Data M✦tters

Palestine's "Think Data Science" Program

National initiative to create a pool of Data Science professionals

الجهاز المركزي للاحصاء الفلسطيني
Palestinian Central Bureau of Statistics

الجامعة العربية الأمريكية
ARAB AMERICAN UNIVERSITY

Fact Sheet

- The data science market is set to reach USD 103 billion by 2023
- In 2019, data science market is expected to grow by 20%
- 97.2% of organizations are investing in big data and Artificial Intelligence
- Big data and analytics market worth $49 billion in 2019.
- Google generated USD 116.4 billions from advertising business mostly via the use of data science.
- In 2012, only 0.5% of all data were analyzed.
- Job listings for data science and analytics will reach around 2.7 million by 2020.
- In 2015, there were between 11,400 and 19,400 data scientists worldwide. McKinsey predicted that in 2018 there should be approximately 2.8 million people with analytical talent.
- Since 2012, the need to manage raw data created 14 million jobs worldwide!
- BM predicts that by 2020, the demand for data scientists will increase by 28%.
- Data scientists average salary will be USD115,000 per year.
- Data scientist is the sexiest job in the 21st century

The National Data Science Initiative

The Palestinian Central Bureau of Statistics (PCBS) and the Arab American University of Palestine (AAUP) plan to invest in a Data Science Initiative over the next three years. The initiative seeks to enhance opportunities for the youth to tap into the enormous potential of data science. The initiative focuses on fundamental challenges related to awareness, skills, data, utilization, methods and coordination mechanisms.

The Data Science Initiative aims to:

- Provide core skills necessary to build a data-enabled culture within different sectors & communities;
- Support interdisciplinary data-related projects and research initiatives;
- Foster new methodological approaches to big data;
- Increase data fluency and contribute to transfer of data science skills to new graduates;
- Support the exploration of the impact of data revolution on society;
- Foster a diverse community of data science practitioners;

The data science initiative encompasses four core projects:

1. Data Matters Festival
2. Think Data Science Program
3. Data Science for Executives & Leaders
4. Data Science for Schools: Schools Without Walls
5. Data Science Society of Palestine

The different projects consist of core activities that include:

- Practical events on data science, big data topics, methods and technologies;
- Interdisciplinary seminars and workshops on cutting-edge aspects of data science;
- Connecting students and new graduates with internship opportunities to gain and practice data science skills;
- Connecting professionals from different disciplines in a shared space to foster interactions, ideas and solutions with focus on data science.
- Providing executives from different fields with key tools and skills to understand and learn basic skills of data science applications to develop businesses, and the importance of hiring data scientists to solve business problems.
- Building capacities of new graduates, civil servants, employees of private sectors in the field of data science, helping them gain new skills to improve their daily tasks and giving them a competitive advantage for employment or career development.
- Raising awareness among schools' students on data science and teaching them basic skills in using of data.

The Project: "Think Data Science Program"

Today, successful data professionals understand that they must advance the traditional skills of analyzing large amounts of data to uncover insights for organizational development. Professionals must master the full spectrum of data life cycle to maximize returns at each phase of the process. The growing demand for data science professionals across industries, big and small, is being challenged by a shortage of qualified candidates available to fill open positions. The need for data scientists shows no sign of slowing down in the coming years. LinkedIn listed data scientist as one of the most promising jobs in 2017 and 2018[1], along with multiple data-science-related skills as the most in-demand skills by companies[2].

[1] https://www.linkedin.com/jobs/blog/most-promising-jobs-2018
[2] Google Trends

The “Think Data Science Program” is a comprehensive and free training course targeting participants from various sectors to build their capacities in a range of necessary skills needed to understand data science. The program is uniquely designed for Palestine based on an international program offered by the university of Harvard. After the addition of some of the curricula to be taught by Palestinian Central Bureau of Statistics and the Arab American University of Palestine. Participants will receive a certificate of accreditation from PCBS and AUUP as well as Harvard University upon completion of the program.

Benefits of the Think Data Science Program:

- Working with aspiring data science projects under the guidance of PCBS and AAUP;
- Networking and interacting with a well-trained and much-sought-after talent pool interested in the field of big data and data science;
- Forming meaningful experiences with data science professionals as they apply data science knowledge in a teamwork based setting with real-world data analytics problems;
- Sharing data science knowledge regarding the terminology, challenges and tool sets prevalent in the business domain particular to the trainee;
- Exploring new techniques and technologies in data science to increase value in the business domain environment;
- Project-based training mechanism that bridges the gap between training and industry by blending classroom experiences with industry-supplied problems.

Program Type:

The program is based on Harvardx Data Science Program with supplements that aim to enrich the knowledge and practical expertise of trainees. After adding some curricula to be taught by PCBS and AUUP.

Program Overview

The demand for skilled data science practitioners in industry, academia, and government is rapidly growing. The HarvardX Data Science program prepares you with the necessary knowledge base and useful skills to tackle real-world data analysis challenges. The program covers concepts such as probability, inference, regression, and machine learning. It also helps you develop an essential skill set that includes R programming, data wrangling with dplyr, data visualization with ggplot2, file organization with Unix/Linux, version control with git and GitHub, and reproducible document preparation with RStudio.

In each course, we use motivating case studies, ask specific questions, and learn by answering these through data analysis. Case studies include: Trends in World Health and Economics, US Crime Rates, The Financial Crisis of 2007 - 2008, Election Forecasting, Building a Baseball Team (inspired by Moneyball), and Movie Recommendation Systems.

Throughout the program, we will be using the R software environment. You will learn R, statistical concepts, and data analysis techniques simultaneously. We believe that you can better retain R knowledge when you learn how to solve a specific problem.

1- Introduction to Data Science Crash Course

The course is meant to provide an overview of data science explaining howthe concept evolved and where it is heading. The course will also address data science as an evolving profession; what a data scientist can do and what is the difference between statistician, data scientist and computer scientist. A brief overview of the major components of data science will be also introduced: Statistic, correlation, data mining, visualization, model building and prediction, machine learning, deep learning and Artificial intelligence.

2- Harvardx Training Program (online)

This is the core of the professional certification. It is based on the Harvard University data science online professional degree. Students are supposed to attend all courses online, do the required exams and pass these exams in order to be certified. This part consists of 8 modules and it is based on the R programming language.

Below is a brief description of each module.

Data science: R Basics

1–2 hours/week, for 8 weeks
Build a foundation in R and learn how to wrangle, analyze and visualize data.

The first course in our Professional Certificate Program in Data Science will introduce you to the basics of R programming. You can better retain R when you learn it to solve a specific problem, so you'll use a real-world dataset about crime in the United States. You will learn the R skills needed to answer essential questions about differences in crime across different states.

We'll cover R's functions and data types, then tackle how to operate on vectors and when to use advanced functions like sorting. You'll learn how to apply general programming features like "if-else," and "for loop" commands, and how to wrangle, analyze and visualize data.

Rather than covering every R skill you might need, you'll build a strong

foundation to prepare you for the more in-depth courses later in the series, where we cover concepts like probability, inference, regression and machine learning. We help you develop a skill set that includes R programming, data wrangling with dplyr, data visualization with ggplot2, file organization with UNIX/Linux, version control with git and GitHub, and reproducible document preparation with RStudio.

The demand for skilled data science practitioners is rapidly growing, and this series prepares you to tackle real-world data analysis challenges.

What you'll learn

- Basic R syntax
- Foundational R programming concepts such as data types, vectors arithmetic and indexing
- How to perform operations in R including sorting, data wrangling using dplyr and making plots

Data science: Data visualization

1–2 hours/week, for 8 weeks
Learn basic data visualization principles and how to apply them using ggplot2.

As part of our Professional Certificate Program in Data Science, this course covers the basics of data visualization and exploratory data analysis. We will use three motivating examples and ggplot2; a data visualization package for the statistical programming language R. We will start with simple datasets and then graduate to case studies about world health, economics, and infectious disease trends in the United States.

We'll also be looking at how mistakes, biases, systematic errors, and other unexpected problems often lead to data that should be handled with care. The fact that it can be difficult or impossible to notice a mistake within a dataset makes data visualization particularly important.

The growing availability of informative datasets and software tools has led to increased reliance on data visualizations across many areas. Data visualization provides a powerful way to communicate data-driven findings, motivate analyses and detect flaws. This course will give you the skills you need to leverage data to reveal valuable insights and advance your career.

What you'll learn

- Data visualization principles
- How to communicate data-driven findings
- How to use ggplot2 to create custom plots
- The weaknesses of several widely-used plots and why you should avoid them

Data science: Probability

1–2 hours/week, for 8 weeks

Learn probability theory — essential for a data scientist — using a case study on the financial crisis of 2007–2008.

In this course, which is a part of our Professional Certificate Program in Data Science, you will learn valuable concepts in probability theory. The motivation for this course is the circumstances surrounding the financial crisis of 2007–2008. Part of what caused this financial crisis was that the risk of some securities sold by financial institutions was underestimated. To begin to understand this very complicated event, we need to understand the basics of probability.

We will introduce important concepts such as random variables, independence, Monte Carlo simulations, expected values, standard errors and the Central Limit Theorem. These statistical concepts are fundamental to conduct statistical tests on data and understanding whether the data you are analyzing is likely occurring due to an experimental method or to chance.

Probability theory is the mathematical foundation of statistical inference which is indispensable for analyzing data affected by chance, and thus essential for data scientists.

What you'll learn

- Important concepts in probability theory including random variables and independence
- How to perform a Monte Carlo simulation
- The meaning of expected values and standard errors, and how to compute them in R
- The importance of the Central Limit Theorem

Data science: Inference and modeling

1–2 hours/week, for 8 weeks
Learn inference and modeling; two of the most widely used statistical tools in data analysis.

Statistical inference and modeling are indispensable for analyzing data affected by chance, and thus essential for data scientists. In this course, you will learn these key concepts through a motivating case study on election forecasting.

This course will show you how inference and modeling can be applied to develop the statistical approaches that make polls an effective tool, and we'll show you how to do this using R. You will learn concepts necessary to define estimates and margins of errors and learn how you can use these to make predictions relatively well and also provide an estimate of the precision of your forecast.

Once you learn this, you will be able to understand two concepts that are ubiquitous in data science: confidence intervals and p-values. Then, to understand statements about the probability of a candidate winning, you will learn about Bayesian modeling. Finally, at the end of the course, we will put it all together to recreate a simplified version of an election forecast model and apply it to the 2016 election.

What you'll learn

- The concepts necessary to define estimates and margins of errors of populations, parameters, estimates and standard errors in order to make predictions about data
- How to use models to aggregate data from different sources
- The very basics of Bayesian statistics and predictive modeling

Data science: Wrangling

1–2 hours/week, for 8 weeks
Learn to process and convert raw data into formats needed for analysis.

In this course, which is part of our Professional Certificate Program in Data Science, we cover several standard steps of the data wrangling process like importing data into R, tidying data, string processing, HTML parsing, working with dates and times and text mining. Rarely are all these wrangling steps necessary in a single analysis, but a data scientist will likely face them all at some point.

Very rarely is data easily accessible in a data science project. It's more likely for data to be in a file, a database or extracted from documents such as web pages, tweets or PDFs. In these cases, the first step is to import data into R and tidy the data using the tidyverse package. The steps that convert data from its raw form to the tidy form is called data wrangling.

This process is a critical step for any data scientist. Knowing how to wrangle and clean data will enable you to make critical insights that would otherwise be hidden.

What you'll learn

- Importing data into R from different file formats
- Web scraping
- How to tidy data using the tidyverse to better facilitate analysis
- String processing with regular expressions (regex)
- Wrangling data using dplyr
- How to work with dates and times as file formats
- Text mining

Data science: Linear regression

1–2 hours/week, for 8 weeks
Learn how to use R to implement linear regression; one of the most common statistical modeling approaches in data science.

Linear regression is commonly used to quantify the relationship between two or more variables. It is also used to adjust for confounding. This course, which is part of our Professional Certificate Program in Data Science, covers how to implement linear regression and adjust for confounding in practice using R.

In data science applications, it is very common to be interested in the relationship between two or more variables. The motivating case study we examine in this course relates to the data-driven approach used to construct baseball teams described in Moneyball. We will try to determine which measured outcomes best predict baseball runs by using linear regression.

We will also examine confounding, where extraneous variables affect the relationship between two or more other variables, leading to spurious associations. Linear regression is a powerful technique for removing confounders, but it is not a magical process. It is essential to understand when it is appropriate to use, and this course will teach you when to apply this technique.

What you'll learn

- How linear regression was originally developed by Galton
- What confounding is and how to detect it
- How to examine the relationships between variables by implementing linear regression in R

Data science: Machine learning

2–4 hours/week, for 8 weeks
Build a movie recommendation system and learn the science behind one of the most popular and successful data science techniques.

Perhaps the most popular data science methodologies come from machine learning. What distinguishes machine learning from other computer guided decision processes is that it builds prediction algorithms using data. Some of the most popular products that use machine learning include the handwriting readers implemented by the postal service, speech recognition, movie recommendation systems and spam detectors.

In this course, which is part of our Professional Certificate Program in Data Science, you will learn popular machine learning algorithms, principal component analysis and regularization by building a movie recommendation system.

You will learn about training data, and how to use a set of data to discover potentially predictive relationships. As you build the movie recommendation system, you will learn how to train algorithms using training data so you can predict the outcome for future datasets. You will also learn about overtraining and techniques to avoid it such as cross-validation. All of these skills are fundamental to machine learning.

What you'll learn

- The basics of machine learning
- How to perform cross-validation to avoid overtraining
- Several popular machine learning algorithms
- How to build a recommendation system
- What regularization is and why it is useful

Data science: Capstone

15–20 hours/week, for 2 weeks
Show what you've learned from the Professional Certificate Program in Data Science.

To become an expert data scientist you need practice and experience. By completing this capstone project, you will get an opportunity to apply the knowledge and skills in R data analysis that you have gained throughout the series. This final project will test your skills in data visualization, probability, inference and modeling, data wrangling, data organization, regression and machine learning.

Unlike the rest of our Professional Certificate Program in Data Science, you will receive, in this course, much less guidance from the instructors. When you complete the project you will have a data product to present to potential employers or educational programs; a strong indicator of your expertise in the field of data science.

What you'll learn

- How to apply the knowledge base and skills learned throughout the series to a real-world problem
- How to independently work on a data analysis project

3- Professional training and Capstone project within the Palestinian context

In this part, students will have some professional training on a nationally generated data. They also have to practice what they learned in the Harvardx courses and complete a project on national data.

www.datamatters.ps