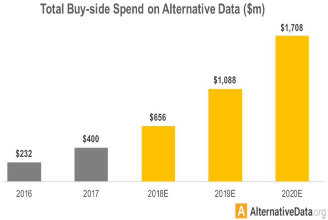
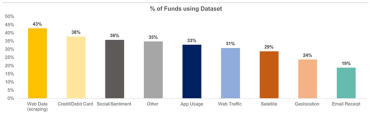
Equilibrium Data Mining and Data Abundance

Jérôme Dugast (Paris Dauphine University)
and
Thierry Foucault (HEC)

Market Microstructure Exchange
April 14, 2020

Data Abundance and Investment



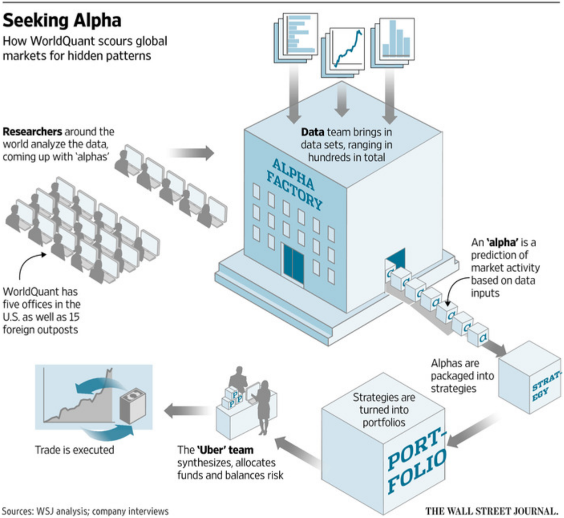


"[Stock pickers] are turning to some of the data-mining techniques pioneered by their "quant" rivals and are investing heavily in programmers and data scientists. They are hoping that a hybrid approach, which combines the judgment of an experienced stockpicker with the insights that big data can offer, will give them a new lease of life" (in "Stock pickers turn to big data to arrest decline" in Financial Times, February 11, 2020)

Abundance of Predictors

The alpha factory [Worldquant] breaks the process of investing into a quantitative trading assembly line. The inputs are data acquired by a special group that scours the globe for interesting and new data sets [...] Researchers around the world attack the data with computing power and mathematical techniques to find patterns [...] They test them intensively [...] The company has "4 millions alphas" to date and is aiming for 100 millions. Each alpha at WorldQuant is an algorithm that seeks to profit by predicting some future change in the price of a stock, futures or other assets. (Wall Street Journal, April 13, 2017)

Industrialization of the search for Predictors



Abundance of Predictors in Aacademic Reseach

- **Yan and Zheng (2017, RFS):** $18,000$ **signals based on combinations of 240 accounting variables from financial statements. Many have predictive power for returns (e.g., more than 500 have highly significant alpha (**$t - stat >\mid 5 \mid$**).**

- **Factor Zoo: The number of factors discovered by academics grow exponentially (see Harvey and Liu (2019): "A census of the factor zoo").**

Data Abundance and Computing Power

- **Two distinct drivers of this evolution:**

  1. **More data:** Data abundance: the search space for predictors get bigger over time (e.g., one can use financial statements and satellite images of retailers' parking lots to forecast firms' earnings (see Katona et al.(2019)).

  2. **More computing power** (it is less time consuming to explore a given set of data): Faster computers and growth in memory capacity.

Computing Power

Two Centuries of Progress 143

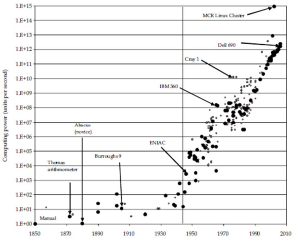


FIGURE 2
THE PROGRESS OF COMPUTING POWER MEASURED IN COMPUTATIONS PER SECOND (CPS)

Source: Nordhaus (2007): «Two centuries of productivity growth in computing », Journal of Economic History

"We apply machine learning and big data analysis to financial economics. The algorithm requires a vast amount of computational power. The average time needed to find the optimum trading rules for a diversified portfolio of ten NYSE/AMEX volatility assets for the 40 year sample using a computer with an Intel® Core(TM) CPU i7-2600 and 16 GM RAM is 459.29 days (11,022.97 hours). For one year it takes approximately 11.48 days."

Source: Brogaard and Zareei (2019)

Questions

- **Are the effects of greater computing power and data abundance on financial markets similar or different? Can we just think of more data as lowering the cost of acquiring information in standard models?**

- **How do speculators optimally search for predictors? Data mining in equilibrium**

- **How does progress in information technologies affect (i) the diversity of predictors, (ii) trading profits, (iii) crowding, (iv) asset price informativeness etc?**

Our Contribution

- **We develop a new model of information acquisition in financial markets which addresses the previous questions**
- **We explicitly formalize the optimal search for predictors by investors and thereby provide a micro-foundation for the cost of acquiring predictors of a given average precision.**
- **The model allows to analyze separately the effects of data abundance from those of computing power.**
- **Main message: Data abundance and computing power do not have the same effects on the equilibrium of the financial market (e.g., asset price informativeness).**

Literature

- **Models of information acquisition**
    1. All investors receive the same signal of a fixed precision (**Grossman and Stiglitz (1980, AER)**) or all investors choose signals of the same precision in equilibrium if investors are identical (**Verrecchia (1982, Eca)**)
    2. In our case, investors are identical and choose signals of different precisions in equilibrium (the distribution of precisions is an equilibrium outcome).
    3. The cost of acquiring a signal with a given minimal precision is endogenous in our model (micro-founded).
- **Effects of big data on financial markets**
    1. **Veldkamp and Farboodi (2019, AER):** Technological progress in information analysis $\Leftrightarrow$ Lower cost of acquiring information.
    2. **Dugast and Foucault (2018, JFE):** Technological progress in information analysis means that signals can be made available faster but fast signals are less precise.
- Our approach is different: It allows to separate the effects of a reduction in the cost of processing data from the effect of increasing the universe of data available to search trading signals.

QUESTIONS?

MODEL

Model

- **A risky asset with payoff:** $\omega \sim \mathcal{N}(0, \tau_\omega^{-1})$
- **A continuum of risk averse (CARA) speculators**
- **Noise Traders**
- **Risk neutral competitive dealers**
- **Timing:**
  1. **Date 0: Speculators search for predictors of the payoff of the risky asset.**
  2. **Date 1: Speculators, noise traders and dealers trade**

Timing

Period 0

Data Mining:

- **Each speculator searches a predictor of the asset payoff .**

Period 1

Trading:

- **Each speculator observes the realization of her predictor (s_θ) and chooses a trading strategy, x(s_θ ,p).**
- **Market clears: The asset price is realized**

Period 2

Asset Payoff, ω, is realized

Predictors

- **There is a continuum of "predictors" of the asset payoff.** Each predictor s_θ is characterized by its type, θ such that:

$$s_\theta = \underbrace{cos(\theta)\,\omega}_{\text{Signal}} + \underbrace{sin(\theta)\epsilon_\theta}_{\text{Noise}} \tag{1}$$

  where $\epsilon_\theta \sim \mathcal{N}(0, \tau_\omega^{-1})$, $\epsilon_\theta \perp \omega$, the ϵ_θs are i.i.d. and $\theta \in [0, \frac{\pi}{2}]$.

- **The signal-to-noise ratio of a predictor, $\tau(\theta)$, decreases with θ on $[0, \infty)$:**

$$\tau(\theta) = (\frac{\cos(\theta)}{\sin(\theta)})^2 = \text{"}\textbf{quality}\text{"}.$$

- **Assumption:** Predictors' types are distributed according to $\phi(\theta)$ on $\left[0, \frac{\pi}{2}\right]$.

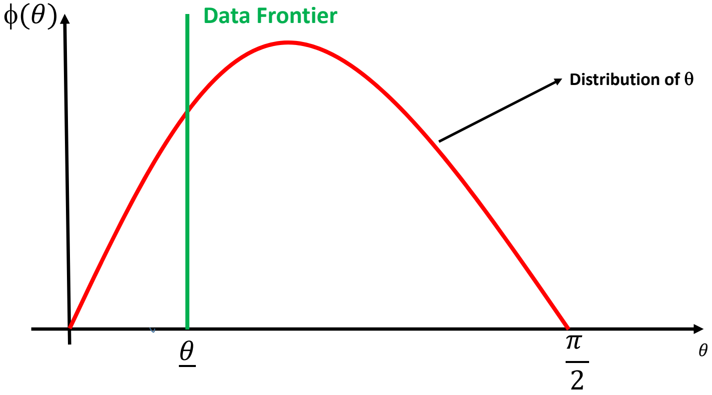
Data Mining (period 0)

- **Information acquisition: A sequential search process with multiple rounds:**

  1. **Each round is a new exploration of data** to find a predictor (e.g., a new dataset).
     1.1 **With probability** $\alpha(1-\phi(\underline{\theta}))$, the exploration is **successful** and returns a predictor of type θ in $[\underline{\theta}, \frac{\pi}{2}]$ with probability $\phi(\theta)$.
     1.2 Otherwise, the exploration is unsuccessful (returns no predictor).
  2. **After observing the outcome of an exploration**, the speculator decides to either (i) stop searching and trade on her latest predictor or (ii) to explore (mine) the data further (move to another search round).
  3. **Each exploration costs** c**.**

- **Assumptions:** No limit on the number of explorations (the search problem is stationary) + No Recall (Results are identical with recall).

Predictors

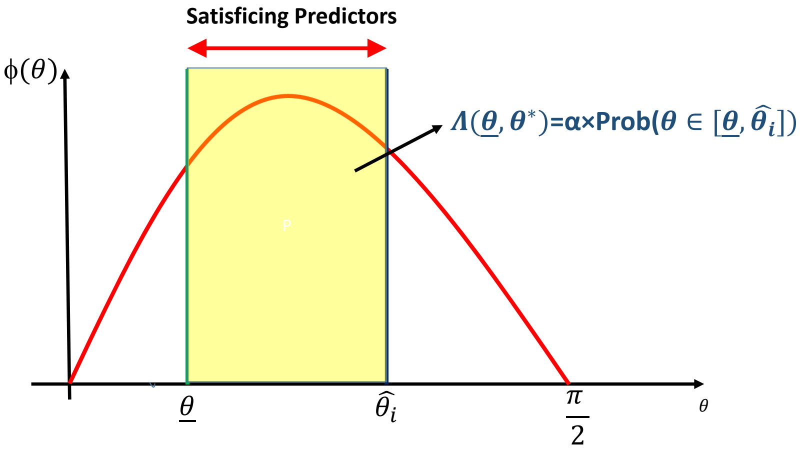


$\phi(\theta)$
Data Frontier
Distribution of θ
$\underline{\theta}$
$\frac{\pi}{2}$
θ

Optimal Data Mining

- **We consider equilibria in which speculators search for predictors using a "stopping rule strategy".** Namely, in a given search round, a speculator i:
  1. Accepts the predictor found in this round if $\theta \leq \hat{\theta}_i$
  2. Mines another round if $\theta > \hat{\theta}_i$
- **Speculator i's stopping rule:** $\hat{\theta}_i$

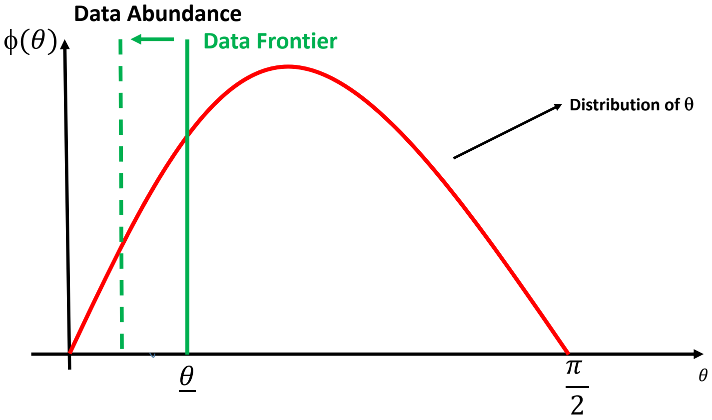
Optimal Data Mining



Interpretation

- **Greater computing power enables speculators to process data at a lower cost $\rightarrow$ c decreases.**
- **Data Abundance:**
  1. **It pushes back the data frontier ("Hidden Gold Nugget Effect):** The quality of the best predictor increases $\rightarrow$ $\underline{\theta}$ decreases.
  2. **It reduces the fraction of informative datasets ("Needle in the haystack effect")** $\rightarrow$ α decreases.
- **Three parameters to capture the effects of progress in information technologies:** c, $\underline{\theta}$, α.

Data Abundance



Trading (period 1)

- **Speculators observe the realization of their predictors (s_θ) and trade on them at date 1.**
- **Trading is modeled as in Vives (1995)**
  1. Each speculator optimally chooses her position $x_i(p, s_\theta)$ given the asset price and the signal.
  2. Noise traders' demand is η, where $\eta \sim \mathcal{N}(0, \nu^2)$
  3. The asset price is set by risk neutral competitive market makers, i.e., such that:
     $$p^* = \mathbb{E}\left[\omega \,|\, D(p^*)\right],$$
     where $D(p)$ is **the aggregate demand from speculators and noise traders.**

Speculators' Objective function

- **Let n_i be the realized number of search rounds for speculator i.**
- **A speculator chooses her search strategy, $\hat{\theta}_i$ and her trading srategy, $x_i^*(s_\theta, p)$ to maximize her expected utility:**

$$\mathbb{E}\left[-\exp(-\rho W_i)\right] =$$

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

- **Key difference with the traditional model: n_i is random and its distribution is endogenous.**

QUESTIONS?

SOLVING FOR THE EQUILIBRIUM

Equilibrium

- **We focus on symmetric equilibria in which speculators all use the same stopping rule: $\hat{\theta}_i = \theta^*$.**

  1. **Same stopping rule ex-ante but different predictors ex-post because the outcome of the search for predictors is random.**
  2. **The ex-post distribution of predictors across speculators is endogenous.** We denote the average quality of predictors used in equilibrium by $\bar{\tau}(\theta^*; \underline{\theta}, c)$.

- **We solve for the equilibrium backward in two steps.**

  1. Solve for the equilibrium of the market for the risky asset given the distribution of predictors used in equilibrium (i.e., for a given θ^*)
  2. Solve for the equilibrium of the search stage, i.e., θ^*.

Equilibrium of the Trading Stage

- **The analysis of the trading stage is standard.**
  1. A speculator with a more precise signal trades more aggressively on his information (i.e., takes a larger position for a given difference between his expectation of the asset payoff and the equilibrium price).

  2. The equilibrium price reflects the information about the asset payoff contained in investors' aggregate demand.
- **The informativeness of the equilibrium price is**:

$$\mathcal{I}(\theta^*; \underline{\theta}, c, \alpha) \equiv \mathbb{V}[\omega \mid p^*]^{-1} = \tau_\omega + \frac{\bar{\tau}(\theta^*; \underline{\theta}, c, \alpha)^2)\tau_\omega^2}{\rho^2 \nu^2},$$

- The equilibrium price is more informative (i.e., "closer" to the asset payoff) if the average quality of predictors, $\bar{\tau}(\theta^*, \underline{\theta})$, is higher ("competition effect").

- **Price informativeness depends on the equilibrium search strategy of speculators:** If they mine the data less intensively (i.e., θ^* increases) then price informativess drops.

Expected Utility from Trading on a Predictor

- Suppose that a speculator finds a predictor with type θ. **Her expected utility from trading** on this predictor is:

$$h(\theta, \theta^*) = -\left(1 + \frac{\tau(\theta)\tau_\omega}{\mathcal{I}(\theta^*; c, \underline{\theta}, \alpha)}\right)^{-\frac{1}{2}}. \qquad (2)$$

- **Thus, other things equal, a speculator's expected utility from trading:**

  1. Increases with the quality of her predictor ($\tau(\theta)$)

  2. Decreases with price informativeness, i.e., the average quality of predictors used in equilibrium.

- $\Rightarrow$ **Search decisions are interdependent.** Other things equal, a speculator has less incentive to search if she expects others to search intensively.

Equilibrium of the Search Stage

- **Let $\hat{\theta}_i$ (possibly different from θ^*) be the stopping rule of speculator i.**
- **If a speculator "rejects" a predictor in a given round, her expected continuation value is:**

$$J(\widehat{\theta}_i, \theta^*; \underline{\theta}, c) = \exp(\rho c) \times$$

$$\left(\Lambda(\widehat{\theta}_i; \underline{\theta}, c) \underbrace{\mathbb{E}\left[h(\theta, \theta^*) | \underline{\theta} \leq \theta \leq \theta^* \right]}_{\text{Expected utility from trading}} + (1 - \Lambda(\widehat{\theta}_i; \underline{\theta}, c)) J(\widehat{\theta}_i, \theta^*; \underline{\theta}, c, \alpha) \right)$$

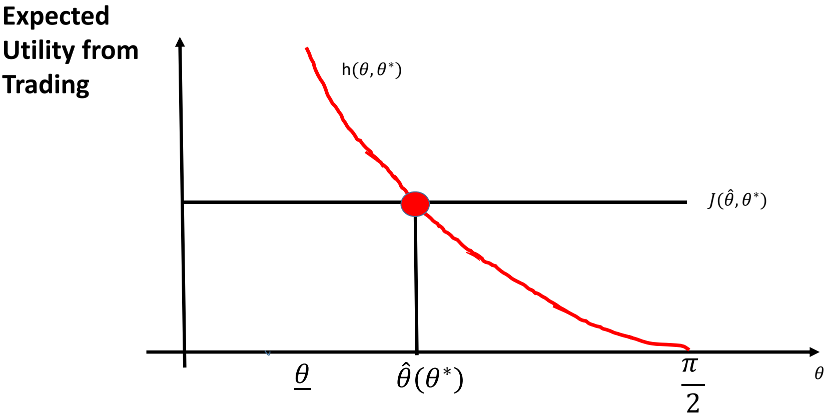
- **Thus:**

$$J(\widehat{\theta}_i, \theta^*; \underline{\theta}, c) = \underbrace{\left[\frac{\exp(\rho c) \Lambda(\widehat{\theta}; \underline{\theta}, c)}{1 - \exp(\rho c)(1 - \Lambda(\widehat{\theta}; \underline{\theta}, c))} \right]}_{\text{Expected Cost of Search}} \underbrace{\mathbb{E}\left[h(\theta, \theta^*) | \underline{\theta} \leq \theta \leq \hat{\theta} \right]}_{\text{Expected utility from trading}}.$$

Equilibrium of the Search Stage

- **After discovering a predictor of type θ, a speculator can:**
  1. Stop searching and obtain her expected utility from trading with this predictor, $h(\theta, \theta^*)$
  2. Or keep searching and obtain the continuation value of searching: $J(\widehat{\theta}, \theta^*; \underline{\theta}, c, \alpha)$.
- **The speculator keeps searching if $h(\theta, \theta^*) < J(\widehat{\theta}, \theta^*; \underline{\theta}, c, \alpha)$ and stops searching otherwise.**

Equilibrium Data Mining



Equilibrium Data Mining

- In equilibrium, a speculator's best response $\hat{\theta}(\theta^*)$ must be θ^*: $\theta^* = \hat{\theta}(\theta^*)$. That is, θ^* solves:

$$h(\theta^*, \theta^*) = J(\theta^*, \theta^*; \underline{\theta}, c, \alpha).$$

- **Result 1: There is a unique symmetric equilibrium of the search stage in which all speculators are active if and only if $c < c^*$.**

- If $c > c^*$, there exist an equilibrium in which (i) not all (or none) speculators search and (ii) $\theta^* = \frac{\pi}{2}$.

- **We focus on $c < c^*$: In equilibrium, the quality of speculators' predictors is distributed over $[\tau(\theta^*), \tau(\underline{\theta})]$.**

IMPLICATIONS

Data Abundance and Search for Predictors 1/2

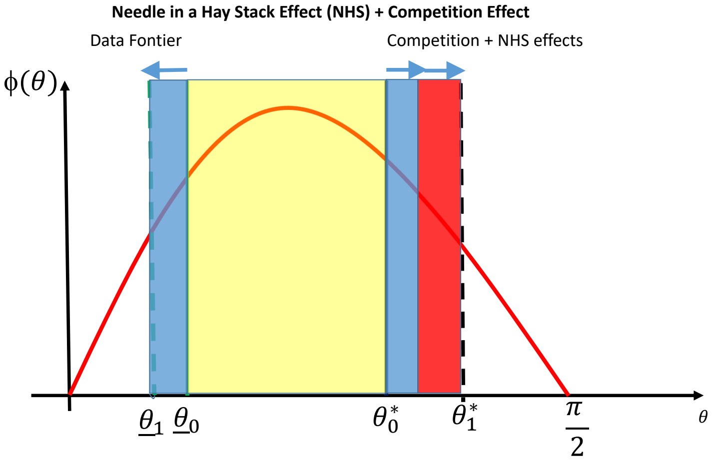
- **Result 2: The effect of data abundance and computing power on speculators' search strategy ("stopping rule") are not the same:**

  1. **Greater computing power** (lower c) always induces speculators to be **more** demanding for the quality of their predictors.

  2. **A push back of the data frontier** (lower $\underline{\theta}$) can induce speculators to be **less** demanding for the quality of their predictors for $\underline{\theta}$ low enough).

  3. **The needle in the haystack problem** (lower α) always induce speculators to be **less** demanding for the quality of their predictors.

Data Abundance and The distribution of Predictors' quality

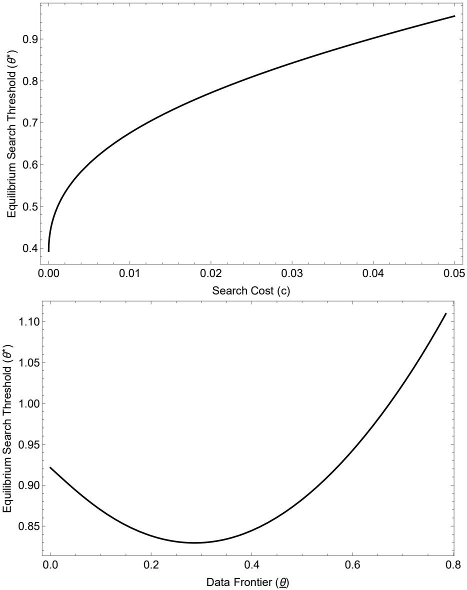


Data Abundance and Search for Predictors 2/2

- **Economic Mechanisms:**

  1. **If the cost of exploration (c) decreases:** The continuation value of searching increases, holding θ^* constant $\Rightarrow$ Speculators are more demanding (search more) in equilibrium (θ^* declines).
  2. **If the quality of the most informative predictor increases:**
     - 2.1 The expected utility of trading for the speculator who finds the best predictor increases ("hidden gold nugget effect"))
     - 2.2 The expected utility of trading for all others speculators decreases (competition effect)
     - 2.3 $\Rightarrow$ **The net effect on the continuation value of searching is ambiguous.**
  3. **Needle in the haystack problem:** Same as an increase in c.

Illustration



Data Abundance and Stock Picking Skills

- **One can measure the quality of predictors used by a speculator (e.g., asset managers by running a regression of her position on the asset return).** The coefficient of this regression is

$$\beta(\theta) = \frac{\text{Quality of the Asset Manager Predictor}}{\text{Risk Aversion}}.$$

- **Let $\Delta\beta$ be the difference between the average β for the top and bottom deciles of βs across speculators (a proxy for the difference between $\beta(\theta^*)$ and $\beta(\underline{\theta})$).**

- **Prediction:** Greater computing power reduces $\Delta\beta$ while data abundance can increase it.

Data Abundance and Price Informativeness

- **Result 3: The effect of data abundance and computing power on asset price informativeness are not the same:**
  1. **Greater computing power** (lower c): **Positive Effect.**
  2. **A push back of the data frontier (smaller $\underline{\theta}$): Positive Effect.**
  3. **The needle in the haystack problem (smaller α): Negative Effect.**
- Over the long run, data abundance can both push back the data frontier **and** increases the needle in the haystack problem $\Rightarrow$ The effect of data abundance on asset price informativeness is ambiguous (as found empirically by Bai, Phillipon and Savov (2015) and Farboodi, Matray and Veldkamp (2019).)

Data Abundance and Asset Price Informativeness

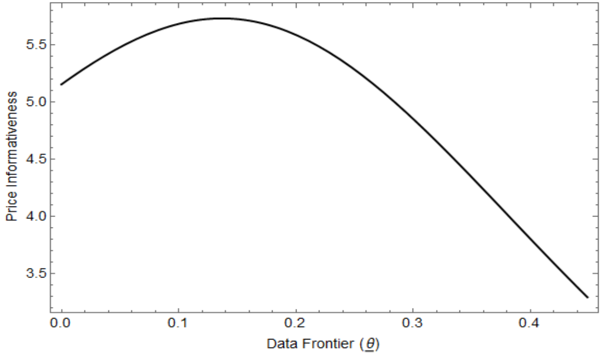


Figure 1: This graph shows the evolution of price informativeness in equilibrium, $\mathcal{I}(\theta^*, \underline{\theta})$ as a function of the data frontier, $\underline{\theta}$ when $\phi(\theta) = 3cos(\theta)sin^2(\theta)$ and $\alpha = Min\{1, 0.32 + 0.8 * \underline{\theta}\}$. Other parameter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

Data Abundance and Trading Profits

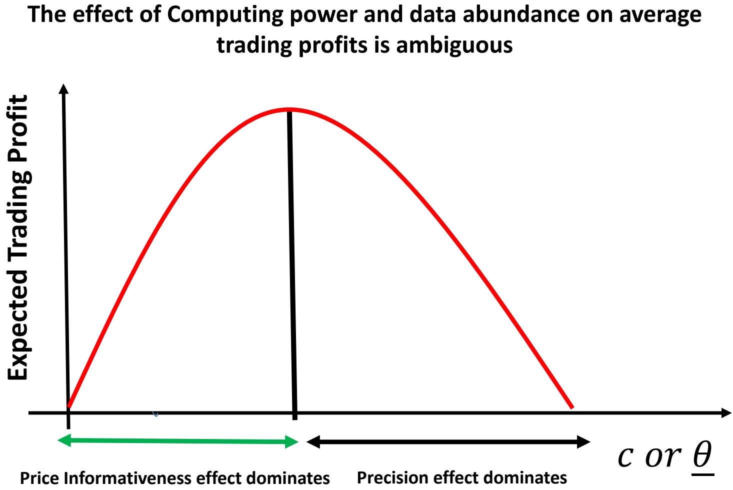
- **Speculators' expected trading profits:**

$$\frac{\text{Average Quality of Speculators' Predictors}}{\text{Risk Aversion} \times \text{Price Informativeness} \times \text{Volatility}}.$$

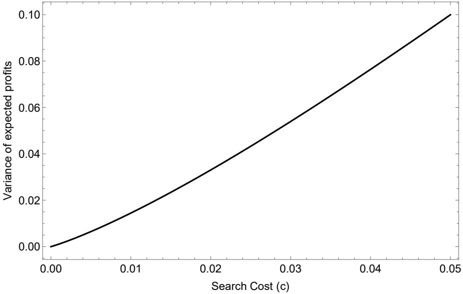
- **Dispersion (variance) of speculators' trading profits:**

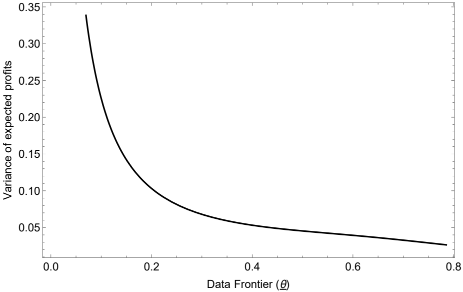
$$\frac{\text{Variance of the Quality of Speculators' Predictors}}{(\text{Risk aversion} \times \text{Price Informativeness} \times \text{Volatility})^2}.$$

Data Abundance, Computing Power, and Trading Profits



Data Abundance, Computing Power and The Dispersion Trading Profits





Data Abundance and Crowding

- **Does data abundance make investors' positions in the risky asset more or less similar?**

- **A measure of similarity in speculators' positions: Pairwise correlation between the positions of speculator j and i given the choice of their predictors:**

$$corr(x(s_{\theta_i}, p^*), x(s_{\theta_j}, p^*)) = \left(1 + \frac{\mathcal{I}(\theta^*, \underline{\theta})}{\tau(\theta_i)}\right)^{-\frac{1}{2}} \left(1 + \frac{\mathcal{I}(\theta^*, \underline{\theta})}{\tau(\theta_j)}\right)^{-\frac{1}{2}}$$

- **Thus, holding speculators' precisions fixed:**

  1. Greater computing power reduces the pairwise correlation in speculators' holdings because it increases price informativeness (speculators' positions differ because the noise in their signals is uncorrelated).

  2. Data abundance *reduces* the pairwise correlation in speculators' holdings when it *reduces* price informativeness (and increases it otherwise).

- **Similar (but more complex pattern) if one considers the "average" pairwise correlation, i.e., $\mathbb{E}(corr(x(s_{\theta_i}, p^*), x(s_{\theta_j}, p^*)))$.**

Conclusion

- **A micro-foundation of information acquisition costs for investors, based on a search model (many possible extensions).**

- **Data abundance and greater computing power are two distinct dimensions of progress in information technologies. They do not affect equilibrium outcomes in financial markets in the same way.**

Thank You! and Take Care!