ARTEFACT

DATA FOR FINANCE

ARTEFACT

We transform **data** into **value** and business **impact**.

14
COUNTRIES

+1000
EMPLOYEES

+300
MAJOR BRANDS

Artefact is a global data-driven services company. Our offers sit at the intersection of consulting, marketing and data science, putting consumers at the heart of enterprises' digital transformation.

DATA CONSULTING | DATA & DIGITAL MARKETING | DIGITAL COMMERCE

TABLE OF CONTENTS

Data for Finance Bank & Insurance

Interview of Vincent Luciani, co-founder & CEO of Artefact

The data consulting specialist is gaining momentum and already has 1,000 employees in 15 countries, only 7 years after its creation.
In a highly fragmented market, Artefact, a consulting firm specializing in data, stands out by offering a particularly popular service around data and digital: consulting, development and integration of AI solutions applied to business, operational support for client teams, even training.

What are the services and solutions offered by Artefact?

Vincent Luciani: Artefact offers the full range of services needed to help a company exploit the potential of their data and create business impact.

Artefact was founded in 2014 by three partners, one with consulting expertise, one with digital marketing expertise and one with AI research expertise, with one goal: to bridge the gap between data and business. Artefact helps companies rethink their organization around the use of data through three types of services:

- **The broadest and most advanced data-driven marketing offering** on the market, based on our historical core business. The objective: to optimize marketing performance through the use of data. Because the digitalization of marketing took place long before that of other departments, we have taken a considerable lead in the development of personalization and performance measurement solutions. This has given us a unique experience in combining data science with digital marketing.

- **Offers around «data readiness»:** organization, data governance, creation of a Data Factory...which aim to transform the company in the long term

- **Business solutions based on advanced AI models,** semantic analysis or image recognition, for example, in order to predict sales volumes or to automate the processing of call center requests. The field of possibilities is very vast, and we study our clients' challenges carefully in order to respond to them in a specific way.

How can AI be used to improve performance?

V. L.: We have turned a corner in the last 2-3 years. Data, coupled with automation or AI models, has successfully demonstrated its profitability in all sectors. It has an important role to play in creating a competitive advantage by attracting more new clients or lowering costs. For example, Artefact has deployed a demand forecasting model with Carrefour for the Bakery and Pastry department, which gives a sales volume to be forecasted for the day, for each hypermarket, from multiple parameters (day, weather, school calendar...), allowing the department manager to adjust their daily production. This solution resulted in a significant reduction in food waste and an increase in client satisfaction as they get fresh bread all day long! Another significant example for L'Oréal is the use of Machine Learning algorithms analyzing millions of documents and images of influencers, in order to identify the weak signals of emerging trends in the field of beauty and cosmetics. Here again, AI proves its ability to be a real vector of product innovation anticipating consumer needs!

What is Artefact's added value?

V. L.: Artefact is Art...and Fact! It is very rare to gather all the expertise needed for these very complex projects under the same roof. Our clients are often forced to multiply the number of service providers: strategy consultants, developers, data scientists, product owners, etc...creating delays and risks of inconsistency. Artefact brings together three main types of professions in a single entity: consultants with business expertise, data experts (data scientists, data analysts, data engineers), and digital marketing experts, who are essential for mastering highly evolving media formats. Another asset: our vast expertise allows us to assemble and create standardized technological bricks tested in many different business contexts, which we make available free of charge to our clients and make available in open-source.

What types of organizations rely on Artefact?

V. L.: All functions and sectors of activity are affected. Our clients are mainly large international organizations that have challenges with organizational silos (data is present in several places in the company), and industrialization and scalability. Our clients include renowned international brands such as Danone, Samsung, Orange, Heineken and Unilever.

Any upcoming project?

V. L.: Our goal is to quadruple the size of the company in 4 years, both through organic growth (we are recruiting 500 people by 2022) and through strategic acquisitions. Hence, to gain flexibility, our recent exit from Euronext Paris and the arrival of two of the best LBO funds (Ardian and Cathay Capital) to help us accelerate our growth. We also want to support the structuring of careers around data: we have created the Artefact School of Data to address the huge shortage of talent in our business, by offering a professional opportunity to anyone wishing to transition to the data industry.

"Artefact bridges data to business to create value."

How Artificial Intelligence is impacting the future of banking and insurance

To meet the challenges presented by digital competition (24/7 service, neobanks, crypto-currencies…), banks and insurers are implementing a wide range of AI technologies, including cloud computing, virtual assistants, machine learning, predictive analytics and natural language processing to gain competitive advantage. By transforming the way they do business, they're increasing customer satisfaction – and retention.

- The global AI fintech market is predicted to reach $22.6B in 2025, up from $6.67B in 2019
- The aggregate potential cost savings for banks from AI applications is estimated at $447 billion by 2023
- 75% of banks with $100 billion or more in assets implement AI, and 46% of smaller banks

The uses of AI in banking and insurance have multiplied since the COVID-19 pandemic dramatically boosted global digital engagement. According to the IMF, banks are increasingly relying on AI systems to improve underwriting processes and fraud detection for the vast number of pandemic-related loan applications and to comply with mandated relief requirements. And in a global survey of Insurance CEOs by KPMG, 85% of respondents say COVID-19 has accelerated the digitization of their operations and the creation of next-generation operating models.

AI can automate processes to reduce costs and increase employee engagement, ensure greater customer loyalty through personalisation, and unlock rapid innovation cycles to improve everything from risk management to cybersecurity, and customer service.

1— Embracing the cloud simplifies banking

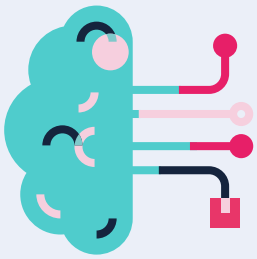
"Adoption of the cloud by banks can lead to improved efficiency, enhanced innovation and greater agility by allowing them to focus on their core business. Cloud solutions can also help banks and insurers gain market share with innovations that positively impact the customer relationship."

Fierce competition within the financial and insurance sectors is forcing companies to find less costly and more effective ways to do business. Enterprises that adopt cloud models can greatly improve employee productivity, rapidly deploy new products and services and significantly reduce operating costs.

For banks, improvements in security and compliance have made cloud-based strategies more lucrative and appealing, but due to strict regulations and client privacy, some executives are still hesitant. Data thefts can occur. In 2019, famously, a hacker obtained the personal data of over 100 million people from the Capital One bank. The bank was fined and ordered to strengthen its security controls. Nonetheless, there's no doubt that the future of banking will be in the cloud: it frees up bandwidth so banks can focus on their core business, which is delivering better customer service, data protection, and regulatory compliance.

Cloud services help insurance companies save money by streamlining time-consuming, repetitive business tasks such as claims processing, which includes multiple tasks, including review, investigation, adjustment, remittance, or denial. This repetitive procedure is prone to human error, but cloud technologies can process documents rapidly, detect fraudulent claims, and check if claims comply with regulations. Interest among insurers is high: according to Aite-Novarica Group, more than 90% of insurance companies were using cloud computing for their business processes as of June 2021, and that figure is expected to increase.

2 — AI detects fraud more efficiently than humans



The global fraud detection and prevention market value is projected to grow from $26.97 billion in 2021 to $141.75 billion in 2028, with a CAGR of 26.7%. This anticipated growth will occur globally across industries including BSFI (Banking Securities Finance Insurance), manufacturing, healthcare, and others. Market growth is also driven by users shifting their focus towards e-commerce platforms and investments by key players to develop secure fraud solutions.

In banking, the enormous processing and learning power of AI is being used to analyze massive quantities of data almost instantaneously to fight against all types of fraud, from identity theft and app scams to counterfeiting and cheque forgery, providing both quantitative and qualitative benefits, and achieving continuous improvement in results.

AI is enabling credit card companies to incorporate predictive analytics into their existing fraud detection workflows to reduce false positives. In the fight against anti-money laundering, algorithms are being used to analyze vast pools of data and raise red flags if unusual transactions or suspicious account activity is detected.

AI can predict and prevent credit risks and identify individuals and businesses who might default on their obligation to repay their loans. It can also identify malicious acts such as identity theft.

For insurance companies, where risk management is their core function, AI is viewed as the future, promising smarter risk assessment and operations, while enabling knowledge workers to make better and faster data-driven decisions using AI-powered text-analytics platforms. AI can automate risk management tasks like recognizing underwriting risks and detecting fraud, while AI's natural language processing and advanced analytics capabilities can be used to extract pertinent risk information from emails to identify underwriting risks.

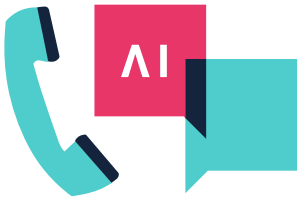
3 — The growing cybersecurity market

Cybersecurity is one of the biggest threats facing organisations today. According to a report by Astute Analytica, global cybersecurity spending will reach $174.7 billion in 2024, with security services the largest and fastest-growing segment.

AI is mainly used to detect and respond to cyber threats (cyberterrorism, ransomware attacks, malvertising, phishing…) thanks to its ability to identify irregular traffic through machine learning or deep learning. AI can also predict critical vulnerabilities, automatically identify their assets and network topology, and continuously improve network defenses against potential cyberattacks.

The personal cybersecurity market is also expanding, as more and more consumers demand protection not merely for their computers, but also against physical threat actors in the form of home monitoring software with real-time support and rapid incident response.

4 — AI-powered chatbots for call centers



When a consumer has a credit card problem or a homeowner has a claim to file, they don't want to wait until office hours to communicate with their banks and insurance. They expect round the clock responses. Thanks to conversational AI, chatbots are helping companies better respond to the needs of clients while saving time and costs.

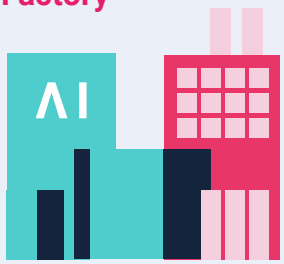
In banking, it's estimated that 826 million hours will be saved through chatbot interactions in 2023, while Insurers, recognizing the promise of AI for streamlining claims processing and underwriting, are set to spend $3.4 billion on AI platforms by 2024.

Artefact's client HomeServe, a world leader in insurance services, had an AI-based conversational solution, but wanted to improve efficiency and customer experience with AI. By implementing speech analytics for topic classification, sentiment analysis and cross-sell potential, Artefact enabled contact center agents to better respond to customers while reducing their workload.

In another use case, Artefact helped French insurer MAIF, when their call center was inundated with questions whose answers could be easily found on the MAIF website. Using Natural Language Processing (NLP) algorithms to analyze over four million calls, and topic modelling to categorize calls into different request types, Artefact help MAIF's teams identify which questions could be solved online and which needed a human response or presented a sales opportunity.

"Building AI software in the Artefact AI Factory is exciting: it's an evolving science at the intersection of several disciplines and is fostering the development of larger AI projects."

5 — Industrializing innovation in the Artefact AI Factory



Transformative AI technologies create new opportunities for growth, improve customer service, and create efficiencies for financial and insurance companies. But who's going to create these technologies, and how? AI development takes months or years, and requires very precise working conditions. Data scientists are great at creating models that represent and predict real-world data, but deployment is another story: a staggering 88% of ML models never reach production.

To respond to the difficulties of moving from concept to industrialization, we created the Artefact AI Factory. It meets all the conditions for scaling up and reducing time to development and can assist banks and insurance companies dealing with competition by building innovative AI software to meet new customer needs and taking them to market fast. We've been seeing returns on investment within 18 to 24 months, which is really impressive compared to traditional IT projects. We work by closing the gap between Proofs of Concept (POC) and production by applying our Machine Learning Operations (MLOps) methodology to all our projects. It delivers scalable AI models quickly and effectively, shortening system development lifecycle and providing continuous delivery with high software quality.

We think "product first" to help companies advance their AI assets smoothly to production while anticipating industrialization constraints and risks. Our MLOps model is based on a solid ecosystem, and we apply the same processes for every AI project we deliver, from POC to product deployment.

Data maturity: Climbing the scale in financial institutions

Athena Sharma,
Director and Global Financial Services Lead
Artefact

Chris Bannocks
Group Chief Data Officer
QBE Insurance

Data maturity – what it is, the role it plays in business today, and why it's a particular challenge for banks, insurance companies, brokerage houses, credit unions and other businesses in the financial services sector.

What do we mean when we talk about data maturity?

In broad terms, data maturity measures the level of an organisation's ability to create value from its data. To achieve a high level of data maturity, data must be deeply embedded throughout the organisation and fully integrated into its every decision and activity. Data maturity is a key factor in successful digital transformation. The higher the level of its data maturity, the more competitive advantages a company will have.

To measure data maturity, a maturity model is used to assess a company's various data and digital capabilities. "*There are many different maturity models available, but they all use a set of questions to rate a company's level of maturity across many different dimensions,*" explains Chris Bannocks. "*These include a company's capabilities – or maturity – in its architecture, people, analytics, ethics, privacy and so on. It's an objective measure that can be delivered either by independent review or self-assessment.*"

"*The result is an aggregation, so companies should decide where they want to acquire more data maturity and where less is acceptable. You don't have to be at the highest level in every dimension. You have to ask yourself, 'Maturity for what purpose?'*"

"To achieve a high level of data maturity, data must be deeply embedded throughout the organisation and fully integrated into its every decision and activity"

Why is data maturity especially important for financial institutions?

Because regulation is a crucial issue in the world of finance, the higher the level of a bank or insurance company's data maturity, the more likely data is to be well-controlled, well-managed, well-governed and secure. "*Data such as risk metrics, client statements or financial reporting may be deemed more accurate as a result of data maturity in those areas,*" says Chris.

But regulation isn't the only driver for data maturity. A data mature financial institution (FI) can leverage the full spectrum of solutions that big data analytics and AI have to offer, leading to better decision making, a more connected business and greater competitive advantage. For example, a more data mature insurance company would be better at using AI to determine personalised insurance risk for each customer, leading to better-tailored policies and significant business benefits. "*Leveraging AI and big data analytics for data driven decision making, enabling greater personalisation and optimising processes is a critical success factor for banks and insurance companies. But overcoming legacy data maturity challenges continues to be a primary blocker,*" says Athena.

Where do financial institutions currently stand on the data maturity scale?

Many FIs still approach data maturity as a zero sum game – either you are data mature or you are not. Instead, data maturity should be viewed as a spectrum with varying degrees of maturity being appropriate for different parts of the business, says Athena. "*This segmented view of maturity makes the data challenge more digestible for FIs, enabling them to target priority areas first and trading off maturity in areas that may not be as important for the business, or that may generate lower ROI.*"

According to Chris, on a data maturity scale of 1 to 5, most financial institutions sit somewhere between 3 and 4. "*FIs have made great strides because they've been hard at work in the maturity and capability space for about 15 years. But if you dive down and look at data quality or architecture or ethics, you'll see varying levels of maturity according to industry.*"

What challenges do financial institutions face in the data maturity journey?

The complexity and age of legacy architecture and systems remain one of the primary challenges faced by banking and insurance companies. Of the 100 global leading banks, 92 still rely on IBM mainframe for their operations. Using an outdated system means a lack of agility and architecture that isn't able to cope with growing workloads, especially when it comes to Big Data. "*As a result of these challenges, FIs need to either grow their processing capacity*

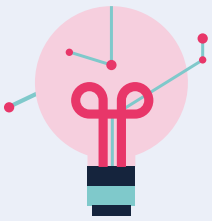
or rebuild their existing architecture, both of which are high effort and time intensive solutions. This is in sharp contrast to nimble FinTechs which can remain agile and customer centric from the get-go," says Athena.

The second key challenge relates to people and organisational complexity. Behavioural change and adoption are hard to achieve across industries, not just financial institutions. But the organisational complexity of FIs makes it harder to drive data accountability and ownership and deliver business value.

According to Chris, "*It may be easier for people in banking and insurance groups to change behaviour as they're used to regulated environments and can adopt things like data governance more easily than those in other industries might. But the size and complexity of the organisational structures themselves may present different demands from a data maturity perspective: a retail bank has a different data maturity need than an investment bank. The difference may be minor, but it requires that the response to each project be tailored to each case in terms of business value, not just regulatory value.*"

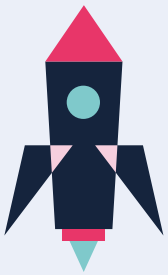
What steps can financial institutions take to become more data mature?

The broad answer is to accelerate towards analytics. Here are some steps FIs can take to achieve this:



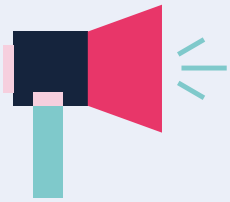
ASSESS THE FOUNDATIONS

FIs should start with a comprehensive audit of their data foundations, not just by assessing technical capabilities but by also asking the right business questions. "*In most cases, the existing foundations are good enough. Once financial institutions begin to put data maturity measures around them, they'll incrementally improve,*" says Chris.



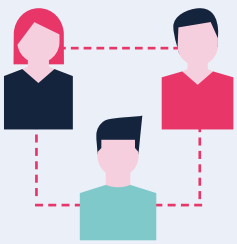
PILOT THE RIGHT USE CASES

FIs need to adopt an agile approach where they "*change by acting now, rather than acting after change has occurred,*" says Athena. In practice, this means selecting use cases that address specific business issues, outlining requirements to enable these use cases and changing data maturity parameters in line with these requirements. This makes change more manageable and effective.



REALISE VALUE EARLY TO MAINTAIN MOMENTUM

A key setback is that large scale transformation projects lose momentum over time and value realisation takes too long or sometimes doesn't take place at all. A use case-driven agile approach to data maturity enables early value realisation, and therefore, greater business buy-in.



KEEP PEOPLE AT THE HEART OF CHANGE

According to Chris, "*Data maturity has to include and understand the human component, as well as the consumption and use of data within the model, not just the way you manage it to its endpoint. To get real value from a data maturity model, all these components must be included.*"

Data-driven Marketing

The Great Data Reckoning How to truly understand Marketing ROI & make smarter decisions in a cookieless world

Bobby Gray
Head of Analytics & Data Marketing

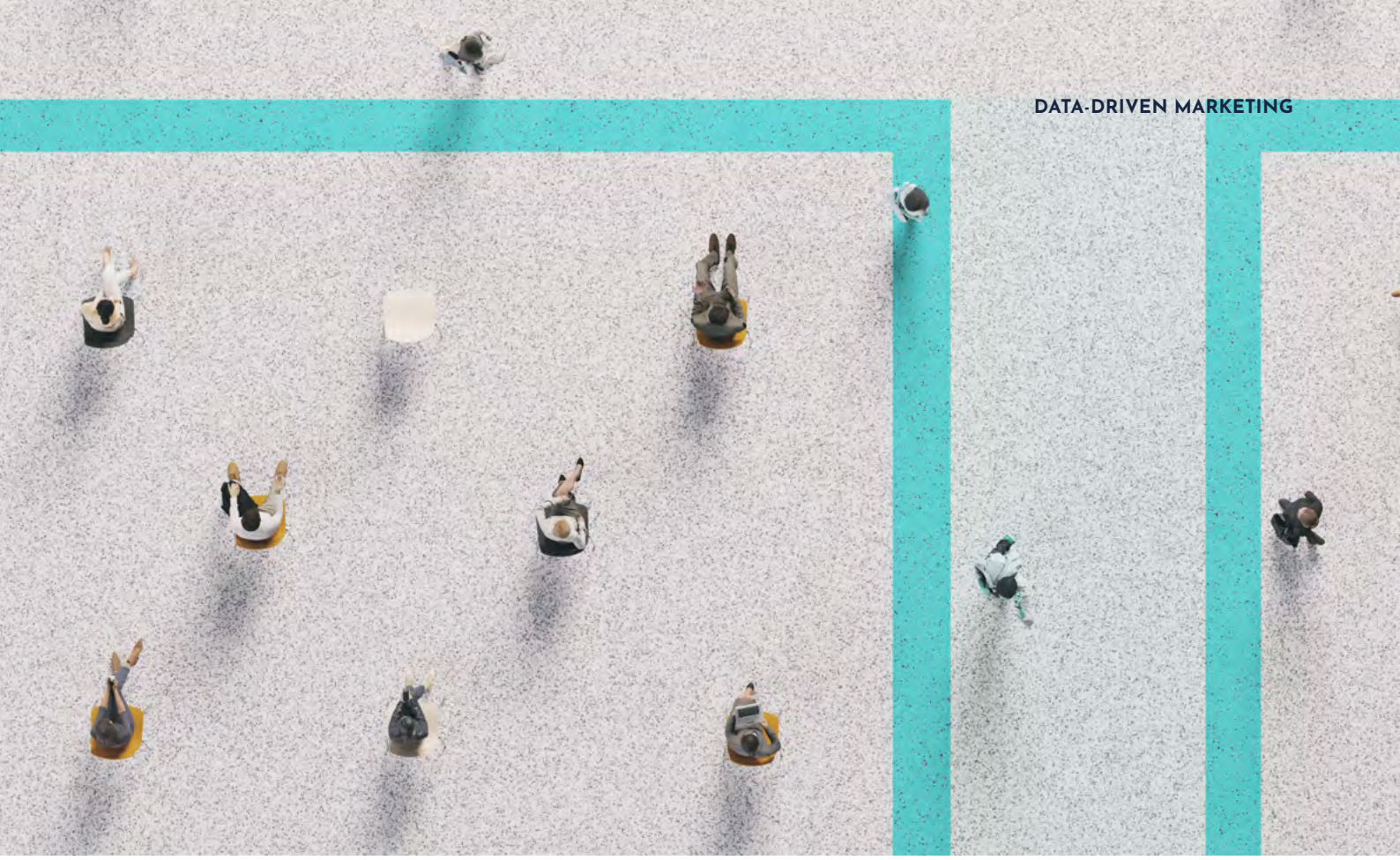
Aleksandra Semenenko
Data Science Senior Manager

The third-party cookie cull: a route to smarter marketing

Organisations are faced with increasingly fragmented advertising channels; since the rise of digital, it's no longer just a case of dividing budget across TV, print, outdoor and possibly radio. Online introduced many advantages, but also made the equation more complicated; the development of so many new addressable digital channels (social, VOD, digital radio, etc) has exacerbated that and the additional rise of activities such as influencers and online communities means that understanding the true impact of marketing is more nuanced than ever.

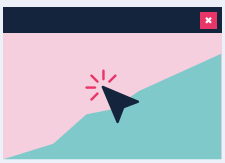
The net result is that it is increasingly complicated to decipher what is happening, both in terms of customer behaviour and the true impact of marketing activities on influencing these behaviours. With so many platforms each providing potentially conflicting views of marketing impact, marketers still do not have a 'single source of truth' to which everyone is aligned.

The impending loss of the third-party cookie exacerbates the marketing challenge still further. The individual customer journey, traditionally used to inform strategies and measure full marketing funnel performance, will be severely impacted. Since Google made its initial announcement in early 2020, various alternatives have been put on the drawing board (although there is a general feeling that progress slowed following the postponement of the deadline).



There are 4 main methodologies for understanding marketing ROI

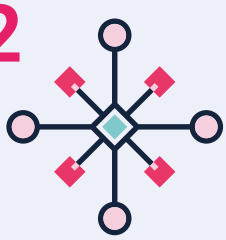
1



Last click / interaction

Simple to use but ignores all other activities throughout the full funnel.

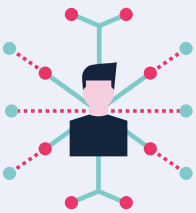
2



Multi Touch Attribution / Rules Based attribution (Individual customer journey based)

Defines a % of credit to each touchpoint in an individual customers journey, but will be severely impacted by cookie deprecation due to fragmentation of those individual customer journeys.

3



Marketing Mix Modelling (Correlation based)

Identifies a relationship between marketing variables and sales, but is used mainly for monthly / quarterly budget planning due to speed of insight.

4

MROI (Causality based)

Determines that individual marketing variables have a direct impact on business outcomes or on success of other marketing variables, allows for daily / weekly optimisation but requires high computational power.

As we know the challenges with last click and that Multi touch attribution has an uncertain future, we will focus on Correlation based analysis and Causality based approaches which will be the most robust solutions in a world post cookie deprecation.

Correlation analysis, such as Marketing Mix Modelling (MMM), which is not reliant on customer data but provides a relatively good view of what is - and isn't – working, has been a key element & has been around for a long time. However, it infers that any fluctuations in marketing delivery are directly correlated (either positively or negatively) with changes to success factors such as sales. This means that it is complex to provide clear explainability of performance uplifts, but rather we can find statistically relevant trends, this can in some cases lead to erroneous decision making as we can be influenced by our own biases.

A concrete example here is that there was an increase in Paid Social spend & traffic during the last 30 days as well as 25% off promotion on site, which led to a surge in sales. In correlation based analyses we can interpret that both of these two factors were likely to have a positive impact on sales but can not necessarily determine the impact of each accurately or the inter-relationship between the increase in paid social spend and 25% off promotion.

In contrast, causality based approaches, which also do not rely on understanding customer journeys at the individual customer level, are being touted as the future-ready solution. The true benefit of these solutions are that they provide clear explainability that A (marketing budget change) caused B (sales uplift) and the % of the uplift that can be directly attributed to A. This is extremely important when looking to make significant budget decisions on your marketing mix. Further to this Causality based approaches also allow marketers to understand the probabilistic impact that one marketing activity is also positively or negatively influencing another's impact on business outcomes.

To explain in a practical sense using the same example as before, causality based approaches are able to identify the % contribution to the increase in sales from Paid Social and the 25% off promotion independently. They are also able to provide a view that defines the impact of the 25% off promotion on paid social traffic and the impact that also has on sales. Clearly explaining which optimisations are the most important

Further to the above MMM studies are also not designed to provide near real time insights and are generally carried

out at a high level of granularity & thus is limited in terms of its ability to tackle the measurement challenges post cookie deprecation.

The causality based approaches are also designed to use real time demand signals to provide insights on the impact that every potential major business driver (media, competition, price, brand demand signals) has on key business outcomes (both digital and offline) at any given moment. Essentially it matches the decisions that businesses need to make with the measurements that they need to support those decisions.

To take another practical example, cookie data provides an individual consumer's journey across, for example, Facebook, followed by YouTube and then a search engine before their eventual conversion and purchase. Correlation, while it shows where they have visited, does not show the sequence – a major disadvantage in view of the different roles played by each channel at each stage of the buying decision process. Causality however enables the customer path to be re-built from a macro level – allowing a brand to see not just that advertising on Facebook works, but its interrelationships with other channels in the marketing mix (the impact on YouTube or PPC for example) and their impact on driving results.

Very simply, causality identifies an event as being the direct consequence of another; it provides the 'why' and 'so what' for outcomes of marketing activity. And understanding 'this is at the root of optimising future activity. Add to these benefits that causality is truly omni-channel, meaning it can marry online and offline activity, as well as cater for retailers without an online presence, and its potential to replace the third-party cookie seems assured.

But more than that, causality is the core principle behind MROI – or Marketing Return on Investment. MROI recognises that marketing measurement must be more than justification of advertising investment after the event. Instead it needs to provide the in-depth insight that enables each level of the decision-making process to be optimised so that it underpins everything, from analysing weekly campaign performance, through quarterly planning, to annual budget allocation.

MROI modelling is individual to each organisation, depending on their specific business objectives, budgets, data & business drivers but the bottom line is that it enables smarter decision-making.

So while the impending death of the cookie is – understandably – a current pain point for many organisations, it also has the potential to pave the way for next-generation marketing.

"In contrast, causality based approaches, which also do not rely on understanding customer journeys at the individual customer level, are being touted as the future-ready solution. The true benefit of these solutions are that they provide clear explainability that A (marketing budget change) caused B (sales uplift) and the % of the uplift that can be directly attributed to A."

CASE STUDY

ORANGE BANK

Enhances its digital marketing with rapid deployment of a full funnel, cross-device activation strategy.

CHALLENGES

Launched in November 2017, Orange Bank offers banking services designed natively around mobile uses so customers can autonomously perform all operations from their app.

The offer includes not only all basic banking services, but many innovative features and Premium offers as well, with a dedicated pack aimed at families, and a fintech designed for professionals. Orange Bank has a total of 1.2 million customers in Europe.

Orange Bank combines digital presence via mobile and the web with human contact through Orange boutiques. This model is an important differentiating factor compared to other neobanks. More than 2,000 qualified sales staff accompany customers in more than 300 Orange shops. Confirmation of account opening is subject to usage checks, which are carried out outside the application.

For this mobile bank, the marketing challenge was to enable subscription process management from start to finish on a mixed web and application path.

SOLUTION

Implementation

Working with Artefact, a GMP & GCP certified partner (Google Marketing Platform and Google Cloud Platform), Orange Bank achieved their steering objective in only one year, thanks to rapid deployment of the following projects:

Migration to Google Marketing Platform and Google Cloud Platform

- Redesign and enrichment of the data collected:
  - Redesign of the datalayer,
  - Definition of tracking plans, especially for compatibility with new Google Analytics,
  - Addition of variables and securing of reconciliation IDs.
- Consolidation of ad-centric and site-centric digital data,
- Consolidation of web and CRM data,
- Deployment of new Google Analytics to handle cross-device use-cases,
- Connection of media buying platforms to Google Analytics.

This deployment allowed Orange Bank to:

- Activate its app audiences in the web universe, thanks to Google Analytics 4 app and web audiences deduplication,
- Enable targeting of “similar users” with exclusion of existing customers,
- Improve bidding strategies,
- Revise steering KPIs to focus on high-value products,
- Refine attribution analyses,
- Profile prospects,
- Internalise SEA, the most profitable lever,
- And activate new marketing levers: affiliation, partnerships, and programmatic.

RESULTS

New Google Analytics and Google Analytics 4 properties

The new version of Google Analytics brings major innovations to marketers and analysts. It offers improved cross-device tracking capabilities, duplicating audiences between app and web environments, and has new features to automate certain activations.

New Google Analytics also includes machine learning capabilities to create audience segments based on user engagement, purchase intent, etc. — segments that can be targeted cross-device.

Attribution is also cross-device and cross-environment, allowing deduplicated reporting regardless of user device.

Data-driven marketing: the rise of the customer data platform

Florian Thiebaut
Partner – Data-Marketing Practice Lead

Everything seems to justify the current explosion of the Customer Data Platform (CDP) market. CDPs' main advantage over older generation Data Management Platforms (DMPs) is that they easily integrate identifiable first-party data (email, phone number) and aren't dependent on using third-party cookies or browsing data to refine customer and prospect knowledge.

CDPs are a true asset in a world that is becoming increasingly cookie- and ad ID-free. At a time when the pandemic is forcing brands to digitise at breakneck speed, and when the transformation of the technical and regulatory environment surrounding advertising trackers is forcing data marketers to revise their approaches, CDPs are here to optimise the customer experience.

A game-changing technical and legal environment

Following Safari's lead in 2016, the world's three main browsers eliminated (or will eliminate) the use of third-party cookies. On the mobile/tablet devices side, Apple's iOS 14 now requires explicit consent for any mobile ID collection.

As for regulation, GDPR laws in Europe have given consumers more control over their personal data, requiring them to give explicit consent for the use of cookies. This regulation represents a major shift in the world of data-driven marketing, as it has reduced the number of cookies placed on European devices by 30%.

This global trend restricting the use of IDs and advertising cookies sharply impacts the targeting capabilities of advertisers, who are often dependent on third party data. The vast majority of them use or have used retargeting and old generation DMPs that rely heavily on segments fed by third party data.

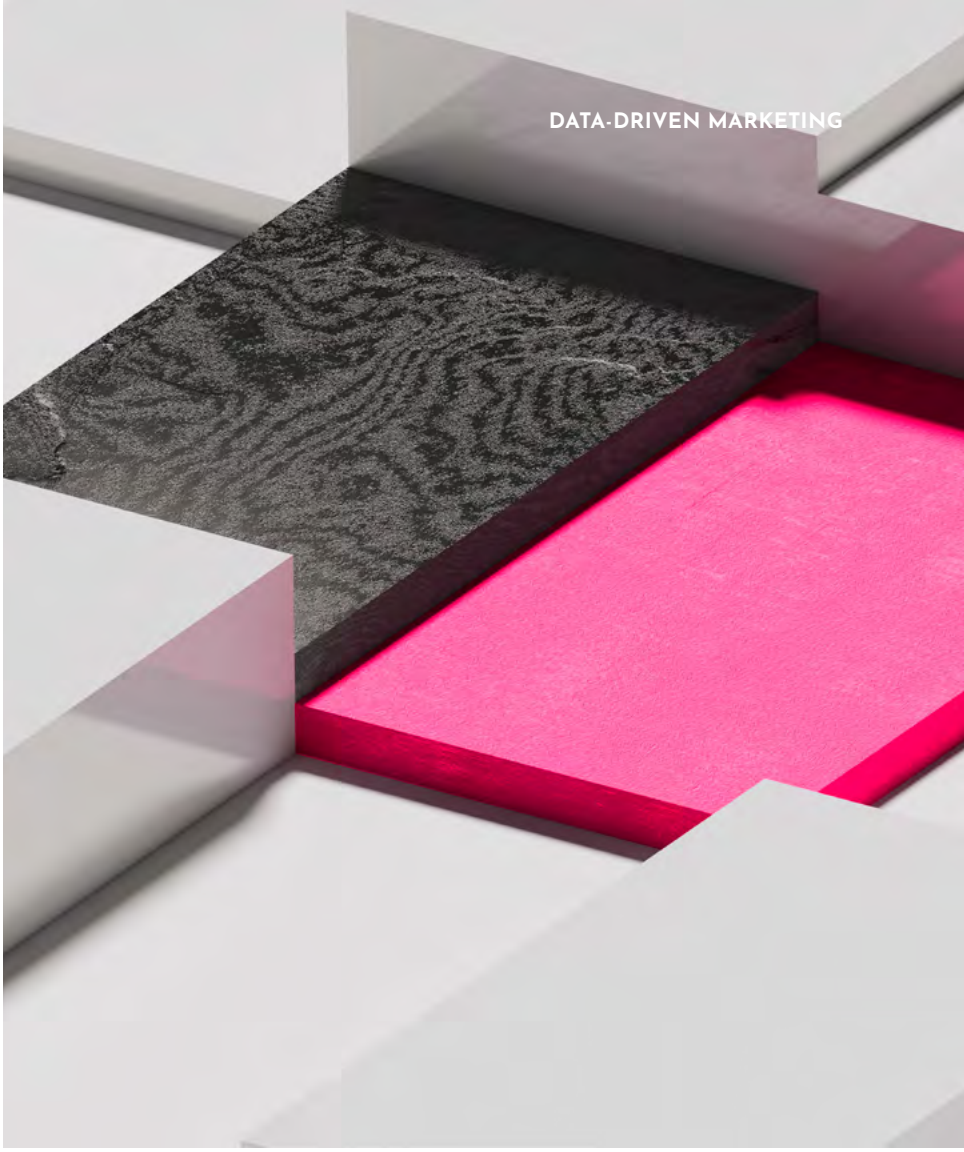
Along with targeting, measurement must also be transformed. With more stringent consent collection requirements, it's more difficult to collect the consumer IDs needed to track impressions, clicks or views, and reconstruct complete customer journeys.

Four pillars for a sustainable data strategy

To maintain the same performance and differentiate themselves from the competition, advertisers must design a sustainable data strategy and exploit their customer and prospect data to its full potential.

This requires focus on four actions:

- **The CDP**: The first step is to establish a CDP environment based on a suite of tools that is both compliant and sustainable. This will enable data to be collected, stored, processed, visualised and activated, whatever the source. From this foundation, the focus must be on first-party data.
- **Data governance**: Brands need to rethink data governance and processes to enable secure and compliant end-to-end data collection.
- **Audience segmentation**: This data, centralised for a unified view of the



consumer, can then be used to create new audience segments and define new metrics for measuring campaign results.

- **Second-party partnerships**: In addition, it's becoming increasingly strategic to form so-called "second party" partnerships with other partner companies to exploit first-party data and create win-win situations.

This data completes a database that is incomplete at certain points in the consumer journey. Examples might be an agreement between an FMCG brand and a retailer, a mobile phone manufacturer with a telco or a hotel chain with an airline.

Three types of data to activate via a suite of tools

First- and second-party data are key to meeting the challenges of the post-cookie world. But what are they and what tools can be used to manage them?

- **PII or Personally Identifiable Information** is essentially CRM (customer relationship management) data. It can precisely identify an individual and is often an email address or a phone number for example. Once anonymised, it can be used via the APIs of media partners (e.g., Google Customer Match, Facebook Custom Audience/ conversion API, Amazon, WeChat, etc.) to build audience segments, perform audience extensions, and reconstruct paths to measure the influence of digital campaigns on offline sales, etc.

- **Non-PII data** can be browsing data that cannot lead directly to the identification of an individual. It can be used to build more granular segments via analytics and audience creation solutions for measuring precision marketing actions without relying on third party data

- **Data that is purely media-related**, such as campaign impressions, video views and click rates, is more voluminous and less granular than the other two types of data. It is more difficult to use but there is a robust market of tools capable of treating it in a secure and compliant manner, such as Google Ads Data Hub, Facebook Advanced Analytics and Amazon Marketing Cloud.

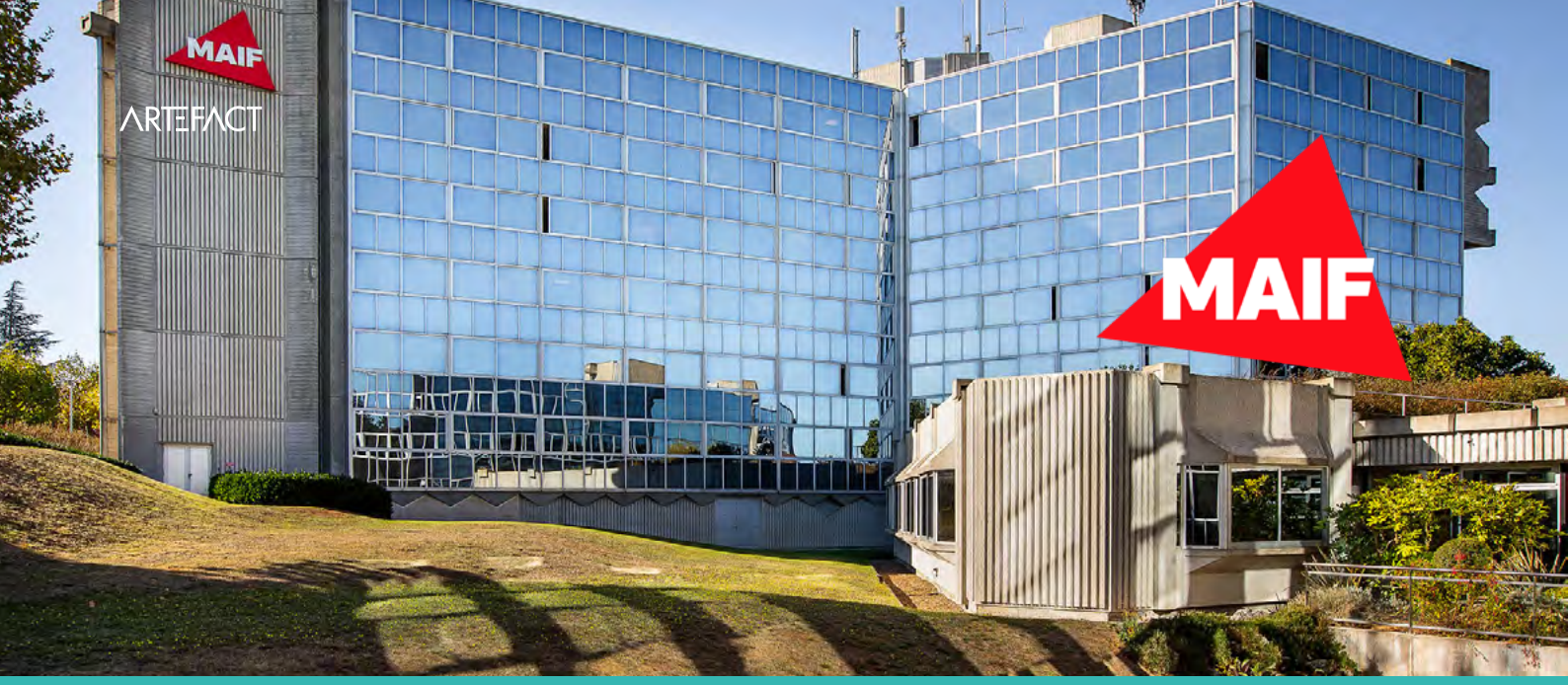
These different data flows are injected into an ecosystem of interconnected tools, which are useful for a range of tasks — from data collection to performance measurement of the actions carried out — and can be activated on all channels, whether media, direct marketing or site personalisation.

This entire ecosystem, the result of all the connections built between the different tools already used by the company (also known as "full-stack" solutions), is what is called the CDP.

When it comes to the adoption of this way of working, the numbers don't lie. Fundraising for CDP providers is soaring, the tech giants are all positioned in the sector, and the number of users is exploding.

In fact, according to the Customer Data Institute, the market increased 30% from $1 billion in 2019 to $1.3 billion last year. Estimates see this figure reaching $1.55 billion in 2021 as conditions are even more favourable for the adoption of CDPs.

As the data-driven world continues to evolve at a rapid pace, there seems little doubt in the business value of the CDP. Now is the time for organisations to consider deploying this future-facing technology.



CASE STUDY

MAIF

Segmentation Marketing -Artificial Intelligence to Enhance Insurance Customer Understanding

MAIF is a leading French mutual insurance company, active for more than 80 years. It has no share capital nor shareholders and works solely for its three million customers, to whom quality protection and perennial services are guaranteed. Created around the needs of education professionals, it is today open to all.

CHALLENGES

A major data transformation project for MAIF

This segmentation marketing project, part of the Artefact-MAIF collaboration, is a phase of the mutual insurance company's transformation as outlined within its new development strategy. MAIF wanted to evolve its development strategy to adapt to the current market reality as well as to the evolution of its portfolio.

"We needed an expert partner, good at clustering [data grouping or partitioning] to build the segmentation, and in marketing strategy to align this data project with our strategic context and its activation, and in the insurance sector to understand the transformation challenges in a market which is forever evolving. Artefact seemed like the ideal partner to fulfill these three roles."

Nathalie MACON
Marketing manager, MAIF

SOLUTION

Better understand clients to put them at the heart of the development strategy

The project's first objective was to understand the reality of the current customer portfolio as well as to anticipate its future evolution so MAIF could continue to talk to its clients according to their needs, keeping them at the heart of their development strategy.

Segmentation based on data and observation of MAIF client behaviour

MAIF's first goal in building the segmentation was to make use of its rich existing internal data pool. This foundation would then be used to create a segmentation based on the real, established behaviours of its clients, not on ad-hoc marketing studies that can't always capture the reality of the portfolio.

RESULTS

Two-level segmentation to respond to strategic and operational challenges

- **A first level comprised of 5 robust and perennial client groups** (Value Customers, Opportunists…) to answer the strategic challenges of MAIF's development plan
- **A second level of 10 sub-groups** which can be adjusted according to portfolio evolution and operational needs

ARTEFACT

Building your own Audience Engine: Turning the Google ecosystem into a cookieless Customer Data Platform (CDP)

Pascal Coggia
Managing Partner UK

Manuela Mesa
Managing Consultant UK

Natacha Zouein
Senior Consultant UK

The race for more privacy, yet greater personalization

The dichotomy between the increasing awareness and concerns about customers' data privacy and the rising need for personalized content and ads is forcing companies to rethink their personalization and segmentation strategies.

Research provides strong arguments for personalization (with 74% of customers feeling frustrated when website content is not personalized for example, according to Instapage). This is mirrored by the feedback we receive from clients:

- More than half, across a variety of industries reported sales uplifts of more than 20% in 2020 as a result of personalization;
- Hypertargeting in online media investments represents around $20bn of additional revenue for consumer packaged goods (CPG) companies;
- The cost of media can be reduced by 5-15% on Facebook and 20-25% on programmatic display.
- Search impression share can be increased across all keywords by around 15% while retaining the same budget.

First-party data ownership is becoming an increasingly significant source of competitive advantage

This situation puts companies under pressure; they have no choice but to enrich their data assets and focus on building a strong first-party data strategy and customer database.

Our data shows us that global players have realized that:

- Category leaders have set objectives requiring them to collect first-party data for millions - and in some cases billions - of customers;
- CPG brands have digital assets that have seen an increase in traffic of more than 20% in 2021, as a result of which a third of their global customer base engaged online with the brands;
- Companies should be collecting at least 10% of first-party data from their most valuable customers and deploying collection tactics as a priority.

Companies need to equip themselves with tools dedicated to the collection and processing of first-party data in compliance with the GDPR regulations, as well as strengthen their audience segmentation capabilities. *But how will the imminent removal of third-party cookies impact user identification?*

CDP platforms are flourishing as off-the-shelf solutions for companies to quickly lift and shift their first-party data maturity

The move towards a cookieless world should be seen as an opportunity for new technologies to emerge. To consolidate and properly manage that data, as well as drive holistic insights and make them actionable across marketing channels, Customer Data Platforms (CDPs) are regarded as the most attractive solution. By using this data-driven technology, marketers can leverage first-party data to extract insights that can help them deliver one-to-one personalization and thus enrich the customer experience.

"The dichotomy between the increasing awareness and concerns about customers' data privacy and the rising need for personalized content and ads is forcing companies to rethink their personalization and segmentation strategies."

- First-party data is the first choice among marketers for customer insight and has the greatest impact on customer lifetime value (CLV);
- By 2025, 20% of direct-to-consumer revenue will come from recurring customer relationships, leading to a tenfold increase in first-party data collection.

Limitations of relying solely on IDs, would they be cookies, emails, IPs or other forms of personal identification

CDPs have traditionally been used to consolidate customers' data through a combination of IDs, but IDs generally raise difficulties:

- Loss of an average 32% in ID matching when sharing them media buying platform, and it ranges from 0 to 95% depending on many factors like browser, OS, user consent, browser settings and add-ons, the type of IDs and the limitations of the media buying platform itself. We also expect this percentage to worsen as regulations become stricter, wall gardens increase privacy rules and browsers add new technical barriers to ID collection and matching.
- We estimate that only 57% of media reach can be targeted with IDs and look-alikes with meaningful accuracy, which leaves an immense opportunity and need for non ID based targeting to extend reach with high relevance.
- Fast growing media formats like YouTube trueview and Bumper, Pinterest,

Without the right tools to translate this first-party data into media and marketing strategies, CDPs can be too much of a blank canvas. Indeed, adopting a CDP approach is only one piece of the puzzle; it is crucial to go beyond IDs and replace more data with more intelligence.

New, rising, and yet big opportunity for brands to leverage first-party data without using names and cookies.

Artefact and cloud providers such as Google help brands to navigate a world without cookies - and Personally Identifiable Information (PII)

The Audience Engine, one of Artefact's bespoke tools for clients, is a CDP that powers personalized digital marketing campaigns by building, sharing, measuring and optimizing audiences from first-party, second-party and third-party consumer data. It was created based on attributes, not IDs, with the aim of building a scalable and AI-driven approach to first-party data. This entailed adding a layer of intelligence on top of the client's data to ensure an enhanced gathering of insights and learnings about their customers.

Google has been promoting this vision over the past few years as it moves away from cookie-based third-party data and behavioral targeting towards the ability to work with emerging aggregates; initial proposals were based on its Federated Learning of Cohorts (FLoC), with this being replaced in early 2022 by Topics, a new proposal for interest-based targeting. *What does this mean for Artefact clients?*

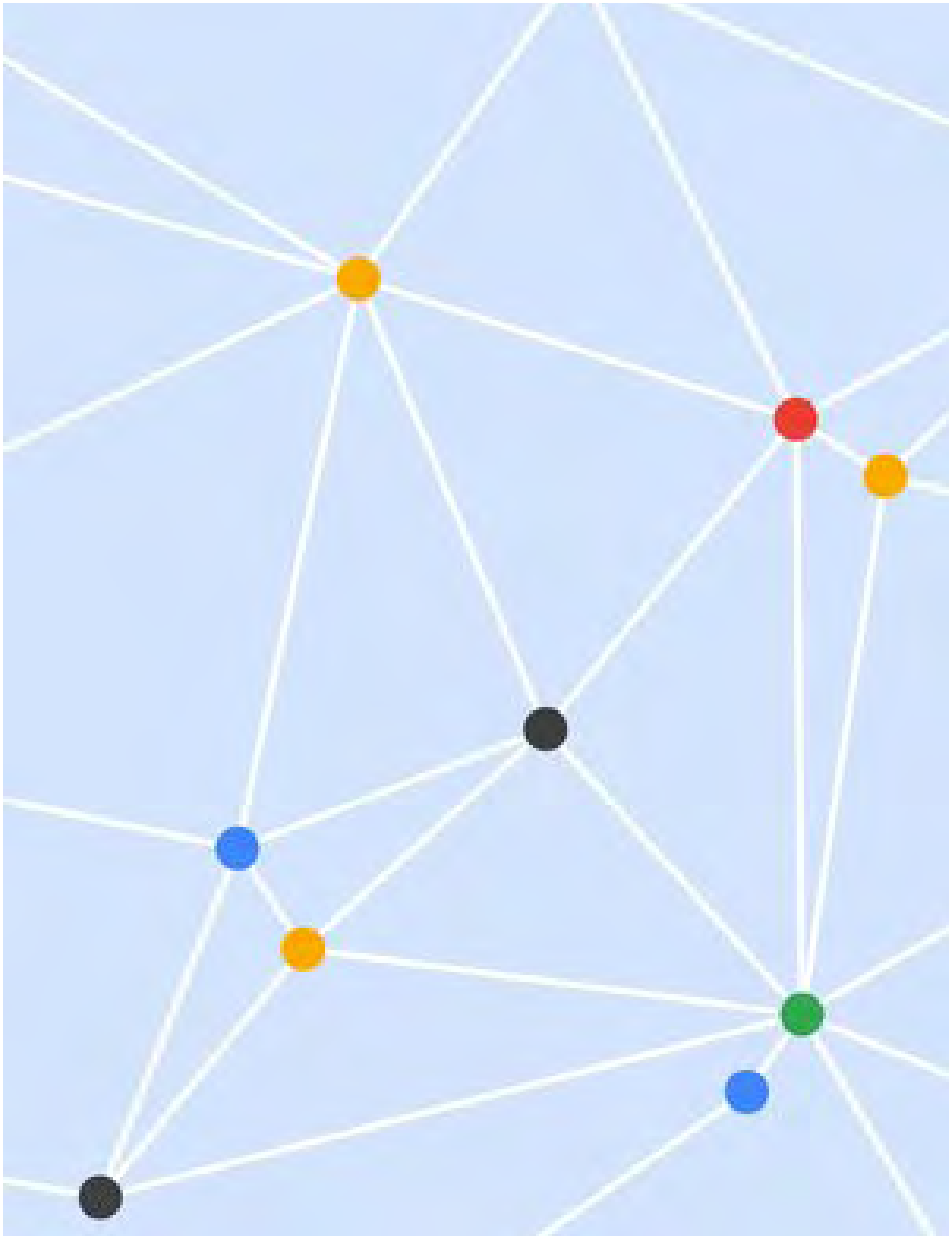
A cookieless and PII-less approach offers large advertisers a new gold mine, delivering high levels of performance with a greater reach

The deployment of cookieless targeting across various of our clients is showing that it can be as efficient as ID targeting, if not more so. With enough data, which is gathered from a combination of sources including market and affinity segments, keywords, online behavior patterns, location and consumption reports, we can create audience segments with the equivalent personalization that meet clients' specific business needs. This type of targeting is delivering effective reach, offering precise and customized personalization on a larger scale.

Data science is only one part of data strategy: companies need to switch from data-driven rules to business-driven machine learning

Personalization is about much more than having a customer's name and email; what the organization understands about them is also key. Rather than simply focusing on identifiable data, it is essential to use machine learning to add more precise insight to the data collected. CDPs and solutions like the Audience Engine enhance the collection and interpretation of data, and are showing how the switch from data to intelligence can help companies' personalization and data privacy strategies reach their full potential.

"The deployment of cookieless targeting across various of our clients is showing that it can be as efficient as ID targeting, if not more so."



CASE STUDY

NEXITY

Ban the Banner. A data-driven advertising campaign to tap into property customers' competitive nature.

CHALLENGES

The real estate market is highly competitive. Great property deals disappear almost instantly, meaning most customers don't even get a chance to bid on their dream property.

As the leading property developer in France, Nexity wanted people to know that they have the best range of property auctions on the market and great deals for everyone.

Our goal was to increase Nexity's website views and registrations for its online property auctions.

SOLUTION

Introducing: 'Ban The Banner' — a campaign which tapped into property customers' competitive nature.

Using an API technology, we served our targets banner ads with offers for properties we thought they'd like.

To increase competitiveness and exclusivity, we gave the first person to click the chance to 'ban the banner' — blocking anyone else from viewing the offer for 30 seconds (until they'd registered their interest). All they had to do was close the ad — like they would with a normal banner ad.

When an offer had been 'banned', we updated the banner ad copy so other people would know that they were too late and had been blocked from viewing this great deal (but that they could unlock deals of their own at Nexity.fr).

RESULTS

Customers spent an average of 13 hours a day blocking ads from other people.

Engagement rate was 1.8% above the average of traditional advertising banners.

These blocked ads piqued curiosity — prompting 15,000 people to visit our site to see what was being hidden.

While they were there, more than 300 registered to receive their own banner offers — growing Netixy's prospect list.

ARTEFACT

AI for Call Centre

AI Requires a Holistic Framework and Scalable Projects

Ghadi Hobeika
CEO Artefact US

Artificial intelligence and digital transformation projects have a low success rate, but best practices help.AI has the potential to drive change in almost every industry. In other words, there are big incentives for organizations to start their AI journey now; there is also the risk that if they don't, playing catch up will be difficult, if not impossible, in a discipline that will become increasingly critical the more widely it is adopted. So, it's no surprise that AI is causing so much interest and excitement. However, a lot of AI projects fail.

Ever since I can remember, artificial intelligence has been the holy grail. Films have portrayed it, from BladeRunner to the more recent Her. In the meantime, business leaders promised it would revolutionize the workplace. In both cases, we've been presented with scenarios in which AI transforms the daily grind.

Indeed, AI has been talked about as a scientific discipline since 1956. And although the math-based fundamentals have existed for more than 70 years, the computing power required has only recently been a reality, with the cloud being the ultimate AI catalyst.

Significant progress has been made — and the sector is no longer in its infancy. According to McKinsey's The state of AI in 2020 survey, 50% of respondents said their companies had adopted AI in at least one business function.

AI has the potential to drive change in almost every industry. In other words, there are big incentives for organizations to start their AI journey now; there is also the risk that if they don't, playing catch up will be difficult, if not impossible, in a discipline that will become increasingly critical the more widely it is adopted. So, it's no surprise that AI is causing so much interest and excitement.

However, a lot of AI projects fail.

POCs Should Be Designed for Long-Term Success

Many proofs of concepts (POCs) are not designed to scale. They do no more than prove that something can be done. Then, they are left to fester because it wasn't determined in advance whether the concept in question was relevant and required by the entire organization, or whether an enterprise-wide roll-out was technically feasible.

Moreover, the cost structure of projects of this nature: Getting to this point is likely to have devoured 70% of the overall budget, without the result ever seeing the light of day. That is bad business on every level. So, what is the alternative?

In short, scale must be an integral part of the POC, and reflected in the metrics that determine if it was successful.

There are some straightforward tactics for achieving this. A good option is running the POC in two regions and requiring both streams to deliver on pre-determined goals before it can be signed off as a success. It is also important to identify parallels and variations between the projects. This approach develops process and structure as part of the initial venture, and it underpins adoption in the wider environment if the project moves ahead.

Skillsets Demolish Silos

Organizational silos, rooted in the traditional business structure, are still commonplace. They are a constant thorn in the side of smooth-running operations, and they can be the death knell for scalable AI implementations. Addressing this means building the right skills into every part of the project.

We need mathematical expertise, IT skills and a coding specialist delivered (respectively) by data scientists, solutions architects, and machine learning (ML) engineers. The business perspective, provided by product managers and owners, is also an essential part of the mix. This multidisciplinary team should have an open and collaborative way of working, with good communication channels and a deep level of trust throughout the lifecycle of the project so they can collectively lay the groundwork, roll out the implementation and, finally, train the people that will run the application on a day-to-day basis once the POC is completed.

Technology Is Important, Too

Cloud computing has made AI projects a reality for many businesses. It does away with the need for big, costly IT implementations, relying instead on agile tools and technologies that are customizable and available on an on-demand basis.

As with the hybrid-team approach, the tech toolbox should comprise the applications and software specific to the project in question. And it goes without saying that it should be scalable.

The AI Risk Paradox

AI presents organizations with a dilemma: Implemented badly, it is likely to fail, creating business risk. However, not implementing AI at all risks falling behind more future-facing competitors as they reap the rewards of exploring this next-generation technology.

The key is to view any AI project in terms of its role in the long-term direction and success of the overall enterprise and its operations. This approach will inform the technical and people-based framework that is essential for successful implementation and a holistic AI vision.

Powering your call centre with artificial intelligence

Matthieu Myszak
Data Consulting Director

In today's competitive business environment, customer experience is becoming a key differentiator for organizations well aware that it is more cost-effective to keep an existing customer than acquiring a new one and that a disgruntled customer well-handled can be turned into an advocate for your brand.

The recent advancements in technology and artificial intelligence can be applied to your own customer interactions and be a tremendous support to help your organization profit from productivity gains, improve customer retention and create additional revenue.

Having your customers talk to a robot can seem out of a scifi novel. But actually, it is already a reality for hundreds of millions of customers that have regular interactions with interfaces powered by artificial intelligence.

Ultimately, the goal of an optimal integration is to have customers talk like they usually do when interacting with a conversational agent that can analyse their sentiment, provide useful information, and answer recurrent standard requests as well as complex issues. Also, the virtual robot can pass on the caller to a human agent when needed. Beyond this conversational agent, Machine Learning technologies record the interaction for further improvement in setting new and sophisticated protocols of problem solving.

1 — The main advantages of using artificial intelligence for call centres

What differentiates customer service boils down to the quality of the relationships and the overall user experience. In that regard, AI can be a useful tool helping organizations achieve unmet customer expectations.

An artificial intelligence-powered customer centre is a crucial asset for three reasons:

- Reduce operating costs (via automatisation, reduction of average handle time)
- Improve the quality of the service and the customer satisfaction (via an increased reactivity and availability)
- Offer opportunities for cross-selling and upselling customers

The revolution of conversational technologies has already begun. Today, internet users are becoming more and more familiar with voice commands and chatbot interactions thanks to natural language understanding capabilities (NLP) wired into mainstream digital products, such as Google Voice Search, Google Assistant or Google Home.

According to a recent Gartner study, by 2023, 70% of consumers will prefer to interact with a vocal interface than a real person and 40% of all customer interaction will be fully automated.

2 — How does it work

Google Cloud has been developing virtual assistants capabilities for years and has created a product that can be used for business purposes.

Artificial intelligence is not always meant to replace humans as it can be utilized to augment real agents. Thanks to the robot, call centre operators can concentrate on complex and higher value situations and be freed from small, repetitive and low-value tasks. The chatbot is also critical in providing assistance to the operator. In certain situations, maintaining a human contact is key.

With the support of artificial intelligence, the customer service agent becomes an "augmented agent", meaning that the virtual assistant listens to calls in real-time and provides contextual assistance letting the human stay focused on the conversation and expressing empathy towards the customer.

3 — Concrete business cases showing tangible benefits in improving customer satisfaction

Businesses of all sorts could benefit on both sides from a positive impact by powering their customer service with artificial intelligence:

FOR CALL SERVICE OPERATORS

Improve efficiency and focus, reduce churn, provide opportunities to upsell and cross-sell

FOR CUSTOMERS

Improve user experience with a customer service available 24/7 with no waiting time, accelerate retention and customer loyalty

"If a customer asks for the status of his order, the virtual assistant has to provide the right request and give the correct information."

4 — Seamless integration into your legacy system

For optimal performance, the Google Contact Centre AI must be integrated into the call centre workflow, work with the existing databases and documentation (via APIs) and the front desk interfaces.

Organizations need to bring a multidisciplinary team to achieve this project according to their needs and their own IT architecture.

Before being set up, a chatbot needs to be fed with customer interaction data. The bot needs to be trained by listening and analyzing past customer interactions. That will enable the virtual assistant to be able to provide value immediately with high levels of customer satisfaction. Existing data can be emails, chat messages or voice calls. The data will help train the model according to the customer journey and the expected optimisations.

"If we have to manage use cases of a client from the banking industry, we will not train the AI tool the same way that if it was for an ecommerce brand, for example."

It can take from one up to three months to integrate the artificial intelligence solution into an existing customer service depending on the number and the complexity of use cases, and number of integration points to access.

"Deploying a virtual agent for an insurance company that can automatically file a damage claim is more complex for example than if you are asking for the status of your order on a ecommerce website."

In the event that a business doesn't have any data to analyze or precise use cases to aim for, it is possible to implement a working solution by asking each caller a prompt such as "Could you please tell us the reason for your call ?" and then letting them access the traditional customer experience journey. By analyzing the initial answer and the human agents' interactions, the artificial intelligence will get trained to quality future interactions.

5 — Why rely on a partner

Before taking the task of implementing a virtual assistant solution into your architecture, it could be useful to bring in an experienced partner that could assist you in the different steps of the project and help you maximise value.

Artefact has been helping clients, in various industries, turbocharge their call centres with artificial intelligence. We provide assistance in different ways:

- Identification and prioritization of use cases
- Setup and training of artificial intelligence solutions
- Development of integrations to collect relevant data

Our company has extensive experience working with both partners and service providers from the digital, data and artificial intelligence industries. Artefact's method is centreed on featured teams, composed of members with complementary skills, from business analysts to data scientists and software engineers, that can help projects come to life.

"We don't usually think of adding a UX Designer to feature teams but this role is important, as it helps give a personality to the virtual agent that reveals your unique brand image"

Our company has extensive experience working with both partners and service providers from the digital, data and artificial intelligence industries. Artefact's method is centered on featured teams, composed of members with complementary skills, from business analysts to data scientists and software engineers, that can help projects come to life.

Setting up a bot that is both well-embedded into your data architecture and provides usefulness to consumers is a goal that can take some effort but can be reached when organizations make it a customer experience priority.

Conclusion

An artificial intelligence, such as Google Contact Centre AI, integrated into the data architecture of your customer service will supercharge your customer experience.

The machine learning capabilities and recordings of interactions provide a constant feedback loop that helps the performance of the virtual assistant and the use cases that can be addressed.

The right partner for an artificial intelligence project can help organizations smooth out the definition and implementation phases and achieve immediate performance and business gains, while providing useful industry benchmarks and key learnings from previous experiences.

CASE STUDY

MAIF

Using Topic Modelling to reduce contact centre bottlenecks.

CHALLENGES

MAIF is one of France's largest home and automotive insurance companies, with more than **3 million members.**

One of the challenges facing its customer services team was managing the volume of calls coming into its call centre — on average, **some 8 million a year.**

With no way of vetting calls before they reached an operator, the team was wasting precious time responding to questions customers easily find the answers to on the MAIF website.

To improve efficiencies, we needed to filter out unnecessary calls and free up more time for MAIF's customer service teams to process more complicated requests.

SOLUTION

To understand why customers were calling MAIF's call centre, we developed **Natural Language Processing** (NLP) algorithms to analyse transcripts of more than 4 million calls.

We then used **topic modelling to categorise every call** into one of 35 different request types.

We liaised with MAIF's business teams to identify which questions could be solved online and which needed a human response or presented a sales opportunity.

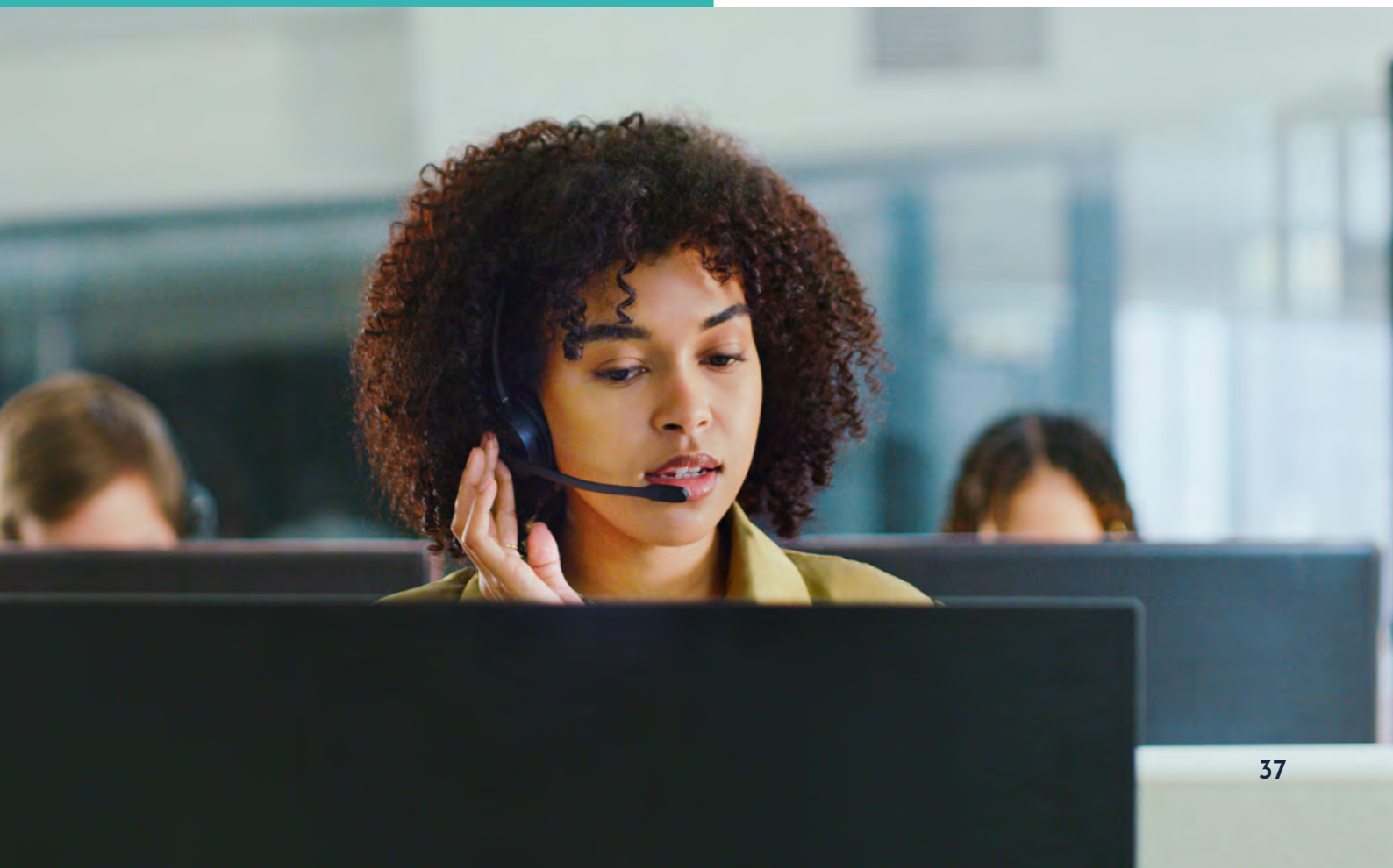
Where calls did not represent an opportunity, we advised how to answer these questions **online.**

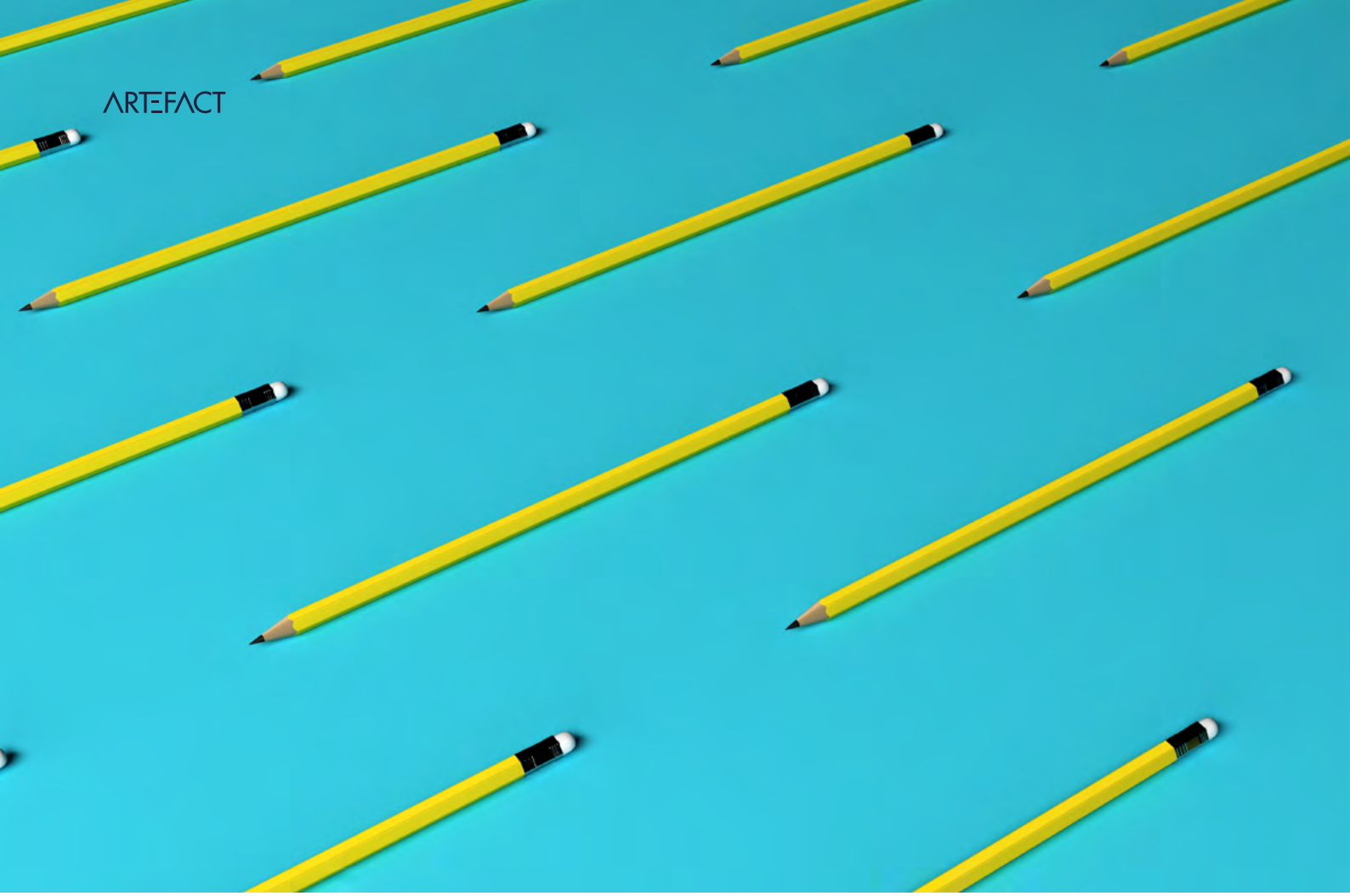
RESULTS

Our analysis showed that 32% of inbound calls were 'low added value requests' — questions that could easily be answered online.

As a result, we built a roadmap advising MAIF how to solve these questions online and direct people to this content to avoid calling.

Digitising these queries has let MAIF's customer services team prioritise cases that require a human touch, improving efficiencies and its round-the-clock service.





Introducing NLPretext, a unified framework to facilitate text preprocessing.

Amale El Hamri
Senior Data Scientist

Working on NLP projects? Tired of always looking for the same silly preprocessing functions on the web, such as removing accents from French posts? Tired of spending hours on Regex to efficiently extract email addresses from a corpus? Amale El Hamri will show you how NLPretext got you covered!

NLPRETEXT OVERVIEW

NLPretext is composed of 4 modules: basic, social, token and augmentation.

Each of them includes different functions to handle the most important text preprocessing tasks.

TEXT AUGMENTATION

The augmentation module helps you to generate new texts based on your given examples by modifying some words in the initial ones and to keep associated entities unchanged, if any, in the case of NER tasks. If you want words other than entities to remain unchanged, you can specify it within the stopwords argument. Modifications depend on the chosen method, the ones currently supported by the module are substitutions with synonyms using Wordnet or BERT from the nlpaug library.

BASIC PREPROCESSING

The basic module is a catalogue of transversal functions that can be used in any use case. They allow you to handle:

- Bad whitespaces in a text, end of line characters
- Encoding issues
- Special characters such as currency symbols, numbers, punctuation marks, latin and non-latin characters
- Emails and phone numbers

```
from nlpretext.basic.preprocess import replace_emails
example = « I have forwarded this email to obama@
whitehouse.gov »
example = replace_emails(example, replace_with=
»*EMAIL* »)
print(example)
# « I have forwarded this email to *EMAIL* »
```

SOCIAL PREPROCESSING

The social module is a catalogue of handy functions that can be useful when processing social data, such as:

- hashtags extraction/removal
- emojis extraction/removal
- mentions extraction/removal
- html tags cleaning

```
from nlpretext.social.preprocess import extract_emojis
example = « I take care of my skin ☺ »
example = extract_emojis(example)
print(example) #[':grinning_face:']
```

CREATE YOUR END TO END PIPELINE

— DEFAULT PIPELINE

Our library provides a Preprocessor object to efficiently pipe all preprocessing operations.

If you need to keep all elements of your text and perform minimum cleaning, use the default pipeline. It normalizes whitespaces and removes newlines characters, fixes unicode problems and removes recurrent artifacts from social data such as mentions, hashtags and HTML tags.

```
from nlpretext import Preprocessor
text = « I just got the best dinner in my life @
latourdargent !!! I recommend ☺ #food #paris \n »
preprocessor = Preprocessor()
text = preprocessor.run(text) print(text)
# « I just got the best dinner in my life !!! I
recommend »
```

— CUSTOM PIPELINE

If you have a clear idea of what preprocessing functions you want to pipe in your preprocessing pipeline, you can add them in your own Preprocessor.

```
from nlpretext import Preprocessor
from nlpretext.basic.preprocess import (normalize_
whitespace, remove_punct, remove_eol_characters,
remove_stopwords, lower_text)
from nlpretext.social.preprocess import remove_
mentions, remove_hashtag, remove_emoji
text = « I just got the best dinner in my life @
latourdargent !!! I recommend ☺ #food #paris \n »
preprocessor = Preprocessor()
preprocessor.pipe(lower_text)
preprocessor.pipe(remove_mentions)
preprocessor.pipe(remove_hashtag)
preprocessor.pipe(remove_emoji)
preprocessor.pipe(remove_eol_characters)
preprocessor.pipe(remove_stopwords, args={'lang': 'en'})
preprocessor.pipe(remove_punct)
preprocessor.pipe(normalize_whitespace)
text = preprocessor.run(text) print(text) # « dinner
life recommend »
```

NLPRETEXT INSTALLATION

To install the library please run pip install nlpretext

Natural Language Understanding (NLU) benchmark for intent detection and named-entity recognition in call centre conversations.

Rafaelle Aygalenq
Senior Data Scientist

Call centre advisors are starting to see NLU emerging in their day to day lives, helping them answering customers' requests more easily. For a tool to do that, it must be able to recognise at the same time the customer request and its characteristics, in other words, intent and named-entities. A lot of frameworks have been created to perform those two tasks and this article aims at describing them and presenting their baseline performances on a real project.

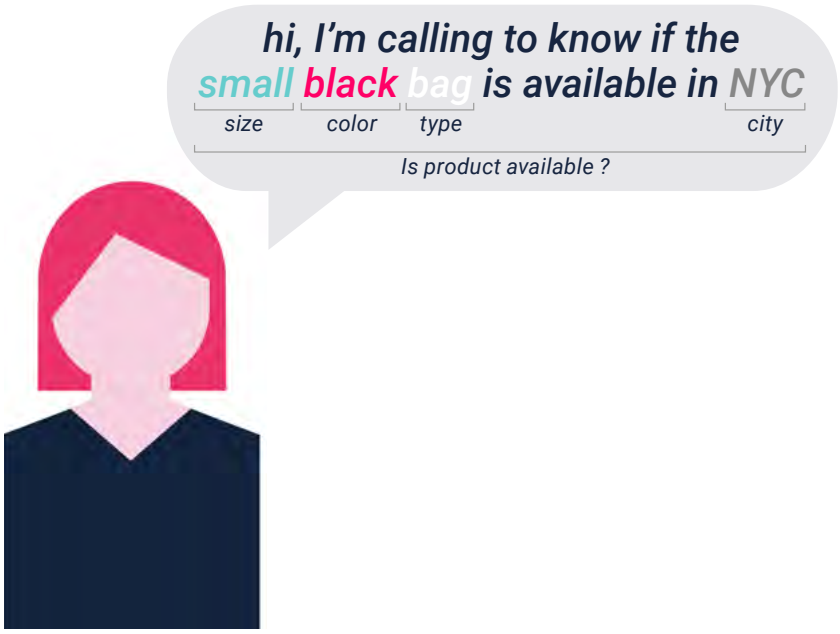
"OK Google, play the Rolling Stones on Spotify.", "Alexa, what is the weather like in Paris today?", "Siri, who is the French president?"

If you have ever used vocal assistants, you have indirectly used some Natural Language Understanding (NLU) processes. The same logic applies to chatbot assistants or automated routing of tickets in customer services. For some time now, NLU has been part of our everyday life and it's probably not about to stop.

Automating the extraction of customer intent, for example, NLU can definitely help us answer our clients' requests faster and more accurately. That is why every large company has embarked on the development of its own solution. Yet, with all the libraries and models existing in the NLU field, all claiming state-of-the-art or easy-to-get results, it is sometimes complicated to find one's way around. Having experimented with various libraries in our NLU projects at Artefact, we wanted to share our results and help you get a better understanding of the current tools in NLU

What is NLU?

Natural Language Understanding (NLU) is defined by Gartner as "the comprehension by computers of the structure and meaning of human language (e.g., English, Spanish, Japanese), allowing users to interact with the computer using natural sentences". In other words, NLU is a subdomain of artificial intelligence that enables the interpretation of a text by analysing it, converting it into computer language and producing an output in an understandable form for humans. If you look closely at how chatbots and virtual assistants work, from your request to their answer, NLU is one layer extracting your main intent and any information important to the machine so that it can answer your request best. Say you call your favourite brand customer service



to know if your dream bag is finally available in your city: NLU will tell the assistant you have a product availability request and look for the particular item in the product database to find out if it is available at your desired location. Thanks to NLU, we have extracted an intent, a product name and a location.

Illustration of a customer intent and several entities that are extracted from conversation

Natural language is instilled in most companies' data and, with the recent breakthroughs in this field, considering the democratization of the NLU algorithms, the access to more computing power & more data, a lot of NLU projects have been launched. Let's look at one of them.

Project presentation

A typical project using NLU is, as mentioned before, helping call centre advisors answering customers' requests more easily as the conversation goes. This would require us to perform two different tasks:

- Understand the customer's intent during the call (i.e.: text classification)
- And catch the important elements that would make it possible to answer the customer's request (i.e.: named-entity recognition), for example contact numbers, product type, product colour, etc...

When we first looked at the simple and off-the-shelf solutions released for both of these tasks we were able to find more than a dozen frameworks, some developed by the GAFAM, some by open-source platform contributors. Impossible to know which one to choose for our use case, how each one of them performs on a concrete project and real data, here call centre audio conversations transcribed into text. That is why we have decided to share our performance benchmark with some tips as well as pros and cons for each solution that we have tested.

It is important to note that this benchmark has been done with English data and transcribed speech text and therefore can be used less as a reference for other languages or applications directly using written text, e.g. chatbot use cases.

Benchmark

INTENT DETECTION

The goal here is to be able to detect what the customer wants, his/her intent. Given a sentence, the model has to be able to classify it into the right class, each class corresponding to a predefined intent. When there are multiple classes, it is called a multi-class classification task. For example an intent can be "wantsToPurchaseProduct" or "isLookingForInformation". In our case we had defined 5 different intents and the six following solutions were used for the benchmark:

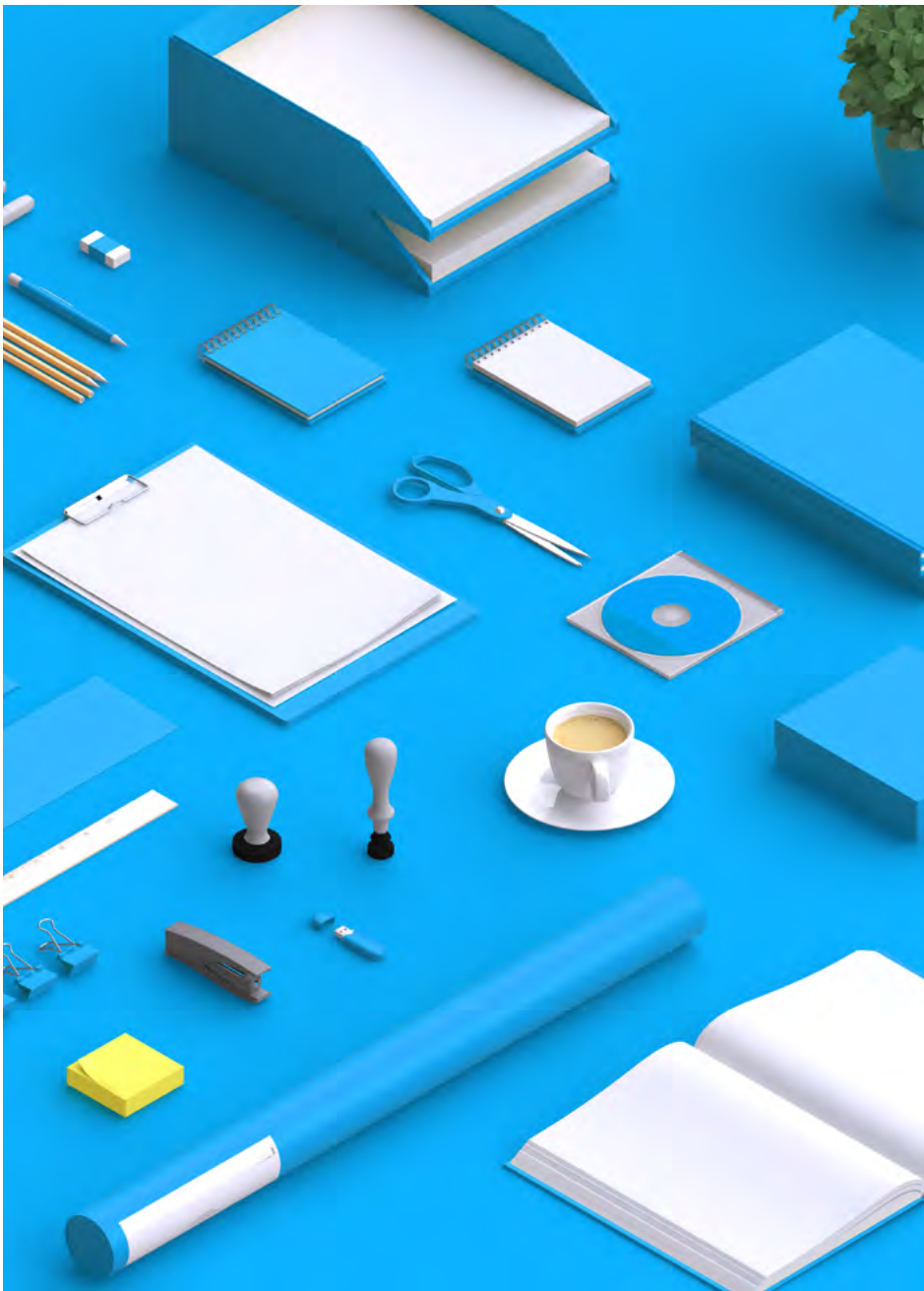
- **FASTTEXT:** library for efficient learning of word representations and sentence classification created by Facebook's AI Research lab.
- **LUDWIG:** a toolbox that allows to train and test deep learning models without the need to write code, using the command line or the programmatic API. The user just has to provide a CSV file (or a pandas DataFrame with the programmatic API) containing his/her data, a list of columns to use as inputs, and a list of columns to use as outputs, Ludwig will do the rest.
- **LOGISTIC REGRESSION WITH SPACY PREPROCESSING:** classic logistic regression using scikit-learn library with custom preprocessing using spaCy library (tokenization, lemmatization, removing stopwords).
- **BERT WITH SPACY PIPELINE:** spaCy model pipelines that wrap Hugging Face's transformers package to access state-of-the-art transformer architectures such as BERT easily.
- **LUIS:** Microsoft cloud-based API service that applies custom machine-learning intelligence to a user's conversational, natural language text to predict intent and entities.
- **FLAIR:** a framework for state-of-the-art NLP for several tasks such as named entity recognition (NER), part-of-speech tagging (PoS), sense disambiguation and classification.

The following models have all been trained and tested on the same datasets: 1600 utterances for training, 400 for testing. Models have not been fine-tuned therefore some of them could potentially have better performances than what is presented below.

Overall, in terms of performance, all the solutions achieve good or even very good results (F1-score > 70%).

One of the inconveniences of Ludwig and LUIS is that they are very "black-box" models which make them more difficult to understand and fine-tune.

LUIS is the only solution tested that is not open source thus it is much more expensive. In addition, the use of its Python API can be complex since it has been initially designed to be used via a click-button interface. However, it can be a solution to prefer if you are in the context of a project that aims to go into production and whose infrastructure is built on Azure for example, the integration of the model will then be easier.



	FASTTEXT	LUDWIG	LOGISTIC REGRESSION WITH SPACY	BERT WITH SPACY	LUIS	FLAIR
PRECISION	94.8%	96.9%	95.3%	92.7%	90.4%	68.4%
RECALL	95.4%	94.3%	87.8%	98.3%	98.4%	73.4%
F1-SCORE	95.1%	95.6%	91.4%	95.4%	94.2%	70.8%
INFERENCE TIME*	0.002 ms	2 ms	1 ms	137 ms	139 ms	533 ms
COST	No	No	No	No	Yes	No
TUNING PERSPECTIVE	High	Low	High	High	Low	Low

Performance results of different models for intent detection.
**Inference time on local Macbook Air (1.6GHz dual-core Intel Core i5–8 Go 1600 MHz DDR3 RAM).*

ENTITIES EXTRACTION

The goal is to be able to locate specific words and classify them correctly into predefined categories. Indeed, one you have detected what your customer would like to do, you may need to find further information in his/her request. For instance, if a client wants to buy something you may want to know which product it is, in which colour or if a client wants to return a product, you may want to know at which date or which store the purchase was made. In our case we had defined 16 custom entities: 9 product-related entities (name, colour, type, material, size, …) and additional entities related to geography and time. As for intent detection, several solutions have been used to make a benchmark:

- **SPACY:** an open-source library for advanced Natural Language Processing in Python that provides different features including Named Entity Recognition.
- **LUIS:** see above
- **LUDWIG:** see above
- **FLAIR:** see above

	SPACY	LUIS	LUDWIG	FLAIR
PRECISION	95.1%	92.4%	100%	100%
RECALL	88.9%	90.3%	18.9%	43.6%
F1-SCORE	91.9%	91.1%	31.8%	60.7%

Performance results of different models for named entity recognition.

The following models have all been trained and tested on the same datasets: 1600 utterances for training, 400 for testing. Models have not been fine-tuned therefore some of them could potentially have better performances than what is presented below.

- Two models perform well on custom named-entity recognition, spaCy and LUIS. Ludwig and Flair would require some fine-tuning to obtain better results, especially in terms of recall.
- One advantage of LUIS is that the user can leverage some advanced features for entity recognition such as descriptors which provides hints that certain words and phrases are part of an entity domain vocabulary (e.g.: colour vocabulary = black, white, red, blue, navy, green, …).

Conclusion

Among the solutions tested on our call centre dataset, whether for intention detection or entity recognition, none stands out in terms of performance. In our experience, the choice of one solution over another should therefore be based on their practicality and according to your specific use case (do you already use Azure, do you prefer to have more freedom to fine-tune your models…). As a reminder, we just took the libraries as they are to produce this benchmark, without fine-tuning the models, so the results displayed are to be taken with slight hindsight and could vary on a different use case or with more training data.

CASE STUDY

HOMESERVE

Using speech analytics to improve customer satisfaction

"The detection of non-compliance in sales calls use case analysis allowed us to prove that AI can be leveraged to better orient the work of the compliance team."

CHALLENGES

Present in France for 20 years, HomeServe is the world leader in home insurance services, with 8 million customers and over one billion in revenue.

When it comes to home emergencies, the most common channel used by customers is the phone – 9 out of 10 customers prefer it. This particularity places the call centre at the heart of every step of the insurance value chain, from sales to customer service, and ultimately assistance.

Although HomeServe has already developed AI-based conversational solutions and is present on Google Assistant and Amazon Alexa, they wanted to explore new ways in which AI could improve efficiency and customer experience in their existing phone channel.

They were especially interested to see what impact speech analytics could have on the vast amounts of unexploited customer data they collected.

SOLUTION

Artefact began by helping HomeServe opt for a "make" over "buy" strategy, as only a proprietary asset tailored to their organisation, combining technology and skills, could meet their many objectives, which include:

We also laid out a plan for developing HomeServe's expertise in natural language, data science algorithms, and AI data-treatment technical structures.

Next, Artefact set up a long-term multidisciplinary team with HomeServe comprised of a business team, a core data team, and an IT team to assess the maturity of speech analytics, the value and feasibility of relevant use cases, and improvements to customer experience and efficiency.

Because we couldn't build the entire architecture right away, we needed to quickly demonstrate the value of speech analytics to all stakeholders via a minimum viable product (MVP), able to expand after its validation with business experts.

To do this, we analysed two high-value use cases in a four-week cross-company workshop. We developed several microservices for data collection and processing and packaged to enable these use cases to be developed and be reused in the future, should the MVP phase prove successful.

1. **Refining understanding of customer contact root causes**
2. **Detecting risks of non-compliance within sales calls**

RESULTS

The most important conclusion for Artefact is that we proved the technology is mature. Speech analytics is ready to produce value for companies right now.

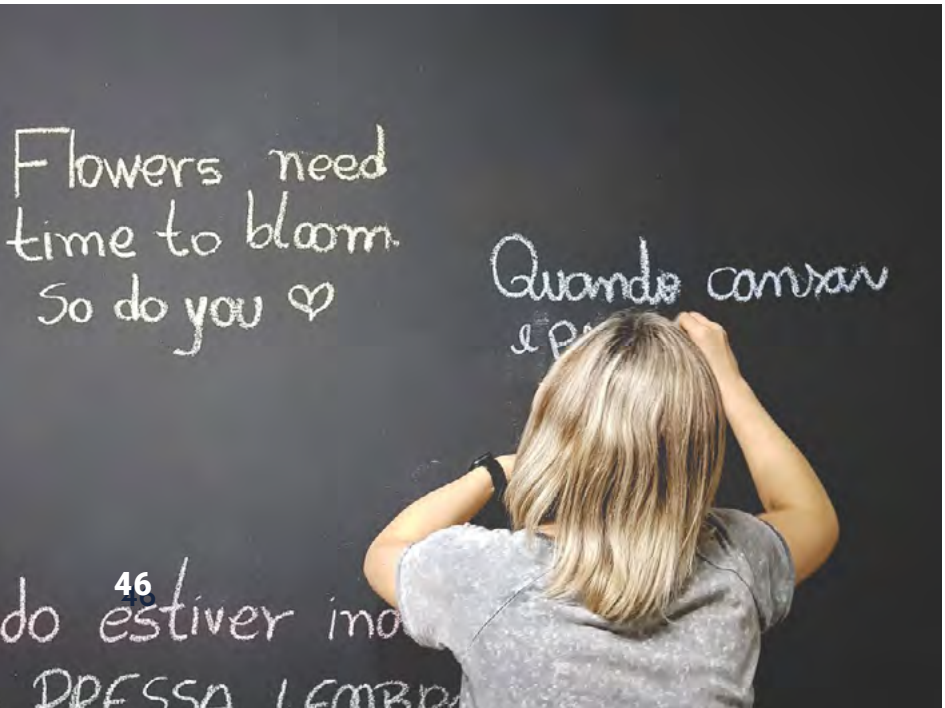
The customer contact root cause use case analysis produced three actionable insights, which could help call centre agents to perform better, sell more contracts, and benefit from a less tedious workload:

The detection of non-compliance in sales calls use case analysis allowed us to prove that AI can be leveraged to better orient the work of the compliance team.

How to train a language model from scratch without any linguistic knowledge.

This article explains how I created my own made language model in Korean, a complex language with limited training data. Here you'll be able to learn how to train a language model without having the luxury of understanding this language yourself. You'll find tips on where to get training data from, how much data you need, how to preprocess your data and how to find an architecture and a set of hyperparameters that best suit your model. My key learnings are:

Amale Elhamri
Senior data scientist



DATA COLLECTION

- When Wikipedia does not have enough volume or that it's not enough used by native speakers of the language you want to train your language model from, a good thing to do is to combine Wikipedia with other data sources such as CommonCrawl.

DATA VOLUME

- Choose documents that best represent the Korean language. Too many documents would not be useful since the marginal performance improvement would be too small compared to the huge training time.
- Choose documents that contain the most used words in the Korean language.
- Find an architecture that manages to modelise the complexity of the training data.
- Find the right combination of regularisation parameters not to overfit.

Introduction

If you don't know it already, NLP had a huge hype of transfer learning in these past 2 years. The main idea is to re-use pre-trained language models for another NLP task such as text classification. A language model is a deep learning model that given part of a sentence can predict the next word of the sentence. The intuition to understand from this is that this kind of model understands the language structure, grammar, vocabulary, and the goal is to "transfer" that knowledge to other downstream models.

	1	2	3
STEPS	**GENRAL DOMAIN LM PRETRAINING** →	**TARGET TASK LM FINE TUNING** →	**TARGET TASK CLASSIFIER**
WHAT IS DONE	LM is pretrained on a large general-domain corpus	The LM is consequently fine-tuned on the date of the target task	We want to provide a sentiment classification
ABILITY OF THE MODEL	The model is able to predict the next word in a sequence	The model has learned target task-specific features of the language	The final output is a probability distribution over the sentiment labels (i.e. positive, negative and neutral)

Example: A simple recipe on how to improve a text classifier using fine-tuning.

"This figure summarises the ULM Fit method that I used for training my language model and therefore fine-tune it and transfer it into a text classifier."

STEP 1 Train a general language model on a large corpus of data in the target language. This model will be able to understand the language structure, grammar and main vocabulary

STEP 2 Fine-tune the general language model to the classification training data. Doing that, your model will better learn to represent vocabulary that is used in your training corpus

STEP 3 Train a text classifier using your fine-tuned pre-trained language model. This method allows your model to understand the words in their context. Furthermore, using a pre-trained language model allows you to train your classifier on very few training examples (as little as 400 texts per label would do the job).

We already know that text classification works nicely on English, French, German, Spanish, Chinese... but what should we do on languages with very few off the shelf language models?

Before going into further details, you may be wondering why a french data scientist like me would want to have a text classifier in Korean ? The reason is that I am part of a project that develops a product to classify social media posts into different categories. After validating the methodology on English and French, we started scaling it to other languages (English, French, Japanese, Chinese and Korean). Only there was a bigger challenge in Korean language because there was no pretrained language model to be found in open source so I had to do it myself with very few Korean linguistic resources.

This article will be focusing on Korean text classification by using the multifit method explained in the following paper.

A lot of languages are very represented in the web such as : English, Chinese, Spanish, Portuguese, French ... but Korean remains very poorly documented and not a lot of content is ready for reuse. So I thought about contributing myself by sharing my key learnings with you, while discovering Korean NLP.

In this article I will tell you about my journey to train a Korean language model without understanding a single word of Korean and how I used it for text classification.

DISCLAIMER

Usually, we consider a language model as good when it reaches an accuracy of about 45–50%. As my goal is not to generate Korean text, I don't need to reach such performances: I only need a model that "understands" the grammar and structure of the Korean language so that I can use it to train a Korean text classifier.

1 — Data collection for language model training

1.1 — DATA SOURCE

Usually, when training a language model from scratch, ULM FiT tutorial suggestions are to download all Wikipedia content in the given language. These guidelines only work if native speakers of this language are used to publishing a lot on this channel.

In Korean, it appears that people are not used to it: not only Wikipedia Korean context has not enough volume, and it is also not representative of native Korean speaking.

Here is a comparison between the number of articles in English and Korean Wikipedia to give some hints *(cf figure1)*.

	ENGLISH	KOREAN
NUMBER OF ARTICLES IN WIKIPEDIA	6,185,131	525,014

Figure1: Wikipedia volume in different languages

MY ADVICE

I combined Wikipedia articles with Common Crawl data.

1.2 — DATA VOLUME

Let's remember that a language model is a model supposed to predict the next word in a text. To do that, our model should have seen a lot of examples to learn the language and be good at speaking it. That being said, it is not useful to go beyond 100 millions of tokens. It only adds complexity to your model as well as a huge training time.

So at first glance, once I had retrieved all Wikipedia and Common Crawl data, I found myself with much more than 100 millions of tokens so I had to pick and choose the most relevant documents to train my model with. The goal of my methodology is to keep the documents that represent in the best way the native Korean language:

- I first performed a weak tokenization on my corpus to approximate the number of tokens I had by splitting the corpus into spaces.
- I removed all numbers, emojis, punctuation and other symbols that are not specific to Korean from my obtained tokens.
- I computed a counter of all tokens in my corpus and retrieved the top 70,000 mentioned tokens
- Then I retrieved documents that mention most of the top used tokens such that my corpus would be built of 100 million tokens and there was my training corpus!

Now that we have our raw corpus of training we can start the real business!

2 — Data tokenization

I guess when I told you earlier that I tokenized with a split function, you started thinking that this article was really a joke but let's reassure you, this was never my end game!

First let's remind you that no further data preprocessing is required for training a language model. A lot of NLP tasks perform some text stripping of numbers, stopwords, lowercasing, stemming … All of those would strip your text from its context and our goal is to learn to speak Korean so we must keep all our text as it was originally written.

To tokenize Korean text I tried two tokenization models:

- Korean spacy model that is a wrapper to Korean mecab tokenizer.
- Sentencepiece subwords tokenizer model trained on my corpus with 28000 maximum tokens.

As it's recommended in the multi-fit article, I went with the second option to have a subword granularity.

3 — Training model

When training a language model as well as training any model, the two things that you want to avoid are underfitting and overfitting.

A model under fits when it is too simple with regards to the data it is trying to model. You can detect that when you find that your model cannot learn on your training data and that your training loss does not converge to 0 at all.

On the opposite, a model over fits when it learns "too well" to model your training data but that performance remains low on the test data. That is a sign that your model is not likely to predict well data that it hasn't seen.

When I started to train my language model, at the beginning I was struggling to learn anything from my data. As you can see on the picture below after 10 epochs of training my training loss was not decreasing by an inch. *(cf figure2)*

What it means is that my model was too simple to represent the complexity of Korean language.

Here is what I did to overcome this issue:

As you can imagine, debugging any deep learning model is not easy as there are so many degrees of liberty. You have to find the right network structure as well as the right set of hyperparameters.

To simplify the problem at the beginning, the right way to go is to try to overfit on a single batch of data. The idea here is to make sure that given some data, your model can interpret its complexity and perform well on the training set.

Here are all the things I tried :

- Increase embedding size
- Increase the number of hidden layers
- Changing optimiser functions
- Changing learning rate

After lots of attempts, here is the structure and hyperparameters that allowed my model to start learning :

Neural network architecture:

- QRNN Structure
- Number of hidden layers: 2500
- Number of layers: 4
- Embedding size: 768

Once your model can predict correctly on your training set, the next thing you want to avoid is overfitting.

Here are some regularizations that I tried to make sure my model would not overfit.

- Add dropout
- Add weight decay
- Add gradient clipping

Here are the regularizers that I used for training my model

- Learning rate: 0.0002
- Weight decay: 1e-8
- Gradient clipping: 0.25

```
%%time
learn.fit_one_cycle(10,lr, wd=wd, moms=(0.8,0.7),
                    callbacks=[ShowGraph(learn),
                               SaveModelCallback(learn.to_fp32(),
                                                 monitor='accuracy',
                                                 name='bestmodel_sp15multifit')])
```

80.00% [8/10 3:32:56<53:14]

EPOCH	TRAIN_LOSS	VALID_LOSS	ERROR_RATE	ACCURACY	PERPLEXITY	TIME
0	9.210296	9.210295	0.999284	0.000716	9999.543945	26:33
1	9.210299	9.210295	0.999284	0.000716	9999.534180	26:31
2	9.210279	9.210295	0.999284	0.000716	9999.534180	26:36
3	9.210284	9.210295	0.999284	0.000716	9999.534180	26:39
4	9.210316	9.210295	0.999284	0.000716	9999.534180	26:38
5	9.210309	9.210295	0.999284	0.000716	9999.534180	26:38
6	9.210290	9.210295	0.999284	0.000716	9999.534180	26:37
7	9.210315	9.210295	0.999284	0.000716	9999.534180	26:38

Figure2

Results

After training my model for 15 epochs, I finally reached an accuracy of 25% and a perplexity of 100. As I said at the beginning, I never intended to use my language model for text generation so I was already satisfied to know that my model can predict correctly one word out of 4.

Then I re-used my pre-trained model for text classification. The dataset I used is a balanced dataset made of 10k social documents coming from Instagram, Facebook, Youtube and websites that were labelled as “label1” or not “not label1”. My goal was to predict that a new publication is about “label1” or not.

	ACCURACY	PRECISION	RECALL
ENGLISH	90%	88%	80%
FRENCH	87%	77%	81%
KOREAN	86%	86%	80%
JAPANESE	90%	85%	85%
CHINESE	91%	82%	81%

Performances of different languages text classifiers

Here are the performances I get for all the languages we developed:

So even without speaking the language and training the pre-trained language model myself, the performances for the Korean text classifier reaches quite well the other languages performances.

I still have a lot of things that I should try to improve the performances I get but still, it was kind of a hail mary to learn to process documents of a complex language like Korean without understanding a word of it and without finding relevant information and advice on the web.

Next steps

I have just described how I could improve a Korean text classification model leveraging a simple language model made from scratch. The initial performance is already good but there is room for improvement. I think what I would like to work on in the short run would be:

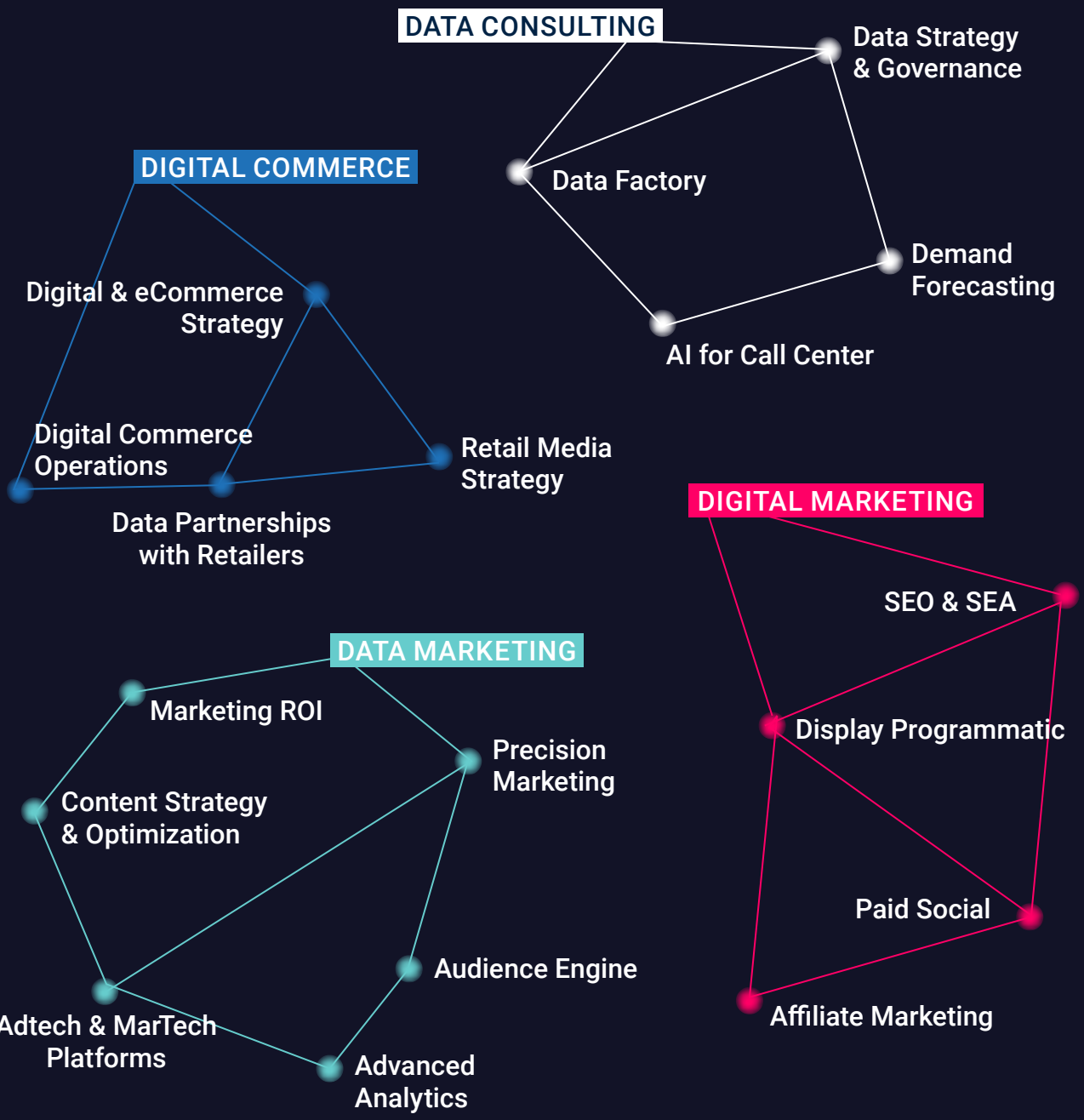
- Proofread the tokenization: as I don’t speak a word of Korean, it would be interesting to have a native Korean speaker have a look at the tokenization and confirm that it makes sense.

Enhance my language model and compare classification performances by:

- Transfer-learning a backward language model, as it appeared to have been more performant on English or French
- Transfer-learning a bi-directional language model
- Having dynamic learning rates during training to avoid being stuck in a local minimum.

ARTEFACT

Every company talks about data.
At Artefact, we don't talk, we act.



DATA CONSULTING | DATA & DIGITAL MARKETING | DIGITAL COMMERCE

ARTEFACT

VALUE BY DATA

CONTACT

hello@artefact.com
artefact.com/contact-us

ARTEFACT HEADQUARTERS

19, rue Richer
75009 – Paris
France

artefact.com