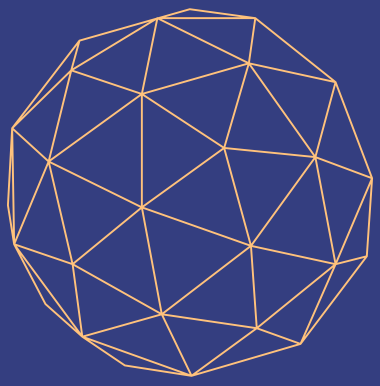
Issue 1 - September 2021

Good In Tech RESEARCH NEWS

Rethinking innovation and technology as drivers of a better world for and by humans



Data management

VOGEL ROBIN, CLEMENÇON STEPHAN, LAFFORGUES PIERRE

Visual Recognition with Deep Learning from Biased Image Datasets 2021

KEVIN MELLET

Relationship marketing and personal data. Loyalty cards, data collection and GDPR compliance

GRAZIA CECERE, FABRICE LE GUEL, VINCENT LEFRERE

Third Parties in the App Market and Economics of Privacy, Economics Bulletin

Institut Mines-Télécom Business School
TELECOM Paris TELECOM SudParis INSTITUT POLYTECHNIQUE DE PARIS
FONDATION DU RISQUE Louis Bachelier
SciencesPo ÉCOLE DU MANAGEMENT ET DE L'INNOVATION

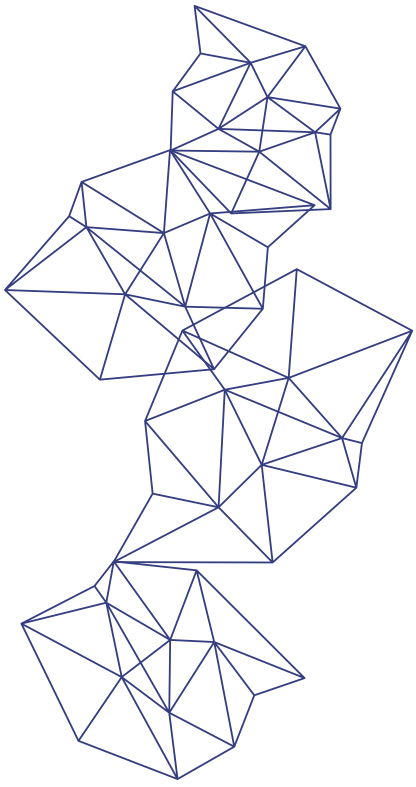
GOOD IN TECH VISION

Good In Tech main objectives are to create knowledge around four research areas and to contribute to the dissemination of this knowledge not only in academic and pedagogical spheres but also to corporations, decision-makers, regulators and the general public.

To this end, the Chair aims to create and develop an ecosystem of interactions between research, companies, students from the two partner academics and political institutions, civil society in order to raise awareness of all stakeholders on this new paradigm on responsible digital technologies and innovation.

The chair also aims to develop international partnerships, particularly in Europe, to share the issues of responsible digital innovation with international committees.

Finally, the Chair aims to share the results of academic works and debates it organizes with national and European political institutions in order to inform and influence public policies.



Sampling Weights for Visual Recognition with Biased Training Datasets

Robin Vogel, Pierre Laforgue, Stéphan Clemençon

Robin Vogel

In 2020, Robin Vogel completed his PhD at the school Télécom Paris, in a collaboration with the computer security firm IDEMIA. He graduated from an engineering degree from ENSAE Paris in 2016 and completed the M2 Data Sciences, a program co-hosted by Ecole Polytechnique, ENS Cachan, Télécom Paris and ENSAE Paris.

Pierre Laforgue

Dr. Laforgue is currently a postdoctoral researcher at the Università degli Studi di Milano in Milan, Italy. He holds a PhD in Machine Learning, prepared at Télécom Paris under the supervision of professors Florence d'Alché-Buc and Stephan Clémençon. His dissertation focuses on Deep Kernel Representation Learning for Complex Data and Reliability Issues. During his PhD, he was awarded a research grant by the chair Good in Tech to study algorithms in presence of selection bias.

Stéphan Clemençon

Stephan Clémençon conducts his research in applied mathematics at the LTCI laboratory of Télécom Paris. He leads the S2A (Statistics and Applications) research team. His research topics are mainly in the fields of statistical learning, probability and statistics.

WHY IS THIS TOPIC IMPORTANT ?

Robin Vogel and Stephan Clémençon, specialized in statistics, see problems in a probabilistic way. For them, the data is the realization of a statistical law and this allows them to draw conclusions about the generative process of the data, so it is a way to propose a mathematical model when doing machine learning, which complements the traditional empirical approach.

The mathematical approach brings an intuition, a direct logic to attack the problems and **correct the biases**. A bias is defined as a problem in the representation of a database.

With the deluge of digitized information in the Big Data era, massive datasets are becoming increasingly available for learning predictive models. However, **in many situations, the poor control of the data acquisition processes may naturally jeopardize the outputs of machine-learning algorithms and selection bias issues are now the subject of much attention** in the literature. It is precisely the purpose of their work to investigate how to extend **Empirical Risk Minimization (ERM)**, the main paradigm of statistical learning, when the training observations are generated from biased models, i.e. from distributions that are different from that of the data in the test/prediction stage.

METHODOLOGY

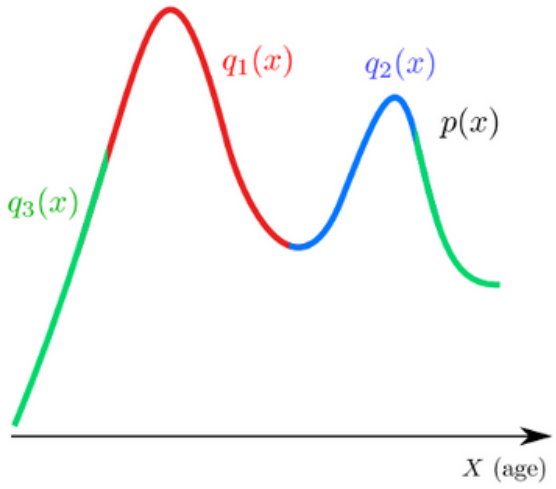
T**he problem was first approached from a statistical perspective** in a first technical paper published by Stéphan Clémençon and Pierre Laforgue. The work "Statistical Learning from Biased Training Samples" proposed theoretical guarantees for learning with a set of biased databases.

Within the Good in Tech Chair, researchers are considering the application of Laforgue and Clémençon (2019) to the case of visual recognition. They propose practical approaches for visual object recognition with several biased datasets, in that their distribution does not match that of the testing data.

Technical paper by Stéphan Clémençon and Pierre Laforgue "**Statistical Learning from Biased Training Samples**" (2019) :

- The purpose of this project is to **be able to learn with several biased databases**.
- How can we compare several different databases with each other? How can we re-weight the databases to fit the application case? This question is complex, especially if we are talking about non-overlapping universes. Our databases may not cover all the space of the test database, so we need to make assumptions to compare our databases.
- · The researchers' assumptions to show that proper weights can be found among the different observations are as follows:
  - There is an overlap between these distributions, a set of data (profiles in biometrics for example) that are found in each of the databases in order to determine their relative weight.
  - We know the bias function. The bias function is the ratio between the test distribution and the training distribution, two distributions used in machine learning.

- We must see machine learning as a process of learning from the law that generated the data. If the test situation and the training situation differ, we must try to understand the reasons.
- Here **there are several training functions, a test function and a bias function that we know**.



$$q_1(x)/p(x) = \mathbb{I}\{15 \leq x \leq 55\}$$

$$q_2(x)/p(x) = \mathbb{I}\{50 \leq x \leq 70\}$$

$$q_3(x)/p(x) = \mathbb{I}\{x \leq 20\} + \mathbb{I}\{x \geq 60\}$$

Following the work of Laforgue and Clemençon (2019), they derive sampling weights for the input data from the knowledge of biasing functions ω. Those biasing functions are proportional to the Radon-Nikodym derivative of the training distribution with respect to the testing distribution. However, in the context of highdimensional signals and particularly images, two distributions generally do not have the same support and the true biaising functions are unknown.

For these reasons, they investigate applying their methodology using sensible approximations of the biasing functions, expressed by combining with simple a priori information on the test data with either 1) auxiliary information on the dataset, or 2) a lower-dimensional transformation of the data.

KEY FINDINGS

In short, this technical paper analyzes **how to weight the distributions between them to obtain a sample that represents the test distribution.** If we learn with this re-weighted representative sample, we will have a good performance with this sample.

Their experiments show that relevant bias functions can improve results by selecting relevant data.

KEY TAKEAWAYS

How can we use this research?

The objective of this research work is to **provide a technical method to correct biases**. Such biases in databases can have undesirable consequences on the algorithms. In order to avoid them, **it is necessary to make machine learning researchers aware of the issue of biases**, so that they can deploy technical approaches to correct their impact.

In the long term, this work could therefore be used to improve statistical models that may be biased. The field of application is wide and can be applied to several fields such as biometrics.

Relationship marketing and personal data. Loyalty cards, data collection and GDPR compliance

Kevin Mellet

Kevin Mellet is an Assistant Professor of Sociology at Sciences Po, and a researcher at the CSO. He is the Scientific head of the "Marketing & society" master's degree within the Innovation Management School of Sciences Po. His research work mainly draws on economic sociology and science & technology studies to study market techniques in the digital age. Current research interests focus on the emerging data marketing landscape and the formation and regulation of the personal data economy. His involvement in the 'Good on Tech' Chair is twofold. First, he promotes and relays the Chair's activities and calls for projects within the community of Sciences Po teachers and researchers, in conjunction with the two chairholders, Dominique Cardon and Christine Balagué. Then, he is involved as a researcher in the activities of the chair, with a research project on the compliance practices of companies in the field of relationship marketing.

WHY IS THIS TOPIC IMPORTANT ?

> **The worlds of marketing are being reconfigured and transformed with big data.**

More and more data is being used to know, predict, track and target customers. This poses a problem because it is personal data that may escape the control of users.

Law is a major instrument for the governance of digital innovation, alongside soft law instruments, such as the development of CSR indicators for responsible innovation.

In order to study its impact, as numerous studies within the "Law & Society" research stream have shown, it is important to observe precisely how compliance with new legislation is constructed. Thus, to **understand the impact of the General Data Protection Regulation (GDPR)**, the author agree to look at how it is interpreted and translated in specific universes, as close as possible to the tools and practices of professionals.

In this paper, Kevin Mellet **focuses on data collection practices in the field of relational marketing**, around customer loyalty programs in retail. **The loyalty card** is considered as a strategic asset. In addition to its effectiveness as a promotional support, the loyalty program is the main tool for building actionable databases within commercial organizations. This usefulness is reinforced in a context of massive investments in big data techniques, in the retail sector as in the rest of the economy.

How is user consent obtained through loyalty programs, and how is it used, in a post-RGPD context? One can observe a very strong tension between a horizon of economic valuation of data and a tightening of the conditions of data extraction under the effect of the regulation. This tension can be seen, for example, in the interfaces for collecting user consent. A detailed analysis of the consent collection interfaces for different loyalty programs would allow them to account for the way in which the retail players are trying to resolve this tension.

Ultimately, how do the universes of relationship marketing, and in particular the universe of loyalty cards, operate and organize themselves? How do they work on customer value? How are they regulated?

METHODOLOGY

To answer these questions, the author uses two main methods.

Interviews:

- Qualitative survey with the professionals involved in extracting and exploiting data. In particular, professionals in charge of data marketing and customer value work.
- Conducted a dozen of interviews

Observational survey:

- Survey on the way customers' consent is collected when they sign up for a loyalty program
- Conducted for a dozen retail brands

KEY FINDINGS

The issue of data is largely understated.

- As in other universes such as online advertising (with cookies), the collection of data is based on user consent, which tends to be euphemistic, discreet, hidden in the world of loyalty programs.
- It therefore appears that users are not fully informed of the uses that will be made of their data, nor of the central place that their data holds in the management of loyalty programs. Loyalty programs are not presented as crucial asset building mechanisms. They are presented only as reward and support programs for customers.
- A retail company does not know its customers, unless they are associated with a customer file. Thanks to loyalty programs, companies are able to record the behavior of customers, their habits and their personal information.

However, it would be **wrong to believe that this universe is totally unregulated**.

- Relationship marketing professionals are constrained and guided in their work by a set of legal and technical regulations.
- This framework raises questions: how do these professionals work? What are the constraints to which they are subjected?
- Interviews with professionals should shed additional light on these questions.

KEY TAKEAWAYS

The author made three recommendations:

- A first action would be to **make the differences between loyalty programs visible and comparable in terms of how they manage the data of loyal customers**: what data is extracted? For what purposes? This could take the form of an ad hoc or permanent observation platform focused on the practices of retail players in this area. This comparative observatory would allow consumers to assess the interest of these programs in terms of what membership implies in terms of privacy. This type of action could be carried out by a consumer association.

- As an extension of the previous action, **the way in which data becomes a lever of economic value in retailing should be made more visible and clear to the end customer**. Too often, this is done at the very least, using agreed-upon terms whose sole purpose is to protect against lawsuits but which mean nothing to the user: improved services, personalization, targeted advertising, etc. Using concrete examples, graphically describing the user's "data life" would allow them to understand the type of relationship that is also involved in engaging in a loyalty program. This type of action should be led by the retail players themselves.

- Consent has become the primary legal basis for data collection and processing in many areas, particularly in marketing and advertising. This consent, which has been made safe and strengthened by the GDPR, also has certain limitations, as shown by the difficulties in constructing an unbiased consent for the collection of cookies or other types of personal data. Of course, the bad faith of some actors is often to blame, but the **material difficulty of producing an explicit, specific, and unbiased consent should not be underestimated**. Consequently, it is also appropriate to **open a debate on alternative forms of regulation of personal data**.

Third Parties in the App Market and Economics of Privacy, Economics Bulletin

Grazia Cecere, Fabrice Le Guel, Vincent Lefrere

Grazia Cecere

Grazia Cecere is Professor of Economics at Institut Mines Telecom, Business School, LITEM. She is research fellow at Université Paris Sud. She completed her Ph.D in Economics at the University of Paris Saclay (France) and the University of Turin (Italy). Her main research interests are digital economy and more particularly the economics of privacy, algorithms and machine learning, economics of mobile applications, and digital marketing. Her research program was partly funded by the Good in Tech chair.

Fabrice Le Guel

Fabrice Le Guel is associate prof. (HDR) at the University of Paris Sud, in the RITM research center in economics of innovation and international economics of the University of Paris-Sud. He has already participated in several research projects financed by the French ANR related to digital economics such as ESPRI (economics of privacy), MOBITICS (mobility and ICT use), EXPERTIC (business model of digital economy) and co-lead an interdiciplinary research project named DAPCODS (2017-2021).

Vincent Lefrere

Vincent Lefrere is an Assistant professor in Economics at Institut Mines-Télécom, Business School, LITEM. He completed his Ph.D in Economics at the University Paris Saclay & Institut Mines -Télécom, Business School LITEM. His study of the governance of technologies was partly funded by the Good in Tech chair. His research interests are digital platform, economics of privacy, industrial organization, machine learning and artificial intelligence, advertising.. He uses Python to generate and manage original database: webscraping, Web API, Jupiter and Pandas program.

WHY IS THIS TOPIC IMPORTANT ?

> **The idea behind it is: when you don't pay, you are the product.**

The goal of the study is to explore **how data is traded in different markets as complementary or substitute goods**.
Vincent began exploring the topic of personal data in markets during his thesis. During this research on the governance of technologies, Grazia and Vincent wondered how to obtain empirical data and especially where it is most present. They also sought to investigate free platforms to understand how personal data is used in the market for free products. Specifically, they sought to further explore how personal data is being exploited for app monetization purposes.

They chose to **explore smartphone apps** because a lot of data flows through them and many apps are free. Websites would also have been a good option, but mobile applications offer more precise information about the data they collect, whereas websites only specify the cookies they gather. Among the data collected by the apps, we find very precise data such as geolocation or contacts that allow the applications to have an extremely precise knowledge of the users. The authors are also trying to understand whether the data is an asset that can be resold as it is to be valued or a complementary asset allowing the application to better know its consumers and to offer them targeted advertising.

> **Ultimately, how does personal data fit into the monetization strategy of companies?"**

METHODOLOGY

Vincent and Grazia use **empirical data that they exploit in a quantitative way**. This raw data was collected **via webscrapping**. This method consists of using a computer language such as Python to gather all the information on a particular subject on the web on a large scale. In this case, the method was used to capture information about the data collected by the application on the presentation page of an app.

This database was complemented by **another research work from Carnegie Mellon University**. This is work that measures the gap between the data that the application needs to function and the data actually collected by the mobile app.

Using these two databases, the authors conducted an empirical study of more than **460,000 apps**, or **one-third of the Android Play Store market in 2016**. This sample is representative of about 90% of the market.

KEY FINDINGS

The research led to three main findings:

- **The amount of personal data collected is negatively associated with the presence of an apparent business model**. In particular, apps that feature ads or contain in-app purchases collect little data.
- Similarly, **apps that are associated with a third party that values data collect less data**.
- Conversely, **apps without an apparent monetization strategy collect more data**. We can therefore assume that these applications use data for monetization purposes. Thus, the thesis put forth by the authors begins to be confirmed: if you don't pay, you are the product.

KEY TAKEAWAYS

How can we interpret these results?

> We can hypothesize that if an application goes through a third party and outsources advertising, it would require less data per individual because the third party has centralized enough data so that the application does not need to collect any more.

> This conclusion opens up the debate: is it better for Google to be able to control all our personal data but for all websites to have less need to collect information? Or is it better that several actors collect more data but without a concentration of data in the hands of a single actor? Finally, **where would our data be the safest ?**

Issue 1 - September 2021

Good In Tech RESEARCH NEWS

Rethinking innovation and technology as drivers of a better world for and by humans

Christine Balagué
Professor at Institut Mines-Télécom Business School
Co-holder of the Good In Tech Chair
christine.balague@imt-bs.eu

Dominique Cardon
Professor at Sciences Po and director of the medialab
Co-holder of the Good In Tech Chair
dominique.cardon@sciencespo.fr

Jean-Marie John-Mathews
Data scientist
Coordinator of the Good In Tech Chair
jean-marie.john-mathews@imt-bs.eu

Jade Vergnes
Writer for Good In Tech Research News
jade.vergnes@sciencespo.fr

Clic here to contact