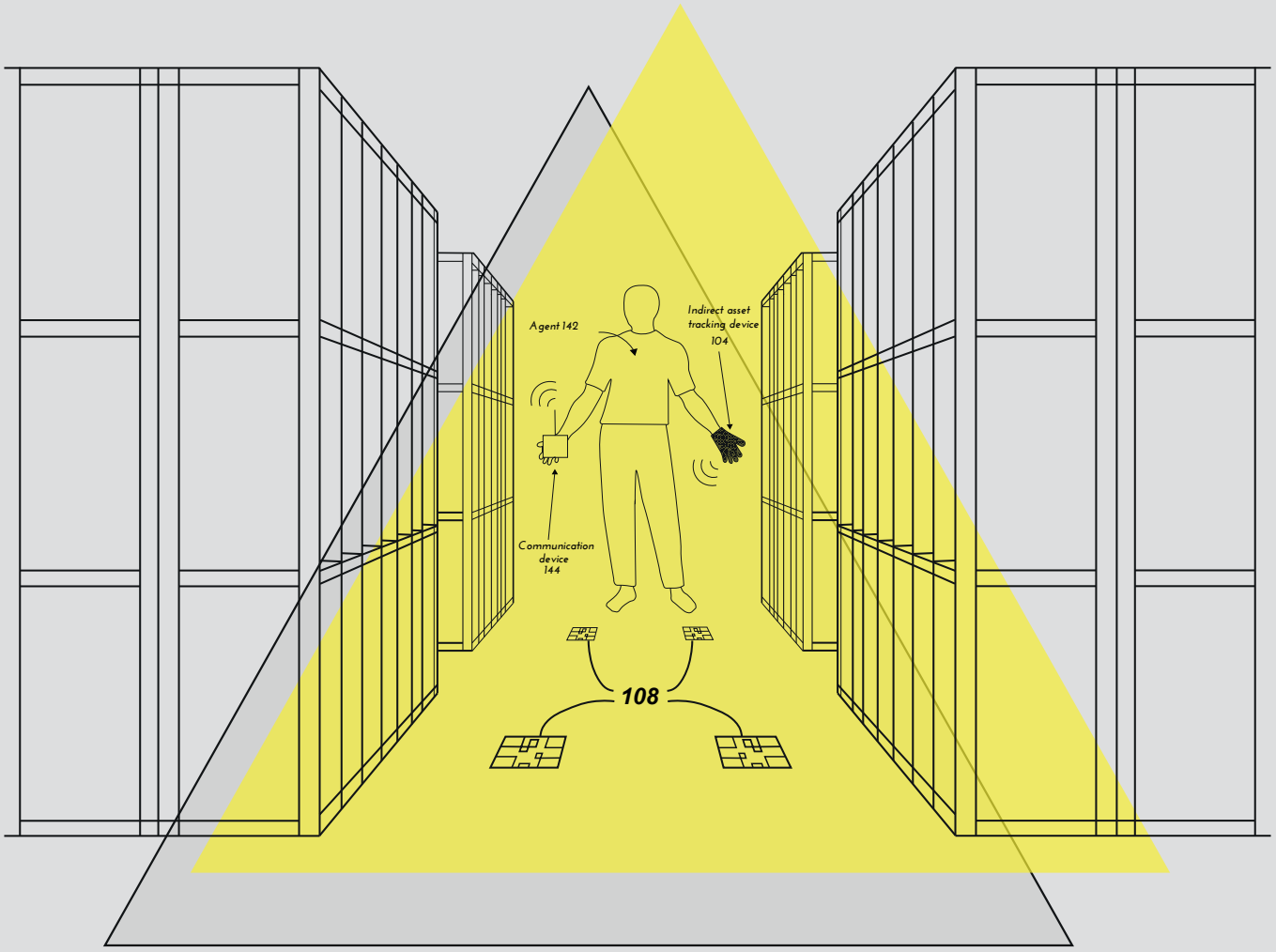
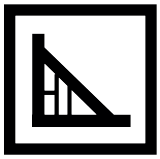
Data Archeogram

Mapping the datafication of work

January 2021



Autonomy



Autonomy is an independent think tank that provides necessary analyses, proposals and solutions with which to confront the changing reality of work today. Our aim is to promote real freedom, equality and human flourishing above all. To find out more about our research and work, visit

autonomy.work

Published 2021 by:

Autonomy Research Ltd
Cranbourne
Pilcot Road
Crookham Village
Hampshire
GU51 5RU

Author:

Armelle Skatulski

This work was supported by
the Alex Ferry Foundation and
the Arts and Humanities Research Council via Techne

Contents

Data Archeogram

1. Mapping the datafication of work

Data – from latin datum, something given (from dare, to give)
Capta – from latin captere, that which is taken or captured

To trace or map a worker's data stream is to situate the worker in the context of the datafication of employment and social relations at large (consumption/communication/work). We might understand 'datafication' as practices serving the purposes of dataveillance, targeted marketing, predictive analytics or algorithmic governance, for instance. These situate the worker at the intersection of vast 'data assemblages' or 'data ecosystems' in which the worker/user/consumer is connected to an array of technologies, databases, analysts and firms. Today's digital economy is constituted by relatively novel infrastructures that mediate these relations, including the Internet of Things (IoT) and its industrial form, **[1]** (big) data analytics, and cloud computing.

•How might we implement an archaeology or critical anatomy of worker data and of data flows and assemblages?
•What tools (conceptual, technological) might be used to do so?
•What might the value of such critical mapping be for an understanding of power in the economy today and for corresponding policy?

1.1. The worker as source of data in today's data assemblages: what datafication means

Such mapping attempts to grasp the web of ramifications linking the worker, as the active or passive source of data, to vast systems, strategies, and infrastructures of data capture, processing, storage, circulation, and monetisation (data rent). It looks to render apparent the processes by which datafication instantiates particular power relations that lead to an increasing informational asymmetry between data collectors and datafied subjects, i.e. digital inequity based on the exclusion of subjects from ownership of the data extracted from them. Mapping also makes visible the processes and techniques that obscure datafication from being unserstood as a practice that captures or extracts data/value from workers/users and the labour processes that constitute it.

---

[1] The Industrial Internet of Things is sometimes referred to as Industry 4.0 and commonly describes the sensor-enabled connectivity of machinery, devices, and workplace architecture more broadly, and their integration to advanced analytics via networks operating through the Internet.

Mapping can reveal in visual form how data is taken from subjects through highly valuable techniques and material infrastructures of capture that then exclude workers/consumers/users from its ownership, all whilst data increasingly becomes the most sought after raw material in the digital economy. **[2]** This is a digital form of 'primitive accumulation' rendered possible by processes of digital enclosures, like the black-boxing mechanisms **[3]** of corporate secrecy borders and patenting.

By helping to highlight these mechanisms, and denaturalise our common sense understandings of data use, mapping can show that data is not 'given' nor 'gratuitous' nor 'abstract' (digital). Rather, it is *taken* (capta) **[4]** and its capture, archiving and processing rely on material infrastructures and procedures with economic and environmental consequences. In this sense, the Data Archeogram represents only the beginning - at a necessarily low resolution - of what mapping the data economy can achieve for those seeking to understand and ultimately change the current system.

1.2. Asking after the flows within the data economy

This project first springs from the following questions:
•How is data collected from workers (where; when; with what technologies/methods)?
•How/where is it stored, processed, analysed, and shared (rented/sold)? How does it move, and to which marketplaces?
•How is worker data to be distinguished from user data, and how are they related?
•Mapping a worker's data stream(s) would allow us to locate strategic spaces to occupy with infrastructures and instruments that would affect and regulate this stream.

Mapping therefore constitutes a form of analysis that allows us to think critically and politically about data infrastructures as socio-technical entities and to account for the diverging, but no less interconnected, modes of intelligibility of data (e.g. what is known and visible to workers vs what is known and visible to data collectors, data scientists, business analysts, etc.) through which power imbalances are established or perpetuated.

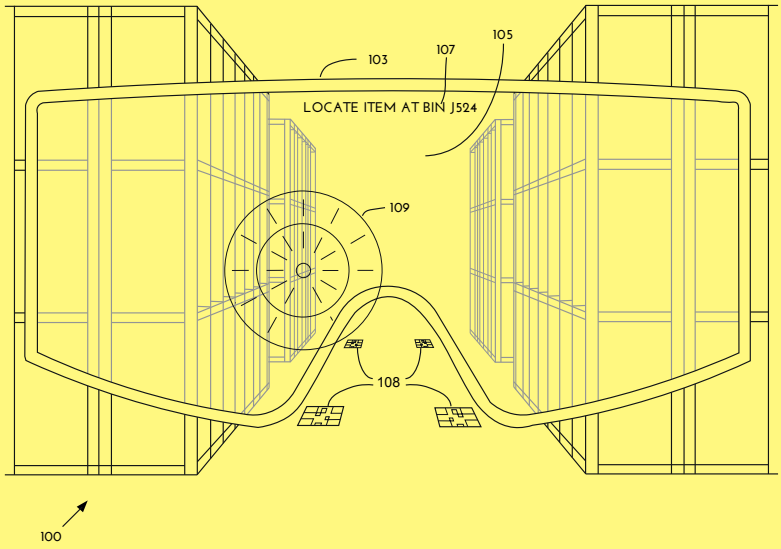
---

[2] UNCTAD. "Digital Economy Report 2019 - Value Creation and Capture: Implications for Developing Countries," 4 Sept. 2019, Geneva: United Nations Conference on Trade and Development, p. 29. Accessed at: <https://unctad.org/webflyer/digital-economy-report-2019>.
[3] On black-boxing mechanisms see, Pasquale, F., *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge, Massachusetts; London, England: Harvard University Press, 2015.
[4] For a distinction between data and capta, see Kitchin, R. "Conceptualising data," in *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage Publications, 2014.

2. Map descriptor

2.1. From the connected workplace to the connected social factory

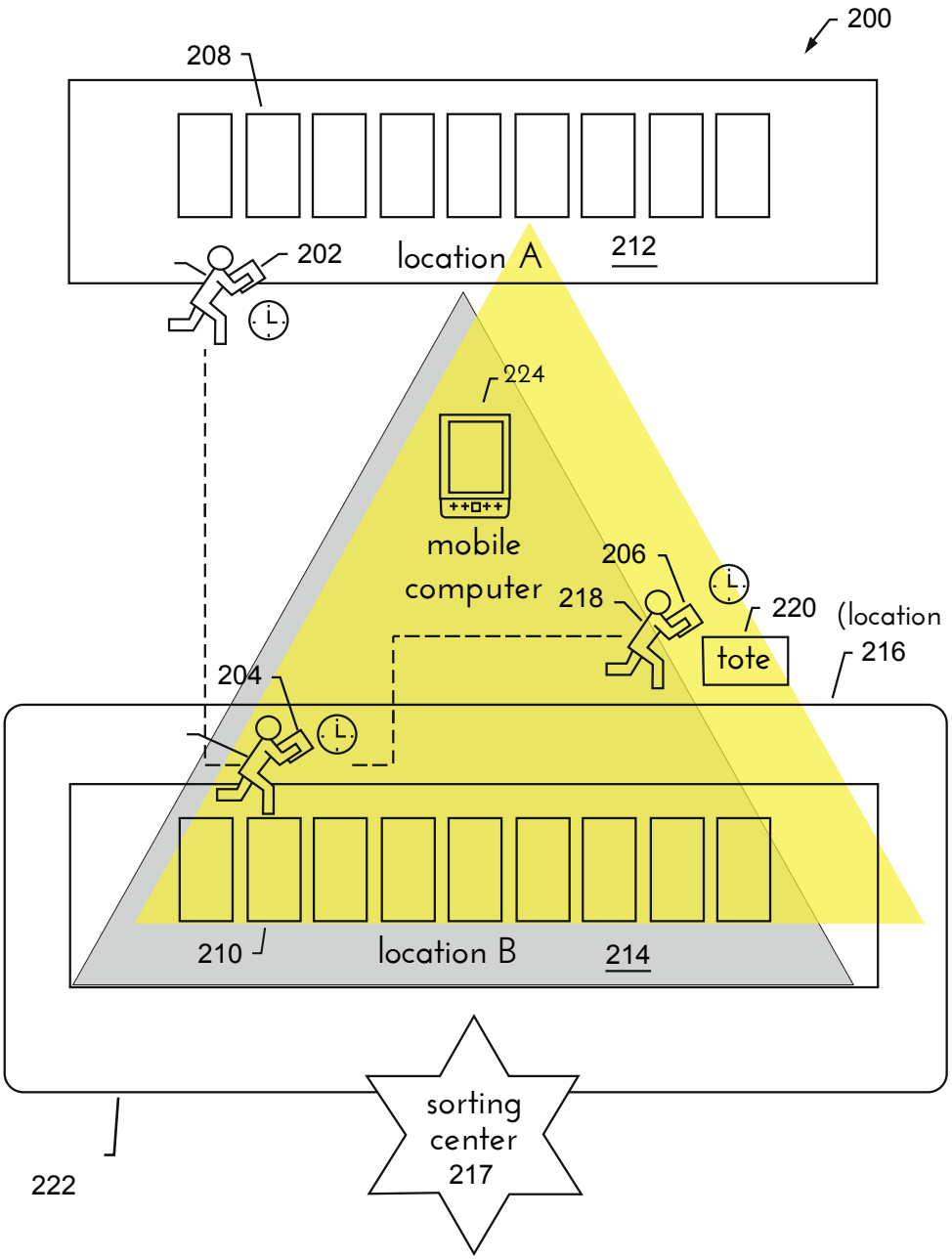
The map presents a distilled rendition of worker data streams in the context of the Industrial Internet of Things (IIoT), such as those occurring in the connected factory, **[5]** the intelligent or smart warehouse **[6]** or the sensor-fused office or shop. **[7]** For instance, IIoT applications are commonly used in smart warehouses to streamline the management of inventories, perform real-time worker activity monitoring (such as through the use of time-based movement maps), **[8]** or augment worker productivity via automated devices and processes.

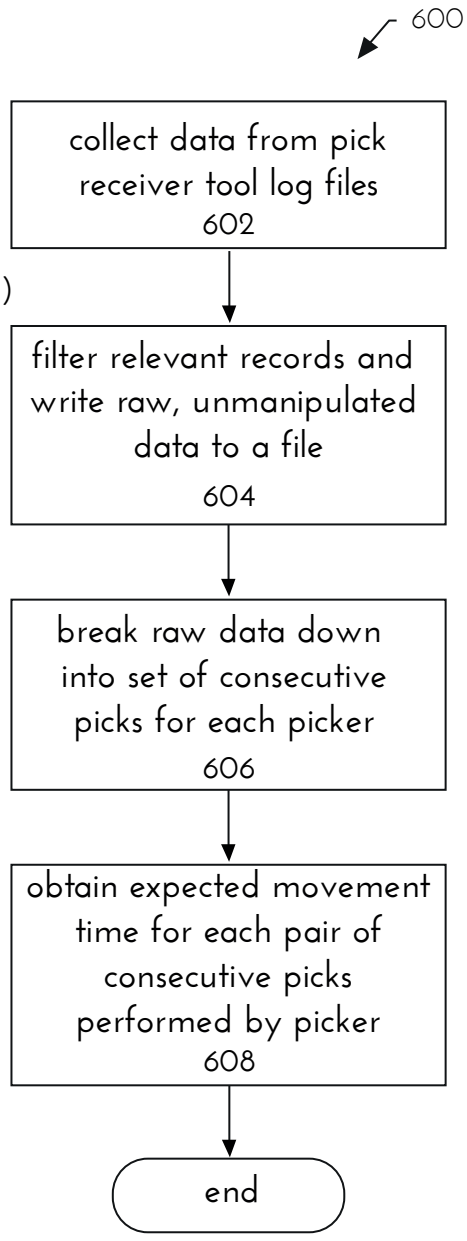


> *From "Augmented reality user interface facilitating fulfillment,"*
> *Amazon Technologies, patent #US 10,055,645 B1, August 21, 2018*
> *Accessed at the United States Patent and Trademark Office (USPTO)*

---

[5] Pooler, M., "Future factories: smart, connected but still with a human touch," Financial Times, 07.10.2019, accessed at: <https://www.ft.com/content/82fdd5-fa-be73-11e9-9381-78bab8a70848>.
[6] Pooler, M., "Amazon robots bring a brave new world to the warehouse," Financial Times, 25.08.2017, accessed at: <https://www.ft.com/content/916b93fc-8716-11e7-8bb1-5ba57d47eff7>.
[7] Sensor technologies are used to produce real-time occupancy metrics in office spaces or shops in the retail industry for instance: see Greenfield, Rebecca, "New Office Sensors Know When You Leave Your Desk," Bloomberg Businessweek, 14 February 2017. Accessed at: <https://www.bloomberg.com/news/articles/2017-02-14/new-office-sensors-know-when-you-leave-your-desk>. The software OfficeSpace is used to produce spatial occupancy metrics captured through sensor technology made by companies like SenzoLive and VergeSense which collect data from a company's security system, its badging systems, networks, and employee databases to "mash it all together" and produce activity reports such as about how workers collaborate, and how they utilize the companys real estate. See: <https://vergesense.com/resources/how-data-science-is-transforming-the-way-we-utilize>.
[8] "Time-based warehouse movement maps," Amazon Technologies, patent #US 7,243,001 B2, 10.07.2007, via the United States Patent & Trademark Office (USPTO).





> *From "Time-based warehouse movement maps"*
> *Amazon Technologies Inc., patent # US 4,243,001 B2, July 10, 2007*
> *Accessed at the United States Patent and Trademark Office (USPTO)*

Sensor-enabled tracking of workers has led to serious concerns around breaches of privacy rights not only in the workplace **[9]** but also outside of working hours, as in some cases data-driven monitoring takes place even while workers are off duty, via GPS devices, health monitoring technologies (e.g. Fitbit bracelets, Apple watches, etc.) **[10]** or mobile devices more generally. **[11]** Such workplace tracking has provoked pushback from organised labour. For instance, Uni Global Union have successfully negotiated a Europe-wide right to disconnect agreement **[12]** with Telefonica and Orange to establish the right for workers to switch-off their devices outside of working hours. In the UK, unions have used collective bargaining to secure protections for workers, such as the inclusion of a "privacy switch" in logistics to stop vehicles being tracked when workers are off duty. **[13]**

While the emphasis of the map is on worker data, it also includes data captured from users/consumers, outside the workplace, to show the generalised form which datafication now takes, affecting both the spheres of work (employed and domestic) and those of consumption, communication/social interaction, leisure, and so on, in ever more expansive and granular forms. As a consequence, the distinction between waged labour and user activity/productivity can only be maintained on formal or rhetorical grounds, as they both constitute prime sources of data, the principal raw material from which digital ecosystems make capital gains today.

Hence, the 'connected' workplace - now joined to vast infrastructures of data collection, storage, analytics and rent via sensor technologies - is also represented in terms of its relation to the 'connected social factory': the connected 'worker' and the connected 'user' constitute two sides of the same coin, so to speak: worker/user.

---

[9] Ong, Thuy. "Amazon patents wristbands that track warehouse employees' hands in real time," The Verge, 01.02.2018, accessed at: <https://www.theverge.com/2018/2/1/16958918/amazon-patents-trackable-wristband-warehouse-employees>.
[10] McGee, Suzanne, "How employers tracking your health can cross the line and become Big Brother," The Guardian, 01.05.2015. Accessed at: <https://www.theguardian.com/lifeandstyle/us-money-blog/2015/may/01/employers-tracking-health-fitbit-apple-watch-big-brother>.
[11] Saner, Emine. "Employers are monitoring computers, toilet breaks – even emotions. Is your boss watching you?", The Guardian, 14.05.2018. Accessed at: <https://www.theguardian.com/world/2018/may/14/is-your-boss-secretly-or-not-so-secretly-watching-you>. See also: Wonil Lee, Ken-Yu Lin, Edmund Seto, Giovanni C. Migliaccio, "Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction," Automation in Construction, Vol. 83, 2017, pp. 341-353. Accessible at: <https://doi.org/10.1016/j.autcon.2017.06.012>.
[12] Pakes, Andrew, "Our always-on work culture hurts everybody – this year, fight for your right to switch off," The Independent, 08.01.2020. Accessible at: <https://www.independent.co.uk/voices/unions-work-life-balance-flexible-working-a9275576.html>.
[13] Pakes, Andrew, "Why Big Data is becoming the front line in the battle for workers' rights," Left Foot Forward, 17.02.2020. Accessible at: <https://leftfootforward.org/2020/02/how-big-data-is-now-the-front-line-in-workers-rights/>; see also: https://congress.tuc.org.uk/composite-06-collective-voice-and-new-technology/#sthash.ndVRcxpu.Rex3vu7p.dpbs.

2.2. Domains of the data economy represented

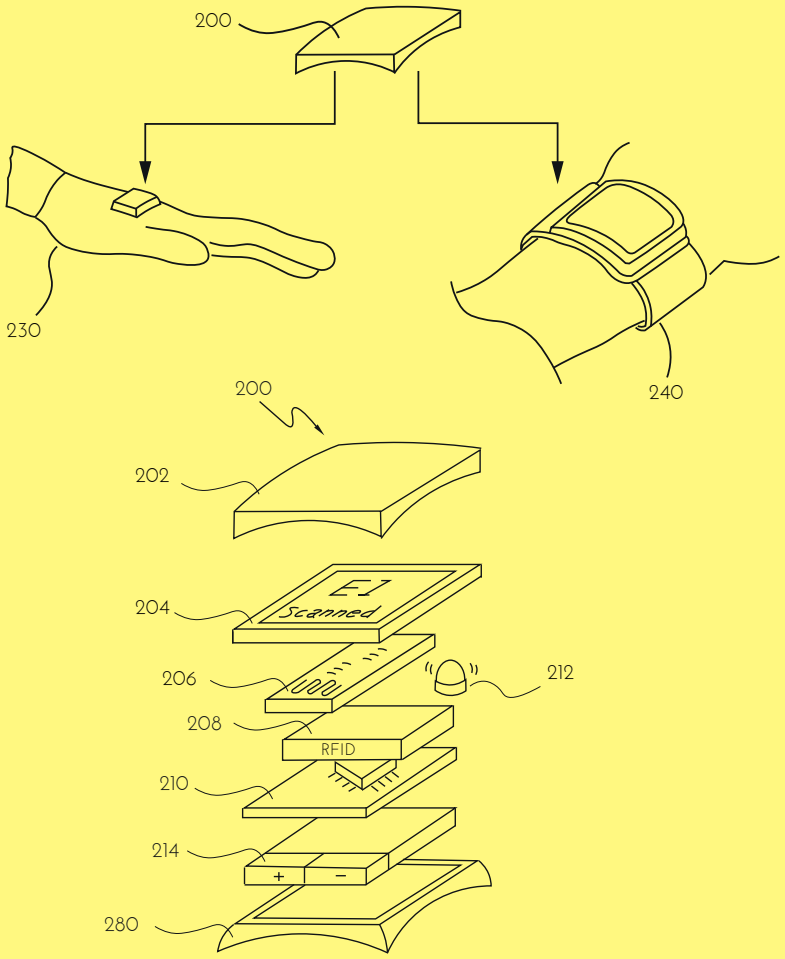
The map represents four domains or levels of the data economy:

1. The level of the worker (and user) where data mining or data extraction is enabled by sensor-fused devices, such as wearables or items embedded with Radio Frequency Identification (RFID) sensors, optical detection sensors, or ultrasonic transducers.

2. The connected workplace and the connected social factory (inscribed within the IIoT and the IoT respectively) which form the contexts for data extraction.

3. Infrastructures of data storage, processing, and analytics, including infrastructures enabling the training of Artificial Intelligence (AI).

4. Infrastructures of data rent and usage.



> *From "Wireless identifier based real time item movement tracking"*
> *Amazon Technologies Inc., patent # US 9,900,061 B1, Feb. 20, 2018*
> *Accessed at USPTO*

2.3. From data mining to algorithmic governance and back: data's life cycle

The map also aims to show the broader life cycle of data from initial data mining, to its processing as part of AI training, which in turn fuels prediction modelling, targeted marketing, behavioural 'futures' (i.e. behavioural predictions) and, more generally, algorithmic governance. **[14]** In other words, the conditioning of behaviour through algorithm-led processes in the (imbricated) spheres of work and consumption is articulated as one outcome tof the same process of datafication.

The map refers to the user as 'product' to express how such conditioning, mainly by way of choice architectures **[15]** (the influencing of consumer choices by design, i.e. nudging) **[16]** leads to the production of behaviour satisfying the interests of marketers or (IoT) corporations.

The term governance is therefore understood in a broad sense to encompass political governance processes as well as data-driven managerial techniques (e.g. data-driven surveillance or dataveillance and productivity monitoring), targeted advertising, predictive policing, etc.

The wider, outer arrows of the map show that algorithmic governance constitutes data mining's ultimate 'product', while in turn conditioning data mining itself, as algorithm-led processes are used to power data collection, processing and analytics, in a cyclical form.

---

[14] On algorithmic governance see for instance: Katzenbach, C. & Ulbricht, L. (2019). Algorithmic governance. Internet Policy Review, 8(4). DOI: 10.14763/2019.4.1424; Eyert, Florian; Irgmaier, Florian; Ulbricht, Lena (2018) : Algorithmic social ordering: Towards a conceptual framework, In: Getzinger, Günter (Ed.): Critical issues in science, technology and society studies: Conference proceedings of the 17th STS Conference Graz 2018, 7th-8th May 2018, ISBN 978-3-85125-625-3, Verlag der Technischen Universität Graz, Graz, pp. 48-57, DOI: 10.3217/978-3-85125-625-3. Accessed at: <https://www.econstor.eu/handle/10419/191592>.

[15] Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2013). Choice architecture. In E. Shafir (Ed.), The behavioral foundations of public policy (pp. 428-439). Princeton, NJ: Princeton University Press.

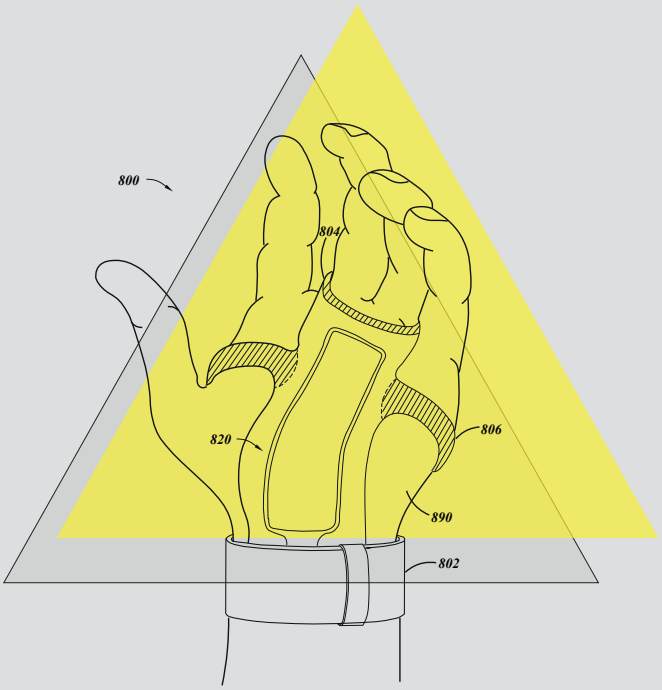
[16] See for instance, Karen Yeung (2017) "'Hypernudge': Big Data as a mode of regulation by design," Information, Communication & Society, 20:1, 118-136, DOI: 10.1080/1369118X.2016.1186713.

2.4. Data capture devices featured

2.4.a. "Wearable RFID device with manually activated RFID tags" (Amazon Technologies Inc.) [17]

Figure 1 on the map represents a wearable *Radio Frequency Identification* (RFID) device which may include "one or more manually activated RFID tags configured to transmit unique RFID signals in response to a manual activation thereof." **[18]** It may be worn about any aspect of a user's body (e.g. hand, a wrist, or an arm) and be manually activated to transmit signals that are consistent with instructions associated with particular tasks. This could, for instance, be integrated into a glove that the worker might wear.

While the general functionality of the RFID device is described as the remote operation of systems, or appliances (i.e. the transmitted signals lead to a sequence of actions such as operating a machine once translated by an application server), the patent emphasises that the enabling of worker identification or authentication constitutes one of the main advantages of RFID wearables over traditional remote controls and is hence, a crucial function of such a device. The patent also describes how manually triggered RFID tags can be used in combination with RFID tags that are actionable without touch (i.e. enabling 'passive scanning') but are associated with a particular worker, thus describing a form of worker surveillance in all but name.



> *From "Wearable passive scanning device"*
> *Amazon Technologies Inc.*
> *Patent # US 9,900,061 B1, Feb. 20, 2018*
> *Accessed at USPTO*

[17] "Wearable RFID devices with manually activated RFID tags," Amazon Technologies Inc., patent # US9,811,955 B2, 07.11.2017, via the United States Patent and Trademark Office (USPTO), accessible at: <https://www.uspto.gov/>.
[18] Ibid.

The device could be used in large inventory systems (e.g. in e-commerce warehouses) to enable the self-authentication or identification of workers during the movement of items from warehouse shelving to packaging and readying for transport and delivery. It may also be used by the worker to confirm her or his performance of a given task. **[19]** If a worker attempts to enter a secure location without manually triggering an RFID tag, they may be identified 'passively' through non-manually activated tags (such as held in identification badges) and permitted to choose to use RFID tags actively and signal their presence voluntarily.

As per the example in Figure 2 described below, the information associated with RFID signals **[20]** is collected in a data store operating in conjunction with one or several application servers. More broadly, the patent describes how the RFID system can be integrated within marketplaces via external and internal networks.

An RFID tag does not need to operate within a line of sight, unlike barcodes or QR codes. Therefore "RFID tags may be concealed or embedded into many different types of objects of any size or shape, as well as humans or other animals." **[21]** Additionally, an RFID tag can transmit signals in many different formats and at many different frequency levels. These elements raise serious ethical concerns and alarms around possible privacy breaches and the question of whether workers are informed of these qualitative differences (i.e. tunable frequencies).

2.4.b. "Ultrasonic bracelet and receiver for detecting position in 2D plane" (Amazon Technologies Inc.)

Figure 2 on the map represents a bracelet enabling the *ultrasonic tracking* of a worker's hands, which may be used to monitor the performance of assigned tasks. The device is integrated in an 'inventory system' which includes a management module operatively coupled with various ultrasonic units. Data collected via monitoring is held within an operational data store, and data access control services and business logics are executed via an application server (see figure 6 on the map) which is capable of generating content (e.g. text, graphics, audio, and/or video) transferrable to users/clients via the Web, for instance. **[22]**

---

[19] Column 7 of patent # US9,811,955 B2.
[20] An RFID reader can be configured to "capture, evaluate, transmit or store any available information regarding signals from one or more RFID tags, including information regarding any attributes of the signals," e.g. "sensed signal strengths or intensities, angular directions or ranges" to the tags. (Ibid, column 9)
[21] Column 4 of patent # US9,811,955 B2.
[22] "Ultrasonic bracelet and receiver for detecting position in 2D plane," Amazon Technologies Inc., patent #US 9,881,276 B2, 30.01.2018, via USPTO.

A data store refers to "any device or combination of devices capable of storing, accessing, and retrieving data." It may include "any combination and number of data servers, databases, data storage media, in any standard, distributed, or clustered environment." **[23]** The data store is integrated with an application server that enables it to obtain, update or process data in response to instructions.

Types of data stored may comprise: production data and user information; inventory information (items identification; storage location identification); log data used for reporting, analysis or other purposes (i.e. monitoring).

Ultrasonic bracelets can be used in inventory systems, such as those operative in mail order warehouses, supply chain distribution centres, airport luggage systems, custom-order manufacturing facilities.

2.4.c. "Physiological data collection" (Fitbit Inc.)

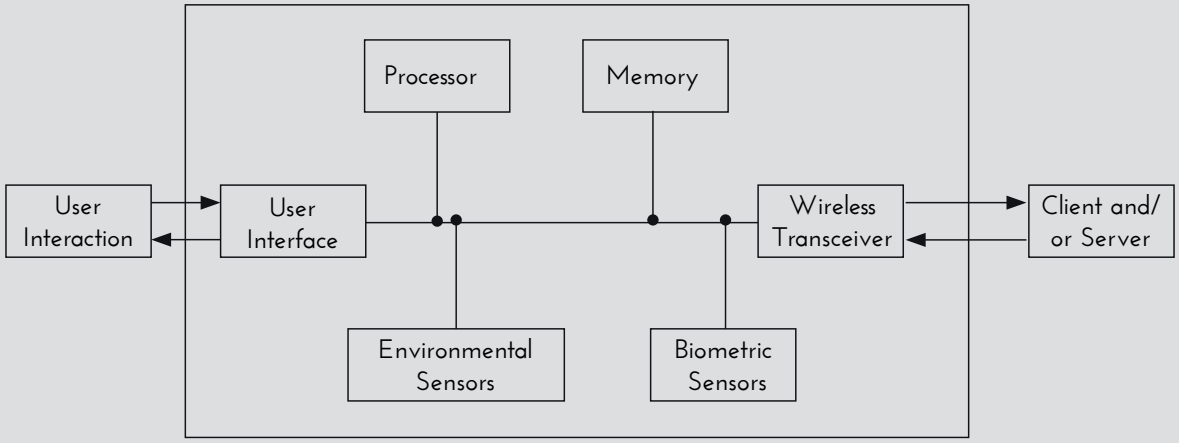
Figure 3 on the map represents "a *portable biometric monitoring device* to take a heart rate measurement from a side-mounted optical heart rate detection sensor." (emphasis added) **[24]** The patent includes several embodiments of biometric data collection devices and physiological information detection systems using algorithms. The main purpose of such devices, in summary, is to find out about the wearer's body.

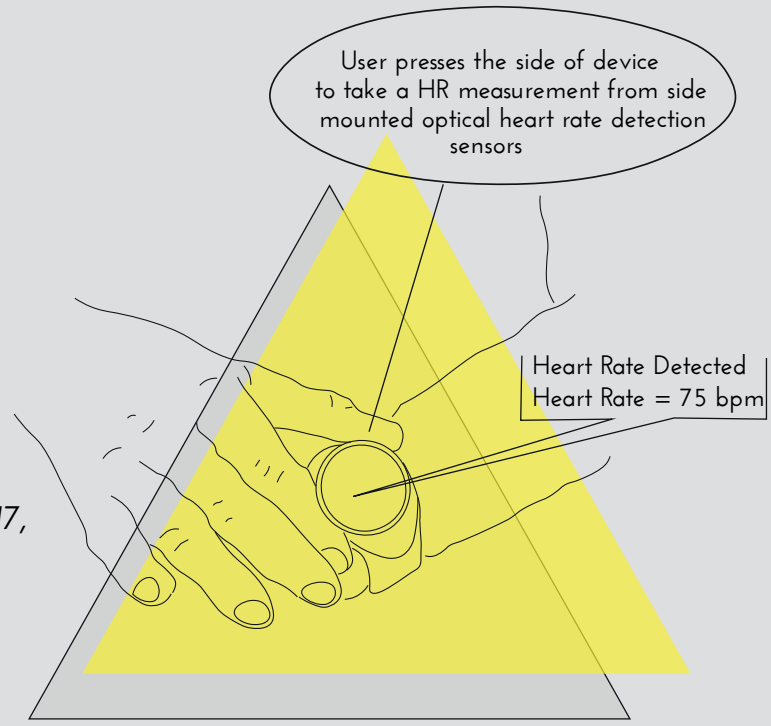
Such devices can collect one or more types of physiological and/or environmental data from embedded sensors and/or external devices and can "relay such information to other devices, including devices capable of serving as Internet-accessible data sources." Biometric data collected can be accessed using a web browser or other network-based application by clients or streamed by employers, real-time, via third-party companies, thus enabling health-based surveillance of workers while they are off-duty. For instance, a US based company, Regal Plastics accessed their employees' fitness and location metrics via their United Health insurance group account, in real-time. **[25]** Data captured from a device on an employee's wrist was delivered through streaming via an app on their boss's mobile phone.

---

[23] Ibid, see columns 21 and 22.
[24] "Physiological data collection," Fitbit Inc., patent #US 9,662,053 B2, Fitbit Inc., 30.05.2017, via USPTO.
[25] For examples, see Rowland, Christopher, "With fitness trackers in the workplace, bosses can monitor your every step — and possibly more," The Washington Post, 16.02.2019, accessed at: <https://www.washingtonpost.com/business/economy/with-fitness-trackers-in-the-workplace-bosses-can-monitor-your-every-step--and-possibly-more/2019/02/15/75ee0848-2a45-11e9-b011-d8500644dc98_story.html>

Biometric and environmental data (e.g. location) produced by consumers can be collected and anonymised by third-parties via applications to produce profiles sold on data marketplaces or to infer health-based behaviours and to market products accordingly. Pharmaceutical companies may for instance direct research and produce new drugs in line with such predictions. **[26]** While offering the prospect of important capital gains for data collectors and brokers, the mining of such data presents privacy risks for consumers. Fitbit, which was acquired by Google via its parent company Alphabet in 2019, **[27]** claims that its privacy policy prohibits it to share any identifiable information, but research has shown that anonymised data can easily be "re-identified." **[28]**





> *From "Physiological data collection"*
> *Fitbit Inc.*
> *patent #US 9,662,053 B2, 30.05.2017,*
> *Accessed via USPTO*

---

[26] Frazee, Gretchen, "Google bought Fitbit. What does that mean for your data privacy?," PBS News Hour, 01.11.2019, accessed at: <https://www.pbs.org/newshour/economy/making-sense/google-bought-fitbit-what-does-that-mean-for-your-data-privacy>.
[27] Chapman, Michelle, "Google buys Fitbit for $2.1 billion," PBS News Hour, 01.11.2019, accessed at: <https://www.pbs.org/newshour/nation/google-buys-fitbit-for-2-1-billion>.
[28] Ibid, see also: Sweeney, Latanya, "k-anonymity: a model for protecting privacy." International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570. Accessed at: <https://epic.org/privacy/reidentification/Sweeney_Article.pdf>.

2.5. The connected workplace

Figure 4 on the map shows how the worker is connected via *sensor-fused devices* to the connected workplace (e.g. the smart warehouse in e-commerce, or the retail shop) where on-premise computing through edge software (see *edge computing* in the glossary) is combined with cloud-based services (see *cloud computing*).

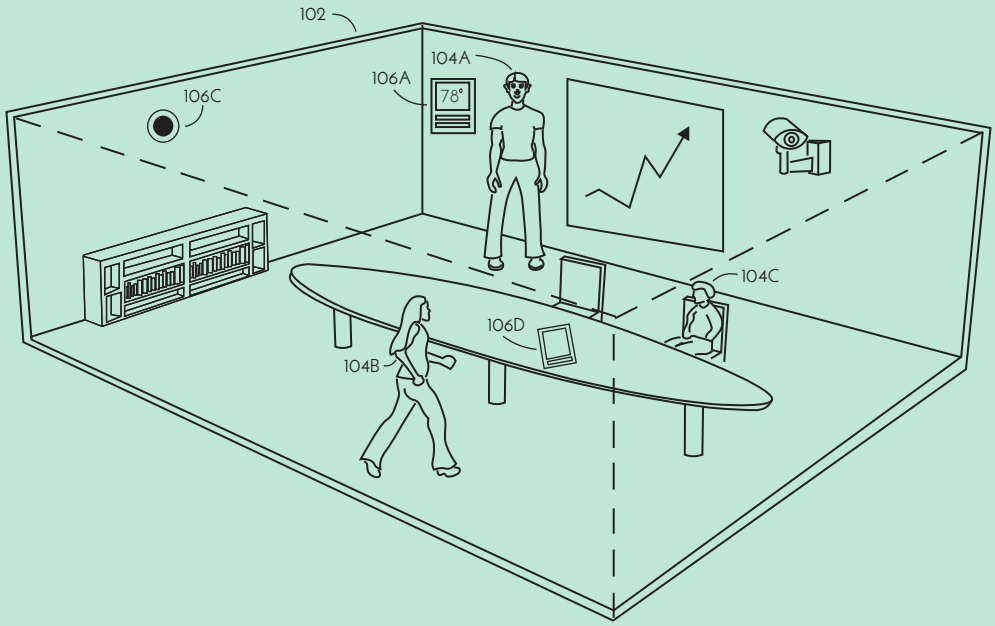
Sensor-based connectivity of the workplace has allowed not only for worker productivity monitoring and streaming from the smart warehouse to the office, but also for predictive analysis relating to "quality monitoring". A patent filed recently by Microsoft under the title "Meeting Insight Computing System," **[29]** describes a system designed to infer and predict "quality scores" for employee meetings from data such as body language, facial expressions, room temperature, time of day, and the number of people gathered. **[30]** Microsoft's quality scoring system may operate through the hybrid integration of sensor-based tracking and cloud computing (e.g. Microsoft's Azure IoT). **[31]**

2.6. The connected social factory

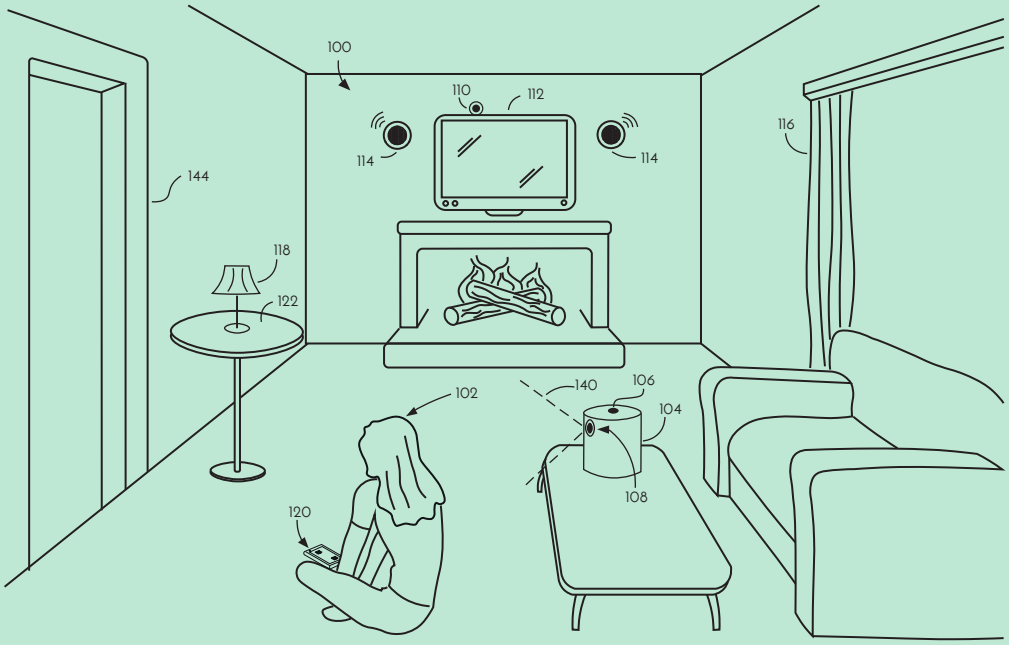
Figure 5 on the map represents the potential connectivity of various - largely domestic - smart objects with respect to larger networks (e.g. cloud-based computing services via the Internet) in the context of the IoT through sensor-based technology. While the diagram refers to the processing of audio data in voice-enabled devices and voice-controlled objects, **[32]** the reach of object and environment connectivity via sensors is ever expanding. For instance, objects may be controlled via *gesture prediction algorithms* allowing for gesture detection as subtle as the closure of an eye.

---

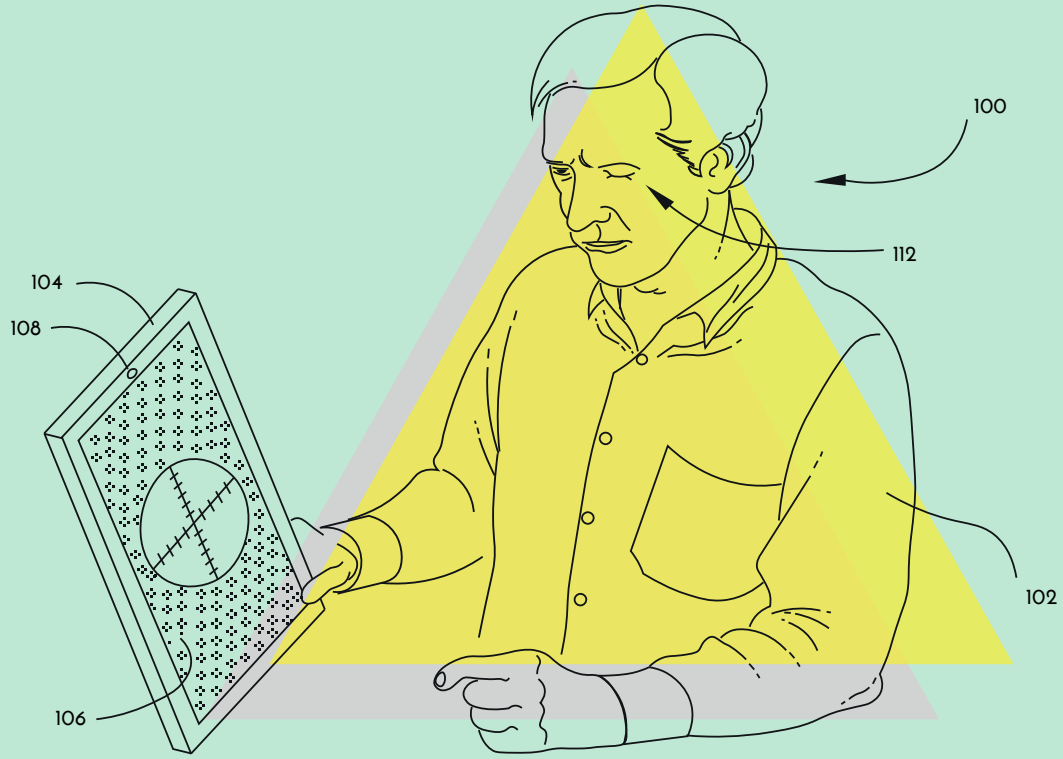
[29] "Meeting Insight Computing System," Microsoft Technology Licensing, patent #US2020/0358627 A1, 12.11.2020, via USPTO, accessible at: <http://appft.uspto.gov/netahtml/PTO/index.html>.
[30] Bishop, Todd, "Microsoft patents tech to score meetings using body language, facial expressions, other data," GeekWire, 28.11.2020, accessed at: <https://www.geekwire.com/2020/microsoft-patents-technology-score-meetings-using-body-language-facial-expressions-data/>.
[31] See: https://azure.microsoft.com/en-gb/services/iot-hub/.
[32] "Processing spoken commands to control distributed audio outputs," Amazon Technologies, patent #US9,858,927 B2, 02.01.2018, via USPTO, accessible at: <http://appft.uspto.gov/netahtml/PTO/index.html>.



> From "Meeting Insight Computing System"
> Microsoft Technology Licensing
> Patent #US 2020/0358627 A1, Nov., 12, 2020, via USPTO



> From "Associating Semantic Identifiers with Objects"
> Microsoft Technology Licensing
> Patent #US 10,783,411 B2, Sep. 22, 2020, via USPTO



> *From "Providing user input to a computing device with an eye closure"*
> *Amazon Technologies Inc., patent # US 9,483,113 B1, Nov. 1, 2016, via USPTO.*

A 2016 Amazon Technologies patent describes a system allowing for user input to a computing device via an eye closure to access additional features (e.g. a graphical overlay enabling the selection of functions relating to an application, game or content). **[33]** Critically, for the device to be able to identify a user's eye closure and its characteristics (e.g. duration), it relies on an algorithm trained on similar type of 'historical' data (i.e. recorded from previous usage). Today, it is possible for such devices to be trained on a continuous basis at the 'edge', i.e. with data collected through imaging sensors by the very device handled by the user and analysed via edge computing software (on the edge training of AI, see *neural networks*).

This means that an event as subtle as an eye closure has become an event of data capture used to train algorithms, possibly 'real-time', and a form of invisible labour of a value now intensified by the capabilities of edge computing, allowing for the spread of new kinds of predictive interfaces. **[34]**

---

[33] "Providing user input to a computing device with an eye closure," Amazon Technologies, patent #US 9,483,113 B1, 01.11.2016, via USPTO, accessible at: <http://appft.uspto.gov/netahtml/PTO/index.html>.
[34] See for instance, Çağla Çığ Karaman, Tevfik Metin Sezgin,"Gaze-based predictive user interfaces: Visualizing user intentions in the presence of uncertainty," International Journal of Human-Computer Studies,Volume 111,2018, pp.78-91,accessible at: <https://doi.org/10.1016/j.ijhcs.2017.11.005>.

2.7. Infrastructures of storage, process, and analytics

Data warehouses, data lakes, and databases constitute infrastructures of storage, processing and analytics, which are increasingly supported by cloud computing. **[35]** Businesses use a combination of these three types of infrastructures to store and analyse data. Their integration is enabled by a variety of applications and services. **[36]**

A data centre or server farm is the physical entity or facility where servers or computing systems are set-up. They are used by organisations to store critical (business) applications and data. **[37]** By distinction, a data warehouse is a data architecture or system on a server, whether in a data centre or cloud-based, which allows an organisation to aggregate data from multiple sources and run powerful analytics. **[38]**

A typical data collection and processing workflow would consist in dropping data in a data lake where data is explored and prepared. The selected data is then moved to a data warehouse where it is analysed. Such data can then be used to inform future reporting activities (e.g. through data visualisation on dashboards). However, data can also be placed in a data warehouse directly. An organisation's workflow will depend on the types of analytics they intend to perform, i.e. big data analytics vs. other forms of analytics.

A data warehouse is a repository where data is stored in a structured manner. It is designed for the aggregation of disparate data and the implementation of analytics. **[39]** Data flows into a data warehouse from multiple sources, from transactional systems to relational databases, **[40]** on a regular basis.

For instance, Acxiom Corporation, said to have amassed the world's largest commercial database on consumers and to be the leader in 'database marketing', may gather data from sources as varied as public records, consumer surveys, online consumer tracking, bank card transactions, etc. to build large relational data infrastructures (see 'relational data' in glossary terms).

---

[35] https://www.snowflake.com/trending/data-lake-vs-data-warehouse.
[36] https://aws.amazon.com/redshift/lake-house-architecture/; https://www.ibm.com/uk-en/products/integrated-analytics-system.
[37] https://www.ibm.com/cloud/learn/data-centers.
[38] https://www.ibm.com/cloud/learn/data-warehouse.
[39] See: IBM (2018) "Data Warehouse platforms demystified," accessed at: <https://www.ibm.com/account/reg/uk-en/signup?formid=urx-32400>; and <https://learn.panoply.io/data-warehousing-trends-report-2018>.
[40] A collection of data items with pre-defined relationships between them through a tabular structure: <https://www.ibm.com/analytics/relational-database>.

Some of their customers include big banks like Wells Fargo and HSBC, investment services companies like E*Trade, large automakers such as Toyota and Ford, department stores like Macy's. **[41]**

In a data warehouse, data is organised in a tabular format to allow for searches using Structured Query Language (SQL) (a standard language used to access, communicate with or manipulate data warehouses or databases). Some applications, such as used in machine learning or AI powered big data analytics, can access data even when semi-structured or unstructured. Data can be accessed from a data warehouse through business intelligence (BI) tools, **[42]** SQL clients, **[43]** and other types of applications, by users such as business analysts, data engineers, data scientists, and decision makers.

A *data lake* is a method of storing data in its native format. Hence in a data lake, data can be found in structured, semi-structured, and unstructured forms (e.g. videos, images, unparsed text forms such as emails, etc.). **[44]** It can store relational data from transactional systems, and non-relational data, e.g. from mobile devices, social media interactions, and IoT devices. **[45]**

As data in a data lake can be sometimes found in unstructured forms which are more difficult to anonymise (i.e. videos and photographs), data lakes require the use of data governance tools that guarantee compliance with data regulation to ensure the protection of privacy rights.

Ungoverned data lakes are referred to as 'data swamps' containing data that cannot be trusted (as it may breach privacy rights or security issues). **[46]** As data in data lakes is often un-curated, data lakes are the playground of data scientists and data analytics experts, not business analysts.

Databases, also referred to as 'transactional databases,' store data captured 'as-is' from a single source, often from transactional systems. They are designed to support the running of production systems – e.g. websites, banks or retail stores. **[47]**

---

[41] Singer, Natasha, "Mapping, and Sharing, the Consumer Genome," New York Times, 16 June 2012, accessed at: <https://www.nytimes.com/2012/06/17/technology-/acxiom-the-quiet-giant-of-consumer-database-marketing.html>.
[42] See for instance <https://aws.amazon.com/quicksight/> or <https://azure.microsoft.com/en-gb/overview/what-are-business-intelligence-tools/>.
[43] An SQL Client is a program written in Java allowing you to view the contents of a database. See: <https://developer.ibm.com/technologies/databases/articles/dm-0312bhogal/>.
[44] IBM, "Five myths about the data lake," accessed at: <https://www.ibm.com/downloads/cas/BOGPM93R>.
[45] See: https://www.ibm.com/uk-en/analytics/data-lake; and for instance: https://azure.microsoft.com/en-gb/solutions/data-lake/.
[46] https://developer.ibm.com/technologies/analytics/articles/-ba-data-becomes-knowledge-2/.
[47] https://looker.com/databases/transactional.

While data warehouses use de-normalised schemas, **[48]** in a database data is stored through highly normalised schemas and organised in tables and columns. Normalization consists in decomposing tables to eliminate data redundancy (repetition) and undesirable characteristics, thus improving data integrity. **[49]**

A data mart is a smaller form of data warehouse that serves the needs of a particular business entity or sector, e.g. finance, marketing, sales, etc. A mart may also constitute a subset of a data warehouse. **[50]**

Data governance tools are used to inform the architecture and management of all types of data infrastructures (data warehouses, data lakes, and databases) to comply with relevant data regulations (e.g. the General Data Protection Regulation (GDPR) in Europe or the California Online Privacy Protection Act, in the USA for instance) **[51]** while the application of such tools also contributes to the valorisation of data, as users will be more inclined to buy or use trusted data. **[52]**

2.8. Hybrid infrastructures

These are infrastructures characterised by the integration of on-premise application servers (i.e. held in corporate data centres) with cloud-based computing services (e.g cloud-based storage and analytics in cloud data warehouses, such as AWS Redshift, Microsoft Azure, or Google BigQuery). Such integration is enabled by applications and services sold by platforms that act as intermediaries between data centres and cloud-based services users. Such mediation takes place via the Internet, while cloud-based services (storage, analytics) can take place in *Virtual Personal Clouds* (VPC) against rent.

---

[48] A schema is the logical description of a collection of database objects, including tables, views, indexes, and synonyms. See: <https://www.tutorialspoint.com/dwh/dwh_schemas.htm>. Examples of de-normalised schemas are the Star schema or Snowflake schema. See: <https://www.ibm.com/support/knowl-edgecenter/en/SS9UM9_9.1.2/-com.ibm.datatools.dimensional.ui.doc/topics/c_dm_dimschemas.html>.
[49] Li, Lorraine. "Database Normalisation Explained," towardsdatascience.com, 02.07.2019, accessed at: <https://towardsdatascience.com/database-normalization-explained-53e60a494495>.
[50] https://try.panoply.io/modern-data-management/.
[51] See <https://globaldatahub.taylorwessing.com/hot-topic-gdpr; and https://globalda-tahub.taylorwessing.com/article/regulation-of-big-data-in-the-united-states>.
[52] https://www.ibm.com/uk-en/analytics/use-cases/governing-data-lake.

2.9. Neural networks

Neural networks are an area of Artificial Intelligence (AI) also known as Deep Learning (DL) (see *AI training* in the glossary). It is a type of machine learning whereby a computer learns tasks by analysing large data sets that usually have been labelled in advance (see *clickwork*). A neural net can consist of thousands or even millions of processing nodes that are densely interconnected in a manner reminiscent of the human brain. **[53]**

Figure 7 is a schematic representation of *adaptive neural networks* used in automatic speech recognition systems. Their performance can be improved by updating or retraining the neural networks during run time. 'Re-training' can be based on the output of a speech recognition system for instance at each utterance or at varying time scales. These networks can be thought of in terms of layers: with an input layer corresponding to the influx of data, an output layer corresponding to the output of predictions, and some hidden layers in between. **[54]**

Edge training of neural networks is now expanding along the trend of edge data centres and edge computing, enabling for faster processing and analysis of data on-premise. Neural networks can be trained on mobile devices for instance to counter latency or bandwidth limitations (see edge computing below). **[55]**

---

[53] Hardestym Larry, "Explained: Neural networks," MIT News Office, 14.04.2017, accessed at: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414; see also: <https://www.ibm.com/uk-en/cloud/deep-learning>.
[54] See "Adaptive neural network speech recognition models," patent #US 9,153, 231 B1, 06.10.2015, via USPTO.
[55] Raj, Bharath, "Deep Learning on the Edge," Medium, 24.06.2018, accessed at: https://towardsdatascience.com/deep-learning-on-the-edge-9181693f466c.

2.10. Infrastructures of data rent and usage

The map refers to data marketplaces as *platforms* where users can buy, rent or sell different types of data sets and data streams from several sources. **[56]** The map shows a distinction between data warehouses and data lakes users. Data Warehouses contain data organised in tabular formats and mainly accessed by business analysts, but also data scientists and data developers. As mentioned above, as data in data lakes is often uncurated, it is therefore mainly accessed by data scientists and data analytics experts, not business analysts. While there are 'giants' (e.g. Acxiom) leading the market of data rent/sale, there is an ever expanding number of data marketplaces at different scales of the market. **[57]**

---

[56] On IoT marketplaces see Deichmann, J, Heineke, K., Reinbacher, T., and Wee, D., "Creating a successful Internet of Things data marketplace," McKinsey, 7 October 2016, accessible at: <https://www.mckinsey.com/business-functions/mckinsey-digital-/our-insights/creating-a-successful-internet-of-things-data-marketplace>.
[57] See for instance the data marketplace platform Datarade acting as an intermediary between data brokers and buyers of commercial data globally: <https://about.datarade.ai/>.

2.11. Platforms

The term refers to a particular business model in which a platform acts as an intermediary between different groups by operating via the Internet and applications: for instance, companies such as Uber and Lyft connect drivers and customers; Facebook and Google bring together consumers, businesses and advertisers; while Amazon, beyond connecting sellers and buyers, builds and owns a great part of the infrastructures upon which digital economic exchanges depend (i.e. cloud computing; data storage; ...). **[58]** Platforms are designed to make capital gains by extracting data from the transactional exchanges between the groups to which they provide infrastructures of intermediation. **[59]** Platforms tend to monopolise a particular market, such as AWS and Microsoft for cloud computing, Google for targeted marketing, etc. The map shows that platforms intervene at every level of the data life cycle: extraction, processing, analytics, renting of hybrid infrastructures, cloud computing, AI-training, data rent or sale, wearables design, etc.

---

[58] Srnicek, Nick. "The challenges of platform capitalism: understanding the logic of a new business model," 20.09.2017, The Institute for Public Policy Research, accessed at: <https://www.ippr.org/juncture-item/the-challenges-of-platform-capitalism>.
[59] Ibid.

2.12. Internet infrastructures (including submarine cables)

The Internet infrastructures include nearly 750,000 miles of fibre optic submarine cables **[60]** which connect the continents to support online communication networks. **[61]** About 99 per cent of total international data transmissions run through these cables, **[62]** while fiber-optic technology constitutes the medium of choice for Internet backbone providers. The IoT and its industrial form operate through the integration of soft and hard infrastructures sustaining the Internet, which in turn enables cloud gateways to operate. The map indicates that Internet infrastructures intervene at most stages of the data life cycle.

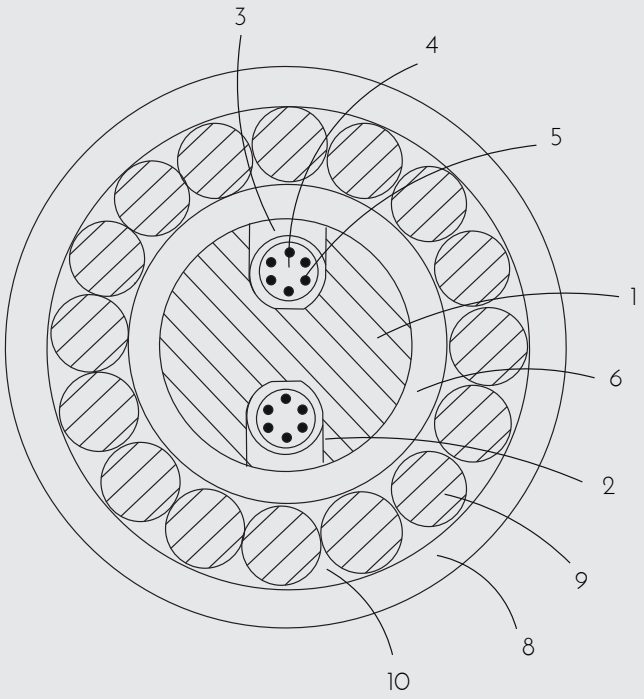
As the market pioneer and leader in cloud services, Amazon Web Services (AWS) has become increasingly active in the subsea cable market. AWS started its investments in the market in late 2016. It has for instance invested in ventures such as the 'Jupiter Cable' connecting Japan and the US (early 2017), became a consortium member of the Bay to Bay Express Cable System (BtoBE) linking Hong Kong and the US (2018), and acquired fiber pairs on MAREA (connecting Virginia and Spain).

The BtoBE consortium includes China Mobile International, Facebook (by its direct subsidiary Edge USA) and Amazon (by its wholly-owned, indirect subsidiary Vadata). **[63]** Google, alongside AARNet, Indosat Ooredoo, Singtel, SubPartners, Telstra, and Alcatel Submarine Networks (ACN) was involved in building a new international subsea cable system in Southeast Asia, called Indigo. **[64]**

---

[60] See Telegeography, The Submarine cable map," accessible at: <https://www.sub-marinecablemap.com/>.
[61] Satariano, Karl, "How the Internet Travels Across Oceans," New York Times, 10 March 2019, accessible at: <https://www.nytimes.com/interactive/2019/03/10/technolo-gy/internet-cables-oceans.html?mtrref=www.google.com&assetType=REGIWALL>.
[62] UNCTAD, "Digital Economy Report 2019," p. 11. Accessed at: <https://unctad.org/webflyer/digital-economy-report-2019>.
[63] See: https://www.submarinenetworks.com/en/systems/trans-pacific/btobe; and https://www.submarinenetworks.com/en/systems/trans-atlantic/marea.
[64] For an overview of subsea cables across the globe, see Ducket, Chris, "Indigo subsea cable made ready for use," ZDNet, 30.05.2019, accessed at: <https://www.zdnet.com/article/indigo-subsea-cable-made-ready-for-use/>.

Figure 8 on the map represents one possible embodiment of a long range submarine optical communication system patented by Alcatel Submarine Networks (ACN) and partner Nokia Solutions & Network Oy, which represents a cable network comprising terrestrial units and submarine units with fibre optic cables designed for different subsea depths. For a detailed description of the system see the patent filed with the European Patent Office (September 2020). **[65]** ACN built the Southeast Asia subsea cable system called Indigo, of which Google is one of the shareholders. **[66]**



> *From "Fiber optic cable having an extended elongation window"*
> *Cross-section view*
> *Alcatel NA Cable Systems Inc.*
> *Patent # US 4, 944,570, Jul. 31, 1990, via USPTO*

---

[65] "Equipment for Long Range Submarine Optical Communication," Alcatel Submarine Networks; Nokia Solutions & Networks Oy, 23.09.2020, via the European Patent Office, accessible at: <https://worldwide.espacenet.com/patent/search/family/066998329/publication/EP3713108A1?q=pn%3DEP3713108A1>.
[66] https://www.offshore-energy.biz/alcatel-submarine-networks-to-build-indigo-subsea-cable/.

3. Additional terminology (glossary)

AI training (and Machine Learning)

This refers to the training of Artificial Intelligence (AI), in particular machine learning (ML) algorithms, through the use of massive data sets and enormous computational capabilities. ML algorithms are programmed to recognise patterns across data sets and train themselves to make 'decisions' without human intervention. Such algorithms are widely active in the financial sector in automated trading for instance. **[67]** Significantly, some training data inputs are only obtainable through human preparation and annotation performed by clickworkers. Such data often plays an essential role in training AI. **[68]**

Analytics

Refers to the process of examining data sets to produce knowledge from them and inform decisions. These aim to uncover patterns from data sets to extract insights that can inform predictive models. Analytics today rely on specialised software and systems that integrate machine learning algorithms and other AI-powered capabilities. There are four broad classes of analytics: analytics based on pattern recognition at the level of data mining; data visualization and algorithm-led visual analytics (e.g. through graphic summaries and time-series maps on dashboards); statistical analysis; and analytics performed for prediction, simulation and optimisation. **[69]**

Analytics (Big Data)

Big data analytics refers to the increasing capacity to analyse and process massive amounts of data. It constitutes a key technology in today's digital economy, characterised by an increased use of advanced robotics, Artificial Intelligence, the Internet of things (IoT) and its industrial form, cloud computing, big data analytics, and three-dimensional printing. **[70]** Big data is characterized by: volume (consisting of terabytes or petabytes), velocity (being created real-time), variety (comprising structured and unstructured formats), exhaustivity (of scope, e.g. capturing systems or large groups), granularity (endowed with fine-grained resolution), and relationality (comprised of large data sets that may be combined together for inferences). **[71]**

---

[67] Yallop, Mark. "Machine Learning: the Big Risks and how to Manage them," Financial Times, 23.12. 2019. Accessed at: <https://www.ft.com/content/90ac19fe-2008-11ea-92da-f0c92e957a96>.
[68] For examples of data prepared through clickwork see: <https://www.clickwork-er.com/machine-learning-ai-artificial-intelligence/>.
[69] Kitchin, R., *The Data Revolution*, 2014, p. 101.
[70] UNCTAD, "Digital Economy Report 2019," pp. 8-9.
[71] For a detailed typology see, Kitchin, R., "Big Data, new epistemologies and para-digm shifts," Big Data & Society, April-June 2014: 1-12.

Augmentation

Refers to the augmentation of devices' capabilities via algorithm-led processes, or assistive forms of AI-based automation. **[72]**

Batch reporting

The practice of running multiple reports at once thanks to algorithmic led correlations from large data sets. **[73]**

Behavioural Futures Market

A type of market in which data collected from workers and consumers is used to produce statistical models designed to predict behavior through algorithm-led pattern recognition across large data sets. "Behavioural futures" are therefore behavioural predictions bought and sold on such a market. **[74]**

Clickwork

A form of work performed virtually by digital workers hired by companies through crowdsourcing platforms such as Amazon Mechanical Turk or Clickworker (one of the largest 'microtask' marketplaces today). Clickworkers constitute a form of distributed workforce. This means that companies today can outsource large pools of freelance digital workers to undertake the same specialised task (or 'microtask') at the same time, ie. in a virtual, distributed manner. Tasks undertaken by such workers may include: correcting data and validating its accuracy, manual input of textual data, implementing search engine optimization or undertaking manual searches. **[75]** As mentioned above valuable data inputs used to train AI are only obtainable through human preparation and annotation performed by clickworkers. Clickwork is a significant source of AI data harvesting enabled by the conjunction of platforms, digital applications and mobile technologies.

---

[72] Thomas, Daniel, "Automation is not the future, human augmentation is," Raconteur, 13.03.2019, accessed at: <https://www.raconteur.net/technology/artificial-intelligence/ai-human-augmentation/>.
[73] https://docs.oracle.com/cd/E26180_01/Platform.94/ATGPersProGuide/html/s1308batchreportingservice01.html.
[74] On "behavioural surplus" and "futures markets" see: Zuboff, S. The Age of Surveil-lance Capitalism, Profile Books, 2019; see also: Sawsan Mahmoud, Ahmad Lotfi, Caro-line Langensiepen, "Behavioural pattern identification and prediction in intelligent envi-ronments," Applied Soft Computing, Vol.13, Issue 4, 2013, pp. 1813-1822. Accessible at: <https://doi.org/10.1016/j.asoc.2012.12.012>.
[75] See: https://www.clickworker.com/solutions/.

Cloud computing

A model of delivering computing capabilities in which various servers, applications, data, and other resources, such as data storage, networking, analytics, etc. are provided as services over the Internet ("the cloud"), i.e. in a virtualised form. **[76]** Cloud computing may exist in the form of open source **[76]** services and resources, but cloud services are mostly privately owned, i.e. by platforms such as Microsoft Azure, Amazon Web Services (AWS), Google Cloud, IBM Cloud, and Alibaba, which in 2019 dominated that market, while AWS were in the lead, accounting for a third of the cloud computing market. **[78]**

Data (structured, semi-structured, and unstructured)

Data can be collected in a variety of forms:

• *structured data* can be easily organised and is associated with a data model, e.g. numbers or text set out in a tabular form and manageable in a relational database (name, date of birth, gender, postal code, etc) – such data can be queried, combined and analysed using algorithms

• *semi-structured data* is a form of structured data that does not follow a defined data model and does not obey a tabular structure (e.g. Extensible Markup Language (XML); HTML and other markup languages)

• *unstructured data* is data that cannot be contained in a tabular structure or relational database and is devoid of a data model (e.g. videos, images, unparsed text forms such as emails, recorded sound, voice tone, data derived from IoT sensors, etc.) **[79]**

Data (aggregated)

Aggregated data is data combined from several measurements. Data aggregation is the process where raw data is collected to produce summary forms of data for the purpose of statistical analysis. **[80]**

---

[76] See for instance: https://azure.microsoft.com/en-gb/overview/cloud-computing-dictionary/.
[77] See for instance: https://www.openstack.org/ or https://cloudstack.apache.org/.
[78] See UNCTAD, "Digital Economy Report 2019," p. 8.
[79] Marr, B., "What's the Difference Between Structured, Semi-Structured and Unstructured Data?" Forbes.com, 18.10.2019, accessed at: <https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/?sh=1f0b31552b4d>.
[80] https://www.ibm.com/support/knowledgecenter/en/SSBNJ7_1.4.2/dataView/Concepts/ctnpm_dv_use_data_aggreg.html.

Data (derived)

Data created by combining varied and large datasets. Derived by correlations.

Data (relational)

Data assembled by data brokers or collectors from a multiplicity of sources to build vast relational data infrastructures. For instance, Acxiom holds a databank derived from 500 million consumers worldwide, with about 1500 data points per person, with servers processing 50 trillion of data transactions a year. **[81]**

Data discovery

The process of obtaining actionable information by finding patterns in data from multiple sources with interactive visual analysis. Data discovery is also referred to as visualization and exploratory data analysis (often in the context of data usage by business analysts). See for instance, IBM's Watson Knowledge Catalog. **[82]**

Data-driven monitoring

In the context of IIoT, data-driven monitoring refers to the monitoring of worker activity and productivity as well as the monitoring of devices (e.g. for safety control) through the collection and analysis of data on a continuous or real-time basis. This is sometimes referred to as data-driven surveillance or dataveillance. **[83]**

Edge computing

Artificial intelligence, which feeds on big data analytics, has facilitated the emergence of "a different paradigm of computing" characterised by "data-crunching" at the network "edge" – i.e. the computing devices that intersect with internet-connected objects, e.g. smart wearables, connected cameras, or autonomous cars. Edge computing allows for the process of data close to where it is collected and to resolve limitations associated with latency (the delay caused by sending information to a distant data centre and waiting for analytics to be returned) or demands put on bandwidth by the processing of large amounts of data. **[84]** Hence the proliferation of edge data centres. **[85]**

---

[81] Singer, Natasha, "Mapping, and Sharing, the Consumer Genome," New York Times, 16 June 2012, accessed at: <https://www.nytimes.com/2012/06/17/technology-/acxiom-the-quiet-giant-of-consumer-database-marketing.html>.
[82]https://mediacenter.ibm.com/media/1_dbwxj60l.
[83]For an overview of worker surveillance, see Ajunwa I., Crawford K and Schultz J., "Limitless worker surveillance," California Law Review 105(3), 2017.
[84]Waters, R. "The Future of Computing is at the Edge." Financial Times, Jun 06, 2018, accessed at: <https://www.ft.com/content/1dba534a-5857-11e8-bdb7-f6677d2e1ce8>.
[85]https://www.pwc.com/us/en/industries/capital-projects-infrastructure/library/edge-data-centers.html.

Exhaust data

The data generated as trails or information by-products resulting from all digital or online activities (e.g. log files, cookies, information generated from transactional processes).

Fleet auditing

Corresponds to the auditing of IIoT devices for compliance with security standards through cloud-based computing capabilities (e.g. AWS IoT Device Defender).

Infrastructure as a service (IaaS)

Infrastructure as a service (IaaS) is an 'instant' computing infrastructure, provided (rented) and managed over the Internet by private corporations (e.g. AWS, Microsoft, Oracle, IBM, etc.). IaaS is one of four types of cloud services, along with Software as a Service (SaaS), Platform as a Service (PaaS) and serverless computing (used by developers to build applications and run code without managing any infrastructure themselves). **[86]** IaaS, SaaS, PaaS and serverless computing reflect how today's digital economy constitutes a digital perpetuation of rentier capitalism and is controlled by large platforms that now have become monopolies.

IaaS-cloud providers supply resources on-demand from large pools of equipment located in data centers (the hard infrastructures enabling virtual or cloud services). IaaS relies on cloud orchestration technology which may be 'open source' applications (e.g. OpenStack, Apache CloudStack or OpenNebula) that manages the creation of virtual machines (virtual computing architectures operating via virtual servers). **[87]** For wide-area networking, users can use either the Internet or carrier clouds, i.e. dedicated Virtual Private Networks (VPN).

Machine Learning (ML)

There are two broad types of ML algorithms:

• *supervised* – where a model is trained to match inputs to other known inputs, such as match handwritten post-codes to typed equivalents (i.e. parsing algorithms)
• *unsupervised* – where a model is programmed to train itself to perform pattern recognition through correlations to shape clusters. **[88]**

---

[86] See for instance: https://azure.microsoft.com/en-gb/overview/what-is-iaas/.
[87] See for instance: https://www.ibm.com/uk-en/cloud/virtual-servers/options.
[88] Han, J., Kamber, M., and Pei, J. Data Mining: Concepts and Techniques, 3rd edition. Waltham, MA: Morgan Kaufmann Publishers, 2011.

Normalization

Consists in decomposing tables to eliminate data redundancy (repetition) and undesirable characteristics, thus improving data integrity. **[89]**

Pre-analytics

All analytics require data to be readied and checked. There are broadly four type of pre-analytics:

• *data selection* - determining subsets of variables with the most utility and identifying redundant data points;

• *data pre-processing* - cleaning selected data to remove noise, errors or biases; identifying missing field or inconsistencies; structuring data for input into analysis

• *data reduction and projection* - applying transformations such as smoothing, aggregation, normalisation, etc.

• *data enrichment* - combining heterogeneous forms of data sets; aided by algorithms that match, combine, repackage and reformat data. **[90]**

Predictive analytics (or predictive modelling)

The use of advanced analytic techniques applied on historical data to formulate predictions about future behaviours. Such techniques combine classical statistical methods with AI. Prediction is a key way through which value is gained from data. **[91]**

Preparation (of data)

Refers to the transformation of raw data into a form that is more suitable for its various usages, e.g. modelling informing machine learning. How data preparation is performed depends on the type of data handled. Preparation tasks include: data cleaning or the identification and correction of mistakes or errors in the data; the selection of variables that are most relevant; changing the scale or distribution of variables. **[92]** As mentioned above, highly valuable data preparation is undertaken through crowd-sourced clickwork.

---

[89] Li, Lorraine. "Database Normalisation Explained," towardsdatascience.com, 02.07.2019, accessed at: <https://towardsdatascience.com/database-normalization-explained-53e60a494495>.
[90] Kitchin, R., 2014, p. 102.
[91] Siegel, Eric, Predictive Analytics, 2nd ed., Hoboken, NJ: Wiley, 2016.
[92] https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/.

Processing

Data processing consists in a series of steps leading to the conversion of raw data to meaningful information. In the processing cycle, data is collected and 'cleaned' or 'prepared' to be input into a sequence of manipulations operated by a computer system, or set of applications etc., that may be AI-powered, to produce an output, i.e. information, interpretation, or insights. **[93]**

Profiling

It is well established that data mining feeds into AI-powered analytics to fuel algorithm-led profiling of workers and consumers/users for the purpose of targeted marketing or algorithmic governance. Profiling based on the collection of worker or user data can be used for behavioural predictions based on pattern recognition through correlated data (such as predicting A-level results via postal codes), which can lead to discriminatory effects. **[94]** EU legislation protects EU citizens from the negative effects of profiling based on personal data to some degree. **[95]** In an organisation, profiling is undertaken in-house or via third parties, or a combination of both. This must happen in compliance with data protection regulation. Data sets gathered by companies from their own customers can be combined with third-party data collected from various sources, to build more comprehensive insights. **[96]**

Real-time data analytics (or 'streaming analytics')

A type of analytics that enables users to see, analyze and understand data as it arrives in a system. It is referred to as *streaming analytics* in 'real time,' which is enabled by the growing capabilities of cloud computing. **[97]**

---

[93] For a legal definition of processing, see: <https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-constitutes-data-processing_en>.
[94] Mann, Monique, and Matzner, Tobias. "Challenging Algorithmic Profiling: The Limits of Data Protection and Anti-Discrimination in Responding to Emergent Discrimination." Big Data & Society, (July 2019). Accessed at: <https://doi.org/10.1177/2053951719895805>.
[95] https://gdpr-info.eu/art-22-gdpr/.
[96] See for instance profiling of UK consumer data produced and sold by Axciom: <https://www.personicx.co.uk/index.html>.
[97] Wilkes, Steve, "Real-Time Data Is For Much More Than Analytics," Forbes Technology Council Post, 16 July 2019, accessed at: <https://www.forbes.com/sites/forbestechcouncil/2019/07/16/real-time-data-is-for-much-more-than-analytics/?sh=4656df06273d>.

Schema

A schema is the logical description of a collection of database objects, including tables, views, indexes, and synonyms. **[98]**

Ubiquitous computing

Refers to the ubiquitous integration of computing capabilities in human environments and resulting in a context where computing is made to appear everywhere and anywhere. **[99]** The IoT (and its industrial form) is one form of pervasive or ubiquitous computing where data is transferred over the Internet.

---

[98] See: https://www.tutorialspoint.com/dwh/dwh_schemas.htm.
[99] Mark Weiser, "The Computer for the Twenty-First Century," Scientific American, Vol. 265, No. 3, Special Issue: Communications, Computers and Networks: How to Work, Play and Thrive in Cyberspace (September 1991), pp. 94-105, accessible at: <https://www.jstor.org/stable/24938718>.