TECHNICAL

NOTES & MANUALS

Tax Administration

Essential Analytics for Compliance Risk Management

Joshua Aslett, Gustavo González, Stuart Hamilton, and Miguel Pecho

TNM/2024/01

TECHNICAL NOTES AND MANUALS

Tax Administration
Essential Analytics for Compliance Risk Managment

Joshua Aslett, Gustavo González, Stuart Hamilton, and Miguel Pecho

Authorized for distribution by Abdelhak Senhadji

This technical note and manual addresses the following questions:

- What is meant by analytics, and why are they used in tax administration?
- How does analytics support compliance risk management (CRM) processes?
- Why are statistics and data science important for modern analytics?
- What data, technology, and tools typically support CRM analytics capabilities?
- For CRM, what analytics services, outputs, and knowledge are essential?

The authors acknowledge the valuable feedback and contributions provided by Debra Adams, Frank van Brunschot, Margaret Cotton, Aquiles Faris, Dmitri Jegorov, Michael Hardy, Andrea Lemgruber, Andrew Masters, John McAlister, and Stefano Pisani (all IMF); as well as Ignacio Gonzalez, Rifat Hyseni, and the staff of the tax administrations that have evaluated the toolkit accompanying this note.

Cover Design: IMF Creative Solutions

Cataloging-in-Publication Data
IMF Library

Names: Aslett, Joshua, author. | González Amilivia, Victor Gustavo, author. | Hamilton, Stuart Gordon, author. | Pecho, Miguel, author. | International Monetary Fund, publisher.
Title: Tax administration : essential analytics for compliance risk management / Joshua Aslett, Gustavo González, Stuart Hamilton, and Miguel Pecho.
Other titles: Essential analytics for compliance risk management. | Technical notes and manuals.
Description: Washington, DC : International Monetary Fund, 2024. | Feb. 2024. | TNM/2024/01. | Includes bibliographical references.
Identifiers: ISBN:
9798400260063 (paper)
9798400260124 (ePub)
9798400256264 (Web PDF)
Subjects: LCSH: Tax administration and procedure. | Public administration and public policy.
Classification: LCC HJ2305.A8 2024

Recommended citation:
Aslett, Joshua, Gustavo González, Stuart Hamilton, and Miguel Pecho. 2024. "Tax Administration: Essential Analytics for Compliance Risk Management." IMF Technical Notes and Manuals 2024/01. International Monetary Fund, Washington, DC.

Please send orders to:
International Monetary Fund, Publication Services
PO Box 92780, Washington, DC 20090, USA
Tel: (202) 623–7430 | Fax: (202) 623–7201
publications@IMF.org
elibary.IMF.org
bookstore.IMF.org

Contents

This page intentionally left blank

Abbreviations

AI	artificial intelligence
B2B	business-to-business
B2C	business-to-consumer
BISEP	business, industry, sociological, economic, psychological
CBC	country by country
CRM	compliance risk management
EDA	exploratory data analysis
ETL	extract, transform, load
G2B	government to business
G2G	government to government
HWI	high-wealth individual
IQR	interquartile range
IT	information technology
LLM	large language model
NLP	natural language processing
OECD	Organisation for Economic Co-operation and Development
OLAP	online analytical processing
RDF	risk differentiation framework
ROC	receiver operating characteristic
RPA	robotic process automation
SAF-T	Standard Audit File for Tax Purposes
SQL	Structured Query Language
VAT	value-added tax
VITARA	Virtual Training to Advance Revenue Administration

This page intentionally left blank

I. Introduction

This note is intended as a starter kit for building practical analytics capabilities that strengthen compliance risk management (CRM) in tax administration. The note opens with an overview of contemporary CRM practices, highlighting principles that influence the use of data. It then provides a short summary of statistics and data science as the primary disciplines that analytics rely on. In the context of CRM, the note subsequently presents illustrative surveys of typical (1) data, tools, and technology; and (2) analytics outputs and services. In its annexes, the note addresses critical knowledge areas, key challenges, and three selected topics—compliance planning, taxpayer profiling, and audit case selection—each critical to CRM but often challenging for analytics groups to address.

A companion toolkit providing templates and tools is available for download from the IMF Fiscal Affairs Department Revenue Portal.[1] Diverse technology for analytics is used by tax administrations around the world.[2] As a matter of policy, the IMF does not endorse specific products or suppliers. If adopted for operational use, the templates provided in the companion toolkit should be converted to use the analytics technology that is most appropriate in local circumstances. While the templates incorporate current good practices, they are provided for educational purposes only and without warranty of any kind. Instructions for accessing and making use of the toolkit are available in Annex 1.

The guidance in this note builds from a broad perspective that defines analytics as "the practice of using mathematics to analyze data." Regardless of the techniques employed, which can range from the use of simple arithmetic to advanced computational algorithms, all forms of analytics rely on mathematics. As mathematics is universally applicable, analytics have the potential to create value in any domain where data exists. To do so, a blend of technical expertise (such as with statistics) and topical knowledge (for instance, of a particular business problem) is required to process data, discover patterns, interpret, and then communicate their meaning. In this note, these and related activities are considered elements comprising *analytics*.

In tax administration, data is pervasive, and the role of analytics has emerged as a topic of strategic importance. The collection of public revenues has always relied on information. Although concepts and techniques have evolved over time, tax collectors have long maintained registers, methods of assessment, and journals for accounting. In recent decades, information technology (IT) has built from these ideas, advancing the development of strong headquarters functions, specialization of staff, and the ability of data to inform decision making at all levels. Because of this, many tax administrations believe that the effectiveness of their analytics work impacts their overall performance, particularly when seeking to efficiently allocate resources.

Apart from strategic implications, most tax administrations cannot effectively deliver on basic mandates without the use of analytics. Examples span the full range of administrative functions. Among supporting other essential work, analytics are today used to help identify candidates for registration, monitor filing and payment, process assessments, understand stocks of arrears, analyze taxpayer segments, identify noncompliance, and predict future behaviors.

As analytics can be applied to these and countless other topics, CRM provides a sharp lens through which their use can be organized. The next sections explore analytics in the context of reasonably developed CRM practices. This exploration begins with a brief overview of CRM itself.

[1] Accessible at https://www.imf.org/revenueportal.

[2] See OECD 2016.

II. CRM Theory, Framework, and Principles

A rich and continually evolving subject, CRM at its core seeks to influence the behavior of taxpayers. Conceptually, its logic is premised on theories and principles of social sciences. In its classic form, practitioners of CRM seek to exert influence by blending service and enforcement activities in a manner consistent with carrot-and-stick metaphors. In other forms, they seek to reinforce dimensions of trust and social norms that exist between taxpayers, their peers, and their collective relationship to public institutions. Regardless of the approach taken, productive uses of CRM tend to rely heavily on risk analysis and intelligence to support the development of focused compliance campaigns, strategies, and plans supported by deliberate, tactical interventions.

The full range of basic CRM practices is best introduced through other resources available online. Those new to CRM may wish to explore (1) the online course on CRM provided through the Virtual Training to Advance Revenue Administration (VITARA) program; and (2) IMF technical notes that describe CRM, the adoption of a CRM framework and the development of compliance improvement plans.[3]

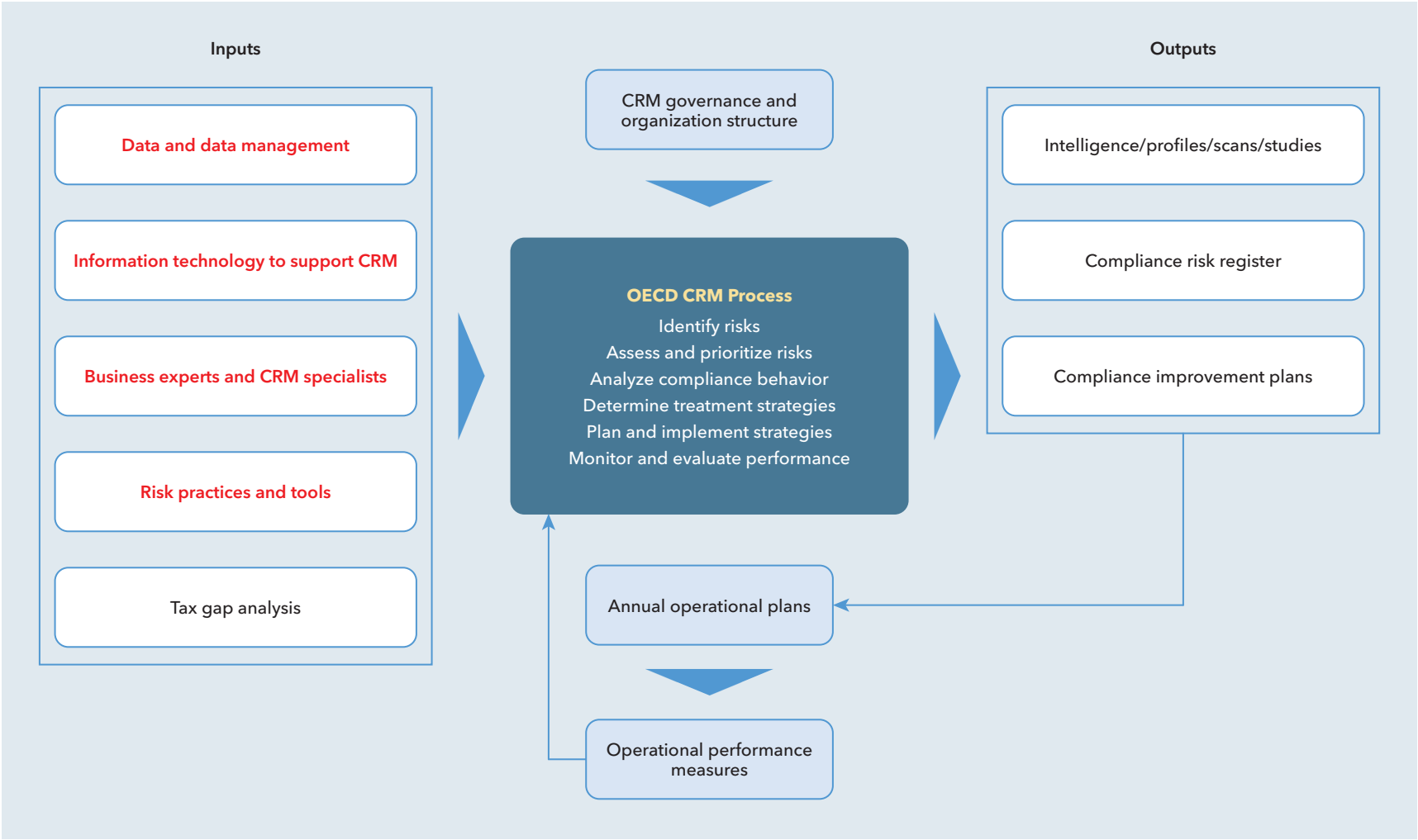
These resources are all premised on the idea that in any given jurisdiction, a state of quasi-compliance equilibrium exists—*and CRM can affect its balance*. Derived from the totality of attitudes, actions, and beliefs among individuals and administrations operating in the same tax system, this equilibrium represents a status quo of behaviors. Its balance results in a compliance gap (that is, a difference between potential tax revenue collectible according to the law and actual revenue). Comprising different components, these gaps vary by industry, segment, and issue. Many administrations view closing gaps as a strategic imperative that CRM can support. Doing so requires resources, higher-order analytical skills, and an understanding of CRM's processes.

As Figure 1 illustrates, CRM lends itself well to analytics by combining data and domain expertise to identify risks, interpret behaviors, and inform interventions. Conceptually, CRM is typically introduced using a process framework published by the Organisation for Economic Co-operation and Development (OECD) in 2004. Outputs from the framework provide a range of intelligence. Supported by analytics, staff in CRM groups identify, qualify, predict, and prevent compliance risks. Beyond describing risks, they aggressively seek to understand why particular behaviors are presented. With deeper insights, responses to risks can move beyond single-action enforcement activities such as audit to comprehensive strategies designed to deal with more structural issues. Figure 1 presents the OECD framework along with key inputs and outputs.

This note focuses on the Figure 1 inputs that are highlighted in red and the analytics that they enable. A topic of special interest among many administrations, the integrated application of CRM inputs can be an opaque area. In many jurisdictions, the specific techniques and methods used are confidential and rarely disclosed. In others, certain details may be published for various purposes (to include demonstrating fairness or supporting related anticorruption initiatives). Important differences also exist across administrations in the levels of maturity and sophistication of both the inputs themselves, and the specific analytics used. Recognizing these issues, the guidance in this note is meant to be illustrative based on good practices consistently observed in operation.

[3] See Chooi and Pemberton (2023), Betts (2022), and Brondolo and others (2022).

Figure 1. Essential Elements of a CRM Framework



Sources: Betts 2022; and IMF staff.
Note: CRM = compliance risk management; OECD = Organisation for Economic Co-operation and Development.

Before exploring inputs and analytics capabilities, it is important to recognize that a core set of CRM principles tends to shape both. Through CRM, tax administrations should:

- ***Start with the core fundamentals.*** The four pillars of compliance—registration, filing, reporting, and payment—provide the foundation for CRM practices. While a need to actively manage compliance with each pillar is constant, with analytics, reporting is often best fully explored after first mastering an understanding of registration, filing, and payment.
- ***Focus on voluntary compliance.*** Recognizing the social aspects of administration, the purpose of CRM is to cost-effectively maximize voluntary taxpayer compliance with the law. Interventions in all forms require resources, some more than others (audit) and the ability to inexpensively influence macro compliance behaviors across each of the four pillars of compliance is paramount.
- ***Differentiate between taxpayers and issues.*** Because segments, industries, and issues all vary in their nature and degree of compliance, a one-size-fits-all approach to risk analysis is rarely appropriate. This logic is also true for compliance planning, requiring that interventions in whatever form they take be appropriately tailored.
- ***Integrate the application of administrative tools.*** Compliance campaigns, strategies, and plans must consider the full spectrum of potential resources available to execute integrated interventions that, collectively, achieve maximum impact (for example, combining data matching, media coverage, services, audit, and referrals for prosecution).
- ***Prioritize modification of current taxpayer behaviors.*** In a reactive CRM approach, work classically focuses on case selection for enforcement by evaluating decisions taken by taxpayers in the past (reflected in tax returns and other data). Although important, better results may be achieved by influencing decisions that are occurring in the present.
- ***Adopt an intelligence-driven posture.*** Once produced, data is immediately a reflection of the past. Because of this, and particularly with jurisdictions that struggle with data management, CRM should lean heavily on human intelligence collected from experts, analysts, partnering agencies, informants, and other resources deployed in the field.

With these principles in mind, exploration of analytics for CRM also requires insight into the potential offered by statistics and data science. The two topics are integral to all aspects of modern analytics, particularly those that are predictive in nature and increasingly used. While tools available today greatly reduce the need for knowledge of the underlying mathematics, a working understanding of the two topics' basic utility is nonetheless essential.

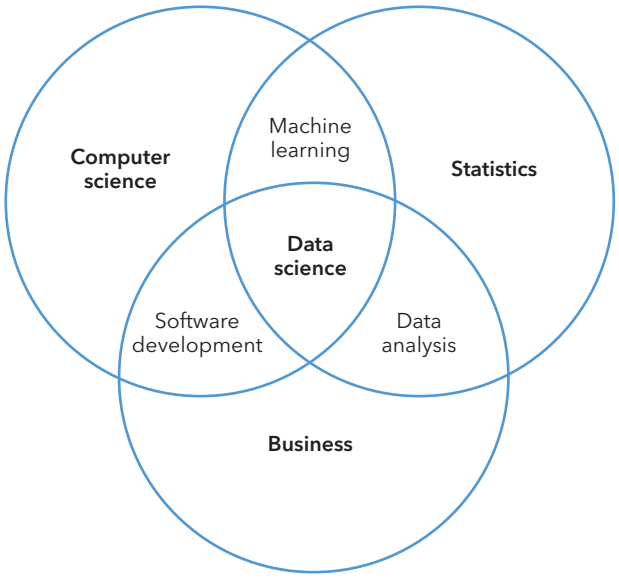
III. Understanding Statistics and Data Science

Statistics is a body of mathematical science that seeks to describe and interpret data using methods that can provide measures of probability and uncertainty. The use of statistics is a bedrock of the scientific method itself. An enabler of modern analytics, classical approaches to statistics generally make use of relatively small, but carefully designed data sets. With a smaller amount of data, statistics have historically been used to model experiments, test hypotheses, and establish relationships between variables. These practices are often used to convey the degree to which conclusions can be inferred and explained. To conduct this work, professional statisticians often have an education in statistics itself, applied mathematics, or a closely related subject.

Data science, by comparison, is a relatively new and interdisciplinary field that draws from statistics, computer science, and business. Emerging as a consequence of modern computing and IT, data science is increasingly recognized as a fundamental knowledge domain or learning discipline. Data scientists work with relatively large volumes of data, making use of the same principles that statisticians rely on. However, building from foundations provided for by statistics, they tend to focus on more operational aspects of the organizations that they support, developing intelligence products, data mining, writing software, and integrating their work into business systems—*a specialization well suited to the CRM inputs in Figure 1*.

While statistics are foundational to data science, clear lines of separation between the two fields are difficult to define and subject to debate. Figure 2 illustrates areas where integration and overlap tend to occur. Far from diluting any one field, the blending of disciplines is leading to the development of new, cutting-edge applications of long-established ideas (particularly, machine learning, which is today a major branch of work within the realm of artificial intelligence [AI]). At the center of these activities, data science has become an engine of rapid innovation. The underlying disciplines that it draws from, while also continually evolving, do so generally with a heightened requirement for rigor and, because of that, less quickly.

Figure 2. Knowledge Domains: Statistics and Data Science



Source: IMF staff.

Building proficiency with statistics and data science requires an awareness of terminology, key concepts, and what they can achieve. Boxes 1 and 2 provide overviews of each. With some variations, the details are consistently used across the domains presented in Figure 2.

BOX 1. Essential Terminology

Data. Facts, figures, measurements, or observations collected and used as a basis for reasoning, analysis, discussion, or calculation.

Population. A similar set of entities, individuals, items, or events that has been selected for study and has at least one shared characteristic.

Sample. A smaller representation of a larger population selected using a predefined method for the purpose of making research and analysis feasible.

Function. A mathematical expression or set of rules that define relationships between a set of variables, often represented by a formula.

Distribution. A collection of data on a particular variable (for example, the amount of income declared) that may conform to a mathematical function.

Confidence interval. A measure of uncertainty, often interpreted as the probability of an experiment producing similar results within a range of values.

Algorithm. Logic, methods, or sets of specific steps for solving a problem that may or may not be codified for use in a computer programming language.

Artificial intelligence. The simulation of human intelligence, carried out using computer systems and software, relying heavily on statistics and data science.

Machine learning. A major branch of artificial intelligence, the use of data and algorithms to enable computers to learn without explicit programming.

Supervised learning. A category of machine learning that relies on humans to label data and train algorithms to make conclusions or predictions.

Unsupervised learning. A category of machine learning that does not require human training to identify patterns in data or to make conclusions or predictions.

Data mining. The use of statistics and data science to analyze large data sets and identify anomalies, correlations, patterns, and trends.

Structured data. Data with clearly defined relationships and data types (for example, normalized database tables) that are easily searchable.

Unstructured data. Data without clear definition, often stored in a format that requires specialized tools for analysis (for example, invoices scanned as images).

Source: IMF staff.

BOX 2. Key Concepts and Potential Utility

Modern analytics build from statistics:

- ***Descriptive statistics*** describe or summarize the features and distribution of a dataset using concepts such as mean, median, mode, range, skew and variance.
- ***Inferential statistics*** apply advanced mathematics to create explicit, structured models that examine sample data, test hypotheses, and draw conclusions about a larger population.
- ***Predictive analytics*** use statistics and machine learning algorithms to analyze often large volumes of data and predict probable future outcomes.
- ***Prescriptive analytics*** build from the conclusions of predictive analytics by analyzing outcomes of prior actions and making suggestions for a next best action.

Statistics and data science produce actionable outputs:

- ***Classification of observations*** using binary or multiclass predictive techniques (as examples, binary—compliant or noncompliant; multiclass—high, medium, or low risk).
- ***Estimation of values*** using both linear and nonlinear estimators to predict outcome variables (for example, audit adjustment amounts).
- ***Estimation of probabilities*** using different techniques (for example, the Random Forest algorithm) to transition from the use of risk scores to more precise probabilities of risk.
- ***Grouping of observations*** using unsupervised techniques (for example, self-organizing maps) to identify and visualize shared traits that are not labeled or immediately obvious.
- ***Analysis of networks*** using specialized tools to identify relationships among entities, transactions, and events across vast amounts of structured and unstructured data.

Performance measurement is essential, starting with two concepts:

- ***A confusion matrix*** for summarizing both the success and error rates of a classification algorithm, which is a foundational data science concept.
- ***A receiver operating characteristic curve*** for visually plotting and confirming an algorithm's prediction power before its operational use.

Machine learning is increasingly important:

The ability of machine learning algorithms to discover patterns in large sets of data is revolutionizing risk practices. Traditionally, analysts have relied heavily on expert rules and explicit statistical models to explore risk top-down using selected ratios within industry and taxpayer segments. The success of these techniques depended largely on the knowledge and skills of analysts involved. By comparison, most basic machine learning algorithms (such as decision trees) can discover relationships in data relating to risk and make bottom-up predictions with a high degree of precision.

Source: IMF staff.

In tax administration, the CRM inputs that make use of these concepts are comprised of a diverse collection of data, technology, and tools. While often varying considerably, consistent patterns in their use do exist. Having a general awareness of the details is essential for exploring the introduction of new capabilities.

IV. Supporting CRM: Data, Technology, and Tools

As a practice, CRM builds from the institutions that support it, many of which are making major investments in digitalization and reform. The inputs to CRM tend to reflect the maturity of the institutions that create them, their operations, and the capabilities of their staff in general. Investments in IT are often transformational across these areas. They enable structural reforms that build from centralization of information, elimination of paper, strengthening of headquarters functions, reengineering of operations, and the greatly improved capture, management, and use of data. As nearly all administrations today have automated core aspects of their operations, many are now exploiting the fuller range of digital possibilities.

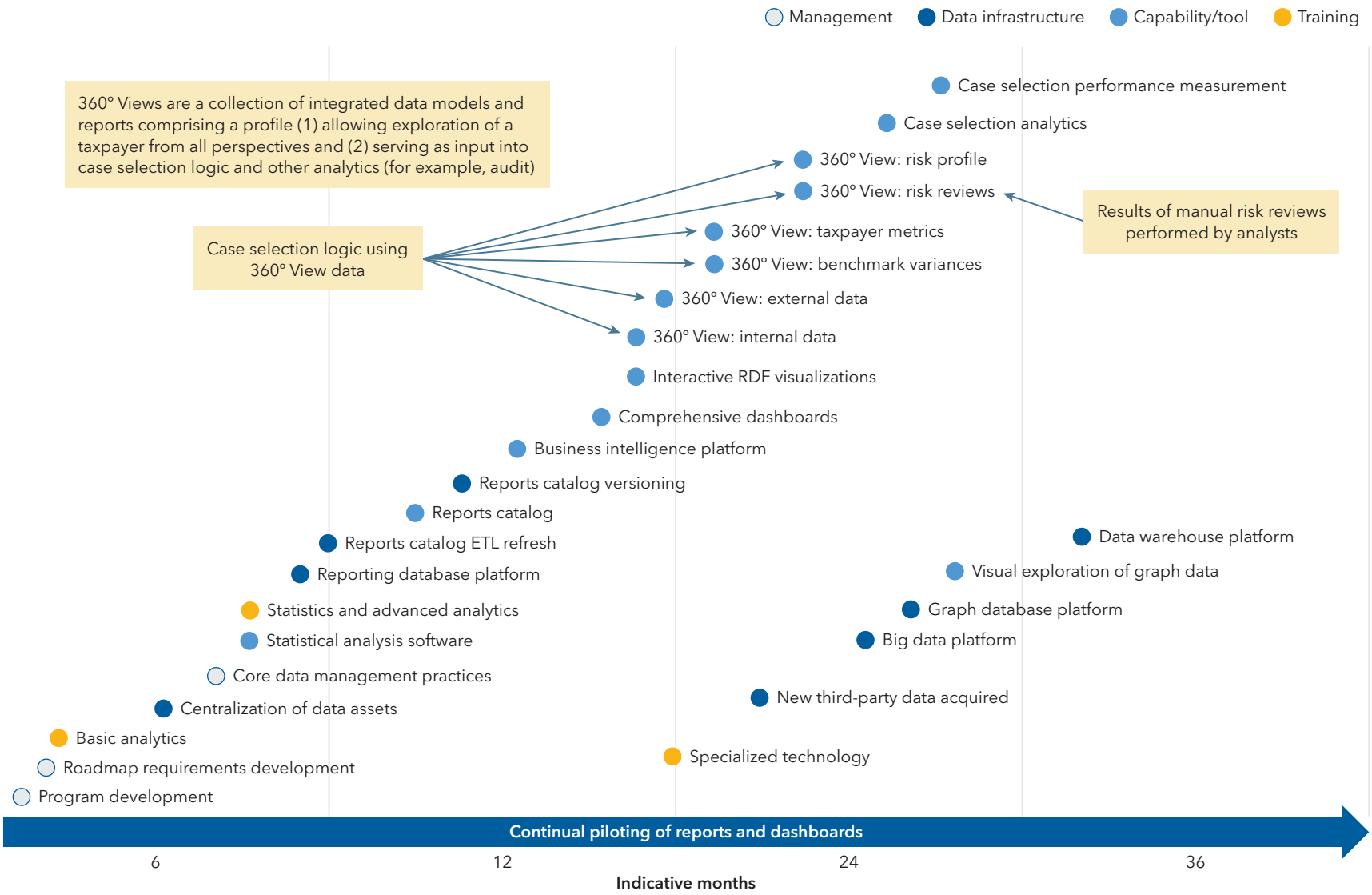
Further insights into digitalization, data, and information technology in general can be explored through other resources available online. These include (1) the forthcoming VITARA course on IT and data management; and (2) IMF technical notes on IT strategic planning in tax administration, core tax administration IT systems, and the implementation of commercial-off-the-shelf options for core systems.[4]

Among the changes possible in tax administration through digitalization, CRM is unique in that its technological inputs do not tend to bundle neatly into a system. For a variety of practical reasons, this can complicate their development. Procurement, as one key example, may be more challenging for CRM compared with other areas as multiple suppliers may be involved. The use of multiple products and IT platforms is typical, and their integration and development usually require the active involvement of specialists from various disciplines, including project management, software engineering, database development, IT administration, and operations.

To aid in exploring these issues, Figure 3 presents an illustrative overview of the typical, technologically oriented CRM inputs from a well-developed administration. Organized visually as a development roadmap, each of the inputs are described further in Table 1. In reviewing Figure 3, note the categorization of inputs and their sequential presentation. Collectively, development of the inputs illustrated is often a more appropriate starting point for scoping a CRM capabilities project than focusing on a standalone risk system. Elements of the roadmap that should be implemented in most administrations today are flagged as essential with a check mark in Table 1. For CRM specifically, most of the technological inputs used are built from layers of technology and data infrastructure introduced over time.

4 See Aslett and others (forthcoming) and Cotton and Dark (2017a, 2017b, 2017c).

Figure 3. Illustrative Roadmap for CRM Support: Data, Technology, and Tools



Source: IMF staff.
Note: CRM = compliance risk management; ETL = extract, transform, load; RDF = risk differentiation framework.

Table 1. Illustrative Roadmap for CRM Support: Data, Technology, and Tools

		Essential?	Milestone	Remarks
Management	1	✓	Program Development	The roadmap should be implemented as a collection of projects.
	2	✓	Roadmap Requirements Development	Formal requirements documents should be created and managed.
	3	✓	Core Data Management Practices	Data lifecycle strategies, architecture, and dictionaries are essential.
Data Infrastructure	4	✓	Centralization of Data Assets	All operational datasets via consolidation of platforms or replication.
	5	✓	Reporting Database Platform	A temporary platform for on-demand reports and ad-hoc analytics.
	6	✓	Reports Catalog ETL Refresh	Automation of "Extract, Transform, Load" (ETL) processing packages.
	7		Reports Catalog Versioning	Providing point-in-time data retrieval capabilities for key reports.
	8		New Third-Party Data Acquired	Acquisition and use of external data subject to appropriate controls.
	9		Big Data Platform	A standalone platform often used as a "data lake" for mass storage.
	10		Graph Database Platform	A database platform designed to support network analysis.
	11		Data Warehouse Platform	Implementation of a traditional data warehouse, a major project.
Capability/Tool	12	✓	Statistical Analysis Software	Used for both statistical analysis and predictive analytics.
	13	✓	Reports Catalog	Provides on-demand access to a catalog of standardized reports.
	14		Business Intelligence Platform	Facilitating the publication of dashboards, reports, and analysis.
	15	✓	Comprehensive Dashboards	Dashboards providing insights into operations and compliance.
	16		Interactive RDF Visualizations	A risk and planning tool for exploration of taxpayer segments.
	17	✓	360° View: Internal Data	Presents a snapshot of a taxpayer's internal tax administration data.
	18	✓	360° View: External Data	Presents a snapshot of a taxpayer's externally sourced data.
	19	✓	360° View: Benchmark Variances	Presents a taxpayer's performance relative to select benchmarks.
	20	✓	360° View: Taxpayer Metrics	Presents a taxpayer's tax, financial and other ratios of importance.
	21	✓	360° View: Risk Reviews	Presents the results of manual risk reviews performed by analysts.
	22	✓	360° View: Risk Profile	Presents the results of risk analytics and suggested next best actions.
	23	✓	Case Selection Analytics	Layers of analytics making use of algorithms and data mining.
	24	✓	Case Selection Performance Measurement	A collection of metrics that monitor case selection performance.
	25		Visual Exploration of Graph Data	Interactive tools that use a graph database for network analysis.
Training	26	✓	Basic Analytics	Topics: spreadsheets, database querying, and reports development.
	27		Statistics and Advanced Analytics	Topics: descriptive and inferential statistics, and predictive analytics.
	28		Specialized Technology	Topics: data warehouse, big data, and graph database platforms.

Source: IMF staff.

Apart from the inputs noted in Figure 3 and Table 1, the following additional tools and capabilities can be important and are observed as actively used by administrations in different regions of the world:

- ***Compliance risk registers.*** In a well-developed framework, CRM groups should at any time be capable of reporting the top compliance risks in their jurisdictions along with the measures being taken for mitigation. Doing so benefits from having a centralized tool for the registration and management of these risks.
- ***Industry risk profiles.*** Similar in concept to the 360° taxpayer views described in Table 1, an industry risk profile has both quantitative and qualitative aspects. In practice, a risk analyst profiling an industry captures the details in a risk database. If integrated, taxpayer risk profiles inherit the relevant risks from the industry that they belong to.
- ***Transaction risk profiles.*** Beyond taxpayers and industries, the profiling of discrete taxpayer transactions using similar concepts is increasingly a topic of interest. It is particularly relevant with electronic invoicing in context of standards for reporting (for example, Standard Audit File for Tax Purposes).
- ***Real-time risk processing.*** Making use of specialized technology, and particularly for high-volume transactions, risk analytics are being applied in real time to optimize and route work (for example, to risk assess invoices within a logical framework, applying an appropriate set of controls and selecting invoices for manual review or action).
- ***Platform integration.*** In recent years, some suppliers have integrated their analytics tools with other products offerings (for example, statistical processing software with workflow and case management platforms). The resulting capabilities can accelerate important aspects of transformation in tax administration, particularly with casework.
- ***Embedded analytics.*** Directly in core processing platforms, many tax administrations are embedding risk analytics in key workflow. Ranging from logic for compliance case selection to refunds processing and registration itself, analysts often manage updates through configuration and programming in the respective software and databases.
- ***Robotic process automation (RPA).*** An emerging topic of interest, robotic process automation offers much potential in tax administration through its ability to combine analytics, and especially machine learning, with bots that mimic human activity. The practical applications of this technology in tax administration are potentially far reaching.

Beneath all the data infrastructure, technology, and tools that can support CRM is the actual data itself, and the range of its potential use is immense. To grasp the potential of analytics in tax administration, an exploration of typical outputs and services can be helpful.

V. Supporting CRM: Essential Analytics

Viewed through the lens of CRM, the analytics possible are best explored by how they are organized, applied, and understood by specialists. Box 3 presents a concept for organization. It is most appropriate for analytics groups in well-developed tax administrations but useful as a reference for others. Its elements are used in strategy, planning, and regular operations.

As a reference, Tables 2 to 5 present illustrative analytics outputs and services for each of the topics in Box 3. The elements that should today be implemented in most tax administrations are flagged as "essential." To help understand how each element is produced, the primary type of analytics used is noted as descriptive, predictive, or prescriptive. For those making use of more than one type, the label "hybrid" is applied. The downloadable toolkit accompanying this note includes functional examples of a selection of the outputs and services described. Instructions for accessing the toolkit are provided in Annex 1.

BOX 3. Organization of Analytics for CRM Support

Compliance Strategy and Planning 1. Cross-Cutting Services 2. Segments and Issues a. Large, Medium, Small b. Sectors and Industries c. Category Risks d. Individuals	**Macrostrategic analysis and targeting** Making use of a variety of techniques and specialized models, outputs support and inform compliance campaigns, strategies, operational plans, and the regular activities of risk analysts.
Active Prevention 1. Compliance Forecasting 2. Targeted Customer Outreach a. Next Best Actions 3. Embedded Analytics 4. Real-Time Analytics 5. Third-Party Cooperation	**Analytics for proactive interventions** Relying on the full range of analytics, descriptive through prescriptive, outputs and models are supplied to operational groups for action and integrated into IT services.
Monitoring and Detection 1. Core Compliance Monitoring 2. Automated Underreporting Detection 3. Exploratory Data Analysis 4. Profiling Analytics 5. Third-Party Cooperation	**Analytics for research and operations** Requiring a solid foundation in descriptive statistics and proficiency with the underlying data sets, outputs inform risk analysts and support tools (360° Views).
Enforcement 1. Compliance Case Selection a. Next Best Actions 2. Case Selection Performance 3. Assessment and Notifications 4. Third-Party Cooperation	**Analytics for reactive interventions** Relying again on the full range of analytics, outputs improve the effectiveness of casework and support enforcement activities and actions taken in cooperation with external agencies.

Source: IMF staff.
Note: CRM = compliance risk management; IT = information technology.

Table 2. Illustrative Analytics: Compliance Strategy and Planning

	Essential?	Analytics	Output or Service	Remarks
1			**Cross-Cutting Services**	
1.1	✓	Descriptive	Strategy and Planning Templates	Provision of templates prepopulated with relevant statistics
1.2		Descriptive	Compliance Coverage Analysis	Identification of gaps in coverage by strategies and plans
1.3		Descriptive	Performance Analysis: Strategies and Plans	Evaluation of the macro impacts of strategies and plans
1.4		Descriptive	Performance Analysis: Directed Workflow	Evaluation of the activities that support strategies and plans
1.5		Hybrid	Cluster Analysis for Segmentation	Identification of unknown patterns of taxpayer segmentation
1.6	✓	Descriptive	Cumulative Count: Turnover and Tax	Identification of the most important taxpayers in a segment
1.7	✓	Hybrid	Identification and Selection of Targets	Regular selection of taxpayers, objects, or events for action
2			**Segments and Issues**	
2.a.i			**Large Taxpayers**	
2.a.i.1	✓	Descriptive	Large Taxpayer Criteria Review	Provision of options for refinement to existing criteria
2.a.i.2	✓	Descriptive	RDF Analysis: All Taxpayers	Evaluation of all large taxpayers in a single analysis
2.a.i.3		Descriptive	RDF Analysis: Accountants	Evaluation of accountants by their aggregated client reporting
2.a.ii			**Medium Taxpayers**	
2.a.ii.1	✓	Descriptive	RDF Analysis: Top Taxpayers	Evaluation of only the top medium taxpayers (for example, ~1,000)
2.a.ii.2		Descriptive	RDF Analysis: Customs Brokers	Evaluation of brokers by their aggregated client reporting
2.a.iii			**Small Taxpayers**	
2.a.iii.1	✓	Descriptive	Threshold Analysis: Bunching at Notches	Analysis of all key thresholds (for example, VAT)
2.b			**Sectors and Industries**	
2.b.1	✓	Descriptive	RDF Analysis: Top Taxpayers	Evaluation of only the top industry taxpayers (for example, ~1,000)
2.b.2		Inferential	FARI Model: Mining	Analysis of mining resources for inclusion in planning
2.b.3		Inferential	FARI Model: Petroleum	Analysis of petroleum resources for inclusion in planning
2.c			**Category Risks**	
2.c.1		Hybrid	Specialized Research on Key Topics	For example, profit shifting, transfer pricing, VAT fraud
2.d			**Individuals**	
2.d.1	✓	Descriptive	HWI Criteria Review	Provision of options for refinement to existing criteria
2.d.2	✓	Descriptive	RDF Analysis: All HWI Taxpayers	Evaluation of all HWI taxpayers in a single analysis

Source: IMF staff.
Note: FARI = fiscal analysis of resource industries; HWI = high-wealth individual; RDF = risk differentiation framework; VAT = value-added tax.

Table 3. Illustrative Analytics: Active Prevention

	Essential?	Analytics	Output or Service	Remarks
1			**Compliance Forecasting**	
1.1		Predictive	Future Registration: Predict Nonregistration	Who will need to be registered, with revenue implications
1.2		Predictive	Future Filing: Predict Late and Nonfilers	Who will file late, not file, or self-finalize after the deadline
1.3		Predictive	Future Payment: Predict Late and Nonpayers	Who will pay late or not pay but has capacity to do so
1.4		Predictive	Future Payment: Predict Response by Type of Intervention	For example: SMS, call center outreach, certified mail, or visit
1.5		Predictive	Future Reporting: Predict Underreporters (Likelihood)	Who will be noncompliant by tax type (output a probability)
1.6		Predictive	Future Reporting: Predict Size of Adjustments (Consequence)	What will be the likely audit result (output a discrete value)
1.7		Predictive	Future Enforcement: Predict Taxpayer Response	Who will be likely to object to amended assessments
1.8		Predictive	Segments and Issues: Predict Aggregate, Future Behavior	Will compliance change and what will the revenue impact be
2			**Targeted Customer Outreach**	
2.1		Predictive	Predict Willingness to Comply (Likelihood)	Based on compliance history (taxpayer and related entities)
2.2		Predictive	Predict Revenue from Outreach (Consequence)	Based on prior outreach and revenue results by channel
2.a			**Next Best Actions**	
2.a.1		Predictive	Options for Customer Outreach	Noting channels, timing, and likely outcome
2.a.2		Predictive	Options for Administrative Actions	Noting the sequence of actions most likely to resolve a matter
3			**Embedded Analytics**	
3.1		Hybrid	Electronic Clearance Certificates: Risk Assess Requests	Using embedded logic or referring to risk profiles elsewhere
3.2		Predictive	Taxpayer Services: Train Chatbots for Tax Matters	Handling routing to staff and navigation of a knowledgebase
3.3		Predictive	E-Filing: Pre-lodgment Reporting Suggestions	Examining keyed values within comparable peer groups
3.4		Predictive	E-Filing: Pre-lodgment Risk Analysis	Examining declared values prior to final taxpayer submission
3.5	✓	Hybrid	Refunds: Evaluate and Risk Assess Requests	Facilitating automatic payment or recommending controls
3.6		Predictive	Financial Facilities: Evaluate and Risk Assess Requests	Indicating whether to grant a financial facility to a taxpayer
4			**Real-Time Analytics**	
4.1		Predictive	B2B & B2C Invoices: Risk Review for Handling Transactions	Considering the entities involved and nature of the invoice
4.2		Hybrid	G2B & G2C Invoices: Risk Review for Handling Transactions	Considering the nature of the invoice and patterns of behavior
5			**Third-Party Cooperation**	
5.1		Hybrid	Electronic Validation: Registration and Certificates	Confirming registration and compliance to third parties

Source: IMF staff.
Note: B2B = business to business; B2C = business to consumer; G2B = government to business; G2G = government to government; SMS = short message service.

Table 4. Illustrative Analytics: Monitoring and Detection

	Essential?	Analytics	Output or Service	Remarks
1			**Core Compliance Monitoring**	
1.1	✓	Descriptive	Current Registration: Unregistered Taxpayers	Relying on third-party data (for example, a business register)
1.2	✓	Descriptive	Current Registration: Required Tax Type Registrations	Focusing on VAT and other important thresholds
1.3	✓	Descriptive	Current Registration: Inactive Registrations with Activity	Relying on current third-party data (for example, public utilities)
1.4	✓	Descriptive	Current Filing: On Time, Late, and Nonfiling	Summary and disaggregation (for example, by taxpayer and tax type)
1.5	✓	Descriptive	Current Payment: On Time, Late, and Nonpayment	Summary and disaggregation (for example, by taxpayer and tax type)
1.6	✓	Hybrid	Current Reporting: Random Selection of Cases for Audit	For monitoring levels of compliance identifying current issues
1.7	✓	Descriptive	Current Arrears: New, Collectable, and Uncollectable Debt	Summary and disaggregation (for example, by taxpayer and tax type)
1.8	✓	Descriptive	Current Credit: Balances, Growth, and Unclaimed Refunds	Summary and disaggregation (for example, by taxpayer and tax type)
1.9		Descriptive	Current Voluntary Compliance: Recidivism Rate	The percentage of taxpayers repeating noncompliant behaviors
2			**Automated Underreporting Detection**	
2.1		Hybrid	Automated Data Matching: Employers	Comparing declared values to employee amounts withheld
2.2		Hybrid	Automated Data Matching: Other Domestic & International	Comparing internal and external data (for example, CBC reports)
3			**Exploratory Data Analysis**	
3.1	✓	Descriptive	Organization of Data Sets	Determination of elements, size, and shape prior to modeling
3.2	✓	Descriptive	Identification of Relationships	Examination of features and variables for possible correlations
3.3	✓	Descriptive	Visualization and Summary	Use of scatterplots, histograms, charts, and supporting statistics
3.4	✓	Hybrid	Outlier and Anomaly Detection	For improved modeling and identification of noncompliance
4			**Profiling Analytics**	
4.1	✓	Descriptive	360° View: Internal Data	Staging data and metrics for presentation (for example, declarations)
4.2	✓	Descriptive	360° View: External Data	Ensuring consistency of formatting with the internal view
4.3	✓	Descriptive	360° View: Benchmark Variances	Precalculating variances from peers and reference data sets
4.4	✓	Descriptive	360° View: Taxpayer Metrics	Sourced from a precalculated catalog of taxpayer metrics
4.5	✓	Descriptive	360° View: Risk Reviews	Including a summary perspective on risk from manual reviews
4.6	✓	Hybrid	360° View: Risk Profile	Integrating results of risk analysis and suggesting next actions
5			**Third-Party Cooperation**	
5.1	✓	Hybrid	Ad Hoc Analysis for Risk Identification	Analysis of third-party data outside of automated processes

Source: IMF staff.
Note: CBC = country by country; VAT = value-added tax.

Table 5. Illustrative Analytics: Enforcement

	Essential?	Analytics	Output or Service	Remarks
1			**Compliance Case Selection**	
1.1	✓	Predictive	Current Registration: Predict Forced Registration Priorities	Case Type: Registration: Estimate revenue implications
1.2		Predictive	Current Registration: Predict Fraudulent Registrations	Case Type: Visitation: Prioritize new VAT registrations
1.3		Predictive	Current Inactivity: Predict Likelihood of Actual Inactivity	Case Type: Deregistration: Account for definition of inactivity
1.4		Predictive	Historic Reporting: Predict Adjustments from Data Mismatch	Case Type: Verification: Automated detection (data mismatch)
1.5	✓	Predictive	Historic Reporting: Predict Other Underreporting Priorities	Case Type: Audit: Estimate likelihood and consequence
1.6	✓	Predictive	Historic Payment: Predict Likely Collection Outcomes	Case Type: Collection: Prioritize new and expiring debt
1.7	✓	Predictive	Historic Reporting: Predict Fraudulent Refund Claims	Case Type: Refund: Segment predictions by size of claim
1.8		Hybrid	Casework Templates	Population of statistics in the tools used by case officers
1.a			**Next Best Actions**	
1.a.1	✓	Prescriptive	Options for Case Configuration: Scope of Action	Suggesting parameters (for example, comprehensive or issue audit)
1.a.2	✓	Hybrid	Options for Case Configuration: Resource Assignments	Suggesting the size and capabilities of staff needed for a case
1.a.3		Prescriptive	Options for Case Configuration: Likely Objections	Suggesting approaches to avoid taxpayer objections (appeals)
2			**Case Selection Performance**	
2.1		Hybrid	Evaluation of Prediction Power	Making use of confusion matrices and ROC curves
2.2		Descriptive	Compliance Outcomes Analysis	Measuring changes in behavior attributed to casework
2.3	✓	Descriptive	Revenue Outcomes Analysis	Measuring changes in revenue attributed to casework
2.4	✓	Hybrid	Options for Improving Selection	For incorporation into future selection logic and analytics
3			**Assessment and Notifications**	
3.1		Predictive	Assessment: Predict Default Assessment Values	Producing estimations within comparable peer groups
3.2		Predictive	Assessment: Predict Default Assessment Objections	Including identification of thresholds for change in behavior
3.3		Prescriptive	Notification: Options for Assessment Notifications	Including identification of channels and variations in messaging
3.4		Prescriptive	Notification: Options for Collection Notifications	Accounting for options in first and subsequent notifications
3.5		Prescriptive	Notification: Options for Other Notifications	Considering the desired outcomes from issuance of the notice
4			**Third-Party Cooperation**	
4.1		Hybrid	Case Selection: Joint Audit with Customs	Making use of specialized criteria and customs intelligence
4.2		Descriptive	Certification for Sharing: Debt Enforcement List	Confirming inclusion of taxpayers with collectible arrears

Source: IMF staff.
Note: ROC = receiver operating characteristic; VAT = value-added tax.

As a starter kit, this note is complemented by the additional resources and background materials provided in the following six annexes. Intended to support practical exploration of the analytics outputs and services presented in Tables 2 to 5, the annexes are the following:

- **Annex 1. Accessing the Companion Toolkit:** Instructions for accessing and making use of the note's companion toolkit, which includes a selection of developed tools and templates configured to make use of synthetic (artificial) data sets, all intended primarily for educational purposes.
- **Annex 2. Critical Domain Knowledge:** Insights into (1) general techniques using data to identify taxpayers of interest, (2) important statistical distributions for analyzing tax data, (3) the use of exploratory data analysis to detect anomalies and outliers in data, and (4) emerging developments in the use of AI.
- **Annex 3. Addressing Key Challenges: Data Quality and Staff Capacity:** General approaches to deal with data quality and staff capacity challenges that have been observed as productive in tax administrations in different regions of the world.
- **Annex 4. Selected Topic: Compliance Planning:** A short conceptual overview of compliance planning and introduction of a risk differentiation framework (RDF), useful for performing a macrostrategic risk analysis to inform compliance activities (strategies and plans). Includes an illustrative output from an Excel-based RDF template provided in the note's toolkit.
- **Annex 5. Selected Topic: Taxpayer Profiling:** To highlight the importance of manually profiling taxpayers (particularly large or otherwise complex), an approach to developing digital tools support for risk reviews is presented. Includes an illustrative output from an Excel-based profiling template provided in the note's toolkit.
- **Annex 6. Selected Topic: Audit Case Selection:** Contemporary concepts influencing the use of analytics for audit case selection are presented, highlighting the special importance of using both risk rules created by domain experts and predictive analytics using algorithms. Includes an illustrative output from a predictive analytics workflow provided in the note's toolkit.

Annex 1. Accessing the Companion Toolkit

The toolkit accompanying this note contains configurable examples of analytics to support three topics that are critically essential in most tax administrations. Downloadable from the Fiscal Affairs Department Revenue Portal, the toolkit includes operationally tested templates that have been configured to make use of synthetic (artificial) data sets. Upon evaluation, the tax administrations seeking to further explore their use may wish to reconfigure the templates to use their own internal data. The toolkit will be updated and expanded over time.

The initial examples provided are expected to include:

- ***An RDF template.*** A Microsoft Excel workbook that contains a completed RDF analysis in the form of a visual scatterplot with summary quadrant statistics, intended to inform compliance planning around large taxpayers.
- ***A template for profiling large taxpayers.*** For the relatively small numbers of high-risk taxpayers identified in a large taxpayer RDF analysis, a Microsoft Excel workbook for use as a manual profiling template is provided to assist with their further evaluation.
- ***An advanced analytics package for case selection.*** For small and medium taxpayers, a workflow using the KNIME Analytics Platform and machine learning is provided for audit case selection, along with logic for measuring prediction power.

Toolkit Requirements

- A current version of Microsoft Office and Excel[5]
- A current version of the KNIME Analytics Platform[6]

Instructions for Accessing the Toolkit

Step 1: Browse to the IMF Fiscal Affairs Department's Revenue Portal (https://www.imf.org/revenueportal).

Step 2: Navigate to the "Analytical and Learning Resources" section of the portal.

Step 3: Scroll down to find "Essential Analytics for CRM."

Step 4: Download the tools and instructions.

At the time of publication, the following URL provides direct access to the page and section of the portal where the toolkit can be downloaded: https://www.imf.org/en/Topics/fiscal-policies/Revenue-Portal/Analytical-and-Learning-Resources#analytical

Should assistance be required, please email revenueportal@imf.org.

5 See Microsoft Excel (https://www.microsoft.com/en-us/microsoft-365/excel).

6 See KNIME (https://www.knime.com).

Annex 2. Critical Domain Knowledge

For the analytics that support CRM to mature, technical proficiency with statistics and data science must be complemented by knowledge of four critically important topics. These are techniques for identifying taxpayers of interest, the use of statistical distributions to analyze tax data, and methods for identifying outliers. In the coming years, advancements in the use of AI will improve the accessibility of these topics—particularly through AI's use of natural language processing (NLP) techniques. The following text provides essential insights.

Techniques for Identifying Taxpayers of Interest

Most uses of analytics for risk analysis can be grouped under one of five broad categories that describe the techniques applied. In practice, these categories comprise a toolkit of options for data analysts and, increasingly, for inclusion as part of automated IT services that make use of embedded analytics. Regardless of specific application, the purpose of all the techniques is to identify taxpayers of interest (both compliant or, potentially, noncompliant) by:

- **Matching**. Matching data and joining data sets, including qualitative intelligence. For some tasks, such as identifying nonregistrants and nonfilers, data matching is commonly used. This could include matching business records (the business license database against the tax database or telephone book database against the tax database). Matching can be either "hard" (using identification numbers) or "soft" (using names). Results from data matching often require prioritization (for example, estimating amounts of tax involved).
- **Measuring**. Determining and calculating ratios and values based on expert judgment and rules. Another common approach used by tax administrations is to identify attributes of interest in a data set and measure their value. These attributes have usually been identified as important for compliance actions by staff with extensive experience in the matter. For example, refunds over a certain value and out of pattern are often flagged for compliance review. Outputs from measuring are often a basis for "rules engines."
- **Mining**. Identifying clusters or segments of interest using various techniques. Usually these are applied to large, complex data sets to identify new groups that may be of interest and those who are outliers from the group. Techniques include k-nearest neighbor (for example, to find outliers in a group or alert taxpayers that values reported on their returns are outside of industry benchmarks) and text mining (for example, to analyze vast amounts of emails and identify those of importance to tax planning arrangements).
- **Modeling**. Building predictive models to classify likelihood and consequences of potential noncompliance. Often relying on machine learning, the results of past cases are analyzed using one or more algorithms to predict the classification status of a taxpayer (compliant or noncompliant) along with the probabilities that the predictions are correct. The use of such approaches is increasingly important as data grow in volume and complexity. Given sufficient data, these techniques often outperform the use of "risk rules."
- **Mapping**. Using specialized software for network analysis to connect and follow data chains. Among a network, these techniques help to identify nodes of interest. Many use cases for tax administration exist. As examples, they include mapping of value-added tax transactions data from business to business to identify missing trader fraud, and use of social network analysis (not to be confused with social media) to identify hidden organizations committing fraud.

Useful Statistical Distributions for Analyzing Tax Data

The correct selection and use of distributions is critical for producing useful analysis. This is especially important when making use of inferential and predictive modeling techniques. While the most accurate distributions to model tax risks will always depend on the specific data and research questions being posed, established concepts and practices for their use do exist:

- **Log-normal distribution**. Commonly used to model data that results from accumulation of many small risk events. While for a log-normal distribution the data cannot be negative, this can be changed by a simple transposition to shift the "zero point" to the largest negative value. Things that multiply usually produce a log-normal distribution. Risk, by definition, is likelihood times consequence.
- **Pareto distribution**. Commonly used to model risks that follow a power-law distribution, which means that a small number of events are responsible for a large proportion of the risk. As such, it is often used to model data such as tax evasion and fraud, and the size of adjustments. It can also be used to model income or wealth distributions or tax avoidance, particularly by large taxpayers.
- **Weibull distribution**. Often used to model time-to-event data. For example, the time until a taxpayer becomes noncompliant or to model the time to detect and correct tax evasion. Both perspectives can be useful when exploring potential outcomes of compliance strategies and plans, and the use of resources.
- **Beta distribution**. A flexible probability distribution that can model data that is bounded between two values, such as a proportion of taxpayers who are noncompliant or the proportion of tax revenue that is lost due to noncompliance.
- **Multinomial distribution**. Can be used to model categorical data, such as the compliance status of taxpayers including compliant, partially compliant (under a threshold), or noncompliant (over a threshold). Population estimates, based on samples, can then be made using Bayesian or Maximum Likelihood Estimators.
- **Poisson distribution**. Used to estimate the probability of a certain number of noncompliant taxpayers in a population. For example, as this distribution can model the number of events that occur in a fixed interval of time, it can be used to predict the number of audits or the number of taxpayers that are non-compliant in a given period.
- **Exponential distribution**. Another distribution to model time-to-event data, such as the time until a taxpayer becomes noncompliant (the "hazard rate" of noncompliance over time), differing from a Weibull distribution in its use of probabilities.
- **Gaussian or normal distribution**. From basic statistics courses, this is likely the most familiar distribution to most staff in tax administration. However, normal distributions do not represent the distribution of most taxpayer data attributes as they reflect underlying economic factors, which are in turn typically Pareto distributed.

Identifying Outliers in Exploratory Data Analysis

A major aspect of analytics for CRM includes the concept of exploratory data analysis (EDA) (also noted in Table 4). The purpose of exploratory data analysis is to identify, understand, and summarize the main characteristics of data sets, often using visual methods. It is an important precursor to making use of statistics for any CRM purposes. In practical terms, before drawing conclusions from specific data, that data should be well-understood first. To do so, exploratory data analysis makes use of different techniques to summarize its features and explore its suitability. This frequently involves identifying patterns through correlation analysis, dimensionality reduction, data clustering, and other methods.

As part of exploratory data analysis, the detection of anomalies and outliers is essential. Their presence can inform decisions regarding the suitability of data (and its potential correction), while also identifying noncompliance. Important for many aspects of CRM, the methods for their detection include:

- **Z-score**. This method calculates the distance (in standard deviations) of each data point from the mean of the data set. Data points that are more than a certain number of standard deviations away from the mean are considered outliers.
- **Interquartile range**. This method calculates the difference between the 75th and 25th percentiles of the dataset. Data points that fall outside of the range defined by the interquartile range (Q3 + 1.5 interquartile range) and (Q1 – 1.5 interquartile range) are considered outliers. Other ranges can also be used, and it is important to explore the impact of changing parameters.
- **Mahalanobis distance**. This method calculates the distance of each data point from the mean of the data set, considering the covariance of the data set. Data points that are farther away from the mean than a certain threshold are considered outliers.
- **Local outlier factor**. This method is density-based and identifies outliers by comparing the density of a data point to the density of its neighbors. Data points that have a significantly lower density are considered outliers.
- **Elliptic envelope**. This is a method based on a multivariate Gaussian distribution. It assumes the data follows a Gaussian distribution and detects the data points that are far away from the center. It requires normalized data transformation.
- **Clustering methods**. Data points that are not well represented in any cluster or that are far away from the centroid of any cluster are considered outliers. Several clustering techniques exist, including k-means clustering, hierarchical clustering, density-based spatial clustering of applications with noise, and Isolation Forest.

Understanding AI for NLP

Building from decades of research, machine learning algorithms have enabled computers to understand and process natural language commands. This concept is generally considered a unique branch of AI referred to as NLP. In its modern form, NLP makes use of algorithms designed for "reinforcement learning" on large-scale neural networks that are trained on vast amounts of textual data (large language models). In tax administration, NLP's capabilities have great potential.

For taxpayer services, adoption of the technology is well underway. This is occurring most prominently using chatbots powered by NLP algorithms. These bots provide automated and personalized assistance, answering questions and delivering guidance on tax-related issues. Similar adoption is rapidly occurring with call center technology.

For CRM support, practical applications for the technology exist and are already being used in limited capacity (for example, to inform development of database queries and other programming for analytics). As it matures, NLP and the tools making use of it are likely to have transformational implications to risk and intelligence work due to their ability to:

- **Reason across knowledge domains**. NLP has the capability to understand natural language questions posed that span vast areas of specialized study (tax, law, and accounting).
- **Dynamically generate programming**. To the extent that questions rely on inference or prediction from data, NLP is capable of producing data queries and scripts for analytics.
- **Apply logic to large data sets**. Making use of the querying requests generated, the tools supporting NLP can themselves retrieve answers to questions from data.

- **Interpret the results**. Beyond retrieving data, the results can be interpreted in the context that the original question was posed in and, using NLP, return an appropriate answer.

While the use of NLP with bots in tax administrations is primarily limited to service aspects at present, their tailoring to support CRM risk analysts is likely. In practice, bots powered by NLP may eventually facilitate accurate and rapid responses to direct questions, such as:

- What are the primary tax compliance risks in my jurisdiction?
- How are my compliance strategies performing?
- Which taxpayers are most likely to be underreporting?

Not without costs, the use of NLP requires navigating the selection of the AI bots and tools used, their training and fine tuning, the appropriate formulation of questions, and quality assurance of responses. While at present these are considerable hurdles, progress in the use of AI is rapidly occurring and the tools that make use of it are quickly becoming more accessible. This progress is benefiting from several initiatives already underway for the purposes of advancing its use, specifically in tax administration.

Annex 3. Addressing Key Challenges: Data Quality and Staff Capacity

Internationally, two recurring challenges are especially difficult for many administrations to deal with: data quality and the capacity of staff. While no easy solutions to either tend to exist, many serious initiatives to improve analytics capabilities for CRM have been observed in recent years. Among the lessons learned from these are several key concepts and practices that, across different regions in the world, are known to be more productive than others.

Data Quality

A reflection of institutional maturity, data quality issues are best addressed through clear definitions and progressive improvement strategies. In their entirety, tax administrations are information-driven organizations. From core tax operations to strategy and planning, the use of data is pervasive. Conceptually, the data collected should result from implementation of defined business processes which are, themselves, subject to quality measurement and control. In scenarios where processes are not defined and fully implemented, data quality is often problematic. Regardless of the reasons why—for example, poor IT systems, staff skills and training—poor process management is usually a consequence of institutional weakness.

The importance of data quality extends beyond tax administration itself and across both government and business. This is particularly true in areas where third-party data is relied on for operational purposes (payments processing) and key aspects of risk analysis or risk-based automation (automated detection of underreporting).

In the realm of analytics for CRM, "data quality" may be defined as the degree to which data is suited to support risk analysis. Accordingly, data quality can be measured. To illustrate, Annex Box 3.1 presents nine principled elements of a data quality framework. In administrations with strong data management practices, data sets are formally registered and evaluated against the types of elements described. Upon doing so, an overall quality rating is determined (for example, low, medium, high), along with separate ratings that

ANNEX BOX 3.1. Elements of a Data Quality Framework

Framework Element	Data Quality Is the Degree to Which …
Accuracy	data correctly represents actual values.
Completeness	the required data is actually present in a data set.
Compliance	policies, controls, and audit trails are in place for a data set.
Consistency	values in a data set are consistent with values elsewhere.
Precision	the level of detail in elements of a data set is sufficient.
Timeliness	data is recent for the time period that a data set covers.
Traceability	the history of changes to data in a data set is maintained.
Uniqueness	data is duplicated in a data set or stored redundantly.
Validity	data is presented in the correct format and within constraints.

Source: IMF staff.

describe a given data set's utility in risk analysis. Despite heavy investments in digitalization, many administrations today have yet to implement these practices.

Because data quality depends on many facets of administration, strategies that focus on core fundamentals first tend to be most successful. Although the technical aspects of data quality issues are important, improvements are an institutional responsibility extending beyond the sole responsibility of an IT department. Accordingly, the engagement of technical specialists and the owners of the respective processes, systems, and data involved is usually required. Through a formal strategy endorsed by leadership, progressive activities could include, in illustrative order:

- ***Introducing basic data management practices.*** These include taking an inventory of data assets that currently exist, classifying their importance and sensitivity, defining a data catalog and data dictionaries, and developing basic data lifecycle strategies.
- ***Identifying and evaluating the basic quality of core data sets.*** A core data set is either essential reference data (tax types, tax periods, tax forms) or data required for core operations (the tax register, assessments, accounts, returns, payments).
- ***Determining root causes of core data set quality issues.*** In doing so, it is important to isolate IT issues from the underlying business processes and the staff and taxpayers that make use of them (whether consistently or not).
- ***Implementing a targeted improvement plan.*** Focused narrowly on core data sets only, the plan should prioritize interventions (for example, revisions of procedures for the use of IT systems) that emphasize improving the quality of current data over historical.
- ***Extending evaluation to noncore data sets.*** This should include all other data sets, each of which is prioritized for evaluation, focusing on third-party data, casework, correspondence, and high-volume electronic reports (for example, invoices).
- ***Implementing an institutional improvement plan.*** Comprised by elements that may represent significant reforms, appropriate actions could include implementation of quality assurance systems, new procedures, training, and acquisition of tools.
- ***Dealing with external partners and third parties.*** Recognizing that entities supplying data may have their own challenges, a formal process for certifying sources, data definitions, and the actual data supplied should include appropriate quality controls.

While improvements are being implemented, analysts should set clear boundaries delineating proven, high-quality data sets from others and lean on human intelligence. In jurisdictions where data is limited or known to be of poor quality, the approach to its use must be selective to avoid negative outcomes. While waiting for data quality strategies to take effect, analysts should rely on basic descriptive statistics to examine filing and payment. For underreporting, the use of expert rules to make predictions can be appropriate in lieu of more advanced techniques. Basic data matching can also be very productive. As higher-quality data sets become available, the full range of analytics capabilities should be considered and explored for introduction.

Staff Capacity

As analytics are multidisciplinary, deficiencies in capacity are best addressed through a combination of recruitment, training, and consolidation of resources. In an ideal setting, CRM should be supported by trained data scientists. As Figure 2 illustrates, data science draws on expertise from three domains of knowledge: computer science, business, and statistics. For a data scientist to be effective for CRM, proficiency in the two mathematically oriented areas (computer science and statistics) as well as knowledge of

tax administration is required. Although expertise in computer science and statistics may be transferable for external hires, a technical understanding of tax matters, tax administration operations, and CRM can take considerable time to develop.

While larger tax administrations are increasingly recruiting data scientists directly and retraining them on taxation, others tend to "grow their own." Both public and private sector organizations in general tend to view data science as an area of strategic importance. The value being placed on the skills required has, in effectively all regions of the world, created competition, moving compensation beyond what many tax administrations can offer. While challenging, this reality has been addressed with some success by using creative recruitment strategies and nontraditional incentives such as:

- ***Marketing the unique scale and volume of data under management.*** Communicating the reality that, in many jurisdictions, tax administrations are among the largest IT operators in a country and house data that may have unique intellectual appeal, particularly in context of intelligence and law enforcement activities.
- ***Marketing the presence of data scientists in key committee meetings.*** Demonstrating the value placed on the skills sought, inclusion in front office work (compliance committees) affirms that analysts will not be relegated to back office support.
- ***Soliciting transfers of staff having an IT education but working in non-IT roles.*** Identifying staff with education or training in the skills needed that are assigned to nontechnical roles and offering a transfer into a CRM analytics group.
- ***Recruiting from public sector agencies, universities, and think tanks.*** Targeting data scientists already in the public sector and academically oriented professionals (for example, university researchers) that may have interest in operational roles.

Regardless of the approach to recruitment taken, a target staff profile and careful thought are needed to promote capacity development. If the quality of data available is insufficient for the use of advanced analytics or CRM practices are not yet mature, starting small can be prudent. In some administrations, having any dedicated resource for CRM analytics is already a major step forward, and away from relying on IT staff for data and information. In these scenarios, only basic training is needed. In others, a full team of data scientists are required. To aid in thinking through scenarios, Annex Box 3.2 presents a range of staff profiles and potential activities to improve their capacity for analytics.

Organizational options also exist that can aid in building staff capacity—consolidation of units is one that is very promising. In all administrations, data analysis occurs in different areas, particularly among headquarters offices. This configuration, which is typical, requires that individuals within each unit separately build capabilities to work with data and manage its use. As a method of building capacity, discussions have emerged in recent years around the concept of consolidating (1) data management; (2) analytics; and (3) research functions into a single group. The benefits of the approach are potentially far reaching for focusing capacity development, and consolidation may be increasingly required for some administrations to scale up their work with analytics.

ANNEX BOX 3.2. Target Staff Profiles and Capacity Development Options

Emerging analyst. A power user or Microsoft Excel expert capable to perform sophisticated analysis in Excel or similar tools using a full range of formulas, macros, and scripting. May or may not have expert business knowledge. Does not have the ability to query databases or analyze large data sets without the support of IT staff. An individual with this profile might:

- Attend new officer and audit training to build knowledge of taxes
- Attend basic SQL training to develop skills for database querying
- Request direct access to databases to speed up exploratory analysis
- Shadow database administrators to learn the data inventory and models

Intermediate analyst. A data specialist with solid database querying skills and basic knowledge of advanced analytics. Has a full and complete understanding of all the major taxes administered, tax administration operations, and the data sets maintained. Is fully capable of producing statistics and reports and risk analysis without the support of IT staff. An individual with this profile might:

- Attend advanced SQL training to ensure full professional proficiency
- Attend advanced analytics training to improve the use of predictive methods
- Attend data visualization training to improve communication of results
- Participate in knowledge exchange events with peer tax administrations
- Collaborate with external agencies to expand insights into related domains
- Document data, tools, and techniques used to support team members

Qualified data scientist. An expert that has full command of computer science, the business of taxation, and statistics who understands the mathematics and logic behind the methods applied and when to use them. Specialized in predictive analytics, produces, tests, and demonstrates the effectiveness of risk analysis using large volumes of data. An individual with this profile might:

- Attend advanced taxation courses to help with the design of new analysis
- Subscribe to journals to maintain awareness of new concepts and methods
- Acquire technology and tools to introduce new, specialized capabilities

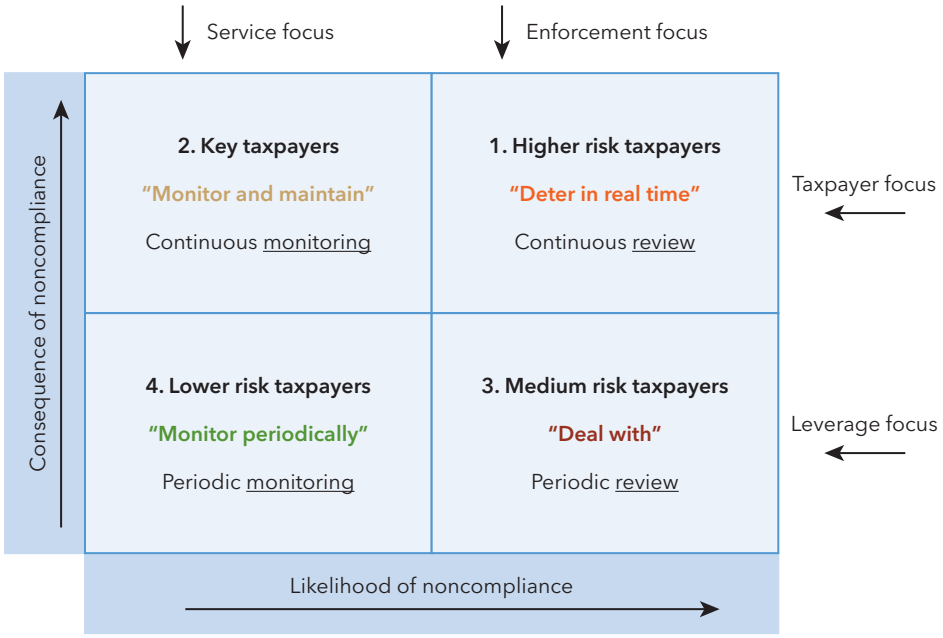
Source: IMF staff.
Note: IT = information technology; SQL = Structured Query Language.

Annex 4. Selected Topic: Compliance Planning

In this note, "compliance planning" is an umbrella term used to describe the formulation of treatment strategies and allocation of resources. With guidance from executive leadership, and informed by risk analysis and intelligence, analysts that support compliance planning select the risks to focus on, the resources to apply, how to apply resources, and where other follow-up may be needed. Their outputs include compliance plans or more robust investment in the form of strategies and campaigns. These activities are often carried out through formal committees.

As input into compliance planning, specialized macrostrategic concepts exist—the use of an RDF for large business is key among them. A contemporary RDF concept is presented here. Its logic is premised on the understanding that not all taxpayers pose the same level of and types of risk, and that resources should be allocated accordingly. In the logical form of a risk matrix, the framework uses analytics to identify higher-likelihood-of-concern taxpayers (often by using low effective tax rates) and focus appropriate compliance efforts on them, while providing lower-likelihood taxpayers with streamlined compliance improvement options (education, assurance, advice, other improved services). As illustrated in Annex Figure 4.1, these actions are organized into four quadrants, each described by a general compliance approach (for example, "Deter in real time").

Annex Figure 4.1. A Risk Differentiation Framework



Source: IMF staff.

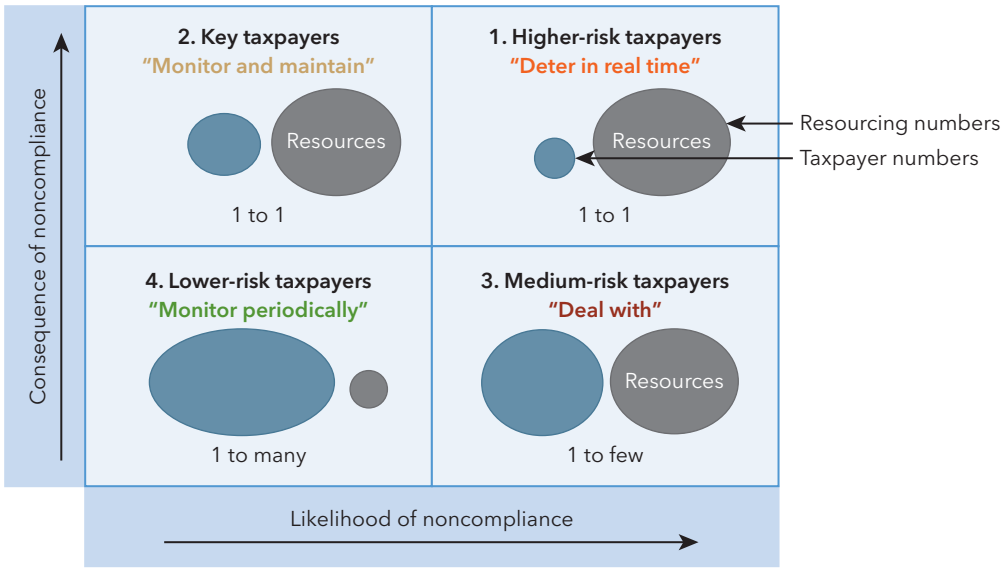
Increasingly adopted internationally, the use of an RDF represents a major evolution in risk practices developed over many decades. Risk differentiation itself is not a new concept. As an example, in the 1980s the Australian Taxation Office introduced SCORE, a ranking system that identified outliers in industries and occupations. Similarly, in the 1990s, the US Internal Revenue Service introduced a form of differentiation when it developed the Discriminant Index Function system to classify income tax returns into high-,

medium-, and low-risk categories for audit. These approaches were and, in many countries, continue as the foundation for risk management. As new CRM practices emerged in the mid-2000s, concepts of risk differentiation evolved by:

- ***Integrating likelihood and consequence.*** In doing so, accounting for both aspects of risk, which together are required to promote productive interventions (but remain absent in the risk analysis performed in many administrations today).
- ***Recognizing patterns of shared behavior.*** By allowing for the visual representation of results in a scatterplot (often visually displaying patterns of interest), and summary statistics for four quadrants, highlighting unique issues in each.
- ***Supporting forward-looking compliance planning.*** Because of its ability to visually aid in identifying patterns and differentiate among behavior (through the four quadrants), an RDF is especially useful for understanding and planning around groups of taxpayers.

When viewed from the perspective of resource planning, the full utility of an RDF concept becomes clear. Represented by the blue spheres in Annex Figure 4.2, a distribution of taxpayers in an RDF analysis should result in few in quadrant 1 (higher-risk taxpayers), more in 2 and 3 (key taxpayers and medium-risk taxpayers), and the majority in quadrant 4 (lower-risk taxpayers). Because of the different issues, likelihood, and consequence of risk in each quadrant, an efficient use of resources (primarily staff) often mirrors the allocation represented by the gray spheres. In quadrant 1, a one-to-one approach is appropriate, where risks are managed on an individual taxpayer basis. This contrasts with quadrant 4, where the focus is the group itself.

Annex Figure 4.2. Resource Allocation, by RDF Quadrant



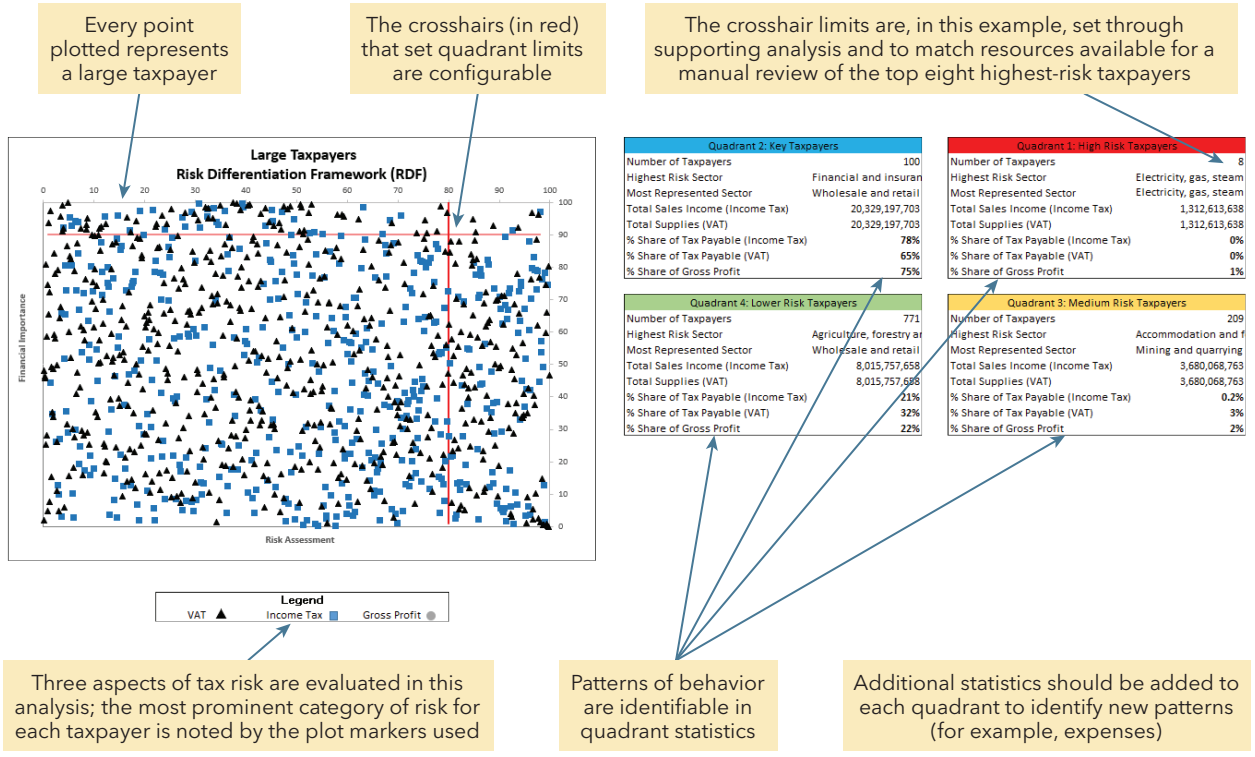
Source: IMF staff.
Note: RDF = risk differentiation framework.

In use, it is important to note that an RDF is intended as a framework for forward planning, not necessarily as a tool for audit case selection. Upon completing the initial analytics to produce an RDF analysis (provided for in the toolkit template), analysts should invest the time required to (1) understand the visual and statistical patterns presented; and (2) as part of the full CRM process, brainstorm through the selection

of the next best step. In some instances, audit may be appropriate. However, a wide range of tools exist that may, especially for large taxpayers, be more effective (for example, key client programs, advanced pricing agreements, rulings, referrals for legislation). Whatever actions are selected should be included in a compliance plan.

Because an RDF is primarily a planning tool for allocating resources to risks, the judgment of analysts is essential for its productive use. The framework can and should be configured to apply different metrics to explore different tax types separately, and in integrated analysis. Annex Figure 4.3 presents an example produced from the template included in this note's toolkit.

Annex Figure 4.3. Example Output from the RDF Toolkit Template



Source: IMF staff.
Note: RDF = risk differentiation framework; VAT = value-added tax.

Annex 5. Selected Topic: Taxpayer Profiling

Essential to support the work of risk analysts, automated profiling and digital tools are increasingly necessary to navigate large volumes of data and manage risk reviews. For large taxpayers, an ideal set of capabilities in most administrations would enable an analyst to (1) examine the large segment through an RDF analysis; (2) interactively explore the detailed profile of each taxpayer in the RDF scatterplot; and (3) select a small number of taxpayers for further, manual risk review (for example, those in an RDF's quadrant 1: higher-risk taxpayers). For those selected for a manual review, digital tools should then allow cataloging of findings for future study and refinement over time. As first illustrated in Figure 3 in the main text and described further in Annex Box 5.1, automated profiling builds from layers of data infrastructure to provide "360° Views" of a taxpayer.

Among the 360° Views that comprise a taxpayer profile, the two relating to risk ("risk reviews" and "risk profile") tend to be underdeveloped or absent. Consistently observed as a limitation in tax administrations globally, the underlying reasons vary from weakness in the use of automation generally to difficulty in developing requirements for the needed IT and digital tools support. Where automation does exist for risk management, it is often limited to resource planning and very basic logic for case selection. Accordingly, while the analytics needed to form a 360° View may be in place and their automation possible, digital tools for active management of a taxpayer profile by an analyst are frequently altogether missing.

A good first step toward improving digital tools for profiling is adoption of a risk review template for large taxpayers—for that, BISEP (business, industry, sociological, economic, psychological) can be helpful. Requiring analysis extending far beyond analytics, large taxpayers are complex. While different methods for their evaluation exist, applications of the BISEP concept are often used. As illustrated in Annex Figure 5.1, BISEP is an acronym comprised of five factors that influence taxpayer behavior (starting from "Business," reading clockwise). A full evaluation of a large taxpayer is itself a research project, and BISEP is actively used by both leading and developing administrations for that purpose. In some, BISEP has been extended to include additional factors for systems of compliance and technology and data.

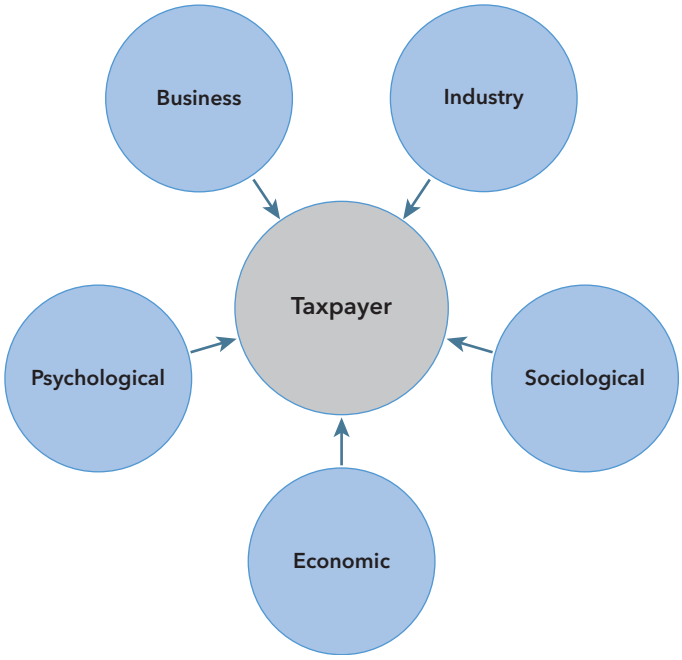
ANNEX BOX 5.1. Authoritative 360° Views of a Taxpayer

Many tax administrations today invest heavily in developing data infrastructure. As the volume of data under management continues to grow, specialized technology, tools, and data modeling techniques are increasingly required. This dynamic is particularly strong in some areas, the concept of "360° Views" among them. Because of a frequent need to quickly retrieve or navigate a complete set of taxpayer records, a collection of supporting data models is often staged on a data warehouse. Although technical approaches vary, conceptually discrete models are often designed explicitly for rapidly retrieving single, authoritative views of:

- **360° View: Internal data**. All data internally generated or captured by the administration itself. Displayed in a structured format (not raw data), data sets span the entirety of records, without restriction (from registration to filing, payment, accounting, enforcement, and a wide range of other subjects). Much of the data is accompanied by visualizations.
- **360° View: External data**. All data generated by external third parties but cataloged and available to the administration. Including data from both external government agencies and partners. External data also include data from key data sets relied on to support risk analysis.
- **360° View: Benchmark variances**. Precalculated variances from comparable peers and reference data sets. Presenting variances visually across the relevant tax periods, primarily by sector, industry, and activity classifications. Often includes trade variances, as well as variances from economic indicators (for example, growth, inflation).
- **360° View: Taxpayer metrics**. Precalculated statistics and measurements, including key tax and financial ratios. Often making use of dimensional modeling techniques by precalculating values, authoritative logic can be consolidated in one location, promoting the integrity of results while improving the performance of underlying technology resources.
- **360° View: Risk reviews**. Results of the manual evaluation of taxpayer risks by analysts. Presents data that is primarily qualitative in nature, reflecting judgment and experience captured from multiple analysts over time. Depending on the capabilities available, also presents next steps for further review and open action items.
- **360° View: Risk profile**. Results from the risk analytics automatically applied, along with analyst additions and remarks. Prioritizing the presentation of results from expert rules, predictive analytics, and data matching, along with suggestions for next steps. Depending on the capabilities available, may present manual adjustments in automated risk scoring by analysts.

Source: IMF staff.

Annex Figure 5.1. Factors That Influence Taxpayer Behavior: BISEP



BISEP Factor	Elements of Evaluation
(B) Business. The nature of the taxpayer itself, whether a single entity or part of an affiliated group.	The extent of business activities, legal structure of the entities involved, location, industry, and capital structure.
(I) Industry. The nature of the industry that the taxpayer operates in.	Geography, size, participants in the industry and the nature of competition, the role of associations, cost structures, established industry norms and regulations.
(S) Sociological. The general perception of the taxpayer by the primary communities that it operates in.	Community norms and expectations, degrees of self-regulation and the reputation of the taxpayer, its executives and management.
(E) Economic. The economic environment that the taxpayer currently operates in and the future conditions that it anticipates.	Domestic or international growth, foreign trade, taxes, interest rates, currency fluctuations, fiscal and monetary reforms.
(P) Psychological. The mindset, disposition, and philosophy of the key individuals and influencers that drive decision making.	Objectives of decision makers, approaches and drivers of risk, governance frameworks, and relationship with tax and other authorities.

Source: The BISEP concept was developed by Dr. Valerie Braithwaite in the early 1990s.
Note: BISEP = business, industry, sociological, economic, psychological.

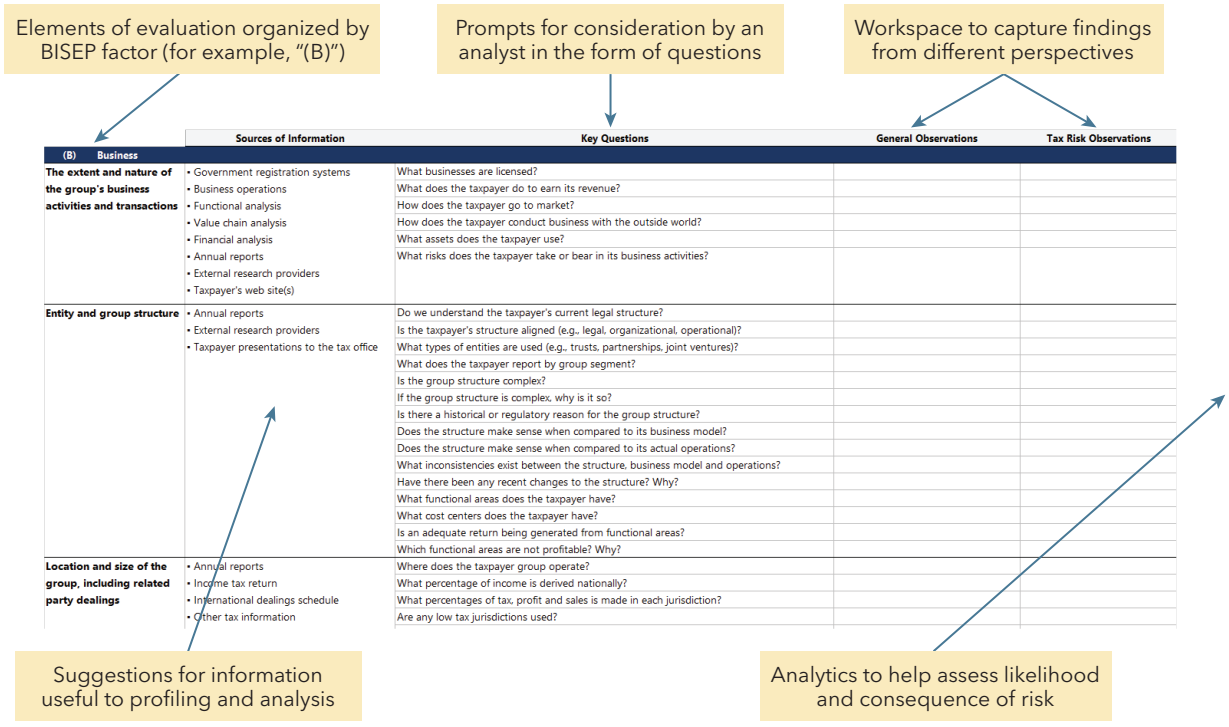
Once a risk review template has been adopted, options for automation and development of digital tools can be easier to explore. Although a need for automated analytics is generally well understood, the additional need for tools to manage profiles is less so. By adopting and using a template, the qualitative aspects of risk management and their value become more obvious. This can help to advance conversations around IT development.

For IT development, the options considered typically include some combination of:

- ***Acquisition or enhancement of a risk module.*** Extending the concept of modules that may be used for other aspects of administration (for example, registration) through existing IT platforms which, while useful for some requirements, may be limiting technologically.
- ***Acquisition of a standalone risk system.*** Developing a new IT system from the ground up to support a wide range of requirements for risk-related functionality, including the need of risk analysts for automation of analytics and tools support.
- ***Integration of commercial products.*** Making use of specialized technology platforms and tools for data infrastructure (for example, data lake, data warehouse), statistical analysis software, profiling, risk-based workflow, and case management.

To help operationalize risk reviews and advance capacity discussions, a template has been included in this note's toolkit. Annex Figure 5.2 presents the key aspects of the template, highlighting its use of the BISEP model to structure the evaluation of a large taxpayer.

Annex Figure 5.2. Excerpt from the Large Taxpayer Profiling Toolkit Template



(B) Business	Sources of Information	Key Questions	General Observations	Tax Risk Observations
The extent and nature of the group's business activities and transactions	• Government registration systems • Business operations • Functional analysis • Value chain analysis • Financial analysis • Annual reports • External research providers • Taxpayer's web site(s)	What businesses are licensed? What does the taxpayer do to earn its revenue? How does the taxpayer go to market? How does the taxpayer conduct business with the outside world? What assets does the taxpayer use? What risks does the taxpayer take or bear in its business activities?		
Entity and group structure	• Annual reports • External research providers • Taxpayer presentations to the tax office	Do we understand the taxpayer's current legal structure? Is the taxpayer's structure aligned (e.g., legal, organizational, operational)? What types of entities are used (e.g., trusts, partnerships, joint ventures)? What does the taxpayer report by group segment? Is the group structure complex? If the group structure is complex, why is it so? Is there a historical or regulatory reason for the group structure? Does the structure make sense when compared to its business model? Does the structure make sense when compared to its actual operations? What inconsistencies exist between the structure, business model and operations? Have there been any recent changes to the structure? Why? What functional areas does the taxpayer have? What cost centers does the taxpayer have? Is an adequate return being generated from functional areas? Which functional areas are not profitable? Why?		
Location and size of the group, including related party dealings	• Annual reports • Income tax return • International dealings schedule • Other tax information	Where does the taxpayer group operate? What percentage of income is derived nationally? What percentages of tax, profit and sales is made in each jurisdiction? Are any low tax jurisdictions used?		

Source: IMF staff.
Note: BISEP = business, industry, sociological, economic, psychological.

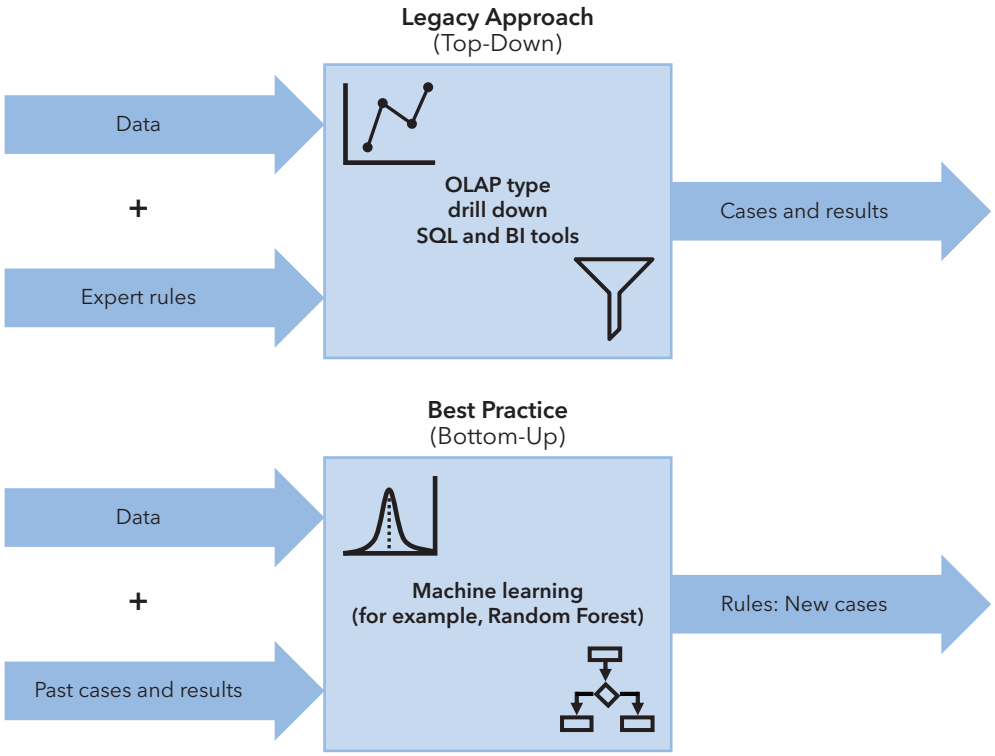
Annex 6. Selected Topic: Audit Case Selection

The selection of cases for audit has long been a foundation of risk practices in tax administration. In leading administrations, sophisticated analytics support the optimization of this casework within robust compliance risk management (CRM) frameworks. Representing the culmination of trial, error and innovation, the selection methods that are increasingly used to support CRM all evolved from humble beginnings. For analytics, those beginnings often focused first on the use of third-party data and developing reliable techniques to identify instances of underreporting from self-assessed declarations. Today, a broad set of ideas shapes the use of audit, many of which have benefited from growing volumes of data and predictive techniques now available to help better understand compliance. Among this set of ideas are principles and established beliefs that:

- ***Audit is primarily a tool to promote voluntary compliance.*** Audit is not generally today considered a tool for direct revenue generation but rather as serving several purposes that include deterrence, monitoring of compliance, detection of noncompliance, education, and intelligence gathering
- ***Selection logic for audit should be defensible.*** A formal selection procedure should be documented, based on reasonable principles, and consistently applied, making use of steps that can be defended when necessary, including during an appeal.
- ***Centralized selection leads to better outcomes.*** Far from its origins in which auditors themselves selected cases, current methodologies effectively centralize all selection (or the majority) through systems and procedures that yield improved results.
- ***Effective centralization accounts for local knowledge.*** To avoid potential overreliance on automation and analytics, procedures need to incorporate the experience of staff to inform risk rules for case selection and capture feedback upon completion of casework.
- ***The analytics applied should mirror scale and levels of maturity.*** In small jurisdictions, or those with poor data quality, the use of basic techniques such as simple matching with third-party data sets may be more practical than the use of predictive analytics.
- ***Advanced analytics are most useful for small and medium taxpayers.*** Because large taxpayers are relatively few and complex, the data that is available from prior casework can have limited value as input into prediction algorithms for classification and case selection.
- ***Performance analysis before deployment is essential.*** Selection logic under development should be applied to historical audit case results to assess accuracy and establish a level of confidence in its effectiveness before actual use.
- ***Monitoring of selection logic after deployment is also critical.*** The prediction power of a selection method may vary from pre-deployment analysis for reasons not reflected in historic case results (for example, changes in economic conditions or tax policy itself).
- ***Expert rules and predictive analytics can coexist.*** Many practical scenarios exist where simple risk rules based on experience may be productive and, until more advanced methods are proven, should remain in regular, active use.
- ***With analytics, identifying the next best case is the ideal objective.*** A single set of integrated logic should be producing ranked lists of taxpayers that, at any time, makes it obvious to an analyst what the next best case for audit is (in lieu of, for example, analysts selecting from a list of potential cases labeled high, medium or low risk and, in the process, introducing biases).

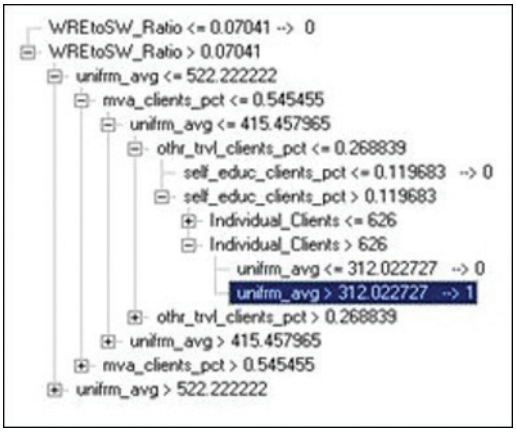
Many contemporary views on selection revolve around the differences between legacy and current best practice approaches. In both, analytics often rely on the use of rules. As illustrated in Annex Figure 6.1, the legacy approach combines data with rules derived by experts. Using Structured Query Language and, increasingly, specialized data models (for example, "OLAP"), analysts apply these rules, assign point scores to each, and navigate the results. In effect, this represents a top-down approach starting first with rules definition. Using advanced analytics, by comparison, algorithms themselves are capable of deriving rules from the bottom up.

Annex Figure 6.1. Top-Down and Bottom-Up Case Selection



Source: IMF staff.
Note: BI = Business Intelligence; OLAP = online analytical processing; SQL = Structured Query Language.

Annex Table 6.1. Legacy Approach and Best Practices

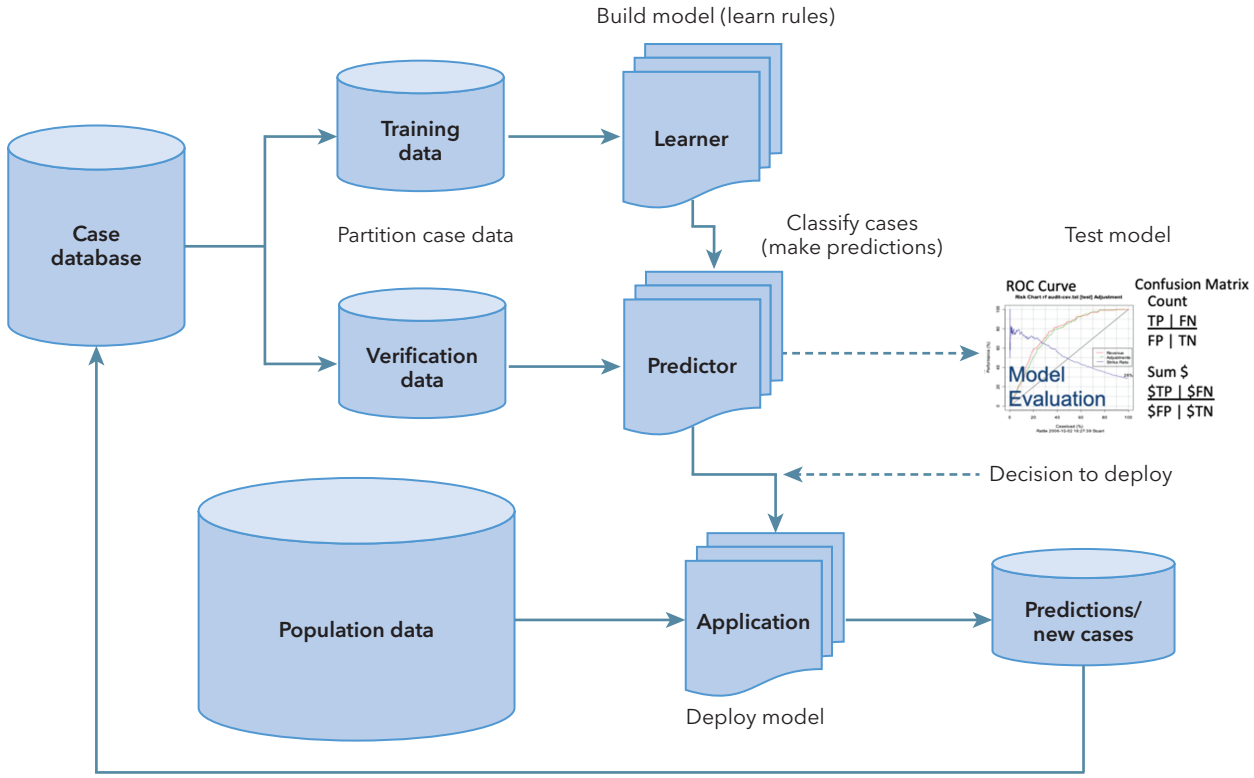
Legacy Approach (Selection by Expert Rules)	Best Practice (Data-Driven Selection)
Subject matter experts use their experience to create case selection rules to filter noncompliant clients out from the broader population in respect of particular risks.	Past cases of noncompliance are divided into "successful" cases (where relevant noncompliance was found) and "unsuccessful" cases (where it was not). The data set is further divided into a "training set" (used to build the analytic case selection model) and a "validation or verification set" to test that the model works.
These experts focus on client features that they believe assist in revealing whether a client is compliant.	Working with subject matter experts, an analytics specialist identifies client features that appear to be associated with noncompliance. These are tested statistically to confirm whether they are associated with noncompliance.
The rules created are often refined over time to enhance the strike rate. Often summary statistics such as strike rate and average adjustment are used to evaluate the outcomes. This is very problematic in skewed distributions. The rules produced are generally subjectively weighted to derive client risk scores for work prioritization.	A training data set is then selected and, essentially, regressed against a target set of successful cases to produce a risk-scoring algorithm that optimizes the probability of predicting a successful case from the data. A machine learning algorithm is then selected and tested against a separate validation data set to see how it performs using techniques such as ROC curves and confusion matrices.
A hypothetical example: • *If client has a WRE claim > $1,500 and uniform > $500 then risk score = 4* • *If client has motor vehicle claim > $5,000 then risk score = 2* • *If client has self-education claim > $2,000 then risk score = 3* • *Add client risk scores to produce total risk score* • *Select for review clients with total risk score > 8*	The rules produced are weighted by the algorithm to produce a client risk score for work prioritization: 
Rules produced this way (top-down) have an advantage of being known to a subject matter expert and are more explainable—but they are generally not optimal.	Rules produced this way (bottom-up) sometimes surprise subject matter experts and may require effort to understand and explain (although generally a simple decision tree can be retrofitted to provide explanatory power to the output of complex model that produces a higher strike rate).
Adding risk scores produces a centrally clustered risk distribution. Multiplying risk scores produces a log-normal distribution. There is a minimal effect on the rank order.	The rules produced from such data-driven approaches will usually outperform rules derived from subjective subject matter view on discriminate features and their weighting.

Source: IMF staff.

Note: ROC = receiver operating characteristic; WRE = work-related expense.

Current best practices for selection make use of predictive analytics and machine learning algorithms—often built from a single, foundational concept. As Annex Figure 6.2 illustrates, historical case results provide the basis for algorithms to make predictions of future behavior. By doing so, an implicit assumption is built into the logic applied that historic case results are sufficiently representative of taxpayer behaviors across different time periods. For many reasons (major changes in tax policy as one example), this may not always hold. Regardless, the concept described can produce results that may be tested in advance of deployment and compared with other approaches (such as the use of expert rules) to objectively identify which has the best prediction power, given the data available.

Annex Figure 6.2. Case Selection Using Predictive Analytics



Source: IMF staff.

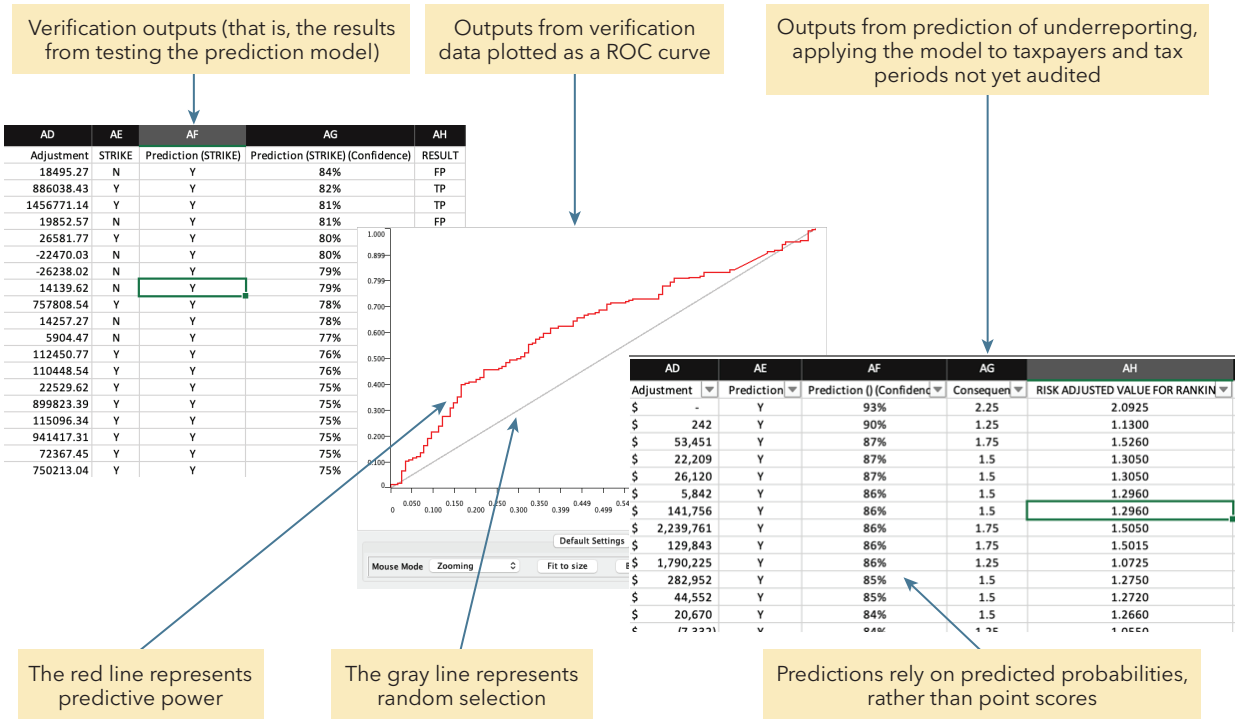
Although variations exist, the core concept that supports predictive analytics generally relies on the same or similar components, including the following:

- ***A case database.*** Containing results from casework and data points that are relied on by algorithms to make predictions (often additional audit assessment amounts).
- ***Training data.*** A relatively large random selection of data from the case database (60 to 80 percent), used by an algorithm to discover patterns and rules.
- ***A learner.*** An algorithm that discovers patterns and rules in historic case data based on the configuration of an analyst (for example, a Random Forest learner).
- ***Verification data.*** A smaller random selection of data from the case database (20 to 40 percent), used to test the accuracy of predictions made by an algorithm.
- ***A predictor.*** An algorithm that uses the patterns and rules identified by the learner to make new predictions using an application and population data (that is, the tax data representing taxpayers not yet selected).
- ***A test model.*** A workflow making use of verification data and applying performance measurement concepts (receiver operating characteristic curves) to determine prediction power.
- ***An application.*** Often used only in analyst tools, models may also be deployed directly in IT platforms that support major operations.
- ***Population data.*** A database of tax data including taxpayers that have not yet been selected for the types of cases and periods being evaluated.
- ***Predictions/new cases.*** Once made, predictions are registered in the case database, assigned for action, and tracked, ensuring that results feed into new analysis.

Greatly improving the accessibility of these concepts, desktop tools for advanced analytics are increasingly designed for low or no coding. Previously a barrier for many administrations, all of the advanced concepts presented can today be used by a data analyst without requiring the need for programming skills. While support from staff having capabilities to work with Structured Query Language is still usually needed (typically from IT personnel to aid in extracting the necessary data sets), once the data required has been provided, its analysis using a full range of techniques and algorithms can be managed through visual design and configuration of the respective models and workflow.

Annex Figure 6.3 presents outputs from the case selection template included in the toolkit accompanying this note. Development and use of the template, while benefiting from a general awareness and insight into the potential of advanced analytics, requires no knowledge of programming languages.

Annex Figure 6.3. Output from the Audit Case Selection Toolkit Template



AD	AE	AF	AG	AH
Adjustment	STRIKE	Prediction (STRIKE)	Prediction (STRIKE) (Confidence)	RESULT
18495.27	N	Y	84%	FP
886038.43	Y	Y	82%	TP
1456771.14	Y	Y	81%	TP
19852.57	N	Y	81%	FP
26581.77	Y	Y	80%	
-22470.03	N	Y	80%	
-26238.02	N	Y	79%	
14139.62	N	Y	79%	
757808.54	Y	Y	78%	
14257.27	N	Y	78%	
5904.47	N	Y	77%	
112450.77	Y	Y	76%	
110448.54	Y	Y	76%	
22529.62	Y	Y	75%	
899823.39	Y	Y	75%	
115096.34	Y	Y	75%	
941417.31	Y	Y	75%	
72367.45	Y	Y	75%	
750213.04	Y	Y	75%	

AD	AE	AF	AG	AH
Adjustment	Prediction	Prediction () (Confidenc	Consequen	RISK ADJUSTED VALUE FOR RANKIN
$ -	Y	93%	2.25	2.0925
$ 242	Y	90%	1.25	1.1300
$ 53,451	Y	87%	1.75	1.5260
$ 22,209	Y	87%	1.5	1.3050
$ 26,120	Y	87%	1.5	1.3050
$ 5,842	Y	86%	1.5	1.2960
$ 141,756	Y	86%	1.5	1.2960
$ 2,239,761	Y	86%	1.75	1.5050
$ 129,843	Y	86%	1.75	1.5015
$ 1,790,225	Y	86%	1.25	1.0725
$ 282,952	Y	85%	1.5	1.2750
$ 44,552	Y	85%	1.5	1.2720
$ 20,670	Y	84%	1.5	1.2660

Source: IMF staff.
Note: ROC = receiver operating characteristic.

References

Aslett, Joshua, and others. 2024. IMFx: VITARA – Information Technology and Data Management. edX. https://www.edx.org/school/imfx

Betts, Sue. 2022. "Revenue Administration: Compliance Risk Management Framework to Drive Revenue Performance." IMF Technical Note 2022/005, International Monetary Fund, Washington, DC.

Brondolo, John, Annette Chooi, Trevor Schloss, and Anthony Siouclis. 2022. "Compliance Risk Management: Developing Compliance Improvement Plans." IMF Technical Note 2022/001, International Monetary Fund, Washington, DC.

Chooi, Annette, and Jonathan Leigh Pemberton. 2023. IMFx: VITARA – Compliance Risk Management. edX. https://www.edx.org/learn/economics/the-international-monetary-fund-vitara-compliance-risk-management

Cotton, Margaret, and Gregory Dark. 2017a. "Use of Technology in Tax Administrations 1: Developing an Information Technology Strategic Plan (ITSP)." IMF Technical Note 2017/001, International Monetary Fund, Washington, DC.

Cotton, Margaret, and Gregory Dark. 2017b. "Use of Technology in Tax Administrations 2: Core Information Technology Systems in Tax Administrations." IMF Technical Note 2017/002, International Monetary Fund, Washington, DC.

Cotton, Margaret, and Gregory Dark. 2017c. "Use of Technology in Tax Administrations 3: Implementing a Commercial-Off-The-Shelf (COTS) Tax System." IMF Technical Note 2017/003, International Monetary Fund, Washington, DC.

Organisation for Economic Co-operation and Development (OECD). 2016. *Advanced Analytics for Better Tax Administration: Putting Data to Work*. Paris: OECD Publishing.

PUBLICATIONS

Tax Administration: Essential Analytics for Compliance Risk Management

TNM/2024/01