OECD INTERNATIONAL ACADEMY FOR TAX AND FINANCIAL CRIME INVESTIGATION (ITALY)

Analytics and AI as a Tool for Investigators

Kong Yew Hon
Jul 2024

INLAND REVENUE
AUTHORITY
OF SINGAPORE

Learning Objectives

At the end of the course, you should be able to understand and appreciate:

(1) Overview of IRAS AI* Strategy

- ✓AI as a Strategic Capability
- ✓IRAS AI Strategy
- ✓Key Pillars of IRAS AI Strategy
- ✓IFD's Data Journey

(2) All About Data

- ✓The different sources of information
- ✓Data governance and limitation of certain data
- ✓Application of concept : Case Study
- ✓Group sharing

(3) Analytics Tool for Investigators

- ✓Leveraging on technology & analytics in financial investigation
  - Analytics as a gatekeeper;
  - Analytics for Tax Crime;
  - Analytics for investigative efficiency;
- ✓Application of concept : Case Study
- ✓Group sharing
- ✓What's next

** AI = Artificial Intelligence*

Survey

www.menti.com

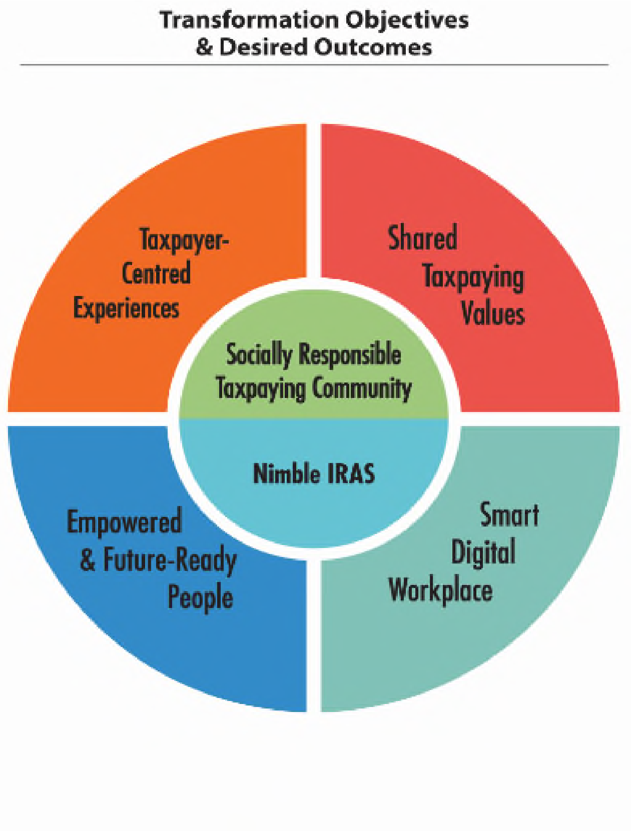
Access via Code :7987433

(1) Overview of IRAS AI Strategy

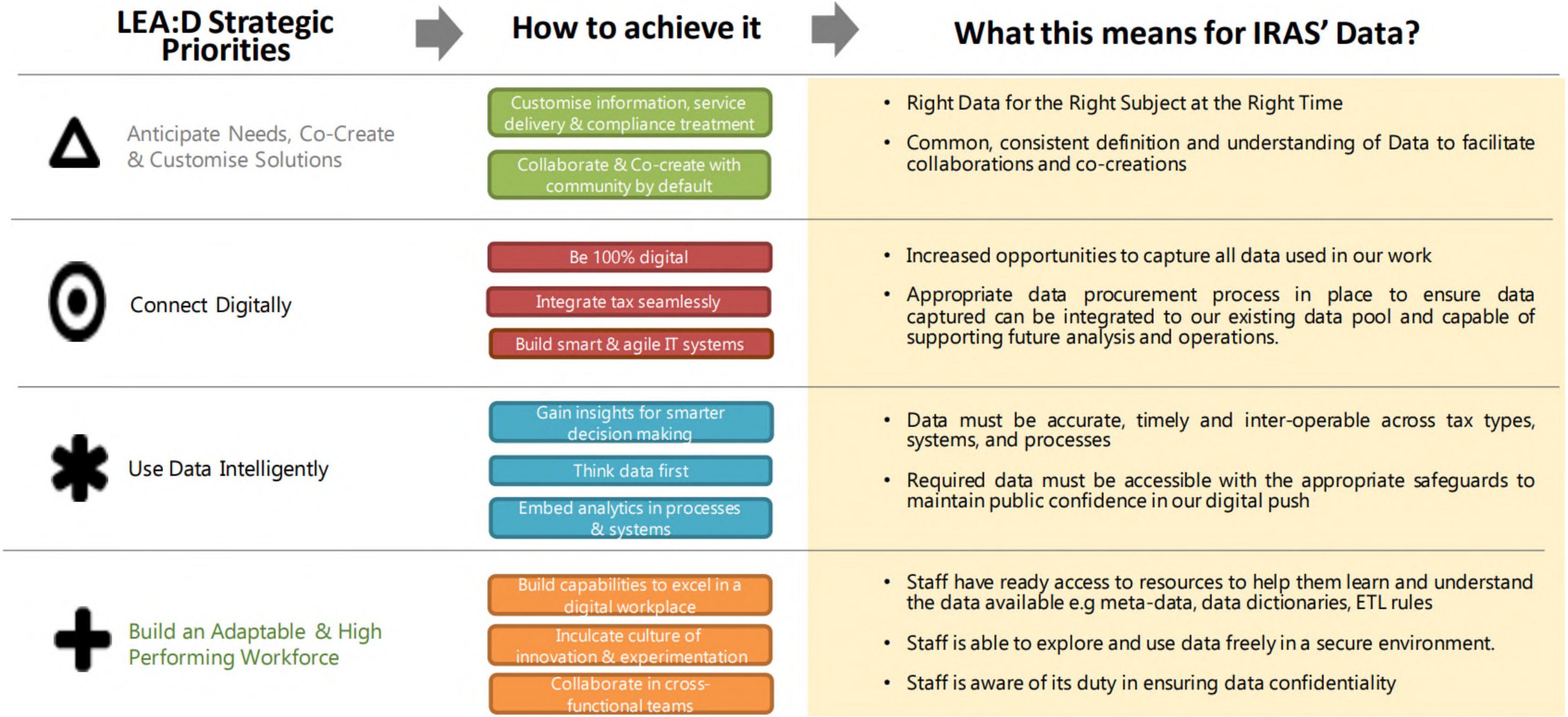
AI as a Strategic Capability

AI is identified as a strategic capability that enables IRAS to pursue desired outcomes in the organisation

- The objectives of AI are aligned with those under the IRAS Leveraging Analytics, Design and Digitalisation (LEA:D) transformation movement
  - To improve the **efficiency** and **effectiveness** of tax administration and enterprise grant disbursements
  - To deliver **anticipatory** and **integrated** services to taxpayers
  - To use **rich, entity-centric data** to develop intelligent applications
  - To build organizational and individual **capabilities** to leverage big data and AI


Transformation Objectives
& Desired Outcomes
Taxpayer-Centred Experiences
Shared Taxpaying Values
Socially Responsible Taxpaying Community
Nimble IRAS
Empowered & Future-Ready People
Smart Digital Workplace

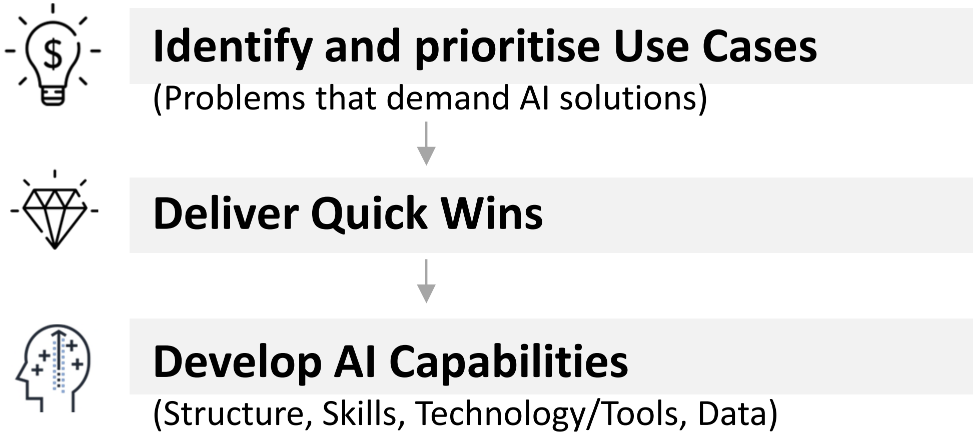
AI as a Strategic Capability



LEA:D Strategic Priorities	How to achieve it	What this means for IRAS' Data?
△ Anticipate Needs, Co-Create & Customise Solutions	Customise information, service delivery & compliance treatment Collaborate & Co-create with community by default	• Right Data for the Right Subject at the Right Time • Common, consistent definition and understanding of Data to facilitate collaborations and co-creations
◉ Connect Digitally	Be 100% digital Integrate tax seamlessly Build smart & agile IT systems	• Increased opportunities to capture all data used in our work • Appropriate data procurement process in place to ensure data captured can be integrated to our existing data pool and capable of supporting future analysis and operations.
✱ Use Data Intelligently	Gain insights for smarter decision making Think data first Embed analytics in processes & systems	• Data must be accurate, timely and inter-operable across tax types, systems, and processes • Required data must be accessible with the appropriate safeguards to maintain public confidence in our digital push
✚ Build an Adaptable & High Performing Workforce	Build capabilities to excel in a digital workplace Inculcate culture of innovation & experimentation Collaborate in cross-functional teams	• Staff have ready access to resources to help them learn and understand the data available e.g meta-data, data dictionaries, ETL rules • Staff is able to explore and use data freely in a secure environment. • Staff is aware of its duty in ensuring data confidentiality

IRAS AI Strategy

In FY2019, the **IRAS AI Strategy** was established and approved. It identified the strategic use-cases and roadmap to extend our analytics capabilities and to scale up the use of AI in IRAS.



Identify and prioritise Use Cases
(Problems that demand AI solutions)

Deliver Quick Wins

Develop AI Capabilities
(Structure, Skills, Technology/Tools, Data)

- Identify quick-wins and medium-term (or “invest”) use cases to kickstart AI roadmap development
- Build and scale AI capabilities for prioritised and future use cases

Scale (FY20 – FY23)

- Identify other use cases and prepare for development
- Scale up use cases, including those identified above, and use of AI in IRAS

Key Pillars of IRAS AI Strategy

The **IRAS AI Strategy** has 4 key pillars:

Applications

Deliver high impact AI use-cases in an agile and continuous manner

Data

Acquire and utilize data effectively to meet AI needs

Infrastructure / Tools

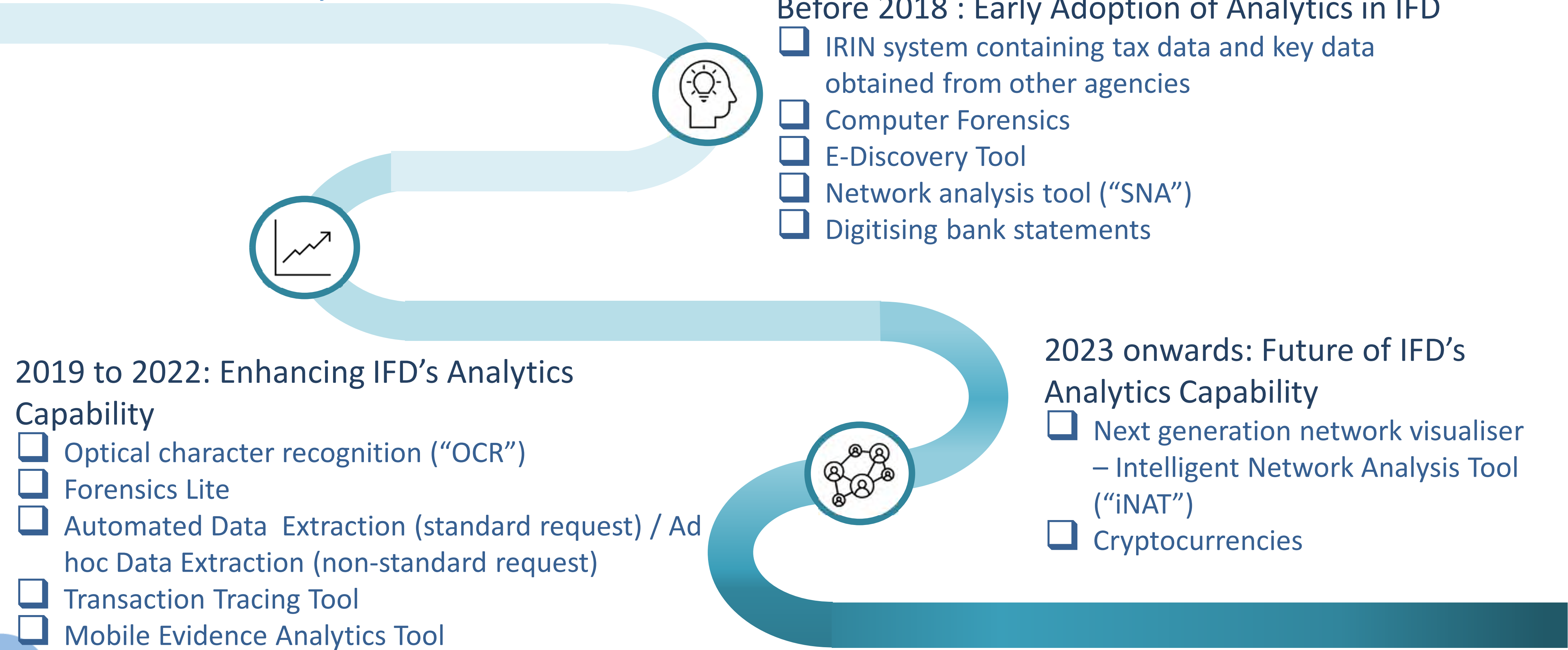
Provide tools and acquire up-to-date AI capabilities in the long-term

People / Organisation

Develop central expertise and in-house AI capabilities

Getting the Investigation & Forensic Division (IFD) ready for the future

*IFD's Data Journey**



Before 2018 : Early Adoption of Analytics in IFD

□ IRIN system containing tax data and key data obtained from other agencies
□ Computer Forensics
□ E-Discovery Tool
□ Network analysis tool ("SNA")
□ Digitising bank statements

2019 to 2022: Enhancing IFD's Analytics Capability

□ Optical character recognition ("OCR")
□ Forensics Lite
□ Automated Data Extraction (standard request) / Ad hoc Data Extraction (non-standard request)
□ Transaction Tracing Tool
□ Mobile Evidence Analytics Tool

2023 onwards: Future of IFD's Analytics Capability

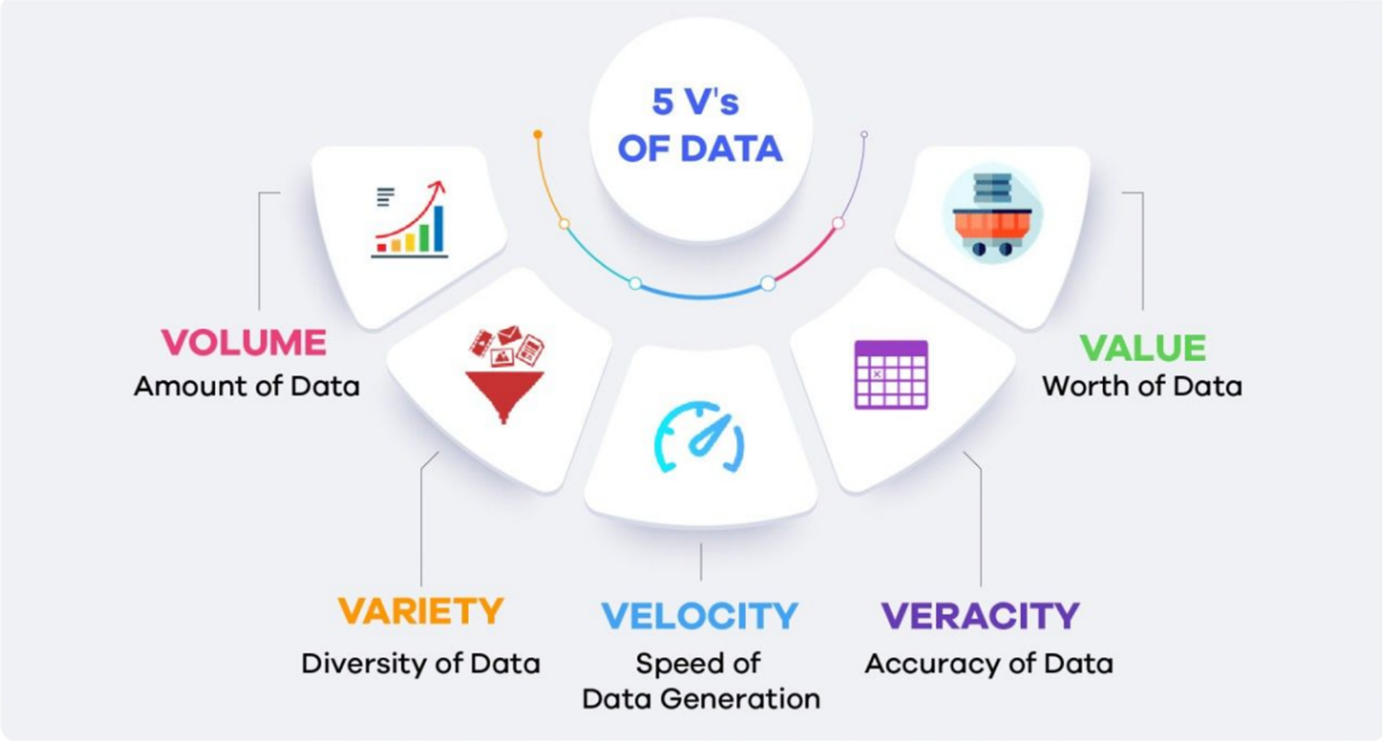
□ Next generation network visualiser – Intelligent Network Analysis Tool ("iNAT")
□ Cryptocurrencies

** Fully implemented for use by all officers including investigators, intel analysts, field officers and auditors*

(2) All About Data *[In the shoe of an Analyst / Investigator]*

5V’s of Big Data

- **Characteristics of Big data-** The five V's are volume, velocity, variety, veracity, and value*.

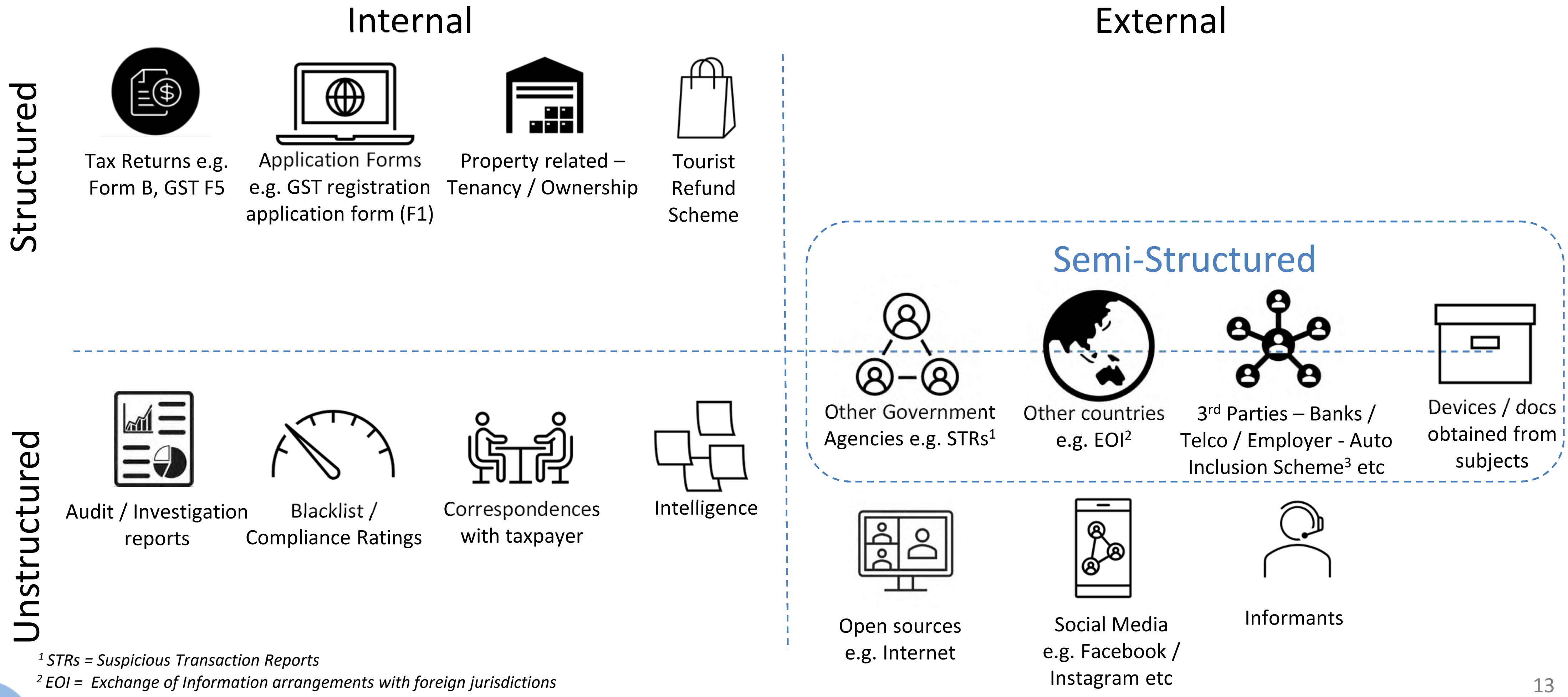


Video : https://youtu.be/bAyrObl7TYE?si=keT0GUC1UsAb5_D9

** Source : www.linkedin.com/pulse/big-data-explained-5v-excelsiorites*

The different sources of information

Sources of Data



Internal

External

Structured

Tax Returns e.g. Form B, GST F5

Application Forms e.g. GST registration application form (F1)

Property related – Tenancy / Ownership

Tourist Refund Scheme

Semi-Structured

Other Government Agencies e.g. STRs[1]

Other countries e.g. EOI[2]

3rd Parties – Banks / Telco / Employer - Auto Inclusion Scheme[3] etc

Devices / docs obtained from subjects

Unstructured

Audit / Investigation reports

Blacklist / Compliance Ratings

Correspondences with taxpayer

Intelligence

Open sources e.g. Internet

Social Media e.g. Facebook / Instagram etc

Informants

[1] STRs = Suspicious Transaction Reports

[2] EOI = Exchange of Information arrangements with foreign jurisdictions

[3] Auto Inclusion Scheme ("AIS"): Under this scheme, employers submits the employment income information of their employees to IRAS electronically

Sources of Data

Internal Structured Data	Internal Unstructured Data
Tends to be tax related data collected by IRAS from the various tax returns e.g. Income Tax returns - Form B / C, GST returns - GST Form 5 etc	Processed data after analysis performed on the data collected e.g. audit or investigation report / correspondences with the taxpayers
Most of IRAS' tax returns / applications are in electronic form where this facilitates data collection in structured format	Data are digitised and stored in IRAS Document Management System or shared folder
Data are stored in IRAS datawarehouses where it can be extracted for easy analysis and manipulation	

Sources of Data

External Structured Data	External Unstructured Data
Data provided by other government agencies e.g. Suspicious Transaction Report Office ("STRO") – STRs or Accounting and Corporate Regulatory Authority ("ACRA") – Singapore's equivalent of the Registrar of Business/ Companies	Subject's Facebook posts generally are collected on adhoc and need-to basis. IRAS also collect data from 3rd parties like banks for deposits / withdrawal data of our subjects during audit / investigation
Data from 3rd parties like employers under IRAS' Auto Inclusion Scheme – "AIS" e.g. Grab	Data would require considerable "cleaning" before use
Data are provided at a fixed intervals (e.g. annually for Central Provident Fund - "CPF"[1] data)	Users need to ascertain accuracy of data especially if sourced from the Internet due to the prevalence of fake news
	For investigation purpose, we have also employed OCR[2] technology to convert paper document seized from our subjects to digital format for ease of case analysis

[1] *Central Provident Fund or CPF is a mandatory social security savings scheme funded by contributions from employers and employees*

[2] *Optical character recognition or optical character reader I.e. OCR is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene photo or from subtitle text superimposed on an image (Source : Wikipedia)*

Internal Sources of Information

Tax Returns e.g. Form B, GST F5

Tax returns submitted by taxpayers. Examples of tax returns submitted to IRAS:

- Income Tax Form (Individuals / Partnership / Company)
- GST Form (Form 5 / Form 7 / Form 8)
- Withholding Tax

Application Forms – GST F1

Application forms submitted by taxpayers to apply for tax status or IRAS' schemes / incentives

- Income Tax Schemes / Incentives (Job Support Scheme etc)
- GST Registration application (Form 1) or scheme (Major Exporter Scheme etc)
- Overseas entities information (with effect from 1 Jan 2020 - under the Overseas Vendor Registration Regime – "OVR") *NEW*
- Withholding Tax

Property related – Tenancy / Ownership

Property related transactions

- Ownership of property (Sales / purchase price of property, name of buyer / seller etc)
- Tenancy related data (Amount of rental, period of tenancy, name of landlord / tenant etc)

Internal Sources of Information

Tourist Refund Scheme

Tourist / retailers for the Tourist Refund Scheme ("eTRS")

- Tourist spending information
- Retailer sales figures to tourist

Audit / Investigation reports

Processed data in the forms of reports

- Audit / investigation findings of taxpayer
- Provides insights to the taxpayer's attitude towards tax

Blacklist / Compliance Ratings

Internal assessment of the taxpayer propensity to commit fraud (blacklist) / non-compliance (compliance rating)

Internal Sources of Information

Correspondences with taxpayer

Past correspondences / enquires from the taxpayers

- Email
- Phone call (Telememo)
- Letter

Intelligence

Information concerning the subject of investigation obtained through intelligence channel

- Source can be from internal (IRAS officers) / external parties (e.g. Informant)
- Relationship linkage of subject
- Location of subject
- Modus operandi

External Sources of Data

Other Government Agencies

Other government agencies. Some examples:

- Business related data (E.g. Name of company director / shareholders / company registered address)
- Housing data (E.g. Public housing ownership information)
- Vehicle ownership data (E.g. Name of vehicle owner etc)
- Licensees data (E.g. Singapore Food Agency – Hawker licensee information)
- CPF data (E.g. Details of taxpayer's employment record)

Other countries e.g. EOI

Other countries:

- Exchange of Information through Avoidance of Double Taxation Agreements (DTAs) & Multilateral Convention on Mutual Administrative Assistance in Tax Matters
- Financial Intelligence Unit ("FIU")

3rd Parties – Banks / Telco etc

3rd parties. Some examples:

- Financial (E.g. Banks / Shares / Insurance)
- Subscriber (E.g. Telco / Internet Service Provider)
- Utilities (E.g. SP Powers / PUB etc)

External Sources of Data

Devices / docs obtained from subjects

Obtained from our subjects:

- Physical document e.g. invoices, Purchase Orders etc
- Digital devices e.g. Cellphone, laptop etc

Open sources e.g. Internet

Open sources like:

- Internet
- News reports
- Research papers
- Published judgment in courts

Social Media e.g. Facebook / Blogs

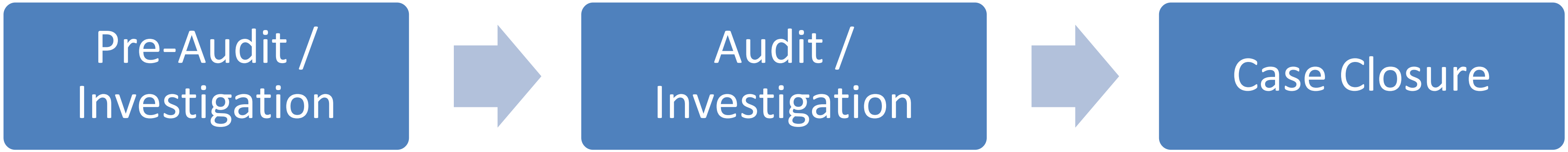
Social media:

- Facebook
- X (former known as "Twitter")
- Instagram

Informants

Whistleblowers on tax offence committed by subject

Use of Data (Syndicated Case)



Pre-Audit / Investigation

Audit / Investigation

Case Closure

Subject / Case Profiling

Analytics Models

Subject / Case Identification

Background of Subject – Location / Income

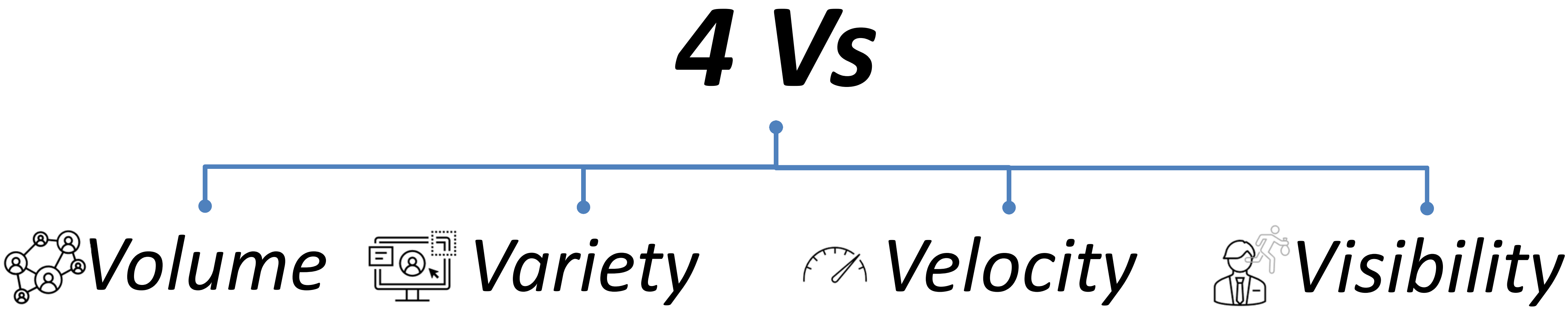
Relationship of subject – Family members, known associates

Unique identifiers of subject – Address / Email / IP address

Modus Operandi of fraud

Audit / Investigation report - Feedback loop to refine Analytics Models / risk indicators

Challenges for Syndicated Case



4 Vs

Volume

For syndicated case, investigators may be looking at more than 50 individuals and / or businesses. How to analyse the large volume of data?

Variety

❖ Diverse variety of internal / external data. What tool to use to extract / analyse data from various sources?

- ✓ Structured : Tax info / Transaction Listing
- ✓ Unstructured : Facebook
- ✓ Semi-structured : STRs / Banking info

Velocity

Thousands of transactions – invoices, funds, goods generated from fraudulent activities within a short time. Best approach to analyse the torrential flow of data?

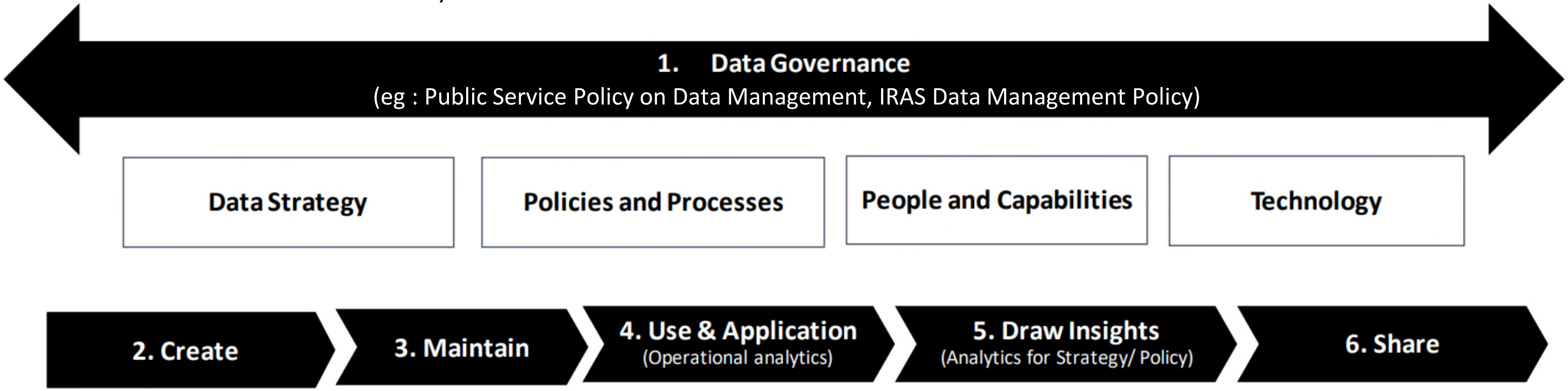
Visibility

Use of fall guy & set-up of shell company to frustrate investigation. Any tool to find the real mastermind?

Data governance and limitations of data

Data Governance

- Overview of Data Lifecycle*



1. Data Governance
(eg : Public Service Policy on Data Management, IRAS Data Management Policy)

Data Strategy

Policies and Processes

People and Capabilities

Technology

2. Create

3. Maintain

4. Use & Application
(Operational analytics)

5. Draw Insights
(Analytics for Strategy/ Policy)

6. Share

Data Governance is the exercise of decision making and authority for data-related matters.

The purpose of data governance is to ensure that data is managed properly, according to policies and best practices. This enables IRAS and the Government to make the best use of data for decision-making. It allows IRAS to achieve its goal of cultivating data- and insights-driven culture.

** Source : Data Steward Reference Materials*

Limitation of Data

Completeness : Data completeness describes whether the data you've collected reasonably covers the full scope of the question you're trying to answer, and if there are any gaps, missing values, or biases introduced that will impact your results*. Example, the data in the employment data table in IRAS system may not be complete as IRAS does not mandate the submission of employment data for businesses below certain staff strength.

**Source : https://www.montecarlodata.com/blog-what-is-data-completeness/*

Relevance : Data relevance is the degree to which data provides insight into the real-world problem or purpose being addressed and contributes to the overall understanding of the business^. For example, IRAS conducted audit on subject from 2014 to 2015. Due to the long passage of time, the audit findings for the subject may not be relevant for the subsequent investigation in 2024.

^Source : www.metaplane.dev/blog/data-relevance-definition-examples

Timeliness : The degree to which data represent reality from the required point in time#. In IRAS' example, certain information are provided at fixed interval e.g. CPF data is provided to IRAS annually. Hence, to get the most up-to-date info, officer would need to write to the relevant authority for the latest information.

#Source : https://dsstream.com/introduction-to-data-quality-terms-definitions-examples-of-use/

Limitation of Data

Reliability and authenticity : Data reliability is the degree to which data, and the insights gleaned from it can be trusted and used for effective decision-making*. For example, certain data provided / obtained by IRAS will need to be assessed for its reliability like financial data of Entity A shared by whistleblower. IRAS would not know if the data provided by the whistleblower is authentic.

**Source : www.thoughtspot.com/data-trends/analytics/data-reliability*

Accuracy : Data is considered accurate if it describes the real world^. One example will be for data obtained from the Internet; it would need to be validated against other sources to confirm its accuracy.

^Source : www.metaplane.dev/blog/data-accuracy-definition-examples

Evaluation Matrix of Data / Information Provided to IRAS (e.g. Whistleblower)

TABLE 1 - SOURCE RELIABILITY

	RATING	DESCRIPTION
A	Reliable	No doubt about the source's authenticity, trustworthiness, or competency. History of complete reliability.
B	Usually reliable	Minor doubts. History of mostly valid information.
C	Fairly reliable	Doubts. Provided valid information in the past.
D	Not usually reliable	Significant doubts. Provided valid information in the past.
E	Unreliable	Lacks authenticity, trustworthiness, and competency. History of invalid information
F	Cannot be judged	Insufficient information to evaluate reliability. May or may not be reliable.

TABLE 2 - INFORMATION RELIABILITY

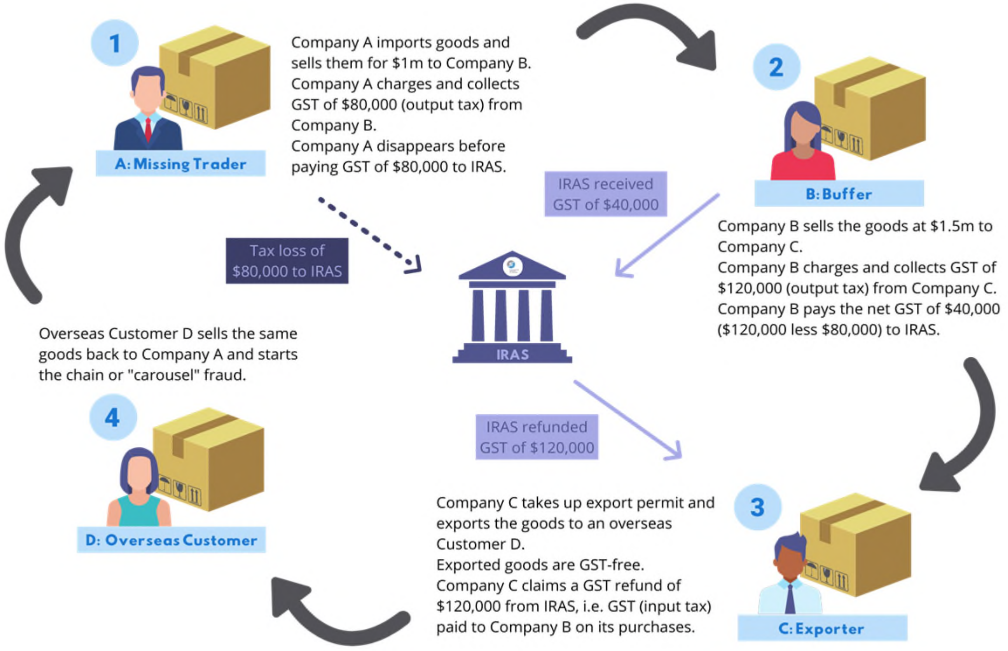
	RATING	DESCRIPTION
1	Confirmed	Logical, consistent with other relevant information, confirmed by independent sources.
2	Probably true	Logical, consistent with other relevant information, not confirmed.
3	Possibly true	Reasonably logical, agrees with some relevant information, not confirmed.
4	Doubtfully true	Not logical but possible, no other information on the subject, not confirmed.
5	Improbable	Not logical, contradicted by other relevant information.
6	Cannot be judged	The validity of the information can not be determined.

Question : Can you share how your agency deals with whistleblowers?

Case Studies

What is Missing Trader Fraud?

MTF* occurs when organised criminal groups abuse the GST/VAT refund system for fraudulent export arrangements and exploit the asymmetry of information in different jurisdictions to avoid being identified/traced.


1
A: Missing Trader
2
B: Buffer
3
C: Exporter
4
D: Overseas Customer
IRAS
IRAS received GST of $40,000
Tax loss of $80,000 to IRAS
IRAS refunded GST of $120,000

Company A imports goods and sells them for $1m to Company B. Company A charges and collects GST of $80,000 (output tax) from Company B. Company A disappears before paying GST of $80,000 to IRAS.

Company B sells the goods at $1.5m to Company C. Company B charges and collects GST of $120,000 (output tax) from Company C. Company B pays the net GST of $40,000 ($120,000 less $80,000) to IRAS.

Company C takes up export permit and exports the goods to an overseas Customer D. Exported goods are GST-free. Company C claims a GST refund of $120,000 from IRAS, i.e. GST (input tax) paid to Company B on its purchases.

Overseas Customer D sells the same goods back to Company A and starts the chain or "carousel" fraud.

What is a Missing Trader Fraud (MTF):

“Under a typical MTF arrangement, a group of businesses would form a supply chain and the same goods/services would be supplied through the chain. To ensure that the final sales of the goods/services are not subjected to GST, the goods/services would ultimately be exported to an overseas customer. A seller upstream in the supply chain would charge GST on the sale of goods to businesses downstream and instead of paying the GST to IRAS, the upstream supplier would ***fail to account*** *in its GST return the GST it had collected.*

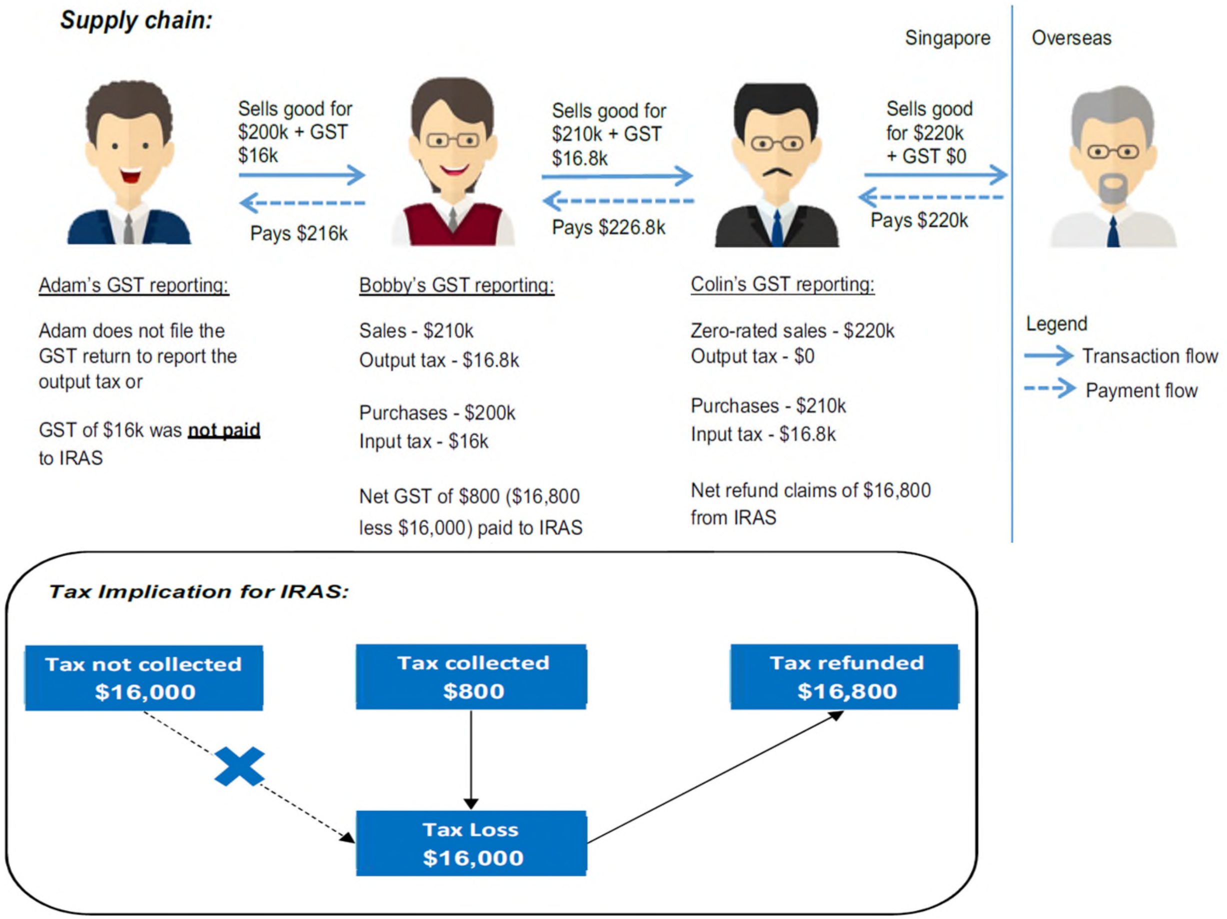
This is termed ***“Missing Trader” fraud*** *as the seller disappears with the GST.”*

** Missing Trader Fraud or MTF is also commonly known as Carousel Fraud in the other parts of the world like the European Union (“EU”)*

What is Missing Trader Fraud?

Illustration of a MTF arrangement

Question : Can share your agency's experience with MTF?



Supply chain:

Singapore | Overseas

Sells good for $200k + GST $16k

Pays $216k

Sells good for $210k + GST $16.8k

Pays $226.8k

Sells good for $220k + GST $0

Pays $220k

Adam's GST reporting:

Adam does not file the GST return to report the output tax or

GST of $16k was **not paid** to IRAS

Bobby's GST reporting:

Sales - $210k
Output tax - $16.8k

Purchases - $200k
Input tax - $16k

Net GST of $800 ($16,800 less $16,000) paid to IRAS

Colin's GST reporting:

Zero-rated sales - $220k
Output tax - $0

Purchases - $210k
Input tax - $16.8k

Net refund claims of $16,800 from IRAS

Legend

→ Transaction flow

⇢ Payment flow

Tax Implication for IRAS:

Tax not collected $16,000

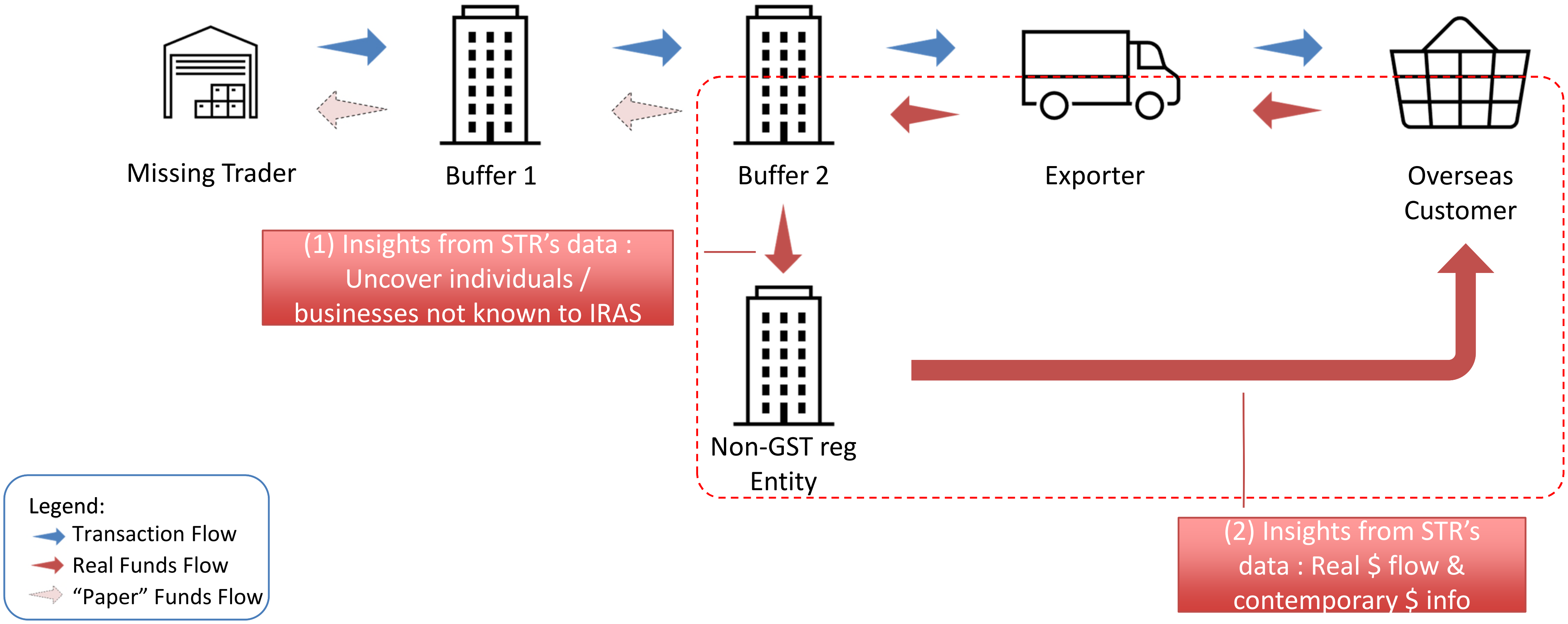
Tax collected $800

Tax refunded $16,800

Tax Loss $16,000

Operation Wand 2 – Missing Trader Fraud

Example : Use of Suspicious Transaction Report ("STR") in Operation Wand 2



(1) Insights from STR's data : Uncover individuals / businesses not known to IRAS

(2) Insights from STR's data : Real $ flow & contemporary $ info

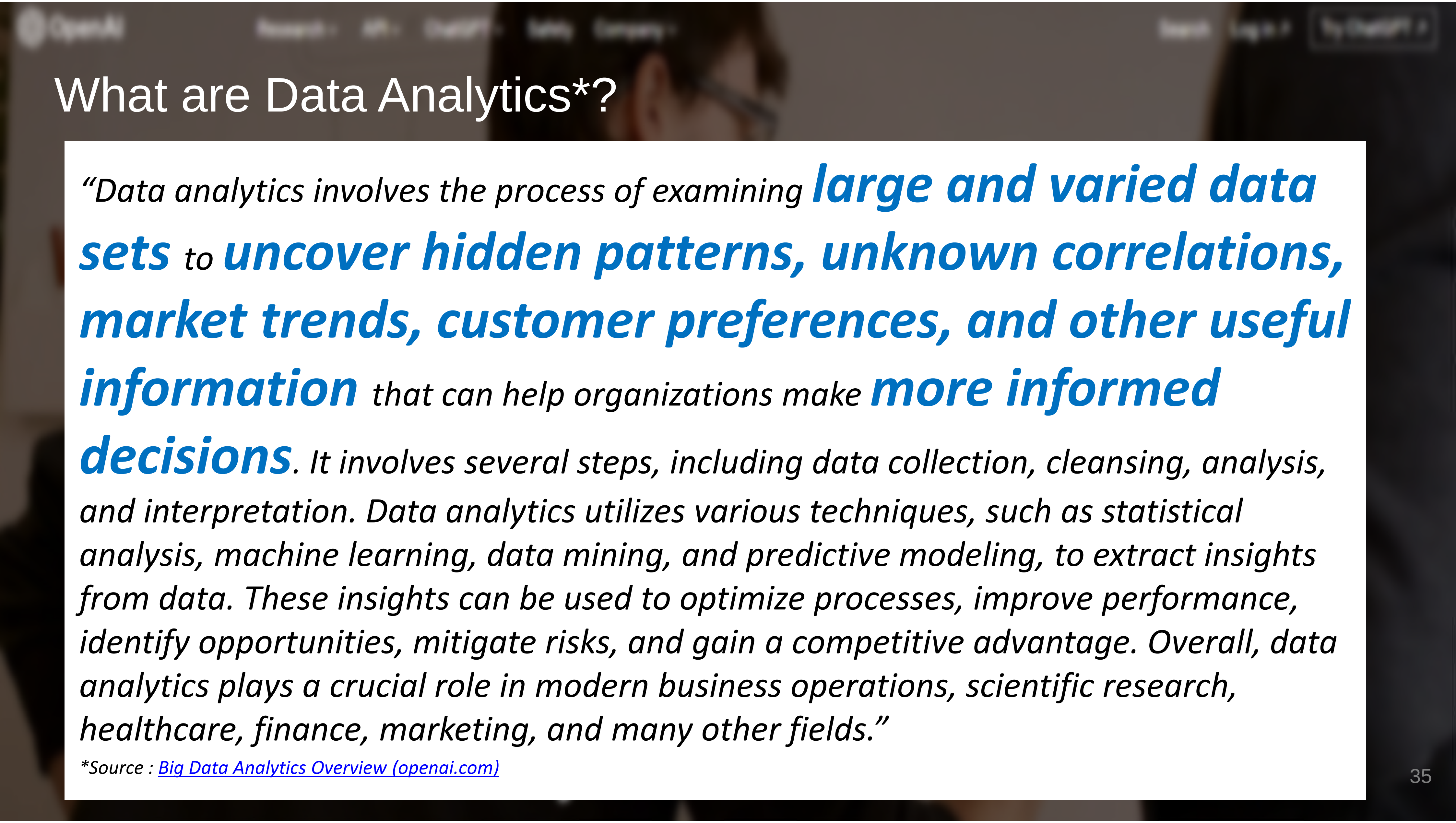
Country Sharing

Country Sharing (20 mins)

Group 1 & 2 : [Current] Any data that has been helpful to combat tax crime?

Group 3 & 4 : [Future] What data would you like to have to enhance your capability to combat tax crime?

(3) Analytics Tool for Investigators



What are Data Analytics*?

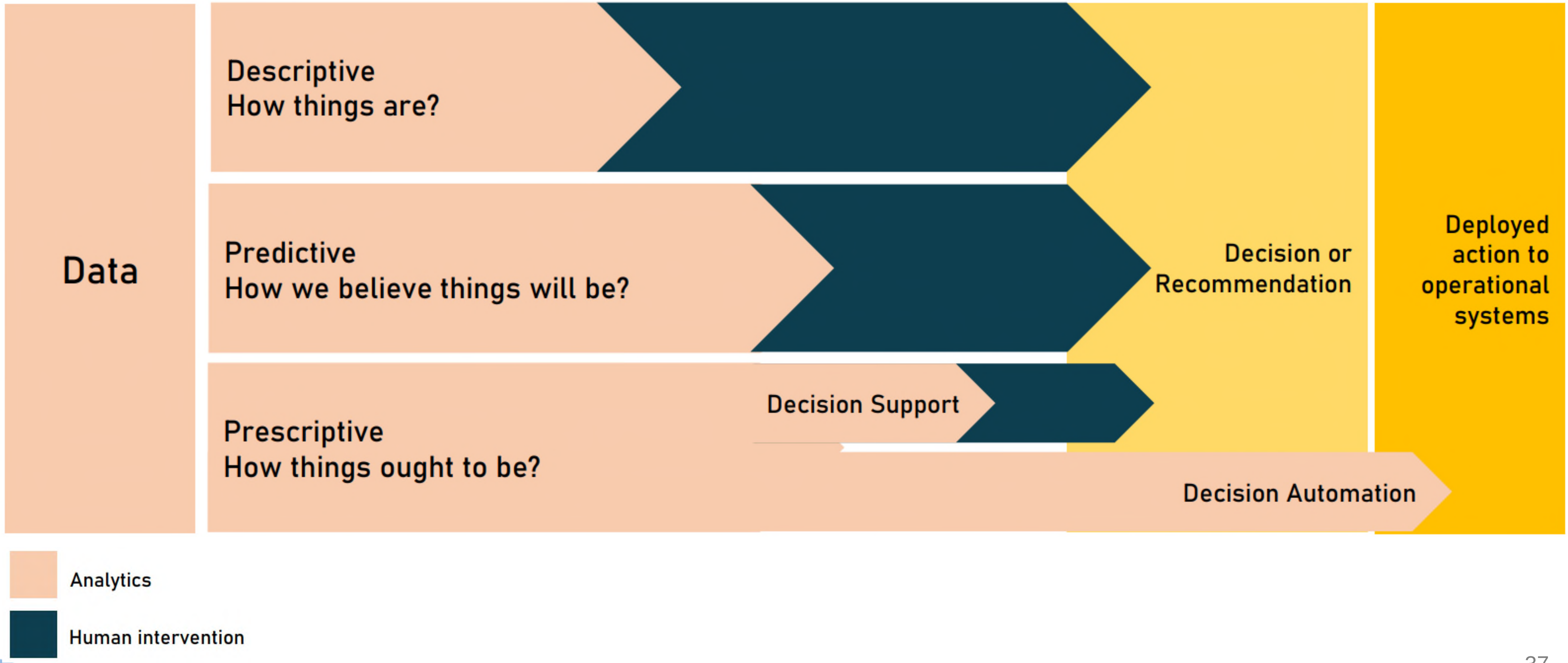
“Data analytics involves the process of examining ***large and varied data sets*** *to* ***uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful information*** *that can help organizations make* ***more informed decisions****. It involves several steps, including data collection, cleansing, analysis, and interpretation. Data analytics utilizes various techniques, such as statistical analysis, machine learning, data mining, and predictive modeling, to extract insights from data. These insights can be used to optimize processes, improve performance, identify opportunities, mitigate risks, and gain a competitive advantage. Overall, data analytics plays a crucial role in modern business operations, scientific research, healthcare, finance, marketing, and many other fields.”*

**Source : Big Data Analytics Overview (openai.com)*

What are Data Analytics (3mins video)?

- https://www.youtube.com/watch?v=nJa20uMtR7I

What are Data Analytics*?



Data

Descriptive
How things are?

Predictive
How we believe things will be?

Prescriptive
How things ought to be?

Decision Support

Decision Automation

Decision or Recommendation

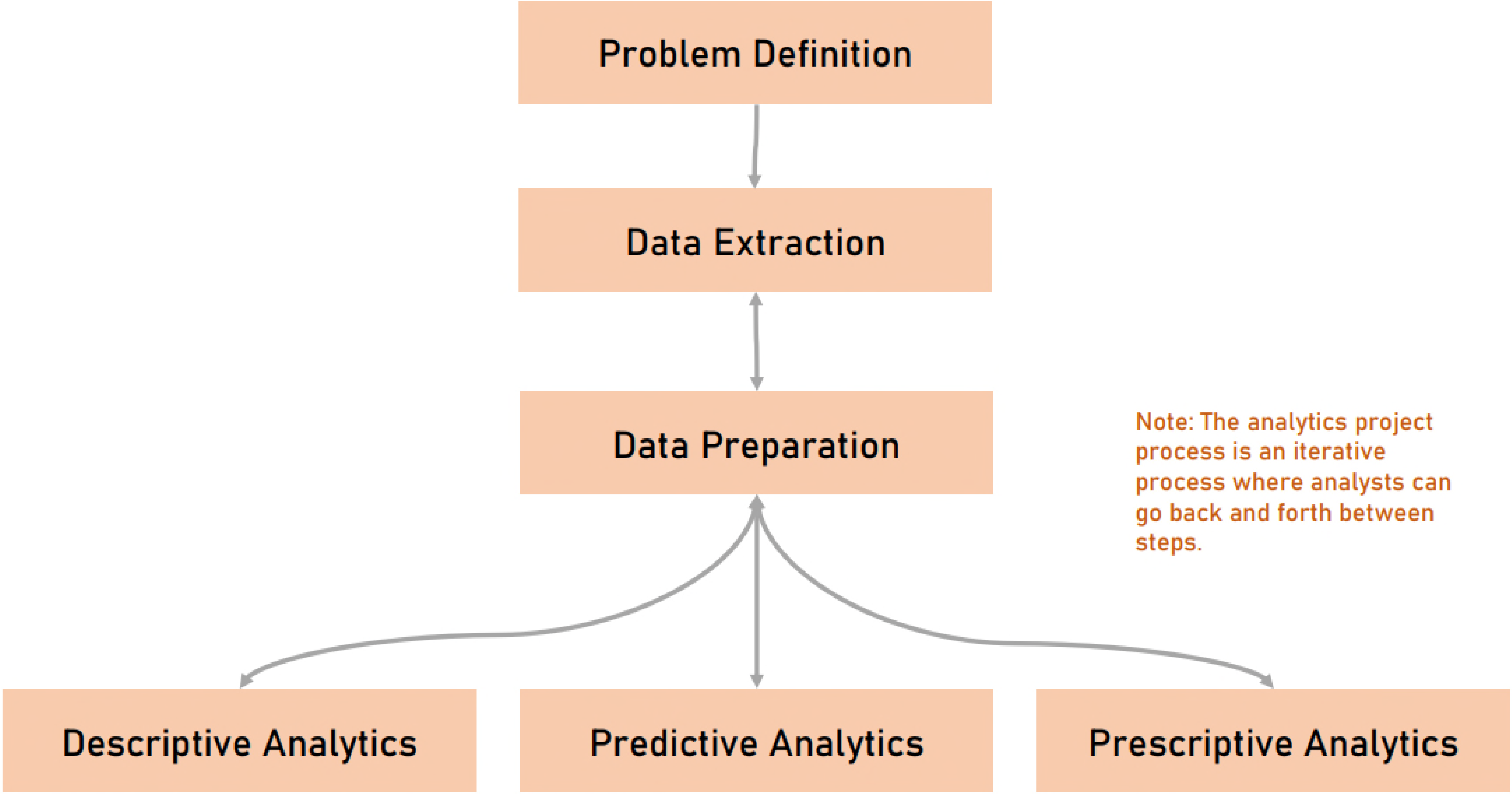
Deployed action to operational systems

Analytics

Human intervention

**Source : IRAS' Analytics Toolkit*

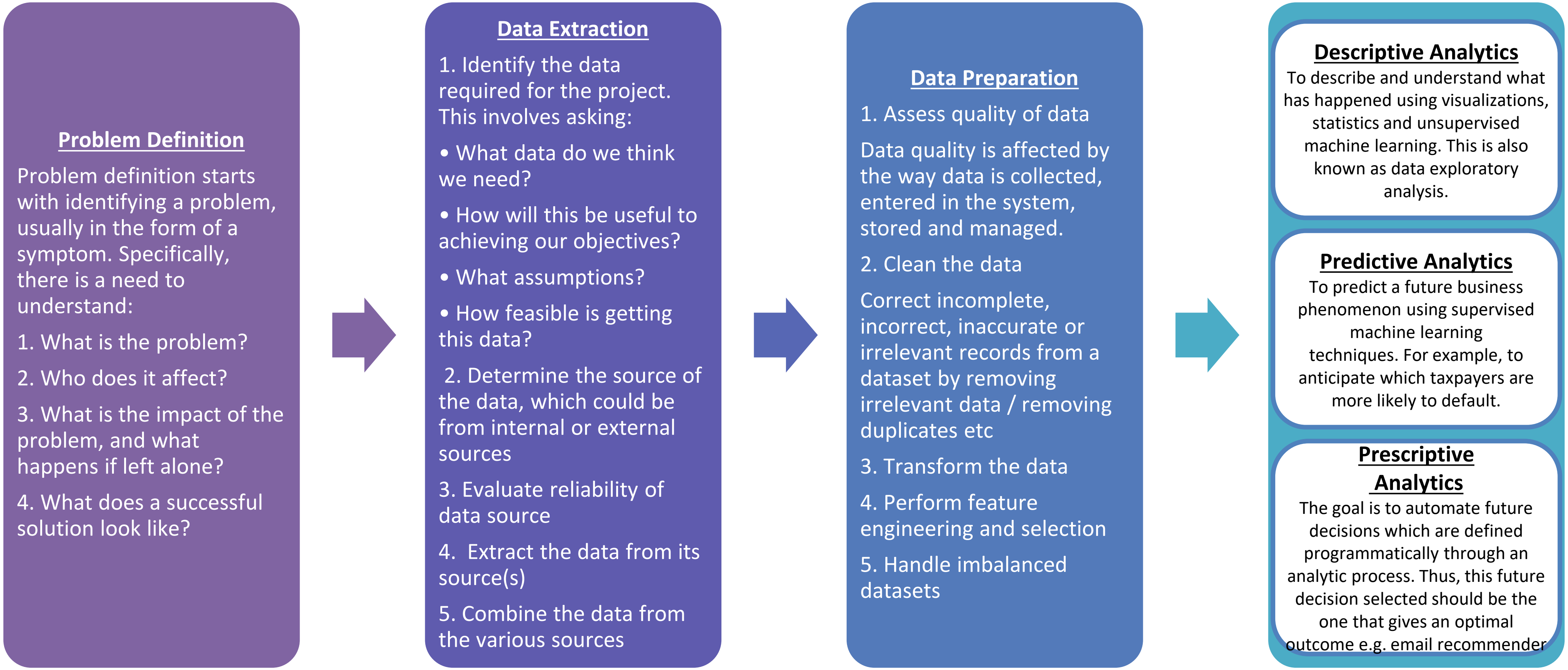
Data Analytics Project Process*



Note: The analytics project process is an iterative process where analysts can go back and forth between steps.

**Source : IRAS' Analytics Toolkit*

Data Analytics Project Process*



Problem Definition

Problem definition starts with identifying a problem, usually in the form of a symptom. Specifically, there is a need to understand:

1. What is the problem?

2. Who does it affect?

3. What is the impact of the problem, and what happens if left alone?

4. What does a successful solution look like?

Data Extraction

1. Identify the data required for the project. This involves asking:

- What data do we think we need?
- How will this be useful to achieving our objectives?
- What assumptions?
- How feasible is getting this data?

2. Determine the source of the data, which could be from internal or external sources

3. Evaluate reliability of data source

4. Extract the data from its source(s)

5. Combine the data from the various sources

Data Preparation

1. Assess quality of data

Data quality is affected by the way data is collected, entered in the system, stored and managed.

2. Clean the data

Correct incomplete, incorrect, inaccurate or irrelevant records from a dataset by removing irrelevant data / removing duplicates etc

3. Transform the data

4. Perform feature engineering and selection

5. Handle imbalanced datasets

Descriptive Analytics

To describe and understand what has happened using visualizations, statistics and unsupervised machine learning. This is also known as data exploratory analysis.

Predictive Analytics

To predict a future business phenomenon using supervised machine learning techniques. For example, to anticipate which taxpayers are more likely to default.

Prescriptive Analytics

The goal is to automate future decisions which are defined programmatically through an analytic process. Thus, this future decision selected should be the one that gives an optimal outcome e.g. email recommender

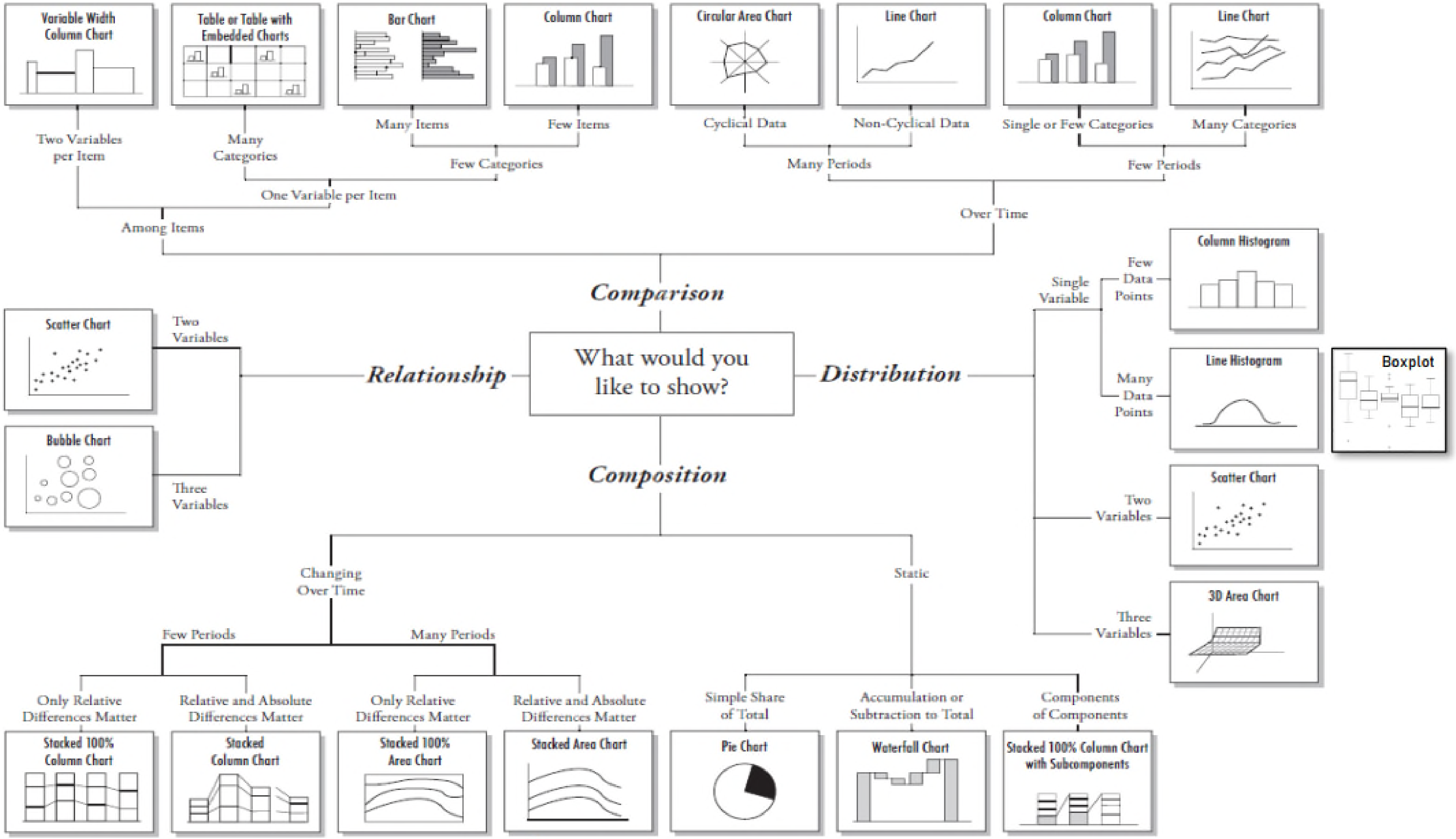
**Source : IRAS' Analytics Toolkit*

Data Analytics Project Process*

Descriptive Analytics

To describe and understand what has happened using visualizations, statistics
and unsupervised machine learning. This is also known as data exploratory analysis.

Depending on what you would like to show, you may use the following charts:



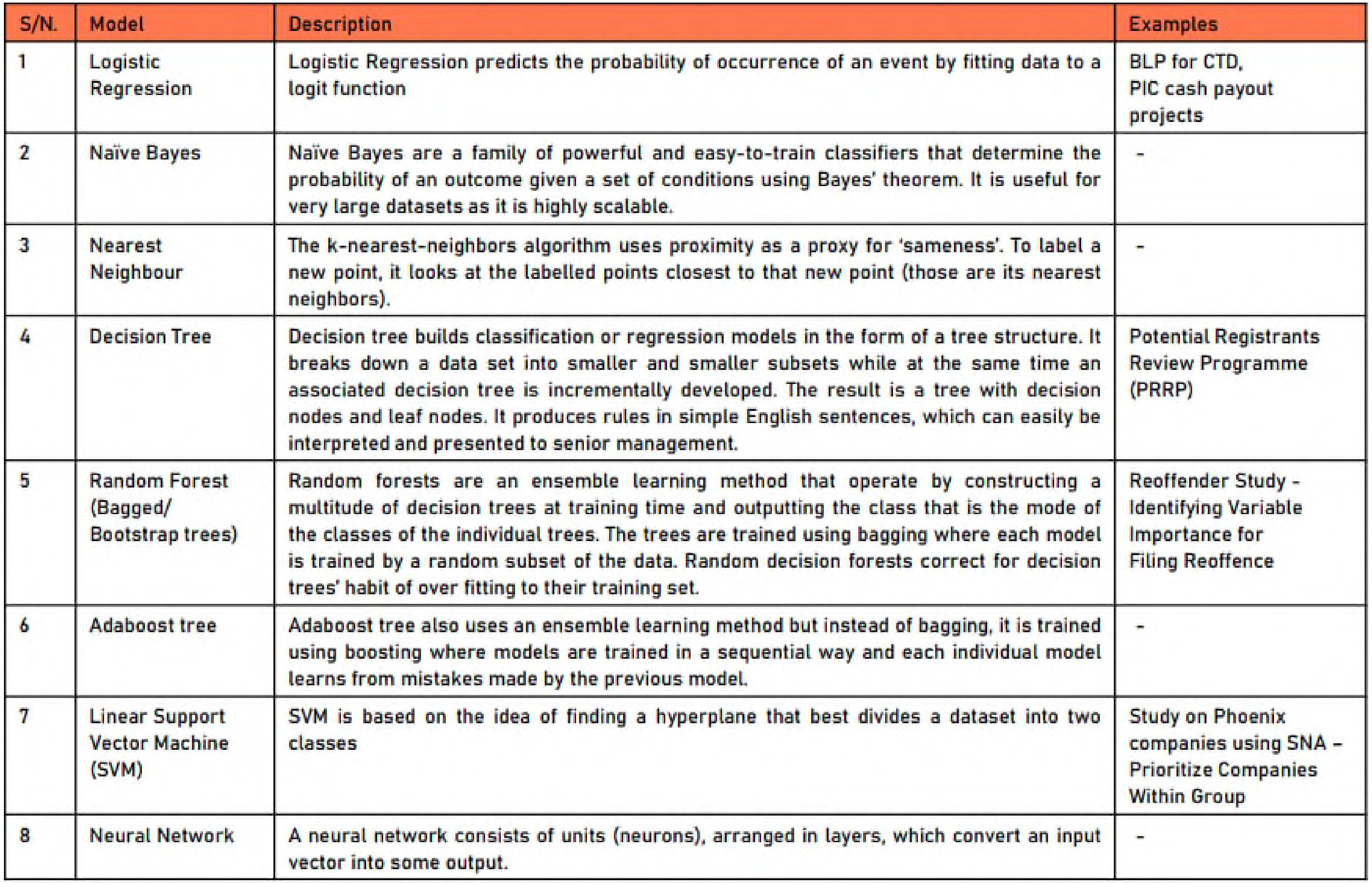
**Source : IRAS' Analytics Toolkit*

Data Analytics Project Process*

Predictive Analytics

To predict a future business phenomenon using supervised machine learning techniques. For example, to anticipate which taxpayers are more likely to default.

Classification Models



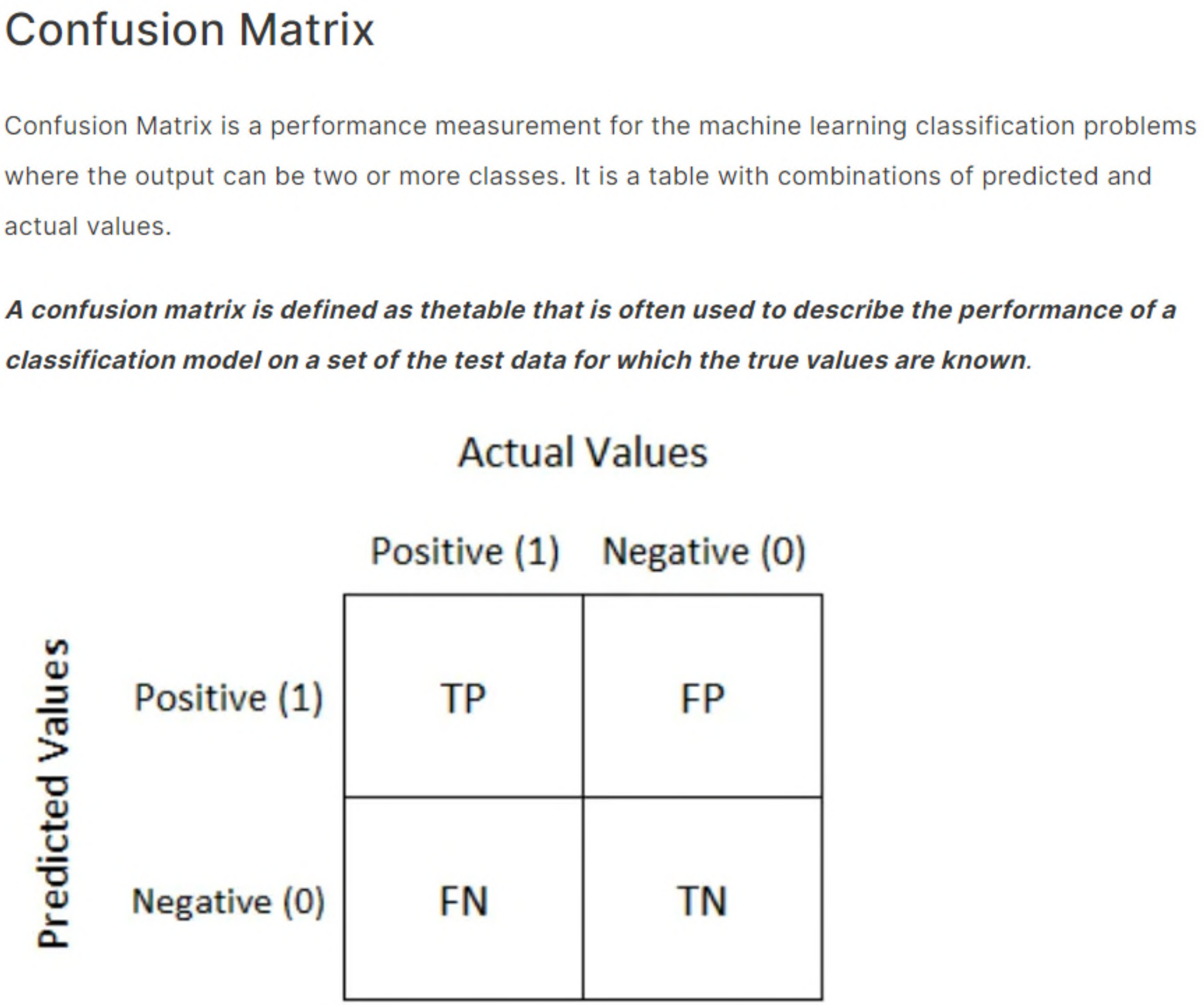
S/N.	Model	Description	Examples
1	Logistic Regression	Logistic Regression predicts the probability of occurrence of an event by fitting data to a logit function	BLP for CTD, PIC cash payout projects
2	Naïve Bayes	Naïve Bayes are a family of powerful and easy-to-train classifiers that determine the probability of an outcome given a set of conditions using Bayes' theorem. It is useful for very large datasets as it is highly scalable.	-
3	Nearest Neighbour	The k-nearest-neighbors algorithm uses proximity as a proxy for 'sameness'. To label a new point, it looks at the labelled points closest to that new point (those are its nearest neighbors).	-
4	Decision Tree	Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. It produces rules in simple English sentences, which can easily be interpreted and presented to senior management.	Potential Registrants Review Programme (PRRP)
5	Random Forest (Bagged/ Bootstrap trees)	Random forests are an ensemble learning method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. The trees are trained using bagging where each model is trained by a random subset of the data. Random decision forests correct for decision trees' habit of over fitting to their training set.	Reoffender Study - Identifying Variable Importance for Filing Reoffence
6	Adaboost tree	Adaboost tree also uses an ensemble learning method but instead of bagging, it is trained using boosting where models are trained in a sequential way and each individual model learns from mistakes made by the previous model.	-
7	Linear Support Vector Machine (SVM)	SVM is based on the idea of finding a hyperplane that best divides a dataset into two classes	Study on Phoenix companies using SNA – Prioritize Companies Within Group
8	Neural Network	A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output.	-

**Source : IRAS' Analytics Toolkit*

Data Analytics Project Process*

Predictive Analytics
To predict a future business phenomenon using supervised machine learning techniques. For example, to anticipate which taxpayers are more likely to default.

Classification Models



Confusion Matrix

Confusion Matrix is a performance measurement for the machine learning classification problems where the output can be two or more classes. It is a table with combinations of predicted and actual values.

A confusion matrix is defined as thetable that is often used to describe the performance of a classification model on a set of the test data for which the true values are known.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

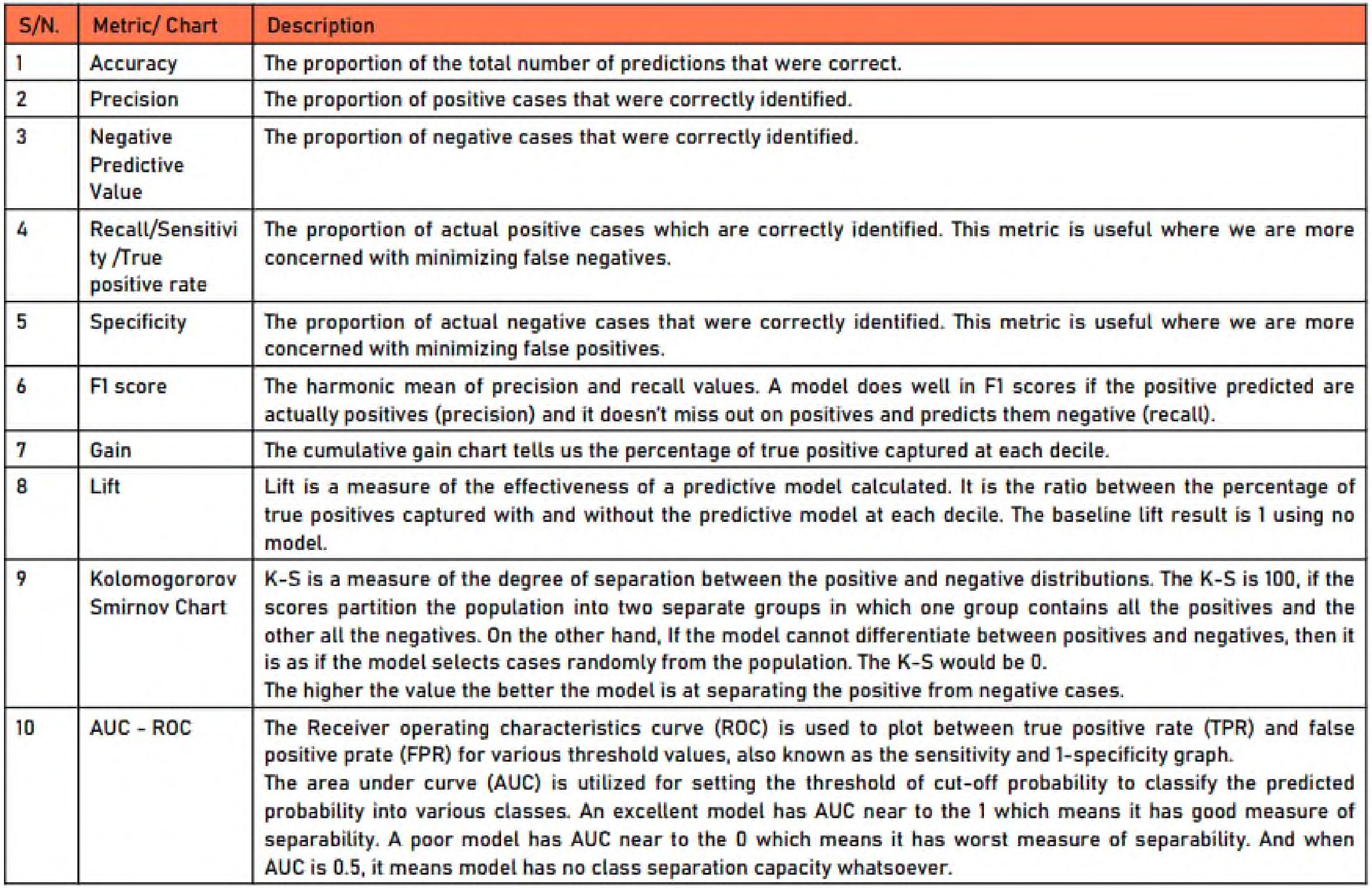
**Source : https://www.analyticsvidhya.com/blo 21/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/*

Data Analytics Project Process*

Predictive Analytics

To predict a future business phenomenon using supervised machine learning techniques. For example, to anticipate which taxpayers are more likely to default.

Classification Models



S/N.	Metric/ Chart	Description
1	Accuracy	The proportion of the total number of predictions that were correct.
2	Precision	The proportion of positive cases that were correctly identified.
3	Negative Predictive Value	The proportion of negative cases that were correctly identified.
4	Recall/Sensitivity /True positive rate	The proportion of actual positive cases which are correctly identified. This metric is useful where we are more concerned with minimizing false negatives.
5	Specificity	The proportion of actual negative cases that were correctly identified. This metric is useful where we are more concerned with minimizing false positives.
6	F1 score	The harmonic mean of precision and recall values. A model does well in F1 scores if the positive predicted are actually positives (precision) and it doesn't miss out on positives and predicts them negative (recall).
7	Gain	The cumulative gain chart tells us the percentage of true positive captured at each decile.
8	Lift	Lift is a measure of the effectiveness of a predictive model calculated. It is the ratio between the percentage of true positives captured with and without the predictive model at each decile. The baseline lift result is 1 using no model.
9	Kolomogororov Smirnov Chart	K-S is a measure of the degree of separation between the positive and negative distributions. The K-S is 100, if the scores partition the population into two separate groups in which one group contains all the positives and the other all the negatives. On the other hand, If the model cannot differentiate between positives and negatives, then it is as if the model selects cases randomly from the population. The K-S would be 0. The higher the value the better the model is at separating the positive from negative cases.
10	AUC - ROC	The Receiver operating characteristics curve (ROC) is used to plot between true positive rate (TPR) and false positive prate (FPR) for various threshold values, also known as the sensitivity and 1-specificity graph. The area under curve (AUC) is utilized for setting the threshold of cut-off probability to classify the predicted probability into various classes. An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

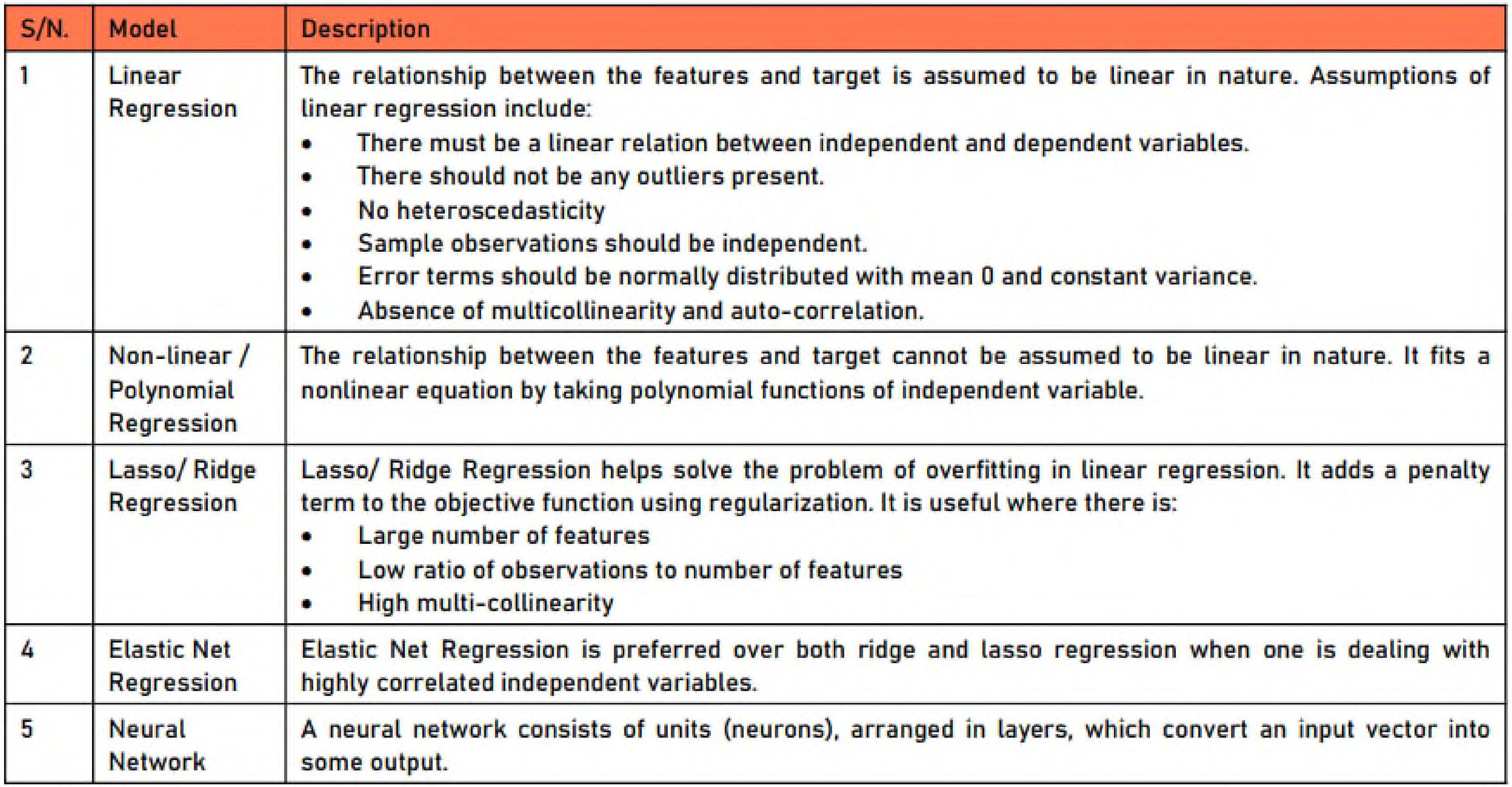
**Source : IRAS' Analytics Toolkit*

Data Analytics Project Process*

Predictive Analytics

To predict a future business phenomenon using supervised machine learning techniques. For example, to anticipate which taxpayers are more likely to default.

Regression Models



S/N.	Model	Description
1	Linear Regression	The relationship between the features and target is assumed to be linear in nature. Assumptions of linear regression include: • There must be a linear relation between independent and dependent variables. • There should not be any outliers present. • No heteroscedasticity • Sample observations should be independent. • Error terms should be normally distributed with mean 0 and constant variance. • Absence of multicollinearity and auto-correlation.
2	Non-linear / Polynomial Regression	The relationship between the features and target cannot be assumed to be linear in nature. It fits a nonlinear equation by taking polynomial functions of independent variable.
3	Lasso/ Ridge Regression	Lasso/ Ridge Regression helps solve the problem of overfitting in linear regression. It adds a penalty term to the objective function using regularization. It is useful where there is: • Large number of features • Low ratio of observations to number of features • High multi-collinearity
4	Elastic Net Regression	Elastic Net Regression is preferred over both ridge and lasso regression when one is dealing with highly correlated independent variables.
5	Neural Network	A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output.

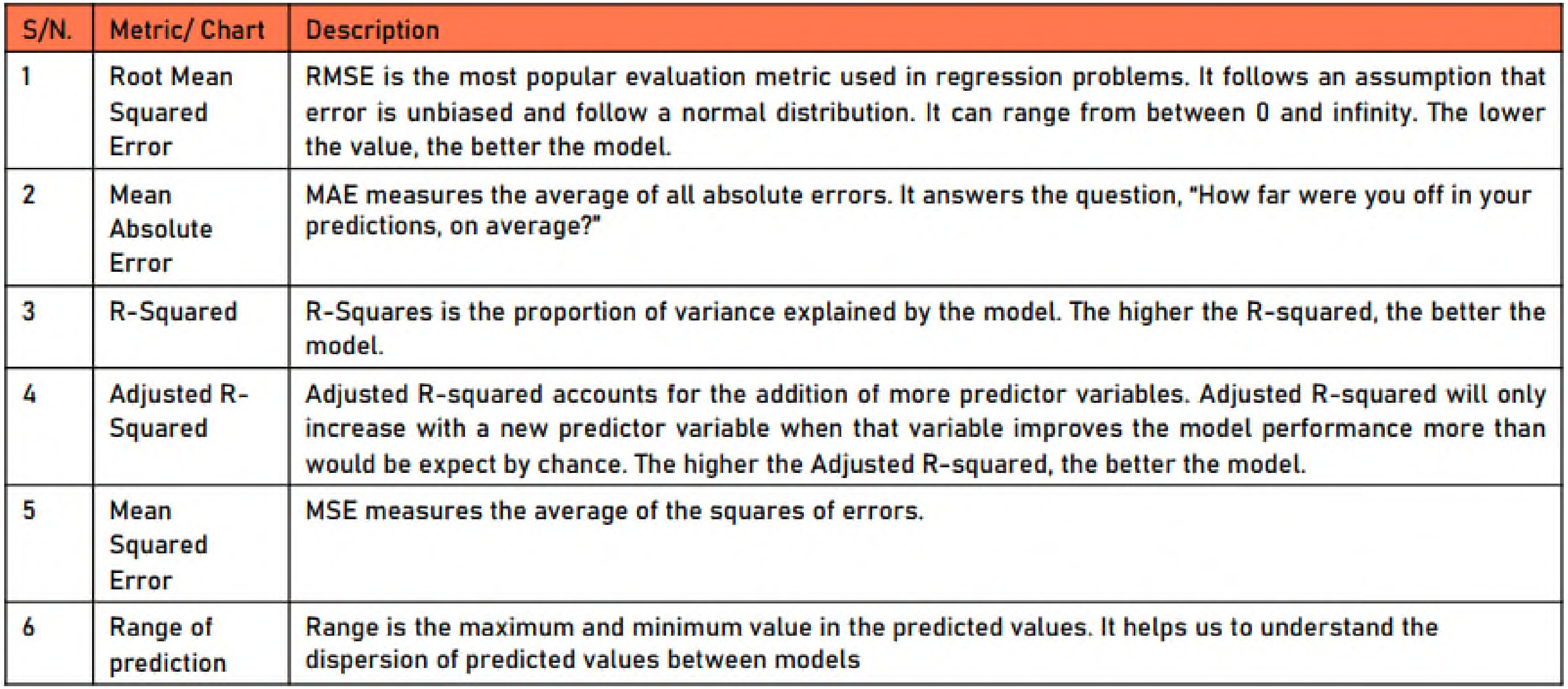
**Source : IRAS' Analytics Toolkit*

Data Analytics Project Process*

Predictive Analytics

To predict a future business phenomenon using supervised machine learning techniques. For example, to anticipate which taxpayers are more likely to default.

Regression Models



S/N.	Metric/ Chart	Description
1	Root Mean Squared Error	RMSE is the most popular evaluation metric used in regression problems. It follows an assumption that error is unbiased and follow a normal distribution. It can range from between 0 and infinity. The lower the value, the better the model.
2	Mean Absolute Error	MAE measures the average of all absolute errors. It answers the question, "How far were you off in your predictions, on average?"
3	R-Squared	R-Squares is the proportion of variance explained by the model. The higher the R-squared, the better the model.
4	Adjusted R-Squared	Adjusted R-squared accounts for the addition of more predictor variables. Adjusted R-squared will only increase with a new predictor variable when that variable improves the model performance more than would be expect by chance. The higher the Adjusted R-squared, the better the model.
5	Mean Squared Error	MSE measures the average of the squares of errors.
6	Range of prediction	Range is the maximum and minimum value in the predicted values. It helps us to understand the dispersion of predicted values between models

**Source : IRAS' Analytics Toolkit*

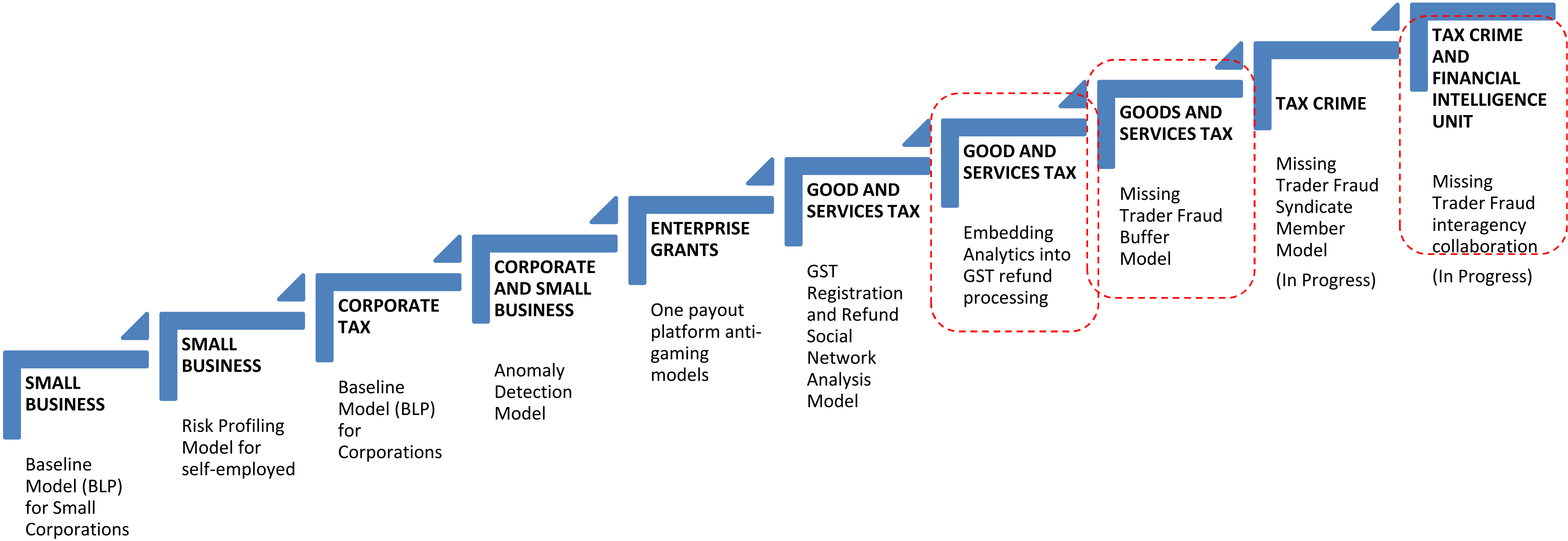
Leveraging on technology & analytics in financial investigation

Analytics as a gatekeeper

Over the years, IRAS has deepened our analytics capability where various analytics tools were developed to detect crimes proactively

Analytics as a gatekeeper - Overview

Tax Fraud / Evasion



SMALL BUSINESS

Baseline Model (BLP) for Small Corporations

SMALL BUSINESS

Risk Profiling Model for self-employed

CORPORATE TAX

Baseline Model (BLP) for Corporations

CORPORATE AND SMALL BUSINESS

Anomaly Detection Model

ENTERPRISE GRANTS

One payout platform anti-gaming models

GOOD AND SERVICES TAX

GST Registration and Refund Social Network Analysis Model

GOOD AND SERVICES TAX

Embedding Analytics into GST refund processing

GOODS AND SERVICES TAX

Missing Trader Fraud Buffer Model

TAX CRIME

Missing Trader Fraud Syndicate Member Model

(In Progress)

TAX CRIME AND FINANCIAL INTELLIGENCE UNIT

Missing Trader Fraud interagency collaboration

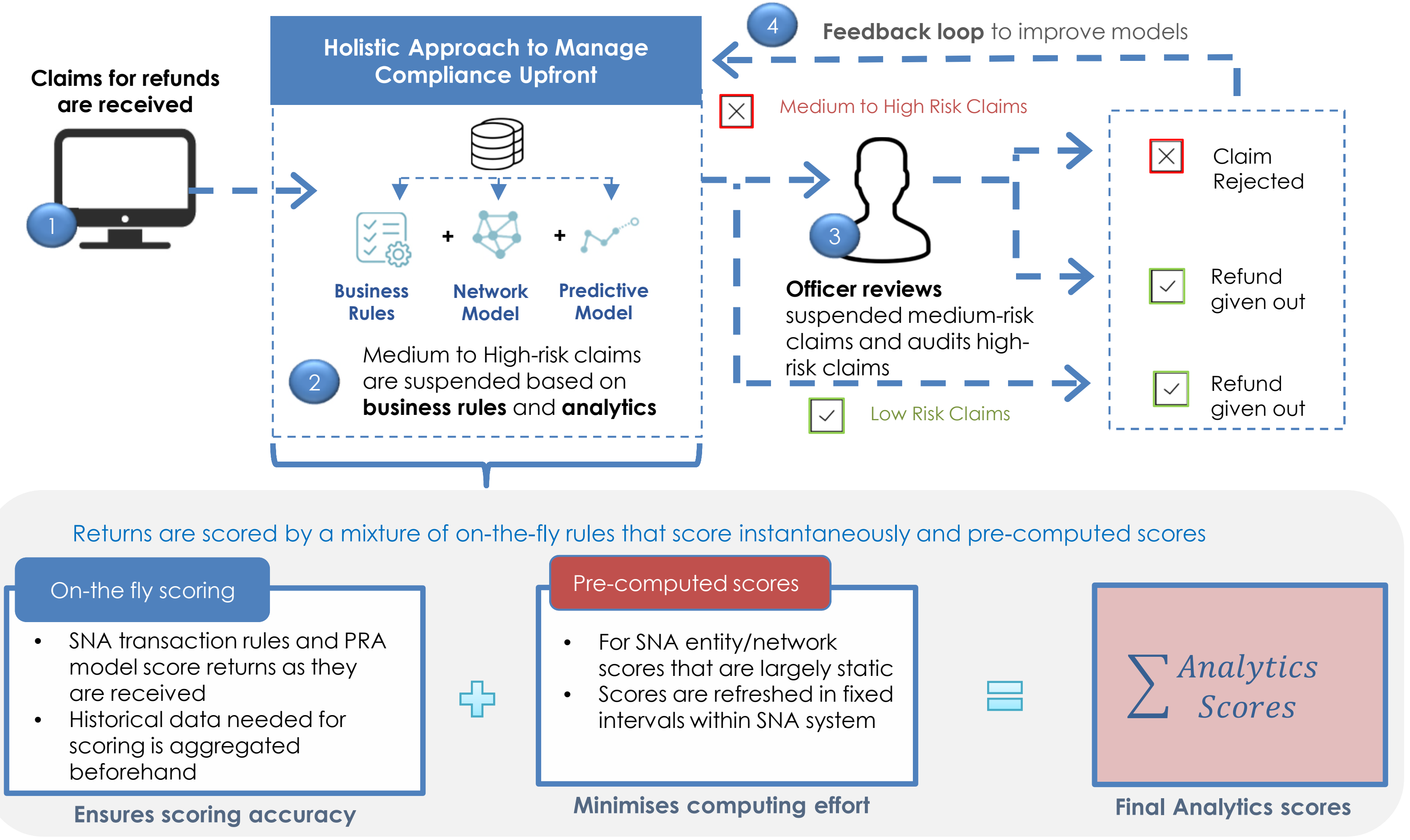
(In Progress)

Tax Non-Compliance

Embedding Analytics into GST Refund Process

Embedding of **2 analytics models** within the GST Refund Process to identify risky refunds:

- ➢Pre-Refund Analytics (PRA) Model
- ➢Social Network Analysis (SNA) Model



Claims for refunds are received

1

Holistic Approach to Manage Compliance Upfront

Business Rules + **Network Model** + **Predictive Model**

2 Medium to High-risk claims are suspended based on **business rules** and **analytics**

☒ Medium to High Risk Claims

3 **Officer reviews** suspended medium-risk claims and audits high-risk claims

✓ Low Risk Claims

☒ Claim Rejected

✓ Refund given out

✓ Refund given out

4 **Feedback loop** to improve models

Returns are scored by a mixture of on-the-fly rules that score instantaneously and pre-computed scores

On-the fly scoring

- SNA transaction rules and PRA model score returns as they are received
- Historical data needed for scoring is aggregated beforehand

Ensures scoring accuracy

+

Pre-computed scores

- For SNA entity/network scores that are largely static
- Scores are refreshed in fixed intervals within SNA system

Minimises computing effort

=

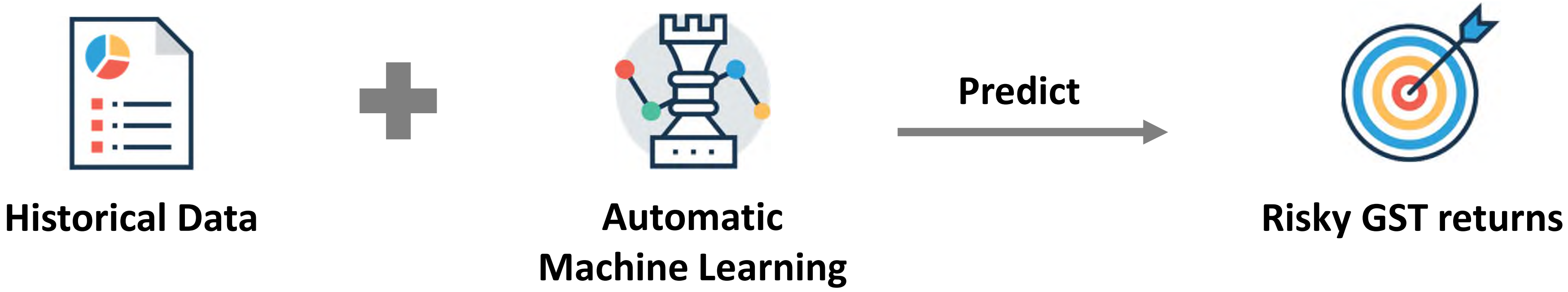
$\sum$ *Analytics Scores*

Final Analytics scores

Missing Trader Fraud (“MTF”) Buffer Model

The MTF Buffer Model augments IRAS’ analytical capabilities with the use of an Auto Machine Learning solution to aid our fight against Missing Trader Fraud

- Enable IRAS to pre-emptively detect buffer entities engaging in MTF activity and **disrupt the supply chain early**
- **AutoML model** trained on known buffers and non-buffers
- Scoring is performed on GST returns to generate a score of **the likelihood that the return displays buffer-like attributes.**



Historical Data

Automatic Machine Learning

Predict

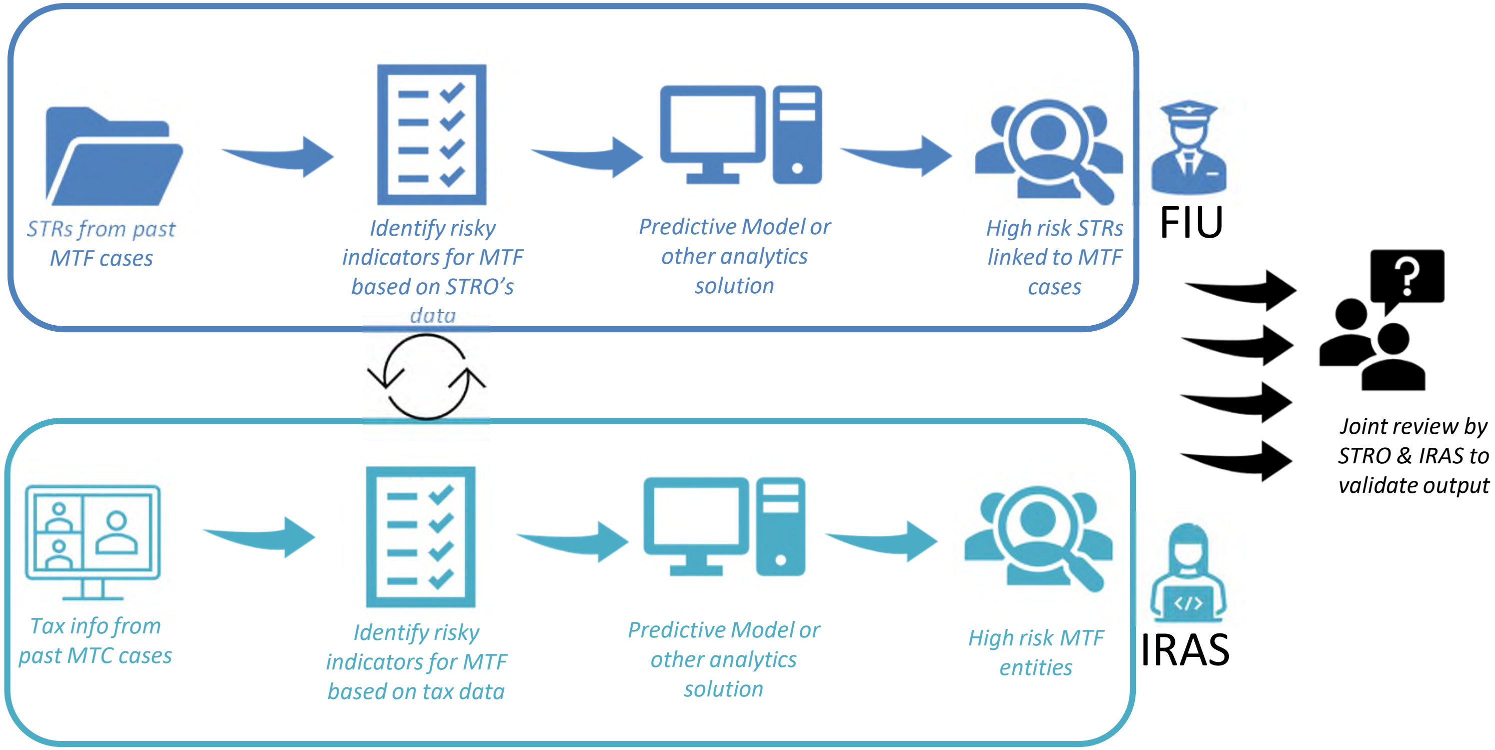
Risky GST returns

Results

Model was **3 times more effective** in detecting MTF buffer entities compared to traditional approaches, allowing IRAS to detect **>$30m of potential tax fraud**.

IRAS / FIU Interagency Analytics Collaboration

The IRAS-FIU interagency analytics collaboration adopts a federated data analytics approach to better identify Missing Trader Fraud ("MTF") in IRAS and STRO's respective systems


STRs from past MTF cases
Identify risky indicators for MTF based on STRO's data
Predictive Model or other analytics solution
High risk STRs linked to MTF cases
FIU
Joint review by STRO & IRAS to validate output
Tax info from past MTC cases
Identify risky indicators for MTF based on tax data
Predictive Model or other analytics solution
High risk MTF entities
IRAS

Progress to date

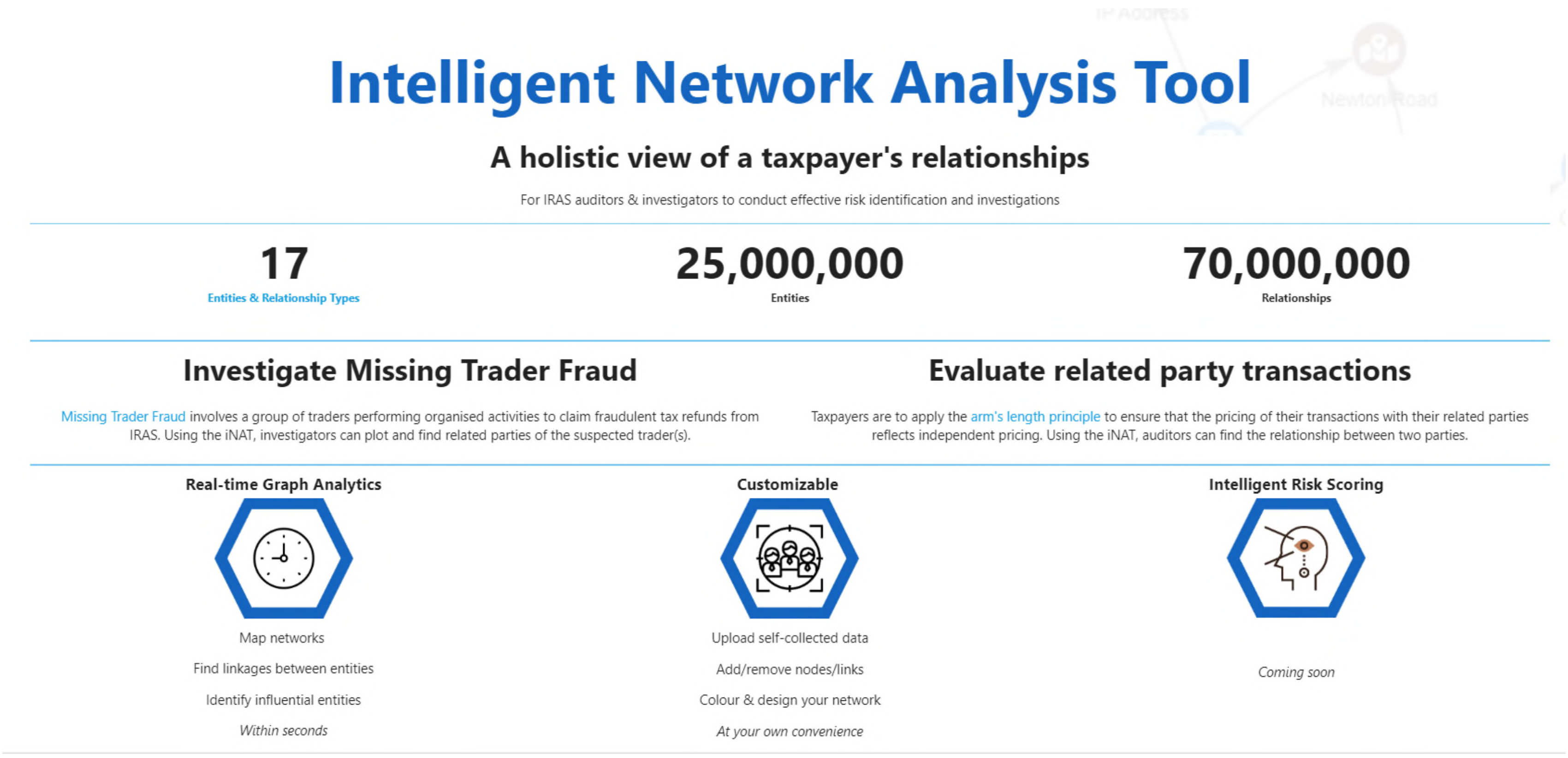
- Collaboration with our Financial Intelligence Unit ("FIU") from the Singapore Police Force to develop analytics models using past STRs filed / tax info from MTF entities to identify unknown MTF cases / clusters.
- Currently, project is at the model deployment stage where both agencies will analyse the cases generated to evaluate the model effectiveness.

Analytics for operational analysis

IRAS employs analytics to amplify intelligence for better sense-making and targeting

Enterprise Network Visualising Tool - Intelligent Network Analysis Tool ("iNAT")

iNAT helps officers easily connect suspicious entities, even if they are linked by a complicated network of entity ownership, business transactions, and personal relations over multiple hops.



Intelligent Network Analysis Tool

A holistic view of a taxpayer's relationships

For IRAS auditors & investigators to conduct effective risk identification and investigations

17
Entities & Relationship Types

25,000,000
Entities

70,000,000
Relationships

Investigate Missing Trader Fraud

Missing Trader Fraud involves a group of traders performing organised activities to claim fraudulent tax refunds from IRAS. Using the iNAT, investigators can plot and find related parties of the suspected trader(s).

Evaluate related party transactions

Taxpayers are to apply the arm's length principle to ensure that the pricing of their transactions with their related parties reflects independent pricing. Using the iNAT, auditors can find the relationship between two parties.

Real-time Graph Analytics

Map networks

Find linkages between entities

Identify influential entities

Within seconds

Customizable

Upload self-collected data

Add/remove nodes/links

Colour & design your network

At your own convenience

Intelligent Risk Scoring

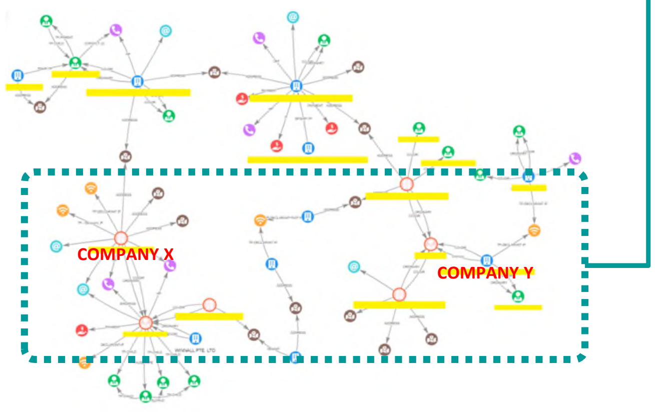
Coming soon

Enterprise Network Visualising Tool - Intelligent Network Analysis Tool (“iNAT”)

iNAT helps officers easily connect suspicious entities, even if they are linked by a complicated network of entity ownership, business transactions, and personal relations over multiple hops.

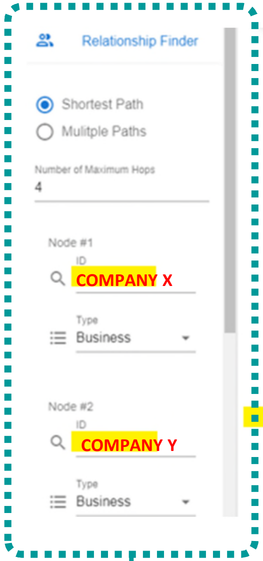
Example of iNAT applied to Missing Trader Fraud (“MTF”) investigations

1 **X** and **Y** are missing traders in a Missing Trader Fraud scheme that involved **$114 Million in fictitious sales** and over **70 entities**.



Part of MTF network involving X and Y

2 With iNAT’s Relationship Finder feature, IOs can access rich **relational databases** and visualize the links between X and Y



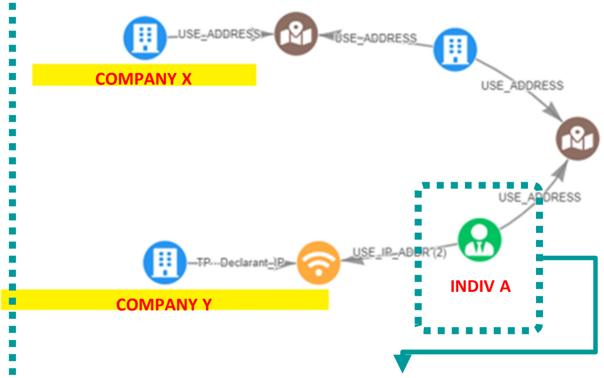
Relationship Finder

● Shortest Path
○ Mulitple Paths

Number of Maximum Hops
4

Node #1
ID
COMPANY X
Type
Business

Node #2
ID
COMPANY Y
Type
Business



3 **A** is quickly uncovered as a suspicious individual who shares location and IP address with X and Y. **IOs can expand targeted investigations on A.**

Enterprise Network Visualising Tool - Intelligent Network Analysis Tool (“iNAT”)

iNAT helps officers easily connect suspicious entities, even if they are linked by a complicated network of entity ownership, business transactions, and personal relations over multiple hops.

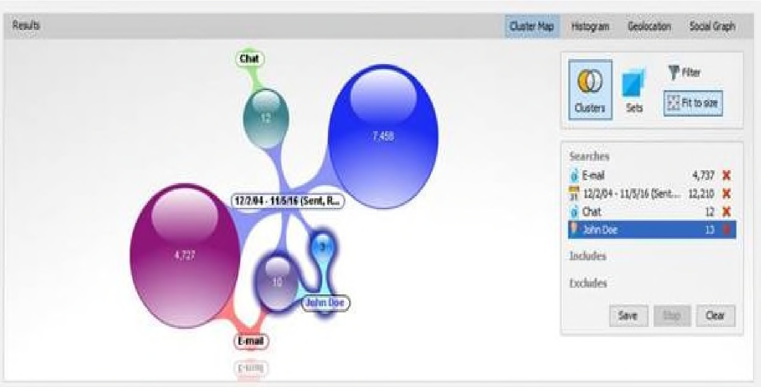
Inhouse iNAT Training Video

Analytics for investigative efficiency

Democratising analytics to all officers for greater speed and effectiveness of casework during investigation

e-Discovery : Self-review Workstation

Investigation and eDiscovery software tool for single users who need to process, search, filter and produce amounts of digital evidence.

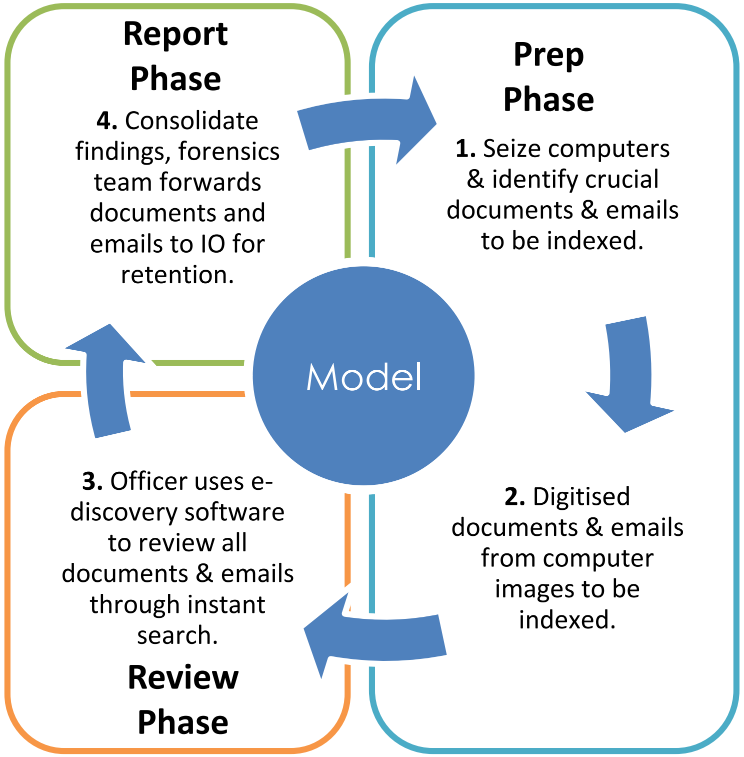


Searches

E-mail	4,737
12/2/04 - 11/5/16 (Sent...	12,210
Chat	12
John Doe	13

Includes

Excludes



Report Phase

4. Consolidate findings, forensics team forwards documents and emails to IO for retention.

Prep Phase

1. Seize computers & identify crucial documents & emails to be indexed.

Model

2. Digitised documents & emails from computer images to be indexed.

3. Officer uses e-discovery software to review all documents & emails through instant search.

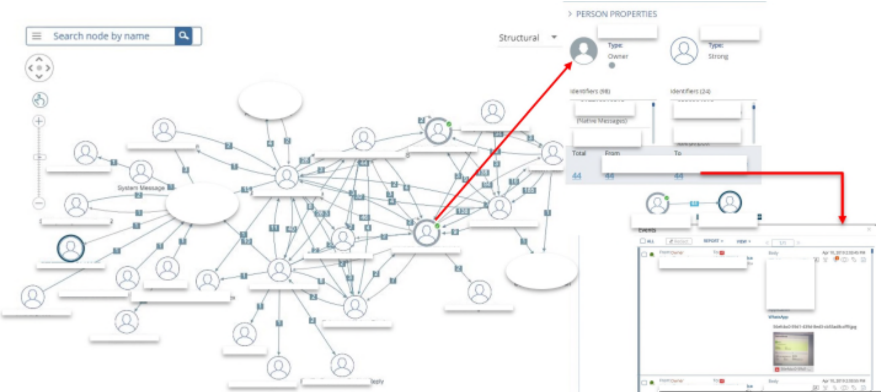
Review Phase

Benefits

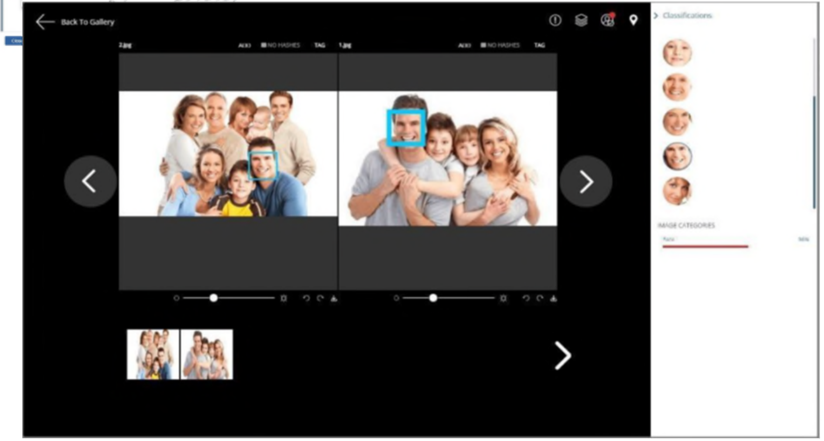
Users can quickly and easily process, search, review, and analyze digital evidence obtained. The easy-to-use interface and Cluster Map allows analysts to quickly find critical data, visualise relevant relationships, and drill down to the most pertinent data.

Mobile Evidence Analytics

Allows multiple mobile devices to be reviewed simultaneously and provides a contextual view of the relationships between devices



Relationship/Graph View



Face/Image Similarity Recognition

Benefits

- Enable multiple users to access at once to look at different facets of a syndicated case
- Other capabilities like facial recognition also allowed users to identify new subjects / relationships

Automated Data Extraction project

Improve productivity of users when reviewing case by automating some of the data extraction process from IRAS' various databases and charts / tables are provided for ease of analysis

Close to **300** data fields were extracted from various databases

*(> **150** data fields for business & > **100** data fields for individual)*

> **10 Key financial ratios** (e.g. Industry Gross Profit Margin) & network analysis are incorporated to supplement background check

Up to **30%** of First Level Report ("FLR" – report submitted by Intel Analyst) is pre-filled with requisite information from extracted data

> **10 charts and tables** are created using POWER BI to facilitate user's review and analysis

Automated Data Extraction project

Key Benefits

Data from IRAS database is extracted automatically by running a code

Report template is pre-filled with information extracted via DW

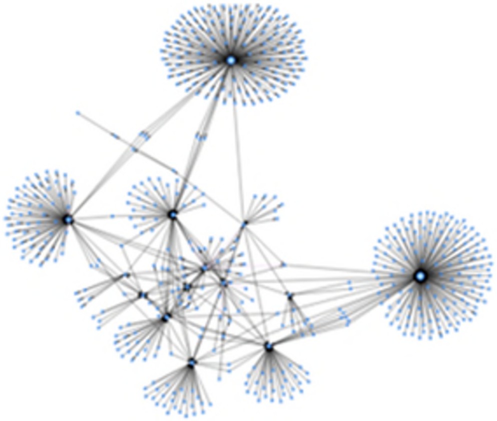
The report template is linked to POWER BI for data visualization

Productivity gained from manual to automated data extraction process (1-3 mandays to 1-2 hours)

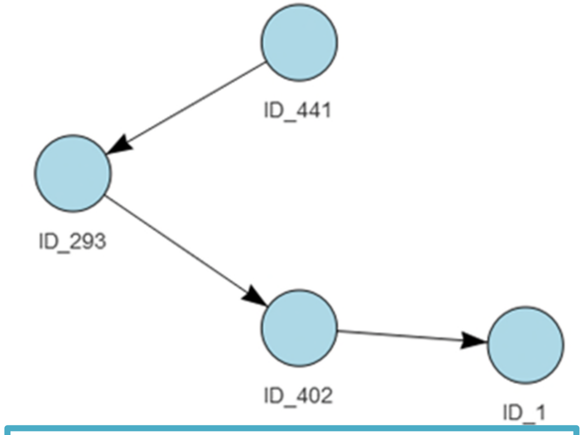
Transaction Tracing Tool (“TTT”)

Facilitate mapping of the supply chains for syndicated cases by providing officers with the entire flow of transactions based on a selected transaction

- Some Missing Trader Fraud ("MTF") clusters involved **large number of entities** (up to 50 entities per cluster) with **high transactional volume** (may be more than 10,000 transactions per cluster).
- After ingesting the transactional data, TTT will provide the supply chain by mapping out the predicted transaction flow for the users



Network Visualisation of transactions



Predicted Transaction Flow in Supply Chain

Transaction Tracing Tool ("TTT")

Key Benefits

Productivity gained by analysing large amount of transactions effectively – can quickly identify the fraudulent supply chains

- Saved effort of reading and matching invoice no & date, long text descriptions, quantity of goods
- Saved time from manually drawing supply chains

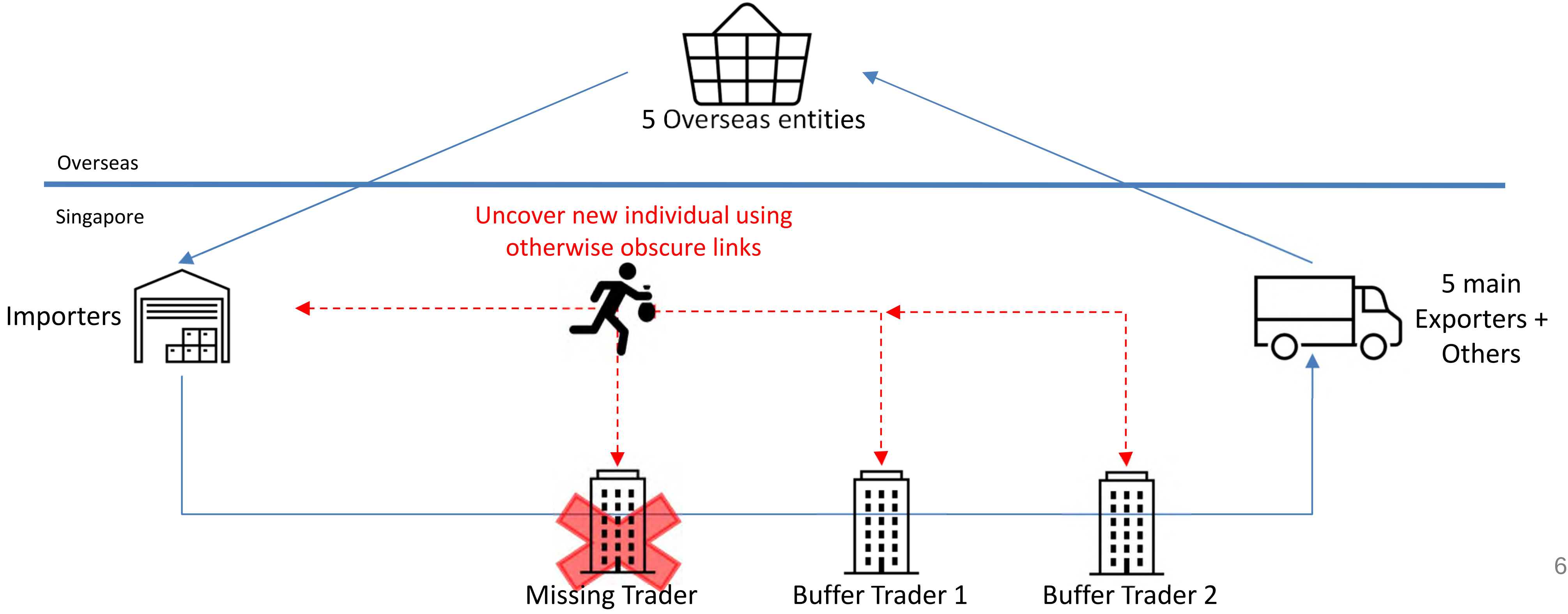
Officers can complete the tracing within 6.5 to 9.5 manhours by using TTT compared to 9.5 to 12.5 manhours when tracing transactions manually

Case Studies

Case Sharing – Using Network Tool to Link Entities

Operation Wand

- Suspected GST Refund Fraud case
- More than 20 entities / individuals in cluster
- Complex network and transactions



Large Language Models (LLMs)

Advisory from the Prime Minister Office ("PMO") on use of LLMs for Government Agencies

Inherent limitations and risks in the use of LLMs

Accuracy & Accountability

Risks due to the tendency of LLMs to produce convincing yet inaccurate output ("hallucinations") or provide inappropriate responses.

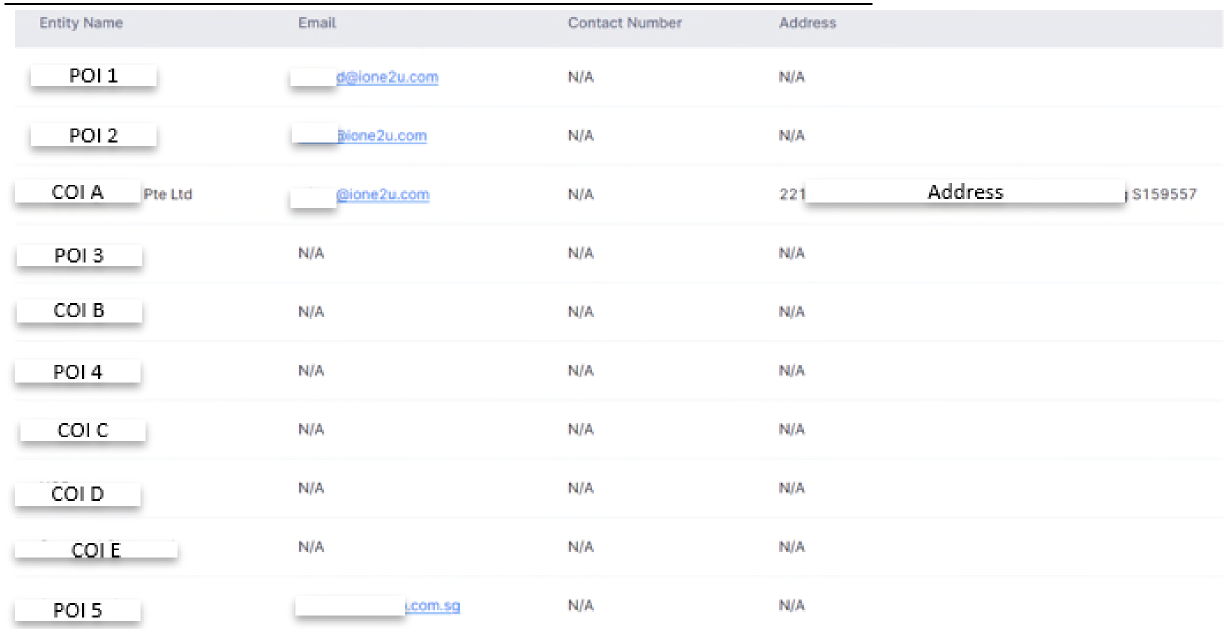
Security

Risks related to the potential compromise of Government data and networks. LLM providers may log data, often in overseas servers, which increases the risk of sensitive data being leaked if included in the user's prompt.

IRAS' Experiment of Large Language Models (LLMs)

IRAS are currently limiting the exploration of LLMs in the domain of productivity aid and information retrieval while managing the risks involved. One potential use case for investigation is to use LLM to extract key names mentioned in statement / document.

Extract Key Names mentioned in the Statement



Entity Name	Email	Contact Number	Address
POI 1	[redacted]d@ione2u.com	N/A	N/A
POI 2	[redacted]@ione2u.com	N/A	N/A
COI A Pte Ltd	[redacted]@ione2u.com	N/A	221 Address S159557
POI 3	N/A	N/A	N/A
COI B	N/A	N/A	N/A
POI 4	N/A	N/A	N/A
COI C	N/A	N/A	N/A
COI D	N/A	N/A	N/A
COI E	N/A	N/A	N/A
POI 5	[redacted].com.sg	N/A	N/A

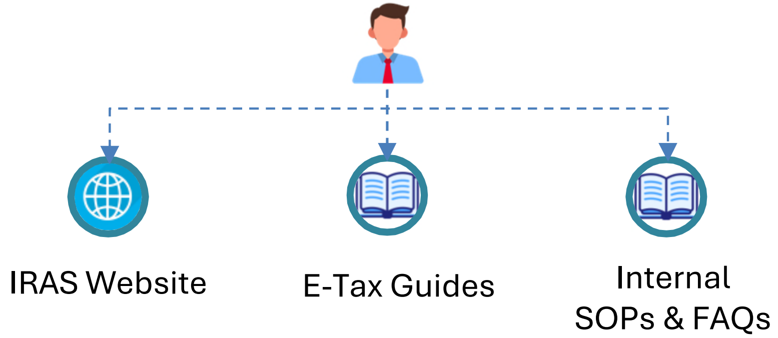
- **Pair Chat** is a free, fast and secure version of ChatGPT, and is currently available and free to use for all public officers. It is currently powered by the same Large Language Model underlying ChatGPT.
- Pair Chat has been cleared by the Smart Nation and Digital Government Office (SNDGO) for public officer use (up to certain data classification).
- Depending on the prompt, the no. of extracted entities extracted differ.

IRAS' Experiment of Large Language Models (LLMs)

- IRASearch (Beta Version)

Challenges

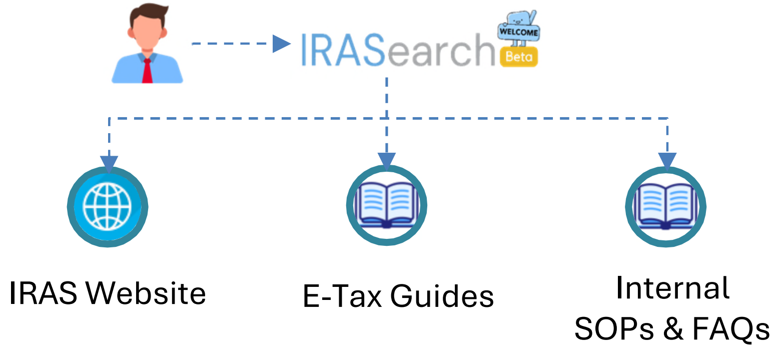
Currently, tax officers need consult multiple sources of information to answer tax inquiries. This could be onerous and time consuming, especially for new tax officers.



Approach

In 2023, IRASearch was developed to assist staff in addressing tax inquiries more effectively. It is an intelligent search engine that leverages Gen AI/LLMs and references authoritative data sources.

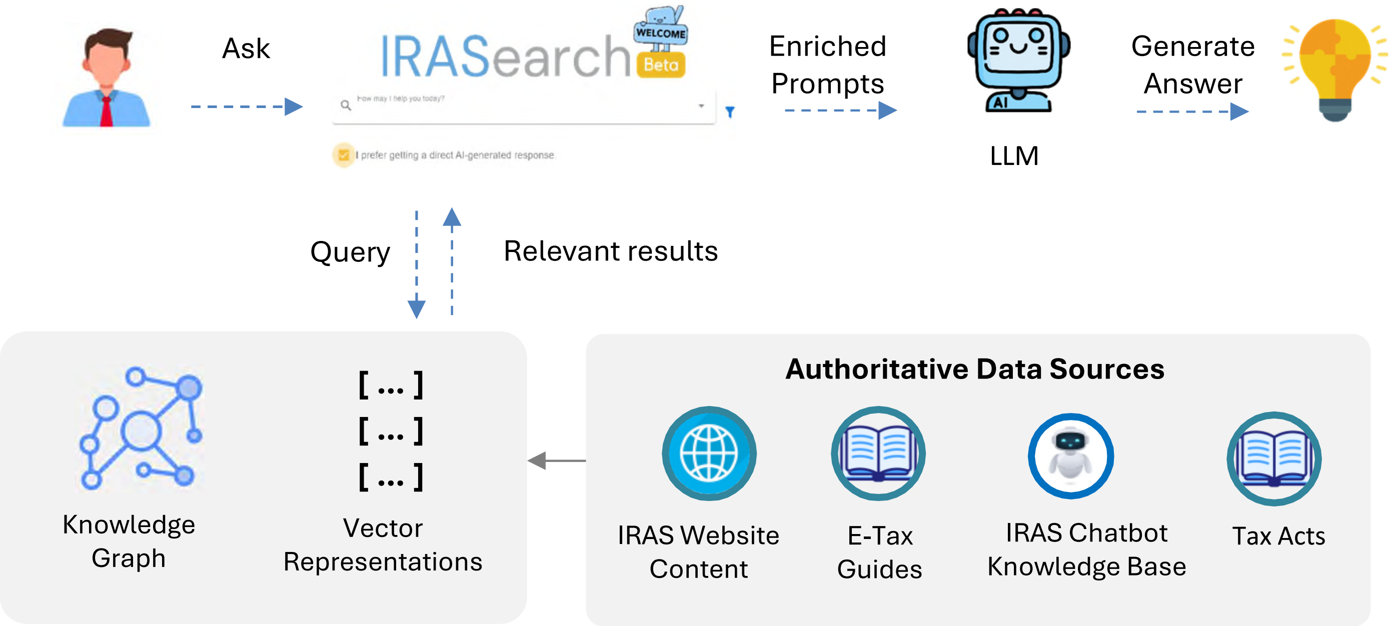
In-House IRASearch



IRAS’ Experiment of Large Language Models (LLMs)

The IRASearch is an in-house search engine and Q&A system that delivers more relevant and reliable responses.

Harnesses the power of Gen AI and LLMs



1 Transforms text queries into intent and enhances search with a hybrid methodology, using keyword search and semantic search.

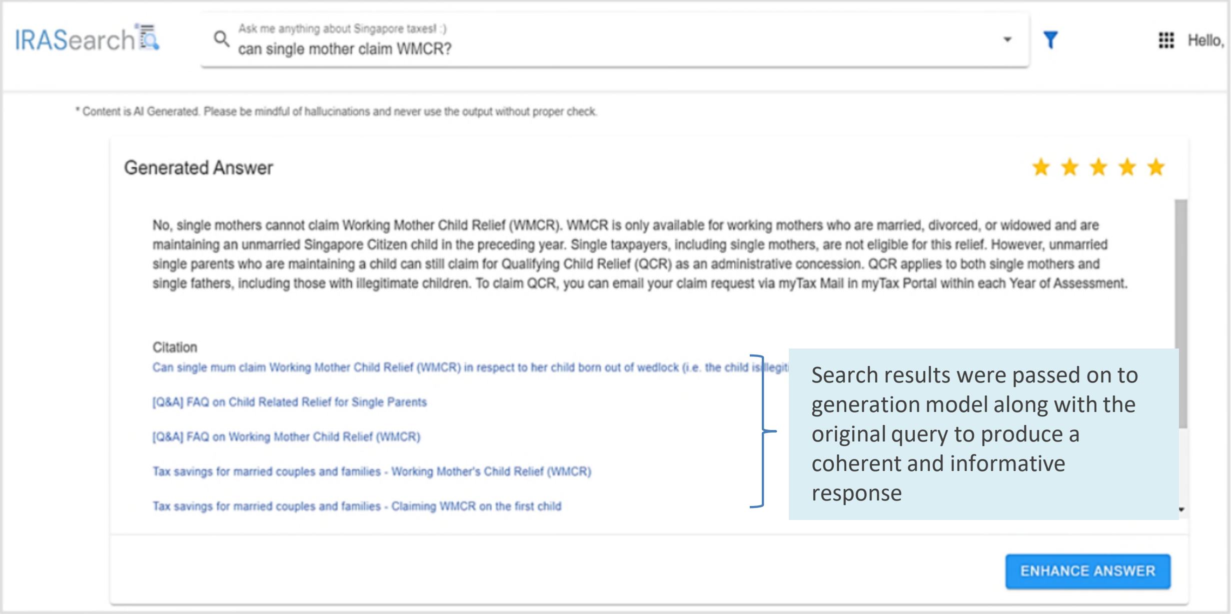
2 Utilises Retrieval-Augmented Generation (RAG) and references authoritative data sources to deliver more reliable and contextually relevant results.

3 Provides customised content generation and assists frontliners with tax inquiries to improve service delivery and operational efficiency.

IRAS’ Experiment of Large Language Models (LLMs)

Pilot testing showed that the Q&A functionality in IRASearch is effective in generating accurate and relevant responses

Integrating Retrieval & Generation Models for Improved Q&A



Ask me anything about Singapore taxes! :)

can single mother claim WMCR?

* Content is AI Generated. Please be mindful of hallucinations and never use the output without proper check.

Generated Answer ★★★★★

No, single mothers cannot claim Working Mother Child Relief (WMCR). WMCR is only available for working mothers who are married, divorced, or widowed and are maintaining an unmarried Singapore Citizen child in the preceding year. Single taxpayers, including single mothers, are not eligible for this relief. However, unmarried single parents who are maintaining a child can still claim for Qualifying Child Relief (QCR) as an administrative concession. QCR applies to both single mothers and single fathers, including those with illegitimate children. To claim QCR, you can email your claim request via myTax Mail in myTax Portal within each Year of Assessment.

Citation

Can single mum claim Working Mother Child Relief (WMCR) in respect to her child born out of wedlock (i.e. the child is illegit

[Q&A] FAQ on Child Related Relief for Single Parents

[Q&A] FAQ on Working Mother Child Relief (WMCR)

Tax savings for married couples and families - Working Mother's Child Relief (WMCR)

Tax savings for married couples and families - Claiming WMCR on the first child

Search results were passed on to generation model along with the original query to produce a coherent and informative response

The Q&A functionality is tailored for tax-related inquiries via customised OpenAI backend system prompts.

94 % **Found Relevant Answers from Generated Responses**

In Sep 2023, pilot testing showed that 94% of users found relevant answers when trying out the new Q&A feature in IRASearch. Enhancements were subsequently made to improve the relevancy and reliability of the AI-generated responses.

IRAS' Experiment of Large Language Models (LLMs)

Deployment of IRASearch for 2024's IIT Filing Season proved useful for newer staff; enhancements needed to increase adoption and productivity gains

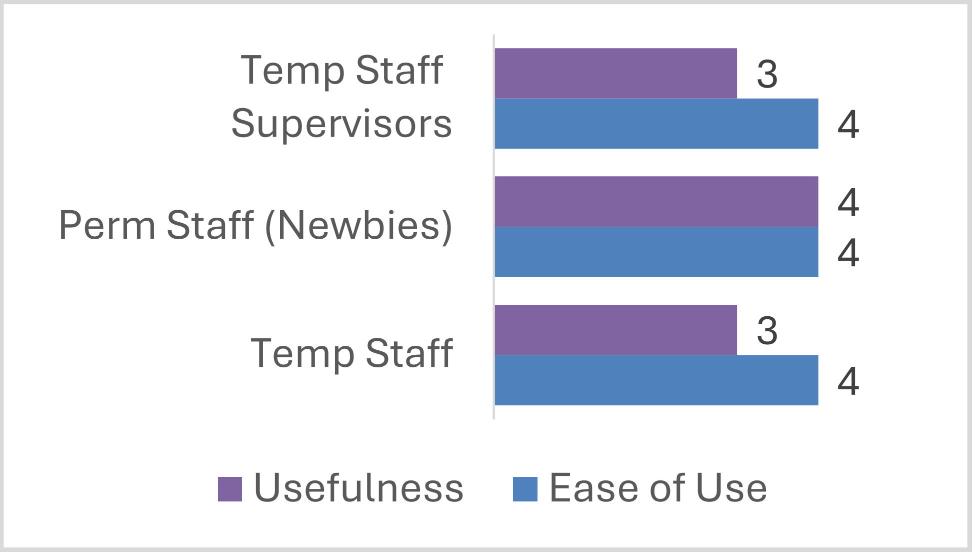
In Feb 2024, we rolled-out IRASearch to a group of 202 frontline staff at Service Experience Centre (SVE), including new perm staff, temp staff for the IIT Filing Season and temp staff supervisors.

Overall, **Perm Staff (Newbies)** found it to be **most useful** (rating 4 out of 5), while most staff found it to be easy to use.

For **experienced frontline staff**, they prefer to **look for their supervisors** as they require end-to-end advice quickly while on the line – the current IRASearch could not handle their enquiries fully due to the lack of data sources covering internal processes and SOPs.

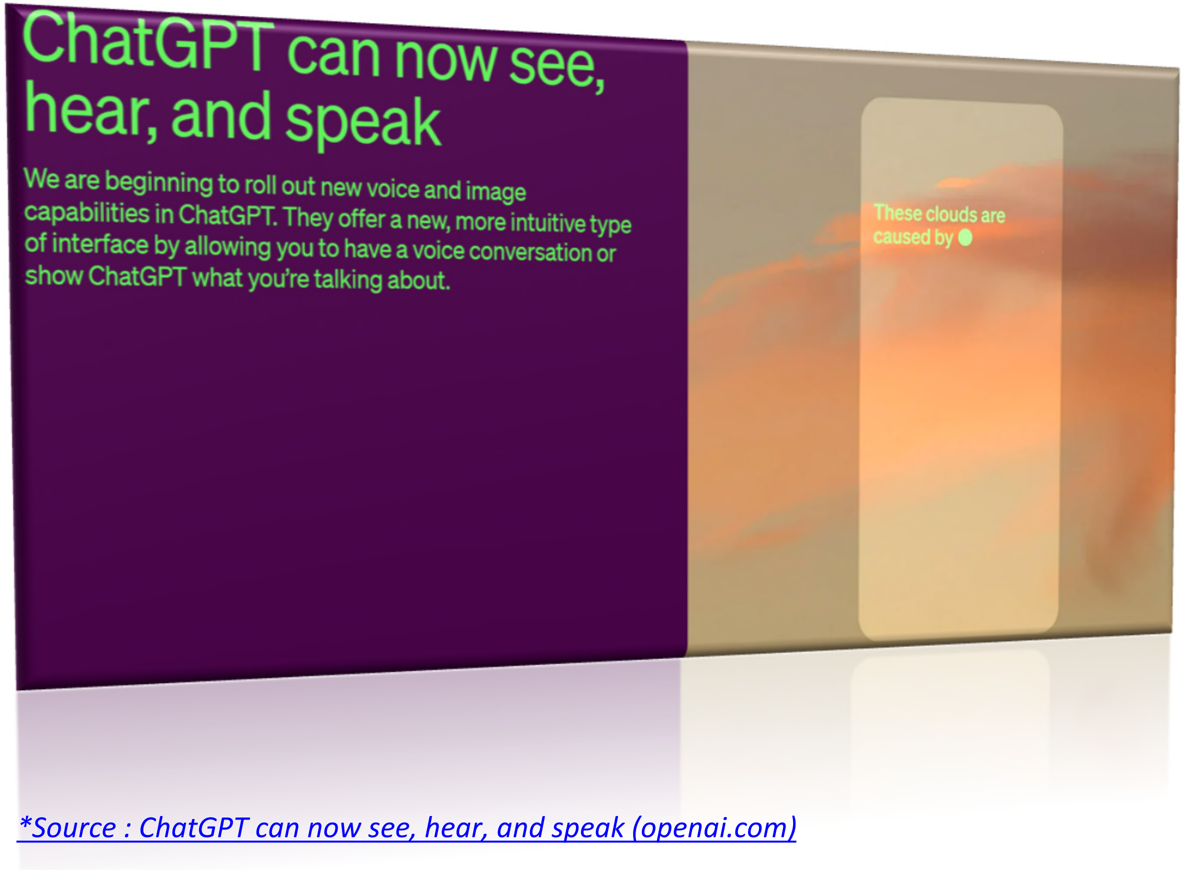
Further enhancement: To Incorporate internal SOPs and other documents into data source

Usefulness & Ease of Use (Average Rating 1 to 5)



New Development in LLM

- New development in LLM*



ChatGPT can now see, hear, and speak

We are beginning to roll out new voice and image capabilities in ChatGPT. They offer a new, more intuitive type of interface by allowing you to have a voice conversation or show ChatGPT what you're talking about.

Snap a picture of a landmark while traveling and have a live conversation about what's interesting about it.

Help your child with a math problem by taking a photo, circling the problem set, and having it share hints with both of you.

Source : ChatGPT can now see, hear, and speak (openai.com)

Country Sharing

Country Sharing (20 mins)

Any AI / analytics tool(s) that have been used by your country to combat tax crime?

- **Group 1 & 2 :** Enhance case detection

- **Group 3 & 4 :** Improve productivity

What’s next

What is next in IRAS AI Strategy

Rich Entity-centric Data Portfolio

Enrich understanding of taxpayers using data/big data

New Data Sources

- To continue to expand our sources of data (e.g. InvoiceNow)
- To tap on new data to refine existing solutions (e.g. fraud detection using bank transaction data) or explore new solutions (e.g. sentiment analysis on audio data)

AI for Everyone

Leverage AI/data science for intelligent tax administration/ enterprise disbursement

Expand AI Applications

- By identifying strategic use-cases with high impact and readiness, to increase the effectiveness and efficiency of our service delivery
- By enhancing and scaling existing AI applications

Future-Ready Data Technology with Optimal Connectivity

Strengthen capabilities to receive, store and use big and real-time data and scale AI solutions

Modernise Data Architecture

- By migrating the UDP to the Cloud to exploit Cloud capabilities and continuously improving it to strengthen AI capabilities
- By exploring and identifying technologies required to enhance governance in the ML/AI development lifecycle

Democratise Data Science and AI Capabilities

Drive pervasive data usage across all staff

Develop Expertise

- By continuously investing in our people and relevant tools
- By exploring partnership opportunities with external stakeholders (e.g. short-term attachment/exchange)

Trusted Data Ecosystem

Use and share AI and data responsibly

AI Ethics & Governance

- By establishing and formalizing a set of principles and measures to inculcate the responsible adoption of AI in IRAS without stifling digital innovation

AI Safety Summit in UK (Nov 2023)

AI Safety Summit (early Nov 2023) at Bletchley Park, in Buckinghamshire

[Extract of BBC article : “Rishi Sunak: AI firms cannot 'mark their own homework'” dated 2 Nov 2023: “The Bletchley Declaration calls for global cooperation on tackling the risks, which include potential breaches to privacy and the displacement of jobs.

*Signed by 28 countries and the EU, it also says AI should be kept "**safe, in such a way as to be human-centric, trustworthy and responsible**".]*

Governance for AI

https://www.bbc.com/news/technology-67285315

Thank You

The information presented in the slides aims to provide a better general understanding of taxpayers' tax obligations and is not intended to comprehensively address all possible tax issues that may arise. This information is correct as at the date of presentation. While every effort has been made to ensure that this information is consistent with existing law and practice, should there be any changes, IRAS reserves the right to vary its position accordingly.