Kandinsky Patterns: An open toolbox for creating explainable machine learning challenges

Heimo Müller, Andreas Holzinger
Institute for Medical Informatics, Statistics and Documentation
Medical University Graz, Austria
heimo.mueller|andreas.holzinger@medunigraz.at

2-page Extended Abstract CiML-2019-Workshop, Vancouver, Canada

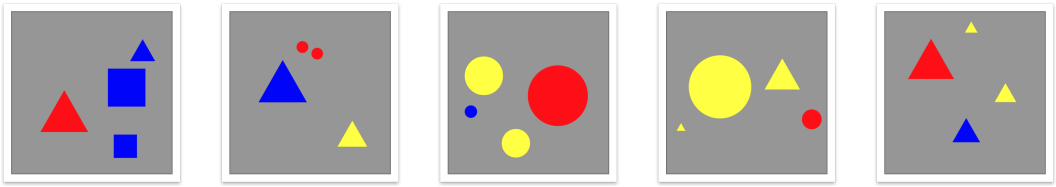

Figure 1: A Kandinsky Pattern following the ground truth *"each Kandinsky Figure has two pairs of objects with the same shape, in one pair the objects have the same color, in the other pair different colors, two pairs are always disjunct, i.e. they don't share a object"*.

Kandinsky Patterns are mathematically describable, simple self-contained, and controllable test data sets for the development, validation and training of explainability in machine learning [1]. While Kandinsky Patterns possess computationally manageable properties, the big advantage is that they are at the same time easily distinguishable from human observers and can therefore also be described by both humans and computers in strictly controllable experimental settings [2]. This is extremely important for the current trend of the international research community in Explainable Machine Learning [3].

The set of all possible Kandinsky Figures - we call this the Kandinsky Universe - is divided by ground truth into two subsets. (A) Kandinsky Figures, which belong to a Kandinsky Pattern, and (B) Kandinsky Figures, which don't belong to the Kandinsky Pattern. Ground truth" can be defined in different ways, e.g. by mathematical functions, as natural language statement or by an algorithm. Different representations of "ground truth" are equivalent, if the resulting Kandinsky Patterns contains exactly the same Kandinsky Figures.

To construct a challenge we generate data sets with different complexity of ground truth and vary the composition and size of the training and test data sets. With this approach we can simulate real word applications and challenge machine learning algorithms to (i) find ground truth by a successful classification of Kandinsky Figures and (ii) to generate human understandable explanations of the algorithms applied.

The process of explanation is the generation and refinement of a hypothesis to find the underlying description. The validation is achieved by the method of asking a question, forming a testable hypothesis, setting up the experimental design, running the experiment and either accepting the hypothesis, rejecting it or, in the third case one cannot make any assumption [4].

With the Kandinsky Patterns method we can generate test data sets to simulate and analyze the process as described above. Our approach allows to:

- Challenge classification algorithms together with explanation,
- Compare human and machine explanation strategies,
- Control the training and test data sets in a fine grained way, e.g. with counterfactuals, which falsifies a wrong hypothesis,
- Simulate domain specific visual pattern, e.g. from medicine or architecture with abstract representations [5].
- Challenge concept mapping and natural language explanation.

Figure 2 shows Kandinsky Figures from the data set "Blue and Yellow Circles"[1], which demonstrates some of the above principles.

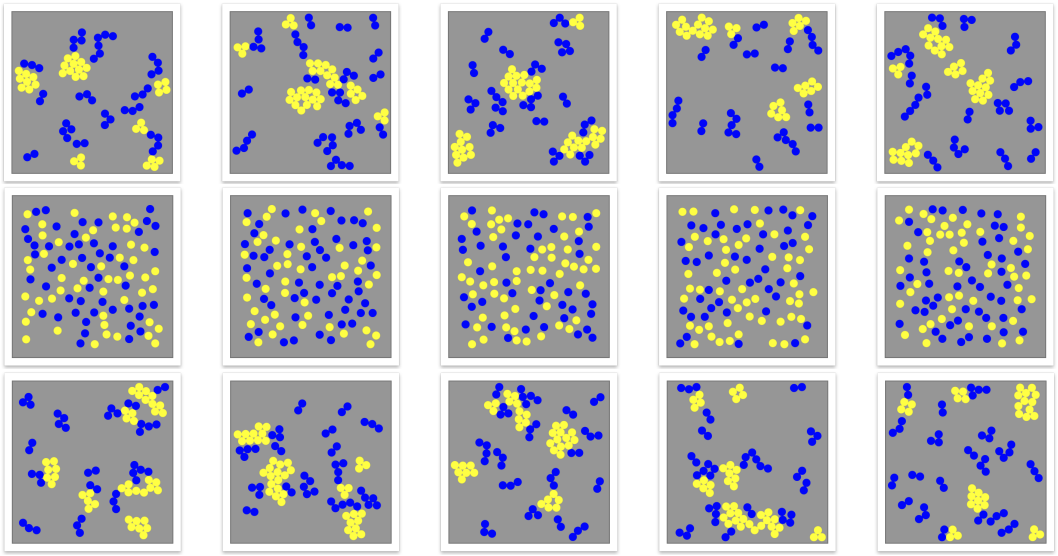

Figure 2: The first row shows Kandinsky Figures according to ground truth, second row shows Kandinsky Figures with approximately the same number of objects not belonging to the Kandinsky Pattern and in the third row Kandinsky Figures, which are "almost true", i.e. they fulfill a hypothesis similar to the ground truth, but falsifies ground truth.

References

[1] Heimo Müller and Andreas Holzinger. Kandinsky patterns. *arXiv:1906.00657*, 2019. URL https://arxiv.org/abs/1906.00657.

[2] Andreas Holzinger, Michael Kickmeier-Rust, and Heimo Müller. Kandinsky patterns as iq-test for machine learning. In *Springer Lecture Notes LNCS 11713*, pages 1–14. Springer Nature Switzerland, 2019. doi: 10.1007/978-3-030-29726-8_1.

[3] Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Jacques, Meysam Madadi, Xavier Baró, Stephane Ayache, Evelyne Viegas, Yaŏmur Güçlútúrk, and Umut Güçlü. Design of an explainable machine learning challenge for video interviews. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966320.

[4] Karl Popper. *Die Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Springer-Verlag, Wien, 1935.

[5] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of ai in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pages 1–13, 2019. doi: 10.1002/widm.1312.

[1] https://human-centered.ai/kandinksy-challenge