The Financial Data Professional Institute (FDPI) was established by CAIA Association to address the growing need in finance for a workforce that has the skills to perform in a digitized world where an increasing number of decisions will be data and analytic driven. The FDP credential is the first of its kind in the industry and reflects expertise in data science and its practical applications in finance.

To earn your FDP Charter:

- Successfully complete the FDP exam
- Complete the online requirements

Learn more at www.fdpinstitute.org
Financial Data Professional Institute

The Evolution of Machine Learning in Investment Management

A conversation with
Rick Roche, CAIA, Managing Director, Little Harbor Investors

Hosted by
Kathryn Wilkens, Ph.D., CAIA, FDP Institute

www.fdpinstitute.org
April 22, 2020
Agenda

• Welcome
• Introductions

Rick Roche, CAIA
Managing Director
Little Harbor Investors

Kathryn Wilkens, Ph.D., CAIA
Sr. Curriculum Advisor
FDPI

Mirjam Dekker
Project Manager
FDPI

• A conversation with Rick Roche
• Q & A (please add your questions to the chat!)
QUANTS’ QUANDARY
Crossing the Chasm

By Richard P. Roche, CAIA®
Evolution of ML in Invt Mgmt.

AI/Machine Learning Birth
Playing Games -- No More
Machine Learners – Asset Mgmt Categories/Tasks
Revenge of the Nerds – ML Pioneers & Practitioners
Why Alternative Invt Data?
Legal Risks of Alt Invt Data
Summary. Action Steps...

(c) 2020 Little Harbor Advisors, LLC
ENAIC: “The 2\textsuperscript{nd} Machine Age”

*Programmed Inequality?*

- ENAIC 6 - All were Women
- 37\% of Computer Science Degrees to women in 1984!
- Today, only 18\% are women
- Only 1 in 5 FB or GOOGL engineers are Women
- 1991. Peak PC Women - 36\%
- Today USA computing force is 25\% female; 76\% male

Playing Games to Major Disruptor!

- ML detect patterns and make predictions by processing data, rather than explicit programming
- ML avoid elaborate programming by defining rules rely on Inputs
- ML algos adapt in response to new data & experiences to “learn” or improve efficacy over time!
- Machine Learning intersection of Computer Science & Statistics to find patterns in datasets
- ML heavily weighted towards predictive modeling vs. inference!
- Uses probability to make product recommendations, speech recog...

“How can we build computer systems that automatically improve with experience...” , Tom Mitchell, PhD. Former Chair, Machine Learning Dept. Carnegie Mellon Univ.

2015 - 67,318 papers on ML in SSRN over 12-months

2017 – 14,000 asset mgmt papers - BIG Data/Analytics

ML = Cheap Prediction

When prices fall, use more!

Use information you have to generate info you don’t have

Cheap Predictions = + Demand

Human Judgment/Expertise!
Features (Attributes) & Labels

Training examples: United States Postal Service Handwritten Digit Database

Training labels

Supervised Learner

Accurate digit classifier

2
ML Model Behavior

Classifier network. Algo classifies new observed data into categories.

- Regression trained with goal of level estimation/numeric value
- Clustering. Assign observations into subset “cluster” where data are similar, e.g. Netflix/AMZN

Image credit: Machine Learning in MATLAB
Learning From Experience

Deep neural networks learn by adjusting the strengths of their connections to better convey input signals through multiple layers to neurons associated with the right general concepts.

When data is fed into a network, each artificial neuron that fires (labeled “1”) transmits signals to certain neurons in the next layer, which are likely to fire if multiple signals are received. The process filters out noise and retains only the most relevant features.
Deep Learning – ANN/DNN

- DNN has 3 layers: 1) Input; 2) Hidden Layer(s); 3) Output
- Analyzes data by passing thru multiple layers (neurons) - software-based calculators – that learn increasingly complex features of data
- Extract data features passing multiple filters in over-lapping segments data/weight data.
- ANN infer patterns in data and make predictions
- Widely-used in image, voice recognition; NLP-text analysis

Biologically inspired programming attempts to mimic human brain
AI-ML After Decades of Promise – Why Now

- Moore’s Law. Cheaper/faster; 1-Gig storage $277 to $0.03
- Cloud Computing – lowered AI barrier of entry for crowd
- 2016 1.49Bn Smartphones Sold! 40% Adults own SmtPhs
- 1,459 Satellites Dec-16/IoT Smart Sensors 2Xs every 2 yrs.
- BIG Data Cambrian Explosion (1,000+ Alt Data Vendors)
Quant Model Evolution

1969
- Statistical Arbitrage
  - Convertible Arbs; Pairs Trading

1970s
- Fully Systematic; Trend Followers
  - Buy Winners; Sell Losers

Mid-1990s
- Hi-Speed Algorithmic Traders
  - The Island-ATD-RenTec Fusion

Mid-2000s
- High Frequency Traders (HFTs); Predictive Algos
  - Trades in Microseconds; 2/3rds U.S. Stock Trades

2012 - Beyond
- Deep Learning/Neural Nets
  - Machine learning algs used to predict stock and indexes. Quants use Alternative Data in models to seek elusive Alpha.

(c) 2019 rickroche
Very Few End-2-End Portfolio Design

<table>
<thead>
<tr>
<th>Investment Process</th>
<th>Possible Machine Learning Applications</th>
<th>Exemplary Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investment Universe</td>
<td>• Identifying uncorrelated assets&lt;br&gt;• Data mapping allowing for interpretation&lt;br&gt;• Proxy assets with more liquid instruments</td>
<td>• Dendrogram&lt;br&gt;• PCA, manifold learning&lt;br&gt;• k-means</td>
</tr>
<tr>
<td>Signal Creation</td>
<td>• Create nonlinear forecast&lt;br&gt;• Forecast based on predefined factors&lt;br&gt;• Forecast without prior factor knowledge&lt;br&gt;• Forecast based on new data sources</td>
<td>• Linear, lasso, ridge regression&lt;br&gt;• k-nearest neighbors&lt;br&gt;• Bag-of-words, term frequency-inverse document frequency</td>
</tr>
<tr>
<td>Portfolio Construction</td>
<td>• Improve estimate of input parameters&lt;br&gt;• Dynamically maximize target value&lt;br&gt;• Scenario and stress testing</td>
<td>• PCA&lt;br&gt;• Lasso, ridge regression&lt;br&gt;• k-means</td>
</tr>
<tr>
<td>Trading</td>
<td>• Pricing based on sparse data&lt;br&gt;• Non-linear relations in transaction costs</td>
<td>• k-means&lt;br&gt;• Lasso, ridge regression</td>
</tr>
</tbody>
</table>

*Figure 2: Artificial intelligence applications in quantitative finance*

ML Diffusion: Hype vs. Reality

FIGURE 18: Machine Learning Adoption and Sentiment

<table>
<thead>
<tr>
<th>Use of Machine Learning</th>
<th>Areas of Use of Machine Learning Techniques</th>
<th>Limitations to ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: Does your firm use machine learning techniques within the investment process?</td>
<td>Data Gathering (finding / scraping, cleaning up, structuring and preparing data for use in modeling process)</td>
<td>“ML allows you the flexibility to do more analyses, but the no free lunch theorem suggests that there is a greater risk involved in relying on the output. You need a lot of data for the results to be reliable.”</td>
</tr>
<tr>
<td>38%</td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>Not Using</td>
<td>Using</td>
<td></td>
</tr>
<tr>
<td>Investment Decision (analyzing data to find statistical relationships, then creating models to trade based upon these relationships)</td>
<td>25%</td>
<td>“ML is just an estimation technique: With it, you will be able to find signals and noise, but it can’t distinguish between them. You will still need to find the rationale for the relationship.”</td>
</tr>
<tr>
<td>Portfolio Construction and Execution</td>
<td>25%</td>
<td>“We tend to prefer a hypothesis-driven research process over relying on ML techniques to find solutions. ML vastly increases the risk of data mining, which we seek to avoid.”</td>
</tr>
</tbody>
</table>

Source: Strategic Consulting Survey results and analysis

“Rise of the Machines: Landscape & Recent Developments in Quant HF Strategies, Products & Managers”, Barclays Capital Solutions, June 2017, page 14
More Realistic ML-AWM Diffusion

“AI Pioneers in Investment Management”, CFA Institute, 2019
Renaissance Technologies LLC (RenTec)

- Founded 1982. $166Bn-AuM
- James Simons-PhD Math, NSA code breaker/Math Champion
- Use Big Data/advanced math models long before other HFs.
- Cryptography/Hidden Markov models seek short-term signals
- Medallion averaged 66.1% ann. Rtn-before fees; 39.1% Net Rtn!
- $100 invested would’ve grown to $398.7 Million (1988-2018)
- RenTec controls 4 Hedge Fund
- Medallion closed to investors -

1993. RT raids IBM’s Speech Recognition/Computational Linguistic Speech recognition akin to equity markets- *More Noise than Signal*
*NLP* maps speech to predict next sound; adapted to securities
Henry Laufer used 5-minute trading bars- easier predict S-T moves
Late 1990’s Peter Brown/Bob Mercer adapted Laufer to Equity Mkt
150-300K Trades Daily; Added foreign-mkt algos/Sharpe ratio - 6.0
“Our goal is to look at the intersection of computers and capital, and find as many interesting and profitable things to do in that intersection as we can.”
– David E. Shaw, PhD

Global Invt & Technology Firm
- PhD/computer science Stanford
- 1986. Morgan’s APT hired D.E. Shaw, parallel processing expert
- 1988. Quit MS; Open shop over communist bookstore/Village
- Statistical Arbitrage pioneer
- Dr. Shaw infamous for keeping DE Shaw’s secret sauce Secret!

David E Shaw is Chief Scientist
- $82B of invt capital (Mar 2020)
- 250 Technologists at DE Shaw

1 The Quants, Scott Patterson, 2010, p. 43
Two Sigma – Applied Data Science

- Co-Founder, David Siegel, PhD, in computer science from MIT
- Co-Founder, John Overdeck, Intl Math Olympiad, MS-Stats-Stanford
- Co-Founders worked at D.E Shaw
- 1,100+ employees. 2/3rds in R&D
- “Harness data from thousands of diverse sources”
- Guided by the Scientific Method
- Hired, Mike Shuster, Google Brain, to consolidate AI research. Shuster was Google Sr. Scientist for 12-yrs
- Host hackathons, sponsor contests, programming challenges to recruit future associates!

(c) 2020 Little Harbor Advisors, LLC
Bridgewater Associates, LP

- Founded 1975. $235Bn-RAuM
- Dalio obsessed with cognitive biases/Principles-Algorithms!
- Believes in expert systems like “Systematized Intelligence Lab” to automate day-2-day management
- Dalio doesn’t like *machine learning* moniker (prefers AI)
- Bridgewater models informed by “Economic Machine” works
- Combine Data Science, Software Engineering + Product Mgmt
- Bridgewater more a technology company than it is a hedge fund.

Dave Ferrucci, PhD Senior Technologist
IBM Fellow led 30 A.I. researchers who built DeepQA Project. In 2011, Watson beat the world’s best Jeopardy! players.

About Ray Dalio: Paul Volker admits to feeling that “he has a bigger staff, and produces more relevant statistics and analyses, than the Federal Reserve.” – Paul Volker in “Man and Machine”, *The Economist*, 03/10/12

Accessed from https://jobs.bwater.com/explore/technology_lounge on 07/04/17
AI IN ASSET MANAGEMENT MAP
(@DiligenceVault)

KNOWLEDGE, CAPITAL, DATA & AMBITIONS

2°  AQR  BRIDGEWATER

CITADEL  D E Shaw & Co

Man  AHL  Point 72

Renaissance  systematica  investments

Winton  WORLDQUANT

NEW & EMERGING IDEAS

Chord  AIDYIA

cerebellum capital  EMMA

euclidean technologies  Katonah Capital Partners

MOV37  Numerai

Rosetta Analytics  Rebellion Research

Simplex  Asset Management  Tech Trader

PLATFORMS, SERVICES, RESOURCES

ABAKA  BINATIX  KENSHO

Kimerick Technologies

ibm  Watson

QuantConnect  Quantiacs  sentient

WALNUT ALGORITHMS
**Traditional Econ Data Unreliable**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualize Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>Number of Months</td>
<td>Annualized</td>
</tr>
<tr>
<td>832.79</td>
<td>2</td>
<td>4,996.74</td>
</tr>
</tbody>
</table>

**Fear & Greed Index**

What emotion is driving the market now?

- **Now:** Fear 29
- **1 Week Ago:** Extreme Fear 8
- **1 Month Ago:** Extreme Fear 21
- **1 Year Ago:** Neutral 48

---

"Does anyone know what is happening?" Paul Donovan, Chief Economist, UBS Mar 30, 2020

---

(c) 2020 Little Harbor Advisors, LLC
Falsehoods Travel Farther/Faster

“A lie can travel halfway around the world while the truth is still putting on its shoes” – Jonathan Swift (not Mark Twain)

The top 1% of falsehoods routinely reach 1,000 to 100,000 people, while accurate reports rarely spread to more than 1,000.

Fake News is retweeted 70% more often than the truth. “The Spread of Truth and False News Online”, Science, Mar 2018

(c) 2020 Little Harbor Advisors, LLC
"Big Data is a Big Deal", Gene Getman, CAIA. Lombard Odier, Mar-2019
Data Sets for Alpha Generation

Estimates that only 0.5% of data currently analyzed. ~80% of Big Data in unstructured/unsupervised data sets (source: Patrick Wolfe, PhD UCL Big Data Institute, 2013)
Eagle Alpha – 1,300 datasets

COVID-19 Report
Using Alternative Data to Track the Impact of the Coronavirus
Edition 2 – April 2nd 2020
Coronavirus For Investors: Using Alternative Data to Unveil Real COVID-19 Insights, RavenPack & Wolfe Research, April 08, 2020
NEWS SENTIMENT IS A GREAT SIGNAL DIVERSIFIER THROUGH CRISIS PERIODS

Trading the spread between positive and negative sentiment stocks offer significant outperformance through periods of market stress.

---

Coronavirus For Investors: Using Alternative Data to Unveil Real COVID-19 Insights, RavenPack & Wolfe Research, April 08, 2020
Legal Risks – Alternative Data
## Legal Risks of Alternative Data

Various potential legal risks are associated with alternative data.

<table>
<thead>
<tr>
<th>Source: Integrity Research Associates</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Availability on exclusive basis</th>
<th>Insider trading</th>
<th>Privacy law</th>
<th>Copyright law</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web data</td>
<td>Unlikely</td>
<td>Lower risk</td>
<td>Lower risk</td>
<td>Higher risk</td>
</tr>
<tr>
<td>Web traffic</td>
<td>Unlikely</td>
<td>Lower risk</td>
<td>Lower risk</td>
<td>Higher risk</td>
</tr>
<tr>
<td>Sentiment data</td>
<td>Unlikely</td>
<td>Lower risk</td>
<td>Lower risk</td>
<td>Higher risk</td>
</tr>
<tr>
<td>Credit card transactions</td>
<td>Possible</td>
<td>Higher risk</td>
<td>Higher risk</td>
<td>Lower risk</td>
</tr>
<tr>
<td>Email Receipts</td>
<td>Possible</td>
<td>Higher risk</td>
<td>Higher risk</td>
<td>Lower risk</td>
</tr>
<tr>
<td>Geolocation</td>
<td>Possible</td>
<td>Higher risk</td>
<td>Higher risk</td>
<td>Lower risk</td>
</tr>
<tr>
<td>Satellite</td>
<td>Possible</td>
<td>Higher risk</td>
<td>Lower risk</td>
<td>Lower risk</td>
</tr>
</tbody>
</table>

Source: Prime Capital AG, October 2018, p.6
"We always overestimate the change that will occur in the next two years and underestimate the change that will occur in the next ten."

Summary

Figure 39: Classification of Machine Learning techniques

Source: J.P. Morgan Macro QDS

Exhibit 2

Asset managers are turning to new sources of investment research.

Rick’s Recommended Reads

1. "Machine Platform Crowd: Harnessing Our Digital Future" by Andrew McAfee and Erik Brynjolfsson

2. "J.P. Morgan: Big Data and AI Strategies Machine Learning and Alternative Data Approach to Investing"

3. "The Quants: How a New Breed of Math Whizzes Conquered Wall Street and Nearly Destroyed It" by Scott Patterson

4. "Inside the Black Box: A Simple Guide to Quantitative and High-Frequency Trading" by Rishi Narang
Recommended Reads – Markets

- **The (Mis)Behavior of Markets**
  - By Benoit Mandelbrot & Richard L. Hudson
  - A Fractal View of Financial Turbulence

- **A Man For All Markets**
  - By Edward O. Thorp
  - From Las Vegas to Wall Street, How I Beat the Dealer and the Market

- **The Man Who Solved the Market**
  - By Jim Simons
  - How Jim Simons Launched the Quant Revolution

(c) 2020 Little Harbor Advisors, LLC
The information contained in this presentation was obtained from sources believed to be reliable, but its accuracy cannot be guaranteed. The information communicated is NOT intended to constitute individual investment advice and is not designed to meet your personal financial situation. Opinions expressed herein are those of the presenter, Rick Roche, and not of Little Harbor Advisors. Such opinions are subject to change without notice. Information may become outdated & there is no obligation to update any such information. Any errors are the sole responsibility of Rick Roche.
Rick Roche (and Little Harbor Advisors) have NO affiliations with the industry sources and firms cited in this presentation. In addition, Rick Roche has not received any compensation for mentioning services or products. His citation of said firm or service is not meant to be an implied endorsement. Rick Roche and Little Harbor Advisors are not affiliated with any of the firms cited. Opinions expressed in this presentation are subject to change without notice. Information may become outdated and there’s no obligation to update any such information.
Online Requirements

- Choose between two (2) Python or two (2) R programming classes (4-6 hours each).
- The online classes can be completed before or after FDP exam.
- No programming background is needed to complete the online classes or the FDP exam.

FDP Exam

1. Introduction to Data Science & Big Data
5. Machine Learning: Classification & Clustering
7. Data Mining & Machine Learning: Naïve Bayes & Text Mining

Next exam window
October 12 – November 8, 2020

Updated curriculum available on May 10th
Registration opens May 10th
Q & A

Kind reminders of upcoming webinars as we go through the Q & A. Add your questions in the chat room please.

For webinar recordings, slide decks and upcoming webinars:

www.fdpinstitute.org/webinars
www.caia.org/caia-infoseries
In Closing

- The Next FDP Exam: October 26 – November 8, 2020
- New curriculum will be available May 10th
- Registration opens May 10th
- For a recent candidate webinar go to www.fdpinstitute.org/webinars

Learn more about the FDP Institute at www.fdpinstitute.org