2 Dataspaces Enabled Mobility

Prof. Dr. Christoph Schlueter Langdon

Data! Finally! The reckoning of the complexity of data has arrived in the C-suite. Data natives saw this coming ages ago. A deep understanding of data and its complexity has always been a key ingredient of good science. But in business it was a bit like fighting windmills – like Don Quixote – until the 2020 Coronavirus tragedy hit. The pandemic unexpectedly put data in the spotlight and turned it into front-page news. Yes, CEOs and business pundits have been talking about “data as the new oil.” Yet, with the Corona crisis, all of us are suddenly debating new variables such as the R0 rate, pronounced “R-naught,” which represents the number of new infections estimated to originate from a single case. And suddenly key concepts to evaluate the quality of scientific research, reliability (consistency of measures), and validity (accuracy) made sense to everyone. The crisis pulled data and its complexity out of the shadow and into the limelight.

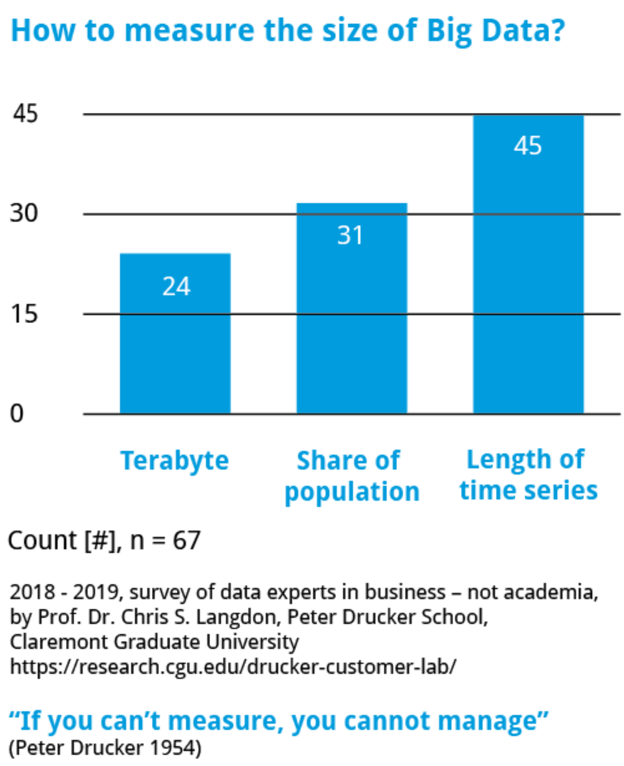

Figure 7: Data measurement dilemma

In the past, if I asked an audience who had taken a statistics class, all hands would go up. But no one would volunteer to quickly explain a t-test. If I asked about data, nobody had taken a class, but everybody would be eager to explain it. So, it was easy to peddle stereotypes, omitting deeper issues, such as data sharing and pooling architectures, and data rights management and governance (e.g., Otto 2011). Even supposedly straightforward issues, such as sizing data, how to measure the quantity of data, has remained surprisingly tricky to this day. We know to buy eggs by the dozen, a pound of butter, and a liter of milk. But how do we buy data? By byte or length of a time series or share of population? Figure 1 illustrates this data sizing dilemma.

2.1 From centralized storage and data lakes to data spaces

The Coronavirus pandemic has upended long-held beliefs in data handling. A prominent example is centralized data storage. It may just be a human habit or a trait and second nature, or even biological. Because humans were hunter-gatherers and stashing food away for bad times was critical for survival (Stephens et al. 2019). Even today any accountant would confirm that there is efficiency in a centralized data stash, as one building is cheaper than two for the same space. So, data was kept centrally because it felt better, and the numbers backed up intuition. Then social media happened, and data became a legal headache: How can we protect data privacy? Well, don't collect it in the first place, or lock it away, the lawyers would advise. Now, it felt better, the numbers were better and … well, who would put a career at risk and argue against lawyers? Then suddenly data scientists emerged waving artificial intelligence (AI) like a magic wand asking for more and more data to create secret formulas or algorithms like those by Amazon, Facebook, and Google. Now it became uncomfortable because who could argue with trillion-dollar market capitalizations? So, something had to be done. Fortunately, none other than Harvard Business School's esteemed strategy guru Michael Porter (Five Forces, Value Chain), a consultant favorite and appreciated in the C-suite, came to the rescue. Together with James Heppelmann he wrote a seminal piece on Smart Products and the Internet of Things (IoT) and advocated: "Create a data lake" (Porter & Heppelmann 2015, p. 109). Today, it has become clear that such centralization is giving way to a more distributed way of creating data pools required for AI.

For one, small companies with only a little data would be left high and dry; this is a no-brainer particularly in Germany with its strong medium-sized companies, referred to as "Mittelstand," which collectively form the backbone of its economy. For another, in Europe, even companies at the other end of the spectrum, the biggest firms in Europe, the 800-pound gorillas, the seemingly unbeatable, are experiencing difficulties in catching up with hyperscalers and first movers from the U.S. and China.

The automotive industry is a case in point. Let's consider autonomous driving. For some time, the industry has been debating whether video data is sufficient to train the best algorithm or whether additional sensors such as LiDAR ("Light Detection and Ranging" to measure ranges) would provide better data. It's a debate about ingredients and nutritional value: Is it better to use butter, or is margarine sufficient? It obscures the fact that regardless of butter or margarine the sheer volume of data is important, particularly for the type of AI used in this field (primarily neural network-based methods). And while the industry has been debating, a new entrant, Tesla, has been collecting. Tesla does not use LiDAR, which is still expensive, but video cameras instead. Tesla uses a large number of them and most importantly, it started doing so years ago. It suddenly leaves industry heavyweights that produce millions of vehicles annually playing catch-up.

How can anyone catch up and create vast data pools quickly? One solution is data ecosystems with data spaces. Business ecosystems are understood as broad configurations of actors operating in a loose setting (Moore 1996 and 1993, for an overview, see Tsujimoto et al. 2018). A data space is a loosely coupled system based on rules or standards for data governance to assure data sovereignty (Jarke et al. 2019). Loosely coupled systems have a rich history in the field of information systems, such as with web services, for example (Schlueter Langdon 2003a). They can provide the flexibility for best business fit with tight integration (Schlueter Langdon 2006, 2003b).

One prominent example of a rule-based data space approach is International Data Spaces (IDS). "International Data Spaces (IDS) is an initiative that promotes a virtual data space leveraging existing standards and technologies, as well as governance models to facilitate secure and standardized data exchange and data linkage in a trusted data ecosystem" (Otto et al. 2019a, p. 37). A core tenant of IDS is a virtual homogenous data layer with decentralized data storage, which in turn "provides a basis for creating smart-service scenarios and facilitating innovative cross-company business processes, while guaranteeing data sovereignty for data owners" (Otto et al. 2019b, p. 9). In 2020 IDS has emerged as a data space standard:

- It became a formal standard published as DIN Spec 27070 "Requirements and reference architecture of a security gateway for the exchange of industry data and services" (link, ordering at: link)
- "IDS is significantly involved in the design of GAIA-X" (BMBF, Federal Ministry of Education and Research, link), an upcoming, multi-year initiative of the German government: GAIA-X strives "to set up a high-performance, competitive, secure, and trustworthy data infrastructure for Europe" (BMWi, Federal Ministry for Economic Affairs and Energy 2019).
- First implementations are available, such as the IDS-based Trusted Connector of the Deutsche Telekom Data Intelligence Hub (link), which was launched as the first "IDS-ready" certified connector at HMI 2019 (Fraunhofer 2019) and which is applied in mobility, specifically in intermodal transportation, in the first laboratory of the National Platform Future of Mobility (NPM) from 2020 until

the end of 2021 in Hamburg (NPM, link).

- So, let's dig deeper and take a look at the digital transformation of the automotive and mobility business to discuss implications for data, specifically its organization and management in the context of data spaces and IDS.

2.2 From automotive to mobility

Automotive is out, mobility is in. People will still buy cars, but they will buy fewer of them, and the excitement is moving on … definitely in densely populated urban spaces and developed countries, maybe less so in rural areas and emerging markets. Such a shift has occurred before with other well-known technologies, like radio, for example. Radio technology and radio broadcasting as a business were a sensation; until television arrived. Video didn't kill the radio star as prophesied in the hit song by The Buggles, the first ever music video to be shown on MTV in 1981, yet TV greatly diminished radio's importance. And today TV itself is being pushed out of the limelight by mobile video platforms such as YouTube and Netflix. So, what should be done in automotive, how can a transition toward mobility be successful?

Merriam-Webster defines mobility as "the quality or state of being capable of moving or being moved." In big cities it is already evident how mobility is evolving from a product-centric and self-organized affair ("my car") toward a more seamless experience involving different modes of travel. New last mile solutions, such as shared bikes and e-scooters, and location data could allow for intermodal offers. Instead of walking to a destination after parking the car, the driver could accept an offer for a shared e-scooter nearby. Combining such last mile solutions with a parking recommendation, the driver could be navigated to an empty parking spot with a scooter or a van shuttle nearby.

Figure 22 illustrates how a simple trip from A to B could evolve toward seamless mobility (Schlueter Langdon 2019). It illustrates the evolution of personal mobility from a self-organized affair toward one that is orchestrated on behalf of a user in a seamless and personalized fashion. For example, a three-segment journey from point A to point B could evolve from car ride-parking-walking to one where the car is navigated to a parking spot that is available (smart parking) and with an electric scooter nearby (smart shared) for the last leg of the journey. The figure also analyses the two sides of the same coin: the user view and the provider perspective. From a user's point of view speed and comfort are important. For a provider the business model and profitability matters.

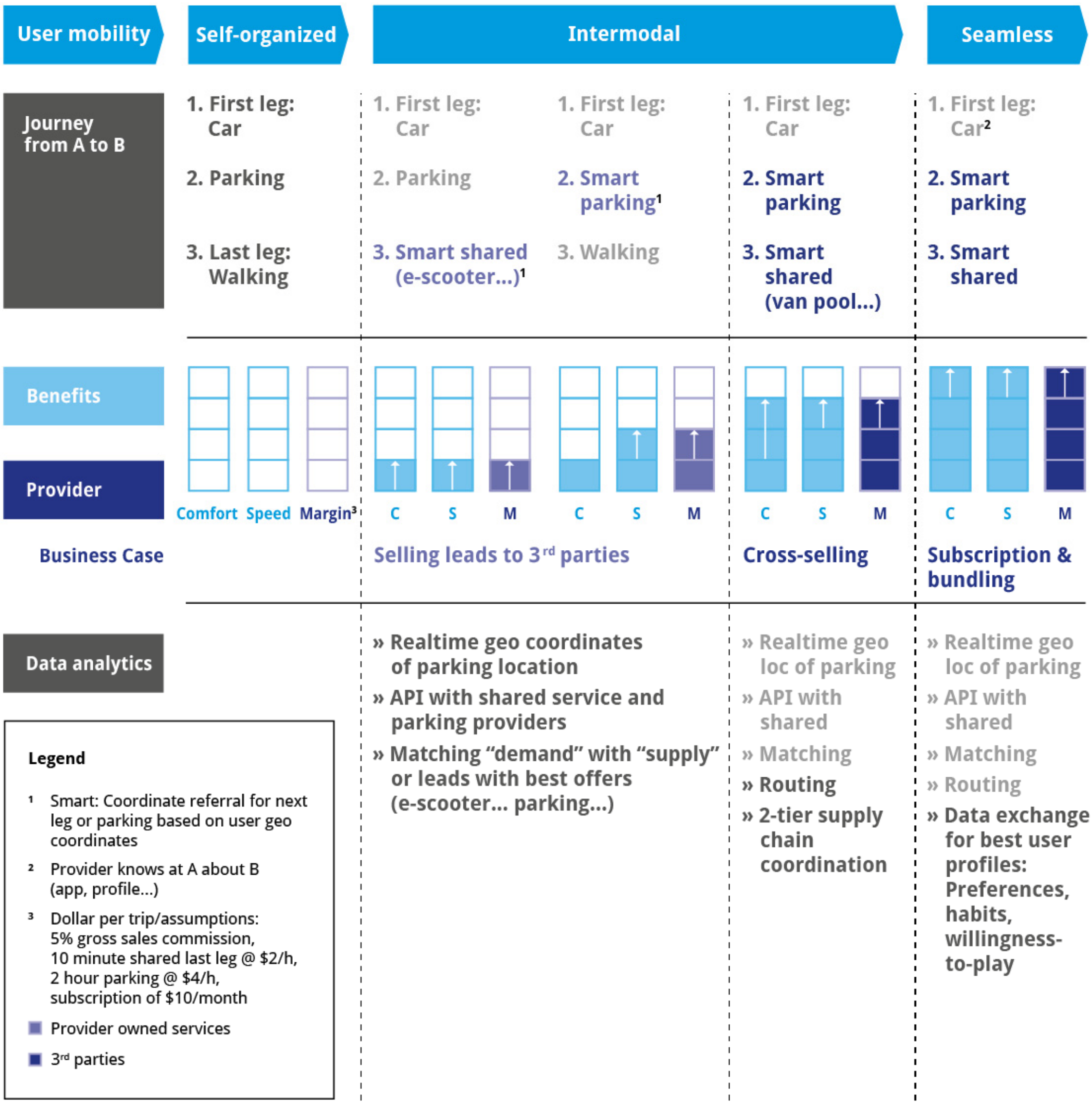

Figure 2: Economics of seamless mobility

The good news: The smarter the solution the bigger the benefits for both sides. This is no surprise as today's system is highly inefficient: cars remain unused for nearly 23 hours per day, are mostly used for single-occupancy trips, and parking often entails circling the block, etc. Our assessment is based on simulation experiments and pilots. The bottom part of the figure is a look under hood that dissects what a smart solution would entail and require.

While high-level, it nonetheless clearly conveys that data needs grow significantly with every step from self-organized to seamlessly orchestrated travel. What is interesting to see is how user benefits increase. Users can save travel time and enjoy increased comfort and convenience. Using a scooter is faster than walking, being navigated to an empty parking spot saves time searching and circling the block; and hopping into a shared van shuttle is probably more comfortable than both walking and riding a scooter. For mobility companies this is terrific news, because higher user benefits can translate into opportunities for higher margins (results based on our "Mobility-as-a-Service Calculator").

However, despite this lucrative outlook, few mobility companies are currently winning financially. Car manufacturers seem to be struggling. What is going wrong? For one, it takes a new success formula. Performance is no longer measured in revenue per vehicle but revenue per trip. For another, creating the right business system und underlying infrastructure seems to be a big challenge, because it is very different from traditional automotive business. Winning is no longer so much about cool cars, mastery of manufacturing, and dazzling dealerships. And it can't be fixed by recruiting star designers, powertrain mavericks, and slick salespeople. Optimizing results per trip requires an entirely different business and information systems infrastructure. A lot of the complexity of winning in mobility is hidden behind the word "*smart*" used in Figure 22 – the notion of an infrastructure designed "*so as to be capable of some independent action [...] having or showing quick-witted intelligence*" (Oxford English Dictionary). The good news – there is a solution that is familiar; it breaks down complexity into building blocks, which makes it manageable. It was invented in the airline business, another transportation business – and with spectacular success. While history won't repeat itself exactly, applying the lessons learned can help avoid mistakes and save time. So what happened and what can be learned from airlines? Lessons from last century's U.S. airline deregulation

In 1978 U.S. President Jimmy Carter signed the Airline Deregulation Act into law (Statue-92-pg1705, link). It was the first piece of deregulation (affecting routes and market entry) that torpedoed business as usual for U.S. carriers. A second piece in 1983 (reducing fares) threatened survival (Kahn 2007). Suddenly, low-cost startups invaded the business with much lower fares. Carriers like American Airlines, Delta, and United looked like dinosaurs facing extinction. Yet, fast forward 40 years, and those very brands are still alive – dominating the U.S. market. *How* did they survive and dominate?

From the outside little seems to have changed. Today, same as 40 years ago, American, Delta, and United buy planes, paint them in their livery, and fly them from A to B. Yet, behind the scenes a lot has changed. Most importantly, carriers started with changing results. As Peter Drucker, the legendary founder of management science, said: "start with results, the rest will follow" (Drucker 1963). 40 years ago, airlines started to *shift from revenue per route to revenue per seat*. What seemed to be a minor tweak in a financial spreadsheet required large investments in new services and processes – and the systems and software to automate it all *("*softwarization*"* Schlueter Langdon 2003c). Fundamentally, the challenge was twofold: First, optimizing the new metric – creating algorithms or analytics *"engines,"* which is an analytics or *data science* task; then secondly, providing the data input to an engine and automating its outcome – deploying the algorithms into a living process, which is a systems and data engineering job.

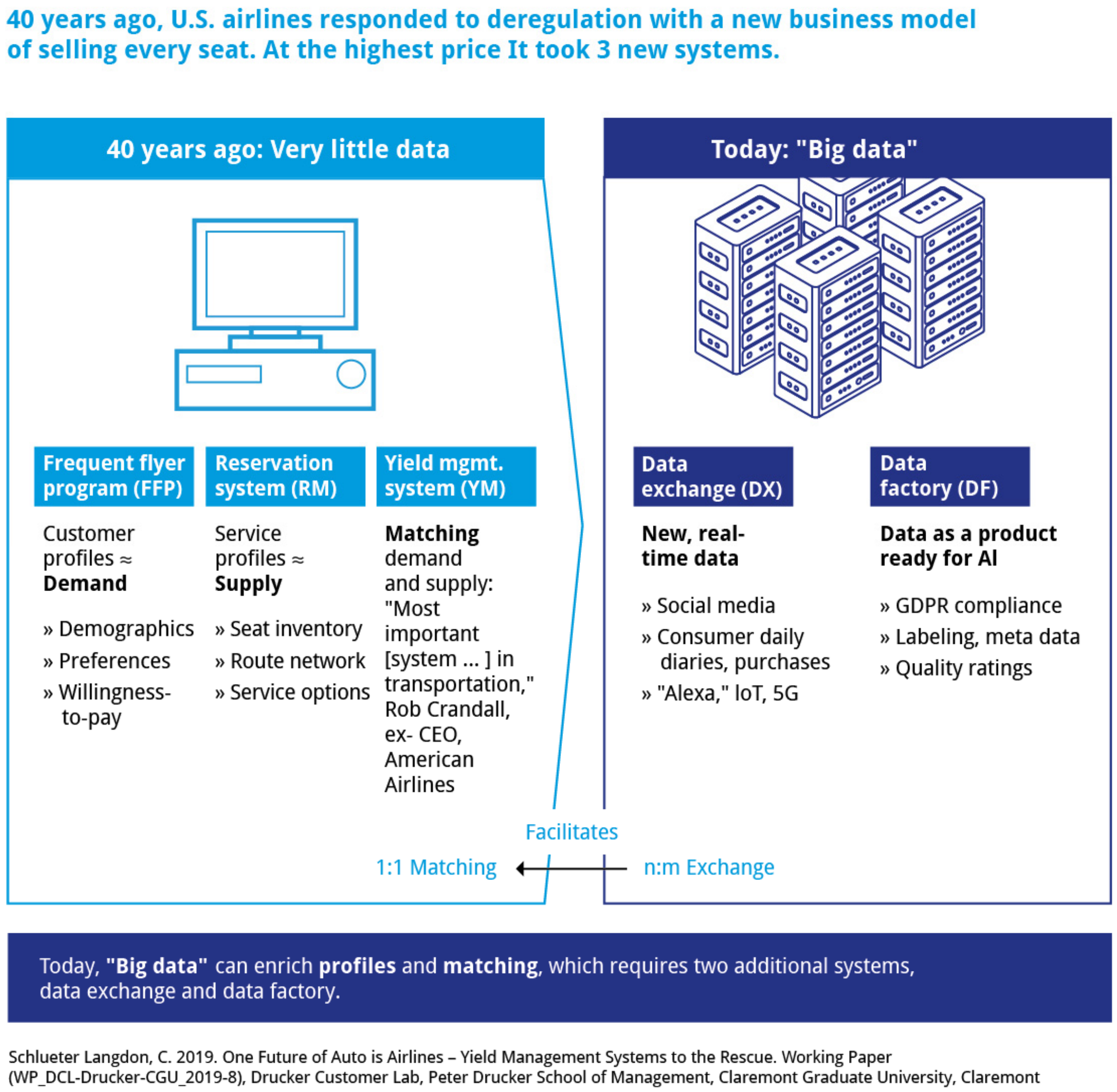

Figure 3: Industrializing the optimization of yield per unit of consumption

2.3 Three systems: two for data, one for analytics

Shifting from revenue per route to revenue per seat gave birth to three types of systems (see Figure 3, Schlueter Langdon 2019):

- An airline reservation system (ARS)
- A loyalty system for a frequent flyer program (FFP)
- A yield or revenue management system (YM)

Why these three systems? At the core of the shift to revenue per seat was the insight that profitability required *selling every single seat and at the highest price possible*, essentially treating seats as perishable goods and customers as NOT created equal. From an analytics perspective the challenge was doable: matching demand with supply, matching customers with seats.

It required coupling inventory management with variable pricing: Use seat inventory to determine supply, use customer profiles to predict demand, then use different price points to clear the market. Finally, learn from results and adjust inventory next time around, for example by using a bigger plane or adding a flight.

The problem back then was less with the analytics but more with the data – or more precisely, the lack of it. Where could the seat inventory for a particular destination be found? How is it possible to keep track of different routes to the same destination? For example, 5 seats in Business Class from Los Angeles/LAX to New York/JFK, and 10 seats for the same destination but via Chicago/ORD, and therefore, with a much longer travel time. Airlines created *reservation systems to manage inventory*.

And where can customer profiles be found to predict demand and establish price points? Selling each seat and at the highest price requires insights into a customer's willingness-to-pay (WTP). Predicting WTP, in turn, requires data on travel event type (leisure or business), budget (income or travel policy), sensitivity to travel time (daytime departure or red eye), travel duration (non-stop or stopover), convenience (economy class or business), and the decision context (traveling alone or with family) – and all of the above not at some aggregate, average level but for each potential traveler. In order to collect this data airlines invented *frequent traveler programs to create traveler profiles.*

Finally, the matching of supply -using the service profile data from the reservation system - with demand – using the traveler profile data from a loyalty system - is automated with a *yield or revenue management system*. Robert Crandall, former Chairman and CEO of American Airlines, gave yield management its name, calling it "the single most important technical development in transportation management since deregulation" (link); for American Airlines, see Smith et al. 1992; for a YM literature review, see McGill & Van Ryzin 1999; for state of YM, Carrier & Fiig 2018).

2.4 More data … two more systems

With a shift in results from revenue per car to revenue per trip, learning from airlines seems a very appropriate first step. Yet, three systems may not be sufficient anymore. The data situation has been reversed. 40 years ago, airlines faced a data drought. Today, there is a glut of data available that can add critical value and should be utilized. 40 years ago, smartphones did not exist. Today, in developed countries almost every adult is using one, and the device itself has evolved into one gigantic data logger (Dezember 2018). Smartphones have been a key enabler of a trend that has been dubbed *"SoLoMo"* by John Doerr, a partner at influential Silicon Valley venture capitalist Kleiner Perkins in 2010 (Guynn 2013). It summarizes the expansion of digitization into social, local, and mobile applications, which has fueled growth of data on consumers and its commercialization for advertising and new service offerings. Companies like Facebook, Google, and Uber exemplify this trend.

Yet those high-profile firms are just the tip of the iceberg. Today, a vast cottage industry of consumer data brokers has emerged that sells consumer data.

For example, a new 2019 law in Vermont that requires data brokers to be registered (Vermont 2018) has already revealed more than 120 vendors of consumer data (Melendez 2019).

All this consumer data allows for better customization of offers by evolving from artificial and fictional *"personas"* with their inherent flaw of bias (systematic error) toward profiles cut from real-life behavioral data of actual and potential customers (McKinsey 2017, Crosby & Langdon 2014). Figure 4 illustrates the different data types available for constructing profiles today:

- data on consumers (traditional demographics, government statistics),
- data on products and services (from vendors),
- user-product interaction data (behavioral data), and a broad category of
- context data.

The latter ranges from capturing a consumer's daily diary and friends & family to environmental settings like weather and traffic conditions. For example, Uber's *"Pulse of a City"* provides a visualization of people traffic data (Belmonte 2015).

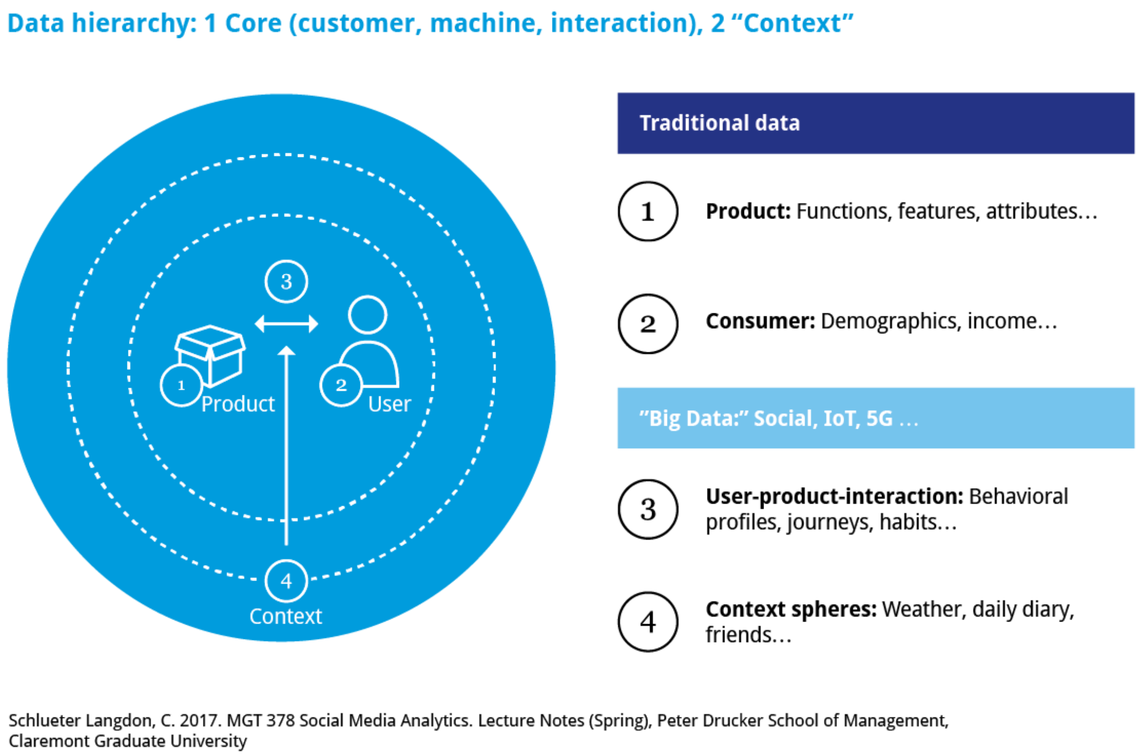

Figure 4: Enriching customization and personalization with Big Data

And much more data is expected. Again, new technology, such as 5G, a cellular mobile communication standard for higher speeds and bandwidth, and the Internet of Things (IoT) with bots and virtual assistants like Amazon's Alexa, will accelerate data creation (Crosby & Langdon 2017). To further complicate the data challenge, this data growth is happening everywhere:

- Within the enterprise
- Across a company's ecosystem of partners
- Externally

Let's take a car manufacturer, for example. Different departments collect data on the same customer: Market Research, Vehicle Development and – as vehicles are being connected for remote, over-the-air (OTA) updates and telematics – also Services & Parts and Financial Services.

And business ecosystem partners are also collecting data on the very same customer: dealerships, aftermarket vendors, insurance companies, payment processors, and various systems operators (concierge services, fleet maintenance providers, etc.). For best consumer profiles and most beneficial matching of users with "*seat inventory*" all this data ought to be considered – otherwise somebody else could make a better offer. It is a bit like battlefield intelligence in military applications: a better sensor, such as a radar system or night vision goggles, can create an immediate advantage.

2.5 Data exchange

This is where a data exchange system could add value. Think of it as a marketplace or "*supermarket*" with data products for data scientists. Today, according to meta-research, more than 80 percent of the time budget of a data analytics project is spent on data wrangling – not with algorithms (Press 2016, Vollenweider 2016). Companies have gone from databases to data warehouses and now to data lakes (Porter & Heppelmann 2015) – and they seem to be drowning in them. The question is – how can all this data be consolidated, organized, and made available to data scientists? An internal data exchange is one solution. Instead of searching for data across departmental silos and country operations, a data scientist could "*shop*" for internal data in a central location. This data hub could also connect with ecosystem partners as well as external, commercial data brokers to provide a single "*storefront*." It could track transactions for ease of auditing. It would also be a smart solution from a compliance and risk management standpoint. Instead of dealing with data regulation in a fragmented fashion, it could be standardized and enforced centrally.

In Europe, one example would be compliance with the General Data Protection Regulation (GDPR, European Commission 2018), which aims to give control to individuals in the EU over their personal data. An exchange could handle data anonymization and consent management of personally identifiable information centrally. With the Deutsche Telekom Data Intelligence Hub (DIH) there could be multiple data exchange options. Data could be exchanged peer-to-peer, directly between a data seller and a data buyer in a transaction brokered by the DIH (see "T-Systems as Pioneer: Implementing IDSA," link). The data could also be persisted on the Deutsche Telekom DIH and made available to buyers on a seller's behalf.

In any case, the DIH relies on IDS standards to ensure that firstly, only trusted partners can transact and that secondly, data will only be traded if binding usage restrictions to the data can be assigned, which creates a secure and trusted data space.

2.6 Data factory

As more and more data will be generated within a company using social media, 5G, or IoT, another system will be required. Call it a data factory. A data factory is needed to refine raw data into data products (Crosby & Schlueter Langdon 2019). Despite the hype surrounding data analytics and artificial intelligence (AI), raw data is still confused with refined data. Machine learning and AI methods require refined data products. This is obvious for data scientists but few in management seem to be aware of it. The food analogy can help illustrate the gap. Very few of us pick food from trees or slaughter animals; most visit a supermarket and pick food off the shelves. The food at the supermarket is processed, packaged, and labeled. Labels inform about product name, vendor, quantity, ingredients, and nutritional value. For example, a "*Nutrition Facts*" label in the U.S. can easily exhibit 20 rows of data (U.S. FDA 2016). These labels are no coincidence but the result of rules. These rules have evolved together with food processing to ensure product quality to protect consumers, because bad food can be a health hazard.

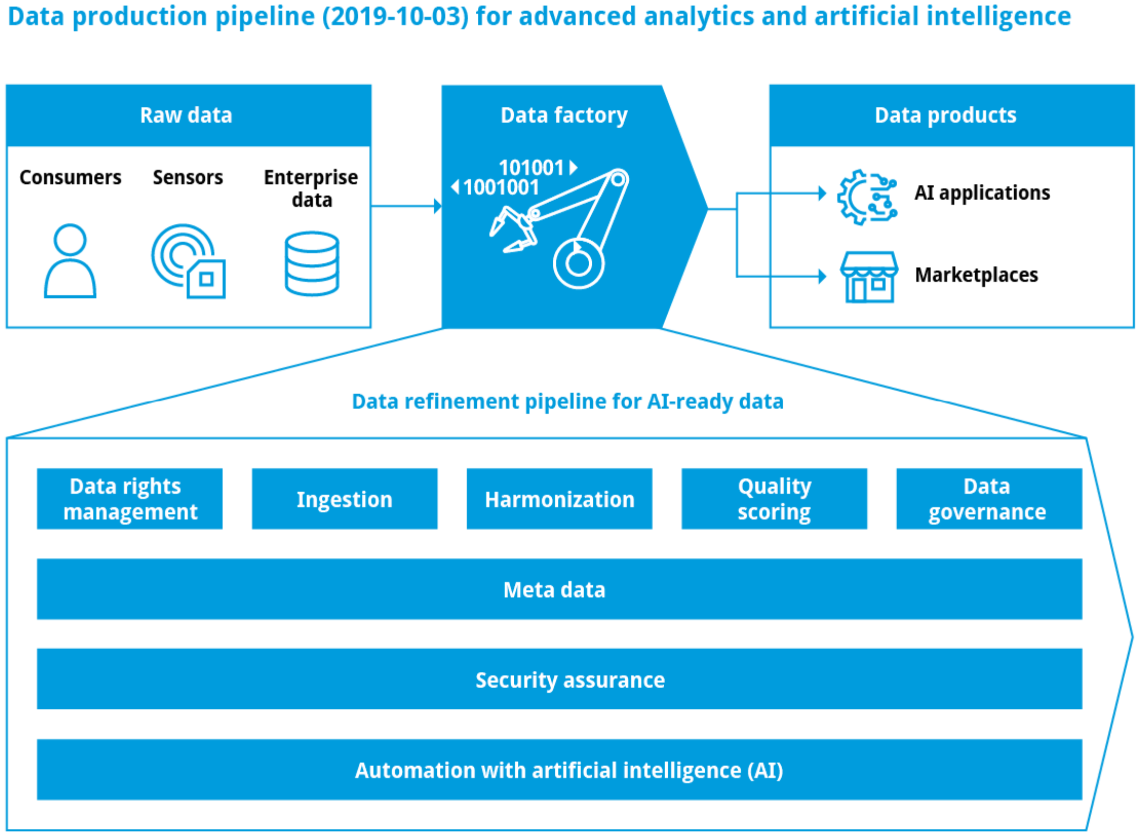

Figure 5: Data factory framework

In a nutshell, the food we buy is a product. It is processed, labeled, and packaged to be safe for consumption and exchanged for money. Machine learning and AI require data products. So, data could learn from food. For data to become a product it needs to be processed, labeled, and prepared to be safe for use and exchange. This data productization can be accomplished economically with a data factory. Figure 5 illustrates such a data factory framework (adapted from Schlueter Langdon & Sikora 2020).

In a nutshell, raw data rights must be verified before any data can be ingested or harvested (rights, licensing, user consent). Then data ought to be properly labeled or tagged for it to be made discoverable through a catalog of categories and search engines (classification). Furthermore, it needs to be scored to provide some indication of quality, because without it any subsequent analytics is pointless – "*garbage in, garbage out*" (GIGO, quality scoring). Finally, data governance mechanisms are required to ensure digital sovereignty for data owners making data available to be exchanged or shared.

2.7 Data spaces for trusted data exchange

Many AI applications, such as predictive maintenance or autonomous driving, can require more data than what is available within a single department and company. Creating data pools across companies would be an advantage (Otto et al. 2019b, Fig. 2.3, p. 15). For example, pooling all data of a particular machine type across all installations (horizontal pooling) would create a rich dataset for anomaly detection and its root-cause analysis. Another use case is pooling data vertically, across the participants along an entire supply chain or channel system in order to better estimate arrival times or ensure proper end-to-end temperature treatment of shipments, for example. In both situations, horizontal and vertical pools, outcomes would be best if most participants were to contribute. However, so far, few companies have been willing to engage in this type of data sharing. On the one hand, data is increasingly seen as a strategic advantage (the value aspect of "*data is the new oil*"), and therefore held closely and protected. On the other hand, more sensor data will only increase data pooling benefits. What has been missing are exchange options with data governance mechanisms that strike a balance between the need to protect one's data and share it with others (Otto et al. 2016; IDSA 2018a, 2018b).

Such data governance solutions are emerging. An important example are data spaces based on the reference architecture model (RAM) of the International Data Spaces Association (Otto et al 2019b). IDSA is an association of industry participants, created to promote data governance architecture solutions based on research carried out by the German Fraunhofer Institute with funding from the German government (Fraunhofer 2015). Members include car makers like Volkswagen, suppliers like Bosch, and traditional information technology specialists like IBM.

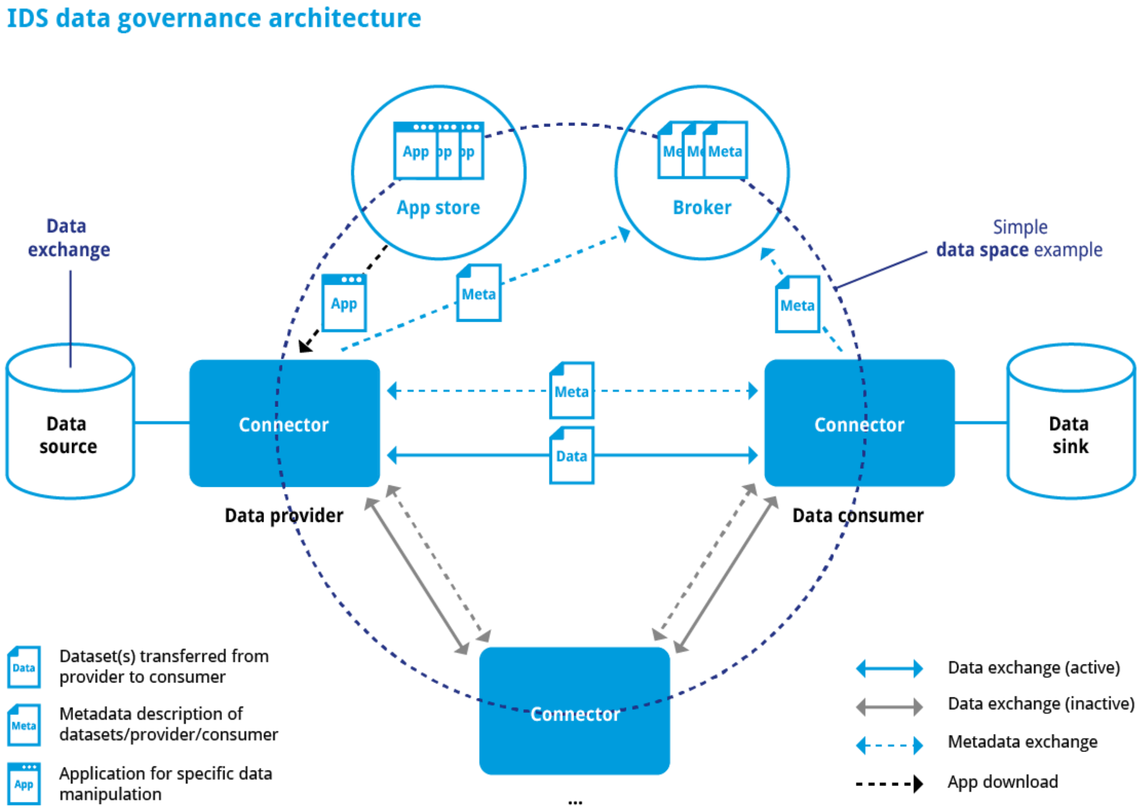

Figure 6: IDS data governance architecture

The core element of an IDS based data space is a "*connector*." It ensures that data rights can be governed. Figure6 illustrates the role of this element in the data flow between source (data provider) and sink (data consumer; Otto et al. 2019b, p. 59). With a connector, any data package or product can be "*wrapped up*" in instructions and rules for use. Technically, it is a dedicated software component allowing participants to exchange, share, and process data such that the data sovereignty of the data owner can be guaranteed. Depending on the type of configuration, the connector's tamper-proof runtime can host a variety of system services including secure bidirectional communication, enforcement of content usage policies (e.g., expiration times and mandatory deletion of data), system monitoring, and logging of content transactions for clearing purposes. As illustrated in Figure6, the functional range of a connector may be extended by (a) custom data apps, such as data visualization, provided in an app store and (b) a broker function to allow for product listings, such as a marketplace menu, and clearing services. A first connector implementation has been certified by IDSA for Deutsche Telekom's Data Intelligence Hub (Fraunhofer 2019).

2.8 “3 plus 2” is happening now

The shift in performance results from revenue per vehicle to revenue per trip can be managed and automated using several systems. U.S. airlines had invented three types of systems – two for data and one for analytics – to master a similar shift 40 years ago:

- An airline reservation system (ARS) for service inventory
- A frequent flyer program (FFP) for customer profiles
- A yield management system (YM) to match customers with service offerings

Today, with the wealth of consumer data from smartphones and social media, two additional systems will be required, particularly considering the anticipated data glut from 5G and IoT technology. Such new systems include:

- A *data exchange* to pool data, and enrich customer and service profiles
- A *data factory* to economize on data refinement and compliance management

Pioneers have already launched into this future. Airlines themselves are evolving systems capabilities to enrich profiles with *“social media sentiment analysis, shopping queries, stated preferences versus actual behavior*” (Sabre 2015, Fig. 4, p. 11) and other “*attribute-level*” data in order to expand personalization into “*ancillary purchases*” of cabin class upgrades, preferred seating with extra leg room, fast-track security screening, onboard food and beverage, in-flight Wi-Fi, lounge access, etc. (McKinsey 2017).

Leading mobility and travel companies are creating the systems required for these personalized (1-to-1) services bundles. Uber has built “*proprietary marketplace […] technologies [… that] include demand prediction, matching and dispatching, and pricing*” (Uber 2019, p. 162). TUI, the world’s largest travel and tourism company, is building its own yield management system (YM) to personalize its hotel offers. According to its CEO, Fritz Joussen, “*first tests have shown that 30 percent of customers are willing to pay five to ten euros more per night for their preferred room*” (Manager Magazin 2019, p. 68). TUI even plans to offer its new YM capabilities as a global platform service to third parties.

In closing and to highlight the YM trend as well as to link back to the “*what data could learn from food*” analogy used earlier for the data factory, McDonald’s, the iconic fast-food pioneer, “*in its largest acquisition in 20 years*” (Patton 2019), has purchased Dynamic Yield, a YM company (Bloomberg 2019). So, mobility companies can benefit from a proven blueprint that decomposes the complexity of automating the shift toward revenue per trip into a manageable set of system modules.

References

Belmonte, N.G. 2015. The Pulse of a City: How People Move Using Uber Engineering. Uber Engineering (August 18), https://eng.uber.com/pulse-of-a-city/