Inside the IoT Data Refinery
Extracting Insights from Telematics
Executive Summary

The ever-growing Internet of Things (IoT) is opening new doors for insurers in sales leads, underwriting, customer engagement, claims, and more.

IoT devices collect vast amounts raw data that can be put to work through data science, predictive modeling, and machine learning. Refining this data is complex and difficult to master, but the rewards are great for insurers that understand its potential value to them and their customers. That includes underwriting and claims, where IoT data may help improve risk assessment, guide fairer pricing, reduce overall risk, and improve workflows.

Motor vehicles are leading the way with driving data collected from multiple sources for use in insurance pricing. Data analysis puts the information to use, but first the data must be prepared through assessment, cleansing, and validation to ensure the analysis is accurate. This report depicts the typical assessment process whereby the Verisk IoT/Telematics team works with data partners of the Verisk Data Exchange™ to enable the creation of new data sets.

Laying a Dependable Data Foundation

![Diagram](image-url)

1. EXAMINE THE DATA
2. IMPORT THE DATA
3. ASSESS THE DATA

- A. Are general statistics sound?
- B. Is data compatible with specs?
- C. Is data complete?
- D. Does different sensor data align?
Gain First-Mover Advantages as a Connected Insurer

Insurers that are slow to leverage Internet of Things (IoT) data risk missing out on new market opportunities across the entire value chain. From attracting millennials, to streamlining claims processing, to developing the ultimate usage-based pricing models, first-mover connected insurers will gain significant competitive advantages.

Projections for 2020

The unrealized potential of IoT data presents vast opportunities
Introduction

In the new world where countless gadgets are connected to the Internet—a world called the Internet of Things (IoT)—a huge amount of raw data is collected, carrying with it a wealth of information that’s not so simple to derive. Data science, predictive modeling, and machine learning point the way, but these disciplines use complicated methods that few understand, very few do well, and even fewer master.

The world of IoT is penetrating every aspect of our lives as various stakeholders understand the value of the data and want to apply it for their or their customers’ benefit. Insurers also want to incorporate IoT data in their underwriting and claims processes, where it may help them improve risk assessment, provide more granular pricing, reduce overall risk, and streamline claims verification and settlement processes.

Insurers are also gradually adopting new, advanced analytics and using the newly available data to replace or supplement their traditional methodologies. Driving data—collected by plug-in devices, mobile apps, and even the cars themselves—provides the first common data elements used for insurance pricing, which will be followed by electronic data recorders (EDRs, or “black boxes”), connected homes, and even wearables.

Alongside the data analysis lies a less glamorous, yet equally important aspect of data science—preparing the data for modeling. Data assessment, data cleansing, and device harmonization are major aspects that must be applied to the data before modeling to ensure accurate analysis. This more laborious phase is the gatekeeper that controls the data feeding the model and prevents “garbage in, garbage out” results.

This report describes the typical assessment process whereby the Verisk IoT team works with potential data partners to the Verisk Data Exchange™ to assess new data sets and their use. Examples are focused on vehicle data because of its accessibility. This process can also be applied to data sets from original equipment manufacturers (OEMs), mobile data loggers, and aftermarket devices, accounting for the characteristics of the data in selecting the tests and key performance indicators.
Reading the specs
The first step in any data assessment or analysis of a data set (even before getting the first byte) is examining the data specifications. This examination can identify any architectural failure at the earliest phase.

When reading the data specs, prior experience and understanding of the various data elements are essential. Every data element comes from a different sensor and thus has unique characteristics. In many cases, different data elements are related and can provide multiple perspectives on the same measurement. Thus, combining data elements provides a holistic, multidimensional picture of the observations needed for modeling.

When reading the specs, it should be evident whether the available elements can provide a minimal viable model and whether the resolution and frequency of the data are sufficient for modeling.

The Verisk IoT team created a data ladder that defines six levels of viable data sets that will be compatible with our analytics, starting from a very basic data set and climbing all the way to a frequent data log with multiple sensors.

Importing the data
After reading, understanding, and approving the specs, it’s time to structure a database and load the data. It’s very important to ensure that the data is loaded into a database that fits the specs. Data specifications always have room for interpretation, and the first examination of the data reveals how the specific data partner interprets it. When building and extracting the data, it’s important to make sure that:

- the data format is compatible with the data set received
- no data is lost because of the database structure
- each data element is traceable and can be isolated when needed
- data is stored in its raw format as well as in a format sensible for analytics
Assessing the data
After a short, yet extremely important extraction process, we start to examine the data. A set of tests and assessments should be created based on the provided specs, but it’s important that we modify them as the assessment unfolds and when new findings on the data emerge.

The assessment steps and expected findings are as follows:

**General statistics**
Sometimes, easy analyses are not taken seriously, but they’re critical to ensuring that the more complex analyses are accurate. General statistics such as number of active vehicles, mileage, and driving time are the most basic extractions. These elements are simple to extract and examine; and with some experience, simple thresholds can be defined to determine which values are acceptable and which are insufficient.

“Basic extractions include number of active vehicles, mileage, and driving time.”

Below is an empirical example in which a simple count of daily trips identified a failure in a file compression, which caused some of the data (on a single day) not to be extracted. Not noticing it could distort our results and lead to incorrect conclusions. Once identified, the solution to the problem was very simple—just re-extract the affected day.

**Figure 1: Daily Log of Trip Count**
Compatibility with specifications
As mentioned above, specs are subject to interpretation, and so a very simple yet extremely important test is to compare the spec with the actual data. Failing to identify mismatches might lead to an incorrect analysis. We test the data for resolution, frequency, completeness, and coverage.

In the following example, we expected significant data only from the United States, where the pilot took place. This simple plot shows that the provided data sample indeed covers most major U.S. cities.

Figure 2: Visual Plot of Data Geographic Coverage

Completeness
Any data set may suffer from collection issues, such as temporary sensor unavailability or temporary failure in delivery. Good analytics should consider these scenarios and to a certain extent must be able to overcome temporarily missing data elements. Qualification of a data set should contain the amount of tolerable missing data, the allowed combination of missing elements, and the allowed duration of missing data streams. For example, if the analysis is based on GPS breadcrumbs in high frequency and the GPS data is incomplete for long periods, the analysis may lose accuracy. But if the analytics can use the speed provided by the vehicle to validate the speed from the GPS, then the affected period may still be usable.

The main challenge in identifying missing data is that “we don’t know what we don’t know,” and if the data set we have is missing large chunks of data completely, it may be difficult to identify and detect them. If a certain vehicle sends no data for a week, it’s very challenging to determine whether the vehicle was not driven during that period or whether
there was a failure in collection or transmission of the data (for example, the vehicle was out of range of cellular coverage). For that, we use the power of statistics: Comparing the data set to an optimal “gold” data set can highlight severe problems that are otherwise problematic to identify.

Another important consideration is the means of data collection. For example, a mobile phone estimates the beginning of a trip by detecting phone movement. Even with advancements in mobile data collection, a dongle or connected vehicle may be able to detect the trip faster.

“A dongle or a vehicle can identify the beginning of a trip sooner than a mobile app.”

The following diagram compares the distributions of “lock distance” for a representative gold data set and a sample data set. Lock distance is related to how long it takes for a GPS to lock onto the appropriate satellites and reliably report its current location. At the end of a trip, the GPS has usually been on long enough for it to lock and report the correct location when the vehicle is turned off. If the GPS instantaneously locks onto the appropriate satellites when the vehicle is turned on again, then it will report the same location as when it was turned off. However, many GPS devices do not instantly lock onto the satellites, so the first reliable position reported can be relatively near or quite far from the ending location of the previous trip due to a variety of factors. The lock distance is simply calculated by comparing the location at the end of the previous trip with the beginning of the current trip and serves as a proxy for how quickly the GPS was able to establish a reliable location. As is evidenced in Figure 3, the gold standard data has a lock distance of less than 100 meters about 90 percent of the time, whereas the comparison data sample has similar lock distances only about 25 percent of the time. We’ll need to consider whether this source meets our requirements depending on the intended use of data.

Figure 3: “Gold Standard” Lock Distance Distribution vs. Sample Lock Distance Distribution
Statistical comparison may not identify problems related to individual observations, but it may identify a systemic problem. We highly recommend searching for these problems as soon as possible and troubleshooting them early because they can be very difficult to resolve later in the project.

**Cross-sensor validation**

A powerful method to increase the accuracy of the assessment and analysis of the data is to cross-check data from complementary sensors, which are different sensors that can provide the same data element with minor differences. Alternatively, one sensor can provide data elements that can be used to derive a different data element. For example, speed can be collected from the vehicle’s CAN (Controller Area Network) bus, but it can also be derived from the GPS breadcrumbs. Another example is acceleration: It can be collected from the accelerometer, but it can also be derived from speed.

When comparing a data element from two sensors, we must understand the sensors’ behavior and their advantages and disadvantages in various scenarios. For example, an accelerometer is very sensitive and reflects acceleration instantaneously, but its sensitivity can make it volatile and noisy; getting a reliable signal from the accelerometer requires applying the proper algorithms to it.

The following graph illustrates the two-dimensional distribution of speed extracted by two different sensors. The x-axis is speed extracted by CAN bus and provided through an OBD-II device, and the y-axis is the speed provided by the GPS.

**Figure 4: Two-Dimensional Speed Distribution (GPS vs OBD)**
Driving data assessment forward
At Verisk, we’ve been assessing telemetry data for more than a decade, examining various technologies and dozens of different devices, data elements, and data sources. In creating the Verisk Data Exchange, we needed to ensure that all known data sources were compliant with the platform and our analytics. Considering all this, we decided to bring an automated data assessment process to the marketplace.

Verisk uses side-by-side validation to test mobile data and its compatibility with the Verisk Driving Score™, originally developed for use with aftermarket installed devices or OEM embedded systems in vehicles. After testing in 150 vehicles over 23 weeks, our analysis validated that mobile data meeting our specifications was sufficient to compute a Verisk Driving Score. However, there are challenges to consider: A phone is correlated to a person, not a vehicle; a phone owner’s actions can impede the collection of data; and a phone is not anchored to the vehicle, which can affect how sensors function.

The Verisk Data Exchange, the first-of-its-kind telematics data exchange for driving history, cuts through the tangle of multiple driving data formats, scores the results with advanced predictive analytics, and leads insurers over the regulatory hurdles to implementation—saving them time and money. They can work with one exchange, rather than developing separate collaborations with the 18 automakers operating in the United States and a growing field of telematics service providers.

“Insurers are learning how driving behavior relates to premiums and claims experience.”

Insurers using the Verisk Data Exchange are gaining insights they’ve never had before, such as how customers’ driving behavior relates to the premiums they pay or their claims experience. For prospective customers, insurers can gain insight into refining pricing decisions that may have previously driven business away.

Additionally, insurers are embedding the Verisk Driving Score into their quoting, binding, and renewal processes so that driving data is an active factor in their individual pricing decisions. There’s less guesswork regarding customers’ driving behavior, even before they become policyholders. No matter the data delivery solution—connected car, mobile, or dongles—Verisk has developed an effective mechanism for thoroughly evaluating the data.
About Verisk

Verisk (Nasdaq:VRSK) is a leading data analytics provider serving customers in insurance, energy and specialized markets, and financial services. Using advanced technologies to collect and analyze billions of records, Verisk draws on unique data assets and deep domain expertise to provide first-to-market innovations that are integrated into customer workflows. Verisk offers predictive analytics and decision support solutions to customers in rating, underwriting, claims, catastrophe and weather risk, global risk analytics, natural resources intelligence, economic forecasting, and many other fields. Around the world, Verisk helps customers protect people, property, and financial assets. Our industry-leading brands include ISO, Xactware, AIR Worldwide, Argus, and Wood Mackenzie.

Learn more about our IoT solutions at verisk.com/data-exchange

This is part one of a two-part paper on operationalizing IoT data. To receive the second part when it is completed or to ask questions about the Verisk Data Exchange, please e-mail telematics@verisk.com.
Author
Avner Freiberger
General Manager, Telematics Innovation Center

Contributors
Brad Huther, Saurabh Khemka, Dawn Mortimer,
Brendan Noonan, Nechama Silberberg,
Jared Smollik, Paulina Yick