ESEM 2023
Empirical Software Engineering and Measurement
October 2023 - New Orleans, LA, USA

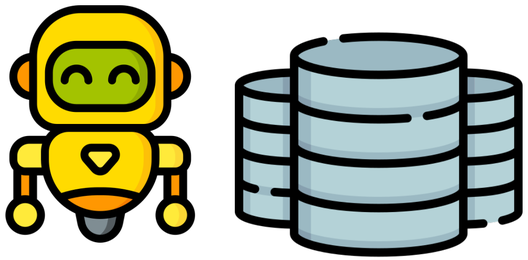
AN EMPIRICAL STUDY ON LOW- AND HIGH-LEVEL EXPLANATIONS OF DEEP LEARNING MISBEHAVIOURS

TAHEREH ZOHDINASAB

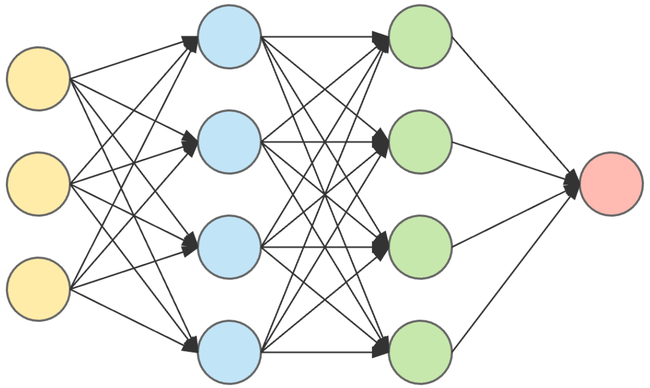
VINCENZO RICCIO

PAOLO TONELLA

DEEP LEARNING (DL) SYSTEM ASSESSMENT



TEST GENERATORS &
ORIGINAL TEST SET



DL SYSTEM
UNDER TEST

ACC = 95%

PERFORMANCE
METRIC

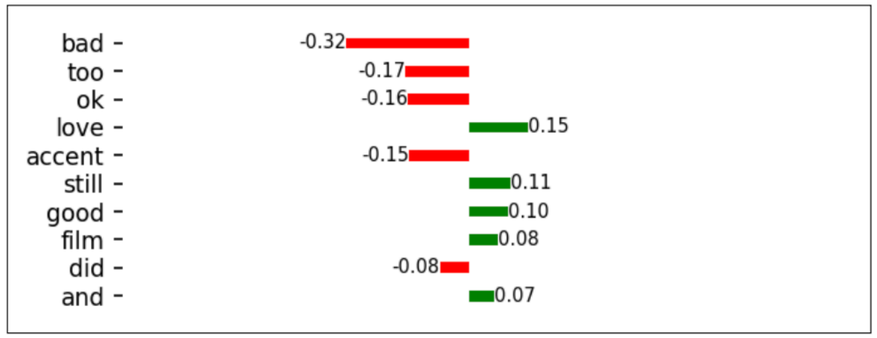
How can we explain the misbehaviours of DL systems?

OPAQUENESS OF DL

LOW-LEVEL EXPLANATIONS

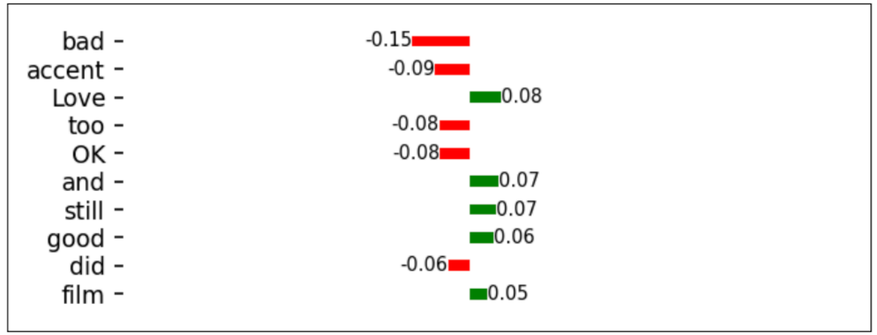
Integrated Gradients

I've taken another look at this film and still consider it pretty good. Chloe is one of the few hardcore stars who really can act. She appears occasionally in soft core such as "Body of Love" and "Lady Chatterly's Stories" on Showtime. I thought Nicole Hilbig did OK too with her nice body and charming accent. Too bad she's not in more films.



LIME

I've taken another look at this film and still consider it pretty good. Chloe is one of the few hardcore stars who really can act. She appears occasionally in soft core such as "Body of Love" and "Lady Chatterly's Stories" on Showtime. I thought Nicole Hilbig did OK too with her nice body and charming accent. Too bad she's not in more films.



HIGH-LEVEL EXPLANATIONS

Feature Maps

Features

MNIST

- Boldness
- Discontinuity
- Orientation

- IMDB
- Positive Words
- Negative Words
- Verbs

COMPARISON BETWEEN LOW-LEVEL AND HIGH-LEVEL EXPLANATIONS

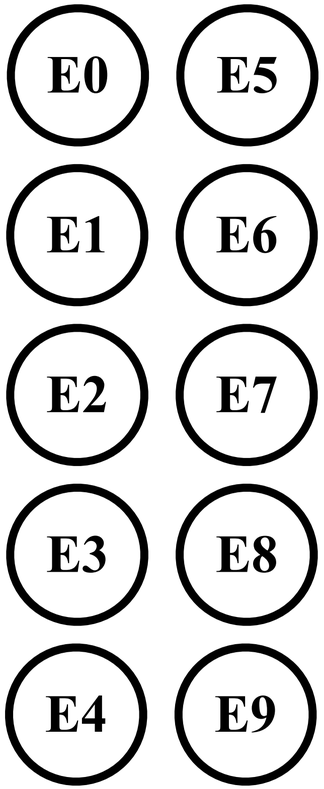
STEP 1

Sample Generation

Original test set + automatically generated

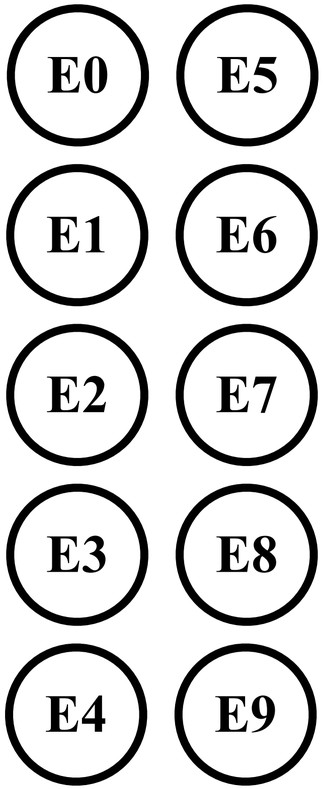
STEP 1

Sample Generation



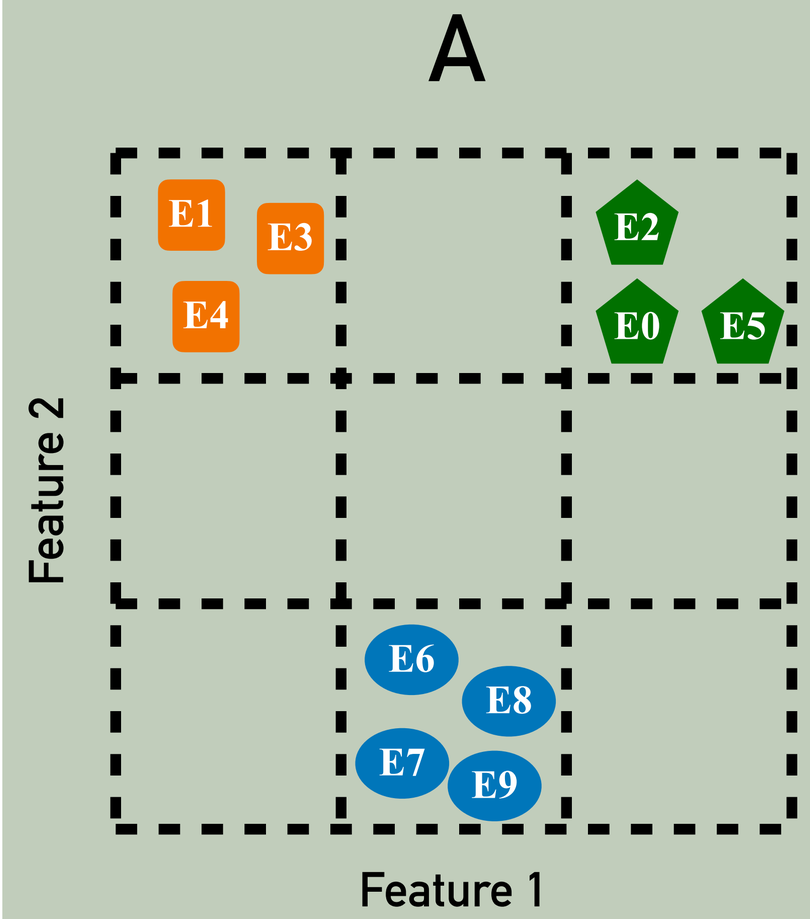
Original test set + automatically generated

STEP 1
Sample Generation

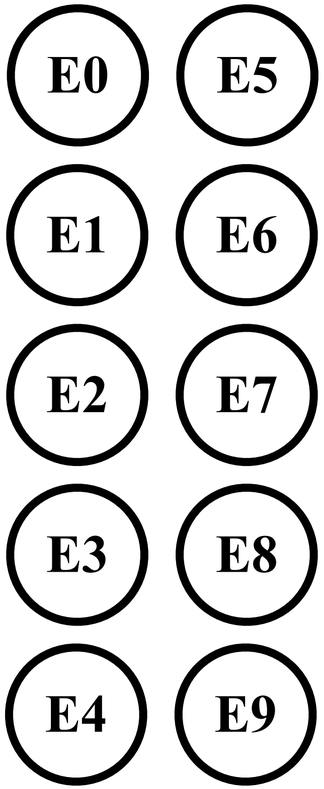


Original test set + automatically generated

STEP 2
Feature Map Computation



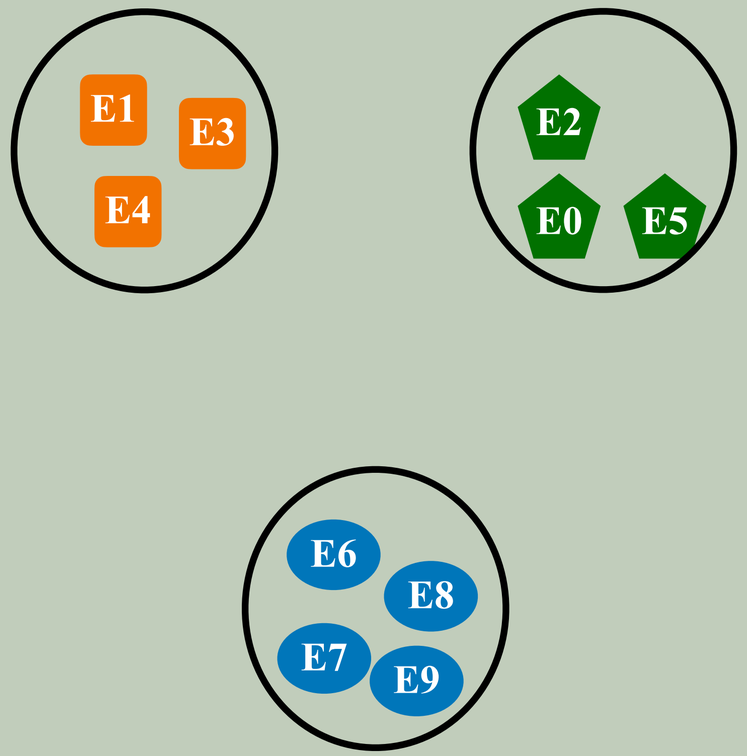
STEP 1
Sample Generation



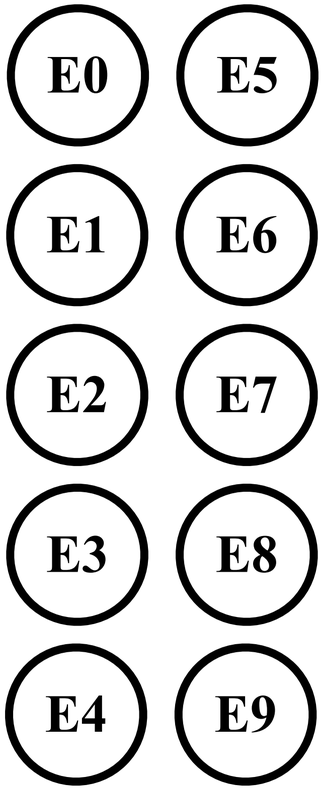
Original test set + automatically generated

STEP 2
Feature Map Computation

A



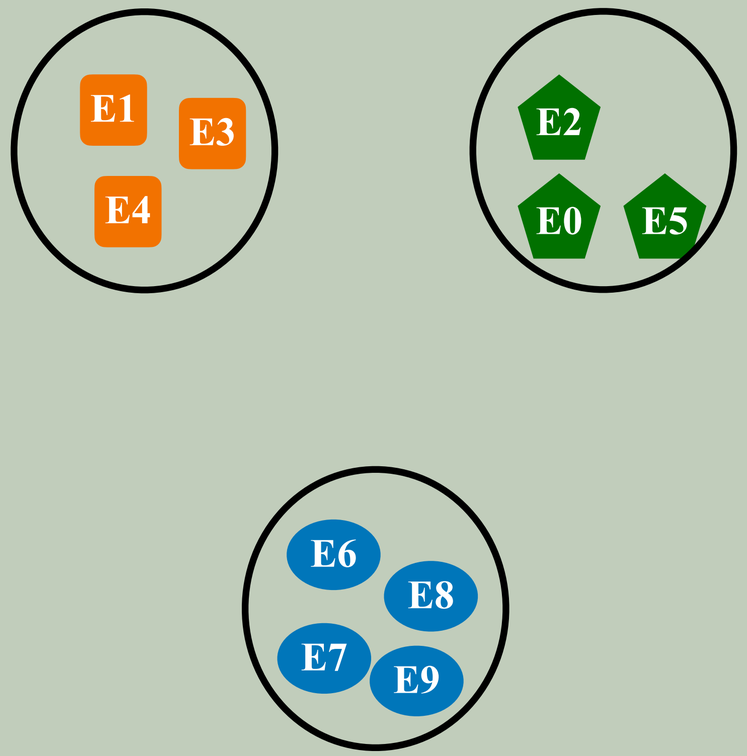
STEP 1
Sample Generation



Original test set + automatically generated

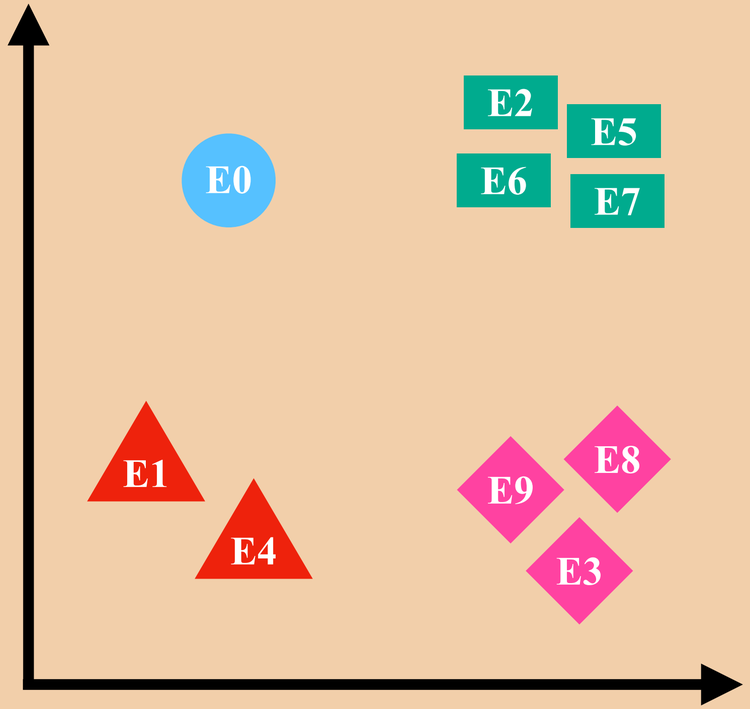
STEP 2
Feature Map Computation

A



STEP 3
LL Explanation Clustering

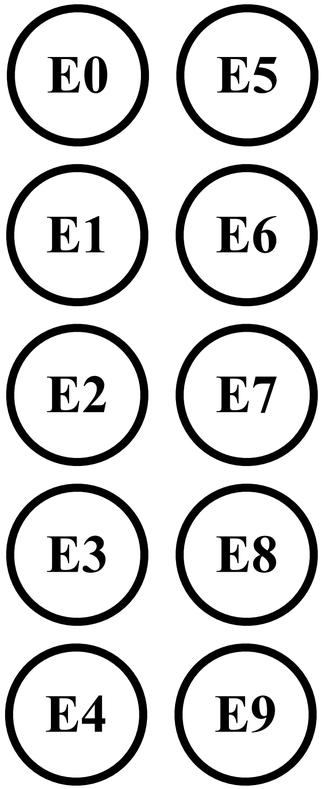
B



Alternative input spaces:

1. Original space
2. Latent Space

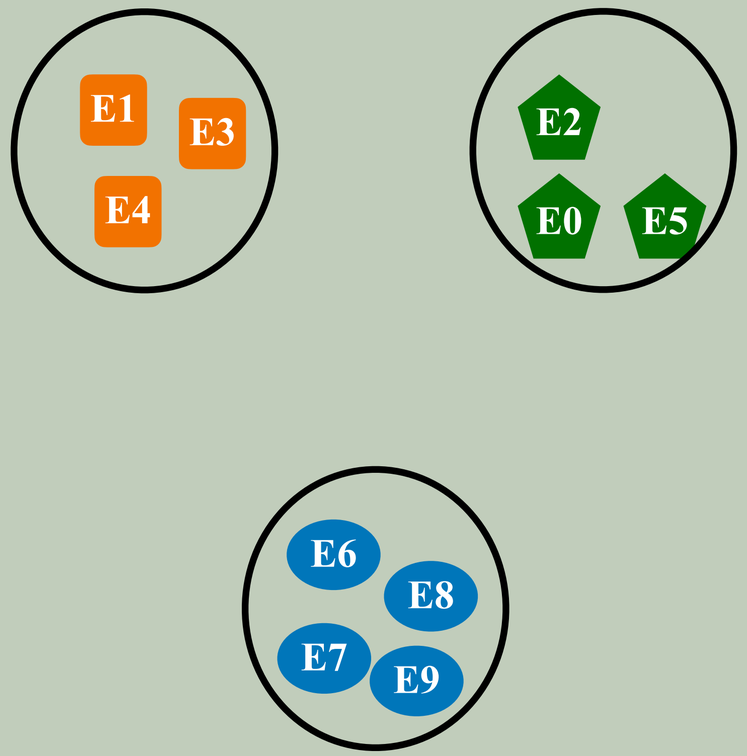
STEP 1
Sample Generation



Original test set + automatically generated

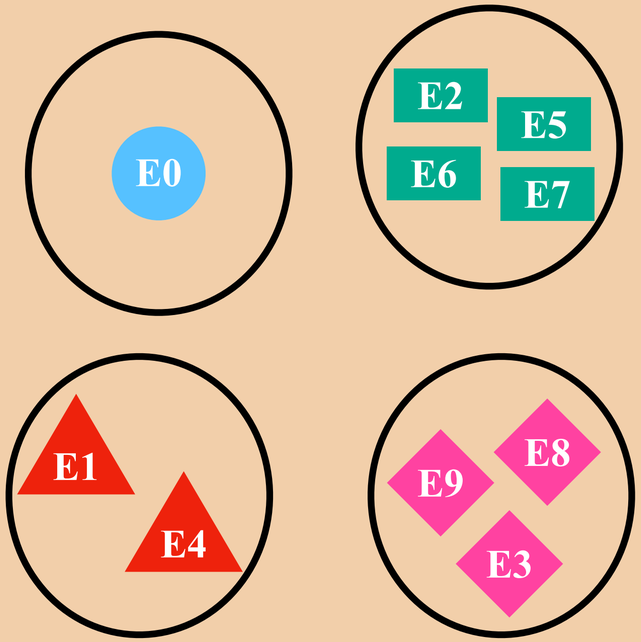
STEP 2
Feature Map Computation

A

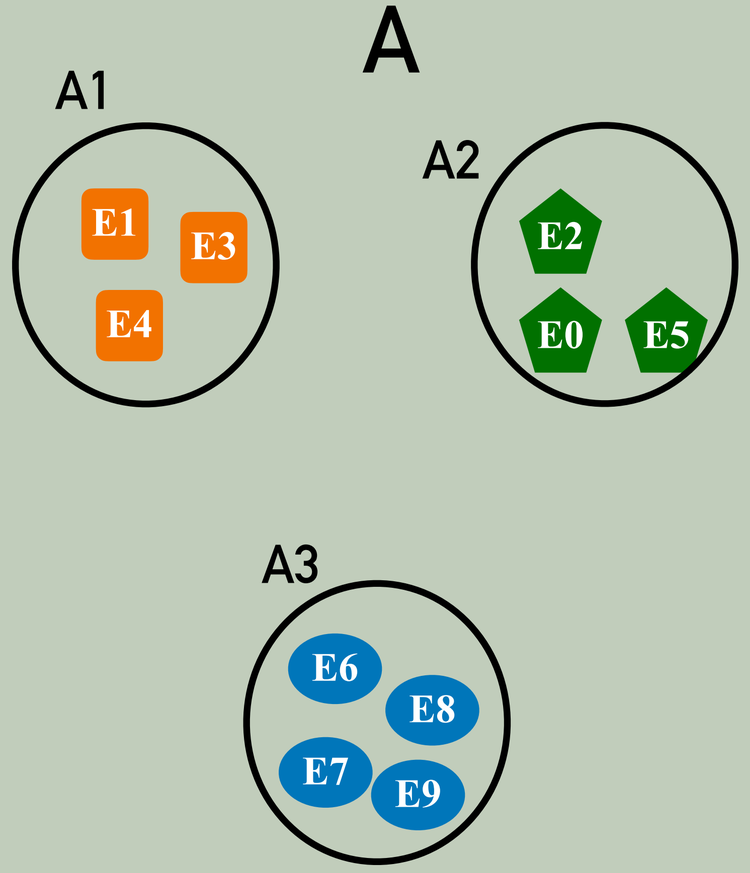


STEP 3
LL Explanation Clustering

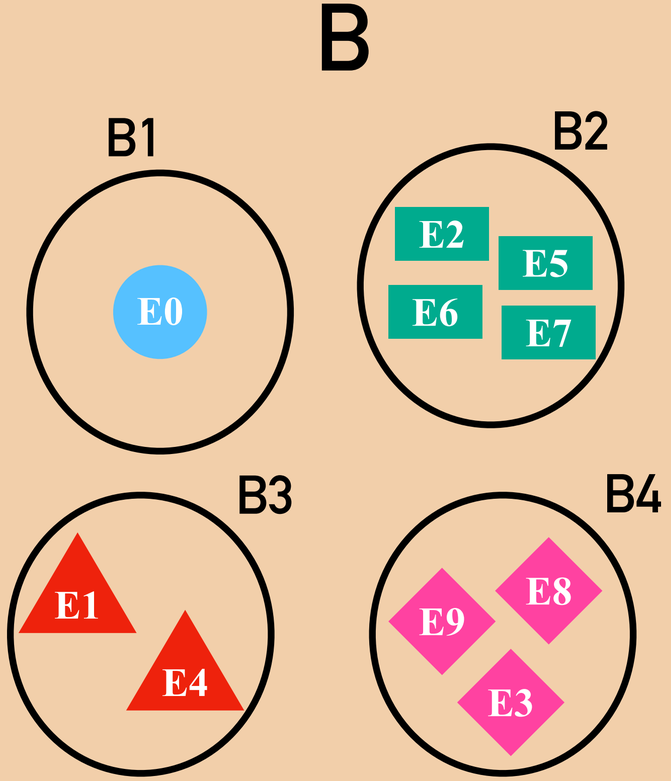
B



STEP 2
Feature Map Computation



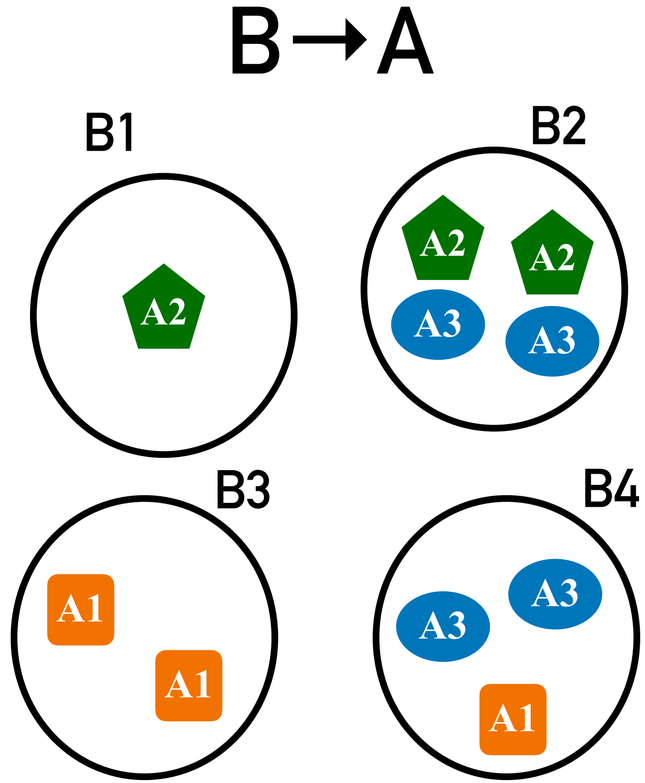
STEP 3
LL Explanation Clustering



Alternative distance metrics:

1. Original space
2. Latent Space

STEP 4
Gini Similarity



$$GS_{(B,A)} = 1 - \frac{1}{|B|}\sum_{i=1}^{|B|} GI(CB_i, A)$$

RQ1: SIMILARITY

RQ1: SIMILARITY

MNIST IMDB

High Level	Low Level	Input space	GSim	GSim
3D	IG	Original	**0.70**	**0.74**
		Latent	0.55	0.68
	LIME	Original	0.55	**0.81**
		Latent	0.53	0.66
2D	IG	Original	**0.76**	**0.76**
		Latent	0.49	0.56
	LIME	Original	**0.62**	**0.80**
		Latent	0.47	0.59
1D	IG	Original	**0.85**	**0.83**
		Latent	0.59	0.66
	LIME	Original	**0.75**	**0.85**
		Latent	0.59	0.68

RQ1: SIMILARITY

MNIST IMDB

High Level	Low Level	Input space	GSim	GSim
3D	IG	Original	**0.70**	**0.74**
		Latent	0.55	0.68
	LIME	Original	0.55	**0.81**
		Latent	0.53	0.66
2D	IG	Original	**0.76**	**0.76**
		Latent	0.49	0.56
	LIME	Original	**0.62**	**0.80**
		Latent	0.47	0.59
1D	IG	Original	**0.85**	**0.83**
		Latent	0.59	0.66
	LIME	Original	**0.75**	**0.85**
		Latent	0.59	0.68

1D Feature Maps and Original Space achieve the highest similarity, but with the highest difference in # of clusters (up to 39)

RQ1: SIMILARITY

MNIST IMDB

High Level	Low Level	Input space	GSim	GSim
3D	IG	Original	**0.70**	**0.74**
		Latent	0.55	0.68
	LIME	Original	0.55	**0.81**
		Latent	0.53	0.66
2D	IG	Original	**0.76**	**0.76**
		Latent	0.49	0.56
	LIME	Original	**0.62**	**0.80**
		Latent	0.47	0.59
1D	IG	Original	**0.85**	**0.83**
		Latent	0.59	0.66
	LIME	Original	**0.75**	**0.85**
		Latent	0.59	0.68

2D Feature Maps and Latent Space achieve the lowest difference in # of clusters, but the lowest similarity

RQ1: CONCLUSIONS

High-level explanations based on human experience and low-level XAI techniques partition inputs in different ways

RQ2: UNDERSTANDABILITY

RQ2: UNDERSTANDABILITY

Survey: 48 SE experts

The digit is bold, oriented to left and very continuous.

The following highlighted pixels:

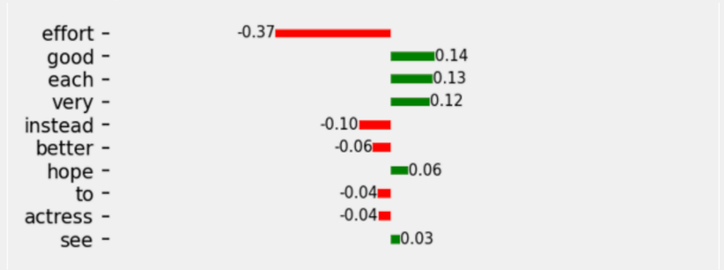
The following highlighted regions:

A very good offering from HBO. Traci Lords is becoming a much-better dramatic actress with each effort. I hope to see this attractive lady in more challenging roles in the future, instead of the "flighty" roles she has been stuck with in the past.

☐ The review contains 3 positive words, 3 negative words and 7 verbs (the number of verbs is an indicator of the text complexity).

☐ The review contains the following words contributing to negative (red) and positive (green) sentiments:

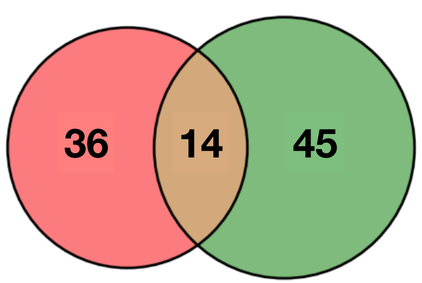
☐ The review contains the following words contributing to negative (red) and positive (green) sentiments:

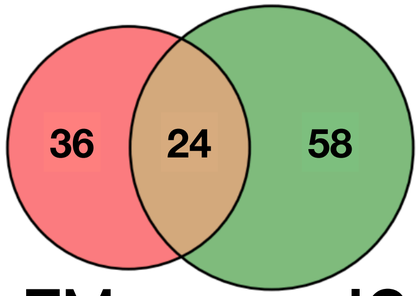


RQ2: UNDERSTANDABILITY

# of times the explanation matches with human expectations

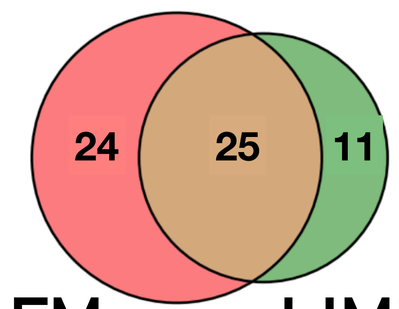
MNIST

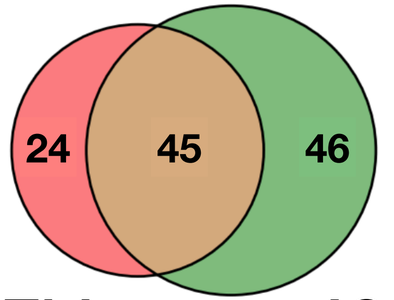




None: 108

IMDB



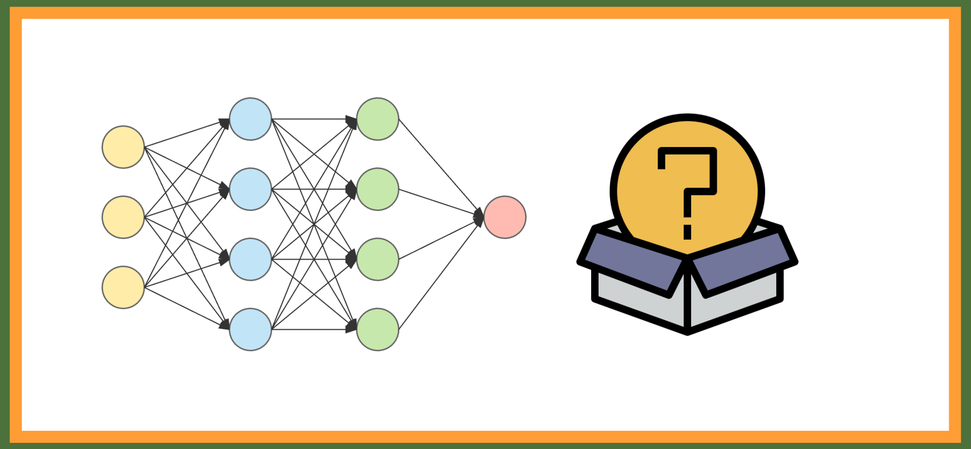


None: 25

RQ2: CONCLUSIONS

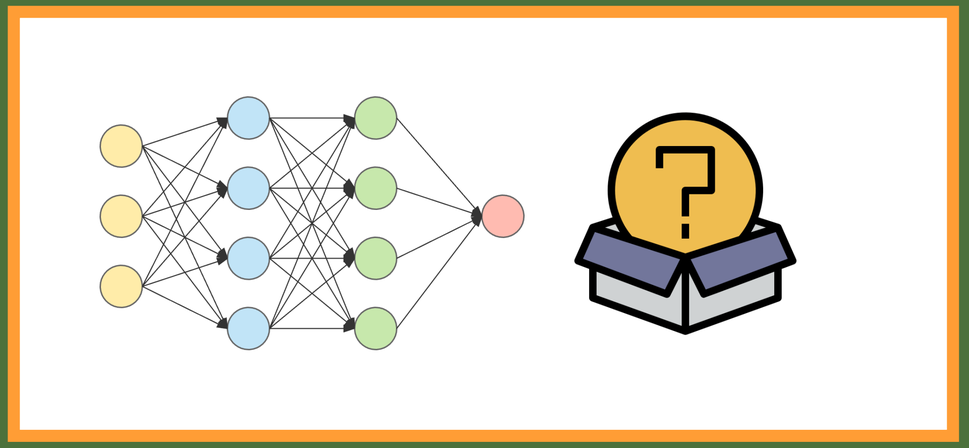
- High- and low-level explanations provide complementary insights
- Current explanations are not always satisfactory

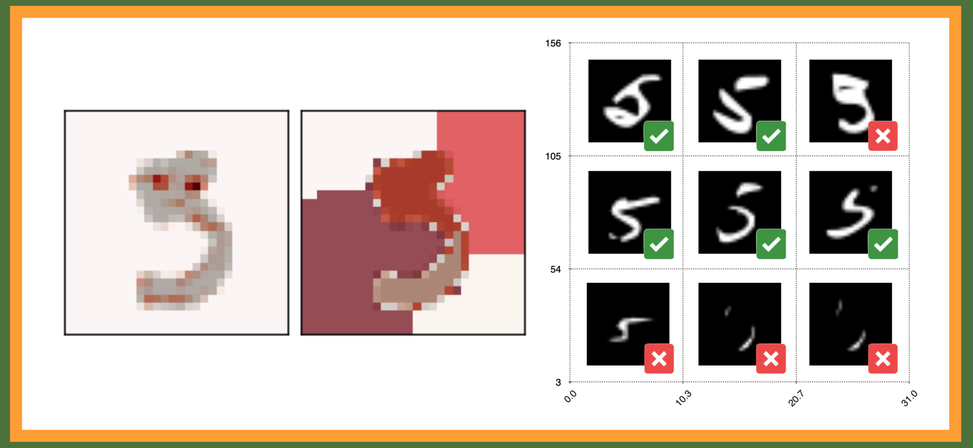
SUMMARY



Icons from www.flaticon.com

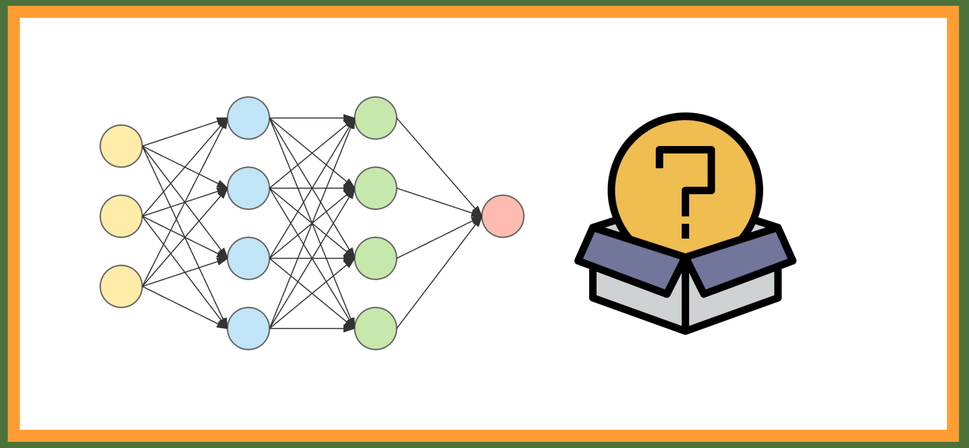
SUMMARY

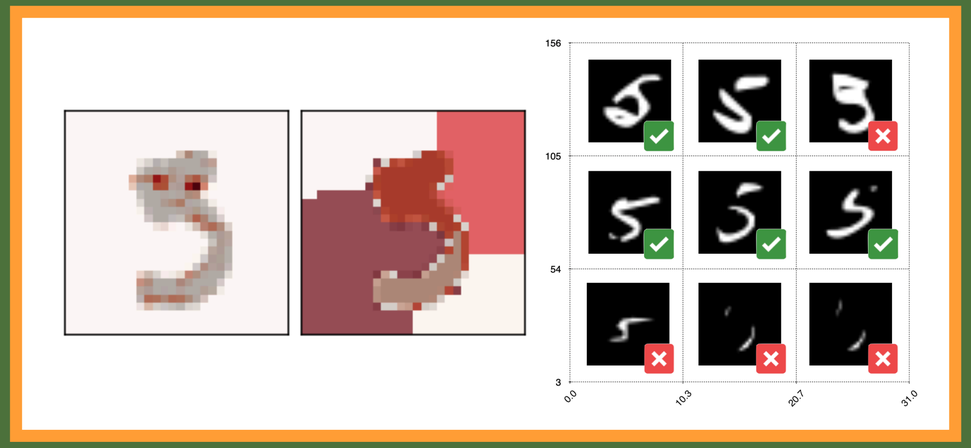


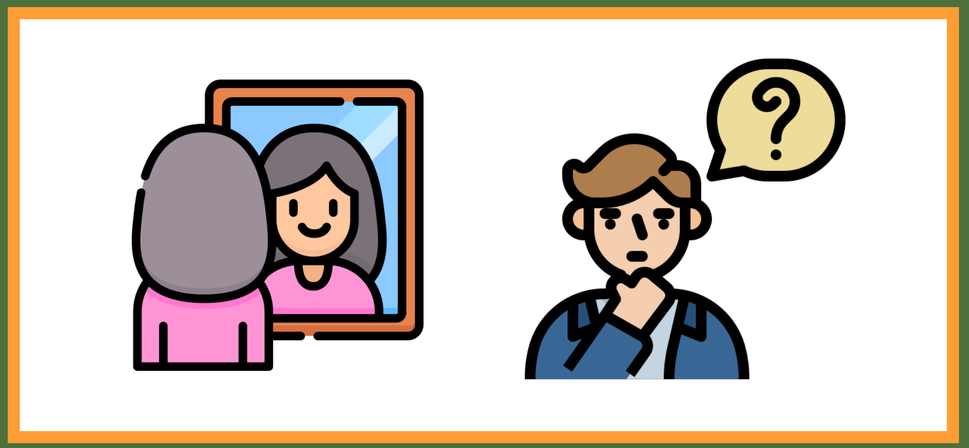


Icons from www.flaticon.com

SUMMARY

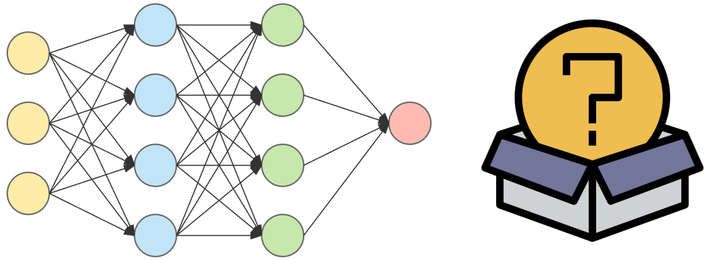


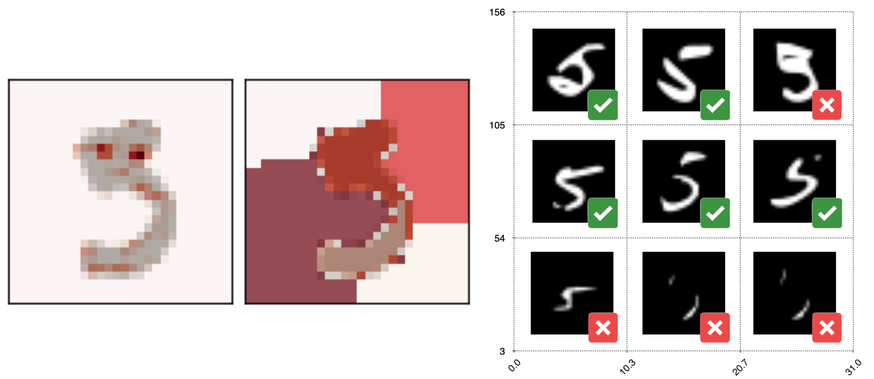


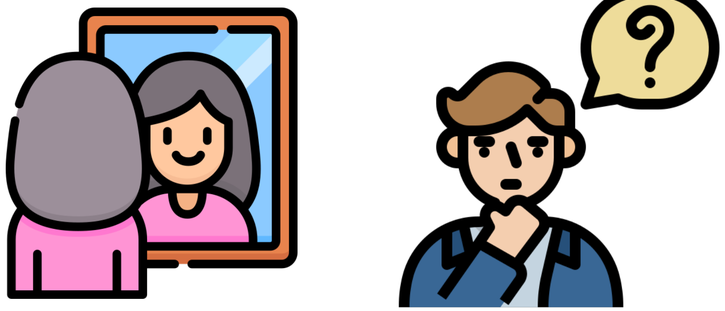


Icons from www.flaticon.com

SUMMARY

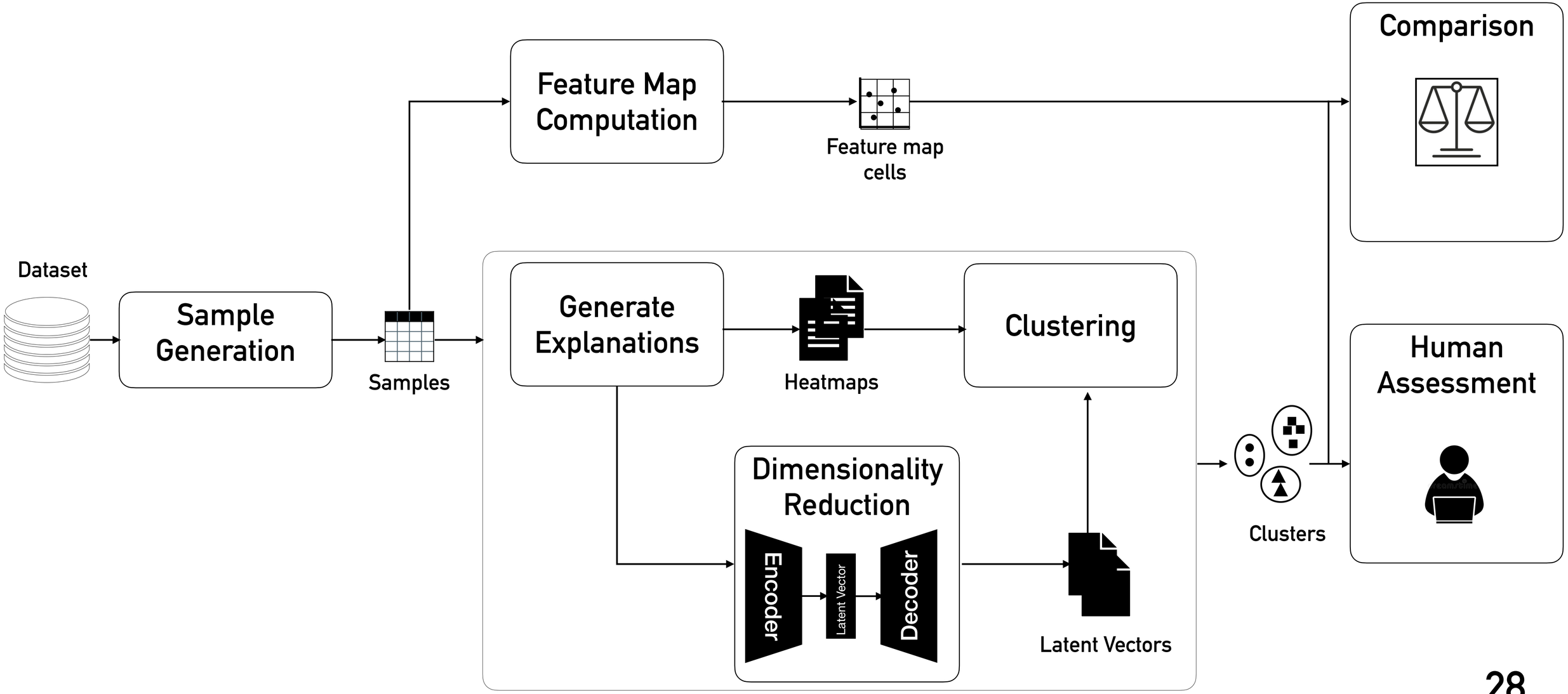






EXTRA SLIDES

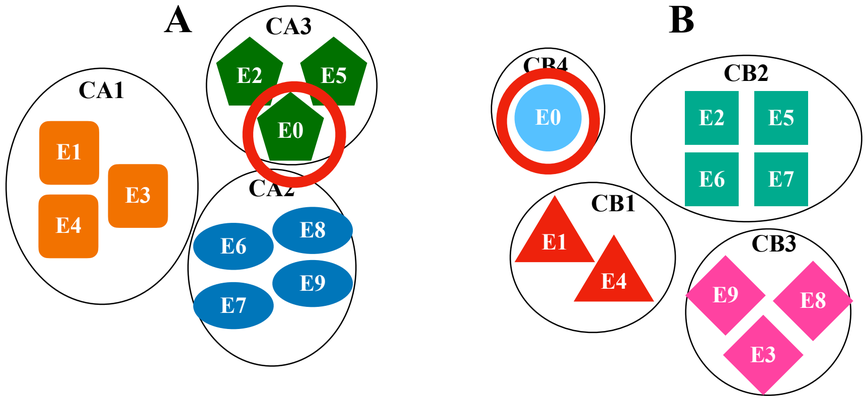
EVALUATION PIPELINE

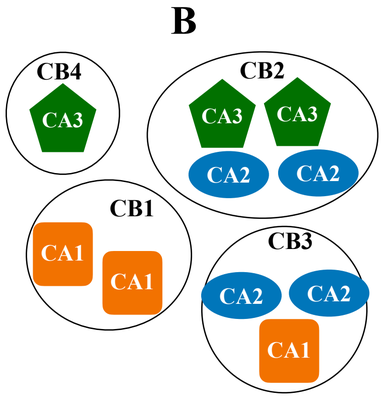


GINI SIMILARITY

$$GI(D,\ A) = 1 - \sum_{i=1}^{|A|} p_{Ai}^2$$

$$GS_{(B,A)} = 1 - \frac{1}{|B|} \sum_{i=1}^{|B|} GI(CB_i, A)$$





$$GS_{(B,A)} = 1 - \frac{1}{4}(GI(CB1,A) + GI(CB2,A) + GI(CB3,A) + GI(CB4,A))$$

$$= 1 - \frac{1}{4}(0 + \frac{1}{2} + \frac{4}{9} + 0) = 0.76$$

RQ1: SIMILARITY

MNIST IMDB

High Level	Low Level	Input space	Sim	Sim
3D	IG	Original	**0.70**	**0.74**
		Latent	0.55	0.68
	LIME	Original	0.55	**0.81**
		Latent	0.53	0.66
2D	IG	Original	**0.76**	**0.76**
		Latent	0.49	0.56
	LIME	Original	**0.62**	**0.80**
		Latent	0.47	0.59
1D	IG	Original	**0.85**	**0.83**
		Latent	0.59	0.66
	LIME	Original	**0.75**	**0.85**
		Latent	0.59	0.68

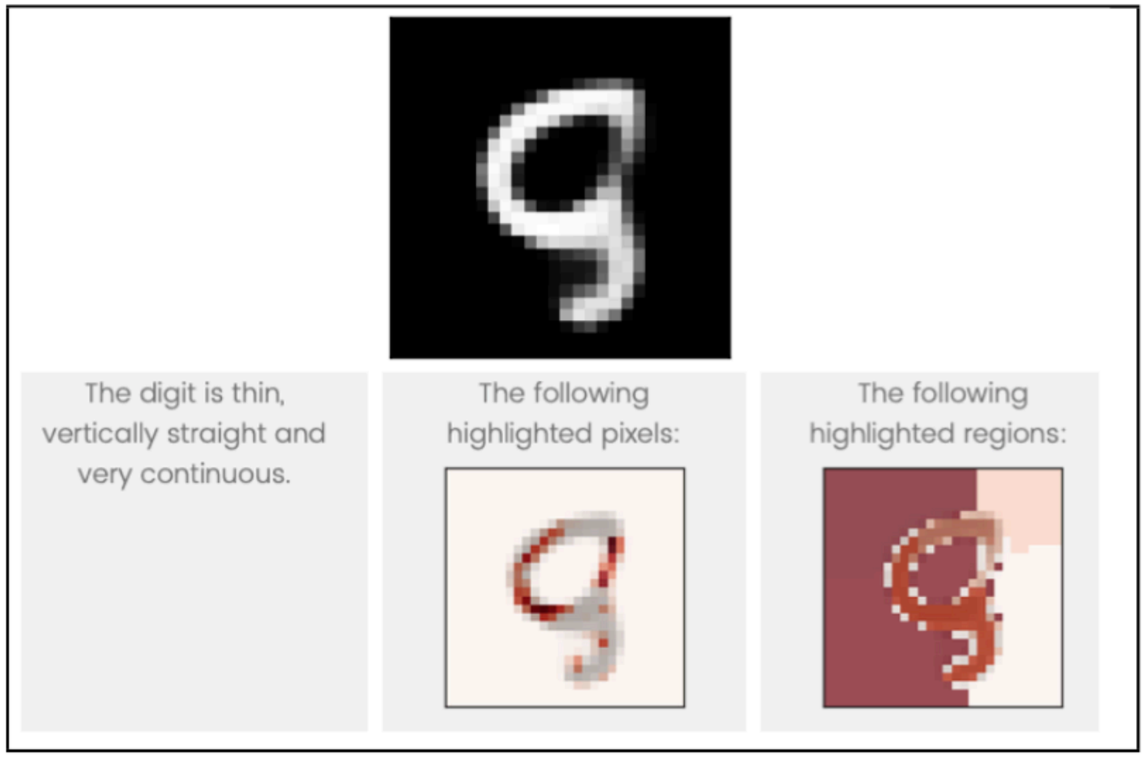
Original space always achieves better similarity

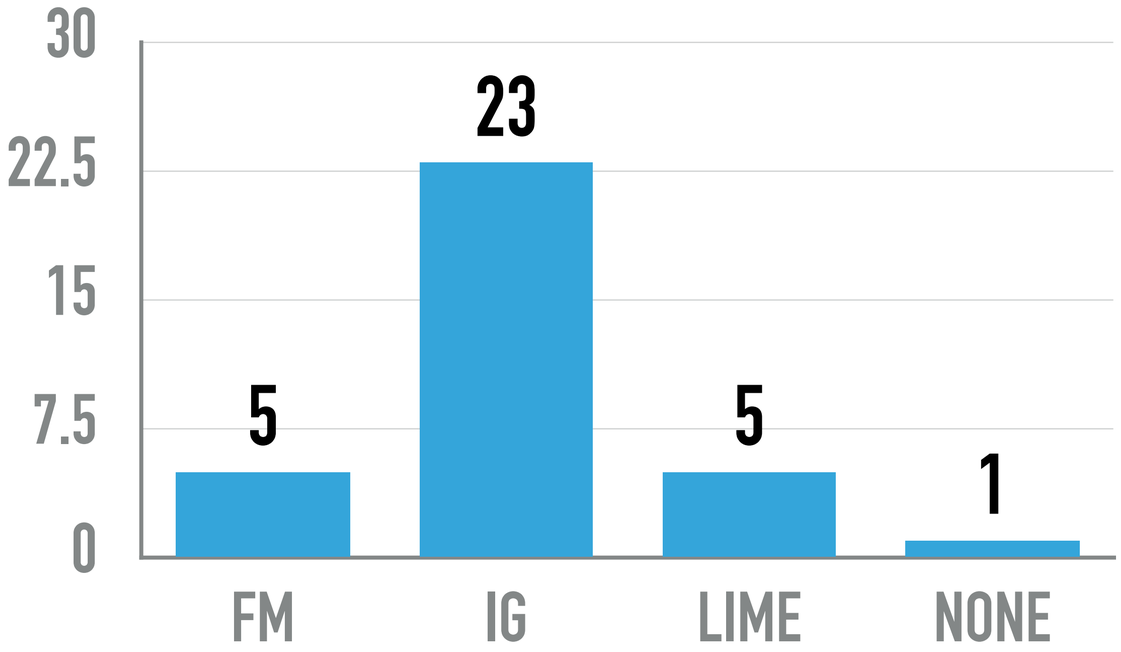
RQ2: UNDERSTANDABILITY

TABLE V: RQ2 - Number of Matches with Human Explanations (MH); 'None' indicates the number of cases when no match was found.

	MH							
	MNIST				IMDB			
Q#	FM 3D	IG	LIME	None	FM 3D	IG	LIME	None
Q1	12	2	10	8	2	13	3	3
Q2	5	23	5	1	5	3	3	9
Q3	4	7	9	12	11	17	8	0
Q4	6	7	5	14	6	15	3	2
Q5	3	15	2	11	10	14	7	1
Q6	7	6	4	14	12	15	8	0
Q7	7	11	7	10	0	14	5	3
Q8	9	5	8	13	11	14	12	1
Q9	5	1	9	17	6	8	4	4
Q10	13	10	5	8	11	12	7	2
Sum	71	87	64	108	74	125	60	25

DISCUSSION


The digit is thin, vertically straight and very continuous.
The following highlighted pixels:
The following highlighted regions:

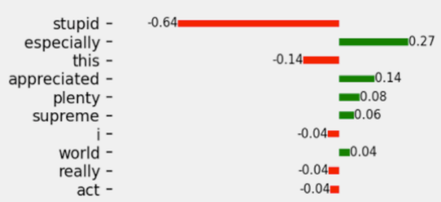

30
22.5
15
7.5
0
5
23
5
1
FM
IG
LIME
NONE

DISCUSSION

If folks were really this stupid I could be the SRW - Supreme Ruler of the World. In this one Knotts plays a dimwitted bean counter for some little jerk water town run by a group of crooked simpletons only slightly brighter than he is. When things appear a bit shaky for the crooks they go for a frame-up of the patsy Figg. Plenty of laughs as Knotts does his usual bumbling, stumbling act. I especially appreciated the extension cord scene; asininity at it's highest level.

The review contains 3 positive words, 5 negative words and 11 verbs (the number of verbs is an indicator of the text complexity).

The review contains the following words contributing to negative (red) and positive (green) sentiments:



The review contains the following words contributing to negative (red) and positive (green) sentiments: