A Cartridge Case Comparison Pipeline

by

Joseph Zemmels

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Heike Hofmann, Major Professor
Alicia Carriquiry
Kori Khan
Danica Ommen
Richard Stone
Susan VanderPlas

Iowa State University

Ames, Iowa

2022

DEDICATION

dedication text (optional)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Acknowledgements go here.

ABSTRACT

Algorithms to compare evidence are increasingly used in forensic examinations to supplement an examiner's opinion with an objective measure of similarity. However, an algorithm must first be thoroughly tested under various conditions to identify its strengths and weaknesses. This experimentation is expedited for algorithms that are accessible to fellow researchers and practitioners. In this work, we discuss an algorithm to objectively measure the similarity between cartridge cases. We have designed this algorithm to be approachable for researchers and practitioners alike. Chapter 2 discusses a modularization of the algorithm into a "pipeline" that enables reproducibility, experimentation, and comprehension. Our goal in this modularization is to lay a foundation upon which improvements can be easily developed. Chapter 3 details a suite of diagnostic tools that illuminate the innerworkings of the algorithm and determine when and why the algorithm "works" correctly. These diagnostics will be useful for both researchers interested in correcting the algorithm's behavior and for practitioners concerned with applying the algorithm to case work. Chapter 4 introduces novel pieces of the pipeline that we demonstrate are improvements to predominant methods. In particular, we introduce a set of features based on the diagnostic tools discussed in Chapter 3 that effectively measure the similarity between two cartridge cases.

CHAPTER 1. LITERATURE REVIEW

1.1 Preliminaries: Forensic Examinations

A bullet casing is found at the scene of a murder. The bullet is recovered from the victim during autopsy. A handwritten letter threatening the victim is found in their pocket. The assailant's shoeprints are discovered fleeing the area. Who left this evidence? Investigators obtain the gun, shoes, and handwriting samples of a suspect. This evidence, along with the crime scene evidence, is sent to a forensic laboratory for analysis. Forensic examiners compare the evidence to establish whether they share a common source. The suspect is charged after the examiners conclude that there is sufficient agreement between the crime scene and suspect's ~~evidence.~~ samples

The procedure described above, in which evidence is ~~compared~~ analyzed to determine ~~their~~ its origin, is called the *source identification* problem (Ommen and Saunders, 2018). Historically, forensic examiners have relied on tools (e.g., microscopes), case facts, and experience to develop an opinion on the similarity of two pieces of evidence. More recently, algorithms to automatically compare evidence and provide an objective measure of similarity have been introduced. These algorithms are used in a forensic examination to supplement and inform the examiner's conclusion. We propose an automatic, objective solution to the source identification problem; specifically in the context of comparing fired *cartridge cases*. Cartridge case comparison is a sub-discipline of Firearms and Toolmarks Identification, which is reviewed in the next section.

evidence is an uncountable noun--use singular verbs and pronouns

are these supposed to be plural?

1.1.1 Firearms and Toolmarks Identification

Firearms and toolmarks (F & T) identification involves studying markings or impressions left by a hard surface, such as the metal of a firearm or other tool (e.g., screwdriver), on a softer surface (Thompson, 2017). For example, a barrel's rifling leaves toolmarks on a bullet as it travels out of the gun.

1.1.1.1 The Firing Process

In this section, we describe the basic process of firing a handgun or rifle using a cartridge. A *cartridge* consists of a metal casing containing primer, gunpowder, and a bullet. Figure 1.1 shows a cross-section of a cartridge featuring these components (30 Magazine Clip, 2017).

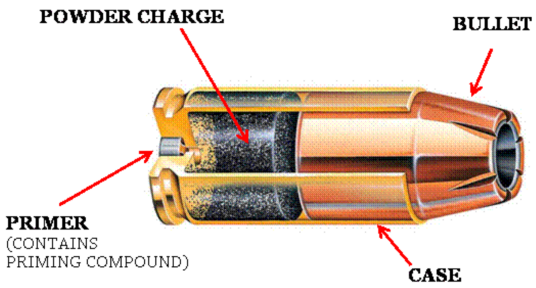

Figure 1.1 A cartridge containing primer, powder, and a bullet. The firing process is initiated by loading a cartridge into the barrel of a firearm.

The first step of the firing process is to load a cartridge into the area in the back of the barrel known as the *chamber*. Figure 1.2 shows an example of a cartridge loaded into the chamber of a pistol (Rattenbury, 2015). In this example, the hammer of the pistol is pulled back such that the firing pin is held back under spring tension. Upon squeezing the trigger, the firing pin is released and travels forwards at a high velocity. The firing pin strikes the primer of the cartridge case, causing it to explode.

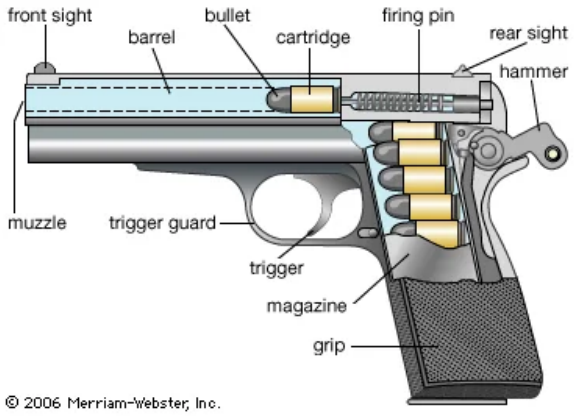

Figure 1.2 Cross-section of a pistol with a chambered cartridge and drawn-back hammer. Pulling the trigger releases the firing pin which strikes the cartridge case primer.

As shown in Figure 1.3, the explosion of the primer ignites the powder in the cartridge (Hampton, 2016). Gas rapidly expands in the cartridge causing the bullet to travel down the barrel. At the same time, the rest of the cartridge is sent towards the back of the barrel.

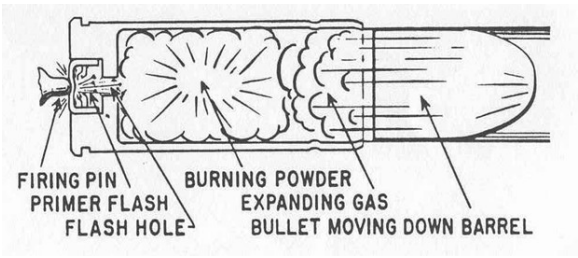

Figure 1.3 A cartridge after a firing pin has struck the primer. The explosion of the primer ignites the powder within the cartridge, causing gas to rapidly expand and force the bullet down the barrel.

As the bullet leaves the barrel, the cartridge case strikes the back wall of the barrel, known as the *breech face*, with considerable force. Any markings on the breech face are imprinted onto the cartridge case, creating the so-called *breech face impressions*. These impressions are analogous to a barrel's "fingerprint" left on the cartridge case. Figure 1.4 shows cartoon examples of breech face markings that appear on cartridge cases (Hampton, 2016).

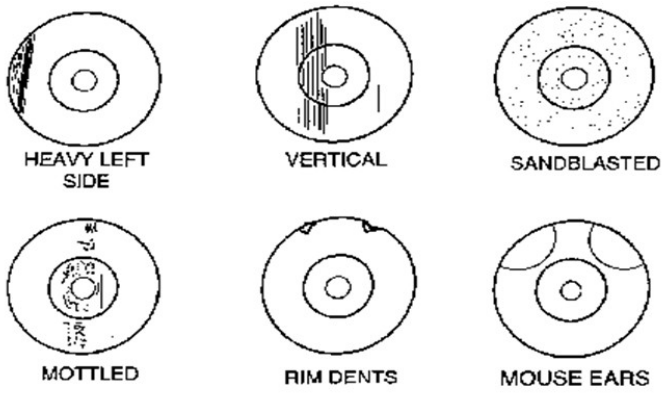

Figure 1.4 Examples of common breech face impression patterns. These are considered analogous to a breech face fingerprint left on the cartridge surface.

Figure 1.5 shows the base of a fired cartridge (Hampton, 2016). The hole to the south-east of the center of the primer is the impression left by the firing pin. Note the horizontal striated breech face markings on the primer to the left of the firing pin impression.

Figure 1.5 A fired 9mm Luger cartridge case with visible firing pin and breech face impressions.

After the bullet has left the barrel, the extractor pin and ejector push the cartridge case out of the chamber. As shown in Figure 1.6, these can leave additional markings on the cartridge (Hampton, 2016). Firing pin, breech face, extractor pin and ejector, and other

possible markings are all used in a forensic examination to determine whether two cartridge cases were fired from the same firearm. This work focuses on the comparison of breech face impressions specifically.

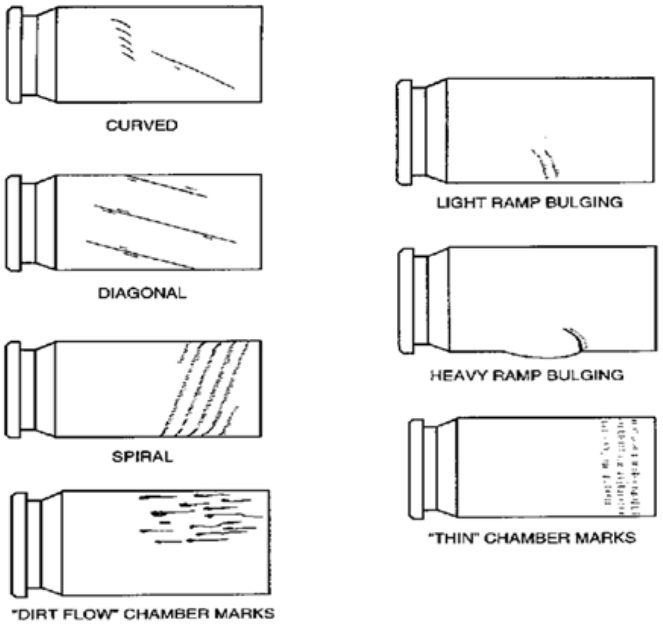

Figure 1.6 Examples of common extractor pin and ejector markings. These, impressions on the cartridge, are used in a forensic examination to determine the source of the fired cartridge.

1.1.1.2 An Overview of Firearms and Toolmarks Examinations

Trained F & T examiners use a *comparison microscope*, such as the one shown in Figure 1.7, to examine two pieces of evidence (Zheng et al., 2014). A comparison microscope consists of two compound microscopes that are joined via an *optical bridge* which allows for viewing of the stages below each microscope simultaneously under the same eyepiece. The right image of Figure 1.7 shows an example of the view under a comparison microscope of two bullets with the white dotted line separating the two fields of view.

Firearm examiners distinguish between three broad categories when characterizing a fired bullet or cartridge case: class, subclass, and individual characteristics. *Class character-*

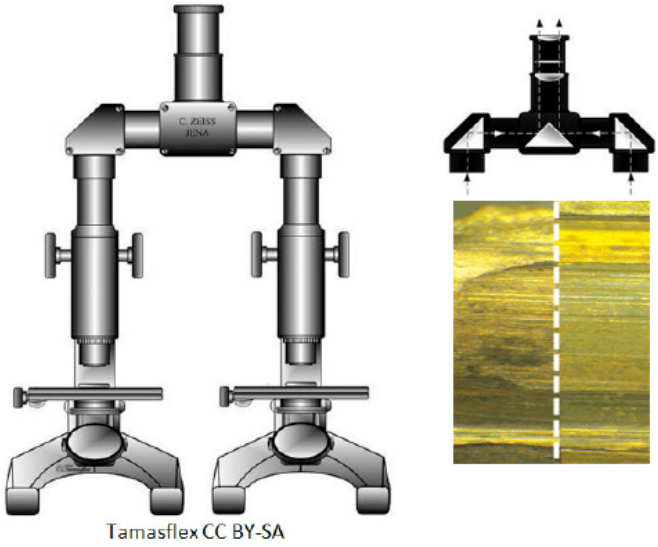

Figure 1.7 A comparison microscope consists of two stages upon which evidence is placed. These stages are placed under two compound microscopes that are joined together via an optical bridge and allow for viewing of both stages simultaneously under a single eyepiece. The image on the right shows an example of a bullet viewed under a comparison microscope.

istics are associated with the manufacturer of the firearm that fired the bullet or cartridge case. These include, but are not limited to, the size of ammunition chambered by the firearm, the orientation of the extractor and ejector, or the width and twist direction of the barrel rifling. Class characteristics are often the first to be examined because they can narrow the relevant population of potential firearm sources (Thompson, 2017). For example, a 9mm cartridge case must have been fired by a firearm that can chamber 9mm ammunition.

If the discernible class characteristics match between two pieces of evidence, for example a cartridge case found at a crime scene and a different cartridge case fired by a suspect's gun, then the examiner uses a comparison microscope to compare the *individual characteristics* of the evidence. Individual characteristics are markings attributed to imperfections on the firearm surface due to the manufacturing process, use, and wear of the tool. For example, markings on the breech face of a barrel may form after repeated fires of the firearm. Individual characteristics are assumed to occur randomly across different

firearms. In an examination, the examiner independently rotates and translates the stages of a comparison microscope to find the optimal matching position of the markings on the two pieces of evidence (Zheng et al., 2014). If the individual characteristics on two pieces of evidence are determined to agree "sufficiently," then the examiner can conclude that they originated from the same firearm (AFTE Criteria for Identification Committee, 1992).

Subclass characteristics exist between the macro-level class and micro-level individual characteristics. These characteristics relate to markings that are reproduced across a subgroup of firearms. For example, breech faces using the same milling machine may include markings that are unique to the milling machine (Werner et al., 2021). As it can be difficult to distinguish between individual and subclass characteristics during an examination, an examiner's decision process may be affected if the existence of subclass characteristics is suspected.

Many F & T examiners in the United States adhere to the Association of Firearms and Toolmarks Examiners (AFTE) Range of Conclusions when making their evidentiary conclusions (AFTE Criteria for Identification Committee, 1992). According to these guidelines, six possible conclusions can be made in a F & T examination:

1. **Identification**: Agreement of a combination of individual characteristics and all discernible class characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool.
2. **Inconclusive**: Formatting a bit weird here.

2.1 Some agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification.

2.2 Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.

2.3 Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination.

3. **Elimination**: Significant disagreement of discernible class characteristics and/or individual characteristics.

4. **Unsuitable**: Unsuitable for examination.

Forensic examinations first involve an examination of a "questioned" bullet or cartridge case for identifiable toolmarks (Thompson, 2017). Markings including breech face, firing pin, chamber marks, extractor pin, and ejector impressions are categorized by their class, individual, or subclass characteristics. If available, this information is compared to "known source" fires obtained from a suspect's firearm. If known source evidence is unavailable, class characteristics from the questioned bullet can be used to narrow the relevant population and provide potential leads. An examiner's decision may be used as part of an ongoing investigation or presented at trial as expert testimony.

Standard operating procedures for assessing and comparing evidence differ between forensic laboratories. For example, some labs collapse the three possible inconclusive decisions into a single decision (Neuman et al., 2022) or prohibit examiners from making an elimination based on differences in individual characteristics (Duez et al., 2017).

1.1.2 Why Should Firearms and Toolmarks Identification Change?

In 2009, the National Research Council released a report assessing a number of forensic disciplines including Firearms and Toolmarks analysis. The report pointed out that F & T analysis lacked a precisely defined process and that little research had been done to determine the reliability or repeatability of the methods. Two of the recommendations from this

when do you define these things?

study were to establish rigorously-validated laboratory procedures and "develop automated techniques capable of enhancing forensic technologies (National Research Council, 2009)."

A number of studies assess the reliability and repeatability of a firearms and toolmarks examination (non-exhaustively: DeFrance and Arsdale (2003); Hamby et al. (2009); Fadul et al. (2011); Stroman (2014); Baldwin et al. (2014); Smith et al. (2016); Mattijssen et al. (2020)). These studies indicate that examiners have a low error rate when comparing evidence obtained under controlled conditions (i.e., for which ground-truth is known). However, as pointed out in a 2016 report from the President's Council of Advisors on Science and Technology, many of these studies, save Baldwin et al. (2014), were not "appropriately designed to test the foundational validity and estimate reliability (President's Council of Advisors on Sci. & Tech., 2016)." The report asserts that additional, properly-designed studies should be performed to more rigorously establish the scientific validity of the discipline.

Due to the opacity in the decision-making process, examiners are referred to as "black boxes" in a similar sense to black-box algorithms (OSAC Human Factors Committee, 2020). Their evidentiary conclusions are fundamentally subjective, and empirical evidence suggests that conclusions differ when examiners are presented with the same evidence on different occasions (Ulery et al., 2011, 2012). Examiners rarely need to provide quantitative justification for their conclusion. Even for qualitative justifications, it can be difficult to determine what the examiner is actually "looking at" to arrive at their conclusion (Ulery et al., 2014). This suggests the need to supplement these black box decisions with transparent, objective techniques that quantitatively measure the similarity between pieces of evidence. As stated in President's Council of Advisors on Sci. & Tech. (2016), efforts should be made to "convert firearms analysis from a subjective method to an objective method" including "developing and testing image-analysis algorithms for comparing the similarity of tool marks." This work focuses on the development of an algorithm for comparing breech face impressions on cartridge cases.

1.2 Forensic Comparison Pipelines

Recent work in many forensic disciplines has focused on the development of algorithms to measure the similarity between pieces of evidence including glass (Curran et al., 2000; Park and Tyner, 2019; Tyner et al., 2019), handwriting (Crawford, 2020), shoe prints (Park and Carriquiry, 2020), ballistics (Hare et al., 2017; Tai and Eddy, 2018), and toolmarks (Hadler and Morris, 2017; Krishnan and Hofmann, 2018). These algorithms often result in a numerical, non-binary (dis)similarity score for two pieces of evidence. A non-binary score adds more nuance to an evidentiary conclusion beyond simply stating whether the evidence originated from the same source as would be the case in binary classification. For example, the larger the similarity score, the "more similar" the evidence. However, a binary (or ternary, if admitting inconclusives) conclusion must ultimately be reached by an examiner. Whether a decision should be reached based solely on results of a comparison algorithm (e.g., defining a score-based decision boundary) or if an examiner should incorporate the similarity score into their own decision-making process is still up for debate (Swofford and Champod, 2021). We view forensic comparison algorithms as a supplement to, rather than a replacement of, the forensic examination.

Forensic comparison algorithms are treated as evidence-to-classification "pipelines." Broadly, the steps of the pipeline include:

1. capturing a digital representation of the evidence,
2. preprocessing this representation to isolate or emphasize a region of interest of the evidence,
3. comparing regions of interest from two different pieces of evidence to obtain a (perhaps high-dimensional) set of similarity features,
4. combining these features into a low-dimensional set of similarity scores, and

5. defining a classification rule based on these similarity features.

Is the above actually taken from Rice? It's a bit unclear to me

This is similar to the structure discussed in Rice (2020). We add to this structure the emphasis that each step of the pipeline can be further broken-down into modularized pieces. For example, the preprocessing step may include multiple sub-procedures to isolate a region of interest of the evidence. Figure 1.8 shows two possible variations of the cartridge case comparison pipeline as well as the parameters requiring manual specification and alternative modules. The benefits of this modularization include easing the process of experimenting with different parameters/sub-procedures and improving the comprehensibility of the pipeline.

In the following sections, we detail recent advances to each of the five steps in the pipeline outlined above. We narrow our focus to advances made in comparing firearms evidence.

1.2.1 Digital Representations of Evidence

Digital representations of cartridge case evidence commonly come in one of two modes: 2D optical images or 3D topographic scans. A common way to take 2D optical images is to take a picture of the cartridge case under a microscope. This implies that the digital representation of the cartridge case surface is dependent on the lighting conditions under which the picture was taken. Some recent work has focused on comparing 2D optical images (Tai and Eddy, 2018; Tong et al., 2014), although the use of 3D microscopes has become more prevalent to capture the surface of ballistics evidence.

Using a 3D microscope, scans are taken at the micron (or micrometer) level that are more light-agnostic than a 2D image (Weller et al., 2012). One common 3D scanning procedure is disc scanning confocal microscopy. This procedure works by shining a focused beam of light on the cartridge case surface. This light is reflected back onto a pinhole allowing a limited height range to pass through. The microscope scans through different height range

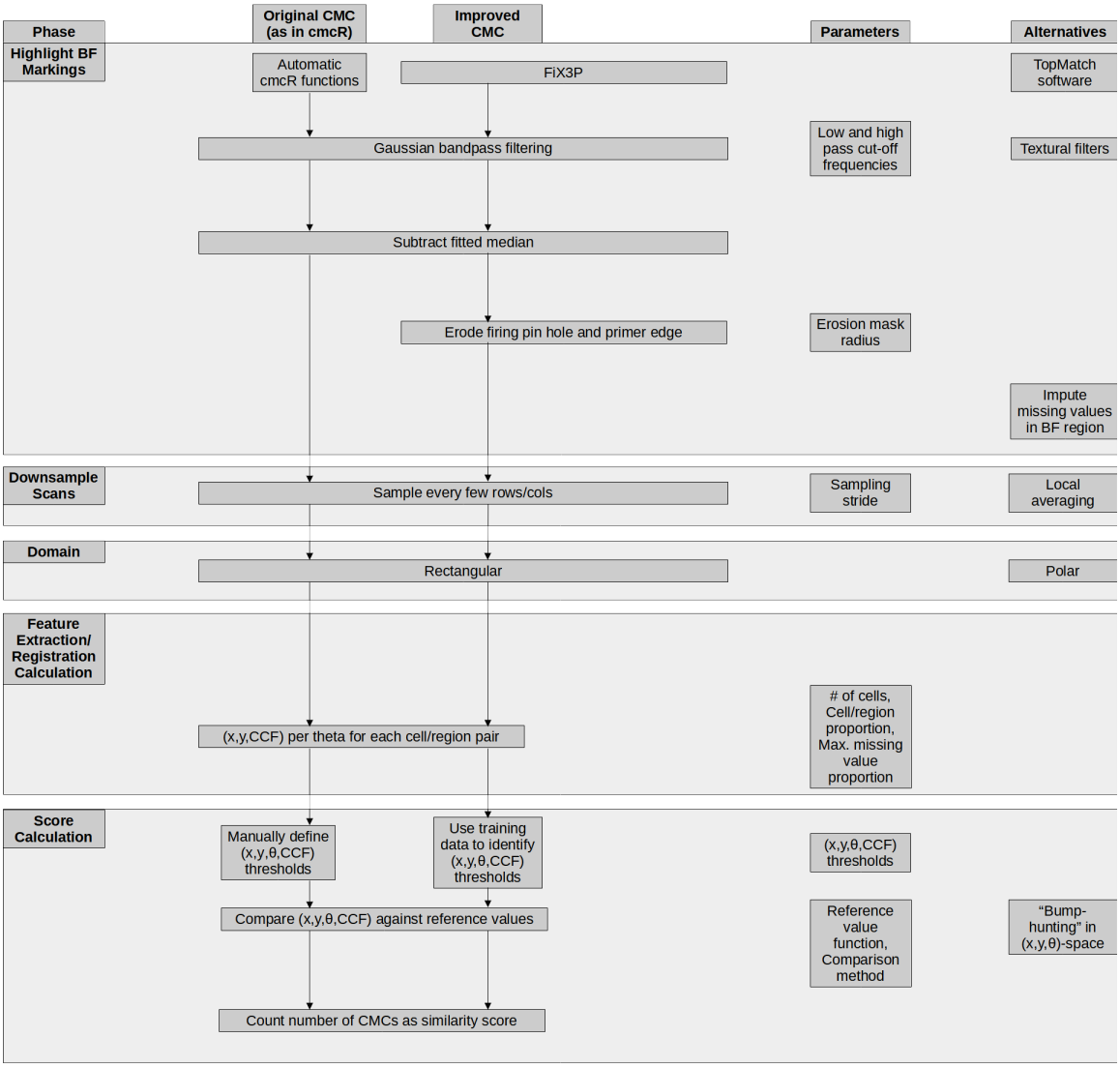

Figure 1.8 Variations upon the cartridge case comparison pipeline. The first two columns detail the pipeline with different sub-procedures. The third columns shows the parameters that require manual specification at each step. The fourth column shows alternative processing steps that could replace steps in the existing pipeline.

"slices" and compiles all these slices into a single 3D topography of the cartridge case primer surface. The Microdisplay Scan Confocal Microscope from Sensofar™~Metrology is shown in Figure 1.9 (Bermudez et al., 2017).

Figure 1.9 The Microdisplay Scan Confocal Microscope from Sensofar™ Metrology. The cartridge case surface is captured by scanning through a range of vertical slices and compiling these slices into a single 3D topography.

Figure 1.10 shows a 2D image and 3D topography of the same cartridge case primer from Fadul et al. (2011).

More recently, Cadre Forensics™~introduced the TopMatch-3D High-Capacity Scanner (Weller et al., 2015). A tray of 15 fired cartridge cases and the scanner are shown in Figure 1.11 (Cadre Forensics, 2019). This scanner collects images under various lighting conditions of a gel pad into which the cartridge case surface is impressed and combines these images into a regular 2D array called a *surface matrix*. The physical dimensions of these objects are about 5.5 mm^2 captured at a resolution of 1.84 microns per pixel (1000 microns equals 1 mm).

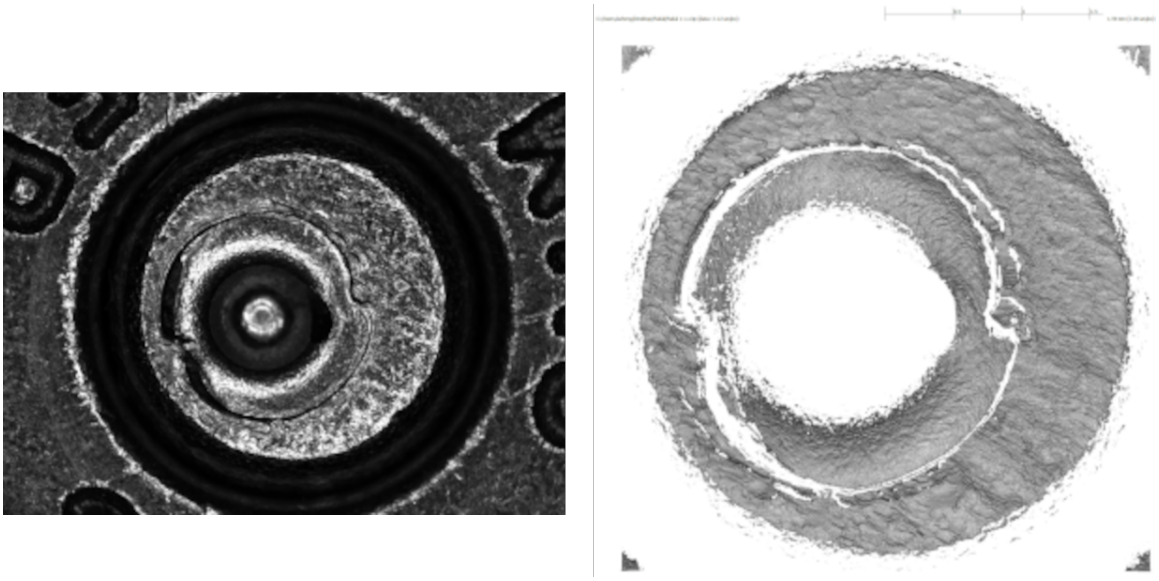

Figure 1.10 A cartridge case captured using 2D confocal reflectance microscopy (left) and 3D disc scanning confocal microscopy (right).

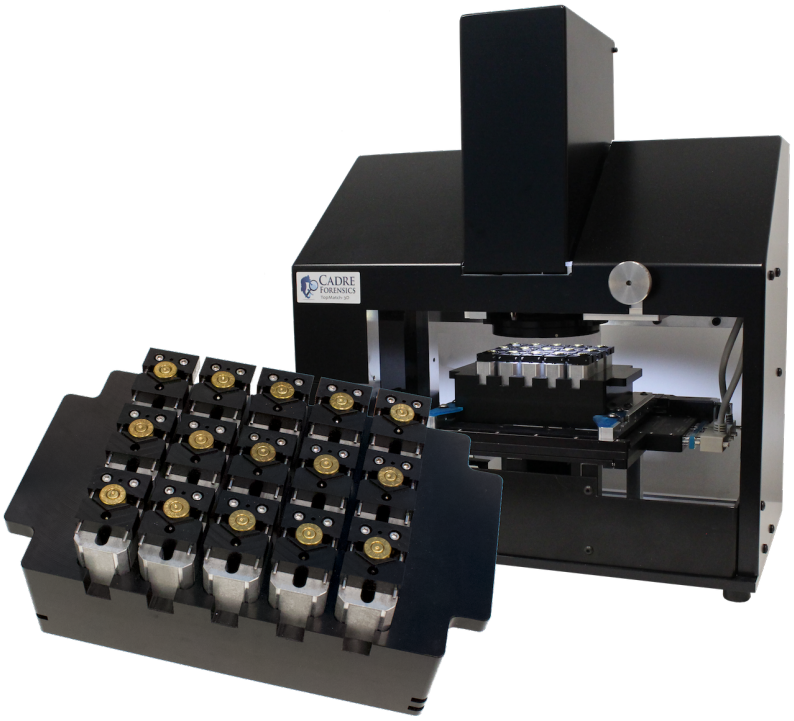

Figure 1.11 The TopMatch-3D High-Capacity Scanner from Cadre Forensics™ . The scanner captures topographic scans of a gel pad into which a cartridge case surface is impressed.

When applied to ballistics evidence, these 3D scans are commonly stored in the ISO standard x3p file format (ISO 25178-72(2017), 2017). x3p is a container consisting of a single surface matrix representing the height value of the surface and metadata concerning the parameters under which the scan was taken as shown in Figure 1.12 (Zheng et al., 2020). It has been empirically demonstrated that comparing 3D topographic scans of cartridge case evidence leads to more accurate conclusions compared to comparing 2D optical images of the same evidence (Tai, 2019; Tong et al., 2014; Song et al., 2014).

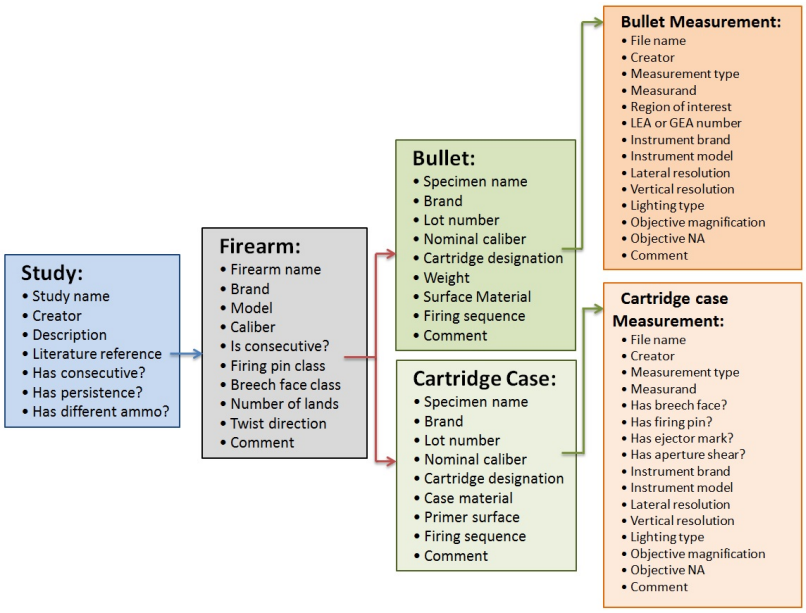

Figure 1.12 The hierarchy of information stored in the x3p file format for both bullet and cartridge case evidence.

1.2.2 Preprocessing Procedures for Forensic Data

Took me a few reads to understand this sentence. maybe reword?

When capturing the surface of a cartridge case, the result is bound to contain extraneous regions due to the incongruity between the circular primer and the rectangular array in which the surface data are stored. Figure 1.10 shows an example of a 2D image and 3D scan of the same cartridge case. The corners of these arrays include non-primer regions of the cartridge case surface. Additionally, the center of the cartridge case primer features an

impression left by the firing pin during the firing process. In most applications, impressions left by the firing pin are compared separately from the breech face impressions (Zhang et al., 2016). Because we are interested in the comparison of breech face impressions between two cartridge cases, only the annular region surrounding the firing pin impression is of interest. The annular breech face impression region must be segmented away from the rest of the captured surface.

Both the 2D optical and 3D topographic representations of cartridge case surfaces are fundamentally pictorial in nature. As such, many image processing and computer vision techniques are used to automatically isolate the breech face impression region. Tai and Eddy (2018) uses a combination of histogram equalization, Canny edge detection, and morphological operations to isolate the breech face impressions in 2D images. Various types of Gaussian filters are commonly employed to remove unwanted structure. Tong et al. (2014) uses a low-pass Gaussian filter that removes noise via a Gaussian-weighted moving average operation. Chu et al. (2013), Song et al. (2018) use a bandpass Gaussian filter, which simultaneously performs the function of a low-pass filter along with a high-pass filter to remove global structure from the scan. Other versions of the bandpass filter are used in Song et al. (2014); Chen et al. (2017); Ott et al. (2017) that accomplish tasks such as omitting outlier surface values or addressing boundary effects (ISO 16610-71(2014), 2014; Brinkman and Bodschwinna, 2003a).

Instead of using automatic procedures, others have used subjective human intervention to isolate the breech face impressions. For example, (Song et al., 2018) indicate that cartridge cases are "manually trimming to extract the breech face impression of interest." In Roth et al. (2015), examiners manually identify the borders of the breech face impression region by placing points around an image of the cartridge case primer.

1.2.3 Forensic Data Feature Extraction

After applying the preprocessing procedures to two cartridge case scans, ~~their breech face impressions are compared and~~ similarity features are extracted. ~~Given that~~ the cartridge cases at this point are represented as high-dimensional matrices, this can be thought of as a dimensionality reduction of the high-dimensional surface arrays to a set of similarity statistics.

from the breech face impressions?
Because
process

A variety of features have been proposed to quantify the similarity between two cartridge case surface arrays. Tai and Eddy (2018) propose calculating the cross-correlation function (CCF) value between two cartridge cases across a grid of rotations. It is assumed that the CCF will to be larger around the "true" rotation for matching cartridge case pairs than for non-matching pairs. Riva and Champod (2014) proposed combining the CCF between the two aligned scans with the element-wise median Euclidean distance and median difference between the normal vectors at each point of the surface. Later, Riva et al. (2016, 2020) applied Principal Component Analysis to reduce these three features down to two principal components onto which a 2D Kernel Density Estimator could be fit.

Pertinent to this work is the cell-based comparison procedure originally outlined in Song (2013). The underlying assumption of Song (2013) is similar to that of Tai and Eddy (2018): that two matching cartridge cases will exhibit higher similarity when they are "close" to being correctly aligned. While Tai and Eddy (2018) measured similarity using the CCF between the two full scans, Song (2013) proposes partitioning the scans into a grid of "correlation cells" and counting the number of similar cells between the two scans. The rationale behind this procedure is that many cartridge case scans have only a few regions with discriminatory markings. As such, comparing full scans may result in a lower correlation than if one were to focus on the highly-discriminatory regions. In theory, dividing the scans into cells allows for the identification of these regions.

After breaking a scan into a grid of cells, each cell is compared to the other scan to identify the rotation and translation, known together as the *registration*, at which the cross-correlation is maximized. Song (2013) assume that the cells from a truly matching pair of cartridge cases will "agree" on their registration in the other scan. Details of this procedure are provided in Chapter 2.

1.2.4 Similarity Scores for Forensic Data

Following feature extraction, the dimensionality of these features is further reduced to a low-dimensional, usually univariate, similarity score.

I would provide a brief summary of the types of approaches/scores you will discuss in ths section. Just a sentence to give the reader an idea of what is coming

After calculating the CCF across various possible registrations, Tai and Eddy (2018) propose using the maximum observed CCF value as the univariate similarity score. In this case, a binary classification can be achieved by setting a CCF threshold above which pairs are classified as "matches" and below which as "non-matches." Tai (2019) proposes setting a CCF cut-off that maximizes the precision and recall in a training set of pairwise comparisons.

Riva et al. (2016, 2020) use a training set to fit two 2D kernel density estimates to a set of features from matching and non-matching comparisons. Using these estimates, they are able to estimate the score-based likelihood ratio (SLR) for a new set of features. This SLR can be viewed as a similarity score (Garton et al., 2020).

I would re-order these sentences. You are referencing "congruent matching" before introducing it. Also, it's been a bit since you introduced the cell-based comparisions, so it might be useful to reference Song etc again. The reference to "above" was a bit confusing to track..

In the case of the cell-based comparison procedure discussed above, the total number of cells that are deemed "congruent matching" is used as a similarity score. The criteria used to define "congruent matching" has changed across papers (Song et al., 2014; Tong et al., 2014, 2015; Chen et al., 2017) and will be discussed in greater detail in Chapter 2. The authors of these papers have consistently used a decision boundary of six "Congruent Matching Cells" to distinguish matches from non-matches.

Zhang et al. (2020) applies the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al., 1996) to the features from the cell-based comparison procedure to determine if any clusters form amongst the per-cell estimated registration values. This is based on the assumption that any cells that come to a consensus on their registration should form a cluster in translation (x, y) and rotation θ space. Zhang et al. (2020) proposes a binary classifier based on whether any clusters are identified by the DBSCAN algorithm (Ester et al., 1996). If a cluster is found for a particular pairwise comparison, then that pair is classified as a "match" and otherwise as a "non-match."

You seem to have collapsed steps 4 and 5 from pages 10-11 here. I would make that more clear and then provide some sort of transition to the next sub-section. It's a bit hard to follow where you are going as a reader.

1.2.5 Reproducibility of Comparison Pipelines

National Academy of Sciences, Engineering, and Medicine (2019) defines reproducibility as "obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis." While not exact in their definition of "consistent," the authors assert that, barring a few exceptions, it is reasonable to expect that the results obtained by a second researcher, after applying the exact same processing steps to the exact same data, be the exact same as the original results. Among the exceptions given is if the original researcher had made a mistake in writing the original source code. In either case, they assert that "a study's data and code have to be available in order for others to reproduce and confirm results." Researchers can easily verify the results given data and code and incorporate the materials into their own research and thus improve or accelerate discovery (Stodden et al., 2018a).

:).. in theory anyway

A number of studies indicate that computationally reproducible research is sparse across various disciplines. Stodden et al. (2018b) and Stodden et al. (2018a) studied the reproducibility of articles sampled from the journals *Science* and the *Journal of Computational Physics*, respectively. In the former, Stodden et al. (2018b) found that only 3 of 204 randomly selected articles from *Science* were "straightforward to reproduce with minimal effort;" despite a journal policy requiring that all code and data used in the paper be made

available to any reader. In the latter, Stodden et al. (2018a) found that zero of 306 randomly selected articles from the *Journal of Computaional Physics* were "straightforward to reproduce with minimal effort" and, at best, that five articles were "reproducible after some tweaking." Similar findings were found in Chang and Li (2022) (29 of 59 economic papers reproducible), Iqbal et al. (2016) (zero of 268 biomedical papers provided raw data and 1 in 268 linked to a full study protocol), Duvendack et al. (2015) (50% or more published articles include data or code in only 27 of 333 economics journals), and Gundersen et al. (2018) (24 of 400 AI conference papers included code). A common recommendation amongst these authors is the establishment of rigorous tools and standards to promote reproducibility. This includes making code and data used in a paper easily-accessible to readers.

Infrastructure already exists to ease the processing of developing, maintaining, and sharing open-source code and data. Data repositories such as the NIST Ballistics Toolmark Research Database (Zheng et al., 2020) provide open access to raw data. Grüning et al. (2018) discuss the use of package managers such as Conda (ana, 2020), container software such as Docker (https://www.docker.com/), and virtual machine software to preserve the entire data analysis environment in-perpetuity. For situations in which VMs or containers aren't available, software such as the `manager` R package allows users to "compare package inventories across machines, users, and time to identify changes in functions and objects (Rice, 2020)." Piccolo and Frampton (2016) reference repositories like Bioconductor (Huber et al., 2015) that make it easy to document and distribute code. Further, software such as the `knitr` R package (Xie, 2014) enable "literate programming" in which prose and executed code can be interwoven to make it easier to understand the code's function. These tools make data, code, and derivative research findings more accessible, in terms of both acquisition and comprehensibility, to consumers and fellow researchers.

1.3 Diagnostic Tools

Forensic examiners often provide expert testimony in court cases. As part of this testimony, an examiner is allowed to provide facts about the outcome of a forensic examination and their opinion about what the results mean. A party to a court may challenge the examiner on the validity of the underlying scientific method or whether they interpreted the results correctly (American Academy of Forensic Sciences, 2021). In these situations, examiners need to explain the process by which they reached an evidentiary conclusion to the fact finders of the case; namely, the judge or jury. As algorithms are more often used in forensic examinations, the technical knowledge required to understand and explain an algorithm to lay-people has increased. While in some cases the authors of the algorithm have been willing to provide testimony to establish the validity of the algorithm (Indiana County Court of Common Pleas, 2009), this will become less viable as algorithms become more prevalent. Indeed, even the most effective algorithms may be moot if an examiner can't explain the algorithm in their testimony.

The resources required to educate examiners on the use of highly technical algorithms makes additional training seem currently implausible. An alternative is to develop algorithms from the ground-up to be intuitive for examiners to understand and explain to others. *Explainability* refers to the ability to identify the factors that contributed to the results of an algorithm (Belle and Papantonis, 2021). For example, understanding "why" a classifier predicted one class over another. Diagnostic tools improve the explainability of a model.

Myriad diagnostic tools exist to explain the results of an algorithm. These range from identifying instances of the training set that illuminate how the model operates (Deng, 2018) to fitting more transparent models that approximate the complex model accurately (Puiutta and Veith, 2020) to explaining the behavior of the algorithm in a small region of

interest (Ribeiro et al., 2016; Goode and Hofmann, 2021). Many of these methods require additional technical knowledge to interpret these explanations.

1.3.1 Visual Diagnostics

A less technical approach is to use visualizations that facilitate understanding of model behavior. Properly constructed visuals enable both exploratory data analysis and diagnostics (Buja et al., 2009), which are critical steps in the data analysis process for anticipating and assessing model fit. Given that many of the procedures by which cartridge case evidence is captured, processed, and compared are based on image processing techniques, a visual diagnostic is an intuitive mode of explanation for researchers and lay-people alike. As stated in Cleveland (1994), "graphical methods tend to show data sets as a whole, allowing us to summarize the behavior and to study detail. This leads to much more thorough data analyses."

Numerical statistics summarize the behavior of data, but miss the detail referenced in Cleveland's quote (Telea, 2014). To illustrate this, consider the famous data sets from (Anscombe, 1973) known as Anscombe's quartet. The two variables in each data set are plotted against one another in Figure 1.13. There are clear differences in the relationship between `x` and `y` across these four data sets.

Despite these differences, Table 1.1 demonstrates that summary statistics, namely the first two moments, are identical. This demonstrates that relying on summary statistics (at least low-order moments) can lead to incorrect assumptions about a data set's behavior.

Data Set	$\bar{x}$	S.D. x	$\bar{y}$	S.D. y
1	9	3.32	7.5	2.03
3	9	3.32	7.5	2.03
3	9	3.32	7.5	2.03
4	9	3.32	7.5	2.03

Table 1.1 Moments of the two variables in Anscombe's quartet.

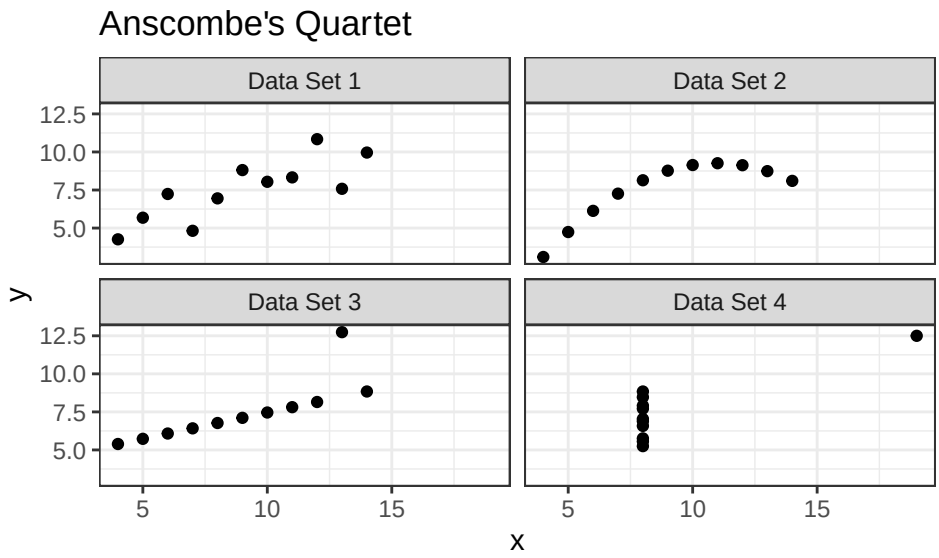

Figure 1.13 A visualization of Anscombe's quartet. Despite there being obvious differences between these four data sets, their summary statistics are nearly identical

Given the pivotal role that visual diagnostics play in the data analysis pipeline, we now consider best practices in creating data visualizations. Human brains are wired for seeing patterns and differences, and for understanding spatial relationships from this (Telea, 2014). As such, an effective visual diagnostic, or data visualization in general, is one that conveys patterns quickly and easily, and with minimal scope for understanding. Arising originally from a pyschological theory of perception, the Gestalt Laws of Perceptual Organization (Goldstein and Brockmole, 2016) are relevant to the construction of statistical graphics. The Gestalt laws are as follows:

- **Pragnanz - the law of simplicity:** Every stimulus pattern is seen in such a away that the resulting structure is as simple as possible.
- **Proximity:** Things that are near each other appear to be grouped together.

- **Good Continuation:** Points that, when connected, result in straight or smoothly curving lines are seen as belonging together, and the lines tend to be seen in such a way as to follow the smoothest path.
- **Similarity:** Similar things appear to be grouped together.
- **Common Region:** Elements that are within the same region of space appear to be grouped together.
- **Uniform Connectedness:** A connected region of visual properties, such as the lightness, color, texture, or motion, is perceived as a single unit.
- **Synchrony:** Visual events that occur at the same time are perceived as belonging together.
- **Common Fate:** Things that are moving in the same direction appear to be grouped together.
- **Familiarity:** Things that form patterns that are familiar or meaningful are likely to become grouped together.

These laws provide guidance on how to construct a visual that concisely conveys a pattern or difference in data. For data visualization, additional laws include (Midway, 2020):

- **Use and Effective Geometry:** Choose a geometry (shape and features of a statistical graphic) that is appropriate to the data.
- **Colors Always Mean Something:** Colors in visuals can convey groupings or a range of values.

As an example of these principles in-action is shown in Figure 1.14. The plot shows the weight over time of chicks fed one of two experimental diets (from the `ChickWeight` base

R data set) (Crowder and Hand, 1990). Individual points represent the weight of a single chick on a particular day. Each set of collected points represents the weight for a single chick over time. This is an example of using an effective geometry (point & line graph to represent time series) along with the Gestalt law of Good Continuation. We further apply the Gestalt law of Common Region by faceting the plot by diet[1]. This implicitly communicates to the audience that the weights of two diet groups of chicks is expected to differ. Indeed, appealing to the Gestalt law of Uniform Connectedness, the "motion" of the grouped time series suggests that chicks given Diet 2 tend to gain weight more rapidly than those given Diet 1. This may suggest a particular modeling structure for these time series (e.g., diet fixed effect) or the need to assess the experimental design to ensure that the assumption that the chicks were randomly sampled from the same population is appropriate. We see how such a plot can be used for both exploratory data analysis or as a post-hoc diagnostic tool. Alternative to faceting, the time series from these two diet groups could be combined into a single plot and distinguished by color.

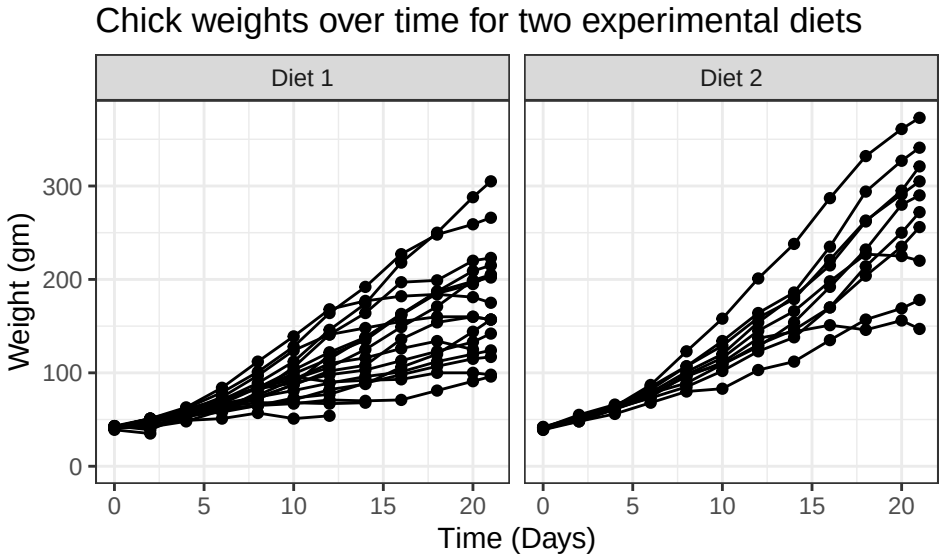

Figure 1.14 An example of using the ggplot2 package to construct a residual plot from a simple linear regression. The features of the statistical graphic are combined layer-by-layer using the + operator.

[1]Note that the encoding for Diet 1 and Diet 2 in the original data set was "1" and "3"

The R programming language (R Core Team, 2017) provides a variety of tools to create visual diagnostics. Among the most robust of these tools is the ggplot2 package (Wickham, 2009). This package extends the "Grammar of Graphics" introduced in Wilkinson (2005) that provides a user-friendly structure by which graphics can be created. Features of a statistical graphic (e.g., elements, transformations, guides, labels) are individually "layered" on a blank canvas using the `+` operator. An example of constructing a residual plot from the `attitude` base R data set is shown in 1.15 (Chatterjee and Hadi, 2006). Such a diagnostic allows the analyst or audience to determine whether the homoscedasticity or linear form assumptions underlying simple linear regression are met. For those willing to learn the "grammar," the code used to create these statistical graphics can easily be re-used and tweaked to fit a specific application.

Is the code supposed to be here? maybe label it as a figure or something with a caption?

```
lmFit <- lm(formula = rating ~ complaints,data = datasets::attitude)


library(ggplot2)


ggplot(data = data.frame(Complaints = datasets::attitude$complaints,
                         Residuals = lmFit$residuals)) +
  geom_point(aes(x = Complaints,y = Residuals)) +
  geom_hline(yintercept = 0,linetype = "dashed") +
  labs(x = "% in-favor of handling of employee complaints")
```

1.3.2 Interactive Diagnostics

While the `ggplot2` package eases the process of constructing visual diagnostics, software such as the `shiny` R package (Chang et al., 2021) enables the consumer of the diagnostic to interact with the visualizations and underlying data. The `shiny` package provides tools for using R to build web applications run on HTML, CSS, JavaScript. Among other

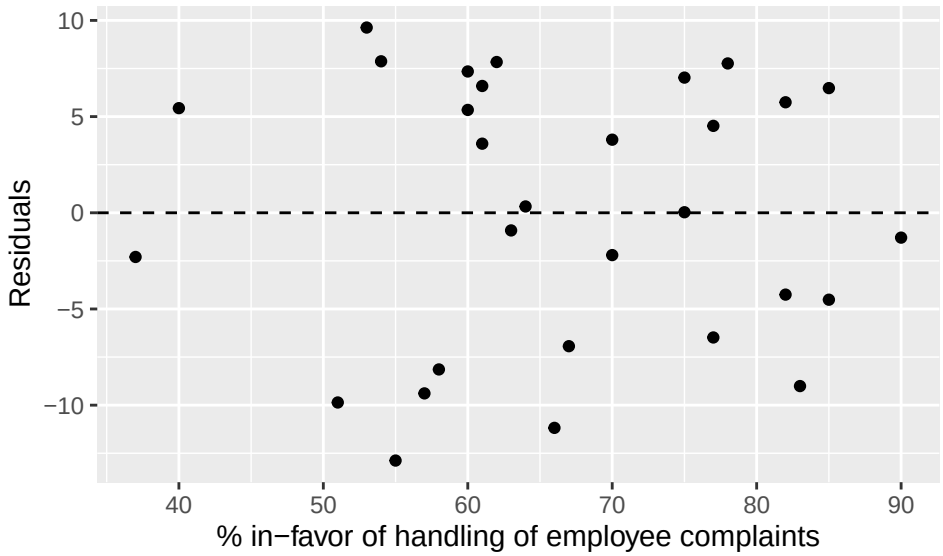

Figure 1.15 An example of using the ggplot2 package to construct a residual plot from a simple linear regression. The features of the statistical graphic are combined layer-by-layer using the + operator.

functionality, these applications allow users to upload or create their own data, set parameters for an analysis, interact with visualizations or data sets (e.g., by hovering to display a tooltip), and export their analyses in various file formats (Beeley and Sukhdeve, 2018). This can be extremely helpful for encouraging non-experts to engage with an analysis pipeline that otherwise may be technically or conceptually inaccessible. Rather than answering a question posed by the author of a plot as a static plot does, such interactive diagnostic tools enable the audience to formulate and answer their own questions. This leads to deeper engagement with the data (Telea, 2014).

Figure 1.16 (Wang et al., 2021) shows a screenshot of the IPDmada shiny application that enables users to perform a meta-analysis of diagnostic test accuracy studies at the individual patient level (an individual patient data meta-analysis or IPD-MA) using a variety of statistical techniques. As seen in 1.16, the user can upload their own data csv file and select parameters that will enable the importing of the data. The other tabs at the top of the application provide statistical tools to analyze the uploaded data. This application is

useful for researchers who are interested in analyzing diagnostic test accuracy data, yet do not necessarily have the coding skills to perform such an analysis in R themselves.

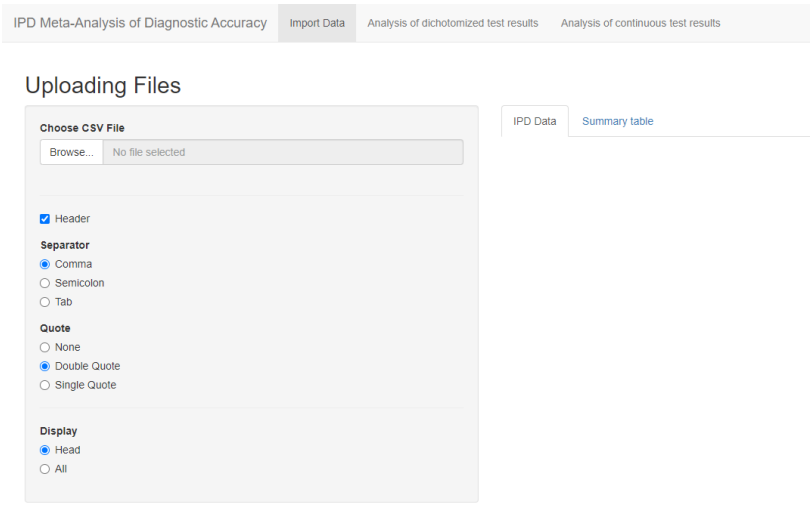

Figure 1.16 The IPDmada shiny application allows users to analyze individual patient data from a diagnostic test accuracy study using a variety of statistical techniques.

Several recently-released software provide interactive diagnostic applications for firarms and toolmarks evidence. Most notable of these software is the Virtual Comparison Microscopy application from Cadre Forensics™. In contrast to Light Comparison Microscopy (LCM; e.g., using comparison microscopes), this software displays 3D topographies saved on a computer (e.g., as an x3p file). An example screenshot of this software is shown in Figure 1.17 (Chapnick et al., 2020). We see that two cartridge case scans are selected from the user's computer for comparison. This user has selected various colors to represent similar or different regions. A major benefit of using VCM over LCM is that these 3D scans can be shared over the internet rather than sending the physical specimen to another lab (which may damage the specimen). Duez et al. (2017), Chapnick et al. (2020), and Knowles et al. (2021) all demonstrate that performing forensic examinations using such VCM technology yields equally, if not more, accurate conclusions compared to traditional LCM methods.

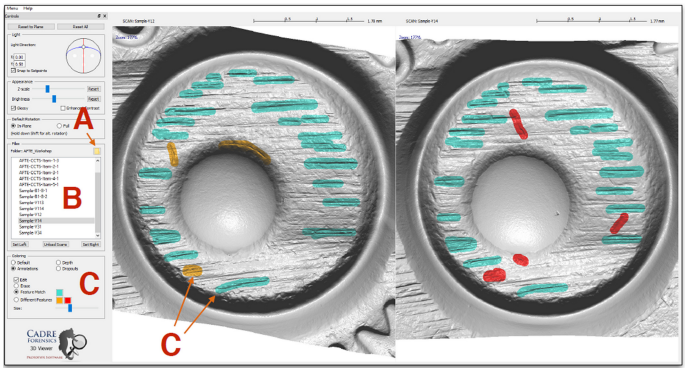

Figure 1.17 A screenshot of the TopMatch-3D™Virtual Comparison Microscopy software. In this example, similar and different markings on the cartridge case scans are manually annotated by the user.

I'm still unclear how you are using "diagnostic" here, so it might be worth clarifying earlier.

We develop a suite of diagnostic tools to explain the behavior of the cartridge case comparison pipeline. These diagnostic tools are created using flexible, open-source tools such as `ggplot2` and `shiny` so that future development and improvement is easy.

1.4 Automating and Improving the Cartridge Case Comparison Pipeline

In this section, we review preliminaries needed to understand various sub-routines of the cartridge case comparison pipeline.

1.4.1 Image Processing Techniques

We first review image processing and computer vision algorithms that are commonly used in cartridge case comparison algorithms. Throughout this section, let A and B denote two images. Define these images to be 2D arrays of a given size where $A[m,n]$ and $B[m,n]$ each map to a spatially-ordered measurement value. For example, the measurement may be the height h value of a cartridge case surface at a particular $[m,n]$ location.

This section is a bit difficult to follow. I'd recommend working on it for clarity. Specifically, I'd simply introduce the mapping you are using specifically (trying to generalize is introducing possible problems that don't matter for how you are using it). Same thing for the similarity metric you are using (just say you will use CCF). You should define notation as you introduce it (not a few sentences later). Finally, I'd introduce the implementation of CCF, and state optimal translation//rotation after giving the order.

1.4.1.1 Image Registration

Image registration involves transforming one image to align with another image (Brown, 1992). For example, in the case of object or facial recognition, one may be interested in finding a template image in another image. For images A and B, image registration can be defined as a mapping between two images:

$$B[m,n] = f(A[m,n])$$

where f is a 2D spatial-coordinate transformation.

f represents the action of an isometry of the Cartesian coordinate space

In our application, ~~f will represent an affine transformation of the Cartesian coordinate space~~ composed of a translation and rotation.[2] This transformation commonly has three parameters: $\Delta x, \Delta y, \theta$ which map a point (x_1, y_1) of the first image to a point (x_2, y_2) of the second image:

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} + \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

what are you using T to refer to? ISO(2)?

A transformation $f(\cdot,\cdot;\boldsymbol{t}^*)$, equivalently a parameter vector $\boldsymbol{t}^* \in \boldsymbol{T}$, is selected such that it optimizes similarity metric $s(\cdot,\cdot)$ between the two images:

Meaning a similarity measure you have previously discussed or are you changing the wording here to indicate something that follows metric axioms?

$$\boldsymbol{t}^* \equiv \arg\max_{\boldsymbol{t}\in\boldsymbol{T}} s(A[m,n], f(B[m,n];\boldsymbol{t})).$$

In our application, the set of possible parameters is $\boldsymbol{T} = \mathbb{Z}\times\mathbb{Z}\times[0,2\pi)$ representing discrete-index horizontal and vertical translations (positive meaning up/right and negative meaning down/left) and a rotation. Commonly, the *cross-correlation function* (CCF) is used as the similarity metric. For a $P\times Q$ "reference" image A and $M\times N$ "template" image B, the cross-correlation function, denoted $A\star B$, is used as a similarity metric. The cross-correlation function measures the similarity between A and B for each translation:

$$(A\star B)[m,n] = \sum_{i=1}^{M}\sum_{j=1}^{N} A[i,j]B[(i+m),(j+n)]$$

[2] We assume that the coordinate spaces do not require scaling.

where $1 \leq m \leq M+P-1$ and $1 \leq n \leq N+Q-1$. By this definition, $A \star B$ is a 2D array of dimension $M+P-1 \times N+Q-1$ in which the $[m,n]$-th element quantifies the similarity between A and B when B is translated m elements horizontally and n elements vertically. For interpretability, the CCF is commonly normalized between -1 and 1.

Using the CCF as a similarity metric, the translations $[m^*, n^*]$ at which images A and B attain the maximum CCF value can be calculated:

$$[m^\dagger, n^\dagger] \equiv \arg\max_{[m,n]} (A \star B)[m,n].$$

To determine the optimal rotation, the maximum CCF value is calculated across a range of rotations of image B. If B_θ denotes image B rotated by an angle $\theta \in [0, 2\pi)$, then the estimated registration $[m^*, n^*, \theta^*]$ is given by:

$$[m^*, n^*, \theta^*] \equiv \arg\max_{[m,n,\theta]} (A \star B_\theta)[m,n].$$

In implementation we consider a discrete grid of rotations $\boldsymbol{\Theta} \subset [0, 2\pi)$. The overall registration procedure is given by:

1. For each $\theta \in \boldsymbol{\Theta}$:

   is the numbering off?

   2.1 Rotate image B by θ to obtain B_θ.

   2.2 Calculate the CCF between A and B_θ.

   2.3 Determine the translation $[m^*, n^*]_\theta$ at which the CCF is maximized. Also, record the CCF value associated with this translation.

2. Across all $\theta \in \Theta$, determine the rotation θ^* at which the largest CCF value is achieved.

3. The estimated registration consists of rotation θ^* and translation $[m^*, n^*]_{\theta^*}$.

Based on the definition given above, the CCF is computationally taxing. In image processing, it is common to use an implementation based on the Fast Fourier Transform

(Brown, 1992). This implementation leverages the Cross-Correlation Theorem, which states that for images A and B the CCF can be expressed in terms of a frequency-domain pointwise product:

$$(A \star B)[m,n] = \mathscr{F}^{-1}\left(\overline{\mathscr{F}(A)} \odot \mathscr{F}(B)\right)[m,n]$$

where $\mathscr{F}$ and $\mathscr{F}^{-1}$ denote the discrete Fourier and inverse discrete Fourier transforms, respectively, and $\overline{\mathscr{F}(A)}$ denotes the complex conjugate (Brigham, 1988). Because the product on the right-hand side is calculated pointwise, this result allows us to trade the moving sum computations from the definition of the CCF for two forward Fourier transformations, a pointwise product, and an inverse Fourier transformation. The Fast Fourier Transform (FFT) algorithm can be used to reduce the computational load considerably.

Figure 1.18 shows an example of two images A and B of dimension 100×100 and 21×21, respectively. The white boxes in both of the images are of dimension 10×10. The box in image A is centered on index [30,50] while the box in image B is centered on index [11,11]. The right image shows the result of calculating the CCF using image A as reference and B as template. The CCF achieves a maximum of 1, indicating a perfect match, at the translation value of $[m^\dagger, n^\dagger] = [22, -2]$. This represents that if image B were overlaid onto image A such that their center indices coincided, then image B would need to be shifted 22 units "up" and 2 units "left" to match perfectly with image A.

1.4.1.2 Gaussian Filters

In image processing, a Gaussian filter (equivalently, blur or smoother) is mathematical operator that imputes the values in an image using a locally-weighted sum of surrounding values. In our application, a Gaussian filter, specifically a *lowpass* Gaussian filter, is used to smooth the surface values of a cartridge case scan. The weights are dictated according to the Gaussian function of a chosen standard deviation σ given by:

$$f(x,y;\sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(x^2+y^2)\right).$$

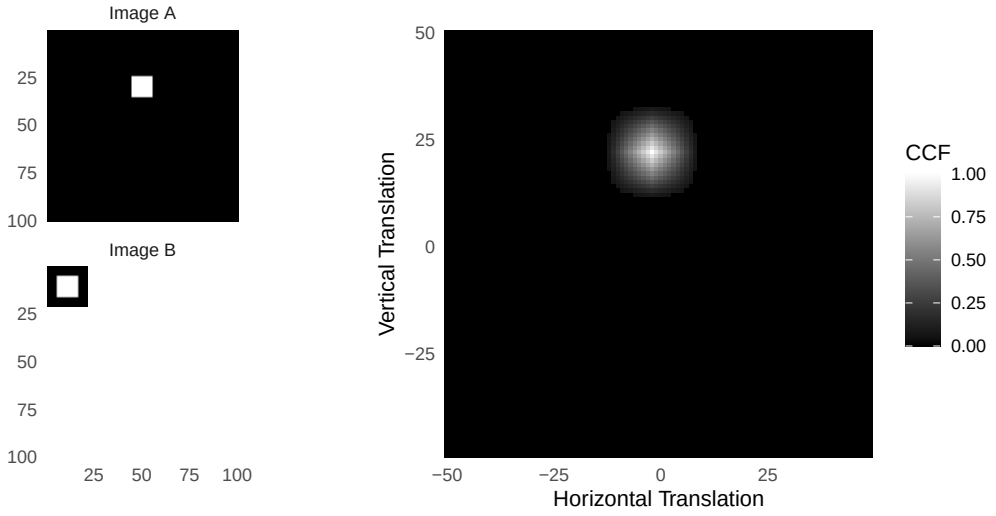

Figure 1.18 (Left) A reference image A and template image B both featuring a white box of dimension 10×10. (Right) The cross-correlation function (CCF) between A and B. The index at which the CCF is maximized represents the translation at which A and B are most similar.

It is common to populate a 2D array with the values of the Gaussian function treating the center index as the origin. Such an array is called a *kernel*. An example of a 3×3 Gaussian kernel with standard deviation $\sigma = 1$ is given below.

$$K = \begin{pmatrix} 0.075 & 0.124 & 0.075 \\ 0.124 & 0.204 & 0.124 \\ 0.075 & 0.124 & 0.075 \end{pmatrix}.$$

For an image A and Gaussian kernel K with standard deviation σ, the lowpass filtered version of A, denoted $A_{lp,\sigma}$ is given by:

$$A_{lp,\sigma}[m,n] = \mathscr{F}^{-1}\left(\mathscr{F}(A) \odot \mathscr{F}(K)\right)[m,n].$$

This operation, known as *convolution*, is extremely similar to the calculation of the CCF given above (ISO 16610-21, 2011).

Figure 1.19 shows an image A of a box undergoing the injection of Gaussian noise (noise standard deviation $\sigma_n = 0.3$) followed by the application of various filters. While

the box is obscured due to noise in the middle image, the lowpass filter (kernel standard deviation $\sigma_k = 2$) recovers some of the definition of the box seen in the original image A.

If a lowpass filter "smooths" the values of an image, then a *highpass* filter performs a "sharpening" operation. More specifically, for image A and kernel standard deviation σ, the highpass filtered version A_{hp} can be defined as:

$$A_{hp,\sigma} = A - A_{lp,\sigma}.$$

The highpass filter therefore removes larger-scale (smooth) structure from an image and retains high-frequency structure such as noise or edges. An example of a highpass-filtered image A is shown in Figure 1.19. The smooth interior of the box is effectively removed from the image while the edges are preserved.

Finally, the bandpass filter performs the highpass sharpening followed by the lowpass smoothing operations. Generally, the highpass kernel's standard deviation will be considerably larger than that of the lowpass kernel. This leads to retaining sharp edges while also reducing noise. An example of a bandpass filtered image A is shown in Figure 1.19. The edges of the box are better-preserved compared to the lowpass filter figure while the interior of the box is better-preserved compared to the highpass filter figure.

Variations on the standard Gaussian filter include the "robust" Gaussian regression filter. This filter fluctuates between a filter step, which applies a Gaussian filter, and outlier step, which identifies and omits outlier observations from the next filter step (Brinkman and Bodschwinna, 2003b). Another alternative, the "edge preserving" filter, adapts the kernel weights when approaching the boundary of an image to mitigate so-called *boundary effects* (Aurich and Weule, 1995).

Transitions would be helpful here, too

1.4.1.3 Morphological Operations

Mathematical morphology refers to a theory and collection of image processing techniques for geometrical structures (Haralick et al., 1987). In our application, these geo-

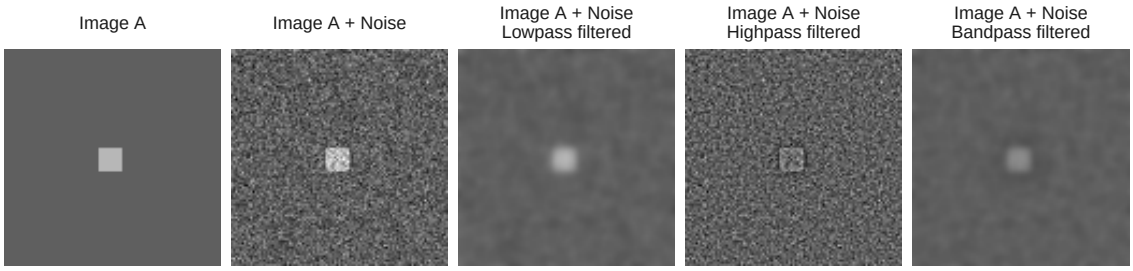

Figure 1.19 An image A of a box undergoing various filtering operations.

metrical structures are cartridge case scans; specifically, binarized versions of these scans representing whether a particular pixel contains part of the cartridge case surface.

Two fundamental operations in mathematical morphology are *dilation* and *erosion* (Haralick et al., 1987). For our purposes, these are both set operations on binary (black and white) images. We classify the set of black and white pixels as the background and foreground of the image, respectively. For an image A, let $W = \{[m,n] : A[m,n] = 1\}$ denote the foreground of A, meaning W^c represents the background. An example of a 7×7 binary image A with $W = \{[3,3],[3,4],[3,5],[4,3],[4,4],[4,5],[5,3],[5,4],[5,5]\}$ is given below.

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

A *structuring element* is a second, typically small, array B of ones that affects the amount of dilation or erosion applied to W within A. For simplicity, the indexing of the structuring element uses the center index as the origin. For example, a 3×3 structuring element is given by $B = \{(-1,-1),(-1,0),(-1,1),(-1,0),(0,0),(0,1),(1,-1),(1,0),(1,1)\}$ or visually:

$$B = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

The dilation of W by B, denoted $W \oplus B$, is defined by

$$W \oplus B = \{[m,n] \in A : [m,n] = [i,j] + [k,l] \text{ for } [i,j] \in W \text{ and } [k,l] \in B\}$$

where the index arithmetic is performed element-wise. Alternatively, if $W_{[k,l]}$ represents the translation of region W within A by k units row-wise and l units column-wise for $[k,l] \in B$

In this example,

$$W \oplus B = \{[3,2],[3,3],[3,4],[3,5],[3,6],[4,2],[4,3],[4,4],[4,5],[4,6],[5,2],[5,3],[5,4],[5,5],[5,6]\}$$

or visually:

$$W \oplus B = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The dilation operation by B has the effect of growing the region W inside of A by one index in each direction.

In contrast, erosion has the effect of shrinking a selected region. More precisely, the erosion of W by B is defined by

$$A \ominus B = \{[m,n] \in A : [m,n] + [k,l] \in A \text{ for every } [k,l] \in B\}.$$

Using the same example as above, $W \ominus B = \{[3,3]\}$ or visually:

$$W \ominus B = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Erosion by B therefore shrinks the region W in A by one index in each direction.

Figure 1.20 shows the example considered here in terms of black and white representations of A undergoing dilation and erosion by B. In practice, there may be two or more disconnected foreground regions in A to which dilation or erosion can be independently applied.

1.4.2 Density-Based Spatial Clustering of Applications with Noise

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is a clustering procedure that assigns observations to clusters if they are in a region of high observation density (Ester et al., 1996). Otherwise they are classified as "noise" points.

Let D represent a $n \times p$ data set (n observations, each of dimension p) and consider observations $x, y, z \in D$. The DBSCAN algorithm relies on the notion of ε-neighborhoods. Given some neighborhood radius $\varepsilon \in \mathbb{R}$ and distance metric d, y is in the ε-neighborhood of x if $d(x,y) \leq \varepsilon$. The *ε-neighborhood* of x is defined as the set $N_\varepsilon(x) = \{y \in D : d(x,y) \leq \varepsilon\}$.

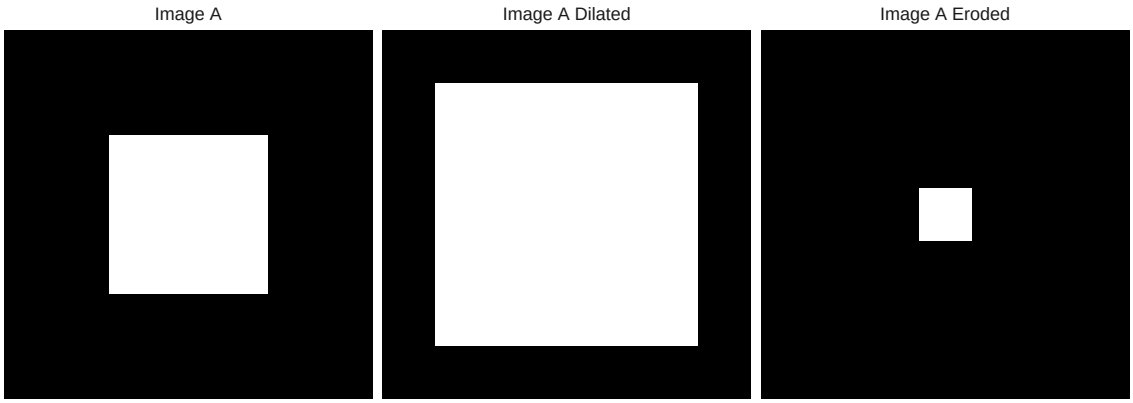

Figure 1.20 A 7×7 image A featuring a 3×3 box undergoing dilation and erosion by a 3×3 structuring element B.

weird notation. is this standard?

Given a minimum number of points $Minpts \in \mathbb{N}$, observation x is called a *core point* with respect to ε and $Minpts$ if $|N_\varepsilon(x)| \geq Minpts$. Both ε and $Minpts$ are selected by the user.

Figure 1.21 shows an example of 10 points on the Cartesian plane. An ε-neighborhood using the Euclidean distance metric and $\varepsilon = 3$ is drawn around an observation x located at $(3,2)$. Points inside the circle are neighbors of x. If, for example, $Minpts = 2$, then x would be considered a core point.

To identify regions of high observation density, two relational notions, *density-reachability* and *density-connectivity*, are used. A point y is *directly density-reachable* to a point x if x is a core point and $y \in N_\varepsilon(x)$. In the example in Figure 1.21, the observation located at $(1,0)$ is directly density-reachable to the observation located at $(3,2)$. More broadly, a point x_m is *density-reachable* to a point x_1 if there exists a chain of observations $x_1, x_2, ..., x_{m-1}, x_m$ such that x_{i+1} is directly density-reachable from x_i, $i = 1, ..., n$. Figure 1.22 shows an example of three density-reachable points located at $(1,0), (3,2)$, and $(4,4)$ using $\varepsilon = 3$ and $Minpts = 2$. All three points are core points and although the points located at

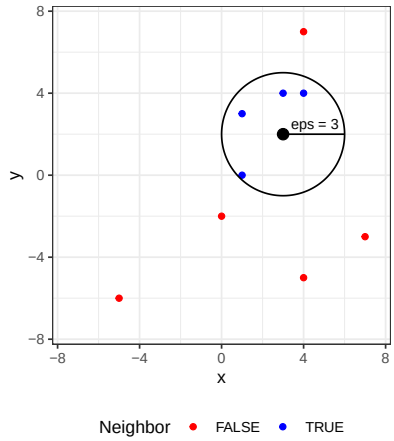

Figure 1.21 An ε-neighborhood around a observation located at $(3,2)$ for $\varepsilon = 3$. Points are labeled based on whether they are neighbors to this observation.

$(4,4)$ and $(1,3)$ are not neighbors, they share a neighbor in the point located at $(3,2)$ and are thus density-reachable.

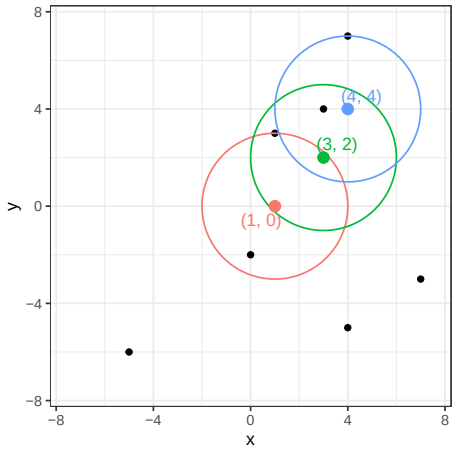

Figure 1.22 An example of three points that are density-reachable with respect to $\varepsilon = 3$ and $Minpts = 2$.

Finally, a point y is *density-connected* to a point x with respect to ε and $Minpts$ if there exists a point z such that both x and y are density-reachable to z (with respect to ε and $Minpts$). While density-reachability requires that all points in-between two points be core points, density-connectivity extends the notion of "neighbors of neighbors" to include points

that are merely within the neighborhood of density-reachable points. Figure 1.23 illustrates how the points located at $(4,7)$ and $(0,-2)$ are density-connected but not density-reachable.

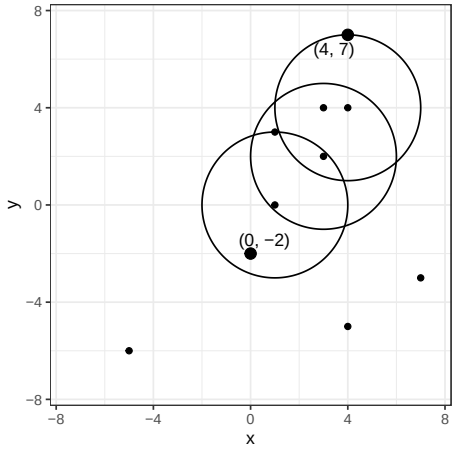

Figure 1.23 An example of two points that are density-connected, but not density-reachable, with respect to $\varepsilon = 3$ and $Minpts = 2$.

A *cluster* $C \subset D$ with respect to ε and $Minpts$ satisfies the following conditions:

1. $\forall x, y$: if $x \in C$ and y is density-reachable from x with respect to ε and $Minpts$, then $y \in C$.

2. $\forall x, y \in C$: x is density-connected to y with respect to ε and $Minpts$.

Points not assigned to a cluster are classified as *noise*.

For a data set D, the DBSCAN algorithm determines clusters based on the above definition. Figure 1.24 shows the labels return by DBSCAN for the example considered above with respect to $\varepsilon = 3$ and $Minpts = 2$.. Seven points are classified in a single cluster and three points are classified as noise.

1.4.3 Features Based on Visual Diagnostics

Much of the "explainable" algorithms literature focuses on black-box machine learning algorithms such as Random Forests or Multi-layer Neural Networks. Less focused is placed

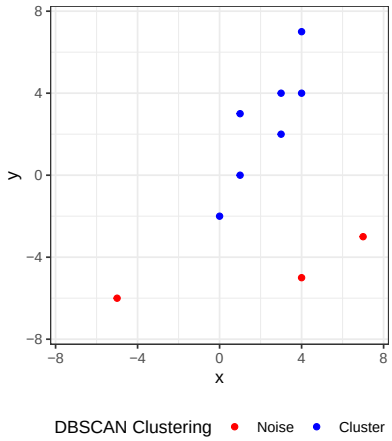

Figure 1.24 Cluster labeling for 10 data points using the DBSCAN algorithm with parameters $\varepsilon = 3$ and $Minpts = 2$. Seven points are assigned to a single cluster and the remaining three are classified as noise.

on constructing explainable features. Feature selection and engineering is a critical, often time-intensive step in the data analysis process that isn't often

The visual diagnostic tools discussed in Chapter 4 are used to develop a set of features. By definition, these features are human-interpretable unlike, for example, features that are calculated in the convolution layer of a convolutional neural network. The interpretability of these features imply that they can be explained to forensic examiners or lay-people. This will make it easier to introduce such methods into forensic labs and court rooms.

1.4.4 Implementation Considerations

This cartridge case comparison pipeline is similar to other data analysis pipelines. Much like other data analysis pipelines, the procedural details can be obscured as the goals of the analysis become more sophisticated. This is helpful neither for the individual performing the analysis nor for any consumer of the results. As such, it is worthwhile to design tools that make the data analysis procedure easier to implement and understand (Wickham, 2014).

Beyond conceptualizing the cartridge case comparison procedure as a pipeline, we also implement the procedure in the R statistical programming as a sequence of algorithms that can programatically be connected together (R Core Team, 2017). In particular, we utilize the pipe operator `%>%` available from the `magrittr` R package (Bache and Wickham, 2022). This operator allows the output of one function to be passed as input to another without assigning a new variable. Data can be incrementally transformed as they move from one function to another.

Implementing a data analysis procedure using the pipe operator allows the user to think intuitively in terms of verbs applied to the data. Table 1.2 illustrates two examples of pipelines that utilize the pipe operator. The left-hand example shows how an R data frame can be manipulated using functions from the `dplyr` package. Functions like `group_by`, `summarize`, and `filter` are simple building blocks that can be strung together to create complicated workflows. The right-hand example similarly illustrates a cartridge case object passing through the comparison pipeline. While the full comparison procedure is complex, the modularization to the `preProcess_`, `comparison_`, and `decision_` steps, which can further be broken-down into simple building blocks, renders the process more understandable to, and flexible for, the user.

Data Frame Manipulation Example	Cartridge Case Comparison Example

`dataFrame %>%`
`  group_by(category) %>%`
`  summarize(x = summary(var)) %>%`
`  filter(x > 0) ...`
	`cartridgeCase1 %>%`
	`  preProcess_func(params1) %>%`
	`  comparison_func(cartridgeCase2,`
	`  params2) %>%`
	`  decision_func(params3) ...`

Table 1.2 Two examples of data analysis workflows that utilize the pipe operator. The left side shows a data frame manipulation while the right side shows a comparison of two cartridge cases.

Figure 1.25, Figure 1.26, Figure 1.27, and Figure 1.28 illustrate how various forensic comparison algorithms use a modularized structure in their preprocessing procedures. In each figure, a sequence of modular procedures are applied to a piece of evidence. Figure 1.25

shows the morphological and image processing preprocessing procedures used to remove the firing pin region from a 2D image of a cartridge case (Tai and Eddy, 2018). Figure 1.26 shows the procedure by which a 2D "signature" of a bullet scan is extracted from a 3D topographical scan (Rice, 2020). Figure 1.27 shows how an image of the written word "csafe" is processed using the handwriter R package to break the word into individual *graphemes* that can be further processed (Berry et al., 2021). Finally, Figure 1.28 shows a 3D topographical cartridge case scan undergoing various procedures to isolate and highlight the breech face impressions. These procedures are discussed in greater detail in Chapter 2.

By breaking the broader preprocessing step into modularized pieces, we can devise other arrangements of these preprocessing procedures that may improve the segmenting or emphasizing of the region of interest. The modularity of the pipeline makes it easier to understand what the algorithm is doing "under the hood" while a modularized implementation enables others to experiment with alternative versions of the pipeline.

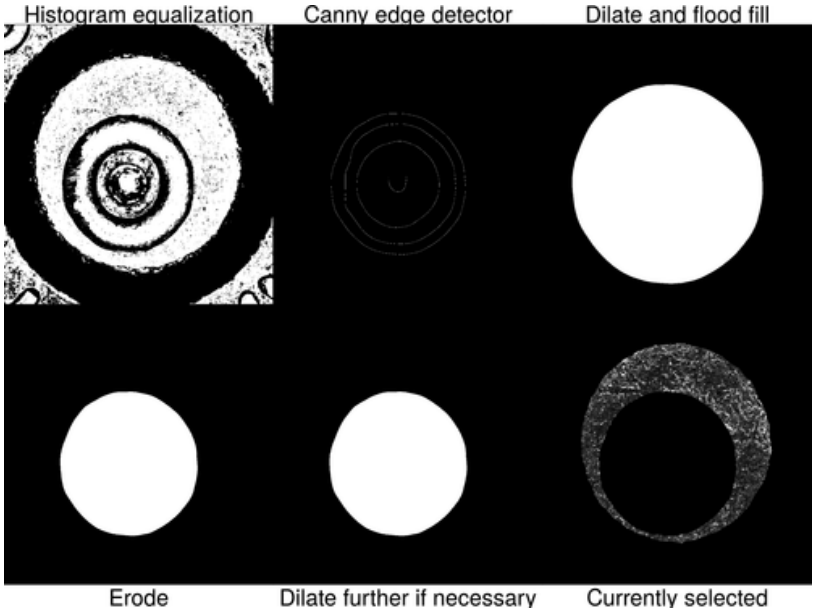

Figure 1.25 A preprocessing procedure applied to a 2D image of a cartridge case to identify the firing pin impression. The procedure results in a 2D image of a cartridge case without the firing pin impression region.

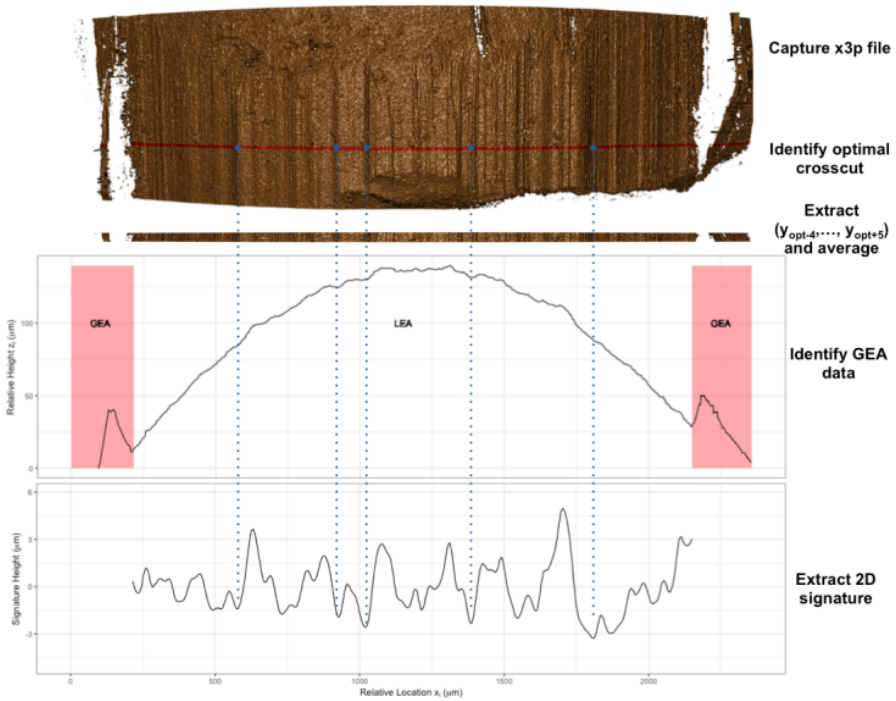

Figure 1.26 A preprocessing procedure for extracting 2D bullet`signatures" from a 3D topographic bullet scan. The procedure results in an ordered sequence of values representing the local variations in the surface of the bullet.

Figure 1.27 A preprocessing procedure applied to a handwriting image of the word "csafe." The procedure results in a skeletonized version of the word that has been separated into graphemes as represented by orange nodes.

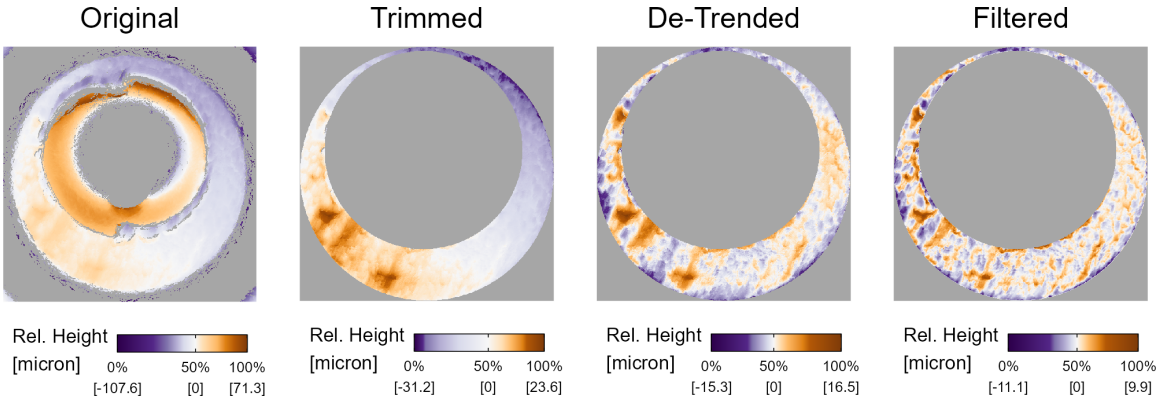

Figure 1.28 A cartridge case undergoing various preprocessing steps. The procedure results in a cartridge case scan in which the breech face impressions have been segmented and highlighted.

Our implementation is structured to adhere to the "tidy" principles of design (Wickham et al., 2019). The `tidyverse` is a collection of R packages that share an underlying design philosophy and structure. The four principles of a tidy API are:

1. Reuse existing data structures.
2. Compose simple functions with the pipe.
3. Embrace functional programming.
4. Design for humans.

Adherence to these principles makes it easier to engage with and understand the overall data analysis pipeline. In our application it also enables experimentation by making it easy to change one step of the pipeline and measure the downstream effects (Zimmerman et al., 2019). Each step of the cartridge case comparison pipeline requires the user to define parameters. These can range from minor, such as the standard deviation used in a Gaussian filter, to substantial, such as choosing the algorithm used to calculate the similarity score. So far, no consensus exists for the "best" parameter settings. A large amount of experimentation is yet required to establish these parameters. A tidy implementation of

the cartridge case comparison pipeline allows more people to engage in the validation and improvement of the procedure.

REFERENCES

(2020). Anaconda software distribution.

30 Magazine Clip (2017). Calibers explained. https://www.30magazineclip.com/the-firearms-crash-course/calibers-explained/.

AFTE Criteria for Identification Committee (1992). Theory of identification, range striae comparison reports and modified glossary definitions. *AFTE Journal*, 24(3):336–340.

American Academy of Forensic Sciences (2021). What is forensic science?

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1):17. https://doi.org/10.2307/2682899.

Aurich, V. and Weule, J. (1995). Non-linear gaussian filters performing edge preserving diffusion. In *Informatik aktuell*, pages 538–545. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-79980-8_63.

Bache, S. M. and Wickham, H. (2022). *magrittr: A Forward-Pipe Operator for R*. R package version 2.0.2.

Baldwin, D. P., Bajic, S. J., Morris, M., and Zamzow, D. (2014). A study of false-positive and false-negative error rates in cartridge case comparisons. Technical report. https://doi.org/10.21236/ada611807.

Beeley, C. and Sukhdeve, S. R. (2018). *Web Application Development with R Using Shiny*. Packt Publishing, Birmingham, England, 3 edition.

Belle, V. and Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4.

Bermudez, C., Matilla, A., and Aguerri, A. (2017). Confocal fusion: Towards the universal optical 3d metrology technology. *Proceedings of the 12th LAMDAMAP, Renishaw Innovation Center, Wotton-Under Edge, UK*, pages 15–16.

Berry, N., Taylor, J., and Baez-Santiago, F. (2021). *handwriter: Handwriting Analysis in R*. R package version 1.0.1.

Brigham, E. O. (1988). *The Fast Fourier Transform and Its Applications*. Prentice-Hall, Inc., USA.

Brinkman, S. and Bodschwinna, H. (2003a). Advanced Gaussian filters. In Blunt, L. and Jiang, X., editors, *Advanced Techniques for Assessment Surface Topography: Development of a Basis for 3D Surface Texture Standards "SURFSTAND"*. Elsevier Inc., United States.

Brinkman, S. and Bodschwinna, H. (2003b). *Advanced Techniques for Assessment Surface Topography*. Elsevier. https://doi.org/10.1016/b978-1-903996-11-9.x5000-2.

Brown, L. G. (1992). A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383. https://doi.org/10.1098/rsta.2009.0120.

Cadre Forensics (2019). Top match-3d high capacity: 3d imaging and analysis system for firearm forensics. https://www.cadreforensics.com/pdf/TopMtopmatchatch-3D-HighCapacity.pdf.

Chang, A. C. and Li, P. (2022). Is economics research replicable? sixty published papers from thirteen journals say "often not". *Critical Finance Review*, 11(1):185–206. https://doi.org/10.1561/104.00000053.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.7.1.

Chapnick, C., Weller, T. J., Duez, P., Meschke, E., Marshall, J., and Lilien, R. (2020). Results of the 3d virtual comparison microscopy error rate (VCMER) study for firearm forensics. *Journal of Forensic Sciences*, 66(2):557–570.

Chatterjee, S. and Hadi, A. S. (2006). *Regression Analysis by Example*. John Wiley & Sons, Inc. https://doi.org/10.1002/0470055464.

Chen, Z., Song, J., Chu, W., Soons, J. A., and Zhao, X. (2017). A convergence algorithm for correlation of breech face images based on the congruent matching cells (CMC) method. *Forensic Science International*, 280:213–223.

Chu, W., Tong, M., and Song, J. (2013). Validation Tests for the Congruent Matching Cells (CMC) Method Using Cartridge Cases Fired with Consecutively Manufactured Pistol Slides. *Journal of the Association of Firearms and Toolmarks Examiners*, 45(4):6.

Cleveland, W. (1994). *The Elements of Graphing Data*. AT&T Bell Laboratories.

Crawford, A. (2020). *Bayesian hierarchical modeling for the forensic evaluation of handwritten documents*. Ph.D thesis, Iowa State University.

Crowder, M. and Hand, D. (1990). *Analysis of Repeated Measures*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Curran, J. M., Champod, T. N. H., and Buckleton, J. S., editors (2000). *Forensic interpretation of glass evidence*. CRC Press, Boca Raton, FL.

DeFrance, C. and Arsdale, M. (2003). Validation study of electrochemical rifling. *Association of Firearms and Tool Marks Examiners Journal*, 35:35–37.

Deng, H. (2018). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, 7(4):277–287. https://doi.org/10.1007/s41060-018-0144-8.

Duez, P., Weller, T., Brubaker, M., Hockensmith, R. E., and Lilien, R. (2017). Development and validation of a virtual examination tool for firearm forensics, ,. *Journal of Forensic Sciences*, 63(4):1069–1084.

Duvendack, M., Palmer-Jones, R. W., and Reed, W. (2015). Replications in economics: A progress report. *Econ Journal Watch*, 12(2). https://EconPapers.repec.org/RePEc:ejw:journl:v:12:y:2015:i:2:p:164-191.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press. doi.org/10.5555/3001460.3001507.

Fadul, T., Hernandez, G., Stoiloff, S., and Sneh, G. (2011). An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides.

Garton, N., Ommen, D., Niemi, J., and Carriquiry, A. (2020). Score-based likelihood ratios to evaluate forensic pattern evidence. https://doi.org/10.48550/ARXIV.2002.09470.

Goldstein, E. and Brockmole, J. (2016). *Sensation and Perception*. CENGAGE Learning Custom Publishing, Mason, OH, 10 edition.

Goode, K. and Hofmann, H. (2021). Visual diagnostics of an explainer model: Tools for the assessment of LIME explanations. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(2):185–200. https://doi.org/10.1002/sam.11500.

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., Goecks, J., Backofen, R., Nekrutenko, A., and Taylor, J. (2018). Practical computational reproducibility in the life sciences. *Cell Systems*, 6(6):631–635. https://doi.org/10.1016/j.cels.2018.03.014.

Gundersen, O. E., Gil, Y., and Aha, D. W. (2018). On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI Magazine*, 39(3):56–68. https://doi.org/10.1609/aimag.v39i3.2816.

Hadler, J. R. and Morris, M. D. (2017). An improved version of a tool mark comparison algorithm. *Journal of Forensic Sciences*, 63(3):849–855. https://doi.org/10.1111/1556-4029.13640.

Hamby, J. E., Brundage, D. J., and Thorpe, J. W. (2009). The identification of bullets fired from 10 consecutively rifled 9mm ruger pistol barrels: A research project involving 507 participants from 20 countries. volume 41, pages 99–110.

Hampton, D. (2016). Firearms identification. a discipline mainly concerned with determining whether a bullet or cartridge was fired by a particular weapon. - ppt download. https://slideplayer.com/slide/9972083/.

Haralick, R. M., Sternberg, S. R., and Zhuang, X. (1987). Image analysis using mathematical morphology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(4):532–550. https://doi.org/10.1109/tpami.1987.4767941.

Hare, E., Hofmann, H., and Carriquiry, A. (2017). Automatic Matching of Bullet Land Impressions. *The Annals of Applied Statistics*, 11(4):2332–2356.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan,

M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, 12(2):115–121.

Indiana County Court of Common Pleas (2009). *Commonwealth of Pennsylvania vs. Kevin J. Foley.*

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., and Ioannidis, J. P. A. (2016). Reproducible research practices and transparency across the biomedical literature. *PLOS Biology*, 14(1):e1002333. https://doi.org/10.1371/journal.pbio.1002333.

ISO 16610-21 (2011). Geometrical product specifications (GPS) - Filtration - Part 61: Linear areal filters: Gaussian filters. Standard, International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/60159.html.

ISO 16610-71(2014) (2014). Geometrical product specifications (GPS) - Filtration - Part 71: Robust areal filters: Gaussian regression filters. Standard, International Organization for Standardization, Geneva, CH.

ISO 25178-72(2017) (2017). Geometrical product specifications (GPS) — Surface texture: Areal — Part 72: XML file format x3p. Standard, International Organization for Standardization, Geneva, CH.

Knowles, L., Hockey, D., and Marshall, J. (2021). The validation of 3d virtual comparison microscopy (VCM) in the comparison of expended cartridge cases. *Journal of Forensic Sciences*, 67(2):516–523.

Krishnan, G. and Hofmann, H. (2018). Adapting the chumbley score to match striae on land engraved areas (leas) of bullets,. *Journal of Forensic Sciences*, 64(3):728–740. https://doi.org/10.1111/1556-4029.13950.

Mattijssen, E. J., Witteman, C. L., Berger, C. E., Brand, N. W., and Stoel, R. D. (2020). Validity and reliability of forensic firearm examiners. *Forensic Science International*, 307:110112. https://doi.org/10.1016/j.forsciint.2019.110112.

Midway, S. R. (2020). Principles of effective data visualization. *Patterns*, 1(9):100141. https://doi.org/10.1016/j.patter.2020.100141.

National Academy of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. National Academies Press.

National Research Council (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC.

Neuman, M., Hundl, C., Grimaldi, A., Eudaley, D., Stein, D., and Stout, P. (2022). Blind testing in firearms: Preliminary results from a blind quality control program. *Journal of Forensic Sciences*, 67(3):964–974. https://doi.org/10.1111/1556-4029.15031.

Ommen, D. M. and Saunders, C. P. (2018). Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, 17(2):179–197. https://doi.org/10.1093/lpr/mgy008.

OSAC Human Factors Committee (2020). Human factors in validation and performance testing of forensic science. Technical report. https://doi.org/10.29325/osac.ts.0004.

Ott, D., Thompson, R., and Song, J. (2017). Applying 3D measurements and computer matching algorithms to two firearm examination proficiency tests. *Forensic Science International*, 271:98–106.

Park, S. and Carriquiry, A. (2020). An algorithm to compare two-dimensional footwear outsole images using maximum cliques and speeded-up robust feature. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(2):188–199.

Park, S. and Tyner, S. (2019). Evaluation and comparison of methods for forensic glass source conclusions. *Forensic Science International*, 305:110003. https://doi.org/10.1016/j.forsciint.2019.110003.

Piccolo, S. R. and Frampton, M. B. (2016). Tools and techniques for computational reproducibility. *GigaScience*, 5(1). https://doi.org/10.1186/s13742-016-0135-4.

President's Council of Advisors on Sci. & Tech. (2016). Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods.

Puiutta, E. and Veith, E. M. S. P. (2020). Explainable reinforcement learning: A survey. In *Lecture Notes in Computer Science*, pages 77–95. Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_5.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rattenbury, R. C. (2015). Semiautomatic pistol. https://www.britannica.com/technology/semiautomatic-pistol#/media/1/44886/66099.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Rice, K. E. (2020). *A Framework for Statistical and Computational Reproducibility in Large-Scale Data Analysis Projects with a Focus on Automated Forensic Bullet Evidence Comparison*. PhD thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2021-05-25.

Riva, F. and Champod, C. (2014). Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases. *Journal of Forensic Sciences*, 59(3):637–647. https://doi.org/10.1111/1556-4029.12382.

Riva, F., Hermsen, R., Mattijssen, E., Pieper, P., and Champod, C. (2016). Objective evaluation of subclass characteristics on breech face marks. *Journal of Forensic Sciences*, 62(2):417–422. https://doi.org/10.1111/1556-4029.13274.

Riva, F., Mattijssen, E. J., Hermsen, R., Pieper, P., Kerkhoff, W., and Champod, C. (2020). Comparison and interpretation of impressed marks left by a firearm on cartridge cases – towards an operational implementation of a likelihood ratio based technique. *Forensic Science International*, 313:110363. https://doi.org/10.1016/j.forsciint.2020.110363.

Roth, J., Carriveau, A., Liu, X., and Jain, A. K. (2015). Learning-based ballistic breech face impression image matching. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8.

Smith, T. P., Smith, G. A., and Snipes, J. B. (2016). A validation study of bullet and cartridge case comparisons using samples representative of actual casework. *Journal of Forensic Sciences*, 61(4):939–946.

Song, J. (2013). Proposed "NIST Ballistics Identification System (NBIS)" Based on 3D Topography Measurements on Correlation Cells. *American Firearm and Tool Mark Examiners Journal*, 45(2):11.

Song, J., Chu, W., Tong, M., and Soons, J. (2014). 3D topography measurements on correlation cells—a new approach to forensic ballistics identifications. *Measurement Science and Technology*, 25(6):064005.

Song, J., Vorburger, T. V., Chu, W., Yen, J., Soons, J. A., Ott, D. B., and Zhang, N. F. (2018). Estimating error rates for firearm evidence identifications in forensic science. *Forensic Science International*, 284:15–32.

Stodden, V., Krafczyk, M. S., and Bhaskar, A. (2018a). Enabling the verification of computational results. In *Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems*. ACM.

Stodden, V., Seiler, J., and Ma, Z. (2018b). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589. https://www.pnas.org/doi/abs/10.1073/pnas.1708290115.

Stroman, A. (2014). Empirically determined frequency of error in cartridge case examinations using a declared double-blind format. *AFTE Journal*, 46:157–175.

Swofford, H. and Champod, C. (2021). Implementation of algorithms in pattern & impression evidence: A responsible and practical roadmap. *Forensic Science International: Synergy*, 3:100142. https://doi.org/10.1016/j.fsisyn.2021.100142.

Tai, X. H. (2019). *Matching Problems in Forensics*. PhD thesis. https://doi.org/10.1184/R1/9963596.V1.

Tai, X. H. and Eddy, W. F. (2018). A Fully Automatic Method for Comparing Cartridge Case Images,. *Journal of Forensic Sciences*, 63(2):440–448.

Telea, A. C. (2014). *Data visualization: principles and practice*. CRC Press.

Thompson, R. (2017). *Firearm Identification in the Forensic Science Laboratory*. National District Attorneys Association.

Tong, M., Song, J., and Chu, W. (2015). An Improved Algorithm of Congruent Matching Cells (CMC) Method for Firearm Evidence Identifications. *Journal of Research of the National Institute of Standards and Technology*, 120:102.

Tong, M., Song, J., Chu, W., and Thompson, R. M. (2014). Fired Cartridge Case Identification Using Optical Images and the Congruent Matching Cells (CMC) Method. *Journal of Research of the National Institute of Standards and Technology*, 119:575.

Tyner, S., Soyoung Park, Krishnan, G., Pan, K., Hare, E., Luby, A., Tai, X. H., Hofmann, H., and Basulto-Elias, G. (2019). sctyner/openforscir: Create doi for open forensic science in r. https://doi.org/10.5281/ZENODO.3418141.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., and Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences*, 108(19):7733–7738. https://doi.org/10.1073/pnas.1018707108.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., and Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE*, 7(3):e32800. https://doi.org/10.1371/journal.pone.0032800.

Ulery, B. T., Hicklin, R. A., Roberts, M. A., and Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. *PLoS ONE*, 9(11):e110179. https://doi.org/10.1371/journal.pone.0110179.

Wang, J., Keusters, W. R., Wen, L., and Leeflang, M. M. (2021). Ipdmada: An r shiny tool for analyzing and visualizing individual patient data meta-analyses of diagnostic test accuracy. *Research synthesis methods*, 12(1):45–54.

Weller, T., Brubaker, M., Duez, P., and Lilien, R. (2015). Introduction and initial evaluation of a novel three-dimensional imaging and analysis system for firearm forensics. *AFTE Journal*, 47:198.

Weller, T. J., Zheng, A., Thompson, R., and Tulleners, F. (2012). Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides. 57(4). https://doi.org/10.1111/j.1556-4029.2012.02072.x.

Werner, D., Berthod, R., Rhumorbarbe, D., and Gallusser, A. (2021). Manufacturing of firearms parts: Relevant sources of information and contribution in a forensic context. *WIREs Forensic Science*, 3(3):e1401. https://doi.org/10.1002/wfs2.1401.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686. https://doi.org/10.21105/joss.01686.

Wilkinson, L. (2005). *The Grammar of Graphics*. Springer-Verlag, Berlin, Heidelberg.

Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In Stodden, V., Leisch, F., and Peng, R. D., editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC. ISBN 978-1466561595.

Zhang, H., Song, J., Tong, M., and Chu, W. (2016). Correlation of firing pin impressions based on congruent matching cross-sections (CMX) method. *Forensic Science International*, 263:186–193. https://doi.org/10.1016/j.forsciint.2016.04.015.

Zhang, H., Zhu, J., Hong, R., Wang, H., Sun, F., and Malik, A. (2020). Convergence-improved congruent matching cells (CMC) method for firing pin impression comparison. *Journal of Forensic Sciences*, 66(2):571–582. https://doi.org/10.1111/1556-4029.14634.

Zheng, X., Soons, J., Thompson, R., Singh, S., and Constantin, C. (2020). NIST ballistics toolmark research database. *Journal of Research of the National Institute of Standards and Technology*, 125. https://doi.org/10.6028/jres.125.004.

Zheng, X., Soons, J., Vorburger, T. V., Song, J., Renegar, T., and Thompson, R. (2014). Applications of surface metrology in firearm identification. *Surface Topography: Metrology and Properties*, 2(1):014012. https://doi.org/10.1088/2051-672x/2/1/014012.

Zimmerman, N., Wilson, G., Raniere Silva, Ritchie, S., Michonneau, F., Oliver, J., Dashnow, H., Boughton, A., Teucher, A., Mawdsley, D., MacDonald, A., Rice, T., Emonet, R., Daigle, R., Mills, B., Bolker, B., Penrose, S., Sloggett, C., Blischak, J., Moore, T. E., Mawdsley, D., Arnold, J., Bridges, D., Becker, E. A., Riva, G. V. D., Ing-Simmons, L.,

Research Bazaar, Bekolay, T., Piaskowski, J., Sze, M., Hadley, M. J., Hejazi, N., Mayer, F., Leinweber, K., Deer, L., Lesniak, N., Burge, O. R., Martinez, P. A., Conrado, A. C., Jankevics, A., Ashander, J., Duckles, J., Zappia, L., Burle, M.-H., Mitchell, N., Bouchet, P., Harris, R. M., Renaut, S., Sparks, A. H., , Daniel, Attali, D., Tyre, D., Morrison, E., McDonald, G., Bar, I., Mickley, J., McDevitt-Irwin, J., Koziar, K., Samuk, K., Marwaha, K., Chatzidimitriou, K., Chang, L., Kardish, M., Potter, N., Boersch-Supan, P., Funkhouser, S. A., Magle, T., Waiteb5, Ahsan Ali Khoja, Lee, A., Berlanga-Taylor, A., Ashwin Srinath, Bippuspm, Beheim, B., Butterflyskip, Harris, D. J., Oliveira, D. R., Balamuta, J., Quan, J., Woo, K., Hertweck, K., Ottoboni, K., Weitemier, K., Nederbragt, L., Lonsdale, A., Johnston, L. W., Frassl, M., Dunning, M., Donovan, M., Clark, M., Jackson, M., Cadzow, M., Narayanan Raghupathy, Sélem, N., Bachant, P., Banaszkiewicz, P., Barnes, R., Bagchi, R., Brosda, S., Munro, S., Lavrentovich, S., Rossum, T. V., Kelly, T. C., Vicken Hillis, West, K. A., and Takemon, Y. (2019). swcarpentry/r-novice-gapminder: Software carpentry: R for reproducible scientific analysis, june 2019. https://doi.org/10.5281/ZENODO.3265164.