RFP Response Form

RFP for Unstructured Data Analytics Solutions

PROPOSER INFORMATION

Proposer Name: Marinus Analytics

Authorized Representative Name & Title: Cara Jones, Chief Operating Officer

Address: PO BOX 23235, Pittsburgh, PA 15222-6235

Telephone: (866) 945-2803

Email: cara@marinusanalytics.com

Website: http://www.marinusanalytics.com/

Legal Status: ☐ For-Profit Corp. ☐ Nonprofit Corp. ☐Sole Proprietor ☒Partnership

Date Incorporated: May 2014

REQUIRED CONTACTS

	Name	Phone	Email
Chief Executive Officer	Emily Kennedy	(866) 945-2803	emily@marinusanalytics.com
Contract Processing Contact	Cara Jones	(866) 945-2803	cara@marinusanaltyics.com
Chief Information Officer	Cara Jones	(866) 945-2803	cara@marinusanaltyics.com
Chief Financial Officer	Cara Jones	(866) 945-2803	cara@marinusanaltyics.com
MPER Contact*	Cara Jones	(866) 945-2803	cara@marinusanaltyics.com

** MPER is DHS's provider and contract management system. Please list an administrative contract to update and manage this system for your agency.*

BOARD INFORMATION

Provide a list of your board members as an attachment or in the space below.
Emily Kennedy
Cara Jones
Artur Dubrawski
Carnegie Mellon University

Board Chairperson Name & Title: Click here to enter text.

Board Chairperson Address: Click here to enter text.

Board Chairperson Telephone: Click here to enter text.

RFP Response Form

RFP for Unstructured Data Analytics Solutions

Board Chairperson Email: Click here to enter text.

REFERENCES

Provide the name, affiliation and contact information [include email address and telephone number] for three references who are able to address relevant experience with your organization.
Please do not use employees of the Allegheny County Department of Human Services as references.
Nic McKinley, Executive Director of Deliver Fund - [redacted]
John Sydow, Detective with Sacramento Sheriff in Sacramento, California - [redacted]
Robert Morgester, Senior Assistant Attorney General with the eCrime Unit, California Attorney General's Office - [redacted]

PROPOSAL INFORMATION

Date Submitted 5/25/2018

Amount Requested: $995,249

Proposal Abstract:
Please limit your response to 750 characters
Marinus Analytics is proposing the CASET solution for actionable intelligence in the field. CASET is an augmentation of Traffic Jam, which is a software-as-a-service (SaaS) offering to provide actionable intelligence to safety and justice organizations from online public records. The strength of our experience is in providing tools based on advanced computing which are operationally trusted, reliable, and impactful. We are proposing to deliver both tracks of the RFP. Marinus Analytics is funded by the National Science Foundation as a Phase II SBIR recipient. A Phase IIB supplement will match DHS and provide up to $250,000. We would like to apply these funds to discount the cost of R&D to Allegheny County for the CASET solution.

CERTIFICATION

Please check the following before submitting your Proposal, as applicable:

☒ I have read the standard County terms and conditions for County contracts and the requirements for DHS Cyber Security, EEOC/Non-Discrimination and HIPAA.

☒ By submitting this proposal, I certify and represent to the County that all submitted materials are true and accurate, and that I have not offered, conferred or agreed to confer any pecuniary benefit or other

thing of value for the receipt of special treatment, advantaged information, recipient's decision, opinion, recommendation, vote or any other exercise of discretion concerning this RFP.

ATTACHMENTS

Please submit the following attachments with your Response Form. Do not provide any other attachments. Forms can be found at http://www.alleghenycounty.us/dhs/solicitations.

- MWDBE documents
- Allegheny County Vendor Creation Form
- 3 years of audited financial reports
- W-9
- Budget and budget narrative, as necessary

REQUIREMENTS

Please respond to the following. Submit only one Response Form, even when proposing Solutions for both tracks. Proposers should leave the section blank that they are not proposing and complete only the sections for the tracks they are proposing. Each track will be scored separately. The maximum score a Proposal can receive for one track is 75 points. You may provide screen shots and visuals, but please insert them in the Response Form and stay within page limits. Do not provide additional attachments that are not listed above.

RFP Response Form

RFP for Unstructured Data Analytics Solutions

1. Mining Information (75 points). If you are proposing Mining Information, fill out the questions below. If you are not proposing Mining Information, leave this section blank and move to section two for Developing Tools and Visualizations. Your response to this section should not exceed 25 pages.

1. Describe your organization's Solution to regularly and efficiently mine and organize unstructured data into meaningful information.

Overview

Marinus Analytics is proposing a comprehensive and fully automated custom data mining solution which will feed and seamlessly integrate with our visualization tools outlined below. We propose a custom solution because in our almost 30 years of combined expertise in developing natural language processing (NLP), artificial intelligence (AI) and statistical tools it has been our experience that each situation and customer is unique and that commercial, off the shelf (COTS) solutions can be fraught with over generalizations and false positives. The use of COTS packages can also stifle innovation, as they often require adherence to a limited number of methods. In short, we view the development of data mining applications as an iterative process that considers the unique circumstances of the area of application as well as the underlying data, and not as a black box, one size fits all solution.

Our solution, dubbed "CASET", will consist of multiple components working together to provide actionable information to DHS personnel and providers in as close to real time as possible. It will merge multiple sources of data, both structured and unstructured, into a centralized data warehouse that will form the basis of our data mining solution. We will then deploy a variety of methodologies to extract structured meta data that to feed into our analytical algorithms and, ultimately into the front-end visualization application.

A major aspect of our work will involve conducting research into which methods work best for DHS data. Not all methods are suitable for all data. The choice of method also depends on what we are trying to accomplish. In the end, the choice of methods will depend on not only data exploration and experimentation, but also on specific details established during the requirements gathering process.

The remainder of this section discuss how we plan on collecting raw data, what the ultimate goals of the data mining system are, the types of methods we intend to use, and some examples of how data will be extracted.

1. Raw Data Collection/Data Warehouse

As outlined in the request for proposals, Allegheny County DHS has a variety of data sources from several different legacy applications. Some of these contain structured data, others do not. In order to be able to effectively and efficiently mine this data, we will first need to design and implement a data warehouse solution that we will deploy in the Amazon cloud. The data warehouse will also store any meta data extracted from unstructured data via data mining, as well as aggregate qualitative and quantitative data that can be used for statistical modeling.

Finally, we plan on adding a geodatabase to the data warehouse, so that we can capture location specific information that can be mapped using the visualization tools, as well as deploy spatio-temporal analytical techniques that can track issues over space and time. A further advantage of including the spatial coordinates of locations such as a client's address, the location of cases or that of a provider, etc. is that we can overlay these with census tracts or blocks, as well as neighborhoods. Base maps and layers for a geodatabase are readily available, either from vendors or the county's planning department. A geodatabase will also allow us to combine DHS data with other data sources, such as census, economic and housing data, which will provide a rich source of data for detailed, location specific analyses. Moreover, this data will not only feed our solution, but could also be used for academic and in-house studies at DHS.

While we are generally database agnostic, given that DHS already uses Oracle, we would suggest going that route. However, if the cost of Oracle is an issue, there are also open source database options that we have deployed in the past which we could use. We are, of course, also open to any other database solutions that DHS might have in mind.

In order to populate the database warehouse in as close to real time as possible, we will build database connectors that access DHS data, transform them and insert them into tables in the warehouse. There are two ways in which this can be accomplished. The first is to pull data from the cloud by directly accessing DHS databases and files, or by pushing data to the cloud from DHS systems. While we would prefer the first method, the second method or some combination would also work. The database connectors would access DHS systems periodically, i.e., on an hourly basis to synchronize data between the various systems. If desired, this can also be accomplished in real time. Documents not contained in databases that are in pdf or other formats will be converted to text using OCR technology.

Depending on the data and the ability to extract and infer entities such as people, places and organizations from case notes and structured data, we anticipate that a graph database would add much value to the solution. For example, it would allow for the linking of events to people and places, as well as allow for grouping of family units and their links to other entities, such as key events or topics, diagnoses and sentiments. Storing data within a graph database opens the door for many methodologies that can be applied to this data, such as social network analytics and identifying networks of families that share similar characteristics. As noted in the RFP, it is difficult to identify named individuals and their roles within a case. However, techniques do exist that can, for instance, disambiguate names and roles based on other clues found within a text. At a minimum it will be worthwhile to explore the feasibility of entity extraction, because if feasible to a reasonable degree of accuracy, it would add much value to the solution.

Using text indexing and/or search engines such as Elastic Search data contained within the data warehouse will be fully searchable almost instantaneously. Compared to other systems we have developed, some of which have contained hundreds of millions or even over a billion records, the anticipated volume of data generated by DHS systems – currently at 6 million historical records - is likely to be very low, so performance will not be an issue whatsoever.

Integrating the structured data currently relied on by DHS with previously unavailable meta data extracted by our data mining solution will not only allow for broader and more detailed insights into the totality of data that the various DHS legacy systems house, but also provide the ability to create analytical reports that were previously impossible on an automated basis. Examples of these will be outlined in more detail below.

2. Data Mining Goals

Allegheny County DHS seeks to extract meaningful information from a combination of both structured and unstructured data, including information about clients and their circumstances, as well as information about overall sentiment of a case. This section addresses how we will meet that requirement, as well as other types of information we would like to mine.

Information about Clients and their Circumstances

This type of data speaks mainly to the categories of events, needs and issues that a client is dealing with. We plan on using a combination of techniques, ranging from regular expressions to machine learning and natural language processing to identify and extract those categories. We anticipate that each category will have major and minor topics associated with it. Some non-exhaustive examples are:

- Major: Substance Abuse
- Minor:
    - Opioids
    - Alcohol
    - Crystal Meth
    - Cannabis
    - Cocaine
    - General, Non Drug Specific
- Major: Natural Disaster
- Minor:
    - Flood
    - Fire
    - Storm Damage
    - Earthquake
    - Prolonged Power Outage
- Major: Children
- Minor:
    - Learning Disability
    - Truancy
    - Poor Educational Performance
    - Physical Handicap

Which major and minor categories to include will be determined by the discovery and requirements gathering process.

Information about Overall Sentiment of a Case.

We will use sentiment analysis to determine how a customer or case worker feels about a case, based on the contents of a report. However, whenever possible, we want to go beyond analyzing just the overall sentiment of a case to also analyzing sentiments about individual categories of events and issues. For example, overall, a caseworker might be happy with a family's progress, but there may still be concerns about certain aspects, such as drug use. Simply taking the overall sentiment at face value would not necessarily capture this fact. We will therefore also examine sentiments at the sentence level. Consider the following example of a sentence contained within a report that might be positive overall:

"The mother states that while things have been improving, she is still worried about her son's continued drug use."

Conducting a sentiment analysis on individual sentences in which categories are extracted will allow for not just the identification of a particular category, but also whether the category is viewed in a positive or negative light within a report. A positive example might be:

"The mother states that she is glad that her son's drug use appears to have ceased."

Because each of those sentiments will have dates and times associated with them, we will be able to track both overall sentiments as well as sentiments about individual categories over time, thus providing more granularity and improved reporting capabilities. We will also be able to identify which aspects of a case are improving, and which are not.

The capabilities of our data mining solution will exceed the functionality suggested in the request for proposals. Specifically, CASET will combine structured and unstructured data to determine the who, what, where, when, why, how and outcome of each case. Each of these categories will refer to specific entities and attributes as follows:

Who?

The who refers to the various individuals and organizations involved with a case. Some of these entities are likely to be contained in structured data, such as assigned case workers and organizations, while others, such as family members and relatives, will have to be extracted via entity extraction from unstructured text. As stated above, there are methods that can potentially identify individuals and disambiguate them from other individuals with the same name.

What?

The what means the types of events, issues and needs associated with a case. This is essentially the same as information about clients and their circumstances.

Where?

The where refers to more than just the street address of a client. It could mean the location of an interview, such as at an office, the location of a provider or location information gleaned from unstructured data via entity extraction. For example, the sentence "*Client stated that he was staying at the homeless camp on Fraser Street"* implies a place on Fraser Street, which can in most circumstances be extracted. As mentioned earlier, having information about the where of a case opens many possibilities with respect to geographic analysis, spatio-temporal modeling, as well as the ability to display the geographic extent of a case and the results of any analysis using a Geographic Information System (GIS).

When?

Dates and times are likely to be contained in both structured and unstructured data and can be readily extracted from both and connected to any other elements that are mined from the data.

Why?

The why of a case is more of an inference rather than an entity or an attribute. Why's can include things such as the type of document, entities, needs, issues, dates and events associated with initial assessments. In other words, why was a caseworker referred to a family in the first place?

How?

The how of a case refers to the treatment(s) applied to the case. Which entities and programs were involved? What type of treatments, such as rehab, counselling, police intervention, etc. were applied? From a data mining perspective, treatments are concepts in the same sense as other entities.

Outcome

Being able to measure the outcome of a case is important in that doing so allows for statistical analysis of cases with respect to the result. Because outcome is very much a qualitative, rather than a quantitative aspect, care will need to be taken exactly how we will quantify outcomes. Some outcomes will be easy to quantify, others not. There are several candidate techniques from social sciences that we could deploy, depending on what is desired.

In summary, classifying features extracted from the data into these categories will allow for the creation of quick overviews of a case, the identification of similar cases, as well as comparisons between cases.

Beyond data mining, it is our goal to provide DHS personnel as well as any potential internal or external researchers with quantitative data that will allow for statistical analyses ranging from the simple to the complex. We plan to include several standard statistical analyses that will be generated automatically, such as using time series analysis and outlier detection to generate early warning alerts for both individual cases as well as emerging trends among all cases, producing periodic maps that highlight

where DHS is most active or where problem areas exist, as well as any other analyses that are identified during the requirements gathering process.

3. Methodology

This section discusses some of the different types of methods we will be using for the solution. Note that while this is primarily an engineering project, and not a research project, there will likely be at least two rounds of data exploration, research and development to identify those methods that work well for DHS data, as well as training those methods and evaluating them. There will likely also be several rounds of recalibration of data mining modules based on feedback from users. We will be using a mix of methods, ranging from the simple, such as regular expressions, to intermediate methods, such as rule-based heuristics to the more complex, such as support vector machines for sentiment analysis.

Preparatory Work

Before we can commence training models and extract meta data, we will need to prepare not just the unstructured text, but also create normalized dictionaries that expand abbreviations and identify which concepts and categories (major and minor) we want to be able to extract.

For machine learning, it will also be necessary to hand annotate unstructured data. There are two basic types of machine learning: supervised and unsupervised. Supervised learning tends to be more accurate but requires training and test data that must be annotated.

Regular Expressions

We will regular expression to extract meta data such as dates, times, and, if applicable, telephone numbers and email addresses. We have several regular expression extractors that we use in our Traffic Jam application that have proven to work very well over the years.

Machine Learning Based Classifiers

The easiest and most common method used to identify documents that talk about specific concepts and categories (also known as features in machine learning parlance) is to use keyword searches. However, keyword searches have several drawbacks that could produce false positives as well as false negatives. For example, simply searching for the term “drugs” could potentially return cases that mention medications and would miss specific mentions of drugs such as cocaine and crystal meth. Another issue is that there many different spelling variations of some keywords. Finally, it would require a user to know all possible permutations of terms to successfully find all documents related to a major or minor category. While our solution will certainly include the ability to perform ad hoc keyword based searches, we will deploy a combination of machine learning and heuristics to extract features from unstructured data.

Figure 1 below is a general depiction of how we will apply machine learning. First, we will create a test data set with hand annotated labels, or categories, that we want to extract. Next, for each document

the unstructured text is preprocessed and transformed into feature vectors. Preprocessing includes expanding abbreviations to their full representation and breaking down a document into individual sentences. It also includes lemmatization, which transforms adjectives, nouns, adverbs and verbs, including misspelled ones, into a single common representation. Examples include:

- Happier, happiest, hapiest -> happy
- Children, kids, kid, child -> child
- Writes, writing, wirting, wrote, written -> write

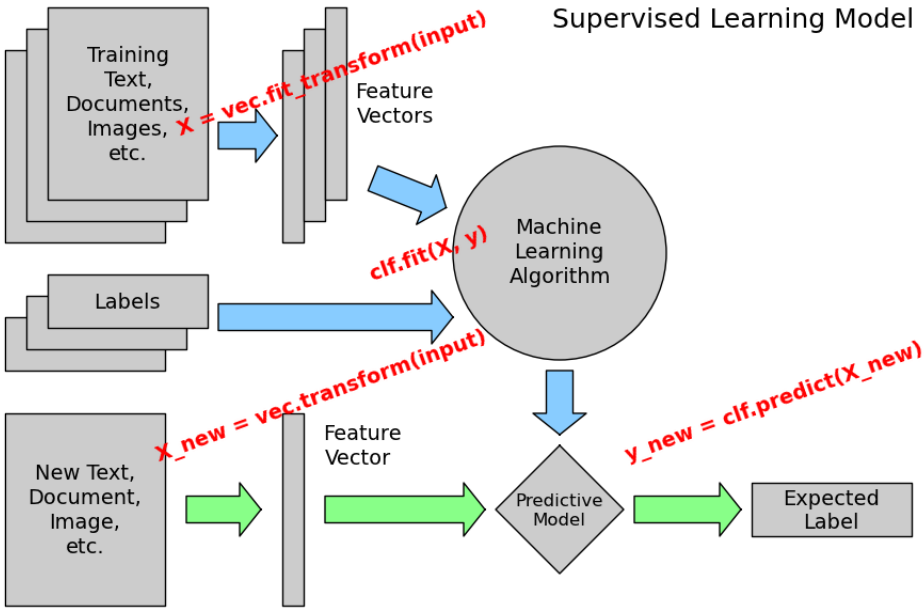


Figure 1: Generalized Supervised Learning Model (Source: http://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/general_concepts.html)

Next, the model is trained using a machine learning algorithm (candidates will include Naïve Bayes classifiers and Support Vector Machines (SVM)) and tested for accuracy against an annotated test data set not used for training that was also preprocessed and featurized. If the model performs well on the test data set and produces a similar degree of accuracy as on the training data set, a model is said to generalize well and can be used in production.

Once a model has been trained, unstructured data is processed by applying the same preprocessing and featurization techniques as the training data. Labels are extracted by applying the trained model to each document.

Sentiment Analysis

Sentiment analysis works similarly to the supervised learning model above, except that the output will be a number between -1 and +1 on a sliding scale. It will likely not be necessary to hand annotate data because there are already many pre-trained models on large corpora of text that we could leverage, although it may be necessary to slightly modify them to accommodate the unique language found in DHS documents.

Figure 2 is an example of how documents could be visualized on the front end based on the sentiment analysis:

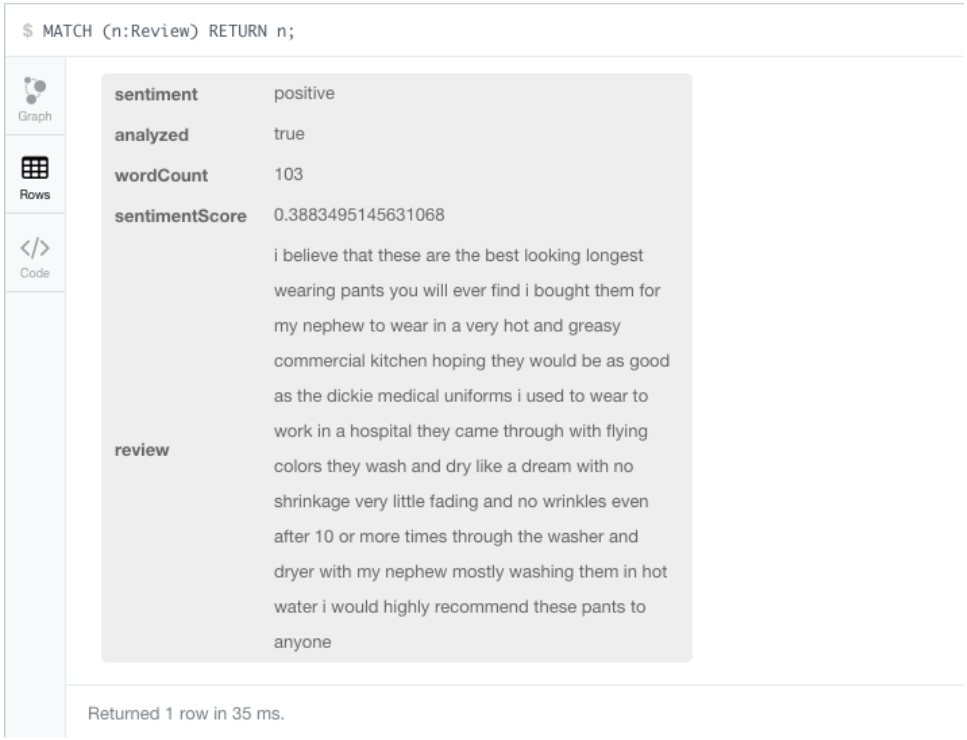


Figure 2: Sentiment Analysis Example (Source : http://kvangundy.com/wp/sentiment-analysis-amazon-reviews-using-neo4j/)

Rule Based Heuristics

Sometimes classification results can be improved by using rule based heuristic algorithms, either on a standalone basis or by combining them with machine learning models. Examples of rule-based heuristics for text classification include decision trees and context sensitive learning methods.

4. Examples of Extracted Data

This section illustrates some examples of how data will be extracted using some of the examples of unstructured text given in the request for proposals. Consider the following original report:

RFP Response Form

RFP for Unstructured Data Analytics Solutions

Original Text

8-8-10--2:44 PM--Message received from Officer Darrel Jones of Pittsburgh P.D. He reported that Leslie Smith was picked up as a RAW. Teen is reportedly drinking and smoking weed. M and MGM are interested in having teen placed in Shuman Center.

later--TC to M, Stephanie Shuster --#412-555-1212. She was at MGMs house. Leslie Smith has been staying with MGM for two months. MGM can no longer handle child. She left on 8-6-10. They found her today in Moon Township with a 19 year old guy. Something has to be done. IW advised M that information would be forwarded to the CW.

later--4:18 PM--Return call from M. Leslie Smith took off again. IW advised M to again reported her daughter missing to the police.

A first step would be to expand all abbreviations as follows:

Expanded Text

8-8-10--2:44 PM--Message received from Officer Darrel Jones of Pittsburgh Police Department. He reported that Leslie Smith was picked up as a runaway. Teen is reportedly drinking and smoking weed. Mother and maternal grandmother are interested in having teen placed in Shuman Center.

later—Telephone call to mother, Stephanie Smith #412-555-1212. She was at maternal grandmother's house. Leslie Smith has been staying with maternal grandmother for two months. Maternal grandmother can no longer handle child. She left on 8-6-10. They found her today in Moon Township with a 19 year old guy. Something has to be done. IW advised mother that information would be forwarded to the case worker.

later--4:18 PM--Return call from mother. Leslie Smith took off again. IW advised mother to again report her daughter missing to the police.

Next, the proposed solution would extract all entities from the text:

Extract Entities

8-8-10--2:44 PM--Message received from Officer Darrel Jones of Pittsburgh Police Department. He reported that Leslie Smith was picked up as a runaway. Teen is reportedly drinking and smoking weed. Mother and maternal grandmother are interested in having teen placed in Shuman Center.

later—Telephone call to mother, Stephanie Smith #412-555-1212. She was at maternal grandmother's house. Leslie Smith has been staying with maternal grandmother for two months. Maternal grandmother can no longer handle child. She left on 8-6-10. They found her today in Moon Township with a 19 year old guy. Something has to be done. IW advised mother that information would be forwarded to the case worker.

later--4:18 PM--Return call from mother. Leslie Smith took off again. IW advised mother to again report her daughter missing to the police.

In this example there are six types of entities extracted: dates (yellow, When?), times (green, When?), people (blue, Who?), organizations (red, Who?), phone numbers (teal, Who?) and places (grey, Where?). Note that for people, we will need to disambiguate, as there are many mothers mentioned in other reports. In this case we will be able to tie this unique mother by the case number or other global identifiers that DHS might have in its data.

The following example illustrates the types of categories, such as events, issues and needs (major and minor) that will be extracted. Note that this is hypothetical and meant as an illustration only. The exact categories to be extracted will be determined during the requirements gathering process.

Met with the family to do the assessment and have mother sign paperwork. The apartment was dirty, smelly and there were flies. There was clutter all over the floor and kitchen. Mother ran out of food for herself and the children. She claimed that she tried contacting the food banks but hasn't heard anything back from them. She seems stressed and overwhelmed. She has not been taking care of herself or her health. She has had incidences where she has passed out and had to be taken to the hospital.

In order of appearance within the text, the following categories were extracted. Sentiments ranging from -1 (negative) to +1 (positive) are also indicated:

- "mother sign paperwork"
    - Major: Task – What?
    - Minor: Paperwork
    - Sentiment: 0 (neutral)
- "dirty, smelly and there were flies"
    - Major: Domestic Conditions – What?
    - Minor: Dirty, Smelly, Flies
    - Sentiment: -0.9 (very negative)
- "clutter all over"
    - Major: Domestic Conditions – What?
    - Minor: Clutter
    - Sentiment: -0.5 (largely negative)
- "Mother ran out of food" – What?
    - Major: Needs
    - Minor: Food
    - Sentiment: -0.6 (largely negative)
- "Contacting the food banks" – What?
    - Major: Treatment
    - Minor: Provide Food
    - Sentiment: -0.4 (negative)

- "Stressed and overwhelmed" – What?
    - Major: Health
    - Minor: Stress
    - Sentiment -0.8 (largely negative)
- "not been taking care of herself or her health"
    - Major: Health – What?
    - Minor: Neglect
    - Sentiment -0.9 (very negative)
- "taken to the hospital" – What??
    - Major: Health
    - Minor: Hospitalization
    - Sentiment: 0 (neutral)

As outlined above, all data extracted from unstructured text will be merged with structured data and will feed the data visualization tool.

2. Describe how your Solution addresses DHS's goals and is scalable to allow for expansion to address future needs.

DHS Goals

We believe that our proposed solution – CASET - will meet and exceed DHS's goals. It will provide functionality that goes beyond that outlined in the request for proposal. Specifically, we will not only mine data from unstructured text, but also provide statistical analyses, mapping capabilities and an early warning system.

Scalability

From a database and data access perspective, each component within CASET is highly modularized and will pass information between all sources and sinks in an open format, such as JSON, which allows for any module to be plugged in, as need be. Additionally, every data storage component will be stored in either clear text or encrypted with a universal key; in other words, no information will be stored in binary format. These two pieces will allow any additional component developed by Marinus Analytics or by a 3rd-party to have easy access to all the results from the APIs. If another application requires data to be processed, it can request the information from the APIs, convert the data into a new format, and send it back to the API to be stored accordingly.

Alternatively, since the data will be stored in industry-standard data sources such as Oracle, SQL Server, PostgreSQL or Neo4J, 3rd-party applications with appropriate security credentials will be able to access the data directly. This will allow them to bypass APIs that transform data in particular ways and also allow them to extend the databases directly to store their new information in a centralized source.

These methods have been proven to work within Marinus Analytics web-based offerings and APIs. Our custom API can return over 100,000 API requests that query a dataset in the millions of records. Our customers have never reported down time or performance issues. The website uses the same data

storage backend the API uses with no impact on each other's performance, which speaks to the ability of our systems to expand in a very scalable way.

For a scalability point-of-comparison, Traffic Jam maintains a growing bank of over 210 million public unstructured records. The volume of data listed in Appendix A of the RFP will not be of scalability concern. The search functionality within the core of Traffic Jam allows for on-the-fly pattern recognition across substantially larger data volumes where respective queries return in a few seconds with no impairment to the user experience. The current user base of Traffic Jam involves hundreds of active users who may concurrently use the system at any moment.

From a data mining perspective, CASET will easily be able to scale up to processing hundreds of thousands of records a day, as we do today with Traffic Jam. Our extractors each run in their own separate process, with minimal lag between the time data is ingested, processed, and available to the front end.

Future Needs

To-date, Marinus Analytics has leveraged cloud architecture to optimize cost and scalability. We are experienced engineers in understanding the full stack of components along with the emerging eco-system of enhancers of the stack which can be procured and incorporated to benefit a given tool or solution.

Should the need arise in the future to add additional data mining functionality, we will be able to incorporate the additional functionality. In our Traffic Jam application, we are adding new data mining functionality on an almost monthly basis.

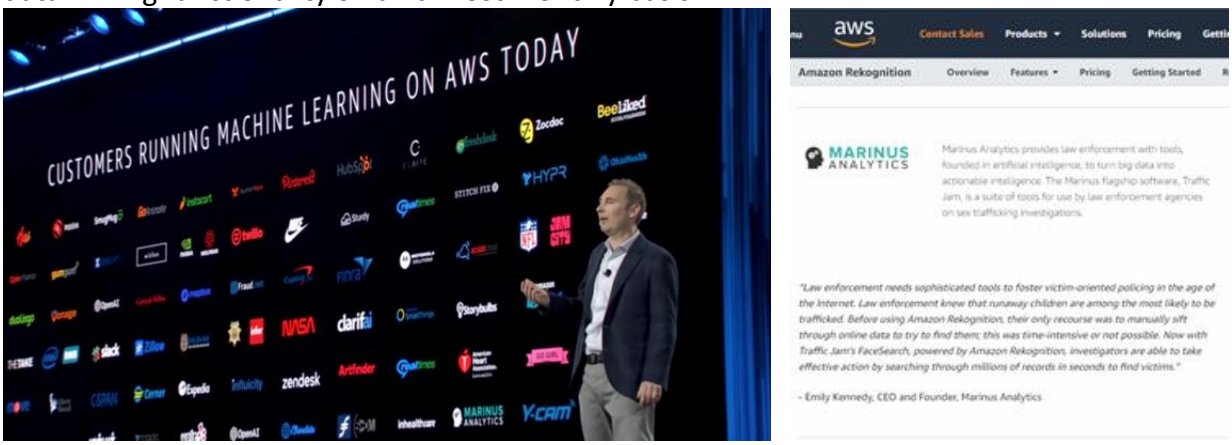


Figure 3 AWS CEO, Andy Jassy, specifically mentioned Marinus Analytics in his keynote speech at re:Invent conference, November 2017 in Las Vegas, NV and Marinus Analytics is listed on AWS website

As mentioned in the answer to the previous question, CASET architecture lends itself to new modules which may provide radical innovations to the traditional case management information system behavior. These are modules which can support speech-to-text transcription, audible play-back of notes, and of-course the latest artificial intelligence products to further the data mining engine.

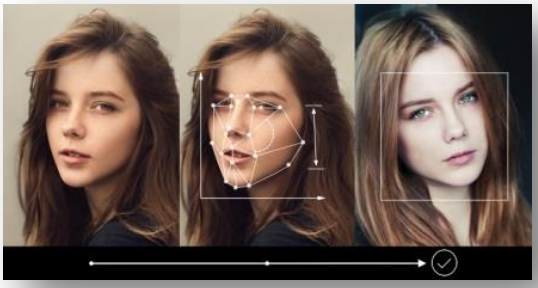
Figure 3 illustrates our relationship with AWS. We strive to maintain the latest knowledge of commercially available cloud services to support delivery of AI-driven capabilities to the public sector. In the world of AI, there is a phenomenon of "two-year-leaps." Maintaining a close pulse of the innovations by leaders like AWS along with our close ties to the academic research communities, allows Marinus Analytics to best prototype and commercialize cutting-edge tools for organizations like Allegheny County DHS. As discussed in the answer to question #4, our team in comprised of a faculty member from Carnegie Mellon University who advises our highly-skilled scientists and engineers on academic breakthroughs which may be translated into practice. These combined aspects of our team will allow for innovative solutions to best serve the future needs of your organization.

3. Describe your organization's experience developing, implementing and evaluating Solutions to regularly and efficiently mine and organize unstructured data.

Marinus Analytics has over 30 years of combined experience in designing, developing and deploying natural language and artificial intelligence-based applications in a variety of fields, including broadcast news, inmate phone calls, law enforcement and intelligence. Practically every project we have worked on, whether at Marinus Analytics or previous employers, has involved creating solutions that are cutting edge, innovative and have transformed the way in which users do their works.

Marinus Analytics is the exclusive provider of Traffic Jam which uses the latest advancements in AI to turn big data online into actionable intelligence, for the rescue of exploited human trafficking victims and protection of at-risk populations. Traffic Jam, on which the core of CASET is based, is an implementation of previous academic research at Carnegie Mellon University that underwent extensive rebuilding and scaling to be productized and available for customers. The majority of this work, performed by Marinus Analytics engineers, took place in 2015 and 2016. As the data archive grew to tens of millions of records, the earlier version of the technology degraded in availability, experiencing uptime at 50% and would often fail during key business hours. This research software was re-written and ported into more stable environments and storage mediums with more reliable web interfaces to increase uptime to 99.9%, and the archive has now exceeded hundreds of millions of records.

Traffic Jam is available on any desktop, handheld, and mobile device. The technology includes a powerful elastic search engine, machine learning algorithms, image feature extraction, image similarity detection, and predictive analytics to make this large amount of data practically useful to public safety and prosecutorial organizations.



Traffic Jam FaceSearch, deployed since June 2017

In June 2017, Marinus Analytics released FaceSearch with Traffic Jam to help detectives improve their ability to find specific missing persons,

and apprehend their exploiters more effectively.[1] The feature has been heavily leveraged by our law enforcement user community and Traffic Jam is used daily by analysts at the National Center for Missing & Exploited Children (NCMEC), an organization which operates a national hotline and CyberTipline for reporting suspicion of crimes including child sex trafficking. It is very important to quickly identify and rescue juveniles who are being trafficked because exploited individuals and groups frequently move across cities and states.

An extensive network of public sector agencies uses Traffic Jam on a daily basis. Other agencies include Department of Homeland Security (DHS) Field Offices, DHS Human Smuggling and Trafficking Center, the Department of State Diplomatic Security Service, U.S. Attorney's offices, and nearly 200 other agencies across the United States, Canada, and recently countries in the United Kingdom. Traffic Jam has proven instrumental, even for agencies with no in-house analyst. Indeed, many of our agencies have given us the same feedback: "The amount of time saved using Traffic Jam replaces the work of a full-time analyst."

Our ability to translate research into practice has been recognized through our participation in the IBM Watson AI for Good XPRIZE Competition. Marinus is one out of 59 teams from around the globe to be accepted into Round 2 of the XPRIZE Competition. The XPRIZE team explains, "The four-year prize competition aims to accelerate adoption of Artificial Intelligence (AI) technologies and spark creative, innovative, and audacious demonstrations of the technology that are truly scalable and solve societal grand challenges."

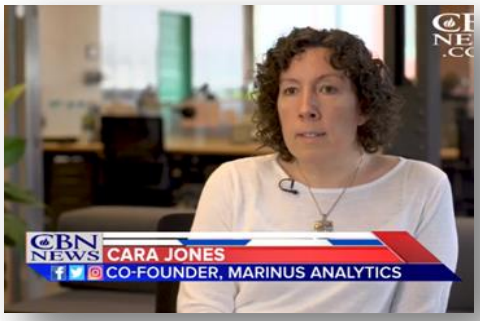
4. Provide staff bios (not CVs) of the key staff who will be implementing your Solution and identify the main point of contact. Describe your management structure and how it will support the goals of your proposed Solution. If you are partnering, describe the structure of the partnership.

Marinus Analytics is a women-owned and managed, small business operating in Pittsburgh, Pennsylvania. The company spun out of the Robotics Institute at Carnegie Mellon University in 2014. The staff at Marinus Analytics has extensive experience managing teams of computer scientists, software engineers, data scientists, and researchers. By embracing today's latest advancements in cloud computing, Marinus Analytics is highly effective in tool building and providing SaaS AI-driven capabilities to the public sector.

[1] Marinus Analytics, "Pittsburgh-Based Technology Company Debuts First Facial Recognition Technology Designed To Halt Global Human Trafficking," *Marinus Analytics Press Release,* Accessible at: http://www.marinusanalytics.com/articles/2017/6/27/face-search-debut

Cara Jones, B.S. Computer Engineering, MBA

Our engineering is led by Chief Operating Officer, Cara Jones, who has over 15 years of experience with high technology implementations. Cara Jones has managed the maturation and commercialization of Marinus Analytics's Traffic Jam technology since 2013. In her career, she has worked a spectrum of projects ranging from novel autonomous material-handling robots for hospitals to complex enterprise information systems. She honed her project management experience while serving in technology consulting at Deloitte, overseeing test, planning, and execution for financial ERP implementations within the Department of Defense. In particular, Cara Jones has significant experience leading testing for complex deployments. System testing is an integrative role which requires an understanding of many aspects of a solution, including the domain-specific functionality to serve the users' needs. At the Department of Defense, she worked with the Joint Interoperability Test Command (JITC) to design the requirements traceability and test methodology for a 13-month deployment to replace the financial accounting and reporting system for an Agency with over $5B in annual transactions. This system included a dozen interfaces with other government applications and required inter-agency coordination, end-to-end testing, defect tracking, and technical resolution under pressure to meet the narrow end-of -fiscal-year cutover window. She also served as technical lead and managed an internationally staffed team on a retail eCommerce deployment. These projects faced many logistical challenges to coordinate, deploy by specific fixed-dates, and mitigate the risk of introducing major new technology to ongoing client operations. Ms. Jones delivers results whether working on project teams involving dozens of members or those involving few resources with minimal funding dollars. In her personal life, she served many years as a youth ice hockey coach and brings a spirit of team work and cohesion to her professional responsibilities. Cara Jones will serve as the main point of contact for the overall awarded contract with Allegheny County DHS and will deliver success in managing the team, engaging the client, and delivering the solution's milestones on-time.



Marinus COO Cara Jones in TV interview, October 2017

Marinus CEO Emily Kennedy at Lincoln Center for Women in the World, May 2018

Emily Kennedy, B.S. Carnegie Mellon Heinz School of Public Policy

Kennedy is CEO of Marinus Analytics, and directs deployment of advanced data mining and machine learning tools to local, state,

and federal law enforcement for use on criminal cases, with an emphasis on human trafficking investigations. She routinely works alongside, advises, and trains stakeholders—such as attorneys general, prosecutors, law enforcement agents, and non-profit victim services organizations—on micro and macro approaches to combating and measuring human trafficking in the United States and abroad. Her work has been covered at the United Nations, Fast Company, NBC News, Vice, CBS News, and Scientific American. She has successfully raised funding from the National Science Foundation, the Bank of New York Mellon, and DARPA. She was recently selected as a Mother of Invention, sponsored by Toyota, to honor Marinus's extensive work in the social impact space. She is intrigued by technology solutions to social problems and is driven to innovate new ways to work with government organizations toward data-driven solutions. She graduated from Carnegie Mellon University in 2012.

Raymond Giorgi, M.S. Computer Science

Senior Software Architect Raymond Giorgi has worked as a software engineer and project manager for 10 years. During his time in the advertising industry, he was often responsible for all aspects of the software development life cycle (SDLC) in events with minimal development time and hard deadlines.

Marinus Senior Software Architect Ray Giorgi speaking about Marinus work in the IBM Watson XPRIZE – AI for Good Competition

His portfolio of rapid creations for clients were sometimes the result of extreme scenarios, such as only one week of development before the client event or developing working architecture and test plans for events in which only one day of testing in a foreign country was available. Raymond was also previously a Managing Data Scientist, where he oversaw projects that provided actionable information from big data resources. Raymond brings these experiences to Marinus Analytics and leads the delivery of new applications and modern tools for policing in the digital era. Raymond incorporates lean and agile methods to innovate and aid in the public sector. His work includes developing game-changing technology to generate leads for proactive policing, to help law enforcement assimilate analytics into their current workflow, and to use facial recognition for finding victims and returning them to safety. Raymond's role also involves coordinating closely with law enforcement officials in the design phase of projects. Prior to joining Marinus Analytics, Raymond served a variety of consulting firms, managing and developing end-to-end solutions for clients in the public and private sectors. Raymond holds a Master's of Computer Science from the University of Pittsburgh.

Research Scientist, **Dr. Andreas "Olli" Olligschlaeger**, holds an M.Phil. and Ph.D. from Carnegie Mellon's Heinz School, an M.A. from the University of British Columbia, and a B.A. from Concordia University. He has over 30 years of experience in productizing and transitioning academic research to the private sector, specifically within law enforcement and public safety. With practical experience in law enforcement, academia, and private industry he has a unique and proven track record of introducing new technologies to law enforcement and integrating them with daily operations. For example, Dr. Olligschlaeger was instrumental in integrating the latest research in geographic information systems with crime analysis units around the country in the early 1990's; today, crime mapping is ubiquitous in law enforcement. His dissertation on using artificial neural network based space/time forecasting techniques and work as a crime and narcotics intelligence analyst within the Pittsburgh Police Department helped establish the field of predictive policing as we know it today.

In the early 2000's he was a systems scientist on the Informedia team at Carnegie Mellon's Robotics Institute and School of Computer Science. There, he worked with DARPA to develop a system that uses AI, natural language processing and speech recognition to extract meta data from broadcast news and allow for this data to be queried and visualized in multiple modalities, including GIS, link charts and timelines. In the mid and late 2000's he worked on integrating research in speech recognition and natural language processing with inmate phone systems, allowing investigators to automatically monitor every single inmate phone call made from a prison facility. The system has been deployed in numerous jails around the country and resulted in hundreds of criminal cases as well as several patents. In the early 2010's Dr. Olligschlaeger worked as a subject matter expert with the FBI and Raytheon to introduce and deploy academic research to N-DEx, which is a system that ingests massive amounts of incident and arrest data on a daily basis from police departments around the country. Specifically, he worked on entity extraction techniques, graph databases and social network analytics (including consulting on how measures of centrality can be applied to law enforcement), all of which are now fully integrated within N-DEx. Dr. Olligschlaeger has also worked on several projects with the Bureau of Alcohol, Tobacco, and Firearms' National Tracing Center, Axon's body worn video and digital evidence program, and served on the FBI's Future's Working Group, where he analyzed emerging technologies and academic research in order to assess their potential impact on policing and crime. Since 2016 Dr. Olligschlaeger has been a member of the Marinus Analytics team, where he has leveraged his expertise to design and develop software that uncovers human trafficking in vast amounts of data mined from the web.

Dr. Olligschlaeger has over 30 years of hands-on experience developing software, both within small startup settings as well as large organizations. He has worked on projects ranging from 100 thousand dollars involving one or two developers to projects costing over 100 million dollars involving teams of 75 developers and numerous project managers working in an agile environment. As such, he is comfortable working in any setting. His development skills include all modern programming languages, including java, C++ and Python, GIS, a wide variety of commercial and open source databases, including Oracle, SQL Server, MySQL and PostGresql, as well as geodatabases and numerous integrated development environments, including Netbeans, Eclipse, and Microsoft Developer Studio. Dr. Olligschlaeger works on a daily basis with code control tools such as Git and SVN, as well as bug tracking tools such as JIRA.

Dr. Olligschlaeger is a member of the International Association of Crime Analysts, the International Association of Law Enforcement Intelligence Analysts, the Society of Police Futurists International (PFI), where he is the immediate past president, the PFI/FBI Futures Working Group, and serves on the advisory board of the High Tech Crime Consortium.

Thomas Wolber is a full-stack engineer whose career started as a graphic artist. Applying design experience to day-to-day work has lead Thomas to have a strong focus on user experience, accessibility, and ADA/Section 508 compliance over his 20 year career. With experience across various modern technology stacks, Thomas is not only skilled in his own work but serves well as a liaison between team members with diverse concentrations.

Julia Deeb-Swihart

Julia Deeb-Swihart is a Computer Science PhD student at Georgia Institute of Technology. She holds a B.S in Computer Science from Georgia Institute of Technology with a focus in Artificial Intelligence and Computational Theory. Her research focuses on applications of machine learning, network science, and user centered design for social good. Her thesis work is focused on building tools that assist with law enforcement investigations into human trafficking. As part of this work, she interviews law enforcement officers to understand their needs and challenges with their jobs and utilizes these insights to design systems that meet their needs. Most law enforcement officers have little or no computer science training, but have clear needs to work with big data to be successful in their work.

“My research is focused on building computing tools that empower individuals in their jobs for social good. Assisting human services aligns with my research visions.”

Dr. Artur Dubrawski, Advisor

Artur Dubrawski, founder and advisor, is a faculty member at the CMU School of Computer Science where he directs the Auton Lab, an applied machine learning research team of 30. He provides guidance on ongoing R&D at Marinus Analytics related to machine learning. Dr. Dubrawski has been researching intelligent systems (that work, are useful, and make economic sense) and ways to effectively build and deploy them for 20+ years. He leads teams at CMU investigating new machine learning algorithms and data structures to facilitate probabilistic modeling, predictive analysis, interactive exploration, and understanding of data.

5. Provide a detailed budget that clearly supports your Solution and implementation plan. Include a narrative that explains and justifies each budget item and how amounts were calculated. You may provide the budget and budget narrative as an attachment.
The budget and narrative is provided as a combined attachment.

6. Describe your understanding of the challenges inherent in implementing your Solution and how you plan to address those challenges.

Challenges to implementing the CASET solution, or any new technology, may fall into the category of scientific, engineering, or an organizational challenge. In this section, we raise some of these challenges and our respective mitigation approach.

Figure 1 Emily Kennedy and Steve Blank, recognized for developing the Customer Development Methodology

When developing data mining solutions, a major challenge managing expectations. Data mining, especially via machine learning and natural language processing, while it can produce highly accurate results, is not perfect. To this end, it is important that this is communicated to users. However, we plan to mitigate any inaccurate results by providing a feedback mechanism in which users can communicate any erroneous results that they find. This feedback will be incorporated into future rounds of model retraining and calibration.

Allegheny County DHS is an early adopter of technology with ambitions to leverage today's artificial intelligence computing to extract actionable intelligence from within its unstructured case notes' data. The potential of AI may be limited by the characteristics and qualities of the data. The results of the data mining research track will be unknown until further into the overall project. One mitigation strategy is the feature agnostic design of the caseworker view of the proposed solution. The overall CASET architecture and visualizations are amenable to react to research phase to appropriately "hook" into the discovered features and attributes on the front end for all views pertaining to caseworkers, along with supervisors and service providers. Marinus Analytics believe there is a base-value in the aggregation of notes across systems and features enabled via software engineering, with added-value stemming from what is produced via the research effort.

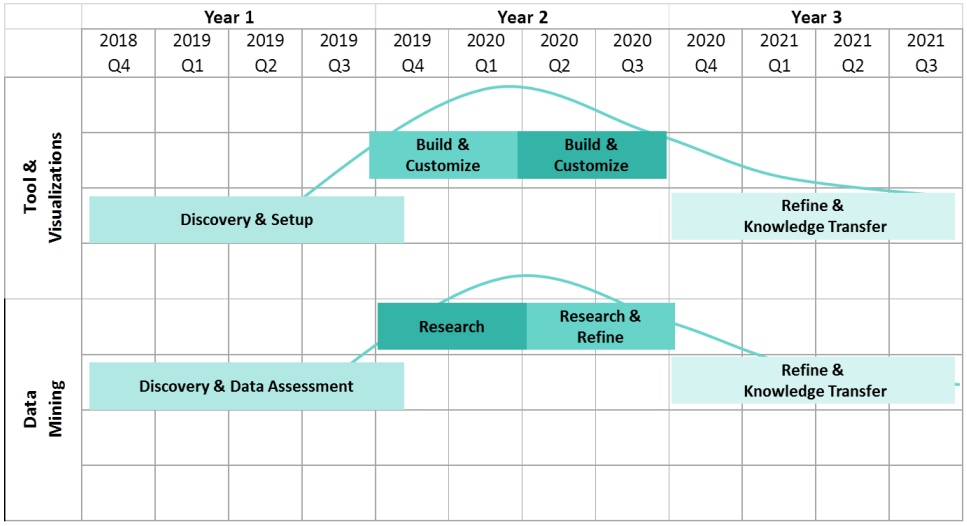
From an engineering perspective, challenges may occur such as data consistency between CASET and the source-of-truth. If the original record gets edited after-inception, the CASET system will need the ability to synchronize the older recorder. This should be manageable assuming the parent database maintains time stamps or awareness of record revisions. If the case note is in a document form, the latest date will reveal the time of edit.

Additionally, we may encounter a challenge in moving the information between the legacy system and CASET. It may be that the older legacy systems are not able to be modified to push data into our system, but, in this case, CASET will be made to pull the data at regular periods from the legacy data sources. Alternatively, security concerns may not allow data to be queried from CASET, but, in this case, the legacy system can be modified to push data into a real-time system from CASET. Finally, there may be

obstacles in convincing existing vendors to push data into our system, but, in this case, we can again pull data directly from the existing vendors into our system and keep the results open so that CASET can be further extended by either Marinus Analytics or another vendor.

Operationally, the goal is to achieve user buy-in and input into the developed system. Later, it is important to attain adoption through good-usage of the technology over time. Our approach to engaging and winning over the target users is by the manner in which we match the solution to their expressed pain-points. This approach is known Customer Development Methodology. A major promoter of this methodology is the National Science Foundation (NSF) who funds academic institutions and small businesses to advance fundamental science in ways which will eventually benefit society. As a part of the NSF I-Corps Berkeley Cohort in 2014, Marinus Executives Emily Kennedy and Cara Jones engaged in the intensive training program which created a catalyst to launch the product Traffic Jam despite known obstacles for innovating in the law enforcement sector. We learned in-depth knowledge and practical application from experts at the NSF and University of California Berkeley; the training was specifically focused on the Steve Blank customer discovery process, which we still currently use for new product development. The core of this process involves "getting out of the building" to test assumptions and hypotheses about user needs, and validate or invalidate them in the field. This has been our guiding methodology to insure that any software solutions we build meet real needs of users on the ground.

7. Provide a timeline for the design and development of your Solution.
The following graph illustrates the two major tracks of this project which will interact harmoniously to deliver the CASET solution to the client. The arch in the illustration emphasizes the intensity of the software engineering and research to fully mature CASET. During the first year, a discovery process will be undertaken to understand the data paradigm and test hypotheses through client specialist interviews. Year 2 will be heavily focused on maturing the CASET tool and conducting research for the extraction of actionable insights from the unstructured data. In the last year of the project, we envision an operational CASET tool and our support will be primarily focused on knowledge transfer, enhancing the robustness of the overall system, remediating any uncovered issue, and/or servicing any feature requests. During this tail end of the implementation, staffing will be reduced to the senior experts of our team to successfully finalize and transfer the solution. Following the end of the project, Marinus Analytics is open to operating the CASET solution for Allegheny County DHS in a SaaS subscription manner, similar to our operation of Traffic Jam. From the start of the project, we will follow a lean development and MVP maturation style which will allow the client specialists to begin exploring the tool while it is still being developed. This approach allows for rich feedback based on “hands on” review by subset of users before the tool is officially rolled out. We anticipate full rollout could be achieved at the beginning of year 3 with the remainder of the project time being spent on refinement through operational feedback and then knowledge transfer based on subsequent maintenance strategy.



8. Describe your organization's plan to collaborate with DHS during development, implementation, knowledge transfer and training about how to use and maintain the Solution.

Critical to the successful implementation of any data mining solution is continued interaction with the customer. We will conduct a thorough requirements review during which specific components of the data mining solution will be identified, as well as the specific major and minor categories of events, issues and needs.

Throughout the research and development and implementation stages, Marinus will provide initial results of data mining accuracy, as well as incremental results for feedback.

During the tool development stage, Marinus Analytics will seek input from DHS specialists and champions on the requirements and design of the system. Marinus Analytics will conduct user discovery interviews to test hypotheses of the functionality of CASET prior to executing the research and engineering to gain confidence in how the system will best support the underserved pain-points of the organization. Marinus Analytics will support DHS in understanding and mitigating hurdles for adoption and to determine ways to align incentives to promote usage of the proposed solution. Marinus Analytics has leveraged this methodology since our participation in the 2014 NSF National I-Corps cohort. Hypothesis testing of project assumptions is a lean approach and is designed to allow DHS to uncover any unforeseen issue before implementation, when the design is most flexible to changes and adjustments.

These interactions will result in the following products:

- User Interface Wire Frame diagrams
- Formal use cases
- Entrance and exit criteria for project milestones

The proposed timeline includes "first look" opportunities to keep DHS informed of the solution long before the formal final release of the tool. This will ensure feature development meets the needs of DHS personnel and contracting agencies. In addition, project champions will be empowered by this additional lead time to proactively prepare for policies, adoption, incentives, and awareness building.

Throughout the course of the project, Marinus will provide bi-weekly progress summaries and status updates. As Marinus is a Pittsburgh-based company, we are available for periodic onsite meetings as part of these touchpoints.

As the project nears end of year 2, we will provide test results and status of any outstanding issues being mitigated. We will create training materials and conduct train-the-trainer sessions to transfer knowledge on the intended usage and best practices for maximizing utility.

Marinus Analytics engineers will either outline a strategy for ongoing SaaS offering of CASET or support the DHS IT team in operating the tool through the delivery of manuals for maintaining the system from an IT perspective. We will also make our team available for in-person walk through of operating materials and transition of responsibilities in the last half of year 3.

9. Describe how your organization will evaluate the success of your Solution. Provide an example of how you measured the impact and success of a similar project in the past.

On past projects we have measured the success of our solutions in a variety of ways. In general, we view the definition of success to transcend merely determining whether we meet the goals of a project. As a company with a social impact mission, for Marinus Analytics the best measurement of success is how our work impacts not only users and their workflow, but ultimately also the customers they serve, whether those are victims of crime or human services clients.

For the purposes of this proposal, there are some specific measures that we will apply with respect to data mining. In data mining in general and text classification in particular, the two most important measures of accuracy are precision and recall. While both are measures of accuracy, each measures a different type of accuracy. Precision refers to the percentage of true positives extracted from unstructured text. For example, if we extract a total of 10 categories from a document, and nine of them are true positives, then the precision is said to be 0.9, or 90%. However, even if we accomplish perfect precision, or 100%, that says nothing about how many categories were missed, or false negatives. This is where recall comes in. Recall is the ratio of categories that were extracted and the total number of categories that are in a document. For example, if a document contains a total of 10 mentions of categories that we want to extract, and our algorithms correctly extract a total of eight, then the recall value is 80%.

A third measure, the F1 score, also known as the Dice Similarity Coefficient (DSC), is the harmonic mean of precision and recall, and ranges between 0 and 1. Trained text classifiers that have an F1 score of 0.85 and above are considered to be reasonably accurate, depending on the area of application.

In order to measure the above, hand annotated documents are divided into a training data set and a test data set. Typical the ratio is about 70% training data and 30% test data. A model is trained on the training data set, and then tested for accuracy on the test data set.

A further measure of success will be how users react to the system. Because this will be new technology for DHS, it is important that users trust the system. A system that produces a lot of false positives and/or false negatives will quickly lead to users losing trust in the system. At the same time, it will be important to convey to users before the beginning of the project that data mining is not 100% accurate and that were will be errors.

10. Describe why you want to serve human services clients, your experience in adapting technology to serve human services clients and your plan to adapt your Solution to meet client needs for this track.

Marinus Analytics was created in 2014 to help populations who could not ask for help, namely victims of human trafficking. Making a positive social impact is part of our company DNA. We came out of research at Carnegie Mellon University, after speaking with hundreds of law enforcement agents about the difficulties of finding these victims and prosecuting their exploiters. We saw a huge need for detectives to be able to harness massive amounts of online data to inform actionable insights for investigations, and we developed research—and ultimately created software tools—to address this issue.

Through our work thus far, we have taken advanced data mining and machine learning technologies and productized tools to deploy them to the average law enforcement or government user. We have insight into the needs, desires, and pain points of local, state, and federal government agencies. Because human trafficking is an evolving problem, we had to make solutions that were adaptable to the changing landscape.

We have seen the impact software can make on the workflow of government workers, and our software for human trafficking empowered the rescue of hundreds of victims. We are passionate about developing advanced technology to support public agencies in doing more with their existing resources.

Marinus Analytics and its Traffic Jam technology are funded in-part by NSF through the Phase II Small Business Innovation Research (SBIR) program. As the proposed CASET solution is an augmentation of the Traffic Jam solution, supplement funding is available to support this project. In effect, the supplement grant, called a Phase IIB, will match the contract award and provide up to $250,000. We are highly motivated to pursue this work with Allegheny County DHS as it is a natural extension of our focus on Traffic Jam. The delivered solution will not only benefit Allegheny County DHS but will benefit human service organizations across the country because Marinus Analytics would provide the resulting technology as a SaaS tool offering. Therefore, it is appropriate to support Allegheny County by applying these funds to an awarded contract to discount the cost of R&D for the CASET solution.

RFP Response Form

RFP for Unstructured Data Analytics Solutions

2. Developing Tools and Visualizations (75 points). If you are proposing Developing Tools and Visualizations, fill out the questions below. If you are not proposing Developing Tools and Visualizations, leave this section blank. Your response to this section should not exceed 25 pages.

Author note: *For purposes of this proposal, we will refer to the solution as "CASET." This name suggests that the solution is a smaller companion tool ("ette") of larger enterprise case management systems, operated at Allegheny County Department of Human Services. Inspired by the pop-techno-icon "cassette," the tool will rewind then play back the case notes in different ways to your liking for actionable intelligence in the field... Sadly, it will not do so to the audible sounds of your favorite mixtape.*

Overview

Marinus Analytics is proposing a cloud-based open architecture solution, CASET for the RFP (Request for Proposals), "Unstructured Data Analytics Solutions." CASET is an augmentation of our flagship tool, Traffic Jam, which is a software-as-a-service (SaaS) offering to provide actionable intelligence to safety and justice government organizations from hundreds of millions of online public records.[2] The strength of our experience is in providing tools based on advanced computing which are operationally trusted, reliable, and impactful.

There are many analogies of the business purpose of Traffic Jam to the needs identified in the RFP. Traffic Jam is an AI-driven tool for information discovery of unstructured data. It is designed to detect patterns of at-risk and exploited individuals in online content. Its powerful capabilities are accessible online and allow users to access critical insights and track individuals from the field, on laptops and mobile devices. Similar to the workflow in human services, users of Traffic Jam have greater knowledge of a subject's history, which empowers the in-person responsibilities of conducting interviews and determining the next steps of an investigation. With over 1,500 police, prosecutor, and analyst users to-date throughout agencies in United States, Canada, and the United Kingdom, Marinus Analytics maintains a professional track record in the design and delivery of a modern analytical tools for unstructured data.

CASET is an expansion of the Traffic Jam core. If awarded, the platform will be tailored to the data paradigm and workflow specifications at Allegheny County DHS. We will work with you as outlined in the answer to Question #8 to ensure parent systems interface with CASET to produce real-time, actionable insights from the mined data. CASET will be a data-driven solution, unlocking the potential of information to benefit caseworkers, service providers, and supervisors' workflows. This section will provide an in-depth technical description of the CASET system including the structure of the application

2

and the primary tool functions to support the operational responsibilities surrounding case notes in human services.

Open Architecture

CASET is a modularly designed system that utilizes a service-oriented-architecture (SOA). The core platform is web-based with support of cloud hosted infrastructure and various sourced and developed components. The design considerations are motivated to deliver the best user functionality and to leverage the latest advancements in computing. In addition, the engineered solution addresses the ease of maintainability (whether the DHS decides to fully operate the technology or prefers Marinus Analytics to manage the tool over-time in a SaaS manner). Further elaboration on the topic of adaptability and scalability are addressed in the "additional options" paragraph below and in the answer to Question #2 of the proposal. An open architecture approach to tool building lends itself to enhancements over time through the capacity to add and exchange components to the core stack.

Conceptually, the CASET system is comprised of four major layers: (1) system interface from the enterprise/legacy source-of-truth, (2) unstructured/structured data storage, (3) data mining processes (and any on-the-fly interpretations), and (4) front end visualizations. Within these layers, multiple components may interact to perform the overall function. The diagram and subsequent narratives below describe the layers in detail.

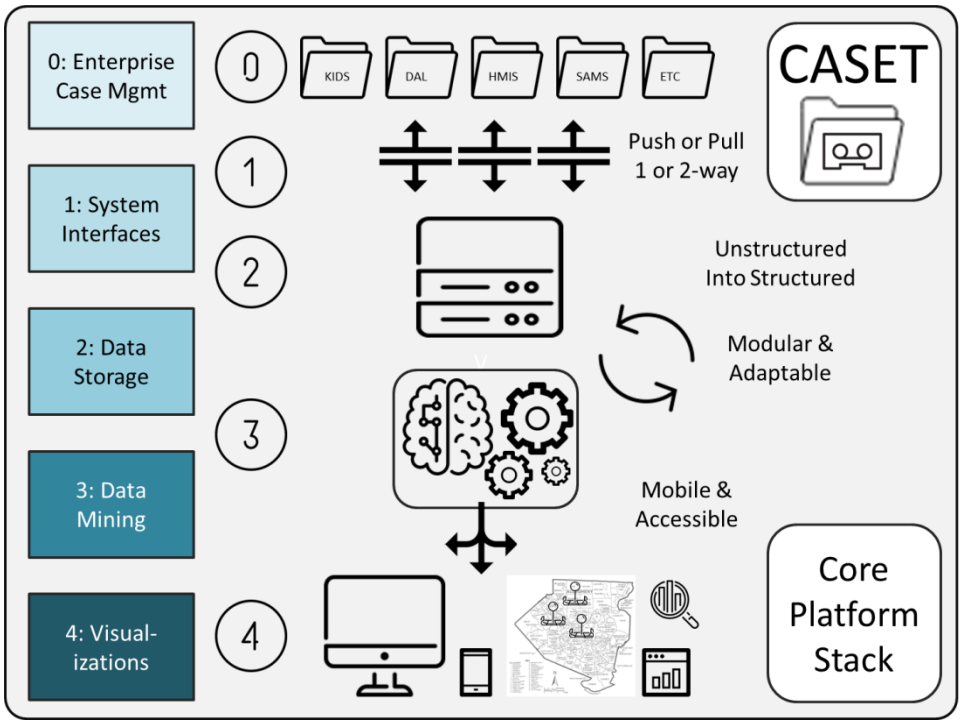


Figure 2 CASET platform architecture and its four layers

Layer 1: System Interfaces

Depending on the interoperability of the enterprise legacy systems at Allegheny County DHS, along with the security and practical considerations, we will build one of two options, with regard to the system interfaces. The interfaces can be constructed to push or pull data between the source-of-truth and the CASET platform. Each option comes with its own advantages.

In a push based system, information from DHS's data sources will be fed into CASET as soon as a record is collected. Because we are receiving information in real time, CASET will then process that information and have it available to the end user within a matter of minutes. Some of DHS's existing data processing systems would have to be modified in order for this to be achieved, and CASET would need to be online at all times. We would recommend these messages be sent in JSON formatting, but CASET could be engineered to accept any type of document.

In a pull based approach, CASET will periodically pull the data storage of the origin case management system to fetch new information that has been entered. CASET will then initiate a batch executable on the newly gathered information and update for the end user. The length of the period between each pulling of information will depend largely on the performance of the origin application servers. If there is shown to be no negative impact on the servers, the batch jobs can be run every few minutes, but, if there is shown to be negative impact on the servers, CASET will wait until a low volume time of day, like midnight or 2AM, to perform a daily pull. Assuming the origin system maintains either timestamps for each record and/or auto-increment identifiers, CASET will be easily able to determine which records are new since the last batch update.

Each of these two approaches would require similar amounts of time and resources to implement. The decision on which to pursue would largely be dependent on the needs of DHS, as well as the technical capabilities of their current systems and development resources.

Layer 2: Data Storage

For technological control over the application, it is our initial recommendation that the CASET tool will include its own database. However, we will strategically consider how user and case information co-exists and synchronizes in this system in accordance with the data landscape at Allegheny County DHS. To optimize the functionality and deliver immediate impact, it will be very important to capture certain available structured information along with the unstructured case notes. For example, the author, client, and supervisor identifiers associated with the note may be required and additional fields may be value-added beyond what can be derived from the unstructured data. The fields per legacy system will be incorporated as part of the application interface (API) and JSON object and made available to CASET, as a result.

CASET access to the relevant records from the legacy management systems will allow for meaningful aggregation and visualization in the near term while the data mining research commences and is pursued over the duration of the contract. This ensures CASET will be guaranteed to have value,

independent of the final research outcomes. Storing a "local" copy of the original case records ensures maximum performance to "drill-down" by users to read through the details, as needed or desired.

Layer 3: Data Mining Engine

The "engine" to mine the unstructured data will be comprised of a portfolio of processes ranging from regular-expressions, statistical methods, and artificial intelligence based feature extractors and classification models. These insights will be stored alongside the original records in the database and provide for greater summary information on the front end. Marinus Analytics is proposing to fulfill the data mining research role of this contract. However, Marinus Analytics is highly motivated to deliver the tool and visualizations functionality and work harmoniously alongside whomever will be awarded to conduct the research aspect of the project. The architecture design is an accommodating modular structure for research products. As the research produces new modules, these components will be tied-in based on our extensive experience to operate AI-based feature extractors and pattern recognition services. As stated in the next section, this deep understanding of the stack and our long history of collaborating with academic teams to productize research directly influences the core platform design. Our approach to visualizations is conducive to unstructured data and, to an extent, agnostic of what text-based features the research may produce. Therefore, we will be ready to harness new features, as they are identified, and "hook" them through to the front end in a lean/dynamic engineering manner.

Layer 4: Visualizations

We briefly mention the architectural construction of the visualization layer here and will go into in-depth scenarios of how the front end will empower actionable insights to serve the needs at Allegheny County DHS, in the section on user perspectives below.

The construction of the front end, which is responsible for rendering the visualizations per user type, will use components including but not limited to a traditional web-stack and web-server, search engine, job queue, build tools (for code pushes), and email notifications.

The core also contains some of the logic which will add important utility to the tool. This may not be served in the data mining portion of the contract but rather in the tool functionality that aggregates or assesses the data on-the-fly. A backend layer will handle these algorithms adjacent to the API calls to the CASET database.

The modularity of the architecture extends to the styling of the application. CASET will easily reflect the appropriate department brand. Per Allegheny County DHS specifications, the "skin" of the interface will be configured to portray the logo graphic, colors, and informative copy to appropriately fit within the existing tools of the organization.

Additional Options

Due to our deep experience incorporating features from data mining research into a productized tool, the proposed architecture in CASET is amenable to future "extra" components to be integrated into the tool. In Traffic Jam today, a service exists to provide chat-based help to users real-time. This service is provided by a company called Chatlio[3] and is an example of a SaaS tool which Marinus Analytics subscribes to in order to enrich Traffic Jam's offering. This same component may be leveraged by Allegheny County DHS if they felt the communication channel is supportive of escalation and workflows.

Recent offerings in SaaS based components allow for some remarkable AI tools". Emerging offerings for speech-based components like AWS's Transcribe and Lex exist to provide radical innovations to the user experience. For instance, challenges like the documentation burden for practitioners to capture notes can be remedied using speech transcription input. Having a tool which allows the notes to be spoken and transformed to text could dramatically reduce any latency input delay and improve caseworkers' day-to-day workload. This RFP only seeks to analyze case notes which imply no input into the legacy information management systems. However, with a tool like CASET in-place, the further ability to prototype speech components in a lean and cost-effective way would be possible.

Other future components involving document optical character recognition (OCR) to convert other sources of case notes into machine readable content is another additional module for consideration. This type of tool along with other novel communication channels like SMS could be leveraged in a similar manner as speech technologies to provide new forms of input going through CASET into the legacy case management systems. Components like these continue to rapidly be available over time, providing the opportunity for even more new feature integrations over the period of performance.

CASET Visualizations

CASET is designed to have visualizations and functionality oriented to the varying roles and responsibilities of the professionals working in human services under Allegheny County. There fundamentally will be three primary perspectives built within the front end of the system, for the caseworkers, service providers, and supervisors, respectively. We provide an in-depth explanation of our view for the caseworker followed by an overview of the supervisors' interface experience.

It is important to note that any part of our proposed solution is open to DHS feedback and requests. This is particularly true with respect to the visualizations. Marinus Analytics is suggesting how the CASET may be developed as a solution based on the Traffic Jam platform which also operates over largely unstructured records. DHS has a more intimate knowledge of the case management systems and the gap in what those systems provide and what should be covered by this new tool and its visualizations. Our team is adept at prototyping and coding per requests and direction by DHS experts. In the answers to Questions #3, #9, and #10, we also highlight our professional experience for performing "user discovery" and lean engineering to maximize resources to deliver a solution which is optimally fit to

[3] https://chatlio.com/

needs of the DHS organization. With this said, the following is our recommendation for the caseworker view which has thoughtfully been crafted per the requirements expressed in the RFP.

User Perspective: Caseworkers

For the remainder of this section, we are suggesting an approach based on Traffic Jam's "trail" feature for the main visualization for caseworker users. CASET functionality does not replace the preparation responsibilities for the caseworker to read through previous case notes. However, CASET will greatly empower caseworkers by intelligently summarizing what may amount to dozens of past interviews and assessments for a given client. Allowing navigation of these notes, with a swipe of a finger, empowers accessibility, so users can effortlessly page through the critical points within each narrative while in the community or at their desks. Combining these characteristics, CASET embodies mobility, accessibility, and explainable intelligence as a modern tool asset for caseworkers and their supervisors.

CASET's embedded logic will aggregate related case notes by client identifier and connect client records across systems. Figure 2 depicts the caseworker view; a timeline shows the temporal history of case notes with each box positioned on the axis according to its creation or service date. The "split-by" feature allows user to view a summary of features by color representation. Features such as document type may contain the attribute value service record, referral snapshot, investigative summary, standardized assessment, etc. This is an example of a structured field that is already available in the current notes. The advanced extraction of structured information using natural language processing, covered in the data mining task order, will generate new features from the unstructured data. Figure 3 illustrates the drop-down menu of selections within the "split-by" feature to allow readers to benefit from the summary information intelligently gathered from the case notes.

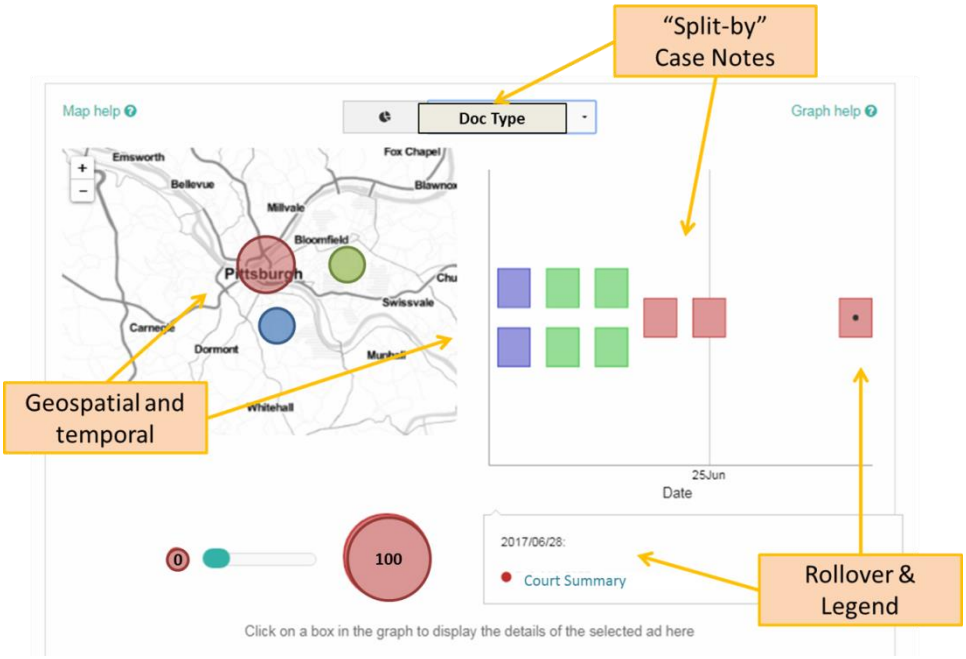


Figure 3 Geospatial and temporal information of notes and their summarized features

The color tagging of attributes of a feature will be strategic. For example, features involving sentiment will have red hues for negative attributes and green hues for positive sentiments. As the machine learning may generate attributes with varying degrees of confidence, a "confidence score slider" will be available under the timeline to adjust the displayed data. For instance, a user may choose to only view attributes for a given feature, where the computer had a confidence over 95% to eliminate noise in the results. By selecting a specific box/case-note, the user may "drill down" and read that note in its entirety. Alternatively, a user may scroll down the page to view the chronological order of case notes.

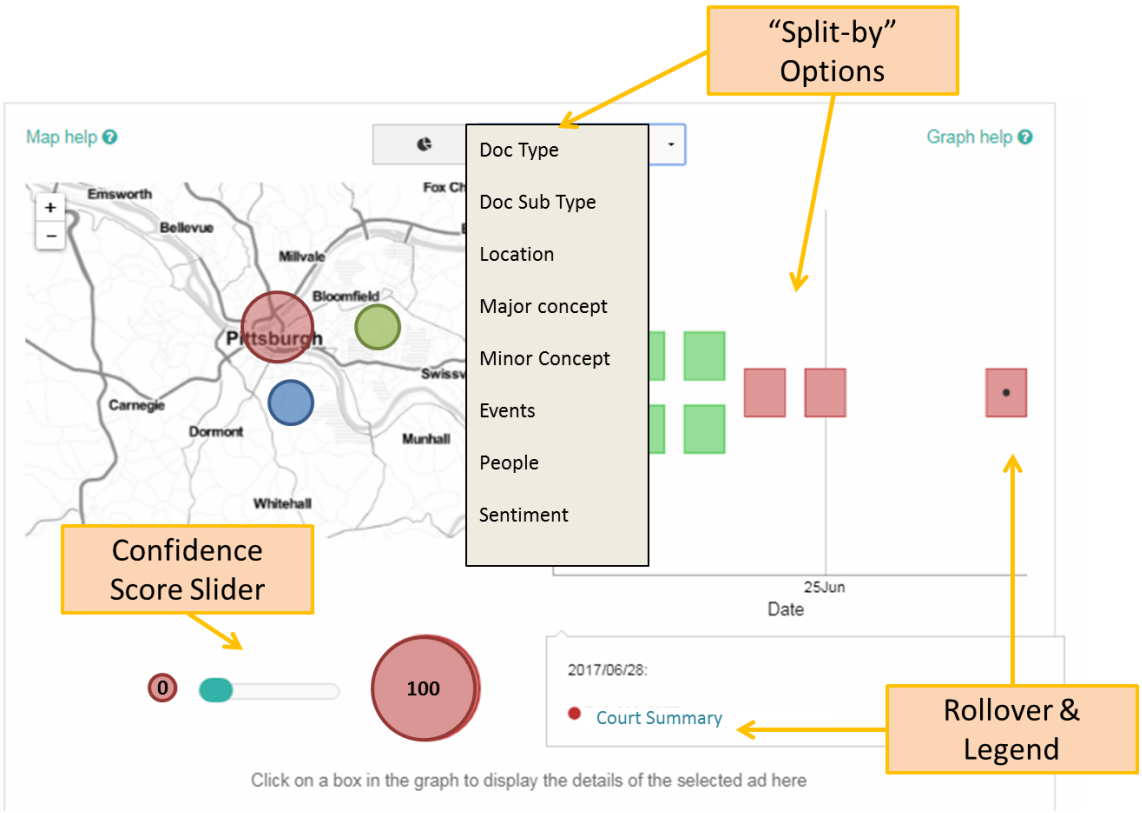


Figure 4 Data mining track will provide portfolio of structured features to be viewed on the front end using "Split-by"

Below in figures 4 and 5, the concept of selecting a specific note on the timeline and drilling-down into the actual case record to read in its entirety is displayed. In this particular scenario, the case worker is scanning the history of note by "minor concept." By rolling over, the history of a couple dozen documents, the user is able to quickly see the computer-identified summary of minor concepts contained within these notes. The most recent note contains three concepts highlighted in red. A rollover of the note brings up a dynamic legend with the actual values, "missed appointment," "mental health issues," and "drug use." The caseworker wants to know more, so she or he clicks on the note to bring up the full narrative. To deliver explainable intelligence, the note has color-highlighted-words and phrases which tie back to the features and attributes discovered by through the data mining process.

These concepts of providing actionable intelligence to the caseworkers, using the unstructured case notes from the enterprise case management systems, is reiterated in a four-minute video, provided along with this proposal response. The video is provided in the submission package and is also posted online on this private YouTube link: https://youtu.be/o1CZbDqJPLE

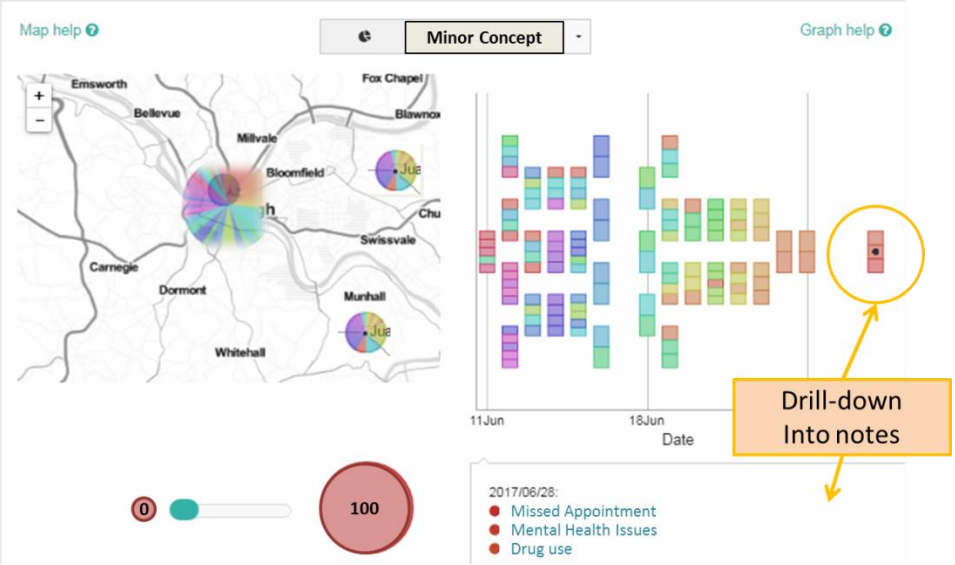
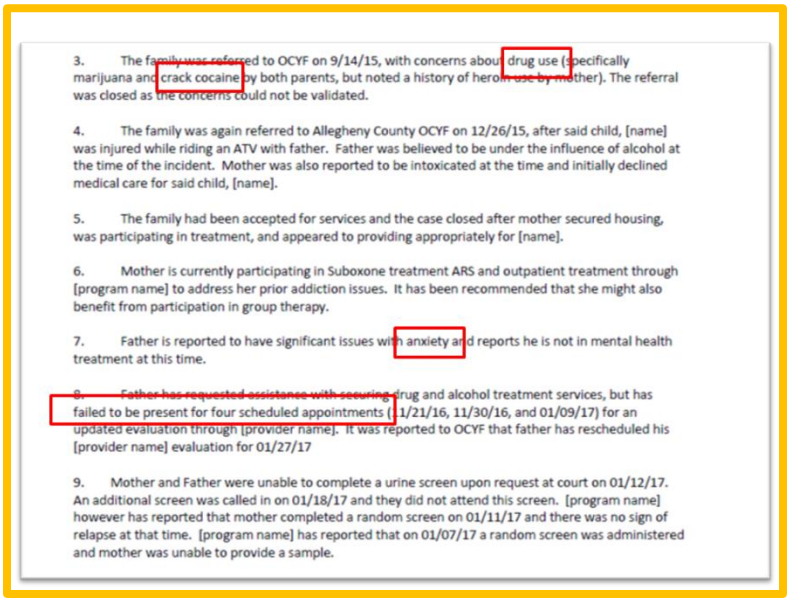


Figure 5 Scroll-down or click on the box to drill-into the actual case notes to read them in their entirety



3. The family was referred to OCYF on 9/14/15, with concerns about drug use (specifically marijuana and crack cocaine by both parents, but noted a history of heroin use by mother). The referral was closed as the concerns could not be validated.

4. The family was again referred to Allegheny County OCYF on 12/26/15, after said child, [name] was injured while riding an ATV with father. Father was believed to be under the influence of alcohol at the time of the incident. Mother was also reported to be intoxicated at the time and initially declined medical care for said child, [name].

5. The family had been accepted for services and the case closed after mother secured housing, was participating in treatment, and appeared to providing appropriately for [name].

6. Mother is currently participating in Suboxone treatment ARS and outpatient treatment through [program name] to address her prior addiction issues. It has been recommended that she might also benefit from participation in group therapy.

7. Father is reported to have significant issues with anxiety and reports he is not in mental health treatment at this time.

8. Father has requested assistance with securing drug and alcohol treatment services, but has failed to be present for four scheduled appointments (11/21/16, 11/30/16, and 01/09/17) for an updated evaluation through [provider name]. It was reported to OCYF that father has rescheduled his [provider name] evaluation for 01/27/17

9. Mother and Father were unable to complete a urine screen upon request at court on 01/12/17. An additional screen was called in on 01/18/17 and they did not attend this screen. [program name] however has reported that mother completed a random screen on 01/11/17 and there was no sign of relapse at that time. [program name] has reported that on 01/07/17 a random screen was administered and mother was unable to provide a sample.

Figure 6 Drill down into the case note. The narrative has color highlights to explain why the system tagged particular features to this note.

User Perspective: Supervisor and Service Provider Screens

In addition to the in-depth review of the proposed visualizations for caseworkers, we will discuss analytical viewpoints that will benefit the supervisors and service providers. During the user discovery phase of the project, we will interview and test our hypothesis for these views' design with the specialists representing these user communities. This phase will hone the engineering direction to best

fit the CASET functionality to the quality assurance needs and case workflow of the supervisors and service providers. Further views will potentially generate new insights for emerging trends and the geospatial occurrences of events. For example, the human services effects of drug addiction may be mapped in CASET to display incidents revealed from case notes over space and time. These views may speak to the Data Analytics organization in addition to the user groups targeted in the RFP. Below are a selection of visualization hypotheses to be considered for the final solution delivery. These include views for quality assurance of service providers, supervisor review of subordinate activity, and macro-intelligence insight over county-wide case notes.

Service Provider Quality Assurance

This module will use statistical analyses to provide an overview of the performance of individual service providers. The types of data displayed might range from aggregate statistics on the number of services provided, to performance over time in order to highlight improvements or declines in performance, as well as possibly sentiments extracted from case notes that involve a specific service provider. The exact functionality of this module will be determined during the user discovery/requirements gathering process.

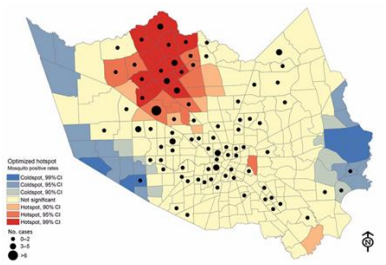
Supervisor Review

The supervisor review module will allow supervisors to query for and review specific cases. It will show structured data extracted from the various in-house databases, the original text of any case notes, as well as data extracted from the unstructured case notes. The module will also allow supervisors to query for those cases that are most in need of attention, as determined by sentiment analysis and categories of issues, events and needs that are deemed most applicable to those cases. Again, the exact features of this module will be determined during the user discovery/requirements gathering process.

Macro-Intelligence Insights

Heat maps can be used to highlight areas particularly susceptible to certain problems, such as drug use. One manner of producing a heat map is by using a Kernel Density Estimation (KDE) approach. Another example is a choropleth map to visualize city neighborhoods that, for example, have higher or lower incidents of a certain kind. The dots are the actual locations of incidents. Allegheny County has its own mapping division which generates geo-files for neighborhoods, census tracts and census blocks, which typically are used for choropleth mapping, as well as street maps. A grid-based approach may be employed, to divide the county into grids; this allows for other types of spatio-temporal analyses. The "boxes" in choropleth map are polygons, representing a neighborhood or census tract, for instance. The advantage of using census tracts is that the demographic information from the census can be combined with Allegheny County DHS data to create statistical models for prediction, outlier detection, regression, etc. In GIS, the incidents are counted in a polygon by geocoding addresses (which attaches coordinates to an address), and then result points are overlaid with census tract or other polygons. In turn, the visualization provides a view into the aggregated incidents for each polygon and each time period,

allowing for space-time analytics. The figure below displays Kernel Density Estimation and Choropleth examples. Again, these visualizations would be possible by incorporating the insights into the generated features and attributes of case notes from the unstructured data, achieved via the data mining track.



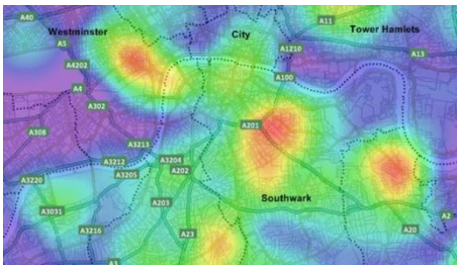


Figure 7 Example of Choropleth Map and Kernel Density Estimation Sources: https://wwwnc.cdc.gov/eid/article/23/8/17-0384-f2 https://www.caliper.com/maptitude/sample_slides/sample_8.htm

CASET Additional Features

In addition to the visualizations, CASET may delivery functionality such as “watchlist” notifications, “early-warning-alerts,” and personalized dashboard of cases and basic case management features to organize workload. These are all existing features in Traffic Jam which were developed to best serve investigators working cases involving exploited victims or proactively identifying at-risk individuals. Watchlist allows a supervisor to list a caseworker, service provider, or client and receive email notification whenever a new record is pulled into CASET which involves any of these entities. If a supervisor is “following” these entities for support or performance review, he or she will be informed of new information to support their oversight role. Alerts work similarly but instead of an email, the notification will be listed on a page within the front end. If certain topics such as suicide-risk, police-involvement, violence, or hard-drug-use are detected in the case notes, the corresponding case records will be ranked on the alerts page for review as the most concerning records out of the daily volume of captured notes. Similarly, quality assurance issues mined from CASET could also be available on its own alert page. These may be issues related to client appointments and complaints denoted in the interviews and other case notes which are extracted into features and attributes. The algorithm or logic will be tailored as recommend by DHS. Dashboard and folders allow any users to “pin” particular notes for ease of recall when logging out and back into the CASET tool.

Maintenance and Operation

In the budget narrative and implementation plan, there is an upfront phase which will involve research and development to apply CASET to the data paradigm at Allegheny County DHS. Marinus Analytics will support DHS in the knowledge transfer and maintenance considerations for CASET in the tail-end phase of the project which is proposed over the third year. Marinus Analytics is interested in maintaining CASET for Allegheny County DHS as a SaaS subscription (similarly to how we maintain Traffic Jam).

Again, this is discussed in the answer to question #5, along with our concluding remarks in the answer to question #10.

Conclusion

Marinus Analytics is proposing the CASET solution, derived from its flagship Traffic Jam SaaS offering for providing actionable intelligence from unstructured case notes. Marinus Analytics is highly motivated to deliver the tool and visualizations and will work harmoniously with whomever is awarded to perform the data mining research (though we have applied to be considered for both responsibilities.) As a local and women-owned small business, serving Allegheny County DHS will incorporate our expertise for delivering AI-driven, modern SaaS tools as part of your information technology portfolio.

2. Describe how your Solution addresses DHS's goals and is scalable to allow for expansion to address future needs.

Scalability

Each component within CASET is highly modularized and will pass information between all sources and sinks in an open format, such as JSON, which allows for any module to be plugged in, as needed. Additionally, every data storage component will be stored in either clear text or encrypted with a universal key; in other words, no information will be stored in binary format. These two pieces will allow any additional component developed by Marinus Analytics or by a 3rd-party to have easy access to all the results from the APIs. If another application requires data to be processed, it can request the information from the APIs, convert the data into a new format, and send it back to the API to be stored accordingly.

Alternatively, since the data will be stored in industry-standard data sources such as PostgreSQL or Neo4J, 3rd-party applications with appropriate security credentials will be able to access the data directly. This will allow them to bypass APIs that transform data in particular ways and also allow them to extend the databases directly to store their new information in a centralized source.

These methods have been proven to work within Marinus Analytics web-based offerings and APIs. Our custom API can return over 100,000 API requests that query a dataset in the millions of records. Our customers have never reported down time or performance issues. The website uses the same data storage backend the API uses with no impact on each other's performance, which speaks to the ability of our systems to expand in a very scalable way.

For a scalability point-of-comparison, Traffic Jam maintains a growing bank of over 210 million public unstructured records. The volume of data listed in Appendix A of the RFP will not be of scalability concern. The search functionality within the core of Traffic Jam allows for on-the-fly pattern recognition across substantially larger data volumes where respective queries return in a few seconds with no impairment to the user experience. The current user base of Traffic Jam involves hundreds of active users who use the system concurrently.

Future Needs

To-date, Marinus Analytics has leveraged cloud architecture to optimize cost and scalability. We are experienced engineers in understanding the full stack of components along with the emerging eco-system of stack enhancers which can be procured and incorporated to benefit a given tool or solution.

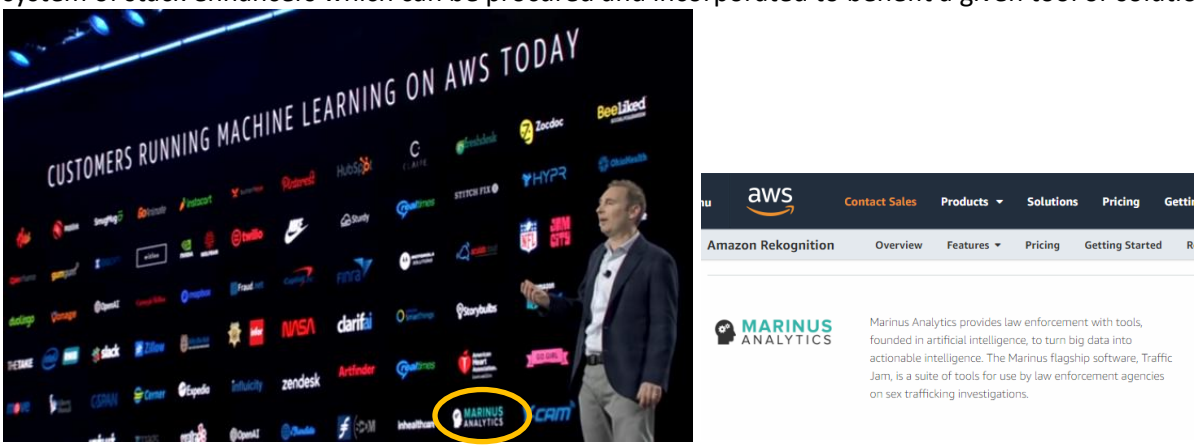


Figure 8 AWS CEO, Andy Jassy, specifically mentioned Marinus Analytics in his keynote speech at re:Invent conference, November 2017 in Las Vegas, NV and Marinus Analytics is listed on AWS website

As mentioned in the answer to the previous question, CASET architecture lends itself to new modules which will provide radical innovations to the traditional case management information system behavior. These are modules which can support speech-to-text transcription, audible play-back of notes, and of course the latest artificial intelligence products to further the data mining engine.

Figure 7 illustrates our relationship with AWS. We strive to maintain the latest knowledge of commercially available cloud services to support delivery of AI-driven capabilities to the public sector. In the world of AI, there is a phenomenon of "two-year-leaps," where progress is rapidly advancing in short timespans. Maintaining a close pulse of the innovations by leaders like AWS along with our close ties to the academic research community, allows Marinus Analytics to best prototype and commercialize cutting-edge tools for organizations like Allegheny County DHS. As discussed in the answer to question #4, our team includes a faculty member from Carnegie Mellon University who advises our highly-skilled scientists and engineers on academic breakthroughs which may be translated into practice. These diverse skills of our team will allow for innovative solutions to best serve the needs of your organization going forward.

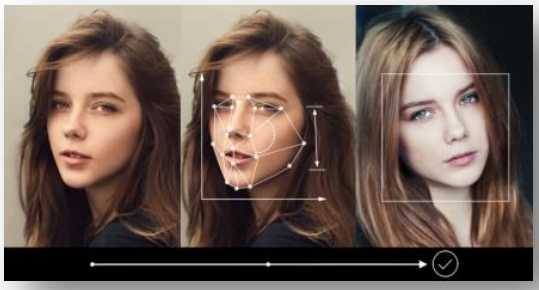
3. Describe your organization's experience developing, implementing and evaluating tools and visualizations for use with unstructured data.

Marinus Analytics is the exclusive provider of Traffic Jam which uses the latest advancements in AI to turn big data online into actionable intelligence, for the rescue of exploited human trafficking victims and protection of at-risk populations. Traffic Jam, on which the core of CASET is based, is an implementation of previous academic research at Carnegie Mellon University that underwent extensive rebuilding and scaling to be productized and available for customers. The majority of this work, performed by Marinus Analytics engineers, took place in 2015 and 2016. As the data archive grew to

tens of millions of records, the earlier version of the technology degraded in availability, experiencing uptime at 50% and would often fail during key business hours. This research software was re-written and ported into more stable environments and storage mediums with more reliable web interfaces to increase uptime to 99.9%, and the archive has now exceeded hundreds of millions of records.

Traffic Jam is available on any desktop, handheld, and mobile device. The technology includes a powerful elastic search engine, machine learning algorithms, image feature extraction, image similarity detection, and predictive analytics to make this large amount of data practically useful to public safety and prosecutorial organizations.

In June 2017, Marinus Analytics released FaceSearch with Traffic Jam to help detectives improve their ability to find specific missing persons, and apprehend their exploiters more effectively.[4] The feature has been heavily leveraged by our law enforcement user community and Traffic Jam is used daily by analysts at the National Center for Missing & Exploited Children (NCMEC), an organization which operates a national hotline and CyberTipline for reporting suspicion of crimes including child sex trafficking. It is very important to quickly identify and rescue juveniles who are being trafficked because exploited individuals and groups frequently move across cities and states.



Traffic Jam FaceSearch, deployed since June 2017

An extensive network of public sector agencies uses Traffic Jam on a daily basis. Other agencies include Department of Homeland Security (DHS) Field Offices, DHS Human Smuggling and Trafficking Center, the Department of State Diplomatic Security Service, U.S. Attorney's offices, and nearly 200 other agencies across the United States, Canada, and recently countries in the United Kingdom. Traffic Jam has proven instrumental, even for agencies with no in-house analyst. Indeed, many of our agencies have given us the same feedback: "The amount of time saved using Traffic Jam replaces the work of a full-time analyst."

Our ability to translate research into practice has been recognized through our participation in the IBM Watson AI for Good XPRIZE Competition. Marinus is one out of 59 teams from around the globe to be accepted into Round 2 of the XPRIZE Competition. The XPRIZE team explains, "The four-year prize competition aims to accelerate adoption of Artificial Intelligence (AI) technologies and spark creative, innovative, and audacious demonstrations of the technology that are truly scalable and solve societal grand challenges."

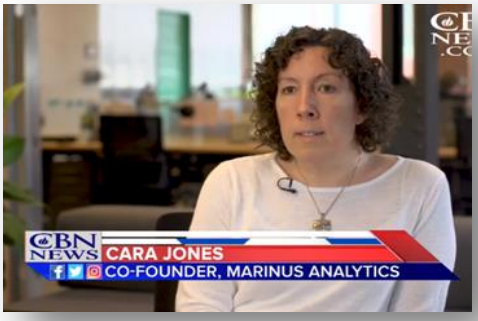
[4] Marinus Analytics, "Pittsburgh-Based Technology Company Debuts First Facial Recognition Technology Designed To Halt Global Human Trafficking," *Marinus Analytics Press Release,* Accessible at: http://www.marinusanalytics.com/articles/2017/6/27/face-search-debut

4. Provide staff bios (not CVs) of the key staff who will be implementing your Solution and identify the main point of contact. Describe your management structure and how it will support the goals of your proposed Solution. If you are partnering, describe the structure of the partnership.

Marinus Analytics is a women-owned and managed, small business operating in Pittsburgh, Pennsylvania. The company spun out of the Robotics Institute at Carnegie Mellon University in 2014. The staff at Marinus Analytics has extensive experience managing teams of computer scientists, software engineers, data scientists, and researchers. By embracing today's latest advancements in cloud computing, Marinus Analytics is highly effective in tool building and providing SaaS AI-driven capabilities to the public sector.

Cara Jones, B.S. Computer Engineering, MBA

Our engineering is led by Chief Operating Officer, Cara Jones, who has over 15 years of experience with high technology implementations. Cara Jones has managed the maturation and commercialization of Marinus Analytics's Traffic Jam technology since 2013. In her career, she has worked a spectrum of projects ranging from novel autonomous material-handling robots for hospitals to complex enterprise information systems. She honed her project management experience while serving in technology consulting at Deloitte, overseeing test, planning, and execution for financial ERP implementations within the Department of Defense. In particular, Cara Jones has significant experience leading testing for complex deployments. System testing is an integrative role which requires an understanding of many aspects of a solution, including the domain-specific functionality to serve the users' needs. At the Department of Defense, she worked with the Joint Interoperability Test Command (JITC) to design the requirements traceability and test methodology for a 13-month deployment to replace the financial accounting and reporting system for an Agency with over $5B in annual transactions. This system included a dozen interfaces with other government applications and required inter-agency coordination, end-to-end testing, defect tracking, and technical resolution under pressure to meet the narrow end-of-fiscal-year cutover window. She also served as technical lead and managed an internationally staffed team on a retail eCommerce deployment. These projects faced many logistical challenges to coordinate, deploy by specific fixed-dates, and mitigate the risk of introducing major new technology to ongoing client operations. Ms. Jones delivers results whether working on project teams involving dozens of members or those involving few resources with minimal funding dollars. In her personal life, she served many years as a youth ice hockey coach and brings a spirit of team work and cohesion to her professional responsibilities. Cara Jones will serve as the main point of contact for the overall awarded contract with Allegheny County DHS and will deliver success in managing the team, engaging the client, and delivering the solution's milestones on-time.



Marinus COO Cara Jones in TV interview, October 2017

Emily Kennedy, B.S. Carnegie Mellon Heinz School of Public Policy

Marinus CEO Emily Kennedy at Lincoln Center for Women in the World, May 2018

Kennedy is CEO of Marinus Analytics, and directs deployment of advanced data mining and machine learning tools to local, state, and federal law enforcement for use on criminal cases, with an emphasis on human trafficking investigations. She routinely works alongside, advises, and trains stakeholders—such as attorneys general, prosecutors, law enforcement agents, and non-profit victim services organizations—on micro and macro approaches to combating and measuring human trafficking in the United States and abroad. Her work has been covered at the United Nations, Fast Company, NBC News, Vice, CBS News, and Scientific American. She has successfully raised funding from the National Science Foundation, the Bank of New York Mellon, and DARPA. She was recently selected as a Mother of Invention, sponsored by Toyota, to honor Marinus's extensive work in the social impact space. She is intrigued by technology solutions to social problems and is driven to innovate new ways to work with government organizations toward data-driven solutions. She graduated from Carnegie Mellon University in 2012.

Raymond Giorgi, M.S. Computer Science

Senior Software Architect Raymond Giorgi has worked as a software engineer and project manager for 10 years. During his time in the advertising industry, he was often responsible for all aspects of the software development life cycle (SDLC) in events with minimal development time and hard deadlines.

Marinus Senior Software Architect Ray Giorgi speaking about Marinus work in the IBM Watson XPRIZE – AI for Good Competition

His portfolio of rapid creations for clients were sometimes the result of extreme scenarios, such as only one week of development before the client event or developing working architecture and test plans for events in which only one day of testing in a foreign country was available. Raymond was also previously a Managing Data Scientist, where he oversaw projects that provided actionable information from big data resources. Raymond brings these experiences to Marinus Analytics and leads the delivery of new applications and modern tools for policing in the digital era. Raymond incorporates lean and agile methods to innovate and aid in the public sector. His work includes developing game-changing technology to generate leads for proactive policing, to help law enforcement assimilate analytics into their current workflow, and to use facial recognition for finding victims and returning them to safety. Raymond's role also involves coordinating closely with law enforcement officials in the design phase of projects. Prior to joining Marinus Analytics, Raymond served a variety of consulting firms, managing and developing end-

to-end solutions for clients in the public and private sectors. Raymond holds a Master's of Computer Science from the University of Pittsburgh.

Research Scientist, **Dr. Andreas "Olli" Olligschlaeger**, holds an M.Phil. and Ph.D. from Carnegie Mellon's Heinz School, an M.A. from the University of British Columbia, and a B.A. from Concordia University. He has over 30 years of experience in productizing and transitioning academic research to the private sector, specifically within law enforcement and public safety. With practical experience in law enforcement, academia, and private industry he has a unique and proven track record of introducing new technologies to law enforcement and integrating them with daily operations. For example, Dr. Olligschlaeger was instrumental in integrating the latest research in geographic information systems with crime analysis units around the country in the early 1990's; today, crime mapping is ubiquitous in law enforcement. His dissertation on using artificial neural network based space/time forecasting techniques and work as a crime and narcotics intelligence analyst within the Pittsburgh Police Department helped establish the field of predictive policing as we know it today.

In the early 2000's he was a systems scientist on the Informedia team at Carnegie Mellon's Robotics Institute and School of Computer Science. There, he worked with DARPA to develop a system that uses AI, natural language processing and speech recognition to extract meta data from broadcast news and allow for this data to be queried and visualized in multiple modalities, including GIS, link charts and timelines. In the mid and late 2000's he worked on integrating research in speech recognition and natural language processing with inmate phone systems, allowing investigators to automatically monitor every single inmate phone call made from a prison facility. The system has been deployed in numerous jails around the country and resulted in hundreds of criminal cases as well as several patents. In the early 2010's Dr. Olligschlaeger worked as a subject matter expert with the FBI and Raytheon to introduce and deploy academic research to N-DEx, which is a system that ingests massive amounts of incident and arrest data on a daily basis from police departments around the country. Specifically, he worked on entity extraction techniques, graph databases and social network analytics (including consulting on how measures of centrality can be applied to law enforcement), all of which are now fully integrated within N-DEx. Dr. Olligschlaeger has also worked on several projects with the Bureau of Alcohol, Tobacco, and Firearms' National Tracing Center, Axon's body worn video and digital evidence program, and served on the FBI's Future's Working Group, where he analyzed emerging technologies and academic research in order to assess their potential impact on policing and crime. Since 2016 Dr. Olligschlaeger has been a member of the Marinus Analytics team, where he has leveraged his expertise to design and develop software that uncovers human trafficking in vast amounts of data mined from the web.

Dr. Olligschlaeger has over 30 years of hands-on experience developing software, both within small startup settings as well as large organizations. He has worked on projects ranging from 100 thousand dollars involving one or two developers to projects costing over 100 million dollars involving teams of 75 developers and numerous project managers working in an agile environment. As such, he is comfortable working in any setting. His development skills include all modern programming languages, including java, C++ and Python, GIS, a wide variety of commercial and open source databases, including Oracle, SQL Server, MySQL and PostGresql, as well as geodatabases and numerous integrated development

environments, including Netbeans, Eclipse, and Microsoft Developer Studio. Dr. Olligschlaeger works on a daily basis with code control tools such as Git and SVN, as well as bug tracking tools such as JIRA.

Dr. Olligschlaeger is a member of the International Association of Crime Analysts, the International Association of Law Enforcement Intelligence Analysts, the Society of Police Futurists International (PFI), where he is the immediate past president, the PFI/FBI Futures Working Group, and serves on the advisory board of the High Tech Crime Consortium.

Thomas Wolber is a full-stack engineer whose career started as a graphic artist. Applying design experience to day-to-day work has lead Thomas to have a strong focus on user experience, accessibility, and ADA/Section 508 compliance over his 20 year career. With experience across various modern technology stacks, Thomas is not only skilled in his own work but serves well as a liaison between team members with diverse concentrations.

Julia Deeb-Swihart

Julia Deeb-Swihart is a Computer Science PhD student at Georgia Institute of Technology. She holds a B.S in Computer Science from Georgia Institute of Technology with a focus in Artificial Intelligence and Computational Theory. Her research focuses on applications of machine learning, network science, and user centered design for social good. Her thesis work is focused on building tools that assist with law enforcement investigations into human trafficking. As part of this work, she interviews law enforcement officers to understand their needs and challenges with their jobs and utilizes these insights to design systems that meet their needs. Most law enforcement officers have little or no computer science training, but have clear needs to work with big data to be successful in their work.

“My research is focused on building computing tools that empower individuals in their jobs for social good. Assisting human services aligns with my research visions.”

Dr. Artur Dubrawski, Advisor

Artur Dubrawski, founder and advisor, is a faculty member at the CMU School of Computer Science where he directs the Auton Lab, an applied machine learning research team of 30. He provides guidance on ongoing R&D at Marinus Analytics related to machine learning. Dr. Dubrawski has been researching intelligent systems (that work, are useful, and make economic sense) and ways to effectively build and deploy them for 20+ years. He leads teams at CMU investigating new machine learning algorithms and data structures to facilitate probabilistic modeling, predictive analysis, interactive exploration, and understanding of data.

5. Provide a detailed budget that clearly supports your Solution and implementation plan. Include a narrative that explains and justifies each budget item and how amounts were calculated. You may provide the budget and budget narrative as an attachment.
The budget and narrative is provided as a combined attachment.

6. Describe your understanding of the challenges inherent in implementing your Solution and how you plan to address those challenges.

Click or tap here to enter text.

Challenges to implementing the CASET solution, or any new technology, may fall into the category of scientific, engineering, or an organizational challenge. In this section, we raise some of these challenges and our respective mitigation approach.

Allegheny County DHS is an early adopter of technology with ambitions to leverage today's artificial intelligence computing to extract actionable intelligence from within its unstructured case notes' data. The potential of AI may be limited by the characteristics and qualities of the data. The results of the data mining research track will be unknown until further into the overall project. One mitigation strategy is the feature agnostic design of the caseworker view of the proposed solution. The overall CASET architecture and visualizations are amenable to react to research phase to appropriately "hook" into the discovered features and attributes on the front end for all views pertaining to caseworkers, along with supervisors and service providers. Marinus Analytics believe there is a base-value in the aggregation of notes across systems and features enabled via software engineering, with added-value stemming from what is produced via the research effort.

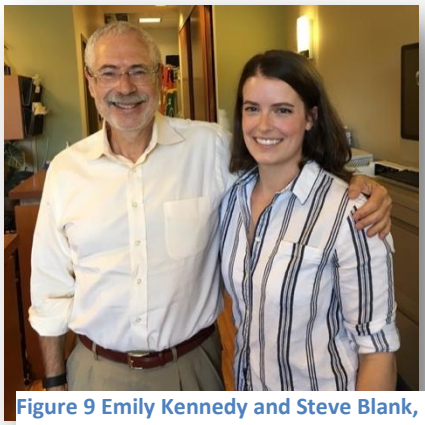


Figure 9 Emily Kennedy and Steve Blank, recognized for developing the Customer Development Methodology

From an engineering perspective, challenges may occur such as data consistency between CASET and the source-of-truth. If the original record gets edited after-inception, the CASET system will need the ability to synchronize the older recorder. This should be manageable assuming the parent database maintains time stamps or awareness of record revisions. If the case note is in a document form, the latest date will reveal the time of edit.

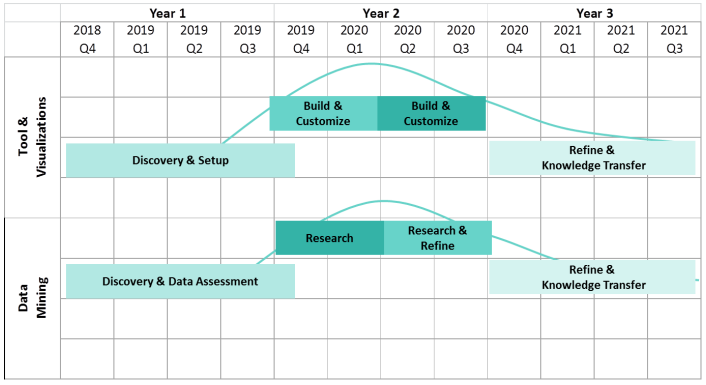
Additionally, we may encounter a challenge in moving the information between the legacy system and CASET. It may be that the older legacy systems are not able to be modified to push data into our system, but, in this case, CASET will be made to pull the data at regular periods from the legacy data sources. Alternatively, security concerns may not allow data to be queried from CASET, but, in this case, the legacy system can be modified to push data into a real-time system from CASET. Finally, there may be obstacles in convincing existing vendors to push data into our system, but, in this case, we can again pull data directly from the existing vendors into our system and keep the results open so that CASET can be further extended by either Marinus Analytics or another vendor.

Operationally, the goal is to achieve user buy-in and input into the developed system. Later, it is important to attain adoption through good-usage of the technology over time. Our approach to engaging and winning over the target users is by the manner in which we match the solution to their expressed pain-points. This approach is known Customer Development Methodology. A major promoter

of this methodology is the National Science Foundation (NSF) who funds academic institutions and small businesses to advance fundamental science in ways which will eventually benefit society. As a part of the NSF I-Corps Berkeley Cohort in 2014, Marinus Executives Emily Kennedy and Cara Jones engaged in the intensive training program which created a catalyst to launch the product Traffic Jam despite known obstacles for innovating in the law enforcement sector. We learned in-depth knowledge and practical application from experts at the NSF and University of California Berkeley; the training was specifically focused on the Steve Blank customer discovery process, which we still currently use for new product development. The core of this process involves "getting out of the building" to test assumptions and hypotheses about user needs, and validate or invalidate them in the field. This has been our guiding methodology to insure that any software solutions we build meet real needs of users on the ground.

7. Provide a timeline for the design and development of your Solution.

The following graph illustrates the two major tracks of this project which will interact harmoniously to deliver the CASET solution to the client. The arch in the illustration emphasizes the intensity of the software engineering and research to fully mature CASET. During the first year, a discovery process will be undertaken to understand the data paradigm and test hypotheses through client specialist interviews. Year 2 will be heavily focused on maturing the CASET tool and conducting research for the extraction of actionable insights from the unstructured data. In the last year of the project, we envision an operational CASET tool and our support will be primarily focused on knowledge transfer, enhancing the robustness of the overall system, remediating any uncovered issue, and/or servicing any feature requests. During this tail end of the implementation, staffing will be reduced to the senior experts of our team to successfully finalize and transfer the solution. Following the end of the project, Marinus Analytics is open to operating the CASET solution for Allegheny County DHS in a SaaS subscription manner, similar to our operation of Traffic Jam. From the start of the project, we will follow a lean development and MVP maturation style which will allow the client specialists to begin exploring the tool while it is still being developed. This approach allows for rich feedback based on “hands on” review by subset of users before the tool is officially rolled out. We anticipate full rollout could be achieved at the beginning of year 3 with the remainder of the project time being spent on refinement through operational feedback and then knowledge transfer based on subsequent maintenance strategy.



8. Describe your organization's plan to collaborate with DHS during development, implementation, knowledge transfer and training about how to use and maintain the Solution.

Critical to the successful implementation of any data mining solution is continued interaction with the customer. We will conduct a thorough requirements review during which specific components of the data mining solution will be identified, as well as the specific major and minor categories of events, issues and needs.

Throughout the research and development and implementation stages, Marinus will provide initial results of data mining accuracy, as well as incremental results for feedback.

During the tool development stage, Marinus Analytics will seek input from DHS specialists and champions on the requirements and design of the system. Marinus Analytics will conduct user discovery interviews to test hypotheses of the functionality of CASET prior to executing the research and engineering to gain confidence in how the system will best support the underserved pain-points of the organization. Marinus Analytics will support DHS in understanding and mitigating hurdles for adoption and to determine ways to align incentives to promote usage of the proposed solution. Marinus Analytics has leveraged this methodology since our participation in the 2014 NSF National I-Corps cohort. Hypothesis testing of project assumptions is a lean approach and is designed to allow DHS to uncover any unforeseen issue before implementation, when the design is most flexible to changes and adjustments.

These interactions will result in the following products:

- User Interface Wire Frame diagrams
- Formal use cases
- Entrance and exit criteria for project milestones

The proposed timeline includes "first look" opportunities to keep DHS informed of the solution long before the formal final release of the tool. This will ensure feature development meets the needs of DHS personnel and contracting agencies. In addition, project champions will be empowered by this additional lead time to proactively prepare for policies, adoption, incentives, and awareness building.

Throughout the course of the project, Marinus will provide bi-weekly progress summaries and status updates. As Marinus is a Pittsburgh-based company, we are available for periodic onsite meetings as part of these touchpoints.

As the project nears end of year 2, we will provide test results and status of any outstanding issues being mitigated. We will create training materials and conduct train-the-trainer sessions to transfer knowledge on the intended usage and best practices for maximizing utility.

Marinus Analytics engineers will either outline a strategy for ongoing SaaS offering of CASET or support the DHS IT team in operating the tool through the delivery of manuals for maintaining the system from an IT perspective. We will also make our team available for in-person walk through of operating materials and transition of responsibilities in the last half of year 3.

9. Describe how your organization will evaluate the success of your Solution. Provide an example of how you measured the impact and success of a similar project in the past.

From the time of participating in the NSF I-Corps program (mentioned in answer to Question #6) and over the years of offering Traffic Jam, Marinus as a company has honed its practice of the user discovery methodology. The nature of serving law enforcement and adapting to changing criminal behavior has required us to routinely check-in with our users to understand their needs and if/how our SaaS offering is fitting with their pain-points. This process is often used in the early stage of formulating an entrepreneurial idea but we believe it is applicable to maintaining relevance, competitive differentiation, and innovation over the life of a software application. With commercial SaaS, it is critical to deliver success through the relationship with end-users or otherwise the attrition from usage will directly result in churn of subscribers hurting the viability of the overall tool. Figure 9 below highlights the main elements of the user discovery process which we recommend as an approach to evaluate the success of the CASET solution. The user discovery process is a cycle that begins with interviewing the users and testing your assumptions around product-to-pain-point fit. This initial element is critical formulating the fundamental goals of the solution. Along the maturation of the solution, user feedback should be solicited. We directly train users (over 1500 to-date). This allows us to observe the reactions and feedback to the latest revisions and features released into the tool. We monitor who logins into the tool which gives us a basic appreciate for the usage. If usage is high, it indicates the utility and relevance of the tool. This is one quantitative success measurement. Finally, we survey our users for statistics on impact. Ultimately, this is the most effective way to know if the tool is improving the users day-to-day workflow and if the AI in the tool translates into operational effectiveness.

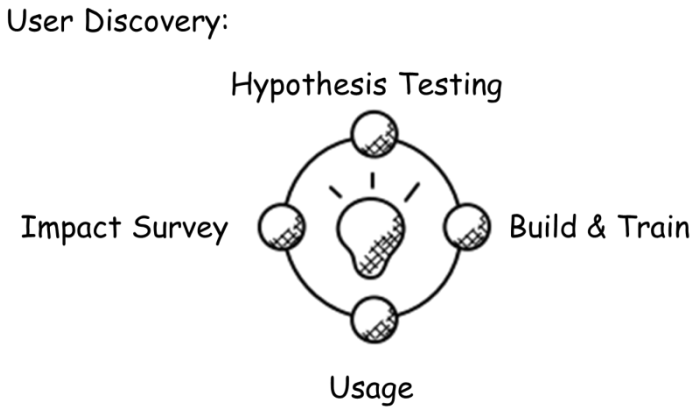


Figure 10 User Discovery to track impact and success

10. Describe why you want to serve human services clients, your experience in adapting technology to serve human services clients and your plan to adapt your Solution to meet client needs for this track.
Marinus Analytics was created in 2014 to help populations who could not ask for help, namely victims of human trafficking. Making a positive social impact is part of our company DNA. We came out of research at Carnegie Mellon University, after speaking with hundreds of law enforcement agents about the difficulties of finding these victims and prosecuting their exploiters. We saw a huge need for detectives to be able to harness massive amounts of online data to inform actionable insights for investigations, and we developed research—and ultimately created software tools—to address this issue.

Through our work thus far, we have taken advanced data mining and machine learning technologies and productized tools to deploy them to the average law enforcement or government user. We have insight into the needs, desires, and pain points of local, state, and federal government agencies. Because human trafficking is an evolving problem, we had to make solutions that were adaptable to the changing landscape.

We have seen the impact software can make on the workflow of government workers, and our software for human trafficking empowered the rescue of hundreds of victims. We are passionate about developing advanced technology to support public agencies in doing more with their existing resources.

Marinus Analytics and its Traffic Jam technology are funded in-part by NSF through the Phase II Small Business Innovation Research (SBIR) program. As the proposed CASET solution is an augmentation of the Traffic Jam solution, supplement funding is available to support this project. In effect, the supplement grant, called a Phase IIB, will match the contract award and provide up to $250,000. We are highly motivated to pursue this work with Allegheny County DHS as it is a natural extension of our focus on Traffic Jam. The delivered solution will not only benefit Allegheny County DHS but will benefit human service organizations across the country because Marinus Analytics would provide the resulting technology as a SaaS tool offering. Therefore, it is appropriate to support Allegheny County by applying these funds to an awarded contract to discount the cost of R&D for the CASET solution.

Marinus Analytics Budget Narrative – CASET solution

In this budget narrative, we provide an overview of the team which will perform the necessary work of the tools & visualizations and the data mining tracks to successfully delivery CASET solution. Our proposed CASET solution is an augmentation of the Traffic Jam system. To meet the needs of Allegheny County, it will require additional research & development to work in the data paradigm of the unstructured case notes and to best function for the case workers, supervisors, and service providers. The budget is detailed by track and by year. We will distribute the workload across the team as the tasks for delivering the solution require. The majority of effort will take place in years 1 & 2. In year 3, the blended rates increase, because we will reduce the staffing to include only our most senior experts on the team to finalize the project, address any unresolved issues, and advise on knowledge transfer to the client. This budget accounts for the following Marinus Analytics team members for each track.

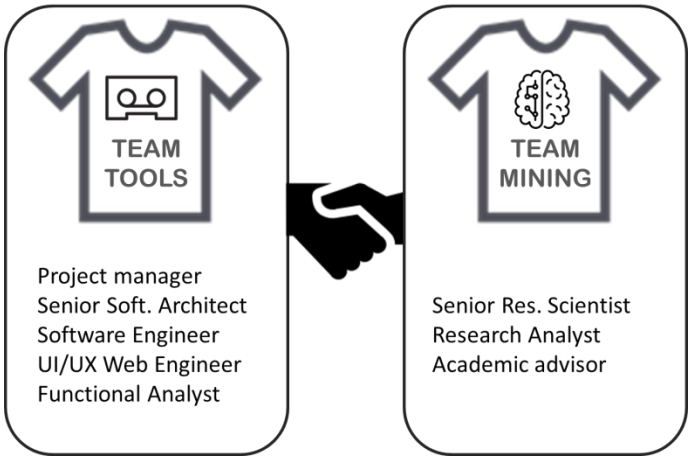


Figure 1 Team composition by track

Marinus Analytics is currently funded by the National Science Foundation as a Phase II Small Business Innovation Research (SBIR) recipient. This program provides a supplement grant, called a Phase IIB, which will match Allegheny County DHS funds by 50%, up to $500,000. We would like to apply these funds to an awarded contract through Allegheny County to discount the cost of R&D for the CASET solution. Marinus Analytics gains from a contract because the final CASET tool will be available for future SaaS offering to other county services around the country.

The following budget is captures two tracks the datamining research and tools and visualizations development. The NSF SBIR PhaseIIB grant is listed as the discount item of the subtotal amount. If

awarded, the manner in which the $250,000 from the National Science Foundation is applied to the project is subject to Allegheny County DHS preference and per the disbursement process instructed by our NSF program director.

Table 1 Project budget by track and year, including grant to offset R&D

	Year	1	2	3	Total
Data Mining	**Labor**	$ 199,680	$ 179,985	$ 47,450	$ 427,115
Tools & Visualizations	**Labor**	213,109	223,834	61,191	498,134
	Materials (Computing infrastructure, services, software, etc.)	20,000	25,000	25,000	70,000
	Sub Total	**$ 432,789**	**$ 428,819**	**$ 133,641**	**$ 995,249**
	NSF SBIR PhaseIIB Grant		$250,000		
	Total	**$ 432,789**	**$ 178,819**	**$ 133,641**	**$ 745,249**

		Year 1	Year 2	Year 3
Data Mining	Day Rate	$512	553.8	730
	Hourly	$68	$74	$97
Tools & Visualizations	Day Rate	497	522	588
	Hourly	$66	$70	$78
Combined	Day Rate	$504	$536	$643
	Hourly	$67	$71	$86

Table 2 Blended day and hourly rates per track and year