To Share or Not To Share?

Alternative views on a 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 may 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.*

Key Points

- All tasks forecasters generally concentrate on implicitly assume that the data can be made available in a centralized manner. But 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 modelling and forecasting of demand and future sales but data sharing has implications, since these datapoints 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 to not 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 w/o compensation -- or monetary incentives. Monetary compensation, if necessary, should be commensurate with the improvements the contributed data make to forecasting performance.

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 quintillion bytes per day (a quintillion being a number with 18 zeros, i.e., a billion of billions!). In its edition of the 6th of May 2017, *The Economist* wrote that “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 data. Quantitative analysts and forecasters consequently focus on challenges related to data cleaning, feature engineering and selection, model building and validation, etc. This is first based on the assumption such that all data can be made available in a centralized manner. It is often not the case in practice.

If the data cannot be gathered and centralized, does that mean that it is not possible to get value out of all that data? Surely not. However, this calls for a paradigm shift, by going towards collaborative forecasting, in its various forms. By collaborative forecasting, we mean ways to collaborate among forecasters and with potential data providers, in order 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 (communication costs, size of databases, etc.) or other reasons which we will detail in the following, this is unlikely to happen by itself. We therefore explore the basis for collaborative forecasting, with and without data sharing.

Importantly, this will lead us to discussing monetization of information and its difficulties, as well as desirable properties of alternative mechanisms one may think of 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?

Conventionally, when mentioning data being distributed, the 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 already, eventually using vector or spatial process modelling among other approaches to get the best out of the data.

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 2 competing distributors. They both collect their own data about sales of their respective products (possibly also online activity related to their webpages), which could be valuable to each other. In principle, if sharing that data, it may allow to improve modelling 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 shoe store example in the above (for which all data was about demand for shoe-related products), the data does not have to be of the same type and for similar variables. Considering tourism-related examples, hotels may be interested in the data of touristic attractions and local transportation companies to better predict demand. Some of the data may be numbers, some of it may consist in 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.

Let us develop further this renewable energy example, based on Figure 1. In this example, 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 market are readily linked to forecast quality: 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, and they do not collaborate. However, if engaging in collaborative forecasting (right side of the figure), based on agreements involving either data sharing or distributed computing, they could all benefit. Indeed, wind farms improving their forecasts would receive higher revenues by improving forecast accuracy (as for 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, this would yield win-win situations.

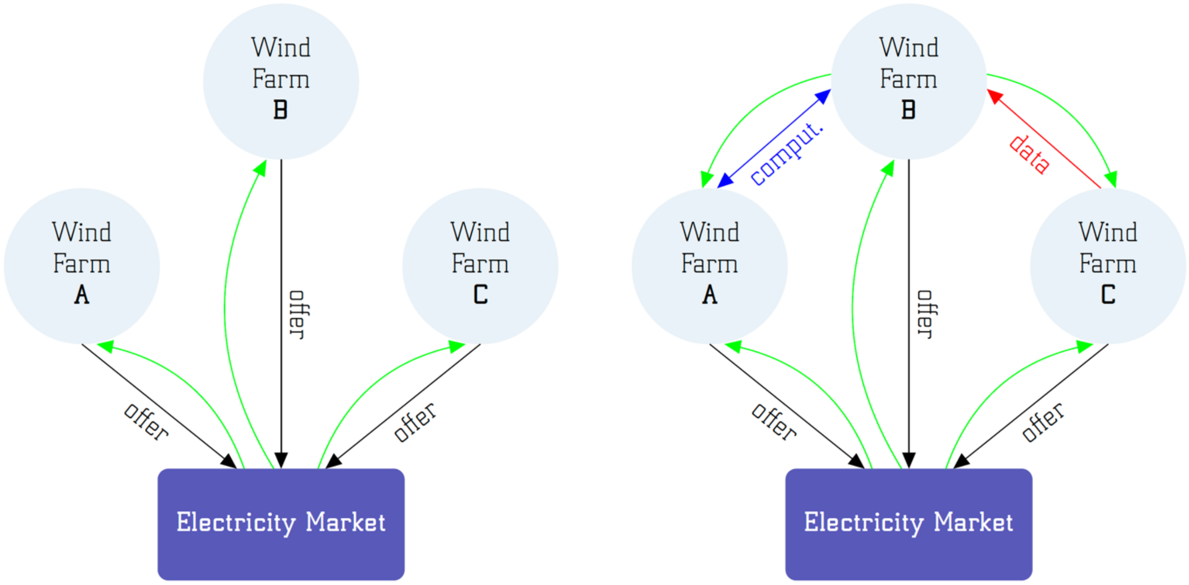

Figure 1: *Wind farms offering in electricity markets (left: status quo, i.e., without collaborative forecasting – right: foreseen as a future and alternative setup, i.e., with collaborative forecasting)*

Many studies have shown that forecast accuracy is significantly improved if valuable data could be shared, or at least taken advantage of. Such improvements are highly dependent upon the problem at hand and time of the year but most likely range from a few percentage points to several tens of percentage points.

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 based on their data? Besides the obvious practical complications in setting up data sharing channels, maintaining large databases, etc., the reasons are to be found elsewhere.

Sharing data has implications, since these datapoints most likely encapsulate private information. If thinking of data collected in relation to people, this private 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) act in Europe for instance. Even if overlooking the advent of such a piece 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 is not linked to people and infringing their privacy. The private data is linked to private information of directly value to a process or a business instead. As a consequence, it is intuitively expected that sharing that information exposes business practices, makes inadvertently public some confidential business information, and most likely leads to a loss of competitivity on a market (market share or revenue). Getting back to the network of shoe stores example, one would easily imagine that the data shared to improve forecasts would readily provide all 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 resulting mutual benefits.

Analysts and forecasters in different fields have all noticed how difficult it is to convinced companies and people to share data, even if being transparent with how the data will be used, while providing them guarantees in terms of privacy protection. Simply speaking, as of today the default attitude of those who own data is to not share it.

HOW TO GET VALUE OUT OF DISTRIBUTED DATA?

If those who collect and own valuable data are reticent to share, it does not mean that it is impossible to find ways to incentivize them to do so. Over the last 5-10 years, the scientific literature is burgeoning with alternative ideas to support collaborative forecasting. And, actually, forecasters should toot their own horn since the concepts of wisdom of the crowd and prediction markets are early forms of what is further developed today in 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 comprising an ideal backbone for these alternative approaches in the form of distributed and linked databases, while allowing for smart contracts as a basis for monetary compensation.

All the following approaches to grasp the value from distributed data require an internet-based platform, to organize communication among agents (forecasters and data owners), perform the necessary analytics, and possibly organize for monetary compensation. You can think of these platforms as blending the functionality of forecast competition platforms (e.g., Kaggle and the likes), 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**), and which may pave the way for a collaborative forecasting future:

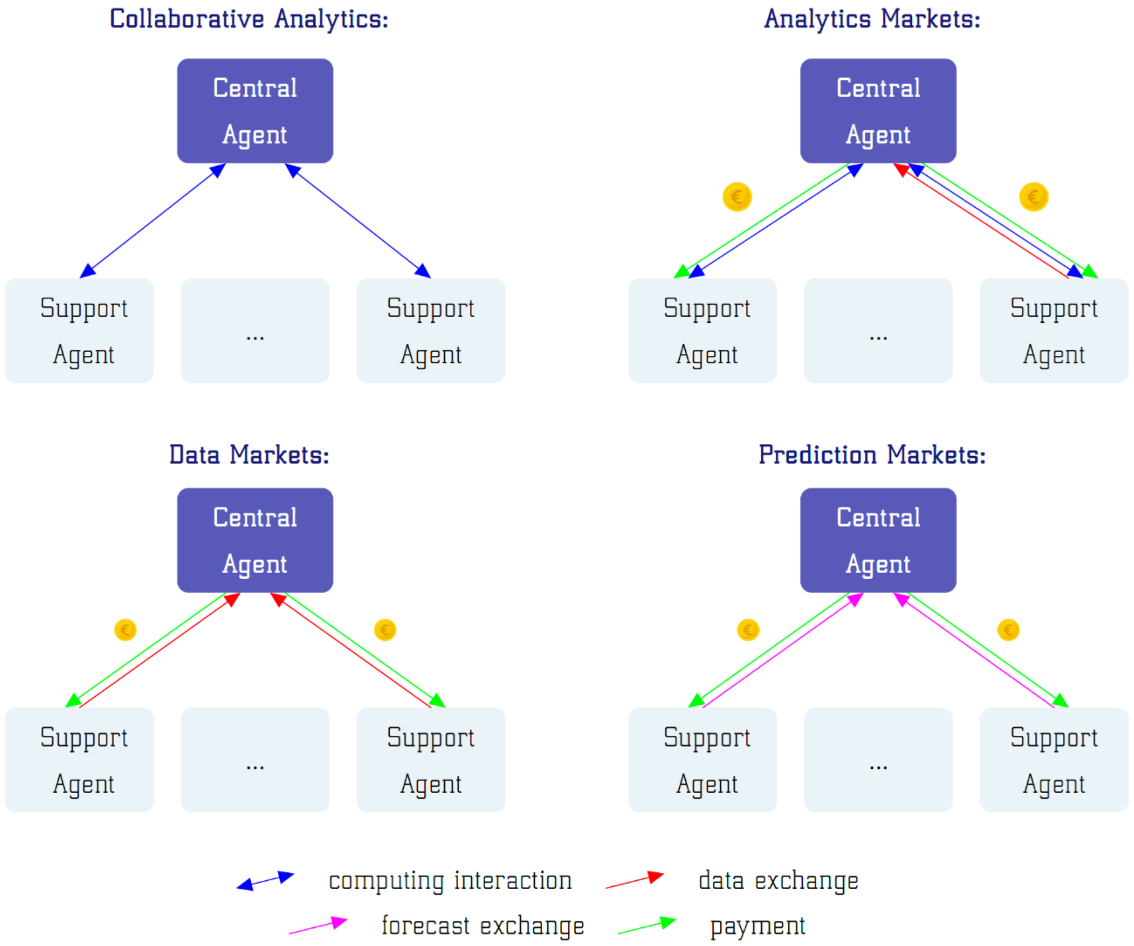

***Figure 2:** Various approaches to collaborative forecasting based on internet-based platforms.*

Collaborative Analytics

Instead of centralizing data to perform analytics and modelling for forecasting, one can distribute the learning and forecasting tasks. This then involves distributed computing and optimization, for which approaches are necessarily iterative. 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”, originally pushed forward by Google in 2016, and which has attracted immense attention since then. Federated learning is a way to denote this idea that the learning is not centralized, but based on a distributed approach instead, while having some level of coordination (hence, the term “federated”). Federated learning was originally based on 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

In many applications, analysts and forecasters may still find it easier to centralize the data, which in our case would involve finding ways for other to be happy to share their distributed data. This is where data markets can play a role, by allowing for data (either raw or in their noisy version, possibly also after feature engineering) to be exchanged and priced through a common marketplace; for example, a pool. It means that data is then treated as a commodity or a good, for which payment implies transfer of ownership. Thinking about it, bilateral data markets have been around for a while already, as for the example of meteorological data companies selling weather information, as well as companies like Bloomberg that sell market intelligence data. The difference with the data markets we consider here is that these are multi-bilateral or within a pool to reflect the large number of players that may be involved, continuously running to reflect the streaming nature of data, etc.

Data markets involve a single communication step, limited computation and an eventual data exchange. The implementation aspects hence appear fairly light. At first, this may sound like a straightforward solution allowing 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 very challenging. A notorious example of a data market that did not work is that of the City Data Exchange hosted by the city of Copenhagen (Denmark) over the period 2016-2018. Today new data markets are being proposed, e.g., based on distributed ledger technology as for the example of the IOTA data marketplace (https://wiki.iota.org/blueprints/data-marketplace/overview).

Analytics Markets

Analytics markets comprise a way to blend the rationale of collaborative analytics with the inducements of monetary compensation for data 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. In the general framework of analytics markets, it is even to blend data sharing and distributed computation, while also allowing to accommodate privacy concerns. These types of markets are not as mature as for the other 3 cases, and mainly the focus of intensive research and development, for instance in the frame of the EU project Smart4RES (www.smart4res.eu).

Eventually, the platform assesses whether the analytics task is performed better thanks to that additional data. If that is the case, this triggers a payment from the central to the support agents. This payment is directly linked to how much the data contributed to improve the analytics task – for example to improve 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. There, the central agent posts a forecasting task on the platform, having already produced a forecast or not. All support agents then keep their private data for themselves, but use it to produce their best forecasts for the event or variable of interest. All these forecasts are gathered at the level of the platform. The platform subsequently applies a well-chosen aggregation operator to obtain a single optimal forecast based on the set of candidate forecasts provided. This single optimal forecast 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.

There is computation performed at the level of both the platform and the support agents, as well as communication between them. However, this does not require multiple iterations as in the case of collaborative analytics and analytics markets. There is a wealth of examples of prediction markets out there: some of them have been active for a long time (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).

These various approaches all have advantages and caveats but they also provide flexibility in implementation for different needs for communication, computing, complexity, etc. For instance, an approach based on federated learning may imply a large number of iterations between the platform and those contributing with their local computation. In contrast prediction markets do not involve iterations with communication and computation, though 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 monetary compensations to the contributors are directly related to the payments of the forecaster or forecast user who obtained an improved forecast (i.e., sum of revenues equal sum of payments).
2. **A zero element** – if no contribution to 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 to receive a monetary compensation if positively contributing to improvement in forecast quality.
5. **Truthfulness** – contributors only get their best monetary compensation if giving their best data, information or forecast.

There may additional properties involved, which are more technical, and possibly related to the specifics of the mechanism involved. In the above, all properties involve monetary compensations. Hence, for the case of collaborative learning in its most basic form, some of these properties may be more difficult to obtain, unless assuming that all agents are altruistic. Indeed, if not receiving monetary compensation to help improving forecasts, why would one try and provide their best information? Truthfulness also is a crucial property since it may also become an incentive to invest in improving data quality and the information content of features to be shared. Alternative approaches can be considered and implemented in order to yield these properties, i.e., they can be at the core of the mechanism design itself, or resulting from contracts and insurance policies. In addition, besides these market properties, aspects related to privacy preservation can be added and embedded into the market, for instance using differential privacy, k-anonymity or ad-hoc data exchange protocols.

NEW BUSINESS MODELS

Such a paradigm shift towards collaborative forecasting could give rise to a wealth of new business models.

Firstly, the collaborative forecasting platforms need to be developed in a way that will make them scalable in order to host many large forecasting tasks, accommodate data streams, etc., while being user-friendly in order to avoid a barrier 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, their revenues could be extremely large. These platforms are to be seen as a generalization of the current approach relying on bilateral data service agreement (e.g., between weather forecasts providers and their users). Such an evolution from ad-hoc bilateral agreement to platforms based on a pool or multi-bilateral agreements for standardized products was already witness in other sectors, as for the case of electric energy for instance.

In parallel though, those contributors who are to help to improve forecast accuracy by monetizing their data, analytics contributions and forecasts, will receive monetary compensation for their contribution. Eventually, this may even translate to revealing the value of each and every data point they collect, yielding a stable additional 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

While we have aimed to keep this article mostly non-technical, readers interested in the topic may want to dig into various papers to better appraise some of the technical concepts mentioned in the above. Generally, considering recent advances in markets for data (and information more generally), the paper by Bergemann and Bonatti (2019) is an excellent starting point.

Two examples of analytics markets are described by Agarwal and colleagues (2019) and by Pinson and colleagues (2022). The first one places more focus on the pricing mechanism and issues with the fact that data may be replicated and sold several times, while also discussing some of the market properties mentioned previously.

The second one 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.

In parallel, Rasouli and Jordan (2021) develop a compelling point about the idea of exchanging data for some other data, in contrast to exchanging data against monetary compensation. This is while those looking for recent developments with decentralized prediction markets based on distributed ledger technologies should have a look at the blueprint for Augur, by Peterson and colleagues (2020).

Finally, even though there are now hundreds of papers looking at federated learning and alternative approaches to decentralized learning, the interested reader should start with the blog post by McMahan and Ramage (2017) which gives a gentle introduction to the topic. Federated learning is today seen as blending collaborative analytics and analytics markets, allowing for monetary compensation, while having privacy-preserving versions.

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), pp. 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 MI (2021) Data sharing markets. Arxiv preprint, available online, arXiv:2107.08630, accessed on 29 July 2022.

Acknowledgements

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), as well as Aida Kharman, Christian Jursitzky, Pietro Ferraro and Robert Shorten at Imperial College London (U.K.).

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, a publication of the International Institute of Forecasters and a leading scientific journal in its field. He has made extensive contributions to probabilistic forecasting, forecast verification, as well as applications to, e.g., 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.
Contact: **p.pinson@imperial.ac.uk**