UDACITY

NANODEGREE PROGRAM SYLLABUS

Data Scientist

Overview

The Data Scientist Nanodegree program is an advanced program designed to prepare you for data scientist jobs. As such, you should have a high comfort level with a variety of topics before starting the program. In order to successfully complete this program, we strongly recommend that the following prerequisites are fulfilled. If you do not have the necessary prerequisites, Udacity has courses and programs that prepare you for this Nanodegree program.

Programming:
- Python Programming: Writing functions, logic, control flow, and building basic applications, as well as common data analysis libraries like NumPy and pandas
- SQL programming: Querying databases using joins, aggregations, and subqueries
- Comfortable with using the Terminal, version control in Git, and using GitHub

Probability and Statistics:
- Descriptive Statistics: Calculating measures of center and spread, estimation distributions
- Inferential Statistics: Sampling distributions, hypothesis testing
- Probability: Probability theory, conditional probability

Mathematics:
- Calculus: Maximizing and minimizing algebraic equations
- Linear Algebra: Matrix manipulation and multiplication

Data wrangling:
- Accessing database, CSV, and JSON data
- Data cleaning and transformations using pandas and Sklearn

Data visualization with matplotlib:
- Exploratory data analysis and visualization
- Explanatory data visualizations and dashboards

Machine Learning:
- Feature Engineering
- Supervised Learning: Regression, classification, decision trees, random forest
- Unsupervised Learning: PCA, Clustering

The following programs can prepare you to take this nanodegree program. There are also several free courses that you can use to prepare.
- Programming for Data Science with Python.
- Data Analyst Nanodegree Program.
- Intro to Machine Learning Nanodegree Program

Educational Objectives: The ultimate goal of the Data Scientist Nanodegree program is for you to learn the skills you need to perform well as a data scientist. As a graduate of this program, you will be able to:
- Use Python and SQL to access and analyze data from several different data sources.

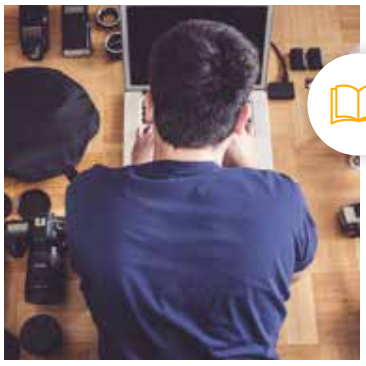
Overview

- Use principles of statistics and probability to design and execute A/B tests and recommendation engines to assist businesses in making data-automated decisions..
- Deploy a data science solution to a basic flask app.
- Manipulate and analyze distributed datasets using Apache Spark.
- Communicate results effectively to stakeholders.

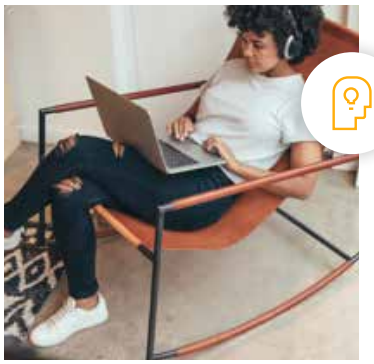
IN COLLABORATION WITH

BERTELSMANN · figure eight · IBM Watson · INSIGHT · kaggle · STARBUCKS

Estimated Time: 4 Months at 10hrs/week



Prerequisites: Python, SQL & Statistics



Flexible Learning: Self-paced, so you can learn on the schedule that works best for you

Need Help?
udacity.com/advisor
Discuss this program with an enrollment advisor.

Course 1: Solving Data Science Problems

Learn the data science process, including how to build effective data visualizations, and how to communicate with various stakeholders.

Course Project
Write a Data Science Blog Post

In this project, you will choose a dataset, identify three questions, and analyze the data to find answers to these questions. You will create a GitHub repository with your project, and write a blog post to communicate your findings to the appropriate audience. This project will help you reinforce and extend your knowledge of machine learning, data visualization, and communication

		LEARNING OUTCOMES
LESSON ONE	**The Data Science Process**	• Apply the CRISP-DM process to business applications • Wrangle, explore, and analyze a dataset • Apply machine learning for prediction • Apply statistics for descriptive and inferential understanding • Draw conclusions that motivate others to act on your results
LESSON TWO	**Communicating with Stakeholders**	• Implement best practices in sharing your code and written summaries • Learn what makes a great data science blog • Learn how to create your ideas with the data science community

Course 2: Software Engineering for Data Scientists

Develop software engineering skills that are essential for data scientists, such as creating unit tests and building classes.

	LEARNING OUTCOMES	
LESSON ONE	**Software Engineering Practices**	• Write clean, modular, and well-documented code • Refactor code for efficiency • Create unit tests to test programs • Write useful programs in multiple scripts • Track actions and results of processes with logging • Conduct and receive code reviews
LESSON TWO	**Object Oriented Programming**	• Understand when to use object oriented programming • Build and use classes • Understand magic methods • Write programs that include multiple classes, and follow good code structure • Learn how large, modular Python packages, such as pandas and scikit-learn, use object oriented programming • Portfolio Exercise: Build your own Python package
LESSON THREE	**Web Development**	• Learn about the components of a web app • Build a web application that uses Flask, Plotly, and the Bootstrap framework • Portfolio Exercise: Build a data dashboard using a dataset of your choice and deploy it to a web application

Course 3: Data Engineering for Data Scientists

Learn to work with data through the entire data science process, from running pipelines, transforming data, building models, and deploying solutions to the cloud.

Course Project
Build Disaster Response Pipelines with Figure Eight

Figure Eight (formerly Crowdflower) crowdsourced the tagging and translation of messages to apply artificial intelligence to disaster response relief. In this project, you'll build a data pipeline to prepare the message data from major natural disasters around the world. You'll build a machine learning pipeline to categorize emergency text messages based on the need communicated by the sender.

	LEARNING OUTCOMES	
LESSON ONE	**ETL Pipelines**	• Understand what ETL pipelines are • Access and combine data from CSV, JSON, logs, APIs, and databases • Standardize encodings and columns • Normalize data and create dummy variables • Handle outliers, missing values, and duplicated data • Engineer new features by running calculations • Build a SQLite database to store cleaned data
LESSON TWO	**Natural Language Processing**	• Prepare text data for analysis with tokenization, lemmatization, and removing stop words • Use scikit-learn to transform and vectorize text data • Build features with bag of words and tf-idf • Extract features with tools such as named entity recognition and part of speech tagging • Build an NLP model to perform sentiment analysis
LESSON THREE	**Machine Learning Pipelines**	• Understand the advantages of using machine learning pipelines to streamline the data preparation and modeling process • Chain data transformations and an estimator with scikit-learn's Pipeline • Use feature unions to perform steps in parallel and create more complex workflows • Grid search over pipeline to optimize parameters for entire workflow • Complete a case study to build a full machine learning pipeline that prepares data and creates a model for a dataset

Course 4: Experiment Design and Recommendations

Learn to design experiments and analyze A/B test results. Explore approaches for building recommendation systems.

Course Project
Design a Recommendation Engine with IBM

IBM has an online data science community where members can post tutorials, notebooks, articles, and datasets. In this project, you will build a recommendation engine, based on user behavior and social network in IBM Watson Studio's data platform, to surface content most likely to be relevant to a user.

	LEARNING OUTCOMES	
LESSON ONE	**Experiment Design**	• Understand how to set up an experiment, and the ideas associated with experiments vs. observational studies • Defining control and test conditions • Choosing control and testing groups
LESSON TWO	**Statistical Concerns of Experimentation**	• Applications of statistics in the real world • Establishing key metrics • SMART experiments: Specific, Measurable, Actionable, Realistic, Timely
LESSON THREE	**A/B Testing**	• How it works and its limitations • Sources of Bias: Novelty and Recency Effects • Multiple Comparison Techniques (FDR, Bonferroni, Tukey) • Portfolio Exercise: Using a technical screener from Starbucks to analyze the results of an experiment and write up your findings

LESSON FOUR	**Introduction to Recommendation Engines**	• Distinguish between common techniques for creating recommendation engines including knowledge based, content based, and collaborative filtering based methods. • Implement each of these techniques in python. • List business goals associated with recommendation engines, and be able to recognize which of these goals are most easily met with existing recommendation techniques.
LESSON FIVE	**Matrix Factorization for Recommendations**	• Understand the pitfalls of traditional methods and pitfalls of measuring the influence of recommendation engines under traditional regression and classification techniques. • Create recommendation engines using matrix factorization and FunkSVD • Interpret the results of matrix factorization to better understand latent features of customer data • Determine common pitfalls of recommendation engines like the cold start problem and difficulties associated with usual tactics for assessing the effectiveness of recommendation engines using usual techniques, and potential solutions.

Course 5: Data Science Projects

Leverage what you've learned throughout the program to build your own open-ended Data Science project. This project will serve as a demonstration of your valuable abilities as a Data Scientist.

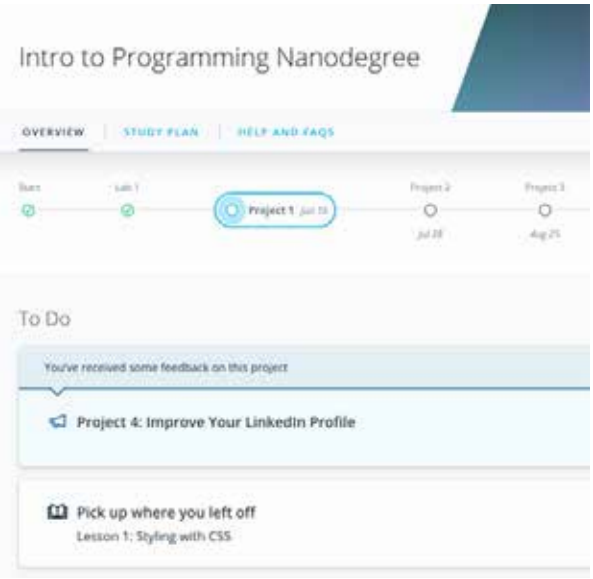
Course Project
Data Science Capstone Project

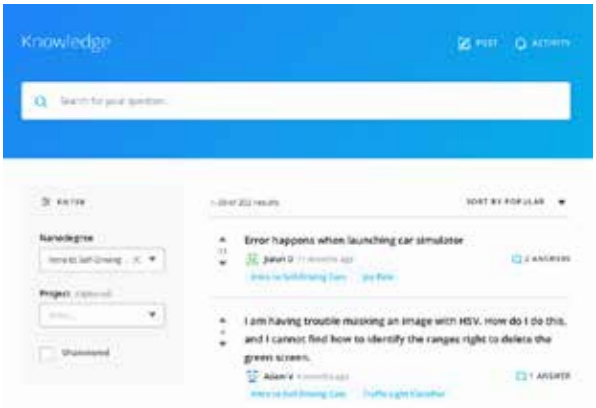
In this capstone project, you will leverage what you've learned throughout the program to build a data science project of your choosing. You will define the problem you want to solve, identify and explore the data, then perform your analyses and develop a set of conclusions. You will present the analysis and your conclusions in a blog post and GitHub repository. This project will serve as a demonstration of your ability as a data scientist, and will be an important component of your job-ready portfolio.

		LEARNING OUTCOMES
LESSON ONE	**Elective 1: Dog Breed Classification**	• Use convolutional neural networks to classify different dogs according to their breeds • Deploy your model to allow others to upload images of their dogs and send them back the corresponding breeds. • Complete one of the most popular projects in Udacity history, and show the world how you can use your deep learning skills to entertain an audience!
LESSON TWO	**Elective 2: Starbucks**	• Use purchasing habits to arrive at discount measures to obtain and retain customers • Identify groups of individuals that are most likely to be responsive to rebates.
LESSON THREE	**Elective 3: Arvato Financial Services**	• Work through a real-world dataset and challenge provided by Arvato Financial Services, a Bertelsmann company • Top performers have a chance at an interview with Arvato or another Bertelsmann company!

LESSON FOUR	**Elective 4: Spark for Big Data**	• Take a course on Apache Spark and complete a project using a massive, distributed dataset to predict customer churn • Learn to deploy your Spark cluster on either AWS or IBM Cloud
LESSON FIVE	**Elective 5: Your Choice**	• Use your skills to tackle any other project of your choice

Our Classroom Experience





REAL-WORLD PROJECTS
Build your skills through industry-relevant projects. Get personalized feedback from our network of 900+ project reviewers. Our simple interface makes it easy to submit your projects as often as you need and receive unlimited feedback on your work.

KNOWLEDGE
Find answers to your questions with Knowledge, our proprietary wiki. Search questions asked by other students, connect with technical mentors, and discover in real-time how to solve the challenges that you encounter.

STUDENT HUB
Leverage the power of community through a simple, yet powerful chat interface built within the classroom. Use Student Hub to connect with fellow students in your program as you support and learn from each other.

WORKSPACES
See your code in action. Check the output and quality of your code by running them on workspaces that are a part of our classroom.

QUIZZES
Check your understanding of concepts learned in the program by answering simple and auto-graded quizzes. Easily go back to the lessons to brush up on concepts anytime you get an answer wrong.

CUSTOM STUDY PLANS
Preschedule your study times and save them to your personal calendar to create a custom study plan. Program regular reminders to keep track of your progress toward your goals and completion of your program.

PROGRESS TRACKER
Stay on track to complete your Nanodegree program with useful milestone reminders.

Learn with the Best

Josh Bernhard

DATA SCIENTIST

Josh has been sharing his passion for data for nearly a decade at all levels of university, and as Lead Data Science Instructor at Galvanize. He's used data science for work ranging from cancer research to process automation.

Juno Lee

DATA SCIENCE INSTRUCTOR

As a data scientist, Juno built a recommendation engine to personalize online shopping experiences, computer vision and natural language processing models to analyze product data, and tools to generate insight into user behavior.

Luis Serrano

INSTRUCTOR

Luis was formerly a Machine Learning Engineer at Google. He holds a PhD in mathematics from the University of Michigan, and a Postdoctoral Fellowship at the University of Quebec at Montreal.

Andrew Paster

INSTRUCTOR

Andrew has an engineering degree from Yale, and has used his data science skills to build a jewelry business from the ground up. He has additionally created courses for Udacity's Self-Driving Car Engineer Nanodegree program.

Learn with the Best

Mike Yi

DATA SCIENTIST

Mike is a Content Developer with a multidisciplinary academic background, including math, statistics, physics, and psychology. Previously, he worked on Udacity's Data Analyst Nanodegree program as a support lead.

David Drummond

VP OF ENGINEERING

David is VP of Engineering at Insight where he enjoys breaking down difficult concepts and helping others learn data engineering. David has a PhD in Physics from UC Riverside.

Judit Lantos

SENIOR DATA ENGINEER

Currently, Judit is a Senior Data Engineer at Netflix. Formerly a Data Engineer at Split, where she worked on the statistical engine of their full-stack experimentation platform, she has also been an instructor at Insight Data Science, helping software engineers and academic coders transition to DE roles.

All Our Nanodegree Programs Include:

EXPERIENCED PROJECT REVIEWERS

REVIEWER SERVICES

- Personalized feedback & line by line code reviews
- 1600+ Reviewers with a 4.85/5 average rating
- 3 hour average project review turnaround time
- Unlimited submissions and feedback loops
- Practical tips and industry best practices
- Additional suggested resources to improve

TECHNICAL MENTOR SUPPORT

MENTORSHIP SERVICES

- Questions answered quickly by our team of technical mentors
- 1000+ Mentors with a 4.7/5 average rating
- Support for all your technical questions

PERSONAL CAREER SERVICES

CAREER SUPPORT

- Resume support
- Github portfolio review
- LinkedIn profile optimization

Frequently Asked Questions

PROGRAM OVERVIEW

WHY SHOULD I ENROLL?

The data science field is expected to continue growing rapidly over the next several years, and there's huge demand for data scientists across industries. Data scientist is consistently rated as a top career.

Udacity has collaborated with industry leaders to offer a world-class learning experience so you can advance your data science career. You'll get hands-on experience running data pipelines, designing experiments, building recommendation systems, and more. You'll have personalized support as you master in-demand skills that qualify you for high-value jobs in the data science field.

By the end of the program, you'll have an impressive portfolio of real-world projects, and valuable hands-on experience. You'll also receive career support via profile and portfolios reviews to help make sure you're ready to establish a successful data science career, and land a job you love.

WHAT JOBS WILL THIS PROGRAM PREPARE ME FOR?

Obtaining the skills required to be a Data Scientist will make you extremely valuable across many industries, and in many roles. Data Scientists work as Analysts, Statisticians, Engineers, and more. Some become Data and Analytics Managers, while others specialize as Database Administrators. As a graduate of this program, you'll be prepared to seek out roles that run the gamut from generalist to specialist, and all points in between.

HOW DO I KNOW IF THIS PROGRAM IS RIGHT FOR ME?

This program offers an ideal path for experienced programmers and data analysts to advance their data science careers. If you're interested in deepening your expertise in the fields of analytics, machine learning, data engineering, and/or data science, this is a great way to get hands on practice with a variety of techniques and learn to build end to end data science solutions.

WHAT IS THE DIFFERENCE BETWEEN THE DATA ANALYST, MACHINE LEARNING ENGINEER, AND THE DATA SCIENTIST NANODEGREE PROGRAMS?

The Data Analyst program is designed for people with some data analysis experience and little-to-no programming experience. Students will learn to analyze data using Python and SQL, to wrangle and clean messy data, to use applied statistics to test hypotheses, and to create data visualizations. Graduates of this program will be prepared for data analyst positions.

The Data Scientist Nanodegree program is designed for students with strong programming and data analysis skills, as it is the next step for graduates of

FAQs Continued

the Data Analyst Nanodegree program. Students will learn to build machine learning models, run data pipelines, design experiments and recommendation engines, communicate effectively, and to deploy data applications. Graduates of this program will be prepared for data scientist positions.

The Machine Learning Engineer Nanodegree program prepares students for machine learning engineering careers. As both data scientist and machine learning jobs require machine learning knowledge, each of these two programs begins with a focus on machine learning. The curriculum diverges in later sections as you begin to focus on more job-specific tools, skills, and techniques.

ENROLLMENT AND ADMISSION

DO I NEED TO APPLY? WHAT ARE THE ADMISSION CRITERIA?

No. This Nanodegree program accepts all applicants regardless of experience and specific background.

WHAT ARE THE PREREQUISITES FOR ENROLLMENT?

The Data Scientist Nanodegree program is designed for students with programming and data analysis experience. Students should have a high comfort level with a variety of topics before starting the program. In order to successfully complete this program, you should meet the following prerequisites:

- Python programming, including common data analysis libraries (NumPy, pandas, Matplotlib).
- SQL programming
- Statistics (Descriptive and Inferential)
- Calculus
- Linear Algebra
- Experience wrangling and visualizing data

IF I DO NOT MEET THE REQUIREMENTS TO ENROLL, WHAT SHOULD I DO?

Udacity's Data Analyst Nanodegree program is great preparation for the Data Scientist Nanodegree program. You'll learn programming with Python and SQL, applied statistics, data wrangling, and data visualization.

You can also prepare by taking a number of Udacity's free courses, such as:

- Introduction to Data Science
- Introduction to Python
- SQL for Data Analysis
- Statistics
- Linear Algebra

FAQs Continued

- Data Visualization with Tableau

TUITION AND TERM OF PROGRAM

HOW IS THIS NANODEGREE PROGRAM STRUCTURED?

The Data Scientist Nanodegree program is comprised of content and curriculum to support four (4) projects. We estimate that students can complete the program in four (4) months working 10 hours per week.

Each project will be reviewed by the Udacity reviewer network. Feedback will be provided and if you do not pass the project, you will be asked to resubmit the project until it passes.

HOW LONG IS THIS NANODEGREE PROGRAM?

Access to this Nanodegree program runs for the length of time specified in the payment card above. If you do not graduate within that time period, you will continue learning with month to month payments. See the **Terms of Use** and **FAQs** for other policies regarding the terms of access to our Nanodegree programs.

SOFTWARE AND HARDWARE

WHAT SOFTWARE AND VERSIONS WILL I NEED IN THIS PROGRAM?

To successfully complete this Nanodegree program, you'll need to be able to download and run Python 3.7.