Recent Studies of XAI - Review

Recent Studies of XAI - Review

Zhongli Filippo Hu*
University of Turin
Italy
zhonglifilippo.hu@unito.it

Tsvi Kuflik*
The University of Haifa
Israel
tsvikak@is.haifa.ac.il

Ionela Georgiana Mocanu*
The University of Edinburgh
Scotland
i.g.mocanu@ed.ac.uk

Shabanam Najafian*
Delft University of Technology
The Netherlands
s.najafian@tudelft.nl

Avital Shulner-Tal*
The University of Haifa
Israel
avitalshulner@gmail.com

ABSTRACT

Over the past years, there has been an increasing concern regarding the risk of bias and discrimination in algorithmic systems, which received significant attention amongst the research communities. To ensure the system's fairness, various methods and techniques have been developed to assess and mitigate potential biases. Such methods, also known as "Formal Fairness", look at various aspects of the system's advanced reasoning mechanism and outcomes, with techniques ranging from local explanations (at feature level) to visual explanations (saliency maps). Another aspect, equally important, represents the perception of the users regarding the system's fairness. Despite a decision system being provably "Fair", if the users find it difficult to understand how the decisions were made, they will refrain from trusting, accepting, and ultimately using the system altogether. This raised the issue of "Perceived Fairness" which looks at means to reassure users of a system's trustworthiness. In that sense, providing users with some form of explanation on why and how certain outcomes resulted, is highly relevant, especially nowadays as the reasoning mechanisms increase in complexity and computational power. Recent studies suggest a plethora of explanation types. The current work aims to review the recent progress in explaining systems' reasoning and outcome, categorize and present it as a reference for the state-of-the-art fairness-related explanations review.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Explainability, Perceived fairness, algorithmic transparency

* The authors contributed equally to this work.

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8367-7/21/06…$15.00
https://doi.org/10.1145/3450614.3463354

ACM Reference Format:
Zhongli Filippo Hu*, Tsvi Kuflik*, Ionela Georgiana Mocanu*, Shabanam Najafian*, and Avital Shulner-Tal*. 2021. Recent Studies of XAI - Review. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3450614.3463354

1 INTRODUCTION

The widespread use of Artificial Intelligence (*AI*) and Machine Learning (*ML*) systems, for convenience we will relate to them as algorithmic systems (*AS*), established the understanding that it is necessary to create "algorithmic regulation" for such systems in order to prevent discrimination, unfairness and harm to individuals as well as groups [9, 64, 70].

According to O'Reilly [54], the following four main features should be included in this type of regulation: "1) A deep understanding of the desired outcome. 2) Real-time measurement to determine if that outcome is being achieved. 3) Algorithms (i.e. a set of rules) that make adjustments based on new data. 4) Periodic, deeper analysis of whether the algorithms themselves are correct and performing as expected".

The most known regulation in this area is the GDPR[1] that was adopted in 2016 and became enforceable in 2018. In particular, Recital 71[2] expands on the right to explanation, which refers to an individual's right to receive an explanation regarding decisions, either legally or financially, that concern or affect the individual. Essentially, this reiterates the need to provide an explanation for the reasoning and outcomes of algorithmic systems. This is a difficult task since there are various types and styles of explanations that can be provided.

In general, the term "explanation" is defined as a set of statements that clarifies something or provides a reason for an action or belief. A single explanation can be grouped into multiple categories such as the purpose of the explanation, the type of the explanation, the presentation form, the context of the explanation, and so on. In this paper, we aim to provide an initial overview of the various categories of explanations that an *AS* can provide. The first three categories (purpose, interpretation method, and context) address the content of the explanation and the additional three categories (presentation format, stakeholder type, and domain) address the configuration of the explanation. We can also view those aspects

[1] https://gdpr-info.eu/chapter-3/
[2] https://gdpr-info.eu/recitals/no-71/

as explanation-related aspects (purpose, interpretation method, context, and presentation format) and as environmental aspects (stakeholder type and domain).

Therefore, our intention in this study is to review the various categories of explanations that can be provided by *AS*s in recent years (starting from 2016, when the need for explanations from *AS*s has debuted).

2 TYPES/ASPECTS OF EXPLANATION

2.1 Purpose

The first, most basic classification can be according to the purpose of the explanation. We can discuss two basic purposes [40, 45]:

- **"How" explanations** – explanations that clarify how something works or facilitates the understanding of something.
- **"Why" explanations** – explanations that clarify why something happened.

Explanations play an important role in our everyday life, they are used for building and storing our knowledge regarding things that happen around us [1]. An explanation is a process used to increase understanding of why and how events/phenomena/concepts occur. It tells how something is done (explaining the different steps in the process) and why something was performed (give reasons for carrying out the process) [1]. Explanations employ four causal patterns [23]:

a) Common cause - explains the initial reason that has implications for the decisions made later. This pattern is often used in the diagnosis of problems, such as medical illnesses, equipment failure, or software bugs (why decisions were made this way).
b) Common effect - explains the set of factors that caused a creation event. Such explanations are common in history, where the factors that enabled the existence of a major event can be explained (how the event happened).
c) Linear causal chains - a special case of common cause and common effect explanations, which represents a series of steps that started from a single initial cause and caused a single effect (why and how an event happen)
d) Causal homeostasis - explains how an integrating set of causes and effects results in a stable set of properties that exist over time.

Understanding why and how something happens enables us to predict and control phenomena and then explain them to others which is the primary goal of science [43].

2.2 Interpretation method

In terms of the interpretation method used to produce the explanations, they can be classified into two main types [37, 46]:

- **Local explanations** – outline why a specific outcome was received for a single instance (explaining what the model will predict for a specific input).
- **Global explanations** – show how the algorithmic model works based on its features and components (explaining the model as a whole).

The reader may notice a soft connection between the purpose of the explanation and the interpretation method category. Local explanations are in general linked to the "why" explanations while global explanations are associated with the "how" explanations.

Local explanations. Local explanations aim to explain individual cases locally. These types of explanations are generally relevant to *ML* algorithms where the decision has an impact on an individual's *"right for explanation"* (GDPR). The local explanations embrace the advantage of being easily implemented and have a low computation complexity [35]. We identify local explanations according to their application: model-specific and model-agnostic (Figure 1).

The model-specific techniques generate more precise explanations as they directly depend on the model to be interpreted. However, this approach does often lead to consistency issues, in particular when comparing the resulting explanations of two models of different structures. The explanations process is linked to the algorithm used by that particular model and any new architecture/models proposed would be required to find new methods of model exploration and explanation.

For these reasons often the model-agnostic techniques are preferred, as they do not assume any information about the model structure. These techniques only analyse the data from the inputs and the result. The main advantage constitutes their flexibility: because the explanation process is dealt with separately from the actual algorithm used, the user has the freedom of choosing any *ML* model desired. Examples of agnostic techniques include *LIME* (Local Interpretable Model-agnostic Explanations) [59], *SHAP* (SHapley Additive exPlanations) [39]. The disadvantage is that often these techniques are based on replacement models (surrogates models) which reduces the quality of explanations provided. LIME is an explanation technique that returns predictions of any classifier in a manner that can be easily interpretable and trustworthy. The core idea behind this technique is that it learns a new simplified model locally around the prediction. LIME produces explanations by locally approximating the selected model with a more interpretable one (such as linear models or decision trees) [59].

Global explanations. Global explanations aim at providing a comprehensive explanation of how the model performs as a whole, as opposed to the local explanations which explain a single prediction. Global explanations give an understanding of the overall function performed and can take various forms, from source code used to training data or simple descriptions of how a search algorithm works or a user's manual, explaining the use and functions of a new product [4]. As a minimum requirement, such examples of user's manual should contain at least information regarding the employment and development of the given intelligent system, the data used in development or training, effectiveness or performance of the system, general logic, etc.

Global explanations can explain feature importance across a population, for example, could aid in the model diagnosis, expose possible biases, or advance feature engineering techniques. Generally, all local explainability methods can be aggregated into global techniques, and also global attributions can reduce Neural Networks decisions to a single set of features [19].

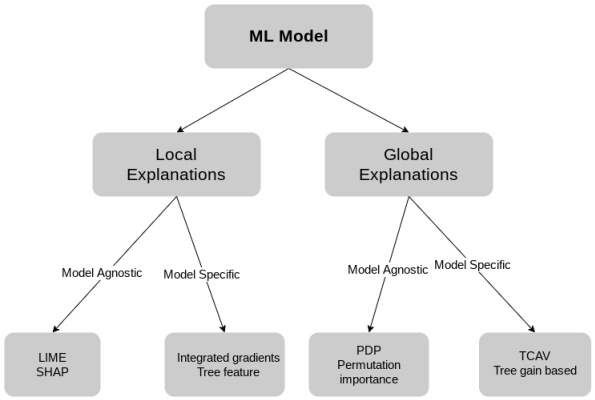


Figure 1: The local and global classification of machine learning models

2.3 Context

The third classification is performed according to the context of the explanation. It can be divided to two types as well [38, 46]:

- **Individual explanations** – explanation that clarifies the outcome of the system for a single instance.
- **Group explanations** – explanation that clarifies the outcome of the system for a group of instances.

Individual explanations. A majority of studies focus on proposing explanation approaches for single users ([11, 13, 16]). A very common technique is to indicate similarity between decision-making system's outcomes and items of the user's preference. In the case of recommender systems, the user's preference could be defined as the items that the user is currently browsing or has expressed a preference for in the past. We identify this kind of explanation being used by Amazon for example *"Users who bought X also bought Y"*. Similar kinds of explanations are applied by, for instance, Netflix and Spotify [29].

Recent works in the context of individual explanations focus on the perception of transparency and trust in the system by augmenting the decision-making system's outcomes with explanations [29, 67].

Group explanations. Which information an explanation should provide to the user can depend on various factors. Group decisions can be seen as a very specific context, that focuses on explaining decision-making processes that are made for a group of users [52]. Initial related work can be found, for example, in [57], where social factors in groups are taken into account to generate tactful explanations for groups, i.e., explanations that avoid, for example, damaging friendships. The existing works on generating explanations for group decision-making primarily consider the need for transparency, i.e., to clarify the reasoning and data behind a decision to help users better understand how the decision-making system works and why a specific decision has been made. Potential goals of group explanations are discussed for example in Najafian and Tintarev [50] and Jameson et al. [21]. However, when generating explanations for groups rather than individuals, privacy becomes of great relevance as well. Initial related work can be found, for reference, in Najafian et al. [49], who investigated the degree of information disclosure that users prefer when it comes to group recommendations in the music domain. Similarly, Najafian et al. [48] studied the factors that influence people's privacy concerns when an explanation is presented to a group in the tourism domain.

2.4 Presentation format

The fourth classification type is the explanation's presentation format. We hereby identify the following types:

- **Textual explanations** – explanations that are presented using natural language (e.g text, audio).
- **Visual explanations** – explanations that are presented using visualizations (e.g. graph, image).

Textual explanations. People usually give each other explanations verbally. The textual approach uses natural language processing techniques to generate sentences that systems can provide to users. In the field of recommender systems, the simplest method of textual display is the listing of the various alternatives, possibly with the show of features of the suggested item. A more elaborate explanation format is the use of canned texts, which are sets of parts of the text that combined form a sentence. Another approach is the usage of templates, which are sentences that need to be completed with a list of topics.

The type of explanation generation depends on the prediction model used. In the case of *black-box* models, for example, models based on ensemble classifiers [69], and deep neural networks [14], the focus is not on creating transparent and easy-to-interpret methods but rather on interpreting already trained complex models. A separate model is then usually created to generate the explanation, usually called a post-hoc justification. Lei et al. [33] proposed a method of generating *rationales*, which are small chunks of sentences extracted from input sentences used to provide justifications. Similarly, Krening et al. [27] proposed a reinforcement learning model to achieve this, but using a recurrent neural network, while McAuley and Leskovec [42] created a model to explain the recommendations of a latent factor model. Musto et al. [47] created three post-hoc generation models, starting from the reviews of the suggested items. The advantage of this approach is that this method is independent of the recommendation model. This paper proposes a natural language processing and sentiment analysis methodology to extract the most important aspects of the reviews and uses templates or text summarization to generate the justifications.

In the case of *white-box* models, a generation of explanation derives from the interpretation of the output without creating a new model. Examples of white-box models are linear ones [59], based on decision trees [30] or rule-based. A type of explanation with white-box models shows the user the inference traces. Other particular explanations are the vocal ones [65], query results [3], and the generation of OWL (Ontology Web Language) knowledge [63]. Iovine et al. [20] recently proposed a Conversational Recommender System, where the explanation was given by a chatbot. This study introduces the possibility of implementing justification by voice assistants.

Visual explanations. Visual explanation offers significant insight into the system's reasoning, in particular for comparison between

system results. In that sense, O'Donovan et al. [53] proposed a movie collaborative filtering recommender system with a graphical interface called PeerChooser, which graphically explains the suggestions in a very intuitive manner (Appx. B.1). In this representation, we have the user in the center of the graph and a set of similar users. The closer the other users are, the more similar their preferences are to the ones of the current user. Each user has some links to the genres of movies they watch. In the domain of music services, Gou et al. [15] offers a graph to explain social friend recommendation (Appx. B.2). The inner part of this graph is a donut plot with subgroups and represents music genres.

Parra et al. [55] uses an interactive Venn diagram for a recommender system on Research Talks or Articles (Appx. B.4.a). Users can filter the list of items by clicking on the area inside. In this paper and other similar, sliders are used to help the user filtering. In the same domain, Tsai and Brusilovsky [68] proposes an interface that uses scatter plots (Appx. B.3).

Similarly, Kouki et al. [25] uses a Venn diagram to explain suggestions in the restaurant domain (Appx. B.4.b). Additionally, they use an interface with concentric circles to explain collaborative filtering suggestions (Appx. B.7.b). In this plot, the user entity is the central part of the graphical representation and is connected with the restaurants' node which is most preferred from the immediately outermost circle. Each restaurant is in turn connected to other users who preferred it. In the external circle, there are other restaurants that these similar users liked and therefore suggested by the system. (See Appx. B.7.a for another proposal of visual explanation, one which denotes: items preferred by the user, recommendation context, and various suggestions for the user.)

Millecamp et al. [44] uses a bar diagram to explore the musical domain (Appx. B.6). Each bar describes a musical feature: for instance, acousticness, popularity, and tempo. It uses the preferences as received from the users, and then it displays a list of songs in the context of the other users' preferences. In the real-estate domain, Mauro et al. [41] proposes INTEREST, a visualization model that uses a bar diagram to support the exploration and analysis of search results by graphical representation of consumer feedback by multiple stages. To summarize information about a single item, Kouki et al. [26] uses dendrograms to represent the information in a tree structure (Appx. B.5). This diagram is closer to sentence format in natural language, in particular to templates.

2.5 Stakeholders

The fifth classification relates to the stakeholders (or stakeholders type). In general, a classical definition of stakeholders of algorithmic systems is: *"anyone who could be affected by the use of an algorithm in some context"* [8].

To exemplify, the authors provide the following example: "Consider an AI algorithm used to automatically assess student writing in an English class. In this case, stakeholders would include the students, parents, teachers, the school or university administering the course, the vendor that created the algorithm, and any regulatory body operating in this domain (i.e., U.S. Department of Education)". Clearly, there is a wide diversity among these stakeholders and this example applies in fact to every domain we consider. [32] provides a similar analysis, showing that different stakeholders may have different views on fairness. When it comes to explanations, they can be given at different levels of detail and using different levels of professional terms, in order to accommodate the needs of the diversity of stakeholders. Therefore, the specific stakeholder (or stakeholder's type) should be considered when creating an explanation, as stakeholders vary in their level of knowledge and understanding of these systems. As we take a closer look at the matter, following past research, we can identify three main types of stakeholders and consider the types of explanations that may be suitable for each one of them [2, 24, 56, 62]:

a) **Developers** – which are people involved in building algorithmic systems (i.e. developers, modellers, *AI/ML* domain experts, data scientists, managers, product owners). Clearly, the explanations that are provided to them may be sufficiently detailed and technical.
b) **Observers** – which are people involved in understanding and promoting the algorithmic systems (i.e. academic or industrial researchers, domain experts, data scientists, theorists, ethicists) or people who are engaged in examining the fairness, accountability and transparency systems (i.e. regulators, auditors, policy-makers, commentators, critics). Explanations that are given to these stakeholders may need to be adapted to their needs and background, less technical and task-related.
c) **Users** – which are people who use the algorithmic systems or that are affected by the systems' results (i.e. domain experts who operate the system, users who are affected by the system results). Explanations that are given to them need to be specific for their case and adapted for their level of technical skills, usually non-experts that need to understand the results of a specific case.

2.6 Domain

And the last classification relates to the domain of the system. There are many aspects of our everyday life that are impacted in various ways such as health, well-being, employment, social, cultural, financial, and many more [61]. Following the increased use of *AS*, we come across such systems in a variety of fields that may require different explanation types that will need further adjustments to the specific domain. We identify *AS*s in the following domains:

- Legal systems (i.e. legal mediation systems, legal recommendations)
- Medical systems (i.e. medical recommendations, identification of medical conditions)
- Recruitment systems (i.e. hiring recommendation, job portals)
- Social systems (i.e. social media, dating systems)
- Entertainment systems (i.e. movie recommendation, music recommendation)

For example, explanations for the medical domain may be used for medical education and research and/or for providing reasons for clinical decision making [18]. Medical explanations should be presented in a way that medical professionals would understand how and why the decision has been made in order to decide and explain the medical treatment process that they recommend to their patients [18, 58].

While in the legal domain, case-based explanations are usually used, they present similarities between a specific case and other past cases in the history of law and they usually present a conclusion of the decision and a rule that the decision was based on [5, 28]. Another example from the entertainment domain is the explanations that present a coherent set of reasons that are in favor or against a movie recommendation [7].

3 DISCUSSION AND CONCLUSIONS

This work proposes a classification of explanations according to six aspects: purpose, interpretation method, context, presentation format, stakeholder type and domain. Various explanations styles from recent years with categorization are shown in Table 1 in Appendix A. Those aspects can be divided according to the content of the explanation (first three aspects) and the configuration of the explanation (last three aspects) or according to explanation-related aspects (first four aspects) and environmental aspects (last two aspects). This initial review of various explanation styles may help in adapting explanations according to the requested aspects of the system. Furthermore, it may contribute to selecting the most appropriate explanation according to the stakeholder's needs. Of course there is a need to expand this review and to categorize more explanation styles that are used in the literature and in various systems. In addition, this classification framework can be easily expanded when other relevant factors are identified as well as sub-aspects such as user models. The reviewed state-of-the-art explanations followed mainly "one size fits all" paradigm, however it is possible that different types of explanations may be preferred or relevant to different users. For example, there are initial works on which factors one should model in the group to generate privacy-preserving explanations [48, 49]. Another example for personalizing explanations can be found in Quijano-Sanchez et al. [57], where they extended work on generating group explanations by including the social factors of personality and tie strength between group members involved in the recommendations decision-making processes. Further work is required for modeling individual or group of users to personalize explanations based on users' needs.

ACKNOWLEDGMENTS

This study was supported by the Cyprus Center for Algorithmic Transparency, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 810105 (CyCAT – Call: H2020-WIDESPREAD-05-2017-Twinning).

Ionela Georgiana Mocanu was supported by the Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Pervasive Parallelism (grant EP/L01503X/1) at the University of Edinburgh, School of Informatics.

Zhongli Filippo Hu was supported by PhD School, Computer Science Department, University of Turin.

REFERENCES

[1] Shela Silviana Augie et al. 2019. *Grammatical Cohesive Devices in Students' Explanation Texts (A Study of the Fourth Semester Students of English Department, Universitas Negeri Semarang Academic Year 2018/2019).* Ph.D. Dissertation. UNNES.

[2] Alejandro Barredo Arrieta, Natalia Diaz Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado González, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, V. Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (12 2019). https://doi.org/10.1016/j.inffus.2019.12.012

[3] A. Basu and R. Ahad. 1992. Using a relational database to support explanation in a knowledge-based system. *IEEE Transactions on Knowledge and Data Engineering* 4, 6 (1992), 572–581. https://doi.org/10.1109/69.180608

[4] Valérie Beaudouin, Isabelle Bloch, David Bounie, Stephan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. 2020. Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach.

[5] Trevor J. M. Bench-Capon. 2020. Explaining Legal Decisions Using IRAC. In *Proceedings of the 20th Workshop on Computational Models of Natural Argument co-located with the 8th International Conference on Computational Models of Argument (COMMA 2020), Perugia, Italy (and online), September 8th, 2020 (CEUR Workshop Proceedings, Vol. 2669)*, Floriana Grasso, Nancy L. Green, Jodi Schneider, and Simon Wells (Eds.). CEUR-WS.org, 74–83. http://ceur-ws.org/Vol-2669/paper10.pdf

[6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. *CoRR* abs/1801.10408 (2018). arXiv:1801.10408 http://arxiv.org/abs/1801.10408

[7] Cristian Briguez, Maximiliano Budán, Cristhian Deagustini, Ana Maguitman, Marcela Capobianco, and Guillermo Simari. 2014. Argument-based mixed recommenders and their application to movie suggestion. *Expert Systems with Applications* 41 (10 2014), 6467–6482. https://doi.org/10.1016/j.eswa.2014.03.046

[8] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8, 1 (2021), 2053951720983865. https://doi.org/10.1177/2053951720983865 arXiv:https://doi.org/10.1177/2053951720983865

[9] Miriam C Buiten. 2019. Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation* 10, 1 (2019), 41–59.

[10] Shruthi Chari, Oshani Seneviratne, Daniel M. Gruen, Morgan A. Foreman, Amar K. Das, and Deborah L. McGuinness. 2020. Explanation Ontology: A Model of Explanations for User-Centered AI. arXiv:2010.01479 [cs.AI]

[11] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 125–150.

[12] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. *CoRR* abs/1901.07694 (2019). arXiv:1901.07694 http://arxiv.org/abs/1901.07694

[13] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[14] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *CoRR* abs/1806.00069 (2018). arXiv:1806.00069 http://arxiv.org/abs/1806.00069

[15] Liang Gou, Fang You, Jun Guo, Luqi Wu, and Xiaolong (Luke) Zhang. 2011. SFViz: Interest-Based Friends Exploration and Recommendation in Social Networks. In *Proceedings of the 2011 Visual Information Communication - International Symposium* (Hong Kong, China) *(VINCI '11)*. Association for Computing Machinery, New York, NY, USA, Article 15, 10 pages. https://doi.org/10.1145/2016656.2016671

[16] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[17] Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*. Association for Computing Machinery (ACM), New York, 219–225. https://doi.org/10.1145/3386392.3399302

[18] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery* 9, 4 (2019), e1312. https://doi.org/10.1002/widm.1312 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312

[19] Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. 2019. Global Explanations of Neural Networks: Mapping the Landscape of Predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 279–287. https://doi.org/10.1145/3306618.3314230

[20] Andrea Iovine, Fedelucio Narducci, and Giovanni Semeraro. 2020. Conversational Recommender Systems and natural language:: A study through the ConveRSE framework. *Decision Support Systems* 131 (2020), 113250. https://doi.org/10.1016/j.dss.2020.113250

[21] Anthony Jameson, Martijn C. Willemsen, Alexander Felfernig, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. 2015. *Human decision making and recommender systems* (2nd ed.). Springer, Germany, 611–648. https://doi.org/10.1007/978-1-4899-7637-6_18

[22] Öykü Kapcak, Simone Spagnoli, Vincent Robbemond, Soumitri Vadali, Shabnam Najafian, and Nava Tintarev. 2018. Tourexplain: A crowdsourcing pipeline for generating explanations for groups of tourists. In *Workshop on Recommenders in Tourism co-located with the 12th ACM Conference on Recommender Systems (RecSys 2018)*, Vol. 2222. CEUR.

[23] Frank C. Keil. 2006. Explanation and Understanding. *Annual Review of Psychology* 57, 1 (2006), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100 PMID: 16318595.

[24] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. arXiv:1806.03281 [stat.ML]

[25] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User Preferences for Hybrid Explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 84–88. https://doi.org/10.1145/3109859.3109915

[26] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 379–390. https://doi.org/10.1145/3301275.3302306

[27] S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl, and A. Thomaz. 2017. Learning From Explanations Using Sentiment and Advice in RL. *IEEE Transactions on Cognitive and Developmental Systems* 9, 1 (2017), 44–55. https://doi.org/10.1109/TCDS.2016.2628365

[28] Teo Kuan, Wah, and Muniandy Manoranjitham. 2014. Courtroom Decision Support System Using Case Based Reasoning. *Procedia - Social and Behavioral Sciences* 129 (10 2014), 489–495. https://doi.org/10.1016/j.sbspro.2014.03.705

[29] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 487.

[30] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[31] Retno Larasati, A. Liddo, and E. Motta. 2020. The Effect of Explanation Styles on User's Trust. In *ExSS-ATEC@IUI*.

[32] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management That Allocates Donations to Non-Profit Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3365–3376. https://doi.org/10.1145/3025453.3025884

[33] Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing Neural Predictions. *CoRR* abs/1606.04155 (2016). arXiv:1606.04155 http://arxiv.org/abs/1606.04155

[34] Qing Li, Sharon Chu, Nanjie Rao, and Mahsan Nourani. 2020. Understanding the Effects of Explanation Types and User Motivations on Recommender System Use. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 83–91. https://ojs.aaai.org/index.php/HCOMP/article/view/7466

[35] Yu Liang, Siguang Li, Chungang Yan, Maozhen Li, and Changjun Jiang. 2021. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing* 419 (2021), 168–182. https://doi.org/10.1016/j.neucom.2020.08.011

[36] Brian Y. Lim, Qian Yang, Ashraf Abdul, and Danding Wang. 2019. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. In *IUI Workshops*.

[37] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (2021), 18.

[38] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[39] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR* abs/1705.07874 (2017). arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[40] Anderson Mark and Anderson Kathryn. 1997. *Text types in English. 2 / Mark Anderson, Kathryn Anderson.* Macmillan Education Australia South Yarra. v, 170 p. : pages.

[41] Noemi Mauro, Liliana Ardissono, Sara Capecchi, and Rosario Galioto. 2020. Service-Aware Interactive Presentation of Items for Decision-Making. *Applied Sciences* 10, 16 (2020). https://doi.org/10.3390/app10165599

[42] Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 165–172. https://doi.org/10.1145/2507157.2507163

[43] Kevin McCain. 2015. Explanation and the Nature of Scientific Knowledge. *Science & Education* 24, 7-8 (2015), 827–854. https://doi.org/10.1007/s11191-015-9775-5

[44] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2020. What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, Genoa Italy, 173–182. https://doi.org/10.1145/3340631.3394844

[45] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[46] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. (2020). arXiv:2010.09337 [stat.ML]

[47] Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2020. Generating post hoc review-based natural language justifications for recommender systems. *User Modeling and User-Adapted Interaction* (June 2020). https://doi.org/10.1007/s11257-020-09270-8

[48] Shabnam Najafian, Amra Delic, Marko Tkalčič, and Nava Tintarev. 2021. Factors Influencing Privacy Concern for Explanations of Group Recommendation. ACM UMAP 2021 ; Conference date: 21-06-2021 Through 25-06-2021.

[49] Shabnam Najafian, Oana Inel, and Nava Tintarev. 2020. Someone really wanted that song but it was not me! Evaluating Which Information to Disclose in Explanations for Group Recommendations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*. 85–86.

[50] Shabnam Najafian and Nava Tintarev. 2018. Generating Consensus Explanations for Group Recommendations: an exploratory study. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 245–250.

[51] Sidra Naveed, Benedikt Loepp, and Jürgen Ziegler. 2020. On the Use of Feature-Based Collaborative Explanations: An Empirical Comparison of Explanation Styles. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 226–232. https://doi.org/10.1145/3386392.3399303

[52] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (2017), 393–444.

[53] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. 1085–1088. https://doi.org/10.1145/1357054.1357222

[54] Tim O'Reilly. 2013. Open data and algorithmic regulation. *Beyond transparency: Open data and the future of civic innovation* 21 (2013), 289–300.

[55] Denis Parra, Peter Brusilovsky, and Christoph Trattner. 2014. See what you want to see: visual user-driven approach for hybrid recommendation. In *Proceedings of the 19th international conference on Intelligent User Interfaces*. ACM, Haifa Israel, 235–240. https://doi.org/10.1145/2557500.2557542

[56] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in Explainable AI. arXiv:1810.00184 [cs.AI]

[57] Lara Quijano-Sanchez, Christian Sauer, Juan A Recio-Garcia, and Belen Diaz-Agudo. 2017. Make it personal: a social explanation system applied to group recommendations. *Expert Systems with Applications* 76 (2017), 36–48.

[58] Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. 2019. A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association* 27 (11 2019). https://doi.org/10.1093/jamia/ocz192

[59] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). arXiv:1602.04938 http://arxiv.org/abs/1602.04938

[60] Johannes Schneider and Joshua Handali. 2019. Personalized explanation in machine learning. *CoRR* abs/1901.00770 (2019). arXiv:1901.00770 http://arxiv.org/abs/1901.00770

[61] Amartya Sen. 1993. *The Quality of Life*. Clarendon Press, Oxford. http://ukcatalogue.oup.com/product/9780198287971.do?keyword=The+Quality+of+Life&sortby=bestMatches Italian translation, Feltrinelli, 1997. Jointly edited with Martha Nussbaum.

[62] H. Sharp, A. Finkelstein, and G. Galal. 1999. Stakeholder identification in the requirements engineering process. In *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99*. 387–391. https://doi.org/10.1109/DEXA.1999.795198

[63] Wanita Sherchan, Seng W Loke, and Shonali Krishnaswamy. 2008. Explanation-aware service selection: rationale and reputation. *Service Oriented Computing and Applications* 2, 4 (2008), 203–218.

[64] Avital Shulner Tal, Khuyagbaatar Batsuren, Veronika Bogina, Fausto Giunchiglia, Alan Hartman, Styliani Kleanthous Loizou, Tsvi Kuflik, and Jahna Otterbacher. 2019. "End to End" Towards a Framework for Reducing Biases and Promoting Transparency of Algorithmic Systems. In *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. IEEE, 1–6.

[65] T. Terano, M. Suzuki, T. Onoda, K. Uenishi, and T. Matsuura. 1989. CSES: an approach to integrating graphic, music and voice information into a user-friendly

interface. *In: International Workshop on Industrial Applications of Machine Intelligence and Vision* (1989), 572–581.

[66] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, Viet Man Le, Ralph Samer, and Martin Stettinger. 2019. Towards Social Choice-based Explanations in Group Recommender Systems. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, 13–21.

[67] Chun-Hua Tsai and Peter Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 391–396.

[68] Chun-Hua Tsai and Peter Brusilovsky. 2019. Exploring Social Recommendations with Visual Diversity-Promoting Interfaces. *ACM Trans. Interact. Intell. Syst.* 10, 1, Article 5 (Aug. 2019), 34 pages. https://doi.org/10.1145/3231465

[69] Abraham Wyner, Matthew Olson, Justin Bleich, and David Mease. 2015. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. *Journal of Machine Learning Research* 18 (04 2015).

[70] Karen Yeung. 2018. Algorithmic regulation: a critical interrogation. *Regulation & Governance* 12, 4 (2018), 505–523.

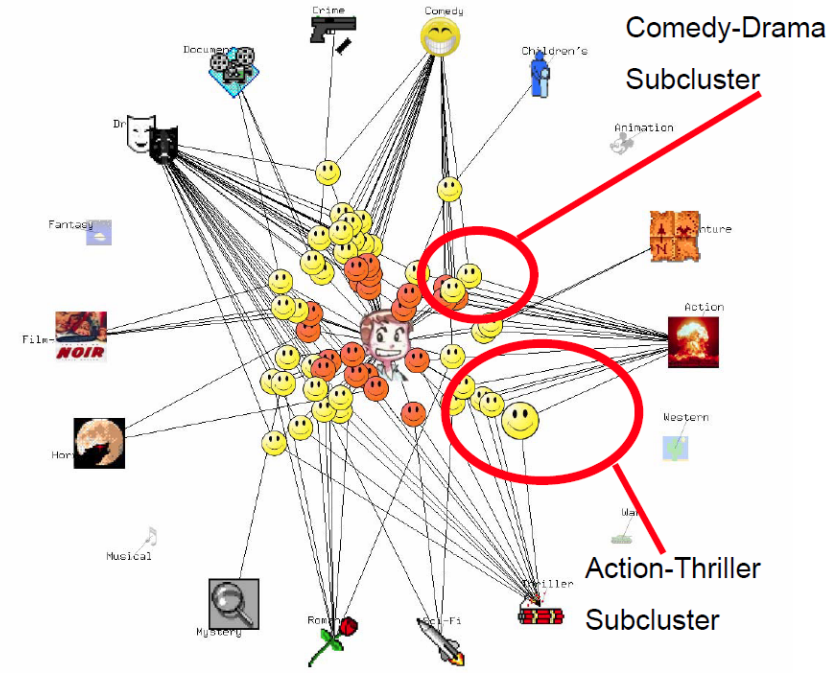
Appendix A EXPLANATION STYLES

Table 1: Explanation styles

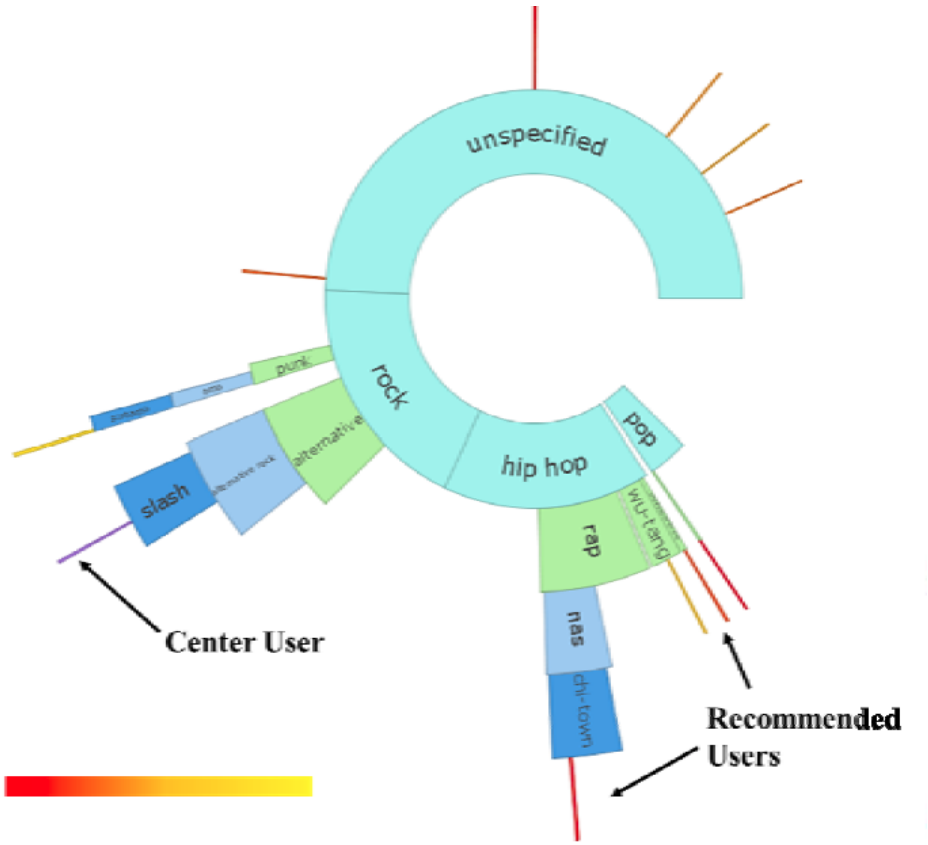
Explanation Style	Purpose	Interpretation	Context	Presentation	Stakeholder	Domain
Inputs explanations [36]	Why	Local	Individual	Textual	User	-
Output Explanations [36]	How	Local	Individual	Textual	User	-
Certainty explanations [36]	How	Local	Individual	Textual	User	-
Why explanations [36]	Why	Local	Individual	Textual	User	-
Why not explanations [36]	Why	Local	Individual	Textual	User	-
What if explanations [36]	How	Local	Individual	Textual	User	-
When Explanations [36]	Why	Local	Individual	Textual	User/Developer	-
Aggregation explanations [17]	Why	Local	Individual	Textual	User	Hotel Recommendation
Summary Explanations [17]	Why	Local	Individual	Textual	User	Hotel Recommendation
Review Explanations [17]	Why	Local	Individual	Textual	User	Hotel Recommendation
Case-Based Explanations [10]	Why	Local	Individual	Textual	User	Clinical Recommendation
Contextual Explanations [10]	Why	Local	Individual	Textual	User	Clinical Recommendation
Contrastive Explanations [10]	Why	Local	Individual	Textual	User	Clinical Recommendation
Counterfactual [10]	How	Global	Individual	Textual	User	Clinical Recommendation
Everyday Explanations [10]	Why	Global	Individual	Textual	User	Clinical Recommendation
Scientific Explanations [10]	Why	Global	Individual	Textual	Expert User	Clinical Recommendation
Simulation-Based[10]	How	Global	Individual	Textual	User	Clinical Recommendation
Statistical Explanations [10]	Why	Global	Individual	Textual	Expert User	Clinical Recommendation
Trace based [10]	How	Global	Individual	Textual	Expert User	Clinical Recommendation
Neighbour Rating [34]	Why	Global	Individual	Textual	User	Movie Recommendation
Profile-Based [34]	Why	Local	Individual	Textual	User	Movie Recommendation
Profile-Based [34]	Why	Local	Individual	Textual	User	Movie Recommendation
Input Influence-Based[12]	How	Global	Individual	Textual	User	Legal Recommendation
Demographic-based [12]	How	Global	Individual	Textual	User	Legal Recommendation
Sensitivity-based [12]	How	Local	Individual	Textual	User	Legal Recommendation
Case-based [12]	Why	Local	Individual	Textual	User	Legal Recommendation
Input Influence-based [6]	How	Global	Individual	Textual	User	Financial
Demographic-based [6]	How	Global	Individual	Textual	User	Financial
Sensitivity-based [6]	How	Local	Individual	Textual	User	Financial
Case-based [6]	Why	Local	Individual	Textual	User	Financial
Feature attribution Explanation [60]	How	Global	Individual	Textual/Visual	User	-
Example-based Explanation [60]	How	Global	Individual	Textual/Visual	User	-
Model internals Explanation [60]	How	Global	Individual	Textual/Visual	User	-
Surrogate model Explanation [60]	How	Global	Individual	Textual/Visual	User	-
Contrastive Explanation [31]	Why	Local	Individual	Textual	User	Clinical Recommendation
General Explanation [31]	Why	Local	Individual	Textual	User	Clinical Recommendation
Truthful Explanation [31]	Why	Global	Individual	Textual	User	Clinical Recommendation
Thorough Explanation [31]	How	Global	Individual	Textual	User	Clinical Recommendation
Content-based explanations [51]	Why	Local	Individual	Textual	User	Item Recommendation
Collaborative explanations [51]	Why	Local	Individual	Textual	User	Item Recommendation
Personalized Social Individual [57]	How	Global	Group	Textual	End user	Movie
Social Choice-based [22]	How	Global	Group	Textual	End user	Tourism
Social Choice-based [49]	How	Global	Group	Textual	End user	Music
Social Choice-based [66]	How	Global	Group	Textual	End user	Restaurant
Context-aware Group [48]	How	Global	Group	Textual	End user	Tourism
Vocal explanations[65]	Why	Local	Individual	Textual	User	-
Conversational RS [20]	Why	Global	Local	Textual	User	Movie/Music/Book
PeerChooser [53]	How	Global	Individual	Visual - Appx. B.1	User	Movie
SFViz [15]	How	Global	Individual	Visual - Appx. B.2	User	Music
Scatter Plot [68]	How	Global	Individual	Visual - Appx. B.3	User	Academic paper
Bar diagram [44]	Why	Local	Individual	Visual - Appx. B.6.a	User	Music
INTEREST [41]	Why	Local	Individual	Visual - Appx. B.6.b	User	Apartment
Venn diagram [55]	Why	Global	Individual	Visual - Appx. B.4.a	User	Academic paper
Venn diagram [25]	Why	Local	Individual	Visual - Appx. B.4.b	User	Restaurants
Dendrogram [26]	Why	Local	Individual	Visual - Appx. B.5	User	Music
Pathways between columns [25]	How	Local	Individual	Visual - Appx. B.7.a	User	Restaurants
Concentric circles paths interface [25]	How	Local	Individual	Visual - Appx. B.7.b	User	Restaurants

Appendix B VISUAL EXPLANATION

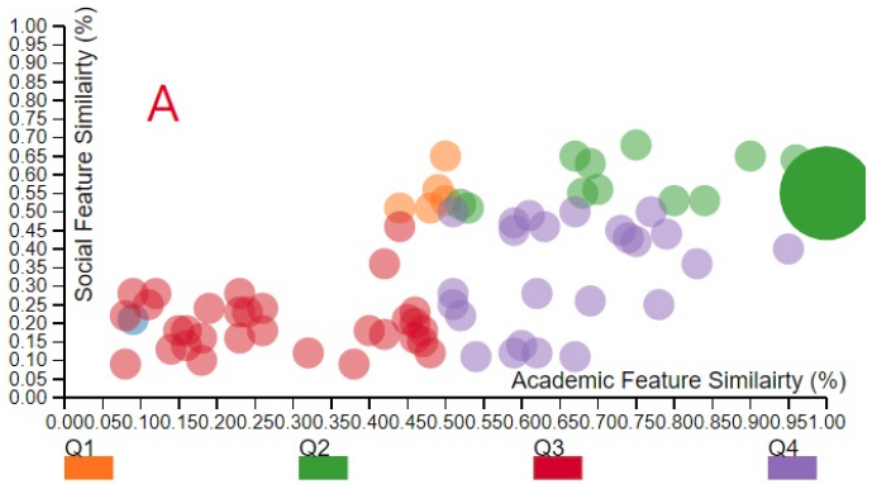
B.1 PeerChooser [53]



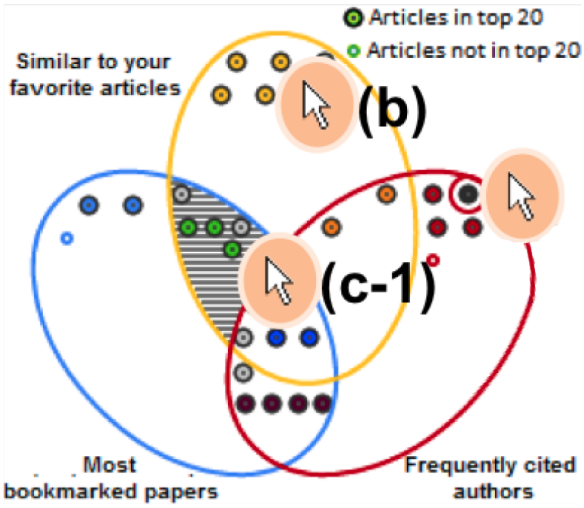
B.2 SFViz [15]



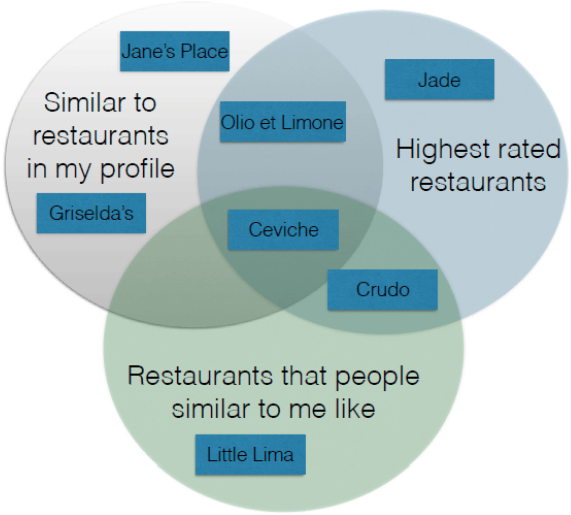
B.3 Scatter Plot [68]



B.4 Venn Diagram

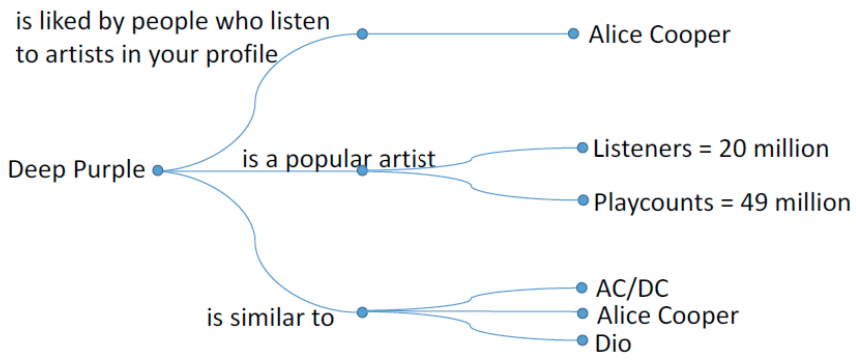


(a) [55]



(b) [25]

B.5 Dendrogram [26]



B.7 Pathways between columns and Concentric circles paths interface [25]