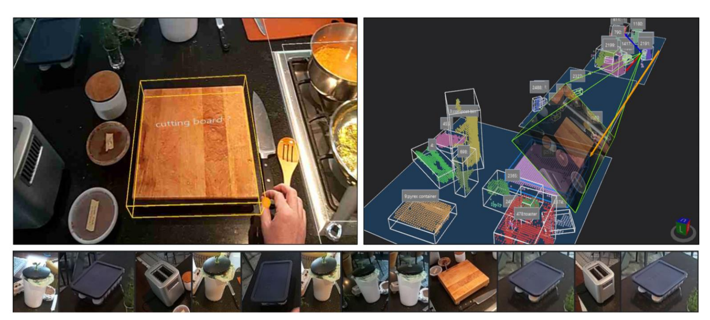
# Streaming Active Learning with Deep Neural Networks

Akanksha Saran, Safoora Yousefi, Akshay Krishnamurthy, John Langford, Jordan T. Ash

## Real-World Applications with Streaming Data Settings

In several real-world applications, data arrive in a stream and the total number of samples are unknown ahead of time.

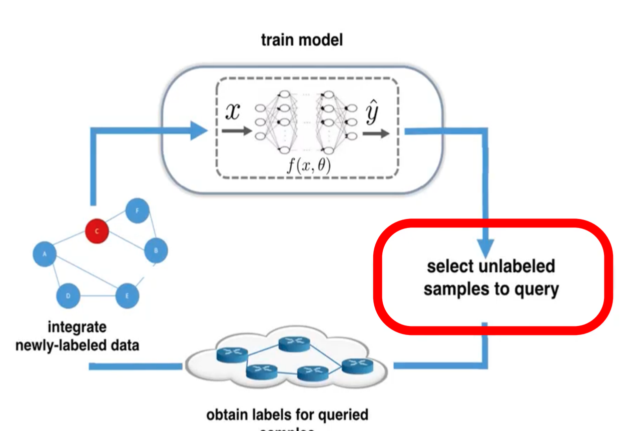
- Interaction-centric AR/VR applications such as continual object/activity learning in the wild
- Fixed datasets that are large, fractured and interacted via streaming, distributed data frameworks



**How can we train deep neural networks in a data efficient manner for streaming applications?**

## Batch Active Learning for Deep Neural Networks

- Batch active learning or pool-based active learning for deep neural networks identifies a batch of k samples from an unlabeled data pool to be integrated into the training set.
- Popular approaches for batch active learning rely on samplers that require all unlabeled data to be simultaneously available.

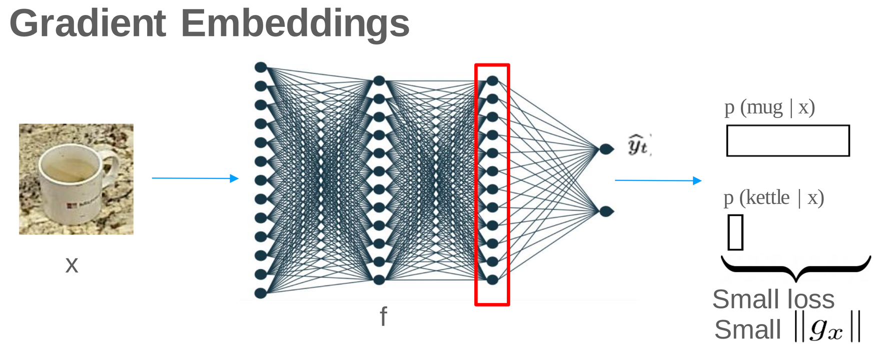



State-of-the-art non-streaming batch active learning method BADGE [1] trades off between the model's **uncertainty** about data labels and **diversity** of samples in the batch.

**Representation: Hypothetical Gradient Embeddings**

$$\hat{y}_t = \arg\max f(x_t;\theta)$$

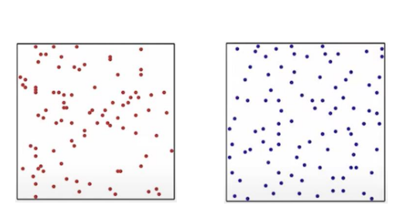
$$g(x_t) = \frac{\partial}{\partial \theta_L}\ell(f(x_t;\theta), \hat{y}_t)$$




**Sampling: Volume Sampling**

$$p_B \propto \det\Big(\sum_{x\in B} g(x)g(x)^\top\Big)$$

The determinant for volume sampling is large for a batch of high magnitude, linearly independent samples, encouraging diversity in the batch.



Random Sampling | Volume Sampling

## Streaming Batch Active Learning for Deep Neural Networks

For streaming batch active learning, it is desirable to approximate volume sampling with the following properties:

**Committal:** Select samples for querying as soon as they arrive in the stream

**Equitable sampling:** Distribute labeling queries evenly across the data stream to match a maximum query rate q

**Approximate Volume Sampling**

Sampling a batch of points

$$p_B \propto \det\Big(\sum_{x\in B} g(x)g(x)^\top\Big)$$

Sampling a single point

$$p_t \propto \det\Big(\widehat{\Sigma}_t + \underbrace{g(x_t)g(x_t)^\top}\Big)$$

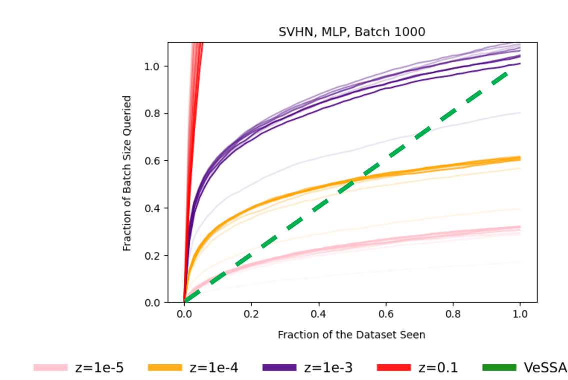
Covariance of points already chosen in the batch

Determinantal contribution of the current sample

$$\mathbb{E}_x[p_t] = \mathbb{E}_x\Big[\boxed{z_t}\cdot \underbrace{g(x_t)^\top\widehat{\Sigma}_t^{-1}g(x_t)}_{\text{Elliptical Potential}}\Big] = \boxed{q}$$

Scaling term

Query Rate




## VeSSAL: VolumE Sampling for Streaming Active Learning

$$\begin{aligned}\mathbb{E}_x\Big[z_t\cdot g(x)^\top\hat{\Sigma}_t^{-1}g(x)\Big] &= z_t\cdot\mathbb{E}_x\Big[\mathrm{tr}\Big(g(x)^\top\hat{\Sigma}_t^{-1}g(x)\Big)\Big]\\ &= z_t\cdot\mathbb{E}_x\Big[\mathrm{tr}\Big(\hat{\Sigma}_t^{-1}g(x)g(x)^\top\Big)\Big]\\ &= z_t\cdot\mathrm{tr}\Big(\hat{\Sigma}_t^{-1}\mathbb{E}_x\Big[g(x)g(x)^\top\Big]\Big).\end{aligned}$$

VeSSAL (algebraically) autotunes the scaling term $z_t$ by disentangling the gradient statistics $\mathbb{E}_x\big[g(x)g(x)^\top\big]$ from the constantly evolving $\hat{\Sigma}_t^{-1}$ .

$$\mathbb{E}_x[p_t] = \mathbb{E}_x\Big[z_t\cdot g(x_t)^\top\widehat{\Sigma}_t^{-1}g(x_t)\Big] = q \quad [1]$$

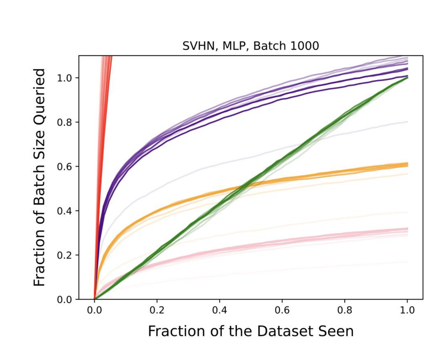
$$= z_t\,\mathrm{tr}\Big(\widehat{\Sigma}_t^{-1}\mathbb{E}_x\Big[\underbrace{g(x_t)g(x_t)^\top}\Big]\Big) \quad [2]$$

Inverse covariance of points already chosen in the batch

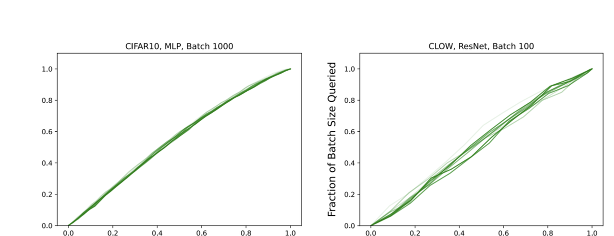
Covariance of all samples seen so far

$$\Longrightarrow \quad p_t = \frac{q\cdot g(x_t)^\top\hat{\Sigma}_t^{-1}g(x_t)}{\mathrm{tr}\Big(\frac{1}{t}\hat{\Sigma}_t^{-1}\sum_{i=1}^t g(x_i)g(x_i)^\top\Big)}$$

Sampling Probability




$$q_t = \frac{k - |B_t|}{n - t}$$



**Algorithm 1** Volume sampling for streaming active learning (VeSSAL)

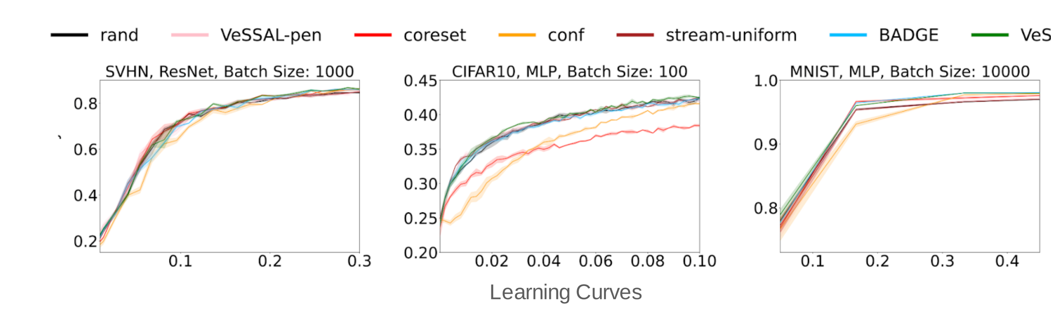
**Require:** Neural network $f(x;\theta)$, unlabeled stream of samples $U$, ideal sampling rate $q$

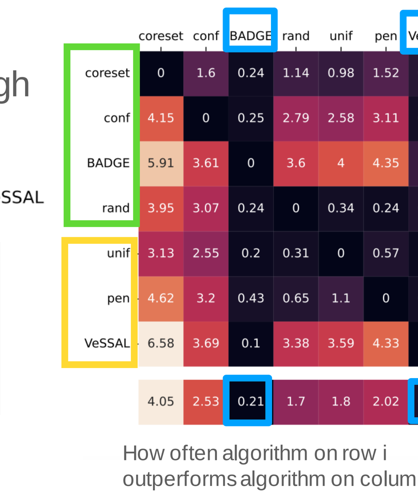
1: Initialize $t = 1$
2: Initialize $\hat{\Sigma}_0^{-1} = \lambda^{-1}I_d$ {regularized by $\lambda$ for stability}
3: Initialize $A_0 = 0_{d,d}$ {covariance over all data}
4: Initialize $B = \emptyset$ {set of chosen samples}
5: **for** $x_t \in U$: **do**
6: $A_t \leftarrow \frac{t-1}{t}A_{t-1} + \frac{1}{t}g(x_t)g(x_t)^\top$
7: $p_t = q\cdot g(x_t)^\top\hat{\Sigma}_t^{-1}g(x_t)\,\mathrm{tr}(\hat{\Sigma}_t^{-1}A_t)^{-1}$
8: **with** probability $\min(p_t, 1)$:
9: Query label $y_t$ for sample $x_t$
10: $B \leftarrow B \cup (x_t, y_t)$
11: $\hat{\Sigma}_{t+1}^{-1} \leftarrow \hat{\Sigma}_t^{-1} - \frac{\hat{\Sigma}_t^{-1}g(x_t)g(x_t)^\top\hat{\Sigma}_t^{-1}}{1+g(x_t)^\top\hat{\Sigma}_t^{-1}g(x_t)}$ {rank-1 Woodbury update}
12: **else**:
13: $\hat{\Sigma}_{t+1}^{-1} \leftarrow \hat{\Sigma}_t^{-1}$
14: $t \leftarrow t+1$
15: **return** labeled batch $B$ for retraining $f$
16: **end for**

## Results

We conduct experiments with 4 datasets x 3 batch sizes x 3 neural network architectures x 7 active learning algorithms (streaming and non-streaming ).

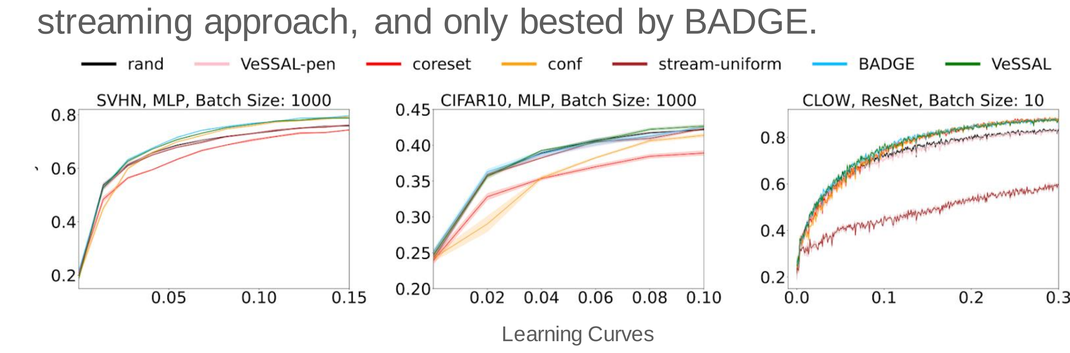
**I.I.D. Data Stream**: VeSSAL produces models with predictive capabilities on par with state-of-the- art approaches, even though they are not restricted to the streaming, committal setting.

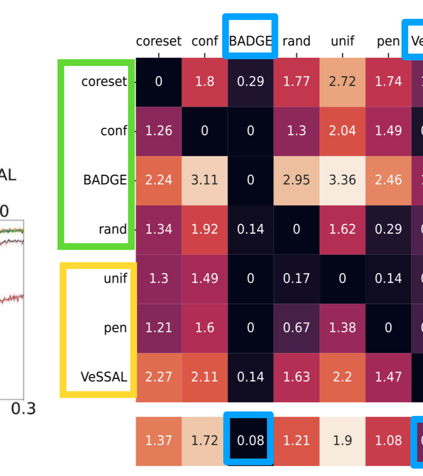


How often algorithm on row i outperforms algorithm on column j

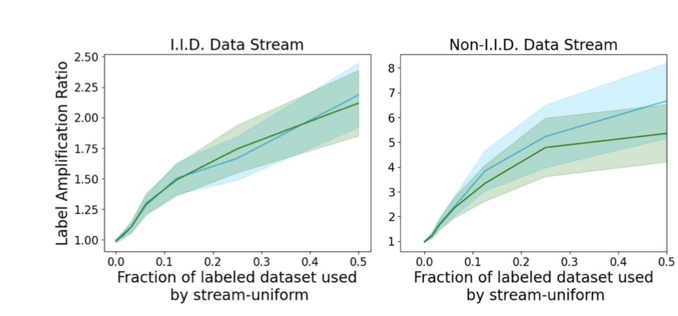
**Non-I.I.D. Data Stream**: VeSSAL suffers minimally under data streams which induce domain drift. It is the highest performing streaming approach, and only bested by BADGE.



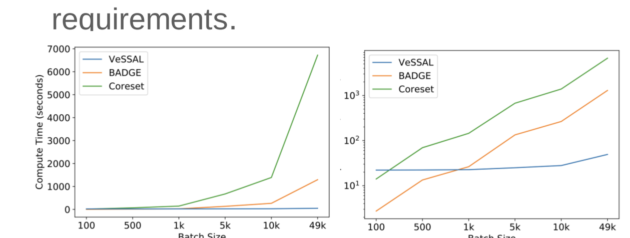



**Predictive Power**: VeSSAL delivers more predictive power (up to 5x) for the same labelling budget compared to uniform sampling in streaming settings.

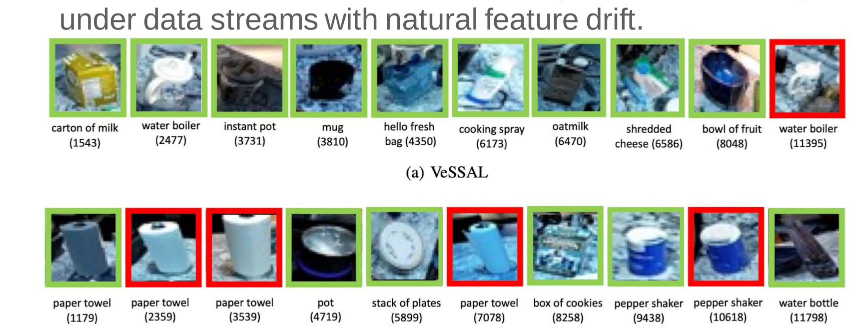
This is evaluated using the **Label Amplification Ratio** which is the number of samples used by a uniform sampling approach divided by the number of samples required by an active sampling approach to reach the same performance.




**Compute Requirements:** VeSSAL enjoys fixed run time with increasing batch sizes, while other non-streaming approaches have super-linear compute requirements.




**Qualitative Results:** VeSSAL samples diverse images under data streams with natural feature drift.




**VeSSAL is a high-performing, hyperparameter free, computationally efficient, committal acquisition function that trades off between diversity & uncertainty from a stream of samples to match a desired query rate.**

## References

[1] Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, un- certain gradient lower bounds. *International Conference on Learning Representations*, 2020.

[2] Ash, J., Goel, S., Krishnamurthy, A., and Kakade, S. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34: 8927–8939, 2021.

[3] MacKay, D. J. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[4] Settles, B. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[5] Bohus, D., Andrist, S., Feniello, A., Saw, N., and Horvitz, E. Continual learning about objects in the wild: An interactive approach. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 476–486, 2022.