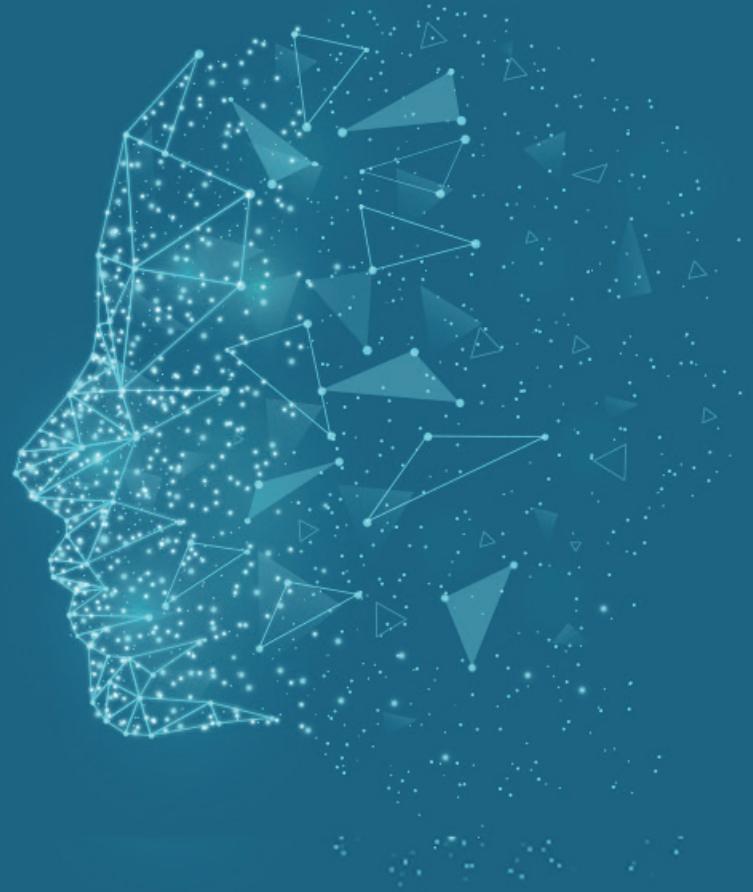


علم البيانات

برنامج ”فکر وتعلم علم البيانات“ الفلسطینی
المبادرة الوطنیة لخلق خبراء فی علم البيانات



الجامعة العربية الامريكية
ARAB AMERICAN UNIVERSITY



الجهاز المركزي للإحصاء الفلسطيني
Palestinian Central Bureau of Statistics

ورقة الحقائق:

- من المقرر أن تصل قيمة سوق عمل علم البيانات إلى 103 مليارات دولار بحلول عام 2023.
- في عام 2019، من المتوقع أن ينمو سوق العمل الخاص بعلم البيانات بنسبة 20%.
- تستثمر 97.2% من المؤسسات في مجال البيانات الضخمة والذكاء الاصطناعي.
- تُقدر قيمة سوق عمل البيانات الضخمة وتحليل البيانات بقيمة 49 مليار دولار في عام 2019.
- حققت جوجل أرباح بقيمة 116.4 مليار دولار من الإعلانات عن خلال استخدام علم البيانات.
- في عام 2012، تم تحليل 0.5% فقط من أصل جميع البيانات.
- من المتوقع أن تصل قوائم الوظائف الخاصة بعلم وتحليل البيانات إلى حوالي 2.7 مليون بحلول عام 2020.
- في عام 2015، كان هناك ما بين 11400 إلى 19400 عالم بيانات في جميع أنحاء العالم.
- توقعت شركة ماكيينزي أنه بحلول عام 2018 سيكون هناك ما يقارب 2.8 مليون شخص لديهم الموهبة التحليلية.
- منذ عام 2012، أدت حاجة إدارة البيانات الخام إلى خلق 14 مليون وظيفة في جميع أنحاء العالم.
- يتوقع مصرف BM بازدياد الطلب على علماء البيانات بنسبة 28% بحلول عام 2020 حيث سيكون متوسط رواتبهم السنوية 115000 دولار.
- مهنة عالم البيانات هي المهنة الأكثر إثارة وتسويق في القرن الحادي والعشرين.

المبادرة الوطنية لعلم البيانات:

يسعى الجهاز المركزي للإحصاء الفلسطيني والجامعة العربية الأمريكية في فلسطين للاستثمار في مبادرة علم البيانات على مدار السنوات الثلاث المقبلة حيث تهدف هذه المبادرة إلى تعزيز الفرص للشباب وذلك للاستفادة من الكم الهائل لعلم البيانات وسوف تُعني المبادرة بالتحديات الأساسية المتعلقة بالوعي والمهارات وأساليب الاستخدام واليات التنسيق للبيانات.

تهدف مبادرة علم البيانات إلى:

- توفير المهارات الأساسية الازمة لبناء ثقافات ممكنة في مختلف القطاعات والمجتمعات.
- دعم المشاريع المتعلقة بالبيانات والمبادرات البحثية.
- تعزيز النهج فيما يتعلق بمنهجيات جديدة للبيانات الضخمة.
- زيادة اتقان وفهم البيانات والمساهمة في نقل مهارات علم البيانات إلى الخريجين الجدد.
- المساهمة في استكشاف أثر ثورة البيانات على المجتمع.
- تعزيز وجود مجتمعي متخصص وممارس علم البيانات.

تشمل مبادرة علم البيانات خمس مشاريع رئيسية:

1. مهرجان علم البيانات.
2. برنامج فكر وتعلم علم البيانات.
3. علم البيانات للمدراء التنفيذيين والقادة.
4. علم البيانات للمدارس: مدارس بلا جدران.
5. جمعية علم البيانات.

ت تكون المشاريع المختلفة من أنشطة أساسية تشمل:

- أنشطة عملية حول علم البيانات، وموضوعات البيانات الضخمة، والأساليب والتكنيات.
- حلقات دراسية وورش عمل متعددة التخصصات حول الجوانب المتطورة لعلم البيانات.
- توفير فرص تدريب للطلاب والخريجين الجدد لاكتساب مهارات علم البيانات وعمارتها.
- ربط المهنيين من مختلف التخصصات في مساحة مشتركة لتعزيز التفاعلات والأفكار والحلول مع التركيز على علم البيانات.
- تزويد المدراء التنفيذيين من مختلف المجالات بأدوات ومهارات لازمة وضرورية لفهم المهارات الأساسية لتطبيقات علم البيانات والتعلم منها لتطوير الأعمال وأهمية توظيف علماء البيانات لحل مشاكل العمل.
- بناء قدرات الخريجين الجدد وطلاب الدراسات العليا وموظفي الخدمة المدنية وموظفي القطاع الخاص في مجال علم البيانات، ومساعدتهم على اكتساب مهارات جديدة لتحسين مهامهم اليومية ومندهم ميزة تنافسية في مجال التوظيف أو التطوير الوظيفي.
- نشر الوعي بين طلاب المدارس في موضوع علم البيانات، وتعليمهم المهارات الأساسية في استخدام البيانات.

المشروع: "برنامج فكر وتعلم علم البيانات"

في يومنا هذا، يدرك خبراء البيانات الناجحون بوجوب النهوض بالمهارات التقليدية لتحليل الكميات الهائلة من البيانات للكشف عن رؤى التنمية التنظيمية. وفي الوقت ذاته، يجب أن يتقن هؤلاء الخبراء النطاق الكامل لدوره حياة البيانات لزيادة العائدات إلى أقصى حد في كل مرحلة من مراحل هذه العملية حيث يواجه الطلب المتزايد على المتخصصين في علم البيانات ضمن الصناعات الكبيرة والصغيرة على حد سواء تحدياً وذلك بسبب نقص المرشحين المؤهلين المتاح لشغل المناصب المفتوحة. ومن الناحية الأخرى، لا يوجد هنالك أي مدلولات إلى أن الحاجة إلى علماء البيانات في تباطؤ خلال السنوات القادمة بل على العكس من ذلك، فقد أدرج موقع LinkedIn¹ أن وظيفة عالم البيانات كأحد أكثر الوظائف الوعادة في الأعوام 2017 و 2018¹، إلى جانب العديد من المهارات المتعلقة بالبيانات العلمية باعتبارها الأكثر طلباً من قبل الشركات².

¹ <https://www.linkedin.com/jobs/blog/most-promising-jobs-2018>

² Google Trends

إن برنامج "فکر وتعلم علم البيانات" عبارة عن دورة تدريبية شاملة ومجانية تستهدف المشاركين من مختلف القطاعات لبناء قدراتهم بمجموعة من المهارات الازفة لفهم علم البيانات، فالبرنامج مصمم خصيصاً لدولة فلسطين وذلك استناداً على برنامج دولي تقدمه جامعة هارفارد، بعد أن تم إضافة بعض المناهج التي سيتم تدريسيها من قبل الجهاز المركزي للإحصاء الفلسطيني والجامعة العربية الأمريكية، حيث سيتم منح المشاركين عند إكمال البرنامج شهادة معتمدة من الجهاز والجامعة بالإضافة لجامعة هارفرد.

فوائد برنامج فکر وتعلم علم البيانات:

- العمل على مشاريع علمية ملهمة فيما يخص علم البيانات تحت إشراف وتوجيهات الجهاز المركزي للإحصاء الفلسطيني والجامعة العربية الأمريكية.
- التشبيك والتفاعل مع مجموعة من المواهب المدربة جيداً والمهتمة في مجال البيانات الضخمة وعلم البيانات.
- إنماء خبرات وتجارب مفيدة وذلك بالعمل مع خبراء وأخصائيي علم البيانات والعمل بروح الفريق لمواجهة الصعوبات في عملية تحليل البيانات.
- المشاركة من حيث المصطلحات والتحديات ومجموعات الأدوات المنتشرة في مجال الأعمال.
- اكتشاف تقنيات تكنولوجية متقدمة ومتقدمة بعلم البيانات من أجل زيادة المعرفة في بيئه مجال الأعمال.
- إيجاد آلية تدريب مبنية على أساس المشاريع التي من شأنها سد الفجوات ما بين التدريب وحاجة سوق العمل من خلال دمج الخبرات المكتسبة من التعليم مع أصحاب العمل.

نوع البرنامج:

تم تصميم هذا البرنامج بالاستناد على برنامج (Harvardx) لعلم البيانات مع ملحقاته الذي يهدف إلى إثراء المعرفة والخبرة العملية للمتدربين، بعد أن تم إضافة بعض المناهج التي سيتم تدريسيها من قبل الجهاز المركزي للإحصاء الفلسطيني والجامعة العربية الأمريكية.

نظرة عامة عن البرنامج:

يتزايد الطلب بشكل سريع على الممارسين والمحترفين لعلم البيانات في القطاعات المختلفة والأوساط الأكademية والحكومية. يقوم برنامج (HarvardX) لعلم البيانات بتجهيز المشارك بأساسيات المعرفة الضرورية والازمة لمواجهة التحديات أثناء عملية تحليل البيانات. كما ويتناول البرنامج مصطلحات ومفاهيم مثل الاحتمالية والاستدلال والانحدار وعلم الآلة والذي بدوره يساعد المشارك على تطوير مجموعة من المهارات الأساسية التي تشمل برمجة (R)، والتعامل مع البيانات وتنقلها بصيغها المختلفة.

لتصور المرئي للبيانات باستخدام (ggplot2)، تنظيم الملفات باستخدام (Unix/Linux)، التحكم بالإصدارات باستخدام (git & GitHub)، وإعداد الوثائق القابلة لإعادة المعالجة باستخدام (Rstudio).

ومنبرز في كل دورة تدريبية دراسات حالات مشجعة ومدفزة، وطرح أسئلة محددة وتعلم الإجابة على هذه الأسئلة من خلال تحليل البيانات. وتشمل دراسات الحالات التالية: الاتجاهات في الصحة والاقتصاد العالمي، معدلات الجريمة في الولايات المتحدة، الأزمة المالية 2007-2008، التنبؤ بالانتخابات، تجهيز فريق لليبيسول مستوحاة من (Moneyball)، والتنبؤ بالطلب على الأفلام السينمائية.

وسيتم خلال هذا البرنامج، استخدام برمجة (R)، حيث سيتعلم المشارك برمجة (R)، المفاهيم الإحصائية، وتقنيات تحليل البيانات في وقت واحد. ونؤمن بأن المشارك في هذا البرنامج سيكون قادرًا على استذكار المعلومات المتعلقة باستخدام برمجة (R) بشكل أفضل وذلك عند تعلم كيفية حل مشكلة معينة.

شهادة برنامج "فك وتعلم علم البيانات":

ت تكون شهادة البرنامج من ثلاثة مكونات: مقدمة لعلم البيانات، برنامج تدريب (Harvardx) عبر الإنترنت، ومشروع تطوير علم البيانات. وفيما يلي وصف لكل مكون:

مقدمة حول البرنامج التدريبي لعلم البيانات (Crash):
يهدف البرنامج إلى تقديم نظرة عامة على علم البيانات، وشرح كيفية تطور المفهوم ومكانه. كما ستطرق الدورة أيضًا إلى علم البيانات كمهنة متقدمة، وما يمكن لعالم البيانات فعله، وما الفرق بين الخبرير الإحصائي وعالم البيانات وعالم الكمبيوتر. كما سيتم تقديم نظرة عامة موجزة عن المكونات الرئيسية لعلم البيانات، الإحصاء، Correlation، Data mining، machine learning، التعلم العميق والذكاء الاصطناعي.

1- 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

2- 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