Financial Data Professional Institute

FDP CANDIDATE ORIENTATION
March 16 – April 4, 2020 FDP Exam

Hossein Kazemi, Senior Advisor, CAIA Association
Keith Black, Managing Director Curriculum & Exams, CAIA Association & recent FDP Charterholders
Mirjam Dekker, Project Manager, FDP Institute

www.fdpinstitute.org
January 15, 2020

Agenda

- Welcome
- Introductions – todays presenters

Mehrzad Mahdavi,
Executive Director,
FDPI

Hossein Kazemi
Sr. Advisor FDPI
Sr. Advisor CAIA

Keith Black
Managing Director
Content Strategy, CAIA
FDP Charter holder

Mirjam Dekker
Project Manager
FDPI

- **Part I - Exam Topics**
  - Curriculum
  - Examination Format
  - Topic 1 – 8 (Required Readings, Sample Keywords, Sample learning Objectives, Sample Question & Answer Key)
- **Part II - Candidate Experience**

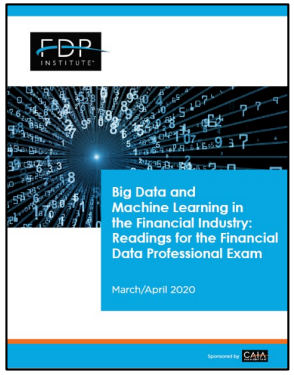
→ Goal: provide you the tools & knowledge that will lead to **a successful exam day**

THE FINANCIAL DATA PROFESSIONAL PROGRAM

MARCH 16 – APRIL 4, 2020

EXAM TOPICS

FDP Curriculum



PMR E-Book

8 required readings

$99

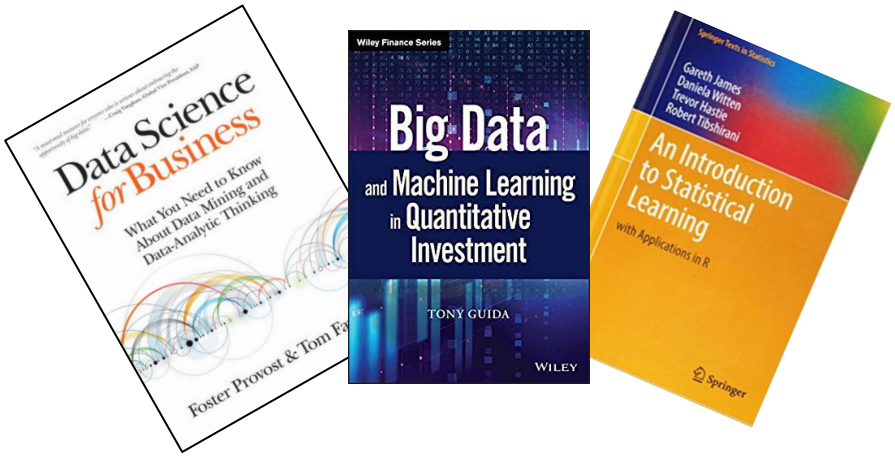
Exam Registrants

Topics in Financial Data Science

12 required readings

Free

Exam Registrants



3 required books

- Data Science for Business
- Big Data and Machine Learning in Quantitative Investment
- An Introduction to Statistical Learning

Exam Registrants can look for best available price.

Examination Format

- 75 multiple choice questions (70%)
- 3-4 essay questions (30%)

Session 1: 70 MC questions

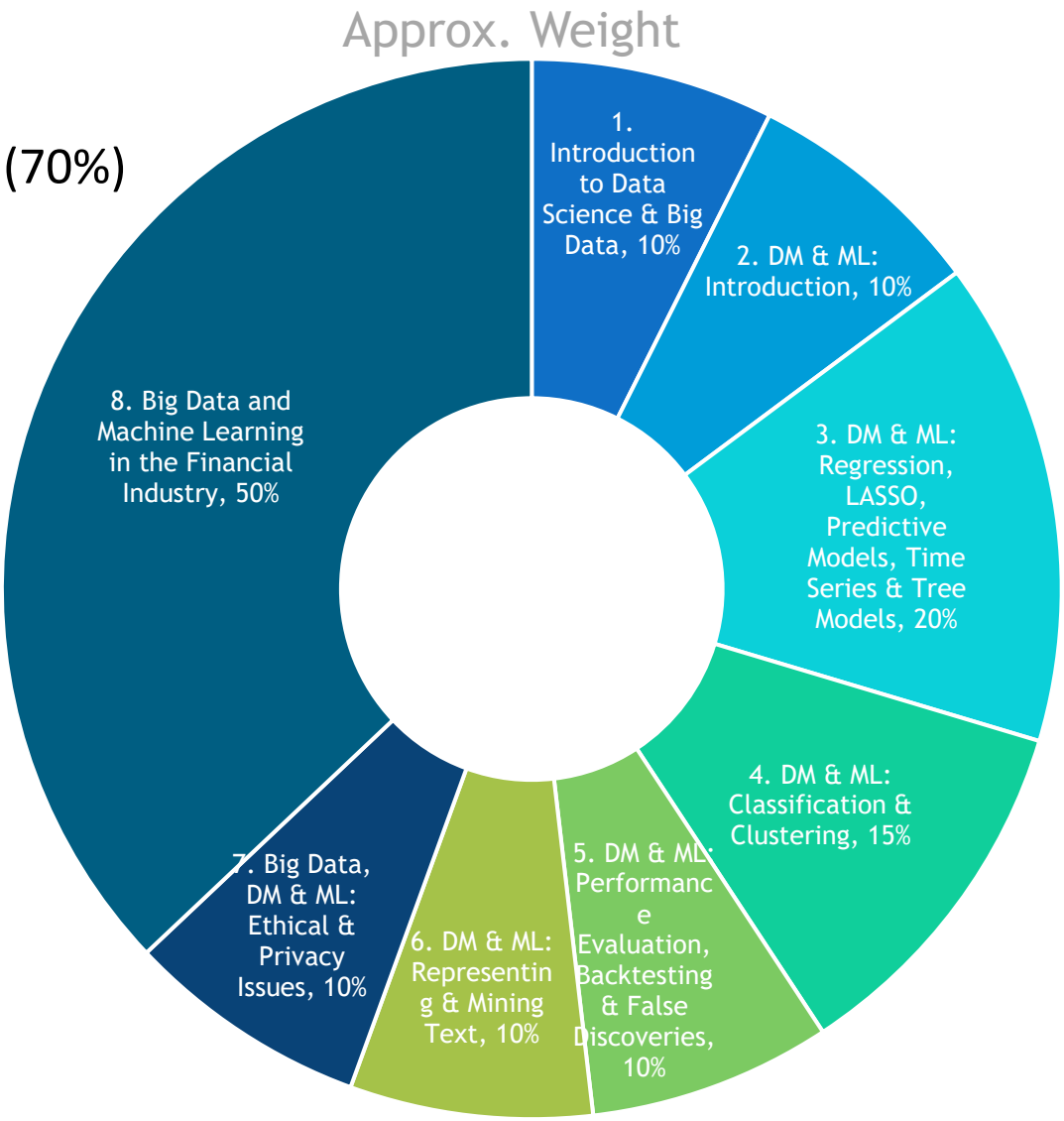
100 minutes

Optional break

Session 2: 5 MC questions

3-4 Essay questions

100 minutes



FDP Curriculum

Topics, Learning Objectives & Keywords

TOPIC 1
Introduction to Data Science & Big Data

- **READING 1** Provost, F. & T. Fawcett. (2019) Data Science for Business Chapter 1 & 2
  - **KEYWORDS**
    - **Keywords**
      - *Data mining (p. 2)*
      - *Churn (p. 4)*
      - *Data engineering (p. 5, 7)*
      - *Target (p. 24)*
      - *Unsupervised data mining (p. 24)*
      - *Data science (p. 4)*
      - *Data-driven decision making (p. 5)*
      - *Data-analytic thinking (p. 12)*
      - *Label (p.24)*
      - *Supervised data mining (p. 25)*
  - **LEARNING OBJECTIVE 1**
    - **Learning Objectives**
      Demonstrate proficiency in the following areas:
      1) Data analytic thinking (Ch. 1)
      For example:
      - o Discuss the ubiquity of data opportunities.
      - o Define data science, engineering, and data-driven decision making.
      - o Explain data and data science capability as a strategic asset.
      - o Describe data-analytic thinking.
      - o Compare data science and the work of the data scientist.
  - **LEARNING OBJECTIVE 2**
    - 2) Business problems and data science solutions (Ch. 2)
      For example:
      - o Describe how one transitions from business problems to data mining tasks.
      - o Compare supervised methods to unsupervised methods.
      - o Describe the difference between data mining and using the results of data mining.
      - o Describe key aspects of the data mining process, including business understanding, data understanding, data preparation, modeling, and evaluation.
- **READING 2**
- **READING 3**
- **READING 4**

FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Reading(s):

1.1 Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA: O'Reilly Media Inc. Chapters 1& 2. Retrieved from http://data-science-for-biz.com/ https://www.amazon.com/Data-Science-Business-Data-Analytic-Thinking-ebook/dp/B00E6EQ3X4

1.2 Getman, G. and Q. Hasan. (2019). *Big Data is a Big Deal: An investor's guide to the applications and challenges of alternative data*, New York, NY: Lombard Odier, Retrieved from https://www.lombardodier.com/contents/news/world-in-transition/2019/april/big-data-is-a-big-deal.html

1.3 Kaul, S. (2016). Chapter 8. Big Data and Investment Management. *In Citi GPS: Global Perspectives & Solutions*, (pages 58-65). Retrieved from https://www.citivelocity.com/citigps/investment-themes-2016/

1.4 Guida, T. (2019). *Big Data and Machine Learning in Quantitative Investments*. West Sussex, UK: John Wiley & Sons Ltd. Chapters 2, 4 & 5. https://www.amazon.com/Machine-Learning-Quantitative-Investment-Finance/dp/1119522196

Sample Keywords:

Keywords

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

FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Sample Learning Objective:

1.2.2 The benefits and limitations of big data for investment decisions

For example:

A. Explain how big data provides a new source of fundamental insights, relative to discretionary and quantitative strategies, by using the four data quadrants of any investment framework (financial data, alternative data, market data and internal data).

B. Describe a process of harnessing data-driven insights including:
- Sourcing
- Backtesting
- Tagging
- Mapping, and
- Visualization

Sample Question:

In the article "Big Data is a Big Deal," there are three main types of manager orientations (legacy fundamental, legacy quantitative, and big data) which use various combinations of four types of data sources. Examples are given for each of the four categories. Which type of manager orientation is most likely to rely on implied volatility data and factor analysis data?

- ○ Fundamental
- ○ Quantitative*
- ○ Big data

Answer key → *LO 1.2.2. p.9*

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Reading(s):

2.1 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). *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

2.2 Nielsen, M. A. (2015). Using Neural Networks to Recognize Handwritten Digits. In *Neural Networks and Deep Learning*, Determination Press. Retrieved from
http://neuralnetworksanddeeplearning.com/chap1.html

Sample Keywords:

Keywords

Statistical learning (p. 1)
Classification problems (p.28)
Semi-supervised learning (p. 28)
Quantitative variables (p. 28)
Qualitative response (p. 28)
Binary response (p.28)
Regression (p. 28)
Predictors (p. 29)
Mean squared error (MSE) (p. 29)
Training MSE (p. 30)
Test data (p. 30)
Test MSE (p. 30)
Flexibility (p. 31)
Degrees of freedom (p. 32)
Cross validation (p. 33)
Expected test MSE (p. 34)
Bias (p. 35)
Bias-variance trade-off (p. 36)
Error rate (p.37)
Indicator variable (p. 37)
Training error (p. 37)
Test error (p.37)
Bayes classifier (p. 37)
Conditional probability (p. 37)
Bayes decision boundary (p. 38)
Bayes error rate (p. 38)
K-nearest neighbors (p. 39)

FDP Curriculum

Sample Learning Objective:

2.2.6 Learning with gradient descent

For example:

A. Recognize a quadratic cost function of weights and biases and alternative terminology for the cost function.

B. Explain why minimizing a quadratic cost function is preferable to working with other types of cost functions.

C. Recognize an equation for an update rule that defines the gradient descent algorithm and explain the purpose of each component in the equation.

D. Explain how quickly stochastic gradient descent can speed up learning given a training set size n and a mini-batch size, m.

Sample Question:

What is the best description of the gradient descent algorithm as it applies to neural networks?

- It iteratively and incrementally changes biases of a neuron.
- It iteratively and incrementally changes weights of a neuron.
- It iteratively and incrementally changes weights and biases of a neuron.*

Answer key → *LO 2.2.6. p. 26*

FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Reading(s):

3.1 Provost, F. and T. Fawcett. (2013). *Data Science for Business*. Sebastopol, CA: O'Reilly Media Inc., Chapters 3, 4, 5, & 9.

3.2 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). *An Introduction to Statistical learning: with applications in R*. New York, NY: Springer. Chapters 3 & 6.

3.3 Aas, K. and X. K. Dimakos. (2004). *Statistical modeling of financial time series: An introduction*. Oslo Norway: Norwegian Computing Center. Retrieved from https://www.nr.no/files/samba/bff/SAMBA0804.pdf

Sample Keywords:

Keywords

Information (p. 43)
Tree induction (p. 44)
Predictive model (p. 45)
Descriptive modeling (p. 46)
Target variable (p. 46)
Attributes or features (p. 46)
Model induction (p. 47)
Deduction (p. 47)
Training data (p. 47)
Labeled data (p. 47)
Supervised segmentation (p. 48)
Information gain (p. 51)
Entropy (p. 51)
Parent set (p. 52)
Child set (p. 52)
Variance (p. 56)
Entropy graph/chart (p. 58)
Classification tree (p. 63)
Decision nodes (p. 63)
Laplace correction (p. 73)
Linear classifier (p. 85)
Linear discriminant (p. 86)
Margin (p. 92)
Support vector machine (p. 92)
Hinge-loss (p. 94)
Zero-one loss (p. 95)
Squared error (p. 95)
Odds ratio (p. 98)
Log-odds (p. 99)
Logistic function (p. 101)
Generalization (p. 112)
Overfitting (p. 113)
Fitting graph (p. 113)
Holdout data (p. 113)
Base rate (p. 115)
Sweet spot (p. 117)
Cross-validation (p. 126)
Folds (p. 127)

FDP Curriculum

Sample Learning Objective:

3.2.4 Subset selection

For example:

A. Describe the best subset selection.

B. List the steps used in the best subset selection.

C. Define deviance.

D. Describe the forward stepwise selection and the backward stepwise selection.

E. List the steps used in forward stepwise selection and backward stepwise selection.

F. Recognize and apply the equations for C^p, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R^2.

G. Describe the model selection criterion using C^p, AIC, BIC, and adjusted R^2.

Sample Questions:

Which of the following approaches to model selection provides a direct estimate of the test error and makes the least number of assumptions about the underlying model?

- C_p
- Cross validation*
- Akaike information criterion (AIC)

Answer key → LO: 3.2.4. p.213

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Reading(s):

4.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 6.

4.2 James, G., D. Witten, T. Hastie and R. Tibshirani. (2013). An Introduction to Statistical Learning: with applications in R. New York, NY: Springer. Chapters 4 & 10.

Sample Keywords:

Keywords

Euclidean distance (p. 144)
Nearest neighbors (p. 144)
Combining function (p. 147)
Weighted voting (p. 150)
Curse of dimensionality (p.156)
Manhattan distance (p. 159)
Jaccard distance (p.159)
Cosine distance (p. 160)
Edit distance or Levenshtein metric (p. 161)
Clustering (p. 164)
Hierarchical clustering (p. 165)
Dendrogram (p. 165)
Linkage function (p. 167)
Cluster center or centroid (p. 170)
k-means clustering (p. 170)
Distortion (p. 173)

Keywords

Logistic function (p. 132)
Odds (p. 132)
Log odds (p. 132)
Likelihood function (p. 133)
Principal component analysis (p. 375)
Loadings (p. 375)
Bottom-up agglomerative clustering (p. 390)
Linkage (p. 394)
Inversion (p. 395)

FDP Curriculum

Sample Learning Objective:

4.1.1 Similarity and distance

For example:

A. Calculate the Euclidean distance.

B. Define nearest neighbors and combining function.

C. Explain how combining function can be used for classification.

D. Define weighted voting or similarity moderated voting.

E. Calculate contributions for weighted voting for classification.

F. Explain how k in k-NN can be used to address overfitting.

G. Discuss issues with nearest-neighbor methods with focus on

i. Intelligibility,

ii. Dimensionality and domain knowledge, and

iii. Computational efficiency.

Sample Question:

In the k nearest neighbors (NN) classification system, let n represents the sample size. For what value of k is a k-NN classifier going to make each prediction be the average value of the data set?

- 0
- 1
- n*

Answer Key → LO: 4.1. 1. p. 153

FDP Curriculum

1. Introduction to Data Science & Big Data

2. DM & ML: Introduction

3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models

4. DM & ML: Classification & Clustering

5. DM & ML: Performance Evaluation, Backtesting & False Discoveries

6. DM & ML: Representing & Mining Text

7. Big Data, DM & ML: Ethical & Privacy Issues

8. Big Data and Machine Learning in the Financial Industry

Reading(s):

5.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapters 7 & 8.

5.2 Arnott, R., C. B. Harvey, and H. Markowitz. (2019). A Backtesting Protocol in the Era of Machine Learning. Journal of Financial Data Science, 1(1), 64-74. DOI: https://doi.org/10.3905/jfds.2019.1.06

5.3 Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, London, U.K.: Royal Society Open Science, Retrieved from https://royalsocietypublishing.org/doi/full/10.1098/rsos.140216

5.4 López de Prado, M. (2019). A Data Science Solution to the Multiple-Testing Crisis in Financial Research. Journal of Financial Data Science, 1(1), 99-110. DOI: https://doi.org/10.3905/jfds.2019.1.099

Sample Keywords:

Keywords

Accuracy (p. 189)	*F-measure (p. 204)*
Confusion matrix (p. 189)	*Profit curve (p. 212)*
False positive (p. 190)	*Base rate (p.214)*
False negative (p. 190)	*ROC graph (p. 215)*
Expected value (p. 194)	*Hit rate (p. 216)*
Class prior (p. 201)	*False alarm rate (p.216)*
Precision (p. 204)	*AUC (p. 219)*
Recall (p. 204)	*Cumulative response curve (p. 219)*

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Sample Learning Objective:

5.3.1 An investigation of the false discovery rate and the misinterpretation of p-values

For example:

A. Define specificity and sensitivity.

B. Describe the false discovery rate with the help of a tree diagram.

C. Calculate the probability of real effect given a result is significant.

D. Define power of a test.

E. Describe the false discovery rate in simulated t-tests.

F. Calculate false discovery rate.

G. Describe underpowered study.

H. Describe the inflation effect in the context of false discovery.

I. Describe what happens when we consider p=0.05 rather than p<=0.05.

J. Describe Berger's approach.

Sample Question:

The performance of 1000 different hedge fund managers is tested. In reality, 10% of the managers actually have abnormal performance. The power of the test is 80% and a p-value of 5% is used. How many managers does the test indicate had abnormal performance?

- 100
- 125*
- 200

Answer Key → LO 5.3.1. p. 10

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Reading(s):

6.1 Provost, F. and T. Fawcett. (2013). Data Science for Business. Sebastopol, CA: O'Reilly Media Inc., Chapter 10.

6.2 Jurafsky, D. and J. Martihttps. (2018). Chapter 4. Naïve Bayes and Sentiment Classification, In Speech and Language Processing. Retrieved from https://web.stanford.edu/~jurafsky/slp3/4.pdf

Sample Keywords:

Keywords

Linguistic structure (p. 252)
Dirty data (p. 252)
Document (p. 253)
Token (p. 253)
Terms (p. 253)
Corpus (p. 253)
Bag of words (p. 254)
Term frequency (p. 254)
Stemmed (p. 255)
Stopwords (p. 255)
Inverse document frequency (p. 256)
n-grams (p. 265)
Latent information model (p. 268)
Information triage (p. 276)

Keywords

Sentiment analysis (p. 1)
Probabilistic classifier (p. 2)
Generative classifier (p. 2)
Discriminative classifier (p. 2)
Linear classifier (p. 5)
Sentiment lexicon (p. 9)
Gold labels (p.11)
Precision (p.12)
Recall (p.12)
F-measure (p.13)
Macroaveraging (p.13)
Microaveraging (p.13)

FDP Curriculum

Sample Learning Objective:

6.1.2 Text representation

For example:

A. Understand the meaning of terms when used in the field of information retrieval.

B. Describe the "bag of words" approach including the following steps:
- Measuring term frequency (TF)
- Measuring sparseness: inverse document frequency (IDF)
- Combining them: TFIDF

C. Apply appropriate methods to search an example set of documents.

D. Express entropy in terms of the IDF.

Sample Question:

When representing and mining text using the "bag of words" approach, which of the following is a common name for a measure of term sparseness?

- ○ Term infrequency
- ○ Inverse document frequency*
- ○ Inverse corpus document frequency

Answer Key → LO 6.1.2. p.256

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. **Big Data, DM & ML: Ethical & Privacy Issues**
8. Big Data and Machine Learning in the Financial Industry

Reading(s):

7.1 Institute of Business Ethics. (2016, June). Business Ethics and Big Data (IBE Issue 52). London, U.K.

7.2 Institute of Business Ethics. (2018, January). Business Ethics and Artificial Intelligence (IBE Issue 58). London, U.K.

7.3 Institute of Business Ethics. (2018, May). Beyond Law: Ethical Culture and GDPR (IBE Issue 62). London, U.K.

Sample Keywords:

Key Terms

Data trust deficit (p. 2) *Veracity (p. 6)*

Keywords

Artificial intelligence (p. 1) *Code of ethics (p. 6)*

Keywords

General Data Protection Regulation (p. 1) *People risk (p. 3)*

FDP Curriculum

Sample Learning Objective:

7.3.1 General Data Protection Regulation (GDPR)

For example:

A. Describe the primary purpose of the GDPR.

B. Describe the key changes in data protection regulation including the meaning of

- rights of the individual
- informed consent
- notification
- data portability
- supervision and enforcement, and
- liability

Sample Question:

According to the General Data Protection Regulation (GDPR) what is the longest period that a company can take to report a security breach to blank without penalty?

- ○ 24 hours
- ○ 48 hours
- ○ 72 hours*

Answer Key → LO 7.3.1, p. 2

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Reading(s):

8.1 Financial Stability Board. (2017). *Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications*. Retrieved from http://www.fsb.org/wp-content/uploads/P011117.pdf

8.2 Phoon, K. and F. Koh. (2018). Robo-Advisors and Wealth Management. *Journal of Alternative Investments*, 20(3), 79-94. DOI: https://doi.org/10.3905/jai.2018.20.3.079

8.3 Monk, A., M. Prins, and D. Rook. (2019). Rethinking Alternative Data in Institutional Investment. *Journal of Financial Data Science*, 1(1), 14-31. DOI: https://doi.org/10.3905/jfds.2019.1.1.014

8.4 Simonian, J., C. Wu, D. Itano and V. Narayanan. (2019). A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model. *Journal of Financial Data Science*, 1(1), 32-44. DOI: https://doi.org/10.3905/jfds.2019.1.032

8.5 Rasekhschaffe, K. and R. Jones. (2019). Machine Learning for Stock Selection. *Financial Analyst Journal*, Forthcoming. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3330946

8.6 Gu, S., B. Kelly, and D. Xiu. (2018). Empirical Asset Pricing via Machine Learning. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281018 https://www.bryankellyacademic.org/ http://dachxiu.chicagobooth.edu/download/ML.pdf The paper posted on the FDP Institute's website will be the version used for exam questions.

8.7 López de Prado, M. (2018). The 10 Reasons Most Machine Learning Funds Fail. *The Journal of Portfolio Management*, 44 (6) 120-133; DOI: https://doi.org/10.3905/jpm.2018.44.6.120

8.8 Harvey, C. R. and Y. Liu. (2014). Evaluating Trading Strategies. [Special 40th Anniversary Issue]. *The Journal of Portfolio Management*, 40(5), 108-118. DOI: https://doi.org/10.3905/jpm.2014.40.5.108

8.9 Guida, T. (2019). *Big Data and Machine Learning in Quantitative Investments*. West Sussex, UK: John Wiley & Sons Ltd. Chapter 10.

8.10 Das, S., S. Kim and B. Kothari. (2019). Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News. *The Journal of Financial Data Science*, 1(2), 8-34. DOI: https://doi.org/10.3905/jfds.2019.1.2.008

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Sample Keywords:

Keywords

Big data (p. 4)
Artificial intelligence (p. 4)
Natural language processing (p. 5)
Machine learning (ML) (p. 4)
Supervised learning (p. 5)
Unsupervised learning (p. 5)
Deep learning (p. 5)
Reinforcement learning (p. 5)
Sentiment indicators (p. 10)
Trading signals (p. 11)
Fraud detection (p. 11)
RegTech (p. 11)
InsurTech (p. 13)
Chatbots (p. 14)
Know your customer (KYC) (p. 20)
SupTech (p. 21)
Auditability (p. 33)
Fintech (p. 35)
Robo-advisors (p.35)
Tonality analysis (p.36)

Keywords

Fintech (p. 79)
Robo-advisor (p. 80)
Work-flow (p. 83)
D2C platforms (p. 86)
Hybrid (p. 86)
B2B platforms (p. 86)

Keywords

Alternative data (p. 14)
Social media (p. 14)
Microdata (p. 14)
Data exhaust (p.14)
Rivalry (p. 16)
Excludability (p.16)
Defensive strategies (p. 17)
Defensible strategies (p. 18)
Operational alpha (p. 19)
Aggregation (p. 19)
Disaggregation (p. 19)
Volume (p. 21)
Velocity (p. 21)
Variety (p. 21)
Veracity (p. 21)
Granularity (p. 21)
Relationality (p. 21)
Flexibility (p. 21)
Actionability (p. 22)
Excludable (p. 28)
Data hoarding (p. 29)

FDP Curriculum

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Sample Learning Objective:

8.10.1 Using linguistic analysis to perform risk analysis of investments.

For example:

A. Explain the difficulties associated with manual parsing of unstructured text.

B. Describe the concept of RegTech.

C. Describe how content and structure of emails could be used for risk analysis.

D. Explain the effectiveness of textual versus hard numbers in corporate risk analysis.

E. Define RegTech.

F. Apply the net sentiment metric to calculate the polarity of a text using its Pos and Neg figures.

G. Calculate the disagreement measure of a text using Pos and Neg figures of a text.

H. Describe the findings of the paper regarding the effectiveness of email length as predictor of risk analysis of Enron.

Sample Question:

According to the article "Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News," what does the decomposition of the 'document term matrix' facilitate?

- Topic sentiment*
- Network activity
- Vocabulary trends

Answer Key → **LO** 8.10.1. p. 28-29

Essay Questions

Sample Question:

Marilyn Taylor is a quantitative analyst responsible for finding new investment ideas in the equity space. In the past, she has used unsupervised learning techniques to filter firms from a larger list, and this technique seems to have worked well for her. She mainly considers firm characteristics, such as P/E ratio, P/B ratio, and size, for her analysis. Recently, she has come across Volta Electric Company, which looks interesting to her. Currently, it has a P/E ratio of 16.5, a P/B ratio of 2.3, and a size value of 4.6 billion. To compare Volta with some other companies, she pulled out information on 3 companies that she had analyzed in the past. Out of the 3 companies, 2 were recommended for investment and 1 was not recommended. The following table provides information on the three companies.

Essay Questions

Sample Question (cont.):

A. Marilyn uses Euclidean distance to measure the difference between a new company and the companies she has analyzed in the past. What are the Euclidean distances between Volta and the 3 companies listed in the table above?

Name	P/E Ratio	P/B Ratio	Size ($ billions)	Recommendation
Northern Healthcare	14	2.1	5.9	Invest
Wholesome Foods	17	1.5	6.7	Invest
Real Tech	21	1.8	7.8	Do not invest

If we take a majority vote of the distances, what would be the recommendation by Marilyn for Volta? Explain how you decided on this.

Constructive Response Questions

Sample Answer:

Euclidean distance between Volta and:	Distance
Northern Healthcare	$\sqrt{(16.5-14)^2+(2.3-2.1)^2+(4.6-5.9)^2}=2.82$
Wholesome Foods	$\sqrt{(16.5-17)^2+(2.3-1.5)^2+(4.6-6.7)^2}=2.30$
Real Tech	$\sqrt{(16.5-21)^2+(2.3-1.8)^2+(4.6-7.8)^2}=5.54$

The recommendation would be to invest in Volta. If we consider the distances between Volta and the three companies analyzed earlier, we see that two of the companies are quite close in terms of the Euclidean distance and the recommendations for both of these companies is to invest in them. Majority voting requires us to look at all the points that are close to point we are trying to classify and assigns the majority class from the closest points to the new point.

Essay Questions

Sample Question:

B. Marilyn also calculates weighted voting or similarity moderated voting for any new company using the companies she has analyzed in the past.

What are the similarity weights and contribution to the probability of investing or not investing for the 3 companies?

What would be the estimated probability of investing and not investing for Volta?

Constructive Response Questions

Sample Answer:

Name	Euclidean Distance	Similarity Weight	Contribution
Northern Healthcare	2.82	$\frac{1}{2.82^2} = 0.125748$	$\frac{0.125748}{0.347366} = 0.362$
Wholesome Foods	2.30	$\frac{1}{2.30^2} = 0.189036$	$\frac{0.189036}{0.347366} = 0.544$
Real Tech	5.54	$\frac{1}{5.54^2} = 0.032582$	$\frac{0.032582}{0.347366} = 0.094$
Sum		0.347366	

Estimated probability of investing = 0.362 + 0.544 = 0.906

Estimated probability of not investing = 0.094

THE FINANCIAL DATA PROFESSIONAL PROGRAM

MARCH 16 – APRIL 4, 2020

CANDIDATE EXPERIENCE

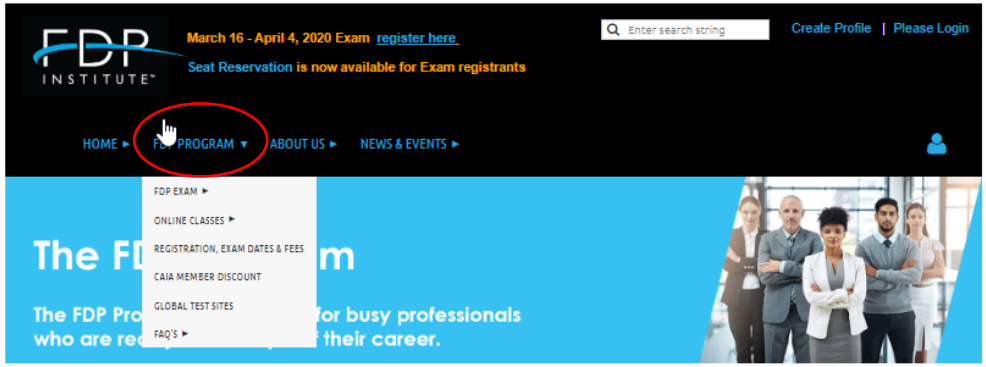
FDP Exam Candidates

FDP Progam

- Exam Registration
- Exam Curriculum
- FAQ's
- Webinars & Events

If you are registered for the exam, we urge you to reserve your seat as soon as possible:

https://www.prometric.com/test-takers/search/fdpi



OVERVIEW

Anyone can register to take the exam. The FDP Charter is awarded after a candidate has successful completed the FDP exam and has submitted the Certificates of Completion for the online classes.

REGISTER FOR THE EXAM

Register here for the Exam. Make sure you have your current passport or driver's license with you when you register. The name on your government issued identification must be exactly the same as on your exam registration. Once you have registered you can log into your FDP account.CAIA members are eligible for a 20% discount of the registration fee. Learn more.

RESERVE YOUR EXAM SEAT ** seat reservation is on a first come, first serve basis.**

Through our partnership with Prometric, FDP has over 450 test locations around the globe. Early seat reservation ensures availability on the date/ time that suits you most.

PREPARE FOR THE EXAM

A candidate who is able to meet all learning objectives in the Study Guide should be well prepared for the exam. Candidates should anticipate approximately 200+ hours of study in preparation for the FDP exam.

PASS THE EXAM

The FDP Institute offers a number of resources to help you prepare. Learn more.

SUBMIT YOUR ONLINE CLASS CERTIFICATES

Upon completion of your online classes, the provider will award you a Certificate of Completion. These can be submitted to certificates@fdpinstitute.org. After verification by the FDP staff you will receive a confirmation of receipt.

EARN YOUR FDP CHARTER

Online Resources (for registered exam candidates)

- Candidate Handbook
- Study guide

Both are exam specific

From the handbook: Prometric tutorials

Flagging Questions

You can flag a question as a reminder to go back and check your answer or attempt it later.

To flag a question, click the **Flag** button displayed at the bottom of the exam screen.

Any questions that are flagged for review will show a flag icon on the numbered button, as shown below. Click the **Flag** button again to remove the flag.

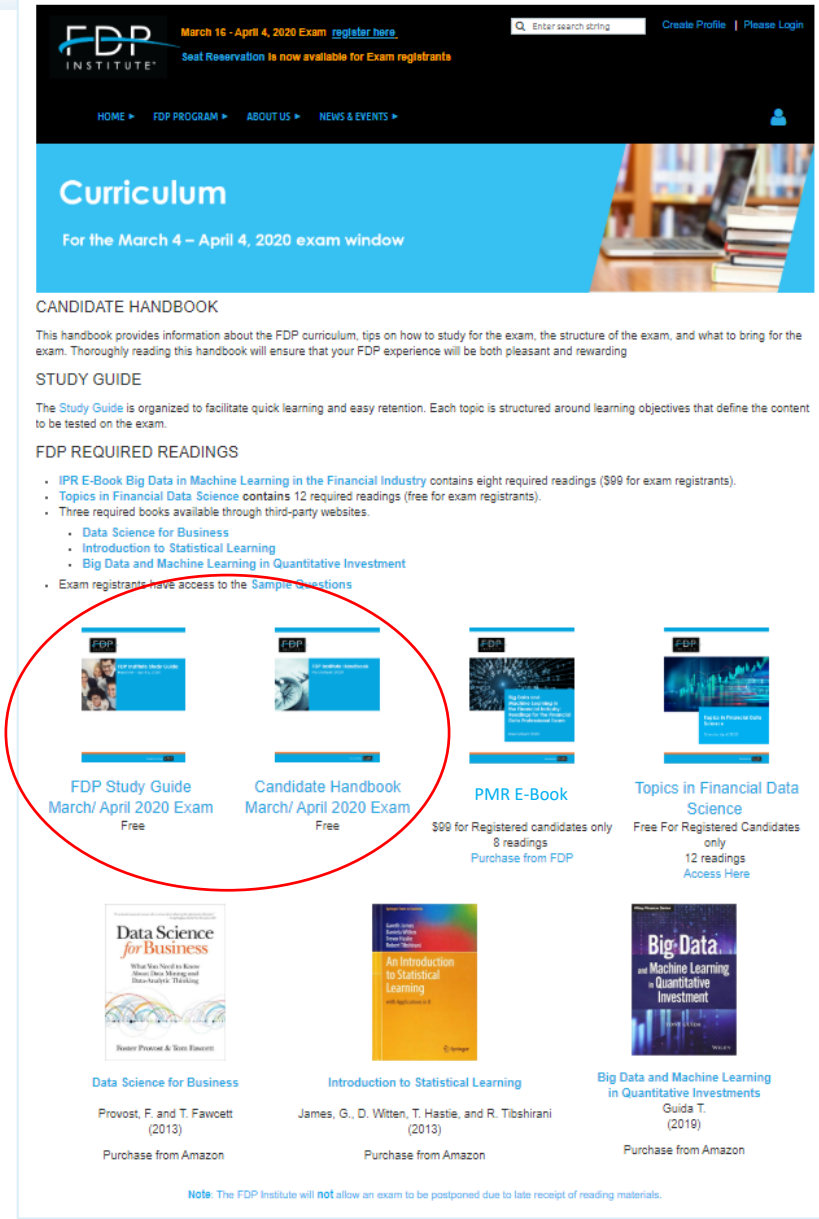
Highlighting Text

During the examination, you will be able to highlight question text that you feel is important to refer back to as you progress through the exam. The highlight will remain present as you navigate through the exam, unless you select to remove it.

To highlight text, click and drag the mouse cursor over the desired text. Click the **Highlight** button, as shown in the image below, that appears after releasing the mouse button. To remove, click on any area of the highlighted text and click the **Highlight** button again.

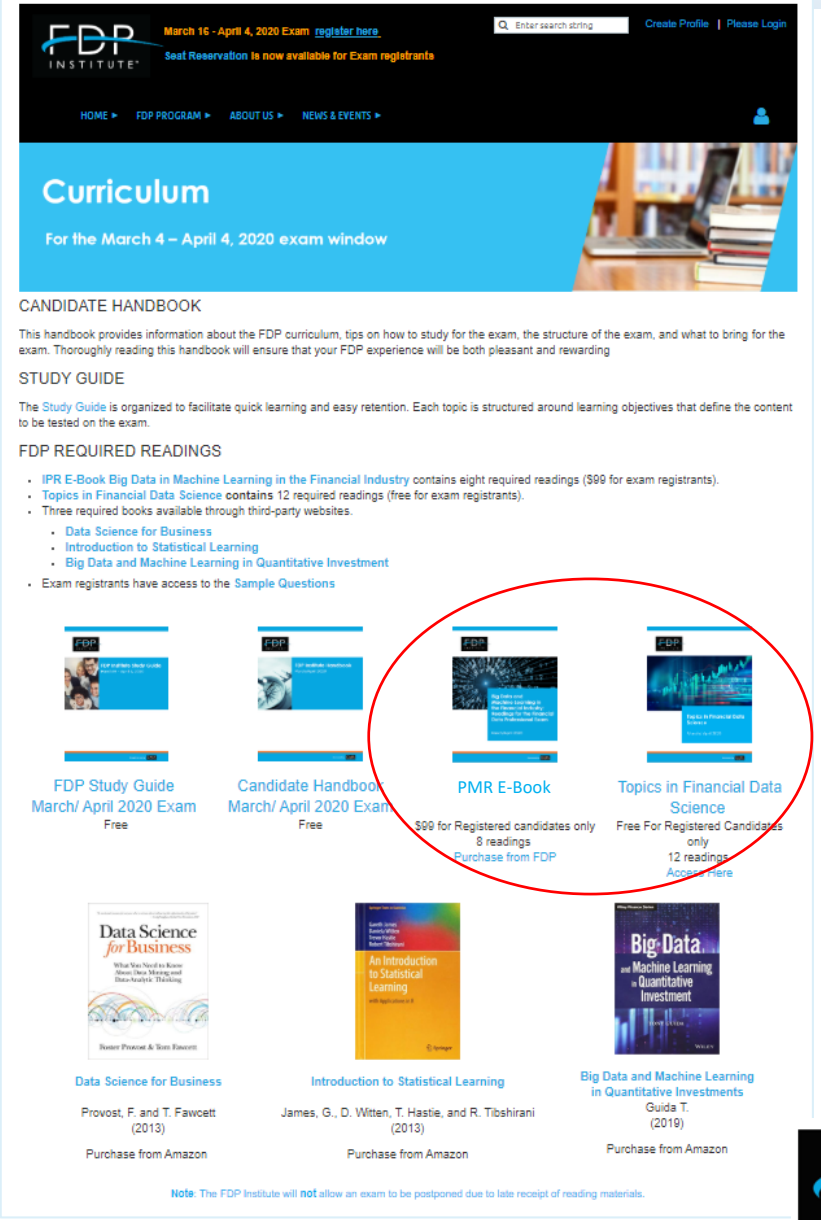
How would you characterize the young Frederick Douglass?

The highlight feature cannot be applied to text within the answer options.



Online Resources (for registered exam candidates)

- Curriculum and Study Tools
- Order the **PMR E-Book**
- Download the (12) readings from **Topics in Financial Data Science**
- **Sample Questions** *(registered exam candidates have access to a webpage with directions)*



Study Time

- Average time for the FDP exam is 200 – 250 hours
- Plan your study time

Exam Day

- **ID Policy**
  - Two valid (not expired) forms of ID
    → Make sure your name in your FDP profile MATCHES that on your ID
  - Both must have a signature
  - One must have a photo
  - Passport preferred
  - The name on your IDs must exactly match the name with which you schedule your exam.

- **Calculator Policy**
  - Texas Instruments BA II Plus (including the TI BA Plus Professional)
  - Hewlett Packard 12C (including the HP 12C Platinum, HP 12C 25th Anniversary Edition, HP 12C 30th Anniversary edition and HP 12C Prestige)
  - Exam proctors will inspect your calculator prior to the start of the exam.

Exam Session

- You will receive a dry-erase white board

Non-disclosure agreement (must be completed within 5 minutes)	5 minutes
Exam instructions	5 minutes
FDP Multiple Choice (MC) Questions	100 minutes
Break (optional)	20 minutes
FDP MC Questions (approx. 5) and 2-4 Constructive Response Questions	100 minutes
Comment Period (optional)	10 minutes
Total Exam Session Time	**4 hours**

Pass Rate

November Exam → 77%

March – April Exam → On our way to double the number of participants

Candidate Performance Report

- The FDP Candidate Performance Report is intended to aid in self-assessment by indicating your areas of relative strength and weakness among the topics in the exam, compared to a reference group.
- The reference group is comprised of all candidates whose total test scores were within the bottom quartile of those who passed the exam during the testing window in which you sat for the exam.

FDP Examination Topics	Relative Performance Level*
1. Introduction to Data Science & Big Data	Above
2. Data Mining & Machine Learning: Introduction	Above
3. Data Mining & Machine Learning: Regression, LASSO, Predictive Models, Time Series & Tree Models	Above
4. Data Mining & Machine Learning: Classification & Clustering	Above
5. Data Mining & Machine Learning: Performance Evaluation, testing & False Discoveries	Above
6. Data Mining & Machine Learning: Representing & Mining Text	Above
7. Big Data, Data Mining & Machine Learning: Ethical & Privacy Issues	Above
8. Big Data & Machine Learning in the Financial Industry	Equal

Q & A

Kind reminders of upcoming webinars as we go through the Q & A

Guen Dondè
Head Of Research at Institute of Business Ethics.
Co-researcher *"IBE Ethics at Work"* and *"Implications of AI on business ethics*
January 22, 2020 - 11am EST

Some of the topics that will be discussed:

- What is the impact of the GDPR from an ethical standpoint?
- Can codes of ethics help?
- How involved should regulators be?

Note: Three readings on the upcoming exam are published by the Institute for Business Ethics
Our conversation will last approximately 35 minutes with the remaining time for Q&A.

REGISTER

Gene Getman
Co-Author of *"Big Data is a Big Deal: An investor's guide to the applications and challenges of alternative data "*
January 29, 2020 - 2pm EST

Gene Getman is a Client Portfolio Manager for the 1798 Alternatives group, focused on Investor Relations for the Hedge Fund business. He is the Product Specialist for the 1798 Q Strategy and other innovative or capacity constrained alternative investment strategies based in New York. Before joining LOIM in April 2013, Gene was an Analyst on the Barclays Capital Solutions team, where he worked with Pensions, Endowments, Foundations and their consultants to make capital introductions to hedge funds. He began his career working for the NYSE Market Makers group within Lehman Brothers and Barclays Capital.

Note: The article "Big Data is a Big Deal" is one of the readings for the upcoming exam.

REGISTER

Seoyoung Kim
Co-Author of *"Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News!"*
February 19, 2020 - 1pm EST

Natural language processing is a fast-growing area of data science for the finance industry. Recent advances in financial technology (FinTech) have dramatically transformed the financial landscape with respect to the way we access, invest, and transfer financial capital. In this article, the authors explore a promising avenue for the use of natural-language processing in an effective yet non-invasive method by which to monitor the health and integrity of financial institutions and corporations in general by analyzing corporate emails and news.

Note: The article "Zero-Revelation RegTech: Detecting Risk through Linguistic Analysis of Corporate Emails and News!" is one of the readings for the upcoming exam.

REGISTER

Contact us !

We wish you continued success with
your exam preparations!

If you have additional questions, please contact us at
info@fdpinstitute.org