

# UX Design for a Data Passionate Community

**Alessandro Pollini**

BSD design

Rome, Italy

alessandro.pollini@bsdesign.eu

**Angela Di Massa**

BSD design

Rome, Italy

angela.dimassa@bsdesign.eu

**Carlo Riccucci**

Engineering

Rome, Italy

carlo.riccucci@eng.it

**Andrea Manieri**

Engineering

Rome, Italy

manieri@eng.it

**Massimo Mazzarini**

Engineering

Rome, Italy

massimo.mazzarini@eng.it

## ABSTRACT

SUBMITTED—15 February 2016.

Data Science competences are of highly increasing relevance either for employer organisations, demanding for Data Science skilled work force, and for the work force supply entities.

However, what are the competences for being a data scientist is still matter of research. The Data Science practitioners come from several paths, they have a broad set of competences, share a respectful attitude towards data lifecycle, and wish to establish and spread a “data culture” in the organizations they are involved in.

As such, it is a very heterogeneous community with currently several forms of aggregation, all focusing on a specific topic. It is considered a challenge to design on-line services that help the various Data Science stakeholders to form an engaged community.

This paper presents how the user experience processes have been used in the context of the EDISON project<sup>1</sup> to conceive and design the online services for this data passionate community.

### Author Keywords

Data Science, design research, service design, user-centred design, mobile devices.

### ACM Classification Keywords

H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces; User-centred design.

## INTRODUCTION

It is widely recognized that terms like “data scientist,” “analytics,” and “big data” are the result of what one might call a “buzzword meat grinder” (Harris et al. 2013) Today data scientists used to come from established fields such as statistics, engineering, physics, machine learning, databases, business intelligence, neuroscience, and more,

- Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

with special competences (skills and knowledge) that let them being able to extract value from the digital data.

All of the stakeholders in the data science domain, ranging from research infrastructure managers to data science researchers and practitioners, until industry training managers, have their own expectations about what a data scientist is, what is able to do (and not do), which projects might join. Furthermore not yet there are well-defined educational and career paths (Manieri et al. 2015), including any certification, that enable the students and practitioners to be recognised as such.

These facts, together with the increasing number of formal or informal, temporary or permanent, distance or in-presence data science communities, prove that we are very far beyond the new trendy buzzword phenomenon.

The current Data Science communities engages in creating, maintaining and curating data products like for example search terms, voice samples, or product reviews, with the users in a feedback loop in which they contribute to the products they use. Data products are then pervasive to the ordinary lives of all of us increasingly interacting with services for mobile devices and continuous interactions with real-time data exchange.

The question facing every company today, every start-up, institution, or project site that wants to attract a community, is how to use data effectively (Loukidis 2010), meaning all the data that might be available and relevant in each implementation scenario.

What differentiates Data Science from other data related disciplines, like statistics, is that Data Science implies a holistic approach. We’re increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.

This work, falling into the application of design research (Faste 2012) in the Data Science domain, investigates the main stakeholders commitments and expectations through user centred design tools and methods to propose service design for mobile devices fostering the rise of an online Data Science-centered community.

<sup>1</sup> [www.edison-project.eu](http://www.edison-project.eu)

In particular, we describe the process leading to the development of the community portal that should attract the main Data Science stakeholders, both on job market demand and education supply sides.

In the following paragraphs are introduced the inspirational benchmark, the sampling of the Data Science Community by the initial set of interviews, and the definition of the scenarios that describe an ideal journeys of each stakeholder interacting with the community portal and services of highest interest. The paper concludes with the design concepts identified in order to attract and engage with the target users.

## **INSPIRATIONAL BENCHMARK**

The inspirational benchmark starts from an analysis of the available web resources and other portal samples, hence providing a state of the art of Data Science communities on-line services.

### *Method*

In the Data Science domain there are several kind of on-line solutions: online information portals and blogs, e-learning platforms, Data and Computing infrastructures and Platforms, marketplaces and sharable services for data seekers and data providers.

The first step has been to define the boundaries of the benchmarking through a clear definition of the benchmark criteria adopted to select the sample cases to be considered for evaluation. In the EDISON context the overall aim is to promote a one-stop on-line service to serve the wider community of Data Scientists and correlated stakeholders, starting from the researchers, students and eScience practitioners, till librarians, eInfrastructures operators, public servants and decision-makers at national or European level. For this reason the quality and completeness of the content exposed was also a relevant selection indicator. Five categories of features (or functionalities offered) have been selected:

- Competence Assessment,
- Knowledge resource marketplace,
- E-learning environment,
- hands-on Virtual Data Labs
- Collaboration services.

The criteria adopted to select the subjects of the analysis for each categories were:

- On-line services with an active learning community that participate, support and contribute to the growth of the human network;
- The degree of growth of the human network;
- Other on-line services generically related the Data Science domain.

### *Key Findings*

The various services analysed can be grouped in three main technical categories:

- Community portals (Data Science Europe, Federated Cloud, D4Science-II, Kaggle, Cross Validated, DataKind);
- Web repositories collecting external resources (On-course, Datascience.community, Data Landscape, EGI Marketplace, Data Science Central),
- E-learning platforms (Moodle, Docebo, Chamilo, DataCamp, DataQuest, DataMonkey, Coursera).

Independently from the above mentioned categories the main features offered are:

- To access online and (redirecting to) on-site training courses, (i.e. Moodle, Docebo, Chamilo, Data Science Europe, DataCamp, DataQuest, DataMonkey);
- To retrieve educational materials for free or against payment, (i.e. EGI Marketplace, Data Science Central, On-course, Datascience.community, Data Landscape, DataCamp, DataQuest, DataMonkey);
- To access EU and worldwide Data Science domain stakeholders map (i.e. On-course, Datascience.community, Data Landscape);
- To contribute to conceptual and technical discussion forum and blog (i.e. Data Science Central, Kaggle, Cross Validated);
- To get certifications of the learning achievements (i.e. Docebo, Coursera and Chamilo).

In terms of target users the selected cases address a variety of Data Science stakeholders, ranging from providers in public and private sectors professionals to educational organizations. In particular the following have been identified:

- Public and private sectors professionals and students (i.e. On-course, Datascience.community, Data Landscape, Moodle, Docebo, Chamilo, Federated Cloud, Kaggle, DataKind, Coursera).
- Educational organizations (i.e. On-course, Datascience.community)
- Data science experts and researchers (i.e. Data Science Central, On-course, Datascience.community, D4Science, Kaggle, Cross Validated, Federated Cloud, DataKind)
- data practitioners and scientists (EGI Marketplace, Data Science Central, DataKind, D4Science).

In most cases users are identified through registration and authentication (as easy as access exploiting Facebook™ or GMAIL™ credential, till the more complex X509 certificates).

Some propose international challenges and competitions (i.e. Kaggle and Cross Validated) as a mean for attracting and engaging with most bright minds. Sometimes the gamification approach and the use of a qualification badge are used to incentivise users familiarize with Data Science topics and gain visibility (i.e. DataCamp, DataQuest, DataMonkey, Kaggle and Cross Validated).

Interestingly, despite the ability of manipulation of data is at the core of Data Science job only few (i.e. DataCamp, DataQuest, Federated Cloud and DataMonkey) allow hands-on experiments and exercises; in some specific cases with the support of a tutor. Some progressive bar for activities (i.e. Moodle, Docebo, Chamilo, DataCamp, DataQuest, DataMonkey, Kaggle).

### SAMPLING THE DATA SCIENCE COMMUNITY

Sampling methods in user research helps in understanding the overall (prospective) community behaviour by selecting an initial set of stakeholders of the Data Science community. In literature user research is described (Holtzblatt, Beyer, 2014):

- as allowing the researchers to grasp the best practices to engagement and fruition of services and content within the Data Science domain;
- as enabling elicitation of specific and tacit knowledge from the various actors involved,
- as inspiring prompt and creative thinking by participatory design and focus groups;
- as providing in-depth analysis and assessment of interaction and service development.

The Data Science community, as exemplified by EDISON, is split in producers of Data Scientists (the supply side) and employers of Data Scientists (the demand side).

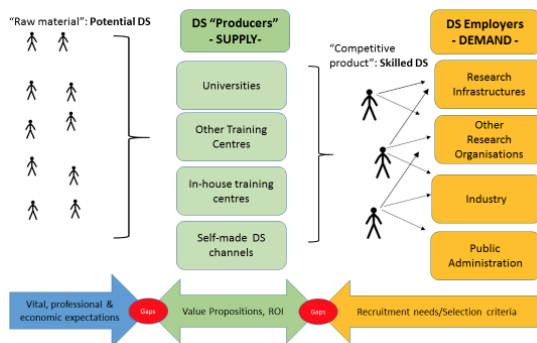


Figure 1- Supply and Demand Side in the EDISON model

The demand side has been qualified with the following entities:

- Digital industries,
- PA, as Government-funded units requiring data scientists or data science competences,
- Research infrastructures and other research organisations,

The following roles have been identified as useful to be interviewed:

- **Top managers (i.e. CIO or Managing directors)**, as top executives in private companies in charge of business development and team planning,
- **Human Resources manager**, accounted for recruitment and enrolment,
- **Training manager**, responsible for internal training of groups and individuals.
- **Data Scientist**, being any of researcher, practitioner or graduate:

The supply side, is qualified by:

- The Universities,
- The In-House training schools,
- The National accreditation agencies,
- The self-made Data Scientist

The roles which have been selected are:

- The Didactic manager, as decision maker responsible for planning and ideating the training program;
- The Learning manager, as university course leader involved in the teaching and training delivery;
- The Accreditation manager, appointed for evaluating and approving the formal training program submitted by the Universities;
- A Researcher skilled in digital data management that is planning or re-orienting his/her career toward a fully-fledged Data Science Profession.

For each of them have been done interviews and analysed the results in the light of the benchmarking. The main findings have been elaborated into a list of themes which describes the design space within which the Community Portal concept is proposed:

### Data Scientist profile

- ‘Passionate about data’: the sample of target users involved in user research reported us that they all share a passion about data and wish to foster a data culture in their own job. Those acting in Data Science, are professionals oriented to enhance their knowledge about the whole science around data, considering all the phases of Data lifecycle management. The effort they have been paying to access and work on Data Science is mainly driven by passion about the various disciplines of this domain. The majority of the users’ sample comes indeed from a PhD in Data Science-related disciplines, being Physics, Mathematics, Statistics or other.
- ‘Data Scientist as communicator’: Data Scientist needs to master the methodologies for Data management lifecycle as well as to own project management, communication and accounting skills. Data Scientist

today has indeed to understand different domain-specific languages in order to gather requirements from different stakeholders. Another very important desired skill for the Data Scientist is the ability to explain, discuss and disseminate the contents, the methods and the results of her work to a non-technical audience (e.g. clients, marketing managers, business analysts).

#### *Learning Community*

- ‘Participation and joint build-up of knowledge’: the sample reported the wish to take an active part in producing DS-related knowledge: they are knowledge-enthusiasts that needs to feel that contents are enriched and built up thanks to the participation of all the professionals acting in this domain, especially where the more informal contents are concerned (experiences, projects, appraisals) which are conceived to be added to the formal contents later.
- ‘Networking’: participation and learning are enhanced in the view of the user sample by the creation of a community which goes beyond individual training events and permits further in-depth study and focalisation of the contents, especially in a project-based learning approach.

#### *Industry Focus*

- ‘Industry focus’: Data Science is reportedly considered as an industry-led field where the application domains play a fundamental role in declining Data Science methods and practices to domain-specific projects. However, several initiatives are guided by research aims, thus opening the opportunity to consider also research driven data science<sup>2</sup> as an unicum.

### **SCENARIO-BASED SERVICE DESIGN**

During the execution of the user research activities, qualitative and quantitative data have been acquired, interpreted and elaborated in relation to the project objectives with the goal of defining initial users profiles in the shape of Personas. The Personas are hypothetical archetypes of real-life users, fictitious characters created to represent different types of target user that are used as input for the design process.

The Personas being developed are used as main actors of the narratives describing the role and functioning of the Community Portal and Services. The portal will be in fact described starting from use cases descriptions whose characters will be drawn on the personas. These use cases are represented in a graphical and textual way in the form of scenarios. Stakeholders with different backgrounds and different interests can relate to stories. The scenarios provide a common ground and engaged all the readers.

Scenario-based design is a methodology that allows the design team to develop scenarios able to describe present

activities, suggest future ones, and for the iterative assessment of users. In particular the design team will develop future and envisioning scenarios that will provide cases for future realities concerning the fruition of the portal services, the development of high quality educational contents and the promotion of learning communities, either temporary and permanent. These scenarios will be used to explore in depth the appropriateness and efficacy of the Community Portal and services concept.

The future and envisioning scenarios are the most useful and viable scenarios conceived to represent the characteristics and role of the future system. The concept thus generated will be described and merged into the reference activity and the scenario will explain how the activity itself may be modified to include the portal. The envisioned scenario thus produced will be enhanced and consolidated as the project evolves.

#### **Example of Scenario**

In this paragraph one example of the service design scenarios produced in the project is described. Such scenario regards personas like the Community Moderator, under the Managing Users, and the Researcher, as Independent Users.

The selected scenario explicitly refers to the following design issues:

- Key value 1: Data science literature as both Encyclopaedia and expert articles
- Key value 2: User-generated content and Community moderator trigger
- Key value 3: Community member reputation and Peer assessment

#### *Scenario “Knowledge-driven community and user-generated content”*

##### *Planning training events*

Marco works as Data Scientist in pectore at Leiden University Medical Center. He is very interested in keeping up to date with the latest in Data Science. Passionate about data, he regularly joins online and onsite training courses to reach a higher specialization.

##### *Evolution of the Community Portal literature*

After having completed a study, Marco finalized his research essay ready for publication. He is already logged with his LinkedIn™ account to Community Portal and posted it in the “Expert Articles” section of the Community Portal Literature. After some days, the content was approved and made accessible to all the members.

##### *The moderator triggers for user-generated content*

Sometime later, Marco was invited by the Community Portal Moderator to publish the updates about his research on regular basis.

---

<sup>2</sup> Cfr. <http://www.codata.org/working-groups/research-data-science-summer-schools>

### User ranking and Community label

Marco is gaining higher and higher ratings for his contributions to the community: his reputation as active Data Science community member is growing up. The quality of his own research advancements was recognized by the second grade Community Label.

### COMMUNITY PORTAL AND SERVICES

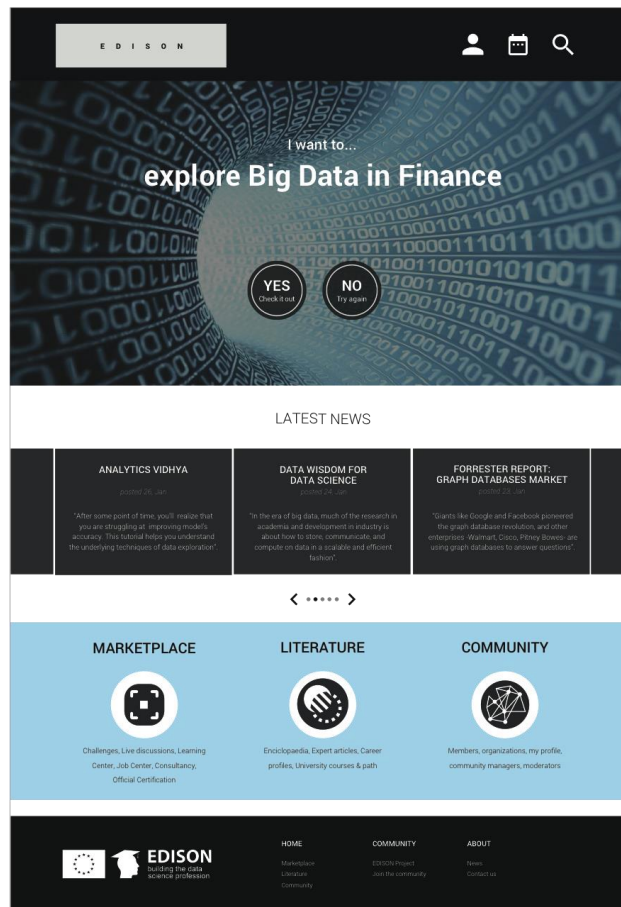
The user requirements as defined in the user research will be used, together with the results of the benchmarking, to design both a web portal and a mobile application specifying the types of interaction with the chosen solutions (Norros et al. 2015). Creative methods and techniques have been used (for example, brainstorming, think-thank, attribute listing) to generate and assess various concepts regarding both the interaction modalities and the organisation of the services and the architecture of the information.

The process to develop the design of the Community Portal generated from the both the development of personas and scenarios as well as from the benchmarking and state-of-the-art on-line portals and services in Data Science.

The Community Portal is then defined as a “one stop / one step shop concept”, since all services are accessible in one portal and all services are available at once. The design concepts are synthesised in the table below.

KNOWLEDGE DRIVEN COMMUNITY	Marketplace and Literature tightly connected for enabling the growth of a data science passionate community
QUICK PROFILING	One-step content customization on the basis of competences' gap analysis
COMMUNITY LABEL	User ranking based on community member reputation and peer assessment
SERVICE PORTAL	Service generation and deployment can be flexibly managed relying upon the EOEE service architecture
THIRD-PARTY INTEGRATION	Portal evolution by third-party contents and services integration

**Table 1. Key Features for a one stop/one step on-line service**



**Figure 2. Community Portal Homepage**

## CONCLUSIONS

The design process described in the paper leverages on user experience, service and interaction design (Hassenzahl, 2010) to provide the conceptual foundation of a successful portal and related web and mobile services for a Data Science-centred on-line community. This is an iterative and incremental process that focused on mapping the needs of initial set of stakeholders, on eliciting knowledge and insights from their current experience in Data Science and on defining the Community Portal and related services future interaction scenarios (Blomberg, Darrah, 2015). The approach taken is of a co-evolutional nature because it foresees the application of a range of analytical and modeling methods following a multi-design pattern into which three different perspectives are woven: user research, service and interaction design and technological potentials, for the definition of a solution at once visionary, innovative and capable of stimulating user engagement and participation and co-built knowledge (Ju 2015).

This activity has been partially supported by the EDISON project with funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675419.

## REFERENCES

1. Harlan D. H., Murphy, S. P., and Vaisman, 2013. Analyzing the Analyzers. Printed in the United States of America.
2. Loukides, M., 2010 What is data science? The future belongs to the companies and people that turn data into products. From <https://www.oreilly.com/ideas/what-is-data-science>.
3. Blomberg, J., Darrah, C. 2015. An Anthropology of Services: Toward a Practice Approach to Designing Services. Synthesis Lectures on Human-Centered Informatics, Morgan&Claypool Publishers
4. Ju, W. 2015. The Design of Implicit Interactions. Synthesis Lectures on Human-Centered Informatics, Morgan&Claypool Publishers
5. Norros, L., Savioja, P., Koskinen, H. 2015. Core-Task Design: A Practice-Theory Approach to Human Factors. Synthesis Lectures on Human-Centered Informatics, Morgan&Claypool Publishers
6. Holtzblatt, K., Beyer, H. 2014. Contextual Design: Evolved. Synthesis Lectures on Human-Centered Informatics, Morgan&Claypool Publishers
7. Hassenzahl, M. 2010. Experience Design: Technology for All the Right Reasons. Synthesis Lectures on Human-Centered Informatics, Morgan&Claypool Publishers
8. Manieri, A. Demchenko, Y. et al. 2015. Data Science Professional uncovered, CloudCom15 Proceedings
9. Faste, T, Faste, H 2012 Demystifying "Design Research": Design is Not Research, Research is Design', IDSA Education Symposium.
10. Data Science Europe, <http://www.datascienceeurope.com/>
11. Federated Cloud, <http://www.egi.eu/infrastructure/cloud/>
12. D4Science-II, <http://www.d4science.eu/>
13. Kaggle, <https://www.kaggle.com/>
14. Cross Validated, <http://stats.stackexchange.com/>
15. DataKind, <http://www.datakind.org/>
16. On-course, <https://www.on-course.eu/>
17. Datascience.community, <http://datascience.community/>
18. Data Landscape, <http://datalandscape.eu/>
19. EGI Marketplace, [http://www.egi.eu/services/training\\_marketplace/](http://www.egi.eu/services/training_marketplace/)
20. Data Science Central, <http://www.datasciencecentral.com/>
21. Moodle, <https://moodle.org/>
22. Docebo, <https://www.docebo.com/>
23. Chamilo, <https://chamilo.org/>
24. DataCamp, <https://www.datacamp.com/>
25. DataQuest, <https://www.dataquest.io/>
26. DataMonkey, <http://datamonkey.pro/>
27. Coursera <http://coursera.com/>