TRUST-LAPSE

When can you trust your model's predictions? A Mistrust Scoring Framework for inference

Nandita Bhaskhar, EE PhD Candidate, Stanford University

Email: nanbhas@stanford.edu
Website: www.stanford.edu/~nanbhas
Paper: https://arxiv.org/abs/2207.11290

Paper

Shout outs to

Daniel Rubin

Christopher Lee-Messer

Co-authors

Rubin Lab

Juan Manuel Zambrano Chaves
Khaled Saab
Siyi Tang
Liangqiong Qu
Florian Dubost

Lee-Messer Lab

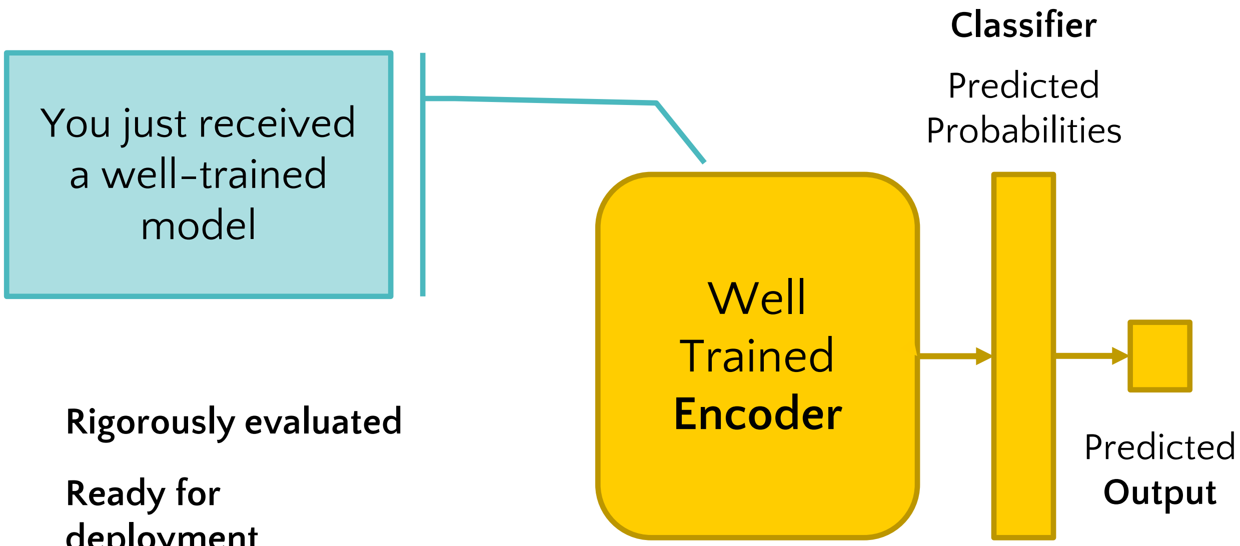
Neurotranslate team

Chaudhari Lab

Akshay Chaudhari
Phil Adamson
Louis Blankemeier
Arjun Desai
Anthony Gatti
Beliz Gunel
Anoosha Pai
Peyman Shokrollahi
Dave van Veen
Rogier van der Sluijs
Ben Viggiano
Jackson Wang

Acknowledgments

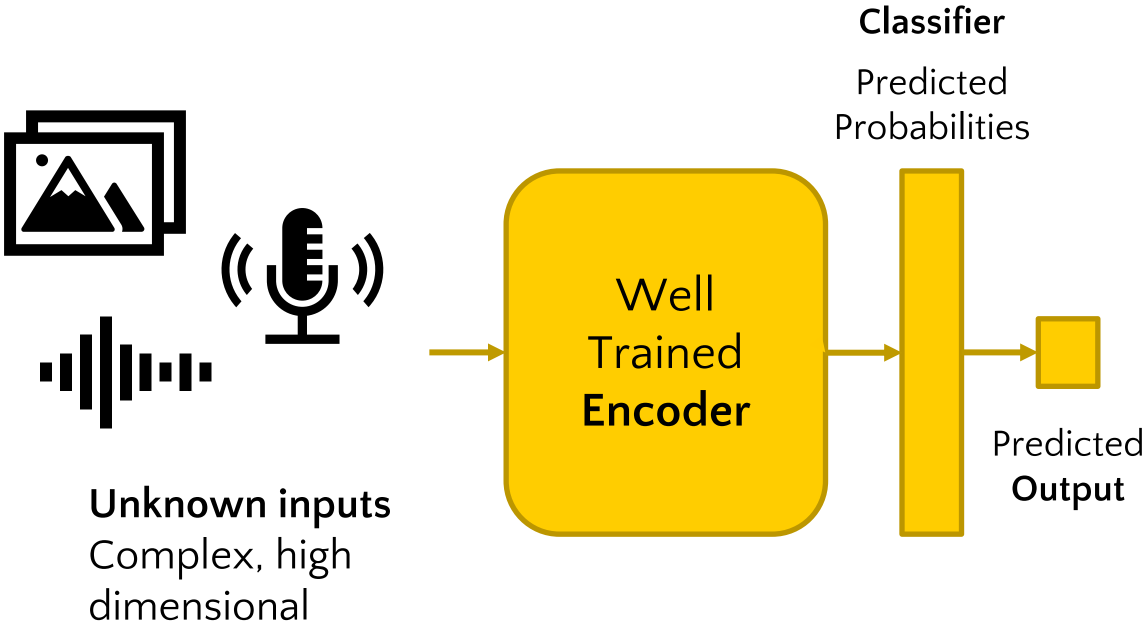
When a model is to be deployed: inference mode



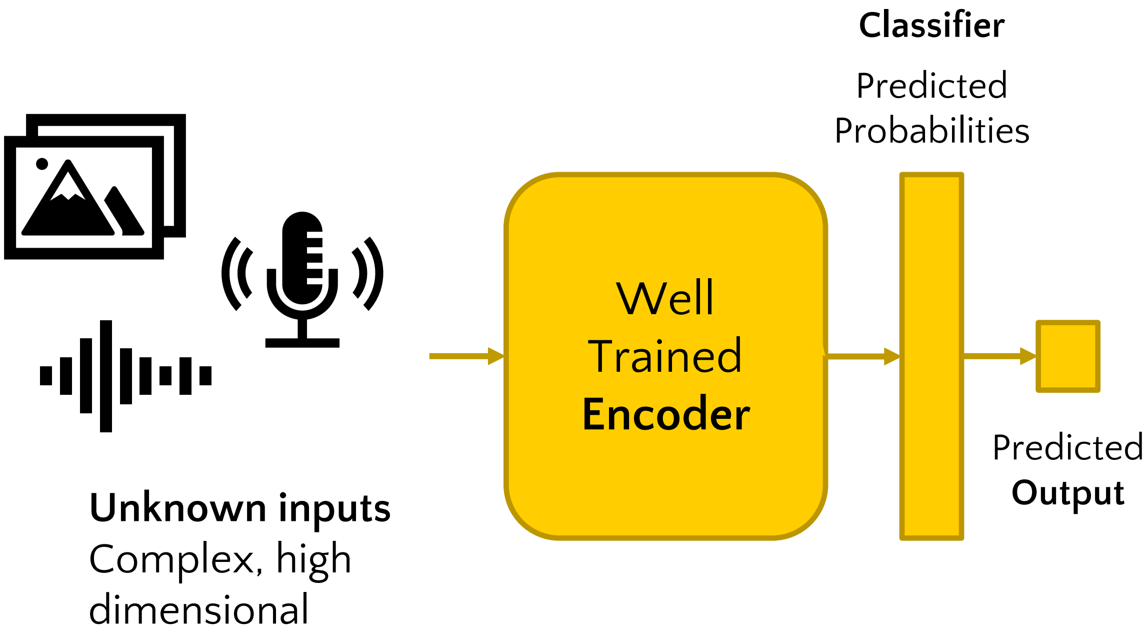
Rigorously evaluated

Ready for deployment

When a model is to be deployed: inference mode



When a model is to be deployed: inference mode



Deep learning models are **GREAT**

but …

They can make **ridiculous** mistakes!

Confidently & Silently, without warning!

When a model is to be deployed: inference mode

Unknown inputs
Complex, high dimensional

Well Trained **Encoder**

Classifier

Predicted Probabilities

Predicted **Output**

They can make **ridiculous** mistakes!

Confidently & Silently, without warning!

NOT

When should we trust this classifier's predictions?

Continuous Model Monitoring

How do humans do it?

We are surprisingly good at knowing when we don't know!

Past learnt knowledge, lived experiences, human intuition, our "Spidey" sense

How do humans do it?

We are surprisingly good at knowing when we don't know!

Past learnt knowledge, lived experiences, human intuition, our "Spidey" sense

For example:
When we encounter a language shift

For example:
Doctors do this all the time! "Something is **weird** in the EEG signal", "I **don't feel comfortable** with this MRI", …

TRUST-LAPSE: Our mistrust scoring framework

Continuous Model Monitoring

Desiderata for Continuous Model Monitoring

Notion of Trust

Complex & nuanced

Has many flavors

POST-HOC

use only the trained, deployed model

P

ACTIONABLE

allow automated, concrete action: accept / reject / flag

A

EXPLAINABLE

to some degree, explain why trust / mistrust

E

HIGH PERFORMING

low false positive rates and false negative rates

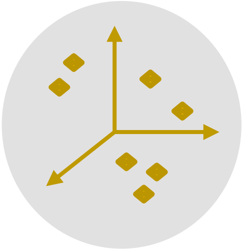
H

Key Insights

Well
Trained
Encoder

encodes the "world" as a **hierarchical, geometric, latent space**

d-dimensional
latent space

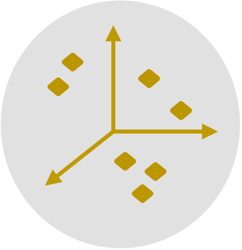


Key Insights

Well Trained DL Encoder

encodes the "world" as a **hierarchical, geometric, latent space**

d-dimensional latent space



Metrics in the latent-space
- how "similar" are two inputs
- how "near" or "far" are two inputs

Track these over time as a sequence!

Key insights

- Different **metrics** capture different aspects of the latent-space embeddings
- **Combining** them has value (as we'll show)
- Continuous model monitoring by **tracking these over time as a sequence**

Key insights

- Different **metrics** capture different aspects of the latent-space embeddings
- **Combining** them has value (as we'll show)
- Continuous model monitoring by **tracking these over time as a sequence**

Results Sneak Peak
SOTA on vision, audio, challenging clinical EEG domains

Results Sneak Peak
Detects SEMANTIC shifts too, unlike other methods

Results Sneak Peak
Evaluate on Drift detection. Very high drift detection rates

TRUST-LAPSE

Coreset (sampled from trainset)

- **Project** complex, high dimensional inputs to the **Latent-Space**
- Compare latent-space embeddings with those of coreset using different **metrics** to get **Latent Space Score**
- Estimate correlations over SEQUENCES of these scores (set-based approach vs instance-based approach) to give **Sequential Mistrust Score**
- Final Decision: Trust / Mistrust

Latent Space Score

Coreset
(sampled from trainset)

$$\text{coreset} = \{h(x_i) \mid x_i \sim \mathcal{D}_{train}\}; \quad |\mathcal{D}_{train}| \geq |\text{coreset}|$$

Distance-based Metric

$$s_{\text{dist}}(x) = \min_{h(x_i) \in \text{coreset}} d(h(x_i), h(x))$$

Angle-based Similarity

$$s_{\text{sim}}(x) = \max_{h(x_i) \in \text{coreset}} \text{sim}(h(x_i), h(x))$$

Latent Space Score

$$s_{\text{LSS}}(x) = s_{\text{dist}}(x) \cdot s_{\text{sim}}(x)$$

More details in the poster! Come, talk to us ☺

Latent Space Score

Coreset
(sampled from trainset)

$$\text{coreset} = \{h(x_i) \mid x_i \sim \mathcal{D}_{train}\}; \quad |\mathcal{D}_{train}| \geq |\text{coreset}|$$

Distance-based Metric

$s_{\text{dist}}(x) =$ Mahalanobis Distance with class-wise separate covariance, no label smoothing

Angle-based Similarity

$s_{\text{sim}}(x) =$ Cosine Similarity $(x_i), h(x))$

Latent Space Score

$$s_{\text{LSS}}(x) = s_{\text{dist}}(x) \cdot s_{\text{sim}}(x)$$

More details in the poster! Come, talk to us ☺

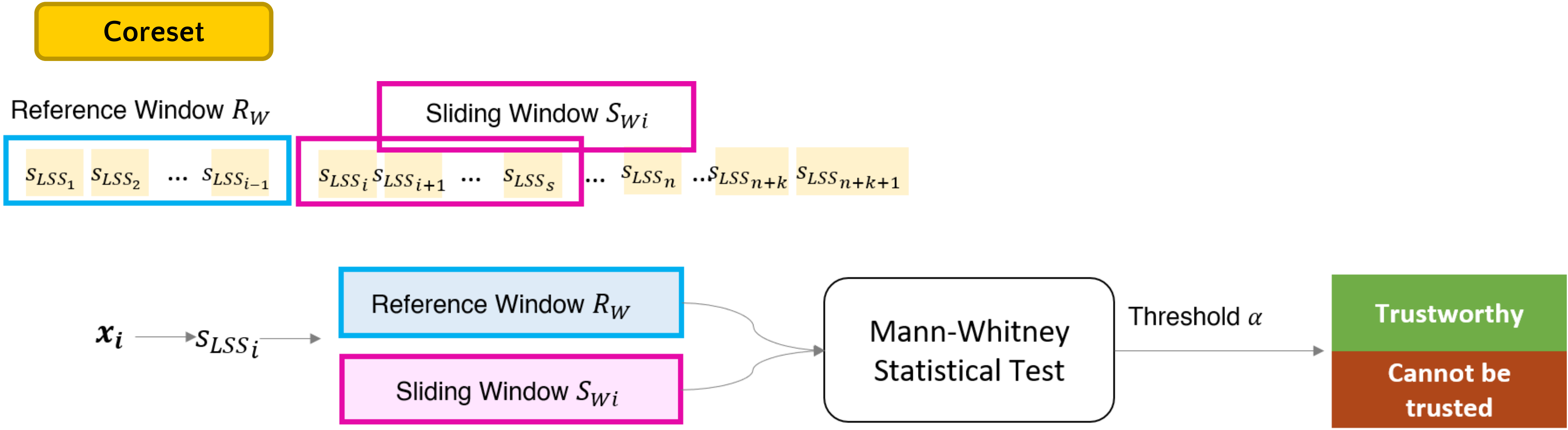
Sequential Mistrust Score

A. Make decisions using just LSS, e.g., via thresholding, etc.

B. Make decisions using sequence of LSS: **Sequential Mistrust Score**

Sequential Mistrust Score

More details in the poster! Come, talk to us ☺



More details in the poster! Come, talk to us ☺

Results: Distributionally Shifted Input Detection

SOTA

AUROC↑ / AUPR↑ / FPR80↓

	Audio	EEG Data	Vision
Task	Speech Classification	Seizure Detection	Image Classification
OOD Sets	Other spoken words	Other institutions	SVHN
MSP	0.626 / 0.527 / 0.515	0.358 / 0.421 / 0.754	0.760 / 0.770 / 0.358
Predictive Entropy	0.615 / 0.515 / 0.515	0.393 / 0.495 / 0.742	0.761 / 0.752 / 0.357
KL_U	0.553 / 0.475 / 0.579	0.390 / 0.472 / 0.719	0.775 / 0.786 / 0.347
ODIN	0.466 / 0.448 / 0.712	0.325 / 0.388 / 0.790	0.748 / 0.776 / 0.402
Vanilla Mahalanobis	0.680 / 0.636 / 0.520	0.633 / 0.651 / 0.525	0.738 / 0.782 / 0.477
Test-Time Dropout	0.649 / 0.619 / 0.523	0.647 / 0.619 / 0.583	0.716 / 0.725 / 0.494
TRUST-LAPSE (ours)	0.739 / 0.704 / 0.439	0.771 / 0.701 / 0.335	0.814 / 0.827 / 0.311

More details in the poster! Come, talk to us ☺

Results Peek: Semantic Shifts

- **Interesting counterfactual experiment** with two spoken word datasets Google Speech Commands (GSC) and Free-Spoken Digits Dataset (FSDD)

ONLY **TRUST-LAPSE flagged GSC WORDS and trusted predictions on FSDD**

Well Trained **Encoder**

Training Data: Subset of GSC 0-9

GSC Spoken Digits 0-9

FSDD Spoken Digits 0-9

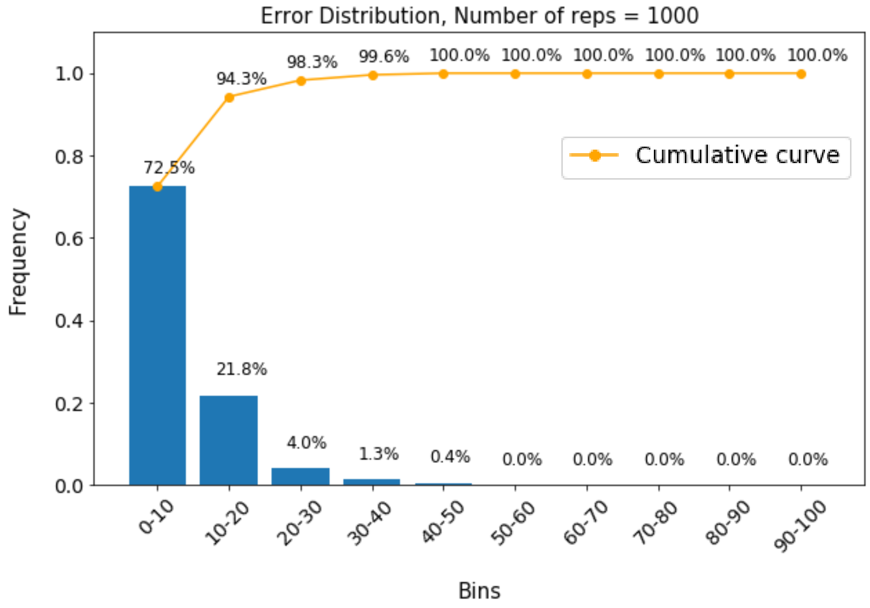
GSC Non-Digit words

Data presented during inference

More details in the poster! Come, talk to us ☺

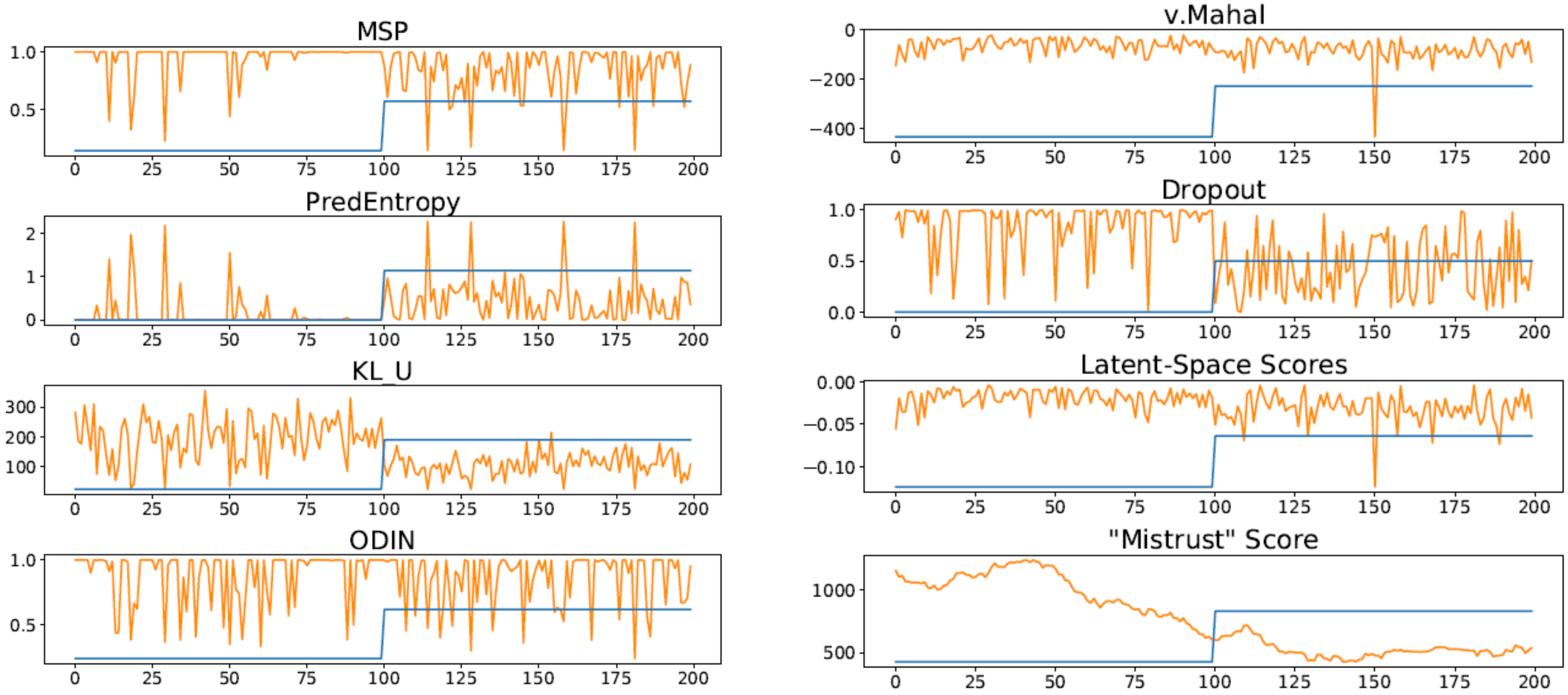
Results: Drift Detection

- **EEG:** Over **73%** of 1000 data streams (of length 10,000) have **less than 10% error** and over **93%** have **less than 20% error**
- **Audio:** over **85%** of the streams have less than **10% error** and over **97%** have **less than 20%** error
- **Vision:** the error distribution is even tighter – near **99% detection accuracy for over 95% of the streams**



Results: Drift Detection

More details in the poster! Come, talk to us ☺



Other Key Results Summary (Details in the poster session)

More details in the poster! Come, talk to us ☺

TRUST-LAPSE detects **SEMANTIC shifts too** on all domains (vision, audio, EEG) unlike other methods

TRUST-LAPSE detects lack of generalization in models

Ablations: TRUST-LAPSE depends on encoder capacity

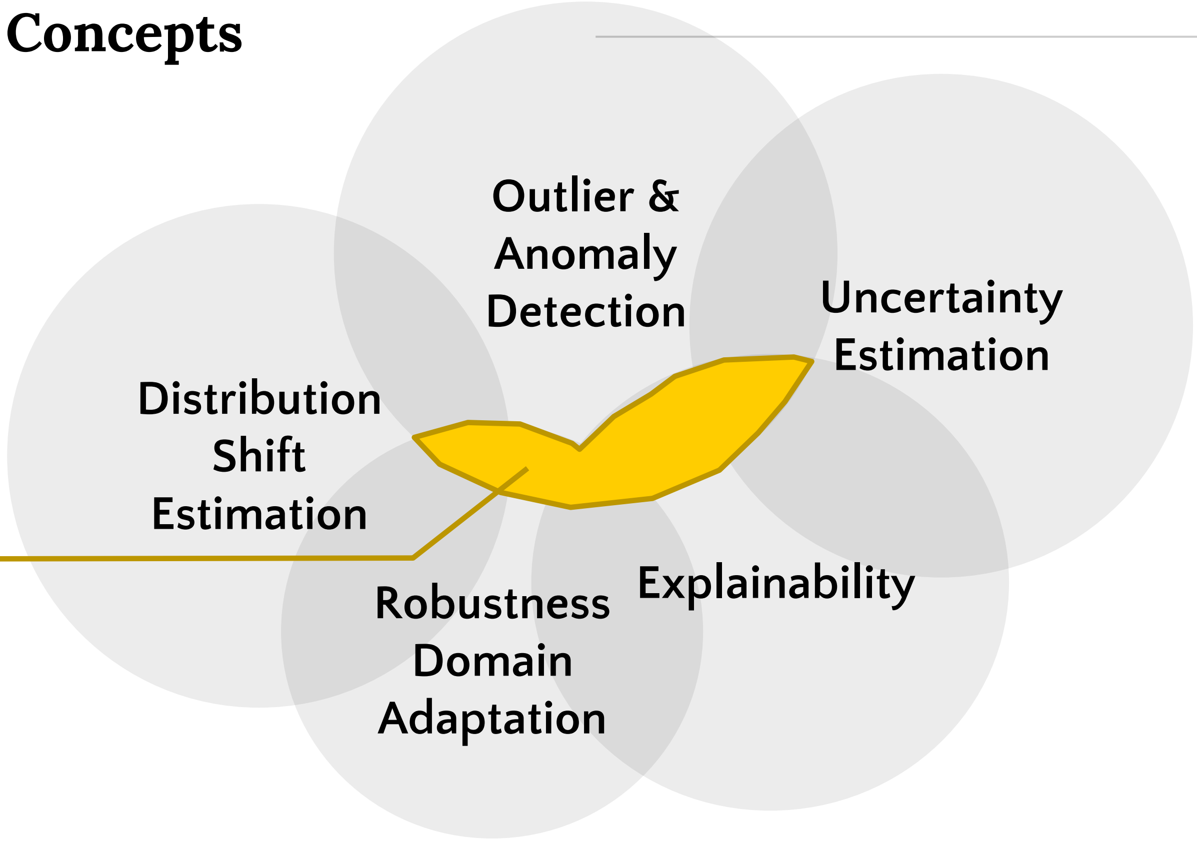
Ablations: Just 1-2% of trainset in the coreset is sufficient for TRUST-LAPSE

Related Concepts

Notion of Trust

Complex & nuanced

Has many flavors



Our work: TRUST-LAPSE

More details in the poster! Come, talk to us ☺

Conclusions

- As deep learning sees more success, there is a need for **continuous model monitoring** to enable **deployment**
- Essential in **safety-critical domains** like healthcare, self-driving, etc.
- **TRUST-LAPSE** is a simple yet powerful and flexible framework that we can use for any model and any task for monitoring a model in deployment
- Provides an opportunity for exploring: metrics, domains, data, etc
- Want to apply it for your models? Come chat with us ☺

More details in the poster! Come, talk to us ☺

Thanks!

Any *questions* ?

Paper: https://arxiv.org/abs/2207.11290

You can find me at

- Email: nanbhas@stanford.edu
- Twitter: @BhaskharNandita
- Website: www.stanford.edu/~nanbhas
- LinkedIn: https://www.linkedin.com/in/nanditabhaskhar/