Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh

3/7/2025

ML in Practice Reading Group

Presented by Mengying Yan

Uncertainty

Epistemic Uncertainty:

This is knowledge-based uncertainty, arising from limitations in data, model complexity, or unknown factors. It is theoretically reducible with more data or improved models

Aleatoric Uncertainty:

This is inherent randomness within the data or system being modeled. It is considered irreducible and intrinsic to the environment

Tasks in uncertainty quantification

- Abstained prediction
  - A model chooses not to make a prediction when its uncertainty is too high
- Out-of-distribution (OOD) detection
  - Epistemic Uncertainty is high for OOD samples as the model has never seen similar data
- Aleatoric uncertainty quantification
  - E.g. predictive variance in regression; entropy
- Uncertainty disentanglement
  - Estimators that tailored to only one source of uncertainty

Summary of the paper

- First benchmark of uncertainty disentanglement
- Reimplement and evaluate a diverse range of uncertainty quantification methods (19) on ImageNet
- No existing approach provides pairs of disentangled uncertainty estimators in practice
- Provide both practical advice for which uncertainty estimators to use for which specific task – specialized estimators work the best

Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

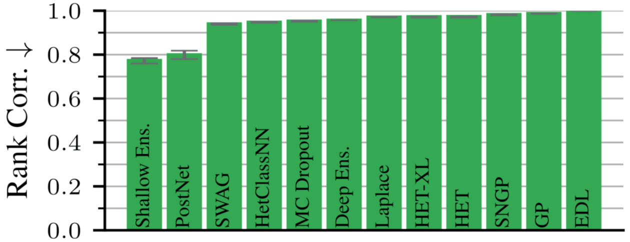
Bálint Mucsányi   Michael Kirchhof   Seong Joon Oh

Summary

- We investigate aleatoric and epistemic uncertainty estimators for ImageNet-1k classification.
- Estimators obtained via decompositions don't work.
- Instead, specialized estimators work best.
- We suggest to define the exact uncertainty task first, then implement an according uncertainty estimator.

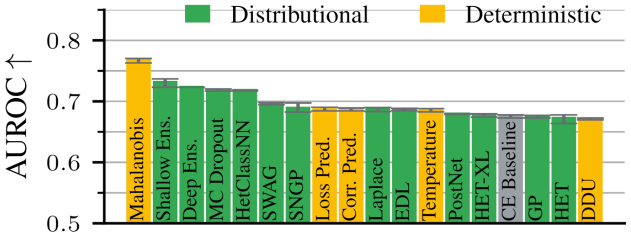
Uncertainty disentanglement fails

Aleatoric/epistemic uncertainties obtained via decomposition formulas are internally correlated, rank corr. ≥ 0.79



Epistemic uncertainty: Specialist wins

Instead, specialized estimators are better. E.g., Mahalanobis is an explicit OOD detector, trained with OOD data.

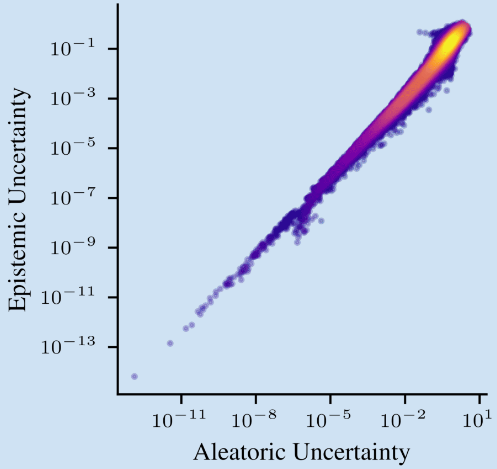


Uncertainty decompositions like this

$$\underbrace{\mathbb{H}_{P(y|x)}(y)}_{\text{predictive}} = \underbrace{\mathbb{E}_{Q(\theta|x)}\left[\mathbb{H}_{P(y|\theta,x)}(y)\right]}_{\text{aleatoric}} + \underbrace{\mathbb{I}_{P(y,\theta|x)}(y;\theta)}_{\text{epistemic}}$$

<u>don't work.</u>

They are highly internally correlated.

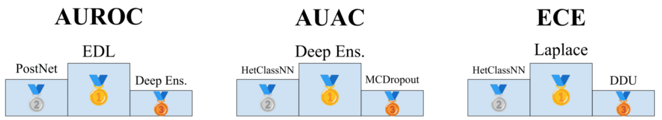


190 figures 📊   Cleanest code 👌   All hyperparams 🤓

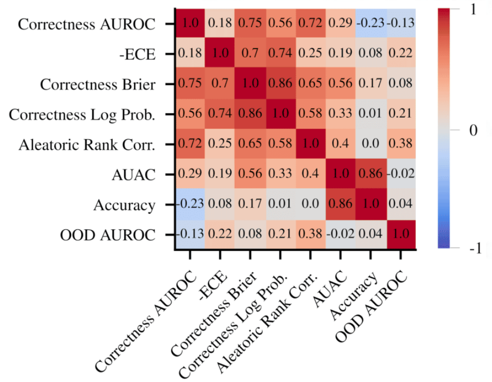
The exact definition of a task/metric matters

E.g., predictive uncertainty: There is not just one best estimator.

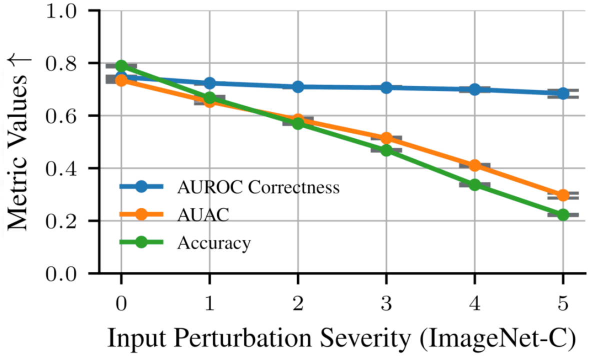
It depends on whether you measure AUROC, AUAC, ECE, …



In general: Model rankings are not highly correlated across tasks.



Good news: Predictive uncertainty is robust



Background

- Giving one predictive uncertainty estimate – relatively easier
- Only theoretical work on uncertainty disentanglement
- Larger scale benchmarks only evaluate one uncertainty component
  - E.g. evaluate OOD samples – higher (epistemic) uncertainty, but such measure could also capture aleatoric uncertainty
- No study that evaluates which component(s) each method captures in practice

Outline

- Benchmark methods
  - Uncertainty quantification methods
  - Uncertainty decomposition formulas
- Experiments
  - Disentanglement using decomposition formulas
  - Epistemic uncertainty
  - Aleatoric uncertainty
  - Predictive uncertainty
- Discussion

Uncertainty Quantification Methods:

Distributional methods

Model a second-order predictive distribution $q(\pi|x)$, where π is class probability

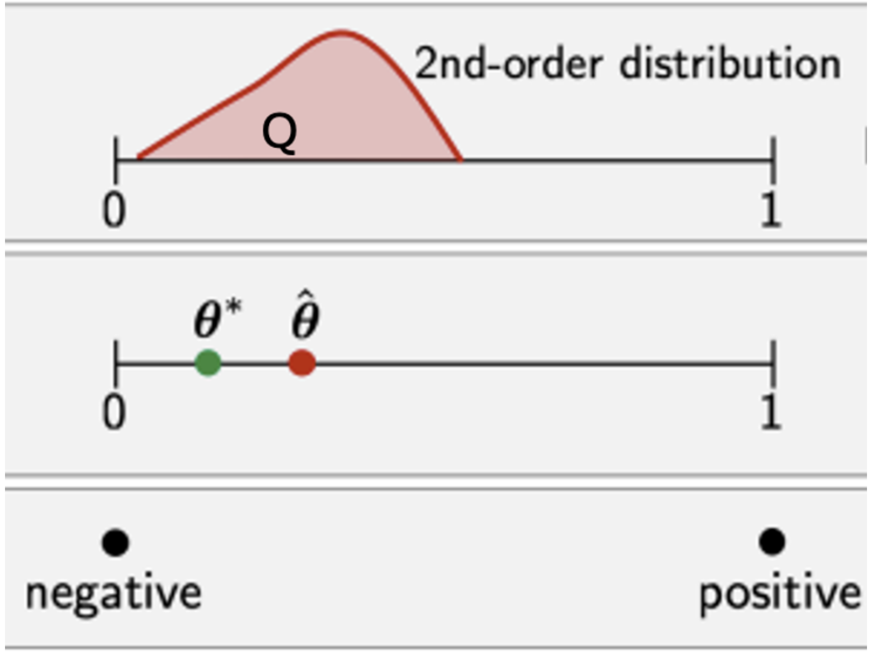
1. Model $q(\pi|x)$ explicitly
    - Approximating a Gaussian process [32]
    - Evidential deep learning (EDL) [45] Aditya presented and PostNet [4]
2. Sample based
    - MC Dropout [12]
    - Deep Ensemble [29]

$$\boxed{p(y|\mathbf{x}) = M^{-1} \sum_{m=1}^{M} p_{\theta_m}(y|\mathbf{x}, \theta_m)}$$

Use uncertainty aggregators to compile distributions into scalar uncertainty estimates – e.g. Bayesian Model Average and use entropy as uncertainty estimate

$$\bar{\boldsymbol{\pi}}(\boldsymbol{x}) := \mathbb{E}_{q(\boldsymbol{\pi}|\boldsymbol{x})}\left[\boldsymbol{\pi}\right]$$

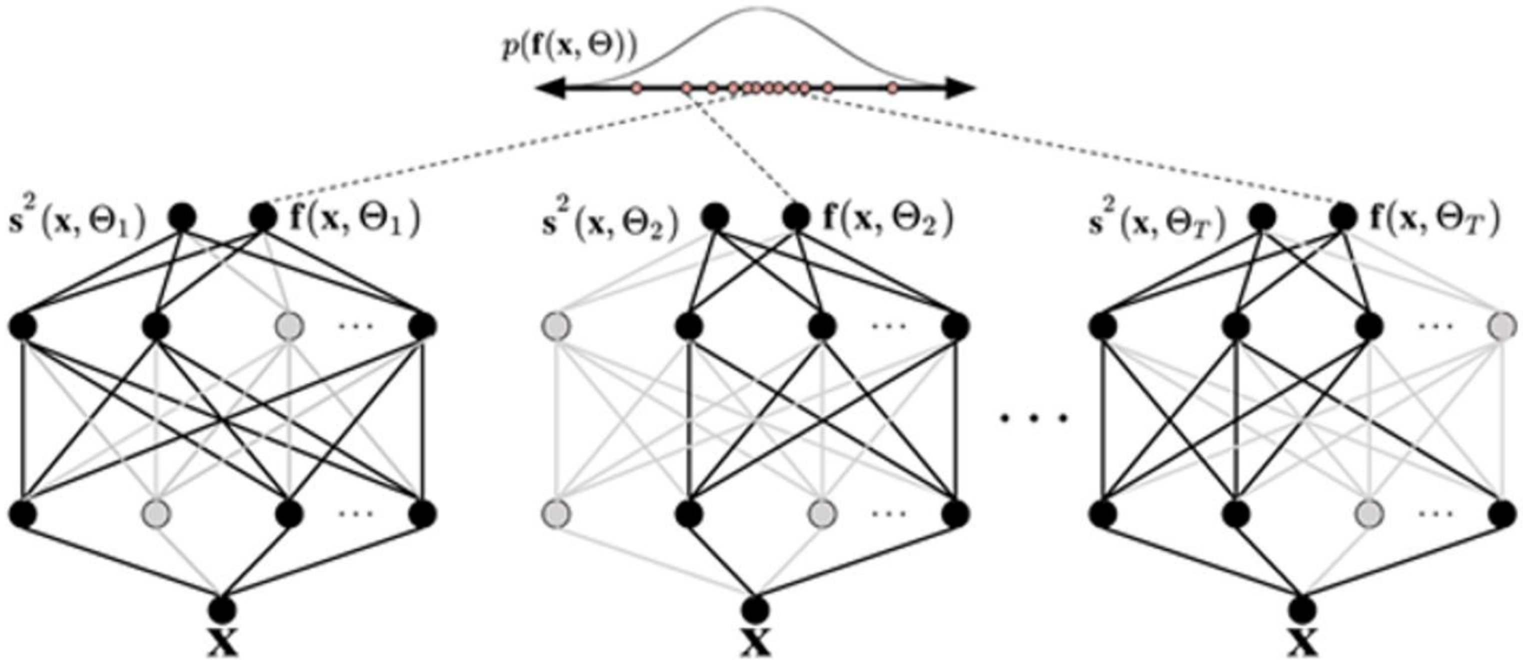
Uncertainty in binary classification



In Bayesian inference, Q is given by the posterior distribution

[57] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In Uncertainty in Artificial Intelligence, pages 2282–2292. PMLR, 2023.

MC Dropout


$p(\mathbf{f}(\mathbf{x}, \Theta))$
$\mathbf{s}^2(\mathbf{x}, \Theta_1)$
$\mathbf{f}(\mathbf{x}, \Theta_1)$
$\mathbf{s}^2(\mathbf{x}, \Theta_2)$
$\mathbf{f}(\mathbf{x}, \Theta_2)$
$\mathbf{s}^2(\mathbf{x}, \Theta_T)$
$\mathbf{f}(\mathbf{x}, \Theta_T)$
$\mathbf{x}$
$\mathbf{x}$
$\mathbf{x}$

Uncertainty Quantification Methods:
Deterministic Methods

Directly output scalar uncertainty estimates $u(x)$

- Cross-entropy (CE) baseline: NN trained with CE loss

- Mahalanobis method [30] for detecting OOD samples
  - Measure how close training and test samples are

$$d_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$\mathrm{Confidence}(x) = -\min_c d_M(x, \mu_c)$$

Uncertainty Decomposition

- Information-Theoretical (IT) decomposition
  - Entropy of predictive distribution

$$H(y \mid x) = -\sum_{i=1}^{C} p(y_i \mid x) \log p(y_i \mid x)$$

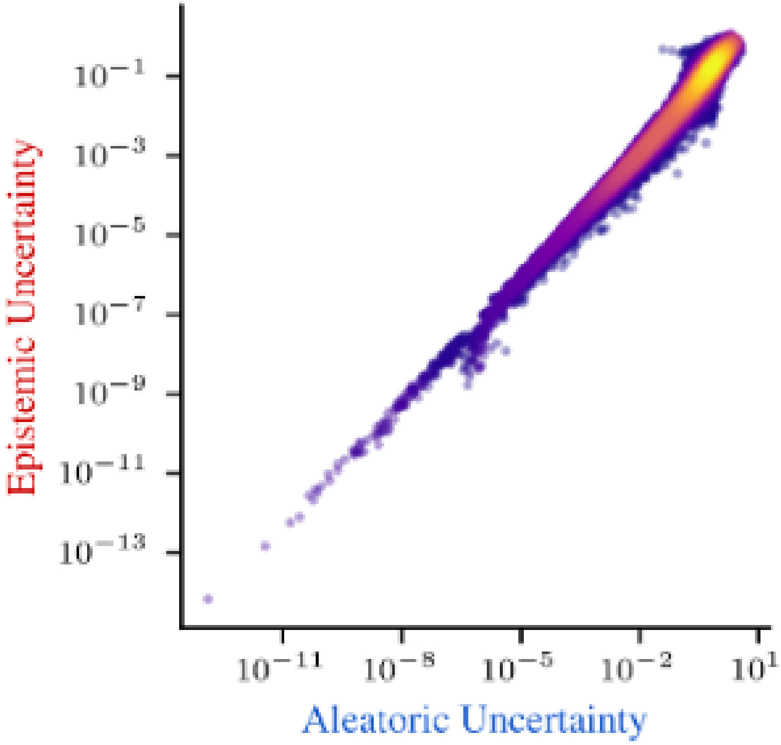
Low entropy: one class has high prob

$$\underbrace{\mathbb{H}_{p(y|\boldsymbol{x})}(y)}_{\text{predictive}} = \underbrace{\mathbb{E}_{q(\boldsymbol{\pi}|\boldsymbol{x})}\left[\mathbb{H}_{p(y|\boldsymbol{\pi},\boldsymbol{x})}(y)\right]}_{\text{aleatoric}} + \underbrace{\mathbb{I}_{p(y,\boldsymbol{\pi}|\boldsymbol{x})}(y;\boldsymbol{\pi})}_{\text{epistemic}},$$

Expected entropy: spread of the labels that the plausible predictions in the posterior have on average

Mutual information: (measure of the mutual dependence between the two variables)
captures the disagreement of the predictions p(y | π, x) in the second order predictive distribution q(π | x)

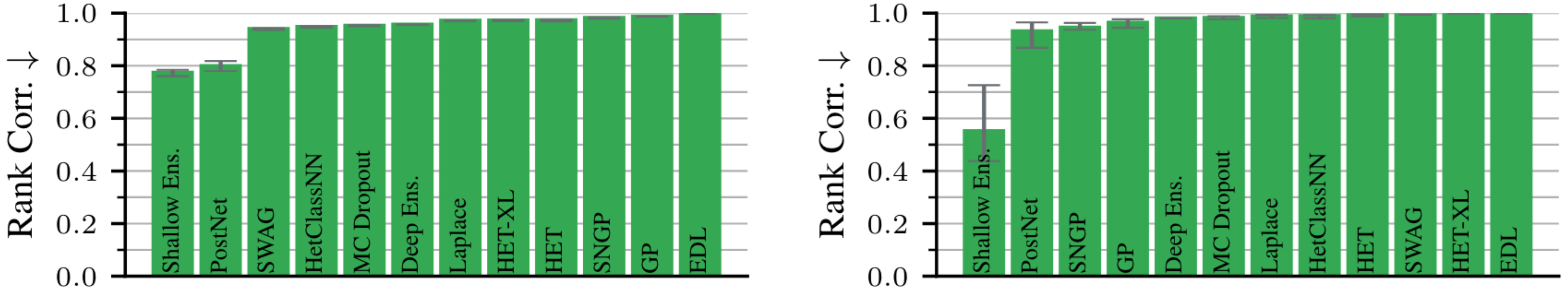
Does any approach give disentangled uncertainty estimators?



(Decomposing deep ensemble uncertainties)

- Decomposition Formulas Fail to Disentangle Aleatoric and Epistemic Uncertainty
- Strongly correlated

Does any approach give disentangled uncertainty estimators?



(a) ImageNet results. All twelve distributional methods exhibit a high rank corr. (≥ 0.78).

(b) CIFAR-10 results. Eleven out of twelve distributional methods exhibit a strong rank corr. (≥ 0.93).

- ImageNet has a level of inevitable correlation between epistemic and aleatoric uncertainty estimates (has fewer samples with higher aleatoric uncertainty)
- But there are estimators that perform well on one uncertainty (show later) – decomposition is not the best choice

Epistemic Uncertainty: Specialized Uncertainty Estimators Detect OOD Inputs the Best

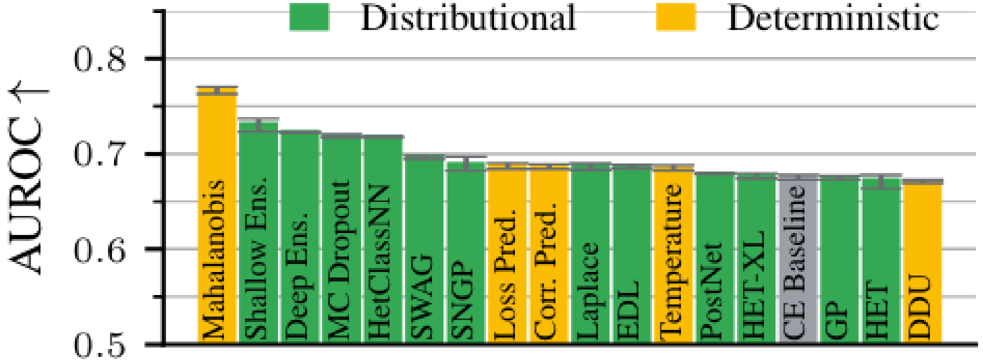
Evaluated on the OOD detection proxy task (standard way to evaluate)

- Label 0: in-distribution samples
- Label 1: out of distribution samples
- Uncertainty should be higher for out of distribution samples
- ImageNet-C severity level 2 as OOD

AUROC

Best: Mahalanobis method

- Specifically trained on OOD



(a) OOD detection AUROC results. OOD samples are perturbed by ImageNet-C corruptions of severity two. Mahalanobis, the best method, is trained specifically to distinguish OOD data of this severity.

ImageNet-C

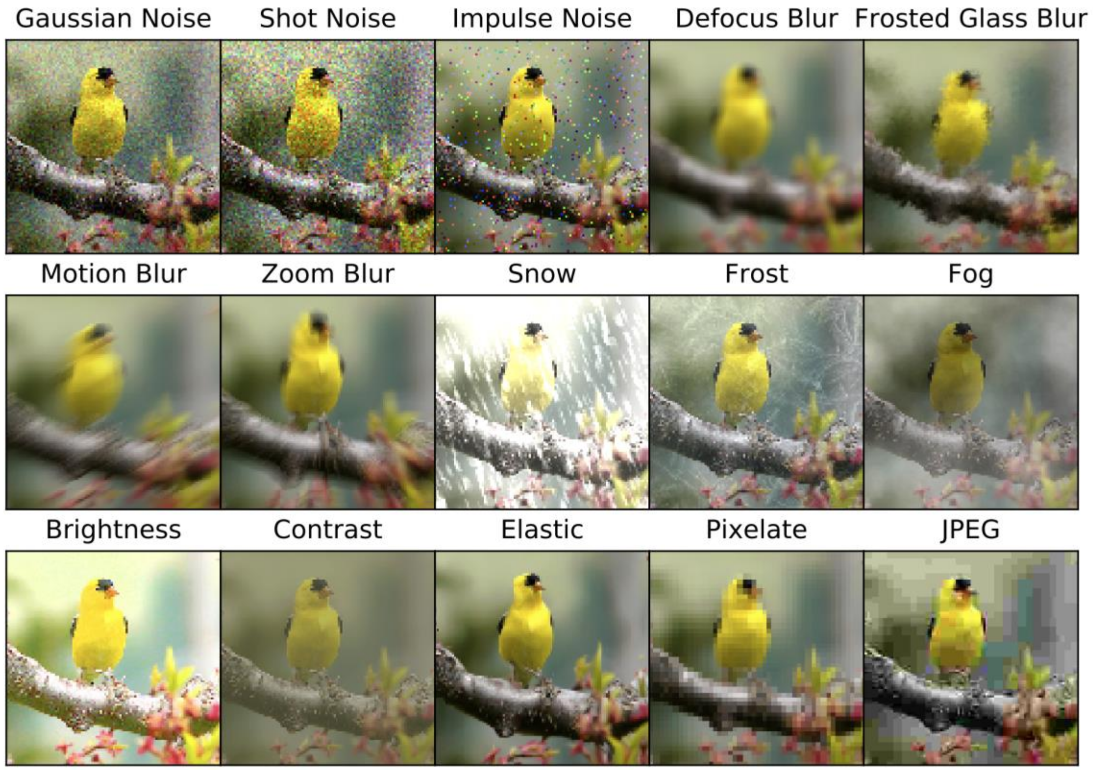


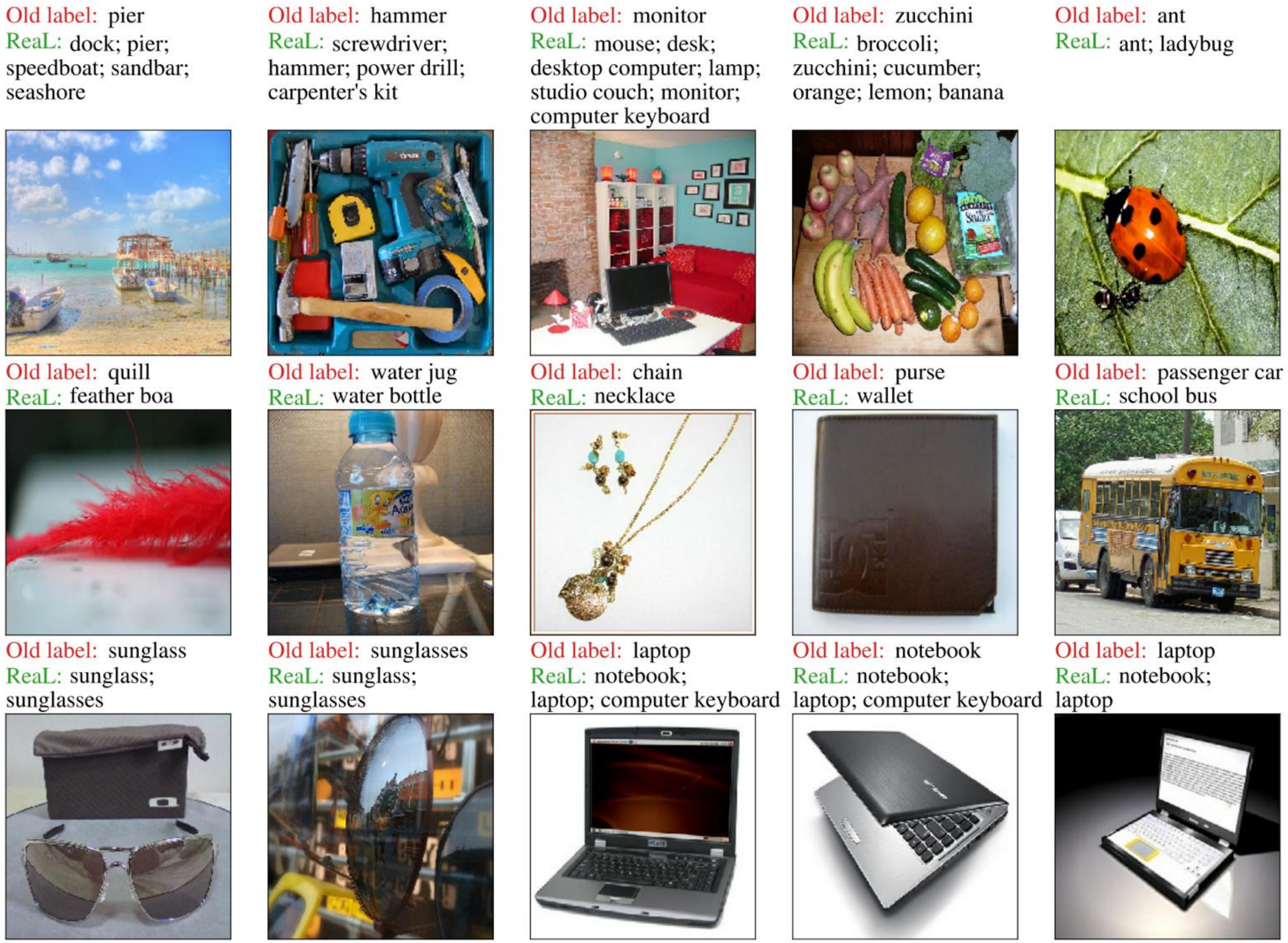
Figure 1: Our IMAGENET-C dataset consists of 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. See different severity levels in Appendix B.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261, 2019.

Aleatoric Uncertainty: No Method With Outstanding Performance

Indicators for aleatoric uncertainty

- Disagreement of human annotators as ground truths (GT) for aleatoric uncertainty

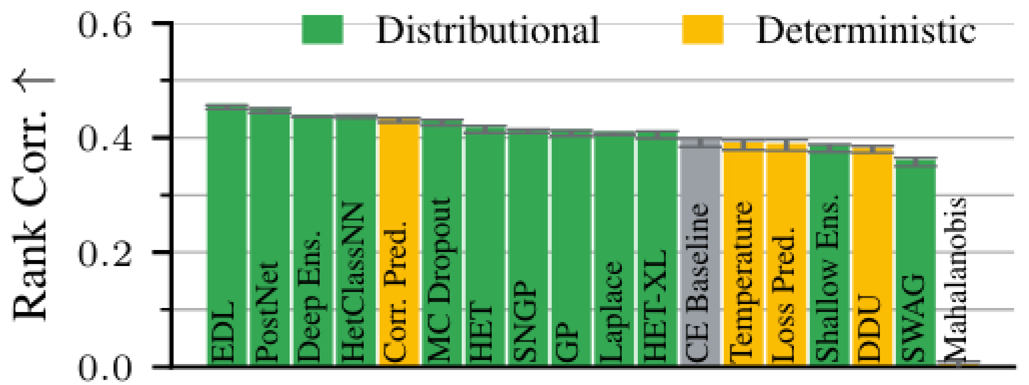


ImageNet-ReaL: Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., & Oord, A. van den. (2020). *Are we done with ImageNet?* (arXiv:2006.07159). arXiv. https://doi.org/10.48550/arXiv.2006.07159

Aleatoric Uncertainty: No Method With Outstanding Performance

Rank correlation

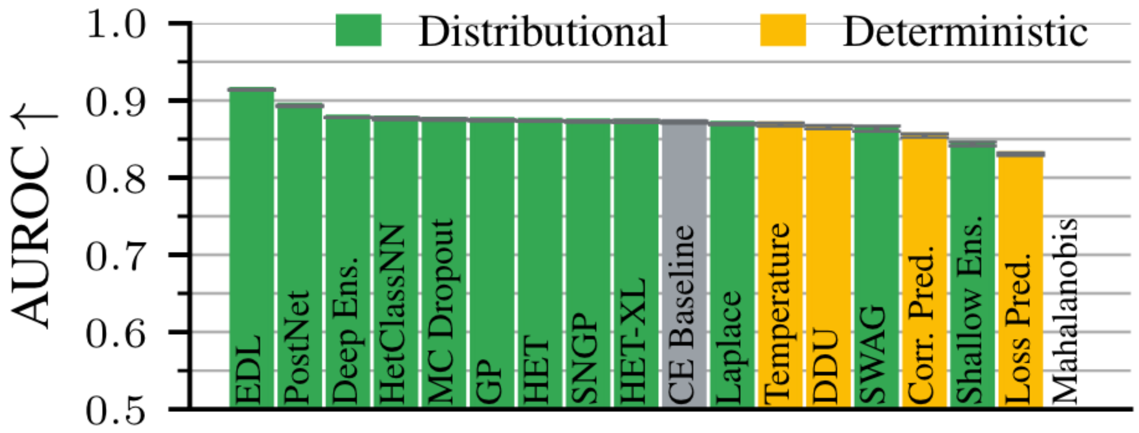
- Best achievable rank correlation is unknown (not 1 due to ties)
- Further performance gains are far from saturated
- Insight into disentanglement
  - Mahalanois: non-informative of aleatoric uncertainty
  - Pair it with an estimator for aleatoric uncertainty (e.g. CE, they have low rank corr 0.15)



(b) Rank correlation of uncertainty estimators and the GT aleatoric uncertainty on ImageNet. The entropy of the ImageNet-ReaL label distributions is used as GT aleatoric uncertainty.

Predictive Uncertainty:

- Correctness prediction
  - Wrong predictions have higher uncertainties
  - Measured by AUROC

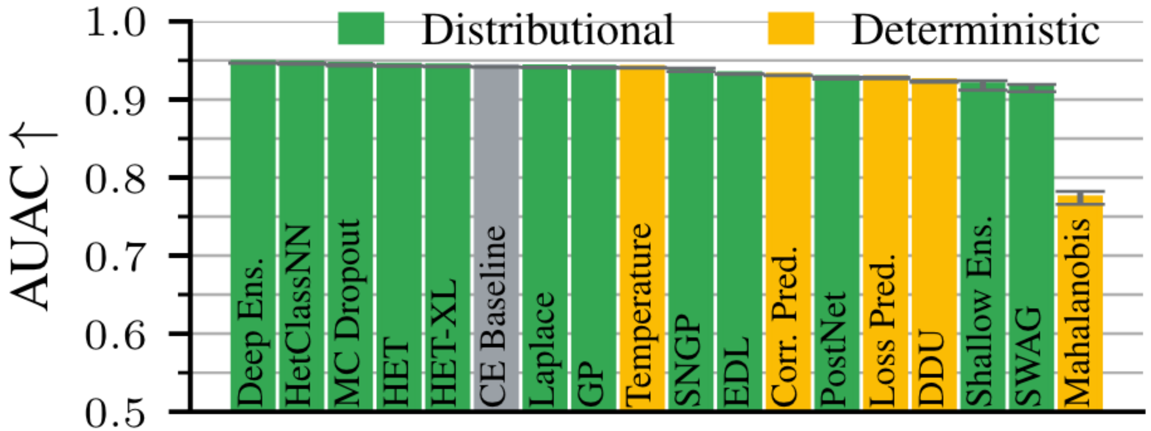


(a) ID correctness prediction results measured by the AUROC w.r.t. model correctness. The evidential deep learning methods, EDL and PostNet, capture predictive uncertainty remarkably well.

Predictive Uncertainty Tasks

- Abstained prediction
  - Refusing to predict on the x% most uncertain examples, and calculate accuracy on the rest
  - Measured by area under the Accuracy-Coverage curve (AC)
    - X-axis: x% abstained samples
    - Y-axis: accuracy on the rest
  - Saturated most >0.91

Different tasks: detect errors or reduce errors?



(b) Abstained prediction results using the AUAC metric. Most methods are within a 0.03 AUAC band. EDL and PostNet lose their advantage as their accuracy is lower.

Predictive Uncertainty Task

- Calibration
  - Predicted probabilities close to truth
  - Measured by ECE

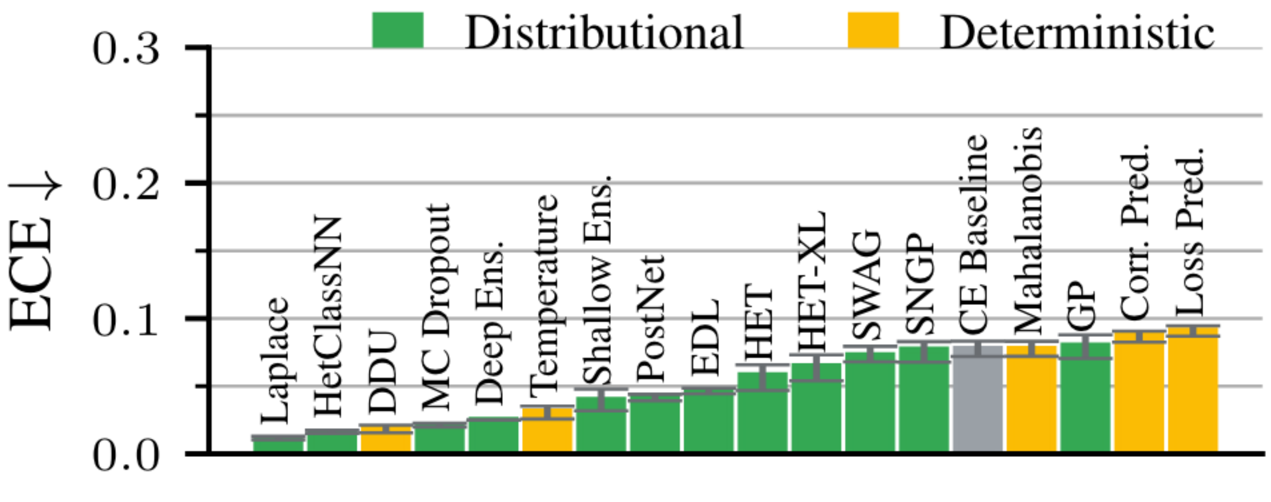
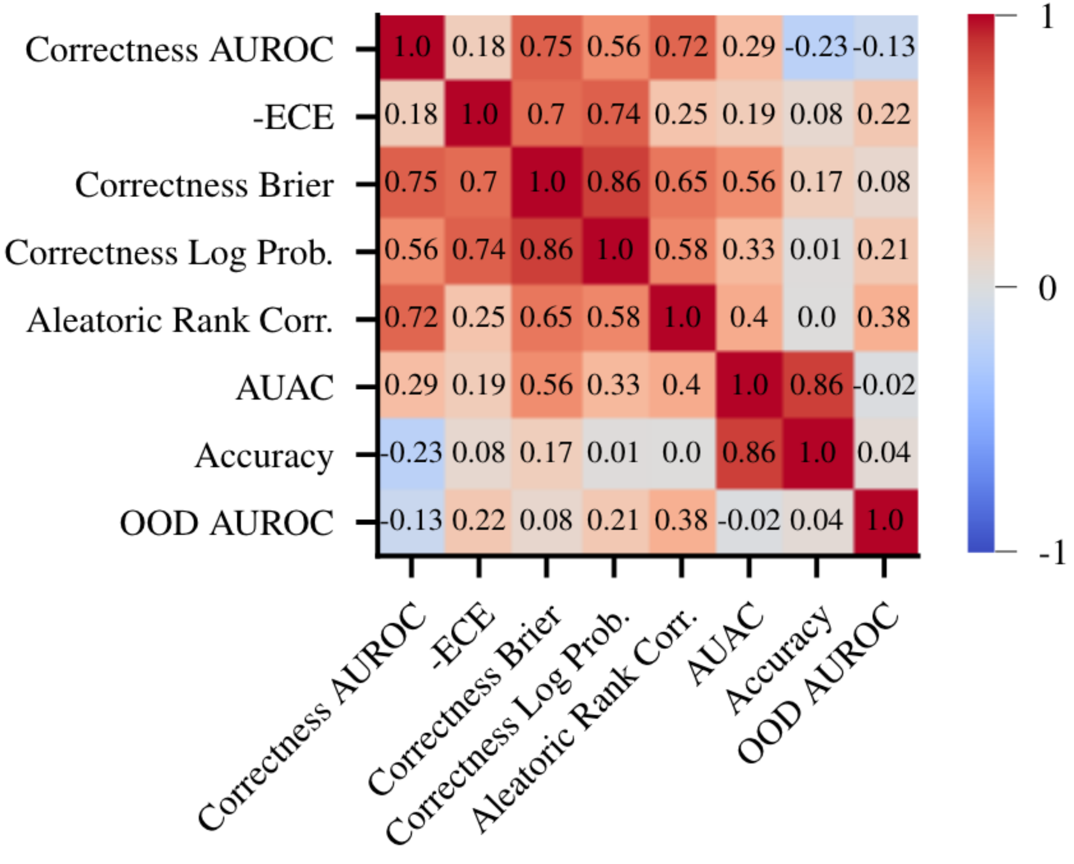


Figure 5: Expected calibration error on ImageNet.

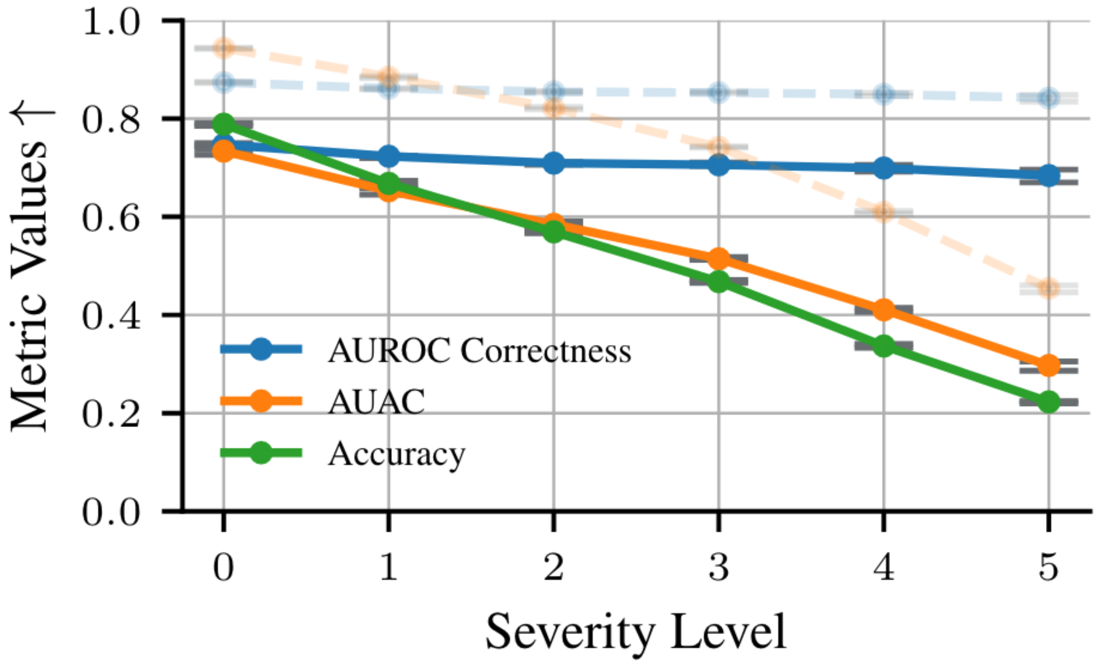
Comparing different Tasks

- Pearson correlation of metric pairs
- No one-fits-all uncertainty estimator
- The Best Method Depends on the Precise Task

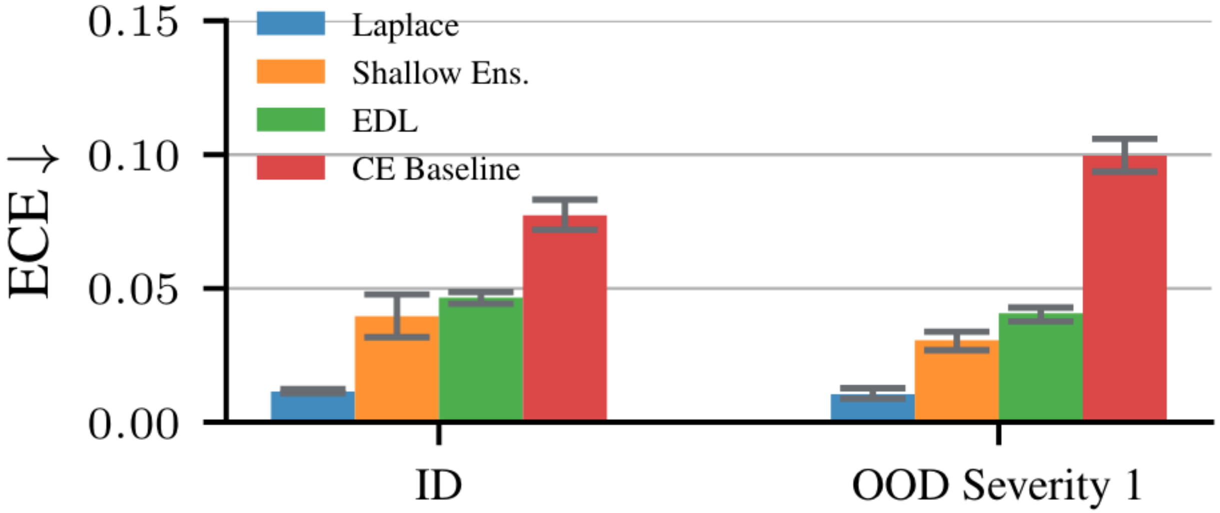


	Correctness AUROC	-ECE	Correctness Brier	Correctness Log Prob.	Aleatoric Rank Corr.	AUAC	Accuracy	OOD AUROC
Correctness AUROC	1.0	0.18	0.75	0.56	0.72	0.29	-0.23	-0.13
-ECE	0.18	1.0	0.7	0.74	0.25	0.19	0.08	0.22
Correctness Brier	0.75	0.7	1.0	0.86	0.65	0.56	0.17	0.08
Correctness Log Prob.	0.56	0.74	0.86	1.0	0.58	0.33	0.01	0.21
Aleatoric Rank Corr.	0.72	0.25	0.65	0.58	1.0	0.4	0.0	0.38
AUAC	0.29	0.19	0.56	0.33	0.4	1.0	0.86	-0.02
Accuracy	-0.23	0.08	0.17	0.01	0.0	0.86	1.0	0.04
OOD AUROC	-0.13	0.22	0.08	0.21	0.38	-0.02	0.04	1.0

Uncertainties are Robust to Distribution Shifts



(a) ImageNet results. The uncertainty estimators' performance in terms of the AUROC degrades much slower than the model's accuracy.



(a) ImageNet results. Methods are generally more robust to OOD perturbations than on CIFAR-10. EDL and Shallow Ensemble even become better calibrated OOD.

Discussion

- Specified uncertainties for specified tasks
- Decomposition does not work in practice, instead, use a pair of estimators for epistemic and aleatoric uncertainties
- Evaluation on more datasets are needed. CIFAR-10 results are slightly different from ImageNet
- Aleatoric uncertainty lacks a standardized testing protocol, and no method can give highly accurate estimates yet
- Predictive uncertainties are saturated

Recommendations

- Key takeaway: specify uncertainty task
- Uncertainty evaluation
- Reading? Yes, focus on key messages
- Implementation of 19 methods and results
  - https://github.com/bmucsanyi/untangle