WORLD CUSTOMS ORGANIZATION
ORGANISATION MONDIALE DES DOUANES
Established in 1952 as the Customs Co-operation Council
Créée en 1952 sous le nom de Conseil de coopération douanière

PERMANENT TECHNICAL COMMITTEE
-
219th/220th Sessions
-
16 - 20 April 2018

PC0513E1a

Brussels, 8 March 2018.

DATA ANALYSIS

o Draft Guidance on Data Analytics

(Item XVII on the Agenda)

I. Background

1. In the growing digital economy and information society, data has become a significant business asset both for governments and business alike. The digital revolution is generating exponential opportunities regarding data access, capture, aggregation, and analysis for meeting organizational goals more effectively.

2. Last year, the WCO introduced the theme - “Data Analysis for Effective Border Management” to encourage the global Customs community to pursue their efforts and activities in this area. Data analysis and development of necessary capabilities and tools are of great importance, given the E-Commerce and Single Window environments as well as growing interaction between and among various stakeholders.

3. It is becoming increasingly apparent that Customs can benefit significantly by harnessing the power of the data and leveraging the advantages of advanced analytics. Applications of big data enable Customs and border regulatory agencies to adopt a proactive rather than a reactive response to supply chain risks, whilst facilitating legitimate trade.

4. Customs administrations could take greater advantage of the use of data analytics as an extremely powerful tool for improving their business processes at both the operational and strategic levels. Enhancing Customs’ ability to perform increasingly sophisticated analytics using the available data will become even more crucial in all future policy-making processes.

5. By bringing all the data together from the entire supply chain, Customs can obtain accurate and holistic pictures enabling them to identify trends - understanding who does what along the chain. Identifying trends can enable Customs authorities to spot suspicious activities that can lead to detecting frauds (e.g., smuggling of goods, under-invoicing) and identifying counterfeits and other restricted/prohibited goods.

For reasons of economy, documents are printed in limited number. Delegates are kindly asked to bring their copies to meetings and not to request additional copies.

6. Advanced data analytics such as predictive analytics can enable Customs to risk rank import and export transactions and create risk scores in real time, thus facilitating compliant traders while intercepting fraudulent shipments. Data analytics can equally help in identifying and prioritizing risk-based audits.

7. Highlighting the WCO theme of the year 2017, various WCO working bodies carried out related work from different perspectives. Reflecting on this topic from the supply chain and facilitation perspective, the March 2017 SAFE Working Group (SWG) agreed that the Secretariat and interested delegates from Member administrations and the private sector would develop resource guidance material on how data analytics could be used to enhance the implementation of the SAFE Framework of Standards and the AEO Programme.

8. The April 2017 Permanent Technical Committee (PTC), through breakouts, discussed the topic of data analysis and its different aspects, namely: 1. objectives; 2. data collection; and 3. IT solutions and provided relevant recommendations that included inter alia the use of data analysis in identifying trading patterns and designing new policies for identifying high-risk economic operators and improving trade facilitation.

II. Draft Guidance on Data Analysis

9. Following the aforementioned discussions, a draft outline Handbook on Data Analysis was presented to the October 2017 SWG meeting. The SWG approved the draft outline of the Handbook on Data Analysis and provided guidance and support for its further development. This draft was further discussed at the October 2017 Information Management Sub-Committee (IMSC) meeting and the December 2017 SAFE Sub-Group meeting.

10. Based on the inputs provided by the delegates of SWG, IMSC, SAFE Sub-Group, the draft Handbook has been further developed. The finalised draft Handbook on Data Analytics was submitted to the 19th SWG meeting held from 21 to 23 February 2018

11. Through breakout and plenary sessions, the SWG examined the draft Handbook on Data Analytics and approved it with additional suggestions. The suggestions include, inter alia, mirror analysis of import/export data, AEO data analysis, significance of data quality for a meaningful analysis, and providing analysis feedback to traders on non-compliance issues in appropriate cases. It also suggested enriching the Handbook by including good practices in this domain. The updated draft Handbook is appended as an Annex to this document.

12. In addition, the SWG suggested to seek the Enforcement Committee (EC)'s feedback on the draft Handbook to enhance operational perspectives.

13. The draft Handbook will be presented to the 37th session of the Enforcement Committee (EC) to be held 19 to 23 March 2018. The outcomes of the EC will be presented at the PTC. PTC Delegates are requested to coordinate with their respective EC delegates in advance of the meeting.

14. The Handbook essentially attempts to identify some key issues that have been elaborated together with a practical implementation approach. Notable among them are data analytics concept, strategy, governance, use cases, capabilities, tools, and associated processes.

15. It is intended to provide guidance and act as a reference resource base about the use of data analytics in the Customs environment, in particular assisting Members and stakeholders in further strengthening supply chain security and facilitation. It will be helpful in developing/enhancing data analytics strategies, operational frameworks, associated skills and capabilities within Customs administrations.

16. Some specific use cases of data analytics include the following (but are not limited to) :

- Effective risk management and trade facilitation,
- Efficient management of AEO programmes,
- Compliance management: identifies irregularities and inconsistencies in the data reported,
- Related party transactions: identifies suppliers for further investigation for related party pricing matters,
- Efficient tariff concession: identifies transactions where concessions may exist or where potential compliance errors arise,
- Incoterm analysis: identifies Incoterms outside of agreed supplier and industry terms for possible over or under valuations of goods,
- Trade lanes: identifies high volume trade lanes for efficient movement and clearance of goods,
- Valuation of goods: assesses valuation methodology against expectations, and
- Customs brokers/agents and other service providers: highlights efficiencies and inefficiencies.

III. **Action requested**

17. The PTC delegates are requested to:

- consider, if appropriate, approve the draft Handbook on Data Analytics; and

- share related good practices and successful working examples/initiatives that could be incorporated into the Handbook.

*

* *

Annex to
doc. PC0513E1

Data Analysis

- Practitioner’s Handbook

World Customs Organization

March 2018

For reasons of economy, documents are printed in limited number. Delegates are kindly asked to bring their copies to meetings and not to request additional copies.

Copyright © 2018 World Customs Organization. All rights reserved. Requests and inquiries concerning translation, reproduction and adaptation rights should be addressed to copyright@wcoomd.org.

Table of Contents

A. Introduction

1. The aim of this Handbook is to present a high-level overview of data analytics, more precisely what it is, how it works, and how useful it may be to Customs and other governmental agencies.

B. Data - A Strategic and Economic Asset

2. Until recently, data was generally perceived by most organizations as a component of an IT system. In the digital era, however, data has become a significant business asset - the more organizations know and use, the better. The digital revolution is generating exponential opportunities in terms of data access, capture, aggregation, and analysis. Growing digitalisation has made it easier and faster to process voluminous data. In recent years, a whole range of new tools has emerged that have the potential to help leverage data in new and powerful ways. Various solution providers could help to maximize uptime and minimize time needed to implementation and maintenance of data analytics systems.

3. Data may be structured or unstructured. Structured data may be collected via internal control systems, reports, and surveys, and presented in an organized manner. In such cases where the data is structured according to rules and categories, traditional data analysis may be applied and enriched with deeper insights via smart Machine Learning options.

4. Unstructured data may be extracted from digital media and communications (e.g. the “cloud,” emails, mobile phones, smart and connected containers), such that data is not properly structured according to rules and categories and does not follow standards. The aggregation of this “Big Data” necessitates the application of new methods of data mining and data analytics.

C. Data Analysis approach

5. All employees of Customs organizations should be empowered to embrace digital era by learning how to think, and create new insights from available data, or should have the possibility to safely experiment by looking proactively into new questions, either by themselves or with the support of assigned colleagues/experts or by leveraging local hackathons and virtual teams working on the more complex insights that cover various data sources.

6. There are different approaches and associated outcomes of data analytics that are based on organizations’ strategic objectives, data availability, and resource availability. They can be broadly grouped into the following :

i. Descriptive – What happened and/or what is happening now based on historical and incoming data. To mine the analytics, real-time dashboard and/or email reports are used.

ii. Diagnostic – A look at past performance to determine why it happened. The result of the analysis is often an analytic dashboard.

iii. Predictive – An analysis of likely scenarios of what might happen. The deliverables are usually a predictive forecast.

iv. Prescriptive – This type of analysis reveals what should be done. This is the most valuable kind of analysis and usually results in rules and recommendations for next steps. Recommendations based on multiple predictive models and complex analytical evaluations as to what options (pros/cons) to choose.

7. Data analysis, be it with small or large data sets, usually consists of the following key steps :

i. Defining the objectives (‘what new insights are needed + value to the organization?);
ii. Developing an analysis plan (what insights will be needed?);
iii. Defining the assumptions (quick wins and insights vs. long development cycles) ;
iv. Running the model;
v. Interpreting the results;
vi. Providing feedback regarding the interpreted data for decision making; and
vii. Monitoring, evaluating and improving the model.

8. There are several aspects that need to be considered when implementing/consolidating data analytics strategy. This may include, among others, the following :

i. Set out strategic priorities involving leadership to build a data-enabled organizational culture;
ii. Ascertain different types of data sources and choose right data relevant to the business domain;
iii. Establish/enhance IT systems that are capable of data sourcing, storage, and analysis;
iv. Examine available data analytics techniques and associated technologies;
v. Identify business-relevant analytics that can be put to an optimal use;
vi. Develop dynamic algorithms by leveraging machine learning and artificial intelligence;
vii. Invest in acquiring expertise relating to data analytics; and
viii. Develop/adapt data governance framework with appropriate legal enablers including those related to data security and privacy.

D. Opportunities

i. Leveraging Big Data

a. Definition of Big Data

9. As per Gartner[1], Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

10. Many organizations refer to Big Data to cover data lifecycle and data maturity requirements to address the need for new insights based on new data sources that are already available or get enabled upon demand outside of the organization. Also many Big Data questions typically relate to Management Information needs, based on new and/or deeper intelligence.

11. Big data refers to a large quantity of structured or unstructured data that may be analysed using advanced statistical analysis. The statistical analysis of a seemingly disparate set of data may reveal patterns, relations, and anomalies that would be difficult to uncover using traditional analytical methods.

b. How to explore Big Data

12. The analysis of Big Data is akin to the traditional process of statistical analysis: collect the data and store it in a practical format (for example as a structured table), cleanse the data (remove the seemingly corrupt or abnormal data that may bias the results), and then analyse it.

13. In data science, the first two steps correspond to a process called data warehousing. The statistical treatment of data is usually done - but not always - through data mining. Data warehousing corresponds to the collection, cleansing, and integration of a large amount of data from multiple sources. Data mining is the automated exploration of the data stored in the “warehouse” using artificial intelligence paradigms such as machine learning or agent-based network modelling.

14. However, in the case of unstructured data—that is data, such as free-form text, that do not fit in pre-established categories or that need to be processed according to rules to give it statistical significance—we first need to proceed to what is commonly called data munging or data wrangling. This is the process of fitting unstructured data into a structure according to rules, and is the first and most important step of data mining and machine learning. Through this step, the machine receives unstructured data, then makes sense of the universe it observes thanks to an initial set of rules and algorithms. Then, the analyst often

[1] An IT-related marketing company headquartered in Stamford, Connecticut, the United States.

conducts data snooping (also called data dredging or data fishing): which is the process of mining the now-structured data to validate a large number of hypotheses, such as correlations. This step usually requires strong understanding of statistical modelling. However, when snooping, a data scientist must always be critical of bogus correlations that are illogical. Such correlations are very likely to appear while data snooping.

15. Data mining may improve the efficiency and rationality of decision-making by providing better information on patterns of the world that may be difficult to uncover and understand by other means.

16. Leveraging Big Data capabilities to turn both internal and external available data into new insights to improve overall operations will provide lots of opportunities in many areas.

17. Examples: Cargo/goods
o Payments of duties / taxes
o Security (drugs / arms / flora & fauna)
o Performance of Customs officers
o Customs entry :
    - Detect misclassification (using artificial intelligence and Big Data).
    - Value of goods
    - Application of Rules + local interpretations+ judicial pronouncements+ WCO Rulings
    - Country of origin (are sanctions in place?)
    - Detect anti-dumping violations / misuse of free trade agreements)

18. Simplest and cost-effective approach would be to leverage highly secure public cloud ‘SaaS’ (Software as a service) solutions that are agile and can be deployed at a global scale in a cost efficient way.

ii. General Use Cases

a. Trend monitoring

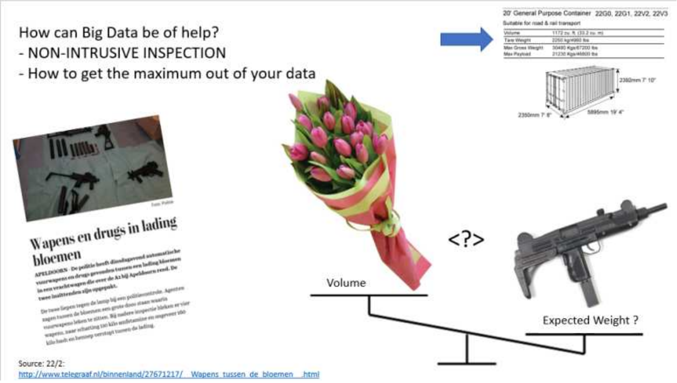
19. Trend monitoring allows organizations to understand variations in the demand of services, such that they may reallocate resources more efficiently. It is also useful, notably, to detect and react against fraud (for example, anomalous data flows may be detected using Benford’s law[2]—a practical notion that assumes that a frequency distribution is likely to be skewed towards smaller digits, with a smooth decreasing probability distribution) and organized crime (using geospatial intelligence to tracks suspicious conveyances and the flow of data).

b. Risk targeting and the prediction of threats

[2] Benford's law, also called the first-digit law, is an observation about the frequency distribution of leading digits in many real-life sets of numerical data. The law states that in many naturally occurring collections of numbers, the leading significant digit is likely to be small.

20. Big data analytics may be used to improve risk targeting by identifying with statistical data the groups that are the more likely to either increase or be subject to risks. In that sense, data may be used to make correlations between incidents and groups that were involved, such that it is easier to target risk according to probability analysis.

21. As example: leverage existing historical and known metrics of goods to enrich targeting of outliers. (e.g. compare typical weight of containers full of flowers shipment, track other containers loaded in same location / region + multiple destination harbours).



How can Big Data be of help?
- NON-INTRUSIVE INSPECTION
- How to get the maximum out of your data

20' General Purpose Container 22G0, 22G1, 22V2, 22V3
Suitable for road & rail transport

Volume	1172 cu. ft. (33.2 cu. m)
Tare Weight	2250 kg/4960 lbs
Max Gross Weight	30480 Kgs/67200 lbs
Max Payload	21230 Kgs/46800 lbs

Wapens en drugs in lading bloemen

APELDOORN - De politie heeft dinsdagavond automatische vuurwapens en drugs gevonden tussen een lading bloemen in een vrachtwagen die over de A1 bij Apeldoorn reed. De twee inzittenden zijn opgepakt.

De twee liepen tegen de lamp bij een politiecontrole. Agenten zagen tussen de bloemen een grote doos staan waarin vuurwapens leken te zitten. Bij nadere inspectie bleken er vier wapens, naar schatting 110 kilo amfetamine en ongeveer 160 kilo hasj en hennep verstopt tussen de lading.

Source: 22/2:
http://www.telegraaf.nl/binnenland/27671217/__Wapens_tussen_de_bloemen__.html

c. Decision-making

22. Using big data, trend monitoring, and risk assessment, to understand the likely consequences of policies and actions, can rationalize decision-making. By evaluating the payoffs of options relative to their probability, decision makers can more effectively make the decisions that involve the least risk and/or offer the best probable outcomes.

E. Data Analytics strategy

i. Key concepts

a. Data Collection and Mining

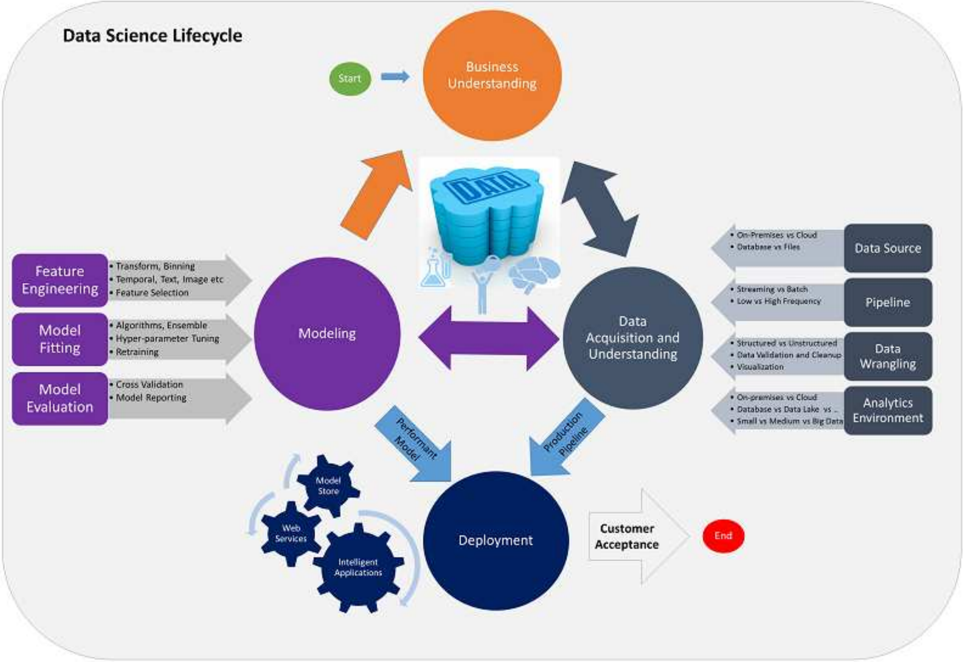
23. Data mining is the automated exploration of the data using artificial intelligence paradigms such as machine learning or agent-based network modelling.

24. Based on the data maturity lifecycle, data mining may start with internal available data and local providers. This could be enriched on case by case basis with data available from other sources, followed by coupling ‘external’ datasets.

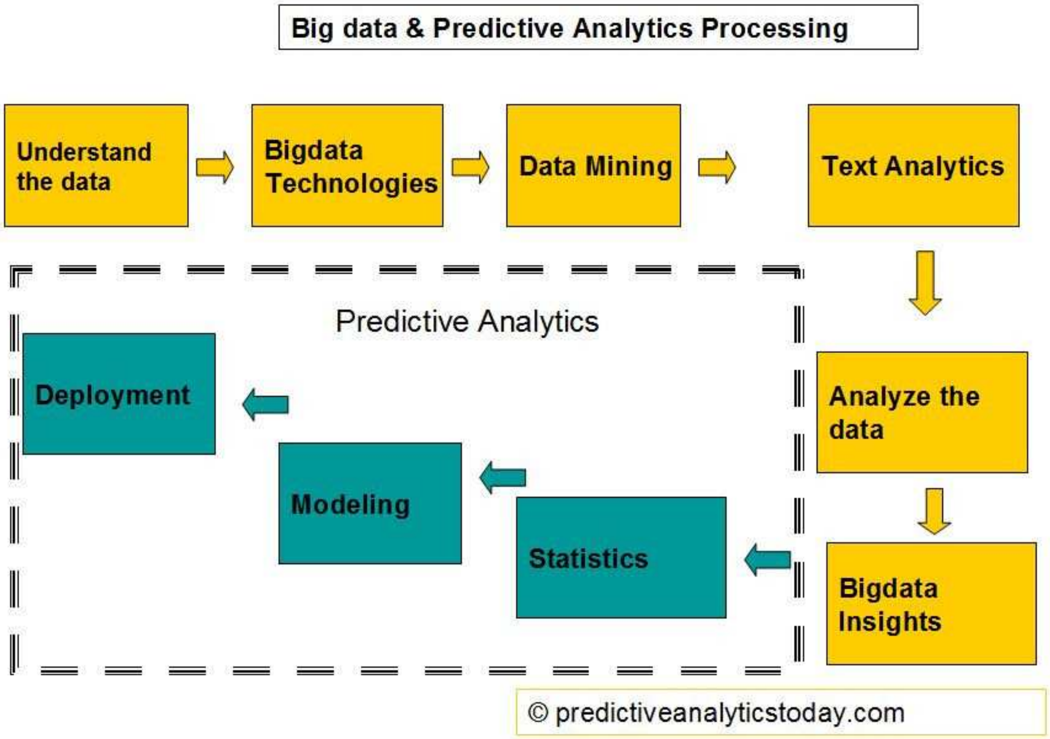
b. Modelling

25. Modelling is the abstraction and simplification of a phenomenon into an input-output system (for linear models) or a system of units that interact with each other (for complex models). Factors of influence are identified as players (objects, agents, actors, or nodes, according to the approach), and algorithms or rules of behaviour are defined by the data modeller to describe the influence of each player on the dynamics of a system or its parts.

26. Below is some visual representations of the Data Science Process lifecycle :



(*Source :* *https://docs.microsoft.com/en-us/azure/machine-learning/data-science-process-overview*)



(Source: https://www.predictiveanalyticstoday.com/big-data-analytics-and-predictive-analytics/)

c. Artificial Intelligence

27. Artificial intelligence is the expression of apparent intelligence akin to human intelligence by machines. Programmes either set pre-defined rules of behaviour that must be followed by a software or a machine in given contexts, or sets rules of machine learning that allow the computer to learn from its environment and respond accordingly.

d. Machine Learning/Deep Learning

28. Machine learning is an artificial learning paradigm that gives computer software the ability to autonomously identify patterns and learn from a set of data. Neural networks analysis, i.e. the analysis of relations between nodes (disparate and apparently independent data elements), is currently the most promising paradigm in machine learning.

29. In its ideal form, machine learning is autonomous and does not require human input. However, to allow such autonomy, humans need to programme adequate learning algorithms that are adapted to the available data, provide a sufficient amount of training data and calibrate the interpretation of the information.

e. Predictive Analytics

30. Predictive analytics is the use of statistical models to predict and/or explain the probable behaviour of actors or systems. Trend analysis is especially helpful to predict the behaviour of actors and systems that follow patterns, whereas agent-based modelling, network analysis and game theory may be more useful to predict the behaviour of actors and systems that follow rules or that are highly interdependent.

ii. Key Enablers

a. IT solutions and applications

31. Various IT solutions may be used for data analytics with or without Big Data. Some commercial and open-source solutions require mathematical modelling skills and/or knowledge of statistical models (e.g. SPSS, SAS, Maple, Wolfram Mathematica), whereas others only need input regarding the logic of a system’s dynamics (e.g. AnyLogic, RapidMiner, Neural Designer). Note that some solutions also offer more easily accessible business intelligence capabilities for more common purposes (e.g. SAP, Dundas BI, Oracle Data Mining), and that it is sometimes useful to design custom-built solutions (using statistical programming languages such as R and Python) to fit more specific intelligence needs.

b. Resources

32. Several data analytics software solutions, which are offered with open sources or proprietary licenses, are available to help Customs to implement data analytics. Considering that the application of appropriate techniques determines the success of an effective data analytics implementation, an investment in acquiring license agreement, sufficient knowledge and domain expertise is critical in the identification of most appropriate techniques/solutions. However, the implementation of data analytics can be equally effective through the use of open source software and techniques in a less costly manner.

c. Skills and Competencies

33. Data analytics is related to multidisciplinary subjects, such as data mining, knowledge discovery in database (KDD), machine learning, database administration, data science, and statistics. The first step in data analysis is to improve data integrity and data quality. Data scientists check the veracity of a data source, correct spelling mistakes, handle missing data and weed out irrelevant/misleading information in order to ensure the accuracy and consistency of data. This is the most critical step in the data value chain as junk data would generate wrong results and misleading business intelligence. Another key aspect is appropriate data modeling: building models that correlate the data with the business outcomes as well as developing algorithms to yield desired business intelligence and predictions. This is where the unique expertise becomes critical to business success.

34. Several types of cross-cutting expertise need to be considered for implementing data analytics, such as data analyst, data engineer, data scientist, quantitative analyst, statistician, econometrician and data-visualization specialist. Based on their policy considerations and human resources policy, Customs administrations need to determine whether in-house expertise needs to be developed or external experts can be engaged for the requisite support. Whilst data analytics indeed is not a subject that exclusively relates to Customs, it is important to tie data analytics technical expertise with an appropriate Customs domain understanding to achieve greater business value.

35. There is no universally accepted definition of a data scientist at the moment, and the process of data analytics requires such a wide variety of skills that it might be impossible to master all skills properly. The whole process needs data modellers, who need to be well versed in linear algebra and either category or set theory, and ideally in one or many analytics and predictive methods such as those of agent-based modelling, game theory, network theory, systems theory. Software engineers need knowledge of programming language that can be used to build applications, notably statistical languages such as R and Python, as well as superficial understanding of advanced statistical methods and linear algebra. Data analysts need superficial understanding of the logic of the models, more advanced knowledge of statistical analysis in order to interpret the data properly, and expertise regarding the field that is being studied in order to make the most logical relations.

F. Use Cases of Data Analytics

36. Data analytics is the process of analysing data sets in order to discover or uncover patterns, associations, and anomalies from sets of structured or unstructured data, and to draw practical conclusions. Big data analytics may reveal information that is not intuitive or difficult to assess by other means.

i. SAFE Framework of Standards

37. The SAFE Framework of Standards (hereafter, the SAFE Framework)[3] was adopted following unanimous agreement by the World Customs Organization in June 2005. It was adopted as means to help WCO members adapt to the increasing movement of trade and recent developments in the digital era, as well as to facilitate trade and the tracking and analysis of its contents. According to the SAFE Framework 2015 edition :

> “The SAFE Framework consists of four core elements. First, it harmonizes the advance electronic cargo information requirements on inbound, outbound and transit shipments. Second, each country that joins the SAFE Framework commits to employing a consistent risk management approach to address security threats. Third, it requires that at the reasonable request of the receiving nation, based upon a comparable risk targeting methodology, the sending nation's Customs administration will perform an outbound inspection of high-risk cargo and/or transport conveyances, preferably using non-intrusive detection equipment such as large-scale X-ray machines and radiation detectors. Fourth, the SAFE Framework suggests benefits that Customs will provide to businesses that meet minimal supply chain security standards and best practices.
> […]
> The SAFE Framework, based on the previously described four core elements, rests on the three pillars of Customs-to-Customs network arrangements, Customs-to-Business partnerships and Customs-to-other Government Agencies co-operation. The three-pillar strategy has many advantages. The pillars involve a set of standards that are consolidated to guarantee ease of understanding and rapid international implementation.
>
> Moreover, this instrument draws directly from existing WCO security and facilitation measures and programmes developed by Member administrations.”

[3] http://www.wcoomd.org/-/media/wco/public/global/pdf/topics/facilitation/instruments-and-tools/tools/safe-package/safe2015_e_final.pdf?la=en

38. In more depth, the SAFE outlines the following :

a. Advance Cargo Information

39. Carriers or their agents should submit advance electronic cargo declarations to the Customs at export and/or import. Advance declaration facilitates both trade and Customs' data analysis.

b. Risk Management

40. Risk management consists of the identification of emerging security risks and associated economic operators, routes, commodities, and the development of mitigation protocols early in the supply chain.

41. Effective risk management should use all the data available to Customs (cargo and goods declarations, supporting documents and other databases), other government agencies, and public domains. Agencies should establish automated risk-management systems to identify potentially high-risk conveyances using relevant assumptions. Currently, enhanced entity resolution algorithms are being used by some Customs agencies to identify shipments of interest. Entity Resolution is the task of disambiguating manifestations of real world entities in various records or mentions by linking and grouping. For example, there could be different ways of addressing the same person in text, different addresses for businesses, or photos of a particular object. This clearly has many applications, particularly in risk management and law enforcement.

c. Targeting

42. Targeting is the action of identifying actors or groups of actors that correspond to a set of criteria. For example, a border services agency may target high-risk individuals and businesses who cross or deal across the border. Customs agencies should cooperate with joint targeting and screening, standardized sets of targeting criteria, and sharing of information. The use of predictive models in the commercial and traveller streams by Customs agencies have resulted in significant seizures in areas such as drugs and agricultural contraband.

43. Usage of non-intrusive inspection (NII) equipment and radiation detection equipment is recommended, when applicable and appropriate, for the swift inspection of cargo and containers.

i. Frauds analysis
ii. Routes
iii. Commodities
iv. Conveyances
v. Analysis of NII images

d. Benefits

44. Benefits for the governmental agencies include: improved security through more efficient risk assessment.

ii. Implementation and management of AEO Programme

45. The Authorised Economic Operator (AEO) programme is based on Pillar 2 of the SAFE Framework and the WCO's Customs-to-Business partnership[4]. The objective is to grant additional facilitation measures to economic operators who meet with a specified set of criteria (as stipulated in the SAFE Framework).

46. The Framework comprises a principle of Mutual Recognition, such that Customs administrations agree to recognise AEO authorisations granted by other Customs agencies and to provide reciprocal benefits to AEOs.

47. The advantage for traders is therefore easier management of supply chains and international trade, whereas Customs agencies witness increased efficiency for processing of trade, treatment of data, and risk assessment.

48. The basic steps for using data analytics in an effective management of the AEO Programme are illustrated by giving examples of the following :

i. Network analytics,
ii. Compliance history/pattern,
iii. Systemic capability of continued compliance,
iv. Latent systemic risks,
v. Identification of potential change in compliance behavior, and
vi. Identification of emerging security risks.

G. Data Governance

49. Data governance is an integral part of good data analytics projects. It encompasses many processes relative to data acquisition, management, and warehousing. In other words, it consists of a set of processes that monitor and ensure the availability, integrity, and usability of data, from collection to modelling and computerized treatment.

50. The implementation of a data governance framework and the creation of a data governance catalogue are necessary for Customs agencies to properly understand, maintain and govern their information. A strong data governance framework can help agencies achieve the following positive outcomes :

[4] World Customs Organization. *Customs-Business Partnership Guidance*. URL: http://www.wcoomd.org/en/media/newsroom/2015/july/~/media/E2B8A58843F44C55AD21BBE9BA2672B3.ashx (Published: June 2015.)

- Improvements in timeliness, consistency, accuracy and accessibility of data which can enable senior management to make evidence based decisions in a timely manner,
- End-to-end data lineage and traceability, and
- Better quality data at the transactional level which can be used by advanced analytics teams to carry out important operational and research-based activities such as targeting, trade fraud detection and scheduling optimization.

i. Identification of Data Sources

a. Revenue Administrations (Customs and Tax)
b. Other Government Agencies
c. Public organizations
d. Private sector
e. Open sources

ii. Aggregation/Integration/Access of Data

51. Data acquisition is the process of bringing data from various internal and external source systems into a single, integrated data warehouse or data lake for the purposes of business intelligence and advanced analytics. Through data acquisition, Customs agencies can optimize these sources of information to assist in setting priorities, decision making, performance management, budget planning, forecasting and operations.

52. The following are examples of datasets that are commonly collected by Customs agencies and can be used to enhance the business intelligence and advanced analytics capabilities of agencies through data integration :

a. Data submitted for the Customs clearance process;
b. Collected data from other government agencies (Single Window, E-government);
c. Commercially available databases;
d. Open source information platforms such as digitized global public records and multilingual news sources;
e. Electronics, software, sensors and network connectivity, like with connected container initiative, which enables these objects to collect and exchange data, a phenomenon known as the “Internet of things”.

53. To properly facilitate the aggregation and integration of data, it is extremely important for Customs agencies to be well aware of the legal, security and privacy policies and regulations that govern the data that they use and collect. For several Customs agencies, a privacy impact assessment (PIA) or similar document is required when conducting activities that use personal information.

54. The following is an example of one process, which has been used by Customs agencies to support the eventual integration of data from numerous internal and external source systems to support business intelligence and operational analytics :

a. Identify which source systems are expected to have high business value when integrated with other systems,

b. Draft business scenarios which can be used to test the value of the integrated data, for example what type of results do we expect to see when you combine enforcement data with importer data?
c. Create an environment where data can be combined for testing of business value in a non-administrative way (no operational decisions can be made off this initial analysis),

d. Use the findings from the testing environment to build business case to justify the movement and combination of multiple data sources,

e. Complete all necessary privacy, legal and security documentation necessary to ultimately access and use this data in a production environment. Draw from the business case to support these documents. It is recommended that privacy, legal and security stakeholders be engaged early on in the process to help anticipate any potential roadblocks to data integration early on.

iii. Data Protection, Privacy and Security

a. Legal arrangements
b. Safeguards against Cybercrime (or cyber-attacks)

H. Tools

i. Predictive Analytics

55. Predictive analytics require models to analyze the data and provide insights relative to the evolution of observable patterns. This generally involves the formulation of algorithms and their operationalization in dedicated data analytics software.

56. Depending on the question that needs to be answered, the availability and nature of the data, and the inclination of data modellers, a wide variety of techniques may be used to formulate the algorithms. Notably, the modeller may make use of network theory, systems theory, game theory, or any other set of methods and general equations that is most adaptable to the case at-hand. These algorithms are then input into IT solutions -such as those presented in Section E above - to run the data through predictive models.

ii. Cognitive computing

57. Cognitive computing is the use of artificial intelligence, often through the use of artificial neural networks, to simulate human thought processes and allow for machine learning. Using self-learning algorithms, node-based models, or a combination of both, cognitive computing therefore imitates human learning within the scope set up by the programmers.

58. Technology solution providers provide such services in the form of Cognitive Services to make applications more intelligent, engaging and discoverable. This would enable

developers to add intelligent features like speech recognition, speech and language understanding into their applications.

iii. Statistical programming languages

59. R and Python are the two most popular statistical programming languages, followed by Java. Although Java is more popular overall among programmers, it is more often used to build the structure and the interface of programmes such that the code may be ported easily to any platform that support Java, rather than to process statistics. Also, among these languages, R is the only “true” statistical programming language; both Java and Python were built and are used for more general purposes such as web and game programming.

60. As these are interpreted programming languages, they either interpret pseudo machine code (e.g. Java), or add a layer of interpretation over an existing language (e.g. Python), unlike lower level languages like C and C++ that executes lower machine code (although the latter are still human-readable languages). Higher-level languages have the advantage of being easier for humans to read and write (interpreted languages such as Python almost mirror a strictly logic-oriented speech pattern), whereas lower-level languages are more difficult for humans to understand but are executed much more quickly and with less computing resources, as they give more direct commands to the hardware. That is why statistical programming languages are handy for computer scientists, but less commonly used in software engineering:
    a. Python and R do not need compiling and can be run directly, which accelerates experimentation.
    b. Both languages are easy to read and write, so they allow for faster development of automated statistical models for data analytics.
    c. As they are easy to read, it is easier to find parts that need to be modified if the data scientist wants to adjust a model.
    d. The syntax of these languages is much more similar to that of set theory, which is natural to mathematicians, and the programmer does not need to worry about memory allocation and the byte limits of operators.
    e. As interpreted languages are more natural for humans, the time savings to write and run the code is translated into cost savings for the enterprise.

61. In summary: it is best if data scientists use interpreted languages such as R and Python for experimentation and non-repetitive tasks. For all tasks that need to be repeated often and by multiple stakeholders at once, as well as tasks that require very high performance or speed in order to reduce the computational burden on servers, it is best if software engineers translate to a lower-level language the prototypes that were built in interpreted languages by data scientists.

Annexes

Model Flow Process

Before building an effective data analytics model, there is need for consistent and reliable data. This may entail the following steps :

a. Simplify access to all types of data, both internally and externally.

b. The data preparation process involves determining what data can best predict an outcome. And because more data generally means better predictors, bigger really is better in this case.

c. Improve the Data Analysts tools and capabilities with Advanced Analytics techniques.

d. A successful visualization is one that emphasizes the information of interest and presents it at a resolution sufficient to perform the task.[5]

e. Scrub data to build quality into existing processes.

f. Data cleansing begins with understanding the data through profiling, correcting data values (like typos and misspellings), adding missing data values (like ZIP code), finding and dealing with duplicate data or customer records, and standardizing data formats (dates, monetary values, units of measure). Cleaning data can also include automated selection of best records and cleaning data in multiple languages.

g. Share Metadata across Data Management and Analytics domains

h. Common metadata also provides lineage information on the data preparation process, which can answer questions like: Where did the data come from? What was its quality? What data was used, and where else has it been used? How was the data transformed? What additional reports or information products are developed using this data?

Good Practices

(to be added)

[5] Information Visualization in Data Mining and Knowledge Discovery", by Usama M. Fayyad, Andreas Wierse and Georges G. Grinstein, page 38.