Unlocking Greater Insights with Data Collaboration

WORLD ECONOMIC FORUM

BRIEFING PAPER

JANUARY 2022

Companies across diverse industries recognize that there is value in collaborating around data and data-driven insights. Whether it is understanding user behaviour across channels and services, improving supply chains through increased transparency, or tracking environmental, social and governance efforts by sharing metrics, there is no question that data collaboration can help drive valuable insight.[1] Why, then, isn't every organization doing it?

In practice, data collaboration is complicated, even within different parts of the same organization. There are security and privacy risks; there are ethical and regulatory considerations. What is more, to capture the benefits of data collaboration at scale, all of the groups involved must make a substantial commitment to enable cross-group sharing and maintain trust.

The World Economic Forum spoke to leaders in the data collaboration space – from individual companies that started their own journeys of data collaboration, to providers of collaboration technologies, to consortiums set up specifically for data collaboration. They were interviewed about their greatest successes and challenges to date as they explored this growing area of enterprise opportunity. Their "lessons learned" offer a jump start for organizations looking to build or expand their data-collaboration capabilities.

Lesson 1: Let's stop talking about "data sharing" – it makes people think they are forced to share the data itself. Instead, let's talk about "data collaboration".

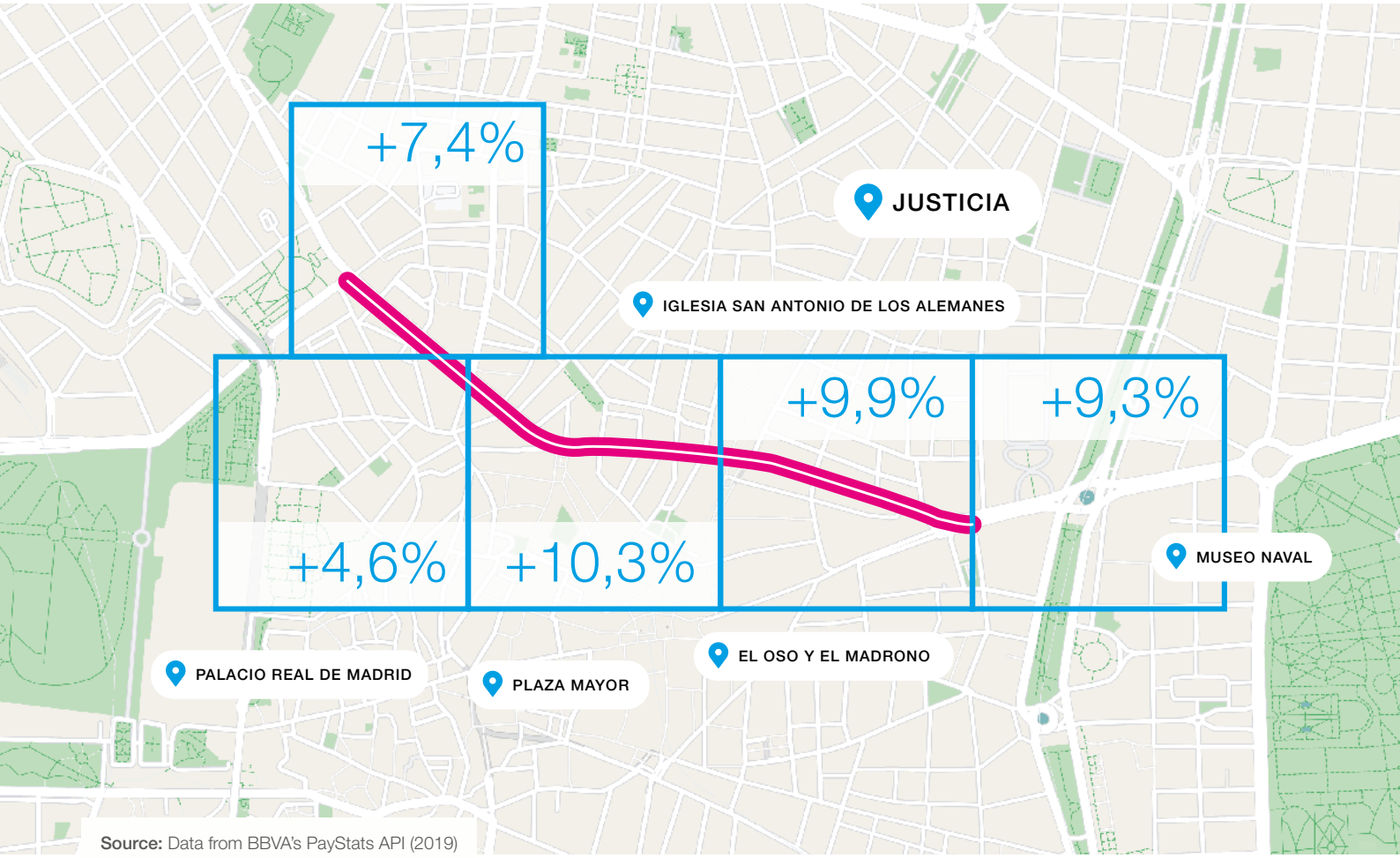
The term "data sharing" implies that data is directly shared with another party on a transactional basis. But that is not correct.

"Using the term 'data sharing' [for all data collaboration] is misleading," explains Rina Shainski of Duality Technologies, "because it implies that you take it, and you give it." But that is not the case.

"Data sharing" also implies that this is the only model available for data collaboration, which is not the case either. It is true that when organizations first began collaborating with their data, sharing data directly with each other was often the only option. This **"share-data"** model, as we refer to it, was one of the easiest methods to implement technology-wise as it simply meant copying the data and transferring it to another organization after data cleaning or data obfuscation. In such cases, calling these collaborations "data sharing" was logical because that is what was happening: the data itself was being transferred between organizations.

There are still some circumstances where direct data sharing is a valuable approach, with the proper technical and legal safeguards in place. The financial services company Banco Bilbao Vizcaya Argentaria (BBVA), for example, in wanting to scale its data collaboration capabilities, made only aggregated and anonymized data accessible through a geofenced, secured application programming interface (API). This enables direct data sharing while protecting against unauthorized access and violation of individual privacy rights. For BBVA, the aggregated and anonymized payment data was made available for Madrid City Council[2] through their API Market[3] to draw a more accurate map of consumer behaviour and gain an objective measurement of the impact on Christmas shopping of restricting car access to Madrid Central.

FIGURE 1 **Results of the analysis of consumer behaviour after changing car access rules**



Source: Data from BBVA's PayStats API (2019)

But in some circumstances, particularly in the case of highly sensitive personal data, the "share-data" model is not viable. Some companies have adopted the **"share-insight"** model. In this approach, an organization's data remains entirely under their control and is not shared with any collaborator. Instead, organizations only share the *insights* produced by their data. In other words, collaborator A can send a query to collaborator B, and B will run the query against their own data, sharing back only the results–none of the data itself is shared. Many market research and data companies such as Nielson and JD Power provide such insights as a service. While this approach protects individual information, it is expensive and slow as it typically requires additional and, in some cases, dedicated human resources that serve as the intermediary between the two collaborators.

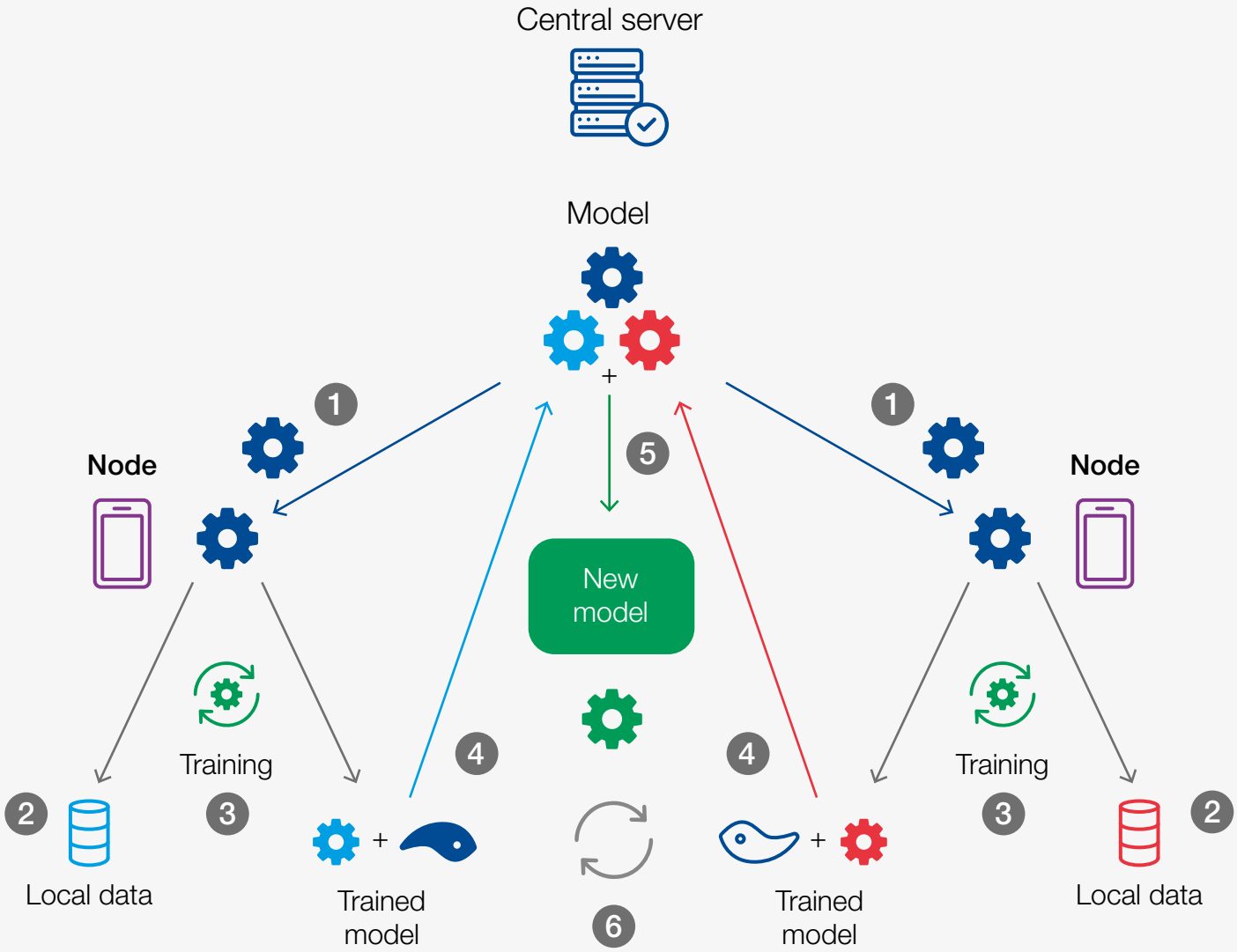
For example, in order to enter new markets, e-commerce companies may be interested in the buying power of specific geographical regions. For this purpose, they don't necessarily need the raw data, such as detailed personal transactional data. The insights alone are sufficient to support their business goals.

The most exciting new model lies between the "share-data" and "share-insight" models. The biggest potential value for organizations is in this spectrum of opportunity.

"Share-computation" gives organizations the right to bring the computation to where another company's data resides, or, in other words, to perform a computation on another company's data without moving it or having full access to it. The "share-computation" model enables the creation of an abstract model of the data, which is being shared across organizations, not the data itself. Enabled by emerging privacy-preserving technology solutions,[4] this approach lets collaborators maintain control of their data, while granting partially limited access and enabling insights. The data itself doesn't move and isn't shared; rather, the organizations needing larger insights bring computation to the data.

An interesting example of a "share-computation" model is a 17-partner consortium in Europe called MELLODDY. The consortium has developed a unique privacy-preserving data collaboration platform based on federated machine learning to extract insights from multiple preclinical datasets for accelerating drug development. The platform allows data contributors to train a machine learning[5] model on data behind their institution's firewall to create more powerful predictive models without compromising data and model privacy. Only the trained model is shared and not the data. The 10 pharmaceutical companies involved in the project – each with strong IT departments and security requirements – have already joined such a federated cycle twice since the start of the project. Because the platform is subject to an extensive external audit before each cycle, the data contributors know that they are not compromising their own security protocols to join. Ultimately, their data always remains within their infrastructure and under their control. MELLODDY announced[6] in September 2021 that it had observed early evidence of their federated learning exercise having boosted predictive performance and the chemical applicability of models used to inform drug discovery programmes.

FIGURE 2 **Illustration of secure aggregation within federated learning**



1) The central server broadcasts the model to the participants (step 1);
2) The participants train the model on their local datasets (steps 2-3) in their own infrastructure;
3) The participants mask thier model updates with pairwise masks and send them to the central server (step 4);
4) The server aggregates the local models (step 5); and finally
5) The whole process repeats (step 6) and needed.

Source: MELLODDY Consortium, *Protected: Privacy @ MELLODDY* (2021)

Beyond federated learning, additional technological solutions can help achieve collaboration without direct sharing, even for those without the in-house capabilities needed to drive it. The provider Points, for example, offers not only solutions based on federated learning but also secure enclaves and homomorphic encryption, among others, to allow for a more general computation beyond machine learning models. Some of these techniques are used in a stand-alone fashion, while others can be combined as needed depending on the use case. A good example of such a use case is banks collaborating in anti-money laundering efforts where sharing a machine learning model alone is insufficient. Regardless of the specific underlying techniques, data collaboration is possible *without* direct data sharing, so it is best to refer to it correctly.

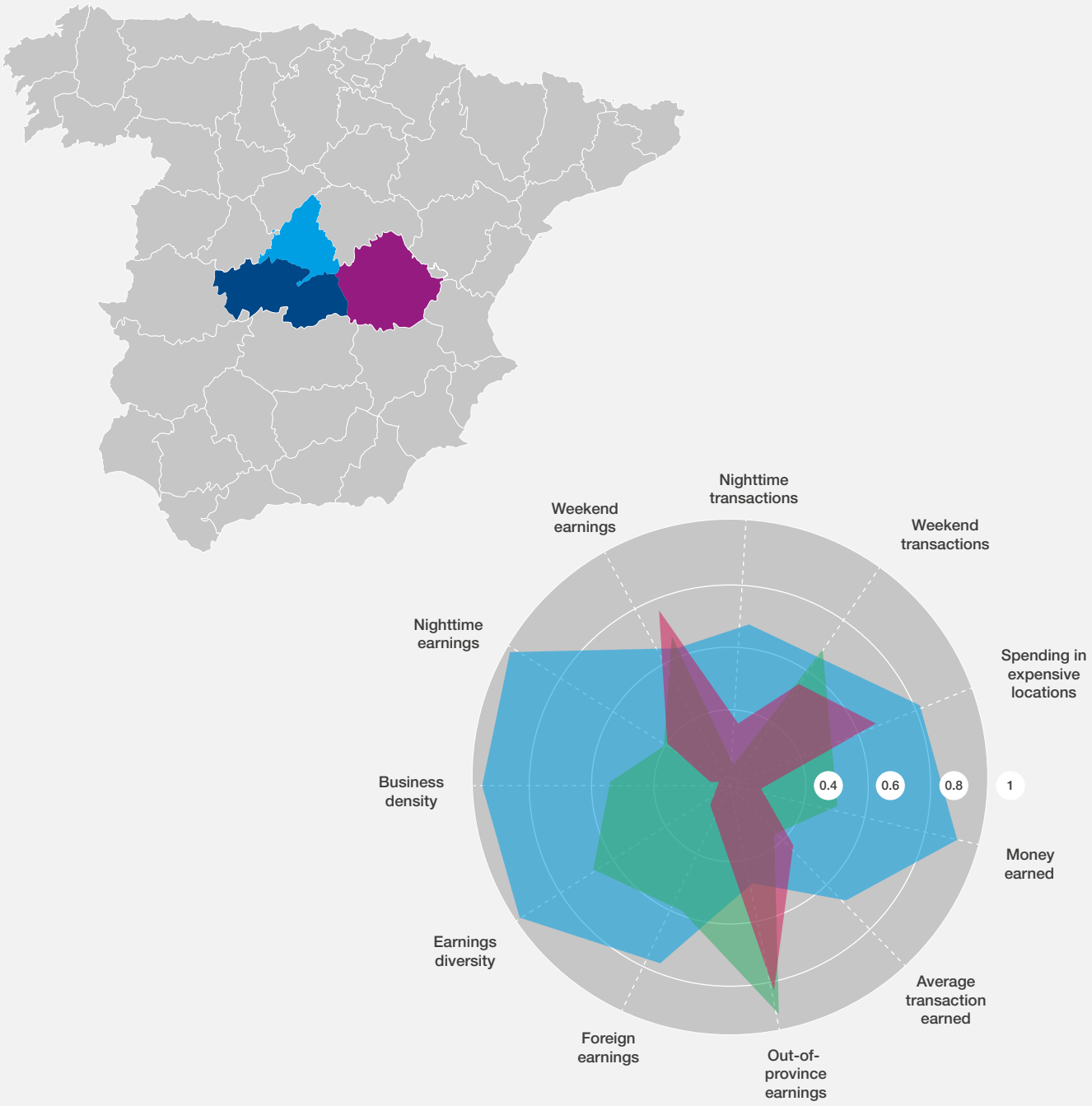
Lesson 2: Be clear about the problem you are trying to solve. Have a clear use case.

With the breadth of data available via data collaboration, it can be tempting to adopt an exploratory mindset and ask the question: “Given all of this data, what can we do?”

Companies such as BBVA started their data collaboration journey in this way and it can be a valuable way to begin. BBVA collaborated with a group of data scientists at the Massachusetts Institute of Technology (MIT) to explore how to best serve their customers based on the data they already had. Our discussions with BBVA and other leaders in the space, however, suggest that beyond the initial exploratory period, companies should shift their mindset to solve specific problems.

As part of the Urban Lens joint project, MIT and BBVA researchers set an objective of performing a comparative analysis of city-level microeconomics through the lens of individual financial spending. Having a clear use case allowed them to build a model that characterizes the socio-economic development in Spanish regions.

FIGURE 3 **Predicting regional economic indices using big data of individual bank card transactions**



Source: Urban Lens, a joint project of researchers at MIT Senseable City Lab and BBVA with an interactive app demonstrating analysis results (2017)

"Ideally, everything starts with the clarity of the use case," says Rina Shainski of Duality Technologies, when asked what companies should consider before starting on a data collaboration journey.

One example is eBay, a corporation that relies on data collaboration for fraud prevention and fraud detection purposes, as well as for anti-money laundering and Know Your Customer efforts. To meet these needs, eBay engages in data collaboration and data-sharing activities with service providers, both with careful attention paid to ensure that data will be shared only with the proper legal basis. The providers then respond to the respective information requests with analysis and insights. By beginning with a clear use case, eBay is well-positioned to understand exactly what data is needed for which purpose and is prepared to build a strong compliance and governance system around it.

Juan Murillo Arias of BBVA, which participates in data collaboration primarily as a data provider, agrees. "Data collaboration could be better if end users think in problems," says Arias. The company notes that they have been surprised by the number of applications people have come up with when challenged to explore their available data, but Arias says that for the best outcomes, it is best if partners identify the proper problems to be solved from the start. From there, the group can determine whether the available data can meet the needs of the use case in question or if additional data is necessary. The use case helps them focus and accelerate their overall data collaboration efforts.

Having a clear use case allows for the right balance of domain and data expertise. If organizations can make their "business" application needs clear to the data scientists – whether they are internal or providers of centralized collaboration solutions – the data collaboration is more likely to be successful.

Starting with a clear use case also gives organizations a better opportunity for future collaboration. Decision-makers are more likely to be swayed by a clear goal and outcomes with well-defined key performance indicators (KPIs) that address business problems they have already experienced. By contrast, using a more exploratory "let's see what we can do with data collaboration" approach makes the solutions feel less mature and less likely to result in newly captured value.

Similarly, eBay's compelling use case for fraud prevention and detection discussed earlier lends itself to strong executive buy-in, given the use case's alignment with its corporate responsibilities. Further explanation about the need for a buy-in can be found in the next lesson.

Lesson 3: Secure executive buy-in from across the organization.

Data collaboration offers much more potential value to business than a single technological tool or strategic approach. But putting it in place in a way that is sustainable and scalable requires executive buy-in across an organization.

Companies see data as one of their most valuable assets, and the notion of collaborating around it with other organizations – perhaps future competitors – can leave executives hesitant to commit. Leadership up to the level of the CEO needs to clearly understand the value (read: quantified risks and benefits) of data collaboration and the shift in mindset required for the organization to participate in it. (This returns to the first lesson: data collaboration does not have to mean data *sharing*.)

With that understanding in place, however, new opportunities abound.

Majid Al Futtaim's Data and Analytics Centre of Excellence, incubated within Corporate Development, was given a broad mandate to maximize the value of the rich data the company collects and transform it into a data-driven organization. Majid Al Futtaim, the leading shopping mall, communities, retail and leisure conglomerate, first started its collaboration efforts internally by bringing together all its customer data to create a single view of its customers. This was no small task as data from hundreds of millions of customer interactions had never been shared between all the different businesses.

Majid Al Futtaim then continued its data collaboration externally with the Dubai government to create insights from multiple data sources that would help the government to gain a better understanding of economic trends in order to shape policies. This further evolved into the vision of building a data and insights company which leverages customer data from Majid Al Futtaim's malls, retail outlets and leisure establishments across the Middle East. The ambition and vision to create and accelerate this central data hub was made possible by the ongoing support from Majid Al Futtaim's leadership team.

In other situations, buy-in is similarly crucial to move forward. Rather than a push for a new business model driven by data collaboration, in a few other companies interviewed by the World Economic Forum team, data collaboration was the obvious answer to the operational needs of the existing businesses involving commercial transactions (see eBay's anti-money laundering example above).

Building the capacity for sustained success requires further buy-in. IT groups need a clear view into intended uses, both short- and long-term, to ensure that data can be made available in the proper formats for direct sharing, insight generation, or computation-in-place. They can also provide critical expertise as to which approaches the organization is ready to participate in based on current technology capabilities, or what capabilities would be needed to expand collaboration in the future.

Lesson 4: Understand the legal and regulatory environment concerning data – both the limitations *and* the opportunities.

While it is encouraging that companies are taking data privacy and security regulations seriously, too many businesses are rushing to lock down everything – leaving both themselves and the people they serve missing out on the valuable insights made possible by data collaboration. Getting clarity on which laws are applicable in which circumstances is a critical step.

"There isn't a blanket restriction on sharing data," Rob Leslie of Sedicii explains, "but talking to many organizations, you would think that there was. Right now, especially small institutions are terrified because it's been drilled into them that data sharing in any form is not acceptable." The General Data Protection Regulation (GDPR) provides a clear set of parameters to comply with. How these parameters are implemented is left to the institutions themselves to determine based on their particular set of circumstances and risks.

There are, of course, types of data that cannot be transferred or shared, and data privacy regulations continue to evolve. But locking the door and pulling down the blinds, metaphorically, is not the answer. As Leslie points out, regulators are attempting to simplify the process. This simplification will be quicker if organizations stay up to date and are involved as new procedures are developed. In the financial sector, for example, regulators *want* to facilitate a space between GDPR protections and the criminal compliance laws financial organizations must follow. Regulators are trying to get to a point where organizations can follow one set of procedures to comply with both. They are currently developing implementation guidelines to help institutions determine how best to comply with what appears to be, on the face of it, a "conflict of laws" situation so that they can comply with both GDPR and the anti-money laundering laws that exist.

There are opportunities for forward-looking organizations to help drive these conversations and ensure that carefully orchestrated data collaborations remain viable under evolving regulations. For instance, IT executives need to collaborate closely with legal and compliance teams from the start of the process. For legal experts unfamiliar with newer technical solutions (such as share-insight and share-computation models, where the technology itself can help prevent data disclosure), giving data collaboration the proper consideration requires in-depth discussions to understand where the technology can help address legal and compliance concerns, and where strict governance is required. Conversely, **a strong legal understanding of the current regulations** concerning data will help other stakeholders get a clear idea of what is possible.

As an example of driving such conversations, through its partnership with Wharton Customer Analytics, Majid Al Futtaim has released sets of anonymized data to 11 research groups from different universities, enabling researchers to access valuable data and business users to benefit from advanced research.

"Collaborating on anonymized data helps address certain aspects of the regulation and lowers confidentiality risks as

the data is masked appropriately," says Guillaume Thfoin, Head of Data & Analytics at Majid Al Futtaim Holding. "However, it does not remove the risk completely as some data can be de-anonymized."

At the other end of the spectrum, Rob Leslie says, are organizations whose leaders aren't thinking carefully enough about the implications of existing data privacy regulations.

"GDPR, for example," says Leslie, "doesn't limit 'personal data' to just the data itself. It also classifies anything that allows you to identify an individual as personal data."

"So, if I write something into a blockchain that allows me to identify an individual off-chain," Leslie continues, "what is now on-chain is personal data even though it contains no directly personal data. When someone says they want their data removed from the blockchain, how do you comply? The only way is to erase the whole chain."

The potential difficulty of having to stop the processing of individual datasets is why global companies like eBay minimize collaboration involving data that requires user consent. The reason is that individual consents can be withdrawn at any time.

Additionally, the treatment of consent is complicated as it varies across different jurisdictions. In some jurisdictions, for example, consent expires automatically after a number of years. And once data is ingested and combined with other information, it can be impossible to untangle and remove it should consent be revoked.

That is doubly important from a data collaboration point of view, where multiple organizations and their respective technology infrastructures are involved. A partner's failure to properly consider the implications of data privacy regulations could create privacy and compliance issues for everyone involved. When it comes to data collaboration, organizations must consider regulations from the perspective of their shared footprint before collaboration begins.

For instance, Majid Al Futtaim leverages a **staged approach**. Collaborations with partners commonly start by sharing insights (e.g. aggregated data) to gauge the value of a collaboration, often followed by a data-matching exercise (e.g. the match rate between customers in different datasets). This can be done in a fully anonymized and privacy-preserving way using technologies such as clean rooms, compute-to-data and homomorphic encryption.

Once the interest of the collaboration has been established, based on the value of the joined datasets, Majid Al Futtaim can then establish different levels of sharing/merging of the datasets based on the abilities and willingness of each party.

"This staged approach allows us to quickly assess the value of the collaboration and its staged approach without starting with lengthy legal discussions. Once we have a clear intent to collaborate, the legal teams can establish a more definitive and comprehensive framework," says Guillaume Thfoin.

Strong data governance complemented by a **data-sharing council** is key. eBay takes this approach seriously: its council comprises representatives from business, technology, product, strategy, legal and privacy groups that make decisions on data collaboration activities.

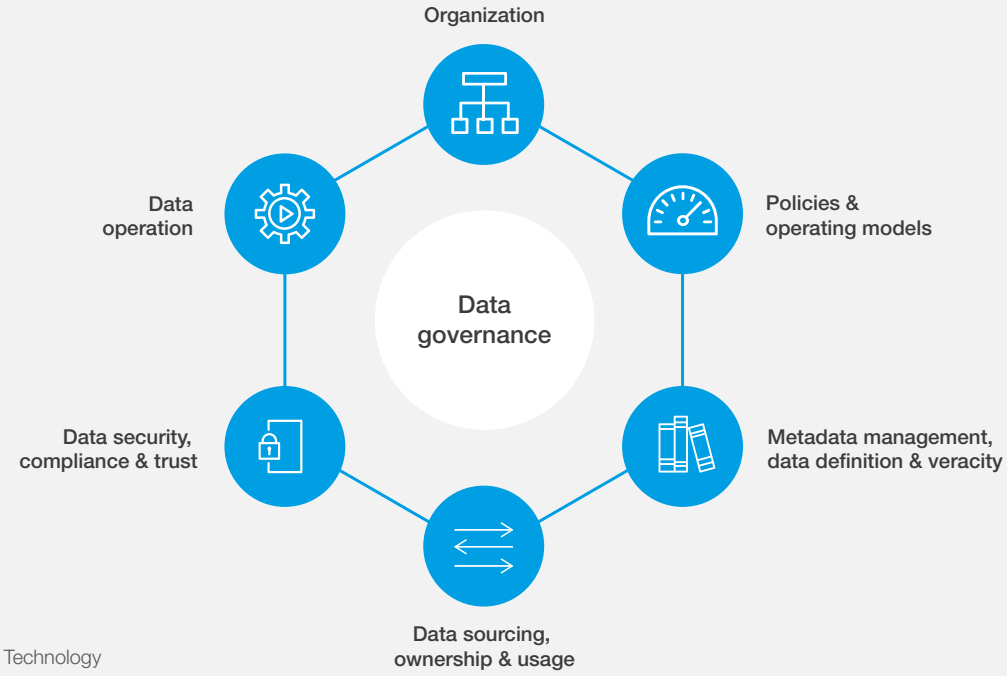
"The council's job is to weigh the risks and benefits of every strategic data sharing decision," says eBay's Chief Privacy Officer, Anna Zeiter.

Lesson 5: Get your own house in order before you begin.

At the beginning of every commercial airline flight comes the reminder that if oxygen masks are deployed, it is essential to put one's own mask on properly before trying to help others. The mandate for data collaboration is a similar one. If you want to collaborate properly with data – whether it's with different groups within your own company or with partners from other organizations – ensure your own house is in order before you begin by having a strong data governance programme in place.

The questions this data governance programme needs to answer most importantly are: What kind of data do you already have? Where does it originate? What can be shared? Or can only insights be shared? And for how long?

FIGURE 4 **Illustrative data governance framework**



Source: Accenture Technology Organization (2021)

Organizations have struggled for years to keep track of their data. Most are still working to understand what kind of data they have and ascertain what data can do for them. When it comes to data collaboration, companies must also have a clear understanding of what format the data is in, how current, accurate and complete it is, what kind of information security is in place, which regulations and potentially legal restrictions apply to which types of data, and so on.

When starting to share data between its businesses, Majid Al Futtaim had to create its own data sharing agreements to facilitate collaborations. Teams quickly realized that a scalable data governance programme needed to be implemented to further enable data exchanges and value creation. Over the past three years, Majid Al Futtaim has invested in modern data governance tools that enable automated data cataloguing, lineage tracking and discovery.

"Having the right data governance frameworks, although a very hard and unexciting task, is critical to set the right foundation for data management internally first and is a must to start any data collaboration," says Guillaume Thfoin.

Beyond regulatory compliance and simple reconciliation of data formats, perhaps the single most important question is one of quality.

"Organizations need to have gone through some form of independent audit in order to meet a standard," says Rob Leslie.

This provides a level of assurance for everyone involved. "If everybody does that," says Leslie, "there's a high degree of certainty to the results you'll produce." At the end of the day, the old maxim still applies: "Garbage in, garbage out." It is therefore imperative to cleanse and standardize any data involved in a collaborative activity as far as possible.

For example, if a group of banks is involved in an anti-money laundering collaboration, each bank needs to ensure to a high degree of certainty that all of the parameters associated with an account or a transaction are as accurate as possible. A simple misspelling of a name could lead to an incorrect identification or the failure to accurately identify someone involved in an illicit transaction of some kind.

Organizations initiating data collaboration efforts must also ensure that they have similar quality reviews in place for future data that is to be collected, ideally as part of the process of collecting and storing the data in the first place, so that collaboration can continue to evolve.

A foundation for success

Data collaboration across well-aligned ecosystem partners has been shown to give rise to new business models, higher operational efficiency and better customer experiences, and drive growth and innovation. Yet outdated defensive postures around collaboration and poor understanding of evolving regulations have left companies hesitant to engage in data collaboration at scale. This hesitancy is leaving real business value untapped at a time when organizations could benefit from collaboration most.

Lessons learned by leaders in data collaboration offer a path to success for those who are not yet capturing the benefits of data collaboration.

First, there are multiple options for collaboration. While some involve sharing data directly with others, there are other less risky models that do not involve moving data from one organization to another.

Second, collaborations that start with a clear use case, business outcomes and well-defined KPIs are more likely to find early success and build it into sustained new sources of value.

Third, executive buy-in along with a strong data governance programme across the organization can significantly alleviate risks and concerns.

Fourth, and similar to the previous point, a clear understanding of the legal and regulatory environment concerning data can help determine the proper scope of data collaboration opportunities, increasing confidence as organizations move forward.

Finally, the work of data collaboration first begins at “home”. Companies need to clearly understand what data is available, its ownership, security and any applicable regulations before beginning a data collaboration journey.

With these five lessons learned by the leaders in the space, organizations have a blueprint for success as they begin their own data collaboration journeys.

Contributors

Hugo Ceulemans
Scientific Director, Discovery Data Sciences, Janssen Research and Development, Janssen Pharmaceutica; MELLODDY Project Lead

Rob Leslie, Chief Executive Officer
Sedicii Innovations

Edy Liongosari
Managing Director, Chief Research Scientist, Accenture Labs

Juan Murillo Arias
Data Strategy and Data Science Innovation Senior Manager, BBVA

Rina Shainski
Chairwoman, Co-founder, Duality Technologies

Grigory Shutko
Platform Curator, Information Technology Industry, World Economic Forum

Michelle Sipics
Editorial Lead, Accenture Technology Innovation

Guillaume Thfoin
Head of Business Analytics, Majid Al Futtaim Holding

Anna Zeiter
Chief Privacy Officer, eBay

Sarah Zhang Jiachen
Founder and Chief Executive Officer, Guangzhishu Technology (Points Technology)

Endnotes

1. World Economic Forum, *New Paradigm for Business of Data,* 29 July 2020, https://www.weforum.org/reports/new-paradigm-for-business-of-data.

2. "Efectos Gasto Navidad 2018/2019, Gran Via y Madrid Central", *Diario de Madrid*, 2019, https://diario.madrid.es/wp-content/uploads/2019/01/MC-gastos-navidad-DEF.pdf.

3. BBVA, *API Market*, no date, https://www.bbvaapimarket.com/es/banking-apis/.

4. Zanussi, Zachary, "A Brief Survey of Privacy Preserving Technologies", *Statistics Canada*, https://www.statcan.gc.ca/en/data-science/network/privacy-preserving. (Updated 1 December 2021).

5. Li, Tan, Talwalkar, Tameet, Kumar Sahu Anit and Virginia Smith, "Federated Learning: Challenges, Methods, and Future Directions, *arXiv*, Cornell University, 21 August 2019, https://arxiv.org/pdf/1908.07873.pdf.

6. "MELLODDY announces first demonstration of federated learning improved model performance in drug discovery", *MELLODDY*, 27 September 2021, https://www.melloddy.eu/y2announcement.

7. "Research Opportunities", *Wharton University of Pennsylvania*, no date, https://wca.wharton.upenn.edu/research/research-opportunity-with-majid-al-futtaim/.

8. "Estimating the success of re-identifications using incomplete datasets using generative models", *Nature Communications,* 10, No. 3069, 2019, https://www.nature.com/articles/s41467-019-10933-3.