To R1: Thank you very much for the positive comments. We are delighted that you appreciate our work. We'll modify the intro to add some background about Information retrieval.

Our baselines differ for Multi-class vs Multi-label datasets. Hence, the tables may show different baselines each time. We'll organize the plots/tables separately for both sub-sections. Thank you for pointing this out.

Great suggestion on choice of B/R. In theory, since there is a direct connection between count-sketches and MACH, we can work that out. Using Cauchy-Schwarz and Markov's inequality, we can get an $\epsilon - \delta$ relation (accuracy-failure probability trade-off) between R and B which goes like: if $max\ p_i >= \alpha$, then $P(|\hat{p_i} - p_i| < \epsilon) > 1 - \delta$ implies that $RB = \frac{1-\alpha^2}{\delta\epsilon^2}$. Based on our tolerance to ϵ and δ, and the ease of data classification (given by α), we can get an estimate of RB. We'll include a discussion about this in the final version.

Thank you for pointing the typo in 'num_trees, n_epochs ...' legend. We've corrected it to 'Parabel, n_epochs ...'.

To R2: Thank you very much for the positive comments. We are delighted that you appreciate our work.

On a second thought we think that changing title to focus on method is a great suggestion. We'll modify the title (if the PC permits) to reflect our proposed method. We'll also modify the intro to better synergize Extreme Classification and Information Retrieval.

To R3: Thank you for the positive comments. We would make necessary changes in the motivation to reflect better synergy between our method MACH and the task of Information Retrieval. Please see the following clarifications:

- Motivation for IR experiment: Posing Information Retrieval as a classification task is not unconventional. The baseline Parabel compared in this paper is a 1-vs all classifier model with some partial tree structure. It has been deployed on Bing Search Engine and it works really well. Posing IR as classification is known in literature and is at least as old as 2008 Li, Burges and Wu (NIPS 2008) " McRank: Learning to Rank Using Multiple Classification and Gradient Boosting" where they showed that classification loss is naturally a good surrogate for ranking (upper bound for DCG).
- Baselines: We compared against two tried and tested baselines from real search engines. The first one, Parabel is discussed in point 1. The second embedding baseline is deployed on our collaborators search engine. Other publicly available Extreme Classification algorithms are inferior to Parabel (as we see the metrics on Extreme Classification Repository). As mentioned in line 265, we tried running the best publicly known embedding model AnnexML on the same dataset. We varied embedding dimension among 256,512, number of learners among 4,8,16 and number of nearest neighbors among 10,50,100. Even the smallest configuration trained for 5 days without any progress. As noted in line 268, another well performing model SLEEC has a MATLAB code that cannot scale beyond 1M classes (as seen on Extreme Classification Repository).
- ODP vs ImageNet: ImageNet has 22K classes most of which are closely related. There are 1000 standard ImageNet classes and 2-hop an 3-hop classes which are fine-grained versions of original 1000. Closely related classes like different types of birds are prone to spurious prediction probabilities in any general ML algorithm. Further, ImageNet has dense features where accuracy takes a hit with approximations. Also, the larger the number of classes, the more the gains with MACH. We'll add a short discussion about this ImageNet disparity in the paper. We omitted it due to space constraints.