What do humans perceive in asset returns?*

Jasmina Hasanhodzic Andrew W. Lo Emanuele Viola

August 5, 2018

Abstract

In this article, the authors run experiments to test if and how human subjects can differentiate time series of actual asset returns from time series that are generated "synthetically" via various processes, including AR1. In contrast with previous anecdotal evidence, they find that subjects can distinguish between the two. These results show that temporal charts of asset prices convey to investors information that cannot be reproduced by summary statistics. They also provide a first refutation based on human perception of a strong form of the efficient-market hypothesis. Their experiments are implemented via an online video game (`http://arora.ccs.neu.edu`). The authors also link the subjects' performance to statistical properties of the data, and investigate whether subjects improve their performance while playing.

*We would like to thank Zvi Bodie for pointing out to us Hung, Heinberg, and Yoong (2010), Michael Coen for helpful discussions on a preliminary version of this work, especially on p-values, and Jason Chen, François Gourio, and Jawwad Noor for helpful feedback and discussions. We are also grateful to Hamid Jahanjou for his help with Amazon Mechanical Turk.

Jasmina Hasanhodzic is assistant professor of finance at Babson College in Wellesley, MA. **jhasanhodzic1@babson.edu**
Andrew W. Lo is Charles E. and Susan T. Harris professor at the MIT Sloan School of Management and principal investigator at the MIT Computer Science and Artificial Intelligence Laboratory in Cambridge, MA. **alo@mit.edu**
Emanuele Viola is associate professor of computer science at Northeastern University in Boston, MA. He was partially supported by a grant from MIT and NSF grants CCF-0845003 and CCF-1319206. **viola@ccs.neu.edu**

One of the most important and complex decisions individuals face is how to save and invest. Choices they make affect not only their own quality of life, but may have an impact on the economy by creating dependencies on government-sponsored benefits. However, it is well noted that when it comes to investing, individuals are not well positioned to make sound decisions. Several reasons have been proposed in the literature, including overload of information about investment products to choose from, marketing strategies designed to mislead, behavioral biases, and financial illiteracy; see, for example, Bazerman (2001), Bodie (2007), Choi, Laibson, and Madrian (2010), and the references therein. The problem of inadequate individual investment decisions is especially acute in the case of retirement savings, where the recent shift from defined benefit pension plans to privatized 401(k) plans has forced individuals to, in effect, manage their own money. As a result, much debate among policymakers and academics has taken place about improving the quality and presentation of data available to investors. For example, Bazerman (2001) and Kozup et al. (2008) call for research on investors' perceptions of investment products and ways of making the information about those products easy to access and comprehend.

An example of a work in this direction is Hung, Heinberg, and Yoong (2010), who evaluate versions of the Department of Labor's proposed Model Comparative Chart, which provides a standard simplified disclosure format for investment information. They conduct an online experiment where subjects are asked to allocate $10,000 among different funds based on funds' performance disclosure. In one version of the disclosure, past returns are presented as a numerical table. In another version, in addition to the numerical table, the disclosure shows a graphical representation of returns over a 10-year period, as a bar chart. Perhaps surprisingly, the authors find that the two disclosures have a statistically significant effect on the retirement investment allocation, although the effect may not be practically significant in terms of investment outcomes.

Together with the prevalence of temporal charts of asset returns in financial media such as Yahoo! Finance and their widespread use by both casual and professional investors, the above brings to the forefront a fundamental question: Just what information can human beings extract from charts of financial returns? This question has several ramifications. For example: Are there any patterns in financial asset returns that humans can actually extract by looking at such charts? Is seeing a chart more informative than just having a few parameters like, say, average and variance? Could Yahoo! and numerous other websites that display charts save space by getting rid of them altogether, with no harm to investors? In Hung, Heinberg, and Yoong's (2010) experiment, is the mere *presence* of some chart biasing the subjects, or are subjects actually gathering information from the contents of the chart?

In this paper we report the results of several experiments designed to test if and how human subjects can differentiate time series of actual asset returns from time series that are generated "synthetically" via various processes. Specifically, we consider time series obtained by permuting at random the samples of actual returns, as well as those arising from first-order autoregressive (AR1) models. Our experiments are implemented via an online video game (`http://arora.ccs.neu.edu`).

The main finding of this paper is that humans can distinguish actual time series from synthetic ones. The results related to random permutations indicate that subjects perceive the temporal order of financial data. The results related to AR1 indicate that subjects are employing more than just first-order autocovariance to differentiate the two time series. We also link the subjects' performance to other statistical properties of the data, and investigate whether subjects improve their performance while playing. For some contests, our results indicate that subjects do improve.

Our findings are in contrast with previous anecdotal evidence. Specifically, it was argued that humans cannot tell price charts from "random," such as charts generated by a random walk. For example, in an experiment (Malkiel 1973, p. 143) students were asked to generate returns (i.e., price differences) by tossing fair coins, and it was argued that those yielded observations that were indistinguishable from market returns to human subjects observing corresponding price charts. For similar arguments in the finance literature see, for example, Roberts (1959), Kroll, Levy, and Rapoport (1988), DeBondt (1993), Wärneryd (2001), and Swedroe (2005). Such anecdotal evidence has also been collected in the computer science literature. For example, Keogh and Kasetty (2003) asked 12 professors at UCR's Anderson Graduate School of Management to look at six time series and determine which three series are random walk, and which three are real S&P500 stocks. They find that "the accuracy of the humans was 55.6%, which does not differ significantly from random guessing."

Our results are also of interest in light of the Efficient Market Hypothesis, according to which "prices fully reflect all available information" and hence must be unforecastable; see, for example, Samuelson (1965), Fama (1965a), Fama (1965b), and Fama (1970). A strong form of this hypothesis presumes asset returns to be independent and identically distributed, see, e.g., Fama (1970). In this case, it would be impossible to distinguish actual asset returns from a random permutation of them. But, again, we show that humans can do that.

Note that works such as Lo and MacKinlay (1988, 1999) and Lo, Mamayski, and Wang (2000), provide compelling evidence that markets are not efficient, i.e. price data does possess statistical properties that noticeably deviate from random models. In fact, they show that autocorrelation is such a property. However, we point out that the data analysis in all of these works is *computer*, not *human*-based. Consequently, the works leave open the question of whether markets look efficient *to human beings*. Our work appears to be the first to provide such an answer.

We note that the idea of testing the ability of human subjects to distinguish random vs. real data using graphical representations is not new. Indeed, this has been studied in depth in the Information Visualization literature; see for example the works by Heer, Kong, and Agrawala (2009), and Wickham, Cook, Hofmann, and Buja (2010), and the references therein. However, we are unaware of any previous work where this idea has been used in a financial setting.

Similarly, we do not view the video game we developed as a main contribution of this paper. This game displays data in a fashion similar to commonly used trading platforms; and similar tools are for example reviewed in the Information Visualization papers just cited. Instead, implementing the experiment as a video game is intended to make the process fun and engaging for the subjects, so that they do not get tired, bored, or frustrated in a way that might affect their behavior. Moreover, the game allows the subjects to make their choices quickly, allowing us to get a large amount of data efficiently, with as little cost to subjects as possible.

Experimental Design

We develop a simple web-based video-game, available at

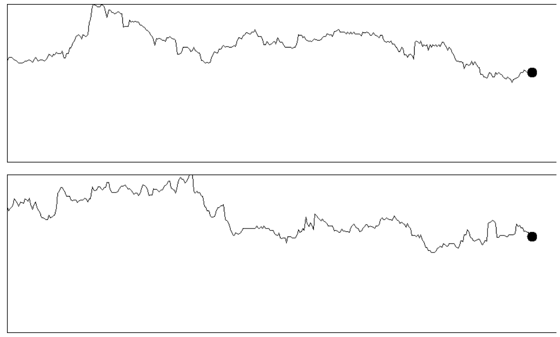

Exhibit 1: Reindeer (real data in top panel).

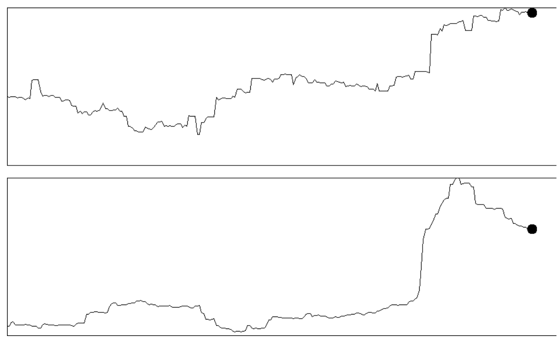

Exhibit 2: Bear (real data in bottom panel).

Exhibit 3: Wrong choice in Beaver contest.

Exhibit 4: Correct choice in Elk contest.

http://arora.ccs.neu.edu. In this game, subjects are shown two dynamic price series (i.e., moving charts) side by side—both of which display price graphs evolving in real time (a new price is realized roughly each second)—but only one of which is a "replay" of actual historical price series. The other series is constructed via a synthetic process. See Exhibits 1 and 2, which are snapshots from our game. Subjects are asked to press a button indicating their selection of the actual price series, and are informed immediately whether they were correct or incorrect, see Exhibits 3 and 4, after which the next pair of price series begins being displayed. Note that the charts are moving, so at any point in time there is a certain number of observations present on the screen for each time series, which is a subset of the total number of observations subjects see on a moving chart before having to make a guess (these parameters are reported for each data set later in the paper). Subjects do not have to wait until the entire moving chart is completed being displayed before making their choice, but can guess at any time prior to its completion (an omnipresent counter informs them of the time left). They have a counter telling them how many seconds they have remaining before the moving chart is done. The game is fast-paced: subjects can observe the charts for 10 to 25 seconds (depending on the data set) before having to make a guess.

For the actual time series we used eight data sets consisting of returns of commonly traded financial assets. These data sets were arbitrarily named after animals, so that users had no knowledge of the specific financial assets used in the experiment. Exhibit 5 summarizes the data used. It also reports how many charts were shown to each subject, how many data points constitute a chart, and, since charts are moving, how many points of the chart fit onto the screen at any given time. Subjects were given 11 seconds to guess in the Bull contest, 15 in Bear, Elk, and Raindeer, 20 in Lynx and Mandrill, 22 in Seal, and 24 in Beaver. The Dow Jones Corporate Bond Price Index was obtained from the Global Financial Database, while all other data series were obtained from Bloomberg. Several statistics of the data sets are presented later. We use them to shed light on the difference in performance between the random permutation and AR1 experiments.

Datasets

Contest	Data	Points on Screen	Charts per Subject	Points per Chart
Mandrill	S&P GSCI Corn Index Spot daily data Jun 1982 - Oct 2009	38	50	125
Bear	Nasdaq Composite Index tick data (about 1 sec.) May - Jul 2009	250	37	400
Lynx	Canada/U.S. Foreign Exchange Rate daily data Aug 1978 - Apr 2009	90	35	190
Reindeer	Gold Spot Price tick data (1-60 sec.) Jun - Oct 2009	350	40	500
Beaver	Dow Jones Industrial Average daily data Sep 1926 - May 2009	300	36	500
Bull	Russell 2000 Index tick data (about 10sec.) May - Dec 2009	110	31	153
Seal	Dow Jones Corporate Bond Price Index daily data Jan 1941 - Apr 2009	250	39	400

Exhibit 5: The correspondence between contest names and data sets and parameters used in the presentation of data to the subjects during the game.

Subjects were recruited via Amazon Mechanical Turk.[1] After registration,[2] a subject can participate in eight different contests, each consisting of the same game applied to different data sets. Participating in a contest consists of the following task. The subject is shown two dynamic price charts on a computer screen, one above the other (Exhibits 1 and 2). Each graph evolves through time—similar to those appearing in computer trading platforms—plotting the price at that point in time as well as the trailing prices over a fixed time window over the most recent past. Of the two moving charts, only one corresponds to the sequence of market prices from the actual data set; we call this graph the "real" chart. The other corresponds to a "synthetic" sequence of prices. We call this graph the "synthetic chart". The computer chooses at random which of the two graphs is placed at the top or the bottom.

The subject is asked to decide which of the two moving charts is the real one by clicking on it. The game registers the subject's choice, and informs the subject immediately whether his/her guess is correct or incorrect, see Exhibits 3 and 4. For each data set, the user is shown approximately 35 pairs of moving charts and asked to make as many choices. The subject is also free to refrain from choosing. This happened rarely, and to err on the conservative side, we recorded the absence of a guess as an incorrect choice for that trial. To provide the participants with some incentive for making correct choices, we paid each participant as follows. We counted the number v of correct guesses made by the participant minus the number of wrong guesses. If v was larger than 0 then we paid v dimes.

To evaluate the robustness of our experimental design, we varied various parameters of the experiment across data sets, as indicated in Exhibit 5. In addition, we presented subjects with data charts in two different ways. For half of the data sets corresponding to transaction-by-transaction (or "tick") data, each subject was shown a fresh set of charts, based on a sequence of returns disjoint from the sequences shown to any other subjects. For the other half of the data, corresponding to

[1] The advertisement read: "The game you are about to play is part of a research project which studies how humans see random data. The game has been designed to be fun to play: we are going to show you pairs of charts; in each pair, one chart is based on real data (such as price fluctuations) and one is randomly generated. You are required to indicate which one you think is real by clicking on it."

[2] In the first iterations, we had subjects fill a short demographic questionnaire which included a question about financial literacy. We found no correlation between their answers and their performance in our experiment, so we discontinued the questionnaire.

daily data, the charts shown to each subject were based on the same sequence of returns.[3]

Finally, for each data set, before entering the contest subjects were required to train on a disjoint set of data.

Synthetic processes and results

In this section we describe the various synthetic processes we considered and the corresponding results. In each case we begin with a time series of actual historical prices $\{p_0, p_1, p_2, \ldots, p_T\}$ and generate from it a synthetic series $\{p^*_0, p^*_1, p^*_2, \ldots, p^*_T\}$. When displayed during the game, each series is scaled so that its maximum and minimum lie on the borders of the window on the computer screen.

Random permutation

Here we want to test the null hypothesis H that human subjects cannot distinguish between actual time series and a time series that is obtained by permuting at random the entries of the actual one. Details follow.

We begin with a time series of actual historical prices $\{p_0, p_1, p_2, \ldots, p_T\}$ and compute the logarithmic returns $\{r_t\}$,

$$r_t \equiv \log(p_t) - \log(p_{t-1}). \tag{1}$$

From this, we construct a randomly generated price series $\{p^*_0, p^*_1, \ldots, p^*_T\}$ by cumulating randomly permuted returns:

$$\begin{aligned} p^*_{t+1} &\equiv p^*_t \cdot e^{r_{\pi(t+1)}} \quad , \quad p^*_0 \equiv 1 \quad , \\ \pi(k) &: \{1, \ldots, T\} \rightarrow \{1, \ldots, T\} \end{aligned}$$

where $\pi(k)$ is a uniform permutation of the set of time indexes $\{1, \ldots, T\}$. A random permutation

[3] However, the data was shifted by a random amount for security reasons, i.e., to avoid the possibility that two subjects could coordinate their guesses, for example by simultaneously playing the same charts on two nearby machines.

of the actual returns does not alter the marginal distribution of the returns, but it does destroy the time-series structure of the original series, including any temporal patterns contained in the data. Therefore, the randomly permuted returns will have the same mean, standard deviation, and moments of higher order as the actual return series, but will not contain any time-series patterns that may be used for prediction. This construction will allow us to test specifically for the ability of human subjects to detect temporal dependencies in financial data.

The results are reported in Exhibit 6. In particular, for each contest we report the p-value of the two-sided t-test of the null hypothesis, according to which the average across subjects of the number of their correct guesses equals the total number of guesses in the contest divided by 2.[4] We also report the correct guesses per subject as percentage of total guesses. The table shows that the null hypothesis is refuted for all eight data sets: p-value is always less than 1%.

A variant

To evaluate the robustness of the results we also considered the following variant of the process, were returns are simply obtained via price differences:

$$r_t \;\equiv\; p_t \;-\; p_{t-1}.$$

From this, we construct a randomly generated price series $\{p_0^*, p_1^*, \ldots, p_T^*\}$ by cumulating randomly permuted returns:

$$p_t^* \;\equiv\; \sum_{k=1}^{t} r_{\pi(k)} \quad , \quad p_0^* \;\equiv\; p_0 \quad ,$$
$$\pi(k) \;:\; \{1, \ldots, T\} \;\rightarrow\; \{1, \ldots, T\}.$$

For this variant we also changed the recruitment and incentive mechanisms. To recruit subjects, an announcement was emailed to Northeastern computer science students, MIT Sloan MBA students in the Fall section of 15.970, members of the American Association of Individual Investors

[4]We use the same test for other synthetic processes considered below.

Results for Distinguishing Price Charts from Their Permutation

Contest	Subjects	p-Value	Correct Guesses per Subject As Percentage of Total Guesses
Mandrill	56	0.00972	50 70 48 54 58 48 64 46 34 54 38 40 50 48 50 60 56 50 54 64 62 40 48 64 46 62 48 46 68 56 52 56 62 56 52 38 54 58 36 58 64 40 62 46 54 56 68 46 52 50 54 66 60 42 56 60
Bear	55	0.00000	78 81 86 73 76 65 76 92 78 43 57 76 76 78 89 46 95 86 54 57 76 46 59 73 81 86 65 86 73 84 46 84 62 78 46 81 86 59 68 65 89 81 92 81 43 92 97 65 51 49 43 97 51 57 68
Lynx	56	0.00045	31 57 60 49 60 57 60 57 54 49 60 43 49 54 66 51 49 60 49 49 46 51 49 69 54 51 60 49 54 63 57 54 54 46 60 51 57 69 54 29 54 66 60 54 63 57 66 37 57 54 66 51 49 46 49 57
Reindeer	56	0.00000	53 63 75 93 95 45 65 85 63 85 78 38 63 65 68 35 48 53 70 45 63 55 73 78 60 70 78 48 60 70 93 70 50 70 45 63 73 65 63 60 80 75 53 60 60 73 43 43 63 43 58 95 70 53 60 55
Beaver	58	0.00013	58 58 64 53 72 47 67 53 42 44 42 61 42 64 56 53 36 61 42 67 64 53 56 44 56 72 53 72 50 72 50 56 53 64 56 47 50 47 56 75 44 50 75 78 75 53 58 61 44 50 50 56 69 39 58 39 69 42
Bull	57	0.00000	81 97 84 97 100 81 84 100 65 61 90 100 90 84 100 100 100 74 39 55 97 48 94 100 29 52 97 61 55 100 97 97 100 100 100 81 97 100 90 71 100 100 94 97 100 68 100 97 35 52 39 61 90 74 29 100 97
Elk	58	0.00000	85 73 68 70 68 68 60 83 48 90 78 83 80 53 85 45 95 40 35 83 30 80 93 48 75 80 70 45 73 60 83 65 83 83 88 45 83 95 100 93 75 78 90 95 78 83 80 80 60 53 50 55 93 73 68 35 85 83
Seal	55	0.00000	77 72 85 54 82 64 64 85 85 62 64 79 38 85 92 59 54 62 77 38 44 85 85 64 87 87 46 74 72 54 79 85 82 69 90 90 85 67 82 74 85 79 79 64 85 85 69 51 51 56 85 64 59 72 90

Exhibit 6: Results for distinguishing price charts from their permutation. For each contest, the p-value is for a two-sided t-test of the null hypothesis that the average across subjects of the number of their correct guesses equals the total number of guesses in the contest divided by 2.

mailing list, Market Technicians Association mailing list, the MTA Educational Foundation mailing list, and the staff and Twitter followers of TraderPsyches. As an incentive, we offered a $100 Amazon gift certificate to the top scorer in each contest.

Results for Distinguishing Price Charts from Their Permutation – Variant

Contest	Subjects	p-Value	Correct Guesses per Subject As Percentage of Total Guesses
Mandrill	17	0.05770	38 42 46 46 48 50 50 52 54 58 58 60 60 64 64 66 70
Bear	29	0.00000	100 70 78 100 92 16 89 86 95 78 100 86 97 84 81 95 100 81 70 73 54 73 73 92 78 89 89 51 57
Lynx	26	0.00156	43 43 46 46 46 46 49 49 51 51 54 54 57 57 57 60 60 63 63 63 63 63 63 63 66 71
Reindeer	22	0.00000	100 40 68 95 75 73 80 68 78 95 78 63 75 68 75 53 63 73 78 85 53 38
Beaver	23	0.00332	33 36 42 50 50 53 53 53 53 56 56 58 64 64 64 64 64 64 67 67 69 81 83
Bull	32	0.00000	100 94 94 100 81 100 97 97 84 94 97 90 100 100 94 97 100 100 100 84 97 97 100 94 97 100 94 100 100 100 97 45
Elk	25	0.00000	100 45 78 88 90 88 55 75 93 90 83 100 83 90 90 88 100 90 95 90 88 93 100 85 53
Seal	38	0.00000	41 44 46 49 54 59 59 59 62 62 64 64 64 67 67 69 69 72 74 74 74 77 77 79 79 79 79 82 82 82 82 85 85 87 90 90 92 100

Exhibit 7: Results for distinguishing price charts from their permutation—variant. For each contest, the p-value is for the two-sided t-test of the null hypothesis that the average across subjects of the number of their correct guesses equals the total number of guesses in the contest divided by 2.

Results for this variant are reported in Exhibit 7. The p-value is less than 6% for all but one data set. We attribute the slightly less decisive outcome for this variant to the smaller number of subjects.

AR1

Here we want to test the null hypothesis H that human subjects cannot distinguish between an actual time series S and a time series that is generated by an AR1 process that is calibrated to match mean, variance, and (first-order) autocovariance of S. Details follow. We refer to Section 3.4 of Hamilton (1994) for background on AR1 processes.

Again we begin with a time series of actual historical prices $\{p_0, p_1, p_2, \ldots, p_T\}$ and compute the logarithmic returns $\{r_t\}$,

$$r_t \;\equiv\; \log(p_t) \;-\; \log(p_{t-1}). \tag{2}$$

Then we compute the sample mean μ, variance v, and (first-order) autocovariance α of the series r. This defines an AR1 process

$$y_t := c + \phi \cdot y_{t-1} + \epsilon_t,$$

where ϵ_t are i.i.d. normal distributions with mean 0 and variance σ^2, as follows:

$$\begin{aligned} \phi &= \alpha/v, \\ c &= \mu(1-\phi), \text{ and} \\ \sigma^2 &= v(1-\phi^2). \end{aligned}$$

The starting point y_0 of the AR1 process is taken to be r_h for an index h chosen uniformly at random.

And finally we set

$$p^*_{t+1} \;\equiv\; p^*_t \cdot e^{y_{t+1}} \quad , \quad p^*_0 \;\equiv\; 1.$$

The results are reported in Exhibit 8. We obtain a p-value less than 0.505% for five of our eight data sets, and higher for the other three.

Comparison of random permutation and AR1 results

In this section we investigate whether subjects do better when presented with the permutation process than with an AR1 process. As a first step, in Exhibit 9 we present the results of one-sided, independent samples t-test for success rate decline between the random permutation and AR1 experiments, reported in Exhibits 6 and 8, respectively. For each contest, the null hypothesis is that

Results for Distinguishing Price Charts from AR1

Contest	Subjects	p-Value	Correct Guesses per Subject As Percentage of Total Guesses
Mandrill	36	0.55792	52 46 40 44 48 54 58 36 58 64 48 56 60 44 36 50 42 56 50 58 52 56 50 50 46 60 44 56 48 54 48 58 50 44 54 54
Bear	40	0.00000	100 59 86 97 49 62 100 54 27 73 54 89 73 92 49 78 59 78 89 89 95 70 62 92 73 97 95 43 86 95 54 100 76 57 92 89 76 92 97 100
Lynx	38	0.14392	54 60 51 43 31 46 40 57 46 63 57 54 51 57 40 60 57 66 60 57 49 66 37 43 37 57 37 51 54 69 60 63 51 54 51 51 46 54
Reindeer	39	0.00033	58 48 90 58 63 55 43 58 53 63 70 45 40 43 48 45 45 43 45 68 48 45 90 58 70 55 43 58 90 43 70 78 83 73 70 55 80 70 65
Beaver	37	0.10729	50 58 39 72 47 64 69 69 58 47 56 56 44 42 56 56 61 44 53 58 44 47 47 69 53 53 50 58 28 44 53 33 33 58 56 67 61
Bull	37	0.00000	77 65 65 77 77 55 81 58 42 77 81 97 81 87 55 61 97 97 87 48 90 74 61 87 58 90 71 55 77 65 55 71 71 87 84 81 87
Elk	36	0.00505	53 55 63 45 40 63 48 43 63 48 48 63 45 48 48 68 48 63 88 60 45 58 63 70 48 65 58 55 60 38 50 55 48 53 75 58
Seal	38	0.00000	67 51 59 67 59 69 64 46 59 67 85 69 79 56 56 49 69 82 59 79 62 33 82 44 79 74 62 77 67 49 62 56 79 69 77 74 59 79

Exhibit 8: Results for distinguishing price charts from AR1. For each contest, the p-value is for the two-sided t-test of the null hypothesis that the average across subjects of the number of their correct guesses equals the total number of guesses in the contest divided by 2.

One-Sided t-Test for Success Rate Decline from Permutation to AR1	
Contest	p-Value
Mandrill	0.067
Bear	0.949
Lynx	0.156
Reindeer	0.084
Beaver	0.096
Bull	0.014
Elk	0.000
Seal	0.009

Exhibit 9: One-sided, independent samples t-test for success rate decline between the random permutation and AR1 experiments. For each contest, the null hypothesis is that the average, across subjects, of their success rates is equal for the two experiments. The alternative hypothesis is that the average success rate in the AR1 experiment is lower than the average success rate in the random permutation experiment. The success rate of a particular subject is defined as the number of their correct guesses divided by the number of charts in the contest.

the average, across subjects, of their success rates is equal for the two experiments. The alternative hypothesis is that the average success rate in the AR1 experiment is lower than the average success rate in the random permutation experiment. The success rate of a particular subject is defined as the number of their correct guesses divided by the number of charts in the contest. For 6 out of 8 data sets the null hypothesis is rejected at least at the 10% level.

To gain insight into differences in performance between the two experiments, in Exhibit 10 we present the first five autocorrelations and the Ljung-Box Q statistic, computed using 20 lagged terms of the actual and synthetic data. The Q statistic tests the null hypothesis that autocorrelations up to lag 20 equal zero, i.e., it tests for "overall" randomness in returns. The first thing to notice is that for the actual data (the top panel of Exhibit 10) we reject the null hypothesis of overall randomness in each of the 8 data sets with high confidence (p-values of the Q statistic are 0.000). Under random permutation, we fail to reject the null (the middle panel of Exhibit 10). The fact that overall randomness of the shuffled data is so different from that of the actual data helps explain why subjects performed so well under the permutation experiment.

Autocorrelations and the Ljung-Box Q Statistic for the Actual and Simulated Data

Contest	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	Ljung-Box Q_{20}	p-value (Q_{20})
				Actual Data			
Mandrill	8.0	1.7	0.4	1.6	-3.0	99	0.000
Bear	11.2	6.5	4.4	4.1	3.8	18845	0.000
Lynx	2.9	-0.3	1.4	1.1	-3.2	66	0.000
Reindeer	-34.8	11.0	-4.0	4.0	-2.4	137260	0.000
Beaver	1.4	-3.0	1.6	1.3	1.1	99	0.000
Elk	31.0	20.2	15.9	13.0	10.6	52615	0.000
Bull	-12.3	1.2	1.2	0.3	0.1	15227	0.000
Seal	8.4	6.2	7.4	4.9	5.5	710	0.000
				Random Permutation			
Mandrill	-2.1	3.1	-2.5	-1.1	-0.5	25	0.189
Bear	-0.2	0.0	-0.1	0.0	0.0	18	0.611
Lynx	0.7	0.9	-0.4	-0.7	0.5	15	0.753
Reindeer	0.1	0.0	0.0	0.1	0.0	10	0.966
Beaver	-0.4	1.7	0.1	-0.8	1.1	20	0.474
Elk	0.1	0.1	0.3	0.0	0.2	16	0.687
Bull	0.1	0.0	0.0	-0.1	-0.1	13	0.868
Seal	0.6	0.8	1.5	0.2	0.1	21	0.420
				AR1			
Mandrill	10.1	1.1	0.7	-1.9	-0.8	80	0.000
Bear	11.2	1.5	0.1	0.0	-0.1	9354	0.000
Lynx	4.3	-1.0	0.6	0.3	-0.9	39	0.006
Reindeer	-35.0	12.2	-4.2	1.4	-0.5	138450	0.000
Beaver	0.9	-0.4	0.1	-0.1	-0.6	29	0.092
Elk	30.8	9.2	2.6	0.9	0.3	24274	0.000
Bull	-12.4	1.7	-0.4	0.0	0.0	15425	0.000
Seal	8.2	0.4	-0.5	-1.6	0.2	137	0.000

Exhibit 10: Autocorrelations and the Ljung-Box Q statistic, computed using 20 lagged terms. The Q statistic tests the null hypothesis that autocorrelations up to lag 20 equal zero, i.e. that returns are random and independent. The corresponding p-values are also presented. These statistics are reported for the actual data, as well as for synthetic processes constructed using random permutation or an AR1 process calibrated to the data. The synthetic processes are constructed based on the entire data sample.

On the other hand, for the AR1 process, we do reject the null hypothesis of overall randomness (the bottom panel of Exhibit 10). In fact, for 6 out of the 8 contests, the p-values of the Q statistic are the same as those of the actual data up to three decimal places. For the remaining 2 contests we reject the null at the 1% level in one case and at the 10% level in the other case. This similarity in overall randomness between the actual and AR1 data helps explain why the subjects had a somewhat harder time in the AR1 test. Interestingly, this similarity appears especially pronounced precisely in the three contests where subjects have the hardest time distinguishing the AR1 process from the actual data: Mandrill, Lynx, and Beaver.

Learning

As our game provides feedback, we investigate whether subjects improve their performance while playing. We do so by comparing the subjects' performance in the first and the last part of each contest. Specifically, for each contest, we consider the subset consisting of the first $\alpha = 1/5$ fraction of guesses, and that consisting of the last α fraction.[5] For each subset, we add up the number of correct guesses across subjects and divide that sum by the total number of guesses in the subset times the number of subjects. We call this the fraction of correct guesses made by the combined pool of subjects. We refer to this fraction in the first (last) part of each contest as "Correct First" ("Correct Last").

Exhibit 11 reports the results for the permutation and AR1 processes. For the permutation process (presented in the left-hand-side panel of the table), in all but one contest the average number of correct guesses increases. To check whether this increase is statistically significant across contests, we conduct one-sided Wilcoxon signed rank test on the results presented in the table. In particular, we test the null hypothesis that "Correct First" minus "Correct Last" comes from a distribution with zero median against the alternative that the median of the "Correct First" column is less than the median of the "Correct Last" column.[6] Exhibit 11 shows that the increase in correct guesses is significant at the 10% level.

The right-hand-side panel of Exhibit 11 reports the results for the AR1 process. Across all

[5]Non-integer numbers are rounded down to the nearest integer.

[6]We use the signed rank test rather than the t-test because of the small sample size.

Performance Improvement in the Combined Pool of Subject

Contest	Random Permutation			AR1		
	Num. Guesses	Correct First	Correct Last	Num. Guesses	Correct First	Correct Last
Mandrill	557	0.506	0.589	359	0.496	0.482
Bear	384	0.693	0.734	280	0.739	0.832
Lynx	391	0.573	0.588	265	0.525	0.536
Reindeer	445	0.620	0.656	312	0.510	0.635
Beaver	406	0.559	0.594	258	0.543	0.500
Bull	342	0.798	0.825	221	0.674	0.819
Elk	461	0.690	0.755	287	0.544	0.568
Seal	385	0.787	0.592	266	0.677	0.519

p -Values of One-Sided Signed Rank Test

	Random Permutation	AR1
All contests	0.098	0.320

Exhibit 11: Performance improvement with the permutation and AR1 processes. For each contest, the column "Correct First" reports the fraction of correct guesses made by the combined pool of subjects in the first one-fifth of guesses. The column "Correct Last" reports the corresponding value for the last one-fifth of guesses. The column "Num. Guesses" is the denominator in the calculation of these fractions of correct guesses. For each synthetic process, the table also reports the p-values of one-sided Wilcoxon signed rank test, which tests the null hypothesis that "Correct First" minus "Correct Last" comes from a distribution with zero median against the alternative hypothesis that the median of the "Correct First" column is less than the median of the "Correct Last" column.

Performance Improvement Subject By Subject

Contest	Random Permutation	AR1
	p-Values of One Sided t-Test	
Mandrill	0.006	0.654
Bear	0.095	0.011
Lynx	0.344	0.399
Reindeer	0.158	0.006
Beaver	0.164	0.801
Bull	0.134	0.001
Elk	0.020	0.317
Seal	1.000	0.998

Exhibit 12: For each contest, for each subject in that contest, we take the fraction of correct guesses in the first one-fifth of guesses and the fraction of correct guesses in the last one-fifth of guesses. For each contest, we then report the p-value of the one-sided t-test of the null hypothesis that subject by subject “Correct First” minus “Correct Last” comes from a distribution with zero mean. The alternative hypothesis is that the mean of the “Correct First” column is less than the mean of the “Correct Last” column.

contests we cannot reject the null hypothesis that the median success rate is the same in the first and the last fraction of guesses. However, in some cases, like Reindeer or Bull, the difference in the average success rates seems significant. Indeed, Exhibit 12 shows that if we conduct a significance test contest by contest (rather than across contests as above in Exhibit 11), we find evidence of learning in three out of eight contests under either random permutation or AR1. Here for each subject in a given contest, we take the fraction of correct guesses in the first one-fifth of guesses and the fraction of correct guesses in the last one-fifth of guesses. For each contest, we then report the p-value of the one-sided t-test of the null hypothesis that subject by subject "Correct First" minus "Correct Last" comes from a distribution with zero mean, against the alternative that the mean of the subject by subject "Correct First" data is less than that of the corresponding "Correct Last" data. The table shows that for the permutation process learning is statistically significant for Mandrill, Bear, and Elk at least at the 10% level, while for the AR1 process, it is significant for Bear, Reindeer, and Bull at least at the 5% level.

Recall also that in our experiment subjects are required to practice before entering a contest. This makes the results in this section less prone to be influenced by extraneous factors such as becoming comfortable with the interface.

Conclusion

A natural question that arises is how were the subjects able to perform so well in seven out of eight data sets? Casual inspection of Exhibits 1–4 shows that distinguishing real data from synthetic data is challenging; for some data sets the real chart tends to be smoother, as in Exhibit 2, while for other data sets the opposite is true, the real chart tends to be spikier, as in Exhibit 4. What complicates the matter further is that, as is evident from the data, the "smoothness" of actual data varies with time. Still, feedback from just a few trials seems sufficient for the user to extract characteristics of the data to be used in classifying charts in the near future. The importance of feedback is supported by the information about winning strategies that some of the subjects volunteered to share with us (anonymously). For example, a subject wrote:

> Admittedly, when first viewing the two data sets in the practice mode, it is impossible to tell which one is real, and which one is random, however, there is a pattern that quickly emerges and then the game becomes simple and the human eye can easily pick out the real array (often in under 1 second of time).

For some contests, our results suggest that indeed subjects improve while playing.

An interesting direction is to compare humans' performance against the performance of computers, following a vast literature, cf. Lawrence et al. (2006). In our experiment, the human eye—as opposed to a computer algorithm—may have an advantage. It is well known that computers still struggle with many image-recognition and classification tasks that are trivial for humans. The same may be the case for distinguishing asset returns from synthetic processes.

Given the recent regulatory push towards ensuring that "consumers have the information they need to choose the consumer financial products and services that are best for them,"[7] the study of optimal ways to present financial data to investors is of current interest. Our paper is a contribution to the growing body of literature on the usefulness of temporal charts in evaluation of financial asset performance.

References

M. H. Bazerman. Consumer research for consumers. *Journal of Consumer Research*, 27(4):499–504, 2001.

Z. Bodie. The challenge of investor education. In Z. Bodie, D. McLeavey, and L. B. Siegel, editors, *The Future of Life-Cycle Saving and Investing*, pages 169–171. Research Foundation of the CFA Institute, February 2008.

J. Choi, D. Laibson, and B. Madrian. $100 bills on the sidewalk: Suboptimal saving in 401(k) plans. *Review of Economics and Statistics*, 2010.

[7] The Consumer Financial Protection Bureau, `http://www.consumerfinance.gov/protecting-you/`.

W. P. M. De Bondt. Betting on trends: Intuitive forecasts of financial risk and return. *International Journal of Forecasting*, 9(3):355–371, 1993.

E. Fama. The behavior of stock market prices. *Journal of Business*, 38(1):34–105, 1965.

E. Fama. Random walks in stock market prices. *Financial Analysts Journal*, 21:55–59, 1965.

E. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25:383–417, 1970.

J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *27th International Conference on Human Factors in Computing Systems (CHI)*, pages 1303–1312, 2009.

A. A. Hung, A. Heinberg, and J. K. Yoong. Do risk disclosures affect investment choice? Technical report, RAND Labor and Population, September 2010.

E. J. Keogh and S. Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371, 2003.

J. Kozup, E. Howlett, and M. Pagano. The effects of summary information on consumer perceptions of mutual fund characteristics. *Journal of Consumer Affairs*, 42(1):37–59, Spring 2008.

Y. Kroll, H. Levy, and A. Rapoport. Experimental tests of the mean-variance model for portfolio selection. *Organizational Behavior and Human Decision Processes*, 42(3):388–410, 1988.

M. Lawrence, P. Goodwin, M. OConnor, and D. Önkal. Judgmental forecasting: A Review of Progress over the Last 25 Years. *International Journal of Forecasting*, 22:493–518, 2006.

A. Lo, H. Mamaysky, and J. Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *Journal of Finance*, LV(4):1705–1765, August 2000.

A. W. Lo and A. C. MacKinlay. Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies*, 1(1):41–66, 1988.

A. W. Lo and A. C. MacKinlay. *A Non-Random Walk Down Wall Street*. Princeton University Press, Princeton, NJ, 1999.

B. G. Malkiel. *A Random Walk Down Wall Street*. W. W. Norton & Company, 1973.

H. V. Roberts. Stock-market 'patterns' and financial analysis: Methodological suggestions. *The Journal of Finance*, 14(1):1–10, 1959.

P. Samuelson. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6:41–49, 1965.

L. E. Swedroe. *The Only Guide to a Winning Investment Strategy You'll Ever Need*. St. Martin's Press, 2005.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

K. E. Wärneryd. *Stock-market psychology: How people value and trade stocks*. Edward Elgar, 2001.

H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Trans. Vis. Comput. Graph.*, 16(6):973–979, 2010.