THE NEW TECH TOOLS IN DATA SHARING

By Massimo Russo and Tian Feng

This is the final article in our series on creating value in the data economy.

"**T**HERE'S GOLD IN THEM thar hills," cried Yosemite Sam in the old Bugs Bunny cartoons, but he never got to enjoy it.

"There's gold in them thar data," Sam might say today. But he'd be equally disappointed unless he was able to navigate the host of data-sharing challenges that we detailed in previous articles in this series. (See the sidebar, page 2.)

Here's Sam's conundrum in a nutshell. Data sharing, by definition, involves multiple parties that tend to coalesce around ecosystems. As these ecosystems grow, they share more types of data, and more detailed data, among the members of an expanding community. They also develop solutions that address an expanding range of use cases, some of which were totally unforeseen when the data was originally generated or shared. Each of these factors introduces its own set of risk-value trade-offs. The extent of the tradeoffs depends on the specific data-sharing capabilities of the underlying platform.

The good news is that the technology companies that enable many ecosystems either have developed or are developing a host of technological solutions to facilitate data sharing by mitigating risk, enhancing value, and reducing the sources of friction that inhibit sharing. These solutions shift the tradeoff frontier between value, on the one hand, and risk and friction, on the other, toward value. Forward-looking management teams should educate themselves on the issues at stake and the technology solutions coming into the marketplace.

Start with Your Needs and Goals

Before they delve into the details of complex technology solutions, management teams need to consider their data-sharing context (for example, their goals, prospective partners, and the potential pathways to value) and their priorities (orchestrating an ecosystem or contributing to one or

more, for example). Three questions bear investigation.

What sharing issues are raised by the underlying data needs, use cases, and scope? Each industry and data-sharing ecosystem has its own challenges. For example, sharing patient data within a network of hospitals poses very different technology challenges than sharing carbon emissions data among companies in a particular industry or country. While both might have clear use cases, the former involves more private, and therefore sensitive, data, which elevates concerns about security. At the same time, the relatively small number of companies sharing the data reduces risk and complexity. In a similar vein, companies working with personal data may need to protect individual data points, while companies sharing enterprise data may want to mask aggregated insights to maintain competitive advantage.

Are good data governance procedures in place? Good data governance practices ensure that technology is used appropriately and consistently. Common challenges such as data breaches often occur not because of technological shortfalls but because of user error. Management should consider data-sharing enablers and data governance as part of one strategic process. Some of the most effective tools, such as those for data management and classification, help companies establish good governance practices by identifying the right level of security for a given type of data. For example,

data consent management and data access controls are automated ways of managing governance.

Where do gaps in trust inhibit data sharing? Trust is a prerequisite in many data exchanges, and a good number of data-sharing technologies function as trust substitutes or trust obviators. Among trust substitutes, technologies such as blockchain, ratings systems, execution environments, and application program interfaces (APIs) decrease risk by imposing controls or by creating transparency, introducing a technological intermediary in which all parties have confidence. Among trust obviators, technologies such as federated learning, edge processing, and private set intersections reduce the need for trust by creating alternatives to direct data sharing. To enhance sharing, it's important to identify trust gaps, determine whether you want to replace or reduce the need for trust, and then adopt the right technological solutions.

Available Technology Solutions

Cloud providers are integrating data-sharing capabilities into their product suites and investing in R&D that addresses new features such as data directories, trusted execution environments, and homomorphic encryption. They are also partnering with industry-specific ecosystem orchestrators to provide joint solutions.

Cloud providers are moving beyond infra-

structure to enable broader data sharing. In 2018, for example, Microsoft teamed up with Oracle and SAP to kick off its Open Data Initiative, which focuses on interoperability among the three large platforms. Microsoft has also begun an Open Data Campaign to close the data divide and help smaller organizations get access to data needed for innovation in artificial intelligence (AI). Amazon Web Services (AWS) has begun a number of projects designed to promote open data, including the AWS Data Exchange and the Open Data Sponsorship Program. In addition to these large providers, specialty technology companies and startups are likewise investing in solutions that further data sharing.

Technology solutions today generally fall into three categories: mitigating risks, enhancing value, and reducing friction. The following is a noncomprehensive list of solutions in each category.

## 1. MITIGATING THE RISKS OF DATA SHARING

Potential financial, competitive, and brand risks associated with data disclosure inhibit data sharing. To address these risks, data platforms are embedding solutions to control use, limit data access, encrypt data, and create substitute or synthetic data. (See Exhibit 1.)

Data Breaches. Here are some of the technological solutions designed to prevent data breaches and unauthorized access to sensitive or private data:

- Data modification techniques alter individual data elements or full data sets while maintaining data integrity. They provide increasing levels of protection but at a cost: loss of granularity of the underlying data. De-identification and masking strip personal identifier information and use encryption, allowing most of the data value to be preserved. More complex encryptions can increase security, but they also remove resolution of information from the data set.

- Secure data storage and transfer can help ensure that data stays safe both at rest and in transit. Cloud solutions such as Microsoft Azure and AWS have invested in significant platform security and interoperability.

- Distributed ledger technologies, such as blockchain, permit data to be stored

EXHIBIT 1 | Technology Can Mitigate Risk
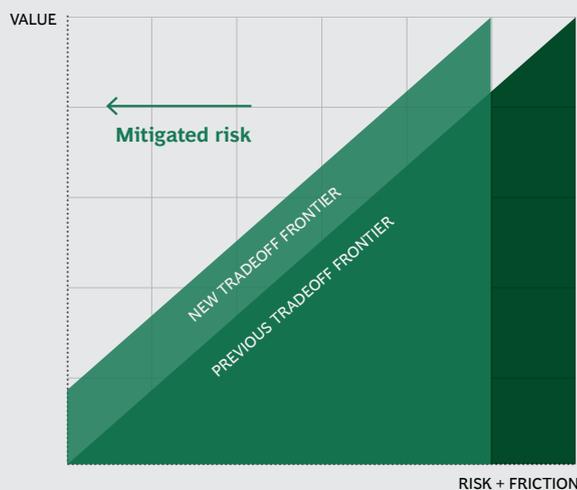
CHALLENGES	SOLUTIONS	EXAMPLES
Data breaches	· Secure data storage & transfer	· Secure cloud · Distributed ledger technologies
	· Secure computation	· Secure multiparty computation · Trusted execution environments
	· Data modifications	· Encryption · De-identification/ masking/anonymity · Differential privacy
Data mis-sharing & misuse	· Usage & access controls	· APIs · Secure database management · Self-destructing data
	· Sharing substitutes	· Federated learning · Synthetic data
Data quality risks	· Validation & authentification	· Distributed ledger technologies · Watermarking

Source: BCG analysis.

and shared in a decentralized manner that makes it very difficult to tamper with. IOTA, for example, is a distributed ledger platform for IoT applications supported by industy players such as Bosch and Software AG.

- Secure computation enables analysis without revealing details of the underlying data. This can be done at a software level, with techniques such as secure multiparty computation (MPC) that allow potentially untrusting parties to jointly compute a function without revealing their private inputs. For example, with MPC, two parties can calculate the intersection of their respective encrypted data set while only revealing information about the intersection. Google, for one, is embedding MPC in its open-source Private Join and Compute tools.

- Trusted execution environments (TEEs) are hardware modules separate from the operating system that allow for secure data processing within an encrypted private area on the chip. Startup Decentriq is partnering with Intel and Microsoft to explore confidential computing by means of TEEs. There is a significant opportunity for IoT equipment providers to integrate TEEs into their products.

Data Mis-Sharing and Misuse. Platforms are embedding a series of functions to control access to and distribution of data:

- APIs are the most widely adopted form of access control. For example, Alibaba's Data Middle Office uses APIs to provide a centralized data and analytics platform that acts as the single source of truth for the company's data. Startups such as Immuta are building fine-grained access controls in addition to smart-query and masking tools to ensure that sensitive data is used only by those with permission to do so.

- Federated learning allows AI algorithms to travel and train on distributed data that is retained by contributors. This technique has been used to train machine-learning algorithms to detect cancer in images that are retained in the databases of various hospital systems without revealing sensitive patient data.

- Synthetic data, a relatively new approach, mirrors the properties of an original data set without disclosing any private information. The data can then be shared with partners to train algorithms or test software. A platform called Mostly AI is designed to generate synthetic data sets.

Data Quality Risks. Before using data, it's critical to understand its provenance, relability, and authenticity. Data watermarking, for example, can document the precise origin of data sourced from third parties, and any tampering will remove or distort the watermark. Adobe recently announced its Content Authenticity Initiative, which embeds an encrypted chain of metadata into digital media to reveal provenance and potential tampering.

2. ENHANCING THE VALUE OF DATA SHARING

Despite the oft-cited analogy, data is not a well-priced commodity like oil. Data has very different values depending on its uniqueness, the specific use case or problem it can help solve, and the degree to which customers are willing to pay for the solution. Valuing data is hard. Data platforms are creating tools to help assess and capture this value. (See Exhibit 2.)

Value Identification. Some data marketplaces are employing AI to match data to potential use cases. For example, data broker Dawex is building AI solutions that match sellers and buyers of data. The company also provides a service to help evaluate data sets using factors such as volume, history, completeness, validity, and rarity. Data competitions are another way to help owners find use cases for their data.

Value Assessment. Once use cases have been identified, how can enterprises

EXHIBIT 2 | Technology Can Enhance Value
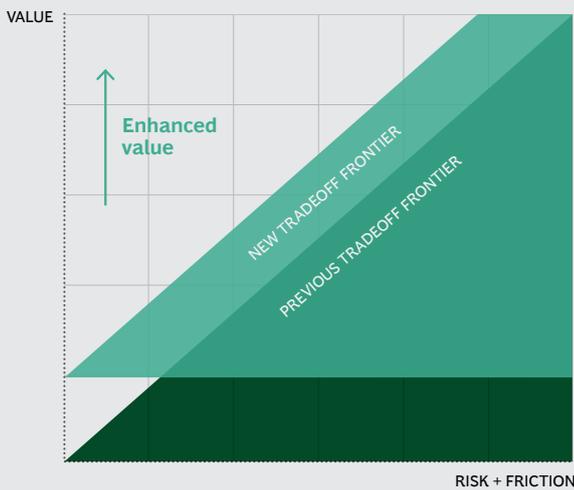
CHALLENGES	SOLUTIONS	EXAMPLES
Value identification	· Data usage mapping	· AI · Data competitions
Value assessment	· Assessment tools	· AI · Data ratings
Value capture	· Usage tracking & control	· APIs · Smart contracts · Data watermarking
	· Payment systems	· Blockchain · Micropayments
	· Data organization	· Data containers · Data networks

Source: BCG analysis.

determine the value they should receive in exchange for sharing their data? While AI is traditionally viewed as a means of deriving insights from data, it can also be used to value the data itself. Collective and Augmented Intelligence Against COVID-19, a new coalition of researchers and nonprofits (including the Stanford Institute for Human-Centered Artificial Intelligence and UNESCO), are leveraging AI to sift through competing data sets in order to organize and evaluate the quality of available data. Other data validation tools, such as ratings, can likewise help with value assessment and are being applied in emerging blockchain data marketplaces.

Value Capture. As data travels from it point of origin to use, it can pass through multiple entities. Similar to raw materials that are transformed along a value chain into a usable product, data can be manipulated and combined—and its value enhanced. Functions such as micropayments, smart contracts, and subscription systems are being embedded into data platforms to capture value. Access control, permissions, and data-tracing solutions ensure that data is routed only for the intended use. New startups such as Digi.Me and Solid, an open-source technology developed in part by World Wide Web inventor Tim Berners-Lee, help consumers control and monetize their data through personal data container technologies. Solid, for example, stores data in "pods" in an interoperable format and provides users with the tools to control which organizations and applications can access their data. This type of technology could be applied to enterprise data as well.

3. REDUCING FRICTION IN DATA SHARING

Data is often hidden behind enterprise walls, so it can be difficult to know what data exists. Solutions are emerging to search for potentially valuable data and then transfer and transform it into usable forms. (See Exhibit 3.)

Dark Data. In a data-sharing ecosystem, there is a need to both discover unknown (or "dark") data behind enterprise walls and aggregate data from disparate sources. Cross-enterprise data directories, such as those provided by marketplaces like Azure and AWS, can provide visibility into the types of data that exist. For example, Dawex is building a data marketplace technology that can search through a variety of data sources (across dimensions

EXHIBIT 3 | Technology Can Reduce Friction
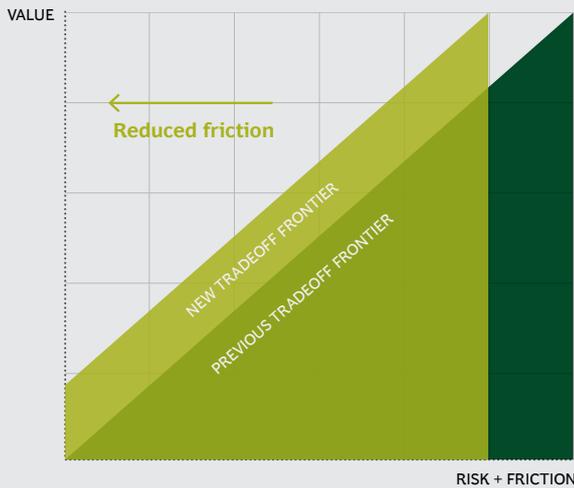
CHALLENGES	SOLUTIONS	EXAMPLES
Dark data **Aggregation complexity**	· Discovery & aggregation tools	· Data curation/wrangling · Data directories · Data fusion tools · Master data management
Barriers to interoperability	· Interoperability support	· RESTful APIs · Semantic technologies · Standards
Barriers to transfer	· Transfer & streaming tools	· APIs · Edge/fog processing · Networking/communication technologies · Stream processing · Software development kits

Plot axes: VALUE (vertical), RISK + FRICTION (horizontal). Labels: Reduced friction, NEW TRADEOFF FRONTIER, PREVIOUS TRADEOFF FRONTIER.

Source: BCG analysis.

such as data types, attributes, and time series) provided by individual contributors.

Aggregation Complexity. Data from different sources will have different definitions, formats, and meanings. Companies such as Tamr and Trifacta are developing "data wrangling" and data curation tools to identify, clean, and interpret data so it can be easily understood and combined.

Barriers to Interoperability. Due to the prevalence of heterogenous data types and data sources, companies are investing in data standards and harmonized data models. For example, in mid-2020, Microsoft acquired ADRM Software, a data analytics company with a large set of industry data reference models that facilitate data aggregation and analysis for enterprises. Other reference architectures, such as LF Edge, the IDS Reference Architecture Model, and iSHARE, can enable secure and scalable peer-to-peer transactions. These initiatives are defining semantic metadata, which attaches context and meaning to a data set, and building registries to facilitate interpretation, aggregation, and analysis of data.

Barriers to Transfer. Fundamental to sharing is the ability to transfer data from its point of origin to its intended use with sufficient speed and without so much overhead that the data's value is compromised. Traditional tools such as APIs, streaming platforms (Apache Kafka, for example), connectivity solutions (5G), and cloud provider features like Azure Data Share have already led to many advances in this area.

THERE IS GOLD in data sharing, which leads quickly to more efficiencies for companies, new services for customers, and new models and revenue streams. Data platform providers and ecosystem orchestrators are investing in solutions to push out the boundary of the risk-value-friction tradeoff in data sharing by enabling controlled access and analytics while protecting the underlying data. It's a complex marketplace that's taking shape, and corporate management teams with data-sharing interests or ambitions need to get up to speed.

About the Authors

Massimo Russo is a managing director and senior partner in the Boston office of Boston Consulting Group. You may contact him by email at russo.massimo@bcg.com.

Tian Feng is a project leader in BCG's Boston office and a BCG Henderson Institute ambassador. You may contact her by email at feng.tian@bcg.com.

Acknowledgments

The authors thank their colleague Aldo Ninassi for contributions to this article.

About BCG

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963. Today, we help clients with total transformation—inspiring complex change, enabling organizations to grow, building competitive advantage, and driving bottom-line impact.

To succeed, organizations must blend digital and human capabilities. Our diverse, global teams bring deep industry and functional expertise and a range of perspectives to spark change. BCG delivers solutions through leading-edge management consulting along with technology and design, corporate and digital ventures—and business purpose. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, generating results that allow our clients to thrive.

About BCG Henderson Institute

The BCG Henderson Institute is Boston Consulting Group's strategy think tank, dedicated to exploring and developing valuable new insights from business, technology, and science by embracing the powerful technology of ideas. The Institute engages leaders in provocative discussion and experimentation to expand the boundaries of business theory and practice and to translate innovative ideas from within and beyond business. For more ideas and inspiration from the Institute, please visit https://www.bcg.com/featured-insights/thought-leadership-ideas.aspx.