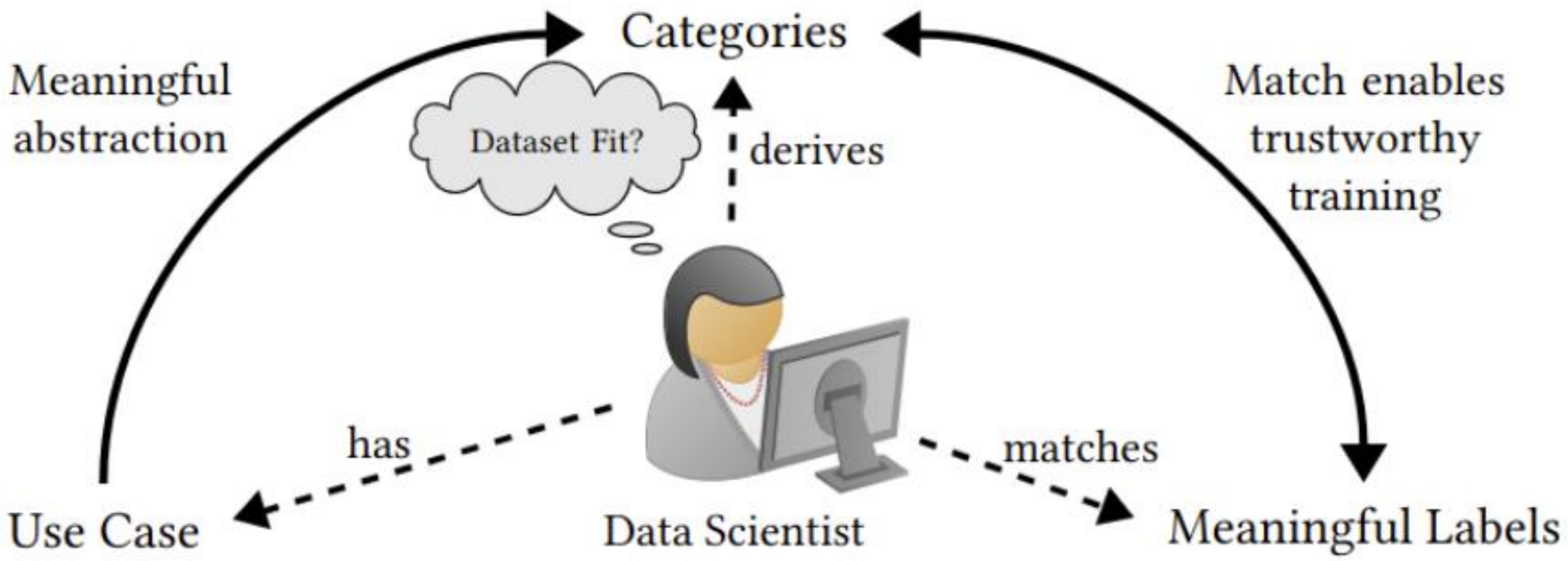
On the Relationship of Explainable Artificial Intelligence and Essential Complexity



Tim Barz-Cech
ELLIIT Focus Period
Digital Engineering Faculty, University of Potsdam & HMS Analytical Software GmbH 06.05.2025

Intro (1/4): How do we evaluate AI techniques?

Standard datasets (Benchmarks) Worse:
With standard metrics

Intro (1/4): How do we evaluate AI techniques?

HPI

Standard datasets (Benchmarks) Worse: With standard metrics

Intro (1/4): How do we evaluate AI techniques?

Standard datasets (Benchmarks) Worse:

With standard metrics

Intro (1/4): How do we evaluate AI techniques?

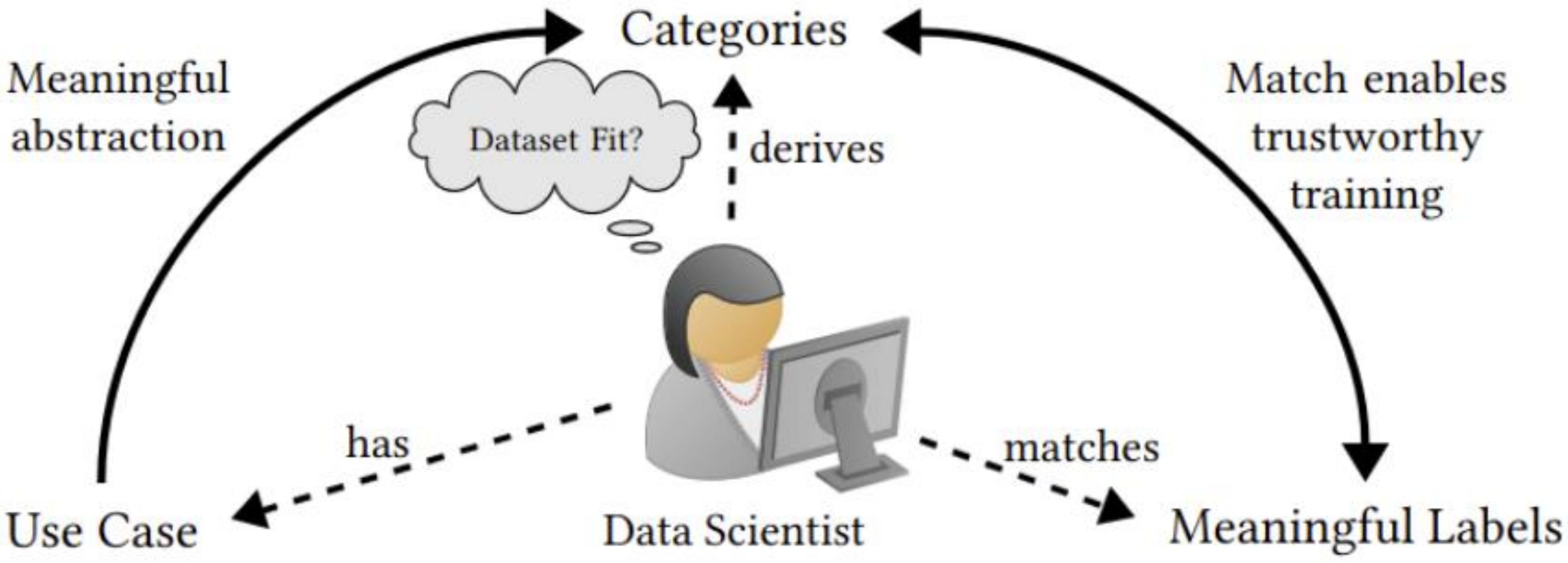
HPI

Standard datasets (Benchmarks) Worse: With standard metrics

Intro (2/4): XAI Supplants Quiddity of Data

Who explains the explanations [Rudin, 2019]?

Intro (3/4): Standardness Fogs Meaning

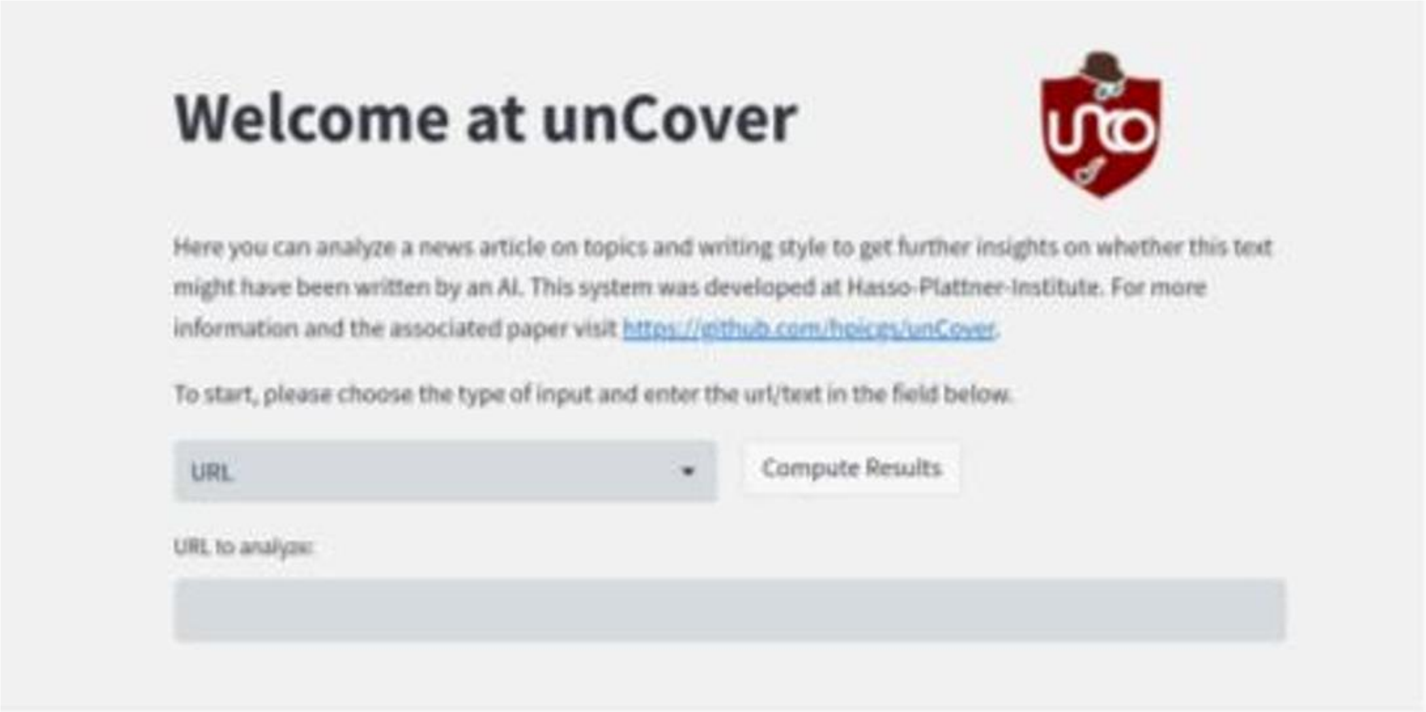


We focus our debate on the relationship between the actual class labels in the dataset and the underlying (implied) categories that are expressed in the essential complexity of the use case [Cech et al., 2025].

Intro (4/4): Essential Complexity

- Essential complexity arises from unknown unknowns within the problem that have yet to be discovered [Werner et al., 2020]
- Essential complexity becomes central when creating explanations
- Three cases
  1. Low essential complexity: Verifying an explanation is easy. Thus, it can be created post-hoc.
  2. High essential complexity: Verifying an explanation is difficult. Use surrogate models instead.
  3. Everything in-between: Use visualization (especially Dimensionality Reduction) to map the use case to the first or second case.

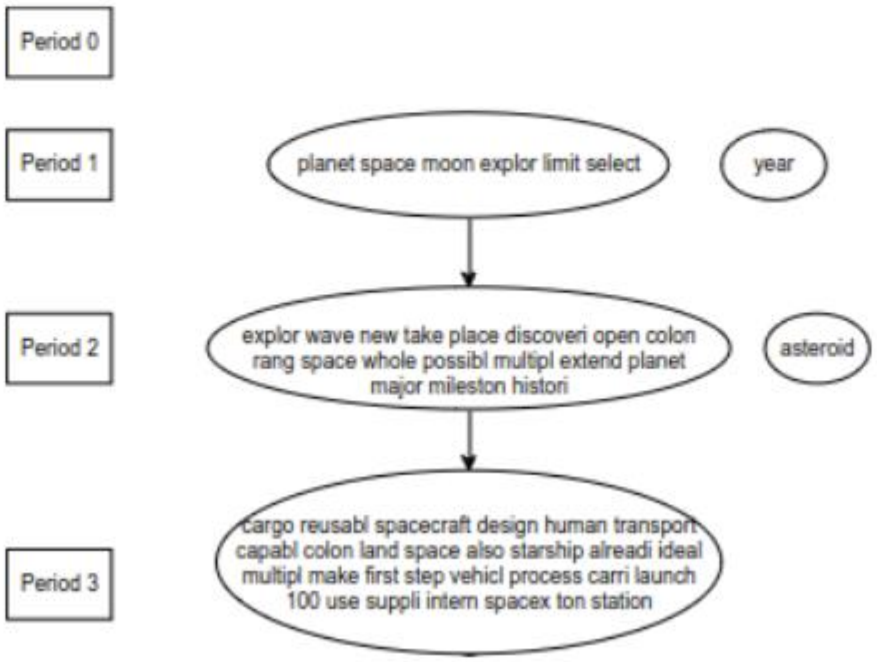
Low Essential Complexity (1/8): UNCOVER



- We proposed UNCOVER to classify whether a news story was written by a human or a Large-Language Model ([Liebe et al., 2023]; github.com/hpicgs/unCover).

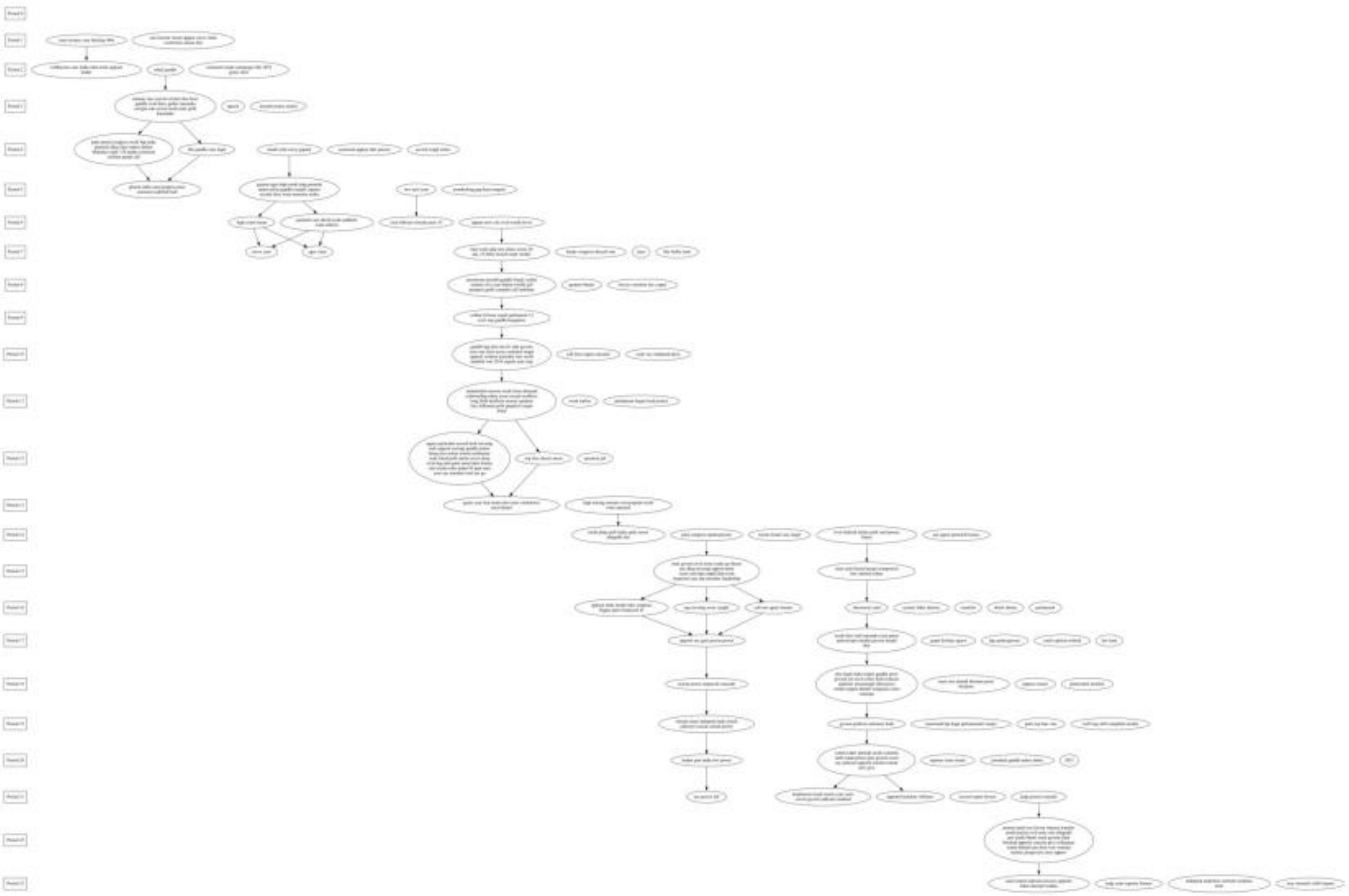
Low Essential Complexity (2/8): TEM

- We adapted the Topic Flow Model from Churchill et al. [Churchill et al., 2018] to our Topic Evolution Model (TEM)
    - Improving the filtering of stop words in early periods
    - Handling short documents where most words would be filtered
    - Handling an edge case of strictly co-occurring terms by merging nodes
    - We developed an alternative technique to matching nodes



An example evolution of our TEM model

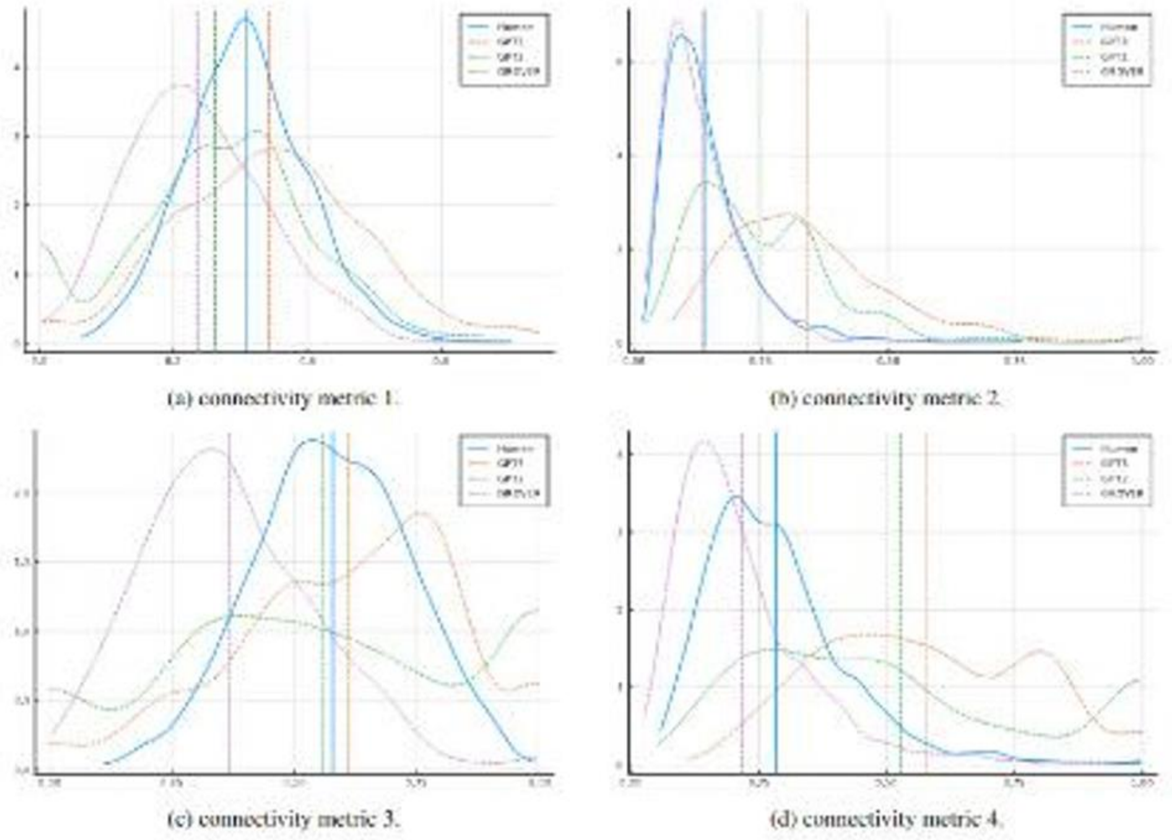
Low Essential Complexity (3/8): Human Output



Low Essential Complexity (4/8): LLM Output

Low Essential Complexity (5/8): Metrics

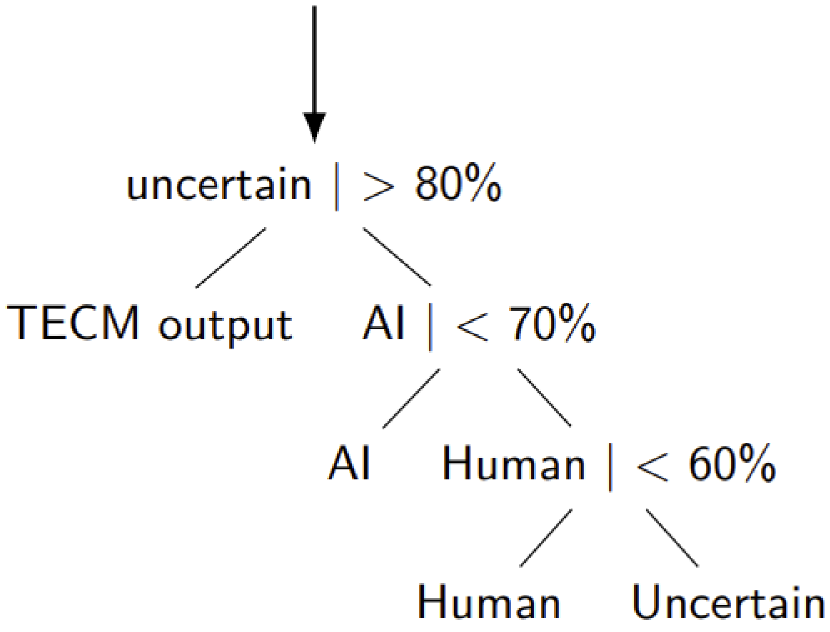
- $\texttt{metric}_1 := \left|1 - \frac{n_{ids}}{n_{topics}}\right|$
- $\texttt{metric}_2 := \left|\frac{n_{mc-ids}}{n_{topics}}\right|$
- $\texttt{metric}_3 := \left|\frac{n_{t-in}}{t-1}\right|$
- $\texttt{metric}_4 := \left|\frac{p_{max}}{t}\right|$



Low Essential Complexity (6/8): Final Approach

- We trained a logistic regression over $\mathtt{metric}_{1-4}$ and over trigrams (stylometry)
- If both regressions agree, output the decision
- Otherwise, combine the outputs:

Stylometry output | TECM confidence



Low Essential Complexity (7/8): Quantitative Eval

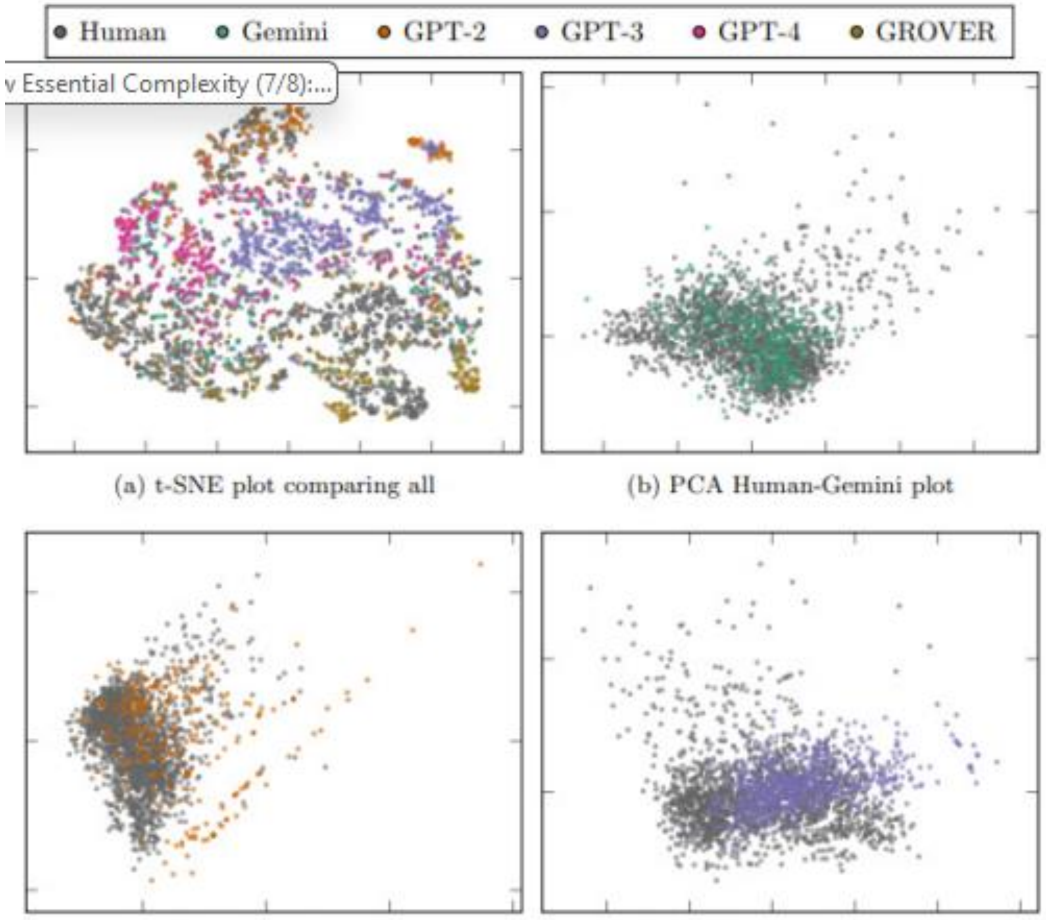
	Character Trigrams			Syntactic Trigrams			Combined Stylometry			TEcm			Final Metric		
	Machine	Human	Unsure	Machine	Human	Unsure	Machine	Human	Unsure	Machine	Human	Unsure	Machine	Human	Unsure
GPT-2	80%	6%	14%	67%	6%	27%	86%	6%	8%	84%	16%	0%	91%	1%	8%
GPT-3	58%	18%	24%	70%	8%	22%	69%	12%	19%	81%	19%	0%	85%	4%	11%
GROVER	24%	37%	39%	76%	7%	17%	53%	14%	33%	55%	45%	0%	62%	21%	17%
Human	14%	54%	32%	57%	19%	24%	32%	33%	35%	52%	48%	0%	35%	45%	20%

100%
75%
50%
25%
0%

Low Essential Complexity (8/8): Qualitative Eval

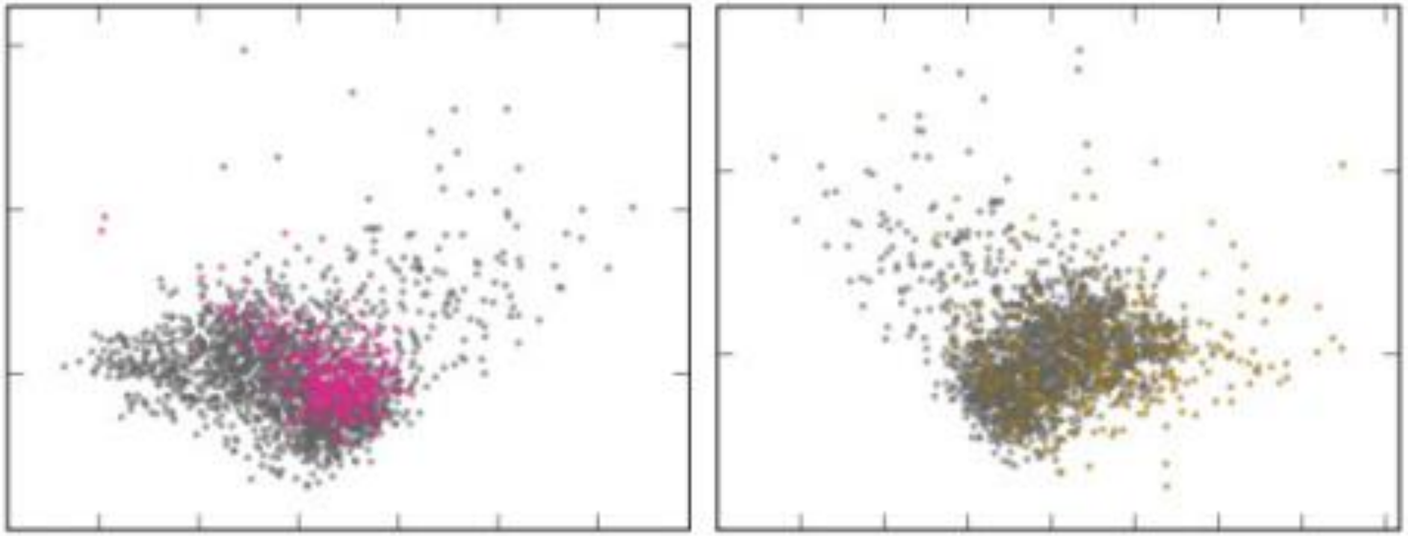
- We interviewed 13 students with experience in Machine Learning or Natural Language Processing and tasked them to correctly discern between generated and human-written news-stories
- 9 out of 13 participants (≈ 69%): UnCover helped “strongly” (4 out of 6 points) or more
- 12 out of 13 participants (≈ 92%): Visualizations are “understandable” (4 out of 6 points) or more with minimal training
- Before being questioned about it, five participants highlighted the explainable aspects of the tool.
- More than half of the participants actively changed their decision by using unCover.
- Most participants would use unCover again.

The Evolution of Language (1/2)



(a) t-SNE plot comparing all
(b) PCA Human-Gemini plot
(c) PCA Human-GPT-2 plot
(d) PCA Human-GPT-3 plot

The Evolution of Language (2/2)



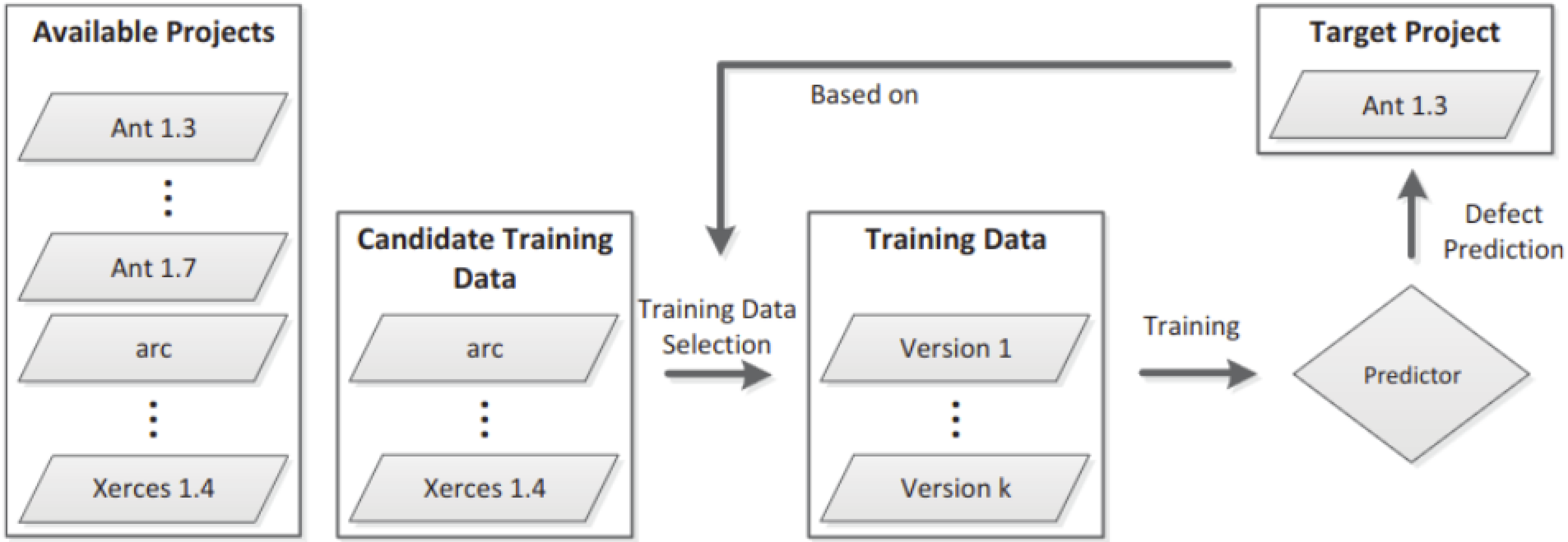
(e) PCA Human-GPT-4 plot (f) PCA Human-GROVER plot

High Essential Complexity (1/6): Software Defects

- Effort awareness [Kamei et al., 2012]
- Just-in-time criterion [Fukushima et al., 2014]
- Labels [Fu and Menzies, 2017]
  - From where?
  - Representativeness?
  - Timeliness?
- t1p.de/SoftwareQualityDays

- Source last checked: 22nd May 2023 15:31

High Essential Complexity (2/6): CPDP



Cross-Project Defect Prediction (CPDP) Aims to Predict Bugs from another project [?]

High Essential Complexity (3/6): Limits of CPDP

- Many algorithms perform poorly when used in a Cross-Project Defect Prediction context [Herbold et al., 2018]

	JURECZKO			
	AUC	*F-measure*	*G-measure*	*MCC*
ALL-RF	**0.66 (1)**	**0.32 (1)**	**0.43 (1)**	**0.17 (1)**
Amasaki15-DT	**0.6 (1)**	**0.38 (1)**	**0.49 (1)**	**0.2 (1)**
CamargoCruz09-DT	**0.58 (1)**	**0.37 (1)**	**0.5 (1)**	**0.18 (1)**
Canfora13-MODEP	0.52 (0.49)	**0.44 (1)**	**0.48 (1)**	**0.19 (1)**
CV-NET	0.71 (0.49)	**0.49 (1)**	**0.46 (1)**	0.29 (0.37)
Herbold13-RF	**0.64 (1)**	**0.39 (1)**	**0.5 (1)**	**0.17 (1)**
Kawata15-RF	**0.65 (1)**	**0.32 (1)**	**0.43 (1)**	**0.17 (1)**
Koshgoftaar08-NET	**0.6 (1)**	**0.32 (1)**	**0.4 (1)**	0.23 (0.37)
Liu10-GP	**0.63 (1)**	0.51 (0.44)	**0.52 (1)**	**0.23 (1)**
Ma12-DT	**0.6 (1)**	**0.37 (1)**	**0.49 (1)**	**0.18 (1)**
Menzies11-RF	0.58 (0.49)	**0.32 (1)**	**0.43 (1)**	**0.15 (1)**
Nam13-NB	-	-	-	-
Nam15-DT	**0.66 (1)**	0.51 (0.44)	0.63 (0.42)	0.29 (0.37)
Nam15-RF	**0.66 (1)**	0.51 (0.44)	0.63 (0.42)	0.29 (0.37)
Panichella14-CODEP-LR	**0.59 (1)**	**0.3 (1)**	**0.38 (1)**	**0.2 (1)**
Peters12-RF	**0.63 (1)**	**0.3 (1)**	**0.38 (1)**	**0.15 (1)**
Peters13-LR	0.72 (0.49)	0.17 (0.44)	0.2 (0.42)	**0.16 (1)**
Peters15-DT	**0.57 (1)**	**0.35 (1)**	**0.46 (1)**	**0.15 (1)**
PHe15-RF	**0.64 (1)**	**0.31 (1)**	**0.43 (1)**	0.13 (0.37)
Random-RANDOM	0.5 (0.13)	**0.37 (1)**	**0.49 (1)**	0.00 (0.08)
Ryu14-VCBSVM	**0.6 (1)**	0.46 (0.44)	**0.5 (1)**	**0.18 (1)**
Ryu15-DT	0.57 (0.49)	**0.29 (1)**	**0.37 (1)**	**0.14 (1)**
Trivial-FIX	0.5 (0.13)	**0.48 (1)**	0.00 (0.13)	0.00 (0.08)
Turhan09-DT	**0.59 (1)**	**0.36 (1)**	**0.47 (1)**	**0.19 (1)**
Uchigaki12-LE	0.74 (0.49)	0.08 (0.44)	0.09 (0.42)	0.1 (0.37)
Watanabe08-DT	**0.59 (1)**	**0.37 (1)**	**0.5 (1)**	0.13 (0.37)
YZhang15-BAG-DT	**0.67 (1)**	**0.37 (1)**	**0.48 (1)**	**0.22 (1)**
ZHe13-NB	**0.62 (1)**	0.46 (0.44)	**0.52 (1)**	**0.23 (1)**
Zimmermann09-LR	**0.62 (1)**	**0.39 (1)**	**0.45 (1)**	**0.17 (1)**

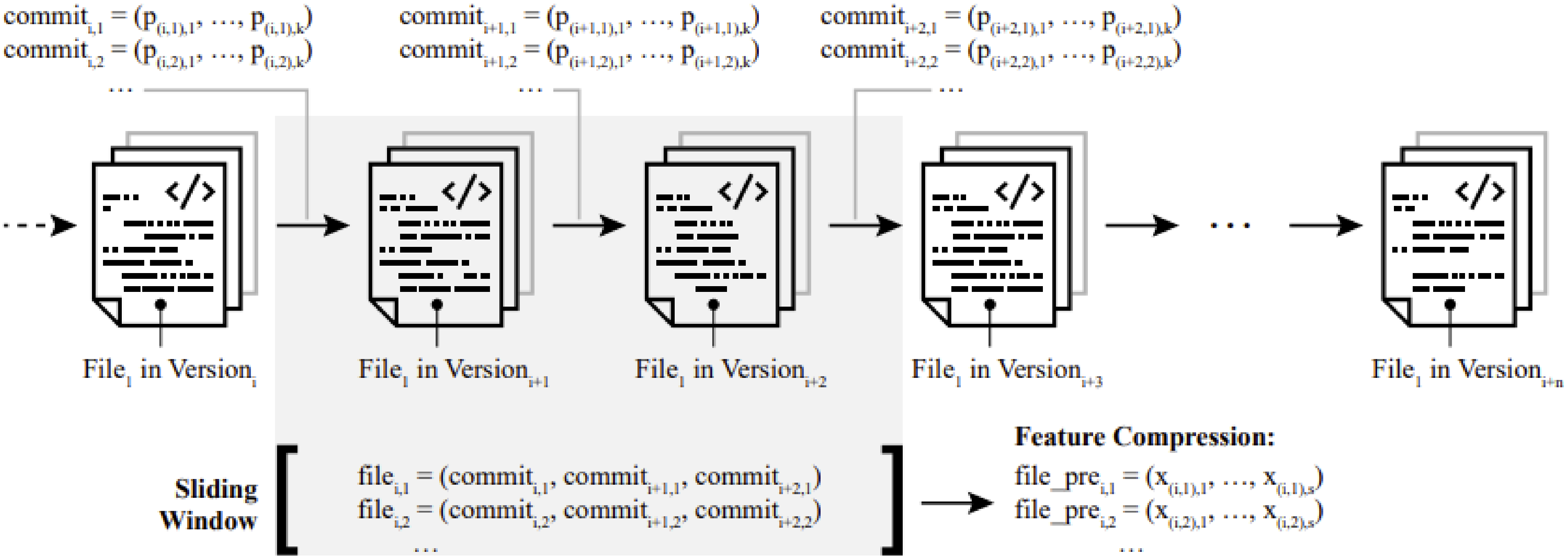
High Essential Complexity (4/6): Our Approach I

Unsupervised techniques! [Yang et al., 2016] (With informed critiques [Fu and Menzies, 2017] [Huang et al., 2017])

- Cluster-Based Techniques
- Outlier Mining Techniques

High Essential Complexity (5/6): Our Approach II

Classifying defects with different models enabled us to discover the predictive model of simple models [Cech et al., 2023a].



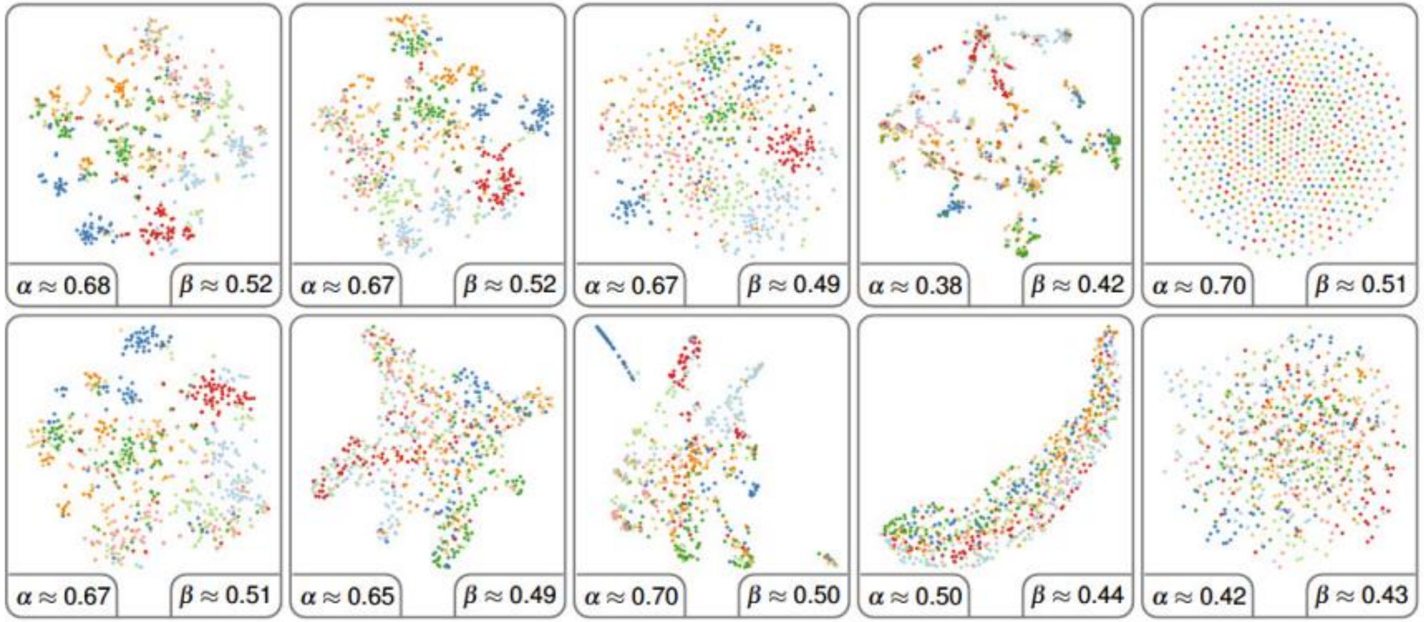
High Essential Complexity (6/6): Results

	ES Ensemble						IF Iso. Forest			LR Logistic Regression					
	FA-EN	FA-MA	FA-SMOTE	VIF-EN	VIF-MA	VIF-SMOTE	AE	FA	VIF	FA-EN	FA-MA	FA-SMOTE	VIF-EN	VIF-MA	VIF-SMOTE
Precision	86	87	86	87	87	87	84	84	85	87	88	87	87	87	87
Recall	87	88	87	78	87	86	84	85	85	79	80	78	78	78	78
F_1-score	86	87	86	81	87	86	84	85	85	82	83	80	81	81	81
Accuracy	87	88	87	78	87	86	84	85	85	79	80	78	78	78	78

100% 90% 80% 70%

	NB Naive Bayes						RF Random Forest						SVM Sup. Vector Machine					
	FA-EN	FA-MA	FA-SMOTE	VIF-EN	VIF-MA	VIF-SMOTE	FA-EN	FA-MA	FA-SMOTE	VIF-EN	VIF-MA	VIF-SMOTE	FA-EN	FA-MA	FA-SMOTE	VIF-EN	VIF-MA	VIF-SMOTE
Precision	86	85	86	86	87	86	86	87	86	87	87	87	86	86	86	85	85	85
Recall	86	87	86	85	85	85	79	78	80	81	76	81	75	88	75	80	87	80
F_1-score	86	86	86	85	86	85	81	81	82	83	79	83	78	86	78	82	83	82
Accuracy	86	87	86	85	85	85	79	78	80	81	76	81	75	88	75	80	87	80

Dimensionality Reduction



Using Dimensionality Reduction (DR) one can make high-dimensional data spaces plain in 2 dimensions ([Atzberger and Cech et al., 2024]; github.com/hpicgs/Topic-Models-and-Dimensionality-Reduction-Benchmark).

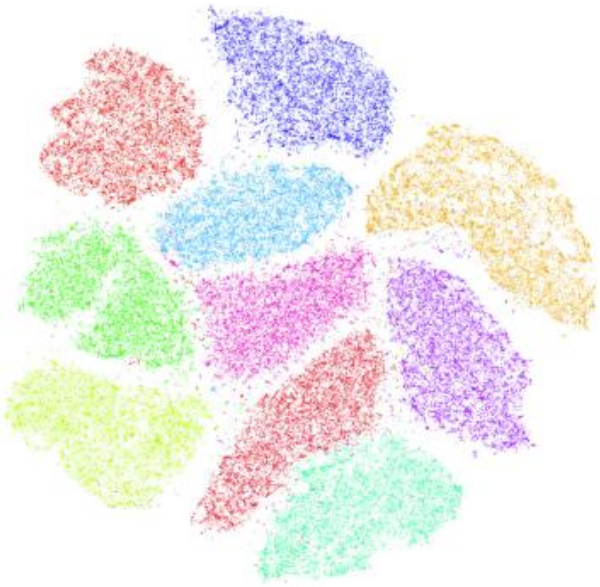
Medium Essential Complexity (1/5): Guidelines

We computed more than 40.000 layouts in the domain of Natural Language Processing and compared them via quality metrics and came up with the following guidelines:

1. Ensuring the independence of high-dimensional axes is recommended (e.g., by applying a linear combination).
2. When optimizing for accuracy metrics, a Topic Model should be applied.
3. When optimizing for perception metrics, a Topic Model might not be necessary.
4. In doubt, use t-SNE with default parameters.

Medium Essential Complexity (2/5): Method I

MNIST is a dataset with hand-written numbers



Medium Essential Complexity (3/5): Method II

```
                    An Introduction to Atheism
                  by mathew <mathew@mantis.co.uk>

This article attempts to provide a general introduction to atheism.  Whilst I
have tried to be as neutral as possible regarding contentious issues, you
should always remember that this document represents only one viewpoint.  I
would encourage you to read widely and draw your own conclusions; some
relevant books are listed in a companion article.

To provide a sense of cohesion and progression, I have presented this article
as an imaginary conversation between an atheist and a theist.  All the
questions asked by the imaginary theist are questions which have been cropped
up repeatedly on alt.atheism since the newsgroup was created.  Some other
frequently asked questions are answered in a companion article.

Please note that this article is arguably slanted towards answering questions
posed from a Christian viewpoint.  This is because the FAQ files reflect
questions which have actually been asked, and it is predominantly Christians
who proselytize on alt.atheism.

So when I talk of religion, I am talking primarily about religions such as
Christianity, Judaism and Islam, which involve some sort of superhuman divine
being.  Much of the discussion will apply to other religions, but some of it
may not.

"What is atheism?"
```

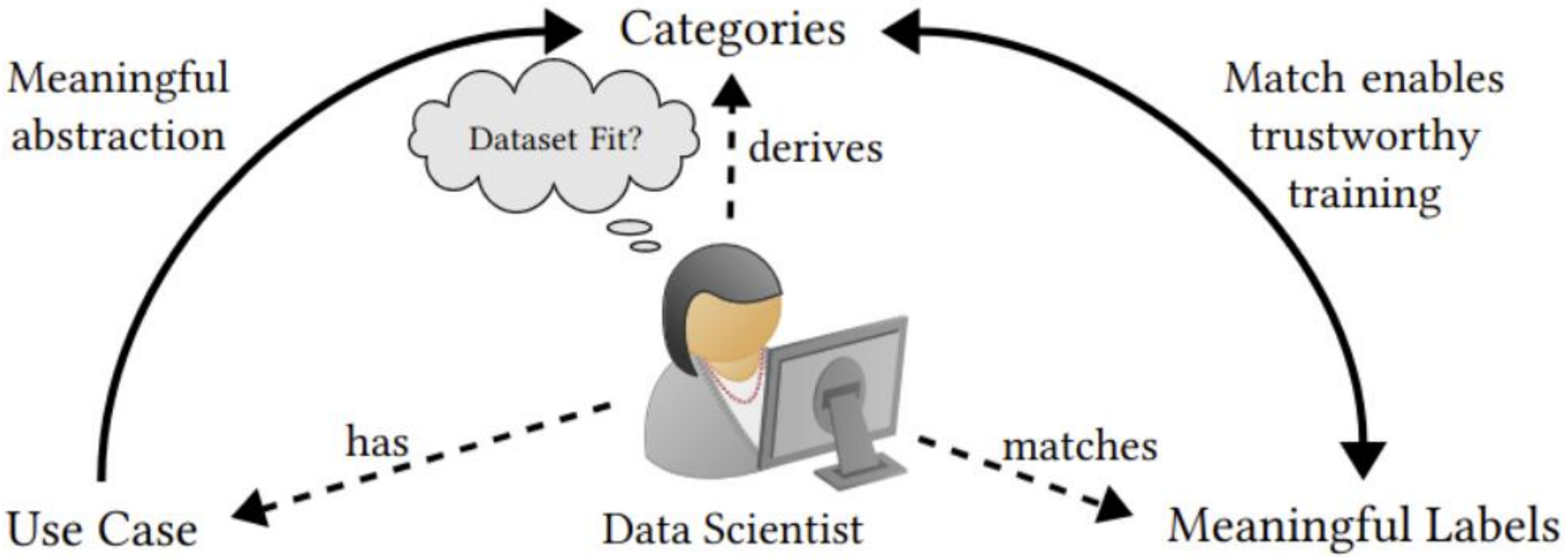
```
>Why is it more reasonable than the trend towards obesity and the trend towards
>depression? You can't just pick your two favorite trends, notice a correlation
>in them, and make a sweeping statement of generality. I mean, you CAN, and
>people HAVE, but that does not mean that it is a valid or reasonable thesis.
>At best it's a gross oversimplification of the push-pull factors people
>experience.

I agree, I reckon it's television and the increase in fundamentalism.. You
think its the increase in pre-marital sex... others thinks its because
psychologists have taken over the criminal justice system and let violent
criminals con them into letting them out into the streets... others think
it's the increase in designer drugs... others think it's a communist plot.
Basically the social interactions of all the changing factors in our society
are far too complicated for us to control. We just have to hold on to the
panic handles and hope that we are heading for a soft landing. But one
things for sure, depression and the destruction of the nuclear family is not
due solely to sex out of marriage.
```

The 20 newsgroups dataset contains text that were gathered on 20 public mailing lists (here two examples of the “atheist” mailing list is shown).

Medium Essential Complexity (4/5): Method III

Medium Essential Complexity (5/5): Method IV



Investigating datasets in a DR layout may enable matching categories in classes [Cech et al., 2025]

Discussion (1/2): No-Free-Lunch Theorem

"The 'No Free Lunch' theorem states that, averaged over all optimization problems, without re-sampling, all optimization algorithms perform equally well." [Adam et al., 2019, p.1] This implies that good results on one dataset may not be applicable on another one [Raji et al., 2021].

Photo by Jezael Melgoza on Unsplash

Discussion (2/2): Curse of Dimensionality

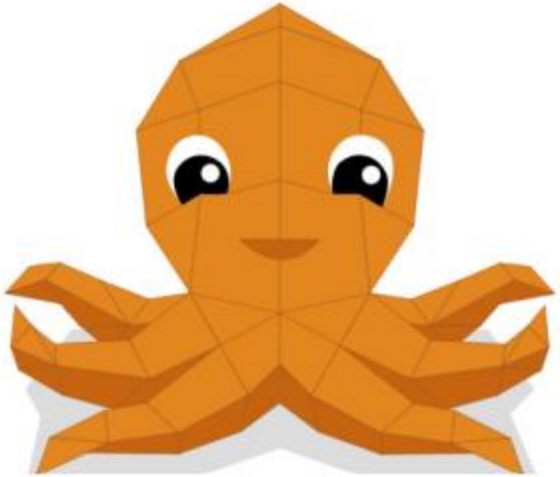
"High-dimensional spaces show surprising, counter-intuitive geometrical properties that have a large influence on the performances of data analysis tools. Among these properties, the concentration of the norm phenomenon results in the fact that Euclidean norms and Gaussian kernels, both commonly used in models, become inappropriate in high-dimensional spaces" [Verleysen and François, 2005, p. 758].

Photo by Alexander Andrews on Unsplash

Outlook (1/4): SDMX

sdmx.io

Governance | Use Cases | Tools | Resources | Events | Contact

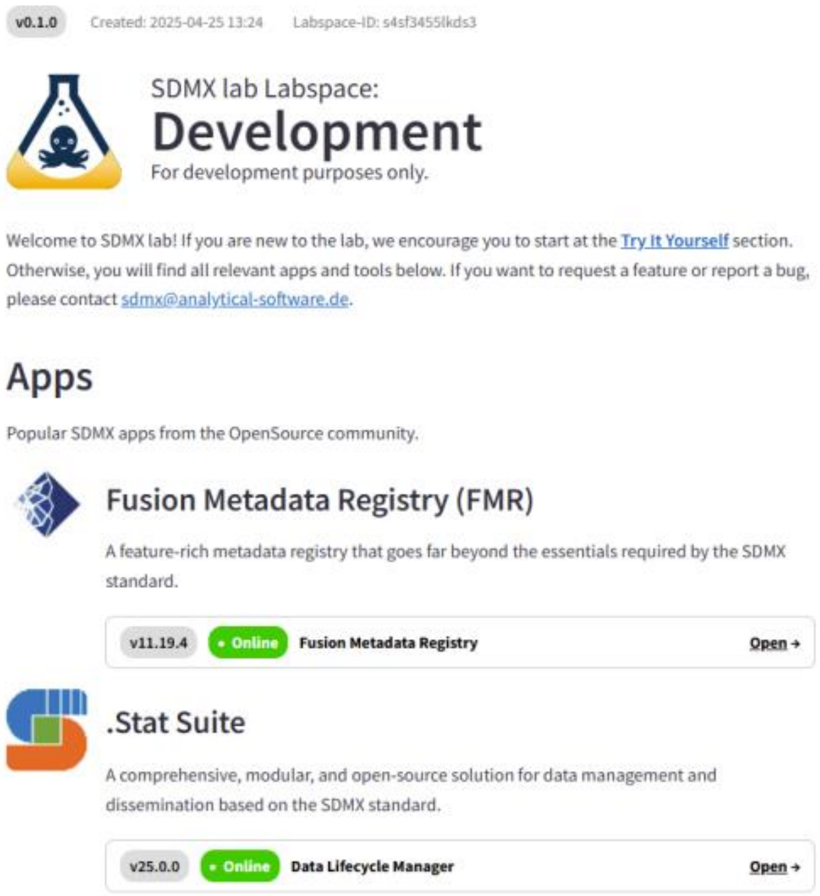


Solutions for official statistics use cases

sdmx.io is not a single project but an ecosystem of open source tools, patterns, guidance, learning materials and other resources like pre-configured containerised environments that make the software quick and simple to deploy.

sdmx.io

Outlook (2/4): SDMX Lab


v0.1.0 Created: 2025-04-25 13:24 Labspace-ID: s4sf3455lkds3

SDMX lab Labspace:

Development

For development purposes only.

Welcome to SDMX lab! If you are new to the lab, we encourage you to start at the Try It Yourself section. Otherwise, you will find all relevant apps and tools below. If you want to request a feature or report a bug, please contact sdmx@analytical-software.de.

Apps

Popular SDMX apps from the OpenSource community.

Fusion Metadata Registry (FMR)

A feature-rich metadata registry that goes far beyond the essentials required by the SDMX standard.

v11.19.4 • Online **Fusion Metadata Registry** Open →

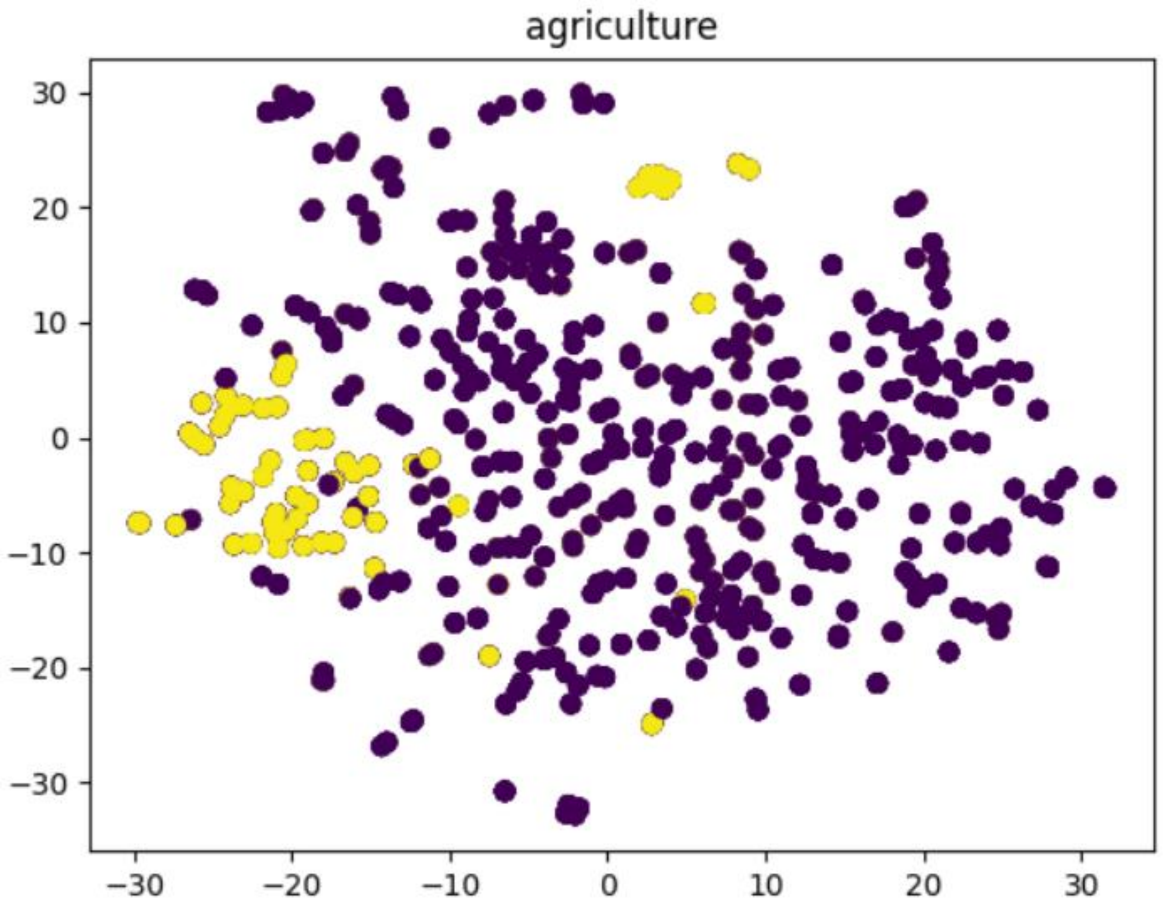
.Stat Suite

A comprehensive, modular, and open-source solution for data management and dissemination based on the SDMX standard.

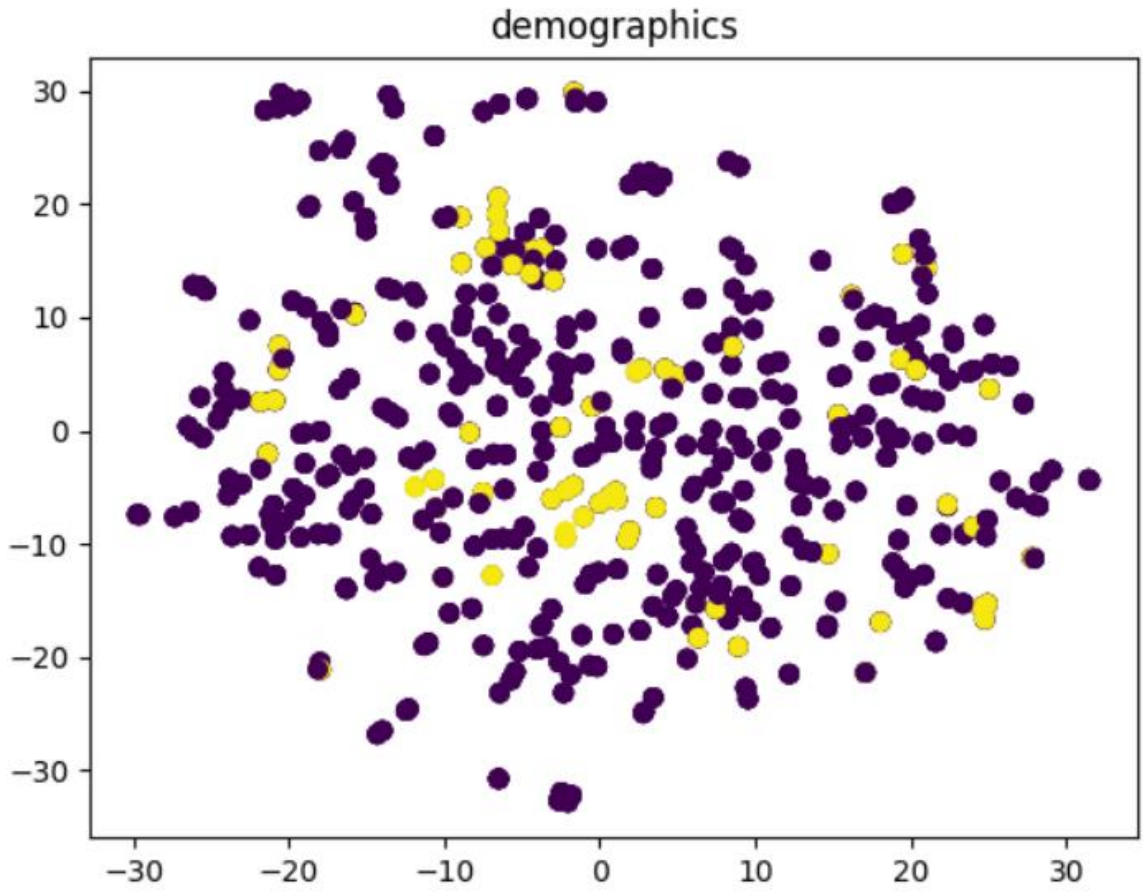
v25.0.0 • Online **Data Lifecycle Manager** Open →

analytical-software.de/en/sdmx/

Outlook (3/4): Keywords I



Outlook (4/4): Keywords II



Conclusions

- We used a tool-focused approach that allows for practical evaluations but also increases the probability of a software defect that alters the data
- In addition, the interpretative approach renders the investigation necessarily incomplete and open for further investigation
- However, we found practical evidence and discovered our central hypothesis from our data
- Thus, essential complexity was found to be a central property when creating explanations
- Contributes to concerns on “sociotechnical frictions” (e.g. [Crisan et al., 2024])

References I

[Adam et al., 2019] Adam, S. P., Alexandropoulos, S.-A. N., Pardalos, P. M., and Vrahatis, M. N. (2019). No free lunch theorem: A review. In Approximation and Optimization: Algorithms, Complexity and Applications, pages 57–82. Springer.

[Atzberger et al., 2025] Atzberger, D., Cech, T., Scheibel, W., Döllner, J., Behrisch, M., and Schreck, T. (2025). A large-scale sensitivity analysis on latent embeddings and dimensionality reductions for text spatializations. Transactions on Visualization and Computer Graphics, 31(1):305–315.

[Atzberger et al., 2024] Atzberger, D., Cech, T., Trapp, M., Richter, R., Scheibel, W., Döllner, J., and Schreck, T. (2024). Large-scale evaluation of topic models and dimensionality reduction methods for 2d text spatialization. IEEE Transactions on Visualization and Computer Graphics, 30(1):902–912.

[Blok and Pedersen, 2014] Blok, A. and Pedersen, M. A. (2014). Complementary social science? quali-quantitative experiments in a big data world. Big Data & Society, 1(2):1–6.

References II

[Böttger et al., 2023] Böttger, F., Cech, T., Scheibel, W., and Döllner, J. (2023). Visual counterfactual explanations using semantic part locations. In Proceedings of the 15th International Conference on Knowledge Discovery and Information Retrieval, KDIR '23, pages 63–74. INSTICC, SciTePress.

[Cech et al., 2023a] Cech, T., Atzberger, D., Scheibel, W., Misra, S., and Döllner, J. (2023a). Outlier mining techniques for software defect prediction. In SWQD 2023: Software Quality: Higher Software Quality through Zero Waste Development, volume 472 of Lecture Notes in Business Information Processing, pages 41–60. Springer.

[Cech et al., 2024] Cech, T., Kohlros, E., Scheibel, W., and Döllner, J. (2024). A dashboard for simplifying machine learning models using feature importances and spurious correlation
analysis. In Proceedings of the 26th EG Conference on Visualization – Posters, EuroVis Posters '24. EG.

References III

[Cech et al., 2023b] Cech, T., Simsek, F., Scheibel, W., and Döllner, J. (2023b). A dashboard for interactive convolutional neural network training and validation through saliency maps. In Proceedings of the 25th EG Conference on Visualization – Posters, EuroVis Posters '23, pages 5–7. EG.

[Cech et al., 2025] Cech, T., Wegen, O., Atzberger, D., Richter, R., Scheibel, W., and Döllner, J. (2025). Standardness clouds meaning: A position regarding the informed usage of standard datasets.

[Churchill et al., 2018] Churchill, R., Singh, L., and Kirov, C. (2018). A temporal topic model for noisy mediums. In Proc. 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD '18, pages 42–53. Springer.

[Crisan et al., 2024] Crisan, A., Butters, N., and Zoe (2024). Exploring subjective notions of explainability through counterfactual visualization of sentiment analysis. In 2024 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV), pages 15–24.

References IV

[danah boyd and Crawford, 2012] danah boyd and Crawford, K. (2012). Critical questions for big data. Information, Communication & Society, 15(5):662–679.

[Fu and Menzies, 2017] Fu, W. and Menzies, T. (2017). Revisiting unsupervised learning for defect prediction. In Proceedings of the 2017 11th joint meeting on foundations of software engineering, pages 72–83.

[Fukushima et al., 2014] Fukushima, T., Kamei, Y., McIntosh, S., Yamashita, K., and Ubayashi, N. (2014). An empirical study of just-in-time defect prediction using cross-project models. In Proceedings of the 11th Working Conference on Mining Software Repositories, pages 172–181.

[Herbold et al., 2018] Herbold, S., Trautsch, A., and Grabowski, J. (2018). A comparative study to benchmark cross-project defect prediction approaches. In Proceedings of the 40th International Conference on Software Engineering, ICSE '18, pages 1063–1088. Association for Computing Machinery.

References V

[Hohman et al., 2020] Hohman, F., Park, H., Robinson, C., and Polo Chau, D. H. (2020). Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. IEEE Transactions on Visualization and Computer Graphics, 26(1):1096–1106.

[Huang et al., 2017] Huang, Q., Xia, X., and Lo, D. (2017). Supervised vs unsupervised models: A holistic look at effort-aware just-in-time defect prediction. In 2017 IEEE International Conference on Software Maintenance and Evolution (ICSME), pages 159–170. IEEE.

[Jo and Gebru, 2020] Jo, E. S. and Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 conference on fairness, accountability, and transparency, pages 306–316.

[Kamei et al., 2012] Kamei, Y., Shihab, E., Adams, B., Hassan, A. E., Mockus, A., Sinha, A., and Ubayashi, N. (2012). A large-scale empirical study of just-in-time quality assurance. IEEE Transactions on Software Engineering, 39(6):757–773.

References VI

[Kitchin and McArdle, 2016] Kitchin, R. and McArdle, G. (2016). What makes big data, big data? exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1):1–10.

[Liebe et al., 2023] Liebe, L., Baum, J., Schütze, T., Cech, T., Scheibel, W., and Döllner, J. (2023). unCover: Identifying ai generated news articles by linguistic analysis and visualization. In Proceedings of the 15th International Conference on Knowledge Discovery and Information Retrieval, KDIR '23, pages 39–50. INSTICC, SciTePress.

[Pappagari et al., 2019] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019). Hierarchical transformers for long document classification. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 838–844.

[Raji et al., 2021] Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). Ai and the everything in the whole wide world benchmark. arXiv preprint arXiv:2111.15366.

[Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5):206–215.

References VII

[Scheuerman et al., 2021] Scheuerman, M. K., Denton, E., and Hanna, A. (2021). Do datasets have politics? disciplinary values in computer vision dataset development. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–37.

[Seaver, 2015] Seaver, N. (2015). The nice thing about context is that everyone has it. Media, Culture & Society, 37(7):1101–1109.

[Speith, 2022] Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (xai) methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, pages 2239–2250. ACM.

[Verleysen and François, 2005] Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In Computational Intelligence and Bioinspired systems: 8th International Work-Conference on Artificial Neural Networks, IWANN '05, pages 758–770. Springer.

References VIII

[Wang et al., 2016] Wang, D., Cui, P., and Zhu, W. (2016). Structural deep network embedding. In 2016 International Conference on Knowledge Discovery and Data Mining (KDD), KDD '16, pages 1225–1234. Association for Computing Machinery.

[Werner et al., 2020] Werner, C., Li, Z. S., Ernst, N., and Damian, D. (2020). The lack of shared understanding of non-functional requirements in continuous software engineering:
Accidental or essential? In 2020 IEEE 28th international requirements engineering conference (RE), RE '20, pages 90–101. IEEE.

[Yang et al., 2016] Yang, Y., Zhou, Y., Liu, J., Zhao, Y., Lu, H., Xu, L., Xu, B., and Leung, H. (2016). Effort-aware just-in-time defect prediction: simple unsupervised models could be better than supervised models. In Proceedings of the 2016 24th ACM SIGSOFT international symposium on foundations of software engineering, pages 157–168.