~~Machine learning revealed~~ Did machine learning reveal symbolism, emotionality, and imaginativeness as primary predictors of creativity?

Assessing the reproducibility and replicability of Spee et al.'s findings on creativity in Western art

Xuanyu Chen, Gabriel Patron, and Tim White
STATS 604 - Project 3

About the study

www.nature.com/scientificreports

scientific reports

Check for updates

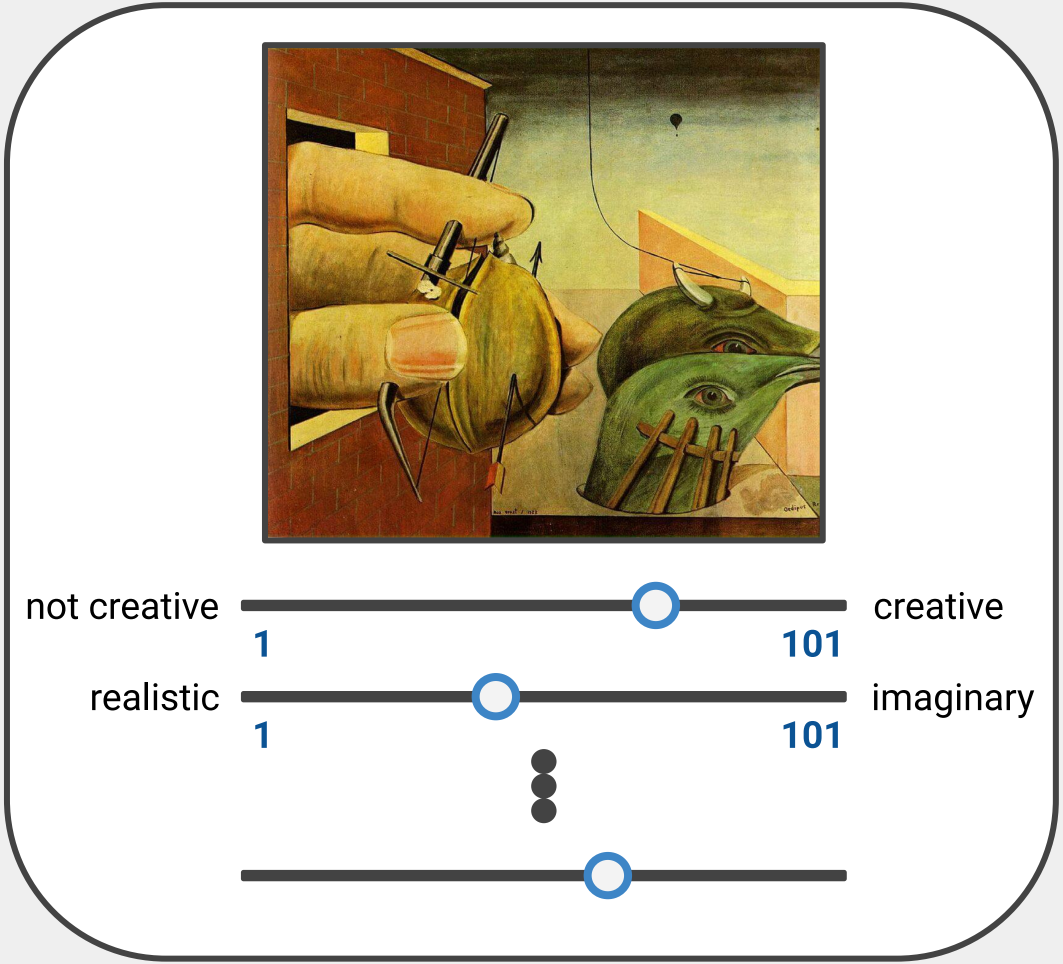
OPEN

Machine learning revealed symbolism, emotionality, and imaginativeness as primary predictors of creativity evaluations of western art paintings

“Which subjectively perceived art-attributes contribute to the judgment of an artwork’s level of creativity?”

Study procedure

- 78 raters (non-experts)
- 54 paintings
- Response:
  `creativity judgment`
- Predictors:
  17 attributes
- Method:
  Random forest



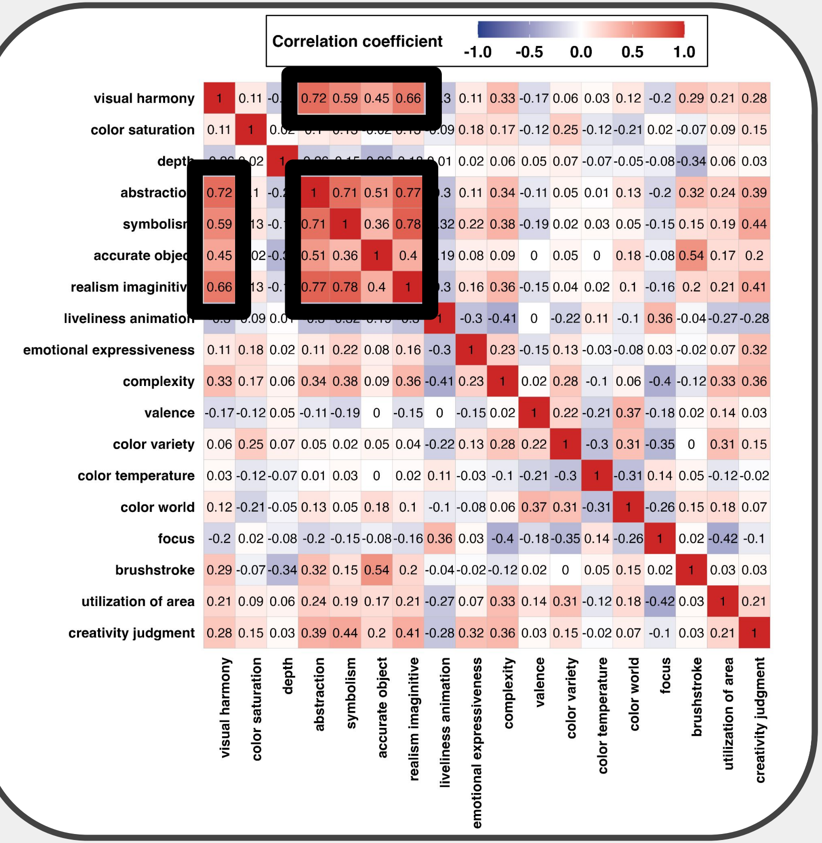
Objectives

Authors' findings:

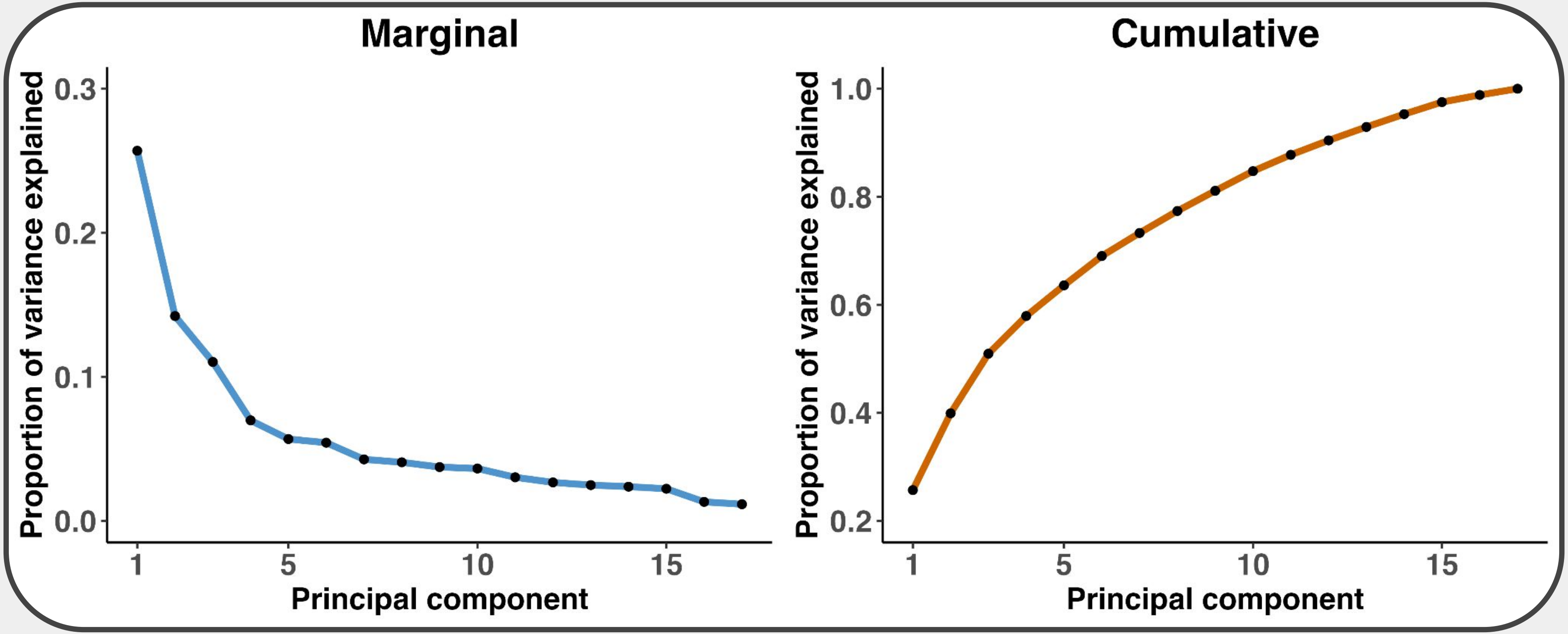
- Attributes explain ~30% of variance in creativity
- Mean absolute error = 17.5 ± 0.9
- Most important attributes: symbolism, emotionality, imaginativeness

1. **Reproduce?**
2. **Decisions, omissions, or ambiguities?**
3. **Replicate?**

Exploratory plots

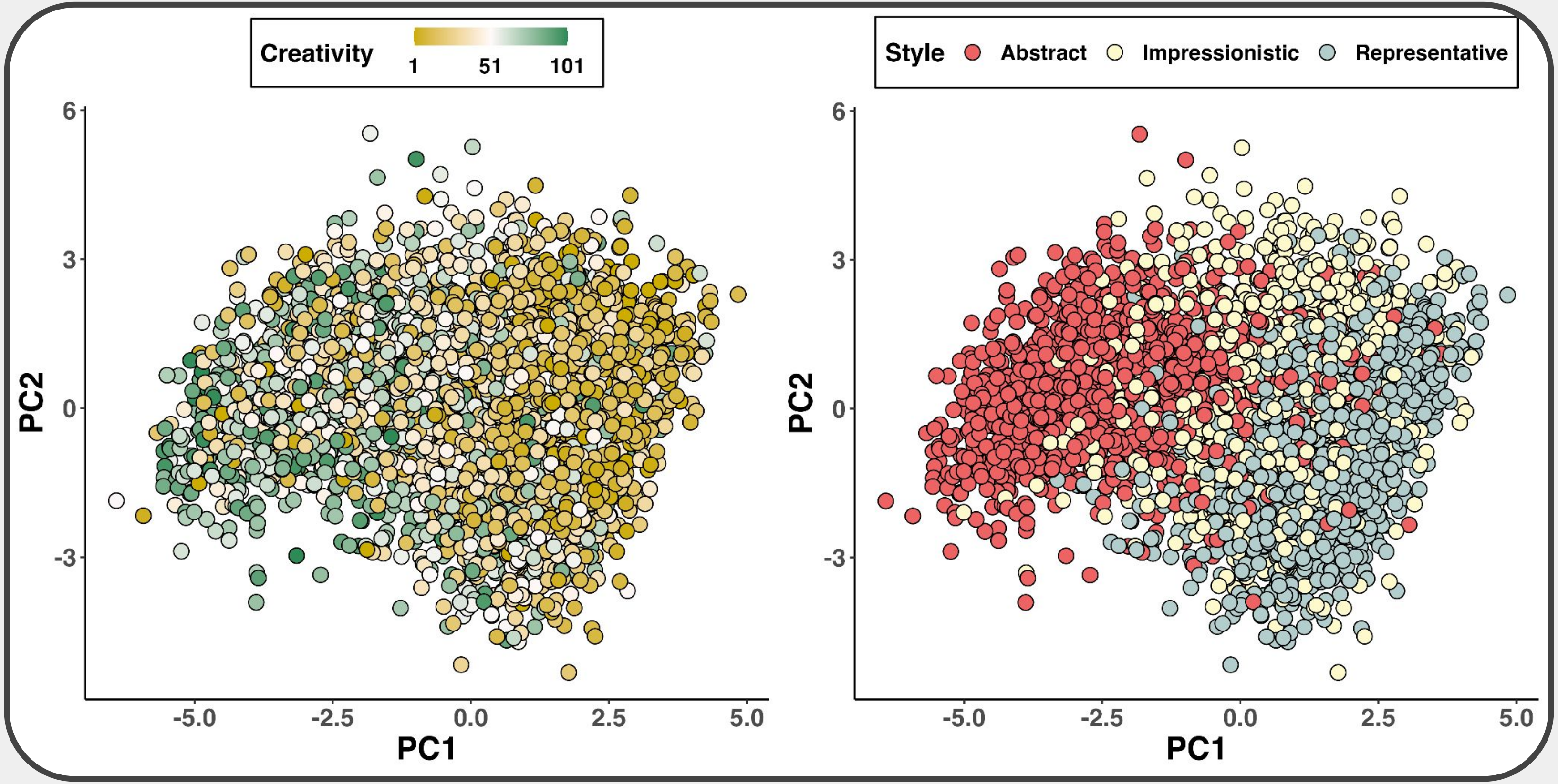


PCA



Marginal
Proportion of variance explained
0.3
0.2
0.1
0.0
1
5
10
15
Principal component
Cumulative
Proportion of variance explained
1.0
0.8
0.6
0.4
0.2
1
5
10
15
Principal component

PCA


Creativity
1
51
101
PC2
PC1
6
3
0
-3
-5.0
-2.5
0.0
2.5
5.0
Style
Abstract
Impressionistic
Representative
PC2
PC1

Reproducing their results

Background I: Nested cross-validation

- Designed to avoid bias in traditional cross-validation

Outer Loop (`n_outer`):

Splits data into `train_data` and `test_data`

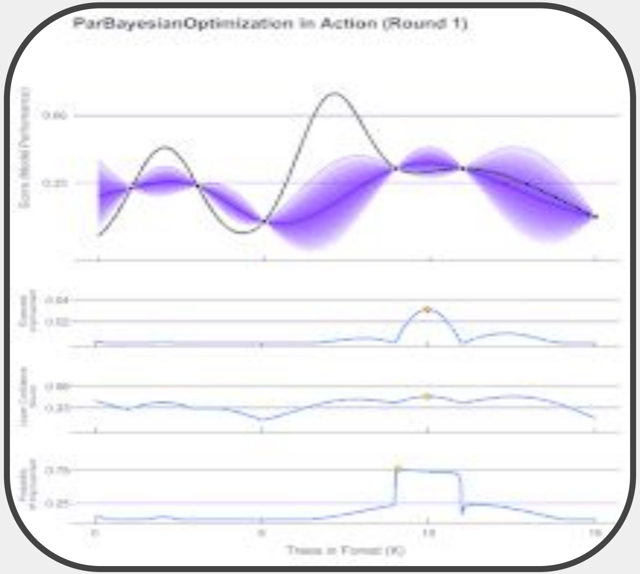
Inner Loop (`n_inner`):

Splits `train_data` into `inner_train` and `inner_validation`

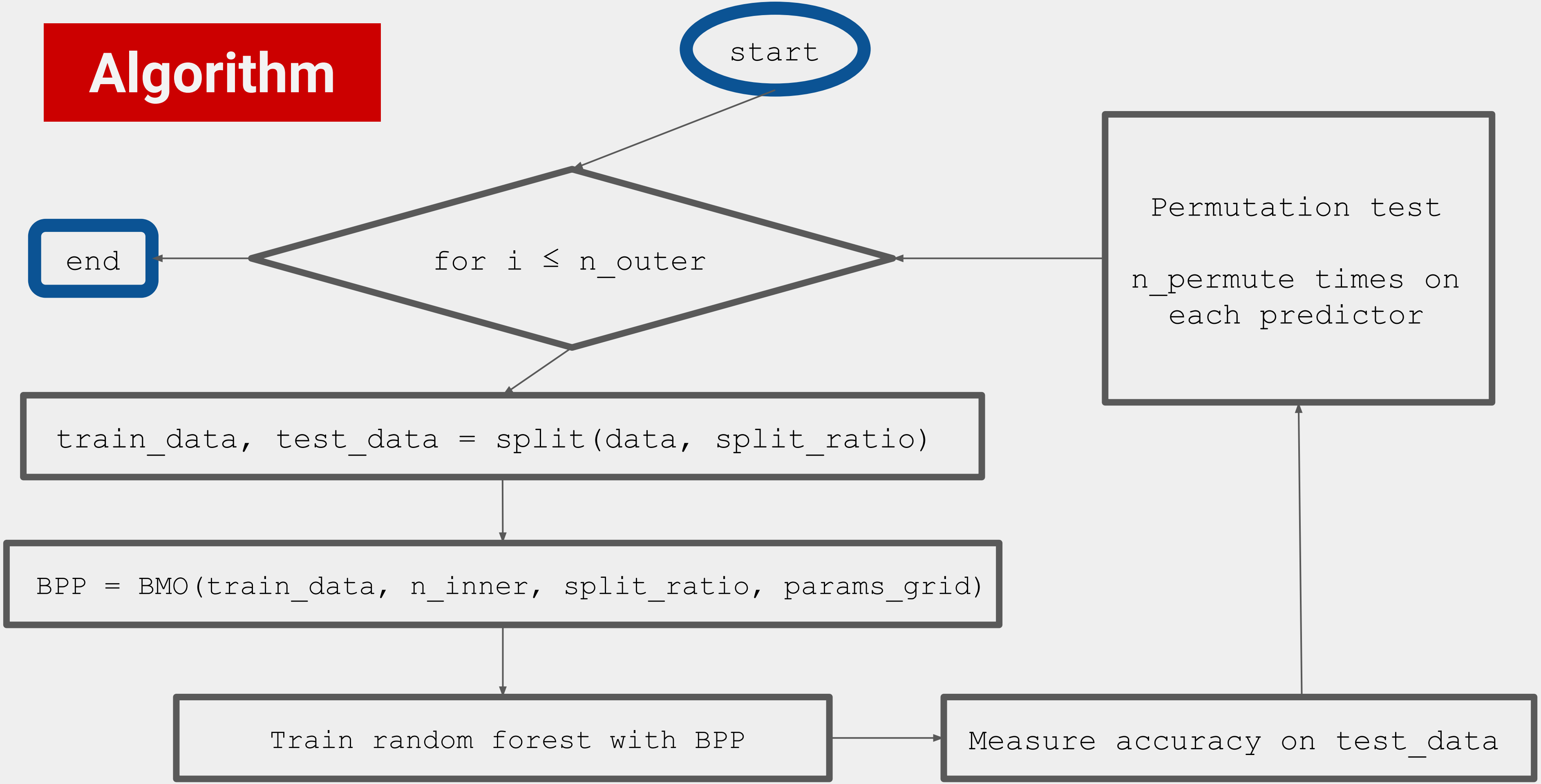
Performs parameter search and outputs best-performing params (`BPP`)

Background II: Bayesian model optimization

- Exhaustive grid-search can be very slow
- BMO maximizes an objective function with iterative Bayesian updates



Algorithm


start
for i ≤ n_outer
end
Permutation test
n_permute times on each predictor
train_data, test_data = split(data, split_ratio)
BPP = BMO(train_data, n_inner, split_ratio, params_grid)
Train random forest with BPP
Measure accuracy on test_data

Unexplained choices

Why optimize over certain **parameters** and not the rest?

Why that **number of iterations**?

Why include certain **predictors** and not others?

Why use random forest instead of **another learning algorithm**?

Summary of the models

- Models were assessed on test set at each outer loop
- *Metrics of prediction performance:* Prediction R^2 and MAE
- *Statistical significance:* Permutation test
  - Shuffle the response on train set, refit the model and recompute the metrics
- *Variable importance:* Reduction of prediction R^2
  - of the original forest when a certain column of the test set is shuffled

Numerical setting

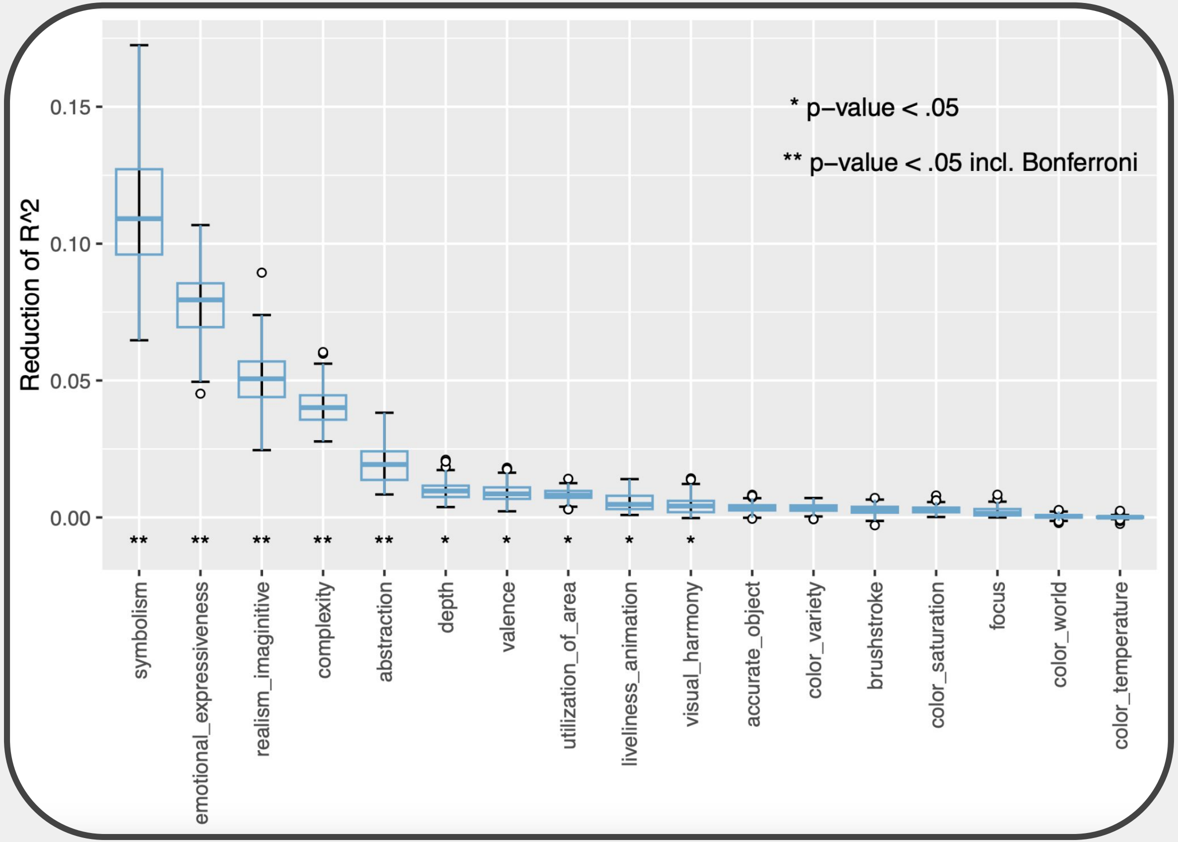
- # of outer loops: 128
- Hyperparameter space being searched over:
  - Minimum sample size of split nodes/leaf nodes: 2-128, 1-128
  - maximum number of features considered at each split: 1-17
- Setting of BMO:
  - 96 initial points and 128 iterations, **12,288** forests in total
  - Too costly, we set them to 24 and 64, respectively
- # of permutations in model assessment: 64

Results: Prediction performance

Response	Average MAE $\pm$ sd	p-value of MAE	Average $R^2 \pm$ sd	p-value of R^2
Creativity	17.31 ± 0.44	$p < 0.001$	0.33 ± 0.02	$p < 0.001$

- Results successfully reproduced, with lower standard deviations
- Predictions differ from the observed responses by 17.3 points, on average
- About 30% of the total variance in creativity judgement explained by the model

Results: Variable importance

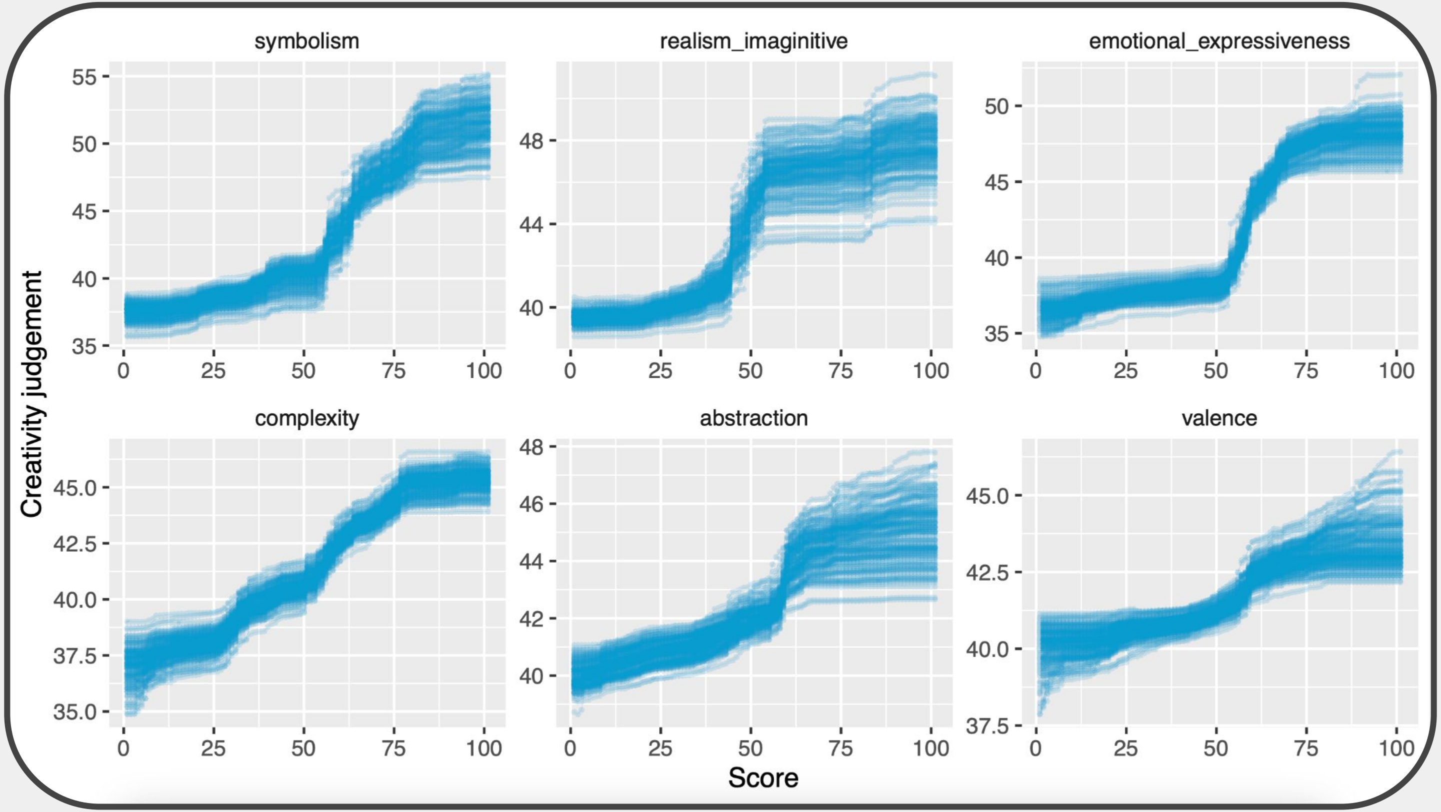


Results: Variable importance

- The three most important attributes:

  symbolism (0.12) > emotionality (0.08) > imaginativeness (0.05)

Results: Partial dependence plots



Results: Partial dependence plots

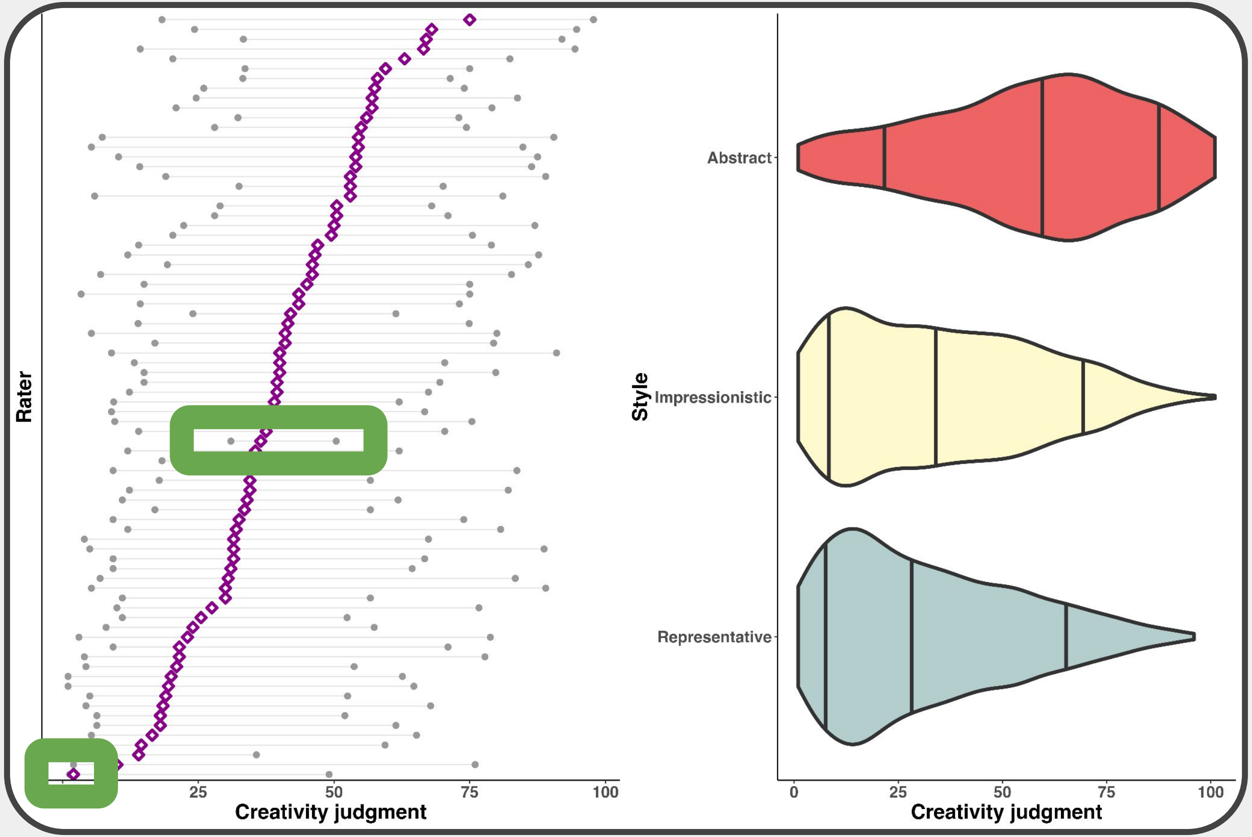
- For all 128 fitted forests and six most important attributes
- Goal: characterize the marginal relationship between creativity judgment and each attribute
- Reproduced their results which supports their claims that
  - important attributes are positively associated with the response
  - these associations cannot be described as linear
    — sudden nonlinear changes are observed

Comments on their analysis

- In general, reasonable and convincing
- Sampling randomness properly addressed by repetition
- Permutation tests are totally distribution-free
- One issue: Some hyperparameters being tuned are highly related

Replicating their results

Overlooked sources of variation



Approach #1: Linear mixed-effects model

Response:

`creativity judgment`

Fixed effects:

17 attributes

Random effects:

`rater, style, rater:style`

Random effects	Variance		
rater	80.06		
style	69.77		
rater : style	44.32		
residuals	315.39		

Unadjusted intra-class correlation = 0.30

Fixed effects	Estimate	Standard error	t-value
(Intercept)	41.43	4.95	8.36
visual harmony	-2.59	0.47	-5.49
color saturation	1.29	0.33	3.86
depth	1.38	0.37	3.75
abstraction	1.10	0.57	1.92
symbolism	4.38	0.52	8.50
accurate object	0.56	0.45	1.23
realism imaginative	1.54	0.56	2.73
liveliness animation	-1.35	0.36	-3.76
emotional expressiveness	4.92	0.33	14.91
complexity	3.93	0.40	9.92
valence	2.04	0.33	6.13
color variety	0.74	0.36	2.06
color temperature	0.20	0.33	0.61
color world	-0.49	0.37	-1.32
focus	0.24	0.38	0.65
brushstroke	-1.20	0.42	-2.89
utilization of area	0.39	0.36	1.08

Marginal $R^2 = 0.20$

Approach #2: Additional random forest predictors

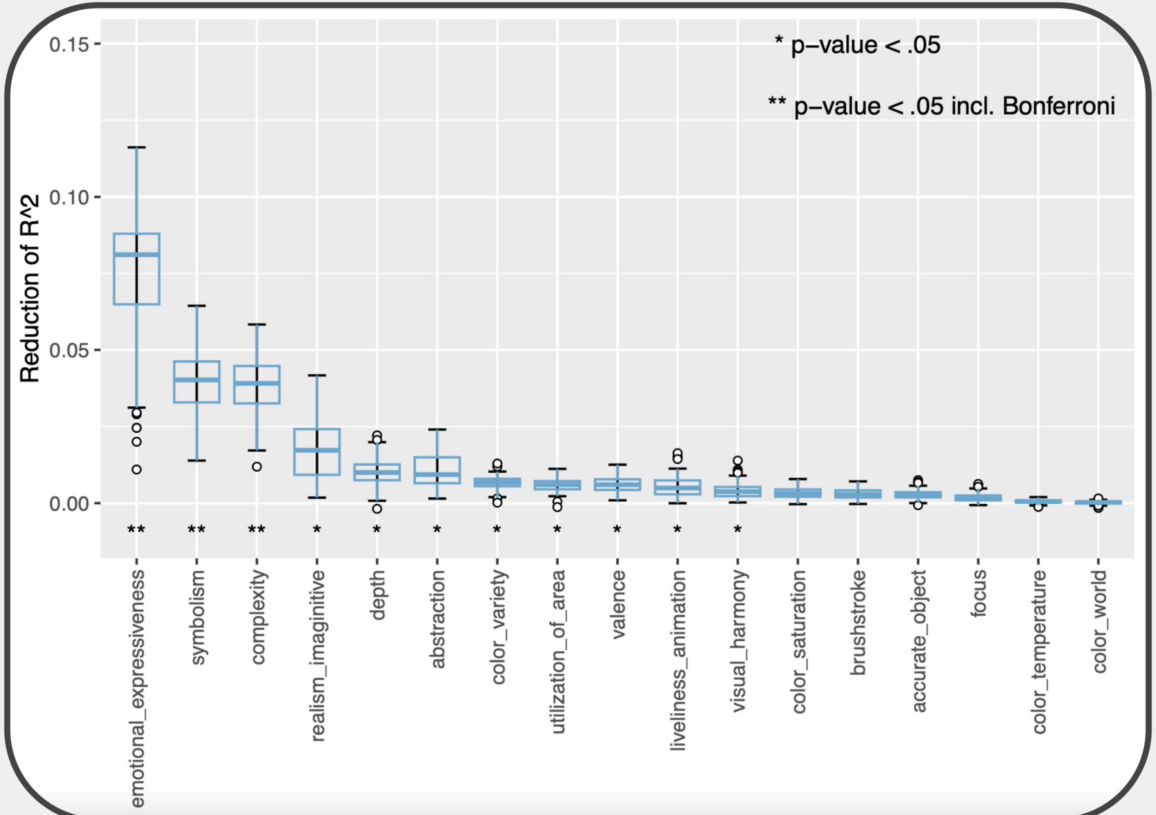
- Two limitations of the linear mixed-effects model
    - Linear assumption, and R^2 on the training set

- Now we include `rater` and `style`, then redo the analysis

- Improvement in prediction R^2 and MAE is negligible, while statistical significance stays the same

- **However**, there is a difference in variable importance!

Approach #2: Additional random forest predictors



Approach #2: Additional RF predictors

- `rater` and `style` explain a non-trivial proportion of total variation (10%)
  + 20% explained by art-related attributes = prediction R^2 (0.3)

- Ordering of most important predictors changes:

symbolism (0.12) > emotionality (0.08) > imaginativeness (0.05) > complexity (0.04)

$\Downarrow$

emotionality (0.08) > symbolism (0.04) > **complexity** (0.04) > imaginativeness (0.02)

- Conclusion: They may not have included all relevant predictors

Discussion

Limitations

- Minimal justification for several design choices
- Failure to account for `rater` and `style`
- Limited generalizability

Closing thoughts

✓ Data ✖ Code ✓ Reproducibility ✓ Replicability

References

Gavin C. Cawley and Nicola L.C. Talbot. “On over-fitting in model selection and subsequent selection bias in performance evaluation”. In: *J. Mach. Learn. Res.* 11 (2010), pp. 2079–2107. issn: 1532-4435.

Marieke Hager et al. “Assessing aesthetic appreciation of visual artworks—The construction of the Art Reception Survey (ARS).” In: *Psychology of Aesthetics, Creativity, and the Arts* 6.4 (2012), p. 320.

Manuela M Marin et al. “Berlyne revisited: Evidence for the multifaceted nature of hedonic tone in the appreciation of paintings and music”. In: *Frontiers in Human Neuroscience* 10 (2016), p. 536.

J. B. Mockus and L. J. Mockus. “Bayesian approach to global optimization and application to multiobjective and constrained problems”. In: *Journal of Optimization Theory and Applications* 70.1 (1991), pp. 157–172. doi: 10.1007/bf00940509. url: https://doi.org/10.1007/ bf00940509.

Shinichi Nakagawa, Paul CD Johnson, and Holger Schielzeth. “The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded”. In: *Journal of the Royal Society Interface* 14.134 (2017), p. 20170213.

Matthew Pelowski, Helmut Leder, and Pablo PL Tinio. “Creativity in the visual arts”. In: *The Cambridge Handbook of Creativity Across Domains* (2017), pp. 80–109.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. “False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant”. In: *Psychological Science* 22.11 (2011). PMID: 22006061, pp. 1359–1366. doi: 10.1177/0956797611417632. url: https://doi.org/10.1177/0956797611417632.

Eva Specker et al. “The Vienna Art Interest and Art Knowledge Questionnaire (VAIAK): A unified and validated measure of art interest and art knowledge.” In: *Psychology of Aesthetics, Creativity, and the Arts* 14.2 (2020), p. 172.

Blanca T. M. Spee et al. “Machine learning revealed symbolism, emotionality, and imaginativeness as primary predictors of creativity evaluations of western art paintings”. In: *Scientific Reports* 13.1 (2023). doi: 10.1038/s41598-023-39865-1. url: https://doi.org/10.1038/ s41598-023-39865-1.

Blanca T.M. Spee et al. “Dataset - How do we identify creative art?” In: (2022). doi: 10. 6084/m9.figshare.19097099.v1. url: https://figshare.com/articles/dataset/Dataset_How_Do_We_Identify_Creative_Art_/19097099.

Jonathan Taylor. Fixed vs. random effects. Lecture slides, Stanford University. 2005.

Eline Van Geert and Johan Wagemans. “Order, complexity, and aesthetic appreciation.” In: *Psychology of Aesthetics, Creativity, and the Arts* 14.2 (2020), p. 135.