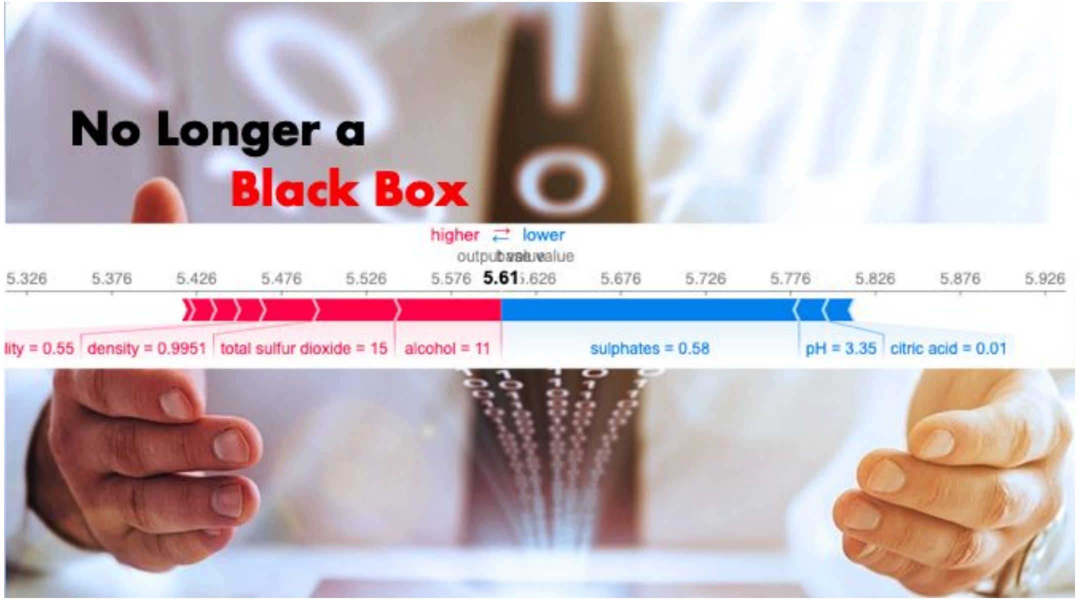
A Unified Approach to Interpreting Model Predictions

Teerapong Panboonyuen, Ph.D.

https://kaopanboonyuen.github.io/

R e f e r e n c e

Lundberg, Scott M., and Su-In Lee.
"A unified approach to interpreting model predictions."
Advances in neural information processing systems. 2017.



https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

Lundberg and Lee (NIPS 2017)

SHAP
SHAPLEY ADDITIVE EXPLANATIONS

Lundberg and Lee (NIPS 2017) - Main (1)

They proposed the **SHAP value** as a united approach to explaining the output of any machine learning model.

Lundberg and Lee (NIPS 2017) - Main (1)

They proposed the **SHAP value** as a united approach to explaining the output of any machine learning model.

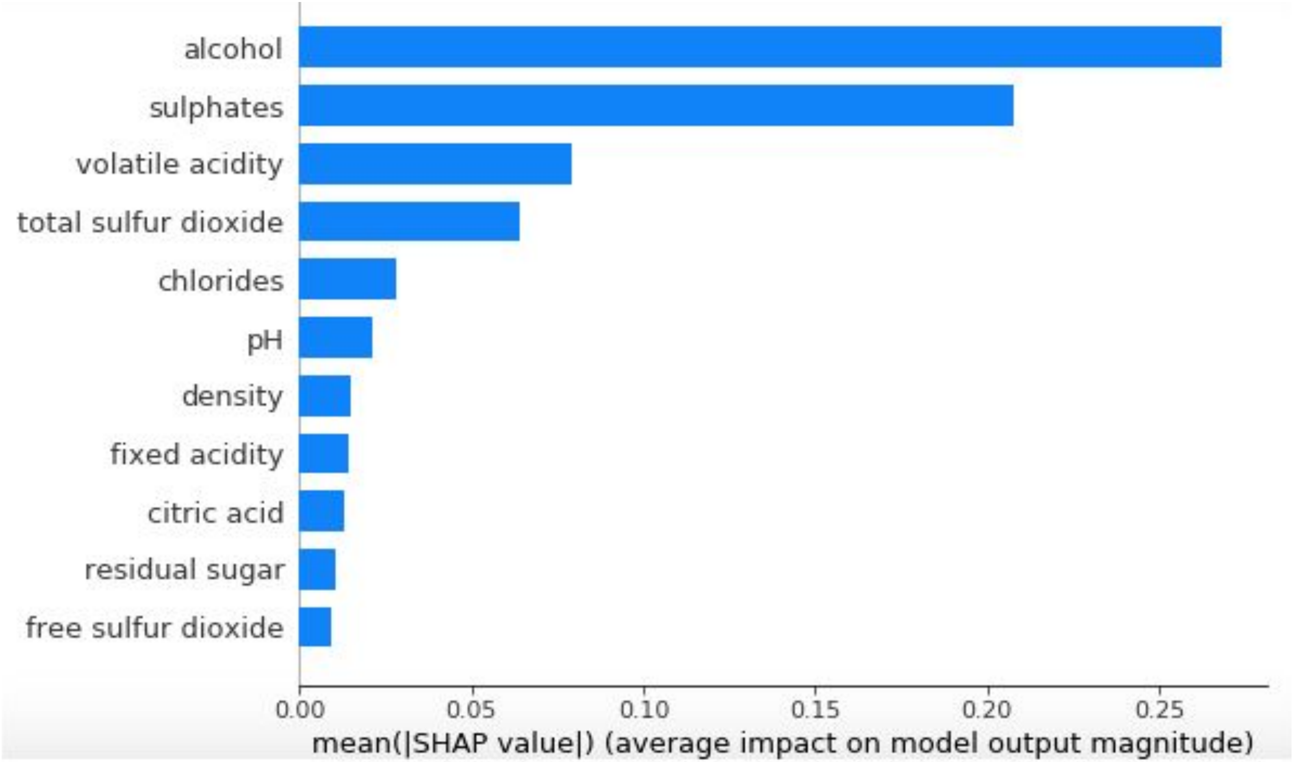
Three benefits worth mentioning here.

The first one is **global interpretability**

— the collective SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable.

(a) This is like the variable importance plot but it is able to show the positive or negative relationship for each variable with the target

Variable Importance Plot



Lundberg and Lee (NIPS 2017) - Main (1)

They proposed the **SHAP value** as a united approach to explaining the output of any machine learning model.

The second benefit is ***local interpretability*** — each observation gets its own set of SHAP values. This greatly increases its transparency.

(a) We can explain why a case receives its prediction and the contributions of the predictors. Traditional variable importance algorithms only show the results across the entire population but not on each individual case.

(b) The local interpretability enables us to pinpoint and contrast the impacts of the factors.

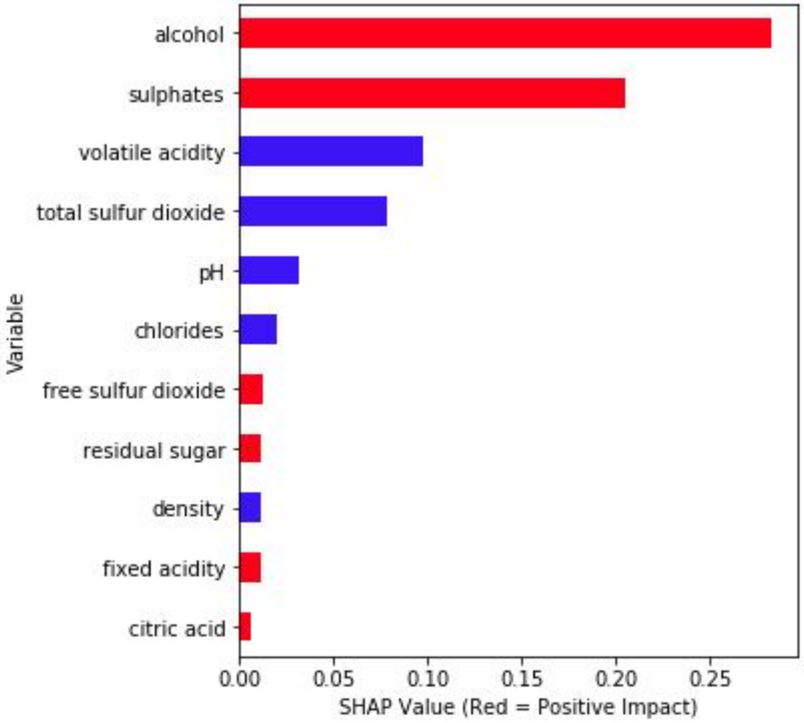
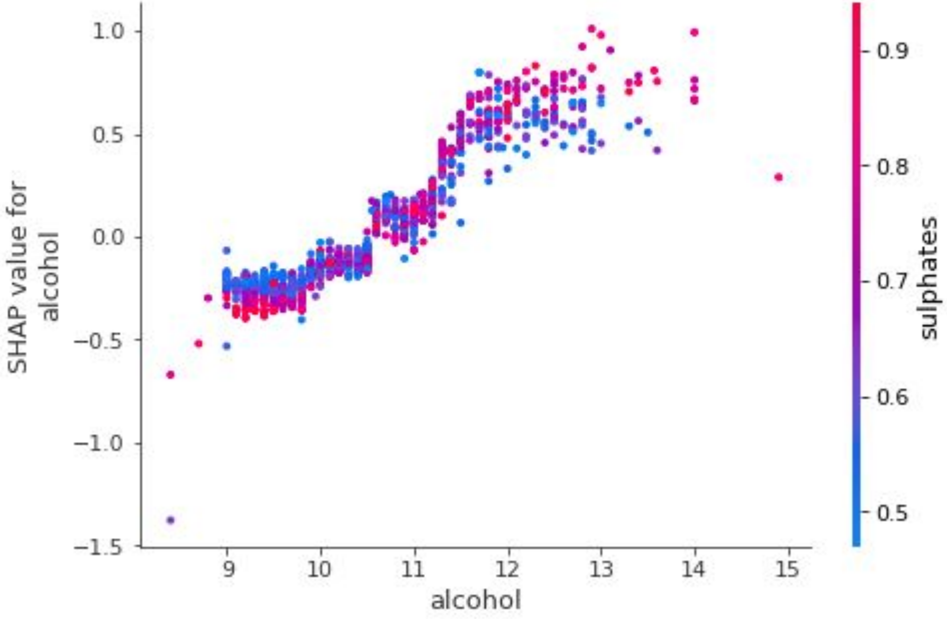


Exhibit (K.1): The simplified version



The SHAP Dependence Plot

NOTE

SHAP values can be calculated for **any tree-based model**,

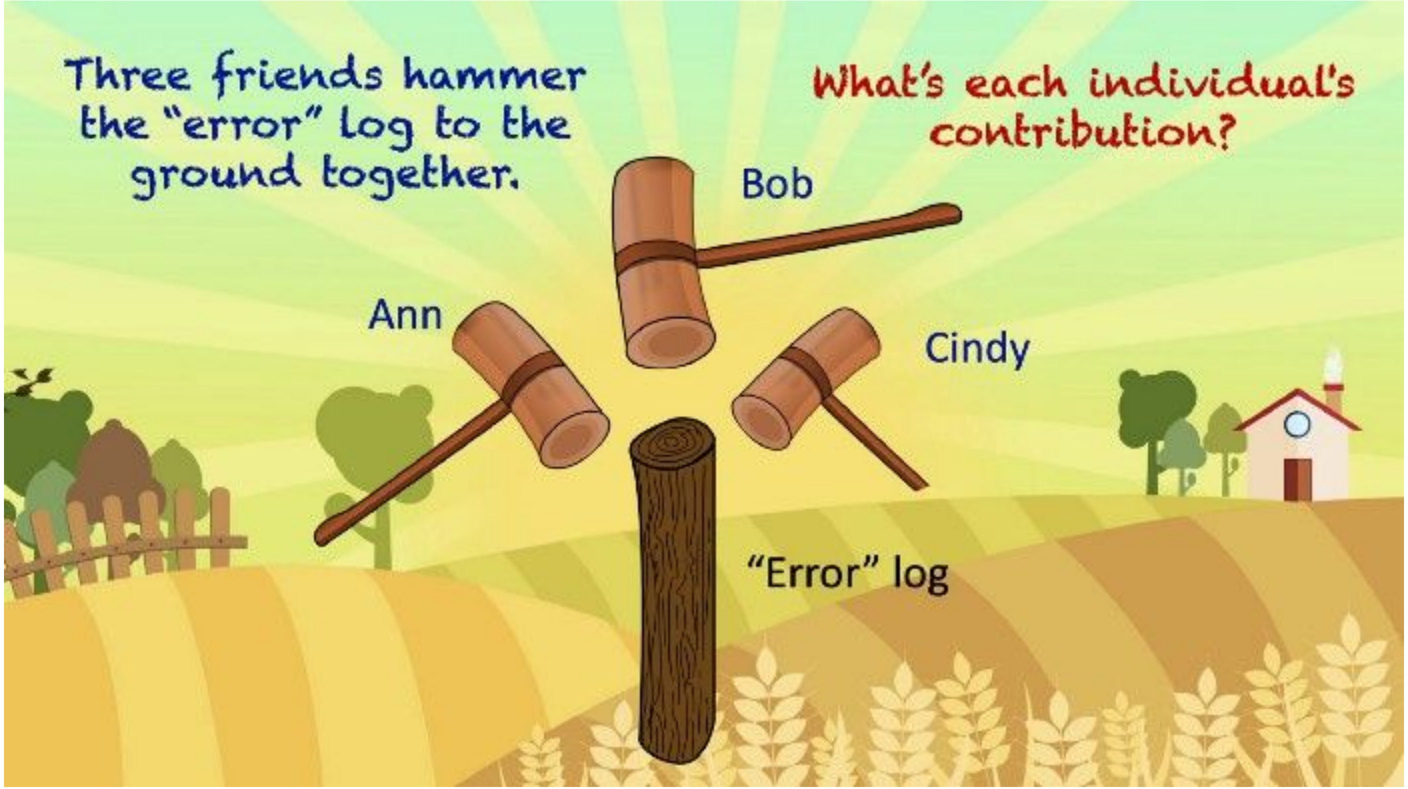
while other methods use linear regression or logistic regression models

as the ***surrogate models.***

Chris Kuo, Ph.D.| Columbia University

Model Interpretability Does Not Mean Causality.

What is the Shapley Value?



Let me explain the Shapley value with a story

Assume Ann, Bob, and Cindy together
were hammering an “error” wood log, 38 inches, to the ground.

After work, they went to a local bar for a drink and I,
a mathematician, came to join them.

I asked a very bizarre question:
“What is everyone’s contribution (in inches)?”

That's the way to calculate the Shapley value

It is the average of the marginal contributions across all permutations.

How do we **measure the contributions of the hammers** (predictors)?

The Shapley values!

	Marginal contribution			inches
Combination	Ann	Bob	Cindy	Total
A, B, C	2	32	4	38
A, C, B	4	34	0	38
B, A, C	2	32	4	38
B, C, A	0	28	10	38
C, A, B	2	36	0	38
C, B, A	0	28	10	38
Average	2	32	4	38

Data Visualization and Model Explainability

Example: with my code

(A) Variable Importance Plot — Global Interpretability

(B) SHAP Dependence Plot — Global Interpretability

(C) Individual SHAP Value Plot — Local Interpretability

Alcohol: has a positive impact on the quality rating. The alcohol content of this wine is 11.8 (as shown in the first row of Table B) which is higher than the average value 10.41. **So it pushes the prediction to the right.**

Example: with my code

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	predict
1575	7.5	0.520	0.40	2.2	0.060	12.0	20.0	0.99474	3.26	0.64	11.8	6.20
1087	7.9	0.190	0.42	1.6	0.057	18.0	30.0	0.99400	3.29	0.69	11.2	6.18
1031	7.3	0.550	0.01	1.8	0.093	9.0	15.0	0.99514	3.35	0.58	11.0	5.61
										0.60	11.2	4.67
										0.54	9.4	5.16
										0.80		
893	7.2	0.660	0.03	2.3	0.078	16.0	86.0	0.99743	3.53	0.57		

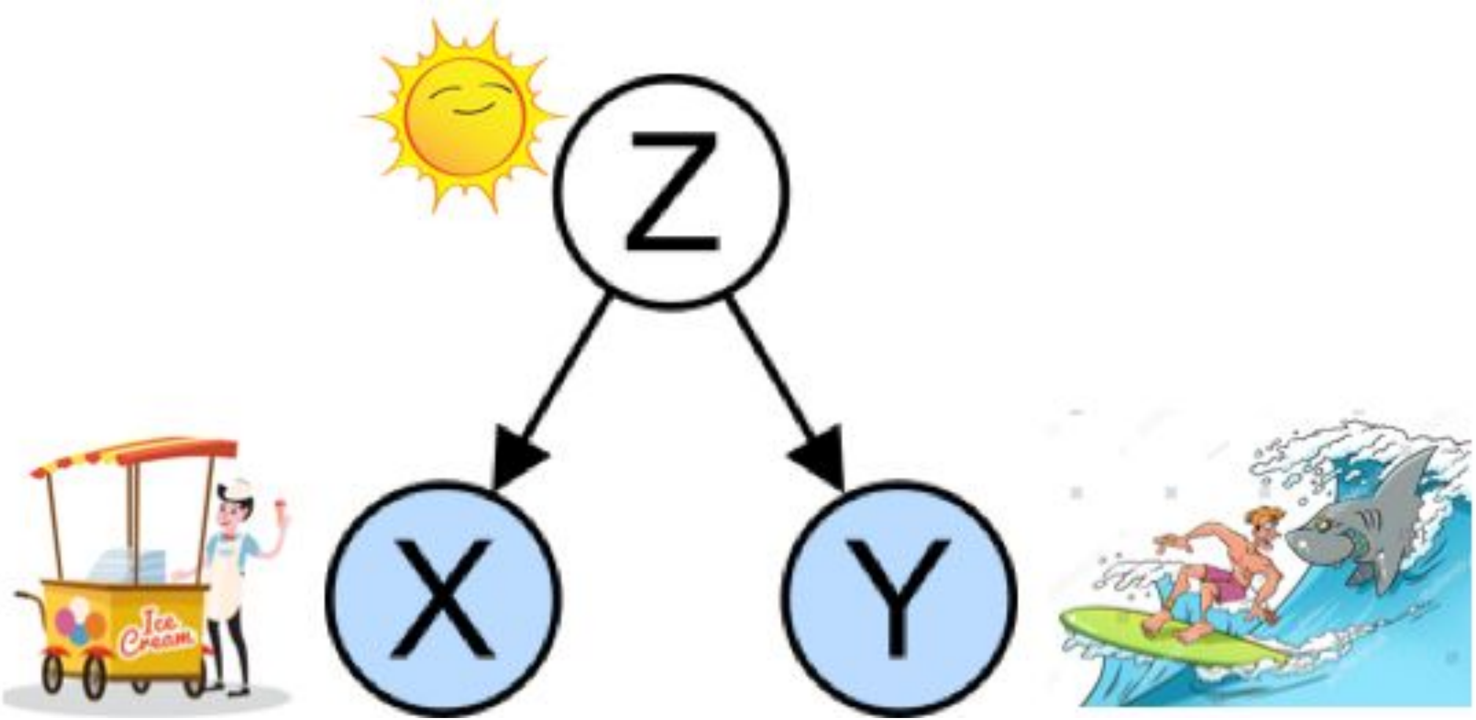
- pH: has a negative impact on the quality rating. A lower than the average pH (=3.26 < 3.30) **drives the prediction to the right**.
- Sulphates: is positively related to the quality rating. A lower than the average Sulphates (= 0.64 < 0.65) pushes the prediction to the left.

```
fixed acidity            8.310164
volatile acidity         0.527392
citric acid              0.268444
residual sugar           2.508444
chlorides                0.087823
free sulfur dioxide     15.885066
total sulfur dioxide    46.455043
density                  0.996726
pH                       3.308702
sulphates                0.658053
alcohol                 10.416302
dtype: float64
```

Correlation Can Mean Causation — Only When Certain Conditions Meet

Inferring Causation from Correlation Is a Scary Thing

Confounding Factor



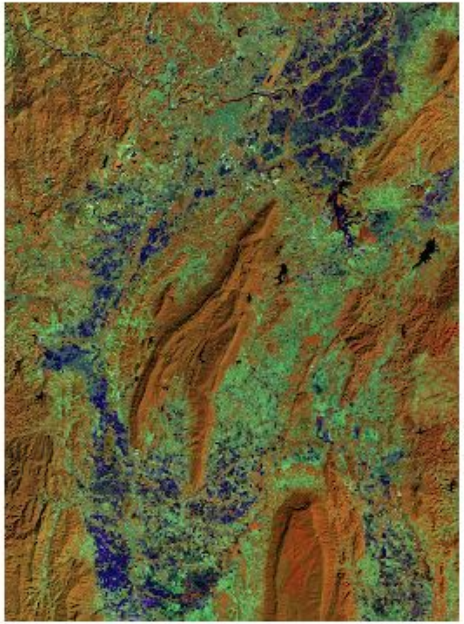
Efficient Uncertainty Estimation for Semantic Segmentation

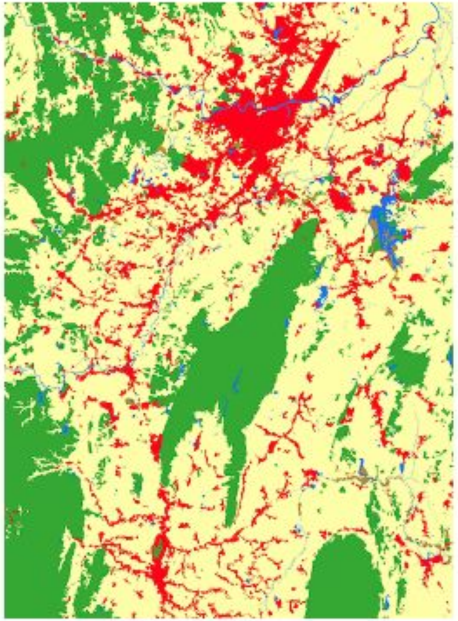
Teerapong Panboonyuen, Ph.D.

https://kaopanboonyuen.github.io/

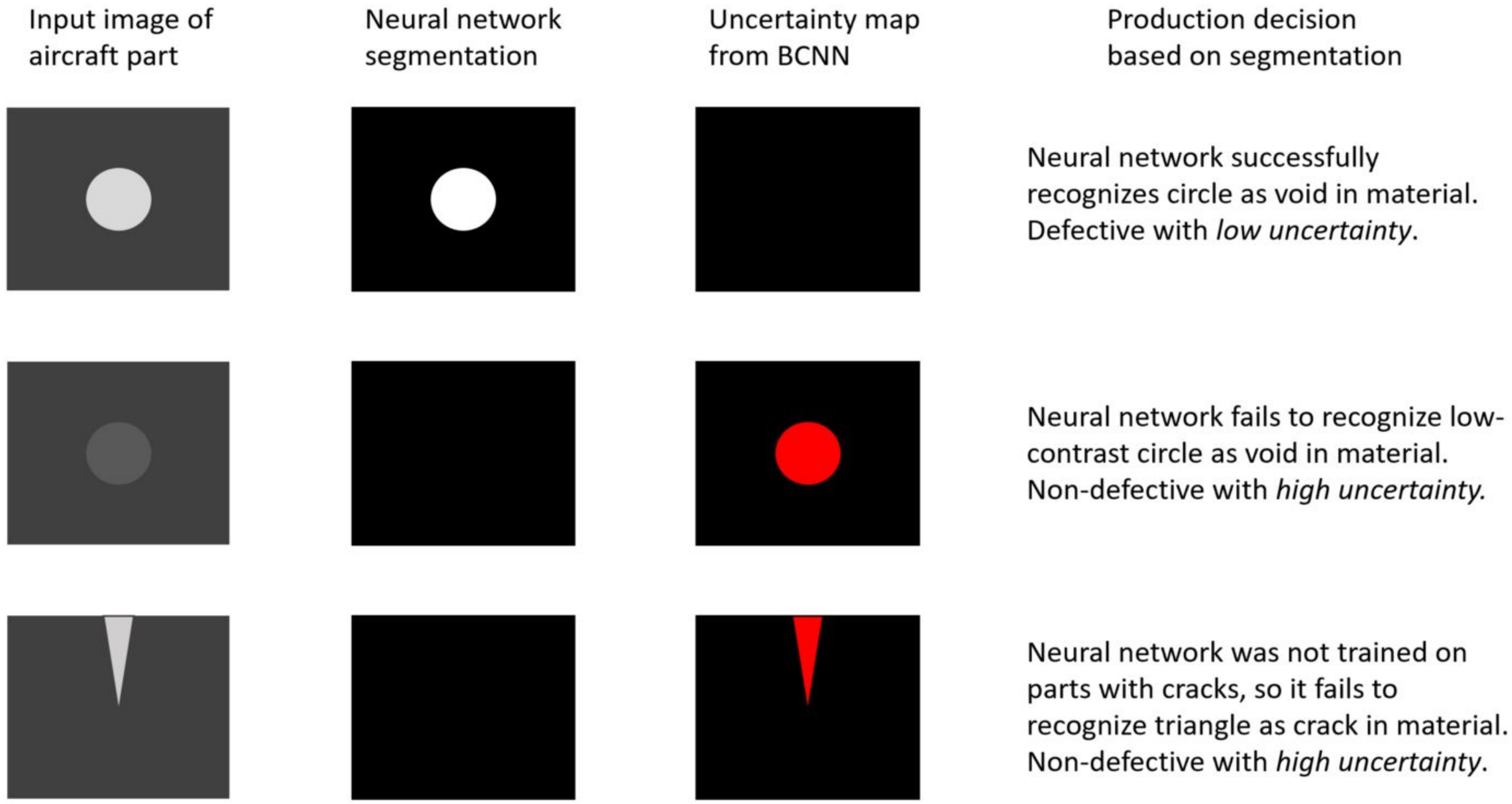
Public and Private Corpora

Private corpus (GISTDA Nan Province Corpus)

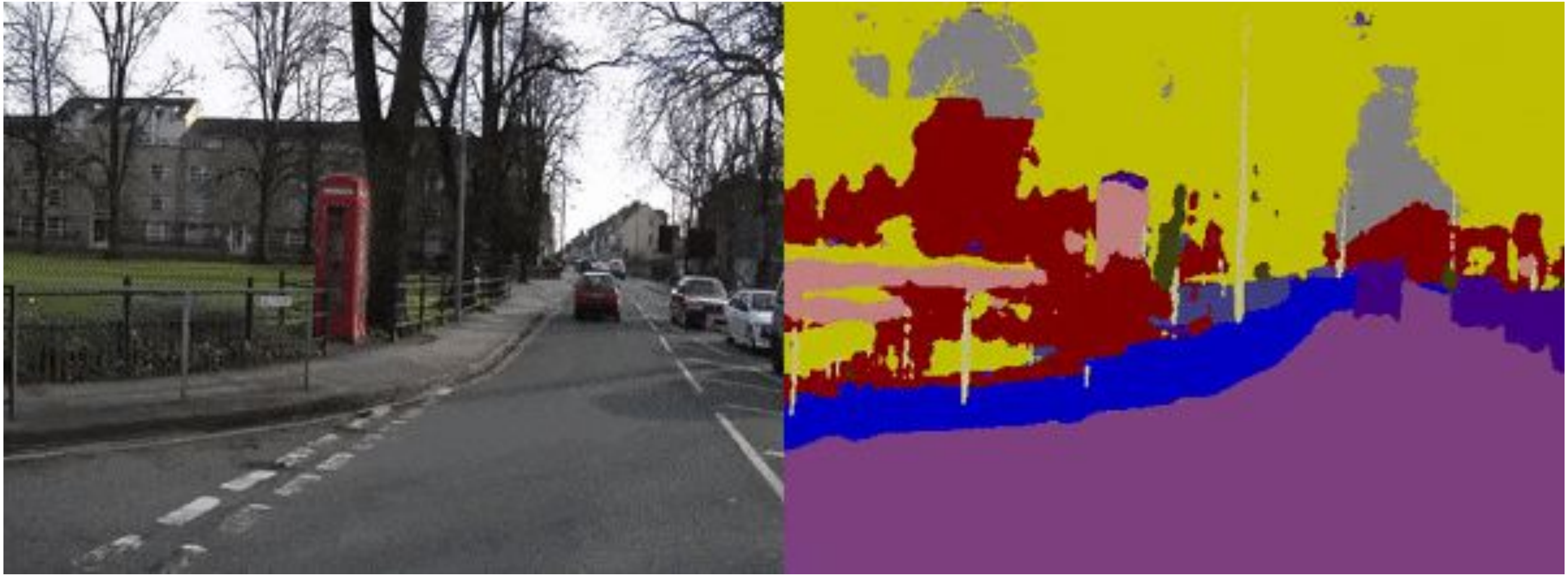




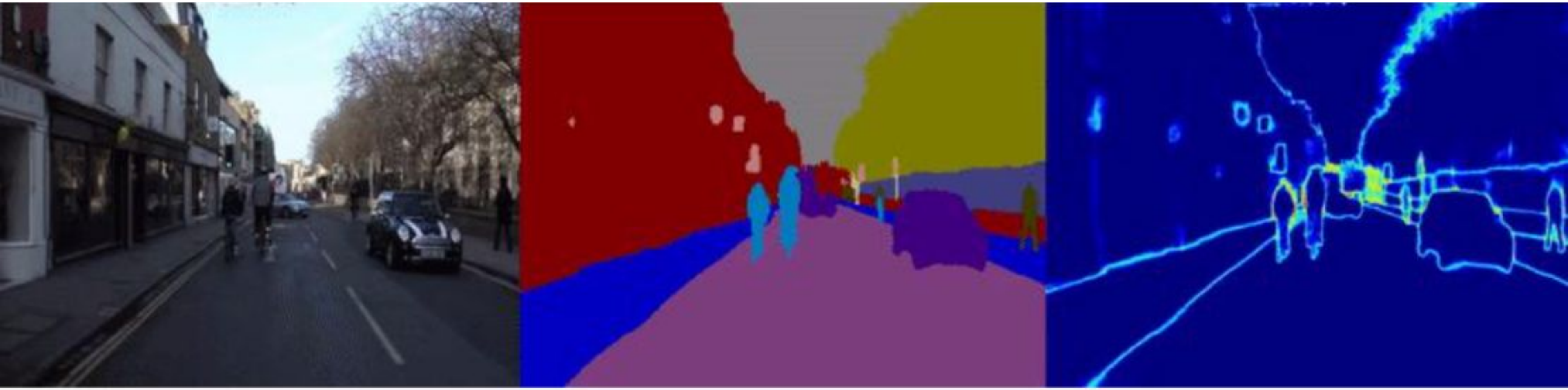
Color	Class
	Agriculture
	Forest
	Miscellaneous
	Urban
	Water


Input image of aircraft part
Neural network segmentation
Uncertainty map from BCNN
Production decision based on segmentation
Neural network successfully recognizes circle as void in material. Defective with *low uncertainty*.
Neural network fails to recognize low-contrast circle as void in material. Non-defective with *high uncertainty.*
Neural network was not trained on parts with cracks, so it fails to recognize triangle as crack in material. Non-defective with *high uncertainty*.

Predictions from Tiramisu on CamVid video stream.

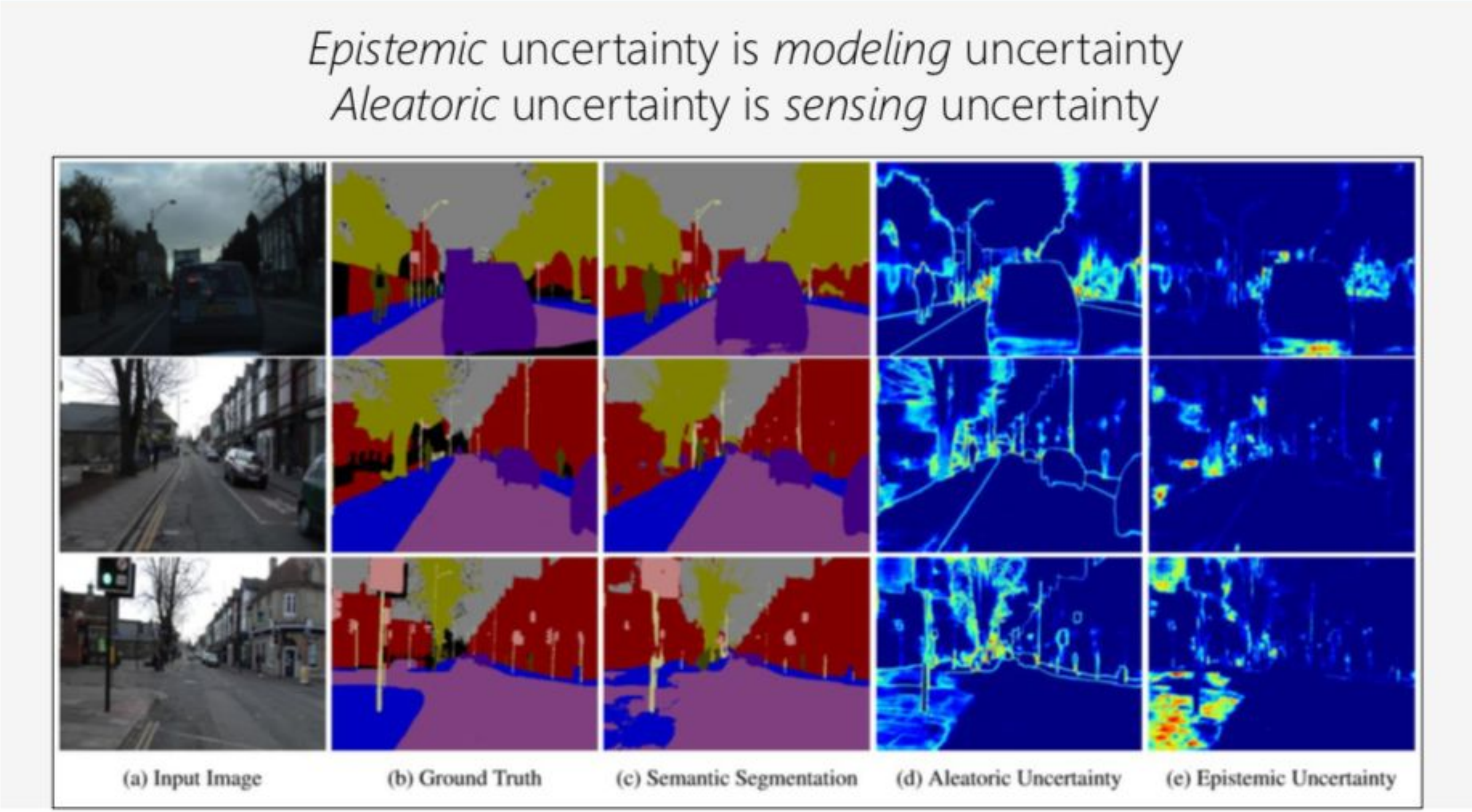


Bayesian SegNet
for probabilistic scene understanding



Input Image | Semantic Segmentation | Uncertainty

What kind of uncertainty can we model?



Modeling Uncertainty with Bayesian Deep Learning

?

Deep learning is required to achieve state of the art results in computer vision applications but doesn't provide uncertainty estimates.

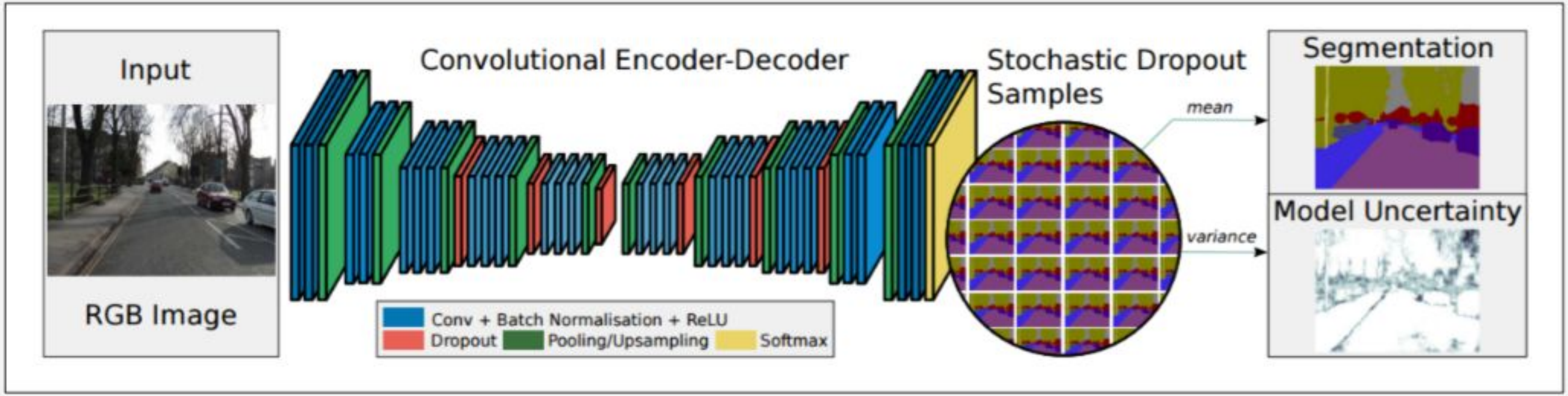
- **Bayesian neural networks** are a framework for understanding uncertainty in deep learning
- They have **distributions over network parameters** (rather than deterministic weights)
- Traditionally they have been **tricky to scale**

MacKay, David JC. **A practical Bayesian framework for backpropagation networks.** Neural Computation, 4(3): 448–472, 1992.

Modeling Epistemic Uncertainty with Bayesian Deep Learning

We can **model epistemic uncertainty** in deep learning models using Monte Carlo **dropout sampling** at test time.

Dropout sampling can be interpreted as **sampling from a distribution over models**.



Alex Kendall, Vijay Badrinarayanan and Roberto Cipolla **Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding**. arXiv preprint arXiv:1511.02680, 2015.

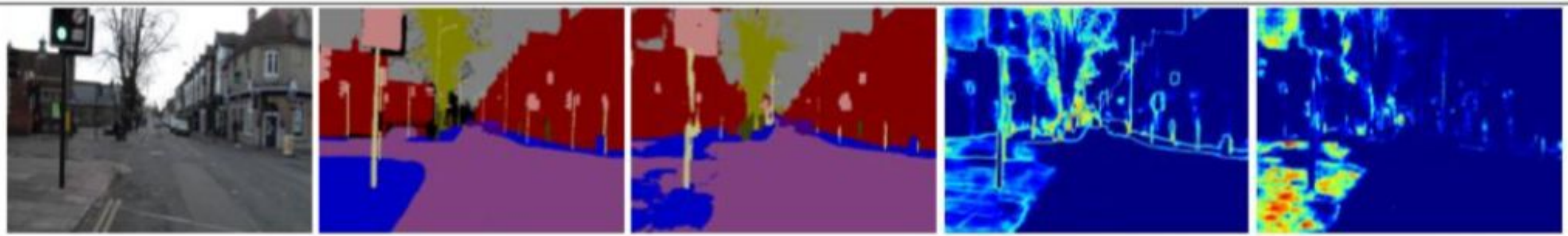
Modeling Aleatoric Uncertainty with Probabilistic Deep Learning

	Deep Learning	Probabilistic Deep Learning
Model	$[\hat{y}] = f(x)$	$[\hat{y}, \hat{\sigma}^2] = f(x)$
Regression	$Loss = \lVert y - \hat{y} \rVert^2$	$Loss = \dfrac{\lVert y - \hat{y} \rVert^2}{2\hat{\sigma}^2} + \log \hat{\sigma}^2$
Classification	$Loss = SoftmaxCrossEntropy(\hat{y}_{\mathrm{t}})$	$\hat{y}_{\mathrm{t}} = \hat{y} + \epsilon_t \qquad \epsilon_t \sim N(0, \hat{\sigma}^2)$ $Loss = \dfrac{1}{T}\sum_t SoftmaxCrossEntropy(\hat{y}_{\mathrm{t}})$

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** arXiv preprint 1703.04977, 2017.

Semantic Segmentation Performance on CamVid

CamVid Results	IoU Accuracy
DenseNet (State of the art baseline)	67.1
+ Aleatoric Uncertainty	67.4
+ Epistemic Uncertainty	67.2
+ Aleatoric & Epistemic	**67.5**



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** arXiv preprint 1703.04977, 2017.