AI AT K1X:

An Interview With Matthew Mahowald, Data Science Technical Lead

CAN YOU INTRODUCE YOURSELF AND EXPLAIN YOUR ROLE AT K1X?

Hi, I'm Matt Mahowald, I lead the machine learning team here at K1x. Here at K1x, we use machine learning primarily for identifying, categorizing, and extracting information from partnership financial and tax documents, such as Schedule K-1 packets. My team is responsible for designing, building, deploying, and maintaining the machine learning models our software uses to read and extract data from these documents.

WHAT INITIALLY ATTRACTED YOU TO WORKING IN AI, PARTICULARLY IN TAX COMPLIANCE SOFTWARE?

Artificial intelligence and machine learning have always been interesting to me. Prior to K1x, I was a partner at a small hedge fund specializing in equities and derivatives trading, so I have a firsthand appreciation for some of the challenges K-1 recipients encounter. The problem domain K1x is tackling is also very technically compelling: **tax and financial documents are a mixture of form data, semi-structured (tabular) data, and unstructured text.** To date most research and literature has focused on only parts of this problem (e.g. unstructured text), and is generally targeted at applications where accuracy does not matter as much as it does in our domain.

CAN YOU DESCRIBE WHAT K1X AI AUTOMATION SOFTWARE DOES AND WHOM IT SERVES?

K1x's software automates the extraction and aggregation of information from K-1 PDFs. We use artificial intelligence and machine learning models to identify data elements

> **Why [are tax forms] even a problem: the answer is that even today most K-1s are still distributed as PDF documents.**

inside a PDF, categorize them, and then map those values to a structured electronic format suitable for analysis and aggregation by other tools (such as by our K1x Aggregator software, but also even more mundane tools like Microsoft Excel). This is useful for anyone who participates in partnerships or other pass-through entities, as well as for the tax accountants that service those partnerships: one of our core value propositions as a company is that **it will be faster, cheaper, and more scalable to use our software to process your K-1s than it will be to do it any other way.**

WHAT WERE SOME OF THE CHALLENGES YOU FACED WHILE DEVELOPING AI TOOLS FOR TAX FORMS, AND HOW DID YOU OVERCOME THEM?

To start, it's perhaps worth pointing out why this is even a problem: the answer is that even today, in 2024, most K-1s are still distributed as PDF documents (or even printed, physical documents that are mailed to the recipient). Adobe created PDFs to solve the problem of consistent visual appearance of documents across any device or printer — that's what the name "portable document format" means — but the format was not intended for data transmission or manipulation. So, one way to look at the fundamental challenge we are trying to overcome at K1x is, we are trying to reconstruct the data that was used to create a K-1 PDF from the PDF itself.

Most of the challenges we encounter follow from this initial challenge: **PDFs are very flexible and support many different ways of displaying and presenting information, which we need to support.** There are almost no limitations on how information can be presented in whitepaper statements attached to a K-1, so we also have spent a lot of time and effort tuning our models to achieve good performance on a variety of different examples of whitepaper statements.

And of course another challenge is ensuring that our models are fast and responsive when our customers upload files for extraction. The more complex a model is, generally speaking, the more difficult it is to deploy and serve in production, and so we devote a lot of thought toward optimizing our model infrastructure to achieve the best experience for our users.

HOW DOES AI AUTOMATION CHANGE THE DAILY LIVES OF K1X CLIENTS WHO MANAGE TAX COMPLIANCE?

At a high level, AI automation is just like any other form of automation: **it takes some of the burden of busywork so that you can focus on higher value, more meaningful tasks.** The users of our software previously would be spending hours on a single partnership's K-1s manually keying in each value into Excel or other tax software before they could aggregate and prepare K-1s for downstream partnerships or taxpayers. Our software changes this from manual entry to a much faster human-in-the-loop review process so that our users can spend more of their time on areas where they provide differentiating value, such as determining reclassifications and adjustments based on the recipient's circumstances.

AI automation is just like any other form of automation: it takes some of the burden of busywork so that

YOU

CAN FOCUS ON HIGHER VALUE TASKS.

WHAT SETS K1X AI SOFTWARE APART FROM OTHER TAX COMPLIANCE SOLUTIONS LIKE OCR?

OCR — that is, optical character recognition — doesn't provide extraction or structured representation of data elements in a document. What OCR is used for is identifying text in images. So, if you have a scanned document, each page may be a bitmap or JPEG, and OCR will convert those images into plain text.

Modern OCR is actually pretty good: under the hood, it typically uses image recognition models based on convolutional neural nets to figure out which pixels correspond to which letters. However, being able to convert an image into something you could open up in Microsoft Word to copy-and-paste from isn't enough by itself, so tax-specific "scanning" solutions have to apply another step to extract the text into your tax software. This is typically done by defining, for each supported

tax form, which regions on the page correspond to which data elements. Then, when a scanned document is received, OCR is run to determine where the words are on the page, and a set of rules are used to check what text (if any) is contained in each of the defined regions for that template.

This approach works pretty well and we use analogous methods to extract data from the facepage of the 1065 and other forms. However**, where it falls apart is the semi-structured and unstructured contents of a K-1, that is, the whitepaper statements.** Here, you **can't** just grab all of the text at particular (x, y) coordinates on the page because each whitepaper statement is different — instead you have to use a model to categorize and map each data element on the page based on the text that surrounds it (for example, its description).

The ability to support data extraction from whitepaper statements is a key differentiating feature of K1x's software, and it is critically important for processing K-1s because, for most partnerships, the bulk of the information is contained in the whitepaper statements, not on the 1065 face page.

GENERATIVE AI GETS A LOT OF PRESS, WHAT ROLE IF ANY SHOULD IT PLAY IN TAX COMPLIANCE SOFTWARE LIKE K1X?

Generative AI is a very trendy topic right now and its capabilities are rapidly evolving, so I want to provide the disclaimer that what I'm about to say may no longer be true in, say, five years. For example, it's pretty likely that economies of scale combined with innovations in areas like edge computing (such as running models client-side rather than server-side) will push the costs of using generative AI (which currently is comparatively high) down over time.

> **The ability to support data extraction from whitepaper statements** is a key differentiating feature of K1x's software, and it is critically important for processing K-1s.

That being said, generative AI as it exists today has a few strengths and weaknesses that I think impact how and where it should be used in tax compliance. First, on the strengths side, **generative AI excels at creative tasks where you aren't especially sensitive to the quality of the results.** By "not sensitive to quality", I mean both cases where you don't care if the model is **bad** or **wrong**, but also cases where it's easy for you to validate and correct the model's output.

This might not sound like the perfect fit for tax compliance — and for many problems it isn't — but there are scenarios where it could be useful. For example imagine you have 10,000 K-1s. You might use generative AI to filter this list of K-1s to those containing cover letters, and then use the model again to summarize the contents of those cover letters.

Hallucinations (a generative model producing a factually incorrect output) are probably the weakness of generative AI that gets the most press, and for good reason. While there are techniques like RAG (retrieval-augmented generation) and, depending on the problem, sometimes postprocessing options as well, it's not currently known how to eliminate hallucinations in general. In fact, some luminaries in the field such as Yann LeCun believe that hallucinations are an intrinsic and unavoidable flaw built into the large language models powering generative AI.

From a practical perspective, the other big issue with generative AI is cost: at any of the major cloud providers, the cheapest GPU-accelerated SKU is commonly 10x or more expensive than a non-GPU option, and because of the size of the models, generative AI solutions often require sharding across multiple GPUs in order to run. Vendored solutions (such as OpenAI's GPT-4 Turbo) may be an option for some use cases, but that can get expensively quickly too — especially if fine-tuning is required to achieve good performance on the task at hand.

> When a generative AI model produces an output, it can be very
>
> **DIFFICULT OR IMPOSSIBLE**
>
> to understand where that output came from.

A final downside is interpretability: when a generative AI model produces an output, it can be very difficult or impossible to understand where that output came from. This is especially important for applications where accuracy really matters: **interpretable models are a better fit for human-in-the-loop workflows** because it's often possible to "peek inside" the model and correct any intermediate mistakes it makes. With a pure black-box solution, it's all or nothing: you either like what you get, or you have do the whole task yourself. Here at K1x, our goal is

> **Data extraction is an accuracy-sensitive task where you're not looking for creativity: you want to pull out data exactly as it appears in the text and map it to a constrained set of destinations.**

not to entirely replace the tax accountant with software — for one thing, the technology just isn't there yet — instead, it is to make them faster and more effective, and to do that you need to support a human-in-the-loop workflow.

That discussion might have gotten a bit in the weeds, so I'll just summarize by saying while there are legitimate use cases for generative AI in tax compliance software, there are also cases where its drawbacks may outweigh its benefits, and performance is not guaranteed to exceed that of more "traditional" machine learning approaches. There are even cases where a blended solution might make sense. It really depends on what problem you're trying to solve.

For K-1 data extraction specifically, I would ask myself, why pay a premium to ask a PDF nicely and hope it gives me an honest answer, when I can just get the answer? Data extraction is an accuracy-sensitive task where you're not looking for creativity: **you want to pull out data exactly as it appears in the text and map it to a constrained set of destinations.**

LOOKING AHEAD, HOW DO YOU SEE AI EVOLVING IN TAX COMPLIANCE FOR ALTERNATIVE INVESTMENTS?

It's still early days for generative AI with a lot of experimentation going on to determine where and how to use AI effectively. As the field matures, I expect to see successful patterns and best practices proliferate, tools mature, and the market consolidate as particular use cases and vendors win out. It's hard to say exactly how this will evolve, though. For example, right now many vendors offering specialist solutions for particular niches are relying on generic foundation

AI IS AN EVOLUTION of our existing tools and capabilities.

models (e.g. GPT-4) published by companies like OpenAI or Anthropic — will domain specialization continue to be a differentiating value for these vendors, or will further advances in the capabilities of foundation models cannibalize these smaller markets?

Alternative investment tax compliance is a complex space requiring significant domain expertise, so I expect that the future looks like AI features and functionality built into tax software to assist and empower practitioners, rather than any fundamental re-alignment of the shape of work to be done or the expertise needed to do it.

WHAT ARE THE CHALLENGES AND OPPORTUNITIES WITH WRANGLING ALL THIS ALTERNATIVE INVESTMENTS DATA AND WHAT OPPORTUNITIES ARISE THE INSIGHTS DERIVED FROM THIS MOUNTAIN OF DATA?

Quantity has a quality all its own: one of the lessons of the "electronification" of so much of finance and capital markets since the 1980s has been the incredible value and power of data. **Data unlocks insights** into portfolio performance, risk exposure, investment trends, and allows investors to be better informed. Data-driven insights can also drive operational efficiencies.

However, these advantages don't come free: **alternative investments encompass a wide range of asset classes**, each with their own data formats and reporting requirements, so data management at any serious volume is itself a challenge.

WHAT DO YOU WISH MORE PEOPLE KNEW ABOUT USING AI SOFTWARE IN TAX COMPLIANCE?

I think there's still a lot of confusion about what the capabilities and limitations of AI systems are. We're still a long ways off from any full automation of tax and compliance workflows: **AI should be viewed as a tool to assist tax professionals, rather than replace them entirely.** Human expertise remains important for navigating tax scenarios, interpreting results, and making strategic decisions. Overall, I think people should be excited by the possibility of removing some of the drudgery of repetitive manual tasks, but also realistic: **AI is an evolution of our existing tools and capabilities.**

ABOUT K1X, INC.

K1x is the leading data distribution platform for alternative investments.

K1x is the leading data distribution platform for alternative investments. The fintech company's patented, AI-powered SaaS solution digitizes and distributes data seamlessly–connecting investors, advisors, tax software, portals, accounting firms, IRS and state taxing authorities–simplifying complex processes, accelerating filings, reducing costs, and delivering greater control, transparency, and accessibility. K1x is battle-tested by the best, and trusted by more than **8000** organizations including **44 of 100** largest institutional investors in the US, **15 of the top 25** accounting firms, **11 of the top 100** private foundations, **36 of the top 100** university endowments, and **8 of the top 40** health systems. Visit us at **K1x.io** and follow us on **LinkedIn**.