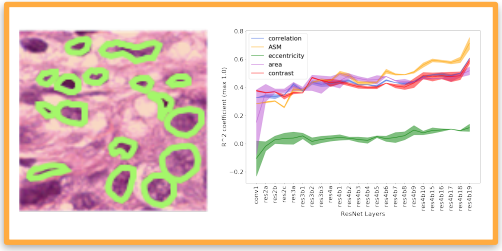
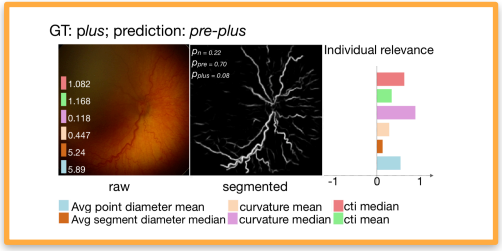
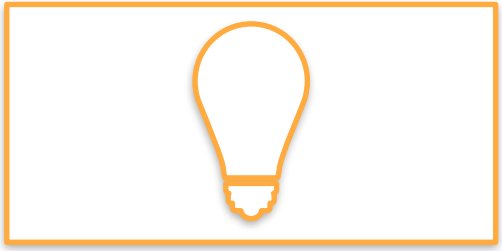
Regression Concept Vectors for explanations of Deep Learning

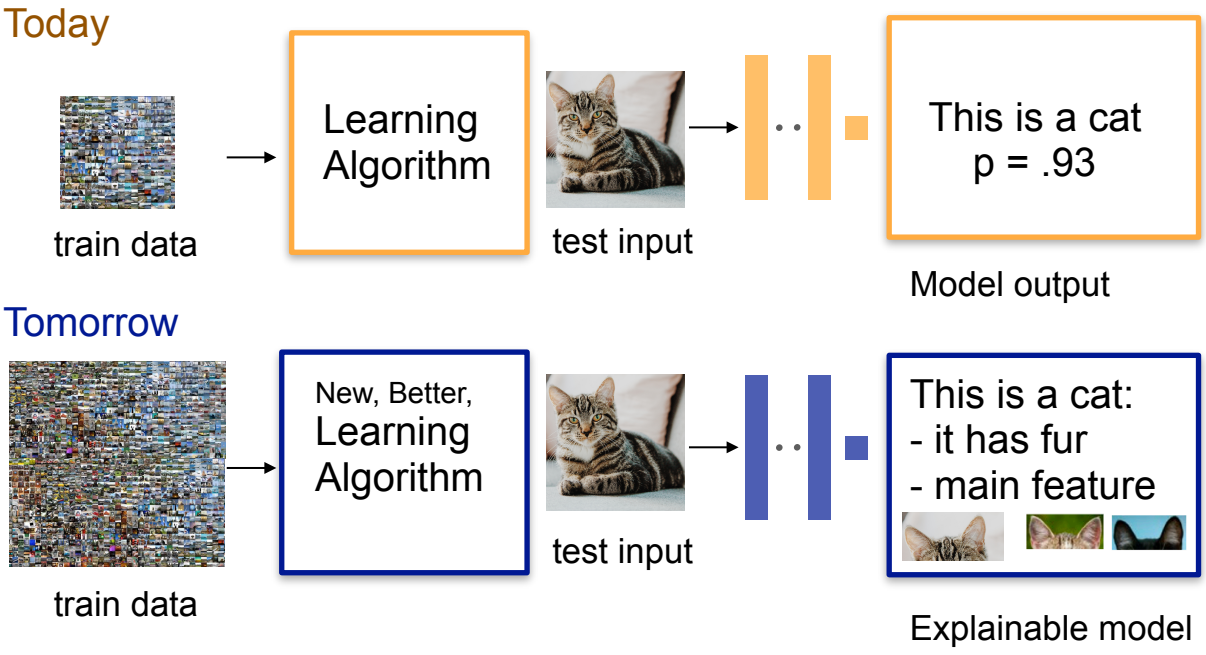
Mara Graziani, SMLD 2018







Why model interpretability?



Why this answer?
Why not something else?
When do you succeed?
When do you fail?
Can we trust you?
How to fix errors?

I see why
I see why not
I know when you will succeed
I know when you will fail
I know when to trust you
I know how to fix mistakes

[DARPA, explainable AI]

Concept Learning

Inferring a Boolean-valued function from training examples

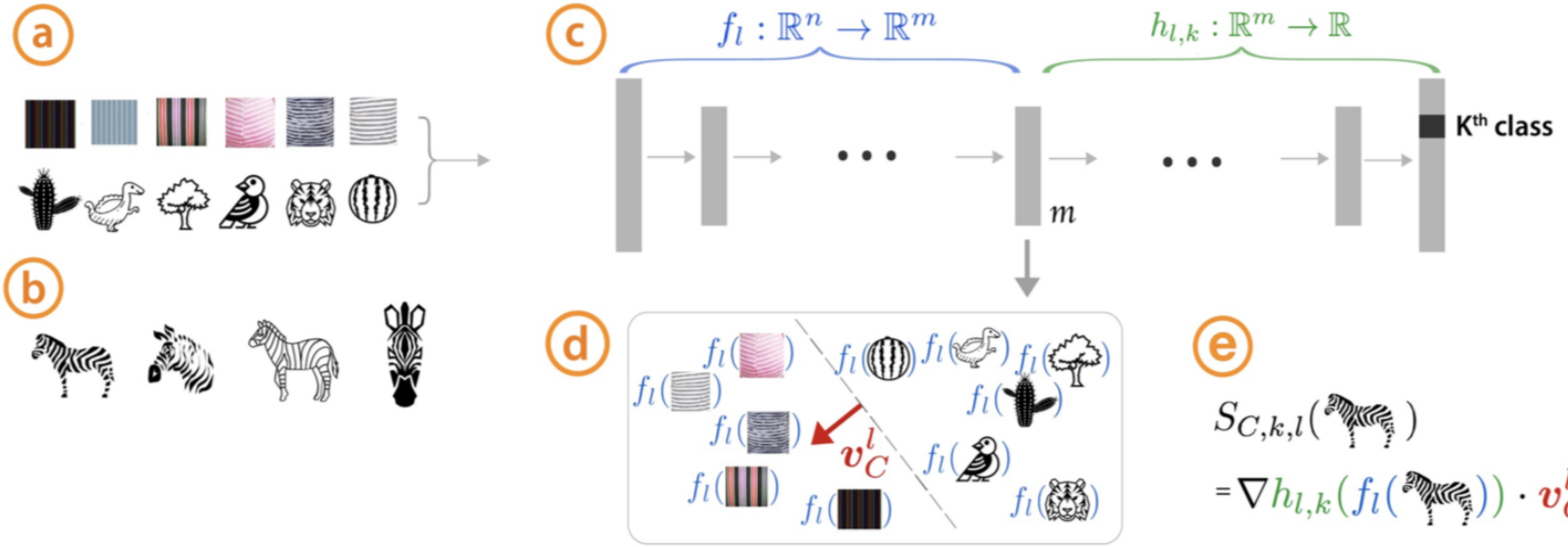
Concept Learning

Inferring a Boolean-valued function from training examples

Concept Learning

Inferring a Boolean-valued function from training examples

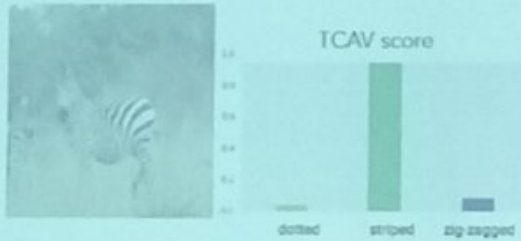
TCAV: influence of a concept to network classification



Testing with Concept Activation Vectors [Kim B. et al., 2017] ICML 2018

TCAV is generalization of saliency maps for concepts

TCAV

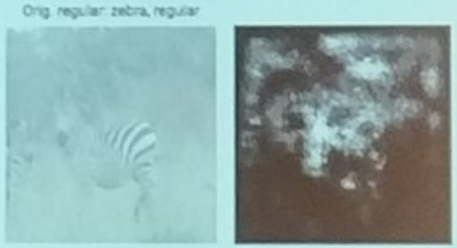


zebra-ness → $\frac{\partial p(z)}{\partial v_C^l}$

striped CAV →

Directional derivative

Saliency Maps

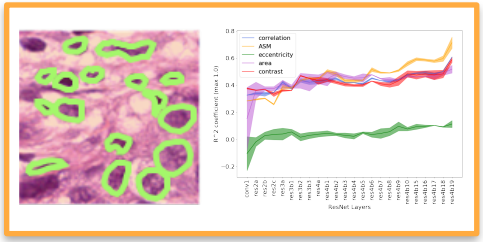


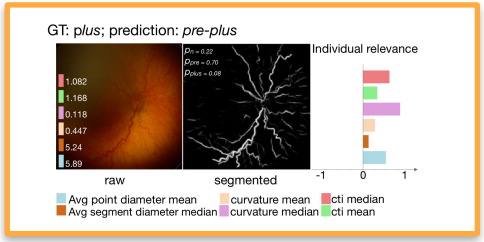
zebra-ness* → $\frac{\partial p(z)}{\partial x_{i,j}}$

One pixel at a time →

What if the concepts are **continuous** measures?

Number of windows,
Window size,
Door height …



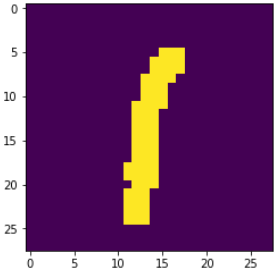


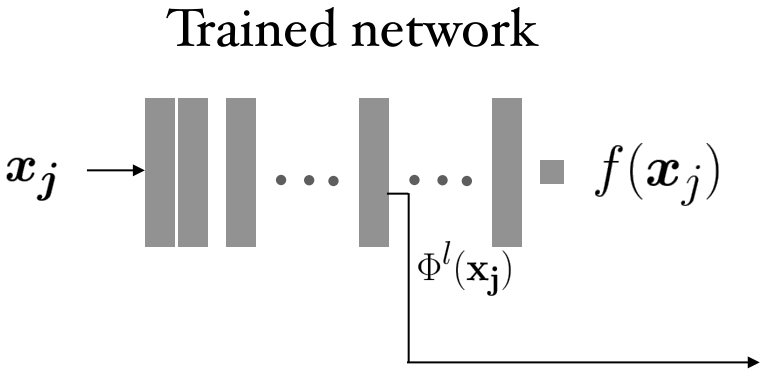
Morphological features

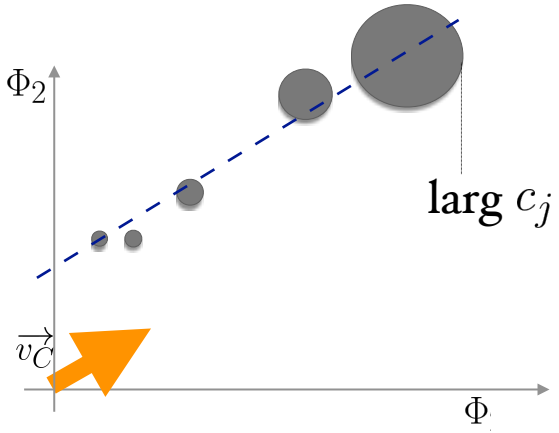
Curvature and tortuosity

RCV: Regression Concept Vectors

Can we find some measures in the data, whose increase is relevant for the classification?

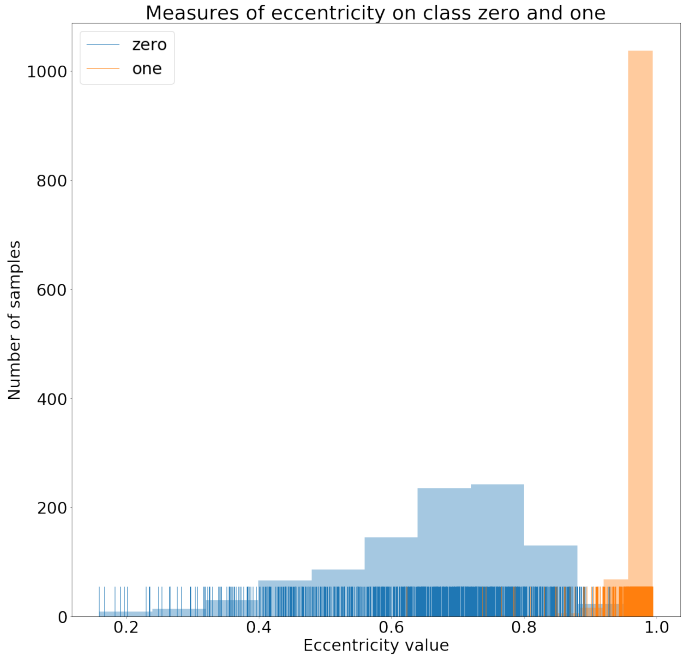


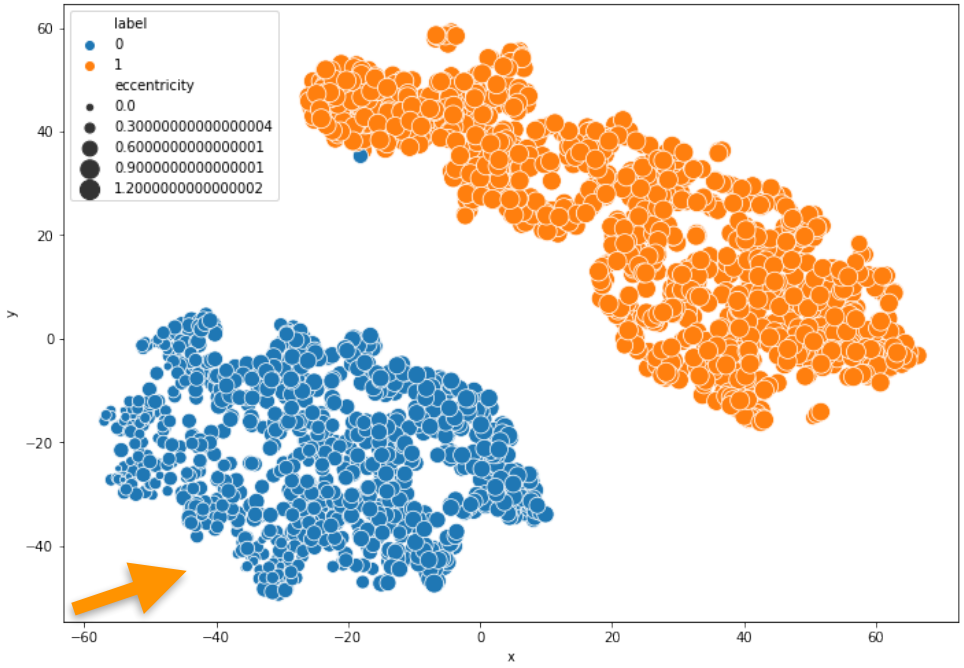




- Pixel counts
- Shape
- Orientation
- Eccentricity…

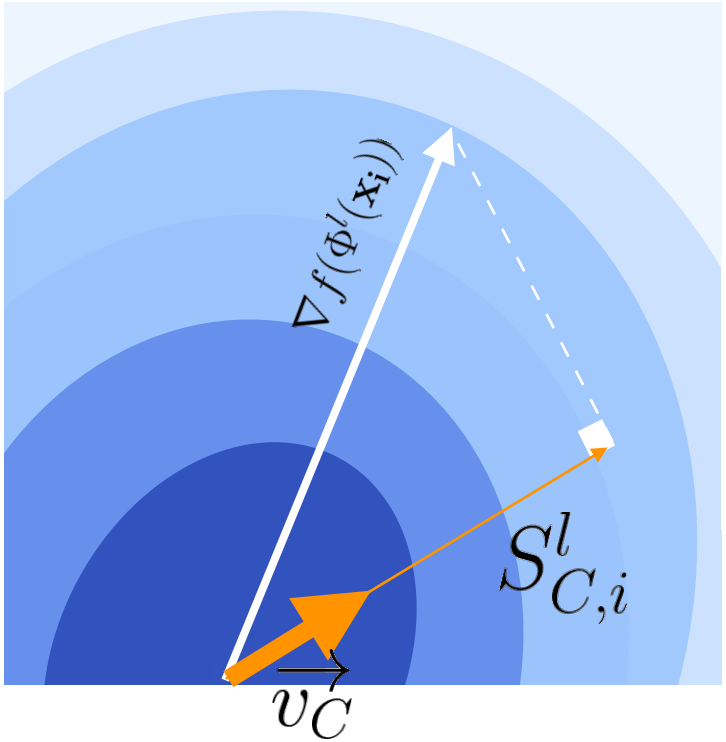
Simple task: is it a 0 or a 1?





We solve **linear regression** in the activation space

Determine the relevance of each concept measure



$S^l_{C,i}$: sensitivity for each testing sample

- Br derive **global explanations**

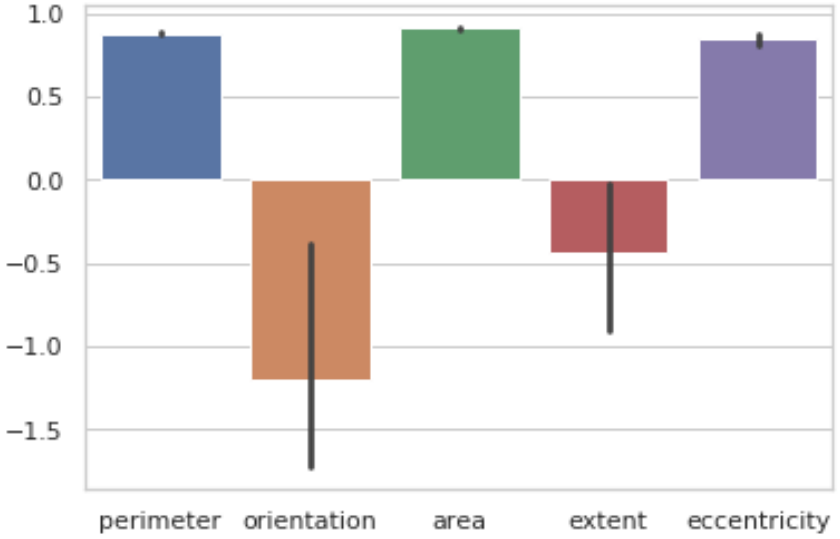
$$Br = R^2 \times \left(\frac{\hat{\mu}}{\hat{\sigma}}\right)$$

Mean of the $S^l_{C,i}$

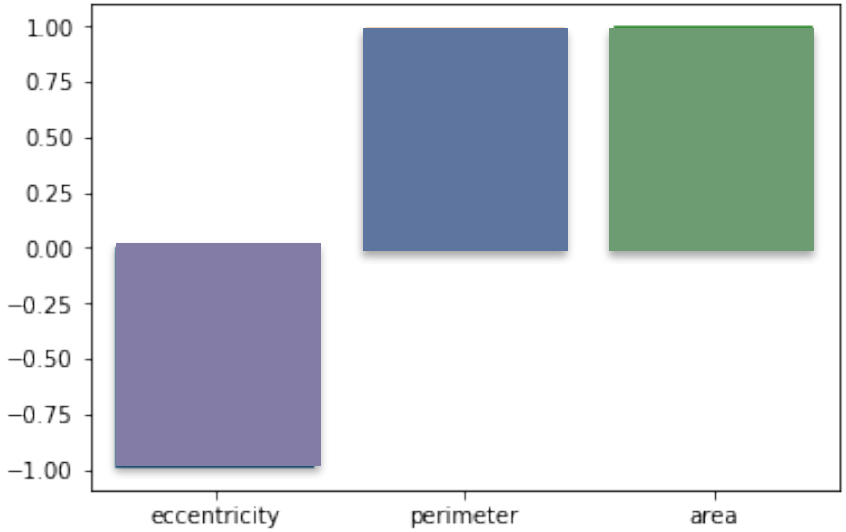
Regression determination coefficient

Standard deviation

Simple task: is it a 0 or a 1?



Regression determination coefficient

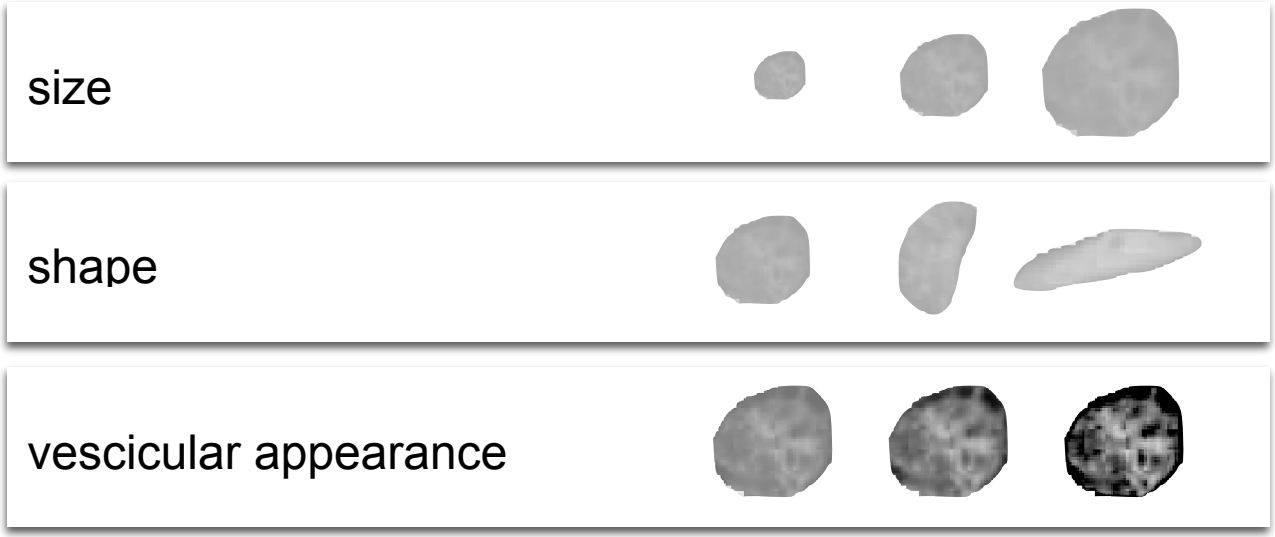


Bidirectional relevance scores

Application to Breast Cancer Histopathology

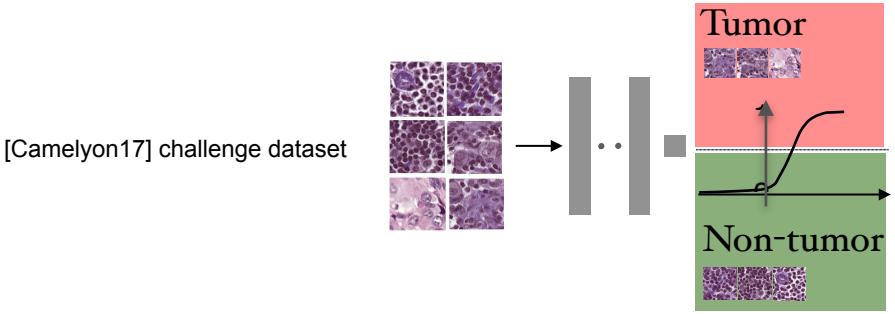
[Graziani M.,et al. iMIMIC at MICCAI 2018]

Concept: nuclei pleomorphism *



*__NGH:__ Nottingham Grading System for breast cancer histopathology [Elston, C.W., 1991]

Three objectives:

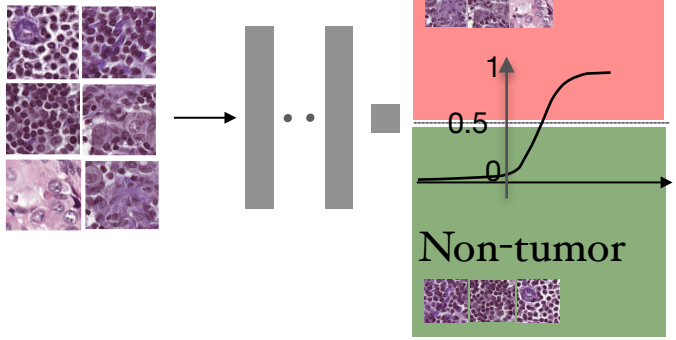


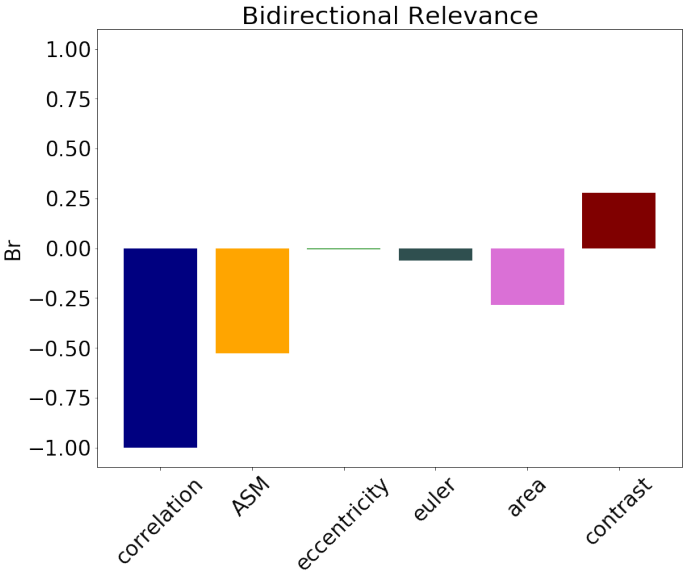
Binary ResNet patch-classifier
(92 % val accuracy)

- Extract concept measures c_j
- Learn the Regression Concept Vectors (RCVs)
- Determine the bidirectional relevance (*Br*) of each concept measure

Results

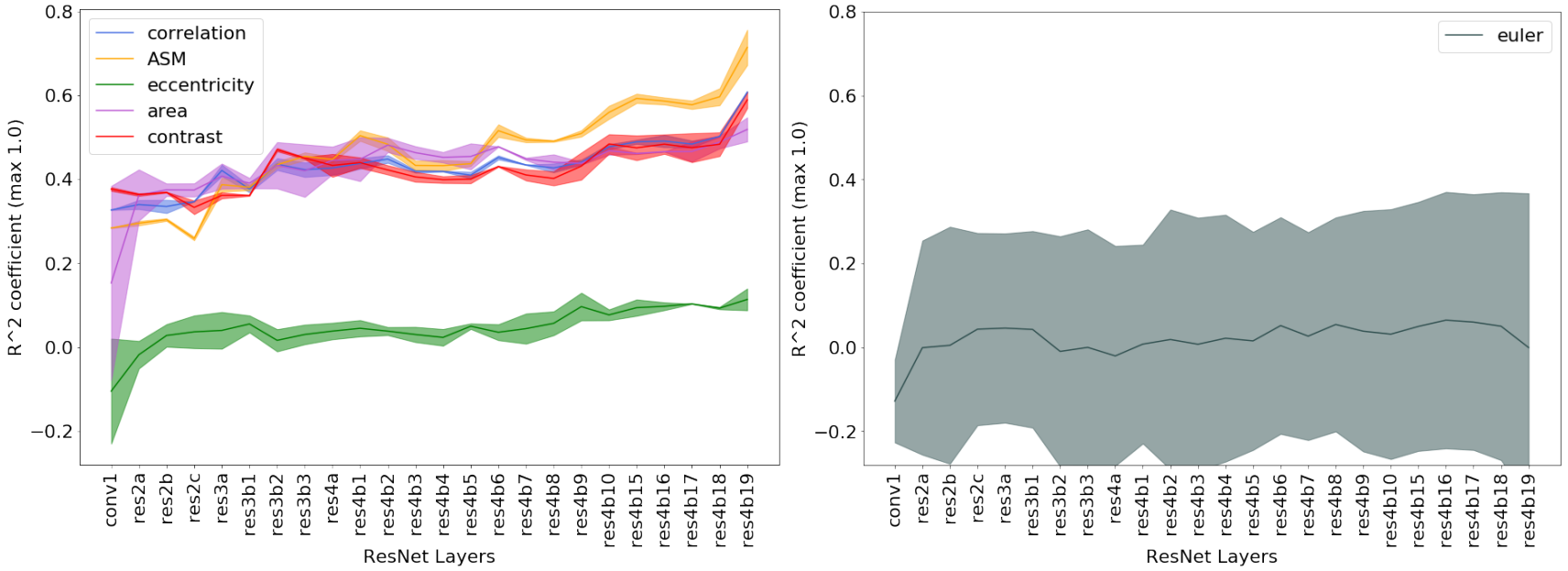
- Binary ResNet patch-classifier (92 % val accuracy)
- Are there clinical factors relevant to classification?





Yes!

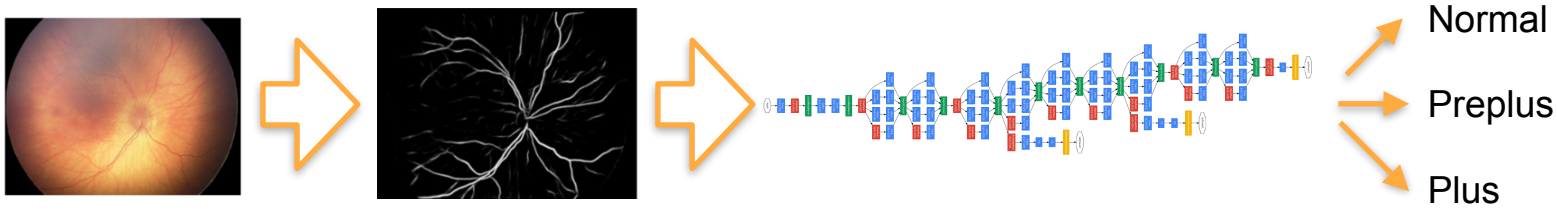
Correlation, ASM and **Contrast**

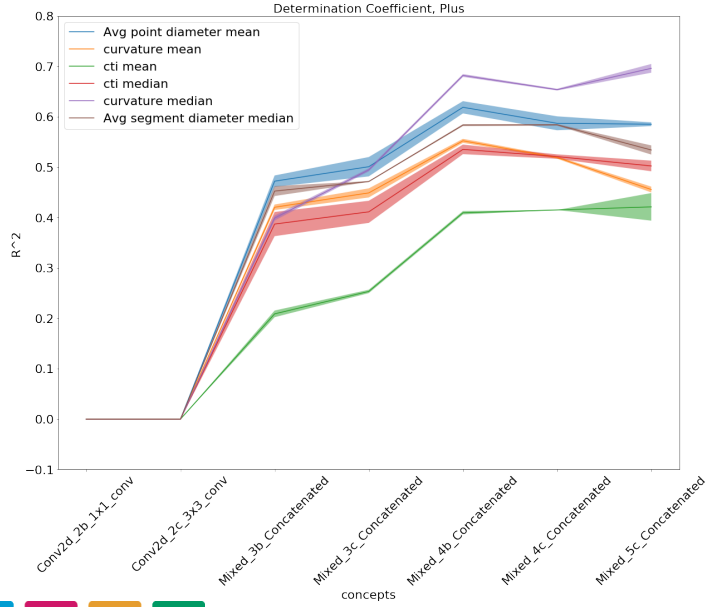


Linear regression determination coefficient at different layers in the network

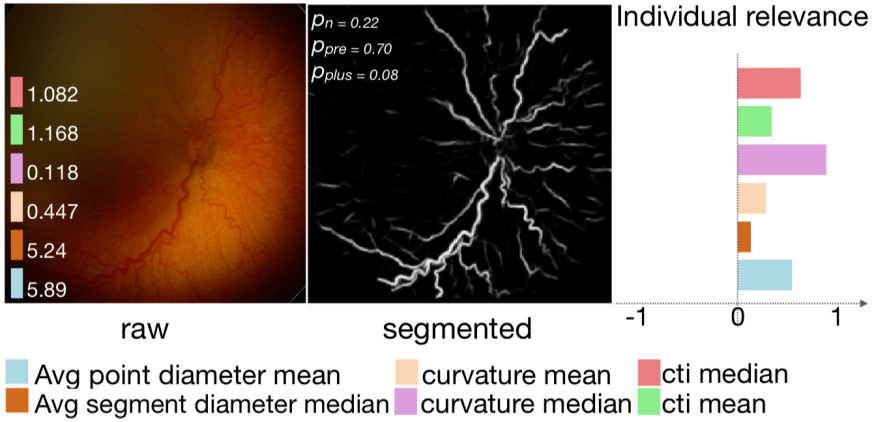
Application to Rethinopathy of Prematurity

Concepts: vessel curvature, tortuosity and dilation





GT: p*lus*; prediction: *pre-plus*



Conclusions

- ✚ Interpretations not at the pixel level
- ✚ Extendable to a variety of concepts and application domains
- ▬ Concept search space very large

Still many more points to address…

Thank you!

More information:

- mara.graziani@hevs.ch
- http://medgift.hevs.ch/wordpress/
- https://github.com/medgift/iMIMIC-RCVs