

IA and the Financial System:

(1) Will Intermediaries be Disintermediated?

(2) Alternative Data For Investment

Université Paris-Dauphine - PSL, MASEF + 203 – 2023

جهاز أبوظبي للاستثمار
Abu Dhabi Investment Authority

ADIA

Charles-Albert Lehalle

Global Head of Quant R&D, ADIA, Abu Dhabi, UAE

(Louis Bachelier Fellow, Paris, France)

(Visiting Professor, Imperial College London)

What Is It About? Data Sciences? Intermediation?

Applications?

Focus On (Alternative) Data

Return To The Source(s)

Post-Stratification

Basics of Data Processing By Machine Learning

Towards a Better Connection to Real Economy

What's Next?

Focus On (Alternative) Data

What Is It About? Data Sciences? Intermediation?

Focus On (Alternative) Data

What's Next?

Focus On (Alternative) Data

Return To The Source(s)

Post-Stratification

Basics of Data Processing By Machine Learning

Towards a Better Connection to Real Economy

Let Start From The Very Beginning: Definitions

What is a data?

- ▶ a data is a recorded information, having a **unit**.
Its unit is very important, since it is the start of a specification of what to do with it.
- ▶ a data is not isolated: it is part of a **sequence of data**,
- ▶ a data has a **timestamp**.

Where do the data come from?

- ▶ usually from a **sensor**: temperature, pressure, etc.
- ▶ a company generates data by its day to day activity: production, mails, interactions with clients and suppliers.
- ▶ your suppliers and your clients generate data,

New: some institutions generate and record data for free, some companies have clients with an only goal: generate data (and sell them)

- ▶ As long as nobody use them, data cannot be considered as clean
- ▶ Changes of units and of recording (or processing) methodologies have to be known.
- 🔍 Your process to record data generates data...
- ▶ Computations are data: there is not difference with the result of computations run on datasets and a new dataset.
- 🔍 Not storing the result of data processing is a choice.
- ▶ You need to maintain a referential about
 - which process generate with data,
 - which process consumes which data.

It will enable you to optimize and maintain your workflow. The origin of all this has been Complex Event Processing.

Most often, datasets are collected for a specific purpose. For instance

- ▶ web traffic to optimize the navigation of web sites
- ▶ credit card tickets to target financial services
- ▶ suppliers and clients as declared to regulators
- ▶ etc.

In all these examples the dataset is (almost) an i.i.d. sample of the distribution of interest, because data are collected close to the populaion of interest: users of web sites, users / usages of credit cards, companies under a specific regulation, etc.

But to implement nowcasting, your distribution is made of all the economic or physical entities of this kind: all web sites, all consumers spending in shops, all companies, etc.

As a consequence these dataset have **a collection bias** : part of the population of interest is not considered at all during the collection process (only users using internet to get awareness on a company, only consumers having credit specific use of their cards, a specific type of companies, etc).

But it is not all: for systematic investment you need to connect (to map) entities in your dataset to tradable instruments. Here again there is another bias, a **blind spot bias**. You will only have traffic on companies using a specific web technology, credit card expenses will not make sense for all companies,

To handle these biases, you need to be creative during your investigation, going back and forth in **testing different comparisons with reference datasets**.

Once it is done, you can

- ▶ estimate the biases and propose a correction, that is the purpose of **post-stratification**,
- ▶ you do it an interactive way, that **introduces an exploration-exploitation process**

Collection Bias

- ▶ collect information only for a type of **consumers**
- ▶ focus on a **subpopulation** (consumers profile, companies type, interm., etc)
- ▶ collect more information with time, or **seasonal** collection of information (non stationarity)
- ▶ **methodological** changes (surveys, ESG, etc)
- ▶ collection related to a **technology** (mobile OS, web site, trucks, etc)
- ▶ country or industry bias

Is a bias always your foe? sometimes **the collection bias can be the information** (job posting, patents).

Blindspot/Mapping Bias

- ▶ country,
- ▶ industry (retail facing only),
- ▶ characteristics (size, age, R&D, etc)
- ▶ vertical vs. horizontal **business type**
- ▶ brand or company name **mappings** (PiT?)
- ▶ financial **symbolology**

Focus On (Alternative) Data

Return To The Source(s)

Post-Stratification

Basics of Data Processing By Machine Learning

Towards a Better Connection to Real Economy

What Kind Of Biases Are We Talking About?

A typical example is a population of geolocalized mobile phones in US malls, you may want to compare

- ▶ the demographics of this population to the one of the US population,
- ▶ the brands presents in the malls of the dataset with the retail brands in the US,
- ▶ the usual overshoot of consumption during the Back Friday with the geolocation.

It is a question of [choosing a reference model](#). For instance which population do you have in mind

- ▶ the whole US population? US consumers?
- ▶ US consumers in malls? US consumers of brands that are in malls?
- ▶ US consumers of brands that are in your dataset?

You have to choose a reference that is not too far away from your sample, to [reflect the natural informational content of the dataset](#).

The ideal reference is related to what you want to do with the data (do you really want to use the data very far away from their natural distribution?).

“ Broadly speaking, post-stratification refers to any method of data analysis which involves forming units into homogeneous groups after the sample has been taken. ” [Zhang, 2000]

The way it is usually expressed in the literature (Survey Theory):

- ▶ You start with strata $s \in \mathcal{S}$ that are disjoint subsets created from a categorical variable (a state, an industry, the age, etc),
- ▶ You want to apply weights $(w_s)_s$ to these strata such that the weighted average of an observations $(x_s)_s$ is as close as possible to the desired expected value \bar{X} : $\sum_s w_s x_s \simeq \bar{X}$. Keep in mind that \bar{X} and each x_s are vectors (to control simultaneously for several biases).
- ▶ But you want the weights w_s to be as close as $a := 1/\#\mathcal{S}$ as possible; you express this using a distance function $\sum_s a \cdot G(w_i/a)$ (see [Deville et al., 1993]).
- ▶ Hence you end up with a constrained optimization that, when $G(r) := (r - 1)^2/2$ boils down to

$$(1) \quad w_i = a \cdot \left\{ 1 + (\bar{X} - \sum_s a x_s) \left(\sum_s a x_s^T x_s \right)^{-1} x_i^T \right\}.$$

The correction w_i is not certain:

- ▶ The previous method allows to estimate the weights w_s to apply to each observation n given it belongs to the stratus s .
- ▶ In practice this weight has an estimation error ε_s ,
- ▶ typically when a fraction p_s of the reference population is expected and the observed sample has a size K , the observed fraction is expected to have a variance of $p_s(1 - p_s)/K$ that can be propagated in the expression (1).

Writing $\tilde{w}_n := w_n + \varepsilon_n$ for the weight to apply to observation n , let's [look at what happens when you use the sample in a linear regression](#):

$$\min_{\beta} \mathbb{E}_{\tilde{w}} \|Y - \beta X\|^2 \Rightarrow \hat{\beta} := (X^T \Delta_{\tilde{w}} X)^{-1} X^T \Delta_{\tilde{w}} Y.$$

It introduces an [uncertainty on the regression coefficients](#): when ε is too high, it may prevent the estimated coefficients to be statistically different from zero...

A lot of variations have been proposed. In essence this approach allows,

- ▶ once you identified groups of observations on which you have an external reference,
- ▶ to adjust weights on your observed sample to get them as close as possible to this external reference.

Open questions

- ▶ the uncertainty due to the size of the sample has to be taken into account,
- ▶ To what extend a sample can be reweighed? for which application?

It is “easy” as long as you restrict yourself to groups/categories of observations.

Focus On (Alternative) Data

Return To The Source(s)

Post-Stratification

Basics of Data Processing By Machine Learning

Towards a Better Connection to Real Economy

- ▶ Improve the service delivered to **clients**
 - 🔊 Scoring, clustering, recommendation engines, etc
- ▶ Have a better connection with **reality**
 - 🔊 Prediction, now casting, natural language processing, etc.
- ▶ Improve your **own processes** : optimization, risk control.
 - 🔊 Portfolio management, Improvement of sensitivities estimations (Monte-Carlo simulations and PDE).

First of all, standard methods do not work on **unstructured data**, why?

► For **geolocated** data:

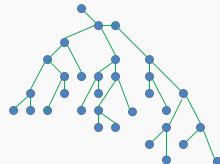
- You need to build masks or selections (possibly on indicators coming from other datasets)
- You need to apply these masks / selections on geographical zones
- 👁 You need to take **distances** into account (on several channels simultaneously)

► Some dataset are structured in **graphs**:

- supply chain
- history of a client
- sequences of transactions

The main difficulty with unstructured data is that **they are not organized in nice rectangles** (filled rows and columns).

Moreover, remember that statistics are about **partially observed distributions**, what is a time series of graphs? of geolocated data? how to process an update of a graph?



Improvement of risk hedging / valuation of financial products.

1. Improving risk management

- ▶ more dimensions, new risks
- ▶ data driven risk assessment (over fitting, etc)
- ▶ asset Management (crowding, factors, products)

🔗 Implementing end-to-end architectures in fragmented infrastructures, governance of models.

2. Simulations

- ▶ small scale (HFT)
- ▶ Long term (climate risk and longevity) - multi scale (improving stress tests)

3. Regulation

- ▶ stress tests and simulations
- ▶ adjust the capital requirements and inventories limits to the need of liquidity

Focus On (Alternative) Data

Return To The Source(s)

Post-Stratification

Basics of Data Processing By Machine Learning

Towards a Better Connection to Real Economy

With the availability of new sources data, it is now possible to build extra financial view on companies. How to include this source of information in the valuation of tradable assets. Where will these information will contribute to the price formation during the life cycle of companies / projects?

1. Accurate Finance and Insurance (ie Finance et assurance de précision)

- ▶ use of extra financial indicators on companies (rating agencies)
- ▶ use of sensors (satellites, GPS on people and cars)

Would need to “parametrized insurance contracts” and valuations in presence of nowcasted information.

2. Financing of projects / companies

- ▶ new sources of financing (crowdfunding, ICO)
- ▶ the lifecycle of financing and associated risks

It is very difficult to train a system to predict future returns, nevertheless it seems easier to perform signal processing tasks with machine learning.

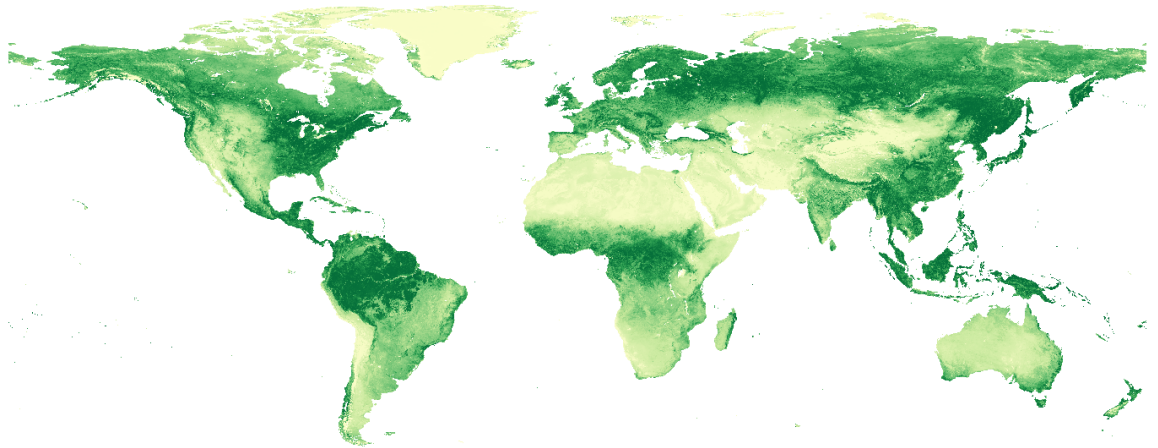
In many cases you have access to images (or a collection of unstructured information, think about premise.com) and you do not want to predict something, but to summarize the information that is currently inside these data. This is a great advantage:

- ▶ you are not guessing something unrelated to the data
- ▶ you try to reconstruct an existing economic indicator, that is just not officially computed yet.

Think about counting cars parked, boats in industrial zones, etc.

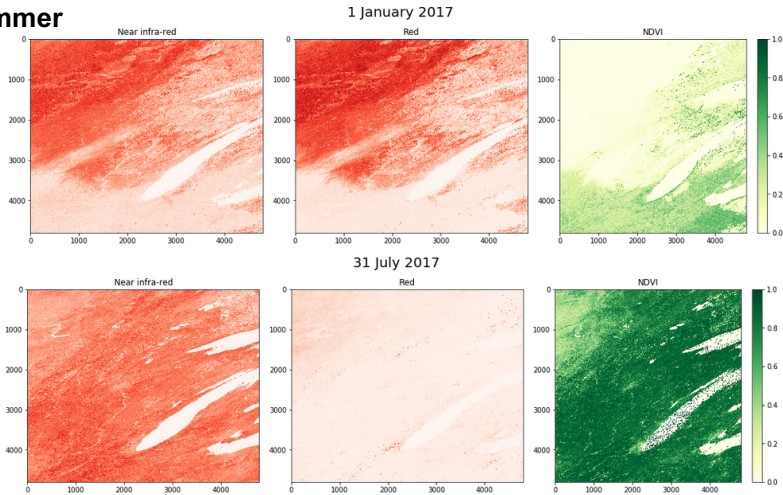
A typical example is to establish crop conditions using satellite images. We will see here an example with the Modis suite of NASA instruments.

$$NDVI = \frac{\text{Near infra red} - \text{Red}}{\text{Near infra red} + \text{Red}}$$



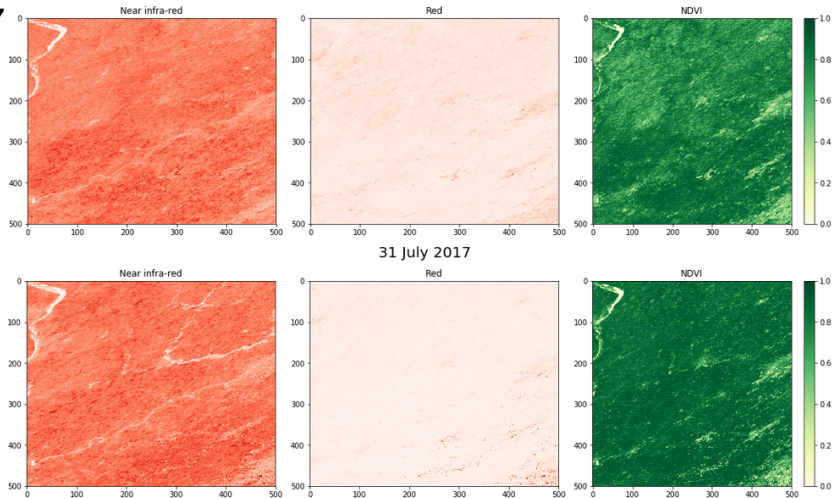
Modis image: 2km x 2km, a map of the globe
(images from public presentations by CFM)

Winter vs Summer



Modis image: 2km x 2km, a typical year

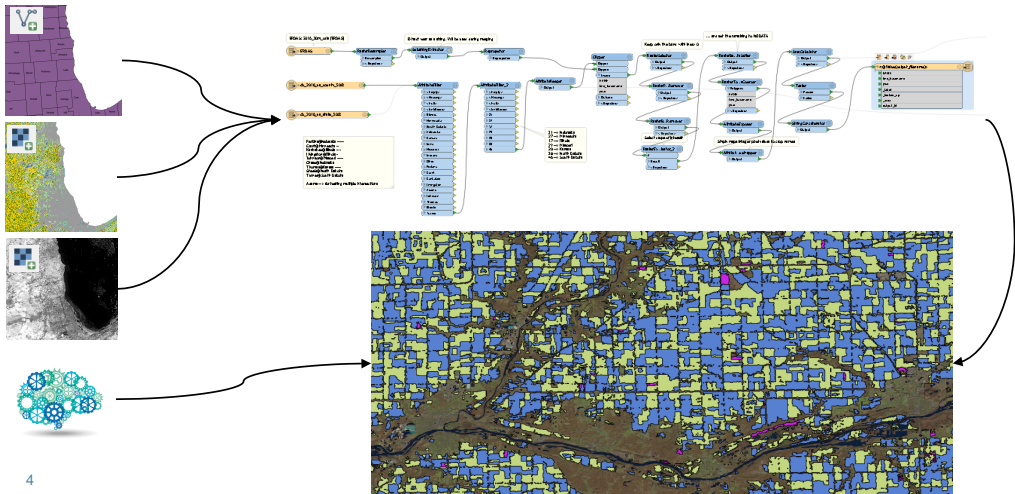
2012 vs 2017



3

Modis image: 2km x 2km, comparing 2 years

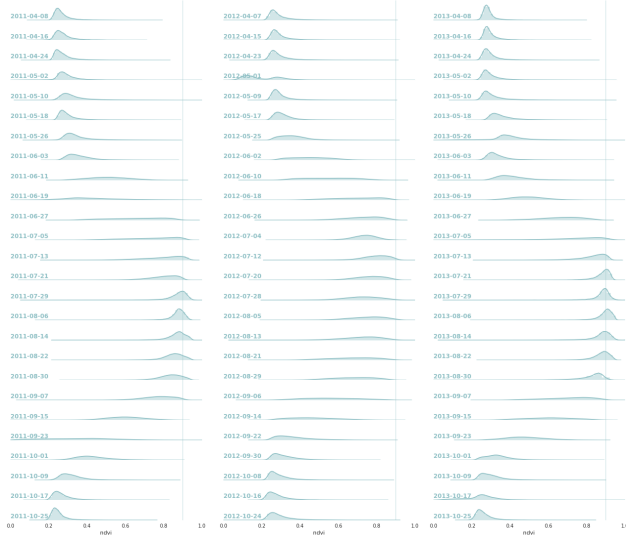
Processing can be automated with geospatial ETL



Modis image: 2km x 2km, a geoprocessing pipeline

Weekly NDVI distributions can be compared over years:

- With 8 days delivery imagery, we have 45 sets of observations per year
- For each date, and the selected area, we have around 1.2M pixels



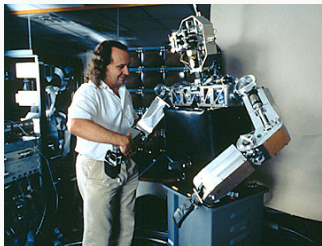
Modis image: 2km x 2km, weekly distribution of biomass

What's Next?

What Is It About? Data Sciences? Intermediation?

Focus On (Alternative) Data

What's Next?



source:
artificialhumancompanions.com

The next challenge is **human-machine interface**

- ▶ What is the information to share with human / supervisors? [Azencott et al., 2014]
- ▶ Never trigger a warning without thinking about corrective options the human will have.
- ▶ Be ready to **audit the machine**: store the code and the data used to learn.

In the scope of **sharing responsibilities with machines**, we will have to solve the following points (cf [Mittelstadt et al., 2016]):

- ▶ inconclusive behaviour
- ▶ unscrutable behaviour
- ▶ misguided behaviour
- ▶ unfair behaviour
- ▶ transformative consequences
- ▶ lack of traceability

We have seen some mechanisms of machine learning, and we went through a lot of potential areas of applications.

The **main themes of applications**

- ▶ automated customization (instrument by instrument, client per client, etc),
- ▶ automated summary of an unstructured collection of information,
- ▶ nowcasting.

An important feature of these technologies is that **it allows you to compose your own menu** from an unstructured set of information offer.

We have seen some mechanisms of machine learning, and we went through a lot of potential areas of applications.

The **main themes of applications**

- ▶ automated customization (instrument by instrument, client per client, etc),
- ▶ automated summary of an unstructured collection of information,
- ▶ nowcasting.






An important feature of these technologies is that **it allows you to compose your own menu** from an unstructured set of information offer.







Open Questions






- ▶ What is Artificial Intelligence?
- ▶ And what about blockchain...

Thank You For Your Attention – Any Question?



-  Almgren, R. (2012).
High-Frequency Event Analysis in Eurex Interest Rate Futures.
Technical report.
-  Azencott, R., Beri, A., Gadhyan, Y., Joseph, N., Lehalle, C.-A., and Rowley, M. (2014).
Realtime market microstructure analysis: online Transaction Cost Analysis.
Quantitative Finance, pages 0–19.
-  Belleflamme, P., Lambert, T., and Schwienbacher, A. (2010).
Crowdfunding: an industrial organization perspective.
In Digital Business Models: Understanding Strategies, Paris, pages 25–26.
-  Brandes, Y., Domowitz, I., Jiu, B., and Yegerman, H. (2007).
Algorithms, Trading Costs, and Order Size.
Technical report, Investment Technology Group.
-  Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993).
Generalized raking procedures in survey sampling.
Journal of the American statistical Association, 88(423):1013–1020.

-  Fermanian, J.-D., Guéant, O., and Rachez, A. (2015).
Agents' Behavior on Multi-Dealer-to-Client Bond Trading Platforms.
-  Froot, K., Kang, N., Ozik, G., and Sadka, R. (2015).
Private information and corporate earnings: Evidence from big data.
Technical report, Harvard Business School.
-  Geeraert, S., Lehalle, C.-A., Pearlmutter, B., Pironneau, O., and Reghai, A. (2017).
Mini-symposium on automatic differentiation and its applications in the financial industry.
-  Lehalle, C.-A. and Capponi, A. (2023).
Machine Learning and Data Sciences for Financial Markets: A Guide to Contemporary Practices.
Cambridge University Press.
-  Lehalle, C.-A., Laruelle, S., Burgot, R., Pelin, S., and Lasnier, M. (2013).
Market Microstructure in Practice.
World Scientific publishing.
-  Lehalle, C.-A. and Raboun, A. (2022).
Financial Markets in Practice: From Post-Crisis Intermediation to FinTechs.
World Scientific.

-  Merton, R. C. (1995).
A Functional Perspective of Financial Intermediation.
Financial Management, 24(2):23+.
-  Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016).
The ethics of algorithms: Mapping the debate.
Big Data & Society, 3(2):2053951716679679.
-  Nakamoto, S. (2011).
Bitcoin: A Peer-to-Peer Electronic Cash System [Illustrated].
Prequel Books.
-  Pages, G., Pironneau, O., and Sall, G. (2015).
Vibrato and Automatic Differentiation for High Order Derivatives and Sensitivities of Financial Options.
working paper or preprint.
-  Poushter, J. and Oates, R. (2015).
Cell Phones in Africa: Communication Lifeline.
Technical report, Pew Research.



Zhang, L.-C. (2000).
Post-stratification and calibrationa synthesis.
The American Statistician, 54(3):178–184.