INTRO TO THE FDP PROGRAM
April 4 – April 18, 2022

Create your FDP profile and gain access to the complete FDP Candidate Study Guide

Brought to you by: CAIA ASSOCIATION®

"If, after reviewing this document, you conclude that the FDP Charter could be an essential step in your continued education and can help you gain a competitive edge in the data-driven finance industry, visit the FDP website, and create your FDP profile, and receive a copy of the complete study guide. Once you have decided to pursue the FDP Charter, register for the exam, and gain access to the sample exam and additional learning tools."
– *Hossein Kazemi*

INTRODUCTION TO THE FINANCIAL DATA PROFESSIONAL (FDP) PROGRAM

The FDP Institute® was founded by the Chartered Alternative Investment Analyst Association® to create the FDP® charter. It is the only globally recognized professional designation in financial data science, an increasingly important part of the financial services industry.

In recent years, the financial industry has been disrupted by the digital revolution. It is critical for industry practitioners to have a working knowledge of the increasingly important roles played by big data, machine learning, and artificial intelligence in the financial industry. The FDP Institute has designed this self-study program to provide the finance professional with an efficient path to learn about financial data science's essential aspects. The FDP curriculum introduces Candidates to the central concepts of machine learning and big data, including ethical and privacy issues and their roles in various financial industry segments. Candidates will earn their FDP Charter once they pass the FDP exam and complete the online class requirements, which can be done before or after the FDP exam.

The university faculty and industry practitioners who have helped create the FDP Charter program bring years of experience in the financial services industry. Consequently, the curriculum is consistent with recent advances in data science applications to the financial industry. Passing the FDP examination is an important accomplishment and will require a significant amount of preparation. All Candidates will need to study and become familiar with the FDP curriculum material to develop the knowledge and skills necessary to be successful on examination day.

This study guide is organized to facilitate quick learning and easy retention. Each topic is structured around learning objectives that define the content to be tested on the exam. The learning objectives are an important way for Candidates to organize their studies as they form the basis for examination questions. All learning objectives reflect the FDP curriculum content, and all exam questions are written to address the learning objectives directly. A Candidate who can meet all learning objectives in the study guide should be well prepared for the exam. For these reasons, we believe that the FDP Institute has built a rigorous program with high standards while also maintaining an awareness of the value of Candidates' time.

Candidates for the FDP Charter are required to complete both the FDP exam and the online requirements. Since the FDP program is designed for finance professionals, it is assumed that Candidates understand the central concepts of financial economics. This includes awareness of the roles and characteristics of various financial institutions and instruments, and the financial models employed by these institutions to value the instruments and measure risk. These concepts are covered in CAIA®, CFA®, and FRM® exams, and dedicated undergraduate or graduate courses covering financial markets, investments, and risk management.

FDP PROGRAM: ONLINE REQUIREMENTS

While the FDP exam will not contain any coding questions, all Candidates must complete the following two components with a passing score before obtaining their FDP charters:

- **FDP Exam**
- **Two online classes covering the basics of Python or R programming, or the single class offered by Metis.**

The online class requirements can be completed before or after a Candidate sits for the FDP exam.

Three approved providers offer online classes. Depending on the Candidate's background, the online classes are estimated to take 8-10 hours. No programming background is required to complete the online classes.

As of now, pre-selected online classes offered by the following organizations have been approved by the FDP Institute.

- **Datacamp:** https://www.datacamp.com
- **Dataquest:** https://www.dataquest.io
- **Metis:** https://www.thisismetis.com

The list of online classes for each approved online provider appears on FDP Institute's website and in this Study Guide.

The approved online classes offered by Dataquest and DataCamp are available as soon as a Candidate registers on their respective sites. The approved class offered through Metis is offered once a month throughout the year. All three providers offer limited free access to their classes. Candidates should take advantage of the limited free access to determine which platform's approach is more suited to their needs. *Candidates cannot mix and match classes from different providers*. Finally, the three platforms' approved classes assume no prior knowledge of Python, R, or any specific computer programming language.

The Candidate Handbook, which can be found on FDP's website, describes the procedure for sending proof of successful completion of the online classes to the FDP Institute. The following classes should be completed to satisfy the FDP Charter requirements.

DataCamp

Candidates can access DataCamp classes through their website at https://www.datacamp.com/. Candidates are responsible for the cost of classes offered at DataCamp. Candidates are encouraged to take advantage of limited free access offered by DataCamp to evaluate its teaching method. The classes listed below are short and, depending on the Candidate's background, each one can be completed within 4 to 6 hours. Candidates can choose between either two (2) R or two (2) Python classes.

DataCamp: Python

1. Introduction to Python

Python is a general-purpose programming language that is becoming very popular for data science. Companies worldwide are using Python to harvest insights from their data and gain a competitive edge. Unlike other Python tutorials, this course focuses on Python specifically for data science. In this Introduction to Python course, you will learn about powerful ways to store and manipulate data and helpful data science tools to begin conducting your own analyses. https://www.datacamp.com/courses/intro-to-python-for-data-science

2. Intermediate Python for Data Science

Intermediate Python for Data Science is crucial for any aspiring data science practitioner. Learn to visualize real data with Matplotlib's functions and get acquainted with data structures such as the dictionary and the pandas DataFrame. After covering key concepts such as Boolean logic, control flow, and loops in Python, you will be ready to blend everything you've learned to solve a case study using hacker statistics. https://www.datacamp.com/courses/intermediate-python-for-data-science

DataCamp: R

1. Introduction to R

In Introduction to R, you will master the basics of this widely used open-source language, including factors, lists, and data frames. With the knowledge gained in this course, you will be ready to undertake your first data analysis. Oracle estimated over 2 million R users worldwide in 2012, cementing R as a leading programming language in statistics and data science. Every year the number of R users grows by about 40%, and an increasing number of organizations are using it in their day-to-day activities. https://www.datacamp.com/courses/free-introduction-to-r

2. Intermediate R

Intermediate R is the next stop on your journey to mastering the R programming language. In this R training you will learn about conditional statements, loops, and functions to power your R scripts. Next, make your R code more efficient and readable using the application functions. Finally, the chapter on utilities gets you up to speed with regular expressions in R, data structure manipulations, times and dates. This course will allow you to take the next step in advancing your overall knowledge and capabilities while programming in R. https://www.datacamp.com/courses/intermediate-r

DataQuest

Candidates can access DataQuest classes through their website at https://www.dataquest.io/. Candidates are responsible for the cost of classes offered at DataQuest. Candidates are encouraged to take advantage of limited free access offered by DataQuest to evaluate its teaching method. The classes listed below are short and, depending on the Candidate's background, each one can be completed within 4 -to 6 hours. Candidates can choose between either two (2) R or two (2) Python classes.

DataQuest: Python

1. **Python for Data Science: Fundamentals**
   In our introductory course on Python for data science, you will get an overview of the Python programming language and how you can use it for data science. You will learn to code using real-world mobile app data while learning key Python concepts such as lists and loops. You will also learn how to update variables, how to work with different kinds of data, how to manipulate Python dictionaries, and how to use custom functions to speed up your workflow. Additionally, we will cover some coding best practices that'll help you build good habits right from the start and show you how to use Jupyter Notebook, a popular tool used in the Data Science workflows, to share data science projects easily. At the end of the course, you will combine all the skills you have learned to create your data science portfolio project. In this guided project, you will analyze different app profiles on the iOS App Store to make recommendations for the most profitable types of apps to develop.
   https://www.dataquest.io/course/python-for-data-science-fundamentals/

2. **Python for Data Science: Intermediate**
   In our Python for Data Science Intermediate course, we will cover some essential techniques for working with the Python programming language for data science. To start, you will learn how to clean and prepare data in Python, a critical skill for any data analyst or data scientist job. To do this, you'll dig into some real-world data about the artwork at the Museum of Modern Art and learn to manipulate text, clean messy data, and more. You will also get to practice summarizing numerical data and formatting strings in Python. Next, you will unlock Python's true power as we dive into object-oriented programming (OOP) and how it relates to data science. Then, you will apply this new understanding by building your class. Finally, you will learn how clean, standardize, and analyze time-series data using Python's datetime module. At the end of the course, you will combine all the skills you learned to create a portfolio project centered around Hacker News post titles to find out what types of posts are most likely to be successful at what times.
   https://www.dataquest.io/course/python-for-data-science-intermediate

DataQuest: R

1. **Introduction to Data Analysis in R**

   In the world of data science, R is a popular programming language for a reason. It was built with statistical manipulation in mind, and there is an incredible ecosystem of packages for R that let you do amazing things – particularly in data visualization – that would be much more difficult in Python.

   This is the first course in the DataQuest Data Analyst in R path, and in it, you will be learning about the fundamentals of R. You will learn to use variables, operators and write logical expressions. You will also learn about the data analysis workflow in R.

   As you learn these new R programming skills, you will be writing your code to practice them right in your browser window. And you will learn all of this while working with real-world data, much as you would for a real data science project.

   We will also cover how to install packages to extend R's functionality for working with dataframes, a crucial skill for extending your data science toolkit. And you will learn the basics of using R Studio, which is a popular free and open-source development environment that is widely used in the R data science community so that you can easily share projects. https://www.dataquest.io/course/introduction-to-data-analysis-in-r/

2. **Data Structures in R**

   This is the second course in the Data Analyst in R path. In it, you will build on the R programming skills you learned in the first course as you start to work with some of the most common data structures in R: vectors, matrices, lists, and dataframes.

   As you learn these new R programming skills, you will be writing code to practice them right in your browser window. And you will learn all of this while working with real-world data, much as you would for a real data science project.

   At the end of the course, you will be ready to dig into your first real data science project: an analysis of COVID-19 virus trends. Our guided projects will challenge you to synthesize and apply everything you have learned while still providing enough direction that you know where to go! https://www.dataquest.io/course/data-structures-in-r/

Metis

Candidates can access the Metis course through their website at https://www.thisismetis.com/. Candidates are responsible for the cost of the course offered at Metis. Candidates are encouraged to take advantage of free sample videos offered by Metis to evaluate its teaching method.

Unlike the classes offered through Dataquest and DataCamp, which consist of pre-recorded videos and texts, Metis offers live online classes with dedicated instructors who are ready to answer your questions during the live sessions and later during office hours. Further, while

Metis continued

Dataquest and DataCamp classes can be taken at any time by a Candidate, Metis's live classes are offered monthly throughout the year. Enrolled Candidates will be able to watch a video of the class should they miss a session. The approved class offered by Metis lasts six weeks.

The single course offered by Metis and approved by the FDP exam is titled Beginner Python and Math for Data Science, and it consists of the following six topics, which are usually covered over 12 live sessions.
https://www.thisismetis.com/courses/beginner-python-and-math-for-data-science/

1. **Python Basics**
   Candidates are introduced to programming in Python. Candidates will learn about Jupyter Notebooks – a popular platform for running Python programs. This part of the course will cover the basics of programming, including data structures, data operations, if-else statements, for and while loops, and logical operations.

2. **Python Advanced**
   This segment of the course covers advanced functionality in Python, including functions, debugging, error handling, string manipulations, and writing efficient code.

3. **Python Mathematical Libraries**
   Candidates will learn about using libraries that are useful for data manipulation and visualization. Candidates will learn to use NumPy, Pandas, and Matplotlib. These libraries will allow Candidates to load and save data, manipulate data such as aggregating, filtering, detecting outliers, and visualizing.

4. **Linear Algebra**
   This segment of the course is a refresher in linear algebra. It will cover the fundamentals of linear algebra, including vectors, and vector manipulations, matrices and matrix manipulations, linear equations and solutions, eigenvalues, and eigenvectors.

5. **Calculus and Probability**
   This module is a refresher in the fundamentals of calculus. It reintroduces students to such central concepts of calculus such as derivatives, integrals, determining local maximum and minimum, and limits. In addition, the module provides a refresher on central concepts of probability, such as random variables, mean, variance, probability mass and density functions, and cumulative distribution functions.

6. **Statistics**
   This final refresher module covers a few important statistical concepts such as ANOVA, hypothesis testing and p-value, and confidence intervals.

FDP EXAMINATION

The FDP examination, administered twice annually, is a four-hour computer-administered examination offered at test centers throughout the world. The FDP examination consists of 80 multiple choice questions weighted as 75% of the total points and three (3) to four (4) constructed response questions (multi-part essay type) weighted as 25% of the total points. The FDP exam will not contain any Python or R programming questions.

The FDP examination is based on this study guide, organized to facilitate quick learning and easy retention. Each topic is structured around learning objectives and keywords that define the content to be tested on the exam. The learning objectives and keywords are an important way for Candidates to organize their studies as they form the basis for examination questions. All learning objectives reflect the FDP curriculum content, and all examination questions are written to address the learning objectives directly.

For additional information about the FDP examination, please see the Candidate Handbook, which can be found on the FDP Institute website.

SAMPLE QUESTIONS

Sample questions are available on the FDP website to assist your study efforts. The set of questions contains more questions than the actual exam. In addition to helping you learn the topic material, the questions can also help you get familiar with the style and conventions used. An example is a simplifying convention of using the **natural logarithm to solve any problem requiring the calculation of a logarithm** on the exam. This convention is announced at the beginning of the sample exam, on the actual exam, and described in the Candidate Handbook.

OTHER STUDY TOOLS AND RESOURCES

In addition to this Study Guide and the Candidate Handbook, the FDP Institute website directs you to the readings covered in the curriculum. The readings are detailed below by topic area and include textbooks, often used across topics, and individual articles that are usually topic-specific. Both types of readings can be purchased from Amazon or the publisher, and whenever possible, they are posted on the FDP Institute website. They will be freely available to registered Candidates.

Page Number References for Keywords
For Candidates' convenience, a set of seven articles published by PMR Journals is provided in one collection titled 'Big Data and Machine Learning in the Financial Industry' and is available at a discounted price of $99 for registered Candidates. There are two sets of page numbers in this collection: one corresponds to the collection's table of contents, while the other one corresponds to each article's page number in the original journal. The page numbers appearing next to the keywords refer to the page numbers as they appeared in the original article. Note: *Check if your organization has a subscription to Portfolio Management Research (PMR) as this might give you free access to the 7 PMR readings.*

THE FDP CURRICULUM: OUTLINE

Candidates for the FDP Charter will have to enroll in the self-study program created by the FDP Institute and follow its carefully designed Study Guide. To become an FDP Charterholder, Candidates must pass the FDP exam and submit their learning certificates of the required online classes. The rest of this document discusses the FDP curriculum.
Below is the outline of the curriculum:

Topics	Approximate Weight %
1. Introduction to Data Science & Alternative Data	4-10
2. Machine Learning: Introduction to Algorithms	4-10
3. Machine Learning: Regression, Support Vector Machine & Time Series Models	4-10
4. Machine Learning: Regularization, Regression Trees, Random Forest & Overfitting	4-10
5. Machine Learning: Classification & Clustering	4-10
6. Machine Learning: Performance Evaluation, Backtesting & False Discoveries	4-10
7. Data Mining & Machine Learning: Naïve Bayes & Text Mining	4-10
8. Big Data & Machine Learning: Ethical & Privacy Issues	4-10
9. Big Data & Machine Learning in the Financial Industry	25-50

THE FDP CURRICULUM: READING LIST

Topic 1: Introduction to Data Science & Alternative Data

1.1   Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA:
        O'Reilly Media Inc. Chapters 1 & 2. Retrieved from
        https://www.oreilly.com/library/view/data-science-for/9781449374273/ and
        https://www.amazon.com/Data-Science-Business-Data-Analytic-Thinking-ebook/dp/
        B00E6EQ3X4
        Make sure that you review the errata page of the book:
        https://www.oreilly.com/catalog/errata.csp?isbn=0636920028918

1.2   Dannemiller, D. and R. Kataria. (2017). Alternative data for investment decisions: Today's
        innovation could be tomorrow's requirement by Deloitte Center for Financial Services.
        • *Link to the original site:* The FDP Institute encourages you to visit the original site,
          so the author's website receives traffic.

        • *Link to the PDF version:* A link to the PDF version is available for free to registered
          exam participants on the FDPI webpage Topics in Financial Data Science.
          The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential
          exam questions will be based on this version.

1.3   Gajjaria, A. (2018). Collective intelligence investing: Alpha generation via alternative
        data brings new risks by Deloitte Center for Financial Services.
        • *Link to the original site:* The FDP Institute encourages you to visit the original site,
          so the author's website receives traffic.

        • *Link to the PDF version:* A link to the PDF version is available for free to registered
          exam participants on the FDPI webpage Topics in Financial Data Science.
          The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential
          exam questions will be based on this version.

1.4   Quandl (2017). Sell Your Data to Wall Street: The comprehensive guide to monetizing
        your data assets for professional investors.
        • *Link to the original site:* The FDP Institute encourages you to visit the original site,
          so the author's website receives traffic.

        • *Link to the PDF version:* A link to the PDF version is available for free to registered
          exam participants on the FDPI webpage Topics in Financial Data Science.
          The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential
          exam questions will be based on this version.

1.5   Fayyad, U. and H. Hamutcu (2020). Toward Foundations for Data Science and Analytics:
        A Knowledge Framework for Professional Standards. Harvard Data Science Review, 2(2).
        • *Link to the original site:* The FDP Institute encourages you to visit the original site,
          so the author's website receives traffic.

        • *Link to the PDF version:* A link to the PDF version is available for free to registered
          exam participants on the FDPI webpage Topics in Financial Data Science.
          The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential
          exam questions will be based on this version.

Topic 2: Machine Learning: Introduction to Algorithms

2.1   James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). Corrected at 8th printing (2017)
       or 2nd edition (August 2021) *An Introduction to Statistical Learning: with applications in
       R*. New York, NY: Springer. Chapters 1, 2.1 & 2.2
       http://www-bcf.usc.edu/~gareth/ISL/
       https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/
       dp/1461471370.
       Make sure that you review the errata page of the book:
       https://www.statlearning.com/errata-first-edition
       https://www.statlearning.com/errata-second-edition

2.2   Nielsen, M. A. (2015). Using Neural Networks to Recognize Handwritten Digits.
       In Neural Networks and Deep Learning, Determination Press.
       • *Link to the original site:* The FDP Institute encourages you to visit the original site,
         so the author's website receives traffic.
       • *Link to the PDF version:* A link to the PDF version is available for free to registered
         exam participants on the FDPI webpage Topics in Financial Data Science.
         The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam
         questions will be based on this version.

2.3   Lee, D. (2019). Reinforcement Learning, Part 1: A Brief Introduction.
       • *Link to the original site:* The FDP Institute encourages you to visit the original site,
         so the author's website receives traffic.
       • *Link to the PDF version:* A link to the PDF version is available for free to registered
         exam participants on the FDPI webpage Topics in Financial Data Science.
         The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam
         questions will be based on this version.

2.4   Lee, D. (2019). Reinforcement Learning, Part 2: Introducing Markov Process.

       • *Link to the original site:* The FDP Institute encourages you to visit the original site,
         so the author's website receives traffic.

       • *Link to the PDF version:* A link to the PDF version is available for free to registered
         exam participants on the FDPI webpage Topics in Financial Data Science.
         The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam
         questions will be based on this version.

2.5   Lee, D. (2019). Reinforcement Learning, Part 3: The Markov Decision Process.

       • *Link to the original site:* The FDP Institute encourages you to visit the original site,
         so the author's website receives traffic.
       • *Link to the PDF version:* A link to the PDF version is available for free to registered
         exam participants on the FDPI webpage Topics in Financial Data Science.
         The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam
         questions will be based on this version.

Topic 3: Machine Learning: Regression, Support Vector Machine & Time Series Models

3.1   Provost, F. and T. Fawcett. (2013). *Data Science for Business.* Sebastopol, CA: O'Reilly Media Inc., Chapters 3 & 4. For the link to the errata page, see *Topic 1*

3.2   James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). Corrected at 8th printing (2017) or 2nd edition (August 2021) *An Introduction to Statistical Learning: with applications in R.* New York, NY: Springer. Chapter 3, Sections 1-3.
        For the link to the errata page, see *Topic 2*

3.3   Aas, K. and X. K. Dimakos. (2004). Statistical modeling of financial time series: An introduction. Oslo Norway: Norwegian Computing Center. Sections 1-4
        • *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

        • *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

Topic 4: Machine Learning: Regularization, Regression Trees, Random Forest & Overfitting

4.1   Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA: O'Reilly Media Inc., Chapter 5. For links to the book and errata pages, see *Topic 1*

4.2   James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). Corrected at 8th printing (2017) or 2nd edition (August 2021) *An Introduction to Statistical Learning: with applications in R*. New York, NY: Springer. Chapters 6 & 8, Sections 6.1, 6.2, 8.1 & 8.2.1-8.2.3 For the link to the errata page, see *Topic 2*

Topic 5: Machine Learning: Classification & Clustering

5.1   Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA: O'Reilly Media Inc., Chapter 6 & 7. For the link to the errata page, see *Topic 1*

Topic 6: Machine Learning: Performance Evaluation, Backtesting & False Discoveries

6.1   Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA: O'Reilly Media Inc., Chapter 8. For the link to the errata page, see *Topic 1*

6.2   Harvey, C. R. and Y. Liu. (2014). Evaluating Trading Strategies. *The Journal of Portfolio Management*, 40(5), 108-118
        • Available for purchase on the FDP webpage Big Data and Machine Learning in the Financial Industry.

6.3   Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, London, U.K.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

Topic 7: Data Mining & Machine Learning: Naïve Bayes & Text Mining

7.1   Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA: O'Reilly Media Inc., Chapters 9 & 10. For the link to the errata page, see *Topic 1*

7.2   Jurafsky, D. and J. Martin. (2018). Chapter 4. Naïve Bayes and Sentiment Classification, in *Speech and Language Processing*.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

Topic 8: Big Data & Machine Learning: Ethical & Privacy Issues

8.1   Institute of Business Ethics. (June, 2016). Business Ethics and Big Data. London, U.K.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

8.2   Institute of Business Ethics. (January, 2018). Business Ethics and Artificial Intelligence. London, U.K.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science.

8.3   Institute of Business Ethics. (2018, May). Beyond Law: Ethical Culture and GDPR. London, U.K.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

8.4   Loukides, M., M., H. Mason and DJ. Patil. *Ethics and Data Science*.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

Topic 9: Big Data & Machine Learning in the Financial Industry

9.1   Financial Stability Board. (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

9.2   AQR Portfolio Group Solutions. (2020). Can Machines "Learn" Finance? *Journal of Investment Management*, Vol 18, No. 2, pp. 23-36.

- *Link to the original site:* The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- *Link to the PDF version:* A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science. The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

9.3   Cerniglia, J. A. and F. J. Fabozzi. (2020). Selecting Computational Models for Asset Management: Financial Econometrics versus Machine Learning – Is There a Conflict? *The Journal of Portfolio Management*, 47(3), 107-118.

- Available for purchase on the FDP webpage Big Data and Machine Learning in the Financial Industry

9.4   Halpin, L. and D. Dannemiller. (2019). Artificial intelligence: The next frontier for investment management firms. Deloitte's Global Investment Management Group.

- ***Link to the original site:*** The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- ***Link to the PDF version:*** A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science.
The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

9.5   Monk, A., M. Prins, and D. Rook. (2019). Rethinking Alternative Data in Institutional Investment. *Journal of Financial Data Science*, 1(1), 14-31.

- Available for purchase on the FDP webpage Big Data and Machine Learning in the Financial Industry.

9.6   Zappa, D., M. Borrelli, G.P. Clemente, N. Savelli. (2019). Text Mining in Insurance: From Unstructured Data to Meaning.

- ***Link to the original site:*** The FDP Institute encourages you to visit the original site, so the author's website receives traffic.

- ***Link to the PDF version:*** A link to the PDF version is available for free to registered exam participants on the FDPI webpage Topics in Financial Data Science.
The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam questions will be based on this version.

9.7   Cowden, C., F.J. Fabozzi and A. Nazemi. (2019). Default Prediction of Commercial Real Estate Properties Using Machine Learning Techniques.
The Journal of Portfolio Management, 45(7), 55-67.

- Available for purchase on the FDP webpage Big Data and Machine Learning in the Financial Industry.

9.8   Sokolov, A, J. Mostovoy, J. Ding and L. Seco. (2021). Building Machine Learning Systems for Automated ESG Scoring. *The Journal of Impact and ESG Investing*, 1(3), 39-50.

- Available for purchase on the FDP webpage Big Data and Machine Learning in the Financial Industry.

9.9   Aldridge, I. and M. Avellaneda. (2019). Neural Networks in Finance: Design and Performance. *The Journal of Financial Data Science*, 1(4), 39-62.

- Available for purchase on the FDP webpage Big Data and Machine Learning in the Financial Industry.

9.10   Rasekhschaffe, K. and R. Jones. (2019). Machine Learning for Stock Selection.
        Financial Analysts Journal, Volume 75, Issue 3.

  • ***Link to the original site:*** The FDP Institute encourages you to visit <u>the original site</u>
     or <u>this site</u>, so the author's website receives traffic.

  • ***Link to the PDF version:*** A link to the PDF version is available for free to registered
     exam participants on the FDPI webpage <u>Topics in Financial Data Science</u>.
     The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam
     questions will be based on this version.

9.11   Gu, S., B. Kelly, and D. Xiu. (2018). Empirical Asset Pricing via Machine Learning.

  • ***Link to the original site:*** The FDP Institute encourages you to visit <u>the original site</u>,
     so the author's website receives traffic.

  • ***Link to the PDF version:*** A link to the PDF version is available for free to registered
     exam participants on the FDPI webpage <u>Topics in Financial Data Science</u>.
     The PDF was accessed on April 5, 2021. The FDP Study Guide and any potential exam
     questions will be based on this version.

9.12   López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail.
        The Journal of Portfolio Management, 44 (6) 120-133.

  • Available for purchase on the FDP webpage <u>Big Data and Machine Learning in the
     Financial Industry</u>.

LEARNING OBJECTIVES

Topic 1: Introduction to Data Science & Alternative Data

Reading 1.1 Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA: O'Reilly Media Inc. Chapters 1 & 2.

Keywords

Data mining (p. 2)	*Data-analytic thinking (p. 12)*
Data science (p. 4)	*Target (p. 24)*
Churn (p. 4)	*Label (p. 24)*
Data-driven decision making (p. 5)	*Unsupervised data mining (p. 24)*
Data engineering (p. 5, 7)	*Supervised data mining (p. 25)*

Learning Objectives
Demonstrate proficiency in the following areas:

1.1.1   Data analytic thinking (Ch. 1)
   For example:

   A. Discuss the ubiquity of data opportunities.

   B. Compare and contrast data science, engineering, and data-driven decision-making.

   C. Explain data and data science capability as a strategic asset.

   D. Describe data-analytic thinking.

   E. Compare data science and the work of the data scientist.

1.1.2   Business problems and data science solutions (Ch. 2)
   For example:

   A. Describe how one transitions from business problems to data mining tasks.

   B. Compare supervised methods to unsupervised methods.

   C. Describe the difference between data mining and using the results of data mining.

   D. Describe key aspects of the data mining process, including business understanding, data understanding, data preparation, modeling, and evaluation.

Reading 1.2 Dannemiller, D. and R. Kataria. (2017). Alternative data for investment decisions: Today's innovation could be tomorrow's requirement by Deloitte Center for Financial Services.

Keywords

Diffusion of innovations curve (p. 2)	*Late majority (p.3)*
Innovators (p.3)	*Laggards (p.3)*
Early adopters (p.3)	*Big data analytics (BDA) (p. 10)*
Early majority (p.3)	*Cognitive technologies (p. 12)*

Learning Objectives

Demonstrate proficiency in the areas of:

1.2.1   Adoption of alternative data
For example:

A. Discuss investment managers in terms of their places on the diffusion of innovations curve.

1.2.2   The risks and rewards of alternative data for investment decisions
For example:

A. Identify components of data risk.

B. Identify four potential risks that may relate to incorporating alternative data in investment selection processes.

C. Identify three risks that late adopters of alternative data in the investment management process face.

1.2.3   Advanced technologies for alpha
For example:

A. Identify two major types of technologies that investment management firms have started to use to realize alternative data's potential.

B. Describe the big data analytics (BDA) groups that are most aligned to support alpha generation.

C. Describe the categories of cognitive technologies that are most aligned to support alpha generation.

1.2.4   Alternative data vendor profiles
For example:

A. Identify the focus of the alternative data vendor profiled.

B. Identify the four fundamental questions the alternative data vendor profiled asks to control the quality of alternative data.

1.2.5   The future of alternative data
For example:

A. Identify the segment of alternative data adoption by investment managers (innovators, early adopters, early majority, late majority, laggards) expected within the next five years.

Reading 1.3 Gajjaria, A. (2018). Collective intelligence investing: Alpha generation via alternative data brings new risks by Deloitte Center for Financial Services

Keywords

Collective Intelligence Investing (p. 1)	*Community engagement risk (p. 5)*
Open communities (p. 3)	*Material nonpublic information risk (p. 5)*
Crowdsourcing platforms (p. 3)	*Model risk (p. 5)*
Digital expert contribution networks (p. 3)	*Information security risk (p. 5)*
Digital expert communication networks (p. 3)	*Data integrity risk (p. 6)*

Learning Objectives

Demonstrate proficiency in the areas of:

1.3.1   Warming up to collective intelligence investing (CII)
For example:

A. Describe the term collective intelligence investing.

1.3.2   Getting started with alternative data and CII
For example:

A. Identify considerations and corresponding actions for traditional investment managers adopting alternative data and collective intelligence investing.

1.3.3   Balancing the risks and rewards of CII from different platform types
For example:

A. Describe four key types of collective intelligence communities: open communities, digital expert contribution networks, digital expert communication networks, and crowdsourcing platforms.

B. Identify investment-related information that can be found in a specified type of CII community.

C. Describe five key risks associated with CII.

D. Describe a CII community in terms of the level of information structure, membership diversity, and risk exposure to each of the five key risks associated with CII.

1.3.4   The road ahead for investment managers
For example:

A. Describe three steps investment manager firms could adopt before jumping wholeheartedly into CII for a potentially smoother takeoff.

Reading 1.4 Quandl (2017). Sell Your Data to Wall Street: The comprehensive guide to monetizing your data assets for professional investors

> **Keywords**
> *Alpha (p. 2)*   *Leading indicator (p. 7)*
> *Buy side (p. 4)*   *Table stakes (p. 9)*
> *Quantitative hedge funds (p. 4)*   *Data hunters (p. 12)*
> *Contemporaneous correlation (p. 7)*   *Personally identifiable information (PII) (p. 14)*

Learning Objectives

Demonstrate proficiency in the areas of:

1.4.1   Understanding why "Wall Street" wants data
For example:

    A. Define the buy-side.

    B. Identify the best customer for alternative data on the buy-side.

    C. Describe what alternative data provides the best customer on the buy-side.

1.4.2   Building a data product
For example:

    A. Describe four stages of the production process.

    B. Assess data quality and describe the criteria used.

    C. Evaluate the predictive power of a data set and describe the criteria used.

    D. Determine the commercial value of a data set and describe factors considered.

    E. Describe factors that go into packaging data for consumption.

1.4.3. Delivering your data
For example:

    A. Describe data delivery capabilities and features required by Wall Street customers.

1.4.4   Marketing and selling your data
For example:

    A. Identify sales and marketing strategies.

    B. Describe good data collateral.

    C. Describe the sales process.

1.4.5   Protecting yourself and your data
For example:

    A. Describe the concept of personally identifiable information (PII) and Wall Street's view of it.

    B. Describe the ability to sell data anonymously.

1.4.6 Pricing your data
For example:

A. Describe what indicators and data attributes are related to the value of the data.

B. Calculate the value of a data set given the assets under management (AUM) of a hedge fund and the expected excess returns the dataset is expected to generate.

C. Describe strategies to balance the trade-offs involved in selling high-priced data to few clients versus more affordable data to many clients.

1.4.7 Case studies
For example:

A. Illustrate how alternative data can help managers produce alpha when trading auto stocks.

B. Illustrate how alternative data can help produce alpha when trading commodities.

C. Illustrate how alternative data can help produce alpha when trading stocks.

Reading 1.5 Fayyad, U. and H. Hamutcu. (2020). Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards.
***Harvard Data Science Review*, 2(2).**

Keywords	
Initiative for Analytics and Data Science Standards (IADSS) (p. 2)	*Docker (p. 21)*
	NumPy (p. 21)
The IADSS Analytics and Data Science Knowledge Framework (p. 18)	*SciPy (p. 21)*
	Octave (p. 21)
Git (p. 21)	*MATLAB (p. 21)*
AWS (p. 21)	*TensorFlow (p. 21)*
Azure (p. 21)	*MatplotLib (p. 21)*

Learning Objectives
Demonstrate proficiency in the following areas:

1.5.1 Explain the motivation for the first article by IADSS

1.5.2 Definitions of data science
For example:

A. Compare and contrast definitions given by:

- Gray (2007)

- Davenport and Patil (2012)

- Hammerbacher (2009)

- Provost and Fawcett (2013)

- Blei and Smyth (2017)

- Donoho (2017)

1.5.3 The IADSS Analytics and Data Science Knowledge Framework
For example:

A. Describe the two primary knowledge domains of the IADSS Analytics and Data Science Knowledge Framework.

B. List subfields within the two primary knowledge domains of the IADSS Analytics and Data Science Knowledge Framework.

C. Identify examples of skills within general purpose computing.

D. Describe the framework's treatment of domain expertise.

ACTION WORDS

In each of the above learning objectives, action words are used to direct your study focus. Below is a list of all action words used in this study guide, along with definitions and two examples of usage, in a question example and a description. Should you not understand what is required for any learning objective, we suggest you refer to the table below for clarification.
NOTE: The question examples in this table are NOT sample questions for the current exam.

Term	Definition	Question Example	Example of Term Use
Analyze	Study the interrelations	George has identified an opportunity for a convertible arbitrage reverse hedge. What risks are associated with this hedge? A. The convertible may remain overvalued, causing the positive cash flow to harm the position's return profile. B. The short convertible may be called in and the position must be delivered, forcing the hedge to be unwound at an inopportune time. C. The implied volatility may decrease, lowering the bond's value.	You have to **analyze** the positions and factors impacting them. *Correct Answer: B*
Apply	Make use of **Note:** If you are asked to apply a model to data, you will be expected to have the appropriate equation memorized, unless the question also contains the action word "recognize"..	Alicia Weeks, CFA, Real Estate Investment Advisor, works in an Asian country where there are no securities laws or regulations. According to CFA Institute Standard I, Fundamental Responsibilities, Alicia: A. Must adhere to the standards as defined in a neighboring country that has the strictest laws and regulations. B. Need not concern herself with ethics codes and standards. C. Must adhere to the CFA Institute's codes and standards.	You have to **apply** CFA Institute Standard I to find the correct answer. *Correct Answer: C*
Argue	Prove by reason or by presenting the associated pros and cons; debate	Why did the shape of the supply curve for venture capital funds change after 1979?	You have to describe how the curve has changed AND **argue** why it changed by providing reasons and supporting the reasons with statements of facts (e.g., change in regulations).
Assess	Determine importance, size, or value	How are lower capital gains taxes expected to impact firm commitments? A. Through increased supply of capital, firm commitments are expected to rise. B. Through decreased supply of capital, firm commitments are expected to rise. C. Through decreased after-tax return on venture investments, firm commitments are expected to rise.	You must **assess** the significance of the change in the tax rate for firm commitments. *Correct Answer: A*

Term	Definition	Question Example	Example of Term Use
Calculate	Determine a value mathematically **Note:** You will be expected to have the appropriate equation memorized, unless the question also contains the action word "recognize".	Consider a set of 100 people. Eighty percent have feature A and twenty percent do not have feature A. What is the entropy for this set? A. 0.72 B. 0.88 C. 0.93	You have to **calculate** entropy based on the given probabilities. *Correct Answer: A*
Compare	Describe similarities and differences	Which of the following least accurately compares the Sharpe and Treynor ratios? A. Both ratios contain excess return in the numerator. B. Both ratios express a measure of return per unit of some measure of risk. C. The Sharpe ratio is the inverse of the Treynor ratio	You must **compare** the ratios based on their most important similarities and their most important differences. *Correct Answer: C*
Compare and Contrast	Examine in order to note similarities or differences	A comparison of monthly payments and loan balances of a constant payment mortgage with a constant amortization mortgage with the same loan terms will show that: A. The initial payment will be the same. B. The payments of the constant payment mortgage are initially greater than those of the constant amortization mortgage, but at some point the payments of the constant payment mortgage become less. C. The present value of the payment streams of the two loan types are the same.	You must **compare** indices to arrive at the answer. *Correct Answer: C*
Construct	Make or form by combining or arranging parts or elements	A reverse convertible arbitrage hedge consists of a: A. Short convertible position plus a long position in the stock. B. Short convertible position plus a put option on the stock. C. Long convertible position plus a put option on the stock.	You must **combine** positions to construct the hedge. *Correct Answer: A*
Contrast	Expound on the differences	Which of the following best characterizes a difference between value at risk (VaR) and modified VaR? A. Modified VaR is expressed as a percent while VaR is a dollar value. B. Modified VaR uses a user defined confidence interval while VaR uses a 99% interval. C. Modified VaR incorporates non-normality while traditional VaR assumes normality.	You have to **contrast** the assumptions of the first model to those of the second model so that the differences are clear. *Correct Answer: C*

Term	Definition	Question Example	Example of Term Use
Define	State the precise meaning	The interest rate charged by banks with excess reserves at a Federal Reserve Bank to banks needing overnight loans to meet reserve requirements is called the: A. Prime rate. B. Discount rate. C. Federal funds rate.	You must **define**, in this case, the federal funds rate. ***Correct Answer: C***
Describe	Convey or characterize an idea	Which of the following words best describes expected return? A. Spread B. Average C. Spread squared	You need to choose the word that best **describes** the concept from a list. **Correct Answer: B**
Determine	Ascertain or establish exactly, typically as a result of research or calculation	Which of the following algorithms is a type of supervised learning? A. Decision tree B. K-means clustering C. Gaussian mixture model	You must **determine** which algorithm represents supervised learning. ***Correct Answer: A***
Differentiate	Constitute the distinction between; distinguish	What type of convertible hedge entails shorting a convertible and going long in the underlying stock? A. Reverse hedge B. Call-option hedge C. Traditional convergence hedge	You must **differentiate** one type of hedge from another. ***Correct Answer: A***
Discriminate	See "Differentiate"		
Discuss	Examine or consider a subject	Discuss the limitations of private equity data.	You must present a **discussion** of a set of ideas in a list or paragraph.
Evaluate	See "Assess"		
Explain	Illustrate the meaning	1. Explain why return on assets (ROA) rather than return on equity (ROE) might be the preferred measure of performance in the case of hedge funds. or 2. Which of the following best explains risk from the standpoint of investment? A. Investors will lose money. B. Terminal wealth will be less than initial wealth. C. More than one outcome is possible.	1. You must place a series of thoughts together as an **explanation** of a term or issue. 2. You need to identify the term that best **explains** a term or issue. ***Correct Answer: C***

Term	Definition	Question Example	Example of Term Use
Identify	Establish the identity	The investments that have historically performed best during periods of recession are: A. Commodities. B. Treasury bills. C. Stocks and bonds.	You must **identify** the term that best meets the criterion of the question *Correct Answer: C*
Illustrate	Clarify through examples or comparisons	For two types of convergence hedges, what situations present profitable opportunities, how are the hedges set up, and what are the associated risks?	You must provide an example for each hedge or compare the two to **illustrate** how they work.
Interpret	Explain the meaning	Your certificate of deposit will mature in one week, and you are considering how to invest the proceeds. If you invest in a 30-day CD, the bank will pay you 4% interest. If you invest in a 2-year CD, the bank will pay you 6% interest. You should choose the: A. 2-year CD if you expect that interest rates will fall in the future B. 30-day CD, no matter what you expect interest rates to do in the future. C. 2-year CD, no matter what you expect interest rates to do in the future.	You must **interpret** the features of an investment scenario. *Correct Answer: A*
List	Create a series of items	List the determinants of real interest rates.	You must differentiate from a **list** those items that are consistent with the question.
Outline	Summarize tersely	Which of the following best characterizes the steps in computing a geometric mean return based on a series of periodic returns from *T* time periods? A. Add one to each return, add them together, divide by T and subtract one. B. Add one to each return, multiply them together, take the Tth root and subtract one. C. Add one to each return, multiply them together, divide by T and subtract one.	You must **outline** the study's most important findings rather than explain them in detail. *Correct Answer: B*

Term	Definition	Question Example	Example of Term Use
Recognize	Recall the purpose of a given equation or term, and its name when appropriate. **Note:** When the action word "recognize" is used and applied to an equation, the equation will be provided within the question stem or the correct answer choice.	What is the following equation called and used for in the context of artificial neural networks? $$\sigma(z) \equiv \frac{1}{1 + e^{-z}}$$ A. It is called a neuron and used to make a NAND gate. B. It is called a sigmoid function and is used to model sigmoid neurons that better enable learning than perceptrons do. C. It is called a perceptron which is used to create a smoother function than a logistic function.	You must **recognize** that this is a sigmoid function, also referred to as a logistic function, which is used as the basis for a sigmoid neuron. *Correct Answer: B.*
Relate	Show or establish logical or causal connection	Which of the following effects does NOT help to explain growth in the venture capital industry? A. Amendments to the prudent man rule B. The rise of limited partnerships as an organizational form C. Decline in the valuations of small capitalization stocks	You must **relate** effects or factors (e.g., the prudent man rule) to another result or concept (e.g.) growth in an industry.) *Correct Answer: C*

FDP EDITORIAL STAFF

Hossein Kazemi, PhD, CFA, Senior Advisor, The CAIA Association
Keith Black, Ph.D., CFA, CAIA, FDP, Managing Director, Program Director, The FDP Institute
Mirjam Dekker, Program Manager, The FDP Institute
Satya Das, Curriculum & Exams Advisor, The FDP Institute
Aaron Filbeck, CFA, CAIA, CIPM, FDP, Director, Global Content Development, The CAIA Association

Design: Gabriele Villamena, Sichtwerk, Inc.