Thoughts on Data Processing

Issues #5-6, 21/2/2023 & 2/3/2023

My Weekly Update

It was a tough week for me, trying to keep the work I planned on schedule. Read 130k+ words over texts this week, and felt overwhelmed by the information overflow and social pressure around me.

I am getting more accustomed to staying alone, as I am never able to share my thoughts with family because each time I will be judged and criticized. Without love or any sense of intimacy to people around, I attempted to keep myself mentally sane by spacing out Sunday for cycling, and spending some evenings taking strolls alone outside the Kowloon harbour.

I treasured hanging out with the few friends I trust on two of the days as they become the sole sources of my security; I realized, it has been that I wanted to spend more time with people/ close frds that I'm interested in knowing a bit more, but in reality everyone's busy with their things and I simply cannot have any feelings of attachment to people around.

So I am learning how to minimize expectations and make myself just a bit happier by exchanging ideas and gaining freshness from strangers – at least I know somewhere on ig there are people actually listening (or reading). How much I wish I could simply be a sun, radiating meaningful positivity for the people around me, without needing others to complement my loneliness and listen to me. As much as, that I know such thoughts is strictly a dream.

A reason why I (re-)created this account, apart from trying to organize my workstream and provide value to people around. Maybe in some sense, it is a good way to process my emotional clutter and fit them into organized containers – at least not a stack overflow over the mind that I need to process things around.

Hoping things to get better, I spent a night thinking about how to process the data and events around me well.

Catching Data that Fly Around

In reality, data is flying around us at an ever-increasing rate. To make data such as social events and digital messages into good use, we all need to summarize things and identify the main ideas – or to be data-savvy.

One may ask the following relevant questions:

- Where to the fetch information one needs?
- How to treat different forms of data well?
- How to summarize things quickly and effectively?
- How to separate the voices inside that matter from noises, and not to get affected by 'nonsense' remarks?

To me, I think it is important to realize what mindsets and tools one could use to process data well. However due to word limit constraints, I would only talk about #2 here (mindsets) and cover #1 and 3 (tools) next week.

#4 is related to personal growth, and I am probably not eligible to comment on given the social diversity.

Mindsets

Academic Data – Inquire Purposefully

Since our early days in HK education, we are taught to remember (recite?) things well on textbooks, and later in our HKDSE days, recite marking schemes or gain insights from tutorial classes to gain full credit in our papers. From Chinese reading passages to question frameworks to LS, we are all the "social output" of a highly standardized education system. Then all of a sudden, the teaching styled changed significantly in university to a more inquisition-oriented mode. But have you wondered why?

The education we have in the first twelve years provided us with chunks of (foundation) academic data of each subject – for us to try make sense of and organize. While teachers and tutors have provided us with specific ways to organize/ interpret the information well early in the stage, it turns out that each of us actually have our unique methods as we grow up further – and we need to identify it to smoothen learning.

Climbing up the education hierarchy, the academic info amount in one uni semester approaches 2-3 HKDSE subjects. With a steepened learning curve, it is important for us to learn how to inquire purposefully into the fields we are interested in. Data streams exist from professors, but also from alternative sources like Youtube or peer materials.

When professors proved inadequate to support learning (e.g. some UG math courses in UST), it is vital to know where to reach out and what to ask ("why?") – a mindset completely different from the Oriental spoon-fed edu systems – to maximize our learning outcomes. Such a thought of overcoming the bubble is also crucial in getting on board for jobs. Over time, it is one's responsibility to discover the specific methods/ tools to help their minds grow.

Social Data – Gather Before Interpret

Naturally, social data contain bias. To obtain a relatively unbiased opinion, it is important to verify data from multiple (independent) sources to make sense of everything. In real life, We trust well-established "news" sources like Bloomberg for financials, or regional prominent outlets like AP (US), BBC (UK) or NHK (Japan) for feed. But what if they are also biased?

And what about for decentralized sources, e.g. Twitter in the crypto world – do we trust "a bull market of BTC 100K in a few months" is coming? Do we firmly believe in FTX proponents' tweets that the exchange would continue to grow robustly, or Do Kwon that he is "deploying capital" and remain "stead" ?

It is noteworthy that everyone has their underlying motives – in a way well-reflected into their accounts. In general, it is not wise to interpret before knowing sb/ sth comprehensively – doing so might risk making decisions/ remarks prematurely (盲人摸象), or making the guy talking to you feel like they are judged pre-emptively (so they wouldn't open up their mind).

By delaying the point of judgement to a time sufficient to gather much important available information, we can make a comprehensive use of the data we obtained for sensible remarks – without getting our minds into bias that can adversely affect our decisions.

Unfortunately, the DSE Liberal Studies exam encourages immediate judgement over the texts we are fed, despite its good intentions to encourage critical thinking.

While preserving the ability to critically analyze social data around us, it is important to put the ideas on relative scale so as to compare and make good interpretations. Although it often means spending a significant amount of time in doing so, being non-judgemental and inclusive to ideas until there are adequate information can help us build more comprehensive pictures, or more convincing bullets.

Just, don't be too early to judge!

Career Data – Harvest Network Effect

Now, coming across data related to career is something rather sensitive/ controversial. For example:

- At a firm level, NDAs are signed to prevent employees/ interns leaking some confidential information (e.g. deals, organizational structures)
- For people within a batch, limited places within a narrow career pathway may hinder information flow: increase in competitiveness of other batchmates is adverse to one's personal standing.
- Sharing personal salary information/ CV to colleagues or batchmates might generate unnecessary judgements and FOMO (comparison gives implicit psychological Pressure).

Some game theory analysis will bring us to the Nash Equilibrium of no-one willing to share things. Sure, proprietary information should be kept strictly confidential, but encouraging the flow of personal items could be beneficial to community growth as a whole, at the slight expense of your own standing.

In balancing between my information 'sacrifice' and expected community impact, I have questioned myself the optimal way to handle career data for long.

Over time I realized the following drivers:

- **Networking effect**: Some people have more insights on their data than you. Conditioned on your willingness to add value to others, they are usually willing to share reciprocally – especially with people whom you trust.
- **Information asymmetry**: Being transparent promotes optimal execution of career decisions by, for example, reducing the salary loss incurred from misinformation over market rates/ job nature. This is why websites like GlassDoor exists.
- **Kickstarter effect**: While keeping some of my info strictly private (directly relevant to hitting aim), giving out less competitive but useful info can encourage mutual growth – building reputation as side benefit.

(For the "tools" part: stay tuned!)

My Weekly Update

Another week down the semester, I dwelled in the world of Excel with thousands of data rows. I delivered multiple Excel/VBA automation solutions to colleagues in Masters CPD office, got myself close to wrapping up step 1 of my quant preparation task and finished the first assessment of LABU course (an interview recorded live) today (1/3). My phone broke down last week, and it made me realize that while I do want a small but stable friend circle that I can talk to every day, I realize it's just not possible, and I better draw towards my own thoughts and execution.

In a mentally compressed state, I spent mornings working in the open space of West Kowloon; afternoons having a Pok Fu Lam hike, cycling in LOHAS park, and watching a sunset on Star Ferry, and evenings looping blue pieces by Japanese vocalist Mitsuki Nakae (中恵光城) to balance my emotions. Solitude is the strongest form of support.

Data Becoming Dynamic

My data-handling work this week prompted me to ponder my unanswered questions last week:

1) How to process/ summarize data effectively?

2) Where to fetch the information needed?

We students might have got used to text skimming in university entrance exams. At some point we were taught about signposting, interpreting synonyms or summarizing the general idea about a certain text in comprehension/reading exams.

But soon it becomes apparent that the information we receive are dynamic as compared to the static exam passages in exams. We might have a constant data influx stream, or just sheer sizes of data (in MB/GBs) to be processed quickly. How do we proceed “in time”? What is “in time” anyway?

Say quants have to process large data sets containing tick-by-tick stock data to derive simple but valuable insights to predict price movements to do trades in seconds. How to be accurate and quick with the data?

Principles for Data Processing

Eventually, we need tools and criteria to quantify data processing effectiveness. In general, the following core principles should be considered:

- **C**onciseness (being relevant/ dense enough),
- **A**ccuracy (unbiasedness, completeness),
- **P**resentability (clarity, reader-friendly),
- **E**fficiency (speed and timeliness), and
- **S**calability (being dynamic)

The way we are educated in HK mainly focus on incorporating “accuracy” and “presentability” in exams; some western education systems or upper layers of education might also instill “conciseness” (e.g. Camb. English C1 Advanced// Word limit in uni assignments), albeit generally to a less extent.

Scaling Effectively with Automation

However, it seems like education systems don't emphasize efficiency and scalability enough in data processing – highly demanded in many roles today as the world goes digital.

Sure, those two points might appear insignificant in small tasks. Say, a student to organize their workstreams may use

- Containers such as excel dashboards, well-formatted websites or business one-pagers to organize qualitative data or project states well (presentability/ conciseness);
- File structures (e.g. Google Drive) to keep things accessible everywhere (presentability) ; or
- Flexible Text Editors like Goodnotes/ Notability on iPads, or Notion, to jot information readily and help organize things (presentability).
- Media (videos and audios) are relatively ineffective sources of data storage due to the need of retrieving; but they include a lot of hidden information (e.g. tone/ pace of voice) => Text preferred for processing

But what if the information amount we receive stacks up? A list of 500 students? A .csv file with 50k entries of 10-minute OHLC data of indices? A whole encyclopedia that answers any random question accurately? To process massive data bases and interpret trends, we rely on automation to scale things effectively:

- **High-Level (more human-facing) Data Analysis**
  We type queries (e.g. SQL/ Excel functions) for trivial data (e.g. # people, counting preferences)
- **Database management**: To go lower level, we design databases to store chunks of structured data effectively and optimize accessing time. Hence we learn DSA, OOP and time complexity in programming.
- **Machine Learning**
  For variables in data between which we want to determine the relationship, we train models, either
  - with human labels (**supervised learning**; to learn patterns, e.g. logistic regression/ NNs),
  - without human labels (**unsupervised learning**; no specific objective, to discover hidden structures/ data clusters)

  to predict changes in the test data set, and reality.

Depending on the fineness of the information required and scale of the projects, we strike a balance between the five indicators to make an optimized compromise; IRL though, most of us focus on microscopic projects, and hence focus on conciseness, accuracy and presentability.

But to quants, efficiency and scalability are highly important – and gaining such specific skills that the education system rarely covers require consistent, in-depth drilling.

Where to Fetch the Information Needed?

Of course, data analysis/ process only comes after gathering the data required – So, where do we actually fetch them?

In ancient times, academics relied on books to gain knowledge and social status – there is a Chinese saying that "reading can take you anywhere" (書中自有黃金屋): advanced, concentrated knowledge (on technical, history, culture…) were only accessible from books. But everything has changed with the Internet.

With information available everywhere, it has become one's wisdom in utilizing those tools that matters:

- For publicly available information, one may "Google" (Of course, now ChatGPT is better) to resolve information asymmetry.
- For private information, there are no readily available answers. We hence need to infer from "**quality benchmarks**", from trustful peers around us to general statistical figures from generally recognized sources like Statista to make estimates. Paid sources (e.g. BBG Terminal/ data vendors) can add credibility.

  → Of course, accuracy depends on quality of such "proxies"; in quant interviews this is how candidates make open-ended estimates and Confidence Intervals for "Fermi Problems".

Remaining Discerning

Yes, one should be data-savvy enough to recognize when to utilize each of those methods and the relevant tools; but it is equally important to remain discerning for the results found, to not get misled by misinformation.

→ e.g. ChatGPT is a strong Natural Language Processing (NLP) algorithm, but it is bad at math ("math skills" not inherent for itself, but rather from the underlying labelled data!); and for non-factual remarks it is likely to also contain human bias and misinformation, based on the underlying texts that it was trained.

It is important to retain (further polish) the ability to critically judge info and data – What are "quality benchmarks"? What if the info is wrong or contradictory? What if our automated data analysis results are actually not complete, because edge information has been lost through the convolutional stages?

We often take ”Accuracy” for granted in the way we studied, but may not be true in real life.