Insuring Data – What Where When Who Why?

Abstract:

As society moves to a digital economy, we need to protect data and its value, calculated in the form of measurable information. How we estimate and manage risk from misuse is a general issue but creates an opportunity for insuring against improper access and exchange. This paper describes a method for measuring value of information in data and how third parties can provide financial strategies to protect data no longer centralized in distributed storage environments, in particular cloud services.

Written By: Andre Szykier

"Information is the oxygen of the modern age"

$$E = -k\sum_{i=1}^{I} p_i \log(p_i)$$

$$E = entropy$$

$$p_i = prob\_of\_i^{th}$$

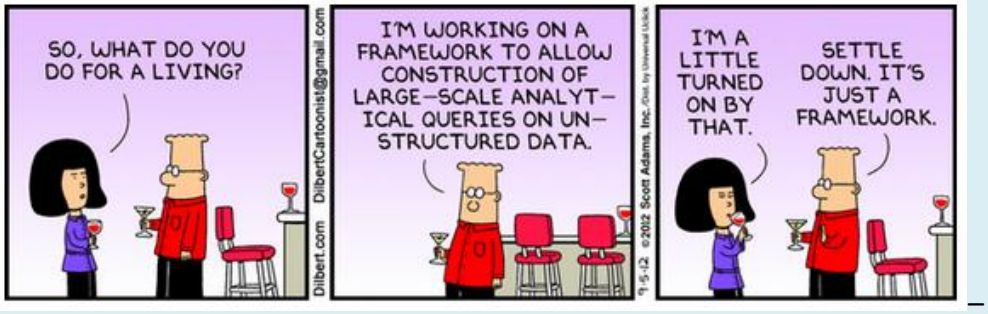
$$k = const$$

$$I = number\_of\_states$$

Background

In the industrial age of machinery, the often-quoted phrase “parts are parts” meant that any process consisted of interchangeable physical components. You build a car, a plane or a factory, everything was steps in a manufacturing process. Physical things have common elements.

The art was in how you built a process that produced a physical good. In the current world, data is an abstraction of a physical event, stored in a binary format (0,1) and readable by a computer program. It is meaningless unless one knows its structure as a series of bytes that represent an object (text, video, sound, program, sensor data and so on.). Often referred to as structured data, these bytes are the fuel for running a digital process.



Dilbert

While data can be seen as a physical object, stored somewhere for access by a program, information is how data is converted to meaningful use. This imparts ***value*** depending on where its used and the consequences of decisions based on its content. In other words, the value of information is in its ability to reduce “***uncertainty***” in decisions. Incomplete or erroneous data leads to decisions where ***risk*** can be higher.

Like any asset, data has an implied value. Your body temperature, heart rhythm, oxygen saturation, blood flow are physical events represented by digital measures. These examples are transient, in that your current readings are of more value to a doctor than those a decade ago.

The word ***transient*** touches on how we measure value – the half-life of data; how important the measurement is as a function of time. You buy a ticket to a concert with high demand and its value is based on what others are willing to pay. Once the concert is done, its value is zero. You record a mortgage and the property market value changes over time, up or down. Your health vitals represent periodic measurements. Seeing how they change is more important. In each case we measure information value over time.

Value – Defined

With anything of value, we try to mitigate risk through some form of protection. Digital value is defined in monetary equivalents. As such it lends itself to buying protection in the form of ***insurance***. Life insurance is a good example requiring actuarial tables that quantify risk as one ages. The older you are the higher the risk of dying and the higher the insurance premium. However, the younger you are and experience a wrongful death, the greater the potential loss of future income; here your value is measured by future rather than current potential.

Opportunity

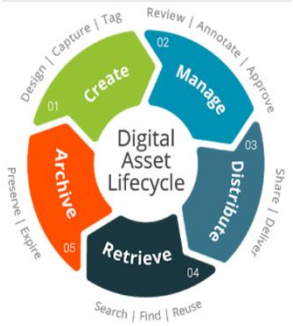
How do we define the value of a digital asset? This is a commonplace issue. As stated earlier we need to have a method that assigns value at any point in time. We also need to know that this value is fungible, namely how do we establish a monetary value to a transaction, current or future, that represents risk.

There are many digital assets whose value is established at the time of a transaction. Some examples are stock swaps, derivatives, contracts, and items with a speculative value. For the last example, alt-currencies have a market value that depends on a third party wanting to transact. It is based on a perception that transacting now is more beneficial from the standpoint of risk than at a future date. The value of the asset is time dependent.

To insure against loss, one needs to make sure that:

- ✓ The value is quantifiable
- ✓ It is transparent to those engaged
- ✓ It is immutable – cannot be modified
- ✓ It can be measured by normal accounting principles
- ✓ The process is secure from unauthorized access

Digital Asset



In principle, one insures a current or future digital event in the same manner that would apply to a physical asset transaction. A ***digital asset*** is a formal representation of some physical or defined asset. A key security requirement is that the digital equivalent represents its physical construct in some acceptable, non-repudiable way. We already do this with trusting the recording of transactions with third parties: your billing information, a record of data transmissions, a digital record of a contract or claim, even the verification of affected parties.

Digital Security

Security is defined as knowing that data is not tampered with and that exchange of such data is not subject to external threats against legitimate parties.

There are four principle of data security protection: access, storage, transmission and modification. For insurance purposes our model (**ASEV**) needs to know,

1. who has **A**ccess,
2. where it is **S**tored and protected,
3. how is information securely **E**xchanged, and
4. its **V**eracity when used for intended purposes.

Only then can you define and measure the cost of insuring data.

1. Access

You need to know who gets to see and use the data, in part or whole. This is a complicated subject. It boils down to identification, authorization and rules for data exchange. Each party may have different permission rules for access and display. Two parties may have different subsets of exposed data and ability to modify content. Keeping track of the who, when, where and why rules is critical. These rules are not universal but differ among engaged actors.

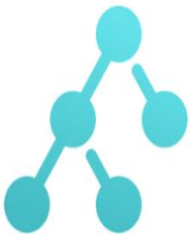
Access depends of ***identity*** and ***permission***. In finance, knowing your customer (***KYC***) is mandatory for legal and taxation requirements. KYC plays a critical role in anti-money laundering (AML) regulations when money or financial documents cross country borders. The movement to multi-factor authentication, devices that support biometric readings and special devices that are required to permit data exchange (cold wallet for transacting alt-currencies), are many ways for managing identity in a digital form

Permission is another set of rules once identity is established. You may belong to multiple communities (typically online), each with different rules. Your activity may modify rules. Your lack of activity may block access over time. Permissions are a broad subject, but they play a strong role in defining the value of information.

2. Store

Even though data is a binary representation, it has one physical property: where to store in a physical medium. Your phone stores images. A copy can be sent to another person. It can be backed up to a cloud service. It can be moved to another medium or printed. One problem is apparent; where the identical images end up. These copies are independently managed, mostly by others offering a service. Another problem emerges; how does one know that all copies are identical to the source?

3. Exchange



Data has one hidden property that imparts value – who needs the data for their purposes. Data with no purpose has no intrinsic information value. If it exists just for the creator, its value is subjective. For example, clinical trials create data useful for the researcher. Once enough data is collected and analyzed, it may or not have value to the pharmaceutical sponsor. At that point it's the information represented by the data that has value. In this example, billions of dollars are at stake.

In a simpler case, ocean sensors can detect marine earthquakes. When enough sensors confirm similar readings, the data has immediate value to allow actions to protect coastal areas Then the data has value to allow actions and protect coastal areas to minimize losses which also may cause billions of dollars in damages.

4. Veracity

Simply put, is it tamperproof? Keeping an original document of your birth, marriage, citizenship or death are legal binding constructs important to protect. But what if it is a digital copy stored somewhere by another party.

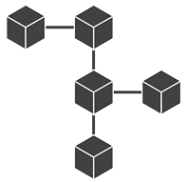
This introduces the principle of ***trust*** – that where the digital version is stored is secure and auditable.

In the case of the drug industry, knowing that the pill or vial contains unadulterated chemicals, or they were produced by a licensed manufacturer is a critical problem today. The same goes for high net worth goods that reach a global market. Fraud is rampant in brand goods that are patently fake copies.

While there are numerous ways of "watermarking" physical goods, doing the same for data is not so simple. Data is protected by encoding through algorithms that hide the content.

Encryption is one method. **Shredding** data into fragments before storing it on a medium is another. **Signing** data using private-public keys is a common way, creating metadata that

confirms the data is original. 2-dimensional bar codes (**QR code**) is a popular method and heavily used for online commerce in China. Even **biometrics** plays a role by signing dat a with information that establishes the veracity of the copy, its originator and its custodian.



More recent, storing data in decentralized systems has made unauthorized access more difficult than data stored on centralized data servers. **Blockchains** are emerging as a way of storing information in multiple copies across nodes in a network where all nodes are untrusted. Recording data in a block occurs with all or a subset of network nodes agreeing to record the block. This process is defined as a ***consensus***.

Blockchains are the foundation that supports alt-currencies such as Bitcoin, equivalent tokens on competing networks such as Ethereum and hybrid public/private networks such as Hyperledger. Because blockchains have a defined limit size, they have spawned support for data stored ***offchain*** land linked to the blockchain data. With sophisticated security methods, complex offchain data is referred to smart distributed ledger technology (**DLT**).

ASEV Risk Function

First, we define the terms that create a measure of risk.

A - When and where and who creates the data

S - What methods are used to secure data at rest in a physical medium

E - What methods are used to secure data in transit during exchange

V - What is the veracity of this data at any time.

β - Volatility of the data per unit of time.

Let's break these out in laymen's terms.

A (Access)

We try to establish that the data source is valid, that its is captured in near real-time and comes from non-random sources. Unvalidated data increases risk because our trust in the source is low. If the data is older, it has a greater chance of modification from its original state. And if there are multiple sources then the chances of different values for the same data increase. The most important part of **A** is how do we authenticate the data source or access to stored data. As we stated earlier, KYC in financial transactions is a standard way to manage this risk. With posts to social media, knowing that the source is valid and not an artificial bot is very difficult.

S (Store)

Cybersecurity is about protecting physical access to binary information, by securing communication paths and protecting the physical infrastructure where information resides. Every computer resource depends on hardware, connectivity and software. In turn they depend on three components: operating systems (OS), programming languages and device characteristics.

A system that is unconnected to other processes (example, a laptop with no Internet connection), is more secure than one that accesses the Web to connect to another. A device that has limited computing power has a lower risk than one with an OS that supports programming languages that in turn may have serious security flaws.

A network process that does not provide data obfuscation (TC/P) like encryption, is more riskier than one that does (SSL/TLS). A system that operates as a private, dedicated resource to its clients (example Salesforce) is less riskier that a cloud service (AWS, Azure, VMWare) that hosts multiple customers with unrelated services (Cryptocurrency, Content Distribution, E-commerce, Social Media.)

E (Exchange)

How data is exchanged between parties plays an important role in establishing risk, hence insurance. First it involves managing identities as referenced in **A**. Authentication reduces risk. Second, it requires permission rules (who can access what in whole or in part.) Third it requires robust protection of data exchange through communication channels. Fourth, it must support for an immutable record of the event for audit and reporting. Fifth, it requires a formal process for identifying what the insurance covers during an exchange.

V (Veracity)

This function depends on the collective risk management in **A**, **S**, and **E**. Ensuring that the data is unchanged at each level is important. The audit function relies on the trustworthiness of what is recorded, stored and exchanged. In practice it can be a cumulative function of the relative risks in each process step.

β (Volatility)

Measuring β is more of an art than science. One trivial example is used by or sell by dates. They mean two different things for the same product but have different impact on the consumer . If the product is a can of soup, the use date can be measured in years. For the retailer, it reflects inventory management strategies. For milk, the quality of the product makes the dates more closely aligned. Same for packaged vegetables where track and trace of origin becomes more important.

What we mean by volatility is the value of information from the data and how it may change at any moment. A stock option is fixed with an exercise date. A warranty is a similar example. But with cryptocurrency volatility and market value can impact interday swings of several percent. An unreleased blockbuster film has extremely high value and risk of unauthorized copies has a major impact on future revenues from distribution networks and theatres. After a time, this risk diminishes until its marginal value is small enough that access is granted for free.

One can apply a few rules to measure volatility. For example, for any unit of time,

- ✓ Will the information value change?
- ✓ What is the size of the market for this information?
- ✓ Are there alternative sources?
- ✓ Are there legal or commercial consequences if data is exchanged?
- ✓ What is the expected useful period for the data (half-life)?
- ✓ Is the process to capture, store and exchange data friction-free?

These are just exploratory ways of defining volatility that impacts the AESV model and are not meant to be exhaustive. That's why its more of a disciplined art.

Insurance Model

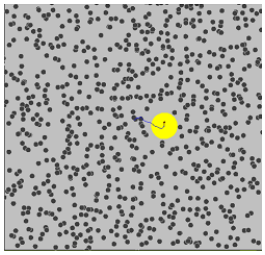
Risk – Price - Insurance

Insurance is defined as a monetary price based on risk in the execution of an event. In order to quantify risk, it requires a model that examines risk over time. We can establish a value at ***t=0*** and make several assumptions:

1. Risk **ρ** remains constant ($\rho_{t=0} = \rho_{t=1} = \ldots$)
2. It changes based on previous values over time (***t-1, t-2, …***)
3. It changes a future value over time (***t+1, t+2, …***)
4. It measures risk in a feedback model that uses 2 and 3.

The model design is not so important as knowing what factors (endogenous and exogenous) are required. Some models use attributes that are actual metrics, while others are either constants or the themselves a result of a function. In principle, the model is ***stochastic***.

> **Stochastic** process involves a randomly determined sequence of observations each of which is considered as a sample of one element from a probability distribution.

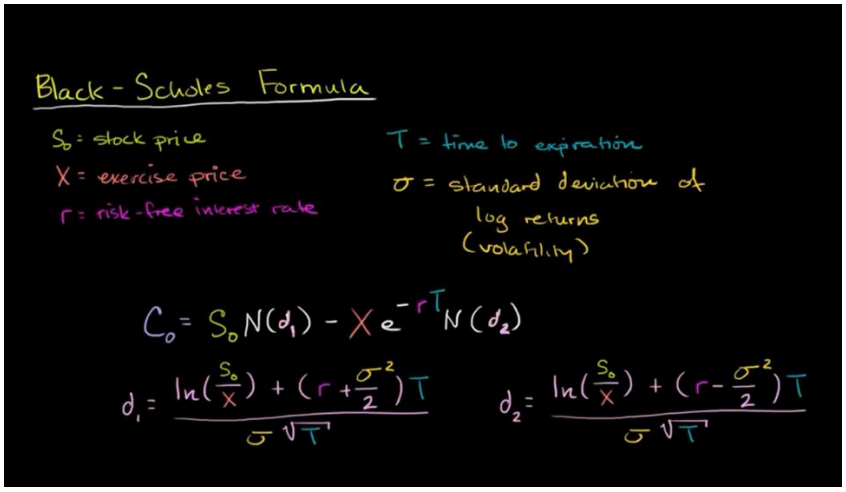


Many risk models adhere to the principles of ***Brownian motion & information entropy***[1]. A classic case is the **Black-Sholes**[2] investment equation. It can be adapted in part to establish a value for measuring risk. While Brownian motion originated in explaining the random movement of microscopic particles in a fluid, it readily describes volatility in financial markets where assets have changing prices that evolve in time.[3] This model requires an assumption of perfectly divisible assets and a frictionless market (no transaction costs occur either for buying or selling). [Click on image to animate.]

The idea of insuring a current or future action as it relates to digital data is somewhat opaque. Our earlier examples of data and information content were somewhat simplified. If the information has low volatility (in information theory, we refer to this as entropy: a measure of the unpredictability of the state, or its average information content,) then the risk is relatively constant over time, regardless of the information value to others over time. But, if its value is volatile, cryptocurrency for example, then insuring becomes riskier if the ASEV model is compromised.

Black Scholes is primarily how to measure risk from buying or selling a financial asset, usually stocks or bonds. If we substitute an asset that is data with an informational content, the equation can apply to measure risk of data loss or change, hence its insurance. Jack Traynor developed CAPM (Capital Asset Pricing Model) which states that risk and return are the same thing. Black focused just on risk.

The volatility of an asset is measured by the standard deviation of a continuously compounded rate of return, nominally over 1 year. Alternatively, we can say it's the standard deviation of percentage change in the asset price - in our case, the value of the information in the data. Our ASEV model (asset, security, exchange, veracity) defines the components that define the composite information value over time.



How this model is described and formulated is presented in another paper. In the Example section we explore how risk and value are established and how insurance can be applied.

Examples

The following are use cases on different data environments and how insurance can be used to quantify the risk of protecting data and its information payload.

Healthcare

This is an ecosystem of complex data types and exchange methods. It includes hospitals, insurers, clinics, treatment centers, pharmacies, clinical trials, pharmaceutical manufacturers and analysts. On of the key problems the industry faces are an accepted data standard that uniformly describes data provided by a stakeholder and used by another party, including the patient. A corollary problem is the isolation of information in the form of data lakes or silos that are not well suited for exchange through a process referred to as ***federated data networks***.

> A federated database system is a meta-database (DBMS), which transparently maps multiple autonomous databases into a single federated database. The constituent data sources are connected in a network. Since the sources remain autonomous, a federated database system is an alternative to the task of merging several disparate databases.

One of the global pressing issues is the ability of a patient to have digital access to their health records. In some countries like Japan and Korea, sharing of information requires a patient's consent. As data is pass between parties, notifying the patient is paramount. A number of trials have focused on using blockchains to record such transactions, improving the veracity of data.

For insurers, understanding the temporal nature of data (half-life) or the longitudinal value is critical as well as the requirements of the AESV model in healthcare.

Cryptocurrency

While cryptocurrency transactions are by design secure (blockchain method) the management of a crypto user's portfolio is not as secure as promised. Users connect with third parties who act as exchanges that connect to the blockchain network. For the most part, the entrust the exchange with their portfolio and using unique access keys perform transactions.

Unfortunately, these exchanges are targets for hacking, the result is a loss of said keys that allow the miscreant to use the portfolio for their own benefit. Since crypto currencies such as Bitcoin by design are anonymous, once stolen, the original portfolio owner cannot recover what they owned.

To address this issue, various solutions have been created. Soft Wallets, Cold Wallets, Multi-factor authorization, adoption of KYC defense mechanisms along with cybersecurity for digital services by exchanges are attempts to mitigate this risk. So far, losses continue which have a big impact on the speculative value of a listed cryptocurrency or token.

Insured vendors could provide protection by keeping the actual transaction processing and storing the portfolio in a system that adopts AESV practices. Already, one company in Japan offers this service to wallet holders and exchanges. The insurance guarantees up to a pre-defined transaction amount that any activity that goes through their system is covered.

Supply Chain

By definition, supply chains involve multiple parties. While carrier insurance is standard for exchange of physical goods, data plays a powerful role in making it efficient and cost-effective. Keeping contracts, invoices, bill of lading, originating source for track and trace, proprietary documents shared among the parties, are all examples of where data insurance plays a role.

Identity

Managing one's digital identity consists of many data components: birth certificate, citizenship, eligibility for government services, passports, marriage and professional licenses, employment history, certifications, education, health, voting and taxation records, memberships, all are examples of data that is collected independently from other data but form a comprehensive identity. This is a special case of federated data networks in that parts of an identity can be exposed to one party but not the whole. Any system that follows the AESV model could insure the digital owner against misuse by authorized and unauthorized parties. A tamperproof digital identity is slowly appearing in some countries such as Finland and Estonia which have adopted advanced technologies to create and protect this data.

Fintech

The global financial system has intentionally maintained a closed business model to attract capital investors and manage a lucrative transaction network that generates substantial fees. With the advent of cryptocurrencies, they are exploring how blockchains will make their systems more efficient while maintaining a closed monopoly.

Because of government regulations, they are already using KYC/AML as a means to identify transactors as well as financial players who attempt to move money and assets outside the system – money laundering.

Fintech is really an agglomeration of various data methods described in the previous use cases. Insuring data in motion and rest is a perfect situation in this market.

Content

Copyrighted content or information that requires permission to access for a price is a natural candidate for insurance scenarios. Content Distribution Networks (CDN) that stream media involve millions of subscribers across the globe. The content owners license their media to these networks and rely on their AESV practice to ensure that it is not illegally copied. This applies to film, music, photographs, publications et al.

In another example, confidential information shared among parties carries potential risk from legal liabilities. Law firms are a classic example as are those sharing confidential government information with various security levels. With the growth in genetic tools to identify health risks based on DNA, such data can be misused without the knowledge of the person.

IoT

So far, we focused on how to employ AESV with people at the center of discussion. Yet the biggest future growth of data will be between systems, sensors or other devices that handle data without human interaction or with only minimal supervision. Your doorbell camera, your remote health diagnostic device, your driver assisted vehicle, your home security system, energy monitoring stations, environmental sensor networks, electronic voting systems, in short anything connected through networks.

Leaving aside protecting data in e-commerce, a big target in itself, IoT data traffic is an enormous opportunity for information insurance.

Summary

The digital economy is creating new classes of assets which lend themselves to monetary protection – insurance being a useful and accepted approach. Protecting digital assets during their life cycle is a complex, multiple environment challenge. The numerous examples of cybertheft, ransomware, malware and data tampering amount to billions of dollars in annual losses.

As business moves towards decentralized, distributed and virtual computer resources like cloud services, the challenge is to manage and protect data in a continuously changing landscape. In this situation, the role of insurance for data is a promising market opportunity for insurers, just as in real world of physical assets and trading.

Calculating risk insurance for data can reuse existing models from finance risk as long as the mechanics of data access, storage, exchange and veracity are defined and measured. This provides insurers the ability to use such models in a mechanistic manner without arbitrary quantifying risk measures.

While we have not introduced the role of AI methods in this analysis, it is obvious that advanced statistical methods based on neural networks and classification will play an important role. Our explorations of the Black-Scholes, CARP and Brownian motion stochastic models are a step in that direction.

Author

Andre Szykier lives in Bermuda Dunes California, is a mathematician and founder of **UbiVault**, a leader in security solutions for a decentralized Web. He is also the CTO of a cryptocurrency ATM network – **BlockchainBTM**, and contributing scientist to **Aegis Health Analytics** in Washington DC.

Andre andre@ubivault.com

References

[1] The measure of **information entropy** associated with each possible data value is the negative logarithm of the probability mass function for the value:

$$S = -\sum_i P_i \log P_i$$

When a data source has a low-probability value due to a low-probability event, the event carries more information (***S***) than when it produces a high-probability value. The amount of information conveyed by each event becomes a random variable whose expected value is the entropy (disorder or uncertainty). Information entropy is analogous to entropy in statistical thermodynamics.

[2] Black Scholes Video (Introduction)

[3] Tsekov, Roumen (2013). "Brownian Markets". *Chin. Phys. Lett.* **30** (8): 088901. arXiv:1010.2061