Equilibrium Data Mining and Data Abundance

Jérome Dugast, Thierry Foucault

Equilibrium Data Mining and Data Abundance*

Jérôme Dugast† Thierry Foucault‡

March 14, 2023

ABSTRACT

We study, using a noisy rational expectations framework, how the availability of new data to forecast asset payoffs ("data abundance") affect the capital allocated to quantitative asset managers ("data miners") relative to other active asset managers, the mean and the cross-sectional dispersion of their performance, and price informativeness. Data miners search for predictors of asset payoffs and trade when they find one with a sufficiently high precision. Data abundance raises the precision of the best predictors. Yet, it eventually induces data miners to lower the bar for their signal precision. Then, their performance becomes more dispersed, and they receive less capital. Overall, data abundance is both a catalyst and an impediment to the rise of quant funds.

Keywords: Big Data, Active Asset Management, Data Mining, Price Informativeness.

*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, Roxana Mihet, Sophie Moinas, Joël Peress, Francesco Sangiorgi, Daniel Schmidt, Andriy Shkilko, 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, EPFL/HEC Lausanne, Frankfurt School of Management, HEC Paris, INSEAD, Mc Gill, NHH, the University of Maryland, the University of New South Wales, and the University of Vienna for very useful comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101018214). All errors are the authors alone.

†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

"In the search for [...] alpha, fund managers are increasingly adopting quantitative strategies. [...] a new source of competitive advantage is emerging with the availability of alternative data sources as well as the application of new quantitative techniques of Machine Learning to analyze these data." (JPMorgan (2017), p.9)

I. Introduction

The "big data revolution" is having a major impact on the financial industry, particularly the active asset management industry. As the opening quote suggests, the proliferation of new datasets ("data abundance") and lower information processing costs (due to increased computing power) fuel the rise of quant funds.[1] Researchers in these funds mine vast amounts of data in search of new trading signals. This evolution raises many questions. Will quants crowd out more traditional ("discretionary") asset managers? How does it affect the performance of active asset managers? And will it increase price efficiency?

Our paper provides a benchmark model to study these questions. This model is unique in its ability to separate the effects of two facets of the big data revolution: (i) data abundance and (ii) lower information processing costs. These are often co-mingled, as if, both were contributing to reduce the cost of information production. Yet, they are conceptually distinct. Alternative data such as sensors or social media data can be useful to predict future cash-flows or returns.[2] Thus, it improves the precision of signals that quant funds can discover. However, it does not in itself reduce data processing costs. A new message of our theory is that the effects of data abundance and lower data processing costs are not necessarily the same.

Our model features risk averse investors who allocate their capital to asset managers. There are two types of asset managers: (i) "Experts" and (ii) "Data miners." Each can manage only a limited amount of capital and the total amount of available capital is equal to the maximum capacity of each category (so that allocating capital only to experts or only to data miners is possible). Experts and data miners invest in the same risky asset but differ in their technology to produce information about the asset. Each expert has a

[1] Abis (2022) finds that the number of quant funds in the U.S. grew from 6.71% to 20.65% of all active U.S. funds from 2000 to 2017. See also Harvey et al. (2017) and Begs et al. (2019) for similar evidence.

[2] Dessaint et al. (2022) identifies more than twenty academic papers showing that social media data (e.g., from twitter), sensors data (e.g., satellite images) or business activity data (e.g., credit card data) can be used to forecast stock returns or earnings.

pre-determined (e.g., via education), specific ability (skill) to obtain a signal of a given precision. In contrast, data miners discover their trading signals through a search process consisting of multiple exploration rounds. In a given round, each data miner can combine variables from different datasets (e.g., accounting data and social media data) to build a predictor of the asset payoff. The quality of this predictor, τ, is drawn from a distribution whose support ranges from zero (noise) to τ_{dm}^{max} (the "data frontier"). After a given round, each data miner can decide either to launch a new round of exploration at cost c (the search cost) or to stop searching. This process is repeated until each data miner finds a satisficing predictor. Then the market for the risky asset opens and trading takes place between asset managers, dealers, and noise traders (trading is formalized as in Vives (1995)).

We interpret experts as "discretionary" funds and data miners as "quant" funds (see Harvey et al. (2017) and Abis (2022)). The former relies on expertise (e.g., industry knowledge) and judgemental analysis to generate investment ideas while the latter systematically explores various datasets to discover and select predictors. As explained in Narang (2013) (Chapters 8 and 9), this process requires (i) obtaining, cleaning and preparing new data for analysis, (ii) using statistical techniques to obtain a predictor with this data and (iii) deciding, via backtesting, whether a predictor is good enough for live trading. One round of exploration comprises all these steps and the exploration cost is the total cost of executing them.

We characterize the equilibrium of the model, that is, (i) the allocation of capital to data miners and experts, (ii) data miners' optimal search strategy, (iii) asset managers' optimal holdings and (iv) the risky asset price such that (a) asset managers and investors make optimal decisions and (b) the asset market clears. In equilibrium, each data miner optimally stops searching once she finds a predictor whose quality exceeds an endogenous threshold, τ^*. Thus, in equilibrium, the quality of data miners' predictors is distributed over τ^* (least informative) to τ_{dm}^{max} (most informative). Hence, even though data miners are ex-ante homogeneous (same preferences and exploration costs), they are ex-post heterogeneous in the quality of their predictors (and therefore performance).

When investors allocate capital to asset managers, they observe experts' skills and correctly anticipate data miners' search strategy (τ^*). Investors prefer to allocate capital to experts with a skill above τ^* than to data miners because their expected utility increases

with experts' quality. Thus, investors optimally allocate capital to experts up to the point at which the marginal expert has a signal quality just equal to τ^*. The rest of investors' capital is then optimally allocated to data miners. This process determines the fraction of investors' capital, denoted μ^*, allocated to data miners.

Our analysis focuses on equilibrium effects of (i) an increase in the data frontier (τ_{dm}^{max}) and (ii) a decrease in data miners' search costs (c). We interpret the former as stemming from the availability of new datasets (data abundance) and the latter as reflecting greater computing power.[3] Conditional on a given capital allocation and for given distributions of signals' quality across data miners and experts (which are all fully determined by τ^*), the analysis of asset managers' trading decisions and the equilibrium of the market for the risky asset is standard. Thus, all the novel implications of the model derives from the effect of the data frontier or data miners' search costs on their search strategy (τ^*) and the resulting capital allocation (μ^*).

Consider a push back of the data frontier (i.e., an increase in τ_{dm}^{max}). It raises the expected trading profit of data miners who find the best predictor (the "hidden gold nugget" effect) since its quality becomes even larger. However, these data miners trade 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 expected profit of all data miners, especially those who do not find good predictors. The first effect raises the value of searching for a better predictor after finding one while the second reduces it. When τ_{dm}^{max} is small, the hidden gold nugget effect dominates. Thus, a push back of the data frontier induces data miners to be more demanding for their predictors (τ^* increases in equilibrium) and capital flows from experts to data miners. However, when τ_{dm}^{max} becomes large enough, a push back of the data frontier induces data miners to search less intensively (τ^* decreases) and capital flows back to experts because the aggressiveness effect dominates. In sum data abundance is a double-edged sword for quant funds: It facilitates their entry initially but, in

[3] One reason why greater computing power reduces quants' search costs is that it reduces the time required to perform one exploration round. For instance, Anthony Ledford, the chief scientist of MAN AHL (a quantitative fund), writes that *"Strategies based on NLP [...] are also live in client trading. Researching such strategies requires [...] a processor called graphical processing unit (GPU) that can complete the calculations [...] in* 1/30*th of the time taken by [...] a standard computer."* See AI Pioneers in Investment Management, CFA Institute, 2019.

the long run, it can eventually reduce capital available to these funds.

In contrast, a decrease in data miners' search cost, c, unambiguously raises the value of searching for another predictor after finding one because it reduces the total expected cost of search without directly affecting data miners' average trading aggressiveness. Thus, a reduction in data miners' search costs always leads them to be more demanding for the quality of their predictors and triggers an increase in capital allocated to these funds.

As shocks to data miners' search costs and the data frontier affect data miners' search intensity, they also affect the average quality of the signals used by both data miners *and* experts and, through this channel, price informativeness and asset managers' average performance (measured by their average trading profits, in a way similar to Berk and van Binsbergen (2015)). In equilibrium, a push back of the data frontier always triggers an improvement in the average quality of the signals used by all asset managers (even though, it can reduce data miners' search intensity) as does a reduction in data miners' search costs. Thus, a push back of the data frontier or a decrease in data miners' search costs raises the informativeness of the risky asset price.

For these reasons, asset managers' (cross-sectional) average performance is in general hump-shaped in data miners' search costs (c) and the data frontier (τ_{dm}^{max}). Indeed, a reduction in c or an increase in τ_{dm}^{max} raise both the average quality of asset managers' signals and asset price informativeness. The first effect has a positive impact on asset managers' average performance while the second has a negative effect (by reducing each asset manager's informational edge). When τ_{dm}^{max} becomes large enough or c small enough, this second effect dominates. Thus, in the long run, the big data revolution should erode active asset managers' performance (quants and discretionary funds alike).

The model also implies that shocks to the data frontier and data miners' search costs should affect the dispersion of performance within a given group of asset managers (experts or data miners). The average performance of an asset manager increases with the quality of her signal, other things equal. Thus, within a given group, the difference between the average performance of top and bottom performers increases in the difference between the quality of their signals. A reduction in data miners' search costs raises the lowest quality of the signals used by both data miners and experts, directly in the former case (through a more intense search process) and indirectly in the latter case (through a reallocation of capital away from the experts with the lowest skills). As a result, the

difference in performance between top and bottom performers within each group declines. In contrast, this difference is a U-shaped function of the data frontier. Indeed, when τ_{dm}^{max} is small, a push back of the data frontier raises data miners' search intensity while it reduces it when τ_{dm}^{max} becomes large enough.

We also analyze how changes in data miners' search costs and the data frontier affect management fees, assuming that these fees are set through Nash bargaining (as in Garleanu and Pedersen (2018)). In equilibrium, data miners charge a fee equal to their reservation wage. Intuitively, they cannot extract rents from investors because the latter always have the outside option to contact another (inactive) data miner and all data miners offer the same ex-ante expected utility to each investor (since data miners are ex-ante homogeneous). In contrast, the higher is an expert's skill, the higher is the expected utility of investors allocating capital to this expert. As experts with a given skill are in short supply, those with a skill above the marginal expert (the expert with skill τ^*) can charge a fee above their reservation wage and this fee increases with their skill. A decrease in data miners' search costs reduces experts' fees because it raises data miners' search intensity and therefore the expected utility that investors obtain by allocating their funds to them (investors' outside option when they negotiate fees with experts). This is also the case for an increase in the data frontier when it induces data miners to search more intensively. However, when it does not, a push back of the data frontier enables low quality experts to charge larger fees because allocating capital to data miners becomes less attractive.

To our knowledge, our predictions regarding data abundance are new and cannot be easily derived from the literature (see Section II). In contrast, the effects of a variation in data miners' search cost (c) are similar to those obtained by varying the cost of information acquisition in standard noisy rational expectations model (e.g., Verrecchia (1982)). However, we are not aware of such models in which, in equilibrium, investors with the same information acquisition technology acquire signals of different precision, as data miners do in our model. This feature of our model enables us to derive implications for the effects of shocks on c (or τ_{dm}^{max}) on the dispersion in asset managers' performance.[4] In Section VIII, we summarize the predictions of the model and discuss how they could be tested by (i) exploiting differences in industry or asset class coverage by alternative

[4]This is important because empirical findings suggest that asset managers have heterogeneous performance due to heterogeneity in the quality of their signals. See, for instance, Barras et al. (2022) and Kacperczyk and Seru (2007).

data providers and (ii) one time shocks to quant funds' search costs (e.g., due to cheaper computing power with the introduction of Amazon Web Services in 2006 or changes in data format facilitating the use of natural language processing techniques).

The paper is organized as follows. Section II positions our contribution in the literature. Section III presents the model and Section IV derives its equilibrium. Section V analyzes how changes in computing power and data abundance affects data miners' search strategy, the allocation of capital to data miners, and the informativeness of the asset price. Section VI derives implications for the effects of the big data revolution on asset managers' average performance and the dispersion of this performance across managers. Section VII endogenizes asset managers' fees while Section VIII summarizes the main testable implications of our theory. Section IX concludes.

II. Contribution to the Literature

Our paper is related to three strands of literature. First, it contributes to the theoretical literature studying trends in the asset management industry (such as the rise of institutional investors, institutional holdings' concentration or indexing) and its effects (e.g., Huang et al. (2019), Kacperczyk et al. (2022), Buss and Sundaresan (2022) or Bond and García (2021)). Here we focus on the rise of quant funds and relates it to (a) the availability of new data and (b) a decline in the cost of processing this data.

The theoretical literature on quantitative investors is scarce. Malikov and Pasquariello (2022) introduces a quantitative informed investor in Kyle (1985). This investor is myopic in the sense that her trading strategy is optimal provided that other informed investors behave as if she was not present. This assumption captures the idea that quantitative investors follow fixed trading rules calibrated on past data ("backtesting"). In contrast, we formalize quants as agents producing information through a search process and we focus on the effects of data abundance and information processing costs on their strategies (these parameters play no role in Malikov and Pasquariello (2022)). Our modeling approach is closer to Garleanu and Pedersen (2018). In their model, active asset managers can be either informed or uninformed and they study how investors allocate capital to both types. Investors can identify informed ones at a cost. Here, we assume that identifying informed asset managers is costless but we introduce heterogeneity in the

technologies used by active asset managers to generate their investment ideas.

Secondly, our paper contributes to the literature on endogenous information acquisition in noisy rational expectations models (see Veldkamp (2011) for a survey). These models assume either that (i) investors can obtain a signal of fixed exogenous precision by paying a cost (e.g., Grossman and Stiglitz (1980) or that (ii) investors choose the precision of their signal, with signals of higher precision being more costly (e.g., Verrecchia (1982)). Neither of these approaches can be used to study how, other things equal, an exogenous shock on the distribution of signals' precision affects agents' efforts to improve the precision of their signals (the intensive margin of information production). With the first approach, only the extensive margin of information production (the mass of investors acquiring information) is endogenous. With the second one, the intensive margin varies endogenously but only through shocks to the cost of information production. Yet, as explained in the introduction, the availability of new datasets enables investors to find signals of higher quality without changing data processing costs. Thus, to analyze the effects of data abundance on investors' efforts to discover signals of high quality, one needs a new modeling approach.

Our methodological contribution is to propose one. Our framework allows to analyze the effect of exogenous shocks on investors' signals precision (shocks to τ_{dm}^{max})–holding the cost of information production (c) constant–on both (i) the intensive margin of information production (characterized by τ^*, which determines the endogenous cross-sectional distribution of the precision of asset managers' signals in our model) and (ii) the extensive margin (the allocation of capital between two types of informed investors). In this way, we can untangle the effects of both dimensions of the big data revolution: (i) a reduction in the cost of processing data and (ii) an increase in the precision of the best signals that investors can obtain due to an expansion of available data. We find that these two shocks can affect efforts to produce information in opposite directions. To our knowledge, this economic insight is new and cannot be delivered in the existing literature (since it does not allow to analyze (ii) separately from (i)).

Han and Sangiorgi (2018) and Banerjee and Breon-Drish (2021) also formalize information acquisition as a search problem but they analyze different questions. Han and Sangiorgi (2018) provide a micro-foundation for the assumption (e.g., Verrecchia (1982)) that information acquisition costs are increasing and convex in precision with a model

in which an agent can repeatedly draw normally distributed signals from an "urn." Each draw is costly, similar to the cost of exploration in our model. In Banerjee and Breon-Drish (2021), one investor optimally alternates between periods in which she searches for information 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. In this model, investors face uncertainty on foregone trading opportunities (due to search delays) rather than signal quality as in our model.

Lastly, our paper contributes to the growing theoretical literature on new information technologies for the production of financial information (e.g., Dugast and Foucault (2018), Farboodi and Veldkamp (2020), Milhet (2020) or Huang et al. (2022)). 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 the improvement in the highest precision of the signals that investors can obtain due to the availability of new data (data abundance).

III. Model

Figure 1 describes the timing of actions in the model. In period 0, a continuum of investors (of mass 1) invest their savings, denoted W_0, in the stock market through active asset managers. As in Garleanu and Pedersen (2018), asset managers can hold a risk free asset (whose rate of return is normalized to zero) and a risky asset. The payoff of the risky asset, ω, is realized in period 3 and is normally distributed with mean zero and variance σ_ω^2.

A Experts and Data Miners

There are two types of asset managers: (i) "Experts" (a continuum of mass 1) or (ii) "Data Miners" (a continuum of mass 1). Each asset manager cannot serve more than a fixed number of clients that we normalize to one. We denote by μ $(1-\mu)$ the fraction of investors who allocate capital to data miners (experts). Henceforth, subscript "dm" refers to "data miners" and subscript "ex" to "expert".

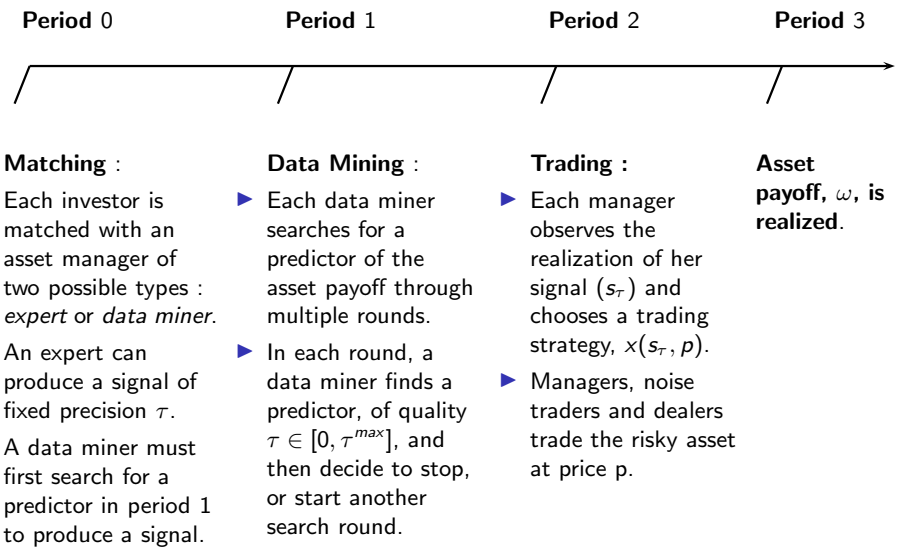

Figure 1: Timing

If a manager receives funding, she builds her portfolio at date 2. Just before doing so, she receives a signal s_{τ_i} about the payoff of the risky asset such that

$$s_{\tau_i} = \omega + \tau_i^{-1/2}\varepsilon_i. \tag{1}$$

The noise in asset managers' signals (ε_i) is normally distributed with mean zero and variance σ_ω^2 and is independent across managers. The higher is τ_i, the higher is the precision ("quality") of asset manager i's signal.

Experts and data miners obtain their signals in different ways. The quality of an expert's signal is fixed. We refer to it as her "skill" and assume it is pre-determined (e.g., by prior education or experience).[5] Experts' skills are distributed over $[0, \tau_{ex}^{max}]$ with a cumulative probability distribution $\Gamma(.)$ (density $\gamma(.)$). In contrast, the quality of data miners' signal is determined in period 1 through a search process. The goal of this process is to discover predictors of the asset payoff via multiple exploration rounds. Each round costs c and yields a predictor of quality τ drawn in $[0, \tau_{dm}^{max}]$ from the cumulative

[5]Chevalier and Ellison (1999) and Li et al. (2011) find that asset manager's education is the most robust variable explaining cross-sectional differences in the performance (alphas) of managers in their samples.

distribution $\Phi(.)$ (density $\phi(.)$), defined as:

$$\Phi(\tau) = \Pr(\tilde{\tau} \leq \tau) = \frac{\Psi(\tau)}{\Psi(\tau_{dm}^{max})}, \tag{2}$$

where $\Psi(.)$ (density $\psi(.)$) is a cumulative probability distribution defined on $[0, \infty]$.[6] After each exploration round, a data miner can decide (i) to stop searching or (ii) to start a new exploration. The quality of the signal she obtains in period 2 is the quality of her *latest* predictor.[7]. We refer to the highest predictors' quality, τ_{dm}^{max}, as the data frontier.

We focus on equilibria in which each data miner follows an optimal stopping rule τ_i^* (there is no other equilibrium with Markovian search strategies; see Section II.B in the online appendix). That is, data miner i stops searching once she finds a satisficing predictor whose quality τ exceeds τ_i^*. We denote by $\Lambda(\tau_i^*; \tau_{dm}^{max})$ the likelihood of this event in a given search round:

$$\Lambda(\tau_i^*; \tau_{dm}^{max}) \equiv \Pr(\tau \in [\tau_i^*, \tau_{dm}^{max}]) = 1 - \Phi(\tau_i^*). \tag{3}$$

An asset manager with a more demanding stopping rule (higher τ_i^*) conducts more exploration rounds on average.[8] For this reason, we refer to τ_i^* as the asset manager's "search intensity".

Interpretation. We interpret "data miners" as quant funds' managers and experts as funds' managers ("discretionary investors") who rely more on judgement and qualitative analysis to generate investment ideas. A central task in quant funds consists in searching for predictors (see Narang (2013), Chapters 8 and 9) . This task requires acquiring new datasets, clearing and preparing them for statistical analysis, writing algorithms to process the data, and finally backtesting trading strategies to assess the economic value of a predictor. Data miners' search cost c is the cost of this task. This information processing cost has declined over time because of digitization (which enables automation of unstructured data processing) and progress in computing power (which enables a research team to complete an exploration round faster).

[6]This specification is analytically convenient to parameterize the effect of τ_{dm}^{max} on the distribution of predictors' quality when we study the effects of τ_{dm}^{max} on equilibrium outcomes in Sections V and VI.

[7]Results are identical (but the presentation more involved) if this is the predictor of highest quality found over all exploration rounds; see Section II.A in the online appendix.

[8]The number of exploration rounds, n_i, performed by data miner i has a geometric distribution with parameter $\Lambda(\tau_i^*; \tau_{dm}^{max})$. Her expected number of explorations is therefore $\mathsf{E}[n_i] = \Lambda(\tau_i^*; \tau_{dm}^{max})^{-1}$, which increases with τ_i^*.

The data frontier, τ_{dm}^{max}, captures a different dimension of the big data revolution, namely the proliferation of new datasets (so called alternative data).[9] As more datasets become available ("data abundance"), data miners can try more diverse variables to predict asset payoffs. Combined with progress in forecasting techniques, this evolution enables quant funds to find ever more powerful predictors, that is to push back the data frontier (raise τ_{dm}^{max}).

In sum, we interpret an increase in τ_{dm}^{max} as reflecting the availability of new datasets and a reduction in c as being due to a decline in the cost of information processing due to greater computing power and digitization. Our analysis focuses on equilibrium effects of these shocks in our model. We explain how one could design tests of the implications of the model in Section VIII.

A predictor can be viewed as a function (e.g., fitted with linear regressions or machine learning techniques) of variables from different datasets (e.g., accounting data, geolocation data and consumer transactions data) that minimizes the predictor's average forecasting error in-sample. The data miner then uses the realization of these variables in period 2 (out-of-sample) to obtain her signal, s_τ, at this date.[10] We formalize this interpretation in Section II.C of the online appendix. Specifically, we assume that data miners obtain a predictor in each round by running a regression of ω on a *fixed* number, N, of variables.[11] The predictive power of one (or several) new variable used in place of one (or several) variable used in former exploration rounds is random. For this reason, as assumed in our model, the quality of a predictor obtained in a given round is random and can drop from one round to the next.[12]

[9] JPMorgan (2017) lists more than 500 different alternative data providers and this number has steadily increased over time (see Dessaint et al. (2022), Figure I)

[10] As usual in rational expectations models, we assume that there is no uncertainty on the quality of a predictor, τ, once it is discovered. In reality, investors are uncertain about this quality (e.g., because they have too few past observations to accurately estimate their predictive model) and learn it over time as they accumulate data (see Martin and Nagel (2022)). Analyzing this case is beyond the scope of our paper.

[11] It is realistic to assume that data miners restrict themselves to using a limited number of variables for building up their predictors. Indeed, quant funds often do so to avoid the risk of overfitting. For instance, Narang (2013) (a quant manager) writes (on p.163) that "*Among quants parsimony implies caution in arriving at a hypothesis. This concept is absolutely central to the research process in quant trading. Models that are parsimonious utilize as few assumptions and as much simplicity as possible [...] As such models with a large number of parameters or factors are generally to be viewed with skepticism, especially given the risks of overfitting.*"

[12] In this approach, s_τ is the predicted value of ω according to the regression and the R^2 of the regression ran in a given round increases with the quality of the predictor obtained in this round (see the online appendix). Thus, searching for predictors of high quality is the same thing as searching for predictors with high R^2s.

B Trading

The market for the risky asset opens in period 2 after *all* data miners find a predictor with satisficing quality. At the beginning of period 2, after observing the realization of her or his signal, s_τ, each asset manager with capital chooses a trading strategy, i.e., a demand schedule, $x_i(s_\tau, p)$, where, p, is the price of the risky asset.

As in Vives (1995), asset managers trade with noise traders and risk-neutral market makers. The noise traders' aggregate demand is price-inelastic and denoted by η, where $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$ (η is independent of ω and errors' in asset managers' signals). Market-makers observe the aggregate demand for the asset, $D(p) = \int x_i(s_\tau, p)di + \eta$ and post a price such that they obtain zero expected profits. Thus, the equilibrium price, p^*, is equal to their expectation of the asset payoff conditional on the aggregate demand for the asset:[13]

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

C Asset managers' objective function.

As in Garleanu and Pedersen (2018), investors have a CARA utility function (with risk aversion ρ) and asset managers invest in their clients' best interest. We assume that asset managers return to investors the value of their portfolio net of their operating costs (we relax the assumption that managers get zero surplus in Section VII). Thus, in period 3, asset manager i with type $j \in \{dm, ex\}$ returns to investors:

$$W_{i,j} = \underbrace{W_0 + x_i(s_{\tau_i}, p)(\omega - p)}_{\text{Portfolio Liquidation Value}} - (n_i c)\mathbb{1}_{\{j=dm\}}, \tag{5}$$

where (i) n_i is the realized number of exploration rounds for asset manager i if she is a data miner and (ii) $\mathbb{1}_{\{j=dm\}} = 1$ if $j = dm$ and zero otherwise.

Thus, the ex-ante (period 0) expected utility of an investor matched with an expert of type τ is:

$$H(\tau_i) = \underbrace{\mathsf{E}\left[-\exp(-\rho(W_0 + x_i(s_{\tau_i}, p)(\omega - p))\right]}_{\text{Expected utility from trading}}, \tag{6}$$

[13] In Section II.D of the online appendix, we show that results are unchanged when we model trading as in Grossman and Stiglitz (1980). In this case, there are no risk neutral dealers and the asset price is such that the net aggregate demand from asset managers and noise traders is zero (so that the market for the risky asset clears). In this case, the expected risk premium of the asset is different from zero. Our approach simplifies the presentation of the results.

and the ex-ante expected utility of an investor matched with data miner i is:

$$V(\tau_i^*) = \underbrace{\mathsf{E}\left[-\exp(-\rho(W_0 + x_i(s_{\tau_i}, p)(\omega - p))\right]}_{\text{Expected utility from trading}} \times \underbrace{\mathsf{E}\left[\exp(\rho(n_i c))\right]}_{\text{Expected utility cost of exploration}}, \quad (7)$$

The ex-ante expected utility of data miner i's client is determined by the data miner's search intensity, τ_i^*, because the latter determines both the distribution of n_i and the distribution of the data miner's optimal trading strategy ($x_i(s_\tau, p)$).

We assume that investors observe the type, τ, of each expert and correctly anticipate data miners' stopping rules in equilibrium. Thus, they allocate capital to data miners and experts by comparing $H(\tau_i)$ and $V(\tau_i^*)$. Last, we assume that investors have no expertise and so no ability to obtain an informative signal about the payoff of the risky asset. It is then never optimal for investors to invest in the risky asset directly.[14]

D Equilibrium Definition.

In equilibrium, experts and data miners face the same portfolio problem: They must choose an investment in the risky asset, x_i, that maximizes their client's expected utility conditional on the price of the asset and the realization of their signal. Thus, in equilibrium, an expert and a data miner who obtain signals of the same quality in period 2 follow the same trading strategy $x^*(s_\tau, p)$. Moreover, as all data miners are ex-ante identical, they will all choose the same search intensity, τ^* in equilibrium.[15] An equilibrium of the financial market is a set $\{\mu^*, \tau^*, x^*(s_\tau, p), p^*\}$ such that:

1. In period 0, the allocation of capital between data miners and experts, μ^*, is stable. That is, investors' capital is allocated in such a way that no investor can increase his expected utility by reallocating his capital to an asset manager without capital.[16]

2. For each data miner i, the search intensity τ^* and the trading strategy $x^*(s_{\tau_i}, p)$ maximize $V(\tau_i^*)$ (given in eq.(7)) when other data miners' search intensity is τ^* and

[14]This is clear given that the expected risk premium on the asset, conditional on public information, is zero. However, this is also the case when the asset provides a risk premium (e.g., when there are no dealers). The reason is that informed asset managers deliver a larger expected utility than what uninformed investors can obtain.

[15]We show in Section II.B of the online appendix that there are no equilibria in which data miners choose different search intensities.

[16]There are always asset managers without capital in equilibrium because there is "excess supply" of asset managers: The total mass of experts and data miners combined exceeds the mass of investors (remember that each asset manager accommodates at most one investor).

other asset managers' trading strategy is $x^*(s_\tau, p)$.

3. The trading strategy $x^*(s_{\tau_i}, p)$ of expert i maximizes $H(\tau_i)$ (given in eq.(6)) when data miners' search intensity is τ^* and other asset managers' trading strategy is $x^*(s_\tau, p)$.

4. The asset price, p^*, satisfies $p^* = \mathsf{E}[\omega | D(p^*)]$ where $D(p^*) = \int x^*(s_{\tau_i}, p^*)di + \eta$.

In the next section, we solve for the equilibrium of the market for the risky asset. Intuitively, in equilibrium, more skilled experts deliver a higher expected utility to their clients. That is, $H(\tau)$ increases with τ (we show that this conjecture is correct below). Thus, if an expert with a given skill, say τ', receives capital then all experts with skills $\tau \geq \tau'$ must receive capital as well. If not then the allocation of capital to asset managers cannot be stable because the investor allocating capital to the expert with skill τ' is better off reallocating his capital to an expert with skill $\tau \geq \tau'$. Thus, the proportion of capital allocated to experts is $(1-\Gamma(\underline{\tau}))$, where $\underline{\tau}$ is the skill of the "marginal expert" (the expert with the lowest skill among those receiving capital). Hence, in a stable equilibrium, it must be the case that:[17]

$$(1-\mu) = (1-\Gamma(\underline{\tau})) \Longleftrightarrow \mu = \Gamma(\underline{\tau}). \tag{8}$$

Using this observation, in the next section, we solve the equilibrium in the following way. We conjecture that, in equilibrium, the quality of the marginal expert, $\underline{\tau}$, is identical to the quality of the data miner with the worst predictor, τ^*, so that $\mu^* = \Gamma(\tau^*)$ (eq.(8)).[18] In a first step (Proposition 1), we solve for the equilibrium of the trading stage in period 2 taking τ^* (and therefore μ^*) as given. From this step, we then derive data miners' optimal search intensity, τ^*, in equilibrium (Proposition 2). Finally (Proposition 3), we show that in equilibrium it must be the case that $\underline{\tau} = \tau^*$ and therefore $\mu^* = \Gamma(\tau^*)$, as conjectured .

[17] Condition (8) holds even if $\mu = 0$ or $\mu = 1$ because the mass of investors is equal to 1. Thus, one can always match all investors with all experts (the case $\mu = 0$) or all investors with all data miners (the case $\mu = 1$). The first case never arises in equilibrium for c low enough (see Proposition 2) while the second never happens for τ_{dm}^{max} finite.

[18] Alternatively, we could solve for the equilibrium search intensity and trading strategy for an arbitrary value of $\underline{\tau}$ (and therefore μ) and then conclude that in a stable equilibrium it must be the case that $\underline{\tau} = \tau^*$. However, this approach lengthens the presentation without adding any insight. Hence we prefer to directly conjecture that $\underline{\tau} = \tau^*$ and then check that the conjecture is correct.

IV. Equilibrium

A Equilibrium of the market for the risky asset

As explained in the previous section, we first solve for asset managers' equilibrium trading strategy and the equilibrium price of the asset in period 2 by taking τ^* as given, under the conjecture that $\mu^* = \Gamma(\tau^*)$. Let $\phi^*(\tau)$ be the distribution of quality for data miners' signals given that they follow the stopping rule τ^*:

$$\phi^*(\tau) = \frac{\phi(\tau)}{\Lambda(\tau^*; \tau_{dm}^{max})} = \frac{\psi(\tau)}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)}, \ \forall \tau \in [\tau^*, \tau_{dm}^{max}]. \tag{9}$$

We denote the *average* quality of signals across (i) all data miners by $\bar{\tau}_{dm}(\tau^*) = \mathsf{E}_\phi\left[\tau \,|\tau^* \leq \tau \leq \tau_{dm}^{max}\right]$ and (ii) all experts by $\bar{\tau}_{ex}(\tau^*) = \mathsf{E}_\gamma\left[\tau \,|\tau^* \leq \tau \leq \tau_{ex}^{max}\right]$. The average quality of signals across all asset managers, denoted $\bar{\tau}(\tau^*; \tau_{dm}^{max})$, is therefore:

$$\bar{\tau}(\tau^*; \tau_{dm}^{max}) = \mu^* \bar{\tau}_{dm}(\tau^*) + (1 - \mu^*)\, \bar{\tau}_{ex}(\tau^*). \tag{10}$$

We obtain the following result.

Proposition 1. *In equilibrium, an asset manager's demand for the risky asset is*

$$x^*(s_\tau, p) = \frac{\mathsf{E}[\omega | s_\tau, p] - p}{\rho\, \mathsf{Var}[\omega | s_\tau, p]} = \beta(\tau)\,(s_\tau - p)\,, \tag{11}$$

where $\beta(\tau) = \frac{\tau}{\rho\sigma_\omega^2}$ *and the equilibrium price of the asset is*

$$p^* = \mathsf{E}[\omega | D(p)] = \lambda(\tau^*)\xi. \tag{12}$$

where $\xi \equiv \omega + \rho\sigma_\omega^2 \bar{\tau}(\tau^*; \tau_{dm}^{max})^{-1}\eta$ *and*

$$\lambda(\tau^*) \equiv \frac{\bar{\tau}(\tau^*; \tau_{dm}^{max})^2}{\bar{\tau}(\tau^*; \tau_{dm}^{max})^2 + \rho^2\sigma_\omega^2\sigma_\eta^2}, \tag{13}$$

An asset manager's optimal holding of the risky asset is proportional to the difference between her signal (s_τ) and the price of the asset (p). Holding this difference constant, asset managers with signals of higher quality (larger τ) take larger positions ($\beta(\tau)$ is larger) because, conditional on their information, their residual uncertainty about the payoff of the asset is smaller. Thus, $\beta(\tau)$ measures the aggressiveness with which an

asset manager exploits her signal.

As in Grossman and Stiglitz (1980), we measure price informativeness by the inverse of the residual uncertainty about the asset payoff conditional on its equilibrium price, which we denote by $\mathcal{I}(\tau^*; \tau_{dm}^{max}) = \mathsf{Var}[\omega \mid p^*]^{-1}$. Using Proposition 1, we obtain that

$$\mathcal{I}(\tau^*; \tau_{dm}^{max}) = \frac{1}{\sigma_\omega^2} + \frac{\bar{\tau}(\tau^*; \tau_{dm}^{max})^2}{\rho^2 \sigma_\omega^4 \sigma_\eta^2}. \tag{14}$$

Thus, in line with intuition, the asset price is more informative when the average quality of asset managers' predictors, $\bar{\tau}(\tau^*; \tau_{dm}^{max})$, increases. The next result is useful for the rest of the analysis.

Lemma 1. *Other things equal, the average quality of asset managers' signals, $\bar{\tau}(\tau^*; \tau_{dm}^{max})$, and therefore price informativeness, $\mathcal{I}(\tau^*; \tau_{dm}^{max})$, increase with data miners' search intensity τ^*.*

Using Proposition 1, we can derive an asset manager's ex-ante expected utility from trading (as defined in eq.(7) or eq.(6)), i.e., before observing the realization of her predictor and the equilibrium price, when her predictor has type τ. We denote this ex-ante expected utility from trading by $g(\tau, \tau^*)$ and refer to it as the trading value of a signal with type τ. Formally:

$$g(\tau, \tau^*) \equiv \mathsf{E}\left[-\exp(-\rho(W_0 + x^*(s_\tau, p^*)(\omega - p^*)) \mid \tau_i = \tau\right]. \tag{15}$$

Lemma 2. *In equilibrium, the trading value of a signal with type τ is*

$$g(\tau, \tau^*) = -\left(1 + \frac{\mathsf{Var}[\mathsf{E}[\omega|s_\tau, p^*] - p^*]}{\mathsf{Var}[\omega|s_\tau, p^*]}\right)^{-\frac{1}{2}} = -\left(1 + \frac{\tau}{\sigma_\omega^2 \mathcal{I}(\tau^*; \tau_{dm}^{max})}\right)^{-\frac{1}{2}}. \tag{16}$$

The trading value of a signal increases with its quality and decreases with the informativeness of the asset price. Thus, it is inversely related to the average quality of asset managers' signals. Hence, holding quality constant, the value of a signal for a data miner (or expert) is smaller if other data miners are more demanding for the quality of their predictors (i.e., τ^* is larger).

B Equilibrium data mining.

Armed with Lemma 2, we can now derive a data miner's optimal search intensity given that other data miners' search intensity is τ^*. Let $\widehat{\tau}_i$ be an arbitrary search intensity for data miner i. The data miner's continuation value after turning down a predictor is

$$J(\widehat{\tau}_i, \tau^*) = \exp(\rho c) \left(\Lambda(\widehat{\tau}_i; \tau_{dm}^{max}) \, \mathsf{E}_\phi \left[g(\tau, \tau^*) \, | \widehat{\tau}_i \leq \tau \leq \tau_{dm}^{max}\right] + (1 - \Lambda(\widehat{\tau}_i; \tau_{dm}^{max})) J(\widehat{\tau}_i, \tau^*)\right), \tag{17}$$

where $\Lambda(\widehat{\tau}_i; \tau_{dm}^{max})$ is the likelihood of finding a satisficing predictor in the next exploration round. The first term $(\exp(\rho c))$ in eq.(17) is the expected utility cost of running an additional exploration round. The second term is the likelihood that the next exploration round is successful $(\Lambda(\widehat{\tau}_i; \tau_{dm}^{max}))$ times the average trading value of a predictor conditional on the quality of this predictor being satisficing (i.e., larger than $\widehat{\tau}_i$). Finally, the third term is the likelihood that the next exploration is unsuccessful times the data miner's continuation value when she turns down a predictor. Solving eq.(17) for $J(\widehat{\tau}_i, \tau^*)$, we obtain

$$J(\widehat{\tau}_i, \tau^*) = \underbrace{\left[\frac{\exp(\rho c)\Lambda(\widehat{\tau}_i; \tau_{dm}^{max})}{1 - \exp(\rho c)(1 - \Lambda(\widehat{\tau}_i; \tau_{dm}^{max}))}\right]}_{\text{Expected Utility Cost from Exploration}} \times \underbrace{\mathsf{E}_\phi \left[g(\tau, \tau^*) \, | \widehat{\tau}_i \leq \tau \leq \tau_{dm}^{max}\right]}_{\text{Expected Utility from Trading}} \tag{18}$$

Now suppose that data miner i just obtained a predictor with quality τ. If she stops searching, her expected utility is $g(\tau, \tau^*)$. If instead she launches a new round of exploration, her expected utility is $J(\widehat{\tau}_i, \tau^*)$. Thus, the data miner optimally stops searching if $g(\tau, \tau^*) \geq J(\widehat{\tau}_i, \tau^*)$ and keeps searching otherwise. As $g(\tau, \tau^*)$ increases with τ, the data miner's optimal stopping rule, $\tau_i^*(\tau^*)$, is the value of τ such that she is just indifferent between these two options:

$$g(\tau_i^*(\tau^*), \tau^*) = J(\tau_i^*(\tau^*), \tau^*). \tag{19}$$

In a symmetric equilibrium, $\tau_i^*(\tau^*) = \tau^*$. We deduce that τ^* solves

$$g(\tau^*, \tau^*) = J(\tau^*, \tau^*). \tag{20}$$

Using the expression for $J(.,\tau^*)$ in eq.(17), we can equivalently rewrite this equilibrium condition as

$$F(\tau^*) = \exp(-\rho c), \tag{21}$$

where

$$F(\tau^*) \equiv \int_{\tau^*}^{\tau_{dm}^{max}} r(\tau,\tau^*)\phi(\tau)d\tau + (1-\Lambda(\tau^*;\tau_{dm}^{max}))\,, \quad \text{for } \tau^* \in [0,\tau_{dm}^{max}]\,, \tag{22}$$

with

$$r(\tau,\tau^*) \equiv \frac{g(\tau,\tau^*)}{g(\tau^*,\tau^*)} = \left(\frac{\tau^* + \sigma_\omega^2 \mathcal{I}(\tau^*;\tau_{dm}^{max})}{\tau + \sigma_\omega^2 \mathcal{I}(\tau^*;\tau_{dm}^{max})}\right)^{\frac{1}{2}}. \tag{23}$$

We deduce the following result.

Proposition 2. *In equilibrium, data miners' search intensity (τ^*) is the unique solution to eq.(21). It is always strictly smaller than τ_{dm}^{max} and it is strictly larger than zero if and only if $F(0) < \exp(-\rho c)$.*

Intuitively, when $\exp(-\rho c) \leq F(0)$ (i.e., when c is large), the expected utility cost of exploration is larger than expected utility of trading for a data miner. Hence, in equilibrium, data miners do not search ($\tau^* = 0$).

In equilibrium, an investor who allocates capital to an expert with skill τ obtains an expected utility equal to $H(\tau) = g(\tau,\tau^*)$. Thus, as conjectured, $H(\tau)$ increases with τ (Lemma 2). In contrast, an investor who allocates capital to a data miner obtains an expected utility equal to $V(\tau^*) = J(\tau^*,\tau^*) = g(\tau^*,\tau^*)$ (eq.(20)). Thus, $V(\tau^*) = H(\tau^*)$. Investors are therefore indifferent between allocating capital to an expert with skill τ^* or a data miner and strictly prefer to allocate capital to experts with skill $\tau > \tau^*$. It follows that, in equilibrium, the skill of the marginal expert (the expert with the lowest skill) is equal to the quality of the worst predictor used by data miners in equilibrium, τ^* and therefore $\mu^* = \Gamma(\tau^*)$ (eq.(8)). We deduce that the equilibrium of the financial market is as described in the next proposition.

Proposition 3. *In equilibrium, τ^* solves: $F(\tau^*) = \exp(-\rho c)$ where $F(.)$ is given by eq.(21). Moreover, x^* and p^* are given as in Proposition 1 and the allocation of capital between data miners and experts is such that $\mu^* = \Gamma(\tau^*)$, with $0 \leq \mu^* \leq 1$. Therefore the skill of the marginal expert is $\underline{\tau} = \tau^* > 0$. Moreover, $\mu^* > 0$ (data miners receive capital)*

if and only if $F(0) < \exp(-\rho c)$*, and* $\mu^* < 1$ *(experts receive capital) if and only if* $F(\tau_{ex}^{max}) > \exp(-\rho c)$.

Proposition 3 provides a full characterization of the equilibrium of the financial market. In the rest of the paper, we focus on thecase in which $F(0) < \exp(-\rho c)$ and $F(\tau_{ex}^{max}) > \exp(-\rho c)$ so that both data miners and experts receive capital (i.e., c small enough and τ_{ex}^{max} large enough). We analyze the effects of data miners' search costs (c) and the data frontier (τ_{dm}^{max}) on the capital allocation between experts and data miners (Section V) and the distribution of performance across asset managers (Section VI). In these analyses, we are particularly interested in the effects that arise when the cost of search (c) becomes very small or data becomes very abundant (τ_{dm}^{max} very large) as these limiting cases are often described as the endpoint of current progress in information technologies (Nordhaus (2021)).

V. Capital Allocation: Data miners vs Experts

In this section we study how a reduction in data miners' search costs (c) or push back of the data frontier (an increase in τ_{dm}^{max}) affect data miners' search intensity (τ^*) and therefore the allocation of capital to data miners relative to experts (μ^*). We also analyze how these effects determine the average quality of asset managers' signals and price informativeness.

Proposition 4. *A decrease in data miners' search costs, c, always increases data miners' search intensity* τ^* *in equilibrium (*$\partial\tau^*/\partial c < 0$*) and* τ^* *goes to* τ_{dm}^{max} *when c goes to zero. For this reason, a decrease in data miners' search costs raises the allocation capital to data miners (*μ^**). Moreover, it increases (i) the average quality of data miners and experts' signals (*$\bar{\tau}_{dm}$ *and* $\bar{\tau}_{ex}$*), (ii) the average quality of all asset managers' signals (*$\bar{\tau}$*) and therefore (iii) price informativeness (*$\mathcal{I}$*).*

Holding τ^* constant, a decrease in data miners' search costs directly reduces the expected utility cost of launching a new exploration after finding a predictor (the first term in bracket in eq.(18)) for data miners. Hence, it raises the value of searching for another predictor after finding one, other things equal. Data miners become therefore more demanding for the quality of their predictor in equilibrium and τ^* increases. As a result, data miners deliver a higher ex-ante expected utility to investors ($V(\tau^*) = g(\tau^*, \tau^*)$

increases). Thus, investors allocate more capital to data miners (μ^* increases) and the quality of the marginal expert increases.

These effects raise the average quality of the signals used by both data miners and experts and therefore price informativeness increases (see 14). The latter effect reduces the trading value of their signals for all asset managers (see eq.(16)). Hence, it dampens the direct positive effect of a decrease in data miners' search costs, c, on the value of searching for a better predictor for data miners. However, in equilibrium, this indirect effect never fully offsets the positive direct effect of a decrease in c on the value of searching for another predictor after obtaining one.[19]

In the next proposition, we describe the effect of the data frontier, τ_{dm}^{max} on the equilibrium. For the proof of this proposition, we need the following (mild) technical assumption:

A.1. The density $\psi(.)$ (defined in eq.(2)), is such that for all $\tau^* > 0$, $\lim_{\tau_{dm}^{max} \to \infty} \bar{\tau}(\tau^*; \tau_{dm}^{max})$ exists and is finite.

Proposition 5. *Under Assumption A.1, there exists a threshold $\tau^{tr}(c)$ such that when $\tau_{dm}^{max} \geq \tau^{tr}(c)$ then a push back of the data frontier reduces (i) data miners' search intensity, τ^* ($\partial\tau^*/\partial\tau_{dm}^{max} < 0$) and (ii) the allocation of capital to data miners ($\partial\mu^*/\partial\tau_{dm}^{max} < 0$). In this case, data abundance reduces the average quality of experts' signals ($\bar{\tau}_{ex}$) while it increases the average quality of data miners' signals ($\bar{\tau}_{dm}$). Nevertheless, for all values of τ_{dm}^{max}, data abundance raises the average quality of asset managers' signals ($\bar{\tau}$), and therefore price informativeness ($\mathcal{I}$).*

Thus, a decrease in data miners' search costs (e.g., due to cheaper computing power) and a push back of the data frontier (e.g., due to the availability of new datasets) do not have the same effects (compare Propositions 4 and 5). While the former always leads to a rise in the allocation of capital to data miners, the latter can have the opposite effect when τ_{dm}^{max} becomes large enough. The reason is that, in contrast to a reduction in search costs, a push back of the data frontier can induce data miners to become *less* demanding for the quality of their predictors.

The reason is the following. Other things equal, a marginal increase in τ_{dm}^{max} directly increases the average quality of data miners' predictors and therefore their average trading

[19]Suppose instead that it does (to be contradicted) and that, as a result, for some values of c, a decline in c reduces τ^*. Then, for these values, asset managers' average aggressiveness and therefore price informativeness would fall when c declines. But then the value of searching for a new predictor would increase. A contradiction.

aggressiveness. As a result, price informativeness increases. This direct effect reduces the average value of trading for data miners ($\mathsf{E}\left[g(\tau,\tau^*)\middle|\,\tau^*\leq\tau\leq\tau_{dm}^{max}\right]$) and therefore the value of searching for a predictor (the "aggressiveness effect"). The improvement in the best predictor acts as a countervailing force because it raises the value of trading on the best predictor (the "hidden gold nugget effect"). However, for τ_{dm}^{max} high enough, the aggressiveness effect always dominates the hidden gold nugget effect. Hence, data miners become less demanding for their predictors in equilibrium.

One can grasp this intuition more formally by differentiating asset managers' ex-ante expected utility from trading with respect to τ_{dm}^{max}, holding τ^* constant:

$$\frac{\partial\,\mathsf{E}_\phi\left[g(\tau,\tau^*)\middle|\,\tau^*\leq\tau\leq\tau_{dm}^{max}\right]}{\partial\tau_{dm}^{max}}=$$

$$\phi^*(\tau_{dm}^{max})\left[\underbrace{g(\tau_{dm}^{max},\tau^*)-\mathsf{E}_\phi\left[g(\tau,\tau^*)\middle|\,\tau^*\leq\tau\leq\tau_{dm}^{max}\right]}_{>0:\text{ Hidden Gold Nugget Effect}}+\underbrace{\mathsf{E}_\phi\left[\frac{\partial g(\tau,\tau^*)}{\partial\tau_{dm}^{max}}\middle|\,\tau^*\leq\tau\leq\tau_{dm}^{max}\right]}_{<0:\text{ Aggressiveness Effect}}\right]. \tag{24}$$

The first term in bracket in the above equation is the difference between the value of trading on the best predictor and the ex-ante value of trading for data miners. It measures the increase in data miners' ex-ante expected value trading following a marginal increase in τ_{dm}^{max}, due to the increase in the value of trading on the best predictor (the "gold nugget"). The second term in bracket is the loss in data miners' ex-ante value of trading due to the increase in their average trading aggressiveness following an improvement in the quality of the best predictor (the "aggressiveness effect").

When τ_{dm}^{max} becomes large, the residual risk faced by data miners who obtain the best predictor vanishes (they are less and less uncertain about the asset payoff). As a result, their trading aggressiveness become very large and the asset price becomes increasingly closer to the asset payoff (more informative). Thus, data miners' expected trading profit vanishes. For this reason, when τ_{dm}^{max} becomes high enough, the aggressiveness effect dominates the hidden gold nugget effect so that a push back of the data frontier (an increase in τ_{dm}^{max}) reduces data miners' ex-ante expected utility. Consequently, the value of searching for a predictor falls and therefore data miners become less demanding for their predictor (τ^* decreases).[20]

[20]Pushing back the data frontier, τ_{dm}^{max}, has a third effect: It increases the chance of finding a satisficing

As data miners' become less demanding for the quality of their predictors, they deliver a smaller expected utility to investors. Thus, a push back of the data frontier triggers a reallocation of investors' capital from data miners to experts (μ^* drops) until the point at which the skill of the marginal expert is identical to the quality of the worst signal for data miners (τ^*). Thus, data abundance indirectly reduces experts' average skill ($\bar{\tau}_{ex}$) via a reallocation of capital from data miners to experts.

When τ_{dm}^{max} is small enough, the effects of an increase in τ_{dm}^{max} on τ^* and μ^* are reversed (the gold nugget effect dominates the aggressiveness effect). Hence, in equilibrium, data miners' search intensity, τ^*, and the fraction of data miners, μ^*, are therefore a hump-shaped function of the data frontier, τ_{dm}^{max}.

Figure 2 illustrate our findings regarding the allocation of capital to data miners and their search intensity for a particular specification of predictors' quality, namely $\Phi(\tau) = \frac{1-(1+\tau)^{-3/2}}{1-(1+\tau_{dm}^{max})^{-3/2}}$.[21] A push back of the data frontier (lower panel in Figure 2) initially triggers a rise in the allocation of capital to data miners. However, at some point, this trend reverses and capital flows back to experts. This suggests that, even though data abundance can be a catalyst for the rise of quant funds (interestingly, the emergence of quant funds coincide with the emergence of alternative data providers at the end of the 90s), it can eventually become a limiting factor.

predictor holding the search strategy, τ^* constant ($\Lambda(\tau^*;\tau_{dm}^{max})$ increases when τ_{dm}^{max} goes up). This effect reduces the expected number of rounds required to find a predictor and therefore the expected utility cost of searching for a new predictor after rejecting one. Thus, like the hidden gold nugget effect, it works to increase data miners' value of searching for another predictor after finding one. However, the combined forces of this effect and the hidden gold nugget effect, are not sufficient to offset the negative impact of the aggressiveness effect on data miners' value of searching for τ_{dm}^{max} large.

[21] This means that $\Psi(\tau) = 1-(1+\tau)^{-3/2}$. This distribution belongs to a more general family for which we can compute $F(.)$ in closed-form and therefore solve for the equilibrium of the model numerically (see Section III in the online appendix for more details). For this family of distributions, $1+\tau$ has a power distribution and Assumption A.1 is satisfied (see the online appendix).

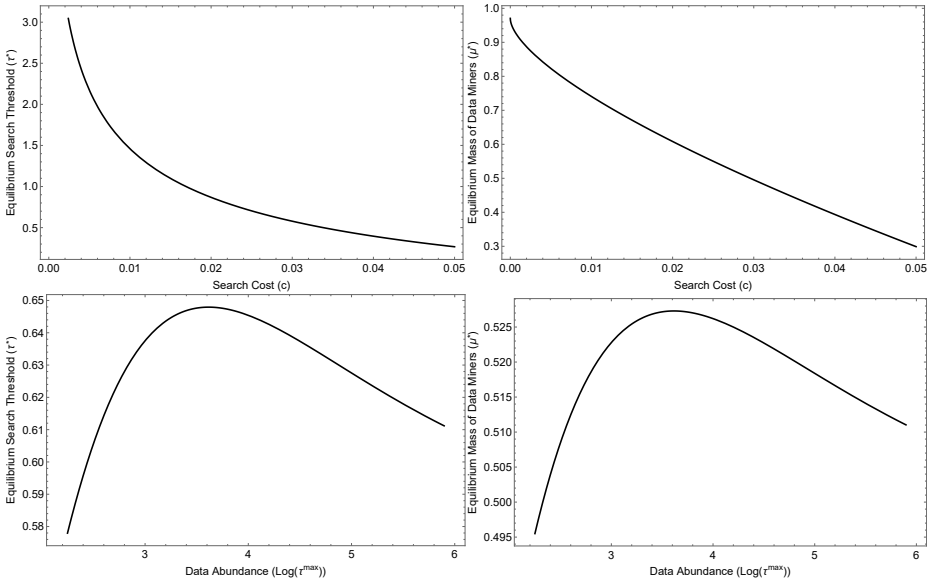

Figure 2: Upper Panels: Equilibrium search threshold, τ^* (left-hand side), and mass of data miners, μ^* (right hand-side), as functions of the search cost, c (other parameter values are $\tau_{dm}^{max} = \cot^2(\pi/10) \simeq 9.5, \rho = \sigma_\omega = \sigma_\eta = 1$). **Lower Panels:** Equilibrium search threshold, τ^* (left-hand side), and mass of data miners, μ^* (right hand-side), as functions of data abundance, $\log(\tau_{dm}^{max})$ (other parameter values are $c = 0.03, \rho = \sigma_\omega = \sigma_\eta = 1$). Underlying distributions of signals' quality are taken as $\Phi(\tau) = \frac{1-(1+\tau)^{-3/2}}{1-(1+\tau_{dm}^{max})^{-3/2}}$ and $\Gamma(\tau) = 1-(1+\tau)^{-3/2}$.

Even though a push back of the data frontier can reduce data miners' search intensity (τ^*), Proposition 5 also implies that data abundance always improves the average quality of data miners' signals, $\bar{\tau}_{dm}$ (see Proposition 5). The reason is that, in equilibrium, the drop in data miners' search intensity is always more than offset by the improvement in the quality of data miners' best signals (τ_{dm}^{max}). Moreover, this improvement is always strong enough to offset the drop in the average quality of experts' signals so that ultimately the net effect of data abundance on the average quality of all asset managers' signals ($\bar{\tau}$) is positive. For this reason, data abundance always improves price informativeness.

For completeness, the next proposition considers the effects of a shock on the volume of noise trading and the volatility of the asset payoff on the allocation of capital to data miners, their search intensity and other variables of interest.

Proposition 6. *In equilibrium, an increase in the volume of noise trading, σ_η^2, or the volatility of the asset payoff, σ_ω^2 trigger (i) a decrease in price informativeness ($\mathcal{I}$), (ii) an increase in data miners' search intensity, τ^*, (iii) an increase in the allocation of capital, μ^*, to data miners, (iv) an improvement in the average quality of data miners'*

signals ($\bar{\tau}_{dm}$) and the average quality of experts' signals ($\bar{\tau}_{ex}$) and (v) an improvement in the average quality of all asset managers' signals ($\bar{\tau}$).

Other things equal (in particular τ^*), 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. Thus, the value of searching for a predictor for data miners increases and therefore they become more demanding for their predictors (τ^* increases). As a result, data miners attract relatively more capital (μ^* increases). This raises the quality of the marginal expert and therefore, ultimately, the average quality of all asset managers. One implication is that quant funds should attract more capital in periods of high volatility and that the dispersion in their performance should become smaller during these periods (an increase in τ^* reduces the dispersion in the performance of data miners; see Section VI.B).

VI. Progress in information technology and performance in active management

We now study how the data frontier and data miners' search costs affect (i) the average performance of all asset managers, (ii) the dispersion in performance across asset managers of a given type, and (iii) the performance of data miners relative to experts. This is of broad interest given the substantial literature on active asset managers' performance (see, for instance, Berk and van Binsbergen (2015), Zhu (2018), Gerakos et al. (2021), or Barras et al. (2022) for recent empirical studies).

One measure of an asset manager's performance is her total dollar return on investment, adjusted for risk (Berk and van Binsbergen (2015) or Stambaugh (2020)). In our model, this corresponds an asset manager's equilibrium trading profit, denoted $\Pi(s_\tau)$:[22]

$$\Pi(s_\tau) = x^*(s_\tau, p^*) \times (\omega - p^*), \tag{25}$$

where $x^*(s_\tau, p^*)$ and p^* are given by eq.(11) and eq.(12), respectively. From eq.(11), we

[22]We do not account for data mining costs in computing the performance of a data miner. That is, we focus on the gross performance of data miners (gross return times investment) before accounting for management costs, as the empirical literature on funds' performance often does.

obtain

$$x^*(s_\tau, p^*) = \frac{1}{\rho\sigma_\omega^2}\left(\tau(\omega - p^*) + \tau^{1/2}\varepsilon_i\right). \tag{26}$$

Thus, the *expected* trading profit of an asset manager with predictor's quality τ is

$$\bar{\Pi}(\tau) \equiv \mathsf{E}[\Pi(s_\tau)|\tau] = \frac{\tau}{\rho\sigma_\omega^2}\,\mathsf{Var}[\omega - p^*] = \frac{\tau}{\rho\sigma_\omega^2\mathcal{I}(\tau^*;\tau_{dm}^{max})}, \tag{27}$$

where the last equality follows from the fact that $p^* = \mathsf{E}(\omega \mid p^*)$ so that $\mathsf{Var}[\omega - p^*] = \mathsf{Var}[\omega \mid p^*] = (\mathcal{I}(\tau^*;\tau_{dm}^{max}))^{-1}$ (by definition of $\mathcal{I}(\tau^*;\tau_{dm}^{max})$). Hence, the average performance of an asset manager increases with the quality of her signal and decreases with price informativeness (as this reduces her informational advantage).

Let $\bar{\Pi}_{dm}$ and $\bar{\Pi}_{ex}$ be, respectively, the (cross-sectional) average data miners' performance and experts' performance in equilibrium. Using eq.(27), we obtain:

$$\bar{\Pi}_{dm} = \mathsf{E}_\phi[\bar{\Pi}(\tau)|\tau^* \leq \tau \leq \tau_{dm}^{max}] = \frac{\bar{\tau}_{dm}}{\rho\sigma_\omega^2\mathcal{I}(\tau^*;\tau_{dm}^{max})}, \tag{28}$$

and

$$\bar{\Pi}_{ex} = \mathsf{E}_\gamma[\bar{\Pi}(\tau)|\tau^* \leq \tau \leq \tau_{ex}^{max}] = \frac{\bar{\tau}_{ex}}{\rho\sigma_\omega^2\mathcal{I}(\tau^*;\tau_{dm}^{max})}, \tag{29}$$

where, $\bar{\tau}_{dm}$ and $\bar{\tau}_{ex}$ are, respectively, the average quality of data miners and experts signals (see Section IV). Hence, the average performance of all asset managers is:

$$\mathsf{E}[\bar{\Pi}(\tau)] = \mu^*\bar{\Pi}_{dm} + (1-\mu^*)\bar{\Pi}_{ex} = \frac{\bar{\tau}(\tau^*;\tau_{dm}^{max})}{\rho\sigma_\omega^2\mathcal{I}(\tau^*;\tau_{dm}^{max})}. \tag{30}$$

Berk and van Binsbergen (2015) (Table 3) estimate the (cross-sectional) mean "value added" of active asset managers in the U.S. (a measure of asset managers' performance conceptually close to $\mathsf{E}[\bar{\Pi}(\tau)]$) to $\$140,000$ per month over the 1977-2011 period, with a significant dispersion across funds. As explained in the rest of this section, our model relates (time-series) variations in this average and the cross-sectional dispersion of asset managers' performance to shocks to data miners costs (c) and the data frontier (τ_{dm}^{max}).[23]

[23]Of course, there are other determinants of asset managers' average performance and its dispersion. For instance, in the model, shocks to noise traders' volume or the volatility of the asset payoff also affect these variables. As explained in the introduction, our paper focuses on shocks to c and τ_{dm}^{max}.

A How does progress in information technology affect active asset managers' average performance?

We first study the effects of the data frontier (τ_{dm}^{max}) and data miners' search costs (c) on asset managers' average performance, $\mathsf{E}[\bar{\Pi}(\tau)]$. Using the expression for $\mathcal{I}(\tau^*;\tau_{dm}^{max})$ in eq.(14), we can rewrite eq.(30) as:

$$\mathsf{E}[\bar{\Pi}(\tau)] = \frac{1}{\rho}\left(\frac{1}{\bar{\tau}(\tau^*;\tau_{dm}^{max})} + \frac{\bar{\tau}(\tau^*;\tau_{dm}^{max})}{\rho^2\sigma_\omega^2\sigma_\eta^2}\right)^{-1}. \tag{31}$$

An increase in the average quality of asset managers' signals, $\bar{\tau}(\tau^*;\tau_{dm}^{max})$, has two opposite effects on their average performance. On the one hand, asset managers make better investment decisions on average (they are more likely to buy the asset when its return is positive and sell the asset otherwise), which raises their performance on average. On the other hand, asset managers trade more aggressively on their signals on average, which raises price informativeness and therefore reduces their average performance. Using eq.(31), we find that the first effect dominates if and only if $\bar{\tau}(\tau^*;\tau_{dm}^{max}) \leq \rho\sigma_\omega\sigma_\eta$. Thus, data miners' average expected profit reaches its maximum for $\bar{\tau}(\tau^*(\tau_{dm}^{max},c);\tau_{dm}^{max}) = \rho\sigma_\omega\sigma_\eta$ if there are values of (τ_{dm}^{max},c) for which this equality holds (we write τ^* as a function of (τ_{dm}^{max},c) to emphasize that it depends on the value of these parameters). We deduce the following result.

Corollary 1. *Denote by $\bar{\tau}_0$ the average predictor's quality in the absence of data miners (i.e $\mu^* = 0$).*

1. *If $\bar{\tau}(\tau^*(\tau_{dm}^{max},0),\tau_{dm}^{max}) > \rho\sigma_\omega\sigma_\eta > \bar{\tau}_0$ then asset managers' average performance is a hump-shaped function of c, which reaches its maximum for $c = \hat{c}$ (characterized in the proof of the proposition). If $\bar{\tau}(\tau^*(\tau_{dm}^{max},0),\tau_{dm}^{max}) \leq \rho\sigma_\omega\sigma_\eta$ then asset managers' average performance decreases with c. And, if $\rho\sigma_\omega\sigma_\eta \leq \bar{\tau}_0$ then asset managers' average performance increases with c.*[24]

2. *If $\bar{\tau}(\tau^*(\infty,c),\infty) > \rho\sigma_\omega\sigma_\eta > \bar{\tau}_0$ then asset managers' average performance is a hump shaped function of τ_{dm}^{max}, which reaches its maximum for $\tau_{dm}^{max} = \hat{\tau}^{max}$ (characterized in the proof of the proposition). If $\bar{\tau}(\tau^*(\infty,c),\infty) \leq \rho\sigma_\omega\sigma_\eta$ then asset*

[24]The two last cases correspond to cases in which there is no value of c in $(0,\infty)$ such that $\bar{\tau}(\tau^*(\tau_{dm}^{max},c);\tau_{dm}^{max}) = \rho\sigma_\omega\sigma_\eta$. Thus, $\mathsf{E}[\bar{\Pi}(\tau)]$ reaches its maximum either for $c = 0$ (first case) or $c = \infty$ (2nd case).

managers' average performance increases with τ_{dm}^{max}*. And, if* $\rho\sigma_\omega\sigma_\eta \leq \bar{\tau}_0$ *then asset managers' average performance decreases with* τ_{dm}^{max}*.*

As explained previously, the big data revolution has reduced quant funds' search costs (c) and pushed back the data frontier, τ_{dm}^{max}. According to Corollary 1, this evolution should first raise the average performance of asset managers but eventually reduce it (see Figure 3). Intuitively, in the long run, the positive effect of a decrease in c or an increase in τ_{dm}^{max} on price informativeness dominates, so that the average performance of asset managers drop. We are not aware of empirical papers analyzing the long run evolution of the average performance of active asset managers and relating it to trends in computing power and the availability of new data.

This implication is related to Stambaugh (2020) (his Proposition 5 and Figure 5) who shows that improvements in the skills of active asset managers can have a negative effect on their average performance if the fraction of asset managers who experience an increase in their skills is large enough. One important difference between his model and ours is that we explicitly relate the shifts in asset managers' skills to shocks to the cost of search for data miners or the data frontier (the rise of new datasets to forecast asset payoffs).

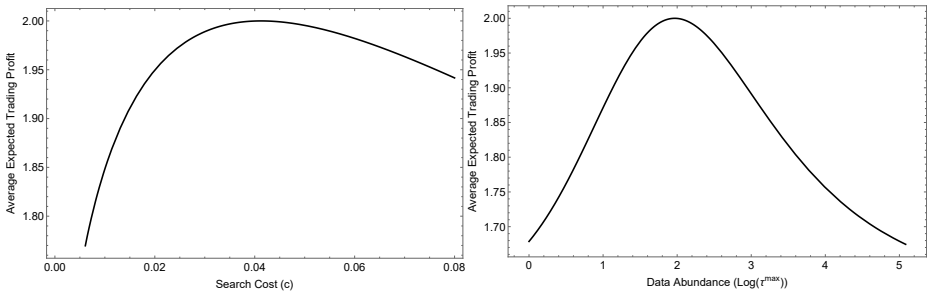

Figure 3: Left: Asset managers' average performance, $\mathsf{E}(\bar{\Pi})$, as a function of the search cost, c (other parameter values are $\tau_{dm}^{max} = \cot^2(\pi/10) \simeq 9.5, \rho = 1, \sigma_\omega = \sigma_\eta = 2$). **Right:** Asset managers' average performance, $\mathsf{E}(\bar{\Pi})$, as a function of the data frontier, $\log(\tau_{dm}^{max})$ (other parameter values are $c = 0.03, \rho = 1, \sigma_\omega = \sigma_\eta = 2$). Underlying distributions of signals' quality are taken as $\Phi(\tau) = \frac{1-(1+\tau)^{-3/2}}{1-(1+\tau_{dm}^{max})^{-3/2}}$ and $\Gamma(\tau) = 1-(1+\tau)^{-3/2}$.

B How does progress in information technology affect the dispersion in asset managers' performance?

Barras et al. (2022) find substantial variations in the profitability (measured by their alphas on the first dollar invested) of investment ideas for active U.S. equity funds.[25]

[25]Barras et al. (2022) (Table II, Panel A) find that the average (gross) alpha of active mutual fund

This is consistent with our model in which asset managers differ in their skill (τ) and therefore their average performance ($\bar{\Pi}(\tau)$). In our theory the dispersion in managers' skills, and therefore performance, is determined by data miners' search intensity (τ^*). Thus, shocks to the data frontier and data miners' search costs also affect the dispersion in asset managers' performance within fund groups (data miners and experts).

We measure the dispersion of performance across data miners by $\Delta\pi_{dm}$, the log difference between the average performance of the data miner with the best signal and that with the worst signal:

$$\Delta\pi_{dm} \equiv \log\left(\bar{\Pi}(\tau_{dm}^{max})\right) - \log\left(\bar{\Pi}(\tau^*)\right) = \log\left(\tau_{dm}^{max}\right) - \log\left(\tau^*\right), \tag{32}$$

where the second equality follows from eq.(28). Similarly, we measure the dispersion in performance across experts by (remember that τ_{ex}^{max} denotes the upper bound of experts' skills distribution):

$$\Delta\pi_{ex} \equiv \log\left(\bar{\Pi}(\tau_{ex}^{max})\right) - \log\left(\bar{\Pi}(\tau^*)\right) = \log\left(\tau_{ex}^{max}\right) - \log\left(\tau^*\right). \tag{33}$$

Corollary 2.

1. *Other things equal, a reduction in data miners' search costs, c, reduces the dispersion in performance for both data-miners and experts ($\Delta\pi_{dm}$ and $\Delta\pi_{ex}$ decrease when c decreases).*

2. *Other things equal, for $\tau_{dm}^{max} \geq \tau_{tr}(c)$, a push back of the data frontier (an increase in τ_{dm}^{max}) increases the dispersion in performance for both data-miners and experts ($\Delta\pi_{ex}$ and $\Delta\pi_{dm}$ increase with τ_{dm}^{max}).*

These implications follow directly from the fact that shocks to computing power and data abundance affect data miners' search intensity, τ^*, in opposite directions when τ_{dm}^{max} is large enough (see Propositions 4 and 5).

A more general approach is to measure the dispersion in asset managers' performance by the interquartile range of $\bar{\Pi}(\tau)$ (rather than the difference between top and bottom

(their sample is the entire population of open-end actively managed US equity funds) is 3% with a *cross-sectional* standard deviation of 4.1%. Moreover, this heterogeneity in funds' skills is a source of dispersion in "value added", that is, a fund's gross alpha times its size (similar to $\bar{\Pi}(\tau)$ in our model). For instance, Barras et al. (2022) find that the mean value added of a fund in their sample is \$1.9 million with a (cross-sectional) standard deviation of \$13.6 million.

performers). Corollary 2 also holds with this approach. To see this, let $\bar{\Pi}^j_\alpha$ be such that α-percent of all asset managers with type $j \in \{dm, exp\}$ have an average profit smaller than $\bar{\Pi}^j_\alpha$ and let ΔQ^j_α denote the log difference between the α and $(1-\alpha)$ (for $\alpha > 0.5$) quantiles of the distribution of average asset managers' performance. That is:

$$\Delta Q^j_\alpha = \log(\bar{\Pi}^j_\alpha) - \log(\bar{\Pi}^j_{1-\alpha}) \quad \text{for} \quad j \in \{dm, exp\}. \tag{34}$$

The next corollary shows that the conclusions of Corollary 2 holds when one measures dispersion in performance with ΔQ^j_α, provided that α is large enough.[26]

Corollary 3. *When $\tau^{max}_{ex} < \infty$ and α large enough then*

1. *Other things equal, a reduction in data miners' search costs (c) reduces ΔQ^{dm}_α and ΔQ^{ex}_α.*

2. *Other things equal, for $\tau^{max}_{dm} \geq \tau_{tr}(c)$, a push back of the data frontier (an increase in τ^{max}_{dm}) increases ΔQ^{dm}_α and ΔQ^{ex}_α.*

Figure 4 provides a numerical illustration of Corollary 3 when $\alpha = 90\%$.

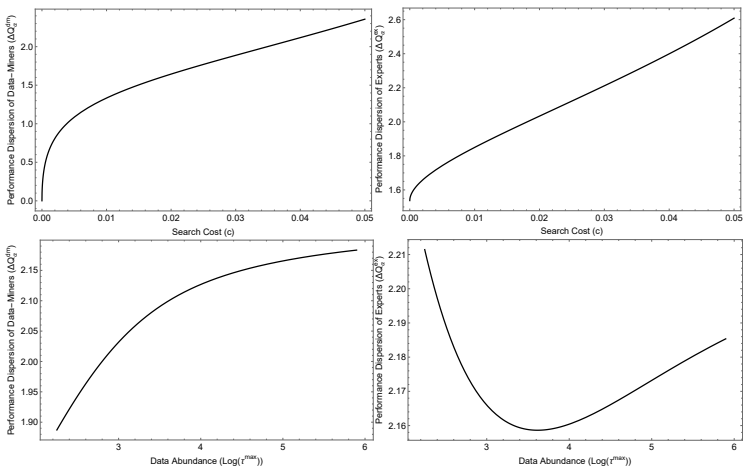

Figure 4: Upper Panels: Performance dispersion of data-miners, ΔQ^{dm}_α (left-hand side), and experts, ΔQ^{ex}_α (right hand-side), as functions of the search cost, c (other parameter values are $\alpha = 0.9, \tau^{max}_{dm} = \cot^2(\pi/10) \simeq 9.5, \rho = \sigma^2_\omega = \sigma^2_\eta = 1$). **Lower Panels:** Performance dispersion of data-miners, ΔQ^{dm}_α (left-hand side), and experts, ΔQ^{ex}_α (right hand-side), as functions of data abundance, $\log(\tau^{max}_{dm})$ (other parameter values are $a = 0.9, c = 0.03, \rho = \sigma^2_\omega = \sigma^2_\eta = 1$). Underlying distributions of signals' quality are taken as $\Phi(\tau) = \frac{1-(1+\tau)^{-3/2}}{1-(1+\tau^{max}_{dm})^{-3/2}}$, and $\Gamma(\tau) = 1-(1+\tau)^{-3/2}$ with $\tau^{max}_{ex} = \infty$.

[26]For the corollary, we assume that τ^{max}_{ex} is finite. We show in the proof of the corollary that Corollary 3 still holds when τ^{max}_{ex} is infinite under a technical condition on $\Gamma(.)$, the distribution of experts' skills.

C Progress in Information Technology: Data Miners vs Experts

When c decreases or τ_{dm}^{max} increases, the average performance of all asset managers first rise and then decline (see Corollary 1). This pattern also holds within a given group of asset managers (expert or data miner). However, the sensitivity of the average performance of each group to a shock on c or τ_{dm}^{max} is not the same. Thus, such a shock can reduce or increase data miners' average performance relative to experts. We measure this relative performance by RP:

$$RP \stackrel{def}{=} \log\left(\bar{\Pi}_{dm}\right) - \log\left(\bar{\Pi}_{ex}\right) = \log(\bar{\tau}_{dm}) - \log(\bar{\tau}_{ex}), \tag{35}$$

where the second equality follows from eq.(28) and eq.(29). In the model, data miners' relative performance can be positive or negative depending on parameter values.[27]

Corollary 4. *When $\tau_{dm}^{max} \geq \tau^{tr}(c)$, the average performance of data miners relative to experts improves when τ_{dm}^{max} increases.*

This result is a direct consequence of Proposition 5. When $\tau_{dm}^{max} \geq \tau^{tr}(c)$, a push back of the data frontier, τ_{dm}^{max}, improves the average quality of data miners' signals ($\bar{\tau}_{dm}$) but it reduces the average quality of experts' signals ($\bar{\tau}_{ex}$) because investors allocate more capital to experts (and therefore the quality of the marginal expert drops). It follows that data miners' relative performance improves (as for Proposition 5, the condition that $\tau_{dm}^{max} \geq \tau^{tr}(c)$ is sufficient but not necessary for this implication; see Figure 5).

In contrast, a decrease in data miners' search cost, c, improves the average quality of the signals used by both groups of asset managers (Proposition 4). Thus, its effect on data miners' relative performance is less clear and one cannot sign it without additional assumptions on the distribution of experts' signals quality ($\Gamma(.)$). When this distribution is similar to the distribution of data miners' signals quality (in the sense that $\Gamma(\tau) = \Psi(\tau)/\Psi(\tau_{ex}^{max})$), we obtain the following result.

Corollary 5. *Suppose that $\Gamma(\tau) = \Psi(\tau)/\Psi(\tau_{ex}^{max})$. When τ_{dm}^{max} and τ_{ex}^{max} are high enough, the average performance of data miners relative to experts decreases when data miners' search cost, c, decreases if and only if and $\tau_{ex}^{max} > \tau_{dm}^{max}$.*

[27] For instance, if τ_{ex}^{max} is high enough, the average precision of experts' signals $\bar{\tau}_{ex}$ tends to exceed the average precision of data miners' signals and therefore RP is negative. Abis (2022) finds that quant funds have a lower average performance than discretionary funds, especially in recessions.

Figure 5 illustrate the results in this section. For the parameter values considered in Figure 5, experts perform better on average than data miners ($RP < 0$). Their relative performance improves when the data frontier (τ_{dm}^{max}) increases and decreases when data miners' search cost is reduced.

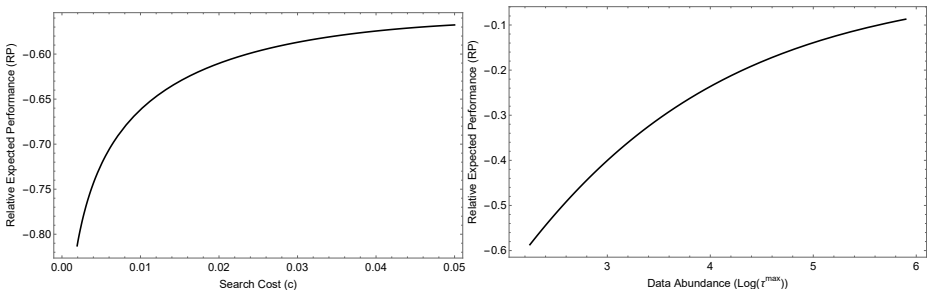

Figure 5: Left: Data miners' relative performance, RP, as a function of the search cost, c (other parameter values are $\tau_{dm}^{max} = \cot^2(\pi/10) \simeq 9.5, \rho = \sigma_\omega^2 = \sigma_\eta^2 = 1$). **Right:** Data miners' relative performance, RP, as a function of the data frontier, $\log(\tau_{dm}^{max})$ (other parameter values are $c = 0.03, \rho = \sigma_\omega^2 = \sigma_\eta^2 = 1$). Underlying distributions of signals' quality are taken as $\Phi(\tau) = \frac{1-(1+\tau)^{-3/2}}{1-(1+\tau_{dm}^{max})^{-3/2}}$, and $\Gamma(\tau) = 1-(1+\tau)^{-3/2}$ with $\tau_{ex}^{max} = \infty$.

VII. Management Fees

To simplify the exposition, we have assumed so far that experts and data miners do not obtain no rents from their services to investors. This assumption can be relaxed without affecting the implications of the model. To do so, consider the equilibrium described in Proposition 3. In this equilibrium, the certainty equivalent of investors who allocate their capital to an expert with skill τ is $w(\tau) \equiv -\rho^{-1}\log(-g(\tau,\tau^*))$ (see Lemma 2). Similarly, the certainty equivalent of investors who allocate their capital to a data miner is $w(\tau^*) \equiv -\rho^{-1}\log(-g(\tau^*,\tau^*))$. Now, following Garleanu and Pedersen (2018), assume that the surplus created by asset managers for investors is shared between these parties through Nash bargaining, at date 0. Thus, the fee charged by an expert with skill τ, denoted $f_{ex}(\tau)$, solves:

$$\max_{f_{ex}(\tau)} \quad f_{ex}(\tau)^\kappa (w(\tau) - f_{ex}(\tau) - w_o^*)^{1-\kappa}, \tag{36}$$

where κ is an asset manager's bargaining power (in the baseline model, $\kappa = 0$) and $w_o^* \leq w(\tau^*)$ is the investor's outside option.[28] Similarly, the fee charged by a data miner, denoted f_{dm}, solves:

$$\max_{f_{dm}} \quad f_{dm}^{\kappa}(w(\tau^*) - f_{dm} - w_o^*)^{1-\kappa}, \tag{37}$$

Thus, the fees charged by experts and data miners are:

$$f_{ex}^*(\tau) = \kappa(w(\tau) - w_o^*), \quad \text{and} \quad f_{dm}^* = \kappa(w(\tau^*) - w_o^*) \tag{38}$$

Hence, net of fees, an investor who allocates her capital to an expert with skill τ obtains $(1-\kappa)w(\tau) + \kappa w_o^*$ while an investor who allocates her capital to a data miner obtains $(1-\kappa)w(\tau^*) + \kappa w_o^*$.

As $w(\tau)$ increases with τ, investors' rankings of asset managers after accounting for fees is the same for all values of $\kappa < 1$. Thus, as in the baseline model ($\kappa = 0$), investors should optimally allocate a fraction $(1 - \Gamma(\tau^*))$ of their capital to all experts with a skill larger than τ^* and the rest to data miners. This means that the equilibrium described in Proposition 3 (and therefore all the implications of the model) remains valid when $\kappa < 1$.[29]

Observe that when $\mu^* = \Gamma(\tau^*) < 1$, some data miners do not receive capital and an investor's best outside option is to allocate capital to such a data miner. If he exerts this option, the investor obtains $(1-\kappa)w(\tau^*) + \kappa w_o^*$. Thus, it is natural to set w_0^* so that $w_o^* = (1-\kappa)w(\tau^*) + \kappa w_o^*$, that is,

$$w_o^* = w(\tau^*). \tag{39}$$

In this case, we deduce from eq.(38) that

$$f_{ex}^*(\tau) = \kappa(w(\tau) - w(\tau^*)), \quad \text{and} \quad f_{dm}^* = 0, \tag{40}$$

so that data miners capture no rents in equilibrium. This is intuitive. Indeed, ex-ante all data miners are identical (they have the same ability to find predictors) and their

[28] As in Garleanu and Pedersen (2018), we normalize asset managers' outside option (reservation wage) to zero. This does not affect the conclusions in this section.

[29] When $\kappa = 1$, investors are indifferent between all asset managers because managers extract all surplus. To break this tie, one can assume that they allocate their capital to asset managers that deliver the largest certainty equivalent gross of fees. The equilibrium is then as described in Proposition 3.

demand for capital exceeds available supply by investors. Thus, competition between data miners should lead them to charge fees equal to their reservation wage, as obtained in eq.(40). In contrast, each expert is "scarce" in the sense that the "mass" of experts with a skill larger than τ $(1-\Gamma(\tau))$ is smaller than the total mass of investors (1). This scarcity enables experts to extract larger fees than data miners. In line with this implication, Abis (2022) finds that quant funds charge smaller fees than discretionary investors. Another implication of the model is that, other things equal, fees should increase with skills, which is in line with empirical findings in Gerakos et al. (2021) (see their Table V).

Interestingly, experts' fees depends on data miners' search intensity. Other things equal, an increase in data miners' search intensity triggers a drop in the allocation of capital to experts and should reduce their fees. Variations in data miners' search intensity are endogenous and driven by shocks to parameters, in particular data miners' search costs and the data frontier. Such shocks affect both $w(\tau)$ (because they affect price informativeness) and $w(\tau^*)$. We obtain the following corollary.

Corollary 6. *Experts' fees decline when data miners' search cost (c) decline. They also decline following an increase in the data frontier if this increase raises data miners' search intensity. When it does not, the fees charged by experts with low skills (τ close enough to τ^*) increases.*

When data miners' search costs decline, price informativeness increases. This effect reduces investors' certainty equivalents ($w(\tau)$ and $w(\tau^*)$), both with data miners and experts. However, this reduction is smaller for data miners because the drop in search costs induces them to search more intensively (τ^* increases). Thus, allocating capital to data miner becomes relatively more attractive ($w_0^* = w(\tau^*)$ rises) and experts must therefore decrease their fees.

The same effects play out when an increase in the data frontier, τ_{dm}^{max}, raises data miners' search intensity. However, when it does not ($\tau_{dm}^{max} \geq \tau^{tr}(c)$), effects are reversed: A push back of the data frontier reduces data miners' search intensity and therefore the certainty equivalent of investors' expected utility with data miners (their outside option). Thus, the net effect of raising τ_{dm}^{max} on experts' fees is ambiguous: It lowers investors' outside option (w_0^*) (a positive effect on experts' fees) but it also increases price informativeness (a negative effect). For experts with relatively low skills (those with a τ close to the marginal expert), the first effect dominates so that their fees increase

(Corollary 6). For highly skilled experts, numerical simulations suggest that enhanced price informativeness is the dominating force so that their fees always drop when τ_{dm}^{max} increases (but we have not been able to prove this result analytically). Figure 6 illustrates these findings for a specific parametrization of the model.

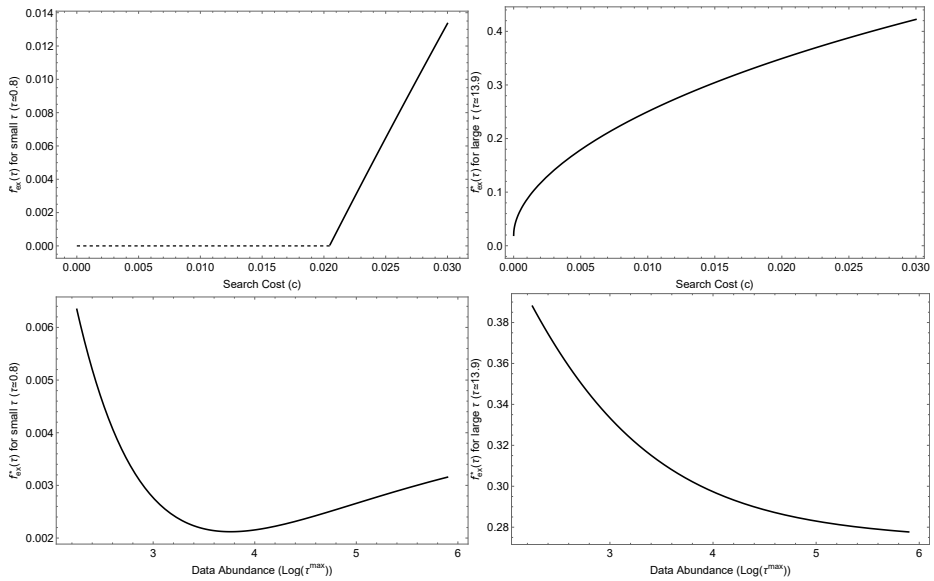

Figure 6: Upper Panels: Experts' fees, $f_{ex}^*(\tau)$ as function of the search cost, c, for two different levels of quality: Low (left-hand side, $\tau \simeq 0.8$), and high quality (right-hand side, $\tau \simeq 13.9$) (other parameter values are $\kappa = 0.5, \tau_{dm}^{max} = \cot^2(\pi/10) \simeq 9.5, \rho = \sigma_\omega^2 = \sigma_\eta^2 = 1$). On the left-hand side, the dashed line corresponds to search cost values for which the low quality expert does not receive capital. **Lower Panels:** Experts' fees, $f_{ex}^*(\tau)$, as a function of the data frontier, $\log(\tau_{dm}^{max})$, for two different levels of quality: Low (left-hand side, $\tau \simeq 0.8$), and high quality (right-hand side, $\tau \simeq 13.9$) (other parameter values are $\kappa = 0.5, c = 0.025, \rho = \sigma_\omega^2 = \sigma_\eta^2 = 1$). Underlying distributions of signals' quality are taken as $\Phi(\tau) = \frac{1-(1+\tau)^{-3/2}}{1-(1+\tau_{dm}^{max})^{-3/2}}$, and $\Gamma(\tau) = 1-(1+\tau)^{-3/2}$ with $\tau_{ex}^{max} = \infty$.

VIII. Empirical Implications

Table I summarizes the main implications of the model regarding the effects of a decrease in information processing costs for data miners (c) and a push back of the data frontier (τ_{dm}^{max}). The direction of these effects are identical for (i) price informativeness ($\mathcal{I}$), (ii) the average quality of active asset managers' signals ($\overline{\tau}$), and (iii) their average performance ($\mathrm{E}(\overline{\Pi})$). Indeed, as explained previously, a push back of the data frontier or lower information processing costs raise the cross-sectional average quality of asset managers'

signals and therefore price informativeness[30] As a result, asset managers' average performance is a hump-shaped function (see Corollary 1) of information processing costs (c) and data abundance (τ_{dm}^{max}).

However, for other variables of interest (e.g., the allocation of capital to data miners) a push back of the data frontier or a reduction in information processing costs for data miners have different effects. A reduction in information processing costs always leads data miners to raise the bar for the quality of their signals (τ^* increases when c declines). As a result, more capital gets allocated to data miners, the average quality of the signals used by each group of asset managers increases and the dispersion in asset managers' performance within each group declines. In contrast, when τ_{dm}^{max} is large enough, a further push back of the data frontier leads data miners to search less intensively for their signals. As a result, more capital gets allocated to experts and the dispersion of performance across both groups of asset managers increases.

To our knowledge, this second set of implications is unique to our model. Indeed, they stem from differences in data miners' choices for the precision of their signals following shocks to (i) information processing costs or (b) the set of feasible precisions, holding information processing costs constant. As explained in Section II, existing models do not allow to study the second type of shocks. To test these predictions, empiricists must (a) identify quant funds in the universe of active funds and (b) use sources of variations in data abundance (τ_{dm}^{max}) that are independent from variations in quant funds' information processing costs (c). To address the first issue (funds' classification), researchers can use methodologies and data similar to those used by Harvey et al. (2017) or Abis (2022). To address the second one, empiricists could use the approaches that we outlined below.[31]

Over the last 20 years, data vendors have introduced new datasets, using data generated by (i) individuals (e.g., when they interact on social media such as Twitter or StockTwits), (ii) business processes (e.g., credit card data) or (iii) sensors (e.g., satellite imagery of land or parking lots utilization). This evolution has generated a steady increase in variables that quant funds can use to increase the predictive power of their trading signals.[32]. Importantly, the availability of such datasets varies across industries

[30]Consistent with this implication, Zhu (2019) and Katona et al. (2019) find that the introduction of new alternative data (satellite imagery and consumer-browsing data)–a positive shock on τ_{dm}^{max}–improves price informativeness for the stocks covered by this data.

[31]We just sketch possible methods. There are certainly others. Our goal is just to show that separating shocks to data abundance from shocks to information processing costs is not impossible.

[32]JP Morgan 2019 handbook of alternative data lists more than 500 hundred alternative datasets

	Lower Search Costs for DMs' (c ↘)		**Data Abundance (τ_{dm}^{max} ↗)**	
Alloc. of capital to DMs (μ^*)	↗		↗↘	
Price Informativeness ($\mathcal{I}$)	↗		↗	
Average Signal Quality ($\overline{\tau}$)	↗		↗	
DMs' Relative Performance (RP)	↘*		↗*	
	Experts	**Data Miners**	**Experts**	**Data Miners**
Quality of the worst signal (τ^*)	↗	↗	↗↘*	↗↘*
Average signal quality ($\bar{\tau}_{ex}$ and $\bar{\tau}_{dm}$)	↗	↗	↗↘*	↗
Average Performance (E($\overline{\Pi}$))	↗↘	↗↘	↗↘	↗↘
Dispersion in Performance (ΔQ_α)	↘	↘	↘↗*	↘↗*

Table I: This table summarizes the implications of the model: (a) ↗ means a positive effect or an increase (b) ↘ means a negative effect or an increase, (c) ↗↘ means that the variable of interest (1st column) is a hump-shaped function of computing power or data abundance. "DM" means "Data Miner". A "*" means that the effect of data miners' search costs or the data frontier when c is large and τ_{dm}^{max} is small has not been derived analytically but numerically.

and asset classes (see Figures 23 and 26 in JP Morgan 2019 and Section IV.C in Dessaint et al. (2022). For instance, there are far more alternative datasets covering firms in the consumer industry than in health care or energy. Similarly, there are far more available alternative datasets covering equity than covering commodities or relevant for macro strategies. Thus, as quant funds cover different industries or asset classes, they are not exposed to the same extent to the rise of alternative datasets (data abundance).[33] Empiricists can use this feature to build a measure of exposure to alternative data of quant funds and use panel data on funds' performance to test whether variations in exposure of a given group of quant funds trigger (i) variations in the dispersion of funds' performance within this group consistent with our predictions and (ii) allocation of capital between quant funds and discretionary funds operating in the same industry or covering the same asset class.

With this approach, identification of the effects of data abundance on some of the variables of interest would come from cross-sectional variations in funds' exposure to alternative data (shocks to τ_{dm}^{max}). Such variation should not be correlated with a downward trend in the cost of processing data (c in our model) since this trend should affect all funds at the same time and can therefore be controlled by time-fixed effects.

To test our predictions regarding the effects of c, empiricists could use technological changes that reduce the cost of data processing for quant funds. One such shock is the introduction of cloud computing by Amazon with its release of Amazon Web Services (AWS) in 2006. This shock significantly reduced the cost of computing power (see Ewens et al. (2018)) and, for this reason, it has been used to study empirically the effect of AI adoption by venture capitalists (see Ewens et al. (2018) and Bonnelli (2022)). Another shock on data processing costs is the SEC's eXtensible Business Reporting Language (XBLR) mandate. This mandate requires firms to provide their regulatory filings in machine-readable form (through the EDGAR website) since 2009. Importantly, this shock did not change the information content of regulatory filings (firms had to provide the same data before and after the mandate). However, it reduced the cost of automated data processing (e.g., using Natural Language Processing techniques; see, for instance, Bhattacharya et al. (2018)) and therefore c for funds using quantitative techniques to

and used cases for asset managers. See "J.P.-Morgan-Alternative-Data-Handbook-2019", retrieved at: https://ea-pdf-items.s3-eu-west-1.amazonaws.com/J.P.-Morgan-Alternative-Data-Handbook-2019.pdf

[33]Empiricists could use CRSP mutual funds holdings dataset to build a measure of funds' exposure to alternative data based on its holdings.

generate trading signals. Interestingly, the XBLR mandate was introduced in a staggered way over a three year periods from 2009 to 2011. Thus, depending on the nature of their holdings, funds were differentially exposed to the XBLR mandate. This offers another way to study the effect of a reduction in information processing costs for quant funds, holding available data constant.

Many of our implications are about the effects of variations in data miners' information processing cost (c) or data abundance (τ_{dm}^{max}) on the average quality of data miners and experts' signals ($\bar{\tau}_{dm}$ and $\bar{\tau}_{ex}$) or the quality of the worst signal used by asset managers (τ^*). One way to test them is to directly estimate the quality of funds' signals (τ) by regressing (in the time series) a fund's holding in an asset on the direction of subsequent returns. In our model, the theoretical coefficient, β_τ, of this regression is

$$\beta(\tau) \equiv \frac{\mathsf{Cov}(x^*(s_\tau, p^*), \omega - p^*)}{\mathsf{Var}[\omega - p^*]} = \frac{\tau}{\rho\sigma_\omega^2}, \tag{41}$$

where the last equality follows from Proposition 1. Thus, $\beta(\tau)$ is strictly positive if the manager is informed ($\tau > 0$) and increases with the quality of the manager's information (τ). Intuitively, $\beta(\tau)$ is a measure of an asset manager's timing ability (see Kacperczyk et al. (2014) and Gerakos et al. (2021) for evidence that active asset managers have timing ability using a measure similar to $\beta(\tau)$). All our predictions regarding the effects of the parameters of the model on the quality of asset managers' signals (e.g. the average quality of asset managers' signals or the quality of the signal of the marginal asset manager) also hold for $\beta(\tau)$.[34]

IX. Conclusion

Our paper studies the effects of the big data revolution on active asset managers using a noisy rational expectations model. Our model contrasts two types of asset managers who invest in a risky asset: (a) experts (which we interpret as discretionary funds) who have a fixed ability to generate trading signals of a given precision about the asset and (b) data miners (which we interpret as quant funds). The latter obtain their trading signals through a search process whose goal is to identify trading signals with high precision. We

[34]For instance, one could test whether an increase in computing power leads to an increase in the quality of the funds with the lowest timing ability (say, in the 10^{th} percentile) as Kacperczyk et al. (2014) do for the effects of recessions (see their Table III).

assume that the highest available precision (the data frontier) increases as data miners have access to more data (data abundance).

Data miners' optimally stop searching for a signal once they discover a signal with a precision exceeding an endogenous threshold (data miners' search intensity), determined by (a) data miners' search costs and (b) the data frontier. We focus on the equilibrium effects of a reduction in data miners' search costs or a push back of the data frontier on (i) their search intensity, (ii) the allocation of capital between data miners and experts, (iii) asset price informativeness, (iv) asset managers' average performance, (v) the cross-sectional dispersion in asset managers' performance and (vi) asset managers' fees.

An important new feature of our model is that it allows to separately analyze the effects of reducing data miners' costs (e.g., due to greater computing power) and the effects of a push back the data frontier (due to new data availability). Our model shows that these two distinct dimensions of the big data revolution do not necessarily have the same implications. For instance, the capital allocated to data miners always rises with a decline in their search costs while it is a hump-shaped function of the data frontier. More generally, we develop a rich set of testable predictions about the effects of the big data revolution on active asset managers.

References

Abis, Simona, 2022, Man vs machine: Quantitative and discretionary equity management, Working paper, Columbia University.

Banerjee, Snehal, and Bradyn Breon-Drish, 2021, Dynamics of Research and Strategic Trading, *The Review of Financial Studies* 35, 908–961.

Barras, Laurent, Patrick Gagliardini, and Olivier Scaillet, 2022, Skill, scale, and value creation in the mutual fund industry, *The Journal of Finance* 77, 601–638.

Begs, William, Jonathan Brogaard, and Austin Hill-Kleespie, 2019, Quantitative investing and market instability, Working paper, University of Utah.

Berk, Jonathan, and Jules van Binsbergen, 2015, Measuring skill in the mutual fund industry, *Journal of Financial Economics* 110, 1–20.

Bhattacharya, Nilabhra, Young Cho, and Jae Kim, 2018, Leveling the playing field between large and small institutions: Evidence from the sec's xbrl mandate, *The Accounting Review* 93.

Bond, Philip, and Diego García, 2021, The Equilibrium Consequences of Indexing, *The Review of Financial Studies* 35, 3175–3230.

Bonnelli, Maxime, 2022, The adoption of AI intelligence by venture capitalists, Working paper, HEC Paris.

Buss, Adrian, and Savitar Sundaresan, 2022, More risk, more information: How passive ownership can improve informational efficiency, Working paper, Imperial College.

Chevalier, Judith, and Glenn Ellison, 1999, Are some mutual fund managers better than others? cross-sectional patterns in behavior and performance, *The Journal of Finance* 54, 875–899.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard, 2022, Does alternative data affect financial forecasting? the horizon effect, Working paper, Forthcoming Journal of Finance.

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

Ewens, Michael, Ramana Nanda, and Matthew Rhodes-Kropf, 2018, Cost of experimentation and the evolution of venture capital, *Journal of Financial Economics* 128, 422–442.

Farboodi, Maryam, and Laura Veldkamp, 2020, Long run growth of financial technology, *American Economic Review* 110, 2485.

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

Gerakos, Joseph, Juhani T. Linnainmaa, and Adair Morse, 2021, Asset managers: Institutional performance and factor exposures, *The Journal of Finance* 76, 2035–2075.

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

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

Harvey, Campbell R., Sandy Rattray, Andrew Sinclair, and Otto Van Hemert, 2017, Man vs. machine: Comparing discretionary andsystematic hedge fund performance, *The Journal of Portfolio Management* 43, 55–69.

Huang, Shiyang, Zhigang Qiu, and Liyan Yang, 2019, Institutionalization, Delegation, and Asset prices, *The Journal of Economic Theory* 186, 1–42.

Huang, Shiyang, Yan Xiong, and Liyan Yang, 2022, Skill acquisition and data sales, *Management Science* 68, 6116–6144.

JPMorgan, 2017, Big data and AI strategies, Handbook.

Kacperczyk, Marcin, Jaromir Nosal, and Savitar Sundaresan, 2022, Market power and price informativeness, Working paper, Imperial College.

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, Working paper, available at https://dx.doi.org/10.2139/ssrn.3222741.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1336.

Li, Haitao, Xiaoyan Zhang, and Rui Zhao, 2011, Investing in talents: Manager characteristics and hedge fund performances, *The Journal of Financial and Quantitative Analysis* 46, 59–82.

Malikov, George, and Paolo Pasquariello, 2022, Quants, strategic speculation and financial market quality, Working paper, available at https://dx.doi.org/10.2139/ssrn.3890275.

Martin, Ian W.R., and Stefan Nagel, 2022, Market efficiency in the age of big data, *Journal of Financial Economics* 145, 154–177.

Milhet, Roxana, 2020, Financial innovation and the inequality gap, Working paper, HEC Lausanne.

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

Nordhaus, William, 2021, Are we approaching an economic singularity? information technology and the future of economic growth, *American Economic Review: Macroeconomics* 13, 299–332.

Stambaugh, Robert, 2020, Skill and profits in active management, Working paper, University of Pennsylvania.

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

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

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

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

Zhu, Min, 2018, Informative fund size, managerial skill, and investor rationality, *Journal of Financial Economics* 130.

A. Proofs

Proof of Proposition 1. We show that $x^*(s_\tau, p)$ and p^* as given by eq.(11) and eq.(12) form an equilibrium. First, suppose that $x^*(s_\tau, p)$ is given by $x^*(s_\tau, p) = \beta(\tau)(s_\tau - p)$. In this case, the aggregate demand for the asset is given by:

$$D(p) = \int x^*(s_\tau, p)di + \eta = \bar{\beta}(\omega - p) + \eta, \tag{42}$$

where $\bar{\beta}$ is the average value of $\beta(\tau)$ across all asset managers ($\bar{\beta} = \mu^* E_{\phi^*}[\beta(\tau) \mid \tau \in [\tau^*, \tau_{dm}^{max}]] + (1-\mu^*)E_\gamma[\beta(\tau) \mid \tau \geq \tau^*]$). Hence, observing $D(p)$ (and p) is informationally equivalent to observing $\xi = \omega + \bar{\beta}^{-1}\eta$. Thus:

$$p^* = \mathsf{E}\left[\omega \mid D(p)\right] = \mathsf{E}[\omega \mid \xi] = \left(\frac{\sigma_\omega^2}{\sigma_\omega^2 + \bar{\beta}^{-2}\sigma_\eta^2}\right)\xi = \left(\frac{\theta_\xi}{\theta_\omega + \theta_\xi}\right)\xi, \tag{43}$$

where $\theta_\omega \equiv 1/\sigma_\omega^2$ is the precision of asset managers' prior about the asset pay-off ω, and $\theta_\xi \equiv \bar{\beta}^2/\sigma_\eta^2$ is the precision of ξ as a signal about ω.

Now consider asset managers. Using standard calculations in the CARA gaussian framework, we obtain that the optimal demand for the risky asset of an asset manager with signal s_τ is:

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

As asset managers have rational expectations on the price, they correctly anticipate that it is linear in ξ, as in eq.(43). Note that the precision of s_τ is $\tau\theta_\omega$. Thus, as all variables are normally distributed and ε_i and η (the noises in s_τ and ξ) are independent, standard calculations yield:

$$\mathsf{E}[\omega|s_\tau, \xi] = \frac{\tau\theta_\omega s_\tau + \theta_\xi \xi}{\theta_\omega + \tau\theta_\omega + \theta_\xi}. \tag{45}$$

and

$$\mathsf{Var}[\omega|s_\tau, p] = \frac{1}{\theta_\omega + \tau\theta_\omega + \theta_\xi}. \tag{46}$$

Thus, we can rewrite eq.(44) as:

$$x^*(s_\tau, p) = \frac{\tau\theta_\omega s_\tau + \theta_\xi \xi - (\theta_\omega + \tau\theta_\omega + \theta_\xi)p}{\rho}, \tag{47}$$

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

$$x^*(s_\tau, p) = \frac{\tau\theta_\omega}{\rho}(s_\tau - p) = \frac{\tau}{\rho\sigma_\omega^2}(s_\tau - p). \tag{48}$$

Thus, $x^*(s_\tau, p)$ is as conjectured (and as in eq.(11)) if and only if $\beta(\tau) = \frac{\tau}{\rho\sigma_\omega^2}$. If follows that $\bar{\beta} = \frac{\bar{\tau}}{\rho\sigma_\omega^2}$. Eq.(12) and eq.(13) in the text immediately follow from substituting this expression for $\bar{\beta}$ in eq.(43).

In sum we have shown that (i) if dealers expect asset managers to follow the trading strategy $x^*(s_\tau, p)$ given by eq.(11) then they set a price given by eq.(12) and (ii) if dealers set a price given by eq.(12) then asset managers follow the trading strategy $x^*(s_\tau, p)$ given by eq.(11). Thus, the trading strategies and the equilibrium given in eq.(11) and eq.(12) form an equilibrium. It is possible to show that this is the unique equilibrium in which asset managers' trading strategy is a linear function of their signal and the price.

Proof of Lemma 1. Using eq(10) and the fact $\mu^* = \Gamma(\tau^*)$, we obtain:

$$\bar{\tau}(\tau^*; \tau_{dm}^{max}) = \Gamma(\tau^*) \int_{\tau^*}^{\tau_{dm}^{max}} \tau\phi^*(\tau)d\tau + \int_{\tau^*}^{\tau_{ex}^{max}} \tau\gamma(\tau)d\tau. \tag{49}$$

Thus:

$$\begin{aligned}\frac{\partial\bar{\tau}}{\partial\tau^*} &= \Gamma(\tau^*)\left(\psi(\tau^*)\frac{\int_{\tau^*}^{\tau_{dm}^{max}}\tau\psi(\tau)d\tau}{\left(\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)\right)^2} - \frac{\tau^*\psi(\tau^*)}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)}\right) + \gamma(\tau^*)\frac{\int_{\tau^*}^{\tau_{dm}^{max}}\tau\psi(\tau)d\tau}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)} - \tau^*\gamma(\tau^*) \\ &= \Gamma(\tau^*)\psi(\tau^*)\frac{\int_{\tau^*}^{\tau_{dm}^{max}}(\tau - \tau^*)\psi(\tau)d\tau}{\left(\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)\right)^2} + \gamma(\tau^*)\frac{\int_{\tau^*}^{\tau_{dm}^{max}}(\tau - \tau^*)\psi(\tau)d\tau}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)} > 0.\end{aligned} \tag{50}$$

Proof of Lemma 2. As in the proof of Proposition 1, we define $\theta_\omega \equiv 1/\sigma_\omega^2$, the precision of asset managers' prior about the asset pay-off ω, and $\theta_\xi \equiv \bar{\beta}^2/\sigma_\eta^2$, the precision of ξ as a signal about ω. Conditional on the realization of the price at date 1 and her signal, s_τ, the expected utility of trading for an investor given her optimal trading strategy is:

$$\begin{aligned}&\mathsf{E}[-\exp(-\rho x^*(s_\tau, p)(\omega - p)) \mid s_\tau, p] \\ &= -\mathsf{E}\left[\exp\left(-\rho\left(x^*(s_\tau, p)(\mathsf{E}[\omega \mid s_\tau, p] - p) - \frac{\rho(x^*(s_\tau, p))^2}{2}\mathsf{Var}[\omega \mid s_\tau, p]\right)\right)\right].\end{aligned} \tag{51}$$

Substituting $x^*(s_\tau, p)$ by its expression in eq.(44), we deduce that:

$$\mathsf{E}\left[-\exp(-\rho x^*(s_\tau,p)(\omega-p)) \mid s_\tau, p\right] = -\exp\left(-\frac{(\mathsf{E}[\omega|s_\tau,p]-p)^2}{2\,\mathsf{Var}[\omega|s_\tau,p]}\right) \tag{52}$$

Thus:

$$g(\tau,\tau^*) = -\,\mathsf{E}\left[\exp\left(-\frac{(\mathsf{E}[\omega|s_\tau,p^*]-p^*)^2}{2\,\mathsf{Var}[\omega|s_\tau,p^*]}\right)\right]. \tag{53}$$

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

$$g(\tau,\tau^*) = -\left(1+\frac{\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau,p^*]-p\right]}{\mathsf{Var}[\omega|s_\tau,p^*]}\right)^{-1/2} \tag{54}$$

Observe that:

$$\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\tau,p^*]-p^*]}{\mathsf{Var}[\omega|s_\tau,p^*]} = \rho^2\,\mathsf{Var}[\omega|s_\tau,p^*]\,\mathsf{Var}[x^*(s_\tau,p^*)]. \tag{55}$$

Now using the expression for $x^*(s_\tau,p^*)$ in eq.(48), we obtain that:

$$\mathsf{Var}[x^*(s_\tau,p^*)] = \frac{\tau^2\theta_\omega^2}{\rho^2}[\mathsf{Var}(s_\tau)+\mathsf{Var}(p^*)-2\,\mathsf{Cov}(s_\tau,p^*)]. \tag{56}$$

Using the expression for p^* in eq(43) and the fact that $s_\tau = \omega + \tau^{-\frac{1}{2}}\varepsilon_i$, we obtain after some algebra that:

$$\mathsf{Var}[x^*(s_\tau,p^*)] = \frac{\tau\theta_\omega(\theta_\omega+\theta_\omega\tau+\theta_\xi)}{\rho^2(\theta_\omega+\theta_\xi)}. \tag{57}$$

Thus, using the expression for $\mathsf{Var}[\omega|s_\tau,p^*]$ in eq.(46), we deduce that:

$$\mathsf{Var}[x^*(s_\tau,p^*)] = \frac{\tau\theta_\omega}{\rho^2(\theta_\omega+\theta_\xi)\,\mathsf{Var}[\omega|s_\tau,p^*]}. \tag{58}$$

Hence, using eq.(55) and the fact that $\theta_\xi = \frac{\bar\tau(\tau^*;\tau_{dm}^{max})^2\theta_\omega^2}{\rho^2\sigma_\eta^2}$, we deduce that:

$$\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\tau,p]-p]}{\mathsf{Var}[\omega|s_\tau,p]} = \frac{\theta_\omega\tau}{\theta_\omega+\frac{(\theta_\omega\bar\tau(\tau^*;\tau_{dm}^{max}))^2}{\rho^2\sigma_\eta^2}} = \frac{\tau}{1+\frac{\bar\tau(\tau^*;\tau_{dm}^{max})^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}} \tag{59}$$

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

Proof of Proposition 2. Using eq.(22), we obtain:

$$\frac{\partial F}{\partial \tau^*} = \int_{\tau^*}^{\tau_{dm}^{max}} \frac{\partial r(\tau, \tau^*)}{\partial \tau^*} \phi(\tau) d\tau, \tag{60}$$

where $r(\tau, \tau^*)$ is defined in eq.(23). As $\mathcal{I}(\tau^*; \tau_{dm}^{max})$ increases with τ^* (see lemma 1, we deduce from eq.(23) that $r(\tau, \tau^*)$ increases in τ^* as well. Thus, $\frac{\partial F}{\partial \tau^*} > 0$. Moreover, we have (i) $F(\tau_{dm}^{max}) = 1$, (ii) $0 < F(0) < 1$ and (iii) $\exp(-\rho c) < 1$ (since $c > 0$). Thus, there is a unique solution to the condition $F(\tau^*) = \exp(-\rho c)$ and this solution is always strictly smaller than τ_{dm}^{max}. Moreover, it is strictly larger than zero if and only if $F(0) \leq \exp(-\rho c)$.

Proof of Proposition 3. The proposition directly follows from Propositions 1 and 2 and the paragraph in the text before Proposition 3. If $F(\tau_{ex}^{max}) \leq exp(-\rho c)$ then $\tau^* > \tau_{ex}^{max}$ (since $F(.)$ increases with τ^*). In this case, $\tau^* \geq \tau_{ex}^{max}$ and therefore $\mu^* = 1$ since $mu^* = \Gamma(\tau^*)$.

Proof of Proposition 4. In equilibrium, $F(\tau^*) = \exp(-\rho c)$. The R.H.S of this equilibrium condition decreases with c and $F(.)$ increases in τ^* (see the proof of Proposition 2). We deduce that τ^* and therefore μ^* (since $\mu^* = \Gamma(\tau^*)$) decrease in c. Moreover when c goes to zero, the R.H.S of the equilibrium condition goes to 1. This implies that $F(\tau^*)$ goes to 1 as well, which (by continuity of $F(.)$) is possible only if τ^* goes to τ_{dm}^{max} (as $F(\tau_{dm}^{max}) = 1$). As τ^* decreases in c, it follows from Lemma 1 that the average quality of asset managers' signals $\bar{\tau}^2(\tau^*; \tau_{dm}^{max})$ and price informativeness also decreases with c (since c affects price informativeness only via its effect on τ^*). The average quality of data miners' signals is

$$\bar{\tau}_{dm}(\tau^*) = E_\phi \left[\tau \,|\tau^* \leq \tau \leq \tau_{dm}^{max} \right] = \frac{\int_{\tau^*}^{\tau_{dm}^{max}} \tau \psi(\tau) d\tau}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)}, \tag{61}$$

and the average quality of experts' signals is

$$\bar{\tau}_{ex} = \mathsf{E}_\gamma \left[\tau \,|\tau^* \leq \tau \right] = \frac{\int_{\tau^*}^{\infty} \tau \gamma(\tau) d\tau}{1 - \Gamma(\tau^*)}. \tag{62}$$

Clearly both are increasing in τ^*. Thus, a decrease in c raises $\bar{\tau}_{dm}(\tau^*)$ and $\bar{\tau}_{ex}(\tau^*)$ since it raises τ^*.

Proof of Proposition 5.

Step 1. Remember that $\mathcal{I}(\tau^*;\tau_{dm}^{max}) = \frac{1}{\sigma_\omega^2} + \frac{\bar{\tau}(\tau^*;\tau_{dm}^{max})^2}{\rho^2\sigma_\omega^4\sigma_\eta^2}$. Denote $\chi = \rho\sigma_\omega\sigma_\eta$. Thus, we can rewrite $r(\tau,\tau^*)$ given in eq.(23) as:

$$r(\tau,\tau^*) = \frac{g(\tau,\tau^*)}{g(\tau^*,\tau^*)} = \left(\frac{\chi^2\tau^* + \chi^2 + \bar{\tau}^2(\tau^*;\tau_{dm}^{max})}{\chi^2\tau + \chi^2 + \bar{\tau}^2(\tau^*;\tau_{dm}^{max})}\right)^{\frac{1}{2}}. \tag{63}$$

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

$$\frac{\partial \bar{\tau}(\tau^*;\tau_{dm}^{max})}{\partial \tau_{dm}^{max}} = \mu^*\phi^*(\tau_{dm}^{max})\left(\tau_{dm}^{max} - \mathsf{E}_{\phi^*}\left[\tau \,|\tau^* \leq \tau \leq \tau_{dm}^{max}\right]\right) > 0. \tag{64}$$

Thus, $\frac{\partial r}{\partial \tau_{dm}^{max}} > 0$. Using the expression for $F(.)$ in eq.(22), we deduce that:

$$\frac{\partial F}{\partial \tau_{dm}^{max}} = \underbrace{-\phi(\tau_{dm}^{max})(1 - r(\tau_{dm}^{max},\tau^*))}_{<0} + \int_{\tau^*}^{\tau_{dm}^{max}} \underbrace{\frac{\partial r}{\partial \tau_{dm}^{max}}}_{>0}\phi(\tau)d\tau + \int_{\tau^*}^{\tau_{dm}^{max}} \underbrace{(r(\tau,\tau^*)-1)}_{<0}\underbrace{\frac{\partial \phi}{\partial \tau_{dm}^{max}}}_{<0}d\tau. \tag{65}$$

Thus, the effect of τ_{dm}^{max} on $F(.)$ and therefore the equilibrium search intensity τ^* is ambiguous. We now show that this effect becomes positive when τ_{dm}^{max} is large enough. To see this, observe that eq.(65) implies that:

$$\frac{\partial F}{\partial \tau_{dm}^{max}} > \phi(\tau_{dm}^{max})\left(\frac{\int_{\tau^*}^{\tau_{dm}^{max}} \frac{\partial r}{\partial \tau_{dm}^{max}}\phi(\tau)d\tau}{\phi(\tau_{dm}^{max})} - 1\right) = \phi(\tau_{dm}^{max})\left(\frac{\int_{\tau^*}^{\tau_{dm}^{max}} \frac{\partial r}{\partial \tau_{dm}^{max}}\psi(\tau)d\tau}{\psi(\tau_{dm}^{max})} - 1\right) \tag{66}$$

We show in Section I.A of the internet appendix that $\frac{\int_{\tau^*}^{\tau_{dm}^{max}} \frac{\partial r}{\partial \tau_{dm}^{max}}\psi(\tau)d\tau}{\psi(\tau_{dm}^{max})}$ goes to ∞ when τ_{dm}^{max} goes to ∞. Thus, $\frac{\partial F}{\partial \tau_{dm}^{max}} > 0$ for τ_{dm}^{max} large enough. Let τ^{tr} be the smallest value of τ_{dm}^{max} such that $\frac{\partial F}{\partial \tau_{dm}^{max}} > 0$. As in equilibrium, $F(\tau^*) = \exp(-\rho c)$ and $F(.)$ increases in τ^*, it follows that τ^* decreases in τ_{dm}^{max} when $\tau_{dm}^{max} > \tau^{tr}$.

Step 2. When an increase in τ_{dm}^{max} improves τ^*, it raises the average quality of predictors (see Lemma 1) and therefore price informativeness. Now consider the other possible case, i.e., the case in which an increase in τ_{dm}^{max} reduces τ^*. We know from Step 1 that this possibility arises when τ_{dm}^{max} is high enough. We prove below, by contradiction, that price informativeness, $\mathcal{I}(\tau^*;\tau_{dm}^{max})$, is also increasing with τ_{dm}^{max} in this case.

Suppose (to be contradicted) that there is a value of τ_{dm}^{max} such that when $\frac{\partial \tau^*}{\partial \tau_{dm}^{max}} < 0$ then $\frac{\partial \mathcal{I}(\tau^*;\tau_{dm}^{max})}{\partial \tau_{dm}^{max}} < 0$. Let $L(\tau_i^*, \tau^*)$ be:

$$L(\tau_i^*, \tau^*) \equiv \int_{\tau_i^*}^{\tau_{dm}^{max}} \frac{g(\tau, \tau^*)}{g(\tau_i^*, \tau^*)} \phi(\tau) d\tau + 1 - \int_{\tau_i^*}^{\tau_{dm}^{max}} \phi(\tau) d\tau. \tag{67}$$

Function L is increasing with τ_i^* because

$$\frac{\partial L}{\partial \tau_i^*} = \int_{\tau_i^*}^{\tau_{dm}^{max}} \frac{\partial}{\partial \tau_i^*} \left(\frac{g(\tau, \tau^*)}{g(\tau_i^*, \tau^*)} \right) \phi(\tau) d\tau > 0. \tag{68}$$

Now, using the expression for $J(.)$ given in eq.(18), we can rewrite the indifference condition (19) as:

$$L(\tau_i^*, \tau^*) = \exp(-\rho c). \tag{69}$$

Moreover: $L(\tau_{dm}^{max}, \tau^*) = 1$ and $0 < L(0, \tau^*) < 1$. Thus, as $L(\tau_i^*, \tau^*)$ increases in τ_i^* , eq.(67) has a unique solution $\tau_i^*(\tau^*)$ when c is small enough. This solution defines the best response of a data miner when other data miners choose the stopping rule τ^*.

Next, for $\tau_i^* \leq \tau \leq \tau_{dm}^{max}$, define

$$l(\tau, \tau_i^*, \tau^*) = \frac{g(\tau, \tau^*)}{g(\tau_i^*, \tau^*)} = \left(\frac{\chi^2 \tau_i^* + \chi^2 + \bar{\tau}(\tau^*; \tau_{dm}^{max})^2}{\chi^2 \tau + \chi^2 + \bar{\tau}(\tau^*; \tau_{dm}^{max})^2} \right)^{\frac{1}{2}} = \left(\frac{\tau_i^* + \sigma_\omega^2 \mathcal{I}(\tau^*; \tau_{dm}^{max})}{\tau + \sigma_\omega^2 \mathcal{I}(\tau^*; \tau_{dm}^{max})} \right)^{\frac{1}{2}}. \tag{70}$$

Clearly, $l(\tau, \tau_i^*, \tau^*)$ increases with $\mathcal{I}(\tau^*; \tau_{dm}^{max})$. Thus, if $\frac{\partial \mathcal{I}(\tau^*;\tau_{dm}^{max})}{\partial \tau_{dm}^{max}} < 0$, then $\frac{\partial l(\tau, \tau_i^*, \tau^*)}{\partial \tau_{dm}^{max}} < 0$ since τ_{dm}^{max} affects $l(\tau, \tau_i^*, \tau^*)$ only through its effect on price informativeness. This implies that:

$$\frac{\partial l}{\partial \tau_{dm}^{max}} + \frac{\partial l}{\partial \tau^*} \frac{\partial \tau^*}{\partial \tau_{dm}^{max}} < 0. \tag{71}$$

As:

$$L(\tau_i^*, \tau^*) \equiv \int_{\tau_i^*}^{\tau_{dm}^{max}} l(\tau, \tau_i^*, \tau^*) \phi(\tau) d\tau + 1 - \int_{\tau_i^*}^{\tau_{dm}^{max}} \phi(\tau) d\tau, \tag{72}$$

we deduce that:

$$\begin{aligned} \frac{dL}{d\tau_{dm}^{max}} &= \frac{\partial L}{\partial \tau_{dm}^{max}} + \frac{\partial L}{\partial \tau^*} \frac{\partial \tau^*}{\partial \tau_{dm}^{max}} \\ &= \underbrace{-\phi(\tau_{dm}^{max})(1 - l(\tau, \tau_i^*, \tau^*))}_{<0} + \int_{\tau_i^*}^{\tau_{dm}^{max}} \left(\frac{\partial l}{\partial \tau_{dm}^{max}} + \frac{\partial l}{\partial \tau^*} \frac{\partial \tau^*}{\partial \tau_{dm}^{max}} \right) \phi(\tau) d\tau. \end{aligned} \tag{73}$$

Eq.(71) implies that the second term is also negative. Thus, $\frac{dL}{d\tau_{dm}^{max}} < 0$. Thus, an

increase in τ_{dm}^{max} results in a lower value of L, holding τ_i^* constant. As $\partial L/\partial \tau_i^* > 0$ and $L(\tau_i^*, \tau^*) = \exp(-\rho c)$, it follows that in this case τ_i^* increases with τ_{dm}^{max}. As, in equilibrium, $\tau_i^* = \tau^*$, this also implies that $\frac{\partial \tau^*}{\partial \tau_{dm}^{max}} > 0$. A contradiction with our starting hypothesis. We deduce that when $\frac{\partial \tau^*}{\partial \tau_{dm}^{max}} < 0$ then $\frac{\partial \mathcal{I}(\tau^*;\tau_{dm}^{max})}{\partial \tau_{dm}^{max}} > 0$. Thus, for all values of τ_{dm}^{max}, an increase in τ_{dm}^{max} improves price informativeness. This implies that for all values of τ_{dm}^{max}, an increase in τ_{dm}^{max} raises the average quality of all asset managers' signals, $\bar{\tau}(\tau^*, \tau_{dm}^{max})$.

Step 3. Finally, we analyze the effect of an increase in τ_{dm}^{max} on the average quality of experts' signals on the one hand and data miners' signals on the other hand. Clearly, when data abundance raises τ^*, it increases both $\bar{\tau}_{dm}(\tau^*)$ and $\bar{\tau}_{ex}(\tau^*)$. Now consider the case in which data abundance reduces τ^*. In this case, it clearly reduces $\bar{\tau}_{ex}(\tau^*)$ because $\bar{\tau}_{ex} = \mathsf{E}_\gamma [\tau \,|\tau^* \leq \tau]$. However, it still raises the average quality of data miners' signals. To see this, we use differentiate the expression for $\bar{\tau}(\tau^*, \tau_{dm}^{max})$ given in eq.(49) to obtain:

$$\frac{d\bar{\tau}(\tau^*, \tau_{dm}^{max})}{d\tau_{dm}^{max}} = \Gamma(\tau^*) \frac{dE_\phi [\tau \,|\tau^* \leq \tau \leq \tau_{dm}^{max}]}{d\tau_{dm}^{max}} + \frac{\partial \tau^*}{\partial \tau_{dm}^{max}} \gamma(\tau^*) \left(\underbrace{E_\phi [\tau \,|\tau^* \leq \tau \leq \tau_{dm}^{max}] - \tau^*}_{>0} \right). \tag{74}$$

We know from Step 2 that $\frac{d\bar{\tau}(\tau^*, \tau_{dm}^{max})}{d\tau_{dm}^{max}} > 0$ for all values of τ_{dm}^{max}. Thus, when $\frac{\partial \tau^*}{\partial \tau_{dm}^{max}} < 0$, it must be that $\frac{d\bar{\tau}_{dm}}{d\tau_{dm}^{max}} = dE_\phi [\tau \,|\tau^* \leq \tau \leq \tau_{dm}^{max}] / d\tau_{dm}^{max} > 0$.

Proof of Proposition 6. It follows from the expression for $r(\tau, \tau^*)$ given in eq.(63) that $r(\tau, \tau^*)$ decreases with σ_ω^2, and σ_η^2 because $\tau > \tau^*$. Thus, from eq.(22), we deduce that $F(\tau^*)$ decreases with σ_ω^2, and σ_η^2. It follows from (i) this observation, (ii) the fact that $F(\tau^*)$ increases with τ^* and (iii) the equilibrium condition $F(\tau^*) = \exp(-\rho c)$ that τ^* and therefore μ^* (since $\mu^* = \Gamma(\tau^*)$ increase with σ_ω^2 and σ_η^2. Moreover, we deduce from Lemma 1 that the average quality of asset managers' signals ($\bar{\tau}$) increases with σ_ω^2 and σ_η^2. This is also the case for the average quality of data miners' and experts' signals ($\bar{\tau}_{dm}$ and $\bar{\tau}_{ex}$) since they both increase in τ^* and depend on σ_ω^2, and σ_η^2 only via τ^* (see eq.(61) and eq.(62)). Last, consider price informativeness. Using eq.(14), one can see that the direct effect of an increase in σ_ω^2 and σ_η^2 on price informativeness is negative. However, there is an indirect effect: it raises asset managers' average signals ($\bar{\tau}$), which affects positively price informativeness. The net effect of an increase in σ_ω^2 and σ_η^2 on price informativeness is therefore ambiguous. However, one can prove by contradiction that it must be positive.

To see why, suppose that there is a value of σ_ω or σ_η for which a marginal increase in either one of these two variables increases price informativeness $\mathcal{I}$. Then, one can see from the expression of $r(\tau, \tau^*)$ in eq.(23) that for this value, a marginal increase in either one of these two variables reduces $r(\tau, \tau^*)$. However, this is impossible since, using the expression of $r(\tau, \tau^*)$ in (63), we have shown that $r(\tau, \tau^*)$ necessarily increases with σ_ω or σ_η. We deduce that for all values of σ_ω or σ_η, an increase in either one of these two variables reduces price informativeness.

Proof of Corollary 1. Consider the effect of τ_{dm}^{max} on asset managers' expected profits. We know from Proposition 5 that $\bar{\tau}(\tau^*(\tau_{dm}^{max}, c), \tau_{dm}^{max})$ increases with τ_{dm}^{max}. Moreover, $\lim_{\tau_{dm}^{max} \to 0} \bar{\tau}(\tau^*(\tau_{dm}^{max}, c), \tau_{dm}^{max}) = \bar{\tau}_0$. Thus, if $\bar{\tau}(\tau^*(\infty, c), \infty) > \rho\sigma_\omega\sigma_\eta > \bar{\tau}_0$, there is a unique value of τ, denoted $\hat{\tau}$, such that $\bar{\tau}(\tau^*(\hat{\tau}, c), \hat{\tau}) = \rho\sigma_\omega\sigma_\eta$. Consequently, when τ_{dm}^{max} varies, holding other parameters constant, asset managers' expected profit reaches its maximum for $\bar{\tau}(\tau^*(\hat{\tau}, c), \hat{\tau}) = \rho\sigma_\omega\sigma_\eta$. If $\bar{\tau}(\tau^*(\infty, c), \infty) \leq \rho\sigma_\omega\sigma_\eta$, then asset managers' expected profit always increases as τ_{dm}^{max} increases. Finally, if $\bar{\tau}_0 \geq \rho\sigma_\omega\sigma_\eta$, then asset managers' expected profit always increases as τ_{dm}^{max} decreases. This proves Part 2 of Corollary 1. The proofs of Part 1 is similar and therefore omitted for brevity. In this case, one obtains that $\hat{c}$ is the unique solution of $\bar{\tau}(\tau^*(\tau_{dm}^{max}, \hat{c}), \tau_{dm}^{max}) = \rho\sigma_\omega\sigma_\eta$.

Proof of Corollary 5. Notice that under the assumption $\Gamma(\tau) = \Psi(\tau)/\Psi(\tau_{ex}^{max})$, which the distribution Ψ conditional on $\tau \leq \tau_{ex}^{max}$, then the equilibrium distribution of experts is distribution Ψ conditional on $\tau^* \leq \tau \leq \tau_{ex}^{max}$, and therefore

$$\bar{\tau}_{ex} = \frac{\int_{\tau^*}^{\tau_{ex}^{max}} \tau\psi(\tau)d\tau}{\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)}. \tag{75}$$

Similarly, we still have

$$\bar{\tau}_{dm} = \frac{\int_{\tau^*}^{\tau_{dm}^{max}} \tau\psi(\tau)d\tau}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)}. \tag{76}$$

To assess the effect of a reduction of c on RP, one can equivalently compute the derivative

of $\log(RP)$ with respect to τ^* which yields

$$\begin{aligned}\frac{\partial \log(RP)}{\partial \tau^*} &= \frac{1}{\bar{\tau}_{dm}}\frac{\partial \bar{\tau}_{dm}}{\partial \tau^*} - \frac{1}{\bar{\tau}_{ex}}\frac{\partial \bar{\tau}_{ex}}{\partial \tau^*} \\ &= \frac{\psi(\tau^*)}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)}\frac{\bar{\tau}_{dm} - \tau^*}{\bar{\tau}_{dm}} - \frac{\psi(\tau^*)}{\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)}\frac{\bar{\tau}_{ex} - \tau^*}{\bar{\tau}_{ex}}.\end{aligned} \tag{77}$$

Next, compute the following derivative,

$$\begin{aligned}&\frac{\partial}{\partial \tau_{ex}^{max}}\left[\frac{\psi(\tau^*)}{\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)}\left(1 - \frac{\tau^*}{\bar{\tau}_{ex}}\right)\right] \\ =& -\frac{\psi(\tau^*)\psi(\tau_{ex}^{max})}{\left(\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)\right)^2}\left(1 - \frac{\tau^*}{\bar{\tau}_{ex}}\right) + \frac{\psi(\tau^*)\psi(\tau_{ex}^{max})}{\left(\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)\right)^2}(\tau_{ex}^{max} - \bar{\tau}_{ex})\frac{\tau^*}{\bar{\tau}_{ex}^2} \\ =& \frac{\psi(\tau^*)\psi(\tau_{ex}^{max})}{\left(\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)\right)^2}\left(\frac{\tau_{ex}^{max}\tau^*}{\bar{\tau}_{ex}^2} - 1\right)\end{aligned} \tag{78}$$

When $\tau_{ex}^{max} = \infty$, then $\bar{\tau}_{ex}$ is finite and τ^* is strictly larger than 0, then for τ_{ex}^{max} large enough $\tau_{ex}^{max}\tau^* > \bar{\tau}_{ex}^2$. For the same reason, for τ_{dm}^{max} large enough $\tau_{dm}^{max}\tau^* > \bar{\tau}_{dm}^2$. Then, both $\frac{\psi(\tau^*)}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)}\frac{\bar{\tau}_{dm} - \tau^*}{\bar{\tau}_{dm}}$ and $\frac{\psi(\tau^*)}{\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)}\frac{\bar{\tau}_{ex} - \tau^*}{\bar{\tau}_{ex}}$ are increasing respectively in τ_{dm}^{max} and τ_{ex}^{max}, for τ_{dm}^{max} and τ_{ex}^{max} large enough. As $\tau_{dm}^{max} < \tau_{ex}^{max}$, then

$$\frac{\psi(\tau^*)}{\Psi(\tau_{dm}^{max}) - \Psi(\tau^*)}\frac{\bar{\tau}_{dm} - \tau^*}{\bar{\tau}_{dm}} < \frac{\psi(\tau^*)}{\Psi(\tau_{ex}^{max}) - \Psi(\tau^*)}\frac{\bar{\tau}_{ex} - \tau^*}{\bar{\tau}_{ex}} \;\Rightarrow\; \frac{\partial \log(RP)}{\partial \tau^*} < 0. \tag{79}$$

Proof of Corollary 2. Direct from the arguments in the text.

Proof of Corollary 3. For $\alpha \in [0,1]$, the a-quantile of type $j \in \{dm, ex\}$ asset managers expected profit distribution is the profit level $\bar{\Pi}_\alpha^j$ such that a mass α of type j asset managers has a profit lower than $\bar{\Pi}_\alpha^j$. One can immediately observe that equivalently a mass α of type j asset managers has a quality lower than τ_α^j with τ_α^j such that $\bar{\Pi}_\alpha^j = \bar{\Pi}(\tau_\alpha^j)$. Thus, we have

$$\frac{H_j(\tau_\alpha^j) - H_j(\tau^*)}{1 - H_j(\tau^*)} = \alpha \iff \tau_\alpha^j = H_j^{-1}\left[\alpha + (1-\alpha)H_j(\tau^*)\right]. \tag{80}$$

with $H_{dm}(.) = \Phi(.)$ and $H_{ex}(.) = \Gamma(.)$. For $\alpha > 0.5$ high enough, one introduce the log-difference between α- and $1-\alpha$-quantiles of type j asset managers expected profits,

ΔQ^j_α, as a performance dispersion measure, that is

$$\Delta Q^j_\alpha = \log(\bar{\Pi}^j_\alpha) - \log(\bar{\Pi}^j_{1-\alpha}) = \log(\tau^j_\alpha) - \log(\tau^j_{1-\alpha}). \tag{81}$$

Compute the derivative of τ^{dm}_α with respect to τ^* as

$$\frac{\partial \tau^{dm}_\alpha}{\partial \tau^*} = \frac{(1-\alpha)\phi(\tau^*)}{\phi(\tau^{dm}_\alpha)} \tag{82}$$

The effect of c on ΔQ^{dm}_α is measured by

$$\frac{d\Delta Q^{dm}_\alpha}{dc} = \left[\frac{(1-\alpha)\phi(\tau^*)}{\tau^{dm}_\alpha \phi(\tau^{dm}_\alpha)} - \frac{\alpha\phi(\tau^*)}{\tau^{dm}_{1-\alpha}\phi(\tau^{dm}_{1-\alpha})}\right]\frac{\partial \tau^*}{\partial c} \tag{83}$$

We want to show that for a large enough we have $d\Delta Q^{dm}_\alpha/dc > 0$, or equivalently that

$$\frac{(1-\alpha)}{\tau^{dm}_\alpha \phi(\tau^{dm}_\alpha)} - \frac{\alpha}{\tau^{dm}_{1-\alpha}\phi(\tau^{dm}_{1-\alpha})} < 0. \tag{84}$$

As τ^{max}_{dm} is finite, one can see that when α goes to 1, and as τ^{dm}_α goes to τ^{max}_{dm} and $\tau^{dm}_{1-\alpha}$ goes to τ^*, then the above expression goes towards $-1/(\tau^*\phi(\tau^*)) < 0$. So for α high enough, the above condition is verified.

Similarly for ΔQ^{ex}_α, we have that $d\Delta Q^{ex}_\alpha/dc > 0$, and also $d\Delta Q^{ex}_\alpha/d\tau^{max}_{dm} > 0$ when τ^* decreases with τ^{max}_{dm} (because the effect of τ^{max}_{dm} is through τ^*), if

$$\frac{(1-\alpha)}{\tau^{ex}_\alpha \gamma(\tau^{ex}_\alpha)} - \frac{\alpha}{\tau^{ex}_{1-\alpha}\gamma(\tau^{ex}_{1-\alpha})} < 0. \tag{85}$$

When $\tau^{max}_{ex} < \infty$, and as for ΔQ^{ex}_α, one can see that for α high enough, the above condition is verified. When $\tau^{max}_{ex} = \infty$, as $\alpha = (\Gamma(\tau^{ex}_\alpha) - \Gamma(\tau^*))/(1-\Gamma(\tau^*))$, the above condition holds for some large α if and only if

$$\lim_{\tau\to\infty} \frac{1-\Gamma(\tau)}{\tau\gamma(\tau)} < \frac{1-\Gamma(\tau^*)}{\tau^*\gamma(\tau^*)} \tag{86}$$

Notice that this condition is verified in the case where the hazard rate $\gamma(\tau)/(1-\Gamma(\tau))$ multiplied by τ is increasing. This is the case for instance when $\Gamma(\tau) = 1 - 1/(1+\tau)^k$, or $\Gamma(\tau) = 1 - \exp(-k\tau)$, with $k > 0$.

To assess the effect of τ_{dm}^{max} on ΔQ_{α}^{dm}, one must notice that

$$\frac{\Phi(\tau_{\alpha}^{dm}) - \Phi(\tau^*)}{1 - \Phi(\tau^*)} = \frac{\Psi(\tau_{\alpha}^{dm}) - \Psi(\tau^*)}{\Psi(\tau_{dm}^{max}) - \Phi(\tau^*)} = \alpha \iff \tau_{\alpha}^{dm} = \Psi^{-1}\left[\alpha\Psi(\tau_{dm}^{max}) + (1-\alpha)\Psi(\tau^*)\right]. \tag{87}$$

Thus we have

$$\frac{d\Delta Q_a^{data}}{d\tau_{dm}^{max}} = \frac{\alpha\psi(\tau_{dm}^{max})}{\tau_{\alpha}^{dm}\psi(\tau_{\alpha}^{dm})} - \frac{(1-\alpha)\psi(\tau_{dm}^{max})}{\tau_{1-\alpha}^{dm}\psi(\tau_{1-\alpha}^{dm})} + \left[\frac{(1-\alpha)\psi(\tau^*)}{\tau_{\alpha}^{dm}\psi(\tau_{\alpha}^{dm})} - \frac{\alpha\psi(\tau^*)}{\tau_{1-\alpha}^{dm}\psi(\tau_{1-\alpha}^{dm})}\right]\frac{\partial\tau^*}{\partial\tau_{dm}^{max}} \tag{88}$$

We want to show that for α large enough an increase in $d\Delta Q_{\alpha}^{dm}/d\tau_{dm}^{max} > 0$, when $\partial\tau^*/\partial\tau_{dm}^{max} < 0$. Notice that when the previous condition is verified we have

$$\frac{(1-\alpha)\psi(\tau^*)}{\tau_{\alpha}^{dm}\psi(\tau_{\alpha}^{dm})} - \frac{\alpha\psi(\tau^*)}{\tau_{1-\alpha}^{dm}\psi(\tau_{1-\alpha}^{dm})} = \frac{(1-\alpha)\phi(\tau^*)}{\tau_{\alpha}^{dm}\phi(\tau_{\alpha}^{dm})} - \frac{\alpha\phi(\tau^*)}{\tau_{1-\alpha}^{dm}\phi(\tau_{1-\alpha}^{dm})} < 0. \tag{89}$$

In addition we want to show that the direct impact of τ_{dm}^{max} on ΔQ_{α}^{dm} is positive, that is

$$\frac{\alpha}{\tau_{\alpha}^{dm}\psi(\tau_{\alpha}^{dm})} - \frac{(1-\alpha)}{\tau_{1-\alpha}^{dm}\psi(\tau_{1-\alpha}^{dm})} > 0. \tag{90}$$

By assumption, the unconditional distribution ψ (over $[0,\infty)$) has a finite first-order moment. It implies first that $\tau\psi(\tau)$ goes to zero when τ goes to ∞. So for τ_{dm}^{max} large enough $\tau_{dm}^{max}\psi(\tau_{dm}^{max}) < \tau^*\psi(\tau^*)$. And then, for α large enough, the above expression is indeed positive.

Proof of Corollary 6 Using the expressions for $w(\tau)$ and $g(\tau,\tau^*)$ in eq.(16), we obtain from eq.(40) that:

$$f_{ex}^*(\tau) = \frac{\kappa}{\rho}\left[\log\left(\frac{1}{r(\tau,\tau^*)}\right)\right]. \tag{91}$$

where $r(\tau,\tau^*)$ is defined in eq.(23). Using the expression for $r(\tau,\tau^*)$ given in eq.(63), it is immediate to obtain that a decrease in c reduces $f_{ex}^*(\tau)$ because such an increase (i) raises τ^* and (ii) raises $\bar{\tau}$ (Proposition 4). The same argument implies that an increase in τ_{dm}^{max} reduces $f_{ex}^*(\tau)$ when τ_{dm}^{max} raises τ^*. To see why the effect is unclear when τ_{dm}^{max}

lowers τ^*, one must take the first-order derivative of $f^*_{ex}(\tau)$ with respect to τ^{max}_{dm},

$$\frac{\partial f^*_{ex}}{\partial \tau^{max}_{dm}} = \frac{\kappa}{2\rho}\left[2\bar{\tau}(\tau^*;\tau^{max}_{dm})\frac{d\bar{\tau}}{d\tau^{max}_{dm}}\left(\frac{1}{\rho^2\sigma^2_\omega\sigma^2_\eta(1+\tau)+\bar{\tau}(\tau^*;\tau^{max}_{dm})^2}-\frac{1}{\rho^2\sigma^2_\omega\sigma^2_\eta(1+\tau^*)+\bar{\tau}(\tau^*;\tau^{max}_{dm})^2}\right)\right.$$
$$\left.-\frac{\partial\tau^*}{\partial\tau^{max}_{dm}}\frac{\rho^2\sigma^2_\omega\sigma^2_\eta}{\rho^2\sigma^2_\omega\sigma^2_\eta(1+\tau^*)+\bar{\tau}(\tau^*;\tau^{max}_{dm})^2}\right], \tag{92}$$

for which the sign is the same as

$$-2\bar{\tau}(\tau^*;\tau^{max}_{dm})\frac{d\bar{\tau}}{d\tau^{max}_{dm}}(\tau-\tau^*)-\frac{\partial\tau^*}{\partial\tau^{max}_{dm}}(\rho^2\sigma^2_\omega\sigma^2_\eta(1+\tau)+\bar{\tau}(\tau^*;\tau^{max}_{dm})^2). \tag{93}$$

When $\tau=\tau^*$, the above expression is positive, as $-\frac{\partial\tau^*}{\partial\tau^{max}_{dm}}>0$. It remains positive for any τ if and only if $-2\bar{\tau}(\tau^*;\tau^{max}_{dm})\frac{d\bar{\tau}}{d\tau^{max}_{dm}}-\rho^2\sigma^2_\omega\sigma^2_\eta\frac{\partial\tau^*}{\partial\tau^{max}_{dm}}>0$. Hence, in the reverse case, it becomes negative for τ large enough.