



# Machine learning, security and the future of fraud

## Table of contents

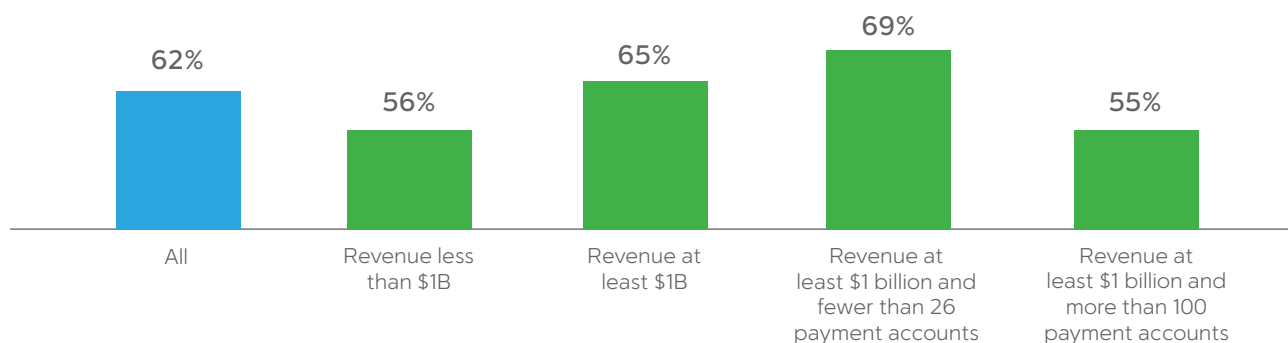
Growing need for real-time fraud identification	3
Machine learning today	4
Big data makes algorithms more accurate	5
Machine learning for fraud prevention	5
Applying machine learning	6
Machine learning engines	7
Beyond fraud prevention	8
Limitations with machine learning	8
The promise of machine learning for fraud prevention	8
First Data Fraud Detect	9

## Growing need for real-time fraud identification

Fraud attacks are getting to be more sophisticated – as technology evolves fraudsters have elevated their game on payment fraud and money laundering. With access to faster and cheaper computing, fraudsters have shifted their targets to more profitable weaker points in the financial services chain.

Sixty-five percent of organizations with annual revenues of at least \$1 billion were victims of payments fraud in 2014 compared to 56 percent of companies reporting annual revenues of less than \$1 billion.<sup>1</sup>

Percent of Organizations Subject to Attempted and/or Actual Payments Fraud in 2014<sup>1</sup>



Newer business models are constantly evolving – from instant delivery of goods to virtual cash to digital downloads. However, the growth in opportunities has led to a corresponding growth in online fraud and fraud losses particularly in eCommerce where it is seven times more difficult to prevent fraud than in the person. According to LexisNexis Fraud Multiplier, in 2016, every dollar of fraud cost merchants \$2.40 up from \$2.23 in 2015.<sup>2</sup>

The ever-faster, ever-bigger cycle of attacks leads to a number of consequences:



### Magnitudes of attacks involve multiple methods

Fraudsters are employing distributed networks, internal knowledge, big data and even machine learning to easily detect vulnerability and maximize the size of the attacks.



### Weakest links create the most exposure

Financial systems are interconnected and consist of a long value chain, a networked ecosystem of multiple entities connecting buyers and sellers. Fraud flows to the least protected components.



### Unexpected attacks can be unsettling and disruptive

Organizations can go from not having a fraud problem to being devastated in just a few days (e.g., Target, Neiman Marcus).

## CONCLUSION

Fraud solutions need to be more sophisticated to keep pace with the fraudsters and react within the short time fraud attacks happen to when they are discovered. Organizations that want to defend themselves against fraud need to have a superior, faster-learning solution that can constantly evolve yet is easy to use and maintain.

## Machine learning today

Machine learning as a data science to uncover patterns and hidden insights is not entirely a new concept. It has been in play with the use of neural networks starting in the 1980's. The question therefore is, "Why is there a big buzz around machine learning today?"

The answer lies in the fact that advancement in technology and science has enabled game-changing differences in how machine learning algorithms have evolved and being applied.

For example, traditionally, human-generated rule sets were the most prevalent approach in fraud management and still continue to be in practice today. But the quantum leap in computing power and availability of big data over the last five years has disrupted how data is being used to identify and prevent fraud. Machine learning uses artificially intelligent computer systems to autonomously learn, predict, act and explain without being explicitly programmed. Simply put, machine learning eliminates the use of preprogrammed rule sets no matter how complex.

### Machine learning enables:



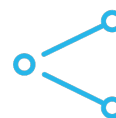
#### Real-time decisions

Advances with in-memory, event streaming technology allow risk scoring and decision making in the sub-second range (i.e., ultra-low latency).



#### Big data set processing

Advances in distributed data processing allow analyzing more data while still maintaining real-time decisions without trade-offs between data and latency.



#### Reduced cycle time

Learning cycles are continuous unlike batch learning where models become out-of-date; with machine learning, the same transactions being scored also update/teach the machine learning models.



#### Increased effectiveness

Extremely subtle patterns and variations can be detected and delivered (e.g. precision, recall) better than humans in many tasks.



#### Error-free processing

Enormous amounts of data can now be processed without human-bias or error.



#### Cost efficiencies

Address long tail "corner case" distribution.

### CONCLUSION

Application of machine learning has redefined previous strategies and tools in fraud management delivering benefits that were previously not possible with traditional methods.



## Big data algorithms more accurate

As businesses continue to evolve and migrate to the Internet and as modern money is transacted electronically in an ever-growing cashless banking economy, commerce is increasingly becoming the business of big data science. Total eCommerce sales in the United States have topped the \$400 billion mark and are projected to surpass \$1 trillion by 2023.<sup>3</sup>

Fortunately, this rapidly expanding “dataverse” also fuels modern artificial intelligence, making big data an inextricable component of today’s fraud management. Much like how IBM’s Deep Blue computer outplayed Garry Kasparov by having learned from millions of chess games, machine learning in general requires access to large amounts of data to be able to learn and generalize knowledge.

Without large amounts of data, a machine learning algorithm cannot learn. The existence of efficient algorithms to process this data very quickly opened up the possibility for sophisticated machine learning algorithms such as spam detection, efficient content recommendations, autonomous driving cars, image recognition, natural language processing, automatic translation, and of course, fraud management.



## Machine learning for fraud prevention

To understand why machine learning is important in fraud management, we need to understand the characteristics of fraud along with the associated business and technical challenges.

### Fraud’s unique characteristics:



**Fraud has a long tail distribution**  
Too many unique cases to pursue.



**Fraud is adversarial**  
Professional opponents actively working to subvert the system at the weakest points.



**Fraud patterns change quickly**  
Slow-learning countermeasures cannot keep up.



**Fraud mimics good customer behaviors**  
Good customers are penalized by over-intrusive countermeasures.

Machine learning directly addresses many business challenges that are time consuming and expensive. Manual reviews and false positives remain issues even with significant investment by large merchants – often creating more work for merchants.<sup>2</sup> Furthermore, new customer channels (e.g., mobile, social), new products and business lines present new risk vectors – fraud through remote channels is up to seven times as difficult to prevent as in-person fraud.<sup>4</sup>

Machine learning can:



Reduce manual review queues through fast iterating machine.



Be channel-agnostic.



Easily adapt to new business lines using experiential data.



Augment human decision making with increased precision.



Reduce false positives with behavior analysis.

#### CONCLUSION

Sophisticated models can reverse engineer machine logic to present human-readable language to explain model decisions.

## Applying machine learning

Machine learning models can be used to very efficiently perform analytics and deliver risk scores in real-time, with greater accuracy by leveraging large amounts of user data. Feedzai's existing model was able to detect +60% of all fraud transactions for a major retailer corresponding to +70% of their fraud money. When trained to include the retailer fraud, the model improved to detect +65% of fraud transactions and +75% of the total fraud money.

Behavior analytics build digital footprints which can then be used to learn from past data in order to make predictions on future, unseen data patterns. For example, in a retail environment, intelligence around user behavior can be used to determine their buying schema – merchandise they buy, stores they frequently visit, times they shop, channel through which they shop. Machine learning algorithms can then synthesize this data collected from multiple sources – online and offline – to baseline behavior profiles. User attributes and other data fields used by machine learning algorithms can automatically learn patterns which are then used to make predictions.

Machine learning can also be used to automatically derive outcome measurements such as a statistical risk (the measurement of the likelihood of incurring loss). The effectiveness of the statistical risk score depends on the model's ability to detect anomalies from known patterns, identify matches to known patterns and uncover new patterns.

## Machine learning engines

Mathematical algorithms power machine learning. The truth is there is not a single best algorithm that is universally better in all situations – choosing the best algorithm depends on the problem type, size, available resources, etc. Having said that, Random Forests (aka Ensemble of Decision Trees) and Deep Learning have been shown to perform very well in a number of scenarios.

Random Forests are more robust for a number of real world problems such as missing data, noise, outliers and errors. In addition, Random Forests also allow multiple types of data (numbers of different scales, text, Booleans, etc.), can scale very well, parallelize very easily, are fast to train and score, and require less effort to achieve the best results. It is no surprise that Random Forests win many machine learning competitions (as described by Kaggle.com, the world's leading machine learning competition site and data science community).

ALGORITHM	PRO	CON
<b>Random Forest, aka Ensemble of Decision Trees</b>	<ul style="list-style-type: none"> <li>• Generalizes patterns well</li> <li>• Robust to different input types (texts, numbers of scales, etc.)</li> <li>• Robust to missing data</li> <li>• Robust to outliers and errors</li> <li>• Fast to train and score</li> <li>• Trivially parallel</li> <li>• Requires less tuning</li> <li>• Probabilistic output (i.e. a score)</li> <li>• Can adjust threshold to tradeoff between precision and recall</li> <li>• Very good predictive power</li> <li>• Found to win many machine learning competitions</li> </ul>	<ul style="list-style-type: none"> <li>• Can become complex to interpret as number of decisions grow (inherent nature of increased capacity to make decisions), but better than all others, especially with Whitebox scoring to demystify decision nodes</li> <li>• Requires labeled data</li> </ul>
<b>Support Vector Machines (SVM)</b>	<ul style="list-style-type: none"> <li>• Does not require labeled data</li> <li>• Reduces feature design tasks</li> <li>• Learns multiple levels of representation (e.g. eyes, head, person)</li> <li>• Highly parallel</li> <li>• Very good predictive power, especially in text and image classification problems</li> </ul>	<ul style="list-style-type: none"> <li>• Very slow train, but benefits from recent architecture advances (e.g. GPU's, large clusters)</li> <li>• Cannot handle different input types</li> <li>• Need scaling inputs</li> <li>• Needs tuning</li> <li>• Does not provide probability estimates</li> </ul>
<b>Deep Learning</b>	<ul style="list-style-type: none"> <li>• Able to detect non-linear and complex patterns</li> <li>• Effective in very high dimensional spaces</li> <li>• Very good predictive power</li> </ul>	<ul style="list-style-type: none"> <li>• Requires labeled data</li> <li>• Cannot handle different input types</li> <li>• Need scaling inputs</li> <li>• Cannot handle missing values</li> <li>• Not scalable</li> <li>• Slow</li> <li>• Needs tuning</li> <li>• Does not provide probability estimates</li> <li>• Lack of interpretability</li> </ul>
<b>Neural Networks</b>	<ul style="list-style-type: none"> <li>• Able to represent complex patterns</li> <li>• Good predictive power</li> </ul>	<ul style="list-style-type: none"> <li>• Requires labeled data</li> <li>• Cannot handle different input types</li> <li>• Need scaling inputs</li> <li>• Cannot handle missing values</li> <li>• Not scalable</li> <li>• Slow</li> <li>• Needs tuning</li> </ul>
<b>K-Nearest Neighbors</b>	<ul style="list-style-type: none"> <li>• Robust to missing data</li> <li>• Robust to outliers</li> <li>• Good predictive power</li> </ul>	<ul style="list-style-type: none"> <li>• Requires labeled data</li> <li>• Cannot handle different input types</li> <li>• Need scaling inputs</li> <li>• Cannot handle missing values</li> <li>• Needs tuning</li> <li>• Lack of interpretability</li> </ul>

## Beyond fraud prevention

Machine learning is not just isolated to identifying and preventing fraud in online retail environment. Machine learning can also be applied wherever large amounts of data can be used to understand and infer behavior for effective decision making.

- Account opening: Validate the authenticity of users signing up online to verify and accept more applicants
- Payment authorization: Score payment requests and authorize payments in real-time
- Checkout scoring: Prevent payment chargebacks by scoring transactions during checkout
- Merchant underwriting: Protect your merchant portfolio through merchant underwriting
- Marketplace: Maintain community trust by connecting buyers and sellers



## Limitations with machine learning

One of the biggest obstacles to machine learning is the steep learning curve. Data science knowledge, plus the amount of time and data needed to create models are beyond reach of many risk teams. A steep learning curve means data scientists who do machine learning need to master many different tools such as R, Weka, Python, DBMS, NoSQL data stores, Hadoop jobs, streaming systems and more. Plus, it is very hard to evolve profiles and models to reflect the ever-changing nature of business, e.g. some companies deploy one-year old models that were trained using two-year old data.

The second biggest challenge is much of machine learning is grounded on black box decision-making. This is a serious limitation as many policy execution or governance requirements need clear explanations of decisions, e.g. explain to customer why transaction was blocked. Finally, increased capacity to process big data creates an inherent tendency towards include irrelevant data. Machines lack common sense so humans are still needed to supervise.

## The promise of machine learning for fraud prevention

While the multiple methodologies in place today to prevent fraud have been successful at keeping fraud rates low for typical payment fraud, the evolving landscape of eCommerce and mCommerce pose newer challenges. These challenges necessitate more innovative solutions that can respond and react quickly to fraud. The need for computational power to process large amounts of data and make decisions is imperative for businesses to reach quickly to fraud attacks.

Machine learning in this aspect is a promising science with potential across multiple environments. From payment fraud to abuse, machine learning can easily scale to meet the demands of big data with greater flexibility than traditional methods.



## First Data Fraud Detect

Explosive growth in eCommerce is creating corresponding growth in card not present transactions – with eCommerce transactions expecting to top 4 trillion by 2020.<sup>5</sup>

At the same time, eCommerce merchants want to grow their business by selling goods and services online. To facilitate growth, merchants are looking to increase their product offering across all channels including online, mobile and international. Additionally, merchants want to capitalize on every potential sale and minimize their overall fraud exposure to realize more revenue and profit – need a robust, integrated and intuitive fraud prevention product offering.

That's where Fraud Detect comes in.



**Fraud Detect** is a comprehensive, fraud prevention solution with real-time fraud scoring and machine learning capabilities designed to reduce a merchant's overall exposure and cost of fraud. With the First Data's global merchant footprint, Fraud Detect utilizes a machine learning technology and big data platform to prevent fraud. Additionally, First Data can access new sources of Dark Web and cyber intelligence to further detect potential fraud patterns.

### Small Medium Business Offering:

Via gateway integration, Fraud Detect is a turnkey solution with minimal custom features providing merchants some control, basic reporting and a template case management platform.

- Dynamic, real-time machine learning enable rapid response
- Confirm customer identity, reject sale if necessary and/or enable setting to automatically reject high-risk scores

### Mid-market/Enterprise Offering:

Via platform integration, Fraud Detect enables merchants who require more customization, workflow and extensive case management screens and integration with third party vendors.

- Dynamic, real-time machine learning based customizable rules engine
- Device fingerprinting and behavior analysis
- Case management with custom workflows and third party integrations
- Reporting/Analytics/Dashboard – near real-time and scenario analysis
- Batch and real-time processing capabilities
- Client services capabilities and timeline – implementation, support, consulting and training

## Sources

- <sup>1</sup> Payments Fraud and Control Survey, 2015 Association of Financial Professionals.
- <sup>2</sup> 2016 True Cost of Fraud Study, LexisNexis.
- <sup>3</sup> Card-Not-Present Fraud: The Merchant Empire Strikes Back, Mercator Advisory Group, 2016.
- <sup>4</sup> 2015 True Cost of Fraud Study, LexisNexis.
- <sup>5</sup> Worldwide Retail eCommerce Sales: The eMarketer Forecast for 2016, eMarketer, August 22, 2016.

### ABOUT FEEDZAI

Feedzai is AI. Feedzai fights fraud with the most advanced risk management platform powered by big data and artificial intelligence. The world's largest issuers, acquirers, processors and merchants across the globe use Feedzai's machine learning technology to reduce risks associated with banking and shopping, whether it's in person, online or via mobile devices.

### ABOUT FIRST DATA

First Data is a global leader in commerce-enabling technology and solutions, serving approximately six million business locations and 4,000 financial institutions in more than 100 countries around the world. The company's 24,000 owner-associates are dedicated to helping companies, from start-ups to the world's largest corporations, conduct commerce every day by securing and processing more than 2,00 transactions per second and \$1.9 trillion per year.