

Operating an Advertising Programmatic Buying Platform: A Case Study

Juan Carlos González and Francisco Mochón

Zed Worldwide, Universidad Nacional de Educación a Distancia (UNED), Spain

Abstract — This paper analyses how new technological developments and the possibilities generated by the internet are shaping the online advertising market. More specifically it focuses on a programmatic advertising case study. The origin of the problem is how publishers resort to automated buying and selling when trying to shift unsold inventory. To carry out our case study, we will use a programmatic online advertising sales platform, which identifies the optimal way of promoting a given product. The platform executes, evaluates, manages and optimizes display advertising campaigns, all in real-time. The empirical analysis carried out in the case study reveals that the platform and its exclusion algorithms are suitable mechanisms for analysing the performance and efficiency of the various segments that might be used to promote products. Thanks to Big Data tools and artificial intelligence the platform performs automatically, providing information in a user-friendly and simple manner.

Keywords — Programmatic Buying, Real Time Bidding, Online Advertising Market, Big Data, Online.

I. INTRODUCTION

THE buying and selling of advertising is no different from transactions carried out in any other market. The central issue is one of supply and demand.

Demand is fuelled by advertisers (directly or via their media agencies), who buy advertising space seeking maximum efficacy for their campaigns and optimal cost efficiency. Supply is driven by media or other formats which offer advertising inventories. The goal as always is to achieve maximum performance while maximizing revenues [11].

The development of the online advertising market, which dates back to 1993, has always been linked to the use and evolution of new technology. New technological developments and the new possibilities generated by the internet have always shaped the market and determined the direction of new developments. Business models also change as technology drives new forms of interaction between supply and demand.

This article focuses on the analysis of a case study of programmatic advertising buying and is structured as follows. The second section analyses the evolution of the online advertising market. The third section looks at how publishers resort to automated buying and selling when trying to shift unsold inventory. Sales are executed automatically according to the usual criteria of ‘buyers’ and ‘sellers’, leading eventually to a programmatic sale. The fourth section introduces a programmatic sales platform for online advertising. This will then allow us to carry out a case study and to watch the platform pinpoint the optimal way of promoting a given product. The platform executes, evaluates, manages and optimizes display advertising campaigns, all in real-time. The fifth section analyses the platform’s architecture and its basic characteristics. The sixth section explains the process followed

to obtain the results and summarizes the main achievements of the case study. The seventh section contains the main conclusions and indicates possible future lines of research.

II. THE EVOLUTION OF THE ONLINE ADVERTISING MARKET

During the early years, almost all inventory was bought for fixed placements. Clients would pay for a certain quantity of a particular placement within a given format or medium. The amount would usually depend on the length of time the advertising would be displayed and the relevance of the placement to that particular medium, either to its scale or to its placement relative to scale.

Advertisers (and agencies which began to branch out into online media) would purchase placements or print-runs directly from existing formats, known as publishers. Available advertising space, known as the ‘inventory’, was sold by print-run (the number of times a creative execution would appear in a given placement). Advertisers would buy print-runs by the thousands, in units known as CPM (Cost per Mile). The development and popularization of internet access gave rise to an increase both in the number of publishers, as well as in the volume of content these publishers generated. The surge in supply meant that a large proportion of inventory remained unsold [18].

This unsold inventory, combined with the newest technological capabilities, gave rise to a new business model and a new player in the value chain: Advertising Networks or Ad Networks, which served as agents or brokers, buying unsold inventory from publishers [8] [9].

Ad Networks make it easier for advertisers to target their campaigns by applying technology to aggregate and segment audiences, packaging and selling advertising accordingly. The technology used by Ad Networks is not particularly sophisticated; in fact it is so negligible as to present a very low barrier to entry into this new segment in the value chain. As a result, the number of Ad Networks soon began to snowball, creating a new problem. The sheer quantity of Ad Networks now marketing unsold advertising inventory in different models or packages soon led to online advertising becoming a highly competitive market, with each actor focused on maximizing their own performance. Buying or selling at the best possible price. The presence of so many players had a negative effect on demand: advertisers could now choose between different Ad Networks to ‘buy’ the same audience more than once. Furthermore, the market soon became in need of efficiency improvements, culminating in yet another new business model and the appearance of another new player in the value chain: Ad Exchanges.

Ad Exchanges allow both advertisers and publishers to exploit audiences rather than print-runs. Ad Exchanges are able to target audiences via publishers’ platforms. Advertisers are then able to select and purchase their target audience. Rather than being booked and purchased directly, audiences are bought via a system of real-time bidding (RTB), in which wins who makes the highest bid [13]. The winning bidder secures the right to position their adverts with the right audience at the right time.

The appearance of this new business model did not entail the disappearance of the previous one; advertisers and publishers were now able to choose between buying and selling inventory via Ad Networks or buying and selling audiences via Ad Exchanges (Fig 1).

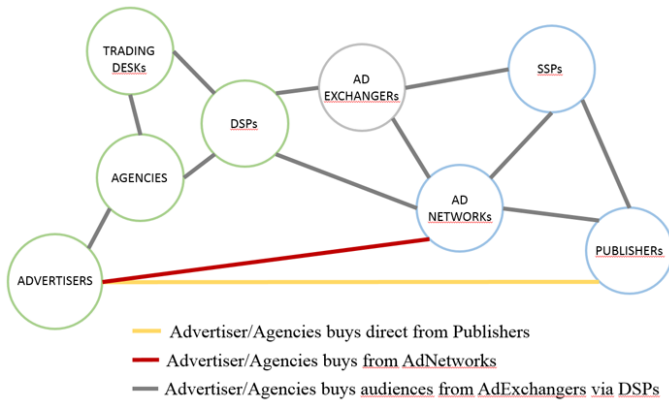


Fig. 1 Online Advertising Business Models.

New opportunities for boosting efficiency and enhancing the buying and selling process began to open up at both ends of the value chain. At the buying end, some advertisers or agencies created their own trading agencies (proprietary agency trading desks), while others joined Demand Side Platforms (DSPs). Both scenarios allow for efficient interaction between advertisers or agencies and Ad Exchanges, via real-time auctions online.

On the selling side, some publishers are able to connect directly with Ad Exchanges, while others connect via Supply Side Platforms (SSPs). The latter process has parallels with the use of DSPs on the purchasing side, in that it allows publishers to interact efficiently with Ad Exchanges, automatically optimizing inventory performance. Nevertheless, the use of Ad Networks remains an effective way of selling inventory.

In addition to the players described above, there are two other crucially important players in the advertising ecosystem: Ad Servers and Data Management Platforms (DMPs). Ad Servers are technological platforms that manage the delivery of creative executions for printing in various formats.

DMPs, as their name suggests, manage data; namely, population segmentation data on groups such as age, gender, location, preferences and so on. This data allows advertisers to select the right audience for each advertising message.

Each and every platform in the advertising ecosystem has been impacted by new developments and technological advances, improving both efficiency and usage rates. This boost has given rise to entirely new forms of interaction, engendering a new purchasing and selling paradigm in the realm of online media [1].

The technological advances responsible for driving this new paradigm have been rolling out over the last five years, precipitating change only when they have coincided efficiently and in concert. Some of these advances include:

- The increase in computational capability: large volumes of complex data need to be processed in milliseconds to allow for real-time decision-making.
- The reduction in storage costs: advertising generates a vast volume of data.
- The application of scientific methods to marketing: marketing professionals make increasing use of Algorithms, Mathematical modelling, Artificial Intelligence or machine learning.
- The increasing speed of data connections.

- Personalization: user recognition mechanisms.
- The use of RTB processes in buying and selling advertising space [17].

Lastly, while the globalization of the advertising market does not in itself qualify as a technological advance, it too has played a role in the birth of the new paradigm. These technologies allow a great deal of rich information to be harvested from advertising activity.

This information can then be applied to boost advertising efficiency, thanks to web analysis and behavioural targeting. As a result, audience acquisition has now become widespread across brands. These brands - no longer exclusively interested in purchasing space - are now also focused on targeting specific audience segments, such as women/men, travel enthusiasts, young people, city dwellers, and so on [4].

On the whole, the online advertising world presents two main options. The first is search engine advertising: advertisers pay to have their creative executions appear when users type certain keywords into search engines such as Google, Bing or Yahoo. The second is display advertising, in which adverts appear when the user visits certain websites. Display advertising normally takes the form of web banners. This paper will focus on display advertising [12].

III. ONLINE ADVERTISING TODAY

As internet penetration advances apace across both computers and mobile devices (smartphones, tablets), online advertising supply continues to outstrip demand. Unsold inventory can constitute a very real risk, and it must be addressed carefully; not least because, if handled properly, it opens up new business possibilities.

Publishers resort to automated buying and selling when trying to shift unsold inventory. Sales are executed automatically according to the usual criteria of 'buyers' and 'sellers', leading eventually to a programmatic purchase [2]. In a programmatic purchase, each party to the transaction (purchaser and vendor) entrusts a machine to complete the sale on their behalves. The purchaser is able to refine the profile selection process - a capacity that every brand demands - while the vendor endeavours to optimize their inventory in the most effective way.

The process has been subject to improvements, thanks to the contribution of data by third parties, which allows to verify the suitability of a particular profile (or profiles) to a particular campaign. Before these third parties arrived on the scene, players had to rely on feedback from only one source (usually cookies from user-identification enabled web browsers) to evaluate their impressions. The presence of neutral third party specialists in tagging and identifying audiences (DMPs and third-party data suppliers) has removed this dependency, and the data they offer is now widely used.

Technology has allowed the automatic buying process to go one step further by enabling the purchasing and selling of media to take place in real-time. Until recently, advertisers used to buy website display advertising, advance-booking the number of imprints that they wanted to display to visitors to those sites. In recent years, the use of APIs has become widespread. These allow advertisers to purchase advertising on an impression basis, with prices negotiated through Real-Time Bidding (RTB) via auction [6].

Instead of purchasing inventory directly from publishers, advertisers now enter into the imprint auction system. Using DSPs, they upload their advertising campaigns, target demographics, and the price they are prepared to pay. Publishers put their inventory and audience at the disposal of Ad Exchanges via SSPs. Agents in this ecosystem, including DMPs and Ad Servers, are integrated into the chain via APIs. Technology takes care of everything else.

While programmatic media selling has solved many of the problems

around supply and demand that were typically encountered in the world of online advertising, particularly when it comes to display advertising, this new form of trading presents a new set of challenges and problems to both sides of the value chain. Programmatic media buying has brought with it the automation of media purchasing, however the system requires something more than automation: intelligence [14].

This article will address some of the problems the purchasing side faces within the new paradigm, and will put forward a solution to some of them [12].

When an advertiser decides to execute a campaign there are three key drivers behind the decision making process: creative executions, target segments, and publishers in whose media these creative executions will appear. The goal, as always, is to refine the campaign and maximise its performance.

Audience selection, however, is no trifling matter, and cannot be rushed. Marketers are not able to predict the effect that the advertising campaign will have on consumer behaviour; on how they will respond, and what effect the stimulus will have on consumer motivation. In other words, initial marketing decisions may not be correct or, at the very least, might be in need of some refinement. However, within this new media sales paradigm, every link in the value chain is capable of sharing information in real-time about a campaign's evolution. This offers a great advantage when it comes to campaign optimization: we now have the chance to analyse a campaign's performance in real-time, and make decisions as and when necessary. Within this environment, real-time decision-making is complex, and calls for skilled analysts and in-house business specialists within the workforce. In addition, the volume of information generated within the new paradigm is vast. This compounds the task; rendering it even more complex, specialised and - ultimately - expensive. In spite of the added complexity, current technological advances supply everything we need to be able to carry out this analysis in a profitable way, without labour-intensive and costly need for direct supervision [15].

This article describes a technological platform and processes that must be in place in order to optimize the three key elements that impact on the performance, efficiency and efficacy of an online campaign: creative executions, audience segments and publishers. We can then begin to maximize the ROI on any campaign [16].

IV. PROPOSAL

Taking what has been mentioned in this paper as a reference point, we will now define and implement a programmatic sales platform for online advertising, which we will call PSP. We will carry out a case study, using the platform to promote a given product. This platform executes, evaluates, manages and optimizes display advertising campaigns, all in real-time.

Campaigns can be optimized according to the following variables: user demographics, the media and distribution channel used by the campaign, and the adverts shown.

The platform continually monitors and analyses the performance of programmed campaigns, as part of a cyclical process that focuses on multiple samples (multiple segments, multiple creativities and multiple publishers), in order to highlight which of these samples are the most efficient. Given that campaigns might need to be tested against a range of variables (segments, creativities, publishers), each of which comes with multiple options, the process has the potential to become very costly from an economic perspective. To mitigate this, we need to execute each campaign and its respective options with a small investment, but enough to generate statistically significant results. This will allow us to make an appropriate decision for each case. We will call this process 'prospection'. Prospection, then, involves running

a campaign through a platform that applies an automated process to determine which combinations of audience, creativity, or publishers are the best fit for a particular product. The investment is relatively small and allows us to carry out advertising test-runs on the necessary segments. The logic employed for deciding which combinations to rule out is based on the hypothesis test, commonly known as the A/B test.

If we find combinations in which the cost of promoting a product is less than the profit from product sales, we can declare the prospection to have been successful, and can then proceed to promote this product on a large scale.

In the event that a prospection is not successful, PSP will continue updating us on promotion costs according to different combinations of audience/creativity/channels/publishers, in a way that ensures that even though the process may not be profitable on this occasion, any loss is by a small margin. It will do this by trying to refine some of the variables, including improving the product's purchasing price, trying out other creative executions, and so on.

The platform operates according to a four-phase process:

Build

PSP initiates the prospection process for a given product. To do this, the platform links a campaign to the corresponding product, uploads the relevant creativities, and assigns a series of basic parameters according to which the prospecting will be carried out; such as start date or budget.

To build the prospection, PSP will create a specific number (N) of demographically segmented campaigns, factor in a given number (M) of creativities (banners) for each, to be published in a limited number (X) of formats.

To do this, PSP will connect with various DSPs via API, and thence gain access to the inventories of different Ad Exchanges that in turn manage the inventory of a large proportion of publishers.

PSP then launches N campaigns in parallel, and begins to gather information from each and every impression.

Measure

Using the information gathered, the platform begins to calculate all of the metrics necessary for applying statistical logic. To lend greater statistical validity, all calculations are based on the concept of a 'unique impression.' The 'unique impression' concept is comparable to the 'user concept.' When the same user sees the same banner three times, Ad Servers will register three discrete impressions. Conversely, the platform will register a single impression. This decision is more onerous in terms of calculus, given that it precludes the counting of impressions according to how many times they are produced. Instead, a more arduous computation must take place each time an impression is produced. This significantly elevates the complexity of real-time processing. It is worth stressing that every single metric handled by the platform is real. The platform never makes use of statistical estimators for calculable variables.

PSP gathers the following metrics:

- Impressions received per user
- Clicks per user
- Unique impressions
- Unique clicks

These metrics are aggregated:

- By segment
- By advert
- By publisher
- By URL

This task is carried out in what could be considered real-time. Five seconds after an impression is produced, or a click is executed, the platform aggregates the event according to each individual criterion.

Optimise

PSP enhances the information it gathers by applying logic according to the following optimization drivers:

- **Bidding Engine**

Every five minutes, each segment is analysed according to ‘speed.’ By ‘speed,’ we mean the number of impressions per hour that a given campaign is achieving. If the speed doesn’t meet the minimum requirements defined at the outset of prospection, we seek to improve it, and vice versa. We control the number of impressions our campaign generates by adjusting our bids in the real-time auction, raising or lowering them as necessary.

- **Fraud Detection Engine**

One of the problems facing the online advertising industry is the ongoing proliferation of automated systems (bots) that generate impressions and execute clicks on adverts.

These impressions and ‘false’ clicks artificially inflate the impression-count, which allows publishers to lay claim to a greater volume of (albeit false) inventory, which they can then sell to Ad Exchanges. Because bot-generated impressions and clicks are clearly unproductive, it is vital that we are able to detect them and, ultimately, eliminate them from our statistical calculations. This is where the Fraud Detection Engine comes in.

Within the platform’s memory, a background program (or daemon) runs an iterative process, reviewing web traffic to seek out patterns that have previously been defined as fraudulent. Some of these patters include detecting the following within a certain period of time:

- An impression-count that is too high to have been generated by a single user.
- A click-count that is too high to have been generated by a

single user.

- A click: impression ratio, or Click-through Rate (CTR) that is too high to have been generated by a single user
- An impression-count that is too high to have been generated by a single IP.
- A click-count that is too high to have been generated by a single IP.
- An impression-count that is too high to have been generated by a single IP.
- Click/impression ratio or Click-through Rate (CTR) that is too high to have been generated by a single IP.
- A break between impressions that is too short to have originated from a single user

Scale

If prospection for a product proves successful, the platform will increase advertising spend on that particular product for the relevant segments.

V. THE PLATFORM: ARCHITECTURE AND OPERATION

Based on all of the above, PSP can be described as an exploration tool that optimizes programmatic inventory sales using real-time bidding (RTB) mechanisms [19].

The process begins when an advertiser decides to create a prospection for promoting a product (Fig. 2).

The platform will then configure the DSP and Ad Server in order to be able to purchase the desired inventory [10].

The impression purchasing process takes place via RTB. Each impression purchased comes with a label supplied by the Ad Server. This label is known as an Ad Tag. Ad Tags are responsible for delivering information to the client’s web browser in order to serve the advert, track user-related information, and monitor the user’s interaction with the advert. Big Data machines then process all of the information gleaned about a user’s interaction with the advert. This process takes

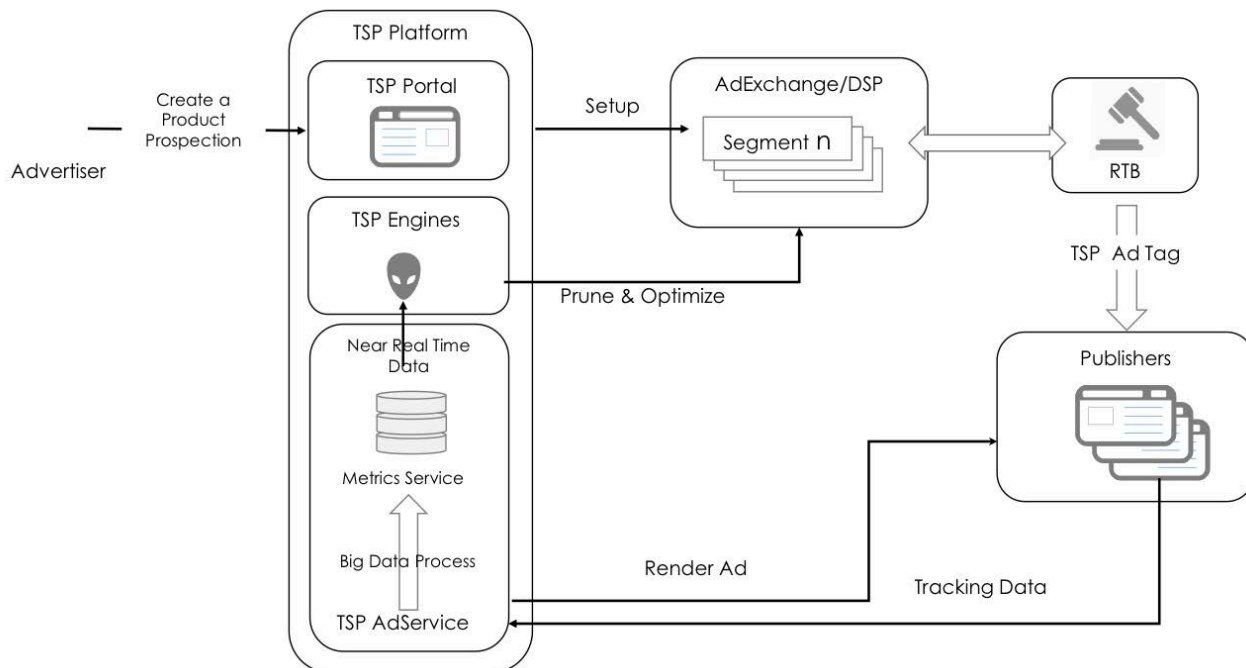


Fig. 2. PSP High Level Architecture

place in Near Real Time (NRT).

Subsequently, the platform's engines then essentially operate in real-time, optimizing inventory purchase, with a view to identifying the segments, adverts and publishers that offer a maximum return on investment (ROI).

Big Data Process

The platform's core is built on Big Data architecture [5]. The rationale for this decision follows on from analysis of the nature and characteristics of the problem at hand. As always, when planning a campaign, we have at our disposal a total of eighty-four possible segmentation combinations; comprised of two gender options, six age ranges and seven HHI (Herfindahl–Hirschman Index) categories. Adding on the various geographical areas (the Designated Market Areas, or DMAs), we would then need to multiply the number of possible combinations by the number of DMAs that exist for a particular country or region. For example, according to Nielsen, the US has 210 DMAs, which would give us 17,640 segmentation combinations. Bearing in mind that digital marketing conventions hold that each individual has between one and one hundred areas of interest, the number of combinations then rockets into millions. In short, the volume of information that needs to be stored and processed is very high [7].

Additionally, given that the platform requires integration into programmatic advertising sales mechanisms, we need to take into account the specific requirements of those mechanisms. The life cycle of a programmatically purchased advertising impression, that is to say, the length of time between a user's arrival at a publisher's page, and the appearance on screen of an advert purchased by auction, is a mere two hundred milliseconds. This means the system has only two hundred milliseconds within which to receive the request, process and analyse the data make a decision and, finally, serve the creative execution. In other words, speed is a key requirement for our platform [3].

In addition, the platform needs to be able to store and process information in a way that satisfies the following requirements:

- Scalability. The platform must be able to execute N prospections simultaneously without requiring any technical adjustments.
- Low latency. Given the fast-paced nature of the RTB market, in which circumstances change from one minute to the next, the platform must be able to analyse events in real-time. Basing a decision on data that are more than five minutes old would, in most cases, be extemporaneous.
- Analytics. While predefined tracking data aggregations do exist (for example the number of impressions and clicks per hour per advert), the platform must also offer a solution that allows data to intersect with the utmost versatility, without needing to pre-process results. This is a fundamental prerequisite for ad hoc analysis – whether seeking to understand user behaviour, find patterns, identify fraud, and so on.

In order to confront these challenges and satisfy these requirements, the platform has been built on Big Data Lambda Architecture. As well as being robust and exact, this architecture allows the platform to combine background (batch) processing with stream processing. While stream processing does not allow for the same level of precision as batch processing, the platform is still able to process and deliver data with very low latency, and continue making decisions practically in real-time.

Flexibility offers further rationale for implementing a Big Data solution; this type of architecture allows virtually infinite scalability. Nevertheless, it is also possible to work with a small and cheap data cluster and incorporate or remove nodes in a straightforward manner as and when necessary. At the same time, technology exists that would

allow the platform to process user demographic information, and to bid higher or lower per impression according to the desirability of a given demographic.

Everything described so far takes place in a programmatic manner, in real time, and by auction (with each impression attracting the interest of a number of different advertisers, and the lot going to the highest bidder).

Objectives

The PSP platform's objective is to help advertisers connect with DSPs in order to maximize their campaigns' ROI. In an automated way, the platform is able to create and launch segmented campaigns and, later, to monitor and manage them in real-time without the need for direct supervision. For each product the advertiser promotes, the platform's goal is to identify the following:

- The most appropriate audience
- The optimal channels and formats
- The best creative execution
- The lowest purchase price

Architecture: modules and their interrelatedness

There are three distinct layers to the platform; each layer is charged with a specific function and is capable of interacting and communicating with the other layers, as well as with any external entities upon which service delivery depends.

PSP Portal

The PSP Portal, the platform's management console, takes the form of a website built in ASP.NET MVC and deployed within Microsoft Azure. The site acts as an interface through which the advertiser interacts with the platform. The console permits campaign configuration (setting benchmarks/thresholds, budget, creative executions and so on). This information is then disseminated directly to the relevant DSPs as well as to the Ad Server, removing the need to access these via their respective interfaces.

The console also tracks the campaign's progress and evolution, and logs this information for each campaign.

PSP Core Engines

The PSP Core Engine is the heart of the platform, and the seat of its intelligence. This layer houses the platform's logic and orchestration capabilities, allowing it to co-ordinate the operation of all the elements involved, both in-house and third-party, including the Ad Server, DMPs, DSPs and so on.

This logic layer is composed of a series of processes (or daemons) that continually run algorithms that process data generated by campaigns. These algorithms allow the advertiser to take timely and wise decisions.

This layer is fully hosted in the cloud (Microsoft Azure Cloud Services). Making use of REST APIs from both the Ad Server and from various DSPs, the layer's daemons interface with the former according to decisions taken by the advertiser.

Ad Server

The role of the Ad Server is to make the adverts appear on the screen by supplying the required information to the client (typically the end-user's web browser, delivered via JavaScript).

The Ad Server also supplies the client's monitoring and identification mechanisms, in order to track both the user's identity, and their interaction with any creativity.

A highlight of the Ad Server's features is its ability to generate near-real-time statistics on ongoing campaigns, thereby providing the PSP with the necessary metrics to make decisions.

The Ad Server is configured as a cluster and, like the PSP, is also cloud-hosted (Microsoft Azure Cloud Services). The cluster's machines are balanced through a layer four load balancer, and continually monitored to verify uptime. The system has an elastic response to the platform's workload, so that if the number of requests rises abnormally within a short space of time, more servers can be automatically added to the cluster. In the same way, if the workload is reduced over a sustained period, machines can be eliminated from the cluster in a tidy manner – without missing out on a deal.

Both the Ad Server and the platform are designed in such a way that the time it takes to process a request and serve an impression is always of the shortest possible duration.

The Platform's Features

1) Initial storage of advert-interaction data.

As soon as the tracking information reaches the Ad Server, it is saved in the Event Ingestor, via a Publisher/Subscriber mechanism that allows the same event to be used by different processes at varying paces (Figure 3).

The platform can either utilize a cloud service, such as Azure Event Hubs or Kinesis, or, alternatively, Storm (part of the Hadoop stack). Given the dynamic nature of workload, the use of one of the aforementioned cloud services is highly recommended.

A daemon consumes events as and when they are generated, caching them in a redundant storage system each time the system logs one thousand requests.

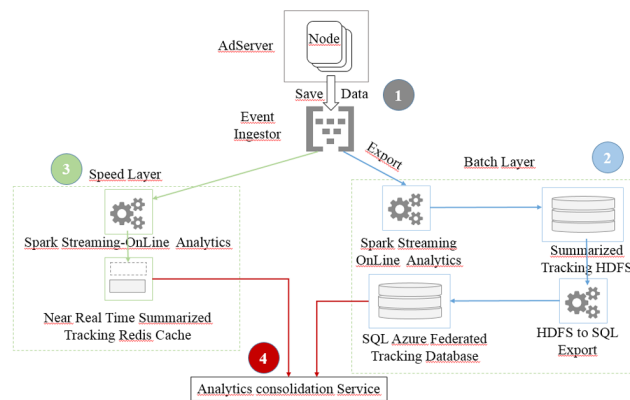


Fig. 3. Platform Features.

2) Batch Layer

Batch processing is tasked with processing all of the information available within a given period of time. This form of information processing delivers accurate aggregated data. The Batch layer consists of an Apache Hadoop cluster distributed by Hortonworks, and all of its processes are orchestrated by Apache Oozie workflows.

The tracking data captured and stored in the previous step is then downloaded in a distributed manner within the cluster's file system. The system proceeds to pre-process, standardize and debug the new logs before finally storing them in HDFS (Hadoop Distributed File System).

In order to improve performance, the file-system is partitioned according to date (via HCatalog). This allows us to easily access and read the logs from a particular time period, without needing to scour through every single log stored in the system.

After the new logs have been stored, the system begins to execute the various processes that are written in Scala on Apache Spark. These processes serve to create aggregated, pre-calculated metrics. Once this data has been calculated, it is inserted into a combined

SQL database. This database is also partitioned by date, thereby allowing the data to be accessed and processed rapidly.

3) Speed Layer (stream)

Stream processing allows us to consume the events stored in the Event Ingesters and to process them with very low latency.

The real-time information-processing application is based on Spark Streaming, using the AMQP protocol for its integration. A Redis database stores information for real-time data processing.

4) Consolidation Layer

In order to be able to take automatic decisions for campaign-optimization, or to be able to display campaign-performance reports, the data must be consumed in near-real-time.

To this end, we have rolled out a REST service, which allows us to check and review the filter and aggregation as desired.

Upon receiving a request, the consolidation layer will receive data while carrying out a parallel consultation on the SQL databases previously fed by the batch processing layer. However, if the query refers to data not yet stored in these databases, the service will access the Redis database, fed by the streaming layer, in order to complete the required information.

VI. TESTING THE PLATFORM: MAIN RESULTS OF THE PILOT

In order to test the platform, we will set up a prospection. To illustrate the platform's operating mechanism with maximum clarity, we will simplify our desired objectives and analyse the results that emerge.

Our prospection focuses on the optimization of two variables: the publishers through which our campaign is to be launched via Ad Exchanges, and the distribution channel. The prospection will be composed of fourteen segments: seven oriented towards web navigation via mobile devices (mobile), and seven oriented towards web navigation via PC (web). The prospection will be endowed with a small but sufficient budget, in order to ensure that the sample is statistically significant. We will not carry out any socio-demographic segmentation, and for didactic purposes will always use the same creative executions.

The platform is integrated with the online advertising market's main Ad Exchanges, giving access to a large quantity of publishers. For our test, we will be using seven Ad Exchanges.

The prospection needs to be able to identify which 'Ad Exchange – Distribution Channel (mobile/web)' combination offers the best performance, and we need to be able to do this in a way that rules out - as quickly as possible - those pairings that would render the campaign's performance significantly lower than average. Campaign performance will be measured according to CTR (Click-Through Rate – the ratio of clicks generated to the total number of impressions).

In order to understand how the platform makes decisions, it is important to understand the statistical logic that is applied. The platform makes decisions by applying contrast hypothesis tests to the population, defined by all segments. Given that we have fourteen segments, we will need to compare each segment's CTR with that of each of the other thirteen, in order to evaluate performance. This means that when analysing segment number one, we will compare its CTR with the aggregated CTR of the other thirteen segments.

When examining segment i , CTR_i will refer to the segment in question, while CTR_{ri} will refer to the aggregated CTR for the remaining thirteen segments, r_i . Note that both segment i and the aggregated remaining segments r_i constitute samples of our population and, furthermore, that all fourteen segments are discrete. This affords us the statistical independence required to carry out our test.

The problem of how to select the best segments can be tackled in a number of ways. We will outline the most straightforward scheme that has been implemented in the PSP. In order to execute the algorithm that we are about to describe, we first need to establish the significance value of the test (α) as well as that of the statistical significance (β) (the Appendix contains a detailed description of these) which we would like to employ in our study. Or, to put it another way, the degree of confidence in our test, and its statistical power. In this case study we have established the values: $\alpha = 0.05$, which translates to a confidence level of 95%, and $\beta = 0.2$, which equates to a statistical power of 80%.

Once we have established these premises, the algorithm for excluding a particular segment i works in the following way:

1. Calculate (or estimate) CTR_i and CTR_{r_i} and thus estimate the $CTR_i - CTR_{r_i}$ magnitude. We will use this magnitude in our hypotheses.
2. Calculate the statistical Z (see Appendix) and ascertain whether the Z value you have obtained falls below the critical value of Z, determined by $\alpha = 0.05$. To achieve the same goal, you can also identify which p-value corresponds to the Z value obtained, and determine whether it is less than -1.645, this being the corresponding p-value to $Z = 0.05$ in queue analytics. If it does, the null hypothesis is rejected and the opposite case is accepted.

Some of the results obtained for total impressions, clicks and CTR by week, for each segment, are collected in table I. These results were obtained following the application of the exclusion algorithm over the course of a five-week-long prospection. The platform's algorithms use this data as a starting point for carrying out the calculations necessary to make the required decisions. Table 1 includes some of the calculations that are most relevant and essential for making such decisions.

TABLE 1. MAIN RESULTS

Segment ID	Week 1		Week 2		Week 3		Week 4		Week 5	
	Z	Power	Z	Power	Z	Power	Z	Power	Z	Power
1655571	0,3548	10%	0,5065	13%	0,3547	10%	-0,2122	8%	0,7213	18%
1655572	-1,6707	51%	0,2315	8%	-0,2551	8%	-0,8562	22%	-1,0543	28%
1655573	7,0197	100%	8,1893	100%	6,9097	100%	5,8508	100%	7,7165	100%
1655574	0,7197	18%	2,9603	91%	3,3045	95%	5,3317	100%	4,8030	100%
1655575	-0,7371	18%	-1,7325	53%	-2,2183	72%	-2,7612	87%	-	-
1655576	0,8309	21%	-1,1253	30%	-1,7253	53%	-2,3646	76%	-3,7834	98%
1655577	-0,4254	11%	-0,7204	18%	0,5512	14%	-0,3437	10%	-0,7585	19%
1655579	-25,8250	100%	-	-	-	-	-	-	-	-
1655580	-1,0663	28%	-4,3884	100%	-	-	-	-	-	-
1655581	4,2081	99%	2,6964	85%	4,5383	100%	5,5031	100%	6,5208	100%
1655582	-6,5587	100%	-	-	-	-	-	-	-	-
1655583	-5,4606	100%	-	-	-	-	-	-	-	-
1655584	-6,9300	100%	-	-	-	-	-	-	-	-
1655585	-9,7202	100%	-	-	-	-	-	-	-	-

■ Ruled out Segments
■ Not ruled out by lack of power

Table 1 shows the values calculated for Z and for statistical power, organised by week and by segment. In the left-hand side of the table, we find the Segment ID column, which shows the identifier that the platform assigns automatically to each segment. We can also see that the data is organised by distribution channel: the first seven lines correspond to campaigns oriented towards the mobile channel, while the final seven correspond to web channel campaigns.

The colour-coding indicates the status of each segment at the end of each week¹. According to this colour code, there are three possibilities:

- **Red:** segments excluded due to low performance. For example, the segment identified by 1655583 had a Z value of -5.4606 in the first week (way below $\alpha = 0.05$), and a

power of 100%, which means that we have power enough to make a decision. Consequently, the segment was excluded. As soon as a segment is excluded due to poor performance, its execution is halted, so that no more impressions are purchased for it.

- **Pink:** segments which produce low performance in comparison with the aggregate, but for which the analysis has not reached sufficient statistical power to be able to reliably declare a verdict of low performance. For example the segment identified by code 1655580, which had a Z value of -1.0663 in the first week (well below $\alpha = 0.05$), nevertheless only has a statistical power of 28% (way below the minimum of 80%). This means that the result cannot be considered conclusive. Therefore, this segment remains active. However, in the second week, the Z value is still below α , but this week its statistical potency is now 100%, which means that the segment is then excluded, and its execution is halted.
- **White:** segments that show similar performance to that of the aggregate. For example the segment identified by code 1655574 had a Z value of 0.7197 in the first week (way above $\alpha = 0.05$). This segment then continues to be executed.

The evolution of calculations across the weeks shows that the platform is able to make conclusive decisions in accordance with the pre-determined significance and power parameters. We can see how the platform excludes those segments whose CTR is significantly lower than that of the aggregate.

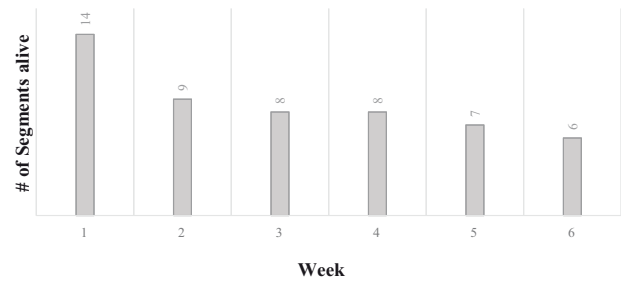


Fig. 4. Segment survival at the beginning of the week.

In figure 4 we can track the survival of different segments on a week-by-week basis, as the algorithm is being executed. At the beginning of week 1, the number of active segments stands at initial fourteen segments. By week 2, the platform has excluded five segments and only nine active segments remain. At the beginning of week 3, eight active segments remain, and the process continues until only six out of the initial fourteen segments have survived by the beginning of week 6.

Of those six active segments that survive past the end of week 5, it is worth highlighting that three (1655573, 1655574 and 1655581) have Z values far above those of the exclusion value α , coupled with high statistical power (100%). One surviving segment (1655571) has high Z values but low power (18%), and two (1655572 and 1655577) have Z values below α , but with power insufficient to lead to exclusion.

The behaviour of the two latter segments is noteworthy. Almost throughout the entire prospection, these two segments maintain consistently low performance, but never attract the statistical power sufficient to lead to their exclusion. What is happening here is either that the segment's CTR values are patchy or uneven, or that they are very close to, or far from, the aggregate CTR. In this situation, we would need to use very large sample sizes if we are to achieve sufficient power to enable us to make a decision. This situation is highly likely to occur, due to the fact that campaigns are subject to price competition within the auction-purchasing model. That is, when the platform goes to an auction in order to purchase impressions for a particular segment, either there is no available inventory or, more

¹ In fact, the platform does not make decisions on a weekly basis, but continuously, however this way of introducing information helps us to understand the mechanism of operation thereof.

likely, the auction concludes with an impression purchase price that exceeds the maximum budget established when the campaign was configured. In other words, the platform does not make the purchase. Ultimately, the platform is either unable to purchase impressions, or the traffic it achieves is not of sufficient quality.

Given that the algorithm excludes those segments whose CTR falls below that of the aggregate, the first effect that we can observe is that the prospecting's global CTR increases on a weekly basis. Thus, in figure 5, we can see how the prospecting's global CTR began at 0.0011 in the first week, concluded at 0.0025 in the fifth week. This is a direct result of the process of excluding lower-performance segments.

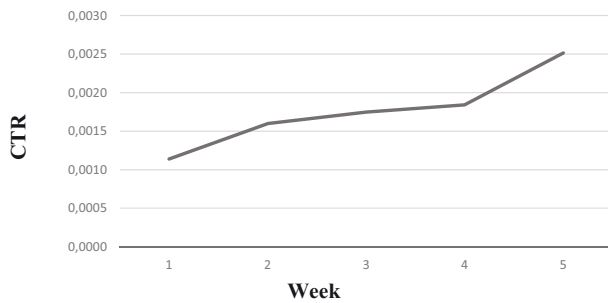


Fig.5. CTR per Week.

A key result of the segment-excluding process is that the mobile communication channel appears to be more appropriate and effective than the web channel when it comes to carrying out campaigns for the product associated with this prospecting. In accordance with the data in table 1, we can see that, of the first seven segments that relate to the mobile channel, five survive until the very end of the prospecting. Of the web channel segments, only one survives the prospecting's full five weeks.

From a business perspective, it is vital that we identify whether the prospecting delivers a clear economic benefit. To this end, we will need to evaluate the cost and revenue that the prospecting generates. Costs are determined by purchases that the platform makes in the various auctions to which it has access. The information available will permit the establishment either of a global prospecting cost for a defined period of time, or of a cost per click.

Revenue will result from conversions; from each and every click that results in a real-life product sale.

In order to gauge the prospecting's benefits or performance, we can simply compare total revenue with total cost. Figure 6 shows the evolution of revenue vs. cost throughout the six weeks of the prospecting.

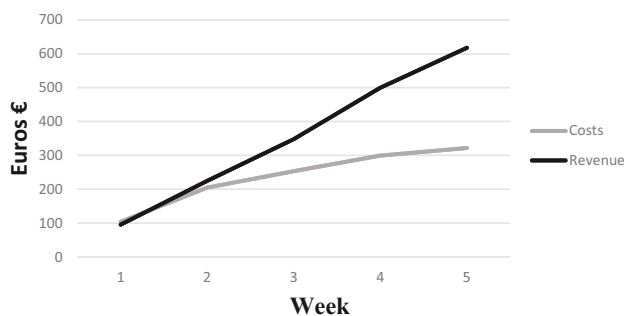


Fig. 6. Costs and Revenue.

As the platform discards segments with a worse outcome the

overall campaign CTR grows, which means that the same number of impressions will give us, week after week, a greater number of clicks (figure 5), and logically a greater number of conversions (clicks that ends in a purchase). In short, the selection process is choosing those segments that are more effective.

Costs increase as we continue purchasing impressions for successful segments. In some cases at a higher price due to the bidding process, but as the effectiveness is better, costs grow at a lower rate than revenues do.

As the platform filters segments that are more efficient, the overall number of impression grow; if the number of clicks is higher the number of conversions is higher, and therefore revenues grow at a higher ratio.

VII. CONCLUSIONS

This paper has tested the application of Big Data techniques and tools, as well as Artificial Intelligence algorithms, to the online advertising purchase process. Specifically, a platform developed in-house and integrated in real time within the online advertising ecosystem, has been employed. The platform aims to find those segments in a particular campaign with the best performance (the champions), i.e. those that maximize ROI.

This system allows advertisers to run prospectings, prior to executing the real campaign, in order to find those champion segments that maximize results. The advantage of running a prospecting is that advertisers will find the better segments with a relatively low investment. Finding the champion segments will allow advertisers to spend their money in proven successful segments.

It is worth highlighting that the platform works without the need for direct supervision, and delivers information in a simple, user-friendly manner. This greatly simplifies the workload of personnel engaged in campaign planning, as well as that of staff on the product side. Traditionally, these professionals would have had to access a range of information sources (the interfaces of various Ad Servers or DSPs), with the added complication that this information would burden non-technical personnel with an unwieldy amount of calculation sheets.

Results from the pilot prospecting have shown that the platform algorithms are able to determine, rapidly and with enough statistical significance (enough statistical power), which segments are more appropriate for a single campaign, because they are more profitable,

The analysis shows that the platform was able to discard segments with poor CTR. As a result of these decisions taken automatically, the overall CTR of the prospecting increased week after week and consequently the overall profitability of the campaign improved significantly.

VIII. FUTURE WORKS

When it comes to possible future investigations, it is worth noting that there are a number of ways of improving the platform, among which these are key:

1. Given that one of the determining factors of the platform's statistical analysis is the minimum sample size, it would be advantageous to identify reliable ways of optimizing this process; establishing smaller samples that maintain reliability and still reach the statistical power necessary for making decisions. In this sense, it is worth pointing out the possibility of using methods based on the Bayes theorem, in order to establish minimum sample size.
2. In spite of the fact that CTR can, as we have seen, be an acceptable method of judging a campaign's performance,

from a business perspective it makes more sense to apply an economic performance criterion, one known as performance display. Nevertheless, this selection criterion, while desirable, would significantly complicate both the volume of data and the calculations required.

Future lines of investigation should be guided towards these two paths of action.

APPENDIX – APPLIED STATISTICS

The central idea of the platform consists of excluding segments as soon as sufficient statistical evidence supports that CTR_i is lower than CTR_{ri}.

It is important to note that, while we have been working with CTR (click-through rate, or the relationship between clicks generated and total impressions), this value is a ratio. We could also affirm that this ratio is in fact a probability, given that the fact that a user clicks on an advert can be considered a success, and similarly, the creative execution's total impressions can be viewed as the total number of possible cases. Given that we are comparing the samples' ratios, and are only interested in knowing whether the ratio of one is lower than another, what we are in fact carrying out is a contrast hypothesis test between the ratios of two samples, within only one-tailed test that of the left.

We can present our hypothesis in this way:

Nul hypothesis H₀: CTR_i = CTR_{ri}

Alternative hypothesis H₁: CTR_i < CTR_{ri}

Or, alternatively, we could present it thus:

Nul hypothesis H₀: CTR_i – CTR_{ri} = 0

Alternative hypothesis H₁: CTR_i - CTR_{ri} < 0

That is to say, we will establish that our nul hypothesis H₀ holds that the sample's CTR (CTR_i) is equal to the CTR calculated for the aggregate of the remaining segments (CTR_{ri}), and that, therefore, there is no difference between the two. Our alternative hypothesis H₁ maintains that the sample's CTR is lower than the CTR calculated for the aggregate of the remaining segments (CTR_{ri}) and that, therefore, there is a difference between the two.

Clear evidence against the null hypothesis and in favour of the alternative hypothesis consists of a CTR_i value lower than that of CTR_{ri} or, in the same way, a difference that is significantly below zero. The reason we need to employ a contrast hypothesis is precisely because it is difficult to pinpoint the meaning of the term 'significantly,' particularly when sample size is neither fixed nor known, and, furthermore, when there is a significant difference between the sample sizes we are comparing [21]. While it is possible to observe differences in both directions, namely CTR_i – CTR_{ri} < 0 as well as CTR_i – CTR_{ri} > 0, in reality we are not concerned with whether the segment is significantly more effective than the others, but only with whether it is less effective, in which case it is excluded. Under the same conditions, a one-tailed test offers greater statistical power than a two-tailed test.

The concept of statistical power refers to the reliability of a test when it comes to preventing misguided decisions. Power refers to the probability of rejecting the null hypothesis H₀ when the alternative hypothesis H₁ is true. This is also known as the probability of committing a type II error. [20]

In our scenario, this means that, provided that the sample size is sufficiently representative, the process can be executed sooner and, therefore, a decision can be made sooner. The power of a test depends on the relationship between the sample sizes compared, and the level of significance α established. It is vital that the test has sufficient statistical power to allow us to make decisions with an acceptable

margin of error.

We are analysing independent events (the fact that a user clicks does not depend, and does not affect, the behaviour of another user), with discrete variables (a user does or does not click; there no intermediary values), and thus our variable - CTR - will follow a Bernoulli distribution. This allows us to establish variance as: σ=p.q or σ= p (1-p). This amounts to saying σ as the product of the probability of success with that of failure. As stated previously, CTR remains a probability. In this way, we could state that CTR is the probability that a user will click, given that it constitutes the ratio between successes (clicks) and all possible cases (impressions), and is nothing more than CTR. In other words, variance can be expressed thus:

$$\sigma = \text{CTR} * (1 - \text{CTR}) \tag{1}$$

Taking standard deviation into account, we can calculate the standard error as:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\text{CTR} * (1 - \text{CTR})}}{\sqrt{n_i}} \tag{2}$$

Given that the sample size of our population n is huge, and in accordance with the central limit theorem, we can reconcile this distribution to a normal distribution provided that the population average and variance have finite values, and as long as n·p > 5 and n·q > 5 is true. In the preceding equation, n is the population size, p is the probability of success, and q the probability of failure.

This fact will facilitate the process of choosing which statistic to employ when carrying out our calculations. For this purpose, we will use the Z statistic (also known as Z-score), unlike the case of two proportions, calculated as:

$$Z = \frac{P_i - P_{ri}}{SE_{i-ri}} = \frac{|CTR_i - CTR_{ri}|}{\sqrt{\frac{CTR_i * (1 - CTR_i)}{n_i} + \frac{CTR_{ri} * (1 - CTR_{ri})}{n_{ri}}}} \tag{3}$$

Once we have calculated the Z value, the following formula will tell us the test's power[22]:

$$Power = 1 - \beta = \Phi(Z - \Phi^{-1}(1 - \alpha)) = \Phi(Z - Z_{1-\alpha}) \tag{4}$$

In which Φ() y Φ⁻¹() represent, respectively, the function of normal standard distribution, and the reverse; Z_{1-α} represents up to the (1-α) quantile of Φ(), that is, the Z value to its left (the area below the curve) is equal to 1-α; β is the type II error, and α is the significance, or type I error.

$$n_{ri_{min}} = \left(\frac{CTR_i(1-CTR_i)}{k} + CTR_{ri}(1 - CTR_{ri}) \right) \left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{CTR_{ri} - CTR_i} \right)^2 \tag{5}$$

In which k is the relation between the studied segment's sample size and the remaining segments, that is:

$$k = \frac{n_i}{n_{ri}} \text{ and } k = \frac{n_{i_{min}}}{n_{ri_{min}}} \tag{6}$$

ACKNOWLEDGMENT

The platform used in this paper has been developed and deployed by the company Zed Worldwide S.A thanks to whom this article was made possible. We would like to thank, as well those people responsible for the platform, for their recommendations, support and help in the

elaboration to make this paper. Especially to Fernando Perez, Head of Big Data, Alberto Ochoa, Data Scientist, Juan Pablo Rizzo, Platform architecture and William B. Henn Innovation Director.

REFERENCES

- [1] S. Adikari and K. Dutta. (2015) Real Time Bidding in Online Digital Advertisement. In *New Horizons in Design Science: Broadening the Research Agenda*. 10th International Conference, B. Donnellan et al. (Eds.), pages 19-38. Springer.
- [2] O. Busch (Editor) (2014). *Programmatic Advertising. The Successful Transformation to Automated, Data-Driven Marketing in Real-Time*. Springer Science+Business Media, 2014. <http://www.springer.com/us/book/9783319250212>
- [3] G. Cormode, and S. Muthukrishnan. (2005). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*. Volume 55, Issue 1, April, Pages 58–75.
- [4] K. Cox. (2015). Programmatic Helps Brands Make the Most of Micro-Moments. Product Marketing Manager. DoubleClick by Google. <https://www.thinkwithgoogle.com/articles/programmatic-helps-brands-make-the-most-of-micromoments.html>
- [5] L.F. Chiroque , H. Cordobés, A. Fernández Anta, R. A. García Leiva , P. Morere , L. Ornella , F. Pérez and A. Santos. (2015). Empirical Comparison of Graph-based Recommendation Engines for an Apps Ecosystem. *International Journal of Interactive Multimedia and Artificial Intelligence*. March, 2015 http://www.ijimai.org/JOURNAL/sites/default/files/journals/IJIMAI20153_2.pdf
- [6] U. M. Dholakia. (2015). The Perils of Algorithm-Based Marketing. *Harvard Business Review*. June 17. https://hbr.org/2015/06/the-perils-of-algorithm-based-marketing&cm_sp=Article_-_Links_-_End%20of%20Page%20Recirculation
- [7] W. Fan and A. Bife. (2012). Mining big data: current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*. Volume 14 Issue 2, December, Pages, 1-5.
- [8] D. Field, P. Zwillenberg, J. Rosenzweig, N. Zuckerman and M. Ruseler. (2015). The Programmatic Path to Profit for Publishers. *BCG Boston Consulting Group*. July.
- [9] D. Field, O. Rehse, K. Rogers, and P. Zwillenberg. (2013). Efficiency and Effectiveness in Digital Advertising. Cutting Complexity, Adding Value. *BCG perspectives*. https://www.bcgperspectives.com/content/articles/media_entertainment_marketing_cutting_complexity_adding_value_efficiency_effectiveness_digital_advertising/
- [10] A. Gandomi, and M. Haider. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* Volume 35, Issue 2, April, Pages 137–144.
- [11] IAB Spain. (2014) Libro blanco de compra programática. IAB Spain, 2014. <http://www.iabspain.net/wp-content/uploads/downloads/2014/09/Libro-blanco-de-Compra-Program%C3%A1tica-y-RTB.pdf>
- [12] P. Minnium. (2014). The Definitive Guide to the Digital Display Ad Ecosystem. Marketing Land. October. <http://marketingland.com/digital-simplified-new-advertising-supply-chain-104734>
- [13] A. Mochón, and Y. Sáez. (2015). Understanding Auctions. Springer Texts in Business and Economics. <http://www.springer.com/br/book/9783319088129>
- [14] J. F. Rayport. (2015). Is Programmatic Advertising the Future of Marketing? *Harvard Business Review*. June. <https://hbr.org/2015/06/is-programmatic-advertising-the-future-of-marketing>
- [15] O. Shani. (2014). Get With The Programmatic: A Primer On Programmatic Advertising. Marketing Land. August. <http://marketingland.com/get-programmatic-primer-programmatic-advertising-94502>
- [16] M. Shields. (2014) CMO Today: Marketers Puzzled by Programmatic Advertising. *The Wall Street Journal*. April <http://blogs.wsj.com/cmo/2014/04/01/cmo-today-marketers-puzzled-by-programmatic-advertising/>
- [17] M. Stange, and B. Funk. (2014) Real-Time Advertising. *Business & Information Systems Engineering*, Volume 6, Issue 5, pages 305-308. <http://link.springer.com/article/10.1007%2Fs12599-014-0346-0>
- [18] P. Zwillenberg, D. Field, M. Kistulínec, N. Rich, K. Rogers, and S. Cohen. (2014). Improving Engagement and Performance in Digital Advertising. Adding Data, Boosting Impact. *BCG perspectives*. September 16. https://www.bcgperspectives.com/content/articles/marketing_digital_economy_improving_engagement_performance_digital_display_advertising
- [19] A. Woodie. (2015). What's driving the Rise of Real-Time Analytics. *Datanami*. September. <http://www.datanami.com/2015/09/15/whats-driving-the-rise-of-real-time-analytics/>
- [20] Ioannidis JPA (2005) Why Most Published Research Findings Are False. *PLoS Med* 2(8): e124. doi:10.1371/journal.pmed.0020124
- [21] Martin Goodson. (2014) Most Winning A/B Test Results are Illusory. *Qubit* http://www.qubit.com/sites/default/files/pdf/mostwinningabtestresultsareillusory_0.pdf
- [22] Matthew S. Nelson, Alese Wooditch, Lisa M. Dario (2015). Sample size, effect size, and statistical power: a replication study of Weisburd's paradox. *Journal of Experimental Criminology* March 2015, Volume 11, Issue 1, pp 141-163



Juan Carlos Gonzalez has a Bachelor Degree in Chemistry from Universidad Autonoma de Madrid and he's got an MBA from IE Business School. He counts on more than eighteen years' experience in Telecom, Technology, Media and Internet sectors. He is currently Chief Innovation Officer at the Zed Worldwide Spanish multinational company, where he is currently working in the following innovation lines: Big Data, Advertising, Mobile Payments and Mobile Financial Services or Security and Privacy among others. He leads several R&D projects, in collaboration with different European universities, funded by various public and semi-public Spanish and European Institutions, under the multiple existing R&D program aids, like Horizon 2020



Francisco Mochón has a PhD in economics from the Autonomous University of Madrid and from Indiana University and is a Fulbright scholar. Currently he is full Professor of Economic Analysis at UNED, Madrid. He has been Advisor to the Ministry of Economy and Finance of Spain, Director General of Financial Policy of the Government of Andalusia, CEO of the research firm ESECA and Chief Financial Officer (CFO) of Telefónica of Spain. He has been the Chairman of the Social Board of the University of Malaga. Currently Prof.dr. Mochón is a member of the advisory committee of U-TAD and member of the advisory committee of the Futures Market of Olive Oil (MFAO). He has published numerous research articles and is the author of more than fifty books on economics, finance and business. Currently his research interests are the Economics of Happiness in the business environment and the Digital Economy. He has been director of the MOOC course "Felicidad y práctica empresarial".