alternative.

We also believe that reaching out to talented high school students who want to excel in data science and internationally recognized data science researchers will strengthen research and teaching in data science here at Bilkent.

5.1) University Operations

We believe that there are many opportunities to increase the efficiency of operational activities pursued by the university. To materialize these benefits, we believe that, a first step is to invest in a technological infrastructure that will generate/store data on all relevant activities (that are policy-relevant) on campus. Using such (historical as well as real-time) data, cost/waste-minimizing or efficiency-enhancing policies can be designed/implemented. Some natural application areas are the followings.

- Provision of shuttle and personnel services,
- Food-waste management and shift-management in cafeterias,
- Heating/electricity or more generally energy consumption,
- Forestation/irrigation,
- Student housing allocations/management,
- Course-registrations,
- Inventory management in various service departments such as the health-center,
- On-campus traffic management

The university administration can even organize innovation contests among students (who could be mentored/advised by faculty members) to find solutions to these problems.

5.2) Outreach Activities

I. **Attracting "Better" Students:** Using ÖSYM and high-school data of students in tandem with their educational data record at Bilkent and employment data after graduation has the potential to bring about valuable insights about the "desired" student profile (possibly varying across departments). With insights gained from such an exercise, the university can design more

precise/customized policies/ways to attract the desired student profile to the university.

II. **International Collaborations:** Some prestigious institutions such as Stanford University, MIT, Carnegie Mellon, UC Berkeley, University of Chicago are –understandably– ahead of the game. Forming collaborations with faculties in such universities may give us a valuable access to a great variety of courses they offer (through online teaching arrangements and satellite-affiliate programs).

6) Summary of Suggestions

1. Attracting new faculty working on all aspects of data science, offering them better salaries and startup packages
2. Offering more technical elective courses related with all aspects of data science
3. Giving incentives to the faculty to offer such new technical elective courses
4. Hiring better TAs, especially in the related fields, offering them better salaries
5. Extending computing resources at Bilkent
6. Constructing a universitywide infrastructure with powerful GPU servers
7. Motivating the faculty to carry on joint research projects with tech companies, improving the overhead policies in favor of these faculty
8. University-wide interdisciplinary data-analytics course
9. Data Analytics Course for Social Sciences
10. A programming course for non-CS students on big data
11. A new Business Analytics Program
12. Moving to open-source Linux operating system more amenable to computing and data analytics and open-source software in teaching
13. Planning university shuttle schedules, student meal-plans, heating and electricity, irrigation, course registration, traffic regulations by means of data analytics techniques applied on historical and online data
14. Attracting better undergraduate and graduate students
15. Reaching out to international collaborators with summer courses, workshops, teaching agreements

Bibliography

[1] National Research Council,. “Frontiers in Massive Data Analysis.” *The National Academies Press*, 27 June 2013, www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis.

[2] Manyika, James, et al. “Big Data: The next Frontier for Innovation, Competition, and Productivity.” *McKinsey & Company*, May 2011, www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.