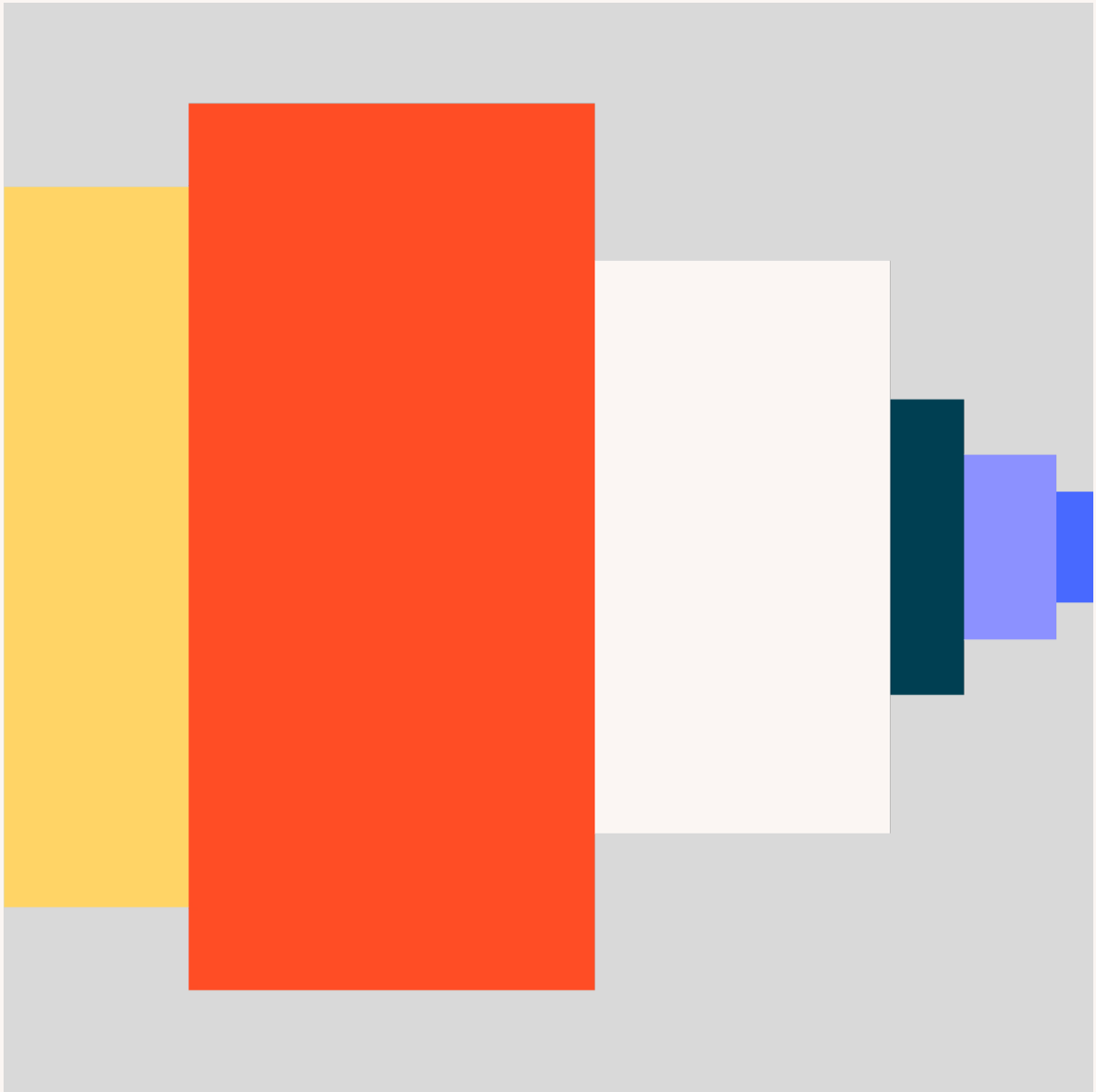
Surmounting a large data dispute



This case study examines our approach to a significant data dispute that has become a model for our clients, giving them a competitive edge.

Reading time	12 minutes	Page count	20 pages	Word count	2,019 words

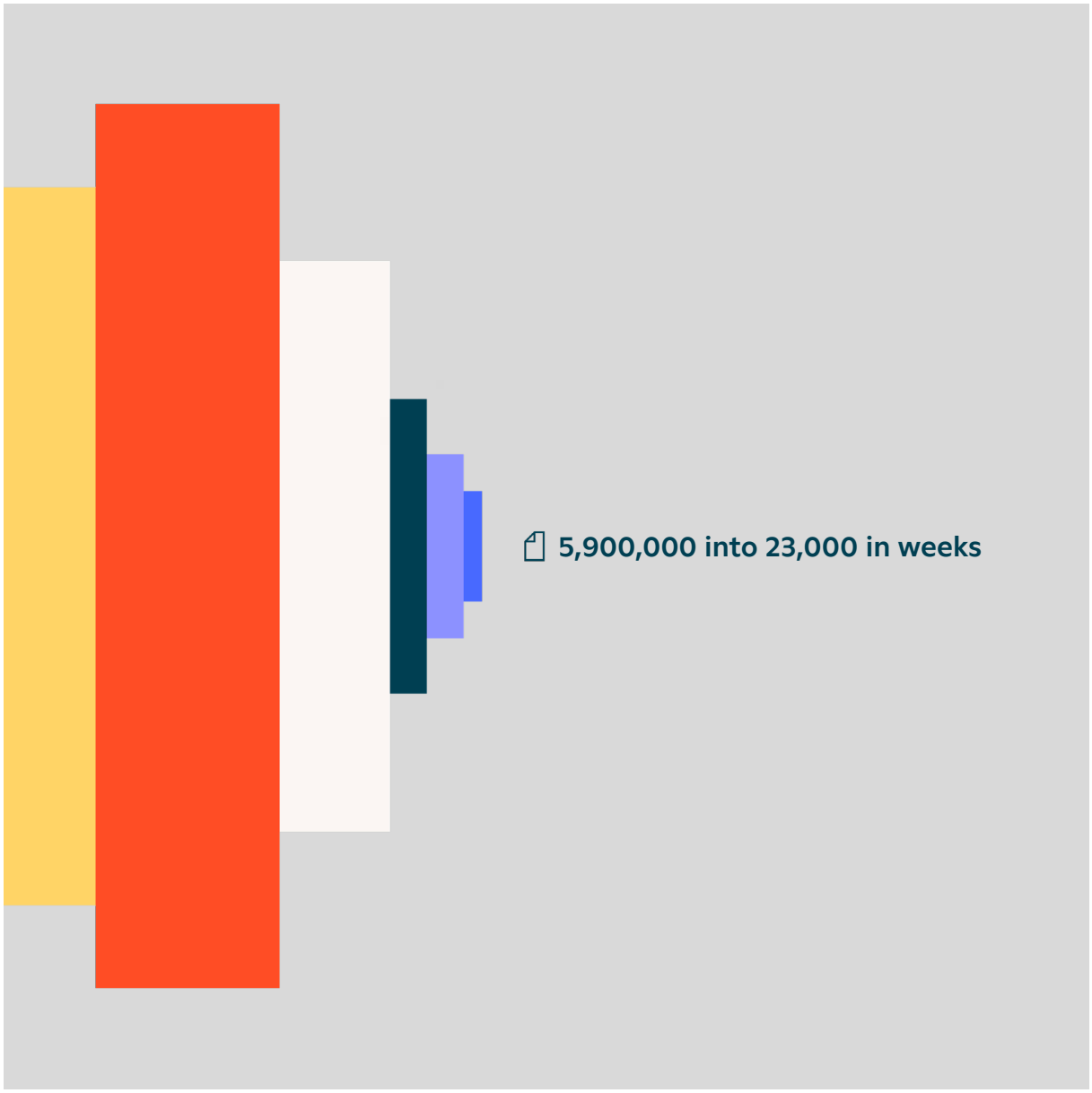
Surmounting a large data dispute with almost 6 million files.

The discovery process for a large Supreme Court matter presented over 1TB of data. After an initial audit, it was agreed that enhancements needed to be made to the collection methods. This increased the amount of data to 2.2TB, which comprised almost 6 million files for processing.

To tackle this gargantuan dataset, the Sky Discovery team deployed tactics like targeted collection, complex keyword strings and date ranges, data cleaning tools, and machine learning to help create review efficiencies. Not only did this increase the efficacy of the discovery process, but it also accelerated the workflow and helped reduce the cost.

Fast Facts

- Sky assisted the client's internal IT with data collection to ensure accuracy and defensibility
- 1.1TB of data was submitted for processing
- Early culling workflows saw 5.9 million files reduced to about 180,000 potentially relevant files
- The data was processed in tranches to allow for quick access to priority files, with the total dataset processed and available for review in just a few weeks
- The final submitted discovery list included 23,000 files
- Due to the efficiencies achieved through file processing and review, actual review fees were $170,000 lower than the low-range estimate and over $600,000 lower than the high-range estimate
- This approach to a significant data dispute that has become a model for our clients, giving them a competitive edge and winning them several new relevant matters.



"Sky Discovery is a true trailblazer in its field and should be commended for its hard work and dedication to providing quality services. I know that the team at Sky work hard and genuinely cares about fostering relationships with the people behind the business."

Senior Associate Large National Class Action Firm

This case study relates to a Supreme Court matter involving multiple parties and a quantum in the hundreds of millions of dollars.

All figures are rounded approximations, and key details have been anonymised to preserve confidentiality.

Approach

To start, Sky Discovery audited all collated data and identified information that was collected incorrectly. The end-client was tasked with re-collection to ensure everything was processed accurately and in a defensible manner. This set the review workflow up for success.

The vast volumes of data for this matter were processed in tranches. The legal team prioritised the tranches to suit their review priorities best. This also meant there was little delay with the review and investigation, which began shortly after the data was provided to Sky Discovery.

Next, Sky Discovery cleaned the entire dataset by using de-duplication (which removes all identical files) and a junk file removal tool. At this juncture, the parties agreed on a discovery exchange protocol. Over 60 data categories were agreed upon. The parties then negotiated critical date ranges and a series of keyword strings to establish relevance within these agreed categories further. Despite delays from the other party, Sky Discovery's crucial technical assistance ensured the protocol aligned with best practices and facilitated an efficient, proportionate discovery process for their client.

Sky Discovery provided technical assistance in formulating keyword strings to match each category. The three key segments for keyword strings were structure, hit rate and redundant terms. These segments helped determine how effective the keywords would be in reducing reviewable document volumes in a proportionate and defensible manner.

The refinement of the keyword searches helped to target and significantly reduce the volume of potentially relevant files. These search terms were also used to tag whether a document was relevant to a particular category. When reviewing these files, the legal team would validate the category tagging as being correct or update it to the correct tag.

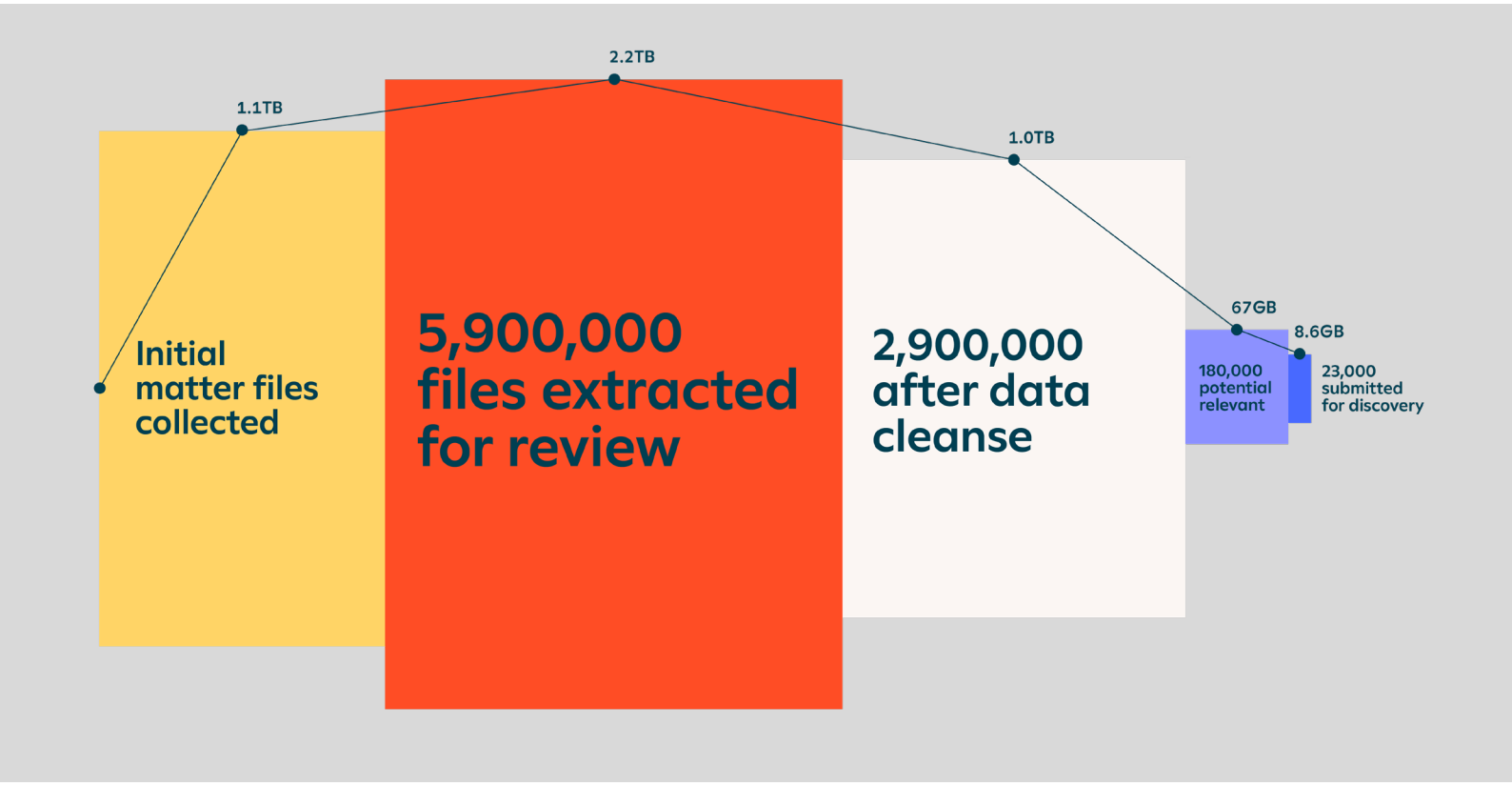
The parties also agreed to use a Continuous Active Learning (CAL) workflow to assist with the review of the relevant files. Essentially, CAL is a workflow where you train a statistical model on what is relevant. It then helps you review and identify highly relevant files early in the review process.

The CAL workflow was incredibly effective in assisting the legal team in reviewing approximately 20,000 files to hit the relevant accuracy metrics agreed upon between the parties for this process. About 10,000 files could not be reviewed by CAL (due to their format and/or content) and were manually reviewed by the legal team for relevance.

Finally, before producing the files for the other parties, several quality assurance steps were taken (such as the elusion test and privilege search checks) to ensure the accuracy of the CAL model and all other processes.

"Sky Discovery continues to combine and apply all the learnings from prior matters to ensure their work is even better with each case. There is simply no other provider in Asia-Pacific who can compete."

Partner Large International Firm



Initially, there were almost 6 million files, of which around 180,000 were run through a CAL workflow (a form of AI), with 32,000 reviewed by humans. The entire project took approximately 7 months; however, when we examine solely the process of managing the data once collected, it took only weeks, yielding the following results.

Results

- 1.1TB of data collected for the matter
- 2.2TB of data required processing, amounting to 5.9 million files
- 2.9 million files remained after the initial data cleanse
- 180,000 files were potentially relevant to the agreed search parameters
- 20,000 files reviewed by the legal team as a part of the CAL workflow
- 10,000 files reviewed outside the CAL workflow
- 1,500 files reviewed to test the accuracy of the overall review
- 23,000 files were listed in the final discovery.

Review process

The initial cleaning of the dataset greatly decreased the number of files needing review.

- 5.9 million files were reduced to 2.9 million files with the initial de-duplication and junk file removal measures
- 2.9 million files were then reduced to around 180,000 files by applying keyword strings and date ranges to the categories.

The legal team and the CAL model then went to work.

CAL continually learns and re-prioritises the order of files that the legal team reviews, based on what the legal team has deemed relevant or irrelevant. This means that the CAL model presents more relevant files to the legal team earlier in the review.

- 15,000 files were first reviewed by the legal team before the relevance rate began to drop off significantly
- A further 5,000 files were reviewed to ensure the team's confidence in the model's accuracy
- To validate the accuracy of the process, 1,500 additional files deemed irrelevant by AI were reviewed by the legal team using an elusion test
- None of these files were deemed relevant, yielding 100% accuracy for this final test.

"Sky Discovery has demonstrated a commitment to innovation and continuous improvement. They regularly inform us of new and improved products that meet the evolving discovery needs in large-scale litigation, enabling us to be more productive and efficient."

Senior Associate Large International Law Firm

Process costs

A total of 90 hours of consulting time were accrued by the Sky Discovery team in supporting this workflow, including collection, collaboration with the legal team for keyword strings development, set up and administration of the CAL workflows, result analysis, quality assurance and preparing the discovery set. In this time:

- 180,000 files were processed through CAL
- 145,000 files were deemed irrelevant by the CAL workflow.

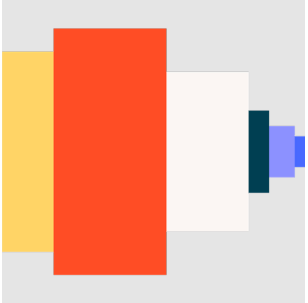
This process significantly reduced the human review costs of the legal team, resulting in savings between $170,000 and $600,000 for the client.

Legal team tasks

Throughout the process, the legal team was required to:

- Assist the end client in re-collecting any inaccurate data
- Finalise the key issues for review
- Draft the case summary and initial keyword search strings
- Work with the Sky Discovery experts to refine keyword search strings in line with initial search results
- Help draft a coding panel for human reviewers to analyse files prioritised using the CAL workflow
- Review the files prioritised by the AI-powered CAL workflow, determine if the document is relevant and confirm category tagging done through the AI coding process.

This all equated to significantly less time than would have been required without the use of AI.



Learnings

Work in stages

Large disputes can benefit from a staged approach. By operating in stages, outcomes from key workflow pieces inform the team's decisions on the next key step and allow them to tailor their approach. Sky Discovery's experts also look for opportunities for elements of the work to occur concurrently but do so only where that delivers true efficiencies for the review team. Regular updates, reporting and communication between the legal team and Sky Discovery's experts ensure that the process stays on track and meets the legal team's needs and objectives.

Cleaning is key

The original dataset was bloated with duplicates, junk and system files. By cleaning this dataset, Sky Discovery was able to reduce the initial number of files from 5.9 million down to 2.9 million. These steps aren't always common practice in the eDiscovery landscape, but with Sky Discovery, they are an essential step in ensuring quality and timely outputs.

Build effective search criteria

The quality of the review output is directly influenced by the preparation of the respective teams. This follows the principle of "garbage in, garbage out". More than 60 categories were used in this case, a considerably large number for any matter.

To efficiently sift through the categories, a critical date range and a series of keyword strings were tested before adoption. The legal team and Sky Discovery developed these keyword strings together. Not only did this help generate accurate results, but the efficiency of the discovery process was maximised.

Know CAL's limitations

Just like humans, machines have limitations. Not all files perform well with CAL. Photos are one example of this. To overcome this limitation, all photos within the dataset were isolated and reviewed manually. In this case, that included approximately 10,000 files.

Leverage new technologies

Modern eDiscovery technology can be used to evaluate the accuracy, completeness, and precision of your review. The elusion test is one such example. It is a judicially accepted practice, typically used to validate the accuracy of a CAL-powered review.

It can also be used to validate other AI-powered review processes and provide assurances, within an agreed margin of error, that a review workflow has uncovered all the relevant files.

Continue to use CAL post-discovery

Once you have built an effective CAL model, it can be used to review additional files you may receive. One such example is the collection of additional data tranches that may be identified and obtained by the client. Another scenario could involve incoming discovery received by other parties.

You can apply the established CAL model to identify files that are more likely to be relevant. You can also compare it to your existing document set to focus on those files that you are less likely to have seen within the existing set.

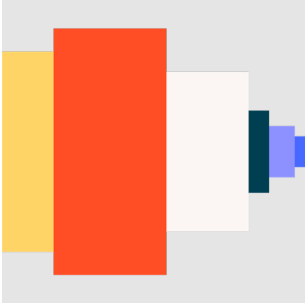
"Based on my contrasting experience with similar competitors, I believe Sky Discovery provides market-leading service and value."

Principal Specialist Construction Law Firm

eDiscovery is scalable

The volume of data in this case presented an unfathomable challenge for the legal team. By utilising a range of eDiscovery tools and AI-powered workflows, the team was able to reduce the number of files they had to review from 5.9 million to 180,000.

We are increasingly seeing cases where the average dataset is around 100GB, with the higher end exceeding 1TB, which translates to over a million files. Each case is different, but the nature of eDiscovery makes it easily scalable. What might work for a smaller matter might not work for a larger matter like this. That's why the Sky Discovery team tailor each step to suit the scope and scale of a case.



Conclusion

When it comes to speed, consistency and processing capacity, there's no comparison between the ability of a human and that of AI and CAL. When all parties work in tandem however, legal teams can effectively deploy machines to drive the discovery process.

The key to using AI in the discovery process is to combine it with existing, court-approved technologies, such as keyword searching and CAL, as well as robust review workflows. CAL does have limitations, such as the ability to review images and other obscure file types. However, other advances in AI may see this issue resolved or minimised in the coming year.

It's also integral to appreciate the role of talented technical experts who have extensive practical industry experience. The productive partnership between this case's legal team and the Sky Discovery team bolstered the quality of the prompts and process. Furthermore, this strong working relationship helped increase the speed and quality of the review.

This case clearly shows how, with our workflows and expertise, a large dataset of nearly 6 million files was efficiently reduced to an accurate and defensible 23,000 files for final discovery within weeks.

The cost and time savings of such a large endeavour cannot be understated. As technology continues to improve, the potential for eDiscovery in your workflow will only strengthen and grow.

"Since its inception, I have worked closely with Sky Discovery and am consistently impressed by their responsiveness, adaptability, and initiative."

Partner Large International Law Firm

At Sky Discovery, we focus on the technical solutions so you can focus on the law.

Almost everything is discoverable

There was a time when lawyers could focus more on the law. Time flies, and so does technology. In addition to the everyday challenges, the demands on lawyers for discovery and deadlines are intensifying with the ever-increasing volumes of data and data locations that may be relevant to a matter.

eDiscovery can be complex

While innovation will continue to be a priority for lawyers, the process of collecting and organising data and documents is often complex and inefficient. Furthermore, it is a specialised area that requires swift and defensible execution. In many ways, it is not actually lawyering.

You can focus on the law

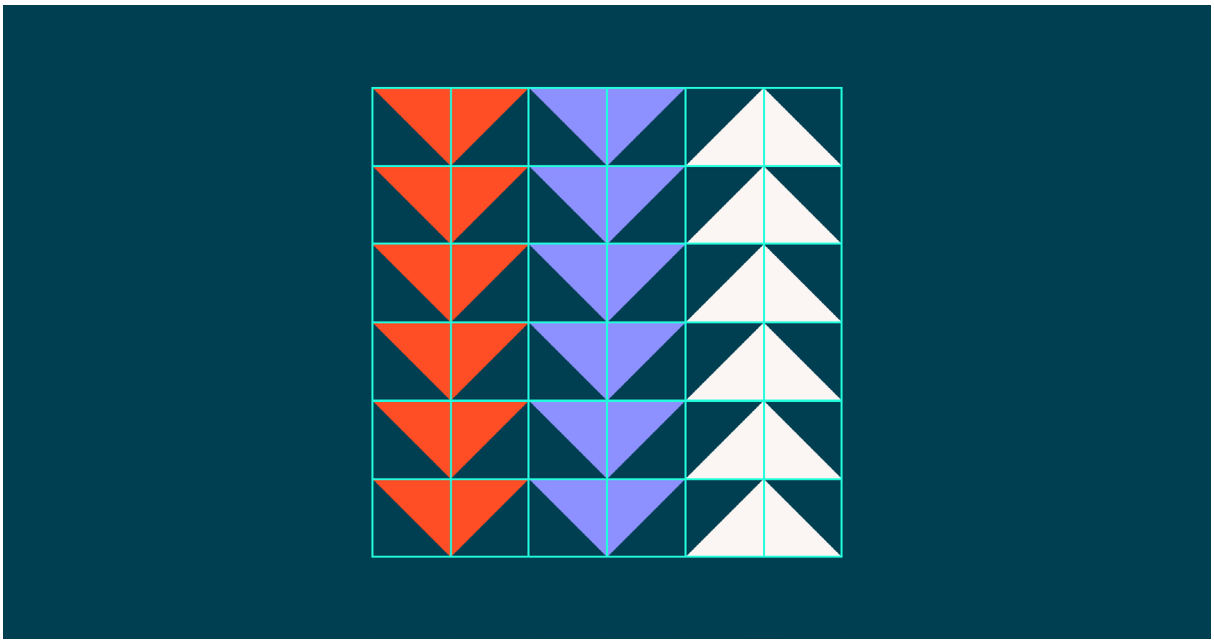
Since 2016, our mission has been to uncover what matters for lawyers and their teams. The volumes of discoverable data will continue to grow, while deadlines shrink. Every day, we see how lawyers who partner with us have more time to focus on the law and their clients.

We know how to implement systems that deliver quality work product faster and for less.

Discovery doesn't need to be an overwhelming, time consuming and expensive process. The solution is about working smarter with technology to turn the process from a burden into a winning advantage. Our proactive strategies start before the matter even begins and continue throughout every stage of the eDiscovery process. We use a mixture of traditional workflows as well as cutting-edge technologies. In some cases, our proven methodology sees us bill 40-60% less than our competitors.

Want to learn more about our proven solutions for reducing the cost of discovery, without compromising quality or compliance?

Start with our guide to reducing the costs of discovery ›



Find the right way to establish your own eDiscovery function

Tailor a model based on successful deployments for our clients across Australia

Projects

Designed to allow teams to access expert eDiscovery support for matter-specific projects.

Small
Matters-specific legal team projects with up to 100gb data.

Large
Matters over 100gb typically involves gigabyte and user charges for better commercial outcomes at this scale.

Largest
Store large data sets at low-cost with limited features, while key data is kept in a full-featured review workspace.

All-in
Customise rates to include some or all support as an alternative billing method, providing greater certainty.

Bundles

Organisational or team-wide support for consistent team and client access and adoption.

Small
Teams running a handful of matters. Includes small GB and user volumes.

Medium
Organisations with 10-20 matters across multiple practice areas. Includes 100s of GB and 50+ users.

Large
Organisations with multiple offices and practice areas, running 20+ matters. Includes 1000s GB and 100s users.

All-in
Tailor rates to include some or all support as an alternative billing method, providing greater certainty.

Includes gigabytes and users for contracted periods.

Partnership

Evolve a fully integrated function, with increasingly more autonomy via a less risky pathway.

No staff
Utilise profit-generating solutions without internal eDiscovery experts.

Small team
Sustainably build internal teams and workflows supported by Sky Discovery experts.

Large team
Respond to the demands of successful systems and scale with flexible solutions designed to foster growth.

Inclusions
Scope additional technology and services depending on the function you want to build.

Includes gigabytes and users for contracted periods.

All inclusions can be tailored to your workflow preferences, commercial appetite and future plans. Learn more ›

Whichever pathway you choose, your eDiscovery is better with our solutions, support and security.

✓ Access to our market-leading experts across consulting, data, innovation, IT and commercial.

✓ RelativityOne or Sky Cloud with advanced custom technology, growing AI capabilities and more.

✓ ISO 27001 ✓ SOC 2 ✓ PCI DSS ✓ HIPAA ✓ GDPR ✓ DPF

Helpful resources

As specialists we continually invest in R&D and best practice so we can advise our partners with confidence. These insights culminate in helpful resources and references for lawyers and decision-makers.

Data Identification Questionnaire

Our questionnaire aims to help you quickly and accurately identify data potentially relevant to your matter. The information captured from key stakeholders will facilitate the development of a collection plan and enable its swift and defensible execution.

Learn what to consider

Draft Exchange Protocol (Australia)

This reference is used by our teams on most disputes in most jurisdictions within Australia. The template provides a starting point for developing a protocol that governs the exchange of documents for Australian disputes.

Learn what to consider

Practice direction by jurisdiction (Australia & UK)

Reference our index of all Australian and UK eDiscovery practice directions.

Learn what to consider

Learn more

Negotiating a document exchange protocol with opposing party	↗
Reducing discovery obligations with another party	↗
Developing an appropriate review workflow	↗

AI or otherwise, when new challenges arise, we find practical, accurate and defensible solutions.

Our growing AI capabilities

Chat

Summarise, translate and label documents using natural language prompts.

Mass Action

Prompt, record and reuse multiple document review queries simultaneously.

Validate

Review, summarise and fact check document references to supporting evidence.

Compare

Review and summarise document similarities or differences in a structured, searchable format.

Chronology

Organise, link, track and review an automated sequence of events from documents.

Retrieve

Gather insights across entire datasets and retrieve quick answers and key references.

Scan

Recognise and tag objects in images and convert them to structured, searchable data.

Extract

Capture and populate data from templated forms into structured, searchable data.

Transcribe

Extract audio and video files and organise them into searchable transcribed data.

Translate

Translate and maintain context in documents in their original format in up to 100 languages.

Review (aiR)

Locate material related to legal issues important to your case strategy.

eDiscoveryAI (eDAI)

First-level AI review, AI-backed early case assessment and automated privilege review.

Leverage AI on your next matter. Our solutions ›

Client success cases

Extracting data from forms using AI workflows

Investigation, Process, Enhance, Review, Sky Solution

Read

Leveraging continuous active learning in large scale document review

Dispute, Analyse, Review, Sky Solution

Read

Migrating an active eDiscovery project from another provider

Dispute, Regulatory, Investigation, Process, Sky Solution

Read

Client success ↗

You need a team with a balance of legal, eDiscovery and technology expertise, this is who we are.

Our expert team of lawyers and technologists are available to assist you with navigating all stages of your matter, from the first meeting, through scoping, to completion. We focus on technical solutions so you can focus on the law. Find out how we help.