Application of Alternative Data and Convolutional Neural Networks for Forecasting Stock Prices

Erez Katz

CEO & Co-Founder, Lucena Research

About

Lucena Research
Predictive Analytics

About Lucena

- Leader in Machine Learning and AI

- Connect big data with investment professionals

Erez Katz, Lucena CEO

Buy Side Alternative Data Adoption

Lucena Research
Predictive Analytics

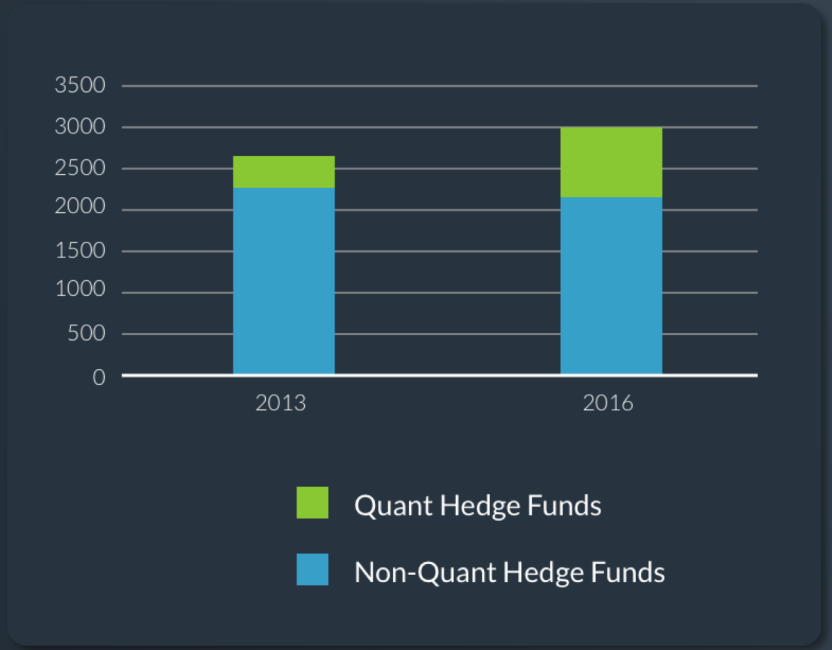


Image Source: Jefferies, Hedge Fund Research

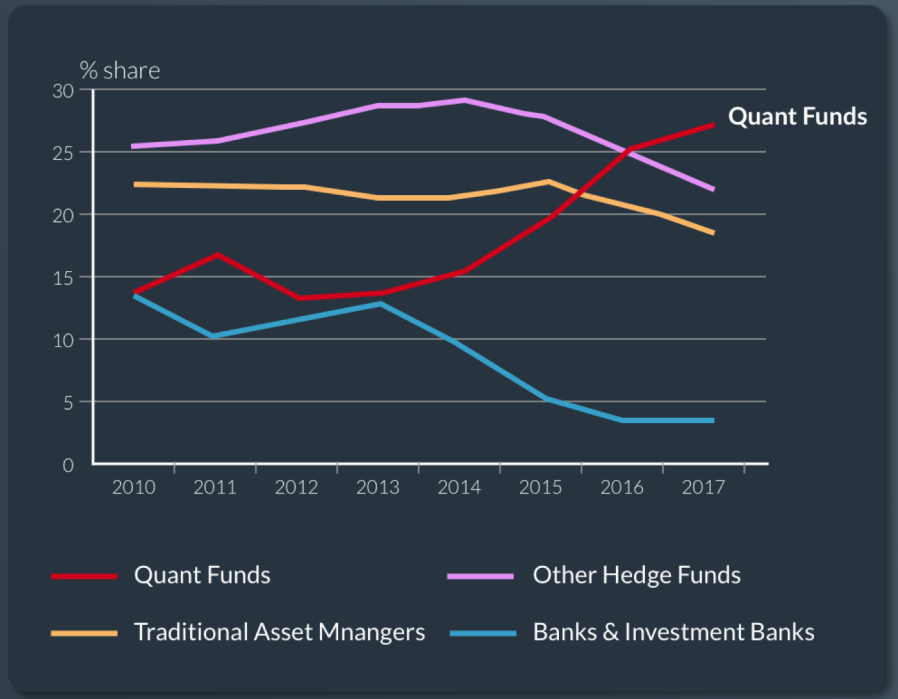
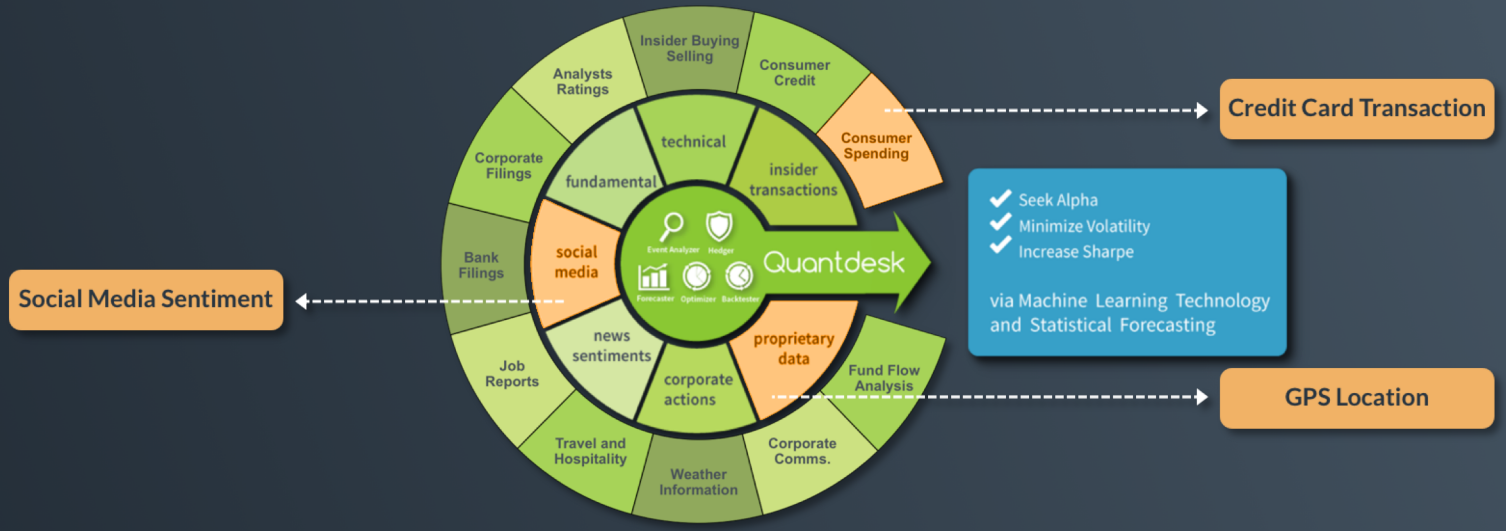


Image Source: WSJ, Tabb Group

Big Data & Alternative Data

Lucena Research
Predictive Analytics

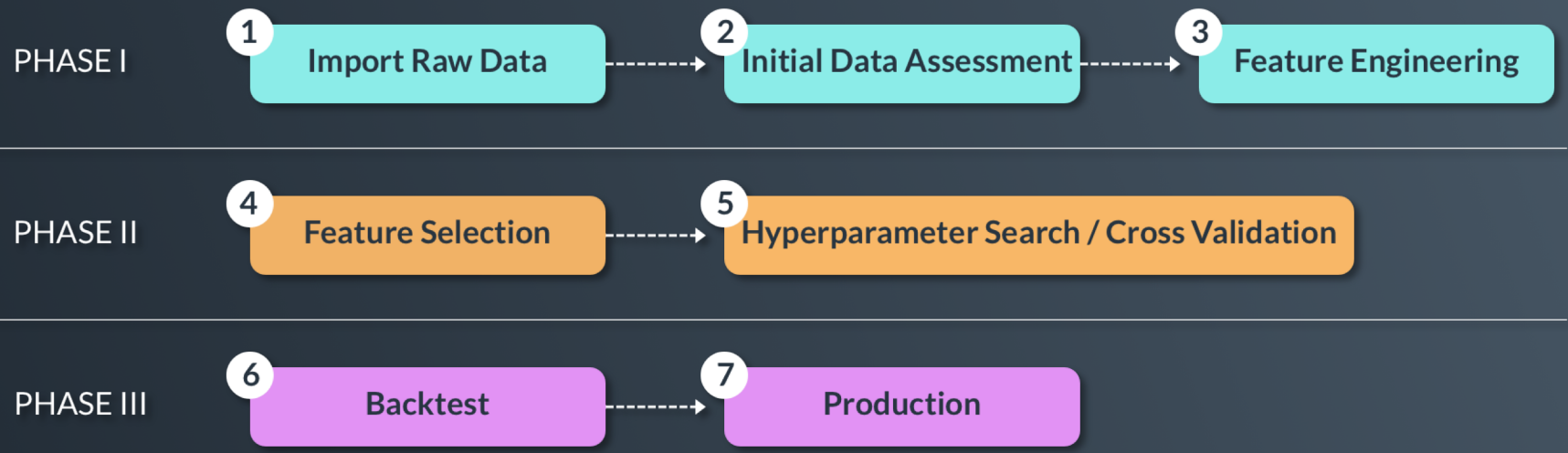
- Traditional Data Types: Mainly Technical and Fundamentals
- Equity Specific Alternative Data: Corporate Action, Insider Buying Selling
- Non Financial Data Repurposed for the financial markets: Credit Card, Weather Data
- Unstructured Data: News Feed, Social Media Feed, Speeches, Earnings Calls



Data Validation Process

Lucena Research
Predictive Analytics

- Critical to follow a regimented process by which you can reliably consume, validate, and enhance data in order to generate predictive signals.

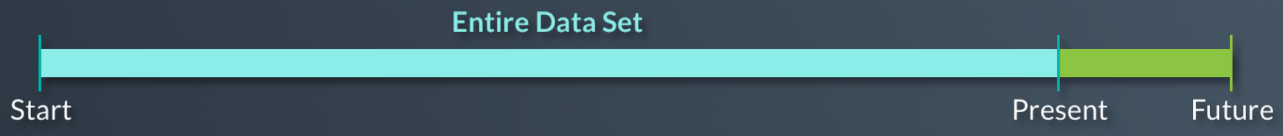


Data Validation Process

Lucena Research
Predictive Analytics

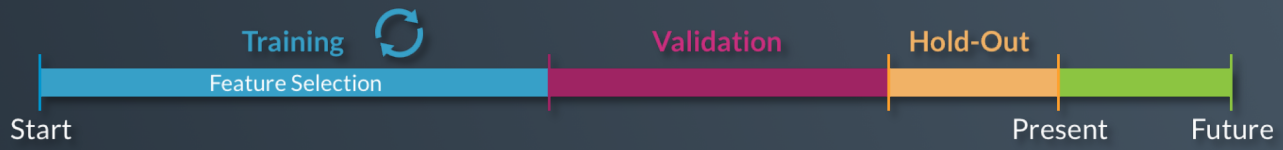
PHASE I

1. Import Raw Data
2. Initial Data Assessment
3. Feature Engineering

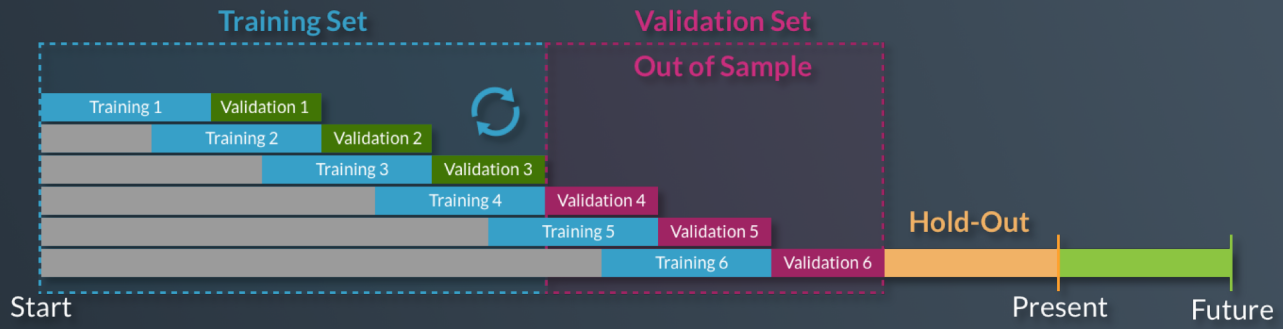


PHASE II

4. Feature Selection



5. Hyperparameter Search
Cross Validation

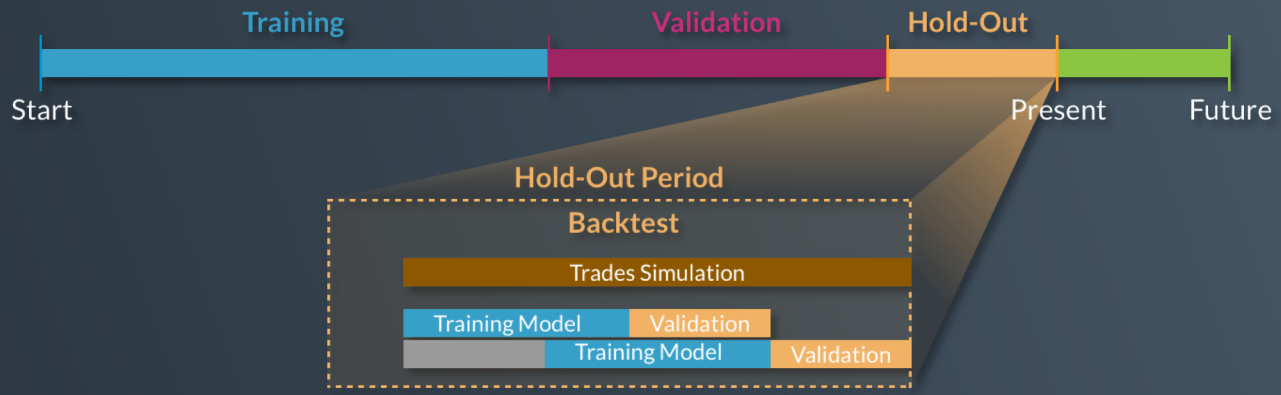


Data Validation Process

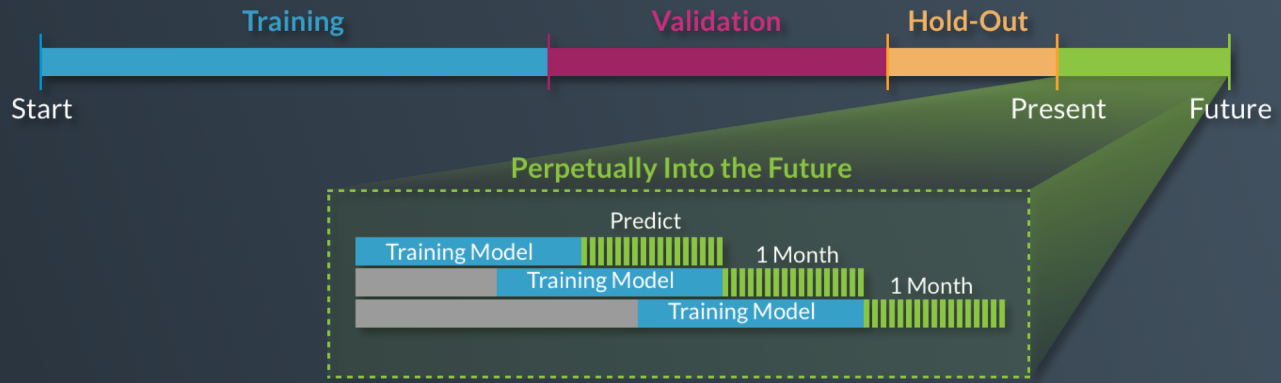
Lucena Research
Predictive Analytics

PHASE III

6 Backtest



7 Production



Automating Data Qualification? Why?

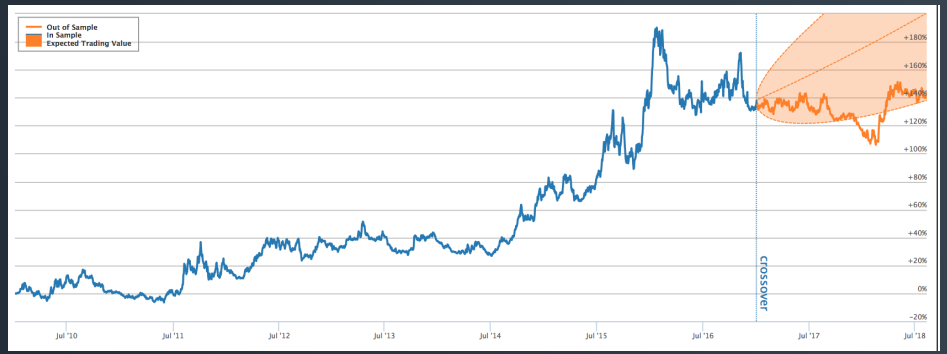
Lucena Research
Predictive Analytics

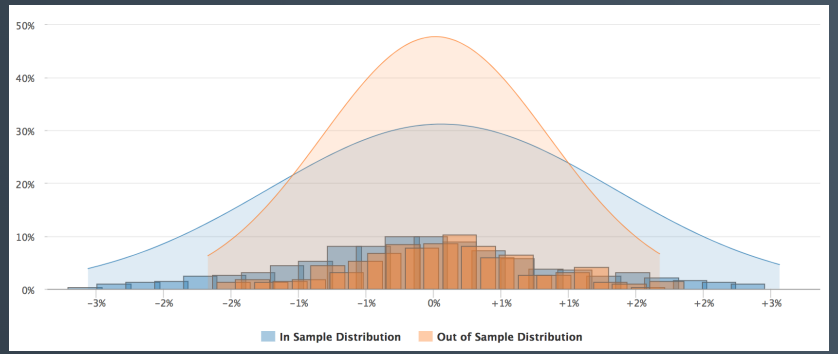
Qualification Statistics

	Value	Best	Grade
Percentage of NaNs	0.0799	Lower	✗
Percentage of Symbols Mapped	0.84	Higher	✓
Likelihood of Survivorship Bias	0.83	Lower	✗
Best Decile Sharpe	0.7353	Higher	✓
Signal Distribution Normality	0.91	Higher	✓
Percentage of Missing Trade Days	0.16	Lower	✗
Percentage of Extra Trade Days	0.01	Lower	✓
Percentage of Days with Multiple Signals per Asset	0.2037	Lower	✗
Average Data Frequency (Days)	65.8316	Closer to 1	✗
Data Type Prevelance	1.00	Higher	✗
Likelihood of Underlying Model Change	0.64	Lower	✗

- **Completeness**
- **Authenticity**
- **Consistency**
- **Survivorship Bias**

Below is an example of an inconsistent model:





Data Qualification What To Look For?

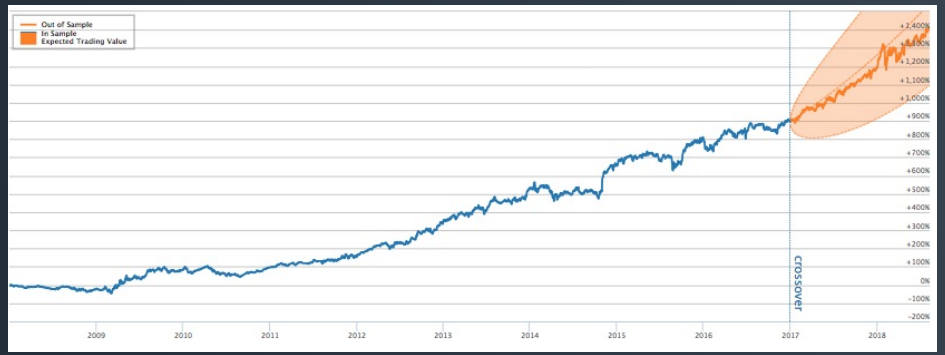
Lucena Research
Predictive Analytics

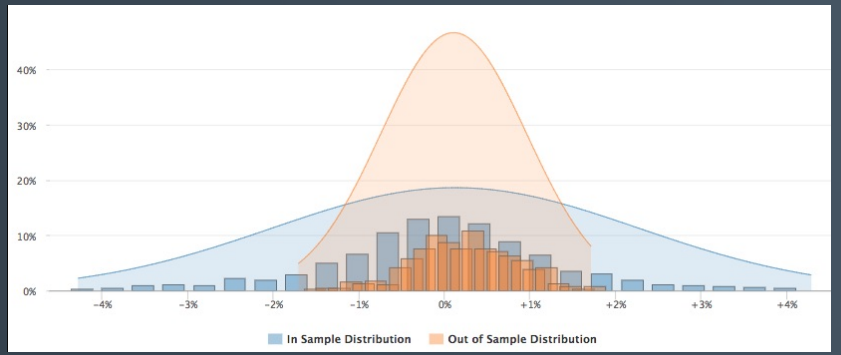
Qualification Statistics

	Value	Best	Grade
Percentage of NaNs	0.0799	Lower	✗
Percentage of Symbols Mapped	0.84	Higher	✓
Likelihood of Survivorship Bias	0.83	Lower	✗
Best Decile Sharpe	0.7353	Higher	✓
Signal Distribution Normality	0.91	Higher	✓
Percentage of Missing Trade Days	0.16	Lower	✗
Percentage of Extra Trade Days	0.01	Lower	✓
Percentage of Days with Multiple Signals per Asset	0.2037	Lower	✗
Average Data Frequency (Days)	65.8316	Closer to 1	✗
Data Type Prevelance	1.00	Higher	✗
Likelihood of Underlying Model Change	0.64	Lower	✗

- **Completeness**
- **Authenticity**
- **Consistency**
- **Survivorship Bias**

Below is an example of a consistent model:





Data Matching Why?

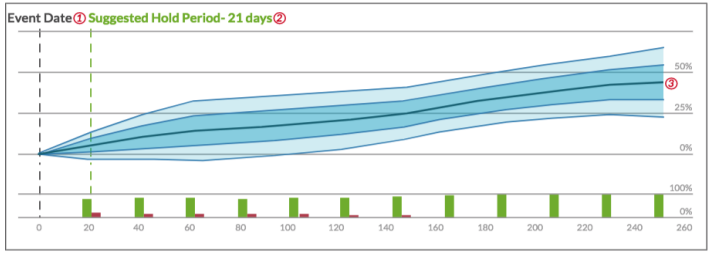
Lucena Research
Predictive Analytics

The Following Data Sources Are Recommended

Company	About	Relevancy Score
Credit Card Transaction	**Company A** provides sentiment scores that can be utilized for market/index timing transactions across time horizons.	80% (0, 20, 100)
Social Media Sentiment	Since 2000, **Company B** has been the world leader in capturing sentiment information from retail investors, owning the world's largest database of retail investor sentiment poll opinions.	70% (0, 30, 100)
GPS Location	**Company C** is a registered investment advisory firm. Over the last thirty years, their research platform, SMA/UMA strategies, indexes, and team of analysts have created more informed, conversations for advisors, investment managers and their clients.	60% (0, 40, 100)

- **Is the data effective for my constituents list?**
- **Is the signal suitable for my investment style**
  - **Hold Time**
  - **Long/Short risk tolerance**
  - **Concentration and turnover requirements**

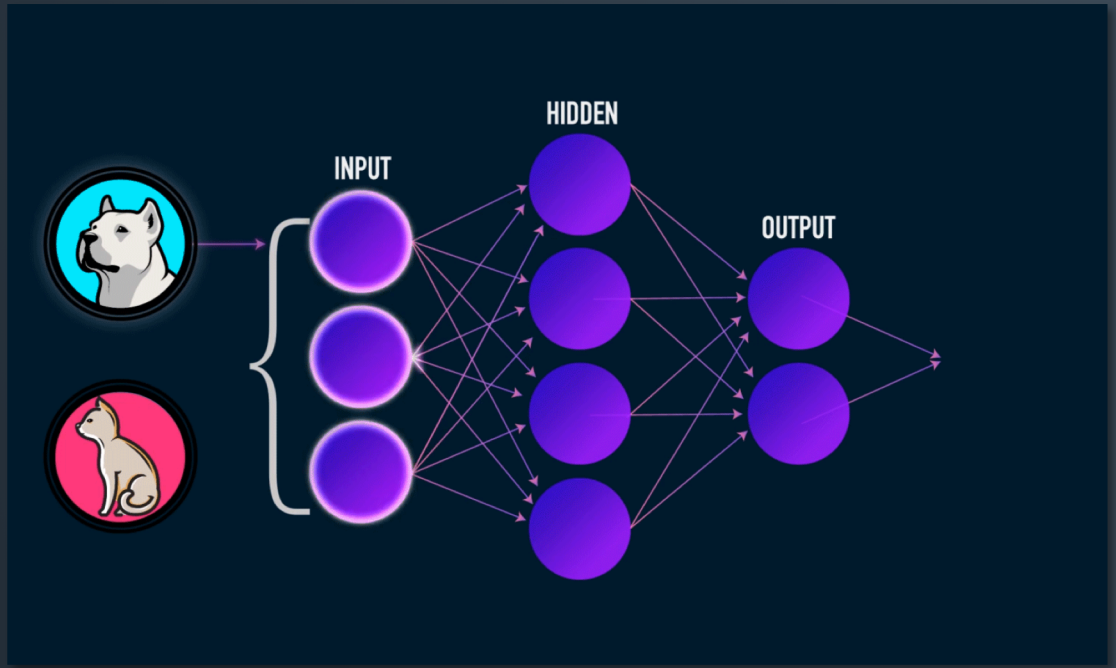
Event Study Analysis



Backtest Charts

What is a Deep Neural Network

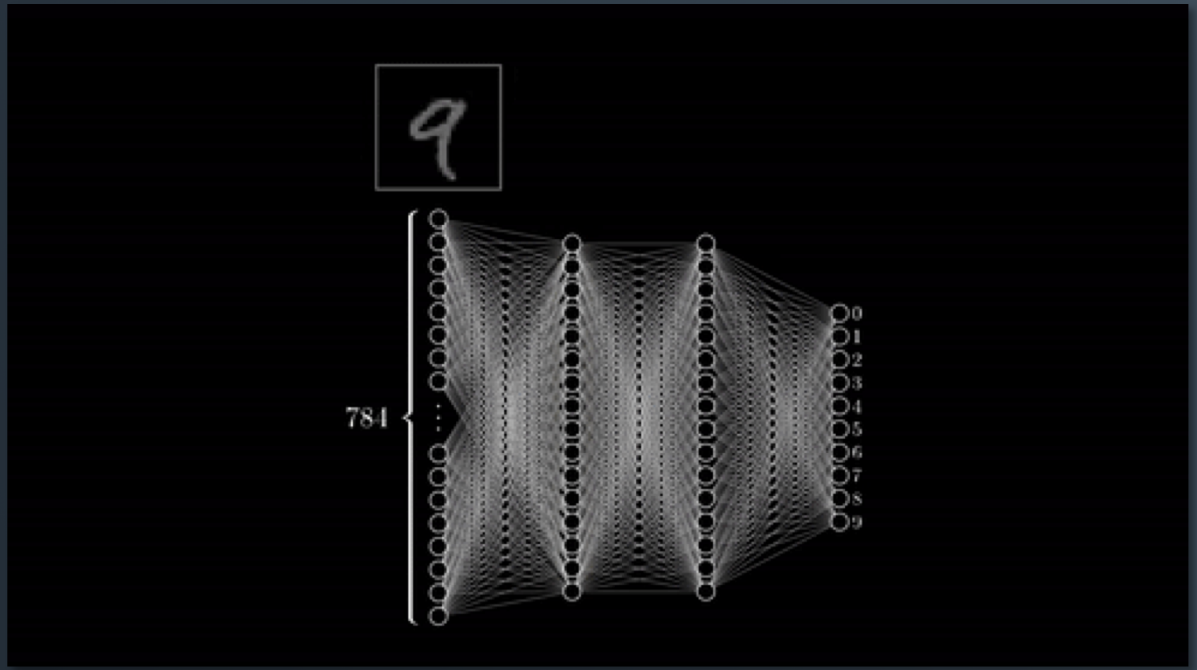
Lucena Research
Predictive Analytics



Animated gif by James Loy. Source: https://towardsdatascience.com/how-to-build-your-own-neural-network-from-scratch-in-python-68998a08e4f6

- Input Layer of Neurons.
- Multiple Hidden Layers.
- Feeds forward information.
- Output layer.

How Does The Network Learns?

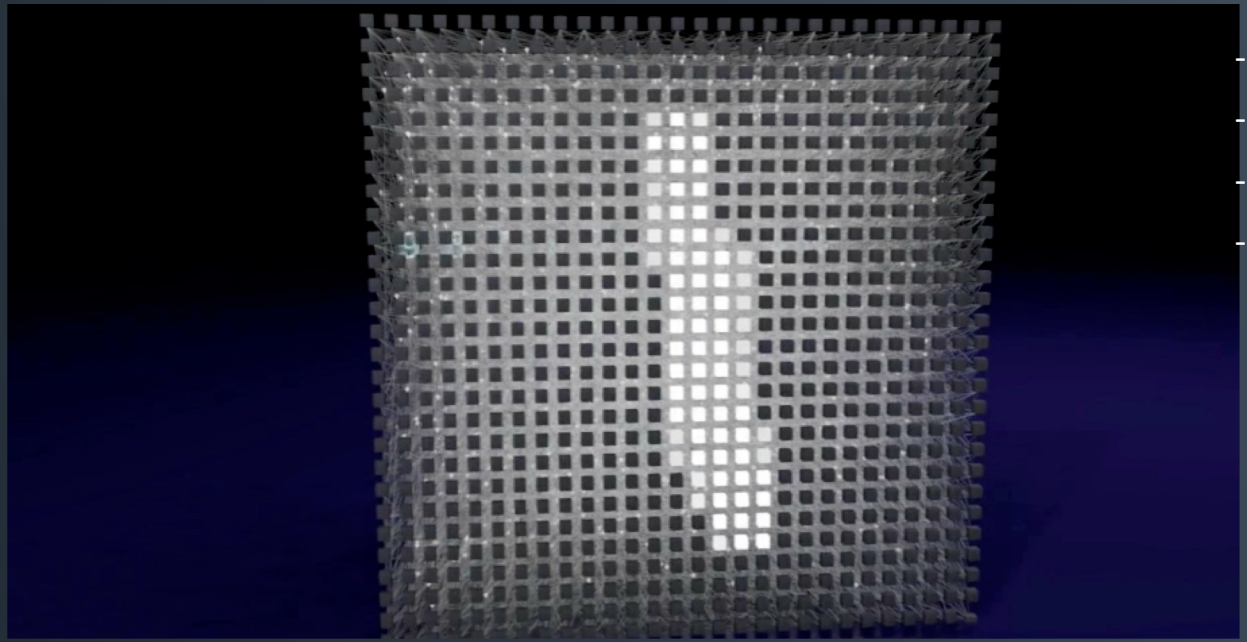


Video by 3Blue1Brown. Source: https://www.youtube.com/watch?v=aircAruvnKk

- An image of hand written digit One layer identifies edges.
- Subsequent layers identify patterns.
- Output layer identifies the full object.

Realistic Visualization of image classification

Lucena Research
Predictive Analytics



- Real world network is complex
- Feed forward Neuron Activation
- Output layer identifies the object
- Backpropagation and gradient descent

Video Source: https://www.cybercontrols.org/neuralnetworks

Example of a Modern CNN

Lucena Research
Predictive Analytics

- CNN forms additional tasks by filtering special patterns.
- Layers are added to down sample a complex image.

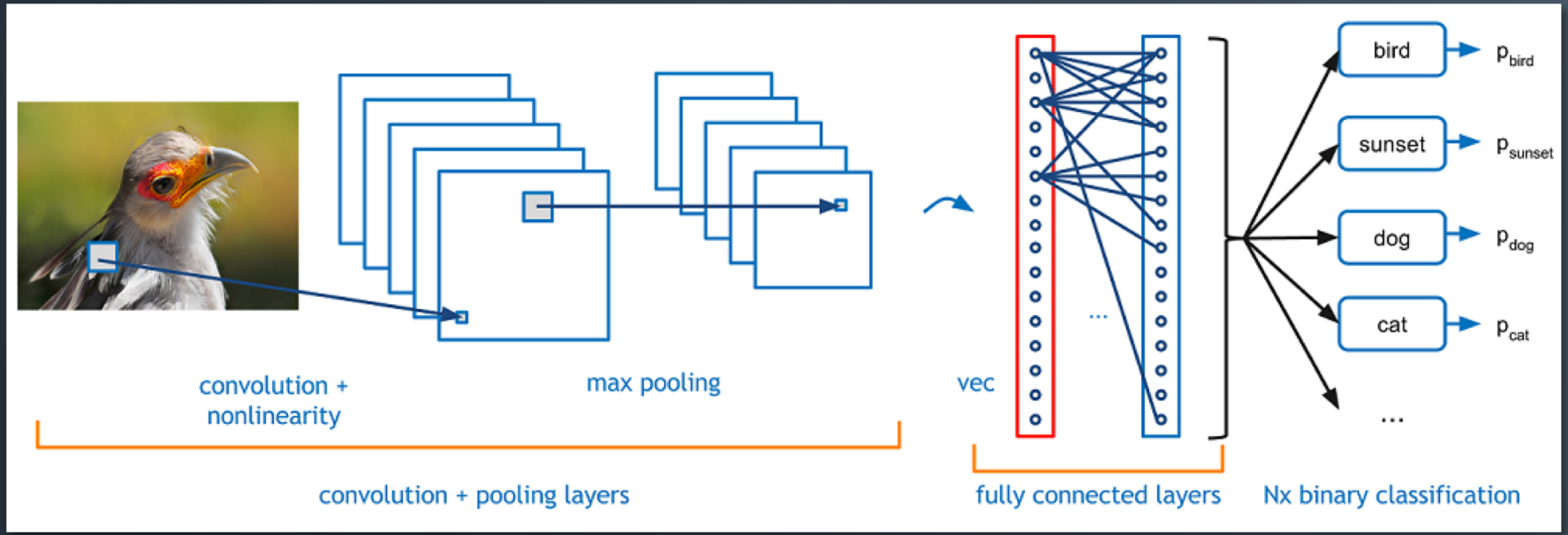
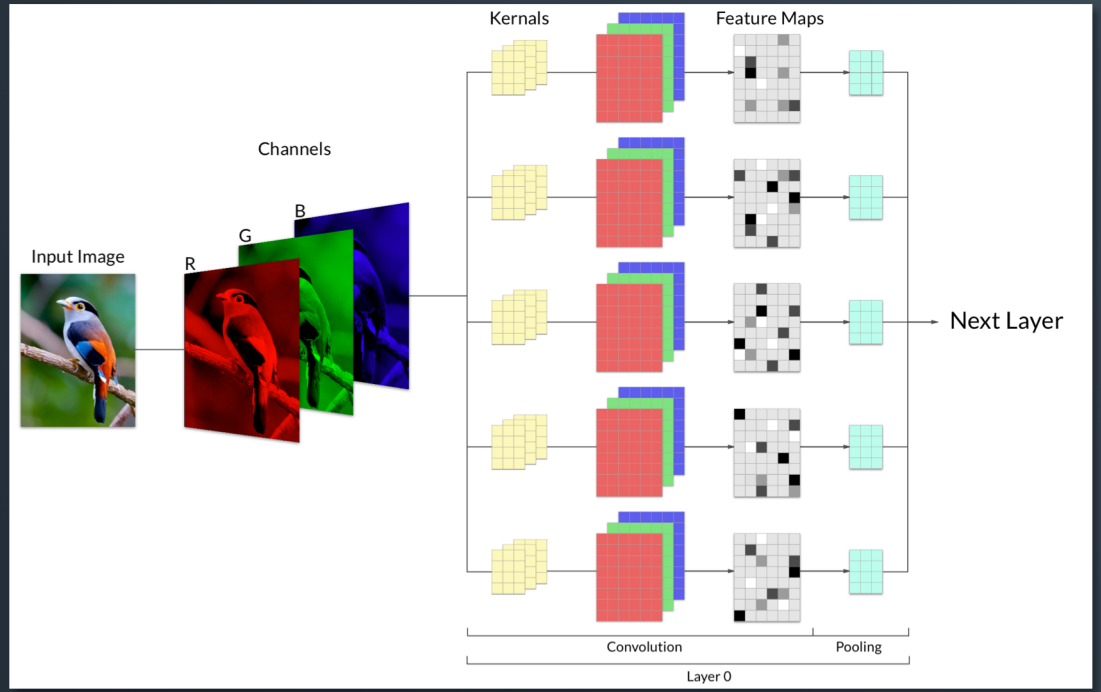


Image by Rob Hess. Source: https://code.flickr.net/2014/10/20/introducing-flickr-park-or-bird/

Multichannel Breakdown of an Image

Lucena Research
Predictive Analytics

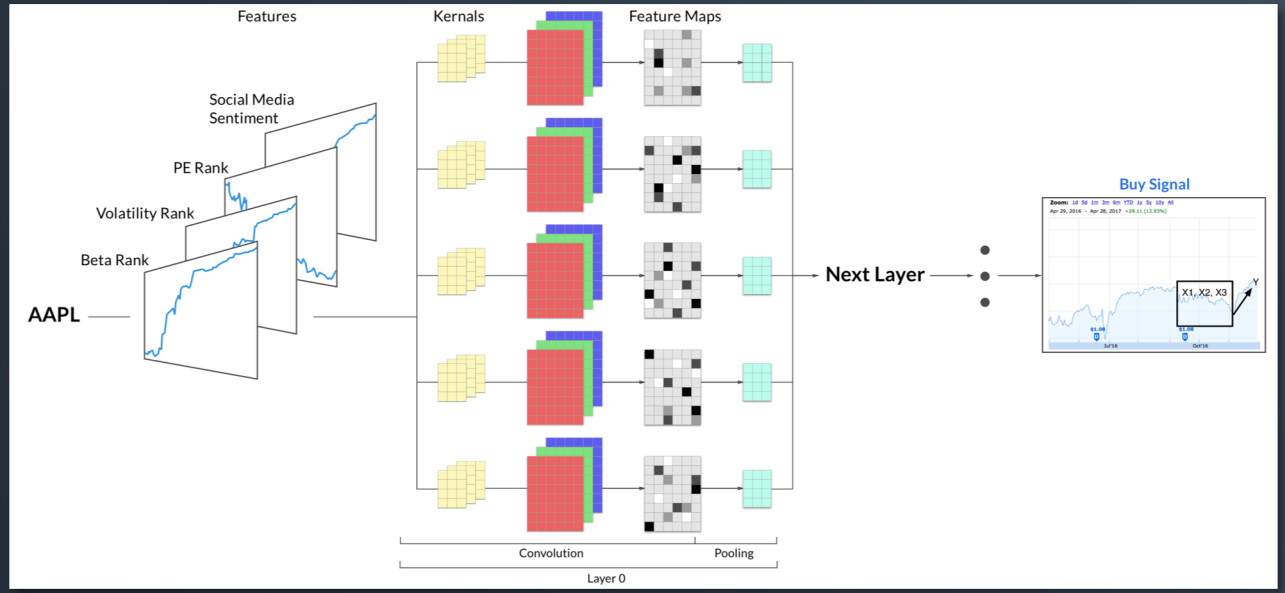


- High degree of accuracy in modern Deep Learning and CNN for image classification or object detection.
  - Autonomous Vehicles
  - Facebook Tagging
  - Amazon purchase habits
  - Netflix
  - Autocomplete Text

Can we apply the same principles to stocks?

We Can Apply the Same Concept to Stocks

Lucena Research
Predictive Analytics



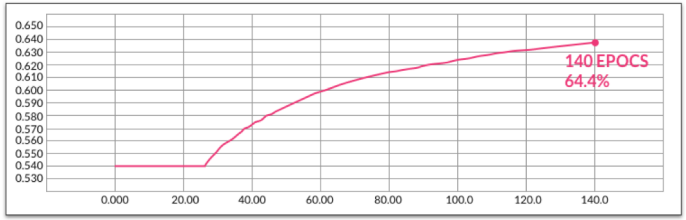
Multiple timeseries images of feature values over time.

- Technical Features
- Fundamental Features
- Macro Econ Features
- Alternative Data Features
- Social Media Sentiment
- Corporate Action
- Analysts Recommendations
- News Feed Sentiment

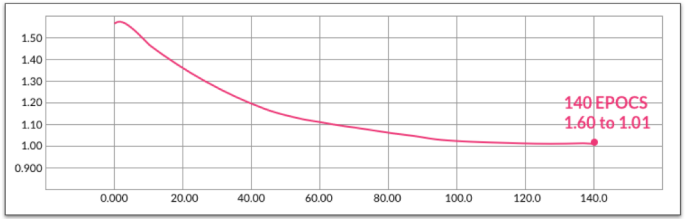
Google's Tensor Board

Lucena Research
Predictive Analytics

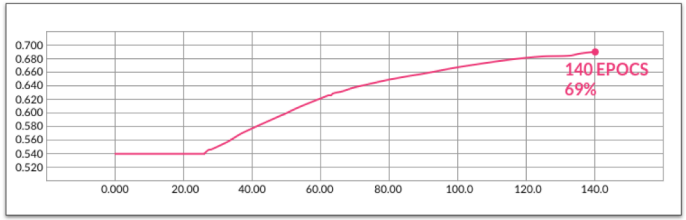
Validation Accuracy



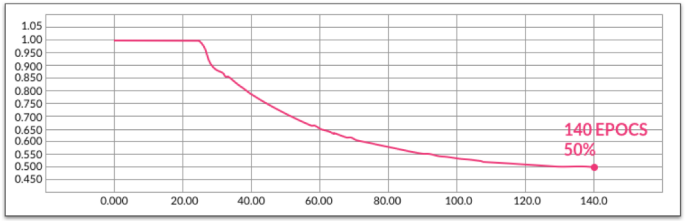
Validation Loss



Validation Precision



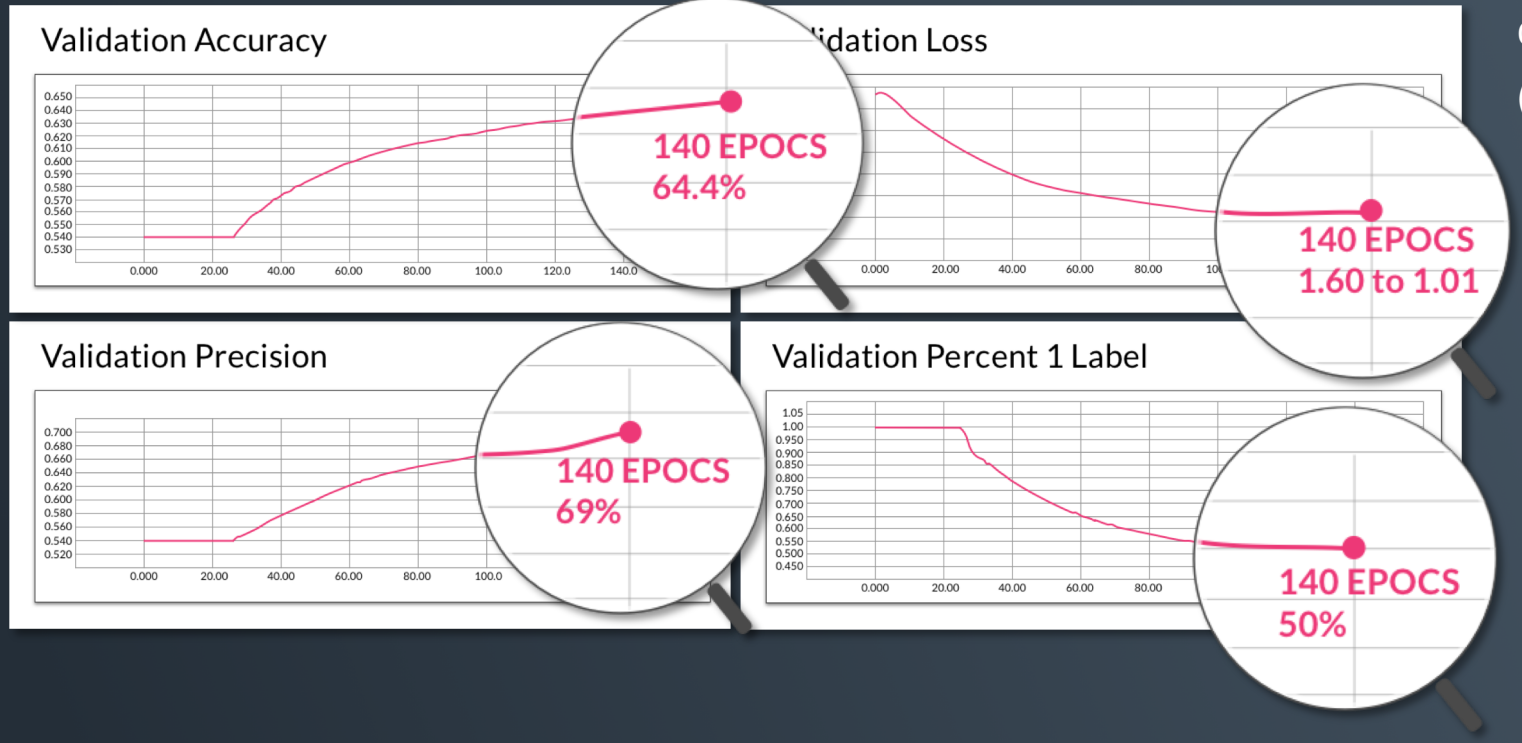
Validation Percent 1 Label



Out of Sample

Google's Tensor Board

Lucena Research
Predictive Analytics



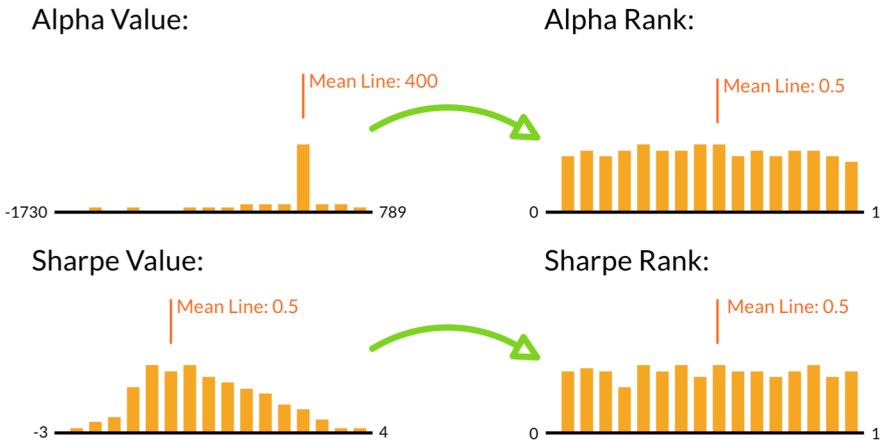
Out of Sample

(Zoom)

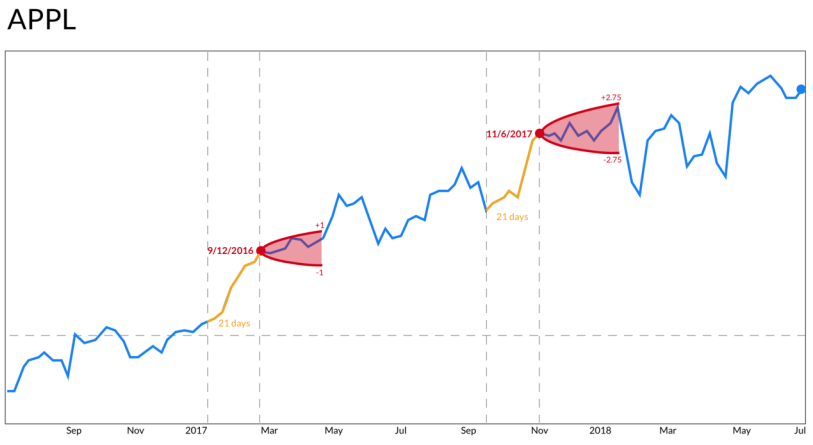
Making It Easier for the Machine to Learn

Lucena Research
Predictive Analytics

Uniformly Distributed Features



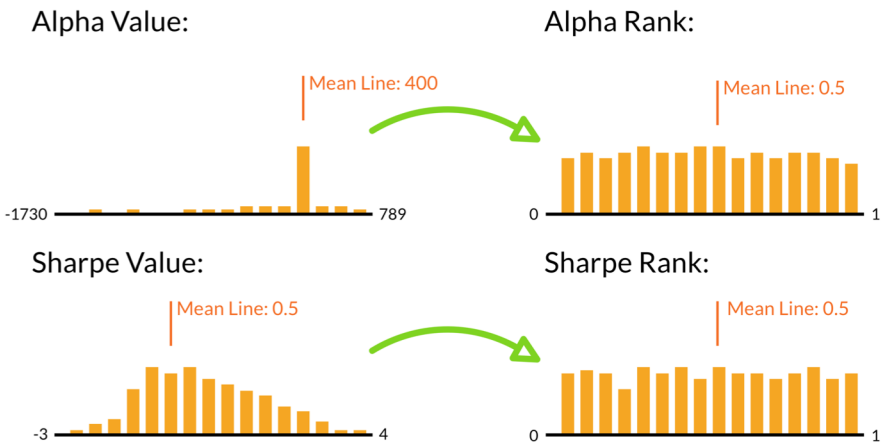
Dynamic Label Data using ATR



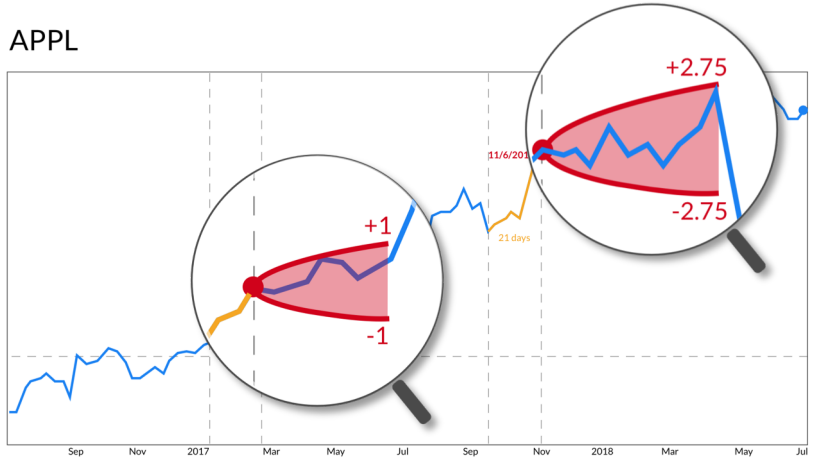
Making It Easier for the Machine to Learn

Lucena Research
Predictive Analytics

Uniformly Distributed Features



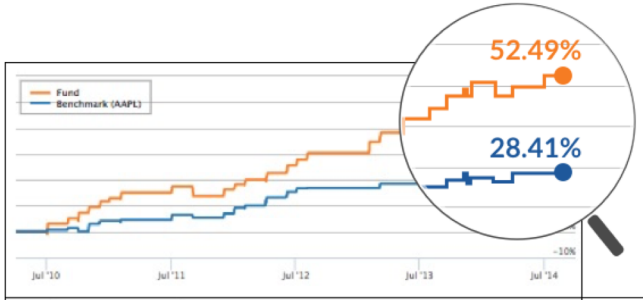
Dynamic Label Data using ATR

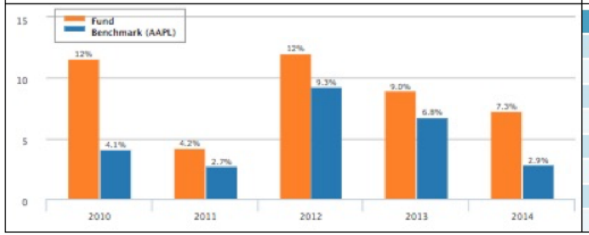


APPL Backtest

Lucena Research
Predictive Analytics

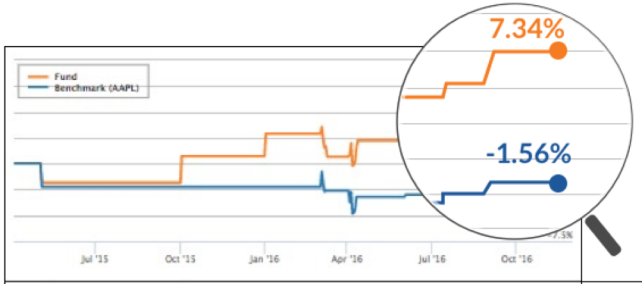
Results are Exposure Adjusted
In Sample: 4/6/2010 to 9/30/2014

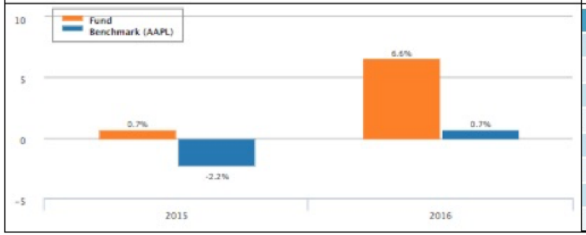




	Strategy Overall	Bench Overall
Abs. Return	52.49%	28.41%
Rel. Return	24.08%	N/A
Beta	0.04	N/A
Std. Dev	0.34%	0.24%
Sharpe	1.76	1.50
Draw Down	-4.38%	-2.22%
IR	-0.72	N/A
R²	0.04	N/A

Results are Exposure Adjusted
Out of Sample: 4/6/2015 to 12/1/2016





	Strategy Overall	Bench Overall
Abs. Return	7.34%	-1.56%
Rel. Return	8.90%	N/A
Beta	0.02	N/A
Std. Dev	0.27%	0.22%
Sharpe	1.01	-0.26
Draw Down	-3.64%	-4.84%
IR	0.30	N/A
R²	0.01	N/A

Thank you!

erez@lucenaresearch.com