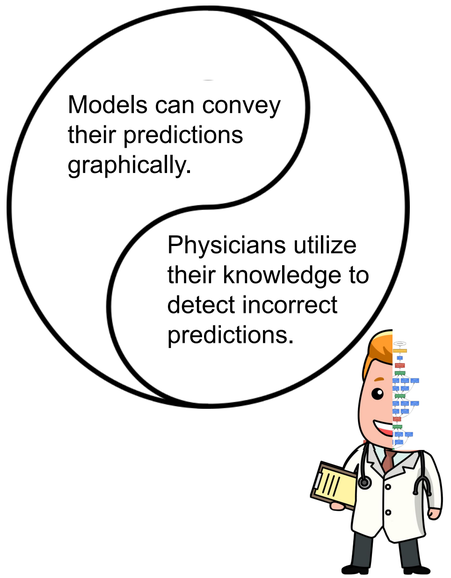
Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays

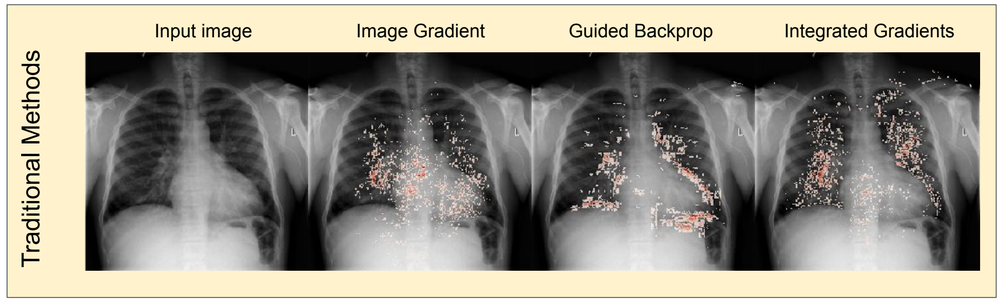
Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, Akshay Chaudhari

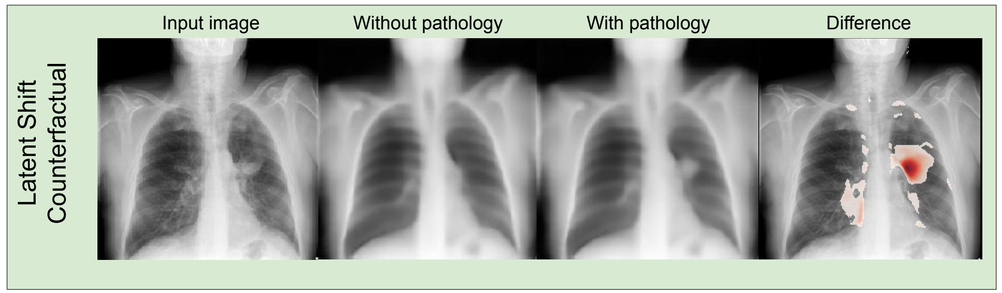
Stanford

Why Counterfactuals?

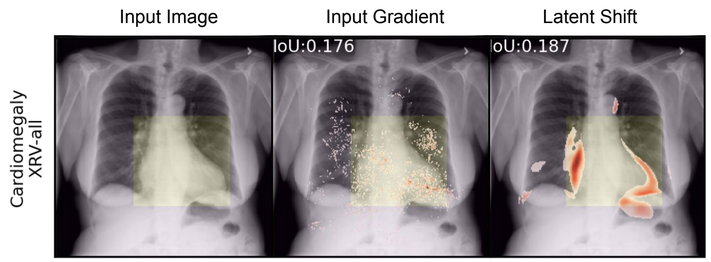
Moving towards the AI+physician symbiosis







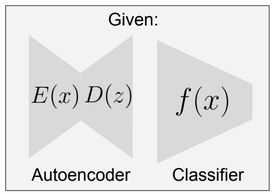
IoU Analysis Inconclusive



IoU is generally low, little variation between methods.
Likely annotations need to be specially created for counterfactuals.

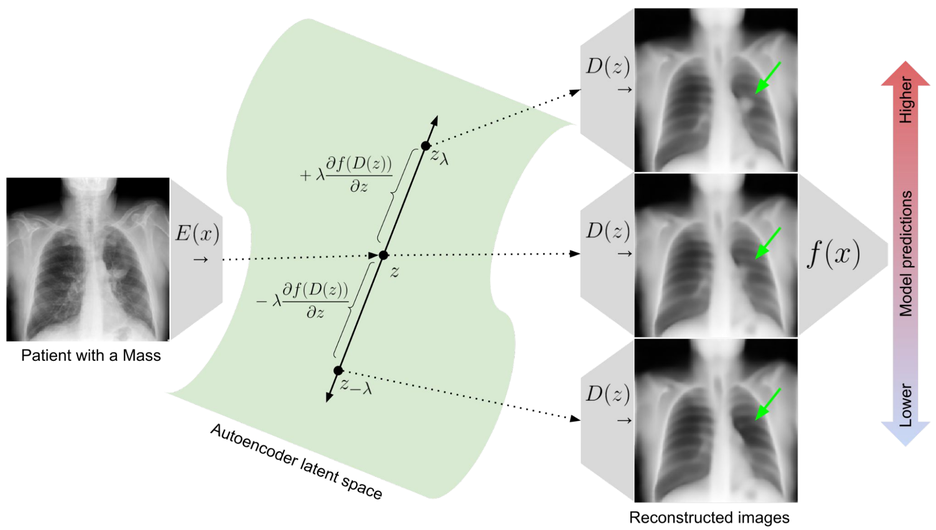
Task	Dataset	Model → 2D Method	XRV-all AUC	XRV-all IoU	XRV-mimic.ch AUC	XRV-mimic.ch IoU
Mass	NIH	grad	0.82	0.16±0.14	Model does not predict	
		guided		**0.19±0.16**		
		integrated		0.13±0.13		
		latentshift-max		0.14±0.17		
Lung Opacity	RSNA	grad	0.84	**0.21±0.11**	0.75	0.13±0.09
		guided		0.21±0.12		0.09±0.07
		integrated		0.17±0.10		0.08±0.07
		latentshift-max		0.20±0.13		**0.15±0.14**
Pneumothorax	SIIM-ACR	grad	0.78	0.01±0.02	0.67	0.01±0.02
		guided		**0.03±0.05**		0.02±0.03
		integrated		0.01±0.02		0.01±0.01
		latentshift-max		0.02±0.04		**0.03±0.07**

The Latent Shift Method

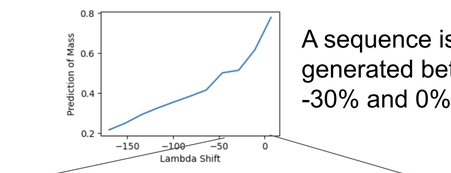


Latent Shift Method:

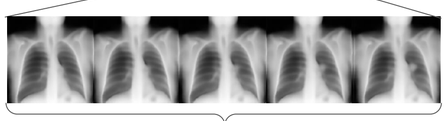
- Opposite of an adversarial attack.
- Perturb the input so the classifier reduces its prediction regularized by the decoder.
- Compute the gradient of the output of the classifier with respect to the latent space.



2D Image Construction



A sequence is generated between -30% and 0%



The maximum pixel wise difference between all images results in a 2D image

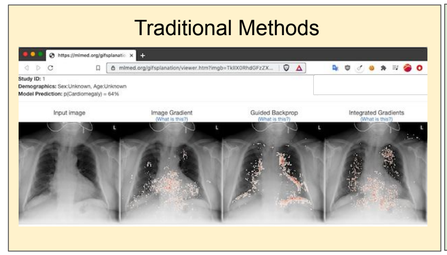
Related work:
[Singla, Explanation by Progressive Exaggeration, 2020]
[Schutte, Using StyleGAN for Visual Interpretability of Deep Learning Models, 2020]
[Joshi, xGEMs: Generating Examplars to Explain Black-Box Models, 2019]

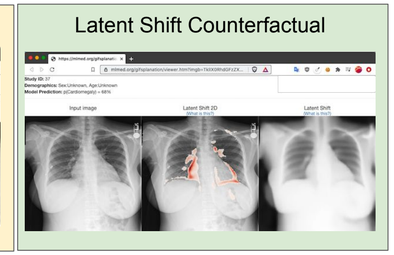
Positive Reader Study Results

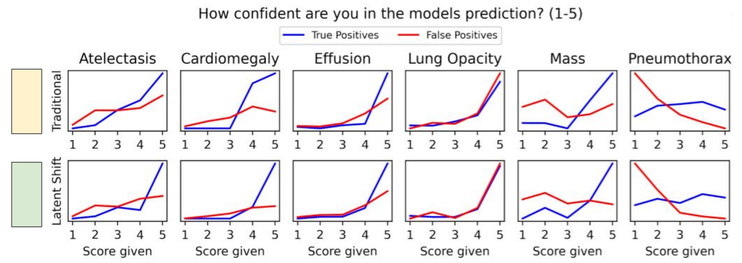
Reader study: Two radiologists evaluated how confident they were in a models predictions.

240 Chest X-ray images
50% are false positives

Radiologists asked:
"How confident are you in the model's prediction? (1-5)"







True Positives: 0.15±0.95 confidence increase using Latent Shift (p=0.01)
False Positives: 0.04±1.06 increase which is not significant (p=0.57)