

1 **Three-dimensional data of wirecut surface scans
2 under the confocal microscope (110 character
3 maximum, inc. spaces)**

4 **Yuhang Lin^{1,2} and Heike Hofmann^{2,3}**

5 ¹Iowa State University, Department of Statistics, Ames,

6 Center for Statistics and Applications in Forensic Evidence (CSAFE), Iowa State University, Ames,

7 ³University of Nebraska-Lincoln, Department of Statistics, Lincoln,

8 *corresponding author(s): Yuhang Lin (yuhanglin@iastate.edu) **it.corresponding returns nothing**

9 *corresponding author(s): Heike Hofmann (hhofmann4@unl.edu) **it.corresponding returns nothing**

10 **ABSTRACT**

Update later

Wire cut data is important in forensic investigations but lacks a systematic way of analyzing the data. We created a data set of 120 scans of aluminum wire cut in x3p format, using 5 wire cutters and 3 locations along the 4 blades, with 2 replicates for each combination. A systematic pipeline with multiple analysis plots was developed to analyze the data and draw conclusions based on numerical measures.

(maximum 170 words) This is a manuscript template for Data Descriptor submissions to *Scientific Data* (<http://www.nature.com/scientificdata>). The abstract must be no longer than 170 words, and should succinctly describe the study, the assay(s) performed, the resulting data, and the reuse potential, but should not make any claims regarding new scientific findings. No references are allowed in this section.

12 Please note: Abbreviations should be introduced at the first mention in the main text no abbreviations lists or tables should be
13 included. Structure of the main text is provided below.

14 **1 Background & Summary**

15 **XXX This section needs a lot more detail - as per instructions, we need:**

16 **overview of the study design: create dataset, manual alignment**

17 **overview of files created (table?):**

- 18 • x3p files from scans: format, explanations, link to sections
- 19 • derivatives from manual processing: profiles, signals
- 20 • aligned signals
- 21 • meta csv

22 **context of previous work and literature**

23 **motivation for the creation of the dataset**

24 Wire cut data is a type of forensic tool mark data used to identify the source of a wire cutter based on the striations left on
25 the surface. There have been cases where the evidence and testimony on wire cut evidence played a crucial role in the criminal
26 investigation and conviction of a defendant. However, there is a lack of a standardized method to analyze it except for visual
27 comparison. ...literature review and motivation... Here, we provide a data set with multiple files in our study, see Table 1.

Table 1. Overview of the whole data set.

Description	Section
-------------	---------

Raw data

x3p files metadata	120 scans of aluminum wire cut in x3p format metadata of the scans in 1 CSV	Cutting wires Cutting wires
Manually processed derivatives		
profiles	profiles extracted from scans in 120 CSVs	Extract profiles
Computational processed derivatives		
signals	signals filtered from profiles in 1 CSV	Filtered signals
aligned signals	pictures of pairwise aligned signals from same sources in PNGs	Align signals
CCF values	CCF values of all pairwise aligned signals in 1 CSV	Align signals

28 breakable table, cannot cross-reference inside table

29 For the reproducibility of all our data and alignment results, we introduce in details in Section 2.1 (cross-reference not
30 working without section number, number-sections: false not working) how we cut the wire and collect the 120 scans with
31 5 tools on 3 locations, in Section 2.2 how we extract profiles from the scans, in Section 2.3 how we filter signals from the
32 profiles, in Section 2.4 how we align signals from different scans, and optimize the alignment with the cross-correlation
33 function (CCF) values. Then, a technical validation was conducted to further compare signals from different sources also
34 match our assumption. We hope this pipeline developed using this data set can be further generalized and applied to real crime
35 scenes to help investigators draw conclusions based on real wire cut data.

36 (unlimited length) An overview of the study design, the assay(s) performed, and the created data, including any background
37 information needed to put this study in the context of previous work and the literature. The section should also briefly outline
38 the broader goals that motivated the creation of this dataset and the potential reuse value. We also encourage authors to include
39 a figure that provides a schematic overview of the study and assay(s) design. The Background & Summary should not include
40 subheadings. This section and the other main body sections of the manuscript should include citations to the literature as
41 needed.

42 2 Methods

43 full descriptions of the experimental design, data acquisition assays, and any computational processing

44 In this study, aluminum wire was used to create an optimal scenario where the most amount of information could be
45 transferred from the tool to the substrate, despite the wire in some real cases being made of lead. The physical property of
46 aluminium wire make it an excellent candidate for keeping marks while being relatively easy to bend and non-toxic.

47 2.1 Cutting wires

48 The aluminum wire used was 16 Gauge/1.5 mm, anodized. In order to cut the wire, 4-inch pieces were unspooled and cut using
49 Kaiweets wire cutters, model KWS-105, as shown in Figure 1(a), for 1 blade location, either inner, middle, or outer, which
50 gives us 1 replicate. Each piece was then cut into half to create 2-inch pieces for each side, AB and CD, with a sharpie line
51 marking the cut ends, giving us 4 samples. Here, we are showing AB sides only in Figure 1(b) (need a different tent figure),
52 and the CD sides are similar from the other side of the cut, with the back of A being C and the back of B being D. Both AB and
53 CD sides form tent structures on the tips of the wire, and we can separate each side of the tent into 2 pieces along the bending
54 position, resulting in 8 scans. We repeated this process for all 3 locations along the blade and 5 wire cutters, with 2 replicates
55 for each tool-edge-location combination, resulting in 120 scans. Each piece was labeled with the naming conventions, T(ool)
56 1/2/3/4/5 (Edge) A/B/C/D W(ire) - L(ocation) I(nner)/M(iddle)/O(uter) - R(epetition) 1/2, with T1AW-LI-R1 being the piece
57 cut by tool 1 on the A edge at the inner location for the first repetition. Then, we can use the standard scanning protocols for
58 the confocal microscope, shown in Figure 1(c) (need an extra pic of the very tip), to scan the wire tip surfaces. The scanned
59 surfaces are saved in a resolution of $0.645\mu\text{m} \times 0.645\mu\text{m}$ per square pixel in an x3p file format.

60 XXX figure 1 - generally, zoom into these images - we do not want to have a hand in the image, nor a view of the
61 crafting aluminum :) - what are the exact rules on visuals in Scientific Data ? XXX hard to put the full requirements here, see
62 <https://www.nature.com/sdata/publish/submission-guidelines#figures>

63 put into quarto layout with (a) (b) (c) on the top left, no tent, add blade C & D

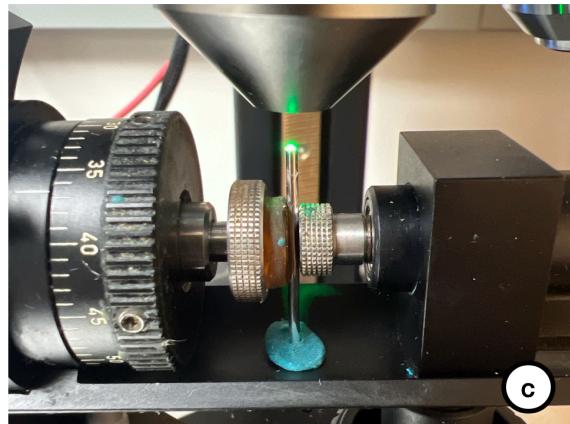
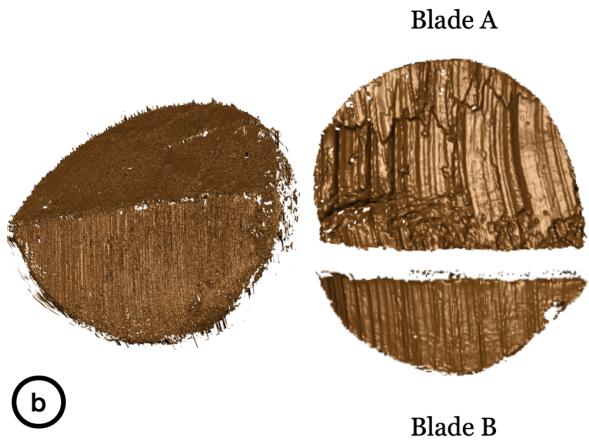
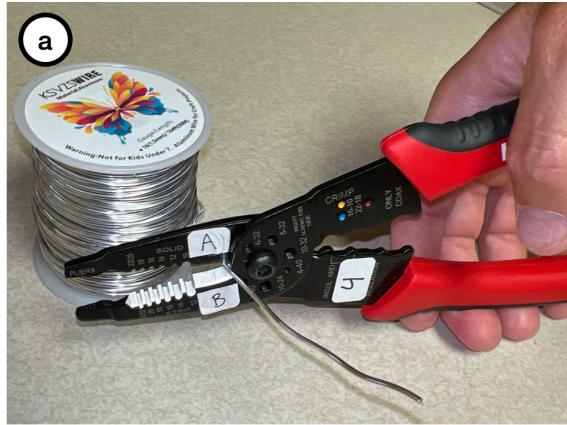


Figure 1. (a) A Kaiweets wire cutter of model KWS-105 was used to cut the wire. (b) A tent structure created by blade AB. After separating 2 tent structures by the connecting position, we obtained 2 samples - 1 sample from blade A and B. (c) A confocal microscope was used to scan the wire surfaces.

64 **2.2 Extract profiles**

65 Numerical comparisons between 2 replicates cannot be done directly on the $\times 3p$ files. We need to extract representative
 66 functions from the scans first. A representative function with the most information is considered as a signal for one scan,
 67 which can be used for comparison later. To obtain this function, we first need a profile of the scan, which is a sequence of
 68 values along a user-drawn line on the surface. The profile should capture most features of the scan, and be orthogonal to the
 69 striation marks of the scan, which are formed by ups and downs of grooves. So, we draw the line across the wide region of
 70 the scan to maximize the feature captured, as shown in dark blue in Figure 2(a). We can then investigate the values under this
 71 profile line. The profile function is along the line is plotted in Figure 2(b).

72 **2.3 Filtered signals**

73 With the profile extracted, we can then obtain the signal. Two Gaussian filters are applied to these resulting profiles. In
 74 particular, we first used a large low-pass filter with bandwidths of 400 microns to remove large trend, as it can overwhelm the
 75 signals, and then used a small high-pass filter of 40 microns to average across noise and remove spikes, as shown in Figure
 76 2(c). [@clevelandLocalRegressionModels1992 ?](#) (add reference: W. S. Cleveland, E. Grosse and W. M. Shyu (1992) Local
 77 regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.).
 78 Finally, the extreme tail values are removed.

79 **2.4 Align signals**

80 Signals extracted from different scans can be put together for comparison, and we maximize the cross-correlation function
 81 (CCF) values between the signals to numerically find the best alignment. For example, we compare T1AW-LI-R1 to T1AW-
 82 LI-R2, T1CW-LI-R1 to T1CW-LI-R2, and so on. That is comparing each row in Figure 3. We know that signals from two
 83 replicates with the same tool-edge-location combination should yield similar signals as in the first and second column of

84 Figure 4, which will give alignments of massive overlapping and high CCF values close to 1. The alignments and values we
85 got in the rightmost column of Figure 4 fulfill our expectations.

86 (unlimited length) The Methods should include detailed text describing any steps or procedures used in producing the data,
87 including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization,
88 image feature extraction). See the detailed section in our submission guidelines for advice on writing a transparent
89 and reproducible methods section. Related methods should be grouped under corresponding subheadings where possible, and
90 methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study.
91 Specific data outputs should be explicitly referenced via data citation (see Data Records and Citing Data, below).

92 Authors should cite previous descriptions of the methods under use, but ideally the method descriptions should be complete
93 enough for others to understand and reproduce the methods and processing steps without referring to associated publications.
94 There is no limit to the length of the Methods section. Subheadings should not be numbered.

95 Authors should review the transparent methods checklist below, and ensure that their manuscript complies with any relevant
96 points. Authors are also encouraged to search FAIRsharing.org for community reporting standards that may be relevant
97 to their specific data-type.

98 **Transparent Methods Checklist**

- 99 • Materials & reagents: Identify commercial suppliers of reagents, instrumentation or kits, when the source is critical to
100 the outcome of the experiments. Declare any restrictions on the availability of unique materials (more information here).
101 Provide catalogue or clone numbers for all antibodies (if available). For primary antibodies, provide proof of validation
102 for the relevant species and applications.
- 103 • Exclusion criteria: If any data or samples were excluded, explain the exclusion criteria and state in the methods whether
104 the criteria were established before the study was conducted.
- 105 • Randomization & blinding: For any studies that involve assigning samples, animals or participants into different groups:
106 State clearly whether randomization methods were used. If randomization was not employed, this should be clearly
107 stated. State clearly whether blinding was employed during data collection. If blinding was not employed, this should
108 be clearly stated.
- 109 • Animal & human studies (full journal policies here): Experiments involving human participants must identify the
110 committee approving the experiments, and include a statement confirming that informed consent was obtained from all
111 participants. Studies employing nonhuman animals should ensure that methods descriptions comply with the ARRIVE
112 checklist.
- 113 • Cell lines: For each eukaryotic cell line used, state the source and whether the cell line has been authenticated or
114 otherwise tested for integrity. If any commonly misidentified cell lines were used (see ICLAC or NCBI Biosample),
115 justify their use. Report whether the cell lines were tested for mycoplasma contamination.
- 116 • Chemistry & materials science: Manuscripts describing chemical syntheses, or characterizing new chemicals or materials
117 should refer to the guidance at Nature Chemistry.

118 **3 Data Records**

119 The complete data set is available on the ISU DataShare repository at <https://iastate.figshare.com/>, which is public and open
120 access for every interested researcher. The data set consists of 120 scans in the x3p file format with the naming convention
121 as described before. ([Explain the x3p header info?](#))

122 (unlimited length) The Data Records section should be used to explain each data record associated with this work, in-
123 cluding the repository where this information is stored, and to provide an overview of the data files and their formats. Each
124 external data record should be cited numerically in the text of this section, for example ?, and included in the main reference
125 list as described below. A data citation should also be placed in the subsection of the Methods containing the data-collection
126 or analytical procedure(s) used to derive the corresponding record. Providing a direct link to the dataset may also be helpful
127 to readers (<https://doi.org/10.6084/m9.figshare.853801>).

128 Tables should be used to support the data records, and should clearly indicate the samples and subjects (study inputs), their
129 provenance, and the experimental manipulations performed on each (please see 'Tables' below). They should also specify the
130 data output resulting from each data-collection or analytical step, should these form part of the archived record.

131 4 Technical Validation

132 a picture of alignment with ccf from different sources to show if different source, our evaluation returns small ccf, which
133 matches what we thought.

134 For the data collection process, two team members did the cutting and labeling together, then one person did the scanning
135 and named according to the naming convention above. The scanning was done in a specific order to ensure consistency across
136 all scans. The data was saved in a consistent format to ensure they could be easily accessed and analyzed. A third person then
137 checked the data to ensure that the data was consistent in naming and accurate.

138 again - the website is not the right place for the validation - instead, move parts from the website here.

139 For the validation of the scans and their processing we investigate the correlation scores of pair-wise aligned signals. For
140 signals from scans of wires cut with a different tool, we would expect a low correlation score. Large scores between signals
141 are indicative of being made by the same tool. Show boxplots and roc curve.

142 For validation of all other tools and locations of scan replicates, see the detailed [report](#).

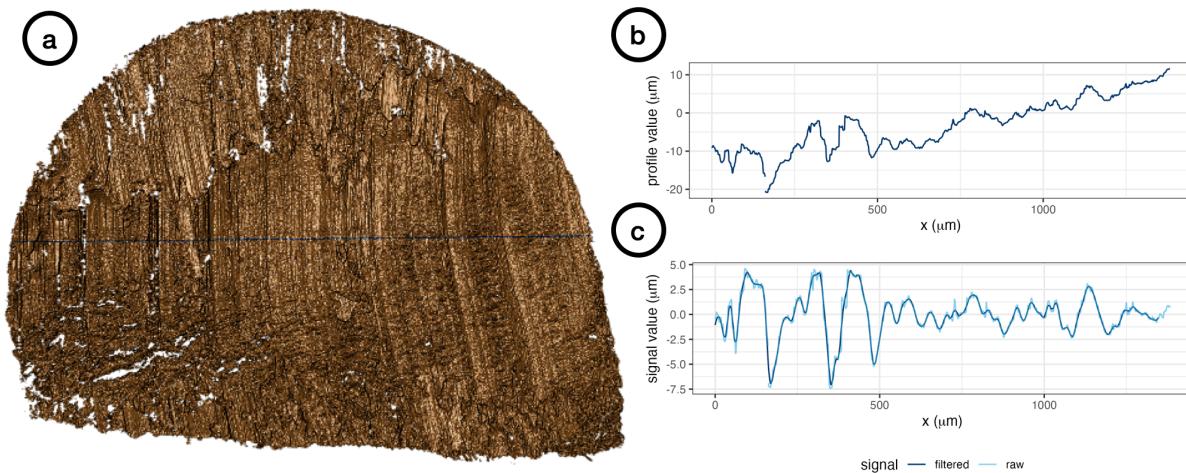


Figure 2. (a) A profile line in dark blue was drawn across the striations of the scan. (b) The profile function extracted along the profile line in (a). (c) The raw signal in light blue is obtained by using the low-pass filter on the profile function in (b) and the filtered signal is obtained by using the high-pass filter on the raw signal.

143 (unlimited length) This section presents any experiments or analyses that are needed to support the technical quality of
144 the dataset. This section may be supported by figures and tables, as needed. This is a required section; authors must present
145 information justifying the reliability of their data.

146 • Measurement of data quality?

147 – Numeric measurements / tests: ?

148 – Visualizations: ?

149 – Check with existing data: ?

150 – Questionable / slur procedures:

151 * [AidDatas Geospatial Global Chinese Development Finance Dataset](#): Second, all data collected is reviewed
152 by at least two individuals. Although this is not a double-blind review procedure, the use of satellite imagery
153 to verify project locations results in far less uncertainty when compared to previous approaches to geocoding
154 where locations were selected entirely based on text descriptions.

155 * [A large open access dataset of brain metastasis 3D segmentations on MRI with clinical and imaging information](#): A medical student (D.R.) double checked and adjusted the revised NIfTI segmentation masks and
156 manually counted the number of lesions with contrast-enhancement, necrosis, and peritumoral edema for
157 each patient.

- * **Time series of freshwater macroinvertebrate abundances and site characteristics of European streams and rivers:** Technical validation of the TREAM dataset was achieved through exclusion of time series data that did not match our inclusion criteria and data standardisation steps (outlined in Methods above). Any noted issues that did not adhere to the outlined standardisation within the datasets from the 41 independent projects included in this dataset were checked with data providers and corrected or removed when standardisation was not achievable (e.g., when collection methods changed over the course of the time series).
- * **3D surgical instrument collection for computer vision and extended reality:** The main issue...Since we store our models in a standard format (STL), they are compatible with a large variety of visualisation and processing software.
- * **Three-dimensional reconstruction of high latitude bamboo coral via X-ray microfocus Computed Tomography :** Regular quality assurance inspections are carried out on the μ -CT scanner to verify its metrological and geometrical (alignments) accuracy for conducting the scans. The geometry of source to object and source to detector distances are verified whenever there is any significant physical interaction with the source such as re-alignment, change of filament, or source anode change. This calibration process involves scanning a specially designed phantom known as an hourglass36, which consists of three pairs of high-sphericity spheres. The sphere sizes are as follows: two spheres with a diameter of 3.000 mm, two spheres with a diameter of 6.000 mm, and two spheres with a diameter of 9.525 mm, and each sphere is kept in contact with its size-counterpart. By using this phantom, it becomes possible to accurately determine a known distance, specifically the centre-to-centre distance of the spheres, in a threshold-independent manner. If the measured distance deviates beyond the acceptable limits of metrological accuracy, the systems calibration parameters are adjusted to ensure agreement between the measured distance and the actual distance.

5 Usage Notes

The R package `x3ptools` (available from CRAN) supports working with files in `x3p` format.

Sample scripts in R for processing scans from `x3p` format to their signal are available from ... [github](#).

Further analysis can be conducted with the GitHub R package `wire` and the GitHub R shiny app `wireShiny` ([citation?](#)). We already conduct between-replicate comparisons in the technical validation section, and we can also conduct across-replicate comparisons to establish error rates threshold and produce other analysis plots.

Suppose we put the CCF values in a tilemap with different tool, location and edge combinations. In that case, we expect only the diagonal to have high CCF values, close to 1 and marked as orange in the tilemap, as the diagonal represents the same source, and the rest of the matrix to have low CCF values, close to 0 and marked as gray. In Figure 5, the behavior is consistent with our expectation overall, except for some rare cases with tool 5 edge D, which is caused by [?????](#). We also put the resulting CCFs in the boxplot, as in Figure 6. We can see that the CCF values for the same sources are close to 1, while the CCF values for different sources are much lower than expected. This difference can be used to establish a threshold for CCF and help us draw conclusions about the similarity between wire cut scans numerically, which can be used in real crime scenes. The density plot in Figure 7 shows the distribution of the CCF values with the same sources and different sources. The overlapping points between the tails of these two distributions can be a rough threshold. Furthermore, the receiver operating characteristic (ROC) curve in Figure 8 shows the sensitivity / true positive rate against the false positive rate (FPR) (1 - specificity). The curve is very close to the upper left corner, which is excellent for classification and drawing conclusion. It gives us a true threshold of 0.589 to control the FPR to be less than 0.05 with false negative rate (FNR) to be 0, ([false positive rate \(FPR\) / false discovery rate \(FDR\)](#) -> define the H0 or call it false identification rate (FIR)[???](#)), and 0.658 to control the FPR to be less than 0.01, with FNR to be 0.02.

(unlimited length) The Usage Notes should contain brief instructions to assist other researchers with reuse of the data. This may include discussion of software packages that are suitable for analysing the assay data files, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Please see our code availability policy for advice on supplying custom code alongside Data Descriptor manuscripts.

For studies involving privacy or safety controls on public access to the data, this section should describe in detail these controls, including how authors can apply to access the data, what criteria will be used to determine who may access the data, and any limitations on data use.

6 Code availability

[table of code-manual?](#)

210 no, we can't use the website as a place for more detailed procedures. This paper is the detailed procedure.
211 README, scanning procedures in another HTML

212 We put together the cutting and the standard scanning procedures mentioned in Section 2.1 cross-ref not working with
213 more pictures for each step into a [README of the GitHub repository heike/Wirecuts](#) (High-res pics needed in the
214 README).

215 The data set can be easily accessed with the CRAN R package `x3ptools`. Further analysis can be conducted with the
216 GitHub R package `wire` and the GitHub R shiny app `wireShiny` ([citation](#)) (again????).

217 For all studies using custom code in the generation or processing of datasets, a statement must be included under the heading
218 "Code availability", indicating whether and how the code can be accessed, including any restrictions to access. This section
219 should also include information on the versions of any software used, if relevant, and any specific variables or parameters used
220 to generate, test, or process the current dataset.

221 **7 End of Body**

222 **(Note that the bibliography style and the name of the bib-file are hard coded in the template file right now.)**

223 LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via
224 the project menu. Use the `cite` command for an inline citation, e.g. ?????. For data citations of datasets uploaded to e.g. `figshare`,
225 please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014`
226 example in the sample bibliography file. For journal articles, DOIs should be included for works in press that do not yet
227 have volume or page numbers. For other journal articles, DOIs should be included uniformly for all articles or not at all. We
228 recommend that you encode all DOIs in your bibtex database as full URLs, e.g. <https://doi.org/10.1007/s12110-009-9068-2>.

229 **Acknowledgements**

230 Acknowledgements should be brief, and should not include thanks to anonymous referees and editors, or effusive comments.
231 Grant or contribution numbers may be acknowledged.

232 **Author contributions statement**

233 Y.L. did all of the work, H.H. made him do the work. But seriously, this paper is the one where we need to cite everybody:
234 Eden Amin, Curtis Mosher, Jeff Salyards. Alicia? Must include all authors, identified by initials, for example: A.A. conceived
235 the experiment(s), A.A. and B.A. conducted the experiment(s), C.A. and D.A. analysed the results. All authors reviewed the
236 manuscript.

237 **Competing interests**

238 (mandatory statement)

239 H.H. is a technical advisor to AFTE (Association of Firearms and Toolmarks Examiners), fellow of the ASA (American
240 Statistical Association), and committee member of the ASA Forensic Science Committee. H.H. has testified as court witness
241 on behalf of judge April Neubauer, NY State Supreme Court Criminal Term in New York City. The corresponding author
242 is responsible for providing a [competing interests statement](#) on behalf of all authors of the paper. This statement must be
243 included in the submitted article file.

244 **Figures & Tables**

245 Figures, tables, and their legends, should be included at the end of the document. Figures and tables can be referenced in
246 L^AT_EX using the `ref` command, e.g. Figure 9 and Table 2.

247 Authors are encouraged to provide one or more tables that provide basic information on the main inputs to the study
248 (e.g. samples, participants, or information sources) and the main data outputs of the study. Tables in the manuscript should
249 generally not be used to present primary data (i.e. measurements). Tables containing primary data should be submitted to an
250 appropriate data repository.

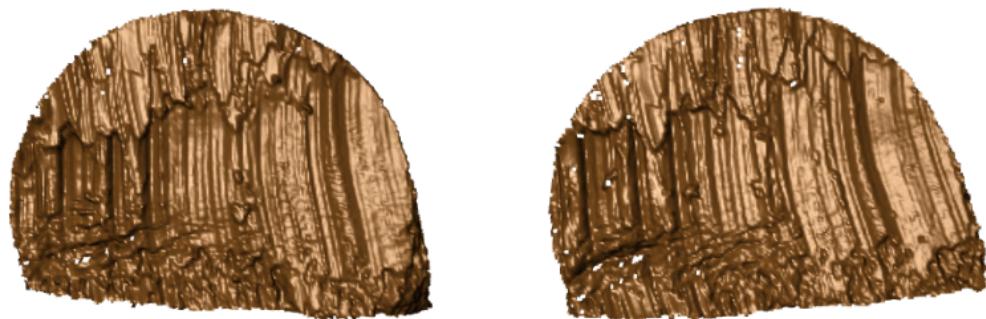
251 Tables may be provided within the L^AT_EX document or as separate files (tab-delimited text or Excel files). Legends, where
252 needed, should be included here. Generally, a Data Descriptor should have fewer than ten Tables, but more may be allowed
253 when needed. Tables may be of any size, but only Tables which fit onto a single printed page will be included in the PDF
254 version of the article (up to a maximum of three).

255 Due to typesetting constraints, tables that do not fit onto a single A4 page cannot be included in the PDF version of the
256 article and will be made available in the online version only. Any such tables must be labelled in the text as Online-only tables
257 and numbered separately from the main table list e.g. Table 1, Table 2, Online-only Table 1 etc.

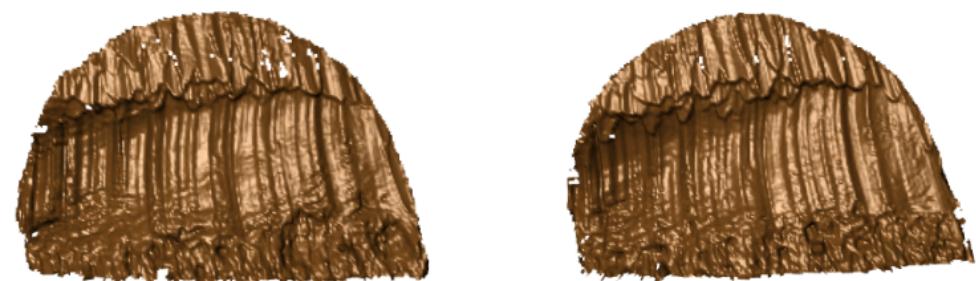
Condition	n	p
A	5	0.1
B	10	0.01

Table 2. Legend (350 words max). Example legend text.

Edge A



Edge C



Edge B



Edge D



Figure 3. Scans from different sides of tool 1 at the inner location.

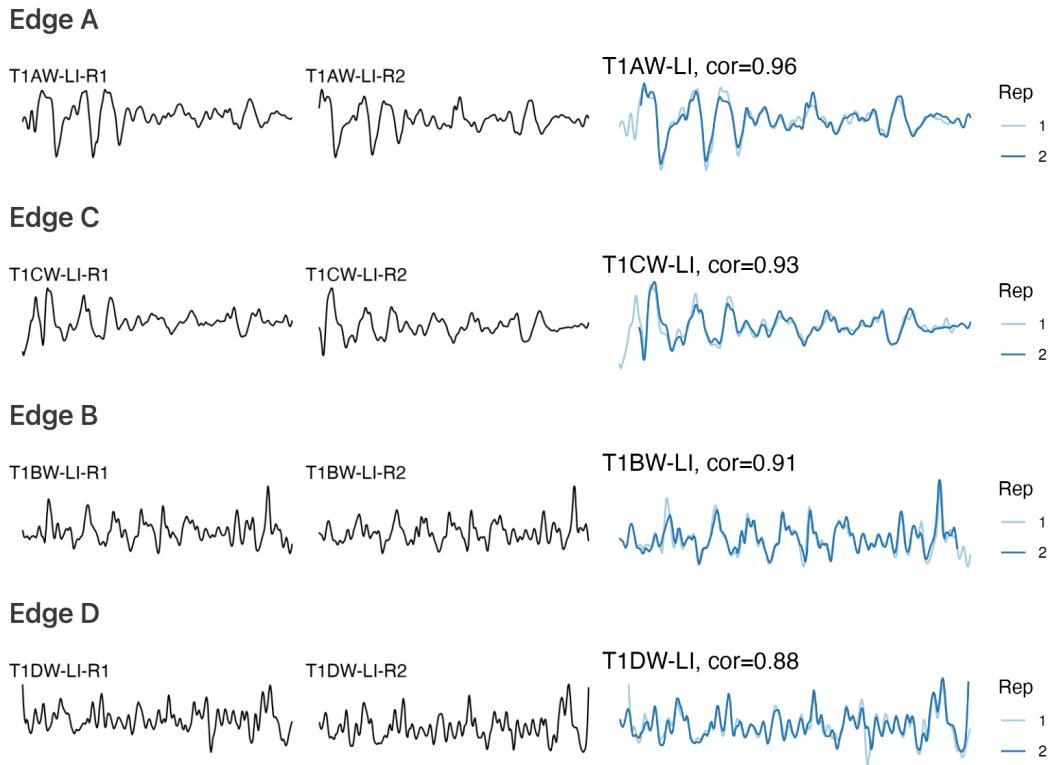


Figure 4. The first and second columns show the signals extracted from Figure 3, and the third column shows the alignments and CCF values between pairs of signals.

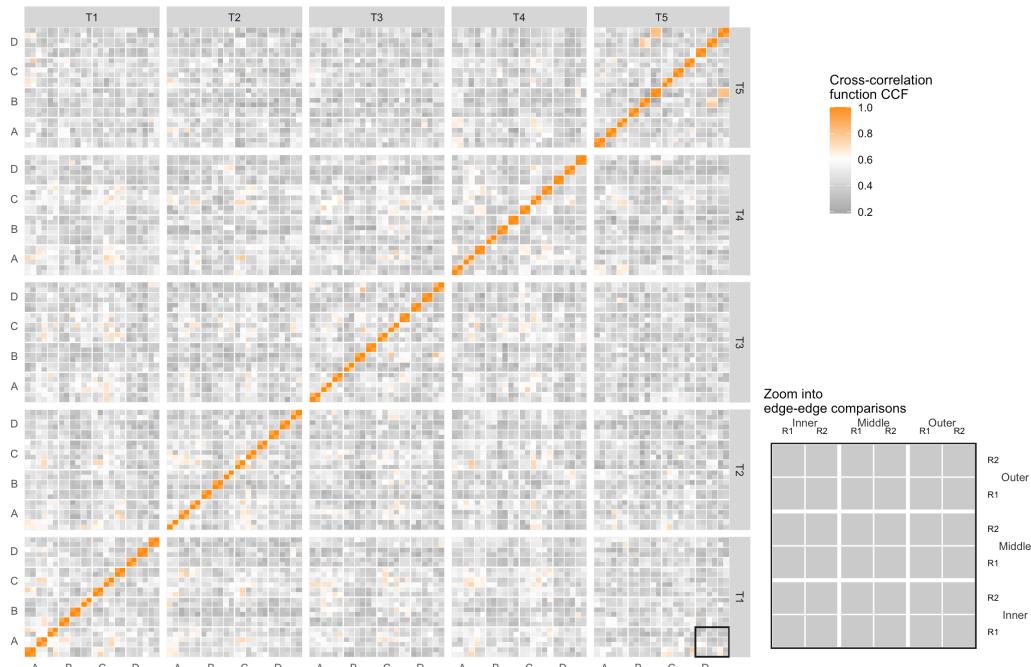


Figure 5. The tilemap shows signals from the same source have CCFs close to 1.

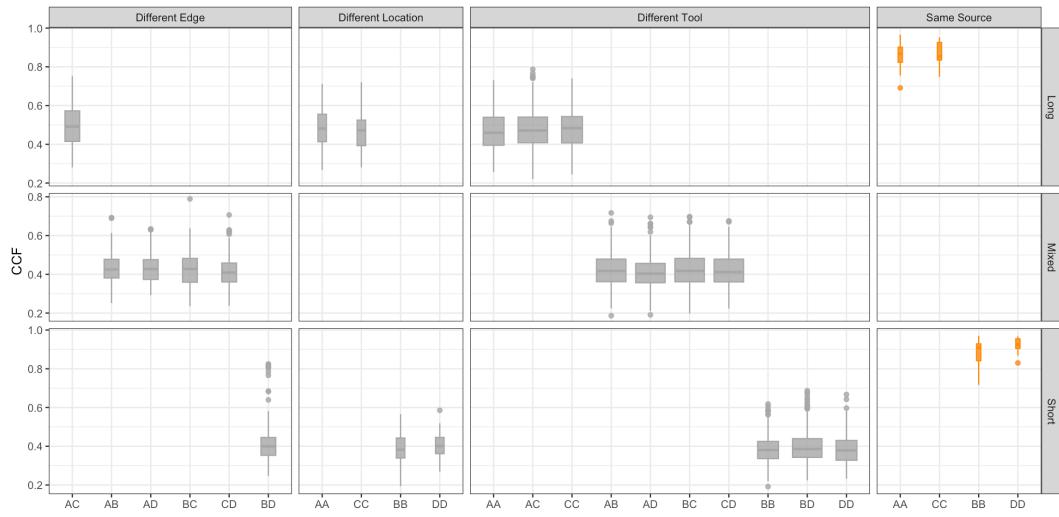


Figure 6. The boxplot shows that signals from the same sources have higher CCFs than those from different sources.

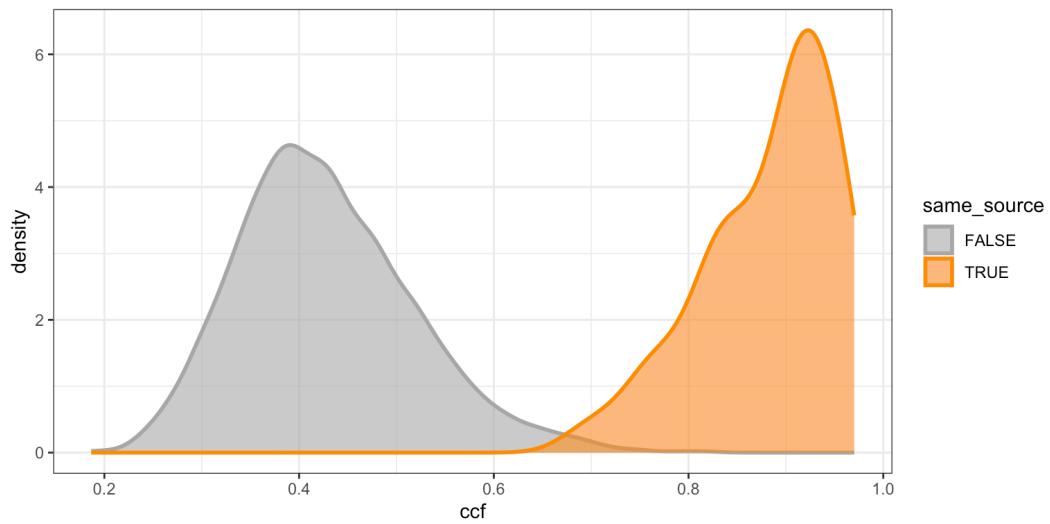


Figure 7. The density plot shows tails of distributions overlap, which can be used as a rough threshold for drawing conclusions.

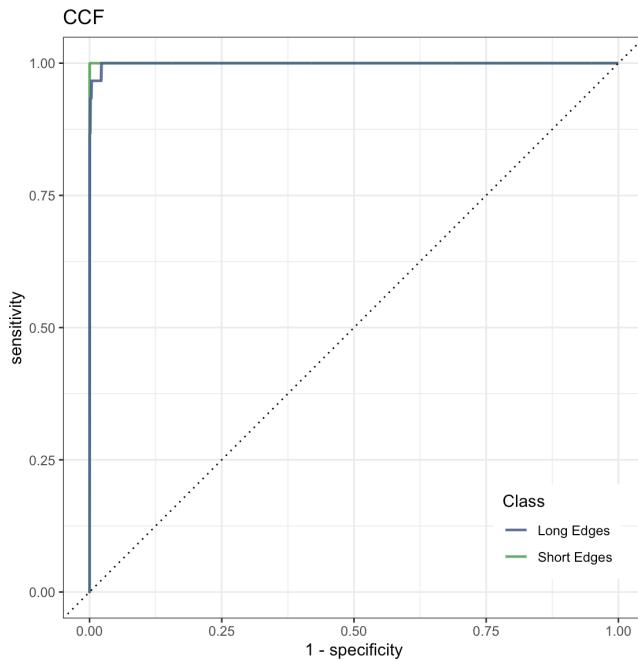


Figure 8. The ROC curve is bending very close to the upper left corner, which means excellent in classification and drawing conclusions.

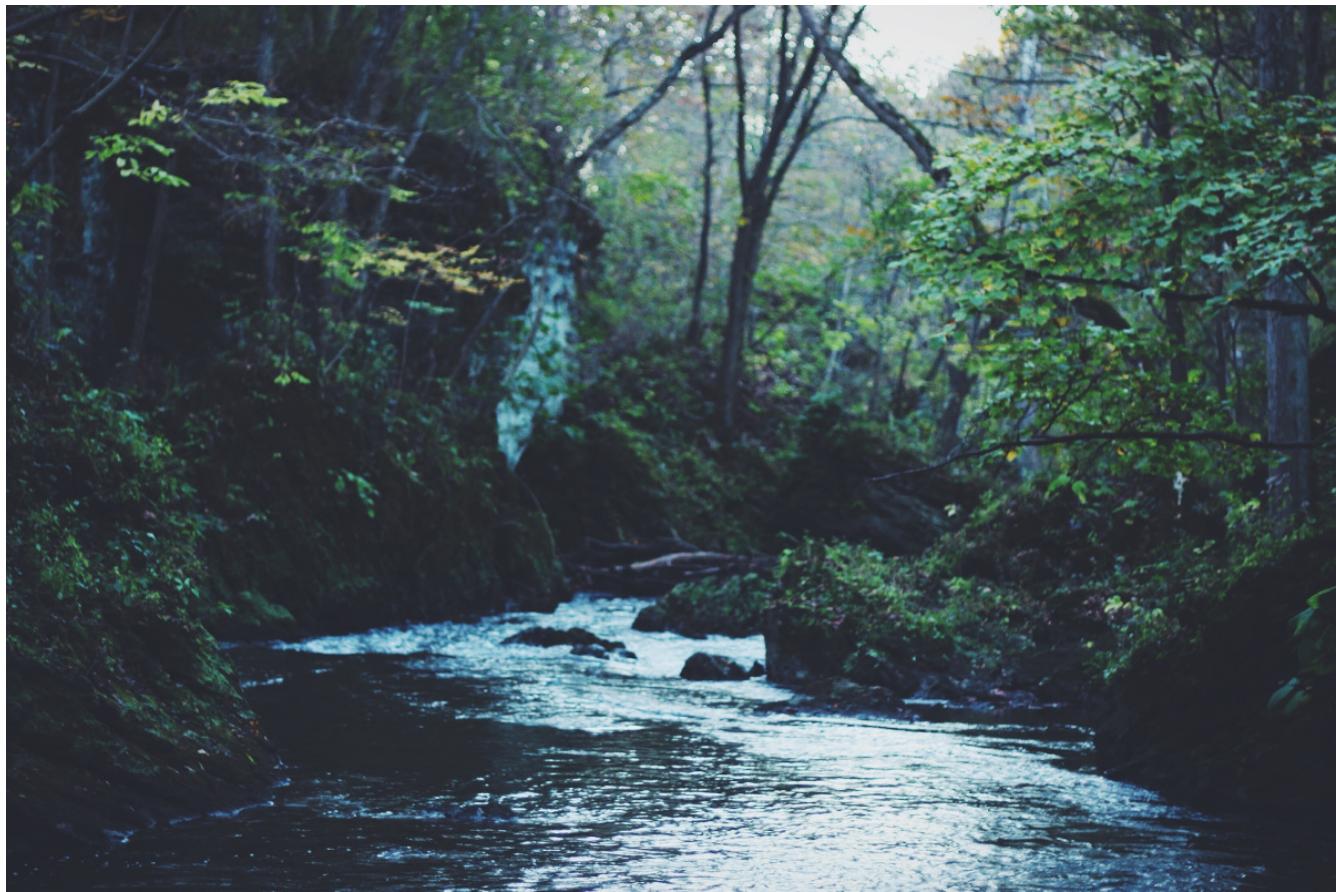


Figure 9. Legend (350 words max). Example legend text.