CROWD-BASED DATA-DRIVEN HYPOTHESIS GENERATION FROM DATA AND THE ORGANISATION OF PARTICIPATIVE SCIENTIFIC PROCESS

Yohann Sitruk, Akin Kazakçi

INTERNATIONAL DESIGN CONFERENCE - DESIGN 2018
Dubrovnik - Croatia, May 21-24, 2018.

CROWD-BASED DATA-DRIVEN HYPOTHESIS GENERATION FROM DATA AND THE ORGANISATION OF PARTICIPATIVE SCIENTIFIC PROCESS

Yohann Sitruk, CGS Mines ParisTech

Akin Kazakçi, CGS Mines ParisTech

Abstract

[Abstract will be inserted automatically]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[---]
[Do not delete or modify the abstract area]

[Keywords will be inserted automatically]
[---]

1. Introduction

During the past decade, the practice of science has been facing profound changes in its organization and its processes. The paper considers two major trends affecting the contemporary science - openness and data-drivenness, and how these trends affect, in turn, the generation of hypotheses in science. The ability to generate the *right* hypothesis is also the mark of a genuinely creative scientist, since a research hypothesis very often determines the *value* of the subsequent process and results.

Hypothesis generation is the least understood and most secluded activity in a scientific process. Despite the recent efforts to open up the scientific process (Franzoni & Sauermann 2013; Wiggins & Cruston 2011), hypothesis generation has remained confined to the laboratory, and most often, to the intellect of the lone researcher. Traditionally, openness in science is manifested by a group of scientists delegating the execution of a part of their scientific activity to the crowd. These activities include data collection and annotation, analysis of existing datasets or building models. One of the main reasons originating the open science movement is to momentarily increase resources available to the scientists with little or no cost. An often-cited case, Galaxy Zoo, that consisted in involving

ordinary citizens to add some common sense, high-level semantic information to images of galaxies, enabled scientists to gather enough data for analysis in a few weeks, whereas it would have taken 83 years to collect that same amount of data for a single individual (Franzoni & Sauermann 2013). Several studies demonstrate that crowdsourcing scientific projects may foster innovation, by harnessing the brainpower and imaginations of the many and leading to a larger variety and quality of solutions (King & Lakhani 2013; Afuah & Tucci 2012; Panchal 2015). Nevertheless, finding new and interesting hypotheses have not been a topic of open science literature so far.

In the meanwhile, the increasing availability of massive datasets seems to affect how hypotheses are being generated in science today. The emergence of Big Data is considered to ease the accessibility to data and the creation of science projects across disciplines and potentially by non-specialists (Laney 2001; Kitchin 2014; Boyd and Crawford 2012). Authors such as Kitchin (2014) and Gray (2009) consider that data-driven science is a fundamental paradigm shift that will transform the way humanity will produce scientific knowledge in the following way. In the previous paradigm, data collection follows the formulation of a research hypothesis (Prensky 2009). In other terms, the hypothesis drives the data collection process. In the new paradigm, researchers are no longer constrained by existing theories for generating research hypotheses - thanks to the advantages offered by the multiplicity of publicly available datasets (Kitchin 2014). Epistemologists define this shift of science as a move from a 'knowledge-driven' science to a 'data-driven' science (Kitchin 2014). Thus, big data will be considered at the very beginning of the scientific process and will play a fundamental role in generating hypotheses, and hence, the value of the scientific output.

Currently, open science and data-driven science literatures have no overlaps. While studies on open-science emphasize the creative potential of involving broader audiences into the scientific process, as far as the authors know, there are no studies or published evidence that the crowd can deploy this creative potential into the most challenging scientific activity where creativity is needed most: the generation of new and valuable hypotheses. On the other hand, while data-driven science literature argues that data will enable the generation of fundamentally different research hypotheses, it does not clarify whether a crowd of non-specialists can accomplish this particular activity.

The central questions we consider in this paper are thus motivated by the lack of intersection between these literatures: Is it possible that a crowd generates useful research hypotheses based on large amounts of data? We shall study this question based on case study of an open scientific community, called Epidemium, working on cancer research. Epidemium, sponsored by Roche Laboratories, is built with the mission to rally a community around publicly available 21,000 datasets related to the epidemiology of cancer. The 2 years research during which Epidemium organized a series of open challenges and built a community of participants demonstrate that hypotheses generation is indeed a non-trivial process for the crowd. First, the availability of data is not enough for generating hypotheses that are consistent with the available data. Difficulties faced by many participant teams points to a need to manage the exploration and the appropriation of data by the crowd for an effective hypothesis generation. Second, when working on a large number of disconnected datasets, it is not possible (nor, necessarily, desirable) to generate all potentially useful hypotheses in one pass. Organizers need to adopt a strategy of exploring the hypotheses space through successive challenges and to capitalize on the intermediary results to become able to help the community to develop better and better hypotheses.

The plan of the paper is as follows. First, we will review the notions of hypothesis generation in the literature and a lack for organizing it through a collective activity from existing data. We present then how crowdsourcing is used for well-defined problems in big data. Then, we introduce our case study in some detail. We will present a practical case of generating hypothesis by crowdsourcing in epidemiology of cancer. An analysis of the crowdsourcing process and its limits will be presented. Finally we will provide some insights for the design problems.

2. Theoretical background

2.1. Hypothesis generation: from individual problem-solving to collective design

Hypothesis generation process in science has been widely studied during the twentieth century by philosophers, epistemologists, logicians or designers (Douven 2017). A notable strand of work on this topic comes from computational models of scientific discovery from Herbert Simon and colleagues (Kulkarni & Simon 1988; Lindsay et al. 1993; Antonsson & Cagan 2005). In these models, hypothesis generation has clearly been described as a part of a more general problem-solving activity. One example for such models is, KEKADA, which automated some tasks of the scientific process by analyzing Hans Krebs' process on urea production research. KEKADA plans sequence of experiments in order to produce observations that can be used to formulate descriptive and explanatory theories of a set of phenomena.

KEKADA conducts a double search process, in an instance space and a rule space. The possible experiments and experimental outcomes define the instance space (e.g. molecular substances), which is searched by performing experiments (run by an external user). The hypotheses and other higher-level descriptions, coupled with the confidences assigned to these, define the rule space.

Hypothesis generation process is conceptualized as a two-step process: the generation of new hypotheses based both on existing knowledge and facts from experiments, and the choice of the hypothesis according a set of rules from a decision-maker process. Each experiment generates new facts that are then evaluated through the set of rules and incorporated in the existing knowledge.

Other researchers conducted experiments to develop automated algorithm to specifically study hypothesis generation. DENDRAL for example is an algorithm to help organic chemists in identifying unknown organic molecules, by analyzing their mass spectra and using knowledge of chemistry (Lindsay et al. 1993). A subsystem incorporates specific knowledge of chemistry and mass spectrometry, accepts a mass spectrum and other experimental data from an unknown compound as input, and produces an ordered set of chemical structure descriptions hypothesized to explain the data.

The underlying problem-solving models to these problems, and to many similar research projects, are clearly based on well-defined problem domains. One might argue that the main difficulty in research, and particularly in hypothesis generation, is to be able to arrive to such a clear structuring of the space. Indeed, Simon himself acknowledges this point (Simon, Langley, & Bradshaw 1981). Hatchuel (2001), on the other hand, argues that what is usually called ill-structured problems are simply design projects, with clearly defined yet under-specified formulations due to a lack of understanding and clarity. As we shall see in our case study, both conditions are satisfied during the hypothesis generation based on a large number of data sets. This, in turn, implies that the whole process can be seen as a collective design activity, in this particular case carried out by a crowd.

Another fundamental shortfall of the above computational models is that hypothesis generation depends on the existing knowledge. As mentioned in the introduction, data-driven approach however is no longer constrained by the theories and existing knowledge in general and data instead forms the basis of the reasoning (Kitchin 2014). The process is fundamentally different since data provides much less insights on a phenomenon than knowledge (Davenport & Prusack 2005).

Finally, these models focus on the individual hypothesis generation, and thus ignores this activity can be accomplished by multiple actors. The major difference between the individual and the collective regimes is that the collective work requires organization. Literature on hypothesis generation is barren from that perspective - the question has not been considered, since historically, this was an activity that has never been opened to the crowd. The opening of this phase to the crowd is even more problematic, since, how to manage or organize is by definition a hard question: according to Oxford Living Dictionaries the crowd is defined as a large number of people gathered together in a *disorganized* or unruly way.

2.2. Crowdsourcing to outsource search process in the solution space

Many studies suggest that crowdsourcing process fosters innovation, by harnessing the brainpower and imaginations of many and leading to larger variety and quality of solutions (King & Lakhani 2013). Often implicitly crowdsourcing literature adopts Simon's original problem-solving metaphor. Afuah & Tucci (2012), for instance, explain that when a contributor conducts a *search* from its current position, he tends to focus on the alternatives around its *neighborhood*. Crowdsourcing thus can be

seen as multiplying the number of contributors to *increase the number of local searches* and the probability of finding the right knowledge and contributor for solving the problem.
It comes thus no surprise that open innovation literature's main finding is that crowdsourcing is particularly efficient in a well-defined problem situation (Afuah & Tucci 2012). In particle physics for example, methods to detect Higgs boson from data were initially developed based on pre-simulated data that were inconsistent with real observations collected from the Large Hadron Collider (LHC). Instead of internalization, physicists decided to outsource the search of a better model through the open challenge HiggsML (Bourdarios et al. 2014). The problem was designed so that no knowledge of particle physics was required to participate. As a result, more than 1700 teams have participated to the challenge, which was the biggest participation to a machine learning challenge at the time.
While crowdsourcing is an effective process for scientists to outsource the search process in the problem space, its use for hypothesis generation in a data-driven approach has not been considered. Indeed hypothesis generation from data can be considered as an ill-defined problem, where hypotheses are hard to structure or to evaluate since we do not know yet whether it leads to an original result nor it can be solved with the existing data. The outcome of a crowdsourcing process based on ill-structured problem is fuzzy since the organizer does not have a clear idea of what he is looking for.

3. Case study: a worldwide open medical project for cancer research

3.1. Method

Our research was conducted from November 2015 to November 2017 with Epidemium, an organization designed for scientific research dedicated to the understanding of epidemiology of cancer. We followed a collaborative management research (Shani et al. 2008), conducted by academics and practitioners in order to create actionable knowledge for the organization and new theoretical models in management research (David and Hatchuel 2008). Other written sources were solicited such as the wiki page of every project, the website and the white book of Epidemium. The purpose of this research was to investigate how an initiative based on Big Data should generate interesting hypotheses through a crowdsourcing process. From the perspective of Epidemium, the goal was to validate or invalidate whether a crowd may help research on epidemiology of cancer by renewing traditional research questions using the availability of disparate data sources. From our research perspective, our first aim was to gain better insight into when and how crowdsourcing maybe an interesting form of organization for generating research hypotheses. Second, we wanted to see what kind of theoretical frameworks would be needed in the crowdsourcing of a creative activity where, traditionally, it is thought that extensive knowledge and expertise is a necessary.

3.2. Big Data in the context of epidemiology for cancer

3.2.1. Epidemiology and Big Data

Epidemiology is a scientific discipline involving both medicine and statistics that studies risk factors associated with the incidence or mortality of diseases. Since the 1950s, epidemiological studies have used statistical methods that allow them to extrapolate results on samples to much larger populations. This approach have led to the emergence of numerous studies on behavioral risk factors such as exposure to alcohol, smoking or nutrition. Statistical biases in sampling, however, affect the extrapolation of local phenomena and several studies highlighted that the results are sometimes contradictory on similar risk factors (for example, a given ingredient in food can both prevent and cause cancer according to different studies; see Schoenfeld and Ioannadis, 2013). The recent emergence of massive databases on the incidence and mortality of diseases is seen in the Epidemiology community as an opportunity for epidemiological studies that could reduce the current problems and limits of existing approaches.

3.2.2. Epidemium as a structure to access knowledge communities

The notion of big data has engendered a wide appeal in health sector and several actors are seeking new opportunities. Roche Laboratories, a pharmaceutical company, wants to evaluate how big data analysis in epidemiology could be a catalyst for a new more preventive and personalized medicine. Although Roche have already participated in epidemiological surveys (e.g. ObEpi 2012 study), in-house experts face a double constraint compared to conventional statistical methods. First, the lab teams are not experts in the data science methods. Second, the scientific method to be implemented differs from conventional statistical methods. While data collection is directly related to a predetermined hypothesis, Big Data Epidemiology seeks to query a database already collected before the hypothesis is defined. Moreover, data have not been collected by epidemiologists but rather heterogeneous institutes and thus restrict an overview given a particular objective. In order to bring together both medical players and data analysis experts, Roche initiated collaboration with a new kind of research laboratory, called La Paillasse, whose objective is to be a research institution open to all citizens. La Paillasse provides Roche laboratories with a culture of open science and access to a community of scientists sensitive to openness in science. To ensure unity and a form of independence, the two entities created the Epidemium project, which is intended as a structure designed for scientific research dedicated to the understanding and epidemiology of cancer.

3.2.3. Available data

The first step of Epidemium was to collect all available open data related to the epidemiology of cancer and prepare the data to make them easily exploitable. A core data set has been compiled on mortality and cancer incidence from the databases available on the OECD and World Health Organization sites. These datasets are extended over periods of about 60 years and specify the type of cancer, country or region, age group and period of death. Data sets on the risk factors related to Sexually Transmitted Infections (STIs), particularly in the United States, as well as a set of datasets on general information (demography, environment and agriculture, climatology, work and working conditions, economic indicators, potential or actual behavioral risk factors, general health data, cancer data) represents the predictor variables to be correlated with the main dataset. In order to extend the scope of possible studies, the Epidemium team has compiled information on all publications in epidemiology in the medical scientific literature. Several datasets have been integrated, including clinical trials gathered on the WHO platform, ClinicalTrials.gov, Clinical Study Data Request, and the full database of PubMed Open Access publications plus publications on PubMed. Finally, Roche laboratories have made available a dataset of studies carried out by the laboratory. In total, Epidemium made it possible to compile a set of about 21,000 datasets accessible to all participants and free of rights and use. In a medical setting, the projects as well as the data used must comply with an ethical framework. The guarantee of anonymization the data is indeed complex to manage with open datasets. An ethics committee has been set up to delimit Epidemium's framework and ethical charter.

3.3. Crowdsourcing contest to explore the data

Epidemium is led by a core team of 6 people, mainly experts in open science and community management. With the overall objective of fostering collaboration through a common scientific purpose in mind, the organizers of Epidemium decided to launch a crowdsourcing contest, named Challenge4Cancer, based on the collected datasets. The declared objective was twofold: identify relevant hypotheses from the available databases and develop methods to test those hypotheses based. Ability to identify missing knowledge and know-how, and identify relevant stakeholders forms the basis competency needed by the Epidemium core team to build the community. In order to promote the project, Epidemium made 115 presentations in a large variety of external organizations seen as potential partners or *hubs* where talents can be recruited for the challenge. Various partnerships with recognized institutions in the medical and scientific field (APHP, Institut Curie, Cancer research Cluster CLARA) as well as a set of technical partners that provide tools for management, storage and analysis of data (Teralab, Dataiku, Hypercube, Center for Data Science) were established. Existing data science communities, such as the RAMP data challenge platform were also involved in order to facilitate access to the existing pool of talents.

Challenge4Cancer took place over 6 months between November 5, 2015 and May 6, 2016. In total, 678 contributors participated. Epidemium defined four challenges from the available datasets:

- Understanding the distribution of cancer over time and space;
- Risk factors and protective factors of cancer;
- Meta-epidemiology: understanding cancer from medical scientific literature;
- Environmental changes and cancer.

The challenges are deliberately under-specified to allow room for a variety of research hypotheses. Note that these clear yet under-specified formulations are the primary source of ill-define problems. Any participant can further specify the problem she or he is trying to solve, thus engages in a design activity where both the objective and the solution should be specified (Hatchuel 2001, 2010). Each participant or team chooses one of the challenges and defines a problem to be solved from the data relating to the challenge. Epidemium designates a scientific committee whose objective is to accompany the contributors during the production process in order to control the compliance between the projects and the proposed challenges. In a community of 678 members including 331 participants registered in the tournament (54% of data scientists, 28% of computer scientists and 18% of health professionals or medical researchers), 75 people took part in one of the 16 projects, with 63 finalists for 8 projects selected by the committees. Epidemium encouraged teams to collaborate with each other by including in the final evaluation the level of cooperation of the project during the tournament and by fostering exchanges between the participants. A weekly meet up held in the premises of La Paillasse that makes easier to integrate new contributors in the projects and to physically meet the various participants for potential collaborations. Project teams are also asked to fill a wiki page on the project run. At the end of the challenge the ethical and scientific committees evaluate the projects. Three winning projects receives a prize: € 5,000 for the first and € 2,000 for the second and third.

4. An analysis of the crowdsourcing process for Challenge4Cancer

The first round of challenge organized by Epidemium was rich of insights. In this section, we will highlight some of observations and analyze them both from the perspective of participants and their hypotheses generation processes and from the perspective of the organizers of Epidemium and their learning in terms of the management of such a process.

4.1. The participants' processes: Specifying and reformulating hypotheses when confronted with data

The participants have encountered several difficulties in conducting their projects. The main factor that was associated with the high failure rate of the projects (only 16 project was submitted in the end out of 678) was the inability of the groups to terminate in time. Those who managed to submit a proposal did not manage to reach the objectives they initially fixed and needed to adapt their final submissions by providing prototypes or simplified versions of the original target. Several aspects of the problem have caused these on-the-fly modifications, such as technical difficulties in the search of an efficient machine learning model, data quality issues and the inability to explore efficiently the large number of data sets.

Only 16 teams proposed a project that fit into the objectives of the challenge. After further screening by the evaluation committee, 8 projects have reached the final of Challenge4Cancer where the number of participants varies across projects from one to several dozen. We can identify several categories. A first category strived to build causal or predictive models between various factors to test certain hypotheses (Baseline, Predictive approach and cancer risk). A second category of teams dealt with data visualization tools, in order to facilitate hypotheses formulation (Viz4Cancer, CancerViz) or to explore the scientific literature (OncoBase, BD4Cancer, Venn). Finally, a unique project proposed to use the data to raise awareness about cancer in a more targeted and data-driven way than the usual solutions (ELSE).

For the first category, we can cite the project 'The Predictive Approaches and Cancer Risk'. This project had to limit its exploration and reformulate their initial hypothesis during the contest. During the process they realized their initial ideas were too broad and difficult to test, so they needed to

restrict the initial scope. Their research for a better-specified hypothesis was hindered though by the data quality problems they discovered progressively as they inspected different datasets more closely.
The Baseline project went through a similar cycle. Initially, project leaders wanted to predict cancer incidence, mortality and survival using risk factors from open data sources (with a global scope and a regional granularity). When they needed to program this question as a prediction problem, they realized the question was too broad to warrant a predictive modeling with the available data. The lack of coherence and substance between the various datasets made it impossible to target a general prediction problem. One of the reformulated hypotheses was " the risk modeling of the data mortality for digestive cancers (gut, colon, rectum and anus, liver, gallbladder) as a function of age and other types of cancer".
While many groups discovered during the process the incompatibility of their initial target with what the available data or analysis tools can deliver, some other groups decided to add new datasets to the database made available by Epidemium. Indeed, data quality issues that were soon realized by several groups led BD4Cancer project to team up with Baseline project in order to create a new database, EpidemiumDB. The data collection was done according to a standardized process designed by the team leaders and the collection was divided between the contributors. Some groups have included this new datasets to continue their work.
As these examples illustrate, several groups needed to abandon or reformulate their original hypotheses once they have started to study the available data in some detail. Before this confrontation, many of the ideas that were put forward were beyond what the Epidemium data could provide. This can be seen as a form of undesired out-of-box phenomena.

4.2. Learning from the challenges: Structuring future work by developing new tools and mechanisms

As with the participants, the organizers have faced several difficulties. One such difficulty was to develop a sound method for evaluating the large variety of proposals. Although the organizers have assumed that this variety would be beneficial to map out the space, they soon realized there could be no easy way to evaluate projects that were very different in their nature. To determine the winners, they have adopted an ad hoc method to compare projects in a pairwise manner. This helped them to figure out which project stood out with what aspect. This gradually became a set of custom-made criteria that could be used for selection, derived directly from the analysis of each project:

- the project clarity and relevance of the proposed approach,
- the originality of the project,
- the working methods (Collaborative work and complementarity, Appropriation of the technologies and tools made available),
- the results and conclusions (Innovative character and work done, Understanding and clarity of the results),
- the impact patients' health (Scientific medical relevance, Use and appropriation by the medical community) and
- perspectives (Long-term vision, Estimated life of the project).

Using these criteria, the three winning projects were selected to be Baseline, CancerViz and ELSE. The Baseline and BD4Cancer projects were the most unifying in the community and the most supported by recognized experts in the medical field and data analysis, which allowed them to spread widely beyond the Epidemium community.
The second major difficulty that was identified concerns the time and effort needed by the participants to develop an understanding of the proposed datasets. Epidemium provided to the contributors both a file to download with the data and a synthesis of what is inside the data. For example, the general data set was presented through categories such as demography, environment and agriculture, work, behavior, or economics factor. Each category was then divided in subcategories: for example, behavior contained data on alcohol consumption, tobacco consumption, coal use, telephone consumption, death road accident. Initially, it was assumed that this was all that is needed. As we have seen, many participants have simply ignored the details of these datasets when generating their initial target - which proved costly since reformulation was needed. These reformulations were not deemed

to be progress since, it was not about further specification of testable hypotheses, but rather, the grounding of initial vague ideas to what was accomplishable with the current time and resources.
Yet another important point affecting the process was the level of specifications of the challenge themes. As we have already pointed out, the challenge themes were under-specified. This led to the lack of ability from the contributors to propose advanced solutions during the challenge. However, this was also inevitable since the organizers themselves did not have very clear ideas about what the more specific research targets could be. In a way, they needed a first round of challenge to promote a breadth-first search strategy, which would give them a better overview of the whole design space. This also implies that successive challenges are needed, where the scope would be reduced to more specific issues and there would be, in theory, more productive.
Finally, an important issue is how to capitalize, not on how to run future challenges, but, how to use the results to contribute to the scientific knowledge on the cancer research. During the process, some projects for example found initial results to contribute to the scientific knowledge by highlighting possible correlations between risk factors and the incidence of certain cancers. *Baseline* for example identified a possible correlation between Black African populations and the incidence of prostate cancer. 'Predictive Approaches and Cancer Risk' project focused its investigations on the incidence of pancreatic cancer. Initial analyses shown that the most discriminating variable for explaining pancreatic cancer, among agro-environmental variables, energy use in the agriculture and forestry sectors as a percentage of total consumption of energy. Currently, there is no scientific publication on this topic, and the result remains a local knowledge detained by the Epidemium community.

4.3. Setting up a second challenge: deepening and centring the search

After the first contest, Epidemium set up a second crowdsourcing in June 2017. Contributors in the first challenge had difficulties to submit a finished product, facing a gap between hypothesis formulation and existing database. Organizers wanted to limit such constraints by lowering the quantity of available data set and reducing the number of challenges proposed to two: constructing a Data-Visualization of the incidence of cancers by exposing the epidemiological factors associated with their dynamics; developing a predictive tool for the progression of cancer in time and space, depending on the known or supposed factors that determine its evolution. Moreover, objectives were readjusted, and organizers asked a final scientific publication to the teams to win. Epidemium generated a great deal of enthusiasm with the first contest and many well-known French engineering schools, such as Centrale-Supélec and Polytechnique, were interested for using the challenge as a platform for student projects. Epidemium therefore set up a student-related challenge on predicting cancer mortality in developing countries. Second Challenge4Cancer has been launched in November 2017 and should be finalized by March 2018.

5. A model of the crowdsourcing process for hypothesis generation

Our analysis of the previous section demonstrate that, in a scientific hypothesis generation context, crowdsourcing should be thought as an iterative activity where the organizers need to capitalize on the results of successive challenges to better learn both the value of emerging hypotheses and the ideal methods and tools for better managing the community in the successive events. In this section, we present a process model that describes how the process can be extended to manage this iterative process. This model was implicitly used by Epidemium although the internal steps were not considered as part of the crowdsourcing process.

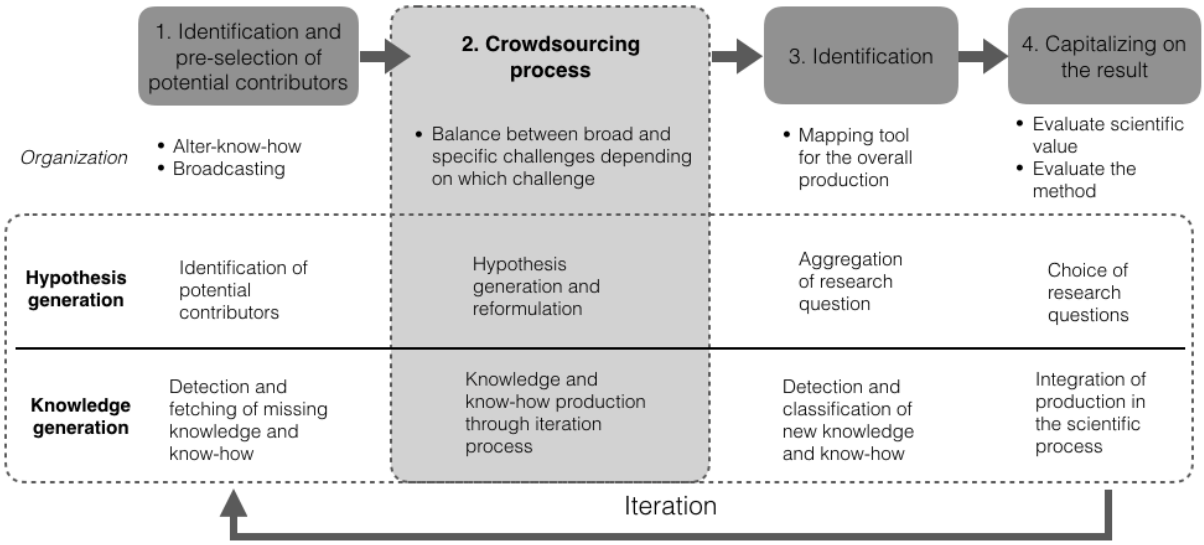

Figure 1. Model of the crowdsourcing process for hypothesis generation

5.1. Identification and pre-selection of potential contributors

As the Epidemium case demonstrate, it is important for the organizers to find the right participants that can bring the necessary expertise to the crowdsourcing process. This is an activity that should not be underestimated by the organizers, since it requires identifying missing knowledge and know-how, as well as identify where those resources can be found. The project leaders should *broadcast* the objectives of their initiatives through strategic events and leverage the intrinsic motivation of the targeted participants. We call this step the search for *alter-know-how*, the identification and fetching of missing know-how. One should be aware that every subsequent contest modify however the initial setup and modify also which kind of knowledge is needed inside the process. There is a need to systematize this action inside the iterative process while it is necessary every time a new contest is designed.

5.2. Crowdsourcing process

The Epidemium case demonstrates that attention needs to be paid to the level of specification of the challenge objectives. At this point, four different objectives compete. First, the objectives should be broad enough to allow room for the generation of a variety of hypotheses. This is particularly true if the organizer do not have a clear vision of the search space and the associated values of potential research questions. Second, the objectives might also need to be specific enough so that the cost of reformulation is not high and the participants remain a certain level of productivity. The organizers may need to reduce this time with tools to improve the appropriation process of the data by the crowd (particularly, in order to avoid the undesirable out-of-the-box effect seen in paragraph 4.1). We have seen in paragraph 4 that Epidemium committee already identified this as a critical part and integrated the generation of new tools as one of the two challenges in the second contest. As Escandon-Quintanilla (2017) suggest in ideation in engineering design processes, the way and the degree to which the participants are allowed to interact with the data has important consequences on the outcome of ideation. Third, the degree of specifications will depend on the current level of advancement in the previous episode of challenge. Epidemium has started with very broad topic which allowed to explore the space very broadly but they soon realized that the next challenge should be more specific and targeted, and possibly a third challenge can be run on a very specific subject that was generated during the second rounds and determined to be highly valuable from a scientific point of view. Fourth, the scientific value of the proposals should be monitored between successive cycles. As the space is likely to be exponentially large on the number of datasets available, care must be taken for steering the exploration process in-between challenges.

5.3. Identification of the crowdsourcing outcome

During the first contest, challenges were too broad and gave not enough specific hypotheses generated by the crowd to be considered as a potential value for the scientific literature. However, some insights emerged providing tacit elements for the next challenge, such as potential correlations between variables or exploration of data analysis. These informations should be categorized to prove its worth and how it can be reintegrated in the next crowdsourcing. A mapping tool like CK theory for categorization might be useful as we first explore what can be done with the data (Hatchuel et al. 2011). The organizers should pay specific attention to what are the specific questions made by the crowd, and their level of specification. C-K provides interesting informations to evaluate the level of specification of a hypothesis and give some insights on the missing knowledge needed for a specific hypothesis. A first experimentation was made during the first challenge (Kokshagina & Sitruk 2017, see figure 2) and investigations should explore how this tool can be used to reintegrate the production of the previous contests.

Analyzing the overall production of the crowdsourcing process through tools for categorization should provide new metrics for evaluating the final contribution of every project.

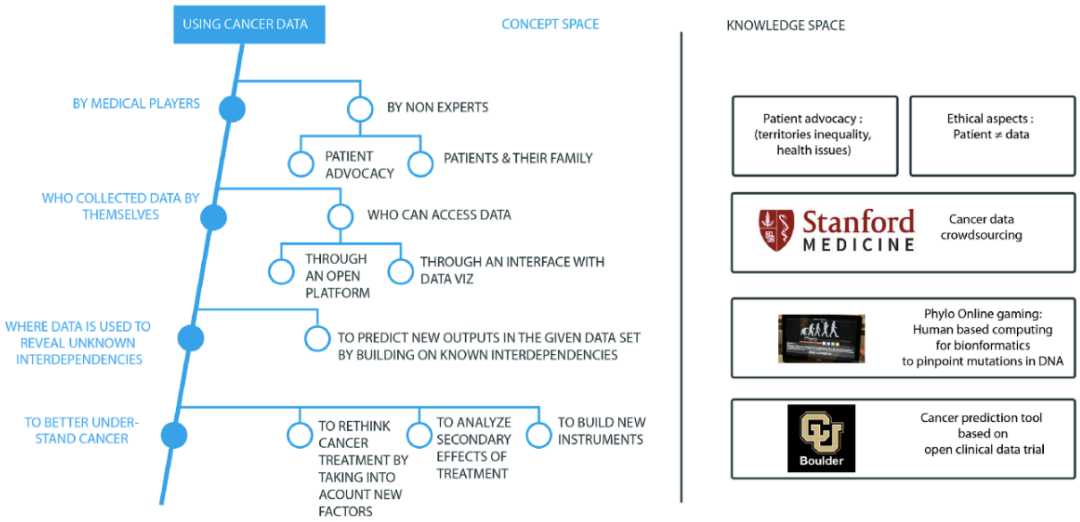

Figure 2. Extract from CK tool (Kokshagina & Sitruk 2017)

5.4. Capitalizing on the results

The organizing committee should be able to mobilize appropriate experts to evaluate the scientific value of those proposals. This will allow to set priorities to determine which challenge to be organized next. Result of step 3 should explicitly guide to the design of the new challenges. Organizers need also to take stock of the methods used and specify which element has been effective and others that need to be improved.

6. Discussion

This paper analyses the crowdsourcing method and identifies a lack of literature on the study of generating hypothesis using the crowd on data-driven projects. We conducted an in-depth analysis at Epidemium that highlights two organizational learning activities that need to be included in the crowdsourcing process: learning from the contributors and learning from the organizers. We propose a process that includes the two forms of learning identified. Further researches should be done to explore the applicability and performance of the proposed process in future Epidemium contests and other scientific contexts.

Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kégl, B., and Rousseau, D. (2014). The higgs boson machine learning challenge. In NIPS 2014 Workshop on High-energy Physics and Machine Learning, volume 42, page 37.

Afuah, A., & Tucci, C. L. (2012). Crowdsourcing as a solution to distant search. Academy of Management Review, 37(3), 355-375.

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. Wired, 23 June 2008. Available at: https://www.wired.com/2008/06/pb-theory/ (accessed 05 December 2017).

Antonsson, E. K., & Cagan, J. (Eds.). (2005). *Formal engineering design synthesis*. Cambridge University Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

Callon, M. (2009). Acting in an uncertain world. MIT press.

David, A., Hatchuel, A., "From actionable knowledge to universal theory in management research", in: Shani, A.B. (Ed), Handbook of Collaborative Management Research, Sage Publications, Thousand Oaks, CA, 2008.

Douven, Igor, "Abduction", *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/abduction/>.

Escandon-Quintanilla, M. L. (2017). Effects of data exploration and use of data mining tools to extract knowledge from databases (KDD) in early stages of the Engineering design process (EDP) (Doctoral dissertation, École de technologie supérieure).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. Research Policy, 43(1), 1-20.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., & Schader, M. (2011, August). Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. In AMCIS.

Hatchuel, A. (2001). Towards Design Theory and expandable rationality: The unfinished program of Herbert Simon. *Journal of management and governance*, *5*(3), 260-273.

Hatchuel, A., P. Le Masson, Y. Reich and B. Weil (2011). A systematic approach of design theories using generativeness and robustness. Proceedings of the 18th International Conference on Engineering Design (ICED11), Vol. 2: 87–97.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 2053951714528481.

King, A., & Lakhani, K. R. (2013). Using open innovation to identify the best ideas. MIT Sloan management review, 55(1), 41.

Kokshagina O., Sitruk Y. Open Science: how to identify exploration axes in a transdisciplinary context? Medium, 17 October 2017. Available at : https://medium.com/epidemium/using-big-data-to-understand-cancer-epidemium-discover-how-c-k-method-can-be-used-to-identify-898120dbfac4 (accessed 5 December 2017)

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. Cognitive science, 12(2), 139-175.

Laney, D. (2001). 3D Data management: Controlling data volume, velocity and variety. Meta Group. http://blogs. gartner.com/doug-laney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Var iety.pdf. Accessed 16 Jan 2013.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1993). DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence*, 61(2), 209-261.

Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engineering applications of artificial intelligence*, *16*(4), 277-291.

Obépi-Roche, R. (2012). Enquête épidémiologique nationale sur le surpoids et l'obésité. *Paris: Inserm/TNS Healthcare/Roche*.

Panchal, J. H. (2015). Using Crowds in Engineering Design–Towards a Holistic Framework. In *2015 International Conference on Engineering Design, Design Society, Milan, Italy, July* (pp. 27-30).

Prensky, M. (2009). H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate: journal of online education*, *5*(3), 1.

Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. Am J Clin Nutr 2013;97:127-34.

Shani, A. B., Mohrman, S. A., Pasmore, W. A., Stymme, B., Adler, N., "Handbook of Collaborative Management Research", Sage Publications, Thousand Oaks, CA, 2008.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. *Synthese*, *47*(1), 1-27.

Shmueli, G. (2010). To explain or to predict?. *Statistical science*, *25*(3), 289-310.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Wiggins, A., & Crowston, K. (2011, January). From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii international conference on* (pp. 1-10). IEEE.