# Numeraire: A Cryptographic Token for Coordinating Machine Intelligence and Preventing Overfitting

**Richard Craib, Geoffrey Bradway, Xander Dunn**

**with Joey Krug**

https://numer.ai

February 20, 2017

### Abstract

Machine learning competitions are susceptible to intentional overfitting. Numerai proposes Numeraire, a new cryptographic token that can be used in a novel auction mechanism to make overfitting economically irrational. The auction mechanism leads to equilibrium bidding behavior that reveals rational data scientists' confidence in their models' ability to perform well on new data. The auction mechanism also yields natural arguments for the economic value of a Numeraire token.

## 1 Motivation

A common approach to verify accuracy in machine learning is to break the dataset into train and test sets. A trained model can be tested for accuracy on the test set, which it has never seen. However, to maintain statistical validity, this test set should only be used once. When a data scientist accesses the test set multiple times and uses that score as feedback for model selection, there's a risk of training a model that overfits the test set. This hurts the model's ability to perform well on new data.

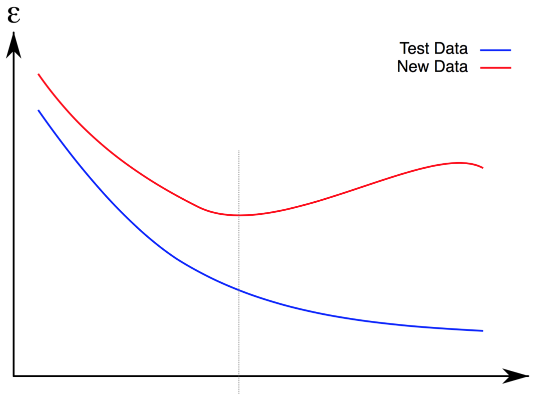

Figure 1: An overfitting curve where the test error continues to decrease with more submissions from data scientists, but the error on new data increases. [2]

This overfitting problem is called adaptive data analysis [3]. Models resulting from adaptive data analysis range from slightly degraded to completely useless [4]. For Numerai, adaptive data analysis occurs when data scientists' models have overfit historical data, at the cost of live performance. In a machine learning competition there is incentive to overfit to the historical data because performance on that data dictates winnings. Overfitting becomes intentional. What Numerai really needs is not a collection of great backtests that work well on historical data, but a collection of great models that work well on new data.

Currently, the state of the art solution to holdout reuse is to limit the amount of information exposed when using the holdout set [1]. While sufficent for scientific discovery, this solution heavily degrades user experience and rankings in machine learning tournaments.

We propose a new system for data scientists to communicate their beliefs about the quality of their models. Data scientists will compete in the new tournament by staking a new crypto-token, Numeraire (NMR), on their predictions. The auction mechanism for resolving these stakes will reward correct predictions of a model's ability to perform well on new data. With Numeraire, data scientists will now be able to express their confidence in their models' live performance. Their expressions of confidence help us to emphasize the right models and improve the performance of our hedge fund.

## 2 Cryptographic Tokens

Numeraire is an ERC20 Ethereum token [6]. Ethereum tokens are represented as smart contracts that are executed on the Ethereum blockchain. The source code to Numeraire's smart contract is publicly available[1].

All minted Numeraire are sent to Numerai. The Ethereum smart contract dictates there will never be more than 21 million Numeraire minted. Numerai will send 1 million Numeraire to data scientists based on their historical ranking on Numerai's leaderboard. After the initial distribution, the smart contract will mint a fixed number of Numeraire each week until the maximum is reached. By performing well in Numerai's machine learning competition, data scientists will earn Numeraire on an ongoing basis.

When data scientists are confident of the predictions they have made, they send Numeraire to the Numeraire Ethereum smart contract. The receiving contract will hold the data scientists' Numeraire for some holding period $t$, with $t$ sufficiently large to judge performance on new data. After $t$ has passed, Numerai will send a message to the contract with information on which data scientists' predictions performed well on new data. Those data scientists whose predictions performed well earn dollars based on the auction mechanism, and their Numeraire are returned. Those data scientists whose predictions did not perform well on new data risk having their Numeraire destroyed. The irreversible destruction of these Numeraire will be publicly verifiable on the Ethereum blockchain.

[1] https://github.com/numerai/contract

# 3 Auction

## 3.1 Overview

Every tournament has a staking prize pool, which is some fixed number of dollars. The auction mechanism allocates the prize pool among data scientists. Data scientists can submit bids to the auction. Bids are tuples $(c, s)$ where $c$ is confidence defined as the number of Numeraire the data scientist is willing to stake to win 1 dollar, and $s$ is the amount of Numeraire being staked. For some time $t$, $s$ is locked in the Ethereum contract, inaccessible to anyone, including Numerai. After $t$ has passed, a variant on the multiunit Dutch auction is used to determine the payouts.

## 3.2 Auction Mechanism

The auction mechanism is a multiunit Dutch auction with some additional rules. Performance is evaluated after time $t$. The performance evaluation metric is logloss[2], a suitable metric for binary classification problems like Numerai's machine learning competition. A model is considered to have performed well if logloss $< -\ln(0.5)$, and badly if logloss $\geq -\ln(0.5)$. The data scientists are ranked in descending order of confidence $c$. In descending order of confidence until the prize pool is depleted, data scientists are awarded $s/c$ dollars if their models performed well or they lose stake $s$ if they perform badly. Once the prize pool is depleted, data scientists no longer earn dollars or lose their stakes.

## 3.3 Example

Assume a prize pool of 3000 dollars, and that time $t$ has elapsed. Assume the staking auction ended as follows:

| Confidence $c$ | Stake $s$ | $s/c$ | Logloss $< -\ln(0.5)$ | Data Scientist |
|---|---|---|---|---|
| 5 | 10000 | 2000 | NO | WSW |
| 4 | 2000 | 500 | YES | XIRAX |
| 1.5 | 3000 | 2000 | YES | PHIL_CULLITON |
| 1 | 5000 | 5000 | NO | DAENRIS |
| 0.5 | 300 | 600 | YES | ABRIOSI |

WSW didn't achieve logloss $< -\ln(0.5)$, so his 10,000 Numeraire are destroyed. XIRAX receives \$500 and his Numeraire are returned. PHIL_CULLITON receives \$2000 and his Numeraire are returned. DAENRIS' Numeraire are destroyed. ABRIOSI receives \$500, \$100 less than his bid because the prize pool is exhausted. Everyone below ABRIOSI will have the Numeraire returned and receive zero dollars.

[2] https://www.kaggle.com/wiki/LogarithmicLoss

# 4 Analysis of Staking

Let $p$ be the probability that the model achieves logloss $< -\ln(0.5)$ on new, unseen data. A low $p$ would imply a high probability that a model is overfit. Let $s$ be a data scientist's total Numeraire staked. Let $e$ be the exchange rate of Numeraire per dollar. $c$ is the confidence. A data scientist will stake Numeraire if the expected value of staking Numeraire is positive. If a data scientist stakes $s$ and achieves logloss $\geq -\ln(0.5)$, the data scientist loses $\frac{s}{e}$ dollars. If a data scientist stakes $s$ and achieves logloss $< -\ln(0.5)$, the data scientist wins $\frac{s}{c}$ dollars. Therefore, the expected value in dollars of staking $s$ with confidence $c$ is

$$E(c, s) = p\frac{s}{c} - (1-p)\frac{s}{e}$$

A data scientist will stake if

$$E(c, s) \geq 0$$

$$p\frac{s}{c} - (1-p)\frac{s}{e} \geq 0$$

This implies

$$p \geq \frac{c}{c+e}$$

This results in self-revelation: Data scientists are moved to reveal their true inner values. Solely in the interest of maximizing winnings, data scientists reveal their knowledge of their models' abilities to generalize to new, unseen data. As we let these tournaments repeat, we expect to see bidding behaviors that accurately reflect $p$, since overbidding and underbidding are both nonoptimal behaviors and the accuracy of estimating $p$ increases with time.

Since having a higher confidence produces greater incentive to participate in an auction, we can make the following observations:

- The higher $p$, the higher $c$ a data scientist will submit, and the more dollars the data scientist can win from the auction.
- For a fixed $p$, a confidence that is too high produces $E(c, s) < 0$, which will deter this strategy.
- Models that perform well on historical data but fail to generalize (low $p$) will either have logloss $< -\ln(0.5)$ or have $E(c, s) < 0$.
- Because Numeraire can be used by data scientists to earn dollars, the exchange rate $e > 0$.
- Numeraire is worth more to data scientists with large $p$ because they can use it to earn dollars with higher confidence.
- A data scientist with $p = 1$ has an expected value in dollars $E(c, s) = \frac{s}{c}$. To this data scientist, the value of all Numeraire is the net present value of all future stake payouts by Numerai.

The purpose of this auction is to get accurate probability estimates, not to maximize Numeraire staked. The auction need not be revenue maximizing, but self-revelation is important. While a weakly dominant strategy in second priced auctions is to bid truthfully, second priced auctions are more susceptible to collusion and first priced auctions are more robust to this [5]. For this reason, and for simplicity, we use a Dutch auction (first priced) rather than an Ausubel auction.

## References

[1] Dwork, Feldman, Hardt, Pitassi, Reingold, Roth. Generalization in Adaptive Data Analysis and Holdout Reuse. http://papers.nips.cc/paper/5993-generalization-in-adaptive-data-analysis-and-holdout-reuse.pdf.

[2] Gringer. Distributed under a CC BY 3.0 License. https://creativecommons.org/licenses/by/3.0/deed.en.

[3] Hardt. Adaptive data analysis. http://blog.mrtz.org/2015/12/14/adaptive-data-analysis.html.

[4] Hardt. Competing in a data science contest without reading the data. http://blog.mrtz.org/2015/03/09/competition.html.

[5] Krishna. *Auction Theory*. Elsevier, Massachusetts, 2010.

[6] Wood. Ethereum: A Secure Decentralized Generalised Transaction Ledger. http://gavwood.com/paper.pdf.