Panorama des outils de visualisation pour l'explicabilité en apprentissage profond pour le traitement automatique de la langue

Présenté par Alexis Delaforge

Alexis Delaforge[1,3]
Jérôme Azé[1]
Sandra Bringay[1,2]
Arnaud Sallaberry[1,2]
Maximilien Servajean[1,2]

[1]LIRMM, Université de Montpellier
[2]AMIS, Université Paul-Valéry
[3]Zortify Labs, Zortify

INTERPRETABILITÉ & EXPLICABILITÉ

Interprétabilité = Transparence + Explicabilité

Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3) :31–57, 2018.

INTERPRETABILITÉ & EXPLICABILITÉ

Interprétabilité = Transparence + Explicabilité

Transparence : facilité avec laquelle un humain peut comprendre et reproduire le fonctionnement d'un modèle, indépendamment d'une prédiction.

Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3) :31–57, 2018.

INTERPRETABILITÉ & EXPLICABILITÉ

Interprétabilité = Transparence + Explicabilité

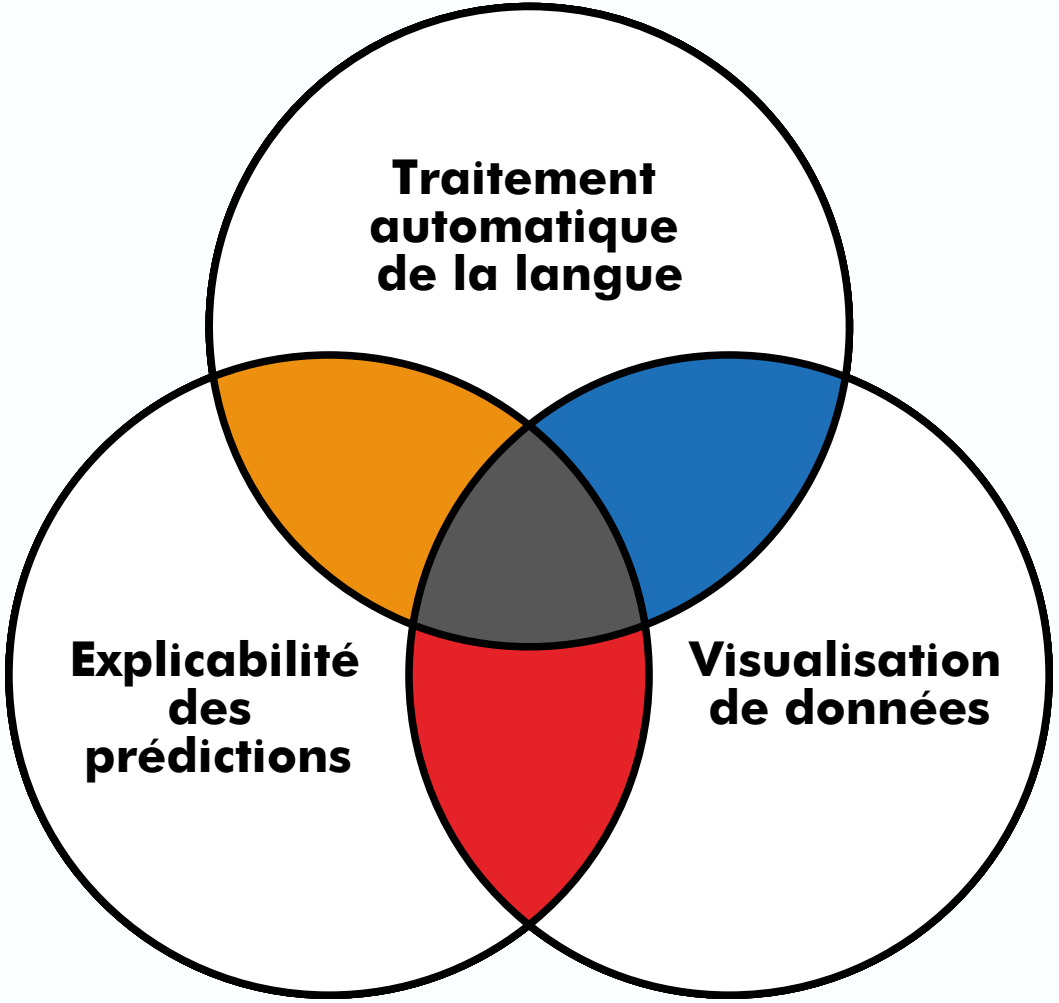
L’explication d’une prédiction, est faite lorsque des indicateurs, issus ou non du fonctionnement d’un modèle, sont utilisés pour expliquer sa décision.

Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3) :31–57, 2018.

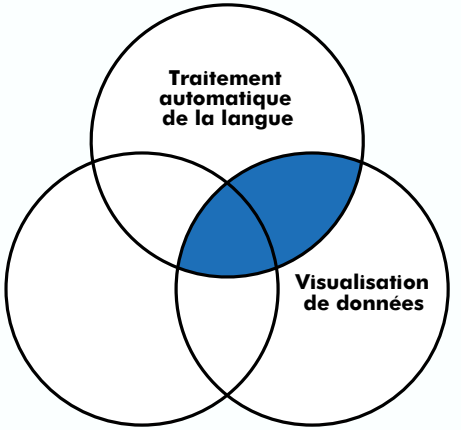
INTERPRETABILITÉ & EXPLICABILITÉ

Pas de consensus :

- Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3) :31–57, 2018.
- Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. Flexible and context-specific ai explainability : a multidisciplinary approach. SSRN, 2020.
- …


Traitement automatique de la langue
Explicabilité des prédictions
Visualisation de données

TAL & VIS



Alharbi, Mohammad, and Robert S. Laramee. 2019. SoS TextVis: An Extended Survey of Surveys on Text Visualization. Computers 8, no. 1: 17. https://doi.org/10.3390/computers8010017

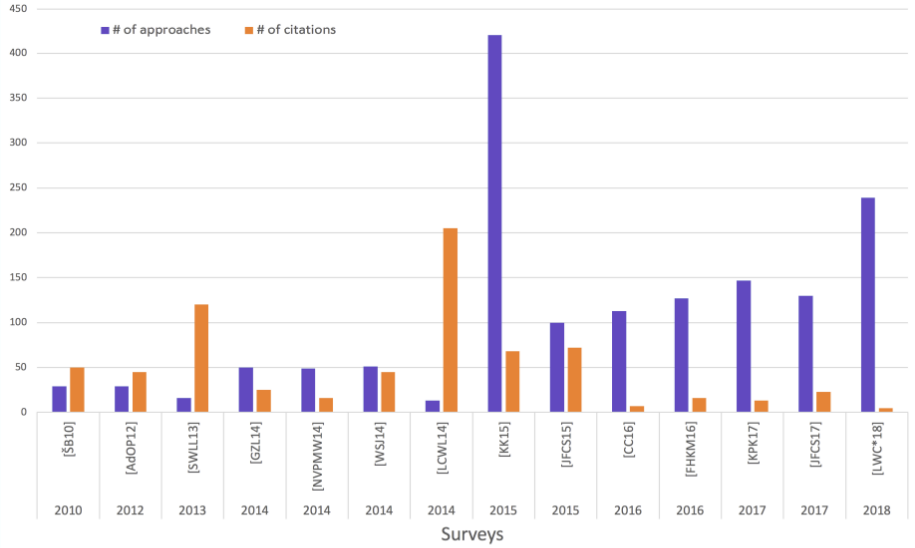


Figure 1. The text visualization surveys from 2010 to 2018. Blue bars indicate the number of methods reviewed in each survey. Orange bars show the number of citations each survey attracts. In term of the number of surveys, 2014 dominates with four surveys. However, with the respect to the number of techniques, surveys from 2015 review 480 methods collectively.

TAL & VIS

Alharbi, Mohammad, and Robert S. Laramee. 2019. SoS TextVis: An Extended Survey of Surveys on Text Visualization. Computers 8, no. 1: 17. https://doi.org/10.3390/computers8010017

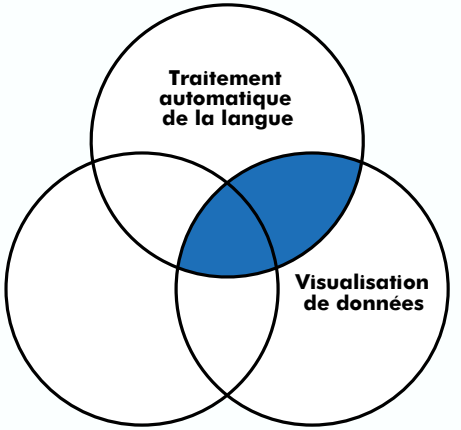
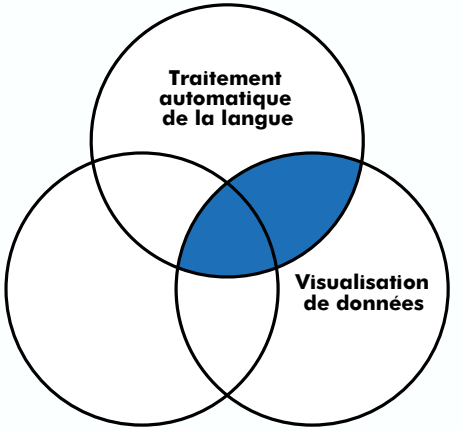


Table 3. Summary of the future challenges reported in our collection of surveys. This list contains the common challenges among the text visualization surveys.

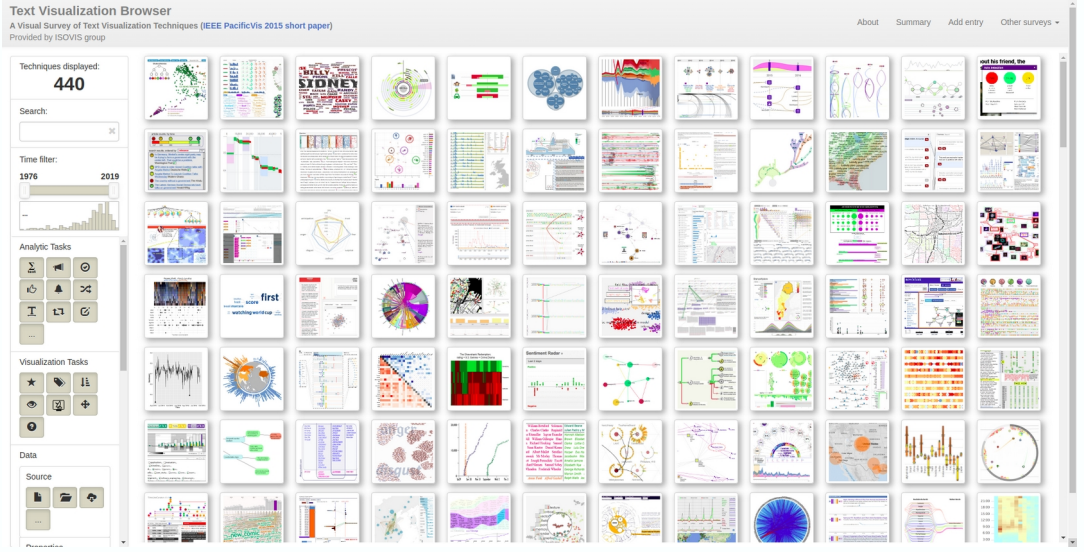
Future Challenges	SoS [7]	[5]	[21]	[15]	[22]	[23]	[26]	[16]	[4]	[27]	[25]	[24]	[9]	[10]	[11]	Total
Scalability	✓	✓	✓	✓	✓			✓			✓	✓			✓	9
Lacking in-depth/effective quantitative or qualitative evaluation	✓	✓	✓	✓	✓		✓			✓	✓			✓		9
Adopting advanced linguistics techniques		✓	✓	✓	✓	✓	✓				✓				✓	8
Natural language ambiguity and uncertainty	✓			✓				✓		✓	✓			✓		6
Lacking user interactivity that support the analysis tasks	✓		✓				✓				✓				✓	5
Designing general models for different tasks (versatility)	✓	✓		✓											✓	4
Multidisciplinary framework										✓	✓			✓		3
Lacking of cognitive and/or psychological analysis		✓	✓													2
Lacking well-defined general concepts						✓	✓									2

TAL & VIS

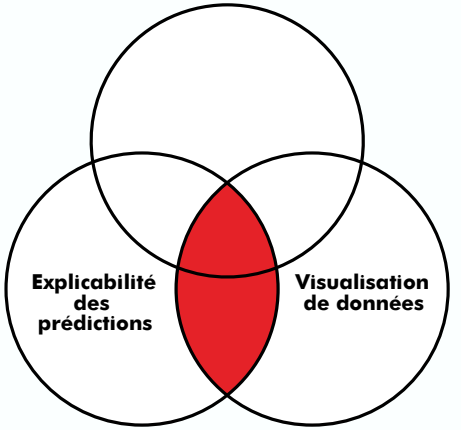


K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," 2015 IEEE Pacific Visualization Symposium (PacificVis), Hangzhou, China, 2015, pp. 117-121, doi: 10.1109/PACIFICVIS.2015.7156366.

https://textvis.lnu.se/ jusqu'à 2019



XAI & VIS



Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning : An interrogative survey for the next frontiers. IEEE TVCG, 25(8) :2674–2693, 2019.

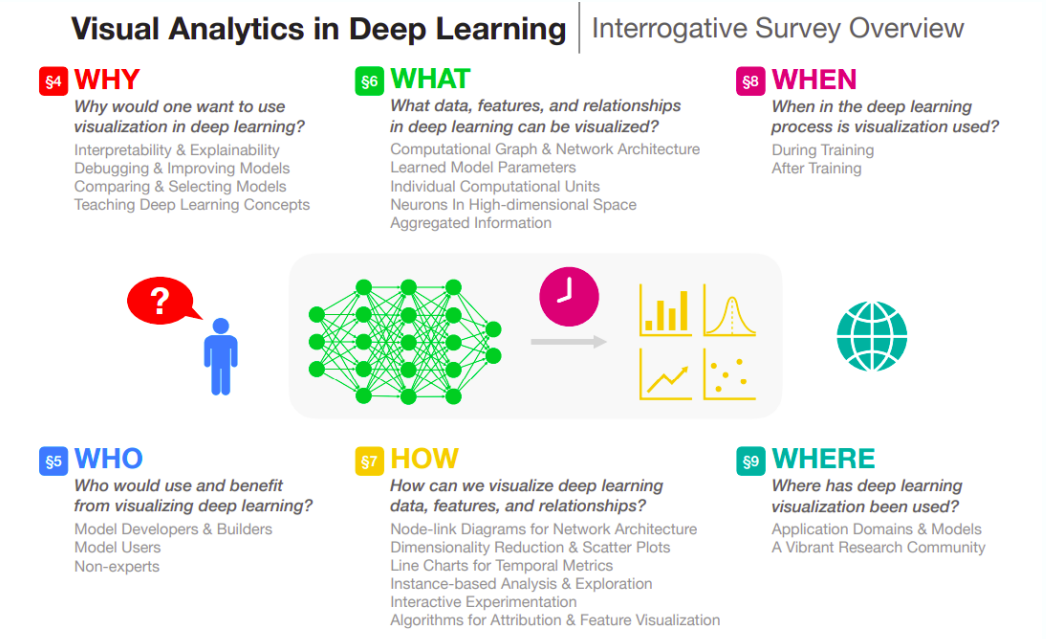
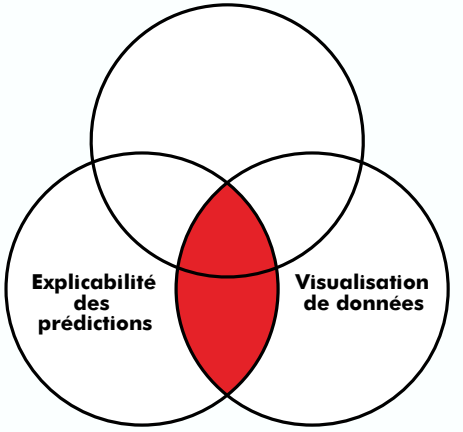


Fig. 1. A visual overview of our interrogative survey, and how each of the six questions, "Why, Who, What, How, When, and Where," relate to one another. Each question corresponds to one section of this survey, indicated by the numbered tag, near each question title. Each section lists its major subsections discussed in the survey.

XAI & VIS

Biagio La Rosa, Graziano Blasilli, Romain Bourqui, David Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini. State of the art of visual analytics for explainable deep learning. Computer Graphics Forum, 42(1) :319–355, 2023.



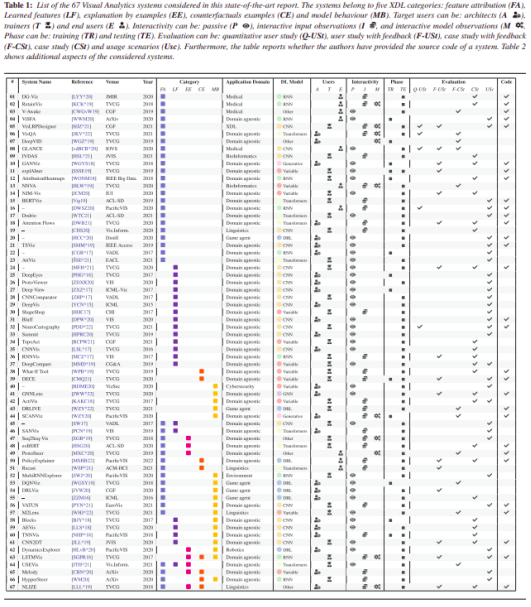
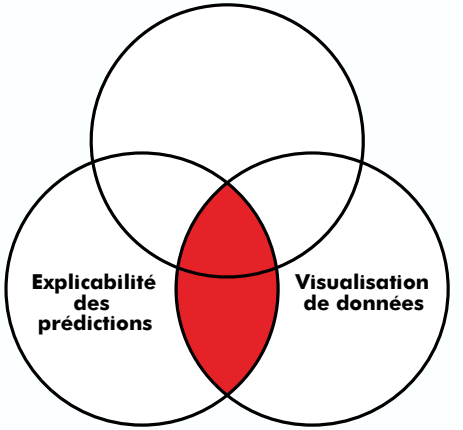


Table 1: *List of the 67 Visual Analytics systems considered in this state-of-the-art report. The systems belong to five XDL categories: feature attribution (**FA**), Learned features (**LF**), explanation by examples (**EE**), counterfactuals examples (**CE**) and model behaviour (**MB**). Target users can be: architects (**A**), trainers (**T**), and end users (**E**). Interactivity can be: passive (**P**), interactive input observations (**I**), and interactive model observations (**M**). Phase can be: training (**TR**) and testing (**TE**). Evaluation can be: quantitative user study (**Q-US**), user study with feedback (**F-US**), case study with feedback (**F-CS**), case study (**CS**) and usage scenarios (**Use**). Furthermore, the table reports whether the authors have provided the source code of a system. Table 2 shows additional aspects of the considered systems.*

XAI & VIS



Biagio La Rosa, Graziano Blasilli, Romain Bourqui, David Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini. State of the art of visual analytics for explainable deep learning. Computer Graphics Forum, 42(1) :319–355, 2023.

- Feature attribution
- Learned features
- Explanations by examples
- Counterfactuals
- Model Behaviour

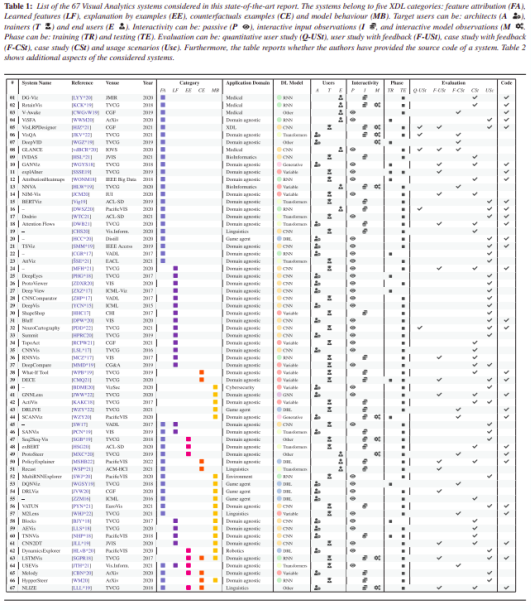
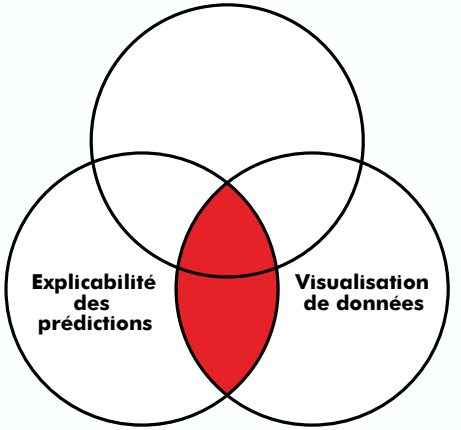


Table 1: *List of the 67 Visual Analytics systems considered in this state-of-the-art report. The systems belong to five XDL categories: feature attribution (**FA**), Learned features (**LF**), explanation by examples (**EE**), counterfactuals examples (**CE**) and model behaviour (**MB**). Target users can be: architects (**A** 👷), trainers (**T** 👩‍🏫), and end users (**E** 👩). Interactivity can be: passive (**P** 👁), interactive input observations (**I** 🖱), and interactive model observations (**M** ⚙). Phase can be: training (**TR**) and testing (**TE**). Evaluation can be: quantitative user study (**Q-US**), user study with feedback (**F-US**), case study with feedback (**F-CS**), case study (**CS**) and usage scenarios (**Use**). Furthermore, the table reports whether the authors have provided the source code of a system. Table 2 shows additional aspects of the considered systems.*

XAI & VIS



Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. Information Visualization, 19(3) :207–233, 2020.

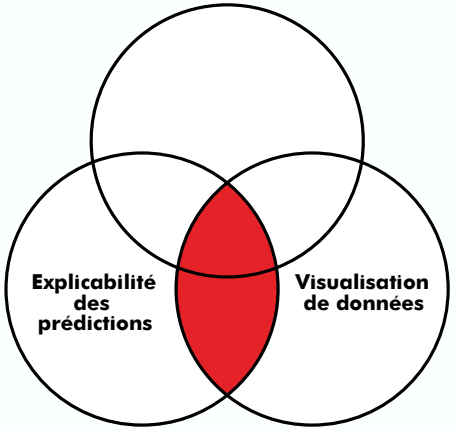
Table 6. Top eight terms for each of the 10 topics generated with latent Dirichlet allocation (LDA) from the collection of the papers of the survey papers.

Topic 1 *NNs for image applications*	**Topic 2** *IML classifiers' training*	**Topic 3** *DR and projections visualization*	**Topic 4** *CNNs (for image applications)*	**Topic 5** *Use of features in models' predictions (for regression)*
Image	Model	User	Part	Feature
Network	User	Visualization	Object	Model
Layer	Learning	Dimension	CNN	Function
Feature	Training	Point	Image	Point
Input	Machine	Analysis	Node	Prediction
Model	Classifier	Visual	Pattern	Value
Neural	System	Projection	Layer	Vector
Deep	Machine learning	Uncertainty	Category	Regression

Topic 6 *Subspace visualization for clustering and classification*	**Topic 7** *Users' feedback in ML*	**Topic 8** *Clustering and algorithm use in feature subset selection*	**Topic 9** *Model and clustering visualization for time-series data*	**Topic 10** *NNs for text applications*
Cluster	User	Feature	Model	Model
Visualization	Topic	Cluster	Cluster	Image
Class	Feature	Selection	User	Network
Clustering	Learning	Clustering	Visualization	Learning
Subspace	Document	Algorithm	Time	Word
Model	Graph	Subset	Visual	Neural
Analysis	System	Search	Analysis	State
Classification	Participant	Learning	Clustering	Training

NN: neural network; IML: interactive machine learning; DR: dimensionality reduction; CNN: convolutional neural networks; ML: machine learning.
The suggested topic titles are displayed in italics. Each topic is represented by one specific color.

XAI & VIS



Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. Information Visualization, 19(3) :207–233, 2020.

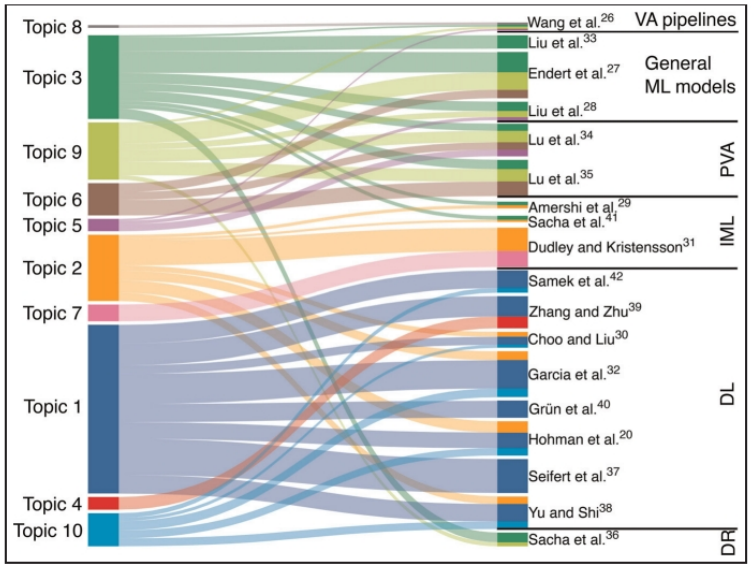
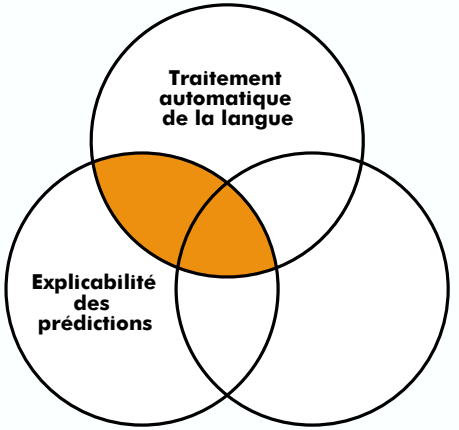
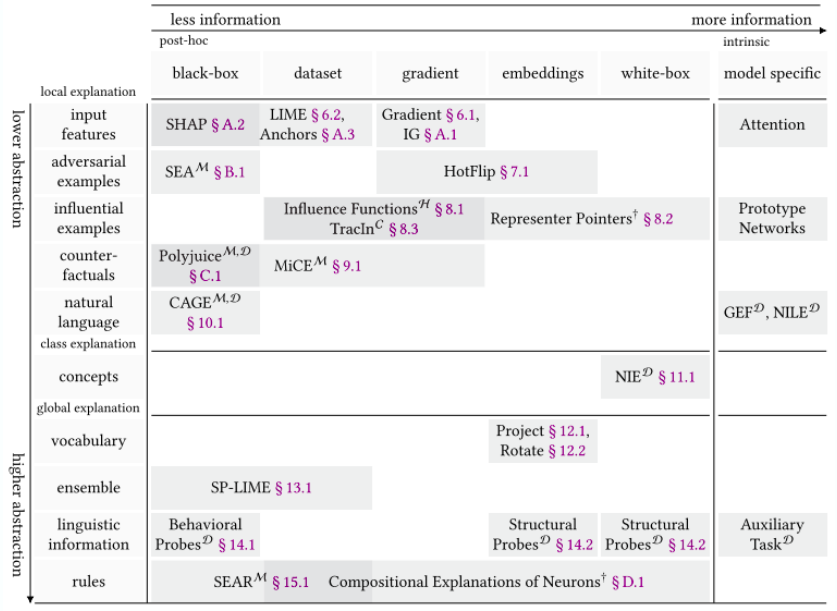


Figure 4. Explicit connections between the survey papers and the LDA topics extracted automatically from the individual papers under each survey. The width of the connections shows the weight of each topic in each survey paper. On the right side, the survey papers are categorized and sorted according to Table 5. On the left side, the topics are ordered to minimize the number of crossing lines.

TAL & XAI



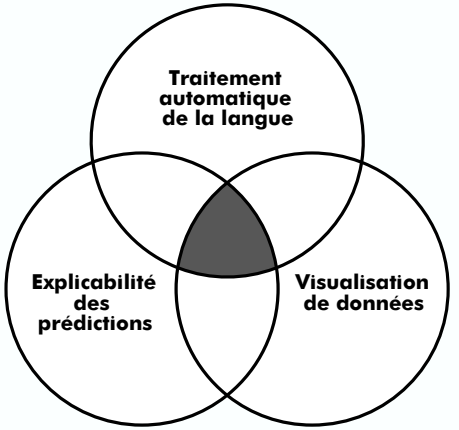
Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp : A survey. ACM Computing Surveys, 55(8) :1–42, 2022.



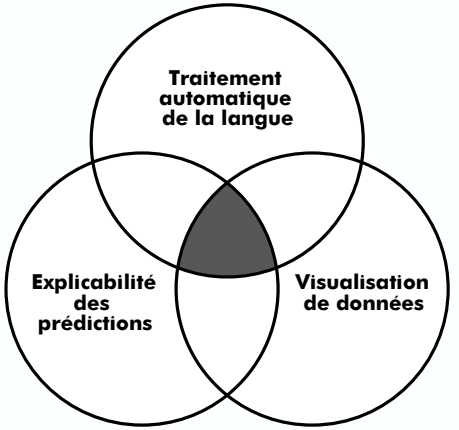
Rows describe how the explanation is communicated, while columns describe what information is used to produce the explanation. The order of both rows and columns indicates a level of abstraction and amount of information, respectively. However, this order is only approximate.
Furthermore, because this survey focuses on *post-hoc* methods, the *intrinsic* section of this table is incomplete and merely meant to provide a few comparative examples. The specifc *intrinsic* methods shown are: *Attention* [9], *GEF* [69], *NILE* [63]. *Prototype Networks* and *Auxiliary Task* refer to types of models.
^{C}: Depends on checkpoints during training. $^{\mathcal{D}}$: Depends on supplementary dataset. $^{\mathcal{H}}$: Depends on second-order derivative. $^{\mathcal{M}}$: Depends on the supplementary model. †: Depends only on the dataset and white-box access.

TAL, VIS & XAI

Il n’existe pas à ce jour de travaux s’intéressant aux techniques de visualisation utilisées pour expliquer les prédictions dans le domaine du traitement automatique de la langue.



TAL, VIS & XAI



Il n’existe pas à ce jour de travaux s’intéressant aux techniques de visualisation utilisées pour expliquer les prédictions dans le domaine du traitement automatique de la langue.

Types de réseaux :

- Perceptrons multicouches et réseaux convolutifs
- Réseaux de neurones récurrents
- Réseaux auto-attentifs
- Méthodes agnostiques

PERCEPTRONS MULTICOUCHES ET RÉSEAUX CONVOLUTIFS

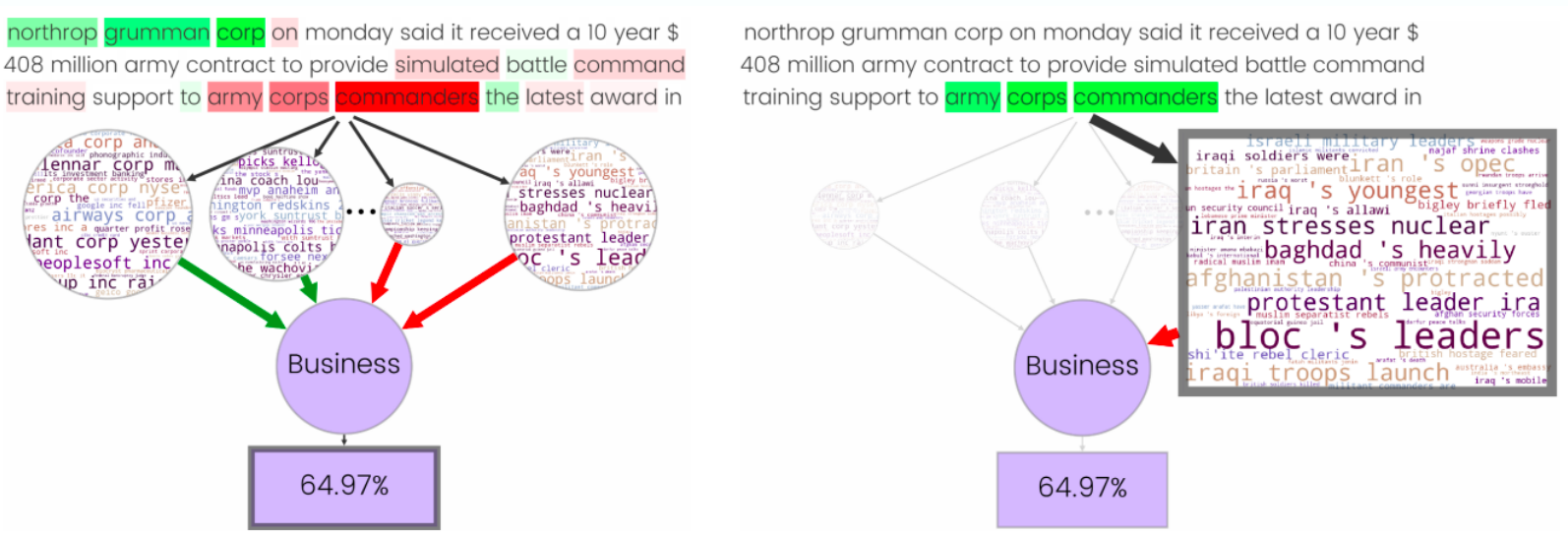


Figure 4: (Left) Graphical DAX for the (correct) output class "Business" (determined with 64.97% probability) by the CNN for text classification with AG-News, for the input article from AG-News at the top. (Right) The view of the DAX after clicking on the rightmost argument/word cloud. Arguments are visualised (by χ) as follows: *input words* with highlighting of a word given, on the left, by the sum of the (dialectical) strengths of all arguments representing that word and, on the right, by its corresponding argument's dialectical strength; *convolutional filters/word clouds* with sizes the corresponding arguments' strengths; and the most probable *class* and its probability. The green/red colours indicate, resp., support/attack (used on edges/relations from intermediate to output arguments, and on words from input to intermediate arguments).

Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. Dax : Deep argumentative explanation for neural networks. arXiv preprint :2012.05766, 2020.

PERCEPTRONS MULTICOUCHES ET RÉSEAUX CONVOLUTIFS

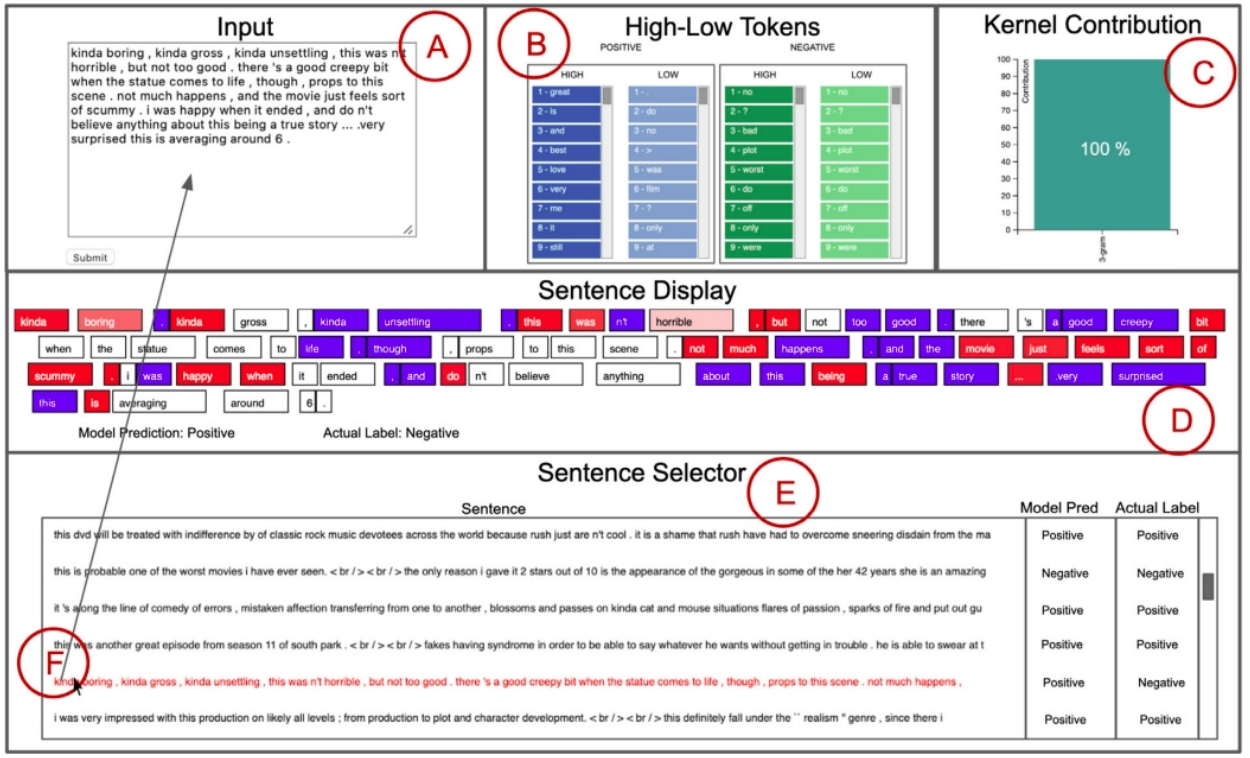


Fig. 6. Visual Analysis System.

Piyush Chawla, Subhashis Hazarika, and Han-Wei Shen. Token-wise sentiment decomposition for convnet : Visualizing a sentiment classifier. Visual Informatics, 4(2) :132–141, 2020.

PERCEPTRONS MULTICOUCHES ET RÉSEAUX CONVOLUTIFS

- Peu de travaux
- Les autres types de réseaux focalisent plus les recherches en explicabilité

RÉSEAUX DE NEURONES RÉCURRENTS

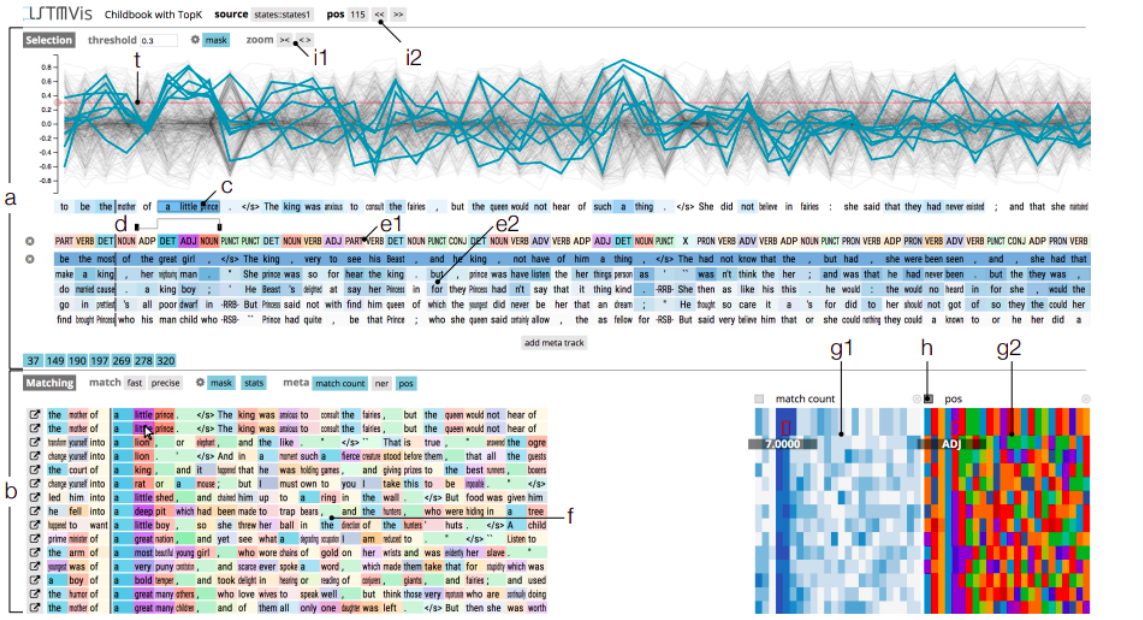


Fig. 1. The LSTMVIS user interface. The user interactively *selects* a range of text specifying a hypothesis about the model in the Select View (a). This range is then used to *match* similar hidden state patterns displayed in the Match View (b). The selection is made by specifying a start-stop range in the text (c) and an activation threshold (t) which leads to a selection of hidden states (blue lines). The start-stop range can be further constrained using the pattern plot (d). The meta-tracks below depict extra information per word position like POS (e1) or the top K predictions (e2). The tool can then *match* this selection with similar hidden state patterns in the data set of varying lengths (f), providing insight into the representations learned by the model. The match view additionally includes user-defined meta-data encoded as heatmaps (g1,g2). The color of one heatmap (g2) can be mapped (h) to the word matrix (f) which allows the user to see patterns that lead to further refinement of the selection hypothesis. Navigation aids provide convenience (i1, i2).

Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis : A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE TVCG, 24(1) :667–676, 2017.

RÉSEAUX DE NEURONES RÉCURRENTS

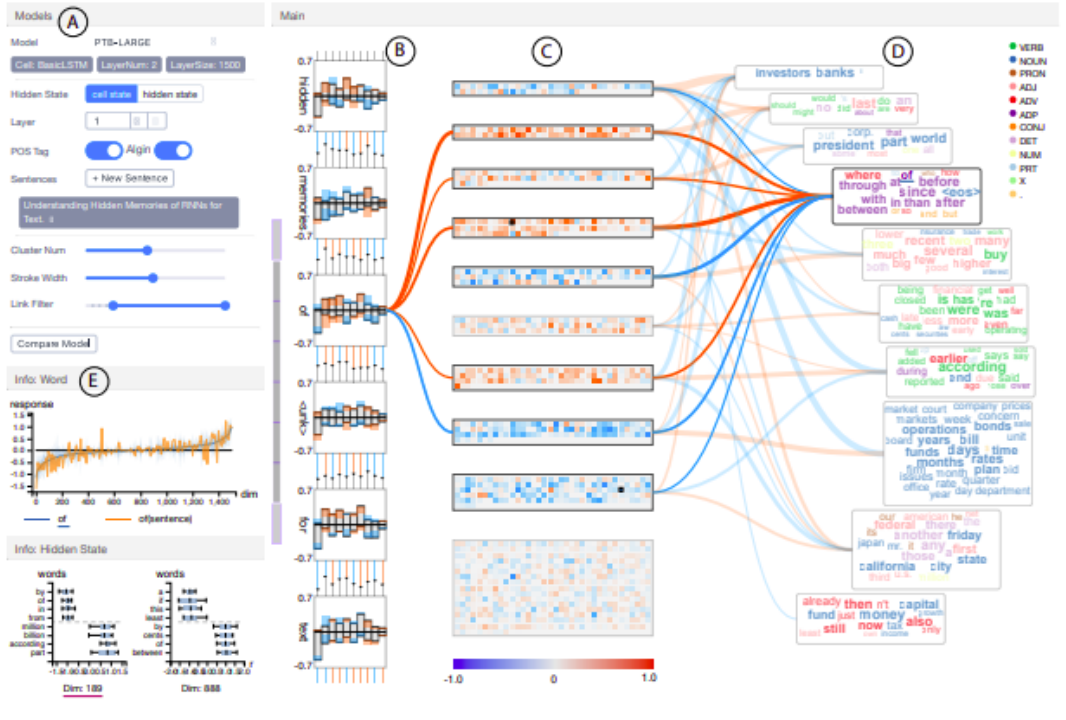


Figure 1: The interface of RNNVis. The control panel (A) shows parameter settings of an RNN and allows users to adjust visualization style. The main view (B-D) contains glyph-based sentence visualization (B), memory chips visualization for hidden state clusters (C), and word clouds visualization for word clusters (D). The detail view (E) shows the distributions of models responses to selected word "of" and interpretations of selected hidden units.

Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. Understanding hidden memories of recurrent neural networks. In 2017 IEEE VAST, pages 13–24. IEEE, 2017.

RÉSEAUX DE NEURONES RÉCURRENTS

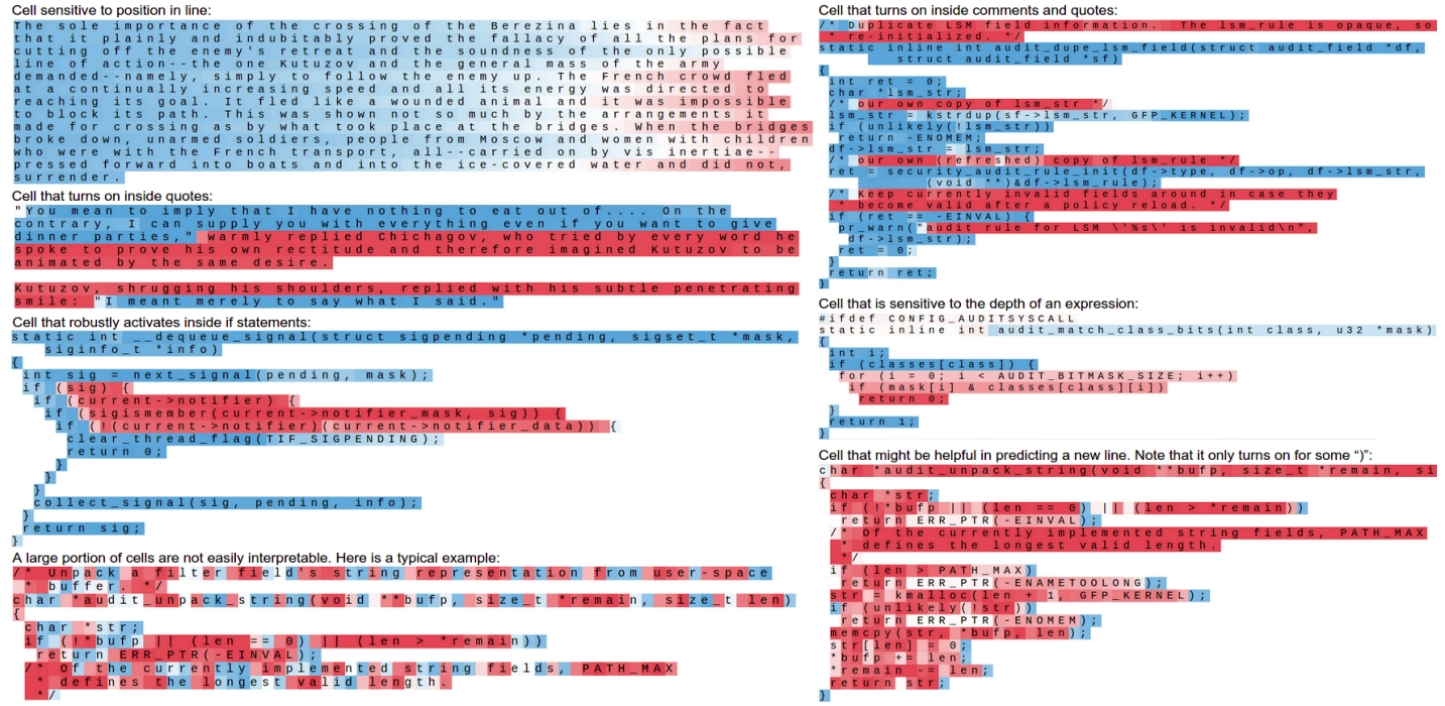


Figure 2: Several examples of cells with interpretable activations discovered in our best Linux Kernel and War and Peace LSTMs. Text color corresponds to $tanh(c)$, where -1 is red and +1 is blue.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. ArXiv preprint :1506.02078, 2015.

RÉSEAUX DE NEURONES RÉCURRENTS

true	predicted	N°	Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive
		1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
	--	6.	the master of disaster - it 's a piece of dreck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
--		10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .
	++	11.	ecks this one off your must-see list .
	-	12.	this is n't a `` friday '' worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork cliches ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is n't funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is n't really good enough to be in theaters .
	++	19.	a worthy entry into a very difficult genre .
++		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Figure 2: LRP heatmaps of exemplary test sentences, using as target class the *true* sentence class. Positive relevance is mapped to red, negative to blue, and the color intensity is normalized to the maximum absolute relevance per sentence. The true sentence class, and the classifier's predicted class, are indicated on the left.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 159–168, Copenhagen, Denmark, September 2017. ACL.

RÉSEAUX DE NEURONES RÉCURRENTS

- Activations, états cachés
- Carte de chaleur sur les mots
- Adaptation de mécanismes issu du traitement des images

RÉSEAUX DE NEURONES AUTO-ATTENTIFS

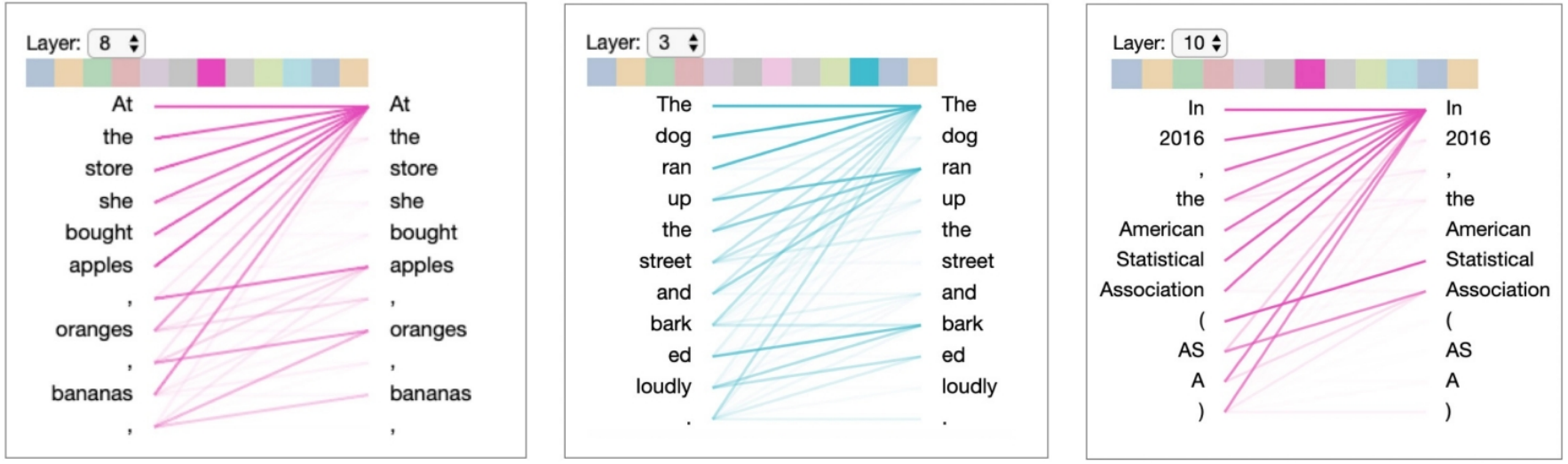


Figure 3: Examples of attention heads in GPT-2 that capture specific lexical patterns: list items (left); verbs (center); and acronyms (right). Similar patterns were observed in these attention heads for other inputs. Attention directed toward first token is likely null attention (Vig and Belinkov, 2019).

Jesse Vig. A multiscale visualization of attention in the transformer model. In 57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations (ACL), pages 37–42. ACL, 2019.

RÉSEAUX DE NEURONES AUTO-ATTENTIFS

Débat autour de l'attention :

- Jain et al. : l'attention n'est pas une méthode d'explication
- Wiegreffe et al. : les travaux que les expérimentations de Jain et al. ne permettent pas de soutenir leur théorie.
- Ethayarajh et al. : il existe une grande similarité entre les représentations des mots dans les transformeurs (similarité cosinus). Ceci produit des matrices d'attention peu informatives.

RÉSEAUX DE NEURONES AUTO-ATTENTIFS

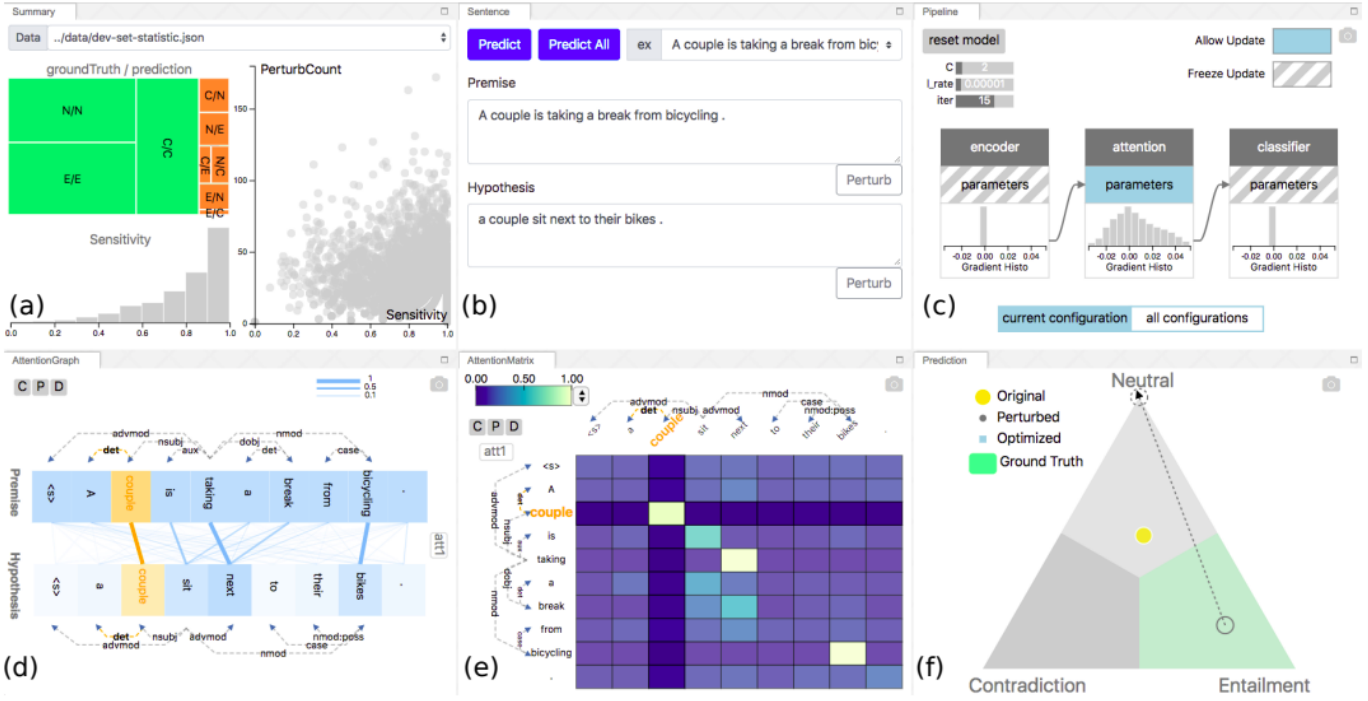


Fig. 1. The interface of the proposed system. During exploration, we can filter through a large number of sentence pairs summarized in (a). the current selected pair is displayed in (b). The model internal information (attention) is displayed in (d) and (e). The predicted probability (one of three labels: *neural*, *entailment*, *contradiction*) is shown in the barycentric coordinate in (f). Finally, the high-level model structure and the updates to the model are summarized in (c).

Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. Nlize : A perturbation-driven visual interrogation tool for analyzing and interpreting natural langage inference models. IEEE TVCG, 25(1) :651–660, 2018.

RÉSEAUX DE NEURONES AUTO-ATTENTIFS

- L’attention est centrale
- Nombreuses tâches de traitement automatique de langue

MÉTHODES AGNOSTIQUES

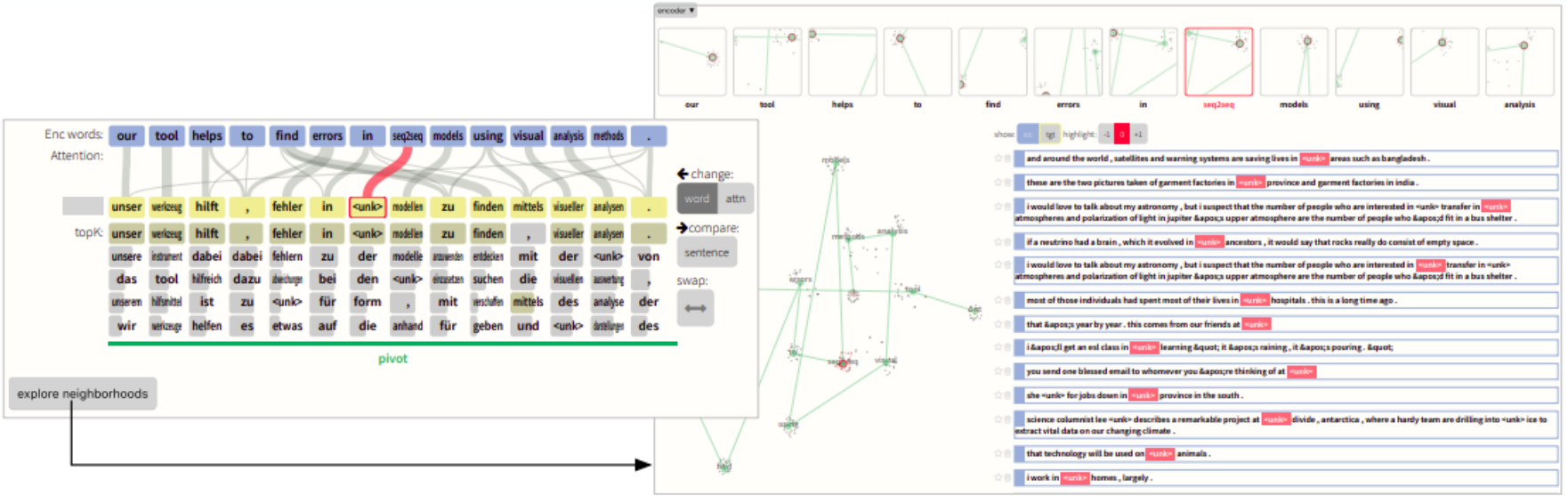


Fig. 1. Example of `Seq2Seq-Vis`. In the translation view (left), the source sequence *"our tool helps to find errors in seq2seq models using visual analysis methods."* is translated into a German sentence. The word *"seq2seq"* has correct attention between encoder and decoder (red highlight) but is not part of the language dictionary. When investigating the encoder neighborhoods (right), the user sees that *"seq2seq"* is close to other unknown words *"⟨unk⟩"*. The buttons enable user interactions for deeper analysis.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. Seq2seq-vis : A visual debugging tool for sequence-to-sequence models. IEEE TVCG, 25(1) :353–363, 2018.

MÉTHODES AGNOSTIQUES

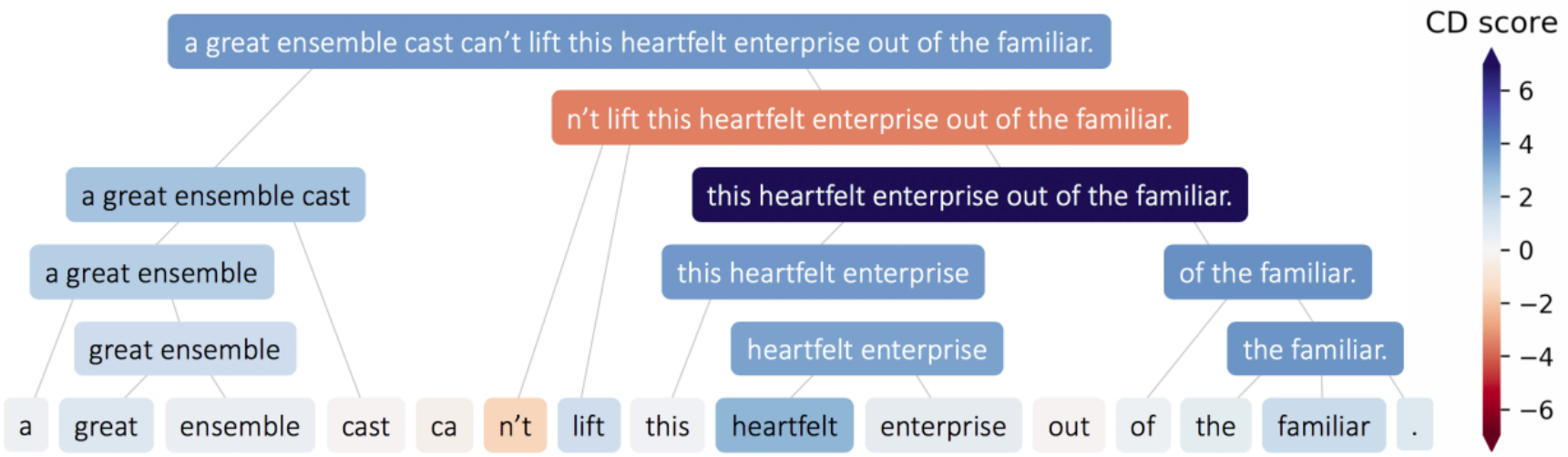


Figure 2: ACD interpretation of an LSTM predicting sentiment. Blue is positive sentiment, white is neutral, red is negative. The bottom row displays CD scores for individual words in the sentence. Higher rows display important phrases identified by ACD, along with their CD scores, converging to the model’s (incorrect) prediction in the top row. (Best viewed in color)

Chandan Singh, W. James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In International Conference on Learning Representations, 2019.

MÉTHODES AGNOSTIQUES

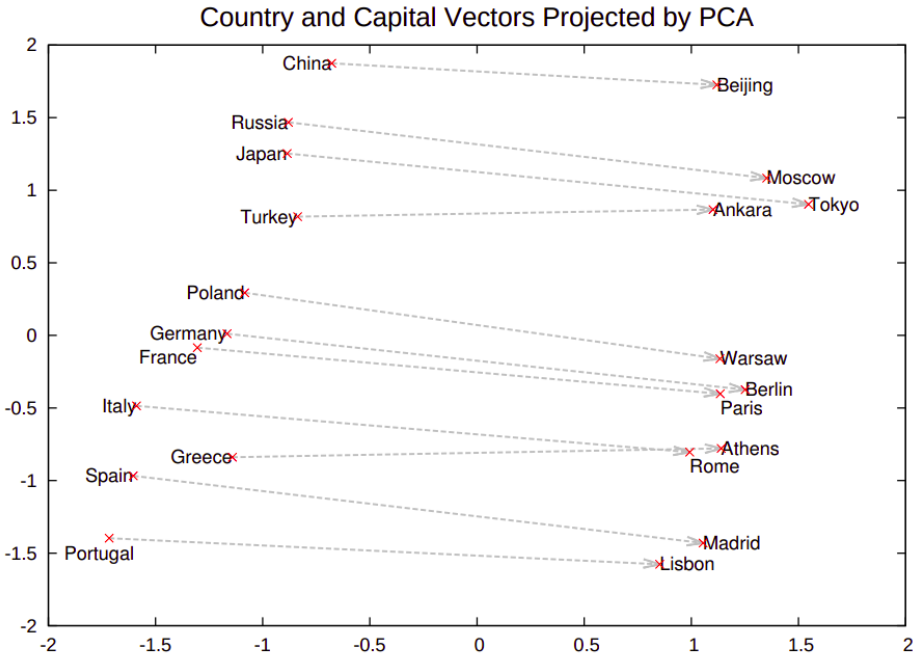


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc., 2013.

MÉTHODES AGNOSTIQUES

- Espace de représentation
- Explication hiérarchique plus récemment

VUE D'ENSEMBLE

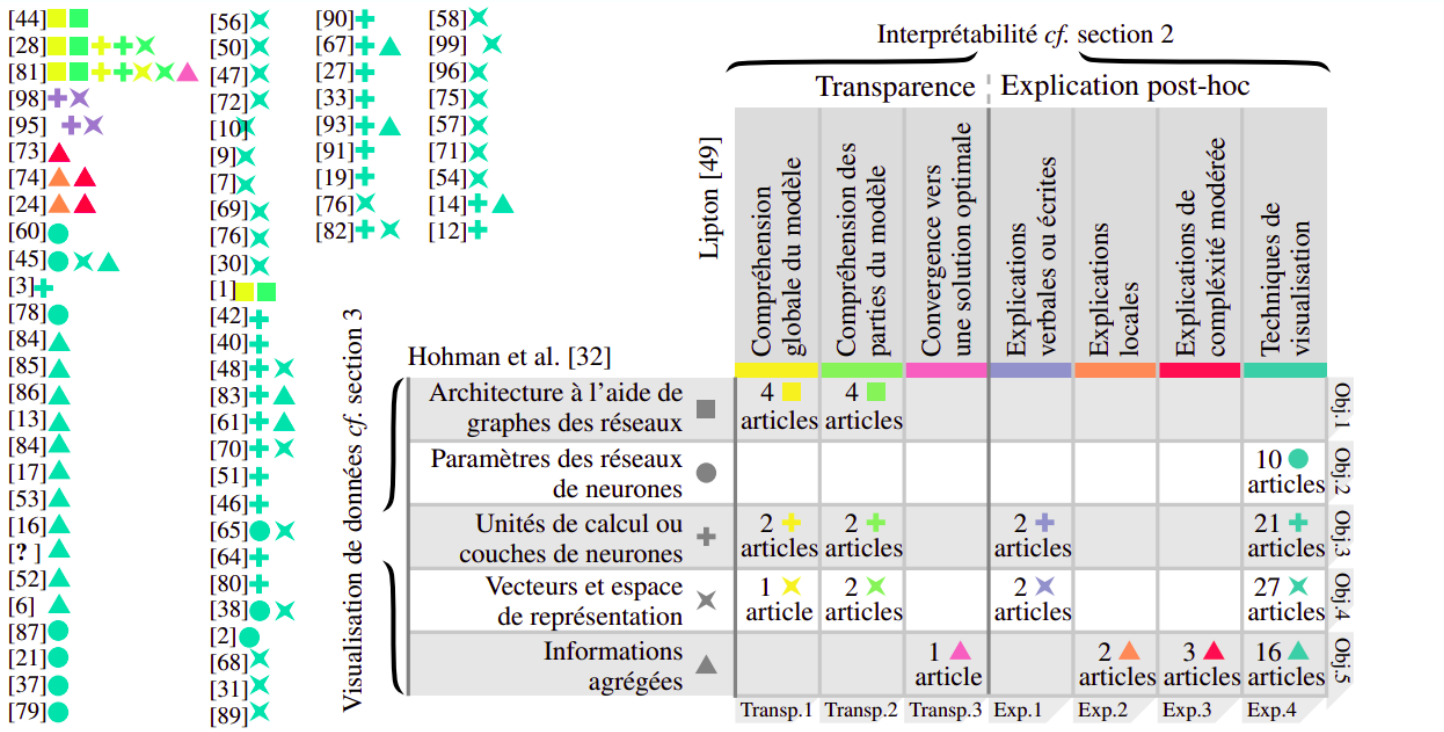


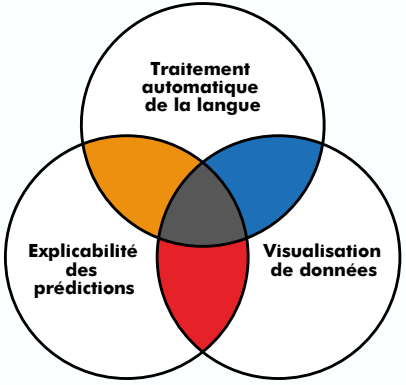
TABLE 1 – Méthodes d'interprétabilité présentées selon les classifications de Hohman et al. [32] et de Lipton [49]. Les catégories ne sont pas exhaustives car, selon le sujet, certaines catégories se confondent (*i.e.*, les unités de calcul et les vecteurs de représentation dans le cas d'utilisation de *RNN* ou réseaux auto-attentifs.)

CONCLUSION

1. Visualisation gagne en intérêt, incontournable
2. Disparité des recherches en fonction des modèles
3. Il n’existe pas de terminologie fixée
4. Méthodes les plus utilisées :
   - Carte de chaleur
   - Graphe bipartie
   - Matrice d’attention
   - Espace de représentation

PERSPECTIVES DE NOS TRAVAUX

- Panorama → État de l'art



- Notions sciences sociales:
  - Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.


LIRMM
UNIVERSITÉ MONTPELLIER
cnrs
dépasser les frontières
La Région
Occitanie
Pyrénées · Méditerranée
ICM
Institut régional du Cancer
Montpellier | Val d'Aurelle
SIRIC
MONTPELLIER CANCER
Site de Recherche Intégrée sur le Cancer

MERCI !

MILLER, T. (2019)

- 1. Explanations are contrastive — they are sought in response to particular counterfactual cases, which are termed foils in this paper. That is, people do not ask why event P happened, but rather why event P happened instead of some event Q. This has important social and computational consequences for explainable AI.
- 2. Explanation are selected (in a biased manner) — people rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation. However, this selection is influenced by certain cognitive biases.
- 3. Probabilities probably don't matter — while truth and likelihood are important in explanation and probabilities really do matter, referring to probabilities or statistical relationships in explanation is not as effective as referring to causes. The most likely explanation is not always the best explanation for a person, and importantly, using statistical generalisations to explain why events occur is unsatisfying, unless accompanied by an underlying causal explanation for the generalisation itself.
- 4. Explanations are social — they are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs.

ANDREAS MADSEN (2022)

- Terminology
- Synergy
- Helpful complex models