# Bicycle Project Crowd Evaluation

Jens Beyermann

September 6, 2021

# The Annotators

# Task 1

Basic data exploration.

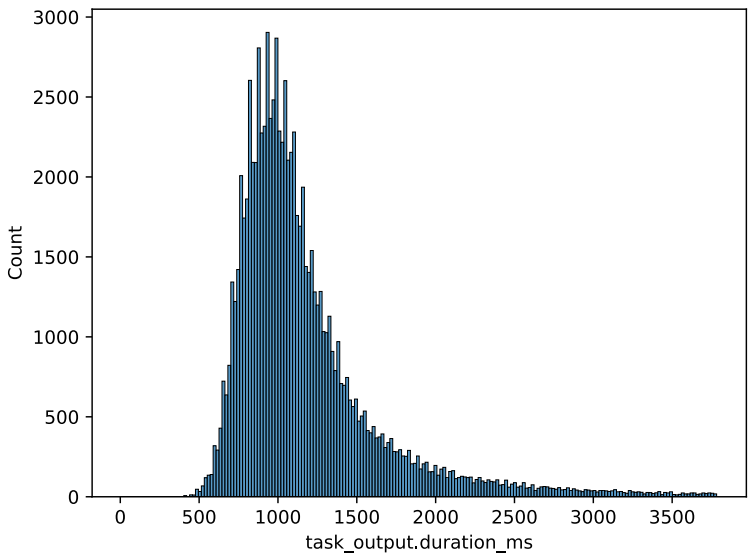
- 9087 images
- 90870 annotations
- 22 annotators
- 1 vendor

# Task 1b

What are the average, min, max annotation times?

- Without sanitization, the annotation times are between $-99999$ms and 40000ms.
- Basic sanitization: Exclude all values $\leq 0$ and a few very large ones.

$\Rightarrow$ Obtain a range from 10ms to 3782ms, with an average of 1176.11ms.



# Task 1b

What are the average, min, max annotation times?

- Without sanitization, the annotation times are between $-99999$ms and 40000ms.
- Basic sanitization: Exclude all values $\leq 0$ and a few very large ones.

$\Rightarrow$ Obtain a range from 10ms to 3782ms, with an average of 1176.11ms.

The lower bound is still inhumanly fast with human reaction times to simple stimuli being usually between 160ms and 190ms, without any decision to make.[**?**] But since the number of annotations with $< 100$ms is 4 this is probably not a sign of some annotators corrupting the process by using a bot for their work. Probably they just accidentally double clicked in the annotation tool.

# Task 1b

What are the average, min, max annotation times?

- Without sanitization, the annotation times are between $-99999$ms and 40000ms.
- Basic sanitization: Exclude all values $\leq 0$ and a few very large ones.

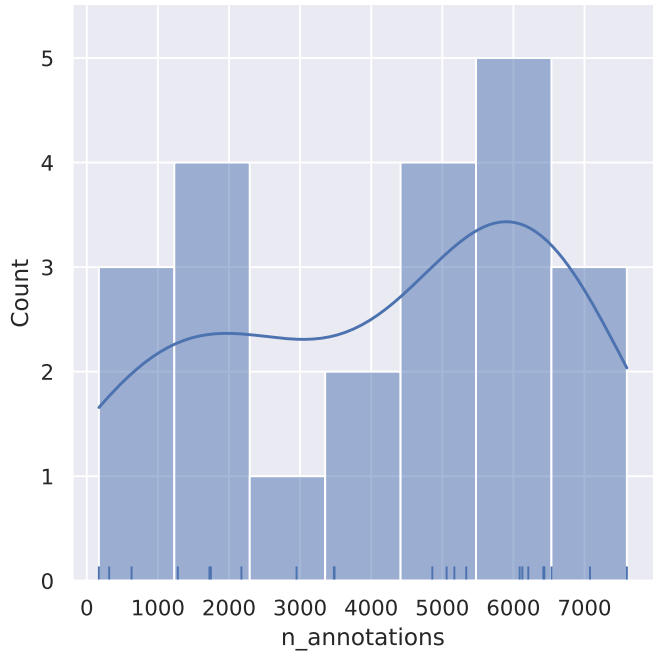
$\Rightarrow$ Obtain a range from 10ms to 3782ms, with an average of 1176.11ms.

**Note**

Technically the observed pattern could be the result of an annotation bot, sampling its own reaction times from a reasonable poisson distribution, but it seems unlikely.

# Task 1c

Did all annotators produce the same amount of results, or are there differences?




⇒ The annotators were involved in very different amounts of annotations, ranging from 170 to 7569 samples per annotator.

# Task 1d

Are There questions on which the annotators highly disagree.

| answer | False | True | invalid | prediction |
|---|---|---|---|---|
| img_id | | | | |
| **img_0000** | 7 | 3 | 0 | False |
| **img_0009** | 6 | 4 | 0 | False |
| **img_0125** | 7 | 3 | 0 | False |
| **img_0126** | 6 | 4 | 0 | False |
| **img_0142** | 6 | 4 | 0 | False |
| **...** | ... | ... | ... | ... |
| **img_8928** | 3 | 7 | 0 | True |
| **img_8942** | 7 | 3 | 0 | False |
| **img_8975** | 6 | 4 | 0 | False |
| **img_9037** | 3 | 7 | 0 | True |
| **img_9054** | 4 | 6 | 0 | True |

385 rows × 4 columns

| answer | False | True | invalid | prediction |
|---|---|---|---|---|
| img_id | | | | |
| **img_0009** | 6 | 4 | 0 | False |
| **img_0126** | 6 | 4 | 0 | False |
| **img_0142** | 6 | 4 | 0 | False |
| **img_0229** | 6 | 4 | 0 | False |
| **img_0341** | 6 | 4 | 0 | False |
| **...** | ... | ... | ... | ... |
| **img_8833** | 6 | 4 | 0 | False |
| **img_8886** | 6 | 4 | 0 | False |
| **img_8894** | 5 | 5 | 0 | False |
| **img_8975** | 6 | 4 | 0 | False |
| **img_9054** | 4 | 6 | 0 | True |

197 rows × 4 columns

As one can see above there are 385 images which are somewhat difficult to classify and 197 ones that are considered as highly difficult.

# Task 2

**Task**

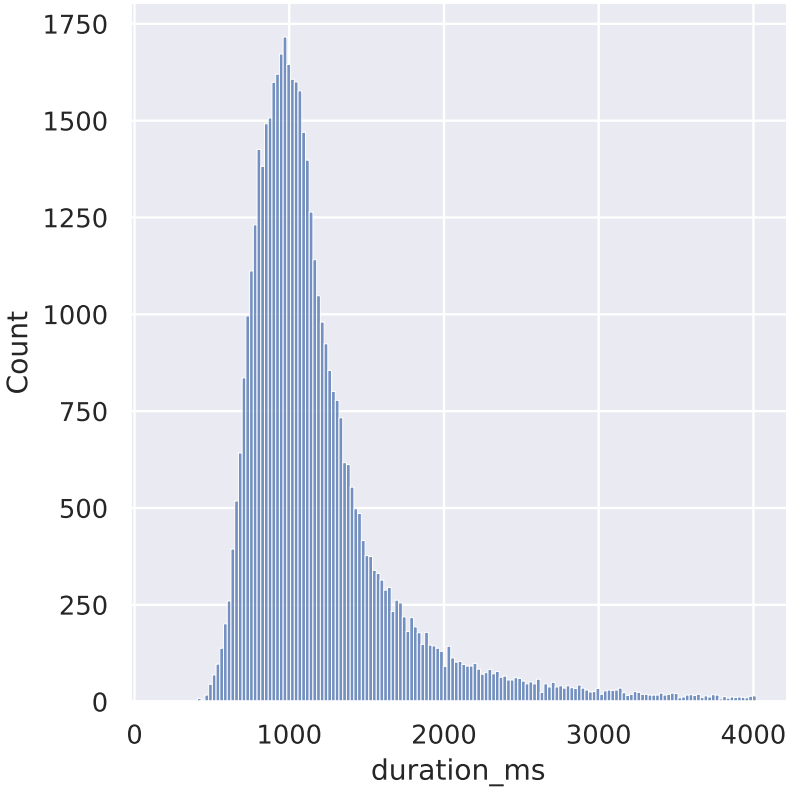
How often does ”cant_solve” and ”corrupt_data” occur in the project respectively?

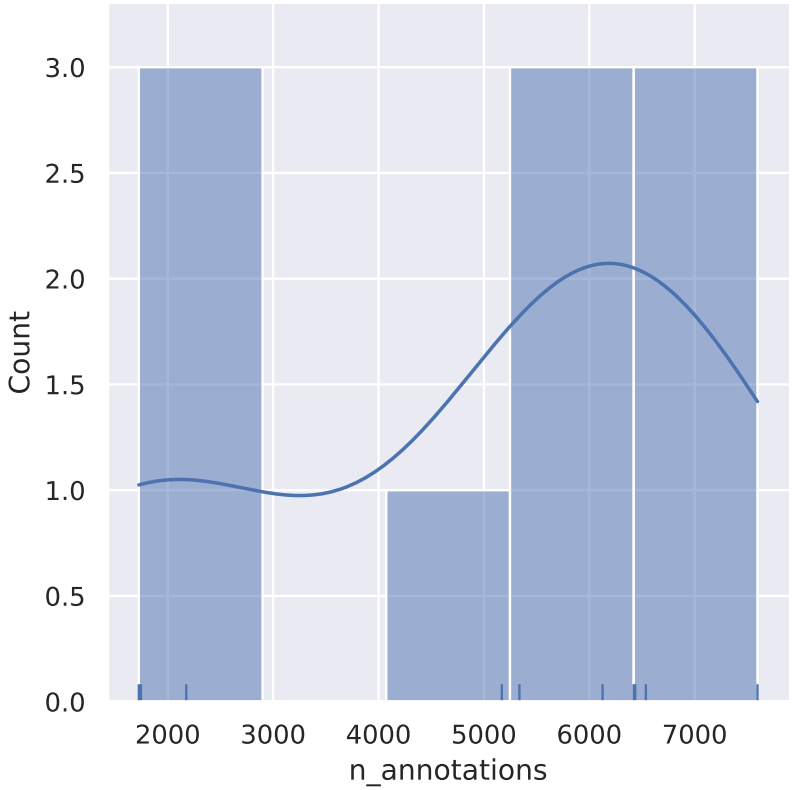
There are:

- 17 occurrences of a task marked as ”cant_solve”
- 4 occurrences of a task marked as ”corrupt_data”
- 10 annotators that made use of those tags.

# Task 2

Can we identify any trends within the annotators using one of the invalid flags?

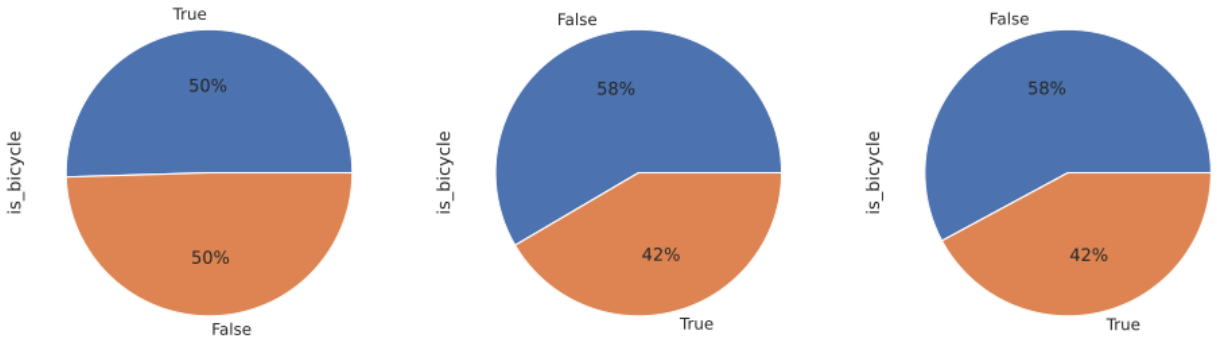



Unfortunately I could not identify any trend here the above distributions seem to represent the overall distributions in annotation times and task numbers pretty well.

# Task 3

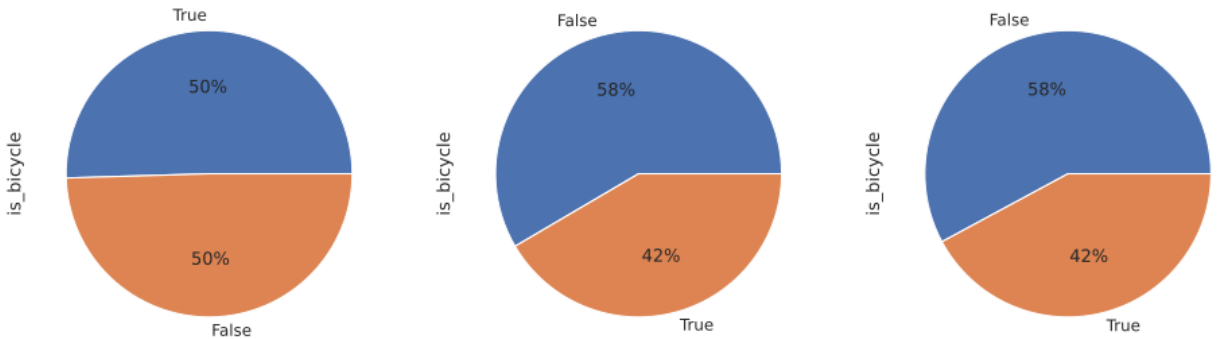
Is the dataset balanced?




As we can see above the dataset is pretty balanced as a whole dataset with nearly exactly as much True cases as False cases. If one restricts the dataset to the more difficult cases, it is not as perfect. But considering that we only talk about 385 difficult and 197 highly difficult cases out of 9087 samples the distribution of 58% to 42% does not seem too bad.

# Task 3

Is the dataset balanced?




Without any further investigation into what does actually make up the difficult cases, I have to conclude that the dataset is reasonably balanced.

# Task 4

**Task**
Identify good annotators.

To complete this task I start with the accuracy, precision and recall of each annotator on the whole dataset.

# Task 4

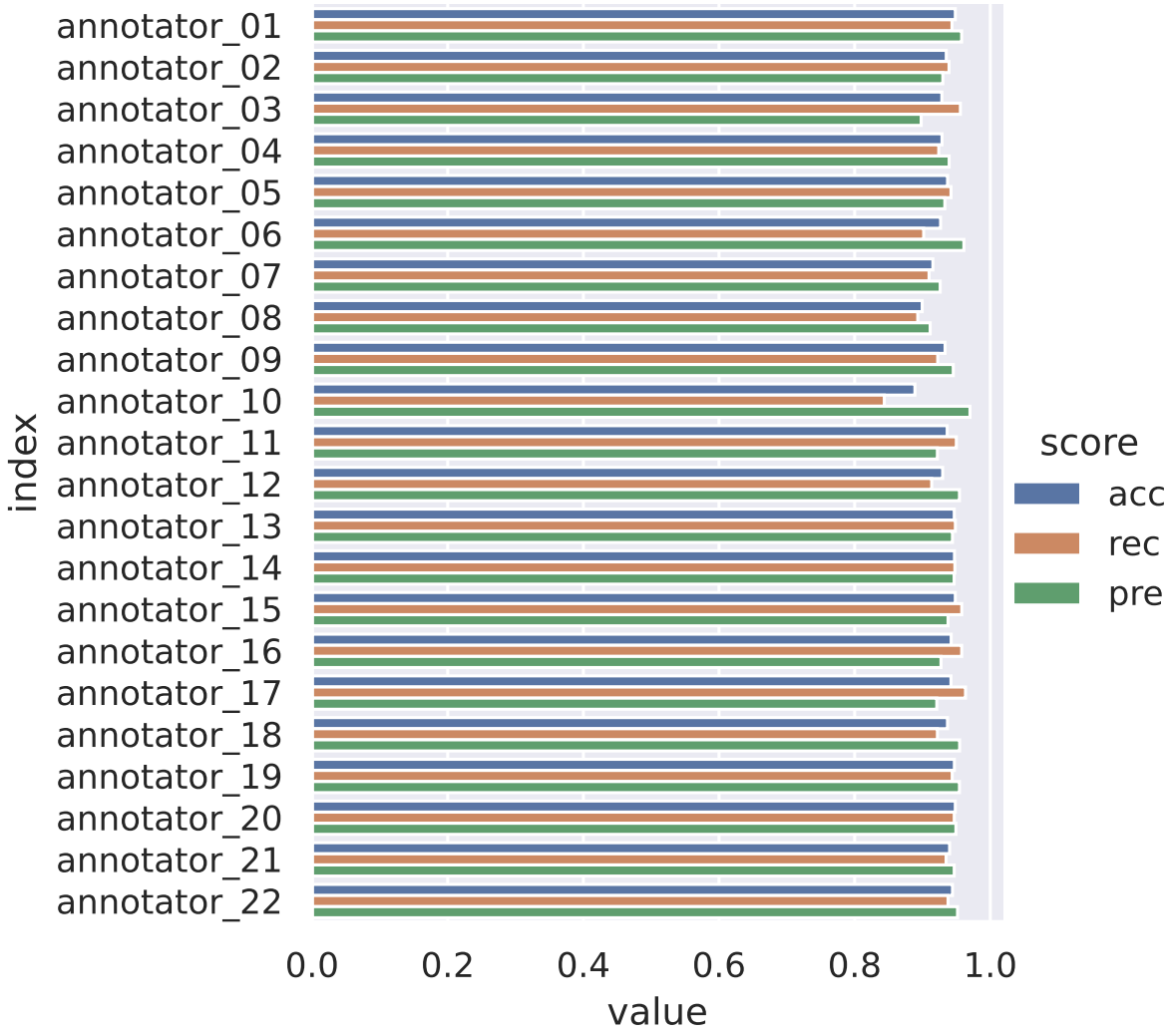



**Figure 1:** accuracy, precision and recall on the whole dataset.

# Task 4

Aside from a few exceptions, most of them achieve very similar results. To be able to differentiate better I focus now on the more difficult tasks.

# Task 4

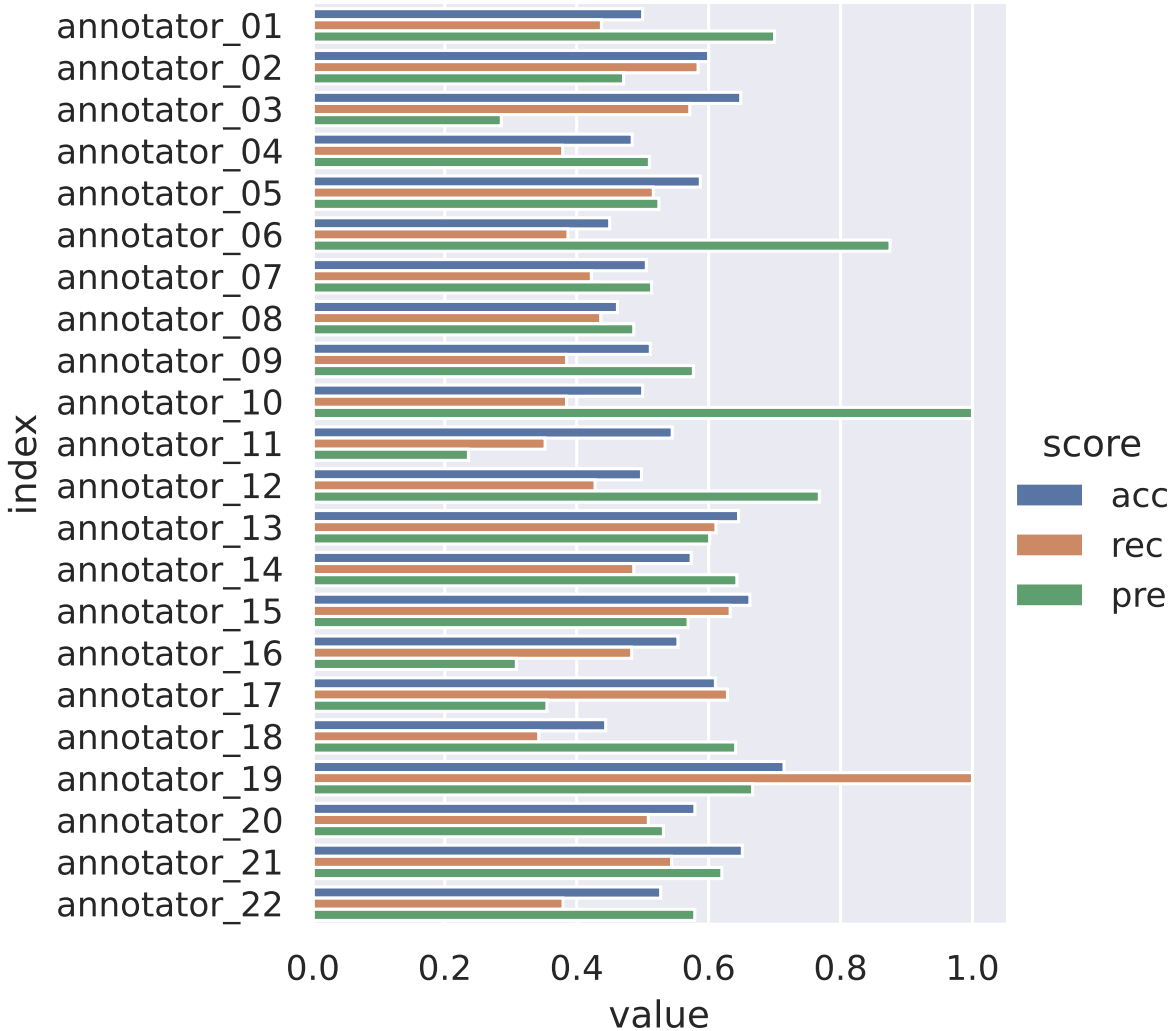



**Figure 2:** accuracy, precision and recall on the somewhat difficult cases.

# Task 4

There are a number of possibilities now, to choose good annotators from this information. E.g. one could build a new weighted score from accuracy, precision and recall but the actual weight would depend on the concrete application. On a self-driving car we would want to make sure to hit all true positives, to avoid hitting bikers with a car, even if we have to sacrifice some accuracy for this. For a traffic surveillance system that wants to estimate the number of bikers passing e.g. a bridge per day we would probably rate overall accuracy highest.

# Task 4

As a very rudimentary measure to identify good annotators I choose those annotators that belong at least to the best 50% in all scores.

| | acc | rec | pre |
|---|---|---|---|
| **annotator_01** | 0.500000 | 0.437500 | 0.700000 |
| **annotator_14** | 0.573529 | 0.486486 | 0.642857 |
| **annotator_19** | 0.714286 | 1.000000 | 0.666667 |
| **annotator_20** | 0.579151 | 0.508621 | 0.531532 |

**Figure 3:** Best performing annotators in the more difficult cases.

| | acc | rec | pre |
|---|---|---|---|
| **annotator_01** | 0.948438 | 0.943536 | 0.957768 |
| **annotator_14** | 0.947216 | 0.947796 | 0.946698 |
| **annotator_19** | 0.947059 | 0.943820 | 0.954545 |
| **annotator_20** | 0.948408 | 0.946850 | 0.949342 |

**Figure 4:** Best performing annotators overall.