Equilibrium Data Mining and Data Abundance

Jérôme Dugast, Thierry Foucault

Equilibrium Data Mining and Data Abundance*

Jérôme Dugast† Thierry Foucault‡

First Version: November 2019.
This version: October 8, 2020

Abstract

We analyze how computing power and data abundance affect speculators' search for predictors. In our model, speculators search for predictors through trials and optimally stop searching when they find a predictor with a signal-to-noise ratio larger than an endogenous threshold. Greater computing power raises this threshold, and therefore price informativeness, by reducing search costs. In contrast, data abundance can reduce this threshold because (i) it intensifies competition among speculators and (ii) it increases the average number of trials to find a predictor. In the former (latter) case, price informativeness increases (decreases) with data abundance. We derive implications of these effects for the distribution of asset managers' skills and trading profits.

Keywords: Alternative Data, Data Abundance, Data Mining, Price Informativeness, Search for Information.

*We are grateful to Bruno Biais, Dion Bongaerts (discussant), Adrian Buss, Jean-Edouard Colliard, Bernard Dumas, Sergei Glebkin, Denis Gromb, Johan Hombert, Joël Peress, Daniel Schmidt, Alberto Teguia (discussant), Josef Zechner and participants at the 2020 European Finance Association Meetings, the 2020 future of Financial Information Conference, The market microstructure exchange workshop, HEC Paris, INSEAD, the University of New South Wales, and the University of Vienna for very useful comments. Future versions will be available at: https://sites.google.com/view/jeromedugast/home or https://thierryfoucault.com/.

†Université Paris Dauphine - PSL. Tel: (+33) 01 44 05 40 41 ; E-mail: jerome.dugast@dauphine.psl.eu

‡HEC, Paris and CEPR. Tel: (33) 1 39 67 95 69; E-mail: foucault@hec.fr

1. Introduction

Asset managers devote considerable effort to find new investment signals (predictors of asset cash-flows or returns). To this end, they take advantage of progress in information technologies. This progress has reduced information processing costs (due to improvement in computing power) and considerably increased the volume and diversity of available data (due to digitization and increase in storage capacities).[1] For instance, asset managers increasingly buy so called "alternative data", such as credit/debit card data, app usage data, satellite images, social media, web traffic data, etc. and use computer-based methods to extract predictors of asset payoffs from these data and design trading strategies exploiting these predictors.[2]

Data abundance and improvements in computing power are related but distinct phenomena. For instance, unstructured data such as satellite images or text from social media expand the set of variables to obtain predictors of future firms' earnings. However, they do not per se reduce the cost of processing data for obtaining these predictors. Thus, to understand the effects of data abundance, one would like a theory of information acquisition in which one can analyze the effect of expanding the search space for predictors holding the cost of data processing *constant* (and vice versa). This is not possible in existing models of financial information acquisition (e.g., Verrecchia (1982)), which captures all dimensions of the progress in information through a single variable, namely the cost of acquiring information of a given precision. Thus, in this paper, we propose an extension of standard models of information acquisition which allows to analyze data abundance, holding the cost of finding information constant. Using this model, we show that the effects of data abundance and progress in computing power on equilibrium outcomes in financial markets are different.

Our model features a continuum of risk averse speculators (asset managers). In the first stage (the "exploration stage"), each speculator optimally scours available data to find a predictor of the payoff of a risky asset. In the second stage (the "trading stage"),

[1] See Goldfarb and Tucker (2019) and Nordhaus (2015) for a discussion of the economic implications of this evolution.

[2] Marenzi (2017) estimates that asset managers have spent more than four billion in alternative data in 2017 (see also "*Asset managers double spending in new data in hunt for edge*", Financial Times, May 9, 2018. Abis (2018) finds that quantitative funds (using computer-driven models to analyze large datasets) have quadrupled in size from 1999 to 2015 and that their growth has been more than twofold that of discretionary funds. Moreover, Grennan and Michaely (2019) find that about 87% of the FinTechs in their sample (190 FinTechs) specialize in producing investment signals using artificial intelligence.

each speculator observes the realization of her predictor and optimally chooses her trading strategy. We formalize the trading stage as a standard rational expectations model (similar to Vives (1995)). The novelty of our model (and its implications) stems from the exploration stage. Here, instead of following the standard approach (e.g., Grossman and Stiglitz (1980) or Verrecchia (1982)), whereby speculators obtain a predictor of a given precision in exchange of a payment, we explicitly model the search for a predictor as a sequential process and we analyze how the optimal search strategy depends on (i) the cost of exploration and (ii) the amount of data available for exploration (the "search space").

We model the search for predictors as follows. We assume that existing data can be combined to generate predictors differing in their signal-to-noise ratios ("quality"). The search space is determined by the quality of the most informative predictor (the "data frontier"), denoted τ^{max}, and the least informative predictor, which is just noise. The distribution of the quality of predictors on this interval is exogenous. Given this distribution, each speculator simultaneously and independently explores ("mines") the data. Each new exploration costs c and returns a predictor whose quality is drawn from the distribution of predictors' quality. After obtaining a predictor, a speculator can decide either to explore the data further, to possibly obtain an even better predictor, or to trade on the predictor she just found.

As a motivation for our approach, consider asset managers using accounting variables to forecast future stock earnings. There are many ways to combine these variables to obtain predictors. For instance, using 240 accounting variables, Yan and Zheng (2017) build more than $18,000$ trading signals and find that many of these yield significant abnormal returns (even after accounting for the risk of data snooping). The data mining cost, c, represents the labor and computing costs of considering a particular predictor (a particular combination of the accounting variables), designing a trading strategy based on this predictor, backtesting it, and thinking about possible economic stories for why the strategy works. After obtaining a predictor, each manager can decide to start trading on it or to keep searching for another, more precise, predictor.

New datasets enable speculators to use new variables to forecast asset payoffs and should therefore push back the data frontier, i.e., increase τ^{max}.[3] In fact, the advent

[3]Recent empirical findings support this conjecture. For instance, Katona et al. (2019) find that combining satellite images of parking lots of U.S. retailers from two distinct data providers improves the

of big data in asset management is often described as a gold rush for this reason: Big data combined with new forecasting techniques (machine learning) enable asset managers to discover more precise predictors.[4] We refer to this dimension of data abundance as the "hidden gold nugget" effect. However, data abundance also creates "a needle in the haystack problem": It results in a proliferation of datasets and only a fraction of these datasets contains useful information for forecasting asset payoffs. Separating the wheat from the shaff can only be done through explorations, which is costly. To capture this dimension of data abundance, we assume that each exploration returns an informative predictor with probability $\alpha < 1$. In sum, we analyze the effect of data abundance on equilibrium outcomes by considering either an increase in τ^{max} (the hidden gold nugget effect) or a decrease in α (the needle in the haystack problem).[5]

As for greater computing power, it reduces the cost of exploring a new dataset.[6] Thus, we study the effect of greater computing power by considering the effect of a decrease in the cost of exploration, c, on equilibrium outcomes.

In equilibrium, each speculator's optimal search strategy follows a stopping rule: She stops searching for a predictor after finding one whose quality (signal-to-noise ratio) exceeds an endogenous threshold, denoted τ^* (we refer to such a predictor as being "satisficing"). This threshold is such that the speculator's expected utility of trading on a predictor of quality τ^* is just equal to her expected utility of searching for another predictor. The latter reflects the prospect of obtaining a larger expected trading profit by finding a predictor of higher quality deflated by the *total* expected cost of search to find such a predictor (i.e., the per-exploration cost, c times the expected number of

accuracy of the forecasts of these retailers' quarterly earnings (see also Zhu (2019)). Also, van Binsbergen et al. (2020) find that, with machine learning techniques, one can obtain more precise forecasts of firms' future earnings than analysts' forecasts (they use random forests regressions combining more than 70 accounting variables with analysts' forecasts).

[4]See, for instance, "*Hedge funds see a gold rush in data mining*", Financial Times, August 28, 2017.

[5]As an illustration, consider searching for medication to cure the Coronavirus in the scientific literature on this topic. There have been more than 23,000 scientific papers written on this topic between January and June 2020 (see da Silva et al. (2020)). As this number grows, the fraction of truly informative papers might drop, even though the chance of a scientific discovery that stops the virus goes up.

[6]For instance, an increase in computing power reduces the time costs of finding predictors. Brogaard and Zareei (2019) use a genetic algorithm approach to select technical trading rules. They note that "*the average time needed to find the optimum trading rules for a diversified portfolio of ten NYSE/AMEX volatility assets for the 40 year sample using a computer with an IntelÂ® Core(TM) CPU i7-2600 and 16 GM RAM is 459.29 days (11,022.97 hours)." For one year it takes approximately 11.48 days.*" They conclude that their analysis would not be possible without the considerable increase in computing power in the last 20 years.

explorations required to find a predictor with a quality higher than τ^*).

All speculators use the same stopping rule because they are ex-ante identical (same preferences, search cost etc.). However, as explorations' outcomes are random, speculators find and trade on predictors of different quality. Thus, in equilibrium, (i) only predictors of sufficiently high quality are used for trading and (ii) speculators endogenously exploit predictors of different quality. Specifically, the quality of predictors used in equilibrium ranges from τ^* (the least informative predictor used in equilibrium) to τ^{max} (most informative).

Greater computing power induces speculators to adopt a more stringent stopping rule in equilibrium, i.e., a decrease in c raises τ^*. Indeed, a decrease in the per-exploration cost, c, directly reduces the total expected cost of launching a new exploration after finding a predictor. Hence, it raises the value of searching for another predictor after finding one and therefore it induces speculators to be more demanding for the quality, τ^*, of the least informative predictor used in equilibrium. An indirect consequence (the "*competition effect*") is that, on average, speculators trade more aggressively on their signal. Indeed, they face less uncertainty on the asset payoff because their predictors are better on average. As a result, price informativeness increases. The competition effect dampens the positive effect of a reduction in the exploration cost on the value of searching for a better predictor. However, it is never strong enough to fully offset it.

The needle in the haystack problem (a drop in α) does not affect the per exploration cost, c. However, it raises the total expected cost of search for speculators because it reduces the chance of finding a satisficing predictor in each exploration. For this reason, it leads speculators to be *less* demanding for the quality of the least informative predictor, τ^*, for the same reasons as an *increase* in the per exploration cost does.

The effect of pushing back the data frontier (an increase in τ^{max}) on speculators' optimal search strategy (τ^*) is more subtle because it directly affects the value of searching for another predictor in two opposite directions. One the one hand, it raises this value for two reasons. First, holding investors' stopping rule constant, it enlarges the range of satisficing predictors, which raises the probability that each exploration is successful. This effect reduces the total expected cost of search. Second, holding price informativeness constant, it increases the expected utility of trading on a satisficing predictor due to the prospect of finding even more informative predictors (the "hidden gold nugget effect").

However, an increase in the quality of the best predictor also has a direct positive effect on price informativeness because it raises the average quality of predictors and therefore the average aggressiveness with which speculators exploit their signals. This competition effect reduces the value of searching for predictors. We show that it dominates when τ^{max} is high enough. Then, a push back of the data frontier leads speculators to follow a less demanding search policy (i.e., τ^* drops). Thus, the model implies an inverse U-shape relationship between the quality of the least informative predictor used in equilibrium (τ^*) and the quality of the most informative predictor.

In sum, the model highlights two channels through which data abundance can reduce the quality of the least informative predictor used in equilibrium: (i) It reduces the trading value of predictors by intensifying competition among speculators (the "competition effect") and (ii) it increases the total expected cost of search, even though it does not change the per exploration cost ("needle in the haystack effect").

The model has several testable implications. First, it has implications for the distribution of investment skills across funds (or managers of these funds). Several papers (e.g., Kacperczyk and Seru (2007) or Kacperczyk et al. (2014)) relate these skills to the quality (precision) of asset managers' signals and interpret heterogeneity in skills as heterogeneity in the quality of these signals. In our model, this distribution is endogenous.[7] In particular, shocks to computing power, data abundance and other parameters of the model affect the lower bound of this distribution and therefore the range of skills (say, the difference in skills between funds in the lowest and top skill deciles). For instance, the model predicts that improvements in computing power should reduce this range (because it increases τ^*) while data abundance (a push back of the data frontier or the needle in the haystack problem) can have the opposite effect (because it can reduce τ^* and, at least weakly, weakly improves τ^{max}). The model also implies that an increase in prior uncertainty (the variance of the asset payoff) or the volume of uninformed (noise) trading should reduce the range of funds' skills because it induces speculators to be more demanding for the quality of their predictors in equilibrium.

[7] In our model, heterogeneity in speculators' skills arises even though speculators are ex-ante identical. They eventually trade on predictors of different quality (appear to have different skills) because the outcome of their search for predictors is random, even though search is optimal. This finding suggests that heterogeneity in asset managers' skills is not necessarily due to innate differences in abilities or differences in efforts (in our model, speculators who happen to pay a larger search cost and therefore seem to exert more effort do not necessarily trade on predictors of higher quality). It might just reflect luck in the search process for predictors.

Our second set of predictions is about asset price informativeness. Our model predicts that greater computing power improves price informativeness because it leads speculators to be more demanding for the quality of their predictors.[8] In contrast, the effect of data abundance on asset price informativeness is more complex. On the one hand, it can lead speculators to be less demanding for the quality, τ^*, of the least satisficing predictor. On the other hand, it pushes back the data frontier and improves the quality of the most informative predictor. The first effect reduces the average quality of predictors used by investors while the second improves it. As a result, the effect of data abundance on price informativeness is ambiguous in our model. In the absence of the needle in the haystack problem ($\alpha = 1$), we show that the second effect dominates and therefore data abundance improves price informativess. In contrast, if data abundance also makes the needle in the haystack problem more severe (α decreases) then the first effect can dominate so that price informativeness drops when more data become available.[9].

Our third set of predictions regards effects of computing power and data abundance on speculators' trading profits (excess returns) and the crowdedness of their strategies (measured by the correlation of their holdings). The model predicts an inverse U-shape relationship between speculators' average trading profits and computing power. Indeed, greater computing power raises the average quality of the predictors used in equilibrium and therefore price informativeness. The first effect raises speculators' expected trading profit while the second reduces it. The former dominates if and only if speculators' cost of exploration, c, is large enough. An improvement in the data frontier has the same effect for the same reasons. The needle in the haystack problem reduces price informativeness and the average quality of predictors used in equilibrium. The second (first) effect dominates when the problem becomes sufficiently severe (α is large enough). Hence, ultimately, the model also predicts an inverse U-shape relationship between speculators' average trading profits and data abundance. So overall the model implies that progress in information technologies initially benefit to all speculators until a point where it starts reducing their profits. Finally, we show that greater computing power or an improvement in the

[8]In line with this prediction, Gao and Huang (2019) find that the introduction of the EDGAR system in the U.S. (which allows investors to have internet access to electronic filings by firms) had a positive effects on measures of price efficiency. One possible reason, as argued by Gao and Huang (2019), is that the EDGAR system reduced the cost of accessing data (a component of exploration cost) for investors.

[9]Given that technological progress has both enlarged the search space and reduced search costs, these implications of our model can explain why the empirical literature on the effect this progress on asset price informativeness reports conflicting results. See Section 5.2 for a discussion.

data frontier reduce the pairwise correlation in speculators' trades while a drop in the proportion of informative datasets (α) has the opposite effect.

2. Related Literature

Our paper contributes to the literature on informed trading with endogenous information acquisition (e.g., Grossman and Stiglitz (1980), Verrecchia (1982); see Veldkamp (2011) for a survey). This literature often takes a reduced-form approach to model the cost of acquiring a signal of given precision. For instance, Verrecchia (1982) (and several subsequent papers) assumes that this cost is a convex function of the precision of the signal. The learning technology in our model is different. Indeed, speculators do not control the exact precision of their signal (which ultimately is random) but only the lower bound of this precision. In raising this bound, they raise the expected precision of their signal but they also raise their total expected search cost (as the expected number of exploration rounds increases when speculators use a more stringent stopping rule). The relationship between a speculator's expected search cost and expected precision is endogenous and micro-founded by an optimal search model.[10] As explained previously, this approach gives us a way to analyze separately the effects of greater computing power (a decrease in the cost of processing data) and data abundance (an expansion of the search space).

Banerjee and Breon-Drish (2020) consider a model in which one informed investor can dynamically control his timing for information acquisition about the payoff of a risky asset. In this model, the informed investor optimally alternates between periods in which she searches for information (when the volume of noise trading is high enough) and periods in which she does not (when the volume of noise trading is low). When she searches for information, the investor finds a signal of a given precision according to a

[10]Han and Sangiorgi (2018) offers an interesting micro-foundation for the specification of information acquisition costs based on a model in which an agent can draw normally distributed signals from a fixed set (an "urn"), with replacement (so that the agent can draw the same signal multiple times). Each draw is costly in their model. They show that the relationship between the precision of the average signal obtained by the agent (a sufficient statistics for all his signals) and her total investment in drawing signals is convex and becomes linear when the number of possible signals goes to infinity. Han and Sangiorgi (2018) use this specification to analyze an optimal forecasting problem. Our approach differs in many respects. In particular, we jointly solve for the equilibrium of the market for a risky asset and speculators' optimal search for predictors (in Han and Sangiorgi (2018), the number of draws by an agent is exogenous and they do not apply their model to trading in financial markets).

Poisson process and starts trading on this signal as soon as she finds it. Interestingly, Banerjee and Breon-Drish (2020) shows that this dynamic model generates predictions different from the standard static model in which the informed investor must decide to acquire a signal before trading. In contrast, we depart from the traditional standard static model by modeling informed investors' search for signals of different precisions (in a static environment since there is no time-variation in parameters affecting the profitability of informed trading over exploration rounds in our model) and we compare the effects (e.g., on the heterogeneity in signals' precisions) of a reduction in search costs with the effects of expanding the search space (data abundance).

Our paper is also related to the recent literature analyzing the economic effects of progress in information technologies (see, Goldfarb and Tucker (2019) and Veldkamp and Chung (2020) for a review) and more specifically theoretical papers analyzing the effects of these technologies for the production of financial information (e.g., Abis (2018), Dugast and Foucault (2018), Farboodi and Veldkamp (2019), or Huang et al. (2020)). These papers analyze this progress as a decrease in the cost of processing information or, similarly, an increase in investors' information processing capacities. In contrast, our model focuses on another dimension of this progress, namely data abundance, i.e., the expansion of investors' search space for predictors. We show that the effects of data abundance and the cost of processing data (c in our model) are different and derive several implications that should allow empiricists to test whether these differences matter empirically. Also, we explicitly analyze the acquisition of financial information as a search problem and consider the effects of reducing the cost of search (c) and increasing the search space on equilibrium outcomes. Goldfarb and Tucker (2019) and Agrawal et al. (2019) highlight the importance of doing so to understand economic implications of digitization and artificial intelligence.

3. Model

We consider a financial market with a unit mass continuum of risk averse (CARA) speculators, a risk neutral and competitive market maker, and noise traders. Investors can invest in a risky asset and a risk free asset with interest rate normalized to zero. Figure 1 describes the timing of the model.

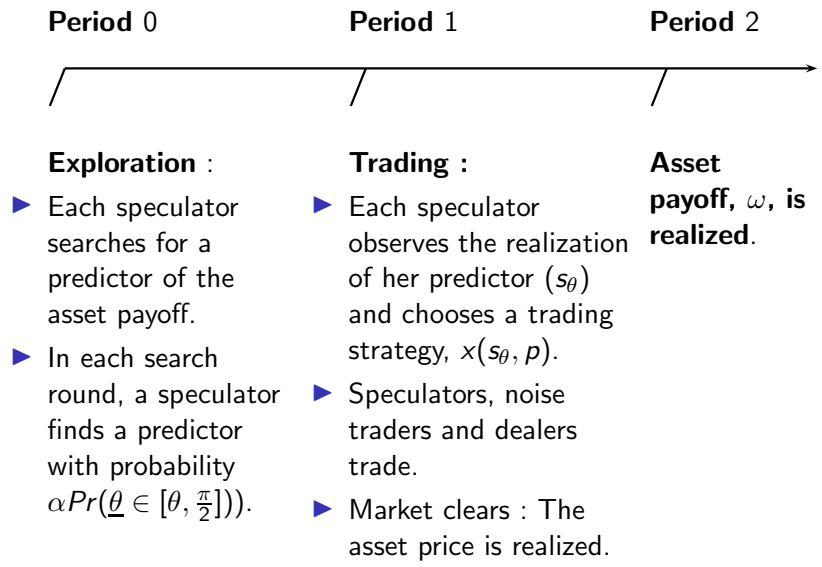

Figure 1: Timing

The payoff of the risky asset, ω, is realized in period 2 and is normally distributed with mean zero and variance σ^2. Speculators search for predictors of the asset payoff in period 0 (the "exploration stage"). Then, in period 1 (the "trading stage"), they observe the realization of these predictors and can trade on them in the market for the risky asset. We now describe these two stages in details.

The exploration stage. In period 0, each speculator i searches for a *predictor* of the asset payoff, ω. There is a continuum of potential predictors. Each predictor, s_θ, is characterized by its type θ and is such that:

$$s_\theta = \cos(\theta)\omega + \sin(\theta)\varepsilon_\theta, \tag{1}$$

where $\theta \in [0, \pi/2]$ and the ε_θs are normally and independently distributed with mean zero and variance σ^2. Moreover, ε_θ is independent from ω. Let $\tau(\theta) \equiv \cos^2(\theta)/\sin^2(\theta) = \cot^2(\theta)$ denote the signal-to-noise ratio for a predictor with type θ. We refer to this ratio as the "quality" of a predictor.[11] The quality of a predictor is inversely related to its type,

[11]Observe that the predictor s_θ is equivalent (in terms of informativeness) to the predictor $\hat{s}_\theta = \omega + \cot(\theta)^{-1}\varepsilon_\theta$, whose precision is $\tau(\theta)/\sigma^2$. Thus, a predictor of high quality is a predictor with high

θ and unrelated to the risk of the asset, σ^2, because $\mathsf{Var}[\varepsilon_\theta] = \mathsf{Var}[\omega] = \sigma^2$. Without this assumption, the quality of all predictors would, counter-intuitively, increase with the uncertainty of the final payoff (σ^2).

We assume that predictors' types, θs, are distributed according to the cumulative probability distribution $\Phi(.)$ (density $\phi(.)$) on $[0, \pi/2]$. Speculators discover predictors' types in period 0 through a sequential search process. Each search round corresponds to a new exploration ("mining") of available data to obtain a new type of predictor. Each exploration costs c. It is unsuccessful, i.e., yields no predictor (or equivalently a predictor that is just noise), with probability $(1 - \alpha \Pr(\theta \in [\underline{\theta}, \frac{\pi}{2}))$, where $0 < \alpha \leq 1$. Otherwise the exploration is successful and returns a predictor of type $\theta \in [\underline{\theta}, \frac{\pi}{2}]$ with probability $\phi(.)$.[12] After each exploration, a speculator can decide (i) to stop searching and trade in period 1 on the predictor she just found or (ii) to start a new exploration in the hope of finding an even better predictor. We assume that there is no limit on the number of explorations.

It is worth stressing that speculators observe the realization of their chosen predictor, s_θ, in period 1, *not* period 0. In period 0, they just choose the type (quality) of the predictor whose realization they will observe at date 1. A predictor can be viewed as a particular combination of variables from various datasets (e.g., past earnings, satellite images and consumer transactions data) that forecast the payoff of the asset. One exploration consists in testing the predicting power of a particular combination with prediction tools (e.g., regression analysis or machine learning techniques). For instance, one can interpret each exploration as collecting various variables and running a regression of the asset payoff (e.g., stock earnings) on these variables. The estimates of the coefficients of this regression can then be used to compute the predicted value of the regression, s_θ, at date 1 after observing the realization of the variables used in this regression at this date.[13]

Thus, a predictor does not need to be interpreted as a single variable. It can be viewed

precision.

[12]When they find a predictor, speculators perfectly observe its quality, $\tau(\theta)$. Thus, there is no uncertainty on whether a predictor is spurious or not. In reality, the quality of predictors is uncertain (see Harvey (2017)). We leave the analysis of this case for future research.

[13]In this approach, the R^2 of the regression is a measure of the quality of the predictor. Indeed, the theoretical R^2 of a regression of ω on s_θ (i.e., $1 - \mathsf{Var}[\omega \mid s_\theta]/\mathsf{Var}[\omega]$) is equal to $\cos^2(\theta)$. Thus, the higher the quality of a predictor, the higher the R^2 of a regression of the asset payoff on the predictor. In other words, searching for predictors of high quality in the model is the same thing as searching for predictors with high R^2s.

as a combination of variables whose weights have been optimally chosen to minimize the predictor's forecasting error in-sample. In this interpretation, speculators can try to improve the quality of their predictors by trying new combinations (e.g., by buying datasets containing new variables).

As more datasets become available ("data abundance"), the number of possible combinations of variables that one can use to predict asset payoffs increases. This evolution has two consequences controlled by parameters $\underline{\theta}$ and α in the model. First, it pushes back the "data frontier", i.e., it increases the chance (at least weakly) of finding even more informative predictors than those existing before. We refer to this dimension of data abundance as the "hidden gold nugget effect." For instance, by combining satellite images of parking lots at Walmart with credit card transactions data and more traditional accounting data, one might be able to find more informative predictors of future earnings for Walmart than using accounting data alone. This dimension of data abundance is controlled by $\underline{\theta}$ in our model: When $\underline{\theta}$ decreases, the quality of the best predictor (the "hidden gold nugget"), denoted $\tau^{max} \equiv \tau(\underline{\theta})$, improves.

Second, the share of combinations that yield informative predictors might fall as the number of all possible combinations explodes. For instance, there are myriads of ways in which one could combine traffic data in large cities with other data to predict economic growth. However, a few are likely to be informative and discovering these combinations take time. We refer to this dimension of data abundance as the "needle in the haystack problem."[14] It is controlled by α in our model: As α decreases, each round of exploration is less likely to be successful as if the share of informative predictors was falling.[15]

Finally, parameter c represents the cost of exploring a specific dataset to identify a predictor. Greater computing power reduces this cost. For instance, with more powerful computers, one can explore more datasets in a fixed amount of time. So the time cost of data mining is smaller. Thus, we analyze the effect of progress in computing power by considering the effect of a decrease in c on the equilibrium.

[14] Agrawal et al. (2019) discusses a related problem for the generation of new scientific ideas. Specifically, as the space of possible combinations of existing ideas to create new ones enlarges, it becomes more difficult to identify new useful combinations. One can think of the search for predictors at date 0 as a search for new "ideas" to forecast asset payoff. Each new idea is characterized by its forecasting power.

[15] See for instance "The quant fund investing in humans not algorithms" (AlphaVille, Financial Times, December 6, 2017), reporting discussions with a manager from TwoSigma noting that: *"Data are noise. Drawing a tradable signal from that noise, meanwhile, takes work, since the signal is continuously evolving [...] Crucially, Duncombe added, there's qualitative data decay going on too. Back in the day, star managers may have had access to far smaller data sets, but the data in hand was of much higher quality."*

We focus on equilibria in which each speculator follows an optimal stopping rule θ_i^*. That is, speculator i stops searching for new predictors once she finds a predictor with type $\underline{\theta} \leq \theta < \theta_i^*$ (a predictor of sufficiently high quality in the feasible range). We denote by $\Lambda(\theta_i^*; \underline{\theta}, \alpha)$ the likelihood of this event (the probability of success) for speculator i in a given search round. That is:

$$\Lambda(\theta_i^*; \underline{\theta}, \alpha) \equiv \alpha \Pr(\theta \in [\underline{\theta}, \theta_i^*]) = \alpha \times (\Phi(\theta_i^*) - \Phi(\underline{\theta})) \tag{2}$$

Thus, a decrease in $\underline{\theta}$ raises the likelihood of finding a predictor in a given exploration, holding α constant. This effect captures the idea that while data abundance might reduce the fraction of informative datasets, it increases the chance of finding a good predictor once one has identified an informative dataset.

As the outcome of each exploration is random, the realized number of explorations varies across speculators (even if they use the same stopping rule). We denote by n_i the realized number of search rounds for speculator i. This number follows a geometric distribution with parameter $\Lambda(\theta_i^*; \underline{\theta}, \alpha)$. Thus, the expected number of explorations for a given speculator (a measure of her search intensity) is:

$$\mathsf{E}[n_i] = \Lambda(\theta_i^*; \underline{\theta}, \alpha)^{-1}. \tag{3}$$

To simplify the exposition, we assume that speculators cannot "store" predictors that they turn down (i.e., the search for predictors is without recall). We show in Section 6.1 of the online appendix that this assumption is innocuous: The equilibrium of the model is identical if, when they decide to stop searching, speculators have the option to pick the best predictor obtained up to this point.

A last remark is in order. In our model, launching a new exploration does not guarantee that one will necessarily obtain a better predictor than in previous explorations. At the first glance, this may look counter-intuitive because one might think that as speculators observe more predictors, they should be able to obtain an increasingly precise signal about the asset payoff (e.g., by just taking the average of all signals). However, at date 0, each exploration returns the type of a particular predictor, *not its realization* (signals are observed only at date 0). And, as previously explained, we see an exploration as experimenting with a new combination of variables (a new "investment idea") to build a

predictor of the asset payoff. As this combination is new, it does not necessarily have a higher forecasting power than previous combinations.

The trading stage. Trading begins after *all* speculators find a predictor with satisficing quality. At the beginning of period 1, each speculator observes the realization of her predictor, s_θ and chooses a trading strategy, i.e., a demand schedule, $x_i(s_\theta, p)$, where, p, is the asset price in period 1.

As in Vives (1995), speculators trade with noise traders and risk-neutral market makers. Noise traders' aggregate demand is price-inelastic and equal to η, where $\eta \sim \mathcal{N}(0, \nu^2)$ (η is independent of ω and errors' in speculators' signals). Market-makers observe investors' aggregate demand, $D(p) = \int x_i(s_\theta, p)di + \eta$ and behave competitively. The equilibrium price, p^* is equal to their expectation of the asset payoff conditional on aggregate demand from noise traders and speculators:

$$p^* = \mathsf{E}\left[\omega \,|D(p^*)\right]. \tag{4}$$

Speculators' objective function. At $t = 2$, the asset pays off and speculator i's final wealth is

$$W_i = x_i(s_\theta, p)(\omega - p) - n_i c. \tag{5}$$

The number of explorations for speculator i, n_i, is independent from the asset payoff, its price, and the realization of the speculator's predictor, s_θ, because n_i is determined in period 0, before the realizations of these variables. Thus, the ex-ante expected utility of a speculator can be written:

$$\mathsf{E}\left[-\exp(-\rho W_i)\right] = \underbrace{\mathsf{E}\left[-\exp(-\rho(x_i(s_\theta, p)(\omega - p))\right]}_{\text{Expected Utility from Trading}} \times \underbrace{\mathsf{E}\left[\exp(\rho(n_i c))\right]}_{\text{Expected Utility Cost of Exploration}} \tag{6}$$

The first term in this expression represents the ex-ante expected utility that a speculator derives from trading gross of her total exploration cost while the second term represents the expected utility of the total cost paid to find a predictor (we call it the expected utility cost of exploration). The expected utility from trading depends both on the investor's optimal trading strategy ($x_i(s_{\theta,i}, p)$) and her optimal stopping rule (θ_i^*) because this rule determines the distribution of s_θ. The expected utility cost of exploration depends on the

speculator's stopping rule, θ_i^*, because it determines the distribution of n_i. In the existing literature (e.g., Grossman and Stiglitz (1980)), $n_i = 1$ (investors pays a cost and gets a signal). In our model, n_i is random and its distribution is controlled by the speculator through her search behavior.

Each speculator chooses her stopping rule, θ_i^*, and her trading strategy, $x_i(s_{\theta,i}, p)$, to maximize her ex-ante expected utility.

4. Equilibrium Data Mining

4.1 Equilibrium

We focus on symmetric equilibria in which all speculators choose the same stopping rule, θ^*. We solve for such an equilibrium as follows. First, we solve for the equilibrium of the trading stage in period 1 taking θ^* as given and we deduce the ex-ante expected utility achieved by speculator i when she chooses a predictor of type θ in period 0. We then observe that a speculator should stop searching as soon as she finds a predictor such that the expected utility of trading on this predictor is larger than or equal to the expected utility she can obtain by launching a new round. The optimal stopping rule of each investor, $\theta_i^*(\theta^*)$, is such that this condition holds as an equality (so that the speculator is just indifferent between searching more or stopping). Finally, we pin down θ^* by observing that, in a symmetric equilibrium, each speculator's best response to other speculators' stopping rule, θ^*, must be identical, i.e., $\theta_i^*(\theta^*) = \theta^*$.

Equilibrium of the asset market in period 1**.** The outcome of the exploration phase is characterized by the distribution of the predictors' types found by speculators. Let $\phi^*(\theta; \theta^*; \underline{\theta}, \alpha)$ be this distribution given that speculators' follow the stopping rule θ^*:

$$\phi^*(\theta; \theta^*; \underline{\theta}, \alpha) = \frac{\phi(\theta)}{\Lambda(\theta^*; \underline{\theta}, \alpha)}. \tag{7}$$

This distribution characterizes the heterogeneity of speculators' predictors in equilibrium. We denote the *average* quality of predictors across all speculators in period 1 by $\bar{\tau}(\theta^*, \underline{\theta}, \alpha) \equiv \mathsf{E}\left[\tau(\theta) | \, \underline{\theta} \leq \theta \leq \theta^*\right]$ and we make the following assumption on the distribution $\phi(\cdot)$:

A.1: The distribution of predictors' type, $\phi(.)$, is such that for all $\theta^* > 0$, $\bar{\tau}(\theta^*; 0, \alpha)$ exists.

This technical condition just guarantees that the equilibrium remains well defined even when $\underline{\theta} = 0$.[16] Proposition 1 provides the equilibrium of the asset market in period 1.

Proposition 1. *In period* 1*, the equilibrium trading strategy of a speculator with type* θ *is:*

$$x^*(s_\theta, p) = \frac{\mathsf{E}[\omega|s_\theta, p] - p}{\rho\, \mathsf{Var}[\omega|s_\theta, p]} = \frac{\tau(\theta)}{\rho\sigma^2}\left(\hat{s}_\theta - p\right), \tag{8}$$

where $\hat{s}_\theta = \omega + \tau(\theta)^{-1/2}\varepsilon_\theta$ *and the equilibrium price of the asset is:*

$$p^* = \mathsf{E}[\omega|D(p)] = \lambda(\theta^*)\xi. \tag{9}$$

where

$$\xi = \omega + \rho\sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^{-1}\eta, \quad \textit{and} \quad \lambda(\theta^*) = \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2}{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2 + \rho^2\sigma^2\nu^2}, \tag{10}$$

This result extends Proposition 1.1 in Vives (1995) to the case in which speculators have signals of heterogenous precisions (determined by their θ in our model). The predictor s_θ is informationally equivalent to the predictor $\hat{s}_\theta = \omega + \tau(\theta)^{-1/2}\varepsilon_\theta$. A speculator's optimal position in the asset is equal to the difference between $\hat{s}_\theta$ and the price of the asset (her expected dollar return) scaled by a factor that increases with the quality of the predictor and decreases with the speculator's risk aversion. The scaling factor measures the speculator's aggressiveness in trading on her predictor. Speculators with predictors of higher quality trade more aggressively on their signal because they face less risk (their forecast of the asset payoff is more precise).

The total demand for the asset ($D(p)$) aggregates speculators' orders and therefore reflects their information. Observing this demand is informationally equivalent to observing the signal ξ, whose informativeness increases with the average quality of speculators' predictors, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$. Thus, the market maker can form a more precise forecast of the asset payoff and the asset price is therefore more informative when the average quality of speculators' predictors, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$, is higher. Formally, let measure the informativeness

[16] Indeed, for some distributions of predictors' type, $\phi(.)$, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$ can diverge because $\tau(\theta)$ goes to infinity when θ goes to zero. Assumption A.1 means that we exclude these distributions from our analysis.

of the asset price by $\mathcal{I}(\theta^*;\underline{\theta},\alpha) = \mathsf{Var}[\omega \mid p^*]^{-1}$ as in Grossman and Stiglitz (1980). Using Proposition 1, we obtain:

$$\mathcal{I}(\theta^*;\underline{\theta},\alpha) = \tau_\omega + \frac{\bar{\tau}(\theta^*;\underline{\theta},\alpha)^2\tau_\omega^2}{\rho^2\nu^2}, \tag{11}$$

where $\tau_\omega = 1/\sigma^2$ is the precision of speculators' prior about the asset payoff. As expected, the asset price is more informative when the average quality of speculator's predictors increases. Thus, the informativeness of the asset price is inversely related to θ^* because $\bar{\tau}(\theta^*;\underline{\theta},\alpha)$ decreases with θ^*. Thus, *other things equal*, price informativeness is smaller when speculators chooses a less stringent stopping rule for the quality of the predictors on which they trade.

Equilibrium of the exploration phase. Using the characterization of the equilibrium of the asset market, we compute a speculator's expected utility from trading ex-ante, i.e., before observing the realization of her predictor and the equilibrium price, when her predictor has type θ and other speculators follow the stopping rule θ^*. We denote this ex-ante expected utility by $g(\theta,\theta^*)$ and refer to it as the trading value of a predictor with type θ. Formally:

$$g(\theta,\theta^*) \equiv \mathsf{E}\left[-\exp(-\rho(x^*(s_\theta,p^*)(\omega-p^*)) \mid \theta_i = \theta\right]. \tag{12}$$

Lemma 1. *In equilibrium, the trading value of a predictor with type θ is:*

$$g(\theta,\theta^*) = -\left(1+\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta,p]-p]}{\mathsf{Var}[\omega|s_\theta,p]}\right)^{-\frac{1}{2}} = -\left(1+\frac{\tau(\theta)\tau_\omega}{\mathcal{I}(\theta^*;\underline{\theta},\alpha)}\right)^{-\frac{1}{2}}. \tag{13}$$

Thus, the trading value of a predictor increases with its quality and decreases with the informativeness of the asset price. Thus, it is inversely related to the average quality of predictors used by speculators. Hence, the value of a given predictor for a speculator depends on the search strategy followed by other speculators: It is smaller if other speculators are more demanding for the quality of their predictors (i.e., when θ^* decreases).[17]

Armed with Lemma 1, we can now derive a speculator's optimal stopping rule given that other speculators follow the stopping rule θ^*. Let $\widehat{\theta}_i$ be an arbitrary stopping rule

[17]This means that speculators' information acquisition strategies are substitutes in our model, as usual in models of information acquisition in finance.

for speculator i. The speculator's continuation utility (the expected utility of launching a new round of exploration) after turning down a predictor is:

$$J(\widehat{\theta}_i, \theta^*) = \exp(\rho c) \left(\Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha) \, \mathsf{E} \left[g(\theta, \theta^*) \left| \underline{\theta} \leq \theta \leq \widehat{\theta}_i \right. \right] + (1 - \Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha)) J(\widehat{\theta}_i, \theta^*) \right) \quad (14)$$

The first term ($\exp(\rho c)$) in eq.(14) is the expected utility cost of running an additional search. The second term is the likelihood that the next exploration is successful times the average trading value of a predictor conditional on the type of this predictor being satisficing (i.e., in $[\underline{\theta}, \widehat{\theta}_i]$). Finally, the third term is the likelihood that the next exploration is unsuccessful time the speculator's continuation utility when she turns down a predictor. Solving eq.(14) for $J(\widehat{\theta}_i, \theta^*)$, we obtain:

$$J(\widehat{\theta}_i, \theta^*) = \underbrace{\left[\frac{\exp(\rho c) \Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha)}{1 - \exp(\rho c)(1 - \Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha))} \right]}_{\text{Expected Utility Cost from Exploration}} \times \underbrace{\mathsf{E} \left[g(\theta, \theta^*) | \, \underline{\theta} \leq \theta \leq \widehat{\theta}_i \right]}_{\text{Expected Utility from Trading}} \quad (15)$$

The continuation value of the speculator when she turns down a predictor does not depend on the outcomes of past explorations because these outcomes do not affect the speculator's opportunity set in future explorations. Thus, $J(\widehat{\theta}_i, \theta^*)$ is also the speculator's ex-ante expected utility before starting any exploration in period 0. As explained previously, it is the product of the expected utility cost from explorations and the expected utility from trading.

Now suppose that speculator i has obtained a predictor with quality θ. If the speculator stops exploring the data at this stage, her expected utility is $g(\theta, \theta^*)$ (her cost of exploration to obtain this predictor is sunk). If instead the speculator decides to launch a new round of exploration, her expected utility is $J(\widehat{\theta}_i, \theta^*)$. Thus, her optimal decision is to stop searching for a predictor if $g(\theta, \theta^*) \geq J(\widehat{\theta}_i, \theta^*)$ and to keep searching otherwise. As $g(\theta, \theta^*)$ decreases with θ, the optimal stopping rule of the speculator, θ_i^*, is the value of θ such that the speculator is just indifferent between these two options:

$$g(\theta_i^*, \theta^*) = J(\theta_i^*, \theta^*). \quad (16)$$

The solution to this equation, $\theta_i^*(\theta^*) = \theta^*$, is unique (see the proof of Proposition 2). In

a symmetric equilibrium, it must be that $\theta_i^*(\theta^*) = \theta^*$. We deduce that θ^* solves:

$$g(\theta^*, \theta^*) = J(\theta^*, \theta^*). \tag{17}$$

Using the expression for $J(., \theta^*)$ in eq.(14), we can equivalently rewrite this equilibrium condition as:

$$F(\theta^*) = \exp(-\rho c), \tag{18}$$

where:

$$F(\theta^*) \equiv \alpha \int_{\underline{\theta}}^{\theta^*} r(\theta, \theta^*)\phi(\theta)d\theta + (1 - \Lambda(\theta^*; \underline{\theta}, \alpha)), \quad \text{for } \theta^* \in \left[\underline{\theta}, \frac{\pi}{2}\right], \tag{19}$$

with

$$r(\theta, \theta^*) \equiv \frac{g(\theta, \theta^*)}{g(\theta^*, \theta^*)} = \left(\frac{\tau(\theta^*)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\tau(\theta)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}\right)^{\frac{1}{2}}, \tag{20}$$

where the second equality in eq.(20) follows from eq.(13). Observe that Assumption A.1 implies that $F(\theta^*)$ is well defined even when $\underline{\theta} = 0$. The next proposition shows that there is a unique interior solution (i.e., $\theta^* \in (\underline{\theta}, \frac{\pi}{2})$) to the equilibrium condition (18) when c is small enough.

Proposition 2. *There is a unique symmetric interior equilibrium of the exploration phase in which all speculators are active (i.e., a unique stopping rule such that $\underline{\theta} < \theta^* < \pi/2$ common to all speculators) if and only if $F(\pi/2) < \exp(-\rho c) < 1$.*

When $\exp(-\rho c) \leq F(\pi/2)$, there is no symmetric interior equilibrium. However, in this case, one can build an equilibrium in which only a fraction of all speculators are active, i.e., search for a predictor and trade (provided that c is not too large of course). In this equilibrium, active speculators search for a predictor with a stopping rule equal to $\theta^* = \pi/2$ while others remain completely inactive (do not search and do not trade). Moreover, the fraction of speculators who are active is such that all speculators are indifferent between being active or not. Henceforth,we focus on the case in which the equilibrium is interior (i.e., $F(\pi/2) < \exp(-\rho c) < 1$ because (i) we are interested in what happens when the cost of exploration becomes small and (ii) this shortens the exposition.

4.2 Data abundance, computing power and optimal data mining.

We now analyze how data abundance (a decrease in $\underline{\theta}$ and/or α) and computing power (a decrease in c) affect the quality of the worst predictor on which speculators trade in equilibrium, i.e., $\tau(\theta^*)$. This is important because this quality determines the range of predictors used in equilibrium and ultimately several equilibrum outcomes of interest (e.g., asset price informativeness).

Proposition 3. *A decrease in the cost of exploration, c, always reduces the stopping rule θ^* used by speculators in equilibrium ($\partial\theta^*/\partial c > 0$). Thus, greater computing power raises the quality, $\tau(\theta^*)$, of the worst predictor used by speculators in equilibrium.*

The economic mechanism for this finding is as follows. Holding θ^* constant, a decrease in the per-exploration cost, c, directly reduces the expected utility cost of launching a new exploration after finding a predictor (the first term in bracket in eq.(15)). Hence, it raises the value of searching for another predictor after finding one (i.e., $J(\theta^*, \theta^*)$). This direct effect induces speculators to be more demanding for the quality of their predictor and therefore works to decrease θ^*. One indirect consequence of this behavior is that, on average, speculators trade more aggressively on their signal (the "competition effect") because their predictors are better on average and therefore they face less uncertainty on the asset payoff. As a result, price informativeness increases. This indirect effect reduces the expected utility from trading on a satisficing predictor (the second term in bracket in eq.(15)) and therefore dampens the direct positive effect of a decrease in c on the value of searching for a better predictor after finding one. However, it is never strong enough to fully offset it.

We now consider the effect of data abundance on speculators' optimal stopping rule. Remember that data abundance has two consequences in the model: (i) it pushes back the data frontier by raising the quality of the best predictor and (ii) it increases the risk for speculators of using datasets which, after exploration, proves to be useless (the needle in the haystack problem).

Proposition 4.

1. *A decrease in the fraction of informative datasets, α, always increases speculators' stopping rule, θ^*, in equilibrium ($\partial\theta^*/\partial\alpha < 0$). Thus, the needle in the haystack*

problem reduces the quality, $\tau(\theta^*)$, of the worst predictor used by speculators in equilibrium.

2. *The effect of a decrease in $\underline{\theta}$ on speculators' stopping rule is ambiguous. However, when $\underline{\theta}$ is less than $\underline{\theta}^{tr}(c)$, a decrease in $\underline{\theta}$ always increases speculators' stopping rule in equilibrium ($\partial\theta^*/\partial\underline{\theta} < 0$ for $\underline{\theta} < \underline{\theta}^{tr}(c)$) and reduces the quality, $\tau(\theta^*)$, of the worst predictor used by speculators in equilibrium.*

When the needle in the haystack problem becomes more acute, speculators become less demanding for the quality of their predictors. Intuitively, a drop in α increases the expected utility cost of launching a new exploration after finding a predictor (the first term in bracket in eq.(15)) because it reduces the likelihood of finding a predictor in a given exploration (Λ). Thus, after turning down a predictor, speculators expect to go through a larger number of explorations rounds before finding a satisficing predictor, which increases their total cost of search. This direct effect induces speculators to be less demanding for the quality of their predictor and therefore works to increase θ^* (reduce $\tau(\theta^*)$). Indirectly, this behavior reduces asset price informativeness and therefore raises the expected utility from trading on a satisficing predictor (the second term in bracket in eq.(15)), which alleviates the direct negative effect of a decrease in α on the value of searching for a better predictor after finding one. However, this indirect effect is never strong enough to fully offset the direct effect. In sum, qualitatively, the effect of a drop in α is similar to that of an increase in the per exploration cost.[18]

The effect of pushing back the data frontier on speculators' stopping rule is more complex. Counterintuitively, it can lead speculators to trade on predictors of worse quality, even though the quality of the best predictor increases. The reason is as follows. On the one hand, pushing back the data frontier increases the chance of finding a satisficing predictor holding the search strategy, θ^* constant ($\Lambda(\theta^*; \underline{\theta}, \alpha)$ increases when $\underline{\theta}$ goes down). This effect reduce the expected number of rounds required to find a predictor and therefore reduces the expected utility cost of searching for a new predictor after rejecting one. Therefore, it increases the continuation value of searching for a predictor (see eq.(15).

[18]Given this, one might be tempted to capture the needle in the haystack effect by just considering the effect of increasing c (on the ground that it becomes more costly to find good datasets). But this approach is inconsistent with the argument that progress in information technology has reduced information processing costs. This point illustrates the importance of having separate parameters to capture the effects of (i) greater information processing power (a decrease in c in our model) on the one hand and (ii) data abundance on the other hand.

On the other hand, a push back of the data frontier affects the expected utility from trading for two reasons. First, it gives the possibility to obtain more informative predictors than those existing before ("the hidden gold nugget effect"), which raises the expected utility from trading on a satisficing predictor. Second, it increases price informativeness (other things equal, $\mathcal{I}(\theta^*;\underline{\theta},\alpha)$ increases when $\underline{\theta}$ decreases) because speculators who obtain the most informative predictors trade even more aggressively than before the change in the data frontier. As a result, speculators' aggregate demand and therefore the asset price are more informative, which reduces the value of being informed ("the competition effect"). This effect reduces the expected utility from trading on a satisficing predictor. Thus, the sign of a change in the data frontier (holding θ^* constant) on the expected utility from trading is ambiguous.

To analyze this more formally, we differentiate the expected utility from trading, $\mathsf{E}\left[g(\theta,\theta^*)|\,\underline{\theta}\leq\theta\leq\theta^*\right]$, with respect to $\underline{\theta}$ (holding θ^* constant):

$$\frac{\partial\,\mathsf{E}\left[g(\theta,\theta^*)|\,\underline{\theta}\leq\theta\leq\theta^*\right]}{\partial\underline{\theta}}$$
$$=\frac{\alpha\phi(\underline{\theta})}{\Lambda(\theta^*;\underline{\theta},\alpha)}\left[\underbrace{\mathsf{E}\left[g(\theta,\theta^*)|\,\underline{\theta}\leq\theta\leq\theta^*\right]-g(\underline{\theta},\theta^*)}_{\text{Hidden Gold Nugget Effect; }<0}+\underbrace{\int_{\underline{\theta}}^{\theta^*}\frac{\partial g(\theta,\theta^*)}{\partial\underline{\theta}}\phi(\theta)d\theta}_{\text{Competition Effect;}>0}\right] \tag{21}$$

When $\underline{\theta}$ becomes small enough, the competition effect dominates the hidden gold nugget effect and the expected utility from trading on a satisficing predictor drops. The second part of Proposition 4 shows that there is always a sufficiently low value of $\underline{\theta}$ such that this drop more offsets the reduction in the expected utility cost of finding a predictor. When this happens, pushing back the frontier further reduces the continuation value of exploration. Hence, speculators choose a less stringent stopping rule in equilibrium and some optimally choose to trade on less informative predictors ($\tau(\theta^*)$ decreases).

We illustrate Proposition 4 by considering two particular specifications of the distribution for θ. In specification 1, we assume that $\phi(\theta)=3\cos(\theta)\sin^2(\theta)$ while in specification 2 we assume that $\phi(\theta)=5\cos(\theta)\sin^4(\theta)$. These specifications are convenient because they enable us to compute all variables of interest in closed forms (see Section 4 in the internet appendix). In the second specification, the distribution of θ has a much fatter right-tail in the first case (see Section 4 in the internet appendix).[19] Figure 2 below shows the effect

[19] Assumption A.1 is satisfied in both examples.

of a change in the exploration cost (c) and the data frontier ($\underline{\theta}$) on the equilibrium value of θ^*. In either case, as implied by Proposition 4, a push back of the data frontier initially raises the quality of the worst predictor used by speculators in equilibrium (reduces θ^*) but, eventually, at some point this effect is reversed.

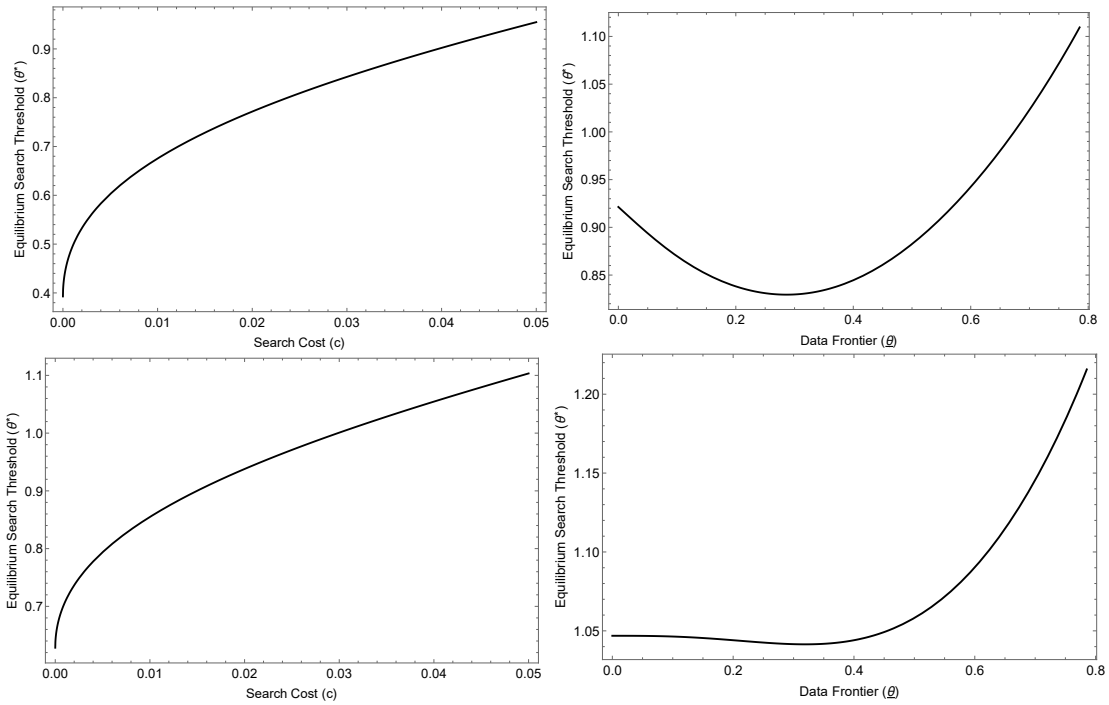

Figure 2: The left hand-side graphs plot the equilibrium search threshold, θ^*, as a function of the search cost, c (other parameter values are $\underline{\theta} = \pi/8, \rho = \sigma^2 = \nu^2 = 1$). The right hand-side graphs plot the equilibrium search threshold, θ^*, as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.03, \rho = \sigma^2 = \nu^2 = 1$). In the two upper graphs, we assume that $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ (Case 1) while in the two lower graphs we assume that $\phi(\theta) = 5\cos(\theta)\sin^4(\theta)$ (Case 2).

Proposition 5. *In equilibrium, the quality of the worst predictor used in equilibrium, $\tau(\theta^*)$, increases with the volume of noise trading, ν^2, or the volatility of the asset payoff, σ^2.*

An increase in the volume of noise trading reduces the informativeness of the equilibrium price. This effect raises the expected value of trading, holding the search policy, θ^*, constant. Thus, the continuation value from searching increases and speculators become therefore more demanding for their predictors (θ^* decreases). The intuition for the effect of the volatility of the asset is identical.

5. Testable Implications

5.1 Data Abundance, Computing Power, and Managerial Skills

As explained in the previous section, the model has implications for the effects of data abundance and computing power on the distribution of the quality of predictors used by speculators in equilibrium, in particular the lower bound of this distribution $\tau(\theta^*)$. To test these implications, one can use data on active funds' holdings and their returns on these holdings (e.g., as in Kacperczyk et al. (2016)) and regress the position of each fund (speculator) in a given asset ($x_i(s_\theta, p^*)$ in the model), at a given point in time on their return on this position ($(\omega - p^*)$ in the model). In the model, the coefficient of this regression, β_θ, is:

$$\beta_\theta = \frac{\mathsf{Cov}(x(s_\theta, p^*), \omega - p^*)}{\mathsf{Var}[\omega - p^*]} = \frac{\tau(\theta)}{\rho}, \tag{22}$$

where the last equality follows from Proposition 1. Intuitively, β_θ is a measure of a speculator's stock picking ability or investment "skills".[20] Equation (22) shows that, holding risk aversion constant, a ranking of speculators based on their stock picking ability (measured by β_θ) is identical to a ranking based on the (unobservable) quality of their predictors, $\tau(\theta)$. This is intuitive: Speculators with better predictors should display a better stock picking ability.

Thus, one could test the implications of Propositions 3 and 4 by ranking speculators (e.g., quantitative asset managers) based on their stock picking ability (measured by βs) and test whether shocks to computing power or data abundance have the effects predicted by Propositions 3 and 4.[21] For instance, one could test whether positive shocks to computing power increase the stock picking ability (measured by β) of the funds with the lowest βs' (say in the lowest decile) while positive shocks to data abundance (e.g., the availability of new alternative data as in Zhu (2019)) have the opposite effect (even though they may increase the stock picking ability of the best performing funds). One

[20] Kacperczyk et al. (2016) measure mutual funds' stock picking ability in a similar way. See Section 2.1 in their paper.

[21] Alternatively, one could proceed as in Kacperczyk and Seru (2007) to measure asset managers' investment skills and rank these. Specifically, Kacperczyk and Seru (2007) measures the precision of asset managers' signals (their "skill") by the sensitivity of their holdings to public information. The higher is this sensitivity, the lower is the precision of a manager's private signals. This would also be the case in a simple extension of our model in which speculators receive a public signal at date 1 in addition to their private signal s_θ.

could also test whether the *difference* between the stock picking ability of speculators with the lowest and highest ability is reduced in periods of heightened fundamental volatility or noise trading, as implied by Proposition 5.

Kacperczyk and Seru (2007) (and others) find that there is considerable heterogeneity in asset managers' skills (see their Table I). Our model suggests that one source of heterogeneity might be managers' luck in their search for a predictor, rather than differences in innate abilities to find investment ideas or effort. Indeed, in our model, all speculators are ex-ante identical and choose the same effort in terms of search in the sense that their stopping rule (and therefore expected total cost of search) is identical. Yet, they end up trading on predictors of different qualities because the outcome of the search process is random. This implies in particular that a speculator might end up paying a large total search cost ($n_i c$) and yet appear as having low skills (trading on a signal of poor quality).

5.2 Data Abundance, Computing Power, and Asset Price Informativeness

Progress in information technologies have improved investors' ability to forecast asset payoffs in two ways. On the one hand, these technologies reduce the cost of filtering out noise from raw data (e.g., greater computing power enables asset managers to use powerful statistical techniques, such as deep neural networks, to form their forecasts). On the other hand, they allow to collect and store increasing volume of data. Propositions 6 and 7 show that these two different distinct dimensions of technological progress do not affect asset price informativeness in the same way.

Proposition 6. *In equilibrium, an increase in computing power (a decrease in c) raises the average quality of speculators' predictors and therefore price informativeness.*

Greater computing power induces speculators to be more demanding for the quality of their predictors (to put more effort in the search of good predictors) because it reduces the cost of exploring new data to obtain a predictor (see Proposition 3). Thus, speculators obtain signals of higher quality on average. Hence, on average, they trade more aggressively on their signals, their aggregate demand for an asset becomes more informative and, for this reason, price informativeness increases (see eq.(11)).

Proposition 7.

1. *In equilibrium, an improvement in the quality of the most informative predictor (a decrease in $\underline{\theta}$) raises the average quality of speculators' predictors and therefore price informativeness.*

2. *In equilibrium, a decrease in the proportion of informative datasets (a decrease in α) reduces the average quality of speculators' predictors and therefore price informativeness.*

Thus the effect of data abundance on price informativeness is ambiguous. Holding α constant, data abundance (a decrease in $\underline{\theta}$) improves asset price informativeness, even when it induces speculators to be less demanding for the quality of their predictors (i.e., when a decrease in $\underline{\theta}$ reduces $\tau(\theta^*)$; see Proposition 4). The reason is that the negative effect of the drop in the quality of the worst predictor used in equilibrium (if it happens) on the average quality of speculators' signals is never sufficient to offset the positive effect of the improvement in the quality of the best predictor in equilibrium. As a result, a push back of the data frontier raises the average quality of predictors and speculators' average trading aggressiveness. In contrast, holding $\underline{\theta}$ constant, data abundance (a decrease in α) leads speculators to be less demanding for the quality of their predictors. As a result, the average quality of predictors drops, speculators' aggregate demand is less informative and therefore price informativeness drops.

In reality, data abundance is likely to both push back the data frontier (reduce $\underline{\theta}$) and exacerbate the needle in the haystack problem (reduce α). As a result, the net effect of data abundance on the long run evolution of asset price informativeness is ambiguous. Figure 3 illustrates this point with a numerical example in which we assume that α increases with $\underline{\theta}$ (specifically, we assume that $\alpha = \min\{1, 0.32 + 0.8 \times \underline{\theta}\}$). Thus, data abundance (a drop in $\underline{\theta}$) generates both an increase in the quality of the best predictor and a needle in the haystack problem. As shown by Figure 3, when these two dimensions of data abundance operate jointly, price informativeness initially rises with data abundance (starting from a large $\underline{\theta}$) until it reaches a peak after which it decreases. The reason is that when $\underline{\theta}$ becomes small, both dimensions of data abundance induce speculators to be less demanding for the quality of their predictors (see Proposition 4), which eventually impairs price informativeness.

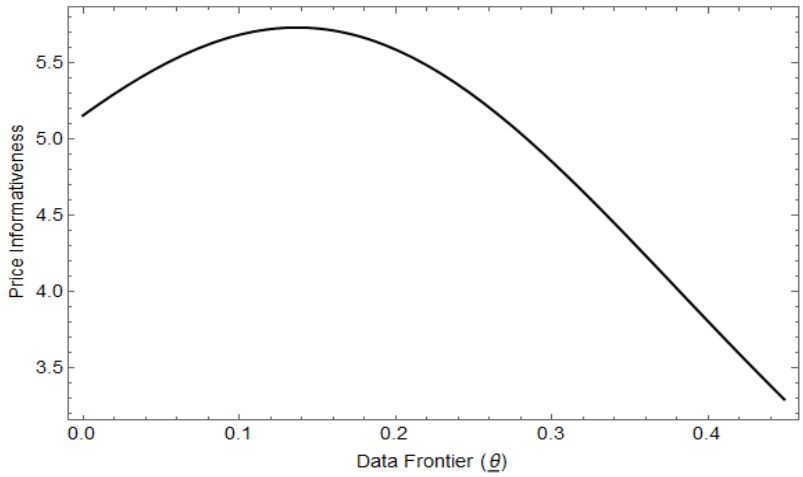

Figure 3: This graph shows the evolution of price informativeness in equilibrium, $\mathcal{I}(\theta^*, \underline{\theta})$ as a function of the data frontier, $\underline{\theta}$ when $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ and $\alpha = \min\{1, 0.32 + 0.8 * \underline{\theta}\}$. Other parameter values, $c = 0.03, \rho = 1, \sigma^2 = 1, \nu^2 = 1$.

Interestingly, consistent with these implications, empirical findings regarding the effect of progress in information technologies on price informativeness are ambiguous. For instance, Bai et al. (2016) find that the price stocks in the S&P500 has become more informative since the 60s while Farboodi et al. (2019) find the opposite patterns for all stocks, except for large growth stocks. Using controlled experiments, Zhu (2019) finds that the availability of alternative data (satellite images and consumer transactions data) improves stock price informativeness while Goldstein et al. (2020) find a drop in the sensitivity of corporate investment to stock prices after the digitization of firms' regulatory filings, which they explain by a decline in the production of private information. Our results suggests that designing tests that only vary computing power holding data abundance constant or vice versa would help to make progress in understanding why information technologies matter for asset price informativeness.

5.3 Data abundance, Computing Power and Trading Profits

In this section, we analyze how data abundance and computing power affects the distribution of trading profits for speculators. In equilibrium, the total trading profit ("excess

return"), $\pi(s_\theta)$, of a speculator with type θ on his position in the risky asset is:

$$\pi(s_\theta) = x^*(s_\theta, p^*) \times (\omega - p^*), \tag{23}$$

where $x^*(s_\theta, p^*)$ and p^* are given by eq.(8) and eq.(9), respectively. Using eq.(8), we deduce that:

$$x^*(s_\theta, p^*) = \frac{1}{\rho\sigma^2}\left(\tau(\theta)(\omega - p^*) + \tau(\theta)^{1/2}\varepsilon_\theta\right). \tag{24}$$

Thus, the *expected* trading profit of a speculator with type θ is:

$$\bar{\pi}(\theta) = \mathsf{E}[\pi(s_\theta)|\theta] = \frac{\tau(\theta)}{\rho\sigma^2}\,\mathsf{Var}[\omega - p^* \mid \theta] = \frac{\tau(\theta)}{\rho\sigma^2\mathcal{I}(\theta^*, \underline{\theta})}. \tag{25}$$

It follows that the unconditional expected trading profit of all speculators (the average trading profit across all speculators) is:

$$\mathsf{E}[\bar{\pi}(\theta)] = \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\rho\sigma^2\mathcal{I}(\theta^*, \underline{\theta})} = \frac{1}{\rho\sigma^2}\left(\frac{\tau_\omega}{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)} + \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\rho^2\nu^2}\right)^{-1}, \tag{26}$$

and the variance of trading profits for speculators (the dispersion of trading profits across all speculators) is:

$$\mathsf{Var}[\pi(\theta)] = \frac{\mathsf{Var}[\tau(\theta) \mid \underline{\theta} < \theta < \theta^*]}{\sigma^4\rho^2\mathcal{I}^2(\theta^*, \underline{\theta})}. \tag{27}$$

Empirically, $\mathsf{E}[\pi(\theta)]$ and $\mathsf{Var}[\pi(\theta)]$ could be measured by the cross-sectional mean and variance of trading profits of active funds (for instance in a given quarter).

An increase in the average quality of predictors ($\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$ has an ambiguous effect on speculators' expected profit. On the one hand, this increase improves speculators's stock picking ability (see Section 5.1). On the other hand, it increases asset price informativeness because it makes speculators' aggregate demand more informative. As shown by eq.(26), the first effect raises speculators' expected profit while the second reduces it. Using eq.(26), it is easily shown that the first effect dominates if and only if $\bar{\tau}(\theta^*; \underline{\theta}, \alpha) \leq \tau_\omega\rho^2\nu^2$. Thus, speculators' average expected profit reaches its maximum for $\bar{\tau}(\theta^*(\underline{\theta}, c, \alpha), \underline{\theta}, \alpha) = \tau_\omega\rho^2\nu^2$ if there are values of $(\underline{\theta}, c, \alpha)$ for which this equality holds (we write θ^* as a function of $\underline{\theta}$,c,α to emphasize that it depends on the value of these parameters). We deduce the following result.

Proposition 8. *Suppose* $\underline{\theta} > \tau_\omega\rho^2\nu^2$.

1. *If* $\bar{\tau}(\theta^*(\underline{\theta}, 0, \alpha), \underline{\theta}, \alpha) > \tau_\omega \rho^2 \nu^2$ *then speculators' expected profit is a hump shaped function of* c*, which reaches its maximum for* $c = \hat{c}$ *(characterized in the proof of the proposition). Otherwise, speculators' expected profit decreases with* c *and reaches its maximum for* $c = 0$

2. *If* $\bar{\tau}(\theta^*(0, c, \alpha), 0, \alpha) > \tau_\omega \rho^2 \nu^2$ *then speculators' expected profit is a hump shaped function of* $\underline{\theta}$*, which reaches its maximum for* $\underline{\theta} = \hat{\underline{\theta}}$ *(characterized in the proof of the proposition). Otherwise, speculators' expected profit decreases with* $\underline{\theta}$ *and reaches its maximum for* $\underline{\theta} = 0$*.*

3. *If* $\bar{\tau}(\theta^*(\underline{\theta}, c, 1), \underline{\theta}, 1) > \tau_\omega \rho^2 \nu^2$ *then speculators' expected profit is a hump shaped function of* α*, which reaches its maximum for* $\alpha = \hat{\alpha}$ *(characterized in the proof of the proposition). Otherwise, speculators' expected profit increases with* α *and reaches its maximum for* $\alpha = 1$

Thus, data abundance or greater computing power do not necessarily improve speculators' expected trading profit. Consider first a decrease in c or $\underline{\theta}$. Such a decrease leads speculators to be more demanding for the quality of their predictors and raises the average quality of their signals. However, for this reason, it raises price informativeness. The first effect has a positive effect on speculators' expected profit while the second has a negative effect. The latter effect always dominates when c or $\underline{\theta}$ are small enough. A decrease in α has exactly the opposite effect: It reduces the average quality of speculators' signals and price informativeness. The first effect has a negative effect on speculators' expected profit while the second has a positive effect. The former effect always dominates when α is small enough.

Overall, these findings mean that there is always a point at which further improvements in computing power or data availability reduces speculators' expected profit, either because price informativeness becomes too high ($\underline{\theta} < \hat{\underline{\theta}}$ or $c < \hat{c}$) or because the needle in the haystack problem has a too large negative effect on speculators' incentive to search for good predictors, so that the average quality of their signals falls by a large amount ($\alpha < \hat{\alpha}$).

Now consider the effect of changes in the cost of processing data and data abundance on the dispersion ($\mathsf{Var}[\pi(\theta)]$) of expected trading profits across speculators. Using eq.(27), we obtain the following result.

Proposition 9.

1. *Other things equal, the dispersion of speculators' expected trading profit decreases when the cost of processing data goes down for c small enough ($d\,\mathsf{Var}[\pi(\theta)]/dc > 0$ for c sufficiently close to zero).*

2. *Other things equal, the dispersion of speculators' expected profit increases when the data frontier is pushed back for $\underline{\theta}$ small enough ($d\,\mathsf{Var}[\pi(\theta)]/d\underline{\theta} < 0$ for $\underline{\theta}$ sufficiently close to zero).*

3. *Other things equal, the dispersion of speculators' expected trading profit increases when the fraction of informative datasets decreases (α decreases).*

To understand the first part of the proposition, suppose that $c = 0$. In this case, all speculators search for a predictor until they find one with the highest possible quality, $\theta^* = \underline{\theta}$. As a result, all speculators trade on predictors of the same quality ($\mathsf{Var}[\tau(\theta) \mid \underline{\theta} < \theta < \theta^*] = 0$) and therefore the dispersion of expected trading profits is nil, as can be seen by inspection of the expression for $\mathsf{Var}[\pi(\theta)]$ (eq.(27)). Now consider a small increase in c starting from the situation in which $c = 0$. This increase raises θ^* and therefore the dispersion of the quality of predictors used by speculators ($\mathsf{Var}[\tau(\theta) \mid \underline{\theta} < \theta < \theta^*]$ increases). As a result, the dispersion of trading profits increases as well. This increase is amplified by the fact that price informativeness goes down, which works to increase the dispersion in trading profits as well (see the expression for $\mathsf{Var}[\pi(\theta)]$ in eq.(27)). As these effects still hold for larger values of c, we conjecture that the first part of Proposition 9 holds for all values of c but we have not been able to show it analytically (numerical simulations suggest that our conjecture is correct; see Figure 4 below for an example).

When $\underline{\theta} < \underline{\theta}^{tr}(c)$, the quality of the best predictor increases while the quality of the worst predictor used by speculator decreases when data become more abundant (see Proposition 4). Thus, the range of quality for the predictors used in equilibrium gets wider. This effect increases the dispersion of the quality of predictors used by speculators ($\mathsf{Var}[\tau(\theta)]$ increases), which increases the dispersion of speculators' expected profits, holding price informativeness constant. In equilibrium, price informativeness improves but for $\underline{\theta}$ small enough, this second effect is not sufficient to offset the first. This explains the second part of the proposition.

The effect of a decrease in α is more straightfoward. Indeed, such a decrease leads speculators to be less demanding for the quality of their predictors (θ^* increases when α decreases). Thus, a decrease in α enlarges the dispersion of the quality of speculators' predictors. As it also reduces price informativeness, it follows from eq.(27) that the dispersion of speculators' trading profits increases.

In sum, data abundance and improvements in computing power have similar effects on speculators' expected profits but can have opposite effects on the dispersion of these profits. Figure 4 illustrates this point using the same specifications for the density of θ as in Figure 2. For these specifications, a decrease in the cost of processing data always reduces the dispersion of expected trading profits across speculators. In contrast, the dispersion expected trading profits increases when $\underline{\theta}$ decreases.

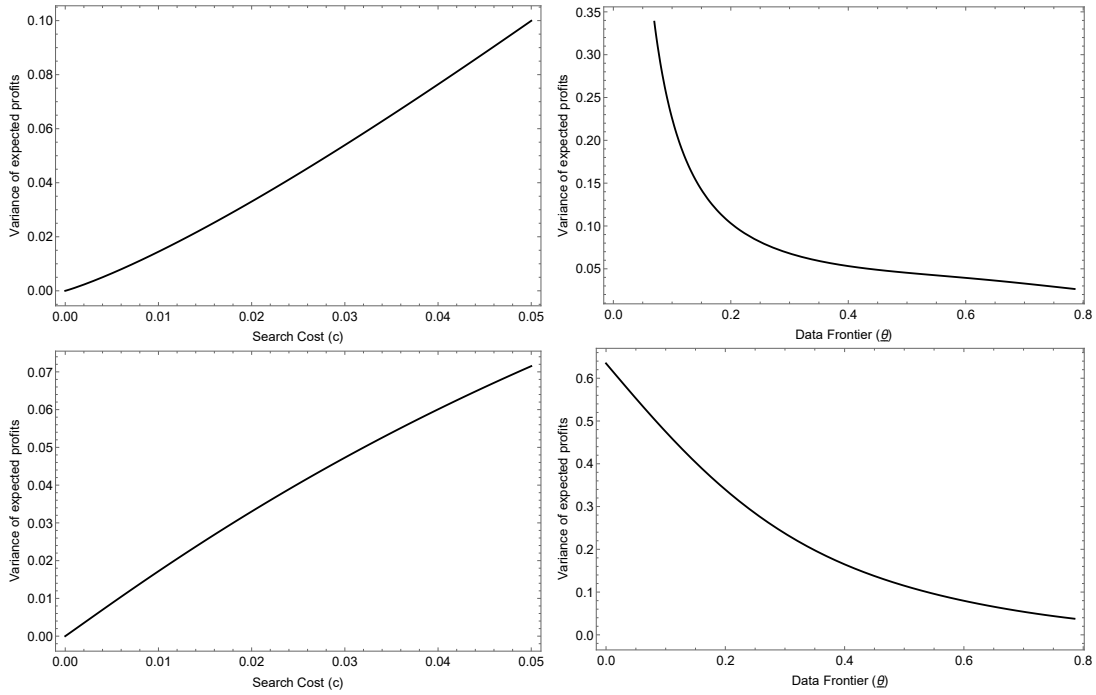

Figure 4: The left hand-side graphs plot the variance of speculators' expected profits, $\mathsf{Var}[\pi(\theta)]$, as a function of the search cost, c (other parameter values are $\underline{\theta} = \pi/5, \rho = 1, \sigma^2 = 1, \nu^2 = 1$). The right hand-side graph plots the variance of speculators' expected profits as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.05, \rho = 1, \sigma^2 = 1, \nu^2 = 1$). In the upper graphs, we assume that $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ (Case 1) while in the lower graphs we assume that $\phi(\theta) = 5\cos(\theta)\sin^4(\theta)$ (Case 2).

5.4 Data Abundance, Computing Power and Crowding

Crowding is the tendency for investors to follow the same trading strategy and exploit the same signals. In this section, we study how data abundance and computing power affect

the correlation between speculators' equilibrium positions, a measure of crowdedness of their trading strategy.[22] Specifically, let $\mathsf{Cov}(x(s_{\theta_i}, p^*), x(s_{\theta_j}, p^*))$ be the covariance between the equilibrium holdings of a speculator with type θ_i and a speculator with type θ_j. Using eq.(24), we obtain:

$$\mathsf{Cov}(x^*(s_{\theta_i}, p^*), x^*(s_{\theta_j}, p^*)) = \frac{\tau(\theta_i)\tau(\theta_j)}{\sigma^4\rho^2}\,\mathsf{Var}[\omega - p] = \frac{\tau(\theta_i)\tau(\theta_j)}{\sigma^4\rho^2\mathcal{I}(\theta^*, \underline{\theta})}. \tag{28}$$

We deduce that the pairwise correlation between the equilibrium positions of a speculator with type θ_i and a speculator with type θ_j is:

$$\mathsf{Corr}(x^*(s_{\theta_i}, p^*), x^*(s_{\theta_j}, p^*)) = \left(1 + \frac{\mathcal{I}(\theta^*, \underline{\theta})}{\tau(\theta_i)\tau_\omega}\right)^{-\frac{1}{2}} \left(1 + \frac{\mathcal{I}(\theta^*, \underline{\theta})}{\tau(\theta_j)\tau_\omega}\right)^{-\frac{1}{2}} \tag{29}$$

Thus, holding the quality of the predictors used by two speculators constant, their positions become less correlated when price informativeness is higher. The reason is that speculators trade on the component of their forecast of the asset payoff that is orthogonal to the price. This component reflects both the component of the fundamental, ω, that is not reflected into the equilibrium price and the noise in speculators' signal. The higher the first component relative to the second, the higher the pairwise correlation in speculators' positions in the asset. As the price becomes more informative, the first component becomes smaller and smaller relative to the noise component and as a result, the pairwise correlation between speculators' positions drops. Using Proposition 7, we deduce the following result.

Proposition 10.

1. *Greater computing power (a decrease in c) reduces the pairwise correlation of speculators' positions.*

2. *Data abundance has an ambiguous effect on the pairwise correlation of speculators' positions. It reduces it if it improves price informativeness but increases it otherwise.*

[22] Shanta Putchler, the CEO of Mannumeric (a quantitative investment fund) notes that: "*The single largest contributor to crowding is the simple fact that investors tend to do the same sorts of things. There is a real propensity for investors to analyse the same datasets, with the same statistical techniques, and hence end up with largely overlapping positions.*" *See https://www.man.com/maninstitute/crowding.*

This proposition suggests again that data abundance and computing power do not necessarily have the same effects. Testing the previous result requires measuring the pairwise correlation of speculators' positions, holding the quality of their signal constant. One possibility is to estimate the cross-sectional distribution of funds' predictors quality using the method described in Section 5.1 and analyze the effect of shocks to computing power or data abundance on the correlation in the positions of funds in different quantiles of the distribution.

6. Speculators' Welfare and Data Abundance

In this section, we analyze how data abundance and computing power affects speculators' welfare, measure by their ex-ante expected utility, which, in equilibrium, is $J(\theta^*, \theta^*) = g(\theta^*, \theta^*)$ (see Section 4.1). That is, each speculator's expected utility is just equal to the expected utility from trading on the worst predictor used in equilibrium. The reason is that the increase in the expected utility from trading associated with further explorations for a speculator who has found a predictor with type θ^* is just offset by the expected utility cost of further explorations.

As can be seen from eq.(13), the data frontier, $\underline{\theta}$ affects speculators' ex-ante expected utility only through its effects on (i) the quality of the worst predictor, $\tau(\theta^*)$ and (ii) the informativeness of the asset price, $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$. Now, when $\underline{\theta} < \underline{\theta}^{tr}(c)$, a decrease in $\underline{\theta}$ raises price informativeness (Proposition 7) and reduces the quality of the worst predictor (Proposition 4). Thus, it unambiguously reduces speculators' expected utility because $g(\theta^*, \theta^*)$ decreases with the informativeness of the asset price and increases with the quality of the worst predictor ($\tau(\theta^*)$.

Proposition 11. *When $\underline{\theta} < \underline{\theta}^{tr}(c)$, pushing back the data frontier (a decrease in $\underline{\theta}$) reduces speculators' expected utility.*

An increase in computing power raises the quality of the worst predictor and price informativeness in equilibrium. Thus, its effect on speculators' welfare is ambiguous. Numerical simulations show that the first effect dominates unless c becomes very small. Thus, in contrast to a push back of the data frontier, an improvement in computing power raises speculators' welfare. Figure 5 illustrates this point using same numerical examples as in Figure 2. For similar reasons, the needle in the haystack problem (a decrease in α)

has an ambiguous effect on speculators' welfare: It reduces price informativeness but also decreases the quality of the worst predictor. The first effect improve speculators' welfare while the second reduces it. Numerical simulations show that the second effect dominates for α low enough.

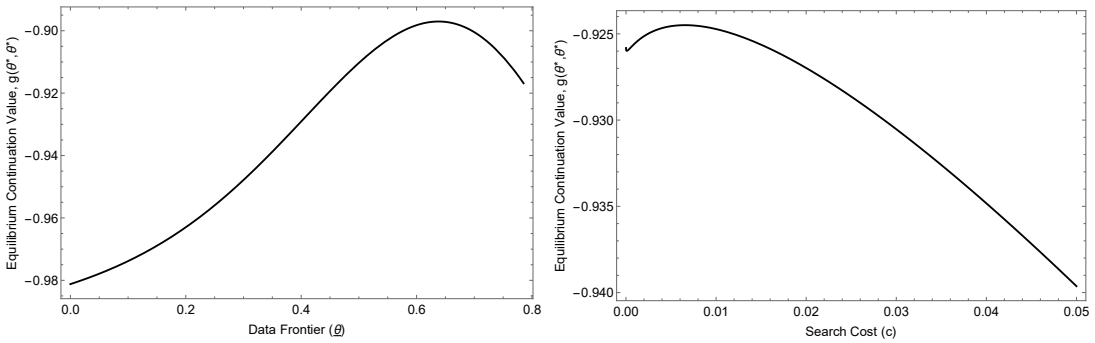

Figure 5: This graph shows speculators' ex-ante expected utility as a function of $\underline{\theta}$ and c when $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ (with other parameter values being set at $\rho = 1, \sigma^2 = 1, \nu^2 = 1$).

Thus, data abundance can make speculators worse off in equilibrium. One might then wonder whether it would not be optimal for a speculator to ignore new data. This is not the case, however. To see this, suppose that $\underline{\theta}$ drops from $\underline{\theta}_0$ to $\underline{\theta}_1 < \underline{\theta}_0$ but that speculators agree not to take advantage of the new data, i.e., to behave as if $\underline{\theta} = \underline{\theta}_0$ (in particular, they use the stopping rule $\theta^*(\underline{\theta}_0)$). A speculator's expected utility is then given by $J(\theta^*(\underline{\theta}_0, c, \alpha), \theta^*(\underline{\theta}_0, c, \alpha))$. When $\underline{\theta}_0 < \underline{\theta}^{tr}$, this decision is collectively optimal for speculators since their welfare decreases when $\underline{\theta}$ is reduced (Proposition 11). However, it is individually optimal for each speculator to deviate from the collective agreement. Indeed, one can show (using eq.(15)) that, holding *holding* $\theta^*(\underline{\theta}_0, c, \alpha)$ *constant*, a speculator's expected utility increases when the quality of the best predictor, $\underline{\theta}$, is improved. Thus, if new datasets open the possibility that the quality of the best predictor improves, each speculator will individually find optimal to use these datasets, if she expects others not to do so. But then the equilibrium stopping rule must shift from $\theta^*(\underline{\theta}_0)$ to $\theta^*(\underline{\theta}_1)$.[23]

Thus, data abundance can be "excessive" from speculators' viewpoint in the sense that they would be better off if the data frontier could not be improved. We now show that speculators' average investment in search is also excessive in the sense that, holding all exogenous parameters constant, they would be better off if they could commit to use

[23]This logic is similar to the arm's race to invest in trading speed for high frequency traders in Biais et al. (2015).

a less demanding stopping rule (and therefore predictors of lower quality on average). To see this, let assume that speculators can collectively choose a stopping rule, θ_r and commit to this choice. In this case, speculators would optimally choose the stopping rule θ_r^{**} such that:

$$\theta_r^{**} = \arg\max_{\theta_r} J(\theta_r, \theta_r). \tag{30}$$

Proposition 12. *In equilibrium, the stopping rule used by speculators is more demanding than the optimal stopping rule with commitment, that is, $\theta^* < \theta_r^{**}$. Thus, in equilibrium, speculators' investment in search for predictors, $E(n_i c)$, is too high relative to the investment that would maximize their welfare if they could collectively choose their stopping rule.*

Thus, there is excessive investment in search in equilibrium from speculators' viewpoint. The reason is as follows. Suppose that all speculators search for predictors with a stopping rule equal to θ^{**}. Now consider a speculator who draws a predictor with quality θ^{**}. Her expected trading profit is less than her expected utility of continuing searching for another predictor, assuming that other speculators keep searching with the same intensity (i.e., the same stopping rule θ^{**}). Formally:

$$g(\theta^{**}, \theta^{**}) < J(\theta^{**}, \theta^{**}).$$

Thus, the speculator has an incentive to deviate from the stopping rule by increasing her search intensity. However, in doing so, the speculator ignores the fact that this must be true for all speculators and that if all speculators deviate, there will all be worse off. Instead, a central planner organizing the search for predictors would internalize this effect.

7. Conclusion

Progress in information technologies enable investors to have access to more data (data abundance), both in terms of volume and diversity, and greater computing power, so that they can deploy more powerful techniques to extract information from raw data. In this paper, we propose a new model of information acquisition to analyze separately the effects of these two distinct dimensions of technological progress.

In our model, speculators search (mine data) for predictors through trials and optimally stop searching when they find a predictor with a signal-to-noise ratio larger than an endogenous threshold. As the outcome of speculators' search process is random, speculators discover different predictors. Thus, even though they are homogenous ex-ante, speculators are heterogeneous ex-post in terms of the quality of their predictors, their performance, their holdings etc. In this way, our model can generate predictions about the effects of data abundance and computing power on the distribution of asset managers' skills (precisions of their signals), the distribution of their trading profits or the correlations in their holdings. Moreover, asset price informativeness is determined by speculators' optimal data mining strategy because this strategy determines the average quality of their signals and thereby the informativeness of their aggregate demand.

The main message of our model is that the effects of data abundance and greater computing power are not the same. For instance, greater computing power always induces speculators to be more demanding for the minimal quality of their predictors while this is not necessarily the case for data abundance. As a result, positive shocks to computing power improve and homogenize predictors' quality across speculators and, for this reason, improve price informativeness. In contrast, data abundance can result in a greater dispersion of predictors' quality across speculators and even a drop in price informativeness.

References

Abis, Simona, 2018, Man vs machine: Quantitative and discretionary equity management, Technical report.

Agrawal, Ajay, John McHale, and Alexander Oettl, 2019, Finding needles in haystacks: Artifical intelligence and recombinant growth, *in The Economics of Artificial Intelligence, the University of Chicago Press* .

Bai, Jennie, Thomas Phillipon, and Alexi Savov, 2016, Have financial markets becomemore informative?, *Journal of Financial Economics* 122, 625–654.

Banerjee, Snehal, and Bradyn Breon-Drish, 2020, Dynamics of research and strategic trading, Technical report.

Biais, Bruno, Thiery Foucault, and Sophie Moinas, 2015, Equilibrium fast trading, *Journal of Financial Economics* 116, 292–313.

Brogaard, Jonathan, and Abalfazl Zareei, 2019, Machine learning and the stock market, Technical report.

da Silva, Jaime Teixeira, Panagiotis Tsigaris, and Mohammadamin Erfanmanesh, 2020, Publishing volumes in major databases related to covid-19, *Scientometrics* .

Dugast, Jerome, and Thierry Foucault, 2018, Data abundance and asset price informativeness, *Journal of Financial economics* 130, 367–391.

Farboodi, Maryam, Adrien Matray, and Laura Veldkamp, 2019, Where has all the data gone?, Technical report.

Farboodi, Maryam, and Laura Veldkamp, 2019, Long run growth of financial technology, *forthcoming American Economic Review* .

Gao, Meng, and Jiekun Huang, 2019, Informing the market: The effect of modern information technologies on information production, *The Review of Financial Studies* 1367–1411.

Goldfarb, Avi, and Catherine Tucker, 2019, Digital economics, *Journal of Economic Literature* 57, 3–43.

Goldstein, Itay, Shiijie Yang, and Luo Zuo, 2020, The real effects of modern information technologies, *Working paper, NBER* .

Grennan, Jillian, and Roni Michaely, 2019, Fintechs and the market for financial aanalysis, *Forthcoming Journal of Financial and Quantitative Analysis* .

Grossman, Sanford, and Joseph Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.

Han, Jungsuk, and Francesco Sangiorgi, 2018, Searching for information, *Journal of Economic Theory* 175, 342–373.

Harvey, Campbell, 2017, The scientific outlook in financial economics, *Journal of Finance* 72, 1399–1440.

Huang, Shyang, Yang Xiong, and Liyan Yang, 2020, Information skills and data sales, *Working paper* .

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp, 2016, A rational theory of mutual funds' attention allocation, *Econometrica* 84, 571–626.

Kacperczyk, Marcin, and Amit Seru, 2007, Fund managers use of public information: New evidence on managerial skills, *Journal of Finance* 62, 485–528.

Kacperczyk, Marcin, Stijn van Nieuwerburgh, and Laura Veldkamp, 2014, Time-varying fund manager skills, *Journal of Finance* 69, 1455–1483.

Katona, Zsolt, Markus Painter, Panos Patatoukas, and JienYin Zengi, 2019, On the capital market consequences of alternative data: Evidence from outer space, Technical report.

Marenzi, Octavio, 2017, Alternative data: The new frontier in asset management, *Report, Optimas Research* .

Nordhaus, William, 2015, Are we approaching an economic singularity? information technology and the future of economic growth, *Yale University Cowles Foundation Discussion Papers 2021* .

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira, 2020, Man vs. machine learning: The term structure of earbnings expectations and conditional biases, *Working paper, NBER* .

Veldkamp, Laura, 2011, *Information choice in macroeconomics and finance* (Princeton University Press).

Veldkamp, Laura, and Cindy Chung, 2020, Data and the aggregate economy, *Forthcoming Journal of Economic Literature* .

Verrecchia, Robert, 1982, Information acquisition in a noisy rational expectations economy, *Econometrica* 1415–1430.

Vives, Xavier, 1995, Short-term investment and the informational efficiency of the market, *Review of Financial Studies* 8, 125–160.

Yan, Xuemin (Sterling), and Lingling Zheng, 2017, Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach, *The Review of Financial Studies* 30, 1382–1423.

Zhu, Christina, 2019, Big data as a governance mechanism, *Review of Financial Studies* 32, 2021–2061.

A. Proofs

Proof of Proposition 1.

We show that $x^*(s_\theta, p)$ and p^* as given by eq.(8) and eq.(9) form an equilibrium. First, suppose that $x^*(s_\theta, p)$ is given by $x^*(s_\theta, p) = a(\theta)(\hat{s}(\theta) - p)$. In this case, the aggregate demand for the asset is given by:

$$D(p) = \int x^*(s_\theta, p) + \eta = \bar{a}(\omega - p) + \eta, \tag{31}$$

where $\bar{a}$ is the average value of $a(\theta)$ across all speculators ($\bar{a} = E(a(\theta) \mid \theta \in [\underline{\theta}, \theta^*]$. Hence, observing $D(p)$ (and p) is informationally equivalent to observing $\xi = \omega + \bar{a}^{-1}\eta$. Thus:

$$p^* = \mathsf{E}(\omega \mid D(p)) = \mathsf{E}(\omega \mid \eta) = (\frac{\sigma^2}{\sigma^2 + \bar{a}^{-2}\nu^2})\xi = (\frac{\tau_\xi}{\tau_\omega + \tau_\xi})\xi, \tag{32}$$

where $\tau_\xi \equiv \frac{\bar{a}^2}{\nu^2}$ is the precision of ξ as a signal about ω.

Now consider speculators. Using standard calculations in the CARA gaussian framework, we deduce that the optimal demand for the risky asset of a speculator with signal s_θ is:

$$x^*(s_\theta, p) = \frac{\mathsf{E}[\omega|s_\theta, p] - p}{\rho\, \mathsf{Var}[\omega|s_\theta, p]}, \tag{33}$$

As speculators have rational expectations on the price, they anticipate that it is linear in ξ, as in eq.(32). Moreover, let $\hat{s}_\theta \equiv \omega + \tau(\theta)^{-\frac{1}{2}}\epsilon_\theta$, so that $s_\theta = \cos(\theta)\hat{s}_\theta$. Then

$$\mathsf{E}[\omega|s_\theta, p] = \mathsf{E}[\omega|\hat{s}_\theta, \xi]. \tag{34}$$

and

$$\mathsf{Var}[\omega|s_\theta, p] = \mathsf{Var}[\omega|\hat{s}_\theta, \xi]. \tag{35}$$

Note that the precision of $\hat{s}_\theta$ is $\tau(\theta)\tau_\omega$. Thus, as all variables are normally distributed and ϵ_θ and η (the noises in $\hat{s}_\theta$ and ξ) are independent, standard calculations yield:

$$\mathsf{E}[\omega|\hat{s}_\theta, \xi] = \frac{\tau(\theta)\tau_\omega \hat{s}_\theta + \tau_\xi \xi}{\tau_\omega + \tau(\theta)\tau_\omega + \tau_\xi}. \tag{36}$$

and

$$\mathsf{Var}[\omega|s_\theta, p] = \frac{1}{\tau_\omega + \tau(\theta)\tau_\omega + \tau_\xi}. \tag{37}$$

Thus, we can rewrite eq.(33) as:

$$x^*(s_\theta, p) = \frac{\tau(\theta)\tau_\omega \hat{s}_\theta + \tau_\xi \xi - (\tau_\omega + \tau(\theta)\tau_\omega + \tau_\xi)p}{\rho}, \tag{38}$$

Using the fact that $p = \frac{\tau_\xi}{\tau_\omega + \tau_\xi}\xi$ we deduce that:

$$x^*(s_\theta, p) = \frac{\tau(\theta)\tau_\omega}{\rho}(\hat{s}_\theta - p) = \frac{\tau(\theta)}{\rho\sigma^2}(\hat{s}_\theta - p). \tag{39}$$

Thus, $x^*(s_\theta, p)$ is as conjectured (and as in eq.(8)) if and only if $a(\theta) = \frac{\tau(\theta)}{\rho\sigma^2}$. If follows that $\bar{a} = \frac{\bar{\tau}(\theta)}{\rho\sigma^2}$. Eq.(9) and eq.(10) in the text immediately follow from substituting this expression for $\bar{a}$ in eq.(32).

In sum we have shown that (i) if dealers expect speculators to follow the trading strategy $x^*(s_\theta, p)$ given by eq.(8) then they set a price given by eq.(9) and (ii) if dealers set a price given by eq.(9) then speculators follow the trading strategy $x^*(s_\theta, p)$ given by eq.(8). Thus, eq.(8) and eq.(9) form an equilibrium. More generally, it is possible to show that this is the unique equilibrium in which speculators' trading strategy is a linear function of their signal and the price.

Proof of Lemma 1.

Conditional on the realization of the price at date 1 and her signal, s_θ, the expected utility of trading for an investor given her optimal trading strategy is:

$$\begin{aligned} &\mathsf{E}(-\exp(-\rho(x^*(s_\theta, p)(\omega - p)) \mid s_\theta, p) = \\ -\mathsf{E}(\exp(-\rho(x^*(s_\theta, p)(\mathsf{E}(\omega \mid s_\theta, p) - p) - &\frac{\rho(x^*(s_\theta, p))^2}{2}\mathsf{Var}(\omega \mid s_\theta, p)). \end{aligned} \tag{40}$$

Substituting $x^*(s_\theta, p)$ by its expression in eq.(33), we deduce that:

$$\mathsf{E}([-\exp(-\rho(x^*(s_\theta, p)(\omega - p))] \mid s_\theta, p) = -\exp[-\frac{(\mathsf{E}[\omega|s_\theta, p] - p)^2}{2\,\mathsf{Var}[\omega|s_\theta, p]}] \tag{41}$$

Thus:

$$g(\theta, \theta^*) = -\mathsf{E}(\exp[-\frac{(\mathsf{E}[\omega|s_\theta, p^*] - p^*)^2}{2\,\mathsf{Var}[\omega|s_\theta, p^*]}]). \tag{42}$$

For a normally distributed variable Z with mean 0 and variance σ_Z^2, $\mathsf{E}[\exp(-Z^2)] = (1 + 2\sigma_Z^2)^{-1/2}$. As $\mathsf{E}[\omega|s_\theta, p] - p$, is normally distributed with mean zero, defining $Z =$

$\mathsf{E}[\omega|s_\theta, p] - p$, we deduce that:

$$g(\theta, \theta^*) = -\left(1 + \frac{\mathsf{Var}\left[\mathsf{E}[\omega|s_\theta, p^*] - p\right]}{\mathsf{Var}[\omega|s_\theta, p^*]}\right)^{-1/2} \tag{43}$$

Observe that:

$$\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta, p^*] - p^*]}{\mathsf{Var}[\omega|s_\theta, p^*]} = \rho^2 \,\mathsf{Var}[\omega|s_\theta, p^*]\,\mathsf{Var}[x^*(s_\theta, p^*)]. \tag{44}$$

Now using the expression for $x^*(s_\theta, p^*)$ in eq.(39), we obtain that:

$$\mathsf{Var}[x^*(s_\theta, p)] = \frac{\tau(\theta)^2\tau_\omega^2}{\rho^2}[\mathsf{Var}(\hat{s}_\theta) + \mathsf{Var}(p) - 2Cov(\hat{s}_\theta, p)]. \tag{45}$$

Using the expression for p^* in eq(32) and the fact that $\hat{s}_\theta = \omega + \tau(\theta)^{-\frac{1}{2}}\epsilon_\theta$, we obtain after some algebra that:

$$\mathsf{Var}[x^*(s_\theta, p^*] = \frac{\tau(\theta)^2\tau_\omega(\tau_\omega + \tau_\omega\tau(\theta)\tau_\xi)}{\rho^2(\tau_\omega + \tau_\xi)}. \tag{46}$$

Finally, using the expression for $\mathsf{Var}[\omega|s_\theta, p^*]$ in eq.(37) and the fact that $\tau_\xi = \frac{\bar{\tau}(\theta^*;\underline{\theta},\alpha)^2\tau_\omega^2}{\rho^2\nu^2}$, we deduce from eq.(44) that:

$$\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta, p] - p]}{\mathsf{Var}[\omega|s_\theta, p]} = \frac{\rho^2\sigma^2\nu^2\tau(\theta)}{\bar{\tau}(\theta^*;\underline{\theta},\alpha)^2 + \rho^2\sigma^2\nu^2} = . \tag{47}$$

This yields the expression for $g(\theta, \theta^*)$.

Proof of Proposition 2.

Step 1. We first show that there is a unique solution $\theta_i^*(\theta^*)$ to the indifference condition (16). Let define the function $L(\theta_i^*, \theta^*)$ as:

$$L(\theta_i^*, \theta^*) \equiv \alpha\int_{\underline{\theta}}^{\theta_i^*}\frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)}\phi(\theta)d\theta + 1 - \alpha\int_{\underline{\theta}}^{\theta_i^*}\phi(\theta)d\theta. \tag{48}$$

Function L is decreasing with θ_i^* because:

$$\frac{\partial L}{\partial\theta_i^*} = \alpha\int_{\underline{\theta}}^{\theta_i^*}\frac{\partial}{\partial\theta_i^*}\left(\frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)}\right)\phi(\theta)d\theta < 0. \tag{49}$$

Now, using the expression for $J()$ given in eq.(15), we can rewrite the indifference condi-

tion (16) as:

$$L(\theta_i^*, \theta^*) = \exp(-\rho c). \tag{50}$$

Moreover: $L(\underline{\theta}, \theta^*) = 1$ and $0 < L(\pi/2, \theta^*) < 1$. Thus, as $L(\theta_i^*, \theta^*)$ decreases in θ_i^* , eq.(48) has a unique solution $\theta_i^*(\theta^*)$ when c is small enough.

Step 2. We now show that there is a unique solution to the equilibrium condition $F(\theta^*) = \exp(-\rho c)$. Note that $F(\theta^*) = L(\theta^*, \theta^*)$. The derivative of $F(\theta^*)$ is

$$\frac{\partial F}{\partial \theta^*} = \alpha \int_{\underline{\theta}}^{\theta^*} \frac{\partial r(\theta, \theta^*)}{\partial \theta^*} \phi(\theta) d\theta, \tag{51}$$

where $r(\theta, \theta^*)$ is defined in eq.(20). As θ^* increases, both $\tau(\theta^*)$ and $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ decreases. We deduce that $r(\theta, \theta^*)$ decreases in θ^*.

Thus, $\frac{\partial F}{\partial \theta^*} < 0$. Moreover, we have (i) $F(\underline{\theta}) = 1$, (ii) $0 < F(\pi/2) < 1$ and (iii) $\exp(-\rho c) < 1$ (since $c > 0$). Thus, there is a unique solution to the condition $F(\theta^*) = \exp(-\rho c)$ and this solution is in $(\underline{\theta}, \pi/2)$ if and only if $F(\pi/2) \leq \exp(-\rho c) < 1$.

Proof of Proposition 3. In equilibrium, $F(\theta^*) = \exp(-\rho c)$. We have shown that $F(.)$ decreases in θ^* in the proof of Proposition 2. It immediately follows from these two observations that θ^* increase in c.

Proof of Proposition 4.

Part 1. In equilibrium, $F(\theta^*) = \exp(-\rho c)$. Moreover, it directly follows from eq.(19) that $F(theta^*)$ decreases in α because $r < 1$ and we know that $F(.)$ decreases in θ^*. It immediately follows from these observations that θ^* increases in α, as claimed in the first part of the proposition.

Part 2. Remember that $\tau(\theta) = cot^2(\theta)$ and that $\mathcal{I}(\theta^*; \underline{\theta}, \alpha) = \tau_\omega + \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2 \tau_\omega^2}{\rho^2 \nu^2}$. Using these two observations, we can rewrite $r(\theta, \theta^*)$ given in eq.(20) as:

$$r(\theta, \theta^*) = \frac{g(\theta, \theta^*)}{g(\theta^*, \theta^*)} = \left(\frac{\rho^2\sigma^2\nu^2 \cot^2(\theta^*) + \rho^2\sigma^2\nu^2 + \mathsf{E}\left[\cot^2(\theta') | \, \underline{\theta} \leq \theta' \leq \theta^*\right]^2}{\rho^2\sigma^2\nu^2 \cot^2(\theta) + \rho^2\sigma^2\nu^2 + \mathsf{E}\left[\cot^2(\theta') | \, \underline{\theta} \leq \theta' \leq \theta^*\right]^2} \right)^{\frac{1}{2}}. \tag{52}$$

Thus, after some algebra, we obtain:

$$\frac{\partial r}{\partial \underline{\theta}} = \frac{\partial \,\mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta} \leq \theta' \leq \theta^*\right]}{\partial \underline{\theta}}\,\mathsf{E}\left[\cot^2(\theta')\Big|\,\underline{\theta} \leq \theta' \leq \theta^*\right] \tag{53}$$

$$\times \frac{\rho^2\sigma^2\nu^2(\cot^2(\theta) - \cot^2(\theta^*))}{\left\{\rho^2\sigma^2\nu^2\cot^2(\theta) + \rho^2\sigma^2\nu^2 + \mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta} \leq \theta' \leq \theta^*\right]^2\right\}^{\frac{3}{2}}} \tag{54}$$

$$\times \frac{1}{\left\{\rho^2\sigma^2\nu^2\cot^2(\theta^*) + \rho^2\sigma^2\nu^2 + \mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta} \leq \theta' \leq \theta^*\right]^2\right\}^{\frac{1}{2}}}. \tag{55}$$

Moreover:

$$\frac{\partial \,\mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta} \leq \theta' \leq \theta^*\right]}{\partial \underline{\theta}} = -\frac{\phi(\underline{\theta})}{\int_{\underline{\theta}}^{\theta^*}\phi(\theta')d\theta'}\left(\cot^2(\underline{\theta}) - \mathsf{E}\left[\cot^2(\theta')\Big|\,\underline{\theta} \leq \theta' \leq \theta^*\right]\right) < 0, \tag{56}$$

where the last inequality follows from the fact $\cot(\theta)$ decreases with θ. We deduce from the expression for $\frac{\partial r}{\partial \underline{\theta}}$ that $r(\theta, \theta^*)$ decreases with $\underline{\theta}$ ($\frac{\partial r}{\partial \underline{\theta}} < 0$). Using the expression for $F(.)$ in eq.(19), we deduce that:

$$\frac{\partial F}{\partial \underline{\theta}} = \underbrace{\alpha\phi(\underline{\theta})(1 - r(\underline{\theta}, \theta^*))}_{>0} + \alpha\int_{\underline{\theta}}^{\theta^*}\underbrace{\frac{\partial r}{\partial \underline{\theta}}}_{<0}\phi(\theta)d\theta. \tag{57}$$

Thus, the effect of $\underline{\theta}$ on $F(.)$ and therefore the equilibrium stopping rule θ^* is ambiguous. We now show that this effect becomes negative when $\underline{\theta}$ is close enough to zero. To see this, observe that eq.(56) implies that:

$$\frac{\partial F}{\partial \underline{\theta}} < \alpha\phi(\underline{\theta}) + \alpha\int_{\underline{\theta}}^{\theta^*}\frac{\partial r}{\partial \underline{\theta}}\phi(\theta)d\theta \tag{58}$$

We show in the internet appendix that $\int_{\underline{\theta}}^{\theta^*}\frac{\partial r}{\partial \underline{\theta}}\phi(\theta)d\theta$ goes to $-\infty$ when $\underline{\theta}$ goes to zero.Thus, $\frac{\partial F}{\partial \underline{\theta}} < 0$ for $\underline{\theta}$ small enough. Let $\underline{\theta}^{tr}$ be the smallest value of $\underline{\theta}$ such that $\frac{\partial F}{\partial \underline{\theta}} < 0$. As in equilibrium, $F(\theta^*) = \exp(-\rho c)$ and $F(.)$ decreases in θ^*, it follows that θ^* increases in $\underline{\theta}$ when $\underline{\theta} < \underline{\theta}^{tr}$, as claimed in the first part of the proposition.

Proof of Proposition 5. It follows from direct inspection of the expression for $r(\theta, \theta^*)$ given in eq.(52) that $r(\theta, \theta^*)$ decreases with σ^2, and ν^2 because $\tau(\theta) > \tau(\theta^*)$. Thus, from eq.(19), we deduce that $F(\theta^*)$ decreases with σ^2, and ν^2. It follows from this observation, the fact $F(\theta^*)$ decreases with θ^* and the equilibrium condition $F(\theta^*) = \exp(-\rho c)$ that $\theta*$

decreases with σ^2 and ν^2.

Proof of Proposition 6. Follows from the text after the proposition.

Proof of Proposition 7.

Part 1. When a decrease in $\underline{\theta}$ reduces θ^*, it is clear that it raises the average quality of predictors and therefore price informativeness.

Now consider the other possible case, i.e., the case in which a decrease in $\underline{\theta}$ raises θ^*, i.e., the case in which θ^* locally decreases with $\underline{\theta}$. We know that such a case arises when $\underline{\theta}$ is low enough (see Proposition 4). We prove below that if price informativeness, $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$, decreases in this case then each investor would individually have an incentive to choose a more demanding stopping rule (i.e., θ_i^* would decrease, which leads to a contradiction since in equilibrium $\theta_i^* = \theta^*$.

Remember that the optimal stopping rule of each speculator solves: $L(\theta_i^*, \theta^*) = \exp(-\rho c)$ where $L(\theta_i^*, \theta^*)$ is given by eq.(48). We have shown that function L decreases in θ_i^* (see (eq:(49)). Next, for $\theta_i^* \geq \theta \geq \underline{\theta}$, define

$$l(\theta, \theta_i^*, \theta^*) = \frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)} = \left(\frac{\rho^2\sigma^2\nu^2\tau(\theta_i^*) + \rho^2\nu^2 + \sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2}{\rho^2\sigma^2\nu^2\tau(\theta) + \rho^2\nu^2 + \sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2} \right)^{\frac{1}{2}} = \left(\frac{\tau(\theta_i^*)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\tau(\theta)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)} \right)^{\frac{1}{2}}. \tag{59}$$

Clearly, $l(\theta, \theta_i^*, \theta^*)$ clearly increases when $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ increases. Thus, if a decrease in $\underline{\theta}$ leads to a decrease in price informativeness, i.e., $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$, it must be that $l(\theta, \theta_i^*, \theta^*)$ increases with $\underline{\theta}$ since $\underline{\theta}$ affects $l(\theta, \theta_i^*, \theta^*)$ only through its effect on price informativeness. This means that:

$$\frac{\partial l}{\partial \underline{\theta}} + \frac{\partial l}{\partial \theta^*} \frac{\partial \theta^*}{\partial \underline{\theta}} > 0. \tag{60}$$

Moreover, from eq.(48), we know that:

$$L(\theta_i^*, \theta^*) \equiv \alpha \int_{\underline{\theta}}^{\theta_i^*} l(\theta, \theta_i^*, \theta^*) + 1 - \alpha \int_{\underline{\theta}}^{\theta_i^*} \phi(\theta) d\theta. \tag{61}$$

Thus, the partial derivative of function L with respect to $\underline{\theta}$, taking into the collective

response of θ^*, can be written as

$$\frac{\partial L}{\partial \underline{\theta}} + \frac{\partial L}{\partial \theta^*}\frac{\partial \theta^*}{\partial \underline{\theta}} = \underbrace{\alpha\phi(\underline{\theta})(1 - l(\underline{\theta}, \theta_i^*, \theta^*))}_{>0} + \alpha \int_{\underline{\theta}}^{\theta_i^*} \left(\frac{\partial l}{\partial \underline{\theta}} + \frac{\partial l}{\partial \theta^*}\frac{\partial \theta^*}{\partial \underline{\theta}} \right) \phi(\theta) d\theta. \tag{62}$$

The effect of $\underline{\theta}$, everything else equal, is given by the first term. This term is positive and reflects the gain of new better predictors. From eq.(60), the second term is also positive if a decrease in $\underline{\theta}$ reduces price informativeness. Overall this means that in this case L increases with $\underline{\theta}$ $(\partial L/\partial \underline{\theta} + (\partial L/\partial \theta^*)/(\partial \theta^*/\partial \underline{\theta}) > 0$. As $\partial L/\partial \theta_i^* < 0$ and $L(\theta_i^*, \theta^*) = \exp(-\rho c)$, we deduce that θ_i^* increases with $\underline{\theta}$. However in equilibrium, $\theta_i^* = \theta^*$. Thus, this implies that θ^* increases with $\underline{\theta}$. A contradiction since we are in the case in which θ^* decreases with $\underline{\theta}$. This means that price informativeness cannot decrease even in this case.

Part 2. When α decreases, $\bar{\tau}(\theta^*)$ decreases (see Proposition 4). Hence, price informativeness goes down since $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ increases with $\bar{\tau}(\theta^*)$ and depends on α only through $\bar{\tau}(\theta^*)$ (see eq.(11). This proves the second part of the proposition. **Proof of**

Proposition 8. Consider the effect of $\underline{\theta}$ on speculators' expected profits. We know from Proposition 7 that $\bar{\tau}(\theta^*(\underline{\theta}, c, \alpha), \underline{\theta}, \alpha)$ decreases with $\underline{\theta}$. Moreover, $\bar{\tau}(\theta^*(\underline{\theta}, c, \alpha), \underline{\theta}, \alpha)$ goes to $\tau(\frac{\pi}{2}) = 0$ when $\underline{\theta}$ goes to $\frac{\pi}{2}$. Thus, if $\bar{\tau}(\theta^*(0, c, \alpha), 0, \alpha) > \tau_\omega \rho^2 \nu^2$, there is a unique value of θ, denoted $\hat{\theta}$, such that $\bar{\tau}(\theta^*(\hat{\theta}, c, \alpha), \hat{\theta}, \alpha) = \tau_\omega \rho^2 \nu^2$. Thus, when $\underline{\theta}$ varies, holding other parameters constant, speculators' expected profit reaches its maximum for $\bar{\tau}(\theta^*, \hat{\theta}, \alpha) = \tau_\omega \rho^2 \nu^2$. If instead, $\bar{\tau}(\theta^*(0, c, \alpha), 0, \alpha) \leq \tau_\omega \rho^2 \nu^2$, then speculators' expected profit always increases as $\underline{\theta}$ decreases. This proves Part 2 of Proposition 8.

Parts 1 and 2 can be proven in the same way. We therefore skip the proofs of these parts for brevity. In these cases, one obtains that $\hat{c}$ and $\hat{\alpha}$ are the unqiue solutions of, respectively, $\bar{\tau}(\theta^*(\underline{\theta}, \hat{c}, \alpha), \underline{\theta}, \alpha) = \tau_\omega \rho^2 \nu^2$ and $\bar{\tau}(\theta^*(\underline{\theta}, c, \hat{\alpha}), \underline{\theta}, \hat{\alpha}) = \tau_\omega \rho^2 \nu^2$.

Proof of Proposition 9. For a given $\underline{\theta}$, when $c = 0$ we have $\theta^* = \underline{\theta}$ and therefore $\mathsf{Var}[\pi(\theta)] = 0$, and when $c > 0$, $\theta^* > \underline{\theta}$ and therefore $\mathsf{Var}[\pi(\theta)] > 0$. Hence, it must be the case that $\mathsf{Var}[\pi(\theta)]$ is strictly increasing with c, for c close enough to 0.

In order to analyze the effect of $\underline{\theta}$ we consider the following expression for $\mathsf{Var}[\pi(\theta)]$:

$$\mathsf{Var}[\pi(\theta)] = \frac{\rho^2\sigma^4\nu^4\left(\mathsf{E}\left[\cot^4(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]-\mathsf{E}\left[\cot^2(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]^2\right)}{\left(\mathsf{E}\left[\cot^2(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]^2+\rho^2\sigma^2\nu^2\right)^2}. \tag{63}$$

For a given $\underline{\theta}$, when $c=0$ we have $\theta^*=\underline{\theta}$ and therefore $\mathsf{Var}[\pi(\theta)]=0$, and when $c>0$, $\theta^*>\underline{\theta}$ and therefore $\mathsf{Var}[\pi(\theta)]>0$. Hence, it must be the case that $\mathsf{Var}[\pi(\theta)]$ is strictly increasing with c, for c close enough to 0.

In order to analyze the effect of $\underline{\theta}$ we consider the following expression for $\mathsf{Var}[\pi(\theta)]$:

$$\mathsf{Var}[\pi(\theta)] = \frac{\rho^2\sigma^4\nu^4\left(\mathsf{E}\left[\cot^4(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]-\mathsf{E}\left[\cot^2(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]^2\right)}{\left(\mathsf{E}\left[\cot^2(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]^2+\rho^2\sigma^2\nu^2\right)^2}. \tag{64}$$

For a given search cost c, we must distinguish two cases. First, if the second moment diverges, that is

$$\lim_{\underline{\theta}\to 0}\mathsf{E}\left[\cot^4(\theta)\Big|\,\underline{\theta}\leq\theta\leq\theta^*\right]=+\infty, \tag{65}$$

Then we also have $\lim_{\underline{\theta}\to 0}\mathsf{Var}[\pi(\theta)]=+\infty$. This necessarily implies that $\mathsf{Var}[\pi(\theta)]$ is strictly decreasing with $\underline{\theta}$, for $\underline{\theta}$ close enough to 0.

Second, if the second moment converges, that is $\mathsf{E}\left[\cot^4(\theta)|\,0\leq\theta\leq\theta^*\right]<\infty$, we will show that the term $\mathsf{E}\left[\cot^4(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]$ still increases (when $\underline{\theta}$ decreases) at a rate that is an order of magnitude larger than the other terms, and that $\mathsf{Var}[\pi(\theta)]$ increases as well. To be more specific, consider a push back of the data frontier from $\underline{\theta}$ to $\underline{\theta}-\delta$, with δ positive and very small. The growth of the variance of profits is equal to $-\delta\times d\log(\mathsf{Var}[\pi(\theta)])/d\underline{\theta}$, that is

$$\delta\times\left\{\frac{-d\,\mathsf{E}\left[\cot^4(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]/d\underline{\theta}+2\,\mathsf{E}\left[\cot^2(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]d\,\mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]/d\underline{\theta}}{\mathsf{E}\left[\cot^4(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]-\mathsf{E}\left[\cot^2(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]^2}\right.$$
$$\left.+\frac{4\,\mathsf{E}\left[\cot^2(\theta)|\,\underline{\theta}\leq\theta\leq\theta^*\right]d\,\mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]/d\underline{\theta}}{\rho^2\sigma^2\nu^2+\mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]^2}\right\}.$$

Then we will show that, for small $\underline{\theta}$'s, $d\,\mathsf{E}\left[\cot^4(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]/d\underline{\theta}$ dominates (by an order of magnitude) $d\,\mathsf{E}\left[\cot^2(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]/d\underline{\theta}$.

Notice first that $\mathsf{E}\left[\cot^4(\theta)|\,0\leq\theta\leq\theta^*\right]<\infty$ implies that $\phi(\theta)\cot^4(\theta)$ can be inte-

grated in 0. Locally around $\theta = 0$, since $\cot(\theta) \sim \sin^{-1}(\theta) \sim \theta^{-1}$, we have

$$\phi(\theta)\cot^4(\theta) \sim \phi(\theta)\cot^2(\theta)\theta^{-2} \tag{66}$$

As θ^{-2} cannot be integrated in 0, it must be the case $\lim_{\underline{\theta}\to 0}\phi(\theta)\cot^2(\theta) = 0$. This is a necessary condition so that $\phi(\theta)\cot^4(\theta)$ can be integrated.

Next, we compute the derivative of the average quality in equilibrium, that is

$$\frac{d\,\mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right]}{d\underline{\theta}} = \frac{\partial\,\mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right]}{\partial\underline{\theta}} + \frac{\partial\,\mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right]}{\partial\theta^*}\frac{\partial\theta^*}{\partial\underline{\theta}} \tag{67}$$

$$= -\frac{\phi(\underline{\theta})}{\int_{\underline{\theta}}^{\theta^*}\phi(\theta')d\theta'}\left(\cot^2(\underline{\theta}) - \mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right]\right) \tag{68}$$

$$-\frac{\phi(\theta^*)}{\int_{\underline{\theta}}^{\theta^*}\phi(\theta')d\theta'}\left(\mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right] - \cot^2(\theta^*)\right)\frac{\partial\theta^*}{\partial\underline{\theta}} \tag{69}$$

According to Proposition 7, we have $d\,\mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right]/d\underline{\theta} < 0$, and according to Proposition 4, we have $\partial\theta^*/\partial\underline{\theta} < 0$ for $\underline{\theta}$ small enough. Hence, for $\underline{\theta}$ close to 0 we have

$$0 < -\frac{\partial\theta^*}{\partial\underline{\theta}} < \phi(\underline{\theta}) \times \frac{\overbrace{\cot^2(\underline{\theta}) - \mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right]}^{(*)}}{\phi(\theta^*)\left(\mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right] - \cot^2(\theta^*)\right)}. \tag{70}$$

The term $(*)$ is dominated by the term $\cot^2(\underline{\theta})$. Then, for $\underline{\theta}$ small, there is a constant $K_1 > 0$ such that

$$0 < -\frac{\partial\theta^*}{\partial\underline{\theta}} < K_1\phi(\underline{\theta})\cot^2(\underline{\theta}). \tag{71}$$

and therefore, pluging inequality (71) in equation (67), we obtain that there exists a constant K_2 such that

$$0 < -\frac{d\,\mathsf{E}\left[\cot^2(\theta')\middle|\,\underline{\theta}\le\theta'\le\theta^*\right]}{d\underline{\theta}} < K_2\phi(\underline{\theta})\cot^2(\underline{\theta}). \tag{72}$$

Finally, we compute the derivative of the second moment in equilibrium and obtain

$$\frac{d\,\mathsf{E}\left[\cot^4(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]}{d\underline{\theta}} = -\frac{\phi(\underline{\theta})}{\int_{\underline{\theta}}^{\theta^*}\phi(\theta')d\theta'}\left(\cot^4(\underline{\theta})-\mathsf{E}\left[\cot^4(\theta')\Big|\,\underline{\theta}\leq\theta'\leq\theta^*\right]\right) \tag{73}$$

$$-\frac{\phi(\theta^*)}{\int_{\underline{\theta}}^{\theta^*}\phi(\theta')d\theta'}\left(\mathsf{E}\left[\cot^4(\theta')\Big|\,\underline{\theta}\leq\theta'\leq\theta^*\right]-\cot^4(\theta^*)\right)\frac{\partial\theta^*}{\partial\underline{\theta}} \tag{74}$$

As the order of magnitude of $\partial\theta^*/\partial\underline{\theta}$ is (at best) $\phi(\underline{\theta})\cot^2(\underline{\theta})$, then the order of magnitude of the second derivative, for $\underline{\theta}$ small is

$$\frac{d\,\mathsf{E}\left[\cot^4(\theta')|\,\underline{\theta}\leq\theta'\leq\theta^*\right]}{d\underline{\theta}}\sim-\frac{\phi(\underline{\theta})\cot^4(\underline{\theta})}{\int_{\underline{\theta}}^{\theta^*}\phi(\theta')d\theta'} \tag{75}$$

Hence, around $\underline{\theta}=0$, $\frac{d\,\mathsf{E}\left[\cot^4(\theta')\left|\underline{\theta}\leq\theta'\leq\theta^*\right.\right]}{d\underline{\theta}}$ dominates $\frac{d\,\mathsf{E}\left[\cot^2(\theta')\left|\underline{\theta}\leq\theta'\leq\theta^*\right.\right]}{d\underline{\theta}}$ by an order of magnitude.

Proof of Proposition 10. Direct from the arguments in the text.

Proof of Proposition 11 Direct from the arguments in the text.

Proof of Proposition 12. First, notice that the search decision, $\widehat{\theta}$ of any given speculator, or the social planner, will always be such that

$$\exp(-\rho c)-1+\alpha\int_{\underline{\theta}}^{\widehat{\theta}}\phi(\theta)d\theta>0 \tag{76}$$

Indeed define θ_{min} such that the former inequality, $\exp(-\rho c)-1+\alpha\int_{\underline{\theta}}^{\theta_{min}}\phi(\theta)d\theta=0$, if any. For $\theta<\theta_{min}$, we have $J(\theta,\theta_r)>0$ while $g(\theta,\theta_r)<0$ for any $\theta>0$, which is inconsistent. And for $\theta\to\theta_{min}^+$, $J(\theta,\theta_r)\to-\infty$, for any θ_r; therefore any optimal search threshold should be larger. Next, consider the derivative of $J(\theta_{r,t}\,heta_r)$ with respect to

θ_r, $\partial J/\partial\theta_r$. It verifies the following

$$\left(\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta\right) \frac{\partial J}{\partial \theta_r} + \alpha\phi(\theta_r)\Pi(\theta_r) = \alpha\phi(\theta_r)g(\theta_r,\theta_r) + \alpha \int_{\underline{\theta}}^{\theta_r} \frac{\partial g}{\partial \theta_r}\phi(\theta)d\theta$$

(77)

$$\Leftrightarrow \left(\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta\right) \frac{\partial J}{\partial \theta_r} = \tag{78}$$

$$\alpha\phi(\theta_r)\frac{\left(\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta\right) g(\theta_r,\theta_r) - \alpha \int_{\underline{\theta}}^{\theta_r} g(\theta,\theta_r)\phi(\theta)d\theta}{\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta} + \alpha \int_{\underline{\theta}}^{\theta_r} \frac{\partial g}{\partial \theta_r}\phi(\theta)d\theta$$

(79)

$$\Leftrightarrow \left(\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta\right) \frac{\partial J}{\partial \theta_r} = \tag{80}$$

$$\frac{-\alpha\phi(\theta_r)g(\theta_r,\theta_r)}{\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta}\left(-\exp(-\rho c) + 1 - \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta + \alpha \int_{\underline{\theta}}^{\theta_r} r(\theta,\theta_r)\phi(\theta)d\theta\right) + \alpha \int_{\underline{\theta}}^{\theta_r} \frac{\partial g}{\partial \theta_r}\phi($$

(81)

$$\Leftrightarrow \left(\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta\right) \frac{\partial J}{\partial \theta_r} = \tag{82}$$

$$\frac{-\alpha\phi(\theta_r)g(\theta_r,\theta_r)}{\exp(-\rho c) - 1 + \alpha \int_{\underline{\theta}}^{\theta_r} \phi(\theta)d\theta} \underbrace{(F(\theta_r) - \exp(-\rho c))}_{>0 \text{ iff } \theta_r \leq \theta^*} + \alpha \underbrace{\int_{\underline{\theta}}^{\theta_r} \frac{\partial g}{\partial \theta_r}\phi(\theta)d\theta}_{>0} \tag{83}$$

Thus, $\partial J/\partial\theta_r \mid_{\theta_r=\theta^*} > 0$. It follows that $\theta^* < \theta^{**}$. The expected investment in search of speculator i for a given stopping rule θ_i^* is $E(n_i)c = \frac{c}{\Lambda(\theta_i^*;\underline{theta},\alpha)}$ (see eq.(3). Now, $\Lambda(\theta_i^*;\underline{\theta},\alpha)$ increases with θ_i^*. It then follows from the fact that $\theta^* < \theta^{**}$ that the expected investment in search is larger when $\theta_i^* = \theta^*$ than when $\theta_i^* = \theta^{**}$, as claimed in the second part of the proposition.