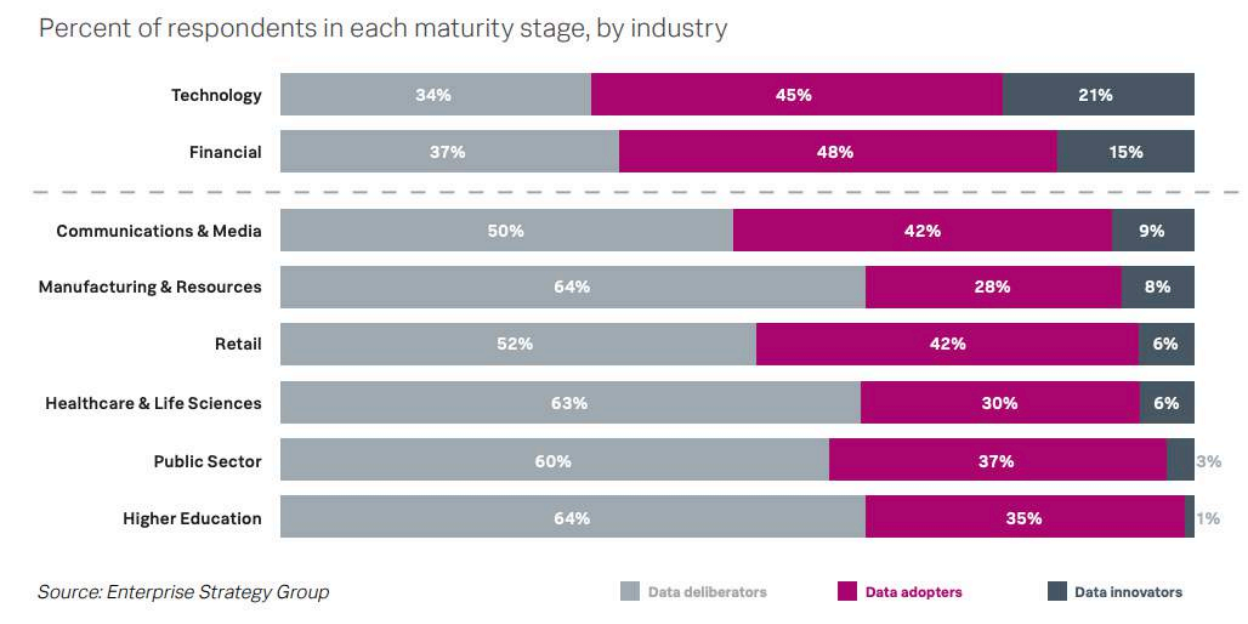
Data Society Training Catalog

Table of Contents

Preparing for a Data Economy

In the last 15 years, 52% of the Fortune 500 companies have disappeared.

In 2015, the average life expectancy of a company was 15 years, compared to 75 years in 1955. The world is changing more rapidly than ever - organizations need to live and breathe data to stay competitive and effective.



Businesses that integrate data into their operations and lead with analytics grow their annual revenue by 5.3 percent and decrease operation costs by 4.85 percent[1]. The organizations that build their data infrastructures and cultivate data-literate workforces will thrive in the new data economy.

Hiring 1 Data Scientist

Training 30 Staff Members

By implementing a data analytics training program, you'll create a workforce that is savvy about data and has demonstrated its dedication to the organization's mission. Data literacy improves efficiency across all levels of an organization, from planning project resources to automating processes.

1. Splunk / ESG “What is your data really worth?”

How to Review This Catalog

There are a lot of courses here – how do I know where to start?

If you feel overwhelmed going through this catalog, you're not alone. Whether you're just exploring our offerings or looking for a specific topic, the following guidelines may be helpful to you in determining which courses are most relevant to your organization's needs:

1. For staff who don't directly manage data but need to understand common data terms and the importance of data, look to our **Data Literacy for All** workshop.
2. For staff who need to better understand how to use data in Excel to gain insights and manage the data quality, look to our **Introduction to Excel and Data Analysis** pathway.
3. For analysts who want to automate data processes and build foundational machine learning processes, look to our **Introduction to R** pathway.
4. For programmers who have a foundation in machine learning and want to build out their predictive capabilities, look to our **Intermediate Machine Learning** pathway.
5. For data scientists who want to embed advanced modeling, look to our **Advanced Machine Learning** pathway.

Data Society can customize all programs and course offerings to fit your use cases and schedule. If you're not sure what makes sense for you or if you want to discuss our customization options, reach out to us at hello@datasociety.com or call us at 202-600-9635.

Included Services

Virtual live-streamed instruction

Course reference guides and cheat sheets

Practical exercises to sharpen skills

Reusable data sets / coding templates

Companion step-by-step workbooks

Capstone project support and guidance

Topics and Tools We Cover

This catalog represents a subset of programs that we provide, but it doesn't capture the full breadth of our content. Curious about what else we teach? Here are some more topics and tools.

Tools	Techniques
• Amazon Web Services (AWS) • Microsoft Azure • Microsoft Excel • Microsoft Power BI • Tableau • SQL / Relational databases • Python • R • Scala • Spark • TensorFlow • Keras	• Data science for executives • Data visualization • Data storytelling • Foundational statistics • Programming fundamentals • Data cleaning • Data transformation • Unsupervised machine learning • Supervised machine learning • Text mining • Network analysis • Neural networks • Deep learning • Cyber security

If you'd like to learn more about the other training programs that we teach or if you want to discuss our customization options, reach out to us at hello@datasociety.com or call us at 202-600-9635.

Capstone Project Plan

The capstone project is a core component of demonstrating the competency and application of employees' skills and can be integrated into your learning pathway. We work closely with students to help them develop a project plan that is scoped appropriately and facilitates success.

Our program will track student progress and set aside time to meet with students individually on scheduled intervals. One of the most critical aspects of the capstone facilitation is ensuring that students frame the right questions and can access the correct datasets. We recommend that students define their key question and hypothesis early on to maximize the time they can spend building and refining the analysis.

One of the main objectives of this program is to teach students the right frameworks and skill sets to understand how to apply their knowledge in different scenarios and problem-solve effectively.

While Data Society instructors will provide students with regular support, we recommend that all participants set aside at least 4-6 hours per week during the program for independent work on the capstone project.

Data Science Literacy Workshops

Building a Data-Literate Workforce

These three workshops are designed for individuals who will be overseeing or working adjacently to data analysts and data scientists. There is no coding involved, but the programs provide a high-level overview of the data's benefits and industry applications and demystify the process behind data analytics. Once attendees complete the workshops, they will feel more confident directing data projects, allocating resources, and working alongside technical staff.

Workshops Include:

Data Science for Executives

Data Science for Managers

Data Literacy for All

Data Sources & Examples Used

- Industry or domain-specific examples
- Customized data that maximizes the attendees' engagement and retention of skills and can be applied directly within your organization

Course Duration

- 6 Hours / 12 Hours
- Available in flexible schedule format

Modality

- Proposed live-streaming (in-person available post-COVID-19)
- Instructor-led

Recommendations

- Up to 20 learners per executive/managers cohort, up to 35 learners per data literacy cohort

Data Science for Executives

1 day / 6 hours of instruction

This workshop teaches executives how to identify opportunities to make better decisions by utilizing new insights from data; which techniques and approaches they can use for various data science projects; how to avoid pitfalls when drawing conclusions from data; and how to create a strategy for improving decision-making across the organization.

Objectives	Prerequisites
1. Identify opportunities to apply data analysis 2. Discuss the elements of a successful data project 3. Recognize misleading data 4. Develop a strategy for building a data-driven culture	No background in math or data analysis is required. We recommend that attendees have experience managing teams.

Syllabus & Topics Covered

1. The importance of data
   a. What is data? What are its sources?
   b. What is a data-driven culture and why is it important?
2. The uses of data
   a. What are the benefits of using data?
   b. What is data analytics?
   c. What do all these buzzwords mean?
3. Components of a data project
   a. What are the principles of data science?
   b. What are the six stages of the typical data science process?
   c. What are common methods data scientists use?
   d. What types of tools do data scientists use to do their work?
   e. How are data teams structured?
4. The pitfalls of data
   a. How can data be deceiving?
   b. How can I avoid being misled by data?
5. Building a data-driven culture
   a. How can I build awareness about the importance and uses of data?
   b. How can I make data-driven decision-making routine?

Software Requirements

None

Data Science for Managers

2 days / 12 hours of instruction

This workshop teaches managers how to identify opportunities to make better decisions by utilizing new insights from data; which techniques and approaches should be used for various data science projects; how to avoid pitfalls when drawing conclusions from data; and how to create a strategy for building a data analytics team and improving decision-making across the organization.

Objectives	Prerequisites
1. Identify opportunities to apply data analysis 2. Discuss the elements of a successful data project 3. Recognize the ethical implications and governance-related challenges of data projects 4. Associate data science methods with their typical applications in data 5. Integrate storytelling and appropriate data visualizations into reporting 6. Develop a strategy for building a data-driven culture	No background in math or data analysis is required. We recommend that attendees have experience managing teams.

Syllabus & Topics Covered

1. The importance of data
   a. What is data? What are its sources?
   b. What is a data-driven culture and why is it important?
2. The uses of data
   a. What are the benefits of using data?
   b. What is data analytics?
   c. What do all these buzzwords mean?
3. Components of a data project
   a. What are the principles of data science?
   b. What are the six stages of the typical data science process?
   c. What are common methods used by data scientists?
   d. What types of tools do data scientists use to do their work?
   e. How are data teams structured?
4. Data governance
   a. What is data governance? Why is it important?
   b. What principles, models, frameworks, and best practices can be used to ensure good data governance?

Software Requirements

None

Data Science for Managers (cont'd)

2 days / 12 hours of instruction

Syllabus & Topics Covered

5. Data ethics
   a. What is data ethics? Why is it important?
   b. What principles, models, frameworks, and best practices apply to the ethical use of data?
6. Foundational data science methods
   a. What is clustering and how is it used?
   b. What is classification and how is it used?
   c. What is regression and how is it used?
7. Advanced data science methods
   a. What is text mining and how is it used?
   b. What is graph analysis and how is it used?
   c. What are neural networks and how are they used?
8. Visualizing results and data storytelling
   a. Why visualize?
   b. What tools can I use for visualization?
   c. How do I select the graphs and charts to use based on my results?
   d. What design principles apply to data visualization?
   e. How do I tell a story with data to inspire action?
9. Building a data-driven culture
   a. How can I build awareness about the importance and uses of data?
   b. How can I make data-driven decision-making routine?

Software Requirements

None

Data Literacy for All

1 day / 6 hours of instruction

This workshop teaches general staff what data science is and builds a shared data vocabulary in order to facilitate communication and collaboration on data science projects. By the end of the workshop, students will be able to identify successful implementations of data analytics, describe data governance best practices, and evaluate data visualizations systematically.

Objectives	Prerequisites
1. Identify what data science is and why it's important 2. Identify proper data collection practices and frameworks 3. Select the appropriate data visualization for the scenario	No background in math or data analysis is required.

Syllabus & Topics Covered

1. The importance of data
   a. What is data? What are its sources?
   b. What is a data-driven culture and why is it important?
2. The uses of data
   a. What are the benefits of using data?
   b. What is data analytics?
   c. What do all these buzzwords mean?
3. The pitfalls of data
   a. How data can be deceiving?
   b. How can I avoid being misled?
4. Data governance
   a. What is data governance? Why is it important?
   b. What principles, models, frameworks, and best practices can I use to ensure good data governance?
   c. What are best practices for data collection and storage?
5. Visualizing and evaluating data
   a. Why is visualization important?
   b. How do I choose the right visualization based on the data?
   c. How do I determine the right visual elements to use to tell my story?
   d. How do I select appropriate combinations of colors to convey results effectively?
   e. How do I assess the validity / accuracy of a visualization?

Software Requirements

None

Learning Pathway:
Introduction to Excel and Data Analysis

Introduction to Excel and Data Analysis

This foundational pathway provides the key building blocks to becoming a successful data analyst. Students will become fluent handling data and understand the nuances of data cleaning. They will learn how to leverage Excel and SQL, one of the most widely used database management languages, to import data, build visualizations, and manipulate and combine multiple data sources.

Courses Include:

Introduction to Excel

Data Analysis in Excel

Introduction to SQL

Data Sources & Examples Used

- Industry or domain-specific examples
- Customized data that maximizes the attendees' engagement and retention of skills and can be applied directly within your organization

Course Duration

- 36 Hours
- Available in flexible schedule format

Modality

- Proposed live-streaming (in-person available post-COVID-19)
- Instructor-led (TA support optional)

Recommendations

- Up to 30 learners per cohort
- Learners should have experience with Excel
- Excel and SQL should be installed on computers

Introduction to Excel

2 days / 16 hours of instruction (Excel)

Microsoft Excel is among the tools professionals and analysts use most widely today. This course will cover how to leverage Excel's capabilities to manipulate and explore data. By the end of this program, students will be able to outline use cases for Excel, import and manipulate data, and build interactive data visualizations.

Objectives	Prerequisites
1. Describe use cases for Excel 2. Import and manipulate data in Excel 3. Build data visualizations 4. Implement best practices for data cleaning	No background in math or data analysis is required, but we recommend that students have some familiarity with Excel.

Syllabus & Topics Covered

1. Introduction to Excel
   a. How is Excel used today?
   b. Use cases for Excel
2. Foundations of Excel functions
   a. Introduction to Excel functions and case studies
   b. Using functions for basic statistics and aggregating data
   c. Building formulas in Excel
3. Advanced functions in Excel
   a. Introduction to case studies and overview
   b. Implementing MATCH(), INDEX()
   c. Finding data with VLOOKUP / HLOOKUP
   d. Auditing formulas, error checking, tracing formula dependents and precedents in Excel
4. Pivot tables
   a. Introduction to PivotTable and PivotChart
   b. Building pivot tables in Excel
   c. Building advanced pivot tables
5. Data visualization
   a. Underlying principles of data visualization
   b. Building multiple charts to communicate data (scatterplots, area charts, etc.)
   c. Adjusting chart design and format
   d. Conditional formatting and sparklines

Software Requirements

Microsoft Excel 2016

Data Analysis in Excel

1 day / 8 hours of instruction (Excel)

Learning Pathway: Excel

This course builds on foundational Excel skills to teach students how to effectively manipulate and analyze data. Students will learn how to build a variety of models and test scenario analyses to make better data-driven decisions. By the end of this program, students will be able to produce multiple scenarios in Excel, optimize data models, and build predictive linear models to test cause and effect relationships.

Objectives	Prerequisites
1. Build data models in Excel 2. Build multiple scenarios in Excel 3. Evaluate and optimize data models	Students must be comfortable using advanced Excel functions, pivot tables, and foundational data visualizations.

Syllabus & Topics Covered

1. What-if analysis and tools
   a. Goal Seek
   b. Data Table
   c. Scenario Manager
2. Goal Seek function overview
   a. Using Goal Seek
   b. Pros and cons of Goal Seek
3. Building 1- and 2-variable data tables and evaluate the outcomes
   a. Sensitivity analysis
   b. Data tables
4. Building multiple scenarios in Excel and evaluate the results
   a. Naming cells
   b. Previewing scenarios
   c. Creating scenarios: Scenario Manager
   d. Summary reports
5. Optimizing data models using Excel Solver
   a. Excel Solver overview
   b. Excel Solver Add-In overview
   c. Optimization modeling
6. Linear regression in Excel
   a. Definition of linear regression
   b. Implementing and running a linear regression model
   c. Evaluating the linear model and interpreting outcomes
   d. Validating the linear regression model

Software Requirements

Microsoft Excel 2016

Introduction to SQL

2 days / 12 hours of instruction (SQL)

This course builds on foundational data skills to teach students how to effectively manipulate data in SQL. SQL is one of the most popular database management languages available and is critical in storing and retrieving data. By the end of this program, students will be able to describe database structures, import and combine multiple data types, and manipulate data to discover new insights.

Objectives	Prerequisites
1. Program proficiently in SQL 2. Manipulate data to prepare for analysis 3. Join multiple tables and data types	This program is best suited to students who are comfortable working with data in Excel.

Syllabus & Topics Covered

1. Introduction to SQL
   a. Identifying use cases for SQL
   b. Defining databases and entity relationship diagrams
   c. Summarizing data types and constraints
2. Manipulating data with SQL
   a. Filtering data based upon conditions
   b. Joining data tables
   c. Defining set operations
   d. Applying various functions
3. Diving into multiple data types
   a. Working with temporal data
   b. Working with views, indexes, and SQL subqueries
   c. Implementing stored procedures and SQL transactions

Software Requirements

SQL

Data Visualization Workshops

Data Visualization

Data visualization is one of the most critical tools in any analyst's toolkit. It inspires action by facilitating impactful communication of data-driven insights. In these workshops, attendees will learn how to effectively manipulate data and bring data stories to life.

Courses Include:

Introduction to Power BI

Introduction to Tableau

Storytelling with Data in Tableau

Data Sources & Examples Used

- Industry or domain-specific examples
- Customized data that maximizes the attendees' engagement and retention of skills and can be applied directly within your organization

Course Duration

- 16 hours per training
- Available in flexible schedule format

Modality

- Proposed live-streaming (in-person available post-COVID-19)
- Instructor-led (TAs are optional)

Recommendations

- Up to 30 learners per cohort
- Learners should have experience with Excel
- Python/R and other packages pre-installed

Introduction to Power BI

2 days / 16 hours of instruction (Power BI)

Data Visualization

This course is designed for students looking to take their data analysis skills to a more advanced level and learn how to communicate effectively with data. By the end of this program, students will be able to build interactive data visualizations, define best practices in data visualization, and manipulate data in Power BI.

Objectives	Prerequisites
1. Build data dashboards in Power BI 2. Import and manipulate financial data in Power BI 3. Describe the best practices for data visualization	While there are no prerequisites for this workshop, it is best suited to attendees with some experience working with data.

Syllabus & Topics Covered

1. Understanding Power BI and its capabilities
   a. What is BI and Power BI?
   b. Logical structure of Power BI
   c. Uses for Power BI across industries
2. Four layers of Power BI
   a. Layer 1: Extract, Transform, Load (ETL)
   b. Layer 2: Data Modeling
   c. Layer 3: Report Design
   d. Layer 4: Web Portal
3. Lab 1: Creating your first Power BI report
4. Lab 2: Developing the Report Layer
   a. Formatting charts
   b. Applying filters such as drillthrough
   c. Reintroducing slicers
   d. Formatting your visual
5. Storytelling with Visualization
   a. Importance of storytelling in BI
   b. Design principles
   c. Examples of storytelling
6. Lab 3: Power BI Web Services
   a. Navigating the portal
   b. Building a report online
   c. Building a dashboard
   d. Additional options

Software Requirements

Microsoft Power BI

Syllabus & Topics Covered

7. Lab 4: ETL layer
    a. Retrieving data from Excel
    b. Retrieving data from a folder of files and combine them into one table
    c. Retrieving data from an existing Power BI data model (dataset)
    d. Getting data from multiple CSV files
8. Lab 5: Navigate through the 3 Power BI data modeling tabs
9. Introduction to Power BI Formulas (DAX)
    a. Introduction to measures
    b. Building a map

Introduction to Tableau

2 days / 16 hours of instruction (Tableau)

Data Visualization

This course is designed for students looking to take their data analysis skills to a more advanced level and learn how to communicate effectively with data. By the end of this program, students will be able to build interactive data visualizations, define best practices in data visualization, and manipulate data in Tableau.

Objectives	Prerequisites
1. Build data dashboards in Tableau 2. Import and manipulate financial data in Tableau 3. Visualize geospatial data in maps	While there are no prerequisites for this workshop, it is best suited to attendees with some experience working with data.

Syllabus & Topics Covered

1. Introduction to Tableau
   a. The use cases for Tableau
   b. Importing data into Tableau
   c. Building basic visuals
   d. Manipulating filters in Tableau
2. Manipulating data in Tableau
   a. Understanding the effectiveness of functions
   b. Manipulating detail functions
   c. Implementing calculations on datasets
3. Visualizing geospatial data
   a. Defining geospatial visualization
   b. Building a data visualization map
   c. Identifying and correcting data errors
   d. Integrating maps with other platforms
4. Advanced data visualizations
   a. Defining use cases for data dashboards
   b. Combining multiple visualizations
   c. Implementing Stories in Tableau
   d. Publishing Stories to Tableau Server

Software Requirements

Tableau

Storytelling with Data in Tableau

2 days / 16 hours of instruction (Tableau*)

Data Visualization

Once students know how to build data visualizations, they can incorporate data storytelling best practices to maximize the impact of their reports and discoveries. This course covers effective techniques to map out a data presentation and practice those skills in Tableau. By the end of the program, students will be able to establish context around data, identify appropriate charts for their messaging, and communicate findings clearly to stakeholders.

Objectives	Prerequisites
1. Learn how to contextualize information 2. Design visualizations to effectively leverage sensory and short-term memory 3. Design visualizations to inform viewers accurately 4. Effectively use chart types to answer specific questions 5. Design dashboards and tell stories with data using visual best practices	Attendees need to have a strong familiarity with Tableau and using the tool to create visualizations and dashboards.

Syllabus & Topics Covered

1. Why tell stories with data?
   a. The power of persuasion, and how to do it
   b. "In God we trust, all others must bring data!" –Prof. Deming
2. Establishing editorial context
   a. Collecting dashboard requirements, managing stakeholders, and balancing priorities
   b. The visual analytics process
   c. Setting up situational context
   d. How people use memory to interpret visualizations
3. Using appropriate chart types and visual elements to convey information
   a. Chart types and what they're best for (line, bar, donut & area charts – why you should avoid pie charts)
   b. How to best visualize data and how to avoid optical illusions and other traps
   c. Balancing visual and cognitive overload – legends, labels, tooltips, and interactivity
   d. The color wheel – beauty and harmony attract attention and create receptivity
4. Reducing clutter and promoting clarity
   a. Gestalt principles of visual perception
   b. The foundations of user experience design and the interaction of form and function
   c. Painting a more concise picture – less is more
5. Speaking clearly to your stakeholders – creating the right user experience
   a. The importance of pre-attentive attributes
   b. Information hierarchy
   c. Start from the end, present from the beginning
6. Developing a dashboard – a capstone project with your data

Software Requirements

Tableau

*This course is also available for instruction in Power BI

Learning Pathway:
Introduction to R and Machine Learning

Introduction to R and Machine Learning

One of the most popular programming languages in data science, R takes care of tedious data collection and cleaning tasks. By learning how to program in R, students will be able to minimize repetitive processes and automate the creation of beautiful data visualizations. Once students have a foundation in R, they will use it to build robust machine learning models to predict trends, identify new patterns, and analyze data.

Courses Include:

Introduction to R and Visualization

Advanced and Interactive Visualization

Introduction to Clustering

Introduction to Regression

Introduction to Classification

Data Sources & Examples Used

- Industry or domain specific examples
- Customized data that maximizes the attendees' engagement and retention of skills and can be applied directly within your organization

Course Duration

- 16 / 48 Hours
- Available in flexible schedule format

Modality

- Proposed live-streaming (in-person available post-COVID-19)
- Instructor-led (TAs are optional)

Recommendations

- Up to 30 learners per cohort
- Learners should have experience with Excel
- R and other packages pre-installed

Introduction to R and Visualization

2 days / 16 hours of instruction (R programming)

Learning Pathway:
Intro to R

Students will automate their intelligence gathering with R, an easy-to-learn programming language that takes care of tedious data collection and cleaning tasks. This course will give students the foundational tools they need to work with data in R, automate monotonous processes, and build insightful visualizations. By the end of this course, students will be able to import and clean data in R, define how data science is used commercially, and quickly build data visualizations.

Objectives	Prerequisites
1. Explain use cases and potential of data science 2. Import, export, clean, and manipulate data in R 3. Create various static visualizations with R	While there are no prerequisites for this course, it is best suited to students with some experience working with data.

Syllabus & Topics Covered

1. Introduction to R and data science
    a. Introduction to R and RStudio
    b. An overview of data science
    c. Performing basic calculations in R
    d. Loading data into R
2. Fundamentals of R programming
    a. Understanding data types, how and when to use them
    b. Reading / writing data in R
    c. Evaluating and addressing missing values in data
    d. Manipulating data types and structures using flow control structures
3. Wrangling and cleaning data
    a. Transforming and cleaning data
    b. Selecting and subsetting data
    c. Summarizing and aggregating data
4. Introduction to data visualization
    a. Basic plotting in R
    b. Introduction to the 'grammar of graphics' structure
    c. Basic plotting in ggplot2
    d. Customizing graphs and adjusting formats

Software Requirements

R & RStudio

Advanced and Interactive Visualization

2 days / 12 hours of instruction (R programming)

Learning Pathway:
Intro to R

Students will take their R skills to the next level and learn how to effectively communicate the results of their analysis with a few lines of code. They'll develop a powerful toolkit to visualize data with static graphs and interactive dashboards. By the end of the course, students will be able to create interactive visualizations and publish dynamic visualizations to websites.

Objectives	Prerequisites
1. Explain use cases and potential of data science 2. Import, export, clean, and manipulate data in R 3. Create various static visualizations with R	Students must be comfortable using R to manipulate data and must know how to create basic visualizations.

Syllabus & Topics Covered

1. Visualizing univariate and bivariate data
    a. Why is data visualization important?
    b. Preparing data for visualization
    c. Visualizing transformed data using compound univariate visualizations
    d. Visualizing transformed data using compound bivariate visualizations
2. Introduction to interactive visualizations
    a. Understanding the integration of the highcharter package
    b. Creating basic interactive visualizations with transformed data
    c. Creating interactive visualizations with transformed summary data
3. Building interactive data maps
    a. Creating interactive maps utilizing JSON files
    b. Summarizing concepts of distance matrix and network visualization
    c. Building and customizing a HTMLwidget using JSON files and JavaScript
4. Introduction to RShiny
    a. Identifying RShiny tools and exploring how they improve user experience
    b. Building an interactive data dashboard
    c. Customizing data dashboards

Software Requirements

R & RStudio

Introduction to Clustering

2 days / 12 hours of instruction (R programming)

Learning Pathway: Intro to ML

Learn how to mine data and uncover patterns within it. Clustering is a foundational unsupervised machine learning technique that is key to discovering underlying patterns and trends. By the end of this course, students will learn how to identify use cases where clustering is relevant, use R to perform clustering on real-world data, and evaluate the results.

Objectives	Prerequisites
1. Mine data to find latent patterns and groups in different types of data 2. Evaluate the accuracy and effectiveness of clustering 3. Understand the purpose and implications of what clustering methods can and cannot achieve 4. Identify use cases where clustering analyses are relevant and where they are not applicable	Students must be comfortable using R to manipulate data and must know how to create basic visualizations.

Syllabus & Topics Covered

1. Introduction to clustering
   a. Commercial applications of data mining
   b. Introduction to clustering, the k-means algorithm used on voting data
2. Implementation of clustering
   a. K-means clustering on multi-dimensional data
   b. Evaluating the quality of clustering
   c. Determining the right number of clusters to use
3. Clustering categorical data
   a. Working with binary data - cosine distance
   b. Clustering binary data - spherical k-means
   c. Assessing quality of spherical k-means clustering
   d. Interpreting clusters of binary data and making recommendations
   e. Pitfalls of clustering
4. Hierarchical clustering
   a. Summary of hierarchical clustering
   b. Description of the process behind AGNES and linkage methods
   c. Implementing hierarchical clustering and visualizing its dendrogram to reveal hidden patterns
5. Principal Component Analysis (PCA)
   a. Description of high dimensional data and its challenges
   b. Definition of the concept of PCA
   c. Choosing the optimal number of dimensions to retain
   d. Identifying common pitfalls of PCA

Software Requirements

R & RStudio

Introduction to Regression

2 days / 12 hours of instruction (R programming)

Learning Pathway: Intro to ML

Learn how to predict trends with regression models, a supervised machine learning technique. This program combines the theoretical basis necessary to perform regression analysis with hands-on practice applying regression to a variety of real-world problems. By the end of this course, students will be able to identify use cases for regression, build regression models, and evaluate their results.

Objectives	Prerequisites
1. Identify opportunities and use cases for regression models 2. Build single and multiple regression models 3. Evaluate a regression model 4. Assess statistical significance and validate models for explanatory power and bias 5. Evaluate causality in variable relationships	Students must be comfortable using R programming to manipulate data and must know how to create basic visualizations.

Syllabus & Topics Covered

1. Supervised machine learning: regression
    a. Summary of regression use cases
    b. Introduction to summary statistics and hypothesis testing
    c. Definition of linear regression
    d. Building and visualizing a linear regression model
    e. Evaluating a linear model and assumptions
2. Advanced regression
    a. Building multiple regression models
    b. Optimizing and evaluating multiple regression models with RMSE
    c. Building nonlinear regression models
    d. Optimizing and evaluating nonlinear regression models

Software Requirements

R & RStudio

Introduction to Classification

2 days / 12 hours of instruction (R programming)

Learning Pathway:
Intro to ML

Classification algorithms are powerful and intuitive data science tools that can predict behaviors and trends. This course teaches students how to implement algorithms like k-Nearest Neighbors and logistic regression, as well as evaluate and interpret their results. By the end of the course, students will be able to build classification models to anticipate events and assess the accuracy of predictive algorithms.

Objectives	Prerequisites
1. Identify opportunities and use cases for predictive analytics 2. Build classification models to anticipate events and behaviors 3. Evaluate accuracy of predictive algorithms	Students must be comfortable using R programming to manipulate data and must know how to create basic visualizations.

Syllabus & Topics Covered

1. Introduction to classification and supervised machine learning
   a. Commercial applications of classification models and predictive analytics
2. Building kNN models and performance metrics
   a. Definition of kNN technique and application
   b. Applying cross-validation to the kNN algorithm
   c. Identifying and evaluating performance metrics
3. Logistic regression
   a. Introduction to logistic regression and its relation to neural networks
   b. Assessing classifier performance (ROC curve, AUC, cutoff value selection)
   c. Tuning & regularization of logistic regression
4. Decision trees
   a. Decision trees: Gini coefficient and information gain
   b. Confusion matrices and misclassification rates
   c. Evaluating the model for overfitting and underfitting

Software Requirements

R & RStudio

Learning Pathway:
Introduction to Python and Machine Learning

Introduction to Python and Machine Learning

This foundational pathway provides the key building blocks to becoming a successful data analyst. Students will learn how to leverage Python—one of the most widely used programming languages—to import data, automate data cleaning, and build powerful visualizations. Then, they will use their new skills to build unsupervised and supervised machine learning models to find new patterns and make predictions.

Courses Include:

Introduction to Python and Machine Learning

Data Visualization in Python

Introduction to Clustering

Introduction to Regression

Introduction to Classification

Data Sources & Examples Used

- Industry or domain specific examples
- Customized data that maximizes the attendees' engagement and retention of skills and can be applied directly within your organization

Course Duration

- 72 Hours
- Available in flexible schedule format

Modality

- Proposed live-streaming (in-person available post-COVID-19)
- Instructor-led (TAs are optional)

Recommendations

- Up to 30 learners per cohort
- Learners should have experience with Excel
- Python and other packages pre-installed

Introduction to Python

2 days / 16 hours of instruction (Python)

Learning Pathway: Intro to Python

Python programming is among the most powerful and widely used tools by data scientists today. This course will cover how to leverage Python's capabilities to manipulate and explore data. By the end of this program, students will be able to outline use cases for Python and gain foundational skills to automate tedious data processes.

Objectives	Prerequisites
1. Identify what data science is and why it's important 2. Program proficiently in Python 3. Automate data cleaning and processing	While there are no prerequisites for this course, it is best suited to students with some experience working with data and programming.

Syllabus & Topics Covered

1. Introduction to Python and programming
   a. Introduction to Anaconda and Jupyter notebook
   b. Using basic operations and variables in Python
   c. Defining and identifying foundational data types and structures
2. Working with data in Python
   a. Understanding data types, how and when to use them
   b. Operations in Python: variables, strings, operations, loops, conditionals, functions
   c. Container objects: tuples, lists, dictionaries, comprehensions
3. Manipulating data sets for analysis
   a. Introduction to data mining: Pandas, NumPy
   b. Selecting and subsetting data
   c. Summarizing and aggregating data

Software Requirements

Python & Anaconda

Data Visualization in Python

2 days / 16 hours of instruction (Python)

Learning Pathway: Intro to Python

Nothing is more powerful than effective data visualizations. Students in this two-day course will develop a powerful toolkit to visualize data with static graphs and interactive charts. The program will help students become expert data communicators to make their analyses shine. By the end of the course, they will be able to manipulate and summarize a variety of file formats and build intuitive data visualizations with Python.

Objectives	Prerequisites
1. Build static data visualizations in Python 2. Build interactive data visualizations in Python 3. Select the appropriate data visualization for the scenario	Students must be comfortable using Python to manipulate data and perform basic operations.

Syllabus & Topics Covered

1. Data visualization in Python
   a. Wrangling data for visualization
   b. Building univariate and bivariate plots
   c. Customizing data visualizations
   d. Saving and publishing plots
2. Advanced data visualization
   a. Building interactive visualizations and graphs
   b. Best practices of data visualization
3. Data visualization best practices
   a. Choosing the right visualization based on the data
   b. Using the right visual elements to tell your story
   c. Selecting appropriate combinations of colors to convey results effectively

Software Requirements

Python & Anaconda

Introduction to Clustering

2 days / 12 hours of instruction (Python)

Learning Pathway: Intro to ML

This program will teach students how to mine data and uncover patterns within it. Clustering is a foundational unsupervised machine learning technique that is key to discovering latent patterns and trends. By the end of this two-day course, students will learn to identify use cases where clustering is relevant, use Python to perform clustering on real-world data, and evaluate the results.

Objectives	Prerequisites
1. Mine data to find latent patterns and groups in different types of data 2. Evaluate the accuracy and effectiveness of clustering 3. Understand the purpose and implications of what clustering methods can and cannot achieve 4. Identify use cases where clustering analyses are relevant and where they are not applicable	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations.

Syllabus & Topics Covered

1. Introduction to clustering
   a. Commercial applications of data mining
   b. Introduction to clustering, the k-means algorithm used on voting data
2. Implementation of clustering
   a. K-means clustering on multi-dimensional data
   b. Evaluating the quality of clustering
   c. Determining the right number of clusters to use
3. Clustering categorical data
   a. Working with binary data - cosine distance
   b. Clustering binary data - spherical k-means
   c. Assessing quality of spherical k-means clustering
   d. Interpreting clusters of binary data and making recommendations
   e. Pitfalls of clustering
4. Hierarchical clustering
   a. Summary of hierarchical clustering and the two types
   b. Understanding the process behind AGNES and linkage methods
   c. Implementing hierarchical clustering on the data and visualizing its dendrogram
5. Principal Component Analysis (PCA)
   a. Understanding high dimensional data and its challenges
   b. Defining the concept of PCA
   c. Choosing the optimal number of dimensions to retain
   d. Identifying common pitfalls of PCA

Software Requirements

Python & Anaconda

Introduction to Regression

2 days / 16 hours of instruction (Python)

Learning Pathway: Intro to ML

This course will teach students how to predict trends with regression models, a supervised machine learning technique. This two-day program combines the theoretical basis necessary to perform regression analysis with a hands-on practical approach to applying the learned methods and techniques to real-world problems. By the end of this course, students will be able to identify use cases for regression, build regression models, and evaluate their results.

Objectives	Prerequisites
1. Identify opportunities and use cases for regression models 2. Build single and multiple regression models 3. Evaluate a regression model 4. Assess statistical significance and validate models for explanatory power and bias 5. Evaluate causality in variable relationships	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations.

Syllabus & Topics Covered

1. Supervised machine learning: regression
   a. Summary of regression use cases
   b. Introduction to summary statistics and hypothesis testing
   c. Definition of linear regression
   d. Building a linear regression model and visualizing it
   e. Evaluating a linear model and assumptions
2. Advanced regression
   a. Building multiple regression models
   b. Optimizing and evaluating multiple regression models with RMSE
   c. Building nonlinear regression models
   d. Optimizing and evaluating nonlinear regression models

Software Requirements

Python & Anaconda

Introduction to Classification

2 days / 12 hours of instruction (Python)

Learning Pathway: Intro to ML

Classification algorithms, one of the most powerful and intuitive data science tools, can predict behaviors and trends. This two-day course teaches how to implement algorithms like k-Nearest Neighbors and Random Forest, as well as evaluate and interpret the results. By the end of the course, students will be able to build classification models to anticipate events and evaluate the accuracy of predictive algorithms.

Objectives	Prerequisites
1. Identify opportunities and use cases for predictive analytics 2. Build classification models to anticipate events and behaviors 3. Evaluate accuracy of predictive algorithms	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations.

Syllabus & Topics Covered

1. Introduction to classification and supervised machine learning
   a. Commercial applications of classification models and predictive analytics
2. Building kNN models and performance metrics
   a. Definition of kNN technique and application
   b. Applying cross-validation to the kNN algorithm
   c. Identifying and evaluating performance metrics
3. Logistic regression
   a. Introduction to logistic regression and its relation to neural networks
   b. Assessing classifier performance (ROC curve, AUC, cutoff value selection)
   c. Tuning & regularization of logistic regression
4. Decision trees
   a. Decision trees: Gini coefficient and information gain
   b. Confusion matrices and misclassification rates
   c. Evaluating the model for overfitting and underfitting

Software Requirements

Python & Anaconda

Learning Pathway: Intermediate Machine Learning

Intermediate Machine Learning

After developing a robust foundation in Python and machine learning skills, it's time to incorporate those algorithms into one of the most powerful and informative topics: text mining. This set of courses builds upon the Introduction to Machine Learning pathway and teaches students more advanced techniques they can practice in either R or Python.

Courses Include

- Advanced Classification
- Introduction to Time Series Analysis
- Advanced Time Series Analysis
- Introduction to Text Mining
- Intermediate Text Mining & NLP
- Feature Engineering and Fraud Detection

Data Sources & Examples Used

- Industry or domain specific examples
- Customized data that maximizes the attendees' engagement and retention of skills and can be applied directly within your organization

Course Duration

- 8 / 16 Hours
- Available in flexible schedule format

Modality

- Proposed live-streaming (in-person available post-COVID-19)
- Instructor-led (TAs are optional)

Recommendations

- Up to 30 learners per cohort
- Learners should have experience with Excel
- Python or R and other packages pre-installed

Advanced Classification

2 days / 12 hours of instruction (Python*)

Intermediate ML

This course covers advanced topics in classification, including classification model parameter tuning and ensemble learning methods – specifically boosting. Students will learn how to make classification models perform better and how to choose the correct methods and parameters based on the problem at hand.

Objectives	Prerequisites
1. Develop advanced classification models with increased accuracy 2. Fine tune parameter weights and perform advanced feature selection	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in classification models and model accuracy measures.

Syllabus & Topics Covered

1. Ensemble methods - random forests and gradient boosting
   a. Random forests
   b. Boosting methods
   c. Grid search and optimization of ensemble methods
2. Support vector classifiers
   a. Defining the concept of a hyperplane
   b. Understanding the concept of support vectors
   c. Building a support vector classifier model
3. Support vector machines
   a. Summary of the key differences between support vector classifier and support vector machine
   b. Building a support vector machine model
   c. Optimizing the support vector machine model using grid search
   d. Identifying the pitfalls of support vector machines

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Introduction to Time Series Analysis

2 days / 12 hours of instruction (Python*)

Intermediate ML

Once analysis reveals patterns in data, the next step is to predict trends about the future. Students will learn how to apply time series models to create accurate forecasts for their organizations and make better decisions when developing strategies. By the end of this course, students will be able to build and visualize time series models, describe key components and pitfalls, and identify hidden patterns within data.

Objectives	Prerequisites
1. Describe how time series can be used effectively in commercial applications and industry 2. Process, clean, and format data for analysis 3. Build time series models to identify and predict trends	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in regression and foundational statistics (i.e. correlation / covariance).

Syllabus & Topics Covered

1. Introduction to time series
   a. Introduction to forecasting and time series analysis use cases
   b. Review of key quantitative forecasting methods
2. Modeling time series data
   a. Visualizing time series
   b. Description of time series modeling basics and key components of time series
   c. Measuring linear relationships within time series
   d. Estimating and visualizing autocorrelation of time series
3. Predicting trends in white noise
   a. Integrating the concepts and implications of a white noise series
   b. Assessing the forecasted models based on comparison criteria
   c. Explaining the concept of a random walk model and stationarity
   d. Making predictions using random walk model and measure error

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Advanced Time Series Analysis

2 days / 12 hours of instruction (Python*)

Intermediate ML

Building time series models is an effective way to predict trends, but how is it possible to anticipate the future when there are additional complexities? That's where seasonal analysis and ARIMA models come in. Students in this program will learn how to decompose and identify seasonal and non-seasonal factors. By the end of this course, they will be able to build and visualize complex time series models.

Objectives	Prerequisites
1. Define key metrics of complex time series models 2. Develop models that account for seasonal trends and other factors 3. Build and evaluate ARIMA models	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in regression models and time series models.

Syllabus & Topics Covered

1. Introduction to seasonality
    a. Understanding the concept of seasonality and trend in a time series model
    b. Visualizing time series data for different time periods using line and boxplots
    c. Definition of moving averages and how they can be used to decompose time series
    d. Deseasonalizing and detrending a time series model using moving averages
2. Building ARIMA models
    a. Review of the concept of stationarity
    b. Testing decomposed series for stationarity
    c. Understanding the components of an ARIMA model
    d. Description of a non-seasonal ARIMA model
    e. Using a non-seasonal ARIMA model to forecast using deseasonalized series

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Introduction to Text Mining

2 days / 16 hours of instruction (Python*)

Intermediate ML

Do you spend your days reading through reports and summaries? Do you want to be able to highlight key phrases and extract meaning automatically? This two-day course will give students the foundational skills they need in order to process, clean, and format text data for analysis. By the end of this course, students will be able to integrate text mining into their work and extract key summary metrics and words from a corpus of documents.

Objectives	Prerequisites
1. Describe how text mining can be used effectively in commercial applications and industry 2. Process, clean, and format text data for analysis 3. Extract key summary metrics and words from a corpus of documents	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in classification models and model accuracy measures.

Syllabus & Topics Covered

1. Preparing data for text mining
   a. What are text mining and Natural Language Processing (NLP)?
   b. Applications and intuition of text mining
   c. The process of cleaning and preparing text for analysis
   d. Working with different data formats – PDF, CSV, TXT
2. Visualizing text and understanding text distributions
   a. N-grams (bi-grams, tri-grams and quad-grams)
   b. Word clouds
   c. Histograms
   d. Summary metrics of corpora
3. Topic modeling within text
   a. Building a term document matrix
   b. Implementing bag-of-words technique on text data
   c. Building out TF-IDF
   d. Summary of topic modeling and implementing LDA
   e. Evaluating results and optimize number of topics

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Intermediate Text Mining and NLP

2 days / 16 hours of instruction (Python*)

Intermediate ML

How can you calculate the underlying emotions of text data? This two-day course will give students the skills to calculate sentiment in text data and identify similar documents. Students will learn how to generate context from language patterns and Natural Language Processing. By the end of this course, they'll be able to analyze the structure and morphology of text, identify similar documents, and visualize the connections.

Objectives	Prerequisites
1. Extract key entities and subject matter from documents 2. Develop sentiment analyses that explain degrees of sentiment with respect to specific topics and ideas 3. Develop accurate topic models that incorporate sentiment	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in text parsing and cleaning and in text feature selection.

Syllabus & Topics Covered

1. Text mining analysis
   a. Description of cosine similarity in text
   b. Building interactive network graphs to visualize similar documents
   c. Applying hierarchical clustering to text data
   d. Visualizing clustering as dendrogram and evaluate results
2. Introduction to sentiment analysis
   a. Introduction to sentiment analysis and word dictionaries
   b. Dispersion and correlation of key words to amplify sentiment analysis
3. Word embeddings
   a. Vectorizing words to identify similarities
   b. Implementing GloVe technique
   c. Implementing BERT technique

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Feature Engineering and Fraud Detection

2 day / 12 hours of instruction (Python*)

Intermediate ML

Identifying the most impactful features for your model and detecting fraudulent data points are powerful machine learning techniques. This class will build upon foundational machine learning techniques to hone predictive skills and discover critical danger points in patterns. By the end of this course, students will be able to determine key features in models and identify anomalous data points.

Objectives	Prerequisites
1. Define use cases for feature engineering and fraud detection 2. Identify and evaluate the most impactful numerical and categorical variables 3. Develop accurate anomaly detection models	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in classification techniques.

Syllabus & Topics Covered

1. Engineering numerical variables
   a. Introduction to feature engineering and its use cases
   b. Summary of feature engineering and its usefulness
   c. Explanation of the feature engineering of numerical variables
2. Engineering categorical variables
   a. Description of the feature engineering of categorical variables
   b. Handling temporal and spatial predictors in the data
   c. Definition of the principle of parsimony and feature selection
3. Introduction to anomaly detection
   a. Definition of anomaly concepts and their uses
   b. Differentiating between types of anomalies
   c. Creation of a problem statement for fraud dataset
4. Implementing anomaly detection
   a. Summary of different techniques for anomaly detection
   b. Identifying SMOTE analysis and its implementation
   c. Description of the isolation forest algorithm
   d. Implementing isolation forest to detect credit fraud

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Learning Pathway: Advanced Machine Learning

Advanced Machine Learning

This pathway dives into a host of algorithms known as artificial neural networks. In addition to learning the foundations of this complex and exciting topic, students will acquire practical skills to implement neural networks and deep learning models using TensorFlow and Keras libraries in order to solve real-world problems. By the end of the program, students will be able to analyze images for the purpose of automatic object classification, automate the analysis of algorithms, and develop neural networks.

Courses Include

- Recommender Systems
- Network Analytics and Community Detection
- Intermediate Scala and Spark
- Introduction to Neural Networks and Deep Learning
- Deep Learning for Image Recognition
- Deep Learning for Text Analysis
- Advanced Deep Learning

Data Sources & Examples Used

- Industry or domain specific examples
- Customized data that maximizes the attendees' engagement and retention of skills and can be applied directly within your organization

Course Duration

- 32 Hours
- Available in flexible schedule format

Modality

- Proposed live-streaming (in-person available post-COVID-19)
- Instructor-led with support from TAs

Recommendations

- Up to 30 learners per cohort
- Learners should have experience with Excel
- Python and other packages pre-installed

Recommender Systems

1 day / 8 hours of instruction (Python*)

Advanced ML

Netflix, Amazon, Facebook – those are just some of the businesses that leverage recommendations to better serve their target markets. This one-day course provides an overview of how recommender systems work and teaches students how to build effective models. By the end of this course, students will be able to explain the key assumptions underlying recommender systems and build and evaluate them based on real data.

Objectives	Prerequisites
1. Identify and define use cases for recommender systems 2. Build and evaluate content-based recommender systems 3. Build and evaluate item-based filtering algorithm	Students must be comfortable using Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in classification techniques.

Syllabus & Topics Covered

1. Introduction to recommender systems
   a. Summary of the usages and types of recommendation engines
   b. Description of the concept of a content-based recommender system
2. Building a recommender system
   a. Building a content-based recommender system
   b. Generating recommendations from the content-based recommender system and discuss the pitfalls
   c. Description of the collaborative filtering recommender system and its types
   d. Building an item-based collaborative filtering algorithm
   e. Evaluating the model using performance metrics

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Network Analytics and Community Detection

2 days / 16 hours of instruction (Python*)

Advanced ML

Learn how to identify communities and like-minded individuals and use the PageRank algorithm (core piece of Google's search engine) to measure a person's significance in their network. By the end of this course, students will be able to quantitatively measure the strength of connections between network nodes, detect communities within networks, and identify the most important nodes in a network.

Objectives	Prerequisites
1. Define networks and identify use cases 2. Build and visualize foundational network graphs 3. Measure and evaluate the strengths and attributes of networks 4. Identify and evaluate communities within networks	Students must be comfortable using R / Python to manipulate data and must know how to create basic visualizations. Additionally, students must have a foundation in classification and clustering algorithms.

Syllabus & Topics Covered

1. Introduction to networks
   a. Definition of networks and their use cases
   b. Identification of case studies where networks are used
2. Building a network graph
   a. Definition of a network graph and its components
   b. Description of network types and visualization of different network graphs
   c. Building a base network, exploring and visualizing it
3. Measuring networks
   a. Ways to measure a network
   b. Calculating in-degree, out-degree, total degree, and net degree of a network
   c. Defining, calculating, and interpreting centrality metrics
   d. Trimming the network based on the attributes of nodes and edges
   e. Visualizing network with attributes and scores
   f. Testing network resilience and summarize findings
4. Introduction to community detection
   a. Introduction to community detection and how it relates to network analysis
   b. Summary of the community detection use case and different algorithms
5. Identifying communities
   a. Introduction and implementation of Louvain community detection algorithm
   b. Discussion of maximal clique percolation algorithm
   c. Explanation and implementation of label propagation algorithm
   d. Comparison of different community detection algorithms

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Intermediate Scala and Spark

3 days / 24 hours of instruction

Advanced ML

Learn how to optimize your code and to speed up current data processing using Scala and Spark. In this course, students will work through best practices of using Scala and how and when to use Spark. They will explore what they can do with Spark and how to use distributed computing within Spark. Finally, students will deploy a Spark application on AWS, learning how to execute a job with Spark on AWS.

Objectives	Prerequisites
1. Be able to write complex queries using Scala and Spark 2. Make correct use of indexing so that queries are fast and effective 3. Deploy applications 4. Introduction to parallel processing with Spark	Students must have experience with Python or another object-oriented language, foundational knowledge of Scala programming, and foundational understanding of Spark.

Syllabus & Topics Covered

1. Introduction to Zeppelin
   a. Working with Zeppelin notebook
2. Collections in Scala
   a. Classification of collections
   b. Explain differences between linear vs. indexed collections
   c. Explain differences between mutable vs. immutable collections
3. Key mutable and immutable collections in Scala
   a. Working with immutable collections: Lists, LazyLists, Queues, Vectors, Maps, and Sets
   b. Working with mutable collections: ListBuffers, Arrays, ArrayBuffers, Stacks, Maps, and Sets
4. Working with data in Spark
   a. RDDs vs. DataFrames vs. Datasets
   b. When to use RDDs over DataFrames & Datasets
5. Features of Dataset in Spark
   a. Optimized queries with Datasets
   b. API for Datasets
6. Workshop: work with Datasets in Spark
7. Using Spark's distributed features
   a. Parallel processing in Spark
   b. Using Scala parallel collections
   c. Spark partitions
8. Building ML pipeline on a cluster on AWS
   a. Deploying a cluster on AWS
   b. ML pipeline architecture using Scala/Spark
9. Spark application
   a. Creating a Spark application
   b. Deploying a Spark application on a cluster
10. Workshop: deploy Spark application on AWS cluster
11. Spark persistence and execution workflow in Spark
    a. Spark execution flow
    b. Batch processing vs. real time processing
    c. The caching mechanism
12. Improving Spark performance
    a. Working with shared variables: broadcast and accumulators
    b. Common performance issues
    c. Troubleshooting performance problems
13. Scheduling & Partitioning
    a. Scheduling and partitioning in Spark to optimize processes
    b. Different types of partitioning
    c. Using Apache ZooKeeper as a a centralized service for maintaining configuration information
14. Workshop: deploy fast applications with Scala and Spark and tie it all together

Software Requirements

Scala / Spark

Introduction to Neural Networks & Deep Learning

Advanced ML

16 hours of instruction (Python*)

These state-of-the-art methods build powerful predictive systems and find latent patterns in large amounts of data. By the end of this course, students will learn the foundations of this complex topic and acquire practical skills to build neural networks in order to solve real-world problems.

Objectives	Prerequisites
1. Define powerful applications and use cases that incorporate deep learning 2. Build foundational neural network models 3. Implement best practices on deep learning models	Any student in this program must have a strong foundation in Python and the common libraries: SciKit-Learn, Pandas, NumPy, and Matplotlib. Additionally, students should have a foundational knowledge of statistics, unsupervised machine learning algorithms, and classification algorithms.

Syllabus & Topics Covered

1. Introduction to the concept of neural networks
   a. Environment setup: TensorFlow 2.0
   b. Why TensorFlow?
   c. Introduction to neural networks
   d. Implement a neural network in Python using sklearn
2. Feed forward networks
   a. Forward and back propagation
   b. Single layer perceptron
   c. Implementation of models in TF
3. Best practices of model building – deep learning
   a. Batch learning
   b. GPU and when it is needed
   c. Data normalization
   d. Accuracy, precision, recall, F1 against “loss”

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Deep Learning for Image Recognition

16 hours of instruction (Python*)

Advanced ML

These state-of-the-art methods build powerful predictive systems and find latent patterns in large amounts of data. By the end of this course, students will be able to work with image data, understanding various methods to work with and recognize images. Starting with simpler methods, this course will span complex methods as well.

Objectives	Prerequisites
1. Define use cases for image analysis 2. Analyze images for the purpose of automatic object classification 3. Build deep learning models	Any student in this program must have a strong foundation in Python and the common libraries: SciKit-Learn, Pandas, NumPy, and Matplotlib. Additionally, students should be comfortable building machine learning models and foundational neural network algorithms.

Syllabus & Topics Covered

1. Autoencoders and Image classification – CNN
   a. Autoencoders and feature reduction
   b. Overview of CNNs and image data
   c. CNN architecture
   d. Simple CNNs – build and implement
   e. Optimizing CNNs
2. Advanced CNNs, transfer learning, and object detection
   a. Understanding the difference between CNNs and advanced CNNs
   b. Review advanced architecture – deep models
   c. VGG16, Inception and ResNet
   d. Transfer learning
   e. YOLO – object detection

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Deep Learning for Text Analysis

16 hours of instruction (Python*)

Advanced ML

These state-of-the-art methods build powerful predictive systems and find latent patterns in large amounts of data. By the end of this course, students will be able to work with text data, understanding various methods to work with text. Starting with more simple methods, this course will span complex methods as well.

Objectives	Prerequisites
1. Build powerful predictive systems using deep learning 2. Build natural language understanding and generation models 3. Analyze images for the purpose of automatic object classification 4. Automate the analysis of time-series data	Any student in this program must have a strong foundation in Python and the common libraries: SciKit-Learn, Pandas, NumPy, and Matplotlib. Additionally, students should be comfortable building machine learning models and foundational neural network algorithms.

Syllabus & Topics Covered

1. Long short-term memory (LSTMs)
   a. Long Short-Term Memory (LSTM) theory
   b. LSTMs for text data
   c. LSTM implementation in TensorFlow
2. Seq2seq and Gated recurrent unit (GRU) models
   a. Gated recurrent unit (GRU) in TensorFlow
   b. Implementing GRU in TensorFlow
   c. Implement the concept of stateful GRU
   d. Seq2seq theory
   e. Implement seq2seq in TensorFlow

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

Advanced Deep Learning

24 hours of instruction (Python*)

Advanced ML

These state-of-the-art methods build powerful predictive systems and find latent patterns in large amounts of data. By the end of this course, students will learn the advanced methods of this complex topic, acquire practical skills to implement neural networks, and build deep learning models using TensorFlow and Keras libraries in order to solve real-world problems.

Objectives	Prerequisites
1. Build powerful predictive systems using deep learning 2. Understand and work with multiple GPUs 3. Make a sequence of decisions by implementing reinforcement learning 4. Use Bayesian probability theory within deep learning to map uncertainties 5. Train models against each other to improve overall performance of the model using GANs	Any student in this program must have a strong foundation in Python and the common libraries: SciKit-Learn, Pandas, NumPy, and Matplotlib. Additionally, students should be comfortable building machine learning models and foundational neural network algorithms.

Syllabus & Topics Covered

1. Deep learning with multiple GPUs and geospatial data
   a. Demonstrate use of cloud computing tools to process and build models on large datasets
   b. Multi GPU training
   c. Dealing with geospatial data
2. Reinforcement learning
   a. Reinforcement learning (RL) theory
   b. Implement RL using TensorFlow
   c. Expand CNNs to accomplish RNN tasks
3. Bayesian leaning
   a. Bayesian learning theory
   b. Naïve Bayes algorithm and application in deep learning
   c. Apply Bayesian inference to shorten training time for neural networks
4. Generative adversarial networks (GANs)
   a. GANs theory and different types of GAN models
   b. Implement GAN using TensorFlow
   c. Main challenges of training GANs

Software Requirements

Python & Anaconda

*This course is also available for instruction in R.

About DATA SOCIETY®

Data Society is a Washington, D.C.-based analytics training and advisory firm that provides data science training to executives and data scientists to eliminate communication gaps, improve data quality, and increase efficiencies through data-driven decision-making.

Clients include: