EPIDEMIUM

WHITE PAPER

OPEN SCIENCE | CANCER EPIDEMIOLOGY | BIG DATA | COMMUNITY

This *White paper* was edited by Epidemium's team:

- Mehdi Benchoufi
- Olivier de Fresnoye
- Karine Lévy-Heidmann
- Ermete Mariani
- Ozanne Tauvel-Mocquet

// **Credit and acknowledgment:**
- Epidemium logo and pictograph: Marie-Sarah Adenis et Lucile Picon
- Drawings: Ermete Mariani (concept) and design Barbara Govin (design)
- Graphic design and page layout: www.ediconcept.fr
- Printed in France - January 2017

// **Disclaimer:** Roche and La Paillasse cannot be held responsible for the content and information included in the White Paper, as well as for any use that infringes on the rights of authors.

// **Print edition:** ISBN 979-10-97214-00-5
// **Digital edition:** ISBN 979-10-97214-01-2

Dedication

We dedicate this book to the patients, who are the reason why Epidemium was created, and to all of you who are willing to make a commitment to help the project grow.

Table of contents

Preface

Pr Cédric Villani

Sometimes progress is only a matter of improvement, about learning how to do things better, faster and more efficiently. And other times, we undergo more radical changes that cause us to change our behavior. When this happens, not only do new techniques emerge, but new ways of collaborating as well.

Today, one of the emerging techniques that is taking the world by storm is big data analysis. Big data is hard to define with accurately; in fact, even the term used to refer to the discipline is unclear: is it big data, data mining, or maybe giga-data? It is also difficult to define its range of possible applications as it combines various aspects of statistics, analysis, geometry, probability, optimization… However, the challenge is clear: expectations are high towards big data techniques, who have already revolutionized material collection, automatic translation, artificial intelligence, along with the business models of many companies, including the GAFA (internet giants). We actually expect so much from big data analysis that statisticians are now very prestigious, to the extent that it has become one of the world's most popular job. As a matter of fact, the American company Careercast ranked the job of data scientist as number 1 in its 2015 list of jobs with the most promising career prospect.

Therefore, it was only a matter of time for big data analysis to tackle one of the world's most dramatic scourge and one of the worst issue in our developed countries: cancer. Which family, in a country like France, has not been affected by this disease? It is a predicament that is especially terrible given how multi-faceted and varied it is and how its causes and risk factors are so extraordinarily diverse.

And this is the exact reason why so much is expected from the alliance between big data and cancerology: so many statistics are so difficult to interpret, with so many variables, that we figure we will have no other choice but to use new methods to come to a breakthrough, and bring to light discoveries and new factors that doctors will be able to use and implement.

However, at the heart of Epidemium's project is the idea that the use of this new technique should also give way to a new form of public and open collaboration, for the same reasons that made Open Software such a successful project in the 1990s, and FabLabs in the year 2000s. It represents a form of collaboration where the notions of platform, Wiki, data-sharing, work-sharing among various organizations, volunteering, cooperation, and synergies, would be meaningful; a framework in which tools, skills, and facilities would be pooled together; and where competition would serve the purpose of selecting which ideas are best.

The fact that La Paillasse and Roche were able to come together to join efforts is somewhat emblematic.

On the one hand, La Paillasse is a militant organization dedicated to the idea of open research. On the other hand, Roche is a major player in the pharmaceutical industry. This is a symbol that tells us that major research institutions have come to grip with the potential that open medical research represents.

After Epidemium defined the basics of its approach, so many obstacles had to be tackled! Identifying relevant databases, defining the contest, harnessing the energy and talent of participants, recruiting a jury, as well as defining the principles and rules governing the contest.

The organizers aimed to follow best practices. With that goal in mind, they recruited an Independent Ethics Committee responsible for determining a variety of safeguards in order to deal with the the sensitive issues, dilemmas and concerns surrounding the question of the use of datasets. This is because cross-referencing data has the potential of increasing efficiency, but it can also lead to the unacceptable risk of breaching people's privacy. I had the pleasure of being a member of this committe: it was a minor yet important task, which led me to ponder when I was asked to assess the projects involved in the competition.

Another best practice was the involvement of a patients association. Patients certainly have a say in the matter. Are they not the first to be concerned? It was therefore justified to let them play an important role in terms of governance.

Above all, there was a great willingness to collaborate, involving the experts of abstraction (mathematicians, statisticians, computer scientists) on the one hand and experts of the human body (cancerologist, doctors, surgeons, ...) on the other. We had to put in place a synergy between people to reflect the existing synergy between the various disciplines. This is an effort that can only be appreciated in the long term, and that Epidemium has sought to accelerate.

Since the project was officially launched, so much work has been accomplished! It was fascinating to experience, even from afar, the activities of the competing teams, in what came to be a great productive chaos.

In due course the time came for us to listen, decide, and award prizes… This represented the end of the first stage, but it was clear for everyone that this was also the beginning of an adventure that would play out over the long-run, and for which we would have to build on in order to help define the future of medicine. ■

Foreword

Isabelle Vitali (Roche) & Thomas Landrain (La Paillasse)

Epidemium: Research innovation in cancer epidemiology

Innovation is not only about developing new technologies or connected objects. It is also about experimenting new ways of working to push the frontiers of collective intelligence, thereby overcoming the barriers standing in the way of open knowledge and skills. Furthermore, we sincerely believe that innovation must have positive and tangible impacts on the lives of people and their health.

This is the visions that the open research laboratory La Paillasse and the pharmaceutical company Roche had in common. With Gilles Babinet acting as a facilitator, this vision ultimately led to a very ambitious project: Epidemium.

Since April 2015, Epidemium, Roche and La Paillasse have joined forces to experiment new ways of conducting medical and scientific research focusing on a major public healthcare issue: cancer. They have done this by leveraging a growing yet under-utilized source of knowledge: big data analysis. We had at heart to prove that in the field of science, the encounter of actors of different nature but driven by the same principles could lead to innovation.

The choice to start a pioneering project in the field of cancer epidemiology, both in terms of form and content, seemed logical to us for two main reasons:

- In spite of recent scientific breakthroughs, Cancer was responsible for 8.2 million deaths in 2012, including 148 000 in France[1] alone. Experts make the terrifying prediction that there will be a 60% increase in the number of deaths for women by 2030 in the world;

- Big data analysis is a precious source of medical and scientific knowledge, but its potential remains insufficiently used in the field of healthcare.

We very quickly understood that if such a project was to be successful, it would have to gather an unusual and diverse group of individuals, each having different skills and coming from a variety of backgrounds but sharing and supporting the same fundamental values: open knowledge, collaboration, interdisciplinarity, and independence.

Since the beginning, we had the intuition, which was later confirmed, that this project would bring us into uncharted scientific and ethical territory. We therefore had to assist and support the community in this approach. This is why we set out to create both an Independent Ethics Committee and a Scientific Committee. These two committees tackled new and unique challenges that are not usually within the perimeters of a traditional research project. Their constant involvement allowed us to find the right balance between the needs of scientific research and the need to safeguard the rights and well-being of all citizens in general, and of patients in particular, that are always at the heart of our commitment.

> *In the world of techno-medicine 2.0., Big Pharma and bio-hackers join efforts against cancer!"*
>
> ***Dominique Nora***
> *(L'Obs, 11/05/2015)*

The diverse and extended Epidemium community, composed of more than 1000 individuals who have taken part in the different events that have taken place over the course of a year, could not have been successful if it were not for the work of the coordination team. They played a key role in mobilizing a vast network of stakeholders, partners and community members.

Since the very beginning, Epidemium became the subject of interest of many players in the medical industry, in scientific research and in big data, who identified with our values and shared our ambition. This led very rapidly to the forging of a significant number of partnerships that have ultimately contributed to the success of the project and that allowed our community to benefit from a set of new skills, tools, and resources.

The time has come to draw the lessons of this first year of program and to share with you, in all humility, our experience to sustain the dynamics started with Epidemium.

Epidemium gave us, Roche and La Paillasse, the opportunity to combine our skills and to discover unexpected complementarities between both our organizations, but most important of all to be astonished by the accomplishments of collective intelligence expressed in the open community that we have supported and accompanied.

As a player in the field of healthcare innovation, Roche's ambition with Epidemium is to reinvent cancer epidemiology, to use it as a decision-making tool for therapy, and to bring innovation closer to patients. When looking at the results we were able to attain a year after the project had started, we are able to measure the immense advances made. Furthermore, Epidemium was a unique opportunity for Roche to experiment with new ways of working and to test new tools designed to stimulate innovation within an open science approach, to the point of opening some of our own data to the community.

As an emblematic center of innovation and interdisciplinarity, La Paillasse advocates in favor of new research methods through the use of an open and collaborative work environment. Epidemium's story is rooted in La Paillasse, where its community expanded and where it had a real space for its members to meet. These proved to be crucial elements in finding the right balance between the virtual exchanges and the face-to-face meetings. Epidemium also enabled La Paillasse to perfect its tools and working methods along with demonstrating the power that collective intelligence holds when it is applied to medical research.

> *Our deepest wish is that the program continues to develop."*

We also wish to take the opportunity that this White Paper offers to share with you our vision of what could be improved. In fact, what we have initiated will certainly not stop here…

Our first observation, which is something that we will work towards within the coming years, is to involve the medical community to a greater degree. Our initial goal was and remains to obtain results and concrete tools that can be utilized by doctors and the medical community in their day-to-day activities. Nevertheless, this cannot occur unless the medical community is involved throughout the program. Though working on data and developing machine learning algorithms is essential to make advances in healthcare today, and this is what we have sought to put at the heart of Epidemium, the input of healthcare experts and clinicians is a fundamental component in this endeavor.

Moreover, we seek to make our program and its challenges more intelligible. After one year of existence, we are conscious of the fact that big data in healthcare is a complex topic and that there are probably different ways to address it. We will always try to adopt a collaborative and open approach for our program and the way we build it. The means we have chosen to reach our goal could have been different; they can certainly be improved.

This is another reason why it was important for us to publish this White Paper, to take a step back and assess the work done, to remain objective in the judgment what has been accomplished, and finally to gain perspective for the future. Because our deepest wish is that the program will continue to develop well into the future.

Let's be grateful: we were astonished by a lot of what has been accomplished. We were impressed by the quality of what was

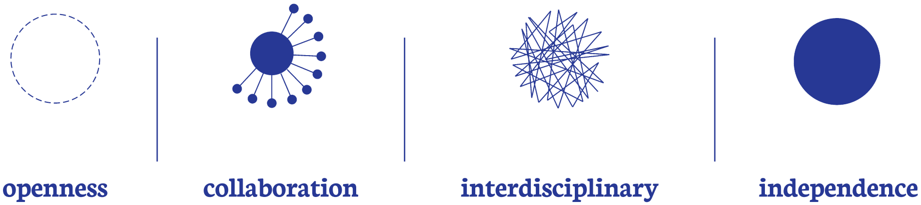
submitted, by the ability of the community to assemble and join efforts to address the challenges of a topic such as cancer, by the quality of experts that have remained committed throughout the entire program and finally by the good conduct of everyone involved in the project, whether it be participants, partners, committees, or the general public. ■

[1] Unicancer <www.unicancer.fr>, last access on November 30, 2016.

Introduction

Epidemium team

THE PROGRAM'S VALUES



Epidemium was born out of one strong will: to work in the field of cancerology thanks to open data using a community-oriented approach. It was also based upon two intuitions. First, that it was possible to obtain relevant results by opening up medical research or, at least, by making it more accessible to a greater number of players, including non-scientific ones. Second, that it was possible to conduct research based on open big data and that they could bring new perspectives in the understanding of our environment and of ourselves: being able to better understand, prevent, and predict. Therefore, Epidemium had to put in place a methodology whose goal was to combine scientific research, a community and open data. In a way, these three components represent the program's DNA.

Building on these intuitions and this initial vision, Epidemium naturally defined itself as an open science program revolving around four fundamental values: openness, collaboration, interdisciplinarity, and independence. These values were embodied in a program that was accessible to all, advocating for and enabling the sharing of knowledge and methods, collaboration and exchange of skills along with interdisciplinarity. Finally, the program is independent with regard to its initiators Roche and La Paillasse, and this is why it surrounded itself with an Independent Ethics Committee. Nevertheless, what does open science really mean and in what ways is Epidemium's program innovative?

Traditionally, scientific research is conducted within the context of dedicated and closed structures, where the production of knowledge is framed and its dissemination limited. However, for a few years now, new structures for the production of scientific knowledge have been emerging that are characterized by two main components. On the one hand, the project's number of participants, and on the other, making issues and results transparent throughout the entire process. We can overcome the barriers standing in the integration of new players or at the level of intellectual property. Every "citizen of science" can contribute and participate in the development and resolution of problems by bringing knowledge and skills to the table. It appears that participants to these open science projects share common values and ideals that gather them even beyond the topic of the project, in a concern of opening up their findings. This is the advent of a new epistemological conception of scientific research. The dynamics of open science encourages interactions between different stakeholders of the project, which improves the crowd's ability to come up with powerful and original solutions. Thus, for Epidemium, having access to heterogeneous skills and knowledge offers a greater statistical variety in the approaches to cancer research as well as new perspectives, that could all potentially lead to new paths to explore. Moreover, the opening up of results throughout the project, makes it possible for any contributor to access the latest results and thus to join the project at any moment, to put forward alternative ones or improvements. This increases the chances for coming up with better research material.

> *"Epidemium has everything to prove that science can be made outside of the academical frame by betting on an open source approach and self-constituted interdisciplinary teams."*
>
> ***Hugo Jalinière***
> *(Sciences et Avenir, 05/30/2015)*

The willingness to conceive Epidemium as an open science program does not only come from the ambition of its two initiators. The idea to tackle the issue of cancer using innovative methods both by working on the program's structure and method, combining openness and big data analysis, was born out of several observations on cancer and the current context. Cancer is a very modern-day issue, it is meaningful and it has the potential of federating a community. In 2012, 8.2 million people died from cancer around the world and this is likely to increase by 70% within the next decade[1]. Each of us is impacted by this disease throughout our lives, whether it be directly or indirectly. Cancer is therefore a major issue for our society, often generating strong emotional reactions and tangible effects. Many patients communities, siblings, or healthcare professionals, have already started playing a role in cancer research and in defending various interests. Epidemium provides the possibility to federate all of those who wish to build a larger movement together.

Epidemium provides players who are usually not solicited at that level the means to assemble, and enables their empowerment by offering them technical solutions: making data accessible and democratizing the processing tools. Open and heterogeneous data are available,

including on intergovernmental websites and in quantities that are large enough to deduce some meaning out of it and conduct research. Moreover, big data has a natural impact on cancer epidemiology. Cancer epidemiology could be one of the fields where it plays the biggest role: it offers data concerning all aspects of society upon the daily lives of individuals and their environment. Complex and full of possibilities, big data requires a significant level of interdisciplinarity to the people who will treat and study data and draw lessons from it. This is why Epidemium is open to all of those who wish to share their skills and expertise, whatever these may be, given that their goal is to make a contribution.

> *With Epidemium, we proved that it was possible to bring together brilliant and motivated people to produce high-quality science and to bring together experts in the Scientific and Independent Ethics Committees, in order to provide guidance and to assess the various projects"*
>
> ***Dr Charles Ferté***
>
> *Member of the Independent Ethics Committee*

Building on the experience of the Challenge4Cancer and convinced of the purpose of a program such as Epidemium, we sought to write this White Paper, that would close this edition and help us conceive the next one. It was build following the same values we favored in the program, integrating a diversity of opinions and disciplines. We took into consideration the recommendations of actors who have had a role in this program: members of the community, members of the Scientific and Independent Ethics Committees, contributors and partners. Therefore, this book advocates in favor of open science and for a methodology based on collaboration.

We present here a composite work, mixing articles - sometimes co-signed - feedbacks and worksheets that all revolve around the themes upon which this program was based: healthcare, open and data. Readers should feel free to read this White Paper from start to finish or to choose to explore topics that they wish to learn more about. From this multiplicity, three parts have emerged. They provide the structure on which this White Paper is based: *An agile and open community*, which presents the general methodology and the experience of the community; *Scientific and medical innovation*, opening the debate on the interaction between data science and medicine, including the benefit for patients that is always a paramount concern; and finally *An open and ethical legal framework*, which explores the issues that are both legal and ethical and that Epidemium has had to tackle when implementing the program. ■

[1] International Research Center on Cancer (CIRC) - World Health Organization (WHO), press release n°223, <www.iarc.fr>, last accessed on November 30, 2016.

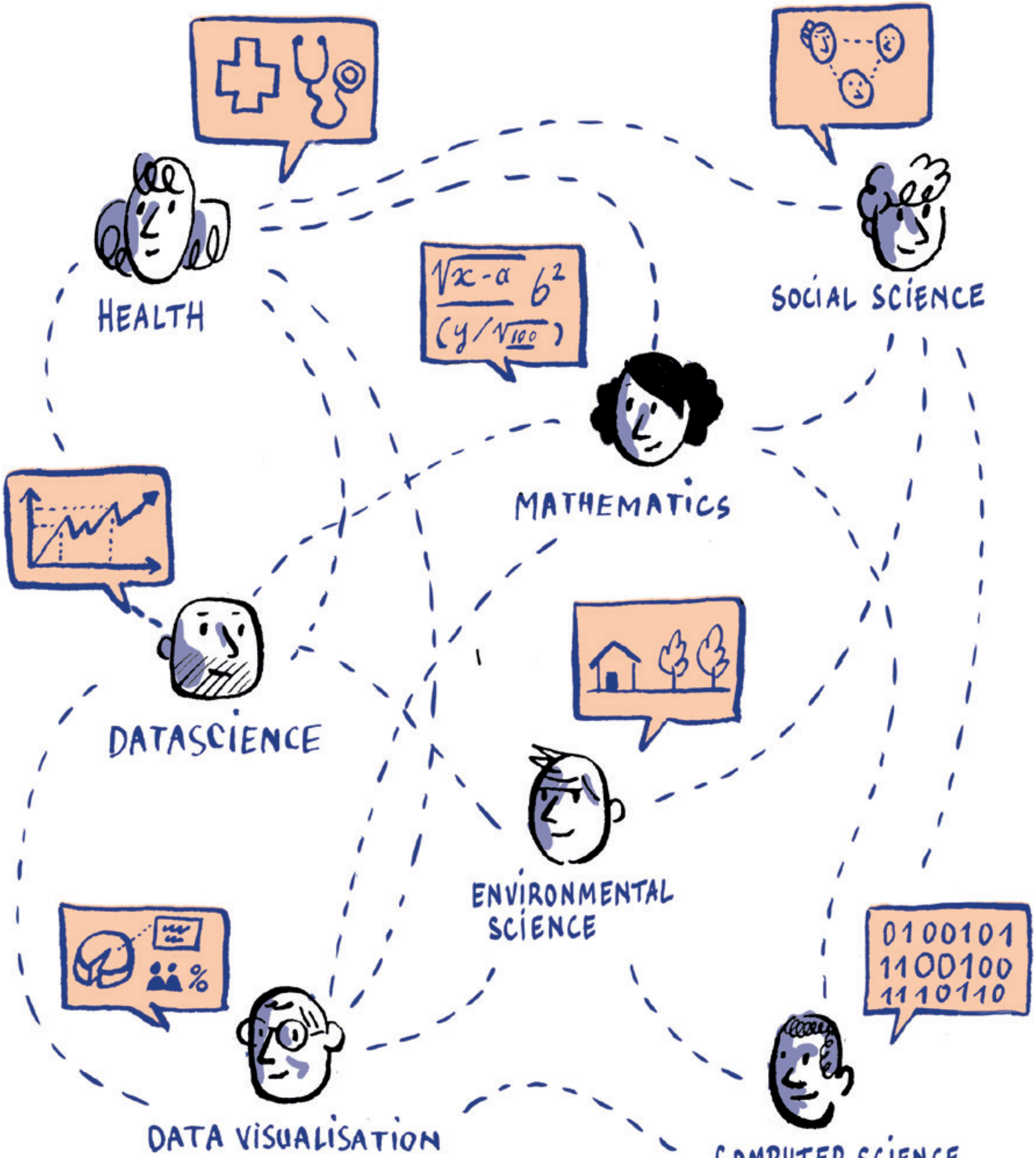
#1

AN AGILE AND OPEN COMMUNITY

Epidemium was only able to take on such an ambitious challenge combining healthcare, data and open science to advance cancer research with the creation and management of an active community of volunteers. What have been the good and bad practices in the organization of this community, whether in the understanding of how it worked or in its animation during the six months of the Challenge? What lessons can the main actors of the program draw to pursue this project with a renewed enthusiasm?

// AUTHORS

Epidemium team / Djalel Benbouzid / Léo Blondel / Marc Santolini
Stéphanie de Haldat / Thomas Landrain


HEALTH
SOCIAL SCIENCE
$\sqrt{x-\alpha}\ b^2$
$(y/\sqrt{100})$
MATHEMATICS
DATASCIENCE
ENVIRONMENTAL SCIENCE
%
DATA VISUALISATION
0100101
1100100
1110110
COMPUTER SCIENCE

Combining cancer and big data using a flexible methodology

EPIDEMIUM CHALLENGE4CANCER (C4C) RAMP

COOPETITION OPEN SCIENCE

How to conduct cancer research using open and big data analysis while adopting an open and community-oriented approach? In this article, we introduce and explain the methodology on which we based Epidemium's program and the implementation of the Challenge4Cancer. With this methodology, we tried to address the issues raised by the matching of the terms health, data and open while trying to promote a growing, trained and dynamic community throughout the six months of the Challenge.

// **AUTHORS**

Mehdi Benchoufi / Olivier de Fresnoye / Karine Lévy-Heidmann
Ermete Mariani / Ozanne Tauvel-Mocquet

#1

AN AGILE AND OPEN COMMUNITY

Introduction

Hosting a community as large and heterogeneous as Epidemium's was a constant challenge for the coordinating team. Since the very beginning, we knew that we had to demonstrate a great deal of flexibility in order to stimulate the creativity of our constantly evolving community, made up of a diverse group of individuals, along with all of the other stakeholders and partners that were collaborating with one another for the first time. Our task was all the more complex since the program's goal was to bring closer together the worlds of healthcare and big data in a spirit of knowledge sharing, and driven by a specific mission: find a new approach to the epidemiology of cancer.

The first step we took was to organize the Challenge4Cancer (C4C), modeled as a real six-month, award-winning competition, but where the values of collaboration and sharing would prevail. In short, and to use a neologism used in the FLOSS (free, libre and open source) movement, this was a *coopetitive*, community-oriented challenge[1]. For this reason, we decided that the C4C would revolve around four main themes, with no pre-determined goal to reach. Therefore, the only constraints that teams had to respect were to follow the ethical and methodological frameworks set up by the ethics and scientific committees, as well as comply with the Rules & Regulations. The Challenge was thus open to every possibility with regards to the nature and purpose of the projects.

We had in mind to remove all possible constraints for participants and make the Challenge as intellectually stimulating as possible. The overarching idea was to maintain a strong level of engagement among participants throughout the entire six months, an engagement that would be mostly taken on their free time. Our vision was to turn their implication into concrete, documented projects, that would then be assessed by a jury composed of the Scientific and Independent Ethics Committees. Yet we had to adapt to a level of participation that was more or less fluctuating depending on the motivation and the availability of participants. Furthermore, by mobilizing a large and heterogeneous community, we had to think about

> *Epidemium, a collaborative research program that is truly unique around the world."*
>
> ***Jean-Bernard Gallois***
> *(01Net, 11/18/2015)*

how to make all projects progress at about the same level, in order to maintain a form of emulation and to address the needs of each participant more efficiently. Technical challenges also arose, linked to constraints of locations and timezones. We wanted to imagine a Challenge that would not be discriminatory against those who were far away in order to welcome participants from all over France, but from abroad as well.

Finally, there were two crucial issues that played a key role in the Challenge's success: the diversity of profiles, and by extension, of skills. How to make the C4C more accessible for beginners while staying attractive to those who were more experienced? How to avoid the risk that beginners become discouraged and the more experienced become jaded? Those with a medical background know the current challenges linked to cancer, the patients' needs, and where research stands; those with a background in data science have the know-how in big data analysis, and have experience processing, interpreting and innovating with data. How to create an environment conducive for people with different yet complimentary backgrounds to meet, so that the final projects would propose comprehensive and feasible solutions?

We have taken into account these various questions in the organization of the program, in the design of the Challenge4Cancer contest and to decide about the tools and resources given to participants.

A community addressing the needs and purposes of the program

1. The committees

One of Epidemium's first step was to establish an Independent Ethics Committee and a Scientific Committee. Committees are necessary in all research endeavors and this was required given the fact that we were initiating a program associating the terms of science, open and big data. These committees were designed so that they would include all of the different sectors and expertise necessary for it to function properly *(see*

> *"Epidemium was a wonderful and multifaceted adventure: it was enriching from both a scientific and social point of view. The fact that the teams were supported by the constant and dynamic efforts of the coordinating team helped a great deal. We used Meetups, RAMPs, and various meetings at La Paillasse to bring together all the various profiles, skills and people. Everyone was working toward the same goal: to conduct cancer research."*
>
> ***Benjamin Schannes, Predictive Approaches and Cancer Risks Project Leader***

***worksheet n°1a** and **worksheet n°1b**)*. We had a few goals in mind: to guarantee the credibility and feasibility of such an initiative, to support an open and enable it to grow, and to innovate in a context of rigorous ethical and methodological framework.

In preparation to the Challenge, the Independent Ethics Committee wrote the Epidemium Charter *(see **worksheet n°3a**, page 137)* setting the boundaries of what was possible so that everyone would respect the rules in terms of use of data within the program. The Committee validated its feasibility by looking into how data was pulled and selected from various databases to be given to participants, as using data to the benefit of the patients necessarily requires to study personal privacy, anonymization and consent questions. The Scientific Committee was in charge of overseeing the quality of what was produced by the community, as no diploma or certificates were required to enter the contest. It made sure that the methodology chosen by the program's coordinating team and the participants was right. During the Challenge, the committees were in charge of making sure every project complied by the rules, of defining assessment criteria, of providing support to projects in the formulation of their hypothesis and finalization work, and of identifying their possible applications and implications, finally composing a jury assessing them. More broadly, they reflected on current practices, the contribution of technology and the use of open data, especially in the field of research.

2. Ecosystem and partners

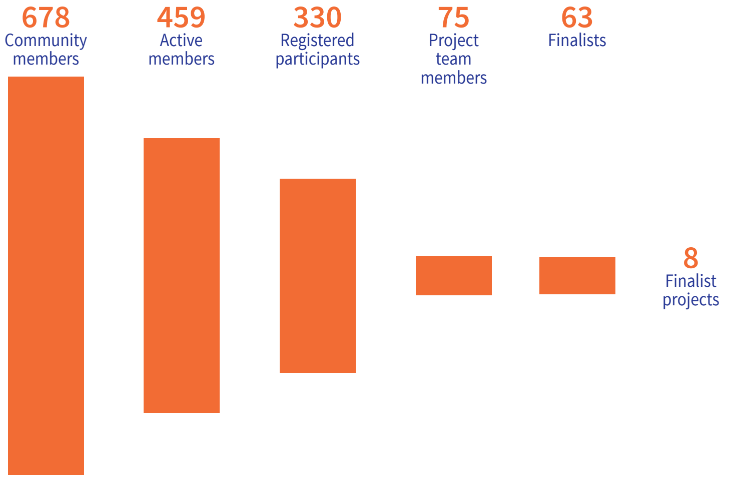
In order to be more than a vain effort, Epidemium sought to create an ecosystem where healthcare professionals, medical researchers, data scientists and open innovators could freely meet and exchange their knowledge as well as their resources and methodologies. This approach allowed us to investigate whether more traditional players were interested in this initiative and to see if they were willing to collaborate with us to help us conceive and define the program, and ultimately to take part in it. These included individuals and organizations that we directly approached or met during the many presentations Epidemium made. They played a variety of roles and helped us spread the program within their ecosystems by accepting to act as community mediators and ambassadors. Being motivated and convinced by the program's approach, they also helped us strengthen our credibility, defend our values and grow the community in the most relevant groups of people. Finally, they brought to the table their expertise, resources and tools and these were indispensable for the success of the program. For instance, by taking part in public conferences held in the context of the Challenge4Debate (C4D), they played a role in managing the community and producing knowledge that was openly shared. The expertise of this ecosystem also benefited the program by providing food for thought, which, in turn, we sought to adapt based upon the needs of the participants.

Technical partners who joined the ecosystem made the Challenge4Cancer possible by ginving access to their tools and by helping participants in the work. This involvement helped to create a high-quality Challenge, capable of producing serious studies with regards to how the data was processed and the methodology that was employed. This type of technical work environment that is usually the priviledge of professional researchers or big companies who can afford it, was optimal for participants. A range of technical partnerships were built with different structures: Hypercube, a structure used to collaborate with with professionals on research papers, that has developed a unique big data analysis tool allowing users to explore, in an exhaustive manner, phenomena whose causes are complex and difficult to predict; Dataiku, that has created a data analysis and data visualization studio equipped with a large range of click-and-go tools designed to prove intuitions and construct hypothesis around datasets; Teralab, a big data cluster conceived to address the immediate needs of researchers, teachers and companies in order to further knowledge and innovations in the field of big data analysis.

3. An interdisciplinary and dynamic community

Diverse in in terms of skills and backgrounds, the Epidemium community is the true source of its wealth. All in all, this community includes all actors who have, at one point or another, made a contribution to Epidemium, including members of the Committees, people who are part of the ecosystem as well as the registered participants to the Challenge. To provide a full picture, the Challenge involved a little bit less than 700

Figure n°1: An efficient, community-oriented effort



Description: *Community engagement conversion rate.*

Source: statitistics provided by Meetup and Epidemium.cc

> *If I had to do it all over again I would not hesitate a second to take part in the Challenge4Cancer contest. I learned a great deal from the challenge and I believe that it represents a new way of conducting research."*
>
> ***Muriel Londres***
> *Member of the Independent Ethics Committee*

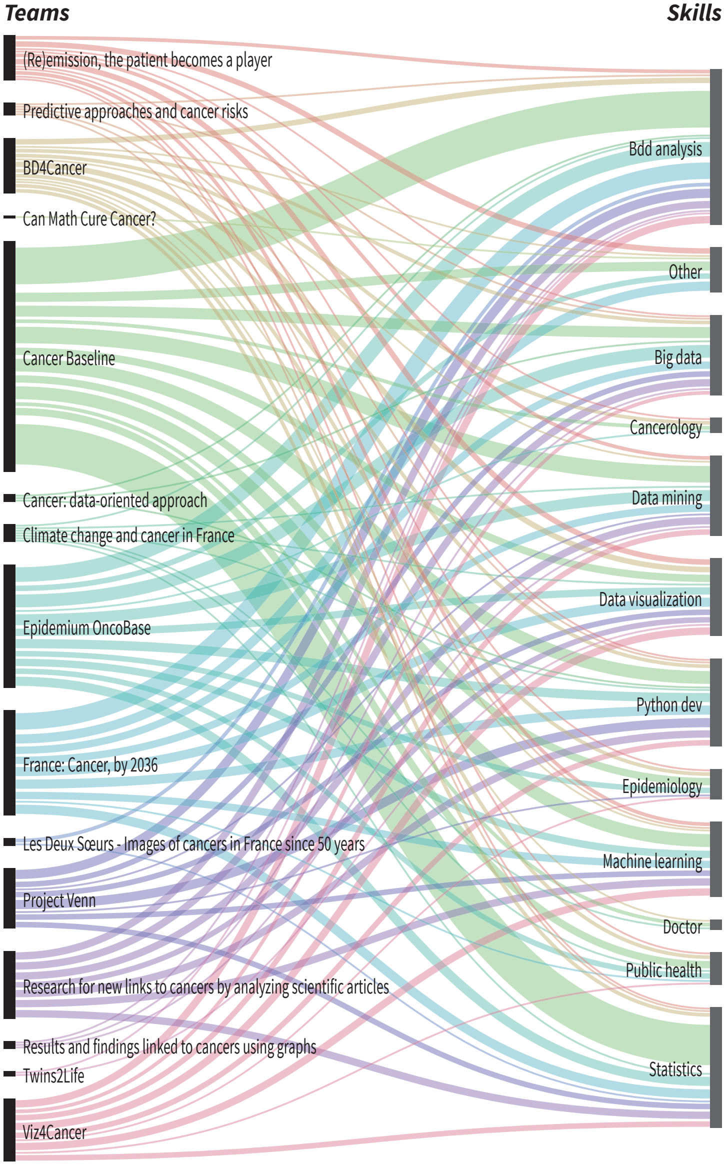
members, 330 registered participants, 75 people who took part in a project, 63 finalists for a total of 8 selected projects in the final round *(see **figure n°1**, page 23)*. These figures demonstrate both the interest that the initiative raised and the level of involvement that it managed to maintain throughout the six months that it lasted.

Throughout the Challenge4Cancer contest, the Epidemium coordination team fostered encounters and collaboration among various groups who came from three main fields: data science, computer science and healthcare. This was done to match the needs and specificities of the Challenge and the topics it dealt with. The C4C being centered around the data, we can acknowledge that this is the reason why the community first attracted data scientists and developed a culture closer to theirs. As a result, the program had to be conceived as an environment where people could learn from this culture's working environment and techniques, especially for the healthcare professionals.

To provide more details on this typology, it should be said that 1 176 skills were accumulated throughout the challenge. By studying the projects and the skills declared in them, we can observe a real circulation of the latter. All of the 15 registered projects have declared having used at least more than one of the 11 key skills, at some point of their development *(see **figure n°2**, following page)*. This enables us to identify the logic of emulation and collaboration that was at play and that was encouraged by the Challenge, thanks to the tools as well as to the numerous encounters and events.

Finally, we can observe that the community has grown throughout the six months, and for every type of skills. This underscores the interest that the Challenge raised among participants and its ability to convince curious individuals. As we have mentioned earlier, the late arrival of healthcare players can be explained by the nature of the Challenge, which required to start with dealing with the significant task of cleaning and aggregating data. Healthcare professionals have mainly taken part in the second phase of the Challenge *(see **figure n°3**, page 26)*, at a moment when their expertise was solicited in order

Figure n°2: Interdisciplinary Teams

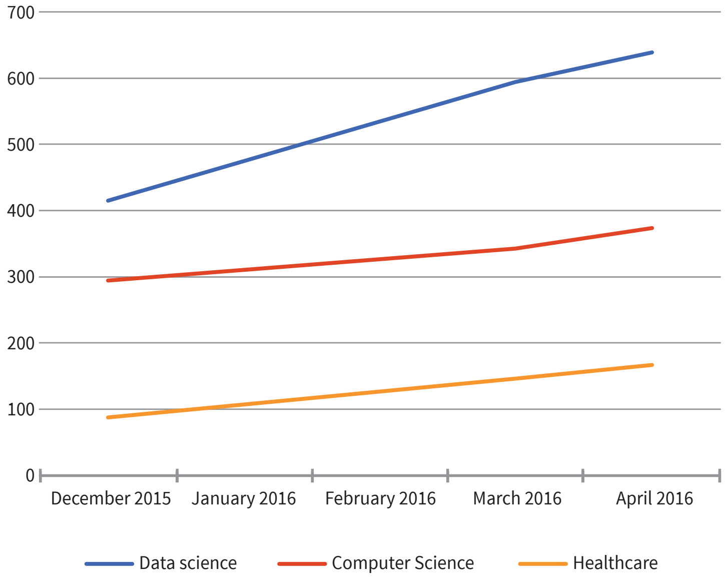


Description: *Skills breakdown within the fifteen project teams.*

Source: Epidemium, visualisation done with <http://raw.densitydesign.org/>

to interpret the first results obtained and to guide medical hypotheses and solutions.

Figure n°3: Areas of expertise and when they were used during the C4C



Description: *Evolution of skills declared on the Epidemium.cc platform during the Challenge4Cancer contest.*

Source: Epidemium.cc

Nevertheless, aggregating a community is obviously not enough; the community has to remain active. In spite of the fact that they all shared the same goal, this group of people still needed to be managed on a daily basis in order to maintain its level of commitment. This is especially true given the fact that they are all volunteers. In light of this, our goal was to create a strong, community-oriented dynamic, leveraging a variety of resources.

Resources to structure, develop and manage the program

In order to make the program known, to engage and manage the community as well as provide it with the means to carry out the projects, the coordinating team has put in place various online and offline tools. Each of them was assigned with a specific role to address the needs that were either identified before or expressed during the challenge by the participants *(see **worksheet n°1d**: Epidemium's toolbox, page 62)*.

If a hierarchy of tools were to be defined, the main online tools of the Challenge that played a structuring role for the community have been the website and the Wiki[2]. The website[3] was the first tool that was conceived before the Challenge and its role was to provide a basis for the community-oriented approach we sought to initiate. It served as a gateway for the participants to the Challenge4Cancer contest. It presented the program, but most important of all it enabled us to manage the community and make it intelligible to everyone. Participants were organized into categories based on a variety of criteria: background, skills, if they were part or not of a project, and in which one of the four challenges. By using such information, a search engine offered users the ability to navigate through the community here gathered and organized online. For example, it was possible to see the projects in each category, to look for or request specific skills and backgrounds in order to carry out one's project, to identify project holders and contact them, etc. Therefore, the website's first goal was to foster interactions between different skills, and to make it easy for those who wished to start or take part in a project to identify them.

The Wiki played an important part given its great modularity. The coordinating team was able to adapt it to the different steps of the program, and cater its needs and those of participants. This is something that is clear when one looks at the following figures: according to the statistics provided by Epidemium's Wiki, from the beginning to the end of the program there, were 3 136 modifications, 10 024 viewed pages and 3 276

> *Cancer & big data: collaborative science is setting up with Epidemium."*
>
> ***Hugo Jalinière***
> *(Sciences et Avenir, 05/30/2015)*

contributions for a total of 117 contributors. It was used as a tool daily and freely. It was accessible by all and anyone could make modifications. It came to be the program's pulse and monitor. It was first used as a means of communication that centralized all of the different information about the program and the Challenge4Cancer, as well as the news. It presented Epidemium and its initiators, its partners, the Committees and the issues they were working on, participation rules, one-time or recurring events, etc. The Wiki was also used to centralize all of the links that redirected users to other tools by explaining how they functioned and could be accessed. Finally, and this is how it was used both by Epidemium's team and its participants: it came to be used as a documentation tool and, as a result, it increased transparency and openness. The entire project was documented by putting in place a weekly log book[4] that included anything important that had happened in the program and the projects, as well as minutes of all events. This approach echoed our willingness to produce knowledge that was freely accessible to all and without any costs. Participants also had the possibility to react to the various contents on discussions windows. Finally, participants were expected to use the Wiki to present and document their projects and their hypotheses. This way, at the end of the Challenge, we froze each project page so that the members of the jury could make their assessments, and the pages included the following information: the project's objectives, the presentation of the team members, the final work submitted, the resources used (datasets, tools, etc.), the methodology they employed, and the future developments they imagined.

Then, Epidemium put in place a series of related tools dedicated to specific community issues, including issues about the program itself. A platform, published in the format of a Q&A[5], allowed members of the community to ask questions linked to the themes, the use of data in cancerology and the methods employed by the Challenge, to which participants, experts and the coordinating team could answer freely. Through the Epidemium group on Facebook as well as its Twitter account, the coordinating team was able to exchange with the broader community on the program's evolution and news, along with the Challenge and the projects. Facebook was intended more for the Epidemium community and Twitter for its larger ecosystem. These two tools played an important role in curating scientific knowledge revolving around the themes of big data and cancer. Let us not forget the use of the account on the Meetup platform[6], that was used to support and promote the events held in the context of the program.

Finally, there were the purely technical tools required for Epidemium and the teams to process the data. Epidemium's Open Data Portal[7], using the open source software CKAN, addressed the technical challenge of making accessible more than 21 000 datasets for the Challenge4Cancer contest through a search engine. It indexes the data based on demographics, environment, agriculture, work, economy, individual behavior, health and cancer, thereby making the large amount of data intelligible, providing a first interpretive framework that made it easier for users to understand. Furthermore, as we have mentioned, various analytical tools were used and provided to participants thanks to the program's partners: a big data cluster by Teralab, a data science analysis tool by HyperCube and a data analysis and data visualization studio by Dataiku.

Far from being a dematerialized program, Epidemium took shape through the organization of many events *(see* ***worksheet n°1e****: Call4Debate 2015-2016, page 63)* with the goal of enabling encounters between participants, fostering synergies and taking the opportunity to exchange with experts. These community events, twenty one in total, were of different formats depending on the targeted audience and the goals they served.

The most widely used was that of conferences, where several experts spoke to present concrete case studies of their past or current work, and to exchange with the public on related topics. These informal conferences played the role of helping participants in the design of their project. All of them were published on the Wiki, which enabled the production of new forms of knowledge that the community could make use of. These events were open to all and free of charge. This was important as we sought to make the program known and to attract a large audience interested in these topics. In total, 926 people attended the conferences.

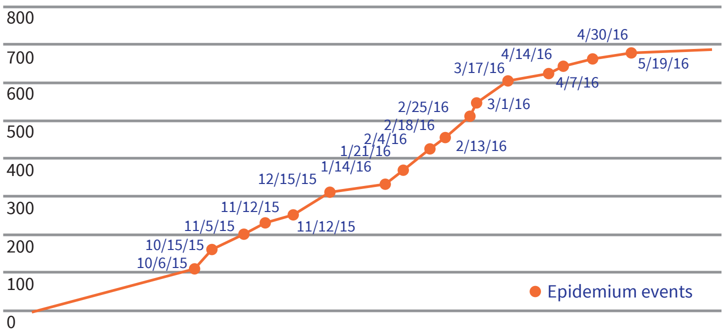
The second format was the RAMPs (Rapid Analytics and Model Prototyping)[8], which are data challenges based on the idea of coopetition. The notion of coopetition echoes the program's mindset, given the fact that all of the tools and knowledge produced by the participants are visible by all. These events lasted one day, and were geared toward data scientists. They were also free and accessible to all, the goal being to create and test prediction models using data collected by Epidemium. Participants who were gathered in the same room, in teams or not, submitted their models on a server in an open approach. Their performance was displayed on a scoring board. All could freely access model codes already submitted and therefore understand those that were efficient, use other elements, combine them, and improve them to submit them once again. This hybrid approach made it possible to accomplish, in a short amount of time, a reliable prediction model that was the result of everyone's efforts.

The third format were the so-called Bocal events. Bocal events were conceived as more restrained events. They acted as one-time events whose goal was for participants to meet up in order to make it easier for everyone to constitute teams, but also to address more specific needs in the presence of experts, especially those coming from the medical field.

Finally, other events took place linked to the conduct of the program. These were important in creating a strong dynamic throughout the six months of the Challenge: the evening launch, which was the very first meeting among members of the ecosystem (partners, Committees members and participants), the mid-term review session where teams were asked to present their projects, methods or hypotheses to the Committees so that they could be offered guidance and support, and the closing evening where the Challenge's final took place, and during which all teams presented their projects in front of the community and the jury. These moments marked the Challenge and were used as milestones for the participants, which helped us sustain the momentum of community-oriented work.

Our various events allowed us to address the issues of being part of a project that lasted a long time, of fluctuating engagement, and of volunteering. Through its event management strategy, Epidemium has sought to create and maintain the interest of participants by offering fun, useful and educational moments. In that sense, we went beyond what we had hoped to achieve. The events we organized represented a real advantage for the program as a whole, which is illustrated by the correlation between the events and the community's growth *(see **figure n°4**, below)*.

Figure n°4: Correlation between the growth of the community and the events



Description: *Evolution of the community with regard to the events organized by Epidemium.*

Source: Meetup statistics

With the tools that were put in place, Epidemium defined itself as a flexible structure, capable of anticipating, identifying and addressing the needs of participants. All of these tools have made it easier for the community, the experts and Epidemium's team to interact. Finally, we observed an interaction between the online and offline formats, with each increasing access to the other.

> *We are extremely proud to be Epidemium's technological partner for this first Challenge. By making our tools available to the consortium with our collaborative platform used for predictive analysis, we had the opportunity to follow-up on and support a large number of incredible cancer research projects. This was both fascinating and inspiring."*
>
> ***Thomas Thus***
> *Dataiku*

The successes and failures of the methodology

The very fact that Epidemium chose to create an open science program on cancerology based on the use of big data analysis was a key success factor. The issues at the heart of the program are contemporary, resonant and known to the public, and sparked the curiosity and interest of many. This made it easier to build a community as well as expand its ecosystem. Since the topics of big data and openness often appeared in the media, Epidemium was able to take advantage of that visibility to bring together the profiles of people it needed for its project. Also, it is something that helped when it came to proving the feasibility of new approaches in this field, especially given the fact that there are still too few existing concrete cases. This was thus an opportunity to position ourselves as a key player of this topic by playing an active role in it through the Challenge, while, at the same time, having experts question it. It is a topic that will certainly become a fundamental issue in the future.

The Challenge4Cancer had no barriers to entry: we asked for no accreditation, technical means were provided and the program offered a framework and an ecosystem that were safe and educational for all participants, even for beginners. More so, the idea of openness structured the entire Challenge, respecting the choice of total transparency and intelligibility for all. This was reflected in the documentation effort, which was made accessible to everyone. Furthermore, working on open and accessible data is what made the program feasible. From a legal perspective, this facilitated its implementation since the selected pool of data that was already accessible and available, respected French law. The pool of data was large and diverse enough to be used as a rich and promising source of research material. Given the fact that anyone could freely access it, the data was a common good that anyone could make use of. In that sense, Epidemium offers us the possibility to truly claim ownership of our own data in a collaborative manner.

Beyond the aspect of taking part in an open initiative and of each partner sharing a common cause characterized by a real social and scientific interest, the program included people with a variety of motivations that were as diverse as its different stakeholders. In spite of the diversity of profiles, including data scientists, doctors, employees, students, job seekers, etc., we can emphasize the shared motivations that encouraged each individual to take part in the Challenge.

It was an opportunity to discover and gain insight in new topics, to build new skills, but also to deepen those already acquired while putting them at the benefit of a shared cause, to test them in the context of a real case, to participate in a collaborative experience, and to meet new people coming from diverse backgrounds, including in the fields of open collaboration, healthcare and data science. Students, who represented about a fourth of all participants, were particularly enthusiastic about the program.

Epidemium's first edition, which truly ends with this White Paper, allows us to test the feasibility of an open and community-oriented approach to address the issue of cancer from a different angle. With this project, we were also able to gauge the growing interest among accredited players coming from the worlds of medical research and big data, public institutions, along with the numerous individuals who identified with our values and our mission.

Staying true to our commitment, we will take into account the feedback of our ecosystem and the lessons drawn from the community that we had the pleasure of coordinating. ■

[1] Coopetition: neologism based on the combination of the terms competition and collaboration. The term captured the essence of the program.
[2] Epidemium's Wiki <http://wiki.epidemium.cc/wiki/Accueil>.
[3] Epidemium's website <http://epidemium.cc>.
[4] Epidemium's log book <http://wiki.epidemium.cc/wiki/Carnet_de_bord>.
[5] Epidemium's Q&A <http://qa.epidemium.cc/>.
[6] Epidemium's Meetup <http://www.meetup.com/fr-FR/Epidemium/>.
[7] Epidemium's Open Data Portal <http://data.epidemium.cc/dataset>.
[8] RAMP (Rapid Analytics and Model Prototyping) is a tool developed by the Paris-Saclay Center for Data Science and the Ecole des Mines, conceived for the management of data challenges, <http://www.ramp.studio/>.

Measuring the program

DATA SCIENCE SOCIAL NETWORKS COLLABORATION ANALYTICS DIGITAL FRAMEWORKS

Thanks to technological advances, it is now possible to analyze quantitatively the activity of teams working within a collaborative framework. In this study case of the Challenge4Cancer organized by Epidemium, participants had access to a number of online tools. Therefore, it is the analysis of activity patterns of users on those different tools, such as the website, the wiki, the Q&A and the Meetup platform, that allows us to frame the member's behaviors within the Challenge. Finally, this allowed us to measure engagement and frame recommendations for the future programs.

// **AUTHORS**

Djalel Benbouzid / Léo Blondel / Marc Santolini

Epidemium's coordination team provided the participants with multiple online tools during the Challenge4Cancer (C4C) in order to create a virtual collaborative work environment. The goal was for the members to be informed about the program, the Challenge, and similar thematics, but also to interact with each other in order to achieve successful projects.

The analysis of the data gathered for each tool allowed us to frame the behaviors of the C4C participants, to understand the patterns of collaborative work within the teams, but also between different teams, and finally, to observe the temporal evolution of the community at large.

Our analysis is based on data collected on the Epidemium Meetup group <www.meetup.com/fr-FR/Epidemium> along with the three main tools utilized by the C4C community:

1. The main website, opened on 5th November 2015, allowed for people to register to the C4C and inform themselves about the latest news;
2. An open wiki, created on 1st October 2015 and opened to the public the 5th November 2015, here to document and share with the community the current state of each projects discoveries;
3. An open Q&A, opened on 23rd February 2016, for the participants to ask and answer questions with the added benefit of a wider community.

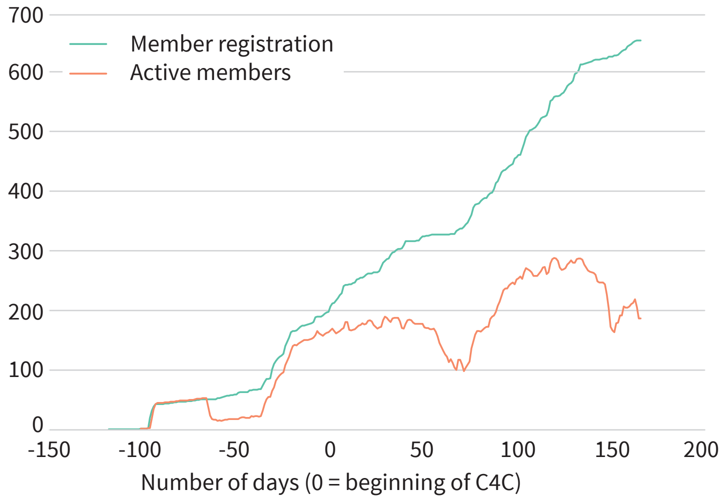
All the collected data and algorithms used in this article are available on GitHub <https://github.com/Epidemium/LivreBlanc>.

A very active community

The first data analysis level we performed looked at individuals' behavior using the online tools put in place during the C4C. At first glance, we can have an accurate observation of the members behavior and measure their level of engagement and usage of the different online tools.

Firstly, in ***figure n°1*** *(see page 36)*, we can observe the temporal evolution of the member's activity on the Epidemium Meetup group from its creation, on 9th September 2015, to the end of

Figure n°1 : Temporal Analysis of member's activity - Aggregated data



Description: *Temporal analysis of the Meetup platform activity. The green curve shows the total number of registered members, and the orange curve corresponds to the number of active members smoothed over the last 30 days.*

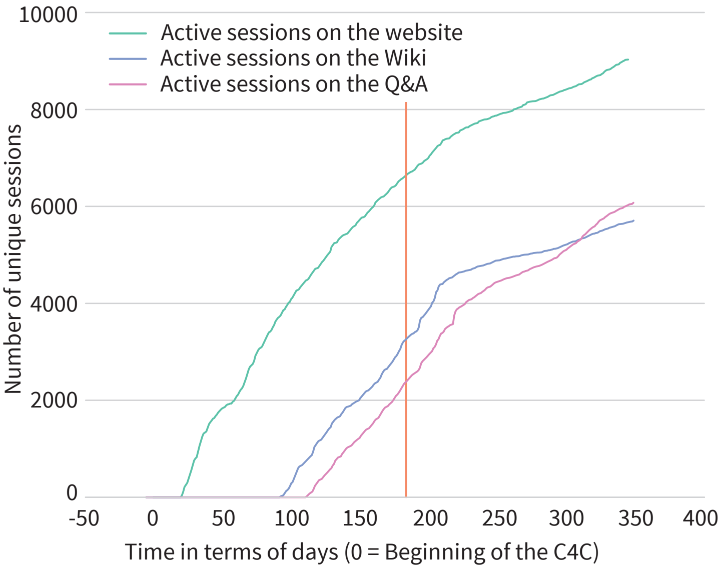
Source : Meetup.com, Epidemium group; figure created by the authors

the Challenge, on 5th May 2016. The number of registered users, here in green, underwent a stable increase for the duration of the Challenge, with a gap which corresponds to the Christmas holiday. Observing the number of active members, orange line, that is users who visited the Meetup group at least once during the last 30 days, we can see two phases emerging. The first one starts just before the beginning of the C4C, and the second one sees a rebound just after the Christmas break. During the later phase, we can observe an increase of about 50% in member's activity. This trend was reinforced by the mid-term session, organized on 12th March 2016 and opened to all teams that were willing to participate. This meeting offered teams a space to confront their project, ideas, methodologies and results with the other teams, along with the members of the ethical and scientific committees. They were required to create a written documentation of their project, with a public presentation of the progress of their research.

Thus, ***figure n°1*** shows a steady recruiting trend of new members into the community to participate to the C4C. Moreover, the registration curve shows no sign of saturation, which implies that the number of individuals willing and interested to participate to the Challenge was probably greater than the one reached.

Next, on ***figure n°2***, we can observe the temporal evolution of the three online platforms installed for the community (website, Wiki and Q&A). Those measurements allow us to analyse the

Figure n°2 : Temporal Analysis of member's activity on the three *online* platforms



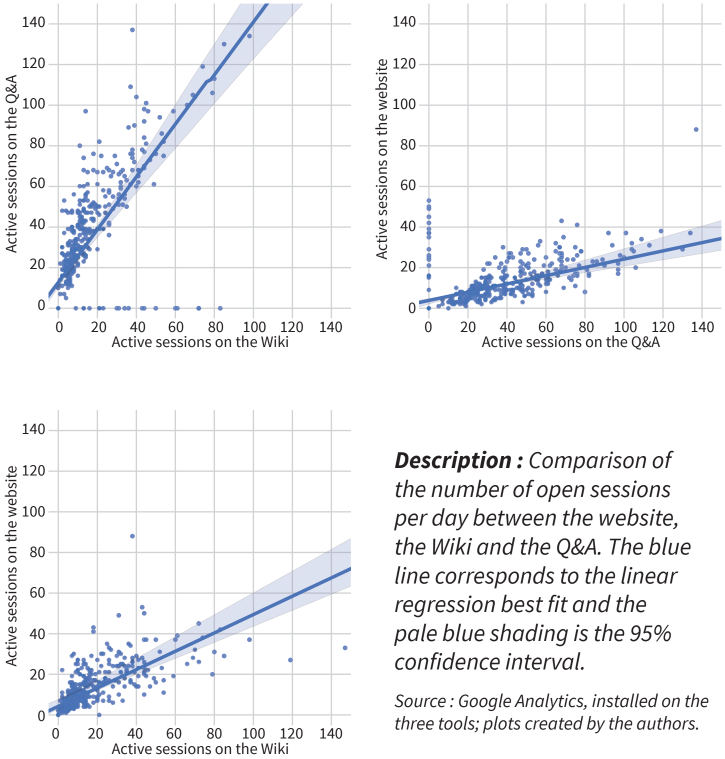
Description : *Temporal evolution of the open sessions on the three online platforms (Wiki, website, Q&A) offered by Epidemium to the participants. The vertical red line corresponds to the end of the C4C.*

Source : Google Analytics, installed on the three tools; plots created by the authors.

trends from the start of the Challenge until 20th July 2016. The shift observed on the three curves is due to the fact that the analytics engine was installed and started at different times on each platform. Indeed, the data collection started on 14th November 2015 for the website, 2nd February 2016 for the Wiki, and 23rd February for the Q&A. We can see that the usage of the three platforms has been similar during the Challenge, with similar slopes (rate of use). Moreover, there is clear decrease in usage after the end of the C4C (red line). Nevertheless, steady traffic continued for up to three weeks after the end of the Challenge, showing a large public and community interest for the results of the Challenge.

Finally, the panels on ***figure n°3*** *(as shown beside)* show the usage correlations between the three different tools during the Challenge, from its start on 5th November 2015, until the day after the end of the Challenge, 6th May 2016. When a user arrives on one of the platforms, a session is opened on Google Analytics, allowing us to track users across the three tools: it is thus possible to know which platforms were mostly used and the relationship between them. Here we analyzed the daily aggregated data of sessions on the tools. Each user counts as one regardless of how many times he used the tools that day. The numbers of visits are then compared between each other. We can observe a very strong correlation between the three different tools, hinting that external users and participants of the C4C used about equally all available tools. Moreover, this shows that the three tools were not only useful and non redundant, but exist in a complementary setting that answers different needs of the participants. Some of which were anticipated before the Challenge, and others that emerged during.

Figure n°3 : Activity correlations between the website, the wiki and the Q&A.



Description : *Comparison of the number of open sessions per day between the website, the Wiki and the Q&A. The blue line corresponds to the linear regression best fit and the pale blue shading is the 95% confidence interval.*

Source : Google Analytics, installed on the three tools; plots created by the authors.

Team building and work methodologies

During the Challenge, team building was completely spontaneous and self-organized. Indeed, the Epidemium coordinating team refused to impose any particular criteria in order not to constrain the innovation potential of the community. Therefore, neither a minimum number of team members, nor obligatory skills quotas per team were imposed, as much as participating in multiple teams was not forbidden.

Measuring the program

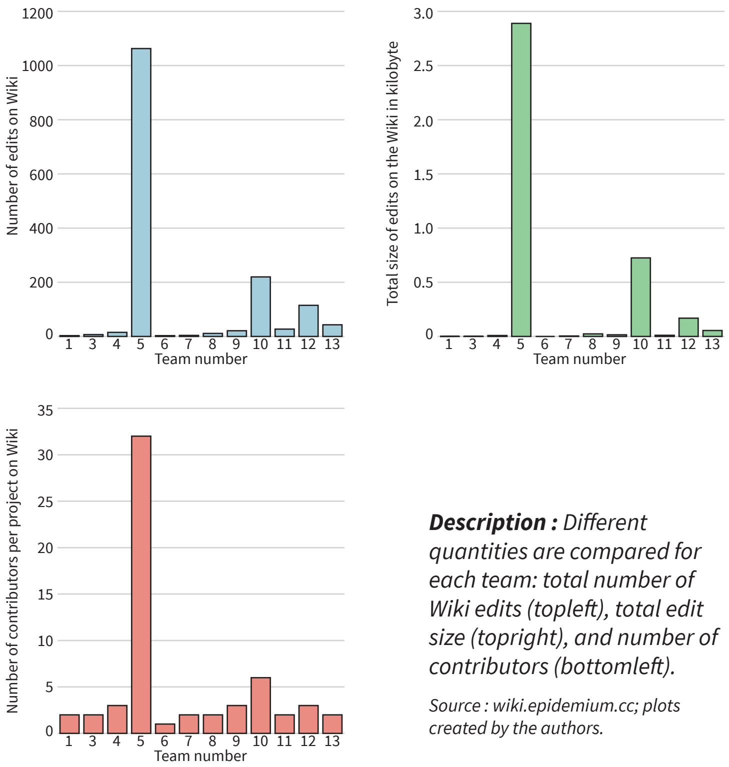
Though, it is by design that the program, looking into the epidemiology of cancer with the alternative methodology of big data analytics, created collaborations between experts in normally very remote fields of study.

Here we present this collaborative dimension of the Challenge and reveal the different aspects of temporal group dynamics and internal team structures.

Team building

At first glance, it is clear that those three plots show that one team mobilized many more contributors than the other ones.

Figure n°4 : Analysis of the Wiki's editions by the C4C teams



Description : *Different quantities are compared for each team: total number of Wiki edits (topleft), total edit size (topright), and number of contributors (bottomleft).*

Source : wiki.epidemium.cc; plots created by the authors.

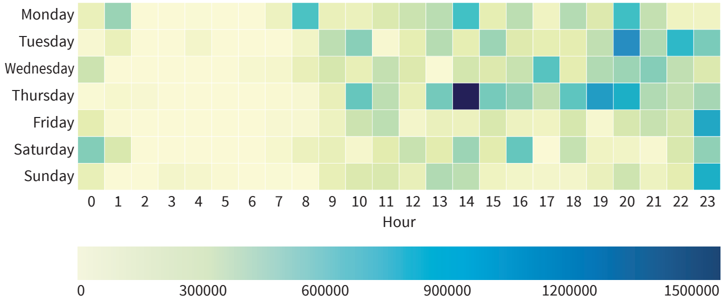
Analyzing the content of the wiki pages which, in the setting of the Challenge, are used to centralize the team's work and discoveries, we can observe that four teams produced the majority of all edits, in numbers as in size.

The format of the Challenge thus leads to a gathering of work forces in a few very productive teams instead of the multiplication of small teams independant from each other. Let us underline that the productivity discussed here is measured by Wiki edits, and does not necessarily reflect the true amount of work performed by a team. Some teams produced codes on Github, created Jupyter[1] notebooks or used other external tools than those provided by Epidemium. Thus, our analysis only shows the tip of the iceberg, and cannot reveal the total productivity of all the teams.

Skills analysis and collaboration dynamics

As can be observed in ***figure n°5*** *(see below)* the community engagement, measured by the activity of participants, has been relatively uniform throughout the week. At the daily level, we note that there is a peak on Thursdays around 2pm, and a general activity increase during the afternoons and evenings, corresponding to periods of free time or lunch break. Although

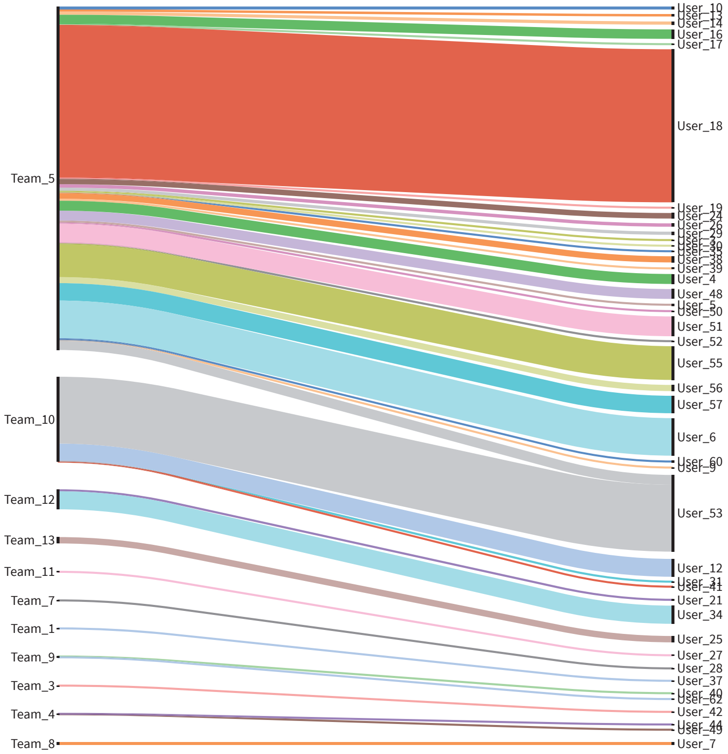
Figure n°5 : Temporal breakdown of members activity



Description : *Heatmap showing the number of edits per day of the week across all teams.*

Source : wiki.epidemium.cc; figure created by the authors this was not explicitly ruled out in the Charter, we can see by looking at the members repartition across teams in ***figure n°6*** *(see below)* that very few participants have contributed to multiple projects. This concentration of individual efforts on single projects has probably reduced energy dispersion, but it might also have limited the development of cross-disciplinary projects involving different teams. Scaling such a challenge

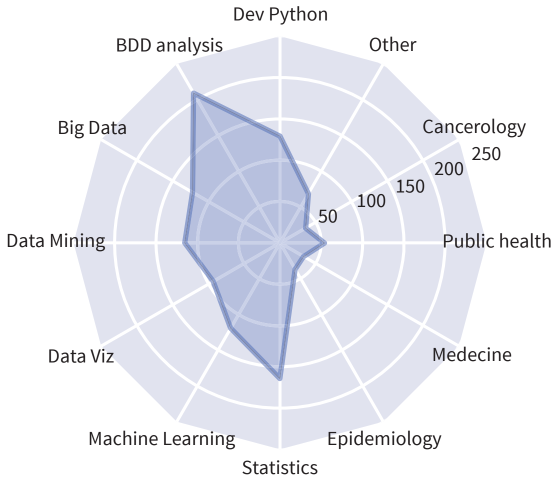
Figure n°6 : Distribution of members across teams



Description : *Sankey diagram showing the distribution of members across teams.*

Source : wiki.epidemium.cc, graph created by the author

Figure n°7 : Skills distribution across teams



***Description :** Radar representation showing the distribution of individual skills across teams.*

Source : epidemium.cc, skills declared by the participants at the moment of their registration; figure created by the authors

will require to think of systematic ways to improve the porosity and cooperation between teams to join forces and synchronize different projects around a common purpose.

Finally, ***figure n°7*** *(as shown above)* provides us with a clear picture of the rich interdisciplinarity that characterized the community. It also demonstrates that the participation of people coming from the world of medicine and health studies should be increased, as the topics of medicine and big data were given equal importance in the program.

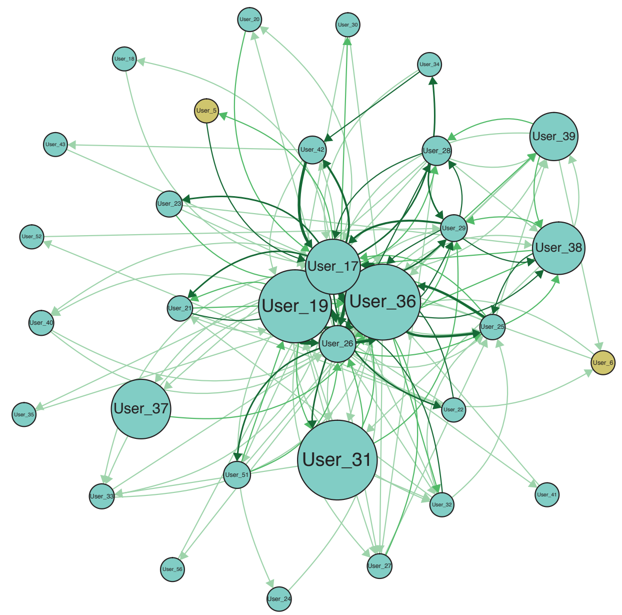
Structure and working dynamics of teams: Analysis of the Wiki edits of team n°5

Here we present the results from the Wiki data analysis for team n°5, the most active team during the Challenge. The team also possesses the most data in terms of Wiki edits.

Measuring the program

We show in ***figure n°8*** *(see below)* the conversation networks arising from the team 5 Wiki project pages. These networks are built by drawing a directed link between two users when they have made edits right after another. The direction goes from the predecessor to the follower. The darker shade of a link indicates a higher number of times a succession has been observed. Node size is proportional to the number of edits for a given user. Many members have made contributions

Figure n°8 : Interaction network of team n°5 members arising from Wiki edits



Description : *An arrow points from user A to user B if B has edited the Wiki after A. Link weight, corresponding to the amount of times B has succeeded to A, is shown with darker color. Node size is proportional to the total number of edits of the user. Yellow nodes correspond to Epidemium organizing team members.*

Source : wiki.epidemium.cc; graph created by the authors to the project's Wiki page, resulting in a particularly dense collaborative network. Within that team, a group of individuals stand out as being more central in the network. They seem to have played a greater role in the management of the project. This network effect can be quantified by looking at the size of nodes in ***figure n°8***. This measure quantifies the importance of a node by summing the total weight of links that connects it to other nodes in the network. This makes it possible to distinguish a form of leadership, if not of the project itself, at least in the edition of the project.

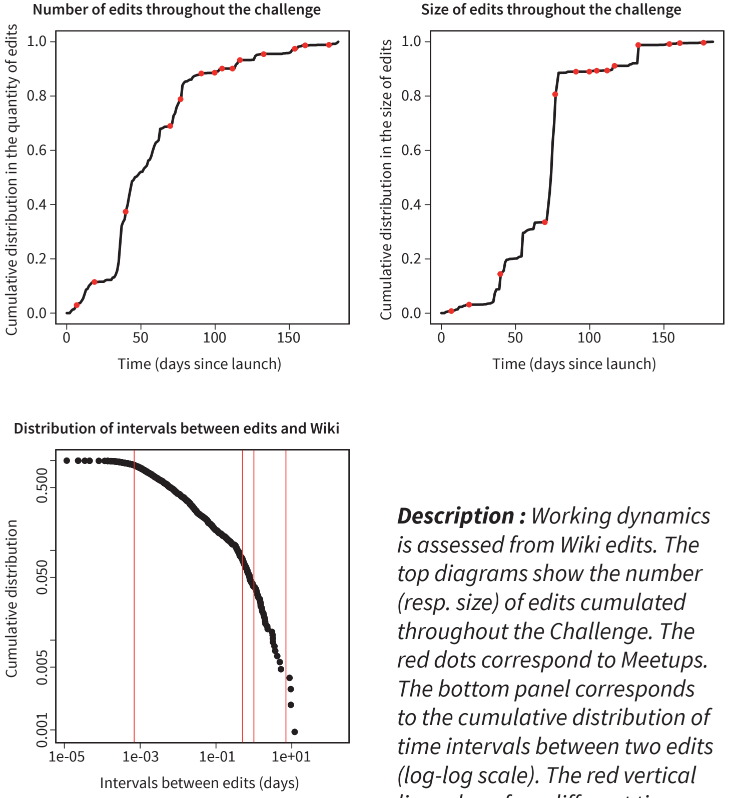
Figure n°9 *(see page 46)* shows the detailed temporal analysis of Wiki edits. The first two graphs show the cumulative distribution of the number and the size of edits in time. An edit is any modification that has been submitted by a user. The edit size corresponds to the quantity of text (in bytes) that has been submitted. The red dots highlight the Meetup events that were organized by the Epidemium organizing team. For team 5, we observed that the majority of edits were done in a relatively short period of time. Moreover, the largest edits generally precede a Meetup, which means that they were subject to a deadline. Finally, the last plot shows the distribution of the time intervals separating two edits. This gives us a hint on how teams collaborate. In particular, when plotted in a logarithmic (log-log) scale, a linear tail is indicative of a bursty working behavior - we call this kind of distribution “scale-free”. Otherly put, while most edits follow each other in bursts of activity, occasionally a long time separates two edits. The red lines show the time intervals corresponding to a minute, half a day (12pm), a day, and a week. In the case of team 5, we observe bursts of activity during a day (the period before the second red bar), and bursts of activity at the level of the week (the period after the second red bar). The different slopes indicate two types of dynamics: one is linked to the Wiki edition during a working day, the other to the time interval between two working sessions, which can span weeks.

These results generalize to other teams with enough Wiki edits to be considered *(**figure n°12**, see page 48)*. We first note that the Epidemium coordinating team should be considered as

Measuring the program

a separate case, since its Wiki edits include specific content related to program documentation and information about the program structure. This results in a steady edition throughout the Challenge, and a working dynamics similar to team 5. This highlights the continuous effort in managing the Wiki, allowing for a favorable and structured work environment. As for the other teams shown in ***figure n°12*** *(see page 48)*, team 11 and team 12 show edition periodicity (bumps in the right panel) whereas team 13 shows a bursty dynamics similar to team 5. As

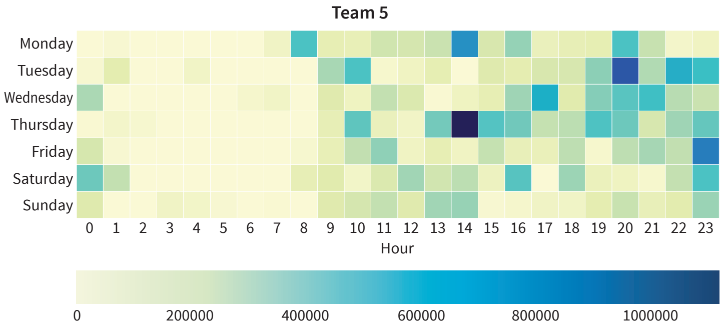
Figure n°9 : Working dynamics of team 5



Source : wiki.epidemium.cc; figures created by the authors

Description : *Working dynamics is assessed from Wiki edits. The top diagrams show the number (resp. size) of edits cumulated throughout the Challenge. The red dots correspond to Meetups. The bottom panel corresponds to the cumulative distribution of time intervals between two edits (log-log scale). The red vertical lines show four different time scales: minute, half a day (12h), day, week.*

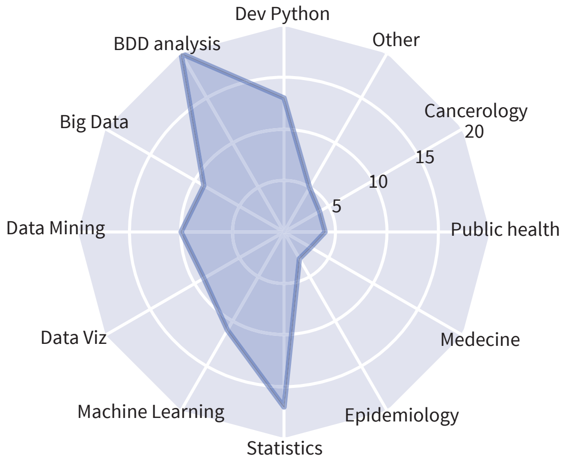
Figure n°10 : Temporal breakdown of team 5 members activity



***Description :** Heatmap showing the number of edits per day and per hour for members of team 5.*

Source : wiki.epidemium.cc; figure created by the authors

Figure n°11 : Skills distribution within team 5

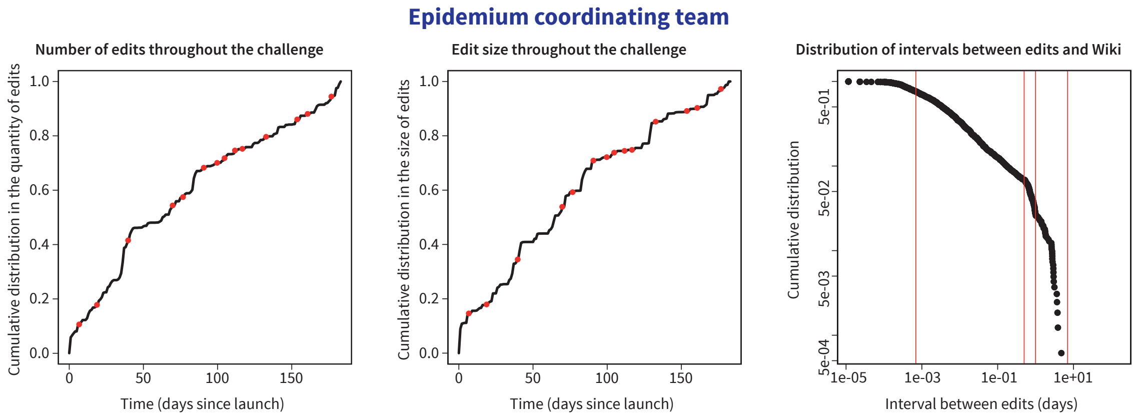


***Description :** Radar representation showing the distribution of individual skills within team 5.*

Source : epidemium.cc, skills declared by participants at the moment of their registration; figure created by the authors previously shown, we observe a meetup effect with a marked acceleration after the mid-term session. Conceived as a first milestone for projects, this session proved highly important in giving pace to the six-month Challenge.

In conclusion, the community saw the emergence of collaborative and productive teams for which team 5 is the symbol. Two types of working dynamics were observed across teams: a bursty Wiki edition dynamics, typical of a self-organized work without prior global synchronization (team 5 and 13) and a more periodic working dynamics, probably indicating pre-arranged meetings (team 11 and 12). The role of the events organized by Epidemium, and in particular the mid-term session, is significant and has contributed to the continuous engagement and productivity of participants throughout the challenge.

Figure n°12 : Comparison of temporal edition dynamics between the Epidemium coordinating team and teams n°11, 12 and 13

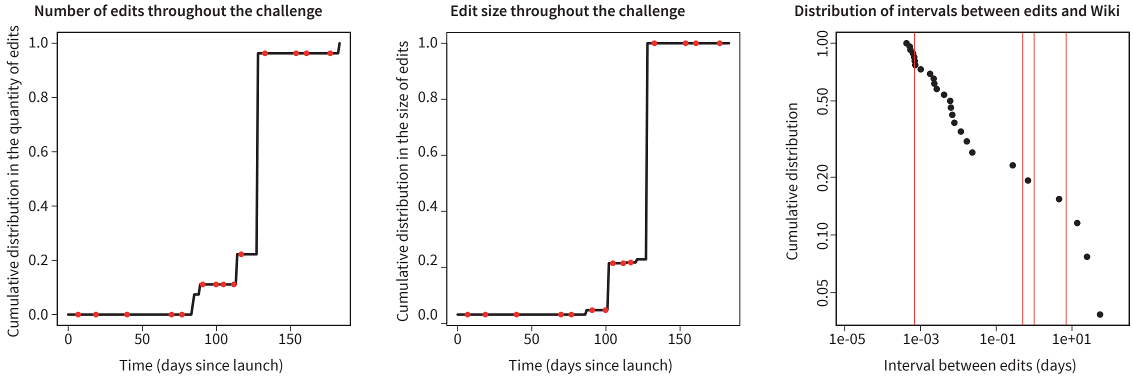


Description : *Diagram showing the temporal edition dynamics. The diagrams on the left and in the middle show the number and size of edits cumulated throughout the Challenge. The red dots correspond to Meetup events. The right panel corresponds to the cumulative distribution of the time intervals between two edits (log-log scale). The red vertical lines indicate different time scales: minute, half a day (12h), day, week.*

Source : wiki.epidemium.cc , graphs created by the authors

Next on page 49...

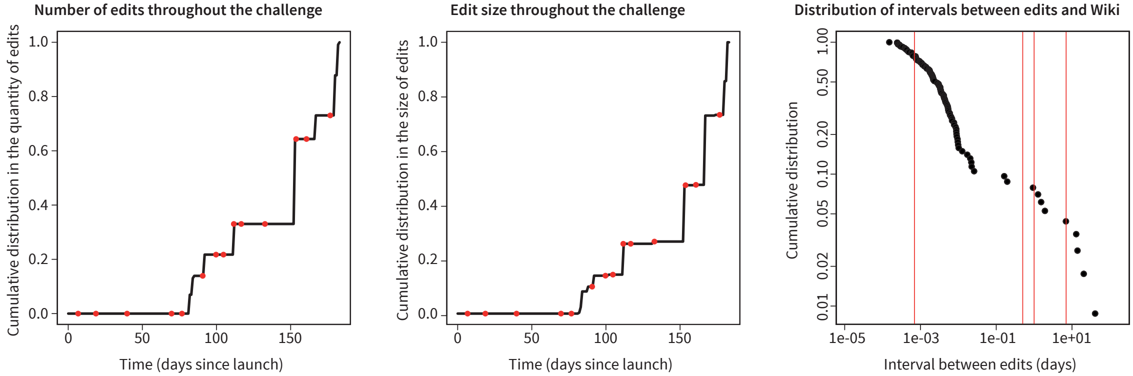
Team 11


Number of edits throughout the challenge

Edit size throughout the challenge

Distribution of intervals between edits and Wiki

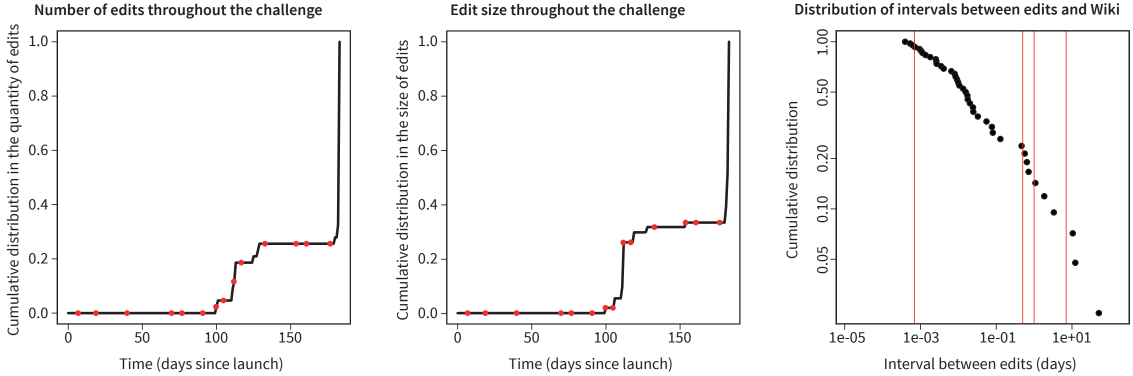
Team 12


Number of edits throughout the challenge

Edit size throughout the challenge

Distribution of intervals between edits and Wiki

Team 13


Number of edits throughout the challenge

Edit size throughout the challenge

Distribution of intervals between edits and Wiki

Discussion

Based on the constant increase in the number of subscriptions to Challenge4Cancer (C4C), we have been able to observe the positive effect of continuously activating the community. This observation is further confirmed by the activity on the different online tools provided to the contributors; it essentially expresses a high interest on the topics of cancer epidemiology and big data technologies, as well as the strategic efficiency of the coordinating team. The community's commitment has been particularly strong during the two synchronization points: the kickoff of the C4C and and the mid-term sync event. Furthermore, the community took advantage of the various tools that we have been studied so far, ie., the Wiki, the Q&A, and the website. The analysis of the Wiki data unveiled a team-based self-organization, wherein different working methods have been adopted. One team, having a large number of contributors (more than 30 overall) emerged: they showed different degrees of internal commitment and were structured around a small central core group. Other smaller teams, however, chose a simpler organizational scheme, with only one person editing most of the Wiki (data not shown). As for the temporal dynamics within the teams, two patterns emerged distinctly, depending on whether the team members collaborated periodically or in a self-organized fashion, ie., with no predetermined temporality. Finally, the Epidemium coordinating team managed to continuously animate the community, by structuring the Wiki content and organizing the different events. This strong supervision shaped the way the community worked together: the events essentially served as deadlines and reignited the engagement of the passive users who were still waiting for an entry point to start contributing.

This Challenge represents an unprecedented event in the open science world. It provides a proof of concept and a potential basis for other similar programs, as well as a roadmap for future improvements. In particular, we strongly recommend to provide the participants with an ecosystem of interconnected tools, in order to conduct real-time analyses of the collaboration within the teams. Such an ecosystem would not only benefit the coordinating team but also empower the participants by making their work and communication more efficient. The coordinating team would then adapt their actions to the needs and dynamics of the community, and the participants would benefit from a better overview of the ongoing interactions, fostering their commitment. Finally, this ecosystem would be necessary in order to scale-up the program to even more participants, which in turn would be a new proof of concept as for animating a massive and open community. ■

[1] Interactive tool allowing to present a comprehensive experimental plan by combining executable code, documentation and interactive visualizations.

Roche's commitment

// AUTHOR

Stéphanie de Haldat

As a subsidiary of the 5th biggest global R&D investor across all sectors, Roche France couples the Group's research model by forging partnerships with tomorrow's healthcare players. Two main convictions are at the heart of Epidemium's project: 1. finding solutions to future healthcare issues requires to approach things differently ; 2. open and big data analysis both represent a unique opportunity for the field of epidemiological research.

Our collaboration with La Paillasse was our very first opportunity. For a major group in the pharmaceutical industry, the ability to work with an open and community-oriented laboratory is far from being a given. However, this first experience was proven to be a success and we have learned a great deal from it.

The project was particularly popular among our employees: a dedicated project group made up of 10 people, 24 ambassadors, more than 20 collaborators involved, etc. At least 50 Roche employees were solicited and invested their time in the project.

When Epidemium was launched, we could feel and grasp the genuine enthusiasm everyone had for the project.

In spite of all of this, we were surprised to see that proportionally only a few of our employees used the collaborative platform. There are a few explanations to this: for non-specialists it was hard to apprehend what was perceived as complex topics such as big data and machine learning, and of course, it was also difficult to find time for those who already had busy schedules.

In light of this past experience, we recommend the following for the next project:

- Train our non-specialists by providing them with accessible information on the topic of big data or through workshops;
- Aim to set up Roche teams within Epidemium and make sure that employees have enough time to participate in the project.

From an organizational point of view, a person was appointed to represent Roche within Epidemium's project team. This person was our main point of contact when it came to dealing with our partner at La Paillasse. We have also created a dedicated project team at Roche with access to an open database, which was an important component for Epidemium. This team was made up of around ten people (doctors, legal experts, biostatisticians, communication specialists, etc.) under the responsibility of Epidemium's project manager.

The collaboration between a major pharmaceutical group and an open and community-oriented laboratory was successful in that enables new forms of approaches and processes. The working processes that we use at Roche are very solid and address the legal requirements that apply to our industry. These processes are also more complex and longer than those used by a project team acting like a startup, which tends to be more agile. This way, Epidemium's project team was able of very quickly referencing 21 000 open data sets. Among other examples, Epidemium's project team was able to create the challenge's infographics quite fast through the use of open source tools. This is an approach we could potentially use for the delivery of our services.

> *Epidemium was capable of combining the creativity of the open science model promoted by La Paillasse with the global e-healthcare objectives in cancerology, Roche"*
>
> ***Ewen Chardronnet***
> *(Makery, 05/31/2016)*

Roche's commitment

In terms of collaboration and working tools, this project taught us how to work differently and using new tools: Slack, Trello whose interest seems obvious for this type of decentralized project.

Our collaboration with La Paillasse has shown similarities in terms of approach and conviction. We both shared the willingness to drive science forward for the benefit of patients, a passion for research, and the need to apply scientific rigor.

Finally, we were able to experiment to what extent the variety of profiles, coupled with collective intelligence, can create value. Diversity and inclusion are aspects that Roche has been seeking to develop for a few years in order to apprehend this type of project. ■

Lessons drawn for La Paillasse

// **AUTHOR**

Thomas Landrain

For La Paillasse, Epidemium was a life-changing opportunity that had to be seized. For over a year we worked extremely hard with our partner Roche to demonstrate that scientific research, for a topic as complex as cancer epidemiology, could be made in an open, cooperative and distributed manner.

Here are the most important lessons drawn from our collaboration with Epidemium

- Open research models represent a major interest and are becoming more and more essential for major healthcare players such as Roche.
- The importance of what we call the "black matter of science" by analogy to what constitutes the majority of the universe and for which traditional matter does not interact with: here, seldom shared data and the time, experience and skills people possess outside of the realm of academia. Such an inclusive approach reduces the distance between the various

> "*It is a completely new approach to epidemiology, which requires the involvements of numerous experts, including data scientists and mathematicians. This is an expertise that we do not possess internally."*
>
> ***Juliette Raynal***
> *(Industry & Technology, 07/08/2015)*

Lessons drawn for La Paillasse

stakeholders involved in the process and allows them to communicate directly with one another. This would have never been possible if not for the transparency and smoother exchanges that the fully open approach enabled. All of the resources offered and produced within Epidemium was entirely open and accessible by all.

- Thanks to the coopetitive environment, most of the participants organized themselves in the form of inter-team cooperation, and the majority of contributions were geared toward the building of knowledge, services and tools that had an interest for the current and future community.

These points give us a clue as to how science without an operator would function. ■

The 6 worksheets

The Independent Ethics Committee

Gilles Babinet	Entrepreneur, *Digital Champion for France.*
Jérôme Béranger	Researcher (PhD), Scientific expert in big data, Information Systems, Ethics and Compliance at Keosys.
Emmanuel Didier	Statistician, PhD in the Socio-economics of Innovation and Professor at the ENSAE.
Muriel Londres	E-patient, Assistant Coordinator at *[im]Patients, Chroniques & Associés* group of associations for chronic diseases, militant and volunteer at the *Vivre Sans Thyroïde* association.
Dr Cécile Monteil	Pediatrician and Emergency Doctor, Medical Director at Ad Scientiam and Founder of the Eppocrate community.
Pr Bernard Nordlinger	Digestive and Oncological Surgery at Ambroise Paré Hospital and Member of the National Academy of Medicine.
Dr Jean-François Thébaut[1]	Cardiologist and Board member of the Haute Autorité de Santé (French National Authority for Health).
Me David Simhon	Health Lawyer and President of the Île-de-France III People Protection Committee.
Pr Cédric Villani	Mathematician, Professor at Lyon University and Director of the Institut Henri Poincaré, Fields Medal winner in 2010.

[1] Jean-François Thébault was a member until January 31, 2016 - resigned for personal reasons.

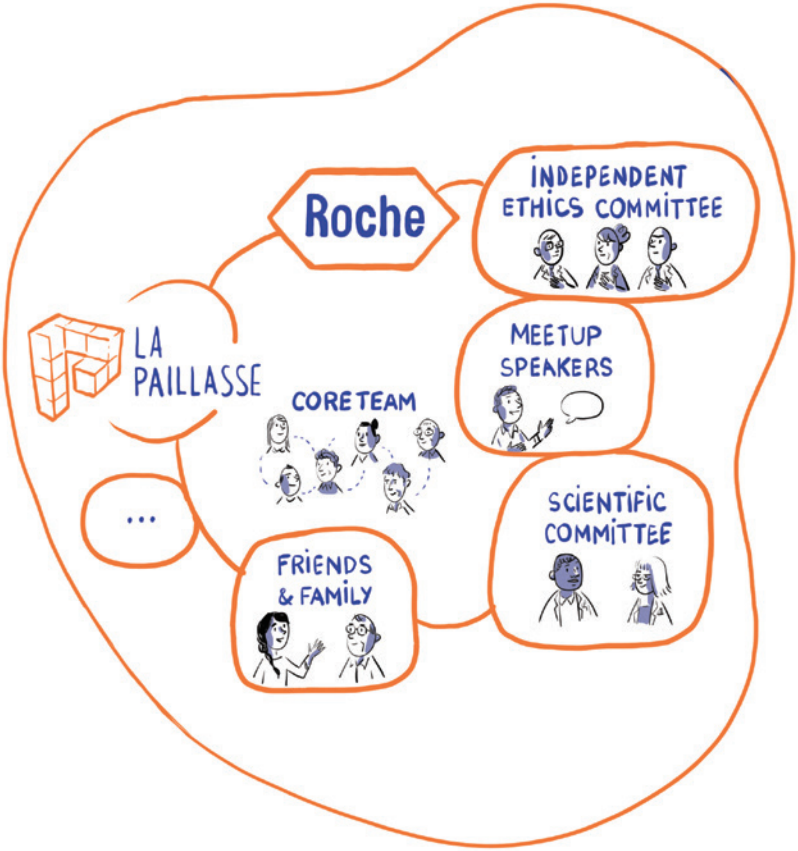
The Scientific Committee

Aurélien Alvarez	Teacher-researcher in Mathematics, Lecturer at the University of Orleans, mainly interested in Dynamic Systems.
Dr Jean-Pierre Armand	Medical Oncology Specialist, Senior Consultant at the Institut Gustave Roussy and at the Institut Curie.
Djalel Benbouzid	PhD in Machine Learning, Post-doc at the LIP6 Laboratory, Pierre et Marie Curie University.
Nicolas de Cordes	Anticipation Marketing VP at the Orange Group, has put in place and managed the project Data4Development in the Ivory Coast and in Senegal.
Dr Charles Ferté	Medical Oncologist, MD PhD at the Institut Gustave Roussy and Expert in Bio-computing.
Pr Thomas Gauthier	Strategy professor at the *Haute école de gestion* in Geneva. His research focuses on the practical application of data science and forecasting on decision-making.
Dr Leila Kockler	Roche Representative, Medical Project Director at the Medical Affairs Division of Roche France.
Thomas Landrain	President & Co-founder of La Paillasse.
Pr Philippe Ravaud	Epidemiology Professor at Paris Descartes University and at Columbia University, Research Director at INSERM, Research Director in Epidemiology and Statistics in Sorbonne Paris University, Director of the Clinical Epidemiology Center at Hôtel-Dieu hospital, Director of Cochrane France, Director of the EQUATOR France Center.

Epidemium's many shapes and sizes

// **Epidemium:**

A collaborative scientific research program open to all, initiated by the pharmaceutical company Roche, and the community-oriented laboratory La Paillasse. Epidemium's ambition is to explore big data's potential in cancer epidemiology thanks to its community.



// **Challenge4Cancer (C4C):**

A major call for projects, from November 5, 2015 to May 5, 2016, in the form of a coopetition, which was structured around four main categories, centered on cancer epidemiology and data science. Teams were free to constitute themselves and develop a project:

1. Understanding the distribution of cancer across time and space.
2. Risk and protective factors of cancer.
3. Meta-epidemiology: understanding cancer from the scientific and medical literature.
4. Environmental changes and cancer.

These topics were complex and the project's formats were not defined beforehand. The teams were encouraged to include a variety of disciplines by leveraging skills in data science, healthcare, social science, or even design.

// The resources:

In order to guarantee the success of the Challenge4Cancer and of its participants, Epidemium provided teams with a variety of resources. First, from a technical perspective, more than 21 000 open data sets were assembled and categorized into themes. They were also associated to a data analysis environment, provided to Epidemium's community thanks to a range of technical partners: HyperCube, Dataiku and Teralab.

Epidemium also developed the Call4Debate, whose aim was to use events to get the community involved, have participants meet, and promote tools developed in other areas that could be useful for Epidemium. Experts or external speakers were involved and supported the program: Institut Curie, Paris-Saclay Center for Data Science, SchooLab, Bress, Quinten, Hacking Health, Global Knowledge, Club Jade, CapDigital, Cancer Campus, Wikimedia, ...

Finally, in order to present and promote the work produced by the community, many digital tools were put in place and were based on the needs of participants: a website, a Wiki, a Q&A, a Facebook group and page, a Twitter account, a Meetup account, ...

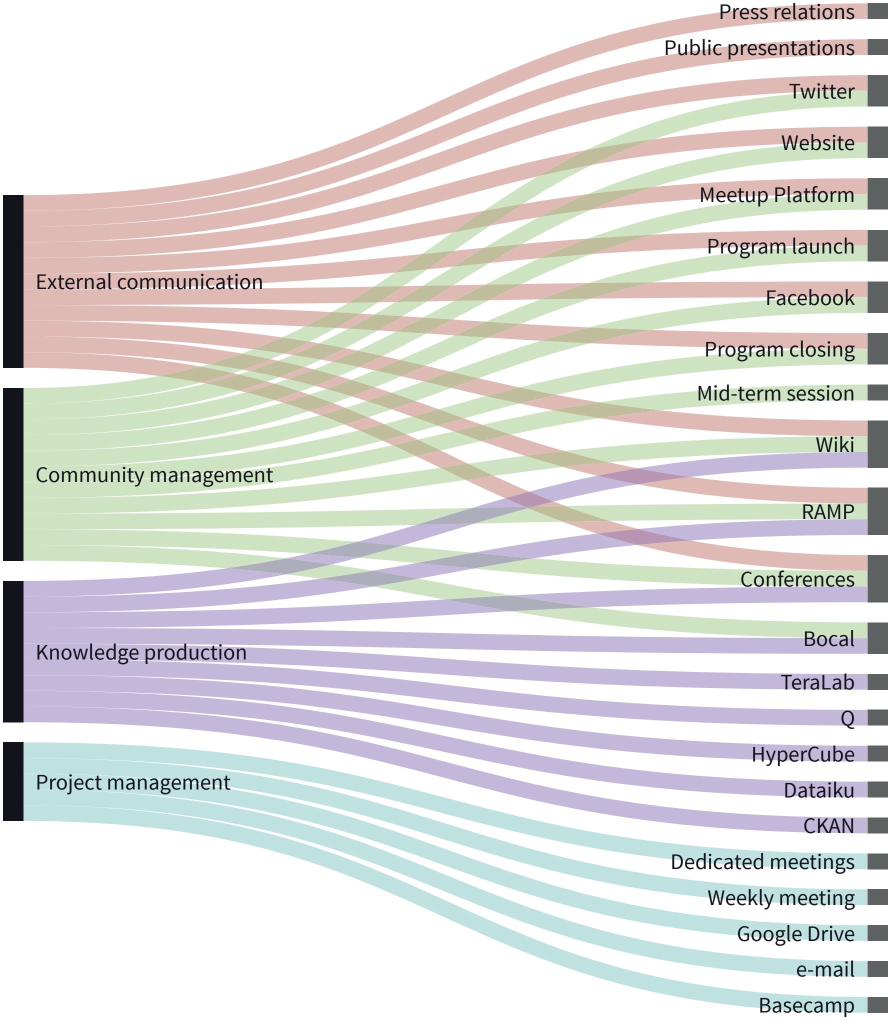
// The committees:

To structure and support the work of teams, two Scientific and Independent Ethics Committees were established before the start of the contest. Their goal was to think about the program's establishment and development. Six months after the program started, the committees formed a jury in order to evaluate the different projects submitted by the teams.

A summary of Challenge4Cancer's first edition in a few figures:

+ 20	678	15	8
events	community members	projects	finalists

Epidemium's toolbox



Description: *Sankey diagram representing tools put in place by Epidemium and their features*

Source and author: Epidemium's coordinating team.

Call4Debate 2015-2016

	Date	Event	Subject & Speaker
🟧	10.06.15	**Conference**	"Ethics and Health Data" **Jean-François Thébaut,** *Cardiologist, Member of the High Authority on Health College*
🟧	10.15.15	**Conference**	"Processing health data, issues and realities" **Alexandre Templier**, *Data science specialist*
⬛	11.05.16	**Challenge4Cancer launch**	Evening dedicated to the Epidemium community marked by the opening of Challenge4Cancer, and the presentation of the four main themes and participation rules
🟧	11.12.15	**Conference**	"Epidemics models on the basis of mobile data" **Nicolas de Cordes**, *VP anticipation marketing at the Orange Group &* **Stefania Rubrichi**, *Biomedical engineer & data scientist*
🟧	11.24.15	**Conference**	"Predicting lung cancer survival rates at an early stage" **Charles Ferté**, *Deputy Clinic Director in Oncology at IGR &* **Mathilde Bateson**, *Data Scientist at the Hypercube Institute*
🟦	12.10.15	**Bocal**	"Setting up teams & projects"
🟧	12.12.15	**Conference**	"Open data in cancerology: case study" **Akpéli Nordor**, *PhD student at the Curie Institute and at the Massachusetts General Hospital*
🟦	01.14.16	**Bocal**	"Developing, advancing, and documenting your projects"
🟧	01.21.16	**Conference**	"Open data in healthcare: issues and debate" **Geoffrey Delcroix**, *in charge of innovation and prospective studies at the CNIL &* **Jonathan Keller**, *Legal Expert at La Paillasse*
🟧	04.02.16	**Conference**	"What's behind the Open era (Big Data, Science, etc.)?" **Guillaume Dumas**, *Co-founder of HackYourPhD &* **Célya Gruson-Daniel**, *Co-founder of HackYourPhD*

▨	02.13.16	**RAMP**	Coopetitive working day on Epidemium data
🟧	02.18.16	**Conference**	"RAMP results analysis"
🟦	02.25.16	**Bocal**	"Oncology and epidemiology"
🟧	03.01.16	**Conference**	"A Google 3.0 for cancer: can it be possible?" **Alain Livartowski**, *Doctor at the Curie Institute*
⬛	03.12.16	**Mid-term session**	Day when participants were invited to present their projects to committee members and discuss their findings
🟧	03.17.16	**Conference**	"Discovering data science + case study in oncology" **Amel Sahli**, *PhD in Mathematics and Market Manager at Global Knowledge*
🟧	04.07.16	**Conference**	"First realizations made by the Epidemium community" **Epidemium DB Meta-project** & **Viz4Cancer Project**
🟧	04.14.16	**Conference**	"Collaborative 3.0 work methodologies: intermediaries to facilitate the conception of innovations" **Olga Kokshagina**, *Researcher at Mines ParisTech* & **Yohann Sitruk**, *Researcher at Mines ParisTech*
▨	04.30.16	**RAMP**	Coopetitive working day on Epidemium & EpidemiumDB meta-project data
🟧	05.19.16	**Conference**	"Meetup & Workshop in writing collaborative scientific articles - with Authorea" **Authorea**
⬛	28.05.16	**Award ceremony & closing of the Challenge4Cancer contest**	Day officially marking the end of the Challenge4Cancer contest and during which team finalists presented their projects to the committees: followed by the award ceremony and mentions

Type of events:

🟧	Conference	▨	RAMP
🟦	Bocal	⬛	Varied

Further details...

// Combining cancer and big data analysis using a flexible and agile methodology

- Olson G. (2000). “Distance Matters” in Human-Computer Interaction, Volume 15, pp. 139–178, available online <http://www.ics.uci.edu/~corps/phaseii/OlsonOlson- DistanceMatters-HCIJ.pdf>, last accessed on November 30, 2016.
- Lakhani K. R. (2016). “Managing Communities and Contests to Innovate with Crowds” in Revolutionizing Innovation, pp. 109-134, Cambridge, MIT Press.
- Houllier F. and Merilhou-Goudard JB. (2016). Collaborative Sciences in France. Report published in February 2016, available online <http://www.sciences-participatives.com/Rapport>, last accessed on November 30, 2016.

// Measuring Epidemium's program

- Barabási A.L. (2016). Network Science. Available online <http://barabasi.com/networksciencebook/>, last accessed on November 30, 2016.
- Nielsen M. (2012). Reinventing discovery: the new era of networked science. Princeton University Press.
- Wuchty S., Jones B.F. et Uzzi B. (2007). “The increasing dominance of teams in production of knowledge” dans Science, May 18; 316(5827):1036-9, available online <http://www. kellogg.northwestern.edu/faculty/jones-ben/htm/Teams.ScienceExpress.pdf>, last accessed on November 30, 2016.
- Klug M. and Bagrow, J.P. (2016). “Understanding the group dynamics and success of teams” in Royal Society Open Science, April 6, 2016, available online <http://rsos.royalsocietypublishing.org/content/3/4/160007>, last accessed on November 30, 2016.
- Börner K., Contractor N., Falk-Krzesinski H.J., Fiore S.M., Hall K.L., Keyton J., Spring B., Stokols D., Trochim W. et Uzzi B. (2010). “A multi-level systems perspective for the science of team science” in Sci Transl Med, September 15, 2, 49cm24, available online <https://www. ncbi.nlm.nih.gov/pmc/articles/PMC3527819/#>, last accessed on November 30, 2016.

#2

MEDICAL AND SCIENTIFIC INNOVATION

Medical progress and the improvement of healthcare for patients are two ideas that are at the heart of Epidemium's project. If we seek to better treat, cure and maybe one day defeat cancer, it will be by tapping into the potential that massive quantities of data represents, including epidemiological data. Public health has everything to gain from utilizing the new technologies and forms of knowledge offered by data science. What's at stake? Building tomorrow's leading-edge, innovative and precision medicine!

// AUTHORS

Dr Charles Ferté / Pr Bernard Nordlinger / Dr Mehdi Benchoufi / Dr Perrine Créquit
Pr Philippe Ravaud / Muriel Londres / Dr Cécile Monteil / Équipe Baseline

DATA
DATA
CANCER

How can we use data science and big data in healthcare?

PATIENT EMPOWERMENT GENOMICS QUANTIFIED SELF

MACHINE LEARNING PRECISION MEDICINE

Healthcare is a major issue for the coming years of our societies. In contrast to many other industries, this field has not yet adopted the digital transition, although the potential that such new technologies hold is huge, for instance when it comes to transforming patient care and the quality of treatment. The digitization of healthcare leads to the daily production of new data, which will eventually allow medicine to usher into a new era where healthcare is more personalized and fair.

// **AUTHORS**

Dr Charles Ferté / Pr Bernard Nordlinger

Soon enough, healthcare in our connected world will be nothing like what we have previously known. It is already a field where the most cutting-edge innovations are being experimented and which we hear about everyday. Connected objects, artificial intelligence, data sharing, big data, blockchain, etc. are terms that are making their way into the world of traditional healthcare and transforming its practice.

Moving toward greater awareness

Massive quantities of health data are generated every day pertaining to the frequency and the death rate of diseases, the efficiency of prescribed treatments, and the medical conditions and living habits of citizens. However, a transformation is currently underway, giving rise to different forms of interactions between the healthcare system and citizens, patients and non-patients alike, helping to address concrete needs. This ranges from the emergence of medical scheduling tools such as Doctolib, or the sharing of medical files, to tools analyzing the voice and fine movements in case of Parkinson disease (mPower project led by Sage Bionetworks [1]), to chatbots offering patients the possibility to ask questions to "conversational robots" and to receive personalized responses, to traceability and epidemics prevention APIs, or even predictive tools offering a form of personalized medicine, such as MammaPrint [2], which orients patients suffering from breast cancer and allows them to choose their treatment.

These new tools allow hospitals to start bridging the digital gap that disconnects them from patients the moment they are physically away. The discontinuity that is experienced when moving from the physical to the digital spaces of hospitals is recognized as a major constraint for the delivery of healthcare services and with regard to the efficiency of systems put in place to follow-up on patients. Hospitals must therefore become hybrid players, with a presence in both the digital and physical worlds, to the benefit of the citizens. As such, as soon as healthcare professionals (health centers, hospitals, healthcare workers) start offering citizens the possibility of staying connected, we are likely to see the implementation of solutions designed to create a sense of continuous interaction for patients. This would help healthcare providers address a vast array of issues that they are currently facing, such as the reporting of side-effects, education and prevention, information on diseases, etc.

Beyond the immediate benefits of experiencing improved interactions between patients and doctors in terms of quality of care and services provided, these tools also enable the medical community to collect important epidemiological and environmental data. These are the fertile ground on which the traditional health system can build upon and evolve, which we will present. Moving closer to the ideal of personalized medicine where medical services are adapted to each patient, and each disease is approached as a unique case, also creates many challenges.

The revolution lies not only in the diagnosis, but in the treatment, the follow-up, the prognosis, and prevention

The first major evolution in the realm of health data is the exponential growth in the number of data available. Genomics represents a vast source of information for doctors and researchers, which they have been seizing for several years now.

Who has not heard about the *Human Genome Project*, one of the, if not the single most important scientific event of our generation? In cancerology especially, The Cancer Genome Atlas (TCGA) [3] and the International Cancer Genome Consortium (ICGC) [4] are two major sequencing programs that have created a considerable amount of publicly accessible data thanks to cohorts of hundreds of patients. TCGA has, for instance, created a multi-dimensional map of genetic mutations for thirty three types of tumors with the sequencing of DNA, RNA, RPPA, etc. These international programs aim at better understanding cancer thanks to the genome sequencing of dozens of different tumors. The scientific community can therefore make use of the immense database at its disposal to accelerate research against cancer.

Genomics is seen as the future in the short and mid term but, at the same time, other types of data are experiencing an increasingly significant rate. Free text for example, which represents all of the written records produced by healthcare professionals, is a new important source of data, with the potential of multiplying the amount of information available on diseases and their treatment. This is the case for the more and more medical imaging data we get as well, thanks to the improvement of medical imaging technologies among other things. In 2015, 80 million CT-scans have been done in the United States. This number was four times less, twenty years earlier.

We are also observing the considerable increase of quantified-self data, which are data collected and provided by citizens themselves, and sometimes without their full consent and knowledge. The data is produced through wearables, also known as 'portable connected objects', and their corresponding applications.

However, when such a vast amount of data is available, the first issue before starting to analyze them, is to make them accessible to everyone, by opening the datasets collected by healthcare and research organizations, public institutions and private companies who own the connected devices, mobile applications and other wearables. Why has the sharing of data become so crucial? Because it greatly increases their impact by way of allowing everyone to use, preprocess, analyze, and interpret them. And it is this analysis that is made possible today by the creation of new and more efficient, more precise and more democratic machine learning algorithms. After having proven their ability in other fields (chess and the game of Go, for example), major projects such as IBM's Watson[5] and Google's Deepmind[6] are now targeting healthcare in order to make use of the immense quantities of circulating data. It is almost as if the fields of applications for these forms of

> *Some of the applications will come to fruition and some will not, either for technical reasons or because progress is more uncertain and difficult when it comes to people than commerce, or personal transportation"*
>
> ***Pr Bernard Nordlinger***

super-intelligence served as a training ground that allowed them to increase their processing power before they started tackling more serious issues such as public health. Cancer is therefore the new big enemy for artificial intelligence and large corporations are allocating vast amounts of resources to better address the needs of citizens. In August 2016 for example, Watson has successfully diagnosed a case of leukemia that had not been detected by human intelligence[7], proving that using artificial intelligence in the field of healthcare will result in major improvements for patients. The University of Tokyo has indicated in September 2016 that Watson had helped diagnose and treat cancer patients in 80% of the cases that it was given for analysis[8]. Overall, this is set to lead to major gains in public health.

At the same time, access to the cloud and the development of new tools offer the possibility for everyone to do what was until now only possible by the companies that owned such technologies. The decrease in prices, and the fact that technologies are becoming more efficient make their use more accessible and widespread. Many hosting services are emerging on the cloud, and are complementing services already offered by large companies such as Amazon, Azure and Google. They are also offering hybrid cloud services that can be personalized to a greater extent and that everyone can use. Nevertheless, we must remain vigilant with regards to data security as a lot of data can be used to identify individuals. This means that by cross-referencing the different data sets, it is possible to reveal the identity of patients. This is why today the hosting of data is only limited to organizations that are ready to comply with article L.1111-8 of the French Public Health Code.

As a consequence, we are witnessing many collaborative initiatives seeking to leverage open data. Very often, they take the form of data challenges and bring together a community organized both online and offline, sharing a common goal and framework of action. Synapse[9], Kaggle[10] or Epidemium are important cases in point, based on the idea of sharing data, using online tools, co-creating algorithms, and most important of all pooling know-how and skills together. By

Article L.1111-8 of the public health code stipulates the conditions in which data can be provided to a hosting company.

- *The person concerned by the health data must have given their full and explicit consent to the hosting of their data.*
- *The hosting company should be authorized to exercise their activity.*
- *The hosting company is subject to rules of confidentiality laid out in article L.1110-4 of the public health code and to principles of interoperability and safety.*
- *When health professionals or institutions host their own health data, they are not subject to authorization and are not obliged to issue request to the person concerned for the preservation of their data.*

developing a collaborative framework structured around cancer epidemiology, Epidemium's program and the Challenge4Cancer contest have demonstrated the extent to which scientific research and patients can benefit from the idea of interdisciplinarity and openness. These types of game-changing collaborative programs can have a genuine impact on the field of traditional research insofar as it provides it exactly with what it is currently missing: a diversity of expertise, given that a graphic designer could play a role that would be as important as a data scientist, the pooling of know-how all the while removing the unnecessary politics usually involved in such endeavors, and finally the unrestricted sharing of results with the rest of the community.

Moving closer to personalized medicine, a journey full of challenges

Nevertheless, no tool and no platform has changed the face of healthcare yet. How can we therefore pave the way for a true personalized medicine? Is it utopian to imagine a world of medicine where each and everyone will be able to access the tools that will allow them to choose their treatment based on the best predictions offered by big data analysis, taking into account a number of different criteria ranging from genetics to

> *I believe that the greatest challenge will be to cross-reference clinical and epidemiological information, which refers to what is represented, that is the phenotype that includes genetic data. We are not there yet. We have more and more biological markers (on a limited number of genes) that allow us to adapt anticancer treatment for those who are lucky enough to benefit from it. Sequencing the human genome has become a common practice, but the molecular diagnosis only has a few applications for the treatment of cancer."*
>
> ***Pr Bernard Nordlinger***

eating habits? Or even to perform early detection of a disease by having access to someone's online searches? These two specific cases exist already but remain isolated. In order to increase access we still have to tackle many challenges. Although personalized treatment programs have shown hopeful results, they remain confidential and the few successes we have known can hardly be generalized for all patients.

The challenges that have to be faced to obtain positive results in personalized medicine are firstly technical: how can we draw useful lessons from data scattered across a variety of sources? The data collected are very heterogeneous by nature (genomics, physiological, biological, social, environmental...), by format (text, numerical value, signals, 2D and 3D images, genome sequences...), and by the fact that they are scattered across different information systems (health institutions, research laboratories, public databases, private corporations...). However, when it comes to big data analysis, there can be no intelligence without learning. The fragmentation of data requires us to invent complex systems in order to tackle the challenge of integrating data of different nature and from a variety of sources. Following the same logic, we are also witnessing an increase in terms of demand for algorithms, storage capacity, and capacity of treatment of databases.

From a technical point of view, another problem comes from the fact that health institutions, trying to come to grips with new innovations, have all developed their own health system independently of other entities. This means that health institutions are not interoperable, which considerably hinders the sharing of data. How to make sure that when a patient has appointments in different health institutions, that their data is not lost or does not have a format problem, and that such data can be used by health professionals who do not necessarily have the time to communicate between themselves?

Giving citizens control over their data and making them true players in the use of their own data is also a major challenge. This is what we call patient empowerment. From a public health perspective, each and everyone should be aware of the wealth of data they generate on a daily basis, and of the potential use that can be made of them in the medical world, for example advancing the understanding and treatment of complex diseases like cancer. In an ideal future, research programs will not need to build armies of volunteers anymore to handle their data, but instead they will be able to tap into the wealth of existing data made possible as a result of the increased awareness of citizens.

In order to have enough data in terms of amount and diversity, the institutions involved in this matter have to come to grips with the topic of ethics. There should be an obligation of information to patients when using data coming from clinical trials or care pathways: it is necessary to explain which data are collected, how they will be anonymized, how they will be stored, and for what purpose they will be utilized. The ability to clarify and educate on the purpose of using data is crucial in

How can we use data science and big data in healthcare?

order to convince the general public of the interest that using such data represents in terms of research and public health. Respecting the ethical principle of transparency in the way data is collected, analyzed and treated can also be difficult to handle due to the extra workload and complex processes that this represents for health institutions.

Finally, collecting large amounts of data while giving confidence citizens to share them represents a security challenge. Not necessarily about data piracy, as we usually come to think of it, but instead about the non-corruption of data. In 1996, the United States issued the Health Insurance Portability & Accountability Act (HIPAA)[11], which requires the implementation of measures designed to secure and respect privacy for the creation, conservation, and sharing of personal data. Likewise, the European Union adopted in 1995 a Directive on the protection of data, putting the entire Union on the same level playing field. However, although the collection and processing of personal data have to comply with a set of rules, the law still cannot prevent technological failures or harmful acts. Blockchain technology represents an interesting innovation in this matter and could technically be used in the healthcare sector[12]. Nevertheless, is this technology mature? And most importantly, are healthcare institutions ready to integrate such innovations into their operations? We would probably need to wait a couple years before the use of blockchain technology becomes widespread in the healthcare sector, but this is still a nice goal to attain over the long run.

Whereas many sectors of the economy are being uberized, we are still waiting to see the emergence of innovative, high-tech, and fast-growing startups worth billions in value in the healthcare sector. It is true that the medical community needs to find the right balance between innovations and their applications, which works to limit the uncertain implementation of new technologies. Of course, big data is increasing the efficiency of the field as a whole, as well as processing and prevention, whether it be for the improvement of prediction, the reduction of costs by limiting unnecessary acts and exams. or through a better understanding of diseases.

However, the use of big data technologies cannot be done at the expense of a lower level of proof, which would mean a lower level in terms of the validation of scientific results, or a lower level of data security. This is the exact reason why the implementation of big data in the healthcare sector takes time. Today, many are getting worried though, because the concrete promises offered by precision medicine have yet to come. Nevertheless, this is the direction in which things are going and there is no going back. We are witnessing a structural trend, of which patients empowerment is a strong component, pushing the medical system towards improved performances, with a greater level of opennes and rationality. We are only seeing the beginning of what is to come. ■

> “The problem relates to the collection of personal data, whether done with or without the consent of individuals, when we know how difficult it is to protect anonymity, but also with regard to the use of data, and not only for insurance companies, to take the most widely used example."
>
> **Pr Bernard Nordlinger**

[1] mPower mobile Parkinson Disease Study website <http://parkinsonmpower.org/>, last accessed on November 30, 2016.

[2] Agendia Website, MAMMAPRINT® 70-GENE BREAST CANCER RECURRENCE ASSAY <www.agendia.com>, last accessed on November 30, 2016.

[3] The Cancer Genome Atlas website <http://cancergenome.nih.gov>, last accessed on November 30, 2016.

[4] International Cancer Genome Consortium website <http://icgc.org>, last accessed on November 30, 2016.

[5] Article “A large range of applications in everyday life" on IBM's website <http:// www-05.ibm.com/fr/watson/>, last accessed on November 30, 2016.

[6] DeepMind is British company specialized in artificial intelligence and bought by Google in 2014 <https://fr.wikipedia.org/wiki/Google_DeepMind>, last accessed on November 30, 2016.

[7] Humanoids, IBM Watson made a successful diagnosis of a case of leukemia in Japan, published online on 10/08/2016 <https://humanoides.fr/ibm-watson-japon-leucemie/>, last accessed on November 30, 2016.

[8] T. KAWAMURA, “Big data system shows promise in helping cancer patients at Todai”, The Asahi Shimbun, September 19, 2016, available online <http://www.asahi.com/ajw/articles/AJ201609190064.html>, last accessed on November 30, 2016.

[9] SAGE SYNAPSE, <https://www.synapse.org>, last accessed on November 30, 2016.

[10] KAGGLE, Your home for data science <https://www.kaggle.com>, last accessed on November 30, 2016.

[11] European Union Agency for Network and Information Security (ENISA), Health Insurance Portability and Accountability Act, available on <www.enisa.europa.eu>, last accessed on November 30, 2016.

[12] Also read D. SCHUYLER, Is the Blockchain a Potential Cure for Securing Healthcare data? <http://leavittpartners.com/2016/09/is-the-blockchain-a-potential-cure-for-securing-health-care-data/>, last accessed on November 30, 2016.

Crowdsourcing cancer epidemiology

EPIDEMIOLOGY CROWDSOURCING CROWDACTING

MICRO-TASKING MEGA-TASKING

Healthcare, and more specifically epidemiology, are profoundly impacted by the technological changes currently underway. Two facts are at play, the unprecedented production of data and the increasing involvement of various communities. Ways and methods of fully expressing their potential are yet to be found; Today, a variety of means are used to engage communities (micro-tasking, mega-tasking) and are promising epidemiological approches that will be original, large, distributed, marked by social and reticular approaches, augmented in real-time and strongly inclusive. Epidemium has started clearing the way to these promises.

// AUTHORS

Dr Mehdi Benchoufi / Dr Perrine Créquit / Pr Philippe Ravaud

The healthcare sector is being disrupted and this is leading to new practices. Two trends are profoundly affecting it. First is the sheer amount of data that is now available. And second is the number of individuals willing to play a role in biomedical research. The emergence of big data analysis coupled with the increasing involvement of communities coming from civil society are making it possible for new players to tackle the many challenges of contemporary medicine. Epidemiology, a field that by its very nature relies on the use of data and external initiatives outside of its traditional body of practitioners, is also perceived as an observation outpost as well as a field of experimentation in this matter. Epidemiology is experiencing shifts that we can qualify (although we still have to wait for this to be confirmed) as paradigmatic[1].

New applications in epidemiology are legion. They are often the results of new and unique approaches initiated players stranger to the healthcare sector. Moreover, they are often presented as significant examples demonstrating the potential of cross-fertilizing gigantic amounts of data, of the processing capacity required for their treatment, and of the number of internet users capable of getting involved in data analysis.

When the potential of big data is catalyzed by a community capable of processing them, crowdsourcing[2] can provide epidemiology with new and interesting means. Crowdsourcing is efficient in that mobilizes and pools together a large and distributed workforce. Its minimal and most widespread form, called micro-tasking, consists in breaking down a complex task into a multitude of more basic tasks. It is also the most frequently used form of crowdsourcing in biomedical research. It is based on rewards, most often of financial nature.

Another form of crowdsourcing called mega-tasking also exists. Mega-tasking is characterized by the willingness to get involved, to make oneself useful by making a contribution, or by offering non-medical skills for the purpose of advancing a major social issue. It is a hallmark of our era. Epidemium is a conclusive example. The desire to get involved, the idea of assertively tackling ambitious challenges, which is something that we were able to witness during the Challenge4Cancer

> *Big data analysis will allow us to identify the factors that are responsible for the emergence of cancer among patients, and to alter our approach in the area of public health in France."*
>
> ***Muriel Londres***
> *Member of the Independent Ethics Committee*

contest, is the expression of a general trend, often called Do It Yourself (DIY), with at its heart the idea that these problems concern us all, and that therefore the solutions should, in turn, also be everyone's business.

From crowdsourcing to crowd-acting

Internet is crowdsourcing. Through the sum of our interactions, we are generating massive amounts of data on a daily basis, that are taken by major platforms to be compressed, refined, and transformed in order to extract its value, which for now is more about its capitalistic value, and not yet its scientific value. The abundance of petabytes of data can be perceived as a natural and passive form of crowdsourcing. In the era of the internet, everything comes in the form of 'data' provided by users. We can therefore distinguish a form of crowdsourcing that is the active substance of the internet, from a form of crowdsourcing that is more voluntary and which we will qualify as crowd-acting.

Epidemiology is, among other things, the study of disease factors. Epidemiology cannot remain unchanged by the evolutions that we are referring to. The means through which it gains knowledge and its applications are today amplified throughout time and space. On the one hand, data widens and deepens our understanding of the genesis of diseases. On the other hand, the dematerialization of platforms arising from digital transition the reformulation of space, which today has no territory, and a time of perpetual present, provide our discipline with a capacity to act in real-time. The moment we imagine being able to control the dissemination of diseases, the potential of this gets all its value.

Given these points, we can understand that the diversity and amount of data being produced by our systems are, on the one hand, allowing us to considerably increase our knowledge of our environment and of our behavior, and on the other hand, allowing us to pool our research efforts together, involving people coming from academic circles as much as expert citizens, which represents a decisive opportunity. These citizens who are eager to play a role in a more open science are both providing support to the project and augmenting its potential by way of delegating micro-tasks. Depending on the type of challenge, of which Epidemium is an avatar, they are also able to offer diverse forms of knowledge.

Crowdsourcing examples

Passive crowdsourcing

// A few examples

The HealthMap[3] initiative used crowdsourcing methods to detect a suspicious form of fever in Africa before public health authorities were alerted by what was later known to be cases of Ebola fever. The method was based on the continuous analysis of a diverse set of collected data from various sources: expert websites, blogs, social networks, health forums. These sources were obtained as a result of a passive crowdsourcing or rather of crowdsourcing in the literal sense: collecting data from the crowd and from the sources from which they were produced.

Let us note that the most inspiring examples are often not the most conclusive successes, but rather they shed light on promising and experimental approaches, and as such, pave the way for new paths and can be the failures leading to victories. This phenomenon is exemplified by the widely commented case of Google Flu[4]. Following this experience, Google Flu did not succeed, as it was intended to, to anticipate or predict the spread of the flu. However, what remains interesting in this case was the idea of forming an intuition about the phenomenon without having to resort to 'manual labor' and the aggregation of data, the meticulous consolidation of information, and the analysis by a group of experts. The ability to complete the work of people using an automated approach based on algorithms leads to significant cost- and effort-savings. It should be mentioned that for two consecutive years Google Flu retrospectively but successfully predicted the evolution of the flu faster than sanitary health monitoring systems in place in the region. This means that the algorithm missed the target, but only slightly.

> “*In fact, as research in epidemiology continues to transform itself as a collective sport and decentralized effort, the epidemiologist of the future will most likely become a community manager."*

// Augmented crowdsourcing: the potential of machine learning

At this point, we should draw a portrait of the essential technological approach that machine learning represents. Machine learning consists in educating computers by feeding them data, which in turn enables them to gain experience and refine their preventive, predictive, and in some cases cognitive analysis capacities. Such techniques require and depend upon a vast amount of data. The massive health data that healthcare systems have at their disposal can be perceived as the result of passive crowdsourcing, thus becoming a valuable component offering us hopes of therapeutic innovation with steady pharmacopoeia.

These techniques have proven to yield promising results: Google's victory in artificial intelligence in the game of Go, the automatic detection of tumors from scanner images, etc.

Admittedly, the volume of data is not in itself an advantage in terms of statistical analysis. However, there are algorithms that become more powerful when fed with more data. This is the case for neuronal networks, for which the results are absolutely fascinating. This somewhat biomimetic approach gathers neurons as calculation units whose rules are precise and whose parameters change as and when data is processed.

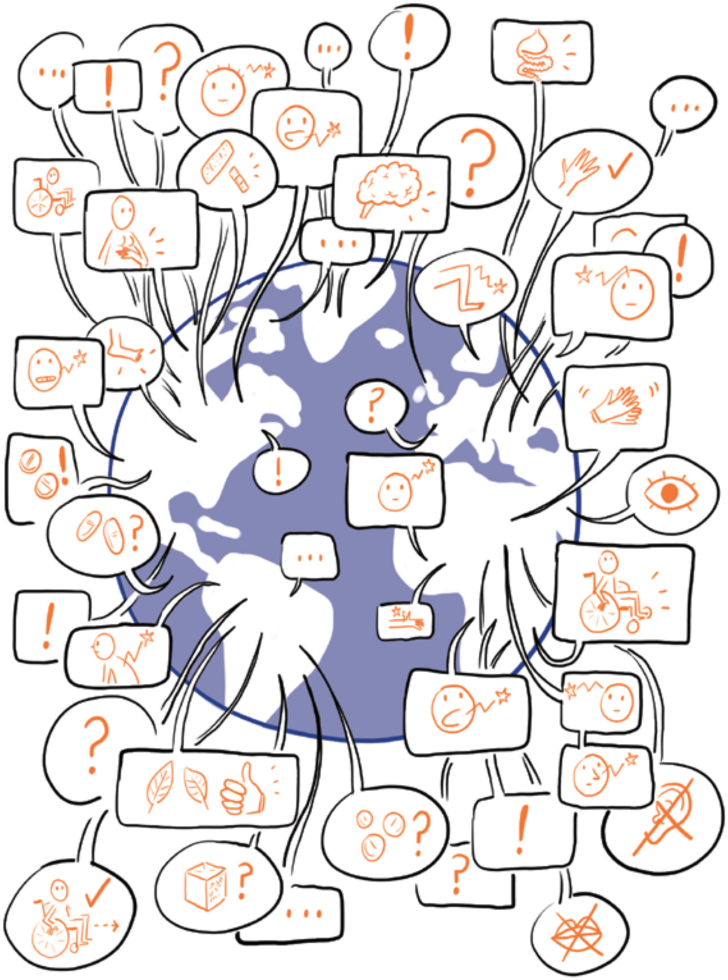
Active crowdsourcing

// Micro-tasking

Micro-tasking is a prime example of communities getting involved in the co-construction of healthcare. It represents an important tool for researchers in the way it can mobilize large crowds to get a vast amount of tasks that were up until now considered to be very difficult to tackle. These tasks are not necessarily complex but usually require a considerable amount of human labor and time. The Embase project conceived by the Cochrane Collaboration, is an excellent example of the division and parallelization of tasks with the aim of identifying clinical trials coming from Embase, which is a vast bibliographical database, in order to publish them in Cochrane's central registry.

It should be noted that no expertise is required of participants, and that they may not have any medical knowledge.

With the aim of creating a more inclusive medical research in mind, the Compare or Nutrinet projects enable patients to take part in online clinical studies and to promote a more open form of medical research, outlining the promise of a more participative and collaborative form of medicine. As a matter of fact, for Compare, patients with chronic diseases share their data, fill out questionnaires, choose research topics, with an objective of mobilizing up to 200 000 patients.



Using an original approach, Epidemium has shown that with a challenge, communities can come up with organization systems that are, if not spontaneous, at least similar to those they should adopt given the skills they gather. Thus, some teams with no medical expertise or computer science skills, undertook the remarkable task of collecting and cleaning up data. For instance, the Baseline project was able to build a rich database among 98 countries, including 107 different risk factors, that can now be utilized by research teams.

Other teams, more experienced in data science, were able to develop algorithms. We observed that many participants were eager to explore topics traditionally requiring a high level of expertise as they were unhindered by the weight of authority.

We can also call to attention how the Challenge encouraged inclusiveness through micro-tasking. Although many participants lacked the skills required, they nevertheless sought to make a contribution to the project. It is in that respect that a great deal of ideas came up to offer to the community more basic tasks, within the reach of anyone: searching for open data sets, putting in place methodological tools, managing communities and recruiting, documenting the Wiki, etc.

// Mega-tasking

The possibilities offered by crowdsourcing are vast, ranging from micro- tasking to complex problems that individuals are capable of resolving without necessarily having the knowledge or background required. They come to form what Jimmy Wales has called "*experts of their own experience*"[5]. We can take an example from a field that is far removed from epidemiology, but that demonstrates the incredible power of crowdsourcing: FoldIt. FoldIt was initiated by the University of Washington in Seattle. The idea was to study the dynamics linking the structure of proteins in space and their functional characteristics, which is something that is not fully understood until this day. The question was highly sensitive and researchers observed that the regular manipulation of these proteins provides practitioners with an intuitive and empirical science on how they fold on themselves and how the folding is actually what gives them their functional property. Given all this, researchers had the idea of launching a contest open to everyone and accessible online via a platform. During the contest participants were asked to resolve a problem linked to the folding shape of proteins which could not be processed by a machine. This way, participants proved to be able to successfully guess the tri-dimensional physical structure of proteins although they hardly had any notion in molecular biology. Understanding the tri-dimensional structure of the retroviral protease enzyme of the M-PMV virus, which provides us with a model that is similar to HIV to test potential inhibiting molecules, was inaccessible to researchers for about 10 years. However, it only took participants three weeks to tackle this issue during the challenge. As a result, a report was published in *Nature Structural & Molecular Biology*[6].

The rise of communities is one of the most significant characteristics marking the history of the internet. In a way, it might even be the force that is driving it forward. In the field of healthcare, a great deal of forums exist and health communities are taking strong initiatives: they report on their diseases and the side-effects of treatments on websites such as CureTogether, they pool together data on pollution from connected objects, they monitor and share physiological features in so-called quantified self communities, they map information with regard to first aid such as the location of automatic defibrillators or access points for the physically challenged on websites like OpenStreetMap. Together, these new forces have come to shape a new form of collective intelligence, based on communities deliberately reporting on the health issues that concern them. These massive forms of crowdsourcing turn them into full-fledged players in the target to build medical knowledge.

All of this illustrates the power of opening up science to dynamic communities. Epidemiologists bear the responsibility of measuring the opportunities arising from potential discoveries and to reveal the full potential that can exist as a result of intertwining the works of experts and laymen.

> *The abundance of data will lead to new epidemiological studies defining new norms [new symptoms that improve the diagnosis]".*
>
> ***Dr Jean-François Thébaut***
> *(Usine Digitale, 04/28/16)*

Comprehensive epidemiology

// Methods

Biomedical research is currently being co-developed. It is a field that must include the power of distributed logics, and change its methodology in order to open up to those who are willing to make a contribution. There is a lot to learn from, re-use and include in terms of suggestions of others. The overarching idea is to stop perceiving biomedical research as an exclusive domain.

Consequently, epidemiology, augmented by the possibilities offered by crowdsourcing, will be capable of building interactions with new participants in what has historically been its reserved domain. It must be capable of creating new interfaces that will allow it to include new contributions, which have up until today, only remained outside of its field of action.

// Aspects

The new epidemiologist must take into account these new aspects: the management of communities, and the sharing and co-development of tools with the help of expert-citizens, in a sustained relationship of built-up reciprocity. It must measure the importance of the diversity of issues at hand and build a comprehensive approach that includes medical aspects as a priority, but that also harnesses the power of communities, all the while taking into account legal and ethical aspects.

// Co-construction

As research in epidemiology continues to transform itself as a collective sport and decentralized effort, the future epidemiologists will most likely become community managers[7]. Future epidemiologists will have to know how to build mutually beneficial ties with individuals whom they know to be worried about the protection of their personal data, and become experts in intellectual property. They will have to play a pivotal role as community managers working to integrate many and diverse contributions.

We will note that the epidemiologist will not need to look very far into the world of the web to build communities. They can start and take into account the necessity of new methods such as opening up their approach by sharing best practices, documentation, engaging in collaborative monitoring and problem co-solving. Crowdsourcing platforms such as a Meta Stack Exchange[8] enable us to explore ways to co-construct knowledge in many fields: computer science (Stack Overflow[9]), mathematics (Mathematics Stack Exchange[10]) and statistics.

Meta-epidemiology

The emerging field of research about research, which is probably one of the most determining fields in medical science is focused on improving the replicability of research. The efforts of numerous communities to bring to the attention of the general public the errors, conflicts of interest and frauds in biomedical research, demonstrates civil society's ability to share and crowdsource transparent results coming from the scientific literature. By doing so, civil society present themselves as the natural, useful and spontaneous allies of clinical research.

Conclusion

"Nothing is lost, everything is transformed"

Far from the pessimistic outlook that citizens would replace experts in their role, and that machines would rise at the expense of everyone else, actually, none of these players involved seem to disappear: to the contrary they are all emerging. Nevertheless, their respective roles are set to evolve, skills are being networked, ideas are being disseminated beyond the frontiers of their own disciplines. Experts have a role to play in the production of knowledge, but also in its transmission and interoperability in a language and form that is within everyone's reach. Therefore they should embrace their new responsibility to make the knowledge they produce as accessible and connected as possible which would, in turn, reduce the barriers separating their discipline from the world.

> *To facilitate this type of research, we need more data to be open and accessible while respecting security and ethical considerations."*
>
> ***Olivier de Fresnoye***
> *(Up Le Mag, 11/09/16)*

It is about finding the right tools and methods to ensure knowledge transmission is done in an accessible way, be it as educational content, or community management. These platforms are essential in enabling the online gathering of somewhat informal communities of challengers.

Thus, epidemiology, just like any other field of knowledge that is not exclusively owned by anyone, must remain open to individuals who are not necessarily certified practitioners to make their contribution and shape the field as a whole.

To summarize, the word that characterize the essence of what epidemiology is to become in the future is openness. Openness is a state of mind, it is the idea that science is by essence open to the thoughts of everyone and that it is not a registered trademark. Science is a practice, it is the way by which we can open up to a much greater number of contributions and efforts. ■

[1] Wiktionary definition <http://fr.wiktionary.org>: "in linguistics, the paradigmatic axis refers to the choice of words, whereas the syntagmatic axis refers to their position in the statement that is being made," last accessed on December 7, 2016.

[2] Harnessing the power of collective intelligence to deal with a topic or an issue.

[3] HealthMap <http://www.healthmap.org/fr/>, built by researchers, epidemiologists and developers at Children's Hospital in Boston, and allows anyone to follow-up in real-time the spread of a disease by collecting data across the web.

[4] Google Flu was an initiative launched in 2008 by Google whose aim was to predict the outbreak of epidemics using searches made through its search engine.

[5] Jimmy Wales, "The wisdom of crowds", The Observer, London, June 22, 2008. The idea that we are all experts of our own experience.

[6] Khatib, F., Di Maio, F., Cooper S., Kazmierczyk M., Gilski M., Krzywda S., ... & Jaskolski M. (2011). "Crystal structure of a monomeric retroviral protease solved by protein folding game players" in *Nature structural & molecular biology*, 18(10), 1175-1177.

[7] Profile whose goal is to manage and federate a community.

[8] Meta Stack Exchange <http://meta.stackexchange.com/>.

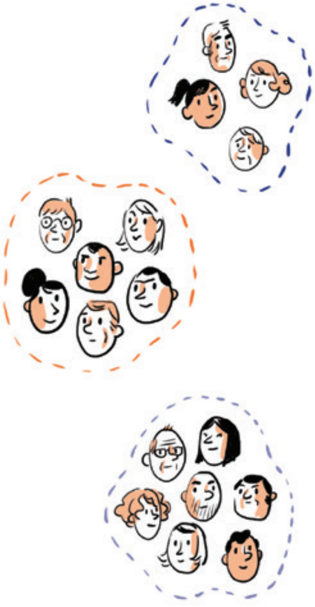
[9] Stack Overflow <http://stackoverflow.com/>.

[10] Mathematics Stack Exchange <http://math.stackexchange.com/>.

Benefits for the patients and the medical community

// AUTHOR

Muriel Londres



Technological disruptions can potentially impact all aspects related to the experience of patients in the healthcare system, and even their entire life experience of people with a chronic disease. It is important that patients and associations representing them become involved in projects and research that have been made possible by the mutation in healthcare linked to digital transformation.

At an individual level, having electronic medical records of patients that are compatible with all of the different care services involved in the medical care will be a big improvement: it will be more secure, more accessible, and easier to share information between healthcare professionals and patients, and healthcare professionals between each other. For people living with a chronic disease who have to relentlessly struggle

Benefits for the patients and the medical community

with the healthcare system and have no choice but to do with it, the prospect of having electronic medical records will enable them to better understand their pathology, and to make more empowered and enlightened choices with their medical team. At a time where it is still difficult for the patient to understand its own medical records, and when a fourth of treatments are not appropriate, we can expect that co-constructed innovations and tools will facilitate the entire patient's journey.

More generally, these datasets organized in the form of databases, becoming increasingly large, will allow us to deepen our understanding of diseases, to better predict them, but it will also help us choose what are the best options in terms of prevention and treatment.

With large datasets becoming more accessible, patients are effectively turned into full-fledged players. The conduct of research in this field cannot be done without their participation. As they closely observe the daily difficulties that they face, patients associations advocate in favor of people living with a chronic disease and put forward proposals to improve the healthcare system. The data collected must be accessible to them, as they are protecting their ethical use, and research should become more collaborative, taking into account the issues put forward by these patient associations.

Epidemium and the Challenge4Cancer contest included the work and voice of patient associations: we took part in the Independent Ethics Committee and the jury. Whereas the presence of patients and their representatives can sometimes be a bit forced in projects like Epidemium - when they actually do play a role - Epidemium asked us to participate and carefully took into account the many questions that arose during the Challenge, especially with regard to the final use of the research material produced by big data analysis, and our concern that scientific research had to be made accessible to the general public. The community aspect of the project should also be noted. Our implication in the Challenge4Cancer lasted more than 6 months. While participants were at work, a number of useful conferences and talks were organized. This was used as an opportunity by teams and representatives of

the Scientific and Independent Ethics Committees to meet with the participants and raise everyone's ideas and concerns. Beyond the results, which are more promising in terms of the work methods that were used than truly revolutionary in terms of scientific findings, Epidemium successfully broke down the barriers separating various big data players, in turn paving the way for more collaborative and open forms of research. ■

// AUTHOR

Dr Cécile Monteil

As a representative of the Eppocrate community, whose objective is to raise awareness among the medical community with regard to new technologies and digital tools, I immediately accepted Epidemium's invitation to become a member of the Independent Ethics Committee. Big data, which illustrates the capacity to collect and analyze massive scales of data, has been on everyone's lips for some time now, but in France, we haven't seen any concrete application in the field of medicine yet! In this context, it was particularly interesting to be able to finally participate in a program designed to experiment with such new techniques.

From the very beginning, I found that the program's focus on collaboration and its defense of open science to be one of the major strength of the project. Epidemium was able to tackle the challenge of gathering multi-disciplinary teams who accepted to publicly share their findings with each other and make them freely accessible for everyone. I believe that systems where disciplines work disconnected from others involved in

Benefits for the patients and the medical community

the research process is one of the biggest obstacles in medical research. A lack of communication and interdisciplinary work have crucial negative impacts on projects. That along with a lack of pooling means and resources due to competition and an obsession with obtaining patents (in China for example they have dropped the patent war for quite a while now) interfere with the efficiency of our research processes.

On top of my involvement with Eppocrate and my position at iLumens, a health simulation department at Sorbonne Paris University, I work part-time at the pediatric emergency center at Robert Debré hospital. Doctors in general rarely meet engineers, developers, or designers during the course of their studies, nor during their professional careers. However, we often find doctors who have come up with great ideas related to the creation or improvement of existing technologies, but because they don't know where to start, it sadly never ends up in anything concrete. How many interesting ideas have never seen the day! In contrast, how many unadapted health gadgets have been put on the market… because they were built by companies or entrepreneurs with no consultation of healthcare professionals nor patients along the process. The original ideas are often interesting, but without answering actual healthcare needs and being perfectly suited for their end users (healthcare professionals and patients) the final product is bound to fail on the long run.

For Epidemium, the role of the Independent Ethics Committee was crucial. The use of big data generated by individuals both healthy and ill, was necessary in order to gain knowledge. However, this cannot be done at the expense of strict ethical, confidentiality and data privacy rules safeguarding the rights of individuals. Our goal was to define an ethical charter that would serve as a framework for the Challenge4Cancer contest and its projects, all the while providing some leeway for the creation of innovative applications, even in a contradictory society who massively shares personal information all over social networks while being very reluctant to share their medical data.

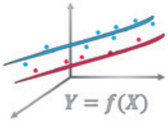
The eight projects that were selected as finalists provide us with an optimistic perspective as to what big data can offer to the world of medical research, especially in the field of cancerology. The projects CancerViz and Viz4Cancer allowed us to see to what extent visualizing data properly can be important. When an enormous amount of data is being processed, it is fundamental to be able to extract relevant data without losing too much after time or getting confused. Moreover, with the Baseline or Risk and Predictive Approach to Cancer projects, we saw the relevance of using specific algorithms to identify new correlations between cancer and risk factors to date unsuspected. Big data is not a solution to each and every problem, but it is a tool that will allow us to work more efficiently in many healthcare fields.

This first edition was a success in its approach, what it achieved, and the prospects it offers for the future. Big data is no longer a buzzword but a reality that brings tangible value to researchers, and therefore, ultimately, to patients. Furthermore, Epidemium met the challenge of making concepts such as open science and collaborative research, still largely unknown the general public, more accessible. Let's continue the good work and develop more ambitious objectives, bring out the best potential of big data and together, push forward medical research. ■

Baseline: creating a model on the incidence and mortality rates of cancer

// AUTHORS

Édouard Debonneuil / Augustin Terlinden
Peter-Mikhaël Richard

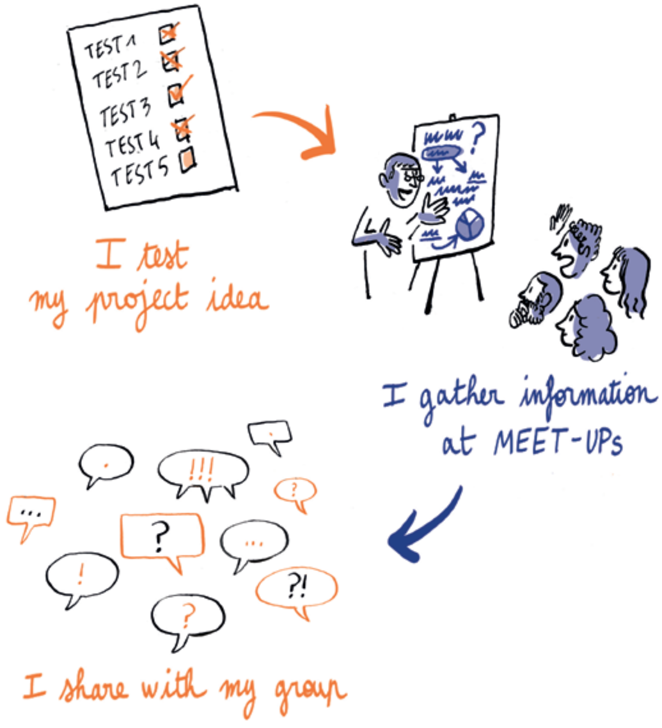


The goal of the Baseline project is to predict the incidence and mortality rates of cancers on the basis of risk factors identified by looking at open data that have a global reach and that can be categorized regionally. Why cancer and why aggregated data? First of all, the diseases linked to cancer are poorly understood although they cause millions of deaths every year. Moreover, individual data are rarely available and aggregated data under-utilized, as epidemiologists believe, among other reasons, that they bring methodological bias. Therefore, our vision was that the solution to one of society's greatest challenges was there for us to grab!

The project's first step was to build a large and robust database to be used by the scientific community. We collected aggregated data from many public websites such as the Centers for Disease

Control and Prevention (CDC), the World Health Organization (WHO), and the French association Seintinelles <www.seintinelles.com>, etc. Working on these, we were then able to develop a predictive epidemiological model whose goal was to identify trends and connections between different variables. This model enabled us to investigate various risk factors including: alcohol consumption, long-term unemployment, blood pressure, cholesterol level, the date of marriage for men and women, and affiliation with a specific ethnic group. Finally, we wanted to have a validation process using anonymous and individual data but this was postponed and later abandoned due to a lack of time. However, we are confident that later on, these data will be used to create decision-making tools for general doctors, or for political advisors in choosing in which therapeutic fields they should allocate budgets.

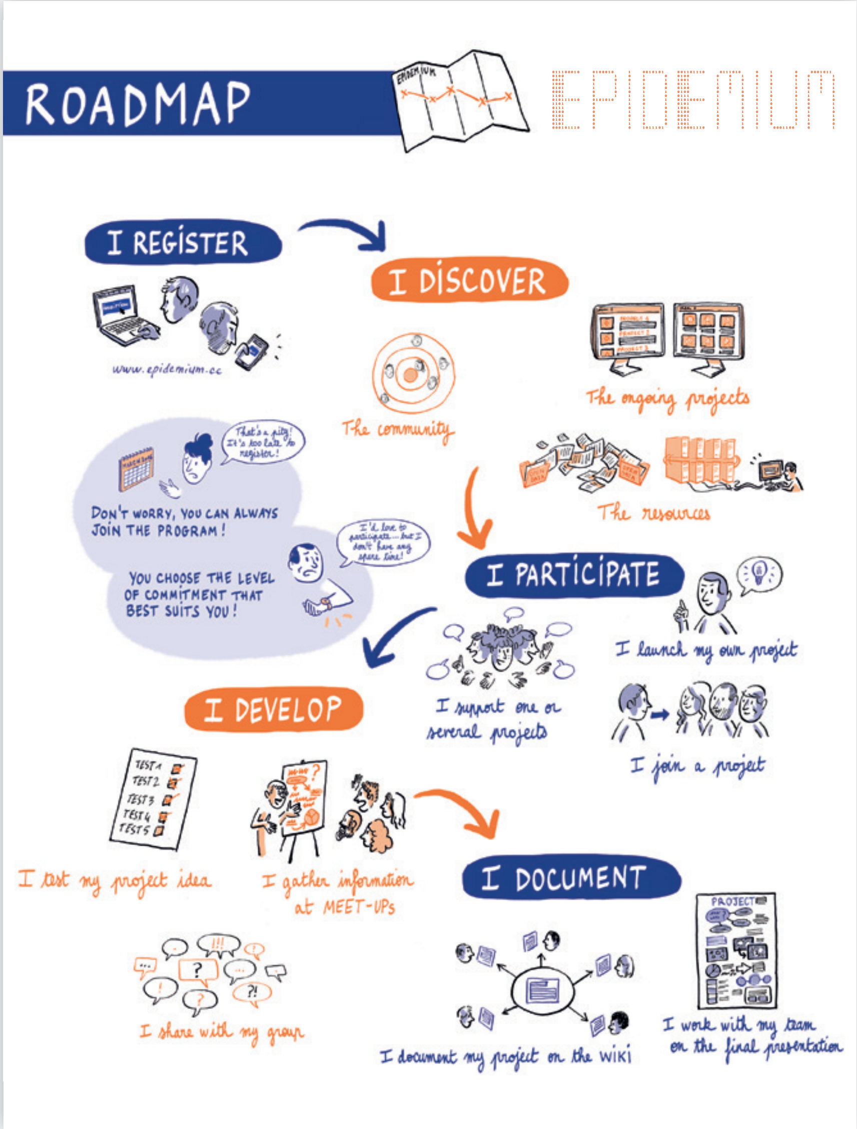
The project attracted many people with complementary backgrounds, including healthcare professionals (general medicine, public health, oncology, epidemiology), statisticians (data architects, economists and actuaries), developers (Web, R / Python, machine learning and data-visualization) and communication professionals. Everyone had the opportunity to participate in this ambitious project. Many lessons were drawn in terms of team management. The ability to motivate a team of fifty people volunteering in top of their main activities constitutes a real challenge. We would also like to use the opportunity of this feedback to thank all of the different contributors who have helped us. ■



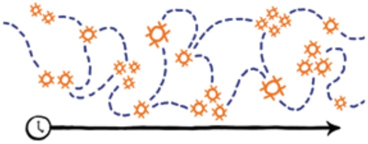
The 4 worksheets

Challenge4Cancer Roadmap

2a


ROADMAP
EPIDEMIUM
EPIDEMIUM
I REGISTER
www.epidemium.cc
I DISCOVER
The community
The ongoing projects
The resources
That's a pity! It's too late to register!
DON'T WORRY, YOU CAN ALWAYS JOIN THE PROGRAM!
I'd love to participate... but I don't have any spare time!
YOU CHOOSE THE LEVEL OF COMMITMENT THAT BEST SUITS YOU!
I PARTICIPATE
I launch my own project
I support one or several projects
I join a project
I DEVELOP
TEST 1
TEST 2
TEST 3
TEST 4
TEST 5
I test my project idea
I gather information at MEET-UPs
?
?!
I share with my group
I DOCUMENT
PROJECT
I document my project on the wiki
I work with my team on the final presentation

2b Challenge4Cancer projects



Topic 1: Understand the distribution of cancer across time and space

Viz4Cancer: interactive data-visualization website representing the different data sets

// Project goal

- Graphically represent the evolution of the different types of cancer in France and its many social-environmental factors.

// Tools

- Data-processing Web API.
- Dynamic graphic visualization.

// What impact will it have?

- Allow multidisciplinary teams to possess a common visual language, available on the website: **viz4cancer.epidemium.cc**

Wiki: http://wiki.epidemium.cc/wiki/Equipe_Quantmetry

CancerViz : Speed up the data analysis exploration phase

// Project goal

- Offer a data-visualization tool that facilitates that data acquisition phase and initiate the first data analysis exploration phase.

// Tools

- A *full-stack technology* based on open-source, front-end data analysis and processing tools.

// What impact will it have?

- A multicriteria data visualization tools accessible via: **cancerviz.weareopensource.me**

Wiki : http://wiki.epidemium.cc/wiki/CancerViz

___ Topic 2: The risk and protective factors of cancer

Baseline: creating a model on the incidence and mortality rates of cancer

// Project goal

- Preventing cancer by comparing living conditions worldwide.

// Tools

- Techniques (Data Science Studio, MySQL and SQLite...).
- Collaborative (Hackathon's modelisation).

// What impact will it have?

- Better understand the risk factors taking into account living conditions and in order to better avoid them in the future.

Wiki: http://wiki.epidemium.cc/wiki/Baseline

Predictive approaches and cancer risks: measuring the influence of environmental factors on cancer risks

// Project goal

- Analyze the impact of carcinogenic environmental factors on the incidences of cancer.
- Build an indicator measuring how a population is exposed to certain factors.

// Tools

- Python codes and R.

// What impact will it have?

- Implementing predictive algorithms on the incidence of cancer.

Wiki: http://wiki.epidemium.cc/wiki/Approches_pr%C3%A9dictives_et_risque_de_cancer

___ Topic 3: Meta-epidemiology: understand cancer by looking at the scientific and medical literature

OncoBase: Produce quality information on cancer can serve as a common foundation for statistical analysis

// Project goal

- Standardize available heterogeneous data using automated data collection, aggregation, homogenization and unification.

// Tools

- Data analysis program.
- Parallelization and aggregation.

// What impact will it have?

- Database based on the scientific literature that is of better quality and streamlined, which makes it easier to avoid erroneous conclusions and improve research.

Wiki: http://wiki.epidemium.cc/wiki/OncoBase

BD4Cancer: combining big data analysis and BioNLP approaches for pharmacovigilance and pharmacogenomics

// Project goals

- Identify the undesirable effects of anticancer treatments.
- Extract knowledge from biomedical literature and clinical trials to predict new medical interactions.

// Tools

- Big data analysis environment
- Machine learning and NLP libraries. • Javascript libraries, ...

// What impact will it have?

- A real-time pharmacovigilance system and the prediction of new interactions between treatments

Wiki: http://wiki.epidemium.cc/wiki/BD4Cancer

__ Topic 4: Environmental changes and cancer

ELSE - Evolutive Life Selection Experience: an educational game based on a person born in 2000 who sees the evolution of their risks of cancer change depending on the choices they make

// Project goal

- Raise awareness on the risks linked to cancer using a fun tool.

// Tools

- Big data analysis.
- Graphic interface.

// What impact will it have?

- A pedagogical app published online and available for everyone to use: **conix.fr/epidemium/else.html**

Wiki: http://wiki.epidemium.cc/wiki/ELSE

Venn: gain a comprehensive vision of research in environmental oncology

// Project goal

- Based on the scientific publication abstracts available in the Pubmed platform, extract and analyze the links between cancer and environmental factors

// Tools

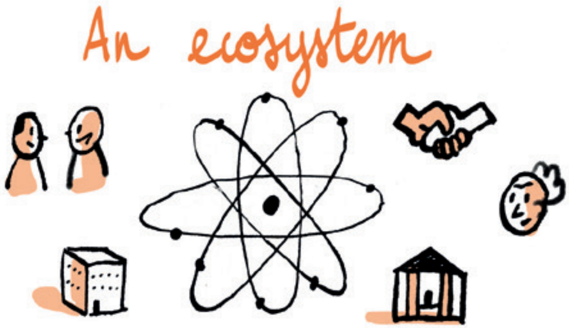
- Smart research tool.
- Machine learning.
- Biomedical text search tool.

// What impact will it have?

- An interactive web app to visualize keywords per topic: **venn-epidemium.github.io**

Wiki: http://wiki.epidemium.cc/wiki/Venn

Challenge4Cancer resources



Further details...

// How can we use data science and big data in healthcare?

- Kawamura T. (2016). “Big data system shows promise in helping cancer patients at Todai”, *The Asahi Shimbun*, September 19, 2016, available online <http://www.asahi.com/ajw/ articles/AJ201609190064.html>, last accessed on November 30, 2016.
- Schuyler D. (2016). “Is the Blockchain a Potential Cure for Securing Health care data?” on the *Leavitt Partners* website, available online: <http://leavittpartners.com/2016/09/is-the-blockchain-a-potential-cure-for-securing-health-care-data/>, last accessed on November 30, 2016.

// Crowdsourcing cancer epidemiology

- HealthMap <http://www.healthmap.org/fr/>, developed by researchers, epidemiologists and developers at the Children’s hospital in Boston, allows to follow in real-time the development and spread of a disease by gathering all kinds of data on the web.
- Wales J. (2008). “The wisdom of crowds” in *The Observer*, London, June 22, 2008, available online <https://www.theguardian.com/commentisfree/2008/jun/22/wikipedia.internet>, last accessed on November 30, 2016.
- Khatib, F., Di Maio, F., Cooper S., Kazmierczyk M., Gilski M., Krzywda S., ... & Jaskolski M. (2011). “Crystal structure of a monomeric retroviral protease solved by protein folding game players” in *Nature structural & molecular biology*, 18(10), 1175-1177.

#3

AN OPEN AND ETHICAL LEGAL FRAMEWORK

Epidemium, a unique, open and collaborative program, enabled its participants to tackle a set of unprecedented questions, both from a legal and ethical point of view. Open science does not mean rules do not exist. It is not just about respecting the law either, but about creating an environment in which participants will be able to share their work. Anticipating the ethical questions raised by Epidemium, a program that sought to use health data, represented an immense challenge.

// AUTHORS

Jonathan Keller / Jérôme Béranger / Me David Simhon
Jean-Frédéric Petit-Nivard

ethics
open
collaboration
legal framework
open science
law

Rules to encourage the open sharing of science and data

CONTRACTUAL ARCHITECTURE RULES

OPEN PERSONAL DATA

This contribution introduces to the readers the reasons of a contractual framework governing an open science program. This type of innovative research is based on the absence of legal framework but needs to take into consideration the question of access to resources. The goal is therefore to make contractual choices allowing this access and to take into account the results found by participants.

// **AUTHOR**

Jonathan Keller

One of the first challenges to tackle for Epidemium was to establish clear rules. In doing so, we had to find a balance between the need to create an environment where it would be easy to share and collaborate and the need to preserve the rights of creators and authors. In fact, it was essential that the rules were effective in encouraging the sharing of knowledge and data in a scientific context, in order to incentivize participants to join a dynamic and innovating community.

Our contribution focuses on the various solutions offered to address the challenges linked to the organization of an open science community. In the absence of a clear legal framework and to respect the principles laid out in the Epidemium Charter, the contractual component was favored.

In the specific case of Epidemium, we focused on the definition of new concepts incorporated in contracts, such as the boundaries separating the community from its members, as well as the rules that had to be followed for the management and sharing of resources, information and findings. The contractual obligations and the legal emptiness will be highlighted before answering the issues related to the management of intellectual property rights.

What type of contract use in the field of collaborative science?

Epidemium seeks to be a program that is open to all, in which all participants are free to choose their level of engagement based on their availability and skills. The contract, or rather the contracts imagined to address this need, must first and foremost define the limits of the community in order to give members the rights to access and use the resources provided by Epidemium and their partners.

For this task, we were not able to use the scientific literature on the collaborative economy and its creative networks of collaboration as a source of inspiration. A few authors have sought to define what a community is. Case law is not of much help either as there is no clear definition of what constitutes a community as no specific legal case has set a precedent. Moreover, social sciences deal greatly with this topic but the approaches are too broad to be useful in our case. Therefore, we proceeded in a pragmatic manner by defining the community and its members based on their needs. Using this approach, we determined individuals who were accredited to get access to the digital tools provided by the program's partners.

Defining the community members

In open source projects, the most conventional interpreters of the legal doctrine define a community as "any user involved in immaterial creation." This vision seemed to be too broad as it only applies to passive users. This passiveness implies that users are not making a contribution, neither directly by taking part in the project, nor indirectly by providing financial support or promoting the project. We are advocating in favor of a radically different version of what constitutes a community: according to us, a member of the community is anyone actively making a contribution to the project, regardless of the fact that their contribution is done directly or indirectly, or that their role is to directly intervene or provide support. The passive end-users, or those working on external projects, do not demonstrate a willingness to participate in the project, that is to say an *affectio communitatis*, which refers to the willingness to become a member of a community.

In order to encourage an open and community-oriented approach, we have distinguished between two different types of members. We started by the "participant" in the Challenge4Cancer contest. The participants must register on the platform[1]. They are also required to approve the Challenge4Cancer Charter and to create or join a team to collaborate on a project. This status is different from the status of "contributor", but one individual can also have both statuses. Therefore, the "contributor" is limited to the activity of contributing directly or indirectly to the program, mainly through their registration in the program's wiki[2], and for which they have to agree with the conditions of use.

Having contributors accept the different contracts fulfills the legal obligations with regard to web hosts, which is mandatory and required by the French Law for trust in the digital economy. This constrains the web host to regulate or sanction contributors, if need be.

Apart from this purely preventive aspect, the identification of contributors and/or participants give them the possibility to enjoy certain rights, such as the ability to take part in the contest, to win it, but most importantly to gain access to the resources graciously offered by third-party partners.

A conceptual framework to reassure partners

A few of the program partners raised the legitimate concern that contributors could pretend to participate to the Challenge4cancer contest in order to use the resources provided for their own personal needs. This fear exists in virtually all open science projects by definition since they have no barriers to entry.

We therefore had to take into account this aspect when we established partner contracts. We adopted three distinct methods, which could sometimes be combined:

- The program's organizers and providers jointly manage the allocation of resources after participants justify that there is a need;
- The provider monitors the resources used by the participants;
- A contractual obligation based on the stipulations included in the contest's rules and in accordance with the end user licence agreement for the use of the resources.

Now that we have, broadly speaking, defined the elements required for the implementation of Epidemium, other related elements, linked to the preparation and launch of the Challenge4Cancer contest, must be explored. These elements, proper to an open source community, refer to the rights of property provided under an open license.

The management of immaterial assets

The main principle of the Challenge4Cancer is to allow participants to use big data analysis techniques and apply them to cancer epidemiology. This means determining trends and correlations by combining different datasets. For these combinations to be possible, this information has to be made lawfully accessible, which depends on whether the data can be qualified as personal data or not. Beyond the issue of data is the question of the intellectual property of work produced collaboratively.

How to open the access to data

During Epidemium's program, the operational team has made the choice to establish various databases, on which participants would work. These were based on open and 'non-viral' data. The notion of virality corresponds to a licence that constrains future users to apply the original licence for all future uses or modifications of the common resource placed under the initial licence. This choice was justified for several reasons.

The first reason was to do away with the contractual obstacles placed by the different data producers. In fact, intellectual property licensing contracts contain destination clauses determining the end uses of data, which refers to the reasons for which they should be used. However, to foresee and to determine the exact end use of data can work to undermine the prospect of open research. Its objective is to discover unknown scientific findings by using common resources provided by participants. Therefore, apart from the framework given through the Epidemium Charter defined by the Independent Ethics Committee, the freedom of participants is absolute.

The second reason was that access to proprietary data, which contain personal data, were limiting in two different ways. First, it was limiting in terms of format. A classic licensing contract restrains the authorized number of copied data. This restriction leads to another. It forces eligible and legitimate users to be defined to access certain datasets. However, by definition, open data is indifferent to the identity of the end user. This indifference becomes relative if the open data includes personal data, especially when it comes to health data. The Epidemium program's goal is to invite participants to solely use open data.

The sharing of results developed by participants

Accepting the contest's rules was done on a purely voluntary basis. This, however, does not imply a renunciation of intellectual property in favor of an open license. In fact, intellectual property law includes rare cases of automatic cessions. As such, participants have voluntarily accepted to contribute to a common good by providing their ideas and research and to have them distributed under an open licence.

Before delving into the question of the different forms of open licences, we have to clarify that for a long time organizers had the intention of encouraging participants to contribute under free license. Nevertheless, this choice would have had a negative impact by potentially psychologically deterring third-party users, that is, anyone interested but who is not formally part

I was invited to present project ConSoRe during a meetup organized by Epidemium. There, I could see the value and interest of the community. I met a lot of people at the event but I was also contacted by others after the event. Nowadays, more people understand the opportunities that new technologies offer. Many seek to utilize these new tools to disrupt science. This is what I discovered at Epidemium."

Dr Alain Livartowski
Oncologist, Co-Director of Data at the Institut Curie

of Epidemium's program. Finally, free and open models differ in terms of obligations. In the free model, there is an obligation for users to place their findings in the public domain when re-using results. Although this offers the possibility to maintain knowledge in an artificial public domain, the obligation of reciprocity included in free licences are perceived, rightly, as an impediment to the conduct of further research.

Finally, the precedent set by the European Union Court of Justice has clarified that any content that is the result of collaborative work must be protected in an autonomous manner. For instance, a data visualization of the project will be protected by a different license than for the text commenting it, or the code needed to embed into the website. As such, depending on the element being protected, the license differs. In these different cases, the starting license will be open, meaning that it will be accessible and reusable under no other condition than having to be attributed by previous contributors. For softwares, we recommend licenses certified by the Open Source Initiative[3] license Expat/ MIT, the Berkeley Software Distribution (BSD) or the CECILL-B; for texts and images, we invite participants to give priority to the Creative Commons Attribution (CC- BY) license; finally, for databases, we suggest selecting the fourth version of the Creative Commons license with the obligation of mentioning the author or using the Open Data license.

Conclusions

Judging by the quality of the discussions we had throughout Epidemium's program as well as the projects launched during the Challenge4Cancer contest, it seems that the rules adopted by the community were successful in the way that they reassured partners and stakeholders without restricting the creativity of participants, and staying in line with the principles expressed in Epidemium's Charter *(see* ***worksheet n°3a****, page 137)* and, of course, with the Law.

Our first concern was to define the different profiles of community members in order to grant them the right to participate in the Challenge4Cancer, to directly and/or indirectly contribute to the program, as well as give them access to the resources while reassuring partners as to whether they will be used properly. Then, we tackled the daunting question of rights governing the use of databases. Finally, in order to increase the positive results obtained during the open science program, participants had to document their work, and provide information on the algorithms developed and the open data used. Final results are also shared through an open license.

Although we were satisfied with the first results, we consider that we are only at the beginning of a process which is meant to last. During the Challenge4Cancer contest, we have raised questions that are at the heart of the open science movement. ■

> *"We had to find a balance between the need to create an environment where it would be easy to share and collaborate and the need to preserve the rights of creators and authors."*

[1] Epidemium website, <www.epidemium.cc>, last accessed on November 30, 2016.
[2] Epidemium website, Wiki part <http://wiki.epidemium.cc/wiki/Accueil>, last accessed on November 30, 2016.
[3] Open Source Initiative website, The Open Source Definition (Annotated), <https://opensource.org/osd-annotated>, last accessed on November 30, 2016.

What should the ethical guidelines be for an open and community-based approach of the use of big data in healthcare?

ACCOUNTABILITY ETHICS OPEN DATA

NTIC EMPOWERMENT

Ethical questions related to digital technology are concerned with the behavior and practices of individuals when using digital tools, along with how digital tools are becoming more and more autonomous in the way they operate. In the context of big data analysis in the healthcare sector, ethics serves the purpose of regulating behavior based on the respect of human values that are deemed essential, and provide a moral framework for the use of digital data.

// **AUTHOR**

Jérôme Béranger

#3

AN OPEN AND ETHICAL LEGAL FRAMEWORK

Are there ethical concepts uniquely associated with digital technology? This vision is always up for debate. It seems natural to associate a human science with a technological science that are in reality very different. Nevertheless, digital technology has led to contradictory injunctions that have had specific ethical repercussions on New Information and Communication Technologies (NICT). Although big data is ethically neutral, their use is not. This explains the need for Epidemium to establish a framework and an ethical charter addressing the issues raised by the use of different data throughout the program. In fact, unique behaviors are generated by the new understanding of space-time created by digital technology. NICTs are both a cultural and anthropological phenomenon. They generate new behaviors, new visions of the world, and new social norms.

We can take the example of anonymization, which raises the question of the responsibility of individuals when their invisibility can free them from certain rules of propriety. The ubiquitous and instantaneous nature of the internet makes our acts of expression consequential and irreversible. From now on, digital technology and ethics cannot be thought of separately. Ethical questions must become an integral part of digital technology and lead to fruitful discussions. As such, we are not referring to an interdisciplinary approach but a merge of the two considerations, in which moral and social impacts are fully integrated into NICTs.

In this context, it becomes essential to clearly define expectations and ethical rules that apply to the use of digital technology. We must also build an entirely new system of ethical and moral values revolving around one central question: can digital technology lead to wrongful ethical behavior?

For a couple of years now, open data has become a new development in the field of big data. This is due to the fact that data is increasingly becoming more reliable, upright, and pure, which effectively addresses the question of data quality (Hamel and Marguerit, 2013). This represents Epidemium's reason for being. Its ambition was to use data in an open and shared framework in order to produce quality results based on the power of interdisciplinarity.

What should the ethical guidelines be for an open and community-based approach of the use of big data in healthcare?

This practice consists in making accessible to all, easily and freely (without any legal, technical or financial restrictions) digital data, generated by a public institution or community. The idea of freedom of access and opening up access to third-parties is part of a larger trend that, at its core, perceives information as a public good, and that the spread of information acts in favor of the public interest. The theory developed by the American sociologist Robert King Merton sheds light on the benefits of opening scientific data. According to the theory, researchers must contribute to the common good and renounce intellectual property rights in order to increase knowledge. Open data in the way it is taking shape today is a direct result of the convergence of this scientific idea and the ideals of open software. The creation of value stems not so much from the sheer volume of existing data, but rather from the sharing of data, the fact that they are given to a third party as a result of participation and collaboration. Open data is an attempt to go against the conventional wisdom: by default, data and public information must be published online, even before they are requested by a third party. This marks a profound shift in terms of cultural practice.

As such, open data can be perceived as a contemporary idea in which the imperative of transparency, accountability are becoming increasingly important. Transparency is linked to the mistrust, or the suspicion toward institutions and their representatives. This movement seeks to address a set of economic and political challenges. Open data is expected to lead to democratic benefits (a more transparent public-policy making process, increased trust with regard to elected officials and their institutions) but also the creation of economic value through the development of new activities based on the use of open data.

Therefore, in the field of research and public health, open data holds many promises: extend the indications for the use of medication, to preserve public health by identifying impactful events, or to the contrary, detect weak signals such as the spread of an epidemic or a biochemical attack, to manage with greater accuracy public health policies in order to better adapt them to national or local needs, or to improve food safety by following recommendations for the use of health products. Consequently, open data facilitates the follow-up of national food safety measures, increases the visibility of risk factors at different levels, and makes it easier to detect and deal with epidemics. From an economic point of view, opening up data improves budget management, ensuring that the treatment prescribed are appropriate, safe, and of quality. Finally, those advocating in favor of open data insist on the democratic aspect related to this practice, and emphasize how important it is that public institutions remain transparent.

Nevertheless, open data is today facing many challenges and interrogations, both in terms of demand and supply. The data offer is underdeveloped: most players have prioritized opening data that is easier to obtain from a technical, legal and political standpoint. Sensitive digital data, or those that have a strong social or societal impact remain largely excluded from open data.

Some data is complex to comprehend if we do not know the context in which they were initially used. Wouldn't it be risky to make them accessible by everyone? Wouldn't we risk distorting the data by interpreting it? As a matter of fact, letting data be used by the greatest number poses the difficult question of the culture of data. A variety of skills are required: knowing how to identify the sources of data, being able to process it, manipulate it, and hold a critical stance with regard to the conditions in which they are being produced and opened, but also having a good command of basic statistical concepts, etc.

The issue of confidentiality and misuse should not be disregarded either. The reason being that it may be possible to "indirectly identify a person" by way of cross-referencing different sources of information. Another fear arises from the possibility of having private players use the data for their own purposes. Security aspects are therefore extremely important and should be taken into account when looking at how to utilize open data.

Moreover, in ethics, the term "value" is prescriptive. It is used as a referential allowing one to assess facts. It presents

> *Apart from the results and their methodological aspects, we look at questions of governance, the ethical considerations related to accessing data, and the sharing of results which will be raised by certain solutions and applications offered by the candidates."*
>
> ***Nicolas de Cordes***
> *Member of the Scientific Committee*

What should the ethical guidelines be for an open and community-based approach of the use of big data in healthcare?

a series of ideals to pursue. The word has a general and dynamic connotation. The first meaning is first and foremost philosophical, even before we start looking at the ethical consideration it implies. Ethics is founded upon the idea of rational agents. This idea revolves around a common understanding of the rules governing coordination, exchange and sharing among the parties involved in the process. Each individual contributes to the search of a common understanding of the situation that needs to be analyzed. This requires a high level of solidarity and consensus in terms of what the final goal should be. If ethics is by nature hard to define, applying it to digital technology is another challenge. Ethics requires a vision, an outlook, an ambition that becomes concrete when it is turned into a sense of direction.

No technology can be perceived as purely instrumental. This is especially true of large and automated information systems designed to help manage and integrate large organizations, such as health institutions. In this context, the setting is mostly made up of individuals. With information systems evolving, the human factor simply governs over technical factors. Although their satisfaction is mandatory, it is never fully enough. In all projects involving big data, the human factor and the interaction of man and machine are fundamental aspects. However, when there are multiple users working simultaneously, interaction between different individuals becomes the main question to solve. Evaluating large datasets, such as the ones found in the field of healthcare, is based on the notion of inter-human relationships (Fessler and Gremy, 2001), which implies the conception, the implementation, and the use of big data. Under these conditions, big data appears as a social system, marked by psychological, sociological and ethical characteristics, which we can use to start formulating our approach and determine the specific ethical guidelines that should be respected in the field of digital technology.

Ethical considerations linked to NICTs can be broken down into three main themes:

- The ethics of data: defining the principles ensuring the fair processing of data and the protection of individual rights when using data for scientific or commercial purposes;

#3

AN OPEN AND ETHICAL LEGAL FRAMEWORK

What should the ethical guidelines be for an open and community-based approach of the use of big data in healthcare?

- The ethics of algorithms: referring to the study of the ethical challenges and the responsibility of authors of scientific data, concerning the unforeseen and undesirable consequences, and the missed opportunities for the conception and implementation of complex and autonomous algorithms;
- The ethics of practices: identifying the appropriate ethical framework required to shape the deontological code related to the governance and management of data, all the while promoting the progress of data science and the protection of the individuals involved.

Subsequently, the technological revolution occurring in the information sector must be done in the interest of patients and for the purpose of improving patient care. In other terms, the only value to take into account is the individual considered in their dignity and as moral beings. The notion of human dignity constitutes an absolute value that we give to the individual. Therefore, ethical, practical, technical and ergonomic principles must be imposed in order to ensure that patients benefit the most from this technological revolution. This is especially true given that any ethical reflection is based on conflicting human values. Regardless of our religious and cultural identity, our political orientation or our field of work, our emotions are what constitute our deepest values. Pierre Le Coz (2010) highlighted during his first ethical day "Cancer and Fertility" at the Institut Paoli-Calmettes, "without emotion, there can be no formal values, and therefore no ethics."

Each main principle can be associated to a specific emotion:

- Respect for: the principle of Autonomy (voluntary and informed consent);
- Compassion for: the principle of Charity (the legitimacy of an action);
- Fear of: the principle of Wrongdoing (not to cause harm);
- Indignation for: the principle of Justice (based on equity and equality).

Moreover, in the case of technology, ethics deal with acts, actions that have a social consequence that cannot be foreseen and that include future knowledge, regardless of whether it is incomplete, and that also influences what was known before. It can be defined as a system of thought used to reflect on the moral meaning of an action. This definition is intentionally broad and fundamental and includes several components borrowed from the field of computer ethics (Waskul and Douglass, 1996). There are five main ethical applications with regard to NICTs:

- The ethics of empowerment: associated with the patient/ player (e-patient) who asks for autonomy and dignity (respect for their rights);
- The ethics of access: as a fundamental right and in respect of the principle of transparency (Universal Design);
- The ethics of dissemination: related to a shift from control-oriented information technology to service-oriented information technology (centralization and distribution);
- The ethics of reappropriation: focusing on shifts as potential (digital literacy);
- The ethics of collaboration: revolving around the sharing of information (on the internet, especially in online forums or social networks).

Finally, as we have just seen, information becomes the main object of moral action. Applying ethics in the realm of digital is a non-natural act due to the fact that NICTs would be bereft of all social and human value. This idea is the result of a common discussion that perceives all technology as ethically neutral, since only human beings can bring meaning to their actions. Nevertheless, we observe that big data brings value to the extent that it influences and conditions the way users behave. Consequently, no digital data can be seen as entirely neutral. This is why it is difficult to boil down digital ethics to the expression of extrinsic values about the good use of technology, but also linked to their intrinsic values. Finally, with the advent of "massive data" in the digital world, entirely new ethical guidelines have to be drawn up since NICTs are bringing about a relational and sociological paradigm shift (Doueihi, 2013).

We do not have the pretension of having re-invented ethics. Our goal is rather to update the existing ethics in view of moving toward what we call "algorithmic ethics," which would be

> *Although big data is ethically neutral, their use is not. This explains the need for Epidemium to establish a framework and an ethical charter addressing the issues raised by the use of different datasets throughout the program."*

exclusively applied to the field of digital technology. This new approach aims to include ethical values and principles for the conception, implementation, and practice of big data analysis in the field of healthcare. ■

What should the ethical guidelines be for an open and community-based approach of the use of big data in healthcare?

Epidemium Charter: when ethics serves the purpose of perfecting the law

// **AUTHOR**

Me David Simhon

What are ethics?

Let's do something crazy. Open a dictionary at the letter E. Not any dictionary. THE Dictionary, that is, the dictionary of the French academy. At the letter E, search for the word ethics, among other things it says "the science of morality." In the same dictionary, the term morality includes "doctrine related to customs and ethics." I can imagine readers wisely replying: this is illogical! Why use two terms when they are referring to the same notion? Are ethics and morality the same thing? I believe (along with a few other philosophers) that frankly, it is not the same thing! There is a world of difference between the two terms.

> *"The alliance between big data and medicine is extremely powerful but must be rigorously regulated by principles used in medicine for a long time."*
>
> ***Pr Cédric Villani***
> *Member of the Independent Ethics Committee*

Morality is about good and evil. And ethics is about right and wrong.

In other terms, ethics is always relative to a given time and place. 80 years ago, abortion in France was not ethical. The death penalty was. Still today, in some societies unknown to the "modern" world, it is ethical to eat the corpses of your ancestors! This seems terrifying and unthinkable for us in the West. However, this seems perfectly ethical for these people.

In contrast, morality is more about what is absolute and universal. An example of this is "Thou shall not kill." This moral imperative **must** apply regardless of the time, place or circumstance.

"What about law, then?" is the question that I would raise, as a legal expert. The Law is neither ethics nor morality. Immoral or amoral actions can be legally authorized (not honoring your parents, etc.). Conversely, regulating exhaust pipes or the size of tomatoes is not a morality issue. And most importantly, we cannot force anyone to be moral, whereas we can constrain someone to respect the Law.

In this game of definitions, ethics seems closer to law than morality. Ethics can, in some cases, surpass Law, surpass itself, or can be complementary.

- On the question of abortion, ethics was probably a bit ahead of law. Before the Loi Veil in 1975, the interruption of pregnancy had already started to be tolerated by society.
- On the question of homosexuality, a lot of time had to pass for mentalities to change and influence our sense of ethics: 1982, decriminalization of homosexuality; 2015, mariage modestly called "for all".
- In France, toward the end of the 70s, was it ethical to guillotine prisoners? Probably. It was actually very legal and precisely laid out in article 12 of the penal code: "Anyone sentenced to death will have their head cut off."

On October 9, 1981 the death penalty was abolished. I am not fully convinced that France's ethical compass changed between the 8th and the 10th. Only a few years later did we collectively consider the death penalty as non-ethical. This time, Law was ahead of ethics.

Sometimes, ethics serves the purpose of perfecting the law. Let's admit it, practically speaking, this is when things get exciting. Ethics can examine on a case by case basis situations that have yet to be examined from a legal standpoint. Where the rule of law cannot spell out the details, ethics can be used to complement it. As such, in the name of ethics, you can prohibit certain acts - the use of data, to use a case relevant for the challenge - that would be authorized by law.

Why think about regulation and ethics beforehand?

On certain topics, once you have violated Law or the ethical norm, you are facing the consequences but it is already too late. The damages have been done, and sometimes this is irreversible.

If you could collect, without control nor filter, the nominative and personal health data of the entire French population, no matter how much of an illegal act this would be, the damage would have already been done. In the era of the internet and the cloud, data can potentially be accessible by anyone for years, if not decades.

It is important to define beforehand what is not acceptable and to make effort not to reach this limit. The idea is to adopt a strategy to anticipate potential damage.

On the importance of creating a multidisciplinary Ethics Committee

It was paramount to draw a line while giving participants enough freedom to work and innovate. We needed a "benevolent hand," the one that holds but does not clench.

Who is allowed to define ethical limits? Researcher themselves? They would then be both judges and jury by

Ethics is not a "parrot" of regulation. We cannot go against the law. But we cannot allow ourselves to go beyond it."

their involvement in the process, which would question their objectivity. Legislators or executive powers? This cannot work. It wouldn't be about ethics anymore but about the rule of law. The program's coordinators and partners? They would not be objective either. From this conclusion came the idea to call on qualified individuals, neutral, trustworthy and independent third-parties assembled in the form of a committee. That is how Epidemium's Independent Ethics Committee was born, at least on paper.

How to establish it? International conventions interested in questions of bioethics insist on the necessity of multidisciplinarity. There needs to be different perspectives looking at the same problem. In France, the ethics committee in the field of biomedical research, the CPPs (committee for the protection of people), are structured into two colleges (scientific and non-scientific). Within those two colleges, there is a willingness to bring together individuals from different backgrounds and fields of expertise *(see box below)*.

The independent ethics committee created during Epidemium did not seek, or could not organized itself so precisely. Epidemium's organizers took the opposite approach to

I. The first college was composed of:

1. Four individuals having in-depth qualification and experience in the field of research related to dealing with people, including at least two doctors and another person qualified due to their skills in the field of biostatistics or epidemiology;
2. A general practitioner;
3. A hospital pharmacist;
4. A nurse.

II. The second college was composed of:

1. A person qualified in ethical matter;
2. A psychologist;
3. A social worker;
4. Two people qualified in legal matter;
5. Two association representatives approved by the healthcare system.

legislators (could we have expected anything different from La Paillasse!) by first determining the individuals, and then defining member categories. But almost instinctively, they sought to call upon **different** skillsets that were in the end **complementary**: mathematicians, legal experts, patient representatives, practicians, big data and innovation specialists, entrepreneurs, and ethical experts, etc.

This committee was as little structured in the way it constituted itself as in the way it functioned: no president (by unanimous decision of its members), free discussions, few meetings but many exchanges by email. An "ethical melting pot" or an organized mess, which was in the end in line with La Paillasse’s DNA.

As explained earlier, tackling ethical issues during Epidemium aimed to define the different limits of what can be prescribed by the law: Ethics is not a "parrot" of regulation. For sure, we could not go against the law, but we could allow ourselves to go beyond it.

Wherever the law prohibits or allows, as a monolith, ethics is used to provide a framework.

To use a concrete example, we had to deal with a question related to the use of ethnical data. The *Loi Informatique et Libertés* allows, under certain conditions, the processing of such information.

Loi informatique et libertés, article 8: It is prohibited to collect or process personal data in which the ethnic or racial origin of the invidividual appear, unless the final use of the process requires it for certain categories of data. This includes the processing of research data in the field of health.

The rule of law is known, or at least accessible, by those who have an interest in it. However, from an ethical point of view, are we allowed to use such data? The answer is not so simple. It is known that for certain pathologies, the black population is

Epidemium Charter: when ethics serves the purpose of perfecting the law

more exposed than the caucasian population. Can we look at the question of processing data from this perspective?

We have tried to adopt a rational and pragmatic approach: we can use the information available, but not in an isolated manner. We asked that the data be correlated with the environment and the population's standard of living, in order to avoid having a purely biological perspective (and be accused of eugenics).

As we received questions from a variety of people, including organizers and participants, the committee decided to create its own "jurisprudence": a set of rules that we deemed important enough to be written down. We considered these principles to be fundamental in 2015-2016 and for Epidemium. But what is true in 2016 in France may not be elsewhere and is most likely to change over the next couple of years. This is what makes the drama, and beauty, of ethics. ■

Roche creates an open database to support Epidemium

// **AUTHOR**

Jean-Frédéric Petit-Nivard

As the project initiator, and Epidemium being an initiative on open data, it seemed essential for us to pave the way and try to open our data for the greater benefit of science. Roche France is proud to be the first pharmaceutical company to open its data to foster research on cancer through the creation of Roche Open Database, an open database made accessible on Epidemium's open data platform.

We would like to share with you our experience and hope that our testimonial will prove to be useful.

The project is based on a strong and shared conviction: open data is an incredible accelerator for science and specifically for cancer epidemiology. By making our data accessible to Epidemium, our ambition was to set a precedent in France and encourage other to try, boosting research on cancer and ultimately serving patient needs.

In order for Roche Open Database to succeed, we had to set up an internal team with expertise in medicine, law, regulation, and data analysis. Our Correspondant Informatique et Liberté played an essential role in the project by liaising with the French local authority in charge of computer technology and freedom , the CNIL (Commission Nationale de l'Informatique et des Libertés). Their assistance and advice proved to be decisive in the project's success.

Before launching the project, we first sought an internal approval from Roche Data Sharing Global team. Their reply was quick and positive which was reassuring since the approach we proposed was not part of Roche's existing Data Sharing Policy guidelines[1].

Once the authorization was received, we were able to get down to work. Roughly speaking, the project can be broken down into two parts: one is legal and the other technical.

Legal aspects

To create Roche Open Database, the first step was to define the project's legal framework.

The legal framework was based on two fundamental notions. The first was the processing of data in accordance with rules laid out by the CNIL and the second was the patient's consent. The law has clear rules concerning the use of health data with the purpose of safeguarding the interest of patients. *"Any operation [...] involving the collection, recording, organization, conservation, modification, extraction, consultation, communication, comparison, interconnection, locking, removing and destruction"* has to be authorized by the CNIL.

To obtain the most comprehensive database possible, our aim was to gather data from different clinical studies and before starting the anonymization process. However, grouping and anonymizing data corresponds to the 'processing' of data as it is laid out in the *Loi Informatique et Libertés*. We therefore asked an authorization to the National Commission on Computer Technology and Freedom (CNIL).

Many elements had to be included such as the project's final goal, how patients would be informed, data included in the scope, and an assessment of data anonymization based on the criteria defined by the G29 (individualization, correlation, inference)[2].

To better understand these terms, here is a brief definition provided by the CNIL[3]:

- **Individualization**: is the possibility to isolate an individual;
- **Correlation**: is the possibility to cross-reference the distincts datasets of one individual;
- **Inference**: is the possibility to deduce information about the individual.

The authorization request must also justify the law it is referring to.

After having studied the various options on the table, it seemed that article 8 offered the best alternative and seemed most appropriate to our project. We identified two possible options: "very short delay anonymization" and a request for anonymization justified by a project serving public interest.

Following the advice of an external counsel, we submitted our request using the “very short delay anonymization". Because these anonymization techniques are mostly used for financial transactions, the CNIL refused our first request.

Based on their recommendation, we submitted a second request justified by public interest and we received the green light in February 2016. Once we had the CNIL's authorization, we were able to start building Roche Open DataBase.

Technical aspects

Building the database was done in four major steps:

1. Choice of data
2. Transforming and regrouping data
3. Anonymization
4. Validation

1. Choice of data

We focused on finalized non-interventional studies conducted in France starting after 1999 in Oncology. We then isolated inclusion data[4] from the studies that were relevant for research in cancer epidemiology.

After our analysis, twelve studies were included in the scope representing about 8000 patients.

2. Transforming and regrouping data

To build this new database including data from twelve different studies, we had to tackle two main challenges: obtain a common structure with the same nomenclature, and preserve the specificities of different pathologies.

To obtain a common structure and a unique nomenclature, we used as reference the standards defined by the Clinical Data Interchange Standards Consortium (CDISC[5]) which is a standard for data storage in clinical studies. This standard is required for submission to the Food and Drug Administration (FDA[6]) in the United States and facilitates the cross-referencing and exploitation of clinical data. It helps to improve the efficiency of clinical research.

The other challenge we had to tackle was to integrate datasets into a common structure while preserving the specificities of each pathology. In Roche Open Database, five pathologies were represented, including colorectal cancer, follicular lymphoma, neoplastic diseases, lung cancer, and breast cancer.

Moreover, data collected within these studies depends on their final goal, and for ethical and efficiency reasons, only patient data that is strictly required is collected. However, apart from standard data such as demographics, most variables are specific to each study. As a consequence, the increment in data we had hoped for by combining the 12 studies fell short of our expectations, and some of the variables in the combined database were provided only for a very small percentage of patients.

3. Anonymization

After our analysis, we identified two ways to anonymize our data : the first would preserve the database's granularity (1 line in the database corresponds to 1 patient) using know masking techniques, whereas the second combines the data from several patients to create an aggregate (1 line in the database corresponds to many patients).

The first option makes it possible to preserve a greater amount of data. This is why we naturally decided to investigate this option first.

To anonymize a database, the first step is to remove any direct personal identifier, and then mask any indirect personal identifier.

Here is a quick definition to better understand these terms:

- **Direct personal identifier** is a data that can be used to re-identify an individual directly, for example by their first/last name or their NIR[7].
- **Indirect personal identifier** is a data point that when combined with other data points can re-identify an individual. For instance, the date of birth is an indirect personal identifier, as the combination of {date of birth, place of birth, zip code, and gender} would be enough to re-identify an individual in a database.

In our case the first step was already done. In fact, all clinical studies are 'pseudonymized', which means that all direct personal identifier are replaced by a random value.

The second step, which consists in hiding indirect personal identifier, seemed much more complex to implement. The available techniques were not deemed sufficiently robust, and we concluded that none would guarantee a perfect anonymization of our database[8].

Therefore, we started working on the second option: aggregation.

The method consists in generating statistics describing a subset of patients sharing common characteristics. This method offers an undeniable advantage in terms of the solidity of the anonymization process, although it limits the extent to which databases can be cross-referenced.

This is the road we decided to resort to for the anonymization process.

This step allowed us to generate Roche Open Database, a new anonymized database made up of aggregates.

4. Validation

The final step in the anonymization process was to validate the database by identifying potential errors or any abnormal distribution that could compromise the anonymization process. Our expert data analysts focused on two main elements. The first was to ensure that each statistic was generated from enough patients, which in our case was a minimum of ten.

The second was to verify that continuous variables had values that were scattered enough in order to avoid particular cases, such as a dirac distribution[9].

This verification marks the final step for the creation of Roche Open Database. Once created, the database was shared with all participants of Epidemium's Challenge4Cancer contest.

Roche Open Database is a truly audacious initiative to support science. By making our data on cancer accessible, Roche is supporting France’s Cancer Plan, which calls for data to become more accessible for it to be shared and put to good use by the greatest number of people possible.

This project would not have succeeded without the support of the CNIL and the involvement of a dedicated and multidisciplinary team. Beyond the creation of the database, the main achievement lies in the ability to demonstrate that these types of initiatives are feasible. We hope that our testimonial will encourage other similar initiatives in the future for the benefit of science. ■

[1] Roche already shares aggregated clinical data with the general public via Clinical Trials, a service provided by the National Health Institute in the US <www.clinicaltrials.gov>, and patient data with other research organizations via Clinical Study Data Request <www. clinicalStudyDataRequest.com>.

[2] G29, is the Article 29 working group on the protection of data for EU state members, <http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083>, last accessed on November 30, 2016.

[3] "G29 published recommendations on anonymization techniques," CNIL, article published on April 16, 2014, available online. <www.cnil.fr/fr/le-g29-publie-un-avis-sur-les-techniques-danonymisation-0>, last accessed on November 30, 2016.

[4] We call "inclusion data" patient data collected at the beginning of a clinical trial. They can be demographic, physiological, age, gender, size, weight, etc.

[5] Souza, T., Kush, R., & Evans, J. P. (2007). “Global clinical data interchange standards are here!” Drug discovery today, 12(3), 174-181.

[6] The Food and Drug Administration is a US agency for the protection of food and medicine, see <https://fr.wikipedia.org/wiki/Food_and_Drug_Administration>, last accessed on November 30, 2016.

[7] INSEE registration number, commonly called "social security number."

[8] G29, Opinion 05/2014 on anonymization techniques, adopted on 10/04/2014, WP 216, available online <http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index_en.htm>, last accessed on November 30, 2016.

[9] Delta function wikipedia page <https://en.wikipedia.org/wiki/Dirac_delta_function>, last accessed on November 30, 2016.

The 2 worksheets

Epidemium's Charter

3a

Epidemium's Ethical Principles

Paris 2015-2016

Access to big data is a major source of progress in understanding diseases and their epidemiological determinants.

As with every innovation, the use of anonymous massive data must be carried out in strict compliance with ethics, confidentiality, the protection of privacy and the legal or regulatory guidelines in effect. This must be carried out in all phases including: collection of data (whether it is public or not), analysis, use for epidemiological purposes, healthcare improvement, etc.

The purpose of the Independent Ethics Committee of Epidemium is to ensure that ethics is respected throughout the life of a project. The major ethical principles that guide Epidemium are outlined below.

// **Project managers must respect:**

- The legal and regulatory guidelines in effect
- Declaration of links of interest
- Confidentiality and privacy of the human being (for both public and private data sources)
- Pre-existing ethical rules of the data used
- Integrity and transparency in the collection, analysis and processing of data
- The principles of beneficence and non-malfeasance through the evaluation of risk vs. benefit
- A commitment to share work documentation as well as year-end results and conclusions.

The committee will be responsible for ensuring that all submitted projects comply with these principles.

// **Signatories:** Gilles Babinet, Jérôme Béranger, Emmanuel Didier, Muriel Londres, Dr Cécile Monteil, Pr Bernard Nordlinger, Me David Simhon, Dr Jean-François Thébaut, Pr Cédric Villani.

The ethics committee reserves the right to modify this charter according to the evolution of Epidemium.

Further details...

// Rules to encourage the open sharing of science and data

- Bensoussan A. (2016). *White Paper: An open Science in a digital republic*, March 2016, available online <http://www.cnrs.fr/dist/Livre_blanc_DIST_CNRS.html>, last accessed on November 30, 2016.
- European Commission Directorate-General for Research and Innovation (2016). *Open innovation, open science, open to the world*, available online <http://bookshop.europa.eu/en/open-innovation-open-science-open-to-the-world-pbKI0416263/>, last accessed on November 30, 2016.
- Jean B. (2011). *Free option: good practices in open licenses*, FramaBook, pp. 307.
- Pelligni F. and Canavet S. (2013). *Software law*, PUF, pp. 616.

// What should the ethical guidelines be for an open and community-based approach of the use of big data in healthcare?

- Doueihi M. (2013). *What is digital technology?* PUF, pp. 150.
- Ericsson White Paper (2011). *More than 50 Billion Connected Devices*. Ericsson, available online <http://www.akos-rs.si/files/Telekomunikacije/Digitalna_agenda/Internetni_ protokol_Ipv6/More-than-50-billion-connected-devices.pdf>, last accessed on November 30, 2016.
- Fessler J-M et Grémy F. (2001). “Ethical problems in health information systems” in *Methods Inf Med*, 40(4), pp. 359-61.
- Hamel M-P et Marguerit D. (2013). *Big data analysis: What are the practices and challenges?* France Stratégie, Analytical note n° 8, pp. 1-12, available online <http://www.strategie.gouv.fr/ publications/analyse-big-data-usages-de s>, last accessed on November 30, 2016.
- Le Coz P. (2010). "Cancer and fertility: the ethical aspects," research findings presented during the symposium organized by the Paoli-Calmettes Institute in Marseille on November 19, 2010.
- Waskul D. and Douglass M. (1996). “Considering the Electronic Participant: some polemical observations on the ethics of online research” in *The Information Society*, vol.12(2), pp. 129-139.

Conclusion

Gilles Babinet

The use of big data for advancement in cancer research is, in itself, breaking the pre-existing mold in the field of science. This discipline is still in its early stages and its development relies heavily on trial and error.

However, what makes Epidemium really stand out is its potential to create an open innovation model not only based on academically recognized experts, but on the widest multitude possible.

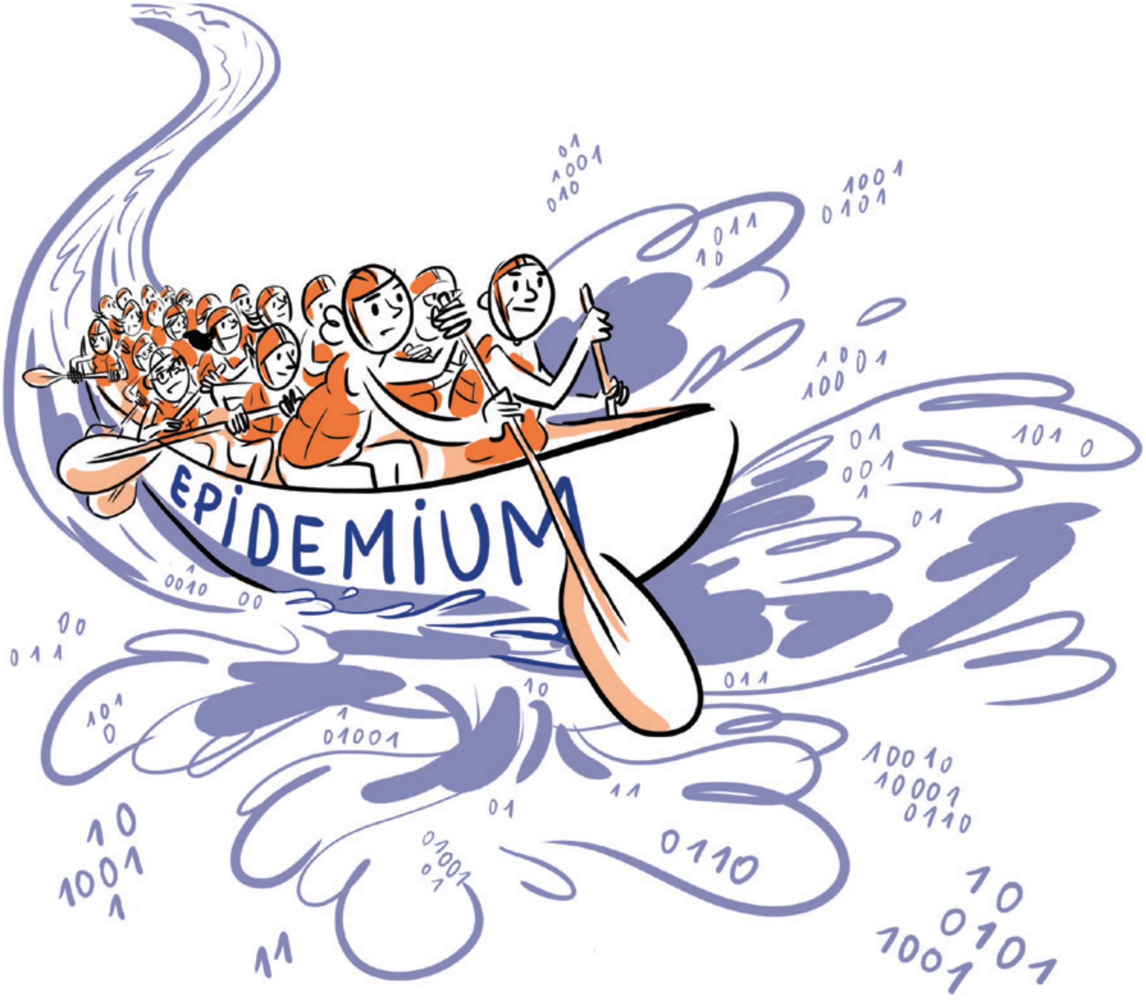
Since the field of scientific knowledge is so vast, it is becoming more and more accepted that even experts can no longer fully embrace their own disciplines. Additionally, in a world where complexity is now the norm, and multidisciplinarity a fundamental dynamic, it is necessary to change the model of innovation. It might even be possible that the future is no longer determined through R&D and centers of vertical expertise but through crowd-based knowledge.

Who could have ever imagined that Wikipedia, an online encyclopedia, would eventually become 100 times bigger than the well respected and universally known Britannica? Even more remarquable, that studies would show that it contains less errors than the Britannica, biographies excluded. The plethora of information, from Wikipedia to Github[1] via Stack Over Flow[2], demonstrates a little bit more every day its capacity to be powerful, quantitative and qualitative.

That is exactly what Roche and La Paillasse have tried to do through Epidemium: create the conditions that allow the multitude to contribute in shaping new models, that often disrupt common and academic models.

Nevertheless, to succeed, it is necessary for its initiators to have faith and believe they have the capacity to manage the many obstacles: regulatory (particularly related to private data), ethical and technological. A year and a half after its launch, we can see that these obstacles have been overcome and that Epidemium is now on its way to success.

Moreover, Epidemium is now a source of inspiration both for large companies and institutions of all types. Recently, during a trip to the European Commission, I was pleasantly surprised to hear a commissioner refer to Epidemium directly as the model of innovation for the future.



Even if we have to acknowledge that Europe is not at the forefront of the digital revolution, it could very well regain its leading position by creating and promoting the innovation model of the future. Whether we are referring to R&D, social innovation or simply public policies, there is little doubt that open innovation will one day dominate any other form of innovation. The challenge for both Europe and France is to seize these dynamics in order to help sustain them over the long-run. ■

[1] GitHub <https://github.com> is a service for web hosting and software development.
[2] Stack Over Flow <http://stackoverflow.com/> is an online community for programmers to learn and share their knowledge.

List of authors

// **BABINET Gilles, *Member of the Independent Ethics Committee*:** Entrepreneur, *Digital Champion France*.

// **BENBOUZID Djalel, *Member of the Scientific Committee*:** PhD in Machine Learning, Post-doc at the LIP6 Laboratory, Pierre and Marie Curie University.

// **BENCHOUFI Mehdi, *Epidemium Coordinating Team*:** Public Health Doctor at Hôtel-Dieu hospital, he received an agrégation in Mathematics, founded the Club JADE, a think tank dedicated to the political and social challenges of digital technology (big data, open culture, open science). He works on collaborative projects to develop open source medical technology.

// **BÉRANGER Jérôme, *Member of the Independent Ethics Committee*:** Researcher (PhD), Scientific Expert in big data, information systems, Ethics and Regulations at KEOSYS.

// **BLONDEL Léo :** Holds a PhD in computational biology at Harvard, passionnate about science. He grew up in the the world of hacking and open-software and advocates for the liberation of science. "Cyborgs also have a soul".

// **CRÉQUIT Perrine (Dr) :** Pneumologist, Network meta-analysis and cancerology.

// **DEBONNEUIL Edouard, *Member of the baseline project team*:** Actuarial Consultant in the pharmaceutical sector.

// **FRESNOYE (de) Olivier, *Epidemium Coordinating Team*:** Humanitarian development and project management specialist. Olivier holds degrees in Science and Economics. He developed several innovative collaborative and community-oriented projects and participates in many working groups on new technologies and innovation.

// **FERTÉ Charles (Dr), *Member of the Scientific Committee*:** Medical Oncologist, MD PhD at the Institut Gustave Roussy and Expert in Bio-computing.

// **HALDAT (de) Stéphanie :** Brand Director at Roche.

// **KELLER Jonathan, *Epidemium Legal Expert*:** Fascinated by law, passionate about new technologies and the legal challenges they raise, he is finishing a PhD in NICT law at the Paris Ouest La Défense University on the notion of authorship in the realm of software.

// **KOCKLER Leila (Dr), *Member of the Scientific Committee*:** Roche Representative, Medical Project Director at the Medical Affairs Division of Roche France.

// **LANDRAIN Thomas, *Member of the Scientific Committee*:** Co-founder and President of La Paillasse.

// **LÉVY-HEIDMANN Karine, *Epidemium Coordinating Team*:** Community Lead of Epidemium's program, in charge of managing, nurturing and developing the community. She is also involved in social entrepreneurship and is a board member of the MakeSense association.

// **LONDRES Muriel, *Member of the Independent Ethics Committee*:** E-patient, Assistant Coordinator at *[im] Patients, Chroniques & Associés* group of associations for chronic diseases, and militant and volunteer at the *Vivre Sans Thyroïde* association.

// **MARIANI Ermete :** Consultant in content strategy and knowledge visualization.

// **MONTEIL Cécile (Dr), *Member of the Independent Ethics Committee*:** Pediatrician and Emergency Doctor, Medical Director at Ad Scientiam and Founder of the Eppocrate community.

// **NORDLINGER Bernard (Pr), *Member of the Independent Ethics Committee*:** Digestive and Oncological Surgeon at Ambroise Paré hospital and Member of the National Academy of Medicine.

// **PETIT-NIVARD Jean-Frédéric:** Innovation Manager at Roche.

// **RAVAUD Philippe (Pr), *Member of the Scientific Committee*:** Epidemiology Professor at Paris Descartes University and at Columbia University, Research Director at INSERM, Research Director in Epidemiology and Statistics at Sorbonne Paris University, Director of the Clinical Epidemiology Center at the Hôtel-Dieu Hospital, Director of Cochrane France, Director of the EQUATOR France Center.

// **RICHARD Peter-Mikhaël, *Member of the baseline project and team*:** Currently pursuing a PhD in the field of Cosmology.

// **SANTOLINI Marc :** Postdoc Researcher at the Center for Complex Network Research of Northeastern University and Affiliated Researcher at the Harvard Medical School in Boston. His research focuses on network science applied to the field of medicine and the analysis of teamwork in the realm of scientific research.

// **SIMHON David (Me), *Member of the Independent Ethics Committee*:** Health Lawyer and President of the Île-de-France III People Protection Committee.

// **TAUVEL-MOCQUET Ozanne, *Epidemium Coordinating Team*:** Graduated in the fields of information and communication. Ozanne is interested in new technologies and in the emerging new forms of organization as well as the collaborative economy through the development of new places such as Fablabs.

// **TERLINDEN Augustin, *Member of the baseline project and team*:** Innovator in the field of medicine.

// **VILLANI Cédric (Pr), *Member of the Independent Ethics Committee*:** Mathematician, Professor at Lyon University and Director of the Institut Henri Poincaré, Fields Medal winner in 2010.

// **VITALI Isabelle:** Head of Innovation and Partnership Department at Roche France (until October 2016).

Acknowledgements

Partners

Hypercube, Dataiku, Teralab, Cancer Campus, CapDigital, Global Knowledge, Club JADE, Hacking Health, l'Institut Mines Telecom, Bress Healthcare, SchooLab, Wikimedia France

Community members

Rajaona H., Humbert G., Sourivong F., Ribas A., Martigny P., Clair B., Zhou A., Lafay E., Kamennoff N., Habert B., Frasca A., Bangoula M., Briand F., Leroy G., Betmont C., Iyeze Y., Dunoyer S., Mouhamadsultane A., Duquesnoy G., Agher D., Chatain O., Perez K., Briere T., Boulahya A., Tran T., Maouche S., Hassan M., Verrier J., Bouin O., Bouin G., Baes O., Rouphael R., Agher D., Thibault C., Couralet P., Bohl S., Clément M., Navarro A., Jourdan F., Ben A., Debonneuil E., Terlinden A., Hebert E., Reverdy V., Rouilly V., Richard G., Bouin E., Dejouy V., Strich I., Sportisse C., Yartseva A., Deponge E., Nekooguyan N., Colas E., Tafat Y., Le Clerc S., Noirel J., Coulonges C., Chekroun M., Neveux P., Kerting C., Boyer S., Thea L., Clavairoly A., Mayer C., Sacepe K., Ramdani S., Elisee R., Fagot G., De Rémacle N., Zlaoui K., Foulquié P., Cupillard E., Graveleau M., Lam J., Choffin B., Imbert A., Staron B., Dansokho F., Allorant A., Estival B., Berdugo R., Ravelomanantsoa M., Salgado M., Jandot C., Schreuder N., Foret P., Djian F., Chauvelot L., Le Tiran L., Bereder J., Muaka Di Mavinga G., Hipeaux K., Baiz S., Duge De Bernonville G., Fournier M., Mandelbrojt P., Desrousseaux C., Guinard C., Hamon A., Balmès I., Benbouzid D., Oussar J., Bossi-Malafosse J., Mackosso c., Essayegh H., Burq P., Roy P., Jean-Théodore A., Hadji Z., Letellier P., Picard F., Merinian N., Loménie n., Hadjidj A., Kegl B., Murarasu T., Plaza A., Trinh B., Saysana J., Hachan F., Lopez S., Mukakanamugire Rwagaju A., Ibnouhsein I., Wanono Y., Neuberger K., Jankowski S., Do Cao L., Funtowiez S., Bertrand N., Sutter P., Thiébau N., Jacob Y., Tazi M., Weiss D., Ben Abdallah M., Silbermann T., Gutu M., Zouaoui M., Cinquin L., Mohamed M., Rodrigues R., Dufour P., Bucchini F., Lafay E., Puig Lombardi E., Kejji M., Mezaache S., Rousselle B., Etchebarne A., Kozlowski F., Shinada J., Rafrafi A., Haubensack M., Zouaoui I., Pouchol C., Samelson L., Chera-Piloyan L., Lampe G., Richard P., Bazzoli C., Vincent Q., Zentici J., Gonidec M., Lefevre K., Gangnard J., Gaudin T., Versini J., Fadili M., Leroy G., Choukroun B., Baough L., Mukakanamugire Rwagaju A., Robellaz H., B M., Lara Ramirez A., Zouaoui I., Fabbro A., Le D., Guggiola A., Roux S., Leloutre W., Mansar Y., De Hemricourt E., Masquelier M., De Moya j., Kali S., De Chanaud N., Seznec B., Adam E., Ngo F., Adanlété R., Touche C., Denoyer L., Rouyer R., Lombardi J., Sinet L., Pradel M., Motola S., Javed S., Fatni E., Fauconnier J., Roussillon M., Epouhe C., Nadim A., Kpochan N., Taillifet E., Yilmaz J., Landre T., Afane A., Estival B., Blondeau V., Jézéquel G., Boosz P., Louis C., Ngo H., Milhaud X., Tran T., Nguyen T., Nguyen T., Planchet F., Barrau D., Clerget A., Haschka T., Koenig P., Racoceanu D., Do H., Bideault T., Do N., Sarr A., Montagnon A., Naimi G., Aljoufi M., Banquet D., Bouvier J., P J., Zenadi M., Bouguira Z., Pineau H., Vigier N., Shum King M., Callegari V., Barbat L., Messan J., Ferraina M., Feldman S., Petit J., Mutschler A., Demilly A., Meunier G., Panou D., Gea M., Sebastien J., Husson H., Lagasse G., Yang M., Philippe C., Riou T., Bouabdallah C., Giolito A., Scheer J., Scheer J., Bencherif C., Blanchard T., Louhaidia E., Beal M., Brouard C., Rollet R., Fauconnier A., Ziletti A., Le Courtois Du Manoir G., Benoumechiara N., Fischer R., Hilliquin L., El Bachiri A., Dub T., Barbé N., Gimenez U., Bah A., Deneche I., Couderc C., Corteel P., Atameklo K., Warnier M., Gautier M., Giraud P., De Rivet S., Caillé Y., Khalaj g., Caoduro C., Choffe M., Impact A., Sallah K., Salhi S., Schannes B., Litwin S., Bertrand N., Mckeon D., Zidane M., Boudraa F., Ay L., Grob V., Humphery C., Vincent R., Razafindrakoto J., Dadi C., De Vinzelles G., Brisorgueil P., Pinquié R., Berthelot M., Frédéric D., Tanguy A., Vincent B., Donzé L., Mustafa R., Truong T., Dalmasso D., Preau M., Bureau L., Pedro L., Asperti F., Tittmann L., Ould Aklouche K., Zeboulon A., Paix A.

Partners / Network leaders

Doutriaux R., Thus T., Sakka L., Massart P., d'Ormesson F., Bateson M., Kokshagina O, Sitruk Y., Gruson-Daniels C., Dumas G., Letélier S., Couraud G., Ayache C., Prévost A.-L., Roset S., Sahli A., Taieb D., Walter A.-L., Fayet J., Treguer S., Hélin V., Templier A., Civet A., Bauvin P., Suarez Valencia J. S., Rodriguez A., Kateb D., Klinger E., Jung N., Lods R., Kamennov N., Pitel G., Lefebvre S., Hispa M., Giacopelli M., Teboul D., Mangin O., Delcroix G., Rousseaux A., De Montjoye Y.-A.

Roche Open Data team

Magrez D., Violleau M., Lassauge A., Croizat B., Caoduro C., Vlamynck G., Gavini F., Herreye A., Essafi F.

“It was only a matter of time for big data analysis to tackle one of the world’s most dramatic scourge and one of the worst issue in our developed countries: cancer. Which family, in a country like France, has not been affected by this disease? It is a predicament that is especially terrible given how multi-faceted and varied it is. And this is the exact reason why so much is expected from the alliance between big data and cancerology: so many statistics are so difficult to interpret, with so many variables, that we figure we will have no other choice but to use new methods to come to a breakthrough and bring to light discoveries and new factors that doctors will be able to use and implement.”

Pr Cédric Villani
Director of the Institut Henri Poincaré
Fields Medal winner in 2010

La Paillasse

The challenge for La Paillasse, a community-oriented laboratory, is to pool together and distribute the resources necessary to carry out Epidemium's project: "*In the era of decentralized and collective intelligence, no one can have a monopoly on great ideas.*"

Epidemium

Epidemium is an open and collaborative scientific research program dedicated to understanding cancer through the use of big data and which takes the form of a data challenge called Challenge4Cancer.

Roche

"The study of open big data is a fascinating field of research. As a player in healthcare innovation, our ambition is to reinvent a new form of cancer epidemiology in view of turning it into a tool for predictive and preventive medicine."

9 791097 214005