INVESTIGATION · METHOD AND THE FUTURE

Analytical Software for Investigators

Modern investigations generate far more data than people can read, so investigators rely on software to organise, search, link and analyse it. This article explains the main categories of analytical software, what each does well, the evidence on technology-assisted review, and the risks of relying on tools whose workings users do not understand.

ModQu research article · 1,928 words · 22 references (APA 7th edition)

A single investigation can now involve millions of emails, terabytes of data from phones and computers, years of financial transactions and thousands of pages of documents. No team of people can read all of this. Investigators therefore depend on software to organise the material, find what matters and reveal patterns. Quick and Choo (2014) documented how the growing volume of digital evidence had created backlogs in forensic laboratories and argued that new methods and tools were essential to keep pace. Garfinkel (2010) predicted that digital forensics would face a crisis of scale and called for tools that could analyse data more intelligently rather than simply extracting more of it.

Analytical software can make investigations faster, more thorough and more consistent. It can also introduce new errors, hide assumptions, and give a false sense of certainty. This article describes the main categories of software used by investigators, the evidence on how well some of them work, and the principles for using them responsibly.

Categories of software

Investigative software falls into several broad categories, each suited to particular tasks.

Main categories of investigative software

Most investigations use several types together

	Main purpose	Key risk
Case management	Track tasks, evidence, decisions and deadlines	Records only as good as what is entered
Digital forensics	Extract and examine data from devices	Tool errors; unverified parsing of new apps
Document review and e-discovery	Search, filter and classify large document sets	Missed documents from poor search terms
Link and network analysis	Map relationships between people, accounts and companies	False links from poor entity resolution
Data analytics	Detect anomalies and patterns in transactions	Spurious patterns; biased training data
Geographic analysis	Map locations, movements and hot spots	Location errors; misleading map scales

Based on Casey (2011), Heuer & Pherson (2015) and Chainey & Ratcliffe (2005) ModQu.com

Figure 1. Six common categories of investigative software, with the main purpose and a key risk of each. The risks are rarely in the software alone, but in how its results are interpreted. Diagram by ModQu.

Case management systems record tasks, evidence, decisions, interviews and deadlines. They support the audit trail that allows an investigation to be reviewed and repeated, and they help supervisors allocate work and spot gaps. Major police investigations in many countries use dedicated systems to manage the flow of information, so that leads are recorded, prioritised and followed up consistently.

Digital forensic tools create verified copies of storage devices, recover deleted files, extract data from phones and computers, and parse records such as browser histories, messages and location data (Casey, 2011; Kent et al., 2006). They save enormous amounts of time, but each tool interprets raw data according to its own rules, and new versions of apps may store data in ways the tool does not yet understand. Analysts therefore validate tools against known data and, for important findings, confirm results with a second tool or by manual examination.

Document review and e-discovery platforms allow investigators and lawyers to load large collections of documents and emails, remove duplicates, search, tag and review them. Many now use machine learning to prioritise or classify documents, an approach known as technology-assisted review, discussed below. Text search remains central, and the field of information retrieval provides the methods on which such searching rests (Manning et al., 2008).

Link and network analysis tools import data such as telephone records, transactions and company registry information, and display the relationships between entities as charts, with measures that identify central individuals and clusters (Xu & Chen, 2005). A key step is entity resolution, deciding which records refer to the same person, company or account.

Record linkage theory, first formalised by Fellegi and Sunter (1969), shows that such decisions involve trade-offs between false matches and missed matches. Christen (2012) describes modern techniques for matching records that contain spelling variations, missing fields and errors.

Data analytics tools search large datasets for anomalies and patterns. In fraud investigations, they may flag duplicate payments, unusual vendors, transactions just below approval limits, or departures from expected digit patterns. Bolton and Hand (2002) reviewed statistical methods for fraud detection and noted that most flag unusual cases for human review rather than proving fraud. Nigrini (2012) describes the use of Benford's Law, which predicts the frequency of leading digits in many naturally occurring datasets, as a screening tool in auditing and forensic accounting.

Geographic analysis tools map locations of incidents, addresses, phone cell sites and movements, and identify concentrations, or hot spots. Chainey and Ratcliffe (2005) and Eck et al. (2005) describe how geographic information systems are used in crime analysis, and how choices about mapping methods affect what the maps appear to show.

Technology-assisted review

Technology-assisted review (TAR) is one of the best-studied uses of machine learning in investigation. In TAR, reviewers code a sample of documents as relevant or not, and software learns from these decisions to rank or classify the rest. Grossman and Cormack (2011) compared TAR with exhaustive manual review in a large evaluation and found that TAR could identify relevant documents at least as effectively as manual review, while requiring people to review far fewer documents. Manual review was itself much less consistent than often assumed, because human reviewers frequently disagreed.

Courts have accepted TAR in litigation. In *Da Silva Moore v. Publicis Groupe* (2012), a United States federal court approved its use for the first time, and in *Pyrrho Investments Ltd v MWB Property Ltd* (2016), the High Court of England and Wales did the same, noting that it could be more accurate and proportionate than manual review of millions of documents. Guidance on electronic disclosure emphasises that parties should be transparent about the methods used and should validate their results, for example by sampling documents the system classified as not relevant (The Sedona Conference, 2018).

Analytics in policing

Police forces increasingly use analytical software to direct resources. Intelligence-led policing, as described by Ratcliffe (2016), uses analysis of crime data and intelligence to decide where to patrol and whom to target. Brayne (2017), in a study of the Los Angeles Police Department, found that data platforms allowed police to link previously separate databases and to track people who had not been suspected of any crime, extending surveillance in ways that were not always visible to the public. Ferguson (2017) argues that such tools can improve efficiency but can also reproduce and amplify existing biases, because the data they rely on reflect past policing decisions. These concerns apply wherever analytical software is used to decide whom to investigate.

Risks of relying on software

The benefits of analytical software are substantial, but several risks recur.

Opacity. Many tools, especially commercial ones, do not disclose how they process data or reach conclusions. Users may not know which records a forensic tool failed to parse, how a matching algorithm decided that two names refer to the same person, or why an analytics tool flagged a transaction. Where results are used as evidence, opacity can make them hard to test.

Automation bias. People tend to trust automated outputs and to look less critically at them. A link drawn by software, or a risk score, can carry more authority than it deserves.

False positives and rare events. Sommer and Paxson (2010) explained why automated detection performs poorly when the events of interest are rare and normal behaviour is highly variable: even a small false positive rate produces many false alerts, which can overwhelm investigators or lead them to dismiss real cases.

Garbage in, garbage out. Software cannot correct errors in the data it is given. Missing records, inconsistent formats and duplicate entities can all produce misleading results.

Deskilling. Investigators who rely heavily on tools may lose the ability to understand and check the underlying data. Heuer and Pherson (2015) emphasise that analytical tools support, but do not replace, structured thinking about hypotheses and evidence.

Using software responsibly

Several principles help investigators gain the benefits of software while managing its risks.

- **Understand the tool.** Know what each tool does, what it does not do and how it can fail.
- **Validate.** Test tools against data with known answers, and check important results by another method.
- **Record settings and versions.** Document which tool, version and settings were used, so that the analysis can be repeated and reviewed.
- **Keep humans in the decision.** Treat software outputs as leads to be checked, not as conclusions.
- **Sample the negatives.** When software filters out material as irrelevant, review a sample to estimate what may have been missed.
- **Be ready to explain.** If results are used as evidence, be able to explain in plain language how they were produced. Courts applying reliability tests, such as that in *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993), may ask whether a method has been tested and has a known error rate.

Choosing and procuring tools

Choosing software is itself an investigative decision, because the tool shapes what the team will see. Before buying or adopting a tool, an organisation should define the questions it needs to answer and the types and volumes of data it expects to handle, then test candidate tools against realistic sample data rather than relying on vendor demonstrations.

Useful questions include whether the tool can export results and audit logs in open formats, whether it records who did what and when, how it handles data it cannot parse, and whether the vendor will explain its methods if results are challenged in court. Licensing terms also matter: a tool that is licensed only for a single workstation, or that stores data on a vendor's servers in another country, may create problems for confidentiality, data protection or later disclosure. Smaller teams can often meet most needs with well-documented general tools, such as spreadsheets, databases and open-source analysis packages, provided they record how they used them.

Training, governance and security

Software is only as reliable as the people who operate it. Investigators should be trained not only in how to use a tool but in how it can mislead, for example by showing them cases in which a tool missed evidence or produced a false link. Organisations should keep a register of approved tools and versions, assign responsibility for testing updates, and require a second person to check key analytical results before they are relied on. Because investigation tools hold sensitive data, access should be limited to those who need it, activity should be logged, and data should be deleted when the case and any retention period end. These controls protect the people whose data are processed and make it easier to defend the investigation if its methods are later questioned.

A short example

The following hypothetical case illustrates these principles. A company's internal audit team suspects that an employee in purchasing has been approving payments to a fictitious supplier. The team loads three years of payment data into an analytics tool, which flags 1,200 transactions for further review based on rules such as round amounts, payments just below the approval limit and suppliers with no tax registration.

Rather than reviewing all 1,200, the team examines the rules and finds that one rule, for round amounts, generated 900 of the flags, most of them routine payments such as rent. After adjusting the rule and testing it on a sample of known legitimate payments, the number of flags falls to 140. Entity resolution then links one supplier's bank account to the same account number as an employee's payroll record, despite differences in the names. The team confirms the match manually against bank statements, interviews the employee, and documents every step, including the tool version, rules and adjustments, in the case management system. When the case reaches a disciplinary hearing and then the police, the analysis can be explained and repeated.

Conclusion

Analytical software is now indispensable to investigation. Case management systems keep investigations organised, forensic tools extract evidence from devices, document review platforms and technology-assisted review handle huge collections, network and geographic tools reveal relationships and patterns, and analytics tools flag anomalies in large datasets. Evidence shows that tools such as technology-assisted review can outperform manual methods. But software also introduces risks: opacity, automation bias, false positives, data errors and loss of skill. Investigators who understand their tools, validate their results, document their methods and keep human judgement at the centre can use software to see more without believing more than the evidence supports.

References

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255. https://doi.org/10.1214/ss/1042727940

Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review, 82*(5), 977–1008. https://doi.org/10.1177/0003122417725865

Casey, E. (2011). *Digital evidence and computer crime: Forensic science, computers and the Internet* (3rd ed.). Academic Press.

Chainey, S., & Ratcliffe, J. (2005). *GIS and crime mapping*. Wiley.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.

Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012).

Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Eck, J., Chainey, S., Cameron, J., Leitner, M., & Wilson, R. (2005). *Mapping crime: Understanding hot spots*. National Institute of Justice.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.

Garfinkel, S. L. (2010). Digital forensics research: The next 10 years. *Digital Investigation, 7*(Suppl.), S64–S73. https://doi.org/10.1016/j.diin.2010.05.009

Grossman, M. R., & Cormack, G. V. (2011). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law & Technology, 17*(3), Article 11.

Heuer, R. J., Jr., & Pherson, R. H. (2015). *Structured analytic techniques for intelligence analysis* (2nd ed.). CQ Press.

Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). *Guide to integrating forensic techniques into incident response* (NIST Special Publication 800-86). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-86

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Nigrini, M. J. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*. Wiley.

Pyrrho Investments Ltd v MWB Property Ltd [2016] EWHC 256 (Ch).

Quick, D., & Choo, K.-K. R. (2014). Impacts of increasing volume of digital forensic data: A survey and future research challenges. *Digital Investigation, 11*(4), 273–294. https://doi.org/10.1016/j.diin.2014.09.002

Ratcliffe, J. H. (2016). *Intelligence-led policing* (2nd ed.). Routledge.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy* (pp. 305–316). IEEE. https://doi.org/10.1109/SP.2010.25

The Sedona Conference. (2018). The Sedona Principles, third edition: Best practices, recommendations & principles for addressing electronic document production. *The Sedona Conference Journal, 19*, 1–207.

Xu, J., & Chen, H. (2005). Criminal network analysis and visualization. *Communications of the ACM, 48*(6), 100–107. https://doi.org/10.1145/1064830.1064834