OREGON DEPARTMENT OF REVENUE

Enhancing the Use of Data and Data Analytics at the Oregon Department of Revenue

Research Paper
July 2021

RESEARCH SECTION

This report was prepared by the Research Section of the Oregon Department of Revenue to contribute to its deliberations on how internal data and data analytics might be leveraged to improve the Department's efficiency and effectiveness.

This research paper and other publications can be found by visiting the following webpage: https://www.oregon.gov/dor/programs/gov-research/

Research Section research papers do not represent the official views of the Oregon Department of Revenue. Comments are welcome and may be sent to the Research Section using the following email address: dor.research@dor.oregon.gov.

Suggested citation:
Oregon Department of Revenue
Research Section
Enhancing the Use of Data and Data Analytics at the Oregon Department of Revenue
July 2021

Enhancing the Use of Data and Data Analytics at the Oregon Department of Revenue

July 2021

Table of Contents

Executive summary

Data and analytics are part of the process of generating insights with the intention of translating them into action. Data analytics might be focused internally—e.g. to increase the Department's efficiency or effectiveness—or externally—e.g. to enhance transparency. It might be used to assist decision-makers describe conditions or phenomena of interest, identify contributing factors or causes of a given issue, predict or otherwise consider future scenarios, and/or choose to automate certain decision-making processes. Analytical tools vary widely and can range from simple descriptive statistics, to data visualization, to experimental design and hypothesis testing, to artificial intelligence.

In a tax administration context, data and data analytics have enhanced several government functions including reporting (e.g. the California Department of Tax and Fee Administration shares its current and historical aggregated tax data with the public using an interactive data visualization tool online), tax return processing (e.g. the IRS automatically identifies discrepancies in returns before issuing invalid refunds), and collecting taxes (e.g. the Australian and Norwegian governments vary payment schedules based on an analysis of taxpayers' propensity and ability to pay). At the Department of Revenue, there are many ways in which data and data analytics might be further used, including assessing and improving data quality, integrating certain databases to streamline reporting, and enhancing the efficiency and effectiveness of audit and collections activities. (See the section "Further applications at the Oregon Department of Revenue" for a longer list of initial ideas from various staff in the Department.) A more rigorous, strategic assessment in this respect, however, would require a more systematic and inclusive review of each division or section's workflows and processes to determine what enhancements using data and data analytics can provide the greatest value.

A few key lessons may be gleaned from governments' increasing experience with data analytics. First and foremost, data and analytics are not a panacea, and it is critical that any work in this area be purpose- and question-driven rather than solely data-driven. Furthermore, data needs to be effectively and consciously managed to ensure it is of sufficient quality and that its limitations are well understood. Also, for analytical tools or models to be successfully used, they need to be developed by skilled individuals in close collaboration with practitioners and subject matter experts. Once built, such tools or models need to be maintained, which requires dedicated staffing.

Finally, cautionary tales of analytical tools, primarily automated systems, having tragic consequences for thousands if not millions of people point to the importance of continually monitoring the effectiveness of such tools, understanding their inner workings and assumptions, as well as developing them and monitoring their performance in light of applicable laws and the broader human system, including the set of financial incentives affecting people's behavior.

Purpose, scope, and research methods

This report was prepared to inform the Director's Office at the Department of Revenue on how data and data analytics might be leveraged to improve the department's work. Given the breadth of this topic, a brief treatment of the many relevant issues is provided in the hope of stimulating and framing thought in this area. Readers interested in learning more about a specific issue may contact the Research Section.

The content of the report is based on a literature review—which included a search of several tax-administration-related websites as well as certain academic databases for any presentations, web-based articles, scholarly papers, and/or case studies on the topic of data analytics, particularly in government, in the last five years—as well as conversations or emails with staff in the Research Section as well as selected individuals across the Department who have an interest in the use of data.

Definition, purpose, and types of data analytics

The term "data analytics" means different things depending on the context in which it is used. For the purposes of this report, data analytics is understood as any means of using data to inform people's perspectives or understanding so that they might take more effective action.

Data itself can refer to many ways of codifying different kinds of phenomena, although for the purposes of this report it is presumed that data is stored digitally. In some cases, data is highly structured, such as in the case of a table of tax return line-item data, or minimally structured, such as in the case of scans of paper-based tax returns. Data can have various levels of granularity and identification. For instance, tax return data is taxpayer-identifiable while property tax data available to the Department of Revenue typically is not. Data can also be generated through active triggers—such as tax return data, which is the result of a person actively filing their taxes—or passive triggers—such as telephone call records, which are automatically recorded by most telephone systems.

Furthermore, computer programs are increasingly used to automate data analysis and even certain classes of decisions. In these cases, the programs themselves may generate additional data about their internal procedures as well as outputs (e.g. the number and classes of daily errors from scans of paper returns). Such data are critical to ensure that these programs are functioning properly and in line with the goals of the organization.

From one perspective, data and data analytics may be used

A. for **public or external purposes**, such as enabling access to revenue statistics and information,

B. or for more **internal purposes and to increase an organization's efficiency**—such as minimizing administrative costs to collect revenue—**or to enhance its effectiveness**—such as collecting the correct amount of revenue across the state.

From another perspective, data and analytics may be leveraged to help people do one or more of the following:

1. **Describe conditions or phenomena of interest**, such as the average effective tax rate by income level, the number of hours worked by unit in the Department over time, or groups of similar taxpayers.

2. **Identify contributing factors or causes of a given issue**, such as selecting tax collection methods that are most impactful or which formats or wordings of tax collection letters are most effective.

3. **Predict or otherwise consider future scenarios**, such as simulating the impact of a tax law change or estimating the future revenue impact of various audit practices.

4. **Choose to automate** certain decision-making processes, such as assessing internal consistency of tax returns using basic arithmetic rules.

The level of sophistication of data analytics can vary widely from straightforward descriptive statistics (e.g. to add up the number of telephone calls made), to various forms of data visualization (e.g. to show average tax rates across counties), to cluster analysis (e.g., to identify groupings of similar taxpayers), to hypothesis testing (e.g. to identify most effective letter wordings), all the way to machine learning (e.g., to identify the best mathematical-statistical model to predict underpayments) and more complex forms of so-called "artificial intelligence" (e.g. to identify individuals to audit based on an evolving model that "learns" over time using new data).

Applications by tax authorities

In 2019 it was estimated that at least 44 states had initiatives to use data for decision making (White, 2018). For instance, in California, an integrated data system was being developed to correlate data from various programs spanning early childhood, higher education, and employment. In Georgia, a Data Analytic Center was created to facilitate more immediate, cross-agency reporting. Not surprisingly, there is a growing number of tax administration entities within the United States and around the world that are leveraging the power of data and data analytics.

According to a few different analyses (Pijnenburg, et al. 2017; Milner & Berg 2017, Ordóñez & Hallo, 2019) there are many classes of analytical tools or techniques that have applicability in a tax administration content. The following table organizes them according to the four purposes described in the previous section.

Describing conditions or phenomena of interest	*Identifying contributing factors or causes of a given issue*	*Predict or otherwise consider future scenarios*	*Choose to automate certain decisions*
▪ Descriptive statistics ▪ Data visualization ▪ Cluster analysis ▪ Classification analysis ▪ Anomaly detection ▪ Network analysis ▪ Optical character recognition ▪ Natural language processing	▪ Hypothesis testing ▪ Experimental design	▪ Forecasting ▪ Prediction ▪ Recommendation systems ▪ Expert support systems	(Combination of tools listed in prior columns as well as simple rules and/or complex algorithms)

Table 1. Classes of data analytical tools for tax administration classified by primary purpose

At the most basic level, descriptive statistics and data visualization can be used to monitor progress within a tax administration according to key performance measures or other relevant indicators. There are numerous examples of organizations developing dashboards and other means to monitor and enhance their performance. In the Washington Department of Revenue, for example, customer survey data was used to evaluate certain tax-related application processes, ultimately leading to measurable improvements in performance (Results Washington, 2019A and B).

Basic summary statistics can also be shared with the public in numeric and graphical forms using web-based systems. The Oregon Open Data Platform (http://data.oregon.gov) provides a readily available avenue for the Department of Revenue to do so. Other state tax administrations, such as in California (https://www.cdtfa.ca.gov/dataportal/visual.htm, see *Box 2* further below) and Ohio (http://checkbook.ohio.gov) have shared selected current and historical tax revenue and expenditure data with the public online.

With respect to performing the core set of tax administration functions, auditing activities, in particular, can be enhanced using data analytics (Ordóñez & Hallo, 2019). The IRS, for instance, has matched "taxpayer filings with third-party information" in 2018 to identify discrepancies, ultimately yielding an additional $5.3 billion dollars in tax revenue (Bernard, 2021; Sarin & Summers, 2019).

Advanced analytics have been used in areas beyond audit case selection, however, to also include techniques to "optimise debt-management processes, secure filing and payment compliance, improve taxpayer service, and understand the impact of policy changes" (OECD, 2016). The Australian and Norwegian governments, for instance, vary their payment arrangements based on an assessment of taxpayers' propensity and ability to pay (ibid.). The Canadian government also uses predictive analytics in ensuring compliance by non-filers (ibid., see *Box 1* for further details on this case). The IRS has developed data-analytical capacities to automatically identify "discrepancies with information returns to prevent the issuance of invalid refunds" (Bernard, 2021).

Some authors have pointed to the benefits of data analytics to craft optimally-worded letters to enhance tax compliance (Keightley, 2019); identify taxpayers who will most likely fail to comply with the tax code and then communicate with them in advance to prevent them from doing so (ibid.); or otherwise find ways to maximize impacts by adjusting methods of communication with different groups of taxpayers (Centre for Public Impact, 2018). The Revenue Agency in Canada has used controlled experiments to measure the effect of automated work processes and different taxpayer communications (OECD, 2016). Other modes of communication can also be analyzed. In Singapore, the Inland Revenue Authority used text-mining techniques to analyze emails from taxpayers more consistently and widely to better understand the nature and trends of their concerns (ibid.).

Data analytics might even be used to streamline data cleaning processes, thereby potentially saving money on expenditures for seasonal workers to process physical forms (ibid.). Finally, some believe that artificial intelligence might even be able to assist with the design of tax policy (Griffith, et al., 2020).

Box 1. Canada's use of data mining models for non-filer programmes

(All of the following text is drawn from OECD, 2016.)

"The Canadian tax system is based on voluntary compliance. In Canada more than 25 million individuals pay and file their tax returns without intervention. The Canada Revenue Agency (CRA) manages its programmes using a risk based approach, to direct resources to cases with the highest risk of failing to file on time.

"The CRA has developed and continues to refine several predictive models to assist in the delivery of its non-filer programmes. The models support improved workload selection and prioritisation for the programmes, and also supply estimates for cases that have not filed returns. In its first year in production, one non-filer model resulted in a total of CAD127.6 million in additional positive assessments. The CRA is now moving away from a pure predicted value to a relative ranking indicator, dynamically scoring accounts on an ongoing basis. The CRA has also developed several other models to improve programme effectiveness and enhance taxpayer services by predicting self-resolution and responsiveness to a specific compliance action.

"In addition to predictive techniques, CRA applies prescriptive analytics to support improved strategic and operational programme delivery. Prescriptive analytics is used to enrich the CRA's understanding of the non-filer population, optimise operational processes, and direct the application of compliance activities, allowing for more fact-based decisions. Complementing the use of predictive models, the non-filer programme is expanding its use of behavioural economics through nudge experiments to influence taxpayer compliance behaviours."

Using a compliance risk management framework, Pijnenburg, et al. (2017) collated a list of analytical techniques that might be used to enhance tax supervisory functions:

Tax function	*Analytical techniques*
Risk Identification	Horizon scans Random audits Identify new risks from data Segment the population of taxpayers Detecting fraud
Risk Analysis	Quantify risks with help of in-house or external data Hit rate scoring Random audits Tax gap estimations Trend analysis Root-cause analysis Estimating costs of treatment
Prioritization	Calculating human and other resources Optimizing resource allocation
Treatment	Easy contacts Desk audits Field audits Real-time checking of tax returns Pre-filled tax returns Administrating in the cloud
Evaluation	Evaluation analysis Experimental design of evaluation

Table 2. Analytical techniques that might be used to enhance tax supervisory functions as outlined by Pijnenburg, et al., 2017.

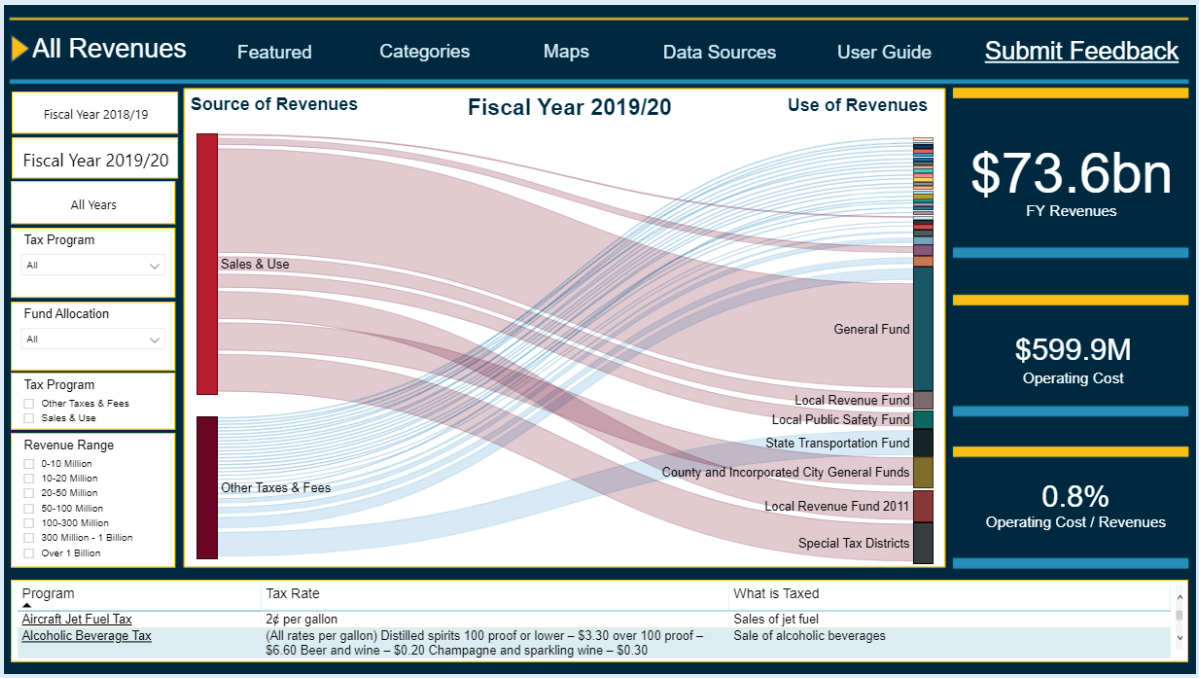
Organized in a different way, data analytics have the potential to enhance the following five categories of operational activities within a tax administration entity: audit case selection, filing & payment compliance, taxpayer service, debt management, and policy evaluation. A study published by the OECD (2016) distilled the relevant experience of over a dozen such entities around the world according to each of these categories. Highlights from that study are included in the following table.

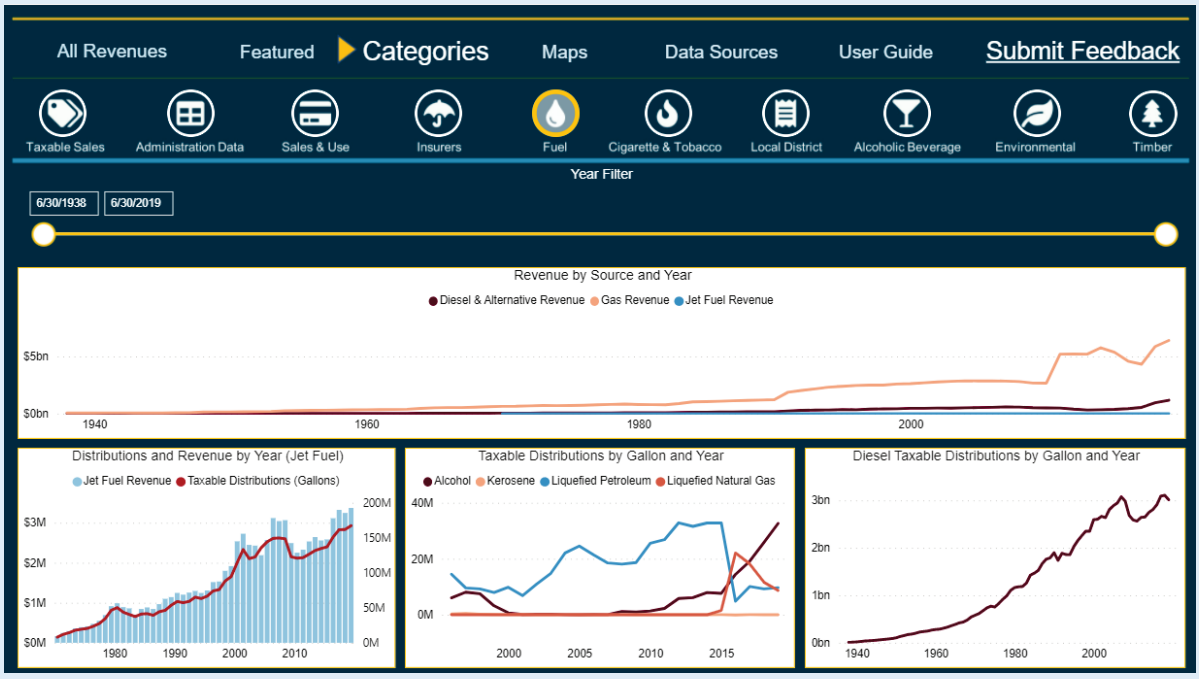
Operational activities	*Experience and insights from seventeen tax administrations*
Audit case selection	▪ Predictive models should be developed and ultimately relied on only after considering their benefit relative to the next best alternative. ▪ Social network analysis shows promise as a means of predicting tax fraud or error for certain risky groups of individuals. ▪ Ideally, tax administrations rely on multiple risk prediction models since each is limited by how it examines data and what kinds of conclusions it can reach. ▪ Whereas "supervised models" are ones that "seek to learn from historical data where the outcome of interest (e.g. whether or not a case was non-compliant) is known", unsupervised models "seek to identify interesting or anomalous patterns in the data, rather than trying to learn from the outcomes of specific cases". The first can "reduce the number of cases wrongly flagged for intervention", "save caseworkers' time and lessen the burden on compliant taxpayers" whereas the latter can "help to identify new or previously unknown types of risk."
Filing & payment compliance	▪ Experimental designs have assisted in designing more effective taxpayer communications. ▪ Predictive techniques—such as ones that aim "to identify which cases are likely to fail to meet payment or filing obligations, and which interventions are likely to remedy the problem"—have been used to inform experimental designs which then test those interventions.
Taxpayer Service	▪ Much of the effort in this area is focused on enhancing communication with taxpayers, including which modes of communication are employed, how they should be designed, and what their content should include.
Debt Management	▪ Similar to "filing & payment compliance", many governments leverage predictive techniques and experimental designs in this area. ▪ A technique that shows promise (and is already widely used in the private sector) is "uplift modelling" through which experiments are conducted to not only estimate the impact of a given intervention but to build a stronger predictive model that can be used for subsequent action.
Policy evaluation	▪ Governments often used data analytics in this area to "conduct tax gap measurements… [and to assess or forecast] the impact of changes in tax policy".

Table 3. Highlights from a study published by the OECD (2016) that distills the experience of over a dozen tax administration entities around the world to leverage data analytics to improve operational activities.

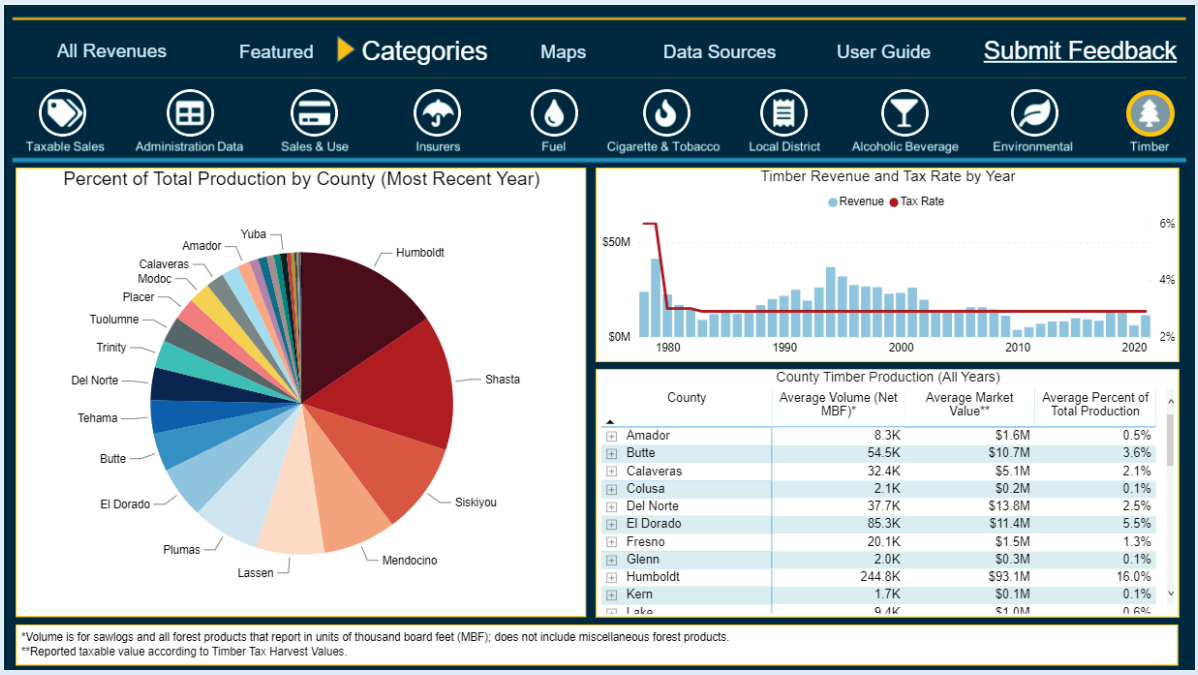
Box 2. California Department of Tax and Fee Administration
Interactive Data Visualizations (Beta)

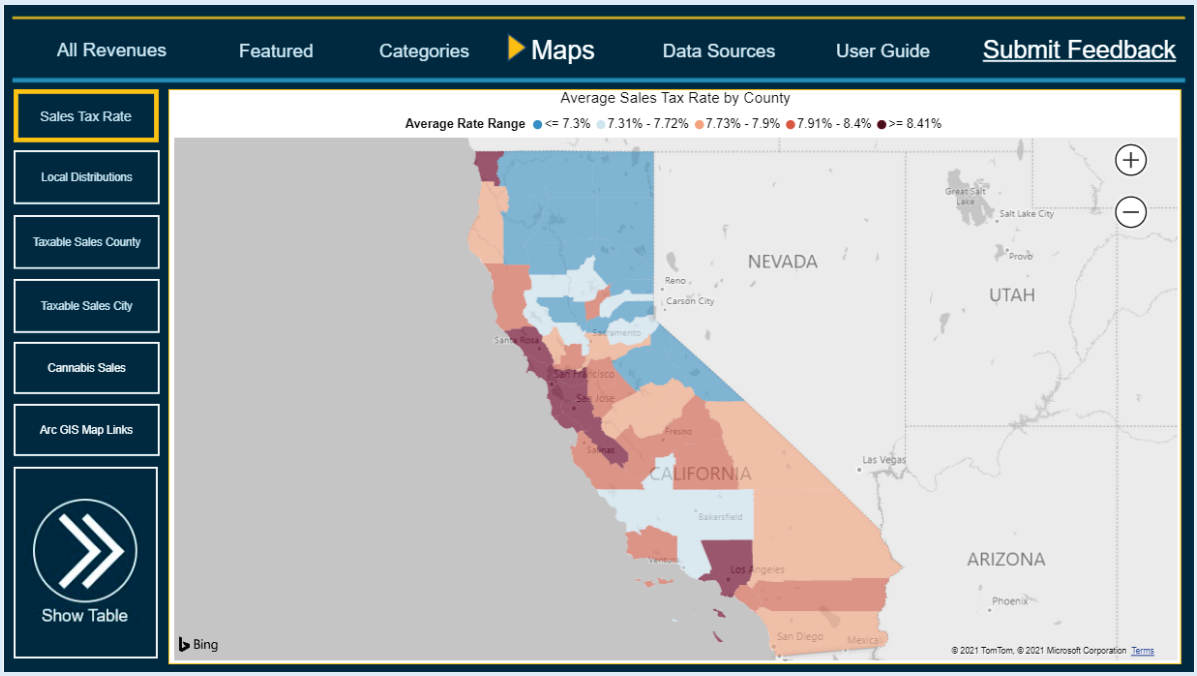
In an effort to increase government transparency, the California Department of Tax and Fee Administration created an interactive data dashboard (using Microsoft PowerBI, available at https://www.cdtfa.ca.gov/dataportal/visual.htm) for citizens to be able to understand the sources and flows of revenue to and within the State government. The following screenshots show what kinds of data visualizations are possible using this web-based tool.





Box 2 (Continued). California Department of Tax and Fee Administration Interactive Data Visualizations (Beta)





Further applications at the Oregon Department of Revenue

Data and data analytics are, of course, already in use across the Department of Revenue. Several sections, for instance, use GenTax as a system to process tax returns and payments as well as a means of storing data from external sources such as the IRS, the Oregon Department of Transportation's Driver and Motor Vehicles division, and the Oregon Employment Department. The data contained in this software package is regularly accessed by various users in the Department using built-in reports but also more customizable queries using the "My Search" function. When users require additional functionalities, they make requests for the development of new reports or other changes to GenTax using an "SQR", the internal ticketing platform to request such changes.

In recent years, an arrangement was made whereby several data tables in GenTax were exported to a separate server so that the Research Section could conduct analysis using in-house custom-built Structured Query Language (SQL) and SAS code. Access to this server, which is also known as the Tax Analysis Reference Database Intermediary Server (TARDIS), has been provided to a few other individuals in the Department who regularly meet with Research Section staff to discuss how to use SQL to access this server and generate their own reports. The same Section has also devised a means to convert e-file data (i.e. electronic tax returns) into a structured format, allowing for more detailed analyses of personal income, partnership, and corporate tax return data.

Other sections in the Department have access to other sources of data. The Processing Center, for example, uses the software INFOPoll to track mail and *Quick* Modules to process paper returns and checks. The Property Tax Division uses data from external sources, such as the assessors' Summaries of Assessment and Levy reports, County Assessment Function Funding Assistance Account (CAFFA) applications and revenues, county tax collector Data Exchange files and annual collections reports, sales ratio study reports, Boards of Property Tax Appeals summary data, ORMAP mapping and fee revenue information, Oregon Department of Forestry timber harvest permits, etc. That Division is also in the process of upgrading and streamlining its current appraisal processes and systems—which include an array of programs, databases and spreadsheets—with the Electronic Valuation Information System (ELVIS). It is anticipated that this will be developed and fully implemented by 2024. There are also other data sources and analytical tools used across the Department.

A few years ago, Deloitte conducted an analysis on how to strengthen "outcome-based management" within the Department. The consulting firm's recommendation was that "DOR establish a framework for organizational metrics, develop a strategic plan, establish an outcome management team for governance of these processes, further engage in efforts to manage its data in order to lay the groundwork for the organizational metrics work, and to better align business processes with GenTax and to provide additional training to staff on the software" (Department of Revenue, 2019).

Toward the end of 2020, a Department of Revenue project was carried out to investigate how data and data analytics might be further leveraged to enhance the work of the Department. As a result, a series of conversations took place with section managers and analysts throughout the agency, many of whom spoke to the importance of better defining quantitative metrics of success so that work quality might be better assessed. As part of the same research effort, conversations with tax administrators in Rhode Island and Michigan who had experience enhancing their States' analytical capacity noted that a) enhancing analytical capacity is a large task and cannot be "half-done", and that b) data analysts not only need support and infrastructure to do their work but should also be

allowed to focus on their work and not be pulled away to perform other duties. Furthermore, administrators from both states emphasized the importance of internally-developed documentation for their systems and close communication between data analysts and the users of reports.

As part of the Research Section's effort to develop this paper, the opinions were sought of a few staff throughout the Department who might have an interest in the use of data and data analytics. The following are the ideas that arose in those exchanges grouped by theme and listed in no particular order.

General comments

- Continue to enhance the capacity of existing internal staff to perform data analytics using SQL (to access GenTax data on TARDIS), Microsoft Access, and Excel. By enhancing existing capacities (or bringing in new staff), then program areas can better assess how severe issues really are before escalating them to the Information Technology section, possibly reducing the backlog of change requests.
- Increase data quality for those issues that have high taxpayer impact or are costly for the state. For instance, it may be important to examine the number of erroneous assessment letters sent to taxpayers due to image processing errors, mailings sent to known bad addresses, or frequent tax form-use errors (which may have implications for form redesign).
- Making the electronic filing (e-file) data, which is cleaned and prepared by the Research Section, accessible to the Corporation and Personal Income Tax programs for the purposes of enhancing audit selection, legislative analysis, and possibly form and instructions development.
- Extract further data from automated processes including associated rule-sets and the reasons for certain automated "decisions" being made, including, for instance, why a case was sent for audit or for suspension, or why a document was flagged for correction following Optical Character Recognition.
- Improve data collection on job or task assignments to identify if changes are necessary.
- Gain access to the GenTax analytics tool to explore its capabilities.
- Consider leveraging Outlook Insights add-on data to track internal performance measures such as email response times by staff.
- Consider ways to integrate data available in the Department to enhance auditing and other processes (e.g. timber tax or cadastral geographic information with Personal Income Tax to better target auditing pools, or similarly using data from the county data exchange program), with due regard, of course, to applicable legal constraints on such sharing.
- Consider ways to standardize the efforts across the Department to build queries using SQL to access data in GenTax. For example, standardized schemas of parts of the database could be shared with data users to ensure the right assumptions are used.
- Create data sets for specific types of analysis by merging fields from multiple tables. For example, the external database in GenTax has a table with combined information about businesses in the "Business Summary", and Research has created a file for analysis of receipts and refunds that combines information from multiple tables.

Personal Tax and Compliance Division

- Decide on the desired outcomes for audits and track relevant data.
- Conduct retrospective analyses to determine whether we are choosing the right people to audit.
- Improve collection scoring, i.e. determining who is most likely to pay.

Processing Center

- Integrate GenTax, *Quick* Modules, and INFOPoll data systems to a) automate creation of the weekly Revenue Leadership Team report (resulting in substantial time savings and more customizability/relevance for different leadership team members), b) analyze bottlenecks to reduce processing times, c) plan for future staffing more effectively.
- Improve data availability to understand why automated Optical Character Recognition processes went to edit and why suspensions took place.
- Enhance image processing of tax returns to reduce errors, such as missing decimal points in dollar figures or incorrectly interpreting addresses. It may be worth noting, in this respect, that image processing for the new Oregon Schedule OR-A, appears to be generating errors.
- Optimize business processes by estimating the tradeoffs between the speed of refunds, accuracy of returns processing, and resources spent on processing.

Property Tax Division

- Correlate demographics in the deferral tax program with external demographic data to better target outreach efforts.
- Determine and track metrics on the effectiveness of valuation-related litigation, such as the number and outcomes of appeals, as well as processing and appraisal work.
- Integrate various spreadsheets and databases into a single, well-structured database so that historical valuation data can be effectively extracted for use in appraisals.
- Investigate additional sources of data to determine real market value.
- Consider ways to overcome obstacles in acquiring better data for valuation, overcoming limited cooperation from experts in the field.

Business Division

- Look into and resolve discrepancies when comparing GenTax versus Finance Section reports.
- Enhance audit task time-use data to be able to calculate return on investment.

Collections Division

- Enable access to data in GenTax to perform basic analytics on letter collections (e.g. the total penalty assessment included in collection letters) and cases (e.g. the amount of penalties waived).
- Fine tune collections approach after examining what works. For instance, analyze the data to identify the optimum frequency and medium of notices, the optimal tone or language of notices, those debtors who need a phone call, and which private collection firm should be used for which kind of collections.

The above ideas are initial thoughts on how data and data analytics might be further used in the Department. A more strategic and considered approach, however, would require a more systematic examination of the question that draws on the viewpoints of more people and carefully examines current practices and workflows. Even so, the above ideas may inform the ongoing work of the Department to enhance use of data through its Data Strategy Group, which is, among other things, considering the implications of the State of Oregon's data strategy (see https://www.oregon.gov/das/OSCIO/Pages/DataStrategy.aspx).

Some ethical and practical considerations

With the exponential increase in data generation, availability, and use in the world, the appeal and promise of data itself and data analytics, but also their potential perils and pitfalls, are gradually being recognized.

Before reviewing some of the benefits and critical considerations in working with data and data analytics, it is important to place them within a wider context. As defined earlier, data analytics is any means of using data to inform people's perspectives or understanding so that they might take more effective action. A simple model, which is depicted visually below, may help frame the discussion in the following subsections.

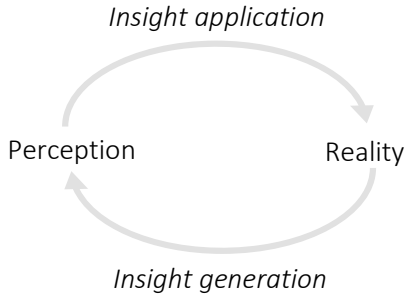


Diagram 1. A simplified depiction of the process through which insights are generated and applied.

It is intuitively obvious that people's understanding or perception of reality is not the same as reality itself. In the context of taxation, for instance, there is a difference between tax payers' actual behavioral response to tax communications and the kind of response that a tax administrator may anticipate, or between the ways auditors in the Department spend their time versus the allocation expected by a supervisor.

People use a variety of practices to understand reality (the "*Insight generation*" arrow in the model above), such as reflecting on personal or collective experience, studying bodies of knowledge, and relying on intuition. Another way to better understand reality is defining, collecting, and analyzing data. The distinction between each practice, is not, however, so sharp, and it is difficult to successfully engage in one without also engaging in another. For instance, defining what data to collect requires an idea as to the kind of information that matters—an idea that is often informed by personal or collective experience. Selecting appropriate techniques with which to analyze data depends on the study of existing bodies of knowledge regarding statistics and probability. In sum, the process of generating insights, regardless of the extent to which data is relied on, is a multifaceted human endeavor.

Once a tentative insight has been identified, it must be applied to effect some change in the world (the "Insight application" arrow in the model above). Returning to the tax administration context, if an analysis of the data suggests that there is a bottleneck at a certain point in the processing of incoming tax returns, then specific next steps within the Department should be considered and implemented. Other analyses might generate more specific insights, such as regarding the kinds of letters that different groups of taxpayers should receive. Even in these cases, an implementation plan would need to be developed and executed to translate this idea into action.

A robust process of learning within an organization, however, depends not only on the quality and strength of the processes of insight generation and insight application, but also on their relationship. To the degree that insights gained are translated into action and new data are collected, it will be possible to gain a better understanding of reality and, perhaps more importantly, effect real and lasting change. Engaging in such an iterative process of learning requires a set of individual, cultural, and institutional capacities that foster conditions such as the following:

- Staff have sufficient clarity, latitude, and support so they can both appreciate existing practices and continuously take steps to improve their mode of operation.
- Staff can, in a service-oriented spirit of inquiry, consult about and suggest changes in practices without fear of being labelled as overly critical or experiencing other forms of (sometimes subtle) reprisal.
- Data and experience, including through experiments (such as trying out new ways of completing tasks), are valued and inform decision-making
- Groups of staff working in similar areas maintain continuity on specific areas of learning, routinely referring to and refining existing documentation.

Benefits and limitations of data

There are many benefits from creating and using data to enhance understanding. The profound contributions of the sciences, which heavily rely on the painstaking generation and analysis of data, testify to it. In an organizational context, data, particularly in its more structured forms, can serve as a repeatedly measurable and comparable unit of information that can be similarly interpreted by many people, in some cases facilitating building consensus. As such, data can sometimes serve as a more reliable benchmark and/or a shared reference point as compared to anecdotal or experiential knowledge. Data, when managed and interpreted appropriately, can help people discover new patterns that would have otherwise remained hidden or at least stimulate curiosity and lead to new questions. Furthermore, the increasing velocity of data storage and computation can, if complemented by the appropriate individual and collective capacities, accelerate the rate at which organizations generate and apply new insights about reality.

For all the possible benefits that may be reaped from data and data analytics, a host of considerations have emerged, particularly in the context of using data to understand social phenomena and make decisions that affect the lives, of sometimes millions, of people. This section provides a mere sampling of the set of issues involved.

A wide array of considerations has to do with the inherent limitations of data as an abstraction and description of reality. The following are just a few of such considerations:

- **Data Framing**. Just as a map is not the same as the territory, data is a specific encoding or representation of some aspect of reality. As such, data includes some information and excludes other information. The selection of what information is codified as data will be limited and informed by the interests of the people designing the data collection mechanisms (Johnson, 2014).

- **Data Representativeness**. Data sometimes "over-represents those already privileged and under-represents those less likely to be part of data producing interactions" (Johnson, 2014), such as in the case of Census data which tends to undercount Black and Hispanic households (ibid.). In the case of tax-return administrative data, there are myriad biases, the direction of which is

sometimes unclear or might sometimes change depending on the population being considered. One of the most well-known cases of this is nonrepresentation of people who do not file a tax return, most of whom have lower incomes. Other kinds of biases include rationally overstating tax liability, such as taxpayers avoiding itemizing deductions due to compliance costs, even when it would be in their financial best interest to do so, as well as willful noncompliance, such as when taxpayers underreport income to avoid tax liability (Slemrod, 2016).

- **Data Quality**. The quality of the data itself also matters significantly if conclusions are to be effectively drawn from it. To this end, data must be carefully and actively managed, including periodically reassessing how data is collected and evaluated (OECD, 2016). Indeed, in many cases analysts spend more time cleaning data than conducting analysis per se (Wiseman, 2017).

Once data has been collected, it must be interpreted to have any useful meaning. To do this, analysts have an array of analytical methods at their disposal, each of which requires its own set of capabilities.

In the case of data visualization as a means of analysis, the designer of the visualization should be well-versed in the many well-documented best practices, for example those having to do with the selection of color, graphs, and scales (see, for example, Midway, 2020; Berinato, 2019; Jones, 2019). Otherwise, data visualization can mislead or, at best, confuse readers.

In the case of analytics that go beyond simple summary statistics and include econometric or other forms of statistical modeling, analysts should be sufficiently trained in probability theory to understand the assumptions and limitations of various models. They would bear in mind, for instance, the difference between theory- versus data-driven models[1], each of which has the potential to generate insights. In the case of data-driven models in particular, however, relationships may be observed in the data which are entirely spurious (Hand, 2018). Absent a basic knowledge of statistics, an analyst can grossly misuse models and misinterpret their results.

Finally, there are other analytical tools and algorithms which make inferences and decisions to automate various processes. Much like the previous case, analysts would do well to have a sound grasp of probability theory and a familiarity with the various tools at their disposal. However, particularly for this class of analytical tools, its creators and maintainers have an added responsibility for due diligence given the deep and wide-ranging impacts that such tools can have.

Indeed, various individuals and organizations have expressed deep concern regarding such tools as their use has proliferated among governments and corporations and have, at times, been grossly mismanaged at the expense of thousands, if not, millions of people. (See *Box 3* below for a short case study about a state government that decided to automate certain unemployment claims processes, unfortunately, at the direct expense of tens of thousands of people.) The public has expressed a variety of concerns regarding such tools. For instance, many people point out that algorithms based

[1] Theory-driven models presume certain relationships within the data, whereas data-driven models do not. For instance, economic theories generally suggest that price has an important (and generally negative) effect on consumer demand. Theory-driven models typically used by economists would, therefore, include price(s) in market analyses. A data-driven approach would, instead, make no assumption about which variables should be included in an analysis. In the latter approach, variables (including price) would be added and removed until the "best" set emerges. This approach is used, for instance, in automated email spam filters, which need to rapidly adjust to changing spammer practices.

or "trained" on biased data (e.g. previously racially motivated policing decisions) will result in biased recommendations or automated decisions. Others have highlighted emerging problems such as determining who is accountable for governmental decisions, ensuring transparency for automated processes, and enhancing the capacity of governments to effectively rely on, assess, interpret, and evolve such algorithms, etc. (see for example Crawford & Schultz, 2019; Citron & Calo, 2020). Some have argued that artificial intelligence and advanced analytical techniques are best used when they enhance rather than replace capacity within government agencies (Citron & Calo, 2020).

Considerations for organizations

The cursory treatment above of the limitations of data and data analytics in a social context sheds light on some critical considerations when data and data analytics are employed in an organization. (See *Box 4* on the next page for recent Federation of Tax Administrators survey results that reinforce the following considerations).

The first is that data and data analytical tools do not replace human judgment (Hoffman, 2021). In the

Box 3. Summary and cursory analysis on the Michigan Integrated Data Automated System and unemployment claims

Around 2010, the Michigan Unemployment Insurance Agency sought to upgrade its 30-year-old system to manage unemployment claims in order to increase accuracy and reduce operational costs. The Michigan Integrated Data Automated System (MiDAS) was then developed over the course of two years with the assistance of private contractors at a cost of approximately $45 million, resulting in increased automation and a reduction in staff at the agency by about one third (i.e., 400 people). Initially, many at the Unemployment Agency felt the System was largely a success. However, it was subsequently steeped in controversy as it was reported that between 2013 and 2015 about 90% of fraudulent cases were mis-identified by MiDAS, resulting in tens of thousands of individuals being wrongfully accused of unemployment fraud and experiencing severe personal and financial difficulties. In some cases, people's homes were foreclosed and they became homeless. In 2015 the Agency stopped using MiDAS for fraud determinations, and in 2017 it completed a review of fraud cases and reversed 70% of them, resulting in refunds totaling $21 million to claimants.

A thorough analysis of this case is beyond the scope of this paper. However, a few possible contributing factors, which are gleaned from various analyses, are listed below in no particular order:

- MiDAS may have been relying on corrupt and inaccurate data, including legacy data that could not be effectively converted and scanned information which could not be read.
- The Automated System's findings appear to have been initially unchecked, despite a rising number of appeals and important evidence lodged by claimants in 2014.
- A substantial number of fraudulent claim designations appear to have been automated, with no human review.
- Despite it becoming easier, through MiDAS, for the Agency to label claims as fraudulent, disadvantaged individuals were ineligible for free representation at appeals hearings.
- Claimants did not appear to receive timely notices that their cases were considered fraudulent.
- Misaligned financial incentives may have encouraged employers to misrepresent, in response to Agency requests for verification, the reason for employees' departures.
- Michigan legislation that allowed for more severe garnishment policies may have exacerbated the effects of automated decisions to regard claims as fraudulent.
- Federal and state law appear to have been insufficiently applied in unemployment claim decision-making and communications.

(Sources: Egan 2017, Citron & Calo 2020, Shaefer and Gray 2015, Cahoo vs. Fast Enters, 2021)

final analysis, it is people who must make sense of the data, design analytical tools, and consider the value of the results in the context of information and knowledge acquired through other means. These people need to adopt the right assumptions, including about the data generating process itself, use the right tools, and develop the right capabilities in order to reach veritable insights from the data.

Second, practitioners have noted how critical it is to focus on the issues that require attention and then identify the data that might assist with deriving actionable insights rather than focusing primarily on existing or readily available data (Wiseman, 2017; de Langhe 2021; West 2020). By focusing exclusively on existing data, for example, erroneous conclusions might be reached that fail to take into consideration the limitations of the data. By focusing instead on the decisions that need to be made or the issues that need to be resolved, analysts can ensure that the right data is selected or collected.

Third, data use can foster and is enriched by a culture of learning and collaboration. Close collaboration between those that are generating the data, those interpreting it, and those acting on it can ensure that interpretations are more accurate and meaningful (Centre for Public Impact, 2018). The IRS, for instance, has data scientists work together with subject matter specialists, including revenue agents, and revenue officers (Lee, 2020). As discussed previously, data analysis begins with data definition and generation. Those who define the data being collected, such as subject matter specialists, as well as those who routinely generate and store data, such as revenue officers, need to collaborate with data scientists, who are well versed in analytical methods and the related research literature. In this way it is more likely for the data generating process to be well understood, for the right analytical techniques to be selected, and for insights with practical importance to be identified correctly.

Box 4. Recent lessons learned from state tax administrations on predictive analytics

Around June of 2021, a few state tax administrators in the United States shared their reflections on the following questions: 1) Has your agency recently completed a project where you have implemented data warehouse and predictive modeling component? 2) If yes, when did you implement it? 3) If yes, what technology are you using? 4) If yes, do you have feedback on how it's working? 5) If yes, are there any lessons learned you would share with another state considering this?

By mid-July, tax administrators from three states replied to the questions indicating that they had relevant experience. The following is a summary of, and in some cases extracts from, their answers to the last question, number 5), on "lessons learned":

- Subject matter experts should be engaged to understand the data and create predictive models.
- Consideration should be given not only to building but maintaining advanced analytical models because they need to be continually updated to remain relevant. Staff with the relevant skills, therefore, must be identified to do this work.
- People make programs, not tools. Sound foundations such as employing a scientific approach, understanding data, and being clear about the goal are more important than the tools selected. This implies managing data quality through data governance; documentation of assumptions and biases; focusing on sufficiently specific questions; allowing data to guide but not restrict thinking; and devising ways to validate model results.
- Buy-in is critical to ensure that the results of data analytics are actually used.
- Define success criteria and determine when an effort should be cancelled. Not all efforts succeed.

Such a collaborative approach has implications for the structure of an organization and how best to situate data analytical capacities. In a study published by the Organisation of Economic Co-operation and Development, which examined the use of advanced data analytics in 16 different countries in 2015, it was found that "[e]stablishing an effective advanced analytics function requires administrations both to create a dedicated, cohesive team (in order to ensure quality control and build capabilities) and to integrate analytics into the wider organisation (in order to establish an effective working relationship between analytics and operational teams). Leaders of advanced analytics functions therefore need to strike a balance between centralisation, which supports internal cohesion, and de-centralisation, which supports integration into the wider organisation. Survey responses suggest that in the early phases of development centralisation may be more appropriate, with activity becoming increasingly de-centralised as the analytics function matures." (OECD, 2016)

A study published by Harvard University examining the functioning of government Chief Data Officers across the United States similarly included a discussion of centralized versus decentralized arrangements for organizing data analytical capacity. In this regard, the author of the study noted that "[w]hile the centralized and decentralized models present opposite ends of a spectrum, some organizations employ a blended or hybrid model that uses the best elements of the two basic models. The hybrid model has been found to be the most effective in the private sector. In government, most CDO offices follow the hybrid model." (Wiseman, 2017)

Additional guiding principles

The comments above scratch the surface when it comes to the insights that a data-driven learning-oriented organization would likely have incorporated into its everyday practice. Indeed, the field of data science is a wide and diverse field that is rapidly evolving both in terms of increasing technical capacities and a myriad of difficult ethical questions. In one study it has been noted that, given such rapid fluctuation in the field, it may be better to rely more on ethical principles to guide action rather than only specific regulations and rules (Hand, 2018).

The Oregon Department of Administrative Services, Enterprise Information Services division published (EIS, 2021) Oregon's Data Strategy in 2021, which includes the following data principles:

Governance and Effective Management

1. Govern: Manage data as a strategic asset for the public good
2. Leverage: Use the State's data to improve the lives of Oregonians through effective and efficient government
3. Protect: Preserve the privacy, quality, and integrity of the data we hold in trust
4. Share: Promote responsible data sharing across agencies and with external partners, including the public

Ethical Use

5. Plan: Be intentional in our collection and use of data and design with equity and the future in mind
6. Engage: Embrace data justice in how we collect, use, and share data for the communities we serve

7. Show: Model transparency in our work to educate others about our data assets and how they are used and seek to build feedback loops between the State and our constituents

Data Informed Culture

8. Learn: Promote a statewide culture of learning and collaboration in the use and analysis of data
9. Autonomize: Educate data leaders within our organization and enable all individuals to use data appropriately, ethically, and effectively
10. Lead: Establish structures for accountability and responsibility for the management of our data for all people we serve

* * *

It is hoped that the concepts, ideas, lessons learned, and guiding principles discussed in this paper may in some way contribute to the ongoing efforts at the Department of Revenue to use its data in a considered and effective way.

Works cited

Berinato, S. (2019). Good Charts Workbook. Harvard Business Review Press.

Bernard, M.J. (2021). Analyze This: Putting Performance Before Analytics. TaxNotes. https://www.taxnotes.com/tax-notes-today-international/tax-technology/analyze-putting-performance-analytics/2021/04/16/3k553

Cahoo v. Fast Enters, Case Number 17-10657 (E.D. Mich. Mar. 25, 2021) https://casetext.com/case/cahoo-v-fast-enters-llc-5

Centre for Public Impact. (2018). Artificial intelligence in taxation A case study on the use of AI in government. https://www.centreforpublicimpact.org/assets/documents/ai-case-study-taxation.pdf

Citron, D.K, R. Calo. (2020). The Automated Administrative State: A Crisis of Legitimacy. https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=1835&context=faculty_scholarship

Crawford, K., & Schultz, J. (2019). AI systems as state actors. Columbia Law Review, 119(7), 1941-1972.

de Langhe, B. & S. Puntoni. (2021). Leading With Decision-Driven Data Analytics in MIT Sloan Management Review Spring 2021 Issue. https://sloanreview.mit.edu/article/leading-with-decision-driven-data-analytics/

Department of Revenue. (2019). Outcome-Based Management Assessment Report. https://olis.oregonlegislature.gov/liz/2019R1/Downloads/CommitteeMeetingDocument/180249

Egan, P. (2017). Data glitch was apparent factor in false fraud charges against jobless claimants. Detroit Free Press. https://www.freep.com/story/news/local/michigan/2017/07/30/fraud-charges-unemployment-jobless-claimants/516332001/

Enterprise Information Services (EIS). (2021). Oregon's Data Strategy, Unlocking Oregon's Potential 2021-23. https://www.oregon.gov/das/OSCIO/Documents/68230_DAS_EIS_DataStrategy_2021_v2.pdf

Griffith, C., A. Benjamin, S. Jeffrey, L. Sarah. (2020). Artificial Intelligence Isn't Here Yet, but It's Already Changing Tax: Transcript. TaxNotes. https://www.taxnotes.com/tax-notes-federal/accounting-periods-and-methods/artificial-intelligence-isnt-here-yet-its-already-changing-tax-transcript/2020/12/21/2d9vd

Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now?. Big data, 6(3), 176-190.

Hoffman, W. (2021). IRS Data Analytics Pros Still See Need for Human Touch. TaxNotes. https://www.taxnotes.com/tax-notes-federal/tax-system-administration/irs-data-analytics-pros-still-see-need-human-touch/2021/03/15/3k5r2?highlight=%22data%20analytics%22

Johnson, J. A. (2014). From open data to information justice. Ethics and Information Technology, 16(4), 263-274.

Jones, B. (2019). Avoiding Data Pitfalls. Wiley, 1st Edition.

Keightley, M.P. (2019). Better IRS Letters Could Improve Tax Compliance, CRS Says. TaxNotes. https://www.taxnotes.com/tax-notes-today-federal/compliance/better-irs-letters-could-improve-tax-compliance-crs-says/2019/08/07/29t7d

Lee, F. (2020). IRS Official Highlights Use of Data Analytics, Hiring Push. https://www.taxnotes.com/tax-notes-federal/compliance/irs-official-highlights-use-data-analytics-hiring-push/2020/12/07/2d8pl?highlight=%22data%20analytics%22

Midway, S. R. (2020). Principles of effective data visualization. Patterns, 100141. https://www.sciencedirect.com/science/article/pii/S2666389920301896.

Ordóñez, P. J., & Hallo, M. (2019, April). Data Mining Techniques Applied in Tax Administrations: A Literature Review. In 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG) (pp. 224-229). IEEE.

Organisation for Economic Co-operation and Development (OECD), (2016) Advanced Analytics for Better Tax Administration: Putting Data to Work, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264256453-en

Pijnenburg, M. G. F., Kowalczyk, W. J., Dijk, H. V., & van der Hel-van, D. E. (2017). A roadmap for analytics in taxpayer supervision. Electronic Journal of e-Government, 15(1), 19-32. https://scholarlypublications.universiteitleiden.nl/access/item%3A3099226/view

Results Washington. (2019A). Strategic Lean Project Report Reporting Period: January – December 2019, Property Tax Exemption Application for Nonprofit Organizations. https://results.wa.gov/sites/default/files/Department%20of%20Revenue%20-%20Property%20Tax%20Exemption%20Application%20for%20Nonprofit%20Organizations.pdf

Results Washington. (2019B). Strategic Lean Project Report Reporting Period: January – December 2019 Warehouse Tax Incentive. https://results.wa.gov/sites/default/files/Department%20of%20Revenue%20-%20Warehouse%20Tax%20Incentive.pdf

Sarin, N. & Summers, L.H. (2019). Shrinking the Tax Gap: Approaches and Revenue Potential. TaxNotes. https://www.taxnotes.com/tax-notes-today-federal/compliance/shrinking-tax-gap-approaches-and-revenue-potential/2019/11/18/2b47g

Shaefer, H.L., & S. Gray. (2015). Michigan Unemployment Insurance Agency: Unjust Fraud and Multiple-Determinations. Memorandum to Gay Gilbert, Administrator at the US Department of Labor. https://waysandmeans.house.gov/sites/democrats.waysandmeans.house.gov/files/documents/Shaefer-Gray-USDOL-Memo_06-01-2015.pdf

Slemrod, J. (2016). Caveats to the research use of tax-return administrative data. National Tax Journal, 69(4), 1003.

West, A. & E. Hume. (2020). Becoming a Data-Driven Decision Making Organization. The CPA Journal. https://www.cpajournal.com/2020/05/25/becoming-a-data-driven-decision-making-organization/

White, K. (2019). States Continue Efforts to Advance the Use of Data and Evidence. https://uidl.naswa.org/handle/20.500.11941/2364_States-Continue-Efforts-to-Advance-the-Use-of-Data-and-Evidence

Wiseman, J. (2017). Lessons from leading CDOs: A framework for better civic analytics. Ash Center Policy Briefs Series. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42372452