STANDARDS OF USE OF DNA DATABASES: PRIVACY AND PUBLIC INFORMATION

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Lily Roark

March 28, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

Deoxyribonucleic acid (DNA) profiling for criminal cases is one of the greatest technological innovations in criminal justice of the past few decades, but uncertainty and controversy still surrounds the standards of DNA collection and analysis. The US employs local, state, and national levels of DNA index systems. All of these systems are supported by the Federal Bureau of Investigation (FBI)-created Combined DNA Index System (CODIS), which includes the DNA database itself and the software suite used to manage it (Butler, 2005, p. 441). Additionally, DNAxs software developed abroad in the recent past builds on CODIS using probabilistic genotyping algorithms (Slagter et al., 2021, p. 1). Public privacy of sensitive biological data is balanced with the need for the system to be accessible, efficient, and accurate.

The STS research topic of this thesis is the standards of use and management of DNA databases, where CODIS is the primary case to study. CODIS has relatively well-defined regulations set by the FBI (July 2020, p. 1). These regulations are highly specific rules for the forensic laboratories analyzing the DNA samples and the investigators and database operators who use CODIS to link cases or suspects; they range from the qualifications required of laboratory personnel and management, through the testing of CODIS software, to the documentation procedures which leave a paper trail in case of malpractice. In addition to the FBI's quality assurance standards for DNA databasing laboratories, the Scientific Working Group on DNA Analysis Methods (SWGDAM), a meeting community of forensic scientists from the United States and Canada, have produced standards for the efficient DNA processing of Sexual Assault Evidence Kits (SAKs) in a laboratory (SWGDAM, 2016, p. 2). Figure 1 on page 2 shows their Direct-to-DNA approach, which recommends analyzing DNA before other kinds of biological forensic evidence such as blood (serological testing), as DNA is not only more unique but also more efficient.

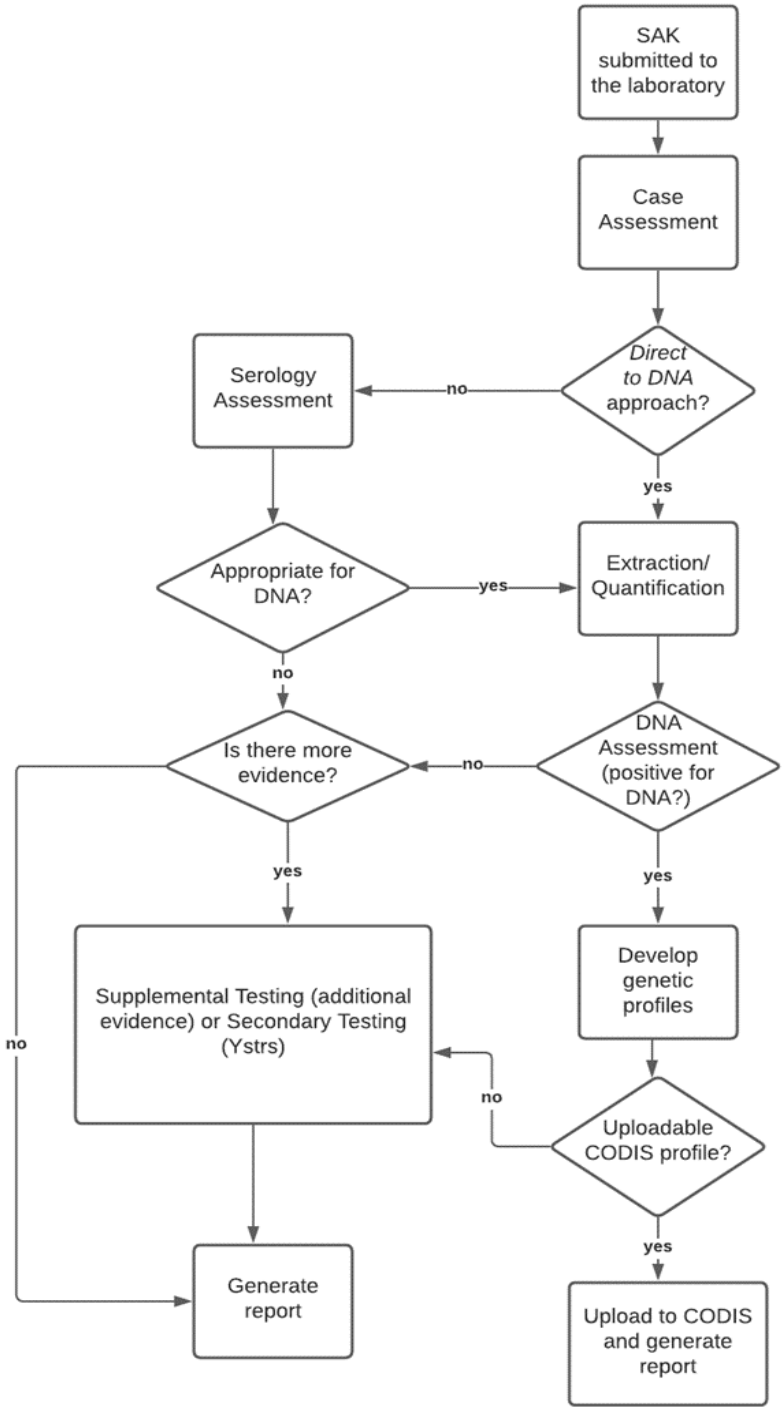

Figure 1: High throughput process flow for sexual assault evidence kits: Adapted from "SWGDAM recommendations for the efficient DNA processing of sexual assault evidence kits", the flow chart tracks the recommended handling of SAKs by a lab, via the *Direct to DNA* approach (SWGDAM, 2016).

Regarding the enforcement of these two standards, the FBI quality assurance standards are enforced via an annual audit, whereas the SWGDAM protocol is merely recommended practice on behalf of the National Institute of Justice (NIJ) for the Sexual Assault Forensic Evidence Reporting (SAFER) Act (SWGDAM, 2016, p. 1). These standards, though rigorous within the confines of the DNA analysis laboratory, need expansion in light of the rapidly developing forensic DNA fingerprinting technology. The STS thesis uses Actor Network Theory as defined by Law and Callon (1988, p. 285) to consider all possible actors that could determine the expansion of standards.

The technical portion of this thesis, under the advice of Professor Daniel Graham, examines how the scalability, efficiency and effectiveness of DNA database software affects the security of the system. The state-of-the-art technical report examining security measures for DNA is especially relevant given developments within the past three years: RapidDNA automated DNA analysis and probabilistic genotyping software. RapidDNA is a fully-automated instrument to be used by law enforcement booking agencies. RapidDNA utilizes buccal (inner cheek) swabs from arrestees and completes the process of DNA entry into CODIS within a time span of only two hours (FBI, September 2020, p. 4). In addition to RapidDNA, DNAxs, developed by the Netherlands Forensic Institute, is a modular and portable software able to apply probabilistic genotyping to complicated case samples (Slagter et al., 2021, p. 2). Although RapidDNA, DNAxs and other probabilistic genotyping software improve the efficiency and efficacy of DNA databasing, their integration into the laboratory analysis software system introduces opportunities for software bugs and security vulnerabilities (Slagter et al., 2021, p. 8). Thus, the technical topic is tightly coupled with the STS topic, both at the surface-level of DNA databases and at the ethical level of protecting the privacy of the public.

ACCURACY, EFFICIENCY OF RESULTS EXTEND BEYOND THE LAB

Within the laboratories analyzing DNA and querying CODIS, the FBI standards are clear and concise, demanding trained managerial and technical staff (FBI, July 2020, p. 14), separation of tasks, a documented chain of custody for evidence, and peer review of scientific validation studies (SWGDAM, 2015, p. 5). They also require software review of functionality, reliability, precision, accuracy, sensitivity and specificity (FBI, July 2020, p. 25). In comparison, the standards offered by SWGDAM are more applicable to the victims, nurses, law enforcement and judiciary involved in the collection of sexual assault evidence than the laboratory analysts (SWGDAM, 2016, p. 1). There is an unfortunately high throughput of SAKs to be analyzed, and this requires measures to increase the efficiency of processing (SWGDAM, 2016, p. 3); SAK processing usually takes two to six months by state, but with *Direct to DNA* and other SWGDAM processing recommendations, this could be cut down to only two to four weeks (SWGDAM, 2016, p. 22).

These existing standards are sufficient in the limited definition of technology; the FBI outlines the setup of labs such that DNA analysis is accurate, and SWGDAM guides the growing efficiency of the practice. However, the practice of a technology is in constant relation with the societal applications of that technology, which reach past the confines of the laboratory. Overlooking this relationship results in real-world consequences; for example, in the Lifecodes band-shifting debacle of 1989, Lifecodes, a privately contracted DNA testing laboratory, withheld their erroneous methodology from the prosecution, resulting in a reversed testimony (Annas, 1990, p. 37). This case would have benefitted from an integrated in-court explanation of the scientific process of DNA analysis, as well as a review of Lifecodes' adherence to standards of practice. It is a mistake to believe that the developers of database software and the analysts

using it are immune to outside pressures to unethically fabricate the authenticity of their work. In addition to standards of use, there must be independent oversight and external code review, as developers are biased towards their own correctness when testing their own code (Buckleton et al., 2021, p. 6). The uncontrollable growth of computer databases containing a wealth of private information means that even if the individual has little control over how their information is used, they should still be informed of what is collected, how it is generally analyzed, and to what purpose it serves (Weiss, 2004, p. 62). The question is: how should additional standards of forensic DNA analysis be set in order to maintain public transparency and avoid malpractice?

DETERMINATION OF SYSTEMIC VIRTUES

In answering the question of how additional standards of forensic DNA analysis could maintain public transparency and avoid malpractice, the primary goal is remediating the belief that developers of database software and the analysts using it are immune to outside pressures to unethically misuse or fabricate the authenticity of their work. A secondary objective within this goal is to identify desirable qualities of the system, such as transparency and consistency, which act in the best interest of the populace. Counter-examples of these qualities, such as the Internal Revenue Service's (IRS) failure to respect the privacy of biometric facial data (Zakrzewski, 2022), set precedent. In this instance, the IRS directed 7 million Americans to sign up with biometric technology company ID.me's face-scan service to confirm identification for viewing tax information on the IRS' website, a move which was unnecessary and dangerous. The risks presented by government-contracted facial recognition include but are not limited to: software bias in race and gender, potential for identity theft, unlawful technological requirements to file one's taxes, and commercialization of a government service (Harwell, 2022). As Rep. Carolyn B.

Maloney (D-N.Y.) commented on the case, "[Biometric] technology remains virtually unregulated, and increasing transparency and accountability is crucial" (Zakrzewski, 2022).

ACTOR-NETWORK THEORY APPLIED TO DNA DATABASES

The academic articles, official government codes, and analogous case studies discussed above provide for context informing how to extend standards of technology-practice in the domain of DNA forensics. The existing standards are constrained to the confines of the analyzing laboratories, which hold the majority accountability for malpractice, but external agents significantly impact the laboratory's ability to securely, accurately, and efficiently analyze private data for public good. Actor-network theory (ANT) as defined by Law and Callon (1988, p. 285) considers all such possible actors that could be affected by the expansion of standards, as shown in the preliminary ANT diagram, Figure 2 on page 8. An appropriate case analogy to ANT-framed forensic DNA analysis is "Blowing against the wind"; a town within the visual scope of a proposed wind-energy farm in France was not included in the preliminary planning stage, resulting in unexpected consequences of actor "overflow", effectively derailing the project (Jolivet & Heiskanen, 2010). Here, overflow is the unintended, often negative, effect of the network on actors that initially seem distant or uninvolved. In order to avoid overflow and the historical mistakes from the other case studies, continuous effort needs to be made to identify and engage with entities that impact the DNA fingerprinting system.

As shown by the myriad of different actors in Figure 2 on page 8, DNA databases are at the intersection of biology and computer science, the law, and public interest, so an ANT perspective untangles those relationships into a digestible model. Government agencies such as the IRS, developers, and legal agents who control the results and must understand the technology

have been overlooked in the real-world lifespan of forensic DNA and the software designed to analyze it. Their perspectives could enlighten potential places for standards refinement, if significant outreach was made to ensure their motivations, desires and frustrations were heard.

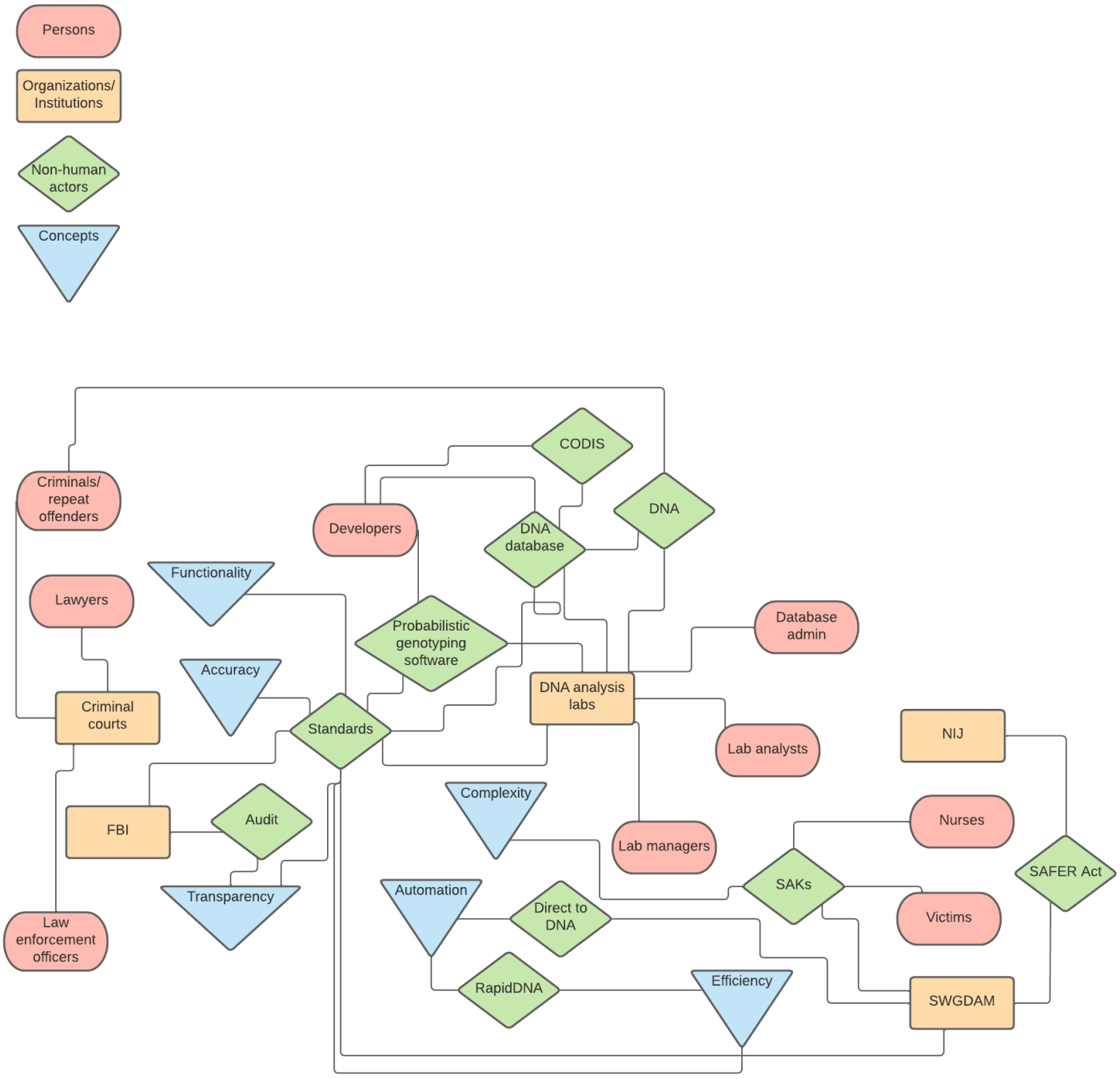

Figure 2: Preliminary DNA database ANT diagram: an overview of key actors influencing the development and standards of practice for forensic DNA databases, featuring probabilistic genotyping software and RapidDNA technologies (Roark, 2021).

In support of using ANT to map out the actors and relationships within the DNA forensic technology network, the issue of biometric data privacy in the ever-accelerating world of information technology likens to scholar Erik van der Vleuten's "grand challenges" facing contemporary society (Van Der Vleuten, 2020, p. 260). Van der Vleuten enumerates a couple "grand challenges", environmental sustainability and widening inequality, but more generally defines them as issues that are "intellectually tricky" and inciting "politically charged debate" (Van Der Vleuten, 2020, p. 261). As these issues scale and become entangled with global systems, they reach a point of crisis, whether real or perceived (Van Der Vleuten, 2020, p. 263). DNA databasing standards are constantly evolving with the technology itself, making them intellectually tricky, and considering the judiciary impact of their results, they are also inherently politically controversial. In terms of crisis level, CODIS is already extremely secure, and mistakes in the system are far from commonplace. CODIS has never been hacked into, and even if it were successfully breached, the data contained within it provides no information directly tying a record to its source name or phenotype ("Fact Sheet"). However, if DNA databasing security lacks the crisis intensity of van der Vleuten's other "grand challenges", his approach to resolving the issue is still helpful considering the issue's complexity and scope.

Van der Vleuten's approach is to initiate cross-disciplinary discourse about the challenge in order to imagine a more sustainable future, and renegotiate the dominant perspectives of the past (Van Der Vleuten, 2020, pp. 265). Through a more diverse set of perspectives, out-of-the box thinking and solutions which transcend the conceptual framework of any one specialized scholar will arise. The key to applying socio-technological historical insights to the present is to overcome these barriers of closed language and methodology between various disciplines of the arts and sciences (Van Der Vleuten, 2020, p. 267). Van der Vleuten's initiative for cross-

disciplinary encounters parallels an ANT-informed revision of standards of use for forensic DNA databases; the major question of data privacy requires communication towards a shared understanding of the issue between legal parties, laboratory analysts, software engineers, law enforcement and the public at large. Each of these actors have a different historical notion of DNA databases and probabilistic genotyping based on their exposure to different aspects of the system, their learning paradigm within their specific profession, and barriers of technical language for communicating to the other disciplines.

The relevant actors for forensic DNA databasing can be directly identified or extrapolated from the standards documents published by the FBI and SWGDAM. Some of the human actors include but are not limited to: law enforcement officers and investigators, lawyers and other members of the judiciary, offenders, developers, employees of the lab ranging from database administrators to analysts and managers, nurses collecting samples, and victims of crime. From mapping out these actors in Figure 2 on page 8, patterns begin to form of their relationships; from left to right on the diagram, the clustered fields of the system are the judiciary, the development of the technology, the laboratory, and the sample data collection. The patterns also apply to the institutional, conceptual, and non-human actors in the map. Therefore, drawing indirectly connected actors from different spots on the diagram results in varying perspectives that can be used in Van der Vleuten's theory of dialogue between specialized fields. For example, facilitated dialogue between the nurses who collect sexual assault kits and the investigators who must testify to the results of those kits could enlighten both parties to come to a new conclusion they would not have reached alone. Additionally, it would foster empathy within the professional network and open a line of communication should either party need clarification on other issues in the future.

Facilitated dialogue between different mindsets within the ANT network of forensic DNA databasing not only helps with out-of-the-box problem solving, but also with reinforcing and expanding the ANT model itself. As a group of actors communicate, they may realize whose expertise they are missing, and identify new actors to introduce to the discourse. Continuing with the investigator/nurse example, the two of them may realize that the transportation associates moving evidence from collection sites to the laboratory or long-term storage have been overlooked in the security of the DNA itself. The actors in transportation would reveal other relevant actors, so the ANT model becomes a positive feedback loop of expanding discourse. The iterative process of talking through overlooked perspectives reveals cracks in the system, opportunities for standards reform.

REFLECTIONS MOVING FORWARD

DNA databasing technology, predominantly the CODIS software system, is an essential tool to modern forensics. As developments in collecting and analyzing DNA samples such as RapidDNA and probabilistic genotyping systems accelerate, standards for actors involved in the system must be continually evaluated and updated to protect public privacy of biometric data. Analogous case studies can illuminate mistakes of the past, as in the Lifecodes evidence debacle, the IRS's contract with ID.me, and the actor overflow in "Blowing against the wind". However, van der Vleuten's theory of "grand challenges" illuminates how Actor-Network Theory can be incorporated into a practical plan for solving such mistakes. In the future, avenues of standards reform should go beyond the laboratory, and the FBI and SWGDAM should facilitate discussion between relevant groups identified by continuous ANT analysis.

REFERENCES

Annas, G.J. (1990). At law: DNA fingerprinting in the Twilight Zone. *The Hastings Center Report*, 20(2), 35-37. https://www.jstor.org/stable/3562618

Buckleton, J.S., Curran, J., Taylor, D., Bright, J.A. (2021, September). What can forensic probabilistic genotyping software developers learn from significant non-forensic software failures? *WIREs Forensic Sci*, 3(2), 1-8. https://doi.org/10.1002/wfs2.1398

Butler, J.M. (2005). Combined DNA Index System (CODIS) and the use of DNA databases. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2nd ed., pp. 435-452). Elsevier Science & Technology. https://www.sciencedirect.com/science/article/pii/B9780123749994000126

Federal Bureau of Investigation (2020, June 30). CODIS and NDIS fact sheet. https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet

Federal Bureau of Investigation (2020, July 1). Quality assurance standards for DNA databasing laboratories. https://www.fbi.gov/file-repository/quality-assurance-standards-for-dna-databasing-laboratories.pdf/view

Federal Bureau of Investigation (2020, September 1). Standards for the operation of Rapid DNA booking systems by law enforcement booking agencies. https://www.fbi.gov/file-repository/standards-for-operation-of-rapid-dna-booking-systems-by-law-enforcement-booking-agencies-eff-090120.pdf/view

Harwell, D. (2022, February 7). IRS abandons facial recognition plan after firestorm of criticism. *The Washington Post*. https://www.washingtonpost.com/technology/2022/02/07/irs-idme-face-scans/

Jolivet, E. & Heiskanen, E. (2010, November 15). Blowing against the wind-An exploratory application of actor network theory to the analysis of local controversies and participation processes in wind energy. *Energy Policy*, 38(11), 6746-6754. https://doi.org/10.1016/j.enpol.2010.06.044

Law, J. & Callon, C. (1988, June). Engineering and sociology in a military aircraft project: A network analysis of technological change. *Social Problems*, 35(3), 284-297. https://www.jstor.org/stable/800623

Roark, L. (2021). *Preliminary DNA database ANT diagram*. [Figure 2]. Prospectus (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Scientific Working Group on DNA Analysis Methods (2015, June 15). SWGDAM guidelines for validation of probabilistic genotyping systems. http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf

Scientific Working Group on DNA Analysis Methods (2016, December 5). SWGDAM recommendations for the efficient DNA processing of sexual assault evidence kits. http://media.wix.com/ugd/4344b0_4daf2bb5512b4e2582f895c4a133a0ed.pdf

Slagter, M., Kruise, D., van Ommen, C., Hoogenboom, J., Steensma, K., de Jong, J., Hovers, P., Parag, R., van der Linden, J., Kneppers, A.L.J., & Benschop, C.C.G. (2021, July). The DNAxs software suite: A three-year retrospective study on the development, architecture, testing and implementation in forensic casework. *Forensic Science International: Reports*, 3(100212), 1-12. https://doi.org/10.1016/j.fsir.2021.100212

Van Der Vleuten, E. (2020). History and Technology in an Age of "Grand Challenges": Raising Questions. *Technology and Culture 61*(1), 260-271. doi:10.1353/tech.2020.0000.

Weiss, M.J. (2004, January 1). Beware! Uncle Sam has your DNA: legal fallout from its use and misuse in the U.S. *Ethics and Information Technology*, 6(1), 55 - 54. https://doi.org/10.1023/b:etin.0000036159.90081.cc

Zakrzewski, C. (2022, February 11). The IRS directed 7 million Americans to sign up with ID.me face-scan service, according to congressional letter. *The Washington Post*. https://www.washingtonpost.com/technology/2022/02/11/letter-maroney-7-million-idme-facial-recognition/