Designing incentives for crowdworker collection of a ground-truth dataset for use in score-image annotation tasks

Eamonn Bell
Ph.D. Candidate
Department of Music
Columbia University

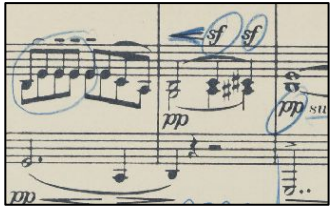
epb2125@columbia.edu
www.columbia.edu/~epb2125

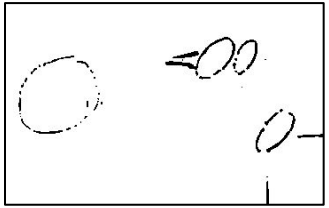
SIMSSA workshop XVII
December 1, 2018 @ McGill

Task

Given an image of a musical score, identify the pixels corresponding to handwritten annotations

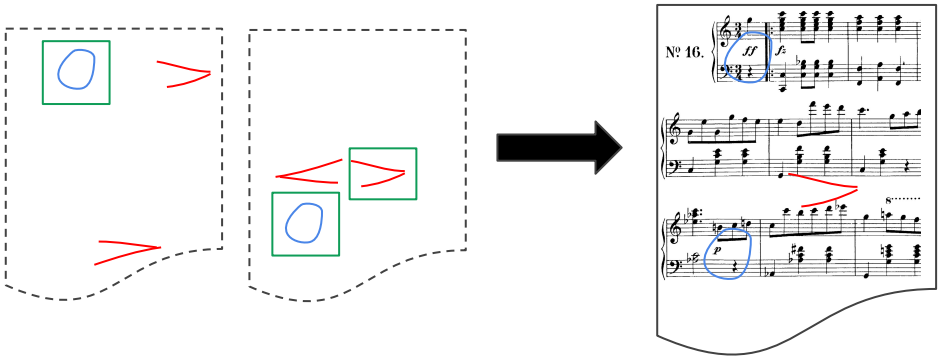
(pixel-level semantic segmentation of images)





Why bother?

- Vectorize annotations and reconcile with MEI-encoded scores as SVG annotations
- MEI-encoded scores can be used to interactively and dynamically visualize different annotation sets
- Preparation of real or virtual performances informed by conductor annotations
- Extracted annotations could be grouped by type using existing shape classification techniques.
- Steps toward authorship attribution in multi-annotator scores



One approach

Supervised machine learning

(e.g. classical classifiers on features, deep CNNs)

dCNNs promising results in other OMR applications

...**requires ground truth**

...laborious to collect/ (semi)expert task

Annotations predicted by RaF classifier trained on GT from different page, same volume and marking artist

Annotations predicted by RaF classifier trained on GT from different page, different volume and different marking artist

One approach

Supervised machine learning

classical classifiers

Bell and Pugin, 2018

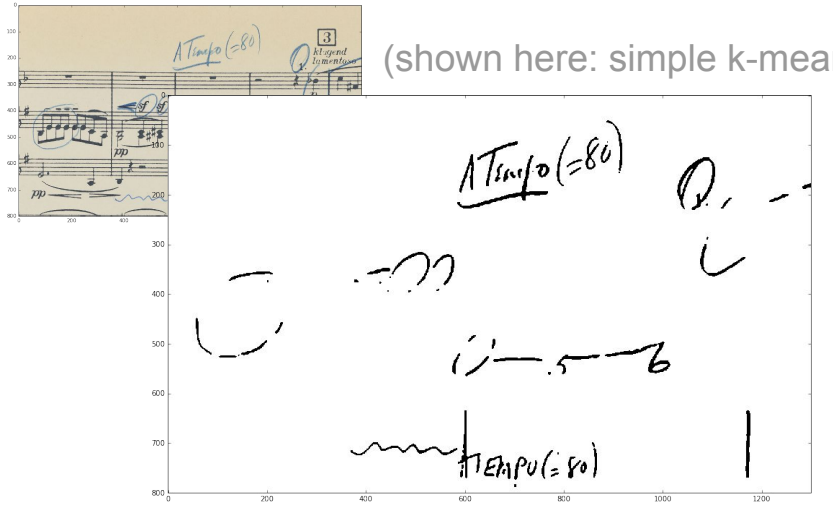
Another approach

Unsupervised machine learning

Doesn’t require labeled data

Another approach

Unsupervised machine learning



(shown here: simple k-means clustering in colorspace)

Doesn’t require labeled data

...but not the full picture (see left)

Bell and Pugin, 2016

Today’s idea

Use unsupervised* approaches to speed up **ground truth** collection

ideas/feedback

ideas/feedback

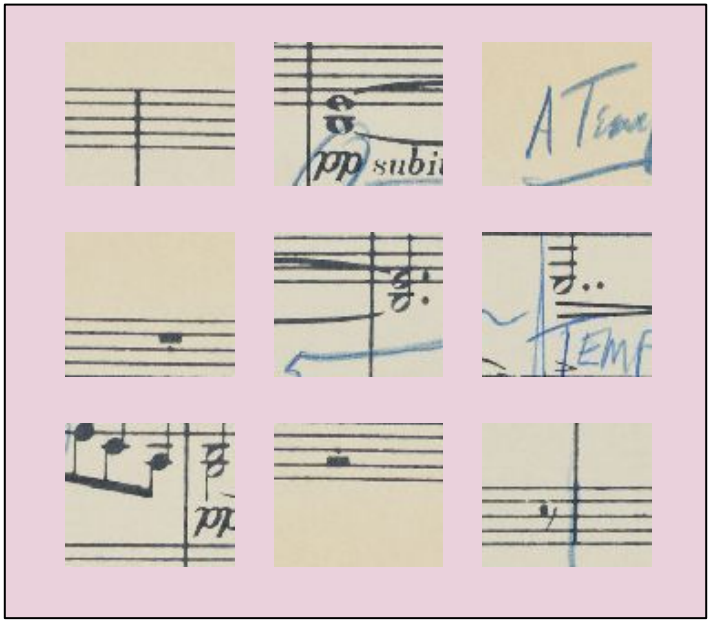
Today’s idea

Use unsupervised* approaches to speed up **ground truth** collection

*we can also use image alignment and comparison to recover annotations by subtracting aligned copies. basically this means anything that can be done cheaper/faster than collecting class labels

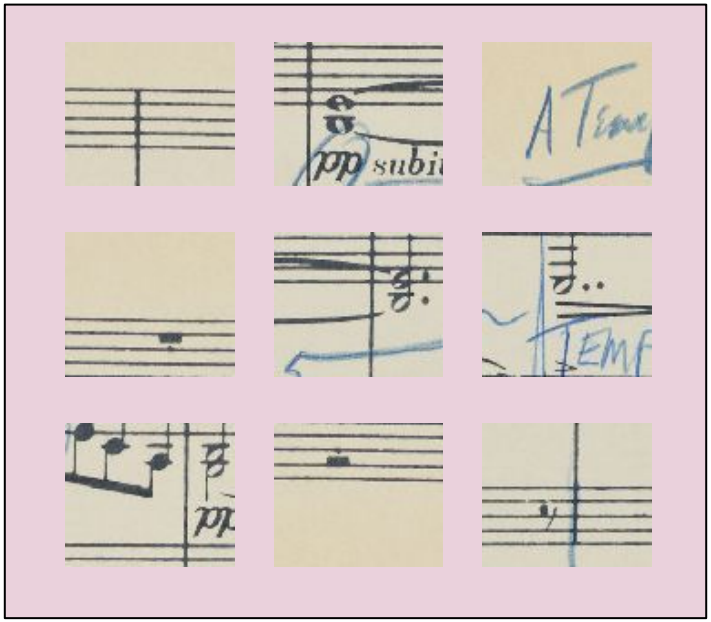
Application 1: Screening tasks

Select all fragments that do **not** contain handwriting



Application 1: Screening tasks

Select all fragments that do **not** contain handwriting



We know (from cheap method):

- probability under the model that tile contains annotation

We ask:

- does tile contain annotation?

Applications:

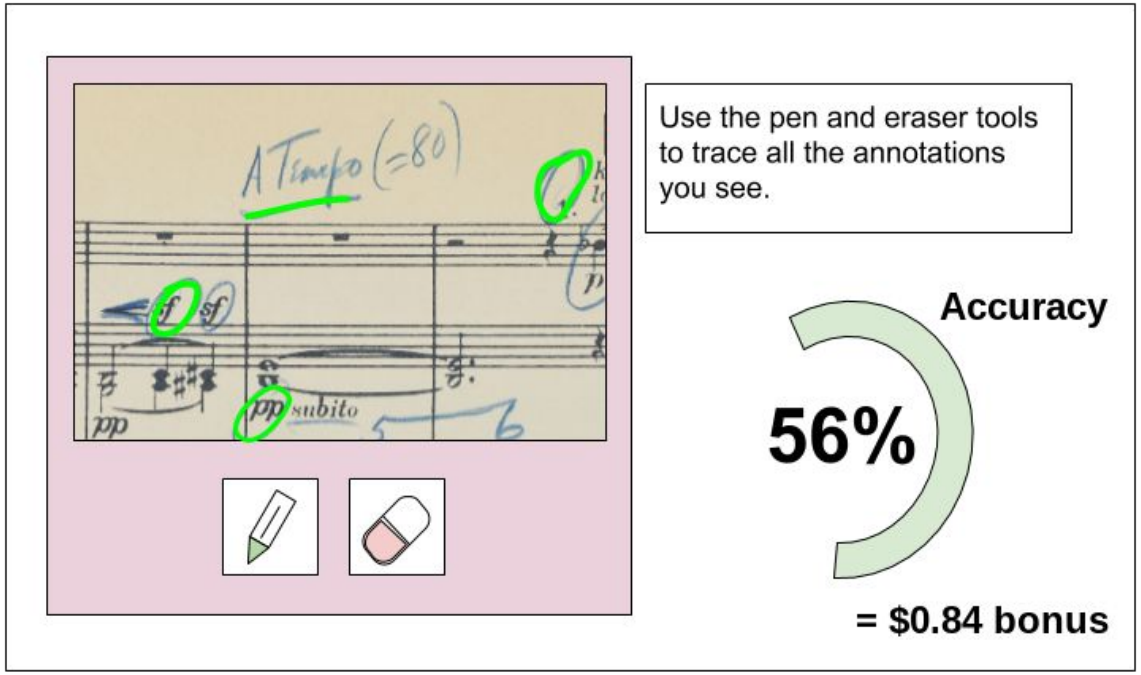
- **worker quality assessment**

  measure: accuracy

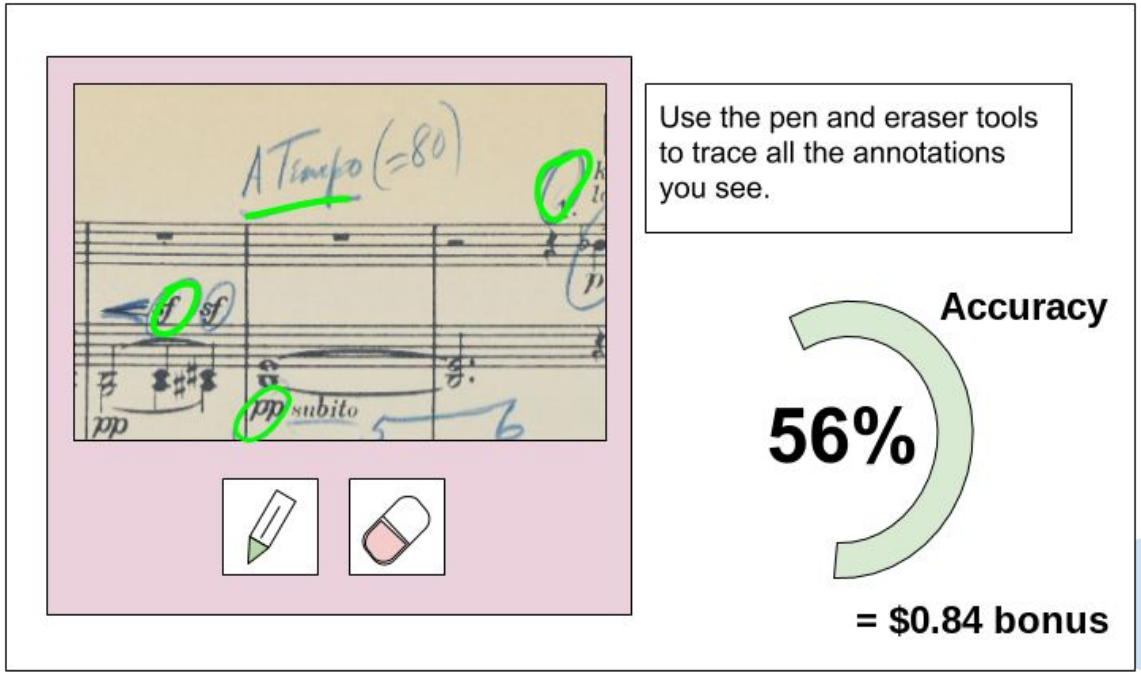
- **identifying tiles with failure cases/contention**

  measure: interworker consensus

Application 2: “Live” segmentation task feedback



Application 2: “Live” segmentation task feedback



We know:

- probability under model whether pixel contains annotation

We ask:

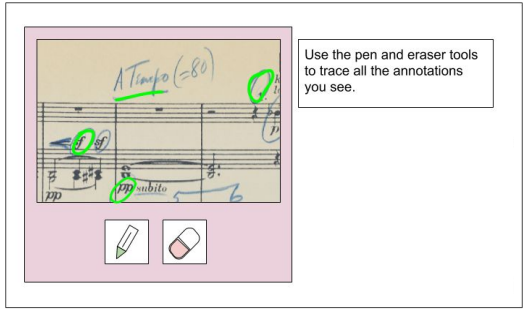
- does pixel contain annotation?

Applications:

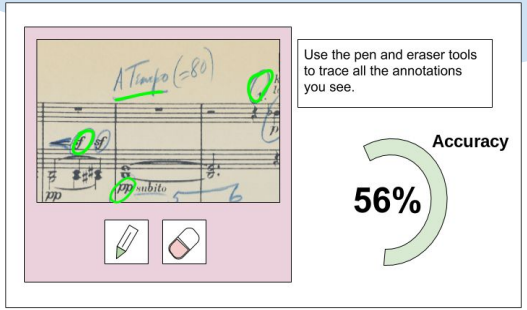
- financial reward (accuracy bonus)
- other rewards (gamification)
- assigning better workers to harder tiles **during a task**

Lots of experiments with UI possible

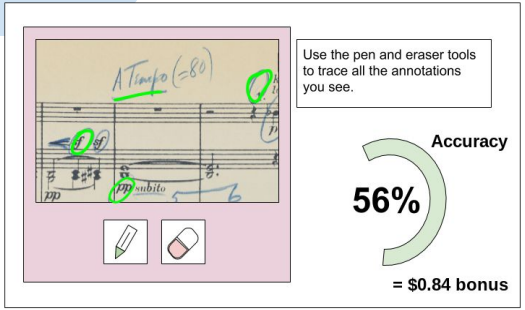
Measure (DV): accuracy, precision, time to completion, engagement, # corrections/undos, satisfaction (!) etc.



vs.

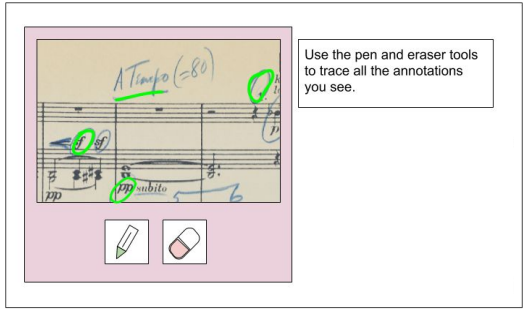


vs.

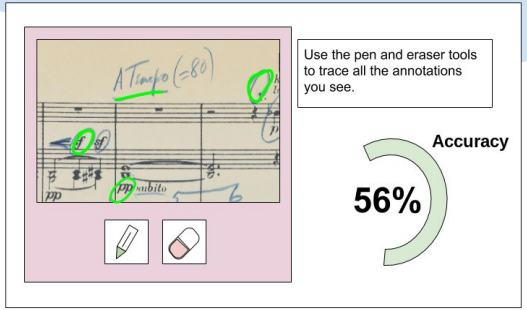


Lots of experiments with UI possible

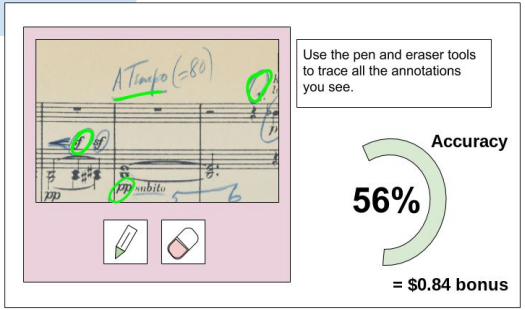
Measure (DV): accuracy, precision, time to completion, engagement, # corrections/undos, satisfaction (!) etc.



vs.



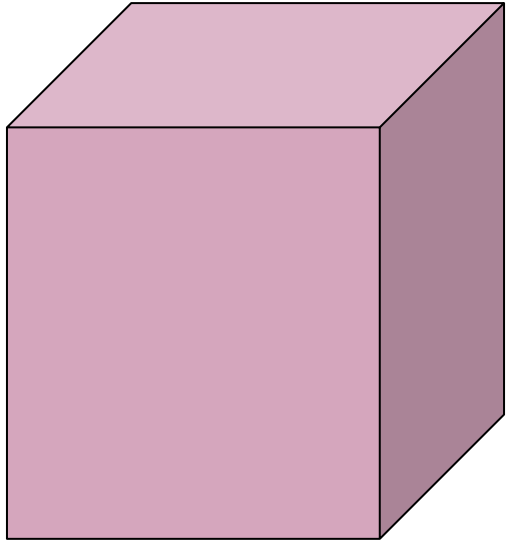
vs.



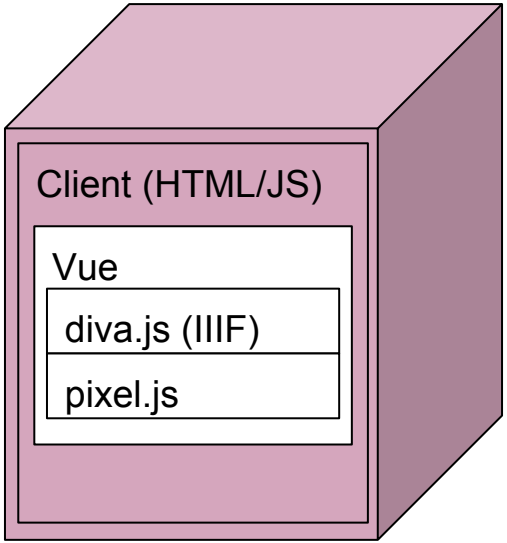
Other things to tweak (IV)

- feedback function
- tile size
- greyscale vs. color
- editor tools
- undo history size
- add “noise” to catch false +ves
- non-financial rewards (facts from LOD)
- reward closures

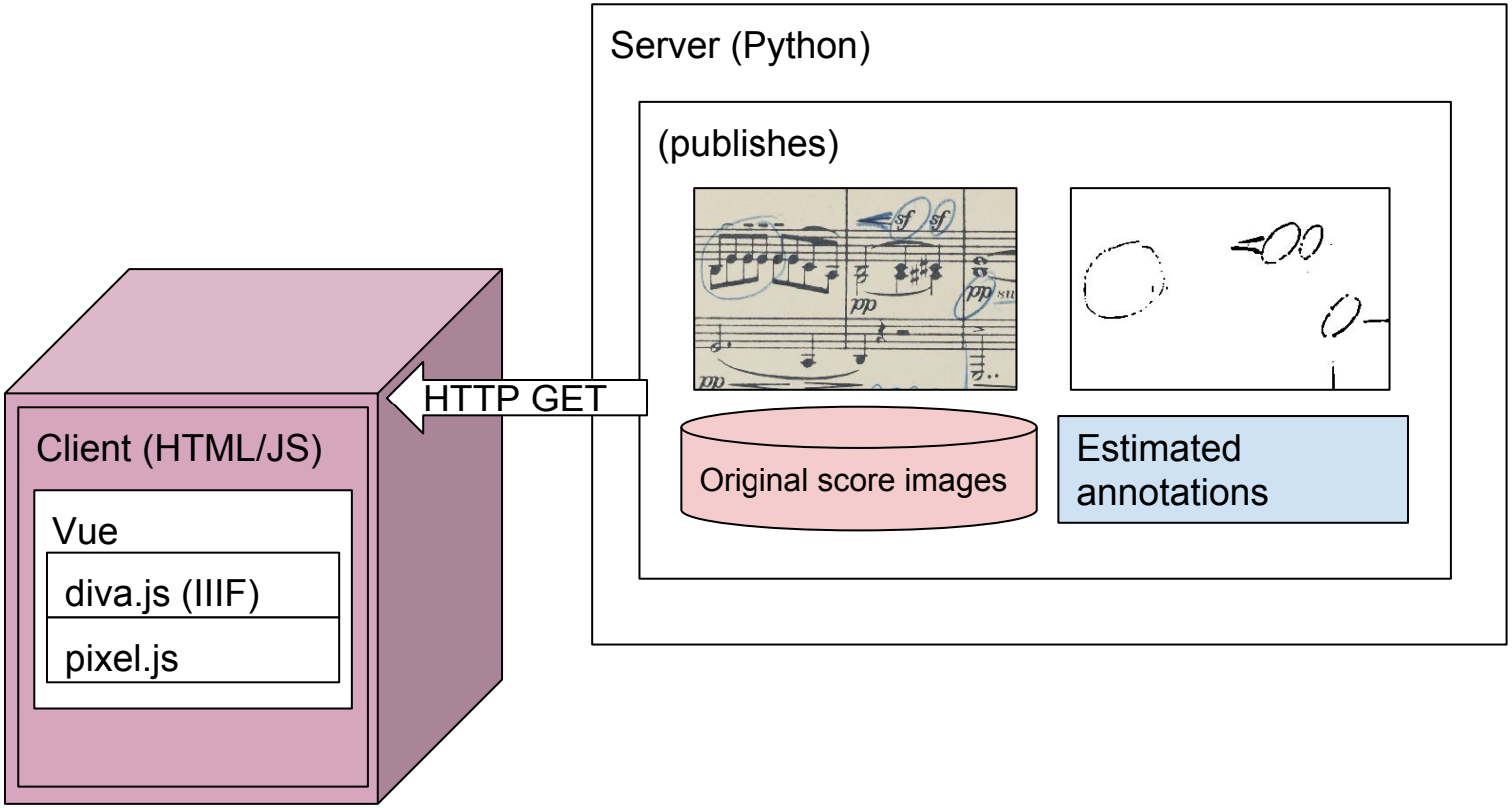
(Of course we need ground truth for accuracy and precision too, but we can share the same image across a pool of workers)

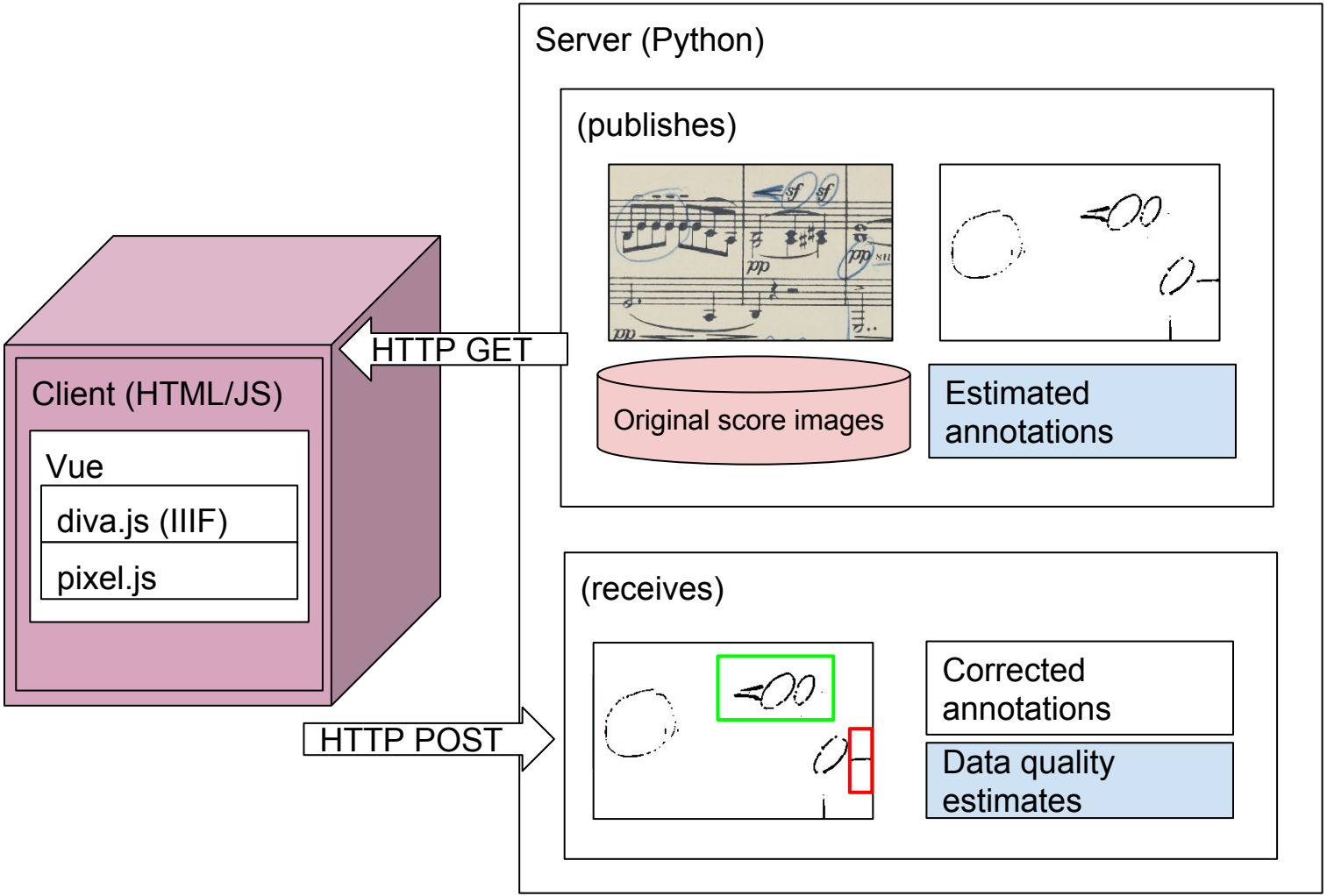


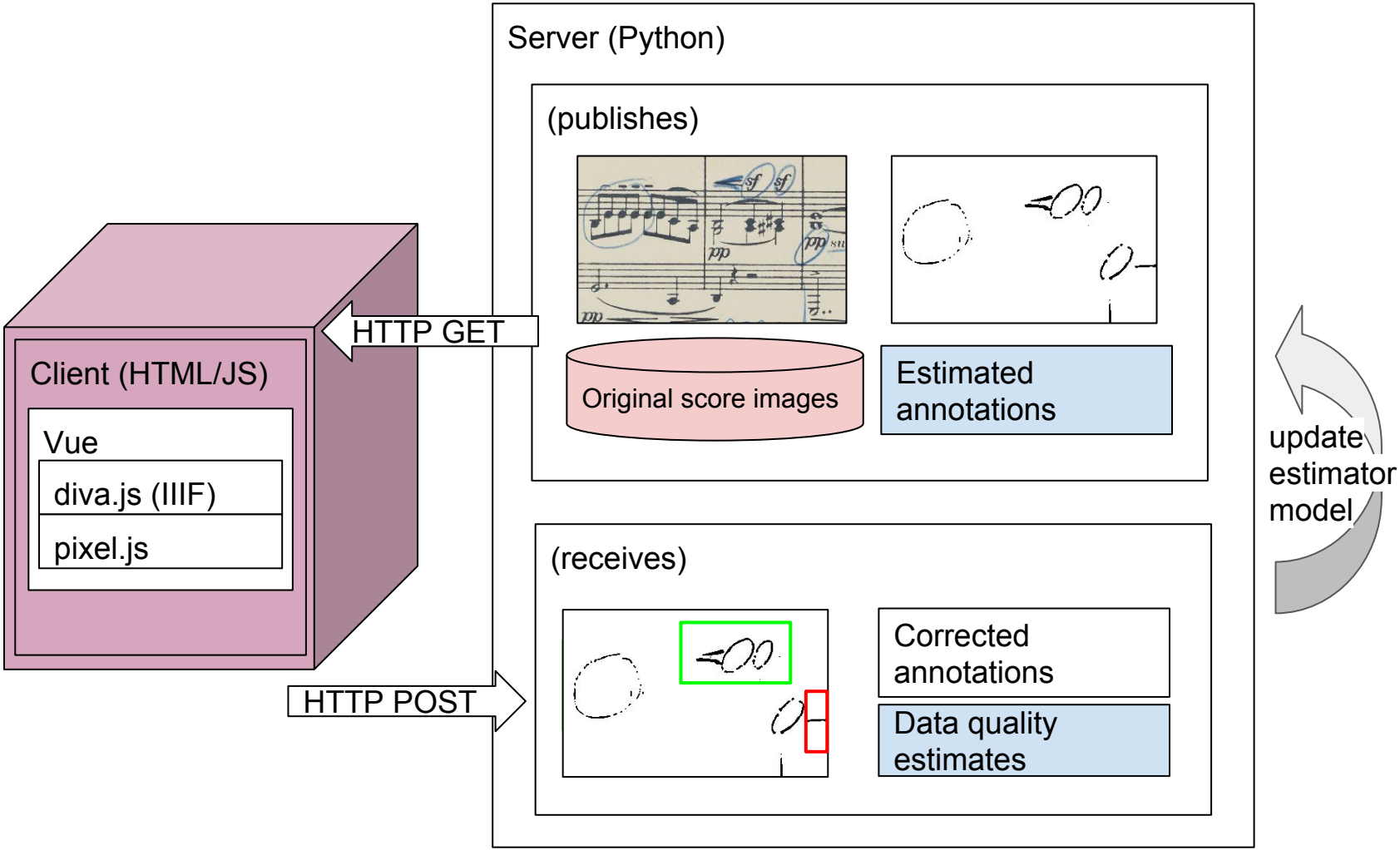
What’s in the red box?

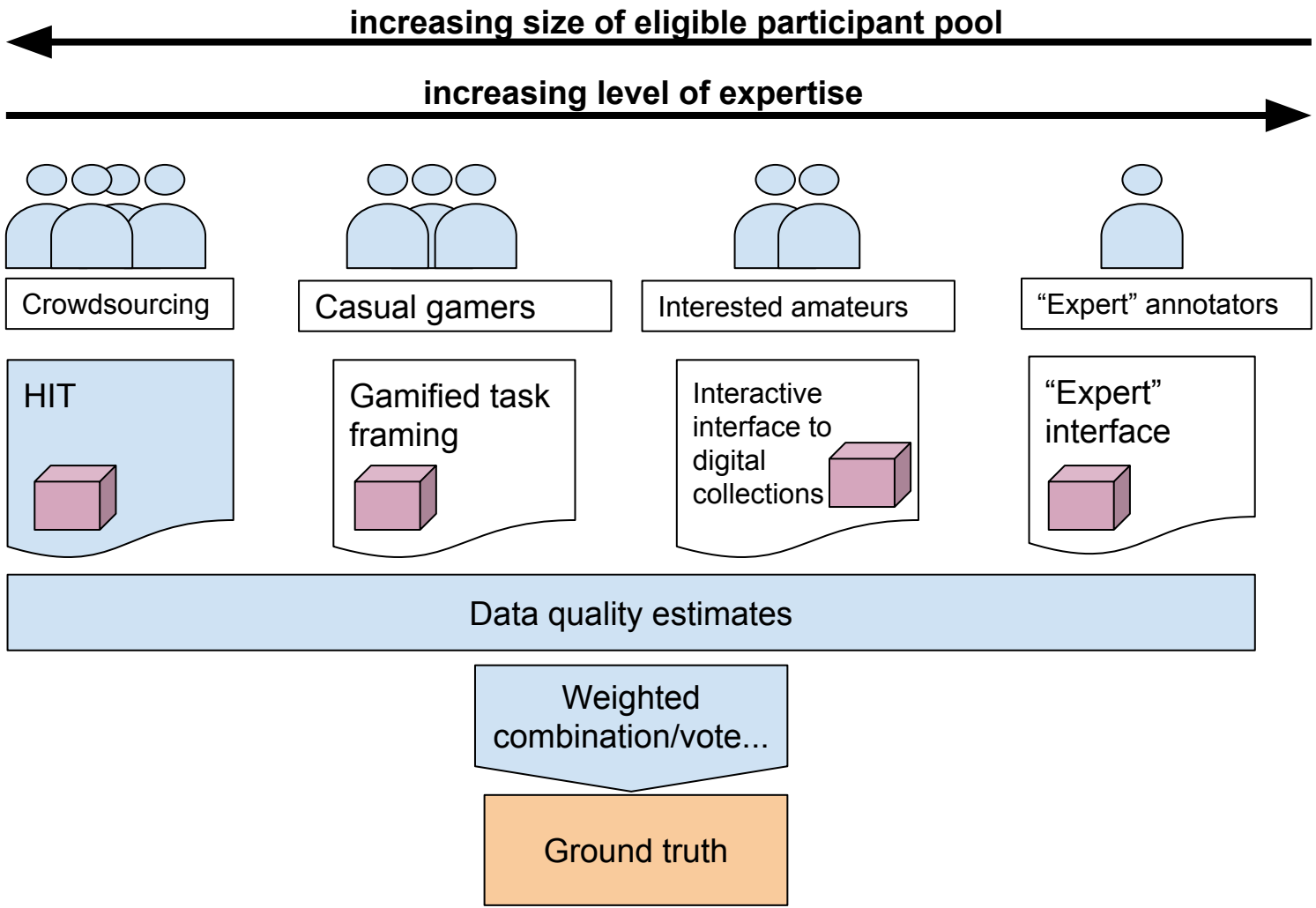


What’s in the red box?


Server (Python)
(publishes)
Original score images
Estimated annotations
HTTP GET
Client (HTML/JS)
Vue
diva.js (IIIF)
pixel.js


Server (Python)
(publishes)
Original score images
Estimated annotations
HTTP GET
Client (HTML/JS)
Vue
diva.js (IIIF)
pixel.js
HTTP POST
(receives)
Corrected annotations
Data quality estimates


Server (Python)
(publishes)
Original score images
Estimated annotations
HTTP GET
Client (HTML/JS)
Vue
diva.js (IIIF)
pixel.js
HTTP POST
(receives)
Corrected annotations
Data quality estimates
update estimator model


increasing size of eligible participant pool
increasing level of expertise
Crowdsourcing
Casual gamers
Interested amateurs
“Expert” annotators
HIT
Gamified task framing
Interactive interface to digital collections
“Expert” interface
Data quality estimates
Weighted combination/vote...
Ground truth

Bibliography

Bell, Eamonn, and Laurent Pugin. 2016. “Approaches to Handwritten Conductor Annotation Extraction in Musical Scores.” In *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology - DLfM 2016*. ACM Press. https://doi.org/10.1145/2970044.2970053.

Bell, Eamonn, and Laurent Pugin. 2018. “Heuristic and Supervised Approaches to Handwritten Annotation Extraction for Musical Score Images.” International Journal on Digital Libraries, July. https://doi.org/10.1007/s00799-018-0249-7.

Bernstein, Michael S., Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. “Soylent: A Word Processor with a Crowd Inside.” In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, 313–322. ACM.

Calvo-Zaragoza, Jorge, Luisa Micó, and Jose Oncina. 2016. “Music Staff Removal with Supervised Pixel Classification.” *International Journal on Document Analysis and Recognition (IJDAR)* 19 (3): 211–19. https://doi.org/10.1007/s10032-016-0266-2.

Dow, Steven, Anand Kulkarni, Brie Bunge, Truc Nguyen, Scott Klemmer, and Björn Hartmann. 2011. “Shepherding the Crowd: Managing and Providing Feedback to Crowd Workers.” In *CHI ’11 Extended Abstracts on Human Factors in Computing Systems*, 1669–1674. CHI EA ’11. New York, NY, USA: ACM. https://doi.org/10.1145/1979742.1979826.

Gallego, Antonio-Javier, and Jorge Calvo-Zaragoza. 2017. “Staff-Line Removal with Selectional Auto-Encoders.” *Expert Systems with Applications* 89 (December): 138–48. https://doi.org/10.1016/j.eswa.2017.07.002.

Gurari, Danna, Mehrnoosh Sameki, Zheng Wu, and Margrit Betke. n.d. “Mixing Crowd and Algorithm Efforts to Segment Objects in Biomedical Images.” Accessed November 28, 2016. http://cs-people.bu.edu/wuzheng/research/publication/MiccaiImic2016_SAVE_System.pdf.

Peng, Xujun, Srirangaraj Setlur, Venu Govindaraju, and Ramachandrula Sitaram. 2011. “Handwritten Text Separation from Annotated Machine Printed Documents Using Markov Random Fields.” *International Journal on Document Analysis and Recognition (IJDAR)* 16 (1): 1–16. https://doi.org/10.1007/s10032-011-0179-z.