General Reply

We thank all reviewers for their constructive comments. Overall, the reviewers appreciated the importance of the problem, the practical utility of our results, the rigorous experiments, and the clarity of our writing. At the same time we recognize that the scores were split, with R3 advocating acceptance and R1, R2 leaning towards weak reject. We are optimistic that our rebuttal can address each reviewer's individual concerns below:

Reply to R1:

Thanks for your thoughtful comments on how we can improve our draft.

"The stud[y] is restricted to "natural" shifts, ie. no attacks." In fact we do run experiments in which shifts are due to adversarial attack, finding, interestingly that here two-sample tests performed in the ambient space are most effective.

"the submission is similar to a review paper." While we agree that our work draws upon many other papers (like a review) would, our contribution here is a rigorous empirical study with results that stem from combining existing techniques in new ways. BBSD has never been applied beyond the label shift setting and its wider applicability (in practice) had never previously been demonstrated. Moreover, the general DR + 2-sample test pipeline has never been systematically investigated. We believe that this constitutes a source of fundamental scientific insights. In short, we wish to assert that the novelty of our paper is in the science, not the introduction of new models.

"the main contribution of this work should be there," While we agree that identifying benign vs malign examples is of paramount importance, we hope to convince you that detection itself, which is better developed in our work is a sufficiently vital problem. We hope to make more progress on detecting malign examples in future work.

"the author could [summarize] the way they select the most anomalous samples" While we considered two approaches, our promising results to date consist of selecting those examples that a domain classifier assigns with highest probability to the target domain.

"Better interpretation of results" We will discuss our intuitions and interpret the results more thoroughly if accepted.

Details: ***a*) Sec. 3.1**: We will clarify that all of our DR + two-sampling methods are non-conventional. The label classifier applied broadly as a shift detector is the most successful technique and indeed our contribution. However, domain classifier approaches have been explored previously. ***b*) L 139**: We chose the Gaussian kernel for MMD following the original paper. Per your suggestion, we will add ablations to show sensitivity to kernel bandwidth. ***c*) L 165-166**: c and r refer to the number of columns and rows of the contingency table. ***d*) L 198**: "difference" classifier refers to the domain classifier (pardon the typo). ***e*) Tab. 1**: Detection accuracy measures how often, across a variety of shifts, a statistically-valid detector identifies the shift. ***f*) Fig. 3**: For this dataset, differences between source and target data arise from a different set of angles from which the items were photographed (source: $0^\circ - 175^\circ$, target: $180^\circ - 355^\circ$). ***g*) Typos and reference bugs**: Thanks for the attention to detail. We will make all appropriate changes in the camera-ready if accepted.

Reply to R2:

***a*) L 196**: You are right. This statement entered the paper due to a miscommunication among collaborators. We apologize for the error and will strike it from the camera-ready version if accepted. You are also right that this heuristic is predicated on strong assumptions (addressed below). ***b*) Access to labeled anomalous samples is a strong assumption**: Our intuition here is that while labeling all examples at deployment would be prohibitive but that a small quality team might be used to selectively label those *top anomalous* examples. Such monitoring teams are increasingly common (Google's search quality team is a famous example). ***c*) Related work and other baselines**: We point out that all standard two-sample tests in the input space constitute baselines as does the domain classifier approach. Notably (Tab. 1) MMD on inputs performs worst among all tested methods. ***d*) Distribution of p-values in experiments**: This is a great question. We will investigate the distribution of p-values and add these results to the paper. ***e*) Why is Tab. 1 averaged over MNIST and CIFAR-10**: In this table, we average over all different perturbation models and datasets because our goals was to see if there exists a shift detection method that consistently outperforms the others. We will make this rationale clearer. ***f*) What sort of intervals are shown in the plots**: the intervals around the (mean) p-value are 1-σ error bars. See L 222 - 225 and the caption of Fig. 2 for details. Time steps used for testing are described in L 229. ***g*) Weaknesses, underlying assumptions and failure cases w.r.t. shift characterization and malignancy detection**: The big assumption here is that *obviously* shifted examples are somehow more likely to be misclassified. This assumption might not hold. We will work to assess the correlation of domain classifier confidence with misclassification in real datasets and to formalize settings in which the assumption is reasonable.

Reply to R3:

Thanks for your strong positive review for our paper. We are glad that you appreciated the usefulness of the approach and the lack of pretense. We agree that further experiments with data from other domains and extensions to handle streaming data are promising next steps.