SILO.AI

ARTIFICIAL INTELLIGENCE AS A SERVICE

BUILDING AI SYSTEMS WE CAN TRUST

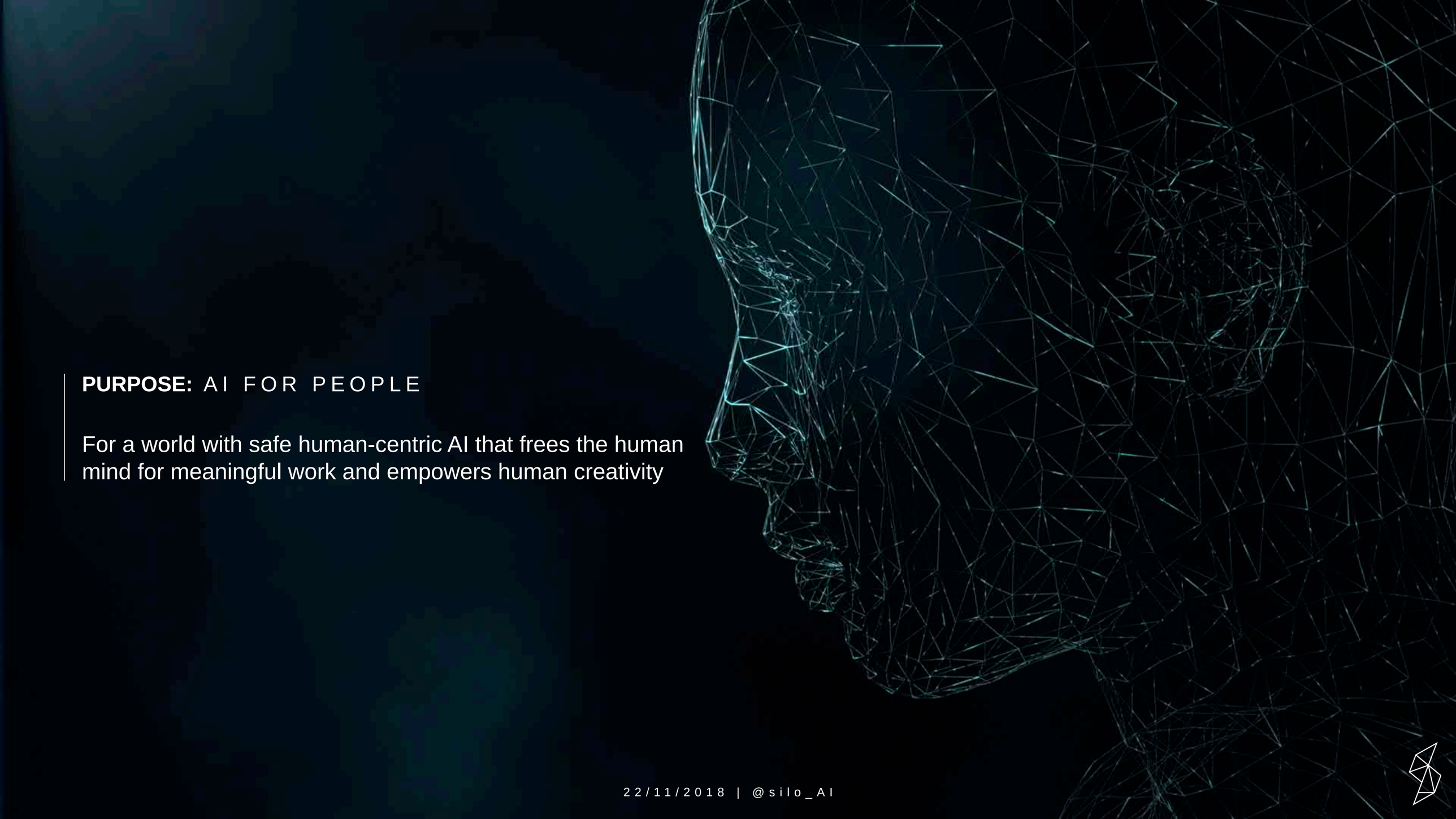
ERLIN GULBENKOGLU

SILO.AI

LARGEST PRIVATE AI LAB IN THE NORDICS

FULL AI SERVICE DESIGN CAPABILITY

20+ PHD:S



PURPOSE: AI FOR PEOPLE

For a world with safe human-centric AI that frees the human mind for meaningful work and empowers human creativity

SILO.AI WORKING MODEL

AI DEPLOYMENT ROADMAP

AI PRESTUDY

2–4 WEEKS

INSIGHTS

AI OPPORTUNITY CO-CREATION

ROADMAP

PRE-PRODUCTION AI

AI SOLUTION

2–4 MONTHS

DESIGN SPRINT & POC

BUSINESS SOLUTION

PRODUCTION-GRADE AI

AI PRIME

PROJECT SPECIFIC

EXPAND AND DEEPEN

MAINTENANCE

AI MODELS WE CAN TRUST…

- Explainability
- Transparency
- Reliability

AI MODELS WE CAN TRUST…

- **Explainability: Justify the predictions of your model**
- Transparency
- Reliability

AI MODELS WE CAN TRUST...

- Explainability
- **Transparency: What is important to your model?**
- Reliability

AI MODELS WE CAN TRUST…

- Explainability
- Transparency
- **Reliability: Ensure your model works in a reliable way**

GDPR Implications on AI

	Decision based ‘solely’ on automated processing	Information on the logic involved	Right to erasure	Data minimisation	Pseudonymisation
Article	Article 22	Article 13	Article 17	Article 5	Article 6
Challenge	AI = autonomy	Complexity of AI models	AI models memorising training data	Need for a feasibility test	Not sufficient
Solution	Human-in-the-loop	Model interpretability techniques	Private Learning: Differential Privacy	Difficult challenge :)	Strong de-identification techniques

GDPR Implications on AI

	Decision based ‘solely’ on automated processing	Information on the logic involved	Right to erasure	Data minimisation	Pseudonymisation
Article	Article 22	Article 13	Article 17	Article 5	Article 6
Challenge	AI = autonomy	Complexity of AI models	AI models memorising training data	Need for a feasibility test	Not sufficient
Solution	Human-in-the-loop	Model interpretability techniques	Private Learning: Differential Privacy	Difficult challenge :)	Strong de-identification techniques

HUMAN-IN-THE-LOOP

AI = MACHINES + HUMANS ≠ AUTONOMY

AI = TD + ML + HITL ≠ BEST ALGORITHM

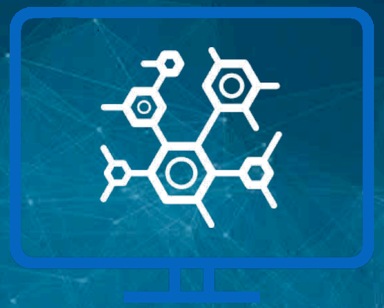
HUMAN-IN-THE-LOOP

DATA

FIND PATTERNS

TRAIN ALGORITHM

MODEL



LEARNING LOOP

DEPLOY MODEL

NEW DATA

MACHINE SIGNALS

HUMAN INTEL & DECISIONS

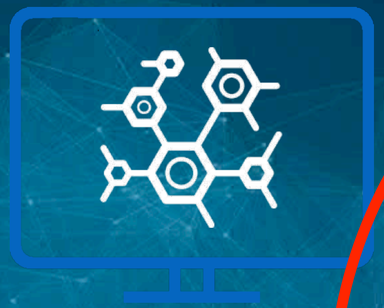
HUMAN-IN-THE-LOOP

DATA

FIND PATTERNS

TRAIN ALGORITHM

MODEL



DEPLOY MODEL

NEW DATA

LEARNING LOOP

INTERACTIVE LEARNING

MACHINE SIGNALS

HUMAN INTEL & DECISIONS

- Fraud detection
- Fault detection
- Assessing documents

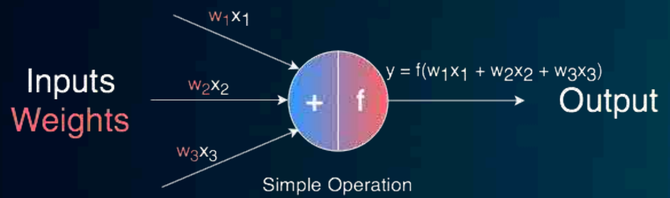
INFORMATION ON THE LOGIC INVOLVED

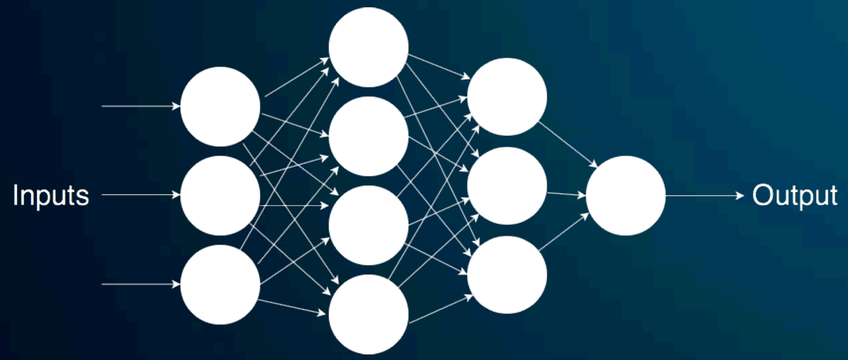
1 Explain how your ML model works

2 Explain the reasons behind your predictions

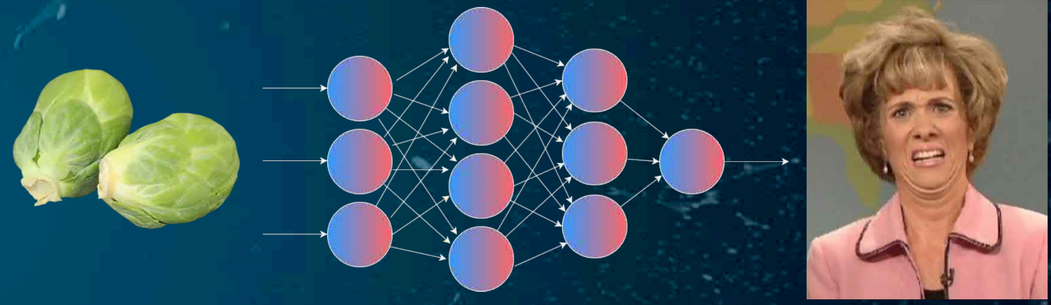
INFORMATION ON THE LOGIC INVOLVED

1 Explain how your ML model works

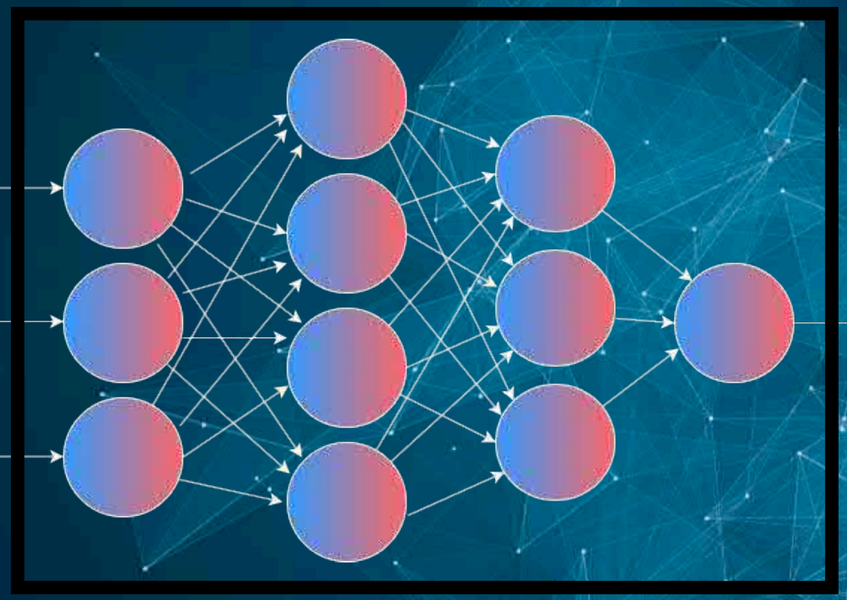


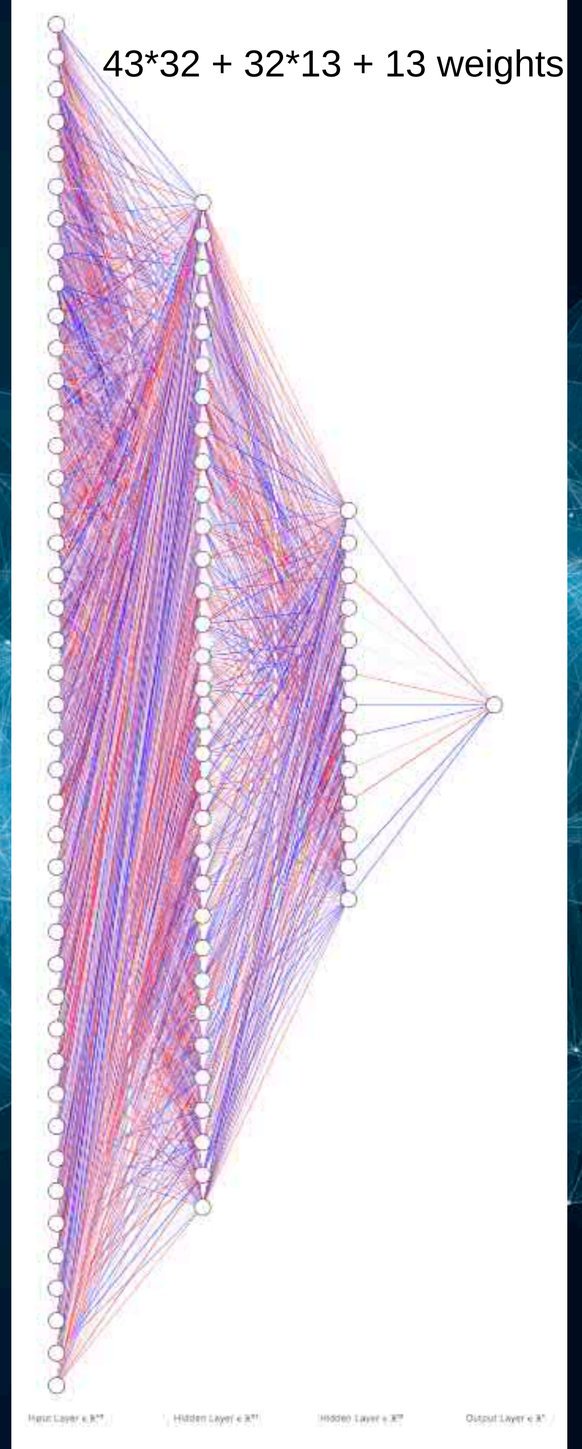


2 Explain the reasons behind your predictions



2 Explain the reasons behind your predictions





- How to explain a prediction when hundreds of variables are involved?

2 Explain the reasons behind your predictions

- Is it possible to interpret any model/method's results?

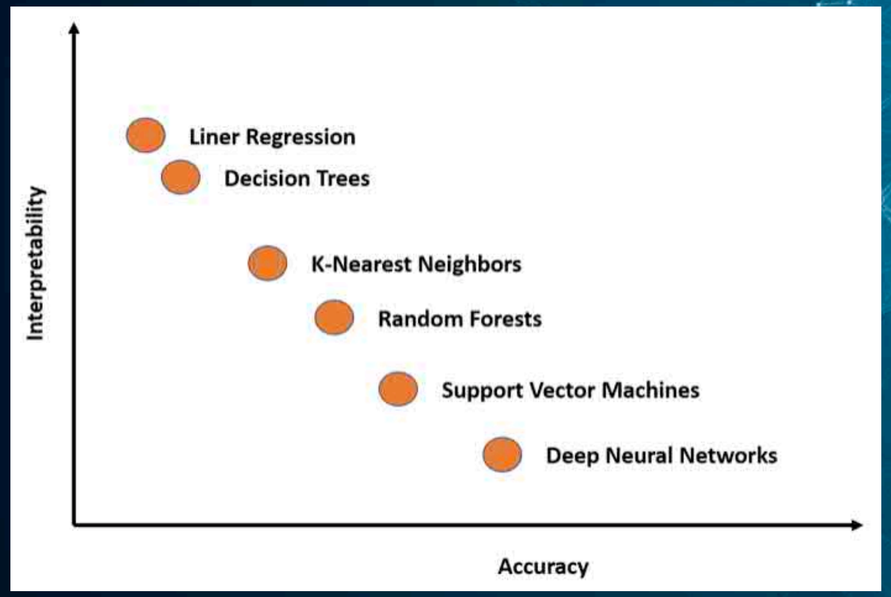
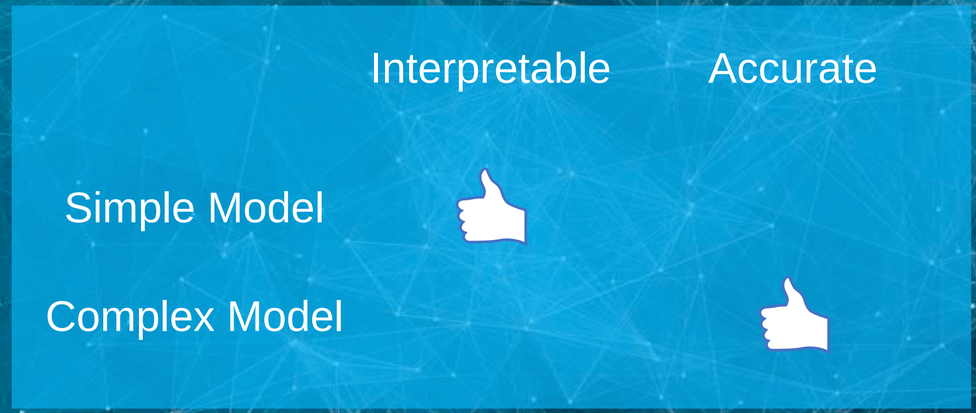


Figure taken from: Rodriguez, J. (2018, June, 6). *Interpretability vs. Accuracy: The Friction that Defines Deep Learning. Retrieved from* https://towardsdatascience.com/interpretability-vs-accuracy-the-friction-that-defines-deep-learning-dae16c84db5c



It is hard to explain complex models!

2 Explain the reasons behind your predictions

- How to be sure that the prediction generated by a black box model is reliable?

  “If the users do not trust a model or a prediction, they will not use it.”

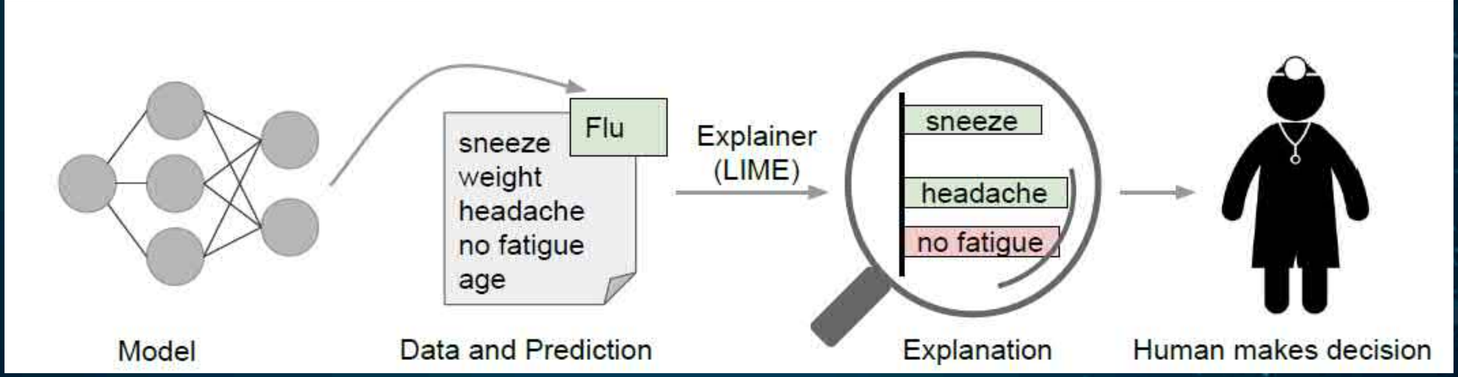


Figure taken from: *Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.*

2 Explain the reasons behind your predictions

- How to explain a prediction when hundreds of variables are involved?
- Is it possible to interpret any model/method's results?
- How to be sure that the prediction generated by a black box model is reliable?

LIME

LRP

Don't explain the whole model just explain one prediction!

SHAP

DeepLIFT

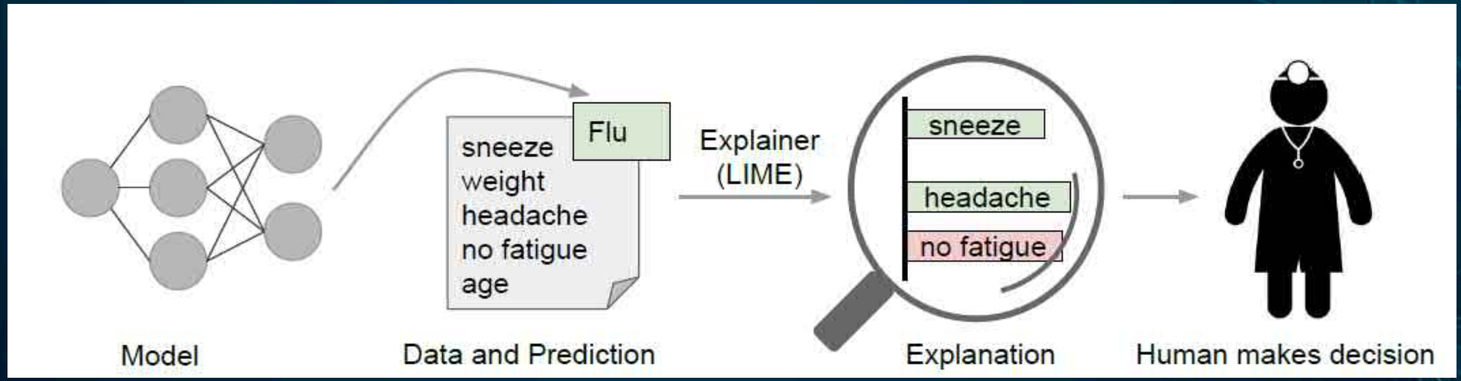


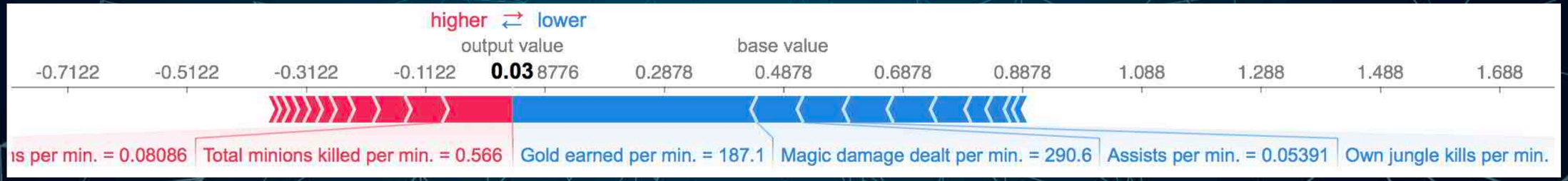
Figure taken from: *Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016.*

2 Explain the reasons behind your predictions

SHAP

Application for predicting LoL winners

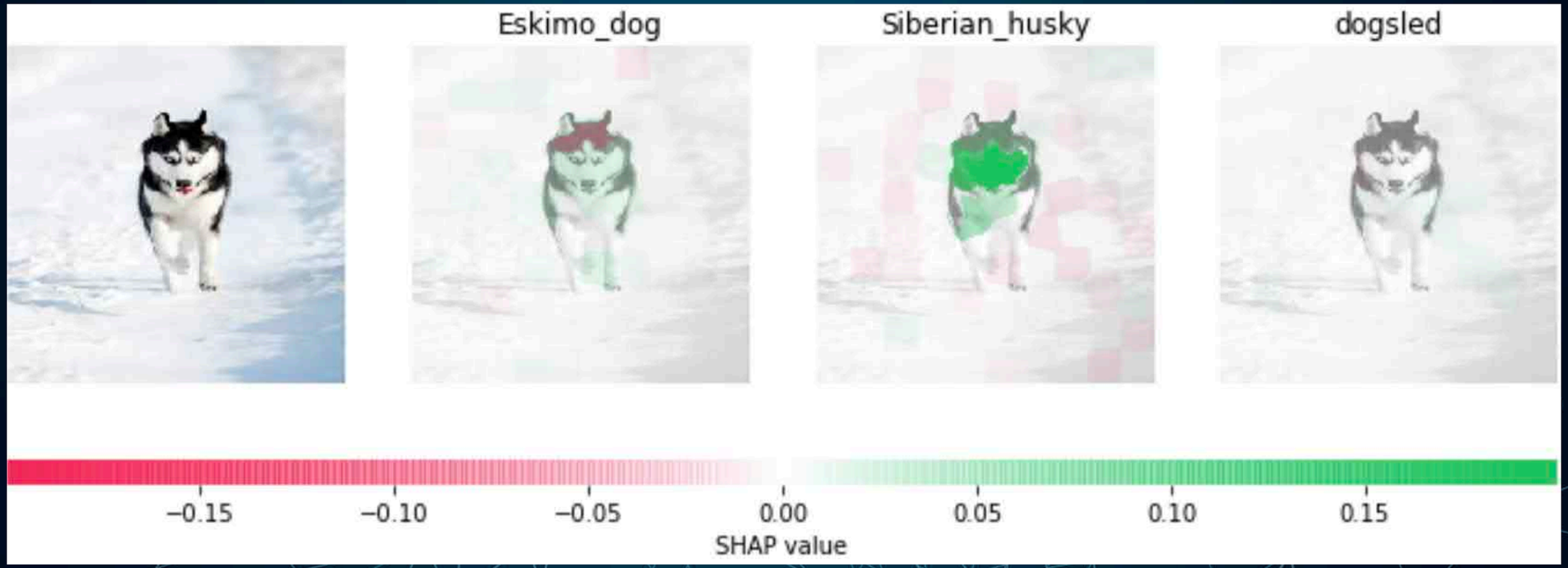
Individualised Explanations



2 Explain the reasons behind your predictions

SHAP

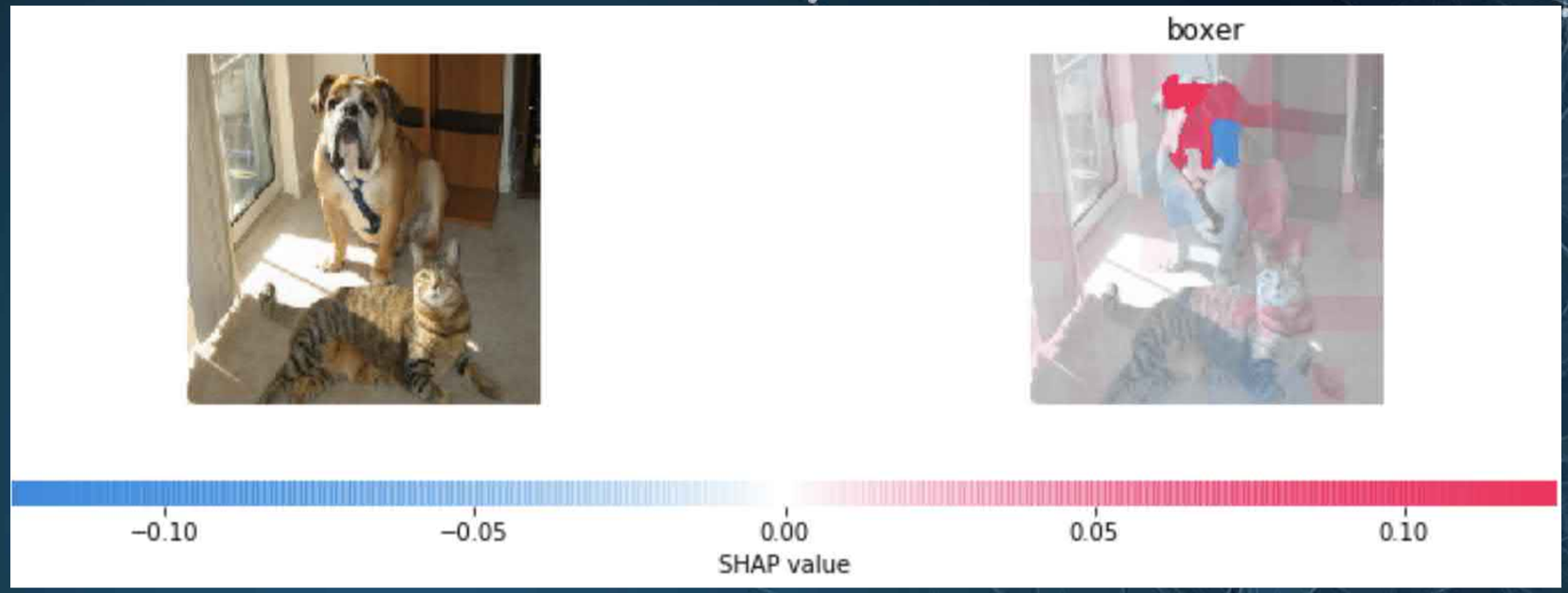
Classifications with InceptionV3

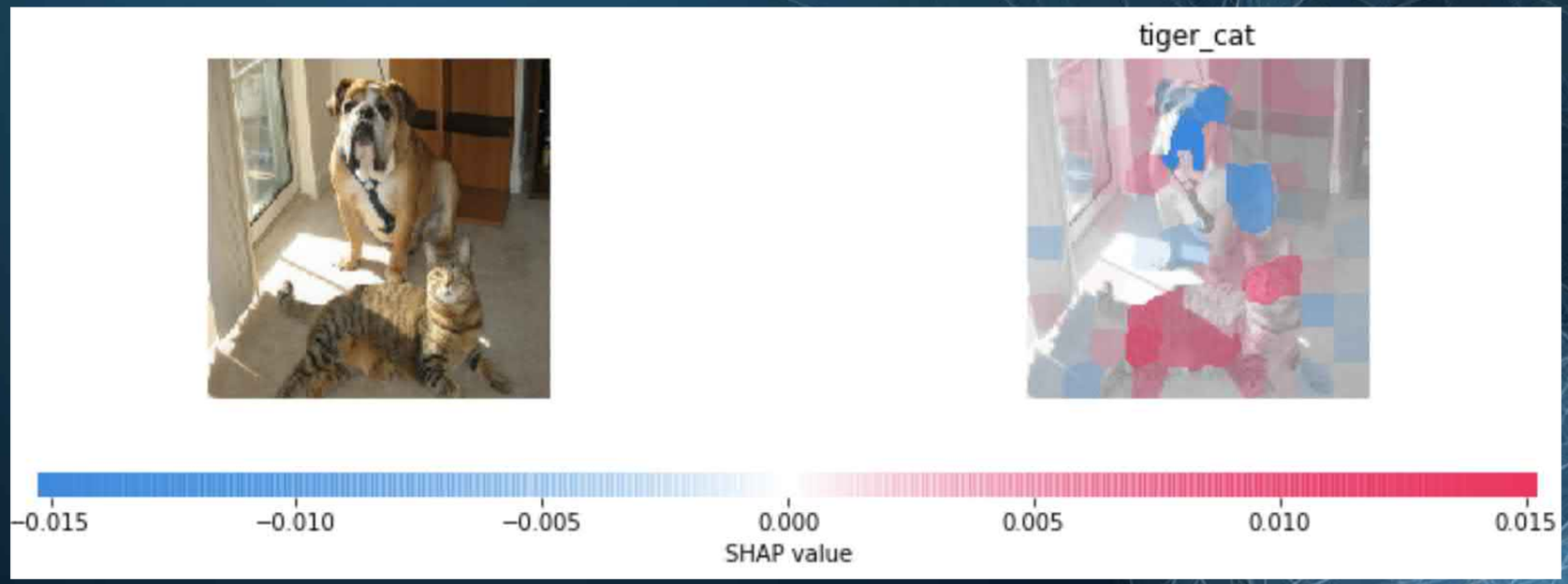


SHAP

Kernel SHAP

Classifications with VGG16

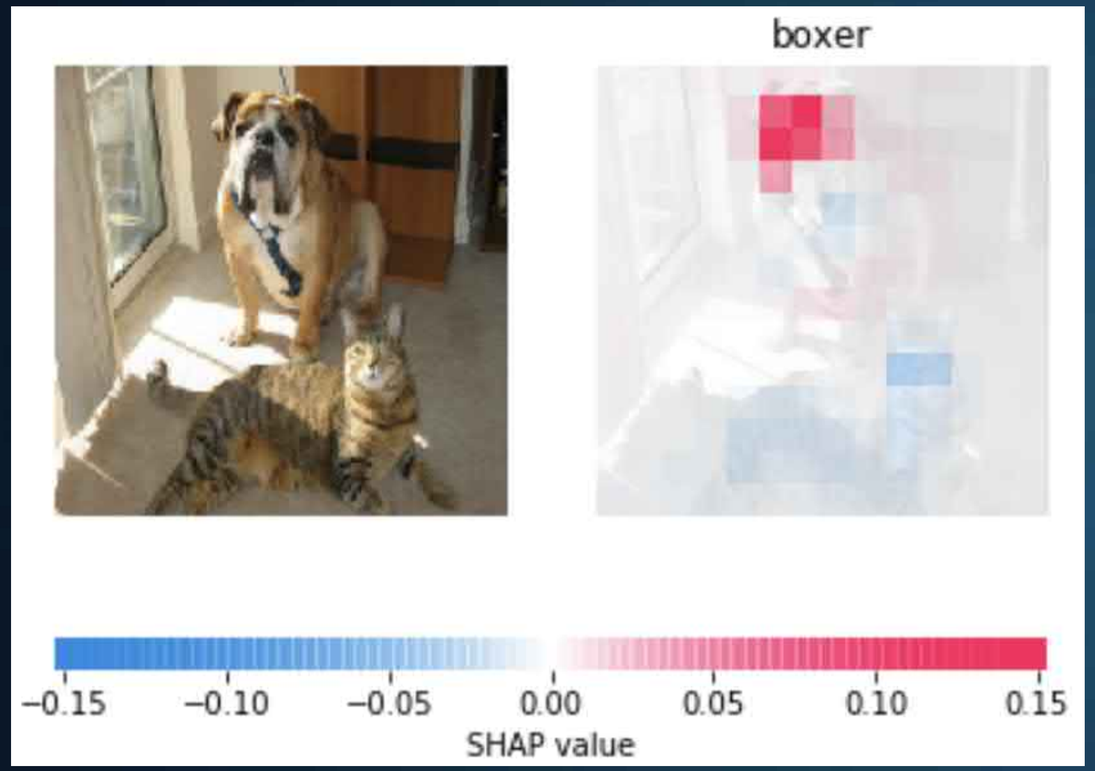




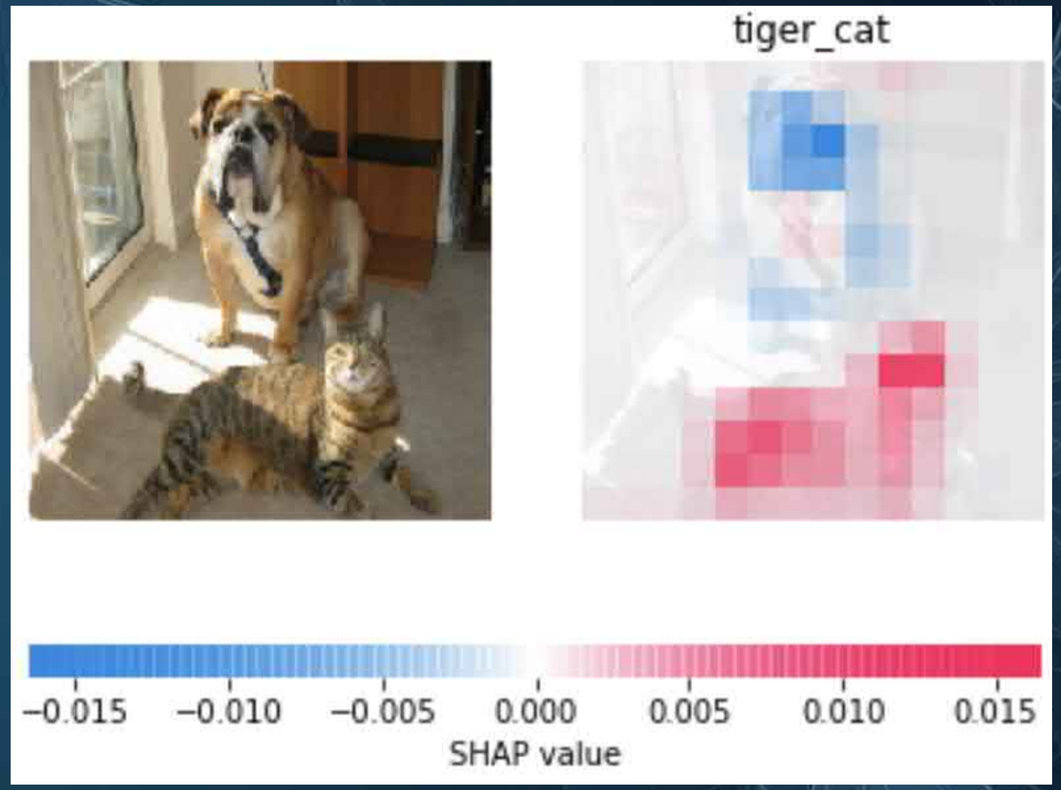
CLASS NAME	INDEX	PROBABILITY
boxer	(242)	0.420
bull_mastiff	(243)	0.282
tiger_cat	(282)	0.053

SHAP

Gradient SHAP for Conv Block 5



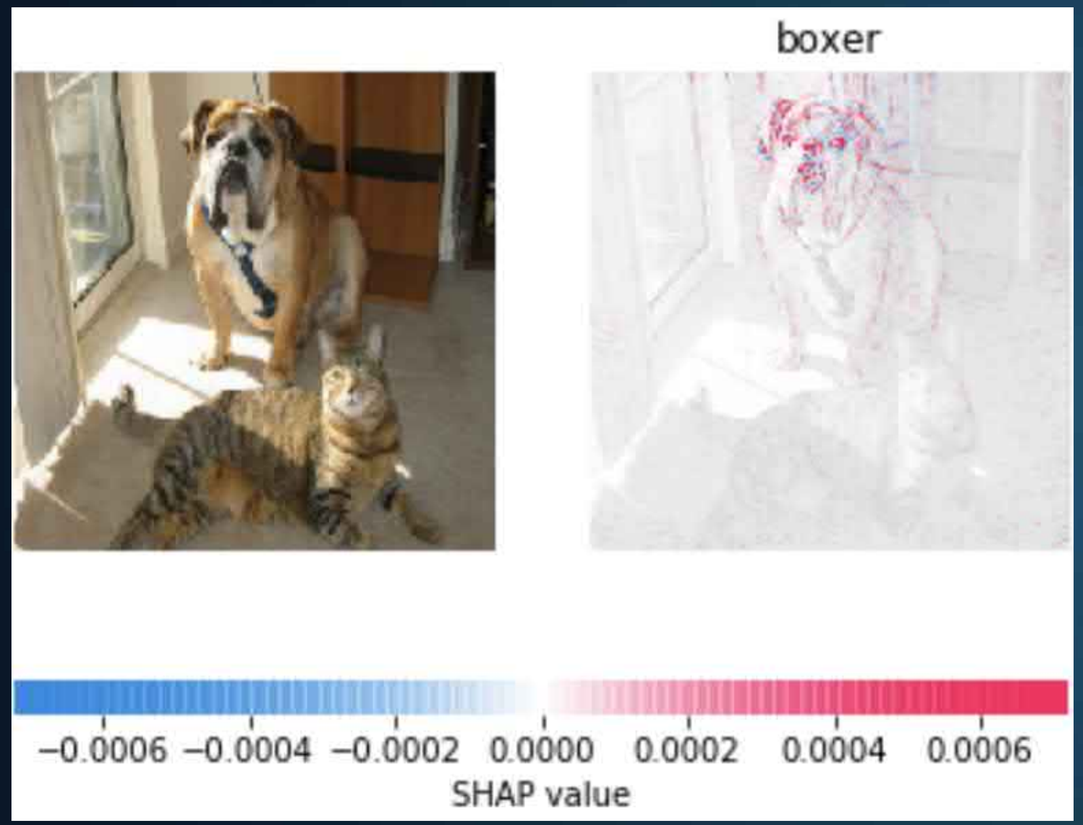
Classifications with VGG16



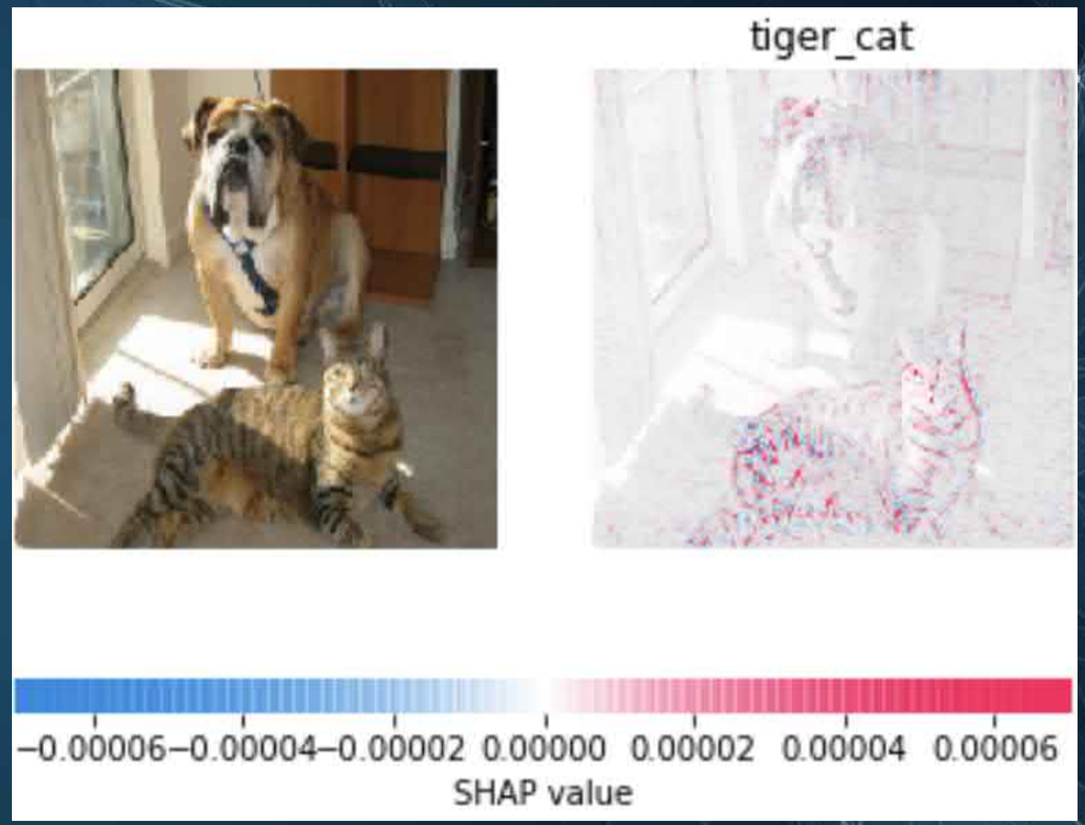
CLASS NAME	INDEX	PROBABILITY
boxer	(242)	0.420
bull_mastiff	(243)	0.282
tiger_cat	(282)	0.053

SHAP

Deep SHAP



Classifications with VGG16



CLASS NAME	INDEX	PROBABILITY
boxer	(242)	0.420
bull_mastiff	(243)	0.282
tiger_cat	(282)	0.053

GradCAM

“boxer”

“tiger_cat”

CLASS NAME	INDEX	PROBABILITY
boxer	(242)	0.420
bull_mastiff	(243)	0.282
tiger_cat	(282)	0.053

LRP

“boxer”

“tiger_cat”

CLASS NAME	INDEX	PROBABILITY
boxer	(242)	0.420
bull_mastiff	(243)	0.282
tiger_cat	(282)	0.053

SHAP for NLP

jupyter SHAP NLP applied to IMDB Sentiment Analysis Last Checkpoint: tunti sitten (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
    num2word[words[w]] = w
x_test_words = np.stack([np.array(list(map(lambda x: num2word.get(x, "NONE"), x_test[i]))) for i in range(10)])
```

In [11]:
```
# plot the explanation of the first prediction
# Note the model is "multi-output" because it is rank-2 but only has one column
shap.force_plot(explainer.expected_value[0], shap_values[0][0], x_test_words[0])
```

Out[11]:

In [40]:
```
color_text(tokens=x_test_words[0], values=shap_values[0][0],
           cf=cf, thrs_neg=-0.025, thrs_pos=0.025, ignore_words=set(['NONE']))
```

68

the wonder own as by is sequence i i jars roses to of hollywood br of down shouting getting boring of ever it sadly sadly sadly i i was then does don't close faint after one carry as by are be favourites all family turn in does as three part in another some to be probably with world uncaring her an have faint beginning own as is sequence

In [42]:
```
# plot the explanation of the prediction
# Note the model is "multi-output" because it is rank-2 but only has one column
shap.force_plot(explainer.expected_value[0], shap_values[0][1], x_test_words[1])
```

Out[42]:

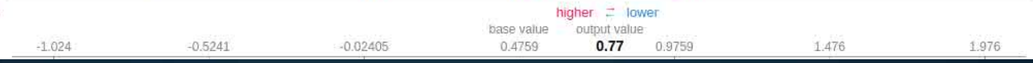
In [41]:
```
color_text(tokens=x_test_words[1], values=shap_values[0][1],
           cf=cf, thrs_neg=-0.025, thrs_pos=0.025, ignore_words=set(['NONE']))
```

80

drugs keep guy i i was throwing room sugar as it by br be plot many for occasionally film verge boyfriend difficult kid as you it failed not if gerard to if woman in launching is police fi spooky or of self what have pretty in can so suit you good 2 which why super as it main of my i i if time screenplay in same this remember assured have action one in realistic that better of lessons

In [43]:
```
# plot the explanation of the first prediction
# Note the model is "multi-output" because it is rank-2 but only has one column
shap.force_plot(explainer.expected_value[0], shap_values[0][2], x_test_words[2])
```

Out[43]:



Why interpretability matters?

1) explaining a prediction, for example, the user can interpret an individual prediction sufficiently to take some action based on it.

2) being able to explain machine learning based decisions improves transparency and acceptance of complex models.

3) assessing a model in detail, for example, the developer can evaluate the model and identify the failure modes.

SILO.AI

ARTIFICIAL INTELLIGENCE AS A SERVICE

TERO OJANPERÄ
CEO
tero.ojanpera@silo.ai
+358 40 558 3096

ERLIN GULBENKOGLU
Data Privacy Expert
erlin.gulbenkoglu@silo.ai
+358 44 924 1961