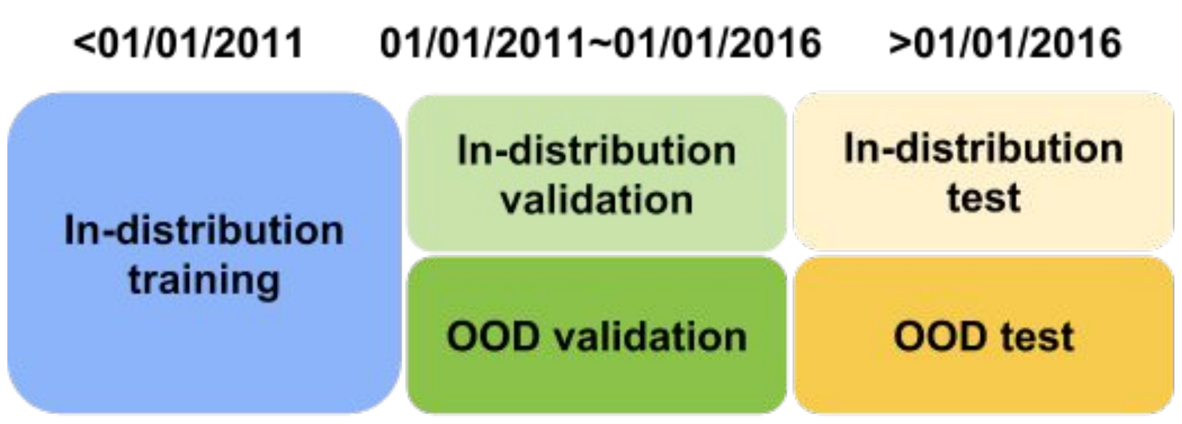
Likelihood Ratios for Out-of-Distribution Detection

Jie Ren*, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, Balaji Lakshminarayanan*

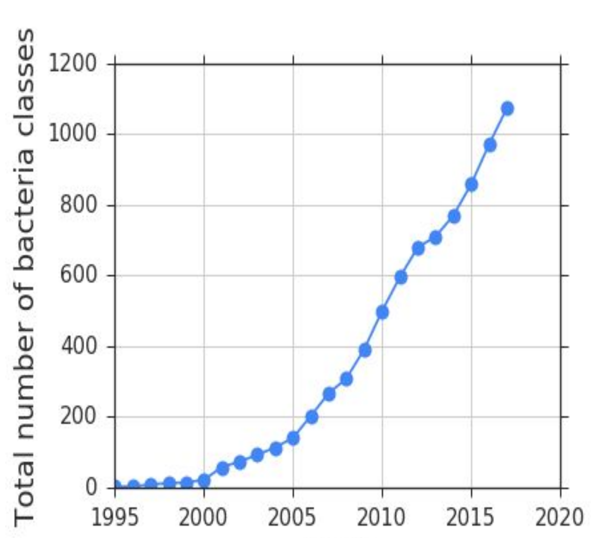
Google AI DeepMind

1. Introduction

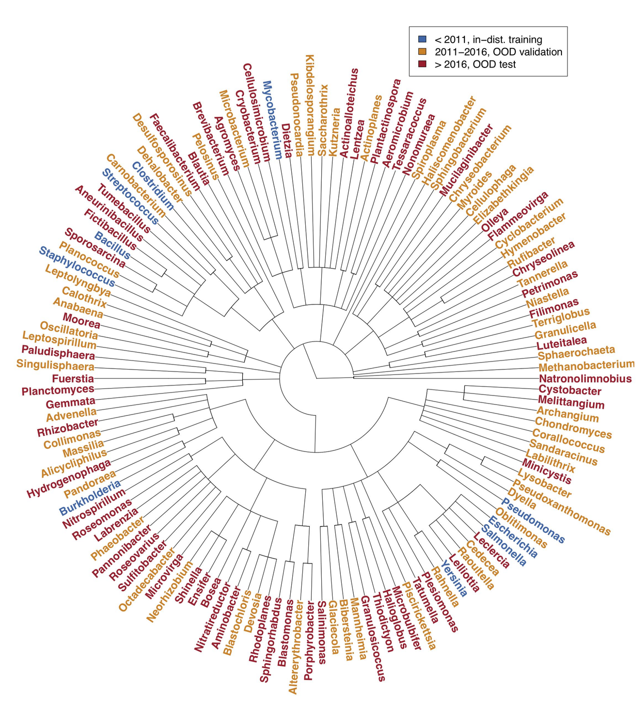
- Discriminative models offer little performance guarantees on **out-of-distribution (OOD) inputs**, limiting the **AI safety** in real-world applications.
- **Bacteria identification based on genomic sequences** holds the promise of early detection of disease.
- ML classifiers perform poorly in real world, because real data **contains 60-80% genomic sequences from unknown bacteria** and other contaminants.
- We create **a realistic benchmark for OOD detection on genomics data**.
- We propose a **Likelihood Ratio** method for OOD detection, achieving SOTA on genomics data



10 in-distribution, 60 OOD validation, 60 OOD test classes.



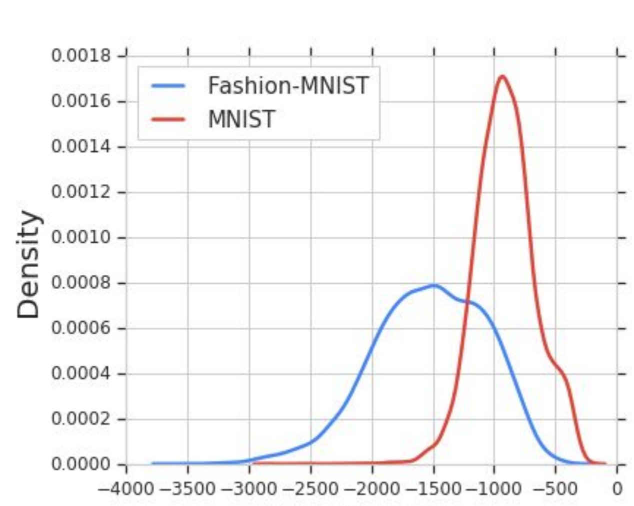
Bacterial classes are discovered gradually over the years (not saturated yet).

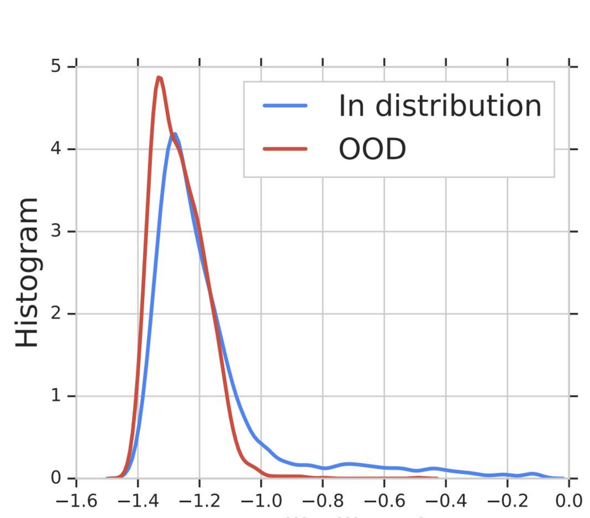


In-distribution and OOD classes are interlaced in phylogeny

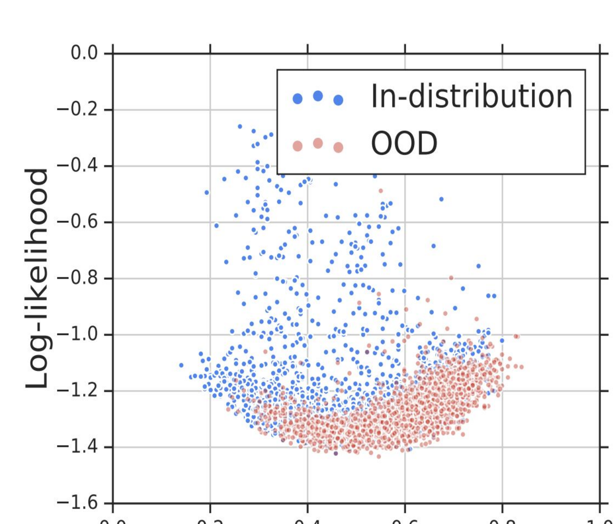
2. Generative Models Can Assign Higher Likelihood to OOD Inputs

- **Generative models:**
  - do not require labeled data
  - model the input distribution $p(x_{TRAIN})$ and then evaluate the likelihood of new inputs.
- **Higher likelihoods for OOD than in-dist.** in Fashion-MNIST (in-dist.) vs. MNIST (OOD) [Nalisnick et al., 2018, Choi et al. 2019].





- We observe a **similar phenomenon on genomic sequences.**
- The likelihood is **heavily affected by the sequence's GC-content (background statistics).**



3. Likelihood Ratios For OOD Detection

- Assumption: An input $\mathbf{x}$ is composed of two components
  - **Background** $\mathbf{x}_B$: population level background statistics
  - **Semantic** $\mathbf{x}_S$: in-dist. specific features. See examples.

$$p(\mathbf{x}) = \boxed{p(\mathbf{x}_B)}\,\boxed{p(\mathbf{x}_S)}$$

can be dominant

the focus

- To focus on $\mathbf{x}_S$ we propose (1) **training a background model** on perturbed inputs and (2) computing the **likelihood ratio**:

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

assuming both models capture background equally well.

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

Examples of Background vs Semantics:

- *Images*: background + objects
- *Text*: stop words + key words
- *Genomics*: GC background + motifs
- *Speech*: background noise + speaker

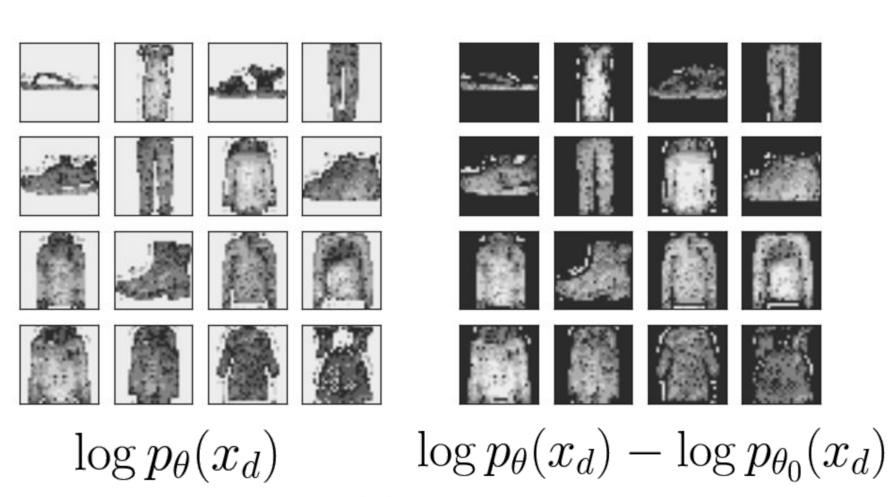
Algorithm

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

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

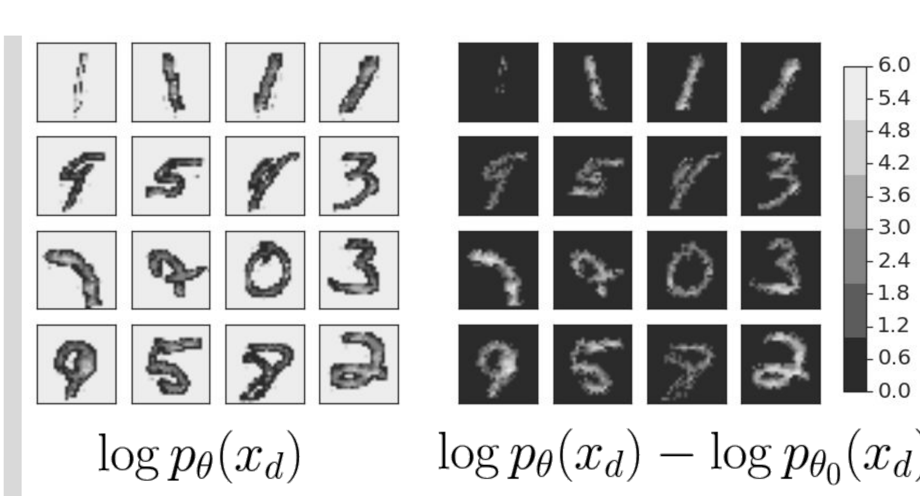
4. OOD Detection For Images

- 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** ⇒ p(Fashion-MNIST) < p(MNIST)
- **Likelihood ratio focuses on the semantic pixels** ⇒ LLR(Fashion-MNIST) > LLR(MNIST)

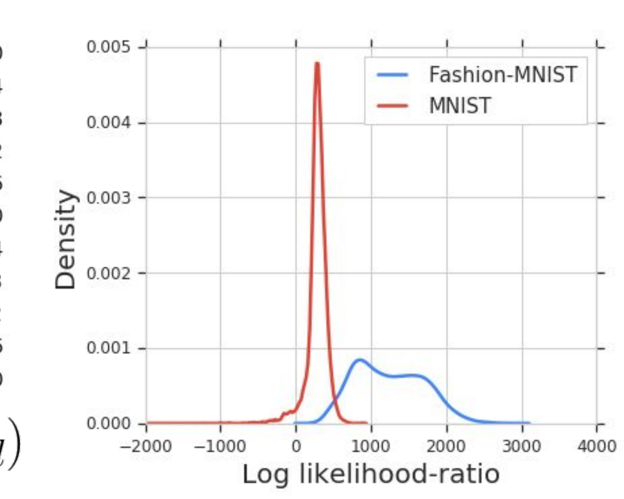


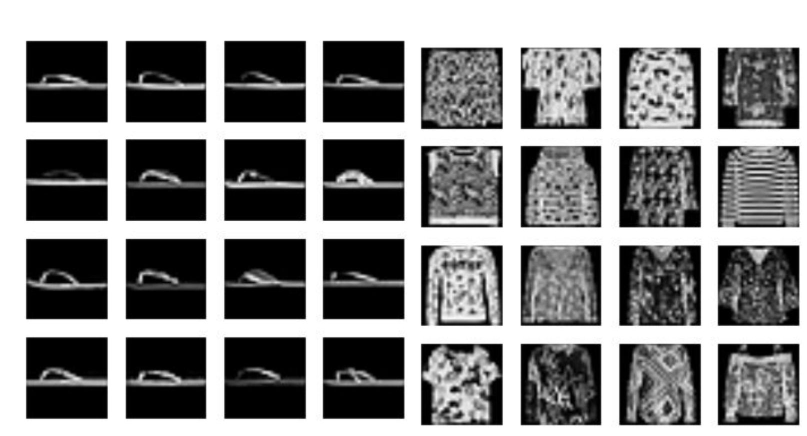
Fashion-MNIST

*(sorted by likelihood from high to low)

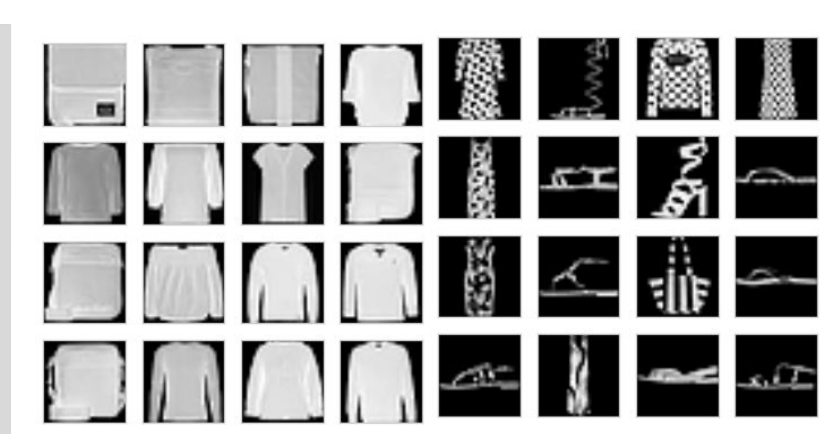


MNIST





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

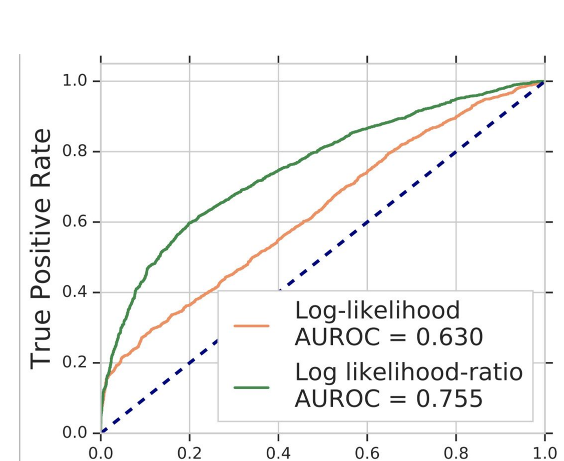


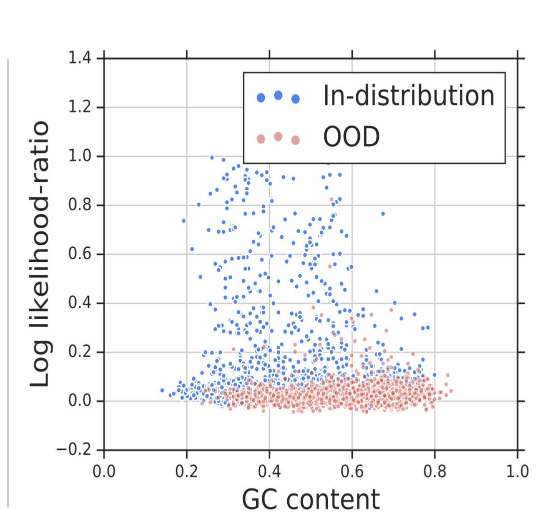
Images with highest (prototypical icons) **& lowest likelihood ratio** (rare patterns)

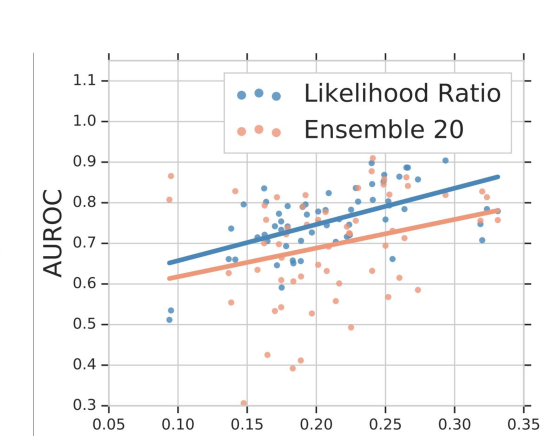
Method	AUROC
Likelihood	0.115
Likelihood Ratio	**0.997**
Classifier-based p(y|x)	0.579
Classifier-based Entropy	0.588
Classifier-based ODIN	0.620
Classifier Ensemble 5	0.832
Classifier-based Mahalanobis Distance	0.986

5. OOD Detection For Genomics

- LSTM model is trained using sequences from in-distribution classes
- Likelihood Ratio significantly improves OOD Detection
- Effect of background GC-content is corrected
- OOD detection correlates with its distance to in-distribution*







*Pearson Correlation 0.569 (LLR), 0.277 (Ensemble)

Method	AUROC
Likelihood	0.630
Likelihood Ratio	**0.755**
Classifier-based p(y|x)	0.622
Classifier-based Entropy	0.622
Classifier-based ODIN	0.645
Classifier Ensemble 5	0.673
Classifier-based Mahalanobis Distance	0.496

Summary

- Create a **realistic benchmark dataset** for OOD detection in genomics
- Show that the likelihood from deep generative models can be **confounded by background statistics**
- Propose a **likelihood ratio method** for OOD detection, outperforming the raw likelihood
- Our method **achieves state-of-the-art performance on genomic dataset**.

Check the ArXiv Version for details

Contact: jjren@google.com, balajiln@google.com