

**Performance Evaluation of TPRI Annotators for Image Segmentation using Labelbox**

To measure the quality of training data is a very critical task. Quality is measured using both consistency and accuracy of labelled data. As per the industry standards, the methods used to calculate the training data quality are benchmarks, consensus and review.

**Quality**: Quality contains consistency and accuracy.

**Consistency**: Consistency is the degree to which annotator annotations agree with one another. It prevents random noise by ensuring that labels are correct or incorrect. It is measured through a [consensus](https://docs.labelbox.com/en/quality-assurance/consensus) algorithm. Since labels can be consistently right or consistently wrong, high consistency alone is not the only criteria to explain the whole quality.

**Accuracy**: Accuracy measures how close a label is to the 'Ground Truth'. Ground truth data is a subset of the training data labelled by the knowledge expert or data scientist to test annotator accuracy. Accuracy is measured through [benchmarks](https://docs.labelbox.com/en/quality-assurance/benchmarks). Benchmarks is the one which is enabled by the data scientists to monitor the overall quality of the training data and then troubleshoot any potential fallouts in quality by providing insight into the accuracy of the annotators job.

[**Review**](https://docs.labelbox.com/en/quality-assurance/review-labels): Review is a method to ensure good accuracy. Once the labels have been completed, an expert review the label accuracy. It is normally conducted by spot checking labels. It is used to identify low-accuracy and inconsistencies in the labelling process but benchmarks are used to get a pulse of annotator’s performance.

Benchmarks tends to be the cheapest quality assurance option since it involves the least amount of overlapping work. It only captures a subset of the training data set.

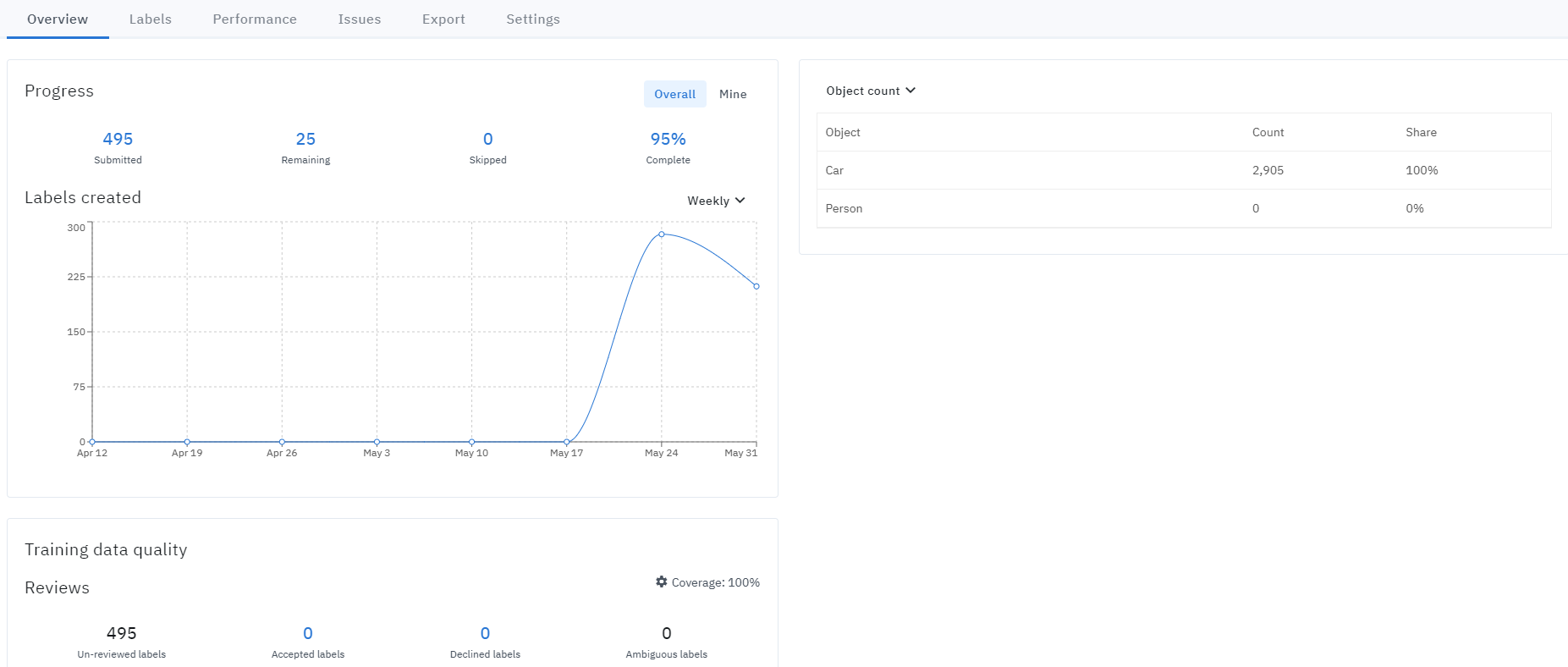
**Quality Assurance**: Quality assurance is an automated process that operates continuously throughout the development of the training data. With the Labelbox consensus and benchmark features, you can automate consistency and accuracy tests. These tests allow you to customize the percentage of your data to test and the number of annotators that will annotate the test data.

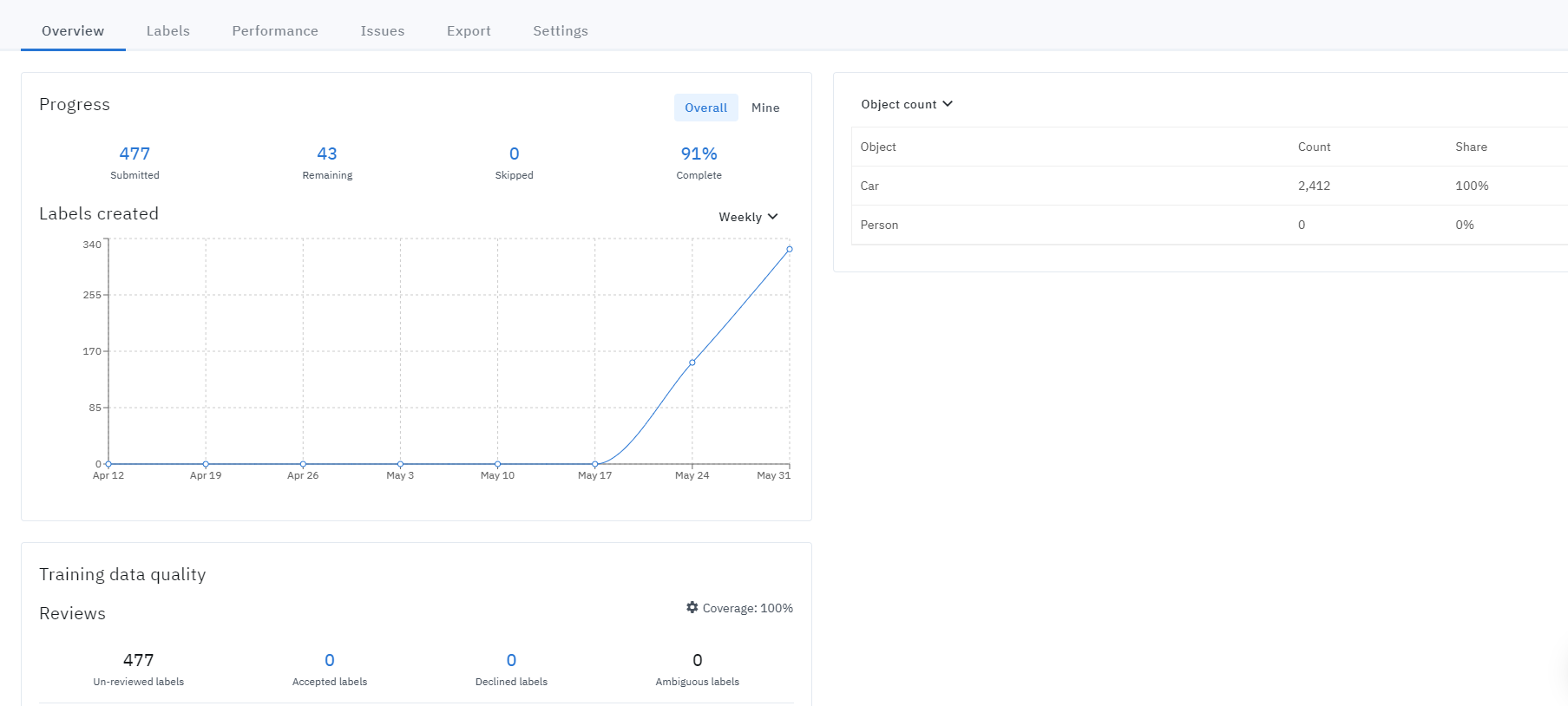
As the annotators are randomly benchmarked, you can monitor project quality with the overall quality graph.

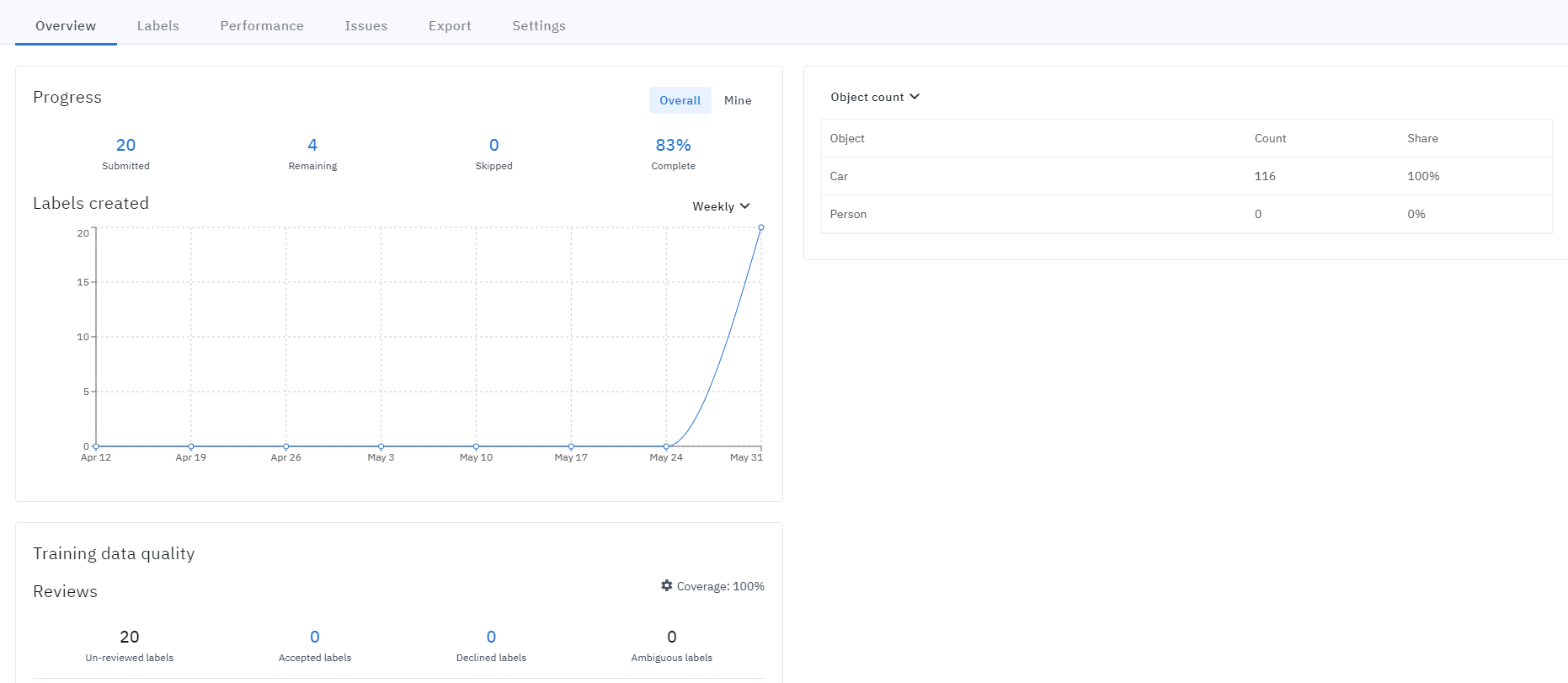
**Benchmarks**:

**Benchmark’s workflow**:

1. Create a new Benchmark by starring an existing label.
2. Automatically annotators get benchmarked at random intervals.
3. Keep track of your project’s overall quality and look into any deviation, either by annotator or by benchmark.

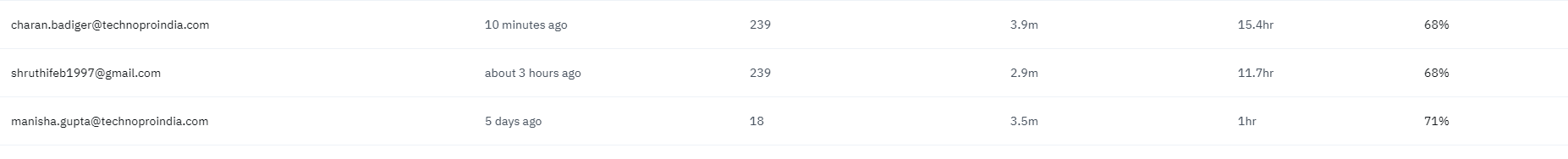
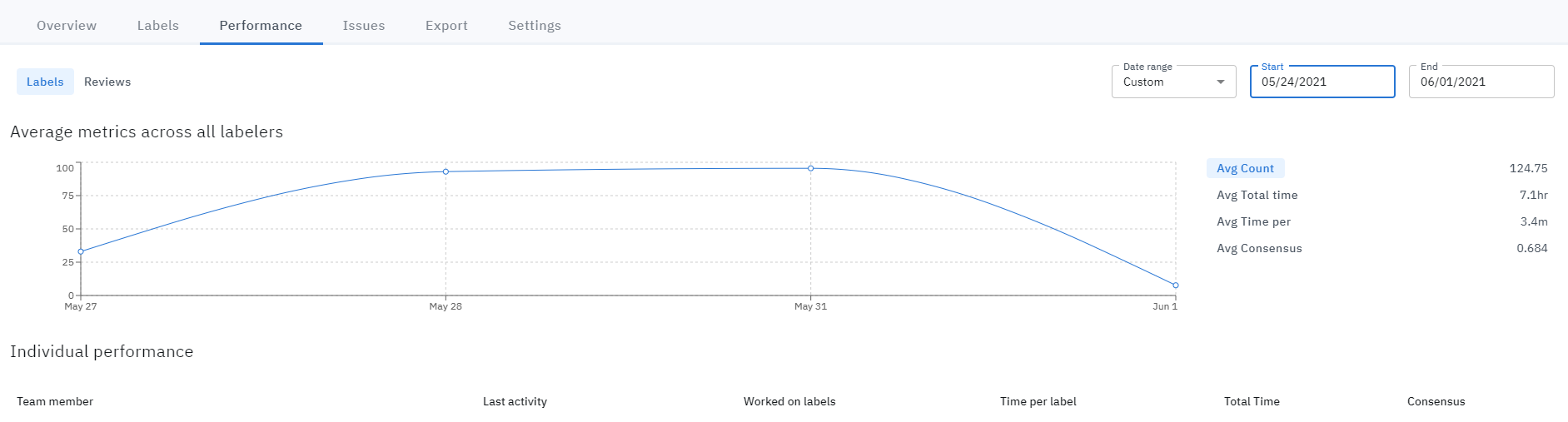


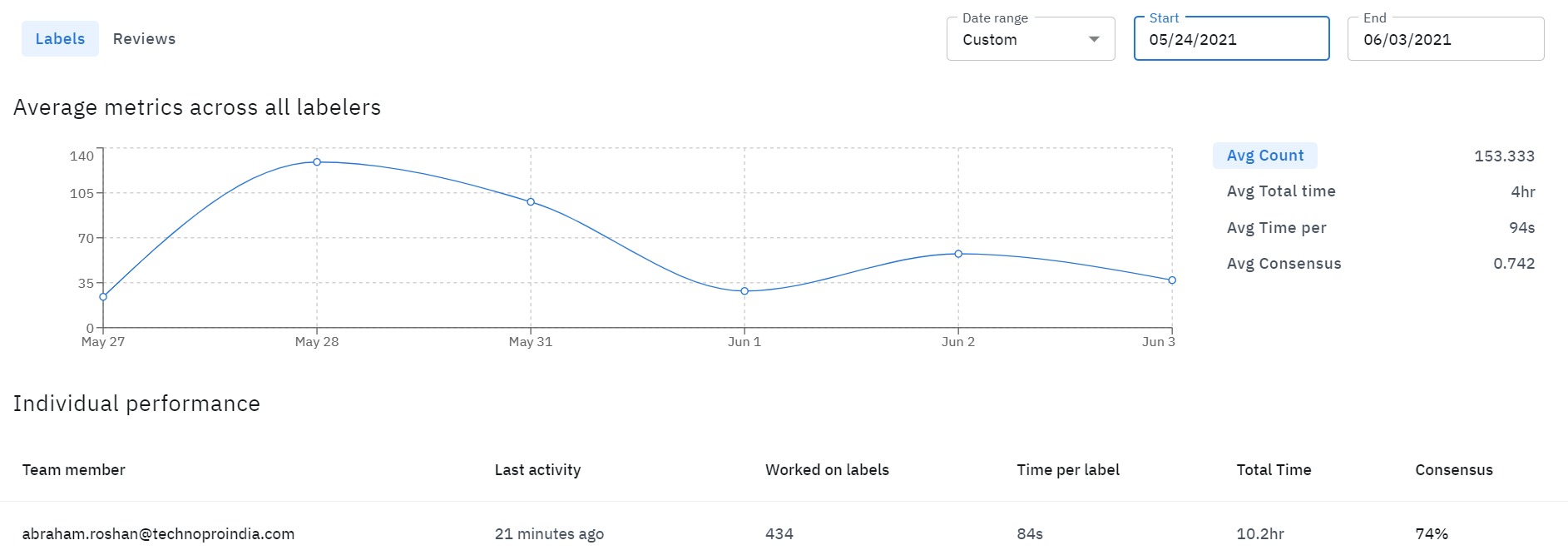


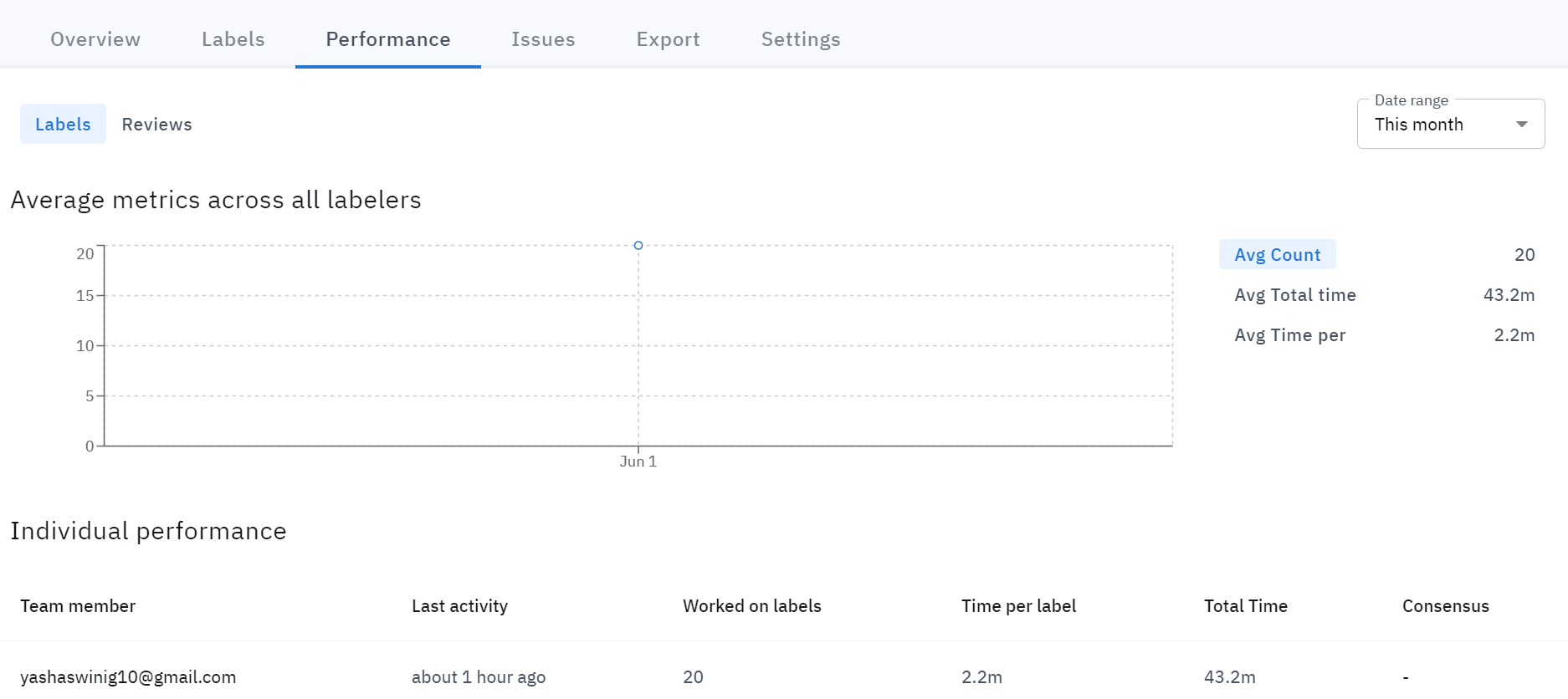


**Figure 1: Graph of TPRI Annotators progress and training data quality**

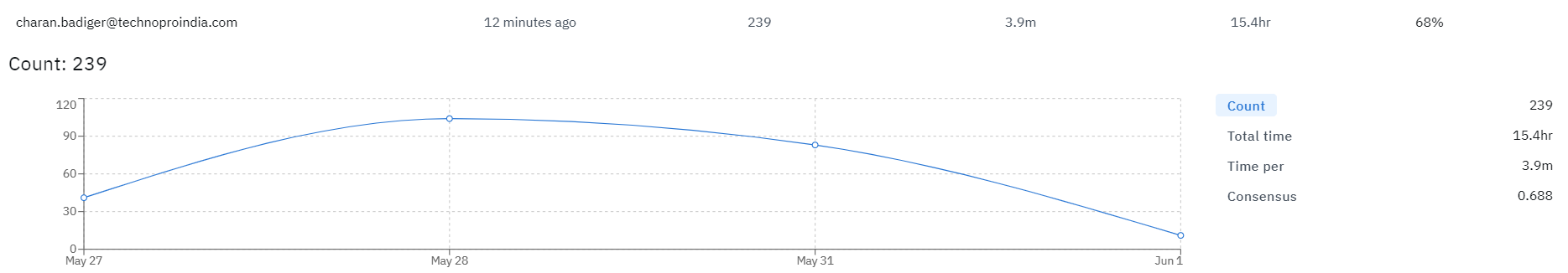
To troubleshoot drops in quality, we can explore performance by annotator or benchmark.

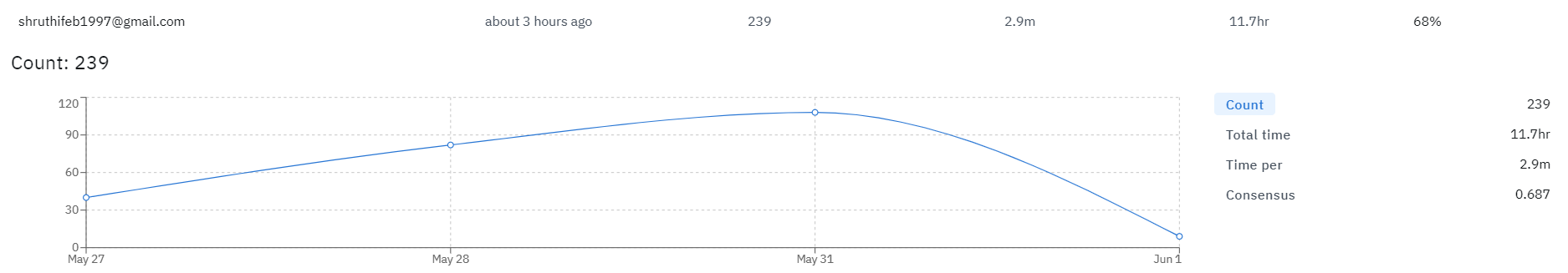


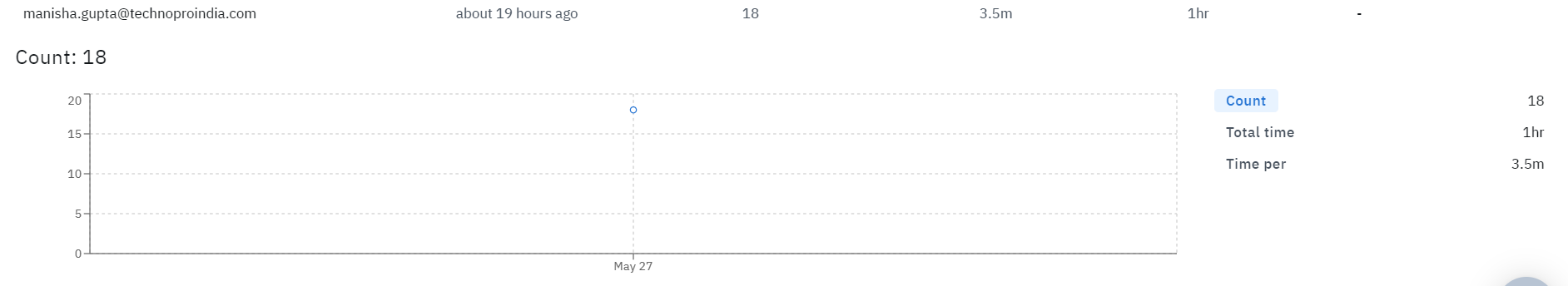
 

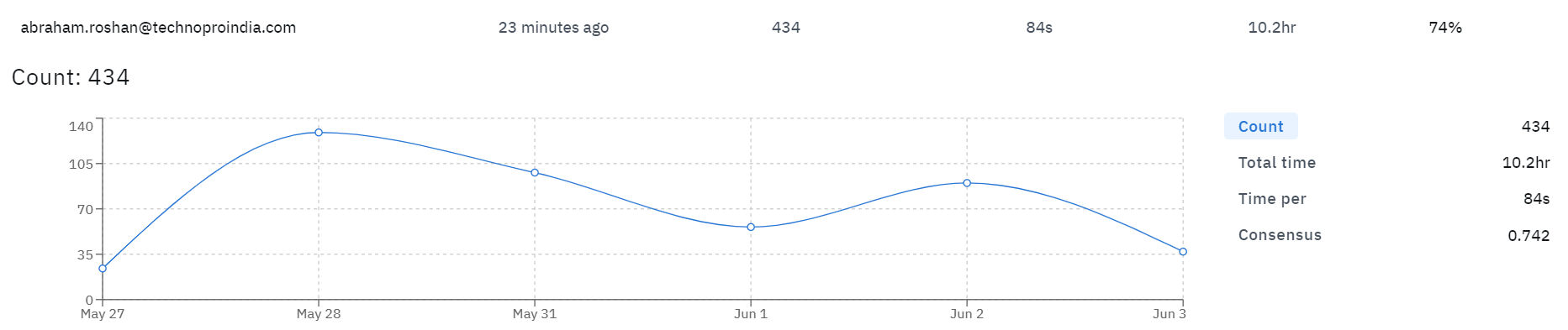


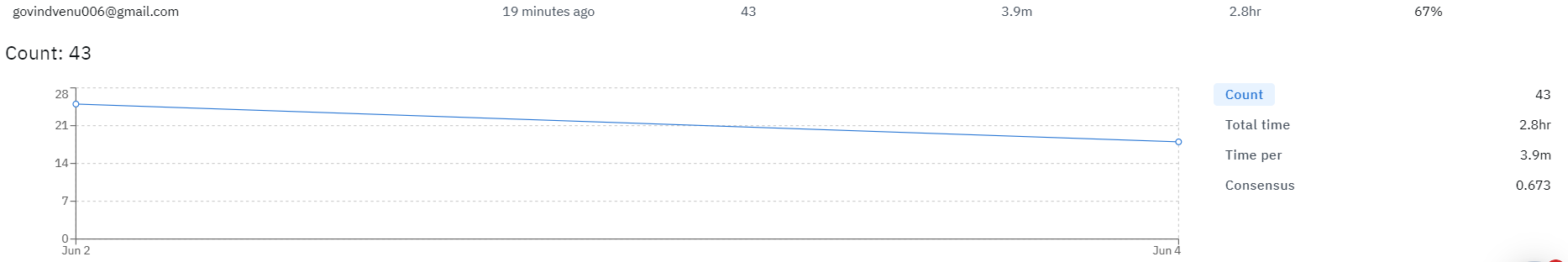
**Figure 2: Graph of the performance of various TPRI Annotators in a team**

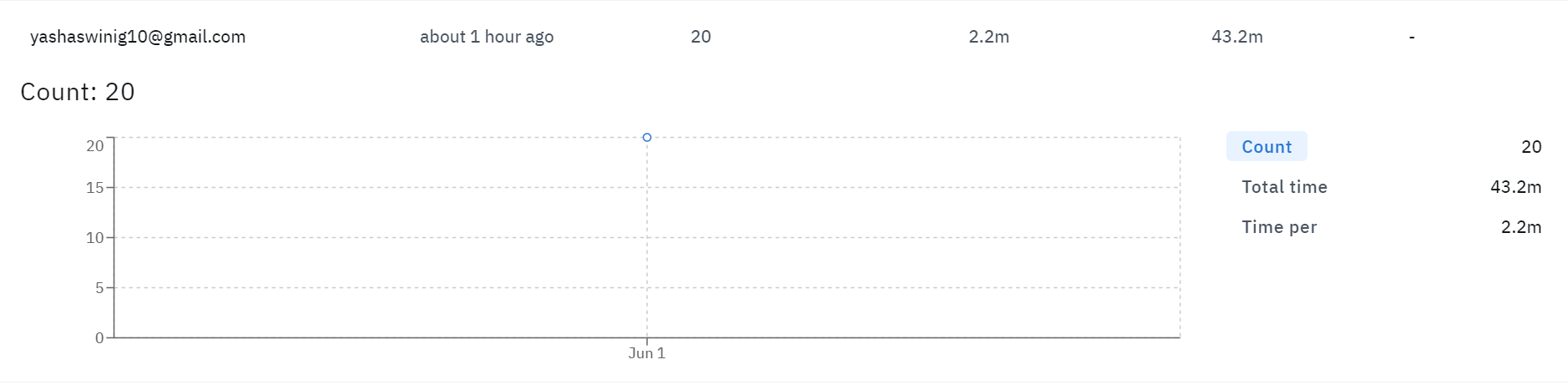












**Figure 3: Graphs of individual performance of each TPRI Annotator**

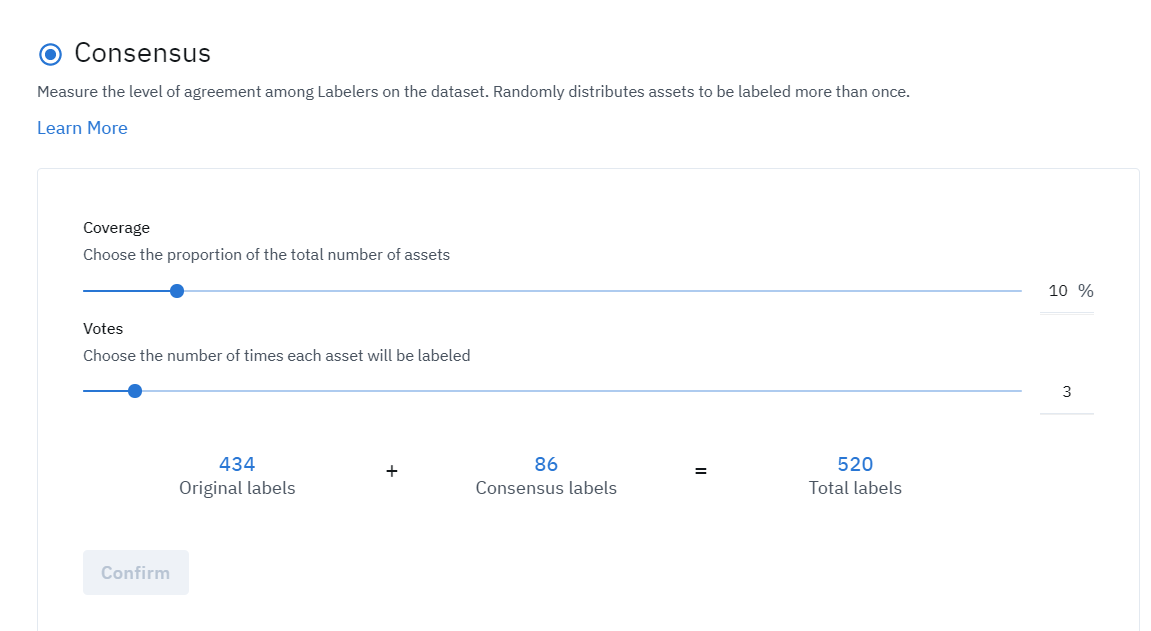
Systemic poor annotator performance is the indication of poor instructions, while poor performance on certain part of data is the indication of edge cases. Data scientists use these values to help them improve annotator on-boarding and education processes.

**Consensus**: [Consensus](https://docs.labelbox.com/en/quality-assurance/consensus) measures the rate of agreement between multiple annotators (human or machine). A consensus score is calculated by dividing the sum of agreeing labels by the total number of labels per asset.

**Consensus Workflow**:

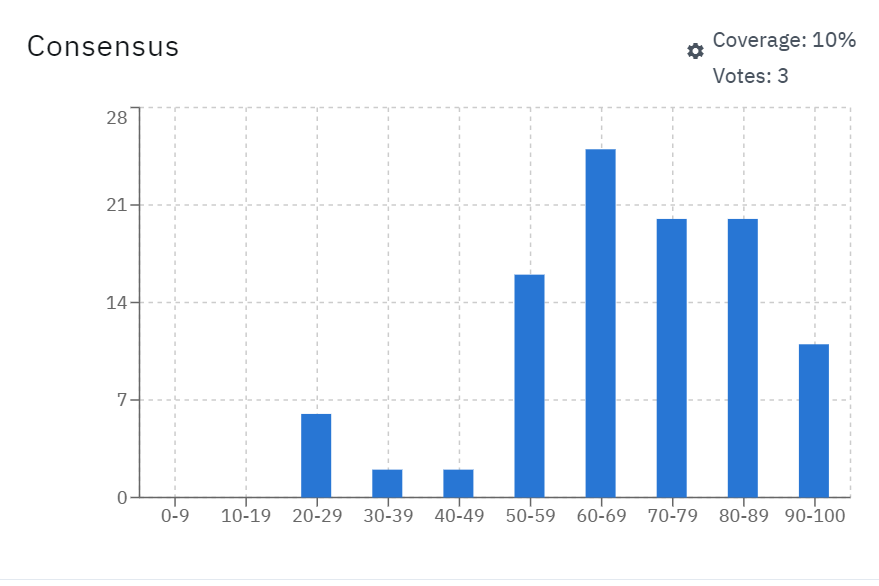
1. Enable consensus and customize the consensus parameters.
2. Automatic random labels are distributed across annotators at random intervals.
3. Keep the track of the overall consistency and investigate any fallouts in quality by examining into individual annotator and label consensus scores.

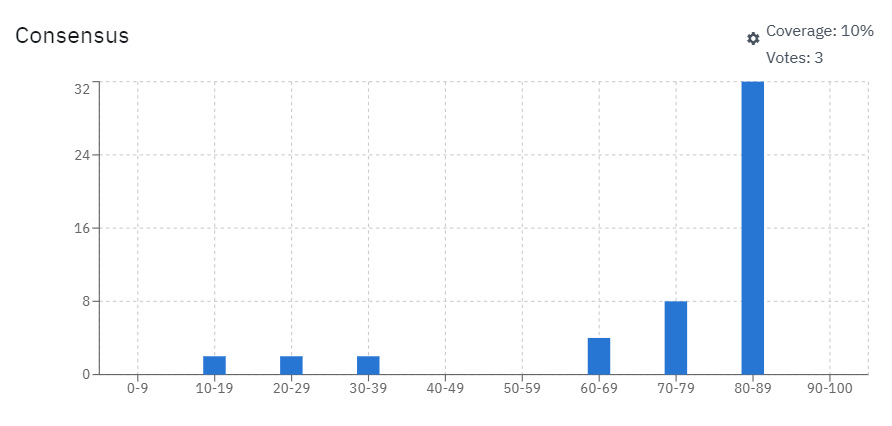
To configure consensus, you can customize the percentage of training data and the number of annotators to test.



**Figure 4: Consensus Workflow**

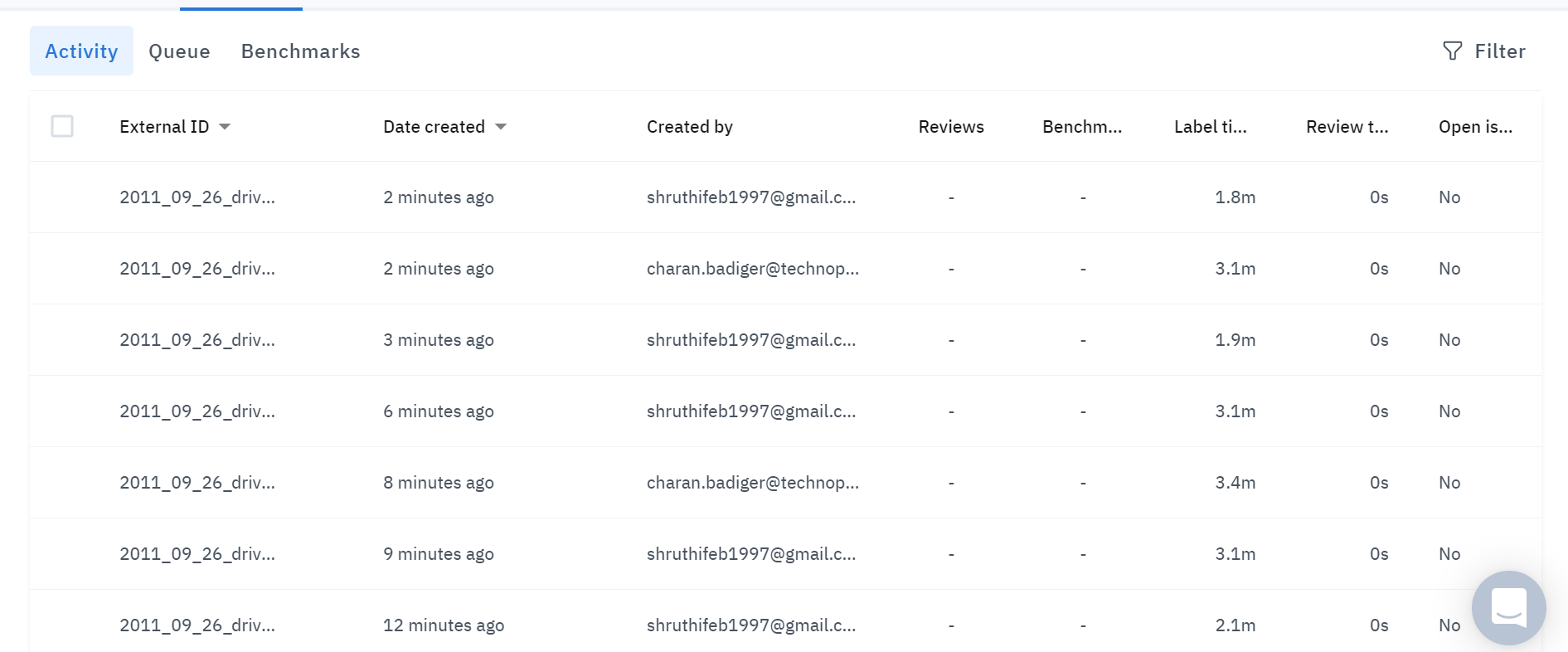
Monitor overall consistency with the Consensus histogram.





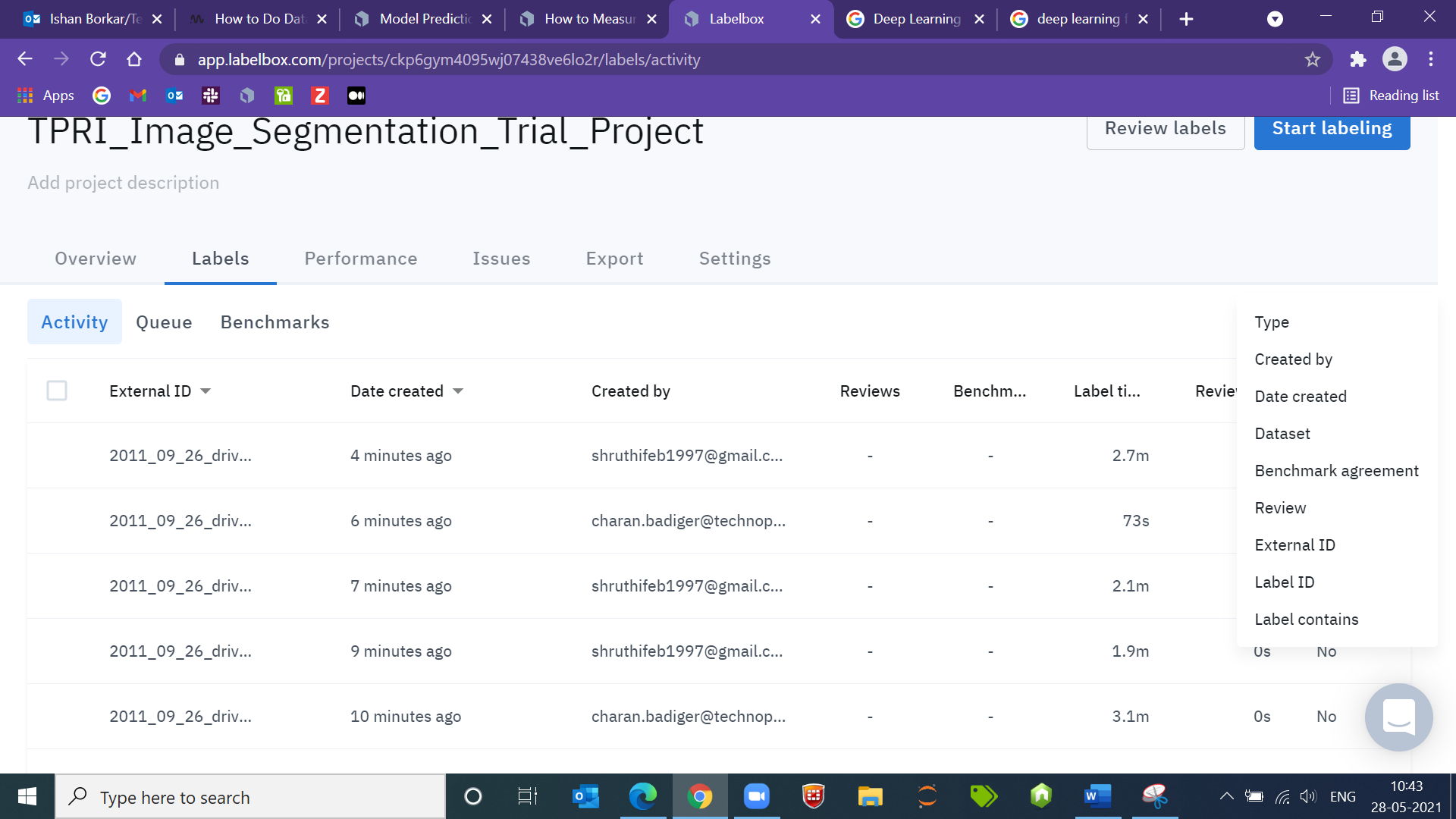
**Figure 5: Consensus Histogram to monitor consistency of the Annotators**

Breakdown the consensus score by asset. We can compare the labels of a particular image.



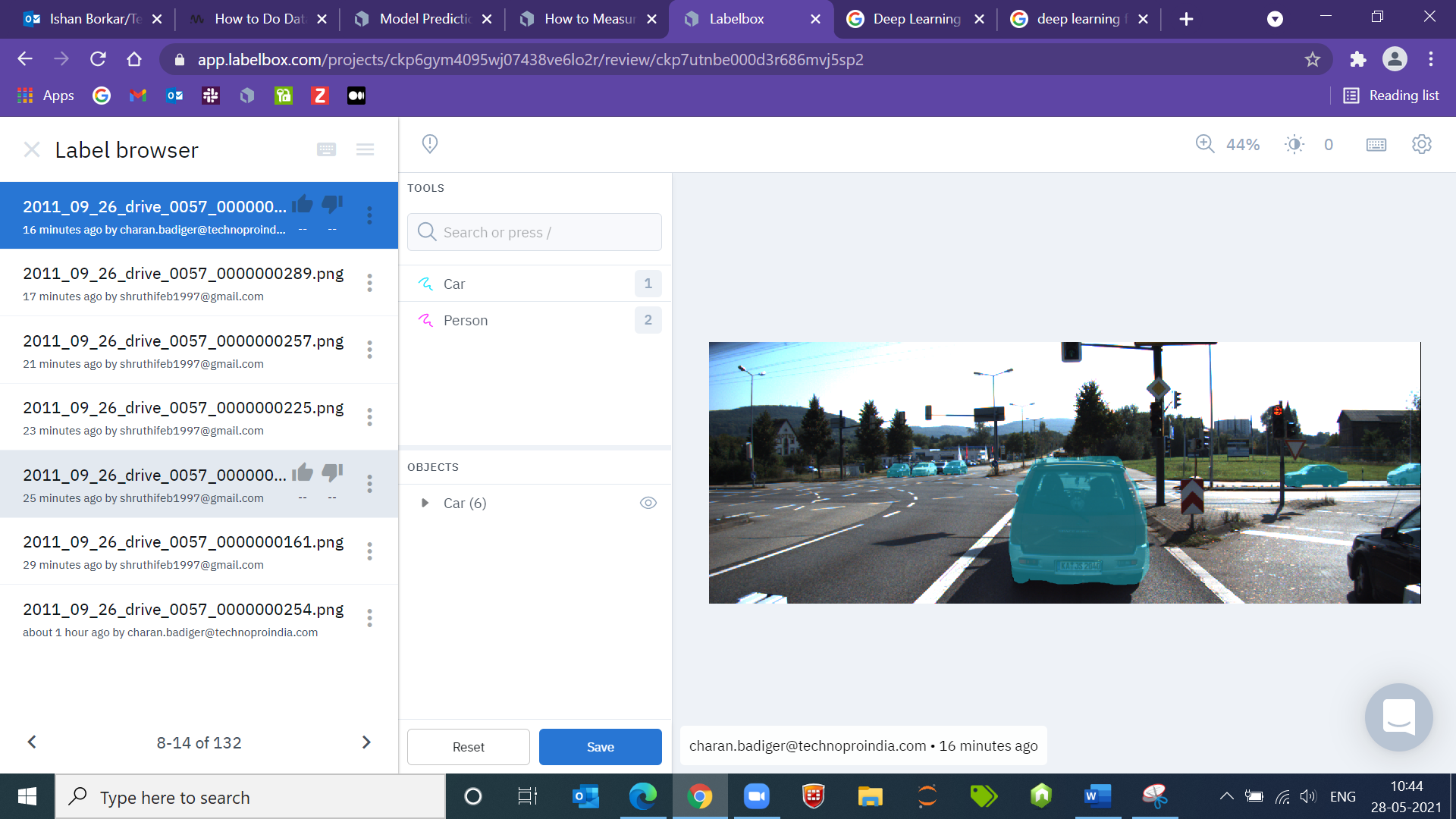
**Figure 6: Comparing the labels of a particular image**

**Review**: Reviewing is a manual process. It is part of keeping a human in the loop. Choose which labels to review. Review, modify or re-enqueue labels. Filtering options help managers prioritize which labels to review. As shown in the image below, available filters include annotator, consensus score, label.



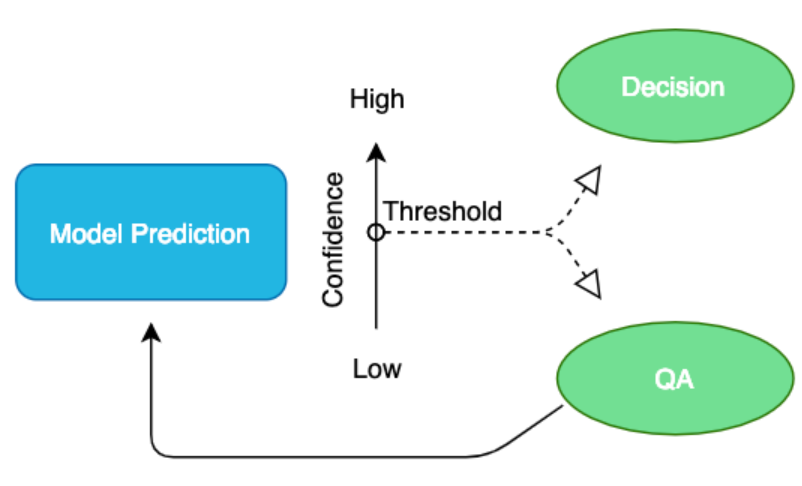
**Figure 7: Filtering options to prioritize the labels for review**

Reviewers are the top performing annotators or internal knowledge experts. To review a label, there are thumbs up and down icons. The reviewer has the option to modify or correct the labels on the spot. Clicking on the three vertical dots allows you to delete and re-enqueue the label, view the benchmark when applicable, and copy the link to send to other collaborators.



**Figure 8: Modify or correct the label option**

**Quality Assurance Prediction Model:**



**Figure 9: Prediction Model for Quality Assurance**

Three basic steps of machine learning: data collection, training and deployment. They operate simultaneously and interactively to form a complex workflow. Once our model is deployed, it still needs to be maintained and updated. Each prediction that a model makes is having a confidence score. We can set a confidence score threshold to determine how to treat a model’s prediction in production.

A prediction with a confidence score above the threshold will operate as a decision without human intervention while a prediction below the confidence threshold will go through a quality assurance. The training data, the predictions and quality assurance data are rendered visually.

**Conclusion**: Creating training data is most tedious task of building a machine learning application. Careful monitoring training data quality increases the chances of having the best model the first time. Getting labels right for the first time is far cheaper than the cost of discovering and reworking to fix the problem.

With Quality Assurance processes data scientists can:

1. Monitor overall consistency and accuracy of the training data.
2. Quickly troubleshoot quality errors.
3. Improve annotator instructions, on-boarding, and training.
4. Better understand the specifications to their project on what and how to label a particular object.

We can also conclude that the time taken by the annotator to annotate an object in the image using image segmentation is more than the time taken by the annotator to annotate the same using the bounding box, polygon etc. since we annotate it with a regular figure in front of the object. Also with increasing training data, time per label gets reduced.

**Reference**:

1. <https://labelbox.com/blog/how-to-measure-quality-when-training-machine-learning-models/>
2. <https://labelbox.com/blog/model-predictions-semi-automatic-labeling-and-quality-assurance-in-production/>