We thank the reviewers for their thoughtful comments. We will incorporate the clarifications provided here in the final paper, add suggested references, and address minor comments such as typos, moving contents between the main paper and supplement (since NeurIPS typically allows an extra page). Citations and line numbers refer back to the paper.

Conceptual Clarifications: [R1] *Meaning of level?* Leaves are level 1, increasing up to the root. [R1] *Will explanations lose fidelity as we move up the tree?* Not always, since closeby train points maybe more realistic than perturbed (local) neighborhoods (see Supp. Figures 3c, 3e, 4b and 4e). [R1] *Weight generation in absence of prior knowledge.* By sorting the labels $f(x_i)$ and setting w_{ij} to 1 whenever $f(x_i)$ and $f(x_j)$ are right next to each other in the sorted list (see lines 218-221). This simple prior (*path graph*) enforces that weights for similar observations must be similar. [R1] *Two step: p, F, and tilda on parameters.* $F \rightarrow$ Frobenius norm, $p \rightarrow$ size of each parameter vector (line 93). Tildas on the parameters differentiate them from MAME solutions. [R1] *SP-LIME relevant because it attempts to provide global explanations by outputting a set of local explanations?* Yes, for comparisons, we control the size of this set to generate same number of explanations as our method for any level. [R2, R3, R4] *Specification of $g()$.* For LIME and our methods, $g()$ is an identity map for numeric and one-hot encoding for categorical features. However, it can be any non-linearity applied on inputs such as squaring of features or multiplying features together to create meaningful interaction terms. [R2] *Why does optimization in Eq. 1 or 2 not involve g? $g()$* is pre-specified and not learned, hence the optimizations do not involve them. [R2] *Regarding α_is: α_i* controls the sparsity of each local (leaf-level) explanation, and are tuned only once to achieve the required individual sparsity for each example. It is not part of main optimization, and does not cause overhead. [R2] *Why not L1 penalty in 3rd term of eq. 1?* L2 is most commonly used in convex clustering literature [21]. We tried L1, but L2 performed better. [R2] *Choice of $\mathcal{N}_i$: $\mathcal{N}_i$* is obtained by randomly perturbing x_i, m times (See lines 98-99 and 215). [R2] *Relation to MAPLE:* We will add more discussion about MAPLE in the paper, however, Two-Step and SP-LIME are closer competitors to our method. MAPLE creates a random forest which itself one might argue cannot be directly interpreted. More importantly though, it does not provide group level explanations as the RF model is learned over the entire dataset. [R3] *Requirement of prior knowledge:* A reasonable prior is important but MAME outperforms other methods when this was chosen in a simple data-driven way. Only in one part of expert study did we incorporate domain knowledge. [R4] *Is example-based grouping a good choice for multi-level?* To explain groups of points in a principled way was our main motivation (like [16, 17]). Pursuing other flavors of multi-level (as in [18]) will be a separate effort. [R4] *Prior knowledge via graph on x or f?* Graphs based on $f(.)$ use the intuition that explanations should be similar if the predictions are similar. It is unclear what the right metric for graphs with (especially a high dimensional) x would be.[R4] *Disconnected prior graph:* This will yield multiple clusters as $\beta \rightarrow \infty$.

Importance of Results:[R4] *Significance of user study:* MAME is statistically significantly better than Two Step and SP-LIME in probability estimate and cluster assignment tasks. Will add ANOVA (for repayment probability MSE, $F(2, 27)=9.48$, $p < .0001$, effect size=0.41) and post-hoc comparison results. [R1, R2, R3] *Regarding dataset results:* With infidelity measure, MAME is better than Two-step in 7/10 cases and overall best in feature importance rank correlation (4/5 cases). With generalized fidelity measure (Supp.), on an average (over all datasets) MAME improves upon SP-LIME and Two Step by 18% and 16% for RF, and 32% and 18% for MLP. We also ran 5-fold CV for the datasets (except ATIS since test partition is pre-defined), averaging over all of them MAME improves upon SP-LIME and Two Step by 31% and 9% for RF, and 118% and 0% for MLP. For MLP, MAME was worse than Two Step only on 1 dataset which hurt the gains. [R4] *Runtime:* We have an efficient implementation in Julia. MAME took ~ 4 hrs on ATIS, where Two-Step took ~ 6 hrs and SP-LIME took ~ 4 days. For other smaller datasets, MAME took < 10 mins.

User study with credit dataset: [R1] *Regarding clusters and guessing them:* The clusters are not the same, but the number is to keep it fair. How well the participants guess the cluster membership shows how homogeneous our clusters are in terms of feature contributions. It also tests whether our explanations are simulatable by the user which is an important metric to judge efficacy of explanations (Lipton 2016). We included the explanation figure for SP-LIME in Figure 8 (supp.) [R1, R3] *Why can users guess outputs better with MAME?* Our method was i) selective, ii) created homogeneous clusters (similar important feature values in each cluster), and iii) was still accurate in terms of the prediction in each cluster. [R4] *Were participants blind to the methods?* Yes, we will clarify this.

Expert study with Oil & Gas: [R1] *Why were there 4 clusters at level 380?* The algorithm successively merges clusters as we go up the tree, and hence resulted in 4 clusters at level 380. [R1] *Semantic relevance of MAME clusters - reason?* MAME is able to ingest prior knowledge effectively and forms clusters whose explanation models are thus not only more intuitive to the expert but also have high fidelity to the black-box model. Two Step does not explicitly control for fidelity which results in less semantically relevant clusters. [R3] *Importance of capturing prior knowledge in explanations:* Behavior of pumps vary widely according to manufacturer, which is why it is important to capture this prior. Remaining useful life estimation of pumps is done separately for each manufacturer.

Quantitative evaluations with public data: [R1] *Comparisons to Pedreschi et al. (2019).* This paper just outlines a high level approach, and does not actually propose an algorithm, so nothing specific to compare against. [R1] *Controlling the amount explained?* In Table 1 (infidelity), the average is computed for all test examples over all levels of the tree/representative explanations to ensure fairness. [R1] *Intuition for infidelity in 4.2?* This measure captures how well our explanation can track changes in black-box prediction when the input goes from a chosen *null* to actual value. Limitation is that this depends on the choice of *null*. See Supp. for another measure (*generalized fidelity*).