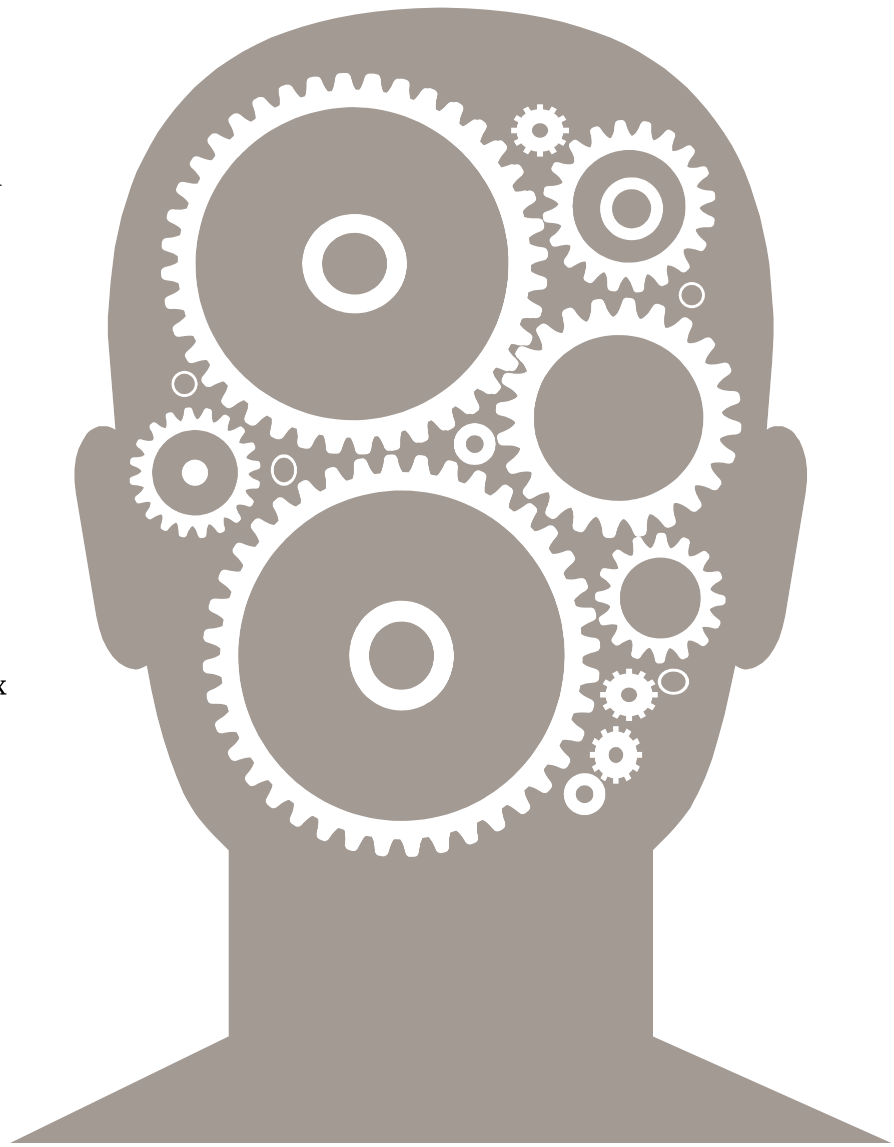
PwC Advanced Tax Analytics & Innovation

Tax Analytics
Artificial Intelligence and Machine Learning–Level 5

By Dr. Cas Milner and Dr. Bjarne Berg

Contents



Introduction

This paper was written for tax professionals and tax data scientists with an interest in machine learning and artificial intelligence.

While the term ‘analytics’ is frequently used by tax practitioners, it is a broad term, used to describe everything from business intelligence, dashboards, predictive and prescriptive tax analytics, to more advanced areas such as machine learning (ML), data mining, and artificial intelligence (AI).

In our Tax Function of the Future (TFoF) series, we have discussed the overall differences in types of analytics done at each level, and in this paper we are exploring “Adaptive Learning” (level 5 shown to the right) in more detail. We also provide some background and examples to better understand the tax and technical capabilities in this emerging field of tax analytics.

So, what is AI? The public perception seems to derive from science fiction movie depictions of malevolent robots or software – in “Terminator,” “2001: A Space Odyssey,” “Ex Machina” and so on. As often is the case, reality differs from the movies, even when based on true stories. With widespread marketplace adoption of AI-enabled applications such as Apple’s Siri, self-driving cars, IoT devices like Rachio, and adaptive Google Search, the contrast is stark: movie fantasy robots have general-purpose function and possess human-like (often evil) characteristics, while real applications are interacting with us unobtrusively, seamlessly, and benevolently to enrich our lives, to simplify repetitive tasks, and to assist us with making complex decisions with fewer errors.

Real AI – the kind that makes life better – has arrived in the tax function, with many usable tax applications beginning to emerge from both academic research and professional service firms. While it may seem unlikely that AI robots (Bots) will ever serve as human-impersonating tax accountants, it is highly likely that AI applications will find a role in the tax function of the future – and even the near future – assisting tax experts in detecting errors, classifying transactions, estimating audit risk, and proposing beneficial tax strategies within increasingly complex global legal frameworks.

***Fig 1:* The Five levels of Tax Analytics**

Tax Analytics generally can be grouped into 5 different levels:

In this review of what is available and what is coming, we discuss recent examples and suggest likely future developments. We emphasize the role of tax resources collaborating with data science experts to build capabilities, and also the importance of improving roles of increasing data quality, detecting missing support documents and details, and reducing human errors in complex tax compliance and analysis/ reporting.

AI and ML applications— Background

In general, AI applications typically are aggregations of several technologies (software, algorithms, big data, cloud computing, and sensory interfaces), that operate in human-like ways to learn, react, perceive, decide, and recommend. While this can be regarded as a working definition of AI, it should be noted that AI is somewhat fluidly defined. So, AI may be seen as part of a computing functional continuum (Figure 2), with a high degree of sophistication.

For example, simple computing – such as the type performed in Excel spreadsheets, arithmetic processing, macros, sorting, data storages, and sometimes the use of very complex formulas – is the mainstay of tax functions. However, much more sophisticated tax analysis can be achieved with ML. This is a category of adaptive algorithms in which the output, or algorithm function, modifies itself as data is processed. By contrast, formulas in Excel do not change when data is processed.

Over the last few years, basic ML has emerged from academia and is now at the leading edge of business analysis. However, a more sophisticated vision of computing is AI, in which systems typically combine several ML technologies, very large datasets, and innovative use of sensors (optical, audio, unstructured data, etc.) to mimic human intelligence.

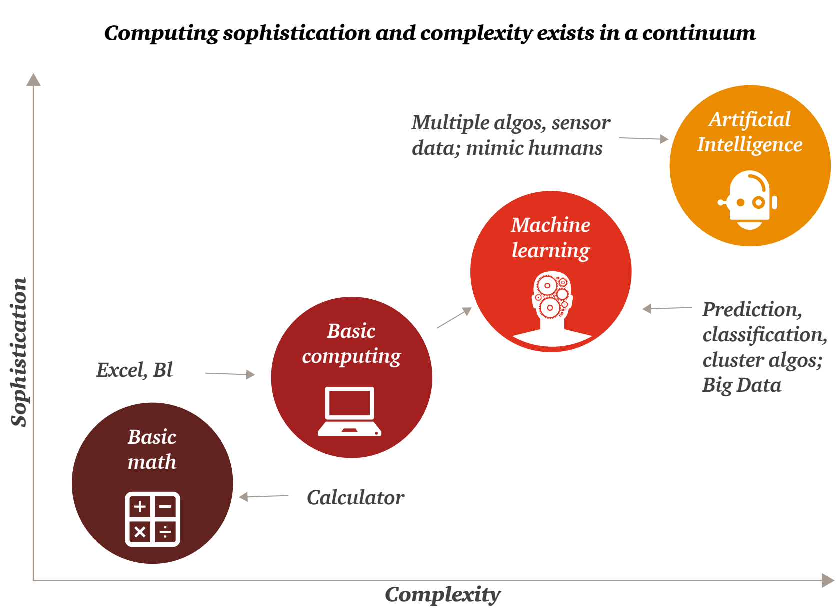
As illustrated in Figure 2, computing sophistication exists in a continuum, with no rigid demarcations between various approaches. At the highest level, AI systems strive to mimic human intelligence and typically combine several distinct ML algorithms, together with innovative sensor information such as visual, audio, or other data. AI systems also interact with and adapt to their environment, which may include other AI systems or humans for input (such as voice recognition systems). We call this the feedback loop (or in human terms, 'learning').

In general, AI is a set of approaches to making software automatically adapt its functions to changing input, without explicit instructions. AI has been developed by university, government, and corporate research laboratories for decades. However, recent breakthroughs allow AI to make a transition to widespread application in everyday life, as well as in heavy data processing areas such as in the tax function.

It is important to note that one distinguishing feature of AI, in comparison to other computer science approaches, is reliance on psychology and neurobiology concepts (actually, many AI system designs are patterned on brain functions).

In addition, the new availability of large-scale parallel computing resources in the cloud has enabled deployment of very process-intensive AI algorithms that previously were impractical due to the time it took to execute them. So, the convergence of faster computers, better networks, and recent research breakthroughs now give tax professionals access to new tools.

Fig 2: **Computing complexity**



What role will AI play in tax?

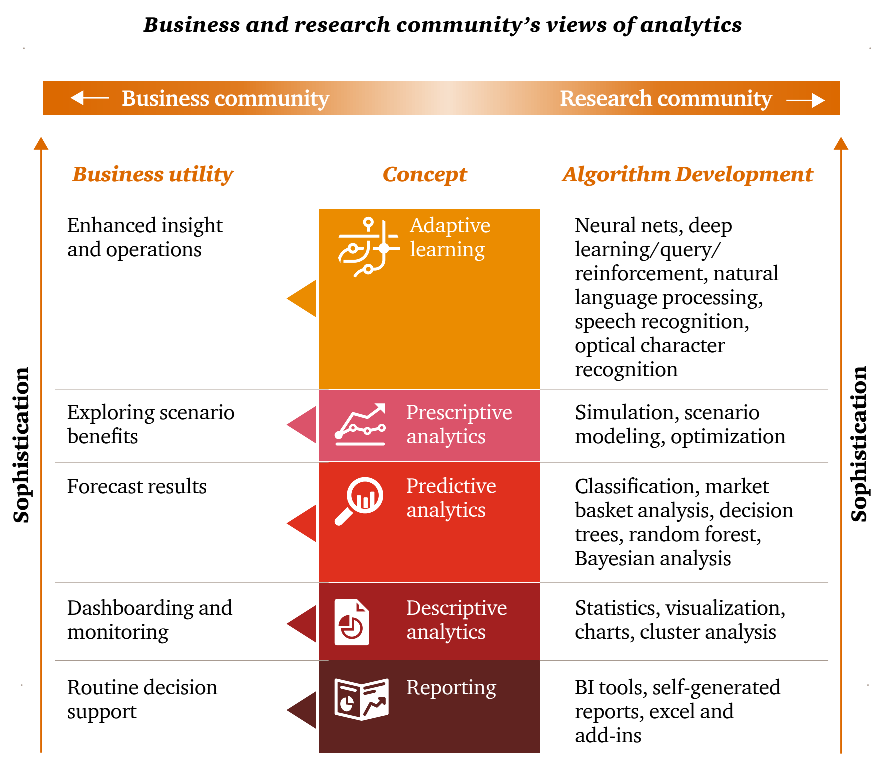
Tax is a complex subject, and AI is highly technical, so discussing the two subjects together requires solid knowledge in both areas. Since people with these skills are hard to find, work in this area normally is done by small, focused teams of experts.

Some mundane, repetitive, and routine tasks are being enabled by AI. These include voice recognition, ATM check amount recognition (optical character recognition – OCR), and text spell checking, supported by underlying technologies that may be readily adapted to other arenas, such as tax systems. Other examples include K-1's, categorizing and processing information that have different formats but leveraging similar forms across tax jurisdictions, classifying data into appropriate buckets such as for categorizing images for fixed asset recognition, or meals and entertainment expenses based on tax rules, vendors, amount, location, and time of day.

Today, a tax professional may develop a standard predictive statistical model, such as a logistic regression, to predict the probability of being audited based upon the internal company data. However, with AI he could analyze all the internal data as well as external, publicly available information like state revenue, number of audits conducted by jurisdiction, legal changes, and political focus, to model tax risks. The latter would be a much more beneficial approach if done properly.

At higher levels of tax functions, tax applications may address more complex, human-judgment tasks like answering subtle legal and taxation questions from legal documents or detecting sophisticated fraud strategies, thereby possibly assisting government oversight.

Fig 3: **Differences in AI perspectives between business and academics**



For the research community, AI and ML are roughly equivalent terms. However, since "learning" is easier to define than "intelligence," most academics prefer the term machine learning. For example, humans change their behavior in response to new information, to become more efficient and have better "judgment" in the future. Similarly, computer algorithms can be made to read new data to adapt and improve their "judgment" and even achieve, in the words of one chess Grand Master who lost a chess tournament to a computer, "*a new kind of intelligence*".[1]

1. Garry Kasparov quoted in Time Magazine: "http://content.time.com/time/specials/packages/article/0,28804,2049187_2049195_2049261,00.html "

The business and research communities view AI differently. These views are summarized in Figure 3 above. The business community views are rooted in history of human decision-making, in contrast to AI-system decision-making. Of course, we do not know as much as we would like about human decision-making, but everyone can make a long list of general deficiencies plaguing humans such as:

- **Unreliability** – Two people using the same data can reach different conclusions.
- **Slowness** – Humans can consider only one decision at a time, often tediously.
- **Inaccuracy** – Many factors can interfere with positive learned behavior by humans such as emotions, illness, and fatigue.

In contrast, computer algorithms are transparent, and their detailed function can be understood. In addition, they have superior speed, consistent results, never fall asleep (well, sometimes they crash), or have "bad days".

Many AI approaches are recognizable because they add the word "deep" to their name, such as Deep Learning, Deep Search, Deep Query, and Deep Reinforcement. This nomenclature indicates a computing approach capable of looking far deeper and wider into data and decision branches than a human can. For example, a chess Grand Master may evaluate five moves into the future, but Deep Blue (a 20-year-old technology) estimated the strength of alternatives for 30 moves into the future.

What topics are good candidates for tax AI development?

How do we recognize a good tax problem that can be solved by AI when we see it? The history of AI suggests we can test a system through the context of a game – a competition with an expert human. A computer "beating" a human gives confidence to business decision-makers that AI is effective. Examples of "gamification" approach include:

- **IBM Deep Blue** – Chess champion.
- **IBM Watson** – General-knowledge answer champion, medical and tax advisor.
- **Google Deep Mind** – Learned to master old Atari games using videos as input.

However, games are a special case of AI, with narrow rules and well-defined goals. A game can be won even if a few points are lost, but in tax, mistakes can lead to business impairments, such as fines, impact to earnings, and missed shareholder expectations, even jail, so the stakes are far higher.

Another approach, which often lends itself better to the tax domain, is goal-oriented AI systems. These often are driven by optimization of a utility function. Note that the term "optimization" does not imply achieving an optimal result, because such a state often is uncertain, ill-defined, and sometimes not legally permitted. Rather, "optimality" is the construct assumed by the system creator in order to establish a necessary goal within legal frameworks and regulations (sometimes the mathematical optimal solution is not allowable under tax law).

Recent trends in tax technology suggest an environment in the near future ripe for AI assistance and deployment, by a variety of organizations including taxation authorities, reporting companies, and tax advisors. The applications of AI in tax include:

- Taxation entities and regulators will implement greatly improved data and analysis platforms; require greater reporting; and have more refined ability to mine and analyze reported data for discrepancies in returns.
- Tax jurisdictions will increasingly share data such as information about transfer pricing, filing inconsistencies, and supporting data feeds for audit purposes.
- Tax data reporting frequency requirements may in the future approach near-real-time, i.e. for areas such as sales, value added tax (VAT) and use-taxes.

An increasing number of tax professionals will be proficient in core data analytic technology through increased focus of education in universities and training programs at leading tax firms and clients. This may manifest itself as:

- AI will become a digital assistant (replacing basic capabilities of first and second-year tax associates).
- AI will begin to make tax determinations of data (e.g. is an expense 50%, 100%, or non-deductible).
- AI will be able to optimize the best outcome and course of action for a company by ingesting a wide variety of supply chain data, SKU level sales data (indirect), tax (direct) data, and external environment data all at the same time to optimize specific outcomes like the effective tax rate (ETR) and tax efficient profitability.

A computer "beating" a human gives confidence to business decision-makers that AI is effective.

Applications of AI in Tax – Right Now

Given that tax is one of the most complicated business processes, how are AI and ML relevant to tax functions? To illustrate the complexity, in the United States, the IRS regulations in printed form are over 75,000 pages long.[2] As a result, many parts of the tax function rely on human expertise developed over long careers to understand the embedded nuance, subtlety, and grey areas in the regulations, administrative rulings, and court cases to make judgments. And because of the complex nature of many of these tasks, AI and ML have been more slowly adopted than in other parts of business. However, computers have advantages even in this area. For example, computers can quickly process this large volume of tax information, and once "consumed" it does not have to consume it again (unlike humans, who forget and have to read the same information over and over).

For all businesses, including the tax function, there are many important issues relevant to the AI adoption trend – and how they will shape businesses in the future.[3] These include:

1. Project design, execution, and investment planning.
2. Worker-resource development and job disruption.
3. Data access.
4. Ethics and privacy.
5. Regulations, law, and government policy.

These issues are discussed below and woven into a discussion of functional prototype tax systems presently under development.

1. Project design execution, and investment planning.

For business problems, the key success factors specific to AI project design are:

- Precise and narrow definition of problem goals:
  - The project should create clear, management-recognized business value early in the first phase – informed by deeply engaged tax subject matter experts, executives, and business leaders. This will energize the execution team and create project momentum.
- Relevant data, including known outcomes:
  - Proprietary business data is valuable, but often difficult and expensive to gather, prepare, and maintain. Access to detailed data for tax professionals represents a competitive advantage. This includes external data that most tax functions have not gathered in the past on regular or recurring basis.
- Statement of goals in terms of a challenge:
  - Devise a competition between the system and an expert human team, and measure each according to speed, accuracy, and efficiency.
- Data science experts on the team who are also experts in the project topic:
  - They may be rare and expensive, but experts with backgrounds in both data science and the subject will create the deepest solutions to the deepest problems.

[2] https://www.irs.gov/tax-professionals/tax-code-regulations-and-official-guidance

[3] https://ai100.stanford.edu/2016-report

Tax authorities do not have many such data-restriction issues. They often use vast databases, and they may build internal models for automating compliance functions. This includes searching for errors, evasions, omissions, and inconsistencies. For this reason, it is possible that in the relationship between tax collectors and some taxpayers, the collectors may soon gain an innovation advantage.

2. Worker-resource development and job disruption

Skilled and experienced data scientists are in short supply for the coming growth in AI and ML. They are needed to drive development, and their scarcity will likely be a limiting factor in initiatives. This is especially true in tax, where deep subject-matter expertise also is required.

Solutions will be available from vendors, or they may be developed in-house by building a proprietary team. Some larger tax firms are developing partnerships with leading university computer science departments to conduct R&D relevant to their special needs. All these approaches are needed to satisfy business needs over the near horizon. In addition, the related mechanisms of job disruption are familiar in the modern age – humans can be replaced by machines, automation, and computing.

AI may also accelerate changes in some areas, especially in knowledge-worker realms that have been historically resistant to competition from machines, such as some tax practitioners and many financial managers.

Few modern jobs entail a single task. Since many AI and ML systems are narrowly focused on single tasks, their use is unlikely to cause a one-to-one human replacement in the near future. Still, across a workforce, AI is likely to cause employment reductions as the number of human-performed tasks decreases. In fact, this is a primary goal of business AI – creating economies of scale, lowering costs, and improving efficiency by replacing humans in jobs that can be performed better by computers.

It is also likely the AI revolution will create new job categories – and not just for robot repair technicians. However, in the short term there may be a mismatch in pace of job destruction and creation that is disruptive to businesses and to individuals in the workforce.

3. Data access

AI and ML, since they are adaptive, typically are enabled by "big data" to "train" the algorithms. For reasons of privacy, the treatment of big data in tax is different and more restricted than for other businesses. Many tax firms have written into their client contracts detailed data privacy and protection measures intended to limit how data may be used in analytics. This may pose some challenges for implementing AI in tax environment.

We can categorize the large datasets relevant to tax as:

a. **Public data** – Generally available data. For example in the United States, data generally available from the Federal Reserve, IRS, SEC, and the Census Bureau may be used.
b. **Data from a single large company** – Can be used if the analysis remains within the company and not in benchmarking studies.
c. **Simulated business data** – The main restriction in this list relates to using specific company data for building (training) general model applications for use outside the company, even if it has been "sanitized" and "anonymized". This is a significant restriction compared to how AI systems are built on large sets of data about individuals. Examples include AI systems such as Netflix recommenders and advertising targeting systems, where data gathered in one venue is used to build a model deployed in a different venue for a different company or individual. Simulated data is the way around these data privacy restrictions, while still allowing developers to leverage the full power of AI systems.

Simulated data is not the same as 'made-up' data. Simulation is widely integrated in engineering and science design and analysis processes, but has been under utilized in business. Simulated data may enable AI model-building and can provide an alternative when tax data access is restricted.

Tax experts, with a deep understanding of company finances, can work together with data scientists to generate a system for simulating detailed tax data for a company or even a large group of companies. This can allow developers to include tax data that is unlikely, but possible, in areas such as error detection and treatment of unusual tax events.

Therefore, simulated data can serve as a proxy for real company data, and can be augmented with errors, inconsistencies, and missing values to provide realism. There are two primary benefits of the simulation approach. First, it creates a platform for business hypothesis testing and management discussion. Second, it enables creating models that then can be deployed in private company settings where "real" data may not be used to build a model, but only as input to system calculations.

After the initial tax AI model is built and tested, it may be directly deployed by substituting real company data in private computational settings.

Massachusetts Institute of Technology (MIT) scientists recently studied a new approach to predicting tax evasion schemes by simulating small changes to known tax avoidance strategies, in the context of proposed and actual tax regulations changes.[4]

The focus was on tax evasion strategies in partnership structures with many layers of ownership and high transaction volumes. In certain cases, these structures theoretically enable strategies for combining legal transactions in new sequences, which then could become evasive and potentially illegal. This MIT study by its Computer Science Department (CSAIL) relied on several technologies, processes, and calculation layers, including:

- Natural Language Processing (NLP) for automating the ingesting, cataloging, and identifying changes in partnership tax regulations;

[4] "Tax Non-Compliance Detection Using Co-Evolution of Tax Evasion Risk and Audit Likelihood", E. Hemberg, et al., in The 15th International Conference on Artificial Intelligence and Law, San Diego, June, 2015.

- Expert knowledge of tax regulations – some team members were CPAs;
- ML decision-tree models of tax evasion schemes, incorporating scheme features, audit score sheets, and tax law rules;
- Simulation process for iterating through transaction sequences and outputting tax liability and audit score; and
- Optimization process examining transaction sequences and audit risk.

The resulting tax AI system was able to detect potential evasive sequencing of events and accurately flagged such entities and filings.

In another recent study, the best sales and use tax candidates for audit flagging were determined in a pilot project for the State of Minnesota.[5] In this study, researchers built a system to mimic the decision process of expert tax auditors. This required them to learn the expert techniques in order to embed them in the analysis. For legal reasons, the authors could not reveal which data elements were most relevant to audit prediction (this is a common conundrum for some AI research in tax – revealing the results could enable tax evasion schemes). Even so, the authors claim a 63% improvement in audit selection efficiency, indicating successful tax audit strategies can be crafted and automated by tax authorities.

4. Privacy

Safeguarding confidential business data is a paramount consideration in the tax function. Since many AI and ML applications will need to access this data, new policies will need to be developed regarding which data can be accessed, how it can be used, and how it may be transmitted. In addition, many non-financial data sources that are in the public domain will be used by AI, in ways we may not recognize today as being relevant to tax. An example could be the use by tax authorities of ubiquitous video-surveillance camera data, to monitor assets and commodity transportation for export taxation, property taxation, and business activity.

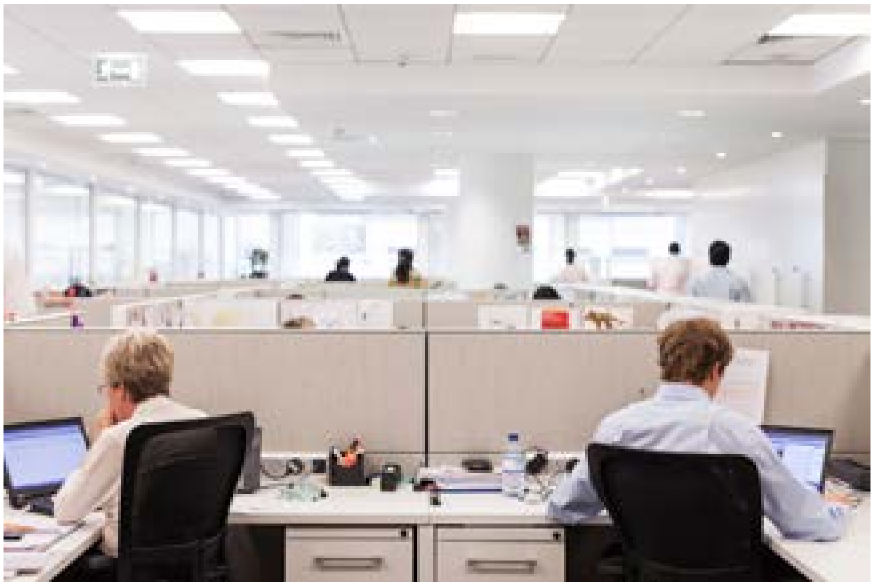
5. Regulation, law, and government policy

Regulatory agencies, lawmakers, and government policy-crafters have not yet addressed the implications of an AI-rich world. Given the pace of AI adoption, they may soon need to accelerate their work. For example, if an AI system gave inappropriate financial advice to an investor, who would be liable? The developer of the system, the operator, or the user? The legal system needs to grapple with this and similar issues, or adequately address liability in contracts and agreements between jurisdictions, providers, and clients.

Even so, the authors claim a 63% improvement in audit selection efficiency, indicating successful tax audit strategies can be crafted and automated by tax authorities.

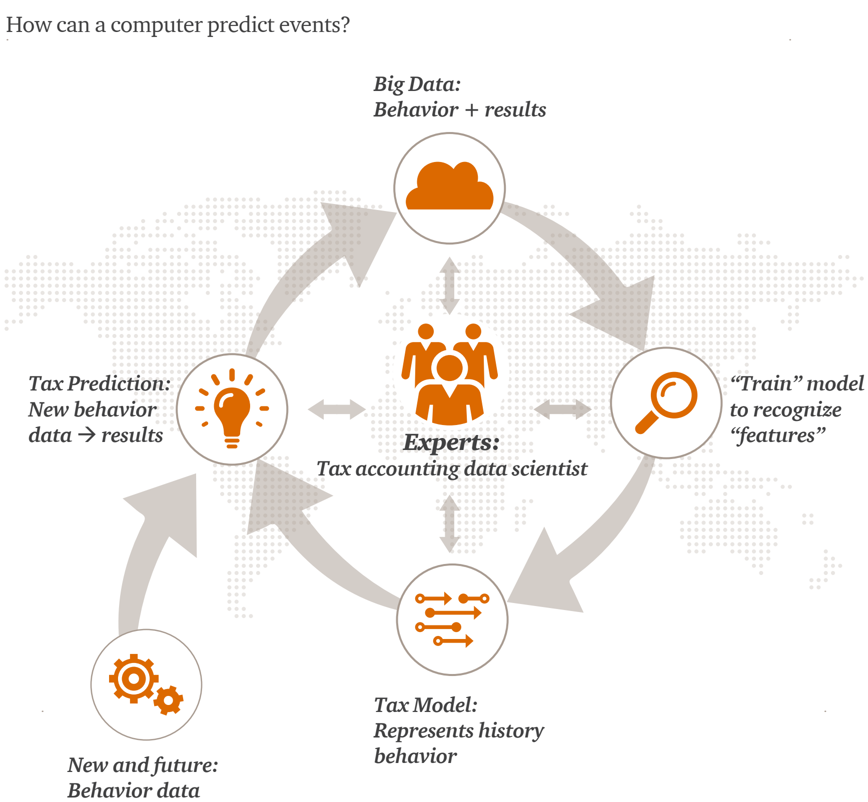
[5] "Data mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue", K.-W. Hsu et al., in Real World Data Mining Applications, Annals of Information Systems 17, 2015.

The Power of Prediction with AI in Tax



One central function of ML is making "predictions." For example, how would a system "predict" the likelihood of a tax audit? There is a general process for building a prediction system, common to most algorithmic approaches (Figure 4):

***Fig 4:* Building a Prediction System**



- Preparing Tax Data.
  - Most tax AI systems require a "big data" set, prepared in a way compatible with specific requirements of the intended algorithm, – for example re-scaling values in a numeric field so they range from zero to one.
- "Training" the AI/ML algorithm on the data to find "features" and results.
  - This is conceptually similar to how a human tax expert might examine a more limited data sample, looking for clues to build ideas on what to look for, and then doing detailed tests. This produces the model.
  - The tax AI/ML model then is a representation of patterns, groups, and behaviors in the data.
- Prediction phase.
  - Analyze new examples or tax records of the same kind of data used in training, except the outcomes or results are not known.
  - Model will associate the new tax data records with behaviors it has seen in the training data.
  - The associated "outcome" in the training data is the "prediction" for the new data.

Historical patterns leading to specific outcomes enable the "prediction" – an outcome the computer has seen in the tax data associated with specific conditions, groupings, and tax events.

Sometimes tax predictions from these systems can seem like intelligence – the basic human ability to anticipate future events, based on experience and present conditions. So in this way we can think of tax predictions as being a kind of intelligence when it is produced by ML algorithms or AI systems.

Tool and Techniques

What Tax Problem Are You Solving, and How Do You Start?

There exist hundreds of ML algorithms, developed by researchers across the world. It probably is not necessary for you to develop your own algorithm; rather, the problem is selecting the right one, consistent with your goals, data, and problem. This selection process is part art and part science, based on tax data scientists' experience, but some general guidelines can be used.

We will now take a look at the major categories of ML, how they are typically used (with examples), and how they could be applied to the tax function.

Forecasting

This is the simplest form of prediction and ML. Basically, the algorithm first computes a trend line, and then can extrapolate, or interpolate, the trend to make a prediction using new data.

Today, many tax professionals do tax provision forecasting using simple techniques such basic regression, rolling averages, exponential smoothing, and "Same As Last Year" (SALY) approaches. However, much better advanced tax forecasting algorithms can be used that can detect trends within tax filing cycles (typically quarterly), annual trends (such as underlying income, inflation, and tax payment growth), and even monthly trends in the data, such as in the areas of indirect taxes, exports/imports, and VAT.

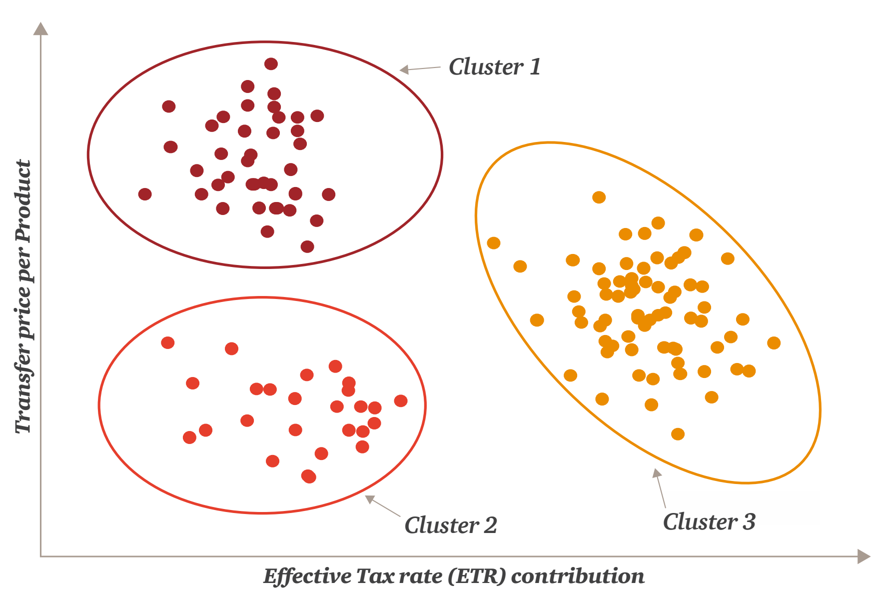
These trends are known by data scientists as alpha, beta, and gamma factors. While they are heavily used in sales, supply chain, manufacturing, and financial modeling, they are seldom used in tax functions. It is important to note that tools and techniques to detect these factors are readily available and can result in significantly better forecasting of tax provisioning, thereby freeing up capital within organizations or avoiding end-of-year tax surprises.

Other specific algorithms include Analysis of Variance (ANOVA), "polynomial regression", "logistic regression", and "multivariate regression". Many of these functions are packaged in spreadsheets. While few tax professionals are using these algorithms, they are applicable in tax in areas such as ETR forecasting and predicting probability of events such as audit/non-audits, and significant deficiency or none. In general, these algorithms address the basic question: "if present trends continue, what will happen?"

Clustering

For humans, recognizing cluster patterns is trivial. It is easy for humans to identify three clusters of points in Figure 5 below. However, computers can perform the same task using precise algorithms to identify these clusters, and do it faster, with higher accuracy, and on vast quantities of transactional and detailed data. In addition, the algorithms can work with a large number of variables, not just two (as shown in the Figure), and to identify groups according to multiple categories.

Fig 5: **Tax Clustering Example**



In general, clustering is a form of "unsupervised" ML, since the only specific goal is identifying groupings of data; a goal or objective is not supplied. It often is the first algorithm used in a new project, since it is exploratory in nature.

Clustering is highly useful for recognizing categories of business behavior, such as

- Companies that tend to have comparable tax characteristics;
- Expense-generating behavior of employees or business units; and
- Grouping of tax implications in transactional data such as R&D credits, sales taxes, capital gains, indirect taxes, foreign tax credits, and export and import taxes, detection of unusual behaviors in expenses and purchasing, examination of legal entity structures, and analysis of transfer pricing impact on tax events.

In general, clustering is a form of "unsupervised" ML, since the only specific goal is identifying groupings of data; a goal or objective is not supplied. It often is the first algorithm used in a new project, since it is exploratory in nature.

However, most clustering algorithms have specific requirements for data preparation. Therefore, an experienced tax data scientist should be used to run them. Insightful behavioral "stories" often emerge when business analysts and tax data scientists interpret the results working together.

Today, there are hundreds of published clustering algorithms, some of general applicability and others that are specialized for specific uses and specific types of data. Since these methods are so well established and almost all analytics software platforms have several clustering algorithms included, it is highly likely that an organization already owns software that can do advanced clustering.

In general, clustering is a fundamentally important technique in tax data analytics. For tax analysts, it often is useful to experiment by applying several different algorithms and interpreting the results to find the best technique before applying it on larger data set. However, the results can be very revealing in finding groups of behaviors and tax events that one would not expect.

Prediction (Classification)

"Prediction" is an important process in ML. Conceptually this is very similar to a human prediction process. First, the algorithm examines a set of data and classifies groups of behaviors leading to various results or outcomes. Think of this as a person who has "seen this before" and predicts what is likely to happen. However, the computerized prediction does better than "gut-feel" predictions of humans, since it can assign the probability of a future result occurring.

For example, humans often expect a successful football team will do well next year if most of the players and coaches return for the next season. While computers also can do that, they are better at looking at much more complex systems. In our example, they can look at all the other teams and explore changes in opponents, stadium and time conditions of planned future games, and injury patterns of each of the opposing players and teams, and assign exact probabilities to how each individual team will perform in the next season. That is the power of analytics over the "gut-feel" of humans.

Although the process can be thought of as prediction, in the ML community it is technically known as classification, because the algorithms operate by associating the historical data (with results) into classes or groups of similar characteristics.

Classification is a type of "supervised learning" because the algorithm is seeking to identify specified results. For example, if you were experiencing financial re-statements due to a significant error, you would be looking for events, or series of events, that caused that outcome.

Another example of ML classification is an email spam filter. It is normally "trained" to recognize spam using a large set of known ("the result") spam emails. It then can be used to recognize the characteristics of spam (unusual sender address, specific words and topics, etc.) and classify the new email as "spam" or "not spam". Spam filters are not perfect, of course. Many of us have missed an important email because it was misfiled by the filter into our spam folder. However, when we find it, and mark it as "not spam", the ML system learns the exception, corrects itself, and improves future operation through this human intervention (i.e., you click the 'not spam' button).

Like the email spam filter, a Tax ML system could be used to classify anything based on large series of tax data and events, and even unstructured data such as tax footnotes, addendums, supporting word documents, emails, and other tax documentation. You have only to envision what you want to classify – that could be items such as R&D credits, tax-deductible meal and entertainment expenses, posting errors, or tax risk. Once the system is running with human feedback, it can remove mundane and repetitive tasks, and perform faster with fewer errors than humans.

While ML systems are not perfect, by most measures they outperform humans. They learn more quickly, do not repeat mistakes once trained, are faster when working, and can work 24/7 without sleep. So the benchmark goal of Tax ML systems is not "perfection" but, rather "better than humans".

As with clustering, there are many kinds of classification algorithms. A popular, readily interpretable type is the decision tree (of which there are also several variants). For example, the probability (based on historical results) of a future tax audit for a high net-worth individual could be assessed using a decision tree (Figure 6). Similar decision trees can also be developed for organizations.

The 'leaves' and 'branches' in figure 6 represent a class along the path to a specific result. In this example (totally imaginary for this paper), the model predicts that high-income taxpayers claiming a higher amount of charitable contributions, concentrated in a few charities, will have a higher likelihood of having personal tax returns audited.

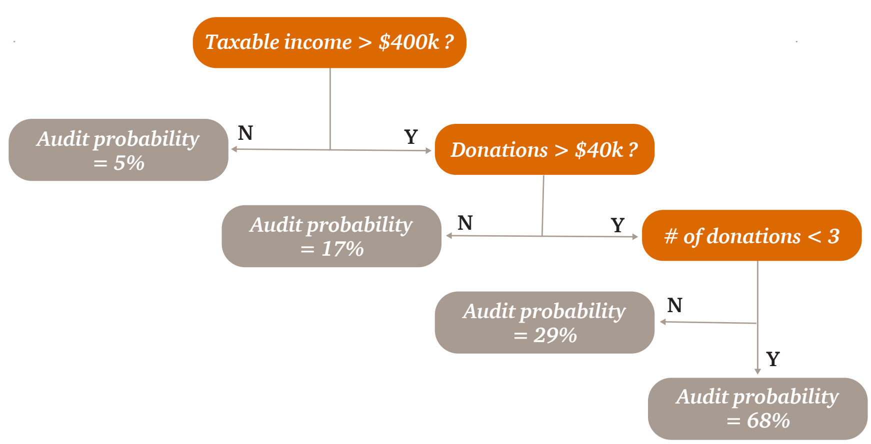
Expert Support-Systems

AI and ML also can be used to assist tax experts by suggesting answers to detailed questions and complex problems. These types of systems are known as "expert-support" systems.

Everyday examples of this are Apple's "Siri", Amazon's "Alexa", and Google's "Home". All are systems that can answer questions. The continued success of these systems may suggest that future technology for answering core tax questions may be on the horizon. There are several approaches to providing expert support, including:

- "Rules-based decision systems", in which decision rules are encoded in the system, and later queried for answers. Some of these systems can "learn" or modify the rules structure, based on new data. A famous example was IBM Blue, the chess-playing computer. Such an approach is of course a natural fit for some tax projects. In a tax setting, rules engines could learn to classify transactions into tax lots.
- The core of a rules system is the knowledge-base, which often is curated with information from human tax experts, with some assistance from automated ingestion of information.

Fig 6: **Decision Tree Example**



- Conditions (orange) with historical audit probabilities (gray)
- Future audit possibilities can be forecast according to conditions

Natural Language Processing (NLP) may be used to interpret complex tax laws and provide guided expert systems for the tax professionals

- The other key component is the reasoning system – how does the system make decisions using the rules?
  - First, the system must “match” the question to its tax knowledge – in a search through the data.
  - Next, it must resolve conflicts and ambiguities because typically several search matches will be found.
  - Finally, it needs to deliver the answer to the expert, and also update the knowledge base with refinements from the search, or correction/feedback from the tax expert.

Other AI expert-support systems do not begin from rules, but rather by compiling a large, searchable knowledge base of tax information from a wide array of text and image sources.

An example of this is IBM Watson, which can be trained to answer specific questions in a specific field, such as medical diagnostics, finance, or general trivia. This area of AI, known as “Deep Query”, is undergoing active development. Even though tax base information consists of over 75,000 pages of regulations and laws, in addition to rulings from courts and the IRS, it is near to nothing when compared to the vast information on the Internet that some AI applications are attempting to address. So, one can hope tax queries will not be more difficult than medical diagnostics.

Optical Character Recognition (OCR)

Can a computer learn to recognize text and numbers on a printed page? The answer, of course is “yes”, and systems can do so with a very high degree of accuracy. For example, bank ATM machines are now capable of scanning checks, recognizing currency amounts, and using the information to complete a deposit transaction – all without the customer entering any information. PwC also uses a system known as ‘Dexter’ to optically read tens of thousands of K-1 filings without human data entry or interventions.

Data entry tasks that used to take hundreds of hours, were error prone, and needed reconciliations now can be done in a few hours with much better data quality.

In general, in OCR two technologies are at work:

a) Recognizing the “features” present in the pixels of the scanned image, such as numbers, lines, decorative patterns, letters, and characters.

b) Matching a wide variety of detected numeric images to a library of known numeric images (such as previously recorded checks). If a company has a “mountain” of paper documents that should be digitized, OCR offers hope.

In this area, the first task is based on image pattern recognition, or “computer vision”, and has been the subject of decades of R&D.[6] Note that the checks submitted to an ATM may have smudges, are handwritten in various styles, and have a variety of formats – all of which the software must recognize in order to reduce the image to a digital summary identifying the currency amount of the check.

In the second part of processing, the characters are matched to a large library in which historical character features have been recorded along with known quantities. Banks, of course, have access to large datasets of old check images together with known check amounts, and this data forms the basis for “training” the system, using AI techniques. This is an important point – having access to a large and known dataset is critical to successfully training the system.

In tax, we can use these systems to digitize tax forms, handwritten notes, images, books, contracts, and any paper-based material. These documents become instantly available for complex searches and use in other AI/ML applications.

[6] “W. Freeman, et al., “Guest Editorial: Machine Learning for Computer Vision” (2008), International Journal of Computer Vision, 77 (1).

Natural Language Processing (NLP)

Natural language processing (NLP), which is at the confluence of AI and linguistics, has the goal of automating identification of "meaning" in text – in other words, not only what the text says, but what does it mean and what does the selection of certain words, over others, imply. For example, if you heard the sentence "you stink wonderfully", there are many ways to interpret the semantic meaning of the sentence, but something is not right in this awkward word selection; so what does it mean?

In an era of growing complexity in understanding regulations and legal rulings, NLP is becoming increasingly important as a user interface. Also, in simpler contexts, NLP is starting to become the foundation for estimating "sentiment" embedded in text. The challenge is to understand what does a sentence "mean"? This is a deep question, and we can start by identifying the related question: what is a sentence "about"? Tax documents often have complexity, such as having:

a) Complex sentences (as measured by reading-level scores),
b) Exceptions and exceptions to the exceptions,
c) Lengthy lists of conditions and qualifiers, and
d) References to other parts of the document, or even to other documents, with supporting data, or information that can seem contradictory.

A knowledgeable human reader will "learn" the document comprehensively by processing the whole text and connecting "meaning" in cross-referenced sections and external references to modify and refine understanding.

A tax NLP system must mimic human learning in this regard, as it refines identified meaning through merged meaning of internal and external indexed references and tax data. So the depth of meaning will grow as the system integrates more tax documents and they are connected through direct references, or connected through the same type of subject. The basic steps in NLP are:

1. Identify the "meaning" of sentences or paragraphs
2. Deploy an algorithm for recognizing similar meanings in references and related subjects
3. Deploy an algorithm for combining or structuring the hierarchy of related meanings.

Assuming the core text of interest is available in digital form, the first technical step is pre-processing or "parsing" the sentences. This involves "tagging" the parts of speech (nouns, noun phrases, verbs, etc.) and also labeling the grammatical relationships or dependencies within each sentence.

There are a variety of open-source tools available for this task, developed by leading research organizations at Google,[7] Stanford,[8] and others – so mature software is available to start tax projects. However, initial testing with generic NLP software has indicated that an "un-natural" language processor may be needed to read the complex sentence structures of the tax code.[9] That is not to say that it cannot be done, just that the off-the-shelf NLP software would need some enhancements.

It should also be noted the parser should be "trained" on each new "body" (or "corpus" as it is termed by researchers) of text it encounters, and this phase is done by experts – a parser trained on "customer reviews" will not be accurate in processing tax rules. Each new body of information, such as tax credits, export taxes, payroll taxes and indirect tax, has to the trained before deployed.

7 https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

8 http://nlp.stanford.edu/software/dependencies_manual.pdf

9 "US Tax Code & Regulations Language Analysis" Reed Coke, Nils Senvalds, Cas Milner, 2016, PwC US-Tax & University of Michigan

For example, how to extract meaning from the sentence ""*Bell, based in Los Angeles, makes and distributes electronic, computer and building products*" is shown below (Figure 7). What is the meaning? Is it Bell, or Los Angeles, that is the actor? What is the core action, "makes" or "distributes"? What is the relative importance of the other words often separated by commas?

In this example, the sentence above was "tagged" and "parsed" automatically by the NLP software – and it evidently is selecting "makes" as the head of the tree.

After parsing, there are several directions for the analysis to take. For example, there are well-known methods for estimating the "readability" or complexity of a passage of text. These were originally developed to categorize reading level to assess the comprehensibility of training manuals, insurance policies, and legal documents.

One of the first reading scores was "Flesch-Kinkaid," which today is included in many word processing software packages. Originally it was used by the Army to measure the effectiveness of training manuals.[10] In general, this type of text readability score can give guidance to how easy, or difficult, tax documents are to understand. This can then be used to classify tax regulations, laws, footnotes, and IRS rulings to determine who is best suited to read and understand the language (non-college graduates, college graduates, CPAs, tax professionals, or tax experts in a specific domain).

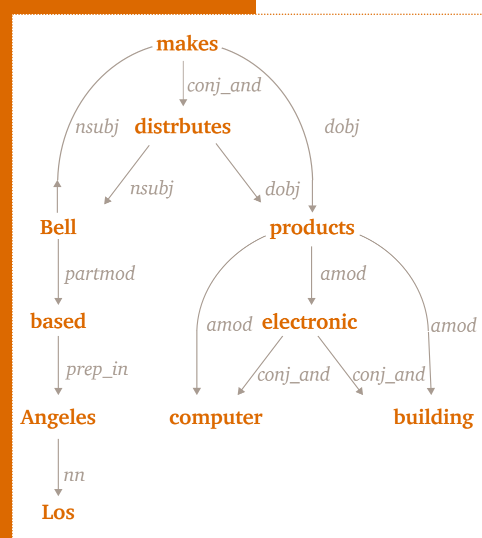
Detailed document searches could be enabled by parsing results, in which similar phrases and meanings may be associated. For example, a refined and enhanced search could be performed for paragraphs in the US tax regulations, showing where discussions occur about "partnership allocations" in related, or similar documents. It could also identify sentence and paragraph "meaning" and clearly enable answering questions posed by tax knowledge-workers, such as "how have legal opinions evolved for the treatment of Effectively Connected Income (ECI)?" Tax applications of this nature already have entered the marketplace, and are likely to improve over the next few years.

Fig 7: **NLP Parsing Example**

"Bell, based in Los Angeles, makes and distributes electronic, computer and building products."

For this sentence, the Stanford Dependencies (SD) representation is:

nsubj(makes-8, Bell-1)
nsubj(distributes-10, Bell-1)
vmod(Bell-1, based-3)
nn(Angeles-6, Los-5)
prep_in(based-3, Angeles-6)
root(ROOT-0, makes-8)
conj_and(makes-8, distriutes-10)
amod(products-16, electronic-11)
conj_and(electonics-11, computer-13)
amod)products-16,computer-13)
conj_and(electronic-11, building-15)
amod(products-16, building-15)
dobj(makes-8, products-16)
dobj(distributes-10, products-16)



[10] "Data mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue", K.-W. Hsu et al., in Real World Data Mining Applications, Annals of Information Systems 17, 2015.

AI is part of the Tax Function of the Future

In summary, AI and ML prototype systems already are being deployed in tax – by academic researchers, taxation authorities, taxable entities, and tax service providers such as PwC. Our prediction? The next few years will see greater deployment of advanced tax analytics – and benefits – as the deployment trends accelerate. Disruption of business plans is possible, and so it is important for all participants to understand what might lie ahead as AI and ML have an impact on the Tax Function of the Future.

To learn more, follow PwC's Advanced Tax Analytics and Innovation practice at:
http://www.pwc.com/us/en/tax-services/tax-innovation.html

Bjarne Berg
Advanced Tax Analytics and Innovation Co-leader
+1 704 344 7720

Rick Reekie
Advanced Tax Analytics and Innovation Co-leader
+1 214 999 2580

Michael Shehab
Tax Technology and Process Leader
+1 313 394 6183

Gulhan Sumer
Advanced Tax Analytics and Innovation Technology Lead
+1 214 979 8009