Paper SAS6560-2016

Pedal-to-the-Metal Analytics with SAS® Studio, SAS® Visual Analytics, SAS® Visual Statistics, and SAS® Contextual Analysis

Manuel Figallo-Monge, SAS Institute Inc.

ABSTRACT

What is it about SAS® Analytics that makes it so *analytic*? SAS provides a suite of web technologies and applications that aid in the discovery and communication of meaningful data patterns to gain insights. One way to accelerate these insights is through "pedal-to-the-metal analytics" or web-based rapid prototypes; this is because SAS web technologies and applications are closely integrated with one another and, further, can integrate with other systems (such as API-based data sources) through the use of SAS software components called SAS macros that are introduced in this paper. Examples of SAS web applications include SAS® Studio, SAS® Visual Analytics, and SAS® Visual Statistics. The integration of web system technologies and applications produces an end-to-end solution invaluable toward prototyping solutions that can be learned from, particularly as organizations begin to capture requirements for an analytics project and understand the risks associated with that project during the elaboration phase.

PREPARATIONS FOR THE ANALYTICS JOURNEY

Every analytics project deals with requirements and risks. Using SAS web technologies and applications helps mitigate against risks while providing a clearer sense of requirements. Although risks are pervasive, so are the rewards of an analytics project; these rewards and benefits are generally the attainment of answers--or clarity of a hypothesis or question--critical to business or mission success. The search for answers to ensure organizational effectiveness generally means heeding a call to adventure fraught with twists and turns and surprises.

Continuing with the theme of a journey, organizations will need three things:

1. Fuel. Simply put, this is data, but, more importantly, clean and usable data.
2. A vehicle. For the purpose of this paper, it is web applications that can facilitate and propel the journey forward with valuable insights on the data.
3. A map. Most analytics-ready projects refer to this as a methodology; it is an iterative and incremental analytic systems development process framework. In this paper's case, it is the analytics lifecycle or roadmap described later.

Figure 1. As with any journey, three things are required for "pedal-to-the-metal analytics": fuel, a vehicle, and a map.

Much like any journey, there are obstacles and bumps on the road that can thwart any analytics project and even cause its end or dissolution. They can be referred to as the four horsemen (of the analytics apocalypse):

1. Chaotic business processes and objectives resulting in requirements risks.
2. Immature capabilities leading to skills risks.
3. Poor technical integration of systems and data or technology risks.
4. Dysfunctional organizations that produce political (organizational) risks.

WHAT IS PEDAL-TO-THE-METAL ANALYTICS?

This paper introduces “pedal-to-the-metal analytics” to accelerate the development of an analytics project, particularly requirements gathering, as well as to quickly identify the risks associated with it.

It has three characteristics:

1. Web software technologies built using SAS, or reusable SAS software components that are modular and flexible enough to be applied to a variety of cases. They are designed to be integrated with one another to run as SAS® Stored Processes or SAS® Studio Projects (Process Flows) on the web.
2. SAS web applications including SAS® Visual Analytics, SAS® Visual Statistics for structured data analytics, as well as SAS® Contextual Analysis for text analytics.
3. Fast iterations through the analytics lifecycle, which serves as an analytics roadmap.

Figure 2. The analytics lifecycle with SAS, which serves as an analytics roadmap, is used for the analytics journey. The SAS® LASR™ Analytic Server and tables are explained later, and source code is SAS code that is produced from creating an analytics model. It is a mathematical formula captured in SAS code that, when applied to a data set, will enrich the data set with additional information.

“Pedal-to-the-metal analytics” is, in other words, rapid prototyping using web technologies and applications. Prototyping can be immensely useful and inform the requirements or design phase as well as the implementation phase of a project especially if itfollows standard system methodologies (for example, Unified Process, Waterfall, Agile, and so on).

WHAT IS THE WEB?

The web is a systems architecture that is built using computer or machine languages for processing data (for example, Javascript, HTML5), publishing content (for example, HTML), and enabling communication between and within systems through the use of HTTP or the Hyper Text Transfer protocol. Traditionally,

the technologies that reside in the web are multi-tier; that is, they consist of a data tier, processing or application tier, and presentation tier.

At its simplest, HTTP is a request-response communication protocol where an application client (for example, browser) makes a request to a server and a response is provided. From any browser, for instance, the address http://www.sas.com will issue a request to the sas.com server and be given a response or a web page in return with SAS.COM front page content for end users.

An extension of HTTP are APIs or Application Programming Interfaces. As the name implies, APIs consist of an interface that will accept inputs to generate one or more outputs.

Figure 3. In clockwise order, an API is analogous to a power outlet (since it is an interface to a larger system power grid), inputs are signified by a plug, and a light generated from the bulb is the output.

Analogous to this is a wall outlet, which is an interface to a larger power grid. The input can be a plug attached to an electric vehicle and the output that results from connecting a plug to a wall outlet is volts used to power a car. Another example is demonstrated in Figure 3 above.

WHY THE WEB?

The web, specifically API-based data, fuels “pedal-to-the-metal analytics” web applications. How? The web or HTTP provides a unified interface to a series of subsystems, such as SAS web applications and, most importantly, API-based data sources (for example, the Health Indicators Warehouse or HIW described later). In other words, the web and APIs can provide data feeds to SAS web applications in a seamless, automated fashion much like an outlet provides electricity.

HTTP not only integrates disparate technologies, applications, and capabilities, it also makes them extensible. SAS stored processes, for example, can be used to extend the collection of visualization objects in SAS Visual Analytics; this will be discussed in more detail later.

Figure 4. HTTP not only integrates disparate applications (for example, Visual Statistics, Visual Analytics, and SAS Studio) but makes them extensible. This is because HTTP is a "spanning layer" among web technologies and applications. These technologies include API-based data sources, such as Health and Human Services (HHS)'s Health Indicators Warehouse (HIW). Note that SAS Visual Analytics can access an API-based data source such as the HIW by running SAS Stored Processes, and SAS Studio can access the HIW with SAS Studio custom tasks. The SAS Web Infrastructure Platform in the figure above will be discussed later—it is the platform on which all SAS web applications rest.

In addition, SAS web applications are easy to use and many provide drag-and-drop user interfaces that make analytics approachable and accessible, meaning it is something that end users can readily learn how to do. Relative to other architectures, web systems are also easy to maintain and scale because IT resources and management are centralized in a single machine or even a distributed system.

Finally, the web (and in particular SAS web technologies and applications) delivers insights that can guide organizations in their analytics journey toward data-driven decision making. This is possible through powerful visualizations that can be displayed and shared through a web browser. Insights are also possible by way of analytic models generated by SAS web applications, such as SAS Visual Statistics.

THE SAS WEB INFRASTRUCTURE PLATFORM AND MIDDLE TIER

SAS web technologies and applications reside in the middle tier, enabling users to access API-based data using SAS code.

The collection of services and applications that provide a common infrastructure and integration for SAS web applications is also referred to as the SAS Web Infrastructure Platform.[1] A platform is a base set of technologies on which other technologies rest.

The SAS Web Infrastructure Platform builds on previous SAS architectures, and as a result of using a web application and server, it is possible to run SAS applications through a browser, thus easing the IT maintenance burden. In addition, business users benefit from easy-to-use, drag-and-drop interfaces. More on that later, but suffice it to say that the user experience through a web browser is a positive one; it is intuitive as well as makes analytics easy to learn, remember, and use.

Two additional components in the SAS Web Infrastructure Platform make it compelling: the SAS Stored Process Server and the SAS LASR Analytic Server.

The SAS Stored Process Server executes SAS programs as required by requesting applications (for example, SAS Visual Analytics) and stores them on a server. These SAS programs can also access any

[1] To better understand this, it is worth examining how SAS typically evolves in organizations. Many organizations who have adopted SAS begin with Foundation (Base) SAS and access engines on either a desktop or mainframe. Foundation SAS is a very rich 4GL that is capable of many things from ETL to reporting; however, users running it on their desktop soon realize capacity constraints; processing is slow and in this era of big data and large data sets, it can be exceedingly slow, even glacial.
To address these performance constraints, projects are better off separating data processing from presentation and utilizing a client/server architecture; this gives users the benefits of a larger machine or server to process large amounts of data while retaining and using the client for human computing interaction on their desktops. Communication between the two can happen through a multitude of protocols (for example, IOM or Integrated Object Model), and in this architecture, a workspace server can be introduced.
Further, in a client/server architecture, SAS Workspace Servers interact with SAS processes by creating a server process for each client connection. The workspace server process is owned by the client user who made the server request. Each workspace server process enables client programs to access SAS libraries, perform tasks by using the SAS language, and retrieve the results. SAS Workspace Servers can also be pooled; that is, a set of server processes can be reused to avoid the processing time that is associated with starting a new process for each connection.
The SAS® Metadata Server can also exist in client/server architectures. It is a centralized resource for storing, managing, and delivering metadata for SAS applications across the enterprise. The SAS Metadata Server enables centralized control so that all users access consistent and accurate data. Access to data and metadata is secured through a metadata-based authorization layer, which supplements protections from the host environment and other systems (for example, authentication).
Recent releases of SAS Metadata Server provide redundancy and high availability that result from clustering. Clustering ensures that the server will continue to operate if a server host machine fails. Also, a load-balancing process distributes work among the clustered nodes. And if a node ceases to operate, the server continues to be available using the remaining nodes.
Maintaining a client/server architecture can, however, be onerous and anathema to IT. Imagine updating hundreds of clients with a software patch. For some organizations, this is simply not feasible or cost effective, hence, the development of web infrastructure and web architectures.

SAS data source or external file and create new data sets, files, or other data targets supported by SAS. Think of a SAS stored process as SAS code that runs on HTTP.[2]

The latest component to be added to SAS Web Infrastructure Platform is SAS LASR Analytic Server.

SAS LASR Analytic Server provides a secure, multi-user environment for concurrent access to data that is loaded into memory. The server can take advantage of a distributed computing environment by distributing data and the workload among multiple machines to perform massively parallel processing. The server can also be deployed on a single machine where the workload and data volumes do not demand a distributed computing environment.[3]

The SAS LASR Analytic Server processes client requests at extraordinarily high speeds due to the combination of hardware and software that is designed for rapid access to tables in memory.

This also brings into sharp relief the benefits of SAS Web Infrastructure Platform:

1. Integration and extensibility.
    - The ability to extract data from a variety of operational data sources on multiple platforms to integrate them.
    - The ability to store large volumes of data efficiently and in a variety of formats while also ensuring scalability.
    - The capability to centrally control the accuracy and consistency of enterprise data through a browser.
2. Ease-of-use.
    - The SAS Web Infrastructure Platform gives business users at all levels the ability to explore data from the warehouse in a web browser, perform simple query and reporting functions, and view up-to-date results of complex analyses. In other words, the enablement of “accessible analytics.”
3. Insights.
    - Using high-end analytic techniques through SAS web applications that run on the Platform provide capabilities such as predictive and descriptive modeling, forecasting, optimization, simulation, and experimental design.

[2] Each SAS Stored Process Server process handles multiple users, and by default each server uses multiple server processes or instances. A load-balancing algorithm distributes client requests between the server processes. Unlike the stored process server, SAS Workspace Servers interact with SAS by creating a server process for each client connection. The workspace server process is owned by the client user who made the server request. Each workspace server process enables client programs to access SAS libraries, perform tasks by using the SAS language, and retrieve the results.

[3] SAS LASR Analytic Server handles both big data and smaller sets of data, and it is designed with a high-performance, multi-threaded, analytic code. By loading tables into memory for analytic processing, the server enables business analysts to explore data and discover relationships in data “at the speed of RAM.”

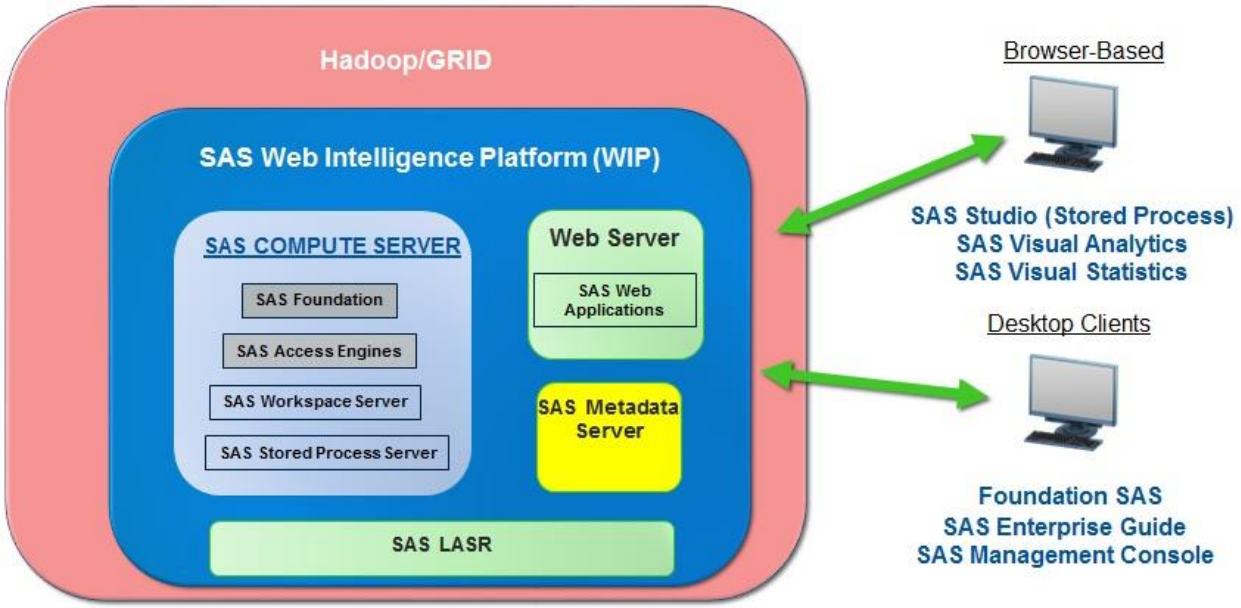


Figure 5. The evolution of the SAS Platform that ends with the SAS Web Infrastructure Platform, which makes browser-based analytics possible (upper right). SAS LASR Analytic Server and SAS Stored Process Server were discussed earlier and are core components of SAS Web Infrastructure Platform. In some cases, SAS users begin with SAS® Foundation products (gray box), move to client/server systems (gray, light blue, yellow boxes sans the SAS Stored Process Server), and eventually adopt a web system (all boxes with an optional distributed environment or the pink box).

This backdrop serves as valuable context for "pedal-to-the-metal analytics." For rapid web-based prototyping, SAS provides a robust architecture and infrastructure, centrally managed through metadata and shared by all web applications. Finally, it is worth noting that the SAS Web Infrastructure Platform produces a hub-and-spoke architecture, with metadata as the hub and several web applications as spokes. These "spokes" include SAS Visual Analytics, SAS Visual Statistics, and SAS Contextual Analysis. What's more, this paper introduces a collection of SAS macros (that is, library of software components) called the AVAF (Applied Visual Analytics Framework), which can extend the capabilities of the aforementioned web applications. AVAF components can be run in SAS Studio or as SAS Stored Processes.

SAS WEB TECHNOLOGIES: THE AVAF (APPLIED VISUAL ANALYTICS FRAMEWORK) AND SAS STUDIO

SAS Studio is a web-based development environment. With SAS Studio, end users can access data files, libraries, and existing programs, and even write new programs.

In a sense, SAS Studio provides the "fuel" or data for "pedal-to-the-metal analytics" web applications since it can be used for data access and preparation.

As mentioned, the AVAF is a collection or library of "reusable SAS software components or macro functions that are modular and flexible enough to be applied to a variety of use cases." It is introduced in this paper to enable access of API-based data sources.

A macro function processes one or more input arguments and produces a result. It consists of inputs, an interface or parameter list, and output. Code logic is, for the most part, encapsulated.

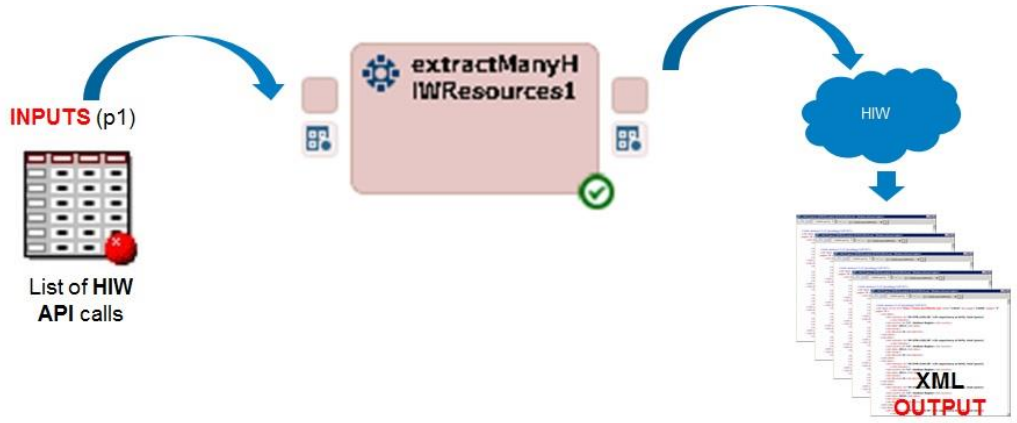


Figure 6. An example AVAF component (SAS Studio custom task with a SAS macro embedded inside) used to extract many HIW indicators. It consists of inputs, an interface, and output. The inputs are RESTful API calls to the HIW, the interface requires a single dataset, and the output consists of several XML files with HIW data retrieved from the cloud

These macros functions or software components can be integrated with one another to form larger applications. It is critical to note that they are reusable. Therefore, they rely heavily on design patterns or repeatable design solutions to commonly occurring system engineering problems to facilitate reuse.

They also perform primarily three tasks:

- Data access (through the use of APIs, for example), including ETL (Extract, Transform, and Load) functions
- Data enrichment
- Data aggregation

This is by no means an exhaustive list of tasks performed as some of the AVAF components can be used for value-added configuration or customizations —that is, augmented capabilities of the SAS Web Infrastructure such as county-level geographic analysis in SAS Visual Analytics (see Appendix). All AVAF components in the use case below can be found at http://tinyurl.com/SGF2016.

Included in this website are, in addition to the SAS macro code, SAS custom tasks as well as the example projects discussed later on.

Here is a summary description of AVAF:

- Applied. Composed of software components based on real-life use case and end-users' experiences and requirements.
- For SAS Visual Analytics to be used in the SAS Web Infrastructure Platform infrastructure or architecture on which SAS Visual Analytics resides.
- A framework. It is a true framework with non-modifiable framework code (really, a gray box), extensibility, and inversion of control --that is, the flow of control is received from the library rather than the user himself.

In sum, the composite applications for data preparation that are created by the AVAF can be run as SAS batch programs, stored processes, or SAS custom tasks in a SAS Studio Process Flow.

SAS WEB APPLICATIONS: SAS VISUAL ANALYTICS, SAS VISUAL STATISTICS, AND SAS CONTEXTUAL ANALYSIS

Until this point, the value of the SAS Web Infrastructure Platform has been explained, and the AVAF as well as SAS Studio have been introduced.

For the purposes of this paper, SAS Visual Analytics, SAS Visual Statistics, and SAS Contextual Analysis also complement the capabilities of SAS Web Infrastructure Platform. Through this collection of web applications, which serve as a vehicle toward insights on data in “pedal-to-the-medal analytics,” projects can make the full breadth of SAS analytics available to users throughout their organization.

The common infrastructure and integration features to be used by SAS web applications provide the following benefits:

- Greater consistency in users' interactions with web applications
- Integration among web applications as a result of the ability to share common resources, particularly as a result of the SAS Web Infrastructure Platform

To start, SAS Visual Analytics is an easy-to-use, web-based product that uses SAS high-performance analytic technologies and complements the capabilities of SAS Web Infrastructure Platform.

Once data has been accessed by SAS Studio and the AVAF, for example, SAS Visual Analytics can be used to reap these benefits:

- Visually explore data, based on any variety of measures, at amazingly fast speeds
- Quickly create reports or dashboards using standard tables, graphs, and gauges
- Share insights with anyone, anywhere, via the web or a mobile device

SAS Visual Statistics is an add-on to SAS Visual Analytics that enables projects to do the following:

- Develop or create analytics models. SAS Visual Statistics also makes analytics accessible because it enables projects to rapidly create powerful statistical models (such as a linear regression model, a logistic regression model, an advanced decision tree, and k-means clustering) in an easy-to-use, web-based interface.
- Test the models.
- Compare models using the in-memory capabilities of SAS LASR Analytic Server.

While SAS Visual Analytics enables users to investigate and visualize data sources to uncover relevant patterns in the exploration phase of the analytics journey, SAS Visual Statistics extends these capabilities by enabling users to create, test, and compare models based on the patterns discovered. SAS Visual Statistics can also be used to export the score code that’s created behind the scenes, before or after performing model comparison, for use with other SAS products and to put the model into production.

Finally, SAS Contextual Analysis allows users to add structure to unstructured data. It is a web-based text analytics application that provides a comprehensive solution to the challenge of identifying and categorizing key textual data. Using this application, end users can build text models (based on training documents) that automatically analyze and categorize a set of documents.

With these web-based applications, fast iterations of the analytics life cycle can be achieved. Each web application, furthermore, is optimally suitable for a specific phase of the analytics lifecycle as demonstrated next.

THE ANALYTICS LIFECYCLE

On one’s analytic journey, a map is critical. The analytics lifecycle is such a map, or, better, a roadmap that automates, operationalizes, and productionalizes the creation, testing, and deployment of analytic models inside an organization. Its purpose is to help analytics projects focus their stakeholders’ time (and not stray into irrelevant pursuits) and provide them with shape and structure for a successful analytics project. The final destination is “insights” that can be used for further elaboration of the project.

As mentioned earlier, it consists of four stages or phases. In order, they are:

- Data access and preparation, specifically API-based data access.
- Data exploration.

- Analytics modeling building.
- Scoring data (with the analytics model) and, further, reporting to evaluate results

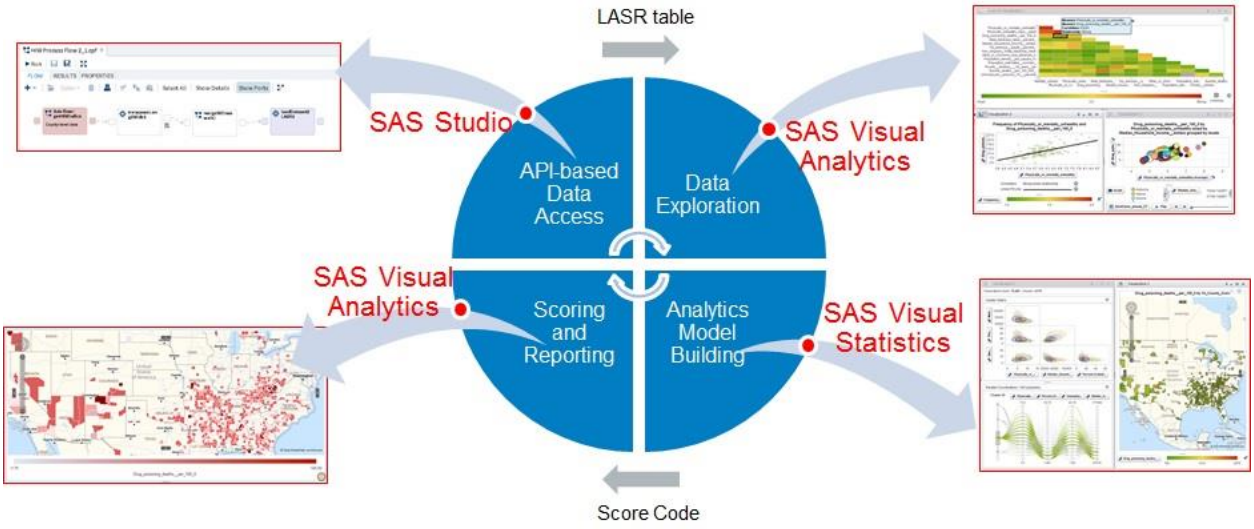


Figure 7. The four phases of the analytics lifecycle or roadmap with their corresponding web technology or application. Note that data is moved into SAS LASR Server before data explorations are achieved, and score code is applied to data in order to evaluate results. Also note what SAS web technology or application can be used for each phase or step in the journey. SAS Contextual Analysis is discussed later.

Each phase ensures that tasks are thorough and complete and serves as a pre-condition for the next. The lifecycle also ensures that better transitions occur to members and stakeholders in a cross-functional analytic team, particularly since roles and responsibilities are defined for each phase.

These are the four roles that are present throughout the lifecycle or roadmap:

- Business Manager (who has domain expertise, makes decisions, and evaluates processes and ROI)
- Data Scientist (who explores the data, identifies visualizations, and creates reports)
- Analyst/Data Miner (who performs exploratory analysis and defines the appropriate analytic models to be used)
- IT Systems Manager (who prepares data, validates models, and deploys them)

Data preparation and access begin with a hypothesis or question that requires data in order to be answered; it also involves the IT Systems Manager and Business Manager. In “pedal-to-the-metal analytics” this will typically require integration with API-based data sources. Ultimately, this first phase allows analytic teams to “conquer the blank page” to get an analytics project started.

The next phase requires exploration of the data in order to derive initial insights that can be used for a modeling exercise. The data scientist has at his or her disposal techniques such as correlations and descriptive statistics to understand the relationship between the different variables in a data set. He or she can also determine whether there is enough data to build a quality model and begin to think about techniques and methods that could be used to actually build a model. A model is nothing more than a formula used for a specific analytic task (such as prediction or clustering) that is applied to data for the purpose of gaining insights into the data.

Creating, testing, and comparing models is evident in the next phase, modeling, and it involves a good understanding of statistical methods that are typically within the ken of the analyst or data miner. In the analytics journey described in this paper, the model can be used on structured numeric data or unstructured text data. Data miners must be given an environment where they are free to store, transform, enrich, integrate, interrogate, and visualize the data in search of valuable relationships and

insights buried across the different data sources. It must be an environment where data models can be refined rapidly in minutes or hours, not days or weeks.

The upshot is score code that is used by the IT Systems Manager to deploy the model on data in order to evaluate results. The evaluation of results also involves the Business Manager, ultimately responsible for publishing results from the analytics lifecycle in order to make actionable data-driven decisions.

As projects go through several iterations of this lifecycle, requirements and risks become clearer. There are, in particular, four risks that must be identified and understood.

RISKS

As with any journey, there are bumps on the road, and even more significantly, dragons and sirens and shape-shifters and tricksters (to borrow from mythology and folklore) that can limit an analytics project's progress forward. Much of the territory on the analytics journey can be terra incognita for stakeholders, and four risks, as mentioned earlier, can be especially stifling:

- Requirements Risks. If requirements and their priorities are not clearly understood, there is a real danger that the wrong system or model will be built. In addition to business functional requirements, technical or non-functional requirements need to be clearly specified.
- Skills Risks. Without the right staff and expertise, an analytics project is doomed to fail.
- Technology Risks. If the technology does not work or does not integrate, expect little return on investment. Data must also be accessible, understood, and of sufficient quality to perform analytics.
- Political Risks. Is there sufficient support and sponsorship for the analytics projects from organizational leadership? Are key stakeholders throughout an organization invested in the analytics project?

One of the purposes of "pedal-to-the-metal analytics" is to take a project from the inception phase to its elaboration. A project's inception is where business rationale for the project and its scope are established, and elaboration is where more details about requirements and risks are gathered.

"Pedal-to-the-metal analytics" is ultimately about creating a prototype using web applications and technologies in order to better understand risks and requirements. Through iterations of the analytics lifecycle, specific use cases will emerge that can be used to solve a business problem. They can include use cases that deal with technology risks or even requirements risks, since the use cases form the basis of communication between sponsors and developers of the analytics project.

The web-based applications in this paper make analytics accessible, mitigating against skills risks through easy-to-use, drag-and-drop user interfaces that can be extended through the use of very flexible 4GL code that includes SAS macros. What's more, the SAS community (composed of user groups and a multitude of support resources) provide the best way of acquiring analytics skills—that is, mentoring. Mentoring requires experienced analytics professionals to impart knowledge and skills.

Because web-based technologies are well integrated through HTTP, which serves as the "spanning layer," they serve as a foundation for a discussion about what specific technologies will best address analytics requirements. These discussions can contribute to additional ones regarding what technologies would be required at a lower layer (for example, metadata) and even serve as the basis for more modern technologies with finer-level software components that communicate via an HTTP layer (for example, the AVAF or a micro-services architecture).

And, since SAS web technologies produce very interactive reports and visualizations to tell a data story[4], they also serve as a means to influence higher-ups in an organization as well as produce "analytics for the masses," thus addressing political or organizational risks.

---

[4] SAS® Visual Analytics, in particular, is useful for data storytelling. It allows end-users to produce graphs that reveal the complex relationship among multiple indicators, making an unwieldly amount of data easily digestible. End-users also have the ability to interact with the visualizations to change the nature of the display, filter out what's not relevant, drill into lower levels of details, and highlight subsets of data across multiple graphs simultaneously.

In conclusion, “pedal-to-the-metal analytics” is effective in dealing with all four risks. The objective is to produce deliverables and an end product where the “work speaks for itself,” rather than one where the “speak works for itself” or, put another way, a situation where shaky and irrational rhetoric of no value for decision making upstages the creation of sound work products, causing analytics projects to fade into quixotic pursuits and eventually fail.

USE CASE WITH STRUCTURED DATA

The United States government is invested in making its data available through the use of APIs, which is aligned with the principles of transparency, participation, and collaboration, thus forming the cornerstone of its open government initiative.

The rise of APIs since 2000 has been impressive indeed, and 2015 was a landmark year in which the number of APIs exceeded 10,000.

One very important API-based data source is the Health Indicators Warehouse (HIW). It is a collection of over 1200 federal health indicators on health outcomes and determinants made available by Health and Human Services (HHS), and it includes Centers for Disease Control (CDC) data.

In recent years, drug poisoning or overdose deaths, as acknowledged by President Obama, have contributed to “one of the deadliest epidemics in America history.” In 2014, over 47,000 deaths in the United States resulted from this epidemic; that is more than a 747 airplane full of passengers crashing every single day for an entire year. The death toll from this epidemic is also approximately one and a half times as many as from motor vehicle crashes, according to the CDC.

Using the SAS web technologies discussed in this paper and API calls to the HIW, this data can immediately be made available for analysis throughout the Analytics Lifecycle.

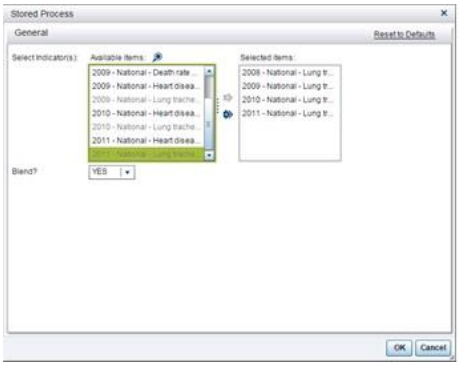
This helps answer the analytics question:

What clusters of US counties demonstrate attributes (income, education, unemployment, and so on) associated with drug poisoning deaths?

Data Access using APIs is possible using SAS Studio custom tasks and will “fuel” SAS web applications with these variables:

- Drug poisoning death data
- Mentally and physically unhealthy days
- Median household income (Income)
- Adults with a bachelor’s degree (Education)
- Unemployment rate

To access the initial batch of state level variables, a stored process can be used. County-level data can be accessed via the SAS Studio custom tasks discussed earlier. Of particular interest are HIW variables or indicators from the NVSS (National Vital Statistics System) or BRFSS (Behavior Risk Factor Surveillance System) systems as well as United States Census Bureau data.



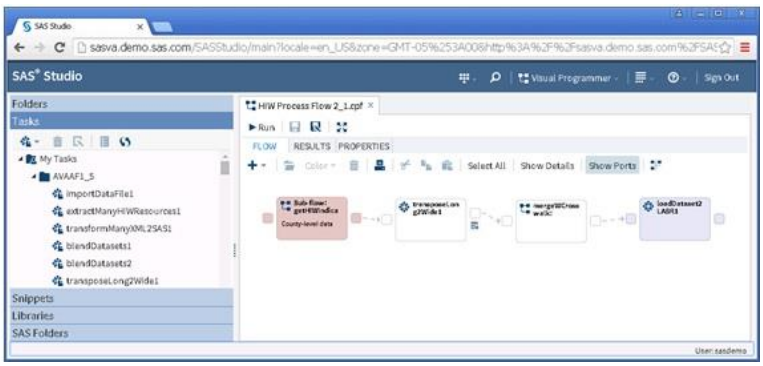


Figure 8. SAS Stored Processes (left) or SAS Studio Process Flow (right) composed of SAS Studio custom tasks (right) can be used to access HIW API-based data. Both utilize AVAF macros under the hood. The red custom tasks in the SAS Studio Process Flow are used for data extraction, the white signifies tasks for transformations, and the blue for loading. SAS Studio custom tasks are point-and-click user interfaces that guide the user through an analytical process.

The initial explorations of data variables show a strong correlation between “Drug Poisoning Deaths” and “Physically and Mentally Unhealthy Days.” Another noteworthy correlation is “Drug Poisoning Deaths” and “Household Median Income.”

These relationships can be used to produce a “Regression Line” and even a “Bubble Plot” to further deepen one’s understanding of the data variables extracted from the HIW.

All of this is shown in Figure 9.

Figure 9. Explorations using SAS Visual Analytics include a correlation matrix (showing a strong correlation between “Drug Poisoning Deaths” and “Physically and Mentally Unhealthy Days”), simple linear regression, and a bubble plot. The bubble plot in the lower right shows that “West Virginia” has the highest number of “Drug Poisoning Deaths” (y-axis) and “Physically and Mentally Unhealthy Days” (x-axis).

The insights gained from exploring the data are invaluable toward building an analytic model, and in this use case, it is a clustering model. Clustering is a method of data segmentation that puts observations into groups that are suggested by the data. The observations in each cluster tend to be similar in some measurable way, and observations in different clusters tend to be dissimilar. SAS Visual Statistics uses the k-means clustering algorithm to determine its clusters displayed as either a Cluster Matrix or Parallel Coordinates Plot.

The Cluster Matrix displays two variables at a time; that is, it is a two-dimensional projection of the cluster.

Of particular use is the Parallel Coordinates Plot. It consists of an x-axis and y-axis. The vertical axis displays the cluster ID, and the number of clusters is configurable by the analyst.

In the case of the horizontal axis, each variable used to create the cluster is a column with its binned range of values displayed vertically, as shown in Figure 10.

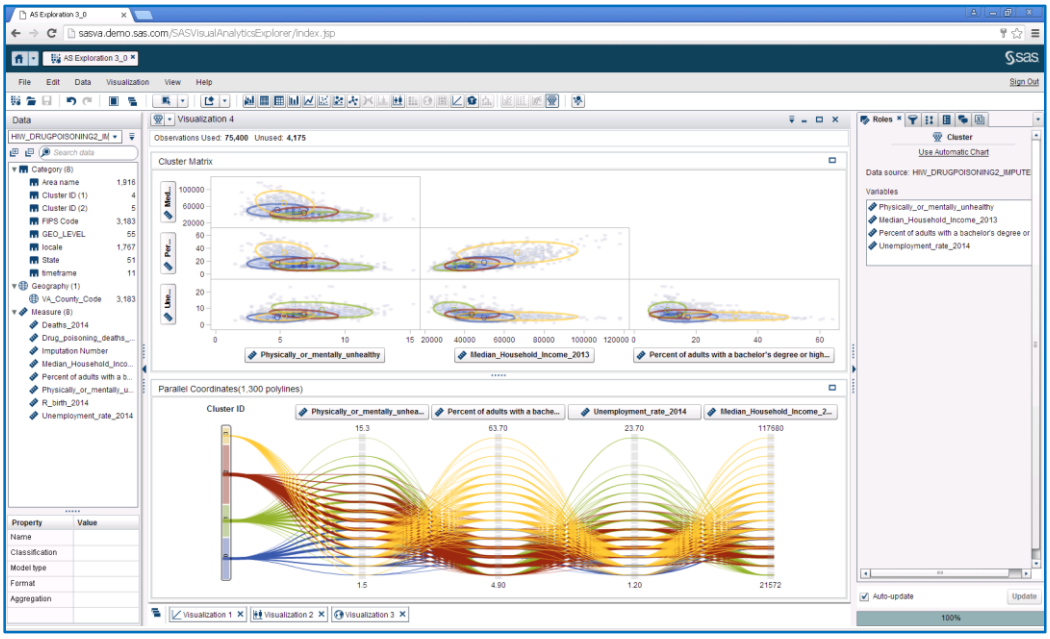


Figure 10. Clustering in Visual Statistics. The top visualization is a Cluster Matrix, and the bottom one a Parallel Coordinates Plot. Of particular importance for this use case is Cluster #1 (in green).

Clustering is a powerful modeling technique, and the source code produced from this modeling exercise can be exported and deployed to a data set, such as the one in the drug poisoning use case.

Results from a cluster model can be used to identify the counties that share the same attributes—for example, Cluster #1 includes counties with a high number of "physically and mentally unhealthy days," a low "number of people with a bachelor's degree," a high "unemployment rate," and low "median household income." See Figure 10. Clusters of US counties with these attributes include those counties at the epicenter of the drug overdose/poisoning epidemic, as shown in Figure 11.

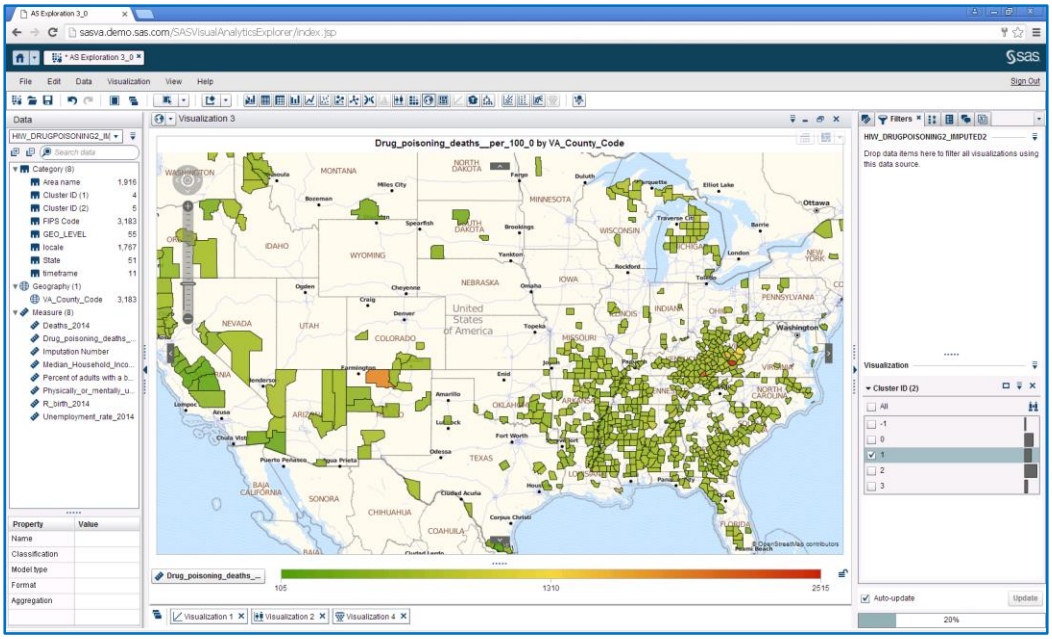


Figure 11. Cluster #1 is composed of counties that exhibit a high number of "physically and mentally unhealthy days," a low "number of people with a bachelor's degree," a high "unemployment rate," and low "median household income."

These results can be closely examined and broadcasted to a large audience by way of the “Report Designer” module in SAS Visual Analytics. See Figure 12.

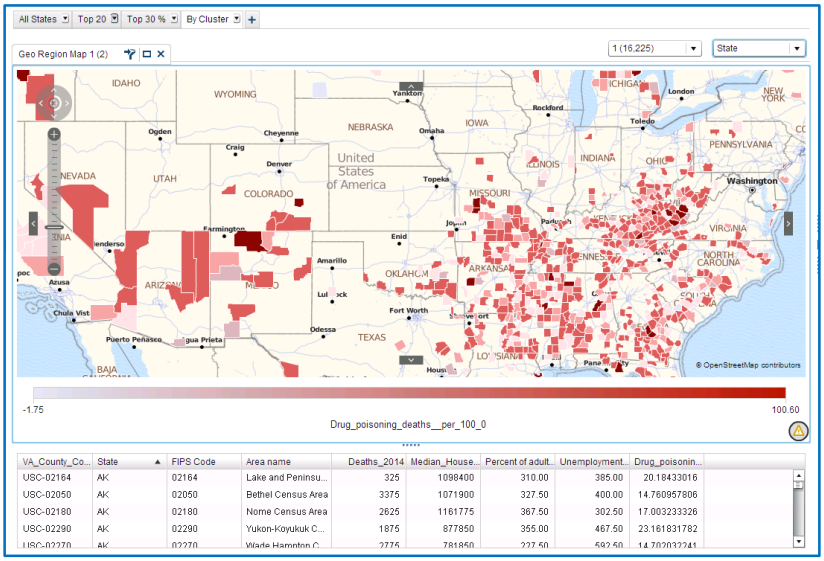


Figure 12. SAS Visual Analytics allows users to move from the x-ray to the cinema, or from close examination of the data in explorations and modeling to dashboards that broadcast results to a wide audience.

A future requirement would likely include analysis that extends beyond a single year of data. As such, insights can be gained in future iterations by applying additional analytic models to the data. Small area estimation techniques, for example, could be particular useful to facilitate supervised imputations of county-level data to achieve longitudinal analysis.

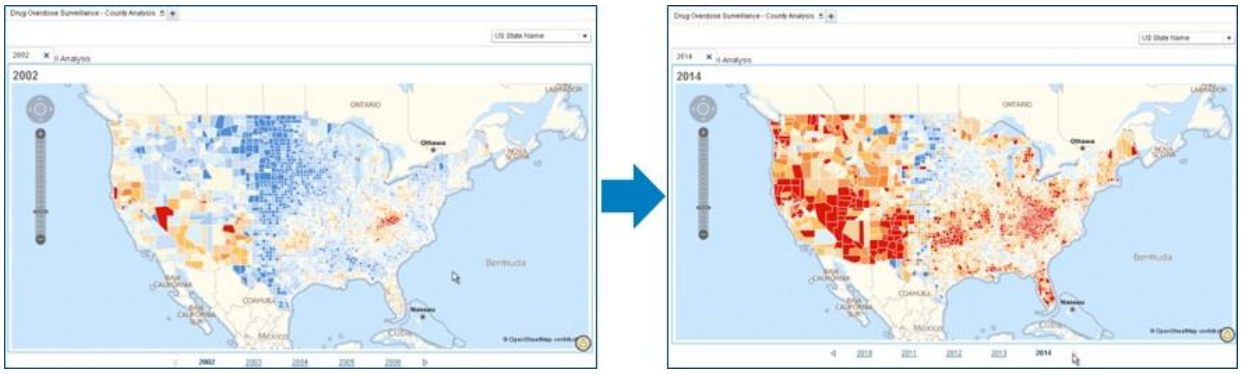


Figure 13. Further iterations of the drug poisoning use case can include small area estimation techniques to derive detailed county-level longitudinal maps from 2002 until 2014 (see Appendix for additional ones). US counties that are in dark red have an age-adjusted death rate greater than 20, a sure sign of an epidemic.

And, text models can be used to produce interventions that are closely aligned with federal government initiatives.

USE CASE WITH UNSTRUCTURED (TEXT) DATA

PubMed is a free database of references and abstracts on medical and health topics, such as drug overdose deaths. It is maintained by the National Institutes of Health (NIH), which provides a public API to allow for programmatic access to the database.

Using SAS Studio custom tasks once again, it is possible to extract abstracts using API calls to PubMed to answer the following text analytics question:

What are the appropriate and effective interventions to prevent, reduce, and reverse drug overdoses?

Since interventions use medical, clinical, and public health approaches designed to moderate biological and physiological factors and as a result improve health outcomes and reduce the risk of drug overdose death, PubMed abstracts are a suitable data source for evaluating and understanding them.

SAS Contextual Analysis is a web-based application that uses text analysis to provide a comprehensive solution to the challenge of identifying and categorizing key textual data. Using this web application, end users can explore unstructured data to derive initial insights (a term map, for example, shows the term of interest as the center node), as well as build models (based on training documents, such as PubMed Abstracts) that automatically analyze and categorize a set of abstracts while also producing sentiment surrounding them. See Figure 14.

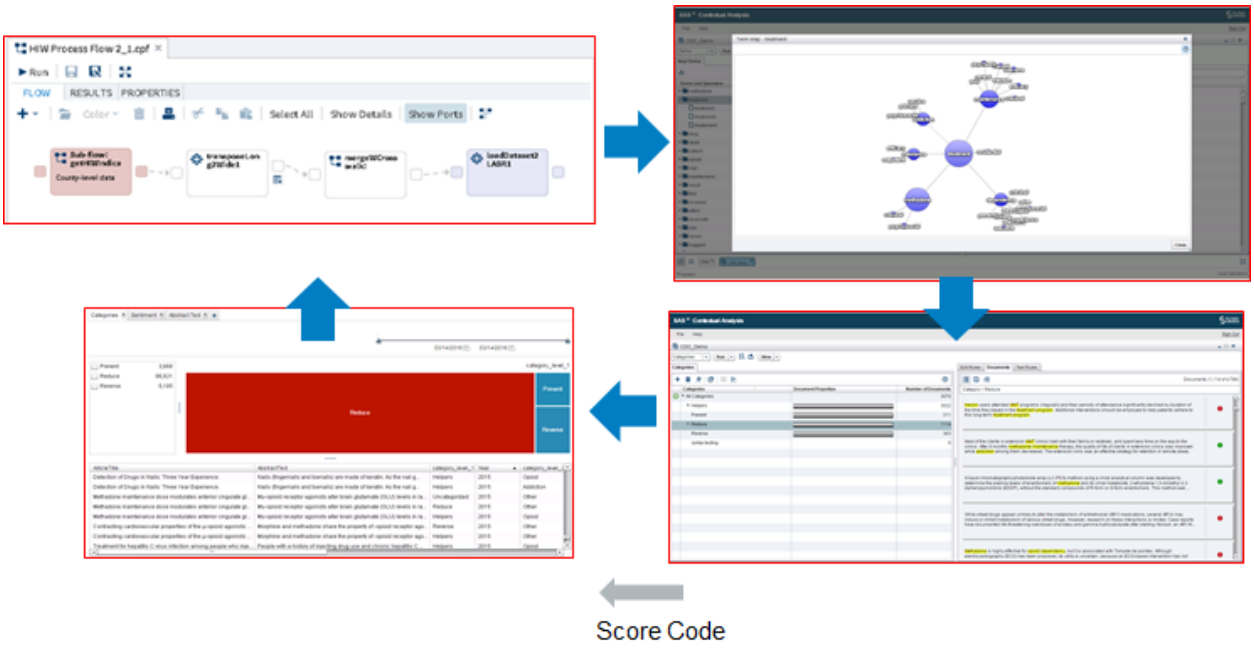


Figure 14. The SAS Analytics Lifecycle applied to unstructured data with SAS Contextual Analysis. In a clockwise order, SAS Studio provides access to Pubmed abstracts (upper-left), SAS Contextual Analysis term maps serve to explore (upper-right), categorizations rules can be built in SAS Contextual Analysis (lower-right), and results are published in SAS Visual Analytics (lower-left). Note that Pubmed abstracts categorized under "Reduce" are the largest (red).

To respond to the drug overdose epidemic in the United States, the CDC has launched a three-pronged strategy to fight the epidemic:

- **Prevent** people from becoming addicted to drugs, particularly opioids, by identifying high-risk prescription drug users early.
- **Reduce** drug addiction through improved access to Medication Assisted Treatment (MAT), whichcombines drug therapies (for example, methadone) with behavioral therapies.
- **Reverse** drug overdose with the opioid overdose antidote or reversal drug called Naloxone (Narcam).

These three “prongs” can also serve as categories for a text categorization model, as shown in Figure 15..

Figure 15. Sentence-level sentiment models built in SAS Contextual Analysis revealing sentiment for Pubmed abstracts categorized under a “Reduce” strategy as shown above. The horizontal bar graph captures sentiment, the blue bar graph frequency, and the line graph frequency trends.

This categorization model as well as (sentence-level) sentiment models can be applied to PubMed abstracts by scoring them to evaluate results in Visual Analytics to ensure a public health response or intervention to the drug overdose epidemic. This response can be targeted to specific high-risk counties and might include the allocation of resources (including Naloxone), prioritization of programs, and even validation of investments made, particularly for U.S. counties hard-hit by the drug overdose epidemic.

CONCLUSIONS: FROM DATA TO INSIGHTS….FAST!

This paper is ultimately about a journey that can be enriched and enhanced through the use of SAS web technologies and applications that are bundled together. It is one whose first steps can be facilitated by providing analytics projects with SAS web technologies and applications that will give them insights into structured and unstructured data, and hopefully lead them to build broad analytic capabilities. Or, at least help them gain momentum toward this goal. An analytics roadmap or lifecycle, furthermore, also provides those who decide to embark on the journey with guidance toward better achieving two things, which makes SAS® Analytics very *analytic* indeed:

- A clearer sense of requirements
- A clearer sense of risks

Although no project is ever free of risks, “pedal-to-the-metal analytics” is an effective prototyping strategy to help mitigate against risks. In addition to helping projects move from data to insights fast, “pedal-to-the-metal analytics” can also, more generally speaking, help organization change and mature into an analytics-ready one where analytics becomes a boon for effective decision making and, in the use case presented in this paper, saves lives.

In the end, this paper hopefully inspires analytics professionals to come closer to the truth. It is out there. Are you ready to seek it? If so, fasten your seat belts and enjoy the ride!

REFERENCES

United States Office of Management and Budget. 2013. “Open Data Policy.” Washington, DC: Office of Management and Budget. Available https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf. Accessed on March 16, 2016.

Health Indicators Warehouse. Available http://www.healthindicators.gov. Accessed on March 16, 2016.

Figallo, Manuel. "Macro Design and Usage in a Multi-Tier Architecture for ETL and Google Visualization API Integration. Available: http://support.sas.com/resources/papers/proceedings12/003-2012.pdf

Fowler, Martin and Kendall Scott. UML Distilled: Applying the Standard Object Modeling Language. Addison-Wesley, 1998.

Mike Porter, Amy Peters, and Michael Monaco, 2015. "What’s New in SAS® Studio?." Available: http://support.sas.com/resources/papers/proceedings15/SAS1832-2015.pdf

Pubmed API: Entrez Programming Utilities. Available http://www.ncbi.nlm.nih.gov/home/api.shtml. Accessed on January 10, 2016.

Health Indicators Warehouse. Available http://www.healthindicators.gov. Accessed on March 16, 2016.

NCHS Data Visualization Gallery. Available http://blogs.cdc.gov/nchs-data-visualization/. Accessed on March 1, 2016.

SAS Institute Inc. 2015. *SAS Stored Processes: Developer's Guide*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/documentation/cdl/en/stpug/68399/PDF/default/stpug.pdf

Abousalh-Neto, Nascif. 2013. “The Forest and the Trees: See it All with SAS Visual Analytics Explorer.” *Proceedings of the SAS Global Forum 2013 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings13/058-2013.pdf

ACKNOWLEDGMENTS

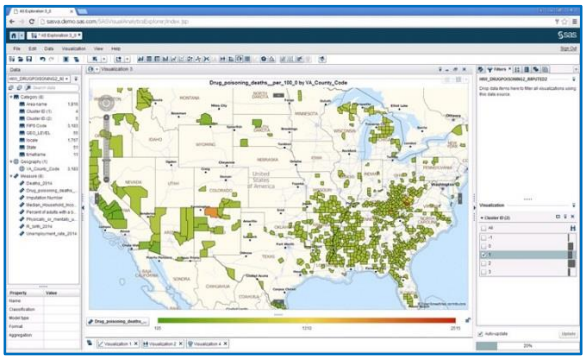
I would like to acknowledge my friends and family for their support during the writing process. It was an incredible journey in itself. Nicholas, Olivia, Sofia, Alexandria, Oscar, Enzo, Milo, Maximiliano, and Liliana, thank you for making life so beautiful! I am also very fortunate to work in an environment with great people, so I am grateful to those who reviewed early versions of this paper and especially those who made a great impression on me through their own work—they include Amy Peters from the SAS Studio team, Bobbie Wagoneer who trains in SAS Visual Statistics, and my colleagues at SAS Federal. Adam Pilz provided invaluable guidance toward mastering key parts of SAS® Contextual Analysis, and meetings with Rachel Dell that not only helped me sharpen my understanding of the software but were a lot of fun! The support staff at SAS, who I consider the best in the world, I am also indebted to. Grace Whiteis and Hunter Tweed were helpful in troubleshooting early version of my work presented in this paper, and Russ Tyndall has been for years my go-to person for macro-related questions. Finally, my deepest gratitude goes to SAS customers at the US Department of Health and Human Services (HHS), chief among them Dr. Lily Chen; they are true paragons of a strong work ethic and models of integrity for me. Without them, this paper would not have been possible. My career has been richer by encountering them on the journey.

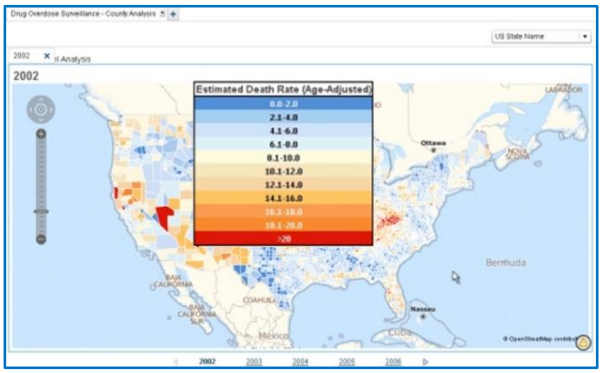
CONTACT INFORMATION

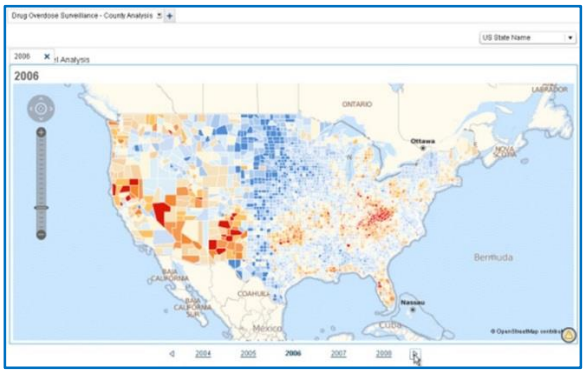
Your comments and questions are valued and encouraged. Contact the author:

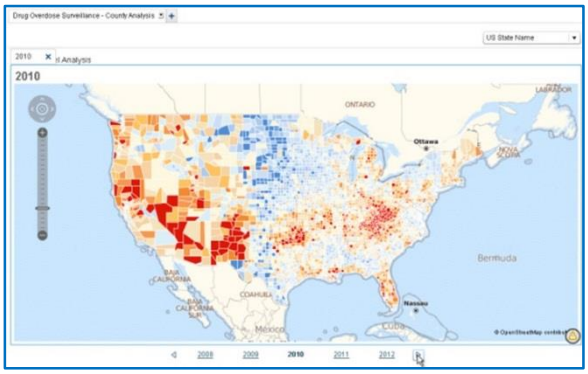
Manuel Figallo
SAS Federal
1530 Wilson Blvd, Suite 800 Arlington, VA 22201
SAS Institute Inc.
Manuel.Figallo@sas.com
http://www.sas.com
https://www.linkedin.com/in/mfigallo

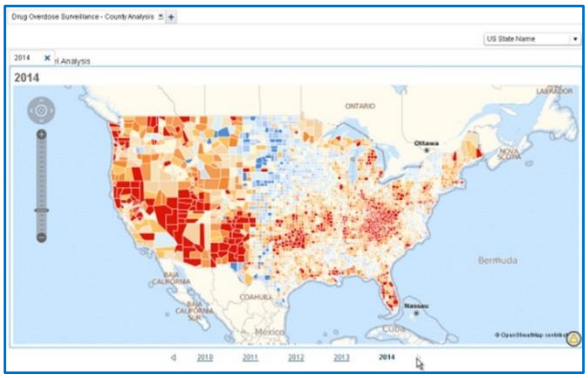
APPENDIX: ADDITIONAL SCREENSHOTS

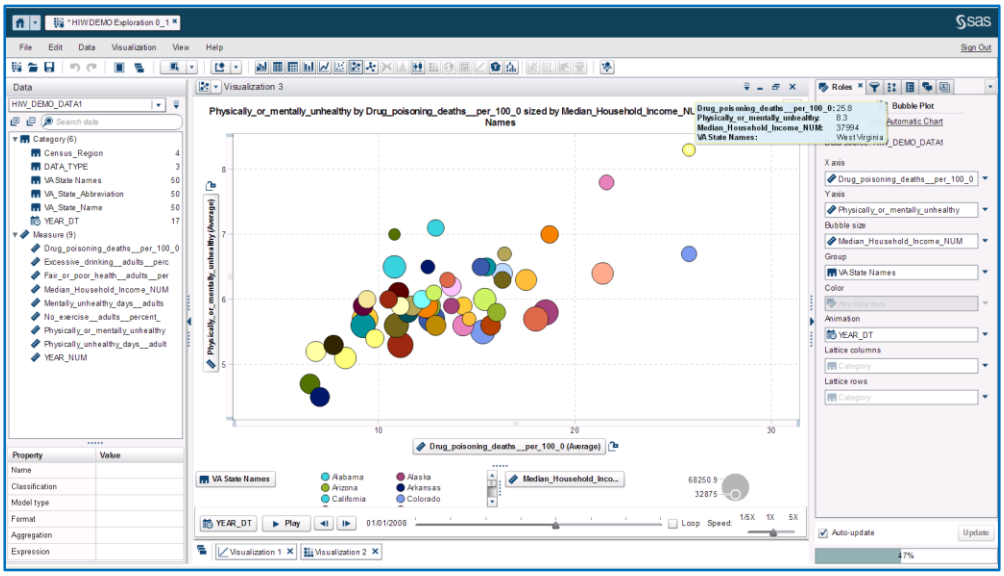


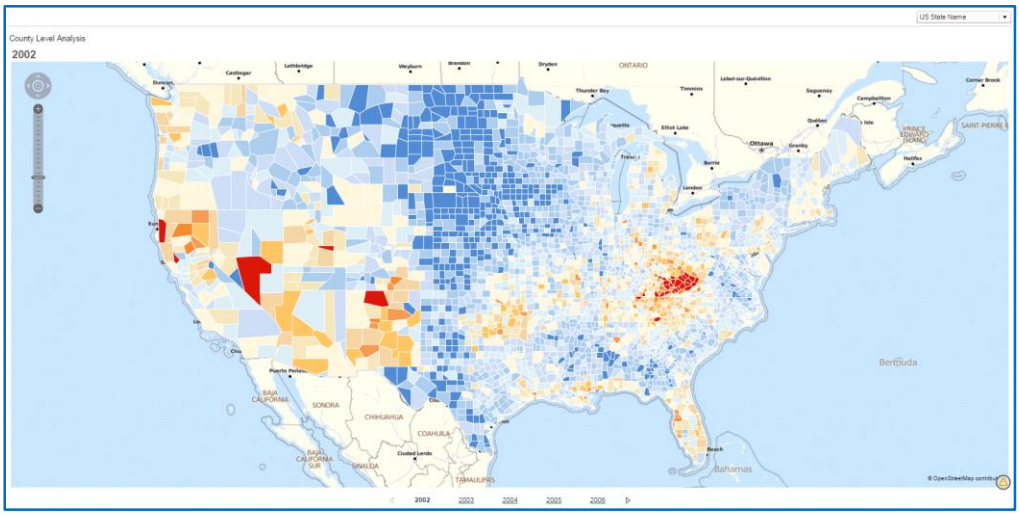


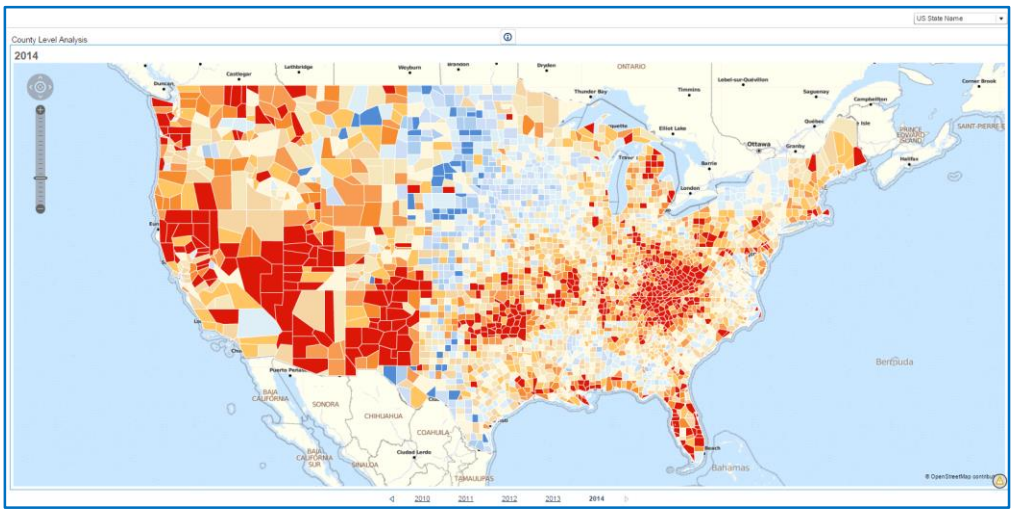







Physically_or_mentally_unhealthy by Drug_poisoning_deaths__per_100_0 sized by Median_Household_Income_NUM Names
Drug_poisoning_deaths__per_100_0 (Average)
Physically_or_mentally_unhealthy (Average)
VA State Names
Alabama
Alaska
Arizona
Arkansas
California
Colorado
YEAR_DT
Play
Visualization 1
Visualization 2


US State Name
County Level Analysis
2002
ONTARIO
Bermuda
Bahamas
SONORA
CHIHUAHUA
COAHUILA
TAMAULIPAS


US State Name
County Level Analysis
2014
ONTARIO
Bermuda
Bahamas
SONORA
CHIHUAHUA
COAHUILA
TAMAULIPAS