UNIVERSITY of VIRGINIA

# Safety: Distribution Shift

Presented by

**Group 7:**
Dhriti Gampa(jwr9ew)
Jing-Ning Su (fzf9mg)
Shafat Shahnewaz(gsq2at)
Jackson Miskill (jcm4bsq)

# Presentation Outline

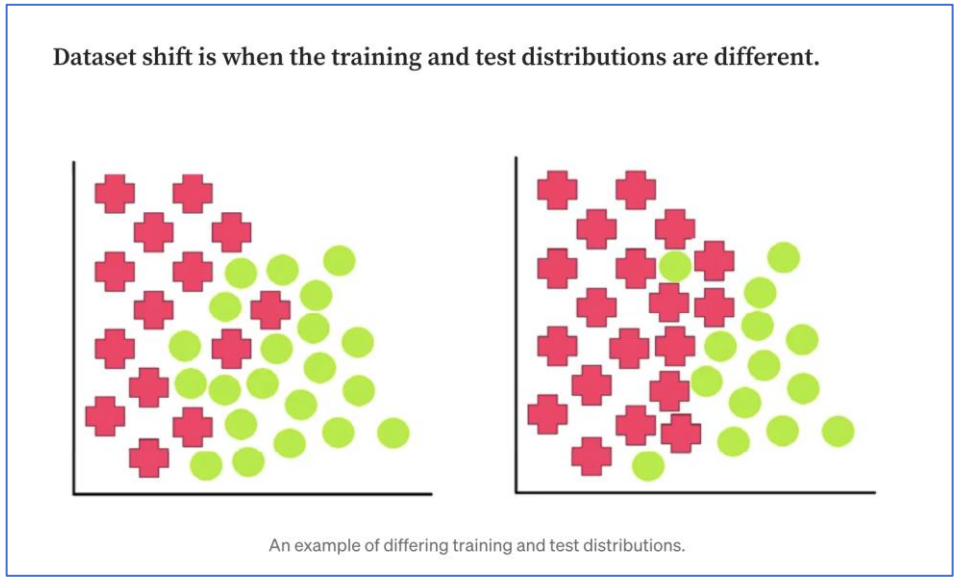

- Introduction – 10 min
  - Problem Statement
  - Motivation
  - Real world example
- ***Discussion 1 – 10 minutes***
- Detecting and quantifying shift (Dhriti) – 10 min
  - *Failing Loudly*
- ***Discussion 2 – 10 minutes***
- Solutions to this problem (Michelle and Shafat) – 15 min
  - Calibration – *Revisiting the Calibration of Neural Networks*
  - Selective Classification- *Deep Gamblers*
- Conclusion
  - Strengths & Weaknesses
  - Future Work
  - Contribution to Responsible AI
- ***Discussion 3 – 10 minutes***

# Presentation Outline


- Introduction – 10 min
  - Problem Statement
  - Motivation
  - Real world example
- ***Discussion 1 – 10 minutes***
- Detecting and quantifying shift (Dhriti) – 10 min
  - *Failing Loudly*
- ***Discussion 2 – 10 minutes***
- Solutions to this problem (Michelle and Shafat) – 15 min
  - Calibration – *Revisiting the Calibration of Neural Networks*
  - Selective Classification- *Deep Gamblers*
- Conclusion
  - Strengths & Weaknesses
  - Future Work
  - Contribution to Responsible AI
- ***Discussion 3 – 10 minutes***

# Problem Statement

- The problem: dataset shift
- Description [4,5]
  - At the most basic level, a difference between the training and test distributions
  - Models are not suited correctly based on their data, based on their targets, or based on a combination of the two
  - This happens due to poor training data, changes in data distributions with respect to time, or events, or anything really
  - It is difficult to monitor in real time (anecdotally)
  - Models can mis-classify
- Current state of the art ML/AI packages do not check rigorously for dataset shift




From [4]

# Types and causes of dataset shift

- 3 Types of dataset shift [4,5]
  - Covariate
    - Concerned with the data distribution (input)
    - Changes in the distribution of the independent variables (the features, or the covariates)
  - Prior probability
    - Concerned with the class distribution (output/target)
    - Changes in the output class (think about boundary shifts for classification problems)
  - Concept shift
    - Relationship between the two variables (both)

- There are lots of potential causes of dataset shift
- Examples:
  - Time
  - Location
  - Group of people
  - Seasonality
  - Market shifts
  - Behavioral shifts
  - etc

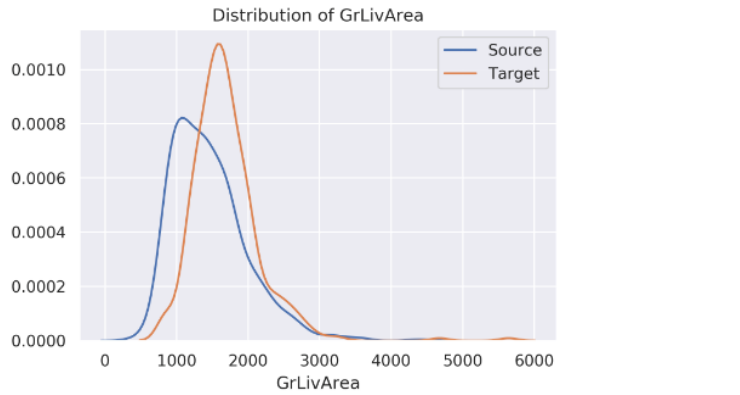



Figure 4: Distribution of source and target based on the GrLivArea attribute

From [7]

# Why is dataset shift a problem?

- Mis-classification can lead to unintended consequences
  - Decreased model accuracy (of course)
  - Bias/unfairness
  - Reduced trust in model
  - Marginalized groups
  - Economic Costs
    - Re-training
    - Litigation
    - Etc
  - Reputation
    - Especially for AI companies now
- It is difficult to detect and manage
- It exists in real world applications now

# A small real-world example

- Health insurance company calculating premiums and monthly costs for customers
- Machine learning classification model for predicting whether a person will develop a disease (and thus adding to premium)
  - Logistic Regression, for example
- Trained on people younger than 30 years old
  - Features
    - Activity level, frequency of junk food consumption, amount of drinking, amount of smoking, genetics, etc
  - But used on all ages of people
- Outcomes:
  - Probable classification errors for those older than 30, probably high error for those much older

# Motivation

- Safety: we want to know when are models are not fit to be presented to the real world
- In this presentation:
  - We answer two questions:
    - How do we detect and classify dataset shifts?
    - How do we solve these problems?


# Discussion 1

- What other causes of dataset shift can you think of (like time, location, etc)? Why would that cause dataset shift?
- In what real-world systems would dataset shift become a potentially huge problem?
    - What would those problematic outcomes be and why would dataset shift cause them?
- What industries might be affected by dataset shift the most? Why? What makes them more apt to be affected by dataset shift?
- What is an example of a system/industry of where dataset shift might not be a problem?

# Dhriti

# Presentation Outline


- Introduction – 10 min
  - Problem Statement
  - Motivation
  - Real world example
- ***Discussion 1 – 10 minutes***
- Detecting and quantifying shift (Dhriti) – 10 min
  - *Failing Loudly*
- ***Discussion 2 – 10 minutes***
- Solutions to this problem (Michelle and Shafat) – 15 min
  - Calibration – *Revisiting the Calibration of Neural Networks*
  - Selective Classification- *Deep Gamblers*
- Conclusion
  - Strengths & Weaknesses
  - Future Work
  - Contribution to Responsible AI
- ***Discussion 3 – 10 minutes***

# Detecting A Data Shift


- Dimensionality Reduction to make data more manageable representation for testing
- Input data is passed through pre-trained classifier to softmax outputs
- Statistical testing → shift detection threshold to determine if the distributions differ significantly
- Domain Classifier, Shift Characterization, Malignancy Assessment

# Types of Dimensionality Reductions

- No Reduction
  - preserves all original information → computationally expensive for high-dimensional data
- Principal Components Analysis
  - find an optimal orthogonal transformation to linearly uncorrelated components
  - captures maximum variance in fewer dimensions → may lose important nonlinear patterns
- Sparse Random Projection
  - uses a sparse random matrix to project data
  - computationally efficient and preserves distances between point → varying performance

# Types of Dimensionality Reductions

- Autoencoders
  - learn nonlinear encodings of data through neural networks
  - struggles with overfitting on training data
  - Trained vs Untrained
- Label Classifiers
  - utilizes supervised information, dependent on quality of classifier
  - struggles to capture shifts unrelated to class distinctions
- Domain Classifier
  - requires training on both source and target domains
  - struggles with subtle shifts

# Statistical Hypothesis Testing

- The DR Technique determines which statistical test they chose
- Multivariate
  - Maximum Mean Discrepancy (MMD) - kernel-based test.
- Multiple Univariate
  - Kolmogorov-Smirnov (KS) Test on each dimension + Bonferroni correction
- Categorical
  - Chi-Squared Test (for BBSD - hard predictions)
- Binomial Testing
  - For Domain Classifier Accuracy against random chance

# Shift Characterization and Malignancy

- Anomalous Samples: Domain classifier identifies samples most likely from the target domain
- Present these samples to the user to understand the nature of the shift
- Shift Malignancy
  - Assess target performance by true labels of the samples
  - Compare predictions on these samples to true labels
  - If accuracy is low, the shift is considered malignant

# Experiments

- Datasets:
  - MNIST and CIFAR-10 image datasets
- Simulated Shifts: Adversarial examples, class imbalance (knock-out), Gaussian noise, image transformations
- Evaluation Metric: Detection accuracy at a significance level of alpha = 0.05
- Varying the number of samples, Shift intensity, and percentage of data affected
- Compared performance of different DR + testing combinations

# Results

- BBSDs
  - performed best overall for shift detection (especially with univariate tests)
- Multiple Univariate Tests
  - comparable to multivariate tests despite the conservative Bonferroni correction
- Untrained Autoencoders
  - performed best for multivariate testing
- Domain Classifiers
  - improve with more samples, good for characterizing what the shift looks like
- The MNIST dataset had a shift between the train/test set

# Potential Biases & Ethical Considerations

- Dataset bias: Results may be specific to MNIST and CIFAR-10.
- Shift Type Bias: Performance depends on the types of shifts considered.
- Potential Misuse: Shift detection could be exploited by adversaries.
- Privacy: Characterizing shifts could reveal sensitive information about the target population.

# Discussion 2

- Do you think that these methods for detecting and classifying dataset shift can be implemented in industrial settings? How would an engineer/data scientist go about doing this?
- Would you expect to see differences in detection and classification techniques in different systems (like text classification or logistic regression for classifying a person)?

# Presentation Outline


- Introduction – 10 min
  - Problem Statement
  - Motivation
  - Real world example
- ***Discussion 1 – 10 minutes***
- Detecting and quantifying shift (Dhriti) – 10 min
  - *Failing Loudly*
- ***Discussion 2 – 10 minutes***
- Solutions to this problem (Michelle and Shafat) – 15 min
  - Calibration – *Revisiting the Calibration of Neural Networks*
  - Selective Classification- *Deep Gamblers*
- Conclusion
  - Strengths & Weaknesses
  - Future Work
  - Contribution to Responsible AI
- ***Discussion 3 – 10 minutes***

UNIVERSITY of VIRGINIA

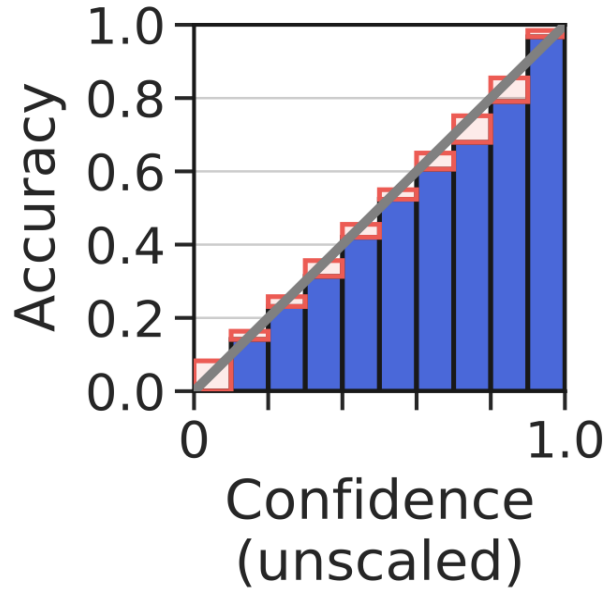
# Calibration

## Revisiting the Calibration of Modern Neural Networks

Jing-Ning Su

# What is Calibration and Why it is Important

- **Definition**: Confidence scores should match actual prediction accuracy.
- **Ideal Case**: 80% confidence → 80% correct predictions.
- **Problem**: Deep learning models tend to be overconfident.
- **Impact**:
  - Healthcare: Misdiagnoses can be fatal.
  - Autonomous Driving: Errors may cause accidents.
  - Finance: Underestimating risk leads to losses.
- **Responsible AI**: Proper calibration improves trust and transparency.



# How Calibration Helps with Distribution Shift

- **Issue**: Models remain overconfident on unseen or corrupted data, making predictions unreliable.
- **Why Calibration Matters**: Helps indicate uncertainty, enabling fallback strategies.
- **Experimental Findings**:
  - Poor calibration → Incorrect predictions with high confidence.
  - Better calibration → Meaningful confidence scores, improving robustness.
- **Real-World Impact**:
  - Autonomous Vehicles: Misreads road signs with high confidence.
  - Medical AI: Flags uncertain diagnoses for human review.

# Experiments Conducted in Research

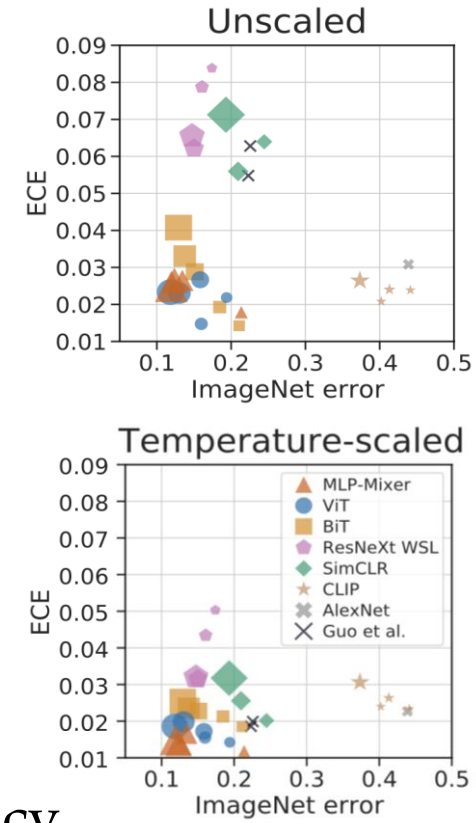
- **Scale of Study**: 180 deep learning models across 16 architecture families.
- **Architectures Studied**: CNNs, Vision Transformers (ViTs), MLP-Mixers, and others.
- **Datasets Used**: ImageNet variants
- **Calibration Metric**: Expected Calibration Error (ECE).
- **Comparison Criteria**:
  - CNNs vs. Transformers: Which is better calibrated?
  - Model Size: Does scaling improve calibration?
  - Pretraining Data: Does more data help?
- **Post-hoc Calibration**:
  - Temperature Scaling: An effective recalibration method.

| Dataset | Description |
|---|---|
| ImageNet | benchmark |
| ImageNet-C | corruptions |
| ImageNet-A | adversarial |
| ImageNet-R | artistic renditions |

# Key Experimental Results

- **Better Calibration in Newer Architectures**:
  - ViTs & MLP-Mixers outperform CNNs in calibration.
  - Contradicts earlier claims that larger models are poorly calibrated.
- **Model Size & Calibration**:
  - Larger models → Worse in-distribution calibration.

    → Better calibration under distribution shift.
- **Pretraining & Calibration**:
  - Larger datasets improve accuracy but not calibration.
  - Models trained on 300M vs. 1.3M images show similar calibration.
- **Temperature Scaling**: Reduces ECE effectively without hurting accuracy.



# Major Insights from the Findings

- **Accuracy ≠ Calibration**: The relationship varies by architecture and data distribution.
- **Generalization & Calibration**:
  - The best in-distribution models also tend to generalize better out-of-distribution.
- **Architecture Matters**:
  - ViTs & MLP-Mixers are naturally better calibrated than CNNs.
- **Beyond Post-hoc Fixes**:
  - Calibration-aware training is essential—temperature scaling alone is insufficient.
- **Comprehensive Evaluation**:
  - Both in-distribution & out-of-distribution calibration must be assessed.
  - A well-trained model may still fail in real-world scenarios.

# Contributions to Distribution Shift

- **Calibration & Robustness:**
  - Poor calibration → High-confidence failures under distribution shift.
- **Impact of Architecture:**
  - ViTs & MLP-Mixers more stable than CNNs.
- **Need for Calibration-Aware Training:**
  - Future work should integrate calibration into training, not rely on post-hoc fixes.
- **Better Calibration = Better Uncertainty Estimation:**
  - Critical for safety-critical AI applications.

# Discussion 3(a)

- Can you think of any other ways to handle distribution shift?

# Presentation Outline


- Introduction – 10 min
  - Problem Statement
  - Motivation
  - Real world example
- ***Discussion 1 – 10 minutes***
- Detecting and quantifying shift (Dhriti) – 10 min
  - *Failing Loudly*
- ***Discussion 2 – 10 minutes***
- Solutions to this problem (Michelle and Shafat) – 15 min
  - Calibration – *Revisiting the Calibration of Neural Networks*
  - Selective Classification- *Deep Gamblers*
- Conclusion
  - Strengths & Weaknesses
  - Future Work
  - Contribution to Responsible AI
- ***Discussion 3 – 10 minutes***

UNIVERSITY of VIRGINIA

# Deep Gamblers

## Learning to abstain with Portfolio Theory

# Classification and the Inadequacy of $nll$ loss

Want to find $\theta = \arg\max \Pr(Y|\theta)$

In practice, **minimize** *negative log loss*→ $nll\ loss: \min \theta[-logp(Y|\theta)]$

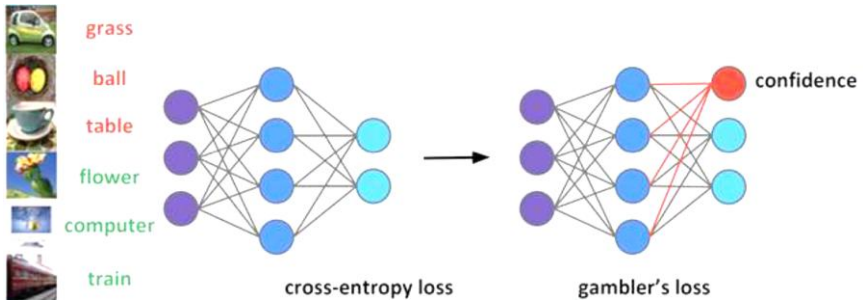
# At a glance: Deep Gamblers Approach

The proposed method: The Gambler's Loss

$$\max E\log(S) = max\sum_{i=1}^{m} p_i \log(o_i b_i + b_0)$$

$p_i$ → Probability of class *i*
$o_i$ → Odds/payoff

$b_i$ →betting amount on class *i*
$b_0$ →reservation amount (not betting)




- A novel framework for selective classification inspired by portfolio theory
- Transforms m-class classification to (m+1)-class with an abstention option
- Introduces a loss function based on the doubling rate in gambling
- Offers a mathematically principled approach satisfying all desiderata
  - ✓ End-to-end trainability
  - ✓ No heavy sampling procedure
  - ✓ No retraining for different uncertainty levels
  - ✓ No model architecture modifications

## Key Insight

**The decision to predict vs. abstain is analogous to a gambler deciding whether to bet or reserve their wealth in a horse race**

# Selective Classification Formulation

Given:

- Feature space X and label space Y
- A prediction model $f_w : X \to Y$ parameterized by weights $w$

❑ A model with rejection option is a pair of functions $(f, g)$:

$$(f, g)(x) := \begin{cases} f(x), & \text{if } g_h(x) \geq h \\ \text{ABSTAIN}, & \text{otherwise} \end{cases}$$

❑ $g_h : X \to \mathbb{R}$ is the selection function

❑ h is a threshold parameter

❑ **Coverage:** Ratio of samples where model makes predictions

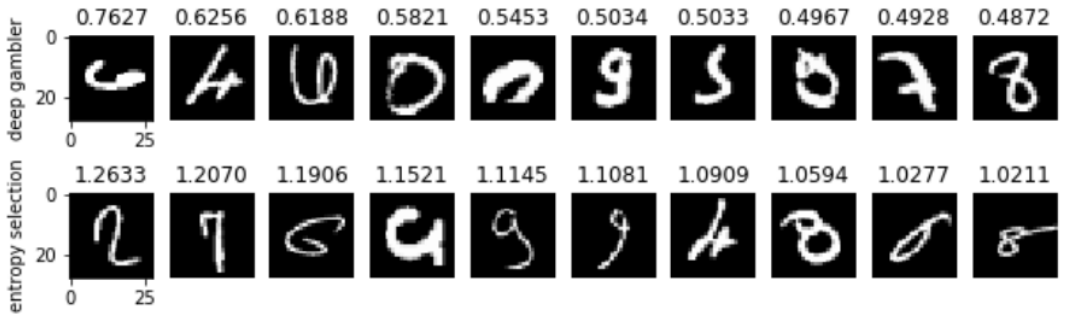
❑ **Selective risk:** Error rates on covered samples

Portfolio Theory - Deep Learning Dictionary

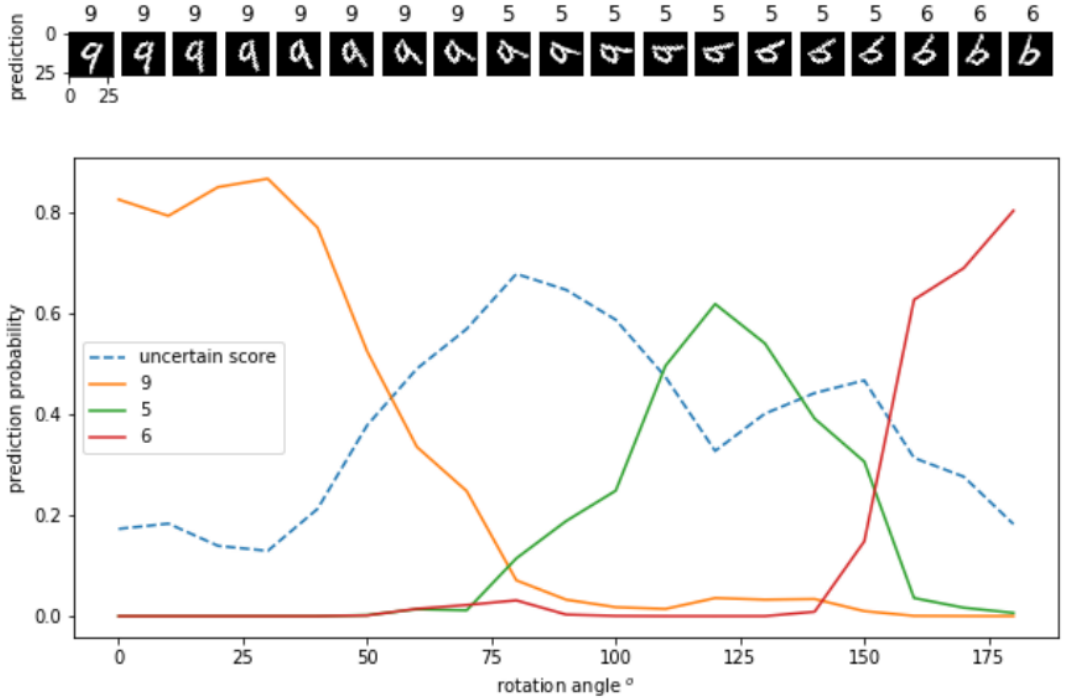
| Portfolio Theory | Deep Learning |
|---|---|
| Portfolio | Prediction |
| Doubling Rate | negative NLL loss |
| Stock/Horse | input data point |
| Stock Market Outcome | Target Label |
| Horse Race Outcome | Target Label |
| Reservation in Gamble | Abstention |

# Toy Example

Identifying Disconfident Images




Predicting Image Rotation

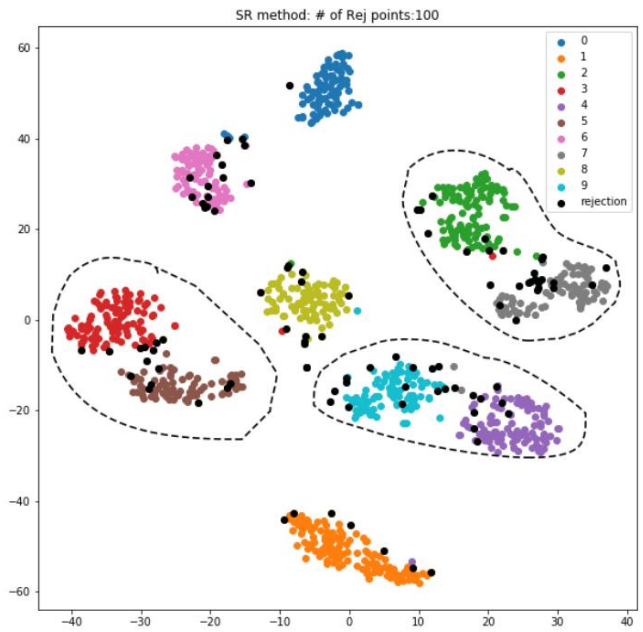


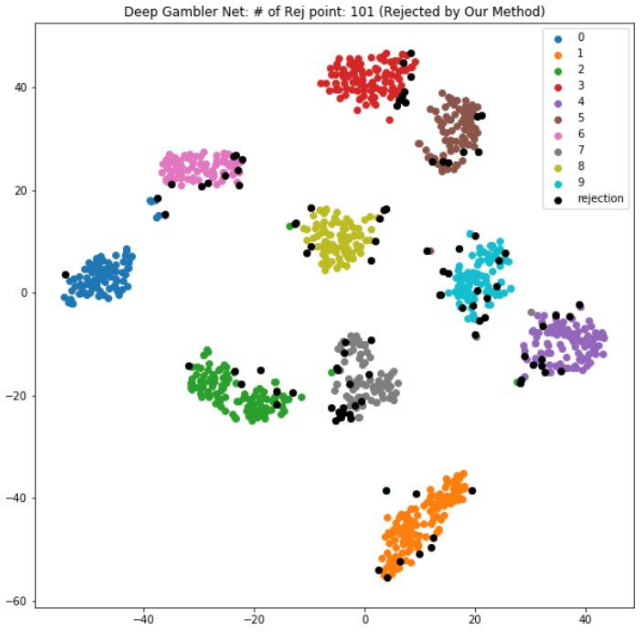

**Surprising Benefits**

The paper claims several advantages of the gambler's loss:

- ❖ Reduced overfitting during training
- ❖ Improved performance when dealing with noisy labels

# The Learned Representation is Better Separable




More distinct & separated clusters

Normal Model

Deep Gambler

## Key Takeaway

**Deep Gamblers enables neural networks to "know when they don't know" by allowing them to abstain from predictions under uncertainty**

# Contribution to Distribution Shift

- ❑ Naturally rejects out-of-distribution samples without explicit OOD training
- ❑ More separable feature representations → Robustness to data shifts
- ❑ Enables single model deployment across varying distribution shift severities via threshold tuning
- ❑ Detects semantic uncertainty when concepts drift between learned categories
- ❑ Provides a principled framework for deciding when to abstain under unfamiliar data distributions
- ❑ Outperforms entropy-based methods in identifying truly anomalous inputs

# Discussion 3(b)

- If AI systems could reliably abstain when uncertain, how would this reshape our approach to high-stakes domains like medical diagnosis or autonomous vehicles, where the cost of error vastly outweighs the cost of saying *'I don't know'*?
- The paper shows Deep Gamblers creates more separable feature representations with clearer decision boundaries. Could this approach be used primarily as a training technique even when abstention isn't needed in deployment?
- Portfolio theory offers an elegant framework for classification with abstention, but how might this gambling-inspired approach extend to other machine learning tasks like regression, reinforcement learning, or generative modeling?

# Presentation Outline


- Introduction – 10 min
  - Problem Statement
  - Motivation
  - Real world example
- ***Discussion 1 – 10 minutes***
- Detecting and quantifying shift (Dhriti) – 10 min
  - *Failing Loudly*
- ***Discussion 2 – 10 minutes***
- Solutions to this problem (Michelle and Shafat) – 15 min
  - Calibration – *Revisiting the Calibration of Neural Networks*
  - Selective Classification- *Deep Gamblers*
- Conclusion
  - Strengths & Weaknesses
  - Future Work
  - Contribution to Responsible AI
- ***Discussion 3 – 10 minutes***

# Strengths & Weaknesses

***Strengths***

***Detection***

1. Practical insights for ML practitioners
2. Discovery of an unnoticed shift in MNIST dataset

***Calibration***

1. Innovation in neural networks could lead to improved mode reliability
2. Provides valuable reference for the design and optimization of future neural networks

***Selective Classification***

1. Mathematical theory foundation is strong and aids in the overall design
2. Performance: state of the art on several benchmarks
3. Suggests benefits beyond selective prediction

***Weaknesses***

***Detection***

1. Unclear generalizability to non-image data such as text or time series data
2. Reliance on simulated shifts may not be accurate to real world distribution changes
3. Assumes access to true labels for anomalous samples

***Calibration***

1. Experimental design limitations: comparing different model families may not directly apply in the real world
2. Mainly targeted at image classification
3. ECE estimates may be affected by bias

***Selective Classification***

1. Hyperparameter sensitivity
2. Theoretical vs practical gap
3. Scope (classification only)

# Future Work and Contribution to RAI

***Future Work***

1. Detection technology that can be implemented in industry
2. ML Standards that include using the techniques in calibration and selective classification
3. ML Standards that state the necessity to publish metrics on how confident the models are
4. More research into the applicability of this in industry

***Contribution to RAI***

- Safety
  - Provides detection methods for lots of models
  - Provides two solutions for ensuring that our models are safe to use
- Confidence
  - In ensuring safety, we bolster confidence in our models as well
- Mitigates (bad) effects of models
- Provides for some focus on model safety within artificial intelligence community

# Discussion 4

- What threshold would you say is good enough for a model to confidently make the prediction they make or abstain from predicting? Why? In what scenarios would you be comfortable raising and lowering that threshold?
- Will calibration and selective classification work in industry? Why or why not? What industries might they not work in? What stakeholders would be affected and how would different stakeholders respond?
- How do we operationalize these findings such that active software engineers/data scientists can pick up these results and implement them into their work?
- How can we check for dataset drift in systems that rely on differential privacy (ie, they add noise to their data via a laplace mechanism)?

# THANK YOU

# Bibliography


1. Rabanser, S., Günnemann, S., & Lipton, Z. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, *32*.
2. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., ... & Lucic, M. (2021). Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, *34*, 15682-15694.
3. Liu, Z., Wang, Z., Liang, P. P., Salakhutdinov, R. R., Morency, L. P., & Ueda, M. (2019). Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, *32*.
4. Stewart, Matthew. "Understanding Dataset Shift." *Medium*, TDS Archive, 29 July 2020, medium.com/towards-data-science/understanding-dataset-shift-f2a5a262a766.
5. Shubham.jain. "Covariate Shift - Unearthing Hidden Problems in Real World Data Science." *Analytics Vidhya*, 21 Oct. 2024, www.analyticsvidhya.com/blog/2017/07/covariate-shift-the-hidden-problem-of-real-world-data-science/.
6. Ataei, Mehdi, et al. "Understanding Dataset Shift and Potential Remedies." *Understanding Dataset Shift and Potential Remedies*, 2021, vectorinstitute.ai/wp-content/uploads/2021/08/ds_project_report_final_august9.pdf.