FORESIGHT

THE INTERNATIONAL JOURNAL OF APPLIED FORECASTING

ARTICLES *by* **FORECASTERS**
for **FORECASTERS: issue 67**

To Share or Not to Share? The Future of Collaborative Forecasting

SPECIAL FEATURE: COLLABORATIVE FORECASTING

To Share or Not to Share? The Future of Collaborative Forecasting

PIERRE PINSON

PREVIEW *Distributed data refers to information that flows from different sources and possibly different owners. Getting top value from distributed data requires a paradigm shift towards collaborative forecasting. Alternative frameworks exist to support collaborative forecasting, from collaborative analytics to data markets, and from analytics markets to prediction markets. While we should accept that not all data will be openly shared, rethinking forecasting processes with modern communication, distributed computation, and a market component could yield substantial improvements in forecast quality while unleashing new business models.*

INTRODUCTION

The quantity of data being collected by individuals and organizations is increasing at a fast pace. Today, we are talking about data volumes in the order of a quintillion bytes per day (a quintillion being a number with 18 zeros, i.e., a billion of billions!). In its edition of May 6, 2017, *The Economist* wrote: "The world's most valuable resource is no longer oil, but data."

Not all that data is valuable for forecasting applications, though. Since the models used for forecasting are increasingly data driven and data hungry, we ought to look for ways to get value out of all this abundance. Quantitative analysts and forecasters consequently focus on challenges related to data cleaning, feature engineering and selection, model building and validation. This is first based on the assumption that all data can be made available in a centralized manner. In practice, though, it is often not the case.

If the data cannot be gathered and centralized, does that mean it is not possible to extract value from them? Surely not. However, this calls for a paradigm shift toward collaborative forecasting in its various forms. By this we mean ways to collaborate among forecasters and with potential data providers to improve forecast quality and value.

One readily thinks about open data sharing, which might be seen as the ideal way to collaborate. For several practical reasons (communication costs, size of databases, etc.), as well as other reasons we will detail, data sharing is unlikely to happen by itself. We therefore explore the basis for collaborative forecasting, with and without data sharing.

This exploration will lead us to discussion of the monetization of information and its difficulties, along with desirable properties of alternative mechanisms to support collaborative forecasting. The field of collaborative forecasting is very active: we expect substantial advances on both methodological developments and application-related problems to make a strong impact on forecasting science and practice in the coming decade.

WHY IS VALUABLE DATA DISTRIBUTED?

When mentioning data being distributed, the conventional first reaction is to understand it in a geographical sense. This is the case of a sensor network, for instance, if collecting information related to traffic and pollution in cities, or if looking at demand for a network of stores. We have been dealing with such distributed data in forecasting processes for decades, eventually using vector or spatial-process modeling, among other approaches, to get the best out of them.

However, data are also distributed in terms of *ownership*; that is, data that may be valuable to improve forecasts for a given forecast user may be collected and owned by someone else. Think for instance about networks of shoe stores in a country, owned and operated by two competing distributors. They both collect their own data about sales of their respective products (possibly also online activity related to their Web pages), which could be valuable to each other. In principle, sharing that data may improve modeling and forecasting of demand and future sales, possibly for all parties involved.

In many applications, we find similar instances of data being distributed in terms of ownership. And, in contrast to the example above (for which all data was about demand for shoe-related products), the data does not have to be of the same type or for similar variables. Consider tourism-related examples; hotels may be interested in the data of tourist attractions and local transportation companies to better predict demand. Some of the data may be numbers, some may consist of images and text. Similarly, operators of renewable-energy assets surely are interested in the data from meteorological stations and remote sensing devices in the area (again, numbers and possibly images), in order to improve their renewable-energy production forecasts.

Key Points

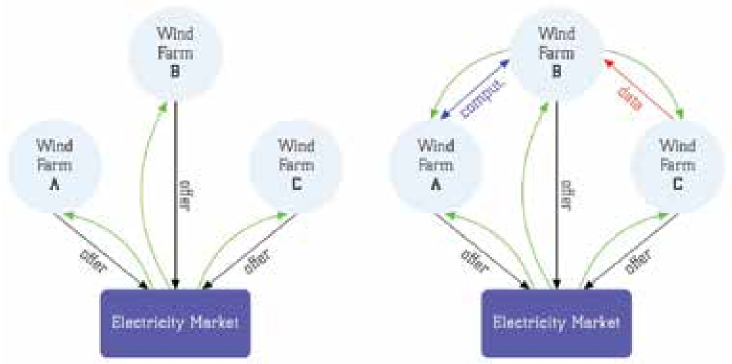
- Most forecasting tasks implicitly assume that the data can be made available in a centralized manner. This is often not the case in practice.
- Valuable data may be distributed among different owners; that is, may be collected and owned by someone else. For instance, networks of shoe stores may be owned and operated by two competing distributors, each collecting their own sales data.
- Sharing that data may allow for improved modeling and forecasting of demand and future sales, but data sharing has implications, since these data points most likely encapsulate private information about people and processes. It can be difficult to convince companies and people to share data, even if they are provided guarantees in terms of privacy protection. Today the default attitude of those who own data is not to share it.
- But there are still ways to extract value from distributed data, thus paving the way for a future of collaborative forecasting. This paper discusses four such approaches:
    1. Collaborative Analytics
    2. Data Markets
    3. Analytics Markets
    4. Prediction Markets

These require either data altruism – a willingness to make data available without compensation – or monetary incentives. Monetary compensation, if necessary, should be commensurate with the improvements the contributed data make to forecasting performance.

Let's develop this example further, based on **Figure 1**. Here, three wind farms participate in electricity markets where they must submit their supply offers in advance, hence based on forecasts. The eventual revenues from the electricity

The field of collaborative forecasting is very active: we expect substantial advances on both methodological developments and application-related problems to make a strong impact on forecasting science and practice in the coming decade.

Figure 1. Wind Farm Offerings in Electricity Markets



Without collaborative forecasting With collaborative forecasting

market are readily linked to forecast quality: in this case, increased forecast accuracy means higher revenues.

The status quo (left side of the figure) is that wind farms produce their own forecasts based on private and public information, but they do not collaborate. However, collaborative forecasting (right side of the figure) based on agreements involving either data sharing or distributed computing could benefit them all. Indeed, wind farms that improve forecast accuracy would receive higher revenues (as wind farm B in the example), while those helping would receive additional payments (as for wind farms A and C in the example). In the case where these mechanisms are properly designed, we have a win-win situation.

Many studies have shown that forecast accuracy is significantly improved if valuable data could be shared, or at least be taken advantage of. Such improvements are highly dependent upon the problem at hand and time of year and most likely range from a few percentage points to several tens of percentage points. An example in the pharmaceutical sector is found in Schachter and Ramoni (2007), and one in supply chain is Van Belle and colleagues (2021).

WHY WON'T THEY SHARE?

If benefits from potentially sharing data on forecast quality improvements are observed and documented (possibly even guaranteed), why is it that we do not see everyone sharing data, or at least trying to find ways to collaborate? Besides the obvious practical complications in setting up data-sharing channels and maintaining large databases, the situation becomes even more complex.

Sharing data has implications, since these data points most likely encapsulate private information. If the data relate to people, this information directly links to an actual privacy component. By sharing data, you then tell a bit about yourself. We have all seen that data and privacy have been a topic of increased interest over the last decade, yielding the now-famous GDPR (General Data Protection Regulation) in Europe, for instance. Even overlooking this type of regulation, many are reluctant to share data if they feel there is any likelihood of this yielding a leakage in personal privacy.

Importantly, some of the valuable data we are thinking of here are not linked to people but to private information of direct value to a process or a business instead. As a consequence, we'd intuitively expect that sharing that information would expose business practices, inadvertently making public some confidential information and most likely leading to a loss of competitiveness, reflected in market share or revenue. In the network of shoe stores example, one could imagine

If benefits from potentially sharing data on forecast quality improvements are observed and documented (possibly even guaranteed), why is it that we do not see everyone sharing data, or at least trying to find ways to collaborate?

Being in a competitive environment most often is the root for this reluctance to share data, whatever the potential mutual benefits.

that the data shared to improve forecasts would expose information about the sales of the competitor. Being in a competitive environment most often is the root for this reluctance to share data, whatever the potential mutual benefits.

Analysts and forecasters in different fields have all noticed how difficult it is to convince companies and people to share data, even if they are transparent with how the data will be used and provide them guarantees in terms of privacy protection. Simply speaking, currently the default attitude of those who own data is not to share it.

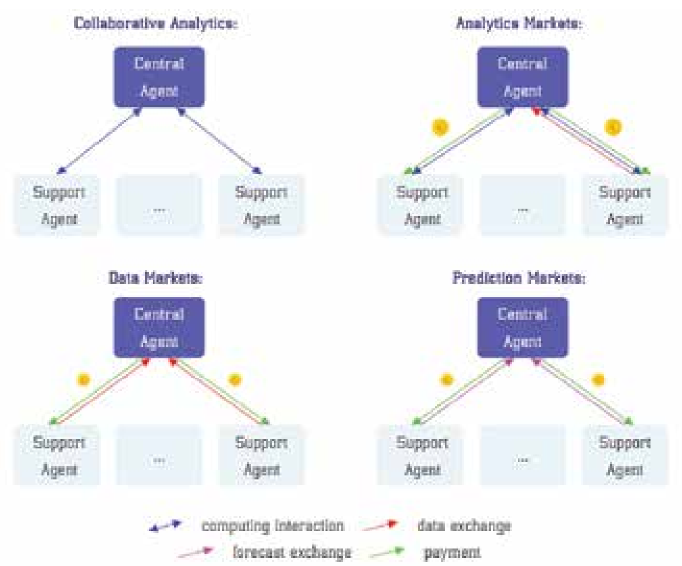
HOW TO GET VALUE OUT OF DISTRIBUTED DATA?

If those who collect and own valuable data are reticent to share, we must find ways to incentivize them. Over the last five to 10 years, the scientific literature is burgeoning with ideas to support collaborative forecasting. Actually, forecasters should toot their own horns here, since the concepts of the wisdom of crowds and of prediction markets are early forms of what is further developed today into the field of collaborative forecasting. In addition, some claim that the recent focus on blockchain and more generally distributed ledger technologies will be of great help, since they comprise an ideal backbone for distributed and linked databases, while allowing for smart contracts as a basis for monetary compensation.

All the following approaches to grasp value from distributed data require an internet-based platform to organize communication among agents (forecasters and data owners), perform the necessary analytics, and possibly arrange for monetary compensation. You can think of these platforms as blending the functionality of forecast competition platforms (e.g., Kaggle), market platforms (e.g., Nasdaq, as one example among many), and distributed computation platforms (e.g., climateprediction.net, among many others). In all cases, the forecaster who is posting the task on the platform is referred to as the "central agent," while those providing support through collaboration based on their data and computation are referred to as the "support agents."

Globally, we see four types of complementary, and possibly linked, approaches to get value out of distributed data (illustrated in **Figure 2**), which may pave the way for a collaborative forecasting future:

Figure 2. Approaches to Collaborative Forecasting Based on Internet Platforms



Collaborative Analytics and Modeling

Instead of centralizing data to perform analytics and modeling for forecasting, we can distribute the learning and forecasting tasks. This involves distributed computing and optimization, for which approaches are necessarily *iterative*, involving repeated steps and review.

In the present case, it translates to having iterative communication between the platform (representing the central agent) and the support agents, as well as local computation at both levels. An instance of this approach is the widely considered case of *federated learning*, originated by Google in 2016, which has now attracted much attention.

Federated learning is based on the idea that learning is distributed, not centralized, while having some degree of coordination (hence, the term "federated"). Federated learning was originally rooted in altruism; that is, those who collect and own data would be willing to help each other, but without directly sharing the data.

Distributing the learning and forecasting tasks instead may then be deemed an appropriate approach. There is no monetary compensation involved, though. Today, many of the leading analytics players (e.g., IBM, Microsoft, NVIDIA, etc.) have some form of federated learning in their offering portfolios, while new unicorns like Owkin have based their original business models on federated learning.

Data Markets

There are many applications where analysts and forecasters still find it better to work with centralized data, which necessitates finding other ways to share their distributed data. *Data markets* can play a role here as they allow data (either raw or after feature engineering) to be exchanged and priced through a common marketplace such as a pool.

In this arrangement, the data are treated as a commodity or a good, for which payment implies transfer of ownership. Bilateral data markets have been around for a while; for example, we've seen meteorological data companies selling weather information, as well as companies like Bloomberg selling market intelligence data. The data markets we're considering here differ from these in that they are multibilateral or lie within a pool of a potentially large number of players, continuously running to reflect the streaming nature of data.

Data markets involve a single communication step, limited computation, and an eventual data exchange. Implementation at first may appear to be straightforward, based on monetary incentives for data sharing. However, with data being a special commodity (it can be reproduced and can be sold several times, for instance), designing such data markets is challenging. A notorious example of a failed data market is that of the City Data Exchange hosted by Copenhagen in Denmark over the period 2016-18. ***https://cphsolutionslab.dk/media/site/1837671186-1601734920/city-data-exchange-cde-lessons-learned-from-a-public-private-data-collaboration.pdf***

New data markets are currently being proposed, some based on *distributed ledger technology*; one example is the IOTA data marketplace (***https://wiki.iota.org/blueprints/data-marketplace/overview***).

Analytics Markets

Analytics markets offer a way to blend the rationale of collaborative analytics with the inducements of monetary compensation, as in data markets. The central agent defines an analytics task that is useful for learning and forecasting, such as regression, and posts this task on the analytics platform. Others (the support agents) can then provide data to the platform. It is even possible to blend data sharing and distributed computation, while accommodating privacy concerns.

These types of markets are not as mature as the other three cases, and are now the focus of intensive research and development, for instance in the frame of the EU project Smart4RES (***www.smart4res.eu***).

The platform assesses whether the analytics task is performed better thanks to those additional data. If that is the case, it triggers a payment from the central agent to the support agents. The payment is directly linked to how much the data improved the analytics task as measured, for example, by improved forecast accuracy. Communication and computation needs may vary widely depending on the type of analytics market and their implementation.

Prediction Markets

Possibly the most pragmatic approach to implementing collaborative forecasting is that of prediction markets. Here, the central agent posts a forecasting task on the platform, possibly having already produced a forecast. All support agents then keep their data private and make their own best forecasts.

All these forecasts are gathered onto the platform, which applies an aggregation operator to combine them into a single optimal forecast which then is delivered to the central agent. Finally, appropriate scoring and allocation functions are used to assess the contribution of individual forecasts to the quality of the aggregate forecast and to decide on a resulting monetary compensation for that contribution.

In prediction markets, computations are performed at the level of both the platform and the support agents, with communication between. Part of the appeal here is that they do not require multiple iterations, as in the case of collaborative analytics and analytics markets.

There are many examples of prediction markets, some of which have long been active (e.g., the Iowa electronic markets, iemweb.biz.uiowa.edu) while some of them appeared following the development of distributed ledger technologies (e.g., Augur, augur.net). However, while prediction markets have become common platforms for political forecasts, they have received limited interest in the business world (Wolfram, 2019).

These various approaches offer flexibility in implementation for different needs with respect to communication, computing, and complexity. For instance, an approach based on federated learning may imply a large number of iterations between the platform and those contributing their local computation; prediction markets do not require such iterations, but at the expense of a potentially lower quality of the resulting final forecasts.

DESIRABLE PROPERTIES AND CHALLENGES AHEAD

Whenever considering collaboration based on coordination and monetization, the field of *mechanism design* ensures that the proposed approach will provide the right incentives for those involved, while yielding the desired outcome. In the case of collaborative forecasting, there are many aspects to consider, since information (either data or forecasts) is a special commodity. The properties we would like to have include

1. **Budget balance** – the payment by a forecaster or forecast user who obtained an improved forecast determines the monetary compensations to the contributors.
2. **A zero element** – if there isn't an improvement in forecast quality, no monetary compensation is given.
3. **Symmetry** – if permuting the names of the contributors, the outcome should be the same, in terms of monetary compensation.
4. **Individual rationality** – contributors should perceive the possibility of receiving a monetary compensation if their data contributes to improvement in forecast quality.
5. **Truthfulness** – contributors only get their best monetary compensation if giving their best data, information, or forecast.

There may be additional properties that depend on the specifics of the mechanism in use. Those listed above involve monetary compensation, and some of these may be more difficult to achieve than others. Collaborative analytics, being without compensation, may require altruism on the part of all agents. Indeed, if not receiving monetary compensation to help improve forecasts, why would anyone provide their best information?

Truthfulness is a crucial property; without it there may be no incentive to invest

A paradigm shift from centralized to collaborative forecasting could give rise to a wealth of new business models.

in improving the quality and information content of the data to be shared.

Many approaches can be considered in order to achieve these properties: they can be at the core of the mechanism design itself or result from contracts and insurance policies. In addition to monetary properties, consideration of privacy preservation can be embedded into the market, using differential privacy, k-anonymity, or ad hoc data-exchange protocols.

NEW BUSINESS MODELS

A paradigm shift from centralized to collaborative forecasting could give rise to a wealth of new business models. But first the collaborative forecasting platforms need to be made scalable, in order to host large forecasting tasks, while remaining user friendly to discourage barriers to entry.

Consequently, one can imagine that these platforms will charge forecasters for the service, in the form of (i) one-off payment per forecasting task; (ii) recurrent payment for the case of repetitive tasks (e.g., in the case of online learning); (iii) all-inclusive subscriptions. Within today's platform economy, and in view of the number of forecasting tasks that could be hosted on such platforms, revenues could be extremely large.

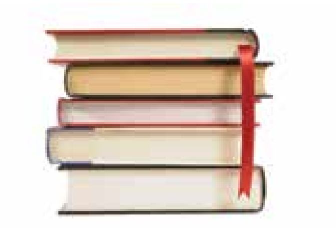
The collaborative platforms can be seen as an extension of current approaches to bilateral data-service agreements (e.g., between weather forecast providers and their users). Such an evolution from ad hoc bilateral agreement to platforms based on a pool or multibilateral agreements for standardized products has already been witnessed, in the case of electric energy.

Contributors who help to improve forecast accuracy by monetizing their data, analytics contributions, and forecasts will receive monetary compensation for their contribution. Eventually, this may reveal the value of each and every data point they collect, yielding a stable new revenue stream for various businesses (and possibly private individuals). Similarly, prospective studies about the potential value of data through such collaborative forecasting platforms could trigger decisions to start collecting data that was not collected previously.

FURTHER READINGS

We have kept this article nontechnical, and readers interested in the topic may want to pursue the more technical concepts involved in the design of these collaborative markets. An excellent starting point is the paper by Bergemann and Bonatti (2019), which also discusses recent advances in markets for data (and information more generally).

Two examples of analytics markets are described by Agarwal and colleagues (2019) and by Pinson and colleagues (2022). The first places more focus on the pricing mechanism and issues with the fact that data may be replicated and sold several times. The second concentrates on the proposal of a market for regression-analytics tasks, such as for batch and online learning, for deterministic and probabilistic forecasts, as well as in-sample (training) and out-of-sample (forecasting) tasks.



Rasouli and Jordan (2021) develop a compelling argument involving exchange of some data for other data, in contrast to exchange of data for monetary compensation. Those looking for recent developments with decentralized prediction markets based on distributed ledger technologies should see the blueprint for Augur, by Peterson and colleagues (2020).

Lastly, even though there are now hundreds of papers examining federated learning and alternative approaches to decentralized learning, the interested reader should start with the blog post by McMahan and Ramage (2017) that gives a gentle introduction to the topic. Federated learning is seen as blending collaborative analytics and analytics markets, allowing for monetary compensation, while maintaining privacy protections.

REFERENCES

Agarwal, A., Dahleh, M. & Sarkar, T. (2019). A Marketplace for Data: An Algorithmic Solution. In Proceedings of the ACM EC'19: ACM Conference on Economics and Computation, Phoenix (AZ, USA), 701–726.

Bergemann, D. & Bonatti A. (2019). Markets for Information: An Introduction, *Annual Review of Economics*, 11, 85–107.

McMahan, H.M. & Ramage, D. (2017). Federated Learning: Collaborative Machine Learning Without Centralized Training Data. ***https://research.googleblog.com/2017/04/federated-learning-collaborative.html***, accessed on 29 July 2022.

Peterson, J., Krug, J., Zoltu, M., Williams, A.K. & Alexander, S. (2020). Augur: A Decentralized Oracle and Prediction Market Platform, Arxiv preprint, available online, ***arXiv:1501.01042***, accessed on 29 July 2022.

Pinson, P., Han, L. & Kazempour, J. (2022). Regression Markets and Applications to Energy Forecasting, TOP, available online: ***https://link.springer.com/article/10.1007/s11750-022-00631-7***, accessed on 29 July 2022.

Rasouli, M., Jordan, M.I. (2021). Data Sharing Markets, Arxiv preprint, available online, ***arXiv:2107.08630***, accessed on 29 July 2022.

Schachter, A. & Ramoni, M.(2007). Clinical Forecasting in Drug Development, *Nature Reviews Drug Discovery*, V6, 107–108.

Van Belle, J., Guns, T. & Verbeke, W. (2021). Using Shared Sell-Through Data to Forecast Wholesaler Demand in Multi-Echelon Supply Chains, *European Journal of Operational Research*, V288:2, 466-479.

Wolfram, T. (2019). Benefits and Challenges of Corporate Prediction Markets, *Foresight*, Issue 54 (Summer), 29-36.

Acknowledgements are due to many who interacted with me on the topic of collaborative forecasting, including Jalal Kazempour and Liyang Han at the Technical University of Denmark; Carla Goncalves and Ricardo Bessa at INESC Porto (Portugal); Aitazaz Raja and Sergio Grammatico at DTU Delft (the Netherlands); Shashi Pandey and Petar Poposki at Aalborg University (Denmark); and Aida Kharman, Christian Jursitzky, Pietro Ferraro, and Robert Shorten at Imperial College London (UK).

Pierre Pinson is the Chair of Data-centric Design Engineering at Imperial College London (UK) and a Chief Scientist at Halfspace, Copenhagen (Denmark). He is also the Editor in Chief of *The International Journal of Forecasting*. He has made extensive contributions to probabilistic forecasting, forecast verification, as well as applications to such areas as energy and meteorology. More recently, he has focused on incentives for data sharing and monetization of information; for instance, through data, regression, and prediction markets.

p.pinson@imperial.ac.uk

Commentary on "To Share or Not to Share"
Asymmetry of Data Ownership

NIELS VAN HOVE

In Pierre Pinson's article "To Share or Not to Share: The Future of Collaborative Forecasting," the comparison between data and big oil is striking. Just like oil, data is not evenly distributed among geographies and owners. Few companies, in even fewer countries, own or have access to most of the consumer data generated in this world.

In the digital winner-take-all markets, we have two mobile operating systems, owned by two companies: Apple and Google. Google has a 90%+ share in online search. In E-commerce, retail sales are dominated by Amazon, with a U.S. share of 57% in 2021. Worldwide digital advertising revenue is shared among a handful of companies.

What do these businesses have in common? They use a data-centric business model, and they are the first point of contact to a consumer in the supply chain. They have access to user data from billions of people, who create more and more personal data every day.

Similar to big oil, data ownership is distributed asymmetrically across the supply chain. OPEC has controlled oil output for decades for their own good. It should be no surprise the digital giants focus on data monetization rather than data altruism. As with big oil, the digital giants go a long way to protect their data and business model.

COLLABORATIVE PLANNING, FORECASTING, AND REPLENISHMENT

We have considerable supply-chain experience with distributed data through Collaborative Planning, Forecasting & Replenishment (CPFR), which has been around since the 1990s as a process developed to reduce supply-chain costs among partners in a single supply chain (***https://www.supplychainsecrets.com/an-introduction-to-cpfr-in-the-supply-chain/***).

The best application of CPFR I've encountered was in 2005, when working for a meat manufacturer that received point-of-sale (POS) data four times a day (10am, noon, 2pm, 4pm) from over 600 stores from the largest retailer in the Netherlands. POS data shared in the morning was used to forecast remaining meat demand for the day, and the afternoon production schedule could be adapted. POS data from the afternoon sales was used to adapt the production schedule for the next day. The meat manufacturer could only react on such short notice because of the POS data from the retailer. It was a win-win situation.

This collaboration made business sense, but it seems to be more an exception than the rule. Although there are reports that CPFR has improved financial and operational performance (Hill and colleagues, 2018), I don't believe CPFR lives up to its potential. One of the reasons is that data in the supply chain is distributed asymmetrically towards the retailer, who has the first point of contact with the customer and owns most of the valuable consumer data.

In Australia, manufacturers can get access to POS scan sales from the largest retailers – but it comes at a cost. The retailer monetizes that data. On top of this, manufacturers can purchase data from market-data aggregators like Nielsen and Aztec, who make selling data and insights their core business.

DATA SUPPLIERS

Recently, data-centric businesses like Fourkites and Project44 have been growing at a very fast pace. They have a global network that provides supply-chain data from ocean freight forwarders and transporters, that they then make available for integration with your own business. It is a great way to use data and create supply-chain visibility.

Although it can be used for short-term forecasting, it is not collaboration. Collaboration from data altruism could happen for several reasons: when power distribution among collaborators is equal; when sharing only small or old datasets; for trials or training, research, charities, or social good; or when there is a strong shared purpose or existential threat among the collaborators.

- I see an example of altruistic data sharing for research purposes. Since 2016 there is even an agreed FAIR (findable, accessibility, interoperability and reusability) data code. ***https://ardc.edu.au/resources/aboutdata/fair-data/***
- The Boston Consulting Group discusses data sharing arrangements for sustainable development and other societal challenges. ***https://www.bcg.com/publications/2021/broad-data-sharing-models***
- TomTom freely publicizes traffic congestion and emissions data for 50 cities around the world. ***https://www.tomtom.com/en_gb/traffic-index/***
- In August, the European Parliament approved new rules boosting intra-EU data sharing to promote data altruism in support of research and health care and to fight climate change.

  https://www.europarl.europa.eu/news/en/press-room/20220401IPR26534/data-governance-parliament-approves-new-rules-boosting-intra-eu-data-sharing
- "Citizen sensing" crowd sources–from digital sensors of urban climates – allow individuals to collect data for the purpose of fact-finding and policymaking.

  https://citizensensing.itn.liu.se/

CONCLUSION

Theoretically, there should be a Nash equilibrium that would make businesses share data to achieve mutual gains. However, if we want to use large datasets for forecasting and analysis that feed into real-life business strategies and commercial business decisions in a competitive environment, we had better get used to the notion that we have to pay for it.

REFERENCES

Hill, C., Zhang, G. & Miller, K. (2018). Collaborative Planning, Forecasting, and Replenishment & Firm Performance: An Empirical Evaluation, *International Journal of Production Economics*, V196, February, 12-23.

Niels van Hove is a Client Engagement Principal at Aera Technology. He uses his more than 20 years of supply-chain experience to guide clients toward a more autonomous process. See our "Forecaster in the Field" interview with Niels in the Summer 2016 issue.

niels.vanhove@aeratechnology.com

Commentary on "To Share or Not to Share"
Legal Ramifications and FVA of Data Sharing

ROBERT STEVENS

Peter Norvig, Google's Chief Scientist, famously said, "We don't have better algorithms. We just have more data" (Cleland, 2011). Having more data is one element of Pierre Pinson's advocacy for "rethinking forecasting processes with modern communication, distributed computation, and a market component." Yet in a sense this is not new. Rather, Pierre's viewpoint represents a modern-day take on ideas that have been around for a long time, but reexamined in light of the latest technology developments and business thinking.

SYNDICATED DATA AND DEMAND SIGNAL REPOSITORIES

A view of data outside one's company dates to the 1930s, when Nielsen introduced the first syndicated data service for packaged-goods companies. This was based on store audits for the entire category. For the first time, managers could see the marketplace performance (and eventually the marketing execution) of their competitors. Yet, this was not a data-sharing service, as Pierre proposes. The privacy issues he describes have prevented such a data marketplace to this day.

The 2000s have seen the emergence of the *demand signal repository* (DSR), where manufacturers and retailers can combine retail point-of-sale data with the manufacturers' shipment and other related data. (See ***https://www.techtarget.com/searcherp/definition/Demand-signal-repository-DSR***.) A DSR enables Collaborative Planning, Forecasting, and Replenishment (CPFR); but again, it is not in the spirit of a distributed data market and does not rise to the level of collaborative analytics or federated learning. Rather, it helps channel partners to mutually improve supply-chain performance.

Tom Davenport, analytics expert and Chairman of my company First Analytics, once asked: "Who owns your data exhaust?" (Davenport, 2013). His idea was that information by-products gathered in the course of a company's normal business (the exhaust) could be used to "informationalize" a business. The company could develop products and services based on information – so called "data products."

LEGAL RAMIFICATIONS OF DATA SHARING

But the final point in Davenport's article is something that Pierre has not mentioned: the legal ramifications of data sharing. Davenport wrote: "We're likely to see lots more data products in the future, and owning the rights to your data exhaust is critical to developing and introducing them without lots of lawsuits."

The collection of data by associated partners already comes with complex agreements with respect to the obligations of the partners. Assuring those obligations are met and the consolidated data are distributed to other third parties represents a liability. For example, in the course of providing data to a third party, besides meeting those obligations, what if erroneous data is provided that causes a bad forecast and incurs severe financial cost?

The collection of data by associated partners comes with complex agreements with respect to the obligations of the partners. For example, what if erroneous data is provided that causes a bad forecast and incurs severe financial cost?

FVA OF DATA SHARING

Even with legal hurdles removed, however, marketplace participants will need to agree on a framework and methodology for pricing the data. The foundation of this framework lies in forecast-value-added analysis. FVA assesses the improvement in accuracy attributable to each step and participant in a forecasting process (Gilliland, 2013). In the present context, FVA analysis would focus on accuracy improvements due to shared data elements, rather than improvement due to process activities such as managerial forecast overrides.

The packaged-goods industry is very familiar with marketing mix or marketing attribution models to quantify the return on investment of various marketing vehicles. These models estimate the incremental sales due to the marketing tactic used and compare these against the costs of execution. Similar methodology can be applied to the valuing of data, in conjunction with FVA.

CONCLUSION

The legal and valuation issues, along with technology considerations that have not been discussed here, can be overcome. The overarching challenge is with so-called "change management" – steering managerial culture and practices away from "the way we have always done it." Perhaps we can begin by developing event tracks and even keynote talks at industry conferences. Digital blogs and podcasts and traditional trade journalists are other forums for addressing data-sharing issues. This would be the soft-selling.

For Pierre's ideas to gain currency, however, profit-seeking needs to be the motivator. Economists tell us that when early adopters begin to reap profits from an innovation, others will follow. Marketplace participants with a focus on innovation and entrepreneurs need to take the initial risk of the commercialization of these ideas. Others will follow the money or, if for no other reason, to not be left behind.

REFERENCES

Cleland, S. (2011, October 3). Google's "Infringenovation" Secrets, *Forbes*.

Davenport, T. H. (2013, November 20). Who Owns Your Data Exhaust? *The Wall Steet Journal*.

Gilliland, M. (2013). FVA: A Reality Check on Forecasting Practices, *Foresight*, Issue 29, 14-18.

Robert Stevens, as part of the leadership team at First Analytics, helps companies develop and execute programs to cultivate their analytics competency. Rob brings experience to bear stemming from 30 years as an analytics professional. His career has consisted of consulting, product development, client service, technical and sales roles within software, market research and advisory services firms.

RStevens@firstanalytics.com

Commentary on "To Share or Not to Share"
Federated Data and Learning in the Supply Chain

RAM GANESHAN

Pierre Pinson's article in this issue and Niels van Hove's subsequent commentary point to two rapidly evolving but conflicting trends in the supply chain.

- First, there is copious data generated in the supply chain that could potentially increase revenue and reduce costs for all the players.
- Second, the data is "siloed"– that is, owned by different entities in the supply chain. Privacy concerns, differing governance structures between entities, inter- and intra-organization changes warranted, poor technological infrastructure, misaligned incentives, and regulatory challenges are all stumbling blocks to sharing the data.

I agree with Pierre's general hypothesis that the mechanisms he describes have emerged as a new paradigm that could overcome many of these problems and promote collaboration in the supply chain. In this commentary, I expand on his article to imagine how these new collaborative tools can improve the supply-chain performance of all entities.

KEY ASSUMPTIONS

These collaborative mechanisms are still in their infancy. Their application assumes that

1. A firm has already leveraged its first-party data: that is, it has reached a point of diminishing results with the data it can collect.
2. Additional data residing elsewhere in the supply chain will substantially improve insights.
3. Specifically for collaborative analytics applications, the firm has deployed machine-learning and artificial-intelligence capabilities across its supply chain.

The number of firms that fit this description is small. And still fewer are using the mechanisms Pierre describes.

ADVANTAGES OF COLLABORATIVE MECHANISMS

So, what can these collaborative mechanisms buy us? I point to five advantages – these are related, and the tools they use significantly overlap.

1. **Hyper-Personalization.**

   Personalization typically uses the customer's attributes, such as past purchasing history, demographic information, geo-localization, browsing history, credit card and investment transactions. Information is often quite granular, tagging behavior by time of day. This information typically sits in different databases and is owned by separate entities.

   Federating this data by any of the mechanisms would help companies offer the right customer the right product at the right price and time. Imagine a shopper walking into a store – store beacons (or cameras) recognize the shopper. Based on their recent shopping, spending, and browsing behavior, customized offers from multiple brands are available at different points in their shopping experience. Indeed, such personalization is a marketer's holy grail.

2. **Pattern Recognition and Trend Identification.** The current practice that firms use is applying advanced statistical tools and machine-learning algorithms to extract patterns from their databases. One example in retail is *basket analysis*, a data-mining technique used by retailers to reveal product groupings, as well as products that are likely to be purchased

together. Retailers and manufacturers can apply this analysis to make better operational decisions.

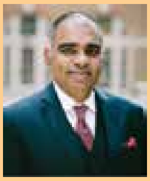
In the digital realm, *recommendation engines* – either other products to buy (e.g., at Amazon.com) or a movie suggestion from Netflix – are everyday examples. Cross-company data or enriched data can make these recommendations much more precise. Another ripe area of application is in health care. Different health-care institutions can federate data and share parameters that affect patient outcomes, without sharing patient data.

3. **Predictive Modeling.** This is a statistical technique to algorithmically predict future outcomes using various data sources (historical data, social media, text mining, and the like). *Failure and fatigue analysis* is a common application – sensor data streams can predict machine failures. Both suppliers and manufacturers have a mutual interest in keeping these machines running and a strong incentive to share such sensor data. *Fraud analytics* is another application in the financial-services industry. Credit card companies and banks have a mutual interest in stopping fraudulent transactions and thus share a strong incentive to federate the data.
4. **Text Mining, Audio, and Video Analytics.** These extract information from text sources such as blog posts, social media, forums, emails, and audio and video feeds. Statistical tools, machine learning, natural language processing, and AI/computer vision are commonly used tools. One well-known application is autonomous driving. While many companies have developed sophisticated computer vision models (Tesla, for example), they have access only to their own databases. While large, these databases are also incomplete. Federating data between companies should make such systems more intelligent.
5. **Track and Trace.** *Track and trace* refers to the ability of a firm to have visibility in the supply chain. The pandemic has heightened its importance – to vet a supplier, ensure the suppliers comply with firm values, and track operations and the eventual flow of parts or products from the supplier. Since firms share suppliers (highly concentrated, for example, in the semiconductor industry), inter-firm sharing of suppliers' data could avert significant supply-chain disruptions.

PITFALLS

While collaborative mechanisms have significant potential, they have some major pitfalls.

- First, it is unclear if a collaborative model will work for all entities. While technological and methodological advancements in decentralized, federated learning models promise high customization for each entity while maintaining privacy, I know no use-case scenarios demonstrating that.
- Second, firms may be wary of losing competitive advantage. They may come into the federation with differing maturity levels in data and computing resources. So, such collaborative mechanisms may not be long lasting without sufficiently aligned motivations.
- Finally, tangible benefits of collaboration are required to justify the cost of data and computing resources. Such justification could take a few years, given how early we are in the implementation – something the management may not have patience for in the short run.



Ram Ganeshan holds the D. Hillsdon Ryan Professorship of Business at the College of William & Mary. He is the coeditor of two books: *Quantitative Models for Supply Chain Management* and *New Directions in Supply Chain Management: Technology, Strategy, and Implementation*. In 2001, the Production & Operations Management Society (POMS) awarded him the prestigious Wickham Skinner Award for Early Career Accomplishments for his research on how supply chains can be efficiently managed. In addition to his academic appointment, Ram is active in several corporate advisory boards, helping these companies embrace the latest developments in supply-chain management and digital strategies.

rxgane@wm.edu

Commentary on "To Share or Not to Share" Third-Party Data Providers

SUJIT SINGH

When I consider collaborative forecasting, I am often thinking of collaboration between different departments of the same company. The average of collective inputs usually is better than any of the individual inputs and, when done well, the idea of "two heads is better than one" really proves itself.

Pierre Pinson has extended this idea to collaboration among organizations. It has great appeal and I am convinced it can create value. The overall market trends that are visible to one company in isolation can be easily enhanced if data visible to other companies can be applied.

However, the challenges Pierre mentions are very real. After all, companies are selfish, profit-making entities, and it's logical for them to ask the all-important question: What's in it for me? Even if a company is sold on the benefits, it might still not share its data for fear that someone else could use that data to win market share or spread rumors. Between "what's in it for me" and "I see how this can hurt me," people typically end up not sharing.

Sujit Singh is the COO of Arkieva, a supply-chain solutions company responsible for managing product delivery and implementations services, customer relationships, and the day-to-day operations of the corporation. He has served on the *Foresight* Advisory Board for more than a decade.

ssingh@arkieva.com

Companies in some industries have formed small entities to mitigate the challenges of sharing data. For example, more than 50 companies have joined hands to form the Alliance to End Public Waste (AEPW) to solve the problem of plastic waste through collaborative and collective action (***https://endplasticwaste.org/en/about#map12***). These consortia do the job of collecting data, creating policies, and establishing standards, and these have been followed by their member companies. Can such a structure be created on the demand side of things? I suspect there already might be such a consortium, although I'm not aware of one.

The challenges for sharing data among companies are very valid. In my mind, the best opportunity in this area is with the third parties (mostly Software as a Service providers) who process large amounts of data from many competing companies and can include ERP vendors, Purchase Order processors, and Third Party Logistics (3PL) ***https://www.datamation.com/cloud/saas-companies***/

These vendors would have to exercise due caution to ensure proper masking of any proprietary data to protect the underlying companies. However, when third parties process data, they have the opportunity to look for trends, seasonality, and other patterns in the aggregate. Based on their existing agreements, they can aggregate and publish the data.

This approach has its own limitations. It might be good at a macro level (i.e., polymer sales are up), but not helpful at the micro level (which colors are trending). But one cannot allow perfect to be the enemy of the good.