

# Women Data Science Leaders in Russia

Alena Suvorova, Valeria Ivaniushina, Alina Bakhitova, Anastasiya Kuznetsova  
National Research University Higher School of Economics  
St.Petersburg, Russia

## ABSTRACT

The project “Women Data Science Leaders in Russia” aims to increase gender diversity and women’s participation in the Russian data science community by means of developing online courses and video materials that present female role models for female students, in order to change the stereotypes that affect the perception of the data science field.

## 1. INTRODUCTION

In Russia mathematics, engineering, and information systems are mainly male-dominated fields, while females tend to select social sciences and the humanities, disciplines considered more suitable for girls. We believe this is a product of stereotypes being translated through school education, and our aim is to take part in changing these stereotypes.

The proposal of our project was based on the well-known benefits of introducing relevant role models. Awareness of successful and qualified female leaders, by female students taking STEM courses, can improve their levels of self-esteem and performance and encourage them in the choice of their field of study.

The initial idea was heavily inspired by our teaching experience. St. Petersburg campus of the Higher School of Economics (HSE University) has a successful record in implementing data science courses for students with social sciences and humanities majors. In 2015, Digital Research in Social Sciences was selected as a strategic focus of our campus, and a minor in data science was introduced. Annually about 180 students with 60% of women from different college majors, from Economics to Oriental Studies, enroll in the two-year system of data science courses that include programming in R, machine learning, basics of network analysis, and computational text analysis. Our current learning analytics show that female students are less confident than the males when it comes to the expectancy of success in the data science course; however, we found no difference in academic achievement between male and female students.

With our project, we want to send a message to female college and high-school students that a data science career is accessible and friendly to women, and we suggest various means for reaching them including promoting data science role models from the international research community and developing an introductory student-organized online course in Data Science.

## 2. PROJECT TEAM

The project team is interdisciplinary and combines data science, sociology of education and educational technology experts. All team members are involved in teaching different aspects of data science to students with non-STEM backgrounds including teaching two-year minor program in data science.

Initially, our team consists of 5 people only, but now the project is still growing and currently more than 10 students volunteers involved in it. Additionally, on our way some of our activities were supported by JetBrains company – the international software development company with offices located in Saint Petersburg.

## 3. PROJECT DETAILS AND OUTCOMES

### 3.1 Data science role models from the international research community

We want to provide visible examples of successful women in data science to encourage female students to enter the field. The first step is to expose these students to experienced female representatives, from both industry and academy.

#### 3.1.1 Biographical interviews

While traditional lectures and workshops usually do not contain personal elements in the narrative, we believe it is essential for female students to understand the possible career trajectories. Personal experiences in the data science field help to realize that there is no exclusive pre-specified “right” way to become a data scientist and different backgrounds can give rise to useful and important ideas. As a part of our project, we recorded a series of short inspirational biographical interviews: 20 English-language interviews and 1 Russian-language. We interviewed the most successful women in Data Science, covering descriptions of their research areas and as well as stories from their life.

Most of the interviews were taken during four major conferences: the Web Conference’18, the WebSci’18, Sunbelt’18 and KDD’18. Due to an extensive selection of conferences we were able to catch the diversity of data science communities, from core KDD community to computational social scientists and quantitative researchers.

Women with different backgrounds and experiences were asked about their career path to Data Science as well as their current work and research projects. They talked about the fashion industry and image recognition, advertising, human genome, user experience, animal movement, exploring inequality, studying obesity using Twitter data and many

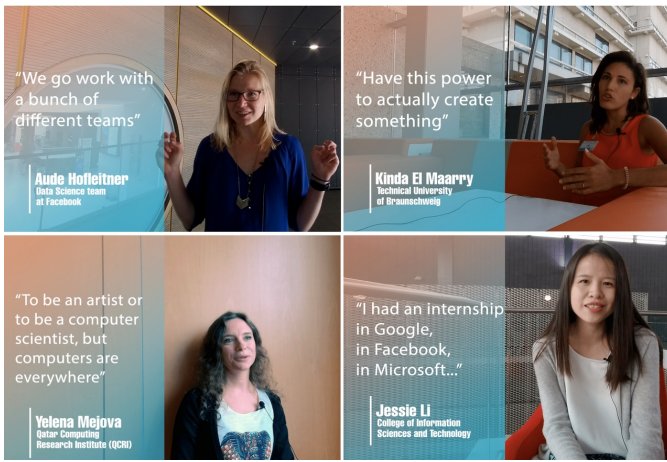


Figure 1: Quotes from the interviews

other fascinating areas of Data Science (Figure 1).

Apart from sharing personal stories, interviews helped draw a new picture of data science and learn important directions for future development including analyzing sensor data for smart home devices, understanding and combating algorithmic discrimination, shifting to prescriptive data science and policy development.

### 3.1.2 Lectures and social events

Besides short inspirational interviews, we organize offline events and invite prominent women, recognized in the international data science community, to give public talks in St. Petersburg. As our team has grown, the project attracted the attention of JetBrains – the international software development company whose biggest offices are located in Saint Petersburg. JetBrains company was particularly interested in offline events and have a successful history of promoting computer science for school kids.

For now, we had one public talk in 2018 and several scheduled in 2019. The first talk by Noa Yehezkel discussed how data science helped to discover new planets and was attended by 200 students from different technical and non-technical universities.

Networking events are essential for building any community and data science is not an exception. This year we host local Women in Data Science 2019 conference (WiDS St. Petersburg). In addition to traditional lectures, we organize a career panel/workshop for female students, oriented on learning about techniques necessary for an effective career launch in data science. Students will have an opportunity to ask questions to the data scientists working in industry and solving practical tasks with real data. An important part of the conference will be the networking and poster session where students of Higher School of Economics and other universities will be presenting their projects. The projects will cover various applications of data science including sociology, logistics, economics, computer science, medicine, oriental studies.

Both interviews and lectures are currently available online on the project website ([wdl-hse.org](http://wdl-hse.org)) and promoted through social media to the intended audience. The Russian educational system is very heterogeneous, with people in remote regions and small cities having fewer opportunities in STEM



Figure 2: Students working on online course

education and in integrating with the data science community, both due to the lack of data science specialists, and because of the language barriers related to English-taught online courses on platforms such as edX and Coursera. That is why we believe the online format of lectures and biographical talks, translated into Russian (supplemented with subtitles both in Russian and English), will have great success with audiences outside the university walls.

### 3.2 Student-organized online course

Between the accomplished specialists in data science and the high school students, there is a vast social distance, and the children may feel that this success is unreachable. To shorten this distance, we are developing short online educational tutorials, that includes interesting examples of research involving data science toolkits, conducted by female students.

At the HSE University, we have examples of research conducted by female students studying social sciences and humanities, including research analysis of the massive communication during esports streaming events, a text-mining-based study of Russian popular song lyrics, and small-scale educational technology research.

We believe that the proposed system, introducing role models on different levels, provides an overview of the possible fields of professional orientation for young girls at school.

Short tutorials, supplemented with exercises and quizzes, are intended to boost the interest in data science and do not require the previous background in Statistics, Mathematics or Programming. Tutorials are developed and delivered by female students who are interested in sharing their knowledge and research experience in data science (Figure 2).

The online course is aimed at high school students and targeted to show potential university applicants how one can study the world around using data science and how this knowledge can be applied to real issues in various areas. The course follows a case-centered model: it gives a small theoretical framework and then offers cases that can be solved with help of basic coding skills. Interactive nature and clear connections with real life make it an engaging form of learning.

The course consists of several units, which are connected to experiences familiar to almost every modern teenager such

	Friendship	Music, Films, and Books	Game studies	Career trajectories	Urban studies
Topic	<ul style="list-style-type: none"> <li>- Social network mechanisms of friendship</li> <li>- Social capital and diversity of social circles</li> <li>- Friendship and Social Media</li> </ul>	<ul style="list-style-type: none"> <li>- Homophily</li> <li>- Conspicuous consumption</li> <li>- Cultural omnivorousness</li> <li>- Dimensions of music</li> </ul>	<ul style="list-style-type: none"> <li>- Players and motivations</li> <li>- Team formation</li> <li>- Toxic behavior</li> </ul>	<ul style="list-style-type: none"> <li>- Virtual communities of practice</li> <li>- Human and social capital</li> </ul>	<ul style="list-style-type: none"> <li>- Citizen movements and social media</li> <li>- Social control in cities</li> <li>- Third places</li> </ul>
Methods	<ul style="list-style-type: none"> <li>- Introduction to Social Network Analysis</li> </ul>	<ul style="list-style-type: none"> <li>- Similarity measures</li> <li>- Multidimensional scaling</li> <li>- Introduction to recommendation systems</li> </ul>	<ul style="list-style-type: none"> <li>- Text Mining</li> <li>- Decision trees</li> </ul>	<ul style="list-style-type: none"> <li>- Association rules</li> <li>- Clustering algorithms</li> </ul>	<ul style="list-style-type: none"> <li>- Introduction to ABM</li> <li>- Visualizing geospatial data</li> </ul>

Figure 3: Outline of the online course

as computer games, friendship, music (Figure 3).

Almost every teenager has a profile on a social networking site. That is why we started the course from exploring social networks concepts and illustrating the strength of weak ties using friendship networks. We discussed how friendship between people is formed, how this connections are changed and why it happens.

Consumption of music, films and books is an integral part of many people’s lives, so this topic will be familiar to almost every student. In the module it is dismantled how cultural preferences are connected to your friends and how we can study dimensions of music based on it. We introduce recommendation systems to illustrate similarity in preferences among friends.

The largest module is devoted to studying human behavior in online games and esports events. We show price formation of virtual items with decision trees and teach text mining on the examples of esports chats.

Next, a module on Career trajectories in IT is introduced aimed to show how we can study skill and their combination using cases of GitHub and StackOverflow.

Lastly, we show how online urban communities are related to the concept of civil society. We shift from individual behavior to patterns taking the case of traffics in the city and learn how to solve it with game theory.

#### 4. EFFECTS, SUSTAINABILITY AND METRICS

We finally decided to focus on predominantly online formats. With Russia being a geographically large and economically disparate country, it is unfeasible without an enormous investment of time and money to organize efficient “offline” forms of data science career promotion. The airplane ticket fare from Omsk (mid-Siberia) to St.Petersburg is around \$400, and \$600 from Vladivostok (Far East), so any format targeting mostly offline participation limits the reach of the project enormously. We think that the online dissemination of talks, materials and courses, under free licenses, supported by online networking activities, will help to increase the impact, not only to geographically remote cities in Russia but potentially far wider than just Russia.

Many women will continue to choose social sciences and the humanities as their majors in college, and our online courses and modules will make data science more attractive to them. In HSE University, these online materials will be prepared with an input of non-STEM students who learned data sci-

ence, R and Python to a certain degree of mastery, and that, we hope, will open the door to the data science community for many women taking non-STEM majors.

Since the project is long-term it is impossible to measure immediate results of the project except for simplest ones such as 200+ people attending the public lecture, 1000+ views on YouTube channel. In general, several basic metrics will be used to track the reach of our project:

- involvement of female students in the online courses (both total number and percentage)
- number of students attending lectures and workshops offline
- representation of different cities and regions in online and offline activities

While for HSE students we expect to have long-term career statistics, via the university alumni office, it is more difficult to track the career influence on students at other universities involved in online and offline events.

By supporting post-event communication, and network sustainability, via social media, however, we expect to build an online survey panel of participants, getting feedback both on the dynamics of data science-specific self-efficacy, and on changes in career choice, including both the next step in the educational system and in industry.

#### 5. LESSONS LEARNED

In this section we summarize the issues/obstacles which we have faced in our project, our ways to solve them, and recommendations to those who might be running similar projects in the future.

The first issue was related to a stigma associated with the “female focused” projects. Quite often such initiatives receive criticism and are perceived as a means of separating women from the society on the whole. Even though our project is aimed at promoting data science regardless of the participants’ gender, we have also received our share of negativity. As we have found out, this attitude can be partly overcome by explaining our position in detail to potential participants and making them adjust their perception.

It should be noted that projects of this kind should be long-term in nature. Firstly, one should not expect fast results of such projects as ours. Secondly, it is difficult to attract participants to a completely new and unknown project: it needs some time to build up. We have found that one of the

ways to make it work is to find and attract a well-known person in the field (in our case, Wendy Hall), who would give weight to the project and make easier to involve other participants. We recommend finding such a person as early as possible, because this makes further steps much easier indeed.

The long-term nature of such projects requires sustainability, which can be achieved by finding partners in other relevant areas, for example, in local data science communities, universities, software companies, and educational initiatives. At first we had difficulty with disseminating the information about our project due to limited connections with other women in Data Science. Thankfully we had social networks at our disposal (including Twitter and social network VKontakte), which we use extensively. We decided against employing advertisement companies, and our promotion campaign was mostly held through VKontakte reposts. This approach has been successful and attracted many interested participants. However, in the future we plan to start advertisement campaigns to increase our audience. Holding local events (such as the Woman in Data Science conference) also helps to increase general awareness and bring in new participants.

There are also language issues. Initially, all our video interviews were in English, which created a language barrier for the Russian-speaking audience. First, we decided to provide subtitles for English-language videos to make them easier for understanding. Second, we plan to shoot more video interviews in Russian, with Russian-speaking women in Data science. Because not all Russian students have sufficient English skills, we duplicated all our materials in two languages - English and Russian; this applies to both website and social media pages.

Sometimes the goal might be achieved by easier ways. For example, initially we planned to make video interviews with invited female speakers during their visits to St. Petersburg. However, quite soon we realized that we could get even better results by interviewing female participants at international conferences, which we were attending anyway. This made our task much easier and at the same time increased the pool of our potential participants.

Finally, we recommend incorporating different aspects of the project into other regular activities (e.g., teaching). For example, at the start of a Data science course, short video interviews can be used for showing different ways that bring women into data science.

This approach also helps to increase the audience coverage and make our work easier in the future.

## 6. CONCLUSION

Our work on this project revealed once again that there is an urgent need for more prominent representation of women in data science. Women work as data scientists in companies, they conduct amazing research but it is challenging to find them as speakers at conferences and meetups, so it makes an impression that there are no women in data science. We hope that our project and similar initiatives will help to make changes.

Here are links to all available materials:

- website (project description, videos, news and event reports): <http://wdl-hse.org/en>
- youtube (videos): [https://www.youtube.com/channel/UCJ\\_QT37\\_2IST3F6wD1rZnGw](https://www.youtube.com/channel/UCJ_QT37_2IST3F6wD1rZnGw),
- twitter: @wdl\_hse, @wdl\_rus
- VK: [https://vk.com/wdl\\_hse](https://vk.com/wdl_hse)

## 7. ACKNOWLEDGMENTS

We would like to express our appreciation to all involved into the project the list of those is not limited to the formal project team. First of all, we would like to acknowledge the KDD Social Impact Program that in many ways gave the start of this project. Also, we are honored that our ideas are supported by JetBrains company and the company becomes an important partner of the project. Special thanks go to Daniel Alexandrov whose support, ideas and advice were crucial for the project from the very beginning, and to Ilya Musabirov who mentors student projects those ideas are later transformed into student-organized online courses. Furthermore, we would also like to acknowledge the role of the volunteers who do amazing work at all parts of the project. Especially, we thank Vsevolod Sushevsky who made a major impact in developing the concept of the interviews, edited videos for all of the interviews in the project, and took two interviews himself. And we really appreciate all our speakers and participants, without them the project would not exist and their great inspirational interviews helped us to continue our work and draw valuable insights.