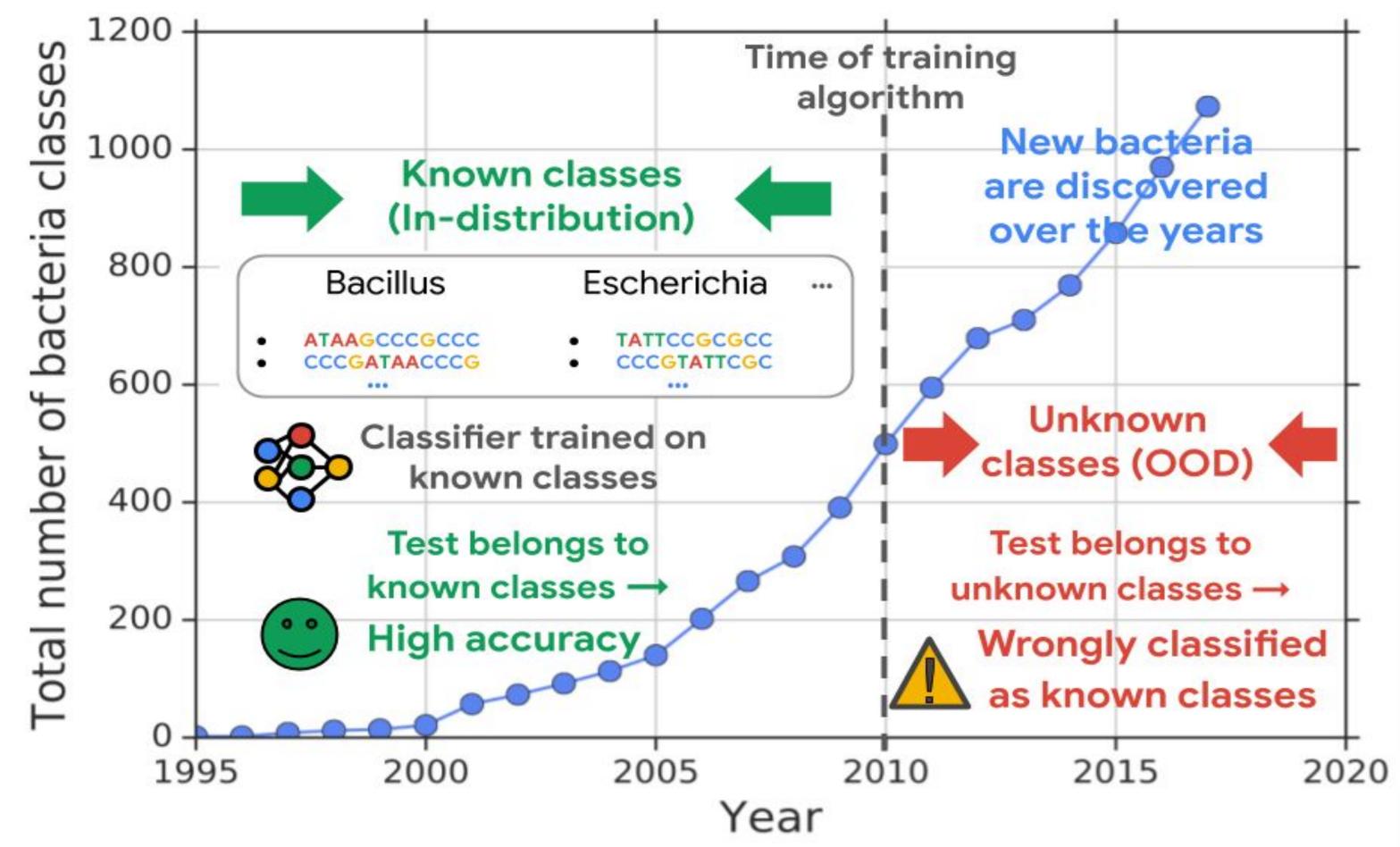
LIKELIHOOD RATIOS FOR OUT-OF-DISTRIBUTION DETECTION Jie Ren*, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, Balaji Lakshminarayanan*

https://github.com/google-research/google-research/tree/master/genomics_ood *Contact: {<u>ijren, peterjliu, balajiln}@qoogle.com</u>

1. MOTIVATION

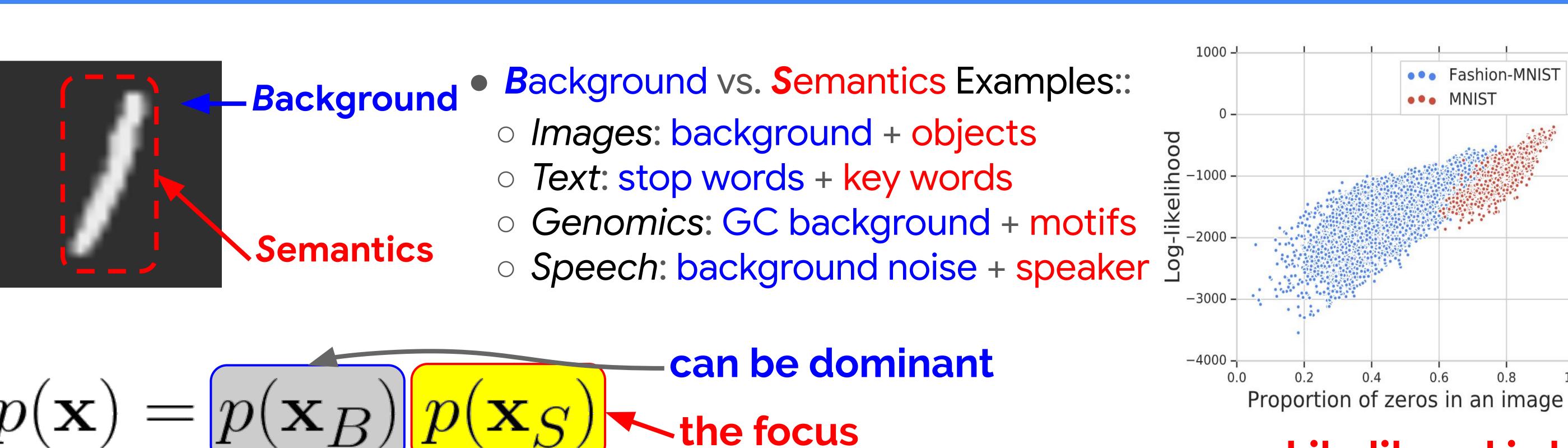
- Bacteria identification based on genomic sequences
- ACGTTAACAACC...GGCTTC ⇒ label
- Promising for early detection of disease
- Classifier can achieve high accuracy on known classes, but perform poorly in real world:
- 60-80% of real-world test inputs belong to as yet **unknown** bacteria
- Ideally, say "I don't know" on OOD inputs than assign high-confidence predictions
- Need accurate OOD detection to ensure safe deployment of classifier



- We create a realistic benchmark for OOD detection on genomics data.
- 10 in-distribution, 60 OOD validation, 60 OOD test classes.
- Classes split by year to reflect challenges faced when classifier trained only on known classes

< 01/01/2011 01/01/2011 ~ 01/01/2016 > 01/01/2016		
In-distribution	In-distribution validation	In-distribution test
training	OOD validation	OOD test

- Challenge: Detect if a test input is OOD (i.e. it does not belong to any of the training classes) Unsupervised: Density-based approaches
- Supervised: Classifier-based approaches



$$p(\mathbf{x}) =$$

for OOD

 $LLR(\mathbf{x}) =$

assuming both models capture background equally well.

New benchmark dataset + code is available at

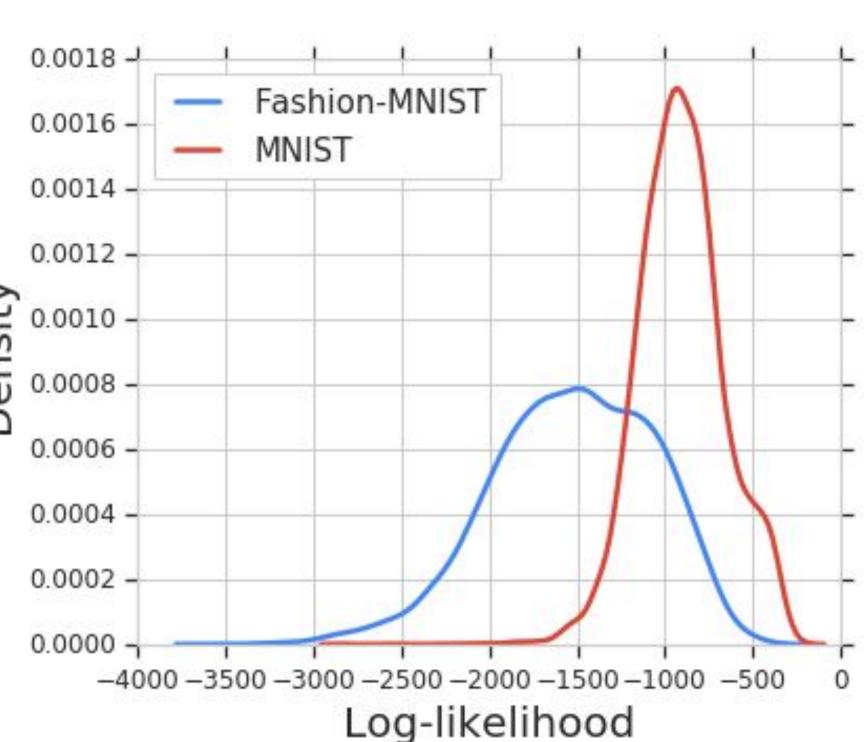
2. GENERATIVE MODELS CAN ASSIGN HIGHER LIKELIHOOD TO OOD INPUTS

• Generative models for OOD detection:

 do not require labeled data • model the input distribution $p_{TRAIN}(\mathbf{x})$ and evaluate the likelihood of new inputs. Prior work [Nalisnick et al., 2018, Choi et al. 2019]

observed failure modes of generative models:

Higher likelihoods for OOD than in-dist. e.g. Fashion-MNIST (in-dist.) vs. MNIST (OOD)



• We observe a similar failure mode on generative models trained on genomic sequences.

3. EXPLAINING WHY DENSITY MODELS FAIL AT OOD DETECTION

• $p(\mathbf{x})$ has to explain both semantic & background components Humans ignore background and focus primarily on semantics

Likelihood is highly correlated with the background

 proportion of zeros in an image • GC-content in genomic sequence

4. PROPOSED SOLUTION: LIKELIHOOD RATIOS FOR OOD DETECTION

 How do we automatically extract the semantic component of p(x)?

• We propose training a background model on perturbed inputs and computing the likelihood ratio:

$$= \log \frac{p_{\theta}(\mathbf{x})}{p_{\theta_0}(\mathbf{x})} = \log \frac{p_{\theta}(\mathbf{x}_B) \ p_{\theta}(\mathbf{x}_S)}{p_{\theta_0}(\mathbf{x}_B) \ p_{\theta_0}(\mathbf{x}_S)} \approx \log \frac{p_{\theta}(\mathbf{x}_S)}{p_{\theta_0}(\mathbf{x}_S)}$$

• LLR is a background contrastive score: the significance of the semantics compared with the background.

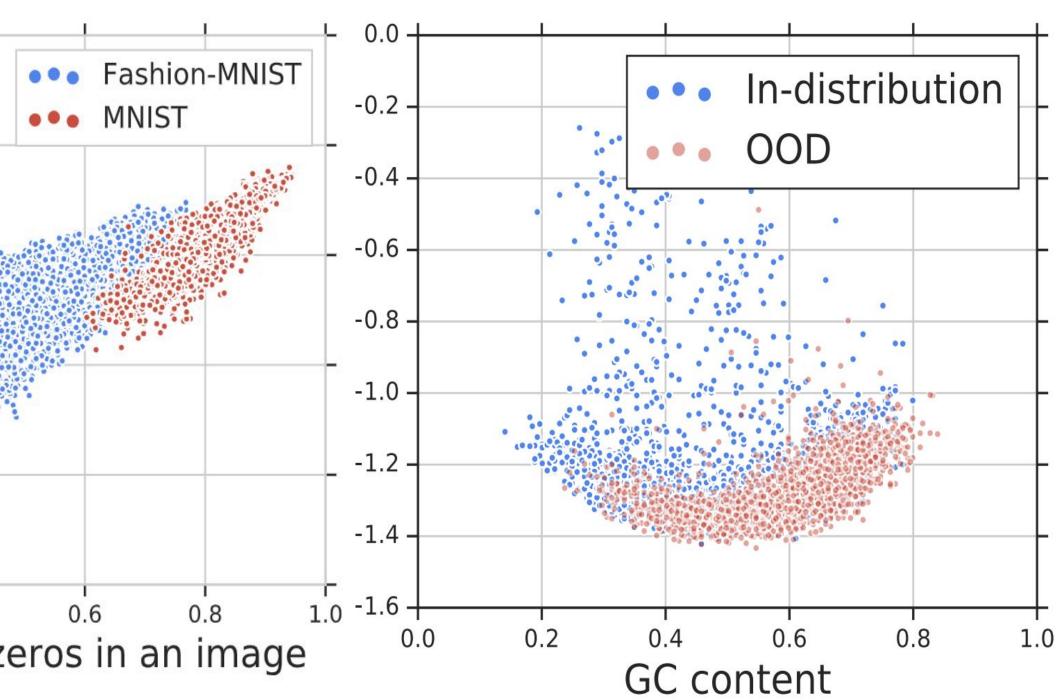
Algorithm

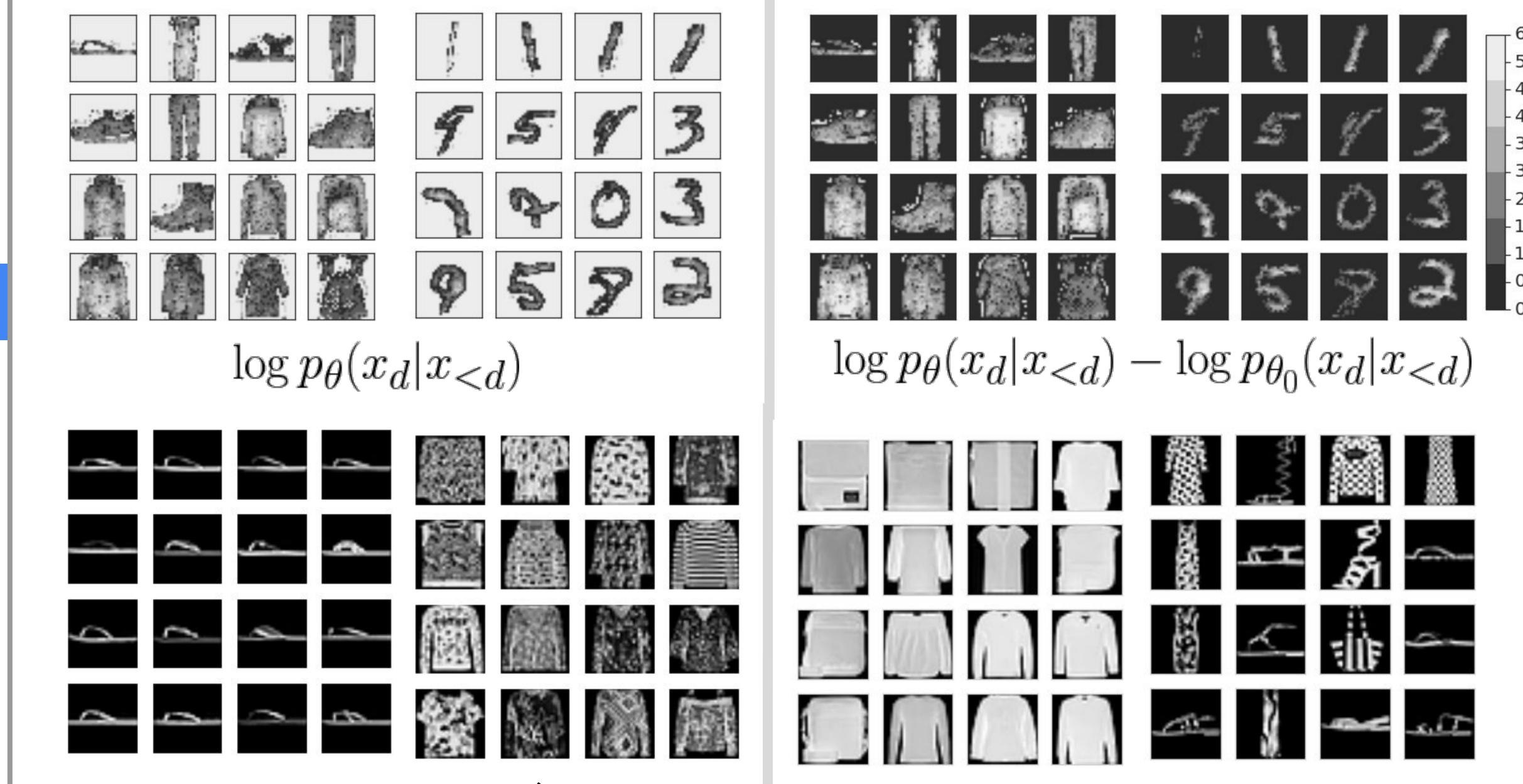
- and optionally model regularization*.
- Fit $p_{\theta_0}(\mathbf{x})$ using perturbed input data
- Fit $p_{\theta}(\mathbf{x})$ using in-distribution data - Compute the likelihood ratio.
- Predict OOD if likelihood ratio is small.

*Hyperparameters (mutation rate and L2 coefficient) are tuned using an independent OOD dataset different from test OOD.



In distribution J - 1.6 - 1.4 - 1.2 - 1.0 - 0.8 - 0.6 - 0.4 - 0.2 0.0 Log-likelihood



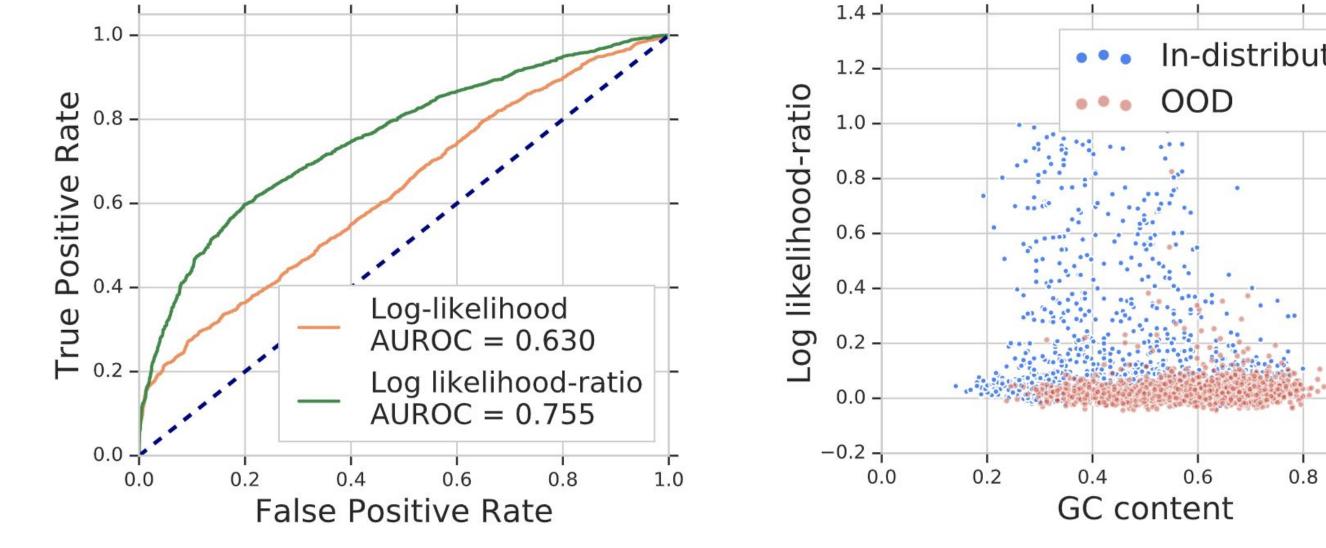


Images with highest (high portion of background) and lowest likelihood

Images with highest (prototypical) & lowest likelihood ratio (rare patterns)

6. OOD DETECTION FOR GENOMIC SEQUENCES

- LSTM model is trained using sequences
- Likelihood Ratio significantly improve
- Effect of background GC-content is cort
- OOD detection correlates with its distar



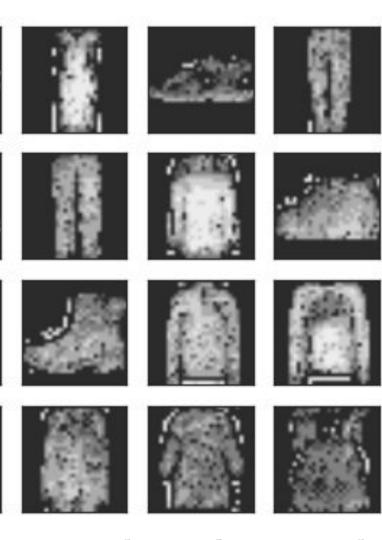
Summary

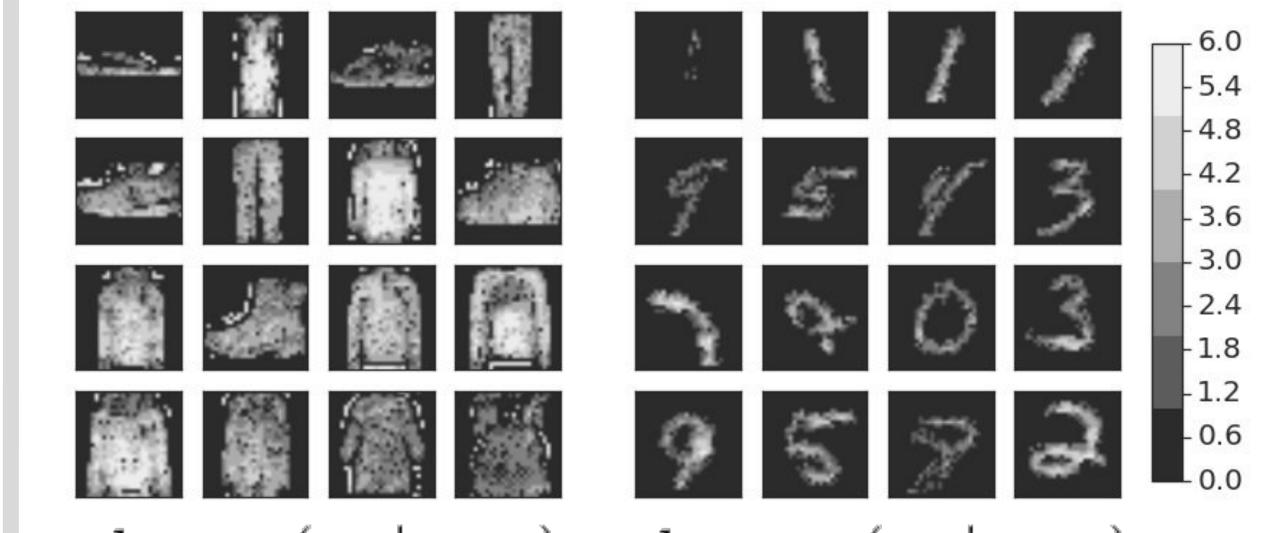


Google Al O DeepMind

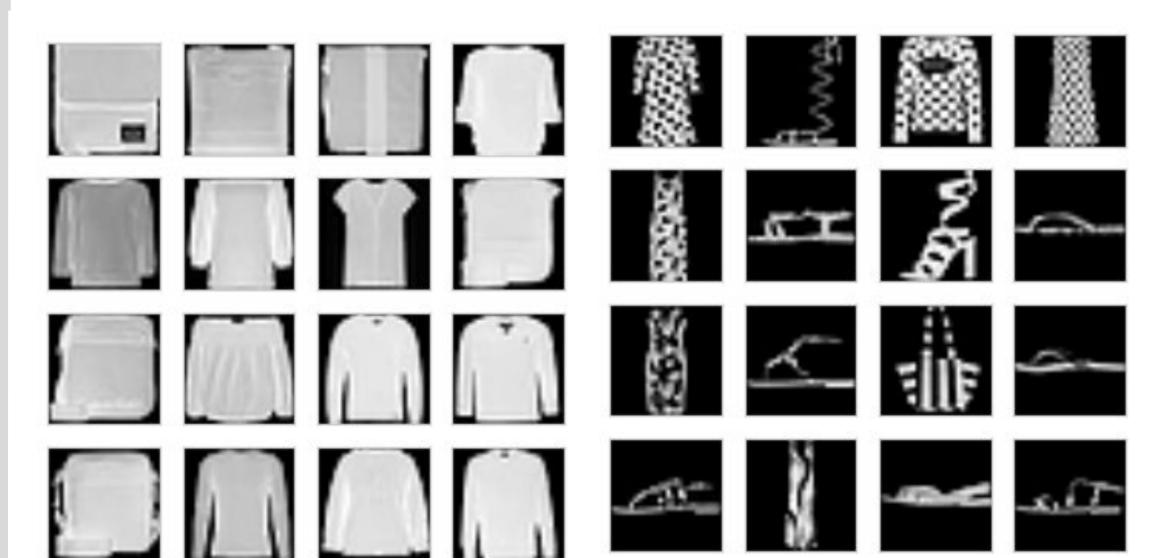
5. OOD DETECTION FOR IMAGES

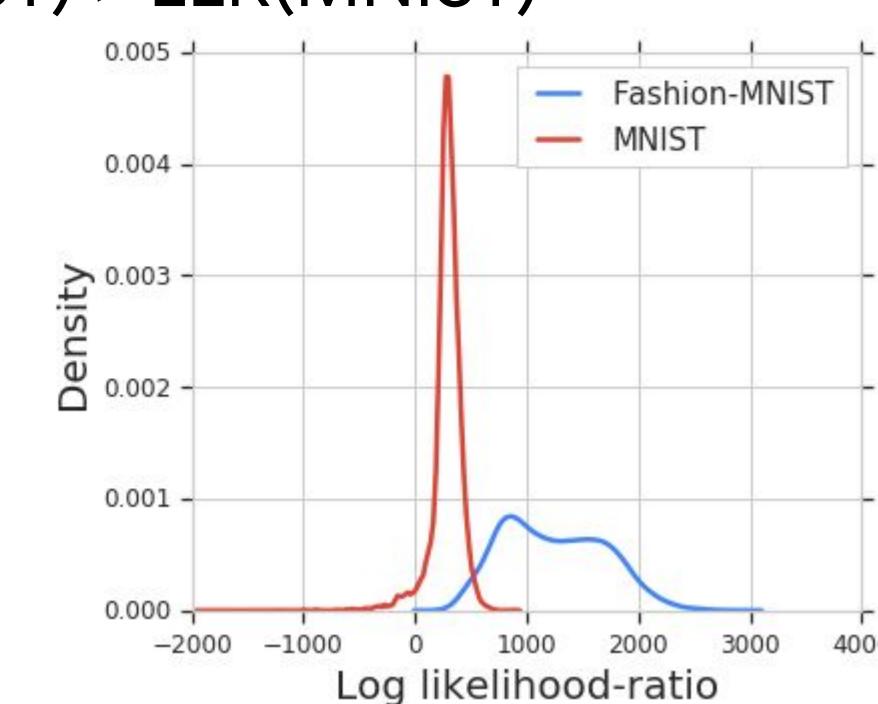
Investigate auto-regressive models: which pixels contribute the most to the likelihood (ratio)? • Fashion-MNIST (in-dist.) vs. MNIST (OOD). PixelCNN++ model is trained on Fashion-MNIST. • Likelihood is dominated by the background pixels $\Rightarrow p(Fashion-MNIST) < p(MNIST)$ • Likelihood ratio focuses on the semantic pixels \Rightarrow LLR(Fashion-MNIST) > LLR(MNIST)





 $\log p_{\theta}(x_d | x_{< d}) - \log p_{\theta_0}(x_d | x_{< d})$





AUROC
0.089
0.994
0.734
0.746
0.752
0.839
0.942

s from in-distribution classes	Method	AUROC
ves OOD Detection	Likelihood	0.626
orrected ance to in-distribution	Likelihood Ratio	0.755
ribution 0.9 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.7 0.8 0.8 0.7 0.8 0.8 0.7 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8	Classifier-based p(y x)	0.634
	Classifier-based Entropy	0.634
	Classifier-based ODIN	0.697
	Classifier Ensemble 5	0.682
0.5 0.4 0.4 0.3 0.5 0.10 0.15 0.20 0.25 0.30 0.35 Minimum distance to in-distribution	Classifier-based Mahalanobis Distance	0.525

• Create a realistic benchmark dataset for OOD detection (and open-set classification) in genomics Show that the likelihood from deep generative models can be confounded by background statistics • Propose a likelihood ratio method for unsupervised OOD detection, outperforming the raw likelihood • Our method performs well on images and achieves SOTA performance on genomic dataset.