James Schorr

11/7/2025

# An Evaluation of Data Science Use Cases and their Adherence to GDPR Guidelines

**Synopsis:**

When the data science community is presented with a use case involving the personal data of European Union (EU) citizens, it must carefully navigate the compliance laws outlined in the General Data Protection Regulation (GDPR). Since its inception in 2016, the GDPR and its supporting guidelines have provided a framework for evaluating the legitimate use of personal data; however, they do not offer extensive and comprehensive practical use cases or examples to guide the determination of legitimate uses across disciplines. The goal of this paper is to illustrate, through scenarios, how to evaluate and determine if the use of personal data for data science projects can be conducted in compliance with GDPR. This paper will include: 1) a high-level overview of GDPR; 2) a summary of the Legitimate Interest Assessment (LIA) test, which is the recommended method for determining legitimate uses of personal data; 3) evaluation scenarios applying the LIA framework to marketing data science use cases; 4) a high-level comparison of guidelines for implementing these scenarios in the United States; and 5) recommendations to improve the efficiency of compliant implementation and conducting LIAs for the specific use cases.

**GDPR Principles:**

GDPR defines three entities in the compliant use of personal data, including the data controller (responsible for compliance with the law including establishing legitimate data uses while fulfilling the rights of the data subject), the data processor (responsible for the compliant processing of data per its identified limited use), and the data subject (the person whose data is being processed). A foundational principle of GDPR is the requirement that data be processed

for its expressed purpose only if the data subject has freely given consent. Furthermore, the rights of the data subject are outlined in Chapter III of the original GDPR regulation[1] and they include:

| Right | Description |
|---|---|
| **Right to be informed** | The right to be informed about the collection and use of your personal data. |
| **Right of access** | The right to access your personal data and to receive a copy of it. |
| **Right to rectification** | The right to have inaccurate or incomplete personal data corrected. |
| **Right to erasure** | The right to have your personal data deleted (also known as "the right to be forgotten"). |
| **Right to restrict processing** | The right to request that the processing of your personal data be limited. |
| **Right to data portability** | The right to receive a copy of your personal data in a machine-readable format and to have it transferred to another service. |
| **Right to object** | The right to object to the processing of your personal data, particularly for purposes such as direct marketing. |
| **Rights related to automated decision-making and profiling** | The right to avoid decisions based solely on automated processing that have legal or similarly significant effects. |

The right to privacy and control of the use of their personal data by data subjects (i.e., EU citizens) is a central tenet of GDPR. The regulation articulates seven principles, each one designed to ensure the protection of personal data and the privacy rights of individuals when sensitive data is being utilized by controllers and processors. The seven principles to be followed by controllers and their processors are summarized as follows[2]:

| Principle | Description |
|---|---|
| **Lawfulness, Fairness, and Transparency** | Having a legal basis, being transparent, and acting in the person’s best interests |

| | |
|---|---|
| **Purpose Limitation** | Only process personal data for the purpose it was intended for |
| **Data Minimization** | Only gather and keep the exact amount of data that is needed |
| **Accuracy** | Take reasonable measures to have the most accurate data possible |
| **Storage Limitations** | Don't store personal data that you don't need any more |
| **Integrity and Confidentiality** | Only people processing the data should have access to it |
| **Accountability** | The data processor is responsible for complying with GDPR |

How do controllers and processors determine if their intended use of the personal data respects the rights of data subjects and complies with the principles of GDPR outlined above? The next section introduces the framework for how this evaluation can be performed.

**The Legitimate Interest Assessment (LIA) Test:**

The Legitimate Interest Assessment (LIA) is a structured test outlined by GDPR to determine whether an organization can lawfully process personal data based on its legitimate interests. The concept and recommended approach are documented in the "Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR", issued by the European Data Protection Board and adopted in October of 2024. There are three recommended steps to conduct a LIA[3]. They include:

- Step 1: Pursuit of a legitimate interest by the controller or by a third party
    - This step defines the legitimate interest as outlined by the controller and processor, such as a commercial objective, and specifies what is being achieved and who benefits.
- Step 2: Analysis of the necessity of the processing to pursue the legitimate interests

    - This step determines if the envisioned processing of the data is necessary to achieve the defined legitimate interest (e.g., “is too much data being used or can other non-personal data sources be used instead?”).

- Step 3: Conduct a “Balancing Test”

    - This step performs a cost-benefit-like analysis to determine if the organization's legitimate interest outweighs the individual's rights, freedoms, and interests of data subjects.

In a December 2024 published opinion[4], the European Data Protection Board recognized that AI models and algorithms can be considered a legitimate interest but stated that their development and application should be evaluated for legitimacy on a case-by-case basis. This presents an opportunity for data scientists to leverage the LIA testing framework to determine if the use of data science and AI methodologies can be done in compliance with GDPR. The next section applies the LIA framework to three data science marketing use cases to illustrate the framework in action.

**Three Scenarios Applying the Legitimate Interest Assessment (LIA) Test Framework:**

***Scenario #1:*** *Building a Recommender System to Promote Specific Products to Known Customers When They Shop on Our Website*

Scenario #1, LIA Evaluation Details:

| LIA Step | Postulation | Analysis |
|---|---|---|
| Legitimate Pursuit Check | To enhance customer shopping experience by showing relevant product suggestions while known customers are shopping on-site. The benefit is twofold: we increase sales and repeat site visitation by our customers over | The commercial interest, scope, and potential outcomes for the business and customers is clear, reasonable, and beneficial for both. |

| | | |
|---|---|---|
| | time and customer satisfaction is improved by creating relevant and expedited shopping experiences. | |
| Data Necessity Check | Recommendation systems require access to identifiable customer behavior (e.g., past purchases, clicks, cart history) to generate recommendations. | Relevant personalization requires access to identifiable customer behavior (e.g., past transactions) to generate meaningful recommendations. Alternative data sources (such as most popular purchase categories in general) will not produce relevant person-level recommendations. |
| Balancing Test | If explicitly opted-in to receive personalized recommendations, customers have reasonable expectation of receiving offers during shopping. The use of data for the recommendations is prior purchase history only and is for the express purpose of producing the recommendations. | The legitimate interest, specifically for cases where customers have clearly opted in to the personalized offers program, is not overridden by the customers rights. In cases where customers have NOT opted-in, the processor cannot use a customer's transaction histories to serve them personalized offers. |

Scenario #1 Result: This is a justifiable application of data science and AI using personal information if opt-in and consumer consent is adhered to.

***Scenario #2:*** *Building a machine learning model using boosting techniques to target non-customers with a direct marketing campaign promoting a new product. The model leverages 3rd party data to find consumers most likely to have high interest in the product.*

Scenario #2, LIA Evaluation Details:

| LIA Step | Postulation | Analysis |
|---|---|---|
| Legitimate Pursuit Check | To efficiently expand our customer base by promoting a new product most effectively with a limited budget by only targeting individuals most likely to purchase the product. | The commercial interest and benefit are clear, but non-customers have low expectation that their personal data is being used by an organization with which they haven't explicitly opted in. |
| Data Necessity Check | As specifically defined, the data processing is necessary (using | In this case, there are alternatives such as geo-based modeling and |

| | | |
|---|---|---|
| | personal information to build the most precise targeting model). | marketing that will achieve profitable results. For example, an aggregate postal code dataset can be used to isolate areas (not specific people) where a high product interest exists among residents of the postal code. |
| Balancing Test | The business benefit is clear, but the consideration of consumer privacy and rights isn't given the 3rd party data required may or may not have fully consented consumers. | The lack of clarity on consent with 3rd party sources and low consumer expectation of companies being able to access their personal data raises compliance concerns. |

Scenario #2 Result: This is unlikely to be a justifiable application of data science and AI using personal information, given consumers likely don't expect profiling and outreach from the company and the risk that comes with using 3rd-party data sources that potentially collected the data non-compliantly.

***Scenario #3:*** *Web scraping and text analytics to identify trending topics that will inform marketing campaign creative, branding, and messaging strategies to key audience segments.*

Scenario #3, LIA Evaluation Details:

| LIA Step | Postulation | Analysis |
|---|---|---|
| Legitimate Pursuit Check | Improve the relevance of our marketing campaigns by creating messaging that mirrors current consumer interests and cultural trends. With more relevant and topical messaging, we will increase our brand awareness and long-term sales. | The commercial interest, scope, and potential outcomes for the business and customers are clear, reasonable, and beneficial for both the business and consumer. |
| Data Necessity Check | The capability to be self-sufficient in identifying trends requires access, at large scale, to publicly available datasets. | Other options do exist such as purchased market research reports from 3rd parties but those are expensive, and not as timely with the insights. |
| Balancing Test | The business benefit is clear but the considerations around fulfilling consumer privacy | There are personal information data points available in public web properties such as usernames, handles, and e-mail |

| | | |
|---|---|---|
| | obligations needs to be more explicit. | addresses raising compliance concerns on collection. |

Scenario #3 Result: As long as the business ensures that the scraping exercise does not collect or store any form of personal data, this is a justifiable application of data science and AI using the collected data.

**A High-Level Assessment of the Scenarios' Legitimacy in the United States:**

In the United States, all three scenarios are feasible, and more onus falls on the data scientist to ensure personal data is used ethically and only when needed. As examples,

- Scenario #1's feasibility and execution of a recommender system is similarly performed in the U.S., with the biggest difference being that the U.S. leverages an opt out approach as a standard, as opposed to an opt in approach as required by GDPR. This will naturally create more scale and reach to personalization programs in the US. In addition, in the U.S., the processor does not have to explicitly outline the data used or the underlying logic to generate the recommendations (there are state by state nuances). This raises the importance of the data scientist's role to identify, and use approved and non-sensitive or protected data in the solutions and to build transparent decisioning solutions.

- While scenario #2's feasibility and execution are unlikely to comply with GDPR, leveraging targeted 3rd party lists for targeted marketing is common practice in the United States. There are several reputable 3rd party data companies that actively maintain opt-out consumer databases and hygiene of the consumer lists. This, again, raises the importance of the data scientist's role in building a machine learning solution (such as a prospective

customer response model using boosting techniques) that does not use highly sensitive, protected class data and is interpretable and transparent in its application.

- Scenario #3's feasibility and execution of web scraping algorithms to identify topical trends in the marketplace can be similarly performed in the U.S., but there is less formal oversight on the capture and use of personal data such as social media usernames or handles when that data is being collected via scraping. A primary example of this was a research study published in 2016 that scraped and then revealed personal information available on the OKCupid dating site[5]. The study resulted in heated debate on the ethics of scraping and releasing such data (consensus was that it was unethical). But the authors were not formally punished, because there is no law in place governing public data scraping activity. This presents a third case and reason that data scientists are essential agents for the ethical deployment of analytical solutions to minimize harm of data collection and use, given the less stringent data capture, governance, and regulation in the United States.

**Recommendations:**

GDPR can benefit from two potential enhancements, one is a formal policy or rules around the legitimate 3rd party data usage, and the other is more explicit guidance or templates around the use of AI and data science that complies with GDPR. Specifically,

- In the spirit of respecting consumer rights and privacy, there is an opportunity to set up a service where consumers are able opt themselves in, provide their personal data to a 3rd party data collector with the express and specific purpose of allowing marketers to

communicate to them about their products. Consumers can pick which types of product families they are interested in hearing about and can also stipulate how frequently they can be communicated to (i.e., consumer A opts-in to receiving marketing messages from electronics providers but doesn't want more than 1 solicitation per week across providers). This will require processing by a select few GDPR-approved processors that need to coordinate amongst themselves to adhere to compliance guidelines and ensure communication frequency caps are respected. GDPR could approve the use of the approved 3rd party data providers as they are opt-in databases by definition and the primary requirement is adherence to the customer category and touch frequency preferences. This will benefit marketers by reducing uncertainty on the use of 3rd party data, all while respecting consumer preferences in alignment with the core principles of GDPR.

- GDPR can also publish a library of explicit, approved use cases and examples (like the exercise in this paper) with the specific goal of helping marketers determine and decipher legitimate uses of AI and data science that comply with regulations. This benefits consumers and businesses alike, as it will reduce the likelihood of breaches which generate financial penalties for businesses and unwanted experiences for consumers.

## Citations


1. European Parliament and Council of the European Union. (2016). General Data Protection Regulation. Official Journal of the European Union, L119, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

2. Sprinto. (n.d.). GDPR principles: The 7 principles of GDPR compliance. Sprinto. https://sprinto.com/blog/gdpr-principles/

3. European Data Protection Board. (2024). Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR (Version 1.0). https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf

4. Wilson Sonsini Goodrich & Rosati. (2024, May 28). EU privacy regulators confirm that legitimate interest is a valid legal basis for AI model training and deployment. https://www.wsgr.com/en/insights/eu-privacy-regulators-confirm-that-legitimate-interest-is-a-valid-legal-basis-for-ai-model-training-and-deployment.html

5. Dewey, C. (2016, May 12). A researcher posted 70,000 OKCupid users' data. The backlash is fierce. Vox. https://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release