Equilibrium Data Mining and Data Abundance

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

October 14, 2020
Progress in Information Technology

Volume of data created per year (in zetabytes). Source: IDC; Seagate; Statista estimates

Cost per computation per second deflated by the price index for GDP in 2006 prices. Source: Nordhaus (2007)

FIGURE 1
“[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)
Progress in information technologies has two distinct dimensions. It allows investors to have:

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: Data processing is less costly. “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 1/30th of the time taken by [...] a standard computer” (A.Ledford, Chief Scientist, Man AHL).
Data Abundance and Search

- **Two effects of data abundance:**

  1. ‘‘Hidden gold nugget effect:’’ Some datasets enable investors to find very good predictors.

  2. ‘‘Needle in the haystack effect:’’ The fraction of truly useful datasets go down (think about scientific publications; More than 23,000 Covid related articles since January 2020).

- **Data abundance creates a search problem.** “The exploration of new data sources in its relative infancy [...] extracting useful signals from the avalanche of data remains challenging”. ([Source: “AI Pioneers in Investment Management”, CFA Institute](https://www.cfainstitute.org/study/library/article/ai-pioneers-in-investment-management/)).
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)
Searching for Predictors

Seeking Alpha
How WorldQuant scours global markets for hidden patterns

Researchers around the world analyze the data, coming up with ‘alphas’

WorldQuant has five offices in the U.S. as well as 15 foreign outposts

Data team brings in data sets, ranging in hundreds in total

An ‘alpha’ is a prediction of market activity based on data inputs

The ‘Uber’ team synthesizes, allocates funds and balances risk

Strategies are turned into portfolios

Alphas are packaged into strategies

Trade is executed

Sources: WSJ analysis; company interviews

THE WALL STREET JOURNAL.
Questions

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

- Equilibrium data mining: How do speculators optimally search for predictors?

- How does progress in information technologies affect (i) the heterogeneity of predictors used by asset managers, (ii) the distribution of their trading profits, (iii) the correlation of their positions (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).

- **Other interesting result**: Endogenous distribution of “investment skills” for asset managers even when asset managers are ex-ante identical.
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

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 expanding the search space for trading signals.
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**

**Exploration:**
- Each speculator searches for a predictor of the asset payoff.
- In each search round, a speculator finds a predictor with probability $\alpha Pr(\theta \in [\theta, \frac{\pi}{2}])$.

**Period 1**

**Trading:**
- Each speculator observes the realization of her predictor ($s_\theta$) and chooses a trading strategy, $x(s_\theta, p)$.
- Speculators, noise traders and dealers trade.
- Market clears: The asset price is realized.

**Period 2**

**Asset payoff, $\omega$, is realized.**
Predictors

- There is a continuum of “predictors” of the asset payoff. Each predictor $s_\theta$ is characterized by its type, $\theta$ such that:

$$s_\theta = \cos(\theta) \omega + \sin(\theta) \epsilon_\theta$$

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

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

$$\tau(\theta) = \left(\frac{\cos(\theta)}{\sin(\theta)}\right)^2 = \text{“quality”}.$$

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


Search Space

\( \alpha : \text{Fraction of informative datasets} \)

\( \alpha \times \phi(\theta) \)

Data Frontier

\( \alpha \times \Pr (\theta \in [\theta, \frac{\pi}{2}]) \)
Data Mining (period 0)

Exploration in Round $n$

Pay exploration cost $c$

$$\lambda = \alpha \times \Pr (\theta \in [\theta, \frac{\pi}{2}])$$

1. Find a predictor with type $\theta \in [\theta, \frac{\pi}{2}]$
   - Keep searching
   - Move to round $n + 1$

2. Find no predictor
   - Stop searching
   - Move to round $n + 1$

- Wait for the market to open

Move to round $n + 1$

Exploration in Round $n$
Interpretation

▶ Three parameters to capture the effects of progress in information technologies: $c$, $\theta$, $\alpha$.

▶ 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 $\rightarrow \theta$ decreases.

2. It reduces the fraction of informative datasets ("Needle in the haystack effect") $\rightarrow \alpha$ decreases.
Data Abundance: Hidden Gold Nugget Effect

\[ \alpha \times \phi(\theta) \]

Data Frontier
Data Abundance: Needle in the Haystack Effect

\[ \alpha \times \phi(\theta) \]

Data Frontier

prop of informative datasets
Optimal Data Mining – a Stopping Rule Strategy

Exploration in Round $n$

Pay exploration cost $c$

$\lambda = \alpha \times \Pr (\theta \in [\theta, \frac{\pi}{2}])$

$1 - \lambda$

Find no predictor

Find a predictor with type $\theta \in [\theta, \frac{\pi}{2}]$

$\theta > \hat{\theta}_i$

Keep searching
Move to round $n + 1$

$\theta \leq \hat{\theta}_i$

Stop searching
Wait for the market to open

Move to round $n + 1$

Exploration in Round $n$
Optimal Data Mining

\[ \alpha \times \phi(\theta) \]

Satisficing predictors

\[ \Lambda(\theta, \hat{\theta}_i) = \alpha \times \Pr(\theta \in [\underline{\theta}, \hat{\theta}_i]) \]
Trading (period 1)

- Speculators observe the realization of their predictors \((s_\theta)\) 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 \(\eta\), where \(\eta \sim \mathcal{N}(0, \nu^2)\)
  3. The asset price is set by risk neutral competitive market makers, i.e., such that:
    \[
    p^* = \mathbb{E}[\omega | D(p^*)],
    \]
    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 strategy, $x^*_i(s_\theta, p)$ to maximize her expected utility:

$$
\mathbb{E}[-\exp(-\rho W_i)] = 
\mathbb{E}[-\exp(-\rho (x_i(s_\theta, p)(\omega - p)))] \times \mathbb{E}[\exp(\rho (n_i c))] 
$$

Expected Utility from Trading
Expected Utility Cost of Exploration

Key difference with the traditional model: $n_i$ is random and its distribution is endogenous.
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 predictors of different quality ex-post because the outcome of the search for predictors is random.

2. The ex-post distribution of predictors’ quality (speculators’ skills) is endogenous. We denote the average quality of predictors used in equilibrium by $\bar{\tau}(\theta^*; \underline{\theta}, c, \alpha)$.

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 $\theta^*$)

2. Solve for the equilibrium of the search stage, i.e., $\theta^*$. 
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 her information.
2. The equilibrium price reflects the information about the asset payoff contained in investors’ aggregate demand.

The informativeness of the equilibrium price is:

\[ I(\theta^*) \equiv \text{Var}[\omega \mid p^*]^{-1} = \tau_\omega + \frac{\bar{\tau}(\theta^*)^2}{\rho^2 \nu^2} \omega^2, \]

A speculator who finds a predictor with type \( \theta \) obtains an expected utility from trading equal to:

\[ g(\theta, \theta^*) = -\left(1 + \frac{\tau(\theta)\tau_\omega}{I(\theta^*)}\right)^{-\frac{1}{2}}. \]

↗ with \( \tau(\theta) \), ↘ with \( \bar{\tau} \) (competition effect)
Equilibrium of the Search Stage

$$J(\hat{\theta}_i, \theta^*)$$

Pay exploration cost $c$

$$\lambda = \alpha \times \Pr (\theta \in [\theta, \pi/2])$$

Find a predictor with type $\theta \in [\theta, \pi/2]$

$$\theta > \hat{\theta}_i$$

$$g(\theta, \theta^*)$$

$$\theta \leq \hat{\theta}_i$$

Find no predictor

$$J(\hat{\theta}_i, \theta^*)$$

1 - $\lambda$

Exploration in a given round
Equilibrium of the Search Stage

- **Solving for** $J(.)$:

\[
J(\hat{\theta}_i, \theta^*; \theta, c) = \frac{\exp(\rho c) \Lambda(\hat{\theta}_i; \theta, c)}{1 - \exp(\rho c)(1 - \Lambda(\hat{\theta}_i; \theta, c))} \mathbb{E} \left[ g(\theta, \theta^*) | \theta \leq \theta \leq \hat{\theta} \right].
\]

**Expected Cost of Search**

**Expected utility from trading**
Expected utility from trading

$g(\theta, \theta^*)$

$J(\hat{\theta}, \theta^*)$

$\hat{\theta}(\theta^*)$

$\pi/2$

Equilibrium Data Mining
Equilibrium Data Mining

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

$$g(\theta^*, \theta^*) = J(\theta^*, \theta^*; \theta, c, \alpha).$$

- 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 $[\tau(\theta^*), \tau(\theta)]$. 
IMPLICATIONS
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 $\theta$) can induce speculators to be less demanding for the quality of their predictors for $\theta$ low enough).

3. The needle in the haystack problem (lower $\alpha$) always induces speculators to be less demanding for the quality of their predictors.
Illustration
Economic Mechanisms

- **If the cost of exploration** \( (c) \) **decreases:** The continuation value of searching increases, holding \( \theta^* \) constant \( \Rightarrow \) Speculators are more demanding (search more) in equilibrium (\( \theta^* \) declines).

- **Needle in the haystack problem:** Same as an increase in \( c \).

- **If the quality of the most informative predictor increases:**
  1. The expected utility of trading for the speculator who finds the best predictor increases (“hidden gold nugget effect”)
  2. The expected utility of trading for all others speculators decreases (competition effect)
  3. \( \Rightarrow \) **The net effect on the continuation value of searching is ambiguous.**
Data Abundance and the Quality of Predictors

Needle in a Hay Stack Effect (NHS) + Competition Effect

Data Frontier pushed back

Competition + NHS effects

\[ \theta_0 \]

\[ \theta_1 \]

\[ \theta_0^* \]

\[ \theta_1^* \]

\[ \pi/2 \]
Data Abundance and Heterogeneity in 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 future asset return (from date 1 to 2 in the model). The coefficient of this regression is

\[
\text{Stock Picking Skill} = \beta(\theta) = \frac{\text{Quality of the Asset Manager Predictor}}{\text{Risk Aversion}}.
\]

- Let \( \Delta \beta \) be the difference between the average \( \beta \) for the top and bottom deciles of \( \beta \)'s across speculators (a proxy for the difference between \( \beta(\theta) \) and \( \beta(\theta^*) \)).

- **Prediction:** Greater computing power reduces \( \Delta \beta \) while data abundance can increase it.
Interpreting Heterogeneity in Stock Picking Skills

- Kacperczyk and Seru (2007): Asset managers’ stock picking skills (precision of asset managers’ private information) are heterogeneous.
- **Standard interpretation:** This reflects heterogeneity in asset managers’ intrinsic ability to process information or effort to acquire information.
- **Our model:** No differences in ex-ante ability or effort (the stopping rule is the same for all speculators). Heterogeneity just stems from randomness in the outcome of the search process for predictors.
- If predictors change over time, there should be **no persistence** in the ranking of asset managers in terms of skills.
- The distribution of skills should vary in predictable way with shocks to volatility, noise trading, data abundance, and information processing costs.
Result 3: The effect of data abundance and computing power on asset price informativeness are not the same:


2. A push back of the data frontier (smaller $\theta$): Positive Effect.

3. The needle in the haystack problem (smaller $\alpha$): Negative Effect.

Over the long run, data abundance can both push back the data frontier and increases the needle in the haystack problem. ⇒ 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).)
Figure 1: This graph shows the evolution of price informativeness in equilibrium, $\mathcal{I}(\theta^*, \theta)$ as a function of the data frontier, $\theta$ when $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$ and $\alpha = \text{Min}\{1, 0.32 + 0.8*\theta\}$. 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:**
  
  \[
  \frac{\text{Variance of the Quality of Speculators’ Predictors}}{(\text{Risk aversion} \times \text{Price Informativeness} \times \text{Volatility})^2}.
  \]
The effect of Computing power and data abundance on average trading profits is ambiguous.
Data Abundance, Computing Power and The Dispersion Trading Profits

![Graph 1: Search Cost (c) vs. Variance of expected profits](image1)

![Graph 2: Data Frontier (θ) vs. Variance of expected profits](image2)
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:

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

▶ 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., $\mathbb{E}(\text{corr}(x(s_{\theta_i}, p^*), x(s_{\theta_j}, p^*))$.
Conclusion

▶ An equilibrium model of financial data mining.

▶ 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.

▶ Extensions:
  1. Spurious predictors (False discoveries; see Harvey (2017)).
  2. Time variations in the distribution of anomalies (Chinco, Neuhierl, and Weber (2020))
  3. Multiple risky assets
Thank You! and Take Care!