Equipping eDiscovery for Data Disruption:

How a new approach to process helps eDiscovery teams reuse data, increase insights, and add value to the company's litigation-adjacent ecosystem.

This publication is a product of a Cowen Group-led cohort of industry experts and dives deep into the evolving challenges in eDiscovery, marking a crucial advancement in understanding and addressing the complexities of modern data types in legal contexts.

Contributors:

Josh Kreamer
Head of Legal Services, Astra Zeneca

Michael Mendola
Director of eDiscovery and Data Governance | Financial Services, Discover Financial Services

Briordy Meyers
Discovery Counsel

Major Baisden
CEO, Lineal

Jeffrey Salling
Chief of Staff & Sr. Director Legal Operations, Moderna

Bradley Johnston
Legal Chief of Staff & Director of Legal Operations, SunPower Corp.

Stephen Aaronson
A Legal Data & Technology Director at a Top 25 global biopharmaceutical company

Sam Davenport
Senior eDiscovery Attorney, Ropes & Gray

THE COWEN GROUP

New Kinds of Data Present New Challenges for eDiscovery Teams

In the kind of four-by-four Zoom grid that passes for a modern conference room, one of the participants is laying out the chief issue:

"Just when we've addressed most of the permutations in workflows we use and we've standardized our processes, the very form of the data – the things on which we built most of our assumptions – changes completely, and it's like we're back to square one. And this keeps happening again and again."

The other participants in this virtual room, all of them heading eDiscovery teams at large organizations, agree in a mix of nodding heads, thumbs-up reactions, and voiced assent. Another says,

"The EDRM model is fine for giving the 10,000-foot view to someone who has never heard of eDiscovery, but it can't help my teams troubleshoot a TAR 2.0 review or help me figure out how to efficiently review a database with 15 years of relevant financial data."

It's clear, at least to this group, that something needs to change. The old ways of working aren't working anymore.

Between August and November of 2023, Lineal commissioned a bench of professionals to delve into the challenges facing today's eDiscovery teams. These professionals donated their time and combined decades of eDiscovery experience in law firms, pharma, tech, life sciences, and other litigation-heavy industries to better identify how the current eDiscovery practice model is insufficient and whether its traditional processes are too ill-defined to manage today's increasingly prevalent data types. After hours of brainstorming, sharing war stories, and collaboration, several strong themes emerged.

The eDiscovery Funnel Is Designed for Single-use and Self-contained Data Sets

The historic model of eDiscovery is a funnel. Data is non-destructively replicated and dumped into a processing engine where the data is pressed through scope filters, and review data called "documents" emerges from the other side. This mass of data-cum-documents is captured – mostly as text, natives, and static images – and teams of attorneys click through records of documents one after the other. In the end, the documents are tagged, stamped, and released as binders for the next step in the legal process.

A funnel is a one-way, linear operation. Everything else that is not produced is locked away and eventually discarded. Insights gained in later mechanisms of a funnel are typically not extracted and re-associated with the original source data.

This process has been faithfully used since the *Zubulake* rulings when electronically stored information (ESI) first began to displace paper documents as the discovery norm. Electronic mailboxes and software office documents were collected, processed, imaged, and reviewed in platforms built for data that is *single-use* and *self-contained*.

The review of the data (documents) is *single-use* in that once the reviewed copies of documents have served the instant purpose of the investigation or litigation, the volumes of productions and the reams of accompanying reviewer coding have no further utility past attesting to how the review happened. In a matter of time, they are virtually shelved or actually destroyed.

The review of the data (documents) is *self-contained* in that only insights inside the hermetic funnel can be applied to the data (documents). Whether a certain tranche of data from the financial system will always be non-responsive to a particular kind of matter is immaterial until the domain or keyword or other kind of filter inside of the funnel makes that non-responsive conclusion real and re-proven. If certain on-face, non-responsive data sets are routinely collected, the funnel exercise must be repeated over and over again.

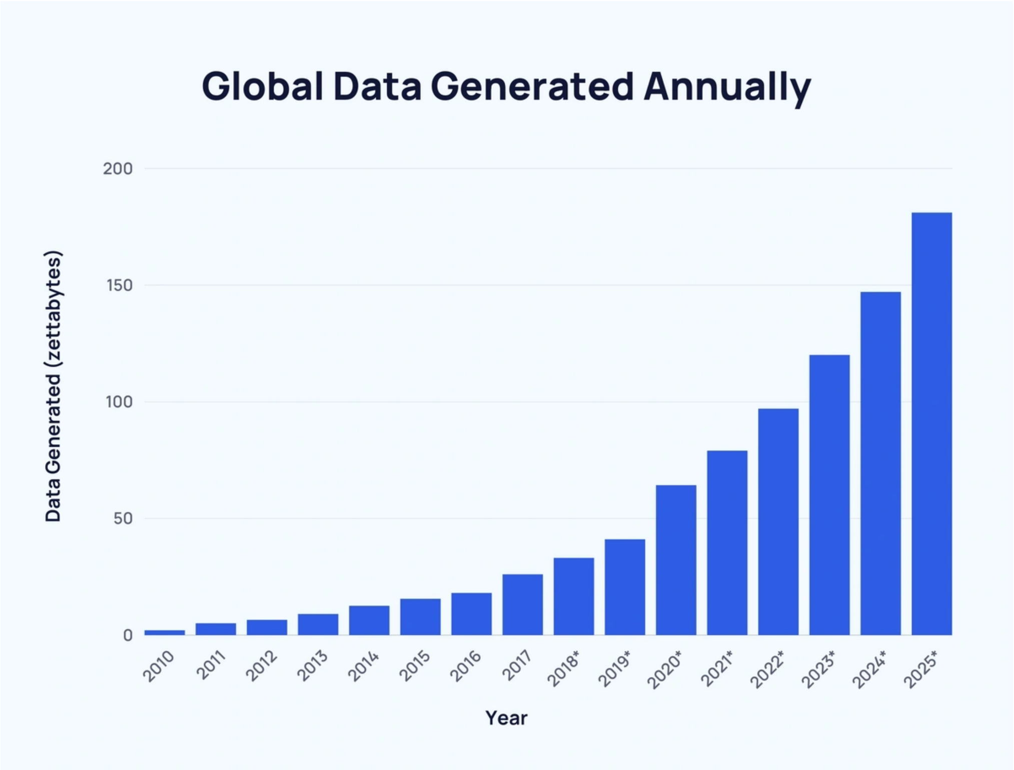
And this historical process has been fit for purpose...until the data changed and disrupted the funnel.

Data Volume, Variety, and Velocity Have Changed Since *Zubulake*

Practitioners built the historic eDiscovery model to convert *unstructured* data (email, images, and office documents) into *structured* data (document review metadata and coding) using metadata, body text, and OCR extraction, and then produce that database as *semi-structured* data (document productions with associated metadata) fit for a fact-finder's consumption.

However, this approach did not envision terabytes of *structured* financial data in formats that cannot easily be captured for a fact-finder's consumption. It did not envision collections of asynchronous, multi-author chats with no clear demarcation of what constitutes the beginning and end of "the document." It did not envision a shared document belonging to multiple authors in multiple time zones wherein, when additions to the document were made, who made them, and in what sequence, are all material to the document's review.

Anyone who has owned one of the nearly 20 generations of smartphones issued since the mid- 2000s doesn't need proof that data volume, variety, and velocity are exploding.



From 2010 to 2023 the global amount of data created grew from 2 to 120 zettabytes. One zettabyte is sextillion bytes – or a one with 21 zeros and seven commas attached!

Put another way: If each terabyte (the amount on a standard 1,000-gigabyte solid-state hard drive that can hold two Apple MacBook Pros) in a zettabyte were equal to a kilometer, then a zettabyte would equal 1,300 trips to the moon and back.

Source-type has also exploded. In 2004, email was converted into images, and some office documents were imaged or produced natively. Today, a review or investigation might have data from collaboration platforms like Slack, Microsoft Teams, Zoom, Salesforce, Yammer, or WebEx; mobile and text messages containing persistent and ephemeral data from WeChat, Snap, WhatsApp, Signal, or traditional SMS data; financial platform data from Bloomberg, SAP, Symphony, or Yieldbroker; audio and video files transcribed from remote meetings; or even audit logs from enterprise tools, SQL databases, and pure computer code and JSON files.

As data volume and type increase, the strategy eDiscovery teams use to process that data must also change. The typical EDRM workflow of duplicating (via forensic image) and serially processing all collected data can easily be an expensive logjam when applied to today's data volume, variety, and velocity.

Filtering out objectively non-responsive data through methods such as early case assessment and technology assisted review (TAR) nibbles at the edges of the problem. However, even these methods typically speed a review, but do little to reduce the collection effort.

Given the exploding volume, variety, and velocity of new data and new data types, the historical process is no longer sufficient. Data volumes are too large, varieties too many, and velocity too rapid to merely copy and pour everything into a funnel that, at best doubles data size, and at worst chokes on its own contents.

Moving Past the Funnel Requires New Process Strategies

Back in the Zoom meeting, more colleagues have shared stories of recollecting the same data multiple times. One member explains that her eDiscovery team worked with a provider to collect the same certain large subset of clearly non-responsive (mobile/chat/finsys) data multiple times.

Each time, it was processed by the provider and subjected to search terms. Each time, it was promoted to the review where it collected hosting costs. Each time, it sat in the review until either the reviewers had marked it all non-responsive (an expensive waste of time) or the TAR model pushed it so far to the bottom of the potentially responsive ranks that no reviewer ever touched it (a slightly less expensive waste of time).

"Of course," she said, "even if we'd known we were going to review it repeatedly, we couldn't just insist it's presumptively non-responsive. We'd need the list of doc IDs or something definitive so we could exclude it from promotion to the review. And if another provider processed the data, then we'd need to do some sort of custom md5 hashing."

That problem set off a flurry of suggestions from the group: ***"If we tracked the source location to the right metadata fields, then we could hash a unique identifier." "If we could link that identifier with the review coding…" "Maybe you could strip out natives and just reassociate images to the right hash values."***

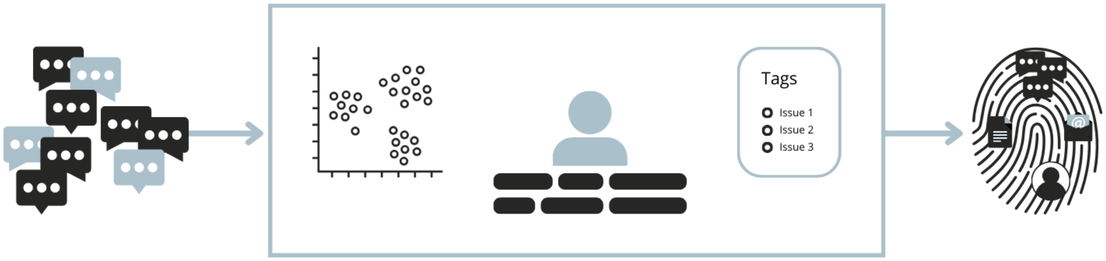
The group dove into the extensive toolbox of the modern eDiscovery team and started dreaming up workflows not yet contemplated by the current EDRM model. They proposed strategies that, if employed, might significantly reduce the cost and complexity a corporation typically suffers during eDiscovery.

One of the themes running through the groups' dream workflows was the idea of keeping data somehow *associated* with the data source. That is, if teams could keep an invisible thread attached between the data source and the data subset pushed through the funnel, then at every step – where new insights are discovered about that data – they could record those insights. That way, if it were collected again, so much more would already be known about the data subset. including whether it should be excluded from the review, whether it contains a large slice of privileged data, whether it could be stripped of natives, and on and on. *This invisible thread that transmitted funnel feedback to the source data could enable many advantages and abilities for the teams to exploit.*

The group conceived of two primary ways that a piece of data, its coding, and its insights could all be linked to its source data. One was a fingerprint. The other, a recipe.

Borrowing from relational database schema, the *fingerprint* is a unique identifier. If the team identified appropriately unique attributes about a given subset of data (say a document, or a range of chats, or table in the financial system), they could use an industry standard hashing algorithm (like SHA12 or MD5) to create a *fingerprint* for the range of data.

This fingerprint could identify a single email, but it could also scale up to identify a collection of emails, or a collection of financial system data. The teams could draw many kinds of boundaries around the data and then follow it into the wild the way ecologists might tag and follow a wild animal of interest and learn how it reacts to its environment of responsiveness coding, concept clustering, or issue tagging.

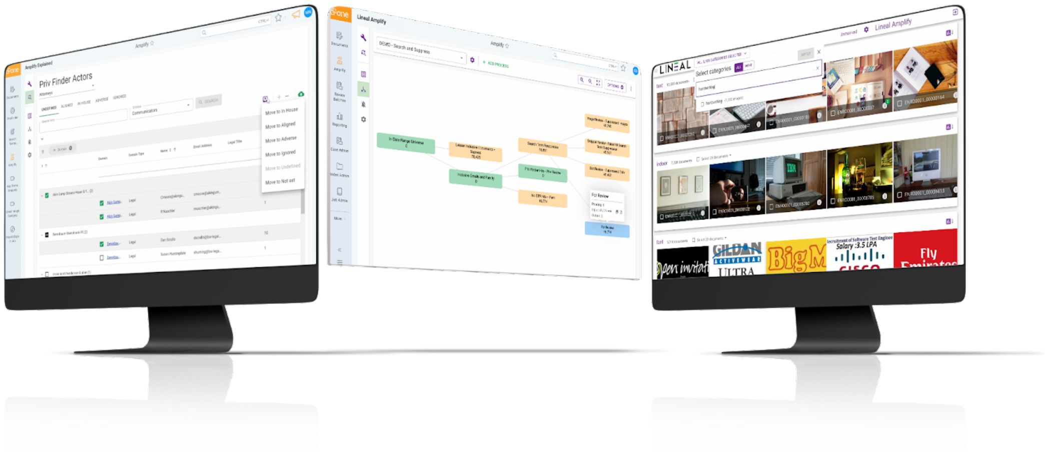


In contrast, but just as traceable and defensible, the *recipe* is a stored procedure like a database engineer might use to call a certain structured query. It is a pre-decided set of steps that always leads to the same location. Starting when the source data is forensically imaged, the data subset is followed and each subsequent action (where it is copied from, where it is copied to, how it is processed, how it is promoted, how it is coded, how it is produced, etc.) is recorded.

Once the string of actions is recorded, the *recipe* can be reverse-engineered to analyze any insight gained in the steps from imaging to production. With such a *recipe*, a team could reliably forecast – years later – how much of an identical collection will end up in review for a completely different matter.

Stephen Aaronson, a Legal Data & Technology Director at a Top 25 global biopharmaceutical company, is hopeful about the chance to reuse data and explains the potential gains for the in-house team from the successful use of fingerprints and recipes.

"If data is labeled/tagged as privileged, it can be reused across numerous matters and can reduce the number of times that the data is reviewed. A data warehouse of fingerprints along with certain metadata is a novel concept that can act as an aggregator working across the numerous eDiscovery tools/databases. A data warehouse with the fingerprints overlayed with an analytical tool could vastly improve the ability to reuse data across matters and reduce over costs."



This idea of subset-to-source-connection by fingerprints and recipes generated excitement in the group. But when asked why they could not do that today, answers popped up of the kind that often plague process engineering.

"I'd need to get buy-in from other risk teams – like information governance – to track the data." "We don't have the kind of database that this would take. And I can't find the type of talent I'd need to run it." "Where a repository like that would get stored, who would build it, and who would own it, all would frustrate multiple members of the steering committee."

Aaronson is well-versed in objections akin to the ones above. He outlines the practical and compelling forces that conspire against adoption of new tools.

"Often in complex multi-source eDiscovery environments, data processing, review, and storage tools are not standardized. Tools in these environments are often siloed, which results in cost erosion and missed-opportunity costs due to the inability to reuse data and decisions across matters.

"The barriers to change are:

> ***1) committing the time and resources to architect and construct the warehouse while incorporating the usage into existing processes and workflow;***
>
> ***2) understanding the data model and how retrieve data from the various eDiscovery tools where data is tagged and ingest it into the warehouse;***
>
> ***3) having the right data science and analytics skills and capabilities to see this endeavor through;***
>
> ***4) data-minimization principles arising from data privacy laws and regulations.***

"Though certain anonymization techniques exist to encode the "fingerprints," regulatory schemes that take an expansive view of data

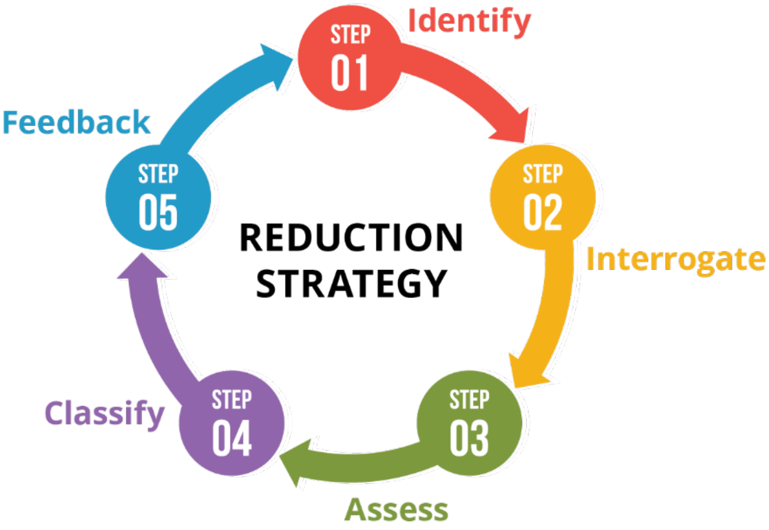
privacy, such as GDPR, may nonetheless limit our ability to preserve such data for reuse if it would not otherwise be subject to legal hold."

While these barriers are real and valid, one objection was more prevalent than the rest. **"My team is too heads down and just focused on the next collection. They tend to look forward and, unfortunately, looking backward really is not part of their process."**

To Find Fingerprints and Record Recipes, Teams Need to Look Backward

eDiscovery can be a reactive activity by nature. Even if both sides agreed to an ESI Protocol, and the Court entered it as an order, that plan will rarely wholly survive contact with reality. New information is a given. Surprises are a mainstay. Add in the cost sensitivity around the activity itself, and it stands to reason that eDiscovery teams are quick to pivot, always trying to see around tomorrow's corner; they rarely have time or space to plan out a detailed meta-project that will track data from imaging to production.

But if teams could find the headspace in which to step back and look at the whole picture, what activities and motions would they need to adopt so that they could create unique identifiers and stored procedures of the kind referenced above? The group discussed five such motions: *identification, interrogation, assessment, classification*, and *feedback*. To some extent, these individual motions are already part of the historic model of eDiscovery, but how they work together and inform each other may be new.



Identification involves locating and recognizing potential relevance in a data subset's source. This location may be a path to a specific folder, or it may be a table in a specific database schema. The location is identified, and the team notes what type of data is adjacent. In addition, the team records the data types contained in the subset and the matter types for which the subset is or is not relevant, all of which is often unknown until after processing.

So, while identification is often the first step, it does not close until after the data is processed and sometimes not until the data is produced. Identification involves adding metadata about the source that helps pinpoint specific documents, emails, databases, or other digital assets that may be pertinent to a matter type, be it a legal case or internal investigation.

Interrogation in this context is the detailed examination and analysis of identified data to extract relevant information. Specifically, it involves understanding the structure, containers, and relationships of the data. For example, consider Enterprise Resource Planning (ERP) software like SAP. Before assessing the data inside the ERP system, teams should ask: What modules is the responder using? Does the HRIS sit in, overlap, or integrate with the ERP? Are there custom Application-Program Interfaces (API) that change the data in a material way? Does the table recording salary "talk" to the table recording reporting relationships? Again, while this step begins after identification, it often does not close until after processing and sometimes review. This step is also a rich source for *feedback*. Linking the source data structure to attributes like likely matter types and common data pulls can dramatically save time and money.

Back in the Zoom meeting, things are heating up again. The leaders are dreaming about what they could do with rich and varied feedback.

"You know, building on this recipe idea, if I could preserve the insight about table location, I could save my team a lot of time! … And a lot of money for the company, too."

The other members want to know how.

"For instance, my team knows that the data in a particular range of our ERP table is always responsive to a certain common RFP. It's expensive to dump that module into the review every time. But the rest of the data in the table – and it's a huge table – is totally non-relevant. If I somehow could store the process of the queries and filters we use after we get that table into the review … if I could apply that process before it gets to review, before it gets to staging… before it's even extracted, then my team could cherry-pick the relevant piece and save time and money on the collection, the processing, and the review."

The process just described is the *recipe* concept from earlier. If a team can record and use the stored procedure of the filtering steps – a procedure they capture in *feedback* – and then apply those steps up front to the collection, processing, or review; and if at the same time the team can "tag" or *fingerprint* snippets of information by creating a custom hash value from different attributes of the data, then the team is perfectly placed to start to reuse all of this data without resorting to re-collecting and re-reviewing.

The team can start to reuse the data because they employed interrogation and the other steps in this cycle.

Interrogation is particularly key when what is being sought for relevancy is not a "smoking gun" document, but instead a nexus of various discrete pieces of information. For example, to find three corresponding data points in a database, a chat log, and a Human Resources Information System (HRIS), requires querying, searching, and filtering, and then some kind of tagging and linking to "join" the data into one relevant fact.

Assessment is the evaluation phase wherein the relevance, importance, and scope of the identified and interrogated data are determined. This step helps in deciding which data should be promoted to review and considers factors like compliance requirements and potential evidentiary value. It is also the step in which the team determines if what they presume is in the data is actually there.

Classification means following the historical model to label data as responsive, non-responsive, confidential, privileged, or whatever coding the review team adopts. However, it also means associating that data *and its coding* with the source of the data.

Feedback is the motion of capturing insights and associating them with the data subset's source. While feedback is the primary "backward-looking" action that the teams follow, it does not have a specific order or place in the steps listed above. Instead, at any step, the team is encouraged to record the results and insights of identification, interrogation, assessment, and classification as *feedback* into a knowledge repository.

"Feedback is critical for clients seeking to accelerate cost and time savings in eDiscovery, especially when applied to an enterprise level," says Discovery Counsel, Briordy Meyers. ***"It represents the constant and historical curation of defensibly retained data through contextual valuation."***

Because the feedback happens throughout the steps listed above, and because the steps do not necessarily have certain and standard beginnings and ends, the process can be understood as a cycle. One rotation of the cycle will see a data subset identified, interrogated, assessed, and classified.

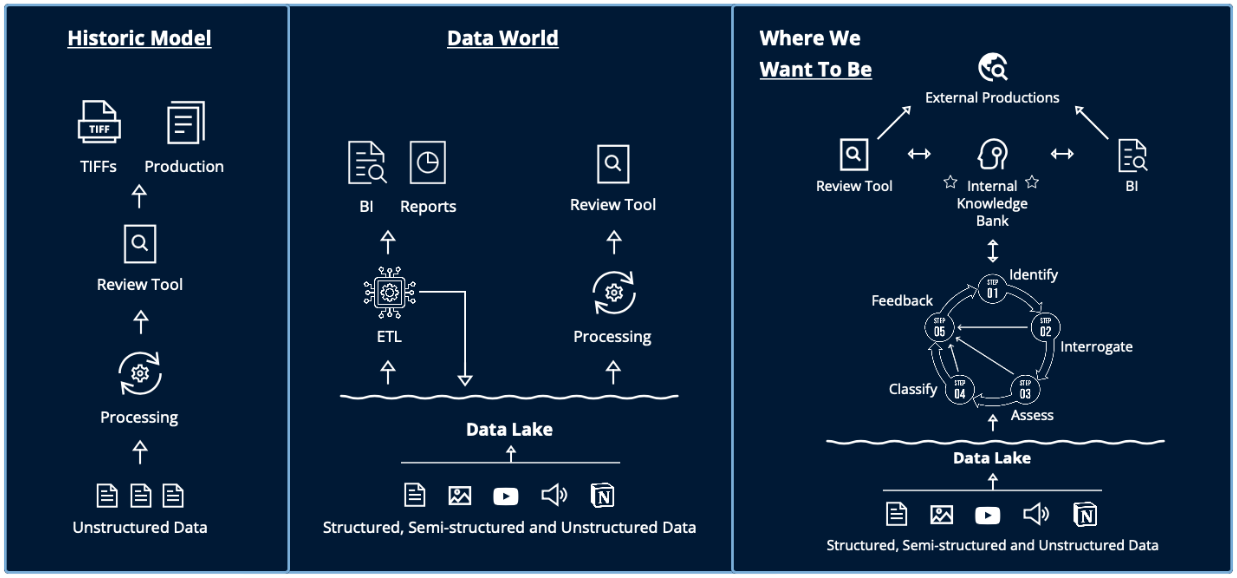
But depending on the emergent events in the legal case or investigation (e.g., an increase in custodians, or scope; a reordering of a TAR model; etc.), a data subset could go through the cycle multiple times. Each time it creates more associations with the subset's data source.

By using these steps, teams can link enough information and insight from the data subset to its source – effectively following it step by step – that they can assemble both the recipes and fingerprints needed to reuse the data.

Myers believes that ***"building feedback into a model allows for accelerated efficiencies both unique to the individual client and utilizable by parties and courts in a discovery context to fulfill the goals of discovery."***

Over time, the associations created in the steps above and associated with the subset's source become a valuable knowledge bank that teams can use to defensibly and traceably keep subsets of data out of review, or reuse coding from prior reviews without the need to process, image, or produce new documents. The cyclical approach promises savings to the companies and efficiency to the eDiscovery teams.

The Reuse of Data Scales eDiscovery Value to the Whole Company



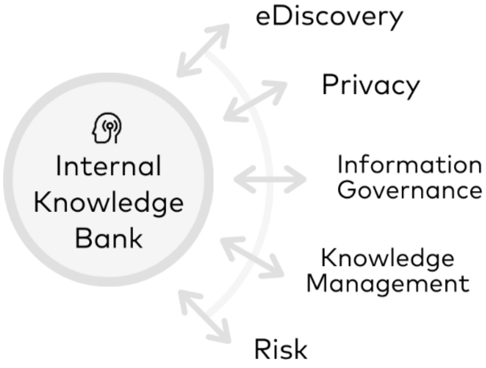
As mentioned above, many eDiscovery teams inherited the historic model and don't routinely practice the motions required to link data subset insights to original data sources. Especially when the team is small, the result is an efficient but largely *tactical* team.

To use a baking metaphor, these teams spend considerable effort gathering the right (custodian data) ingredients, perfecting measurements (ECA), and honing techniques (filtering and scoping) to bake a complex and difficult cake (review). But once it comes out of the oven and is pain-takingly frosted (for privilege) and decorated (for redactions), the team carves out a small piece (for production) and shoves the rest of the cake off the table and into the trash. Then they do it again with the next legal hold they receive. They bake and finish the whole cake just to keep the one slice.

As the team matures and grows, it develops standard templates and processes for custodian interviews, ESI protocols, coding templates, and other predictable pieces of the process. This standardization allows the team to ensure projects stay inside control parameters and do not get off track. It enables more predictable review budgets and dashboards that track crucial information. The team that plans, predicts, tracks, and adjusts is largely a *strategic* team.

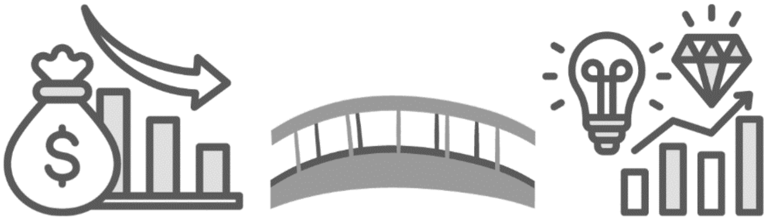
The strategic state is the goal state for most of today's eDiscovery teams. Until the last few years, the stabilized data variety, volume, and velocity did not call for anything else. Now, those goalposts have moved down the field, and a new standard is emerging.

The new standard for eDiscovery teams is one where eDiscovery sits shoulder to shoulder with Privacy, Information Governance, Knowledge Management, and Risk to actively manage data across the organization. One area where this convergence is already happening is in the application of mobile-device management policy to many areas of a company. Another is the process used to execute data-subject access requests from interested third parties or aggressive plaintiffs.



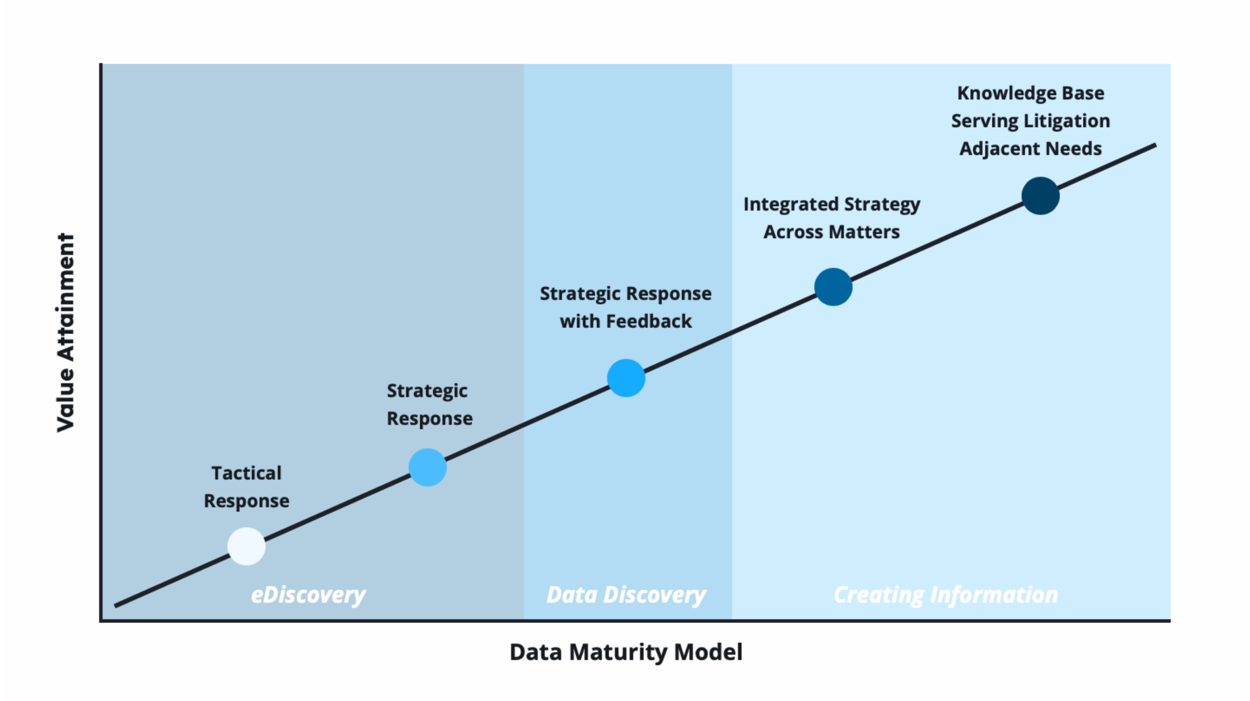
In each of these examples, the stakeholders must decide who owns which decision rights for what data at what point in the lifecycle. eDiscovery teams who are part of this ecosystem will face confusion and complexity around definitions, ownership, budget, standards, and processes. But those teams that can integrate successfully will bring subject matter expertise across matters and serve a variety of litigation-adjacent needs for the company.

Servicing these needs – customer privacy; data rights; cost management; environmental, social, and governance, etc. – will necessarily move eDiscovery teams from cost center to value creation. However, the historic model is insufficient to bridge a path into this new, elevated standard.



The bridge is the reuse of data. The reuse of data can move the eDiscovery team from a cog in the legal operations machine relentlessly chasing zero-overhead growth (ZOG) to a resource steward with a cross-functional seat at the strategy table and generating real value.

A strategic eDiscovery team that can link the insight and information gleaned in the imaging, processing, review, and production processes back to the source data as it exists and sits in the company’s day-to-day infrastructure, can make the transition from manager of the discrete review to co-manager of the company’s data.



Reusing data is the missing piece between a traditional eDiscovery team that stays strategic yet siloed, and one that elevates its value by managing large knowledge bases that integrate legal strategy across matters and serve many critical litigation-adjacent needs. In this latter model, not only does the explosion in data variety, volume, and velocity not frustrate, it in fact fuels the eDiscovery team to become an organization-wide stakeholder and true steward of company resources.

Quote from Brad Johnston – Director of Legal Operations at SunPower Corporation and former Senior eDiscovery Counsel at Cardinal Health.

"A distinct advantage of being a discovery professional sitting as co-manager of the company's data is that, as the company contemplates changes in the data infrastructure, the results can be fashioned in a way that is eDiscovery friendly. That is, the peculiar needs of the eDiscovery process can be addressed at the core of the data structure rather than as an afterthought when collection and processing occur, which nearly always is incrementally more expensive and time-consuming.

"Insisting on the reuse of data and the presumptive storage of the gleaned insights at the source data facilitates forging new and more advanced relationships with business partners across the enterprise. There are many benefits from these arrangements, including advanced notice of new data and/or business practices which allow pre-planning new eDiscovery processes, collaboration with other data stakeholders in the space in which they operate, identifying common themes for data hygiene across the enterprise and developing new skillsets for the legal professionals as they get broader exposure, as a team player, to how the managers of data in the company actually operate."

Conclusion

Reusing data is the missing piece between a traditional eDiscovery team that stays strategic yet siloed, and one that elevates its value by managing large knowledge bases that integrate legal strategy across matters and serve many critical litigation-adjacent needs. In this latter model, not only does the explosion in data variety, volume, and velocity not frustrate, but it in fact also fuels the eDiscovery team to become an organization-wide stakeholder and true steward of company resources.