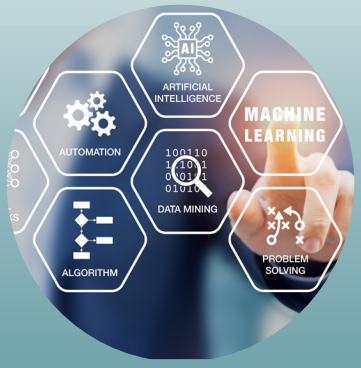
mosaic
DATA SCIENCE

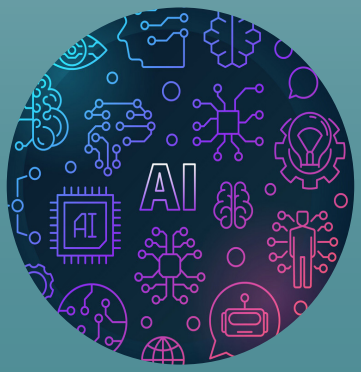
# *The* **VALUE** *of* **INFORMATION** *in the* **AGE** *of* **BIG DATA**

MOSAIC DATA SCIENCE | WHITE PAPER




## MACHINE LEARNING

Mosaic data scientists collaborate with customers, digging deep into the data to inform design and deployment of custom ML tools that make a difference.




## ARTIFICIAL INTELLIGENCE

Mosaic integrates powerful AI tools into clients' existing technology stack to solve complex business challenges.

## BUSINESS ANALYTICS

Mosaic helps corporations of all shapes and sizes take advantage of their data, transforming their decision-making processes.

## INTRODUCTION

This white paper explores how traditional models of the value of information (VoI) can be extended effectively to account for uncertainties presently inherent in gathering and analyzing big data. To illustrate the challenge, we explore the VoI an automobile manufacturer may derive by engineering a telematics system into its vehicles.

## BACKGROUND

Let's start by reviewing a few key basic concepts.

**Big data.** We define **big data** as data that an organization cannot store, process, or analyze economically using traditional storage technologies such as relational databases or text files. This definition has three implications that matter to us here.

1. One organization's big data may be another organization's small data.

2. Data is usually big because it has a low value per unit of storage (bit, byte, etc.), compared to the data that organizations have stored in relational databases for decades.

3. The point of gathering big data is the same as gathering small data: to capture transactions represented by the data (**online transaction processing,** or OLTP); or to analyze the data in a way that lets the organization improve its operations (**business intelligence,** or BI). Sometimes these two purposes merge in **operational BI**. Big data is not an economic end in itself.

**Decision analysis.** Next, **decision analysis** is the practice of formally modeling a decision to determine rigorously the best course of action available to an individual or group **decision maker.** The practice dates back at least to the mid-twentieth century axiomatics of researchers such as Stanford Professors Kenneth Arrow and Ron Howard.[2] And while decision analysis relies on basic concepts of probability and utility that are much older, it continues to be an active area of research. For example, in recent decades cognitive scientists have catalogued **cognitive biases,** frequently observed departures from decision science's prescriptions about rational decisions.[3] Some decision scientists have created ways to account for a decision maker's attitude towards risk.[4] Others continuing in Professor Arrow's tradition explore how best to model and improve necessarily imperfect group decision processes.[5]

**Decision analysis** usually represents decisions graphically as **decision trees** or **influence diagrams.** In this paper's examples we'll stick with the former, because they're likely more familiar. (The two are formally equivalent.[6])

**The value of information.** Finally, VoI is the difference between the expected value of a given decision in the absence of some piece of information, and the expected value of the same decision in the presence of that information (presumably having paid some price to receive the information). So a decision problem involving VoI models two decisions: whether to acquire the information, and the nominal decision the information may inform.

Here is a trivial example. Suppose you consider purchasing a used car priced at $20,000. If the car is in excellent shape (and it seems to be), the Kelley Blue Book Web site tells you the car is worth $21,000. If it's in good condition, requiring only minor repairs, it's worth $20,000. If it's in poor condition and requires a major repair, it's worth $16,000. A reliable used-car valuation Web site tells you that 50% of vehicles with the same make, model, and model year are in excellent condition, 30% in good condition, and 20% in poor condition.

Here's the decision tree representing this decision:

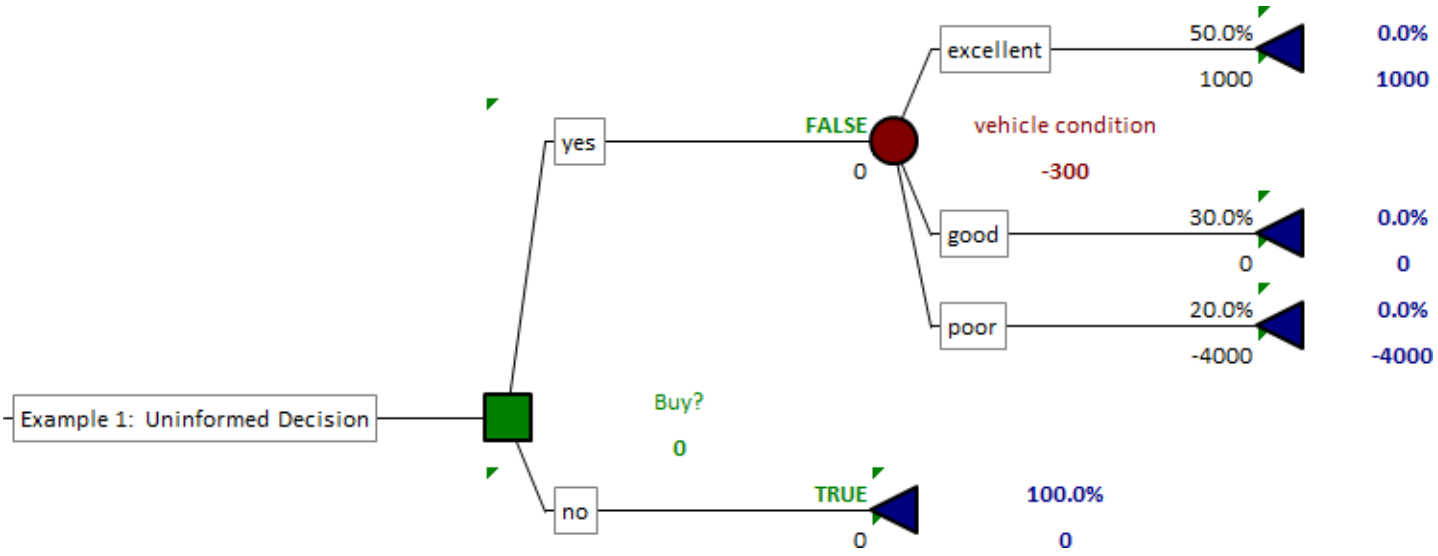



*Figure 1: Uninformed Purchasing Decision*

The expected value of this uninformed decision to buy is

0.5 * ($21,000 - $20,000) + 0.3 * ($20,000 - 20,000) + 0.2 * ($16,000 - 20,000)
= 0.5 * $1,000 + 0.3 * $0 + 0.2 * -$4,000
= $500 + $0 + -$800
= -$300

So buying the car without knowing what condition it's really in would "in expectation" leave you worse off by $300. Compared to doing nothing (at an expected value of $0), that would be a bad choice.

Before making an offer on the car, you could insist on taking the car to a mechanic for an independent inspection. The inspection costs $200. The inspection would reveal the car's condition with certainty. Now your decision looks like this:

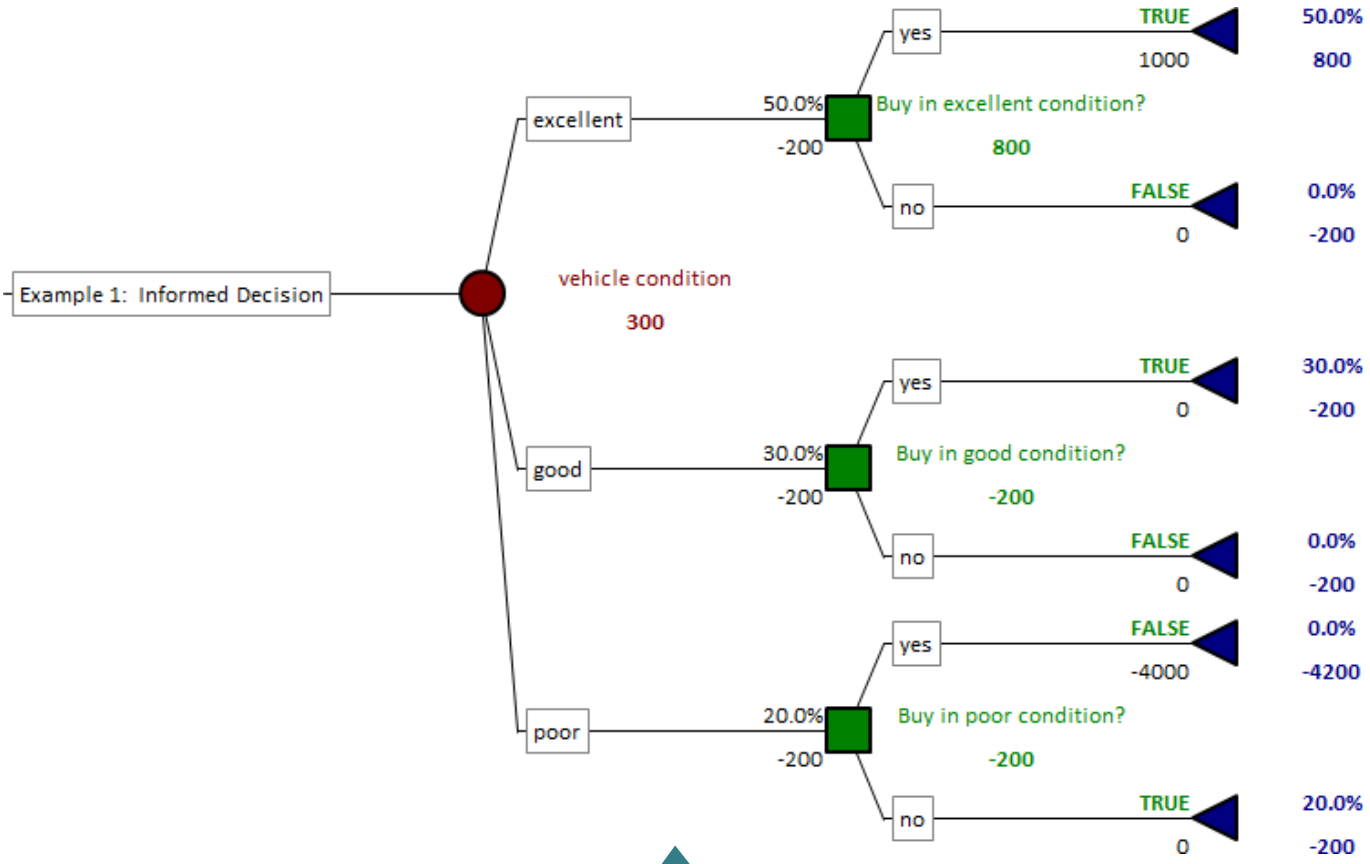



*Figure 2: Informed Purchasing Decision*

The expected value is now:

**-$200 +**
0.5 * ($21,000 - $20,000) + 0.3 * ($20,000 - 20,000) + **0.2 * ($0)**
= -$200 + 0.5 * $1,000 + 0.3 * $0 + 0.2 * $0
= -$200 + $500 + $0 + $0
= $300

As the terms in bold suggest, you trade away a certain $200 to avoid a 20% chance of losing $4,000 (a -$800 expected value). That tradeoff is favorable enough to make your overall expected value $300, which is now better than doing nothing. So in this case the information is worth buying. (In fact, a little reflection tells you that it would be worth paying the mechanic up to $500 to inspect the car before you decide whether to buy it.)

## BIG DATA'S VOL UNCERTAINTIES

BI specialists have long known that data warehouse (DW) end users do not reliably predict how they will eventually use the DW's historical data. Initially they often overlook use cases entirely, and for years later may persist in significantly misjudging the value of even well-known use cases having well-understood business cases.[7]

The same uncertainties become even more acute when working with big data. Because end users generally lack experience with the data, the organization may fail entirely to recognize a valuable use case. Or, the organization may be uncertain about whether and to what degree a known use case is valuable. Furthermore, unlike BI projects, which have fairly predictable development costs (if they're well managed), big data projects introduce two new uncertainties: whether the selected big data storage and analysis technologies will work at all, and if so how much it will cost to use them. In sum, big data is fraught with uncertainty as to implementation feasibility and cost, use cases, and use-case value. Merely deciding which of these uncertainties to reduce, by what means, and by how much, can be intimidating. This is where decision modeling incorporating VoI analysis really shines.

## THE EXAMPLE OF MOTOR VEHICLE TELEMATICS

**What are telematics?** Let's now explore VoI opportunities in the context of an automobile manufacturer considering adding telematics to its new vehicles. **Telematics** in this context combine telecommunications and informatics capabilities to make an automobile part of the "Internet of things." The vehicle sends operational data to the manufacturer, and may also receive operational instructions, software updates, etc. from the manufacturer. Ultimately the manufacturer must decide whether, when, and how to implement telematics. To inform that decision, the manufacturer may wish to gather additional information about the costs and benefits of telematics.

**Telematics mean big data.** Automotive telematics quickly become a big data challenge as the frequency with which a vehicle transmits data increases. To illustrate: one telematics system currently in production for fleets offers real-time delivery of diagnostic fault codes, fuel consumption, idle vs. work time, engine hours, odometer, temperatures, and pressures.8 So it would not be unrealistic to suppose an automotive telematics system transmits 10 four-byte numbers in a message that overall requires 100 bytes, once per time period. Suppose further that the manufacturer sells 10 million vehicles worldwide each year, and wants to track five years of telematics history (e.g. to cover the entire warranty period). At steady state the manufacturer would have 50 million vehicles reporting 100 bytes, or 5GB total, per time period. Table 1 below presents different periods and the steady-state data storage they require under these assumptions.

| Period | Transmissions Over Five Years | Total Steady-State Storage |
|---|---|---|
| Month | 60 | 300 GB |
| Week | 261 | 1.3 TB |
| Day | 1,826 | 9.1 TB |
| *Hour* | *43,830* | *219 TB* |
| *Minute* | *2,629,800* | *13 PB* |
| *Second* | *157,788,000* | *789 PB* |

*Table 1: Storage Requirements for Various Data-Transmission Periods*

The *petascale* numbers in the bottom three rows are big data for *any* organization.[9] So a well-informed telematics decision will involve all four uncertainties we cite above.

## POTENTIAL TELEMATICS USE CASES

Automotive telematics have several potentially profitable use cases. Each of them deserves enough study to determine how real and valuable the use case is, assuming the data are stored in a way that supports the relevant analyses. Here are brief discussions of eight possible use cases Mosaic has identified, including possible VoI analyses that might help the manufacturer gain an accurate understanding of the use case's value.

Design and engineering improvements. Telematics can provide the engineering function with a statistically precise, detailed characterization of the real-world loads experienced by key vehicle components. Those loads may vary geospatially or seasonally, as well as by vehicle type or option package. The manufacturer might review opportunities to improve component performance, reliability, or manufacturing cost, assuming it has this sort of detailed, localized knowledge of component loads.

Suppose, for example, that the telematics metrics include transmission temperature. This metric could result in the manufacturer discovering that certain vehicle models in specific geographies routinely run their transmissions at or above a high-wear temperature threshold, explaining a high rate of transmission warranty service for these vehicles. This discovery could lead to improvements in the transmission cooling systems for these vehicles in specific geographies (perhaps making a supplementary transmission cooler a required feature in these areas), to reduce expensive warranty repairs at a modest cost and in a highly targeted fashion.

This example illustrates a natural approach to investigating the value of specific telematics metrics. Rather than starting with the metric, data scientists would start with the manufacturer's historical warranty claims, to identify expensive classes of claims that might be dramatically reduced in a cost-effective way, if engineering had a more precise understanding of the loads experienced by the relevant components, and the markets in which those loads occur.[10] Warranty claims can represent several percentage points of product sales, and can be more than twice as high in bad years as they are in good years. For example, Figure 3 depicts the claims and accruals rates for one major automobile manufacturer, for the past decade:[11]

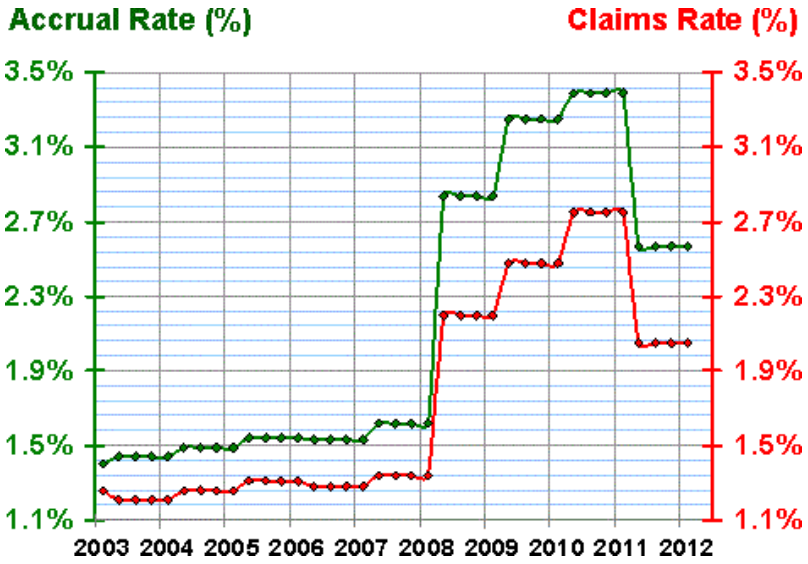



*Figure 3: Annual Claims and Accruals Rates for an Auto Manufacturer*

Clearly the opportunity for claims reductions through improved reliability engineering appears substantial. Possibly better operating metrics would let the manufacturer work far more proactively to eliminate quality problems early in the vehicle lifecycle, to reduce spikes in annual warranty claims and accruals. Targeted marketing. Automobile dealerships routinely send their customers direct-mail offerings related to vehicle service. The effectiveness of these campaigns has been limited historically, because the dealership has no way to time an individual mailing so it reaches a vehicle owner exactly when the vehicle needs the services offered in the mailing. Telematics can change that, potentially increasing dramatically the effectiveness of vehicle-service marketing campaigns. Besides producing additional service revenue, such campaigns can increase vehicle owners' compliance with recommended maintenance schedules, which will reduce the frequency of component failures that give rise to warranty claims. Telematics might also let the manufacturer increase revenues from warranty-extension offers, by letting the manufacturer tailor the offers to a vehicle's operating history. (Poorly maintained vehicles would require more costly extended warranties, or would not be offered warranty extensions at all.)

The increased profitability of targeted marketing can be modeled from known cost and revenue figures related to historical marketing campaigns. The main unknown in these models is the amount of uplift produced by the telematics data. These unknowns can be estimated from general knowledge of what constitutes good uplift in targeted marketing. For example, a 1% response rate is typical for a direct-mail marketing campaign based solely on household income and distance to the advertiser. More targeted campaigns can easily double the response rate.

Warranty-claim review. Vehicle owners sometimes fail to maintain their vehicles properly. When failures to maintain result in product failures during the warranty period, the owner, not the manufacturer, should be responsible for the consequences of the failure to maintain. Telematics can provide a basis in fact for denying warranty claims under these conditions, especially when the owner has received timely telematics-based marketing communications encouraging the owner to maintain the vehicle properly. A VoI study of warranty-claim validity and vehicle-maintenance patterns would help determine the magnitude of this opportunity.

Product-liability reduction. Product-liability claims fall into three classes.

- Manufacturing defects are deviations from design that result in bodily injury or death.
- Design defects are aspects of a product's design that are inherently hazardous.
- Failures to warn occur when the manufacturer knows of a hazardous defect but does not timely warn the consumer.

Telematics can help a manufacturer become aware of potential defects early enough to re-engineer, recall, and replace the relevant components, thereby reducing customer exposure to the defects. At the same time telematics can help warn consumers about the defect early enough to avoid failure-to-warn claims.

Some product-liability claims arise from vehicle misuse, either failure to maintain or improper operation. Telematics might help the manufacturer discover these facts and resist a product-liability claim on the basis of product use at variance with manufacturer design and recommendations.

A VoI study of historical product-liability cases could help quantify the proportion of liability cases that telematics could help the manufacturer avoid by each means: reducing defect impact, warning consumers timely about known hazards, and resisting claims based on improper product use or maintenance.

**Supply-chain management.** Telematics can improve supply-chain management in several ways. First, wear metrics could reveal disproportionate early wear in certain components. This might let the manufacturer anticipate and prepare for heightened demand for those components. If the premature wear were limited to a specific supplier or factory, the manufacturer could work with the supplier or factory to solve its quality problem far earlier than the problem might otherwise come to light through demand patterns for replacement parts. A VoI analysis could translate known frequencies of premature component wear (perhaps by component source) into an expected value of using telematics to detect such problems while they develop.

Finally, as we remark above, telematics could improve the service organization's ability to level demand and anticipate demand surges. A VoI study could translate known demand patterns into opportunities to improve the service organization's resource utilization through demand leveling.

**Leasing program cost reduction.** We have already remarked how telematics might improve warranty-claim expenses. That argument applies equally to leased vehicles. Moreover, the value of a leased vehicle at the end of the lease depends on the vehicle's condition. Telematics can help the manufacturer encourage its leasing customers to bring their leased vehicles to a dealership on time for routine maintenance, or apply appropriate surcharges when customers fail properly to maintain a leased vehicle. A VoI study of present maintenance-schedule compliance rates among leasing customers would suggest the size of this opportunity.

**Competitive advantage or parity (consumer perception).** Telematics create several feature/function opportunities that attract consumers. These are well known, and we do not review them here. Marketing researchers can project how much demand uplift a set of product features create, or how much loss of market share results from a failure to maintain feature/function parity with competitive products. These benefits do not relate to the information gathered by telematics per se, but they are nevertheless benefits that should be part of the analysis of the decision whether and when to implement telematics.

## BIG DATA TECHNOLOGY UNCERTAINTIES

Having reviewed the use cases that assume the existence and adequacy of a telematics data infrastructure, we must also investigate the uncertainties surrounding the infrastructure itself. Again, these uncertainties revolve around feasibility and cost. This is because, while the vehicle-side sensor and communications technologies are commodities, the manufacturer-side big data storage and analysis systems required by petascale telematics data require cutting-edge technologies. These technology-based uncertainties represent much of the total uncertainty of a telematics program.

**Feasibility.** A feasibility study must identify at least provisional solutions to many technical problems. Without meaning to be exhaustive, we suggest some of them here, and outline opportunities to improve technology certainty through VoI analysis.

**Transmission frequency by use case:** Each use case may require telematics transmission of specific metrics at specific frequencies. For example, supporting targeted marketing of maintenance services might only require daily transmission of total mileage. In contrast, supporting design and engineering improvements might require transmission of several metrics each minute. An accurate prediction of overall data volumetrics must account for these differences, and for volumetrics uncertainties for each use case. A VoI analysis could help specify plausible metrics and reporting frequencies for each metric, for each use case.

**Overall volumetrics:** Merging the use-case volumetrics into a single overall volumetrics estimate requires determining when it is reasonable to assume that one use case's volumetric uncertainties are independent of another's. If there are correlations in the use cases' volumetrics, the overall estimate must account for them in combining the individual use cases' estimates. Otherwise the overall estimate may underestimate its worst case. VoI analysis can help identify these risks.

**Analytical requirements by use case:** Each use case will have different query and analysis requirements for its metrics. Some will amount to simple queries, such as identifying vehicles reaching a specific mileage since the last transmission period. Others will just involve computing summary statistics such as average mileage per time period. Others still will require far more complex analytics, such as running a classifier to detect probable cases of warranty claim fraud.

Furthermore, query and analysis performance must be efficient enough to keep pace with the flow of data—a nontrivial requirement in big data contexts.
Adequate certainty about these requirements is imperative to ascertain feasibility.

**Overall analytical requirements:** The combination of analytical processing requirements imposed by the use cases may be too much to ask of a single storage technology. This can result in exploring several courses of action, such as

- abandoning a marginally profitable use case to avoid its requirements,
- adding an in-memory data store to support specific requirements, or
- partitioning the telematics data into separately stored subsets such that no single processing requirement reads data from multiple partitions.

VoI analysis might be valuable in determining whether real incompatibilities among individual analytical requirements exist.

**High-level data architecture:** The question of how to organize the data in its storage servers requires understanding how the server technology manages data and queries; but also how the server technology, analytical requirements, and expected access patterns interact. The feasibility study should address this question enough to identify at least one plausible data architecture for each proposed storage technology.

**Storage technologies:** Choosing a storage technology that fails to

- scale to required data volumes or throughput rates,
- support an important class of data-processing requirements,
- satisfy the application's uptime and recovery requirements, or
- support the application's CAP etc. tradeoffs[12]

can be a very expensive error, when it is not discovered until after the organization has invested substantial resources in the storage technology.13 The feasibility study should document all of these requirements and identify at least one storage technology, or one set of storage technologies (if more than one appears necessary), that satisfies all of them. If satisfying all of the requirements appears uncertain, VoI analysis can help determine

which requirements are at greatest risk.

**Analytical technologies:** In most cases choosing a storage technology supports several analytical tools. For example, Hadoop supports the Mahout library of machine learning and data mining algorithms, the R language, Java, and several SQL-like languages. The feasibility study should compare available analytical technologies on each candidate storage platform with the overall analytical requirements, to ensure the storage system will support all required analytics.

Some of the use cases may have uncertain analytical requirements. If so, VoI analysis can help reduce these uncertainties by determining (at least) which *classes* of analytical techniques each use case requires, and the likelihood that at least one supported analytical tool will provide a satisfactory technique in the required class.

**Hardware platforms:** Big data storage systems are generally designed to run on commodity hardware, so that they "scale out" cheaply. The feasibility study should ensure that the storage and analysis tools' hardware requirements, and the telematics system's overall reliability requirements, are consistent with at least one hardware option's processing power, I/O and network bandwidth, storage capacity, failover technology, etc.

**Cost.** Assuming the feasibility study has identified a provisional technology stack, the remaining area of uncertainty is implementation cost. Outlining the activities and concerns involved in a big data implementation project deserves a separate white paper.[14] Experienced petascale practitioners such as Shutterfly (over 30PB of big data[15]) report that continual hardware failures, uptime guarantees, and recovery processes become fundamental concerns. The cost of implementation depends strongly on the types and frequencies of hardware failures, as well as the number of hardware devices and software server instances that the solution must instantiate and administer. VoI can be an invaluable approach to identifying which cost projections are most uncertain, and to determining which of these uncertainties can be effectively reduced.

## PUTTING IT ALL TOGETHER

The decision-analysis process is less linear than this white paper suggests. In particular, there are significant interactions between different sources of uncertainty. Technology limitations can lead to decisions to scrap a use case that viewed independently appears attractive. The decision to implement one use case requiring frequent transmission of a set of metrics may dramatically reduce the cost of another, otherwise marginal use case, by satisfying its metrics requirements at little or no incremental cost.

Most important, several marginally attractive use cases can combine to make a much more attractive overall business case for vehicle telematics, in part because (as long as their outcomes are generally independent) aggregating many use cases pools their risks, making it unlikely that a properly executed telematics implementation will prove unprofitable. On average each use case will be marginally profitable. And over time, additional use cases for telematics data already in storage will accrue, improving the return on the telematics big data investment. This is perhaps the most surprising consequence of the fact that big data is, by definition, data that one cannot store and analyze profitably in traditional databases. In contrast with data in traditional databases, which mostly have very small sets of use cases, a key virtue of telematics data is the abundance of modestly attractive use cases these data enjoy.

1 Roger Magoulas and Ben Lorica. "Introduction to Big Data." Release 2.0 (O'Reilly, 2009).

2 Kenneth Arrow, Social Choice and Individual Values (Yale University Press, 1951). Ronald A. Howard, "Decision Analysis: Applied Decision Theory." Proceedings of the 4th International Conference on Operational Research (1966), pp. 55-77.

3 This tradition began in the 1970s with the work of Amos Tversky and Daniel Kahneman. See e.g. Daniel Kahneman, Paul Slovic, and Amos Tversky, Judgment Under Uncertainty: Heuristics and Biases (Cambridge University Press, 1982).

4 A decision maker can be risk averse or risk prone, paying to avoid risk or paying to experience it.

5 Professor Arrow is most famous for his impossibility theorem, which very roughly speaking says that no formal group decision process other than dictatorship or consensus can be fully rational.

6 R.M. Oliver and J.Q. Smith, eds. Influence Diagrams, Belief Networks, and Decision Analysis (Wiley, 1990).

7 These truisms give rise to several rules of "defensive" DW architecture.. For example, data architects are told to store data at the finest level of granularity available, even if known reporting and analysis requirements do not compel it. DW project management is likewise very oriented towards iterative process, anticipating the inevitable rapidly evolving business requirements.

8 http://www.aemp.org/category/sponsred-article. Retrieved January 28, 2014.

9 By way of comparison: as of May 2013, eBay had about 90PB of data supporting analytics. About 8PB was in a relational data warehouse. The rest was in Hadoop or Singularity, both noSQL storage/analysis platforms. True real-time vehicle telematics could thus produce an order of magnitude more data than eBay analyzes. See http://www.itnews.com.au/News/342615,inside-ebay8217s-90pb-data-warehouse.aspx, retrieved January 28, 2014.

10 The manufacturer would no doubt have performed a Pareto analysis of its warranty claims, and so would already know which few classes of claims account for the bulk of warranty expenses. The VoI analysis would then take the Pareto analysis as its point of departure.

11 http://www.warrantyweek.com/archive/ww20120628.html, retrieved January 28, 2014.

12 That is, consistency, availability, and partition tolerance. See http://www.cs.cornell.edu/courses/cs6464/2009sp/papers/brewer.pdf for a description and proof of Eric Brewer's "CAP theorem." See also http://cs-www.cs.yale.edu/homes/dna/papers/abadi-pacelc.pdf on the tradeoff between consistency and latency. The very large database (VLDB) literature documents many such tradeoffs.

13 We have seen this specific mistake cost one organization millions of dollars in rework and lost revenue opportunities. The organization chose a big data technology that was poorly suited to its processing tasks. By the time the organization recognized the mistake, the storage technology was in production, supporting several million users. The organization ultimately layered two other noSQL technologies over the original, to help support the system's loads—resulting in a far more complex architecture than the application would have required, had the organization chosen a better suited storage technology in the first place.

14 Or indeed a whole book. See e.g. Tiffani Crawford, Big Data Analytics Project Management (2013).

15 http://www.cio.com/article/704354/How_to_Implement_Next_Generation_Storage_Infrastructure_for_Big_Data, retrieved January 28, 2014.

**FOR MORE INFORMATION**

*Want to learn more?*

Please contact info@mosaicdatascience.com

mosaic
DATA SCIENCE


## ABOUT MOSAIC DATA SCIENCE

We provide innovative machine learning, AI and analytics consulting across organizations.

Mosaic is a leading data science consulting company focused on helping organizations build and deploy actionable analytics solutions. Our customers are as varied as the techniques we use — some just starting their first predictive analytics project; others with deep in-house machine learning expertise.

## HOW WE WORK WITH YOU

We work in a highly collaborative partnership with our customers to ensure you get only the best results to consistently drive business value.

### MACHINE LEARNING

*We design and deploy predictive algorithms to solve the most challenging problems facing businesses today*

### ARTIFICIAL INTELLIGENCE

*We bring a wealth of knowledge on how to tune AI models to deliver the maximum business value*

### BUSINESS ANALYTICS

*We leverage technical expertise and experience across a swath of industries, bringing fresh approaches to challenging problems*

mosaic DATA SCIENCE

mosaicdatascience.com

info@mosaicdatascience.com

(866) 202-8600