IMPLEMENTING AI-AIDED CONTENT DISTRIBUTION STRATEGIES IN THE GDPR ERA

Alain Nochimowski, Alice Wittenberg, David Zucker

Viaccess-Orca, Israel

ABSTRACT

While the mathematical foundations of artificial intelligence (AI) have been used for quite some time to make predictions and analyse data, the recent explosion in volume, variety, and velocity of viewership data collected from multiple connected devices is turning TV into a new frontier for predictive algorithms. With General Data Privacy Regulation (GDPR) coming into force across the EU in May 2018, extracting actionable insights from TV service providers' viewership data will become both increasingly appealing and challenging. This is expected to make a significant impact on the economics of targeted advertising and marketing, which have traditionally relied heavily on third-party data. Our research indicates that the utilization of state-of-the-art AI techniques could help alleviate this reliance.

INTRODUCTION

European TV service providers are progressively waking up to a fundamental dichotomy that will dramatically impact their way of doing business. Never before has the business imperative for collecting customer data and implementing AI-aided services been so strong. On the one hand, with the rapid migration from legacy broadcast protocols with no return path to ever more personal and immersive media formats (now accessed via natural interfaces such as voice and gesture control), viewership data is growing exponentially in volume, variety, and velocity (Gartner's "3V's") (1). On the other hand, the dramatic shift in viewers' attention that occurred in recent years in the favour of digital platforms whose businesses heavily rely on data has caused TV service providers to accelerate the "datafication" of their business operations as they look to compete on equal terms.

Yet never before have the stakes of navigating the intricacies of private data protection regulations been so high. With GDPR coming into force in May 2018 in each EU member state, the mismanagement of customers' personal data can result in fines of up to €10 million or 2 percent of global turnover, whichever is higher (serious offenses are double those amounts, up to €20 million or 4 percent) (2). The recent Facebook and Cambridge Analytica controversy, along with Facebook's subsequently stated intent to apply the same privacy protections throughout its footprint, signals that GDPR will likely become the gold standard of privacy regulations with real impacts beyond the strict borders of the EU.

The story of the GDPR tidal wave impacting traditional marketing and advertising practices is just starting to unfold before our eyes. One does not have to be an expert on the EU law system to foresee that reliance on third-party data will be made riskier and the incentives for extracting new insights from first-party data will become stronger due to stringent user consent management obligations under GDPR. For EU TV service providers, this could induce few possible strategies with regards to the collection and monetisation of TV viewership data (assuming full GDPR-compliant opt-in) and renewed interest for state-of-the-art machine learning techniques. This paper will explore some of the possible strategies and shed light on concrete insights that could be derived by TV service providers from their viewership data through advanced predictive algorithms.

EUROPEAN TV SERVICE PROVIDERS' VIEWERSHIP DATA MONETISATION STRATEGIES

The new life of TV viewership data in the age of GDPR

The purpose of this paper is not to provide a comprehensive analysis of the 99 articles setting out the rights of individuals and obligations placed on organizations covered by the regulation, which is the result of four years of discussion and negotiation (3). Neither do we aim to exhaustively present the foreseeable impacts of GDPR on the TV industry, especially since they are subject to expert controversies: beyond advertising, the whole domain of profiling and targeting (including the cases of content recommendations and service personalization will be impacted. This includes service and algorithm designs (e.g., privacy and security by design, increased attention to algorithm fairness and transparency) and service operations (e.g., collection and management of viewers' consents). However, highlighting some of the expected impacts of GDPR in relation to first-party and third-party data management practices (4) may be useful to get a feeling of the tectonic changes that are at work in the marketing and advertising industries and the potential value of viewership data.

Besides stringent obligations related to the collection of explicit and withdrawable consent (Article 7) and the required transparency on the purposes of the intended processing (Article 13), the regulation reaffirms a long list of data subjects' rights. In particular, it defines right of access (Article 15), to rectification (Article 16), to erasure (Article 17) and to restriction of processing (Article 18). As it is difficult to enforce these rights throughout a chain of intermediaries, these stipulations will undoubtedly turn traditional practices around third-party data into a riskier business. Mechanically, due to their expected scarcity (5), GDPR-compliant third-party datasets will become more expensive, thereby transforming the economics of targeted advertising and first-party data validation as well as enrichment (two key use cases enabled by third-party data). For TV service providers, these changing economics, combined with a wider definition of personal data (which unambiguously encompasses IP addresses, the identifier commonly used to create and sell granular audiences out of a TV operator's viewership in the U.S. and UK markets), cast serious doubt as to the applicability of the same existing addressable TV solutions in post-GDPR continental Europe.

In light of these expected changes, there is little doubt that the “data centre of gravity” will rapidly move toward those in the digital value chain that retain opted-in first-party data. In addition, extracting additional insights from underexploited sources such as viewership data may rapidly generate renewed interest. Because they sit on a treasure trove of viewership data that could enrich their first-party data, TV service providers may soon be in position of leverage.

Besides the technical feasibility of targeted advertising and marketing, the very nature of personal targeting is also brought into question. While under the pre-GDPR paradigm, digital marketers have frequently hailed behavioural targeting (measured by the granularity of its attributes) as being superior to contextual targeting. For example, we can expect contextual targeting techniques will find a new appeal among advertisers in order to reach untrackable viewers (i.e., non-consenting to personal data collection) or in the face of new regulatory constraints imposed on profiling (GDPR Article 22). Here again by its very nature, the processing of viewership data can constitute a formidable asset for TV service providers.

Enabling inbound viewership data monetisation

With the renewed interest around insights extracted from viewership data now posited, let us look at possible TV service provider monetisation strategies. Collecting, processing, and analysing viewership data can add real value in terms of increased process efficiencies within TV service providers’ domains (i.e., “inbound” monetisation). For example, service providers gain the ability to predict the appetency of viewer segments for certain content and can implement targeted content promotions or upselling of service packages.

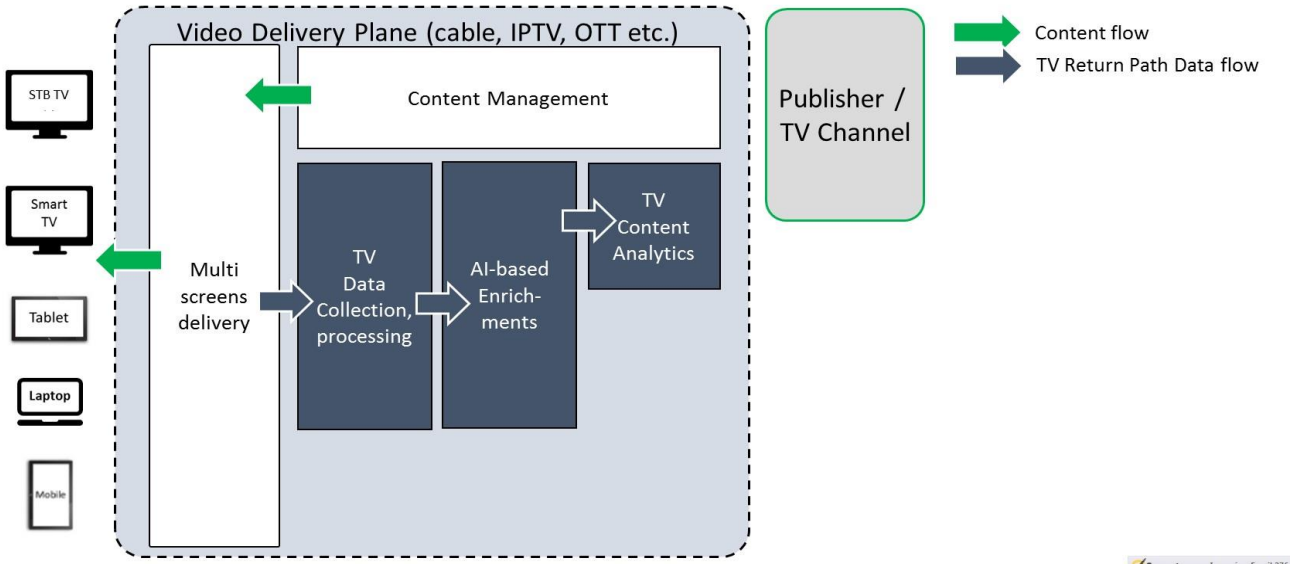

Figure 1 – TV service provider inbound monetisation high-level architecture

This inbound monetization workflow is schematically described in Figure 1, with data collected directly from multiple IP-enabled devices and/or TV/video delivery enablers including service delivery platform (SDP) or conditional access systems (CAS). This covers data cleansing, normalization and ultimately data enrichment through AI techniques as well as the provision of instant and predictive analytics use cases, such as audience measurement and churn prediction.

Outbound monetisation (advertising)

For TV service providers in continental Europe, the most promising TV data monetisation opportunity may actually relate to programmatic and addressable TV. Despite tightly regulated and pretty diverse environments in Europe, positive evolutions are expected in the coming years (6), and we should see operators progressively moving from experimentations limited in scale to wider addressable TV deployments. As explained above, we can expect these deployments to differ in many ways from existing ones in North America or the UK, in part due to the impact of the GDPR regulation and the constraints on data collection and profiling.

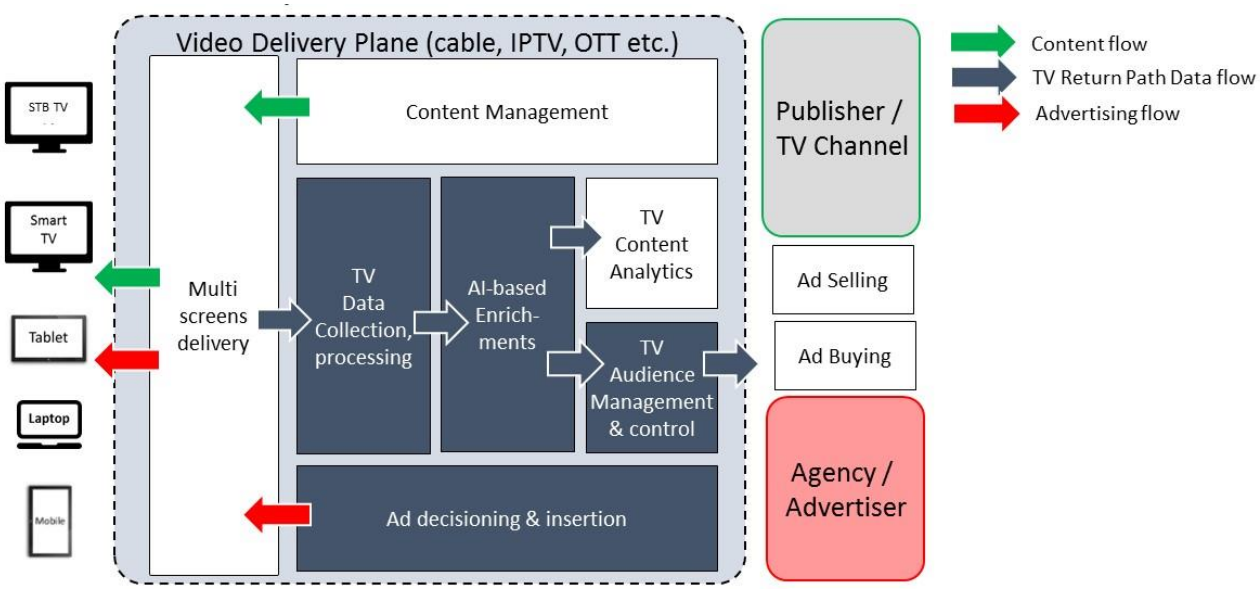

Figure 2 – TV service provider outbound monetization (advertising) high-level architecture

As described in Figure 2, enabling an addressable TV strategy requires additional infrastructure components in order to integrate with the programmatic advertising chain of the channels (which own most of the inventories), or perform dynamic ad insertion (DAI). Creating, managing, and targeting audience segments (outbound audience monetisation) in full compliance with GDPR will constitute one of the main challenges faced by TV

service providers. Companies like Viaccess-Orca (VO) provide solutions to address this challenge, the extent of which goes beyond the framework of this paper.

Enriching such audience segments with AI-based techniques implemented on fully optedin GDPR compliant viewership data (see illustration in Figure 3) may not only represent an elegant way to reduce reliance on third-party data in a TV audience management platform (e.g. think of predicting household socio-demographics to enrich audience segments or simply validate first-party data), but it can also reveal unique insights such as “life moments” through changes in household TV consumption such as a child joining the household.

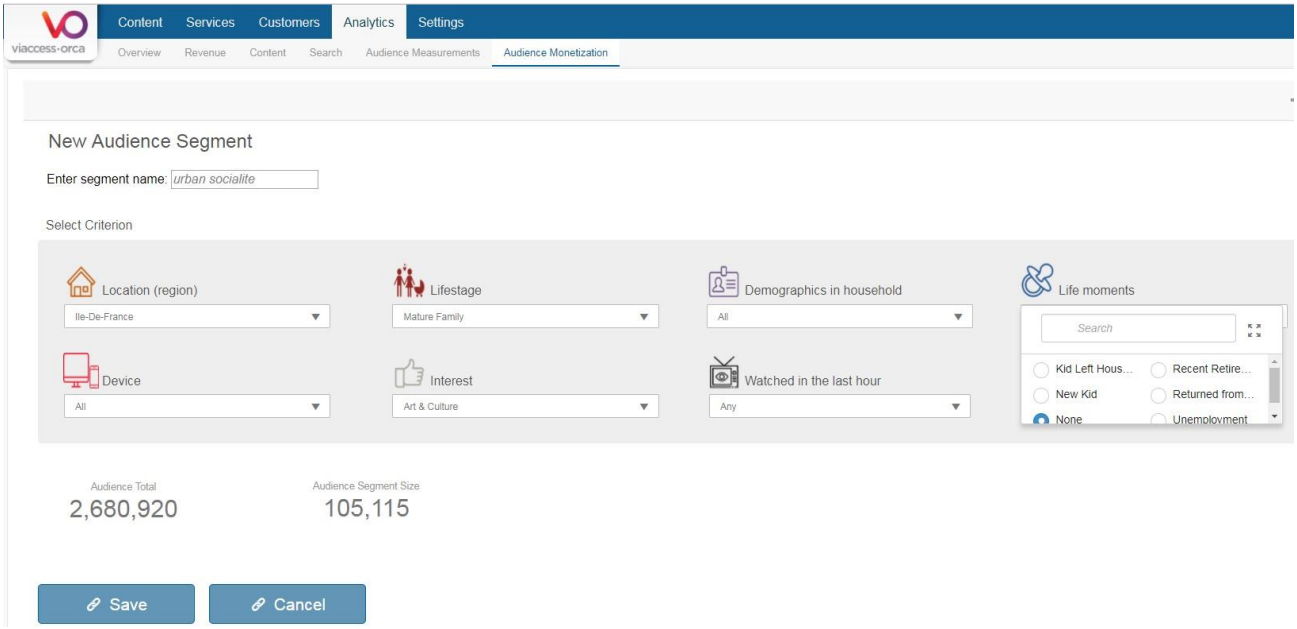

Figure 3 – AI-enriched TV audience management by Viaccess-Orca

EXAMPLE OF AI ENRICHMENTS

The following sections present some of VO’s research results around socio-demographics predictions obtained using advanced machine learning techniques in the framework of a collaborative project gathering business and academic data science experts under the Israel Innovation Authority. The dataset used for this research covers one year (2015) of viewership data for ~300,000 cable-equipped households across the U.S. It consists of a fully labelled dataset (i.e., each household associated with socio-demographic information (24 age segments i.e., third-party data from brokers), and coverage of all TV consumption events (including linear channel zaps and on-demand events such as DVR) from set top boxes (STB) at the resolution of a second with the corresponding EPG/catalogue information.

Description of our dataset

Although in the most advanced machine learning techniques (i.e., deep learning), feature selection is performed by the system itself, a basic exploration of the data may provide some interesting information as to the representativeness or potential biases of our dataset. It appears that a number of features may help in the viewers' profiling process besides the nature of content being viewed or household socio-demographic information, starting with the zapping behaviour on linear content. Figure 4 provides a view of the distribution of linear TV zapping events over one month (interval between two zaps), which shows that it takes a lot of zaps for viewers to reach a program they are willing to view.

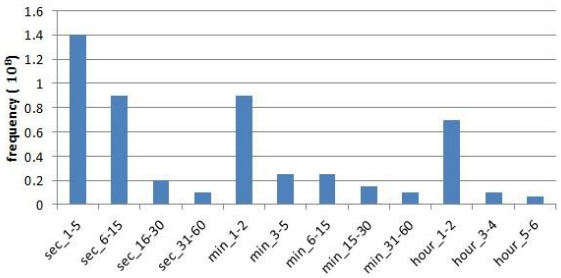

Figure 4 – Distribution of linear zapping events over one month of viewership

The number of devices (STB) per household is also a meaningful piece of information when it comes to viewers' profiling. Figure 5 shows the distribution of STB devices per household across our dataset.

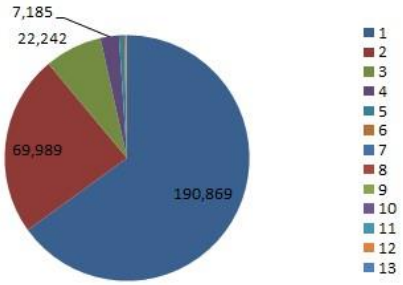

Figure 5 – Number of STB device distribution per household

Another important feature consists of the geographic distribution of our households, across U.S. Designated Market Areas (DMAs). These are U.S. regions where the population can receive the same broadcast TV offering. As described in Figure 6, our data covers all U.S. geographies with certain biases.

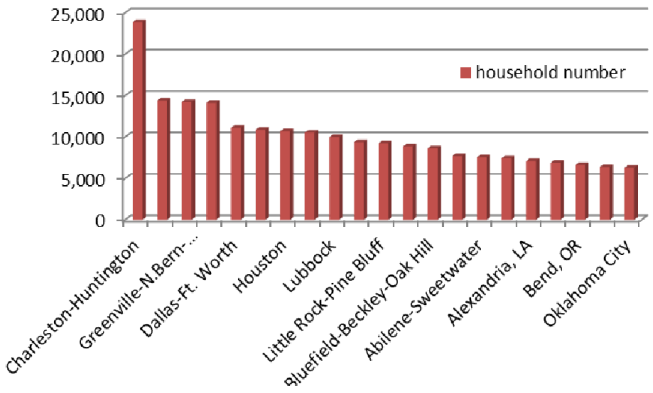

Figure 6 – DMA distribution demographic

The combination of all those features (and a few others) together with the nature and sequence of viewed content and socio-demographics attributes allows us to associate a unique spectral signature to clusters of households.

Each household is therefore described as a vector combining all the above features, the entry of which corresponds to the specific socio-demographic attributes we want to train the model on (e.g., prediction of age and genders in a household). The particular case of a household containing a single individual could be modelled in the exact same way.

In a typical real-life deployment, a minimal labelled dataset from TV service provider's first party data would suffice to perform such training (the general assumption being that we'd be working with fully opted-in GDPR compliant data).

Connecting content items to demographics

Our dataset therefore connects households to demographics of the household members and households to viewership data. However, the data doesn't reveal which household member watched which content item, and therefore doesn't offer a clear connection between content items and viewers' demographics. We will start by demonstrating how to associate demographic information with content items from the data.

The first step is to represent each household in our labelled dataset as a combination of demographics. In the data, a household is accompanied by gender and age range for each household member. We represent the household as a vector, where each entry corresponds to a specific gender and age range, and the value at said entry is the number of household members from the corresponding demographics, as illustrated in Figure 7.

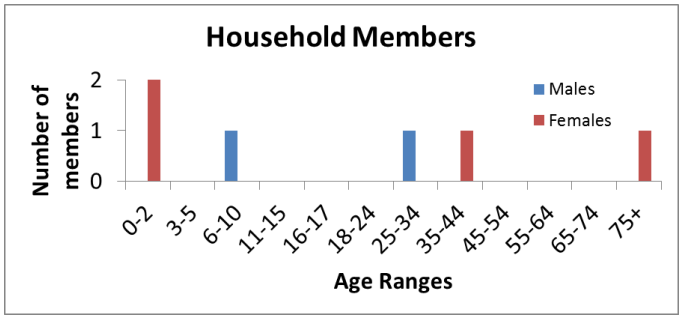

Figure 7 – Vector representation of household members demographics

Then, for each linear TV program we take the household that viewed it and calculate the average of its representing vector to obtain a vector representation of the content item and thereby a connection between the content item and demographic information, as shown in Figure 8, where the ordinate indicates the percentage of households that watched the program among those households that contain the corresponding demographic class. Therefore, we represented both households and content items as vectors with 24 entries (age segments), where each entry connects to a specific gender and age range.

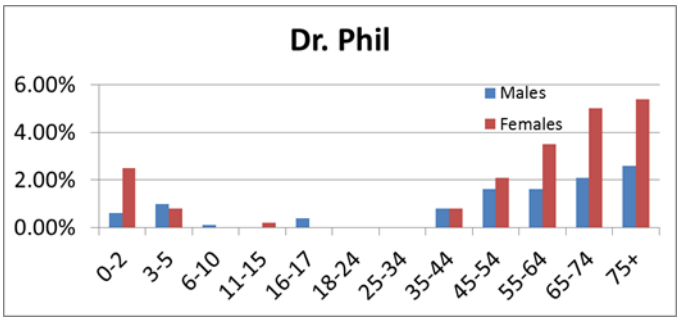

Figure 8 – Content representation

In order to illustrate the validity of our representation of content items, we take the vector representations for a number of content items and use dimensionality reduction through Principal Component Analysis (PCA) to extract the two leading principal components of each content item, therefore embedding the content items in a two-dimensional space, as depicted in Figure 9.

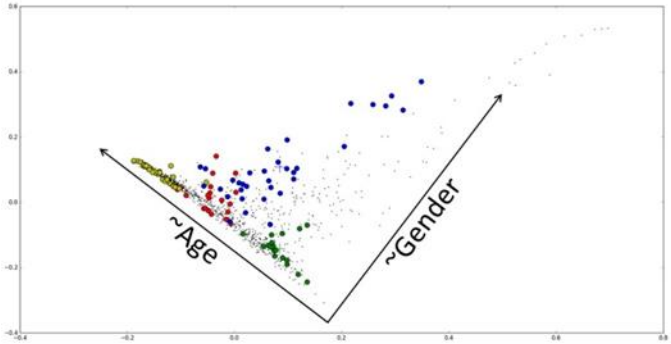

Figure 9 – Content representation

In the figure, each point corresponds to a content item, and four groups of content items can be identified in different colours: in green is a group of kid programs (such as Calliou"), in blue is a number of sport programs (such as "college football"), in red are programs statistically more popular with female viewers (such as "Dr. Phil"), and in yellow are programs statistically more popular with elderly viewers (such as "Jeopardy!"). Examining the embedding of the content items, it is visually apparent that two main axes are revealed, the more significant one correlates to age, and the second one loosely correlates to gender. This provides a clear validation that TV content can constitute a strong marker of a household's socio-demographic attributes.

Using deep learning to predict household members' demographics

To put this into practice and extract concrete predictions of socio-demographics attributes (typically age and gender) out of a household viewing behaviour, data science offers a wide range of techniques. Recent advances in computing power and the increased availability of data makes the usage of the decades old technologies of deep learning, and Deep Neural Networks (DNN) in particular, feasible.

Figure 10 depicts an example of a DNN model. In a classic DNN model, a network of neurons is arranged in layers. Inputs enter the network at the first layer. Each neuron at each layer obtains its inputs from neurons of the previous layer, processes the inputs and outputs a result to the neurons of the next layer, until the final layer where the output is the output of the DNN model.

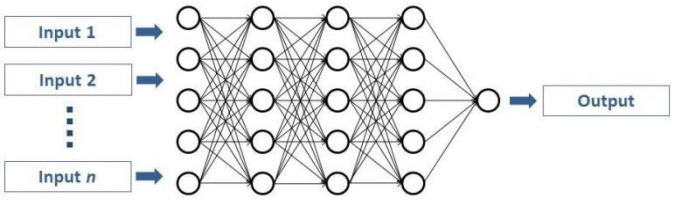

Figure 10 – Example of a DNN model

A DNN model may be trained to automatically set the processing that the different neurons perform. Starting with initial guess for the DNN, training examples with known desired output are provided to the DNN model, and the certain actual outputs are compared to the desired outputs. Then, parameters of the neurons of the DNN model are updated to bring future actual outputs closer to the desired outputs. This update is performed by using a procedure called back propagation to calculate the gradient of the parameters for the training example inputs, and using a (stochastic) gradient descent step to update the parameters. This process is repeated again and again to obtain a DNN model that is more and more accommodated to the training examples.

Different variants of deep learning models are specifically designed to handle different types and structure of data. Some important variants are especially suited to inputs arranged in sequences, including Recurrent Neural Network (RNN).

Through the application of advanced techniques, our research indicates we can reach pretty promising accuracy levels when predicting certain household socio-demographic classes out of a viewing behaviour (in the range of 80 percent).

CONCLUSIONS

This paper examined the expected renewed interest for TV viewership data in the context of the implementation of GDPR, and some operator monetisation strategies (inbound and outbound), which could greatly benefit from the application of AI techniques. This is illustrated by some concrete results that can be derived from TV viewership data (prediction of a household's socio-demographics) through algorithmic works. It brings within reach the promise of a "living" predictive model, updated by continuous streams of viewership data. Because they sit on such a treasure trove of viewership data that could enrich their first-party data, TV service providers may soon be in an enviable position. Interestingly, as much as we believe GDPR may provide an incentive for TV service providers to explore state-of-the-art algorithmic approaches and extract additional insights from the data they collect by themselves, these incentives may actually be limited by the regulation itself when it creates some sort of a "right to explanation" (i.e., the option for a data subject to ask for an explanation of an algorithmic decision that was made about him so as to avoid discriminations) (Articles 13 and 22) (see 'Goodman and Flaxman (7)'). A maximalist interpretation of these stipulations may constitute a real challenge to some of the most advanced AI approaches that are often implemented in a black-box approach, a challenge that we believe data science research will progressively find ways to address.

REFERENCES

1. Laney, D., 2001, "3D Data Management: Controlling Data Volume, Velocity, and Variety, MetaGroup Research," (https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf)
2. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (https://eurlex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679)
3. Burgess M., 2018. "What is GDPR? The Summary Guide to GDPR Compliance in the UK," Wired
4. Marshall J., 2014. "What is Third-Party Data?" Digiday
5. Tiku N., 2018. "Europe's New Privacy Law Will Change the Web and More," Wired.
6. Southern L., 2017. "The State of Addressable TV in 4 Charts," Digiday.
7. Goodman B. and Flaxman S., 2016. "European Union Regulation on Algorithmic Decision-Making and a 'Right to Explanation,'" 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY

ACKNOWLEDGEMENTS

The authors would like to thank Ron Zass for his contributions to their work.