|  |  |  |
| --- | --- | --- |
| **Preliminary Project Plan for the Creation of an**  **Industry-Leading Machine Learning Program**  **(Draft 1/28/2019)** | | |
| **Project Name** | Project Learned Hand |
| **Mission** | To create an industry-leading machine learning ("ML”) platform built around a highly automated ML pipeline designed to permit (i) rapid prototyping and evaluation of ML models, (ii) easy deployment of successful ML models within the firm’s existing enterprise architecture, and (iii) seamless integration of deployed ML solutions with other core systems (*e.g.*, document management database, email exchange server, accounting and time keeping software). The platform’s efforts will generally focus on advanced research involving the use of ML technology and natural language processing (“NLP”) to improve the delivery of legal services. The firm’s efforts will initially focus on the development of a number of individual ML models and solutions that are each designed to perform a narrow set of well-defined tasks. Development priority will be based largely on expected return on investment, which will be measured broadly to capture benefits from (i) improved efficiency, to the extent of improved margins, (ii) our ability to deliver better and more valuable services to our clients, and (iii) the reduction of loss prevention risks through improved quality control.  Where compelling economic or financial circumstances exist, the firm may also collaborate with one or more traditional providers of legal technology (*e.g.,* Thomson Reuters) to jointly develop, market and commercialize certain ML solutions. |
| **Purpose** | The Project’s primary objectives are to (i) provide the firm with a competitive advantage through the deployment of ML solutions not otherwise available through traditional vendor channels, (ii) generate additional revenue by licensing internally-developed ML solutions to corporate law groups, AmLaw 200 law firms and other providers of legal services unable or unwilling to make the investment of capital and resources to develop similar solutions, and (iii) jumpstart a sustained effort to deliver high-value practice tools to our lawyers, without further delays. |
| **Expected Benefits** | In addition to increasing the firm’s competitiveness and expanding potential sources of revenue, the firm can expect to benefit from the Project’s success in several other ways, including:   1. reassuring the firm’s partners and other stakeholders that firm management is responding aggressively to changes in the legal industry, especially those influenced by advances in technology; 2. positioning the firm’s technology-focused practices to grow revenue through engagements advising clients on the development, licensing and deployment of ML technology, which is an area expected to produce significant work (likely recession-proof) for both transactional practices (*e.g.*, M&A, technology licensing, outsourcing) and litigation practices (*e.g.*, intellectual property, data security and privacy) over the next several years, which also furthers existing efforts to grow our practice around autonomous vehicles; 3. generating positive media coverage and providing an opportunity to solidify the firm’s brand as a leader in legal innovation, which can be leveraged for business development opportunities and attracting new law school graduates and lateral candidates who desire to work for a firm that promotes a culture of innovation and provides its lawyers the best available technology to support their practices; 4. an opportunity to promote the fair and ethical use of ML technology by educating legal professionals and clients on the risks posed by ill-defined and/or poorly-managed ML projects. These risks, among others, include (A) failing to identify racial, gender or other inappropriate bias within training data that may carry over to the model’s predicted outcomes, (B) the development of ML technology for legitimate purposes that may also be used for purposes detrimental to society (*e.g.,* the same ML model used to develop gene-therapy treatments for viral infections might be used by others to create weaponized strains of genetically altered viruses, and (C) providing users with access to ML models without adequate instruction on their deployment, such as failing to advise users about known limitations on a model’s performance, or alternatively, any safeguards or other processes that are intended to be used by users to improve the model’s performance to acceptable levels; 5. reducing loss prevention risks relating to our own deployment of ML technology through the development of industry best practices, including (A) standards for evaluating the precision and recall of ML models designed to perform legal tasks, (B) guidelines for determining, on a task by task basis, the minimum precision and recall scores required to allow a model to perform a given task with little or no human supervision; and 6. educating and improving our lawyers’ understanding of ML technology; and thereby, improving our lawyers’ ability to identify potential opportunities to deploy ML technology within our practices. |
| **Background** | The legal industry continues to experience pressure on its traditional business model. These changes have left many firms seeking alternative ways to increase revenue. For some, the solution may be to match their subject matter expertise with technology. Firms that are successful in offering their services through technology offerings may be able to generate new sources of revenue and remain competitive even as demand for traditional legal services from law firms remains flat. The development of trained, machine learning models presents one of the most promising opportunities for law firms to leverage subject matter expertise through technology.  Over the last five years, the explosion of new entrants into the legal technology market has not led to any fundamental change in the practice. Law firms may hold the key to changing this dynamic. For example, most machine learning models require large amounts of training data. This training data is often difficult to compile. Law firms, however, generally have an abundance of documents pertaining to legal tasks are perfect for training models on legal subject matter. In addition, law firms have the subject matter experts needed to properly prepare and annotate such materials for training exercises.  The historic lack of law firm participation is indicative of the legal profession’s struggle to adopt an engineering approach to problem solving, which is the result of several factors. Of all these factors, three reasons stand out. The first is a misalignment of incentives created by the traditional method of billing on an hourly basis. An engineering approach requires a firm to make a capital investment that will ultimately allow it to perform client services better, faster and cheaper. A law firm, however, loses revenue if it utilizes the same process. In other words, inefficiency can often lead to greater profits. Unfortunately, the environment in which legal services have been bought and sold, especially in corporate America, did not promote efficient outcomes. The second factor was a cartel structure that was effective at keeping non-lawyer competitors out of the marketplace and effectively regulating the number of lawyers. The natural consequence was to create an industry less sensitive to price theory and basic microeconomics. These forces have given way to increased competition from the ‘Big Four’ accounting and consultancy firms, and countless others. The final factor is culture. The legal profession remained largely unchanged for decades, with each successive generation training the next one to practice in the same way as the one that taught it. The weakening of the first two barriers and a younger generation of lawyers unwilling to sacrifice quality of life has begun to soften this culture.  For those firms willing to make the investment, legal technology could be a significant new source of profits. Successful law firms are often a collection of highly productive lawyers, which means, as a group there is an enormous amount of subject matter knowledge and expertise. As such, well-respected law firms with thought leaders in their respective practice areas that are able to convert and organize this sought-after knowledge into software and other technology solutions will be able to increase margins and remain more competitive than their peer firms. Ultimately, and regardless of the success of such offerings, such efforts help ensure that a firm’s lawyers have access to the best solutions.  This opportunity exists in large part because of improvements in access to technology. In fact, it’s hard to overstate the improvement in performance of deep learning models over the last several years. Several factors have contributed to this improvement: (1) exponential increases in computational resources, (2) the development of simple code libraries and abstraction of complexity in model development, and (3) larger and better processed data sets. Like most advances in technology, this explosion in capabilities has also seen the cost of access plummet. Today, many of these algorithms and models have become standardized. Image classification, text classification, text to speech, speech to text, sentiment analysis, recommender systems and many other tasks; all have standard algorithms and models that are available to users as open source code. A word of caution—law firms exploring the internal development of technology should generally proceed only if a sufficient nexus exists between the proposed technology solution and the firm’s core practices and expertise. Law firms won’t fare well competing against the likes of Google, IBM, AWS, Microsoft and the other technology companies. We have no intention in fundamentally changing the firm’s core business, only the form in which we deliver those services.  One of the biggest challenges to the use of machine learning for NLP tasks has been the shortage of training data. Because NLP is a diversified field with many distinct tasks, most task-specific datasets contain only a few thousand or a few hundred thousand human-labeled training examples. However, the most promising areas of research for improving NLP are modern deep learning-based NLP models. Unfortunately, these models generally require even larger amounts of data, only improving when trained on millions, or billions, of annotated training examples.  For years, researchers unsuccessfully attempted to bridge this gap by training models on general purpose language using the enormous amount of unannotated text on the web (known as pre-training) with the hope that this general purpose model could then be trained on a more manageable amount of domain specific data (effectively transferring the prior training to a model then fit with incremental data building on the initial training data. Up until the last few months, these efforts were generally unsuccessful.  Last November, however, Google open sourced a new technique for NLP pre-training called “Bidirectional Encoder Representations from Transformers”, or BERT. With this release, anyone in the world can train their own state-of-the-art question answering system (or a variety of other models) in about 30 minutes on a single Cloud TPU, or in a few hours using a single GPU. The release includes source code built on top of TensorFlow and a number of pre-trained language representation models. In an associated research paper, Google demonstrated state-of-the-art results on 11 NLP tasks, including the very competitive Stanford Question Answering Dataset.  Firms will also need to be aware of the risks that the use and development of artificial intelligence by law firms raises. Many of these risks involve ethical issues with which lawyers are only starting to grapple. These include determining the appropriate liability for a poorly engineered solution. In the U.S., lawyers can’t generally limit liability. On the other hand, such limitations imposed by software engineers are generally enforced. What if a law firm designs a flawed system that causes a loss, but there is a limitation on liability in the engagement letter? What if the loss is not because of incorrect legal advice but rather a bug in the code? What if the software was separately licensed to the client by a company that the law firm formed to develop the software, which was fully disclosed to client?  In addition, lawyers must be aware of the risk of models that are biased against certain groups of individuals because the data used to train the model was itself the product of bias, or otherwise flawed. Care must be taken not to perpetuate continued bias. This issue will become more important as machine learning is used with greater frequency to influence credit decision, sentencing in criminal proceedings and other decisions that can have a significant impact on people’s lives. While these are all challenges, the potential reward easily justifies expending the resources to overcome them. Those that do will likely become the market leaders of the future. |
| **Strategy** | Educating our Lawyers about Innovation.  We will provide more education on the technology driving disruption in the legal industry. We hope that by having a better understanding of how a technology works and its potential uses and limitations, our lawyers will be better equipped to identify weaknesses in current processes and identify solutions to improve them. Topics will cover the basic principles of machine learning and natural language processing. Subsequent programs will cover (1) robotics and automation, (2) knowledge management, including intelligent Q&A systems, and (3) data science applied to the law—how to become a data driven lawyer.  Knowledge Management and Q&A Systems.  We have developed, internally, a Q&A platform that leverages state-of-the-art machine learning and other technology recently opened sourced by a company acquired by Facebook. The platform can be deployed in a number of different contexts, with its primary functionality being the ability to extract answers to natural language inquiries from documents stored in a designated database. Some examples of prototype systems we have developed include (i) email utility that scans inbound emails for questions and attempts to extract the answer from its corpus of documents, with the platform even autogenerating a proposed reply email, (ii) a compliance tool that allows a client to upload all of its compliance policies and manuals and be able to quickly answer questions about its policies, and (iii) uploading voluminous land development regulations and then being able to query the corpus regarding permitted uses and other entitlements under each specific zoning district—avoiding the need to search through hundreds or even thousands of pages. We expect to begin deploying the platform at scale middle of this year.  Development of domain specific ML Models.  After spending the last several months prototyping machine learning models designed to assist our litigators, we will be transitioning to a development program capable of providing machine learning solutions to them at scale. More specifically, we have been developing and training about a half-dozen state-of-the-art machine learning models, each for a specific litigation domain. Some of these models have achieved precision and recall scores higher than those developed by third-party vendors. Examples of models deployed or under active development, include:   1. models to extract relevant information from pleadings. Once classified, information regarding the complaint (e.g., accuser) is automatically extracted and passed to a simple document templating engine, which based on the classification of the claims, will generate an initial draft of the response, using the appropriate standard paragraphs and inserting party details and other extracted information into the template’s variable fields. 2. A similar technique was designed to increase the efficiency of several teams handling high-volume consumer finance defense work and insurance claim matters. We are expanding our library of trained models to several that can ingest PDFs of requests for production and interrogatories and draft initial responses, which include objections identified by the model based on training data consisting of thousands of prior responses with similar requests. 3. Sequence of models trained to identify, classify and draft separate amendments to LIBOR priced debt. 4. An automated workflow for the ingestion, organization and extraction of information from large collections of unstructured documents. The workflow uses a series of machine learning processes, including (i) Google's newly-released computer vision model, which achieves state of the art OCR performance, (ii) Latent Dirichlet Allocation (LDA) using Scikit Learn's open source framework, which we use to perform unsupervised, document-level classification, (iii) an ensemble of pre-trained NER models (SpaCy, NLTK, IBM’s NLU and AllenNLP) for the extraction of names, organizations, dates, and other information, and (iv) pre-trained NLP models offered through Amazon AWS’s Comprehend Medical service, which were trained on an extensive corpus of medical records. |
| **Short-Term**  **(1-2 months)** | Initial Organization. Finalize project plan and initiate project tracking through Department of Project Management.  Consolidation of ML/NLP Projects. Consolidate existing ML/NLP projects to ensure individual projects are aligned with overall strategic plan and prioritized accordingly.  All Partner Meeting (“APM”). Use APM to create initial buzz around eﬀort and increase willingness of partners to participate in idea generating workshops, pilots, and ultimately, the adoption of solutions developed under the Project (e.g., Machine learning and litigation hot topic table).  Project Sunset (ML-based LIBOR review tool). Complete development of prototype suﬃcient to demo to clients. Commence marketing campaign to sell solution to existing financial institution clients.  Project McGruff. Complete usable prototype for insurance fraud recovery team. | |
| **Mid-Term**  **(3-6 months)** | Initial ML Models. Establish aggressive roadmap for the development of a small number of high value (but low-complexity) pre-trained ML models (collectively, the “Initial ML Models”). The intention is to make the Initial ML Models available to public (via usable model endpoints (APIs) or in ‘pickle’ format, but no source code). The target release date is the start of the CLOC conference (mid-May 2019) to ensure maximum media coverage and exposure with corporate law groups and legal practice operations professionals attending the conference. We will also be hosting a workshop/panel at the CLOC conference on the practical considerations for developing and deploying successful ML-based solutions inside law firms and corporate law groups.  H&K ML Guidelines. Initial draft of H&K guidelines and policies for use of ML-based technology for legal tasks. While the guidelines will be primarily focused on best practices for the deployment of ML technology for use by lawyers, it will also include extensive discussions around the ethical and fair use concerns described in item (iv) under ‘Project Justification’.  Research University Partnership. Establish one or more working relationships (formal or informal) with research universities (*e.g*., MIT, Georgia Tech, Stanford) for the development of an annotated corpus of training materials intended to become the “benchmark” corpus for evaluating the performance of ML-models trained to perform legal tasks (the "Legal Domain Corpus”). A Legal Domain Corpus would be particularly valuable given recent advances in transfer learning in NLP/ML research and state-of-the-art performance of domain specific models initially pre-trained with word embeddings (*e.g.*, BERT, ELMo) from only a generic corpus. If successful, H&K has an opportunity to be credited with contributing towards a benchmark corpus that will likely be used for years as the standard by which legal domain ML models performing NLP tasks are evaluated.  Education and Training Program. Initial program materials prepared.  Ideation Campaign. Establishment of internal programs, workshops and other idea generating events (and incentives) to promote the identification of high ROI use cases for integrating ML solutions into legal processes. | |
| **Long-Term**  **(6-18 months)** | Legal Domain Corpus. Launch release of Legal Domain Corpus and associated research paper(s).  H&K ML guidelines. Finalize H&K ML guidelines and policies.  Public Release of Resources. Release more robust suite of open source ML/NLP modules trained for legal domain.  Commercial Launch. Launch first commercial products for license to corporate law groups and AmLaw 200 firms. | |
| **Resourcing** | **Technology/Technical** - Josias N. Dewey, Jeff Seul, Joshua Strickon, Ziggy Williamson, Marty Durkin, Jake Schneider, Mark Francis.  **Legal Ethics/Diversity** - [Allison Rhodes] [CK – Who was the other fellow you suggested?], Tiffani Lee, Paul Bond.  **Business Strategy** - Sanjay Kamlani.  **Education** - Deb Bernhard.  **IT resources** – IT group will be needed for security, integration and systems support and maintenance.  **Project management** – Project management support will be needed for overall project and several sub-projects. | |
| **Measuring ROI** | Return on investment will be measured broadly to capture benefits from (i) improved efficiency, but only to the extent of improved margins, (ii) our ability to deliver better and more valuable services to our clients, and (iii) the reduction of loss prevention risks through improved quality control.  [***DEVELOP PROJECTIONS***] | |