# Equilibrium Data Mining and Data Abundance[*]

Jérôme Dugast[†] Thierry Foucault[‡]

October, 2021

**Abstract**

We analyze how computing power and data abundance affect speculators' search for predictors. Speculators search for predictors through trials and optimally stop searching when they find a predictor with a signal-to-noise ratio greater than an endogenous threshold. Greater computing power raises this threshold by reducing search costs. In contrast, data abundance can reduce this threshold because (i) it reduces rents from informed trading, except for the best informed speculators and (ii) it increases the average number of trials to find a predictor. We study implications of these findings for active asset managers' performance, the similarity of their signals and positions and the informativeness of asset prices.

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

[*] We are grateful to Simona Abis, Snehal Banerjee, Bruno Biais, Dion Bongaerts, Maxime Bonnelli, Adrian Buss, Jean-Edouard Colliard, Bernard Dumas, Maryam Farboodi, Sergei Glebkin, Denis Gromb, Johan Hombert, Pete Kyle, Sophie Moinas, Joël Peress, Francesco Sangiorgi, Daniel Schmidt, Andriy Shkilko (discussant), Alberto Teguia, Mao Ye, Josef Zechner and participants at the ILB Rising Talents in Finance and Insurance Conference, the Paris Finance December Meeting 2020, the 2020 European Finance Association Meetings, the 2020 future of Financial Information Conference, the 2021 Western Finance Association Meetings, the 2021 FIRS Conference, the Microstructure Exchange virtual seminar, Frankfurt School of Management, HEC Paris, INSEAD, Mc Gill, the University of Maryland, the University of New South Wales, and the University of Vienna for very useful comments. All errors are the authors alone. All rights reserved by Jérôme Dugast et Thierry Foucault

[†] Université Paris-Dauphine, Université PSL, CNRS, DRM, Finance, 75016 PARIS, FRANCE. 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

Active asset managers play a central role in securities markets. They make substantial investments to discover new predictors of asset payoffs and, by trading on these predictors, they make securities prices more informative. Technological progress changes how asset managers discover predictors by enabling them to use (i) more diverse data ("alternative data" generated by social media, web traffic, online transactions or mobile phones etc.) and (ii) more powerful computer-based methods (e.g., machine learning) to mine these data.[1] As a result, the set of potential predictors has considerably increased.[2] This evolution raises many interesting questions: How does it affect managers' search for predictors? How does it affect asset managers' performance? Does it make asset managers' signals and holdings more similar? Does it make asset prices more informative?

In this paper, we address these questions. A unique feature of our theory is to allow a separate analysis of the effect of expanding the set of available data to find predictors (data abundance) and the effect of reducing information processing costs (greater computing power). These two aspects of the big data revolution are related but conceptually distinct.[3] To develop predictions about the effects of progress in information technologies, it is therefore important to develop models of information acquisition, like ours, in which data abundance and computing power are different parameters.

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"), each speculator observes the realization of her predictor and optimally chooses her trading strategy. The trading stage is similar to other rational expectations models (in particular 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 at a fixed cost, we explicitly model the search for a predictor and we analyze how the

[1] See Goldman Sachs Asset Management (2016): "*The role of big data in investing.*" 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).

[2] For instance, Martin and Nagel (2020) note (on p.2) that: "*As technology has improved, the set of available and potentially valuation-relevant predictor variables has expanded enormously over time.*"

[3] For instance, social media data expands the set of variables that investors can consider to find predictors but does not per se lower the cost of processing these data)

optimal search strategy depends on (i) the cost of exploration (computing power) and (ii) the "search space" (data abundance).

We model the search for predictors as follows. Each speculator can combine variables (e.g., past returns, accounting variables and social media data) from different datasets to build predictors. A predictor is characterized by its signal-to-noise ratio ("quality"). The quality of a given predictor is a priori unknown but speculators know the distribution of quality across predictors. The lowest quality is zero (just noise) while the highest quality (the "data frontier") is denoted $\tau^{max}$. Given this distribution, each speculator simultaneously and independently discovers predictors during the exploration phase. Discovering a predictor and its quality requires launching a round of exploration, which costs "$c$." A round returns a predictor with probability $\alpha$ and fails otherwise. After obtaining a predictor, a speculator can decide to trade on the predictor or to search for another one, which requires paying the exploration cost again. This process goes on until the speculator finds a satisficing predictor.

In practice, discovering and selecting a predictor requires (a) buying and preparing new data for analysis, (b) building a predictor with these data and assessing its quality with statistical techniques and (iii) deciding, via extensive backtesting, whether a predictor is good enough for live trading.[4] One round of exploration comprises all these steps and the exploration cost can be interpreted as the labor cost of executing them. We assume that automation and greater computing power reduces this cost.[5]

In contrast, we assume that data abundance affects the distribution of predictors' quality in two ways. Firstly, data abundance enables speculators to discover new predictors by using combination of variables that previously were not available (e.g., data from social media). This possibility pushes further the data frontier, $\tau^{max}$, i.e., improves the quality of the best predictors (the "hidden gold nugget" effect).[6] Secondly, data abun-

[4] See Chapters 8 and 9 in Narang (2013) for a practitioner's account of the way quant funds generate predictors.

[5] 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.

[6] This effect is often discussed in the financial press (e.g., "*Hedge funds see a gold rush in data mining*", Financial Times, August 28, 2017) and supported by recent empirical findings. For instance, Katona, Painter, Patatoukas, and Zengi (2019) find that combining satellite images of parking lots of U.S. retailers from two distinct data providers improves the accuracy of the forecasts of retailers' quarterly earnings (see also Zhu (2019)). Also, van Binsbergen, Han, and Lopez-Lira (2020) find that,

dance creates a "needle in the haystack" problem: It results in a proliferation of datasets, of which only a fraction is useful for forecasting asset payoffs. This effect increases the likelihood that a particular dataset proves useless after being tested, i.e., reduces $\alpha$. In our model, we analyze these two effects separately by varying either $\tau^{max}$ or $\alpha$.

In equilibrium, each speculator optimally stops searching for a predictor after finding one whose quality exceeds an endogenous threshold, $\tau^*$. This threshold is such that the speculator's expected utility of trading on a predictor of quality $\tau^*$ is just equal to her expected utility of searching for another predictor (her continuation value). Thus, the quality of predictors used in equilibrium ranges from $\tau^*$ (least informative) to $\tau^{max}$ (most informative). Hence, even though speculators are ex-ante homogeneous (same preferences and exploration costs), there is endogenous heterogeneity in the quality of speculators' predictors (and therefore performance).

We show that data abundance and computing power do not have the same effects on speculators' optimal search policy, $\tau^*$. To understand why, it useful to contrast the effect of a decrease in the cost of exploration, $c$, and the effect of an increase in the quality of the best predictor, $\tau^{max}$ on the value of launching a new round of exploration after finding a predictor (the continuation value), holding the search policy ($\tau^*$) constant. An increase in $\tau^{max}$ has two countervailing effects. On the one hand, it raises the continuation value because the expected utility of trading on the best predictor becomes even larger. On the other hand, speculators who obtain the best predictor now trade even more aggressively on their signal (i.e., they make larger bets for a given deviation between the asset price and their forecast of its payoff) because they face less risk (the "aggressiveness effect"). As a result, the asset price is more informative (closer to the asset payoff), which reduces the value of searching for a predictor. When $\tau^{max}$ is large enough, the aggressiveness effect dominates and speculators optimally react by adopting a less demanding search policy (so that $\tau^*$ decreases in equilibrium).

In contrast, a decrease in $c$ unambiguously raises the value of searching for another predictor after finding one because it reduces the total expected cost of search without affecting speculators' average trading aggressiveness. Thus, speculators optimally react

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). Last, Gu, Kelly, and Xiu (2020) consider 900+ predictors of stock and market returns and find that machine learning techniques (trees and neural networks) considerably increase out-of-sample $R^2$ of predictive models.

to a decrease in $c$ by adopting a more demanding search policy (which, in equilibrium, eventually raises their average aggressiveness). The effect of a decrease in $\alpha$ (the "needle in the haystack" effect) is symmetric to that of a decrease in $c$ because it increases the total expected cost of search without affecting speculators' average trading aggressiveness.

In sum, greater computing power always reduces the difference between the quality of the best and the worst predictor used in equilibrium while data abundance (an increase in $\tau^{max}$ or a decrease in $\alpha$) has the opposite effect (in the case of $\tau^{max}$ when it is large enough). These contrasting effects yield several testable implications.

First, several papers (e.g., Kacperczyk and Seru (2007) or Kacperczyk, van Nieuwerburgh, and Veldkamp (2014)) show that asset managers' skills (measured by their stock picking or timing abilities) are heterogeneous. In our model, as in other papers, an asset manager's skill is determined by the quality of their signal. Even though speculators are identical ex-ante and follow the same search policy in equilibrium, their skills are heterogeneous because the outcome of their search is random. One implication is that heterogeneity should persist even after controlling for characteristics that may explain heterogeneity in skills (e.g., fund size, investments in technology or asset managers' educational background). A second implication is that shocks to data abundance or computing power should affect the dispersion of asset managers' skills (e.g., the difference in skills between funds in the lowest and top skill deciles). Specifically, improvements in computing power should reduce heterogeneity in funds' skills (because it increases $\tau^{*}$) while data abundance should have the opposite effect (for $\tau^{max}$ large enough). We also predict that an increase in prior uncertainty (the variance of the asset payoff) or the volume of uninformed (noise) trading should reduce heterogeneity in funds' skills because it induces speculators to be more demanding for the quality of their predictors in equilibrium.[7]

Our second set of predictions is about the informativeness of asset prices for fundamentals. 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

[7] In our model, managers are ex-ante identical and heterogeneity in their skills only stem from randomness in the search process. Of course, in reality, asset managers are probably not identical ex-ante (before discovering predictors). Heterogeneity in their skills may reflect characteristics such fund size (possibly correlated with resources to find information, fund type (quant vs. discretionary), fund style etc. Our predictions should be tested after controlling for these characteristics.

[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.

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, $\tau^*$, of the least satisficing predictor. On the other hand, via its effect on $\tau^{max}$, it increases the trading aggressiveness of speculators who find the best predictors (and thereby the average aggressiveness of all speculators). The first effect reduces the average trading aggressiveness of all speculators while the second effect increases 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 ($\alpha$ decreases) then the first effect can dominate so that price informativeness drops when more data become available.[9].

Our third set of predictions regards the effects of computing power and data abundance on speculators' trading profits (excess returns) and the similarity 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 small enough. An increase in the quality of the most informative predictor, $\tau^{max}$, has the same effect for the same reasons. A decrease in $\alpha$ reduces price informativeness and the average quality of predictors used in equilibrium. The second (first) effect dominates when the needle in the haystack problem becomes sufficiently severe ($\alpha$ is low enough). Hence, the model also predicts an inverse U-shape relationship between speculators' average trading profits and data abundance. Overall the model implies that progress in information technologies initially benefit to all speculators until a point where it starts reducing their profits.

We also show that greater computing power or an improvement in the quality of the most informative predictor reduce the pairwise correlation in speculators' trades (i.e., their similarity). The reason is that, in equilibrium, speculators optimally trade on the component of their predictors that is orthogonal to the equilibrium price. As $c$ decreases

[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.

or $\tau^{max}$ increases, this component increasingly reflects the noise in speculators' signals. As this noise is independent across speculators, speculators' holdings become less correlated when $c$ decreases or $\tau^{max}$ increases (a decrease in $\alpha$ has the opposite effect because it reduces price informativeness). Interestingly, this happens even though speculators may become more similar in terms of the quality of their signal (e.g., in the case of a decrease in $c$).

## 2. Contribution to the Literature

Our paper contributes to the literature on informed trading in financial markets when information acquisition is endogenous (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. The relationship between a speculator's total expected cost of obtaining information and the expected precision of her signal is endogenous and micro-founded by an optimal search model. As explained previously, this approach enables us 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). To our knowledge, our paper is the first to offer this possibility.

A few other papers have formalized information acquisition as a search problem (Garleanu and Pedersen (2018), Han and Sangiorgi (2018), Banerjee and Breon-Drish (2020)) but they analyze different questions. In Garleanu and Pedersen (2018), investors can invest in passive or active funds and pay a search cost to discover whether an active asset manager is informed or not about a risky asset. In contrast to our model, informed managers have a signal of the same precision obtained as in Grossman and Stiglitz (1980). In Han and Sangiorgi (2018), an agent can draw, with replacement, normally distributed signals from an "urn." Each draw is costly, similar to the cost of exploration in our model. Interestingly, 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, which provides a microfoundation for the assumption that information acquisition costs are convex in precision. 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 (Han and Sangiorgi (2018) do not apply their model to trading in financial markets). In Banerjee and Breon-Drish (2020), one investor dynamically controls her timing for information acquisition about the payoff of a risky asset. She 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 she searches for information, the investor finds a signal of a given precision according to a Poisson process and starts trading on this signal as soon as she finds it. 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.

More broadly, our paper is related to the growing literature on the theoretical effects of new information technologies for the production of financial information (e.g., Abis (2018), Dugast and Foucault (2018), Farboodi and Veldkamp (2019), Milhet (2020) or Huang, Xiong, and Yang (2020)). This literature assumes that progress in information technologies reduces the cost of processing information or relaxes investors' attention constraints and explores ramifications of this assumption. Our model accounts for another dimension of this progress, namely data abundance (the expansion of speculators' 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.

# 3. Model

## 3.1 Searching predictors

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. Speculators have no initial endowments in these assets.

Figure 1 describes the timing of the model. The payoff of the risky asset, $\omega$, is realized in period 2 and is normally distributed with mean zero and variance $\sigma^2$. Speculators search a predictor of the asset payoff in period 0 (the "exploration stage"). Then, in period 1 (the "trading stage"), they observe the realization of their predictor and can trade in the market for the risky asset.

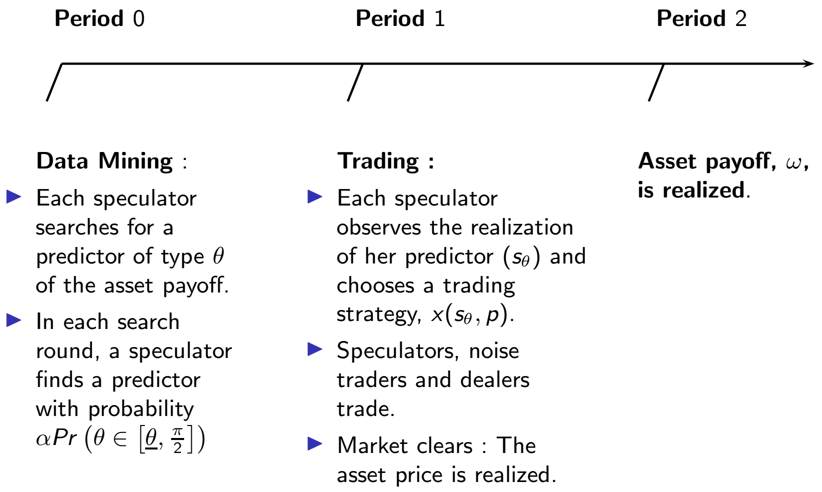
**Period** 0
**Period** 1
**Period** 2
**Data Mining** :
Each speculator searches for a predictor of type $\theta$ of the asset payoff.
In each search round, a speculator finds a predictor with probability $\alpha Pr\left(\theta \in [\underline{\theta}, \frac{\pi}{2}]\right)$
**Trading :**
Each speculator observes the realization of her predictor ($s_\theta$) and chooses a trading strategy, $x(s_\theta, p)$.
Speculators, noise traders and dealers trade.
Market clears : The asset price is realized.
**Asset payoff,** $\omega$**, is realized**.

**Figure 1:** Timing

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

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

where $\theta \in [0, \pi/2]$ and the $\varepsilon_\theta$s are normally and independently distributed with mean zero and variance $\sigma^2$ and $\varepsilon_\theta$ is independent from $\omega$. Let $\tau(\theta) \equiv \cos^2(\theta)/\sin^2(\theta) = \cot^2(\theta)$ denote the signal-to-noise ratio for a predictor with type $\theta$. We refer to this ratio as the "quality" of a predictor. The quality of a predictor decreases with its type, $\theta$ and varies from zero ($\theta = \frac{\pi}{2}$) to infinity (when $\theta$ goes to zero). It is unrelated to the uncertainty about the asset payoff, $\sigma^2$, because $\mathsf{Var}[\varepsilon_\theta] = \mathsf{Var}[\omega] = \sigma^2$.[10] We assume that predictors' types, $\theta$s, are distributed according to the cumulative probability distribution $\Phi(.)$ (density $\phi(.)$) on $[0, \pi/2]$.

[10] Without this assumption, the quality of all predictors would, counter-intuitively, increase with uncertainty.

The predictor $s_\theta$ is equivalent (in terms of informativeness) to the predictor $\hat{s}_\theta = \omega + \tau(\theta)^{-\frac{1}{2}}\varepsilon_\theta$. As there is a one-to-one mapping between $\theta$ and $\tau$, we could use this (more standard) specification for speculators' predictors and use the distribution of $\tau$ (rather than that of $\theta$) as a primitive of the model without changing any results (see Section A.II in the online appendix for a formal proof). Our approach just parameterizes $\tau$ in a way that it is convenient for some calculations.

Speculators discover predictors' types in period 0 via a sequential search process comprising multiple rounds of exploration. Each round costs $c$ and possibly yields a new type of predictor in $[\underline{\theta}, \frac{\pi}{2})$, i.e., speculators cannot find predictors with quality higher than $\tau^{max} \equiv \tau(\underline{\theta})$. More specifically, with probability $\alpha(1 - \Phi(\underline{\theta}))$ $(0 < \alpha \leq 1)$, an exploration round is successful and returns a predictor of type $\theta$ (picked according to the distribution $\phi(.)$) in $[\underline{\theta}, \frac{\pi}{2})$. Otherwise, it returns a predictor that is just noise. After each exploration round, a speculator can decide (i) to stop searching and to trade in period 1 on the predictor she just found or (ii) to start a new exploration. There is no limit on the number of exploration rounds.

It is worth stressing that speculators observe the realization of their chosen predictor, $s_\theta$, in period 1, *not* in 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 function (e.g., chosen with linear regressions or machine learning techniques) of variables from different datasets (e.g., past earnings, satellite images and consumer transactions data) that minimizes the predictor's average forecasting error in-sample based on past realizations of the payoffs and the variables used to build the predictor). The speculator then uses the realization of these variables at date 1 to compute the predictor, $s_\theta$, at this date (out-of-sample).[11]

As more datasets become available ("data abundance"), investors can try more diverse variables to predict asset payoffs (even holding the number of variables used to build

[11] For instance, the predictor could be obtained by running a regression of $\omega$ on some variables. In this approach, the $\mathrm{R}^2$ of the regression is a measure of the quality of the predictor. Indeed, the theoretical $\mathrm{R}^2$ of a regression of $\omega$ on $s_\theta$ (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 $\mathrm{R}^2$ of a regression of the asset payoff on the predictor. In other words, searching for predictors of high quality is the same thing as searching for predictors with high $\mathrm{R}^2$s. Note that, as usual in rational expectations model, we assume that there is no uncertainty on $\theta$, i.e., on the true predictive model relating the payoff of the asset to the predictor. In reality, investors might be uncertain about the true $R^2$ of a predictive model (e.g., because of too few past observations for past cash-flows relative to the number of variables used to forecast these cash-flows) and learn it over time (see Martin and Nagel (2020)). In our model, this means that speculators would learn about the true $\theta$ of a predictor (e.g., after observing an estimate of $\theta$). We leave this extension for the future.

predictors constant). This evolution has two consequences controlled by parameters $\underline{\theta}$ and $\alpha$ in the model. First, it pushes back the "data frontier", i.e., it improves(at least weakly) the quality of the most informative predictor (the "hidden gold nugget effect.") 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"), $\tau^{max}$ improves.

Second, while the number of combinations of variables that one can consider to build predictors becomes very large, the number of combinations that actually yield informative predictors might fall. 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, only 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."[12] It is controlled by $\alpha$ in our model: As $\alpha$ decreases, each round of exploration is less likely to be successful as if the share of informative predictors was falling.[13]

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.

We focus on equilibria in which each speculator follows an optimal stopping rule $\theta_i^*$. 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 for speculator $i$ in a given search round:

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

Thus, a decrease in $\underline{\theta}$ raises the likelihood of finding a predictor in a given exploration, holding $\alpha$ constant. This effect captures the idea that while data abundance might reduce

[12] Agrawal, McHale, and Oettl (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.

[13] 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."*

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). Let $n_i$ be 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:

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

**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, $s_\theta$ 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 denoted by $\eta$, where $\eta \sim \mathcal{N}(0, \nu^2)$ ($\eta$ is independent of $\omega$ 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 exploration rounds for speculator $i$, $n_i$, is independent from the asset payoff, its price, and the realization of the speculator's predictor, $s_\theta$, 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 ($\theta_i^*$) because this rule determines the distribution of $s_\theta$. The expected utility cost of exploration depends on the speculator's stopping rule, $\theta_i^*$, 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 one signal of known quality). In our model, $n_i$ is random and its distribution is controlled by the speculator via her stopping rule.

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

## 3.2 Discussion of the assumptions on the learning technology

In our framework, a new round of exploration does not necessarily yield a predictor of better quality than in a previous round. This captures the idea that, in reality, asset managers limits the number of variables that they use to build their predictor to reduce the cost of data and the risk of overfitting. In such a scenario, launching a new round of exploration does not guarantee that one will find a predictor of better quality. For instance, suppose that in each round, a speculator uses $K$ variables to predict the asset payoff and that each variable is equal to the asset payoff plus noise. The precision of each variable is determined according to some distribution, and therefore possibly different across variables. Now suppose that after obtaining a predictor, a speculator launches a new round of exploration by replacing the $K^{th}$ variable with a new one whose precision is a priori unknown. In this case, the quality of the predictor obtained in the new round is not necessarily higher than that of the predictor obtained in the first round (see Section A.III in the online appendix).

An alternative learning technology is one in which a new round of exploration necessarily yields a predictor of better quality than that found in the previous round. In this case, the model is significantly less tractable because each speculator's optimal stopping time is path dependent (i.e., specific to the history of the predictor she found until some point). We therefore leave this case for future analysis.

We also assume that if a speculator turns down a predictor, she cannot recall it. This assumption simplifies the exposition but it can be relaxed. To this end, in Section A.I of the online appendix, we consider the more realistic case in which when speculators decide to stop searching, they can use the best predictor they found so far (they do not forget predictors). We show that the results in this case are identical to those obtained when speculators cannot recall the best predictor. The reason is that the search problem faced by a speculator is stationary because there is no limit on the number of exploration rounds for a speculator.

Last, we assume that speculators draw the type of their predictors according to the unconditional distribution of predictors' type ($\phi(.)$) in the interval $[0, \pi/2]$ but that they cannot exploit predictors with a type $\theta < \underline{\theta}$. Alternatively, we could assume that speculators draw the type of their predictors in $[\underline{\theta}, \pi/2]$, conditional on this type being in this interval. We show in the online appendix (Section A.IV) that this approach yields the same results.

# 4. Equilibrium Data Mining

## 4.1 Equilibrium

We focus on symmetric equilibria in which all speculators choose the same stopping rule, $\theta^*$. We solve for such an equilibrium as follows. First, we solve for the equilibrium of the trading stage in period 1 taking $\theta^*$ as given and we deduce the ex-ante expected utility achieved by speculator $i$ when she chooses a predictor of type $\theta$ in period 0. We then observe that a speculator should stop searching when 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 exploration. 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 $\theta^*$ by observing that, in a symmetric equilibrium, each speculator's best response to other speculators' stopping rule, $\theta^*$, 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 $\theta^*$:

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

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 guarantees that the equilibrium remains well defined even when $\underline{\theta} = 0$.[14] 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 $\theta$ 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*

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

This result extends Proposition 1.1 in Vives (1995) to the case in which speculators have signals of heterogenous precisions (determined by their $\theta$ in our model). The predictor $s_\theta$ 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).

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

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 $\xi$, 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 about this payoff when the average quality of speculators' predictors, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$, is higher. Formally, let measure the informativeness 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 $\theta^*$ because $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$ decreases with $\theta^*$. 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 $\theta$ and other speculators follow the stopping rule $\theta^*$. We denote this ex-ante expected utility by $g(\theta, \theta^*)$ and refer to it as the trading value of a predictor with type $\theta$. 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 $\theta$ is:*

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

The trading value of a predictor increases with its quality and decreases with the informativeness of the asset price.[15] 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 $\theta^*$ decreases).

Armed with Lemma 1, we can now derive a speculator's optimal stopping rule given that other speculators follow the stopping rule $\theta^*$. Let $\widehat{\theta}_i$ be an arbitrary stopping rule for speculator $i$. The speculator's continuation utility (the expected utility of launching a new round of exploration) after turning down a predictor is:

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

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 times the speculator's continuation utility when she turns down a predictor. Solving eq.(14) for $J(\widehat{\theta}_i, \theta^*)$, we obtain:

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

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 $\theta$. If the speculator stops exploring the data at this stage, her expected utility is $g(\theta, \theta^*)$ (her cost of

[15] Observe that $\frac{\mathsf{Var}[\mathsf{E}[\omega | s_\theta, p^*] - p^*]}{\mathsf{Var}[\omega | s_\theta, p^*]} = \frac{\mathsf{E}[(\mathsf{E}[\omega | s_\theta, p^*] - p^*)^2]}{\mathsf{Var}[\omega | s_\theta, p]}$ because $\mathsf{E}[\omega | s_\theta, p^*] - p^* = 0$. Thus, eq.(13) implies that, $\frac{\tau(\theta)\tau_\omega}{\mathcal{I}(\theta^*; \underline{\theta}, \alpha)} = \mathsf{E}\left[ (\frac{\mathsf{E}(R_\theta | s_\theta)}{\sigma_{R_\theta | s_\theta}})^2 \right]$, where $R_\theta = \omega / p^* - 1$ is the excess return of a speculator with type $\theta$ (the riskless rate ofd return is normalized to zero) and $\sigma_{R_\theta | s_\theta}$ is the standard deviation of this return conditional on the observation of $s_\theta$. In other words, $\frac{\tau(\theta)\tau_\omega}{\mathcal{I}(\theta^*; \underline{\theta}, \alpha)}$ is the equilibrium value of the expected square Sharpe ratio of a speculator trading on a predictor with type $\theta$.

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 $\theta$, the optimal stopping rule of the speculator, $\theta_i^*(\theta^*)$, is the value of $\theta$ such that the speculator is just indifferent between these two options:

$$g(\theta_i^*, \theta^*) = J(\theta_i^*, \theta^*). \tag{16}$$

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

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

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). Assumption A.1 guarantees 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 $\exp(-\rho c) \leq F(\pi/2)$ (i.e., $c$ large enough), 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 (if $c$ is not too large). 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 $\alpha$) and computing power (a decrease in $c$) affect the quality of the worst predictor on which speculators trade in equilibrium, i.e., $\tau(\theta^*)$. Indeed, the quality of this predictor determines the range of predictors used in equilibrium and ultimately several equilibrum outcomes of interest (see next section).

**Proposition 3.** *A decrease in the cost of exploration, c, always reduces the stopping rule $\theta^*$ 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 $\theta^*$ 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 $\theta^*$. One indirect consequence of this behavior is that, on average, speculators trade more aggressively on their signal (the "competition effect") because they face less uncertainty on the asset payoff (their predictors are better on average). 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, $\alpha$, always increases speculators' stopping rule, $\theta^*$, 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 $\alpha$ 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 ($\Lambda$). 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 $\theta^*$ (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 $\alpha$ 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 $\alpha$ is similar to that of an increase in the per exploration cost.[16]

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,

[16] 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.

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, $\theta^*$ 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)).

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 $\theta^*$ 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 $\theta^*$ constant):

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

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 data 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 with two particular specifications of the distribution for

$\theta$: (i) $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ and (ii) $\phi(\theta) = 5\cos(\theta)\sin^4(\theta)$. With these specifications, one can compute all variables of interest in closed forms (see Section A.VI in the internet appendix, where we also derive the corresponding distribution for predictors' quality, $\tau(\theta)$).[17] Figure 2 below shows the effect of a change in the exploration cost ($c$) and the data frontier ($\underline{\theta}$) on the equilibrium value of $\theta^*$. 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 $\theta^*$) but, eventually, at some point this effect is reversed.

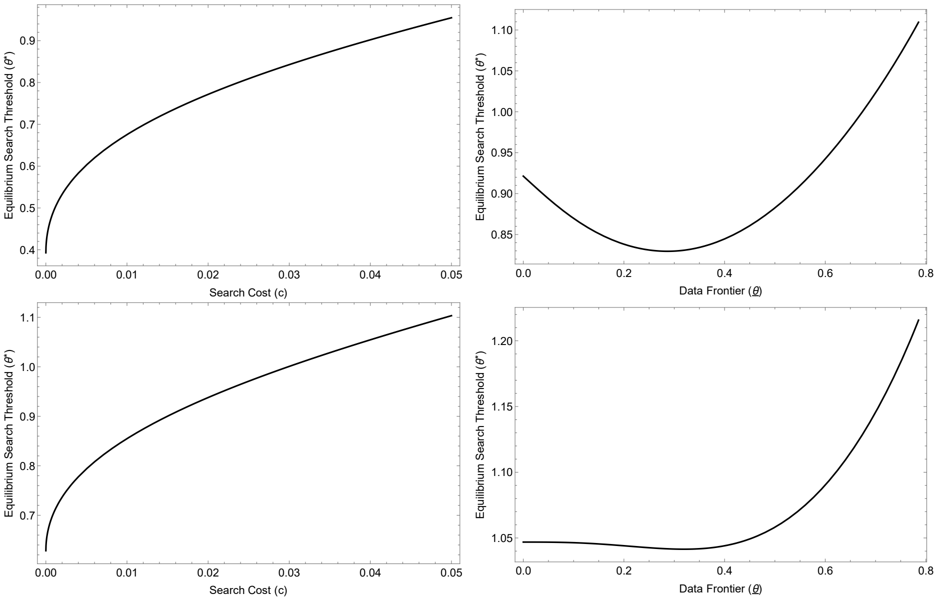

**Figure 2:** Left-hand-side: Equilibrium search threshold, $\theta^*$, as a function of the search cost, $c$ (other parameter values are $\underline{\theta} = \pi/8, \rho = \sigma^2 = \nu^2 = 1$). Right-hand-side: Equilibrium search threshold, $\theta^*$, as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.03, \rho = \sigma^2 = \nu^2 = 1$). Upper graphs: $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$. Lower graphs: $\phi(\theta) = 5\cos(\theta)\sin^4(\theta)$.

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

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

[17] Assumption A.1 is satisfied in both examples. The main difference is that the mass is shifted to the left in the first case. That is, the likelihood of finding a predictor of high quality in a given exploration is higher with the first distribution than with the second. See Section A.VI in the internet appendix.

# 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 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, $\beta_\theta$, is:

$$\beta_\theta = \frac{\mathsf{Cov}(x(s_\theta, p^*), \omega - p^*)}{\mathsf{Var}[\omega - p^*]} = \frac{\tau(\theta)}{\rho\sigma^2}, \tag{22}$$

where the last equality follows from Proposition 1. Intuitively, $\beta_\theta$ is a measure of a speculator's stock picking ability or investment "skills".[18] Equation (22) shows that, holding risk aversion constant, a ranking of speculators based on their stock picking ability (measured by $\beta_\theta$) is identical to a ranking based on the (unobservable) quality of their predictors, $\tau(\theta)$.

Thus, one could test the implications of Propositions 3 and 4 by ranking speculators (e.g., managers of active mutual funds or hedge funds) based on their stock picking ability (measured by $\beta$s) and test whether shocks to computing power or data abundance have the effects predicted by Propositions 3 and 4.[19] For instance, one could test whether positive shocks to computing power increase the stock picking ability (measured by $\beta$) of the funds with the lowest $\beta$s' (say in the lowest decile) while positive shocks to data abundance (e.g., the availability of new alternative data as in Zhu (2019) or Dessaint, Foucault, and Frésard (2021)) have the opposite effect (even though they may increase

[18] Kacperczyk, van Nieuwerburgh, and Veldkamp (2014) measure mutual funds' stock picking ability in a similar way. See Section I.B in their paper. More generally, this measure is related to holdings based measures of mutual funds performance; see Grinblatt and Titman (1993) and Daniel, Grinblatt, Titman, and Wermers (1997)

[19] 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 $s_\theta$.

the stock picking ability of the best performing funds). One 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 $\alpha$) reduces the average quality of speculators' predictors and therefore price informativeness.*

Thus, the effect of data abundance on price informativeness is ambiguous. Holding $\alpha$ 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 (when 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 $\alpha$) 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 $\alpha$). As a result, the net effect of data abundance on the long run evolution of asset price informativeness is ambiguous, as shown in Figure 3 (in which we assume that $\alpha = \min\{1, 0.32 + 0.8 \times \underline{\theta}\})$).

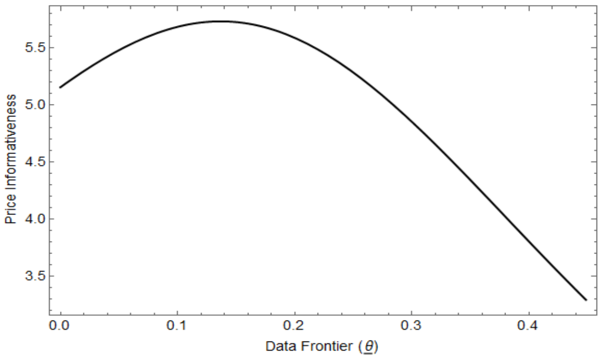

**Figure 3:** This graph shows the evolution of price informativeness in equilibrium, $\mathcal{I}(\theta^*, \underline{\theta})$ as a function of the data frontier, $\underline{\theta}$ when $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ and $\alpha = \min\{1, 0.32 + 0.8 * \underline{\theta}\}$. Other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

Consistent with these implications, empirical findings regarding the effect of progress in information technologies on price informativeness are ambiguous. For instance, Bai, Phillipon, and Savov (2016) find that the price stocks in the S&P500 has become more informative over time while Farboodi, Matray, and Veldkamp (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, Yang, and Zuo (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 considering shocks that only affect computing power or abundance (rather than both dimensions simultaneously) would help to make progress in understanding how progress in information technologies affect asset price informativeness.

## 5.3 Data abundance, Computing Power and Trading Profits

In equilibrium, the total trading profit ("excess return"), $\pi(s_\theta)$, of a speculator with type $\theta$ 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:

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

Using eq.(23), the *expected* trading profit of a speculator with type $\theta$ is therefore

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

where the last equality follows from the fact that $p^* = \mathsf{E}(\omega \mid p^*)$ so that $\mathsf{Var}[\omega - p^*] = \mathsf{Var}[\omega \mid p^*] = (\mathcal{I}(\theta^*, \underline{\theta}))^{-1}$ (by definition of $\mathcal{I}(\theta^*, \underline{\theta})$).

Thus, 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, $\mathsf{E}[\bar{\pi}(\theta)]$ and $\mathsf{Var}[\bar{\pi}(\theta)]$ could be measured by the cross-sectional mean and variance of trading profits of active funds (for instance in a given quarter). Another possibility is to consider the distribution (across funds) of the squared Sharpe Ratio (the ratio of average excess returns for a fund divided by the standard deviation of returns) of active funds. Indeed, $\bar{\pi}(\theta)$ is equal to the expected squared Sharpe ratio of a speculator with type $\theta$, divided by her risk aversion (see Footnote 15). Thus, $\mathsf{E}[\bar{\pi}(\theta)]$ and $\mathsf{Var}[\bar{\pi}(\theta)]$ can also be interpreted as the mean and variance of the distribution of squared Sharpe ratios across funds.

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 term equal to $\frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\rho\sigma^2\mathcal{I}(\theta^*, \underline{\theta})}$), the first effect raises speculators' expected profit while the second reduces it. Using eq.(26), we find that the first effect dominates if and only if $\bar{\tau}(\theta^*; \underline{\theta}, \alpha) \leq (\tau_\omega\rho^2\nu^2)^{1/2}$. Thus, speculators' average expected profit reaches its maximum for $\bar{\tau}(\theta^*(\underline{\theta}, c, \alpha), \underline{\theta}, \alpha) = (\tau_\omega \rho^2 \nu^2)^{1/2}$ if there are values of $(\underline{\theta}, c, \alpha)$ for which this equality holds (we write $\theta^*$ as a function of $\underline{\theta}$,$c$,$\alpha$ to emphasize that it depends on the value of these parameters). We deduce the following result.

**Proposition 8.**

1. *If $\bar{\tau}(\theta^*(\underline{\theta}, 0, \alpha), \underline{\theta}, \alpha) > (\tau_\omega \rho^2 \nu^2)^{1/2}$ 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 $\bar{\tau}(\theta^*(0, c, \alpha), 0, \alpha) > (\tau_\omega \rho^2 \nu^2)^{1/2}$ 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 $\bar{\tau}(\theta^*(\underline{\theta}, c, 1), \underline{\theta}, 1) > (\tau_\omega \rho^2 \nu^2)^{1/2}$ then speculators' expected profit is a hump shaped function of $\alpha$, which reaches its maximum for $\alpha = \hat{\alpha}$ (characterized in the proof of the proposition). Otherwise, speculators' expected profit increases with $\alpha$ 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 (see Figure 4 for a numerical example). A decrease in $\alpha$ has the opposite effects: It reduces the average quality of speculators' signals and price informativeness. The first effect reduces speculators' expected profit while the second increases this expected profit. The former effect always dominates when $\alpha$ is small enough. Overall, these findings suggest that there can be a point at which further improvements in computing power or data availability reduces speculators' expected profit.

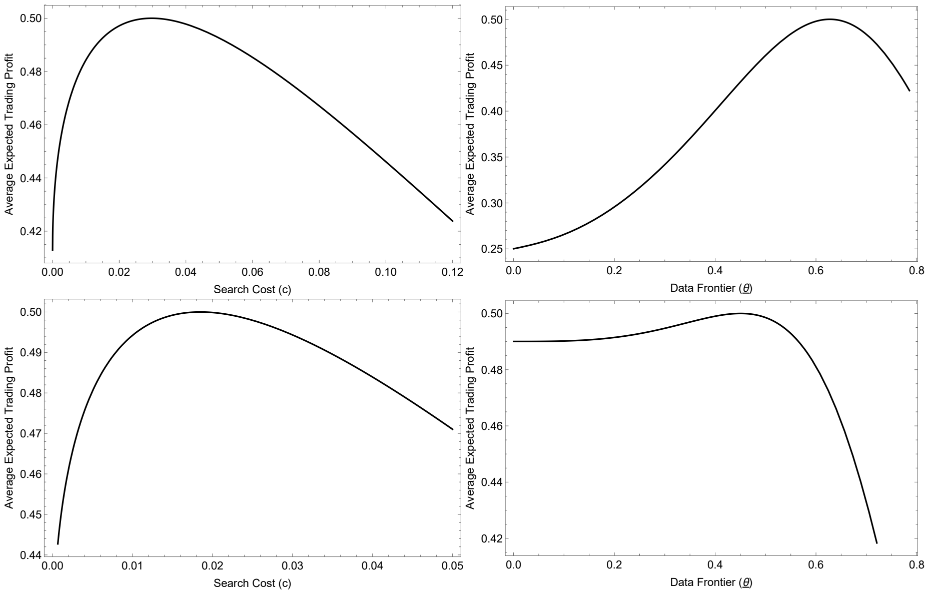

**Figure 4:** Left: Speculators' expected profits, $\mathsf{E}(\bar{\pi})$, as a function of the search cost, $c$ (other parameter values are $\underline{\theta} = \pi/5, \rho = \sigma^2 = \nu^2 = 1$). Right: Speculators' expected profits, $\mathsf{E}(\bar{\pi})$, as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.05, \rho = \sigma^2 = \nu^2 = 1$). Upper graphs: $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$. Lower graphs: $\phi(\theta) = 5\cos(\theta)\sin^4(\theta)$.

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).*

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, i.e., $\theta^* = \underline{\theta}$. 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 (see (eq.(27)). Now consider a small increase in $c$ starting from the situation in which $c = 0$.

This increase raises $\theta^*$ 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 5 below for an example).

When $\underline{\theta} < \underline{\theta}^{tr}(c)$, pushing back the data frontier further raises the quality of the best predictor and reduces the quality of the worst predictor used by speculators (see Proposition 4). Thus, the range of quality for the predictors used in equilibrium widen. 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, which dampens the previous effect (since $\mathsf{Var}[\pi(\theta)]$ is inversely related to price informativeness; see eq.(27)). However, for $\underline{\theta}$ small enough, this second effect is not sufficient to offset the first. This explains the second part of the proposition.

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 (see Figure 5).

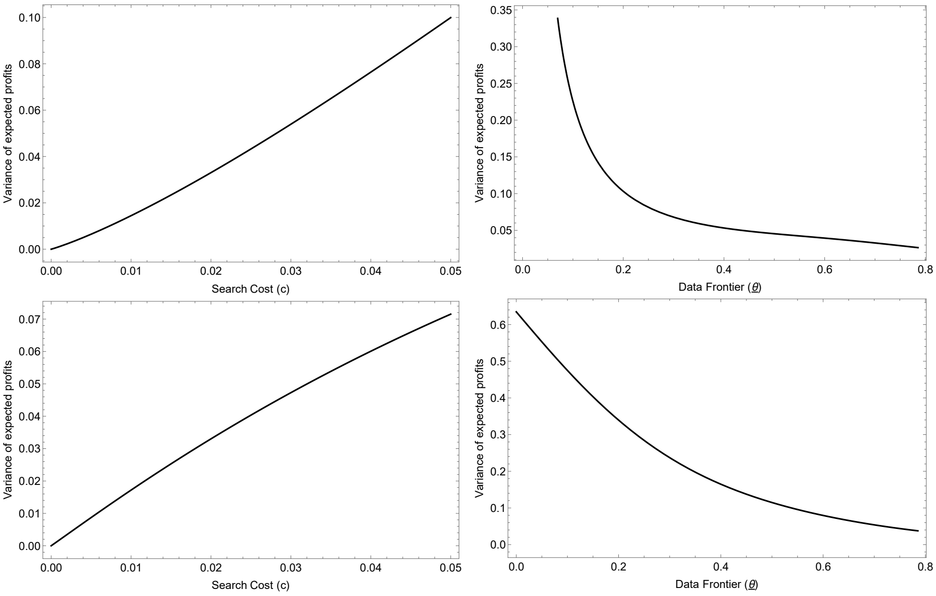

**Figure 5:** Left: Variance of speculators' expected profits, $\mathsf{Var}[\pi(\theta)]$, as a function of the search cost, $c$ (other parameter values are $\underline{\theta} = \pi/5, \rho = 1, \sigma^2 = 1, \nu^2 = 1$). Right: Variance of speculators' expected profits as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.05, \rho = 1, \sigma^2 = 1, \nu^2 = 1$). Upper graphs: $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$. Lower graphs: $\phi(\theta) = 5\cos(\theta)\sin^4(\theta)$.

## 5.4 Data Abundance, Computing Power and Crowding

Practitioners refers to the tendency for investors to follow the same trading strategy and exploit the same signals as "crowding".[20] 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 $\theta_i$ and a speculator with type $\theta_j$. Using eq.(24) and the fact that $\mathsf{Var}[\omega - p^*] = (\mathcal{I}(\theta^*, \underline{\theta}))^{-1}$, 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 $\theta_i$ and a speculator with type $\theta_j$ (a measure of crowding) is:

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

[20] 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.*

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, $\omega$, 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 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.*

Testing Proposition 10 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' 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 $\theta^*$ 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 via 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, a decrease in $\underline{\theta}$ always raises price informativeness (Proposition 7) and, when $\underline{\theta} < \underline{\theta}^{tr}(c)$, it 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 (even though, it can reduce their average gross trading profits; see Proposition 8). Figure 6 illustrates this point. For similar reasons, the needle in the haystack problem (a decrease in $\alpha$) has an ambiguous effect on speculators' welfare: It reduces price informativeness but it 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 $\alpha$ low enough.

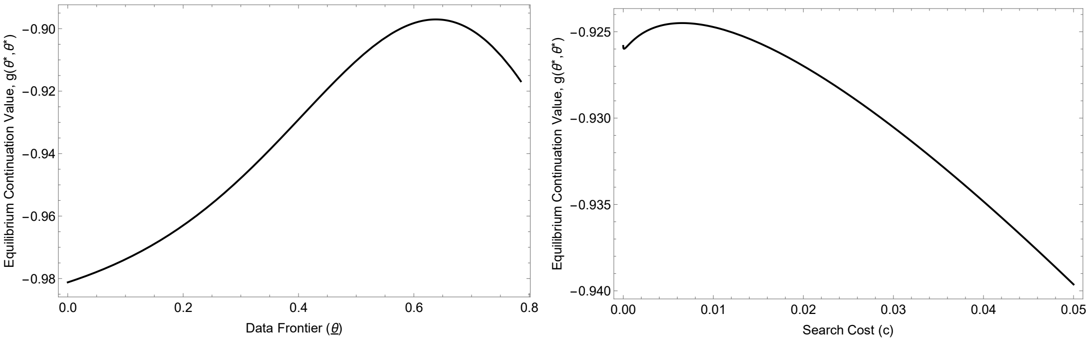

**Figure 6:** Speculators' ex-ante expected utility as a function of $\underline{\theta}$ and $c$ when $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ (other parameter values: $\rho = 1, \sigma^2 = 1, \nu^2 = 1$).

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 the emergence of new datasets enable investors to reduce $\underline{\theta}$ from $\underline{\theta}_0$ to $\underline{\theta}_1 < \underline{\theta}_0$ but that speculators agree not to take advantage of the new datasets. In this case, each speculator obtains an expected utility equal to

$J(\theta^*(\underline{\theta}_0), \theta^*(\underline{\theta}_0))$. If an investor secretely deviates by acquiring the new data, she does not affect the equilibrium of the trading stage. Hence, price informativeness is unchanged. It follows that if the investor finds a predictor with a type in $[\underline{\theta}_0, \theta^*(\underline{\theta}_0)]$, her expected utility of trading on this predictor is unchanged. However, in addition, the speculator has the possibility to finds a predictor with a type in $[\underline{\theta}_1, \theta^*(\underline{\theta}_0)]$ and her expected utility from trading on a predictor with a type in this range is strictly higher than her expected utility from trading on a predictor with a type in $[\underline{\theta}_0, \theta^*(\underline{\theta}_0)]$. Thus, the deviation is profitable for the speculator. In other words, each speculator individually finds optimal to use the new datasets, if she expects others not to do so. Hence, unless speculators can credibly commit not using the new datasets, all of them do, which makes them collectively worse off than if they did not.

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 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, $\theta_r$ and commit to this choice. In this case, speculators would optimally choose the stopping rule $\theta_r^{**}$ such that:

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

**Proposition 12.** *In a symmetric interior equilibrium of the exploration phase, 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, $\mathsf{E}(n_i c)$, is higher than the investment that would maximize their welfare if they could collectively choose their stopping rule.*

Thus, there is excessive investment in search for predictors in equilibrium from speculators' viewpoint. The reason is as follows. When speculators choose a more stringent stopping rule, they expect to trade on a predictor of better quality on average, which raises their expected utility. However, this choice raises their expected utility cost of exploration (as it will take more exploration rounds to find a predictor) and price informativeness (as speculators trade more aggressively on more precise signals). Both effects reduce their expected utility. When they individually choose their stopping rule, each speculator accounts for the first cost but ignores the second (as each speculator is too "small' to affect price informativeness). In contrast, a central planner organizing the search for predictors in speculators' interests internalizes both costs and therefore choose a less stringent stopping rule than that chosen individually by speculators.

# 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 via 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 generates 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 correlation 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 a drop in price informativeness.

# References

Abis, Simona, 2018, Man vs machine: Quantitative and discretionary equity management, Discussion paper, .

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, Discussion paper, .

Brogaard, Jonathan, and Abalfazl Zareei, 2019, Machine learning and the stock market, Discussion paper, .

Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers, 1997, Measuring mutual fund performance with characteristic based benchmarks, *Journal of Finance* 52, 1035–1058.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard, 2021, Does alternative data affect financial forecasting? the horizon effect, Discussion paper, Working Paper.

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?, Discussion paper, .

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* pp. 1367–1411.

Garleanu, Nicolae, and Lasse Heje Pedersen, 2018, Efficiently inefficient markets for assets and asset management, *Journal of Finance* 78, 1163–1711.

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

Grinblatt, Mark, and Sheridan Titman, 1993, Performance measurement without benchmarks: An examination of mutual fund returns, *Journal of Business* 66, 42–68.

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

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *Review of Financial Studies* 33, 2223–2273.

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

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

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, Discussion paper, .

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

Martin, Ian, and Stefan Nagel, 2020, Market efficiency in the age of big data, *Working paper, LSE and University of Chicago.*

Milhet, Roxana, 2020, Financial innovation and the inequality gap, Discussion paper, .

Narang, Rishi, 2013, *Inside the Black Box: A simple guide to quantitative and high-frequency trading* (Wiley: New-York).

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).

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

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

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:

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

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

Now consider speculators. Using standard calculations in the CARA gaussian framework, we obtain that the optimal demand for the risky asset of a speculator with signal $s_\theta$ 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 $\xi$, as in eq.(32). Moreover, let $\hat{s}_\theta \equiv \omega + \tau(\theta)^{-\frac{1}{2}}\epsilon_\theta$, so that $s_\theta = \cos(\theta)\hat{s}_\theta$. Thus,

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

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 $\epsilon_\theta$ and $\eta$ (the noises in $\hat{s}_\theta$ and $\xi$) 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, $s_\theta$, the expected utility of trading for an investor given her optimal trading strategy is:

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

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

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

Thus:

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

For a normally distributed variable $Z$ with mean 0 and variance $\sigma_Z^2$, $\mathsf{E}[\exp(-Z^2)] = (1 + 2\sigma_Z^2)^{-1/2}$. 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:

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

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:

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

Thus, using the expression for $\mathsf{Var}[\omega|s_\theta, p^*]$ in eq.(37), we deduce that:

$$\mathsf{Var}[x^*(s_\theta, p^*] = \frac{\tau(\theta)\tau_\omega}{\rho^2(\tau_\omega + \tau_\xi) \, \mathsf{Var}[\omega|s_\theta, p^*]}. \tag{47}$$

Hence, using eq.(44) and the fact that $\tau_\xi = \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2 \tau_\omega^2}{\rho^2 \nu^2}$, we deduce that:

$$\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta, p] - p]}{\mathsf{Var}[\omega|s_\theta, p]} = \frac{\tau_\omega \tau(\theta)}{\tau_\omega + \frac{(\tau_\omega \bar{\tau}(\theta^*; \underline{\theta}, \alpha))^2}{\rho^2 \nu^2}}. \tag{48}$$

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

**Proof of Proposition 2.** The derivative of $F(\theta^*)$ is

$$\frac{\partial F}{\partial \theta^*} = \alpha \int_{\underline{\theta}}^{\theta^*} \frac{\partial r(\theta, \theta^*)}{\partial \theta^*} \phi(\theta) d\theta, \tag{49}$$

where $r(\theta, \theta^*)$ is defined in eq.(20). As $\theta^*$ increases, both $\tau(\theta^*)$ and $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ decreases. We deduce that $r(\theta, \theta^*)$ decreases in $\theta^*$. 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 $\theta^*$ in the proof of Proposition 2. It immediately follows from these two observations that $\theta^*$ increase in $c$.

**Proof of Proposition 4.**

**Part 1.** It directly follows from eq.(19) that $\frac{\partial F}{\partial \alpha} = -\int_{\underline{\theta}}^{\theta^*}(1 - r(\theta, \theta^*)\phi(\theta)d\theta) < 0$, since $r < 1$. Thus, $F(\theta^*)$ decreases in $\alpha$. As $F(.)$ also decreases in $\theta^*$ and, in equilibrium, $F(\theta^*) = \exp(-\rho c)$, it immediately follows that $\theta^*$ increases in $\alpha$, as claimed in the first part of the proposition.

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

$$r(\theta, \theta^*) = \frac{g(\theta, \theta^*)}{g(\theta^*, \theta^*)} = \left(\frac{\rho^2\sigma^2\nu^2\tau(\theta^*) + \rho^2\sigma^2\nu^2 + \bar{\tau}^2(\theta^*; \underline{\theta}, \alpha)}{\rho^2\sigma^2\nu^2\tau(\theta) + \rho^2\sigma^2\nu^2 + \bar{\tau}^2(\theta^*; \underline{\theta}, \alpha)}\right)^{\frac{1}{2}}. \tag{50}$$

The ratio $(a + x)/(b + x)$ increases with $x$ iff $a < b$. Thus, as $\tau(\theta) > \tau(\theta^*)$, the sign of $\frac{\partial r}{\partial \underline{\theta}}$ is the same as the sign of $\frac{\partial \bar{\tau}}{\partial \underline{\theta}}$ because $\tau(\theta) > \tau(\theta^*)$. We obtain:

$$\frac{\partial \bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} = -\phi^*(\underline{\theta})\left(\tau(\underline{\theta}) - \bar{\tau}(\theta^*; \underline{\theta}, \alpha)\right) < 0, \tag{51}$$

where the last inequality follows from the fact $\tau(\theta)$ decreases with $\theta$. Thus, $\frac{\partial r}{\partial \underline{\theta}} < 0$.

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:

$$\frac{\partial F}{\partial \underline{\theta}} = \underbrace{\alpha\phi(\underline{\theta})(1 - r(\underline{\theta}, \theta^*))}_{>0} + \alpha \int_{\underline{\theta}}^{\theta^*} \underbrace{\frac{\partial r}{\partial \underline{\theta}}}_{<0} \phi(\theta)d\theta. \tag{52}$$

Thus, the effect of $\underline{\theta}$ on $F(.)$ and therefore the equilibrium stopping rule $\theta^*$ is ambiguous. We now show that this effect becomes negative when $\underline{\theta}$ is close enough to zero. To see this, observe that eq.(52) implies that:

$$\frac{\partial F}{\partial \underline{\theta}} < \alpha\phi(\underline{\theta})\left(1 + \frac{\int_{\underline{\theta}}^{\theta^*} \frac{\partial r}{\partial \underline{\theta}}\phi(\theta)d\theta}{\phi(\underline{\theta})}\right) \tag{53}$$

We show in Section 4 of the internet appendix that $\frac{\int_{\underline{\theta}}^{\theta^*} \frac{\partial r}{\partial \underline{\theta}}\phi(\theta)d\theta}{\phi(\underline{\theta})}$ 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 $\theta^*$, it follows that $\theta^*$ increases in $\underline{\theta}$ when $\underline{\theta} < \underline{\theta}^{tr}$, as claimed in the second part of the proposition.

**Proof of Proposition 5.** It follows from direct inspection of the expression for $r(\theta, \theta^*)$ given in eq.(50) that $r(\theta, \theta^*)$ decreases with $\sigma^2$, and $\nu^2$ because $\tau(\theta) > \tau(\theta^*)$. Thus, from eq.(19), we deduce that $F(\theta^*)$ decreases with $\sigma^2$, and $\nu^2$. It follows from this observation, the fact $F(\theta^*)$ decreases with $\theta^*$ and the equilibrium condition $F(\theta^*) = \exp(-\rho c)$ that $\theta^*$ decreases with $\sigma^2$ and $\nu^2$.

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

**Proof of Proposition 7.**

**Part 1.** When a decrease in $\underline{\theta}$ reduces $\theta^*$, 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 $\underline{\theta}$ increases $\theta^*$. We know that this possibility arises when $\underline{\theta}$ is low enough (see Proposition 4). We prove below, by contradiction, that price informativeness, $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$, is also inversely related to $\underline{\theta}$ in this case.

Suppose (to be contradicted) that there is a value of $\underline{\theta}$ such that when $\frac{\partial \theta^*}{\partial \underline{\theta}} < 0$ then $\frac{\partial \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} > 0$. Let $L(\theta_i^*, \theta^*)$ be:

$$L(\theta_i^*, \theta^*) \equiv \alpha \int_{\underline{\theta}}^{\theta_i^*} \frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)} \phi(\theta) d\theta + 1 - \alpha \int_{\underline{\theta}}^{\theta_i^*} \phi(\theta) d\theta. \tag{54}$$

Function $L$ is decreasing with $\theta_i^*$ because:

$$\frac{\partial L}{\partial \theta_i^*} = \alpha \int_{\underline{\theta}}^{\theta_i^*} \frac{\partial}{\partial \theta_i^*} \left( \frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)} \right) \phi(\theta) d\theta < 0. \tag{55}$$

Now, using the expression for $J(.)$ given in eq.(15), we can rewrite the indifference condition (16) as:

$$L(\theta_i^*, \theta^*) = \exp(-\rho c). \tag{56}$$

Moreover: $L(\underline{\theta}, \theta^*) = 1$ and $0 < L(\pi/2, \theta^*) < 1$. Thus, as $L(\theta_i^*, \theta^*)$ decreases in $\theta_i^*$ , eq.(54) has a unique solution $\theta_i^*(\theta^*)$ when $c$ is small enough. This solution defines the best response of a speculator when other speculators choose the stopping rule $\theta^*$.

Next, for $\theta_i^* \geq \theta \geq \underline{\theta}$, define

$$l(\theta, \theta_i^*, \theta^*) = \frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)} = \left(\frac{\rho^2\sigma^2\nu^2\tau(\theta_i^*) + \rho^2\nu^2 + \sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2}{\rho^2\sigma^2\nu^2\tau(\theta) + \rho^2\nu^2 + \sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2}\right)^{\frac{1}{2}} = \left(\frac{\tau(\theta_i^*)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\tau(\theta)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}\right)^{\frac{1}{2}}. \tag{57}$$

Clearly, $l(\theta, \theta_i^*, \theta^*)$ increases when $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ increases. Thus, if $\frac{\partial \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} < 0$, then $\frac{\partial l(\theta, \theta_i^*, \theta^*)}{\partial \underline{\theta}} > 0$ since $\underline{\theta}$ affects $l(\theta, \theta_i^*, \theta^*)$ only through its effect on price informativeness. This implies that:

$$\frac{\partial l}{\partial \underline{\theta}} + \frac{\partial l}{\partial \theta^*}\frac{\partial \theta^*}{\partial \underline{\theta}} > 0. \tag{58}$$

As:

$$L(\theta_i^*, \theta^*) \equiv \alpha \int_{\underline{\theta}}^{\theta_i^*} l(\theta, \theta_i^*, \theta^*) + 1 - \alpha \int_{\underline{\theta}}^{\theta_i^*} \phi(\theta) d\theta, \tag{59}$$

we deduce that:

$$\frac{dL}{d\underline{\theta}} = \frac{\partial L}{\partial \underline{\theta}} + \frac{\partial L}{\partial \theta^*}\frac{\partial \theta^*}{\partial \underline{\theta}} = \underbrace{\alpha\phi(\underline{\theta})(1 - l(\underline{\theta}, \theta_i^*, \theta^*))}_{>0} + \alpha \int_{\underline{\theta}}^{\theta_i^*} \left(\frac{\partial l}{\partial \underline{\theta}} + \frac{\partial l}{\partial \theta^*}\frac{\partial \theta^*}{\partial \underline{\theta}}\right) \phi(\theta) d\theta. \tag{60}$$

Eq.(58) implies that the second term is also positive. Thus, $\frac{dL}{d\underline{\theta}} > 0$. Thus, a decrease in $\underline{\theta}$ results in a smaller value of $L$, holding $\theta_i^*$ constant. As $\partial L / \partial \theta_i^* < 0$ and $L(\theta_i^*, \theta^*) = \exp(-\rho c)$, it follows that in this case $\theta_i^*$ increases with $\underline{\theta}$. As, in equilibrium, $\theta_i^* = \theta^*$, this also implies that $\frac{\partial \theta^*}{\partial \underline{\theta}} > 0$. A contradiction with our starting hypothesis. We deduce that when $\frac{\partial \theta^*}{\partial \underline{\theta}} < 0$ then $\frac{\partial \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} < 0$. Thus, for all values of $\underline{\theta}$, a decrease in $\underline{\theta}$ improves price informativeness.

**Part 2.** By definition, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha) = \int_{\underline{\theta}}^{\theta^*} \tau(\theta)\phi^*(\theta)\theta$. Using the definition of $\phi^*(\theta)$, we deduce that $\frac{\partial \bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\partial \alpha} = \frac{\partial \theta^*}{\partial \alpha}(\phi^*(\theta^*)(\tau(\theta^*) - \bar{\tau}(\theta^*; \underline{\theta}, \alpha)) > 0$, where the last inequality follows from the fact that $\tau(\theta)$ decreases with $\theta$ and $\frac{\partial \theta^*}{\partial \alpha} < 0$ (see Proposition 4). Hence, price informativeness increases with $\alpha$ because (i) $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ increases with $\bar{\tau}(\theta^*)$ and (ii) depends on $\alpha$ 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, $\lim_{\underline{\theta} \to \frac{\pi}{2}} \bar{\tau}(\theta^*(\underline{\theta}, c, \alpha), \underline{\theta}, \alpha) = \tau(\frac{\pi}{2}) = 0$. Thus, if $\bar{\tau}(\theta^*(0, c, \alpha), 0, \alpha) > (\tau_\omega \rho^2 \nu^2)^{1/2}$, there is a unique value of $\theta$, denoted $\hat{\theta}$, such that $\bar{\tau}(\theta^*(\hat{\theta}, c, \alpha), \hat{\theta}, \alpha) = (\tau_\omega \rho^2 \nu^2)^{1/2}$. Consequently, 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. The proofs of Parts 1 and 3 are similar and therefore omitted the proofs for brevity. In these cases, one obtains that $\hat{c}$ and $\hat{\alpha}$ are the unqiue solutions of, respectively, $\bar{\tau}(\theta^*(\underline{\theta}, \hat{c}, \alpha), \underline{\theta}, \alpha) = (\tau_\omega \rho^2 \nu^2)^{1/2}$ and $\bar{\tau}(\theta^*(\underline{\theta}, c, \hat{\alpha}), \underline{\theta}, \hat{\alpha}) = (\tau_\omega \rho^2 \nu^2)^{1/2}$.

**Proof of Proposition 9.**

**Part 1.** 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.

**Part 2.** In order to analyze the effect of $\underline{\theta}$, it is useful to rewrite $\mathsf{Var}[\pi(\theta)]$ as follows (using eq.(27) and the definition of $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$):

$$\mathsf{Var}[\pi(\theta)] = \frac{\rho^2 \sigma^4 \nu^4 \left(m_2(\theta^*, \underline{\theta}, \alpha) - \bar{\tau}(\theta^*, \underline{\theta}, \alpha)^2\right)}{(\bar{\tau}(\theta^*, \underline{\theta}, \alpha)^2 + \rho^2 \sigma^2 \nu^2)^2}. \tag{61}$$

where $m_2(\theta^*, \underline{\theta}, \alpha) \equiv \mathsf{E}\left[\cot^4(\theta) | \, \underline{\theta} \leq \theta \leq \theta^*\right]$ is the second order moment of the variable $\tau(\theta)$ (the distribution of the quality of speculators' predictors). The first moment of this distribution is $\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})$. For a given search cost $c$, we must distinguish two cases. First, if the second moment of the distribution for the variable $\tau(\theta)$ diverges when $\underline{\theta}$ goes to zero (that is, $\lim_{\underline{\theta} \to 0} m_2(\theta^*, \underline{\theta}, \alpha) = +\infty$,), then we also have $\lim_{\underline{\theta} \to 0} \mathsf{Var}[\pi(\theta)] = +\infty$. Thus, $\mathsf{Var}[\pi(\theta)]$ is strictly decreasing with $\underline{\theta}$, for $\underline{\theta}$ close enough to 0.

If the second moment of the distribution for the variable $\tau(\theta)$ converges when $\underline{\theta}$ goes to zero, the analysis is more complex.[21] Indeed, as shown below, both the second and the first moments of the distribution for $\tau(\theta)$ decreases with $\underline{\theta}$. If the effect on the second moment dominates then $\mathsf{Var}[\pi(\theta)]$ decreases with $\underline{\theta}$ while if the effect on the first moment dominates then $\mathsf{Var}[\pi(\theta)]$ increases with $\underline{\theta}$ (see eq.(61)). We show below that for $\underline{\theta}$ sufficiently close to zero the first effect dominates.

We have:

$$\frac{d\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{d\underline{\theta}} = \frac{\partial \bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{\partial \underline{\theta}} + \frac{\partial \bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{\partial \theta^*} \frac{\partial \theta^*}{\partial \underline{\theta}} \tag{62}$$

[21] Notice first that $m_2(\theta^*, \underline{\theta}, \alpha) < \infty$ (the second moment converges) implies that $\phi(\theta)\cot^4(\theta)$ can be integrated in 0. Locally around $\theta = 0$, since $\cot(\theta) \sim \sin^{-1}(\theta) \sim \theta^{-1}$, we have $\phi(\theta)\cot^4(\theta) \sim \phi(\theta)\cot^2(\theta)\theta^{-2}$ As $\theta^{-2}$ 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.

Thus:

$$\frac{d\bar{\tau}(\theta^*,\underline{\theta},\hat{\alpha})}{d\underline{\theta}} = -\left(\phi^*(\underline{\theta})(\tau(\underline{\theta}) - \bar{\tau}(\theta^*,\underline{\theta},\hat{\alpha})) + \phi^*(\theta^*)\left(\bar{\tau}(\theta^*,\underline{\theta},\alpha) - \tau(\theta^*)\right)\frac{\partial\theta^*}{\partial\underline{\theta}}\right) \tag{63}$$

According to Proposition 7, we have $\frac{d\bar{\tau}(\theta^*,\underline{\theta},\hat{\alpha})}{d\underline{\theta}} < 0$, and according to Proposition 4, we have $\partial\theta^*/\partial\underline{\theta} < 0$ for $\underline{\theta} < \underline{\theta}^{tr}(c)$ small enough. Hence, using eq.(62), we deduce that for $\underline{\theta}$ close to 0 we have

$$0 < -\frac{\partial\theta^*}{\partial\underline{\theta}} < \phi(\underline{\theta}) \times \frac{\overbrace{\tau(\underline{\theta}) - \bar{\tau}(\theta^*,\underline{\theta},\hat{\alpha})}^{(*)}}{\phi(\theta^*)\left(\bar{\tau}(\theta^*,\underline{\theta},\alpha) - \tau(\theta^*)\right)}. \tag{64}$$

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

$$0 < -\frac{\partial\theta^*}{\partial\underline{\theta}} < K_1\phi(\underline{\theta})\tau(\underline{\theta}). \tag{65}$$

and therefore, inserting inequality (65) in equation (62), we obtain that there exists a constant $K_2$ such that

$$0 < -\frac{d\bar{\tau}(\theta^*,\underline{\theta},\hat{\alpha})}{d\underline{\theta}} < K_2\phi(\underline{\theta})\tau(\underline{\theta}). \tag{66}$$

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

$$\frac{dm_2(\theta^*,\underline{\theta},\alpha)}{d\underline{\theta}} = -\left(\phi^*(\underline{\theta})\left(\tau^2(\underline{\theta}) - m_2(\theta^*,\underline{\theta},\alpha)\right) + \phi^*(\theta^*)\left(m_2(\theta^*,\underline{\theta},\alpha) - \tau^2(\theta^*)\right)\frac{\partial\theta^*}{\partial\underline{\theta}}\right) \tag{67}$$

As the order of magnitude of $\partial\theta^*/\partial\underline{\theta}$ is (at best) $\phi(\underline{\theta})\tau(\underline{\theta})$, we deduce from the previous equation that:

$$\frac{dm_2(\theta^*,\underline{\theta},\alpha)}{d\underline{\theta}} \sim -\phi^*(\underline{\theta})\tau^2(\underline{\theta}), \tag{68}$$

when $\underline{\theta}$ is small. Hence, around $\underline{\theta} = 0$, $\frac{dm_2(\theta^*,\underline{\theta},\alpha)}{d\underline{\theta}}$ dominates $\frac{d\bar{\tau}(\theta^*,\underline{\theta},\hat{\alpha})}{d\underline{\theta}}$ by an order of magnitude. Indeed, the ratio between the second derivative and the first is bounded by $1/\tau(\underline{\theta})$, which goes to zero when $\underline{\theta}$ goes to zero.

**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.** In a symmetric interior equilibrium of the exploration phase, we have $\theta^* < \frac{\pi}{2}$, i.e., $F(\pi/2) < \exp(-\rho c)$ (see Proposition 2). Remember, from eq.(15), that any stopping rule $\hat{\theta}$ (e.g., $\theta^*$ or $\theta_r^{**}$) must satisfy:

$$J(\hat{\theta},\hat{\theta}) = \left[\frac{\Lambda(\hat{\theta};\underline{\theta},\alpha)}{H(\hat{\theta})}\right] \times \mathsf{E}\left[g(\theta,\hat{\theta})\middle|\underline{\theta} \leq \theta \leq \hat{\theta}\right] \tag{69}$$

where $H(\hat{\theta}) = \exp(-\rho c) - (1 - \Lambda(\hat{\theta};\underline{\theta},\alpha))$. Observe that $H(\hat{\theta})$ increases with $\hat{\theta}$. Moreover, under the condition $F(\pi/2) < \exp(-\rho c)$, we have: $H(\pi/2) > 0$. Let $\theta^{min}$ be the value of $\hat{\theta}$ such that $H(\theta^{min}) = 0$ (if any). The equilibrium stopping rule $\theta^*$ must be strictly greater than $\theta^{min}$. Indeed, if it was smaller then for $J(\theta^{*\text{‘}},\theta^*)$ would be strictly positive, which is impossible since $J(\theta^{*\text{‘}},\theta^*)$ is the ex-ante expected utility of a speculator and it cannot exceed zero (since speculators have CARA utility functions). Moreover, for $\hat{\theta} \to \theta_{min}^+$, $J(\hat{\theta},\hat{\theta}) \to -\infty$; A speculator can avoid such a low expected utility by choosing a stopping rule equal to $\hat{\theta} = \frac{\pi}{2}$. Thus, it must be that the equilibrium stopping rule, $\theta^*$, is strictly larger than $\theta^{min}$. Therefore, $H(\theta^*) > 0$.

Now consider $J(\theta_r,\theta_r)$. Eq.(69) and the fact that $\frac{\partial H}{\partial \theta_r} = \alpha\phi(\theta_r)$ implies that:

$$H(\theta_r)\frac{\partial J}{\partial \theta_r} + \alpha\phi(\theta_r)J(\theta_r) = \alpha\phi(\theta_r)g(\theta_r,\theta_r) + \alpha\int_{\underline{\theta}}^{\theta_r}\frac{\partial g}{\partial \theta_r}\phi(\theta)d\theta \tag{70}$$

$$\Leftrightarrow H(\theta_r)\frac{\partial J}{\partial \theta_r} = \alpha\phi(\theta_r)\frac{H(\theta_r)g(\theta_r,\theta_r) - \alpha\int_{\underline{\theta}}^{\theta_r} g(\theta,\theta_r)\phi(\theta)d\theta}{H(\theta_r)} + \alpha\int_{\underline{\theta}}^{\theta_r}\frac{\partial g}{\partial \theta_r}\phi(\theta)d\theta \tag{71}$$

$$\Leftrightarrow H(\theta_r)\frac{\partial J}{\partial \theta_r} = \frac{-\alpha\phi(\theta_r)g(\theta_r,\theta_r)}{H(\theta_r)}\underbrace{(F(\theta_r) - \exp(-\rho c))}_{>0 \text{ iff } \theta_r \leq \theta^*} + \alpha\underbrace{\int_{\underline{\theta}}^{\theta_r}\frac{\partial g}{\partial \theta_r}\phi(\theta)d\theta}_{>0} \tag{72}$$

Thus, at $\theta_r = \theta^*$, $\partial J/\partial\theta_r\ |_{\theta_r=\theta^*} > 0$ since the equilibrium stopping rule $\theta^*$ solves: $F(\theta^*) = exp(-\rho c)$ and $H(\theta^*) > 0$. This implies that by slightly increasing $\theta_r$ at $\theta_r = \theta^*$, one can make all speculators better off. Thus, the value of $\theta_r$ that maximizes $J(\theta_r,\theta_r)$ is strictly larger than $\theta^*$ ($\theta^* < \theta_r^{**}$. The expected investment in search of speculator $i$ for a given stopping rule $\theta_i^*$ is $\mathsf{E}(n_i)c = \frac{c}{\Lambda(\theta_i^*;\underline{\theta},\alpha)}$ (see eq.(3)). Now, $\Lambda(\theta_i^*;\underline{\theta},\alpha)$ increases with $\theta_i^*$. 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.