DESIGN CULTURE(S) | CUMULUS ROMA 2021
JUNE 08.09.10.11, SAPIENZA UNIVERSITY OF ROME

Towards a visual-based survey on explainable machine learning.

Beatrice Gobbo[a]*

[a]Politecnico di Milano
*beatrice.gobbo@polimi.it

Abstract | The increasing use of Machine Learning in people's everyday life raised the need for solutions aimed to reveal the work done by those models when transforming an input into an output. In the field of Computer Science, techniques of Explainable Machine Learning have been developed for unveiling algorithms' inner workings at different degrees of sophistication. The current status of the research on Machine Learning Explainability is still empowering the creators of those models but is not informing the people affected by them. Being information visualisation considered a good mean to show these processes, it is legitimate that tools able to help designers to browse visual models used in the past are designed. The paper proposes a visual-based methodology for displaying and analysing images in-groups as a support for designers in the observation, investigation and selection of visual models and solutions to be adopted in the area of Explainable Machine Learning.

KEYWORDS | INFORMATION DESIGN, VISUAL MODELS, EXPLAINABLE MACHINE LEARNING, DATA VISUALISATION, VISUAL DEVICE

1. Introduction

Explaining machine learning is the ability to present in understandable terms machine learning models to a human (Doshi-Velez & Kim, 2017). In the field of Computer Science, techniques of *Explainable Machine Learning* have been developed for unveiling and analysing algorithms' inner workings at different degrees of sophistication. However, the current status of the research on *Machine Learning Explainability* is still empowering the creators of Machine Learning models but is not informing the people affected by these models (Correll, 2019). After a few years that Explainable Machine Learning techniques have emerged in the field of Computer Science (Lipton, 2016), the need to explain the internal functioning of Machine Learning algorithms has spread and amplified, starting to involve other fields, such as information design and ethics (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). In the Computer Science field, different surveys (Garcia, Telea, Castro da Silva, Tørresen, & Dihl Comba, 2018; Hohman, Kahng, Pienta, & Chau, 2018) have been carried out to investigate and evaluate the state of the art and future research directions, especially for *deep learning* and *neural networks techniques* (Garcia et al., 2018; Hohman et al., 2018; Yu & Shi, 2018). Indeed, data visualization and visual analytics are considered good means (El-assady, Jentner, Kehlbeck, & Schlegel, 2019; Offert, 2017) for both communicating and explaining internal states of machine learning models. However, in the examined literature there does not seem to be a collection showing all the visual models used to represent processes regardless of the type of machine learning model.

The paper presents the initial phase of a research process that aims to investigate how information visualisation and design can be exploited to communicate to a lay audience the internal processes of machine learning algorithms that impact society. The presented stage of the research proposes a visual-based methodology for displaying and analysing groups of data visualisations as a support for designers in the observation, investigation and selection of visual models when dealing with machine learning explainability.

For distantly investigating the aesthetic patterns of visualisations in the field of explainable Machine Learning, in the paper, a method coming from the study of digital platforms has been used, leveraging tools from Digital Methods (Rogers, 2017). Digital Methods are a collection of tools and techniques that have been originally designed to analyze online text content with a sociological approach by experimenting on different online platforms — Twitter, Facebook, and Google –. Indeed, in the last years, the increasing dissemination of visual content on the web has driven the desire to develop techniques related to the study of images *en groupe* by using Digital Methods (Niederer & Colombo, 2019; Ricci, Colombo, Meunier, & Brilli, 2017). Most of the efforts have been made to study images from social networks and search engines to create *networked images systems* able to display multiple images connected by common attributes.

Since scientific articles, usually associated with images displaying visualisations and screenshots of interactive interfaces, are available on the web, this paper proposes an experimental approach for studying visual models in groups, for distantly investigating the aesthetic patterns of visualisations in the field of explainable machine learning by using techniques coming from the study of digital platforms (Pearce et al., 2018).

As machine learning spreads across domains, the aim is to draft a method that can help designers understand how visual patterns, colours, and shapes and other visual variables (Bertin, 1967) are used trying to visually foment and answer the question: *How to visualise explainable machine learning process.*

The intention of the paper is not to propose an alternative system for performing surveys, but to promote an integrative device that supports the exploration of case study collections and help information designers in browsing visual models during the design process.

In Section 2 related works will be presented, considering both existing surveys presenting explainable machine learning examples supported by visual analytics tools and image-based research; Section 3 presents a list summarising the design requirements; Section 4 describes the process that led from the construction of the dataset to the development of experimental visual devices; Section 5 briefly summarises results coming from the performed visual network analysis and in Section 6 strengths and weaknesses of the presented methodology are listed.

Being this ongoing research the result of an interdisciplinary approach to the study of aesthetics and visual patterns in the field of Interpretable Machine Learning, related works will be analysed by following both:

- *surveys* presented as tools for clarifying and systematising the state of the art on techniques and visual analytics for Interpretable Machine Learning. (Section 2.1)
- *research* carried out for analysing visual content (images) coming from the web presented in the form of networked images. (Section 2.2)

In the last two years, many surveys on Explainable Machine Learning techniques have been presented and discussed. Among these, some are specific to a group of machine learning models, while others differ by applying a more general approach. Moreover, examples of visual-based surveys coming from neighbouring areas will be taken into consideration.

2.1 Surveys on Explainable Machine Learning

While there is a substantial literature of surveys for visualisation and visual analytics applied in predictive analytics (Lu, Wang, Landis, & Maciejewski, 2018) and deep learning (Hohman et al., 2018; Yu & Shi, 2018) that takes into account: the reason why visualisation is used as a tool, the type of user to which it is addressed and in which context is visualisation applied, however, it seems that there is a gap in presenting a comprehensive and visual survey on visual models adopted in visual analytics for Explainable Machine Learning.

Often, evaluation grids and evaluation matrices (Garcia et al., 2018; Hohman et al., 2018; Lu, Garcia, Hansen, Gleicher, & Maciejewski, 2017)are proposed as tools for browsing case studies and evaluation parameters can vary according to researchers' own interests.

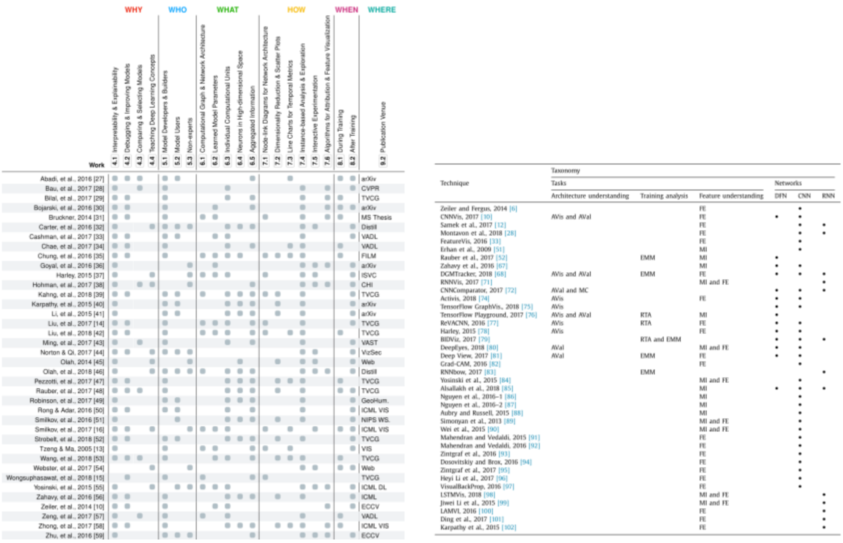

Figure 1. Two examples of evaluation matrices used for displaying case studies of visual analytics in deep learning. From the left: (Garcia et al., 2018; Hohman et al., 2018)*.*

One of the most effective functional feature of evaluation matrices is that they provide a detailed and systematic vision ensuring a good exploration of contents, however, when it is necessary to list the visual models used to display explanations, they are presented through labels referring to visual model names (for instance "scatterplot", "line chart" or "networks") (left - Figure 1) together with screenshots of relevant use cases.

Surveys related to other fields propose diverse and interactive tools for browsing projects, Liu et al. (Liu et al., 2018) propose a comprehensive, interactive task-driven survey on text visualisation taking into account tasks, visual models and data mining techniques and how their use has changed over time. Unlike the latter, Kucher (Kucher, 2014)presents his ongoing, collaborative survey on text visualisation as an interactive browser that allows users to study and inspect research and projects by using a wide spectrum of filters that enhance the analytical process of the users.

Similarly, Isenberg et al. (Isenberg et al., 2017) are working on the *Vispubdata* project which provides free access to metadata and statistics on IEEE Visualization publications (IEEE VIS) from 1990-2018. From this repository, researchers can work on data and design tools for exploring them.

Nonetheless, as far as the literature related to explainable machine learning is concerned, the production of visual-driven survey methods seems to be still lacking.

The experiment presented in the paper provides a visual-based approach for collecting and displaying used visual models, proposing a method of representation, display and reading that departs from the evaluation matrix and the interactive browser proposed by Kutcher.

2.2 Visualising and analysing images *en groupe*

As mentioned in the Introduction, the described design approach leverages Digital Methods Tools. Although Digital Methods have been initially designed for allowing domain-experts in the field of STS (Science and Technology Studies) to investigate textual content on the web, in recent years they have been used also for analysing and exploring visual content — such as images and memes — coming from the web (Social Network, Online News, Search engines results).

According to the theory of composite images (Niederer & Colombo, 2019) observing relationships among digital images deals with the study of them *en group*e and a recent approach for studying images *en groupe* refers to the use of existing metadata such as hashtag on Twitter or metadata attributed by Vision APIs Services as semantic tags. (Niederer & Colombo, 2019; Ricci et al., 2017) (Figure 3)

Vision APIs offer image content recognition based on predictions by returning a list of concepts with corresponding probabilities of how likely some concepts are contained within each image. (Figure 2)

When studying images *en groupe*, networks and networked images are the most adopted visual models (Ricci et al., 2017). This type of representations of contents allows researchers to conduct visual network analysis which focuses on the description of the network according to the topological distribution of nodes – images – in the space. (Pattern and cluster recognition, identification of bridging nodes, the relation between centre and periphery of the resulting network). (Venturini, Jacomy, & Pereira, 2016)

	A	B	C
1	url	concept	confidence
2	https://github.c	calendar	0.9903412
3	https://github.c	planner	0.9880015
4	https://github.c	date	0.98659444
5	https://github.c	daily occurence	0.9848956
6	https://github.c	monthly	0.9832363
7	https://github.c	almanac	0.9815499
8	https://github.c	schedule	0.97923803
9	https://github.c	agenda	0.9755267
10	https://github.c	vector	0.96764016
11	https://github.c	journal	0.96008015
22	https://github.c	time	0.9824935

Figure 2. Vision API Service (Clarifai) results. Images are labeled with concepts according to their visual content. The confidence value shows how likely some concepts are contained within each image.

In declaring their limits and peculiarities, we used Clarifai (Clarifai Inc., 2013) a free Vision APIs Service for automatically tagging images of visualizations extracted from databases of peer-reviewed scientific articles — Scopus (Elsevier, 2004), Google Scholar (Google, 2004) and Web of Science (Clarivate Analytics, 1997) –.

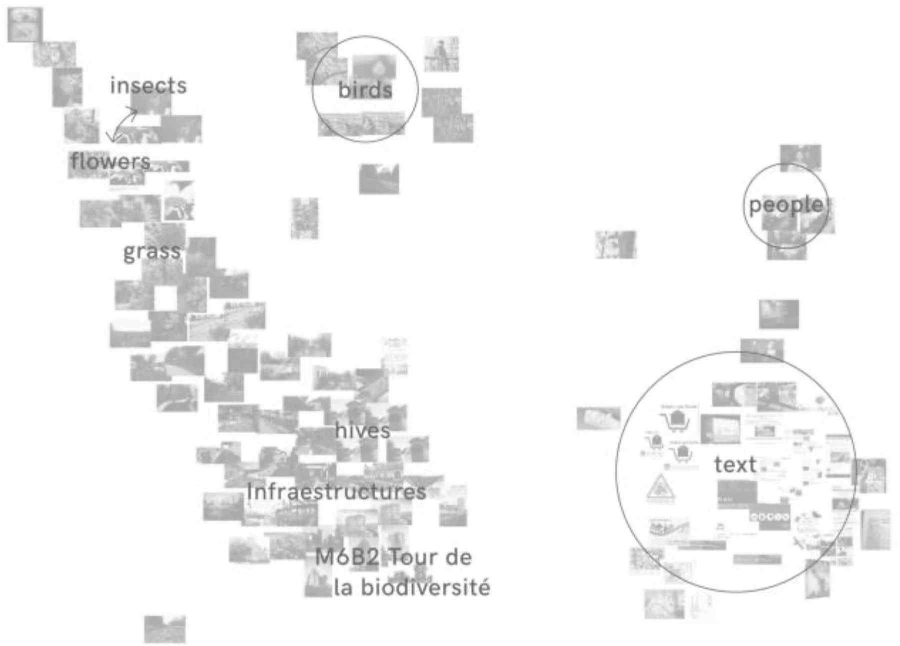

Figure 3. An example of visual network analysis on a network coming from the literature (Niederer, 2019) related to the analysis of pictures shared on Twitter about the topic of "nature" in Paris.

3. Design Requirements

Part of the process of choosing the most appropriate visual model to represent machine learning models' inner working is usually dictated by literature analysis, experimentation and evaluation.

The design of the tool was then guided by three points:

- **R1 - Visual-based approach:** Design a system able to provide information about the aesthetics and visual appearance of visualisation adopted for visualising Explaining Machine Learning processes.
- **R2 - A visual-based representation:** Design a method for visually representing a survey which differs from the evaluation matrix and the interactive browser.
- **R3 - A repeatable approach**: Design a system and a method that can be applied on similar conditions but referring to other (or filtered) data. (for instance, for analysing a specific group of machine learning models).

Final users involved are designers, whose aim is to provide new tools for explaining machine learning models' inner workings, with an experience both in the field of design and in the field of computer science.

Below, the defined action to demonstrate the effectiveness of the methodology:

- **T1 - Pattern recognition**: Recognise cluster of contents and identify visual peculiarities.
- **T2 - Inference**: Inferring by exploring, annotating and browsing the visualisation which type of machine learning models are presented.

4. Methodology

The presented experiment has been carried on with the intention of ensuring its reproducibility, also thanks to the support of the protocol diagram (Figure 5) that, acting like a visual recipe, shows progressively each step needed to achieve the final visualization: from data gathering to final results. **(Design Requirement n° 3)**

4.1 Data Gathering

Scientific articles have been gathered from three databases of peer-reviewed scientific articles (Scopus, Google Scholar and Web of Science) by using a consistent methodology. Each platform has been queried using *"explainable machine learning" OR "interpretable machine learning" AND "visual analytics*" and for each article the following information — if present — has been recorded:

- *Metadata* (title, authors, year, venue, link to the original document)
- *Images* including visualisations, illustrations, graphs, schemata.

While *Scopus* and *Web of Science* offer the possibility to perform advanced searches and download files in .csv format containing the aforementioned metadata (images excluded), for gathering data from *Google Scholar* the *WebScraper* (Magnetic Latvia, n.d.) an extension of *Google Chrome* has been used. Following this gathering process, a corpus of *238* articles has been obtained (*82* results from *Google Scholar*, *96* results from *Scopu*s and *60* results from *Web of Science*). Among these, only *6* papers are shared by all the three sources, *40* papers are shared by two of them and the remaining *192* are univocal.

By qualitatively browsing the results, emerged that some of these papers contain surveys and state of the art analysis, while others describe single case studies.

Then, scientific articles listed in the gathered surveys (Adadi & Berrada, 2018; Garcia et al., 2018; Hohman et al., 2018; Ming, 2017) have been added to the corpus and following this process a final collection of *257* scientific articles has been built.

However, not all the papers contained images, thus an additional filter was applied, allowing only papers containing images to be selected.

The final body of articles to be analysed contained a total amount of *194* documents whose pictures have been extracted using *Adobe Reader*. A final collection of *1059* images has been obtained. The final dataset is open source and available at this link.

4.2 Image tagging

Starting from the collection of images obtained through the qualitative and quantitative data gathering process, images have been tagged by Clarifai's Vision APIs Service. Clarifai offers ready-to-use image recognition models which allow to recognize different classes of semantic content in images to suit researchers' specific needs and, in declaring their limits and peculiarities, the general model has been tested being images contained in the corpus mainly visualisations and graphs. (An image recognition model for visual models and data visualisation has not yet been designed and would be the subject of further works).

To accelerate the process of image tagging, an interface[1] of Clarifai API has been used. By uploading a .csv file containing the list of URLs referring to each single image of the corpus the interface output another .csv with the results of the predictive model on each image. As mentioned in Section 2.2, observing and representing relations among digital images is practiced by using networks and networked images. Moreover, generating image-keywords networks using descriptive tags provided by the vision APIs has become a common approach. Starting from the .csv output file a network where images are connected by the same descriptive tag has been generated (.gexf file) and visualised using the open-source software Gephi by setting *ForceAtlas2* as spatialisation algorithm (Jacomy, Venturini, Heymann, & Bastian, 2014) and running modularity statistics to identify communities of nodes. Thanks to the Gephi's *ImagePreview* plugin (Chrisxue815, 2019) nodes have been replaced with images and the network has been printed out on paper to be annotated with pens by data visualisation designers.

[1] "This tool is an easier interface to the *Clarifai* API. What this API does in short is that it takes a list of images (in the form of URLs) and outputs what an algorithm, also called a model, sees in these images"

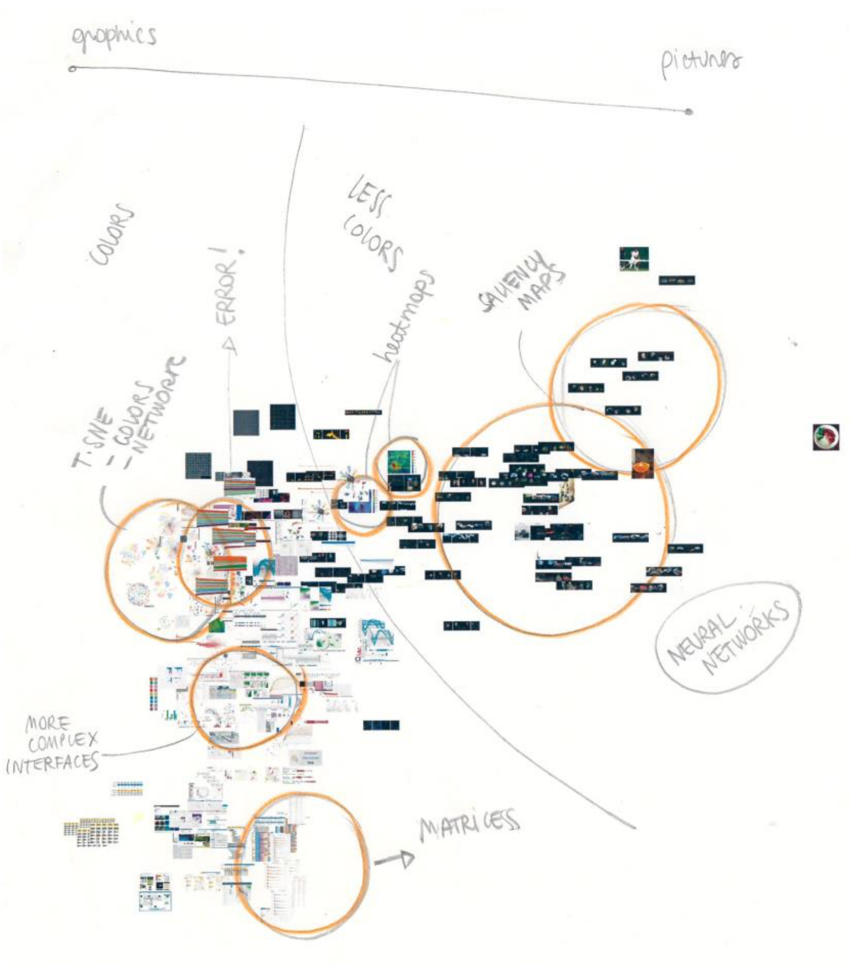

Figure 4. An example of visual network analysis performed by a designer on a network displaying images retrieved from a selection of the corpus

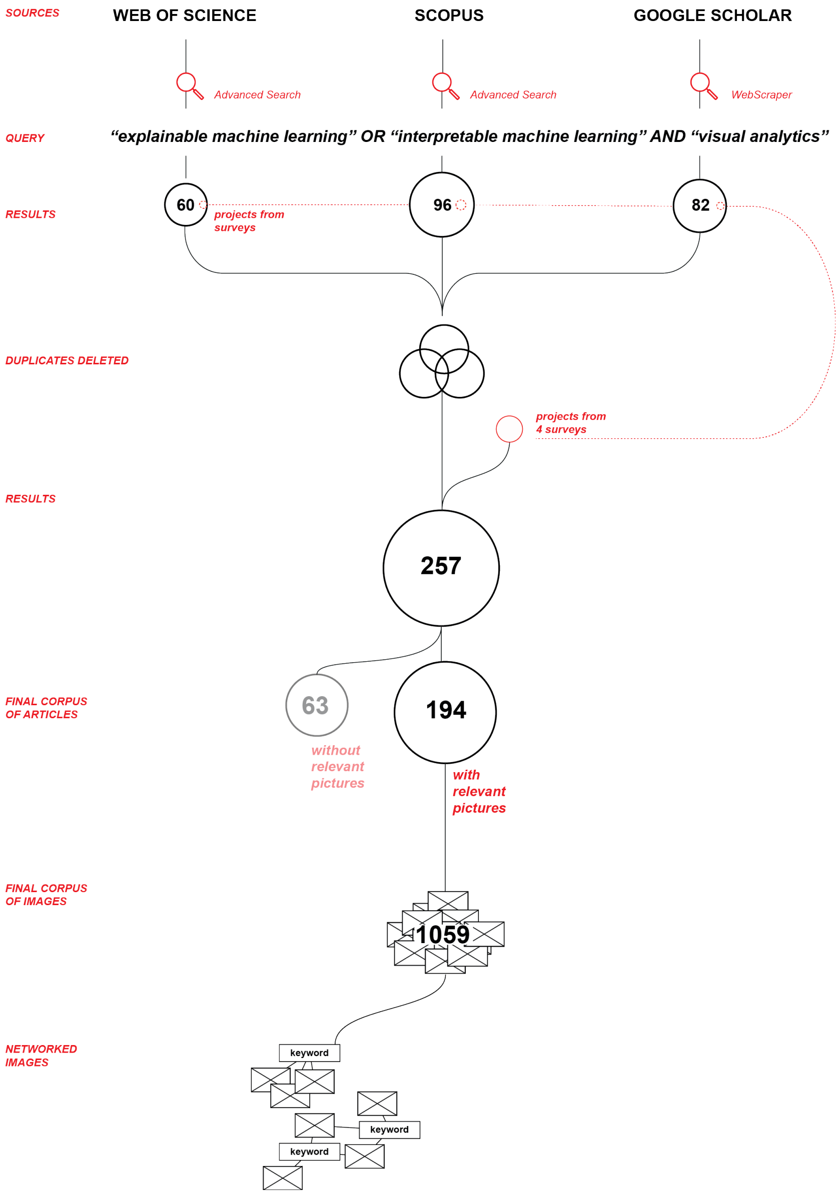

Figure 5. The visual protocol explaining the methodology.

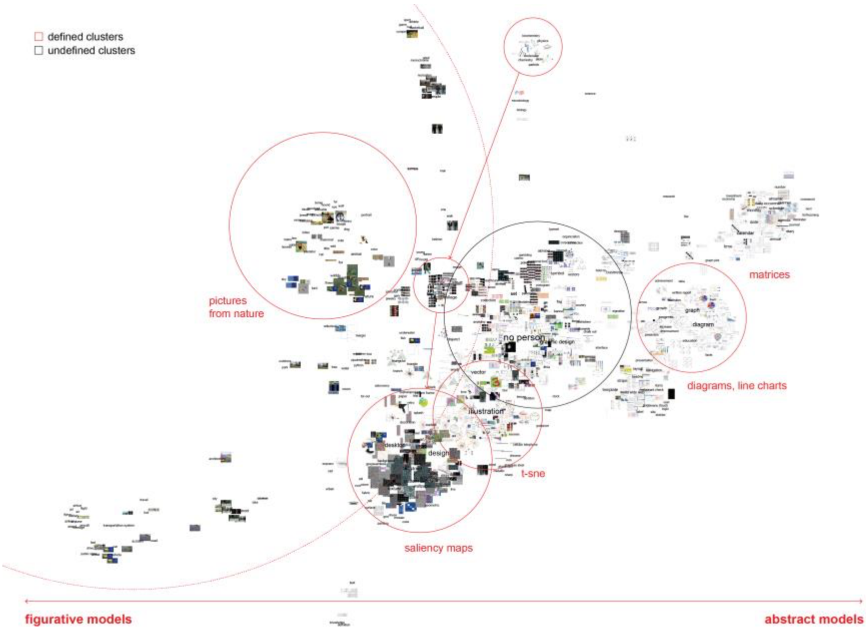

*Figure 6: A bipartite network showing the relations among pictures and corresponding labels gathered from scientific articles filtered by tags whose confidence is greater than 0.98 (Figure 2). The network can be divided in two main areas: the right side where **figurative models** are grouped and the left side where chart and graph models are clustered (**abstract models**).*

5. Results

The presented methodology has been tested on bipartite networks since they show connection between semantic tags and images. As the API module is not yet specifically designed to identify visual models, semantic tags are not very precise, however, it is possible to identify some clear clusters grouping similar visual models.

By observing the bipartite network (Figure 6) immediately emerges its bipolarity being the left-side populated by figurative visual models (pictures and saliency maps) while the right side is occupied by chart and graphs. Therefore, at first glance, emerges a central and generic cluster around the tag “no person” which is extremely generic, revealing the limits of the adopted vision APIs model but, at the same time, identifying a group of images to be tested with other vision APIs services. (Figure 7)

By looking more closely appears that the right side is populated by two well defined clusters: the first, on the top-right (Figure 7 and 8b) which mainly displays matrices that the Vision APIs Service recognised as calendar", "crosswords", while the latter, slightly below, (Figure 7 and 8a) contains charts and graphs.

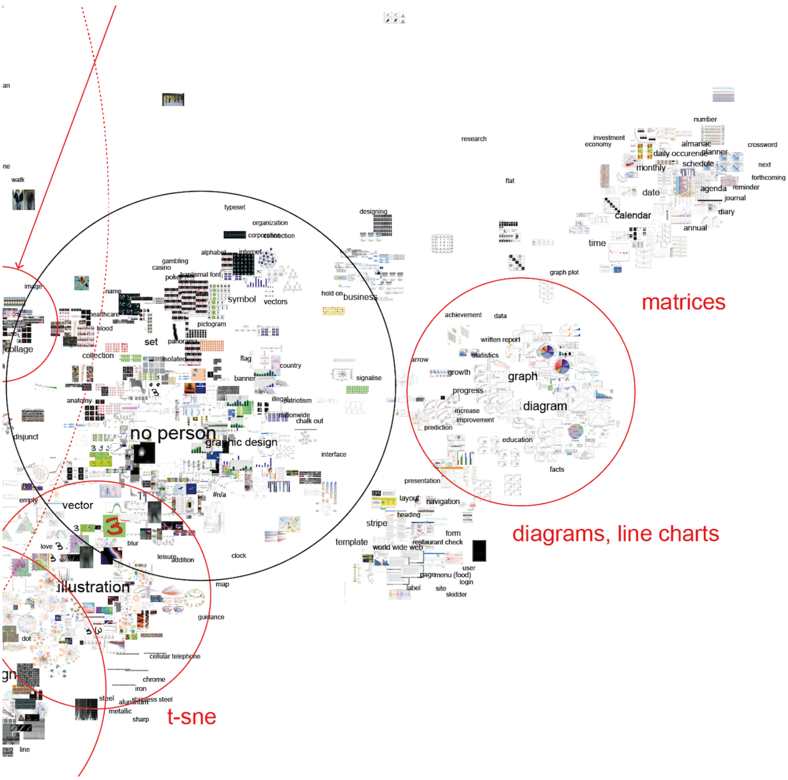

Figure 7: In a bipartite network, two different types of nodes — images and keywords — are connected. Here a portion of Figure 6 is shown, highlighting the most evident clusters.

Moving to the left side, even the cluster of t-SNe2 representation is evident (Figure 7 and 9a) where visual models containing "dots" are clustered. Instead, in Figure 9b which is located in the figurative side of the bipartite network, users can browse picture portraying animals and objects.

Moreover, the bipolar structure of the network also allows the observer to jump from one side to the other. For instance, users can see how the saliency maps cluster could be ideally linked to the one related to graphical representations of neural networks on the top (Figure 6, circles connected by the straight line).

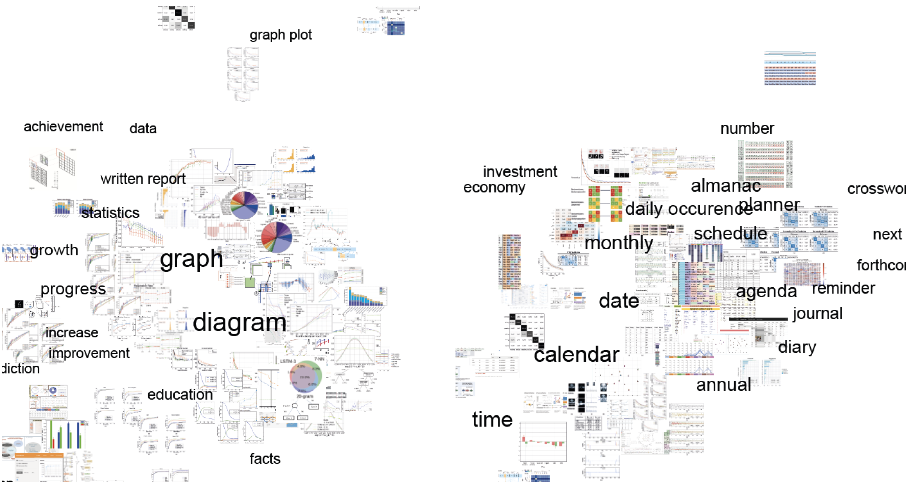

Figure 8: This is a zoom of Figure 6. On the left the group of bar charts and line charts. On the right, the cluster of matrices.

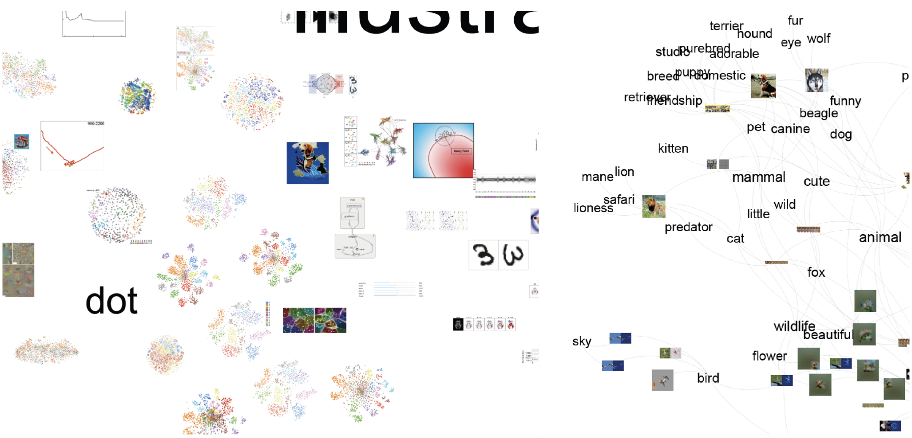

Figure 9: This is a zoom of Figure 6. On the left the group of t-SNe. On the right the cluster of animals.

6. Discussion

In this paper, we propose an experimental and interdisciplinary methodology for supporting designers in the action of browsing and choosing visual models through the presentation of a visual-based survey displaying images.

To our knowledge, this is the first attempt to design a data visualisation for showing metadata collections and images related to scientific publications.

The intention of the paper is not to propose an alternative system to the evaluation matrices seen in the literature but to promote an integrative device that supports the exploration of surveys and help designers in the choice of visual models**. (R1).**

A visual-based representation of contents **(R2)** helps designers in entering and integrating into the explainable machine learning field.

Moreover, the visual protocol allows designers to repeat the operation and apply the methodology even on other bibliographic collections. **(R3)**

However, we want to point out some weaknesses that may be addressed and fixed in the future.

- The paper selection has been carried out both qualitatively and quantitatively. Thus, the final corpus is subjective but what is relevant for this research is the methodology applied and its flexibility.
- The number of images for each scientific article is slightly different.
- The Vision API Service selected is biased too, the same process could be run on
- different APIs services (for instance: *Google Vision APIs, Microsoft Azure, IBM Watson, Imagga*) being their results potentially different. In the future could be interesting to design a training dataset able to identify visual models.
- The final networked images visual model could be interactive, allowing a dynamic browsing of contents and an augmented visual network analysis.

The presented organic and complete representations of visual models help designers in the act of browsing visual contents contained in scientific articles, and give the possibility to have, at first sight, an overall picture of the state of the art.

Moreover, visual network analysis helps in understanding relations between clusters and immediately identify relevant visual references in the design process

6. References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Bertin, J. (1967). Sémiologie graphique. *Image (Rochester, N.Y.)*, 7–10.

Chrisxue815. (2019). *Gephi Image Preview Plugin*.
Clarifai Inc. (2013). *Clarifai*.
Clarivate Analytics. (1997). Web of Science.
Correll, M. (2019). *Ethical Dimensions of Visualization Research*. 1–13.
Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. (Ml), 1–13. Retrieved from http://arxiv.org/abs/1702.08608
El-assady, M., Jentner, W., Kehlbeck, R., & Schlegel, U. (2019). *Towards XAI : Structuring the Processes of Explanations Towards XAI : Structuring the Processes of Explanations*. (May).
Elsevier. (2004). *Scopus*. Retrieved from www.scopus.com
Garcia, R., Telea, A. C., Castro da Silva, B., Tørresen, J., & Dihl Comba, J. L. (2018). A task-and-technique centered survey on visual analytics for deep learning model engineering. *Computers and Graphics (Pergamon)*.
Google. (2004). Google Scholar.
Hohman, F. M., Kahng, M., Pienta, R., & Chau, D. H. (2018). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 1–20. https://doi.org/10.1109/TVCG.2018.2843369
Isenberg, P., Heimerl, F., Koch, S., Isenberg, T., Xu, P., Stolper, C., … Torsten, M. (2017). *vispubdata . org : A Metadata Collection about IEEE Visualization (VIS) Publications To cite this version : HAL Id : hal-01376597 vispubdata . org : A Metadata Collection about IEEE Visualization (VIS) Publications*.
Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE*, *9*(6), 1–12. https://doi.org/10.1371/journal.pone.0098679
Kucher, K. (2014). Text Visualization Browser : A Visual Survey of Text Visualization Techniques. *InfoVis*, (November 2014), 1–3.
Lipton, Z. C. (2016). *The Mythos of Model Interpretability*. (Whi).
Liu, S., Wang, X., Collins, C., Dou, W., Ouyang, F., El-Assady, M., … Keim, D. (2018). Bridging Text Visualization and Mining: A Task-Driven Survey. *IEEE Transactions on Visualization and Computer Graphics*, *2626*(c).
Lu, Y., Garcia, R., Hansen, B., Gleicher, M., & Maciejewski, R. (2017). The State-of-the-Art in Predictive Visual Analytics. *Computer Graphics Forum*, *36*(3), 539–562.
Lu, Y., Wang, H., Landis, S., & Maciejewski, R. (2018). A Visual Analytics Framework for Identifying Topic Drivers in Media Events. *IEEE Transactions on Visualization and Computer Graphics*, *24*(9), 2501–2515. https://doi.org/10.1109/TVCG.2017.2752166
Magnetic Latvia. (n.d.). Webscraper.
Ming, Y. (2017). *a Survey on Visualization for Explainable Classifiers*. (October). Retrieved from http://www.myaooo.com/projects/survey-xml/survey-1103.pdf
Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967.
Niederer, S., & Colombo, G. (2019). Visual Methodologies for Networked Images: Designing Visualizations for Collaborative Research, Cross-platform Analysis, and Public Participation. *Diseña*, *14*, 40–67.
Offert, F. (2017). *“I know it when I see it”. Visualization and Intuitive Interpretability*. (Nips). Retrieved from http://arxiv.org/abs/1711.08042
Pearce, W., Özkula, S. M., Greene, A. K., Teeling, L., Bansard, J. S., Omena, J. J., & Rabello, E.