HERMENEUTIC INVESTIGATIONS

Dean Allsopp

PyData London

November 6, 2018

INTERPRETABLE MACHINE LEARNING – WHY?

- *COMMERCIAL*
- *REGULATORY*
- *SOCIAL*

INTERPRETABLE MACHINE LEARNING – WHY?

INTERPRETABLE MACHINE LEARNING – WHY?

- *EPISTEMIC*
- *MOTIVATIONAL*

- *COMMERCIAL*
- *REGULATORY*
- *SOCIAL*

HERMENEUTICS

'The branch of knowledge that deals with interpretation'

UNDERSTANDING VIA

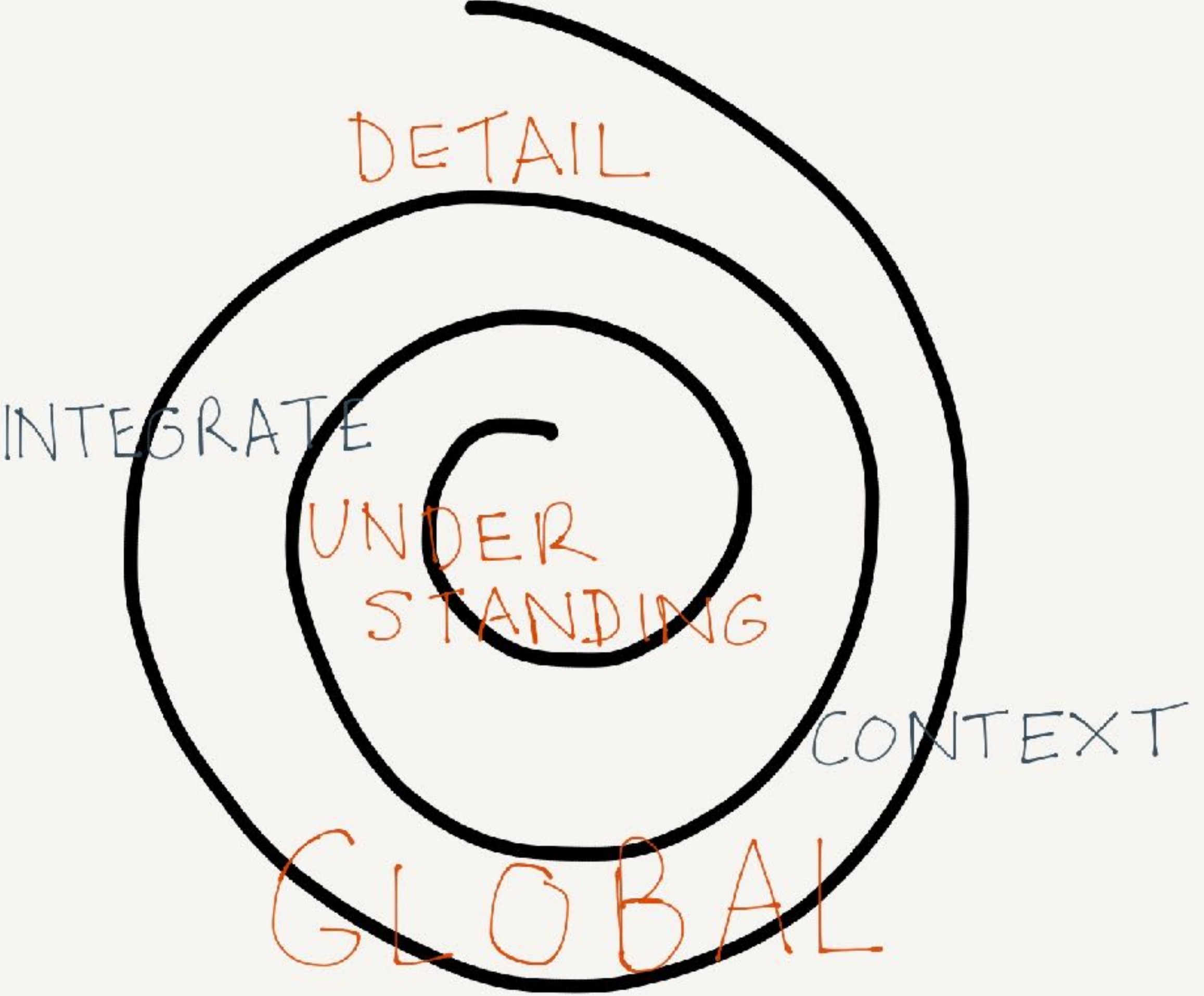
- ITERATIVE INTERACTION

UNDERSTANDING VIA

- ITERATIVE INTERACTION
- COMPARISON OF WHOLE & PART

UNDERSTANDING VIA

- ITERATIVE INTERACTION
- COMPARISON OF WHOLE & PART
- FUSION OF HORIZONS


DETAIL
INTEGRATE
UNDER
STANDING
CONTEXT
GLOBAL

INTERPRETABLE MACHINE LEARNING

“

Global ->

Model characteristics

Local ->

Individual predictions

Terms

“ Interpretability ->

Improving understanding

Explainability ->

Getting towards why

Terms

LIME

Linearity in Complexity

LIME slices (local; agnostic)

Generate data points based on training data

Compute complex model predictions from the generated data to find the 'most useful features'

Fit a local linear model for the 'most useful features' and use the feature coefficients as reason codes

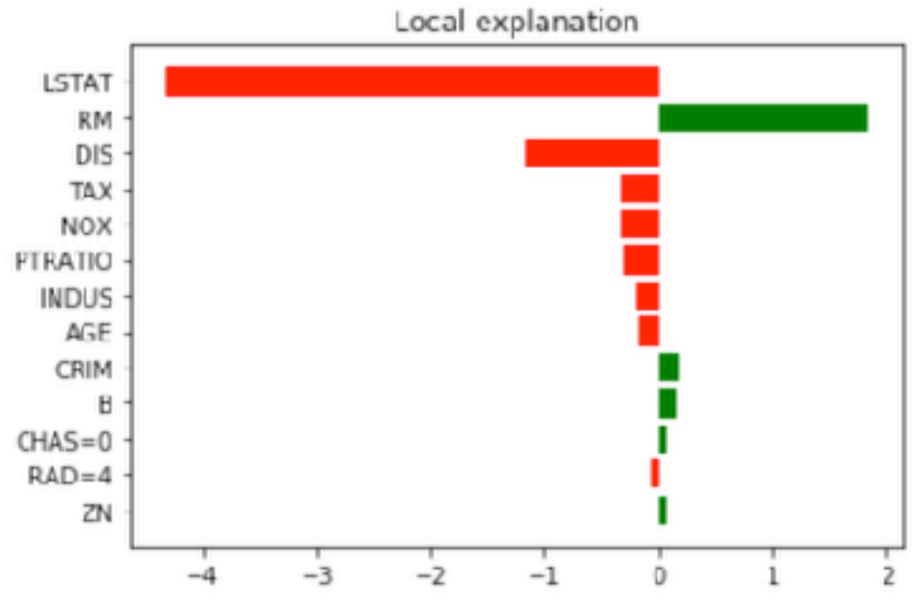
A few lines of code…

In [2]:
```
# import lime tools
import lime
import lime.lime_tabular

# generate an "explainer" object
categorical_features  = np.argwhere(np.array([len(set(boston.data[:,x])) for x in range(boston.data.shape[1])]) <= 10).
explainer = lime.lime_tabular.LimeTabularExplainer(train, feature_names=boston.feature_names, class_names=['price'], ca
```

In [3]:
```
#generate an explanation
i = 13
exp = explainer.explain_instance(test[i], rf.predict, num_features=14)
```

In [4]:
```
%matplotlib inline
fig = exp.as_pyplot_figure();
```



Limits of LIME

Local non-linearity and kernel width…

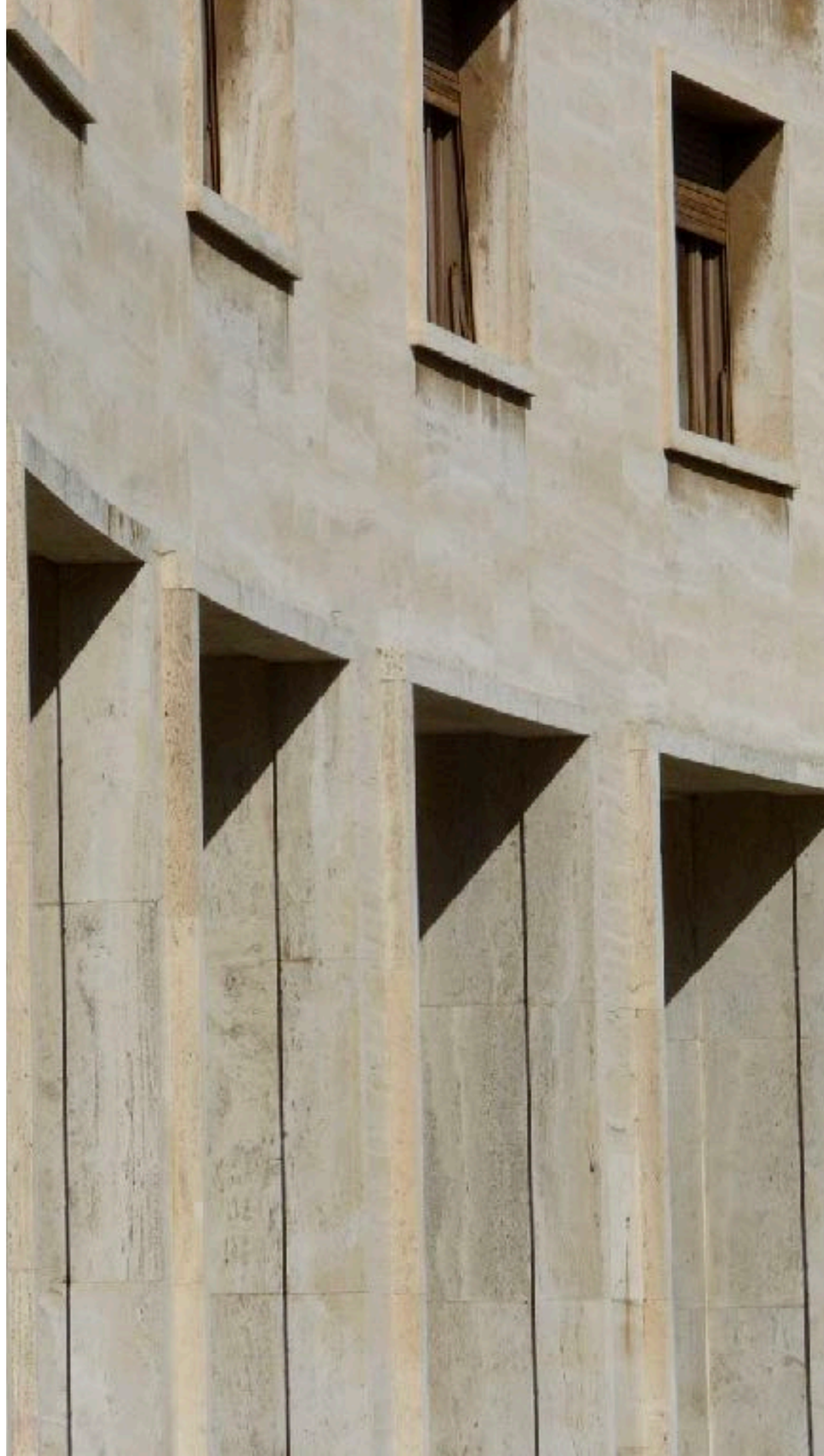
Linearity versus feature interactions?

Latency of results due to computation

*Check the LIME prediction probability**

SKATER

GENERATE TREES
RULES
LIME



Skater tree surrogate (regression)

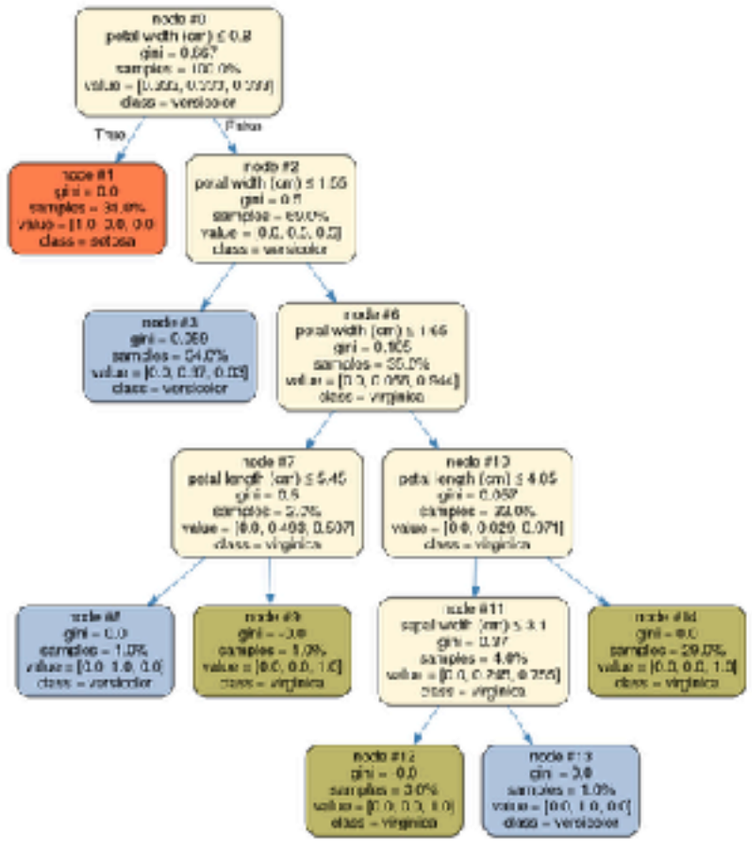
```
In [49]: surrogate_explainer_reg.decisions_as_txt('local', X_test.iloc[sample_index])
```

```
As RM[5.914] <= 6.837500095367432 then,
  As LSTAT[18.33] > 14.399999618530273 then,
    As CRIM[0.31827] <= 7.084139823913574 then,
      As CRIM[0.31827] <= 0.6147900223731995 then,
        As DIS[3.9986] > 1.979900024795532 then,
          As AGE[83.2] > 73.30000305175781 then,
            As CRIM[0.31827] > 0.17127001285552979 then,
              As CHAS[0.0] <= 0.5 then,
                As B[390.7] <= 391.875 then,
                  As NOX[0.544] > 0.4864999499320984 then,
                    As B[390.7] > 389.96002197265625 then,
                      Value: [[19.36]]
```

Skater tree surrogate (classification)

In [28]:
```
surroqate_explainer2.plot_global_decisions(colors=['coral', 'lightsteelblue','darkkhaki'],
                                           file_name='simple_tree_post.png')
show_in_notebook('simple_tree_post.png', width=400, height=300)
```

```
2018-09-23 18:31:23,476 - skater.util.dataops - INFO - File Name: simple_tree_post.png
```



Skater Bayesian rules list (classification)

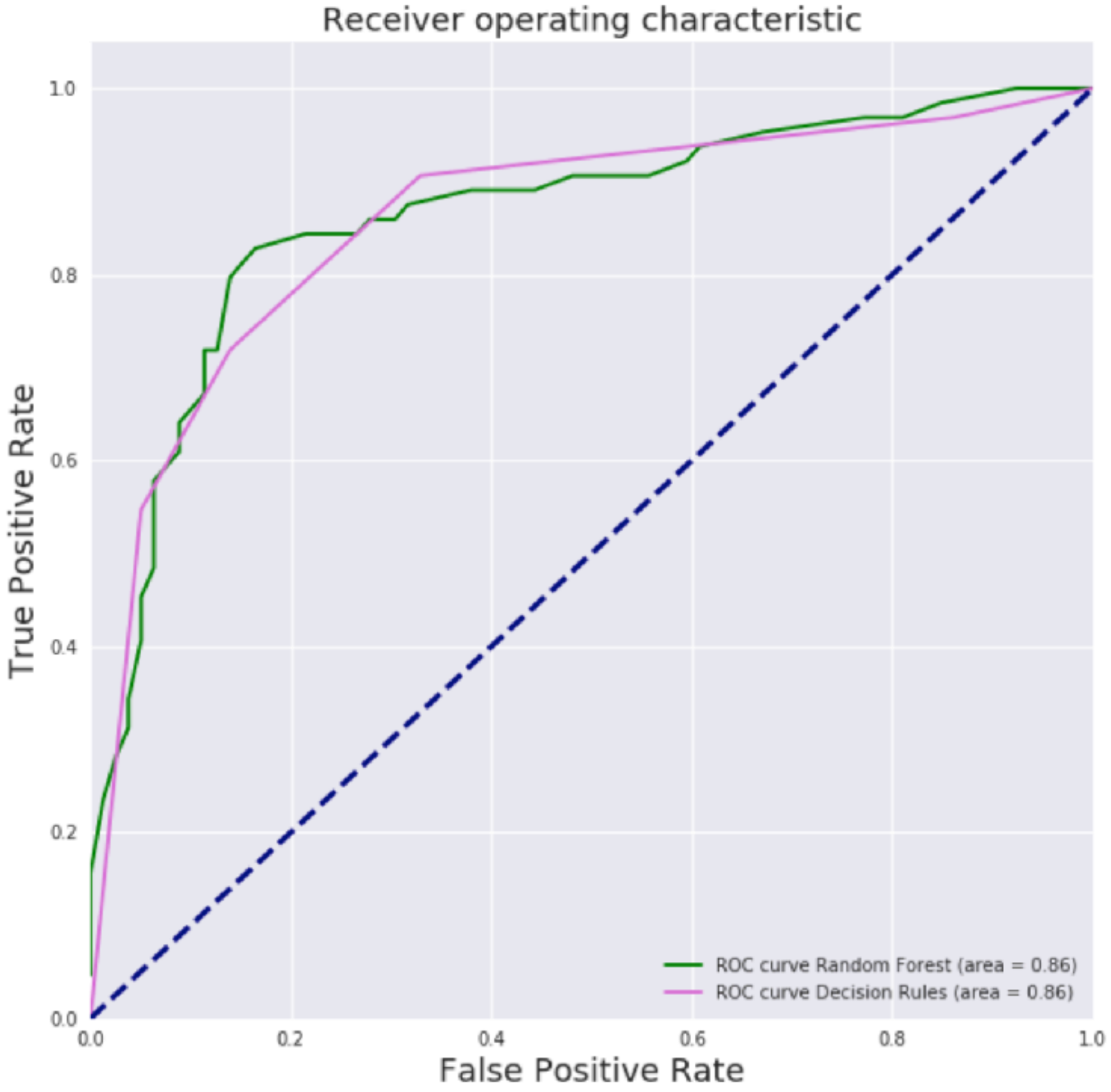
```
In [31]: sbrl_big.print_model()

The rules list is :

If      {Sex_Encoded=1} (rule[276]) then positive probability = 0.18709677

else if {Pclass=3} (rule[237]) then positive probability = 0.32352941

else  (default rule)  then positive probability = 0.94029851
```


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve Random Forest (area = 0.86)
ROC curve Decision Rules (area = 0.86)

SHAP

INTERPRETING TREES
KERNELS
GRADIENTS

Cooperative Game Theory!

Shapley Values (approximations)

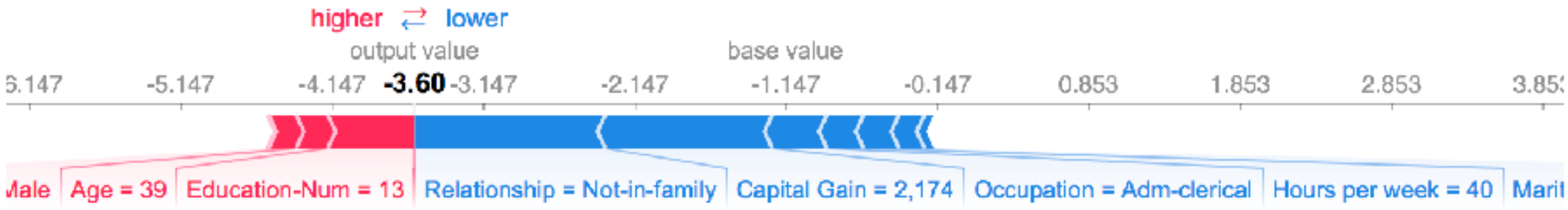
Calculate the value of a member to a group by comparing combinations of coalitions

Beware the effect of multicollinearity…

Tree SHAP(local, groups)

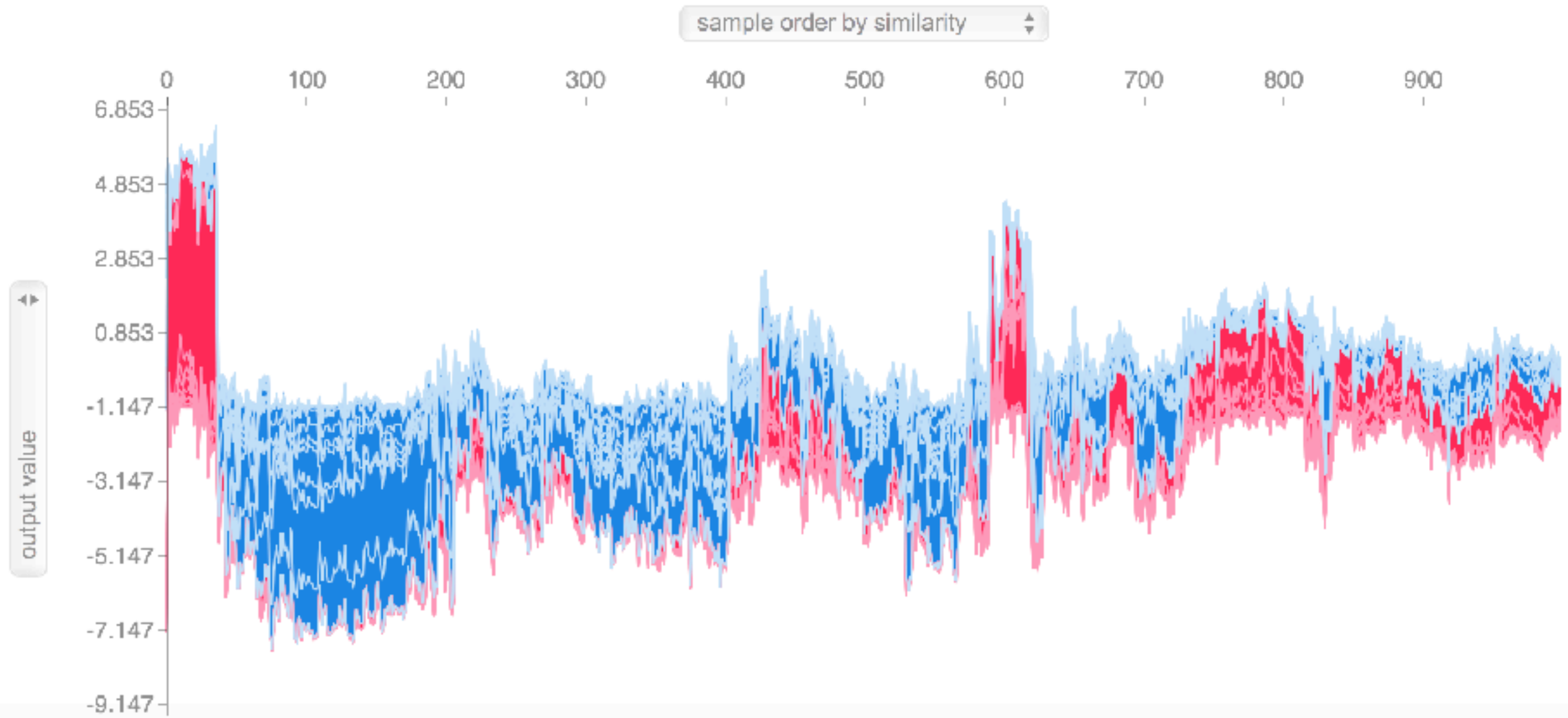
```
In [9]: shap.force_plot(explainer.expected_value, shap_values[0,:], X_display.iloc[0,:])
```

Out[9]:




```
In [10]: shap.force_plot(explainer.expected_value, shap_values[:1000,:], X_display.iloc[:1000,:])
```

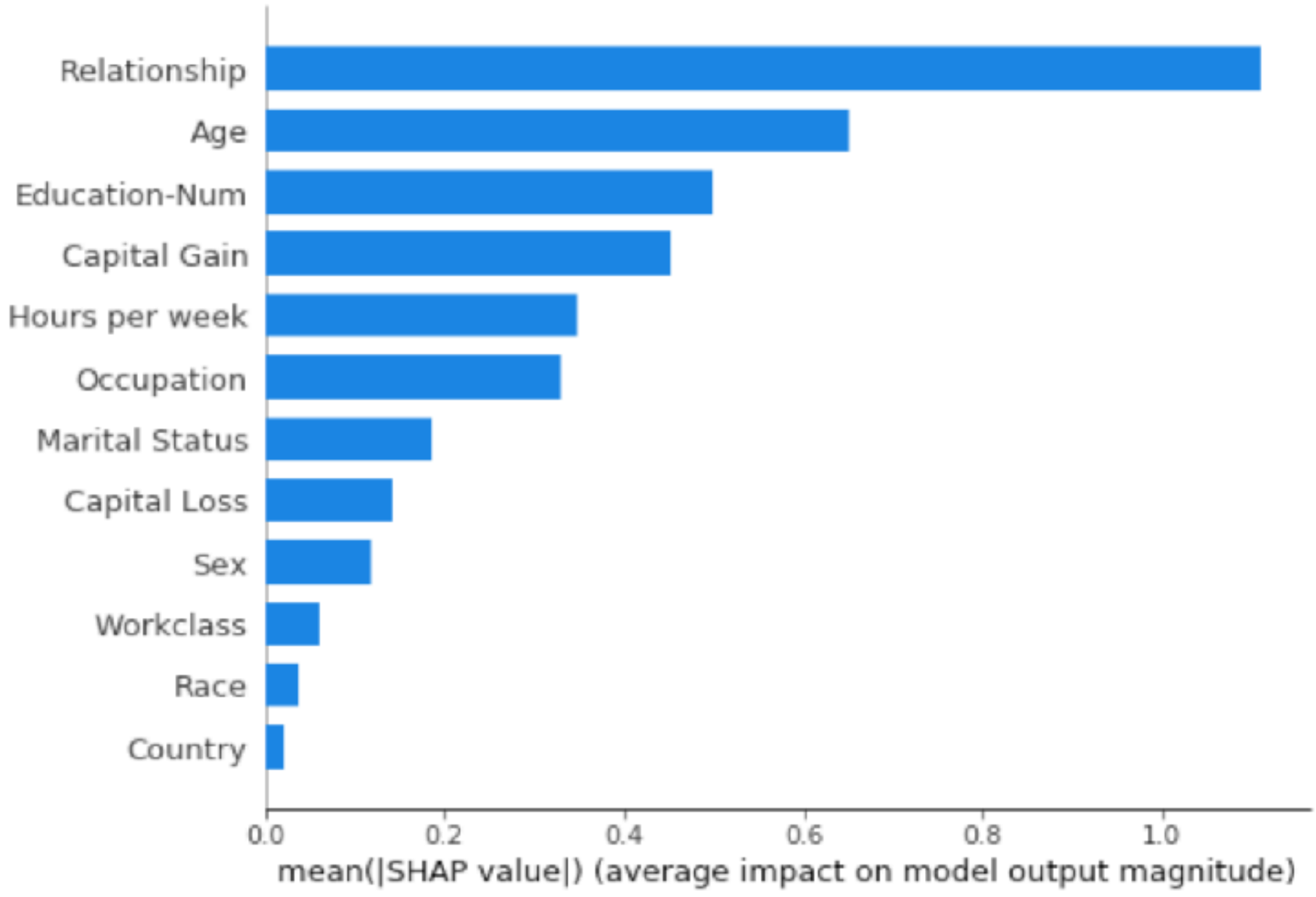
Out[10]:



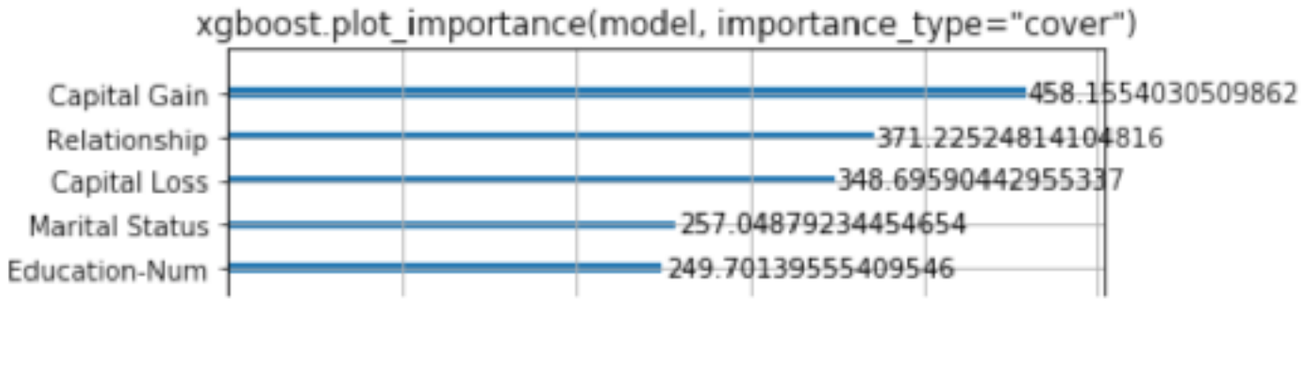
Tree SHAP(global)

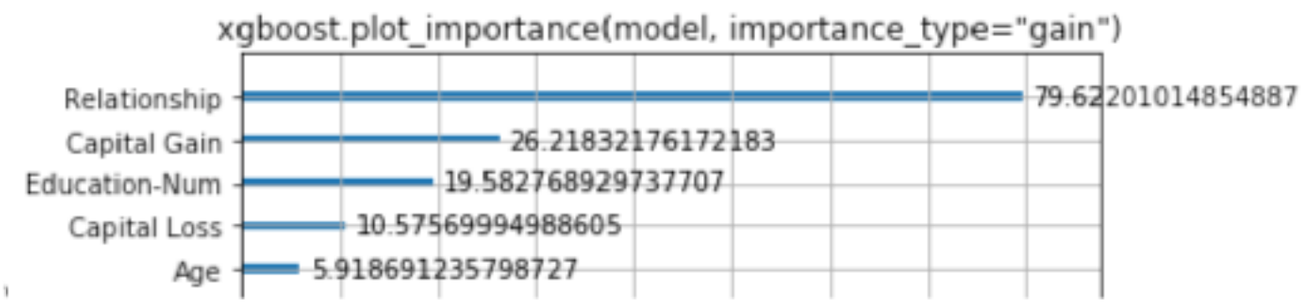
Variable Importance

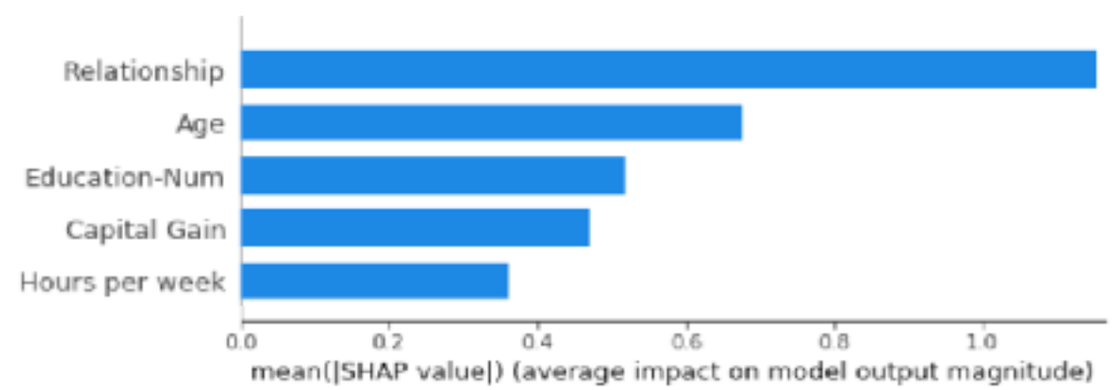
```
In [11]: shap.summary_plot(shap_values, X_display, plot_type="bar")
```



Variable Importances (global…)

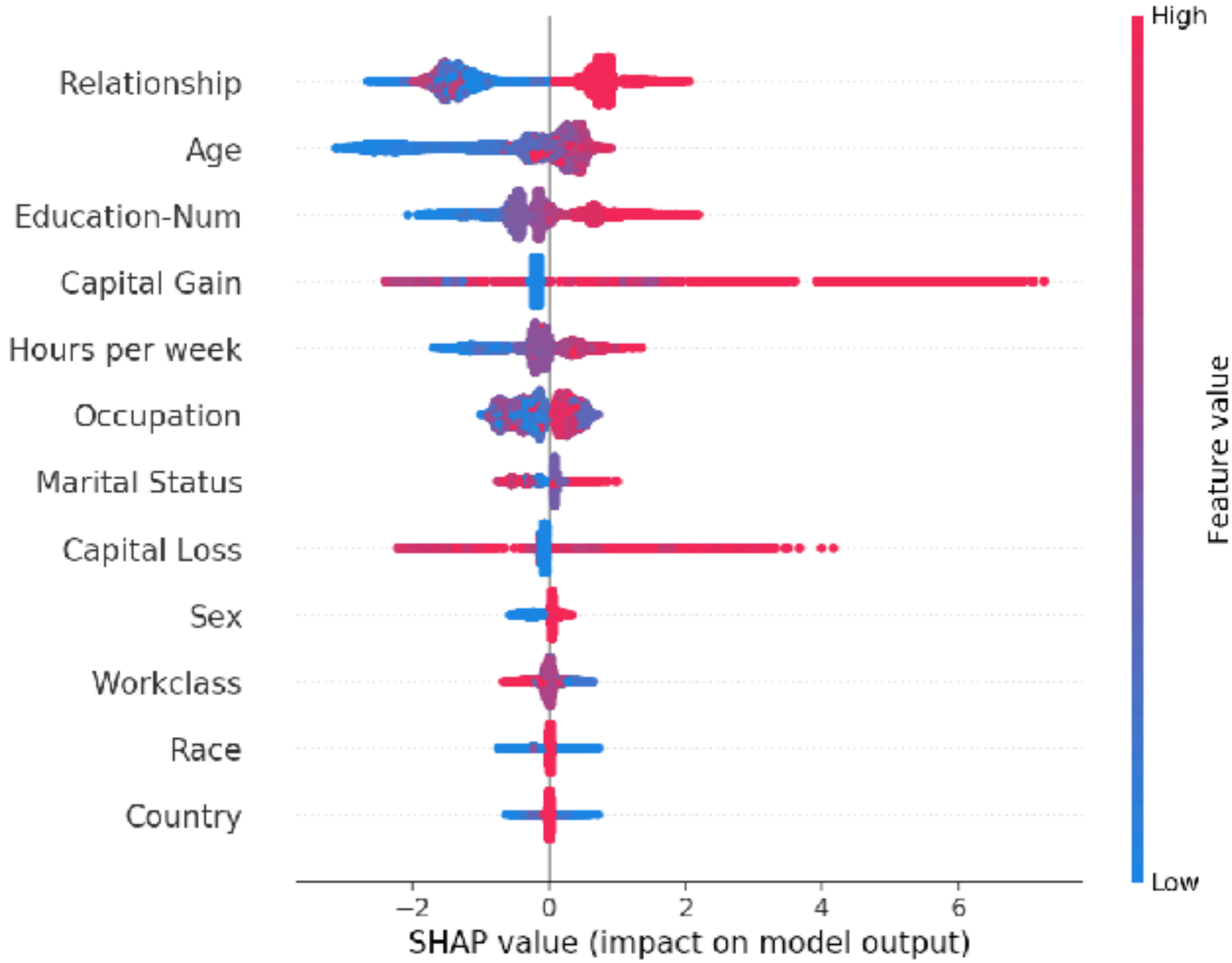






Tree SHAP(global and local) Overall and Individual

```
In [9]: shap.summary_plot(shap_values, X)
```



NOTE: XGBOOST/CATBOOST TREE SHAP INSIDE

CONTRASTIVE

COUNTERFACTUAL

“

He who differentiates well teaches well

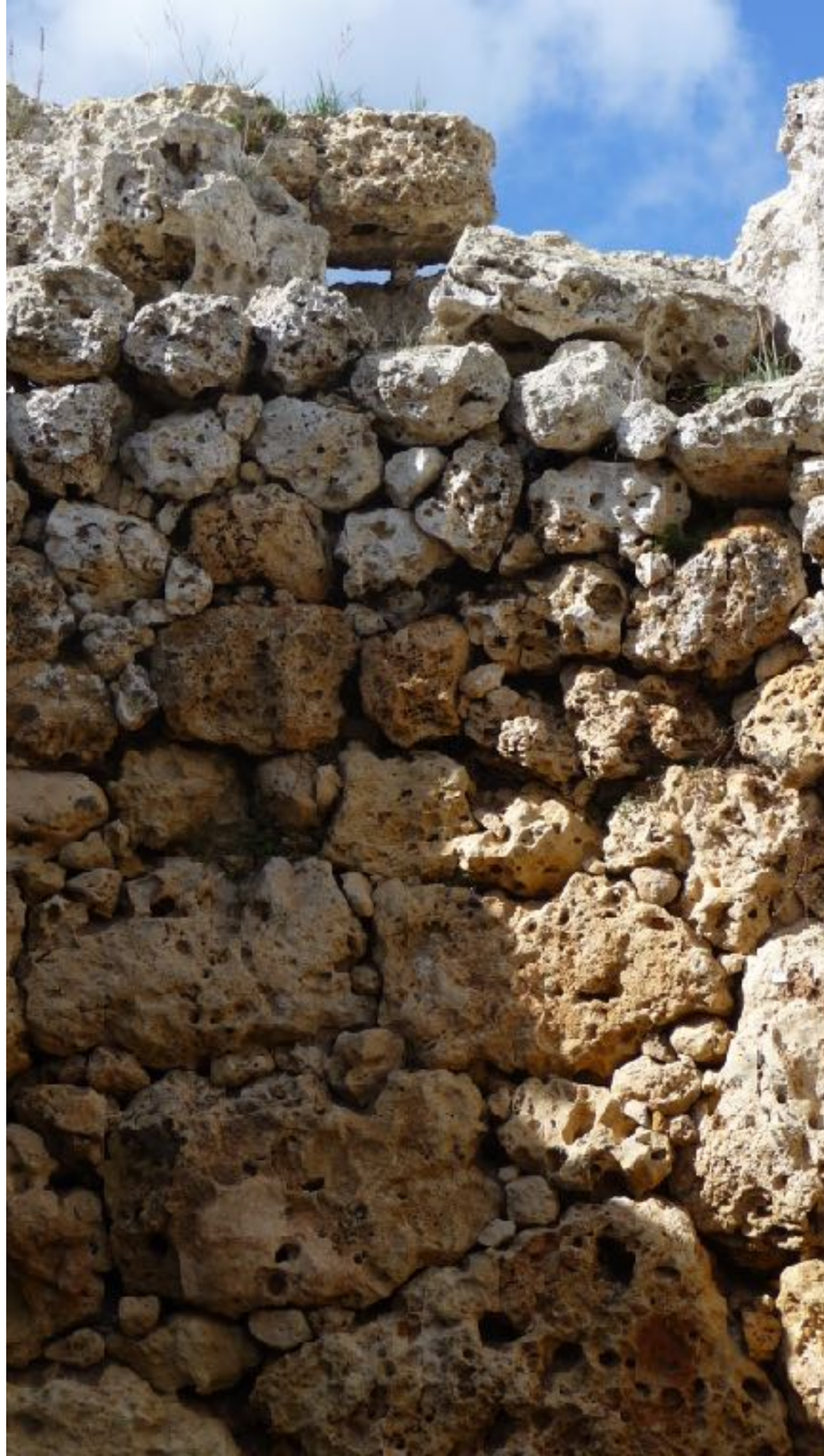
-John Amos Comenius

ContrastiveExplanation (local)

```
The model predicted '92.414' instead of 'more than 92.414' because 's3 <= 0.1
The model predicted '92.414' because 'bmi <= 0.006 and s5 <= -0.003 and s3 <=
 s1 <= -0.058 and s1 > 0.023 and age > -0.03 and s4 > 0.016 and s2 <= -0.002


of 'more than 92.414' because 's3 <= 0.106 and bmi <= -0.021'",
'bmi <= 0.006 and s5 <= -0.003 and s3 <= 0.02 and bp > 0.054 and s4 <= 0.024
> -0.03 and s4 > 0.016 and s2 <= -0.002 and s1 > 0.113 and bp <= -0.018'")
```

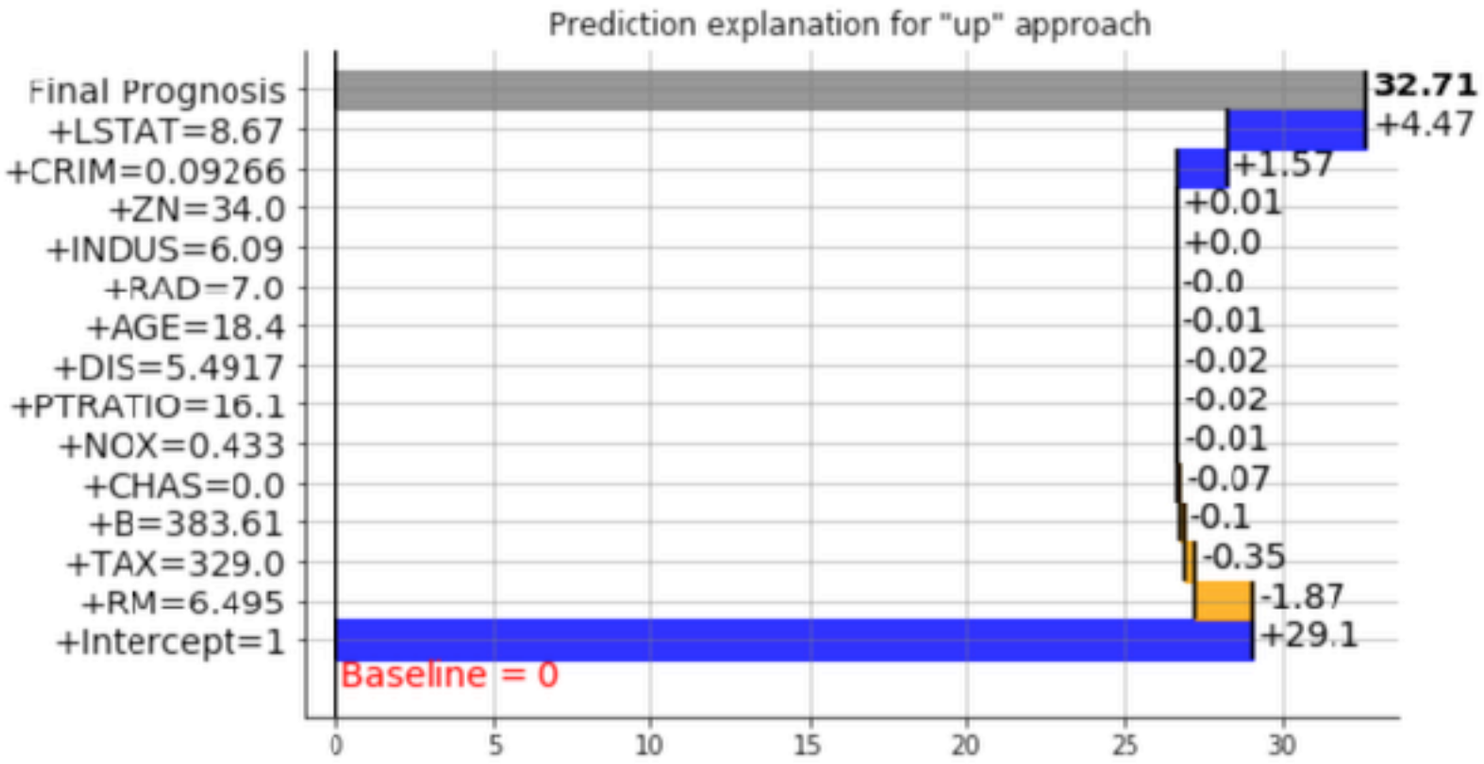
PYBREAKDOWN



Breakdown (Global)

In [7]:
```
# visualisation

explanation.visualize(figsize=(8,5),dpi=100)
```



RULEFIT

linearity with interactions

RuleFit

Fit a sparse linear model with the training set and additional features which are decision rules

The decision rules are generated by decision trees and reveal interactions between the original features

RuleFit

```
rf = RuleFit(tree_size=4,sample_fract='default',max_rules=2000,
             memory_par=0.01,
             tree_generator=None,
            rfmode='classify',lin_trim_quantile=0.025,
             lin_standardise=True, exp_rand_tree_size=True,random_state=1
rf.fit(X, y_class, feature_names=features)
```

```
      lstat > 19.775001525878906 & lstat <= 19.82999...    rule  8.465416
      lstat <= 7.71999979019165 & lstat > 7.68499994...    rule -6.059011
```

L2X

LORE

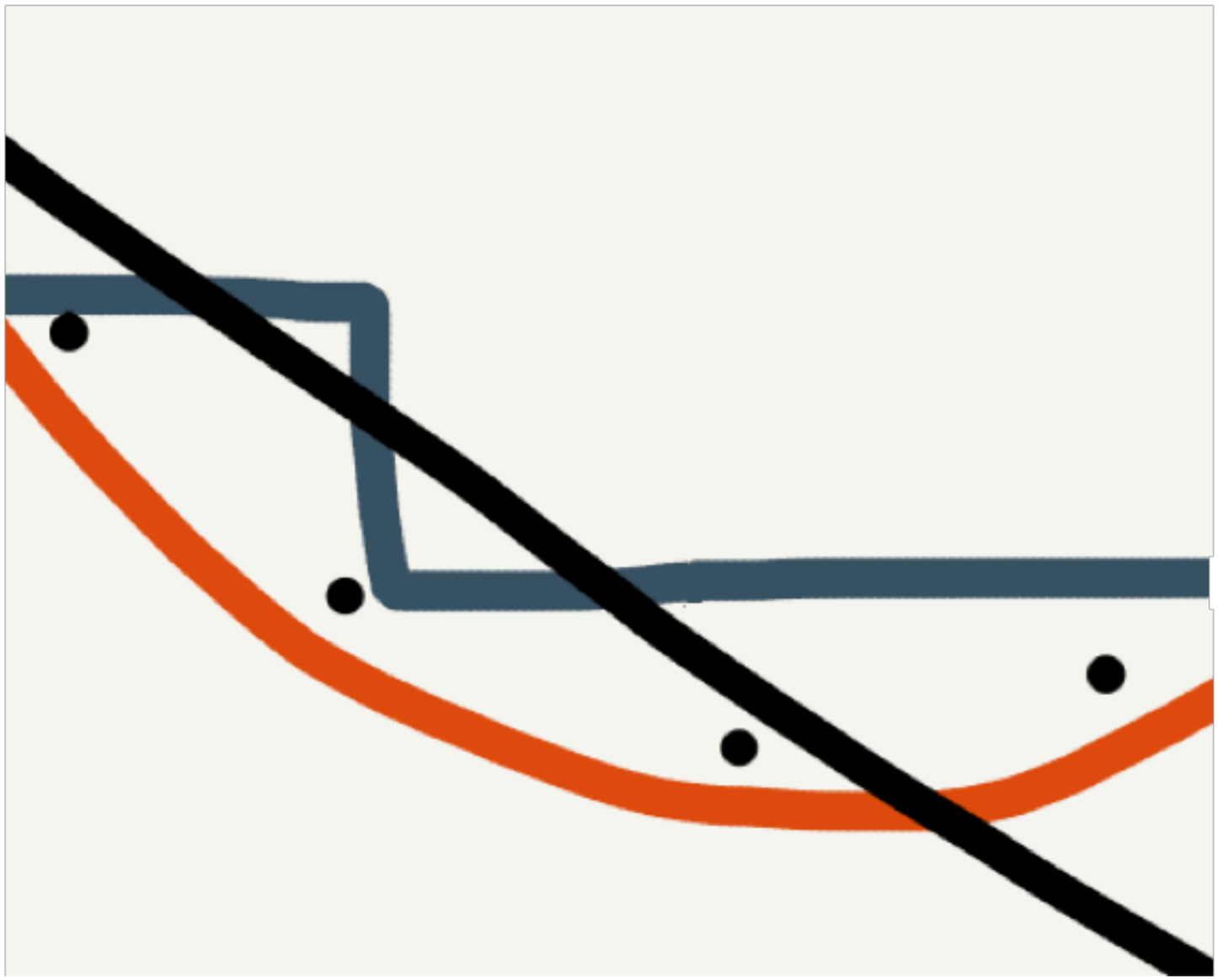
NONCONFORMIST

INFORMATION THEORETIC
LOCAL EXPLANATIONS
CONFORMAL

XGBOOST

Model Monotonicity

Monotonic models

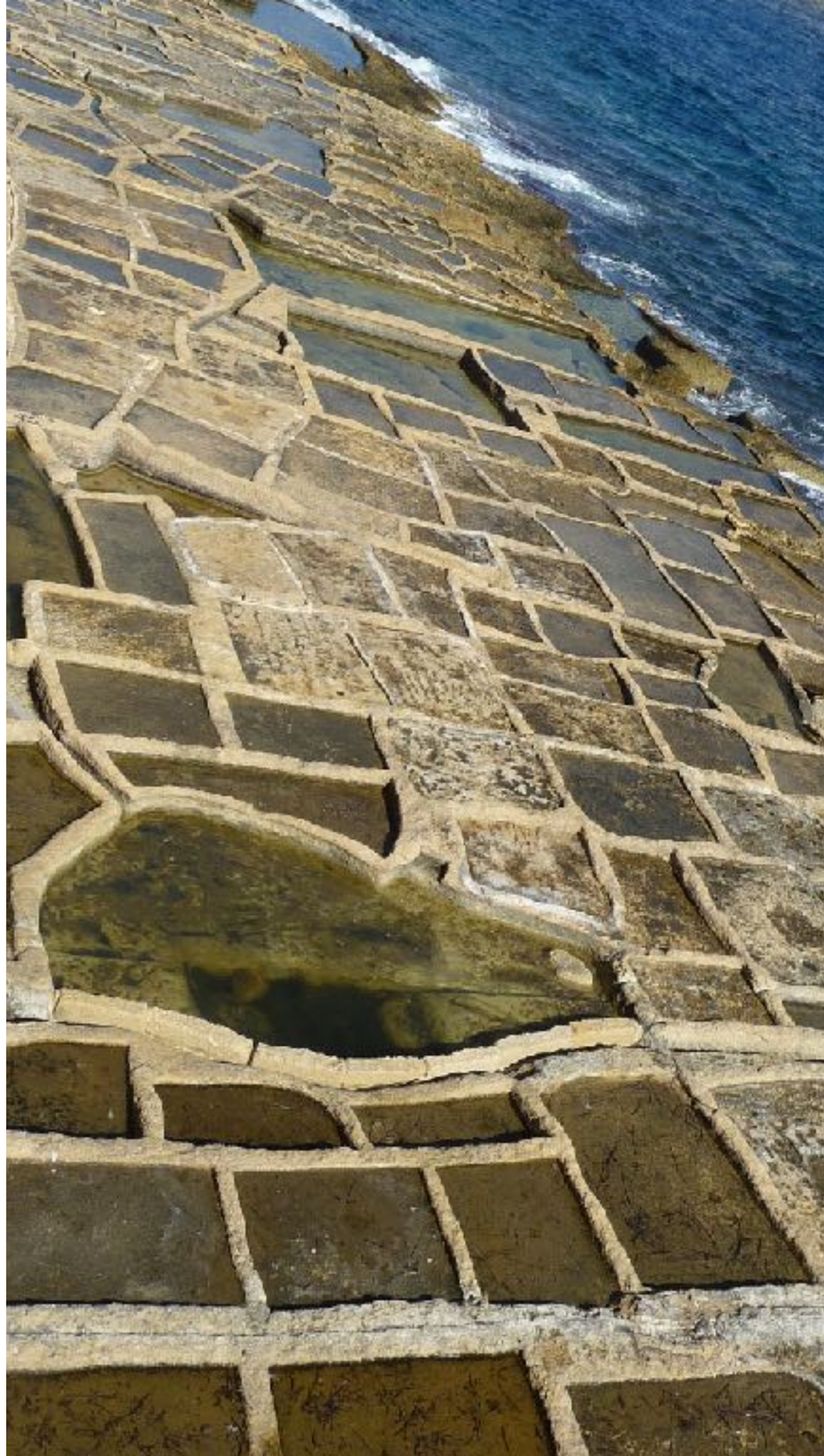


INTO THE DEEP

'what about neural networks?'

TENSORFLOW LATTICE

Model Monotonicity



DKNN

DARKSIGHT

INNVESTIGATE!

DEEP k-NEAREST NEIGHBOURS

MODEL DISTILLATION

DEEP TAYLOR DECOMPOSITION

Deep Taylor Decomposition

Saliency

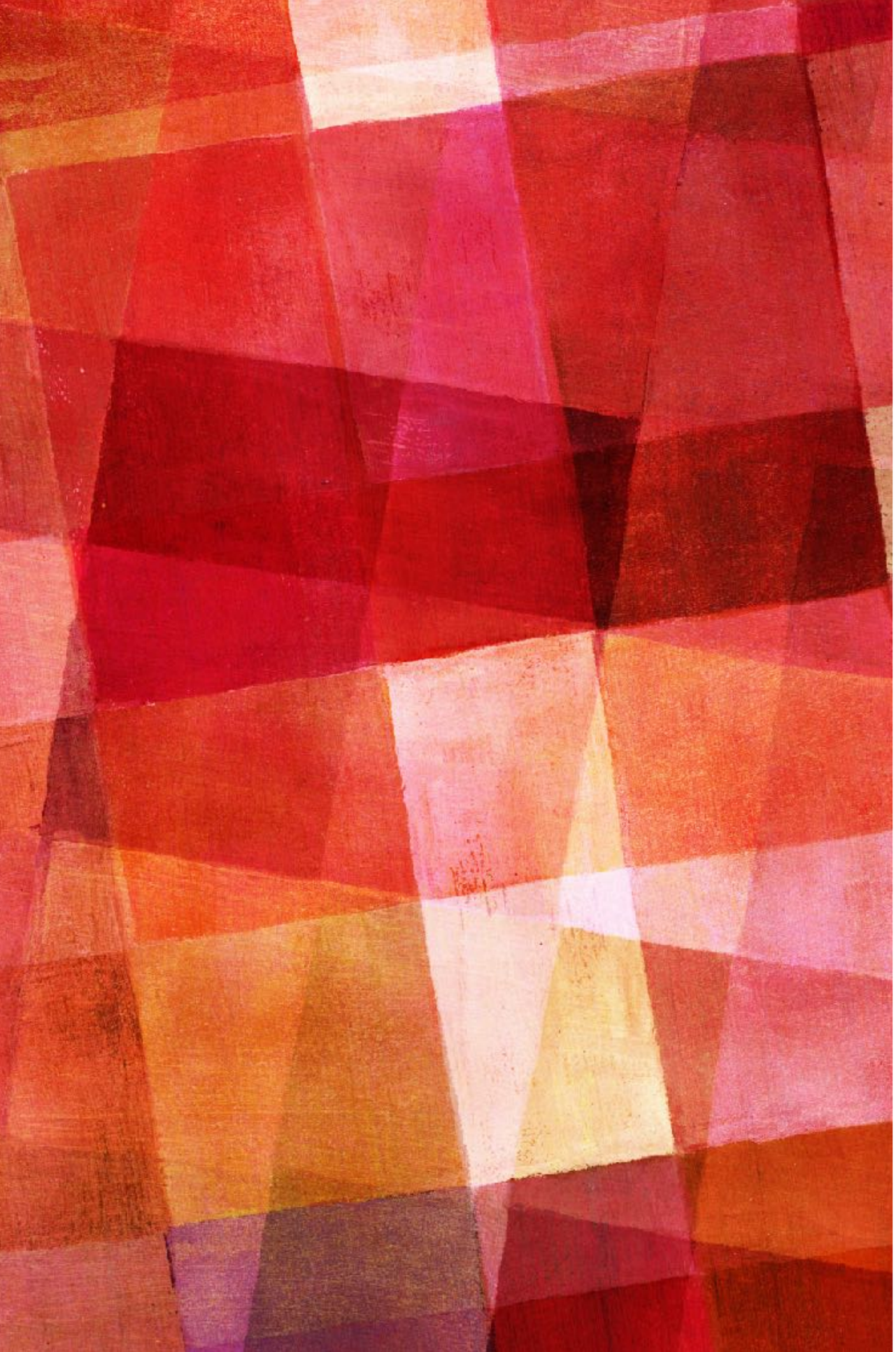
Sensitivity

*Relevance**

To the reading of DL research papers there is no end…

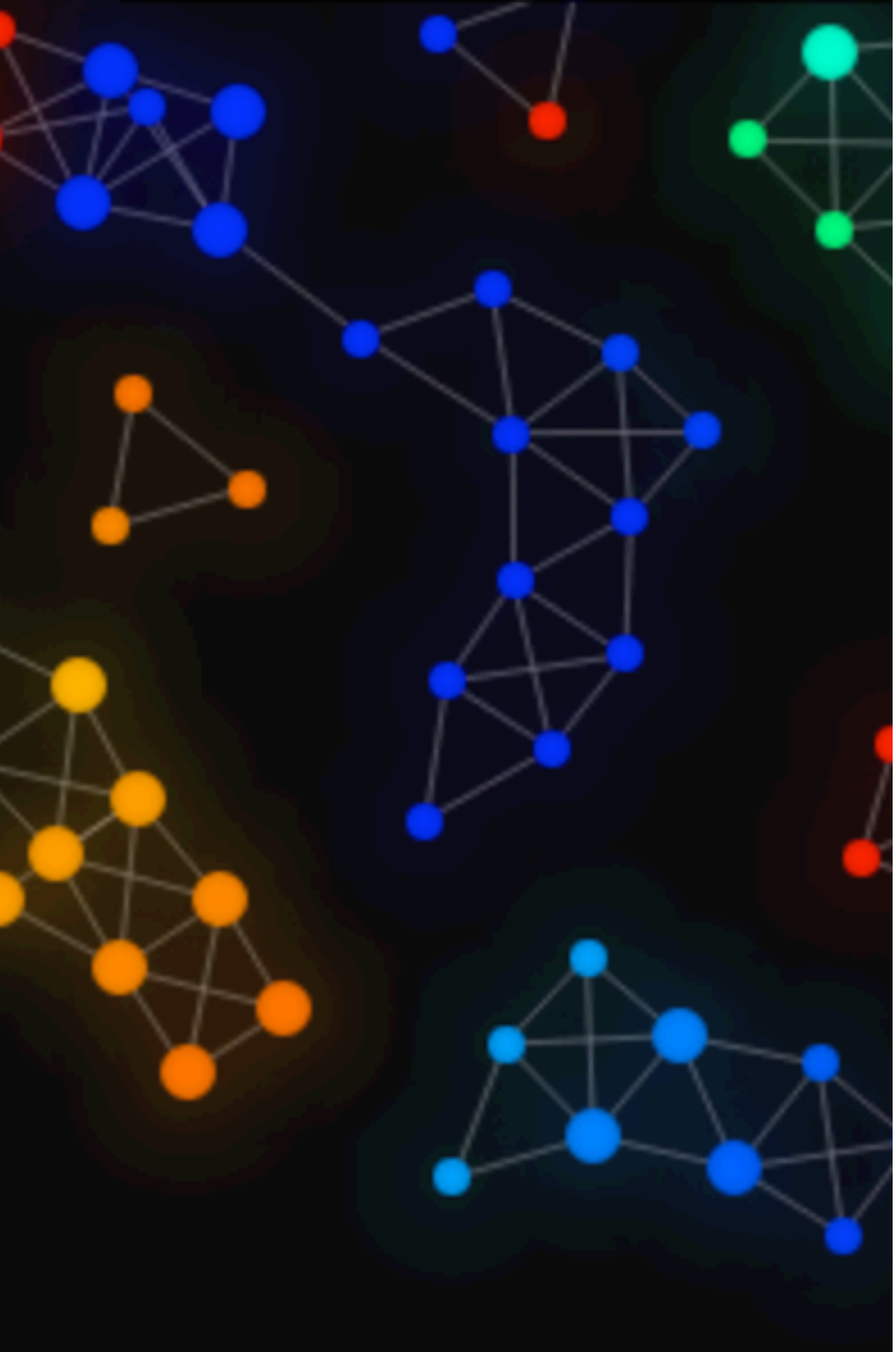
VISUALISATION

‘biological neuron activation’



PLOTTING CONTRIBUTIONS

- PDPBOX
- PARTIAL DEPENDENCE (AVERAGE)
- INDIVIDUAL CONDITIONAL EXPECTATION

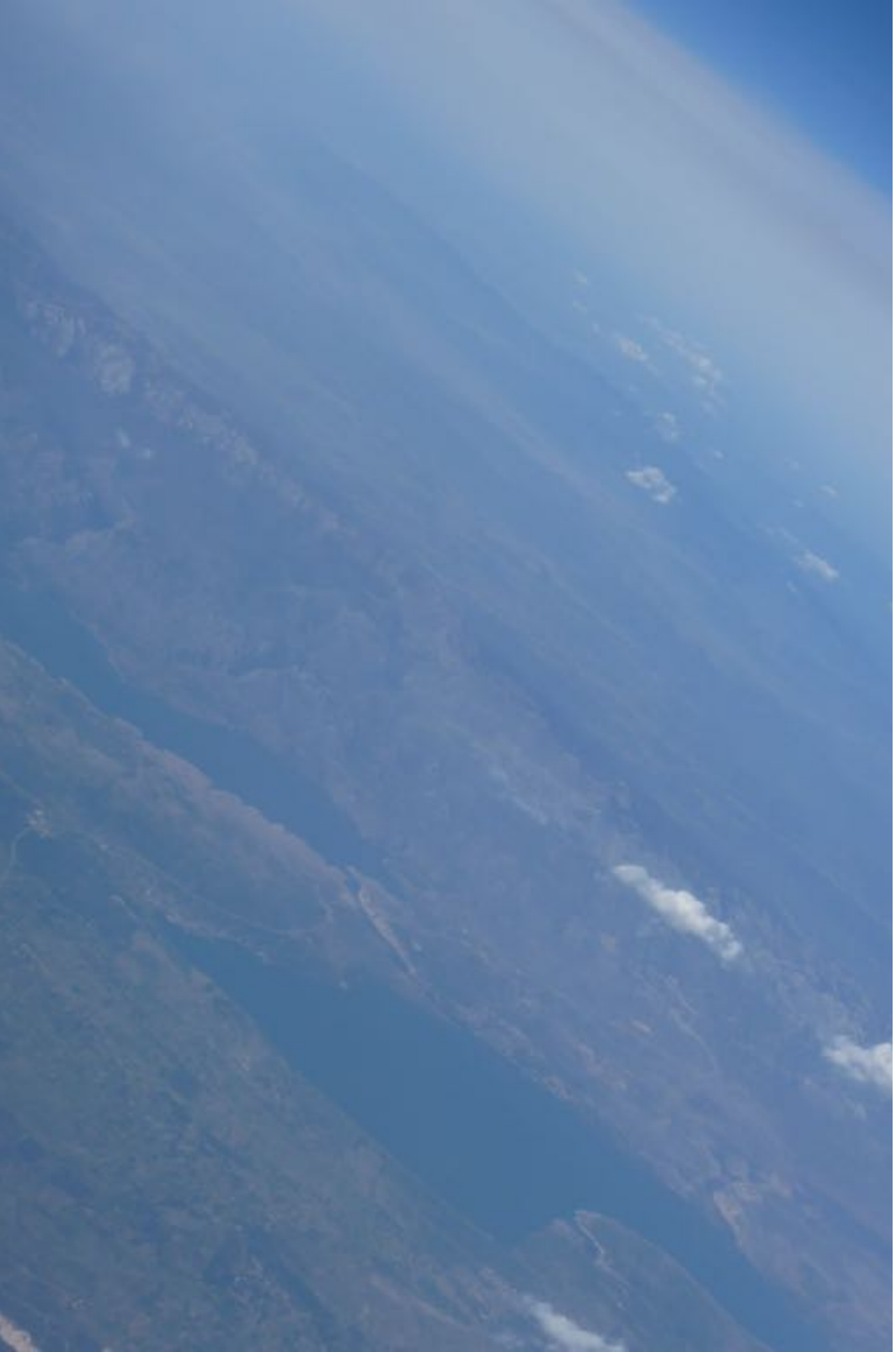


RELATIONSHIPS

- YELLOWBRICK (DIAGNOSTICS)
- KEPLER MAPPER (TOPOLOGICAL DATA ANALYSIS)

ANOTHER ANGLE

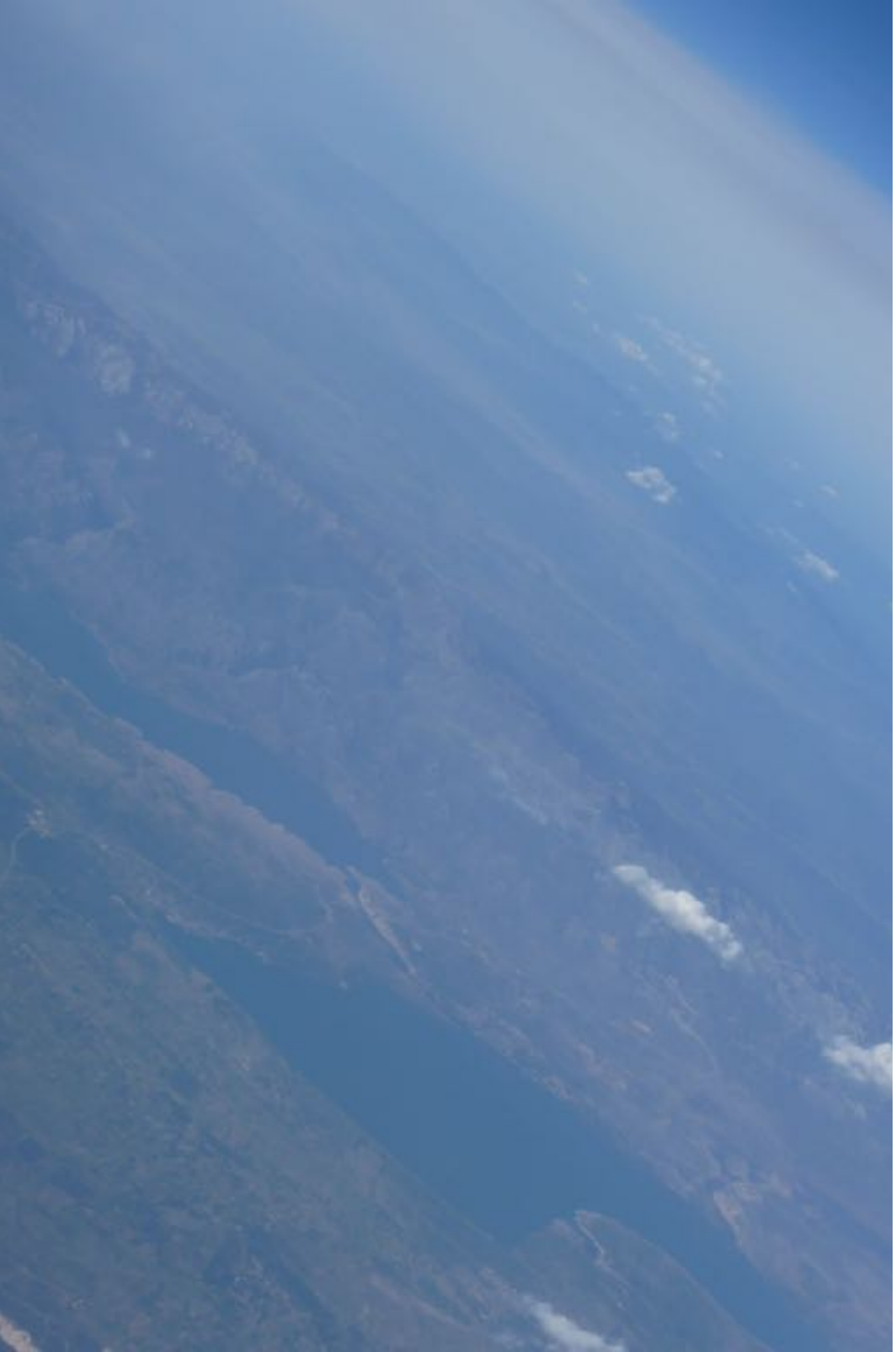
'R there other options?'



PACKAGES

- FFTREES
  FAST AND
  FRUGAL TREES

- GAMSEL
  GENERAL
  ADDITIVE
  MODEL
  SELECTION



PLOTS

- DALEX (DESCRIPTIVE MACHINE LEARNING EXPLANATIONS)
- ACCUMULATED LOCAL EFFECTS
- CETERIS PARIBUS

ONE MORE THING

‘All that coding…’

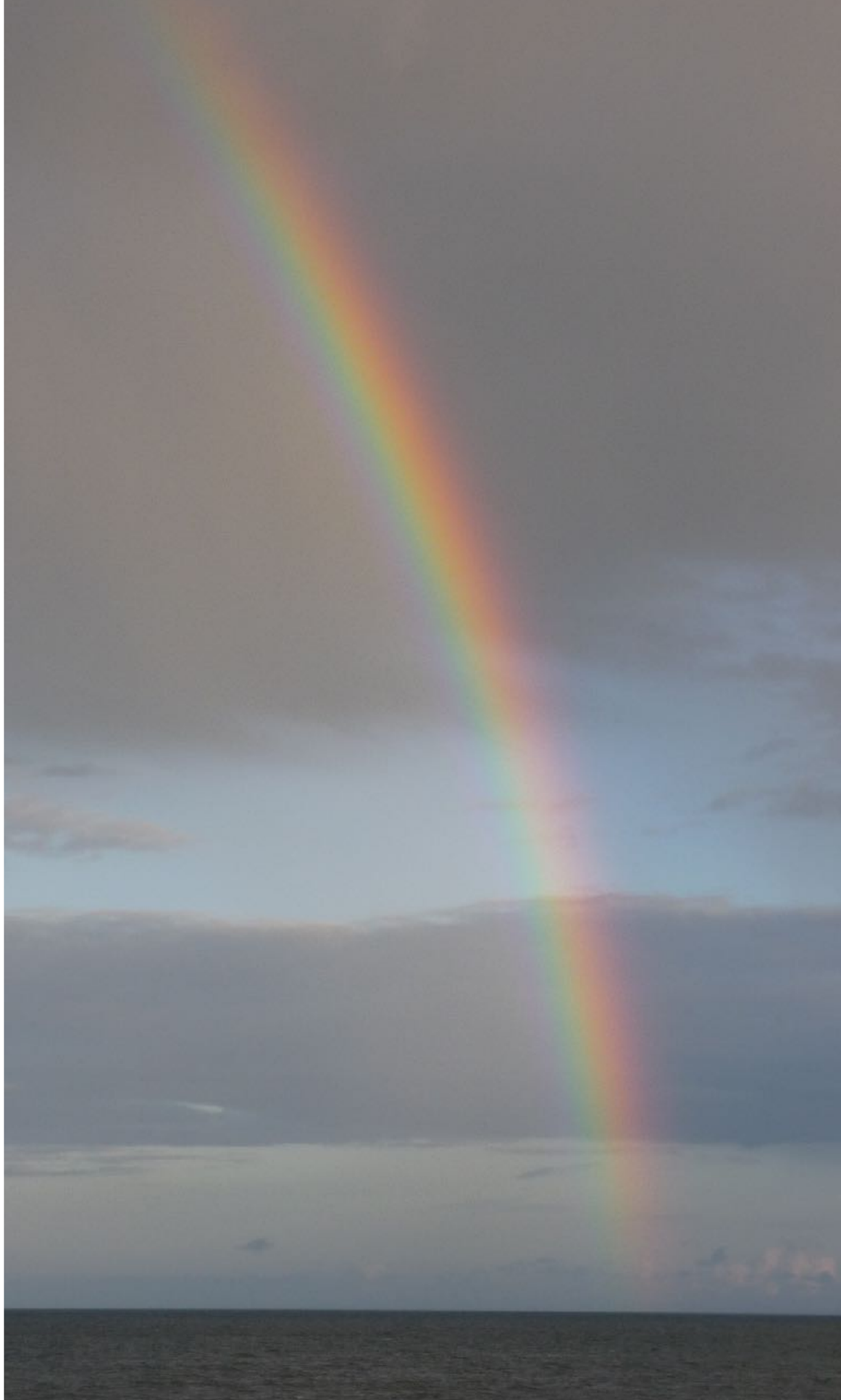
WHAT IF?

VIEW SLICES OF PREDICTIONS

VISUALISE INTERACTIVE EFFECTS

COMPARE COUNTERFACTUAL EXAMPLES

APPLY ALGORITHMIC FAIRNESS CONSTRAINTS



Partial dependence plots | Compute distance | Show nearest counterfactual: L1 | L2 | ⓘ | 500 examples loaded

DATAPOINT EDITOR | PERFORMANCE + FAIRNESS | FEATURES

Example ID: 15 | Run inference | Search f...

Feature name	Selected datapoint	L1 counterfactual
age	43	39
capital-gain	0	0
capital-loss	0	0
education	Masters	Bachelors
education-num	14	13
hours-per-week	50	50
marital-status	Married-civ-spouse	Separated
native-country	United-States	United-States
occupation	Exec-managerial	Prof-specialty
over_50k	1	1
race	White	White
relationship	Husband	Not-in-family
sex	Male	Male
workclass	Private	Private

+

Inference Results: Example 15

Run	Label	Score	Delta
1	1 (>50k)	0.809	
1	0 (<=50k)	0.194	

L1: Example 191

Run	Label	Score	Delta
1	0 (<=50k)	0.668	
1	1 (>50k)	0.330	

Binning l... age | Binning l... (none) | Color By Inf... | Label By Ex... | Scatter l... (defa) | Scatter l... (defa)

17 – 23 | 24 – 29 | 30 – 35 | 36 – 42 | 43 – 48 | 49 – 54 | 55 – 61 | 62 – 67 | 68

Legend

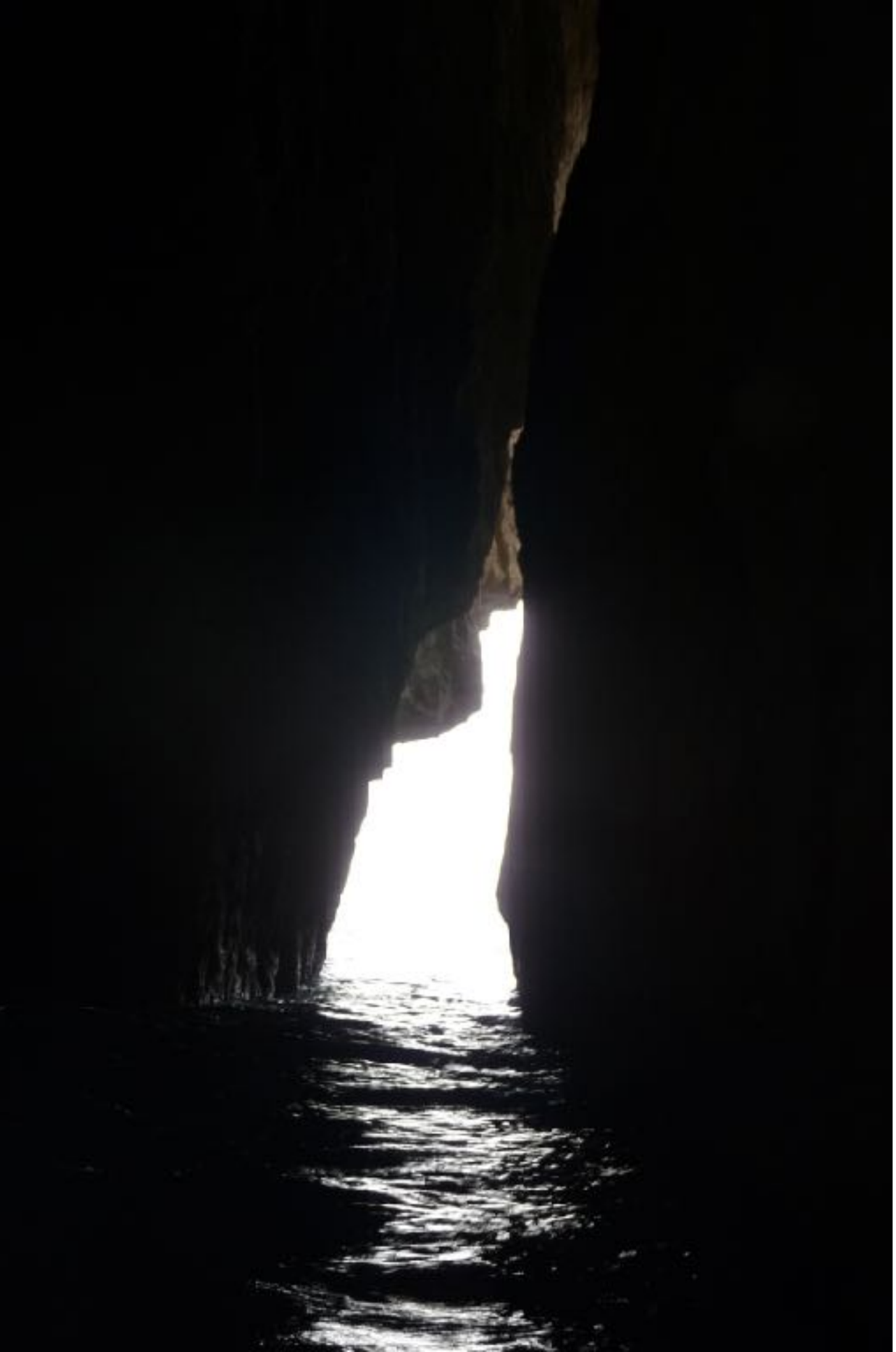
Colors

by Inference label

<=50k

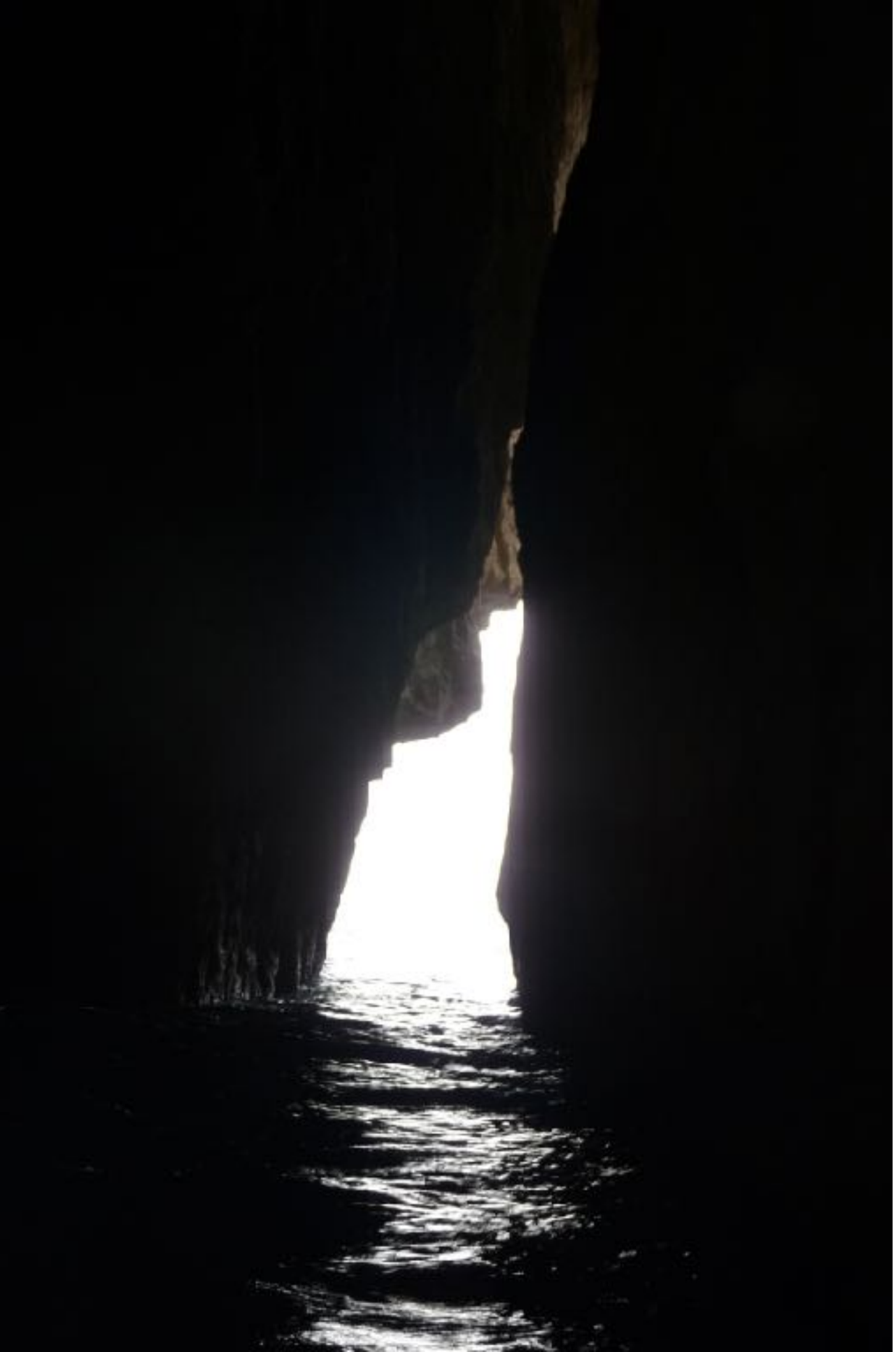
>50k

REVIEW



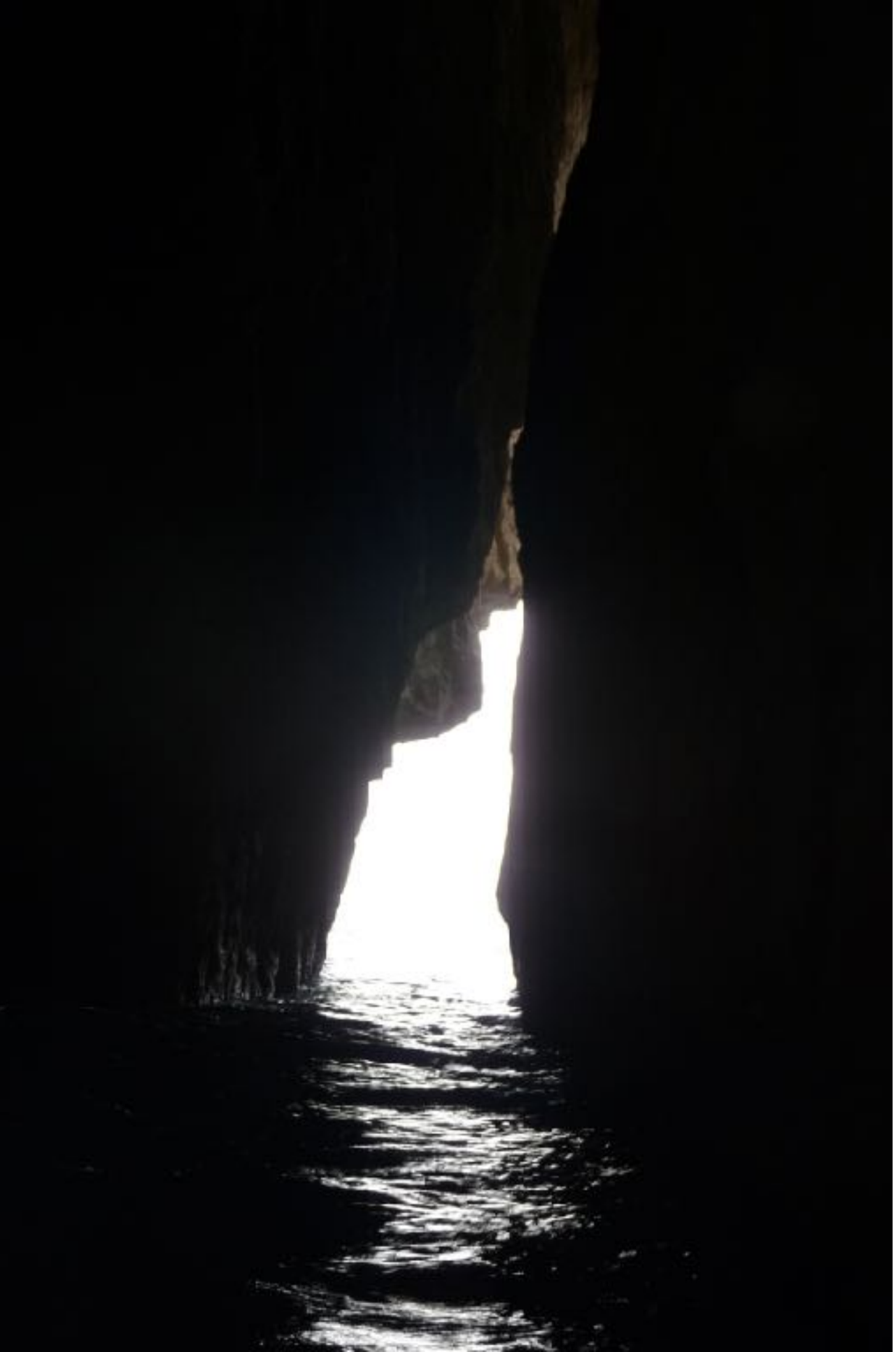
UNDERSTANDING VIA

- COMBINING MULTIPLE TECHNIQUES



UNDERSTANDING VIA

- COMPARISON OF GLOBAL & LOCAL CONTRIBUTIONS
- COMBINING MULTIPLE TECHNIQUES

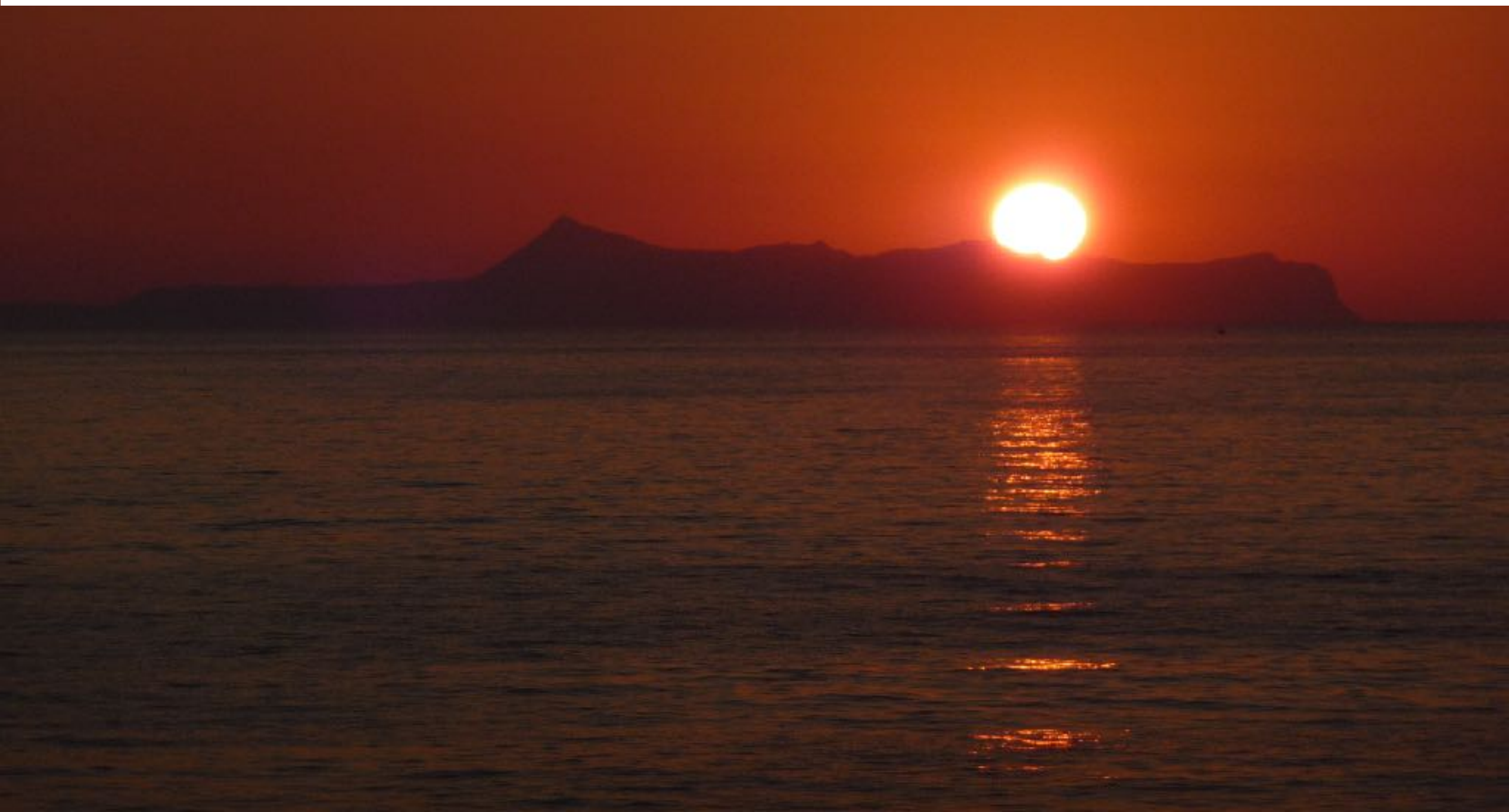


UNDERSTANDING VIA

- ITERATIVE INTERACTION
- COMPARISON OF GLOBAL & LOCAL CONTRIBUTIONS
- COMBINING MULTIPLE TECHNIQUES

HOW FAR DOES INTERPRETABILITY GO?

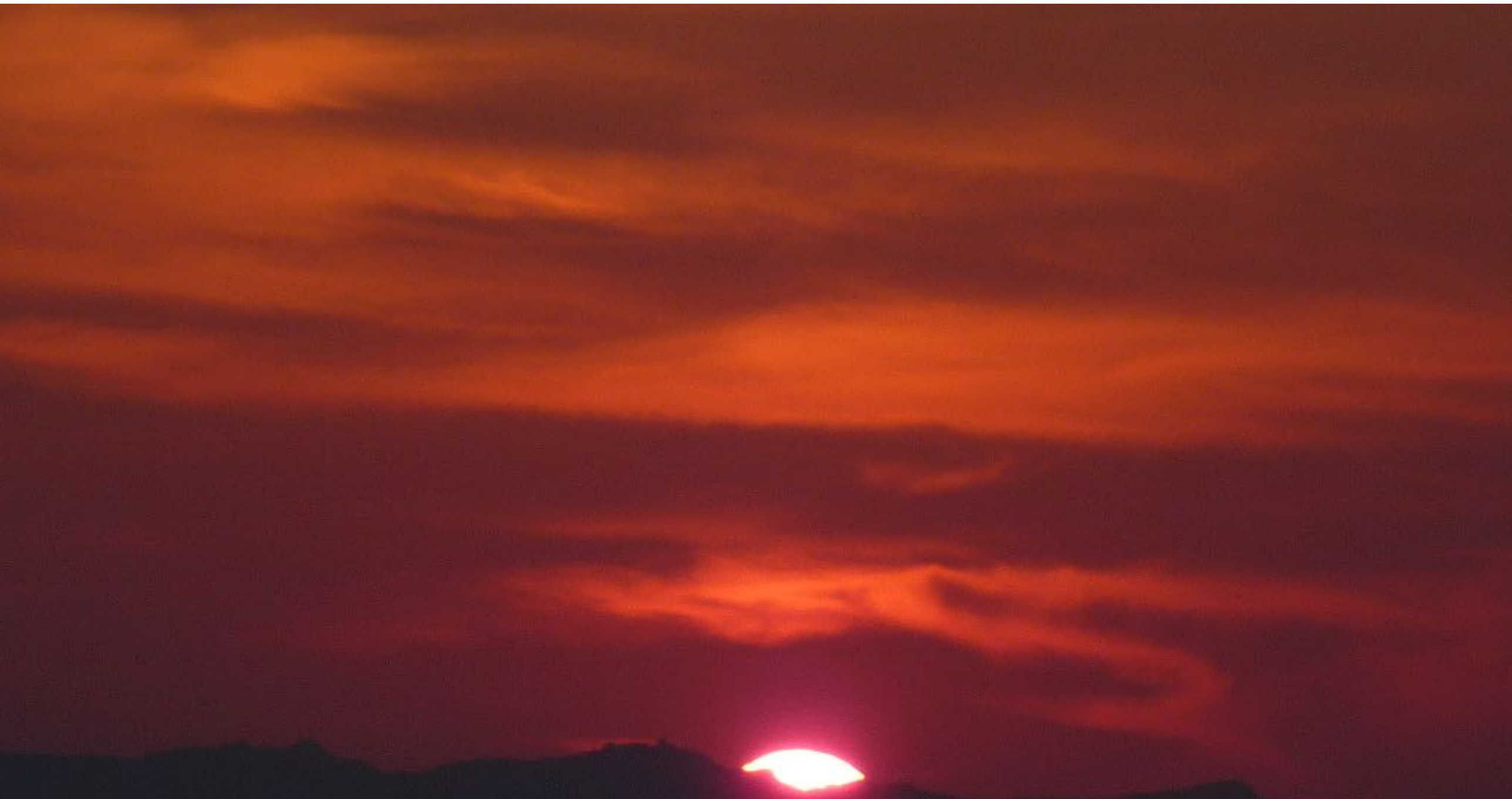
Data, Preprocessing, Feature Engineering,
Model Building, Model Selection, Drift Monitoring



INTERPRETABILITY – THE BEST OF TIMES?

Model: Sensitivity? Fidelity? Specific? Implementation compatible?

Code: Robustness? Licensing? Scalability?



dean_allsopp@hotmail.com

```
➜  iml ls
ContrastiveExplanation PDPbox                anchor                lime                  nonconformist
LZX                    SHAP                  defragTrees           manual_pdp_ice        pyBreakDown
LORE                   SKATER                kepler-mapper         monotonic             rulefit
```

Questions?

THANKS