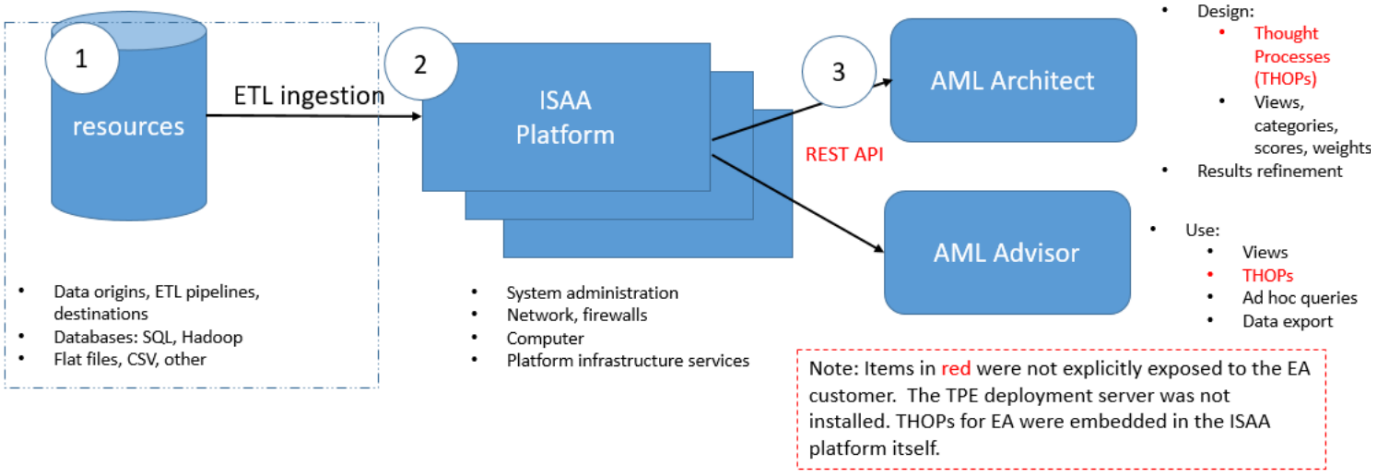
ISAA AML Advisor Guide, EA edition

- Revision history: ISAA AML Advisor Guide, EA edition

ISAA conceptual overview

The Intel Saffron Anti-Money-Laundering Advisor (ISAA) is a cognitive computing system for financial institutions to discover *actionable insights* in to possible crime. Based on systematic analysis of your data, you can tailor your analyses to your data and your needs to progressively refine analyses and improve insights.

Below is a simplified, at-a-glance logical view of the ISAA and its subsystems.



1	2	3
Your data is central to ISAA. A data source is called a *resource*. AML Architects design *pipelines* (which have an *origin* and a *desti nation*) to run data transformations via an *ETL* (Extract, Transfer, Load) process called *ingestion*. During ingestion, data are normalized, sent to a destination, and made available for queries via the AML Advisor for further refinement and investigation.	The ISAA platform is the central hub of the ISAA system. You configure clusters of *leader nodes* and *worker nodes*. You can also setup *zones* and *spaces* to secure containers, segregate your data, and isolate ISAA processes.	AML investigators work with ISAA's web user interfaces: the AML Architect and the AML Advisor. • With AML Architect, you design specific queries (called *views*) and *THOught Processes* (THOPs) for use via the AML Advisor. Types of views include *anomaly views* and *customer risk views*. • With the AML Advisor, users can also create ad hoc queries with LiveSearch.

Notes on Early Adopter deployment

Throughout these guides, specific details bout the ISAA deployment at a customer site are indicated with this marker:

Note on EA deployment

In general, hese notes indicate where the EA deployment varied from the ISAA system design, where the ISAA system itself might have been immature, or where additional manual steps had to be taken to successfully deploy.

Early Adopter (EA) audiences and ISAA documentation

This collection of guides describes the Early Adopter (EA) release of ISAA, which was deployed at a customer site.

The ISAA documentation is grouped into usable collections of information by roles (or personas).

In practice at your site, these roles might be combined. For instance, in test/evaluation, these roles are often a single person.

ISAA AML Advisor Guide

The *ISAA AML Advisor Guide* is for “data explorers”, persons using the AML Advisor web interface to investigate, query, and analyze results in ISAA, results based on the work of data analysts.

Recommended skills

- Curiosity
- Knowledge of AML
- Understanding of your specific goals for AML

ISAA AML Architect Guide

The *ISAA AML Architect Guide* is for the “data analyst” (sometimes called "data scientist" or "programmer") who designs the Extract, Transform, Load (ETL) programming, ingestion, categories, attributes, application of algorithms via THOught Processes (THOPs), and query design. With the AML Architect web interface, ETL tools, and JavaScript programming, the data analyst acts as “power user” in preparing data for use by data explorers via the AML Advisor.

Recommended skills

- Deep knowledge of your data and the desired goals/result of your design
- Comfort with ETL
- Familiarity or prior experience with machine learning
- Knowledge of computer programming with REST APIs and JavaScript
- Experience with StreamSets helpful

ISAA Administrator Guide, with Installation

The *ISAA Administrator Guide, with Installation* is for system administrators involved installing ISAA and third-party software, configuring servers and services, data resources, network design, maintenance and upgrades, and administration of databases, clusters, and all ISAA components.

Recommended skills

- Comfort with Linux operating system
- Familiarity with TCP/IP networks, firewalls, ports
- Familiarity with software installation using tar, gzip, RPM
- Familiarity with LDAP
- General system administration
- Familiarity with Hadoop-based systems helpful

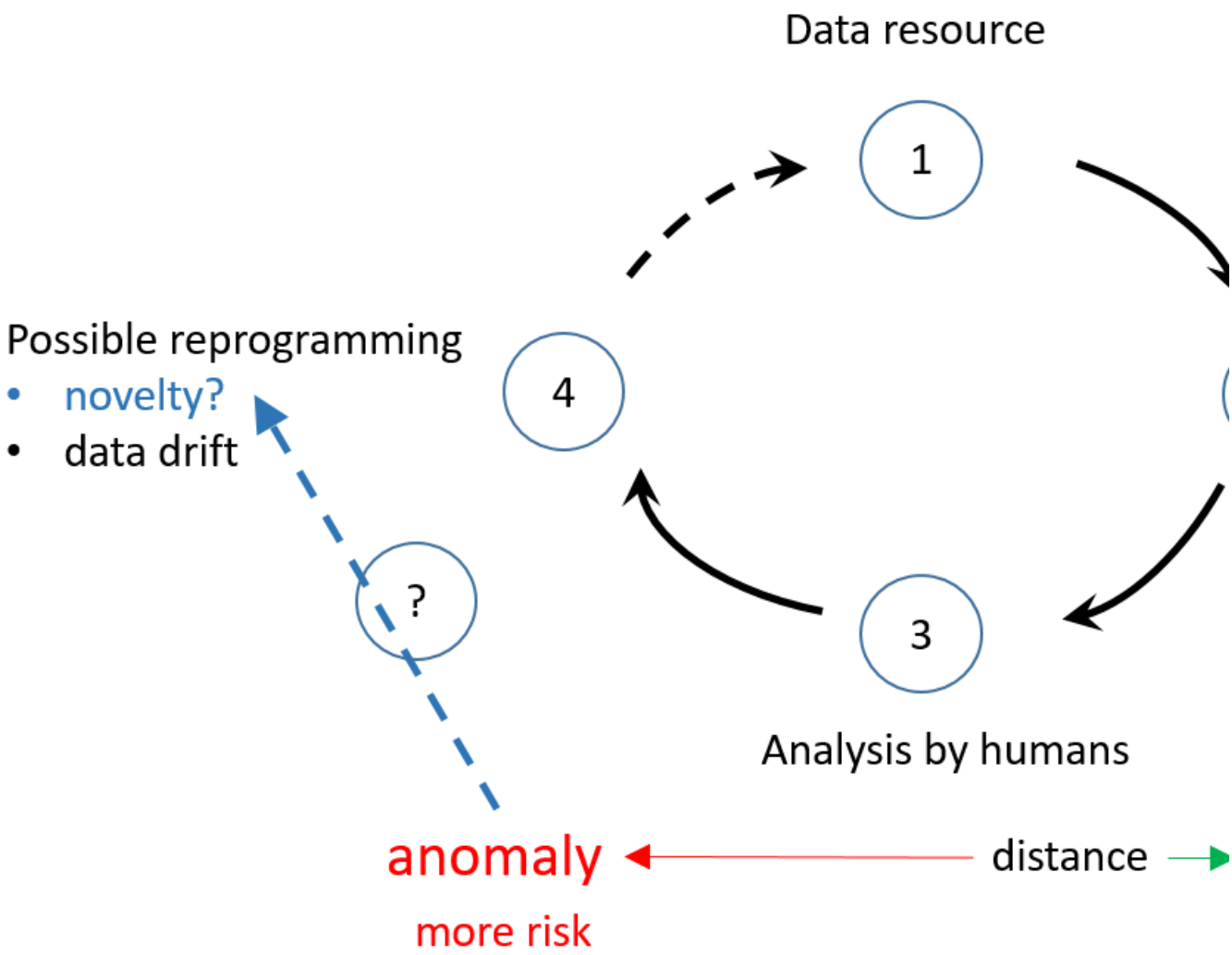
ISAA Glossary

The ISAA Glossary, EA edition includes definitions of frequent terms in AML and ISAA.

The cycle of AML analysis, ISAA, and computer/human division of labor

To help you visualize your work ISAA, the diagram below shows a cycle of the tasks in AML analysis with ISAA and the relation between work by computers and work by you and your team members.

Many terms in this section are formally defined in the ISAA Glossary, EA edition. In addition, we use the definitions of roles from the ISAA conceptual overview.



Phase	Description	Who does task?
1. Data resource	All work in AML analysis begins with a source of data.	The tasks associated with the data resource fall under Phase 2.
2. Computation	Programming the system to create mechanisms for human analysis: • Extract, Transform, Load (ETL) is to normalize data, ingest it from the resource, and recurringly reingest it as more data becomes available. • THOught Processes (THOPs) can compute custom calculations you design for your needs. NOTE on EA deployment: THOPs per se THOPs were not explicitly exposed to the EA customer. • Views of the data • Various reports A common task in Phase 2 is to programmatically classify the data into measures of risk: • Similarity is the close matching of transactions to the established patterns of customer behavior and therefore a measure of lesser risk. • Distance of the classification is based on computed probability of similarity vs. anomaly. The less the distance, the greater the similarity. The greater the distance, the greater the possible anomaly. • Anomaly is an unusual variance from the established patterns of customer behavior.	Data analyst/data scientist/programmer

3. Analysis by humans	Phase 3 is based on the mechanisms created in Phase 2. Your "Know Your Customer" work via the results of Phase 2 is one of the primary focuses of human analysis. Some questions you can ask: • How do the results from Phase 2 compare with other information sources you have? • Are the results reliably verifiable? For example, are they the same over time? Or do they vary considerably? • Is the flagged anomaly truly an anomaly? Or is it a novelty (see Phase 4)? • What are other questions pertinent to your data and goals?	Data explorer, AML investigator
4. Possible reprogramming	A return to Phase 2, reprogramming is often necessary, especially with data that grows or changes over time: • Novelty is an exception to the established pattern of customer behavior but not necesarily an anomaly. The system might need reprogramming to account for the novelty so it is no longer a novelty. • Data drift is an inevitable phenomenon: data changes over time. For example, new information about customers might be added to your data resource. The system might might need reprogramming to account for the data drift and take advantage of it for future analysis.	You and the data analysts/programmers

Accessing the ISAA AML Advisor

You access the ISAA system with your web browser. Consult your administrators for the exact URL of the AML Advisor. It is like this:

```
https://some_host_name:port
```

For example::

```
https://amladvisor.somecompany.com:8080
```

Logging in with LDAP

Your account has been set up by the ISAA administrators to allow access to the ISAA system. Your account credentials (user name and password) come from your company's LDAP system.

If you are unsure, contact your administrators for the user name and password you should use.

Working with views

A view is a pre-defined query that has been stored for your use.

To see views that have already been prepared:

- Click **Views** in the menu bar.

Results:

- Views that have been previously prepared are displayed for you to select.
- Details about the view are displayed, including the type of view, the owner's name, the status of the view, and action.

Type Column	Status Column	Action Column ...
• An *anomaly view* shows data that does not match certain identified patterns that have been defined in advance. Anomalies are sometimes called "outliers": they are variations from expected behavior or patterns in the data. • A *customer risk view* ranks customers according to the Saffron risk score and presents its results in various presentations.	• Red if the running of the view is still in progress • Green if the processing is complete and the view is ready to run.	Note on EA deployment: T he only action in the EA deployment was **Open**.

Running a view

To run a view:

1. In the list of views, click the view you want to run.
2. For that view, click the ... **Action** column, and select **Open**.

Result: The results of running the view are shown under the headings **List**, **Map**, and **Related**.

Viewing the list

Click the **List** tab for a presntation of the specific data fields that the view retrieved.

Viewing the map

Click the **Map** tab to see a diagram of the geographic locations (*geocodes*) constructed according to the similarity, anomaly, or novelty of the data points.

Viewing “Related”

Click the **Related** tab to see another presentation that includes the Saffron risk score with the data points of the map and other possibly related correlations, such as industry. These additional factors are listed on the left.

Creating an anomaly view

You must have sufficient privilege to create views. If you do not have this privilege, contact your administrator.

NOTE on EA deployment: Roles and privileges were not part of the Early Adopter deployment.

To create an anomaly view:

1. **Preparation**: Before you begin creating a view, prepare a CSV file containing IDs that represent the customers whose anomalies you want to see. Put all the comma-separated IDs on a single row.
2. From the main screen, click **Views**.
3. In the upper right, click **Create View**.
4. From the pulldown, select **Create Anomaly View**.
5. Enter a descriptive name for the view.
6. For **Date Range**, use the calendar pickers to select the starting and end dates of the data to analyze.
7. Under **Upload**, click **Choose File**, navigate your computer, and select the file of IDs you have prepared.
8. For **Dimensions**, click the pulldown menu and select any optional dimensions to constrain the analysis.
9. For **Signatures**, click the pulldown menu and select any optional signatures you want.
10. In the **Business justification/rationale** field, enter a description of the reason for this view or any other justification.
11. Click **Save and Run**.

Result:

- The view is saved to the list of views. To run it again, see Running a view.
- The view’s results are displayed. See Viewing the list, Viewing the map, or Viewing "Related".

Creating a customer risk view

The customer risk view is a predictive model based on identified characteristics of a customer's past behavior.

You must have sufficient privilege to create views. If you do not have this privilege, contact your administrator.

NOTE on EA deployment: Roles and privileges were not part of the Early Adopter deployment.

To create a customer risk view:

1. **Preparation**: Before you begin, prepare a CSV file containing IDs that represent the customers whose risk you want to see. Put all the comma-separated IDs on a single row.
2. From the main screen, click **Views**.
3. In the upper right, click **Create View**.
4. From the pulldown, select **Create Customer Risk View**.
5. Enter a descriptive name for the view.
6. Click **Upload** and navigate your computer, and select the file of IDs you have prepared.
7. In the **Context** field, enter the context you want.
8. Click the **Priority** box to enter the relative priority of this context in relation to all other contexts you plan on entering.
9. To add more contexts, click **+ Add Signature**.
10. Repeat the previous steps to add as many contexts as you want.

11. In the **Business justification/rationale** field, enter a description of the reason for this view or any other justification.
12. Click **Save and Run**.

Result:

- The view is saved to the list of views. To run it again, see Running a view.
- The view's results are displayed. See Viewing the list, Viewing the map, or Viewing "Related".

Working with Live Search

Instead of prepared views, with Live Search you can query the data yourself.

There are several ways to use Live Search:

- The displayed columns, fields, and categories
- The Live Search text box

Running a Live Search query with columns, fields, and categories

To run a Live Search query via presented fields in columns:

1. In the menu bar, click **Live Search**.
2. Select from the columns, fields, and categories (under the eye icon).

Result:

- As you select from columns, fields, and categories, the view of the data is "constrained" (that is, filtered) by the criteria you select.
- Live Search displays the data that matches your selections.
- You can copy and save this query from the text box for use in the future to avoid having to select again from the columns, fields, and categories. See Writing your query in Live Search query box.

Writing your query in Live Search query box

If you know the exact syntax of the query you want to make, including field names and values, you can enter a copied query directly in the Live Search text box. For example, you can paste a query directly into the text box.

ISAA Glossary, EA edition

ISAA Glossary

The glossary is oriented to *AML* and specific uses of the *ISAA*.

Note: The glossary does not include definitions of many common programming/computing terms, such as HTML, JavaScript, JSON, or R.

AML

Anti-Money-Laundering

anomaly

An unusual pattern that does not conform to expected behavior, sometimes also called an *outlier*. Examples of anomalies include:

- Any sudden and substantial increase in funds
- A substantial increase in the velocity (frequency) of transactions
- A large withdrawal
- Moving money to a bank secrecy jurisdiction.
- Smaller transactions that meet certain criteria might also be flagged as suspicious.

Compare *similarity* and *novelty*.

attribute

A value and a *category* with which the value is associated. Each category can be assigned a type as part of a space definition; the type is not stored in the *resource*. The supported types are string (default) and number. Example:

Category	Value	Attribute
ocean	atlantic	ocean.atlantic

An attribute is sometimes called an "entity".

Compare the programming construct *name/value pair*.

Bank Secrecy Act

US law for combating money laundering and terrorist financing. Codified in Title 31 USC 5311.

BSA

See *Bank Secrecy Act*.

category

A classification of a value. The left hand side of a *name/value pair*. Sometimes a category is a *hypernym*. See also *attribute*.

CDH

Cloudera open source big data software with integrated Apache Hadoop

data drift

A common phenomenon in a machine learning or other AI systems: data changes over time, requiring re-evaluation and perhaps redesign or reprogramming.

destination

StreamSets term for where data that has been transformed via *processors* is sent. The end of a *pipeline*, the sink for ouput from ETL. See also *origin*.

dimension

An ordered relationship in a data continuum, such as time or physical space. A secondary aspect that modifies or constrains another datum. Typically described with the word "by", as in "transactions **by time**" or "outgoing transfers **by location**".

distance

The result of a calculation of the *similarity* between two or more objects. Some kinds of distance are:

- inherent, such as with time or numbers
- geographical distance-based
- feature-based
- psychological

entity

Synonym for *attribute*.

ETL

"Extract, Transform, Load." A process in computing for pulling data out of source systems, changing the data, and making it available to other systems (sometimes by placing it into a data warehouse).

FQDN

Fully qualified domain name of an Internet-connected computer

geocode

Formal notation for the longitude and latitude of a location on the surface of the Earth.

Geocode information is supplied by the GeoNames postal and city downloads available under the Creative Commons Attribution 4.0 License.

hypernym

A word with a broad meaning that more specific words fall under. A superordinate. For example, "color" is a hypernym for the following:

- red
- green
- blue

ingestion

Transferring data from one system to another, usually transforming it for use in the new system. See also *ETL*.

ISAA

Intel Saffron AML Advisor

Know your customer

A key goal of AML involving analysis of patterns of customer behavior to establish common financial characteristics about that customer, such the kinds of transactions in which the customer is likely to engage. By knowing one's customers, financial institutions can often identify unusual or suspicious behavior, termed *anomalies*, which may be an indication of money laundering.

lemmatization

Part of *NLP*, a subtask for processing text with the use of a vocabulary and morphological analysis of words. See also *stemming*.

Lemmatization, like stemming, tries to group related words, but it goes farther than stemming in that it tries to resolve ambiguity by grouping words by their word sense, or meaning, not by their specific grammatical form. The same word may represent two meanings—for example, “wake” can mean “to wake up“ or a “funeral”.

name/value pair

In programming, a data structure that assigns a value to a variable. The name of the variable is similar to a classification or *category* for the value.

The left-hand-side is the name. The right-hand side is the value.

The name is often a *hypernym*, a superordinate of the value.

Arrays of name/value pairs are often combined to form a *namelist*, which is useful in *Named Entity Recognition*.

namelist

Programming construct for input or output of whole groups of variables, or input of selected items in a group of variables, usually in the form of an array. It specifies a group name to list the variables and arrays belonging to that group.

NER

Named Entity Recognition. Part of *NLP*, a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined *categories* such as the names of persons, organizations, locations, expressions of times, and so on.

NLP

Natural Language Processing. Some terms in NLP include:

- *hypernym*
- *lemmatization*
- *NER*
- *stemming*

novelty

A previously unnoticed observatiion of a pattern in the data not originally included or accounted for by *processors*. Distinct from *anomaly*. The novel pattern is typically added back to the data transform processors to account for the previously unobserved pattern and thus remove the novelty. Compare *similarity* and *anomaly*.

origin

StreamSets term for where particular input data comes from, a data source. The start of a *pipeline,* which ends in a *destination*.

outlier

Synonym for *anomaly*.

path

In machine learning, a probability path is designed for humans who require a deep understanding of advanced probability for their research or applied use in statistics, biology, operations research, mathematical finance (such as *AML*), engineering, and other disciplines.

In topology, a path is a continuous mapping, with an initial point, a final point, and the space of continuous functions between them. In a topological space *X*, a path is a continuous function *f* from the unit interval *I* = [0,1] to *X*. *f* : *I* ? *X*. The initial point of the path is *f*(0) and the terminal point is *f*(1).

In graph theory, a path in a graph is a finite or infinite sequence of edges which connect a sequence of vertices which, by most definitions, are all distinct from one another.

See also *signature*.

pipeline

StreamSets term for a communications/transformation channel for incoming data. With an *origin* and a *destination*, a pipeline includes discrete *stages* that run *processors* to perform a particular change (or "transformation") on the incoming data.

processor

A defined programatic function that transforms incoming data, included as a *stage* in a *pipeline*. From StreamSets.

regex

Regular expression, a text pattern matching mask. See https://en.wikipedia.org/wiki/Regular_expression.

resource

A collection of *attributes* and optional structural information. The *origin* of data for a *pipeline*.

Saffron risk score

A measure of the probability of risk based on the distance from the established pattern of customer behavior, based on specific attributes. See Metrics and Scores.

SAR

See *Suspicious Activity Report*.

segment

An ordered list of *attributes* or other segments. Segments are identified by a label, which is a string.

signature

A mathematical expression that quantifies a *path*, an evolving or time-ordered sequence of events, parameterized by a continuous variable.

similarity

The state of “likeness” between two or more objects expressed by a mathematical formula. The formula is a quantification of the degree of similarity, which is called *distance*. See also *anomaly* and *novelty*.

space

In a Docker multi-tenancy deployment, a *zone* containing spaces is a segregated area for protecting and isolating processes and data for specific purposes and specific groups of users.

In analogy with a physical apartment building with many tenants, a zone is a single, locked apartment. The zone/apartment is further subdivided into individual rooms, one per person (or group of users). The rooms an analogy for Docker spaces, which protect data specific to that group of users.

stage

A discrete, identified portion of a *pipeline* where *processors* transform incoming data. From StreamSets.

stemming

In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word base or root form, which is generally a written word form. Example: "send" is the stem of:

- send
- sending
- sent

See also *lemmatization*.

Suspicious Activity Report

After a suspected incident of money laundering or fraud, financial institutions must file a SAR report with the Financial Crimes Enforcement Network (FinCEN) of the US government. These reports are required by the United States Bank Secrecy Act (BSA) of 1970.

THOP

THOught Process. A JavaScript program you write for computing results from a Saffron memory store, relying on algorithms you implement to produce meaningful results. These THOPs are packaged into a library you create and load into the *TPE* deployment service for use with the *AML* Advisor.

THOught Process

See THOP.

TPE

Thought Process Engine. *ISAA*'s computing service that processes *THOP*s.

zone

In a Docker multi-tenancy deployment, a zone is a segregated area for protecting and isolating processes and data for specific purposes and specific groups of users.

In analogy with a physical apartment building with many tenants, a zone is a single, locked apartment. The zone/apartment is further subdivided into individual rooms, one per person (or group of users). The rooms an analogy for Docker *spaces,* which protect data specific to that group of users.

Revision history: *ISAA AML Advisor Guide, EA edition*

Date	Description
2018-01-30	Inspection session
2017-12 – 2018-01	Working drafts for internal reviews