Data protection by design and data analytics: can we have both?

Sophie Stalla-Bourdillon Senior Privacy Counsel with Immuta and Professor in Information Technology Law and Data Governance at the University of Southampton, discusses how to practically meet the requirements of DPbD, and how to pursue a DPbD approach alongside conducting big data analytics

Data Protection by Design ('DPbD') is a core data protection requirement introduced by the General Data Protection Regulation ('GDPR') in its Article 25. The concept, however, is not completely novel. DPbD builds upon the privacy-by-design approach developed by Ann Cavoukian in her seven foundational principles back in the 90's. Plus, the Data Protection Directive of 1995 (95/46/EC) had already called in its Recital 46 for the taking of organisational and technical measures at the time 'of the design of the processing system'.

That said, the GDPR is the first instrument to make DPbD a legal requirement of its own, with an express sanction. Under Article 83(4) of the GDPR, violations of Article 25 could lead to the issuance of fines up 'to 2 % of the total worldwide annual turnover of the preceding financial year, whichever is higher'.

DPbD is thus a serious matter. This is all the more true considering that DPbD appears to be the backbone of the GDPR — Article 25 refers to both the Data Protection Principles and data subject rights. Pursuing a DPbD approach means 'implement [ing] appropriate technical and organisational measures, such as pseudonymisation' with a view 'to implement Data Protection Principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects'. In other words, DPbD should lead to compliance with the Data Protection Principles (Article 5) and data subject rights (Articles 12 to 22), bearing in mind that complying with Article 5 implicitly means meeting other key provisions such as Articles 30 (records of processing activities) and 32 (technical and security measures).

Research-born methodologies

So how can controllers demonstrate their compliance with Article 25?

Whilst the literature on Privacy by Design is getting richer every year (e.g., the preliminary opinion of the European Data Protection Supervisor or EDPS of 2018), legal experts and compliance teams have been struggling to offer clear guidance on the matter. At the same time, industry has not always warmly welcomed research work in the field and, in fact, often sees it as overly complex and academic.

By way of example, the EDPS refers to the LINDDUN methodology (https://linddun.org/) a highly-technical privacy threat modelling methodology that relies upon two key steps: data flow mapping through the creation of a data flow diagram incorporating four types of building blocks (entities, data stores, data flows, and processes); and privacy threat modelling through the identification of privacy threats for each building block included in the data flow diagram.

However, for those interested in building a compliance strategy, a methodology of this type falls short of offering a real answer, as its list of threats (linkability, identifiability, non-repudiation, detectability, information disclosure, content unawareness, and policy and consent non-compliance) does not exactly match the list of principles found in Article 5. In addition, it is primarily addressed to developers.

The methodology of German Professor Jaap-Henk Hoepman is interesting in that it identifies two types of privacy design strategies: data-oriented strategies and process-oriented strategies. It has even inspired the European Union Agency for Network and Information Security's guidance. (See ENISA reports on Privacy and Data Protection by Design of January 2015.)

That being said, it is arguable whether the strategies actually capture all Article 5 principles. Moreover, once again privacy design strategies are meant to guide the work of developers, not to demonstrate compliance.

There thus appears to be a slight gap between research and practice.

How to practically meet the requirement of DPbD, then? More crucially, how to pursue a DPbD approach when conducting data analytics activities in an era of Big Data, where the common assumption is that vast amounts of data (input data) need to

be processed in order to derive valuable insights?

How to do DPbD on the ground

A useful starting point is the German Standard Data Protection Model (the 'SDM', copy at www.pdpjournals.com/docs/887982). The SDM is relatively clear and accessible in its attempt to match technical and organisational measures (i.e., controls) to Data Protection Principles.

The SDM correctly conceives Data Protection Principles as goals, because the GDPR does not offer an exhaustive list of controls for each principle, and ultimately, the choice of the applicable controls should depend upon a trade-off between privacy and utility set in context.

Let's take an example. If a data analyst thinks it is impossible to derive insights from aggregates only, she could implement methods such as global differential privacy (based on the randomised injection of noise within data sets in order to generate aggregates) as controls for the principle of data minimisation and/or confidentiality.

Importantly, deriving insights or building models from aggregates only can be done in many instances, even if the objective is to create user profiles. It might not always be the easiest solution, but it is certainly worth the effort for accountability purposes. What is more, even if a data scientist needs access to low-level data to derive insights, methods of local differential privacy often appear to be the perfect candidates for meeting the goals of data minimisation and/or confidentiality. Indeed, methods of local differential privacy produce individual-noised records, when methods of global differential privacy produce noised aggregates.

Once one has identified controls for each data protection principle, what should one do? The exercise is incomplete, because the selection of controls should be risk-based. In practice, this means that for each control, controllers should identify the events likely to jeopardize the controls at stake, their likelihood of materialising, and the potential resulting harm for data subjects. This risk-based approach should enable the controller to make sure the strength of the control is actually tailored to the risks posed to the rights and freedoms of data subjects. Here, technical experts and compliance personnel should sit together and create a risk matrix.

> ***"Following a DPbD approach when conducting data analytics projects is not mission impossible. However, it is resource-intensive, and requires a combination of different expertise. The good news is that once controls are scaled, workflows accelerate."***

Guidance on how to perform this task can be found in methodologies on risk assessment produced by Supervisory Authorities, in particular the useful guidance produced by the French CNIL entitled 'Methodology for Privacy Risk Assessment' (copy at www.pdpjournals.com/docs/887983).

For a list of controls, it is worth reading both the SDM and the knowledge base produced by the CNIL for Privacy Impact Assessments. Reports produced by ENISA may also be helpful and relevant.

How to do DPbD for data analytics

If the initial premise is that a vast amount of data are needed to build a data analytics model, how can one really follow a DPbD approach?

This area of the GDPR has been intensely lobbied. Many stakeholders have an interest in arguing that the framework is doomed to fail from the start and will never work, in particular when applied to data analytics, machine learning, and artificial intelligence ('AI').

In reality, the GDPR is both risk and process-based, meaning that it offers a framework to make sure controllers ask themselves the right questions as early as possible. The GDPR is therefore only one stepping stone on the path to AI for good. Importantly, this process-based approach is not necessarily incompatible with data analytics or the building of machine learning model. On the contrary in fact.

At a high level, data analytics model lifecycles can be divided into three stages:

- documentation of model objectives, assumptions, limitations and relevant data;
- model development and testing; and
- model deployment and usage.

What is surprising to note is that all Data Protection Principles make sense for all three stages.

Let's take an example. At stage 1 when business owners set the objectives and data scientists identify model assumptions, limitations and relevant data, all of the Data Protection Principles can support the documentation activities. Purpose limitation is relevant and puts useful constraints for determining the objectives of the model (e.g., analyse usage pattern/zip codes to determine premium). Fairness is also relevant for assessing model assumptions, deriving model limitations, and evaluating the quality of the data (i.e., whether the data are biased or not). Accuracy is also important to assess the quality of the data.

Data minimisation helps determine whether the data are actually relevant and necessary for the purpose pursued. Integrity and confidentiality

(Continued on page 10)

(Continued from page 9)

ensure that data are unaltered and will remain unaltered over the time, which is another key requirement for data quality. Transparency helps to ensure that model pipelines are transparent and information is logged so that it is possible to monitor and audit behaviour and ultimately demonstrate compliance (i.e., accountability). Finally, storage limitation helps ensure data analytics projects are conducted within pre-defined timeframes and are actually closed when models are built.

More specifically regarding data minimisation, different techniques can be used to tailor the amount of data to the pursued purpose. They are usually captured under the heading masking: e.g., creating samples of data sets, hiding rows or columns within datasets, rounding values, hashing, using differential privacy methods to produce aggregates, etc.

Scaling controls for a variety of data analytics projects

In the end, it becomes clear that following a DPbD approach should facilitate the deployment of GDPR good practice. It is true, however, that a DPbD approach is resource-intensive, in particular if one organisation intends to conduct not one but many data analytics projects. This is where it is worth considering legal tech solutions or trusted architectures. This is the best way to scale controls for an increasing number of data analytics projects.

Trusted architectures should comprise three types of layers to properly enforce DPbD:

Access and control layer — A layer for standardised data access should be the first building block. It should be agnostic to data storage and third-party tools used to consume data, and should aim at reducing data movements between stakeholders, ensuring security and confidentiality of communications while making it possible to set common rules for data access across projects.

In several instances, a trusted architecture is a better option than a trusted third party. This is because trusted architectures reduce movements of data within organisations while making access to data faster and maintain data usage within one controlled environment. From a risk-based standpoint, resorting to a third party to conduct the analysis is not always the best option, as it increases data movements and makes it more difficult for the controller to argue that the processing is taking place within a controlled environment.

Project layer — A layer of project-specific controls should make it possible to tailor the trade-off between privacy and utility in context once a comprehensive risk assessment has been conducted.

Auditing layer — A third layer comprising monitoring and auditing tools should also be added to facilitate accountability.

Conclusion

Following a DPbD approach when conducting data analytics projects is not mission impossible. However, it is resource-intensive, and requires a combination of different expertise. The good news is that once controls are scaled, workflows accelerate.

Why invest in DPbD? Avoiding fines is certainly part of the answer, but ultimately what is really at stake is the building of good practice for developing Artificial Intelligence.

And in the end, it is likely that we'll need more than DPbD, for it is only the first stepping stone in the process.

Sophie Stalla-Bourdillon
Immuta
sstalla-bourdillon@immuta.com